TSF Project - Coded

Sparkling

Isha Shukla

26 May 2024

INDEX

Plots

Problem Statement

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

The main goal of this project is to study and predict wine sales trends from the 20th century using historical data from ABC Estate Wines. We want to give ABC Estate Wines useful insights to improve sales, take advantage of new market opportunities, and stay competitive in the wine industry.

1. Define the problem and perform Exploratory Data Analysis

(I) Read the data

- There are two columns in the dataset Sparkling.csv
- The dataset has 187 rows and 2 columns.
- Columns - YearMonth(datatype as object) and Sparkling (datatype as int)
- Sparkling column has no null values.

```
(187, 2)
```

Shape of 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: 3.1+ KB
```

Information about the dataset

Table 1 - Rows 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

Top 5 rows of the dataset

	YearMonth	Sparkling
82	1995-03	1897
83	1995-04	1862
84	1995-05	1670
85	1995-06	1688
86	1995-07	2031

Last 5 rows of the dataset

- We can see Year from 1980 - 1995 in the dataset.

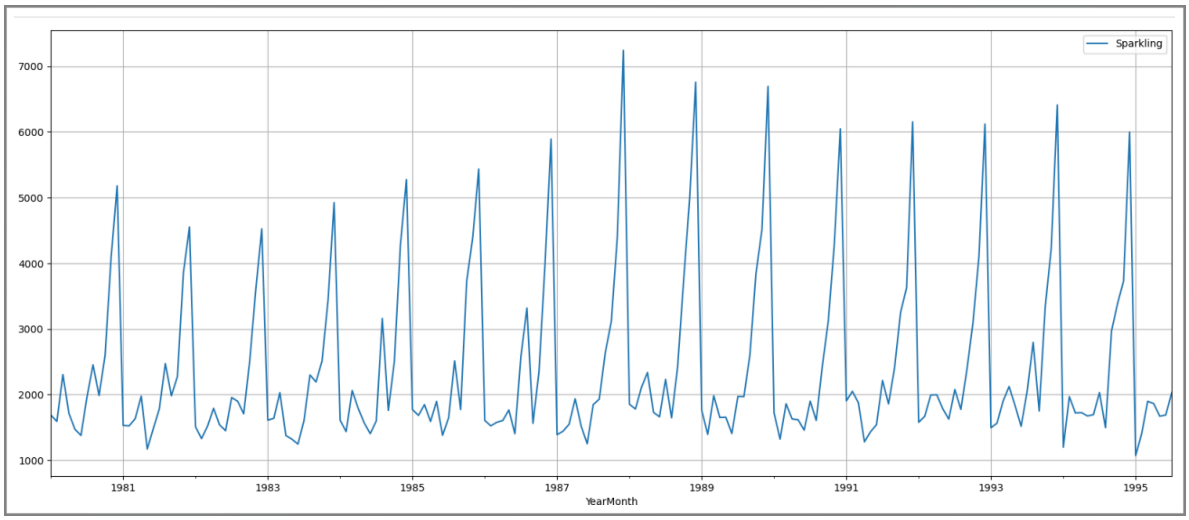
Table 2 - Statistical Summary of the dataset.

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

Comprehensive Summary of the dataset.

II. Plot the data

Plot 1 - Plot the data



Plot of the dataset - Sparkling vs YearMonth

For enhanced analysis of the dataset, we have segmented it further by extracting the month and year components from the 'YearMonth' column. This division allows for more granular examination of the data based on month-to-month and year-to-year trends. Now, there are 3 columns and 187 rows.

Table 3 - After extraction of Year and Month.

YearMonth	Sparkling	Year	Month
980-01-01	1686	1980	1
980-02-01	1591	1980	2
980-03-01	2304	1980	3
980-04-01	1712	1980	4
980-05-01	1471	1980	5

Top 5 rows after extraction of Year and Month

IV. Perform Exploratory Data Analysis (EDA)

- There is no null values.

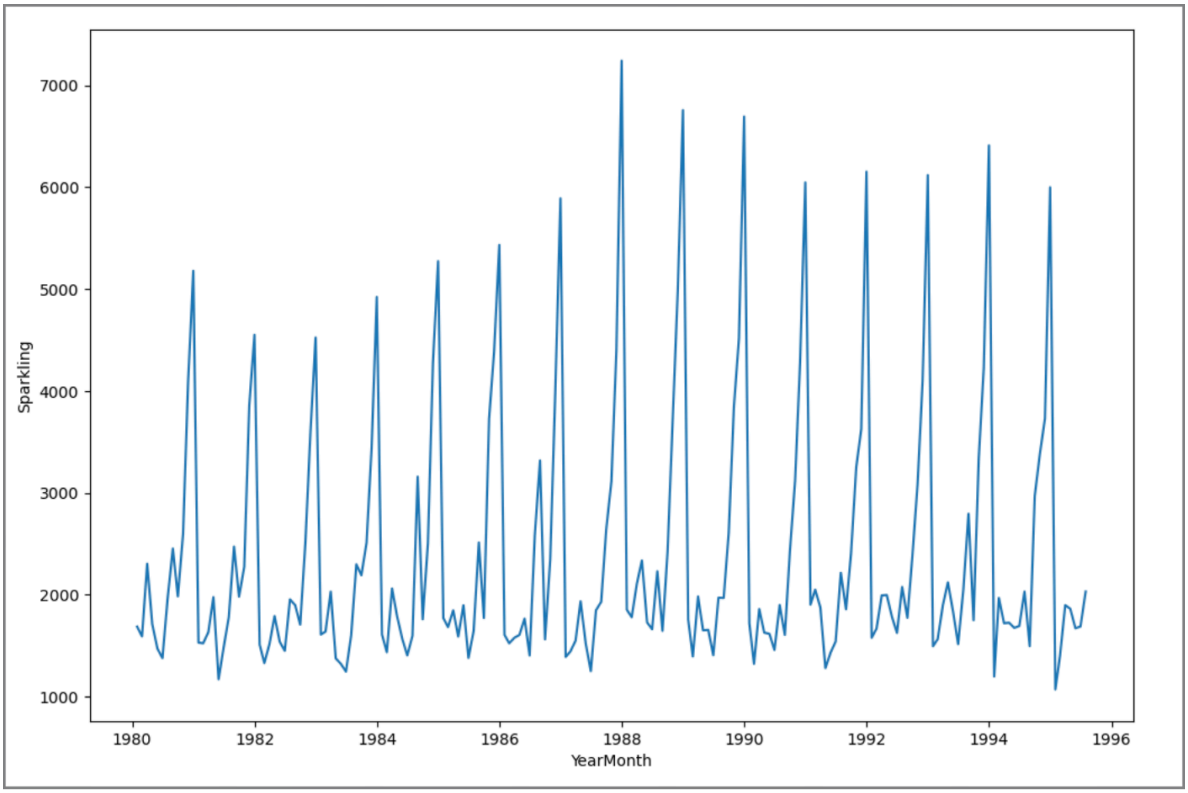
Treating missing values is crucial for maintaining data integrity, ensuring accurate analyses, and deriving reliable insights, thereby enabling informed decision-making and valid conclusions.

- We'll resample the data to aggregate values at a monthly level from the daily-level data, computing the average for each month.

YearMonth	Sparkling	Year	Month
1980-01-31	1686.0	1980.0	1.0
1980-02-29	1591.0	1980.0	2.0
1980-03-31	2304.0	1980.0	3.0
1980-04-30	1712.0	1980.0	4.0
1980-05-31	1471.0	1980.0	5.0

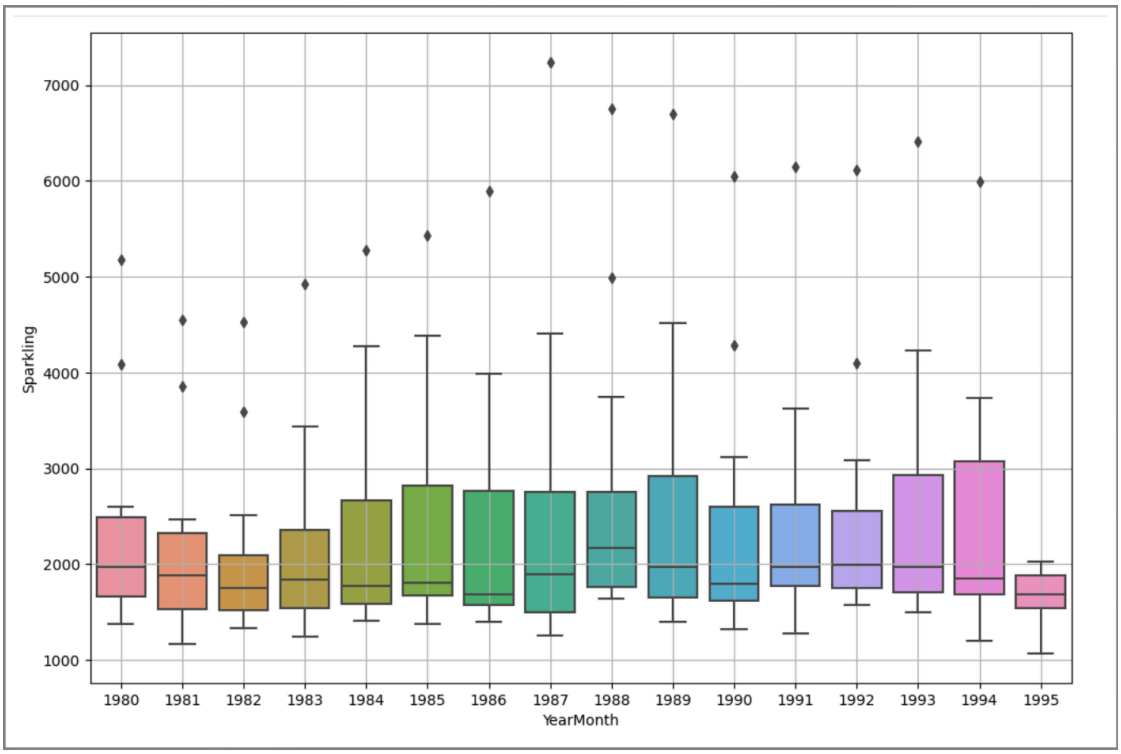
After resampling of the dataset

Plot 2 - The trend of Sparkling at year level



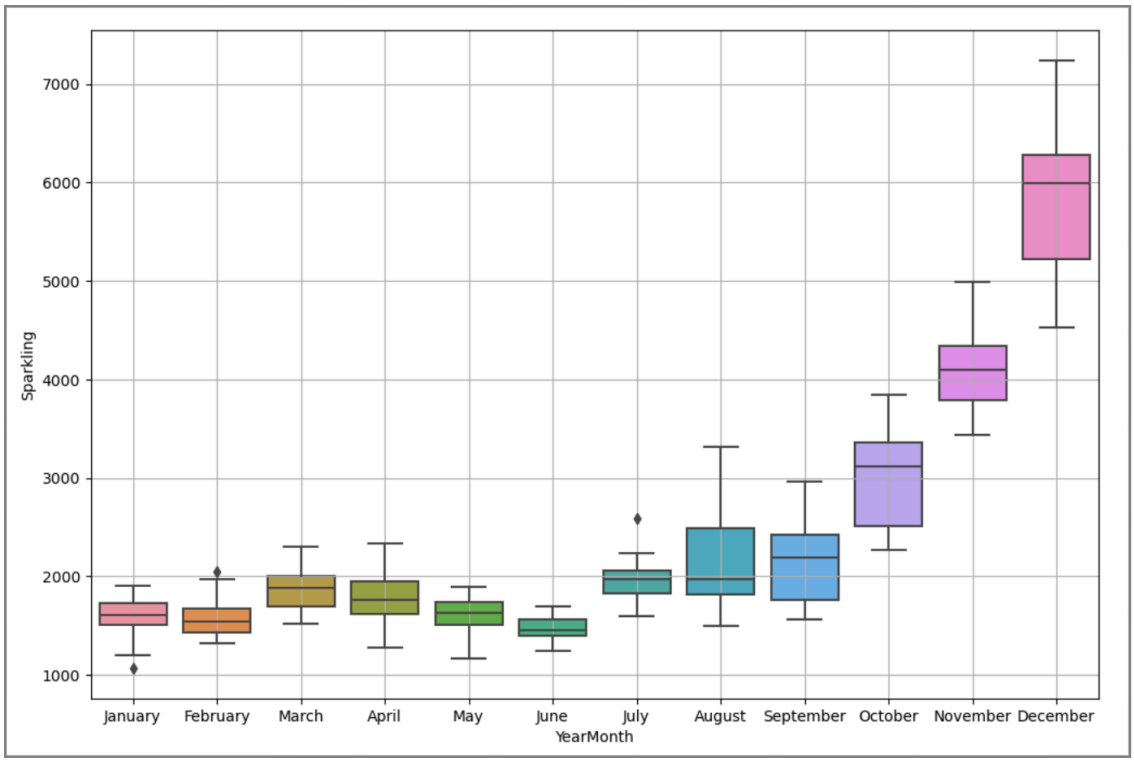
- There was a peak in 1988. The plot shows that there is trend and seasonality.

Plot 3 - Yearly Box-plot



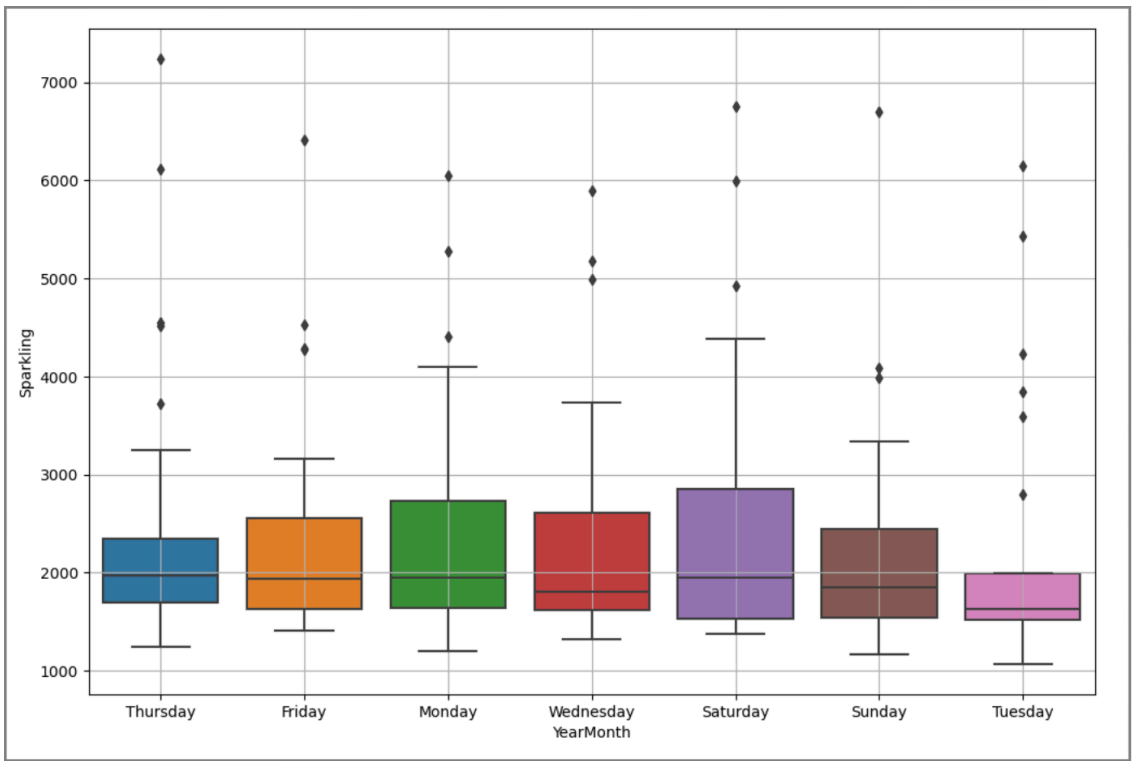
Yearly Box-plot

Plot 4 - Monthly Boxplot



Monthly Boxplot

Plot 5 - Weekly Box-plot



Weekly Box-plot

Outliers Observation

- **Yearly Boxplot** - Outliers persist across nearly all years. There was a peek in 1988.
- **Monthly Boxplot** - The graph indicates that wine sales peak in December and hit their lowest point in January. Sales remain steady from January to June, but then begin to rise steadily from July onwards. However, there are some outliers in January, February, and July.
- **Weekly Boxplot -** There are outliers present on all days. Tuesday has the lowest sales.

Table 4 - Pivot table displays monthly price across years

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

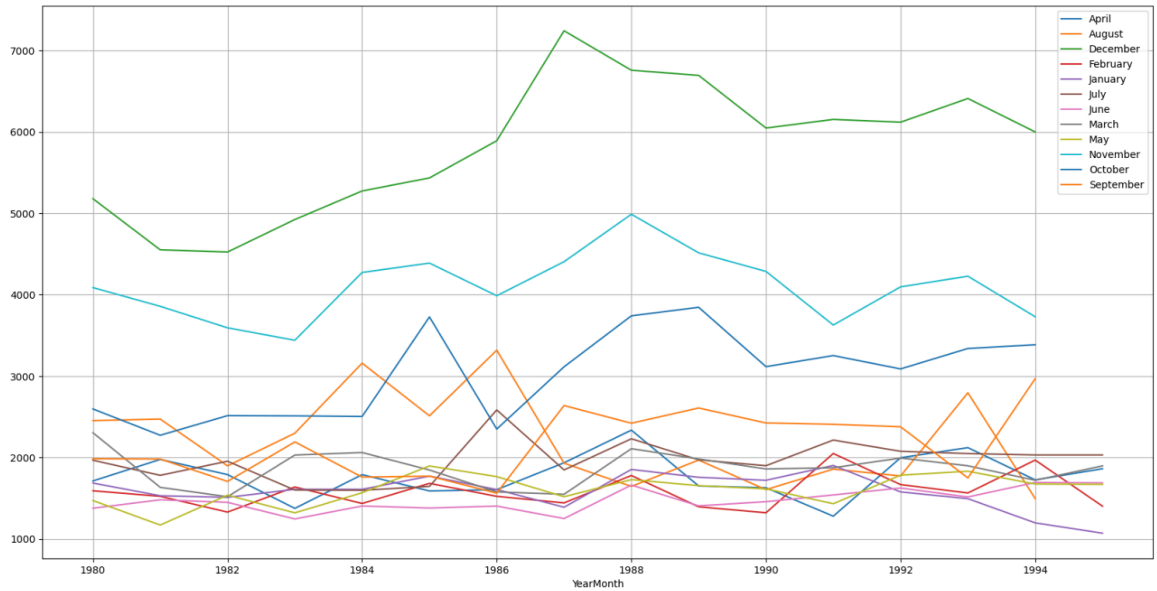
Pivot Table

Here are observations from the pivot table:

- December consistently has the highest sales across all years, peaking at 7242 units in 1987.

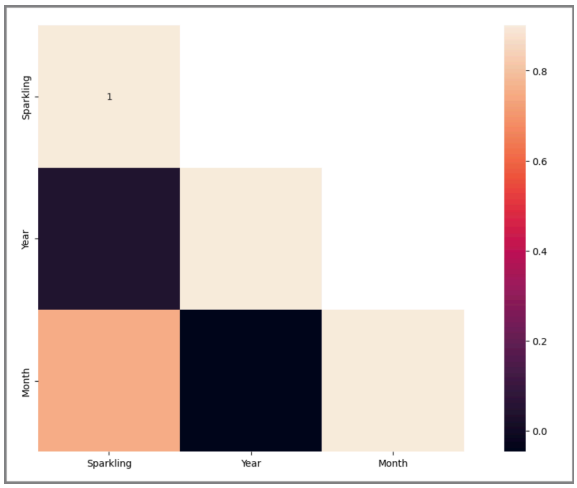
- January consistently has the lowest sales, with the lowest recorded sales of 1070 units in 1995.
- Sales are stable from January to June.
- Sales increase from July onwards, peaking in December.
- In 1987, December marked the highest sales at 7242 units.
- In 1995, January had the lowest sales at 1070 units.
- August and November consistently show high sales but not as high as December.
- April and May generally have lower sales, indicating potential for targeted marketing efforts.
- Notable spikes in sales occur in June, July, August, September, and December across various years.
- Missing data for August, October, and November 1995 suggest potential data collection issues or significant disruptions.
- December peaks and January lows are observed consistently across the years.
- From March to June, sales are moderate and stable compared to other months.
- Sales noticeably increase from July onwards, leading up to December.

Plot 6 - Pivot table plot for monthly wine sale across year



Graph to display monthly price across years

Plot 7 - Correlation Plot



Correlation Plot

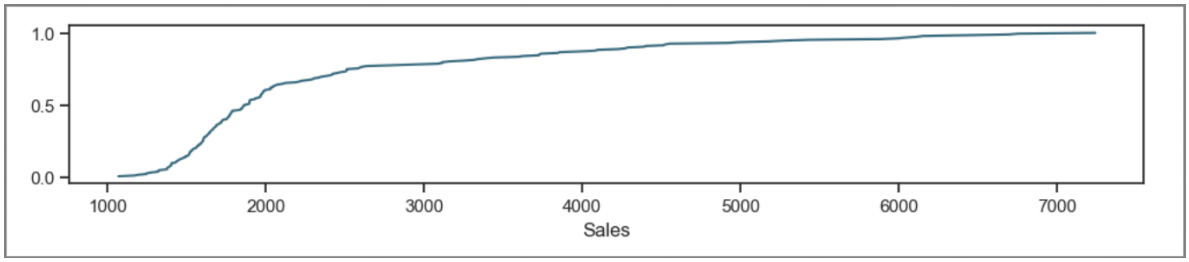
	Sparkling	Year	Month
Sparkling	1.000000	0.039035	0.752337
Year	0.039035	1.000000	-0.046502
Month	0.752337	-0.046502	1.000000

Sparkling Wine sales Correlation with respect to year and month

Observations from the correlation table and plot:

- The weak positive correlation (0.039) suggests that there is no significant trend in sparkling wine sales over the years, indicating sales are relatively stable without a clear upward or downward trend.
- The strong positive correlation (0.752) indicates a pronounced seasonality in sparkling wine sales, with sales peaking during certain months of the year.
- The very weak negative correlation (-0.046) confirms that the variables "Year" and "Month" are independent, indicating that the trend over the years is separate from monthly seasonal patterns.

Plot 8 - Empirical Cumulative Distribution Function (ECDF)



Graph for ECDF

From the ECDF plot of wine sales observations, we can observe the following:

- The x-axis represents the range of wine sales observations, while the y-axis represents the cumulative probability.
- The plot shows how the cumulative probability increases as we move along the sorted wine sales values.
- By examining the slope of the curve, we can infer the density of the observations at different values. Steeper slopes indicate higher density of observations, while flatter slopes indicate lower density.
- The ECDF plot provides a comprehensive overview of the distribution of wine sales observations, allowing us to assess characteristics such as central tendency, spread, and percentiles.
- Overall, the ECDF plot helps us understand the empirical distribution of wine sales observations and can provide insights into the underlying patterns and variability in the data.

```
[ -inf 1070. 1170. 1197. 1245. 1250. 1279. 1320. 1321. 1329. 1375. 1377.
 1379. 1389. 1394. 1402. 1403. 1404. 1406. 1432. 1435. 1442. 1449. 1457.
 1471. 1480. 1494. 1495. 1510. 1515. 1518. 1518. 1523. 1523. 1530. 1537.
 1540. 1548. 1562. 1564. 1567. 1577. 1577. 1589. 1591. 1597. 1600. 1605.
 1605. 1606. 1609. 1609. 1615. 1625. 1628. 1633. 1638. 1645. 1645. 1650.
 1654. 1661. 1667. 1670. 1674. 1682. 1686. 1688. 1693. 1706. 1712. 1720.
 1720. 1725. 1728. 1749. 1757. 1759. 1765. 1771. 1771. 1773. 1779. 1781.
 1783. 1789. 1790. 1831. 1846. 1847. 1853. 1857. 1859. 1862. 1874. 1896.
 1897. 1897. 1898. 1899. 1902. 1930. 1935. 1954. 1966. 1968. 1968. 1971.
 1976. 1981. 1982. 1984. 1993. 1997. 2030. 2031. 2031. 2048. 2049. 2061.
 2076. 2108. 2121. 2191. 2214. 2230. 2273. 2298. 2304. 2336. 2349. 2377.
 2408. 2421. 2424. 2453. 2472. 2504. 2511. 2512. 2514. 2584. 2596. 2608.
 2638. 2795. 2968. 3088. 3114. 3116. 3159. 3252. 3318. 3339. 3385. 3440.
 3593. 3627. 3727. 3729. 3740. 3845. 3857. 3987. 4087. 4096. 4227. 4273.
 4286. 4388. 4405. 4514. 4524. 4551. 4923. 4988. 5179. 5274. 5434. 5891.
 5999. 6047. 6119. 6153. 6410. 6694. 6757. 7242.]
```

ECDF for X- axis(Wine sales)

```
[0.         0.00534759 0.01069519 0.01604278 0.02139037 0.02673797
 0.03208556 0.03743316 0.04278075 0.04812834 0.05347594 0.05882353
 0.06417112 0.06951872 0.07486631 0.0802139  0.0855615  0.09090909
 0.09625668 0.10160428 0.10695187 0.11229947 0.11764706 0.12299465
 0.12834225 0.13368984 0.13903743 0.14438503 0.14973262 0.15508021
 0.16042781 0.1657754  0.17112299 0.17647059 0.18181818 0.18716578
 0.19251337 0.19786096 0.20320856 0.20855615 0.21390374 0.21925134
 0.22459893 0.22994652 0.23529412 0.24064171 0.2459893  0.2513369
 0.25668449 0.26203209 0.26737968 0.27272727 0.27807487 0.28342246
 0.28877005 0.29411765 0.29946524 0.30481283 0.31016043 0.31550802
 0.32085561 0.32620321 0.3315508  0.3368984  0.34224599 0.34759358
 0.35294118 0.35828877 0.36363636 0.36898396 0.37433155 0.37967914
 0.38502674 0.39037433 0.39572193 0.40106952 0.40641711 0.41176471
 0.4171123  0.42245989 0.42780749 0.43315508 0.43850267 0.44385027
 0.44919786 0.45454545 0.45989305 0.46524064 0.47058824 0.47593583
 0.48128342 0.48663102 0.49197861 0.4973262  0.5026738  0.50802139
 0.51336898 0.51871658 0.52406417 0.52941176 0.53475936 0.54010695
 0.54545455 0.55080214 0.55614973 0.56149733 0.56684492 0.57219251
 0.57754011 0.5828877  0.58823529 0.59358289 0.59893048 0.60427807
 0.60962567 0.61497326 0.62032086 0.62566845 0.63101604 0.63636364
 0.64171123 0.64705882 0.65240642 0.65775401 0.6631016  0.6684492
 0.67379679 0.67914439 0.68449198 0.68983957 0.69518717 0.70053476
 0.70588235 0.71122995 0.71657754 0.72192513 0.72727273 0.73262032
 0.73796791 0.74331551 0.7486631  0.7540107  0.75935829 0.76470588
 0.77005348 0.77540107 0.78074866 0.78609626 0.79144385 0.79679144
 0.80213904 0.80748663 0.81283422 0.81818182 0.82352941 0.82887701
 0.8342246  0.83957219 0.84491979 0.85026738 0.85561497 0.86096257
 0.86631016 0.87165775 0.87700535 0.88235294 0.88770053 0.89304813
 0.89839572 0.90374332 0.90909091 0.9144385  0.9197861  0.92513369
 0.93048128 0.93582888 0.94117647 0.94652406 0.95187166 0.95721925
 0.96256684 0.96791444 0.97326203 0.97860963 0.98395722 0.98930481
 0.99465241 1.        ]
```

ECDF for Y-axis(Cumulative Probability)

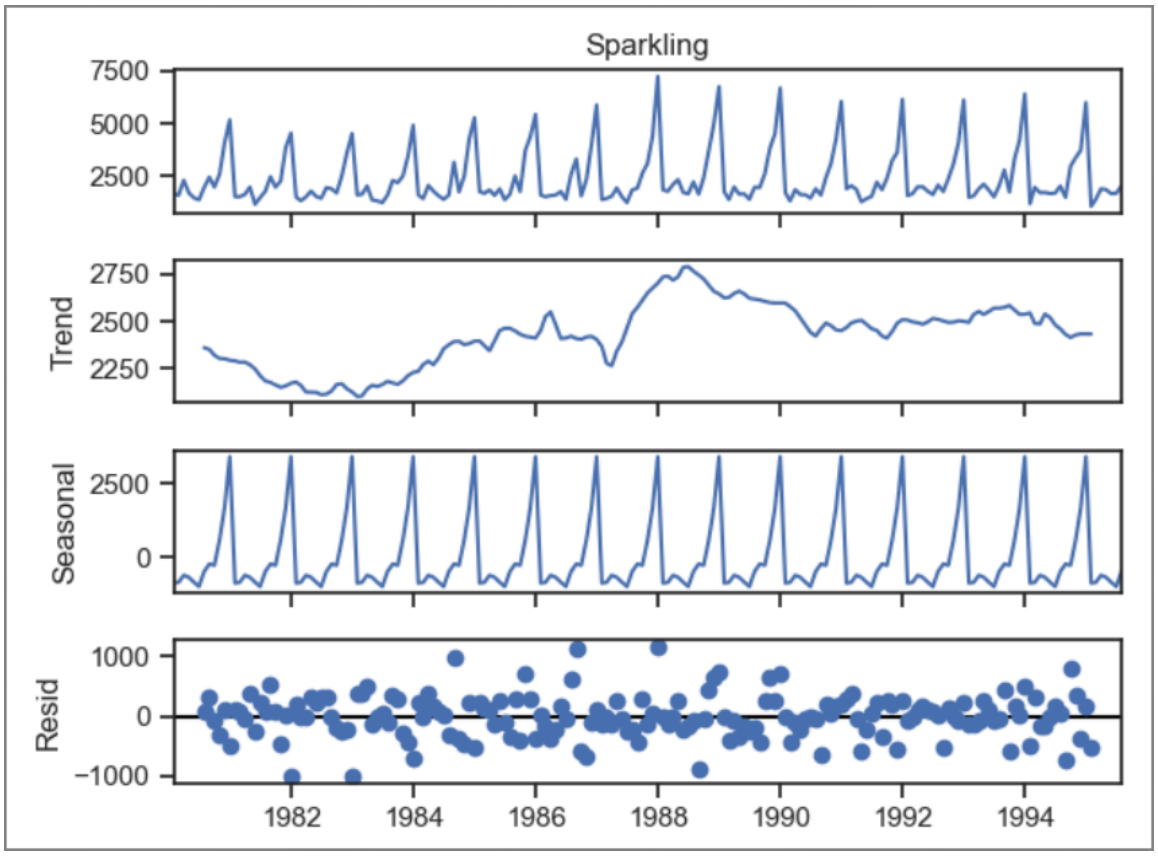
- Highest Wine sale is 7242.
- Lowest wine sale is 1070.

- More than 50% has sales less than 2000.

IV. Decomposition

a. Additive

Plot 9 - Additive Decomposition



Decomposition - Additive

- Trend and Seasonality is present.
- Residue/Noise is also there.
- Trend is decreasing after 1988 with respect to year.
- Sparkling wine sales increases as the month progresses, implying a seasonal pattern within each year.
- Peak year was 1988. Afterward sales is decreasing over the time.

```
Trend
 YearMonth
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
Freq: M, Name: trend, dtype: float64

Seasonality
 YearMonth
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
Freq: M, Name: seasonal, dtype: float64

Residual
 YearMonth
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
Freq: M, Name: resid, dtype: float64
```

Additive Decomposition - Trend, Seasonality and Residual in year 1980

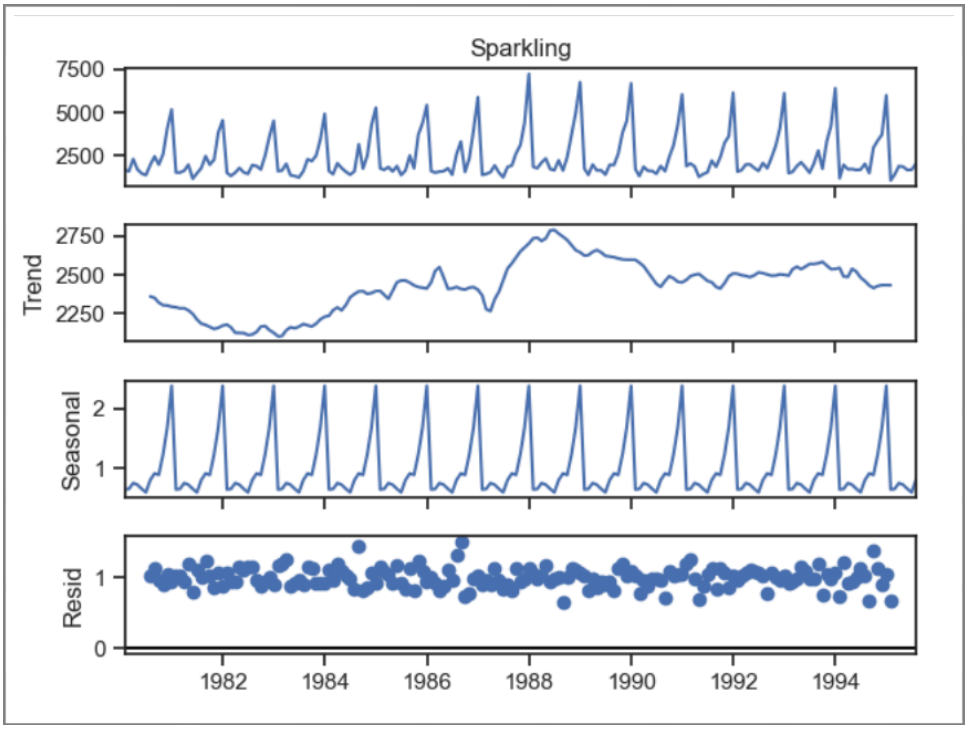
```
Trend
 YearMonth
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31    2360.666667
1980-08-31    2351.333333
1980-09-30    2320.541667
1980-10-31    2303.583333
1980-11-30    2302.041667
1980-12-31    2293.791667
Freq: M, Name: trend, dtype: float64

Seasonality
 YearMonth
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
Freq: M, Name: seasonal, dtype: float64

Residual
 YearMonth
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
Freq: M, Name: resid, dtype: float64
```

Multiplicative Decomposition - Trend, Seasonality and Residual in year 1980

b. **Multiplicative -** Plot 10



Decomposition - Multiplicative

- Trend and Seasonality is present.
- Residue/Noise ranges from 0 to 1, whereas additive noise ranges from 0 to 1000.
- Trend is decreasing after 1988 with respect to year.
- Sparkling wine sales increases as the month progresses, implying a seasonal pattern within each year.
- Peak year was 1988. Afterward sales is decreasing over the time.
- The multiplicative model is preferred over the additive model for decomposing Sparkling wine sales because of its narrower residual range.

2. Data Pre-processing

I. Train-test split

- The data from 1980 to 1990 is used as the training set, while the data from 1991 to 1995 is used as the testing set. This separation allows us to use the earlier data for training models and the later data for testing their performance.

Table 5 - Train and Test rows and columns

First few rows of Training Data

YearMonth	Sparkling	Year	Month
1980-01-31	1686.0	1980.0	1.0
1980-02-29	1591.0	1980.0	2.0
1980-03-31	2304.0	1980.0	3.0
1980-04-30	1712.0	1980.0	4.0
1980-05-31	1471.0	1980.0	5.0

Last few rows of Training Data

YearMonth	Sparkling	Year	Month
1990-08-31	1605.0	1990.0	8.0
1990-09-30	2424.0	1990.0	9.0
1990-10-31	3116.0	1990.0	10.0
1990-11-30	4286.0	1990.0	11.0
1990-12-31	6047.0	1990.0	12.0

Train Dataset

```
Train:  (132, 3)
Test:  (55, 3)
```

Shape of Test(55 rows and 3 columns) and Train (132 rows and 3 columns)

First few rows of Test Data

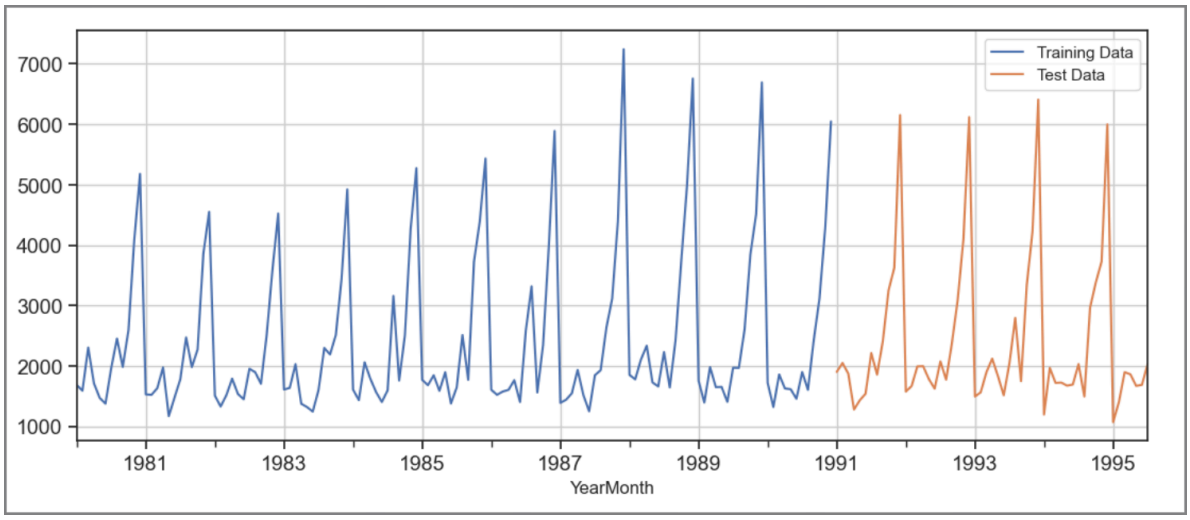
YearMonth	Sparkling	Year	Month
1991-01-31	1902.0	1991.0	1.0
1991-02-28	2049.0	1991.0	2.0
1991-03-31	1874.0	1991.0	3.0
1991-04-30	1279.0	1991.0	4.0
1991-05-31	1432.0	1991.0	5.0

Last few rows of Test Data

YearMonth	Sparkling	Year	Month
1995-03-31	1897.0	1995.0	3.0
1995-04-30	1862.0	1995.0	4.0
1995-05-31	1670.0	1995.0	5.0
1995-06-30	1688.0	1995.0	6.0
1995-07-31	2031.0	1995.0	7.0

Test Dataset

Plot 11 - Plot of train and test

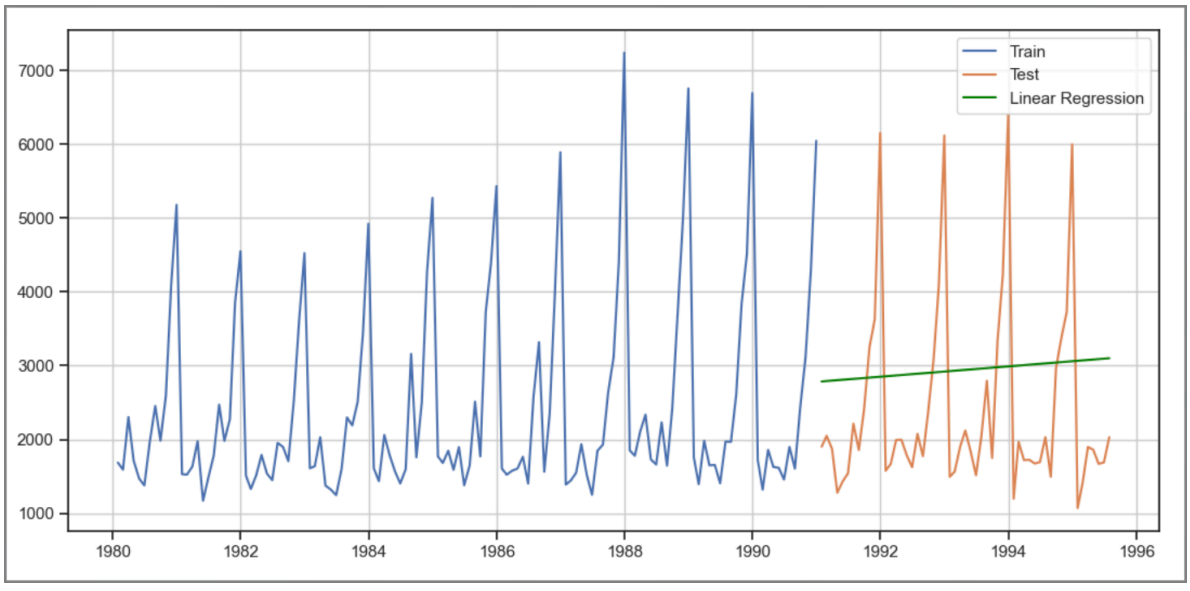


Plot of train and test

3. Model Building

(1) **Linear Regression**

Plot 12 - Linear Regression



Green Line indicates Linear Regression Prediction

RMSE calculated for Linear Regression: ***1386.84***

```
For RegressionOnTime forecast on the Test Data,  RMSE is 1386.84
```

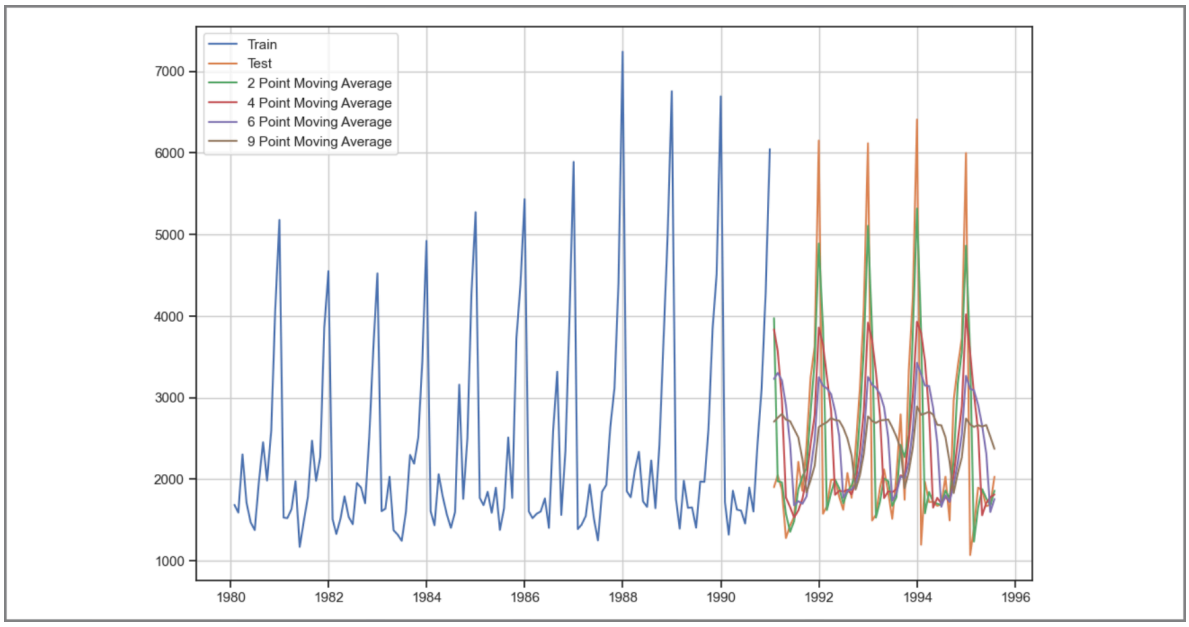
(2) **Moving Average (MA)**

For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here. For Moving Average, we are going to average over the entire data.

YearMonth	Sparkling	Year	Month	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	1686.0	1980.0	1.0	NaN	NaN	NaN	NaN
1980-02-29	1591.0	1980.0	2.0	1638.5	NaN	NaN	NaN
1980-03-31	2304.0	1980.0	3.0	1947.5	NaN	NaN	NaN
1980-04-30	1712.0	1980.0	4.0	2008.0	1823.25	NaN	NaN
1980-05-31	1471.0	1980.0	5.0	1591.5	1769.50	NaN	NaN

Top 5 rows for Trailing Moving average for 2, 4, 6 and 9

Plot 13 -Moving Average (MA)

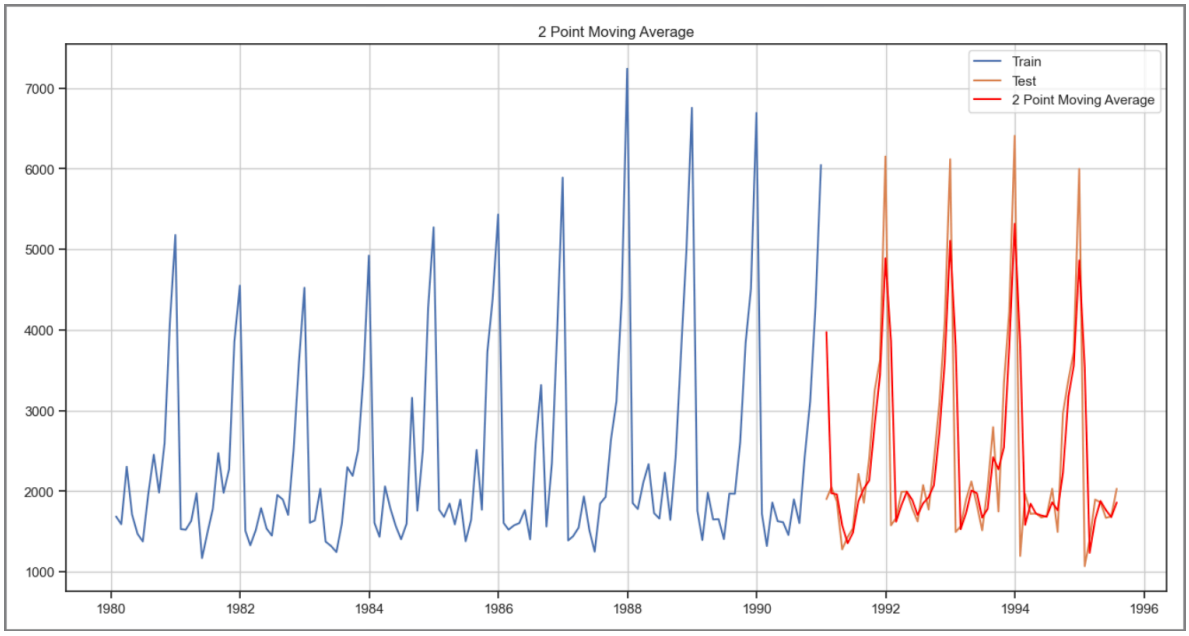


Moving Average plot - 2 point Moving Average is best

RMSE calculated for Moving Average:

```
For 2 point Moving Average Model forecast on the Testing Data,  RMSE is 813.401
For 4 point Moving Average Model forecast on the Testing Data,  RMSE is 1156.590
For 6 point Moving Average Model forecast on the Testing Data,  RMSE is 1283.927
For 9 point Moving Average Model forecast on the Testing Data,  RMSE is 1346.278
```

We created several moving average models with rolling windows ranging from 2 to 9 points. The best model was the 2-point moving average, with a RMSE value of 813.401.



Plot for the best Moving Average that is rolling window 2.

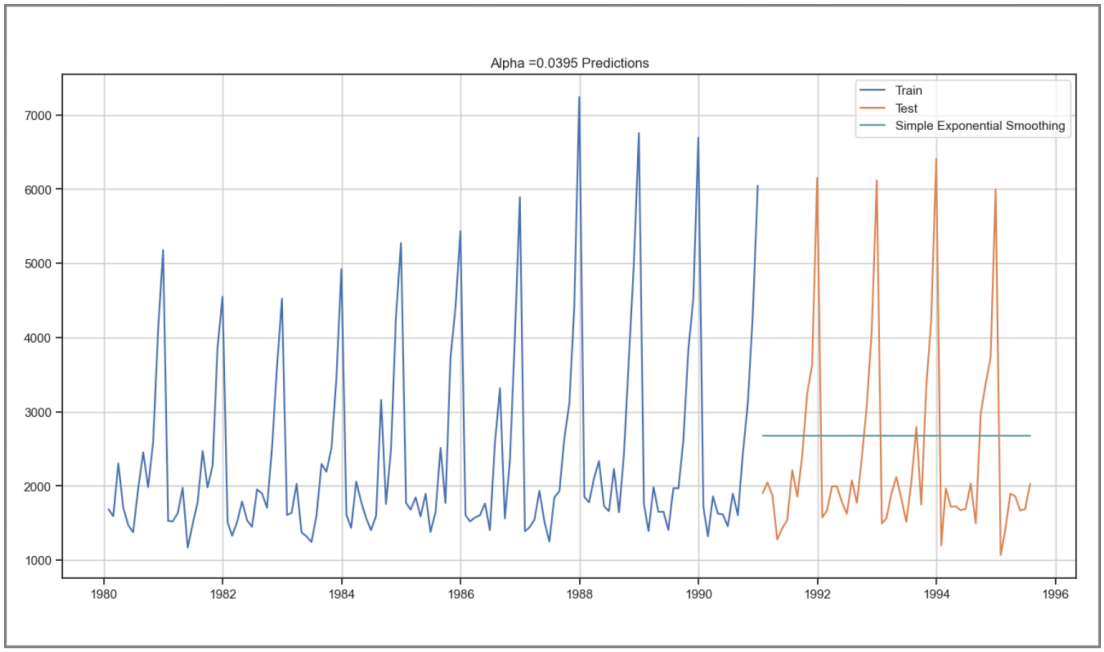
(3) **Simple Exponential Smoothening Model - Plot 14**

A Simple Exponential Smoothing (SES) model is a time series forecasting technique that applies weighted averages of past observations to make future predictions. In SES, more recent observations are given exponentially more weight compared to older observations, allowing the model to adapt quickly to changes in the data.

The SES model is particularly useful for data with no clear trend or seasonal pattern, as it effectively smooths out short-term fluctuations to reveal longer-term trends or patterns.

YearMonth	Sparkling	Year	Month	predict
1991-01-31	1902.0	1991.0	1.0	2676.676366
1991-02-28	2049.0	1991.0	2.0	2676.676366
1991-03-31	1874.0	1991.0	3.0	2676.676366
1991-04-30	1279.0	1991.0	4.0	2676.676366
1991-05-31	1432.0	1991.0	5.0	2676.676366

Table for forecast of the model



'smoothing_level': 0.03953488372093023

RMSE

```
Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1304.927
```

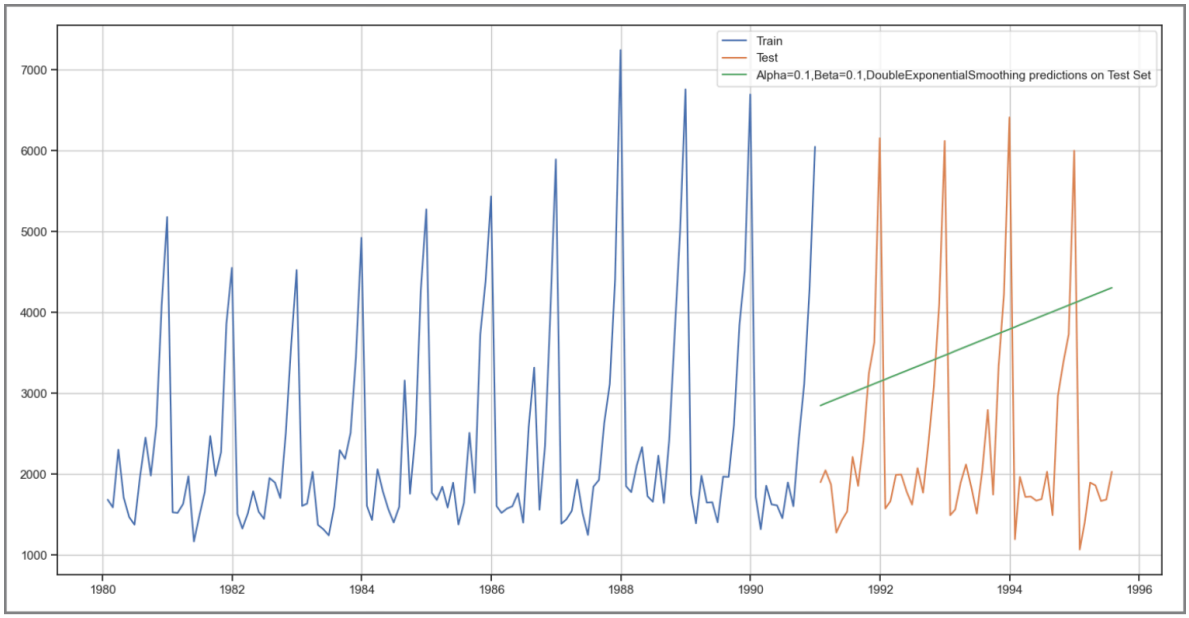
(4) **Double Exponential Smoothening (Holt's Model)**

Double Exponential Smoothing (DES), also known as Holt's Exponential Smoothing, is an extension of Simple Exponential Smoothing that incorporates both level and trend components to handle time series data with trends.

The DES method helps in capturing both the level and the trend in the time series, making it suitable for datasets where trends are present, thus providing more accurate forecasts compared to Simple Exponential Smoothing when trends exist in the data.

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

Plot 15 - Holt's Model



Holt's Model

RMSE

α-0.1 and β-0.1 Double Exponential Smoothing Model forecast on the Test Data, RMSE is 1778.565

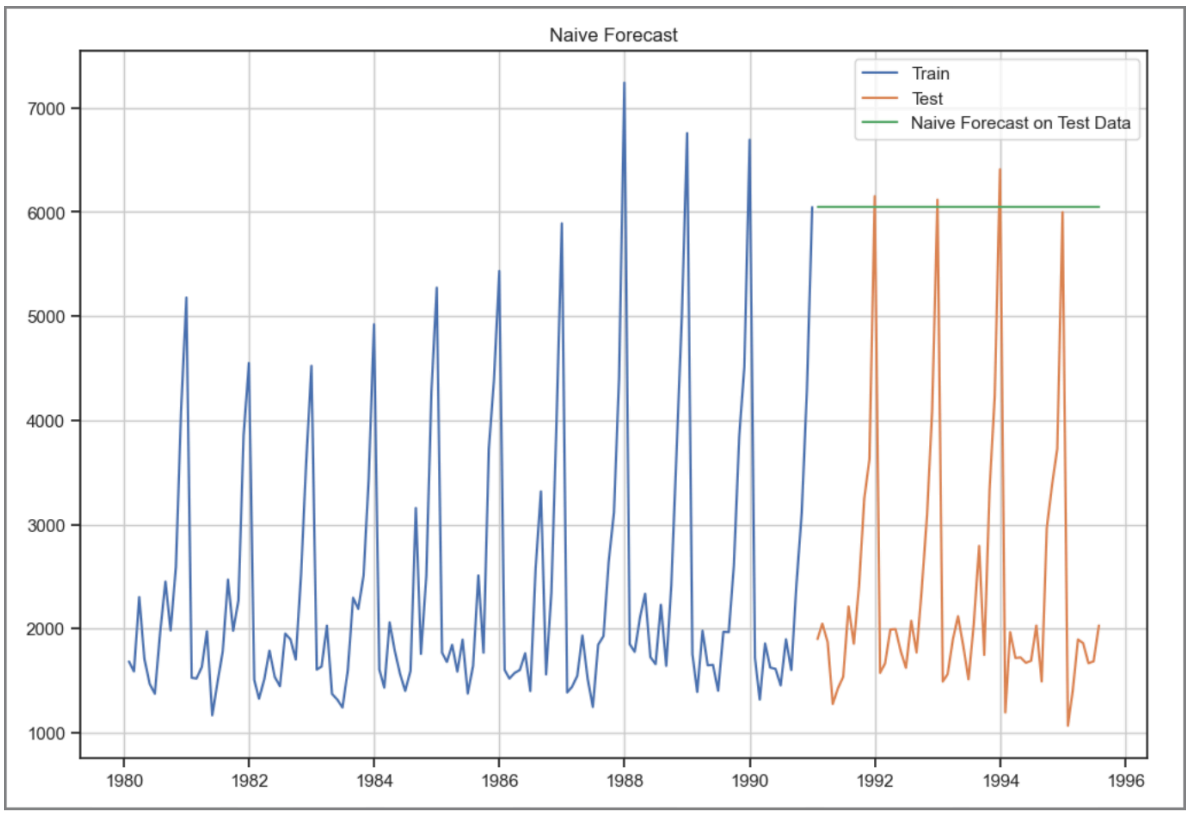
(5) **Naive Approach**

The Naive Approach is a simple and straightforward time series forecasting method where the forecast for any future period is assumed to be equal to the most recent actual observation.

```
YearMonth
1991-01-31    6047.0
1991-02-28    6047.0
1991-03-31    6047.0
1991-04-30    6047.0
1991-05-31    6047.0
Freq: M, Name: naive, dtype: float64
```

forecast for test prediction is assumed 132

Plot 16 Naive Approach



Plot for Naive Approach

RMSE for Naive Approach

```
Naive Model forecast on the Test Data, RMSE is 3864.279
```

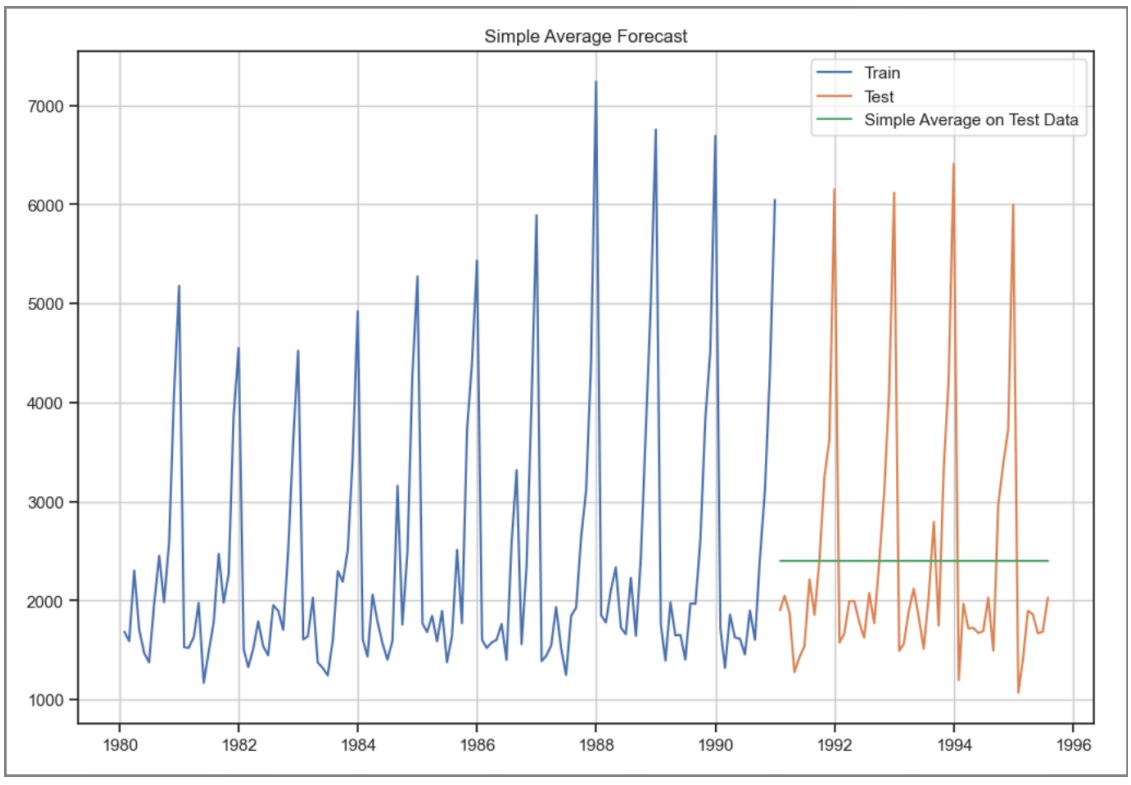
(6) **Simple Average**

The Simple Average Time Series Forecasting (TSF) model is a basic yet effective method for predicting future values in a time series. It operates on the principle that the forecasted value for a given period is the simple average (arithmetic mean) of all previous observations. This model is particularly useful for data with a consistent level over time and without significant trends or seasonal patterns.

YearMonth	Sparkling	Year	Month	mean_forecast
1991-01-31	1902.0	1991.0	1.0	2403.780303
1991-02-28	2049.0	1991.0	2.0	2403.780303
1991-03-31	1874.0	1991.0	3.0	2403.780303
1991-04-30	1279.0	1991.0	4.0	2403.780303
1991-05-31	1432.0	1991.0	5.0	2403.780303

Simple Average Test mean forecast predicted values

Plot 17 Simple Average



Simple Average Plot

RMSE for Simple Average

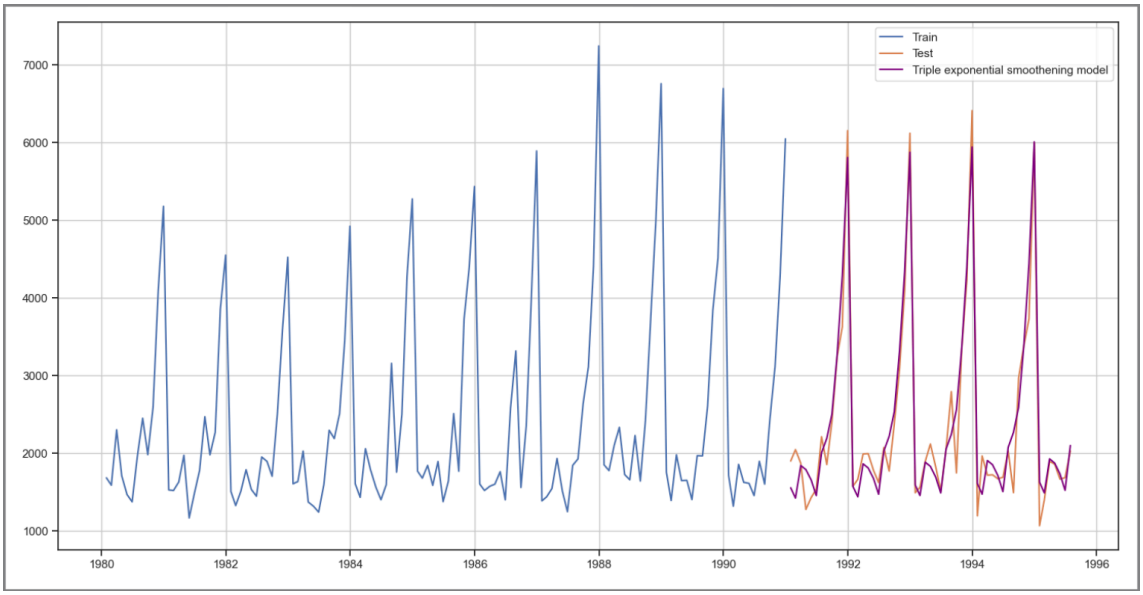
```
Simple Average forecast on the Test Data, RMSE is 1275.082
```

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

Triple Exponential Smoothing, also known as the Holt-Winters method, is an extension of Exponential Smoothing that accounts for trends and seasonality in time series data. This method involves three components: level, trend, and seasonality. There are two main variations of the Holt-Winters method: additive and multiplicative. The additive model is used when seasonal variations are roughly constant over time, while the multiplicative model is used when seasonal variations change proportionally to the level of the time series.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE	Method
1301	0.4	0.1	0.2	384.467709	317.434302	ta_sm
2245	0.4	0.1	0.3	381.106645	326.579641	tm_sm
1211	0.3	0.2	0.2	388.544148	329.037543	ta_sm
1200	0.3	0.1	0.1	388.220071	337.080969	ta_sm
1110	0.2	0.2	0.1	398.482510	340.186457	ta_sm

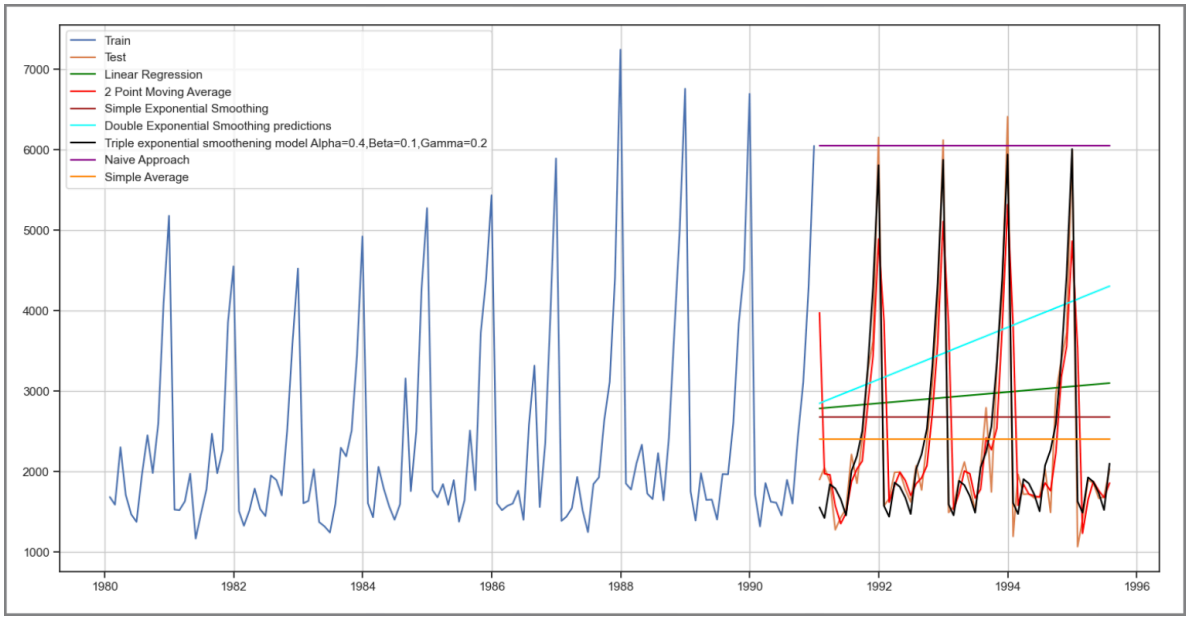
Plot 18 - Triple Exponential Smoothing



RMSE = 317.434

The optimal Triple Exponential Smoothing (Holt-Winters) model, featuring an additive trend and multiplicative seasonality, has been identified. The best smoothing parameters alpha, beta and gamma have been determined, making this the most effective model so far.

Plot 19 - Model Building for all forecast



Plot of all the forecasting models

```
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing Trend - Additive and Seasonality - Multiplicative 317.43430200265055
```

RMSE = 317.434

Sorted by RMSE values on the Test Data:

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing Trend - Additive and Seasonality - Multiplicative	317.434302
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
Simple Average	1275.081804
6pointTrailingMovingAverage	1283.927428
Alpha=0.0395,SimpleExponentialSmoothing	1304.927405
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,Beta=0.1, Double Exponential Smoothing prediction	1382.520870
Linear Regression	1386.836243
α-0.1 and β-0.1 Double Exponential Smoothing Model	1778.564670
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	1778.564670
Naive Model	3864.279352

RMSE Value in sorted way for all the building

3. 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.

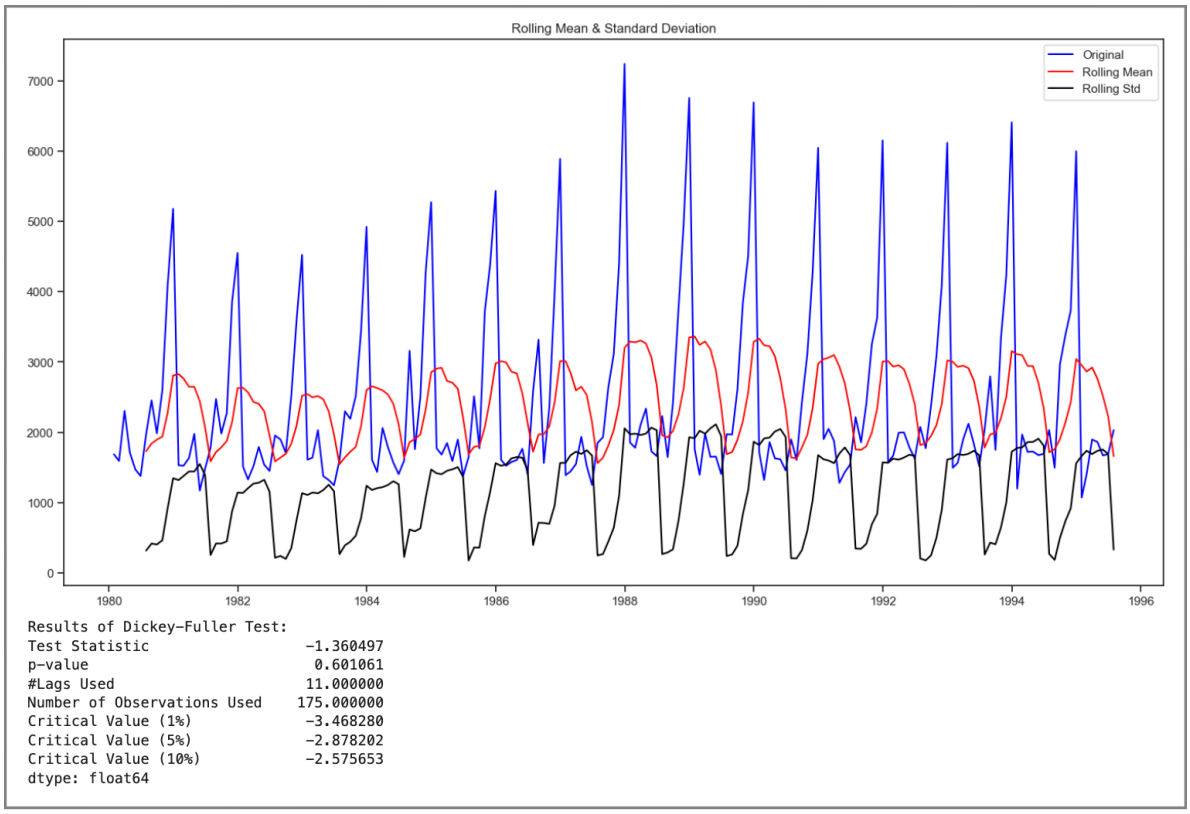
The hypothesis in a simple form for the ADF test is:

- 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.

We see that at 5% significant level the Time Series is non-stationary.

Plot 20 - The Augmented Dickey-Fuller test

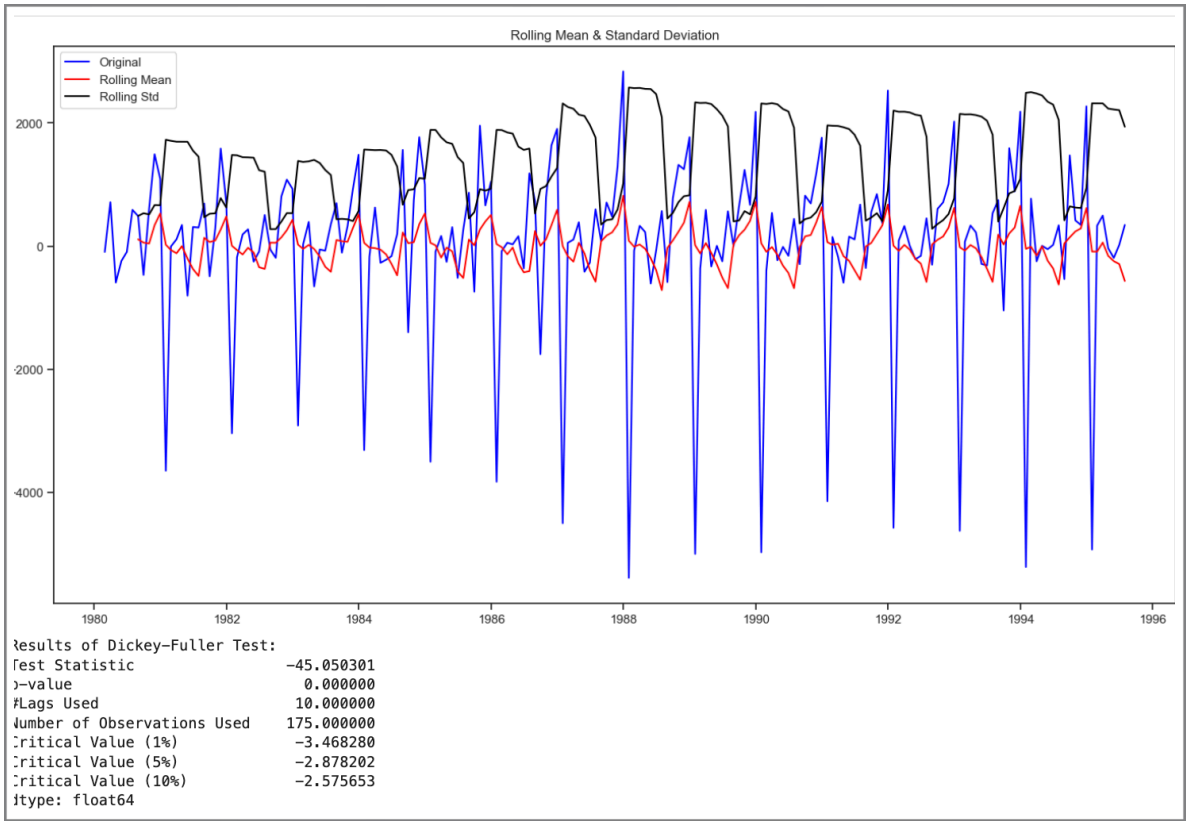


Dickey-Fuller Test

The Dickey-Fuller test results indicate:

- The test statistic value is -1.360497.
- The p-value is 0.601061.
- 11 lags were used in the test.
- 175 observations were used.
- Critical Value (1%): -3.468280, Critical Value (5%): -2.878202 and Critical Value (10%): -2.575653
- These results suggest that the time series data is not stationary, as the p-value is greater than the significance level of 0.05. Additionally, the test statistic is higher than the critical values at all confidence levels, further supporting non-stationarity. Therefore, we fail to reject the null hypothesis, indicating that the series is likely non-stationary.

Plot 21 - The Augmented Dickey-Fuller test after difference



AdF test after difference

The Dickey-Fuller test, performed after differencing the data, is used to test the null hypothesis that a unit root is present in the time series sample. Here are the key points regarding the null hypothesis and the interpretation of the test results:

1. Null Hypothesis (H0): The time series has a unit root (i.e., it is non-stationary).
2. Alternative Hypothesis (H1): The time series does not have a unit root (i.e., it is stationary).

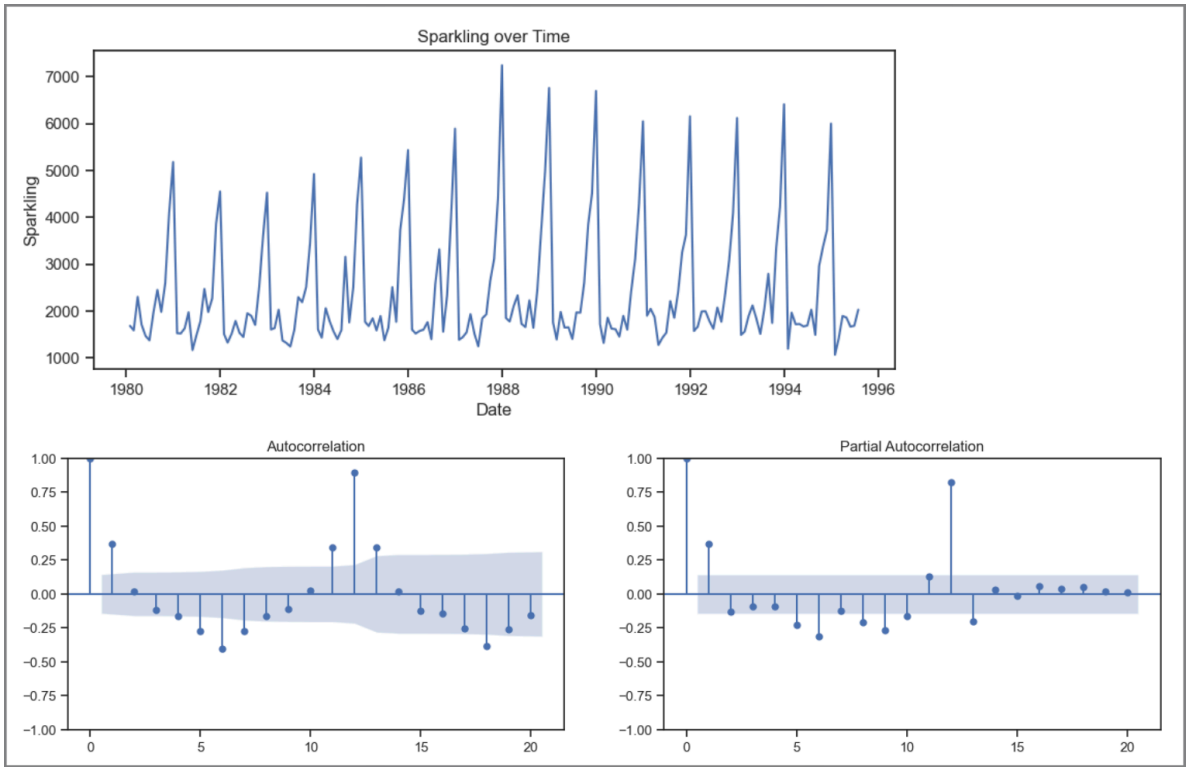
Interpretation of the Results:

After differencing the time series data, the results of the Dickey-Fuller test are as follows:

- Test Statistic: -45.050301
- p-value: 0.000000
- Lags Used: 10
- Number of Observations Used: 175
- Critical Value (1%): -3.468280
- Critical Value (5%): -2.878202
- Critical Value (10%): -2.575653
- With a significantly low p-value of 0.000000, which is less than the typical significance level of 0.05, we can reject the null hypothesis. The test statistic is also much lower than the critical values at all confidence levels. These results indicate that after differencing, the time series data becomes stationary.
- This indicates that the time series is stationary, meaning its statistical properties such as mean and variance remain constant over time.

5. Model Building - Stationary Data

Plot 22 - Plot the Autocorrelation and the Partial Autocorrelation function plots on the whole data.



Plot the Autocorrelation and the Partial Autocorrelation function plots on the whole data.

(1) Auto ARIMA (Auto-Regressive Integrated Moving Average)

The Auto ARIMA (Auto-Regressive Integrated Moving Average) model is a statistical analysis technique used for time series forecasting that automatically selects the best-fitting ARIMA model by optimizing the parameters. ARIMA models are widely used for forecasting data that show evidence of non-stationarity and require differencing to achieve stationarity. The ARIMA model comprises three components: the auto-regressive (AR) part, which regresses the variable on its own lagged values; the integrated (I) part, which involves differencing the observations to make the time series stationary; and the moving average (MA) part, which models the error term

as a linear combination of past error terms. The Auto ARIMA model automates the identification of optimal values for these parameters (p, d, q) by evaluating multiple ARIMA models with different combinations and selecting the best one based on information criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). This process includes differencing the series to achieve stationarity, exploring a range of p and q values, and evaluating each model to find the one with the lowest criterion value.

For Sparkling wine sales analysis, the parameter d represents the differencing required to render the series stationary. The for loop iterates over p and q values ranging from 0 to 3, while a fixed value of 1 is assigned to d. This choice is made because we had previously determined through the Augmented Dickey-Fuller (ADF) test that a differencing order of 1 was necessary to achieve stationarity.

Some parameter combinations for the Model:

```
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)
```

Parameters for Auto ARIMA

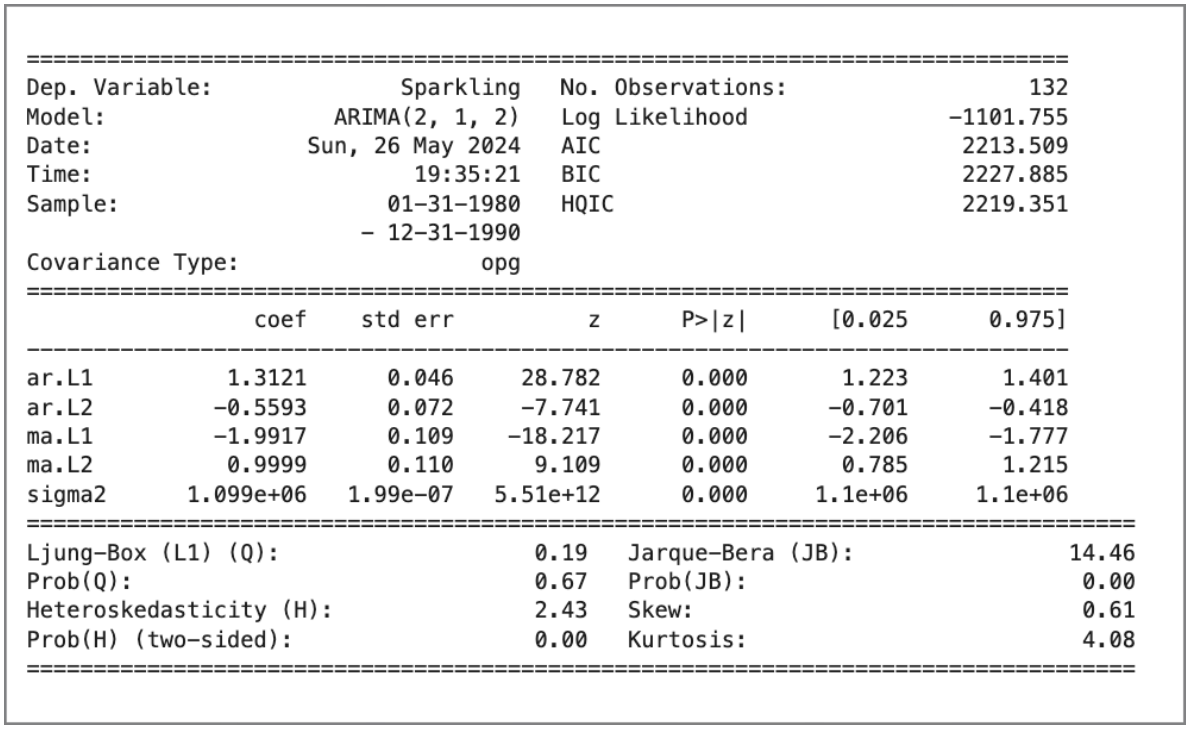
To obtain the parameters corresponding to the minimum AIC value, we need to sort the AIC values in ascending order and then select the parameters associated with the lowest AIC value.

	param	AIC
10	(2, 1, 2)	2213.509212
15	(3, 1, 3)	2221.452537
14	(3, 1, 2)	2230.816576
11	(2, 1, 3)	2232.917659
9	(2, 1, 1)	2233.777626
3	(0, 1, 3)	2233.994858
2	(0, 1, 2)	2234.408323
6	(1, 1, 2)	2234.527200
13	(3, 1, 1)	2235.499067
7	(1, 1, 3)	2235.607814
5	(1, 1, 1)	2235.755095
12	(3, 1, 0)	2257.723379
8	(2, 1, 0)	2260.365744
1	(0, 1, 1)	2263.060016
4	(1, 1, 0)	2266.608539
0	(0, 1, 0)	2267.663036

Arranged the AIC values in ascending order to identify the set of parameters that yield the minimum AIC value.

p=2, d=1 and q=2 has the minimum AIC value of 2213.509212.

We will generate the summary report for this.



```
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sun, 26 May 2024   AIC                           2213.509
Time:                        19:35:21   BIC                           2227.885
Sample:                    01-31-1980   HQIC                          2219.351
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.782      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.741      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.217      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.109      0.000       0.785       1.215
sigma2      1.099e+06   1.99e-07   5.51e+12      0.000     1.1e+06     1.1e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.19   Jarque-Bera (JB):                 14.46
Prob(Q):                              0.67   Prob(JB):                          0.00
Heteroskedasticity (H):               2.43   Skew:                              0.61
Prob(H) (two-sided):                  0.00   Kurtosis:                          4.08
===================================================================================
```

Auto ARIMA Summary Report for p=2, d=1 and q=2

The summary report for the Auto ARIMA model offers a detailed overview of the model's performance and diagnostics. It begins by identifying the dependent variable, labeled as "Sparkling," and specifies that 132 observations were utilized in the analysis. The chosen ARIMA model is denoted as ARIMA(2, 1, 2), indicating auto-regressive and moving average orders of 2 and 2, respectively, with a differencing order of 1. The log likelihood, AIC (Akaike Information Criterion) - 1274.695 , and BIC (Bayesian Information Criterion) - 2213.509212 values provide measures of model fit, with lower AIC and BIC values indicating better fit. Additionally, the report includes parameter estimates for the model coefficients, standard errors, and statistical significance. Diagnostic tests such as the Ljung-Box (Q) - 0.19 and Jarque-Bera (JB) - 14.46 tests assess the goodness of fit, while the Heteroskedasticity (H) - 2.43 test evaluates the constancy of residual variance. Skewness and kurtosis measures provide insights into the distributional properties of the residuals. Overall, this comprehensive summary aids in the interpretation and evaluation of the Auto ARIMA model, helping to understand its effectiveness in capturing the underlying patterns in the time series data.

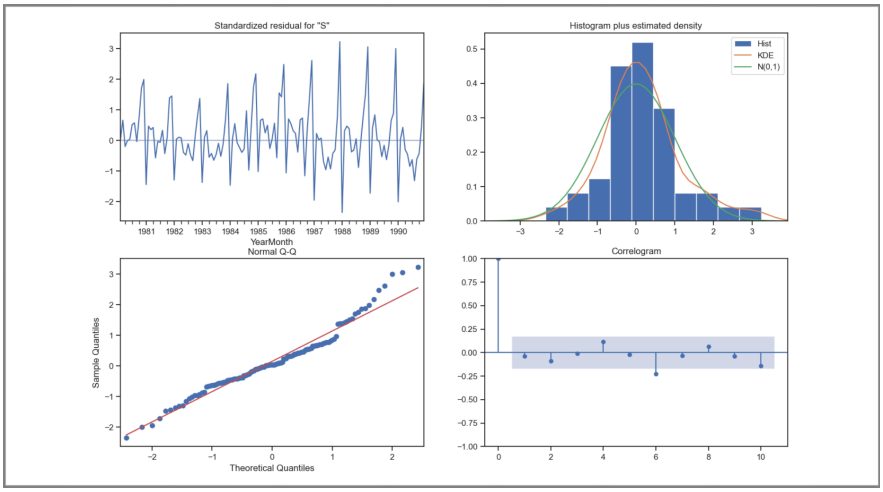
Forecast for Auto ARIMA model before we evaluate the RMSE

```
1991-01-31    4252.361455
1991-02-28    2863.104969
1991-03-31    2043.981140
1991-04-30    1746.206002
1991-05-31    1813.625098
1991-06-30    2068.631102
1991-07-31    2365.520097
1991-08-31    2612.447507
1991-09-30    2770.394735
1991-10-31    2839.533406
1991-11-30    2841.911681
1991-12-31    2806.363215
1992-01-31    2758.389507
1992-02-29    2715.324800
1992-03-31    2685.650644
1992-04-30    2670.800795
1992-05-31    2667.912797
1992-06-30    2672.428892
1992-07-31    2679.969760
1992-08-31    2687.338381
1992-09-30    2692.789250
1992-10-31    2695.820152
1992-11-30    2696.748380
1992-12-31    2696.271145
1993-01-31    2695.125805
1993-02-28    2693.889907
1993-03-31    2692.908856
1993-04-30    2692.312842
1993-05-31    2692.079503
1993-06-30    2692.106686
1993-07-31    2692.272858
1993-08-31    2692.475691
1993-09-30    2692.648891
1993-10-31    2692.762705
1993-11-30    2692.815171
1993-12-31    2692.820357
1994-01-31    2692.797818
1994-02-28    2692.765343
1994-03-31    2692.735338
1994-04-30    2692.714132
1994-05-31    2692.703089
1994-06-30    2692.700459
1994-07-31    2692.703185
1994-08-31    2692.708233
1994-09-30    2692.713331
1994-10-31    2692.717198
1994-11-30    2692.719420
1994-12-31    2692.720173
1995-01-31    2692.719918
1995-02-28    2692.719162
1995-03-31    2692.718313
1995-04-30    2692.717622
1995-05-31    2692.717190
1995-06-30    2692.717010
1995-07-31    2692.717015
```

RMSE value for Auto ARIMA = 35.96

```
RMSE for Auto ARIMA model for test data: 1299.9798975363012
```

Plot 23 - Diagnostic plot for auto ARIMA for the best auto ARIMA model



Diagnostic Plot for auto ARIMA

(2) Auto SARIMA (Seasonal Auto-Regressive Integrated Moving Average)

SARIMA, which stands for Seasonal Auto-Regressive Integrated Moving Average, extends the ARIMA model to account for seasonal patterns in the data.

Components of SARIMA

1. **Auto-Regressive (AR) part**: Represents the correlation between the current observation and a lagged (past) observation within the same series.
2. **Integrated (I) part**: Involves differencing the raw observations to make the time series stationary. This accounts for trends present in the data.
3. **Moving Average (MA) part**: Represents the correlation between the current observation and a residual error from a moving average model applied to lagged observations.

Additionally, SARIMA includes seasonal components:

1. **Seasonal Auto-Regressive (SAR) part**: Represents the correlation between the current observation and a lagged observation within the same series, but over seasonal intervals.
2. **Seasonal Integrated (SI) part**: Involves seasonal differencing to remove seasonal trends from the data.
3. **Seasonal Moving Average (SMA) part**: Represents the correlation between the current observation and a residual error from a moving average model applied to lagged observations over seasonal intervals.

Overall, Auto SARIMA helps you forecast future values in your data easily and accurately by automatically finding the best way to do it.

For Sparkling wine sales analysis, the parameter d represents the differencing required to render the series stationary. The for loop iterates over p, d and q values ranging from 0 to 2. The parameter m represent number of seasonals months. We are keeping seasonal month as 12. This choice is made because we had previously determined through the Augmented Dickey-Fuller (ADF) test that a differencing order of 1 was necessary to achieve stationarity.

To obtain the parameters corresponding to the minimum AIC value, we need to sort the AIC values in ascending order and then select the parameters associated with the lowest AIC value.

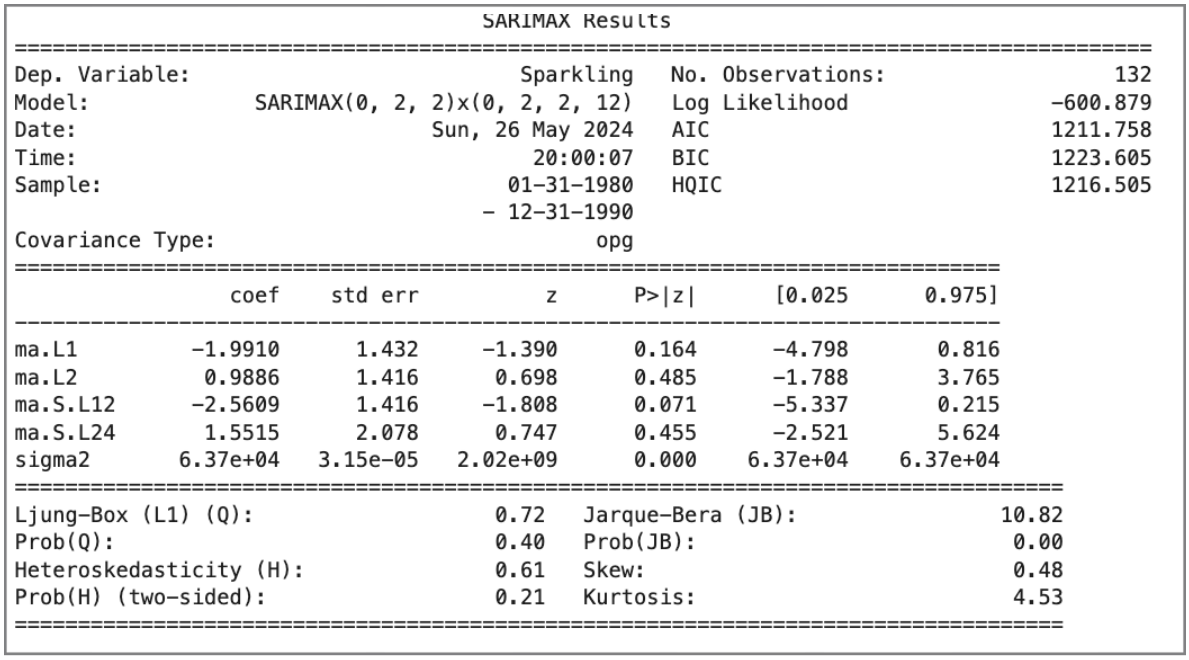
p=0, d=2 and q=2 has the minimum AIC value of 1211.69870
Seasonal p=0, d=2 ,q=2 and m=12.

	param	seasonal	AIC
224	(0, 2, 2)	(0, 2, 2, 12)	1211.698701
467	(1, 2, 2)	(0, 2, 2, 12)	1212.339656
233	(0, 2, 2)	(1, 2, 2, 12)	1213.528381
710	(2, 2, 2)	(0, 2, 2, 12)	1214.130998
476	(1, 2, 2)	(1, 2, 2, 12)	1214.336021

Top 5 rows for Auto SARIMA based on the minimum AIC value

The summary report for Auto SARIMA

The summary report provides a detailed overview of observations derived from a SARIMAX model. The analyzed data comprises 132 observations of the dependent variable labeled as "Sparkling." The SARIMAX model, characterized as SARIMAX(0, 2, 2)x(0, 2, 2, 12), encompasses seasonal and exogenous factors in its formulation. Evaluating the model's fit, the log



```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                          Sparkling   No. Observations:                  132
Model:             SARIMAX(0, 2, 2)x(0, 2, 2, 12)   Log Likelihood                -600.879
Date:                            Sun, 26 May 2024   AIC                           1211.758
Time:                                    20:00:07   BIC                           1223.605
Sample:                                01-31-1980   HQIC                          1216.505
                                     - 12-31-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -1.9910      1.432     -1.390      0.164      -4.798       0.816
ma.L2          0.9886      1.416      0.698      0.485      -1.788       3.765
ma.S.L12      -2.5609      1.416     -1.808      0.071      -5.337       0.215
ma.S.L24       1.5515      2.078      0.747      0.455      -2.521       5.624
sigma2       6.37e+04   3.15e-05   2.02e+09      0.000    6.37e+04    6.37e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.72   Jarque-Bera (JB):                 10.82
Prob(Q):                              0.40   Prob(JB):                          0.00
Heteroskedasticity (H):               0.61   Skew:                              0.48
Prob(H) (two-sided):                  0.21   Kurtosis:                          4.53
===================================================================================
```

Summary Report for Auto SARIMA
p=0, d=2 and q=2 has the minimum AIC value of 1211.76

Seasonal p=0, d=2 ,q=2 and m=12

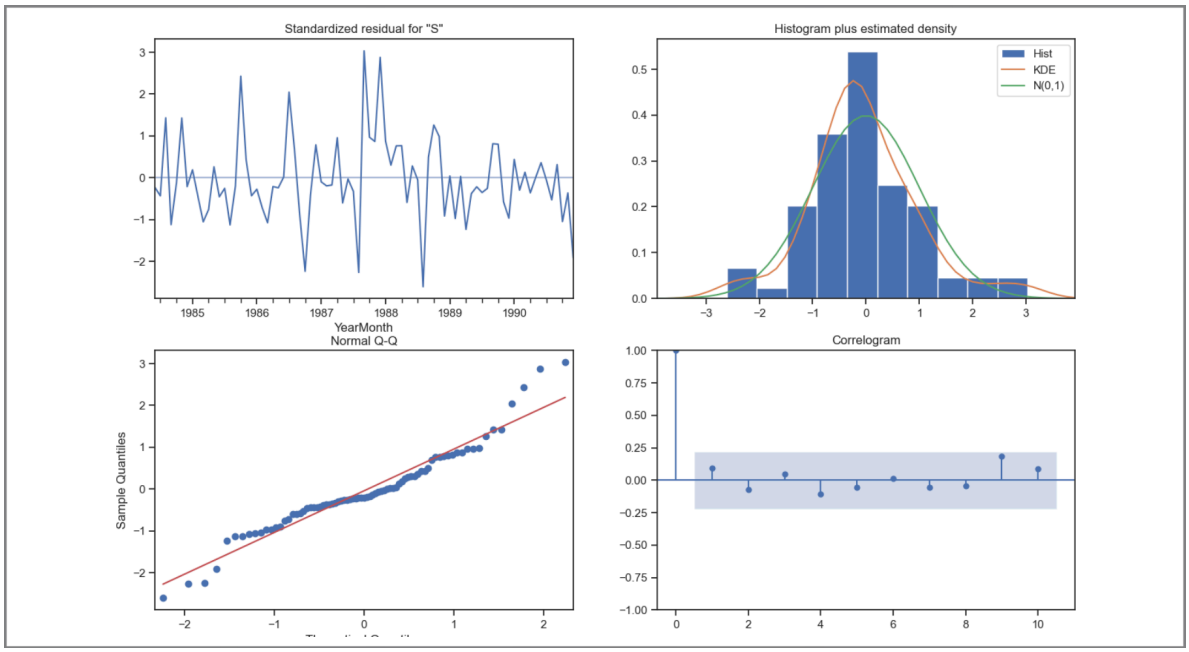
likelihood is reported as -600.879, indicating how well the model aligns with the data, while the AIC and BIC stand at 1211.758 and 1223.605, respectively, serving as measures of model fit and complexity. The temporal span of the dataset extends from January 31, 1980, to December 31, 1990. Covariance estimation of the model is identified as "opg." Parameter estimates offer insights into the coefficients of model terms, alongside their associated standard errors and statistical significance. Diagnostic tests encompass the Ljung-Box (Q) test, Jarque-Bera (JB) test, and a test for heteroskedasticity (H), assessing various assumptions underlying the model. Additionally, skewness and kurtosis measures provide further characterization of the distributional properties of residuals. Overall, the summary furnishes a comprehensive assessment of the SARIMAX model's performance, encompassing its alignment with data, parameter significance, and adherence to underlying assumptions.

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1991-01-31	1190.581498	448.873039	310.806508	2070.356488
1991-02-28	968.313061	448.618354	89.037245	1847.588877
1991-03-31	1400.634232	452.178984	514.379709	2286.888755
1991-04-30	1274.802863	455.664763	381.716339	2167.889387
1991-05-31	1190.214253	459.458800	289.691553	2090.736954

Auto SARIMA forecast values

```
RMSE for Auto SARIMA model for test data:  2297.6862838489883
```

RMSE



Diagnostic Plot for Auto SARIMA

Plot 24 - Diagnostic plot for auto SARIMA for the best auto SARIMA model

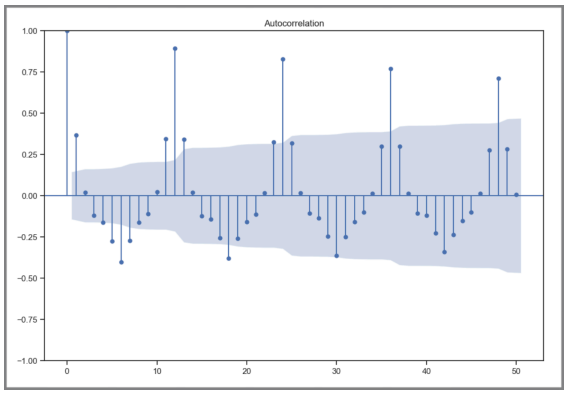
(3) Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE

- **Manual ARIMA**

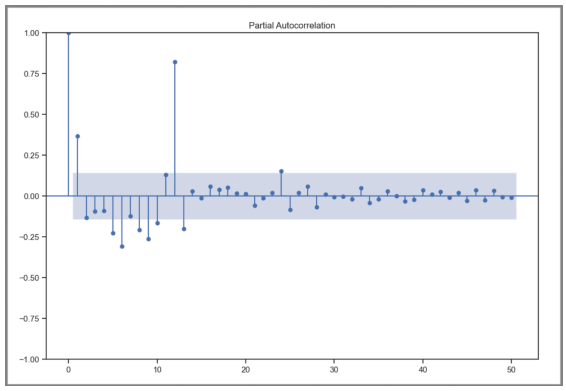
In manual ARIMA, the user manually selects appropriate values for p, d and q based on prior knowledge, domain expertise, or iterative testing.

This approach requires a deep understanding of the data and the underlying patterns to choose the most suitable parameters. Manual ARIMA is often used when automated methods like Auto ARIMA or Auto SARIMA are not available or when users prefer a more hands-on approach to model selection. However, it can be time-consuming and may not always yield the best results compared to automated approaches.

Plot 25 - ACF Plot on train data

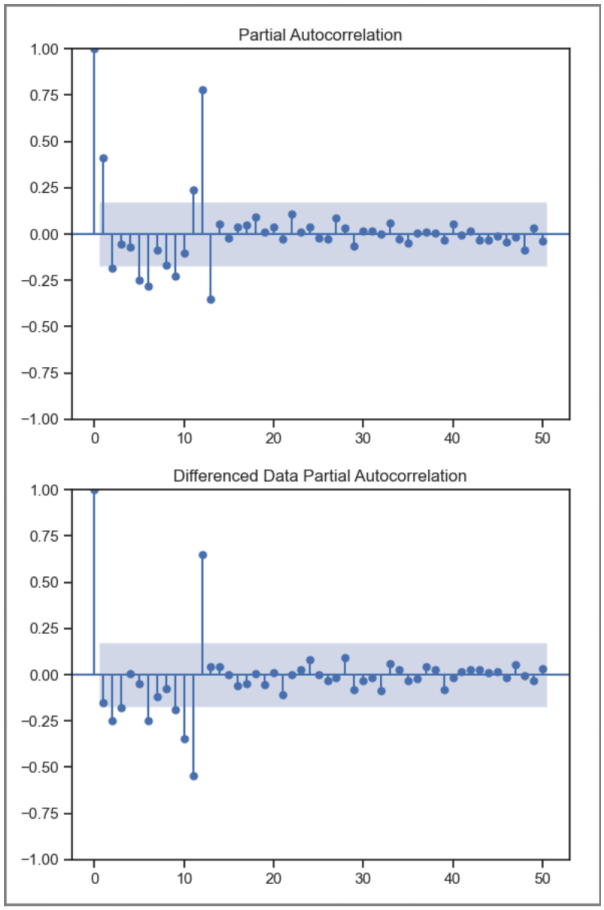


ACF Plot on train data



PACF Plot on Partial Auto-correlation

Plot 26 - PACF Plot on train data



PACF Plot on train data

Value selected for manual ARIMA: p=3, q=1 and d=1

```
==============================================================================
Dep. Variable:                Sparkling   No. Observations:                  132
Model:                 ARIMA(3, 1, 1)   Log Likelihood               -1112.750
Date:                Sun, 26 May 2024   AIC                           2235.499
Time:                        20:13:48   BIC                           2249.875
Sample:                    01-31-1980   HQIC                          2241.341
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.5102      0.119      4.294      0.000       0.277       0.743
ar.L2         -0.1560      0.206     -0.756      0.449      -0.560       0.248
ar.L3         -0.0469      0.195     -0.241      0.810      -0.429       0.335
ma.L1         -0.9982      0.250     -3.988      0.000      -1.489      -0.508
sigma2      1.353e+06   2.92e+05      4.630      0.000     7.8e+05    1.93e+06
==============================================================================
Ljung-Box (L1) (Q):                   0.06   Jarque-Bera (JB):                12.63
Prob(Q):                              0.80   Prob(JB):                         0.00
Heteroskedasticity (H):               2.74   Skew:                             0.51
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.13
```

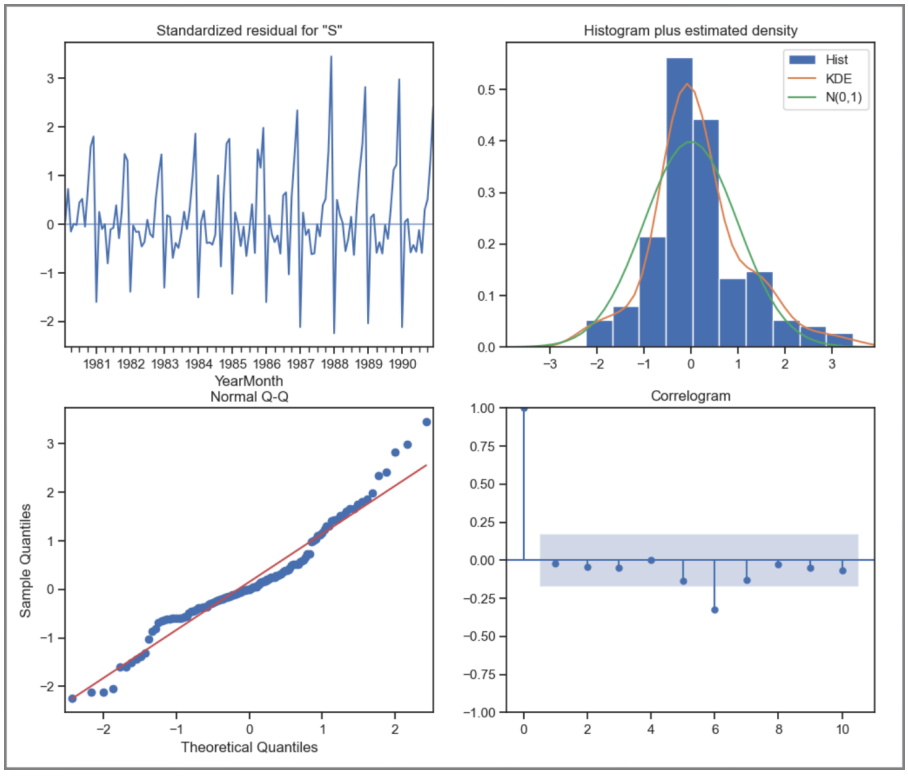
Manual ARIMA p=1, q=1 and d=1

The manual ARIMA model with parameters (3, 1, 1) was fitted to the Sparkling wine sales data. The model suggests a moderately positive coefficient for the first autoregressive term (AR.L1) and a negative coefficient for the first moving average term (MA.L1). However, the coefficients for the second and third autoregressive terms (AR.L2 and AR.L3) are not statistically significant. The log likelihood is -1112.750, and the AIC is 2235.499, indicating a relatively good fit of the model to the data. The Ljung-Box test indicates no significant autocorrelation at lag 1, and the Jarque-Bera test suggests non-normality in the residuals. Overall, the model provides insights into the dynamics of Sparkling wine sales but may benefit from further refinement. the manual ARIMA model provides insights into the relationships between the variables and their predictive capabilities within the dataset.

RSME value for Manual ARIMA

```
RMSE for manual ARIMA model: 1297.1152862087922
```

Plot 27 - Manual ARIMA diagnostic



Manual ARIMA - p-3, q-1 and q-1

- **Manual SARIMA**

The manual SARIMA (Seasonal Auto-Regressive Integrated Moving Average) model is a technique for time series forecasting where the user manually selects the values of the SARIMA parameters to capture both the seasonal and non-seasonal patterns present in the data.

Once the parameters are selected, the manual SARIMA model is fitted to the data, and forecasts can be generated for future time points. Diagnostic tests and evaluation metrics are then used to assess the model's performance and determine if adjustments to the parameter values are necessary.

Manual SARIMA offers flexibility and control over the modelling process, but it requires expertise in time series analysis and a deep understanding of the data to select appropriate parameter values that result in accurate forecasts.

Value selected for manual SARIMA: p=1, q=2 and d=2

Seasonal p=1, q=2 m= 12 and d=2

The summary report for manual SARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(1, 2, 2)x(1, 2, 2, 12)	Log Likelihood	-600.168
Date:	Sun, 26 May 2024	AIC	1214.336
Time:	20:22:39	BIC	1230.922
Sample:	0	HQIC	1220.981
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.1457	0.146	0.998	0.318	-0.140	0.432
ma.L1	-2.0023	7.977	-0.251	0.802	-17.637	13.632
ma.L2	1.0030	8.006	0.125	0.900	-14.689	16.695
ar.S.L12	0.0017	0.056	0.030	0.976	-0.107	0.111
ma.S.L12	-1.6007	7.944	-0.201	0.840	-17.171	13.970
ma.S.L24	0.6011	4.755	0.126	0.899	-8.718	9.920
sigma2	1.591e+05	8.32e-05	1.91e+09	0.000	1.59e+05	1.59e+05

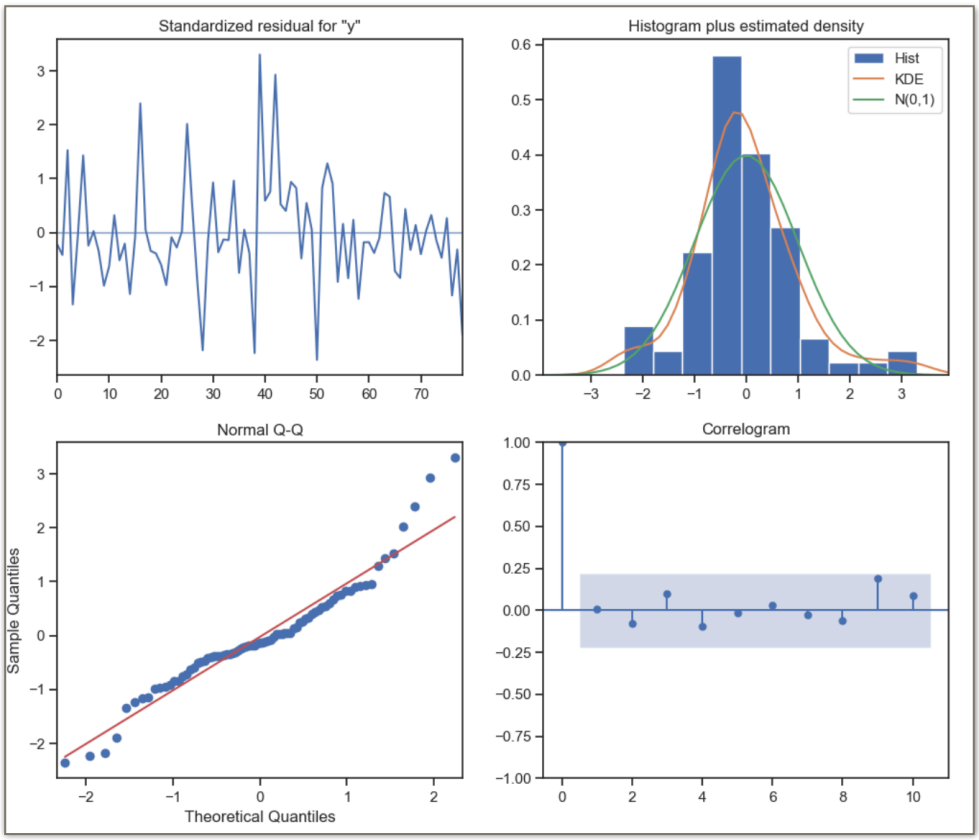
Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	16.14
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	0.58	Skew:	0.62
Prob(H) (two-sided):	0.18	Kurtosis:	4.83

Manual SARIMA (1, 2, 2)x(1, 2, 2, 12)

The manual SARIMA model with parameters (1, 2, 2)x(1, 2, 2, 12) was applied to the dataset, indicating the presence of seasonal differences. The log likelihood is -600.168, with an AIC of 1214.336, suggesting a reasonable fit of the model to the data. However, some coefficients, such as ar.L1 and ar.S.L12, are not statistically significant, indicating potential areas for improvement. The Ljung-Box test shows no significant autocorrelation at lag 1, and the Jarque-Bera test suggests non-normality in the residuals. Further refinement of the model may be beneficial for better predictive performance.

RSME for Manual SARIMA

```
RMSE for manual SARIMA model: 1976.9518
```



Diagnostic Plot for Manual SARIMA

6. Compare the performance of the models

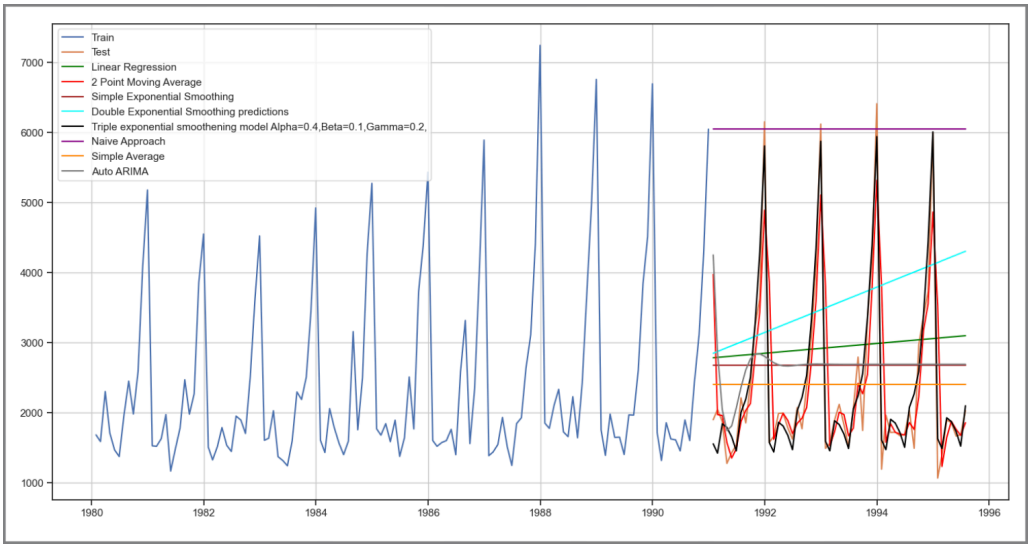
- **Compare the performance of the models**

We can see that Alpha = 0.4, Beta = 0.1 and Gamma = 0.2 Triple Exponential Smoothing has the lowest RSME value. So this will be considered as the best mode

Sorted by RMSE values on the Test Data:

	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing Trend - Additive and Seasonality - Multiplicative	317.434302
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
Simple Average	1275.081804
6pointTrailingMovingAverage	1283.927428
order=(1, 1, 1) - Manual ARIMA	1297.115286
Auto ARIMA(2, 1, 2)	1299.979898
Alpha=0.0395,SimpleExponentialSmoothing	1304.927405
9pointTrailingMovingAverage	1346.278315
Alpha=0.1,Beta=0.1, Double Exponential Smoothing prediction	1382.520870
Linear Regression	1386.836243
α-0.1 and β-0.1 Double Exponential Smoothing Model	1778.564670
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	1778.564670
order=(1, 2, 2),seasonal_order=(1, 2, 2, 12),Manual SARIMA	1976.951890
order=(0,2,2),seasonal_order=(0, 2, 2, 12),Auto_SARIMA	2297.686284
Naive Model	3864.279352

Sorted by RMSE values on the Test Data



Plot for all test models

- **Rebuild the best model using the entire data - Make a forecast for the next 12 months**

After comparing all the models we constructed, it's evident that the triple exponential smoothing or Holt-Winters model yields the lowest RMSE. Therefore, it emerges as the most optimal choice. We will rebuild the best model using triple exponential smoothing for the next 12 months prediction.

```
DatetimeIndex(['1995-08-01', '1995-09-01', '1995-10-01', '1995-11-01',
               '1995-12-01', '1996-01-01', '1996-02-01', '1996-03-01',
               '1996-04-01', '1996-05-01', '1996-06-01', '1996-07-01'],
              dtype='datetime64[ns]', freq='MS')
```

Next 12 months dates that we will be using for prediction

Forecasts and confidence intervals into a DataFrame.

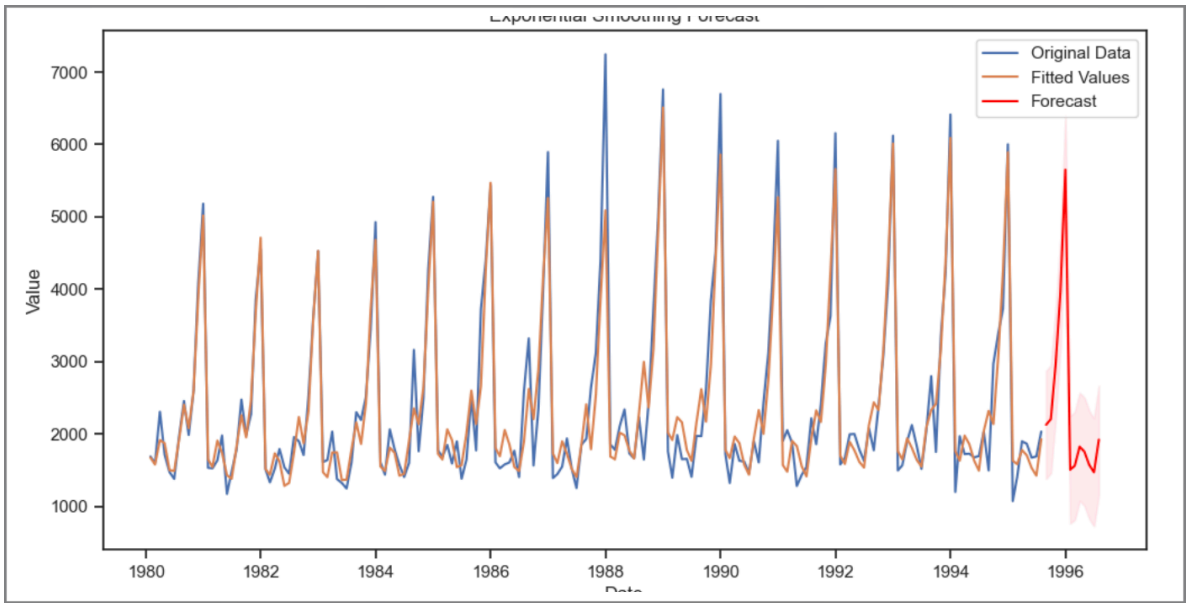
```
1995-08-31    2124.572947
1995-09-30    2203.468063
1995-10-31    2974.840500
1995-11-30    3962.065362
1995-12-31    5649.728868
1996-01-31    1504.266000
1996-02-29    1558.771788
1996-03-31    1821.925572
1996-04-30    1753.628758
1996-05-31    1570.069307
1996-06-30    1467.952120
1996-07-31    1918.096927
```

Sparkling Wine Sales predicted values made with Triple Exponential Smoothing

	Forecast	Lower Bound	Upper Bound
1995-08-31	2124.572947	1381.254048	2867.891847
1995-09-30	2203.468063	1460.149163	2946.786963
1995-10-31	2974.840500	2231.521600	3718.159400
1995-11-30	3962.065362	3218.746462	4705.384262
1995-12-31	5649.728868	4906.409968	6393.047768
1996-01-31	1504.266000	760.947100	2247.584900
1996-02-29	1558.771788	815.452888	2302.090688
1996-03-31	1821.925572	1078.606672	2565.244472
1996-04-30	1753.628758	1010.309858	2496.947658
1996-05-31	1570.069307	826.750407	2313.388207
1996-06-30	1467.952120	724.633220	2211.271020
1996-07-31	1918.096927	1174.778027	2661.415827

Forecasted value for next 12 months

Plot 29 - Prediction for future 12 months



Future predicted plot

7. Actionable Insights & Recommendations

Actionable Insights

- Despite a slight decline, Sparkling wine remains consistently popular among consumers, especially since its peak in the 1988.
- Sales of Sparkling wine experience notable seasonality, with a slower start in the first half of the year followed by a surge from August to December.
- The weak positive correlation suggests stable sales trends without significant upward or downward trends over the years.
- The very weak negative correlation between "Year" and "Month" confirms their independence, highlighting separate trends over years and months.

Recommendations

- Consider pairing Sparkling wine with a less popular option like Rose wine in promotional offers to encourage sales and diversify customer choices.

- Implement targeted marketing campaigns from April to June to stimulate sales during slower periods.
- Ensure product availability and quality during peak festival seasons to capitalize on high demand.
- Investigate the reasons behind the long-term decline in sales and adjust marketing strategies accordingly.
- Further analyze significant spikes and drops to gain insights into consumer behavior and preferences.