Business Report on
Time Series Forecasting for Sparkling Wine Sales data

Yedupati Venkata Yamini

06 Aug 2023

Table of Contents:

List of tables:

List of figures:

Problem 1:
For this particular assignment, the data of different types of wine sales in the 20th century is to be analysed. Both of these data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

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

Ans:
Viewing the first and last 5 rows of given time series data:

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

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

Viewing the basic information of the dataset:

- The given data set has 187 rows and 2 columns including Month-Year column and the sales of Sparkling Wine during the months.
- Viewing the column and data types information of the features 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    datetime64[ns]
 1   Sparkling   187 non-null    int64         
dtypes: datetime64[ns](1), int64(1)
memory usage: 3.0 KB
```

Figure 1: Basic information of the dataset

Insights:
1. The columns present in the dataset in the initial glance contain monthly data of Sparkling Wine sales.

2. There are no null values present in the dataset.

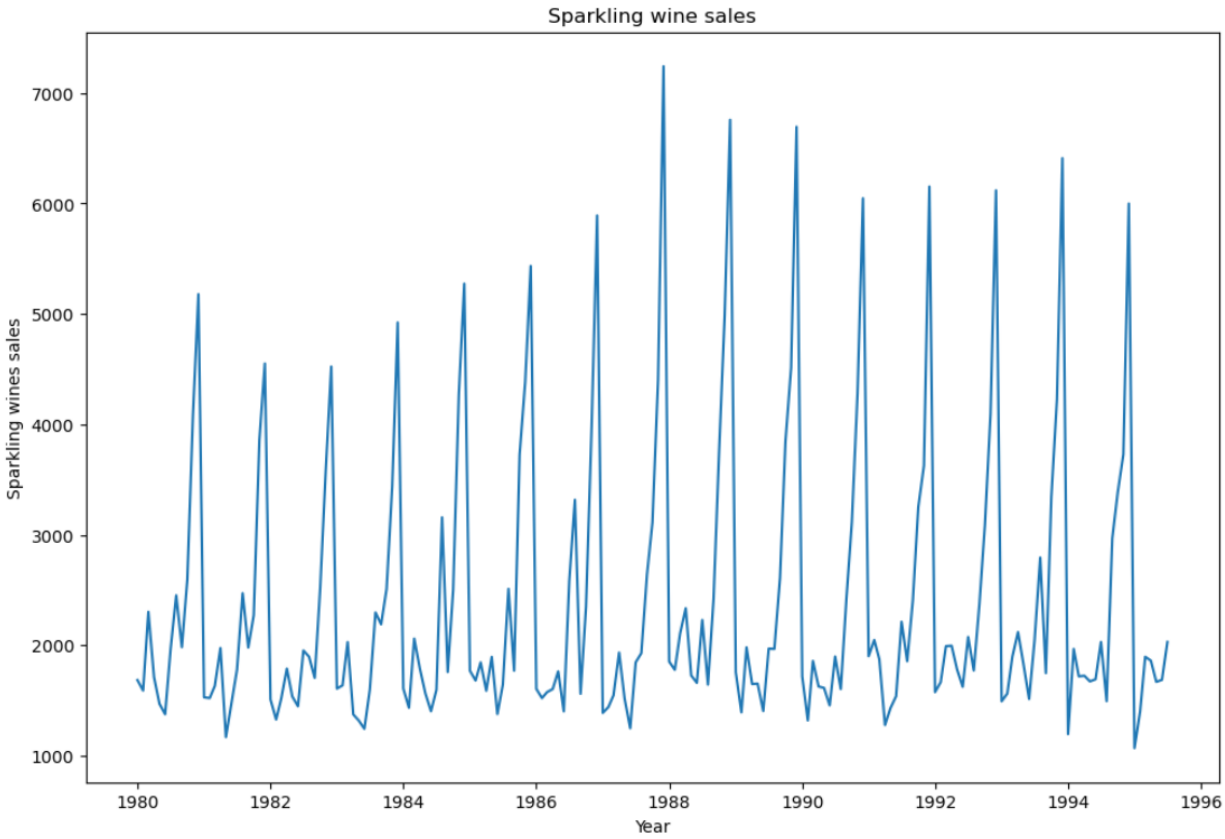
Viewing the summary of the dataset:

```
count     187.000000
mean     2402.417112
std      1295.111540
min      1070.000000
25%      1605.000000
50%      1874.000000
75%      2549.000000
max      7242.000000
Name: Sparkling, dtype: float64
```

Table 4: Summary of the dataset

Insights:
1. Average sales of sparkling wine over the years is around 2400$ monthly.
2. Highest sales recorded for a month is 7242$.

Plotting the time series data:



Insights:

1. The wine sales does not show any increasing/decreasing trend in terms of Sparkling Wine sales over the years.
2. However, there is a strong seasonality observed in the plot. The sales of wine seem to reach peaks during particular times of the year.

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

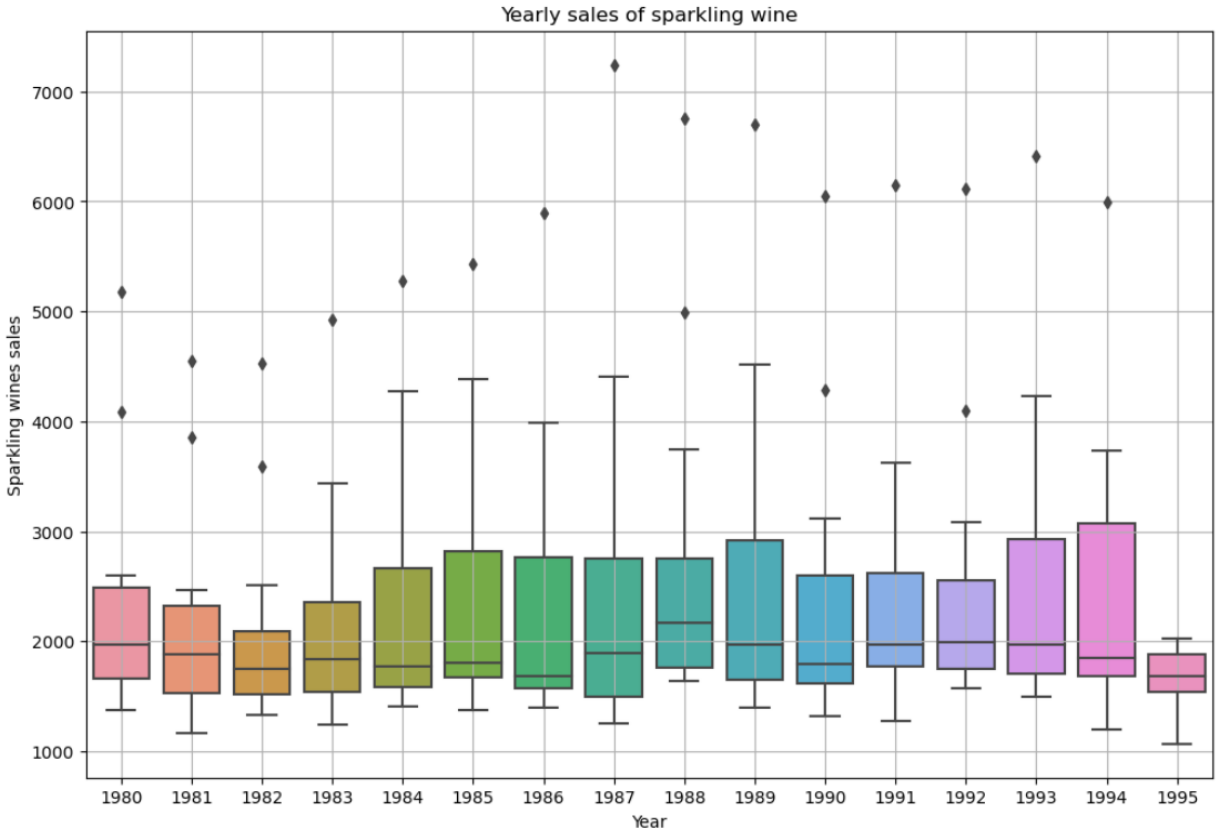
Univariate analysis:

```
count     187.000000
mean     2402.417112
std      1295.111540
min      1070.000000
25%      1605.000000
50%      1874.000000
75%      2549.000000
max      7242.000000
Name: Sparkling, dtype: float64
```

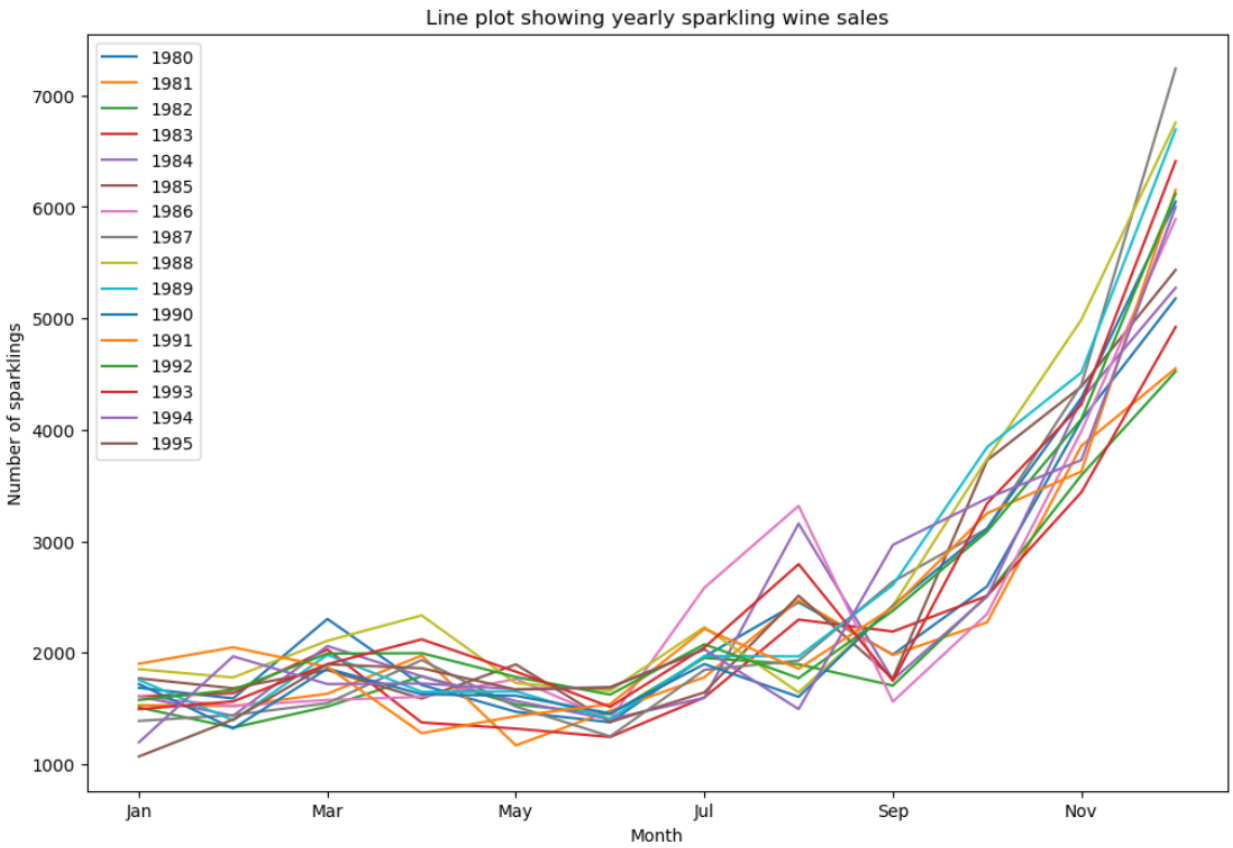
Insights:
1. Average sales of sparkling wine over the years is around 2400$ monthly.
2. Highest sales recorded for a month is 7242$.

Bivariate analysis:

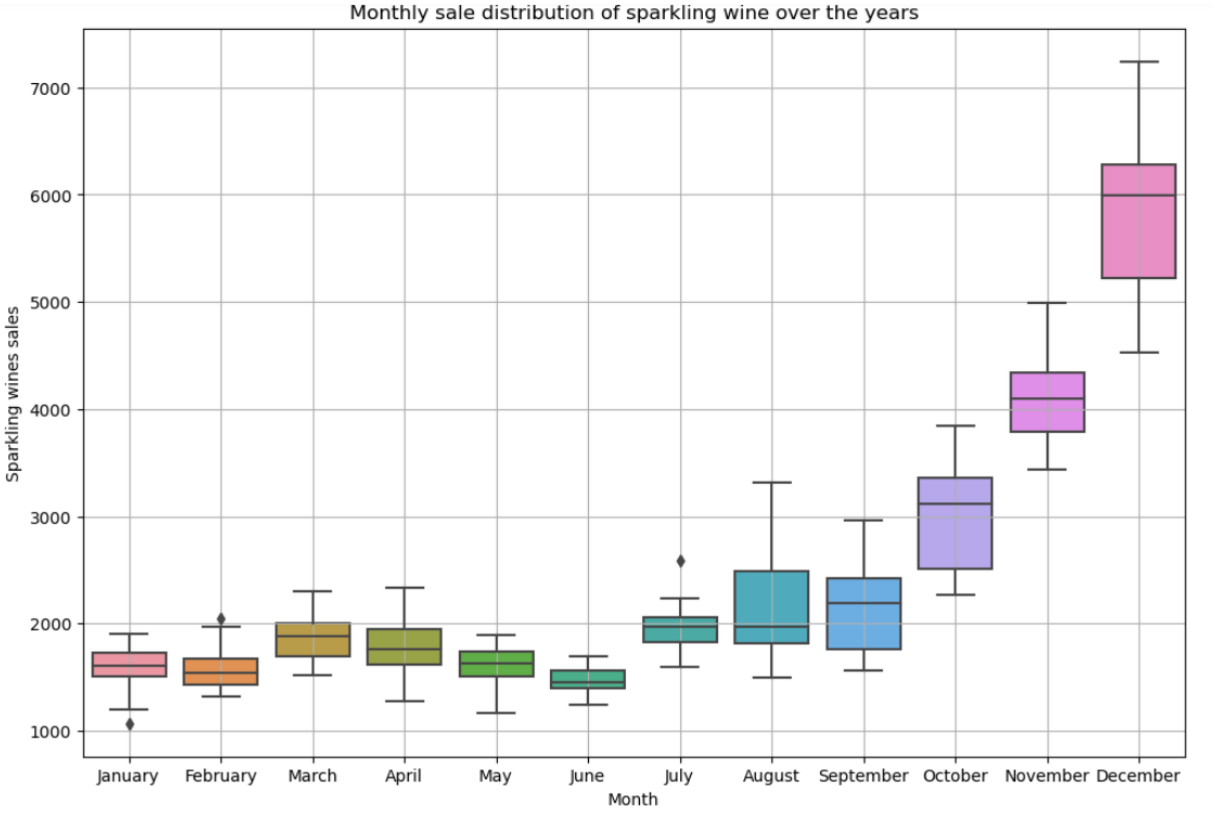
Yearly sales of Sparkling wine:



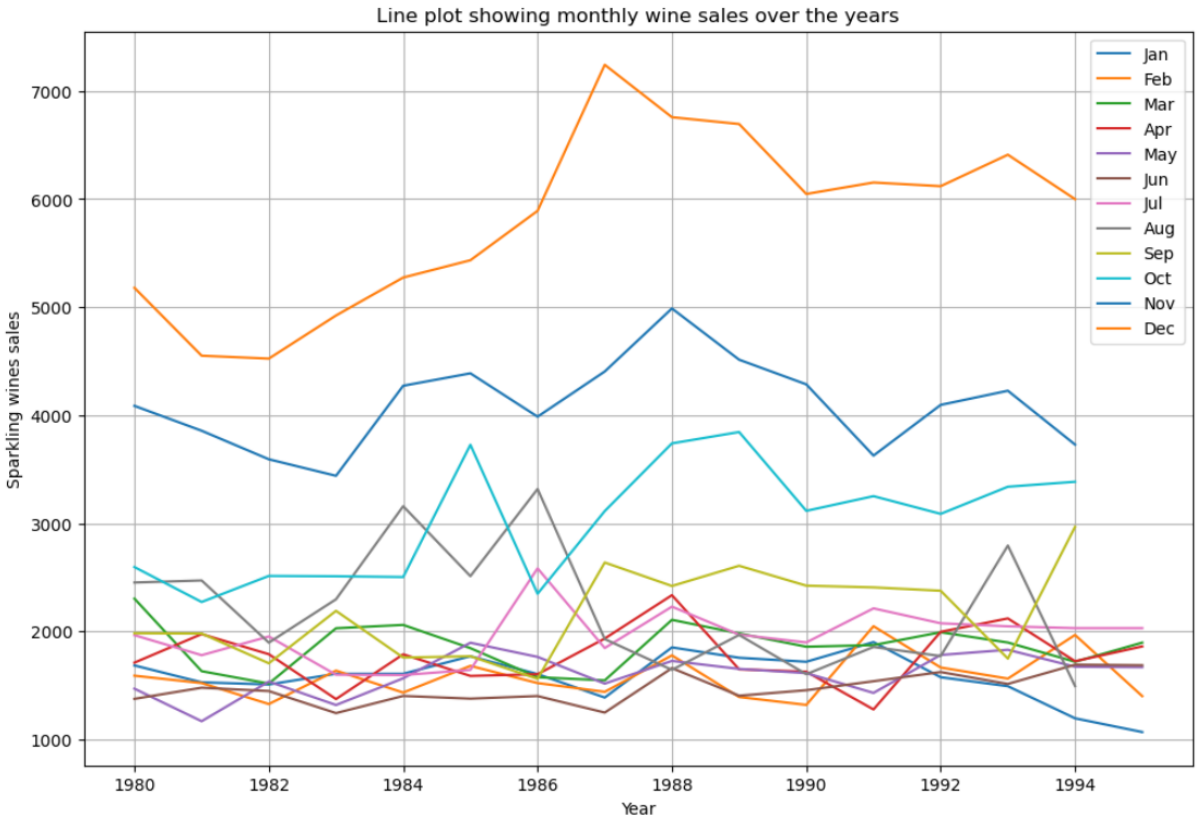
- Sales for the wine have not seen drastic changes over the years and have almost followed a linear horizontal trend.
- The same can be visualized in a line plot as shown below



Monthly sale distribution of sparkling wine:



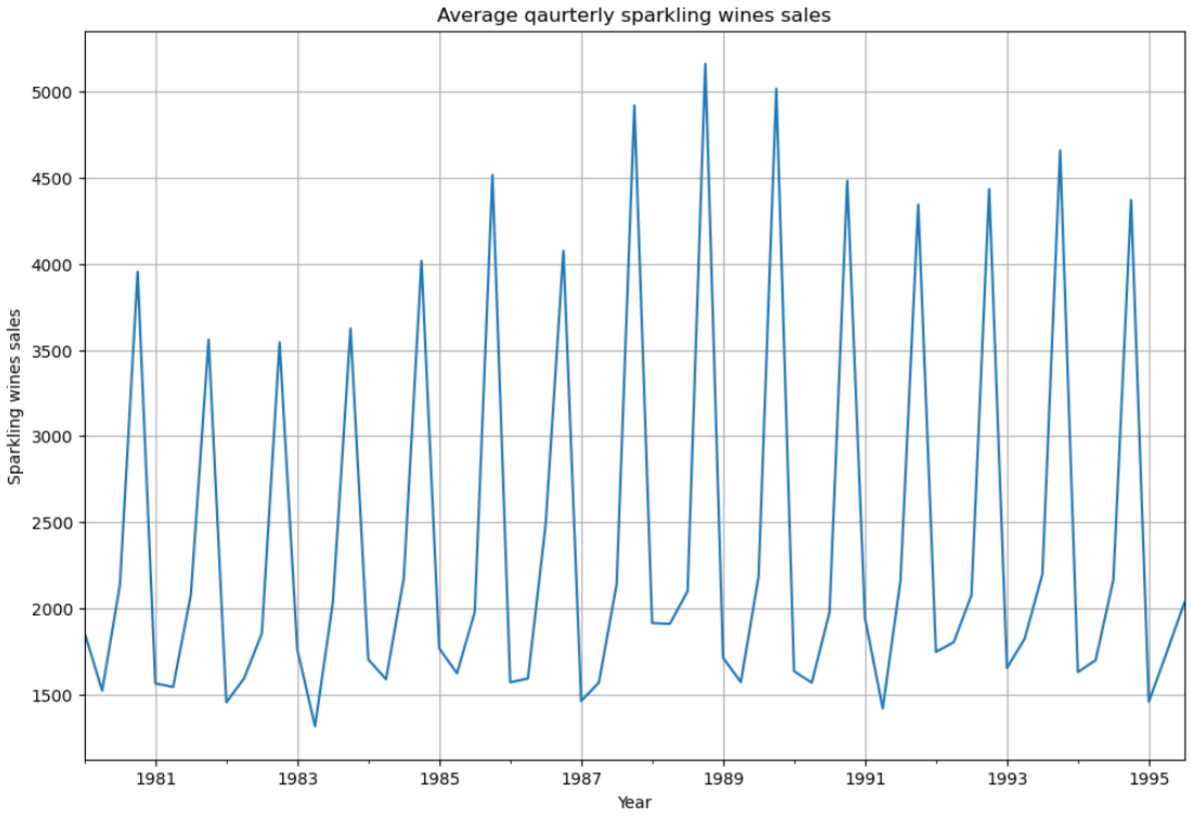
- Every year, the company is experiencing a huge spike in Wine sales during the end of the year due to occasions like Christmas, New Year etc.
- The same can be visualized in a line plot as shown below:



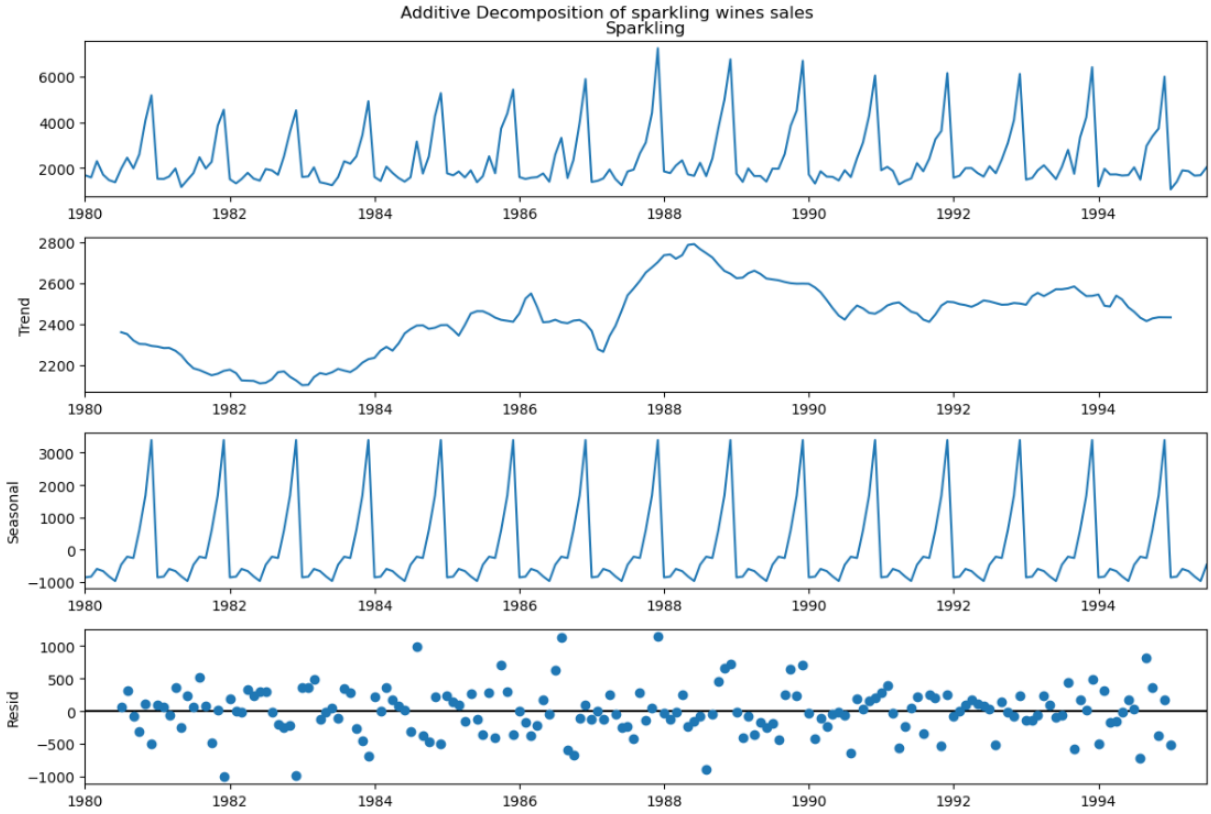
Distribution of the wine sales indexed by months and years:

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

Average quarterly sparkling wines sales:

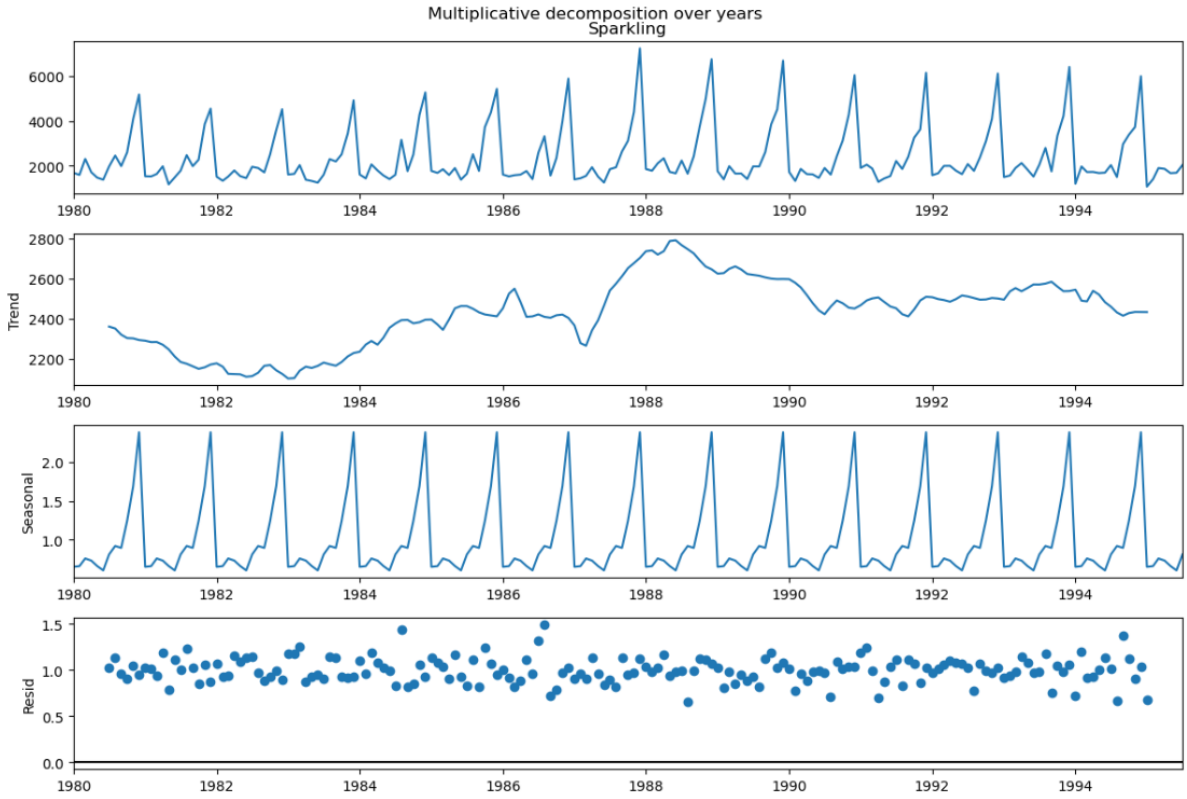


Additive seasonal decomposition:



- The initial glance on the time series plot shows us no significant increasing or decreasing trend.
- But the data shows clear signs of seasonality.
- Wine sales seem to be increasing at the end of every year and low at the start of every year.
- The graph with the 'Trend' label shows the linear trend of the sales over the years.
- As expected, it is neither continuously increasing nor continuously decreasing but is fluctuating.
- The seasonal component has also been captured in the third sub graph.
- The residual ideally should not show any pattern. But the residual graph for additive seasonal decomposition shows some pattern, which indicates that the data is not following additive seasonality. Seasonality trend must have been multiplicatively increasing over the years.

Multiplicative seasonal decomposition:



- Wine sales seem to be increasing at the end of every year and low at the start of every year.
- The graph with the 'Trend' label shows the linear trend of the sales over the years.
- As expected, it is neither continuously increasing nor continuously decreasing but is fluctuating.
- The seasonal component has also been captured in the third sub graph.
- The error or residual component does not show any pattern. It indicates that the data follows multiplicative seasonality.

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

The data has been split into training and test data.
Date range for training time series: Jan-1980 to Dec-1990 (132 months)
Data range for test time series: Jan-1991 to July-1995 (55 months)

First and last rows of training and test data set:

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

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

YearMonth	Sparkling
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

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

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.**

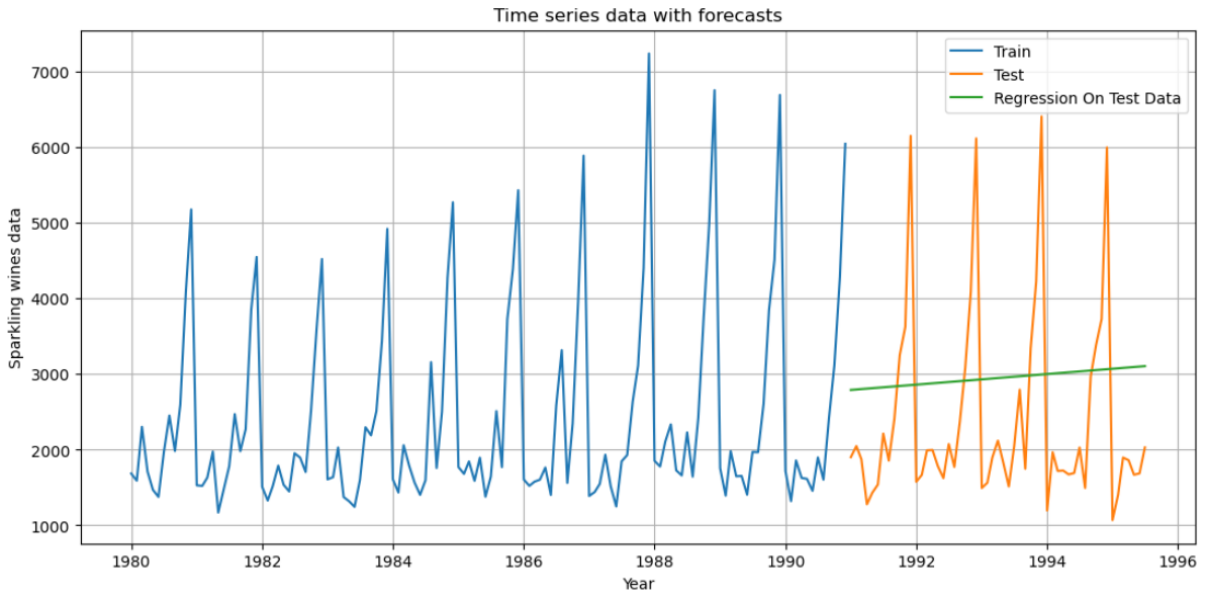
4.1 Linear regression model:

- A plain linear regression model has been applied to the training data set.
- All default parameters are used.
    - fit_intercept=True,
    - normalize='deprecated',
    - copy_X=True,
    - n_jobs=None,
    - positive=False
- Predictions done by the linear regression model:

:

YearMonth	Sparkling	time	RegOnTime
1991-01-01	1902	133	2791.652093
1991-02-01	2049	134	2797.484752
1991-03-01	1874	135	2803.317410
1991-04-01	1279	136	2809.150069
1991-05-01	1432	137	2814.982727
1991-06-01	1540	138	2820.815386
1991-07-01	2214	139	2826.648044
1991-08-01	1857	140	2832.480703
1991-09-01	2408	141	2838.313361
1991-10-01	3252	142	2844.146020
1991-11-01	3627	143	2849.978678

-
- Root mean squared error for the predictions obtained on this model(on test data): 1389.135175
- Visualizing the line built by linear regression:



-

- The equation built by linear regression model is 2015.908 + (5.832 * time_step)

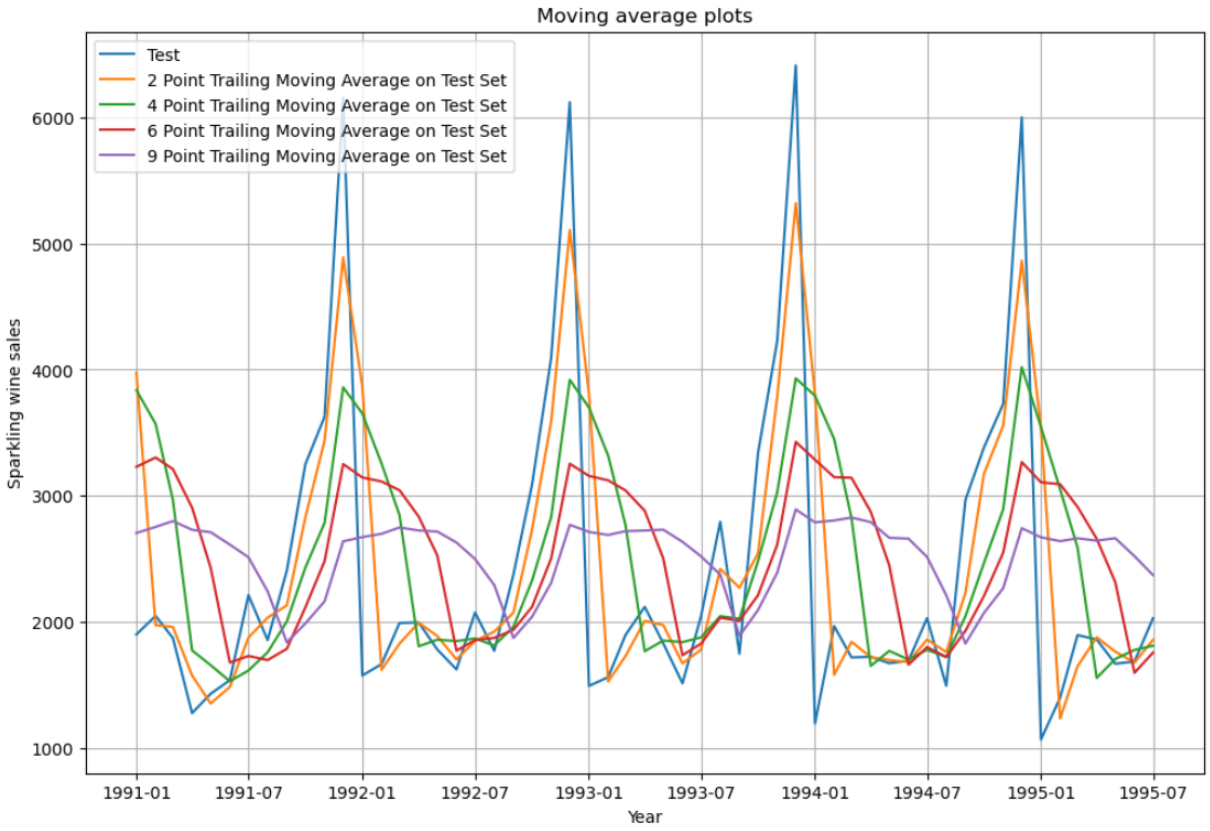
4.2 Moving Average(MA) model:

- For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals.
- The best interval can be determined by the maximum accuracy (or the minimum error) over here.
- The data constructed with different moving averages for the original is shown below

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

YearMonth	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1995-03-01	1897	1649.5	2592.00	2913.666667	2664.000000
1995-04-01	1862	1879.5	1557.75	2659.833333	2645.222222
1995-05-01	1670	1766.0	1707.75	2316.666667	2664.666667
1995-06-01	1688	1679.0	1779.25	1598.166667	2522.444444
1995-07-01	2031	1859.5	1812.75	1758.333333	2372.000000

The plots for different moving averages on the test sset is shown below:

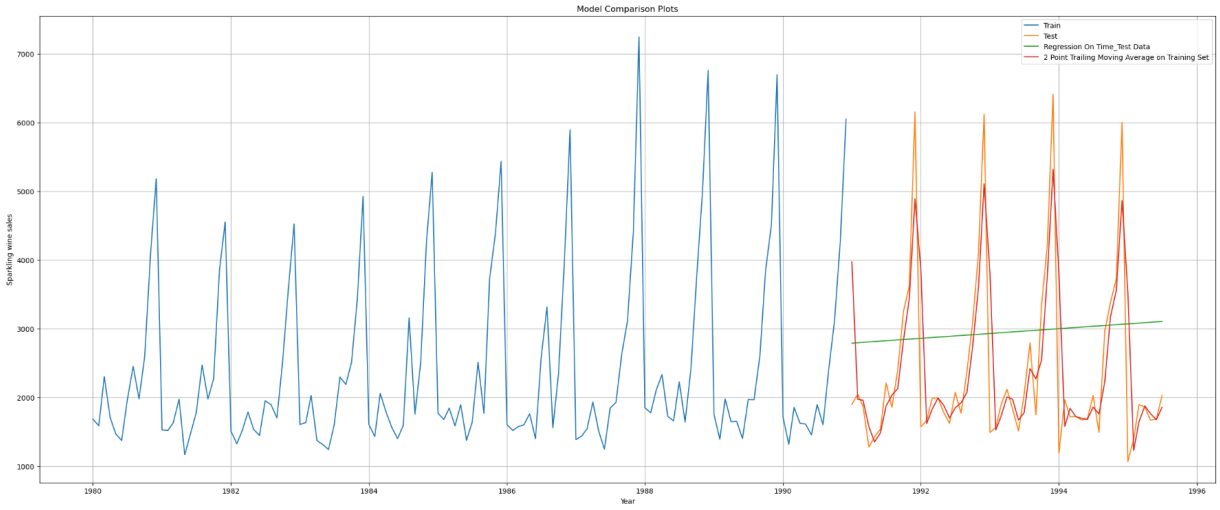


RMSE values for different averages:

	Test RMSE
RegressionOnTime	1389.135175
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315

- As we can see the 2 point trailing rolling average has the least error term compared to other rolling averages.
- This indicates that the data is mostly predictable using 2 most recent instances.
- Plot showing linear regression forecast and 2 point moving average forecast on test set:

-



4.3 Single exponential smoothing:

Simple Exponential Smoothing is a time series forecasting method used to predict future values in a time series by giving more weight to recent observations while assigning exponentially decreasing weights to past observations. It is particularly useful for time series data that exhibit a trend or seasonality.

Mathematical Formula:

The formula for calculating the forecasted value using Simple Exponential Smoothing is as follows:

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

Where:

- F_{t+1} is the forecasted value for the next time period $t + 1$.
- Y_t is the actual value at time period t.
- F_t is the forecasted value for the current time period t.
- α is the smoothing parameter, a value between 0 and 1 that determines the weight given to the most recent observation. A smaller α gives more weight to past observations, while a larger α gives more weight to the most recent observation.

- Simple exponential smoothing has been applied with optimized approach and a brute force approach

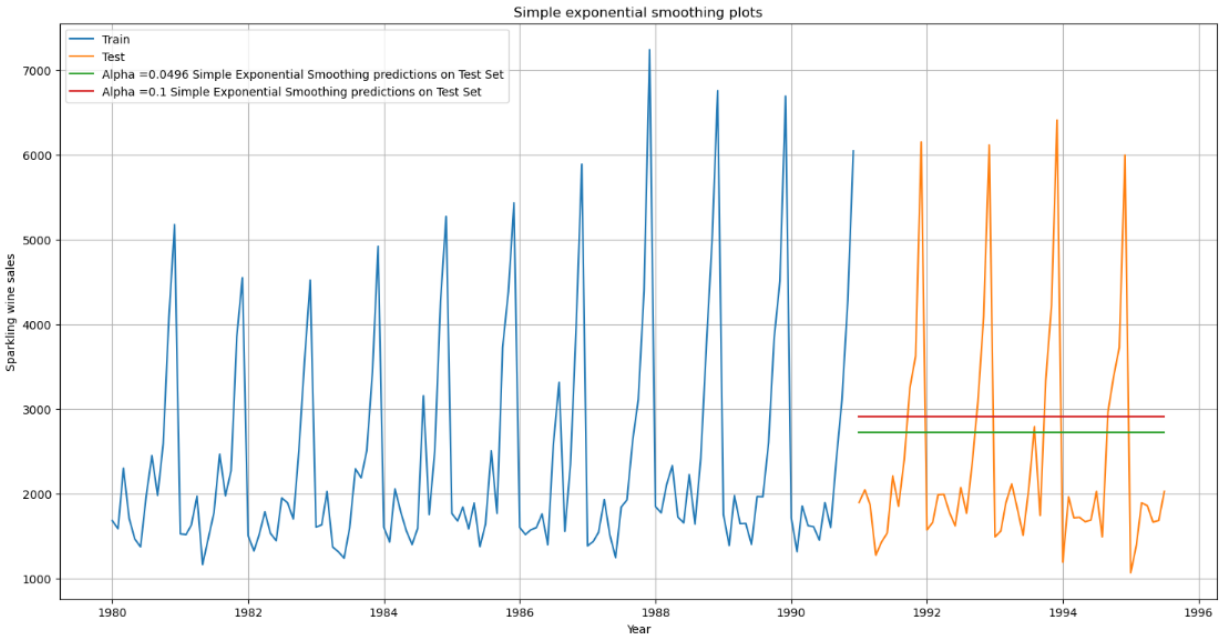
- The value for alpha obtained with optimized approach: 0.0496
- Parameters used:
    - {'smoothing_level': 0.049607360581862936,
    - 'smoothing_trend': nan,
    - 'smoothing_seasonal': nan,
    - 'damping_trend': nan,
    - 'initial_level': 1818.535750008871,
    - 'initial_trend': nan,
    - 'initial_seasons': array([], dtype=float64),
    - 'use_boxcox': False,
    - 'lamda': None,
    - 'remove_bias': False}

- The value of alpha obtained with brute force approach: 0.1
- RMSE values for the simple exponential smoothing applied on the dataset with the above alpha values:

	Test RMSE
RegressionOnTime	1389.135175
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.0496,SimpleExponentialSmoothing	1316.035487
Alpha=0.1,SimpleExponentialSmoothing	1375.393398

-

Plotting the forecast obtained by above simple exponential smoothing equations:



- It is observed that the model performs better when alpha is 0.0496 as compared to 0.1.

4.4 Double exponential smoothing

Double Exponential Smoothing, also known as Holt's Linear Exponential Smoothing, is an extension of Simple Exponential Smoothing that takes into account both the level and trend of a time series. This method is particularly useful when the time series exhibits a linear trend.

Mathematical Formula:
The formula for calculating the forecasted value using Double Exponential Smoothing is as follows:

1. **Level Smoothing:**

   $$L_t = \alpha \cdot Y_t + (1 - \alpha) \cdot (L_{t-1} + T_{t-1})$$

   Where:

   - L_t is the smoothed level (or the estimate of the level) at time t.
   - Y_t is the actual value at time t.
   - L_{t-1} is the smoothed level at time $t - 1$.
   - T_{t-1} is the trend at time $t - 1$.
   - α is the smoothing parameter for the level, similar to Simple Exponential Smoothing.

2. **Trend Smoothing:**

   $$T_t = \beta \cdot (L_t - L_{t-1}) + (1 - \beta) \cdot T_{t-1}$$

   Where:

   - T_t is the smoothed trend (or the estimate of the trend) at time t.
   - β is the smoothing parameter for the trend.

3. **Forecast Calculation:**

   $$F_{t+h} = L_t + h \cdot T_t$$

   Where:

   - F_{t+h} is the forecasted value for h periods beyond time t.

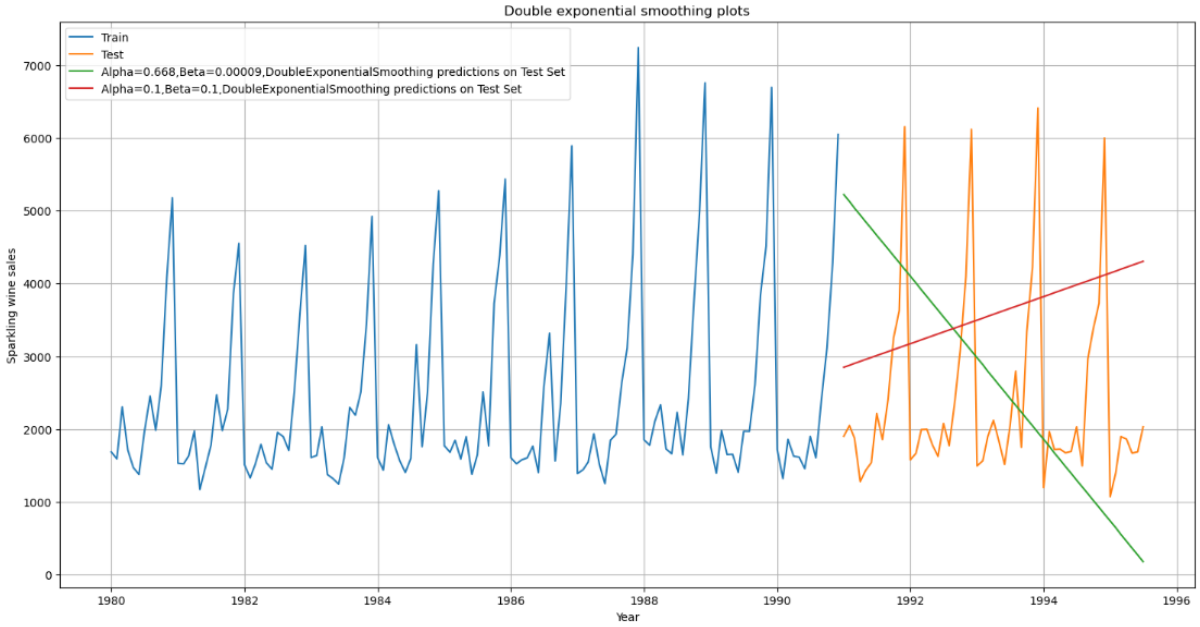
- Double exponential smoothing has been applied with optimized approach and a brute force approach
- The value for alpha obtained with optimized approach: 0.688
- The value for beta obtained with optimized approach: 0.00009
    - {'smoothing_level': 0.6885714285714285,
    - 'smoothing_trend': 9.999999999999999e-05,
    - '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}

- The value of alpha obtained with brute force approach: 0.1
- The value of beta obtained with brute force approach: 0.1

- RMSE values for the simple exponential smoothing applied on the dataset with the above alpha values:

	Test RMSE
RegressionOnTime	1389.135175
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.0496,SimpleExponentialSmoothing	1316.035487
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha =0.688, Beta = 0.00009,DoubleExponentialSmoothing	1316.035487
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670

Plotting the forecast obtained by above simple exponential smoothing equations:



- It is observed that the model performs better when alpha - 0.688 and beta - 0.0009 with RMSE value of 1316.035

- However, till the current status, 2 point moving average seems to give the best performance.

4.5 Triple exponential smoothing:
Triple Exponential Smoothing, also known as Holt-Winters Exponential Smoothing, is an extension of Double Exponential Smoothing that includes a seasonality component. This method is particularly useful when the time series exhibits trend and seasonality.

Mathematical Formula:
The formula for calculating the forecasted value using Triple Exponential Smoothing is as follows:

1. **Level Smoothing:**
   $$L_t = \alpha \cdot (Y_t - S_{t-m}) + (1 - \alpha) \cdot (L_{t-1} + T_{t-1})$$
2. **Trend Smoothing:**
   $$T_t = \beta \cdot (L_t - L_{t-1}) + (1 - \beta) \cdot T_{t-1}$$
3. **Seasonal Smoothing:**
   $$S_t = \gamma \cdot (Y_t - L_t) + (1 - \gamma) \cdot S_{t-m}$$
4. **Forecast Calculation:**
   $$F_{t+h} = L_t + h \cdot T_t + S_{t-m+h_m}$$

Where:

- L_t is the smoothed level (or the estimate of the level) at time t.
- T_t is the smoothed trend (or the estimate of the trend) at time t.
- S_t is the smoothed seasonal component (or the estimate of the seasonality) at time t.
- Y_t is the actual value at time t.
- S_{t-m} represents the seasonal component at time $t - m$, where m is the number of seasons (e.g., for monthly data, $m = 12$ for yearly seasonality).
- h is the number of periods ahead for forecasting.
- h_m is the corresponding seasonal index for h.

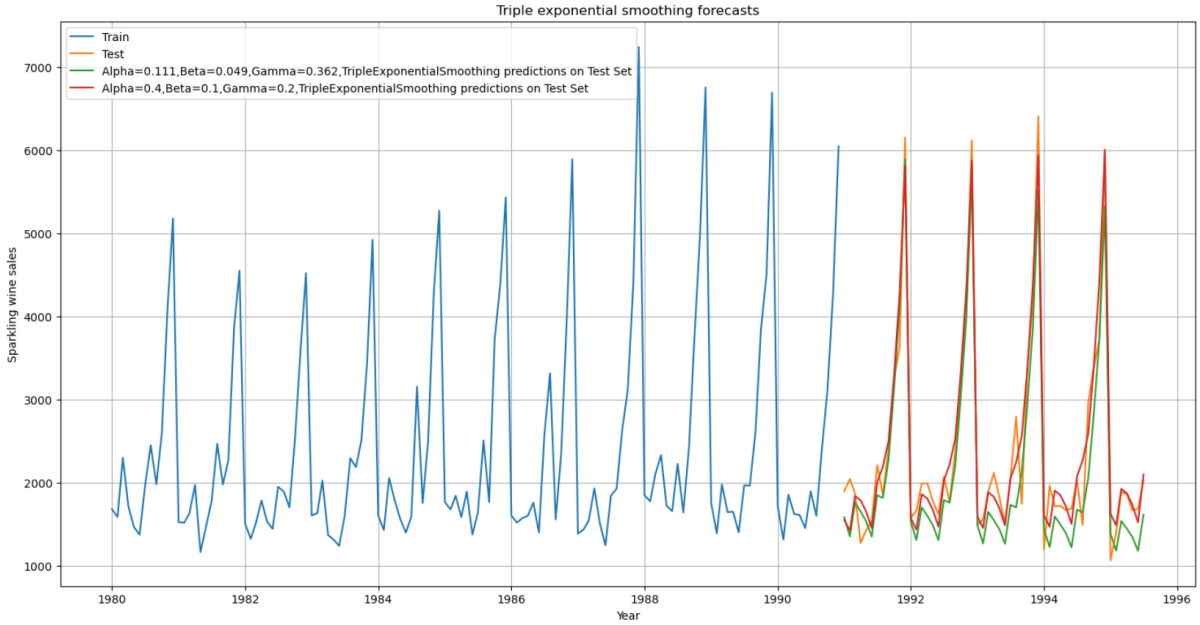
- Triple exponential smoothing has been applied with optimized approach and a brute force approach
- The value for alpha obtained with optimized approach: 0.111

- The value for beta obtained with optimized approach: 0.049
- The value for beta obtained with optimized approach: 0.362
- Parameters used:
    - {'smoothing_level': 0.11133818361298699,
    - 'smoothing_trend': 0.049505131019509915,
    - 'smoothing_seasonal': 0.3620795793580111,
    - 'damping_trend': nan,
    - 'initial_level': 2356.4967888704355,
    - 'initial_trend': -10.187944726007238,
    - 'initial_seasons': array([0.71296382, 0.68242226, 0.90755008, 0.80515228, 0.65597218,
    - 0.65414505, 0.88617935, 1.13345121, 0.92046306, 1.21337874,
    - 1.87340336, 2.37811768]),
    - 'use_boxcox': False,
    - 'lamda': None,
    - 'remove_bias': False}
- The value of alpha obtained with brute force approach: 0.4
- The value of beta obtained with brute force approach: 0.1
- The value of beta obtained with brute force approach: 0.2
- RMSE values for the simple exponential smoothing applied on the dataset with the above alpha values:

:

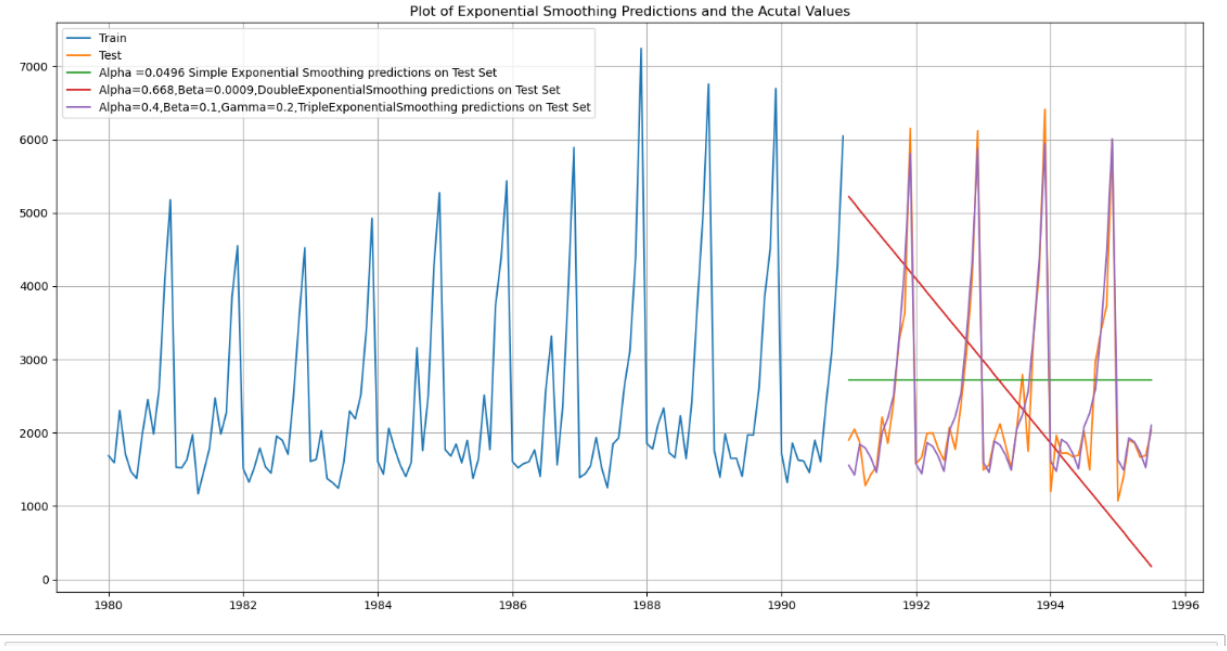
	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothing	404.286809
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
Alpha=0.0496,SimpleExponentialSmoothing	1316.035487
Alpha =0.688, Beta = 0.00009,DoubleExponentialSmoothing	1316.035487
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670

Plotting the forecast obtained by above simple exponential smoothing equations:



- It is observed that the model performs better when alpha - 0.4, beta - 0.1 and gamma - 0.2 with RMSE value of 317.434

Plotting the best variants of all exponential smoothing models built till now:



- Till the current status, triple exponential smoothing with alpha - 0.4, beta - 0.1 and gamma - 0.2 seems to give the best performance with RMSE 317.434

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**

- Models like ARIMA and SARIMA work on data that is stationary. Data is said to be stationary when it shows no trend and significant variance in the distribution of data over time.
- If data is found to be non-stationary we use differencing to make the data stationary, if the data shows significant variance, we use transformations like log transformation in order to stabilize the magnitude and variance of the data.
- Dicky Fuller Test on the timeseries is run to check for stationarity of data.

    - **Null Hypothesis** *H*0: Time Series is non-stationary.
    - **Alternate Hypothesis** ***Ha***: Time Series is stationary.

- *So Ideally if p-value < 0.05 then null hypothesis: TS is non-stationary is rejected else the TS is non-stationary is failed to be rejected .*
- Applying a dickey-fuller test on the data gives the following results. (considering 0.05 to be the level of confidence)

```
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
dtype: float64
```

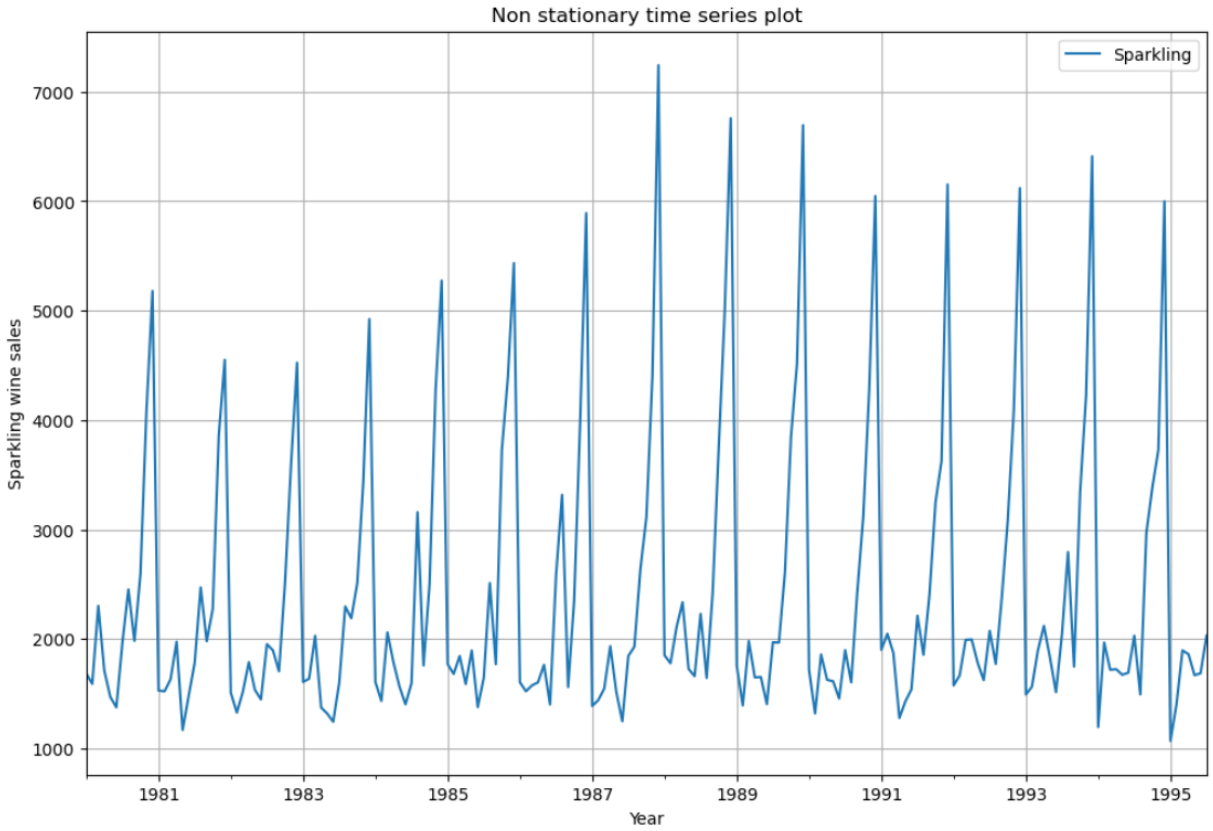
- P-value is > 0.05 which means we fail to reject null hypothesis. The data is non stationary.

- Applying differencing on the data and verifying for stationarity (considering 0.05 to be the level of confidence)

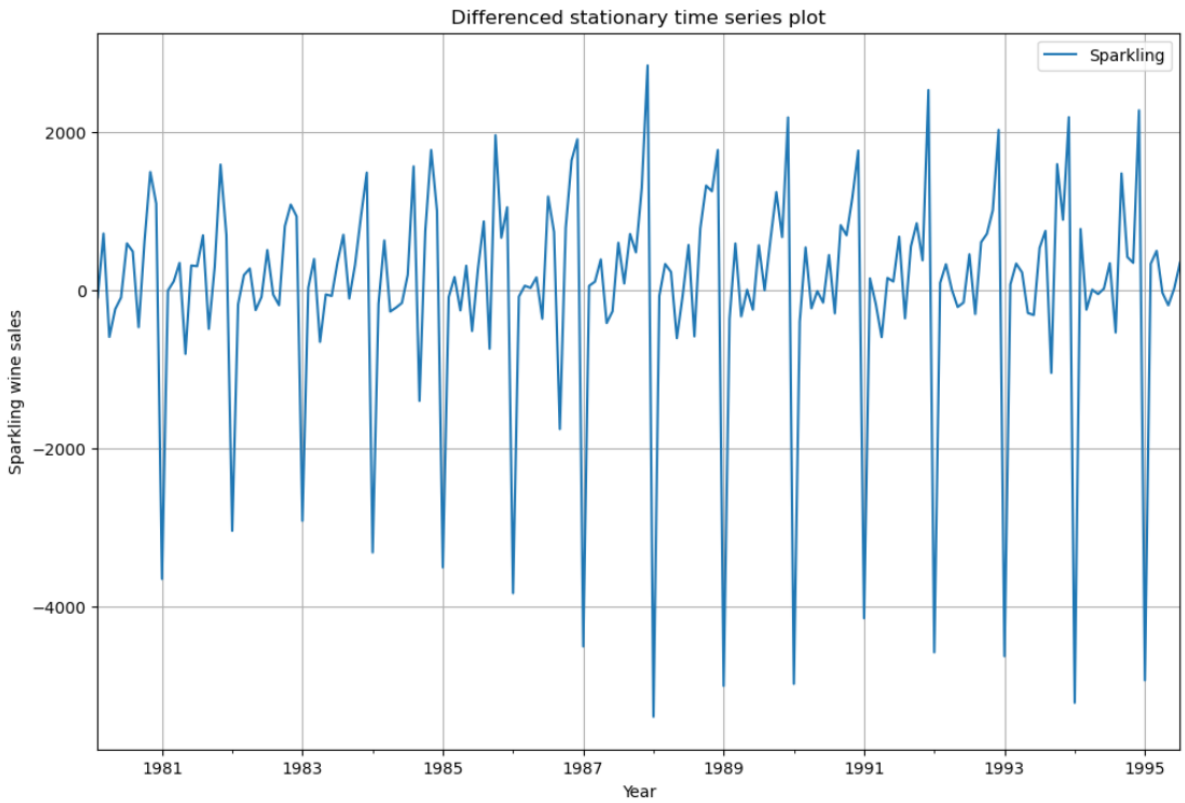
```
Test Statistic                  -45.050301
p-value                           0.000000
#Lags Used                       10.000000
Number of Observations Used     175.000000
Critical Value (1%)              -3.468280
Critical Value (5%)              -2.878202
Critical Value (10%)             -2.575653
dtype: float64
```

-
- P-value obtained now is 0 which means we reject null hypothesis.
- Applied first order differencing on the data and the data is now stationary.

Plot of non stationary time series data:



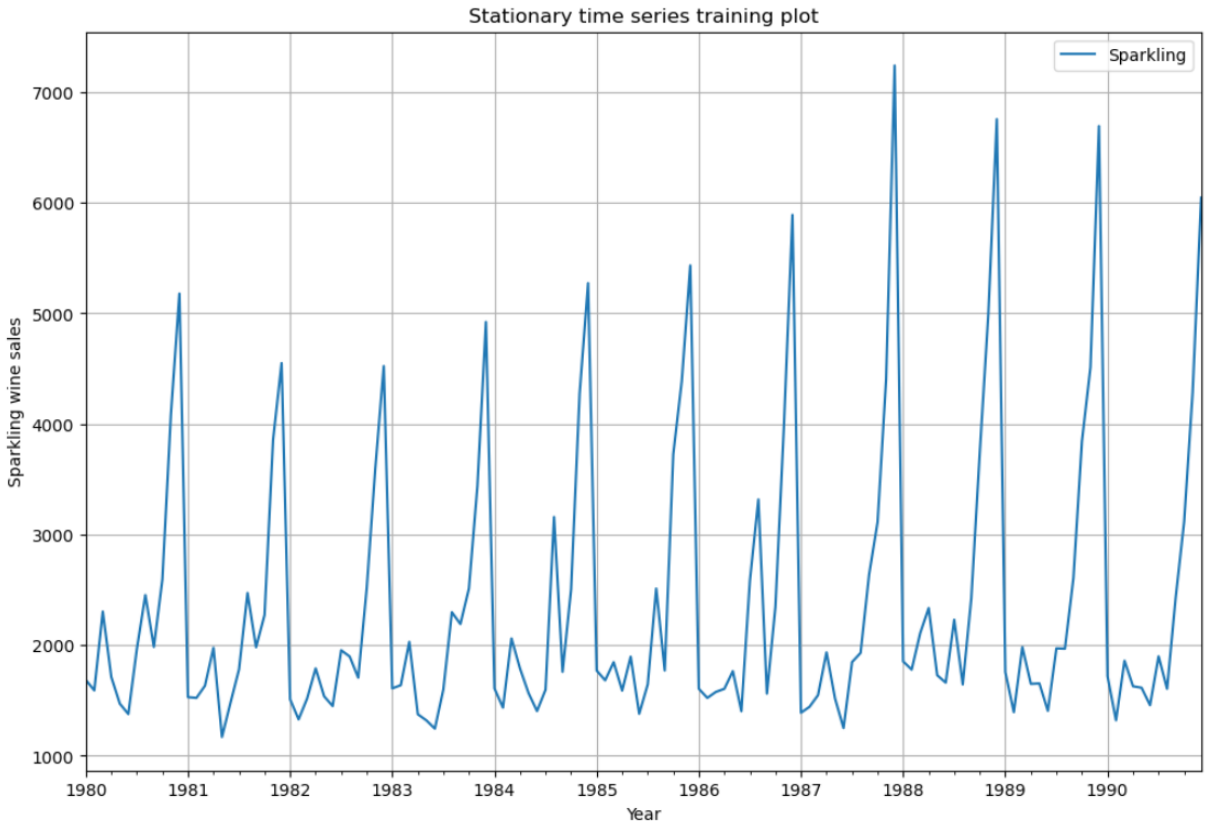
Plot of first order differenced time series data:



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.**

Prerequisite: The training data has to be determined to be stationary or not.

- Given training time series data is stationary.
- Plotting the distribution of stationary time series data:



- First and last rows of training and test data:

First few rows of Training Data

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

First few rows of Test Data

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

Last few rows of Training Data

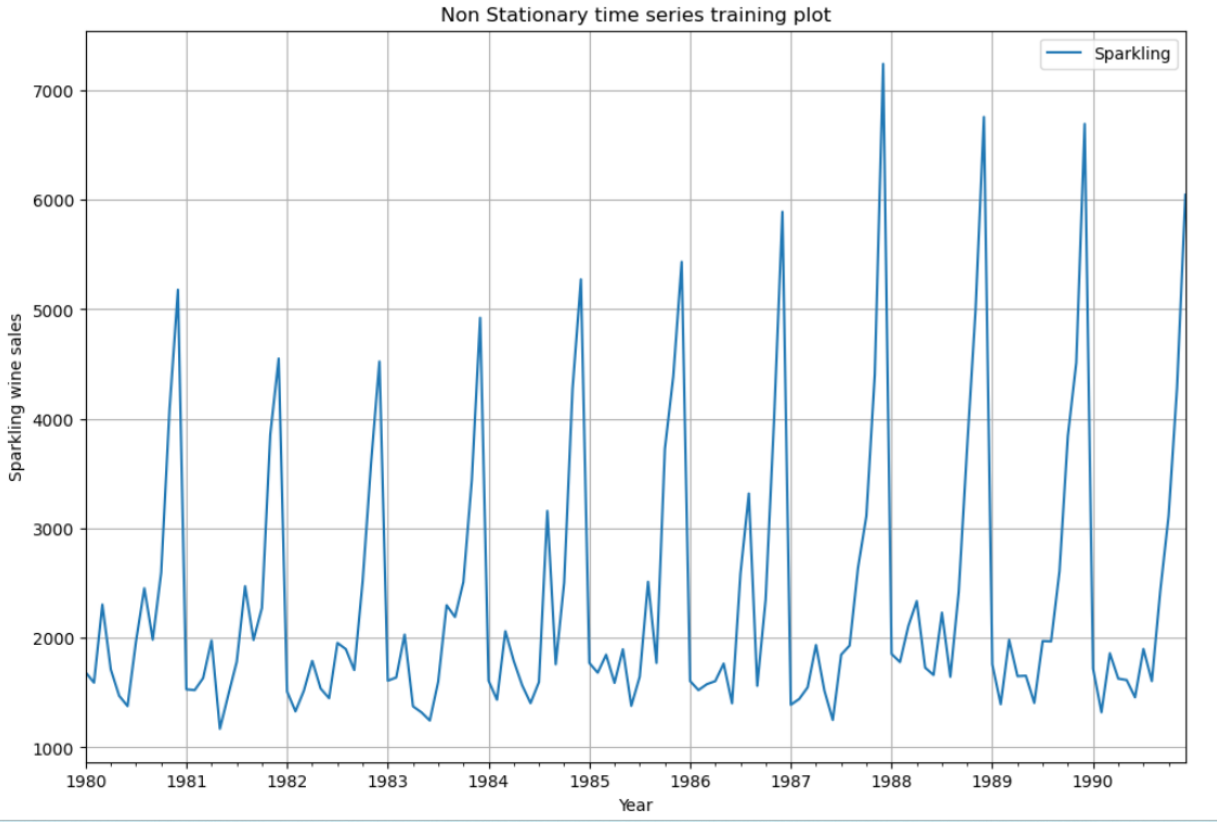
YearMonth	Sparkling
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

Last few rows of Test Data

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

-

- Training time series plot:



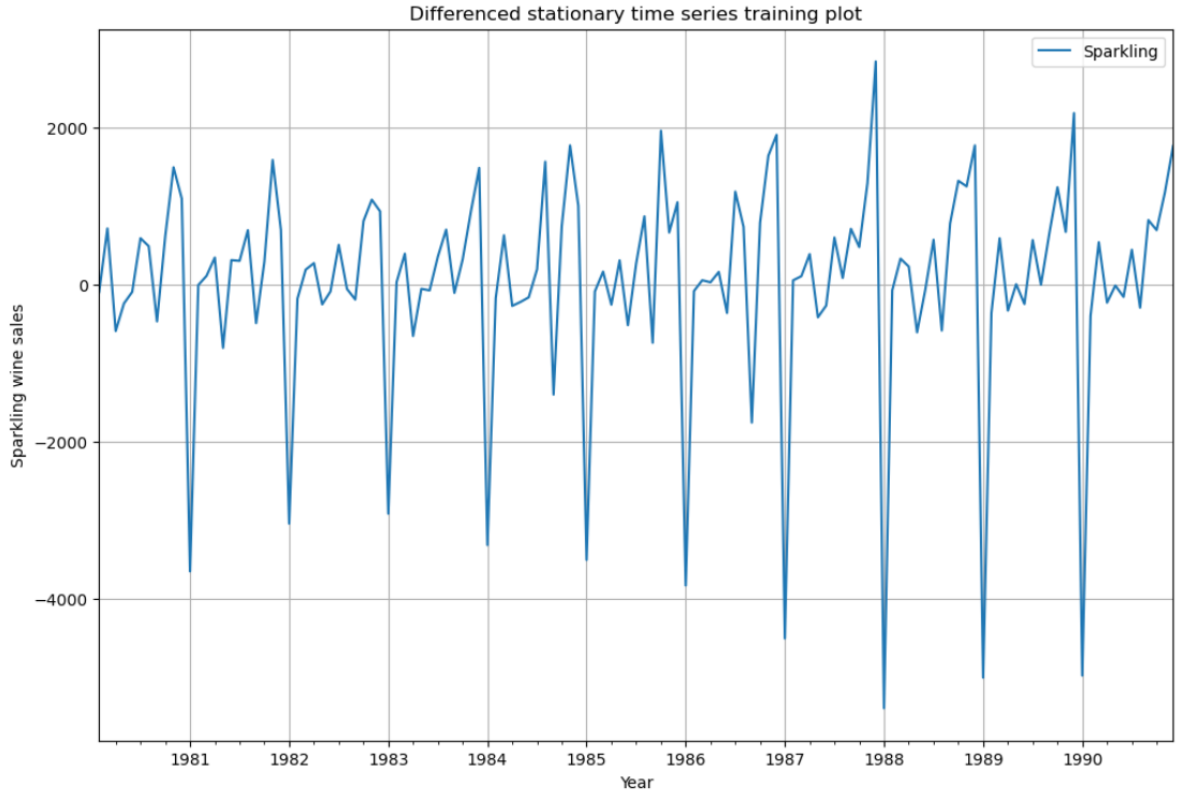
- Applying dickey-fuller test on training data:

```
Test Statistic                  -1.208926
p-value                          0.669744
#Lags Used                      12.000000
Number of Observations Used    119.000000
Critical Value (1%)             -3.486535
Critical Value (5%)             -2.886151
Critical Value (10%)            -2.579896
dtype: float64
```

-
- The p-value obtained is > 0.05 which means the training time series is non stationary. Need to make the data stationary in order for ARIMA and SARIMA models to use it.
- Applied first order differencing on the training data.
- Applied dickey - fuller test on training data after differencing

```
Test Statistic                 -8.005007e+00
p-value                         2.280104e-12
#Lags Used                      1.100000e+01
Number of Observations Used     1.190000e+02
Critical Value (1%)            -3.486535e+00
Critical Value (5%)            -2.886151e+00
Critical Value (10%)           -2.579896e+00
dtype: float64
```

- Since p-value is < 0.05, we reject null hypothesis and time series is considered to be stationary.
- Differenced stationary time series training plot:



- We use the training dataset directly because we are using ARIMA and SARIMA models which have an inbuilt option to difference the time series.

ARIMA model:

- ARIMA:- **Auto Regressive Integrated Moving Average** is a way of modeling time series data for forecasting or predicting future data points.

- ARIMA:- **Auto Regressive Integrated Moving Average** is a way of modeling time series data for forecasting or predicting future data points.
- Improving AR Models by making Time Series stationary through Moving Average Forecasts
- ARIMA models consist of 3 components:-
    - **AR model:** The data is modeled based on past observations.
    - **Integrated component:** Whether the data needs to be differenced/transformed.

- For the given data set, we consider p,d,q to represent level component, differencing component and trend component respectively.
- p: The number of lag observations included in the model (order of autoregressive terms).
- d: The degree of differencing, which is the number of times the data is differenced (order of integration).
- q: The size of the moving average window (order of moving average terms).
- Ranges considered for p,d,q are (0,5).
- An automated ARIMA model has been built with the following combinations/parameters.
    - AR component(p): Range(1,5)
    - MA component(q): Range(1,5)
    - Differencing component(d): 1
    - Enforce_stationarity: False (not to enforce stationarity on the AR components of the model)
    - Enforce_intvertibility: False(not to enforce invertibility on the MA components on the model)
- It notes the AIC scores for the data with each set of p,q,d values.

- The top 5 combinations with least AIC scores obtained are

:

	param	AIC
104	(4, 0, 4)	2192.432126
79	(3, 0, 4)	2201.407315
53	(2, 0, 3)	2205.695049
78	(3, 0, 3)	2209.251589
103	(4, 0, 3)	2211.840116

- Therefore, best hyper parameters for p,d,q for the given data set are taken as 4,0,4
- AIC metric for this combination is 2192.057
- Applying ARIMA on the model with the above obtained parameters.
- Summary obtained:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(4, 0, 4)   Log Likelihood               -1086.216
Date:                Sat, 05 Aug 2023   AIC                           2192.432
Time:                        12:17:13   BIC                           2221.260
Sample:                    01-01-1980   HQIC                          2204.147
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       2403.7689     99.522     24.153      0.000    2208.710    2598.828
ar.L1          0.6996      0.133      5.277      0.000       0.440       0.959
ar.L2          0.0574      0.185      0.310      0.756      -0.305       0.420
ar.L3          0.2193      0.169      1.297      0.195      -0.112       0.551
ar.L4         -0.6910      0.125     -5.528      0.000      -0.936      -0.446
ma.L1         -0.5145      0.178     -2.890      0.004      -0.863      -0.166
ma.L2         -0.4359      0.079     -5.545      0.000      -0.590      -0.282
ma.L3         -0.4496      0.126     -3.558      0.000      -0.697      -0.202
ma.L4          0.9608      0.147      6.530      0.000       0.672       1.249
sigma2      9.785e+05      0.000   9.71e+09      0.000    9.78e+05    9.78e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.37   Jarque-Bera (JB):                17.95
Prob(Q):                              0.54   Prob(JB):                         0.00
Heteroskedasticity (H):               2.28   Skew:                             0.61
Prob(H) (two-sided):                  0.01   Kurtosis:                         4.34
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 9.23e+27. Standard errors may be unstable.
```

- The equation obtained by ARIMA model can be written as follows:

ARIMA(4,0,4)

$$= (0.6996 * Yt - 1 + 0.0574 * Yt - 2 + 0.2193 * Yt - 3 - 0.6910 * Yt - 4)$$

- $(-0.5145 * et-1 - 0.4359 * et-2 - 0.4496 * et-3 + 0.9608 * et-4)$

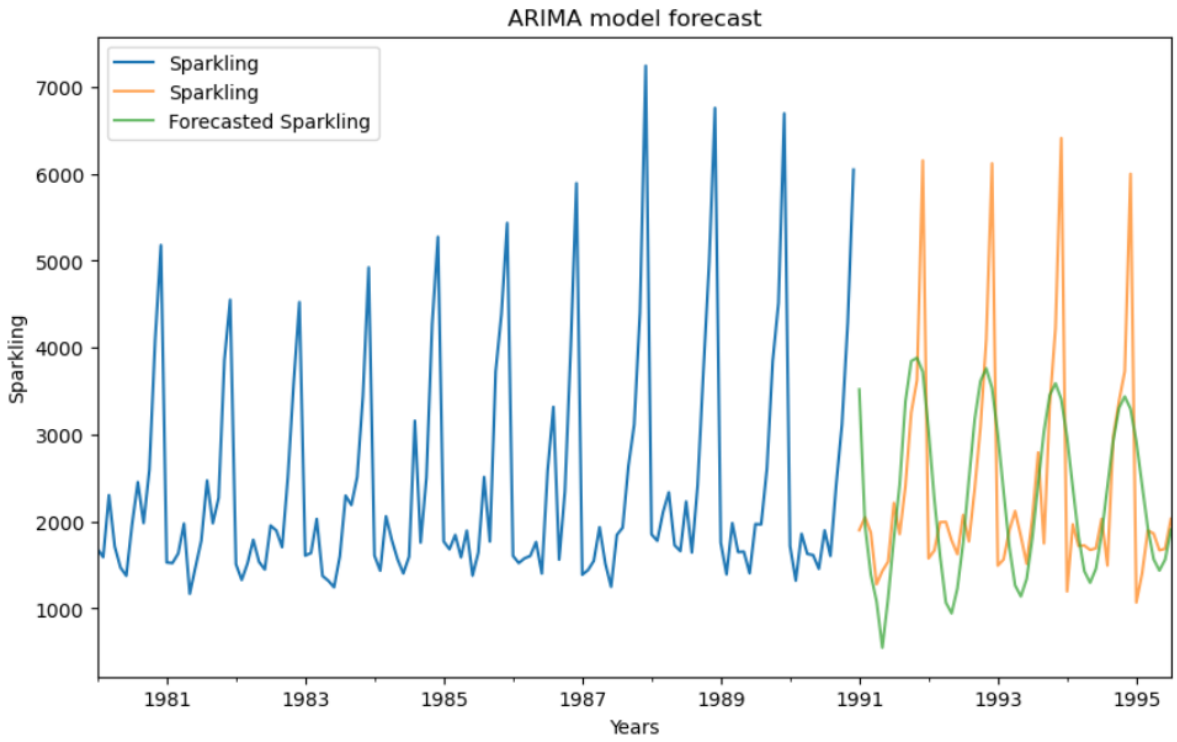
+ 2403.7

- A sample of the forecasted values on the actual test data is shown below:

YearMonth	Sparkling	sparkling_forecasted
1991-01-01	1902	3519.390427
1991-02-01	2049	1951.626611
1991-03-01	1874	1389.978322
1991-04-01	1279	1080.924594
1991-05-01	1432	550.144243

-

- Root mean squared error for the above built ARIMA model is 1011.472
- Forecast plot by ARIMA model on test data:



-

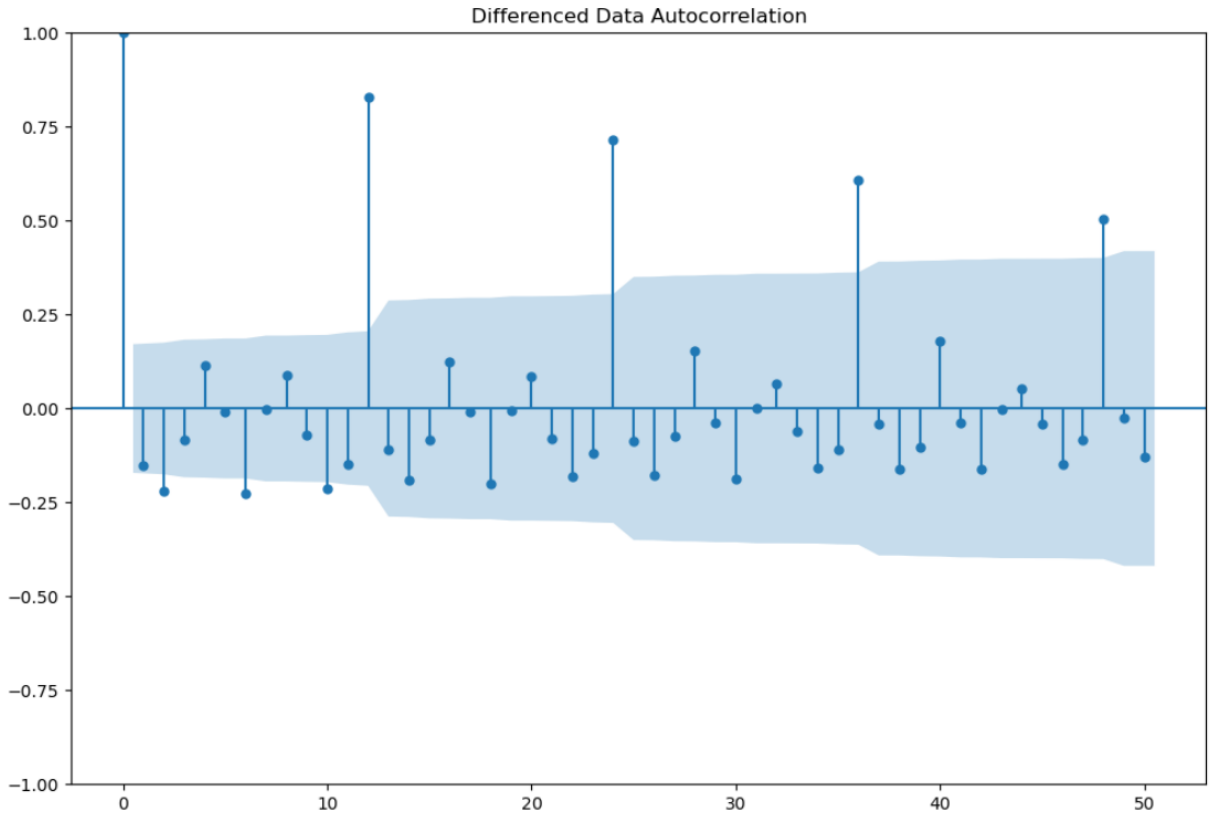
- Table showing all models built till now and their best RMSE values:

- It is observed that the ARIMA model has captured the test data but has failed to accurately predict the test values when compared to 2-point moving average or exponential smoothing models.

SARIMA model:

- **The ARIMA models can be extended/improved to handle seasonal components of a data series.**
- The seasonal autoregressive moving average model is given by
  - **SARIMA (p, d, q)(P, D, Q)F**
- The above model consists of:
  - Autoregressive and moving average components (p, q)
  - Seasonal autoregressive and moving average components (P, Q)
  - The ordinary and seasonal difference components of order ‘d’ and ‘D’
  - Seasonal frequency ‘F’
- The value for the parameters (p,d,q) and (P, D, Q) can be decided by comparing different values for each and taking **the lowest AIC value** for the model build.
- **The value for F can be consolidated by ACF plot**

Let’s plot an ACF plot on the given data:



- ACF plots help us understand the correction of time series values with their own past values.
- From the ACF plot, we observed there is a trend in the data which is not significantly increasing or decreasing.
- It is also observed that there is strong correlation among the time series values at regular lag intervals. For example at lag = 6 and lag = 12, there is strong correlation.
- This indicates that there is strong seasonality in the data that needs to be considered when choosing our final model.

- SARIMA model also takes into account the seasonality along with the AR, MA and differencing terms.
- An automated SARIMA model has been built with the following combinations/parameters.
- AR component(p/P): Range(1,5)
- MA component(q/Q): Range(1,5)
- Differencing component(d): 1
- Seasonal component(F): (6,12)
- Enforce_stationarity: False (not to enforce stationarity on the AR components of the model)
- Enforce_intvertibility: False(not to enforce invertibility on the MA components on the model)

Results with SARIMA model having seasonality = 6:

- It notes the AIC scores for the data with each set of (p,q,d)(P,Q,D,F) values.
- The top 5 combinations with least AIC scores obtained are

	param	seasonal	AIC
364	(2, 1, 4)	(2, 0, 4, 6)	1534.798401
114	(0, 1, 4)	(2, 0, 4, 6)	1535.344780
239	(1, 1, 4)	(2, 0, 4, 6)	1535.645847
489	(3, 1, 4)	(2, 0, 4, 6)	1536.507068
124	(0, 1, 4)	(4, 0, 4, 6)	1537.199250

- Therefore, best hyper parameters for p,d,q for the given data set are taken as (2,1,4) (2,0,4,6)
- AIC metric for this combination is 1534.798
- Applying SARIMA on the model with the above obtained parameters.

- Summary obtained:

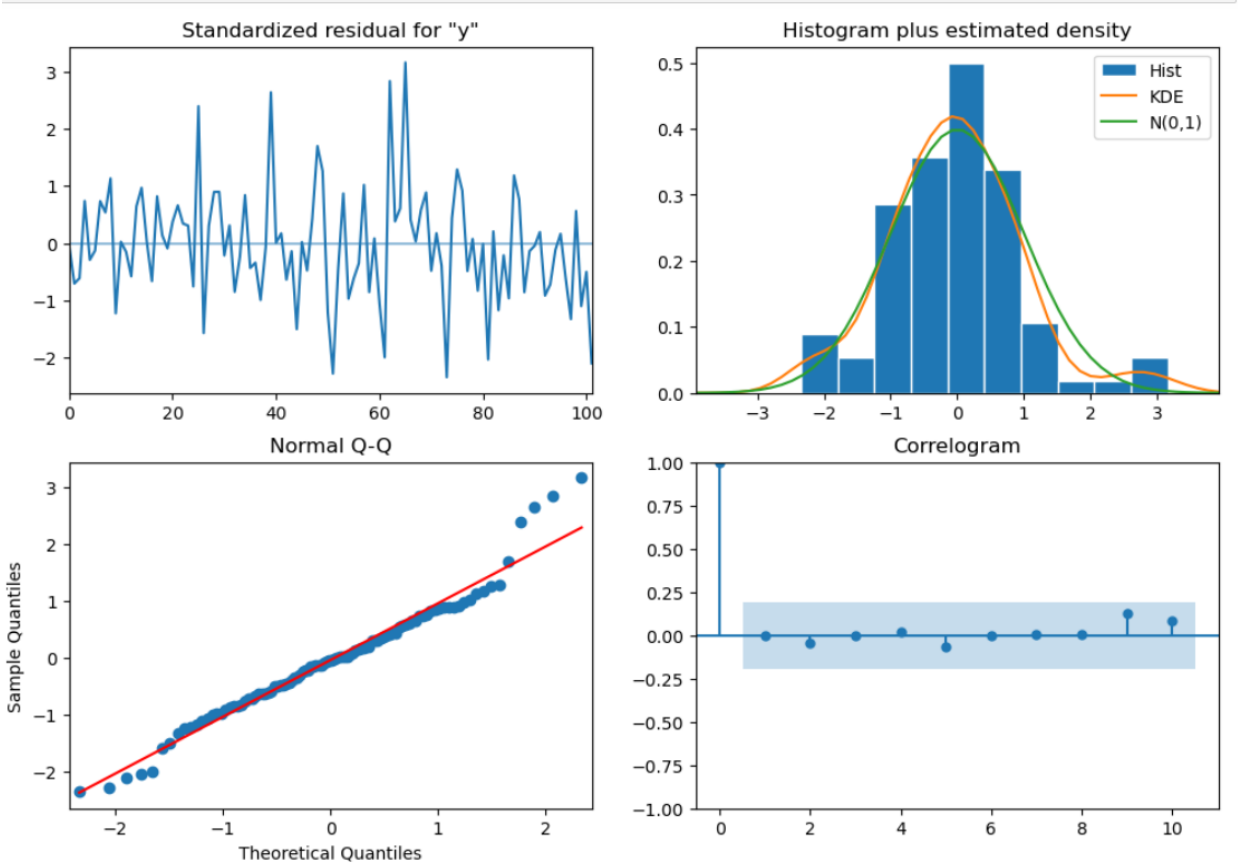
```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(2, 1, 4)x(2, 0, 4, 6)   Log Likelihood                -754.399
Date:                           Sat, 05 Aug 2023   AIC                            1534.798
Time:                                   12:45:54   BIC                            1568.923
Sample:                                        0   HQIC                           1548.617
                                             - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6592      0.140    -11.885      0.000      -1.933      -1.386
ar.L2         -0.8481      0.120     -7.065      0.000      -1.083      -0.613
ma.L1          0.9474      0.197      4.821      0.000       0.562       1.333
ma.L2         -0.6008      0.173     -3.478      0.001      -0.939      -0.262
ma.L3         -0.8712      0.199     -4.376      0.000      -1.261      -0.481
ma.L4         -0.1533      0.146     -1.051      0.293      -0.439       0.133
ar.S.L6       -0.0122      0.033     -0.371      0.710      -0.077       0.052
ar.S.L12       1.0341      0.026     39.086      0.000       0.982       1.086
ma.S.L6        0.2701      0.241      1.120      0.263      -0.202       0.743
ma.S.L12      -0.6350      0.153     -4.164      0.000      -0.934      -0.336
ma.S.L18       0.0618      0.149      0.413      0.679      -0.231       0.355
ma.S.L24      -0.0465      0.150     -0.310      0.756      -0.341       0.248
sigma2      1.347e+05   2.01e-06   6.69e+10      0.000    1.35e+05    1.35e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 8.82
Prob(Q):                              0.99   Prob(JB):                         0.01
Heteroskedasticity (H):               1.47   Skew:                             0.42
Prob(H) (two-sided):                  0.26   Kurtosis:                         4.17
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 6.92e+25. Standard errors may be unstable.
```

-
- The equation obtained by ARIMA model can be written as follows:
- The terms ar.L1, ar.L2, ma.L1, ma.L2, ma.L3, ma.L4 indicate the orders of AR and MA components of the actual data.
- The terms ar.S.L6, ar.S.L12, ma.S.L6, ma.S.L12. ma.S.L18, ma.S.L24 indicate the AR and MA components of the previous seasonal data.

- Root mean squared error for the above built ARIMA model is **666.321**
- Table showing all models built till now and their best RMSE values:

	Test RMSE
RegressionOnTime	1389.135175
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.0496,SimpleExponentialSmoothing	1316.035487
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha =0.688, Beta = 0.00009,DoubleExponentialSmoothing	1316.035487
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothing	404.286809
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
Best ARIMA Model : ARIMA(4,0,4)	1011.471574
SARIMA(2,1,4)(2,0,4,6)	666.320817

- The diagnostics for the above built SARIMA model with seasonality = 6 is:



-

- It is observed that the SARIMA model has captured the test data more accurately than the ARIMA model as we have included seasonal components to forecast the values. This made the forecasts more accurately predictable.

<u>Results with SARIMA model having seasonality = 12:</u>

- It notes the AIC scores for the data with each set of (p,q,d)(P,Q,D,F) values.
- The top 15 combinations with least AIC scores obtained are

:

	param	seasonal	AIC
2	(1, 1, 1)	(1, 0, 3, 12)	14.000000
20	(1, 1, 3)	(1, 0, 3, 12)	18.000000
47	(2, 1, 3)	(1, 0, 3, 12)	20.000000
65	(3, 1, 2)	(1, 0, 3, 12)	20.000000
68	(3, 1, 2)	(2, 0, 3, 12)	22.000000
77	(3, 1, 3)	(2, 0, 3, 12)	24.000000
23	(1, 1, 3)	(2, 0, 3, 12)	95.963384
5	(1, 1, 1)	(2, 0, 3, 12)	883.122682
69	(3, 1, 2)	(3, 0, 1, 12)	1388.602614
60	(3, 1, 1)	(3, 0, 1, 12)	1388.681497
61	(3, 1, 1)	(3, 0, 2, 12)	1389.195899
70	(3, 1, 2)	(3, 0, 2, 12)	1389.702000
78	(3, 1, 3)	(3, 0, 1, 12)	1390.535977
79	(3, 1, 3)	(3, 0, 2, 12)	1392.632812
51	(2, 1, 3)	(3, 0, 1, 12)	1400.119862

-

- Therefore, best hyper parameters for p,d,q for the given data set are taken as (3,1,1) (3,0,1,12)
- AIC metric for this combination is 1388.602
- Applying SARIMA on the model with the above obtained parameters.
- Summary obtained:

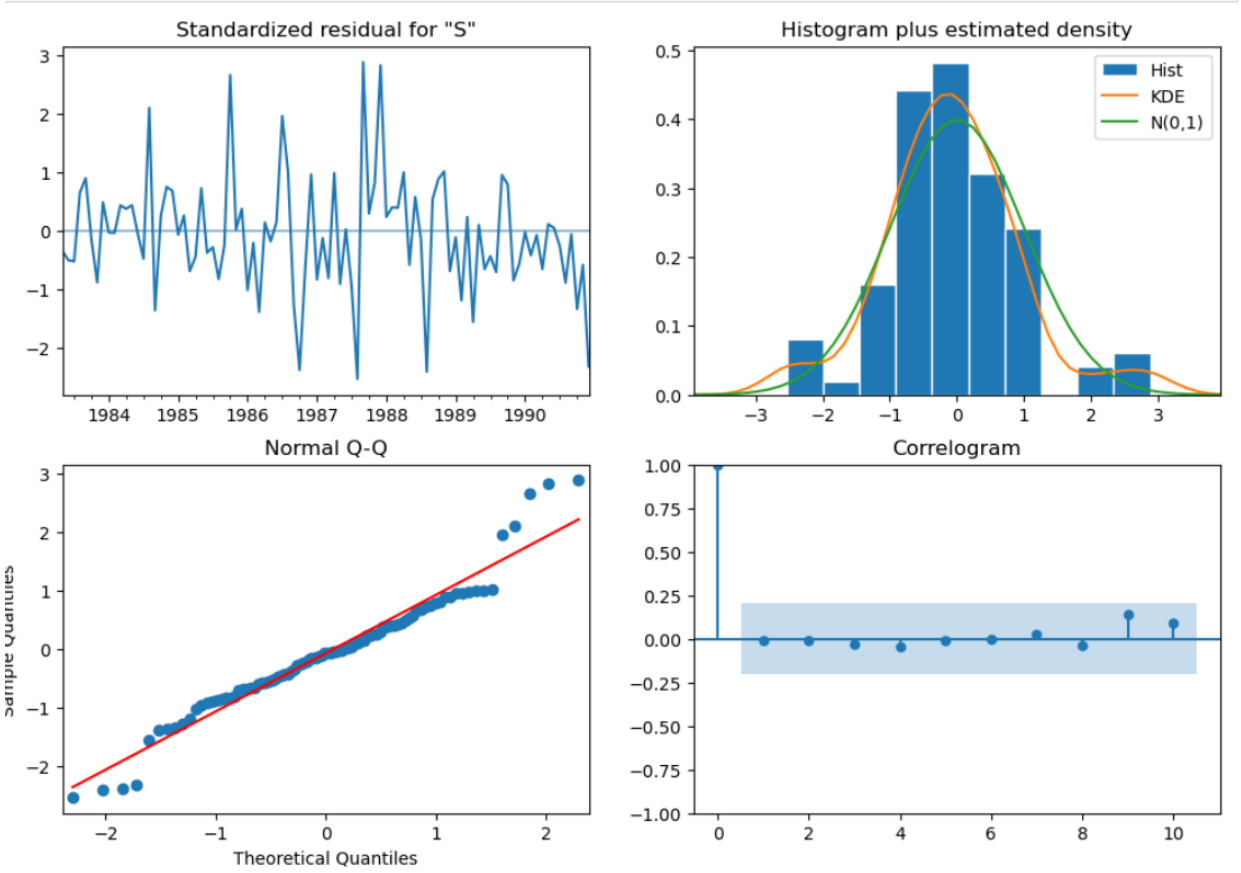
```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                           Sparkling   No. Observations:                  132
Model:             SARIMAX(3, 1, 2)x(3, 0, [1], 12)   Log Likelihood                -684.301
Date:                              Sat, 05 Aug 2023   AIC                           1388.603
Time:                                      14:48:06   BIC                           1413.820
Sample:                                  01-01-1980   HQIC                          1398.781
                                       - 12-01-1990
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5433      0.416     -1.306      0.192      -1.359       0.272
ar.L2         -0.0077      0.198     -0.039      0.969      -0.396       0.381
ar.L3          0.0636      0.140      0.452      0.651      -0.212       0.339
ma.L1         -0.1993      0.405     -0.493      0.622      -0.992       0.594
ma.L2         -0.6547      0.327     -2.004      0.045      -1.295      -0.014
ar.S.L12       0.7654      0.448      1.707      0.088      -0.113       1.644
ar.S.L24       0.1090      0.330      0.330      0.741      -0.537       0.755
ar.S.L36       0.1764      0.187      0.946      0.344      -0.189       0.542
ma.S.L12      -0.2431      0.451     -0.540      0.589      -1.126       0.640
sigma2      1.664e+05   2.63e+04      6.324      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).
```

-
- The equation obtained by ARIMA model can be written as follows:
- The terms ar.L1, ar.L2, ar.L3, ma.L1, ma.L2 indicate the orders of AR and MA components of the actual data.
- The terms ar.S.L12, ar.S.L24, ma.S.L36, ma.S.L12 indicate the AR and MA components of the previous seasonal data.

- Root mean squared error for the above built ARIMA model is **580.057**
- Table showing all models built till now and their best RMSE values:

	Test RMSE
RegressionOnTime	1389.135175
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.0496,SimpleExponentialSmoothing	1316.035487
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
Alpha =0.688, Beta = 0.00009,DoubleExponentialSmoothing	1316.035487
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothing	404.286809
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
Best ARIMA Model : ARIMA(4,0,4)	1011.471574
SARIMA(2,1,4)(2,0,4,6)	666.320817
SARIMA(3,1,2)(3,0,1,12)	580.056846

- Observing the diagnostics for the above model:



- Viewing the sample predictions of SARIMA model with predicted range of values on a confidence interval of 95%:

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1991-01-01	1320.319125	407.878186	520.892570	2119.745680
1991-02-01	1298.568720	421.171952	473.086862	2124.050578
1991-03-01	1604.435992	421.172386	778.953284	2429.918701
1991-04-01	1625.976998	429.668416	783.842378	2468.111618
1991-05-01	1397.847814	430.136023	554.796701	2240.898928

- It is observed that the SARIMA model has captured the test data more accurately than the ARIMA model as we have included seasonal components to forecast the values. This made the forecasts more accurately predictable.

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

Forecasting with parameter values	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothing	404.286809
SARIMA(3,1,2)(3,0,1,12)	580.056846
SARIMA(2,1,4)(2,0,4,6)	666.320817
2pointTrailingMovingAverage	813.400684
Best ARIMA Model : ARIMA(4,0,4)	1011.471574
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
Alpha=0.0496,SimpleExponentialSmoothing	1316.035487
Alpha =0.688, Beta = 0.00009,DoubleExponentialSmoothing	1316.035487
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670

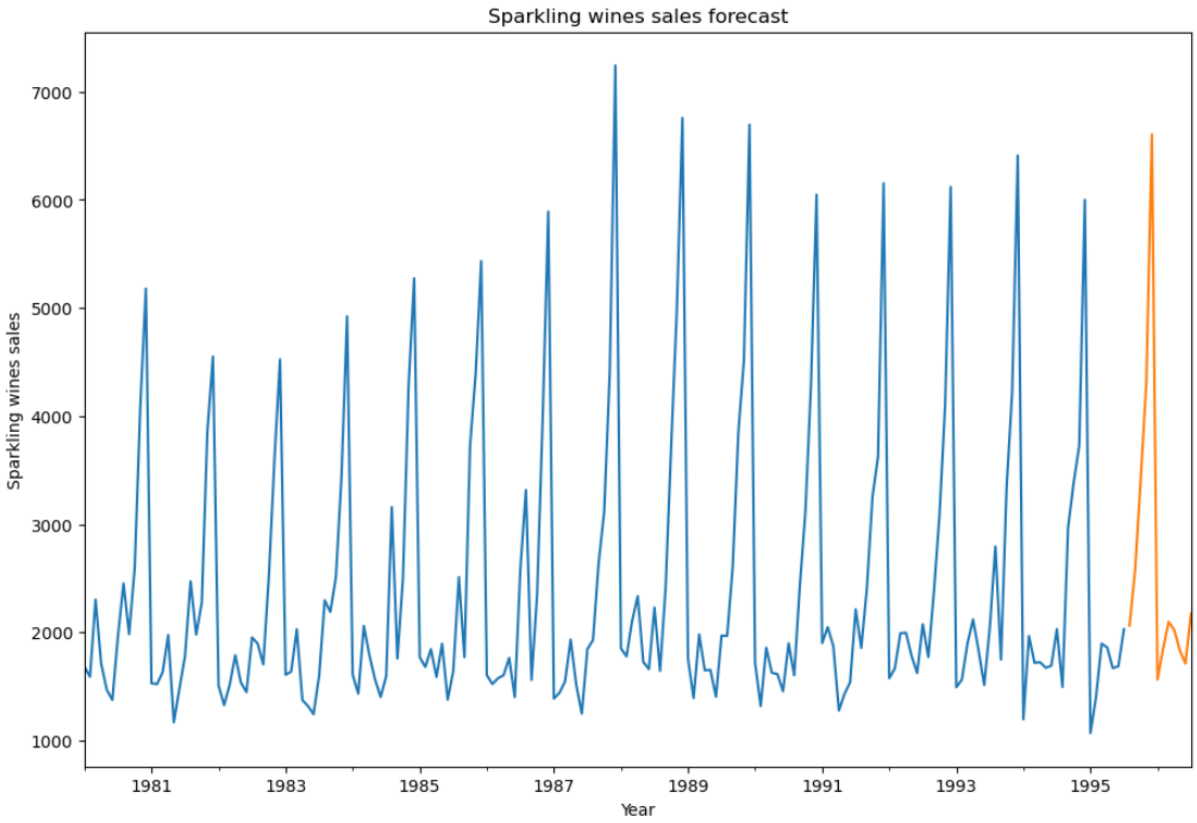
- The above indicates all models used for forecasting Sparkling wines sales with parameters specified.
- The data has been sorted according to the ascending RMSE values.
- It is evident that the Triple exponential smoothing model with parameters alpha = 0.4, beta = 0.1, gamma = 0.2 renders the most optimal model with the best accuracy among all.

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

- Since, triple exponential smoothing is the most optimal model, constructed a triple exponential model on full data.
- Parameters used: alpha = 0.4, beta = 0.1, gamma = 0.2
- RMSE obtained for the model built: 376
- This model is used to forecast the values for the next 12 months from the last month of the dataset.
- The forecast ranges from '1995-08-01' to '1996-07-01'
- Below is the forecasted data:

	Sparkling future predictions
1995-08-01	2063.448803
1995-09-01	2579.407389
1995-10-01	3416.654268
1995-11-01	4304.477169
1995-12-01	6604.876647
1996-01-01	1564.539777
1996-02-01	1849.759980
1996-03-01	2098.878830
1996-04-01	2022.428830
1996-05-01	1834.540687
1996-06-01	1712.408933
1996-07-01	2176.425361

-
- The forecast along with current data looks like below:

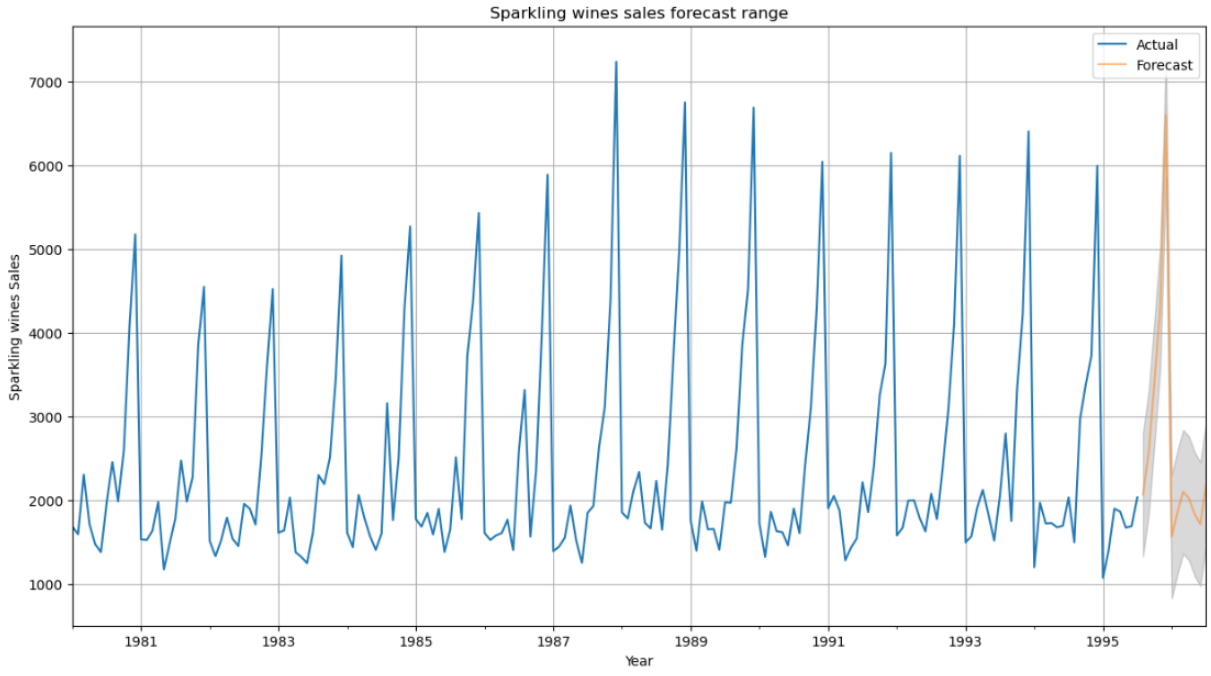


-

- Plotting the range of forecasted values for the next 12 months with a confidence interval of 95%

	lower_CI	prediction	upper_ci
1995-08-01	1322.989138	2063.448803	2803.908469
1995-09-01	1838.947723	2579.407389	3319.867055
1995-10-01	2676.194602	3416.654268	4157.113934
1995-11-01	3564.017503	4304.477169	5044.936834
1995-12-01	5864.416981	6604.876647	7345.336312
1996-01-01	824.080112	1564.539777	2304.999443
1996-02-01	1109.300314	1849.759980	2590.219645
1996-03-01	1358.419164	2098.878830	2839.338496
1996-04-01	1281.969164	2022.428830	2762.888495
1996-05-01	1094.081021	1834.540687	2575.000352
1996-06-01	971.949267	1712.408933	2452.868599
1996-07-01	1435.965695	2176.425361	2916.885027

- The plot for the same is shown below:



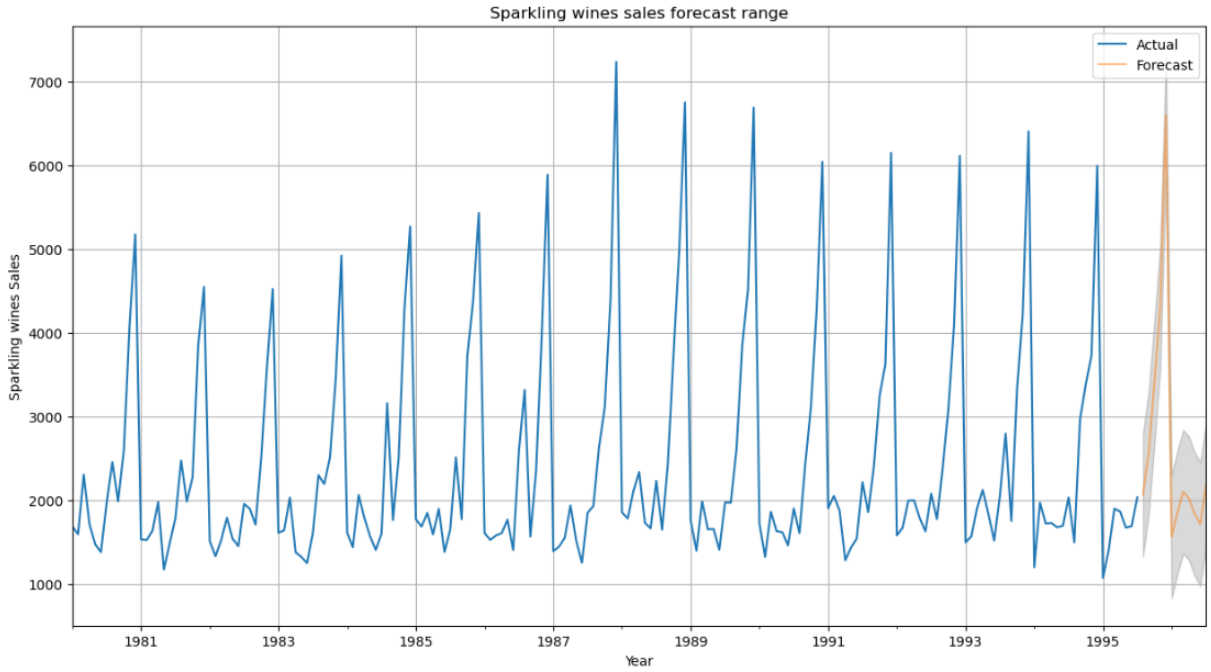
9. **Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.**

Ans:

- The given dataset depicts sparkling wines sales on a monthly basis for the years 1908-1995.
- The same has been read in proper datetime format. Exploratory data analysis has been done on the project.
- It showed that the data did not have any missing values.
- The data has been decomposed. There is no significant upward or downward trend observed but a significant multiplicative seasonality observed in the data. The sales grow rapidly at the end of the year.
- Several models like Linear Regression, Exponential smoothing methods, Moving average methods, ARIMA, SARIMA have been trained with the dataset given.
- Of all the models, the models which account for seasonality have performed better compared to other models.
- The RMSE scores obtained for all the models built are shown below.

Forecasting with parameter values	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	317.434302
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothing	404.286809
SARIMA(3,1,2)(3,0,1,12)	580.056846
SARIMA(2,1,4)(2,0,4,6)	666.320817
2pointTrailingMovingAverage	813.400684
Best ARIMA Model : ARIMA(4,0,4)	1011.471574
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
Alpha=0.0496,SimpleExponentialSmoothing	1316.035487
Alpha =0.688, Beta = 0.00009,DoubleExponentialSmoothing	1316.035487
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,SimpleExponentialSmoothing	1375.393398
RegressionOnTime	1389.135175
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	1778.564670

- The best model for forecasting sales for Sparkling wines data is triple exponential smoothing.
- The forecast range for the sales of next 12 months has obtained as below:



-

- The forecast range indicates that the data is going to see the same peaks as it has seen recently during the end of the year.
- Mid year sales seem to perform better compared to the very recent lows experienced during those seasons.