PROJECT: TS FORECASTING

MODULE:7

Manikandan Thiyagarajan

Table of Content

	simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.	
2.5	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.	48
2.6	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.	49
2.7	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.	52
2.8	Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.	58
2.9	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.	58
2.10	Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.	60

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. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.
Data set for the Problem: Sparkling.csv and Rose.csv

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

First 5 rows:

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

Last 5 rows:

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

Info:

```
<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
```

Missing value: No missing value in Dataset.

Create Time stamp:

```
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')
```

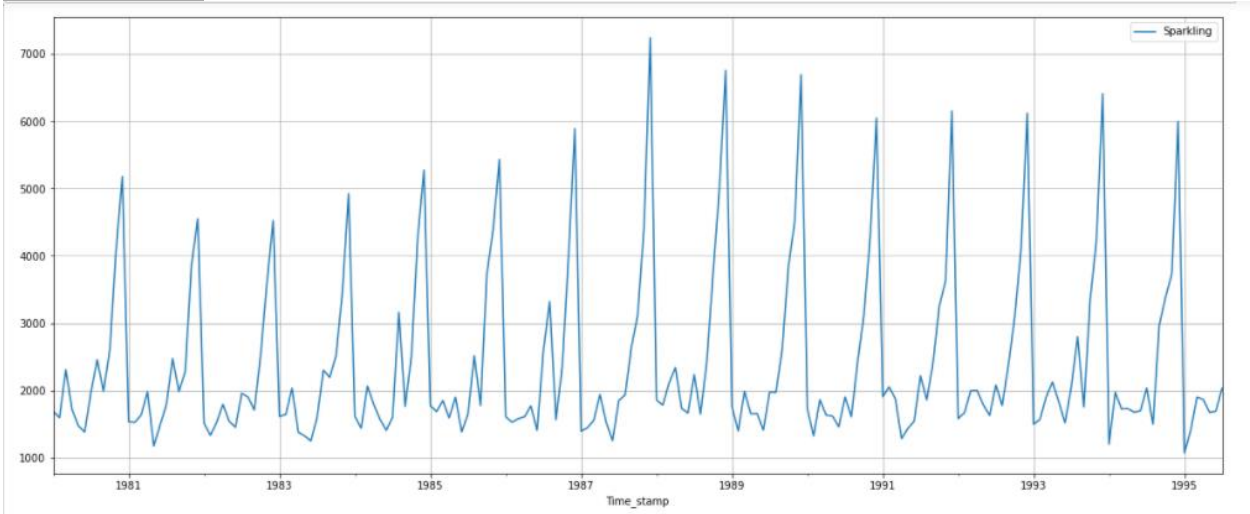
Time Stamp added to Data frame:

	YearMonth	Sparkling	Time_stamp
0	1980-01	1686	1980-01-31
1	1980-02	1591	1980-02-29
2	1980-03	2304	1980-03-31
3	1980-04	1712	1980-04-30
4	1980-05	1471	1980-05-31

Timestamp made as index:

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

Plot the Dataset:

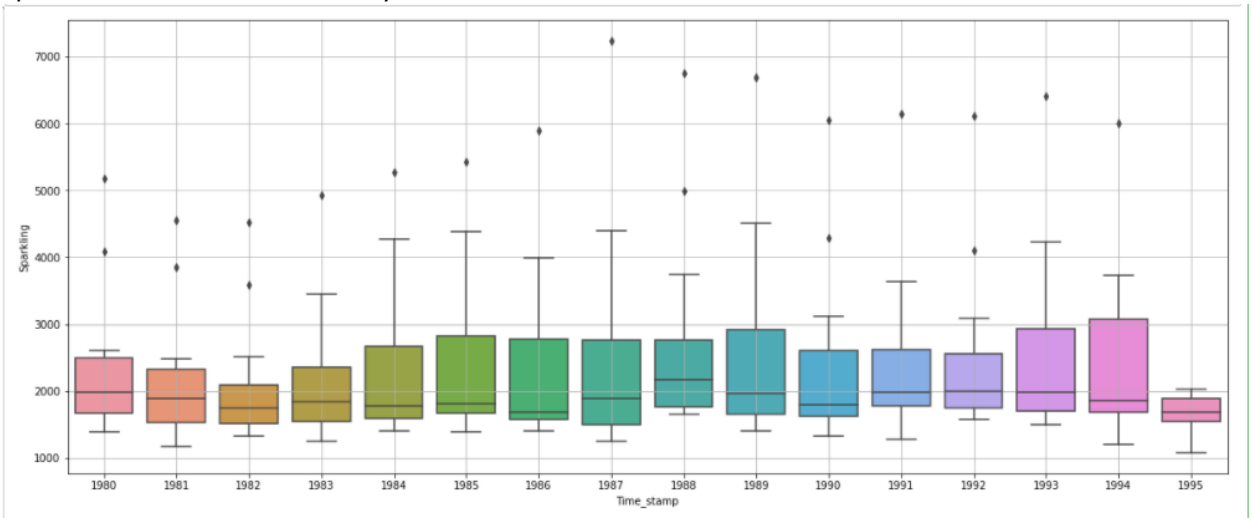


Dataset Summary:

	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

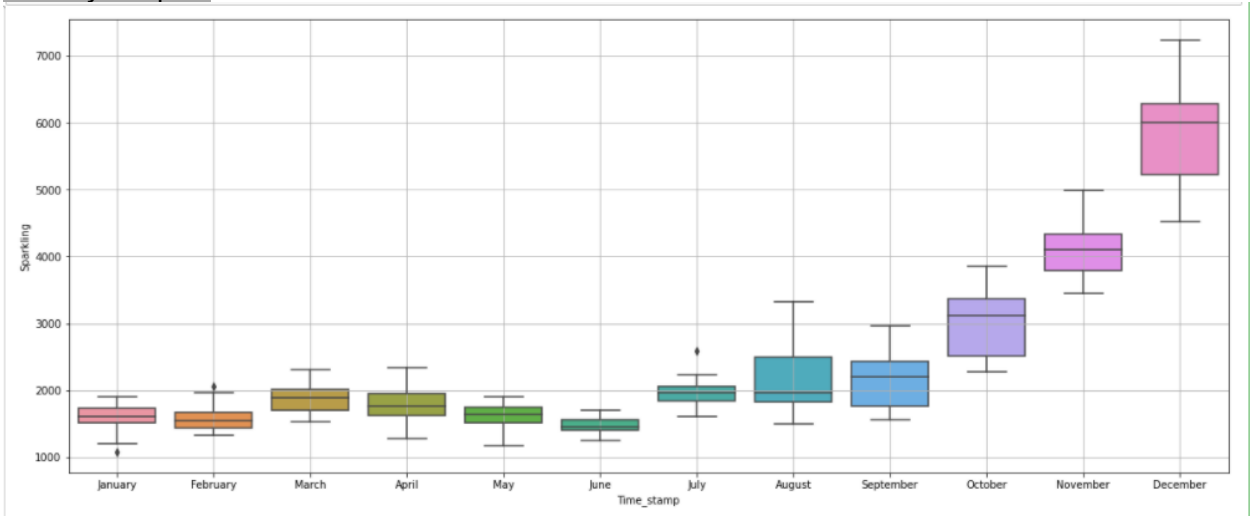
1.2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Spread of sales across different years.



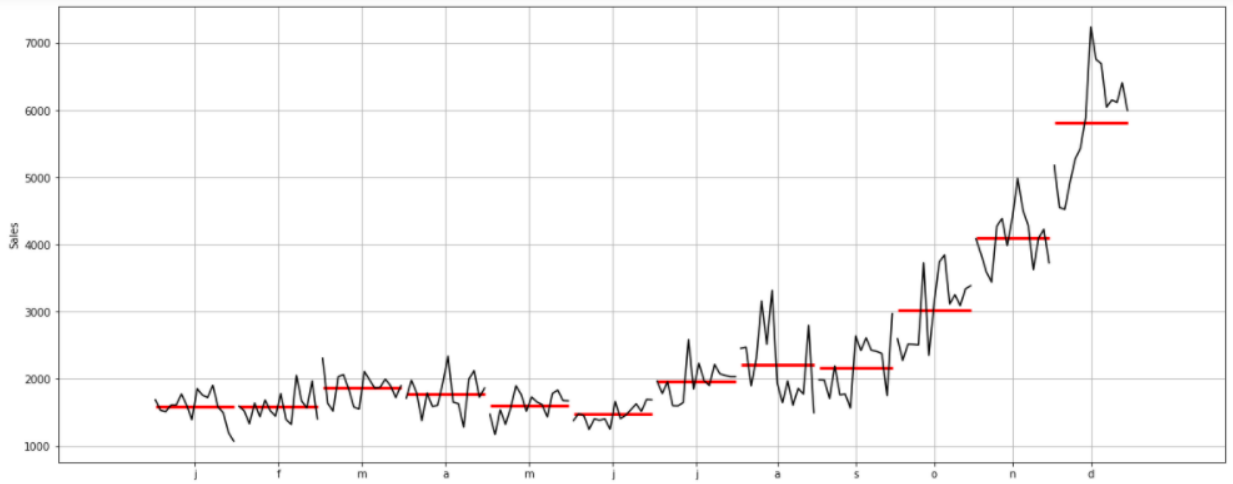
From year plot, we can't say much about trend because median value across the years are same.

Monthly Boxplot:



From monthly plot it is evident that sales are increasing from July to December then slightly decreases from January to June. June month records the lowest sales across the years.

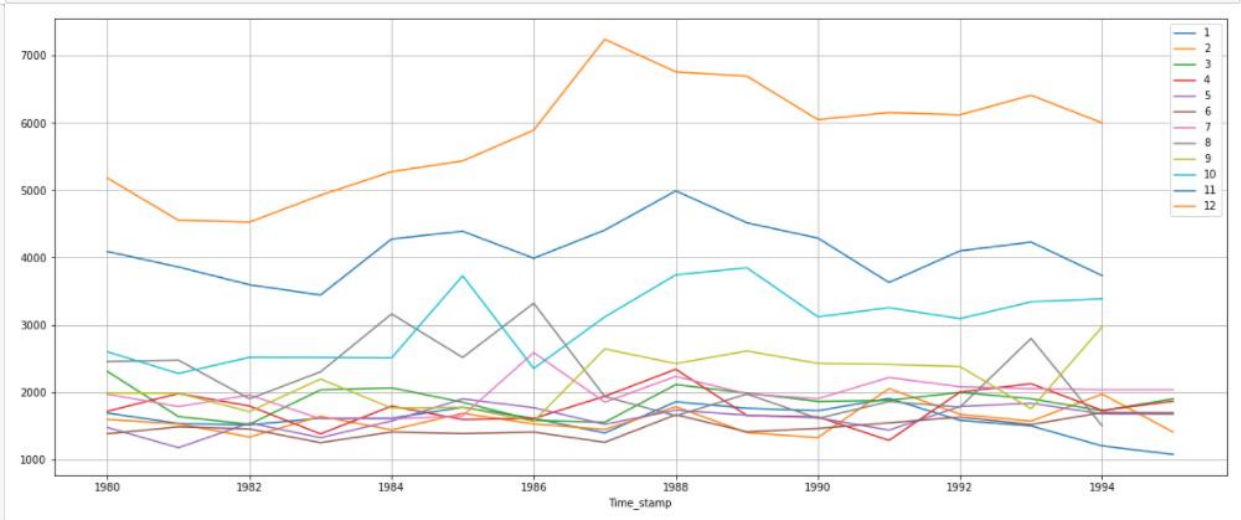
Monthly sales within different months across years.



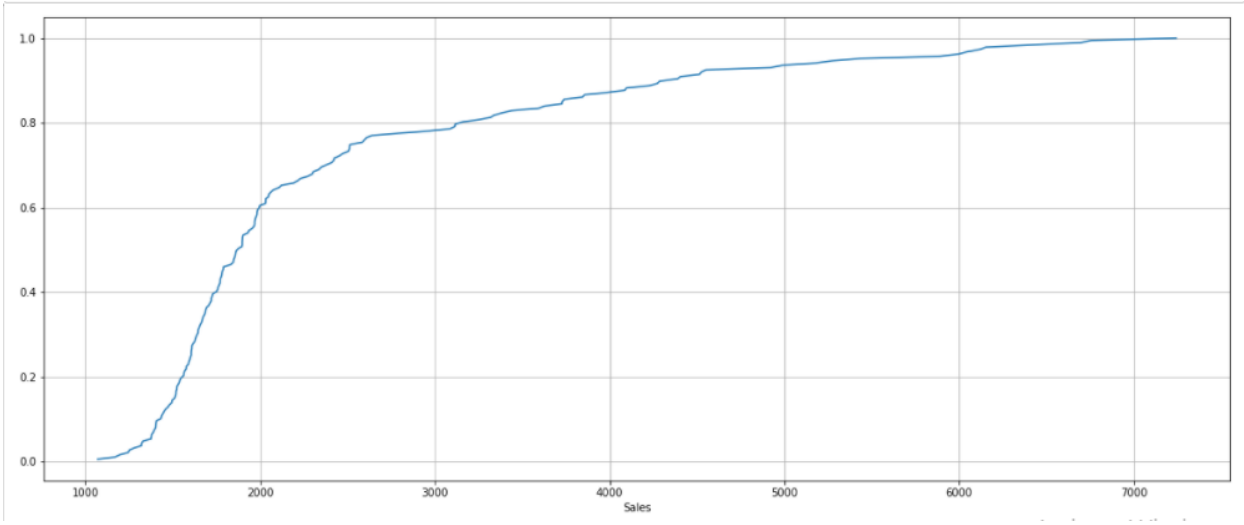
This plot shows us behavior of time series across months.

Plot a graph of monthly sales

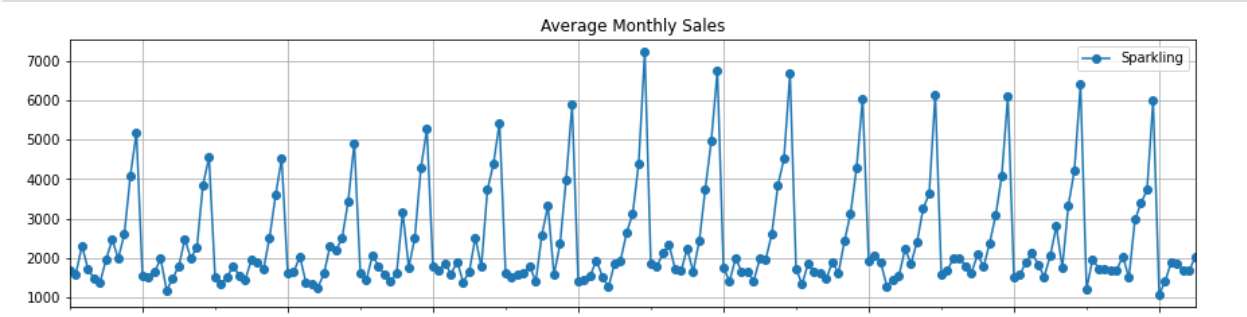
Time_stamp	1	2	3	4	5	6	7	8	9	10	11	12
Time_stamp												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	1510.0	1329.0	1518.0	1790.0	1537.0	1449.0	1954.0	1897.0	1706.0	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

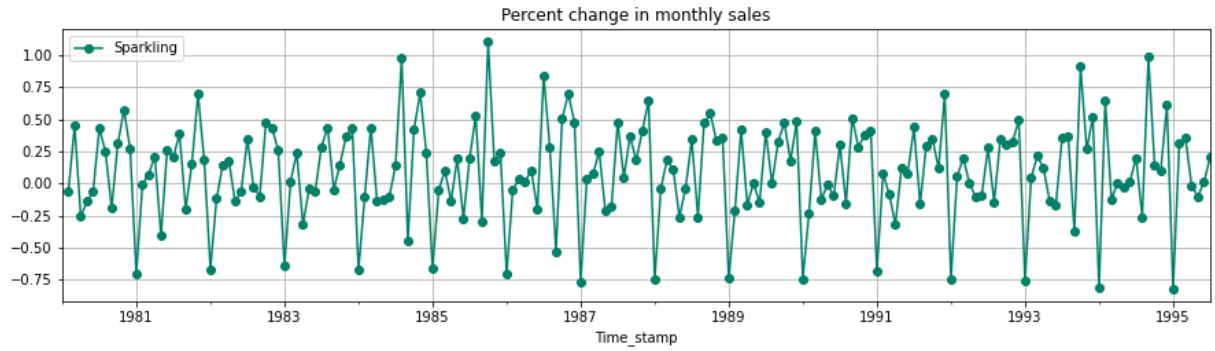


Empirical cumulative distribution:

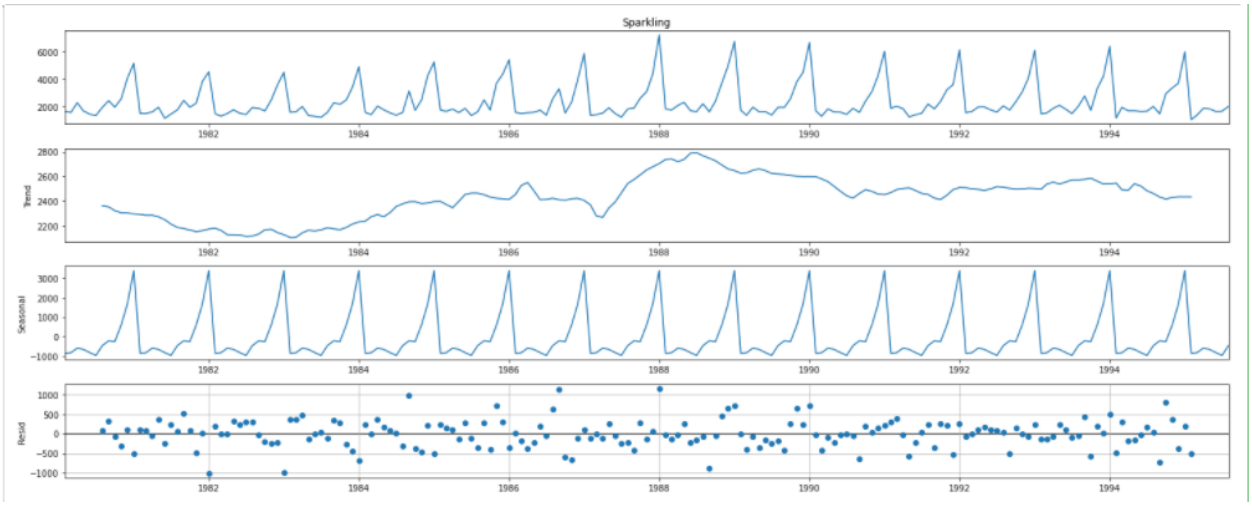


This graph tells us what percentage datapoint refers to what number of sales. Only 20% of sales is between 3000 to 6000.



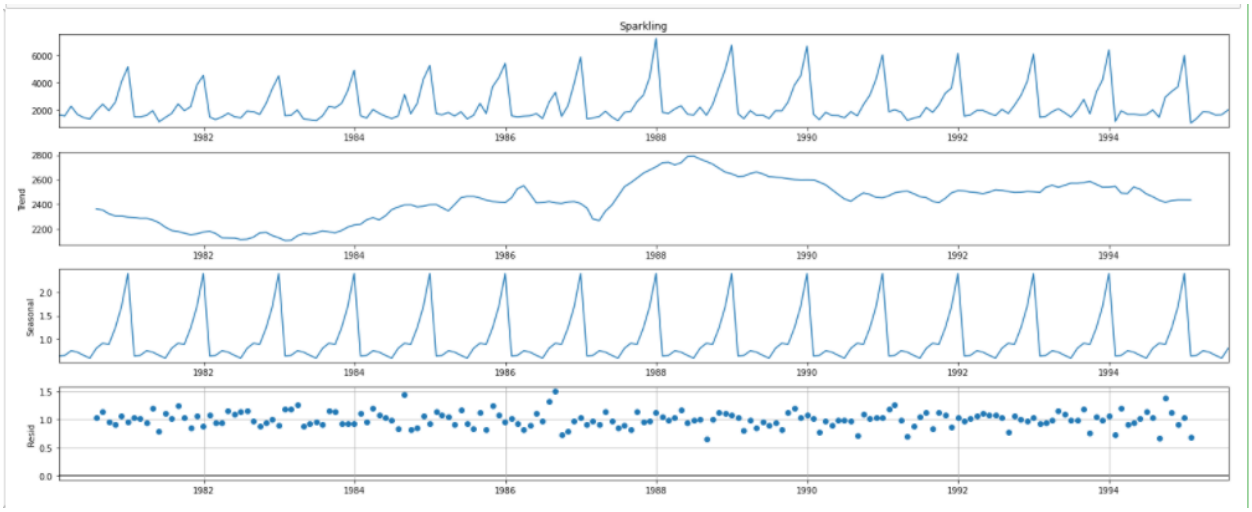


Additive decomposition:



From the above plot trend is not so clear, it is fluctuating (increasing, decreasing then increasing). A strong seasonality is observed. Residual has some patterns. Multiplicative model will be analysed before taking any decisions.

Multiplicative decomposition:



Multiplicative model's trend and seasonality are same. Residual has some patterns. so additive model is considered for further analysis.

```
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
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
Name: seasonal, dtype: float64
```

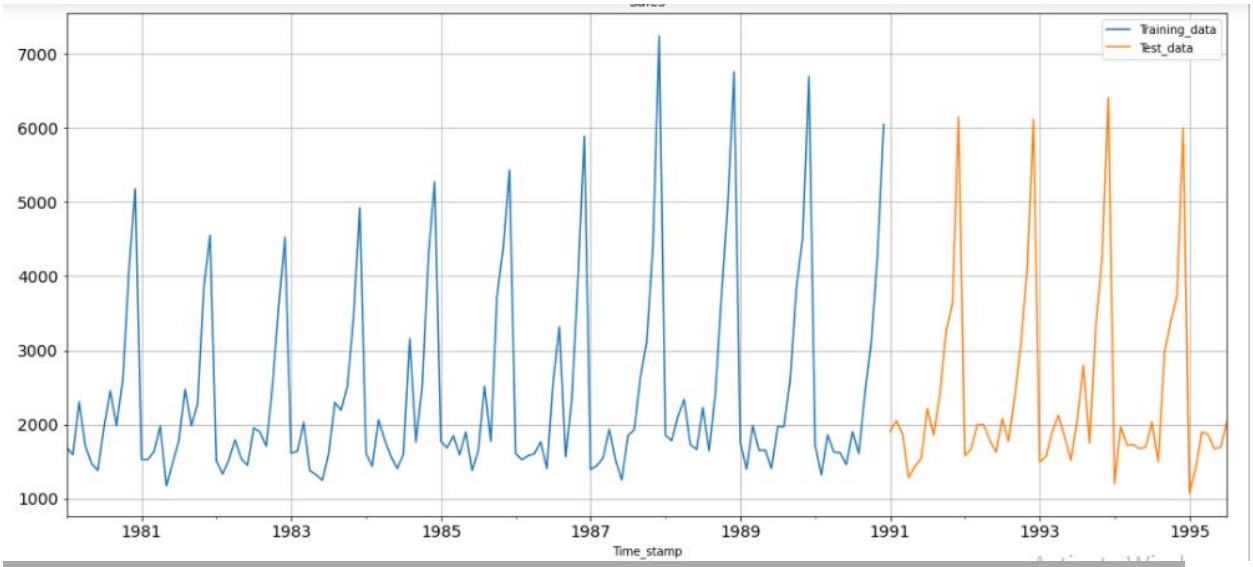
```
residual
 Time_stamp
1980-01-31           NaN
1980-02-29           NaN
1980-03-31           NaN
1980-04-30           NaN
1980-05-31           NaN
1980-06-30           NaN
1980-07-31     70.835599
1980-08-31    315.999487
1980-09-30    -81.864401
1980-10-31   -307.353290
Name: resid, dtype: float64
```

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

Dataset was split between Training and test dataset. Test dataset should start from 1991.

Shape of training and test dataset:

```
(132, 1)
(55, 1)
```



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 and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

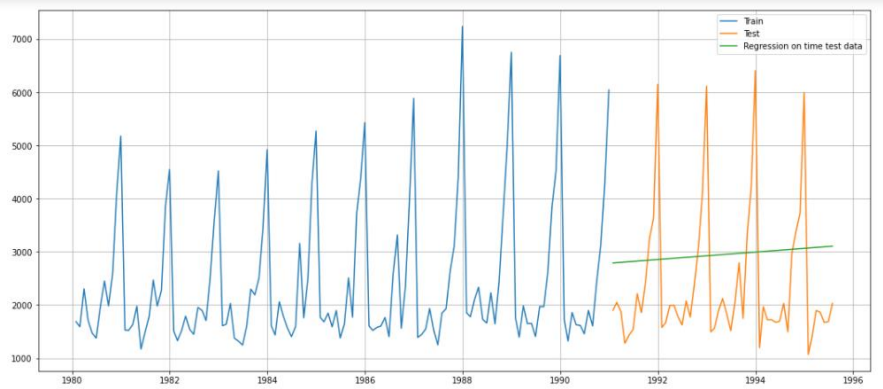
Model 1 : Linear Regression

For this particular linear regression, we are going to regress the 'Sales' variable against the order of the occurrence. For this we need to modify our training data 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]
```

Copy separate training and test data for Linear Regression.

Create variable for Linear regression. Fit the data and predict it. Plot the regression on training and test set.

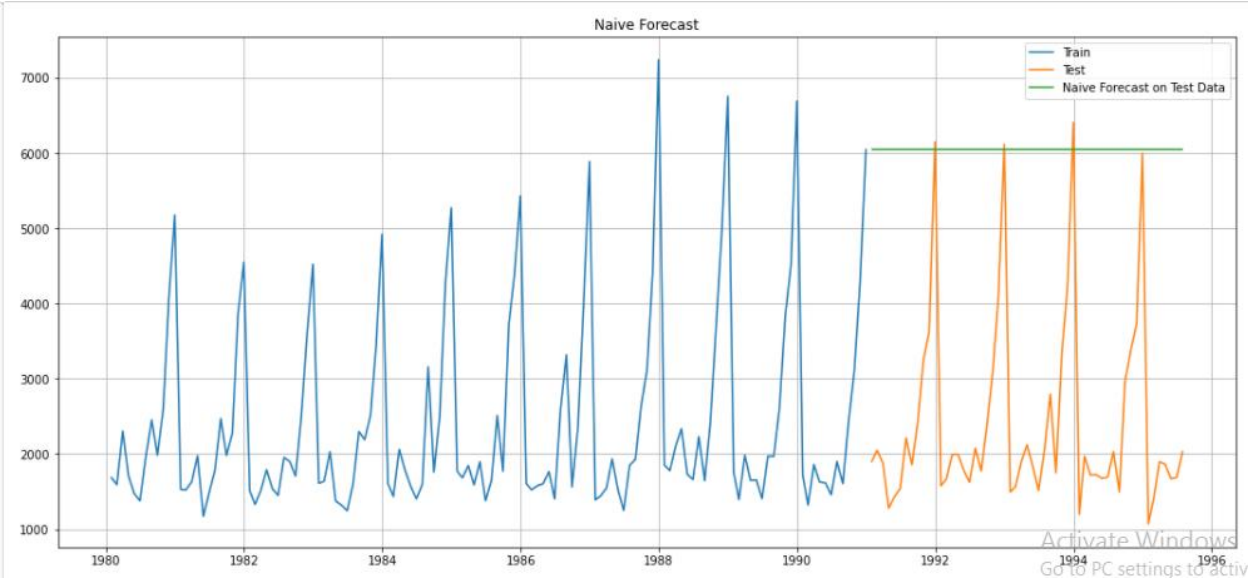


For Regression On Time forecast on the Test Data, RMSE is 1389.135

Model 2: Naive Approach: $\hat{y}\ t+1=yt$

For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

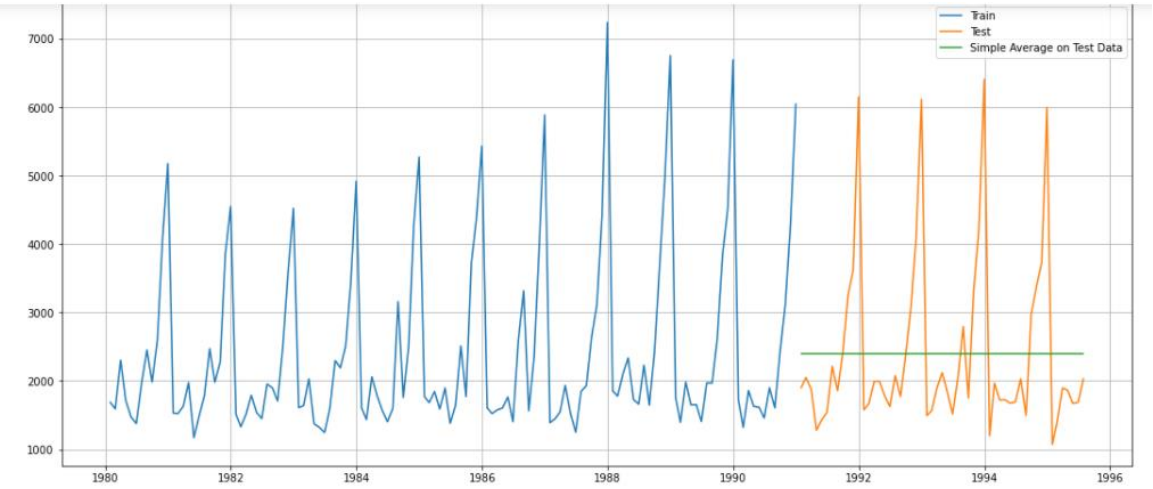
Copy separate Training and test data for Naïve approach.
Fit , predict and plot it.



For Naive forecast on the Test Data RMSE is 3864.279

Model 3: Simple Average

For this particular simple average method, we will forecast by using the average of the training values.
Copy separate Training and test data for Simple Average.
Fit , predict and plot it.



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

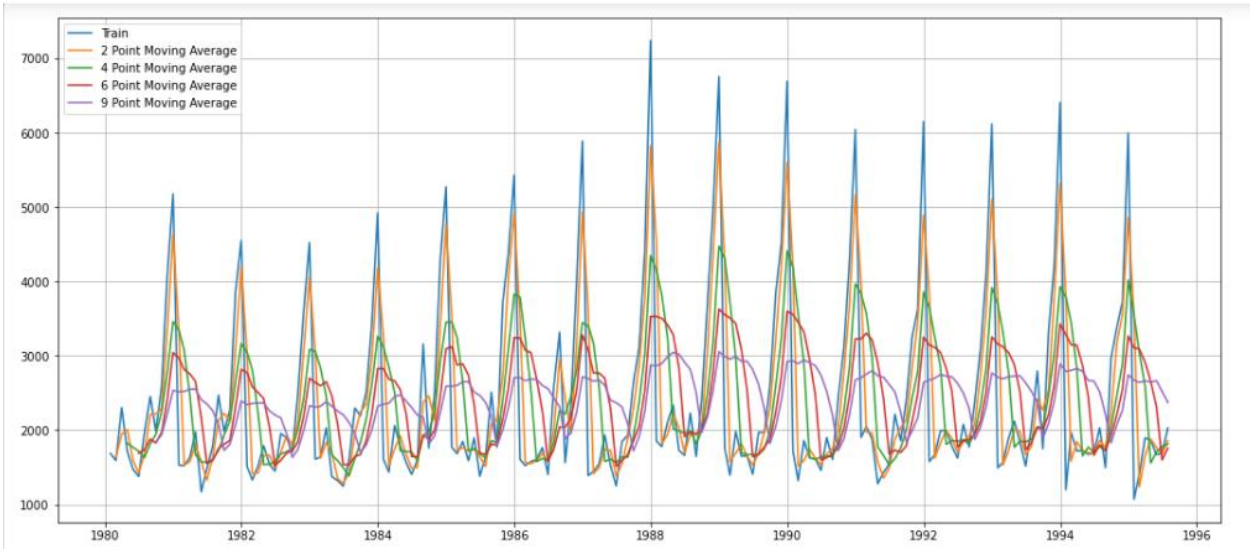
Model 4: Moving Average(MA)

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

Copy separate dataset for moving Average.

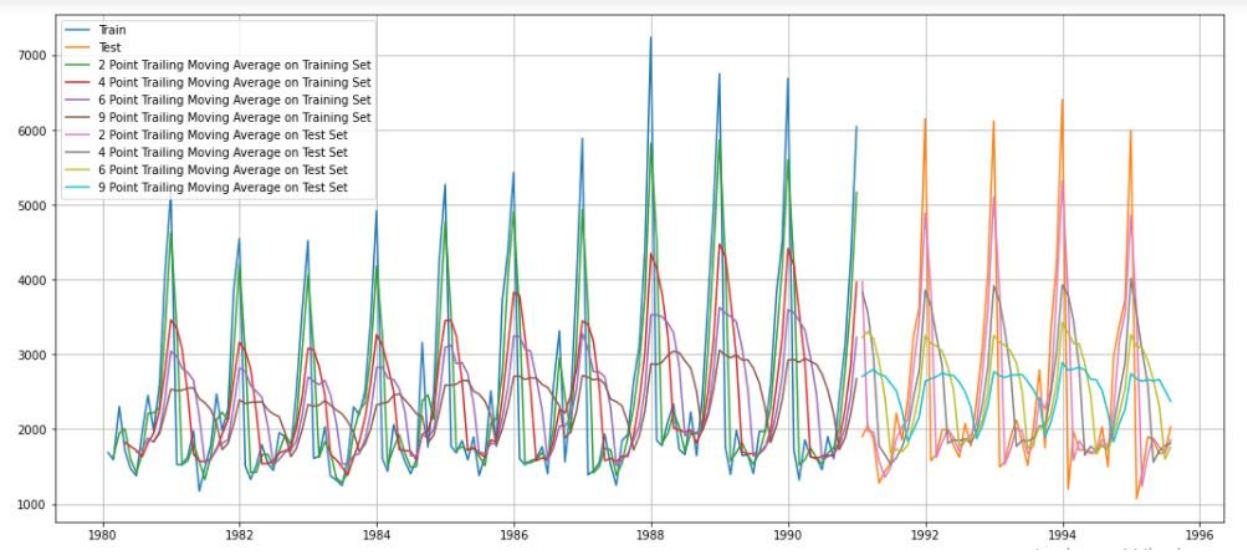
Trailing moving averages

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



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.

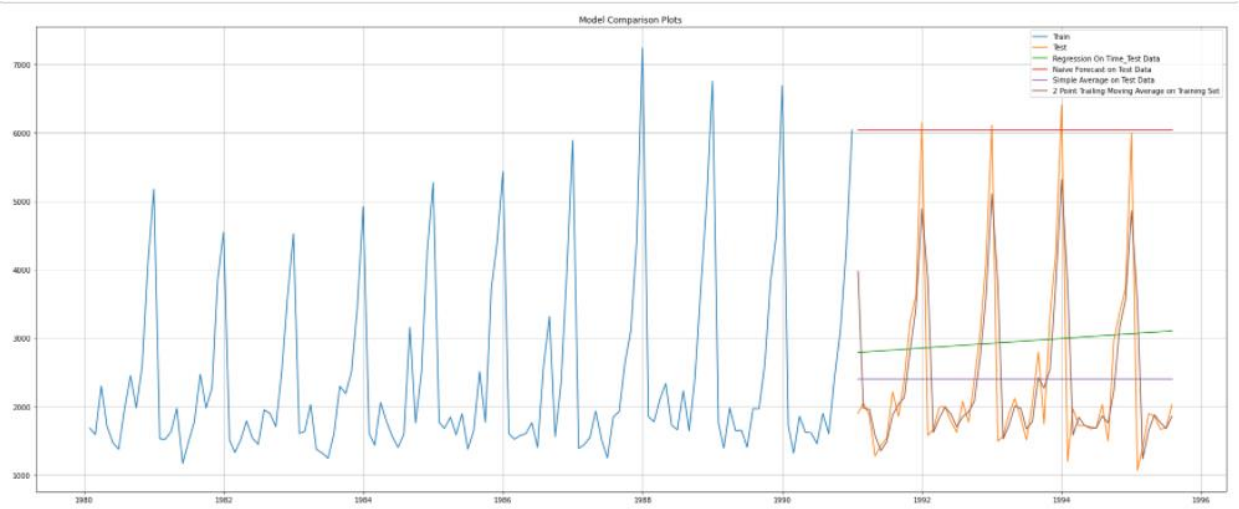
Create training and test dataset.



Model Evaluation done only on the test data.

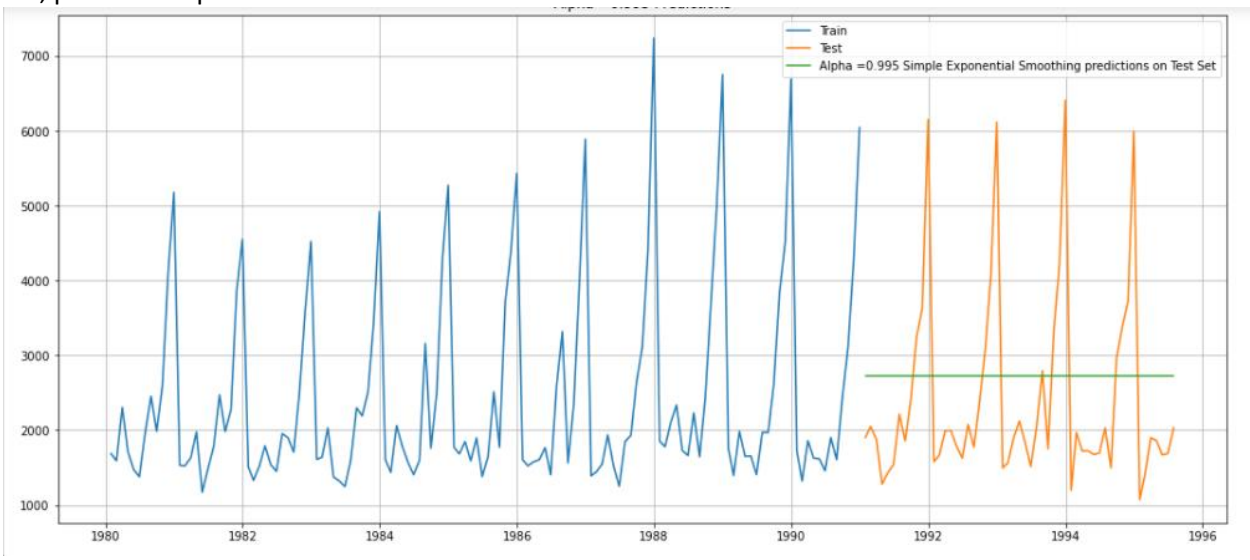
For 2 point Moving Average Model forecast on the Training Data, RMSE is 813.401
For 4 point Moving Average Model forecast on the Training Data, RMSE is 1156.590
For 6 point Moving Average Model forecast on the Training Data, RMSE is 1283.927
For 9 point Moving Average Model forecast on the Training Data,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

Before we go on to build the various Exponential Smoothing models, let us plot all the models and compare the Time Series plots.



Model 5: Simple Exponential Smoothing

Copy separate Training and test data for Simple Exponential smoothing.
Fit , predict and plot it.



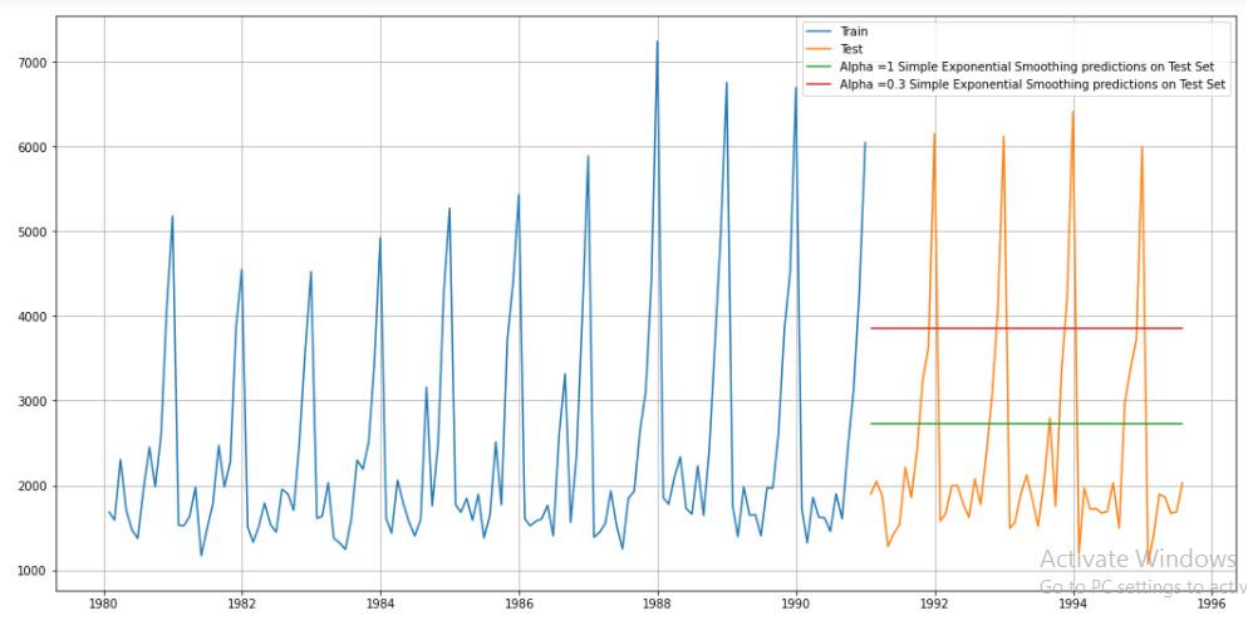
Model Evaluation for α = 0.995 : Simple Exponential Smoothing

For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 1316.035

Setting different alpha values. Higher the alpha value more weightage is given to the more recent observation. That means, what happened recently will happen again. We will run a loop with different alpha values to understand which particular value works best for alpha on the test set.

Model Evaluation

	Alpha Values	Train RMSE	Test RMSE
0	0.3	1359.511747	1935.507132
1	0.4	1352.588879	2311.919615
2	0.5	1344.004369	2666.351413
3	0.6	1338.805381	2979.204388
4	0.7	1338.844308	3249.944092
5	0.8	1344.462091	3483.801006
6	0.9	1355.723518	3686.794285
7	1.0	1373.082528	3864.279352

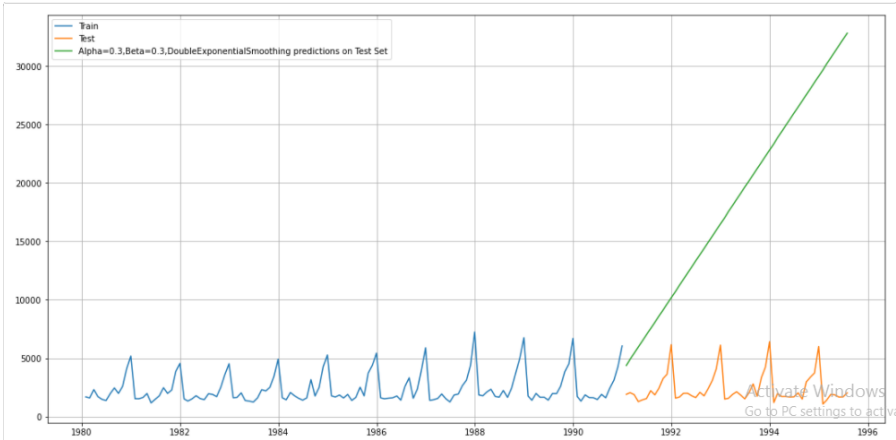


Model 6: Double Exponential Smoothing (Holt's Model)

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

Copy separate Training and test data for Double Exponential smoothing.
Test RMSE are calculated for different α and β . then sorted by Test RMSE values.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	1592.292788	18259.110704
8	0.4	0.3	1569.338606	23878.496940
1	0.3	0.4	1682.573828	26069.841401
16	0.5	0.3	1530.575845	27095.532414
24	0.6	0.3	1506.449870	29070.722592

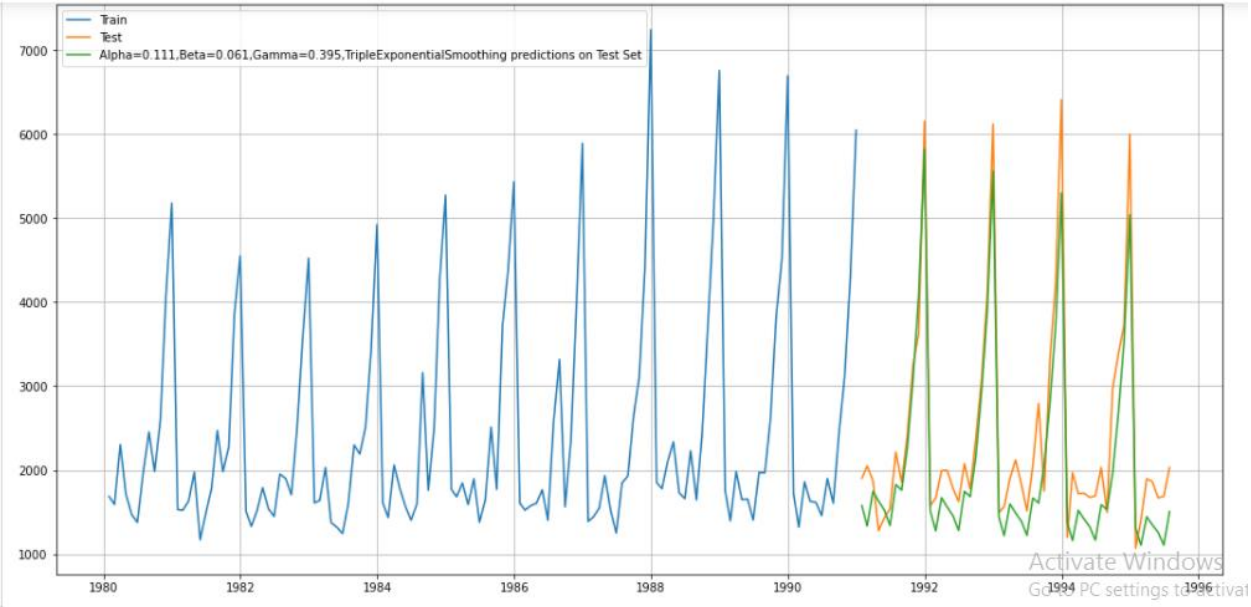


Method 7: Triple Exponential Smoothing (Holt - Winter's Model)

Three parameters α, β and γ are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

Copy separate Training and test data for Triple Exponential smoothing.
Fit the data (autofit) and predict it.

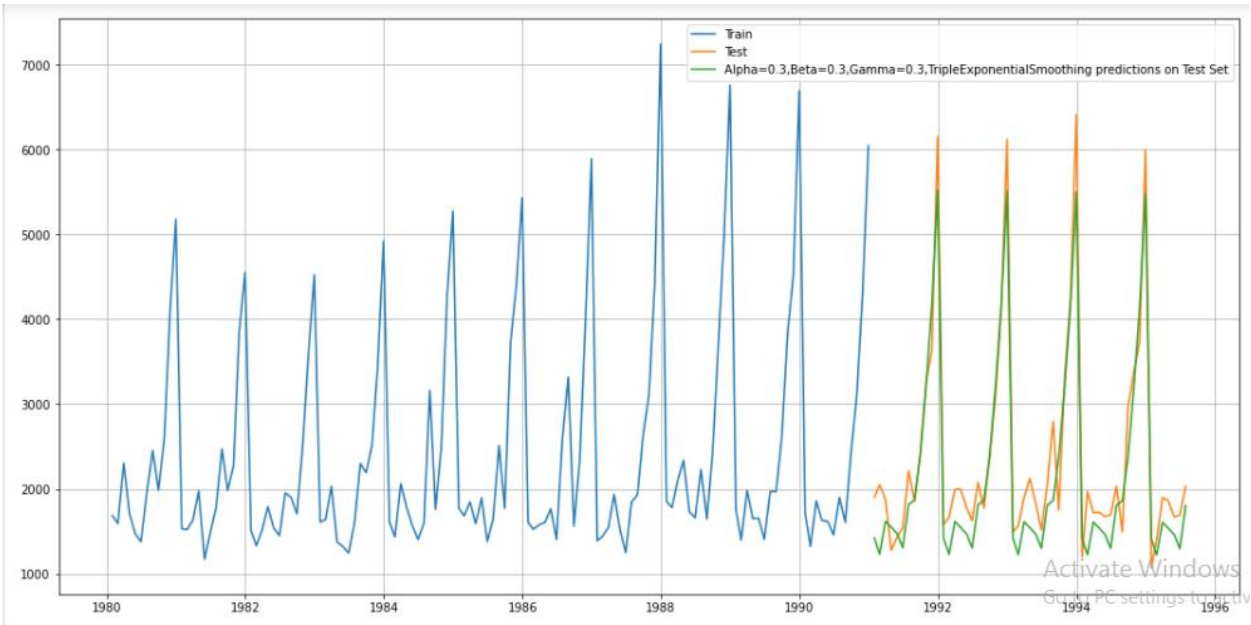
Time_stamp	Sparkling	auto_predict
1991-01-31	1902	1577.247616
1991-02-28	2049	1333.624996
1991-03-31	1874	1746.043741
1991-04-30	1279	1630.568737
1991-05-31	1432	1523.310226



For Alpha=0.111,Beta=0.061,Gamma=0.395, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 469.593

Test RMSE are calculated for different α , β and γ . then sorted by Test RMSE values.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
0	0.3	0.3	0.3	404.513320	392.786198
8	0.3	0.4	0.3	424.828055	410.854547
65	0.4	0.3	0.4	435.553595	421.409170
296	0.7	0.8	0.3	700.317756	518.188752
130	0.5	0.3	0.5	498.239915	542.175497

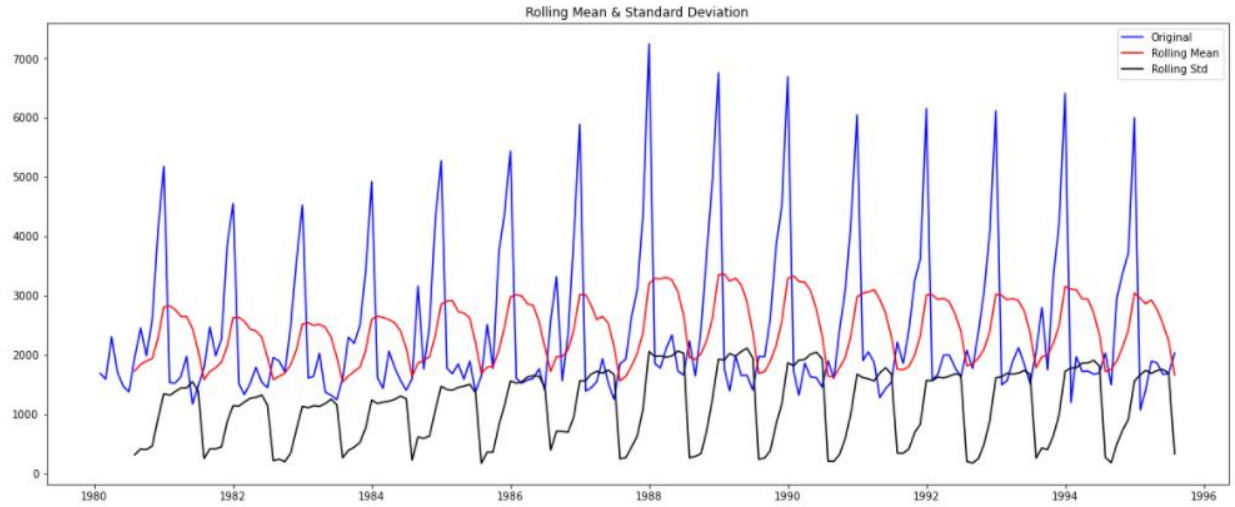


1.5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

Dickey Fuller Test

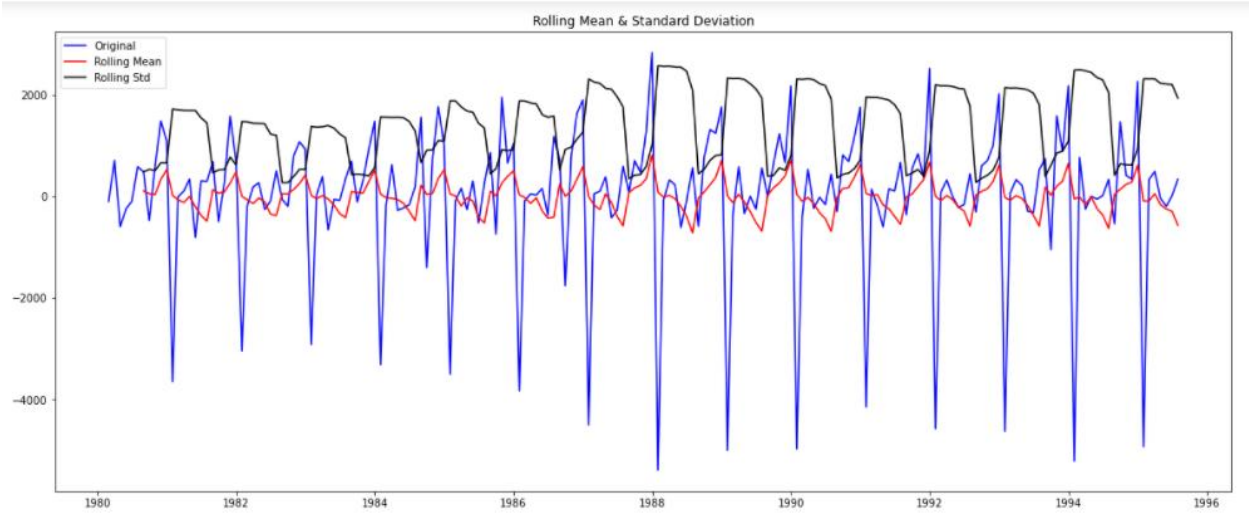
Null Hypothesis $H0$ - Series is not Stationary

Alternative Hypothesis $H1$ - Series is Stationary




```
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 see that at 5% significant level 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 taking a difference of order 1 the series have become stationary at α = 0.05.

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.

Automated Version of ARIMA

The following loop helps us in getting a combination of different parameters of p and q in the range of 0 and 2 We have kept the value of d as 1 as we need to take a difference of the series to make it stationary.

```
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)
```

Model calculated for different p and q values and sorted with lowest AIC values.

	param	AIC
8	(2, 1, 2)	2210.622756
7	(2, 1, 1)	2232.360490
2	(0, 1, 2)	2232.783098
5	(1, 1, 2)	2233.597647
4	(1, 1, 1)	2235.013945
6	(2, 1, 0)	2262.035600
1	(0, 1, 1)	2264.906438
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:            D.Sparkling   No. Observations:                  131
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1099.311
Method:                       css-mle   S.D. of innovations            1013.444
Date:                Sat, 04 Sep 2021   AIC                            2210.623
Time:                        20:02:15   BIC                            2227.874
Sample:                    02-29-1980   HQIC                           2217.633
                         - 12-31-1990
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                     5.5850      0.518     10.780      0.000       4.570       6.600
ar.L1.D.Sparkling         1.2698      0.075     17.042      0.000       1.124       1.416
ar.L2.D.Sparkling        -0.5601      0.074     -7.617      0.000      -0.704      -0.416
ma.L1.D.Sparkling        -1.9963      0.043    -46.962      0.000      -2.080      -1.913
ma.L2.D.Sparkling         0.9963      0.043     23.384      0.000       0.913       1.080
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1335           -0.7074j            1.3362           -0.0888
AR.2            1.1335           +0.7074j            1.3362            0.0888
MA.1            1.0002           +0.0000j            1.0002            0.0000
MA.2            1.0035           +0.0000j            1.0035            0.0000
-----------------------------------------------------------------------------
```

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

RMSE VALUES: 1374.2065102508311

Automated Version of SARIMA

The following loop helps us in getting a combination of different parameters of p, q, P, Q in the range of 0 and 3 We have kept the value of d in the range (1,2) and D in range (0,1) .

```
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)
```

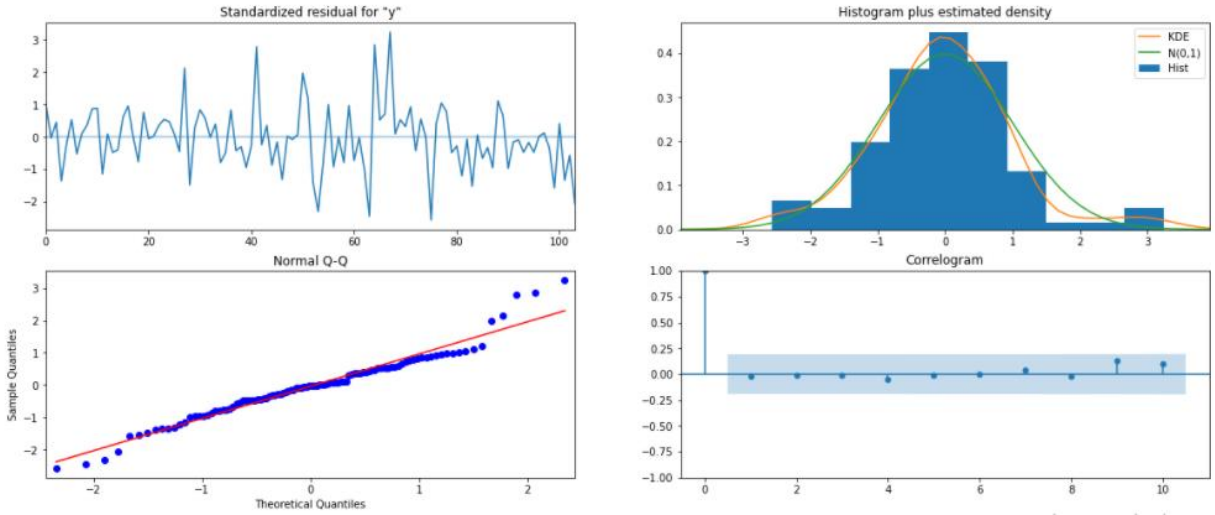
Model calculated for different p, q, d, P, Q,D values and sorted by least AIC values.

	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

```
                                 SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -770.792
Date:                            Sat, 04 Sep 2021   AIC                           1555.584
Time:                                    23:26:23   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).
```

Diagnostic plot:



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

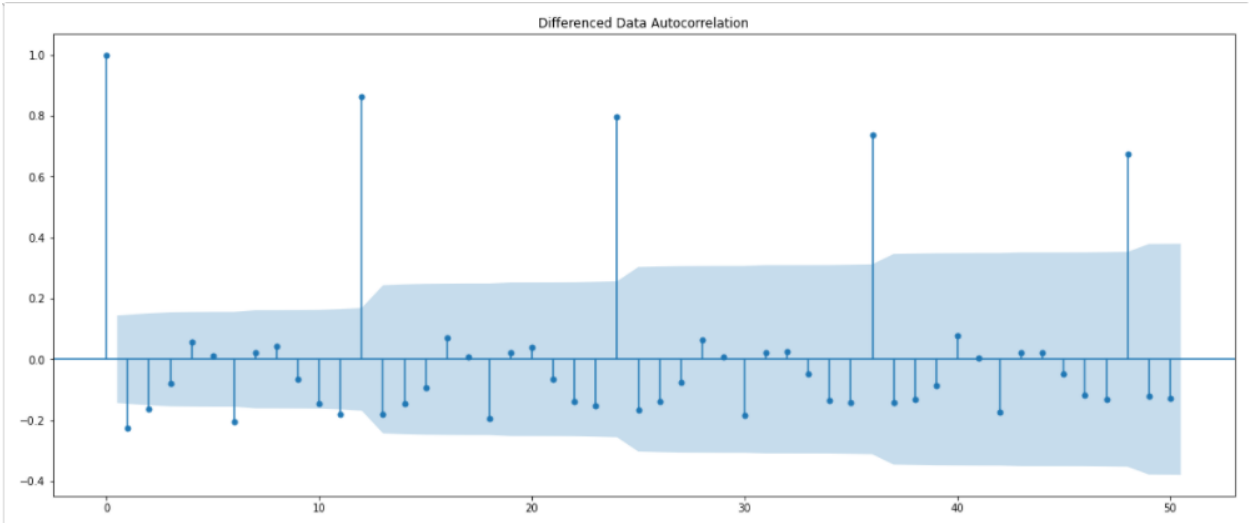
Forecast the model

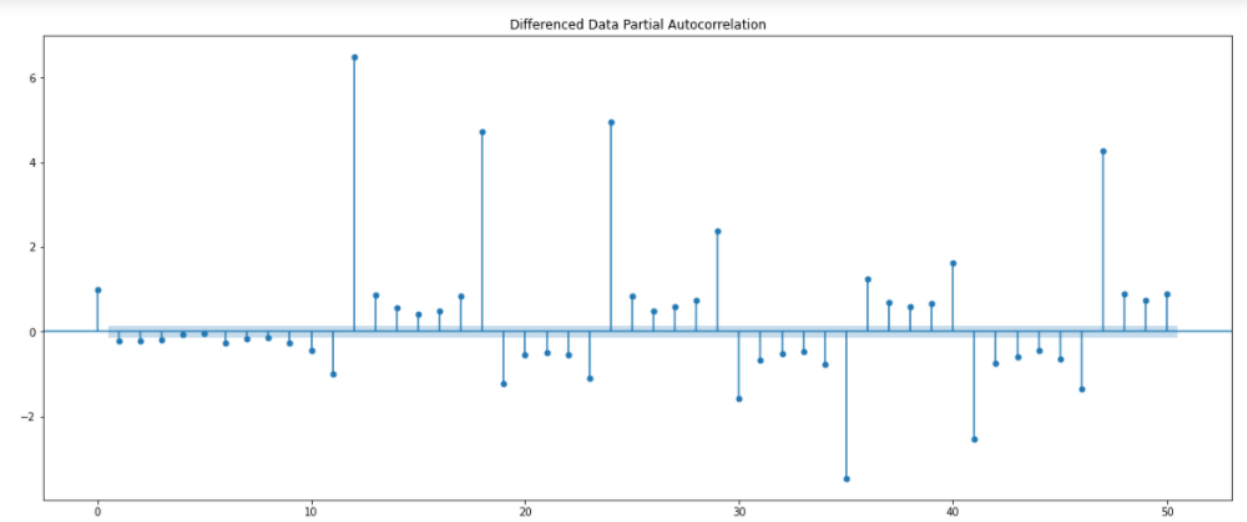
y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1327.386418	388.344800	566.244597	2088.528239
1	1315.110768	402.007729	527.190097	2103.031440
2	1621.588857	402.001336	833.680717	2409.496997
3	1598.867465	407.239037	800.693619	2397.041311
4	1392.688227	407.969106	593.083472	2192.292982

RMSE: 528.62

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.

ARIMA Model based on ACF and PACF





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 0.

The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.

By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 3 and 2.

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(3, 1, 2)	Log Likelihood	-1107.463
Method:	css-mle	S.D. of innovations	1105.893
Date:	Sat, 04 Sep 2021	AIC	2228.927
Time:	23:26:40	BIC	2249.053
Sample:	02-29-1980 - 12-31-1990	HQIC	2237.105

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.9796	7.57e-05	7.89e+04	0.000	5.979	5.980
ar.L1.D.Sparkling	-0.4420	nan	nan	nan	nan	nan
ar.L2.D.Sparkling	0.3079	nan	nan	nan	nan	nan
ar.L3.D.Sparkling	-0.2501	nan	nan	nan	nan	nan
ma.L1.D.Sparkling	-0.0002	nan	nan	nan	nan	nan
ma.L2.D.Sparkling	-0.9998	nan	nan	nan	nan	nan

Roots

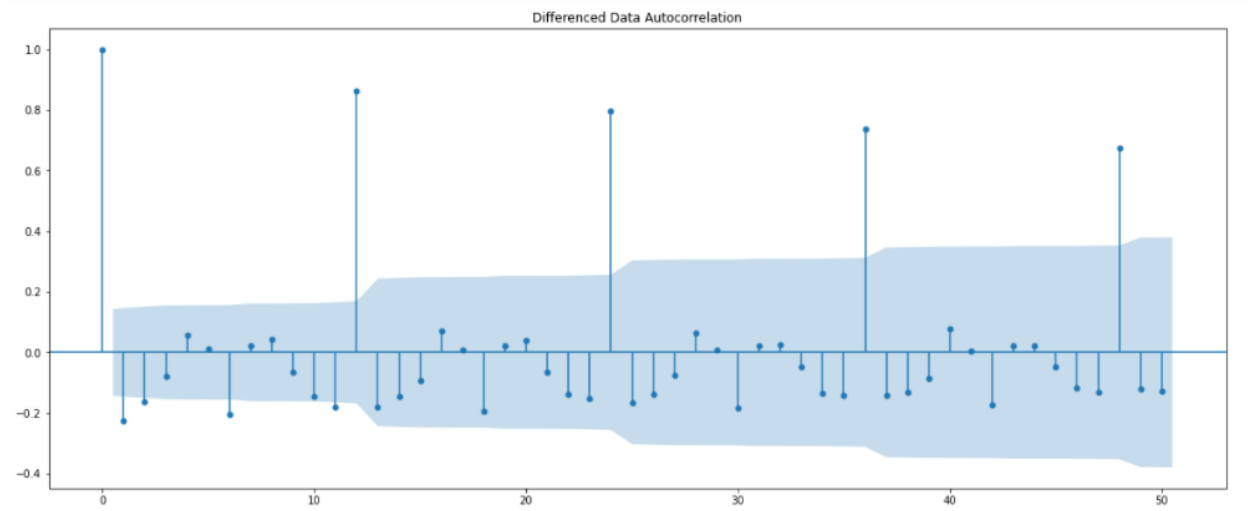
	Real	Imaginary	Modulus	Frequency
AR.1	-1.0000	-0.0000j	1.0000	-0.5000
AR.2	1.1156	-1.6595j	1.9996	-0.1558
AR.3	1.1156	+1.6595j	1.9996	0.1558
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-1.0002	+0.0000j	1.0002	0.5000

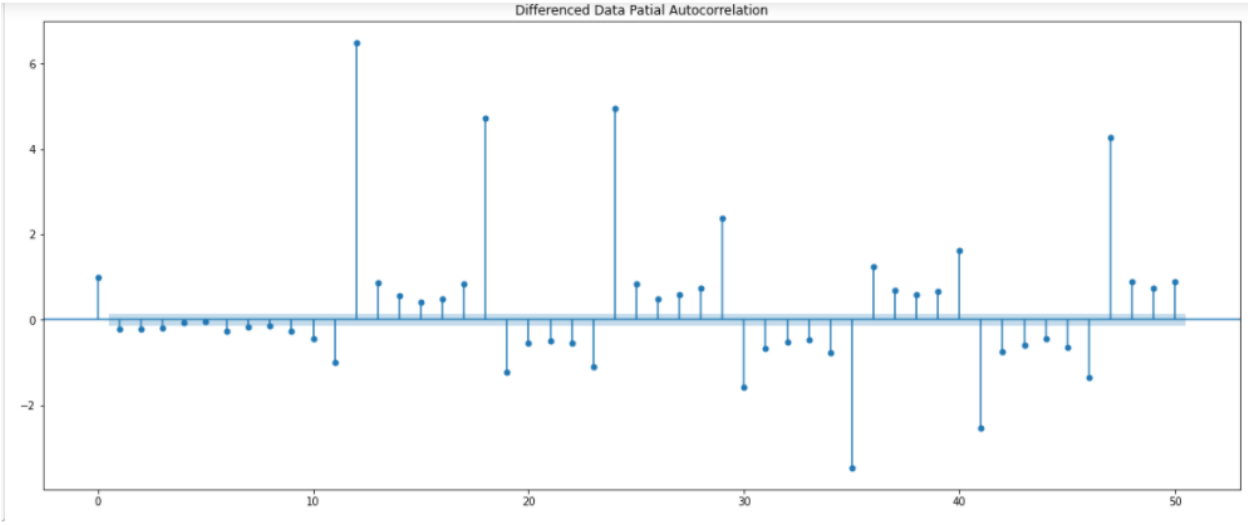
Predict on the Test Set using this model and evaluate the model.

RMSE: 1378.656099152732

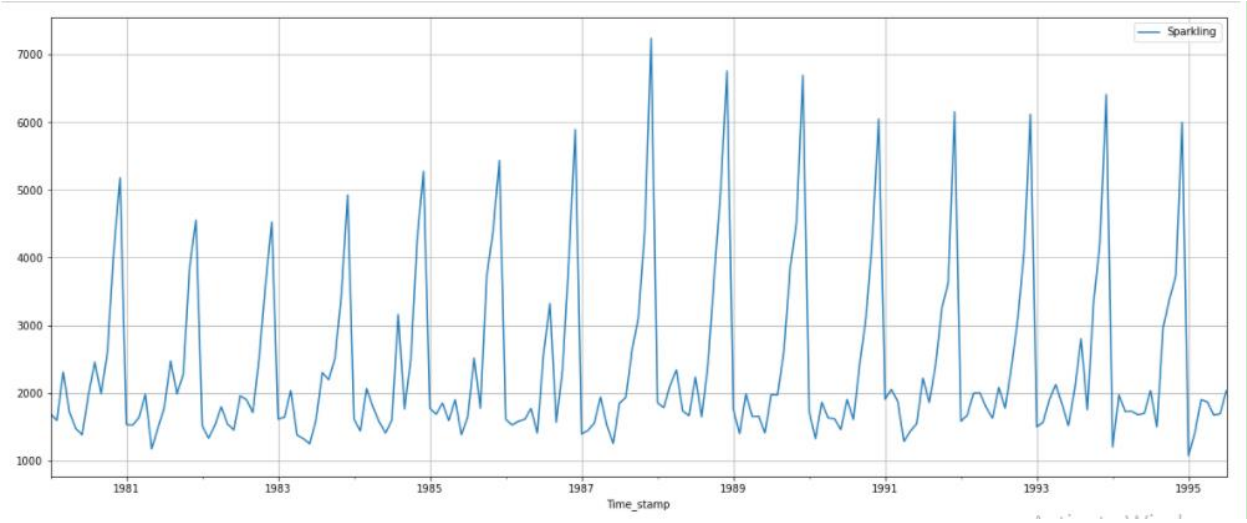
SARIMA Model : Manually looking at ACF and PACF

Let us look at the ACF and the PACF plots once more.

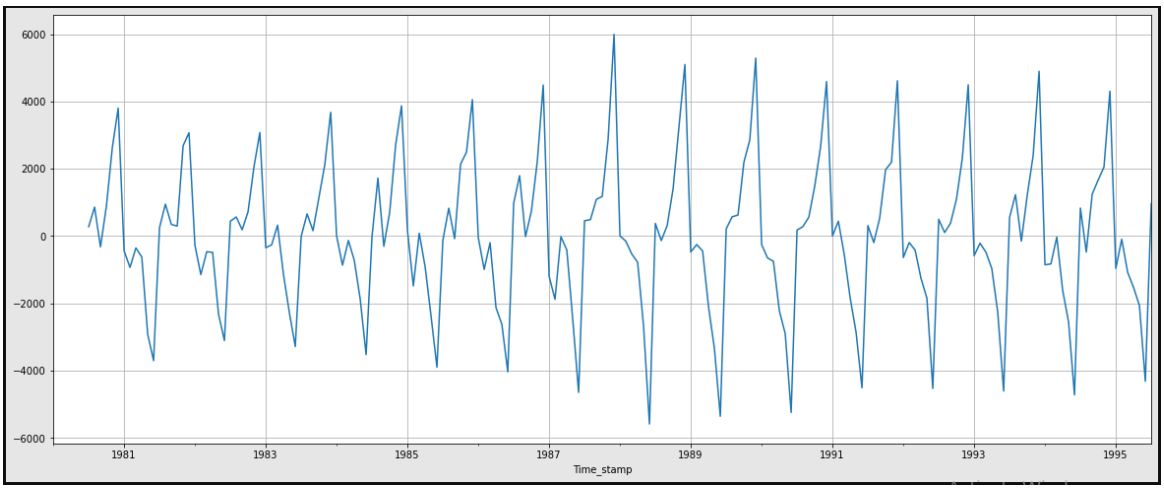


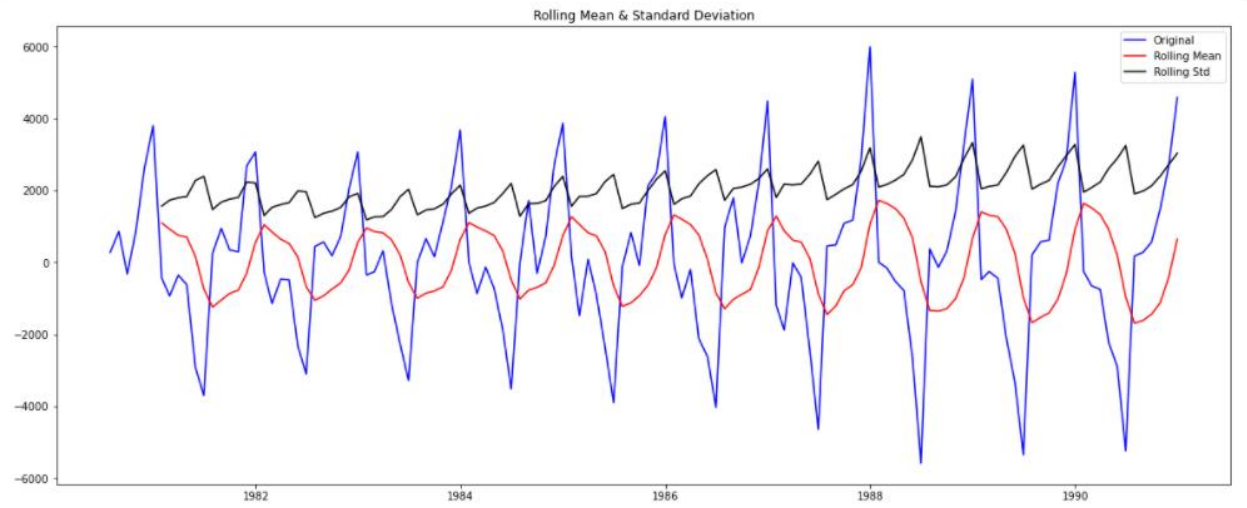


Plot the dataset:



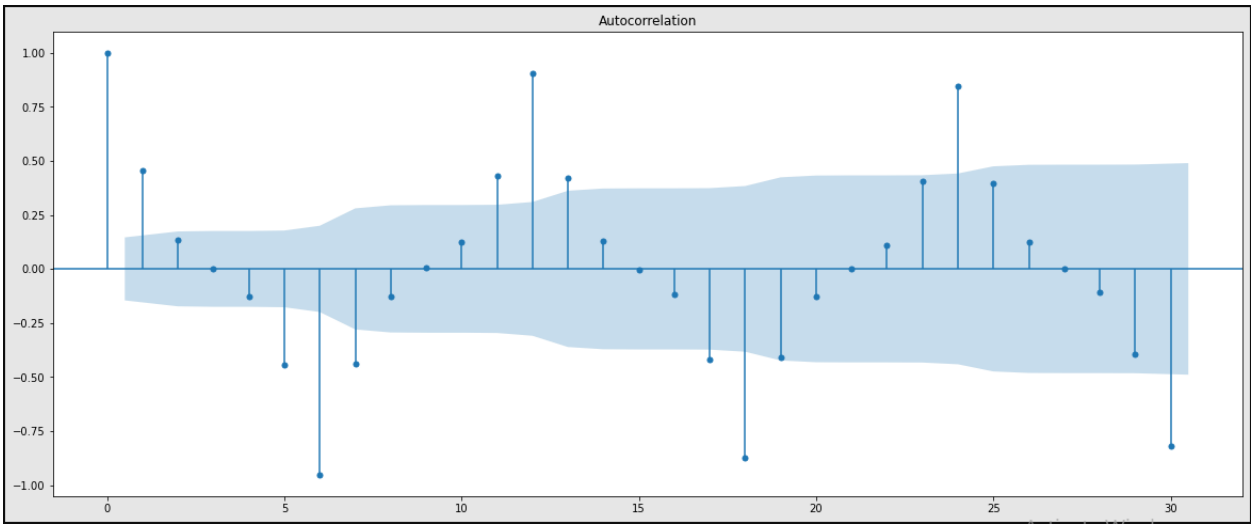
We see that there is a seasonality. So, now we take a seasonal differencing and check the series.

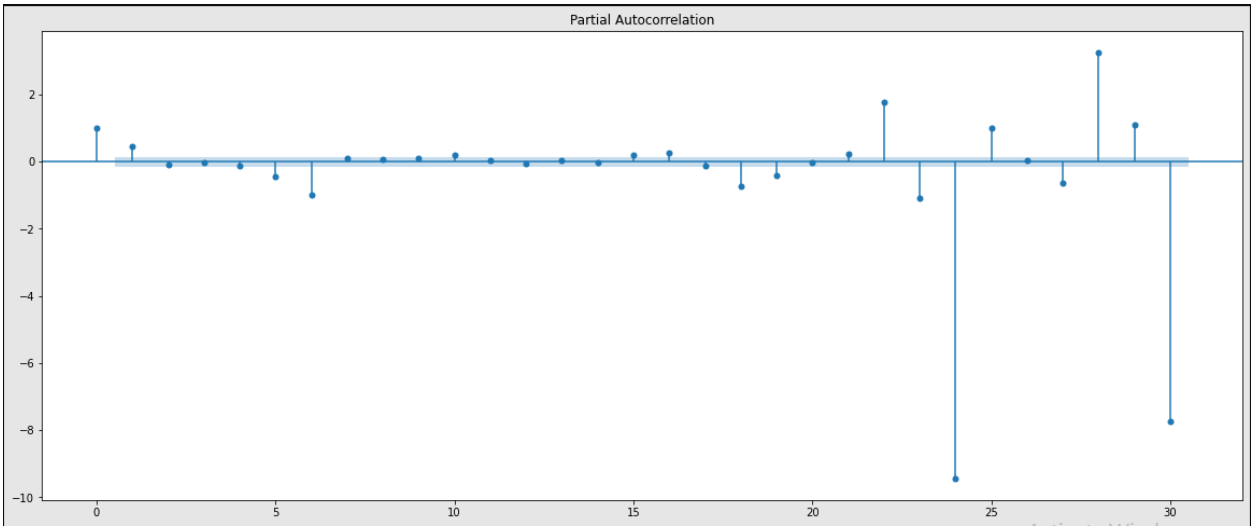





```
Results of Dickey-Fuller Test:
Test Statistic                -8.181919e+00
p-value                        8.088278e-13
#Lags Used                     6.000000e+00
Number of Observations Used    1.190000e+02
Critical Value (1%)           -3.486535e+00
Critical Value (5%)           -2.886151e+00
Critical Value (10%)          -2.579896e+00
dtype: float64
```

ACF and PACF Plot after differencing:





Here, we have taken alpha=0.05.

We are going to take the seasonal period as 6. We will keep the p(1) and q(1) parameters same as the ARIMA model.

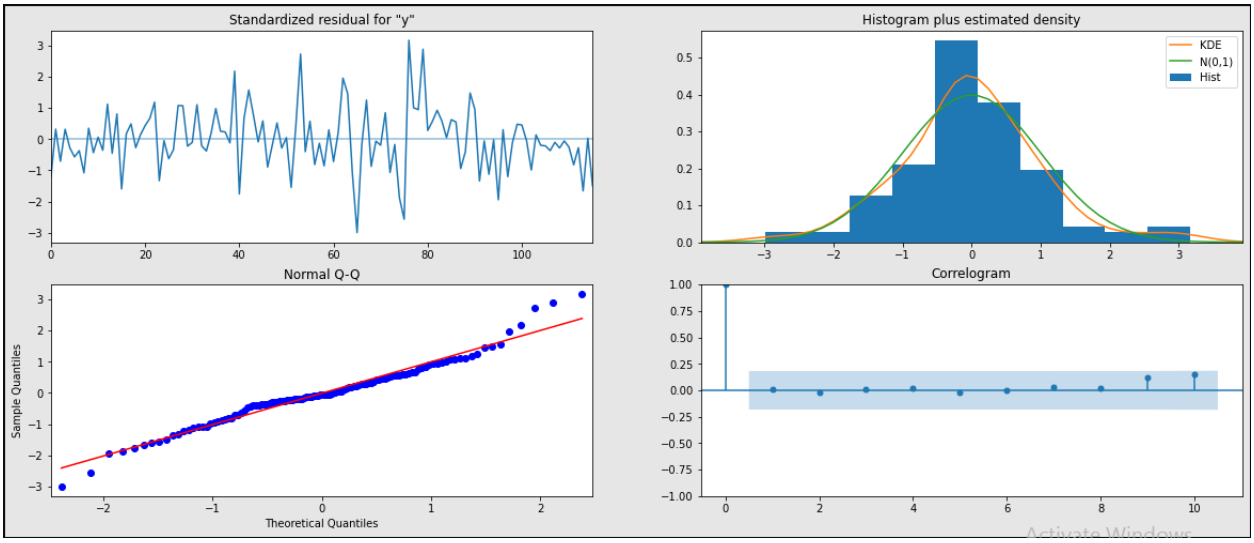
The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.

The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 0. Remember to check the ACF and the PACF plots only at multiples of 6 (since 6 is the seasonal period).

By looking at the plots we see that the ACF and the PACF cut-off at 1 & 1 .

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(3, 1, 2)x(1, 1, [1], 6)   Log Likelihood                -865.947
Date:                             Sat, 04 Sep 2021   AIC                           1747.895
Time:                                     23:26:44   BIC                           1769.924
Sample:                                          0   HQIC                          1756.837
                                               - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6129      0.105     -5.858      0.000      -0.818      -0.408
ar.L2          0.1361      0.091      1.491      0.136      -0.043       0.315
ar.L3          0.0095      0.085      0.112      0.910      -0.157       0.176
ma.L1         -0.0889      0.153     -0.582      0.560      -0.388       0.210
ma.L2         -0.9111      0.127     -7.180      0.000      -1.160      -0.662
ar.S.L6       -0.9839      0.021    -46.306      0.000      -1.026      -0.942
ma.S.L6       -0.0023      0.129     -0.018      0.986      -0.255       0.250
sigma2      1.717e+05   1.29e-06   1.33e+11      0.000    1.72e+05    1.72e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 8.37
Prob(Q):                              0.93   Prob(JB):                         0.02
Heteroskedasticity (H):               1.76   Skew:                             0.21
Prob(H) (two-sided):                  0.08   Kurtosis:                         4.24
===================================================================================
```

Diagnostic plot:



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

RMSE: 312.69

Forecast test set with confidence interval.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	1497.341917	416.088483	681.823476	2312.860359
1	1407.561546	435.169378	554.645239	2260.477853
2	1803.845549	435.404617	950.468182	2657.222917
3	1723.210856	436.955384	866.794041	2579.627670
4	1628.760545	437.261973	771.742827	2485.778263

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.

	Test RMSE
SARIMA(3,1,2)(1,1,1,6) based on ACF & PACF	312.690768
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponential Smoothing	392.786198
Alpha=0.111,Beta=0.061,Gamma=0.395,TripleExponential Smoothing	469.593384
SARIMA (1,1,2)(1,0,2,12)	528.621309
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
SimpleAverageModel	1275.081804
6pointTrailingMovingAverage	1283.927428
Alpha=0.995,SimpleExponential Smoothing	1316.034674
9pointTrailingMovingAverage	1346.278315
ARIMA (2,1,2)	1374.206510
ARIMA (3,1,2) based on ACF & PACF	1378.656099
RegressionOnTime	1389.135175
Alpha=0.3,SimpleExponential Smoothing	1935.507132
NaiveModel	3864.279352
Alpha=0.3,Beta=0.3,DoubleExponential Smoothing	18259.110704

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.

From the above analysis, SARIMA model based on ACF, PACF plot would be optimum for this dataset because it has low RMSE value. Building the most optimum model on the Full Data.

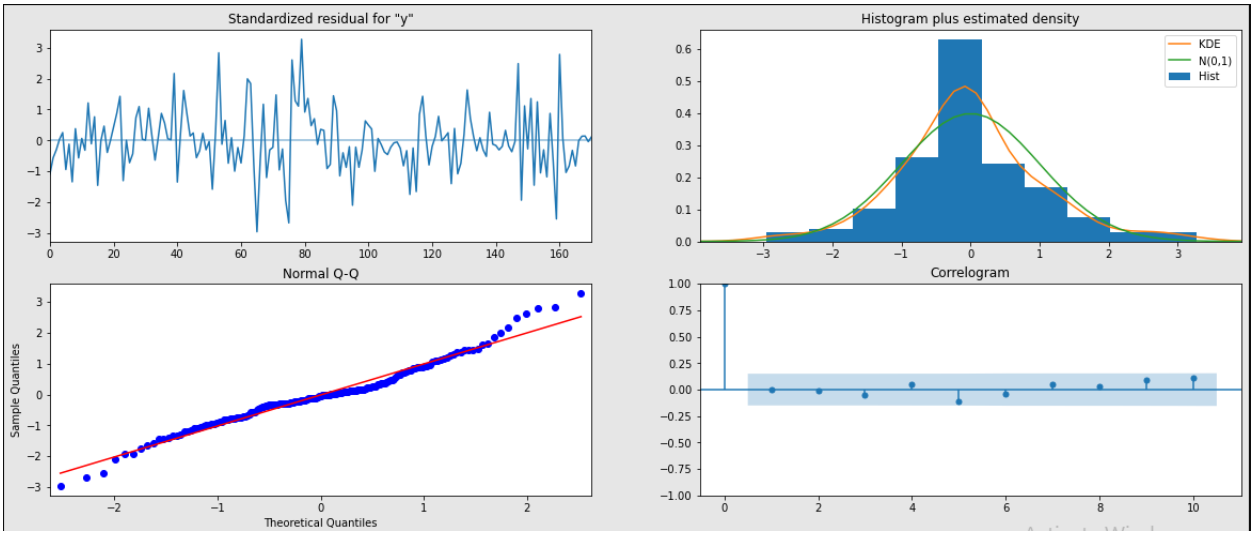
SARIMAX Results

Dep. Variable:	y	No. Observations:	187
Model:	SARIMAX(3, 1, 2)x(1, 1, [1], 6)	Log Likelihood	-1274.645
Date:	Sat, 04 Sep 2021	AIC	2565.290
Time:	23:26:49	BIC	2590.424
Sample:	0	HQIC	2575.488
	- 187		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8299	0.068	-12.261	0.000	-0.963	-0.697
ar.L2	0.0225	0.079	0.283	0.777	-0.133	0.178
ar.L3	-0.0101	0.069	-0.146	0.884	-0.145	0.125
ma.L1	8.361e-06	31.882	2.62e-07	1.000	-62.488	62.488
ma.L2	-1.0000	0.094	-10.619	0.000	-1.185	-0.815
ar.S.L6	-0.9906	0.014	-68.493	0.000	-1.019	-0.962
ma.S.L6	0.1698	0.093	1.817	0.069	-0.013	0.353
sigma2	1.668e+05	1.27e-05	1.31e+10	0.000	1.67e+05	1.67e+05

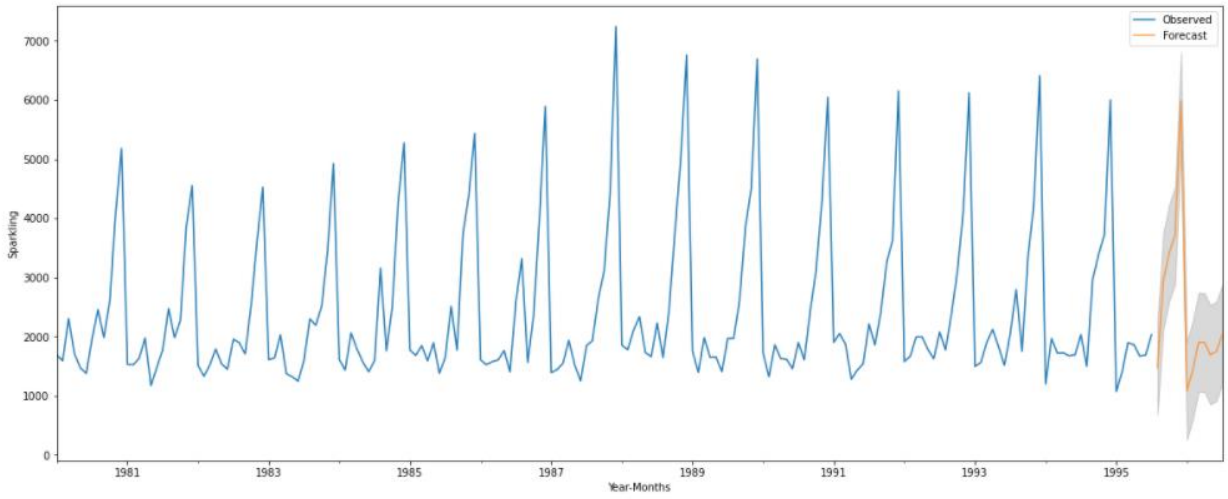
Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	12.01
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	1.28	Skew:	0.31
Prob(H) (two-sided):	0.35	Kurtosis:	4.14

Diagnostic plot:



Evaluate the model on the whole :RMSE: 571.871846931169

Predict 12 months into future and plot it with confidence interval:



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. Please explain and summarize the various steps performed in this project. There should be proper business interpretation and actionable insights present.

- Sales of Sparkling Wine (1980 - 1995) were analysed. Hidden details are captured by doing EDA.
- From 1981 to 1983 trend is decreasing , 1983 to 1988 increasing, then decreasing throughout 1995.
- Median : 1900.
- Highest Sale on 1987, whereas lowest sale on 1995.
- It is evident from monthly plot that sales has been increased from September to December. Stock has to be more during these time frame. January recorded the lowest sale which is right after the month of December of previous years.
- Trend and Seasonality are there in the dataset.
- Dataset were splitting for training and test set. Various models such as Linear regression, Naïve bayes, Simple Exponential smoothing, Double exponential smoothing, Triple exponential smoothing, ARIMA and SARIMA built on training data and tested on test data.
- Since the dataset has seasonality, SARIMA model would be best suited model. Same was evident through RMSE value. SARIMA has the lowest RMSE value. SARIMA model was applied on full data.
- Sales for next 12 month is predicted with confidence interval. Sales are varying drastically across the month, since shelf life of the wine is more, company should do production on median values which close to 2000. Which will meet peak demand without additional resource during the month of highest sales.

Problem 2: Analyzing with rose wine dataset.

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

First 5 rows:

	YearMonth	Rose
0	1980-01	112.0
1	1980-02	118.0
2	1980-03	129.0
3	1980-04	99.0
4	1980-05	116.0

last 5 rows

	YearMonth	Rose
182	1995-03	45.0
183	1995-04	52.0
184	1995-05	28.0
185	1995-06	40.0
186	1995-07	62.0

Info:

```
<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   Rose        185 non-null    float64
dtypes: float64(1), object(1)
memory usage: 3.0+ KB
```

Dataset has missing values:

```
YearMonth    0
Rose         2
dtype: int64
```

Dataset has missing values. These values has to be imputed.

Create Time stamp:

```
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')
```

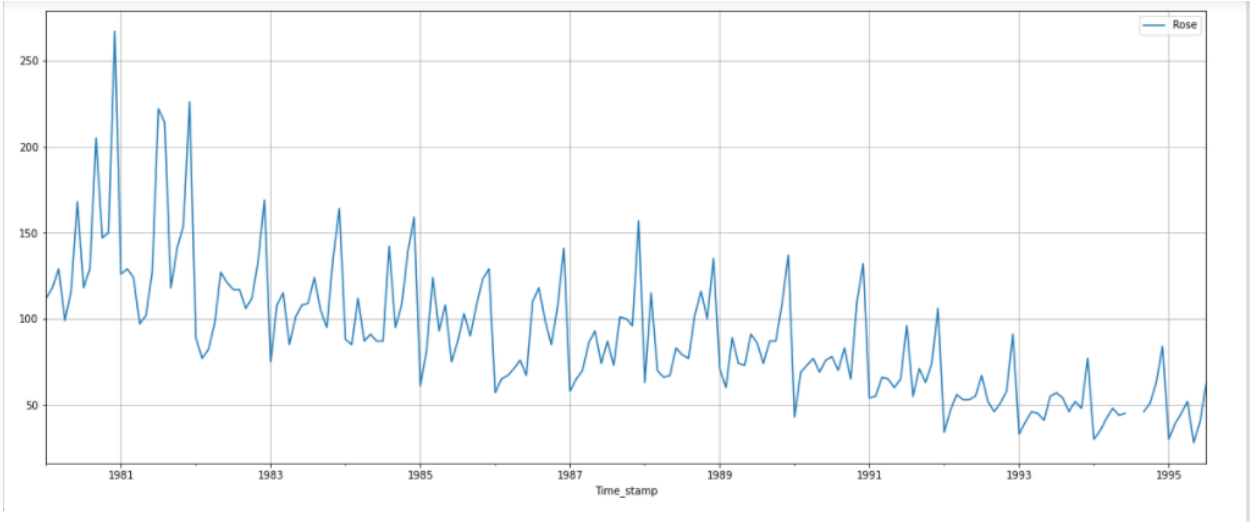
Add time stamp to dataset:

	YearMonth	Rose	Time_stamp
0	1980-01	112.0	1980-01-31
1	1980-02	118.0	1980-02-29
2	1980-03	129.0	1980-03-31
3	1980-04	99.0	1980-04-30
4	1980-05	116.0	1980-05-31

Make timestamp as index and remove Year month column:

Time_stamp	Rose
1980-01-31	112.0
1980-02-29	118.0
1980-03-31	129.0
1980-04-30	99.0
1980-05-31	116.0

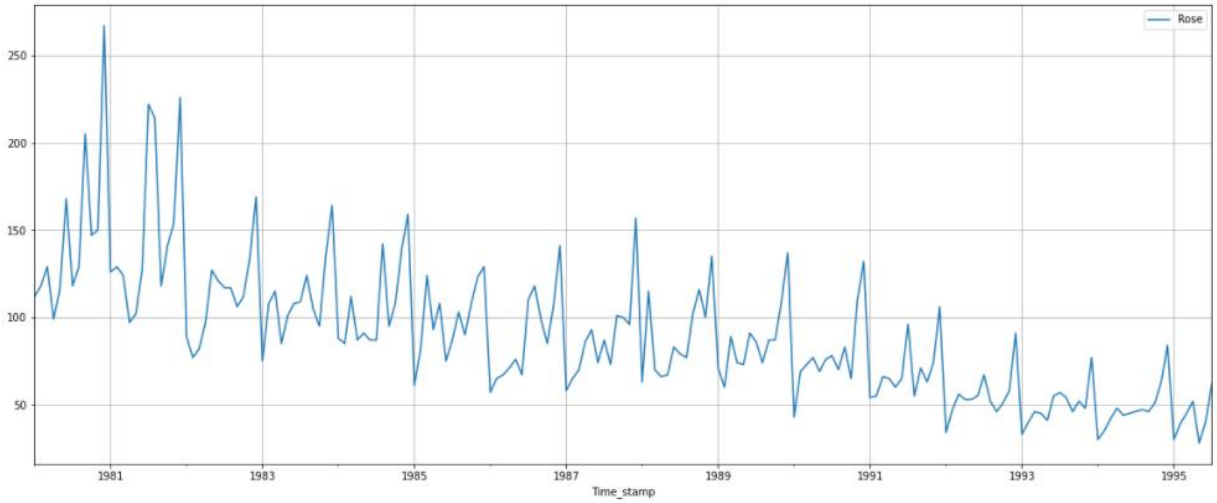
Plot the Dataset:



Data has decreasing trend and Seasonality.

Data has missing values in the year 1994, Same is observed in the plot. July and August of 1994 has missing values.

Interpolate the data with spline method and plot it again to visualize the missing values.



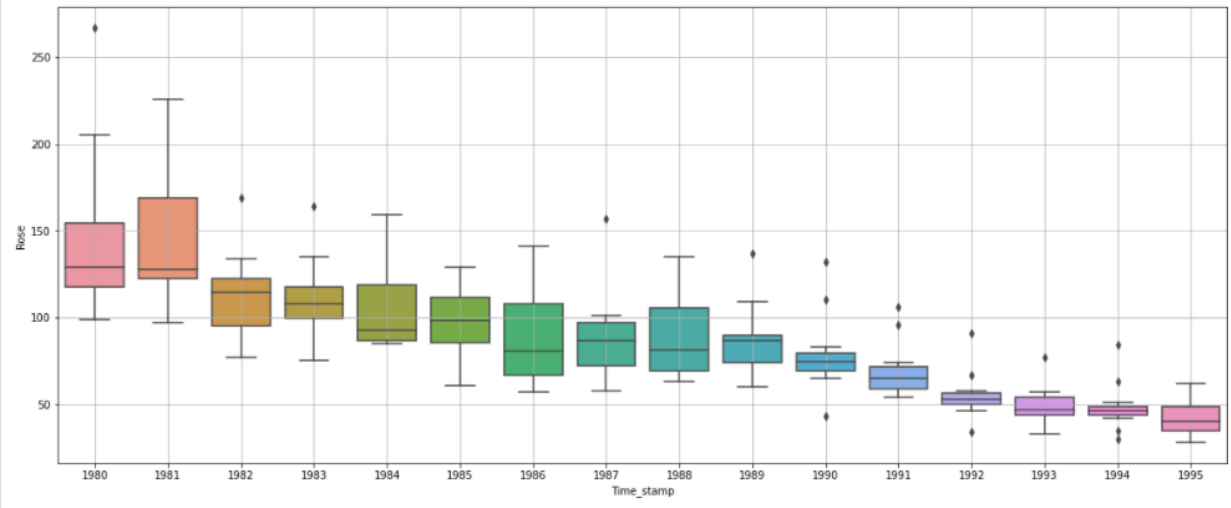
Continuous plot is observed.

Describe the dataset:

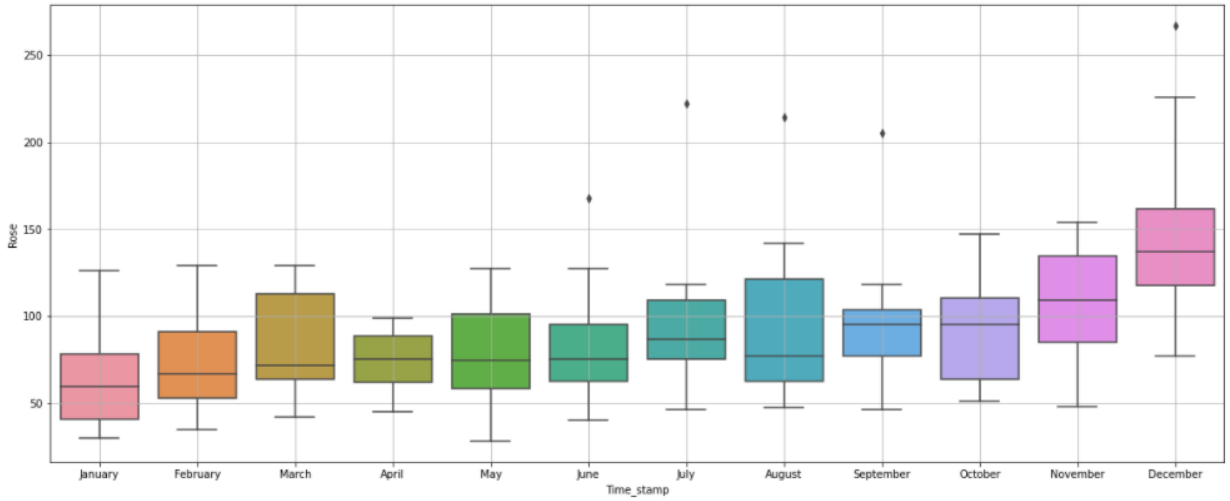
	Rose
count	187.000000
mean	89.927152
std	39.224081
min	28.000000
25%	62.500000
50%	85.000000
75%	111.000000
max	267.000000

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

Spread of sales across different years:

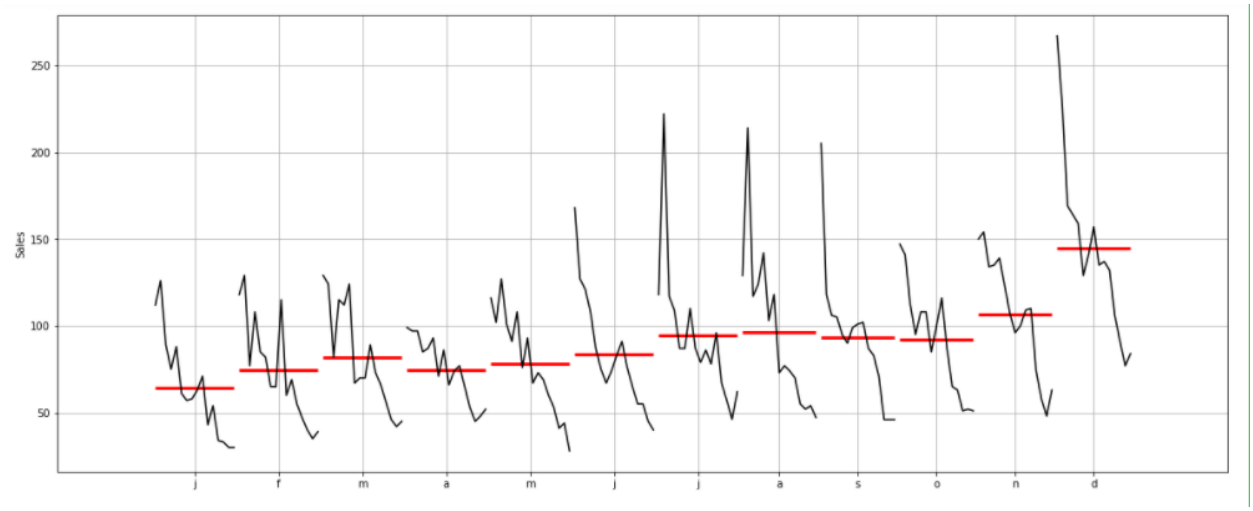


Spread of sales across different Months:



From Year plot, we can notice trend is decreasing.

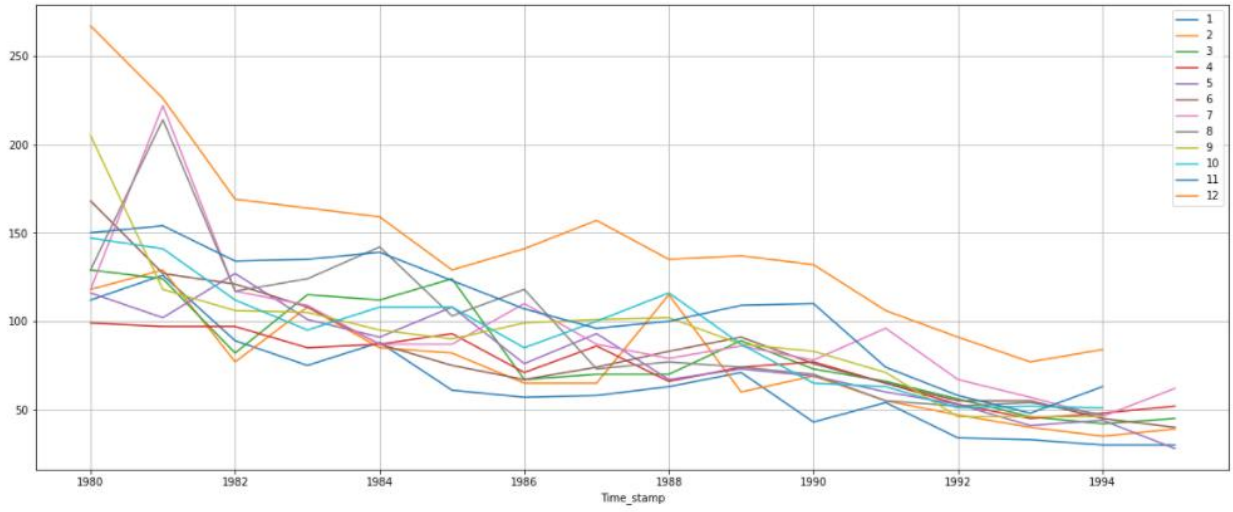
Sales across different years and within different months across years



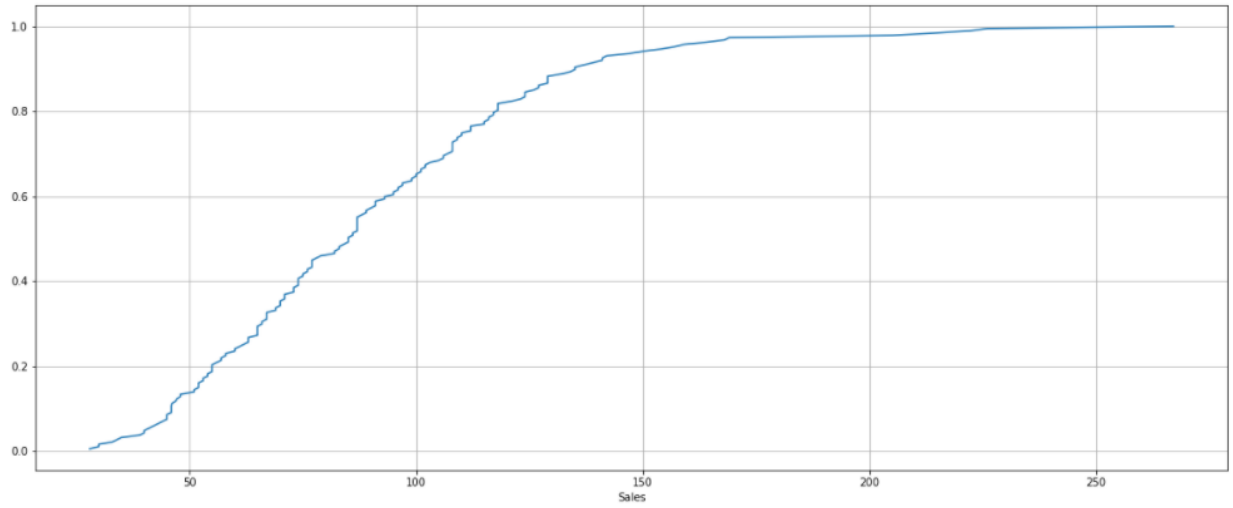
This plot shows us behavior of time series across months.

Graph of monthly sales:

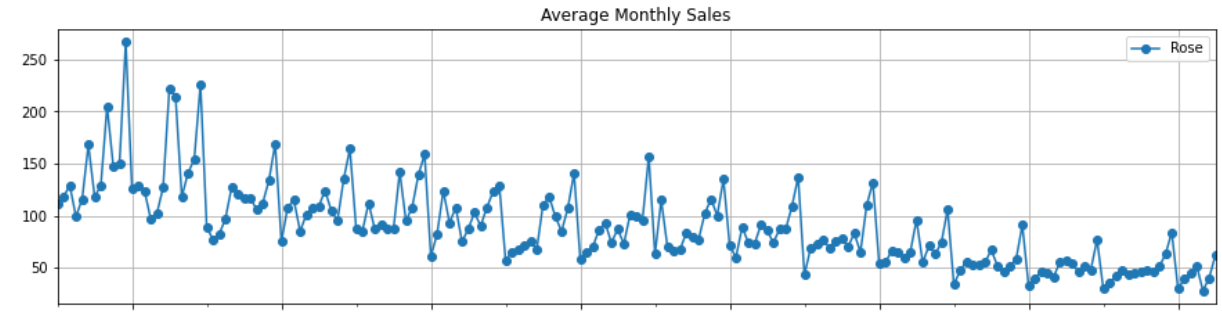
Time_stamp	1	2	3	4	5	6	7	8	9	10	11	12
Time_stamp												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.000000	129.000000	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.000000	214.000000	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.000000	117.000000	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.000000	124.000000	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.000000	142.000000	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.000000	103.000000	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.000000	118.000000	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.000000	73.000000	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.000000	77.000000	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.000000	74.000000	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.000000	70.000000	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.000000	55.000000	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.000000	52.000000	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.000000	54.000000	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	46.155493	47.221907	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.000000	NaN	NaN	NaN	NaN	NaN



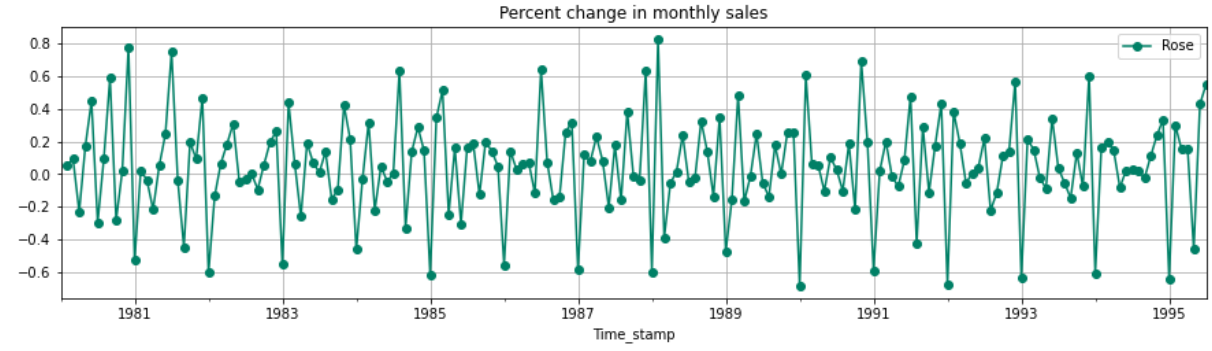
Empirical cumulative distribution



Average Monthly sales:

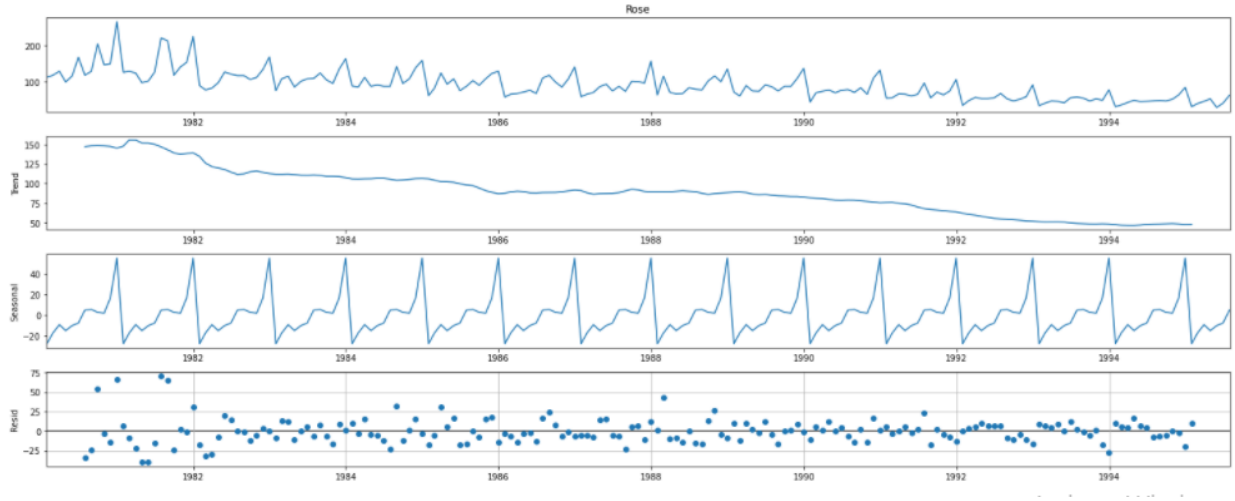


Percentage change in monthly sales:



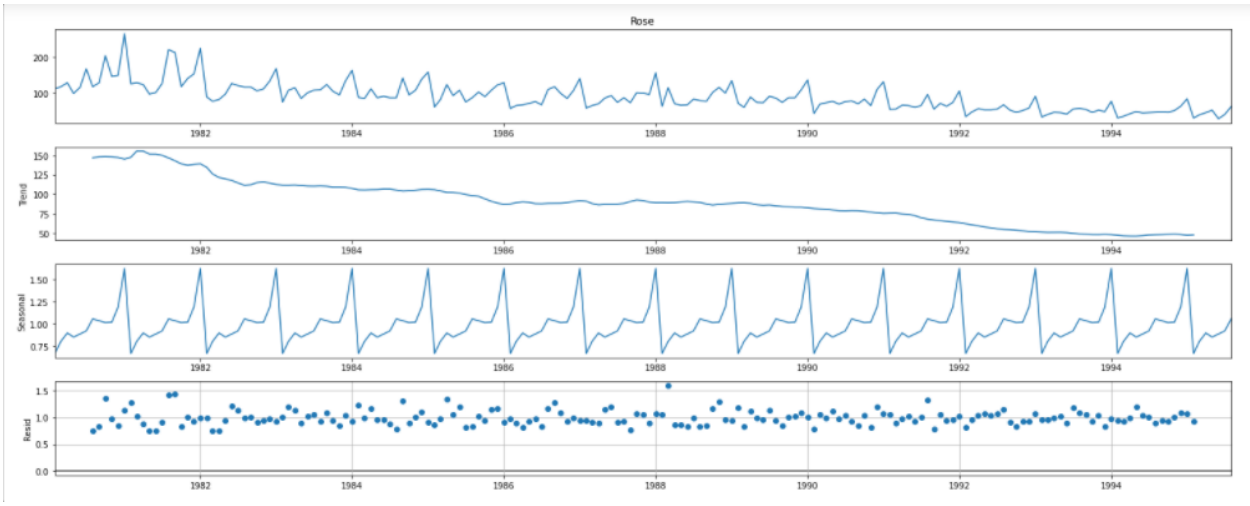
Decompose Time series and plot different components

Additive Decomposition:



From the above plot, it is clear that data has decreasing trend and seasonality. Residual has some patterns. so let's explore multiplicative model.

Multiplicative decomposition:



Multiplicative model's trend and seasonality are same. Residual has some patterns. so additive model is considered for further analysis.

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    147.083333
1980-08-31    148.125000
1980-09-30    148.375000
1980-10-31    148.083333
Name: trend, dtype: float64

seasonality
 Time_stamp
1980-01-31   -27.921848
1980-02-29   -17.445147
1980-03-31    -9.299974
1980-04-30   -15.112474
1980-05-31   -10.210688
1980-06-30    -7.692831
1980-07-31     4.938518
1980-08-31     5.590168
1980-09-30     2.761485
1980-10-31     1.858708
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   -34.021852
1980-08-31   -24.715168
1980-09-30    53.863515
1980-10-31    -2.942041
Name: resid, dtype: float64
```

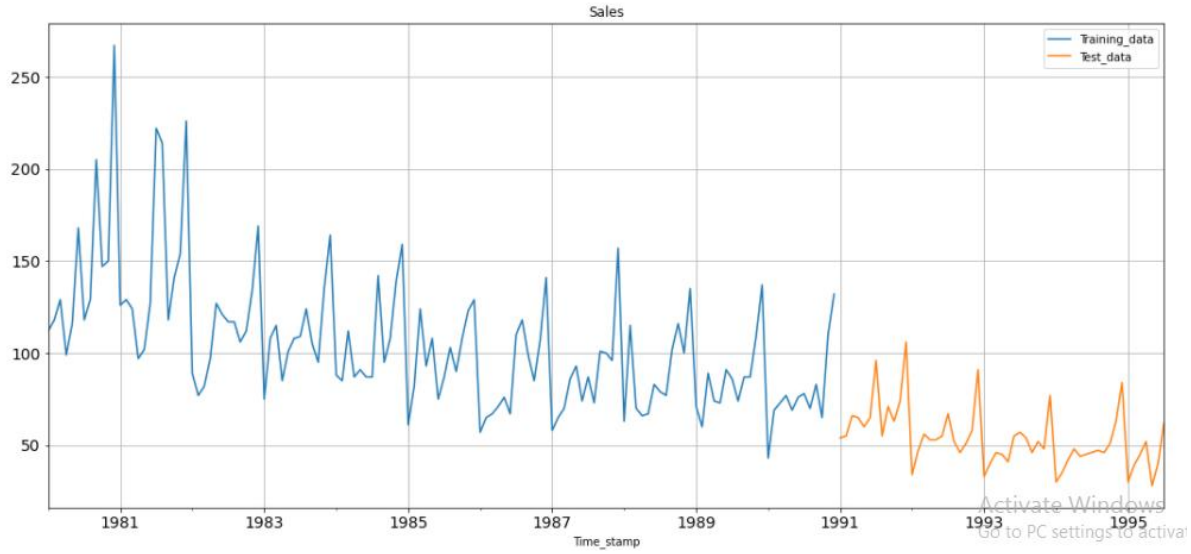
2.3. Split the data into training and test. The test data should start in 1991.

Dataset was split between Training and test dataset. Test dataset should start from 1991.

Shape of Train and test dataset:

```
(132, 1)
(55, 1)
```

Plot as Training and test dataset:



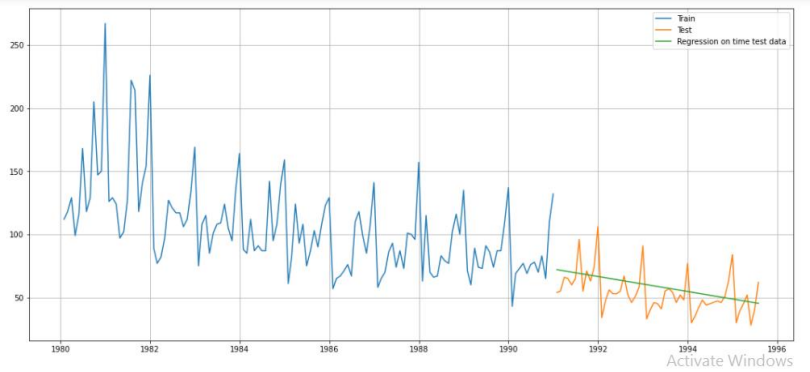
2.4. Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression ,naïve forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

Model 1 : Linear Regression

For this particular linear regression, we are going to regress the 'Sales' variable against the order of the occurrence. For this we need to modify our training data 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, 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,
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
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,
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,
158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 18
3, 184, 185, 186, 187]
```

Copy separate training and test data for Linear Regression.
Create variable for Linear regression. Fit the data and predict it. Plot the regression on training and test set.

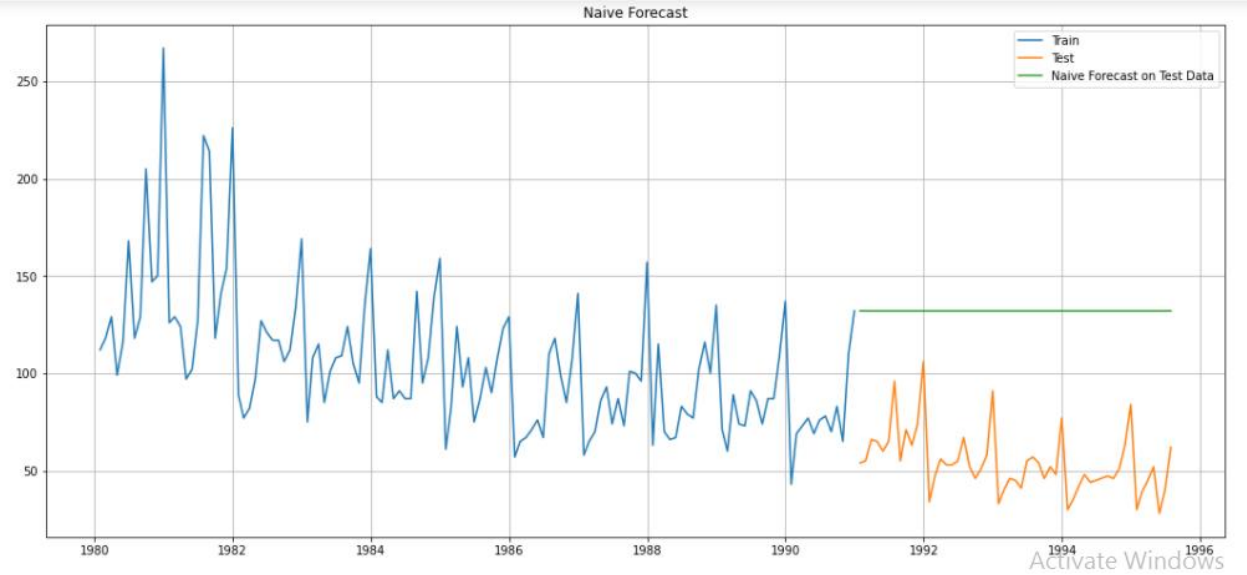


For Regression On Time forecast on the Test Data, RMSE is 15.255

Model 2: Naive Approach: $\hat{y}\ t+1=yt$

For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today, therefore the prediction for day after tomorrow is also today.

Copy separate Training and test data for Naïve approach.
Fit , predict and plot it.



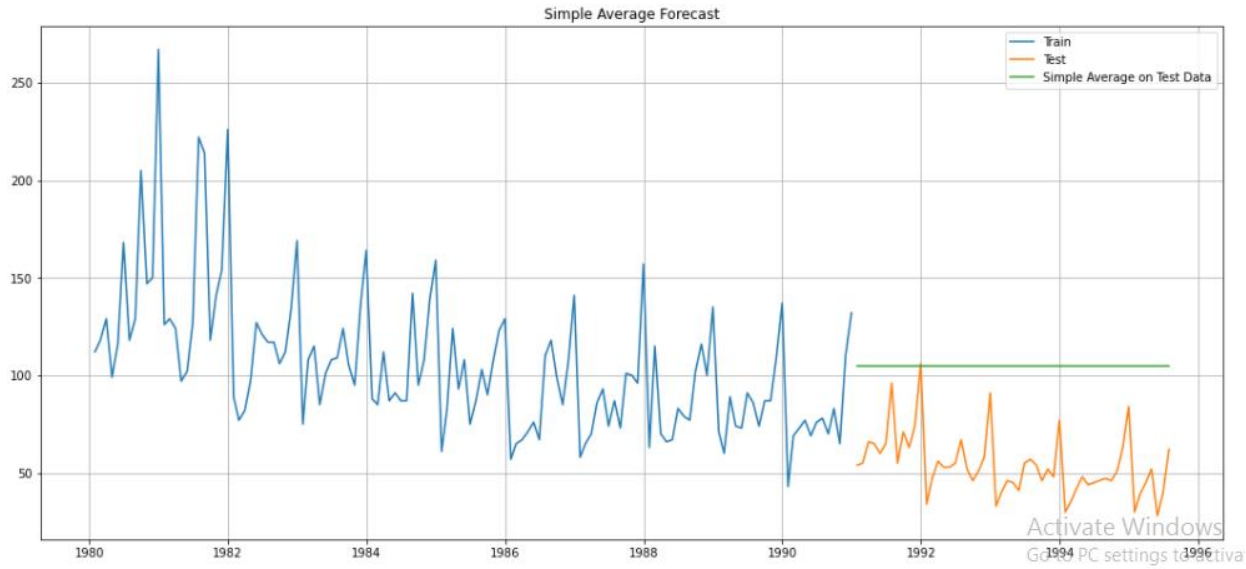
For Naive forecast on the Test Data RMSE is 79.672

Model 3: Simple Average

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

Copy separate Training and test data for Simple Average.

Fit , predict and plot it.



For Simple Average forecast on the Test Data, RMSE is 53.413

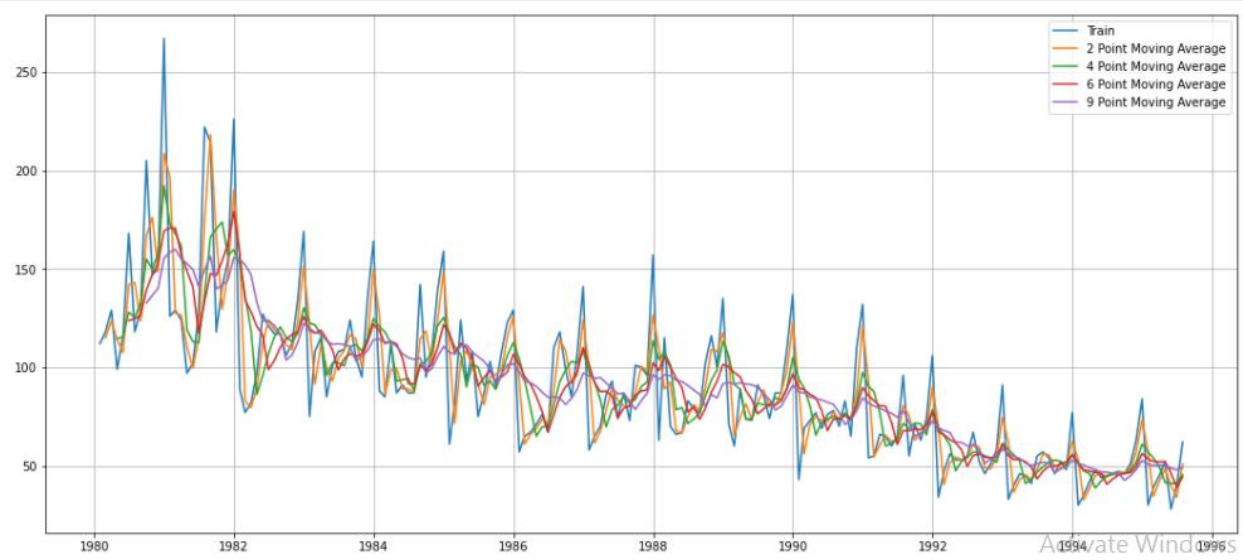
Model 4: Moving Average(MA)

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

Copy separate dataset for moving Average.

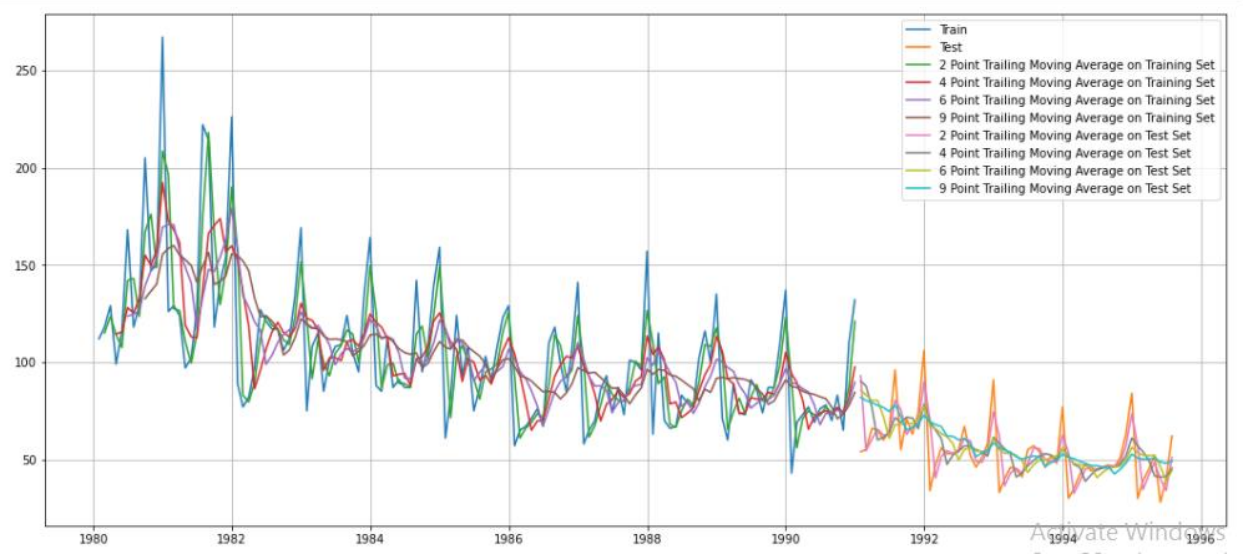
Trailing moving averages

Time_stamp	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-31	112.0	NaN	NaN	NaN	NaN
1980-02-29	118.0	115.0	NaN	NaN	NaN
1980-03-31	129.0	123.5	NaN	NaN	NaN
1980-04-30	99.0	114.0	114.5	NaN	NaN
1980-05-31	116.0	107.5	115.5	NaN	NaN



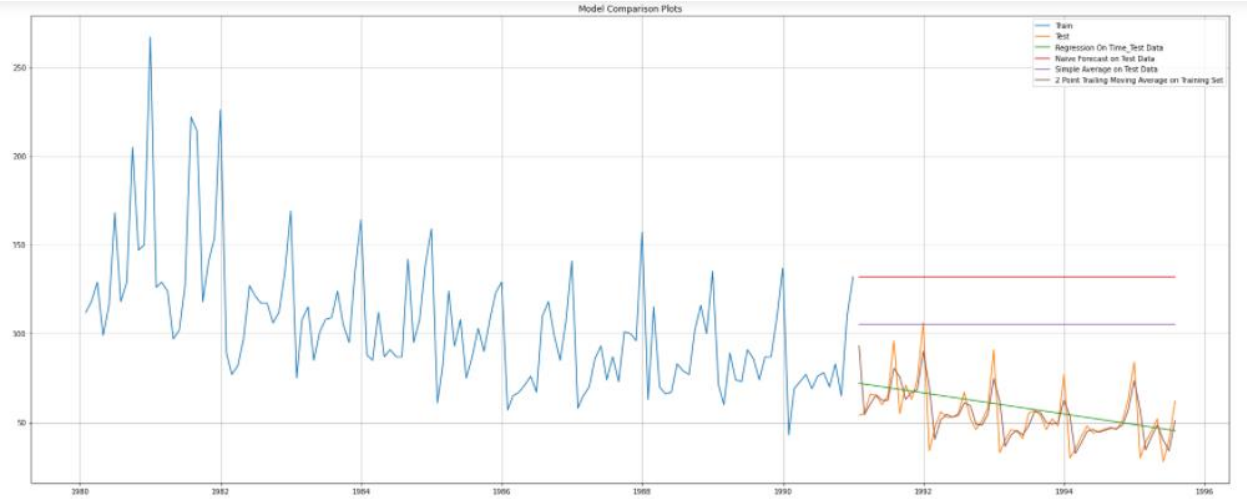
Let us split the data into train and test and plot this Time Series. The window of the moving average is need to be carefully selected as too big a window will result in not having any test set as the whole series might get averaged over.

Create training and test dataset.



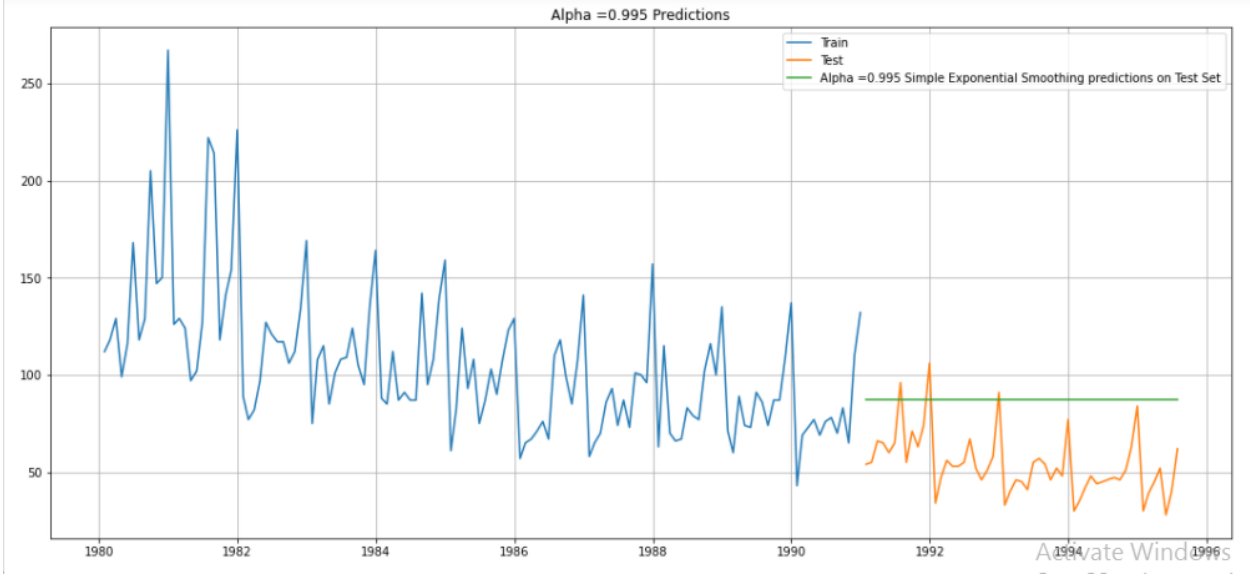
Model Evaluation done only on the test data.

For 2 point Moving Average Model forecast on the Training Data, RMSE is 11.530
For 4 point Moving Average Model forecast on the Training Data, RMSE is 14.444
For 6 point Moving Average Model forecast on the Training Data, RMSE is 14.555
For 9 point Moving Average Model forecast on the Training Data, RMSE is 14.721
Before we go on to build the various Exponential Smoothing models, let us plot all the models and compare the Time Series plots.



Model 5: Simple Exponential Smoothing

Copy separate Training and test data for Simple Exponential smoothing.
Fit , predict and plot it.



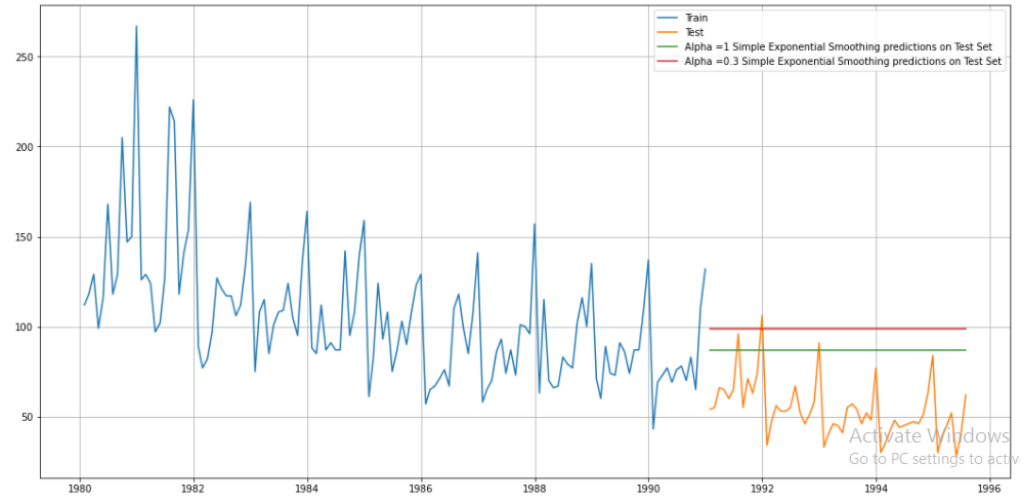
Model Evaluation for α = 0.995 : Simple Exponential Smoothing

For Alpha =0.995 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 36.748

Setting different alpha values. Higher the alpha value more weightage is given to the more recent observation. That means, what happened recently will happen again. We will run a loop with different alpha values to understand which particular value works best for alpha on the test set.

Model Evaluation

	Alpha Values	Train RMSE	Test RMSE
0	0.3	32.470164	47.457057
1	0.4	33.035130	53.719906
2	0.5	33.682839	59.594532
3	0.6	34.441171	64.924245
4	0.7	35.323261	69.651295
5	0.8	36.334596	73.727266
6	0.9	37.482782	77.092660
7	1.0	38.783783	79.672238

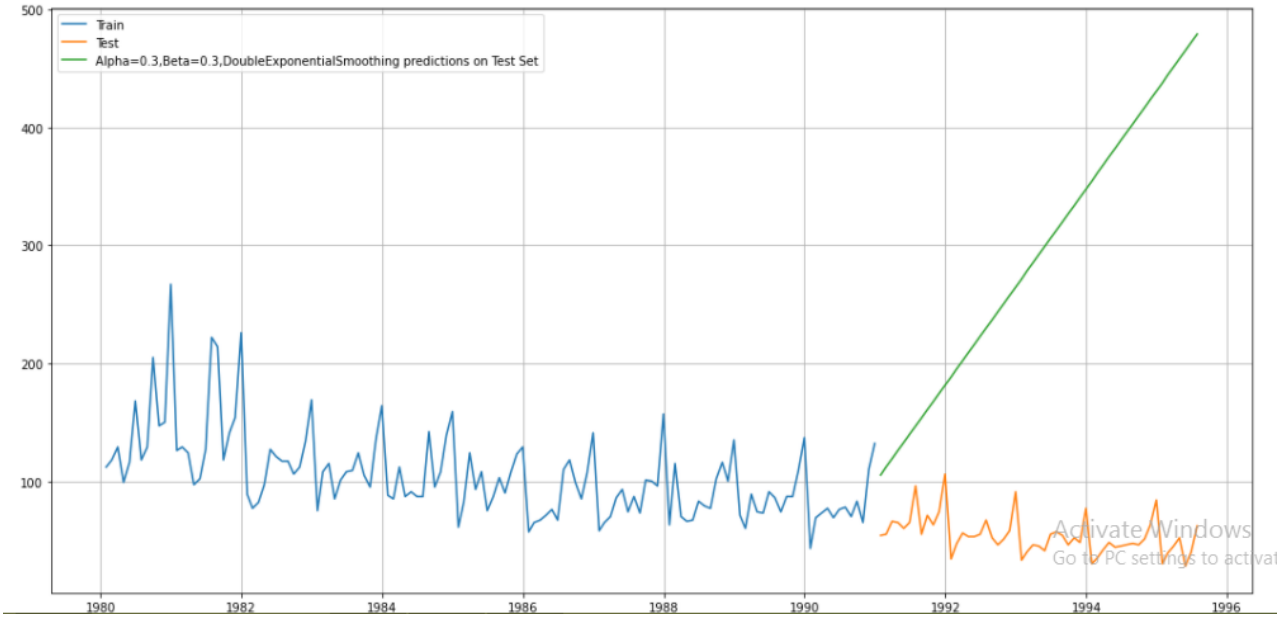


Model 6: Double Exponential Smoothing (Holt's Model)

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

Copy separate Training and test data for Double Exponential smoothing.
Test RMSE are calculated for different α and β . then sorted by Test RMSE values.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	35.944983	265.509912
8	0.4	0.3	36.749123	339.248849
1	0.3	0.4	37.393239	358.693008
16	0.5	0.3	37.433314	394.214956
24	0.6	0.3	38.348984	439.238366

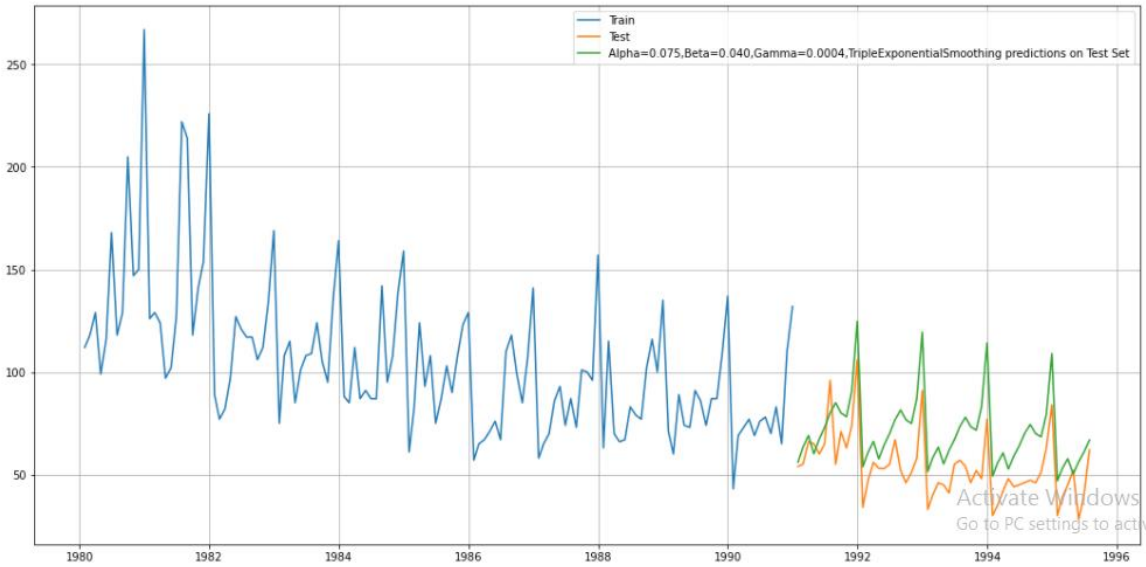


Method 7: Triple Exponential Smoothing (Holt - Winter's Model)

Three parameters α, β and γ are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

Copy separate Training and test data for Triple Exponential smoothing.
Fit the data (autofit) and predict it.

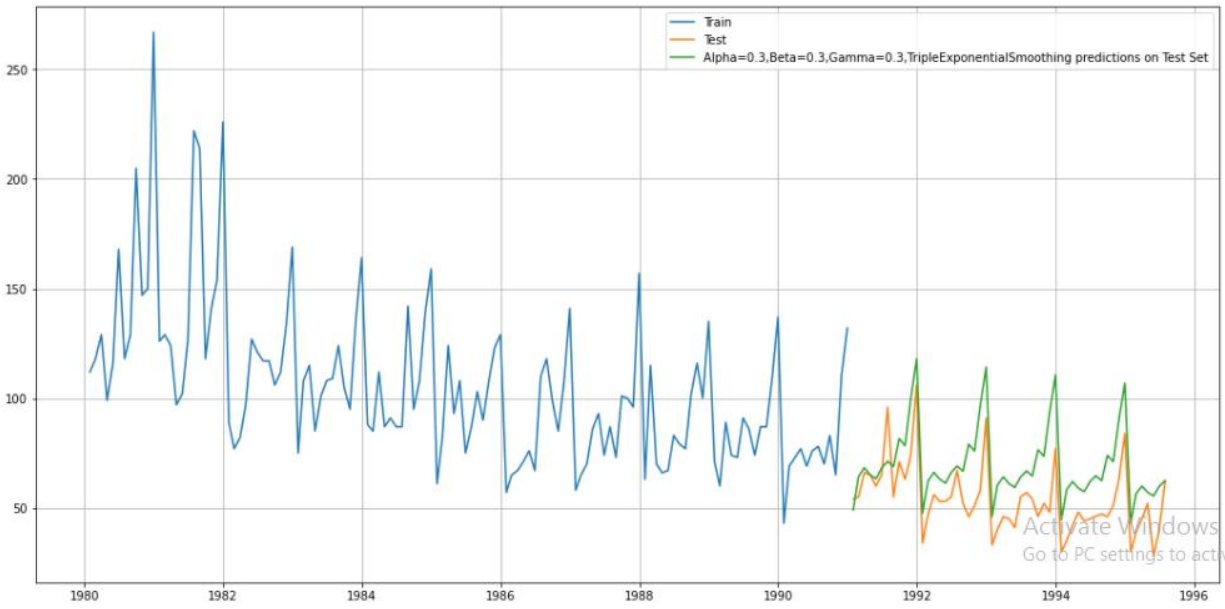
Time_stamp	Rose	auto_predict
1991-01-31	54.0	56.042556
1991-02-28	55.0	63.341718
1991-03-31	66.0	69.018232
1991-04-30	65.0	60.105997
1991-05-31	60.0	67.377021



For Alpha= 0.075,Beta=0.040,Gamma=0.0004, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 19.322

Test RMSE are calculated for different α , β and γ . then sorted by Test RMSE values.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
8	0.3	0.4	0.3	28.111886	10.935749
1	0.3	0.3	0.4	27.399095	11.243009
69	0.4	0.3	0.8	32.601491	12.624534
16	0.3	0.5	0.3	29.087520	14.461672
131	0.5	0.3	0.6	32.144773	16.762882

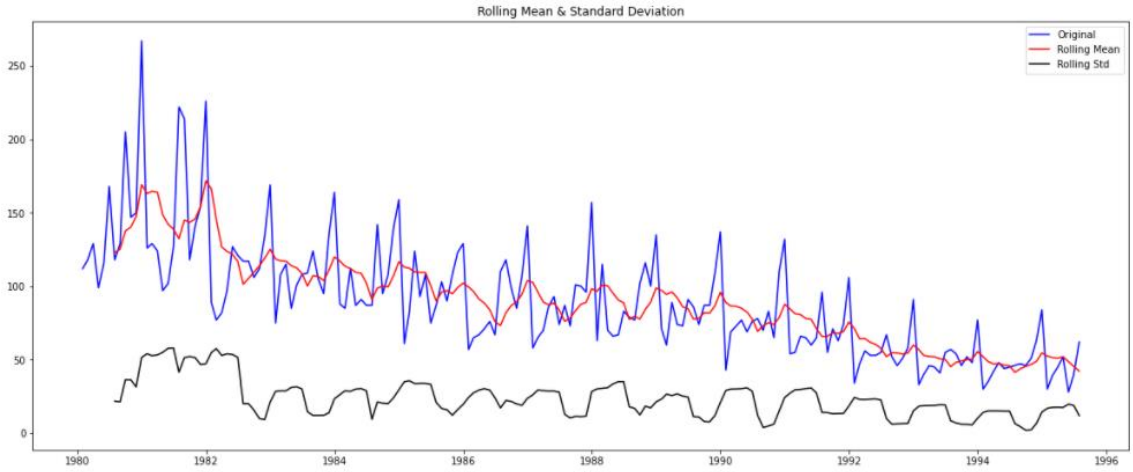


2.5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and also mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

Dickey Fuller Test

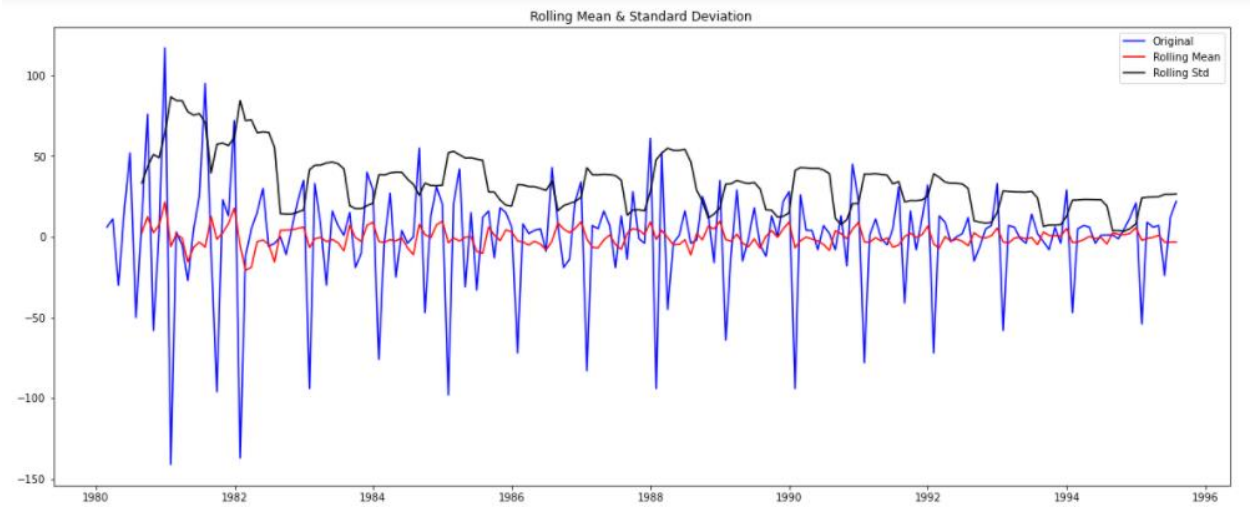
Null Hypothesis $H0$ - Series is not Stationary

Alternative Hypothesis $H1$ - Series is Stationary




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.880931
p-value                          0.341084
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

We see that at 5% significant level 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                -8.044820e+00
p-value                        1.806363e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical Value (10%)          -2.575756e+00
dtype: float64
```

We see that after taking a difference of order 1 the series have become stationary at α = 0.05.

2.6. 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.

Automated Version of ARIMA

The following loop helps us in getting a combination of different parameters of p and q in the range of 0 and 2 We have kept the value of d as 1 as we need to take a difference of the series to make it stationary.

```
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)
```

Model calculated for different p and q values and sorted with lowest AIC values.

	param	AIC
2	(0, 1, 2)	1276.835382
5	(1, 1, 2)	1277.359227
4	(1, 1, 1)	1277.775747
7	(2, 1, 1)	1279.045689
8	(2, 1, 2)	1279.298694
1	(0, 1, 1)	1280.726183
6	(2, 1, 0)	1300.609261
3	(1, 1, 0)	1319.348311
0	(0, 1, 0)	1335.152658

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                 D.Rose   No. Observations:                  131
Model:                 ARIMA(0, 1, 2)   Log Likelihood                -634.418
Method:                       css-mle   S.D. of innovations             30.168
Date:                Sat, 04 Sep 2021   AIC                           1276.835
Time:                        20:02:33   BIC                           1288.336
Sample:                    02-29-1980   HQIC                          1281.509
                         - 12-31-1990
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          -0.4885      0.085     -5.742      0.000      -0.655      -0.322
ma.L1.D.Rose   -0.7600      0.101     -7.500      0.000      -0.959      -0.561
ma.L2.D.Rose   -0.2398      0.095     -2.517      0.012      -0.427      -0.053
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
MA.1            1.0001           +0.0000j            1.0001            0.0000
MA.2           -4.1695           +0.0000j            4.1695            0.5000
-----------------------------------------------------------------------------
```

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

RMSE VALUES: 15.60

Automated Version of SARIMA

The following loop helps us in getting a combination of different parameters of p, q, P, Q in the range of 0 and 3 We have kept the value of d in the range (1,2) and D in range (0,1) .

```
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)
```

Model calculated for different p, q, d, P, Q,D values and sorted by least AIC values.

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	887.937509
80	(2, 1, 2)	(2, 0, 2, 12)	890.668798
69	(2, 1, 1)	(2, 0, 0, 12)	896.518161
53	(1, 1, 2)	(2, 0, 2, 12)	896.686897
78	(2, 1, 2)	(2, 0, 0, 12)	897.346444

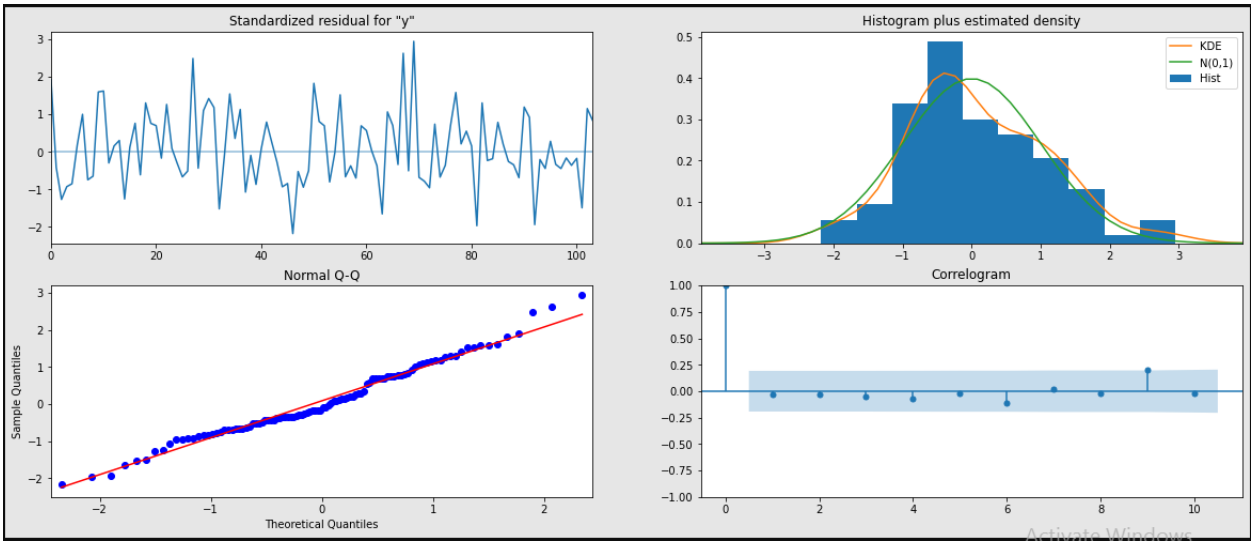
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-436.969
Date:	Sat, 04 Sep 2021	AIC	887.938
Time:	20:03:38	BIC	906.448
Sample:	0	HQIC	895.437
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.8427	189.512	-0.004	0.996	-372.279	370.593
ma.L2	-0.1573	29.773	-0.005	0.996	-58.512	58.197
ar.S.L12	0.3467	0.079	4.375	0.000	0.191	0.502
ar.S.L24	0.3023	0.076	3.996	0.000	0.154	0.451
ma.S.L12	0.0767	0.133	0.577	0.564	-0.184	0.337
ma.S.L24	-0.0726	0.146	-0.498	0.618	-0.358	0.213
sigma2	251.3136	4.76e+04	0.005	0.996	-9.31e+04	9.36e+04

Ljung-Box (L1) (Q):	0.10	Jarque-Bera (JB):	2.33
Prob(Q):	0.75	Prob(JB):	0.31
Heteroskedasticity (H):	0.88	Skew:	0.37
Prob(H) (two-sided):	0.70	Kurtosis:	3.03

Diagnostic plot:



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

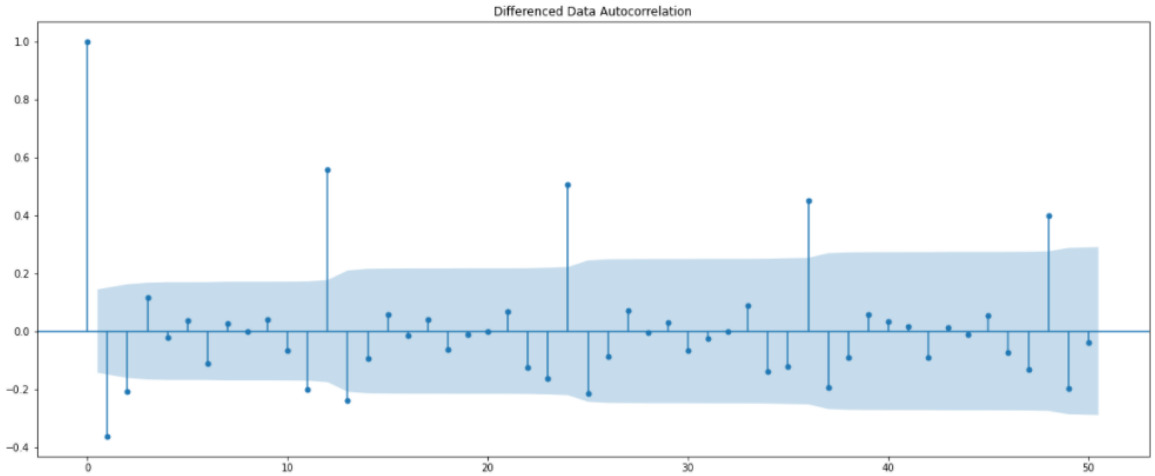
Forecast the model

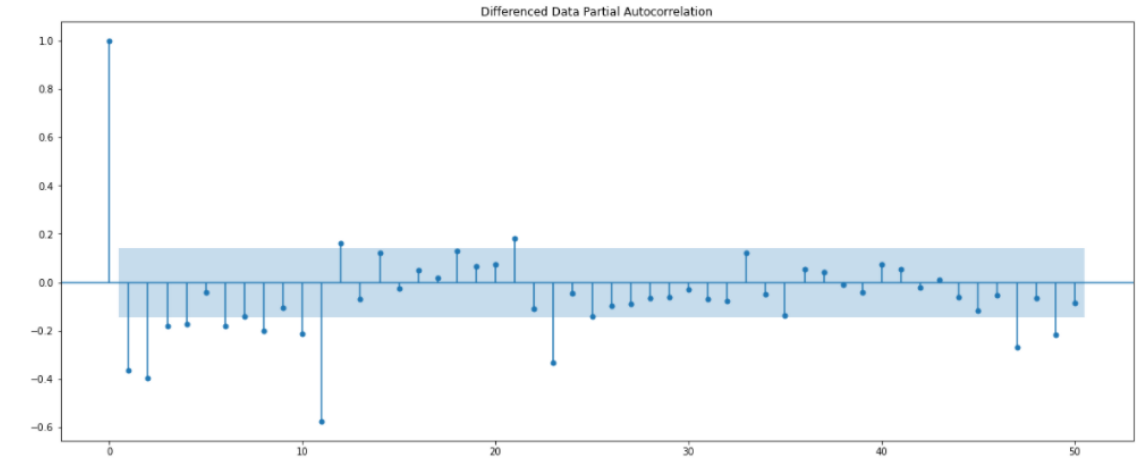
y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	62.867261	15.928500	31.647975	94.086547
1	70.541189	16.147658	38.892362	102.190017
2	77.356410	16.147655	45.707587	109.005233
3	76.208813	16.147655	44.559990	107.857637
4	72.747397	16.147655	41.098574	104.396220

RMSE: 26.88

2.7. 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.

ARIMA Model based on ACF and PACF





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 0.
The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 0.
By looking at the above plots, we can say that both the PACF and ACF plot cuts-off at lag 4 and 2.

ARIMA Model Results

Dep. Variable:	D.Rose	No. Observations:	131
Model:	ARIMA(4, 1, 2)	Log Likelihood	-633.876
Method:	css-mle	S.D. of innovations	29.793
Date:	Sat, 04 Sep 2021	AIC	1283.753
Time:	20:03:41	BIC	1306.754
Sample:	02-29-1980 - 12-31-1990	HQIC	1293.099

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.1905	0.576	-0.331	0.741	-1.319	0.938
ar.L1.D.Rose	1.1685	0.087	13.391	0.000	0.997	1.340
ar.L2.D.Rose	-0.3562	0.132	-2.693	0.007	-0.616	-0.097
ar.L3.D.Rose	0.1855	0.132	1.402	0.161	-0.074	0.445
ar.L4.D.Rose	-0.2227	0.091	-2.443	0.015	-0.401	-0.044
ma.L1.D.Rose	-1.9506	nan	nan	nan	nan	nan
ma.L2.D.Rose	1.0000	nan	nan	nan	nan	nan

Roots

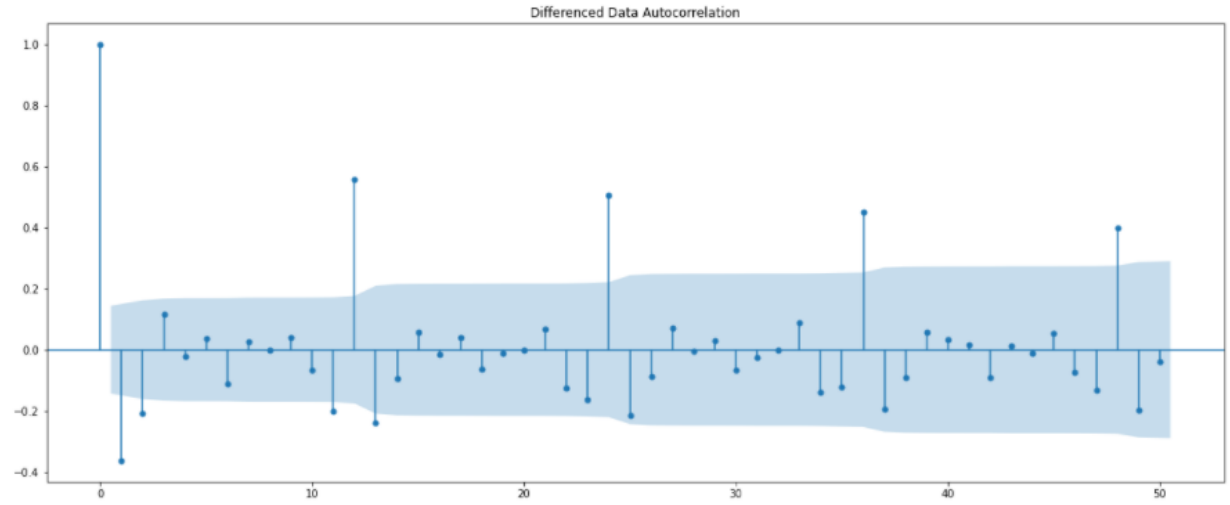
	Real	Imaginary	Modulus	Frequency
AR.1	1.1027	-0.4116j	1.1770	-0.0569
AR.2	1.1027	+0.4116j	1.1770	0.0569
AR.3	-0.6862	-1.6643j	1.8002	-0.3122
AR.4	-0.6862	+1.6643j	1.8002	0.3122
MA.1	0.9753	-0.2209j	1.0000	-0.0355
MA.2	0.9753	+0.2209j	1.0000	0.0355

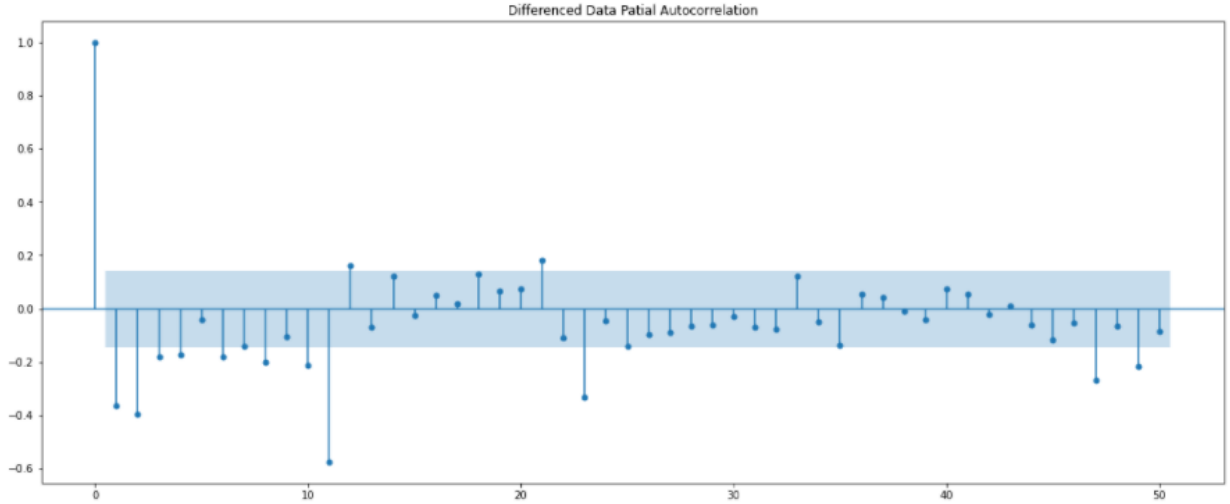
Predict on the Test Set using this model and evaluate the model.

RMSE: 33.90

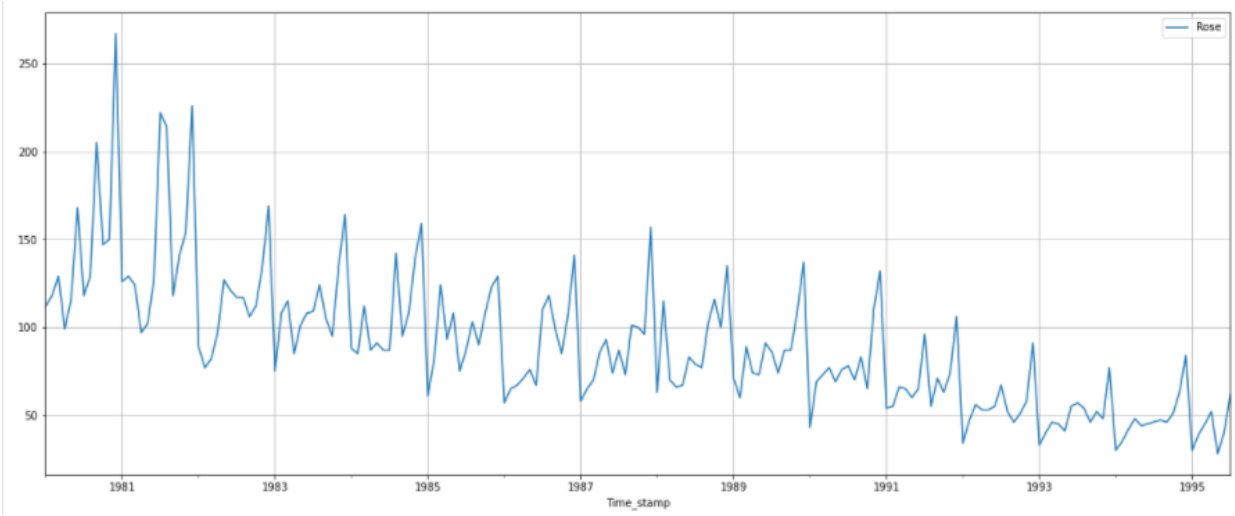
SARIMA Model : Manually looking at ACF and PACF

Let us look at the ACF and the PACF plots once more.

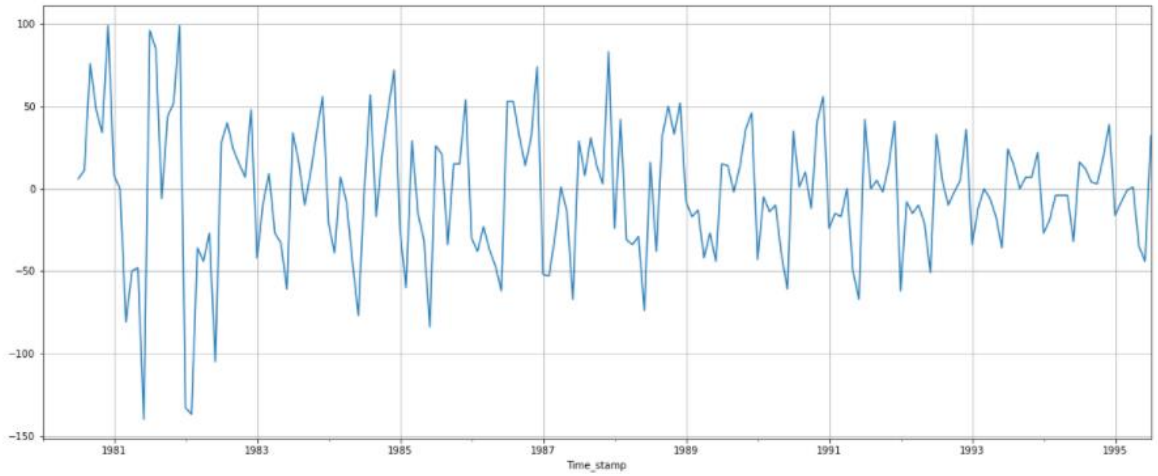


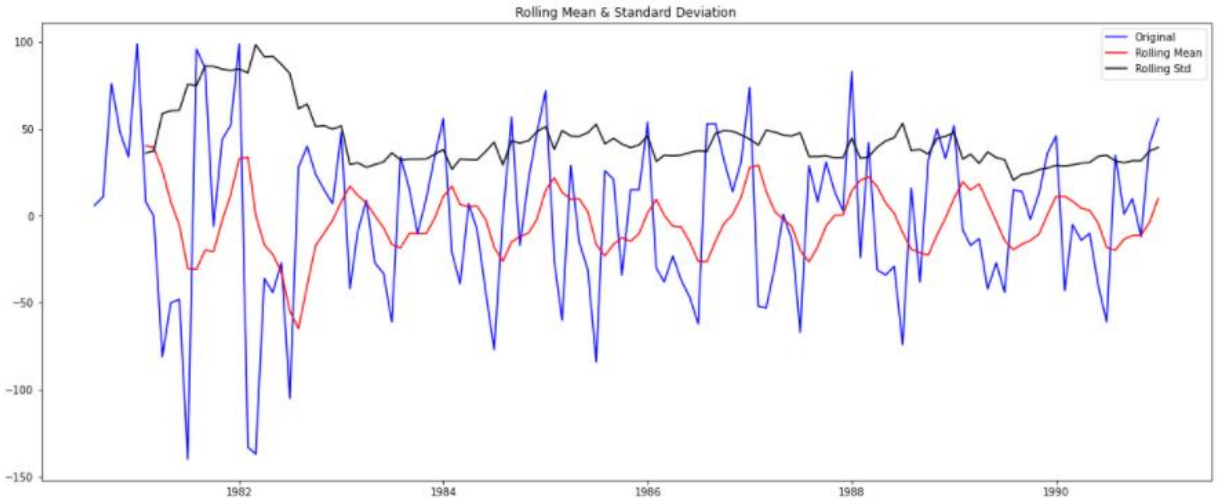


Plot the dataset:



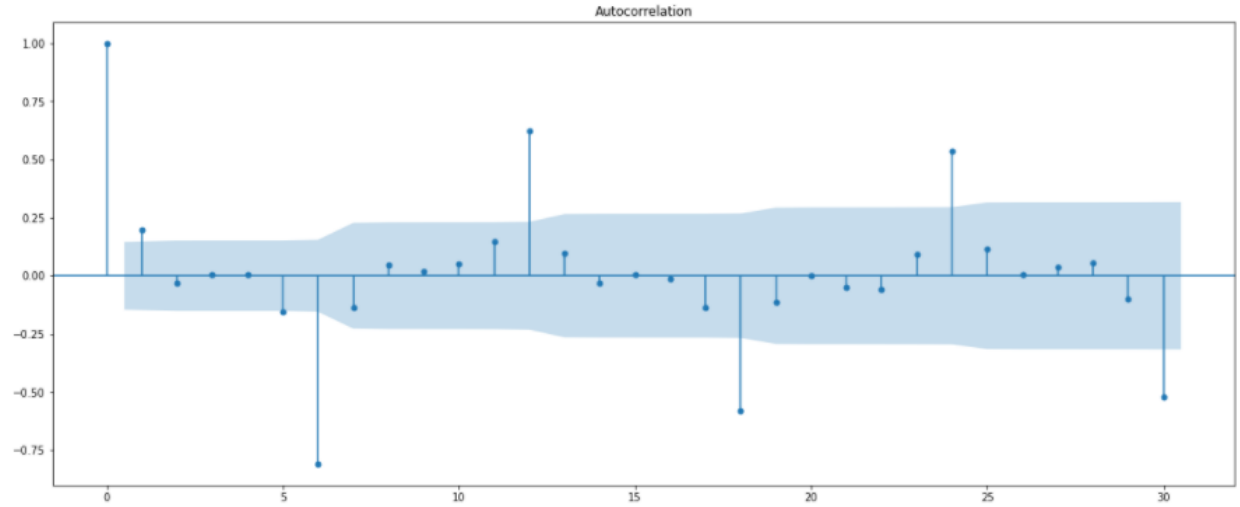
We see that there is a seasonality. So, now we take a seasonal differencing and check the series.

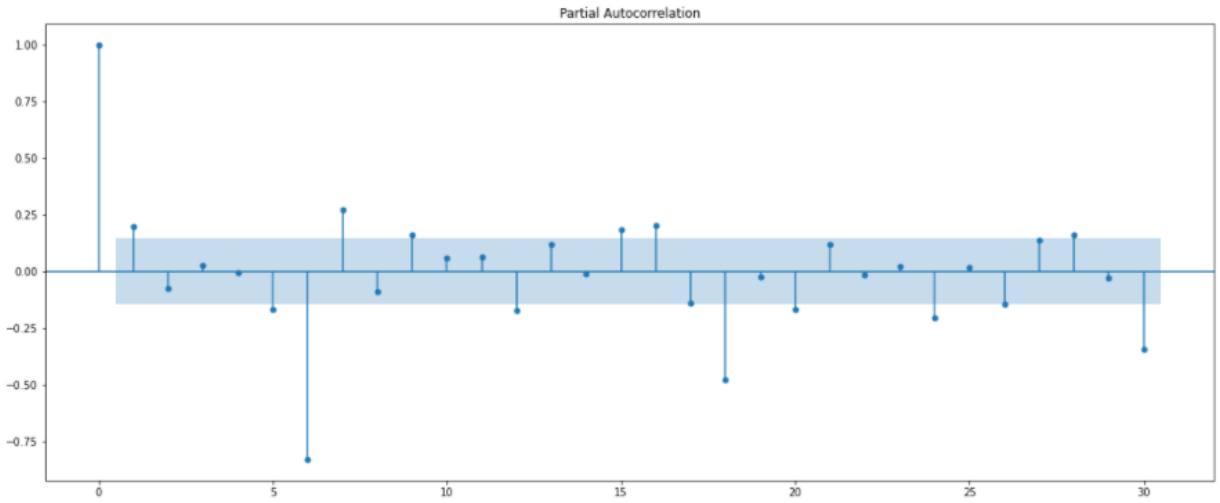





```
Results of Dickey-Fuller Test:
Test Statistic                -7.442449e+00
p-value                        5.956534e-11
#Lags Used                     7.000000e+00
Number of Observations Used    1.180000e+02
Critical Value (1%)           -3.487022e+00
Critical Value (5%)           -2.886363e+00
Critical Value (10%)          -2.580009e+00
dtype: float64
```

ACF and PACF Plot after differencing:





Here, we have taken alpha=0.05.

We are going to take the seasonal period as 6. We will keep the p(1) and q(1) parameters same as the ARIMA model.

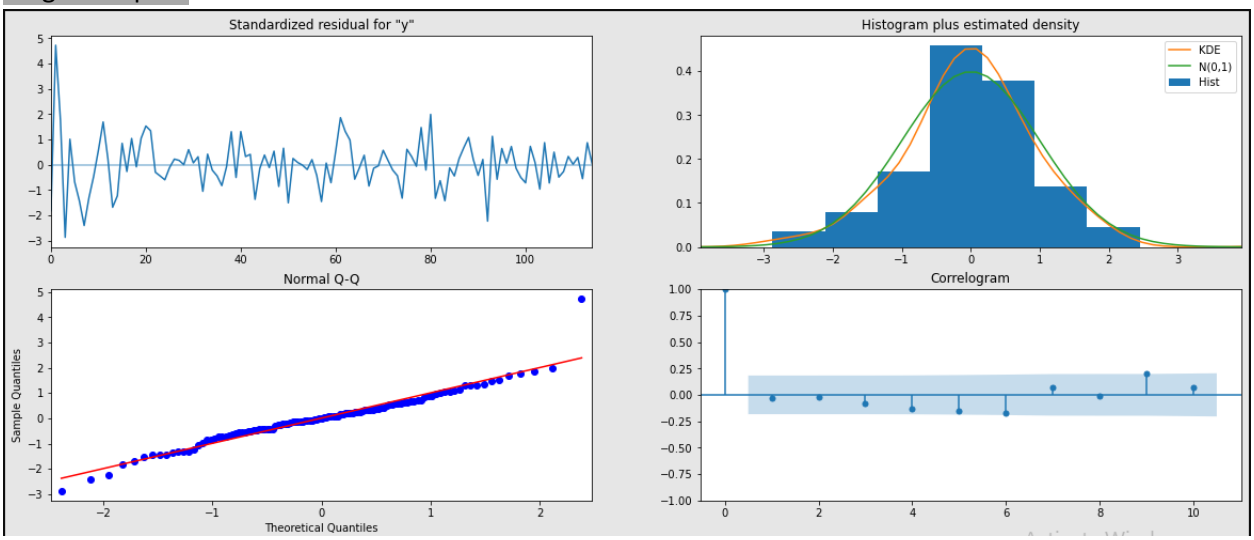
The Auto-Regressive parameter in an SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.

The Moving-Average parameter in an SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 0. Remember to check the ACF and the PACF plots only at multiples of 6 (since 6 is the seasonal period).

By looking at the plots we see that the ACF and the PACF cut-off at 1 & 1 .

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(4, 1, 2)x(1, 1, [1], 6)   Log Likelihood                -540.333
Date:                             Sat, 04 Sep 2021   AIC                           1098.665
Time:                                     20:03:46   BIC                           1123.370
Sample:                                          0   HQIC                          1108.693
                                               - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4668      0.356     -1.310      0.190      -1.165       0.231
ar.L2         -0.6301      0.196     -3.207      0.001      -1.015      -0.245
ar.L3         -0.3466      0.204     -1.702      0.089      -0.746       0.053
ar.L4         -0.2476      0.116     -2.140      0.032      -0.474      -0.021
ma.L1         91.3511      4.508     20.263      0.000      82.515     100.187
ma.L2        -12.3925     34.059     -0.364      0.716     -79.147      54.362
ar.S.L6       -0.8436      0.047    -18.037      0.000      -0.935      -0.752
ma.S.L6       26.3998    108.892      0.242      0.808    -187.025     239.825
sigma2         0.0001      0.001      0.122      0.903      -0.002       0.002
==========================================================================================
Ljung-Box (L1) (Q):                   0.09   Jarque-Bera (JB):                68.56
Prob(Q):                              0.76   Prob(JB):                         0.00
Heteroskedasticity (H):               0.38   Skew:                             0.57
Prob(H) (two-sided):                  0.00   Kurtosis:                         6.61
==========================================================================================
```

Diagnostic plot:



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

RMSE: 23.03

Forecast test set with confidence interval.

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	43.074435	26.574002	-9.009651	95.158521
1	65.094398	28.707918	8.827913	121.360883
2	70.508113	28.742646	14.173563	126.842663
3	73.747771	29.470888	15.985893	131.509649
4	75.722800	30.800047	15.355818	136.089782

2.8. 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.

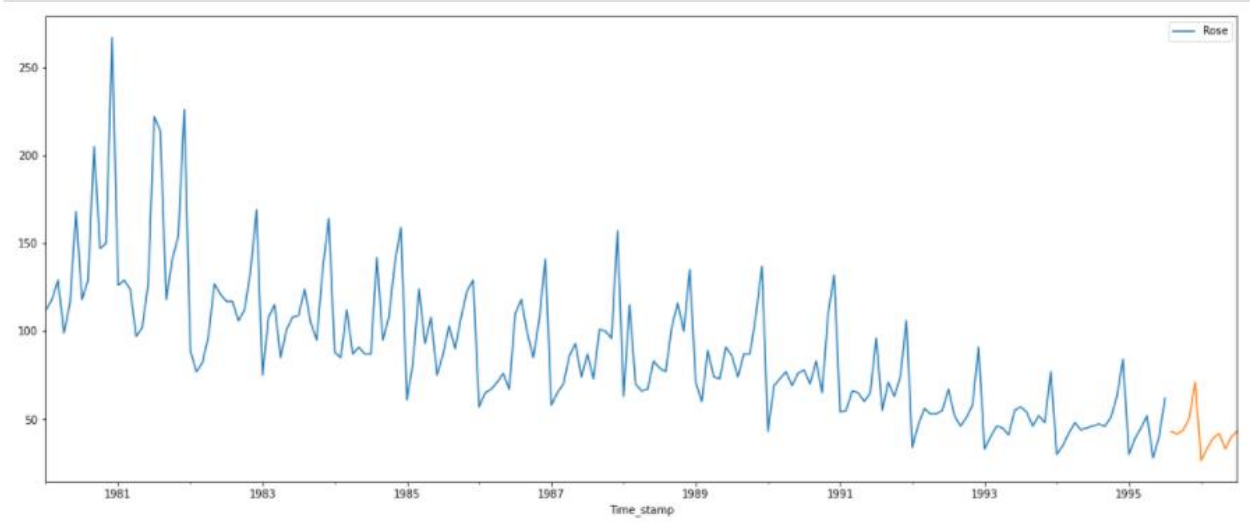
	Test RMSE
Alpha=0.3,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	10.935749
2pointTrailingMovingAverage	11.529994
4pointTrailingMovingAverage	14.444342
6pointTrailingMovingAverage	14.554944
9pointTrailingMovingAverage	14.721499
RegressionOnTime	15.255435
ARIMA(0,1,2)	15.605942
Alpha=0.075,Beta=0.040,Gamma=0.0004,TripleExponentialSmoothing	19.322173
SARIMA(4,1,2)(1,1,1,6) based on ACF & PACF	23.037273
SARIMA(0,1,2)(2,0,2,12)	26.880625
ARIMA (4,1,2) based on ACF & PACF	33.903744
Alpha=0.995,SimpleExponentialSmoothing	36.748162
Alpha=0.3,SimpleExponentialSmoothing	47.457057
SimpleAverageModel	53.413057
NaiveModel	79.672238
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	265.509912

2.9. 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.

Best model is the Triple Exponential Smoothing with multiplicative seasonality with the parameters α = 0.3, β = 0.3 and γ = 0.3.Building the most optimum model on the Full Data.

RMSE: 19.95

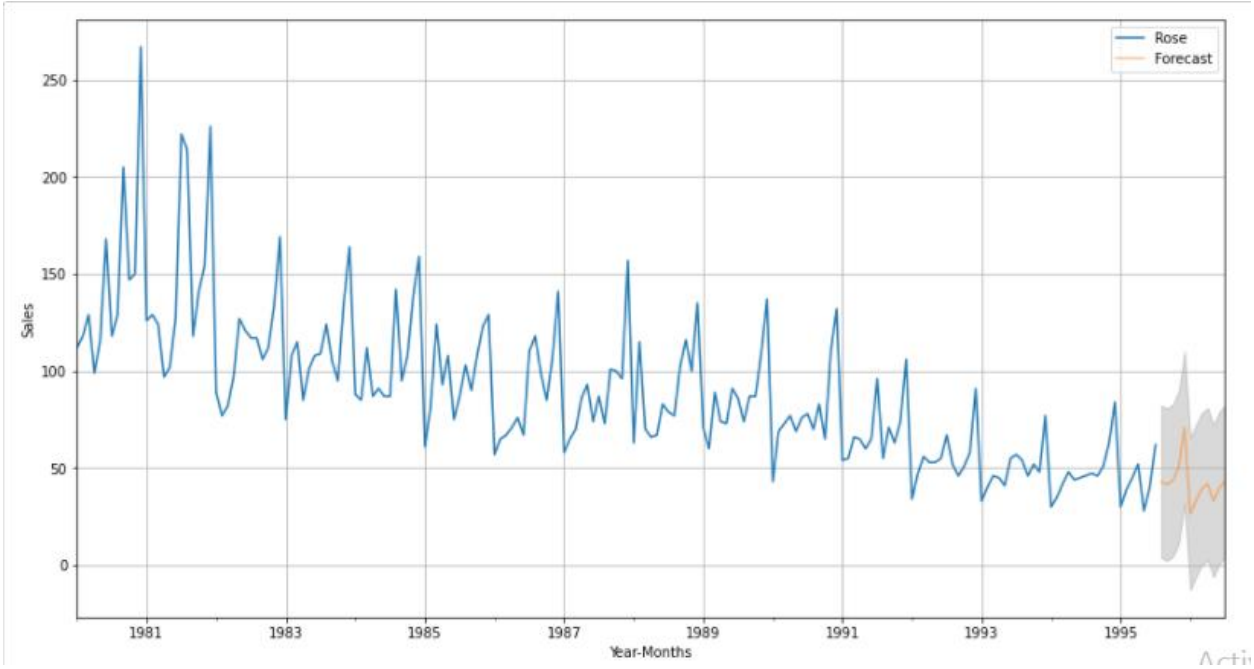
Getting the predictions for future in 12 steps and plot it.



Calculated the upper and lower confidence bands at 95% confidence level. Here multiplier to be 1.96 as we want to plot with respect to a 95% confidence intervals.

	lower_CI	prediction	upper_ci
1995-08-31	3.801629	43.008486	82.215344
1995-09-30	2.303920	41.510778	80.717636
1995-10-31	4.472140	43.678998	82.885855
1995-11-30	11.061361	50.268219	89.475076
1995-12-31	31.744942	70.951800	110.158658

Plot the forecast along with the confidence band



2.10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. Please explain and summarize the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Prediction:

```
1995-08-31    43.008486
1995-09-30    41.510778
1995-10-31    43.678998
1995-11-30    50.268219
1995-12-31    70.951800
1996-01-31    26.448111
1996-02-29    33.419382
1996-03-31    39.277688
1996-04-30    41.829250
1996-05-31    33.017518
1996-06-30    39.751866
1996-07-31    43.310090
Freq: M, dtype: float64
```

Describe Prediction:

```
count    12.000000
mean     42.206016
std      10.956609
min      26.448111
25%      37.813111
50%      41.670014
75%      43.402317
max      70.951800
dtype: float64
```

- Sales of Rose Wine (1980 - 1995) were analysed. Hidden details are captured by doing EDA.
- From 1981 sales decreasing throughout 1995.
- Median : 85.
- Highest Sale on 1980, whereas lowest sale on 1995.
- It is evident from monthly plot that sales has been increased from August to December. Stock has to be more during these time frame. January recorded the lowest sale which is right after the month of December of previous years.
- Trend and Seasonality are there in the dataset.
- Dataset were split for training and test set. Various models such as Linear regression, Naïve bayes, Simple Exponential smoothing, Double exponential smoothing, Triple exponential smoothing, ARIMA and SARIMA built on training data and tested on test data.
- Since the dataset has seasonality, SARIMA model would be best suited model. But from RMSE values of various models Triple exponential smoothing is best for the dataset. And same was applied on full data.
- Sales for next 12 month is predicted with confidence interval. Sales are varying drastically across the month. Trend for Rose wine is decreasing drastically. it recommended to conduct the survey about the wine, accordingly action should be taken care.
- Quality of Rose wine has to get increased to improve the sales.