Problem 1 Time Series Forecasting - Sparkling

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

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

Data set :

YearMonth	Sparkling
1980-01	1686
1980-02	1591
1980-03	2304
1980-04	1712
1980-05	1471
1980-06	1377
1980-07	1966
1980-08	2453

We are provided with the above data set of 187 rows and 02 columns. Of the above columns, one column is object data type and one is integer data type.

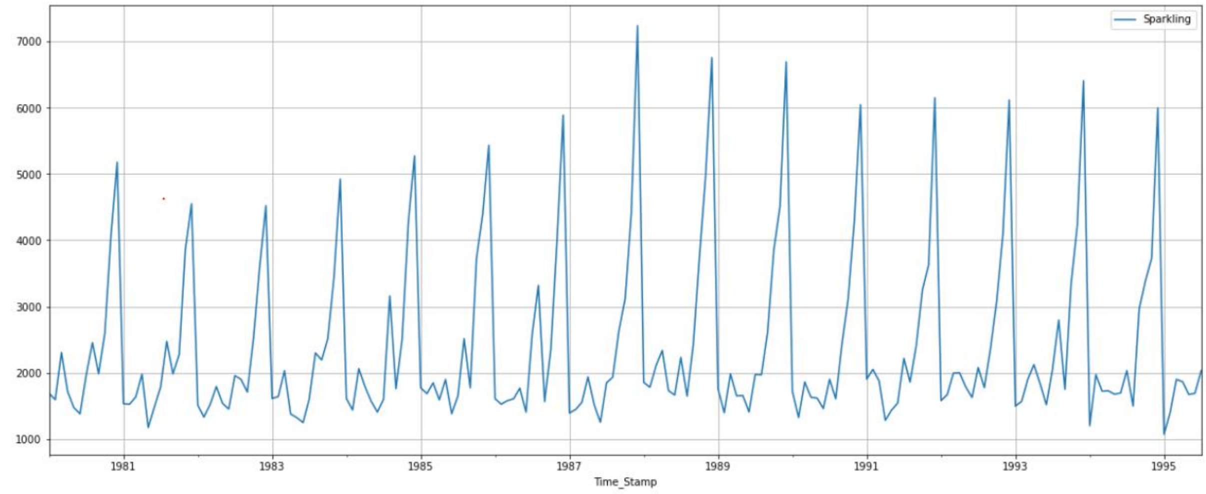
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   YearMonth  187 non-null    object
 1   Sparkling  187 non-null    int64
dtypes: int64(1), object(1)
memory usage: 3.0+ KB
```

There are **no** Null values in the given dataset.

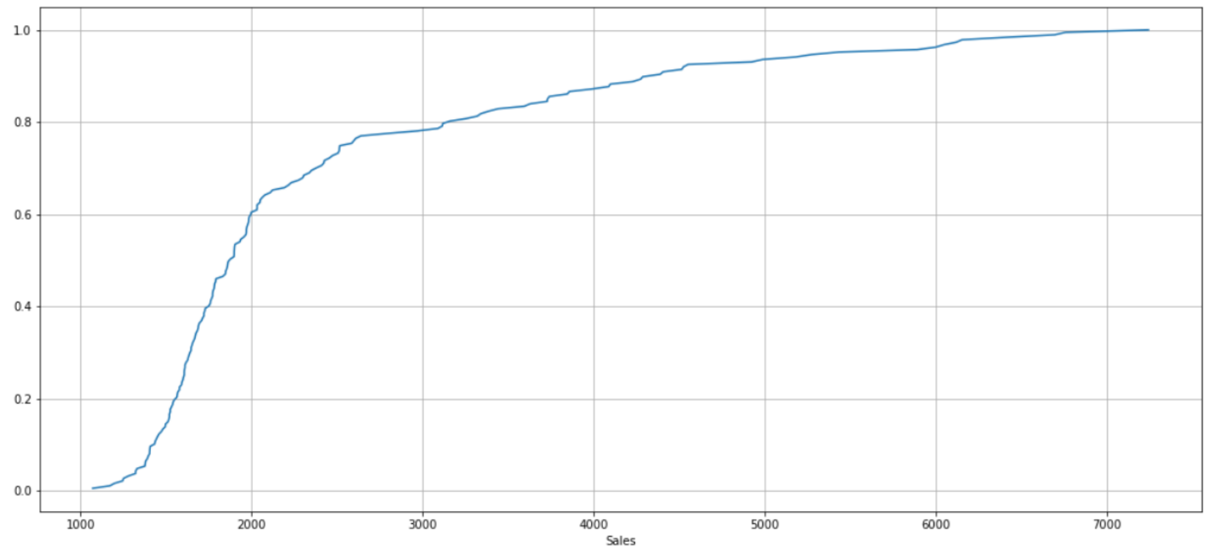
We have read the **YearMonth** column as date type and assign it as index.

```
<class 'pandas.core.frame.DataFrame'>
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
dtypes: int64(1)
memory usage: 2.9 KB
```

By plotting the Time Series to understand the behaviour of the data. We have the following curve



The given data has no identifiable trend and it has seasonality associated with it.

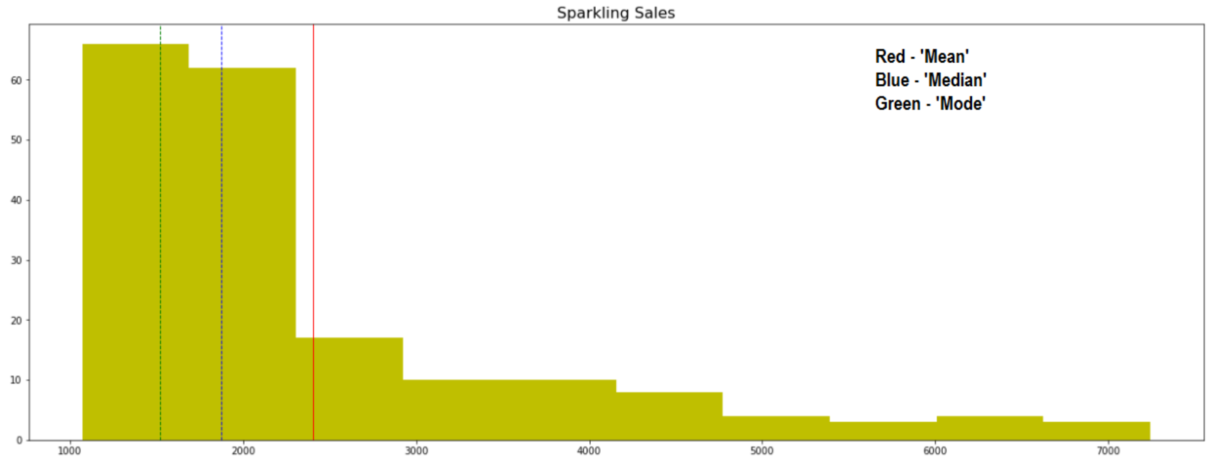


From the above plot , we can see that 60% of the values lie below value 2000 and 80% of values lie below 3200 respectively.

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

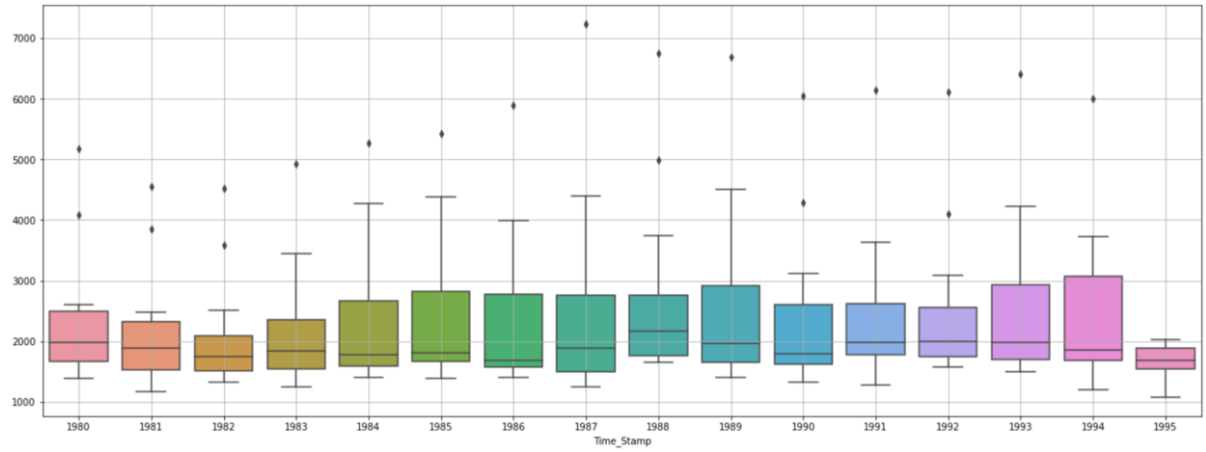
Descriptive statistics of the given time series:

	Sparkling
count	187.000
mean	2402.417
std	1295.112
min	1070.000
25%	1605.000
50%	1874.000
75%	2549.000
max	7242.000



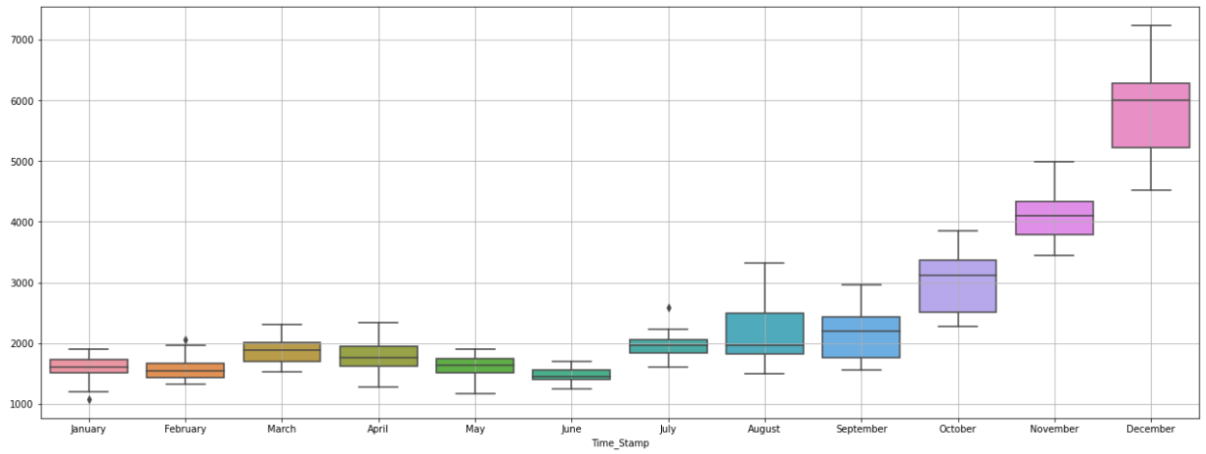
The given data set has mean of value – '2402.417' and median value –'1874'

Spread of sales across different years:

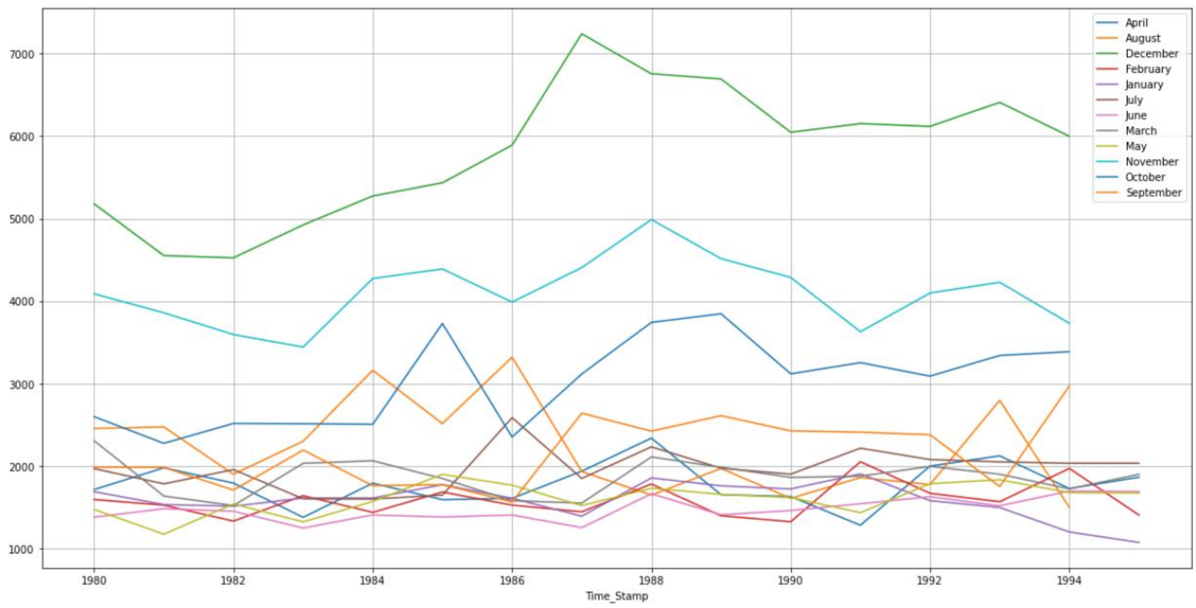


We can see that sales have increased from start to middle and decreased from middle to last. All most all years are showing outlier values of the data set.

Spread of sales across different months:



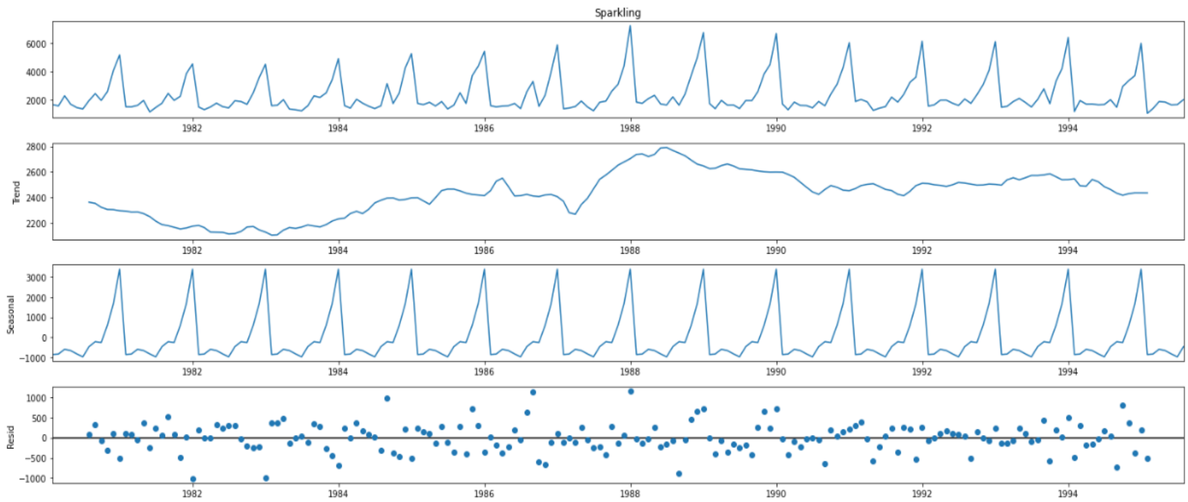
We can understand that **December month** is having the highest sales among all the months.



From above plot also, we can see that December has the highest sales across years.

Decompose the Time Series:

Additive Decomposition –



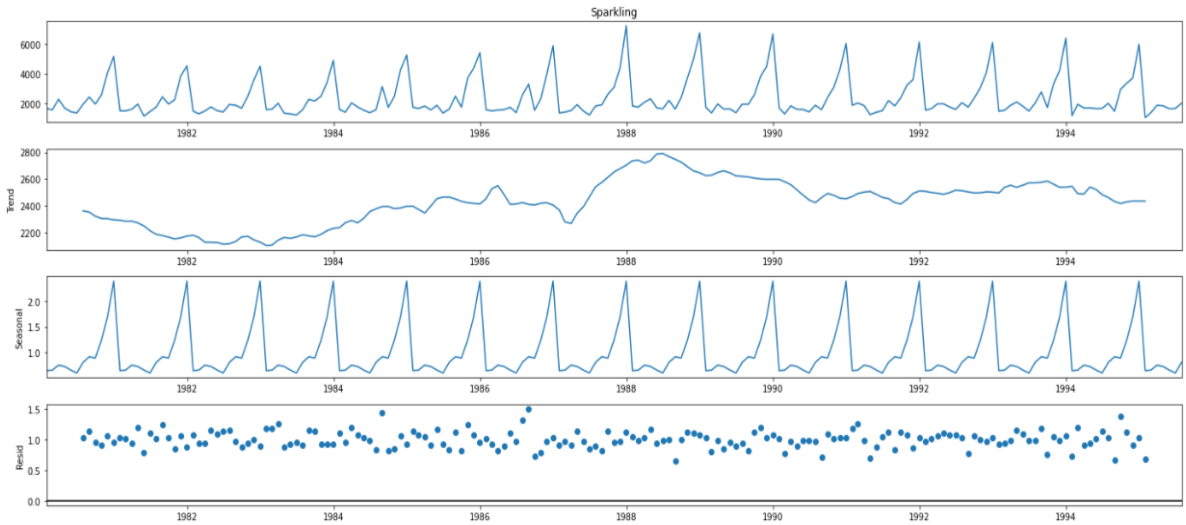

```
Trend
 Time_Stamp
1980-01-31    NaN
1980-02-29    NaN
1980-03-31    NaN
1980-04-30    NaN
1980-05-31    NaN
Name: trend, dtype: float64
```

```
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
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
Name: resid, dtype: float64
```

As per the 'additive' decomposition, we see that there is a increased trend in the earlier years of the data and decreased trend in latest years. There is a seasonality as well. We see that the residuals are located around 0 from the plot of the residuals in the decomposition.

Multiplicative Decomposition:




```
Trend
 Time_Stamp
1980-01-31    NaN
1980-02-29    NaN
1980-03-31    NaN
1980-04-30    NaN
1980-05-31    NaN
Name: trend, dtype: float64
```

```
Seasonality
 Time_Stamp
1980-01-31    0.649843
1980-02-29    0.659214
1980-03-31    0.757440
1980-04-30    0.730351
1980-05-31    0.660609
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
Name: resid, dtype: float64
```

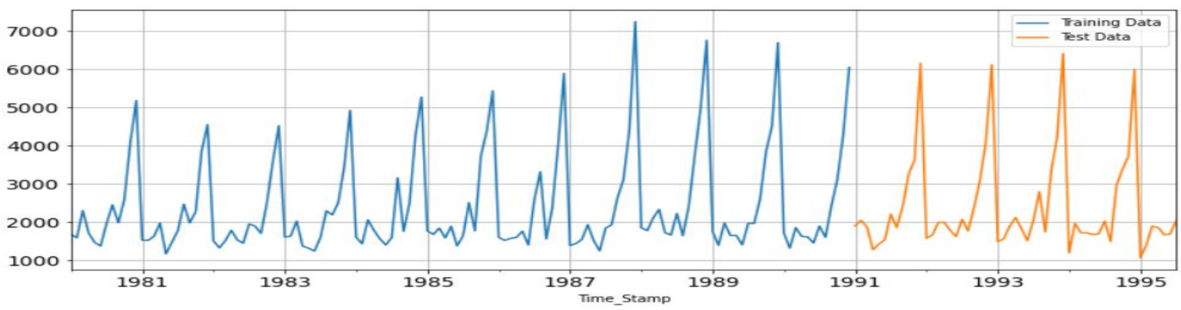
As per the 'Multiplicative' decomposition, we see that there is a increased trend in the earlier years of the data and decreased trend in latest years. There is a seasonality as well. We see that the residuals are located around 1 from the plot of the residuals in the decomposition.

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

```
Shape of Training Data
 (132, 1)

Shape of Testing Data
 (55, 1)
```

```
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
```

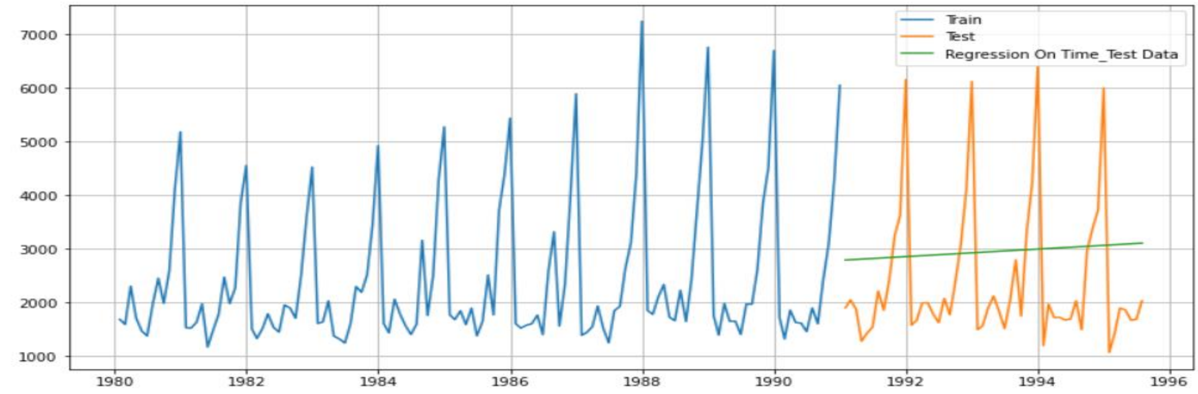
```
First few rows of Test Data
             Sparkling
Time_Stamp
1991-01-31        1902
1991-02-28        2049
1991-03-31        1874
1991-04-30        1279
1991-05-31        1432

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
```



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

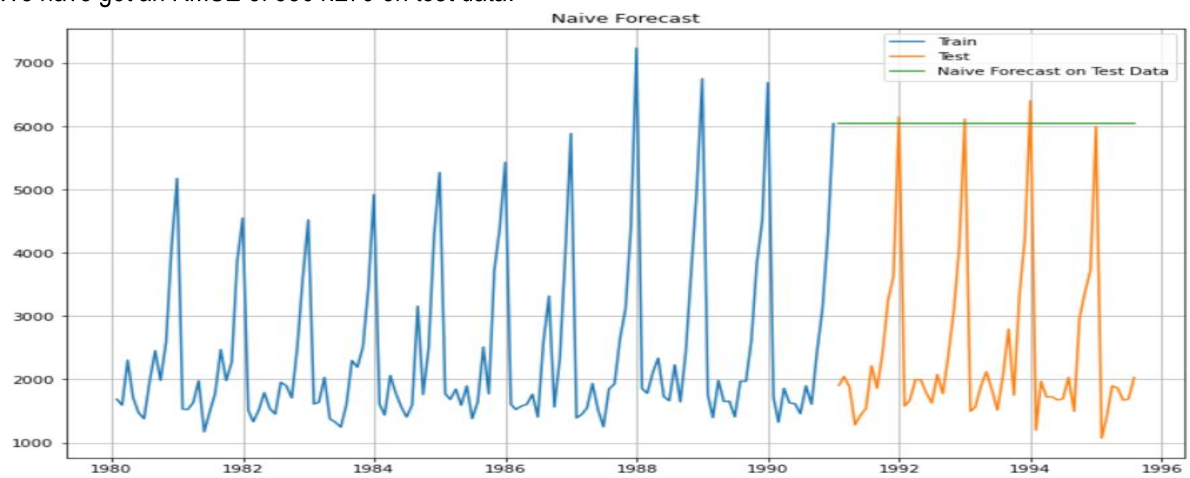
<u>Linear Regression Model :</u>

We have got an RMSE of 1389.135 on test data.



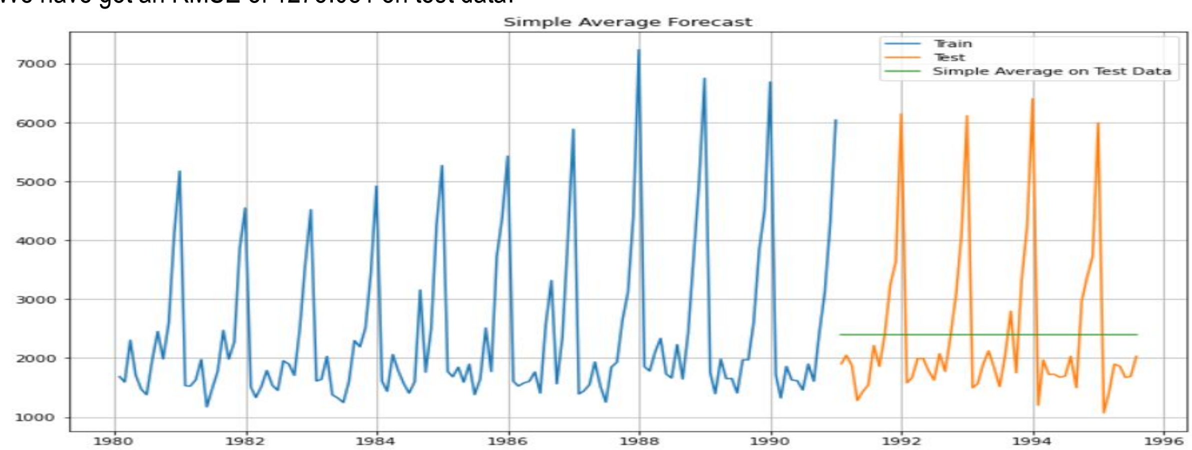
<u>Naive Model:</u>

We have got an RMSE of 3864.279 on test data.



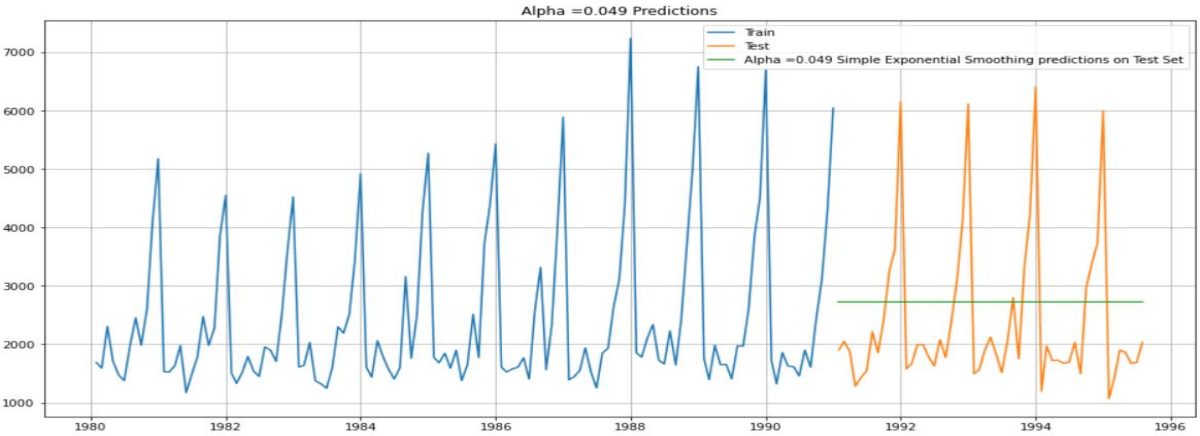
<u>Simple Average Method :</u>

We have got an RMSE of 1275.081 on test data.



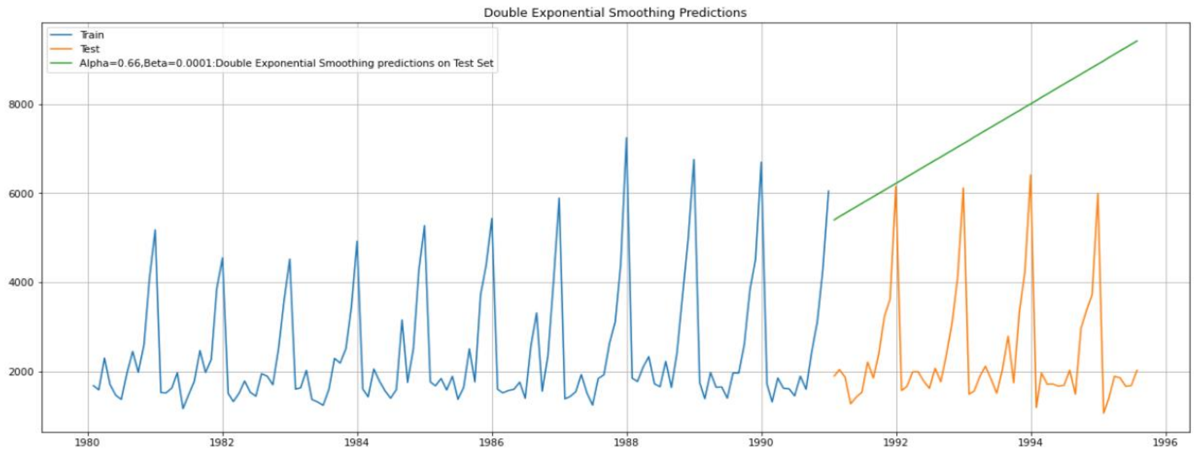
Simple Exponential Smoothing:

We have got an RMSE of 1316.034 on test data.



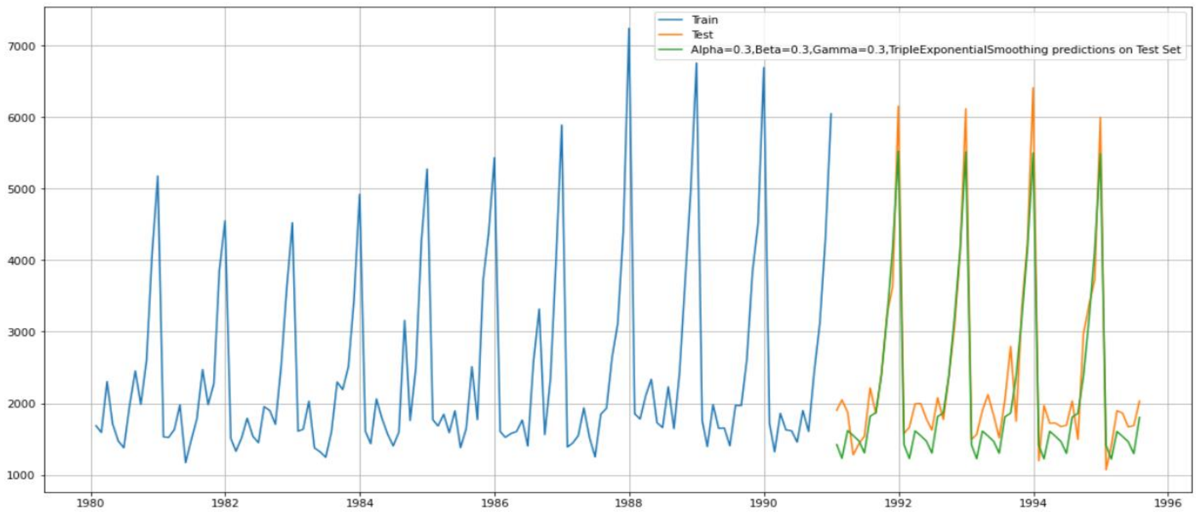
Double Exponential Smoothing :

We have got an RMSE of 5291.879 on test data.



Triple Exponential Smoothing:

We have got an RMSE of 392.786 on test data.



Comparing RMSE values for all the above three models , we have got the following table

	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
Alpha=0.049,SimpleExponentialSmoothing	1316.034674
Alpha=0.66,Beta=0.0001:Double Exponential Smoothing	5291.879833
Alpha=0.111,Beta=0.061,Gamma=0.395,TripleExponentialSmoothing	469.593384
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.786198

We have built several models got an idea as to which particular model gives us the least error on our test set for this data. As the dataset has both trend and seasonality , Triple Exponential Smoothing works best with this model among all the above models.

1.5. 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.

Augmented Dickey –Fuller test is used to test whether a time is non-stationary.

Null hypothesis Ho : Time series is non stationary
Alternative hypothesis Ha : Time series is stationary.
Rejection of null hypothesis implies that the series is stationary.

For the dataset, we have following results :

```
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
```

As the p-value is greater than 0.05 , we fail to reject the null hypothesis. So the time series is non stationary.
Let us take a difference of order 1 and check whether the Time Series is stationary or not.

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

We see that after the difference of order 1 , the time series is stationary.

1.6. 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.

As the data shows seasonality , we use SARIMA model on the training data.

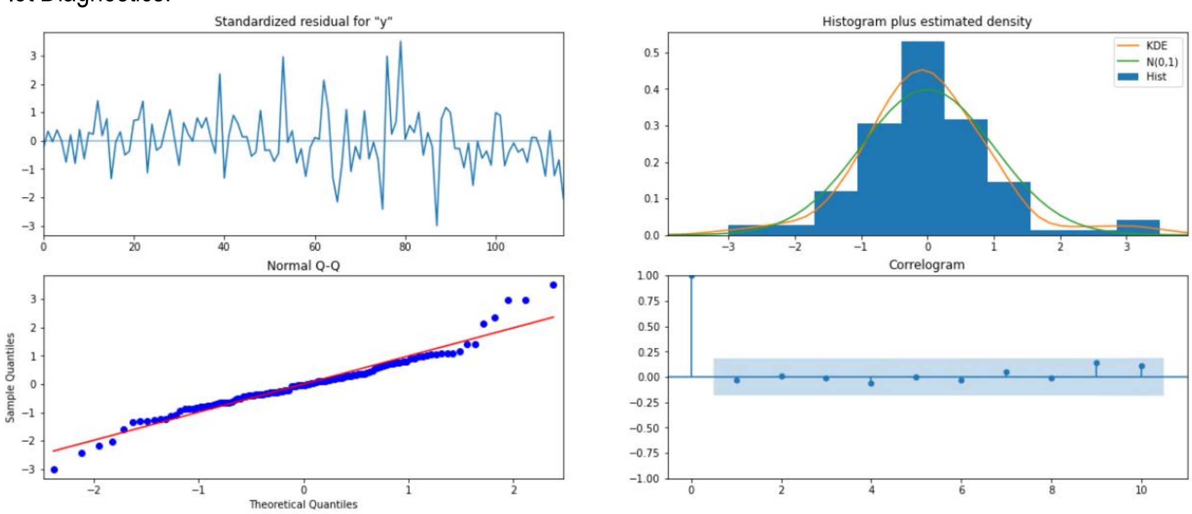
Seasonality as 6 for the model , we have got lowest AIC for on the training data for the model with paramaters

param	seasonal	AIC
(1, 1, 2)	(2, 0, 2, 6)	1727.670866
(0, 1, 2)	(2, 0, 2, 6)	1727.888818
(2, 1, 2)	(2, 0, 2, 6)	1729.192582
(0, 1, 1)	(2, 0, 2, 6)	1741.641478
(1, 1, 1)	(2, 0, 2, 6)	1743.379778

SARIMAX 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, 20 Dec 2020   AIC                           1727.671
Time:                                    17:24:50   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
===================================================================================
```

Plot Diagnostics:



We have an RMSE of value **626.898** on test data

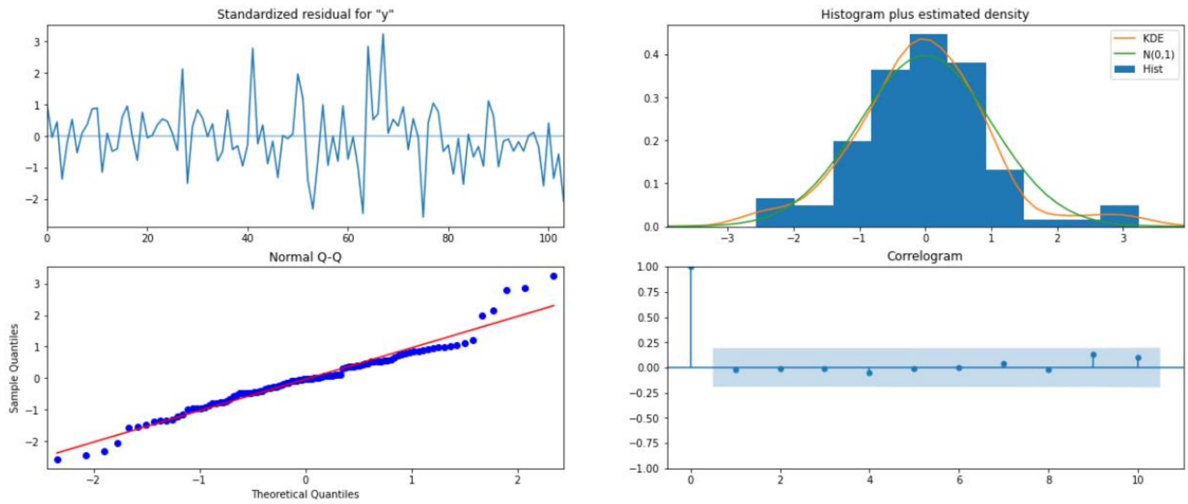
Seasonality as 12 for the model , we have got lowest AIC for on the training data for the model with paramaters

param	seasonal	AIC
(1, 1, 2)	(1, 0, 2, 12)	1555.584247
(1, 1, 2)	(2, 0, 2, 12)	1555.929659
(0, 1, 2)	(2, 0, 2, 12)	1557.121564
(0, 1, 2)	(1, 0, 2, 12)	1557.160507
(2, 1, 2)	(1, 0, 2, 12)	1557.340402

SARIMAX 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, 20 Dec 2020   AIC                           1555.584
Time:                                    17:31:56   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
===================================================================================
```

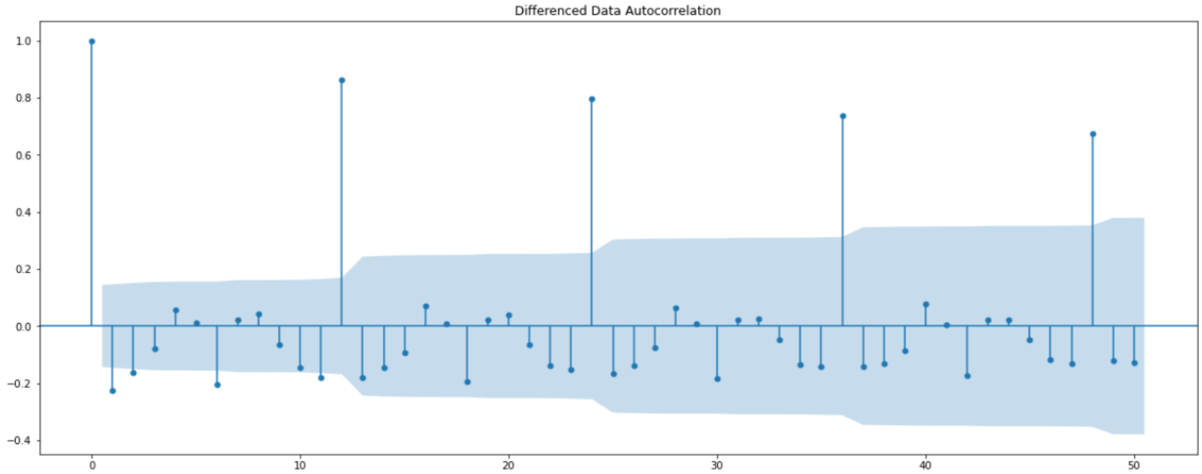
Plot Diagnostics:

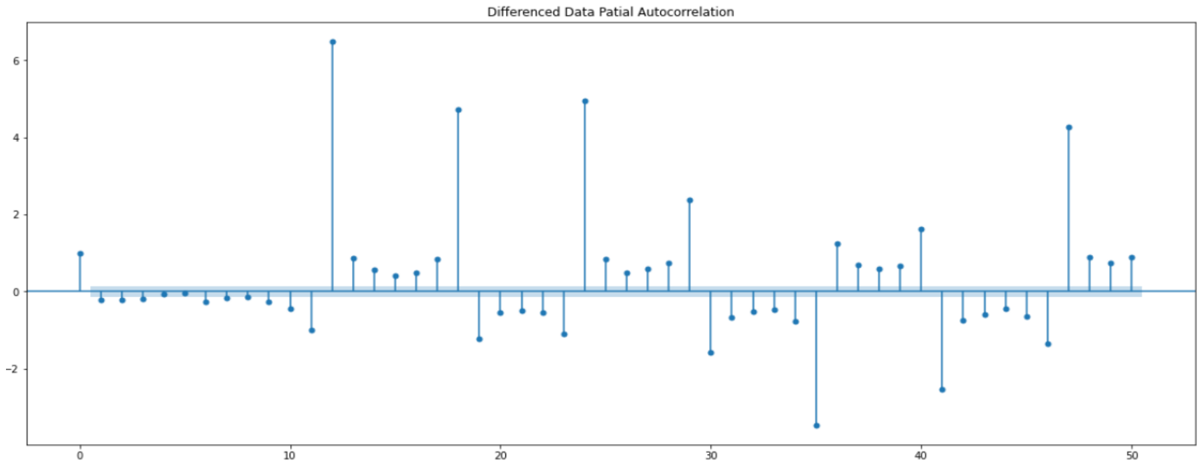


We have an RMSE of value **528.621**on the test data

1.7. 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.

We have got the following ACF and PACF plots

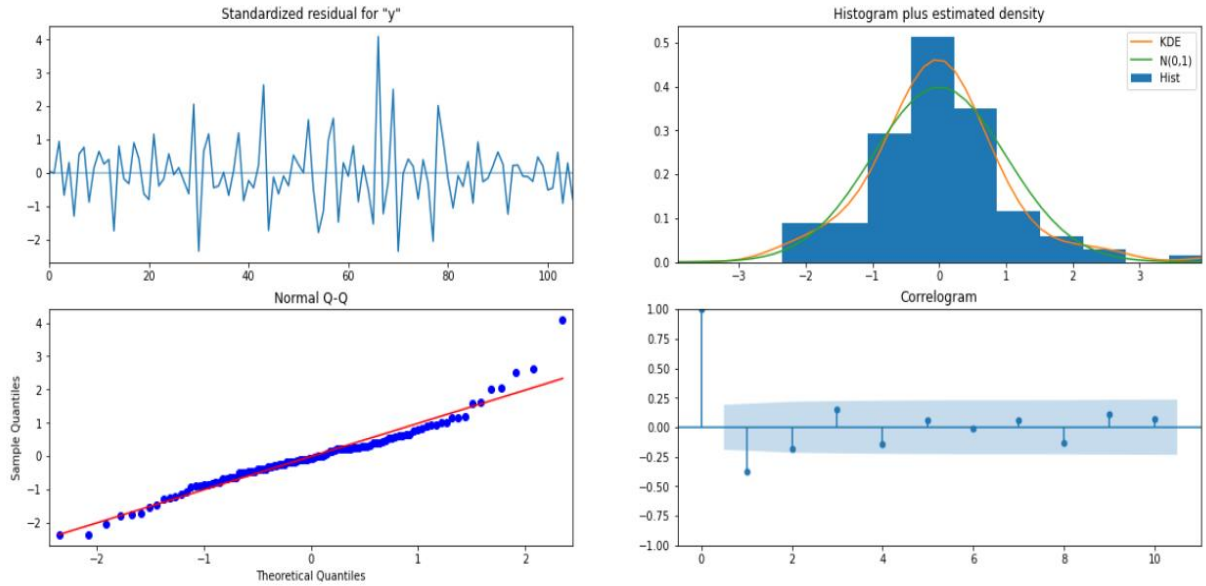




SARIMAX results:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                     y   No. Observations:               132
Model:             SARIMAX(0, 1, 0)x(1, 1, [1, 2, 3], 6)   Log Likelihood              -811.726
Date:                               Sun, 20 Dec 2020   AIC                        1633.452
Time:                                       17:41:11   BIC                        1646.770
Sample:                                            0   HQIC                       1638.850
                                               - 132
Covariance Type:                                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.S.L6       -1.0176      0.015    -68.689      0.000      -1.047      -0.989
ma.S.L6        0.0335      0.176      0.190      0.849      -0.312       0.379
ma.S.L12      -0.4660      0.081     -5.772      0.000      -0.624      -0.308
ma.S.L18       0.0764      0.164      0.465      0.642      -0.246       0.399
sigma2      2.608e+05   2.85e+04      9.148      0.000    2.05e+05    3.17e+05
===================================================================================
Ljung-Box (L1) (Q):                  15.59   Jarque-Bera (JB):                33.69
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.72   Skew:                             0.68
Prob(H) (two-sided):                  0.34   Kurtosis:                         5.41
===================================================================================
```

Plot Diagnostics:



We have got an RMSE value 1914.95 on the test data

1.8. 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.

We can summarize the results of all the different models through the following table:

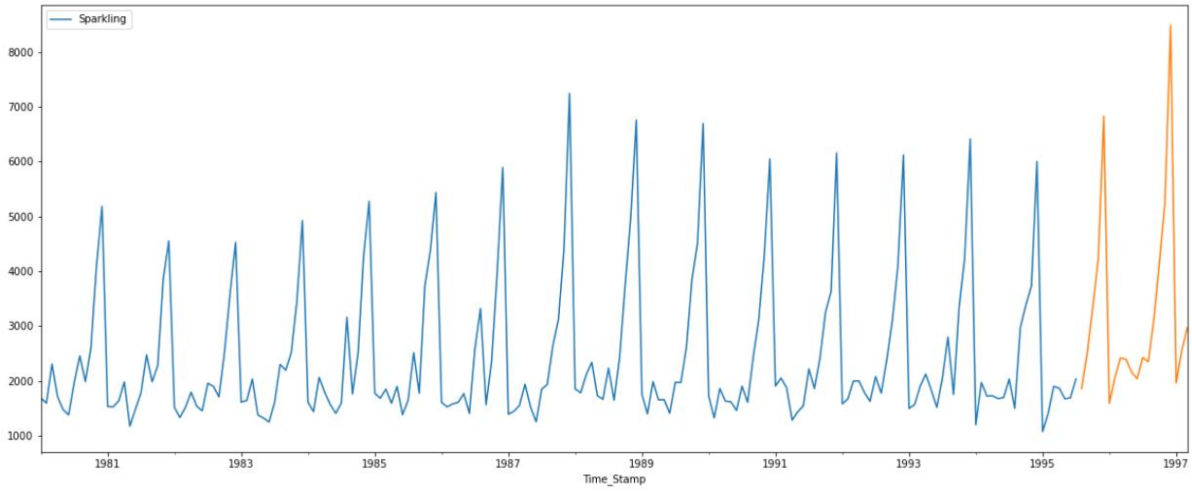
	Test RMSE
RegressionOnTime	1389.135175
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
Alpha=0.049,SimpleExponentialSmoothing	1316.034674
Alpha=0.66,Beta=0.0001:Double Exponential Smoothing	5291.879833
Alpha=0.111,Beta=0.061,Gamma=0.395,TripleExponentialSmoothing	469.593384
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	392.786198
SARIMA(1,1,2)(2,0,2,6)	626.898233
SARIMA(1,1,2)(1,0,2,12)	528.621309
SARIMA(0,1,0)(1,1,3,6)	1914.957852

From above table, we can see that Triple Exponential Smoothing with Alpha = 0.3,Beta = 0.3 and Gamma = 0.3 has the lowest Test RMSE of value 392.786

1.9. 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.

As the Triple Exponential Smoothing with Alpha = 0.3,Beta = 0.3 and Gamma = 0.3 has the lowest Test RMSE of value 392.786, we use this model to prediction.

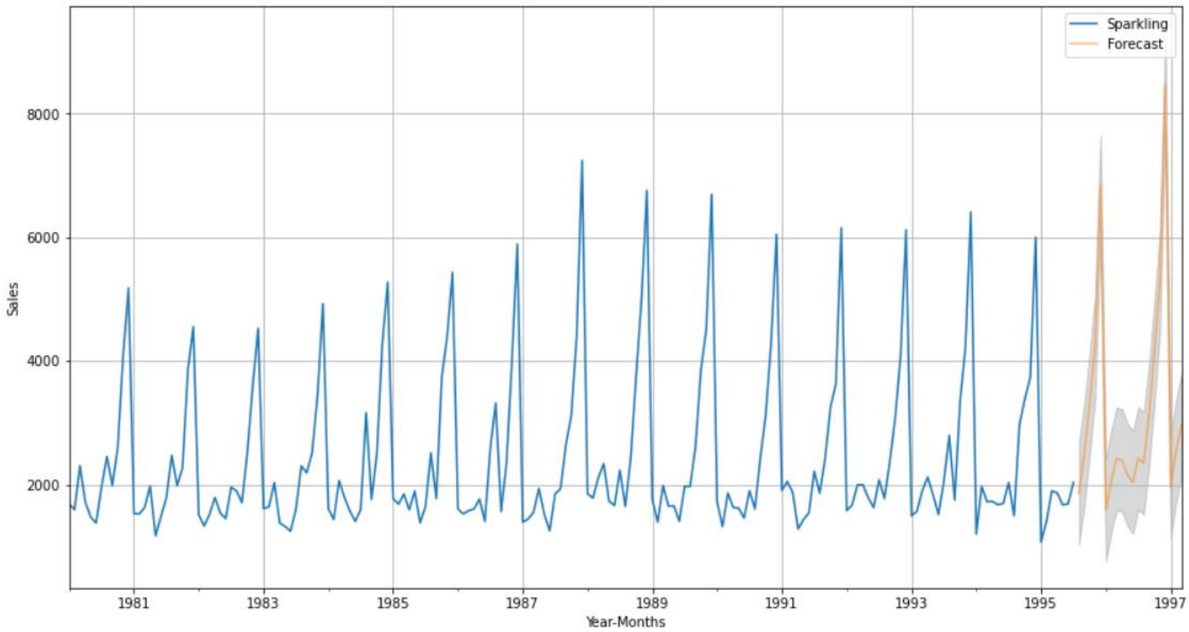
This model gives RMSE of **422.284** on the full data.



Confidence bands for prediction:

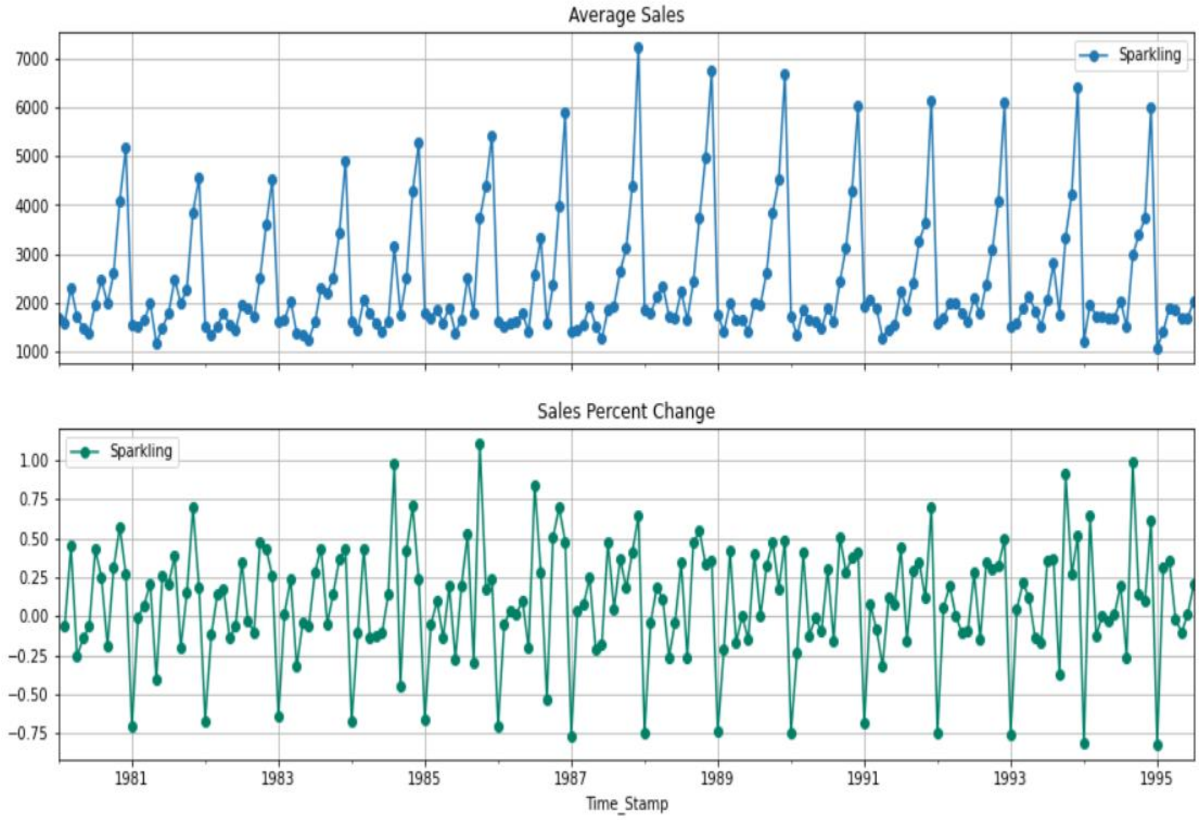
lower_CI	prediction	upper_ci
1025.541131	1855.439826	2685.338521
1656.976007	2486.874702	3316.773397
2493.261044	3323.159740	4153.058435
3395.278907	4225.177602	5055.076298
5998.104872	6828.003567	7657.902262

Forecast along with the confidence band :

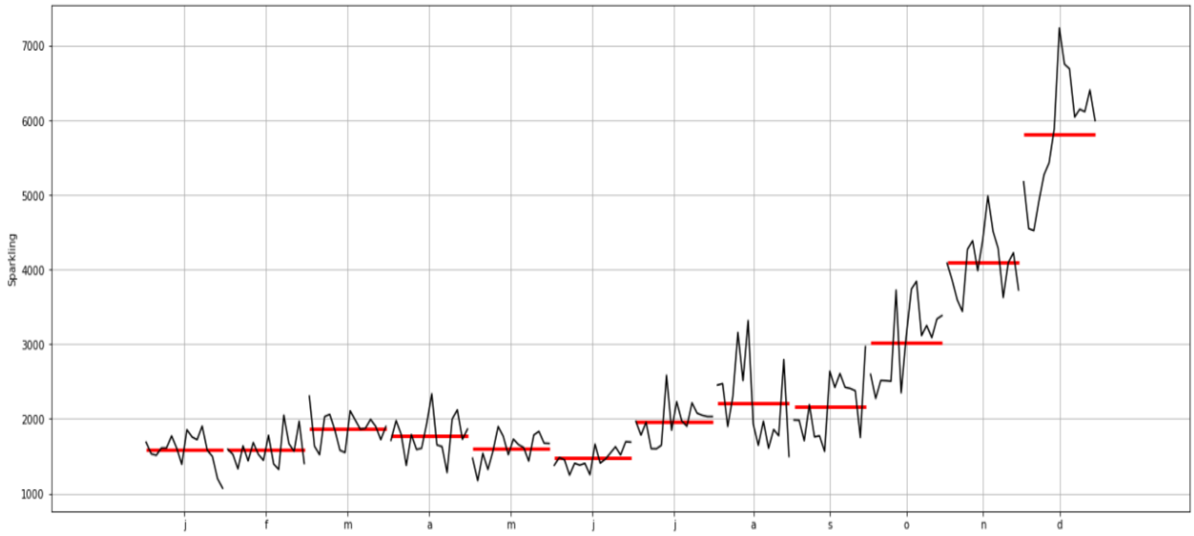


1.10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

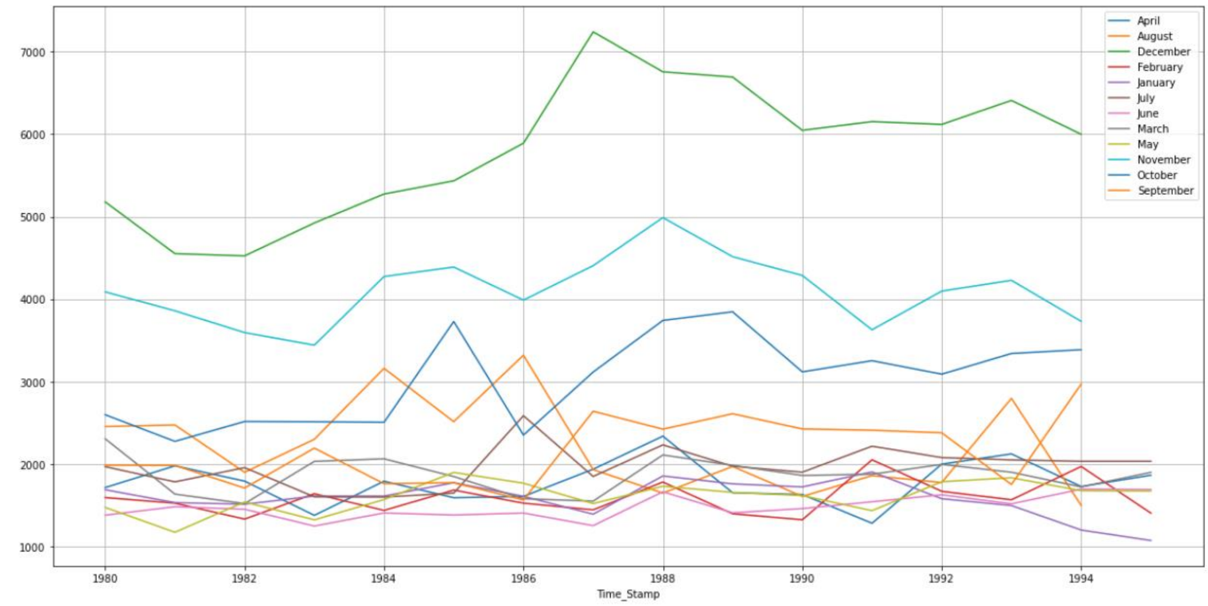
The Triple Exponential Smoothing model with parameters Alpha = 0.3,Beta = 0.3 and Gamma = 0.3 will be helpful in making best forecasts for the given time series data.



The average sales are almost showing stable values year by year from 1981 to 1995 and also sales percent change is higher at the start and end of year.



Looking at sales for different months across all the years, there is minimal change in sales in july month and maximum change in august and December months. The company can use this data in maintain stocks for the respective periods based on the demand.



From the plot we can see that sales are higher in December month for all the years.

This could be because off festival events such as Christmas & New year. So the company can increase sales in this month by increasing sales qtys and providing any offers to attract more wine consumers.

Therefore from the above forecasting values based on the trend and seasonality of the time series data, the company can make best decisions in increasing it sales.