Model Solution

Time Series Forecasting Assessment

DSBA

Contents

TSF - Sparkling

1. Time Series Forecast for Sparkling Dataset

1.1. Get the data and start analysis

Let us get started.

Load the required packages, set the working directory and load the data file.

We would be reading the file, by proving the index_col as 'YearMonth'

Dataset has 210 rows and 7 features. There is one features which is float64 type.

Let us start the data exploration step with the head function to look at first 5 initial rows.

Let us check the head and tail of the dataset

Head

	Sparkling
YearMonth	
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

Tail

	Sparkling
YearMonth	
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

Let us check the data types of the different variables present in the dataset

Datetime Index: 187 entries, 1980-01-01 to 1995-07-01

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

There are 187 data points, with no null or missing values and the 'Sparkling' variable is of integer type.

The number of rows: 187
The number of columns: 1

The data is loaded with the index as YearMonth

5 point summary for the dataset

	Sparkling
count	187.00
mean	2402.42
std	1295.11
min	1070.00
25%	1605.00
50%	1874.00
75%	2549.00
max	7242.00

The mean value is 2402.42 and the median is 1874. Range is 6171.

1.2. Exploratory Data Analysis

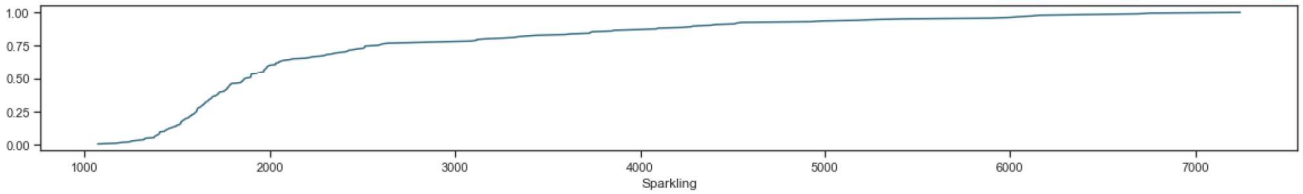
1.2.1. Univariate Time Series

A univariate time series is a series with a single time-stamped variable at time t.

Here the dataset belongs to the Sparkling Wine sales from the January of 1980 to July of 1995. Here, Sparkling is the time-dependent variable.

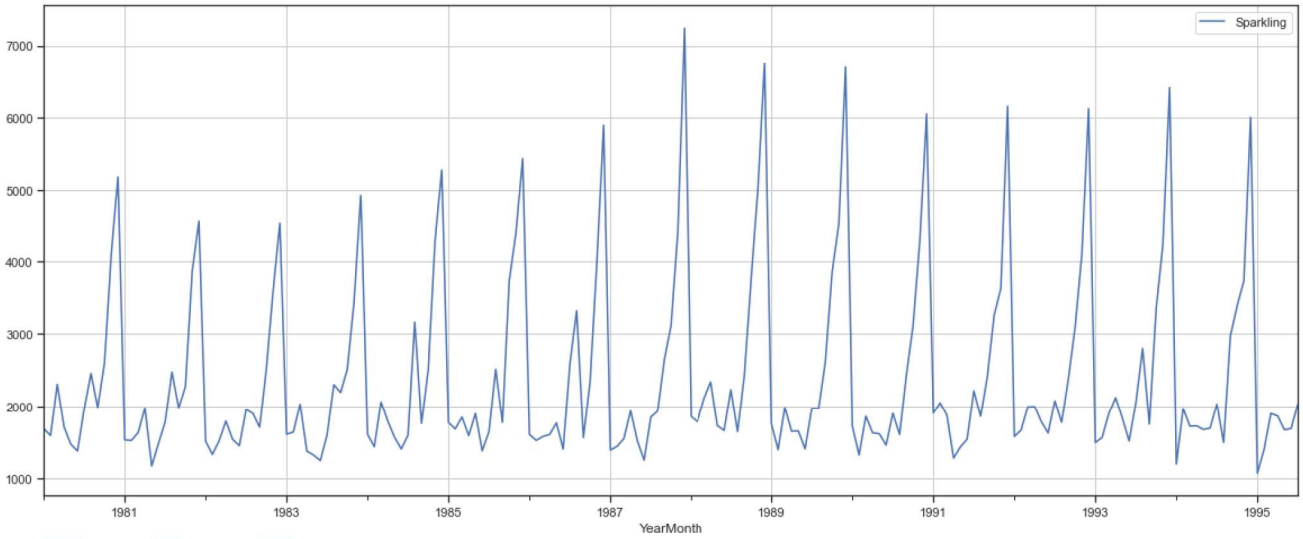
The series is a monthly series, wherein for each month between Jan-1980 and Jul-1995 a datapoint is recorded.

1.2.2. Plot ECDF: Empirical Cumulative Distribution Function



An ECDF is an estimator of the Cumulative Distribution Function. The ECDF essentially allows you to plot a feature of your data in order from least to greatest and see the whole feature as if is distributed across the data set. The data ranges from 1070 to 7242.

1.2.3. Plot for the Sparkling timeseries.



Suppose we predict the Sparkling for the next few months, we will look at the past values and try to gauge and extract a pattern.

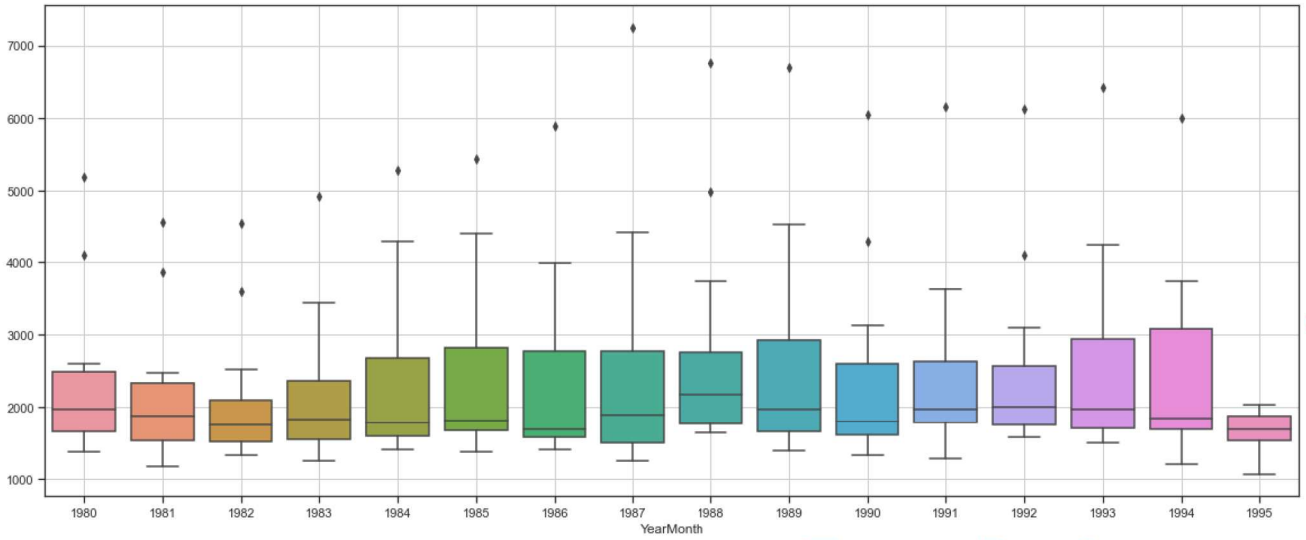
Here we observe a pattern within each year indicating a seasonal effect. Such observations will help us in predicting future values.

Note: We have used only one variable, Sparkling (the Sparkling sale of the past 15 years).

Hence, this is will be Univariate Time Series Analysis/Forecasting.

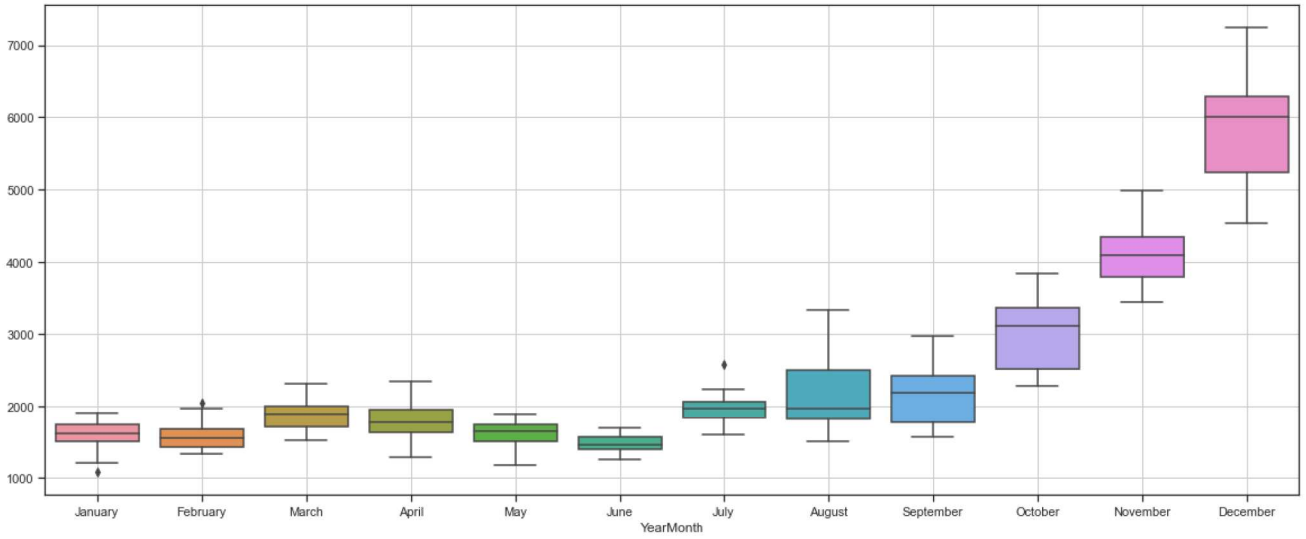
Trend might be absent but a Seasonal effect looks to be present. These can be inferred from the Time Series plot.

1.2.4. Box plot per year wise



Across the year there are not much noticeable difference, though there are few high ranges in middle years. Also the last year is showing less sale due to the fact that the data is recorded for 7 out of 12 months in year 1995.

1.2.5. Box plot per month wise



There is a maximum sale in month of December, followed by November and then October. These fluctuations can be attributed to the months being the holiday month or months. June shows the minimum sale.

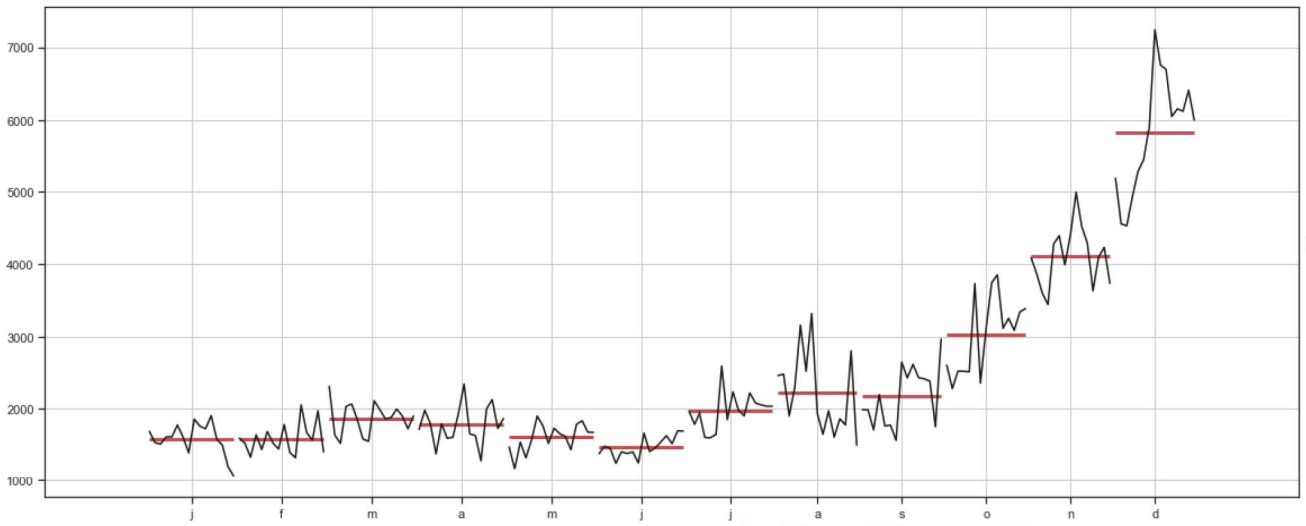
Do remember that there is one less observation for a few months in the year of 1995.

1.2.6. Year/Month Table

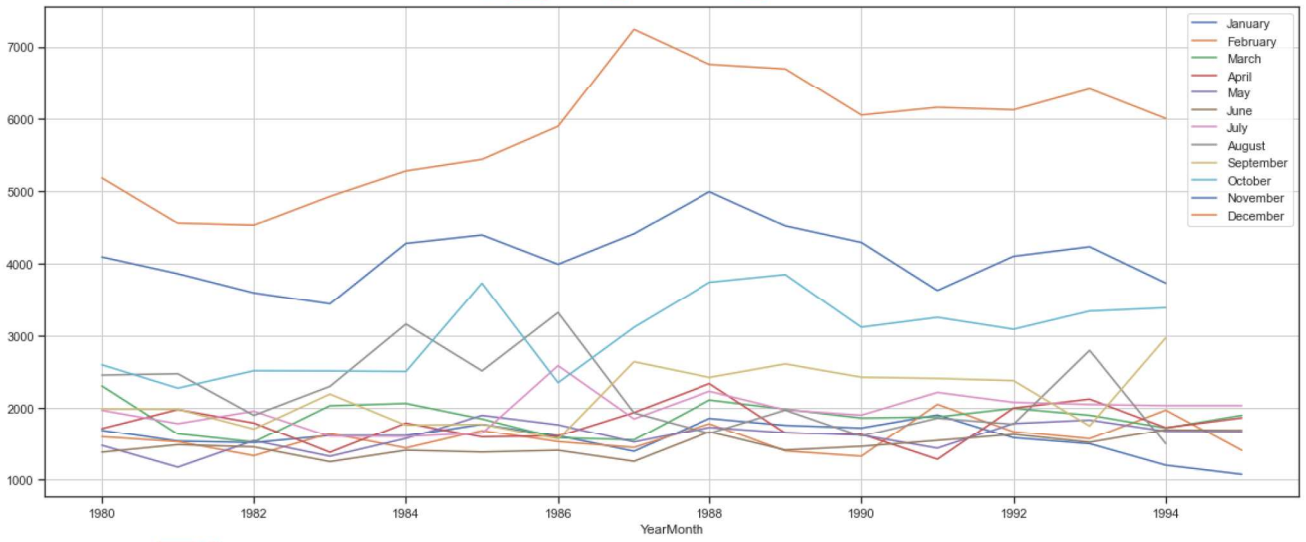
YearMonth	January	February	March	April	May	June	July	August	September	October	November	December
YearMonth												
1980	1686.0	1591.0	2304.0	1712.0	1471.0	1377.0	1966.0	2453.0	1984.0	2596.0	4087.0	5179.0
1981	1530.0	1523.0	1633.0	1976.0	1170.0	1480.0	1781.0	2472.0	1981.0	2273.0	3857.0	4551.0
1982	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

Post July 1995 there is no datapoint for 1995 year. This Aug to Dec are having one less datapoint than other months.

1.2.7. Month plot for each month distribution



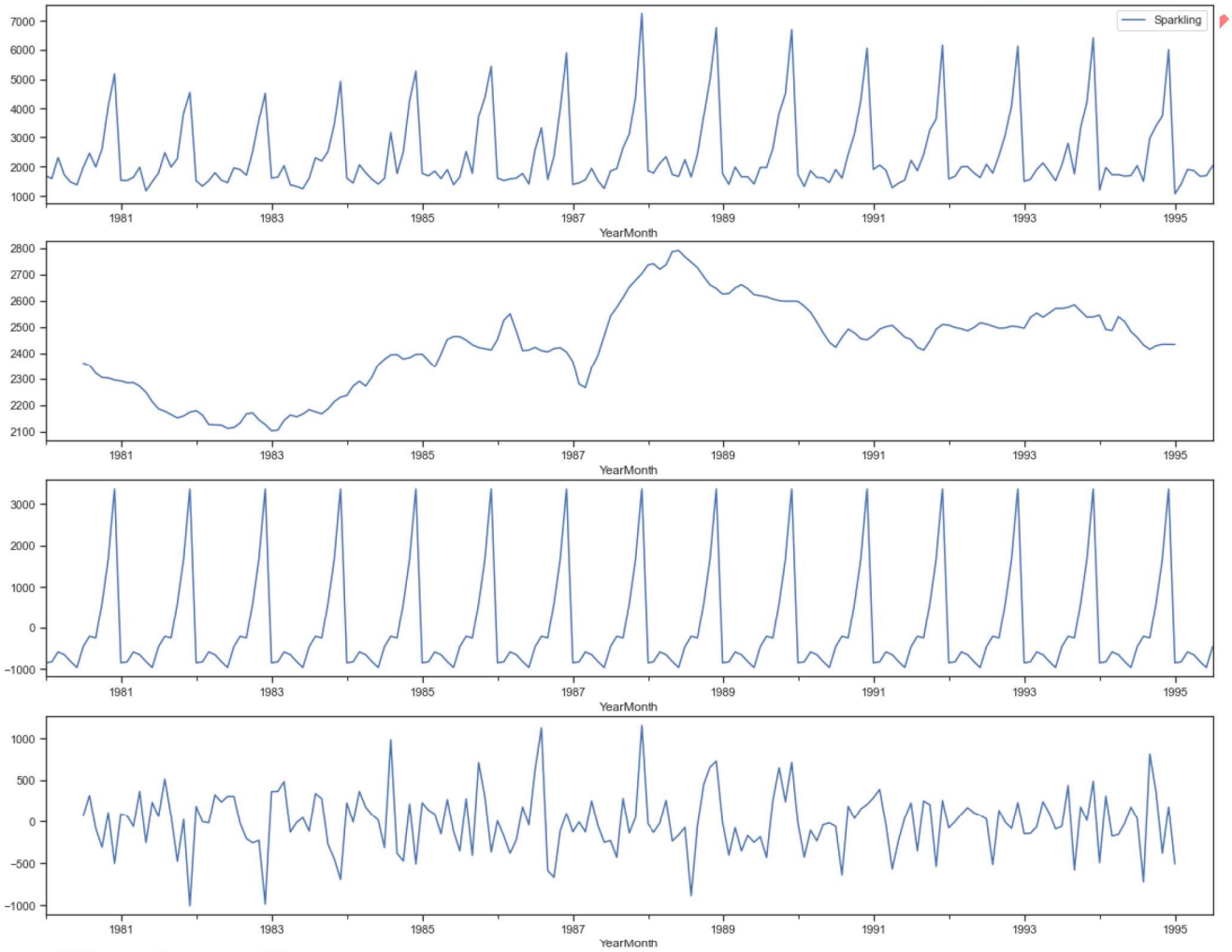
1.2.8. Line plot for comparison between each month of each year



As previously seen Month of December is outperforming for almost every year and it is another indicator that there is high out of seasonal sale in Oct-Dec.

1.2.9. Additive Decomposition

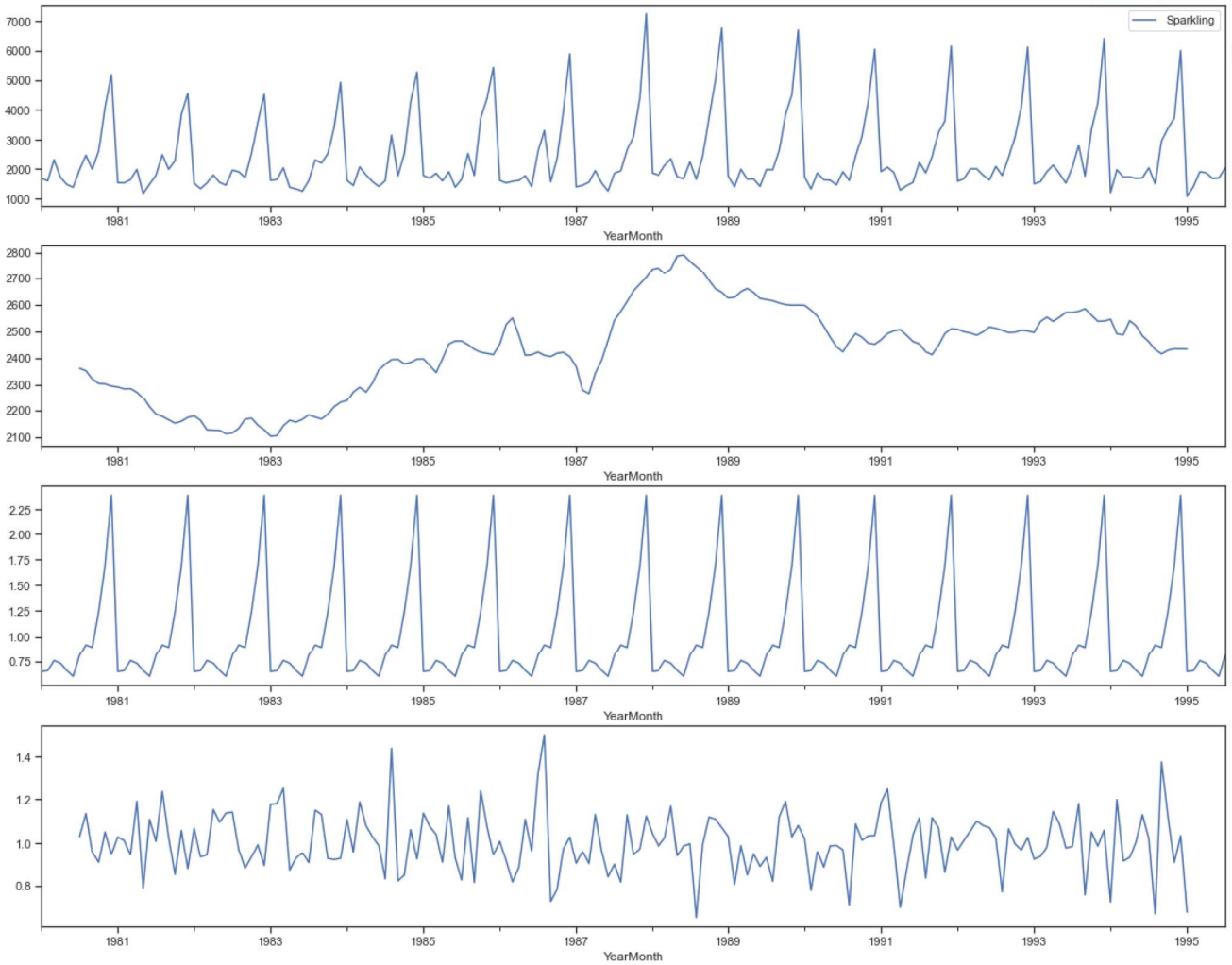
- An additive model suggests that the components are added together.
- An additive model is linear where the changes over time are consistently added by the same amount. The seasonal correction is added with the Trend.
- A linear seasonality has the same frequency (width of the cycles) and amplitude (height of the cycles).



Running the above code performs the decomposition and plots the 4 resulting series. We observe that the trend and seasonality are clearly separated.

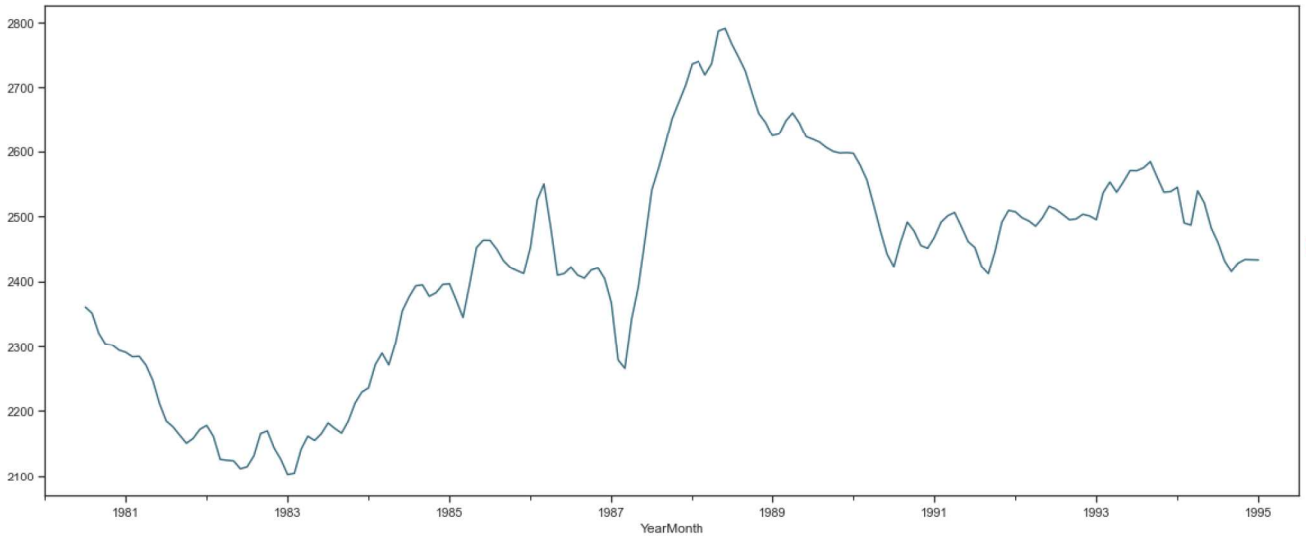
1.2.10. Multiplicative Decomposition

- A multiplicative model suggests that the components are multiplied together.
- A multiplicative model is non-linear.
- The seasonal correction is multiplied with the trend..
- A non-linear seasonality has an increasing or decreasing frequency (width of the cycles) and / or amplitude (height of the cycles) over time.



Running the above code performs the decomposition and plots the 4 resulting series. We observe that the trend and seasonality are clearly separated.

1.2.11. Closer look to the trend in the data set



There is an upward trend in the initial half which seems to be reach at a peak and then move in downward position. This will help in the forecast as in the most recent time it seems to have downward trend.

1.3. Split the data into train and test and plot the training and test data

- Split the data into train and test.
- Build different time series models on train data set and test it on test data set
- Compare the models' performances.
- The test data begins 1991 onwards. Anything before 1991 can considered in the training data so long they are contiguous.
- Size of the training dataset is (132, 1)
  Size of the test dataset is (55, 1)

First few rows of Training Data

YearMonth	Sparkling
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

Last few rows of Training Data

YearMonth	Sparkling
1990-08-01	1605
1990-09-01	2424
1990-10-01	3116
1990-11-01	4286
1990-12-01	6047

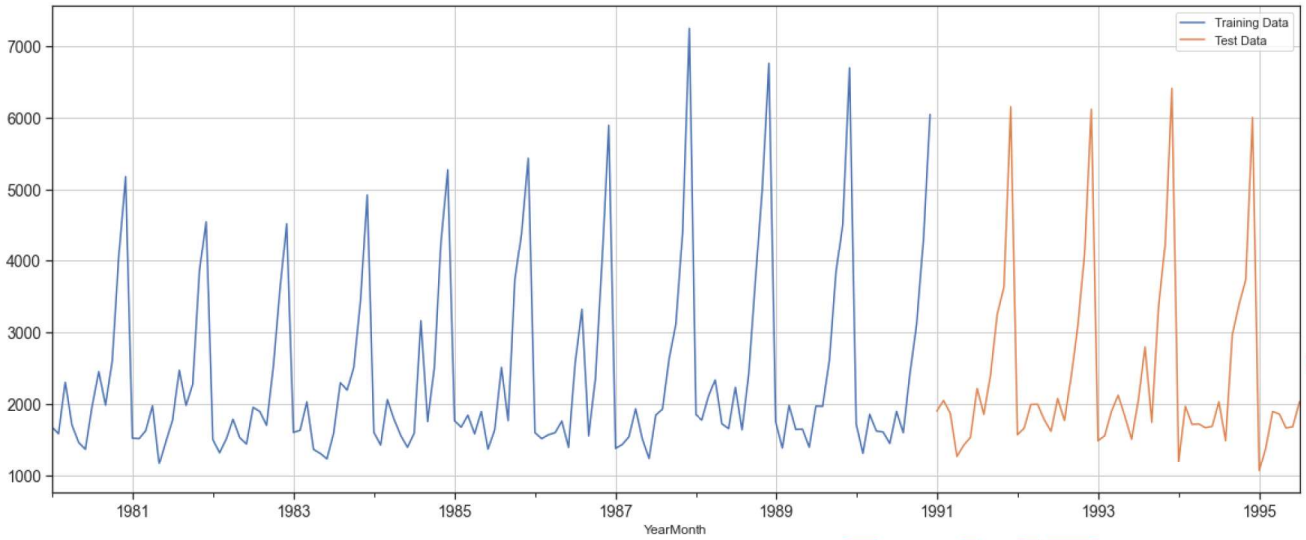
First few rows of Test Data

YearMonth	Sparkling
1991-01-01	1902
1991-02-01	2049
1991-03-01	1874
1991-04-01	1279
1991-05-01	1432

Last few rows of Test Data

YearMonth	Sparkling
1995-03-01	1897
1995-04-01	1862
1995-05-01	1670
1995-06-01	1688
1995-07-01	2031

1.3.1. Joint plot for training and test data

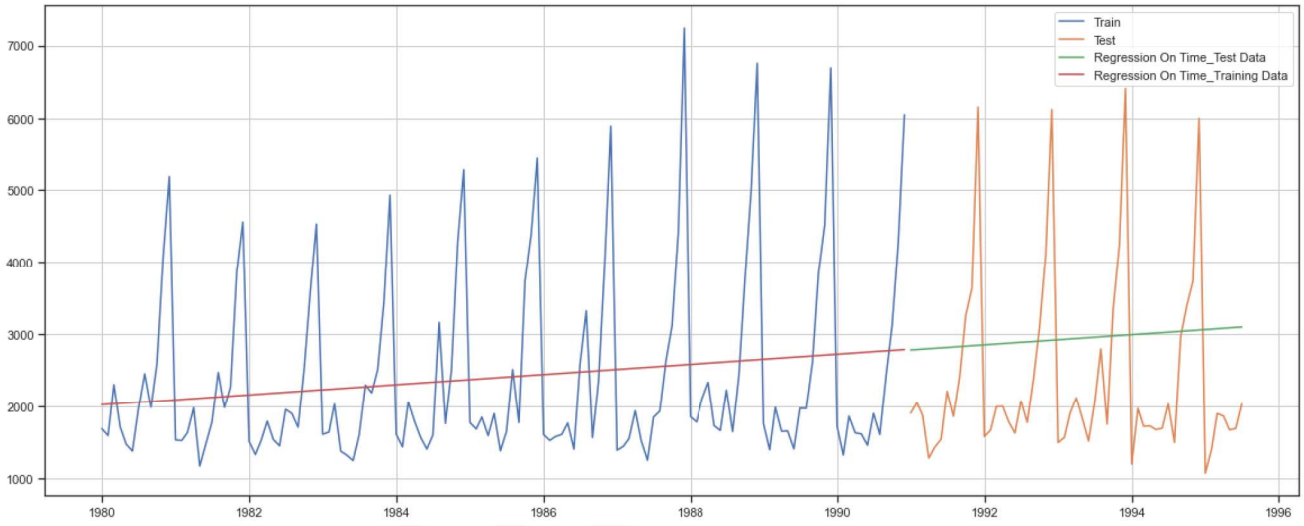


1.4. Building different models and comparing the accuracy metrics.

1.4.1. 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.

Model 1 Plot - Linear Regression



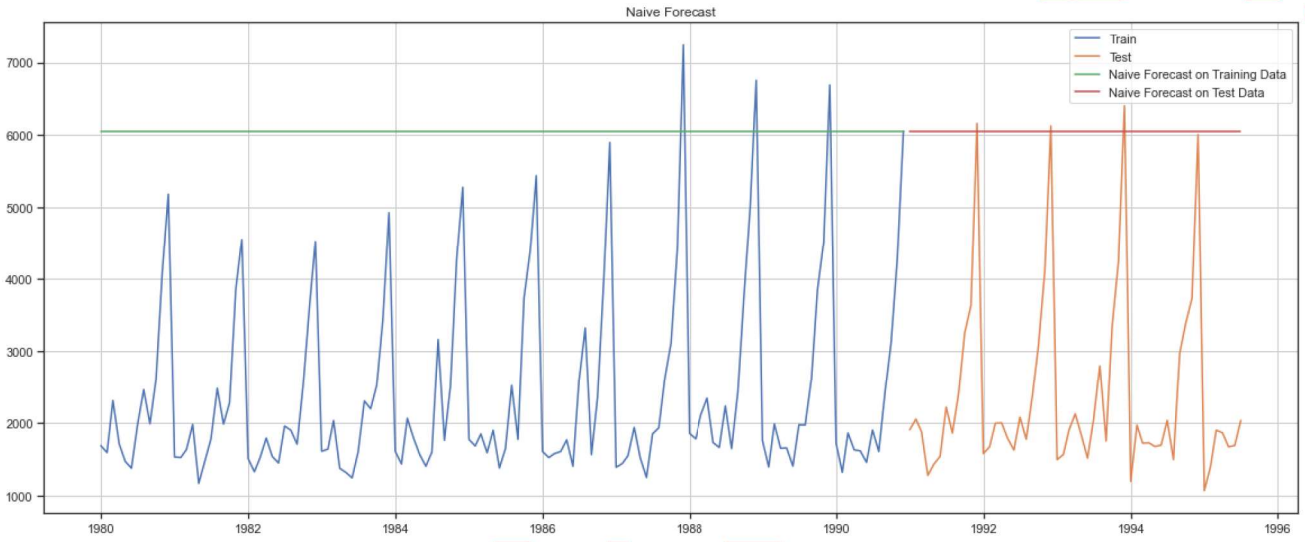
Model 1 Evaluation - Linear Regression

For RegressionOnTime forecast on the Training Data, RMSE is 1279.322
For RegressionOnTime forecast on the Test Data, RMSE is 1384.558

1.4.2. Model 2: Naive Approach

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.

Model 2 Plot - Naive



Model 2 Evaluation - Naive

For Naive Model forecast on the Training Data, RMSE is 3867.701
For Naive Model forecast on the Test Data , RMSE is 3864.279

We can infer from the RMSE values and the above graphs that the Naive method and Regression on Time models might not be suited for datasets with high variability.

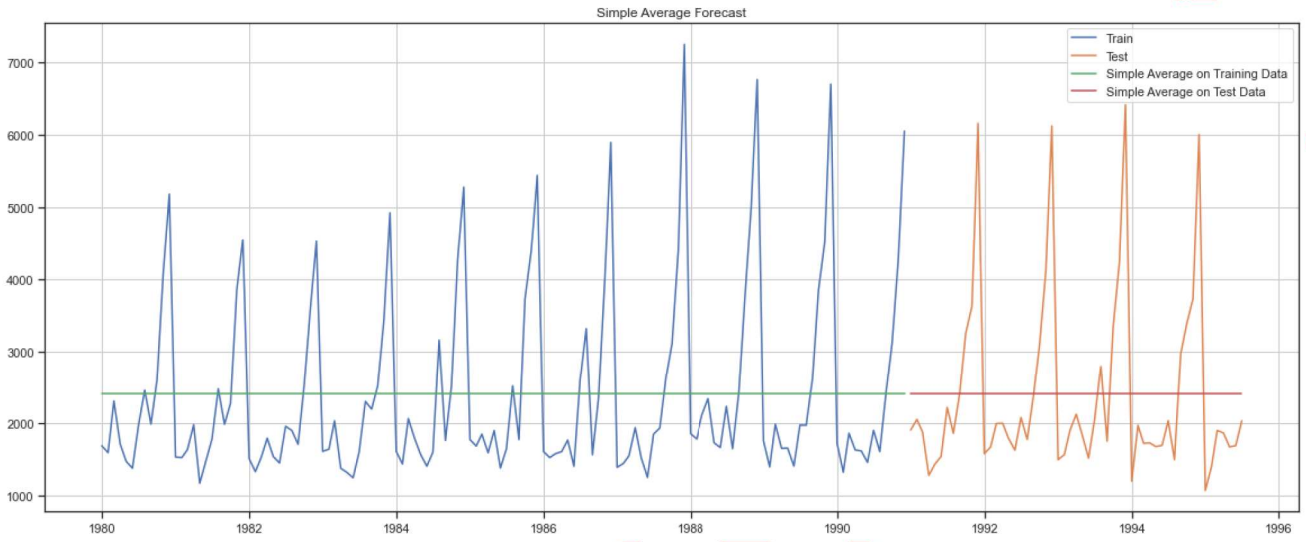
Naive method is best suited for stable datasets. We can still improve our score by adopting different techniques.

Now we will look at another technique and try to improve our prediction accuracy.

1.4.3. Model 3: Simple Average

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

Model 3 Plot - Simple Average



Model 3 Evaluation - Simple Average

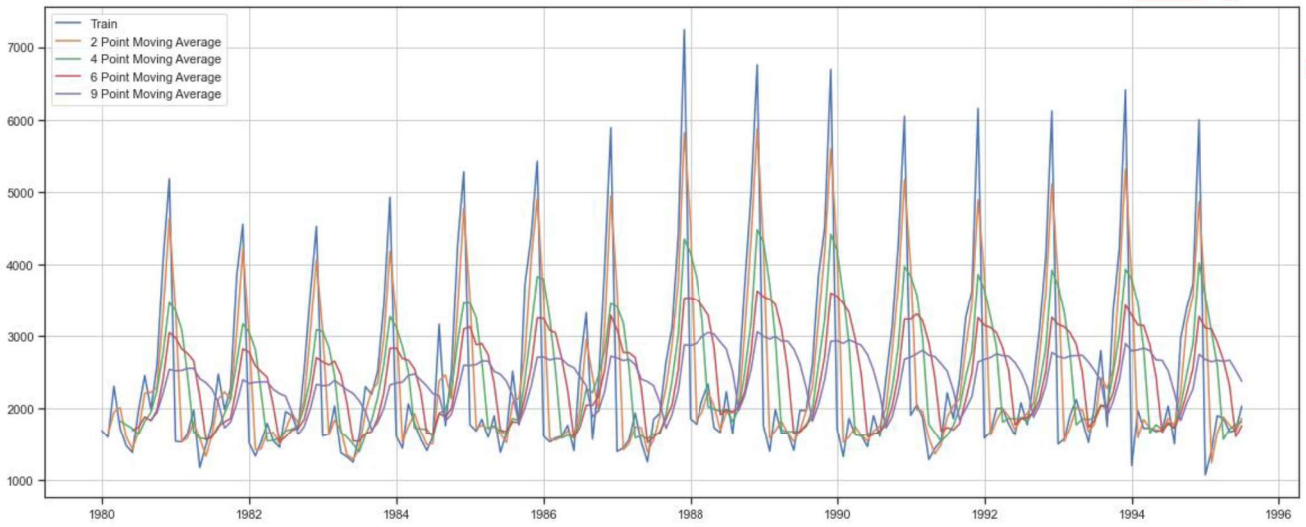
For Simple Average Model forecast on the Training Data, RMSE is 1298.484
For Simple Average forecast on the Test Data, RMSE is 1275.082

1.4.4. 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.

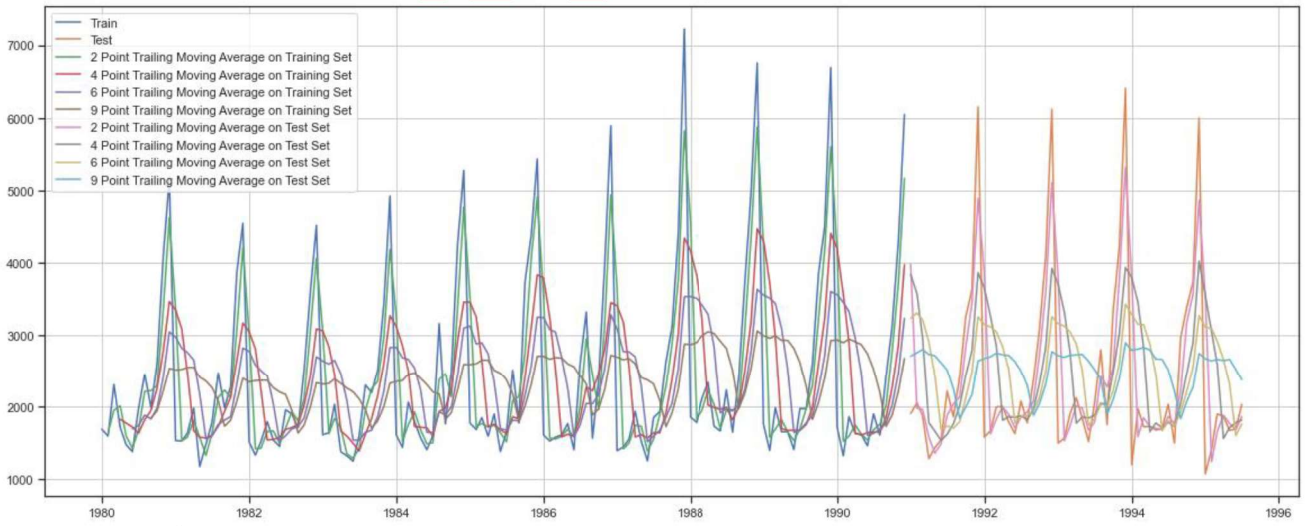
For Moving Average, we are going to average over the entire data.

Model 4 Plot - Moving Average on entire data



For Moving Average, we are going to average over the training data.

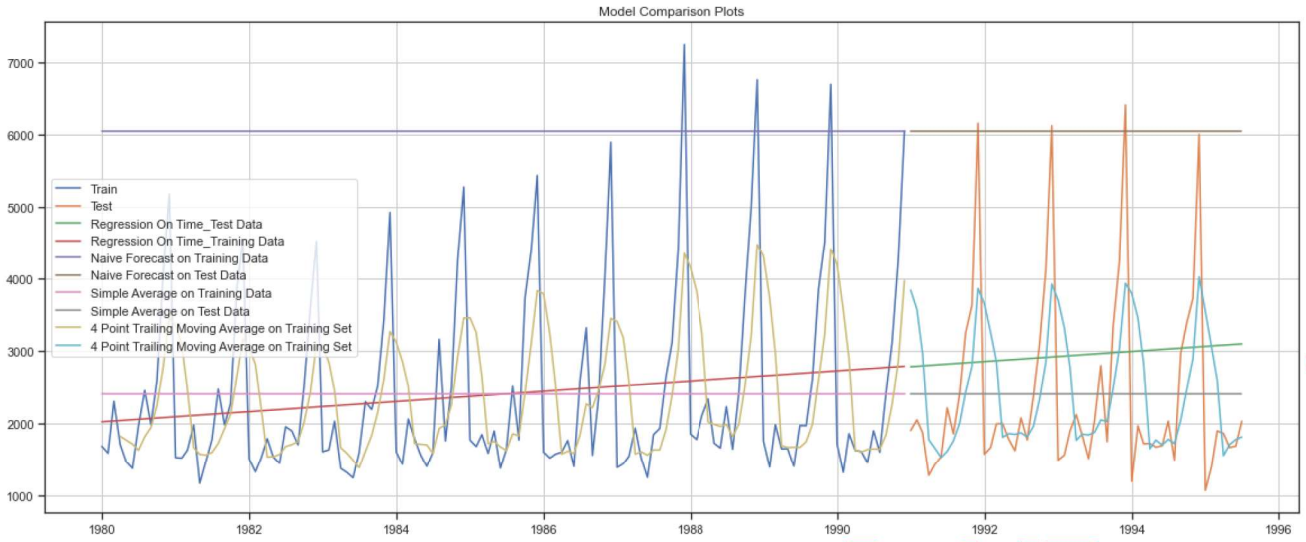
Model 4 Plot - Moving Average on Train and test Data separately



Model 4 Evaluation - Moving Average

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

Let us plot all the models done so far and compare the time Series plots.



1.4.5. Model 5: Double Exponential Smoothing (Holt's Model)

Two parameters Alpha and Beta are estimated in this model. Level and Trend are accounted for in this model.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	1592.292788	18259.110704
1	0.3	0.4	1682.573828	26069.841401
2	0.3	0.5	1771.710791	34401.512440
3	0.3	0.6	1848.576510	42162.748095
4	0.3	0.7	1899.949006	47832.397419
...	...	...	...	...
59	1.0	0.6	1753.402326	49327.087977
60	1.0	0.7	1825.187155	52655.765663
61	1.0	0.8	1902.013709	55442.273880
62	1.0	0.9	1985.368445	57823.177011

	Alpha Values	Beta Values	Train RMSE	Test RMSE
63	1.0	1.0	2077.672157	59877.076519

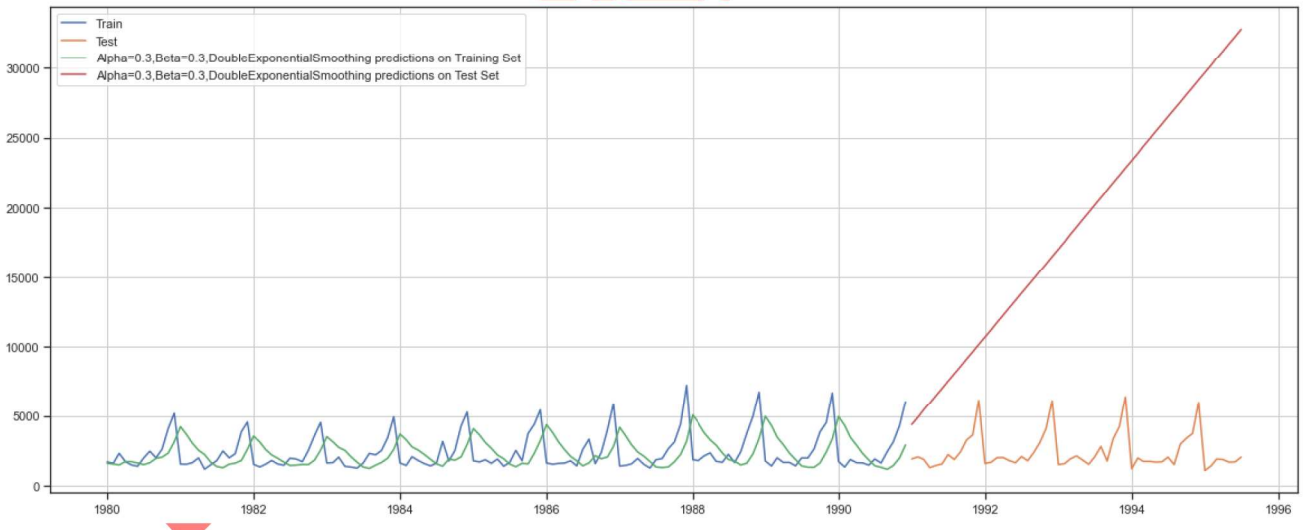
64 rows × 4 columns

Let us sort the data frame in the ascending ordering of the 'Test RMSE'

	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

Model 5 Plot – Double Exponential Smoothing

Plotting on both the Training and Test data



Model 5 Evaluation – Double Exponential Smoothing

	Test RMSE
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704

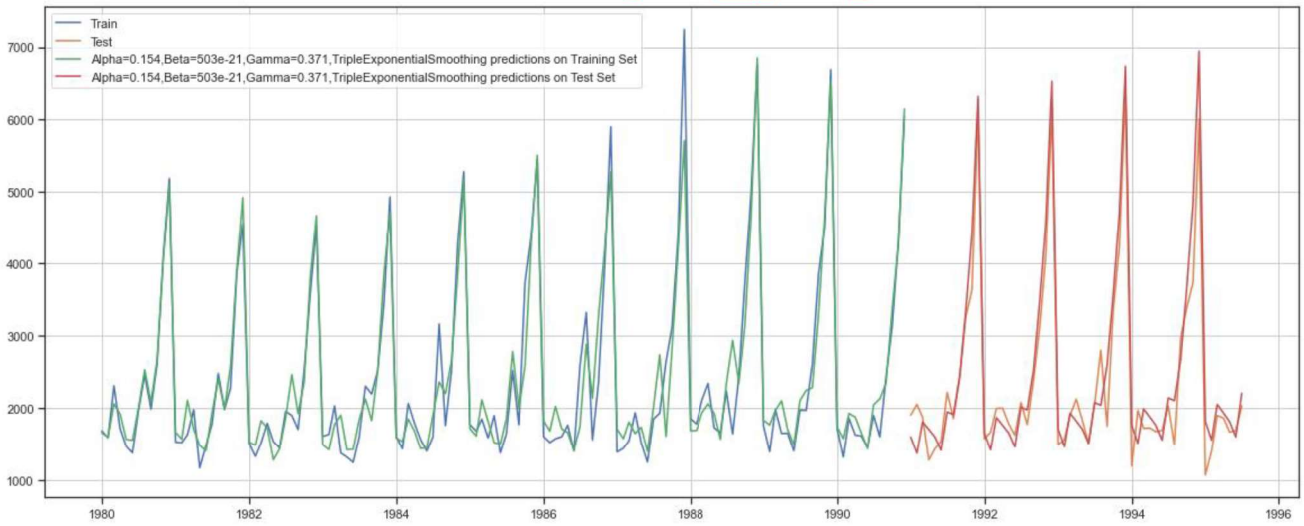
1.4.6. Model 6: Triple Exponential Smoothing (Holt - Winter's Model)

Three parameters alpha, beta and gamma are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

Model 6 - Auto-Fit Parameters

```
{'smoothing_level': 0.1541983771982372,
 'smoothing_slope': 5.189414295988849e-21,
 'smoothing_seasonal': 0.3713299858378782,
 'damping_slope': nan,
 'initial_level': 1639.9993278777176,
 'initial_slope': 4.84642057386791,
 'initial_seasons': array([1.0084339 , 0.96899852, 1.24181877, 1.13206906, 0.93979327,
        0.93811531, 1.22458354, 1.54428601, 1.27335854, 1.63198403,
        2.48293255, 3.11863164]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Model 6 Plot – Triple Exponential Smoothing



Model 6 Evaluation – Triple Exponential Smoothing

For Alpha=0.154,Beta=503e-21,Gamma=0.371, Triple Exponential Smoothing Model forecast on the Training Data, RMSE is 353.379

For Alpha=0.154,Beta=503e-21,Gamma=0.371, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 383.122

1.4.7. Model 7: Brute Force - Triple Exponential Smoothing

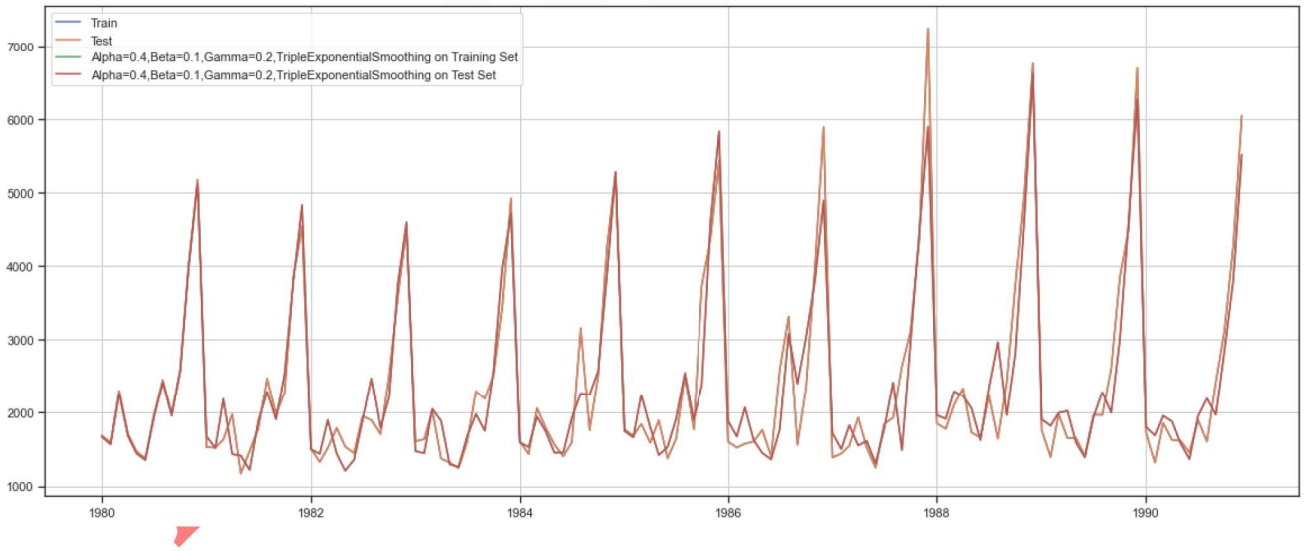
	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
0	0.0	0.0	0.0	2117.794609	3.934417e+03
1	0.0	0.0	0.1	5185.426811	2.722161e+04
2	0.0	0.0	0.2	8214.592317	4.554097e+04
3	0.0	0.0	0.3	10670.142261	5.805787e+04
4	0.0	0.0	0.4	12566.129842	6.495477e+04
...	...	...	...	...	...
1326	1.0	1.0	0.6	153394.791827	7.989790e+05
1327	1.0	1.0	0.7	94040.964957	1.074413e+06
1328	1.0	1.0	0.8	102196.953755	5.010607e+06
1329	1.0	1.0	0.9	77924.294413	4.318265e+05
1330	1.0	1.0	1.0	239917.432848	1.254280e+05

1331 rows × 5 columns

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
497	0.4	0.1	0.2	389.772245	336.715250
387	0.3	0.2	0.2	395.529174	350.145204
265	0.2	0.2	0.1	405.333164	352.571689
375	0.3	0.1	0.1	394.630053	352.607849
155	0.1	0.3	0.1	414.423963	354.534561

With Brute force a better RMSE value is found.

Model 7 Plot – Brute Force for Triple Exponential Smoothing

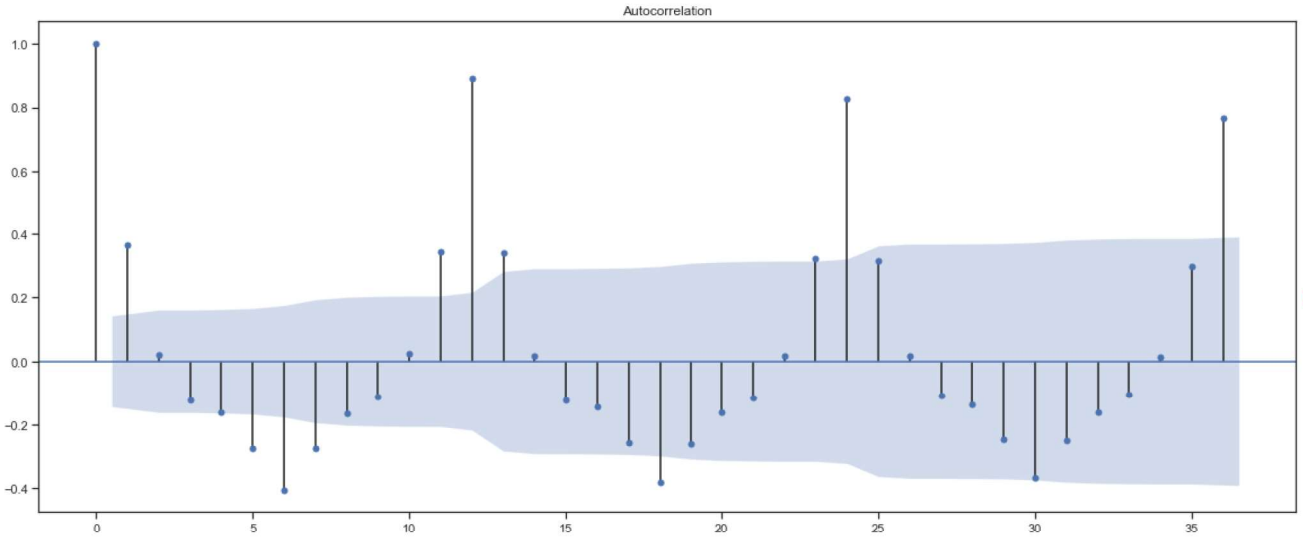


Model 7 Evaluation – Brute Force for Triple Exponential Smoothing

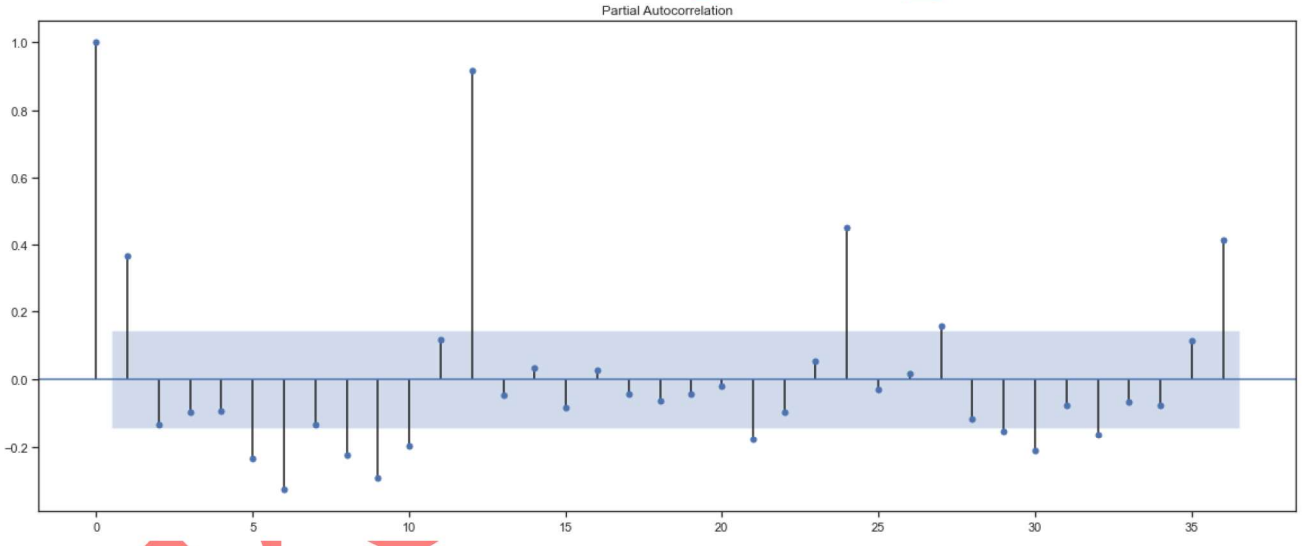
	Test RMSE
Alpha=0.4,Beta=0.1,Gamma=0.2,BruteForce TripleExponentialSmoothing	336.715250

1.4.8. ACF/ PCF Plots

ACF Plot



PACF Plot

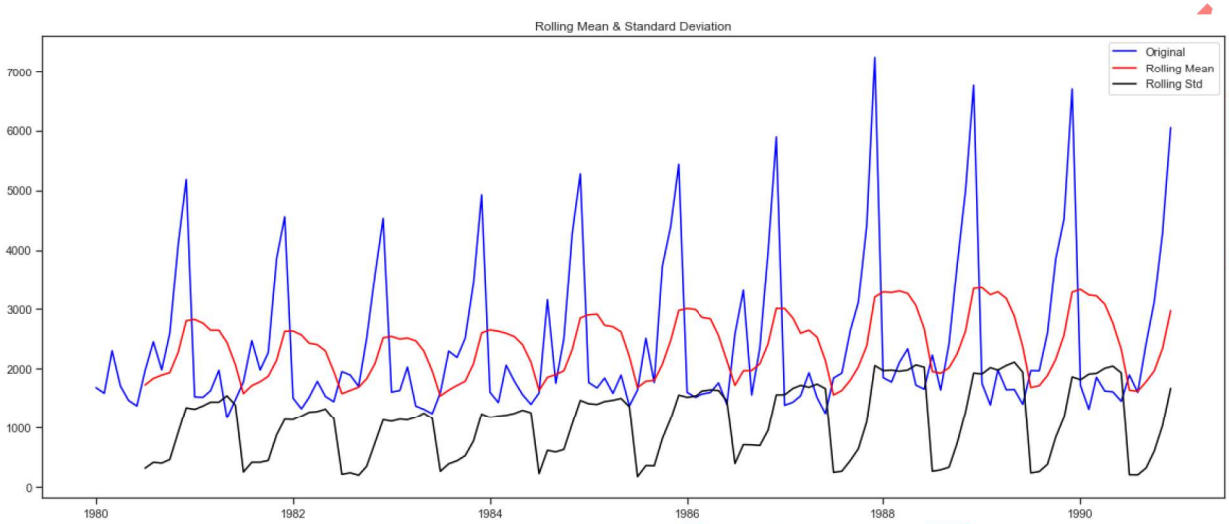


Seasonality after certain lags is visible. Every 12th Month. ACF and PACF plots are done with 95% confidence interval bands.

Test for stationarity of the series - Dicky Fuller test

Null and Alternate Hypothesis for the Augmented Dickey Fuller Test.

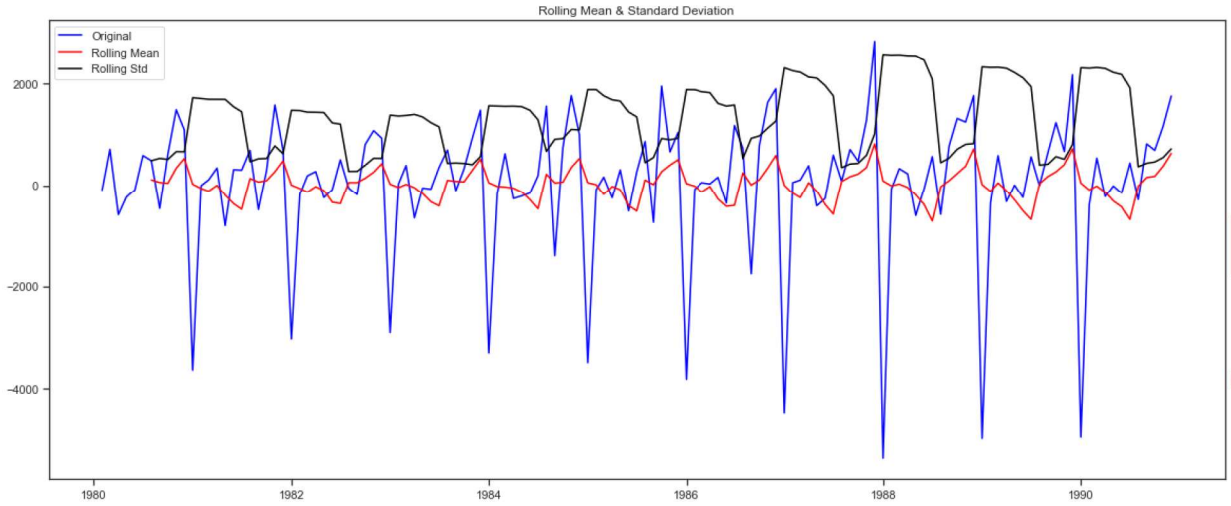
H0: The series is not stationary.
H1: The series is stationary.



Series is not stationary with original form at alpha = 0.05.

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.208926
p-value                          0.669744
#Lags Used                      12.000000
Number of Observations Used    119.000000
Critical Value (1%)             -3.486535
Critical Value (5%)             -2.886151
Critical Value (10%)            -2.579896
```

Let us try check for stationarity after taking first order differencing.



Series is stationary post first order differencing at alpha = 0.05

```
Results of Dickey-Fuller Test:
Test Statistic                -8.005007e+00
p-value                        2.280104e-12
#Lags Used                     1.100000e+01
Number of Observations Used    1.190000e+02
Critical Value (1%)           -3.486535e+00
Critical Value (5%)           -2.886151e+00
Critical Valuc (10%)          -2.579896c+00
```

1.4.9. Model 8: ARIMA Model by picking the pdq values from the ACF/ PACF plot

- p = 3
- d = 1
- q = 2

Summary of ARIMA (3, 1, 2) Model

ARIMA Model Results

Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(3, 1, 2)	Log Likelihood	-1107.464
Method:	css-mle	S.D. of innovations	1106.125
Date:	Mon, 20 Jul 2020	AIC	2228.927
Time:	12:51:19	BIC	2249.054
Sample:	02-01-1980 - 12-01-1990	HQIC	2237.106

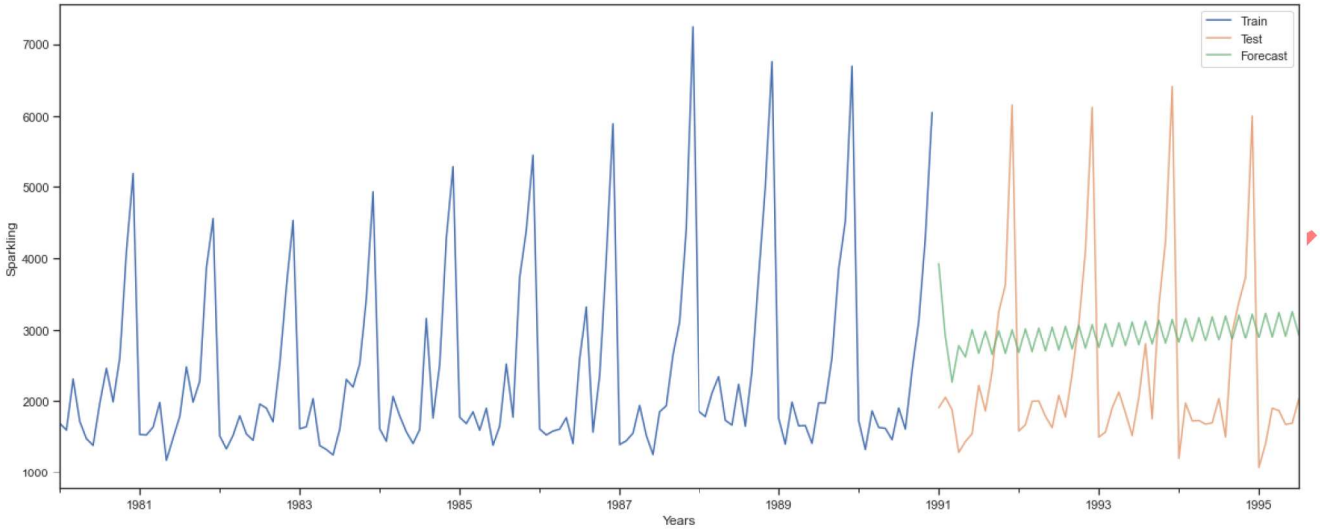
	coef	std err	z	P>|z|	[0.025	0.975]
const	5.9843	3.643	1.643	0.100	-1.156	13.125
ar.L1.D.Sparkling	-0.4420	5.95e-06	-7.43e+04	0.000	-0.442	-0.442
ar.L2.D.Sparkling	0.3079	1.57e-05	1.96e+04	0.000	0.308	0.308
ar.L3.D.Sparkling	-0.2501	1.36e-05	-1.83e+04	0.000	-0.250	-0.250
ma.L1.D.Sparkling	-0.0006	0.020	-0.028	0.977	-0.040	0.039
ma.L2.D.Sparkling	-0.9994	0.020	-49.385	0.000	-1.039	-0.960

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-1.0000	-0.0000j	1.0000	-0.5000
AR.2	1.1156	-1.6594j	1.9996	-0.1558
AR.3	1.1156	+1.6594j	1.9996	0.1558
MA.1	1.0000	+0.0000j	1.0000	0.0000
MA.2	-1.0006	+0.0000j	1.0006	0.5000

For this particular Auto-Regressive Integrated Moving Average we are regressing the original series on itself at the lags of 1,2 and 3. We are also considering the errors from the auto-regression of the first two lags. The values of p and q are calculated by looking at the ACF and the PACF plots.

Model 8 Plot – ARIMA Model by picking the pdq values from the ACF/ PACF plot



Model 8 Evaluation – ARIMA Model by picking the pdq values from the ACF/ PACF plot

	Test RMSE
ARIMA(3, 1, 2) Looking at ACF/PACF	1378.973814

1.4.10. Model 9: SARIMA Model by picking the pdq and PDQ values from the ACF/ PACF plot

- p = 3
- d = 1
- q = 2
- P = 1
- D = 0
- Q = 1

Summary of SARIMA(3, 1, 2)(1, 0, 1)12

SARIMAX Results

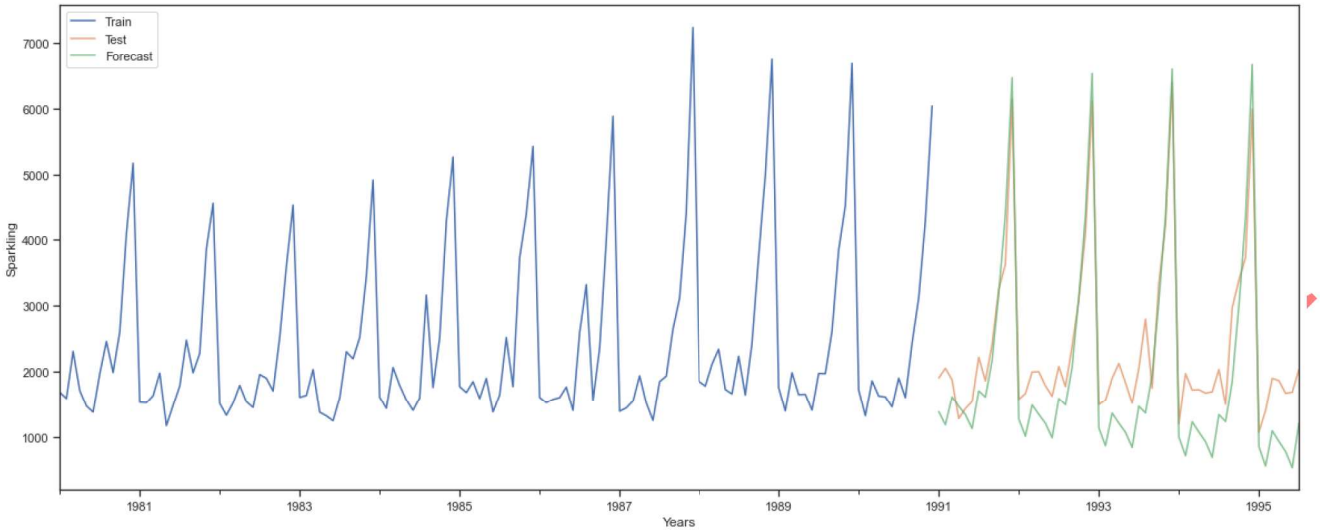
Dep. Variable:	Sparkling	No. Observations:	132
Model:	SARIMAX(3, 1, 2)x(1, 0, [1], 12)	Log Likelihood	-856.815
Date:	Mon, 20 Jul 2020	AIC	1729.629
Time:	12:51:21	BIC	1751.658
Sample:	01-01-1980	HQIC	1738.572
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2895	1.800	0.161	0.872	-3.239	3.818
ar.L2	-0.1155	0.190	-0.607	0.544	-0.489	0.258
ar.L3	0.0195	0.210	0.093	0.926	-0.392	0.431
ma.L1	-1.0734	1.806	-0.594	0.552	-4.613	2.466
ma.L2	0.1812	1.562	0.116	0.908	-2.881	3.243
ar.S.L12	1.0384	0.013	78.634	0.000	1.012	1.064
ma.S.L12	-0.6018	0.091	-6.620	0.000	-0.780	-0.424
sigma2	1.47e+05	1.56e+04	9.423	0.000	1.16e+05	1.78e+05

Ljung Box (Q):	29.29	Jarque Bera (JB):	38.45
Prob(Q):	0.89	Prob(JB):	0.00
Heteroskedasticity (H):	2.83	Skew:	0.53
Prob(H) (two-sided):	0.00	Kurtosis:	5.61

For this particular Seasonal Auto-Regressive Integrated Moving Average we are regressing the original series on itself at the lags of 1,2 and 3. We are also considering the errors from the auto-regression of the first two lags. For the seasonal parameters, we are considering the regression of the series on itself one with a lag of one year (or 12 months) and consi dering the errors from that particular auto-regression. The values of p, q, P and Q are calculated by looking at the ACF and the PACF plots.

Model 9 Plot – SARIMA Model by picking the pdq and PDQ values from the ACF/ PACF plot



Model 9 Evaluation – SARIMA Model by picking the pdq and PDQ values from the ACF/ PACF plot

	Test RMSE
SARIMA(3, 1, 2)(1, 0, 1, 12) Looking at ACF/PACF	633.013284

1.4.11. Model 10 Auto ARIMA

Series is not stationary and hence differentiation would be required. For an Auto-ARIMA, we calculate the best p and q parameters by looking at the lowest corresponding Akaike Information Criterion (AIC) values.

	param	AIC
8	(2, 1, 2)	2210.621009
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)	2264.906437
3	(1, 1, 0)	2268.528061
0	(0, 1, 0)	2269.582796

ARIMA(2,1,2) has the lowest AIC

Summary of ARIMA(2, 1, 2)

ARIMA Model Results

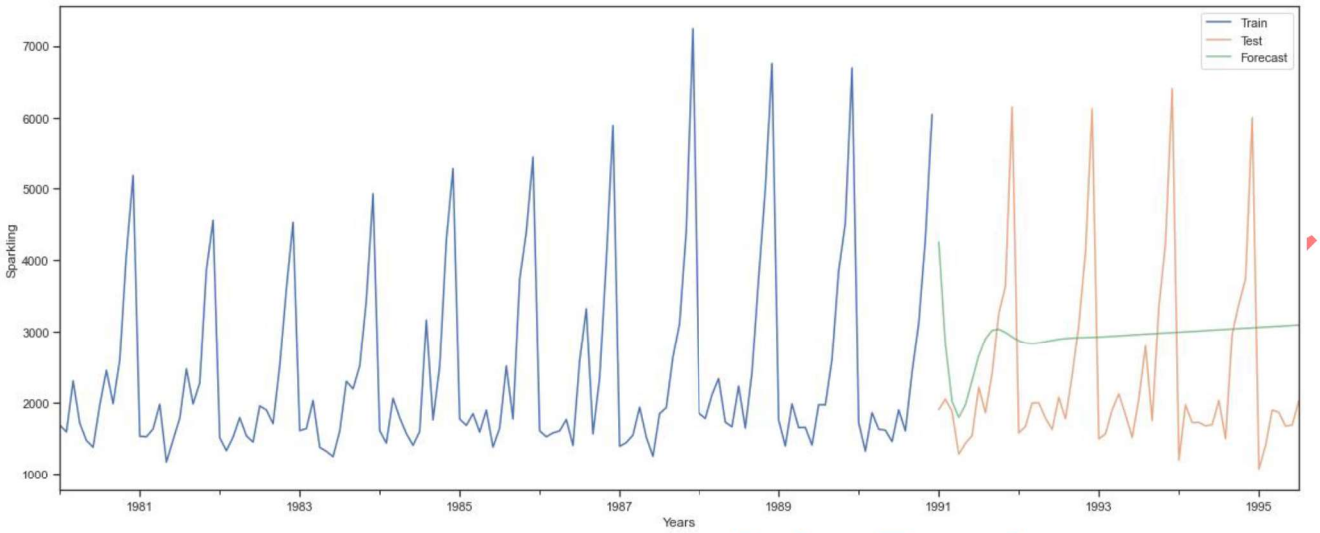
Dep. Variable:	D.Sparkling	No. Observations:	131
Model:	ARIMA(2, 1, 2)	Log Likelihood	-1099.311
Method:	css-mle	S.D. of innovations	1013.139
Date:	Mon, 20 Jul 2020	AIC	2210.621
Time:	12:51:25	BIC	2227.872
Sample:	02-01-1980 - 12-01-1990	HQIC	2217.631

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.5854	0.517	10.794	0.000	4.571	6.600
ar.L1.D.Sparkling	1.2699	0.075	17.044	0.000	1.124	1.416
ar.L2.D.Sparkling	-0.5602	0.074	-7.617	0.000	-0.704	-0.416
ma.L1.D.Sparkling	-1.9969	0.042	-47.066	0.000	-2.080	-1.914
ma.L2.D.Sparkling	0.9969	0.042	23.458	0.000	0.914	1.080

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.1335	-0.7074j	1.3361	-0.0888
AR.2	1.1335	+0.7074j	1.3361	0.0888
MA.1	1.0001	+0.0000j	1.0001	0.0000
MA.2	1.0029	+0.0000j	1.0029	0.0000

Model 10 Plot – Auto ARIMA



Model 10 Evaluation – Auto ARIMA

	Test RMSE
ARIMA(2, 1, 2) Best AIC	1374.381356

1.4.12. Model 11: Auto SARIMA

As the dataset has seasonality.. Let's build the model with SARIMA. For an Auto-SARIMA, the parameters p, q, P and Q are selected based on the lowest Akaike Information Criterion (AIC).

Top 5 best AIC values for Auto SARIMA

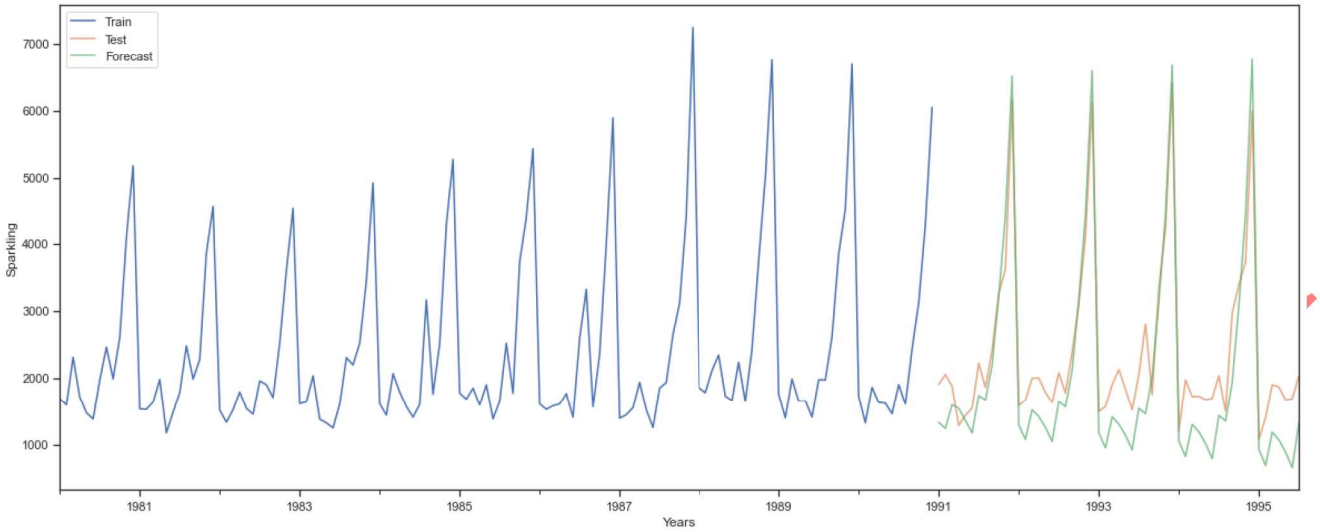
	param	seasonal	AIC
50	(1, 1, 2)	(1, 0, 2, 12)	1555.646564
23	(0, 1, 2)	(1, 0, 2, 12)	1557.160319
26	(0, 1, 2)	(2, 0, 2, 12)	1557.559840
53	(1, 1, 2)	(2, 0, 2, 12)	1562.131238
14	(0, 1, 1)	(1, 0, 2, 12)	1570.131967

Lowest AIC value is for SARIMA(1, 1, 2)(1, 0, 1, 12)

Summary of SARIMAX(1, 1, 2)(1, 0,1)12

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                            Sparkling   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 0, [1], 12)   Log Likelihood                -855.998
Date:                              Mon, 20 Jul 2020   AIC                           1723.995
Time:                                      12:52:43   BIC                           1740.517
Sample:                                  01-01-1980   HQIC                          1730.702
                                       - 12-01-1990
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6202      0.281     -2.211      0.027      -1.170      -0.070
ma.L1         -0.1259      0.258     -0.488      0.626      -0.632       0.380
ma.L2         -0.6740      0.203     -3.322      0.001      -1.072      -0.276
ar.S.L12       1.0379      0.013     81.380      0.000       1.013       1.063
ma.S.L12      -0.6168      0.084     -7.325      0.000      -0.782      -0.452
sigma2      1.452e+05   1.72e+04      8.468      0.000    1.12e+05    1.79e+05
==========================================================================================
Ljung-Box (Q):                       27.81   Jarque-Bera (JB):                 23.83
Prob(Q):                              0.93   Prob(JB):                          0.00
Heteroskedasticity (H):               2.70   Skew:                              0.44
Prob(H) (two-sided):                  0.00   Kurtosis:                          5.04
==========================================================================================
```

Model 11 Plot – Auto SARIMA



Model 11 Evaluation – Auto SARIMA

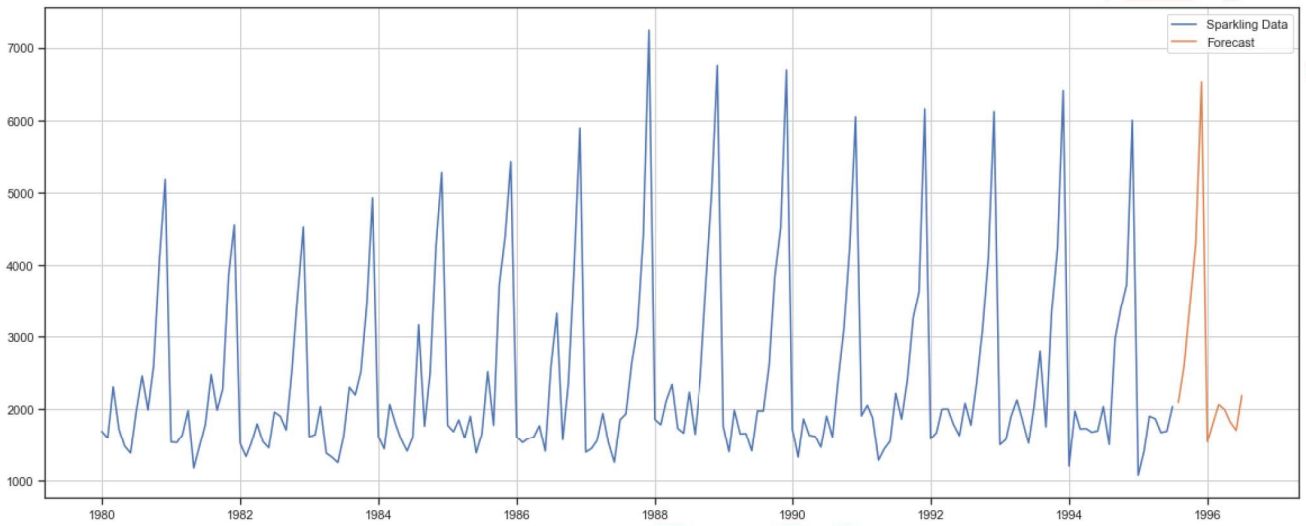
	Test RMSE
SARIMA(1, 1, 2)(1, 0, 1, 12) Best AIC	583.659809

2. Forecast Sparkling using the best fit BruteForce TripleExponentialSmoothing

Let's apply Brute Force Triple Exponential Smoothing on Full Data

The best found parameters for Triple Exponential Smoothing are Alpha=0.4, Beta=0.1, Gamma=0.2

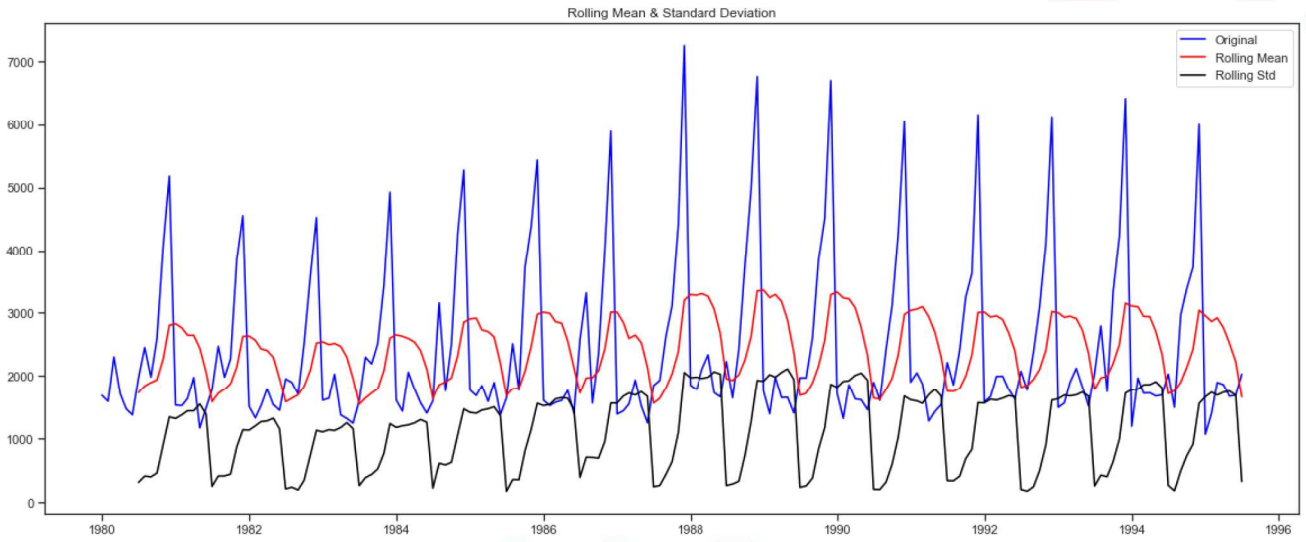
Plot for Best Next 12 month forecast



3. Forecast Sparkling using the best fit SARIMA model

Once we have found the parameters that best fit to the data, we are going to use these parameter to train of full Sparkling dataset and then forecast for the 12 months in future. To fit a SARIMA model, we need to check for stationarity on the whole data.

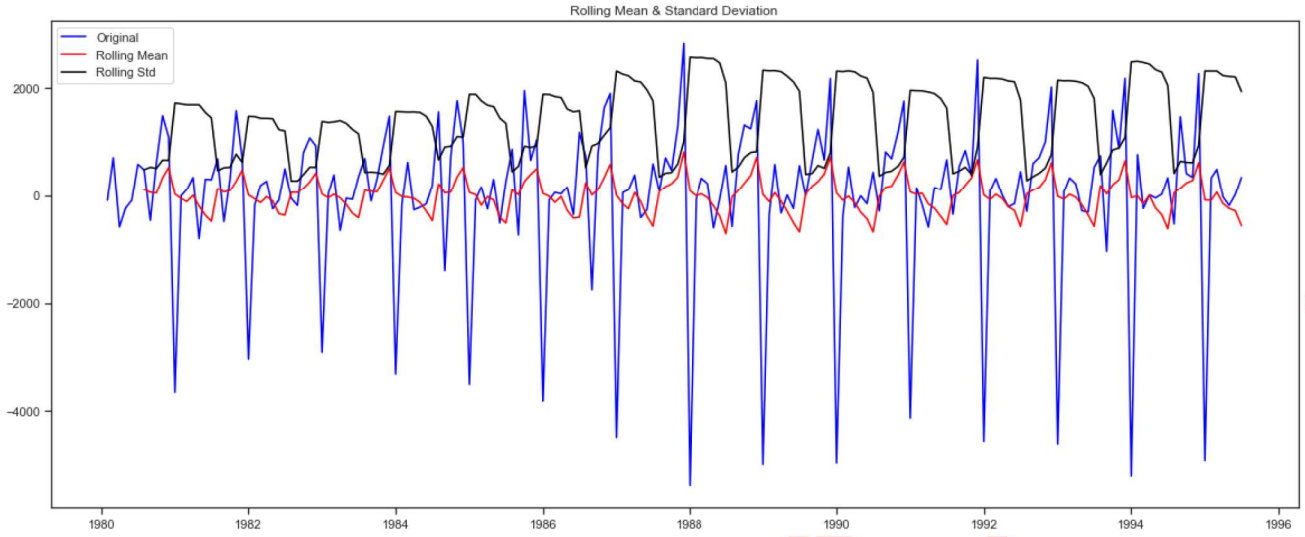
Stationarity test on the whole of original data



The whole series is non-stationary at alpha = 0.05 using the Augmented Dickey Fuller test.

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

Stationarity test post differencing of first order



The whole series is stationary at alpha = 0.05 using the Augmented Dickey Fuller test.

```
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
```

Summary of SARIMA(1, 1, 2)(1, 0, 1)12 on the full data

SARIMAX Results

Dep. Variable:	Sparkling	No. Observations:	187
Model:	SARIMAX(1, 1, 2)x(1, 0, [1], 12)	Log Likelihood	-1258.312
Date:	Mon, 20 Jul 2020	AIC	2528.624
Time:	12:52:46	BIC	2547.474
Sample:	01-01-1980 - 07-01-1995	HQIC	2536.272
Covariance Type:	opg		

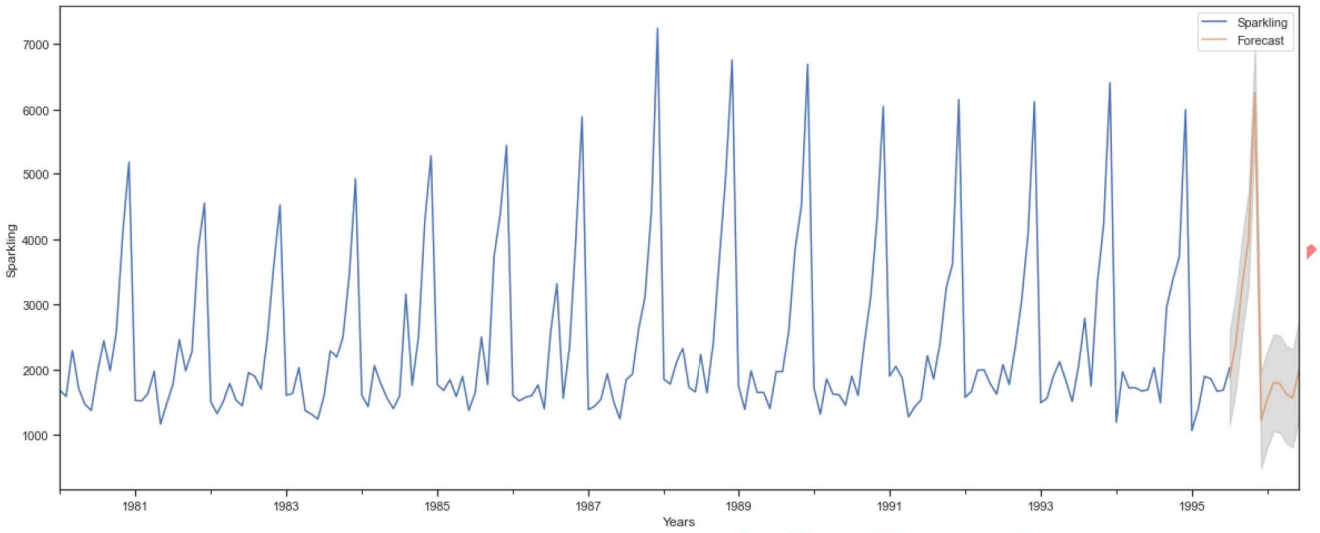
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.1514	0.719	-0.211	0.833	-1.561	1.258
ma.L1	-0.6747	0.716	-0.942	0.346	-2.078	0.729
ma.L2	-0.2537	0.664	-0.382	0.702	-1.555	1.047
ar.S.L12	1.0131	0.011	92.046	0.000	0.992	1.035
ma.S.L12	-0.6101	0.067	-9.148	0.000	-0.741	-0.479
sigma2	1.387e+05	1.14e+04	12.173	0.000	1.16e+05	1.61e+05

Ljung-Box (Q):	19.97	Jarque-Bera (JB):	54.77
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	1.25	Skew:	0.61
Prob(H) (two-sided):	0.41	Kurtosis:	5.49

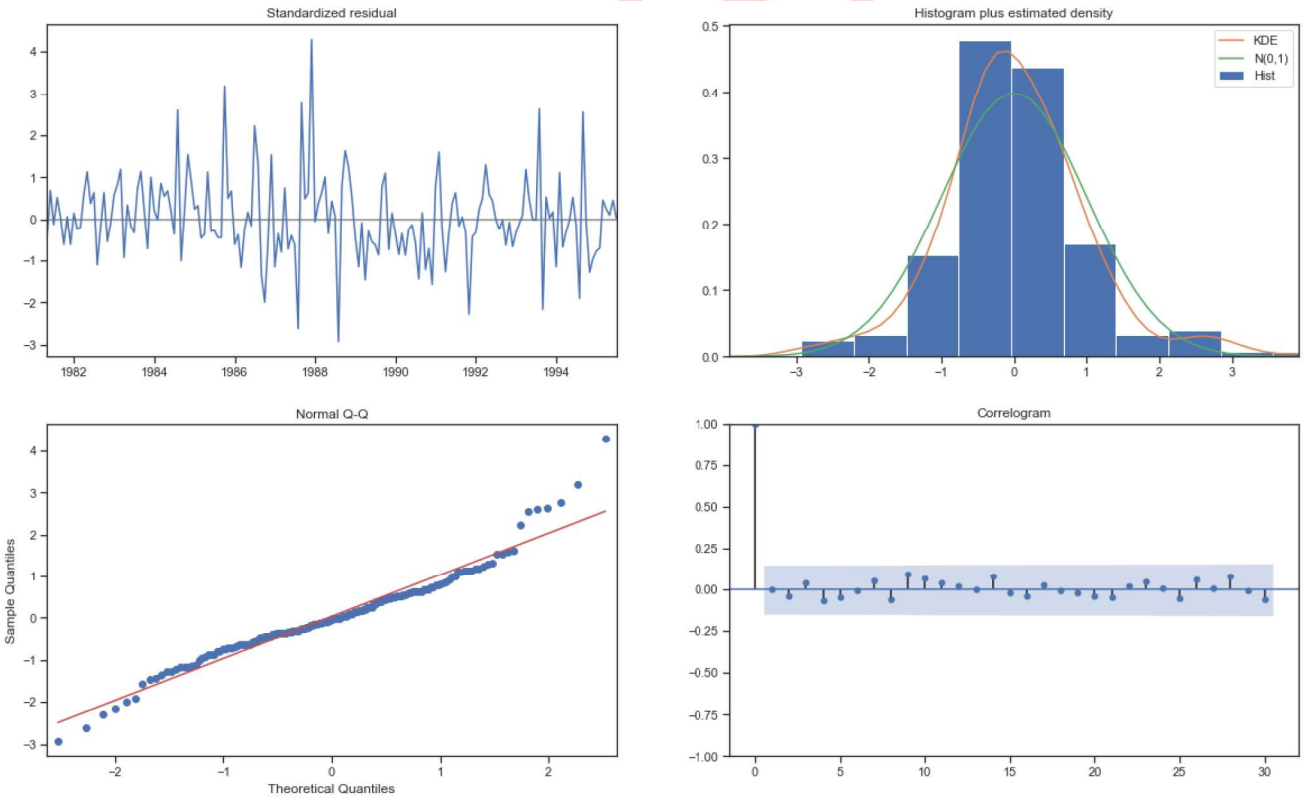
95 and 99 percent confidence interval table

	forecast	lower_ci_95	upper_ci_95	lower_ci_99	upper_ci_99
1995-07-01	1878.132895	1148.268721	2607.997069	1148.268721	2607.997069
1995-08-01	2427.589431	1686.776777	3168.402084	1686.776777	3168.402084
1995-09-01	3304.116929	2562.568111	4045.665747	2562.568111	4045.665747
1995-10-01	4003.095993	3260.043554	4746.148431	3260.043554	4746.148431
1995-11-01	6262.087766	5517.668214	7006.507319	5517.668214	7006.507319

Plot for Best Next 12 month forecast



SRIMAX diagnostic plot on full data



4. Inference

We need to ensure that the residuals of our model are uncorrelated and normally distributed with zero-mean. If it is not that it signifies that the model can be further improved and we repeat the process with the residuals.

In this case, our model diagnostics suggests that the model residuals are normally distributed based on the following:

1. The KDE plot of the residuals on the top right is almost similar with the normal distribution.
2. The qq-plot on the bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with N(0, 1). Again, this is a strong indication that the residuals are normally distributed.
3. The residuals over time (top left plot) don't display any obvious seasonality and appear to be white noise. This is confirmed by the autocorrelation (i.e. correlogram) plot on the bottom right, which shows that the time series residuals have low correlation with lagged versions of itself.

Those observations coupled with the fact that there are no spikes outside the insignificant zone for both ACF and PACF plots lead us to conclude that that residuals are random with no information or juice in them and our model produces a satisfactory fit that could help us understand our time series data and forecast future values. It sems that our ARIMA model is working fine.

5. Appendix - Summary of Testing RMSE

	Test RMSE
RegressionOnTime	1384.558065
NaiveModel	3864.279352
SimpleAverageModel	1275.081804
2pointTrailingMovingAverage	813.400684
4pointTrailingMovingAverage	1156.589694
6pointTrailingMovingAverage	1283.927428
9pointTrailingMovingAverage	1346.278315
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	18259.110704
Alpha=0.154,Beta=503e-21,Gamma=0.371,TripleExponentialSmoothing	383.122273
Alpha=0.4,Beta=0.1,Gamma=0.2,BruteForce TripleExponentialSmoothing	336.715250
ARIMA(3, 1, 2) Looking at ACF/PACF	1378.973814
SARIMA(3, 1, 2)(1, 0, 1, 12) Looking at ACF/PACF	633.013284
ARIMA(2, 1, 2) Best AIC	1374.381356
SARIMA(1, 1, 2)(1, 0, 1, 12) Best AIC	583.659809