SPARKLING WINE SALES

BUSINESS REPORT

DIVYA BALASUBRAMANYAM

9/3/23

TIME SERIES ANALYSIS

Table of Contents

INTRODUCTION

The data of Sparkling wine sales in the 20th century is to be analysed. As an analyst in the ABC Estate Wines, we will analyse and forecast Wine Sales in the 20th century. The purpose of this whole exercise is to analyse the wine sales, do the exploratory data analysis build Models using the dataset to forecast Wine Sales for the next 12 months.

Problem1: Sparkling Wine Sales

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

Below is the sample of the dataset

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

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

Table 1: Top 5 and bottom 5 records of Sparkling Dataset

Shape of the Dataset:

There are 187 records with 2 columns in the dataset. The data consists of 187 monthly records of wine sales starting from January 1980 to July 1995 .

Information on the Dataset :

```
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   YearMonth  187 non-null    object
 1   Sparkling  187 non-null    int64 
dtypes: int64(1), object(1)
```

Table 2: Data Information of Sparkling Wine Dataset

There are 2 columns YearMonth and Sparkling in the dataset . YearMonth is an object type and Sparkling is integer type column. There are **no null values** in the dataset .

'Sparkling' column represents the monthly sales of the sparkling wine across all the years 1980 to July 1995

Since we are doing time series analysis , we will convert the column YearMonth to datetime column and make it an DatetimeIndex

```
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
```

Table 3 Data Information after transformation of YearMonth Column

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

Table 4: Top 5 records of Sparkling Dataset after transformation of YearMonth Column

Plotting the Time series data of Sparkling Wine sales

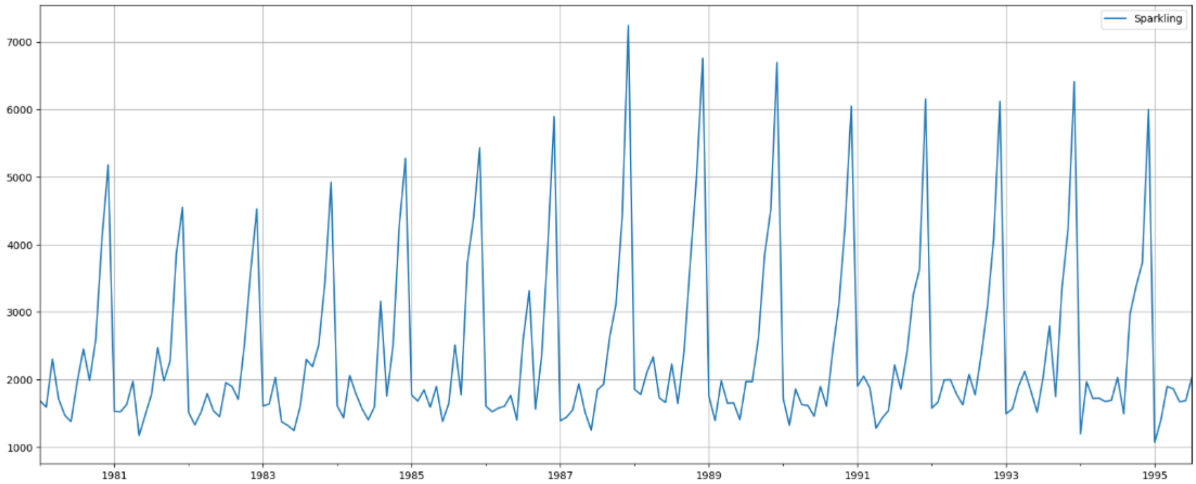


Figure 1 Time series plot of Sparkling Wine Data

As we can see from the time series plot above , there is no trend in the sales but there seems to be a seasonality in the sales data

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

Univariate Analysis

As there is only one numerical variable 'Sparkling' , let's see data description and boxplot of this variable

count	mean	std	min	25%	50%	75%	max
187.000000	2402.4171	1295.1115	1070	1605	1874	2549	7242

Table 5: Data description of the dataset

Findings from the Data description

Mean of the sales is 2402.42 and Median is 1874.

Standard Deviation of the Sales is 1295.11

Minimum sales recorded for a month is 1070.

Maximum sales recorded for a month is 7242.

25% of the sales is below 1605

50% of the sales is below 1874

25% of the sales is below 2549

Boxplot of the 'Sparkling' variable

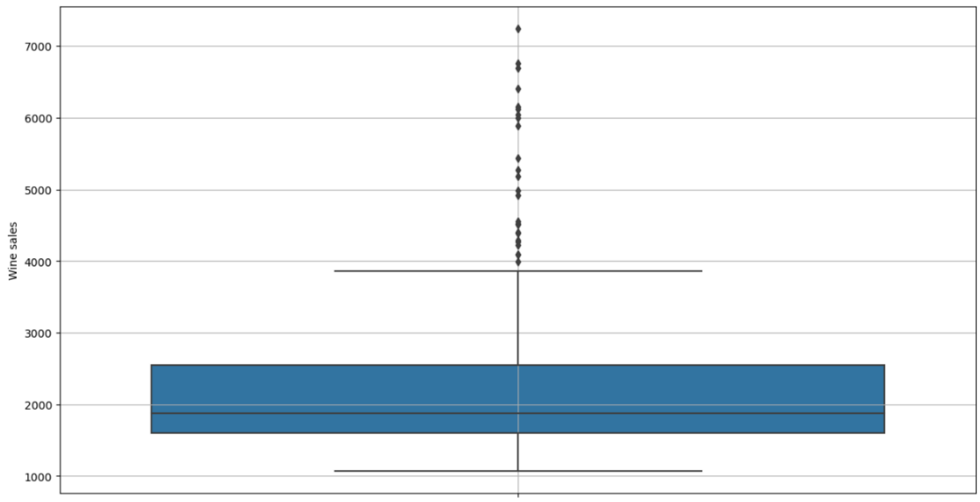


Figure 2 Boxplot of the 'Sparkling' Variable

It's a right skewed distribution . There are outliers in the sales data.

Yearly Sales Plot

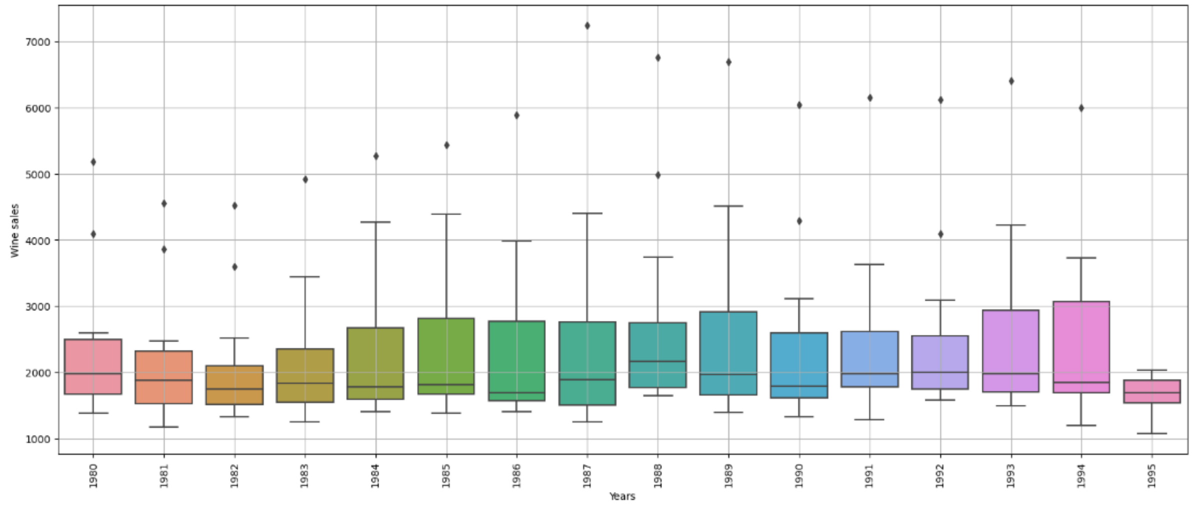


Figure 3: Yearly plot of Sales

As we can see from the yearly plot , year 1988 had the highest average sales of 2770.50 and year 1995 had the lowest average sales of 1660.

Monthly sales plots across all years

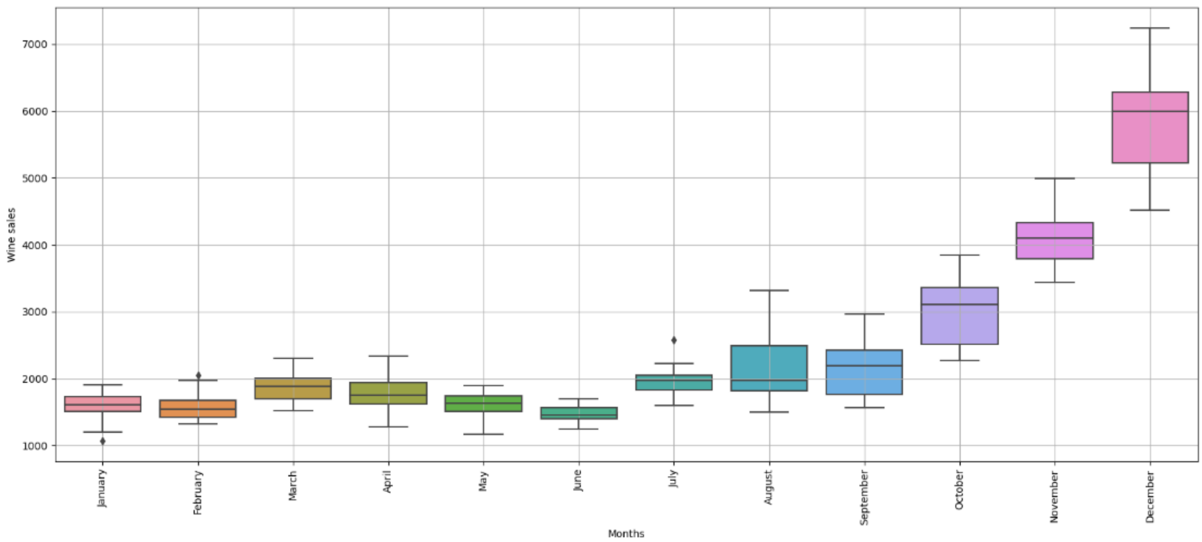


Figure 4: Monthly plot of Sales

As we can see from the Monthly sales plot , December month had the highest average sales year on year and June had the lowest average sales.

Month plot of sales

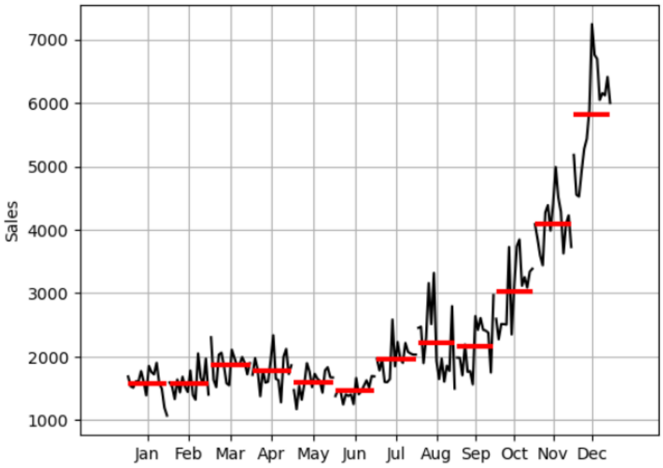


Figure 5:Month Plot of sales

This plot shows us the behaviour of the Time Series ('Wine Sales' in this case) across various months. The red line is the median value. As observed earlier , December month had the highest sales year on year and June had the lowest sales.

Bivariate Analysis

Let's see a how the yearly sales have been across all the months

Time_Stamp	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Time_Stamp																
January	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
February	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
March	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
April	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
June	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
July	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
August	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
September	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
October	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
November	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
December	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

Table 6: Tabular column of yearly sales across all months

As we can see from the tabular column above December month clocks highest sales across all years and June the lowest . Also, we should note that that sales data is unavailable for months of August, September, October , November, December for the year 1995

Sales vs month plot across the years

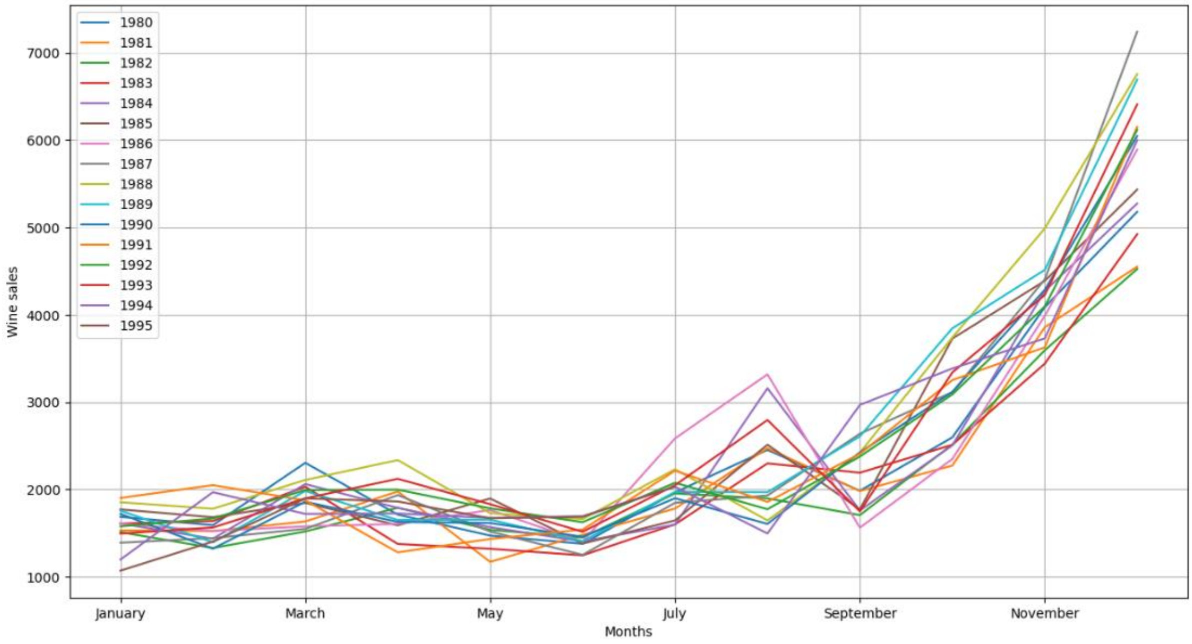


Figure 6: Months Vs Sales across all years

After plotting the Sales vs month across the years, we can note that December had the maximum sales across all years .

Sales vs Years Plot across the months

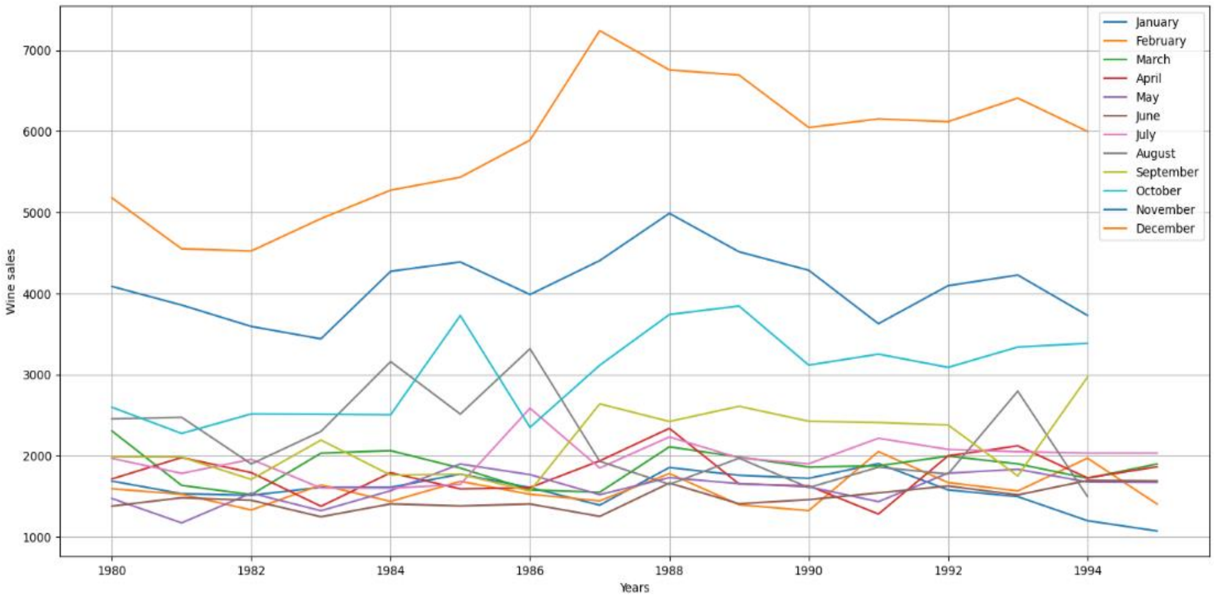


Figure 7: Years Vs Sales across all months

After plotting the Sales vs Years across the months, we can note that 1988 had the maximum sales across all years whereas 1995 had the minimum sales considering the fact that data is unavailable for months of August, September, October , November, December for the year 1995.

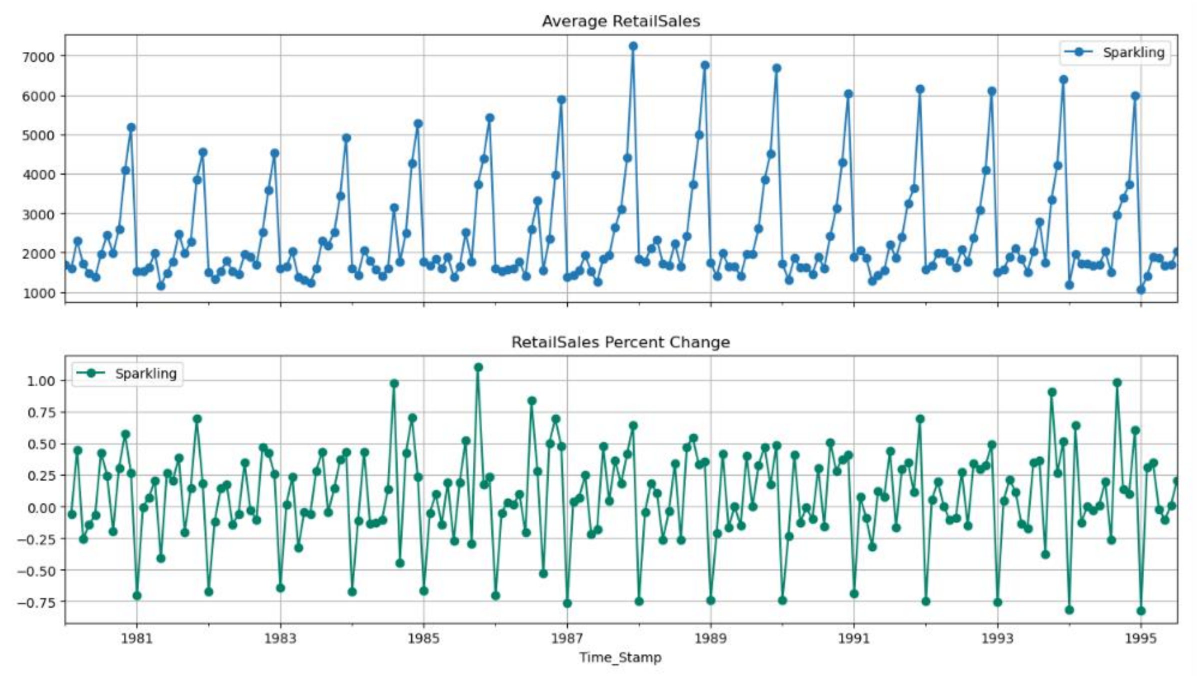


Figure 8:Average sales across years and percentage change is sales .

From the Average sales across years and percentage change sales plot , we can see that the mean doesn't seem to be changing much across the years and variance is also not changing

Decomposition of the Time series data

- We decompose the time series to understand revenue generation without the quarterly effects
- De-seasonalize the series to estimate and adjust by seasonality
- Compare the long-term movement of the series (Trend) vis-a-vis short-term movement (seasonality) to understand which has the higher influence

Decomposition Model can be Additive or Multiplicative

- Additive model: Observation = Trend + Seasonality + Error

  Yt= Tt+ St+ It

- Multiplicative model: Observation = Trend * Seasonality * Error

  Yt= Tt* St* It

Yt: time series value (actual data) at period t.

St: seasonal component (index) at period t.

Tt: trend cycle component at period t.

It: irregular (remainder) component at period

Let's decompose the data and check the trend, seasonality and the irregular/residual/error component.

Additive Decomposition

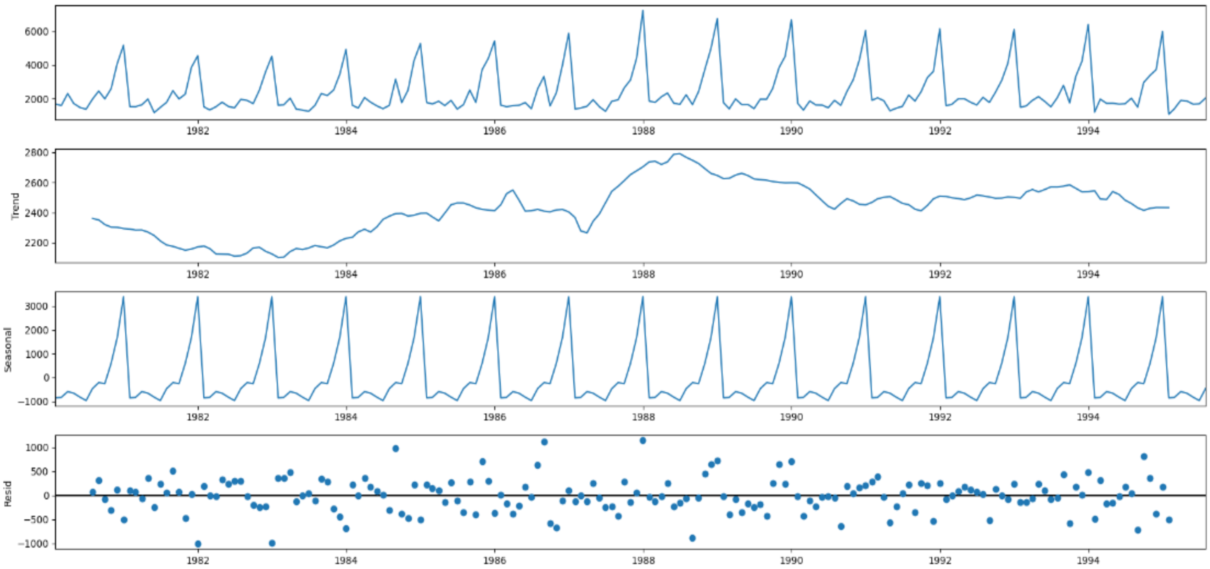


Figure 9: Additive decomposition of Time Series Data

The plot above shows breakup of time series into Trend, Seasonality and Residual components using Addtitve decomposition model.Time series value at period t can be obtained by adding the Trend, Seasonality and Residual Data . As we can see from the plot there is no visible trend in the data , but there is a seasonality component . The residual component doesn't seem to have a pattern .

The Sales data is broken down into all Trend, Seasonality and Residual components using Addtitve model as shown in table below:

Trend Time_Stamp	
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	2360.67
1980-08-31	2351.33
1980-09-30	2320.54
1980-10-31	2303.58

Seasonality Time_Stamp	
1980-01-31	0.65
1980-02-29	0.66
1980-03-31	0.76
1980-04-30	0.73
1980-05-31	0.66
1980-06-30	0.60
1980-07-31	0.81
1980-08-31	0.92
1980-09-30	0.89
1980-10-31	1.24

Residual Time_Stamp	
1980-01-31	NaN
1980-02-29	NaN
1980-03-31	NaN
1980-04-30	NaN
1980-05-31	NaN
1980-06-30	NaN
1980-07-31	1.03
1980-08-31	1.14
1980-09-30	0.96
1980-10-31	0.91

Multiplicative Decomposition

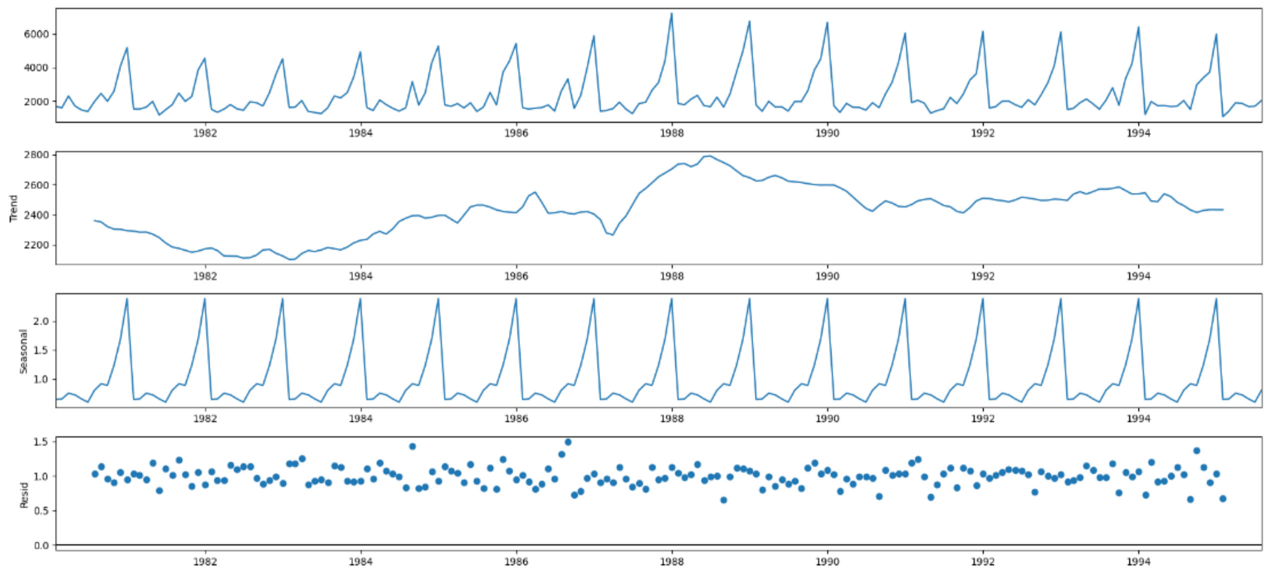


Figure 10: Multiplicative decomposition of Time Series Data

The plot above shows breakup of time series into Trend, Seasonality and Residual components using Multiplicative decomposition model .Time series value at period t can be obtained by multiplying the Trend, Seasonality and Residual Data . As we can see from the plot there is no visible trend in the data , but there is a seasonality component . The residual component doesn't seem to have a pattern

The Sales data is broken down into all Trend, Seasonality and Residual components using Multiplicative model as shown in table below:

Trend Time_Stamp		Seasonality Time_Stamp		Residual Time_Stamp	
1980-01-31	NaN	1980-01-31	0.65	1980-01-31	NaN
1980-02-29	NaN	1980-02-29	0.66	1980-02-29	NaN
1980-03-31	NaN	1980-03-31	0.76	1980-03-31	NaN
1980-04-30	NaN	1980-04-30	0.73	1980-04-30	NaN
1980-05-31	NaN	1980-05-31	0.66	1980-05-31	NaN
1980-06-30	NaN	1980-06-30	0.60	1980-06-30	NaN
1980-07-31	2360.67	1980-07-31	0.81	1980-07-31	1.03
1980-08-31	2351.33	1980-08-31	0.92	1980-08-31	1.14
1980-09-30	2320.54	1980-09-30	0.89	1980-09-30	0.96
1980-10-31	2303.58	1980-10-31	1.24	1980-10-31	0.91

Time series value at period t can be obtained by multiplying the Trend, Seasonality and Residual Data .

Multiplicative Model seems to be giving a better prediction of Time series value at period t.

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

The dataset is split into training and testing data with testing data starting from Year 1991.

After the train and test split of 187 records, there are 132 records in the training dataset and

55 records in the testing dataset .

Training Data is used to train (develop) the model. Training Data is used to identify a few working models. The forecasts for training data are called fitted values. Each of the models is tested against the observed values of the series for hold-out period.

The model is selected to be the best were observed and forecasted values are the closest.

Predictive power of a model is estimated by comparing its forecasting performance on a Test Data

Let's see a sample of Training and Testing Dataset

Training dataset is ending at 1990 December

```
First few rows of Training Data
             Sparkling
Time_Stamp
1980-01-31        1686
1980-02-29        1591
1980-03-31        2304
1980-04-30        1712
1980-05-31        1471

Last few rows of Training Data
             Sparkling
Time_Stamp
1990-08-31        1605
1990-09-30        2424
1990-10-31        3116
1990-11-30        4286
1990-12-31        6047
```

Testing dataset is starting at 1991 January

```
First few rows of Test Data
             Sparkling
Time_Stamp
1991-01-31        1902
1991-02-28        2049
1991-03-31        1874
1991-04-30        1279
1991-05-31        1432

Last few rows of Test Data
             Sparkling
Time_Stamp
1995-03-31        1897
1995-04-30        1862
1995-05-31        1670
1995-06-30        1688
1995-07-31        2031
```

Plot of the training and Testing Dataset

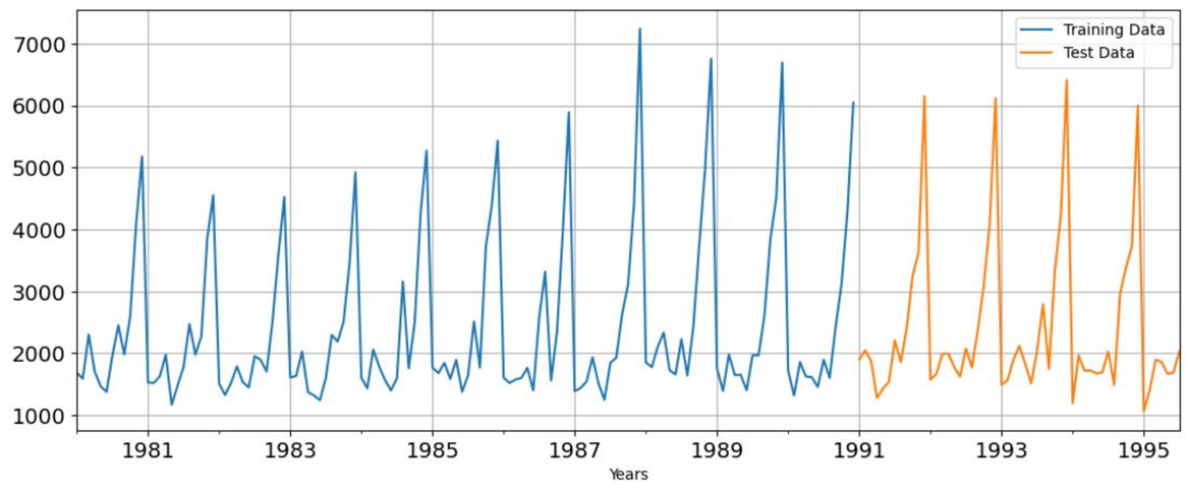


Figure 11: Time series Data split into Train and Test Data

As we can see from the plot above , the training data is marked in blue and testing data is marked in orange starts from 1991 and goes on till the end of the timeseries dataset .

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

Exponential Smoothing Models

Exponential Smoothing Models take weighted averages of past observations **,w**eights decay as observations get older**.**One or more parameters control how fast the weights decay **.**These parameters have values between 0 and 1.

Simple Exponential Smoothing (SES)

SES model is used If the time series neither has a pronounced trend nor seasonality:

Performance of the smoothing parameter α controls performance of the method. If α is closer to 1, forecasts follow the actual observations more closely. If α is closer to 0, forecasts are farther from the atual observations and the line is smooth .

The SES model gives the parameters α as 0.0496

The smoothing level α is 0.0496 which is close to 0 , hence forecasts are farther from the actual observations and the line is smooth .

The predictions are all the values are same 2724.932

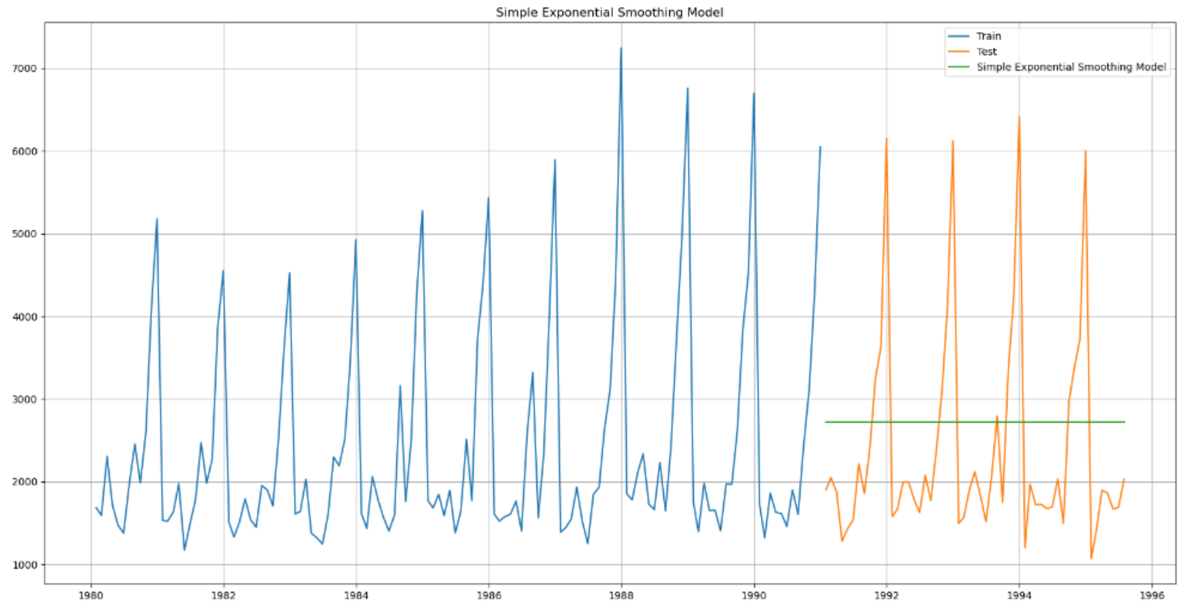


Figure 12: Time series plot with Testing and Training Data. The green line is the SES prediction data values

As we can see from the plot above , the SES Model represented by the green line is not good at predicting the test data values . It's a smooth line with a constant value .

Let's evaluate the model using RMSE

Test RMSE score for SES Model with Alpha 0.0496 which is 1316.008384

Simple Exponential Smoothing with Iteration (SES)

Using Iterative Method to find the best values for smoothing parameter α and we get following values for Test RMSE for different α values

	Alpha Values	Test RMSE
1	0.02	1279.495201
0	0.01	1281.032699
2	0.03	1293.110073
3	0.04	1305.462953
4	0.05	1316.411742
...	...	...
94	0.95	3778.432623
95	0.96	3796.048620
96	0.97	3813.437370
97	0.98	3830.602869
98	0.99	3847.548965

Table 7: SES Parameters after iteration

With best α value 0.02, so we can build SES model

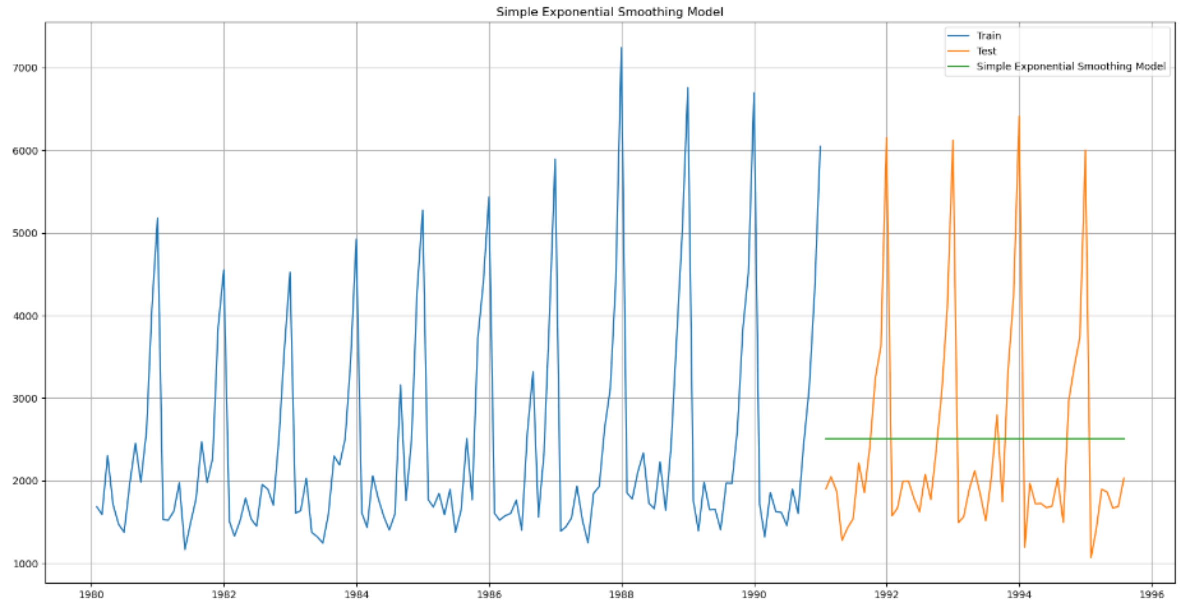


Figure 13: Time series plot with Testing and Training Data. The green line is the SES prediction data values

The predictions are all the values are same 2505.42

Let's evaluate the model using RMSE

Test RMSE score for SES Model with Alpha 0.02 which is 1279.495

Double Exponential Smoothing (DES)

- DES is applicable when data has Trend but no seasonality .Its an extension of SES
- Two separate components are considered: Level and Trend
- Level is the local mean
- One smoothing parameter α corresponds to the level series
- A second smoothing parameter β corresponds to the trend series
- Also known as Holt mode

The DES model gives the following parameters :

:

	name	param	optimized
smoothing_level	alpha	0.688571	True
smoothing_trend	beta	0.000100	True
initial_level	l.0	1686.000000	True
initial_trend	b.0	-95.000000	True

Table 8: DES Parameters

The smoothing level α is 0.688 which is significant but β is 0.0001 which is almost 0 , hence we can say that there is no trend factor in the series .

Following table shows the predictions on testing dataset

1991-01-31	5221.278699
1991-02-28	5127.886554
1991-03-31	5034.494409
1991-04-30	4941.102264
1991-05-31	4847.710119
1991-06-30	4754.317974
1991-07-31	4660.925829
1991-08-31	4567.533684
1991-09-30	4474.141539
1991-10-31	4380.749394

Table 9: Test Data predictions using DES Model

As we can see from the above table , the predicted value is not the same for all the data points as in SES model and the trend component is almost 0

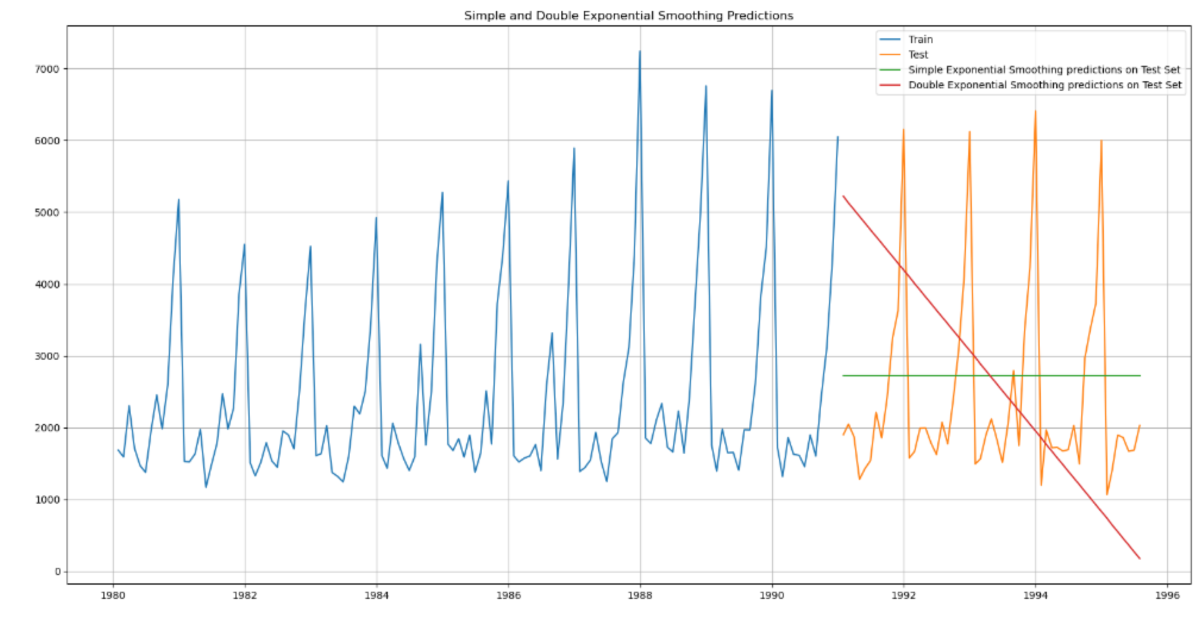


Figure 14: Time series plot with Training, Testing, SES and DES Model

As we can see from the plot above , the DES Model represented by the red line is also not good at predicting the test data values .

Let's evaluate the model using RMSE

The Test RMSE score for DES model is 2007.23

	Test RMSE
Alpha=0.0496:SimpleExponentialSmoothing	1316.035487
Alpha=0.688,Beta=0.0001:DoubleExponentialSmoothing	2007.238526

Table 10:RMSE score for SES and DES Models

Double Exponential Smoothing with Iteration (DES)

Using Iterative Method to find the best values for smoothing parameter α and we get following values for Test RMSE for different α and β values

	Alpha Values	Beta Values	Test RMSE
148	0.02	0.50	1274.630824
115	0.02	0.17	1275.105310
254	0.03	0.57	1276.025836
255	0.03	0.58	1278.425944
253	0.03	0.56	1278.585750
...	...	...	...
2175	0.22	0.97	60335.137153
2077	0.21	0.98	60589.909084
2176	0.22	0.98	60740.944412
2177	0.22	0.99	61104.414936
2078	0.21	0.99	61161.469936

Table 11: DES Parameters after iteration

With α value 0.02 and β values 0.5, we can build DES model

The predictions are all the values are

```
1991-01-31    2370.481106
1991-02-28    2371.765873
1991-03-31    2373.050640
1991-04-30    2374.335407
1991-05-31    2375.620173
1991-06-30    2376.904940
1991-07-31    2378.189707
1991-08-31    2379.474474
1991-09-30    2380.759241
1991-10-31    2382.044007
```

Let's evaluate the model using RMSE

Test RMSE score for SES Model with Alpha 0.02 which is 1274.63

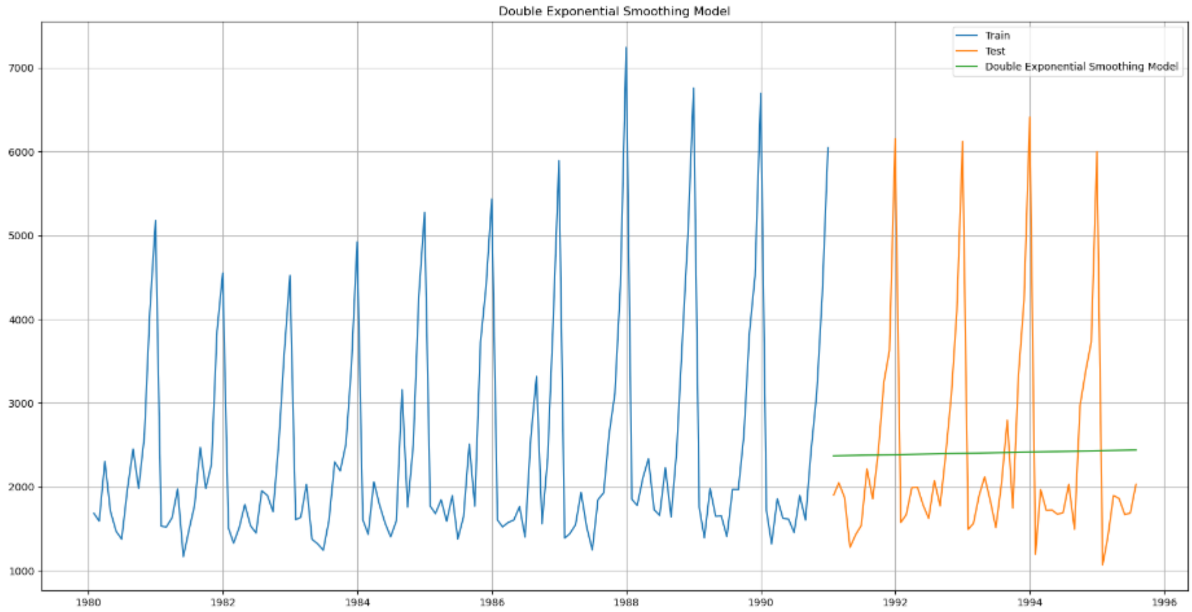


Figure 15: Time series plot with Testing and Training Data and DES prediction data values

Triple Exponential Smoothing (TES) or Holt-Winters' Model

- TES is applicable when data has Level, Trend and Seasonality, Its an extension of DES
- Three separate components are considered: Level , Trend and Seasonality
- Because Seasonality can be additive or multiplicative, TES model can be additive or multiplicative
- Simultaneously smooths the level, trend and seasonality

Three separate smoothing parameters

α: Smooths level; $0 < \alpha < 1$

β: Smooths trend; $0 < \beta < 1$

γ : Smooths seasonality; $0 < \gamma < 1$

TES Additive Model

The TES Additive model gives the following parameters

	name	param	optimized
smoothing_level	alpha	0.111272	True
smoothing_trend	beta	0.012361	True
smoothing_seasonal	gamma	0.460718	True
initial_level	l.0	2356.577981	True
initial_trend	b.0	-0.102437	True
initial_seasons.0	s.0	-636.233193	True
initial_seasons.1	s.1	-722.983201	True
initial_seasons.2	s.2	-398.644108	True
initial_seasons.3	s.3	-473.430454	True
initial_seasons.4	s.4	-808.424733	True
initial_seasons.5	s.5	-815.349914	True
initial_seasons.6	s.6	-384.230650	True
initial_seasons.7	s.7	72.994844	True
initial_seasons.8	s.8	-237.442260	True
initial_seasons.9	s.9	272.326083	True
initial_seasons.10	s.10	1541.377371	True
initial_seasons.11	s.11	2590.076923	True

Table 12:TES Parameters

The smoothing level α is 0.111272 , Smoothing trend β is 0.012361 and smoothing seasonal γ is 0.460718.

From the above values we can say that since γ component is having a significant value, there is a seasonal component .

Following table shows the predictions on testing dataset

1991-01-31	1490.402890
1991-02-28	1204.525152
1991-03-31	1688.734182
1991-04-30	1551.226125
1991-05-31	1461.197883
1991-06-30	1278.646707
1991-07-31	1804.885616
1991-08-31	1678.955032
1991-09-30	2315.373126
1991-10-31	3224.976222

Table 13: Test Data predictions using TES Additive Model

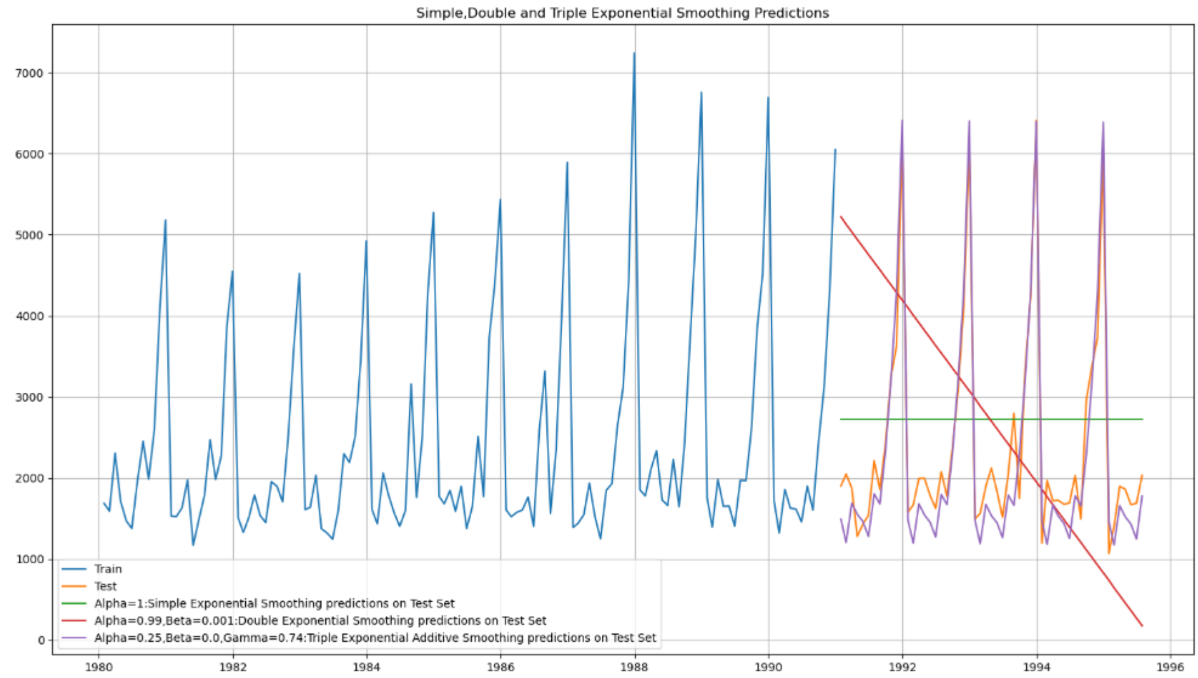


Figure 16: Time series plot with Training, Testing, SES,DES and TES Additive Model

As we can see from the plot above , the TES Additive Model represented by the purple line is good at predicting the test data values as it is following the test data variations .

Let's evaluate the model using RMSE

The Test RMSE score for TES Additive model is 378.951

TES Multiplicative Model

The TES Multiplicative model gives the following parameters

	name	param	optimized
smoothing_level	alpha	0.111338	True
smoothing_trend	beta	0.049505	True
smoothing_seasonal	gamma	0.362080	True
initial_level	l.0	2356.496789	True
initial_trend	b.0	-10.187945	True
initial_seasons.0	s.0	0.712964	True
initial_seasons.1	s.1	0.682422	True
initial_seasons.2	s.2	0.907550	True
initial_seasons.3	s.3	0.805152	True
initial_seasons.4	s.4	0.655972	True
initial_seasons.5	s.5	0.654145	True
initial_seasons.6	s.6	0.886179	True
initial_seasons.7	s.7	1.133451	True
initial_seasons.8	s.8	0.920463	True
initial_seasons.9	s.9	1.213379	True
initial_seasons.10	s.10	1.873403	True
initial_seasons.11	s.11	2.378118	True

Table 14: TES multiplicative model Parameters

The smoothing level α is 0.111338 , Smoothing trend β is 0.049505 and smoothing seasonal γ is 0.362080.

From the above values we can say that since γ component is having a significant value, there is a seasonal component .

Following table shows the predictions on testing dataset

1991-01-31	1587.497468
1991-02-28	1356.394925
1991-03-31	1762.929755
1991-04-30	1656.165933
1991-05-31	1542.002730
1991-06-30	1355.102435
1991-07-31	1854.197719
1991-08-31	1820.513188
1991-09-30	2276.971718
1991-10-31	3122.024202

Table 15: Test Data predictions using TES Multiplicative Model

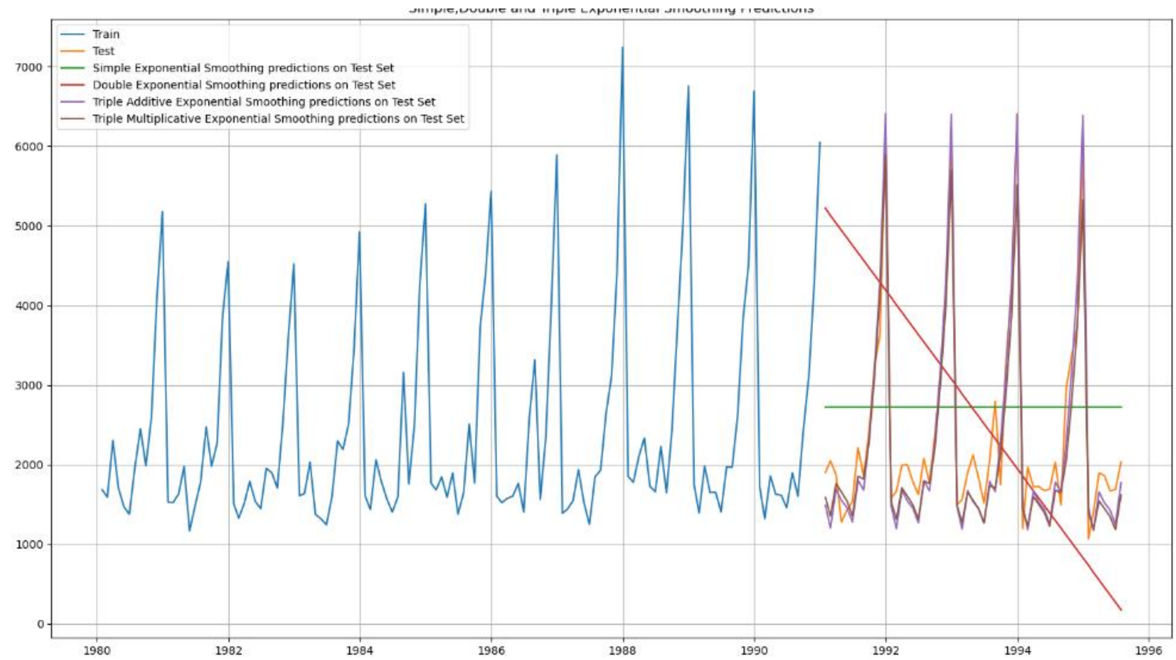


Figure 17: Time series plot with Training, Testing, SES,DES and TES Additive and Multiplicative Model

As we can see from the plot above , the TES Multiplicative Model represented by the brown line is good at predicting the test data values as it is following the test data variations .

Let's evaluate the model using RMSE

The Test RMSE score for TES Multiplicative model is 404.286

	Test RMSE
Alpha=0.0496:SimpleExponentialSmoothing	1316.035487
Alpha=0.688,Beta=0.0001:DoubleExponentialSmoothing	2007.238526
Alpha=0.111338,Beta=0.049505,Gamma=0.362080:TripleExponentialSmoothingMultiplicative	404.286809
Alpha=0.111272,Beta=0.012361,Gamma=0.460718:TripleExponentialSmoothingAdditive	378.951023
Alpha=0.02:SimpleExponentialSmoothing	1279.495201
Alpha=0.02,Beta=0.50,IterativeDoubleExponentialSmoothing	1274.630824

Table 16: Test RMSE for various Exponential smoothing Models

Linear Regression Model(LR)

For this particular linear regression, we are going to regress the 'Sparkling' variable against the order of the occurrence. For this we need to modify our training data by adding a new variable 'time' before fitting it into a linear regression.

```
Training Time instance
 [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65,
66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97,
98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
124, 125, 126, 127, 128, 129, 130, 131, 132]
Test Time instance
 [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 18
3, 184, 185, 186, 187]
```

We generate the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

Let's see the first and last few rows of training and test data

```
First few rows of Training Data
```

Time_Stamp	Sparkling	time
1980-01-31	1686	1
1980-02-29	1591	2
1980-03-31	2304	3
1980-04-30	1712	4
1980-05-31	1471	5

```
Last few rows of Training Data
```

Time_Stamp	Sparkling	time
1990-08-31	1605	128
1990-09-30	2424	129
1990-10-31	3116	130
1990-11-30	4286	131
1990-12-31	6047	132

Table 17: Sample of Training Data for LR model

First few rows of Test Data

Time_Stamp	Sparkling	time
1991-01-31	1902	133
1991-02-28	2049	134
1991-03-31	1874	135
1991-04-30	1279	136
1991-05-31	1432	137

Last few rows of Test Data

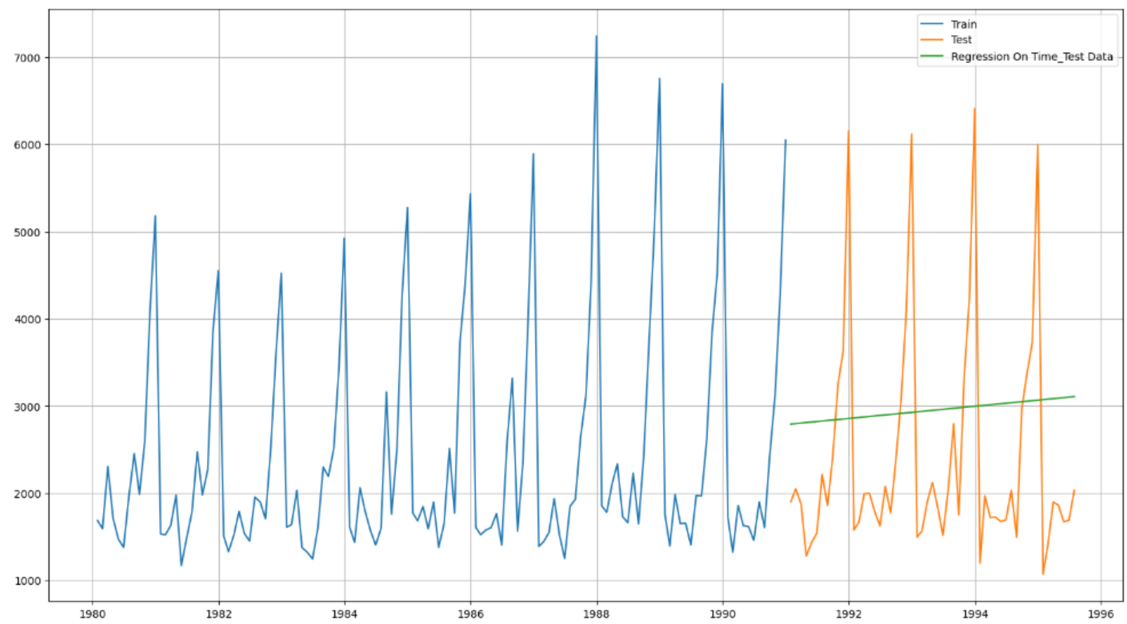
Time_Stamp	Sparkling	time
1995-03-31	1897	183
1995-04-30	1862	184
1995-05-31	1670	185
1995-06-30	1688	186
1995-07-31	2031	187

Table 18: Sample of Testing Data for LR model

Now that our training and test data has been modified, let us go ahead use Linear Regression to build the model on the training data and test the model on the test data and plot the time series data with predictions using LR Model alongside Training and Testing Data . As we can see from the plot below , the Linear regression Model represented by the green line is not good at predicting the test data values as it is not following the test data variations

Let's evaluate the model using RMSE

The Test RMSE score for Linear Regression model is 1389.135



.

Naïve Model : y^t+1=yt+1

For this particular Naïve model, we say that the prediction for next month is the same as current month and the prediction for month after next is same as prediction for next month and since the prediction of next month is same as current month, therefore the prediction for month after next is also same current month.

Since Naïve model prediction for next month is the same current month, the prediction for first record in test data is same as last record of training data , so let's see the last record of training data and the predictions on test data

Time_Stamp	Sparkling
1990-08-31	1605
1990-09-30	2424
1990-10-31	3116
1990-11-30	4286
1990-12-31	6047

Table 19: Last 5 records of training data

Using Naïve Approach to build the model on the training data and test the model on the test data we get the following predictions. As we can see last record of training data has value 6047, therefore the test data will have the same value for all records

Time_Stamp	
1991-01-31	6047
1991-02-28	6047
1991-03-31	6047
1991-04-30	6047
1991-05-31	6047

Table 20: First 5 records of predicted test data

Let's plot the time series data with predictions using Naïve Model alongside Training and Testing Data

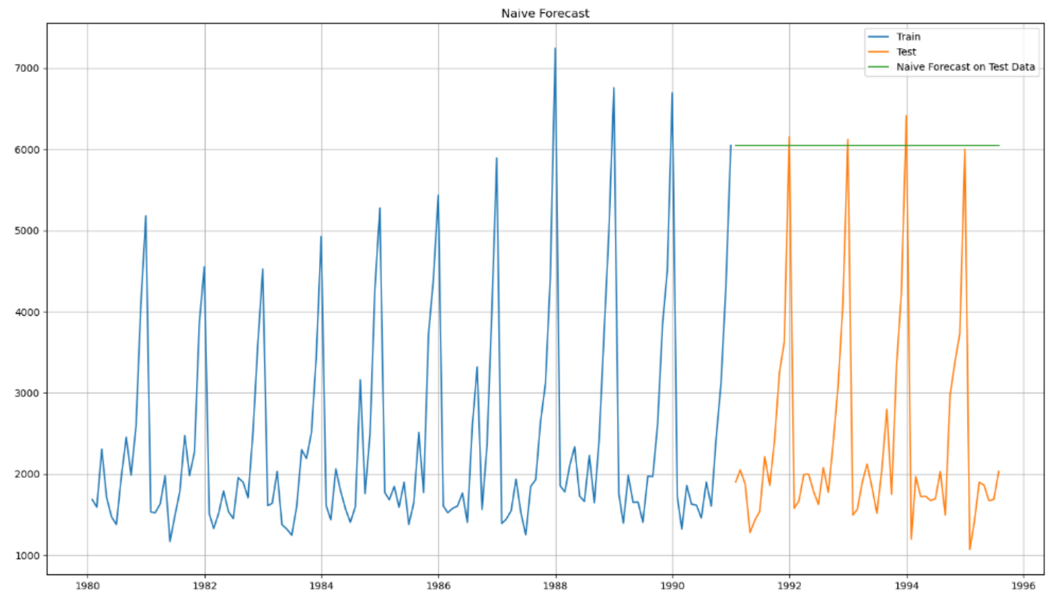


Figure 18: Time series plot with Training, Testing and Naïve Model

As we can see from the plot above , the Linear regression Model represented by the green line is not good at predicting the test data values as it is not following the test data variations .

Let's evaluate the model using RMSE

The Test RMSE score for Linear Regression model is 3864.279

Simple Average Model

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

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

Table 21:Mean Forecast for simple average model against the actual values of test data

Let's plot the time series data with predictions using Simple Average Model alongside Training and Testing Data

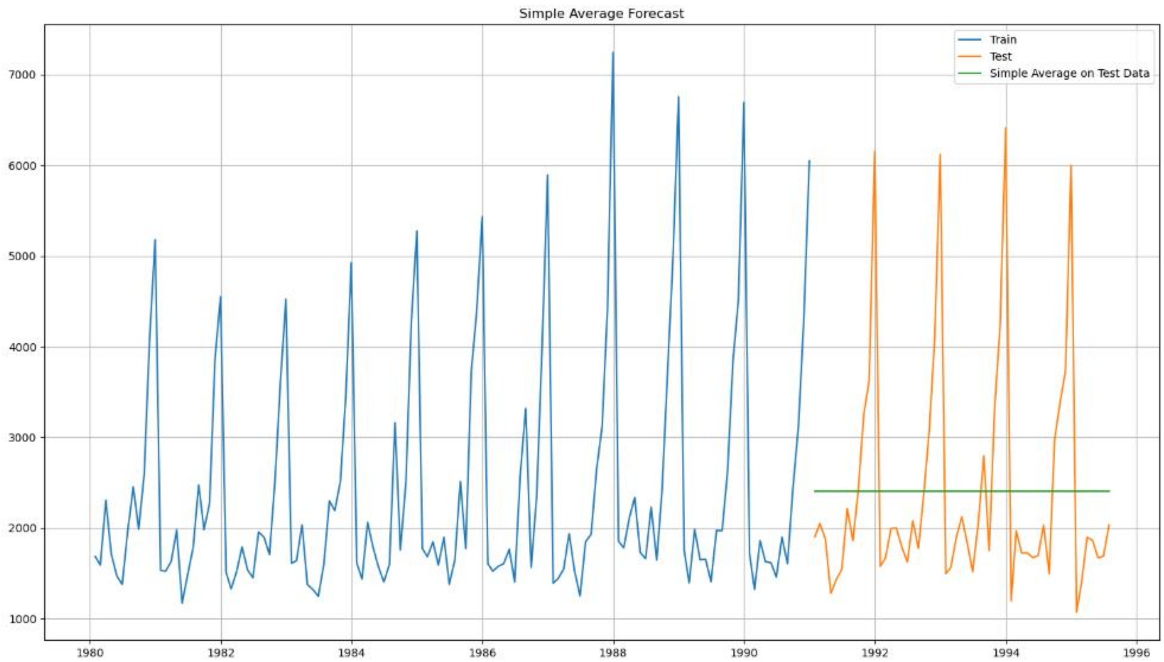


Figure 19: Time series plot with Training, Testing and Simple Average Model

As we can see from the plot above , the Simple Average Model represented by the green line is not good at predicting the test data values as it is not following the test data variations .

Let's evaluate the model using RMSE

The Test RMSE score for Linear Regression model is 1275.081

Moving Average Model(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.

Let's compute moving averages with intervals 2,4,6,9 on the entire dataset:

First 10 records of the dataset with moving average at different intervals

Time_Stamp	Sparkling	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	1686	NaN	NaN	NaN	NaN
1980-02-29	1591	1638.5	NaN	NaN	NaN
1980-03-31	2304	1947.5	NaN	NaN	NaN
1980-04-30	1712	2008.0	1823.25	NaN	NaN
1980-05-31	1471	1591.5	1769.50	NaN	NaN
1980-06-30	1377	1424.0	1716.00	1690.166667	NaN
1980-07-31	1966	1671.5	1631.50	1736.833333	NaN
1980-08-31	2453	2209.5	1816.75	1880.500000	NaN
1980-09-30	1984	2218.5	1945.00	1827.166667	1838.222222
1980-10-31	2596	2290.0	2249.75	1974.500000	1939.333333

Figure 20: 2,4,6,9 point Moving Average

Plotting the Moving average vis-a-vis training data

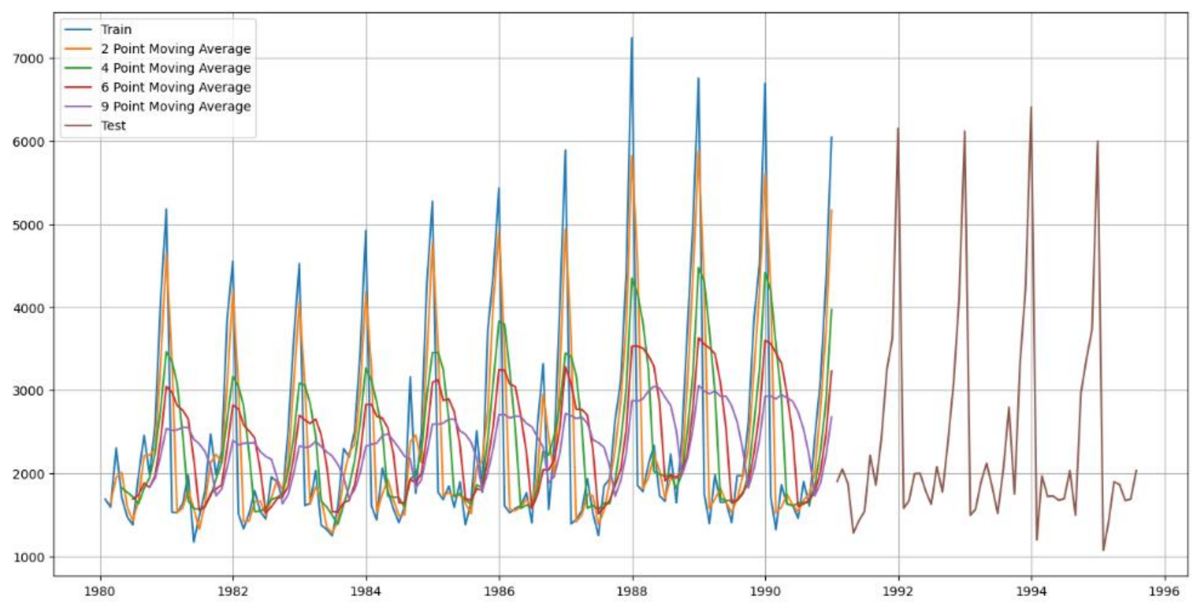


Figure 21: Time series plot with Training Dataset and Moving Average Model with different intervals

Let's apply the each of the moving average model on the Testing dataset

Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.

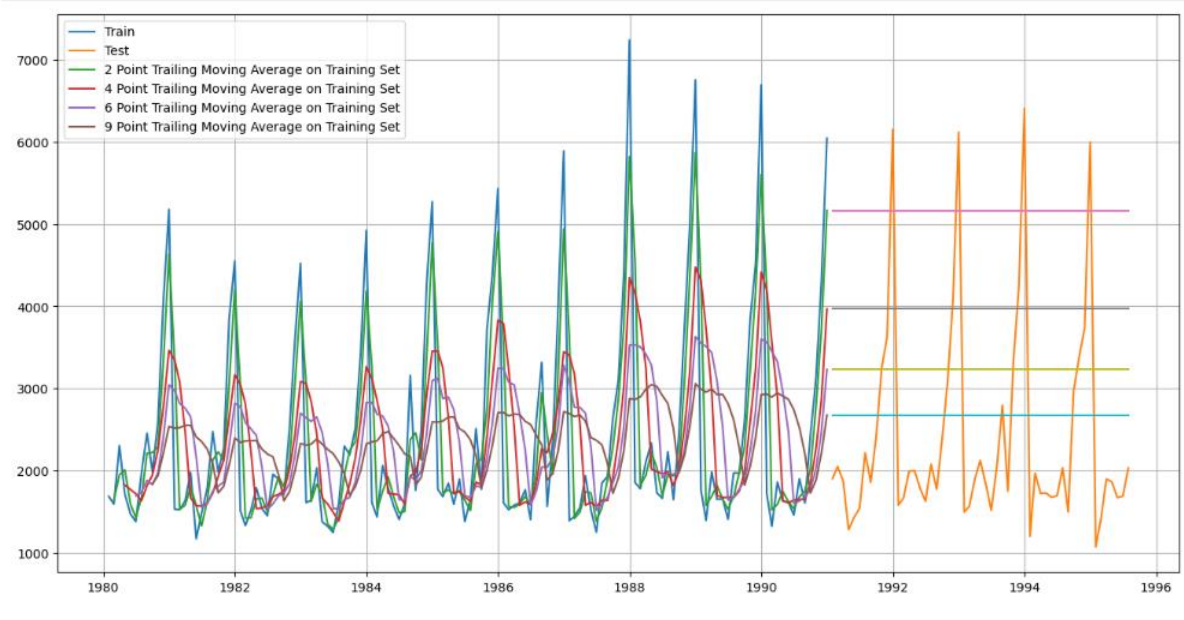


Figure 22: Time series plot with Training and Testing Dataset and Moving Average Model with different intervals

As we can see from the plot above , the Moving Average Model for different intervals is plotted vis- a-vis testing dataset . 2 point MA model is represented by green line, 4 point MA model is represented by red line, 6 point MA model is represented by purple line and 9 point MA model is represented by brown line. 2 point MA model seems to be performing best on the testing dataset

Let's evaluate the model using RMSE for all the MA models

RMSE for 2 point Moving Average Model forecast on Testing Data is 3046.976

RMSE for 4 point Moving Average Model forecast on Testing Data is 2021.856

RMSE for 6 point Moving Average Model forecast on Testing Data is 1521.611

RMSE for 9 point Moving Average Model forecast on Testing Data is 1304.618

Among Moving Average models RMSE score of 4 point Moving Average Model has the leas RMSE score , hence that is the best model among all other MA models

RMSE score of all the models

	Test RMSE
Alpha=0.0496:SimpleExponentialSmoothing	1316.035487
Alpha=0.688,Beta=0.0001:DoubleExponentialSmoothing	2007.238526
Alpha=0.111338,Beta=0.049505,Gamma=0.362080:TripleExponentialSmoothingMultiplicative	404.286809
Alpha=0.111272,Beta=0.012361,Gamma=0.460718:TripleExponentialSmoothingAdditive	378.951023
Alpha=0.02:SimpleExponentialSmoothing	1279.495201
Alpha=0.02,Beta=0.50,IterativeDoubleExponentialSmoothing	1274.630824
RegressionOnTime	1389.135175
NaiveModel	3864.279352
Simple Average	1275.081804
2pointTrailingMovingAverage	3046.976092
4pointTrailingMovingAverage	2021.855880
6pointTrailingMovingAverage	1521.611250
9pointTrailingMovingAverage	1304.618442

Figure 23 RMSE of ALL Models

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.

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

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

Stationarity means that the autocorrelation of lag 'k' depends on k, but not on time t.

Let Xt denote the time series at time t.

Autocorrelation of lag k is the correlation between Xt and $X(t-k)$

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 H_a: 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 .

Dicky Fuller Test on the entire dataset to check stationarity

DF test statistic is -1.798

DF test p-value is 0.7056

- As the p value is larger than 0.05, we fail to reject the null hypotheses that Time Series is non-stationary.
- The dataset is non-stationary at 95% confidence level. Differencing 'd' to make time series stationary

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

(d=1) 1st order differencing is done where the difference between the current and previous (1 lag before) series is taken and then checked for stationarity using the ADF(Augmented Dicky Fueller) test. If differenced time series is stationary, we proceed with AR modelling.

Else we do (d=2) 2nd order differencing, and this process repeats till we get a stationary time series

1st order differencing equation is : y_t=y_t-y_{t-1}

2nd order differencing equation is : y_t=(y_t- y_{t-1}) − (y_{t-1}- y_{t-2}) and so on…

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

Let's apply differencing of order 1 on the dataset and check for stationarity

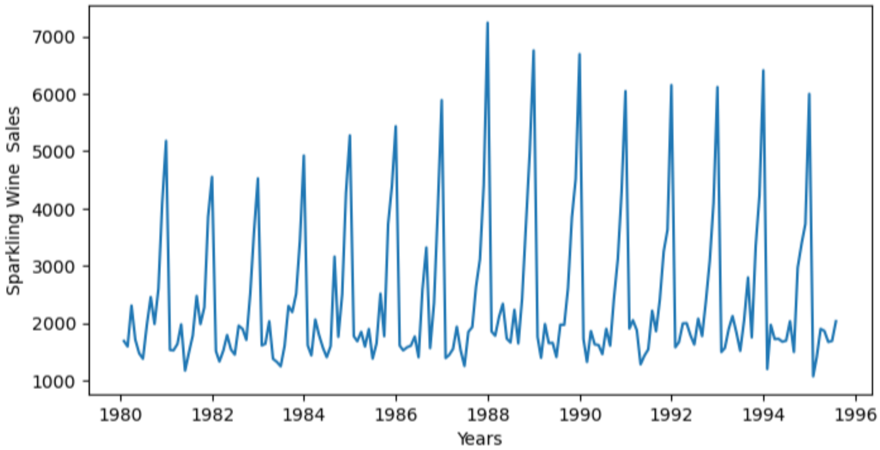


Figure 24:Original Time series

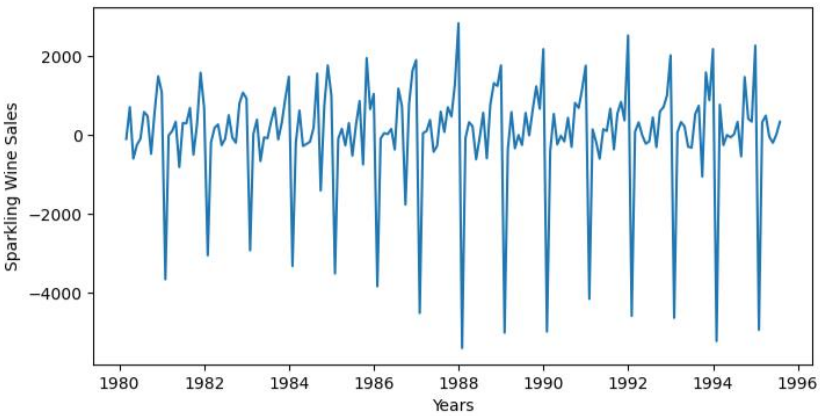


Figure 25:Time series after differencing d=1

Dicky Fuller Test on the differenced dataset to check stationarity

DF test statistic is -44.912

DF test p-value is 0.0

Observations:

• As the p value is less than 0.05, we reject the null hypotheses that Time Series is non-stationary.

• The Training data is stationary at 95% confidence level.

Dicky Fuller test on the Training dataset to check stationarity

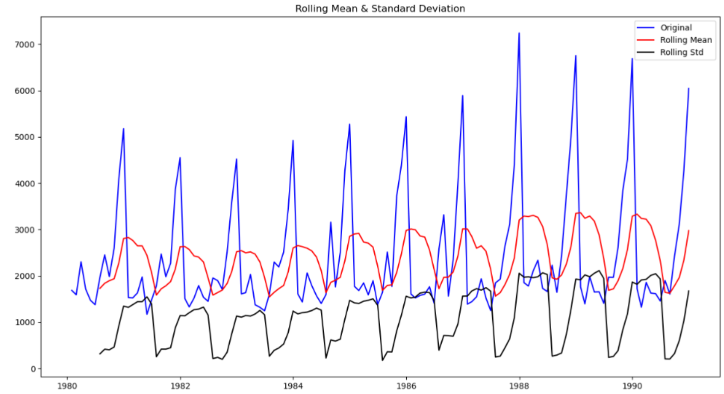


Figure 26: Original Training Time series

DF test statistic is -2.062

DF test p-value is 0.5674110388593698

• As the p value is larger than 0.05, we fail to reject the null hypotheses that Time Series is non-stationary.

• The dataset is non-stationary at 95% confidence level. Differencing 'd' to make time series stationary

Dicky Fuller Test on the differenced dataset to check stationarity

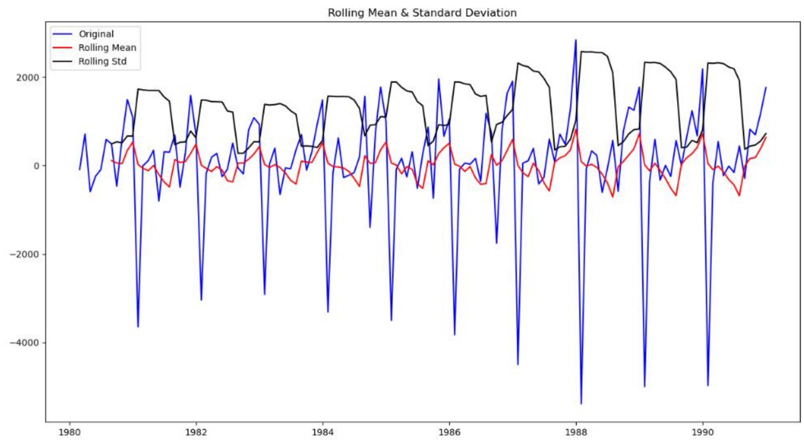


Figure 27: Training Time series after differencing d=1

DF test statistic is -7.968

DF test p-value is 8.479210655514579e-11

Observations:

• As the p value is less than 0.05, we reject the null hypotheses that Time Series is non-stationary.

• The Training data is stationary at 95% confidence level.

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.

ARIMA model

ARIMA:- Auto Regressive Integrated Moving Average is a way of modelling 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 modelled based on past observations.

Integrated component: Whether the data needs to be differenced/transformed.

MA model: Previous forecast errors are incorporated into the model.

ARIMA Model building to estimate best 'p' , 'd' , 'q' parameters (Lowest AIC Approach)

We split the dataset into Training and Testing set , with recent observations in the testing dataset. Training Data is used to train (develop) the ARIMA model. Training Data is used to identify a few working models with different values of p,d,q. Estimate p,d,q by looking at the lowest AIC for the models built on training data

The model built parameters is then used on the training data to forecast the test data and calculate model evaluation parameters like RMSE .

After the best model is selected , model is checked using diagnostics on the whole data and forecast for the desired future time points using this model .

The table below shows AIC scores for different values of p,d,q listed in ascending order of AIC scores

	param	AIC
10	(2, 1, 2)	2213.509213
15	(3, 1, 3)	2221.459399
14	(3, 1, 2)	2230.757366
11	(2, 1, 3)	2232.831964
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)	2235.500194
7	(1, 1, 3)	2235.607812
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

Figure 28 ARIMA AIC values

AS we can see Param 2,1,2 has the lowest AIC score. Let’s build the model using the param 2,1,2

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sun, 27 Aug 2023   AIC                           2213.509
Time:                        15:01:43   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     28.783      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.740      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.216      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.108      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
===================================================================================
```

Figure 29:Summary of ARIMA model

We can see from the summary above that ar.L1, ar.L2, ma.L1,ma.L2 are significant variables in building the model equation .

Let's see the diagnostics plot.

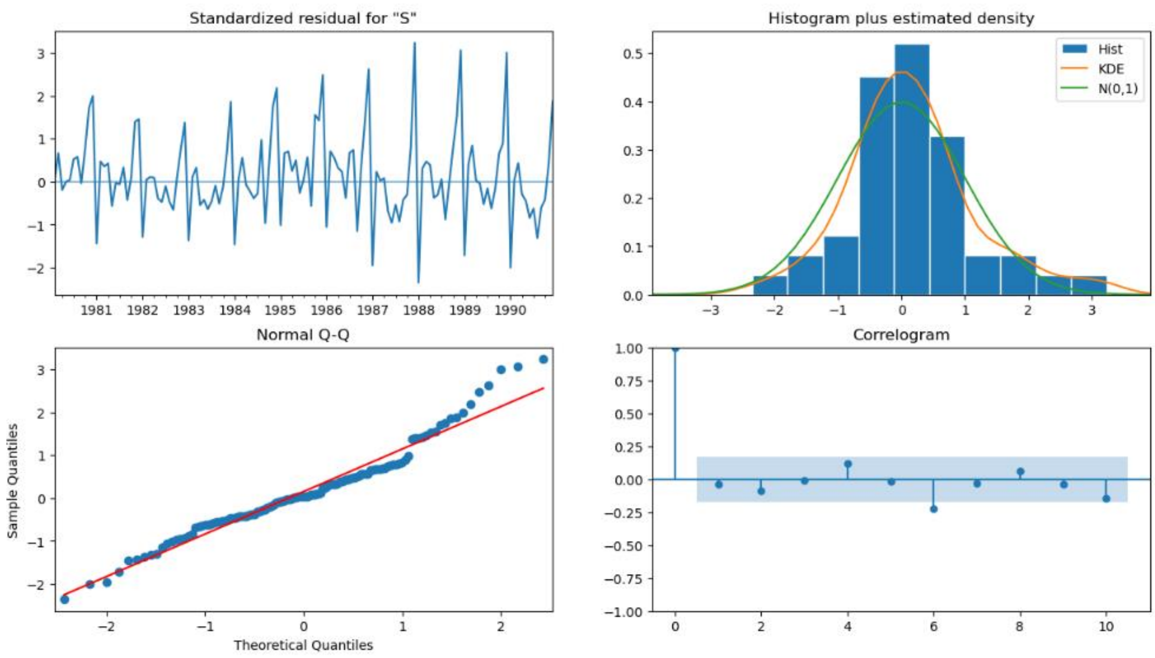


Figure 30: Diagnostics plot of ARIMA model

Now we can predict on the Test Set using this model and evaluate the model.

Let's evaluate the model using RMSE

The Test RMSE score for ARIMA 2,1,2 model is 1299.979.

ARIMA model on the training data for which the best parameters are selected by looking at the ACF and the PACF

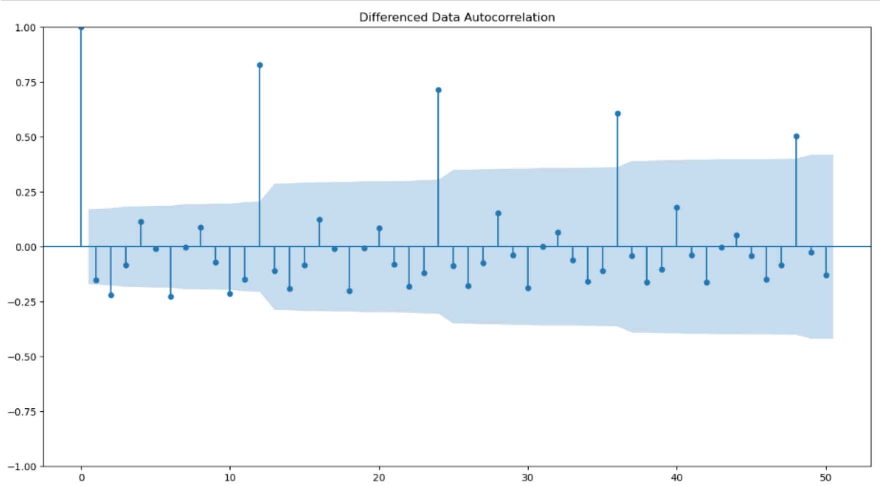


Figure 31: ACF of Training dataset

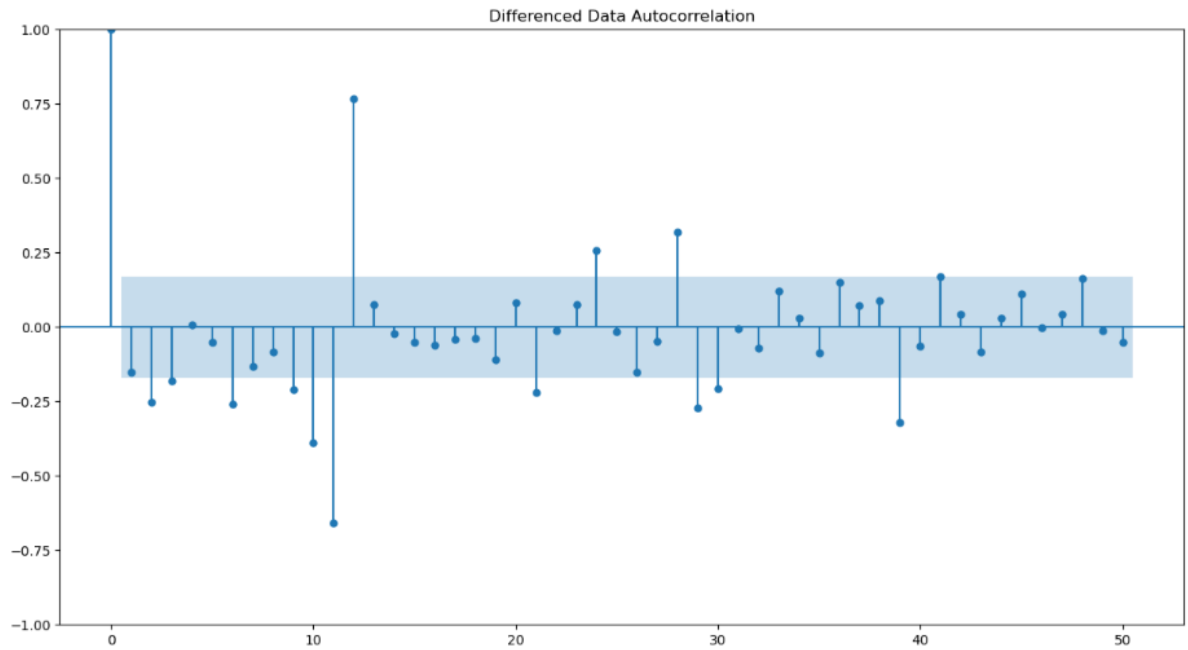


Figure 32: PACF plot of Training Dataset

There are no significant peaks in the ACF and PACF plots, so let's take p and q as 0 because significant peaks are only at 12, 24, 36 in ACF plots

Manual ARIMA(0, 1, 0)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(0, 1, 0)   Log Likelihood               -1132.832
Date:                Fri, 01 Sep 2023   AIC                           2267.663
Time:                        18:37:42   BIC                           2270.538
Sample:                    01-31-1980   HQIC                          2268.831
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      1.885e+06   1.29e+05     14.658      0.000    1.63e+06    2.14e+06
===================================================================================
Ljung-Box (L1) (Q):                   3.07   Jarque-Bera (JB):               198.83
Prob(Q):                              0.08   Prob(JB):                         0.00
Heteroskedasticity (H):               2.46   Skew:                            -1.92
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.65
-----------------------------------------------------------------------------------
```

Table 22: Summary of ARIMA(0, 1, 0) model

. As we can see there are error component in the model that is significant

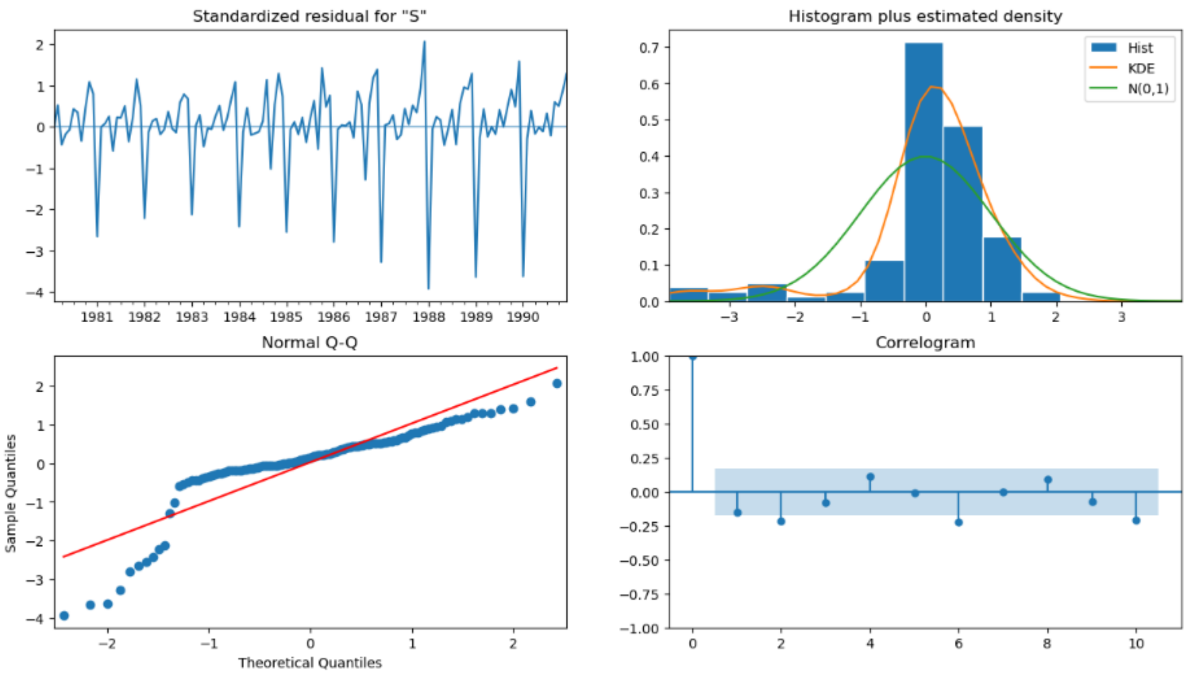


Figure 33: Diagnostics plot of Auto ARIMA(0, 1, 0)

Predict on the Test Set using these models and evaluate the model.

RMSE score for the Manual ARIMA model(0,1,0) is 3864.27

Lets plot the ARIMA predictions timeseries against Training and Testing Dataset

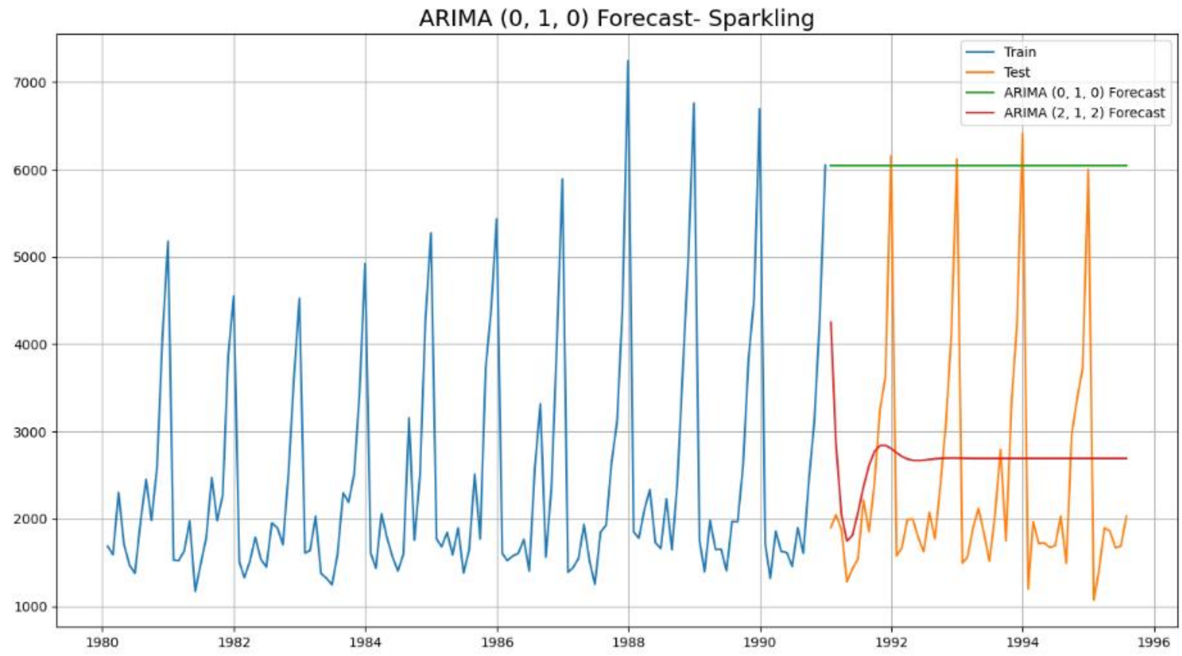


Figure 34: plot of ARIMA(2,1,2) and ARIMA(0,1,0) model

As we can see from plot above ARIMA model 01,0 and 2,1,2 are not very good at predicting slaes on testing data as its not taking seasonality into effect

SARIMA model

Although ARIMA method can handle data with a trend, it does not support time series with a seasonal component. An extension to ARIMA that supports the direct modelling of the seasonal component of the series is called SARIMA. It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

Configuring a SARIMA requires selecting hyperparameters for both the trend and seasonal elements of the series.

Trend Elements

There are three trend elements that require configuration.

They are the same as the ARIMA model; specifically:

p: Trend autoregression order.

d: Trend difference order.

q: Trend moving average order.

Seasonal Elements

There are four seasonal elements that are not part of ARIMA that must be configured; they are:

P: Seasonal autoregressive order.

D: Seasonal difference order.

Q: Seasonal moving average order.

F: The number of time steps for a single seasonal period.

Together, the notation for a SARIMA model is specified as:

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

Training Data is used to identify a few working models with different values of 'p' , 'd' , 'q' and 'P','D','Q' . Estimate p,d,q and P,D,Q by looking at the lowest AIC for the models built on training data

The model built parameters is then used on the training data to forecast the test data and calculate model evaluation parameters like RMSE .

After the best model is selected , model is checked using diagnostics on the whole data and forecast for the desired future time points using this model .

The following are the top AIC values

67	(3, 1, 3)	(3, 1, 1, 12)	1215.21335
155	(2, 1, 2)	(2, 0, 4, 12)	1215.85096
259	(3, 1, 1)	(3, 1, 0, 12)	1215.898777
251	(3, 1, 3)	(3, 1, 2, 12)	1216.480059
163	(3, 1, 2)	(3, 1, 0, 12)	1216.859179
783	(2, 1, 2)	(3, 0, 4, 12)	1216.883594
35	(2, 1, 2)	(1, 0, 4, 12)	1217.171842
43	(3, 1, 1)	(3, 1, 1, 12)	1217.713895
27	(3, 1, 2)	(3, 0, 4, 12)	1218.240069
260	(3, 1, 1)	(3, 1, 2, 12)	1218.416044
784	(3, 1, 2)	(3, 1, 1, 12)	1218.991384
131	(3, 1, 2)	(3, 1, 2, 12)	1219.259979
719	(3, 1, 2)	(1, 0, 4, 12)	1220.254175
261	(2, 1, 4)	(1, 0, 3, 12)	1222.362945

Table 23: AIC scores of SARIMA model

The table above shows AIC scores for different values of (p,d,q) and (P, D, Q) and listed in ascending order of AIC scores.

Let's build model with different AIC scores and see the diagnostics plot to determine the best model

Model building using order(3, 1, 3)(3, 1, 1, 12)

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(3, 1, 3)x(3, 1, [1], 12)   Log Likelihood                -596.607
Date:                            Fri, 01 Sep 2023   AIC                           1215.213
Time:                                    20:05:20   BIC                           1241.416
Sample:                                         0   HQIC                          1225.719
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6125      0.187     -8.639      0.000      -1.978      -1.247
ar.L2         -0.6125      0.302     -2.030      0.042      -1.204      -0.021
ar.L3          0.0826      0.161      0.513      0.608      -0.233       0.398
ma.L1          0.9841      0.473      2.079      0.038       0.057       1.912
ma.L2         -0.8777      0.168     -5.226      0.000      -1.207      -0.549
ma.L3         -0.9470      0.489     -1.939      0.053      -1.904       0.010
ar.S.L12      -0.5581      0.746     -0.749      0.454      -2.019       0.903
ar.S.L24      -0.2765      0.331     -0.835      0.404      -0.926       0.373
ar.S.L36      -0.1251      0.192     -0.650      0.515      -0.502       0.252
ma.S.L12       0.1072      0.772      0.139      0.890      -1.406       1.620
sigma2      1.839e+05   8.81e+04      2.087      0.037    1.12e+04    3.57e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 3.92
Prob(Q):                              0.93   Prob(JB):                         0.14
Heteroskedasticity (H):               0.73   Skew:                             0.47
Prob(H) (two-sided):                  0.42   Kurtosis:                         3.54
===================================================================================
```

Figure 35: Result Summary of Auto SARIMA(3, 1, 3)(3, 1, 1, 12) Model

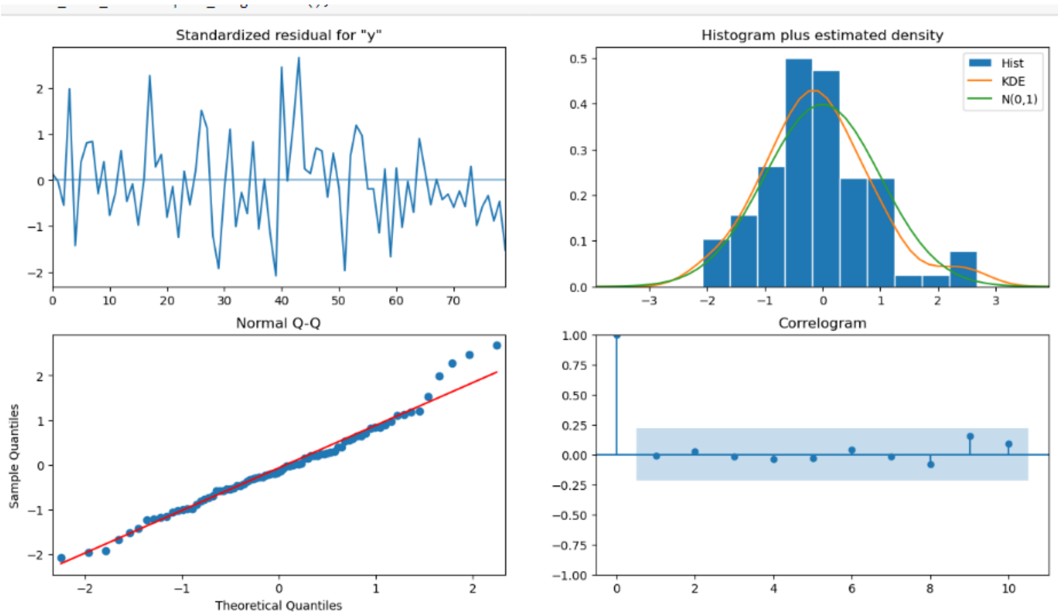


Figure 36:Diagnostic plot of Auto SARIMA(3, 1, 3)(3, 1, 1, 12) Model

By looking the SARIMA model summary, we can see that the coefficients for components and p-values of the components like – ar.L3,ma.L3 , ar.S.L12 , ar.S.L24 and ar.S.L36 of the SARIMA model are more than 0.05 so these are insignificant variables in prediction whereas the coefficients for components and p-values of the components like – ar.L1,ma.L1 , ar.L2 , ma.L2 of the SARIMA model are less than 0.05, hence significant in prediction

Model building using order(3, 1, 3)(3, 1, 0, 12)

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(3, 1, 3)x(3, 1, [], 12)   Log Likelihood                -596.641
Date:                            Fri, 01 Sep 2023   AIC                           1213.283
Time:                                    20:06:01   BIC                           1237.103
Sample:                                         0   HQIC                          1222.833
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6130      0.176     -9.175      0.000      -1.958      -1.268
ar.L2         -0.6103      0.299     -2.040      0.041      -1.197      -0.024
ar.L3          0.0867      0.161      0.540      0.589      -0.228       0.402
ma.L1          0.9853      0.464      2.124      0.034       0.076       1.895
ma.L2         -0.8740      0.166     -5.260      0.000      -1.200      -0.548
ma.L3         -0.9464      0.481     -1.966      0.049      -1.890      -0.003
ar.S.L12      -0.4520      0.142     -3.192      0.001      -0.730      -0.174
ar.S.L24      -0.2338      0.144     -1.620      0.105      -0.517       0.049
ar.S.L36      -0.1003      0.121     -0.825      0.409      -0.338       0.138
sigma2      1.839e+05   8.84e+04      2.081      0.037    1.07e+04    3.57e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 4.06
Prob(Q):                              0.93   Prob(JB):                         0.13
Heteroskedasticity (H):               0.73   Skew:                             0.48
Prob(H) (two-sided):                  0.41   Kurtosis:                         3.54
```

Figure 37: Result Summary of Auto SARIMA(3, 1, 3)(3, 1,0, 12)

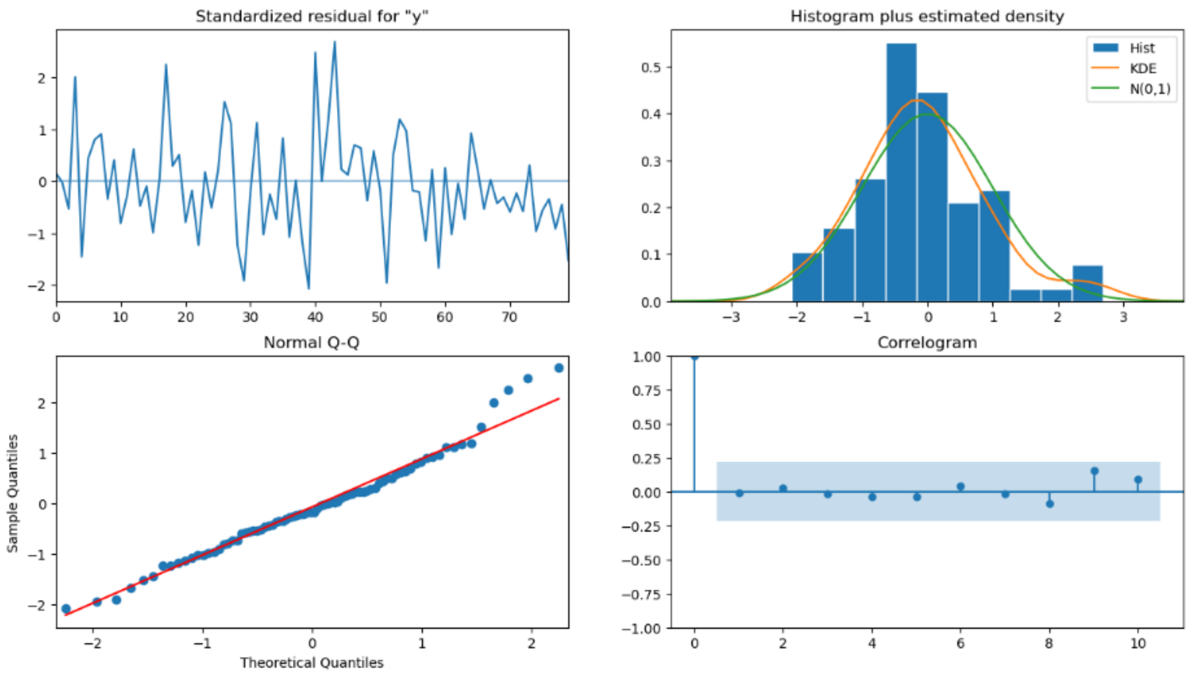


Figure 38: Diagnostics plot of Auto SARIMA(3, 1, 3)(3, 1,0, 12)

By looking the SARIMA model summary, we can see that the coefficients for components and p-values of the components like – ar.L3, ar.S.L24 and ar.S.L36 of the SARIMA model are more than 0.05 so these are insignificant variables in prediction whereas the coefficients for components and p-values of the components like – ar.L1,ma.L1 , ar.L2 , ma.L2.ma.L3 and ar.S.L12 of the SARIMA model are less than 0.05, hence significant in prediction.

Predict on the Test Set using this model and evaluate the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1434.482690	431.129516	589.484366	2279.481013
1	1539.016472	458.380029	640.608125	2437.424819
2	1713.498310	460.373757	811.182327	2615.814293
3	1858.333162	466.877144	943.270774	2773.395549
4	1505.223215	467.208128	589.512111	2420.934320

Table 24: Predictions on the test set with Auto SARIMA(3, 1, 3)(3, 1, 1, 12) Model

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1430.292563	431.167218	585.220345	2275.364781
1	1540.081835	458.496284	641.445631	2438.718040
2	1708.093756	460.259647	806.001423	2610.186088
3	1858.409087	466.808162	943.481902	2773.336273
4	1502.118483	467.105606	586.608317	2417.628648

Table 25: Predictions on the test set with Auto SARIMA(3, 1, 3)(3, 1, 0, 12) Model

RMSE score for the Auto SARIMA(3,1,3)(3,1,1,12) models is 331.710

RMSE score for the Auto SARIMA(3,1,3)(3,1,0,12) models is 331.610

SARIMA model on the training data for which the best parameters are selected by looking at the ACF and the PACF

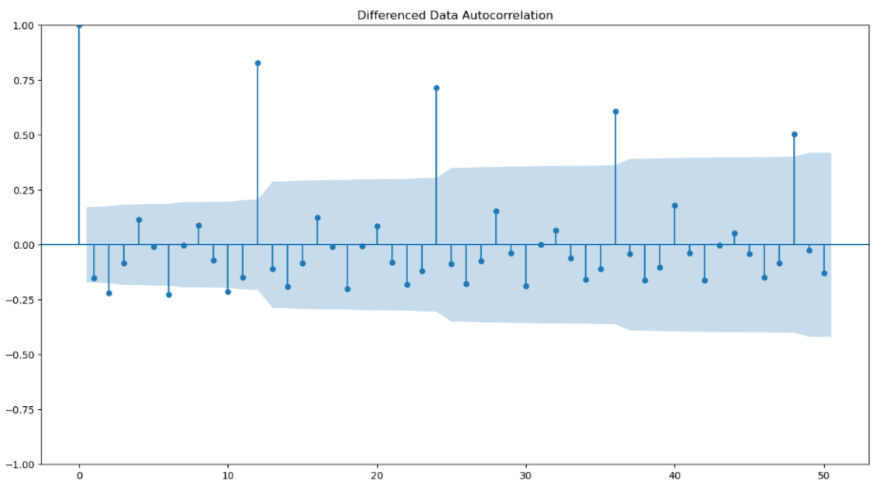


Figure 39: ACF of Training dataset

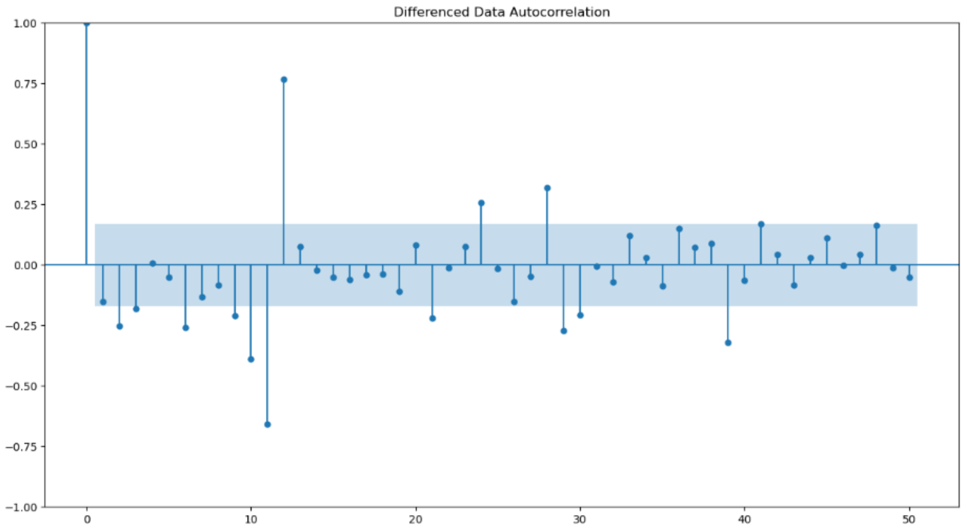


Figure 40: PACF plot of Training Dataset

There are no significant peaks in the ACF and PACF plots, so let's take p and q as 0 because significant peaks are only at 12, 24, 36 in ACF plots

To determine P and Q let's take values as 2 and 0 for one model and 4 and 2 for another model

Manual SARIMA(0, 1, 0)(2, 1, 4, 12).

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                        y   No. Observations:          132
Model:             SARIMAX(0, 1, 0)x(2, 1, [1, 2, 3, 4], 12)   Log Likelihood        -538.663
Date:                                  Fri, 01 Sep 2023   AIC                       1091.326
Time:                                          19:39:08   BIC                       1107.066
Sample:                                               0   HQIC                      1097.578
                                                  - 132
Covariance Type:                                    opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L12      -0.5734      0.253     -2.266      0.023      -1.070      -0.077
ar.S.L24      -0.5548      0.108     -5.147      0.000      -0.766      -0.344
ma.S.L12       0.3449      0.391      0.882      0.378      -0.422       1.111
ma.S.L24       0.5798      0.191      3.040      0.002       0.206       0.954
ma.S.L36      -0.5033      0.117     -4.306      0.000      -0.732      -0.274
ma.S.L48      -0.0809      0.349     -0.232      0.816      -0.764       0.602
sigma2      2.044e+05   1.02e-06      2e+11      0.000    2.04e+05    2.04e+05
===================================================================================
Ljung-Box (L1) (Q):                   7.81   Jarque-Bera (JB):                32.02
Prob(Q):                              0.01   Prob(JB):                         0.00
Heteroskedasticity (H):               0.32   Skew:                             0.95
Prob(H) (two-sided):                  0.01   Kurtosis:                         5.72
===================================================================================
```

Figure 41: Result Summary of Auto SARIMA(0, 1, 0)(2, 1,4, 12).

As we can from the summary parameters ma.S.L12 and ma.S.L48 are insignificant because the P values is greater than 0.05. So only ar.S.L12, ar.S.L24, ma.S.L24 and ma.S.L36 are significant in model building .

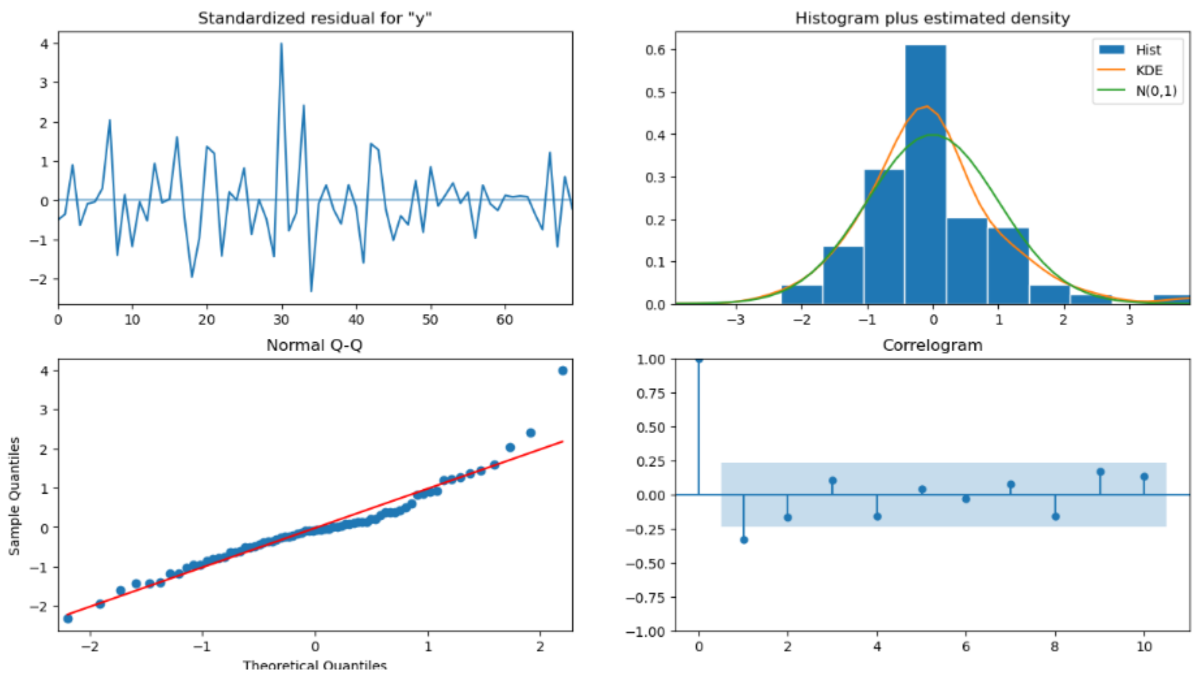


Figure 42: Diagnostics plot of Auto SARIMA(0, 1, 0)(2, 1,4, 12)

Manual SARIMA(0, 1, 0)(2, 1, 0, 12).

There are no significant peaks in the ACF and PACF plots, so let's take p and q as 0

P and Q let's take values as 2 and 0 for this model

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(0, 1, 0)x(2, 1, 0, 12)   Log Likelihood                -730.311
Date:                            Fri, 01 Sep 2023   AIC                           1466.621
Time:                                    19:39:12   BIC                           1474.283
Sample:                                         0   HQIC                          1469.717
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L12      -0.3653      0.084     -4.327      0.000      -0.531      -0.200
ar.S.L24      -0.2042      0.105     -1.946      0.052      -0.410       0.001
sigma2      2.785e+05   2.85e+04      9.778      0.000    2.23e+05    3.34e+05
===================================================================================
Ljung-Box (L1) (Q):                  12.26   Jarque-Bera (JB):                38.99
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.78   Skew:                             0.77
Prob(H) (two-sided):                  0.48   Kurtosis:                         5.73
===================================================================================
```

Figure 43: Result Summary of Auto SARIMA(0, 1, 0)(2, 1,0, 12)

As we can from the summary parameters ar.S.L24 is insignificant because the P values is greater than 0.05. So only ar.S.L24 is significant in model building .

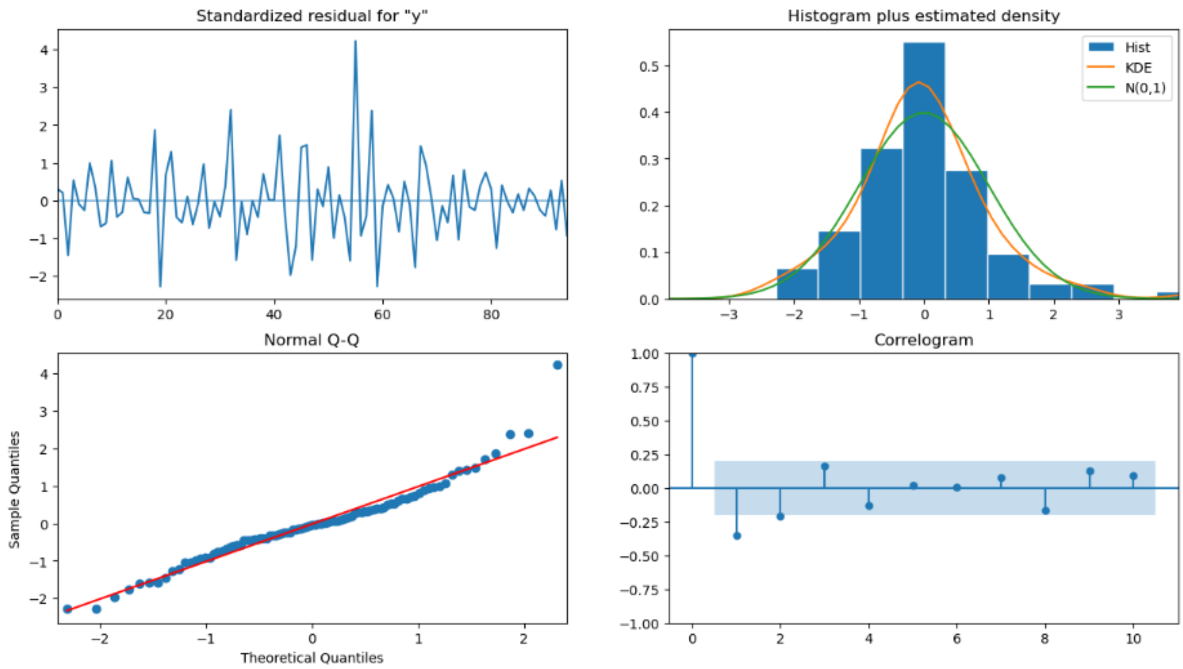


Figure 44: Diagnostics plot Auto SARIMA(0, 1, 0)(2, 1,0, 12)

Predict on the Test Set using these models and evaluate the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1258.379360	501.571361	275.317557	2241.441163
1	890.013778	709.162809	-499.919787	2279.947344
2	1352.591554	868.475647	-349.589437	3054.772544
3	1241.696847	1002.790113	-723.735658	3207.129352
4	1232.913127	1121.127143	-964.455694	3430.281949

Table 26: Predictions on the test set with SARIMA(0, 1, 0)(2, 1,4, 12) Model

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	984.085029	527.706960	-50.201608	2018.371666
1	657.236488	746.290340	-805.465701	2119.938677
2	1160.621583	914.015267	-630.815421	2952.058588
3	1007.060876	1055.413921	-1061.512397	3075.634150
4	875.324006	1179.988636	-1437.411222	3188.059234

Table 27: Predictions on the test set with SARIMA(0, 1, 0)(2, 1,0, 12) Model

RMSE score for the Manual SARIMA(0, 1, 0)(2, 1,4, 12) models is 937.54

RMSE score for the Manual SARIMA(0, 1, 0)(2, 1,0, 12) models is 1779.214

Let's plot the best SARIMA models

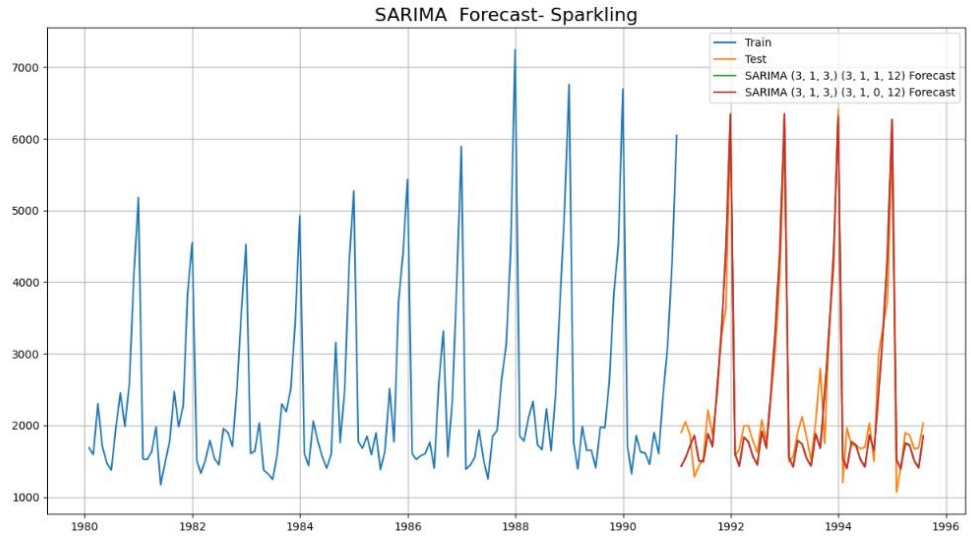


Figure 45: Plot of the SARIMA model vis-à-vis Training and Testing Graphs

The SARIMA 3,1,3(3,1,0,12) and 3,1,3(3,1,1,12) are performing well on the testing dataset and as indicated in the plot above they fit well with testing data

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

After building various time series forecasting models like Linear Regression , Naïve Forecast ,Simple Average , Moving Average and exponential smoothing models like (Simple , Double , Triple Exponential) and ARIMA / SARIMA models on the Sparkling Wine Sales dataset , here are the models and their RMSE on the test data

	Test RMSE
Alpha=0.0496:SimpleExponentialSmoothing	1316.035487
Alpha=0.688,Beta=0.0001:DoubleExponentialSmoothing	2007.238526
Alpha=0.111338,Beta=0.049505,Gamma=0.362080:TripleExponentialSmoothingMultiplicative	404.286809
Alpha=0.111272,Beta=0.012361,Gamma=0.460718:TripleExponentialSmoothingAdditive	378.951023
Alpha=0.02:SimpleExponentialSmoothing	1279.495201
Alpha=0.02,Beta=0.50,IterativeDoubleExponentialSmoothing	1274.630824
RegressionOnTime	1389.135175
NaiveModel	3864.279352
Simple Average	1275.081804
2pointTrailingMovingAverage	3046.976092
4pointTrailingMovingAverage	2021.855880
6pointTrailingMovingAverage	1521.611250
9pointTrailingMovingAverage	1304.618442
ARIMA(2,1,2)	1299.979753
ARIMA(0,1,0)	3864.279352
SARIMA(0, 1, 0)(2, 1, 4, 12)	937.540131
SARIMA(0, 1, 0)(2, 1, 0, 12)	1779.214720
SARIMA(3, 1, 3)(3, 1, 1, 12)	331.710438
SARIMA(3, 1, 3)(3, 1, 0, 12)	331.610287

Table 28: RMSE values for all the models

As we can see Test RMSE of SARIMA models (3, 1, 3)(3, 1,0, 12) and models (3, 1, 3)(3, 1,1, 12) and Triple Exponential Smoothing Additive Models is least among all the models.

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.

After building various time series forecasting models like Linear Regression , Navie Forecast ,Simple Average , Moving Average and exponential smoothing models like (Simple , Double , Triple Exponential) and ARIMA / SARIMA models on the Sparkling Wine Sales dataset and on comparing their RMSE on the test data we deduce that the Test RMSE of SARIMA models (3, 1, 3)(3, 1,0, 12) , models (3, 1, 3)(3, 1,1, 12) and Triple Exponential Smoothing Models is least among all the models with different parameters.

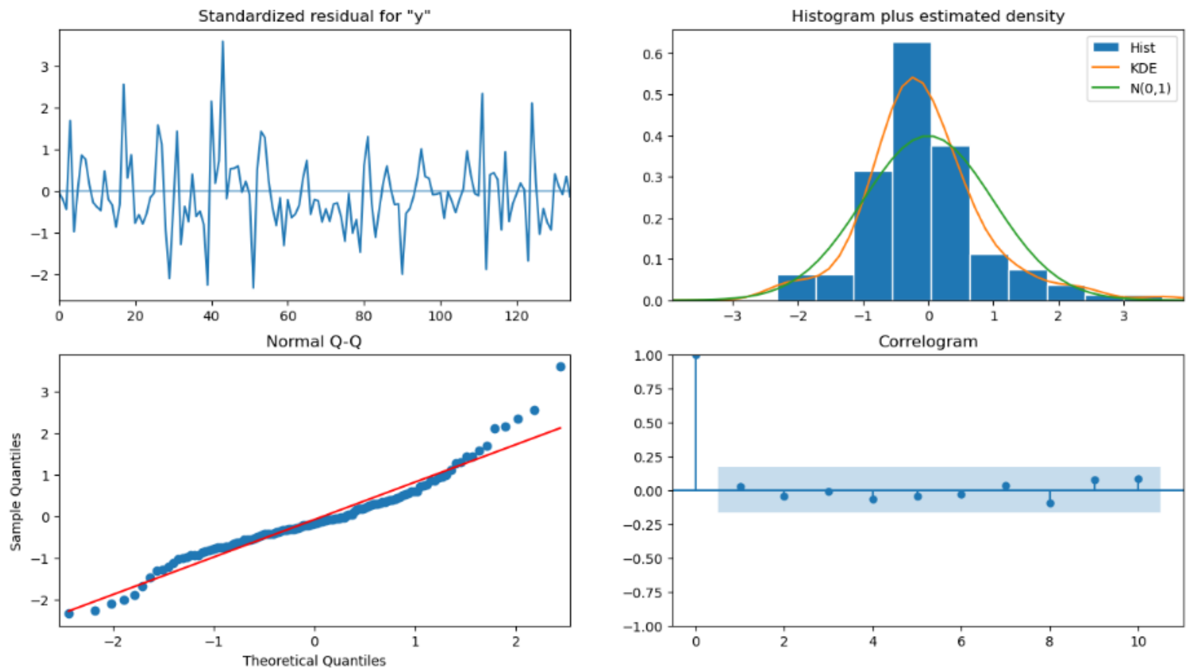
Lets build Model on complete data and predict 12 months into the future with appropriate confidence intervals/bands using SARIMA models (3, 1, 3)(3, 1,0, 12)

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  187
Model:             SARIMAX(3, 1, 3)x(3, 1, [], 12)   Log Likelihood                -999.553
Date:                           Fri, 01 Sep 2023   AIC                           2019.106
Time:                                   21:26:05   BIC                           2048.159
Sample:                                        0   HQIC                          2030.912
                                           - 187
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.0415      0.132     -7.907      0.000      -1.300      -0.783
ar.L2         -0.8621      0.129     -6.664      0.000      -1.116      -0.609
ar.L3          0.0708      0.113      0.628      0.530      -0.150       0.292
ma.L1          0.1977      0.149      1.330      0.183      -0.094       0.489
ma.L2         -0.1380      0.132     -1.044      0.297      -0.397       0.121
ma.L3         -0.9528      0.213     -4.471      0.000      -1.371      -0.535
ar.S.L12      -0.5531      0.093     -5.919      0.000      -0.736      -0.370
ar.S.L24      -0.2657      0.142     -1.865      0.062      -0.545       0.014
ar.S.L36      -0.1412      0.106     -1.337      0.181      -0.348       0.066
sigma2      1.852e+05   4.06e+04      4.560      0.000    1.06e+05    2.65e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.11   Jarque-Bera (JB):                40.64
Prob(Q):                              0.74   Prob(JB):                         0.00
Heteroskedasticity (H):               0.55   Skew:                             0.71
Prob(H) (two-sided):                  0.05   Kurtosis:                         5.28
===================================================================================
```

Figure 46 Result summary of SARIMA3,13(3,1,0,12)

As we can from the summary parameters ar.L3,ma.L1,ma.L2,ar.S.L24 and ar.S.L36 are insignificant because the P values is greater than 0.05. So only ar.L1, ar.L2 , ma.L3, ar.S.L12 are significant in model building

Lets plot the diagnostics plot to check residuals



Residuals are not normally distributed , but can be used for model building

Figure 47:Diagnostics plot of SARIMA 3,1,3(3,1,0,12)

Prediction on the Data using these models and evaluate the model.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1937.839435	431.220508	1092.662771	2783.016100
1	2396.749831	436.282219	1541.652394	3251.847267
2	3331.407580	436.500682	2475.881965	4186.933196
3	3877.722482	436.517633	3022.163642	4733.281322
4	6095.066296	437.814878	5236.964903	6953.167689

Table 29: Predictions on the Entire data set with SARIMA(3, 1, 3)(3, 1,0, 12) Model

RMSE score for the Manual SARIMA(3, 1, 3)(3, 1,0, 12) models is 614.656

Let's plot the future prediction of 12 months alongside original time series

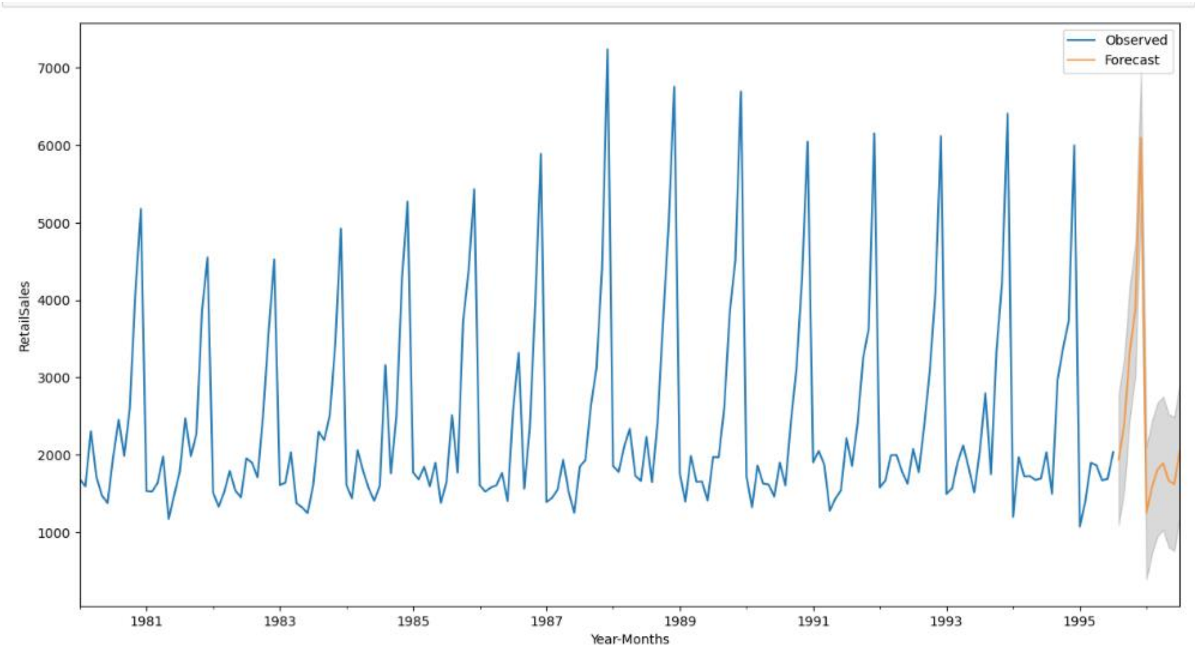


Figure 48: Prediction for the next 12 months with confidence intervals

Plot above shows Model building on complete data and predict 12 months into the future with appropriate confidence intervals/bands using SARIMA model . The orange line traces the next 12 months forecast and as we can see the forecast indicates there the peak sales in December, and a drop in sales in the following months as the original dataset

Let's build Model on complete data and predict 12 months into the future with appropriate confidence intervals/bands using TES model

After fitting the TES Additive model on the entire dataset, the following are the predicted values and confidence intervals for the next 12 months

	lower_CI	prediction	upper_ci
1995-08-31	1148.760806	1852.913618	2557.066429
1995-09-30	1755.095900	2459.248711	3163.401523
1995-10-31	2479.740231	3183.893042	3888.045854
1995-11-30	3081.197419	3785.350230	4489.503041
1995-12-31	5227.478286	5931.631097	6635.783909
1996-01-31	512.344503	1216.497315	1920.650126
1996-02-29	884.697157	1588.849969	2293.002780
1996-03-31	1143.731494	1847.884305	2552.037117
1996-04-30	1126.250580	1830.403392	2534.556203
1996-05-31	963.386057	1667.538868	2371.691680
1996-06-30	916.633460	1620.786272	2324.939083
1996-07-31	1266.867591	1971.020402	2675.173214

Table 30: Forecasted values for 12 months in the future

Let's plot the future prediction of 12 months alongside original time series

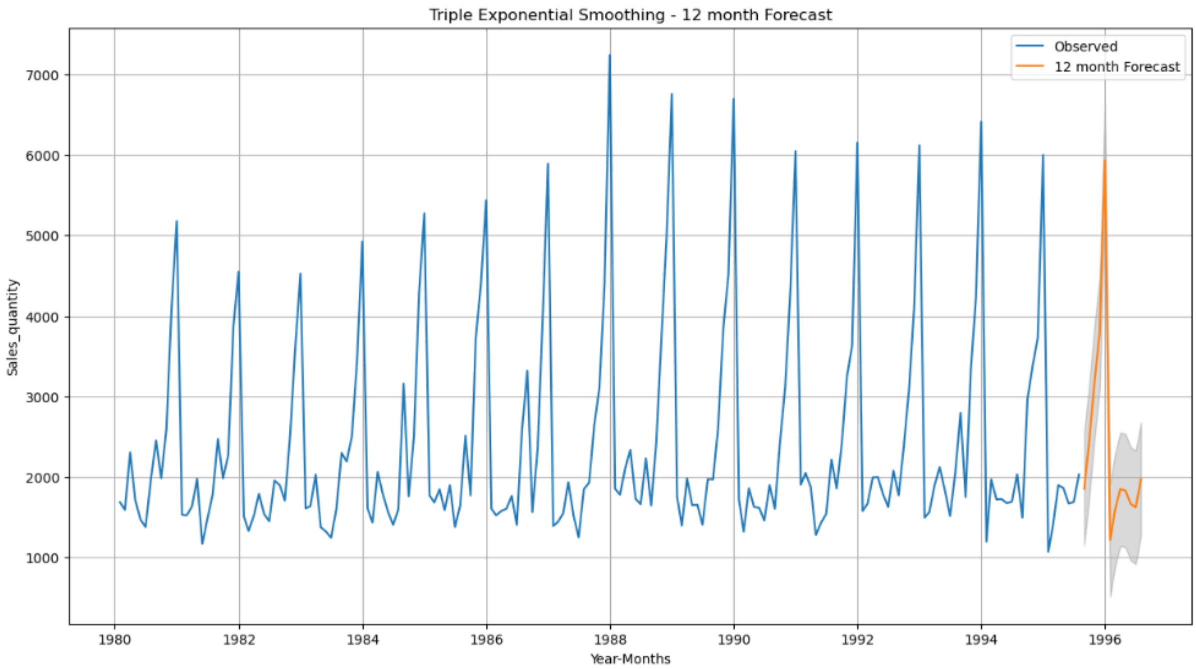


Figure 49: Forecasted time series for next 12 months

The orange line traces the next 12 months forecast and as we can see the forecast indicates there the peak sales in December, and a drop in sales in the following months as the original dataset

RMSE score for the TES models is 370.30

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

Comments on Final Model

After building various time series forecasting models like Linear Regression , Naive Forecast ,Simple Average , Moving Average and exponential smoothing models like (Simple , Double , Triple Exponential) and ARIMA / SARIMA models on the Sparkling Wine Sales dataset and on comparing their RMSE on the test data we deduce that the Test RMSE of SARIMA (3, 1, 3)(3, 1,0, 12) is least among all the models with different parameters.

So, we take Manual SARIMA(3, 1, 3)(3, 1,0, 12) and Triple Exponential Additive model to build best fit mode on complete data taking into considerations of order and seasonality and predict 12 months into the future with appropriate confidence intervals/bands. The forecast also indicates there the peak sales in December, and a drop in sales in the following months as the original dataset.

RMSE score of the SARIMA Model is 614.656

RMSE score for the TES models is 370.**3**0

From the Plot of the forecast on full data along with the 95% confidence interval we infer that forecast also follows the same pattern as original sparkling wine sales series follows. Forecast indicates there the peak sales in December, and a drop in sales in the following months as the original dataset

Findings based on EDA / Data Visualization and Time Series Forecasting Models

- The data is from year 1980 to 1995 .
- The highest sales are recorded in the month of December across all years.
- The lowest sales are recorded in the month of June across all years.
- Sales was on upward trend till 1988 and from then on there hasn't been an upward trend.
- From September to December the Sparkling Wine Sales increases so this is the period where the Sparkling Wine Sales is highest which shows the seasonality in Sparkling Wine Sales.
- Boxplot of sales indicates Mean of the data is 2402.42 and Median is 1874.
- Minimum sales recorded for a month is 1070.
- Maximum sales recorded for a month is 7242.
- There are outliers in the sales data.
- From the Plot of the forecast on full data along with the confidence band we infer that with 95% of the confidence level we found that forecast also follows the same pattern as original sparkling wine sales series follows.

Measures for Future Sales

- Sales is highest in December and lowest in June, it could be due to holiday and tourist season coming to an end. But there could be other factors due to which sales is dropping which is not available as part of the dataset .
- Various factors like wine quality, wine supply and demand, pricing, availability of better alternatives, not enough marketing of the product, shelf life could be reasons for drop in sales over the years. Company should take measures to address these factors
- The ABC Estate Wines company should develop marketing strategies to promote Sparkling Wine Sales . During the months wine sales is low , company can run various offers during this period to boost their wine sales to attract more customers.
- Wine pairings are a great way to introduce customers to new choices. So the company can let customers sample the Sparkling wine and have culinary experiences to promote Sparkling wine
- Give a variety of Sparkling bottle sizes to offer to a customer. Having different bottle sizes is a great way to appeal to large groups and couples. Selling bottles to groups if they only have to buy 2 or 3, and couples can commit to smaller half bottles. Marketing studies state that if you make your product more accessible to your customers, they're more likely to buy. Serving different-sized bottles does exactly that
- Tying up with restaurants and hotels that serve alcohol to run offers on Sparkling wine for customers to try and suggest wine pairings, so that customers take a liking towards the Wine
- The Staff should be approachable and knowledgeable enough to make informed recommendations and conversations about the wine. The staff members who are well informed will be more likely to confidently make a recommendation or upsell the wine to the customer. Teach them to sell the story and not just the wine and dramatically increases wine sales.

- Offering a box of miniature-wines which allows clients to purchase several of your wines as tasting samples. This allow customers to taste wine before committing to large purchases.
- Online marketplaces and Platforms, like Google Products, can make your wine more visible when someone googles a specific type of wine or similar products. Its free marketing and potential sales, to anyone interested in the exact product that you offer.
- Offering an efficient Wine Delivery-Service that delivers quickly and efficiently by allowing customers the opportunity to purchase wine online.
- Paid Ad's are a way of targeting a particular group of people .Facebook ads are relatively affordable considering how targeted the advertising can be.
- Host Tasting Events and Offer Wine Subscription Boxes could be another way to boost sales

End of the report