TIME SERIES FORECASTING BUSINESS REPORT – SPARKLING WINE SALES

TEJAS PADEKAR
PGP-DSBA Online
FEB’ 22
Date: 20/02/2022

CONTENTS:

List of Figures

List of Tables

PROBLEM 1

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

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

Dataset Head:

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

Dataset Tail:

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

The dataset contents 187 observations across 02 columns in total.

Date Time Index:

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

Time Stamp:

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

We do not require the column of "YearMonth" as we have created a Time Stamp for the same and made it as our index column as well. Hence, we have dropped "YearMonth" from our dataset.

Dataset Description:

	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

We observe from a historical record of 187 months since Jan 1980 until July 1995 that average sales over the period of Sparkling wine was 2402 bottles. The least being 1070 bottles and highest being 7242 bottles. Now, we have our data ready for the Time Series Analysis.

Time Series Plot

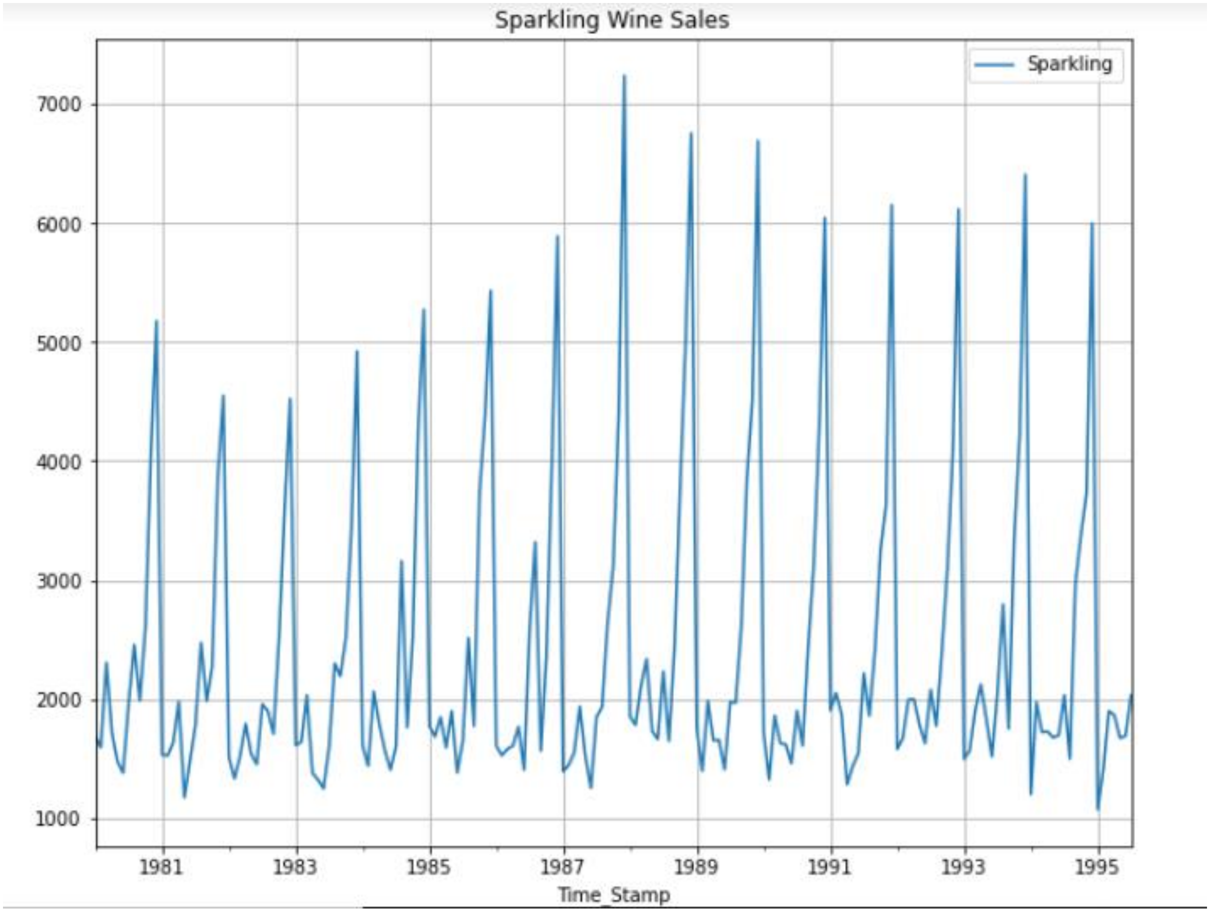


Figure 1: Time Series Plot

For above figure, we observe presence of seasonality throughout the time series. However, there is presence of trend as well at various time frames for eg; we see an uptrend from the year 1983 until 1987 and then from 1987 until 1991 it has reversed into a downtrend. From 1991 until the end of the time series, the trend is somewhat static with no major changes observed in the sales figures of the Sparkling Wine.

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

Yearly Boxplot:

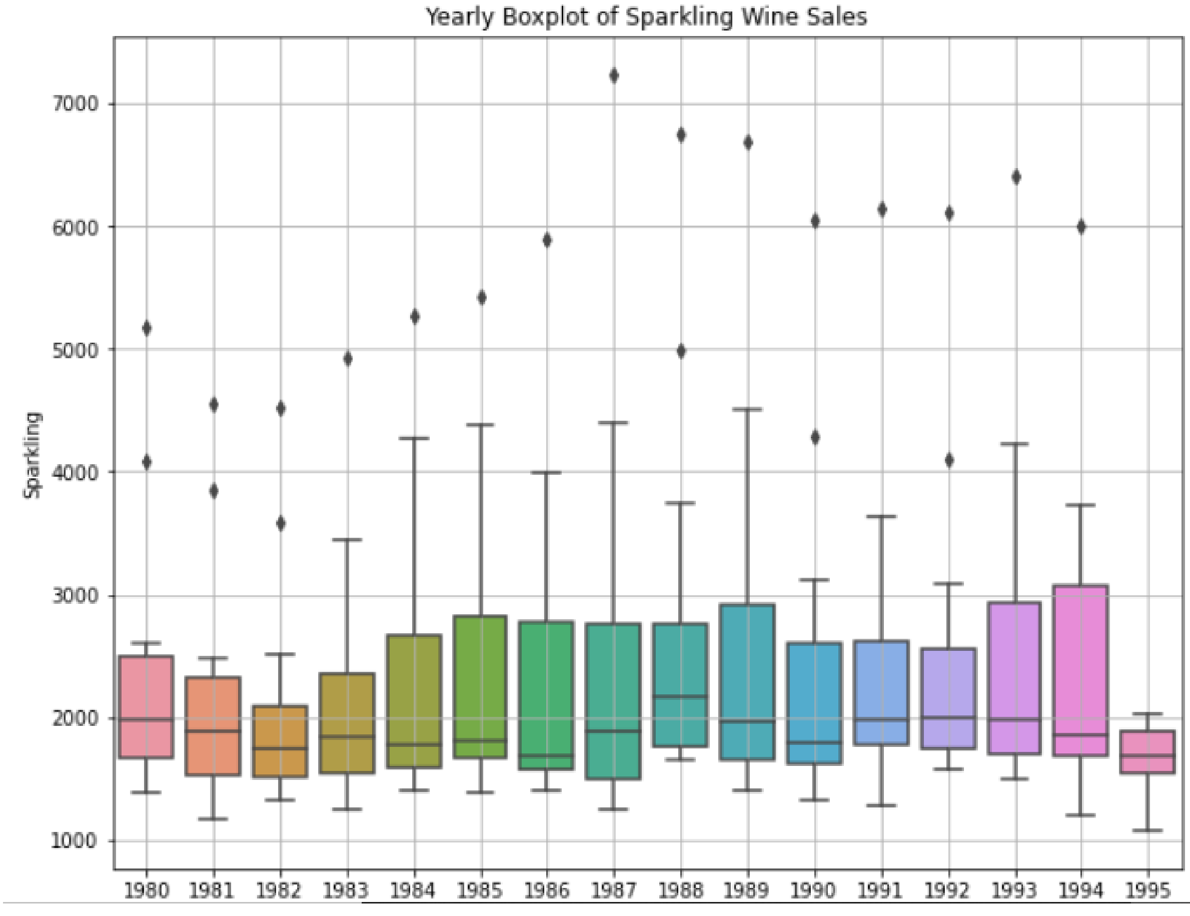


Figure 2 – Yearly Boxplot of Sparkling Wine Sales

The sales figures can be seen to pretty much ranges between 1000 bottles to 5000 bottles with outlier months where they have exceeded 5000 bottles as well. The least sales figure is seen for the last year of our time series which is also the only year where there are no outliers observed.

However, it mainly also could be seen as least as the data was provided only until the month of July and the boxplot shows us that the sales have been better in the 2nd half of the year throughout the time series (See Below Figure 3: Monthly Boxplot of Sparkling Wine Sales)

Also, the median sales for year 1995 until July was better or almost similar to the year of 1986 and we can see that the sales increased massively in the 2nd half of 1986.

Monthly Boxplot:

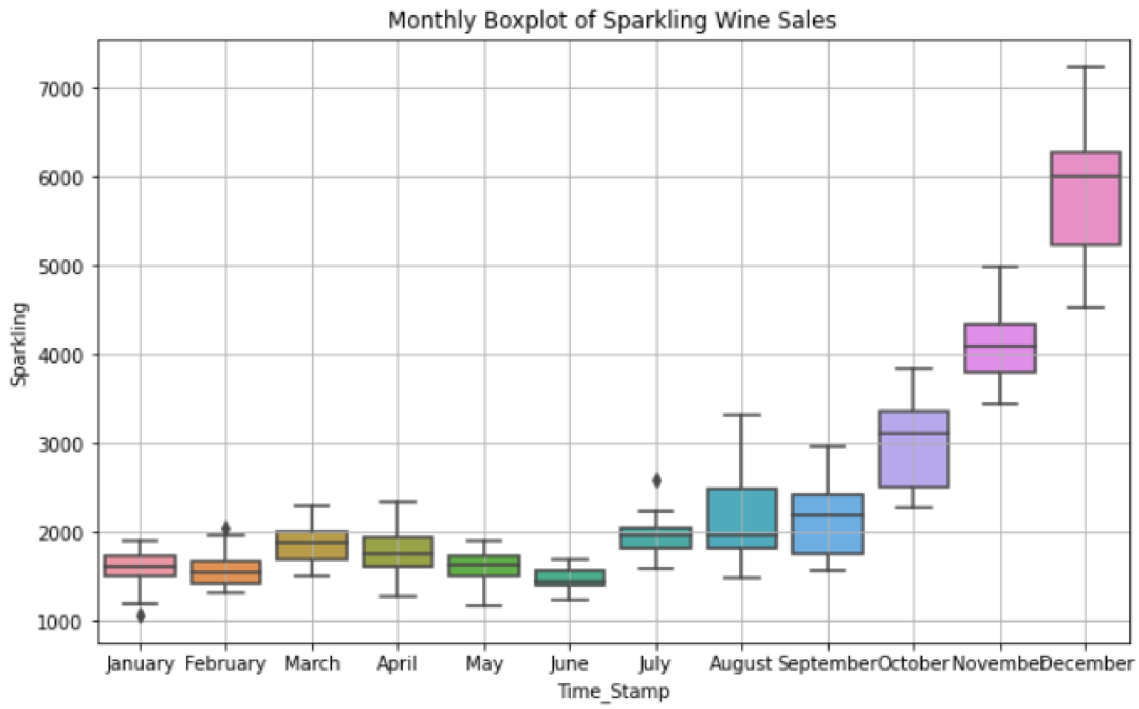


Figure 3 – Monthly Boxplot of Sparkling Wine Sales

As stated earlier, the sales have been better in the 2nd half of the year throughout the time series especially from August onwards. There is a clear upward trend in sales from September onwards. The least sales have come in the month of June and highest from December throughout various years!

Outliers are present for January, February and July months.

Monthly Boxplot with Median Values as Red Line:

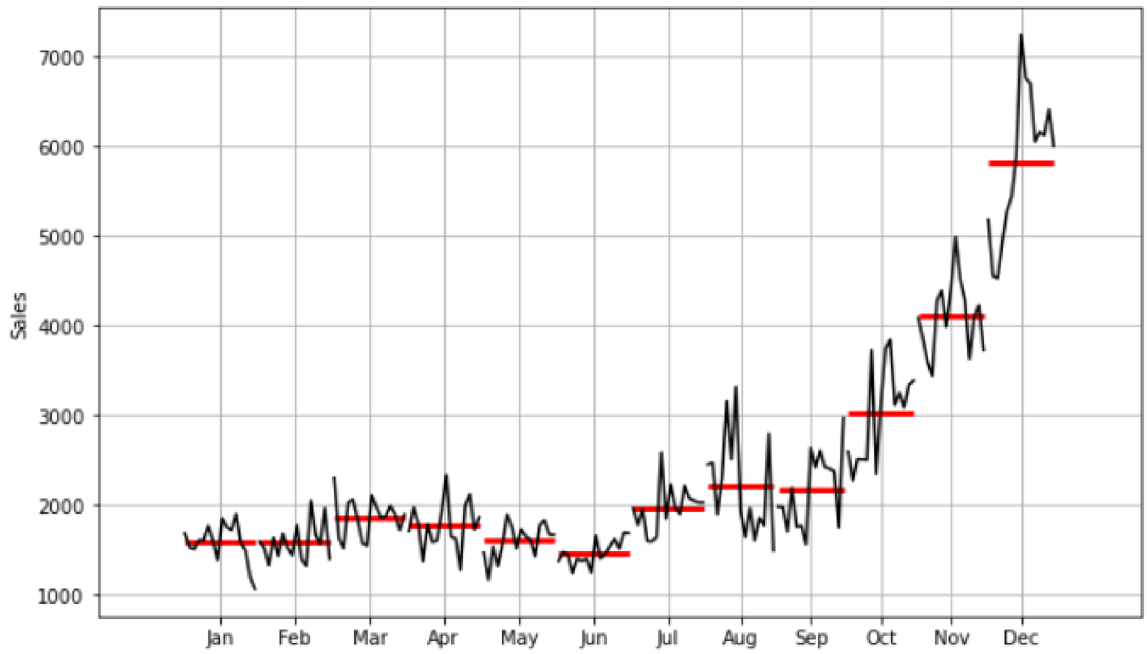


Figure 4 – Monthly Boxplot with Median Values as Red Line

The median values are within 2000 bottles until July and picks up slightly in August and September. However, from October until December there are huge spikes in the median values with 3000 bottles in October, 4000+ in November and almost 6000 in December.

Pivot Table Monthly Sales:

Time_Stamp Time_Stamp	1	2	3	4	5	6	7	8	9	10	11	12
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	2514.0	3593.0	4524.0
1983	1609.0	1638.0	2030.0	1375.0	1320.0	1245.0	1600.0	2298.0	2191.0	2511.0	3440.0	4923.0
1984	1609.0	1435.0	2061.0	1789.0	1567.0	1404.0	1597.0	3159.0	1759.0	2504.0	4273.0	5274.0
1985	1771.0	1682.0	1846.0	1589.0	1896.0	1379.0	1645.0	2512.0	1771.0	3727.0	4388.0	5434.0
1986	1606.0	1523.0	1577.0	1605.0	1765.0	1403.0	2584.0	3318.0	1562.0	2349.0	3987.0	5891.0
1987	1389.0	1442.0	1548.0	1935.0	1518.0	1250.0	1847.0	1930.0	2638.0	3114.0	4405.0	7242.0
1988	1853.0	1779.0	2108.0	2336.0	1728.0	1661.0	2230.0	1645.0	2421.0	3740.0	4988.0	6757.0
1989	1757.0	1394.0	1982.0	1650.0	1654.0	1406.0	1971.0	1968.0	2608.0	3845.0	4514.0	6694.0
1990	1720.0	1321.0	1859.0	1628.0	1615.0	1457.0	1899.0	1605.0	2424.0	3116.0	4286.0	6047.0
1991	1902.0	2049.0	1874.0	1279.0	1432.0	1540.0	2214.0	1857.0	2408.0	3252.0	3627.0	6153.0
1992	1577.0	1667.0	1993.0	1997.0	1783.0	1625.0	2076.0	1773.0	2377.0	3088.0	4096.0	6119.0
1993	1494.0	1564.0	1898.0	2121.0	1831.0	1515.0	2048.0	2795.0	1749.0	3339.0	4227.0	6410.0
1994	1197.0	1968.0	1720.0	1725.0	1674.0	1693.0	2031.0	1495.0	2968.0	3385.0	3729.0	5999.0
1995	1070.0	1402.0	1897.0	1862.0	1670.0	1688.0	2031.0	NaN	NaN	NaN	NaN	NaN

Figure 5 – Pivot Table of Sparkling Wine Sales

Monthly Sales across Years:

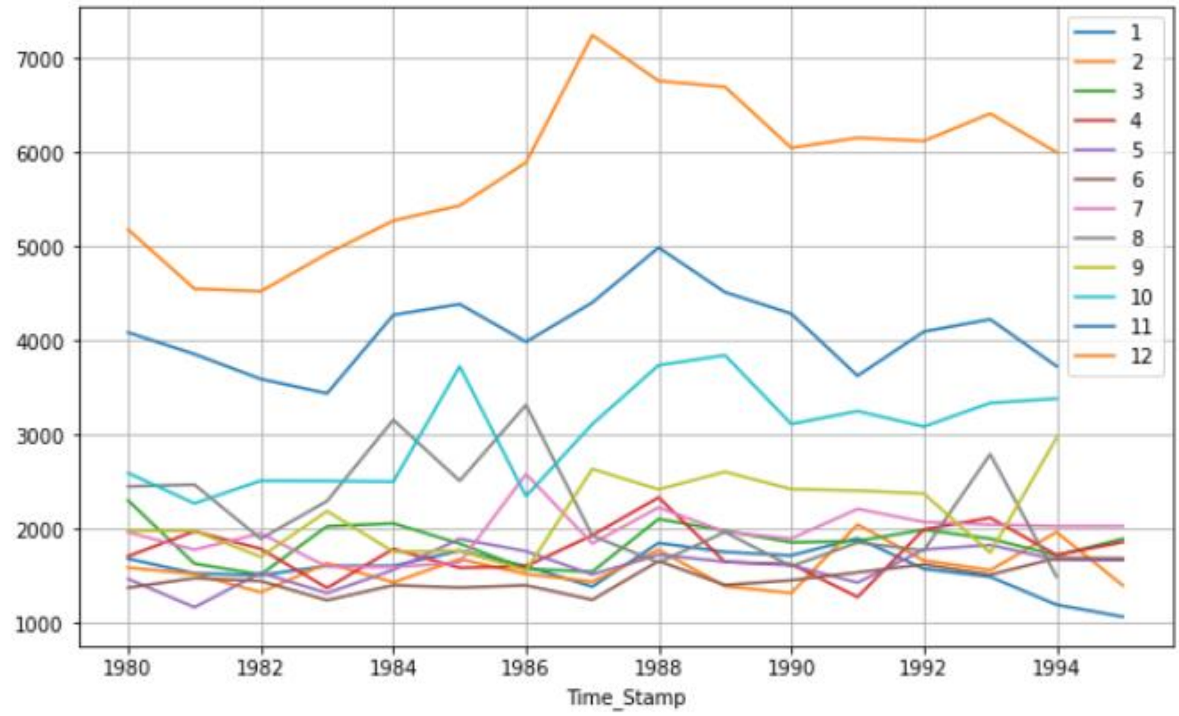


Figure 6 –Sparkling Wine Monthly Sales across Years

There is a rise in sales figures post 1986 for the months of September, October, November and December but post 1988 it had lowered again. However, for July and August we can observe that it was the other way and the sales figures were lower from 1986 onwards.

December has the highest sales across all years followed by November.

Empirical Cumulative Distribution:

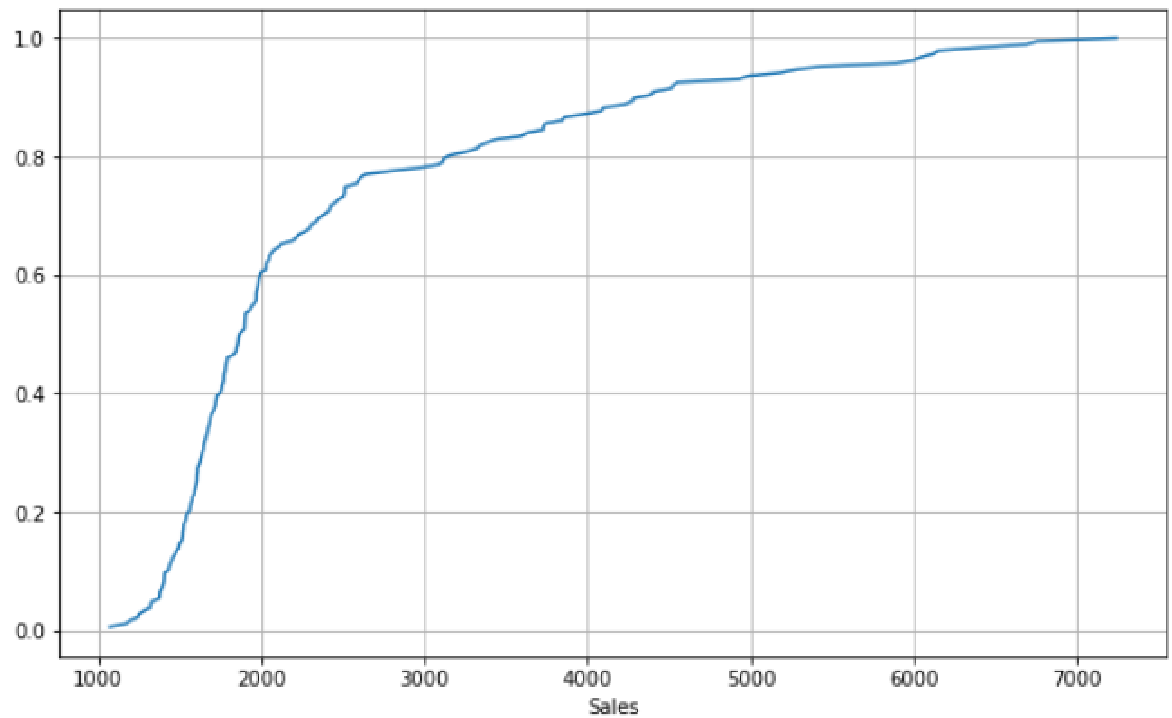


Figure 7 –Sparkling Wine Sales Empirical Cumulative Distribution Plot

Average Sales and Percentage Change of Sales:

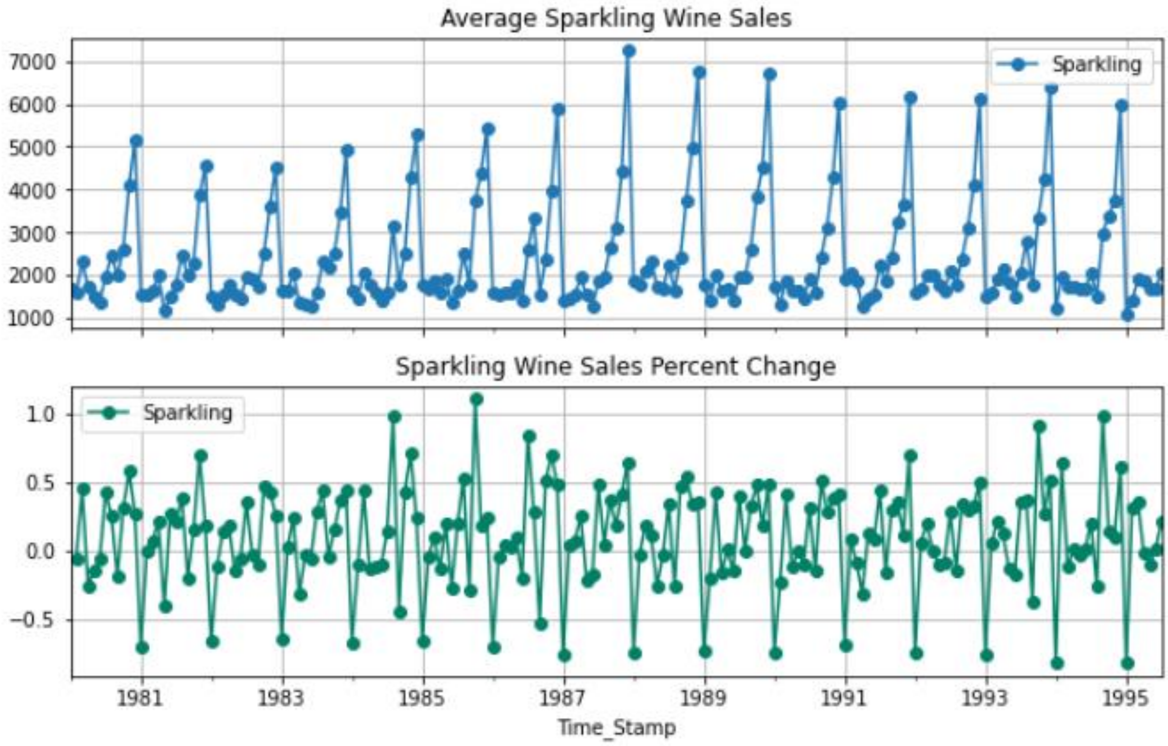


Figure 8 –Average Sales and Percentage Change of Sales

Sum of Sales of each year in Plot:

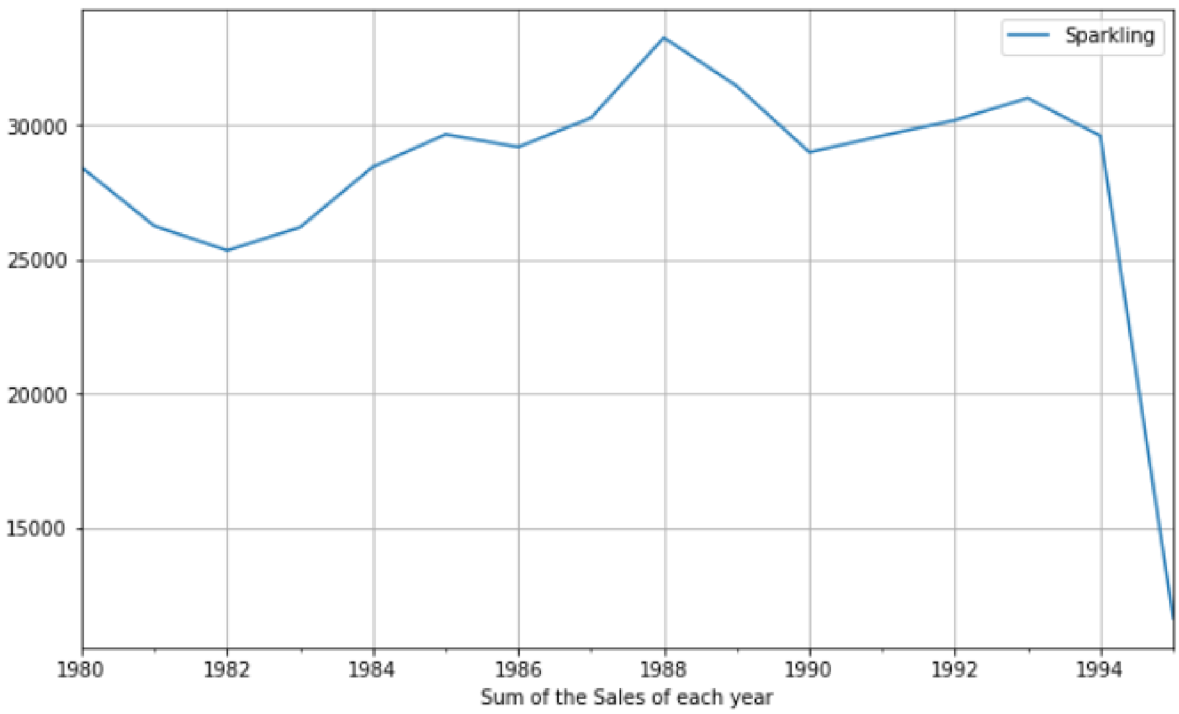


Figure 9 –Sum of Sparkling Wine Sales for each Year

Sum of Sales of each year in Figures:

	Sparkling
Time_Stamp	
1980-12-31	28406
1981-12-31	26227
1982-12-31	25321
1983-12-31	26180
1984-12-31	28431
1985-12-31	29640
1986-12-31	29170
1987-12-31	30258
1988-12-31	33246
1989-12-31	31443
1990-12-31	28977
1991-12-31	29587
1992-12-31	30171
1993-12-31	30991
1994-12-31	29584
1995-12-31	11620

The highest sales were recorded in the year 1988 with total 33246 bottles sold and the least sales were for 1995 with total 11620 bottles sold within the first seven months. On a 12 monthly basis, the least sales can be seen for the year 1982 with 25321 bottles sold in total.

Mean of Sales of each year in Plot:

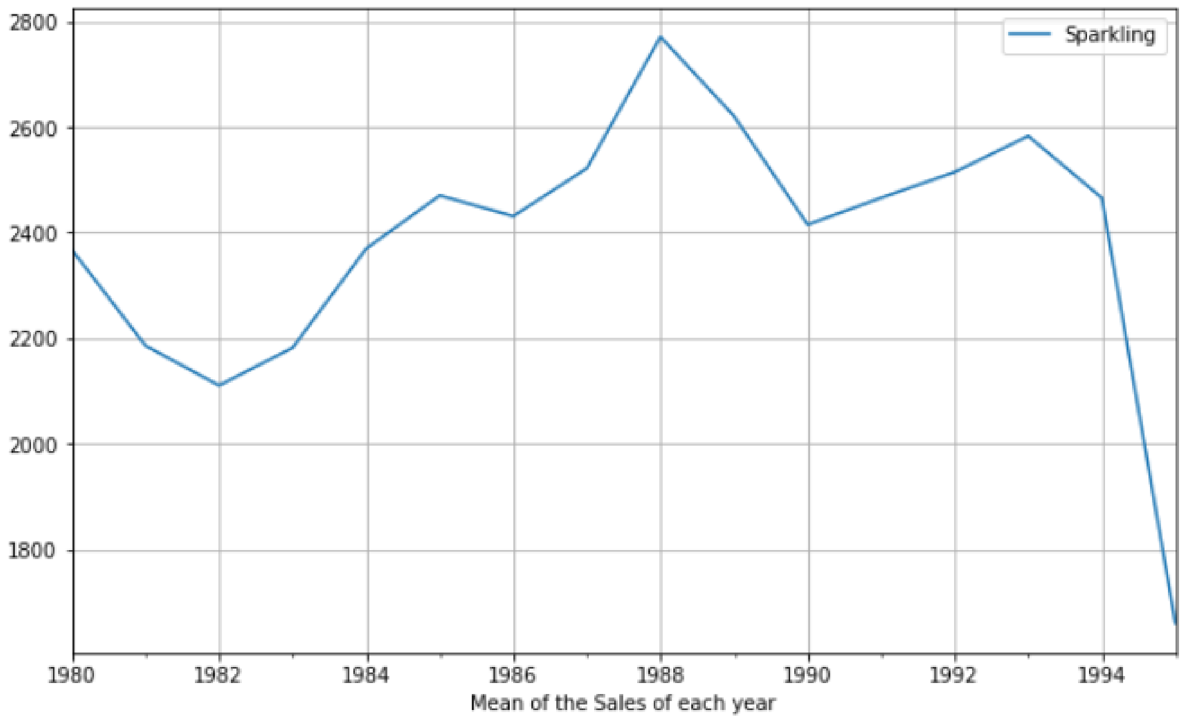


Figure 10 –Mean of Sparkling Wine Sales for each Year

Mean of Sales of each year in Figures:

Time_Stamp	Sparkling
1980-12-31	2367.166667
1981-12-31	2185.583333
1982-12-31	2110.083333
1983-12-31	2181.666667
1984-12-31	2369.250000
1985-12-31	2470.000000
1986-12-31	2430.833333
1987-12-31	2521.500000
1988-12-31	2770.500000
1989-12-31	2620.250000
1990-12-31	2414.750000
1991-12-31	2465.583333
1992-12-31	2514.250000
1993-12-31	2582.583333
1994-12-31	2465.333333
1995-12-31	1660.000000

The mean sales ranged between 2100 and 2800 from 1980 till 1994. With lowest in 1982 and highest in 1988 similar to earlier observation for sum of monthly sales. Lowest mean sales were for 1995

Decomposing the Time Series: Additive Method

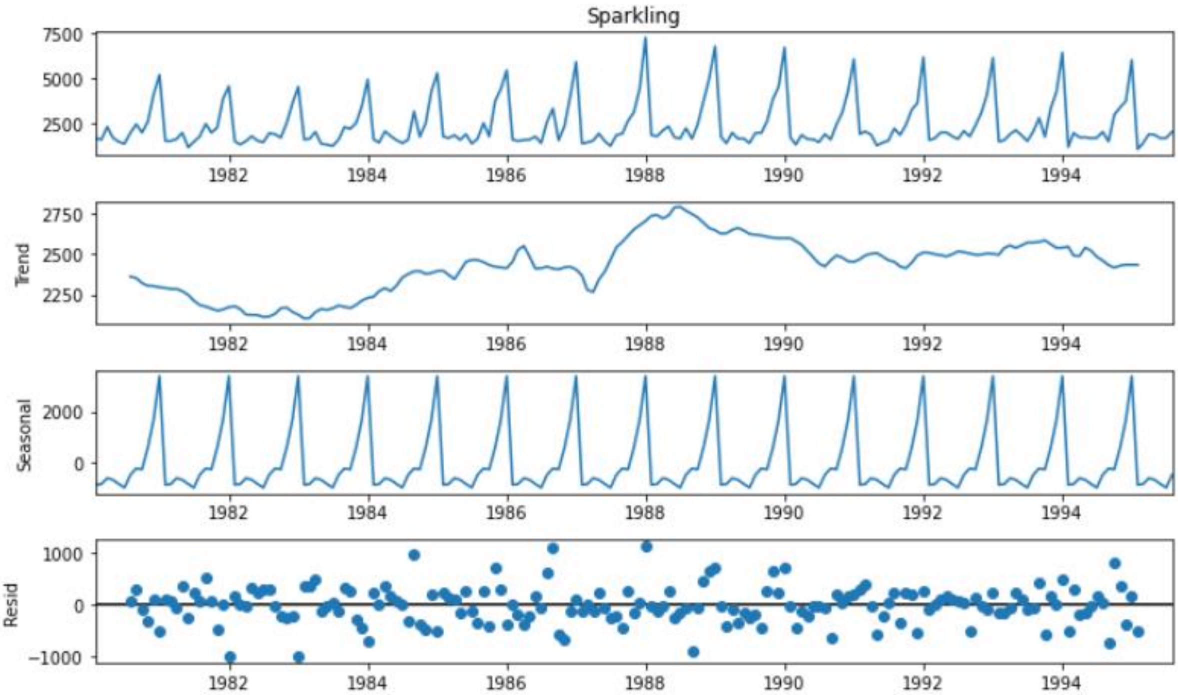


Figure 11 – Additive Decomposition

As per the 'additive' decomposition, we see that there is a pronounced trend until 1991. There is a seasonality as well. A lot of residuals are located around 0 from the plot of the residuals in the decomposition.

Trend, Seasonality and Residual:

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

```
Seasonality
 Time_Stamp
1980-01-31    -854.260599
1980-02-29    -830.350678
1980-03-31    -592.356630
1980-04-30    -658.490559
1980-05-31    -824.416154
1980-06-30    -967.434011
1980-07-31    -465.502265
1980-08-31    -214.332821
1980-09-30    -254.677265
1980-10-31     599.769957
1980-11-30    1675.067179
1980-12-31    3386.983846
Name: seasonal, dtype: float64
```

```
Residual
 Time_Stamp
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31     70.835599
1980-08-31    315.999487
1980-09-30    -81.864401
1980-10-31   -307.353290
1980-11-30    109.891154
1980-12-31   -501.775513
Name: resid, dtype: float64
```

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

```
Training Data Shape  (132, 1)
Testing Data Shape  (55, 1)
```

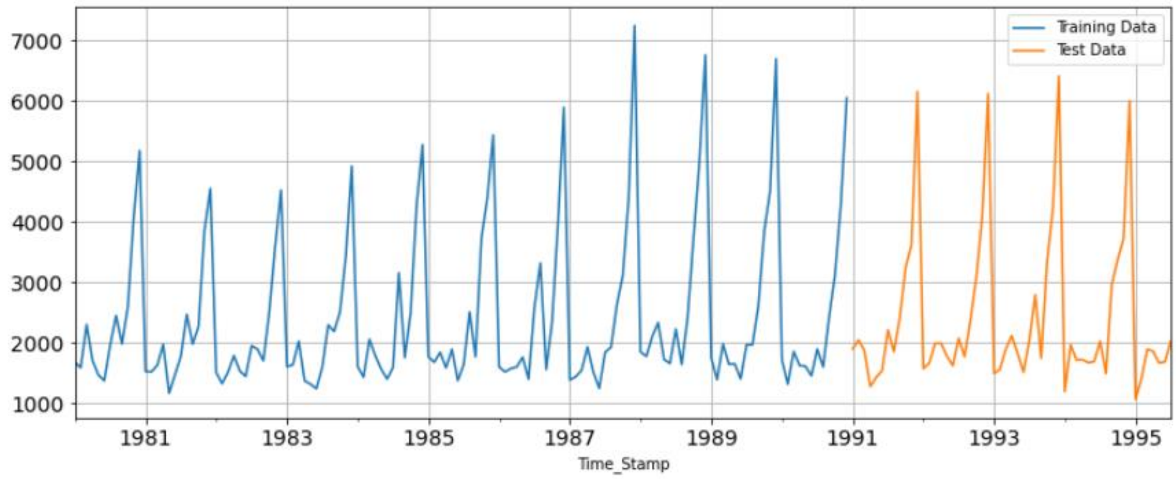


Figure 12 – Train Test Dataset

Training and Test Time Instances:

```
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 see that we have successfully the generated the numerical time instance order for both the training and test set.

1.3 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.

MODEL 1 - Linear Regression Model:

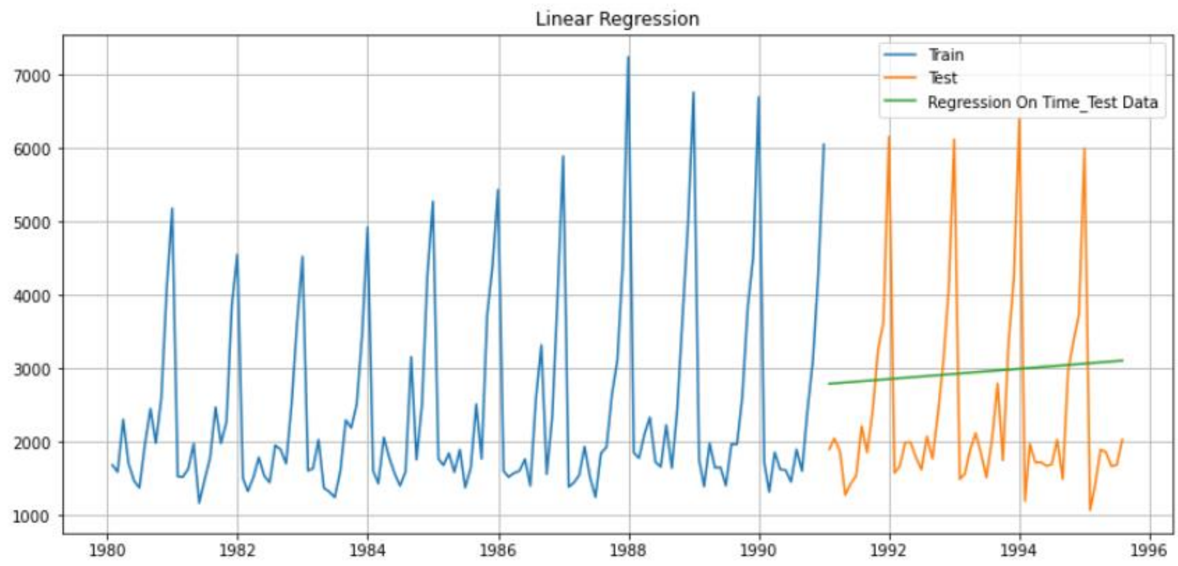


Figure 13 – Linear Regression Plot

	Test RMSE
RegressionOnTime	1389.135175

MODEL 2 - Naïve Model:

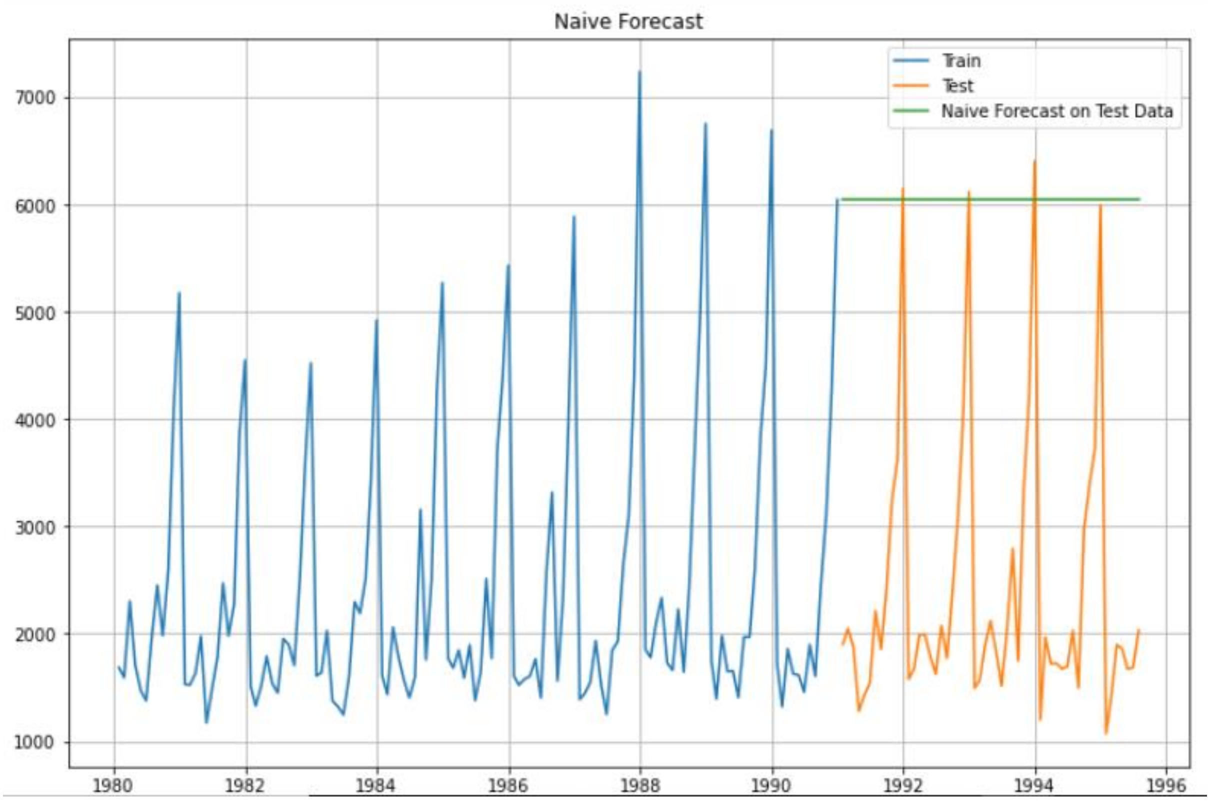


Figure 14 – Naïve Forecast Plot

	Test RMSE
NaiveModel	3864.279352

MODEL 3 - Simple Average Model:

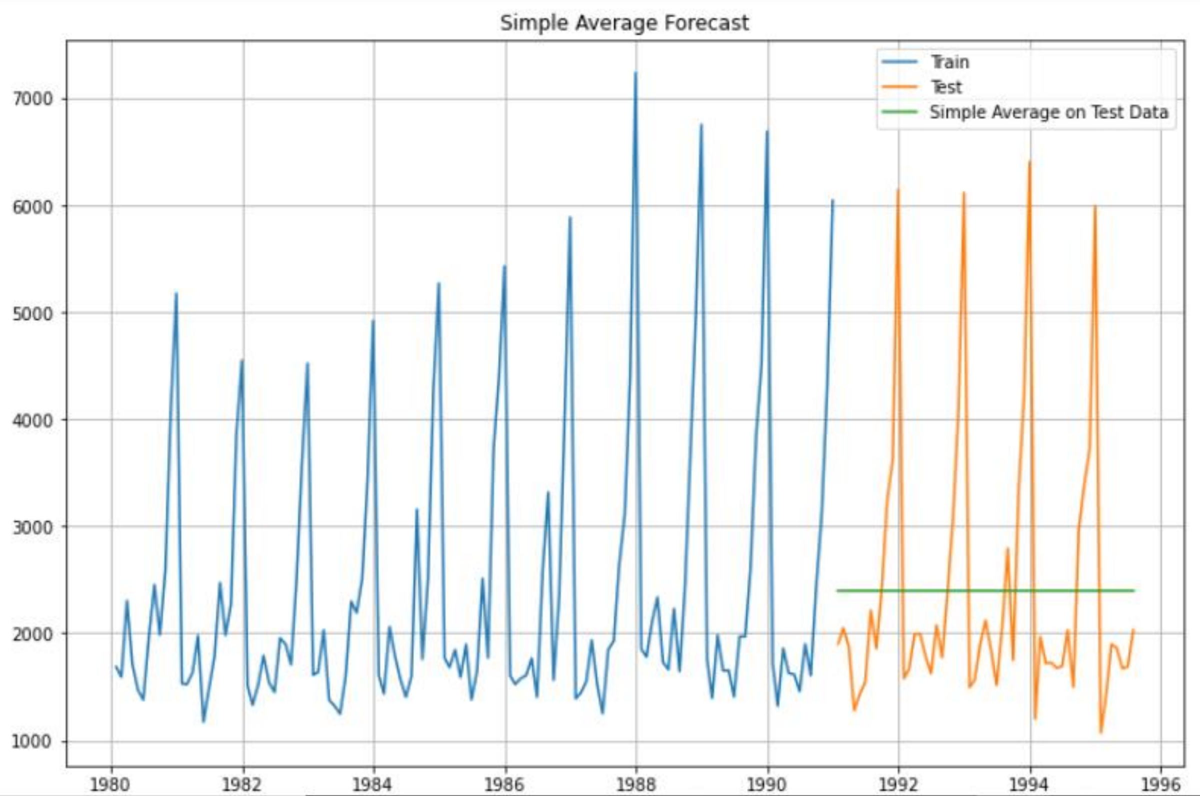


Figure 15 – Simple Average Forecast

	Test RMSE
SimpleAverageModel	1275.081804

MODEL 4 - Moving Average Model:

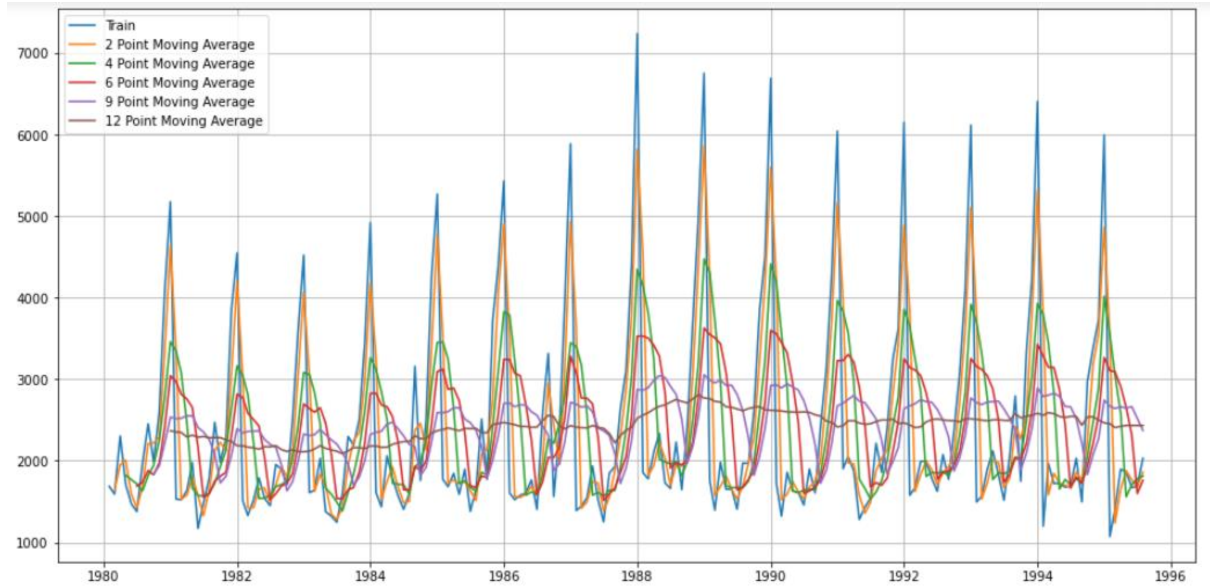


Figure 16 – Point-Wise Moving Average Forecast

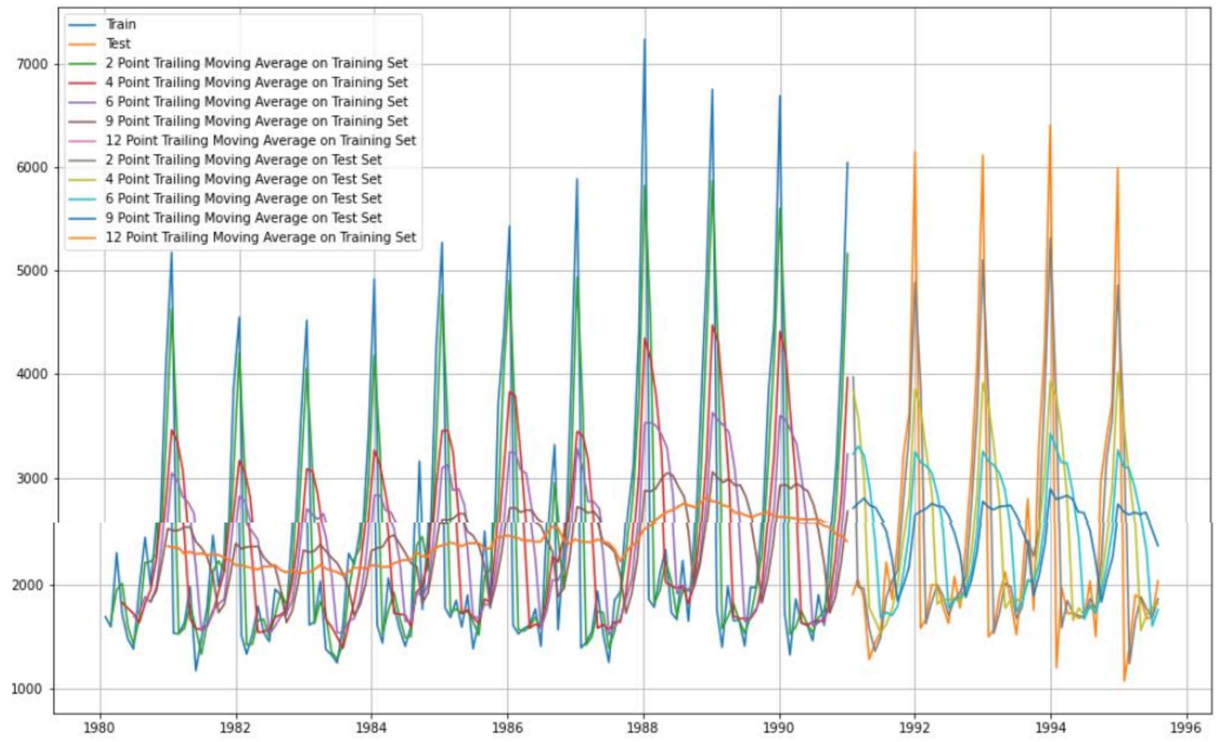


Figure 17 – Point-Wise Trailing Moving Average Forecast

	Test RMSE
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
12pointTrailingMovingAverage	1267.925330

Plotting of all the models:

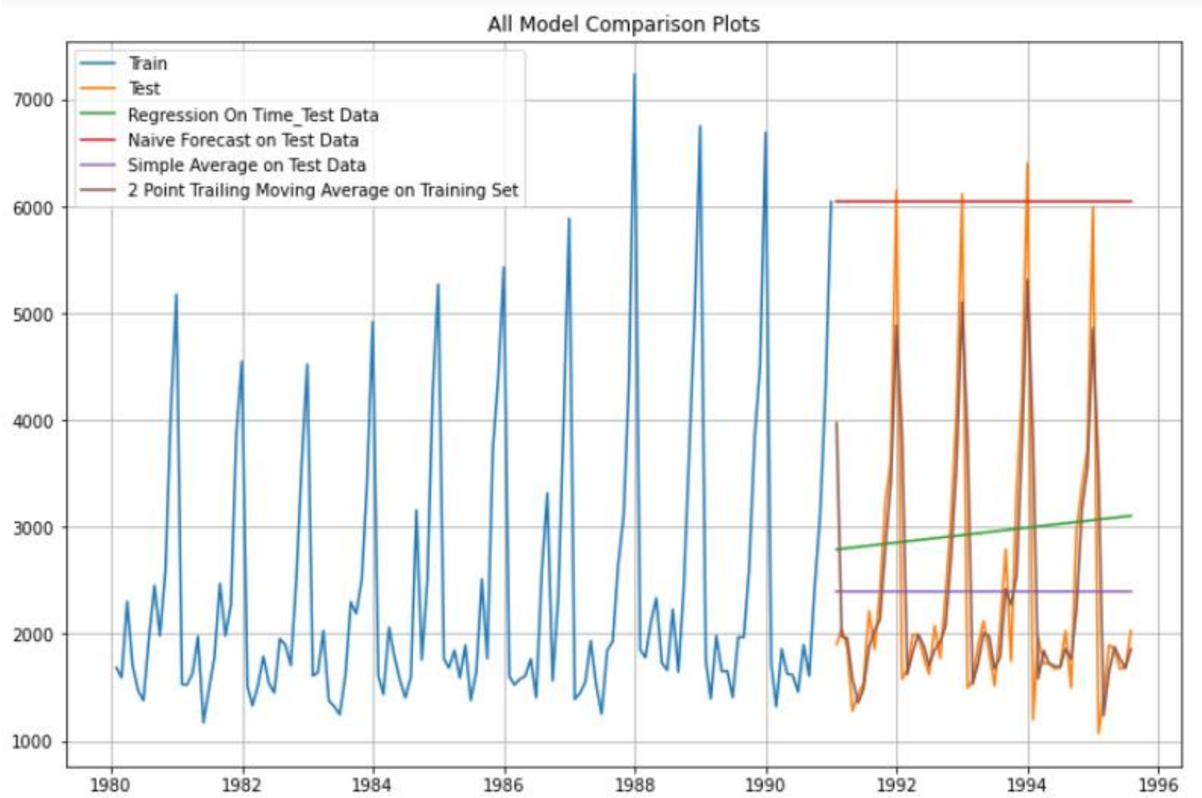


Figure 18 – All Model Comparison Plots

MODEL 5 - Simple Exponential Smoothing Model:

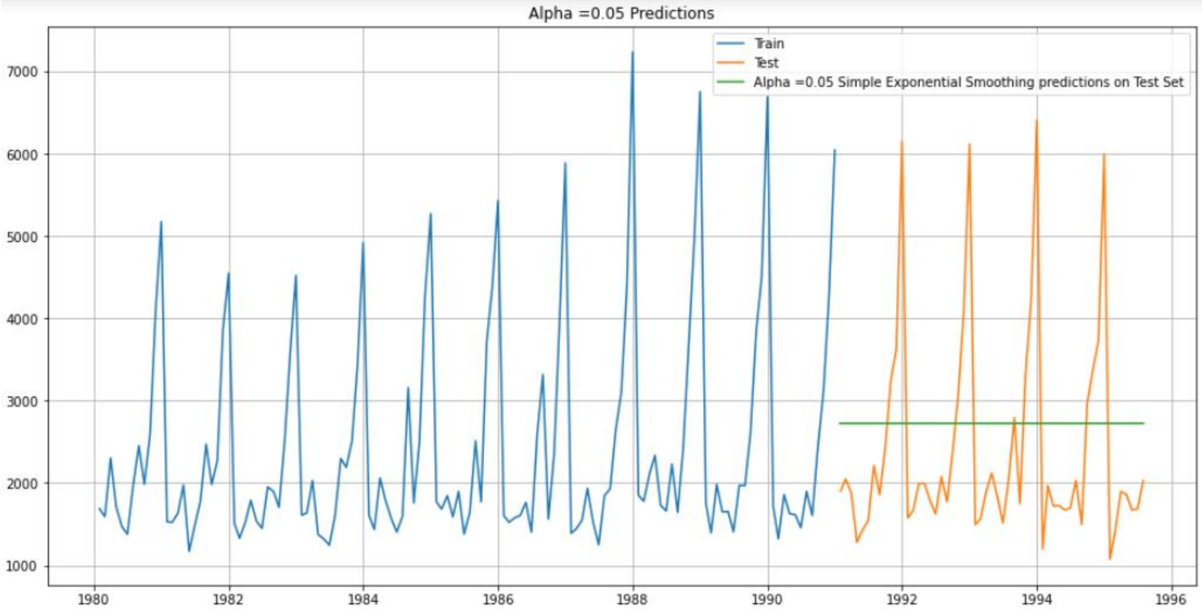


Figure 19 – Simple Exponential Smoothing Model

	Test RMSE
Alpha=0.05,SimpleExponentialSmoothing	1316.034674

MODEL 6 - Double Exponential Smoothing:

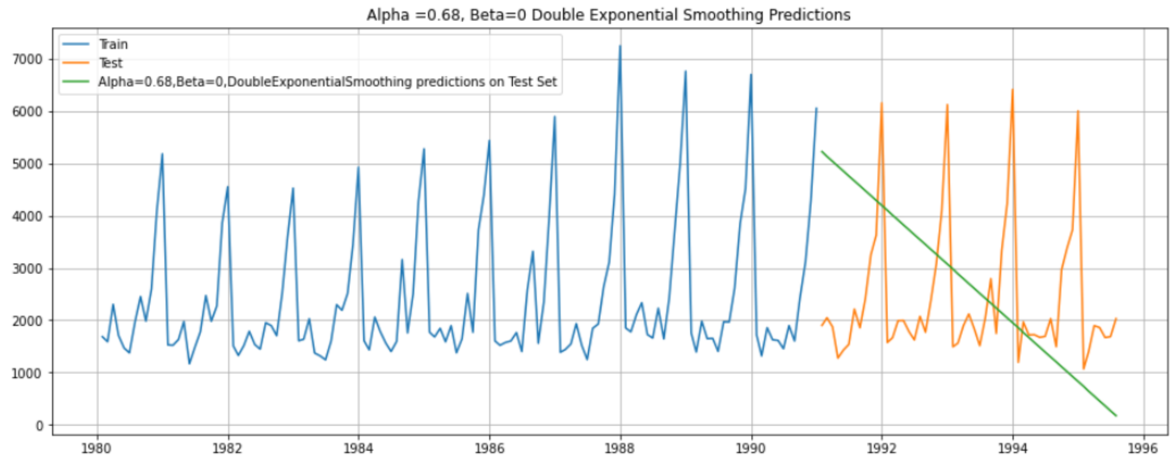


Figure 20 – Double Exponential Smoothing Model

	Test RMSE
For Alpha =0.68, Beta = 0 DoubleExponentialSmoothing	2007.238526

MODEL 7 - Triple Exponential Smoothing Additive:

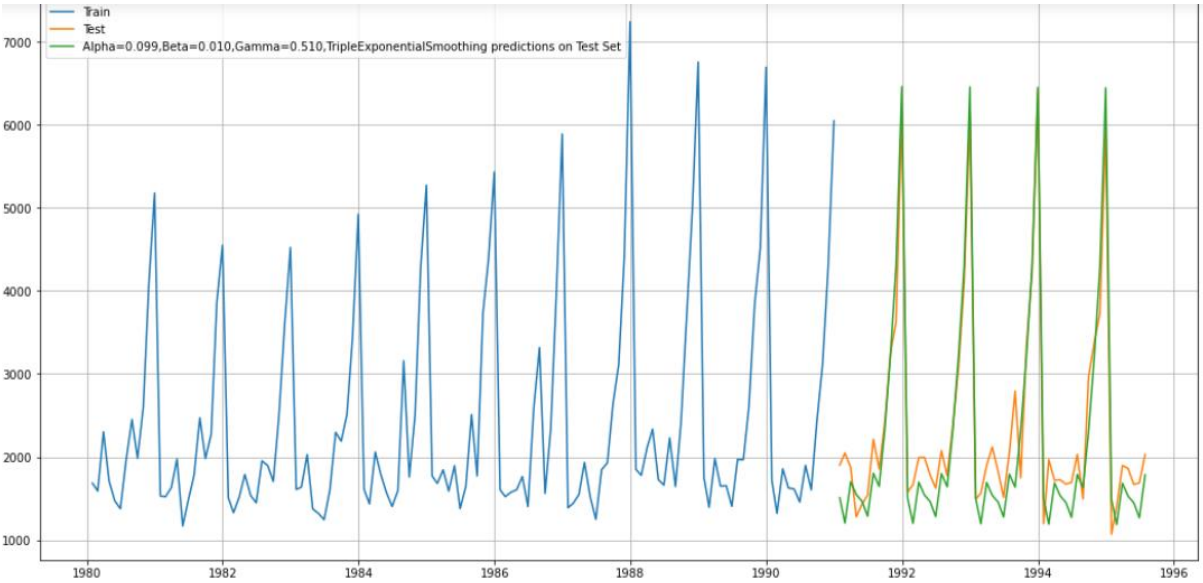


Figure 21 – Triple Exponential Smoothing Additive Model

	Test RMSE
Alpha=0.099,Beta=0.010,Gamma=0.510,TripleExponentialSmoothing	379.981727

MODEL 8 - Triple Exponential Smoothing Multiplicative:

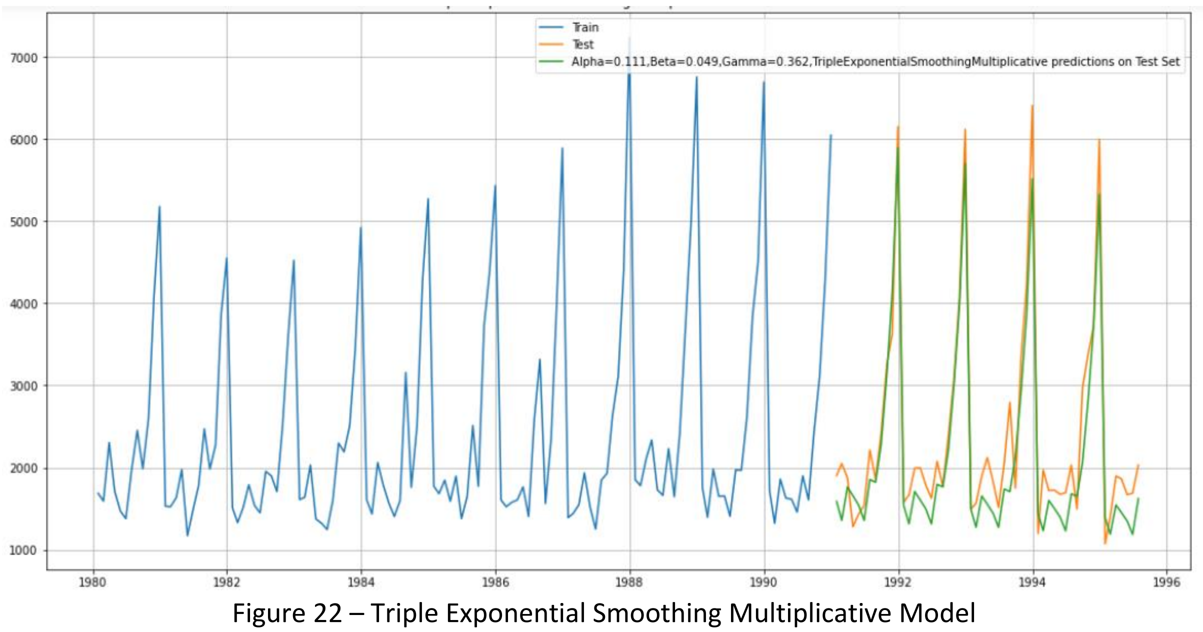


Figure 22 – Triple Exponential Smoothing Multiplicative Model

	Test RMSE
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothingMultiplicative	403.319631

From the observations so far, we can clearly see that exponential smoothing techniques/models have performed better than the other models as they have a lower RMSE on the test data. The Triple Exponential Smoothing Additive Model has performed the best due to its lower RMSE of 379.98.

1.4 Build all the 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.

Hypothesis for Statistical Test:

Null Hypothesis - H0 = Time Series is not Stationary
Alternative Hypothesis - HA = Time Series is Stationary

Stationarity Check Using Dickey-Fuller Test:

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.360497
p-value                          0.601061
#Lags Used                      11.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

We observe the Time Series is non-stationary for alpha = 0.05 as the p-value is > alpha at 0.60 . Hence, we fail to reject the null hypothesis.

Let us take a difference of order 1 and check whether the Time Series is stationary or not.

Stationarity Check Using Dickey-Fuller Test by taking difference of Order 1:

```
Results of Dickey-Fuller Test:
Test Statistic                 -23.500036
p-value                          0.000000
#Lags Used                      10.000000
Number of Observations Used    175.000000
Critical Value (1%)             -3.468280
Critical Value (5%)             -2.878202
Critical Value (10%)            -2.575653
dtype: float64
```

We observe the Time Series is now stationary for alpha = 0.05 as the p-value at less than alpha.

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

MODEL 9 - ARIMA Model:

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
```

Above is some combination of different parameters of p and q in the range of 0 and 2

ARIMA AIC SCORES for Parameters in range of 0 & 2:

	param	AIC
8	(2, 1, 2)	2213.509213
7	(2, 1, 1)	2233.777626
2	(0, 1, 2)	2234.408323
5	(1, 1, 2)	2234.527200
4	(1, 1, 1)	2235.755095
6	(2, 1, 0)	2260.365744
1	(0, 1, 1)	2263.060016
3	(1, 1, 0)	2266.608539
0	(0, 1, 0)	2267.663036

We can see the best AIC is for ARIMA (2,1,2) of 2213.50. Below, we will built our ARIMA Model for this parameter and check the performance.

ARIMA (2,1,2) Model Results:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Sun, 27 Feb 2022   AIC                           2213.509
Time:                        18:41:16   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     -7.740      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
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.15e+29. Standard errors may be unstable.
```

Figure 23 – Arima Model

	Test RMSE
ARIMA(2,1,2)	1299.979832

MODEL 10 - SARIMA Model with seasonality 6:

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
```

Above is some combination of different parameters of p and q in the range of 0 and 2

SARIMA AIC SCORES for Parameters in range of 0 & 2:

	param	seasonal	AIC
53	(1, 1, 2)	(2, 0, 2, 6)	1727.670866
26	(0, 1, 2)	(2, 0, 2, 6)	1727.888818
80	(2, 1, 2)	(2, 0, 2, 6)	1729.192582
17	(0, 1, 1)	(2, 0, 2, 6)	1741.641478
44	(1, 1, 1)	(2, 0, 2, 6)	1743.379778

We can see the best AIC is for SARIMA (1,1,2) (2,0,2,6) of 1727.67. Below we will built our SARIMA Model for this parameter and check the performance.

SARIMA (1,1,2) (2,0,2,6) Model Results:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                y   No. Observations:                    132
Model:             SARIMAX(1, 1, 2)x(2, 0, 2, 6)   Log Likelihood                  -855.835
Date:                           Sun, 27 Feb 2022   AIC                             1727.671
Time:                                   19:21:10   BIC                             1749.700
Sample:                                        0   HQIC                            1736.613
                                           - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6451      0.286     -2.256      0.024      -1.206      -0.085
ma.L1         -0.3355      0.227     -1.475      0.140      -0.781       0.110
ma.L2         -0.8805      0.277     -3.180      0.001      -1.423      -0.338
ar.S.L6       -0.0045      0.027     -0.165      0.869      -0.057       0.049
ar.S.L12       1.0361      0.018     56.096      0.000       1.000       1.072
ma.S.L6        0.0675      0.152      0.444      0.657      -0.231       0.366
ma.S.L12      -0.6125      0.093     -6.592      0.000      -0.795      -0.430
sigma2      1.153e+05   1.79e+04      6.456      0.000    8.03e+04     1.5e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                25.26
Prob(Q):                              0.77   Prob(JB):                         0.00
Heteroskedasticity (H):               2.63   Skew:                             0.47
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.09
===================================================================================

Warnings:
```

Figure 24 – Sarima Model with seasonality 6

Summary Frame for Alpha = 0.05:

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1330.347607	380.569348	584.445390	2076.249823
1	1177.284748	392.119860	408.743945	1945.825551
2	1625.868709	392.314443	856.946530	2394.790887
3	1546.370547	397.718345	766.856914	2325.884179
4	1308.633296	398.937917	526.729347	2090.537244

	RMSE
SARIMA(1, 1, 2)(2,0,2,6)	626.898233

MODEL 11 - SARIMA Model with seasonality 12:

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

Above is some combination of different parameters of p and q in the range of 0 and 2

SARIMA AIC SCORES for Parameters in range of 0 & 2:

	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.584247
53	(1, 1, 2)	(2, 0, 2, 12)	1555.929659
26	(0, 1, 2)	(2, 0, 2, 12)	1557.121564
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160507
77	(2, 1, 2)	(1, 0, 2, 12)	1557.340402

We can see the best AIC is for SARIMA (1,1,2) (1,0,2,12) of 1555.58. Below we will built our SARIMA Model for this parameter and check the performance.

SARIMA (1,1,2) (1,0,2,12) Model Results:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                            Sun, 27 Feb 2022   AIC                           1555.584
Time:                                    19:43:20   BIC                           1574.095
Sample:                                         0   HQIC                          1563.083
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6282      0.255     -2.463      0.014      -1.128      -0.128
ma.L1         -0.1041      0.225     -0.463      0.643      -0.545       0.337
ma.L2         -0.7276      0.154     -4.734      0.000      -1.029      -0.426
ar.S.L12       1.0439      0.014     72.840      0.000       1.016       1.072
ma.S.L12      -0.5550      0.098     -5.663      0.000      -0.747      -0.363
ma.S.L24      -0.1354      0.120     -1.133      0.257      -0.370       0.099
sigma2      1.506e+05   2.03e+04      7.401      0.000    1.11e+05     1.9e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                11.72
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               1.47   Skew:                             0.36
Prob(H) (two-sided):                  0.26   Kurtosis:                         4.48
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```
Figure 25 – Sarima Model with Seasonality 12

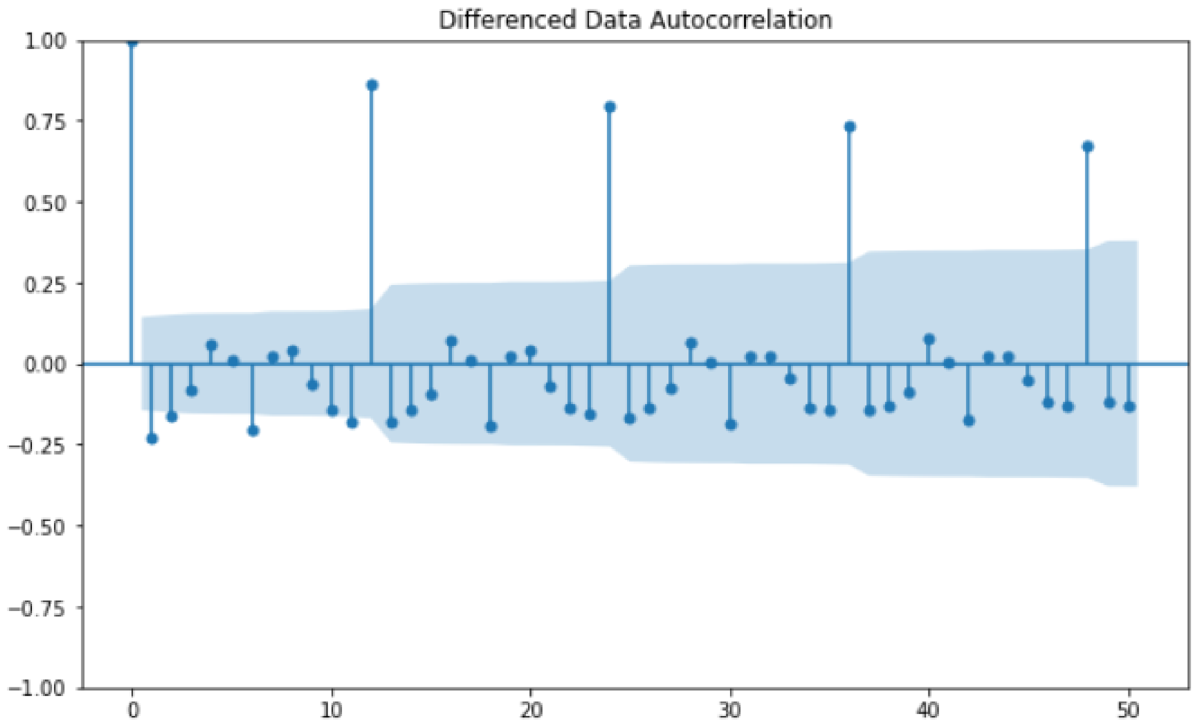
Summary Frame for Alpha = 0.05:

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1327.386418	388.344800	566.244597	2088.528239
1	1315.110768	402.007729	527.190097	2103.031440
2	1621.588857	402.001336	833.680717	2409.496997
3	1598.867465	407.239037	800.693619	2397.041311
4	1392.688227	407.969106	593.083472	2192.292982

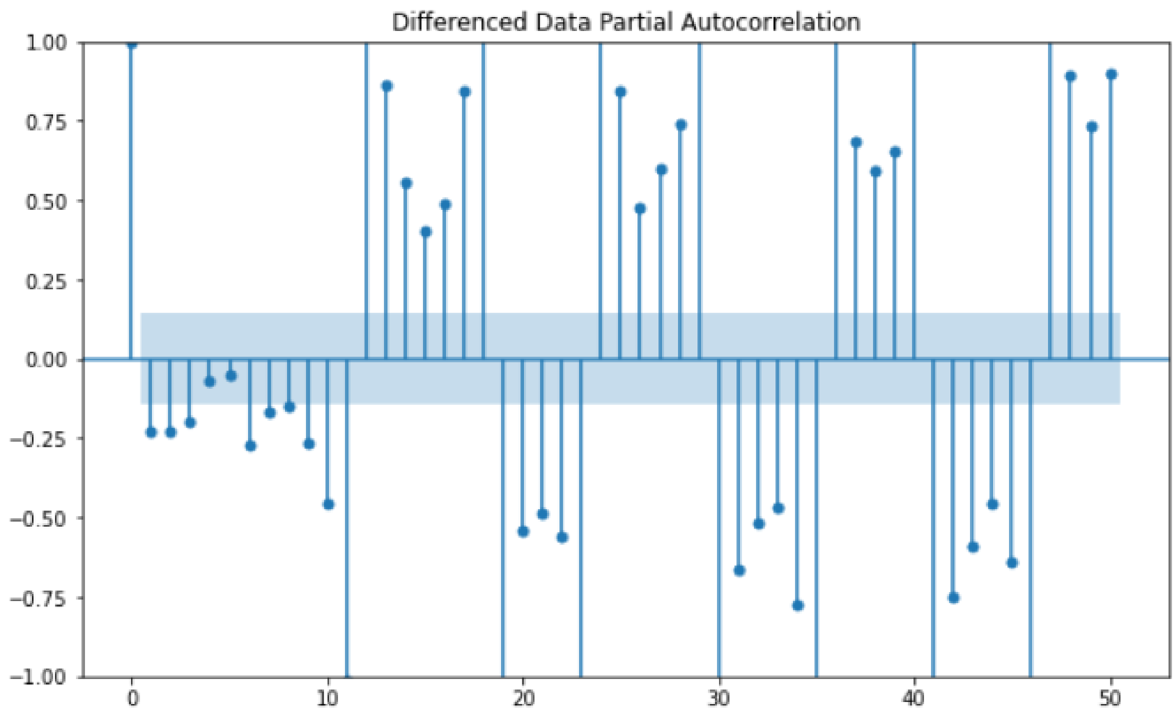
	RMSE
SARIMA(1, 1, 2)(1, 0, 2, 12)	528.621309

1.6 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.

ACF Plot:



PACF Plot:



Here, we have taken alpha=0.05.

* The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 3.

* The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2.

MODEL 12 - Manual ARIMA Model based on cut-0ff point (3,1,2):

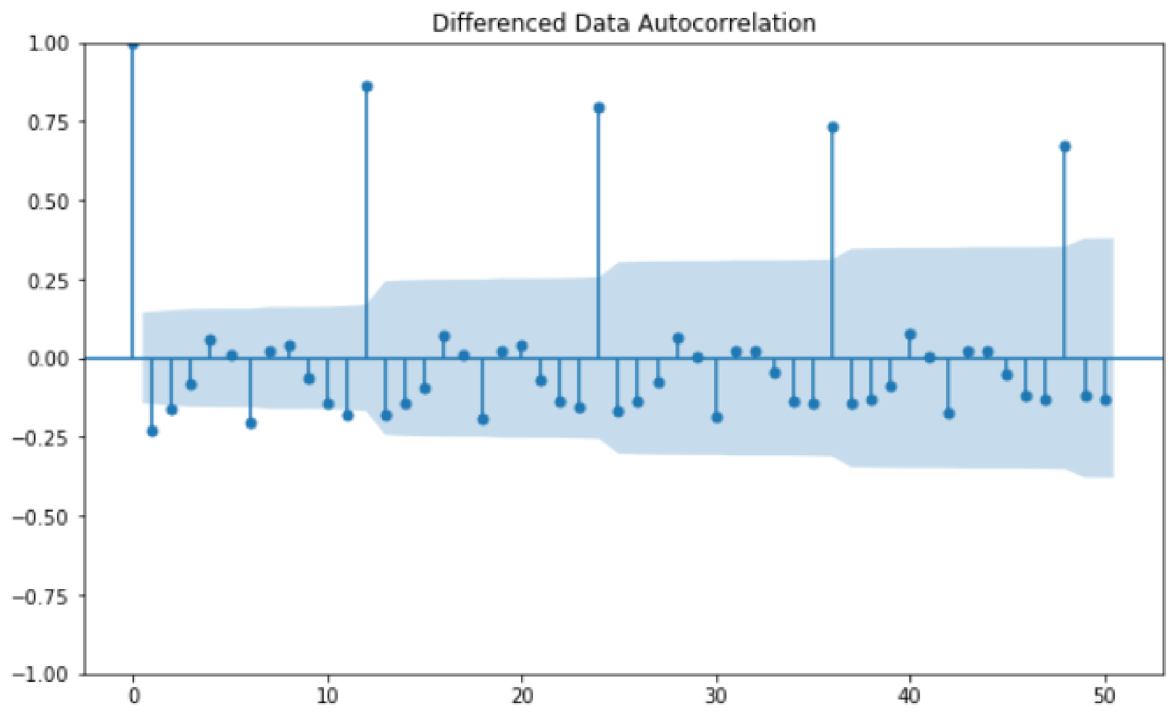
```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(3, 1, 2)   Log Likelihood               -1109.378
Date:                Sun, 06 Mar 2022   AIC                           2230.756
Time:                        15:36:14   BIC                           2248.007
Sample:                    01-31-1980   HQIC                          2237.766
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4323      0.044     -9.794      0.000      -0.519      -0.346
ar.L2          0.3303      0.109      3.021      0.003       0.116       0.545
ar.L3         -0.2374      0.065     -3.639      0.000      -0.365      -0.110
ma.L1          0.0175      0.128      0.136      0.892      -0.234       0.268
ma.L2         -0.9823      0.136     -7.245      0.000      -1.248      -0.717
sigma2      1.273e+06   1.95e-07   6.54e+12      0.000    1.27e+06    1.27e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                 4.63
Prob(Q):                              0.88   Prob(JB):                         0.10
Heteroskedasticity (H):               2.72   Skew:                             0.37
Prob(H) (two-sided):                  0.00   Kurtosis:                         3.55
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 3.76e+27. Standard errors may be unstable.
```

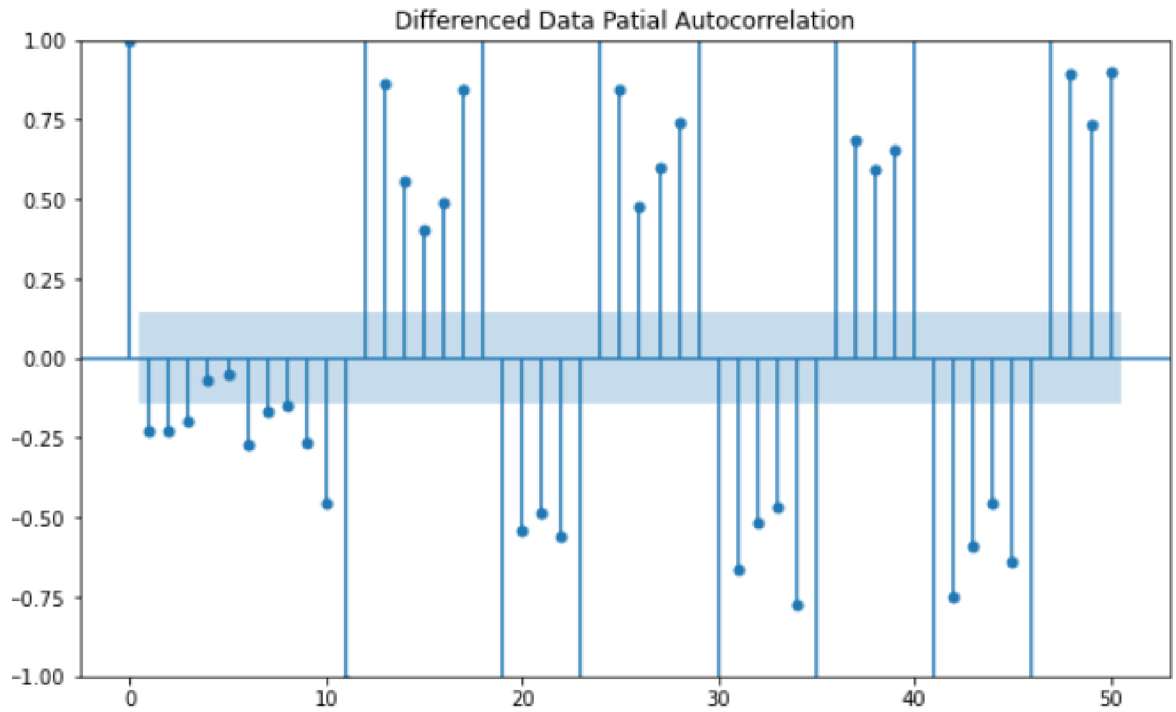
Figure 26 – Arima Model based on ACF and PACF cut-off points

	Test RMSE
ARIMA(3,1,2)	1281.482077

MODEL – 13 Manual SARIMA Model based on cut-0ff point (3,1,2):

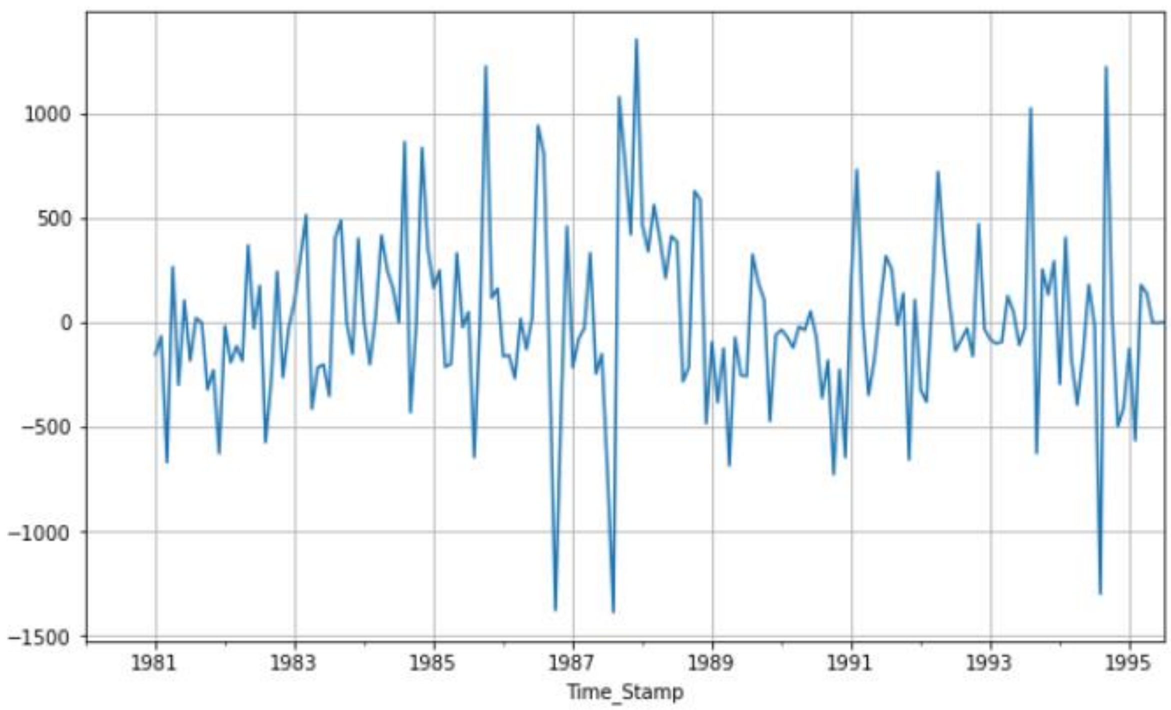
ACF Plot:



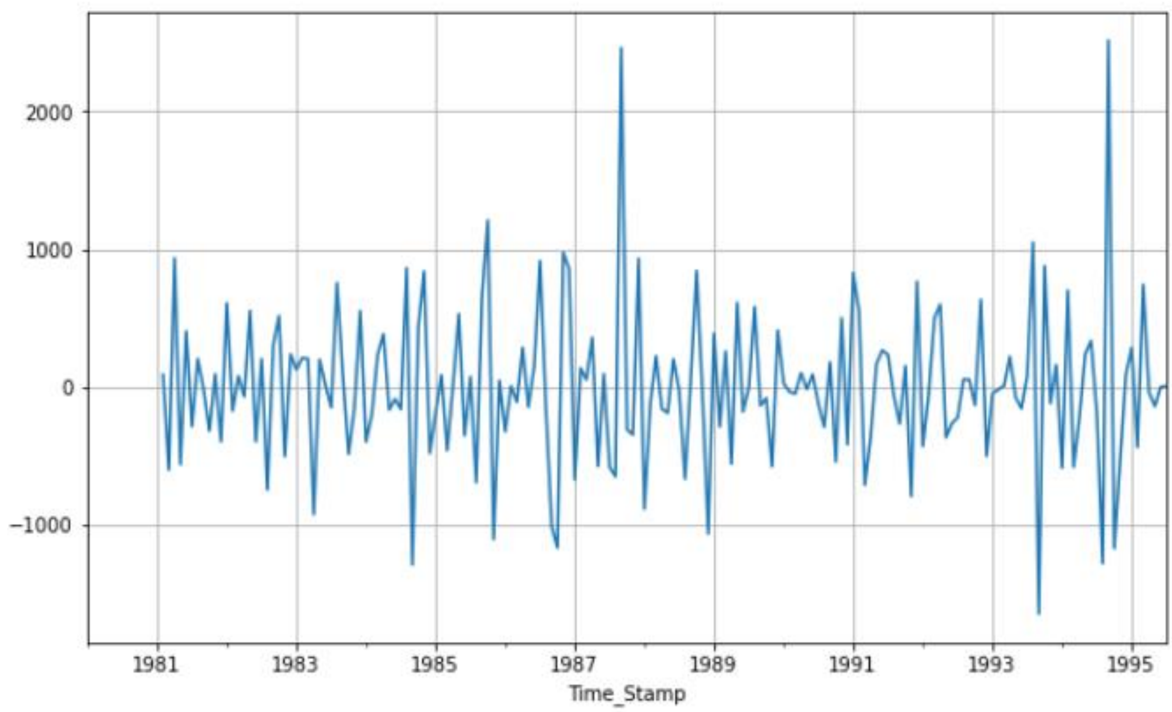


We see that there can be a seasonality of 12. We will run our auto SARIMA models by setting seasonality for 12.

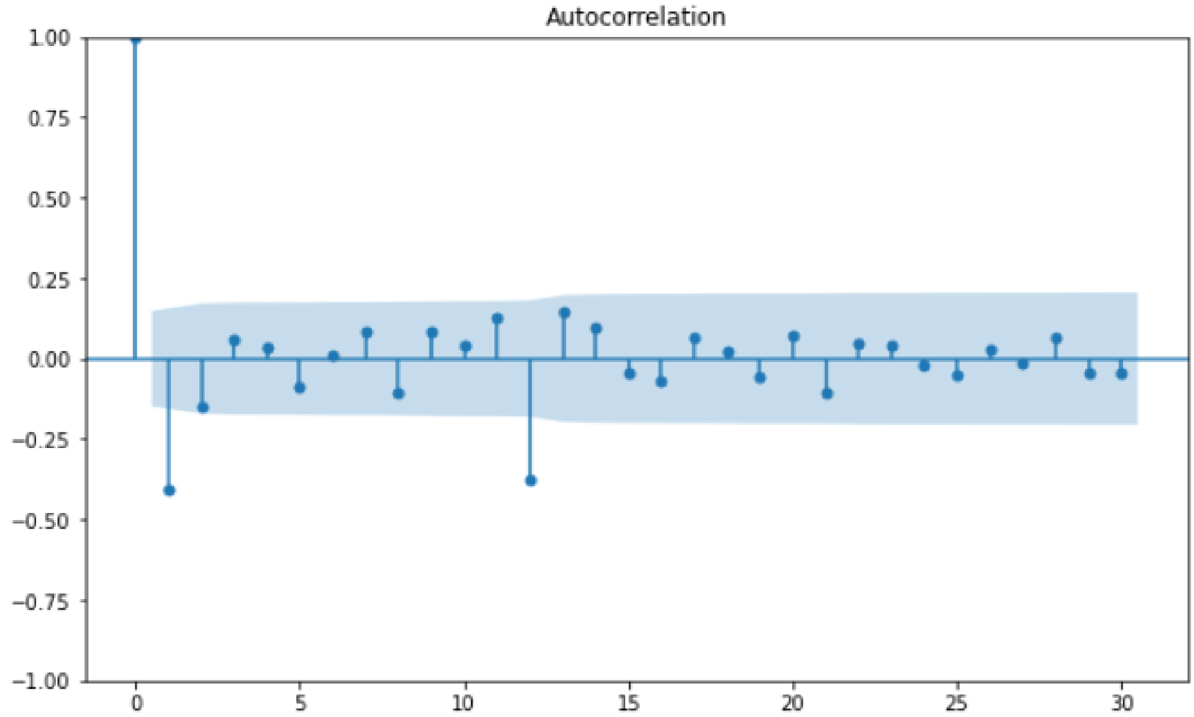
Plot with Seasonal Difference of 12:



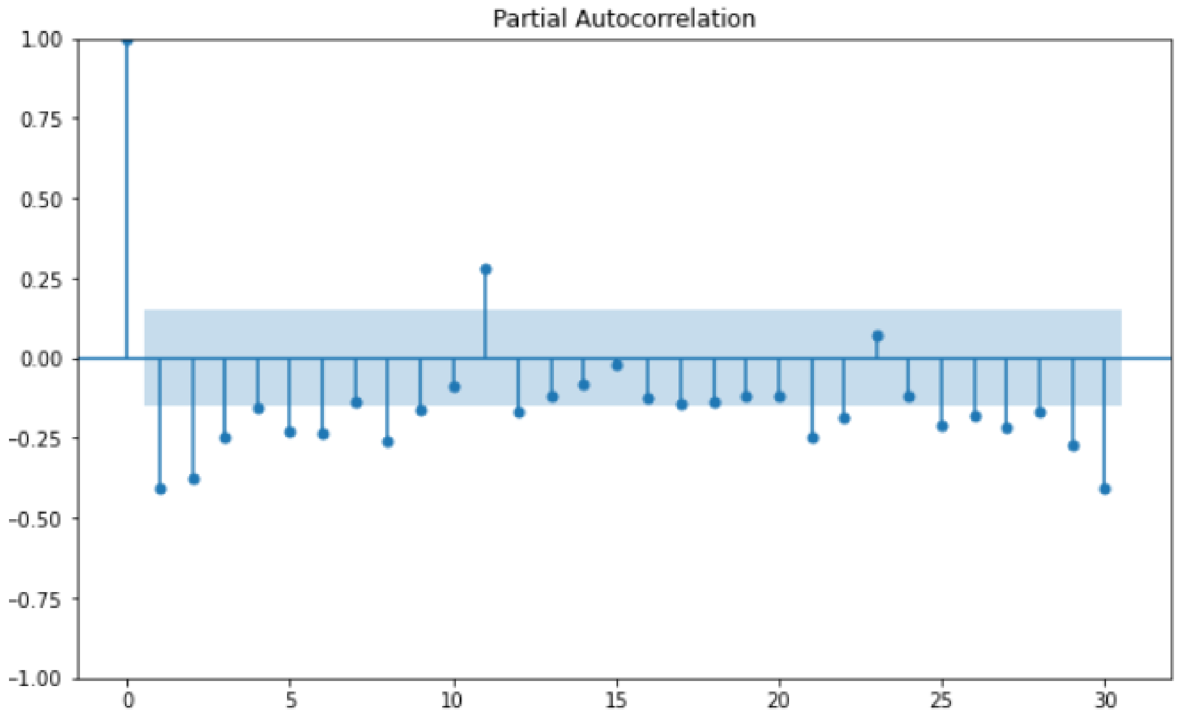
Plot without Trend and Only Seasonality:



ACF plot for the new modified Time Series with Stationarity:



PACF plot for the new modified Time Series with Stationarity:



Manual SARIMA 3,1,2) (6,1,1,12) Model Results:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                    y   No. Observations:                  132
Model:             SARIMAX(3, 1, 2)x(6, 1, [1], 12)   Log Likelihood                -323.675
Date:                              Sun, 06 Mar 2022   AIC                            673.349
Time:                                      15:48:00   BIC                            696.543
Sample:                                           0   HQIC                           681.951
                                              - 132
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5243      0.236     -2.217      0.027      -0.988      -0.061
ar.L2          0.3109      0.422      0.736      0.462      -0.517       1.139
ar.L3          0.3246      0.224      1.451      0.147      -0.114       0.763
ma.L1      -3.709e-05    363.116  -1.02e-07      1.000    -711.694     711.694
ma.L2         -1.0000     84.416     -0.012      0.991    -166.452     164.452
ar.S.L12      -0.8795      0.206     -4.260      0.000      -1.284      -0.475
ar.S.L24      -0.3472      0.218     -1.592      0.111      -0.775       0.080
ar.S.L36      -0.1852      0.173     -1.068      0.286      -0.525       0.155
ar.S.L48      -0.2829      0.260     -1.087      0.277      -0.793       0.227
ar.S.L60      -0.5926      0.336     -1.765      0.078      -1.251       0.065
ar.S.L72      -0.2040      0.273     -0.748      0.454      -0.739       0.331
ma.S.L12       0.9979     84.697      0.012      0.991    -165.005     167.001
sigma2      1.046e+05      0.001     9.8e+07      0.000    1.05e+05    1.05e+05
===================================================================================

Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                 3.11
Prob(Q):                              0.84   Prob(JB):                         0.21
Heteroskedasticity (H):               0.34   Skew:                             0.49
Prob(H) (two-sided):                  0.04   Kurtosis:                         3.85
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 4.82e+26. Standard errors may be unstable.
```

Figure 27 – Sarima Model with Seasonality 12 based on ACF and PACF cut-off points

Summary Frame for Alpha = 0.05:

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1416.434911	359.390446	712.042580	2120.827241
1	1566.580697	393.811151	794.725026	2338.436369
2	1728.591852	394.330094	955.719069	2501.464635
3	1657.547915	426.835267	820.966163	2494.129666
4	1670.469879	425.908252	835.705046	2505.234713

	Test RMSE
SARIMA(3,1,2)(6,1,1,12)	369.677413

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

MODELS	TEST RMSE
SARIMA(3,1,2)(6,1,1,12)	369.677413
Alpha=0.099,Beta=0.010,Gamma=0.510,TripleExponentialSmoothing	379.981727
Alpha=0.111,Beta=0.049,Gamma=0.362,TripleExponentialSmoothingMultiplicative	403.319631
SARIMA(1, 1, 2)(1, 0, 2, 12)	528.621309
SARIMA(1, 1, 2)(2,0,2,6)	626.898233
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
12pointTrailingMovingAverage	1267.92533
Simple Average	1275.081804
ARIMA(3,1,2)	1281.482077
6pointTrailingMovingAverage	1283.927428
ARIMA(2,1,2)	1299.979832
Alpha=0.05,SimpleExponentialSmoothing	1316.034674
9pointTrailingMovingAverage	1346.278315
RegressionOnTime	1389.135175
For Alpha =0.68, Beta = 0 DoubleExponentialSmoothing	2007.238526
Naive Model	3864.279352

Table 1 – All Models with RMSE

1.8 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.

The best model built is **SARIMA(3,1,2)(6,1,1,12)** with Test RMSE of **369.677.** Now we will built the best optimum full model on the same parameters

SARIMA (3,1,2) (6,1,1,12) Full Model Results:

```
                              SARIMAX Results
============================================================================================
Dep. Variable:                          Sparkling   No. Observations:                  187
Model:             SARIMAX(3, 1, 2)x(6, 1, [1], 12)   Log Likelihood                -728.128
Date:                            Sun, 06 Mar 2022   AIC                           1482.255
Time:                                    16:24:43   BIC                           1515.992
Sample:                                01-31-1980   HQIC                          1495.905
                                     - 07-31-1995
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.8419      0.138     -6.084      0.000      -1.113      -0.571
ar.L2          0.1370      0.180      0.760      0.447      -0.216       0.490
ar.L3          0.0813      0.129      0.630      0.528      -0.172       0.334
ma.L1          0.0251      0.122      0.206      0.837      -0.214       0.264
ma.L2         -0.9519      0.112     -8.472      0.000      -1.172      -0.732
ar.S.L12      -1.0134      0.208     -4.878      0.000      -1.421      -0.606
ar.S.L24      -0.6075      0.202     -3.005      0.003      -1.004      -0.211
ar.S.L36      -0.4308      0.177     -2.430      0.015      -0.778      -0.083
ar.S.L48      -0.2907      0.174     -1.667      0.096      -0.633       0.051
ar.S.L60      -0.2565      0.152     -1.690      0.091      -0.554       0.041
ar.S.L72      -0.2406      0.095     -2.529      0.011      -0.427      -0.054
ma.S.L12       0.5822      0.249      2.341      0.019       0.095       1.070
sigma2      1.396e+05   2.39e+04      5.838      0.000    9.28e+04    1.87e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 8.60
Prob(Q):                              0.98   Prob(JB):                         0.01
Heteroskedasticity (H):               0.60   Skew:                             0.45
Prob(H) (two-sided):                  0.15   Kurtosis:                         4.13
===================================================================================

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

Figure 28 – Sarima Full Model

Summary Frame for Alpha = 0.05:

Sparkling	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-31	1670.837412	373.926389	937.955158	2403.719667
1995-09-30	2585.710558	380.056472	1840.813560	3330.607556
1995-10-31	3274.133458	380.648957	2528.075211	4020.191705
1995-11-30	4025.299179	383.803351	3273.058433	4777.539924
1995-12-31	6013.424308	383.839762	5261.112198	6765.736417

Future 12 Months Sales Forecast :

```
1995-08-31    1840.381271
1995-09-30    2491.896629
1995-10-31    3258.142750
1995-11-30    3857.215381
1995-12-31    6092.517723
1996-01-31    1187.145902
1996-02-29    1587.383038
1996-03-31    1857.722056
1996-04-30    1843.773531
1996-05-31    1681.470365
1996-06-30    1642.727017
1996-07-31    1996.474937
Freq: M, dtype: float64
```

Forecast for next 12 months along with confidence band:

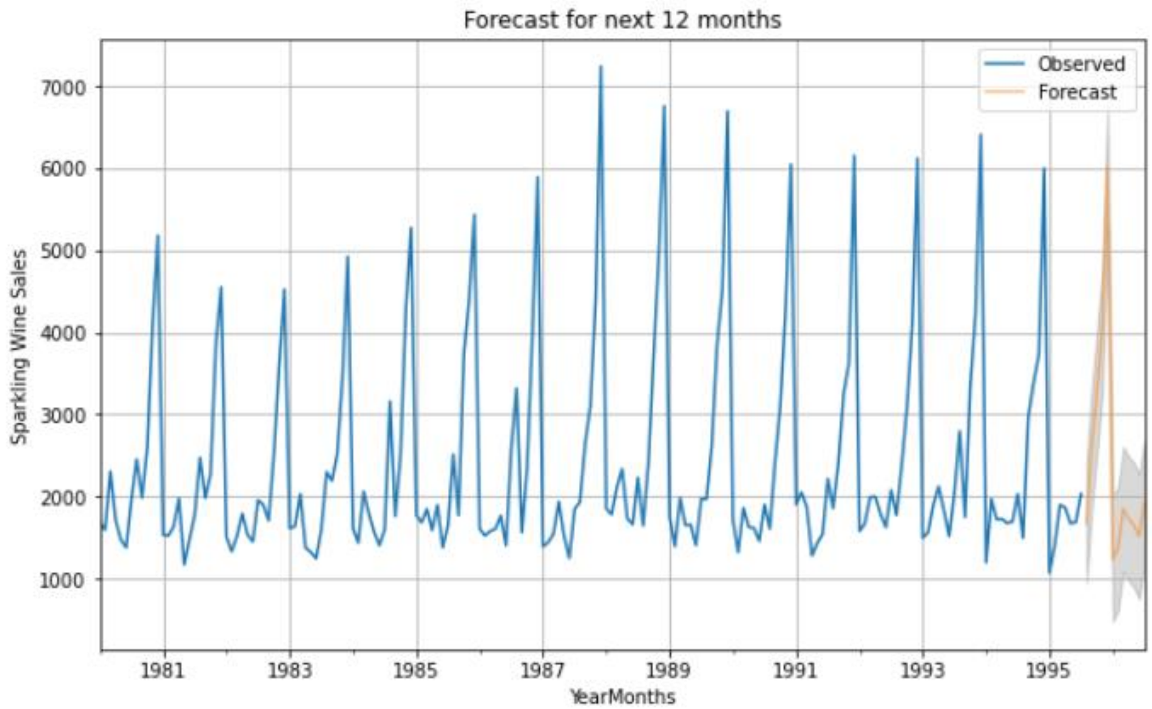


Figure 29 – Future 12 months Forecast with confidence band

	RMSE Full Model
SARIMA(3,1,2)(6,1,1,12), Full Model	626.083007

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

The Sparkling Wine is very much in demand for sure. Hence, the company should try and capitalize this into increasing sales during the off-season through marketing, co-offers, 2+2 offers etc.

The peak season is not a problem for the company. However, the sales are not increasing at a desired proportion YOY. Hence, they should look out for introducing new stores in domestic and foreign markets. They can also try and look too expand their factory capacity by starting a new plant.

They can also tie up with mega event companies to create more brand exposure and also can promote the brand on social, digital as well as offline media.