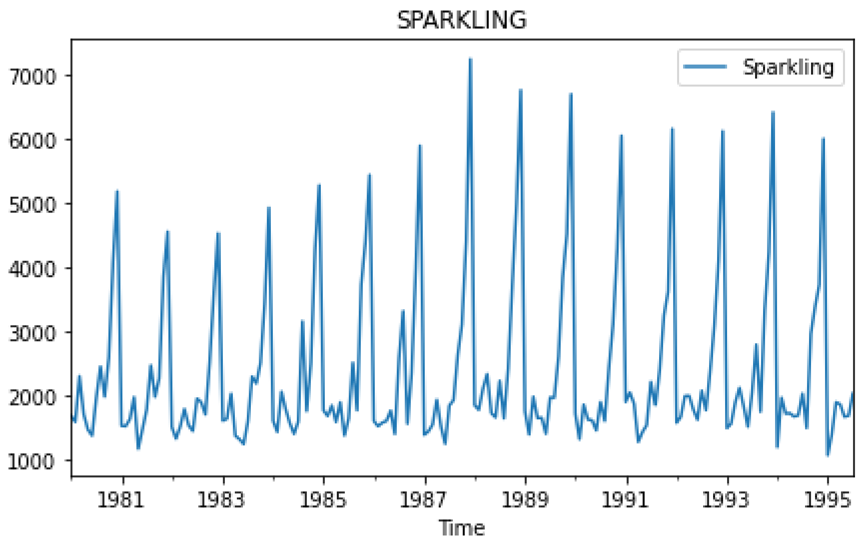
Q1. Read the data as an appropriate Time Series data and plot the data.

The csv file "Sparkling" is imported and the index is rest to a time stamp explicitly to get a sense of a time-series. Here is the head of the data.

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

After plotting, the data looks like following: -



The data has some sense of seasonality but the trend is almost constant.

Q2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Following is the description of the data: -

```
DatetimeIndex: 187 entries, 1980-01-31 to 1995-07-31
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
```

It is quite clear from the info. the data contains only one column of int type.

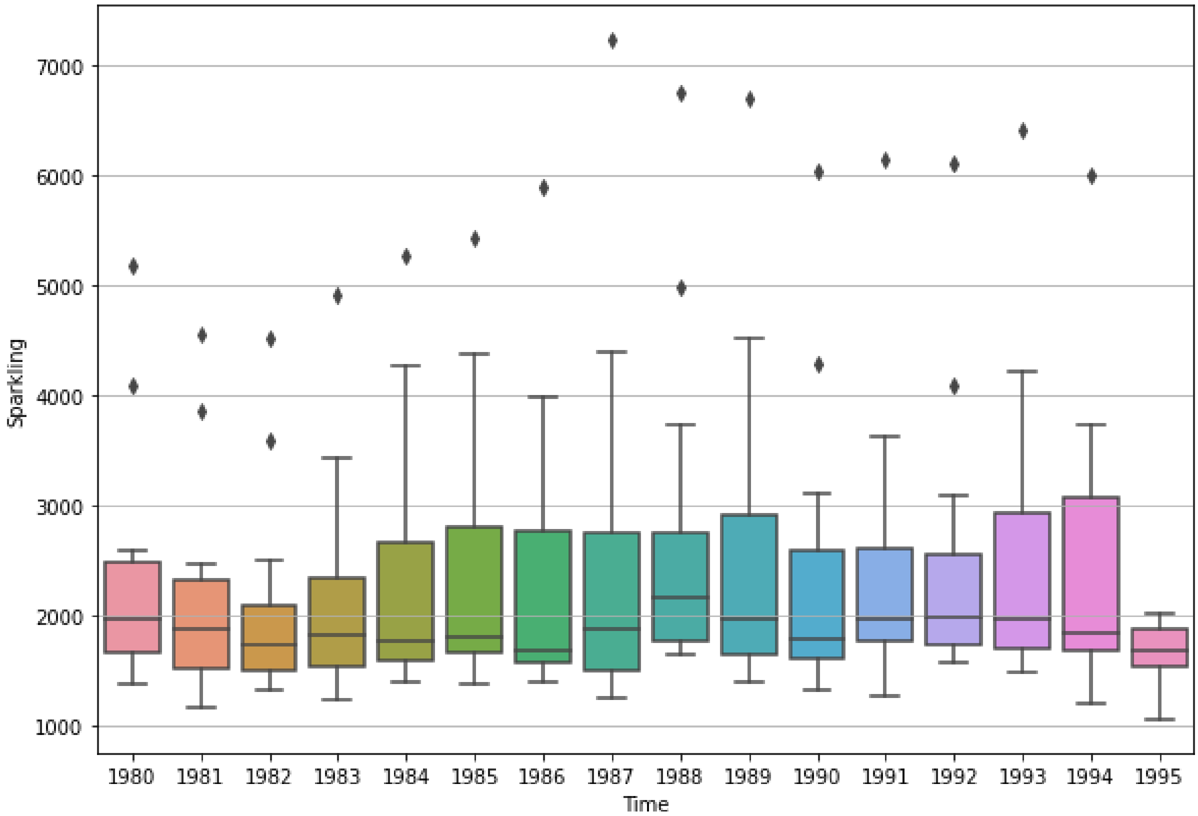
	Sparkling
count	187.000000
mean	2402.417112
std	1295.111540
min	1070.000000
25%	1605.000000
50%	1874.000000
75%	2549.000000
max	7242.000000

```
df.isnull().sum()
```

```
Sparkling    0
dtype: int64
```

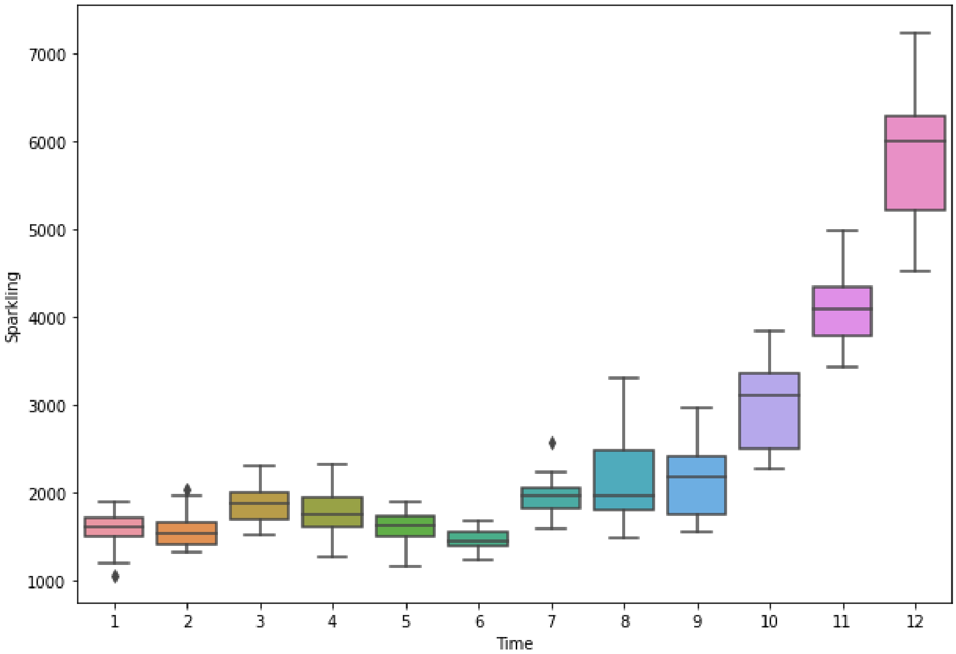
The data has no missing values.

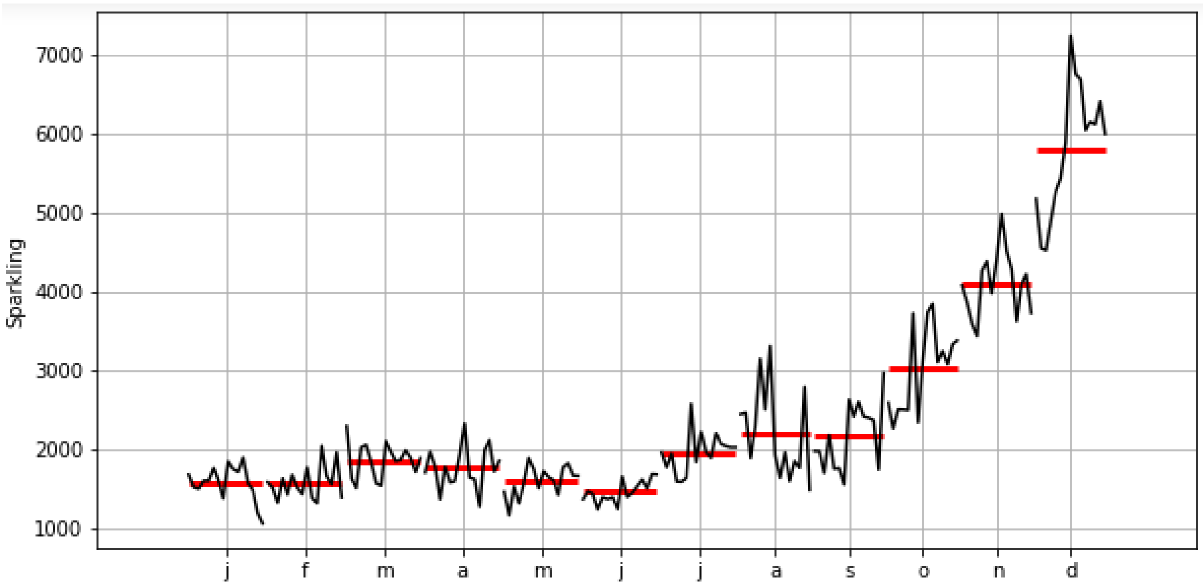
Yearly Plot: -



The above plot depicts the yearly sales of the wine. The sales is fairly high in 1984, 1985, 1989 and 1994.

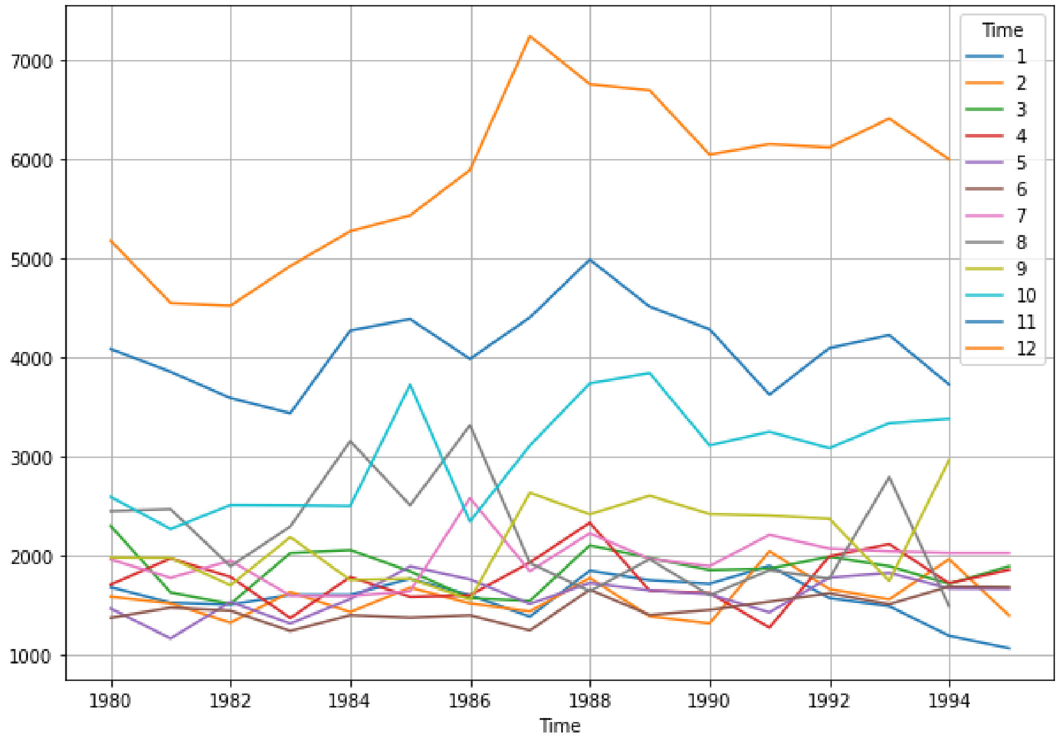
Monthly Plot





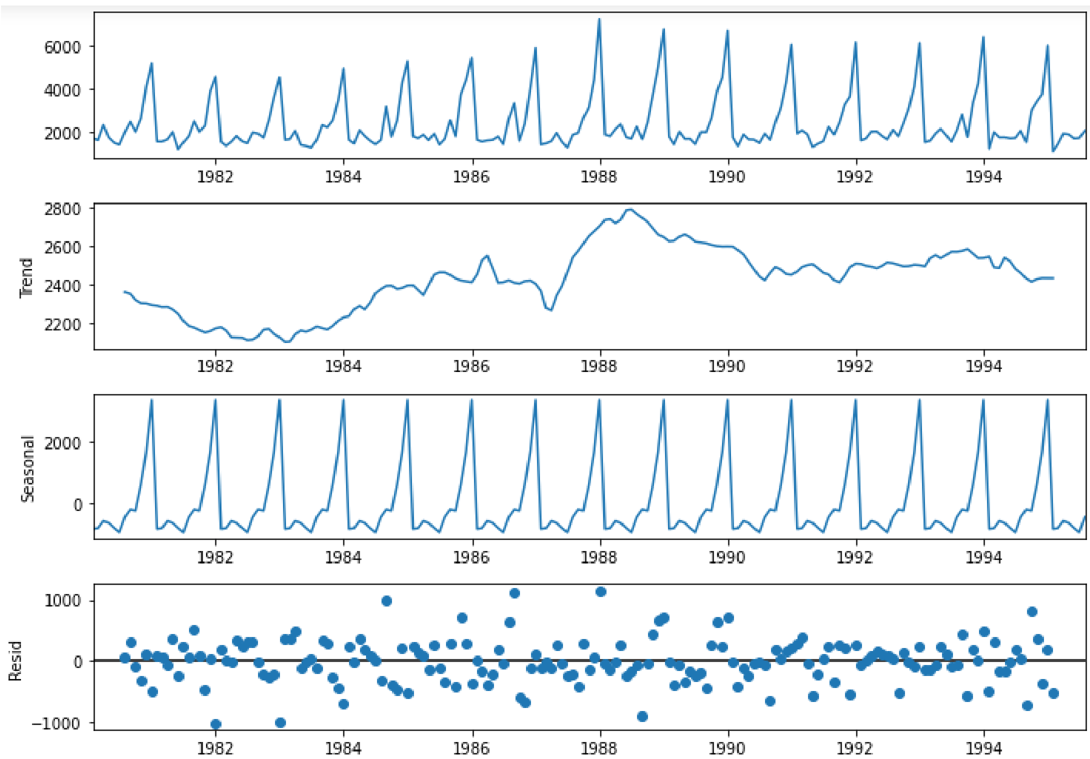
From the monthly plot we can infer that the sales are particularly skyrocketing in the month of December. It should not come as a surprise as the month of December is a festive month (Christmas + New Year).

Year-Month Plot



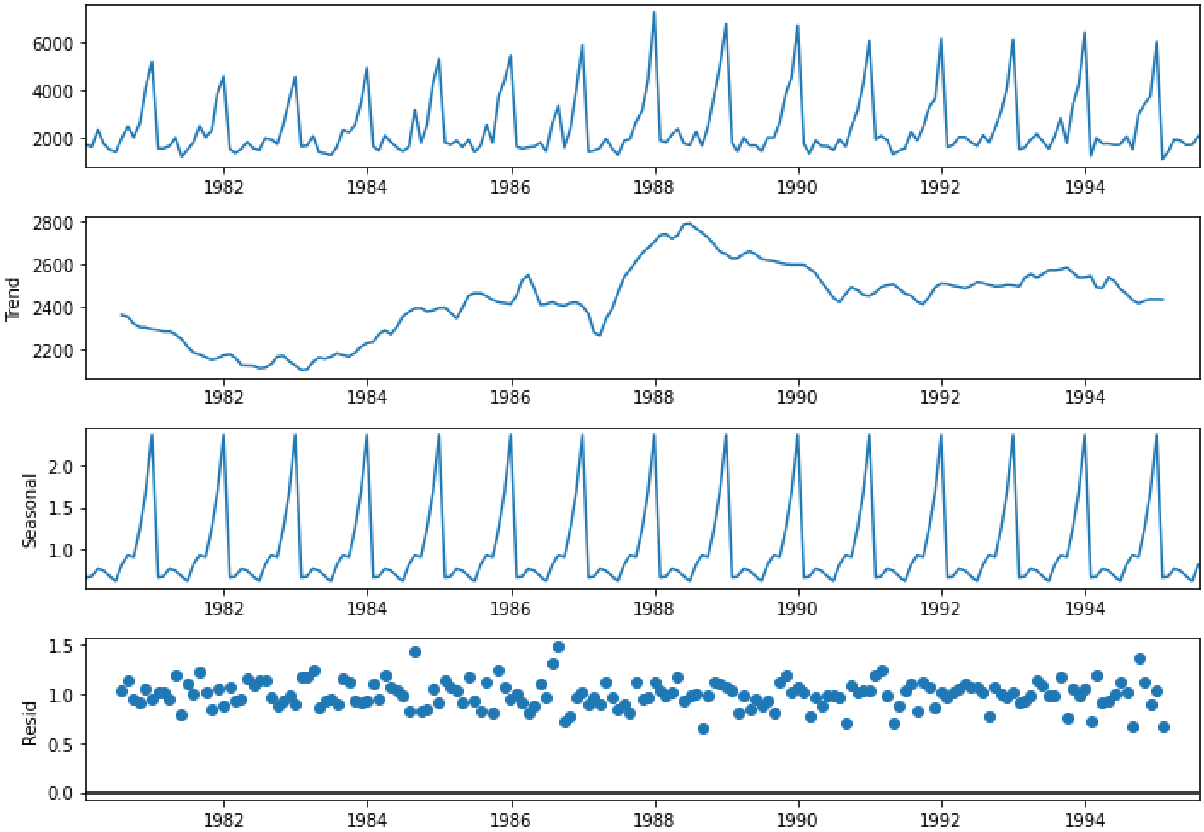
It is evident from the above pot that the sales in the month of December across all years have remained highest.

Data decomposition: Additive



The residuals, do not show any particular pattern, they are completely random, thus the additive decomposition holds good here. However, we can’t move ahead with the multiplicative decomposition.

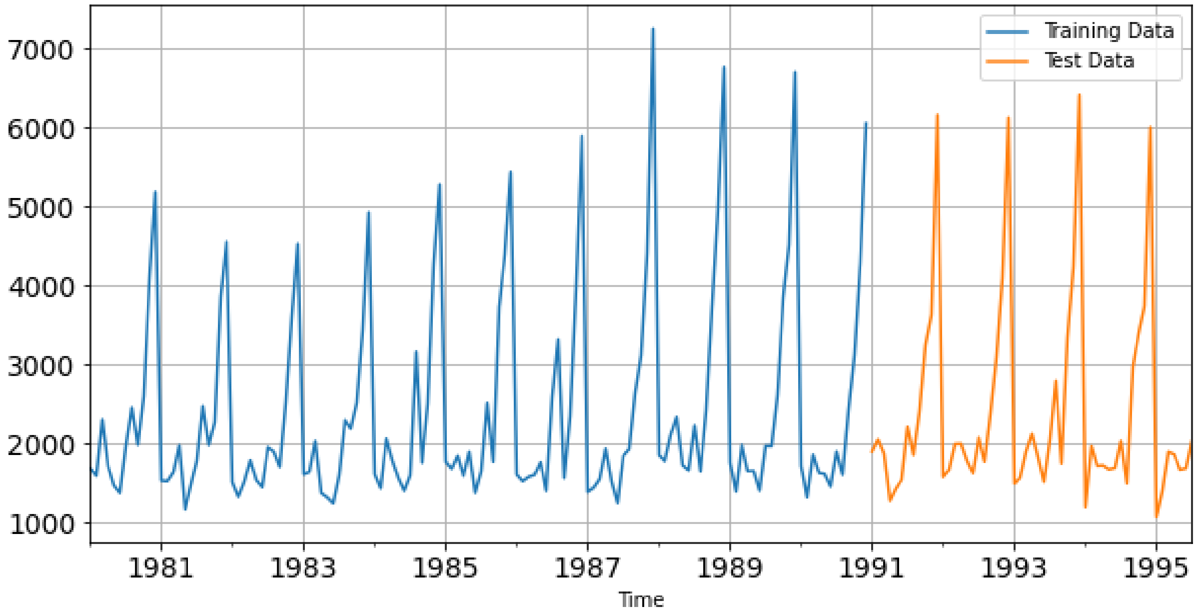
Multiplicative Decomposition:



Mostly the residuals are between 1 and 1.5. Multiplicative model too holds good here.

Q3 Split the data into training and test. The test data should start in 1991.

The data is split accordingly: -

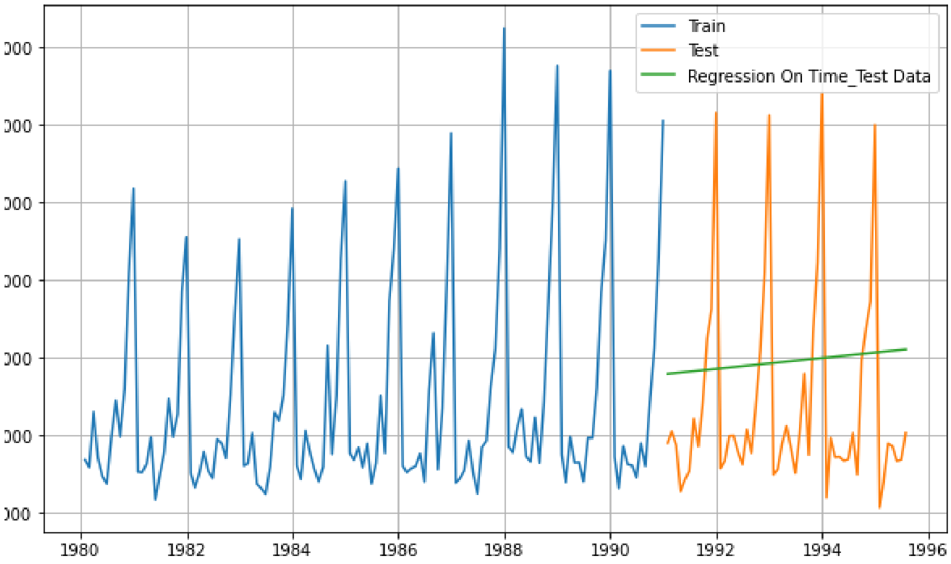


Q4 Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data.

Following models are built: -

1. Linear Regression
2. Naïve Model
3. Simple average
4. Exponential Smoothing

Linear Regression: -



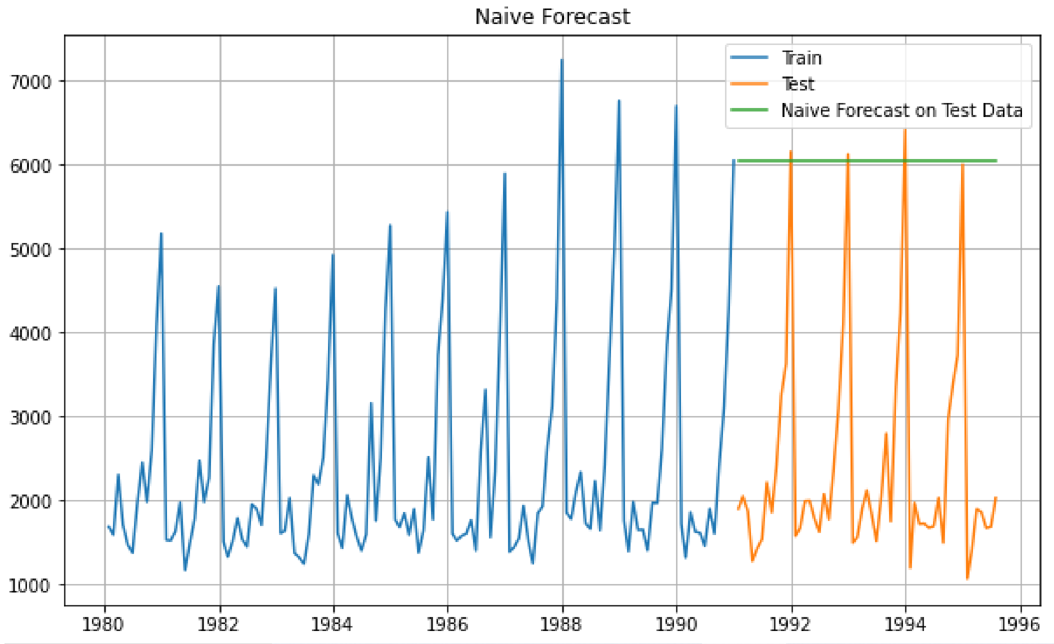
The regression model considers the sales as the target variable and the time stamp as the independent variable. It just takes into account the possible trend and predicts accordingly. Not very accurate.

	Test RMSE
RegressionOnTime	1389.135175

The RMSE score is also too high, and can't be used for final model building.

Naïve Approach

The Naïve approach just takes the latest value and presents it as upcoming forecast.

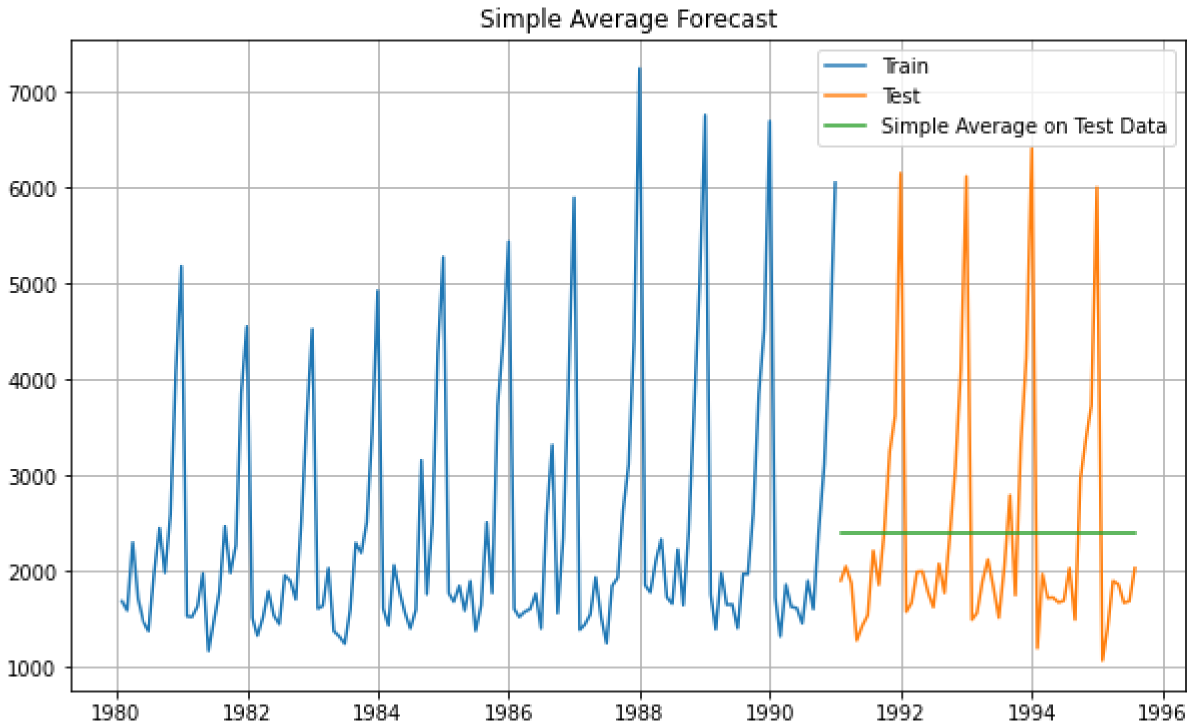


	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352

Evident that the model seems almost of no use, with an even higher RMSE.

Simple average

It considers the average on the whole data and presents it as the forecast.

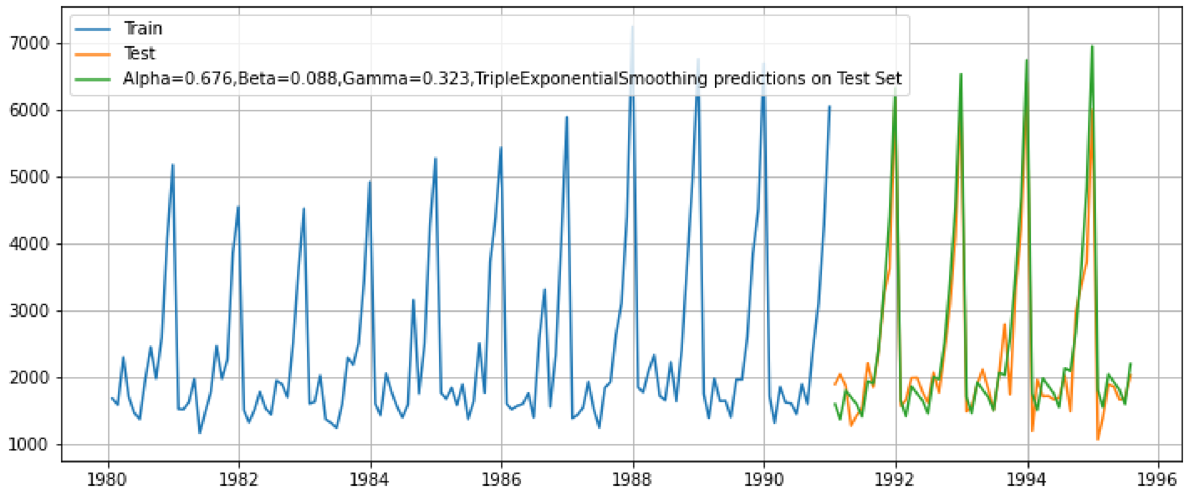


	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804

The simple average has better RMSE.

Exponential Smoothing

The data has some amount of trend and an evident seasonality. A triple exponential smoothing is the one that could fit it the best.



As can be seen, the testing data and the predictions are so close.

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
Alpha=0.676,Beta=0.088,Gamma=0.323,TripleExponentialSmoothing	383.157627

The RMSE has significantly decreased.

Q5. 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.

H0: The time series has a unit root and is not stationary.

H1: The time series does not have a unit root and is stationary.

```
dftest = adfuller(df,regression='ct')
print('DF test statistic is %3.3f' %dftest[0])
print('DF test p-value is' ,dftest[1])
print('Number of lags used' ,dftest[2])
```

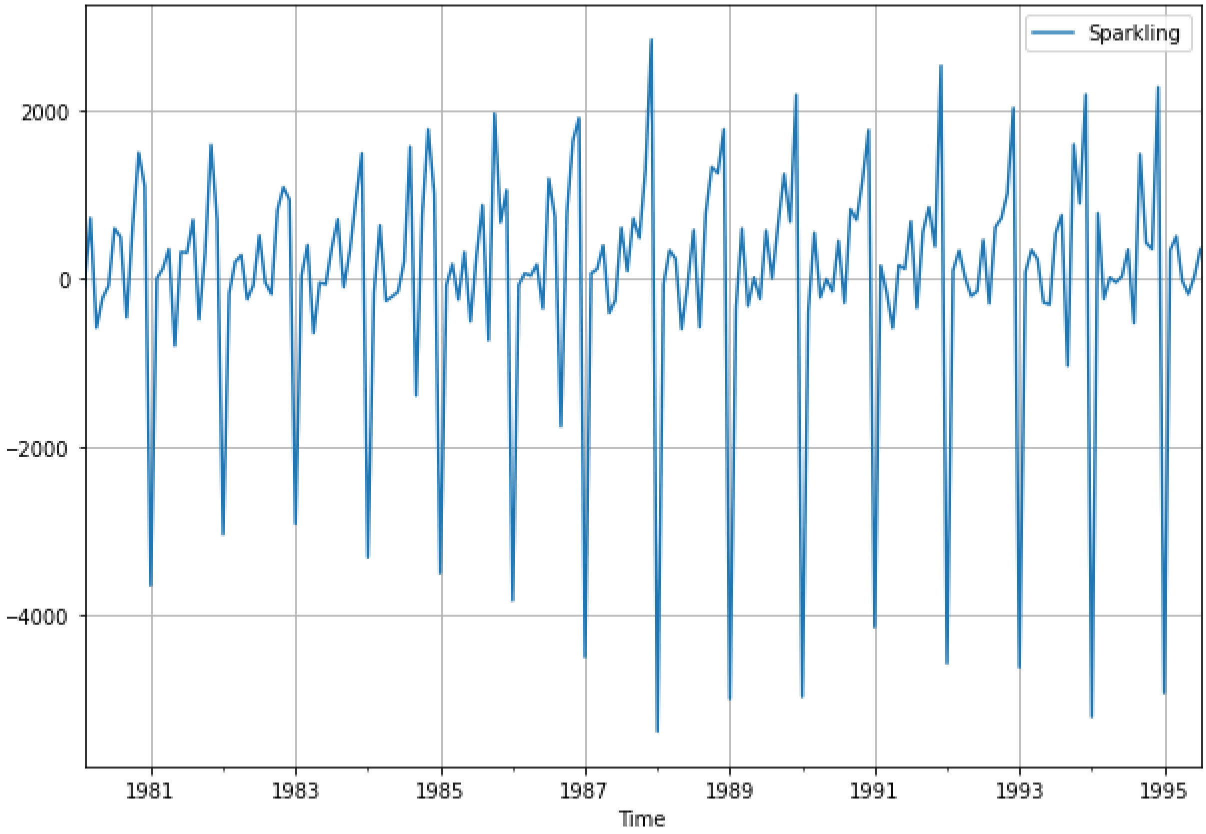
```
DF test statistic is -1.798
DF test p-value is 0.7055958459932417
Number of lags used 12
```

From the ADF test, we can arrive at a conclusion that p>.05 and we can't reject NULL, thus the series is not stationary. We need to follow certain steps of differencing to make it stationary.

```
dftest = adfuller(df.diff().dropna(),regression='ct')
print('DF test statistic is %3.3f' %dftest[0])
print('DF test p-value is' ,dftest[1])
print('Number of lags used' ,dftest[2])
```

```
DF test statistic is -44.912
DF test p-value is 0.0
Number of lags used 10
```

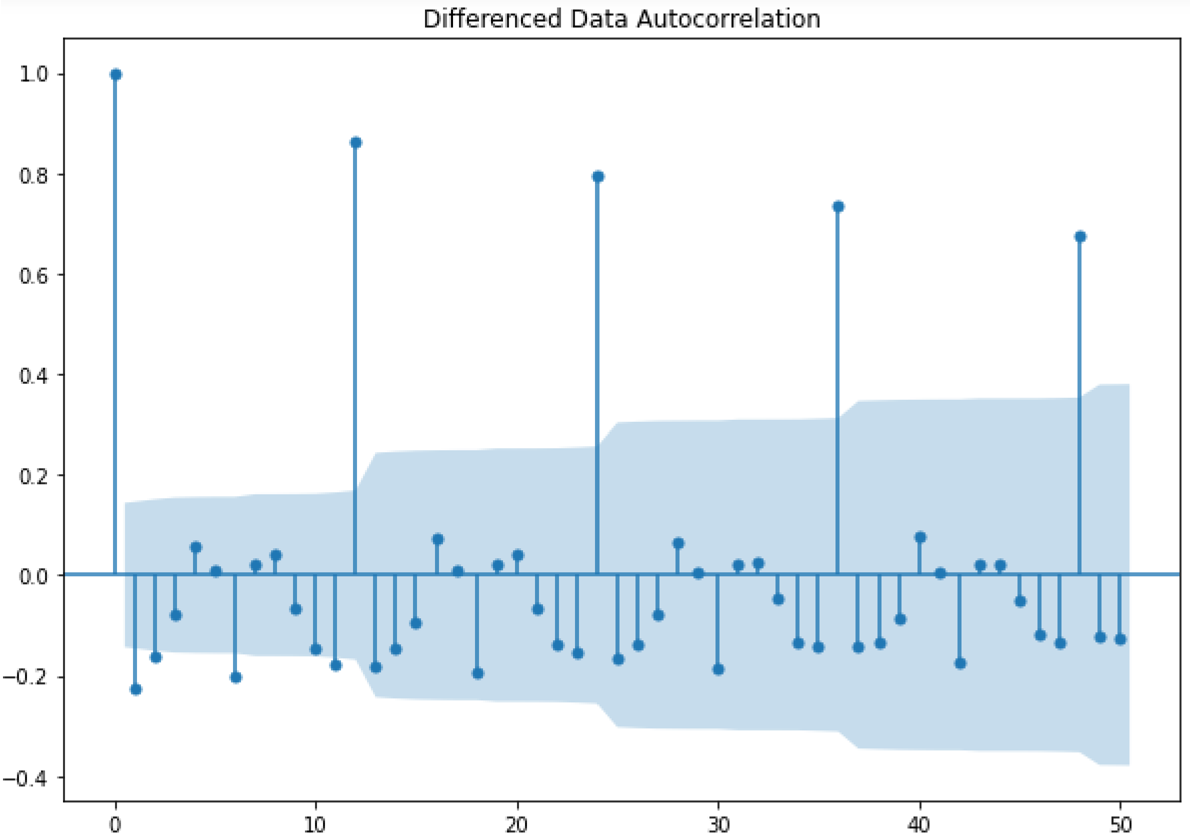
After one level of differencing, we can see the p value is<0.05 and we can thus reject the NULL hypothesis. Now the series is stationary.



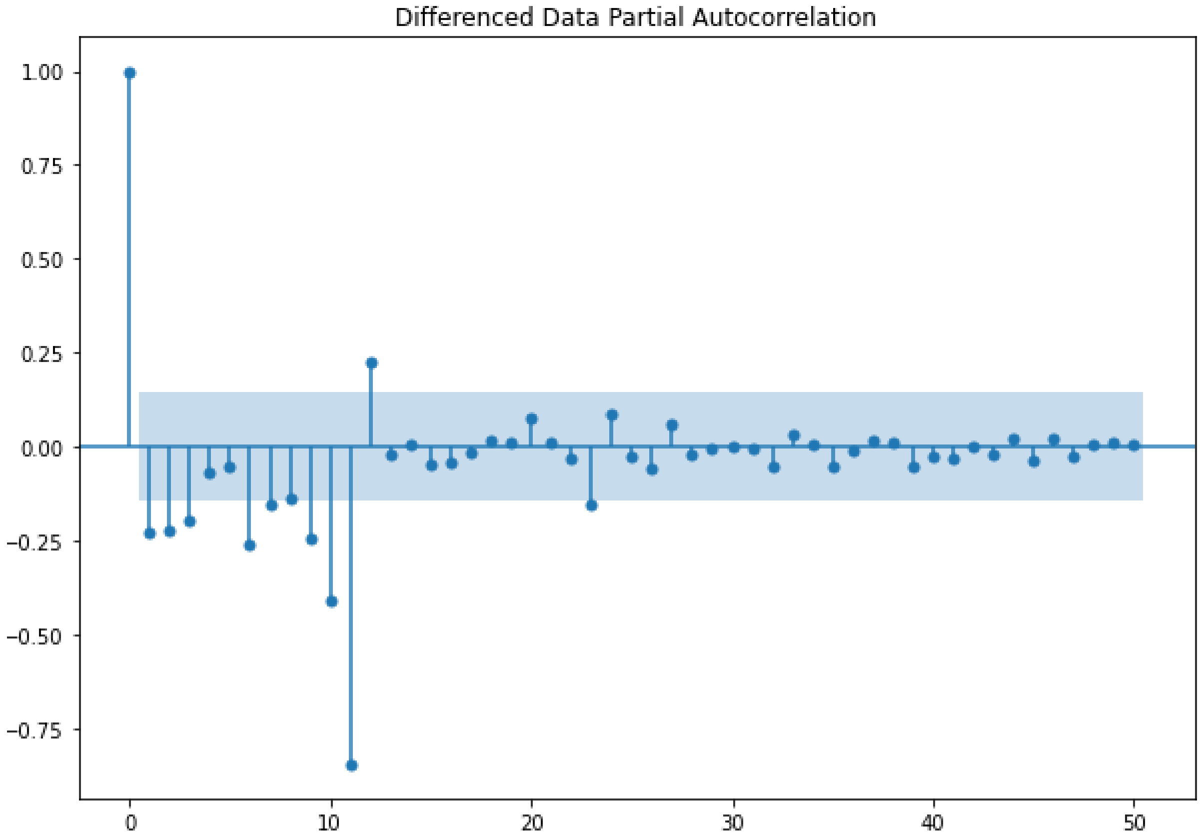
As can be seen, after one level of differencing, the series is stationary.

Q6. 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.

To begin with one needs to check the ACF and PACF plot to get a rough idea about the p and q values. (this has been plotted on the entire data)



From the ACF plot, we can predict the q value to be 2.



From the PACF plots, the p value comes around 3.

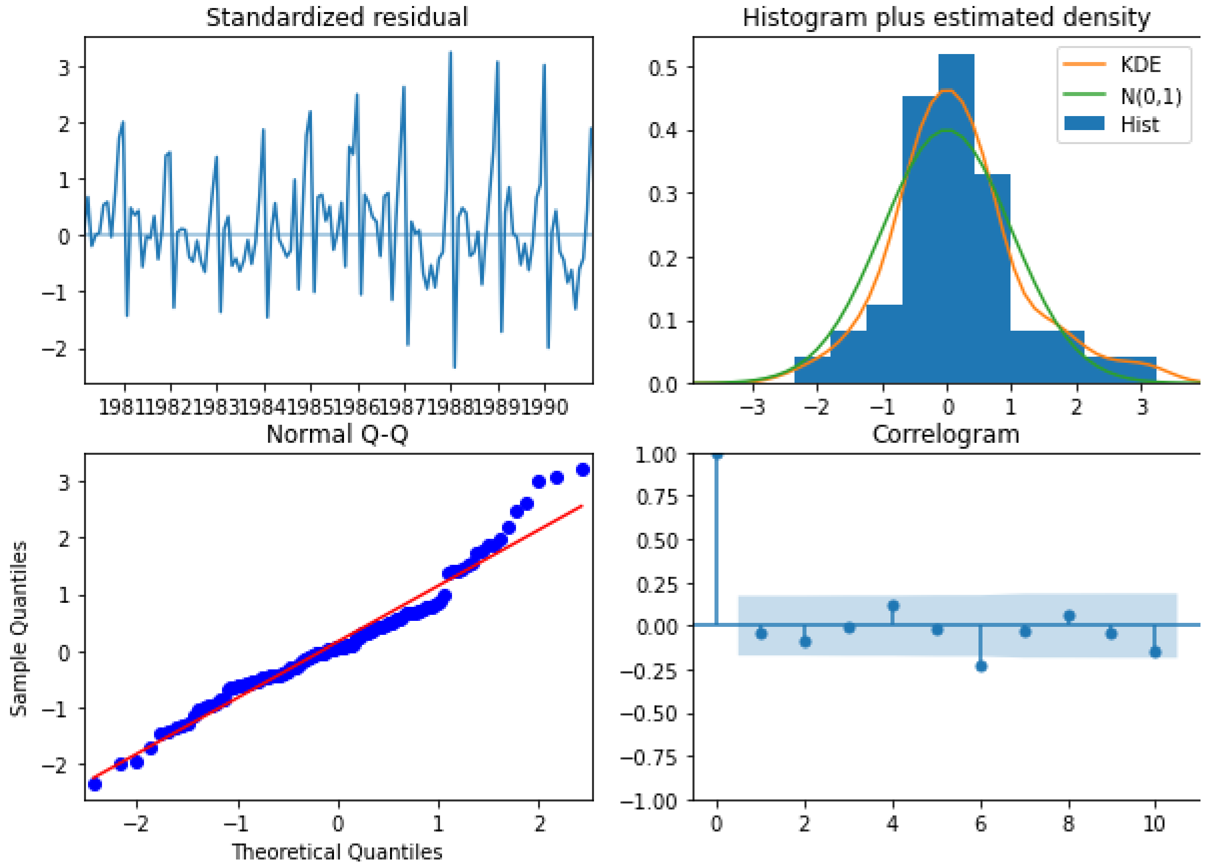
Building the Automated ARIMA model: -

	param	AIC
10	(2, 1, 2)	2213.509212
15	(3, 1, 3)	2221.455906
14	(3, 1, 2)	2230.775698
11	(2, 1, 3)	2232.960945
9	(2, 1, 1)	2233.777626

The least AIC comes for parameters (2,1,2). We already know, data is not stationary, and we need to apply differencing to make it stationary, thus d = 1.

Building, the model yields the following summary: -

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1101.755
Date:                Thu, 22 Apr 2021   AIC                           2213.509
Time:                        19:56:58   BIC                           2227.885
Sample:                    01-31-1980   HQIC                          2219.351
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.3121      0.046     28.782      0.000       1.223       1.401
ar.L2         -0.5593      0.072     -7.741      0.000      -0.701      -0.418
ma.L1         -1.9917      0.109    -18.217      0.000      -2.206      -1.777
ma.L2          0.9999      0.110      9.109      0.000       0.785       1.215
sigma2      1.099e+06   1.99e-07   5.51e+12      0.000     1.1e+06     1.1e+06
===================================================================================
Ljung-Box (Q):                      293.72   Jarque-Bera (JB):                14.46
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               2.43   Skew:                             0.61
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.08
===================================================================================
```



The residuals show a little deviation from the original.

Prediction on the test data: -

	RMSE	MAPE
ARIMA(2,1,2)	1299.979402	47.099871

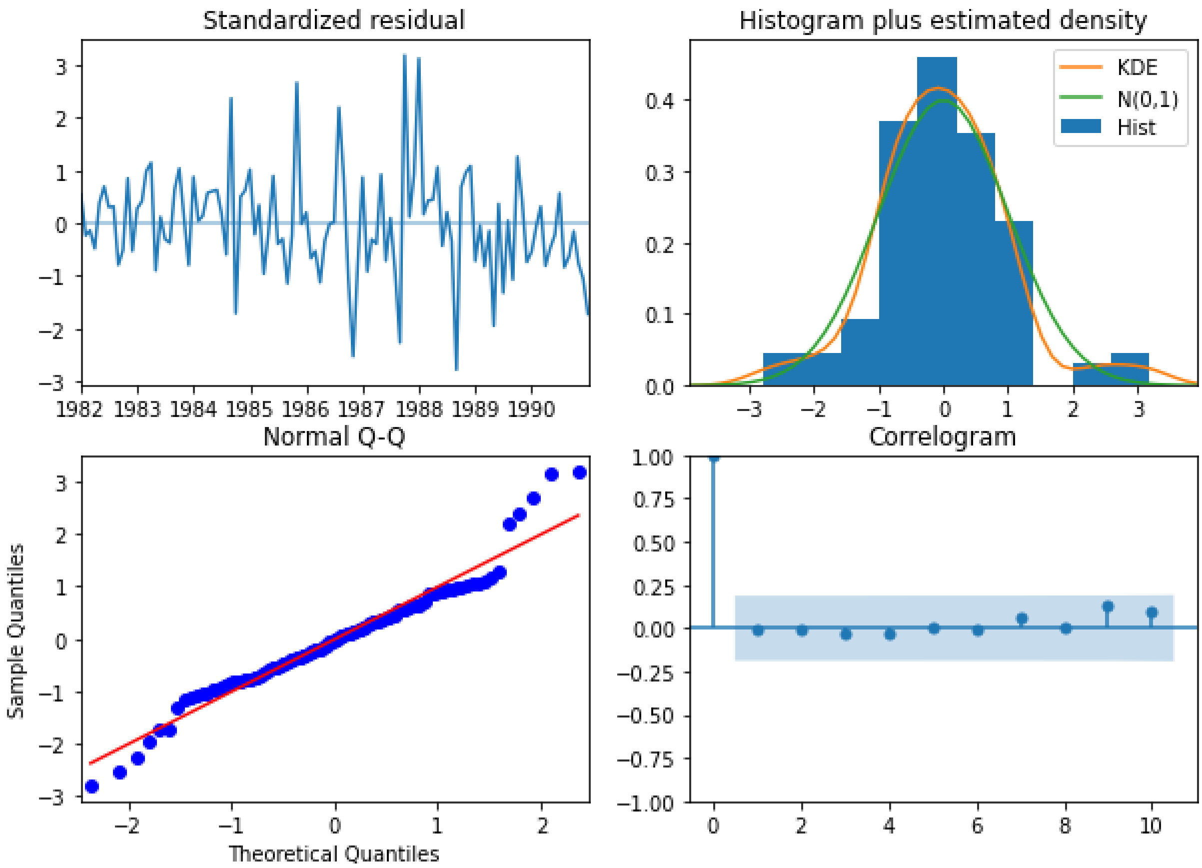
Yields a bit higher RMSE.

Building automated SARIMA model: -

	param	seasonal	AIC
187	(2, 1, 3)	(2, 0, 3, 6)	1629.282423
251	(3, 1, 3)	(2, 0, 3, 6)	1631.005091
59	(0, 1, 3)	(2, 0, 3, 6)	1633.327872
123	(1, 1, 3)	(2, 0, 3, 6)	1633.965380
63	(0, 1, 3)	(3, 0, 3, 6)	1635.101039

Building model based on the least AIC values: -

```
                                     SARIMAX Results                                      
==========================================================================================
Dep. Variable:                           Sparkling   No. Observations:                  132
Model:             SARIMAX(2, 1, 3)x(2, 0, 3, 6)   Log Likelihood                -803.641
Date:                             Thu, 22 Apr 2021   AIC                           1629.282
Time:                                     20:54:50   BIC                           1658.887
Sample:                                 01-31-1980   HQIC                          1641.288
                                      - 12-31-1990                                         
Covariance Type:                               opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.7438      0.089    -19.547      0.000      -1.919      -1.569
ar.L2         -0.7859      0.085     -9.229      0.000      -0.953      -0.619
ma.L1          1.0835      4.387      0.247      0.805      -7.515       9.682
ma.L2         -0.7543      0.556     -1.356      0.175      -1.844       0.336
ma.L3         -0.8886      3.952     -0.225      0.822      -8.634       6.857
ar.S.L6       -0.0133      0.031     -0.428      0.669      -0.074       0.047
ar.S.L12       1.0360      0.024     43.438      0.000       0.989       1.083
ma.S.L6       -1.1428      1.265     -0.904      0.366      -3.622       1.336
ma.S.L12      -1.4244      0.593     -2.401      0.016      -2.587      -0.262
ma.S.L18       0.3348      0.804      0.417      0.677      -1.241       1.910
sigma2      3.973e+04   1.66e+05      0.240      0.811   -2.85e+05    3.65e+05
===================================================================================
Ljung-Box (Q):                       22.76   Jarque-Bera (JB):                15.24
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               1.47   Skew:                             0.38
Prob(H) (two-sided):                  0.26   Kurtosis:                         4.66
===================================================================================
```



Quite a bit deviation can be seen in the sample vs theoretical quantities.

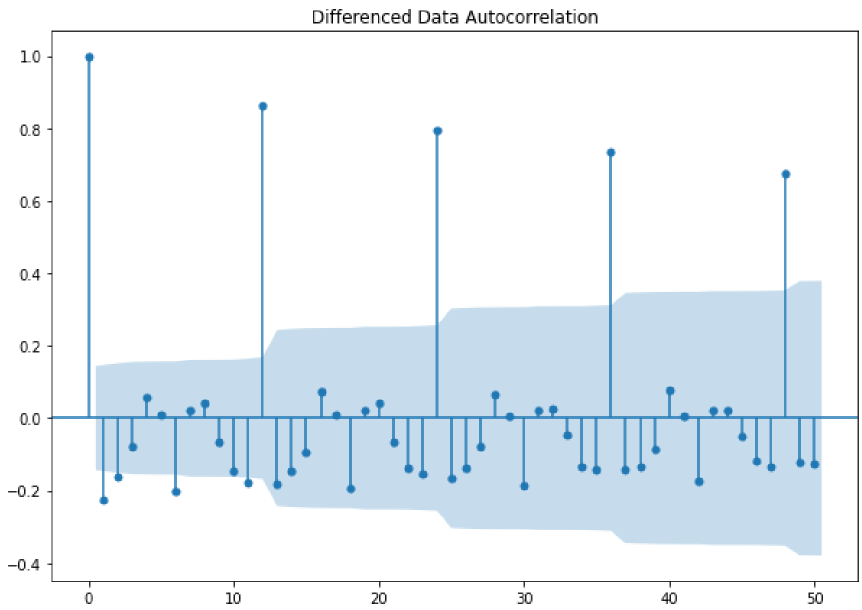
Predicting on the test data: -

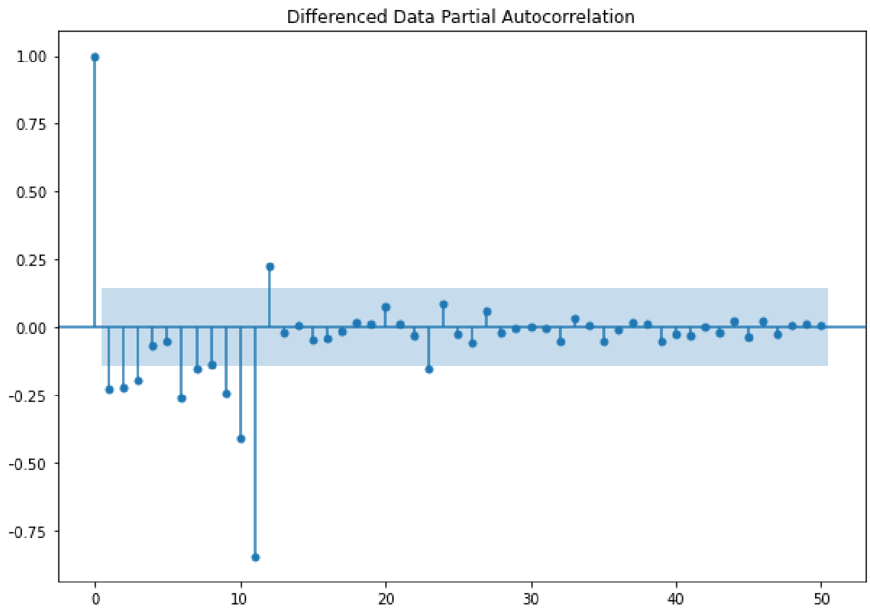
	RMSE	MAPE
ARIMA(2,1,2)	1299.979402	47.099871
SARIMA(2,1,3)(2,0,3,6)	838.941329	36.867492

The RMSE have significantly dropped with the SARIMA Model.

Q7. 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.

Looking at the ACF and PACF plot to determine p and q: -

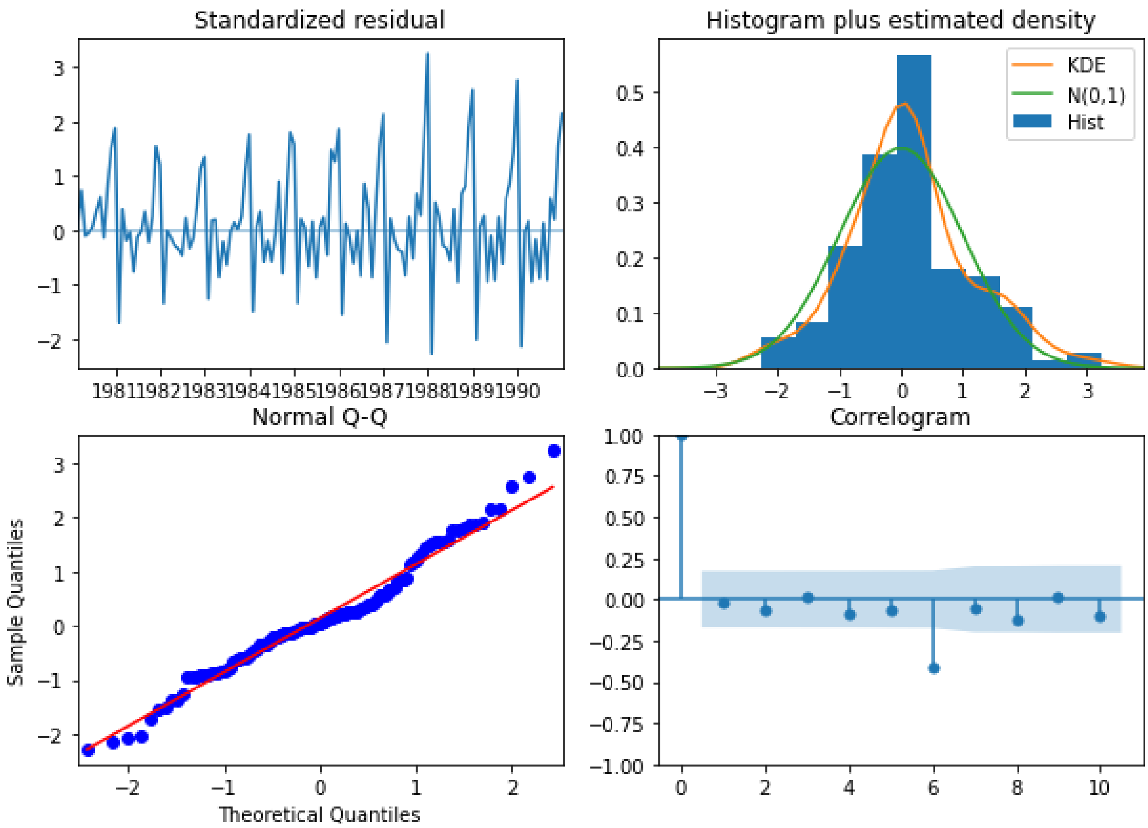




From the above two plots q = 2, and p = 3.

The manual model yields the following results: -

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  132
Model:                 ARIMA(3, 1, 2)   Log Likelihood               -1109.388
Date:                Thu, 22 Apr 2021   AIC                           2230.776
Time:                        21:56:03   BIC                           2248.027
Sample:                    01-31-1980   HQIC                          2237.786
                         - 12-31-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4286      0.047     -9.053      0.000      -0.521      -0.336
ar.L2          0.3357      0.106      3.163      0.002       0.128       0.544
ar.L3         -0.2357      0.059     -4.006      0.000      -0.351      -0.120
ma.L1          0.0160      0.130      0.123      0.902      -0.238       0.270
ma.L2         -0.9838      0.136     -7.226      0.000      -1.251      -0.717
sigma2      1.273e+06   1.96e-07    6.5e+12      0.000    1.27e+06    1.27e+06
```



	RMSE	MAPE
ARIMA(2,1,2)	1299.979402	47.099871
SARIMA(2,1,3)(2,0,3,6)	838.941329	36.867492
ARIMA(3,1,2)	1280.666115	43.399032

Prediction on the test data, shows lower value than automated ARIMA, but still not the best.

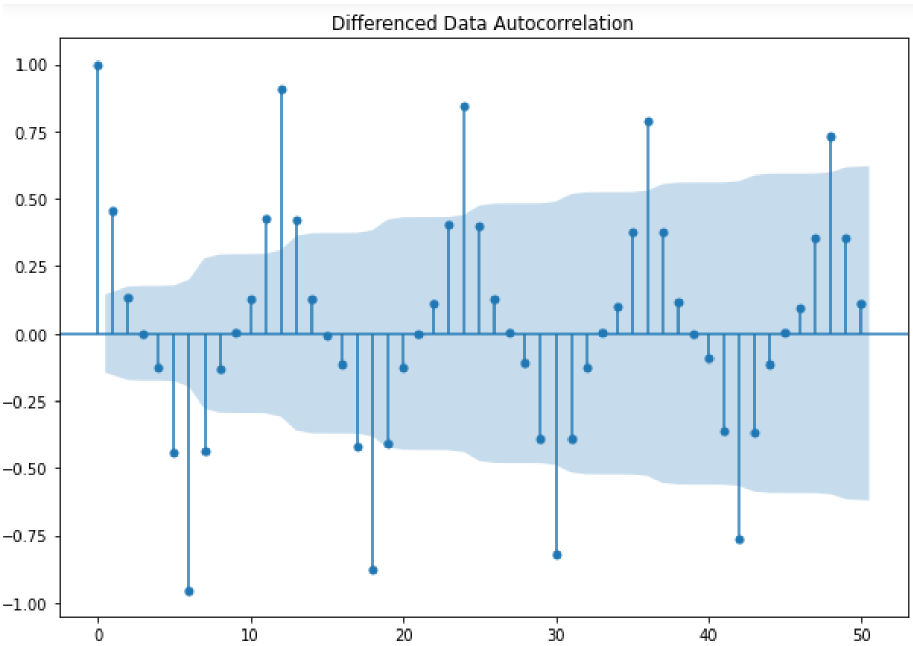
Moving ahead with the manual SARIMA.

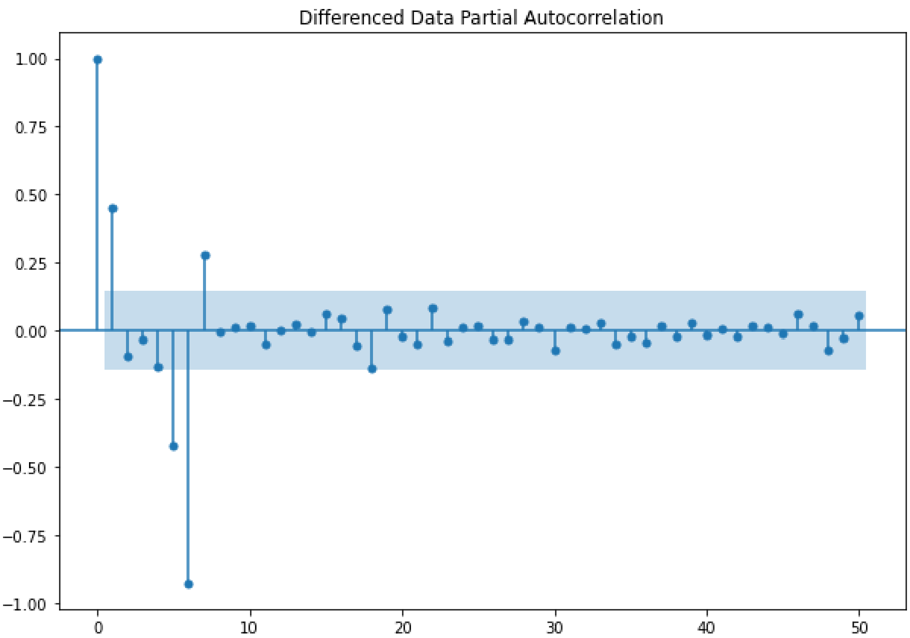
Looking at the stationarity: -

```
DF test statistic is -11.364
DF test p-value is 4.720421360314017e-18
Number of lags used 6
```

The data is stationary and thus D = 0.

Looking at the seasonal ACF and PACF with differencing 6: -





From the above two plots P = 2 and Q = 1.

Finally, we have: -

p = 1

q = 2

d = 1

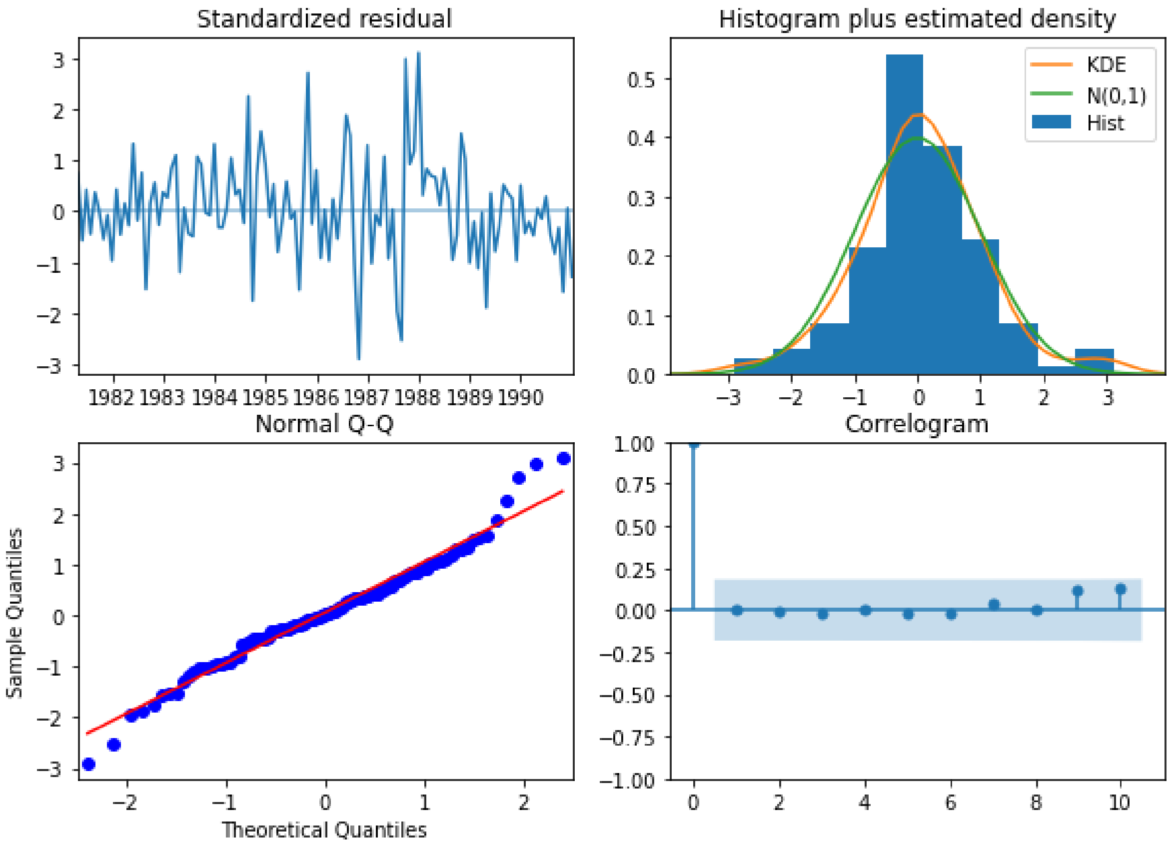
P = 2

Q = 1

D = 0

The final model yields the following results: -

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                            Sparkling   No. Observations:                132
Model:             SARIMAX(2, 1, 2)x(2, 0, [1], 6)   Log Likelihood                -872.413
Date:                              Thu, 22 Apr 2021   AIC                           1760.827
Time:                                      22:03:10   BIC                           1782.924
Sample:                                  01-31-1980   HQIC                          1769.798
                                       - 12-31-1990
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5704      0.166     -3.440      0.001      -0.895      -0.245
ar.L2          0.0875      0.094      0.932      0.351      -0.097       0.272
ma.L1         -0.1410      0.182     -0.774      0.439      -0.498       0.216
ma.L2         -0.8589      0.171     -5.036      0.000      -1.193      -0.525
ar.S.L6       -0.0246      0.053     -0.464      0.643      -0.129       0.080
ar.S.L12       0.9520      0.031     30.797      0.000       0.891       1.013
ma.S.L6        0.0663      0.139      0.476      0.634      -0.207       0.339
sigma2      1.704e+05   1.24e-06   1.37e+11      0.000     1.7e+05     1.7e+05
===================================================================================
```



Finally, plotting predicting on the test data, the least value for manual SARIMA is obtained: -

	RMSE	MAPE
ARIMA(2,1,2)	1299.979402	47.099871
SARIMA(2,1,3)(2,0,3,6)	838.941329	36.867492
ARIMA(3,1,2)	1280.666115	43.399032
SARIMA(2,1,2)(2,0,1,6)	335.216828	12.323205

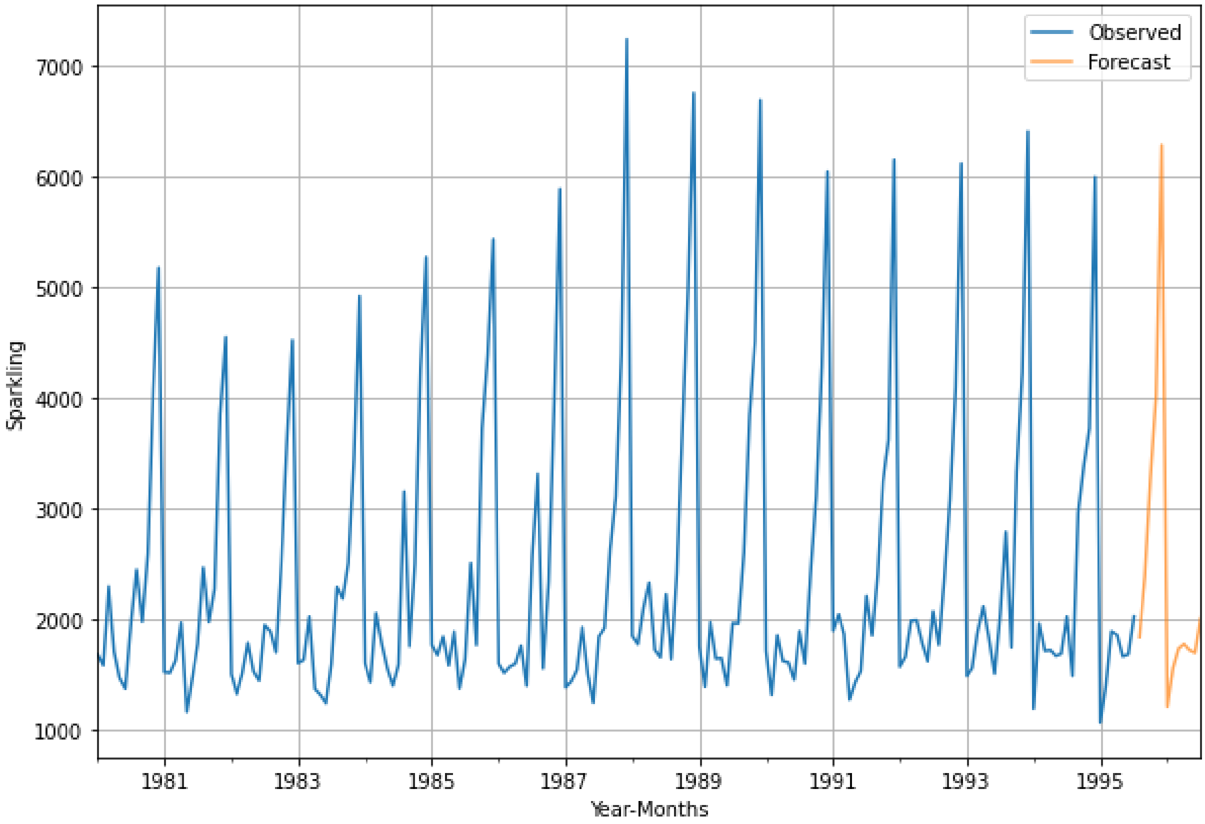
Q8 Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

MODEL	RMSE
Linear Regression	1389.13
Naïve Approach	3864.27
Simple Average	1275.08
Exponential Smoothing	383.15
Automated ARIMA (2,1,2)	1299.97
Automated SARIMA (2,1,3)(1,0,3,6)	839.48
Manual ARIMA(3,1,2)	1280.66
Manual SARIMA(2,1,2)(2,0,1,6)	335.2

Q9. 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.

Building the model on the whole data: -

```
                                   SARIMAX Results
==========================================================================================
Dep. Variable:                              Sparkling   No. Observations:            187
Model:             SARIMAX(2, 1, 2)x(2, 0, [1, 2, 3], 6)   Log Likelihood       -1215.503
Date:                            Thu, 22 Apr 2021   AIC                       2451.006
Time:                                    21:41:40   BIC                       2482.066
Sample:                                01-31-1980   HQIC                      2463.614
                                     - 07-31-1995
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6904      0.272     -2.534      0.011      -1.224      -0.156
ar.L2          0.0102      0.111      0.092      0.927      -0.207       0.227
ma.L1         -0.1380      0.264     -0.522      0.602      -0.656       0.380
ma.L2         -0.7552      0.249     -3.036      0.002      -1.243      -0.268
ar.S.L6        0.0097      0.018      0.530      0.596      -0.026       0.046
ar.S.L12       1.0176      0.011     93.111      0.000       0.996       1.039
ma.S.L6        0.6550      0.374      1.753      0.080      -0.078       1.387
ma.S.L12      -1.1519      0.167     -6.886      0.000      -1.480      -0.824
ma.S.L18      -0.1468      0.268     -0.548      0.583      -0.672       0.378
sigma2      7.133e+04    2.5e+04      2.851      0.004    2.23e+04     1.2e+05
===================================================================================
Ljung-Box (Q):                       19.20   Jarque-Bera (JB):                33.74
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.15   Skew:                             0.53
Prob(H) (two-sided):                  0.60   Kurtosis:                         4.94
===================================================================================
```



Finally, the prediction into the future looks like above.

Q10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

The best model and the least RMSE was shown by the manual SARIMA, with seasonality as 6. The sales peak every 6 months, it is important for the company to encash the same.

Company must bring out exciting offers and discounts for the wine, to increase the sales.

The company could give certain credit points on every purchase which can be used during the next sales.

Additional discounts could be provided on the peak season, i.e., month of December where sales is the highest.

As far as the future is concerned, the next 12 months show the similar pattern, and the above-mentioned steps must be followed to increase sales and profit.