2023

TIME SERIES FORECASTING GRADED PROJECT REPORT

DSBA

Girish Chadha
4/04/2023

Contents

List of Figures

List of Tables

EXECUTIVE SUMMARY

Problem 1 for the Data Set : Shoesales.csv

You are an analyst in the IJK shoe company and you are expected to forecast the sales of the pairs of shoes for the upcoming 12 months from where the data ends. The data for the pair of shoe sales have been given to you from January 1980 to July 1995.

Problem 2 for the Data Set : SoftDrink.csv

You are an analyst in the RST soft drink company and you are expected to forecast the sales of the production of the soft drink for the upcoming 12 months from where the data ends. The data for the production of soft drink has been given to you from January 1980 to July 1995.

Please do perform the following questions on each of these two data sets separately.

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

Solution:

Shoesales Data

Read the Shoesales data from the '.csv' file as a monthly Time Series

YearMonth	Shoe_Sales
1980-01-01	85
1980-02-01	89
1980-03-01	109
1980-04-01	95
1980-05-01	91

Table 1

Checking the number of observations
187 Rows & 2 columns

Plot the Time Series to understand the behaviour of the data

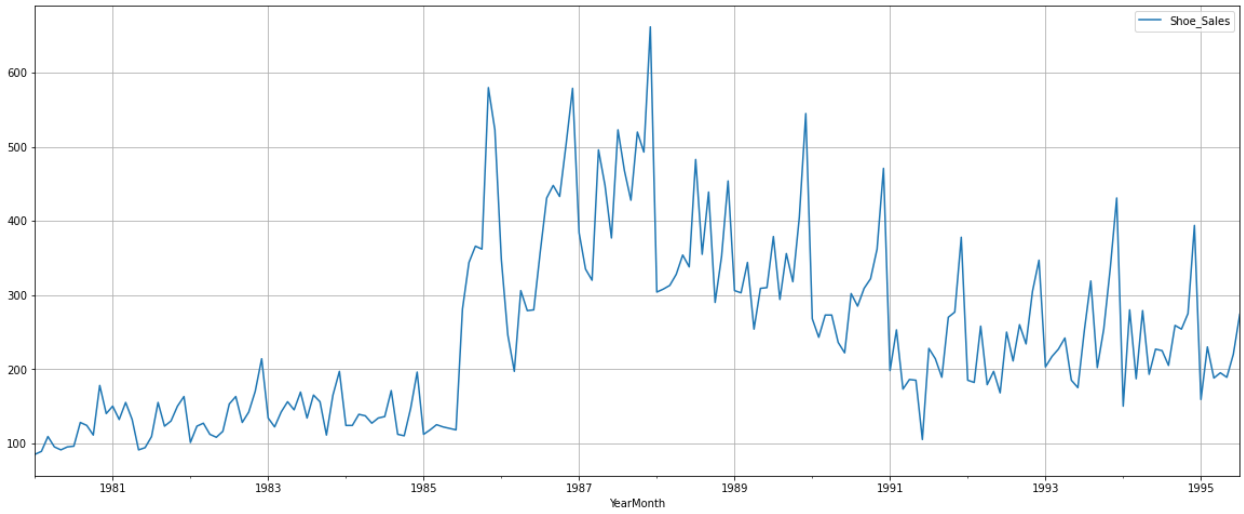


Fig 1

The plot depicts an increasing trend of sales over the period of 1980 to 1995.

Soft Drink Data

Read the Soft Drink data from the '.csv' file as a monthly Time Series.

YearMonth	SoftDrinkProduction
1980-01-01	1954
1980-02-01	2302
1980-03-01	3054
1980-04-01	2414
1980-05-01	2226

Table 2

Checking the number of observations
187 Rows & 2 columns

Plot the Time Series to understand the behaviour of the data.

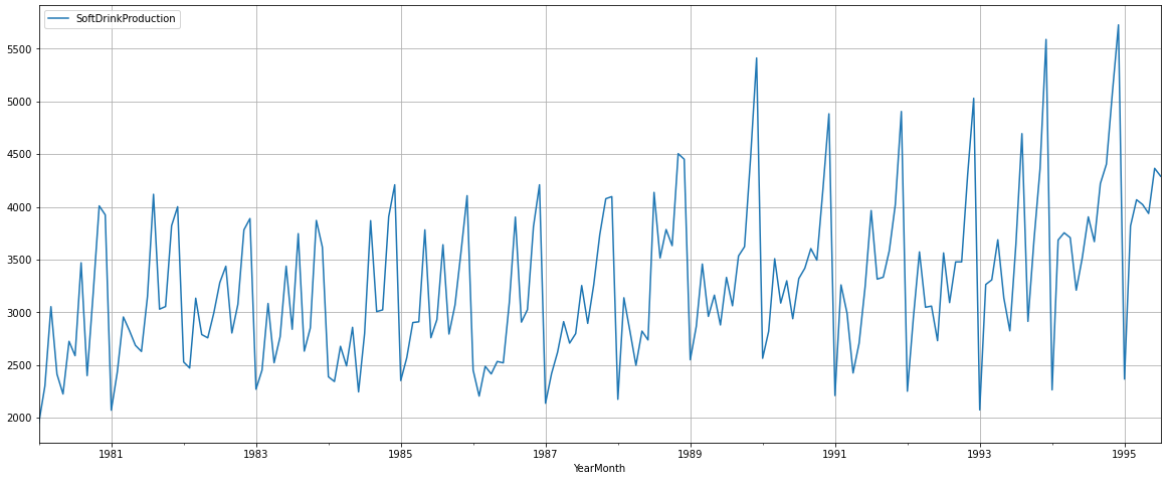


Fig 2

The plot depicts an increasing trend of sales over the period of 1980 to 1995.

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

<u>Shoesales data</u>

Check the basic measures of descriptive statistics

	Shoe_Sales
count	187.000000
mean	245.636364
std	121.390804
min	85.000000
25%	143.500000
50%	220.000000
75%	315.500000
max	662.000000

Table 3

Checking the null values

```
Shoe_Sales    0
dtype: int64
```

Total 187 records and min sales of 85 and max sales of 662 with 50 percentile is 220.

0 null values.

Plot a boxplot to understand the spread of sales across different years and within different months across years.

Yearly Boxplot

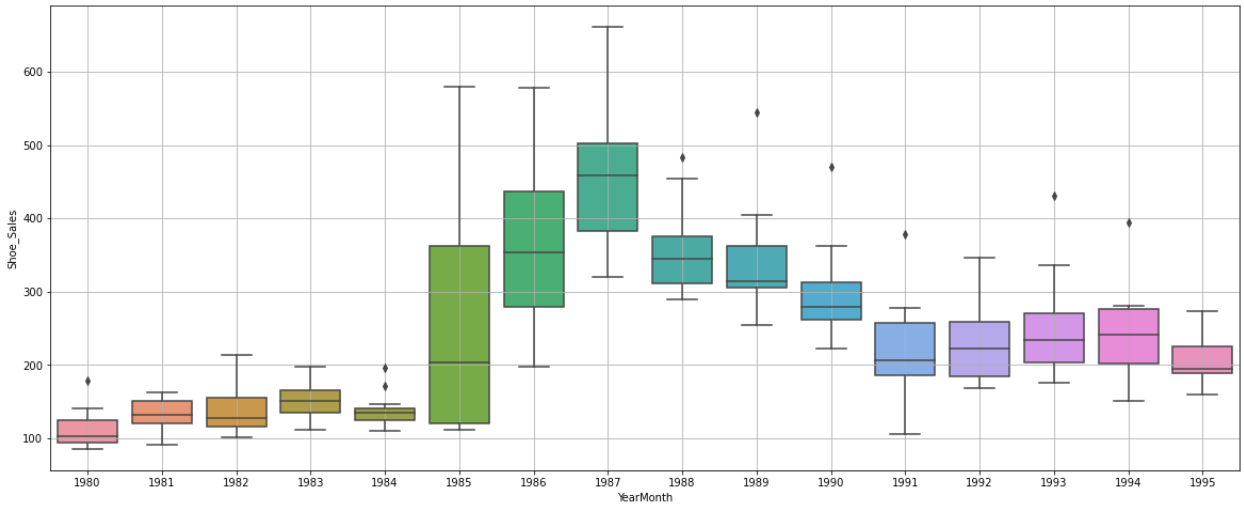


Figure 3

The above picture shows the trend over the period between 1980 to 1995 with outliers in most of the years. Highest sales in 1987 and lowest sales in 1980.

Monthly Plot

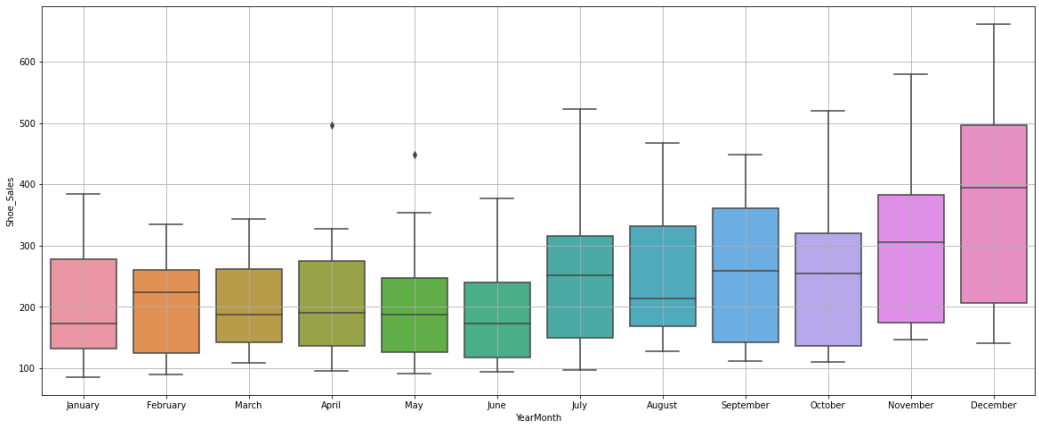


Fig 4

The above picture shows the trend of sales in months , December seems to have the highest sales and January seems to have lowest sales.

Plot a time series monthplot to understand the spread of Sales across different years and within different months across years

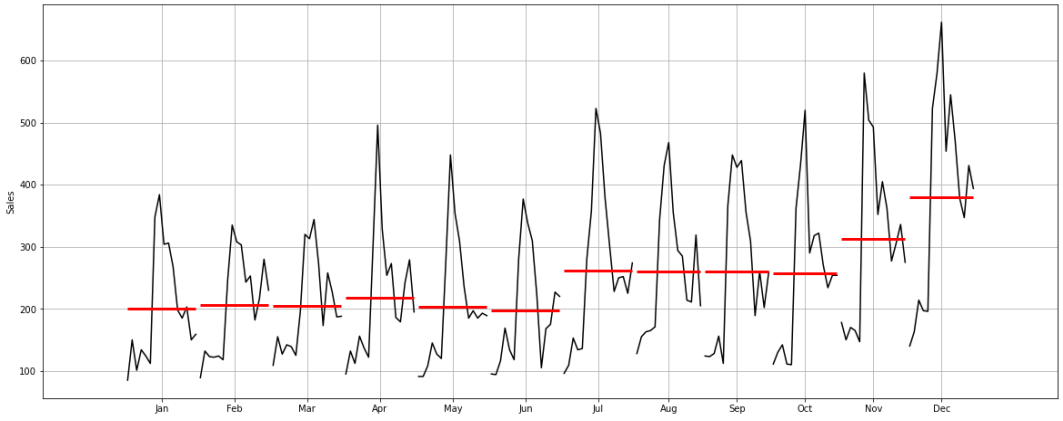


Fig 5

This plot shows us the behaviour of the Time Series ('Shoe_Sales' in this case) across various months. The red line is the median value.

Plot a graph of monthly RetailSales across years

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	85.000000	89.000000	109.000000	95.000000	91.000000	95.000000	96.000000	128.000000	124.000000	111.000000	178.000000	140.000000
1981	150.000000	132.000000	155.000000	132.000000	91.000000	94.000000	109.000000	155.000000	123.000000	130.000000	150.000000	163.000000
1982	101.000000	123.000000	127.000000	112.000000	108.000000	116.000000	153.000000	163.000000	128.000000	142.000000	170.000000	214.000000
1983	134.000000	122.000000	142.000000	156.000000	145.000000	169.000000	134.000000	165.000000	156.000000	111.000000	165.000000	197.000000
1984	124.000000	124.000000	139.000000	137.000000	127.000000	134.000000	136.000000	171.000000	112.000000	110.000000	147.000000	196.000000
1985	112.000000	118.000000	125.000000	122.000000	120.000000	118.000000	281.000000	344.000000	366.000000	362.000000	580.000000	523.000000
1986	348.000000	246.000000	197.000000	306.000000	279.000000	280.000000	358.000000	431.000000	448.000000	433.000000	504.000000	579.000000
1987	384.000000	335.000000	320.000000	496.000000	448.000000	377.000000	523.000000	468.000000	428.000000	520.000000	493.000000	662.000000
1988	304.000000	308.000000	313.000000	328.000000	354.000000	338.000000	483.000000	355.000000	439.000000	290.000000	352.000000	454.000000
1989	306.000000	303.000000	344.000000	254.000000	309.000000	310.000000	379.000000	294.000000	356.000000	318.000000	405.000000	545.000000
1990	268.000000	243.000000	273.000000	273.000000	236.000000	222.000000	302.000000	285.000000	309.000000	322.000000	362.000000	471.000000
1991	198.000000	253.000000	173.000000	186.000000	185.000000	105.000000	228.000000	214.000000	189.000000	270.000000	277.000000	378.000000
1992	185.000000	182.000000	258.000000	179.000000	197.000000	168.000000	250.000000	211.000000	260.000000	234.000000	305.000000	347.000000
1993	203.000000	217.000000	227.000000	242.000000	185.000000	175.000000	252.000000	319.000000	202.000000	254.000000	336.000000	431.000000
1994	150.000000	280.000000	187.000000	279.000000	193.000000	227.000000	225.000000	205.000000	259.000000	254.000000	275.000000	394.000000
1995	159.000000	230.000000	188.000000	195.000000	189.000000	220.000000	274.000000	nan	nan	nan	nan	nan

1987 had the highest sales of all years in December.

1980 January had the least sales

Table 4

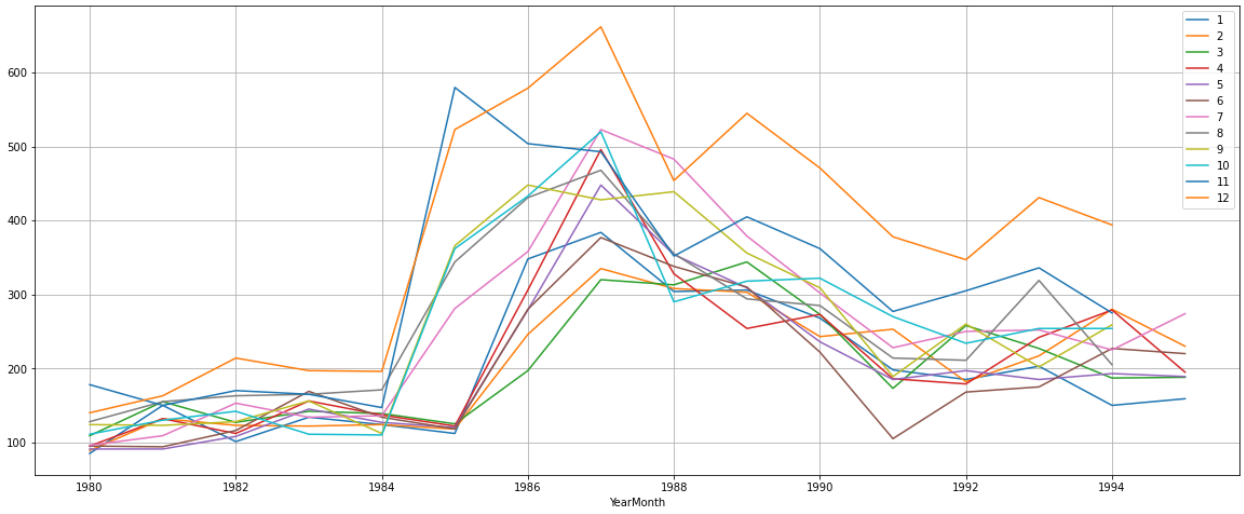
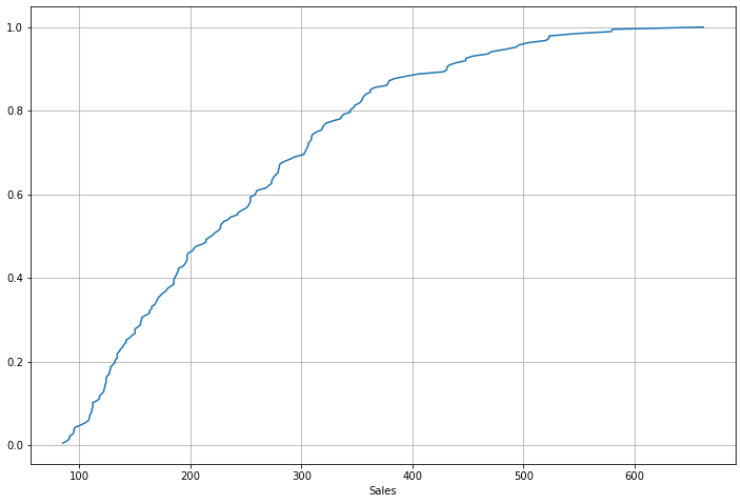


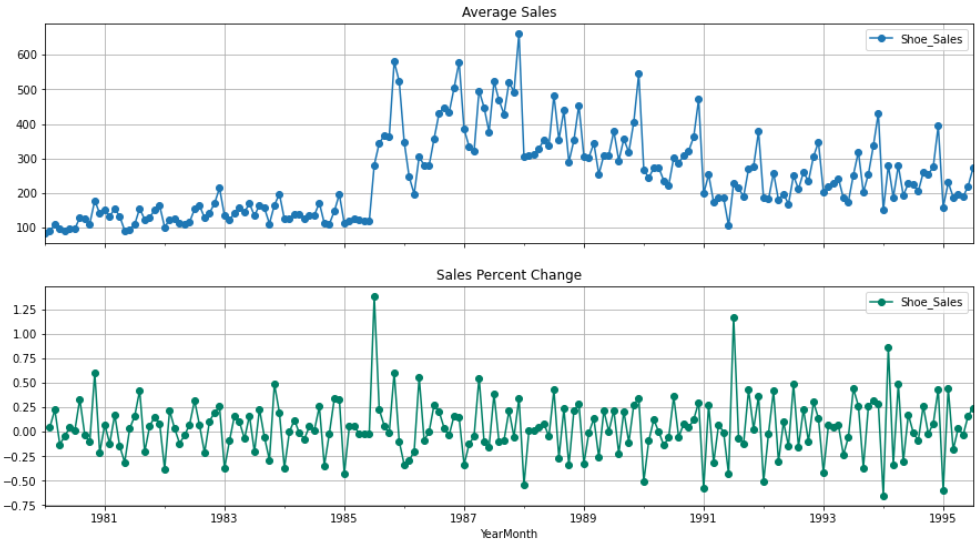
Fig 6

Plot the Empirical Cumulative Distribution



This particular graph tells us what percentage of data points refer to what number of Sales. Sales increases over the years.

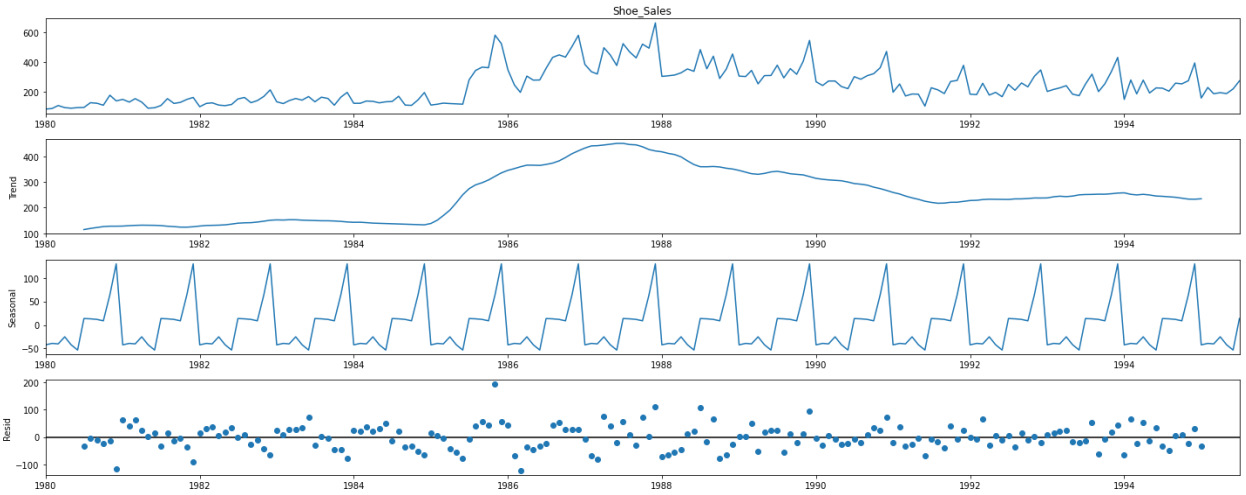
Plot the average Sales per month and the month on month percentage change of Sales



The above two graphs tells us the Average 'Sales' and the Percentage change of 'Sales' with respect to the time.

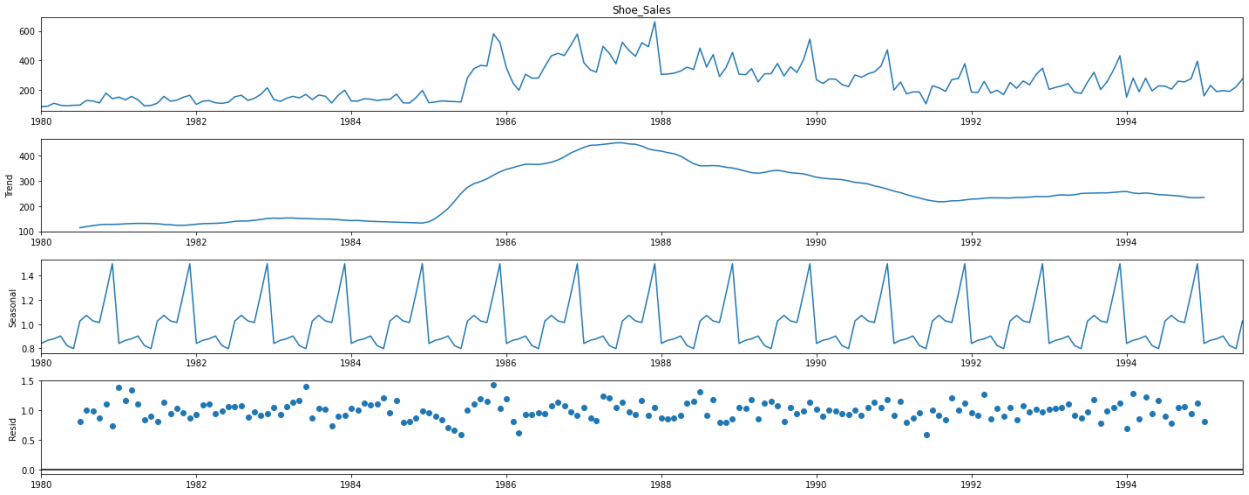
Decompose the Time Series and plot the different components

Additive Decomposition



As per the 'additive' decomposition, we see that there is an increasing decreasing trend in sales of Shoes starting from 1980 to 1994. Seasonality is also clearly visible from the seasonal graph where trend lines are forming the peaks each year. Residuals seems to be scattered from the 0 level. Indicating that the series is not additive.

Multiplicative Decomposition



The trend and seasonality are present same as in case of additive model. But residuals plot is clearly showing the concentration of data towards 1 point. Hence it can be concluded that series is multiplicative.

Soft drink Data

Check the basic measures of descriptive statistics

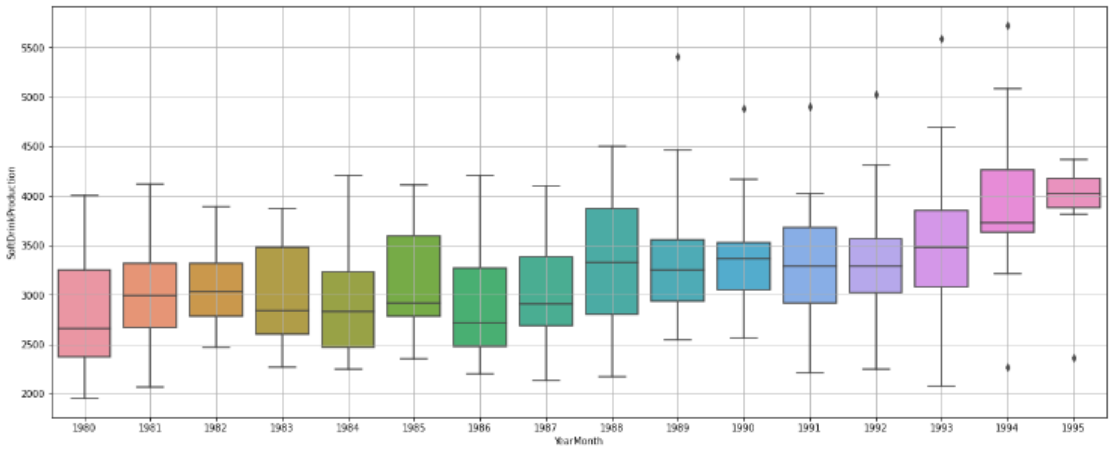
	SoftDrinkProduction
count	187.000000
mean	3262.609626
std	728.357367
min	1954.000000
25%	2748.000000
50%	3134.000000
75%	3741.000000
max	5725.000000

Total 187 records and min sales of 1954 and max sales of 5725 with 50 percentile is 3134.

0 missing/null values.

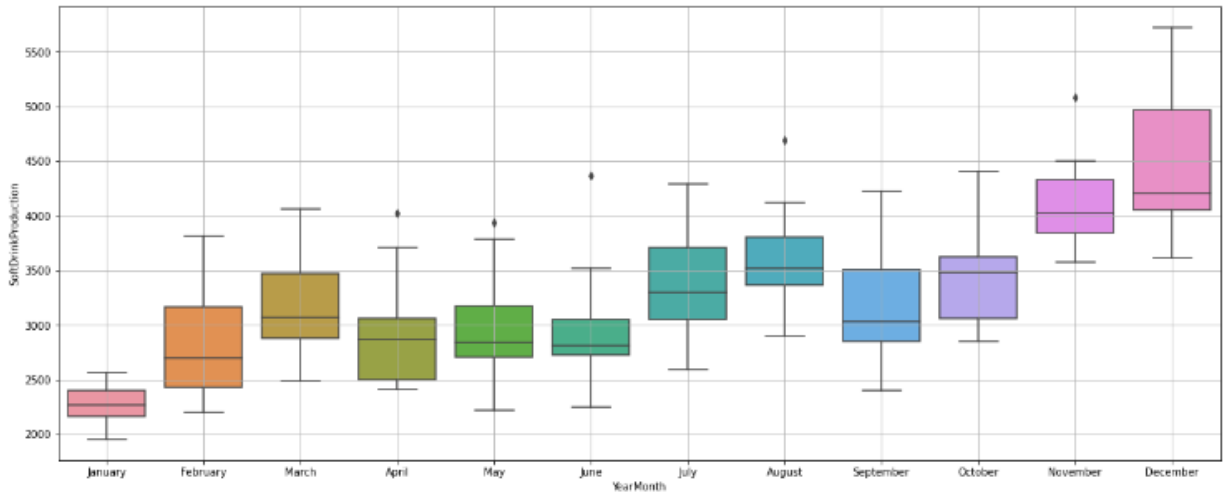
Plot a boxplot to understand the spread of sales across different years and within different months across years.

Yearly Boxplot



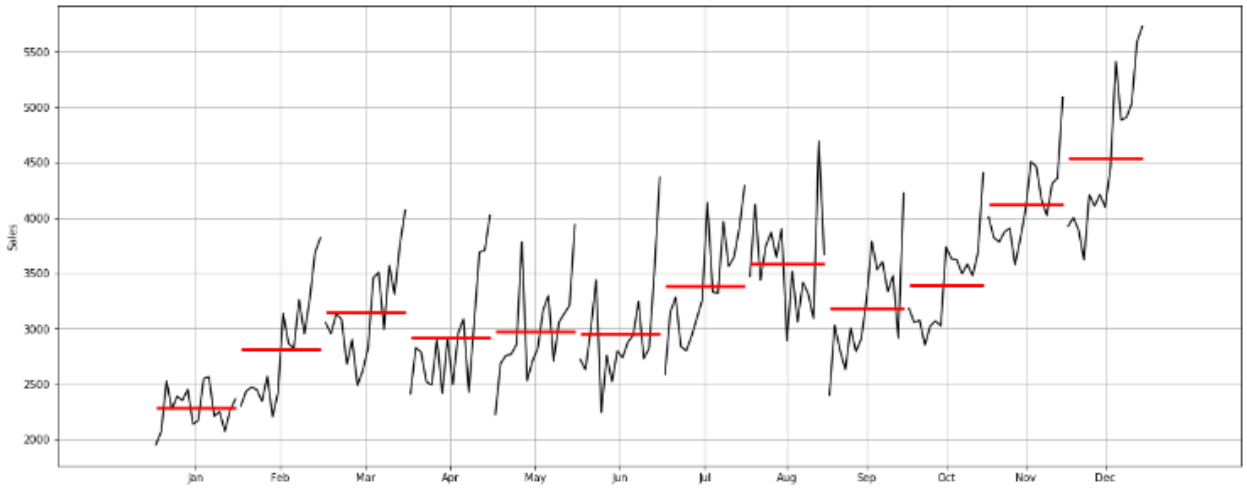
The above picture shows the trend over the period between 1980 to 1995 with outliers in 1989 to 1995. 1994 - highest sales, 1980 - lowest sales.

Monthly Plot



The above picture shows the trend of sales in months , December seems to have the highest sales and January seems to have lowest sales.

Plot a time series monthplot to understand the spread of Sales across different years and within different months across years.

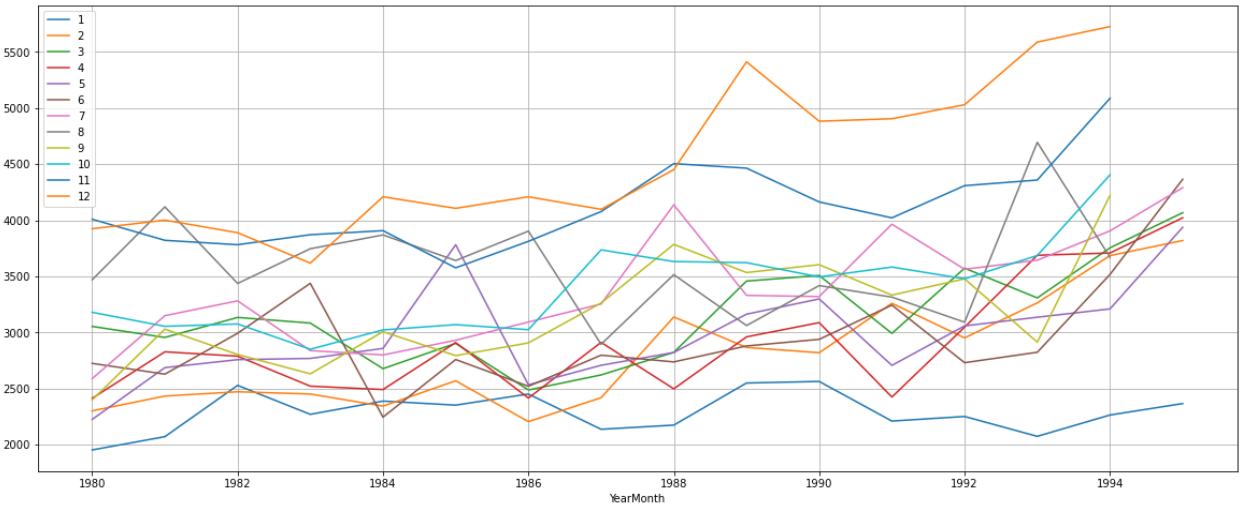


This plot shows us the behaviour of the Time Series ('SoftDrinkProduction' in this case) across various months. The red line is the median value.

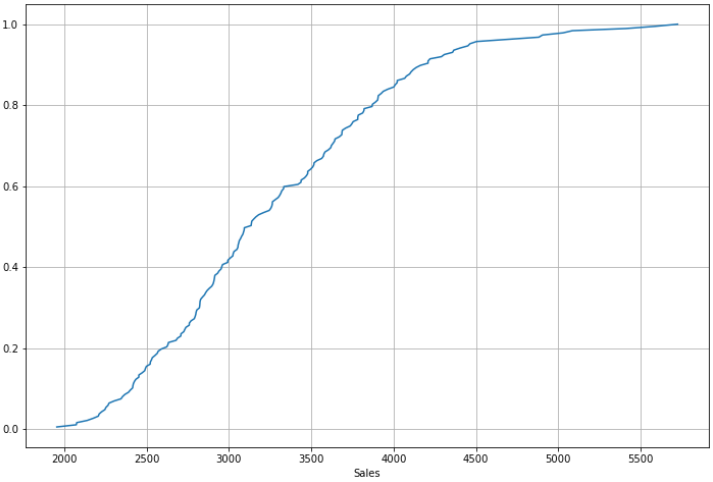
Plot a graph of monthly RetailSales across years

Month	1	2	3	4	5	6	7	8	9	10	11	
Month												
1980	1954.000000	2302.000000	3054.000000	2414.000000	2226.000000	2725.000000	2589.000000	3470.000000	2400.000000	3180.000000	4009.000000	3924.00
1981	2072.000000	2434.000000	2956.000000	2828.000000	2687.000000	2629.000000	3150.000000	4119.000000	3030.000000	3055.000000	3821.000000	4001.00
1982	2529.000000	2472.000000	3134.000000	2789.000000	2758.000000	2993.000000	3282.000000	3437.000000	2804.000000	3076.000000	3782.000000	3889.00
1983	2271.000000	2452.000000	3084.000000	2522.000000	2769.000000	3438.000000	2839.000000	3746.000000	2632.000000	2851.000000	3871.000000	3618.00
1984	2389.000000	2344.000000	2678.000000	2492.000000	2858.000000	2246.000000	2800.000000	3869.000000	3007.000000	3023.000000	3907.000000	4209.00
1985	2353.000000	2570.000000	2903.000000	2910.000000	3782.000000	2759.000000	2931.000000	3641.000000	2794.000000	3070.000000	3576.000000	4106.00
1986	2452.000000	2206.000000	2488.000000	2416.000000	2534.000000	2521.000000	3093.000000	3903.000000	2907.000000	3025.000000	3812.000000	4209.00
1987	2138.000000	2419.000000	2622.000000	2912.000000	2708.000000	2798.000000	3254.000000	2895.000000	3263.000000	3736.000000	4077.000000	4097.00
1988	2175.000000	3138.000000	2823.000000	2498.000000	2822.000000	2738.000000	4137.000000	3515.000000	3785.000000	3632.000000	4504.000000	4451.00
1989	2550.000000	2867.000000	3458.000000	2961.000000	3163.000000	2880.000000	3331.000000	3062.000000	3534.000000	3622.000000	4464.000000	5411.00
1990	2564.000000	2820.000000	3508.000000	3088.000000	3299.000000	2939.000000	3320.000000	3418.000000	3604.000000	3495.000000	4163.000000	4882.00
1991	2211.000000	3260.000000	2992.000000	2425.000000	2707.000000	3244.000000	3965.000000	3315.000000	3333.000000	3583.000000	4021.000000	4904.00
1992	2252.000000	2952.000000	3573.000000	3048.000000	3059.000000	2731.000000	3563.000000	3092.000000	3478.000000	3478.000000	4308.000000	5029.00
1993	2075.000000	3264.000000	3308.000000	3688.000000	3136.000000	2824.000000	3644.000000	4694.000000	2914.000000	3686.000000	4358.000000	5587.00
1994	2265.000000	3685.000000	3754.000000	3708.000000	3210.000000	3517.000000	3905.000000	3670.000000	4221.000000	4404.000000	5086.000000	5725.00
1995	2367.000000	3819.000000	4067.000000	4022.000000	3937.000000	4365.000000	4290.000000	nan	nan	nan	nan	

1994 had the highest sales of all years in December. 1980 January had the least sales.

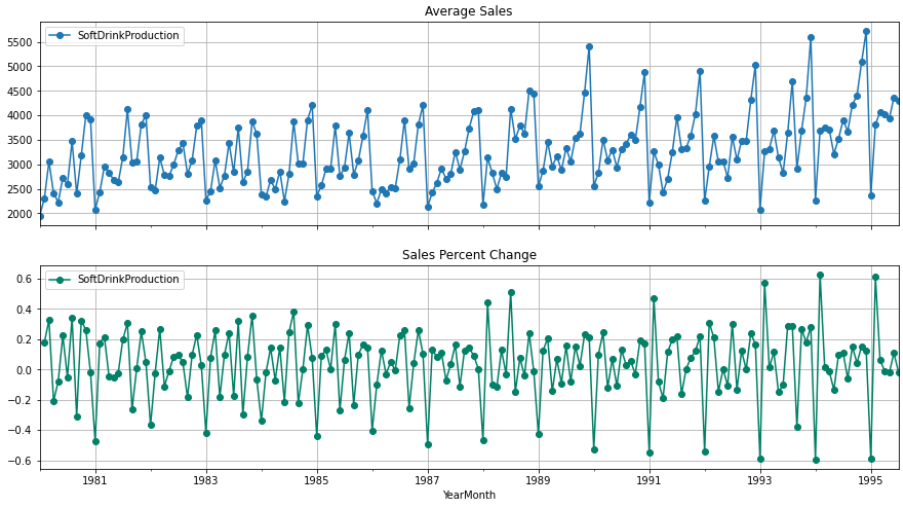


Plot the Empirical Cumulative Distribution



This particular graph tells us what percentage of data points refer to what number of Sales. Sales increases over the years.

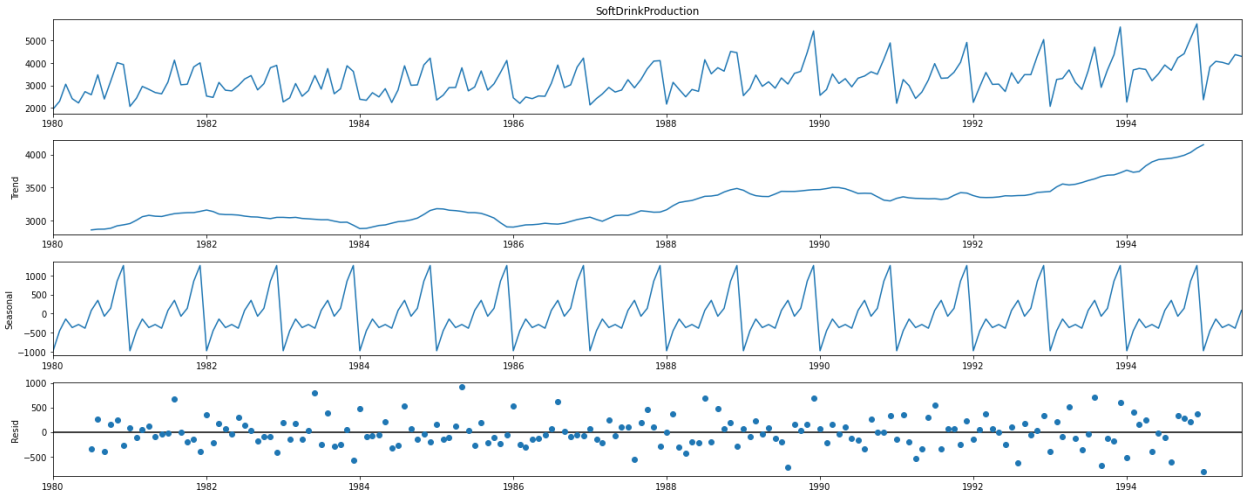
Plot the average Sales per month and the month on month percentage change of Sales



The above two graphs tells us the Average 'Sales' and the Percentage change of 'Sales' with respect to the time.

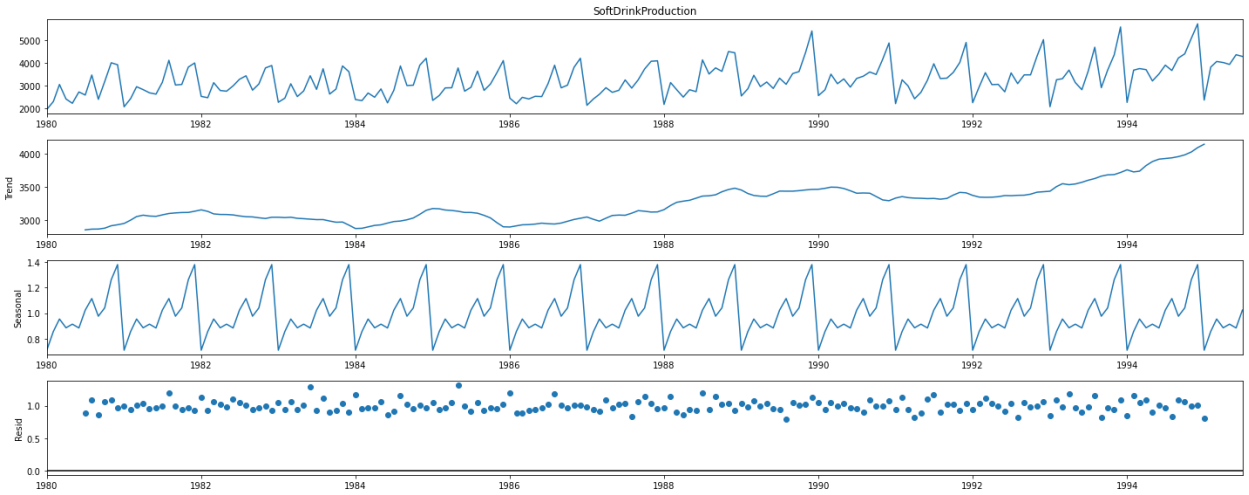
Decompose the Time Series and plot the different components

Additive Decomposition



As per the 'additive' decomposition, we see that there is an increasing trend in sales of SoftDrinkProduction starting from 1980 to 1995. Seasonality is also clearly visible from the seasonal graph where trend lines are forming the peaks each year. Residuals seems to be scattered from the 0 level. Indicating that the series is not additive.

Multiplicative Decomposition



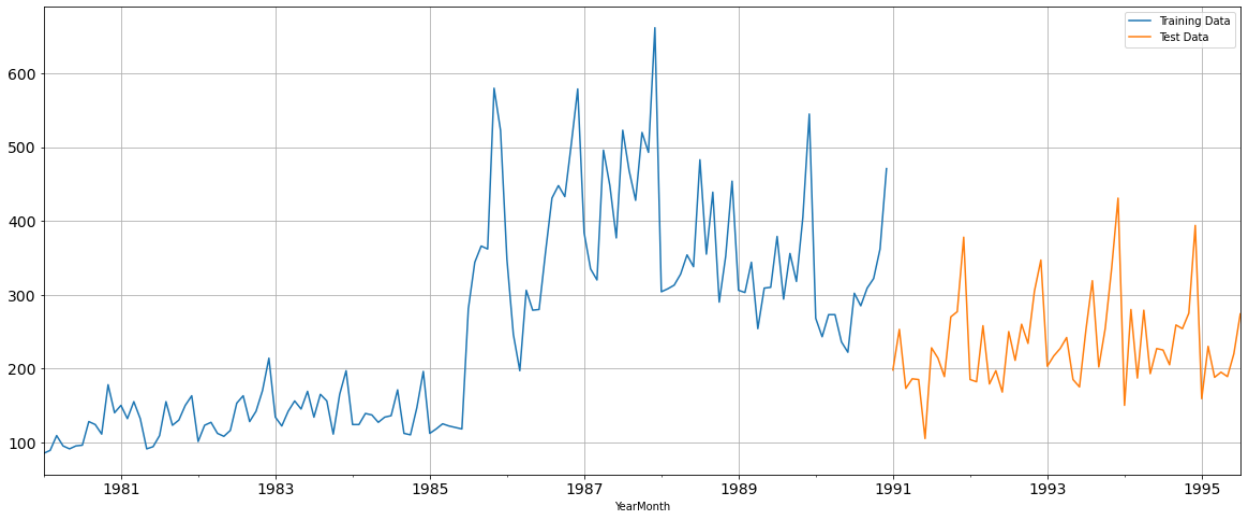
The trend and seasonality are present same as in case of additive model. But residuals plot is clearly showing the concentration of data towards 1 point. Hence it can be concluded that series is multiplicative.

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

<u>Shoesales Data</u>

Train has been split with data before 1991 and test with data after 1991. Train has 132 rows and test has 55 rows.

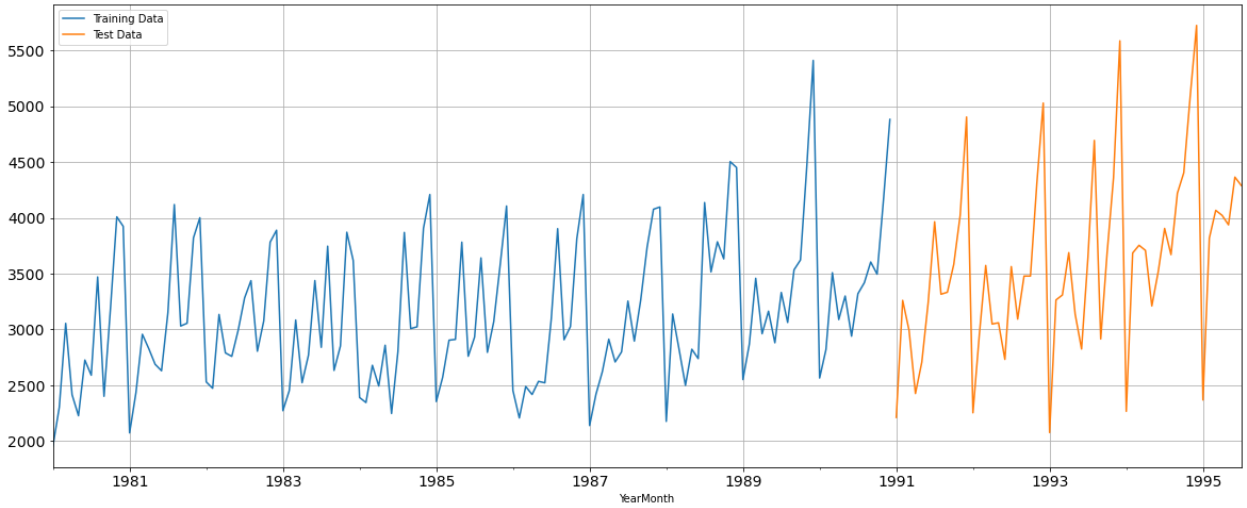
Graphical representation of test and train data



Soft Drink data

Train has been split with data before 1991 and test with data after 1991. Train has 132 rows and test has 55 rows.

Graphical representation of test and train data



4. Build all the 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.

Shoe Sales Data

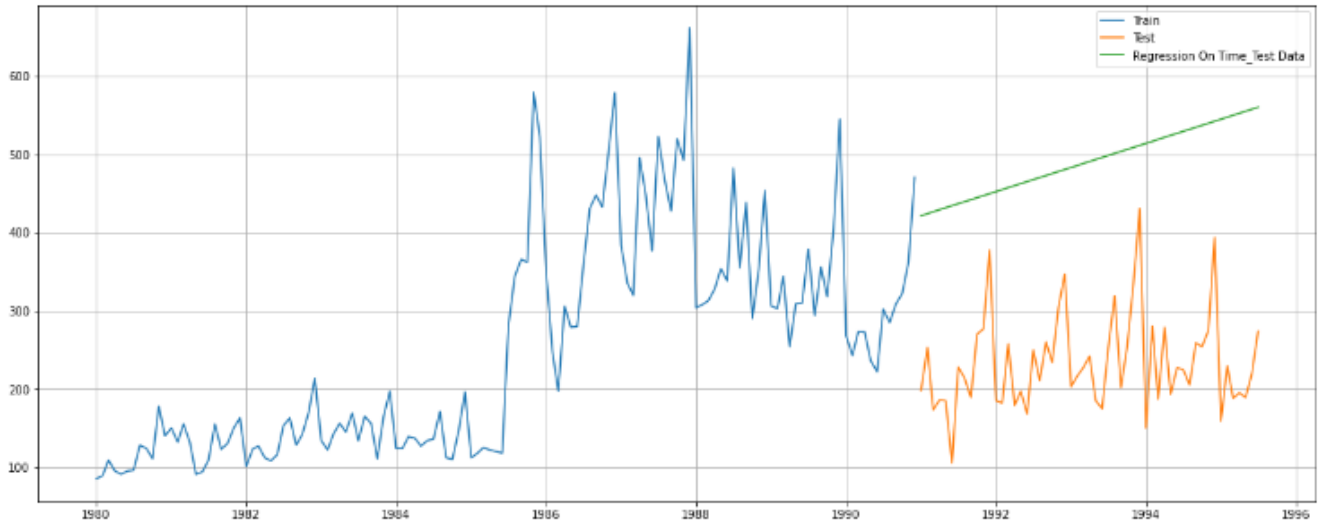
Model 1: Linear Regression

For this particular linear regression, we are going to regress the 'Shoe_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]
```

We see that we have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

Now that our training and test data has been modified, let us go ahead use *LinearRegression* to build the model on the training data and test the model on the test data.



The predicted trend is increasing.

Defining the functions for calculating the accuracy metrics.

```
For RegressionOnTime forecast on the Test Data,  RMSE is 266.276
```

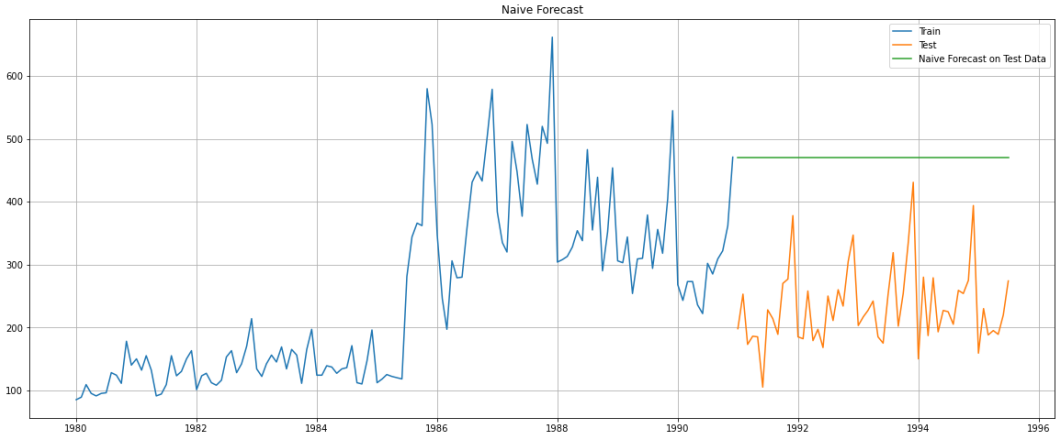
The RSME on test data value is 266.276, value is not very high but since seasonality is also not taken care by model this model is not suitable predictions on Shoe_Sales time series data.

	Test RMSE
RegressionOnTime	266.276472

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

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.

```
YearMonth
1991-01-01    471
1991-02-01    471
1991-03-01    471
1991-04-01    471
1991-05-01    471
Name: naive, dtype: int64
```



Model Evaluation

```
For Naive forecast on the Test Data,  RMSE is 245.121
```

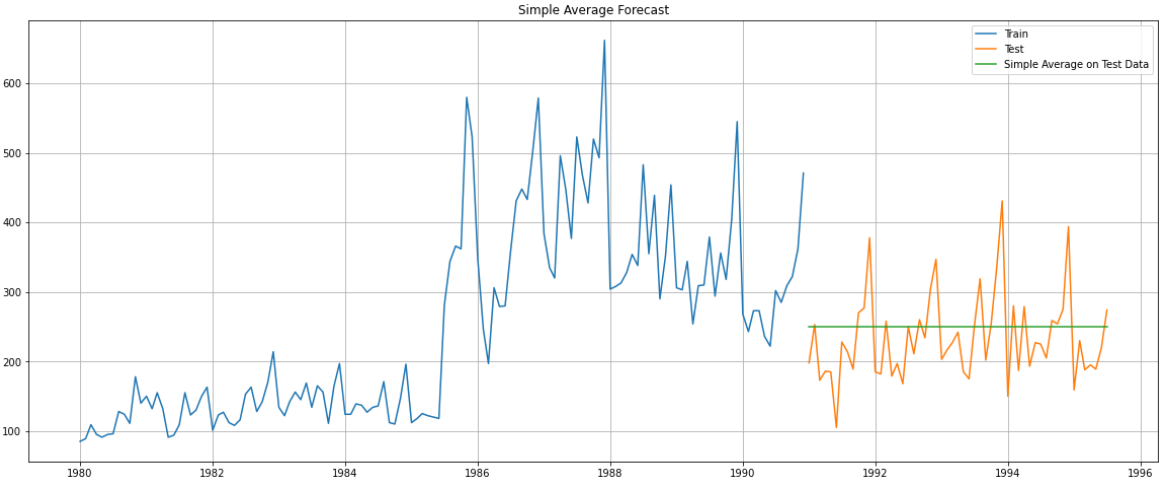
RSME is 245.121 which is lower than the linear regression model. But this model being too simple is not taking care of seasonality.

	Test RMSE
RegressionOnTime	266.276472
NaiveModel	245.121306

Method 3: Simple Average

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

YearMonth	Shoe_Sales	mean_forecast
1991-01-01	198	250.575758
1991-02-01	253	250.575758
1991-03-01	173	250.575758
1991-04-01	186	250.575758
1991-05-01	185	250.575758



The plot shows a flat line

```
For Simple Average forecast on the Test Data,  RMSE is 63.985
```

RMSE is 63.985 which is less than Naïve model & regression model but without seasonality component.

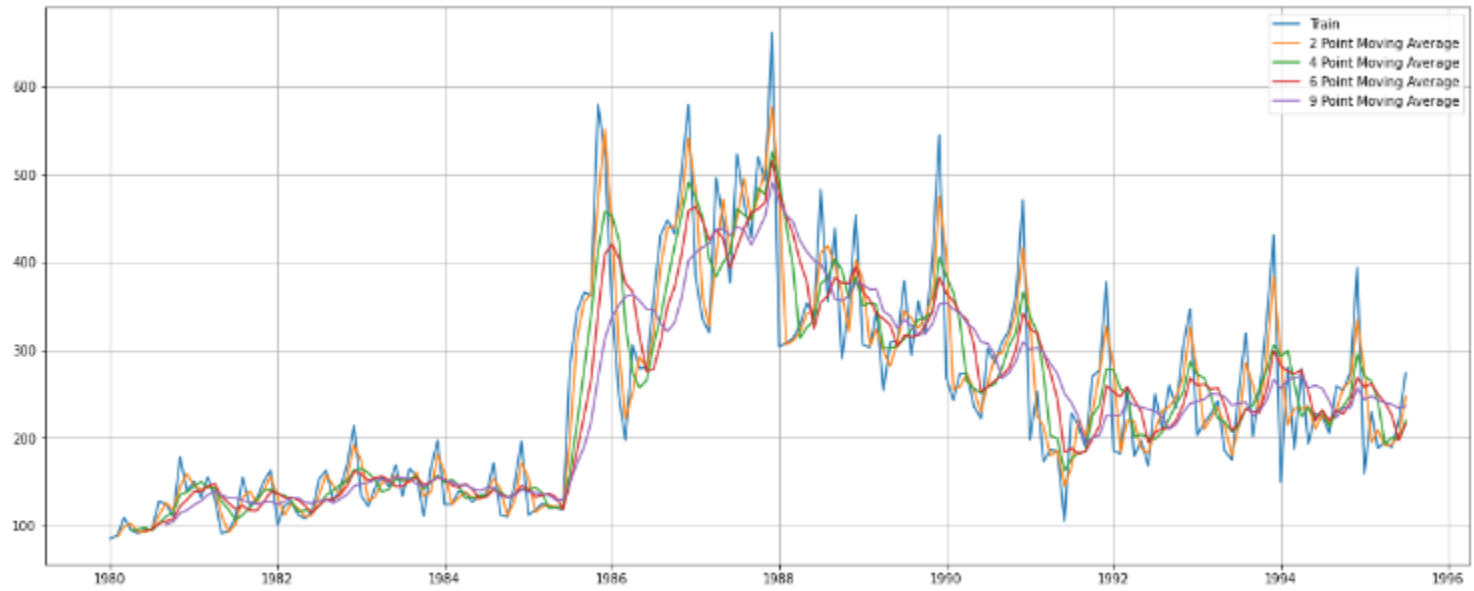
	Test RMSE
RegressionOnTime	266.276472
NaiveModel	245.121306
SimpleAverageModel	63.984570

Method 4: Moving Average(MA)

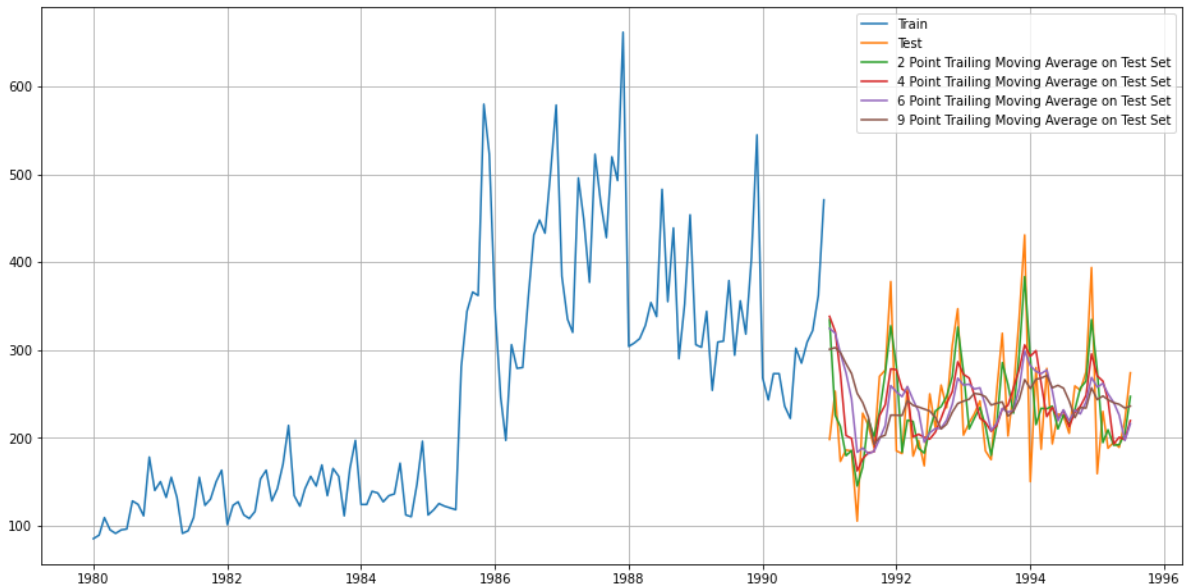
For the moving average model, we are going to calculate rolling means (or moving averages) for different intervals. The best interval can be determined by the maximum accuracy (or the minimum error) over here.

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

YearMonth	Shoe_Sales	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	85	NaN	NaN	NaN	NaN
1980-02-01	89	87.0	NaN	NaN	NaN
1980-03-01	109	99.0	NaN	NaN	NaN
1980-04-01	95	102.0	94.5	NaN	NaN
1980-05-01	91	93.0	96.0	NaN	NaN



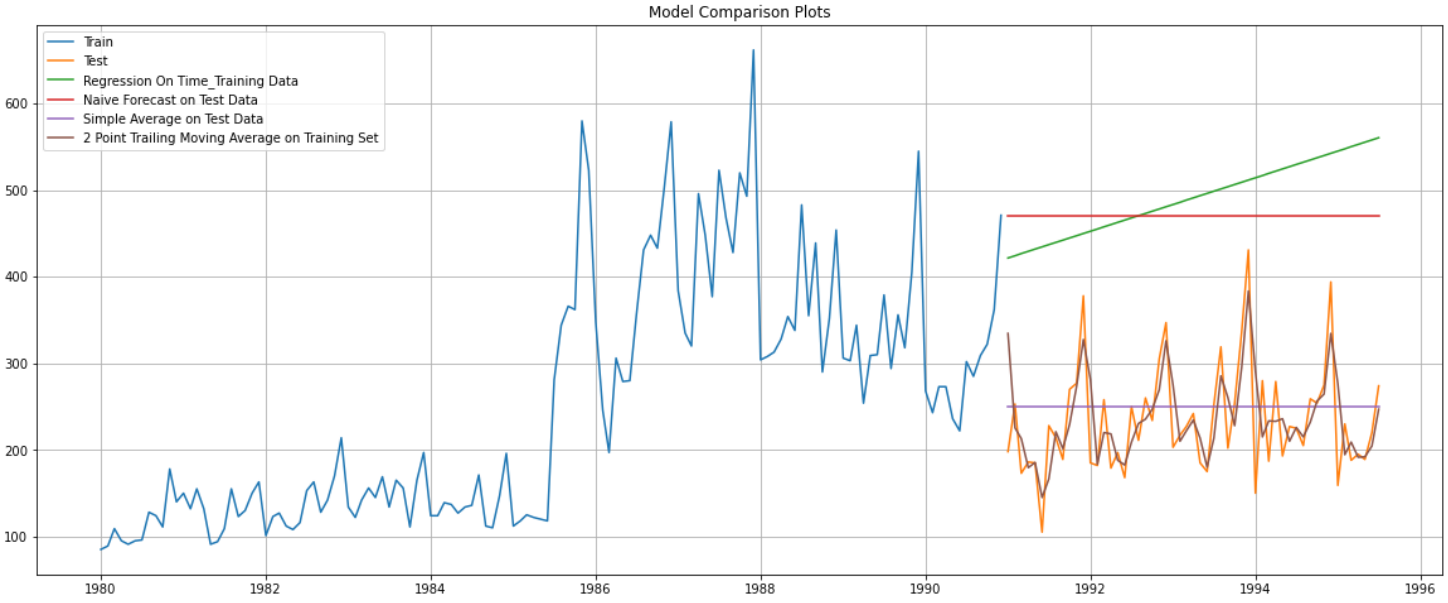
Considering criteria that testing data should start from 1991 onwards, training and testing data is prepared.



```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 45.949
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 57.873
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 63.457
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 67.724
```

Out of Linear regression, Naïve , Simple average & moving average , best performing model with lowest RMSE is 2 point moving average.

	Test RMSE
RegressionOnTime	266.276472
NaiveModel	245.121306
SimpleAverageModel	63.984570
2pointTrailingMovingAverage	45.948736
4pointTrailingMovingAverage	57.872688
6pointTrailingMovingAverage	63.456893
9pointTrailingMovingAverage	67.723648



Comparison plot shows the best fit model in brown color line for 2 point moving average appropriately fitting on the actual test values.

Method 5 : SES - ETS(A, N, N) - Simple Exponential Smoothing with additive errors

The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing (SES).

This method is suitable for forecasting data with no clear trend or seasonal pattern.

- $F_{t+1} = \alpha Y_t + (1 - \alpha) F_t$

Parameter α is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing.

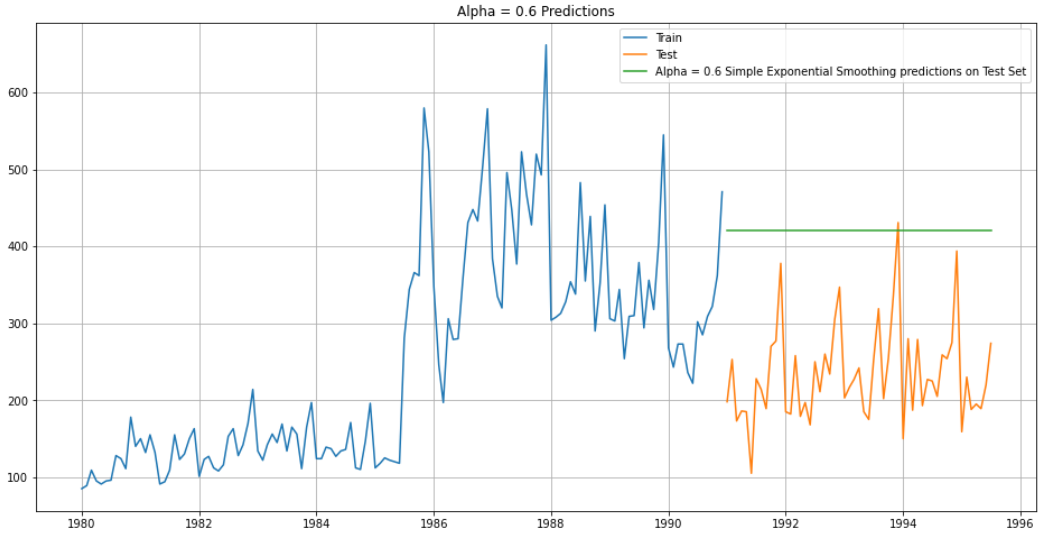
SimpleExpSmoothing class must be instantiated and passed the training data.

The fit() function is then called providing the fit configuration, the alpha value, smoothing_level. If this is omitted or set to None, the model will automatically optimize the value.

```
{'smoothing_level': 0.6050493159152485,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 88.82865380886527,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Alpha value is 0.60

YearMonth	Shoe_Sales	predict
1991-01-01	198	420.229871
1991-02-01	253	420.229871
1991-03-01	173	420.229871
1991-04-01	186	420.229871
1991-05-01	185	420.229871




```
For Alpha = 0.6 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 196.405
```

	Test RMSE
RegressionOnTime	266.276472
NaiveModel	245.121306
SimpleAverageModel	63.984570
2pointTrailingMovingAverage	45.948736
4pointTrailingMovingAverage	57.872686
6pointTrailingMovingAverage	63.456893
9pointTrailingMovingAverage	67.723648
Alpha=0.6,SimpleExponentialSmoothing	196.404850

Setting different alpha values. Remember, the 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.

	Alpha Values	Train RMSE	Test RMSE
0	0.3	74.555356	143.400350
1	0.4	73.062722	162.553211
2	0.5	72.200617	180.072484
3	0.6	71.902349	195.663327
4	0.7	72.131707	209.658339
5	0.8	72.846955	222.417584
6	0.9	74.023429	234.188186

0.3 alpha value seems to have low RMSE score than 0.6 alpha value.

Method 6 - Holt - ETS(A, A, N) - Holt's linear method with additive errors

Double Exponential Smoothing

Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecating the short term avarage value or level and the other for capturing the trend.

- Intercept or Level equation, L_t is given by: $L_t = \alpha Y_t + (1 - \alpha)F_t$
- Trend equation is given by $T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$

Here, α and β are the smoothing constants for level and trend, respectively,

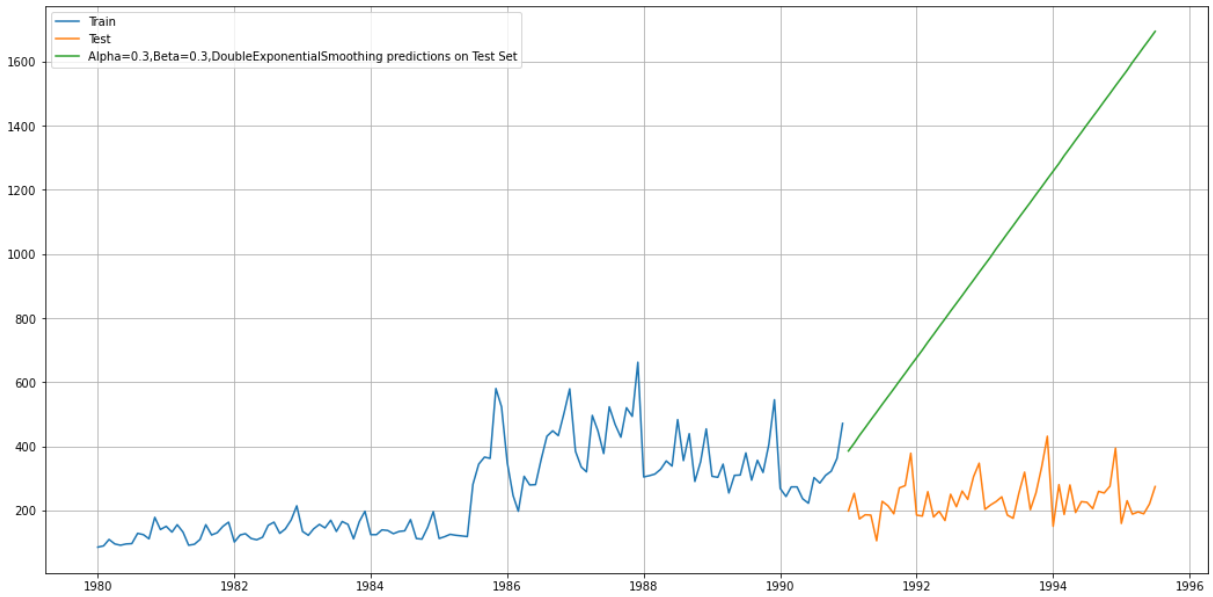
- $0 < \alpha < 1$ and $0 < \beta < 1$.

The forecast at time t + 1 is given by

- $F_{t+1} = L_t + T_t$
- $F_{t+n} = L_t + nT_t$

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	84.736667	890.968504
8	0.4	0.3	82.660727	1132.467007
16	0.5	0.3	80.640171	1264.035724
1	0.3	0.4	88.649551	1270.606989
24	0.6	0.3	79.699269	1355.955337

Plotting on both the Training and Test data



	Test RMSE
RegressionOnTime	266.276472
NaiveModel	245.121306
SimpleAverageModel	63.984570
2pointTrailingMovingAverage	45.948736
4pointTrailingMovingAverage	57.872686
6pointTrailingMovingAverage	63.456893
9pointTrailingMovingAverage	67.723648
Alpha=0.6,SimpleExponentialSmoothing	196.404850
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	890.968504

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

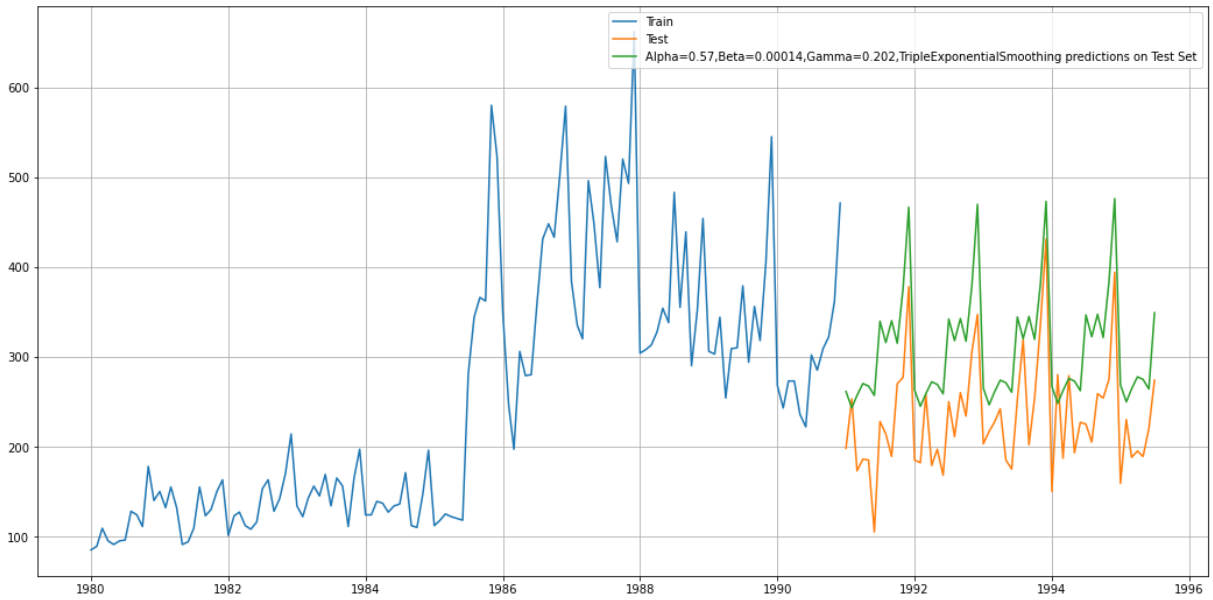
Three parameters α, β and γ are estimated in this model. Level, Trend and Seasonality are accounted for in this model.

```
{'smoothing_level': 0.5711286329525818,
 'smoothing_trend': 0.00014781930867568429,
 'smoothing_seasonal': 0.20294733706077994,
 'damping_trend': nan,
 'initial_level': 116.35529208070726,
 'initial_trend': 0.11219854465675648,
 'initial_seasons': array([1.05679343, 1.01130311, 1.2337466 , 1.40663129, 1.32162715,
       1.07936886, 1.18018187, 1.50183082, 1.72369093, 1.4704132 ,
       1.75485304, 1.92101444]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Prediction on the test data

YearMonth	Shoe_Sales	auto_predict
1991-01-01	198	261.342543
1991-02-01	253	243.085370
1991-03-01	173	256.996702
1991-04-01	186	270.198135
1991-05-01	185	267.375606

Plotting on both the Training and Test using autofit



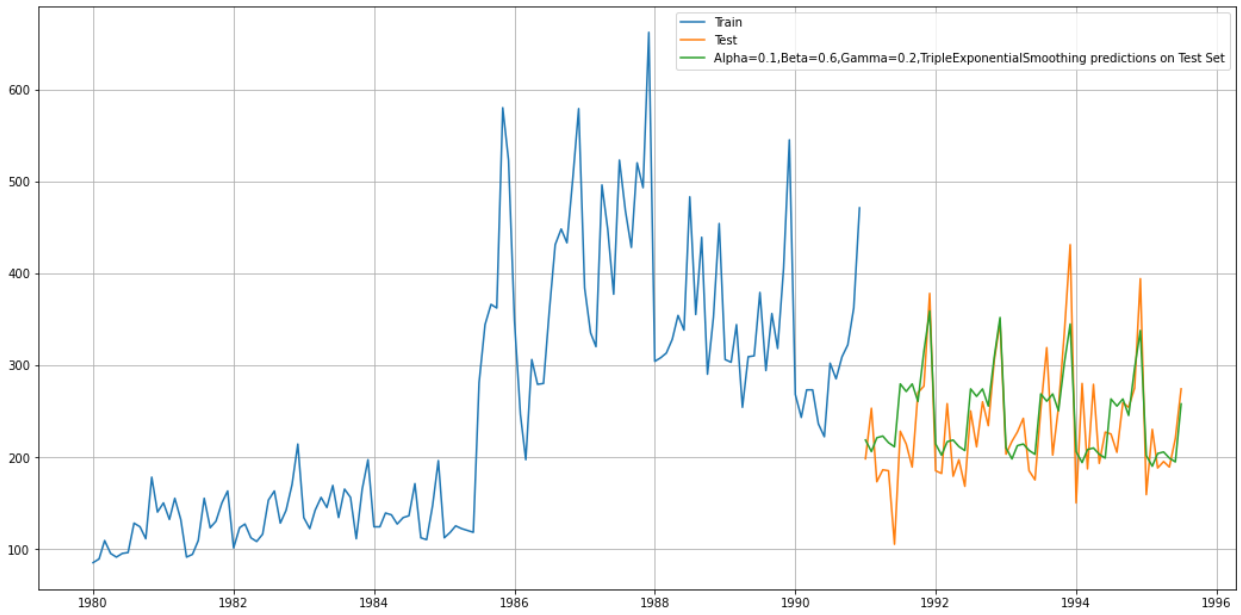
For Alpha=0.57,Beta=0.00014,Gamma=0.202, Triple Exponential Smoothing Model forecast on the Test Data, RMSE is 83.734

	Test RMSE
RegressionOnTime	266.276472
NaiveModel	245.121306
SimpleAverageModel	63.984570
2pointTrailingMovingAverage	45.948736
4pointTrailingMovingAverage	57.872686
6pointTrailingMovingAverage	63.456893
9pointTrailingMovingAverage	67.723648
Alpha=0.6,SimpleExponentialSmoothing	196.404850
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	890.968504
Alpha=0.57,Beta=0.00014,Gamma=0.202,TripleExponentialSmoothing	83.734048

RMSE for Alpha=0.57,Beta=0.00014,Gamma=0.202,TripleExponentialSmoothing is 83 which is very less as compared to other methods.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
51	0.1	0.6	0.2	74.696236	41.237522
91	0.1	1.0	0.2	101.394646	43.047089
44	0.1	0.5	0.5	94.006921	43.849368
50	0.1	0.6	0.1	70.879943	44.415011
37	0.1	0.4	0.8	120.751189	45.132153

Plotting on both the Training and Test data using brute force alpha, beta and gamma determination



Sorted by RMSE values on the Test Data:

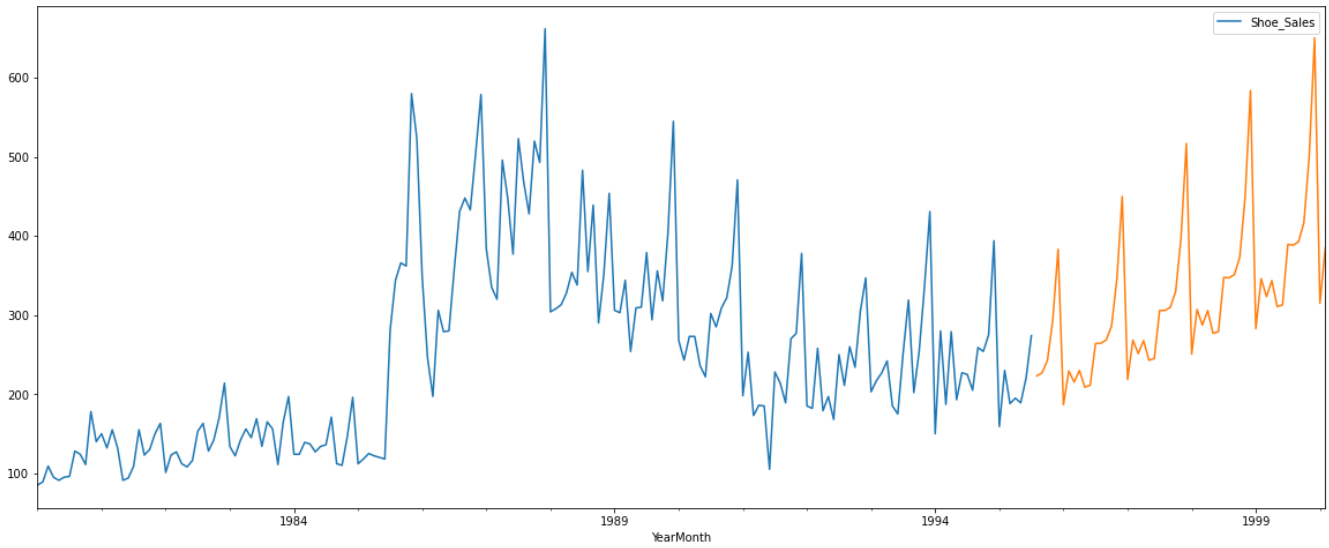
	Test RMSE
Alpha=0.1,Beta=0.6,Gamma=0.2,TripleExponentialSmoothing	41.237522
2pointTrailingMovingAverage	45.948736
4pointTrailingMovingAverage	57.872686
6pointTrailingMovingAverage	63.456893
SimpleAverageModel	63.984570
9pointTrailingMovingAverage	67.723648
Alpha=0.57,Beta=0.00014,Gamma=0.202,TripleExponentialSmoothing	83.734048
Alpha=0.6,SimpleExponentialSmoothing	196.404850
NaiveModel	245.121306
RegressionOnTime	266.276472
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	890.968504

Alpha=0.1,Beta=0.6,Gamma=0.2,TripleExponentialSmoothing method is the best model whose RMSE is 41.2 which is very less as compared to other methods/models.

Full Model Forecasting

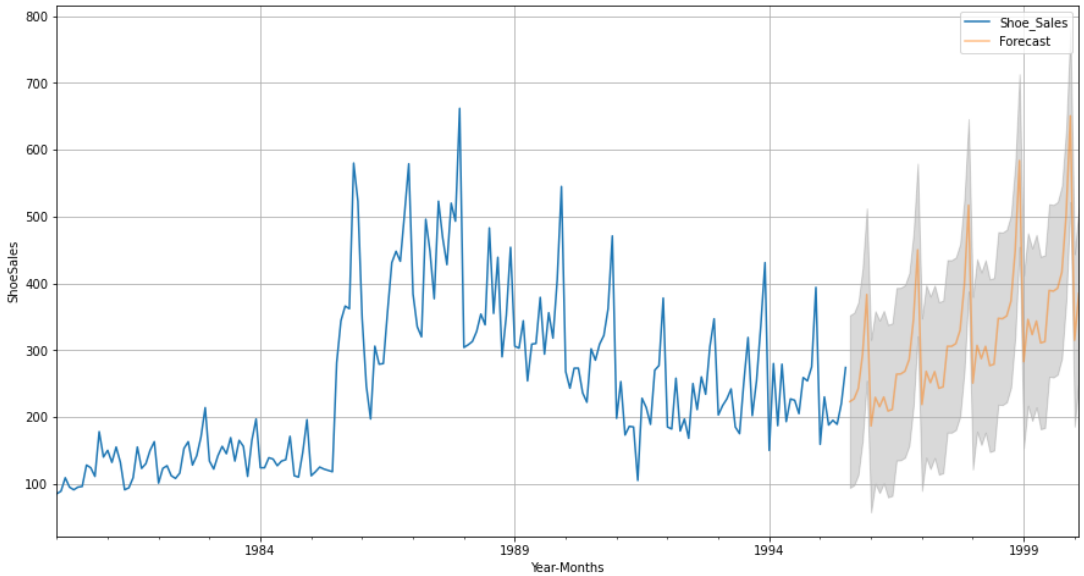
RMSE: 65.72957833447327

Getting the predictions for the same number of times stamps that are present in the test data



We have calculated the upper and lower confidence bands at 95% confidence level.

	lower_CI	prediction	upper_ci
1995-08-01	93.929879	223.070887	352.211856
1995-09-01	97.890414	227.031403	356.172392
1995-10-01	113.444210	242.585199	371.726188
1995-11-01	164.320143	293.461131	422.602120
1995-12-01	254.106116	383.247105	512.388093



The above picture shows an increasing trend of sales till 2000

Soft Drink data

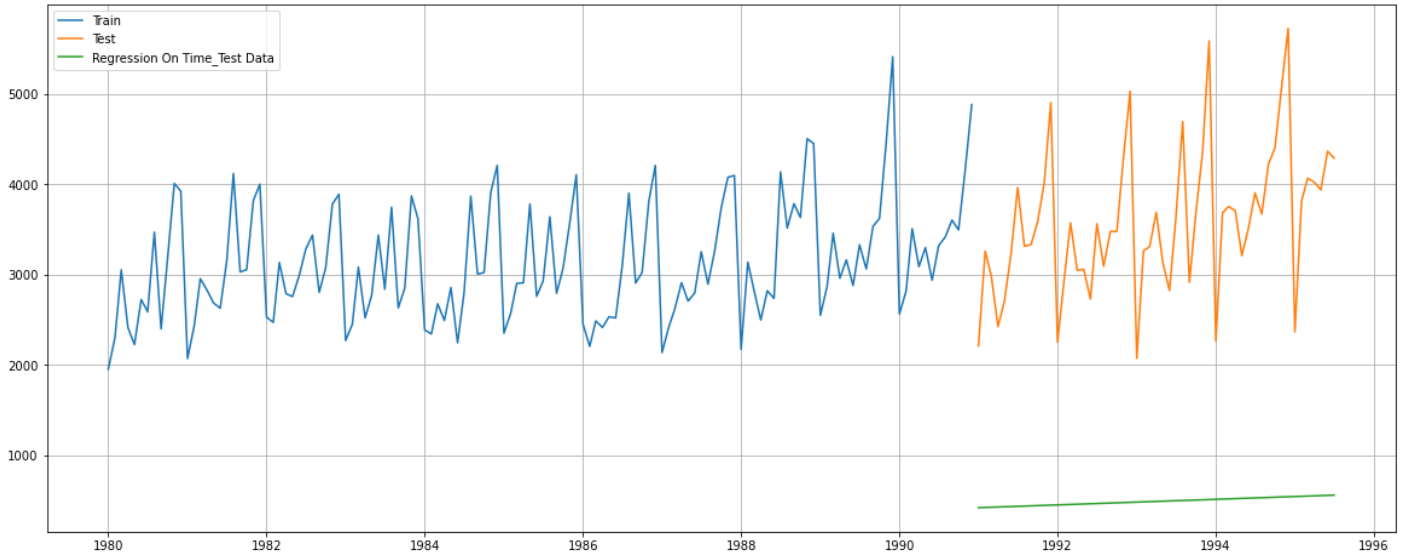
Model 1: Linear Regression

For this particular linear regression, we are going to regress the 'Soft Drink' 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]
```

We see that we have successfully the generated the numerical time instance order for both the training and test set. Now we will add these values in the training and test set.

Now that our training and test data has been modified, let us go ahead use *LinearRegression* to build the model on the training data and test the model on the test data.



The predicted trend is increasing.

Defining the functions for calculating the accuracy metrics.

```
For RegressionOnTime forecast on the Test Data,  RMSE is 3202.844
```

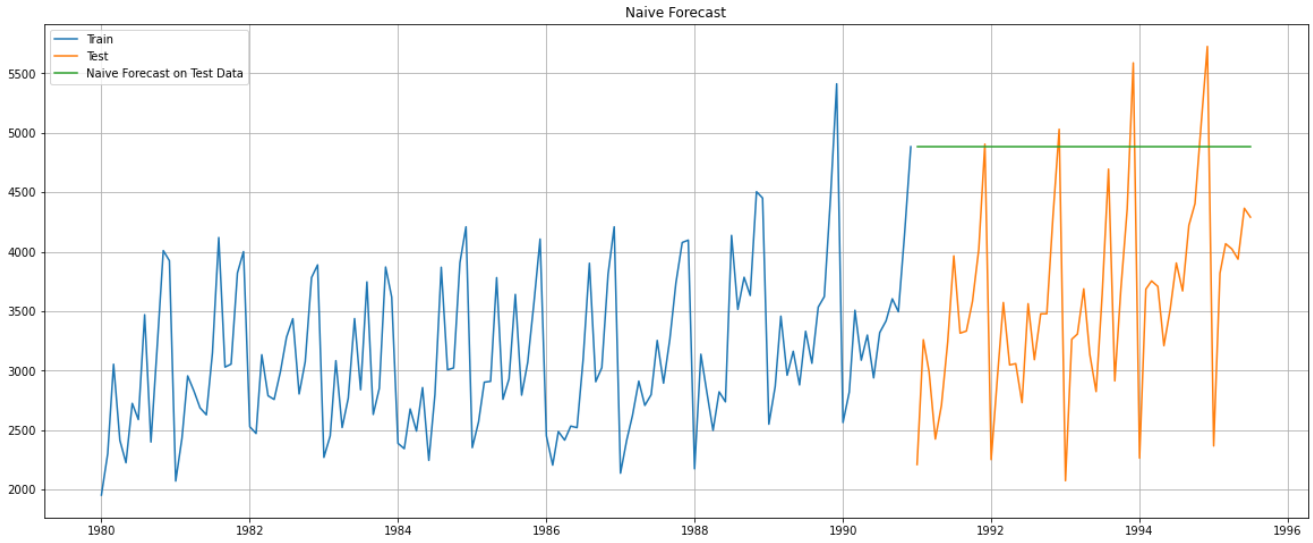
The RSME on test data value is 3202.844, value is not very high but since seasonality is also not taken care by model this model is not suitable predictions on Soft Drink time series data.

	Test RMSE
RegressionOnTime	3202.844447

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

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.

```
YearMonth
1991-01-01    4882
1991-02-01    4882
1991-03-01    4882
1991-04-01    4882
1991-05-01    4882
Name: naive, dtype: int64
```



Model Evaluation

```
For Naive forecast on the Test Data,  RMSE is 1519.259
```

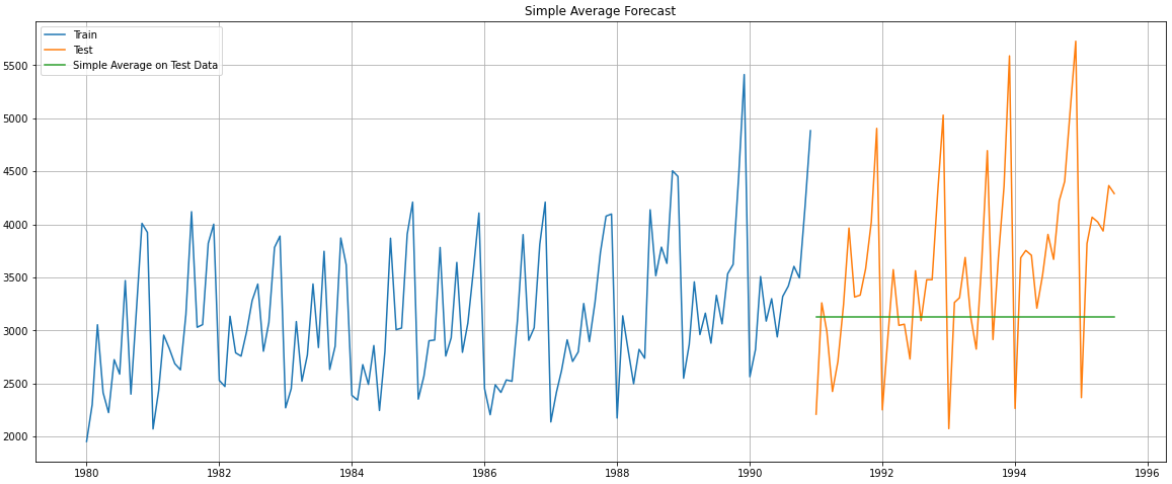
RSME is 1519.259 which is lower than the linear regression model. But this model being too simple is not taking care of seasonality.

	Test RMSE
RegressionOnTime	3202.844447
NaiveModel	1519.259233

Method 3: Simple Average

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

YearMonth	SoftDrinkProduction	mean_forecast
1991-01-01	2211	3124.166667
1991-02-01	3260	3124.166667
1991-03-01	2992	3124.166667
1991-04-01	2425	3124.166667
1991-05-01	2707	3124.166667



The plot shows a flat line

```
For Simple Average forecast on the Test Data,  RMSE is 934.353
```

RMSE is 934.353 which is less than Naïve model & regression model but without seasonality component.

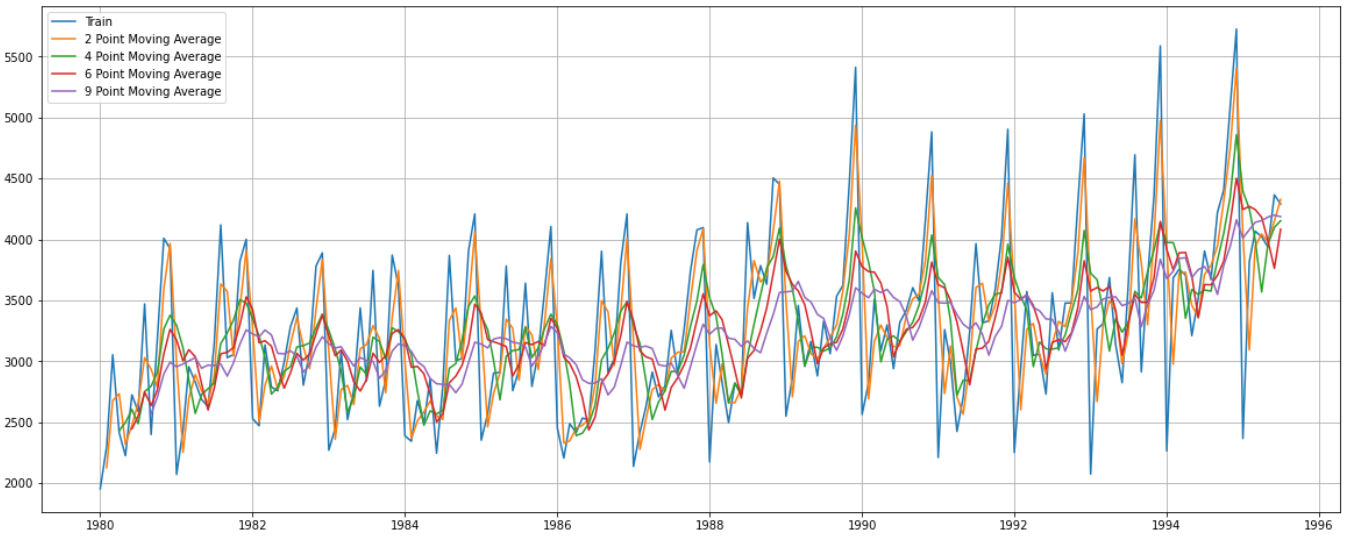
	Test RMSE
RegressionOnTime	3202.844447
NaiveModel	1519.259233
SimpleAverageModel	934.353358

Method 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.

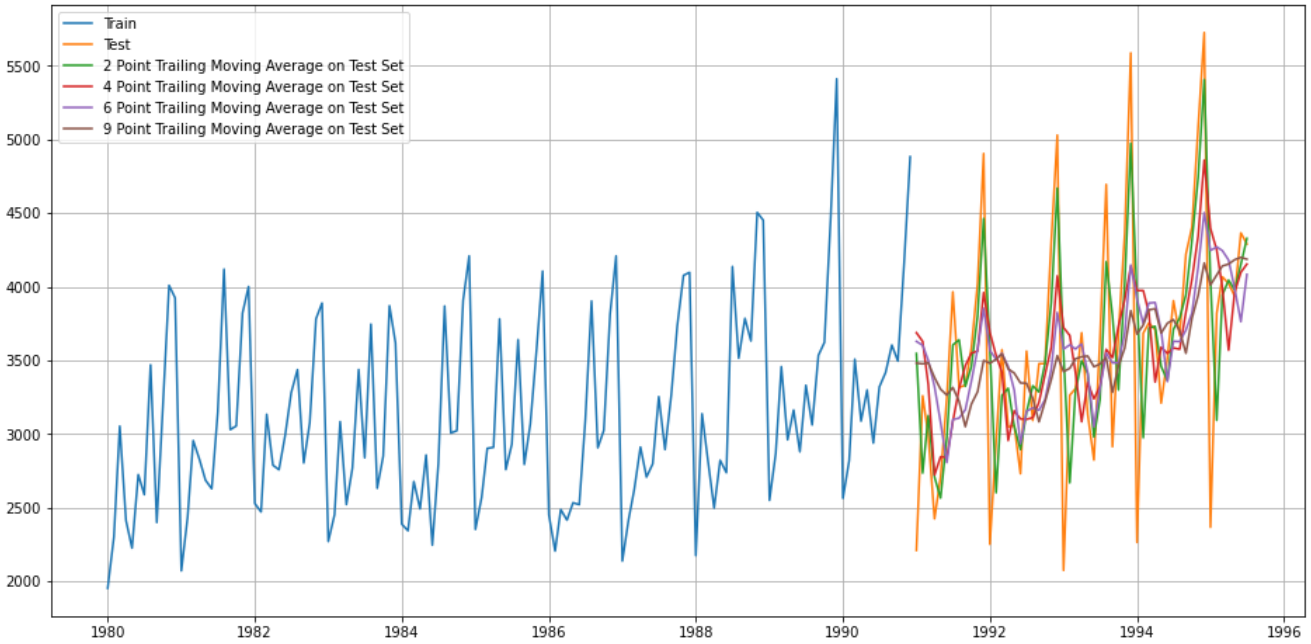
For Moving Average, we are going to average over the entire data.

YearMonth	SoftDrinkProduction	Trailing_2	Trailing_4	Trailing_6	Trailing_9
1980-01-01	1954	NaN	NaN	NaN	NaN
1980-02-01	2302	2128.0	NaN	NaN	NaN
1980-03-01	3054	2678.0	NaN	NaN	NaN
1980-04-01	2414	2734.0	2431.0	NaN	NaN
1980-05-01	2226	2320.0	2499.0	NaN	NaN



Considering criteria that testing data should start from 1991 onwards, training and testing data is prepared.

Plotting on both the Training and Test data

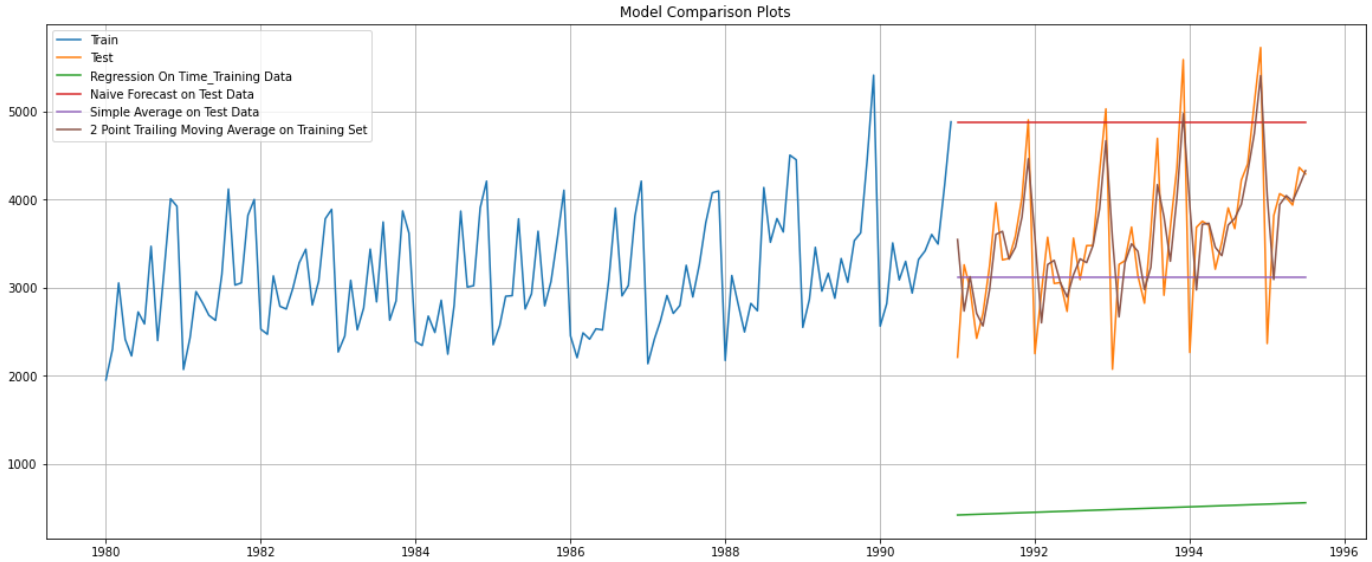



```
For 2 point Moving Average Model forecast on the Training Data,  RMSE is 556.725
For 4 point Moving Average Model forecast on the Training Data,  RMSE is 687.182
For 6 point Moving Average Model forecast on the Training Data,  RMSE is 710.514
For 9 point Moving Average Model forecast on the Training Data,  RMSE is 735.890
```

Out of Linear regression, Naïve , Simple average & moving average , best performing model with lowest RMSE is 2 point moving average.

	Test RMSE
RegressionOnTime	3202.844447
NaiveModel	1519.259233
SimpleAverageModel	934.353358
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827

Plotting on both Training and Test data



Comparison plot shows the best fit model in brown color line for 2 point moving average appropriately fitting on the actual test values.

Method 5 : SES - ETS(A, N, N) - Simple Exponential Smoothing with additive errors

The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing (SES).

This method is suitable for forecasting data with no clear trend or seasonal pattern.

- $F_{t+1} = \alpha Y_t + (1-\alpha) F_t$

Parameter α is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing.

SimpleExpSmoothing class must be instantiated and passed the training data.

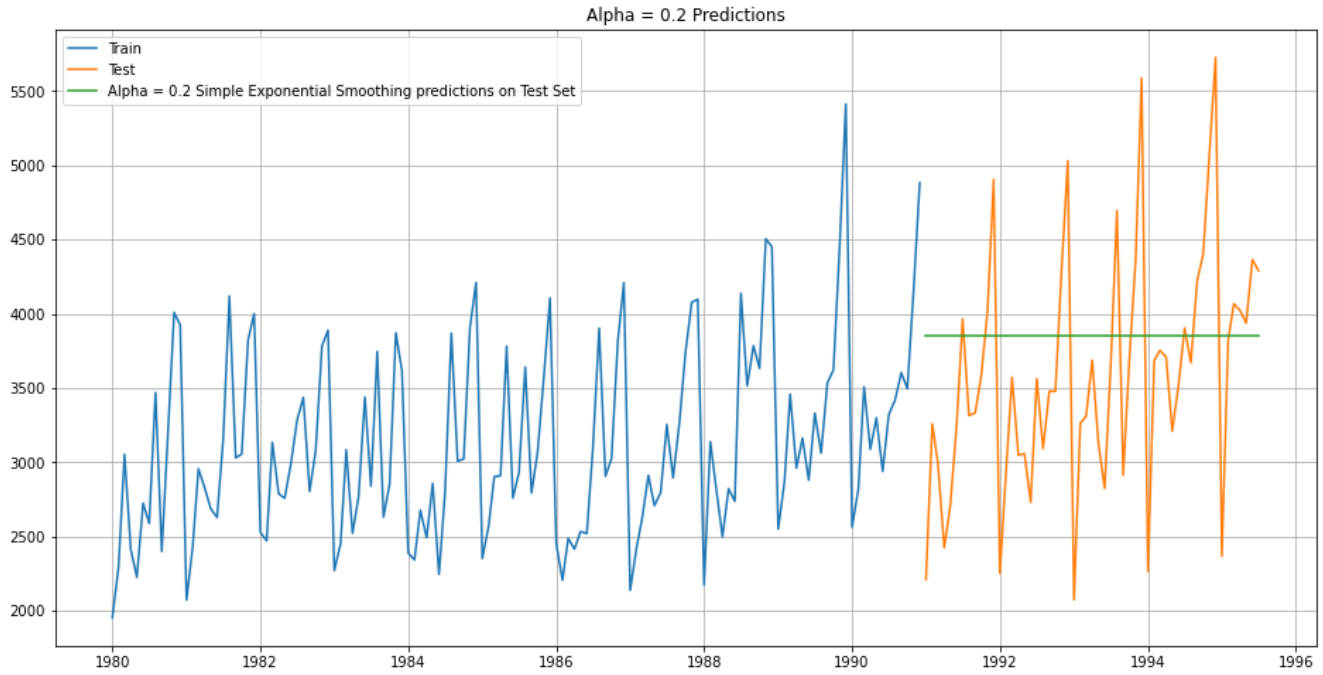
The fit() function is then called providing the fit configuration, the alpha value, smoothing_level. If this is omitted or set to None, the model will automatically optimize the value.

```
{'smoothing_level': 0.21628856026090065,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 2297.422897653051,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

Alpha value is 0.21

YearMonth	SoftDrinkProduction	predict
1991-01-01	2211	3853.781071
1991-02-01	3260	3853.781071
1991-03-01	2992	3853.781071
1991-04-01	2425	3853.781071
1991-05-01	2707	3853.781071

Plotting the Training data, Test data and the forecasted values



For Alpha = 0.2 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 847.635

	Test RMSE
RegressionOnTime	3202.844447
NaiveModel	1519.259233
SimpleAverageModel	934.353358
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827
Alpha=0.2,SimpleExponentialSmoothing	847.635259

Setting different alpha values. Remember, the 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.

	Alpha Values	Train RMSE	Test RMSE
0	0.1	648.758846	807.346865
1	0.2	645.186884	838.357158
2	0.3	650.458591	910.187416
3	0.4	656.803409	1005.179377
4	0.5	664.777265	1105.985227
5	0.6	674.988238	1203.565956
6	0.7	687.376817	1294.680933
7	0.8	701.579829	1378.198740
8	0.9	717.287681	1453.359494

0.1 alpha value seems to have low RMSE score than 0.2 alpha value.

Method 6 - Holt - ETS(A, A, N) - Holt's linear method with additive errors

Double Exponential Smoothing

Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecating the short term avarage value or level and the other for capturing the trend.

- Intercept or Level equation, L_t is given by: $L_t = \alpha Y_t + (1 - \alpha)F_t$
- Trend equation is given by $T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$

Here, α and β are the smoothing constants for level and trend, respectively,

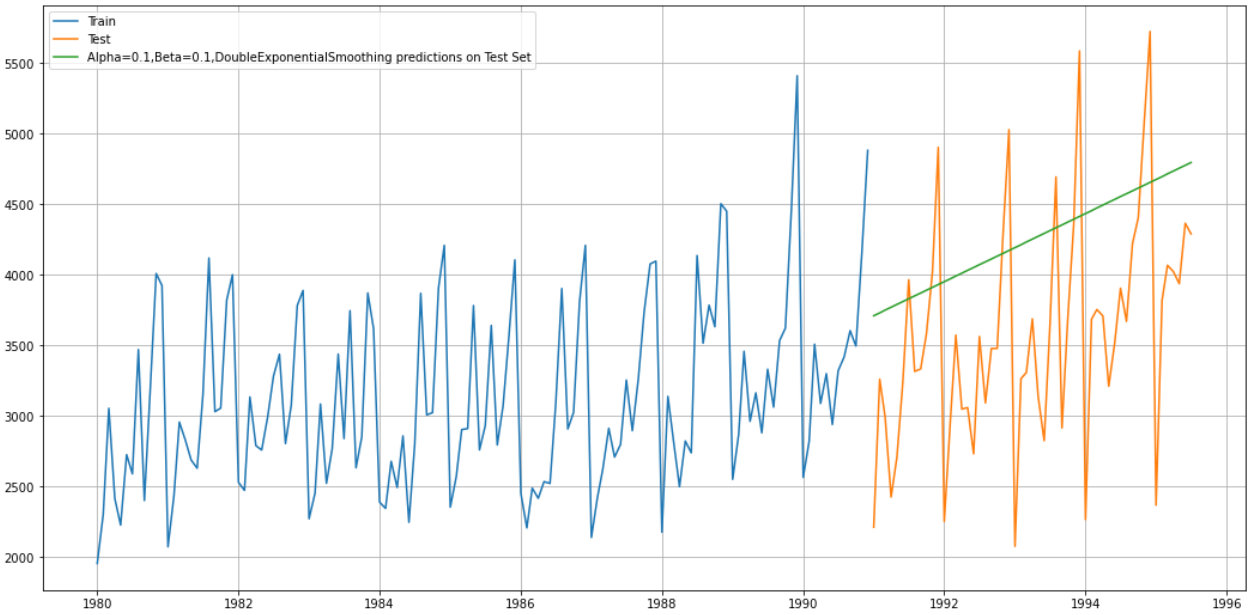
- $0 < \alpha < 1$ and $0 < \beta < 1$.

The forecast at time t + 1 is given by

- $F_{t+1} = L_t + T_t$
- $F_{t+n} = L_t + nT_t$

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	893.178173	982.938364
1	0.1	0.2	765.760815	1082.798709
10	0.2	0.1	721.404860	1524.533365
2	0.1	0.3	731.718292	1706.868339
20	0.3	0.1	698.165128	2306.003981

Plotting on both the Training and Test data



	Test RMSE
RegressionOnTime	3202.844447
NaiveModel	1519.259233
SimpleAverageModel	934.353358
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827
Alpha=0.2,SimpleExponentialSmoothing	847.635259
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	982.938364

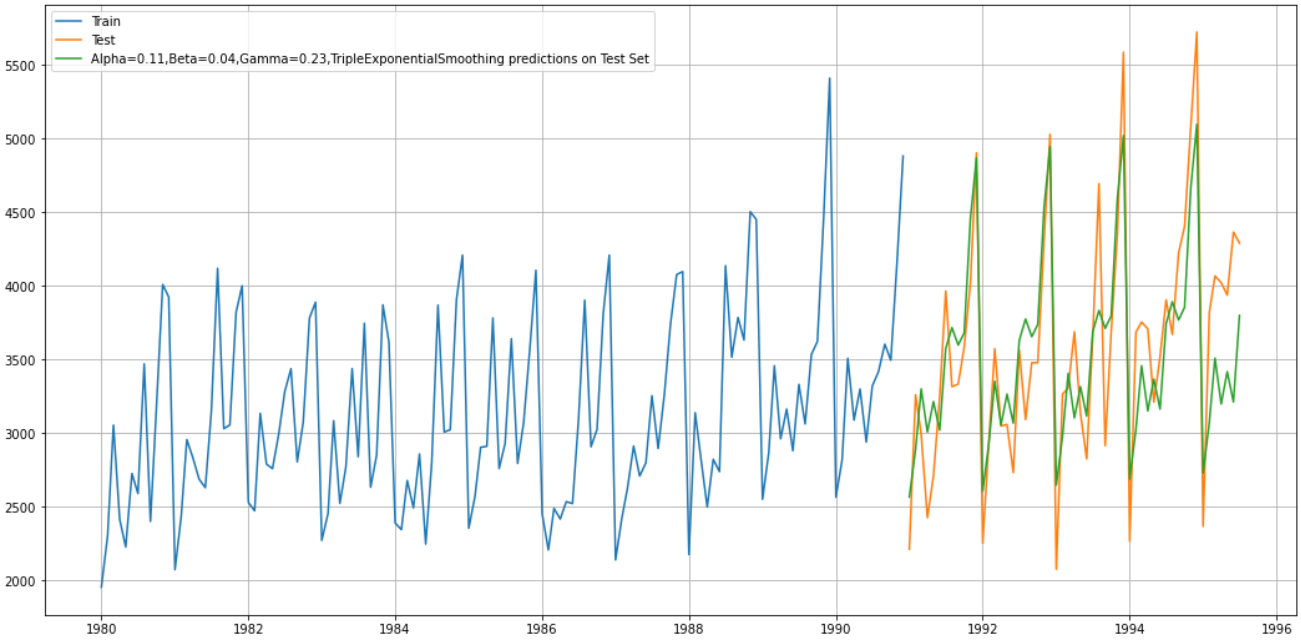
Method 7: Triple Exponential Smoothing (Holt - Winter's Model)

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

```
{'smoothing_level': 0.11128429736328378,
 'smoothing_trend': 0.04947326762762311,
 'smoothing_seasonal': 0.23037194388521623,
 'damping_trend': nan,
 'initial_level': 2803.0168193984414,
 'initial_trend': 10.486286228443715,
 'initial_seasons': array([0.80284001, 0.86968748, 1.08266033, 0.93954787, 0.96331944,
        0.98854326, 1.0654188 , 1.28504436, 1.0083707 , 1.0929922 ,
        1.36460606, 1.41709466]),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

YearMonth	SoftDrinkProduction	auto_predict
1991-01-01	2211	2584.869242
1991-02-01	3260	2887.850537
1991-03-01	2992	3300.056912
1991-04-01	2425	3007.348662
1991-05-01	2707	3213.743610

Plotting on both the Training and Test using autofit



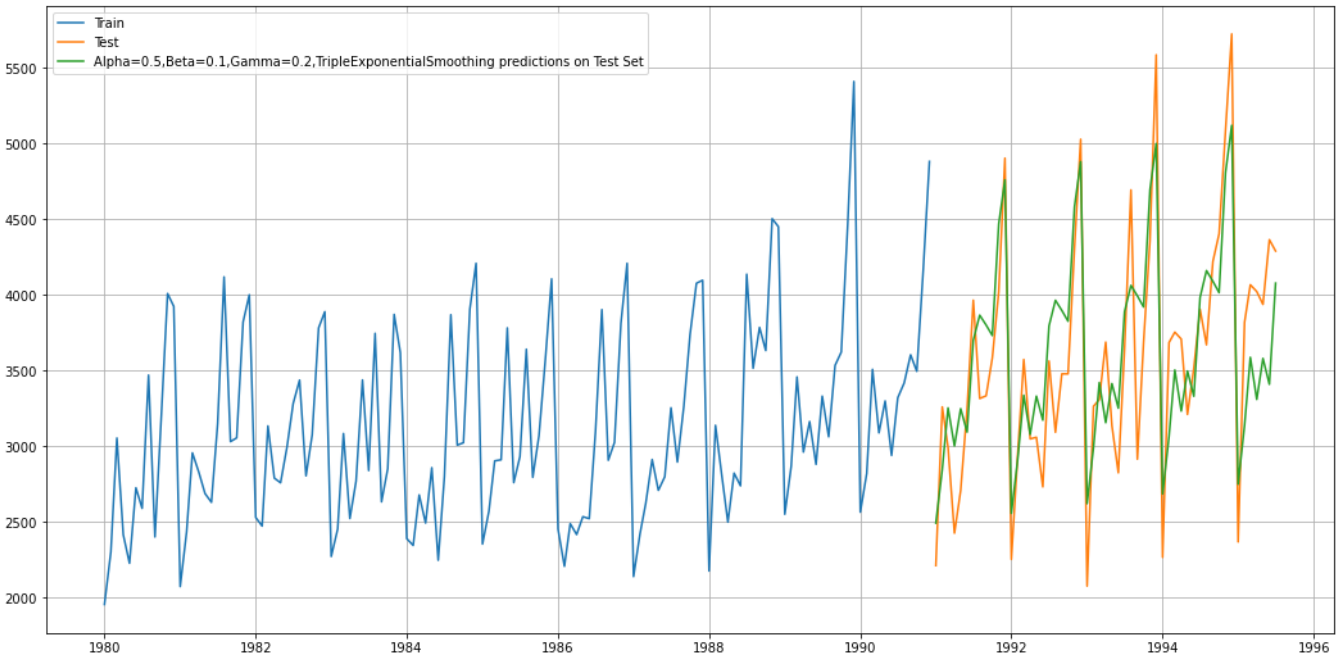

```
For Alpha=0.11,Beta=0.04,Gamma=0.23, Triple Exponential Smoothing Model forecast on the Test Data,  RMSE is 447.723
```

	Test RMSE
RegressionOnTime	3202.844447
NaiveModel	1519.259233
SimpleAverageModel	934.353358
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827
Alpha=0.2,SimpleExponentialSmoothing	847.635259
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	982.938364
Alpha=0.11,Beta=0.04,Gamma=0.23,TripleExponentialSmoothing	447.722581

RMSE for Alpha=0.11,Beta=0.04,Gamma=0.23,TripleExponentialSmoothing is 447 which is very less as compared to other methods.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
451	0.5	0.1	0.2	344.935215	442.214551
180	0.2	0.7	0.1	389.554491	447.999128
362	0.4	0.3	0.3	373.320057	453.599111
562	0.6	0.1	0.3	356.289387	461.108902
340	0.4	0.1	0.1	342.285498	462.751707

Plotting on both the Training and Test data using brute force alpha, beta and gamma determination



Sorted by RMSE values on the Test Data:

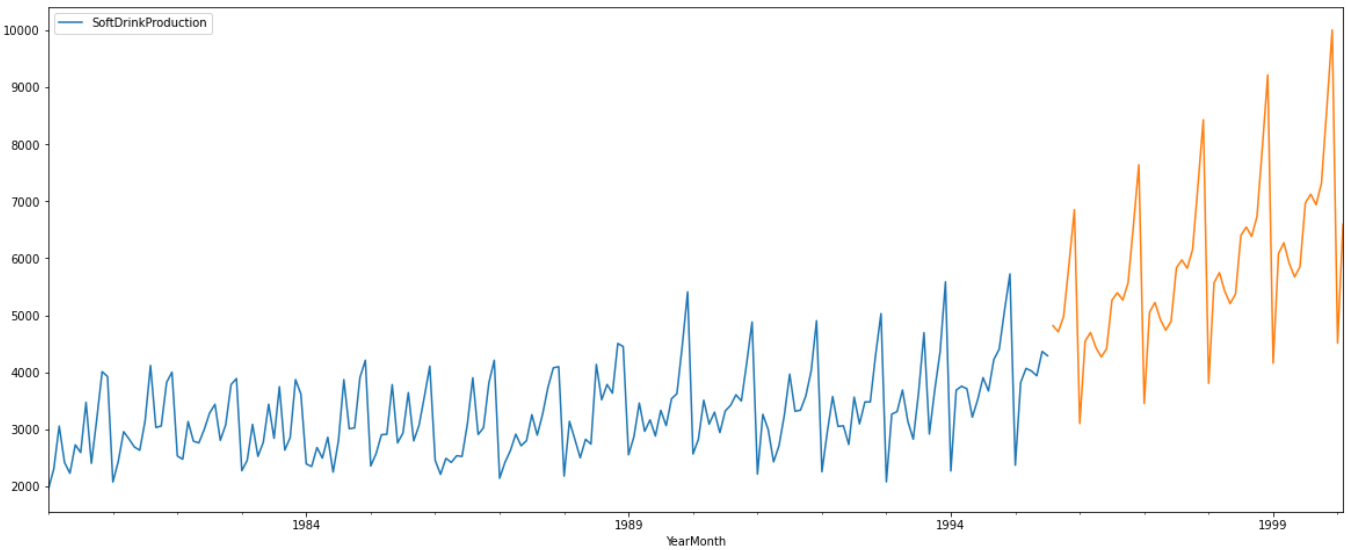
	Test RMSE
Alpha=0.5,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	442.214551
Alpha=0.11,Beta=0.04,Gamma=0.23,TripleExponentialSmoothing	447.722581
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827
Alpha=0.2,SimpleExponentialSmoothing	847.635259
SimpleAverageModel	934.353358
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	982.938364
NaiveModel	1519.259233
RegressionOnTime	3202.844447

Alpha=0.5,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing method is the best model whose RMSE is 442.2 which is very less as compared to other methods/models.

Full Model Forecasting

RMSE: 378.8789809329386

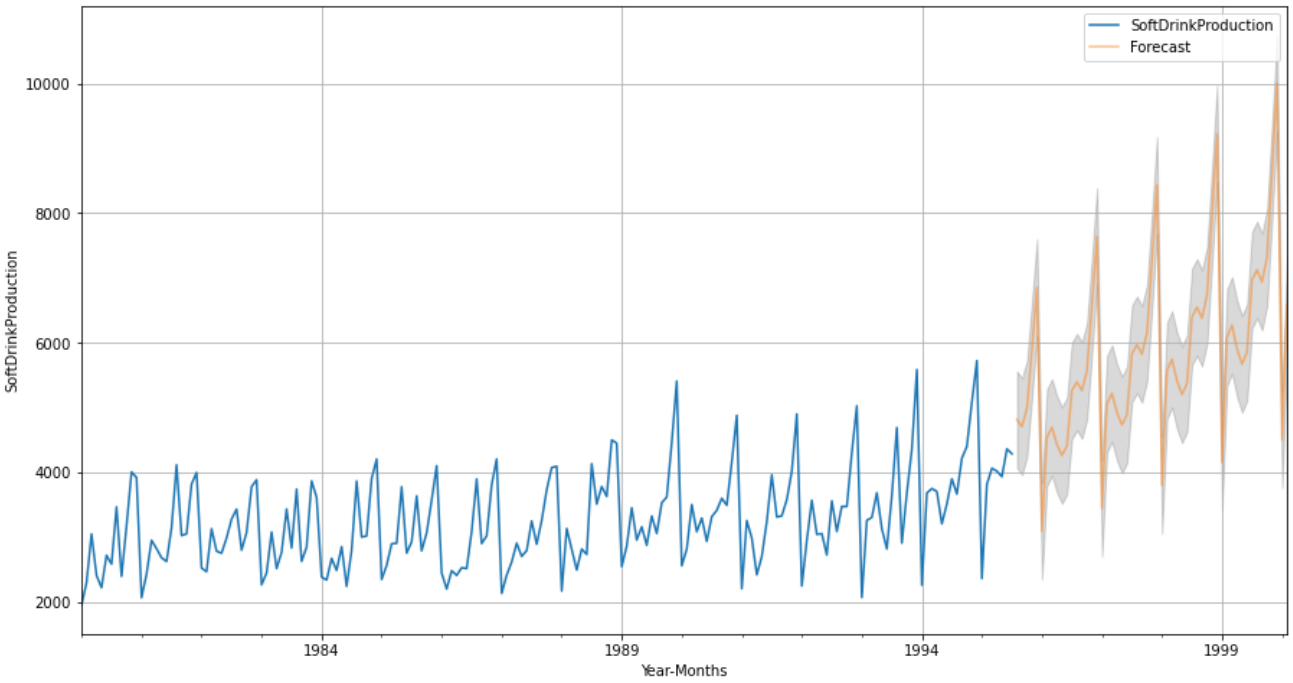
Getting the predictions for the same number of times stamps that are present in the test data



We have calculated the upper and lower confidence bands at 95% confidence level

	lower_CI	prediction	upper_ci
1995-08-01	4074.433203	4819.028412	5563.623622
1995-09-01	3964.461957	4709.057166	5453.652376
1995-10-01	4237.527942	4982.123152	5726.718362
1995-11-01	5154.522741	5899.117950	6643.713160
1995-12-01	6109.807328	6854.402538	7598.997748

Plot the forecast along with the confidence band



The above picture shows an increasing trend of sales till 2000

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.

ShoeSales Data

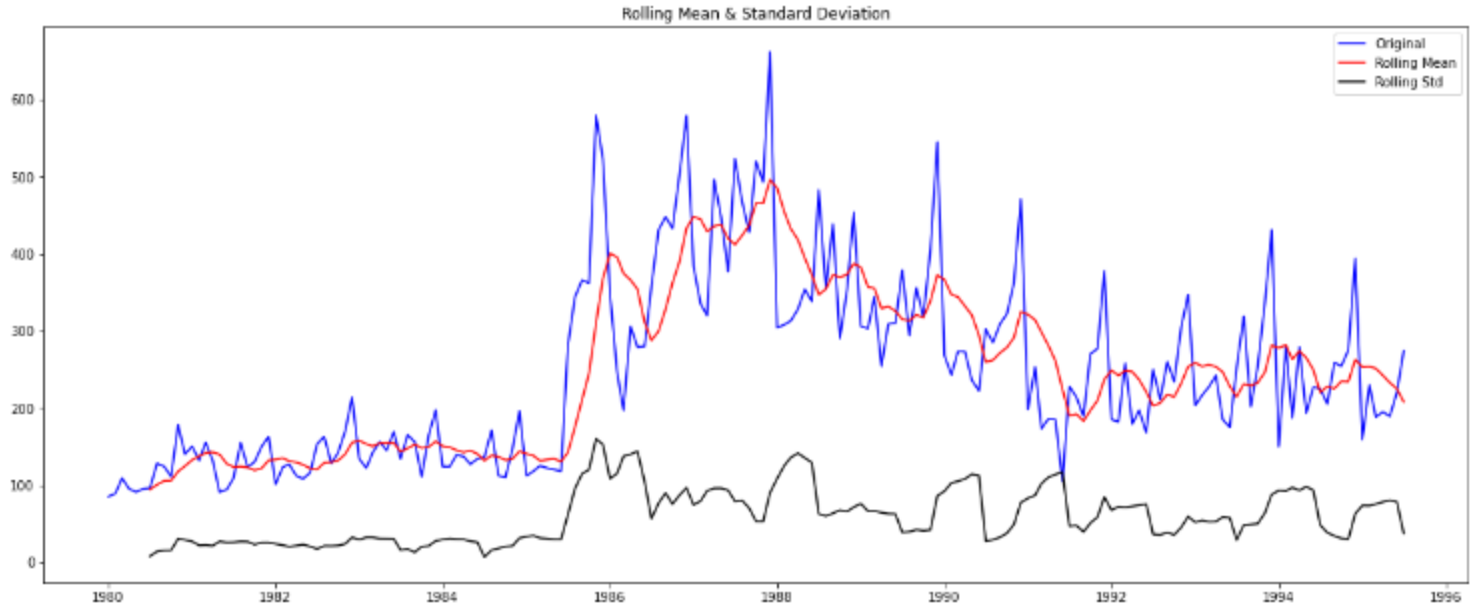
Check for stationarity of the whole Time Series data.

Dicky Fuller Test

Null Hypothesis H0- Series is not Stationary

Alternative Hypothesis H1- Series is Stationary

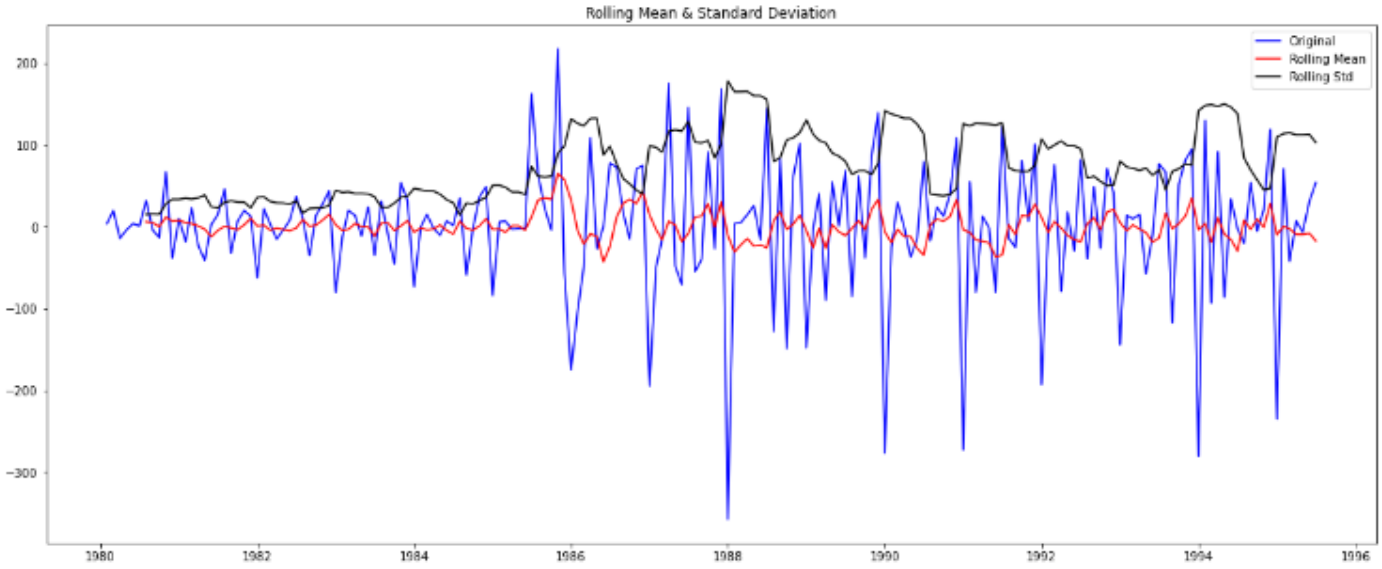
When p value is greater than 0.05 we accept the null hypothesis otherwise we reject the null hypothesis and accept the alternative hypothesis.




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.717397
p-value                          0.422172
#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.

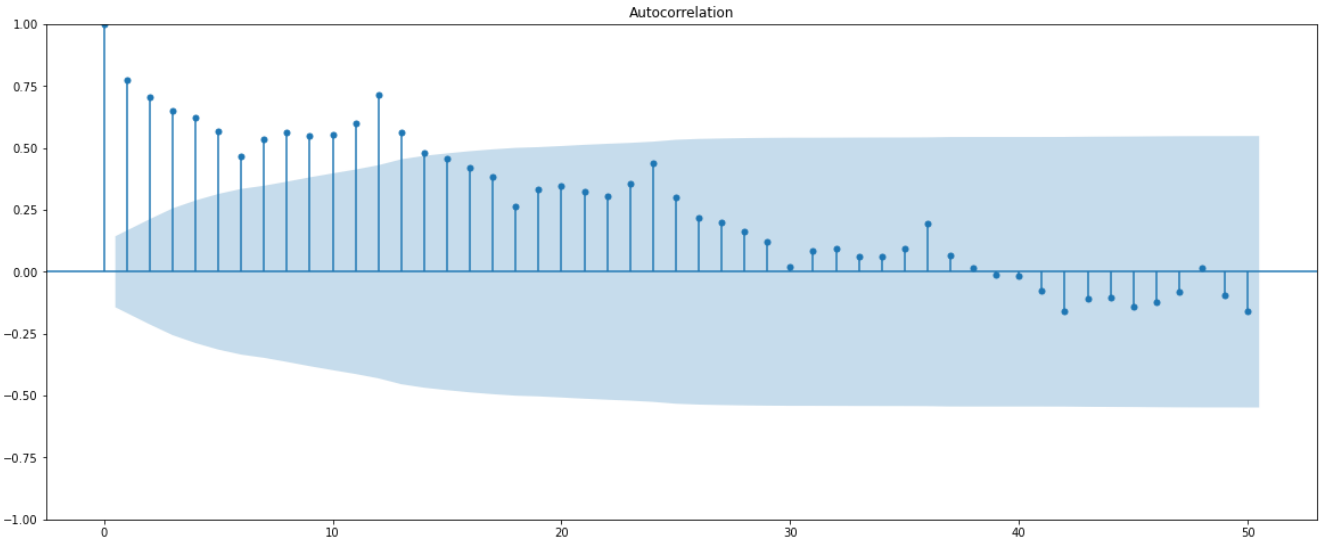
Let us take a difference of order 1 and check whether the Time Series is stationary or not.

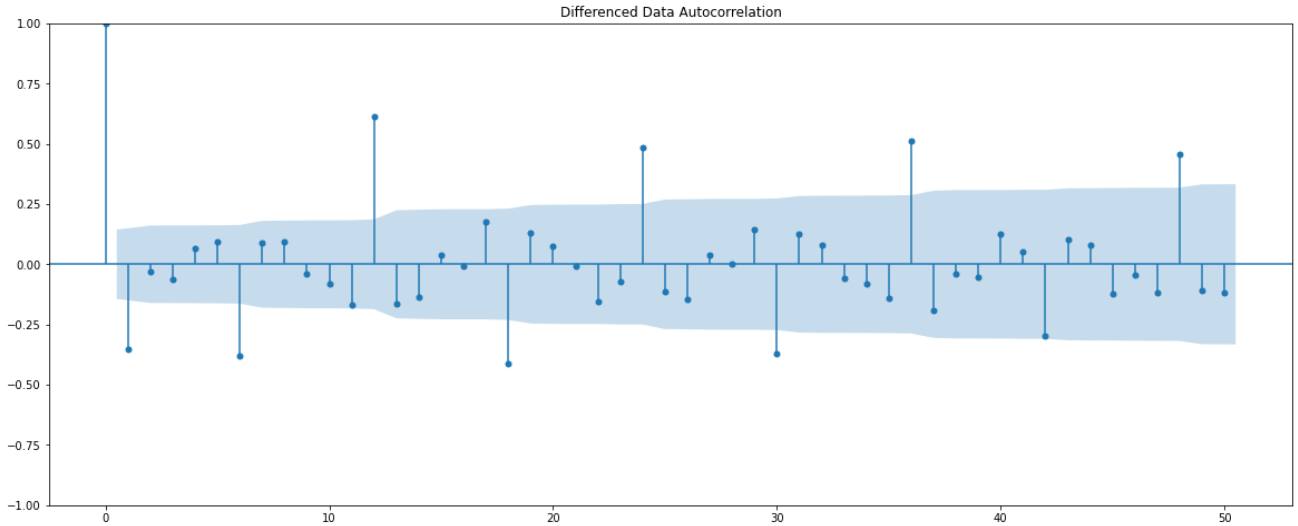


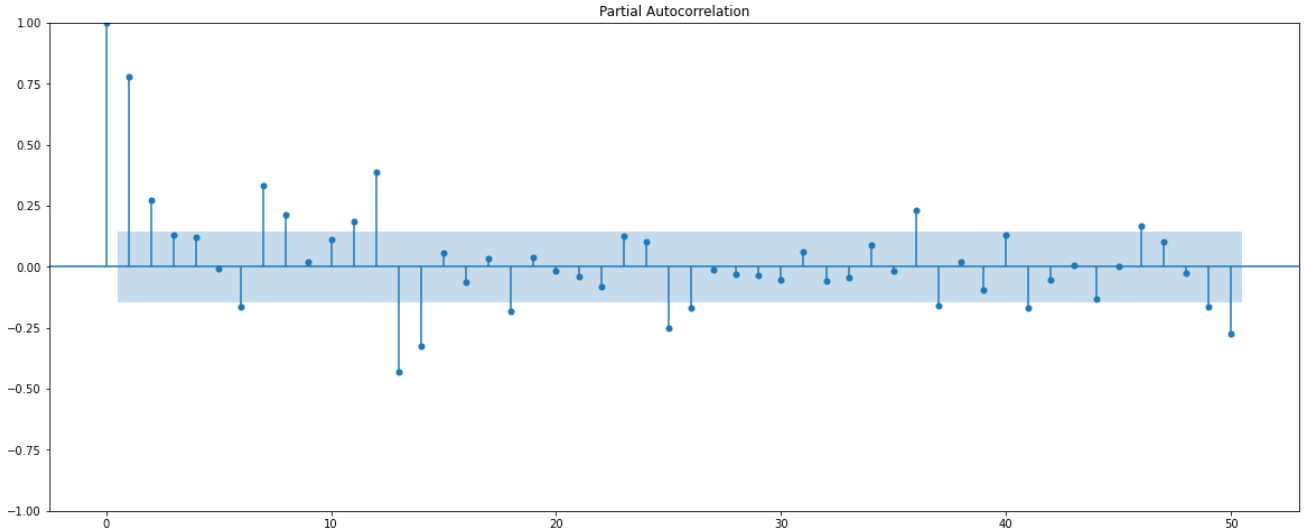

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.479160
p-value                          0.008539
#Lags Used                      12.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

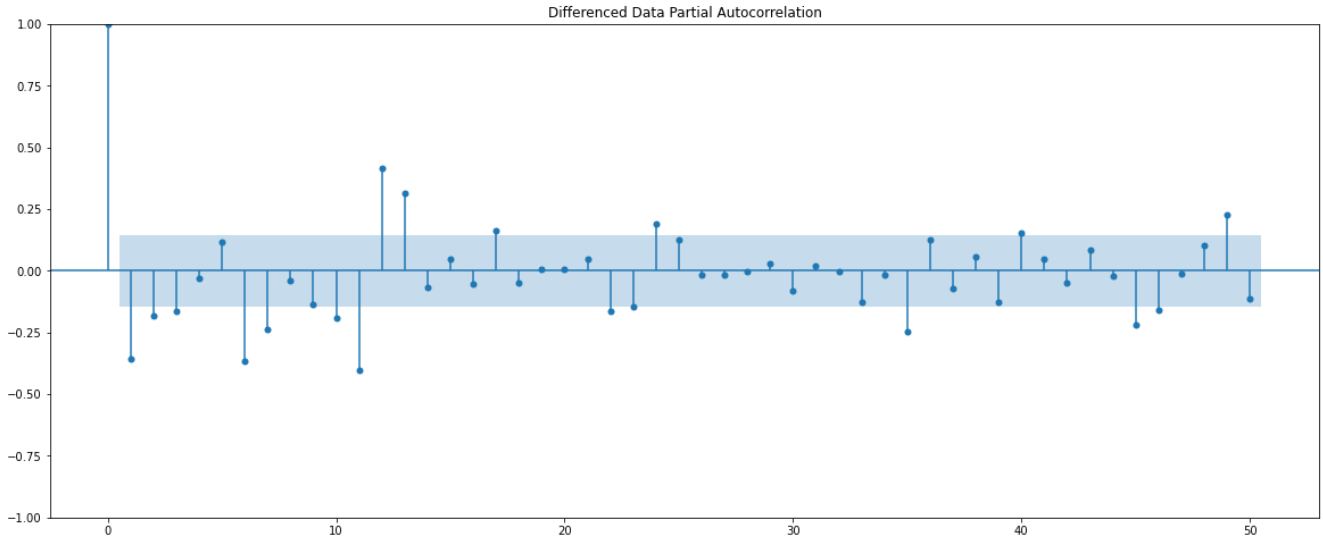
We see that at $\alpha = 0.05$ the Time Series is indeed stationary.

Plot the Autocorrelation and the Partial Autocorrelation function plots on the whole data



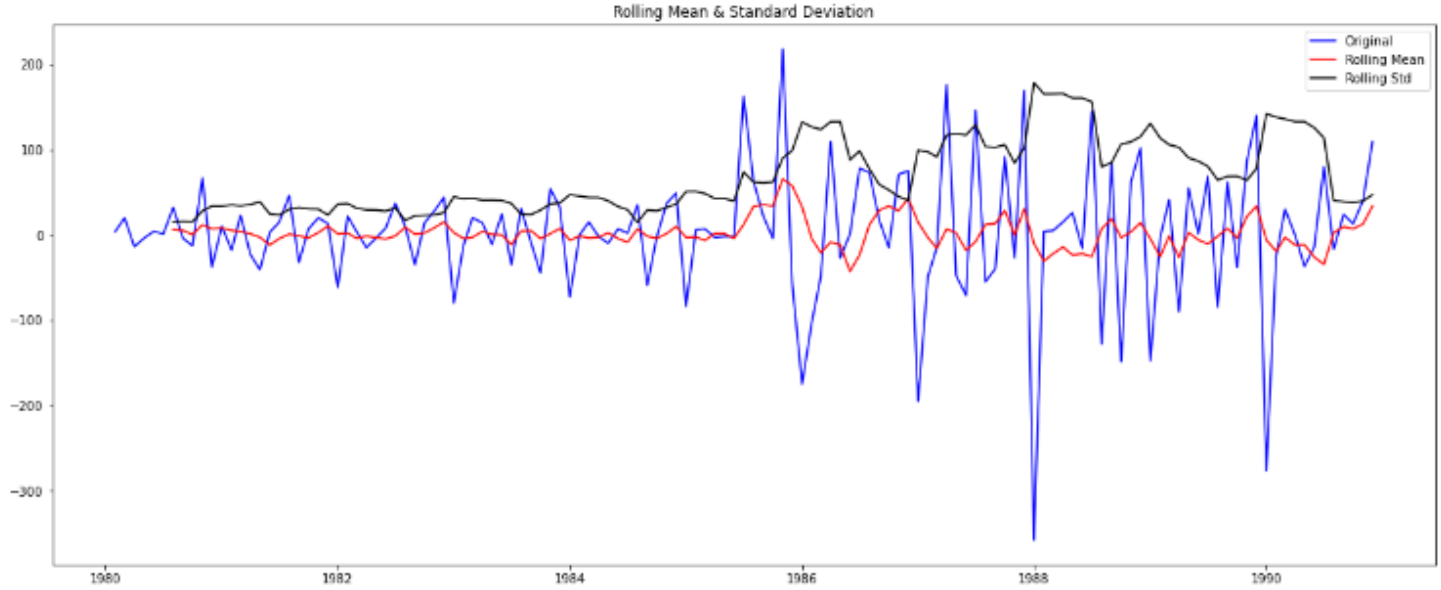






From the above plots, we can say that there seems to be a seasonality in the data.

Check for stationarity of the Training Data Time Series




```
Results of Dickey-Fuller Test:
Test Statistic                  -3.144211
p-value                          0.023450
#Lags Used                      13.000000
Number of Observations Used    117.000000
Critical Value (1%)             -3.487517
Critical Value (5%)             -2.886578
Critical Value (10%)            -2.580124
dtype: float64
```

We see that after taking a difference of order 1 the series have become stationary at α = 0.05. As p value is less than 0.05.

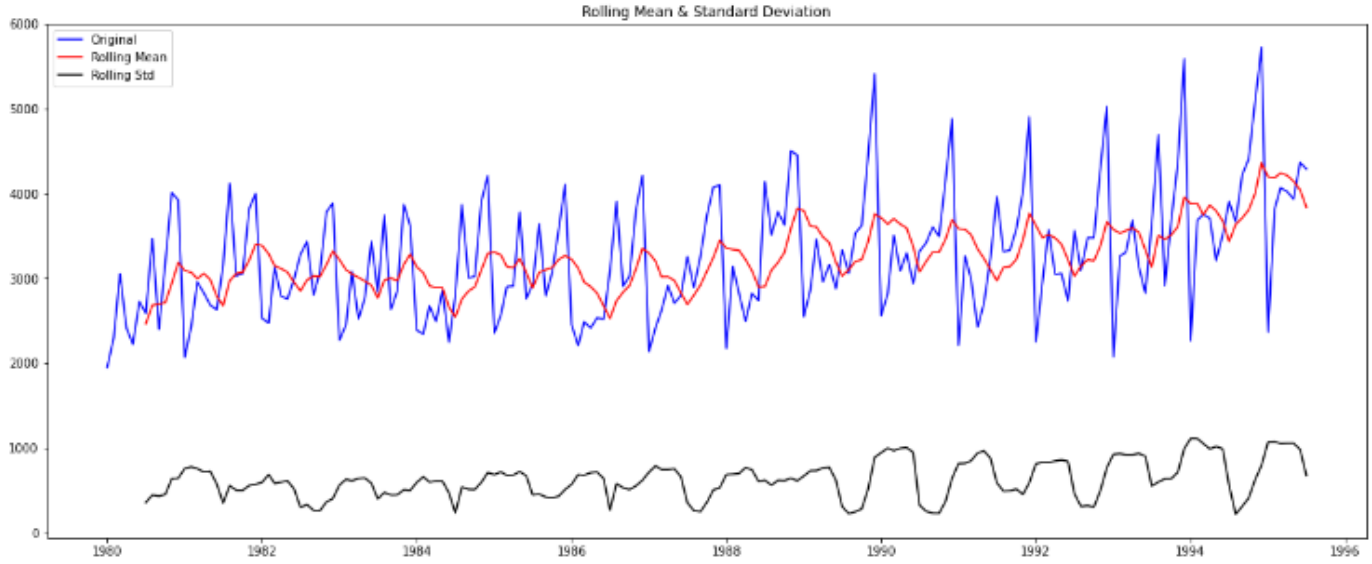
Soft Drink Data

Check for stationarity of the whole Time Series data.

Dicky Fuller Test

Null Hypothesis H0- Series is not Stationary

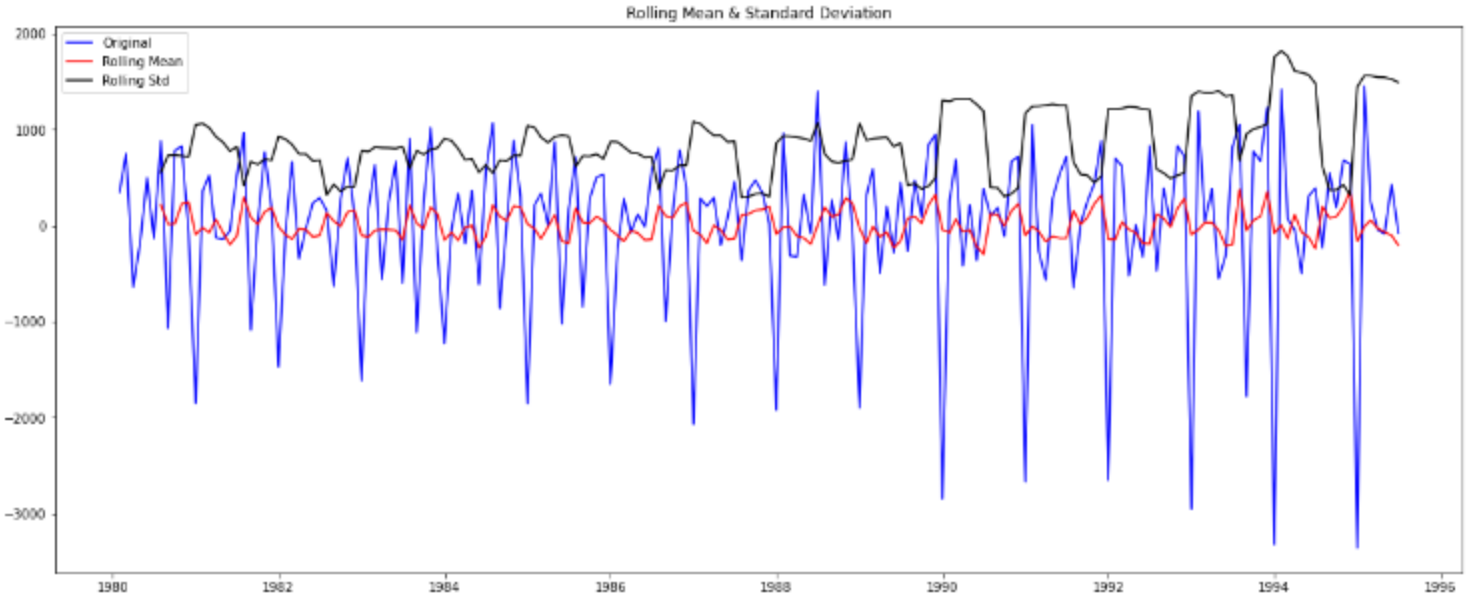
Alternative Hypothesis H1- Series is Stationary




```
Results of Dickey-Fuller Test:
Test Statistic                   1.098734
p-value                          0.995206
#Lags Used                      12.000000
Number of Observations Used    174.000000
Critical Value (1%)             -3.468502
Critical Value (5%)             -2.878298
Critical Value (10%)            -2.575704
dtype: float64
```

We see that at 5% significant level the Time Series is non-stationary.

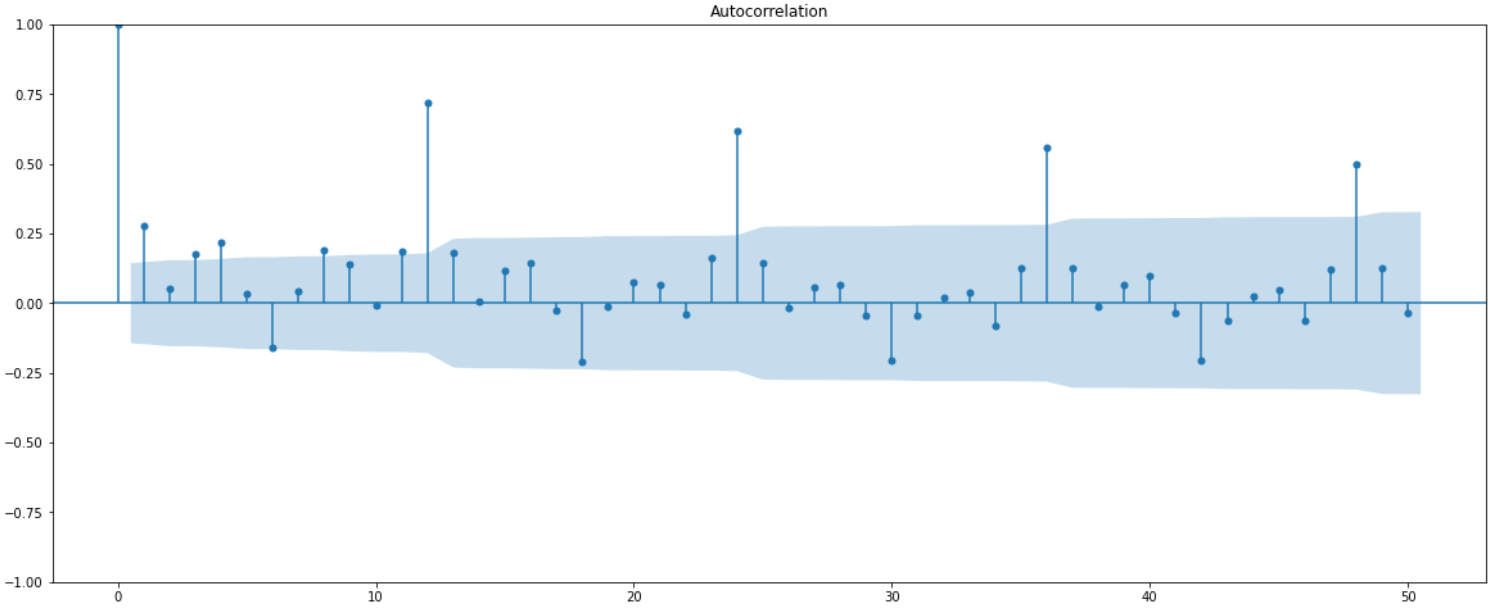
Let us take a difference of order 1 and check whether the Time Series is stationary or not.

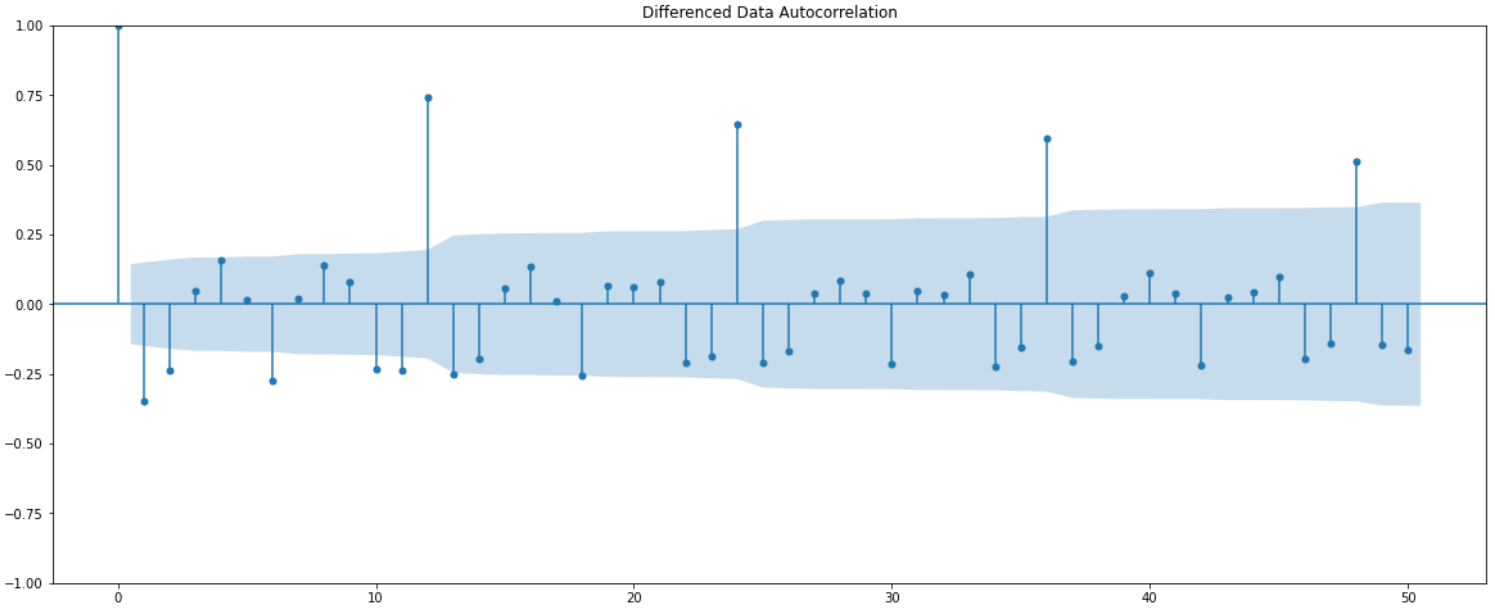


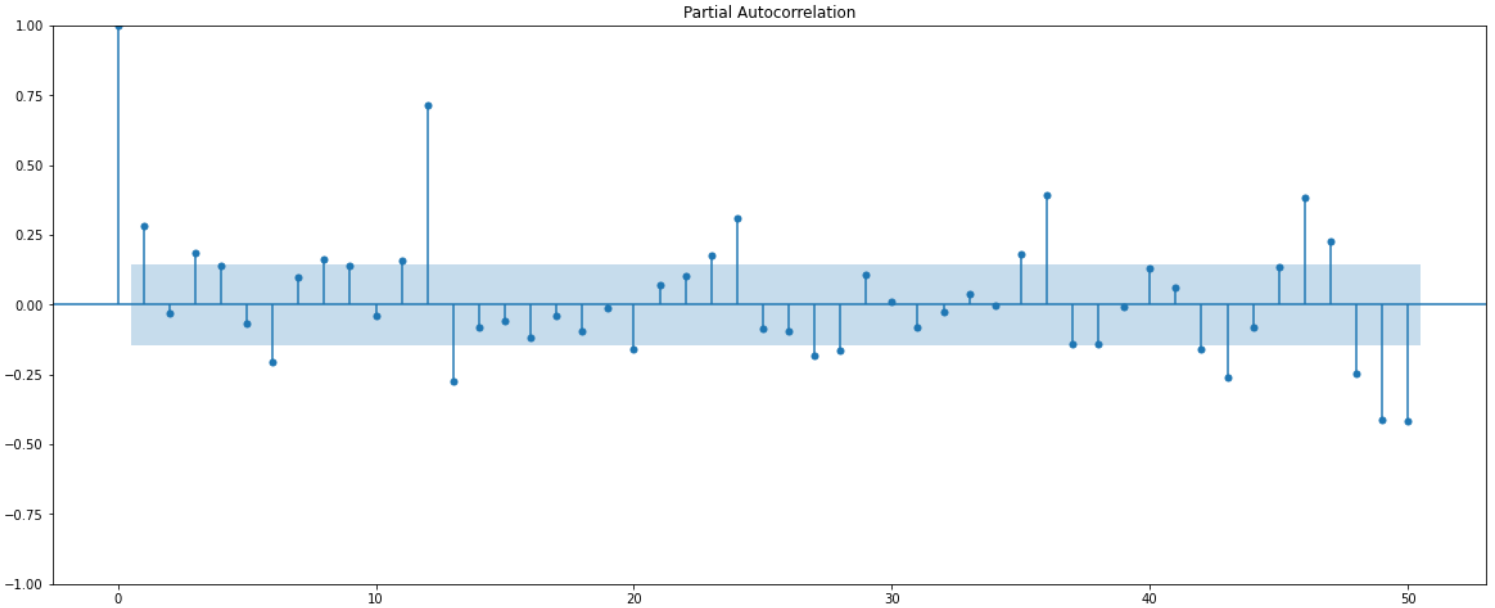

```
Results of Dickey-Fuller Test:
Test Statistic                -9.313527e+00
p-value                        1.033701e-15
#Lags Used                     1.100000e+01
Number of Observations Used    1.740000e+02
Critical Value (1%)           -3.468502e+00
Critical Value (5%)           -2.878298e+00
Critical Value (10%)          -2.575704e+00
dtype: float64
```

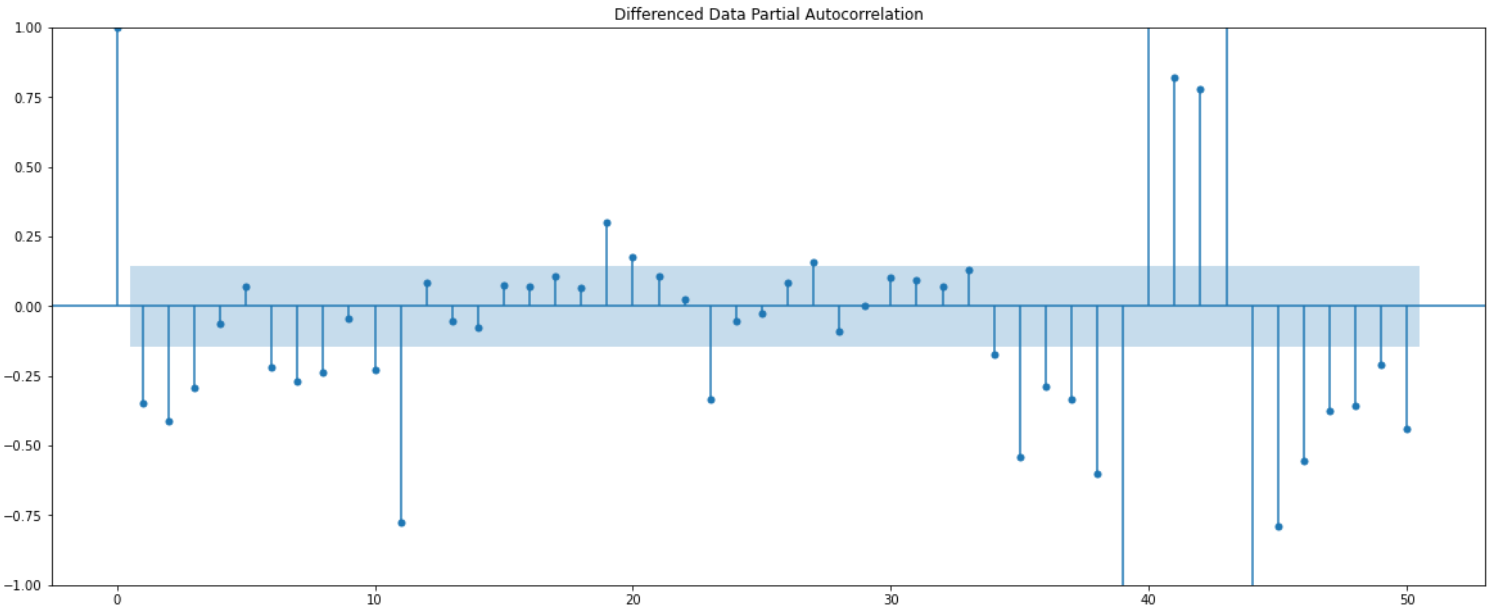
We see that at α = 0.05 the Time Series is indeed stationary.

Plot the Autocorrelation and the Partial Autocorrelation function plots on the whole data



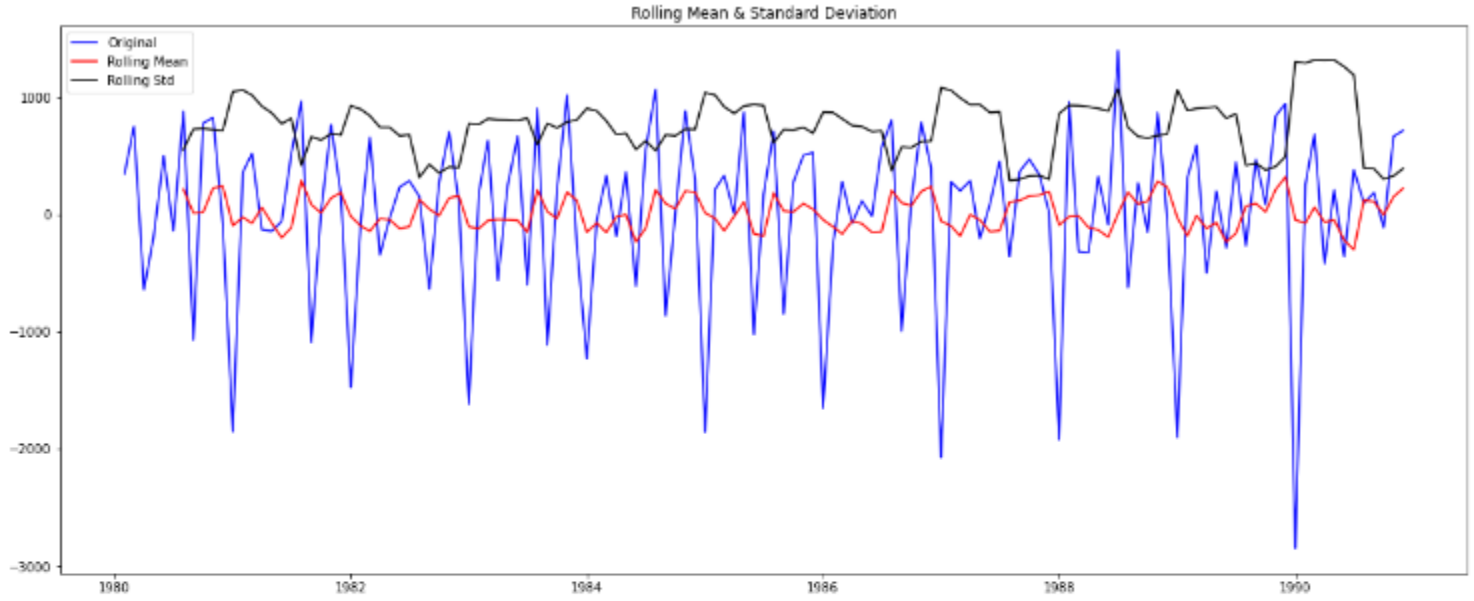

Differenced Data Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50


Partial Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50



From the above plots, we can say that there seems to be a seasonality in the data.

Check for stationarity of the Training Data Time Series




```
Results of Dickey-Fuller Test:
Test Statistic                -7.299886e+00
p-value                        1.347278e-10
#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 Value (10%)          -2.579896e+00
dtype: float64
```

We see that after taking a difference of order 1 the series have become stationary at α = 0.05. As p value is less than 0.05.

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.

ShoeSales Data

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

Sort the above AIC values in the ascending order to get the parameters for the minimum AIC value

	param	AIC
4	(1, 1, 1)	1492.487187
5	(1, 1, 2)	1494.423859
7	(2, 1, 1)	1494.431498
2	(0, 1, 2)	1494.964605
8	(2, 1, 2)	1496.410739
1	(0, 1, 1)	1497.050322
6	(2, 1, 0)	1498.950483
3	(1, 1, 0)	1501.643124
0	(0, 1, 0)	1508.283772

Apply the ARIMA Model

 # Top box - Gives the overall summary.

 # Middle box - Gives the coefficients and tells if the variables are significant.

 # Lower box(Roots) - If the model is stable and good for forecasting. Between Imaginary & Real you draw a circle and check if the values/roots are inside the circle.

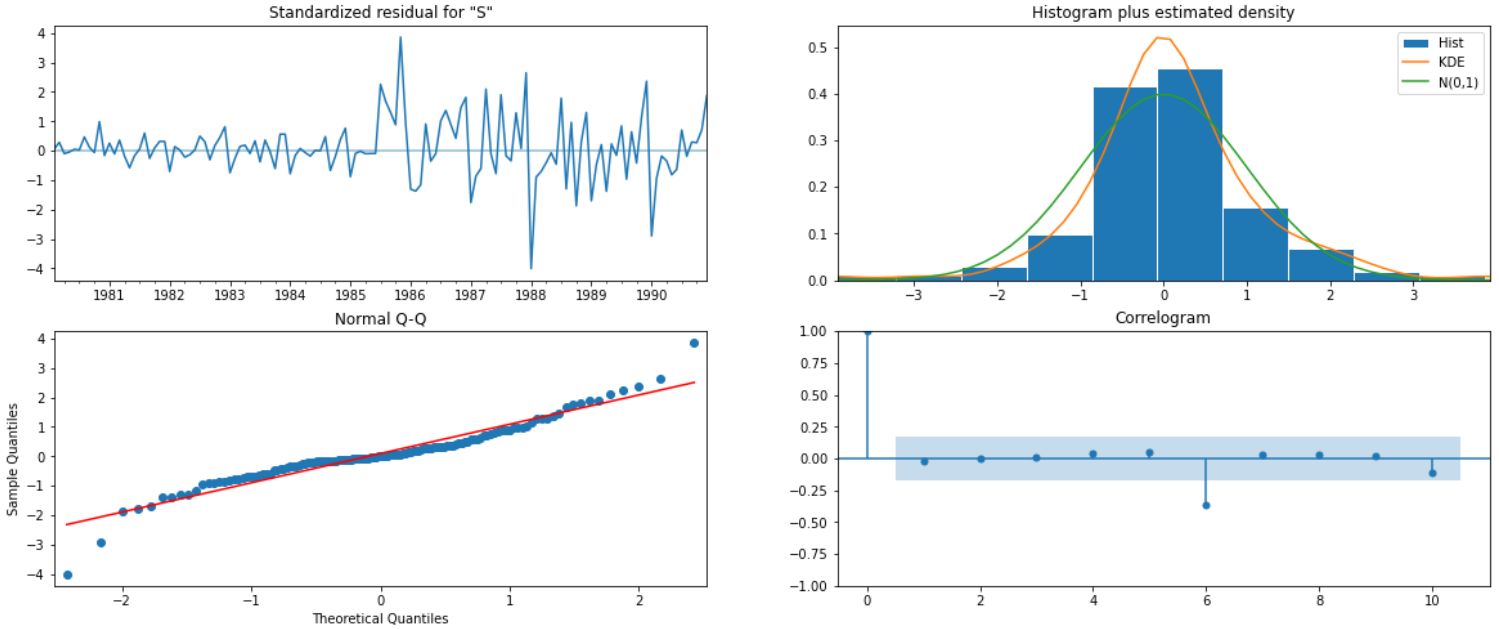
SARIMAX Results

Dep. Variable:	Shoe_Sales	No. Observations:	132
Model:	ARIMA(1, 1, 1)	Log Likelihood	-743.244
Date:	Sat, 01 Apr 2023	AIC	1492.487
Time:	05:22:06	BIC	1501.113
Sample:	01-01-1980 - 12-01-1990	HQIC	1495.992
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4699	0.111	4.235	0.000	0.252	0.687
ma.L1	-0.8347	0.068	-12.261	0.000	-0.968	-0.701
sigma2	4944.0868	405.583	12.190	0.000	4149.158	5739.015

Ljung-Box (L1) (Q):	0.05	Jarque-Bera (JB):	57.30
Prob(Q):	0.83	Prob(JB):	0.00
Heteroskedasticity (H):	12.81	Skew:	0.01
Prob(H) (two-sided):	0.00	Kurtosis:	6.24

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



Predict on the Test Set using this model and evaluate the model
RMSE : 142.82

	Test RMSE
RegressionOnTime	266.276472
NaiveModel	245.121306
SimpleAverageModel	63.984570
2pointTrailingMovingAverage	45.948736
4pointTrailingMovingAverage	57.872686
6pointTrailingMovingAverage	63.456893
9pointTrailingMovingAverage	67.723648
Alpha=0.6,SimpleExponentialSmoothing	196.404850
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	890.968504
Alpha=0.57,Beta=0.00014,Gamma=0.202,TripleExponentialSmoothing	83.734048
Alpha=0.1,Beta=0.6,Gamma=0.2,TripleExponentialSmoothing	41.237522
ARIMA(1,1,1)	142.820730

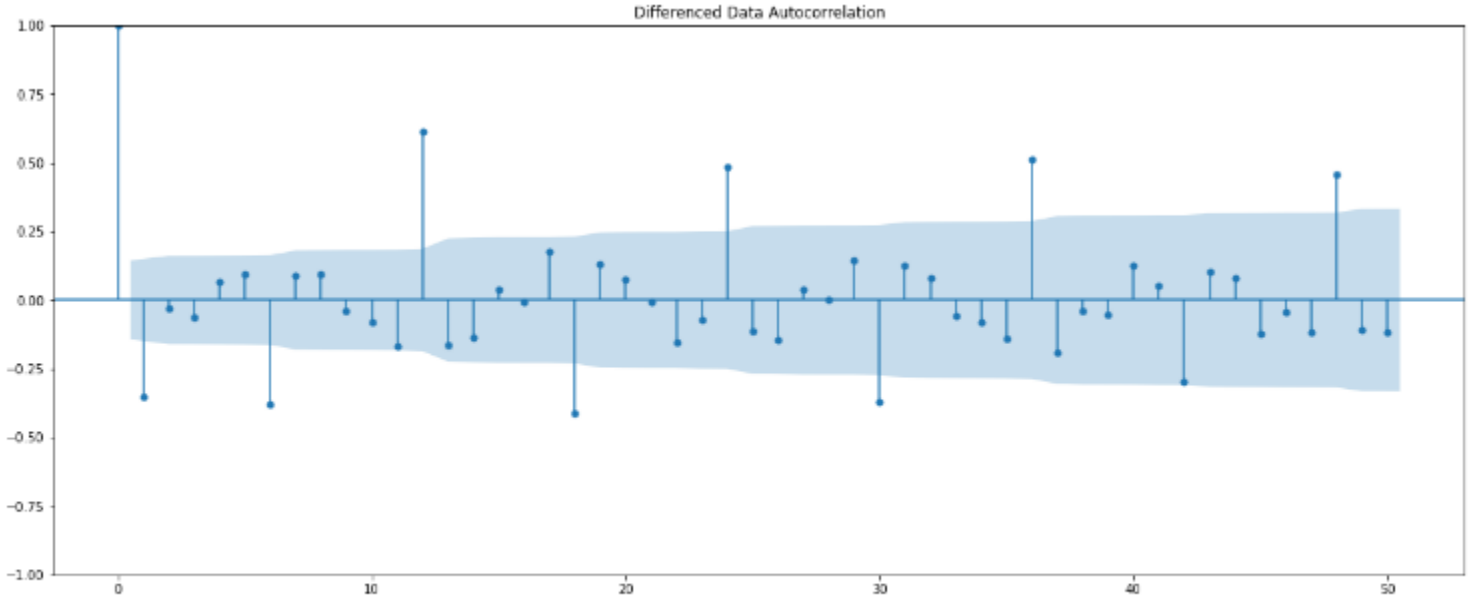
Ran the automated model getting a combination of different parameters of p and q in the range of 0 and 2. We have kept the value of d as 1 as we need to take a difference of the series to make it stationary.

Sort the AIC values in the ascending order to get the parameters for the minimum AIC value

The ARIMA model(1,1,1) is with the lowest AIC value and the test RMSE for the model is 142.82.

Build an Automated version of a SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).

Let us look at the ACF plot once more to understand the seasonal parameter for the SARIMA model.



We see that there can be a seasonality of 6 as well as 12. We will run our auto SARIMA models by setting seasonality both as 6 and 12.

Setting the seasonality as 6 for the first iteration of the auto SARIMA model

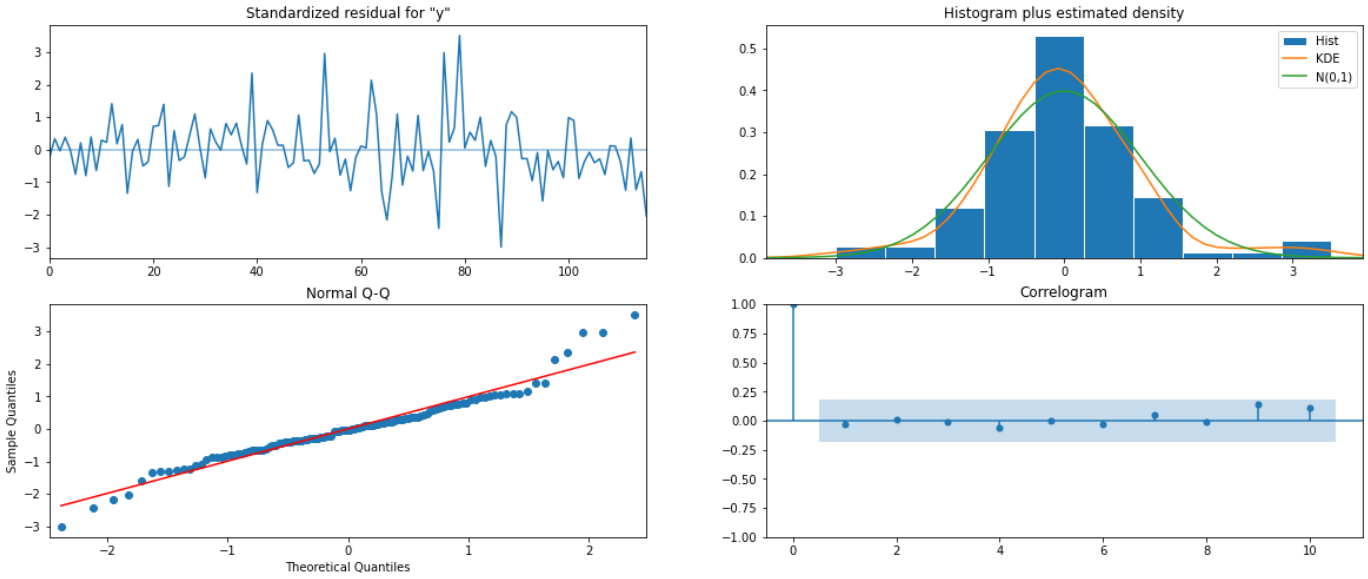
```
Examples of some parameter combinations for Model.
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
```

Sort the above AIC values in the ascending order to get the parameters for the minimum AIC value

	param	seasonal	AIC
80	(2, 1, 2)	(2, 0, 2, 6)	1280.778864
26	(0, 1, 2)	(2, 0, 2, 6)	1281.026602
53	(1, 1, 2)	(2, 0, 2, 6)	1282.065372
17	(0, 1, 1)	(2, 0, 2, 6)	1288.975863
50	(1, 1, 2)	(1, 0, 2, 6)	1289.791748

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(2, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -631.389
Date:                           Sat, 01 Apr 2023   AIC                           1280.779
Time:                                   06:28:59   BIC                           1305.561
Sample:                                        0   HQIC                          1290.839
                                             - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0600      0.474      0.127      0.899      -0.869       0.989
ar.L2          0.3977      0.161      2.464      0.014       0.081       0.714
ma.L1         -0.4650      0.492     -0.946      0.344      -1.428       0.498
ma.L2         -0.3238      0.295     -1.096      0.273      -0.903       0.255
ar.S.L6       -0.1732      0.136     -1.278      0.201      -0.439       0.092
ar.S.L12       0.7914      0.130      6.103      0.000       0.537       1.046
ma.S.L6        0.1151      0.184      0.626      0.531      -0.245       0.475
ma.S.L12      -0.3010      0.179     -1.683      0.092      -0.652       0.050
sigma2      3080.4669    348.197      8.847      0.000    2398.013    3762.921
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                21.55
Prob(Q):                              0.92   Prob(JB):                         0.00
Heteroskedasticity (H):               7.76   Skew:                             0.18
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.08
===================================================================================

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



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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	257.242892	55.501954	148.461061	366.024723
1	257.174888	64.584820	130.590967	383.758810
2	265.714452	73.831864	121.006658	410.422247
3	263.007366	78.615069	108.924662	417.090070
4	240.046301	83.309369	76.762937	403.329664

RMSE : 57.03

	Test RMSE
RegressionOnTime	266.276472
NaiveModel	245.121306
SimpleAverageModel	63.984570
2pointTrailingMovingAverage	45.948736
4pointTrailingMovingAverage	57.872686
6pointTrailingMovingAverage	63.456893
9pointTrailingMovingAverage	67.723648
Alpha=0.6,SimpleExponentialSmoothing	196.404850
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	890.968504
Alpha=0.57,Beta=0.00014,Gamma=0.202,TripleExponentialSmoothing	83.734048
Alpha=0.1,Beta=0.6,Gamma=0.2,TripleExponentialSmoothing	41.237522
ARIMA(1,1,1)	142.820730
SARIMA(2,1,2)(2,0,2,6)	57.030694

We can see that the RMSE value has decreased on the test data by including the seasonal parameters as well.

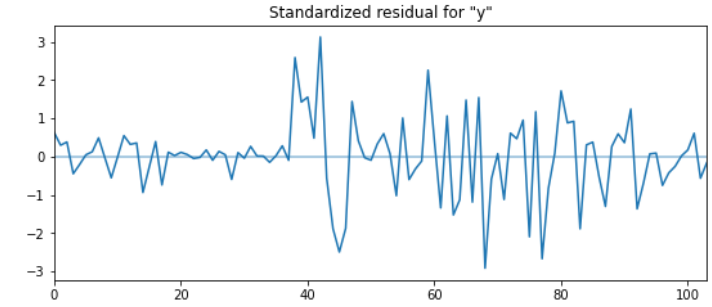
Setting the seasonality as 12 for the second iteration of the auto SARIMA model.

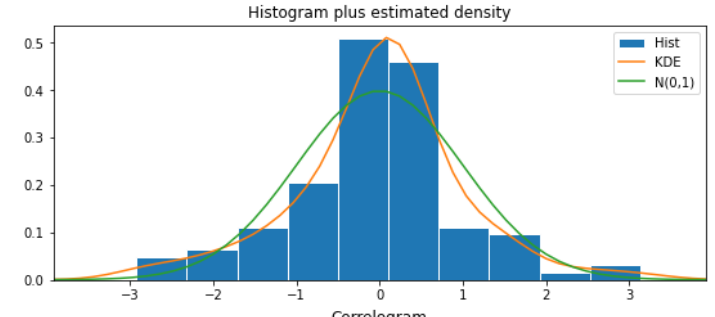
```
Examples of some parameter combinations for Model
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

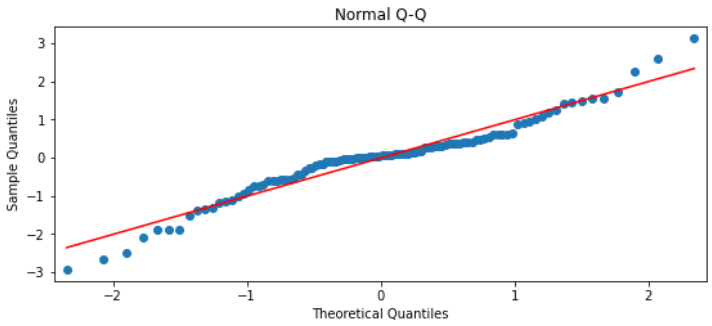
	param	seasonal	AIC
23	(0, 1, 2)	(1, 0, 2, 12)	1156.165429
50	(1, 1, 2)	(1, 0, 2, 12)	1157.082589
26	(0, 1, 2)	(2, 0, 2, 12)	1157.772313
77	(2, 1, 2)	(1, 0, 2, 12)	1158.491006
80	(2, 1, 2)	(2, 0, 2, 12)	1158.630324

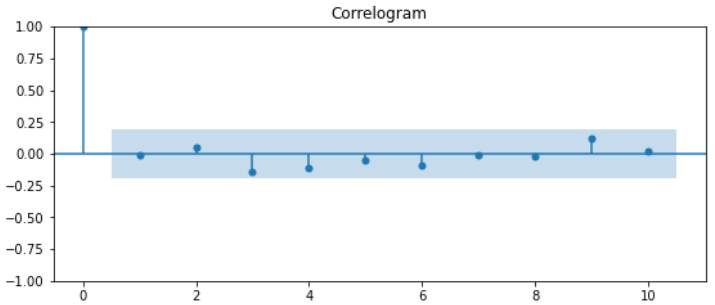
```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(1, 0, 2, 12)   Log Likelihood                -572.083
Date:                            Sat, 01 Apr 2023   AIC                           1156.165
Time:                                    06:31:30   BIC                           1172.032
Sample:                                         0   HQIC                          1162.593
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.3742      0.081     -4.632      0.000      -0.533      -0.216
ma.L2          0.0616      0.077      0.803      0.422      -0.089       0.212
ar.S.L12       1.0635      0.051     21.042      0.000       0.964       1.163
ma.S.L12      -0.7636    291.960     -0.003      0.998    -572.994     571.467
ma.S.L24      -0.2365     68.957     -0.003      0.997    -135.390     134.917
sigma2      2818.0938   8.23e+05      0.003      0.997   -1.61e+06    1.62e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 8.60
Prob(Q):                              0.91   Prob(JB):                         0.01
Heteroskedasticity (H):               7.92   Skew:                            -0.15
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.38
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 8.63e+14. Standard errors may be unstable.
```









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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	225.292059	55.845596	115.836701	334.747416
1	199.179251	65.878234	70.060284	328.298218
2	213.554159	76.229342	64.147394	362.960924
3	238.613122	85.270200	71.486601	405.739642
4	215.720842	93.440542	32.580745	398.860939

RMSE : 69.03

	Test RMSE
RegressionOnTime	266.276472
NaiveModel	245.121306
SimpleAverageModel	63.984570
2pointTrailingMovingAverage	45.948736
4pointTrailingMovingAverage	57.872686
6pointTrailingMovingAverage	63.456893
9pointTrailingMovingAverage	67.723648
Alpha=0.6,SimpleExponentialSmoothing	196.404850
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	890.968504
Alpha=0.57,Beta=0.00014,Gamma=0.202,TripleExponentialSmoothing	83.734048
Alpha=0.1,Beta=0.6,Gamma=0.2,TripleExponentialSmoothing	41.237522
ARIMA(1,1,1)	142.820730
SARIMA(2,1,2)(2,0,2,6)	57.030694
SARIMA(0,1,2)(1,0,2,12)	69.030660

RMSE value has increased to by 69 taking seasonality as 12.

Soft Drink Data

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

Sort the above AIC values in the ascending order to get the parameters for the minimum AIC value

	param	AIC
2	(0, 1, 2)	2056.489263
5	(1, 1, 2)	2056.715682
8	(2, 1, 2)	2058.712702
7	(2, 1, 1)	2059.100672
4	(1, 1, 1)	2061.523084
1	(0, 1, 1)	2069.59963
6	(2, 1, 0)	2073.234861
3	(1, 1, 0)	2097.872122
0	(0, 1, 0)	2103.733834

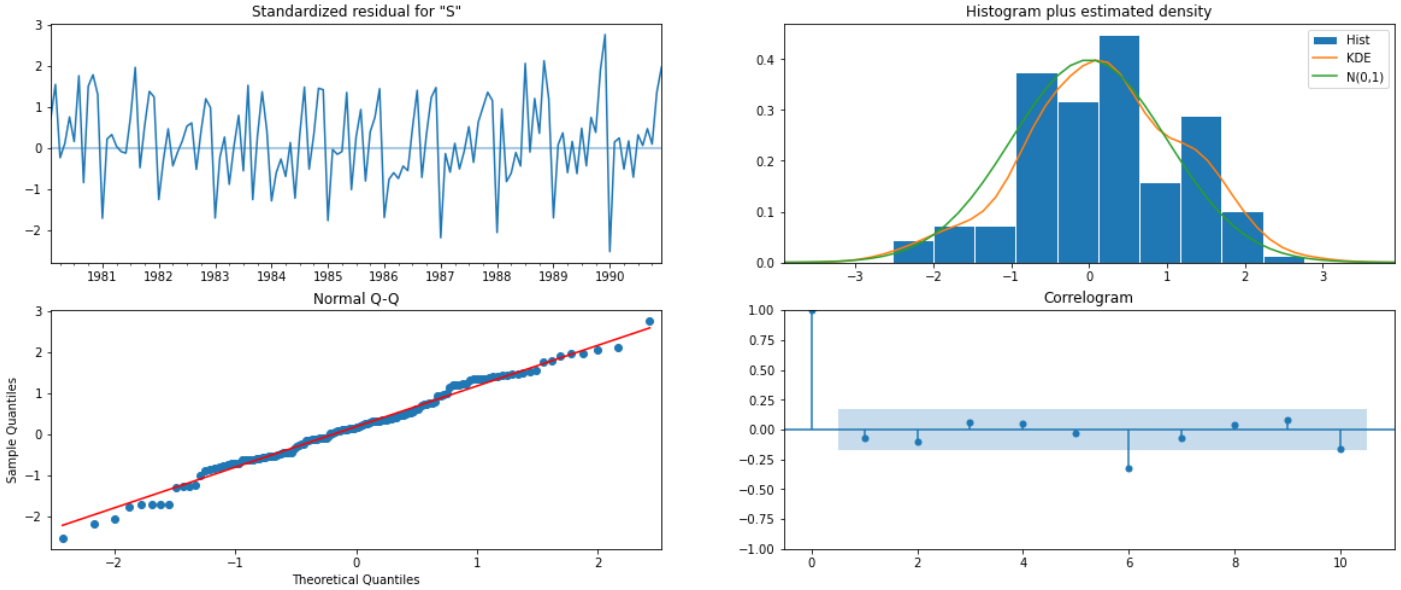
SARIMAX Results

Dep. Variable:	SoftDrinkProduction	No. Observations:	132
Model:	ARIMA(0, 1, 2)	Log Likelihood	-1025.245
Date:	Sat, 01 Apr 2023	AIC	2056.489
Time:	06:35:02	BIC	2065.115
Sample:	01-01-1980 - 12-01-1990	HQIC	2059.994
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.5407	0.085	-6.392	0.000	-0.707	-0.375
ma.L2	-0.3913	0.113	-3.475	0.001	-0.612	-0.171
sigma2	3.572e+05	4.62e+04	7.725	0.000	2.67e+05	4.48e+05

Ljung-Box (L1) (Q):	0.61	Jarque-Bera (JB):	0.39
Prob(Q):	0.44	Prob(JB):	0.82
Heteroskedasticity (H):	1.31	Skew:	-0.13
Prob(H) (two-sided):	0.37	Kurtosis:	2.91

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



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

RMSE : 831.61

	Test RMSE
RegressionOnTime	3202.844447
NaiveModel	1519.259233
SimpleAverageModel	934.353358
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827
Alpha=0.2,SimpleExponentialSmoothing	847.635259
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	982.938364
Alpha=0.11,Beta=0.04,Gamma=0.23,TripleExponentialSmoothing	447.722581
Alpha=0.5,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	442.214551
ARIMA(0,1,2)	831.615852

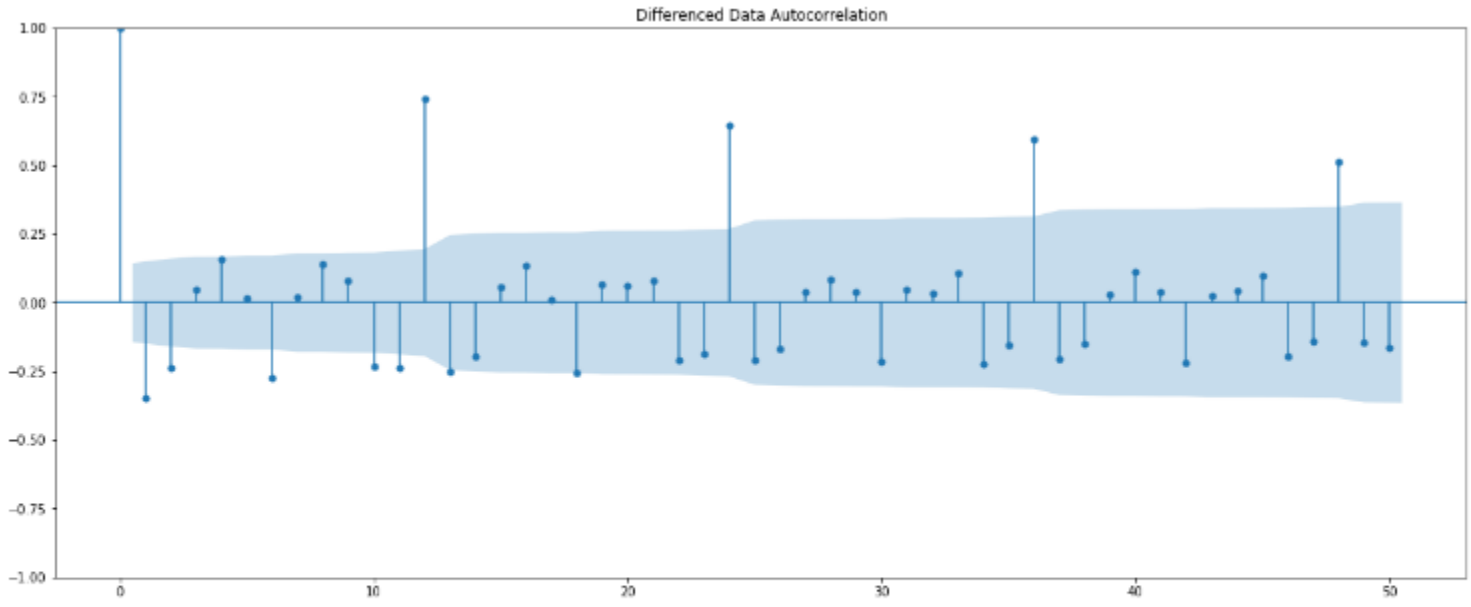
Run the automated model getting a combination of different parameters of p and q in the range of 0 and 2. We have kept the value of d as 1 as we need to take a difference of the series to make it stationary.

Sort the AIC values in the ascending order to get the parameters for the minimum AIC value

The ARIMA model(0,1,2) with the lowest AIC values and the test RMSE for the value is 831.6.

Build an Automated version of a SARIMA model for which the best parameters are selected in accordance with the lowest Akaike Information Criteria (AIC).

Let us look at the ACF plot once more to understand the seasonal parameter for the SARIMA model.



We see that there can be a seasonality of 6 as well as 12. We will run our auto SARIMA models by setting seasonality both as 6 and 12.

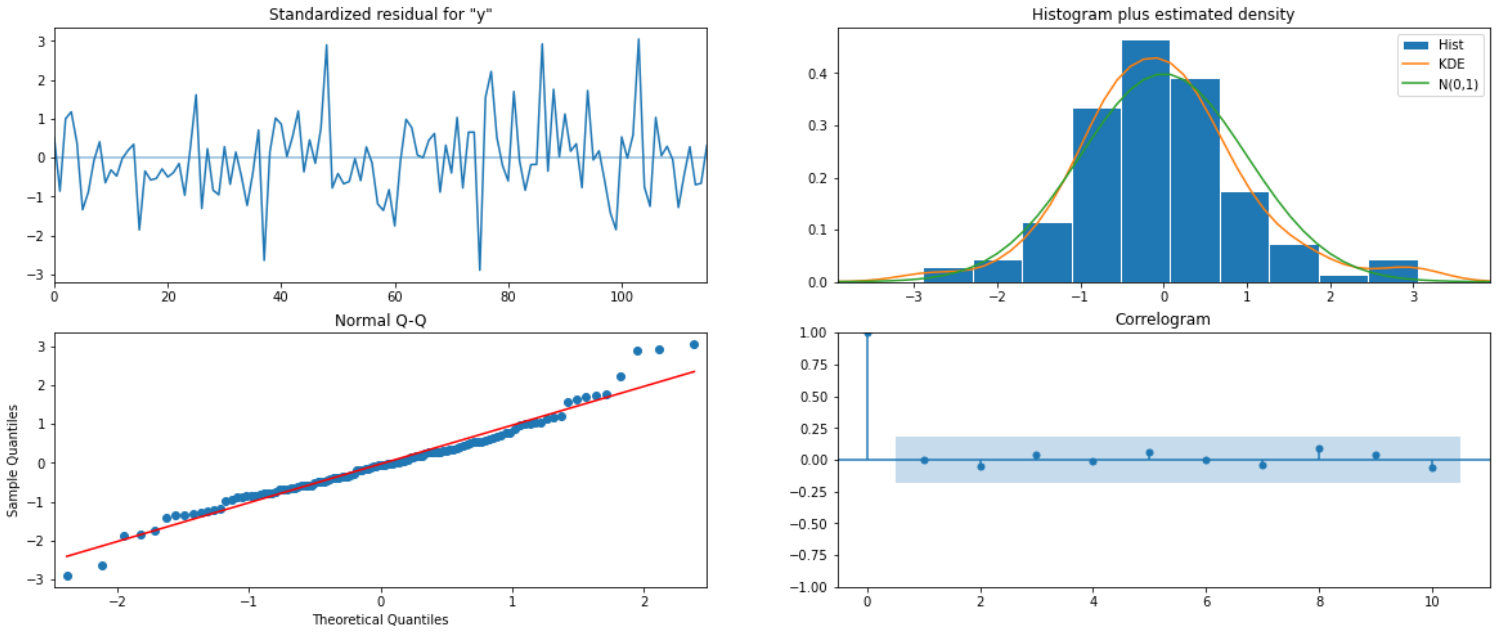
Setting the seasonality as 6 for the first iteration of the auto SARIMA model.

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

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 6)	1686.172022
53	(1, 1, 2)	(2, 0, 2, 6)	1688.105594
80	(2, 1, 2)	(2, 0, 2, 6)	1689.372223
17	(0, 1, 1)	(2, 0, 2, 6)	1698.846967
44	(1, 1, 1)	(2, 0, 2, 6)	1700.331864

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(2, 0, 2, 6)   Log Likelihood                -836.086
Date:                           Sat, 01 Apr 2023   AIC                            1686.172
Time:                                   06:39:15   BIC                            1705.447
Sample:                                        0   HQIC                           1693.997
                                             - 132
Covariance Type:                             opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7796      0.116     -6.725      0.000      -1.007      -0.552
ma.L2         -0.0866      0.094     -0.926      0.354      -0.270       0.097
ar.S.L6        0.0053      0.023      0.228      0.820      -0.040       0.051
ar.S.L12       0.9803      0.029     33.262      0.000       0.922       1.038
ma.S.L6       -0.1663      0.113     -1.469      0.142      -0.388       0.056
ma.S.L12      -0.6096      0.098     -6.203      0.000      -0.802      -0.417
sigma2      1.012e+05   1.11e+04      9.126      0.000    7.94e+04    1.23e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                11.94
Prob(Q):                              0.96   Prob(JB):                         0.00
Heteroskedasticity (H):               1.67   Skew:                             0.41
Prob(H) (two-sided):                  0.12   Kurtosis:                         4.34
===================================================================================

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



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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	2716.673533	318.171394	2093.069061	3340.278006
1	3110.199797	325.815150	2471.613838	3748.785756
2	3344.806208	328.583954	2700.793493	3988.818924
3	3103.976767	331.329063	2454.583737	3753.369797
4	3290.403888	334.052398	2635.673219	3945.134558

RMSE : 447.94

	Test RMSE
RegressionOnTime	3202.844447
NaiveModel	1519.259233
SimpleAverageModel	934.353358
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827
Alpha=0.2,SimpleExponentialSmoothing	847.635259
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	982.938364
Alpha=0.11,Beta=0.04,Gamma=0.23,TripleExponentialSmoothing	447.722581
Alpha=0.5,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	442.214551
ARIMA(0,1,2)	831.615852
SARIMA(0,1,2)(2,0,2,6)	447.942606

We can see that the RMSE value has decreased on the test data by including the seasonal parameters as well.

Setting the seasonality as 12 for the second iteration of the auto SARIMA model.

```
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)
```

	param	seasonal	AIC
26	(0, 1, 2)	(2, 0, 2, 12)	1517.207903
23	(0, 1, 2)	(1, 0, 2, 12)	1518.229381
53	(1, 1, 2)	(2, 0, 2, 12)	1518.328976
50	(1, 1, 2)	(1, 0, 2, 12)	1519.197016
80	(2, 1, 2)	(2, 0, 2, 12)	1520.313657

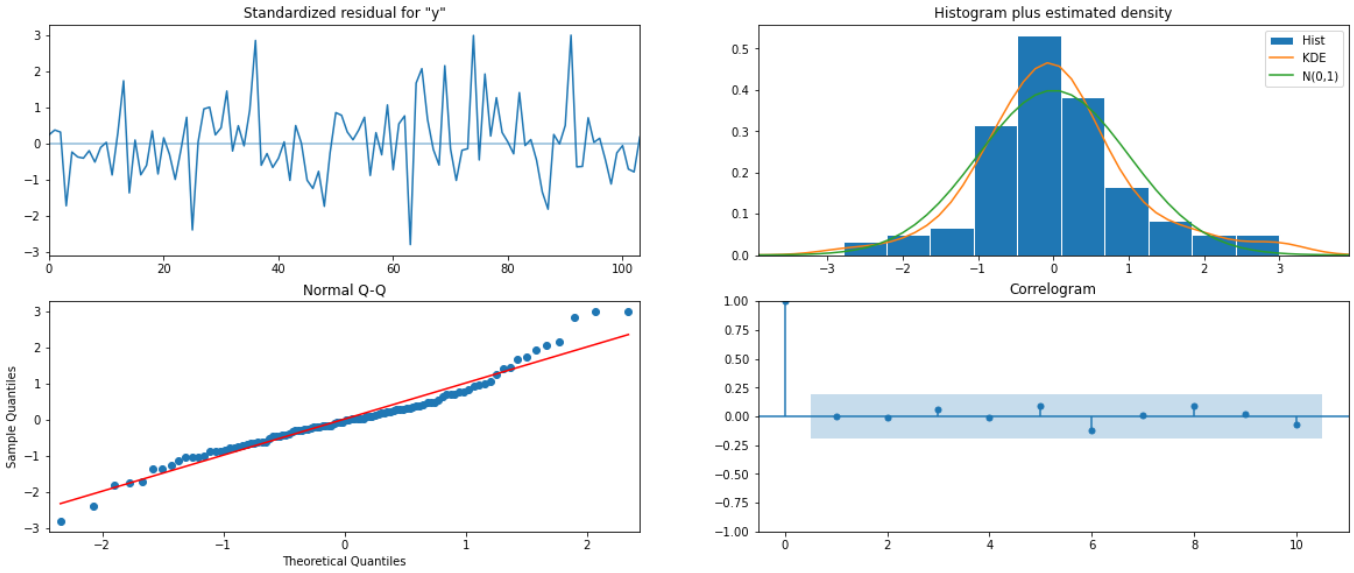
SARIMAX Results

Dep. Variable:	y	No. Observations:	132
Model:	SARIMAX(0, 1, 2)x(2, 0, 2, 12)	Log Likelihood	-751.604
Date:	Sat, 01 Apr 2023	AIC	1517.208
Time:	06:42:40	BIC	1535.719
Sample:	0	HQIC	1524.707
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.9981	0.141	-7.092	0.000	-1.274	-0.722
ma.L2	-0.1064	0.121	-0.879	0.379	-0.344	0.131
ar.S.L12	0.6097	0.437	1.396	0.163	-0.246	1.466
ar.S.L24	0.3745	0.438	0.855	0.392	-0.484	1.233
ma.S.L12	-0.2179	0.438	-0.498	0.619	-1.076	0.640
ma.S.L24	-0.2300	0.273	-0.841	0.400	-0.766	0.306
sigma2	9.002e+04	1.8e+04	4.993	0.000	5.47e+04	1.25e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	12.45
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	1.74	Skew:	0.49
Prob(H) (two-sided):	0.10	Kurtosis:	4.39

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



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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	2657.024987	328.613851	2012.953674	3301.096299
1	2979.306649	334.136589	2324.410969	3634.202328
2	3404.286347	335.603086	2746.516385	4062.056309
3	3086.458765	337.063859	2425.825741	3747.091788
4	3284.532830	338.518355	2621.049047	3948.016613

RMSE: 437.70

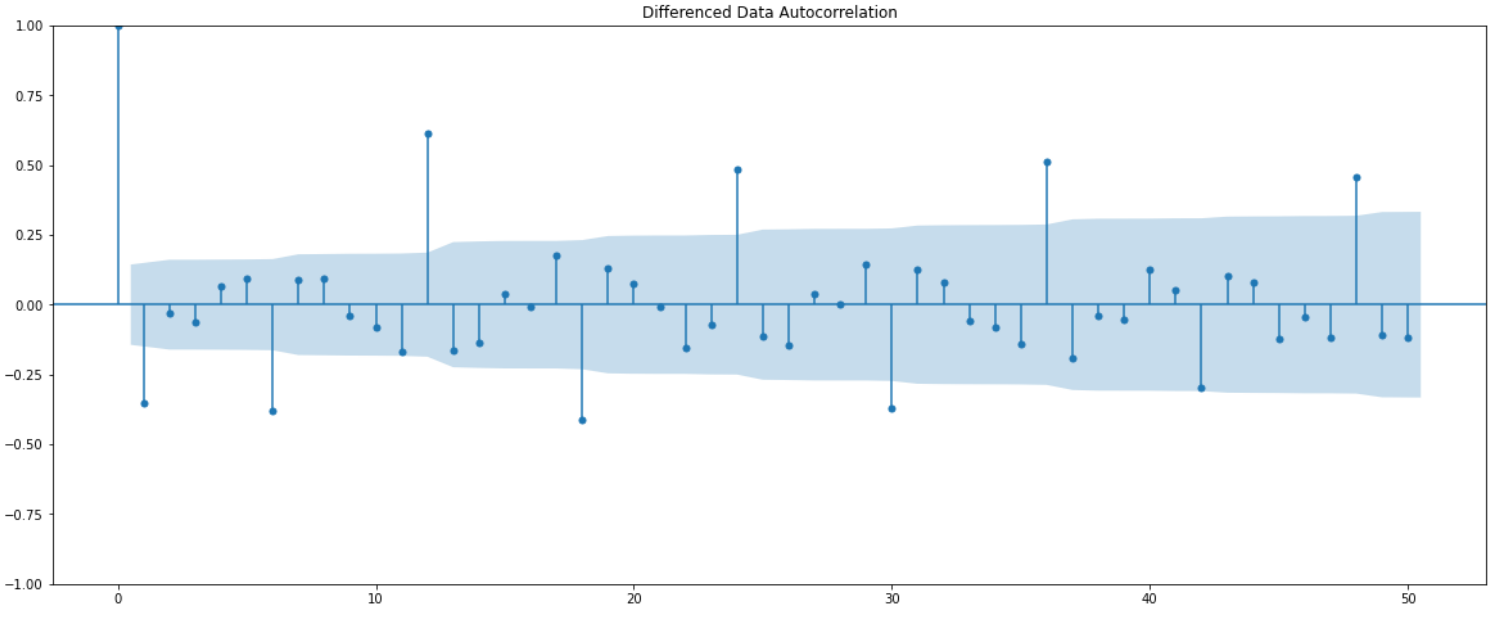
	Test RMSE
RegressionOnTime	3202.844447
NaiveModel	1519.259233
SimpleAverageModel	934.353358
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827
Alpha=0.2,SimpleExponentialSmoothing	847.635259
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	982.938364
Alpha=0.11,Beta=0.04,Gamma=0.23,TripleExponentialSmoothing	447.722581
Alpha=0.5,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	442.214551
ARIMA(0,1,2)	831.615852
SARIMA(0,1,2)(2,0,2,6)	447.942606
SARIMA(0,1,2)(2,0,2,12)	437.706534

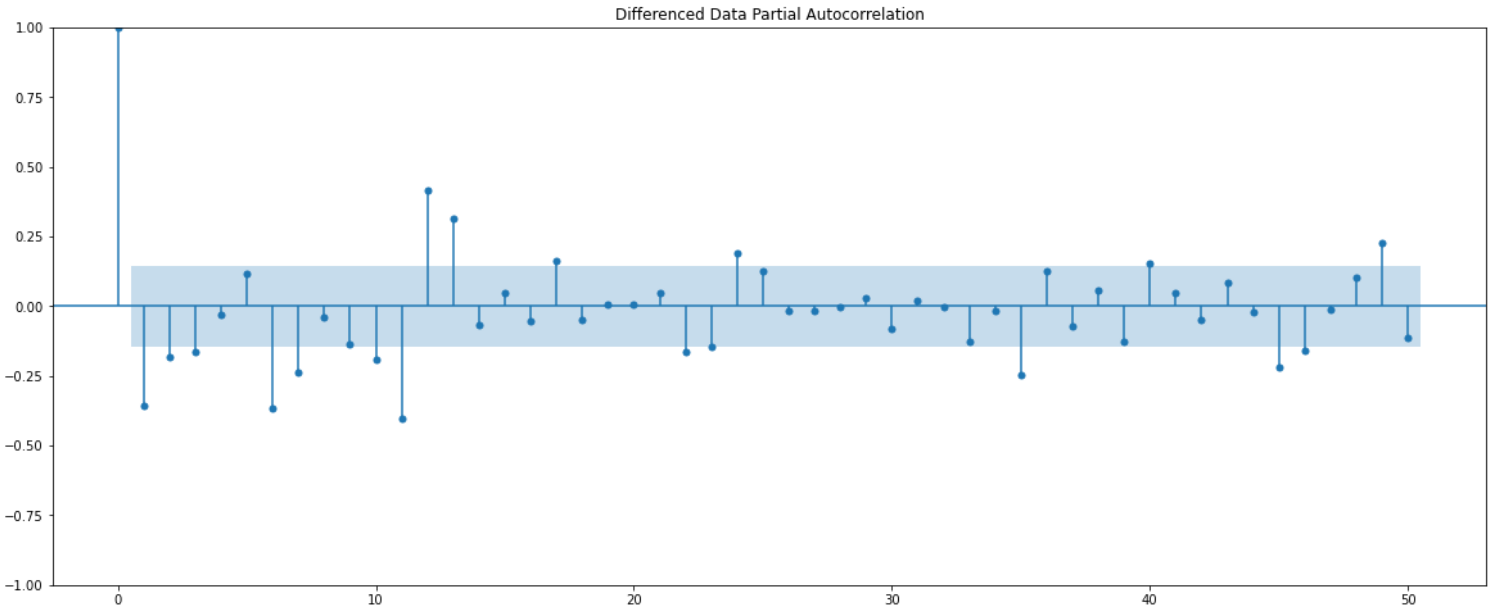
RMSE value has reduced further down to 437 by taking seasonality as 12.

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.

ShoeSales Data

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





Here, we have taken alpha=0.05.

By looking at the above plots, we will take the value of p and q to be 3 and 1 respectively.

SARIMAX Results

Dep. Variable:	Shoe_Sales	No. Observations:	132
Model:	ARIMA(3, 1, 1)	Log Likelihood	-743.173
Date:	Sat, 01 Apr 2023	AIC	1496.347
Time:	06:47:38	BIC	1510.723
Sample:	01-01-1980	HQIC	1502.188
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.4794	0.125	3.850	0.000	0.235	0.723
ar.L2	0.0166	0.111	0.150	0.881	-0.200	0.234
ar.L3	0.0281	0.108	0.260	0.794	-0.184	0.240
ma.L1	-0.8529	0.109	-7.792	0.000	-1.067	-0.638
sigma2	4939.3625	432.933	11.409	0.000	4090.830	5787.895

Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	54.46
Prob(Q):	0.90	Prob(JB):	0.00
Heteroskedasticity (H):	12.74	Skew:	-0.01
Prob(H) (two-sided):	0.00	Kurtosis:	6.16

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

We get a comparatively simpler model by looking at the ACF and the PACF plots.

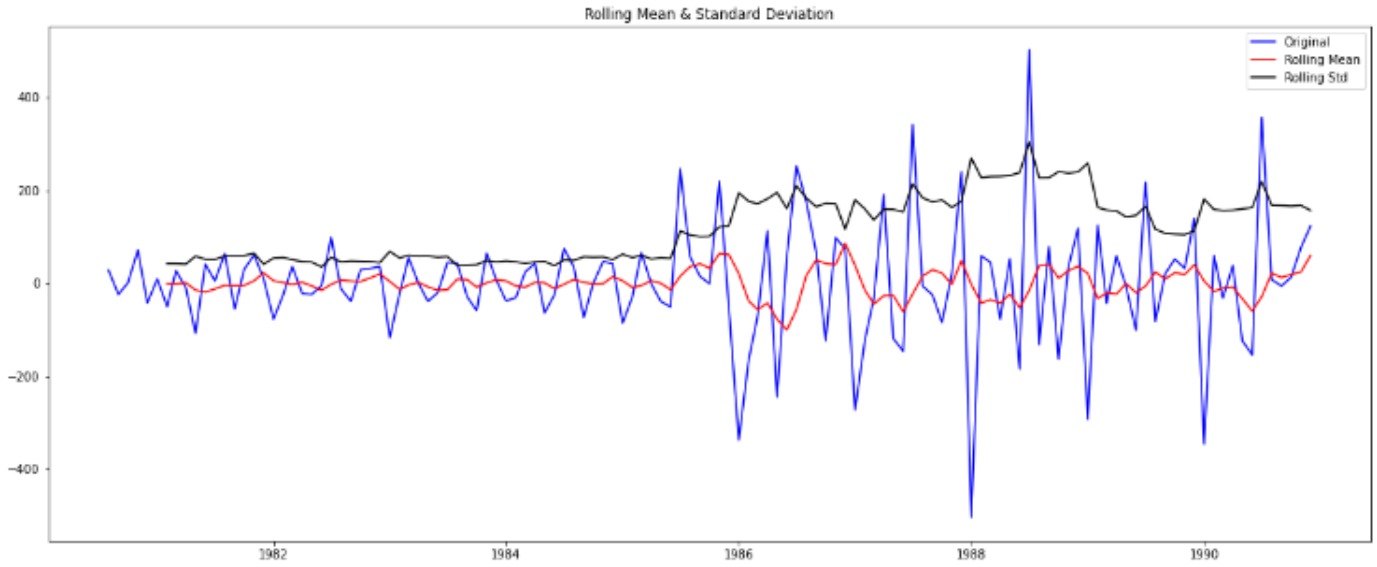
Predict on the Test Set using this model and evaluate the model.

RMSE : 144.18

	Test RMSE
RegressionOnTime	266.276472
NaiveModel	245.121306
SimpleAverageModel	63.984570
2pointTrailingMovingAverage	45.948736
4pointTrailingMovingAverage	57.872686
6pointTrailingMovingAverage	63.456893
9pointTrailingMovingAverage	67.723648
Alpha=0.6,SimpleExponentialSmoothing	196.404850
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	890.968504
Alpha=0.57,Beta=0.00014,Gamma=0.202,TripleExponentialSmoothing	83.734048
Alpha=0.1,Beta=0.6,Gamma=0.2,TripleExponentialSmoothing	41.237522
ARIMA(1,1,1)	142.820730
SARIMA(2,1,2)(2,0,2,6)	57.030694
SARIMA(0,1,2)(1,0,2,12)	69.030660
ARIMA(3,1,1)	144.183934

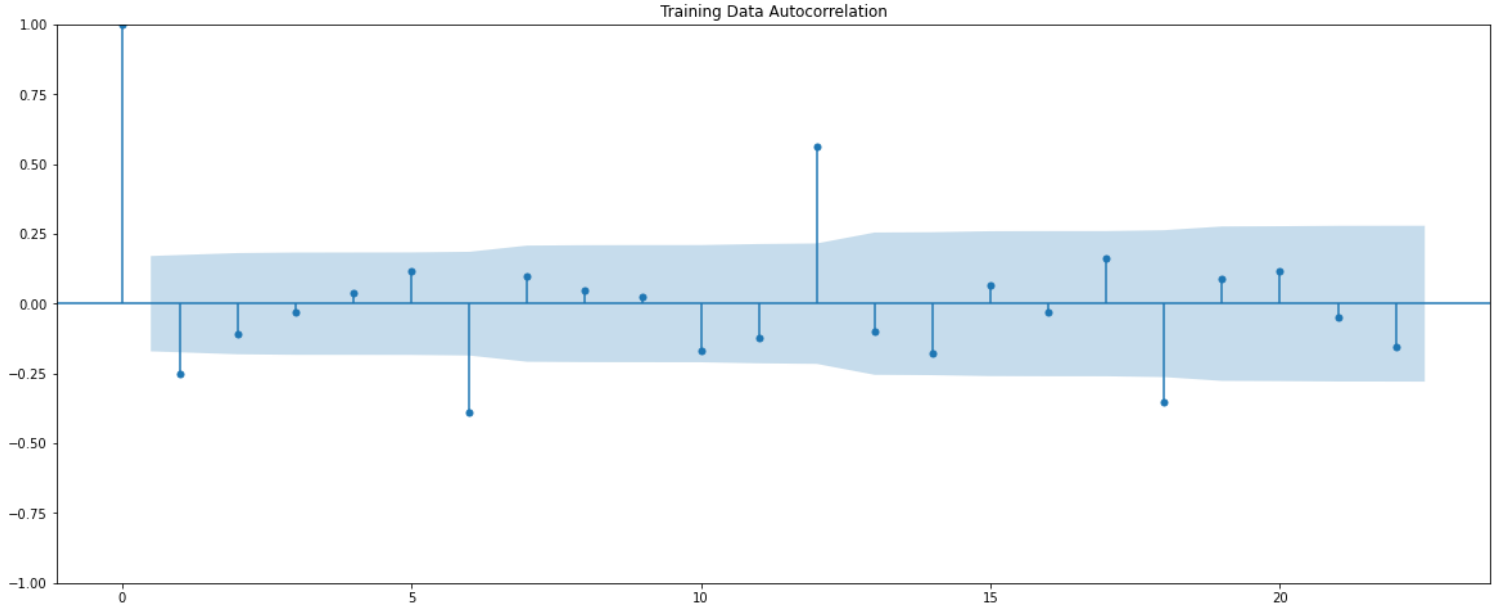
RMSE value is higher than Arima model built on the basis of lowest AIC.

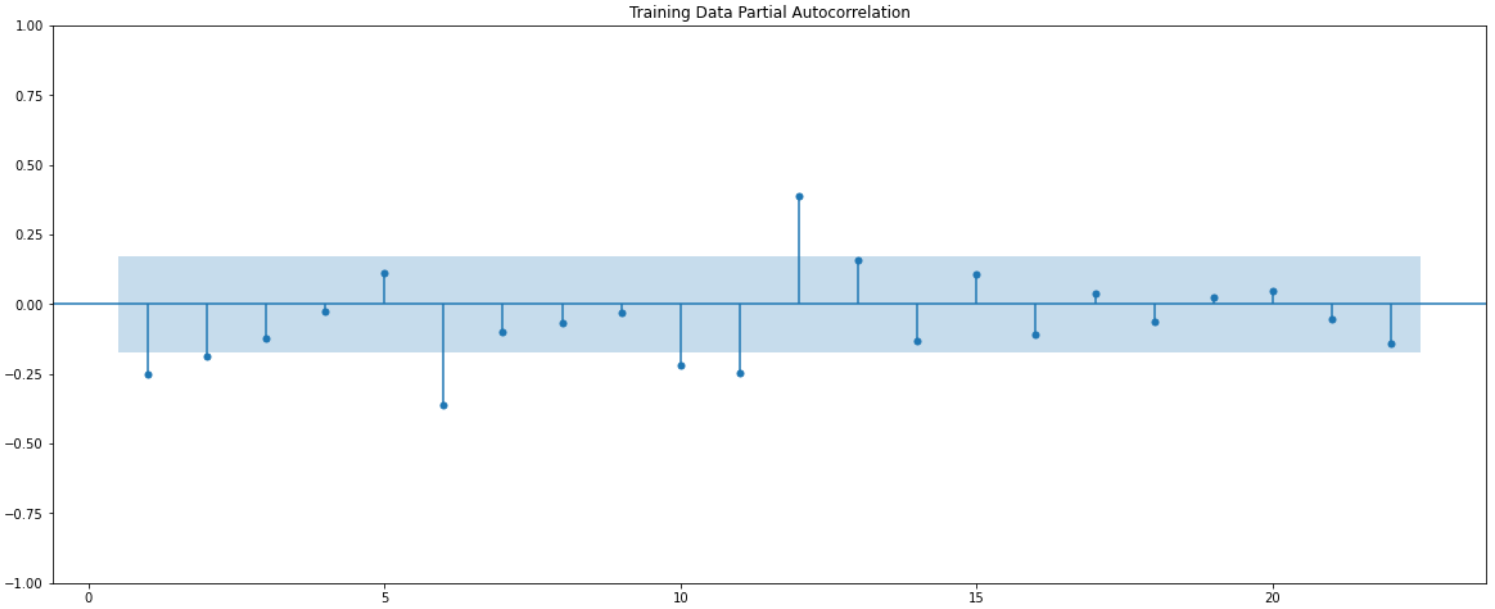
Build a version of the SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots. - Seasonality at 6 or 12.




```
Results of Dickey-Fuller Test:
Test Statistic                -1.133336e+01
p-value                        1.098825e-20
#Lags Used                     6.000000e+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
```

As p value is less than 0.05 hence the data is stationary thus we can build SARIMA model by looking at ACF & PACF plot taking seasonality as 6.





Here, we have taken alpha=0.05.

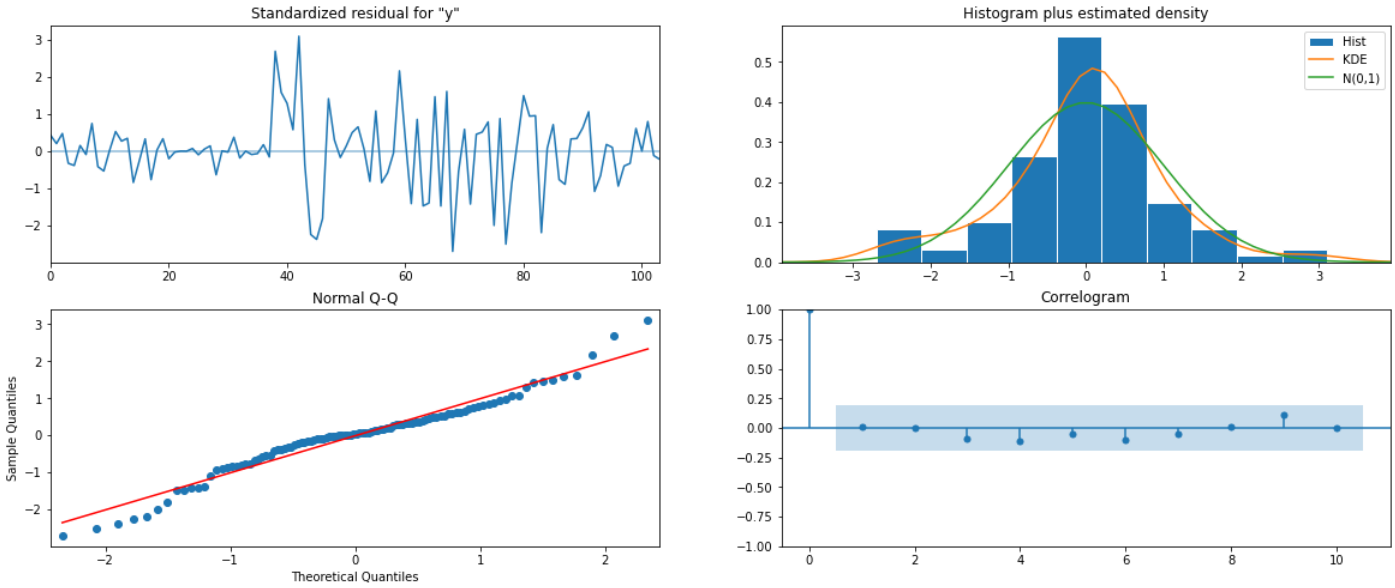
We are going to take the seasonal period as 12. We will keep the p(3) and q(2) parameters same as the ARIMA model.

- 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 6.

Manually choosing p,d,q & P,D,Q & seasonality=12 , values with lowest RMSE.

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(1, 1, 2)x(1, 1, [1], 12)   Log Likelihood                -571.985
Date:                              Sat, 01 Apr 2023   AIC                           1155.970
Time:                                      08:41:36   BIC                           1171.837
Sample:                                           0   HQIC                          1162.398
                                                - 132
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.5545      0.523     -1.059      0.289      -1.580       0.471
ma.L1          0.1751      0.531      0.329      0.742      -0.867       1.217
ma.L2         -0.0983      0.218     -0.450      0.653      -0.526       0.330
ar.S.L12       0.0388      0.156      0.249      0.804      -0.267       0.344
ma.S.L12      -0.6066      0.129     -4.703      0.000      -0.859      -0.354
sigma2      3442.2390    442.944      7.771      0.000    2574.084    4310.394
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 6.27
Prob(Q):                              0.96   Prob(JB):                         0.04
Heteroskedasticity (H):               7.13   Skew:                            -0.14
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.17
===================================================================================

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



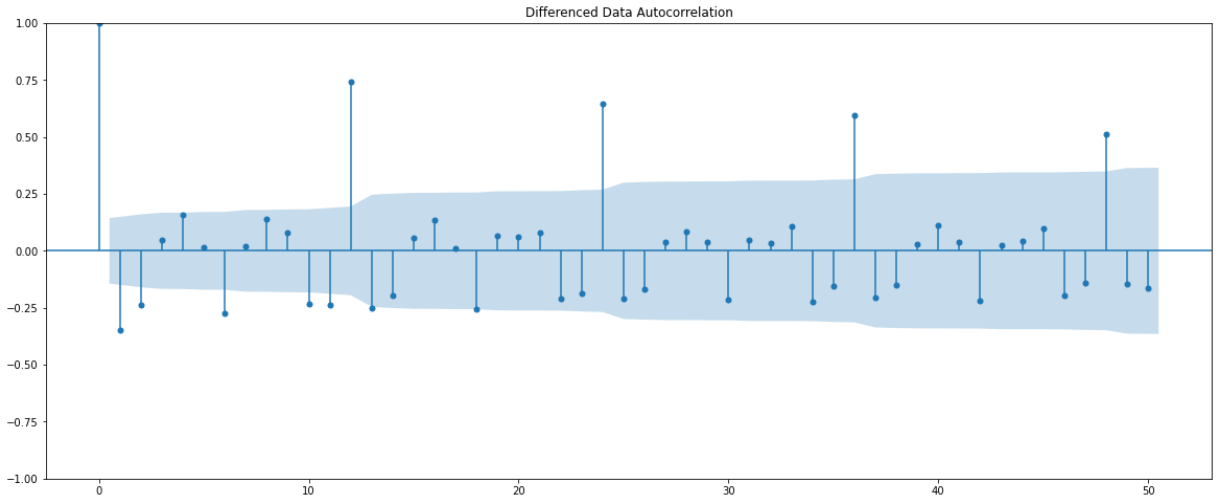
Predict on the Test Set using this model and evaluate the model

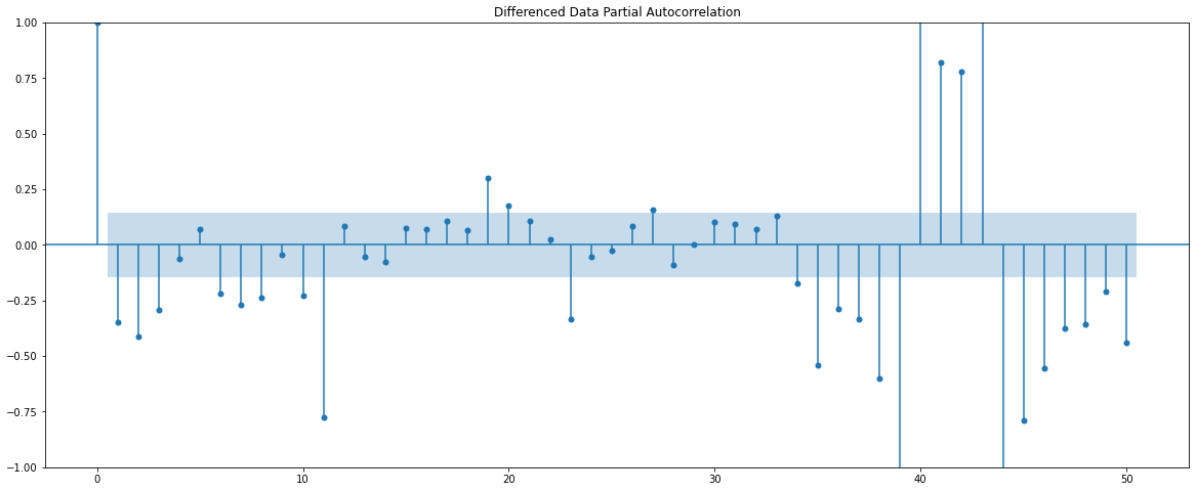
	y	mean	mean_se	mean_ci_lower	mean_ci_upper
0		239.571888	58.671615	124.577635	354.566140
1		220.744805	69.050105	85.409087	356.080523
2		240.138209	81.336553	80.721494	399.554923
3		242.114318	90.349711	65.032137	419.196498
4		235.798486	99.366705	41.043323	430.553649

RMSE : 50.25

Soft Drink Data

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





Here, we have taken alpha=0.05.

By looking at the above plots, we will take the value of p and q to be 3 and 2 respectively.

SARIMAX Results

Dep. Variable:	SoftDrinkProduction	No. Observations:	132
Model:	ARIMA(3, 1, 2)	Log Likelihood	-1024.340
Date:	Sat, 01 Apr 2023	AIC	2060.680
Time:	08:30:44	BIC	2077.931
Sample:	01-01-1980	HQIC	2067.690
	- 12-01-1990		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3216	0.460	-0.700	0.484	-1.223	0.579
ar.L2	-0.0001	0.189	-0.001	0.999	-0.371	0.370
ar.L3	-0.0227	0.218	-0.104	0.917	-0.449	0.404
ma.L1	-0.2633	0.453	-0.581	0.561	-1.151	0.624
ma.L2	-0.6405	0.400	-1.599	0.110	-1.425	0.144
sigma2	3.514e+05	4.97e+04	7.066	0.000	2.54e+05	4.49e+05

Ljung-Box (L1) (Q):	0.07	Jarque-Bera (JB):	0.43
Prob(Q):	0.80	Prob(JB):	0.81
Heteroskedasticity (H):	1.29	Skew:	0.04
Prob(H) (two-sided):	0.40	Kurtosis:	2.73

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

We get a comparatively simpler model by looking at the ACF and the PACF plots.

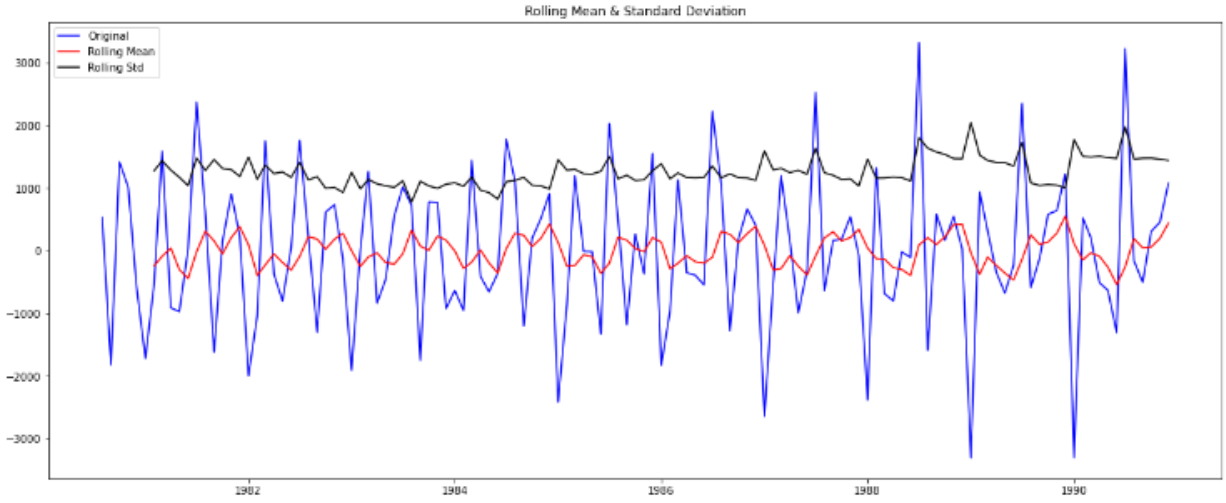
Predict on the Test Set using this model and evaluate the model.

RMSE : 822.21

	Test RMSE
RegressionOnTime	3202.844447
NaiveModel	1519.259233
SimpleAverageModel	934.353358
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827
Alpha=0.2,SimpleExponentialSmoothing	847.635259
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	982.938364
Alpha=0.11,Beta=0.04,Gamma=0.23,TripleExponentialSmoothing	447.722581
Alpha=0.5,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	442.214551
ARIMA(0,1,2)	831.615852
SARIMA(0,1,2)(2,0,2,6)	447.942606
SARIMA(0,1,2)(2,0,2,12)	437.706534
ARIMA(3,1,2)	822.217445

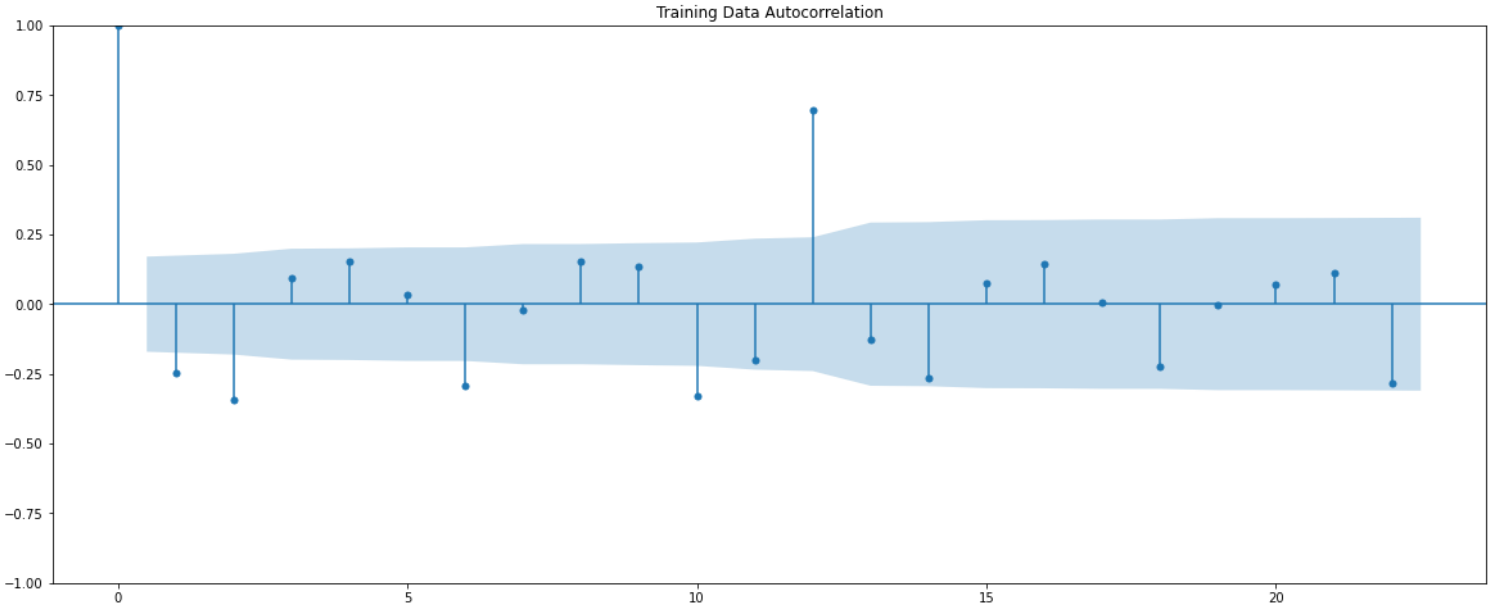
RMSE value is lower than Arima model built on the basis of lowest AIC.

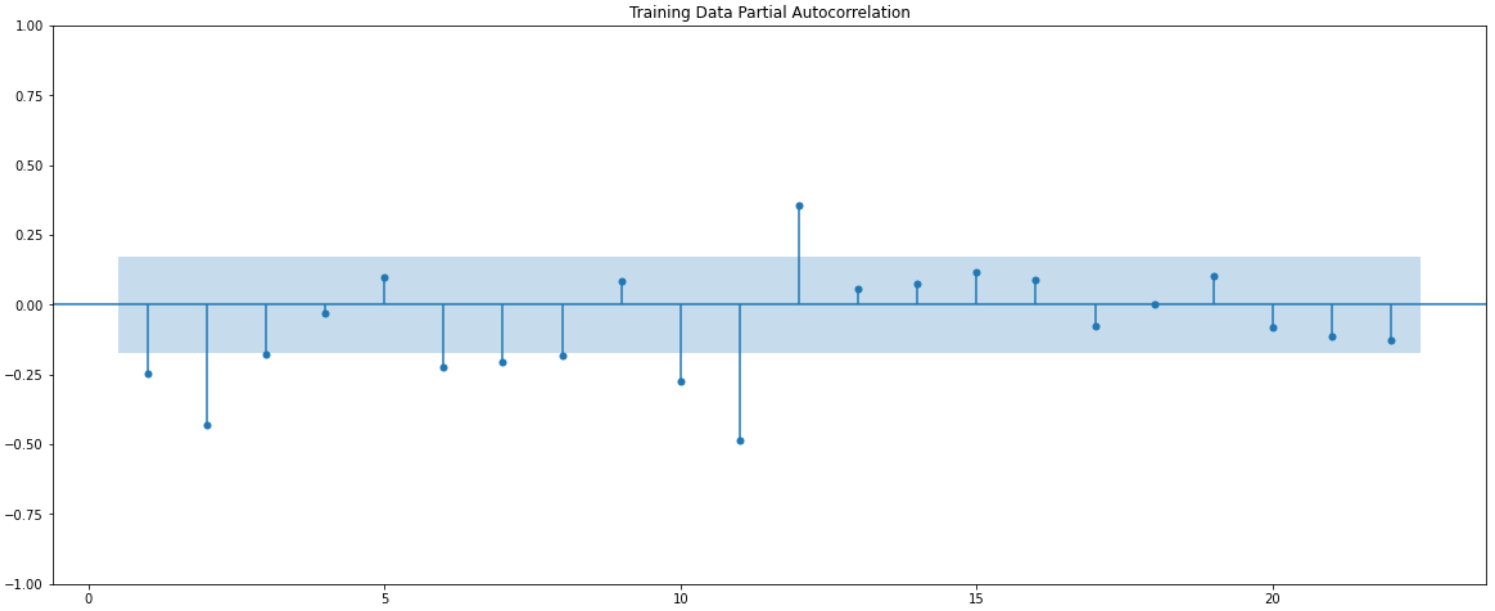
Build a version of the SARIMA model for which the best parameters are selected by looking at the ACF and the PACF plots. - Seasonality at 6.




```
Results of Dickey-Fuller Test:
Test Statistic                -7.167226e+00
p-value                        2.865418e-10
#Lags Used                     1.200000e+01
Number of Observations Used    1.120000e+02
Critical Value (1%)           -3.490131e+00
Critical Value (5%)           -2.887712e+00
Critical Value (10%)          -2.580730e+00
dtype: float64
```

As p value is less than 0.05 hence the data is stationary thus we can build SARIMA model by looking at ACF & PACF plot taking seasonality as 6.





Here, we have taken alpha=0.05.

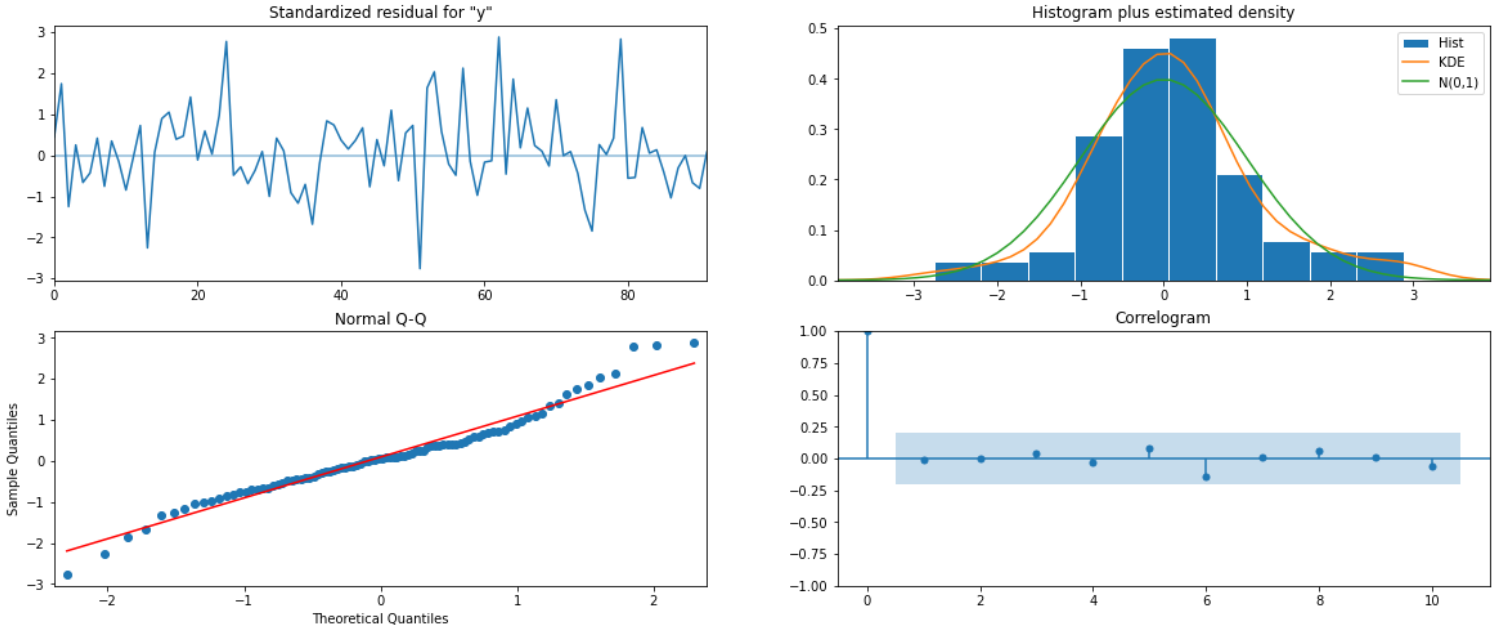
We are going to take the seasonal period as 12. We will keep the p(3) and q(2) parameters same as the ARIMA model.

- 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 2.

Manually choosing p,d,q & P,D,Q & seasonality=12 , values with lowest RMSE

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  132
Model:             SARIMAX(0, 1, 2)x(0, 1, 2, 12)   Log Likelihood                -667.910
Date:                            Sat, 01 Apr 2023   AIC                           1345.821
Time:                                    08:41:06   BIC                           1358.430
Sample:                                         0   HQIC                          1350.910
                                              - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8157      0.120     -6.807      0.000      -1.051      -0.581
ma.L2         -0.0855      0.106     -0.806      0.420      -0.293       0.122
ma.S.L12      -0.5959      0.114     -5.235      0.000      -0.819      -0.373
ma.S.L24      -0.0273      0.157     -0.174      0.862      -0.334       0.280
sigma2      1.157e+05   1.43e+04      8.083      0.000    8.76e+04    1.44e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 7.15
Prob(Q):                              0.95   Prob(JB):                         0.03
Heteroskedasticity (H):               1.16   Skew:                             0.37
Prob(H) (two-sided):                  0.68   Kurtosis:                         4.15
===================================================================================

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



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

y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	2627.038254	340.185198	1960.287518	3293.788990
1	2950.201535	345.919223	2272.212316	3628.190754
2	3387.943087	347.546294	2706.764849	4069.121285
3	3069.619553	349.170227	2385.258485	3753.980622
4	3274.185842	350.786855	2586.656240	3961.715444

RMSE : 434.61

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.

ShoeSales Data

	Test RMSE
Alpha=0.1,Beta=0.6,Gamma=0.2,TripleExponentialSmoothing	41.237522
2pointTrailingMovingAverage	45.948736
SARIMA(1,1,2)(1,1,1,12)	50.251949
SARIMA(2,1,2)(2,0,2,6)	57.030694
4pointTrailingMovingAverage	57.872686
6pointTrailingMovingAverage	63.456893
SimpleAverageModel	63.984570
9pointTrailingMovingAverage	67.723648
SARIMA(0,1,2)(1,0,2,12)	69.030660
Alpha=0.57,Beta=0.00014,Gamma=0.202,TripleExponentialSmoothing	83.734048
ARIMA(1,1,1)	142.820730
ARIMA(3,1,1)	144.183934
Alpha=0.6,SimpleExponentialSmoothing	196.404850
NaiveModel	245.121306
RegressionOnTime	266.276472
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	890.968504

Alpha=0.1,Beta=0.6,Gamma=0.2,TripleExponentialSmoothing has the least Test RMSE 41.23. Hence best model/method for complete data to predict 12 months into future.

Soft Drink Data

Sorted by RMSE values on the Test Data:

	Test RMSE
SARIMA(0,1,2)(0,1,2,12)	434.614623
SARIMA(0,1,2)(2,0,2,12)	437.706534
Alpha=0.5,Beta=0.1,Gamma=0.2,TripleExponentialSmoothing	442.214551
Alpha=0.11,Beta=0.04,Gamma=0.23,TripleExponentialSmoothing	447.722581
SARIMA(0,1,2)(2,0,2,6)	447.942606
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827
ARIMA(3,1,2)	822.217445
ARIMA(0,1,2)	831.615852
Alpha=0.2,SimpleExponentialSmoothing	847.635259
SimpleAverageModel	934.353358
Alpha=0.1,Beta=0.1,DoubleExponentialSmoothing	982.938364
NaiveModel	1519.259233
RegressionOnTime	3202.844447

SARIMA(0,1,2)(0,1,2,12) has the least Test RMSE 434. Hence best model/method for complete data to predict 12 months into future.

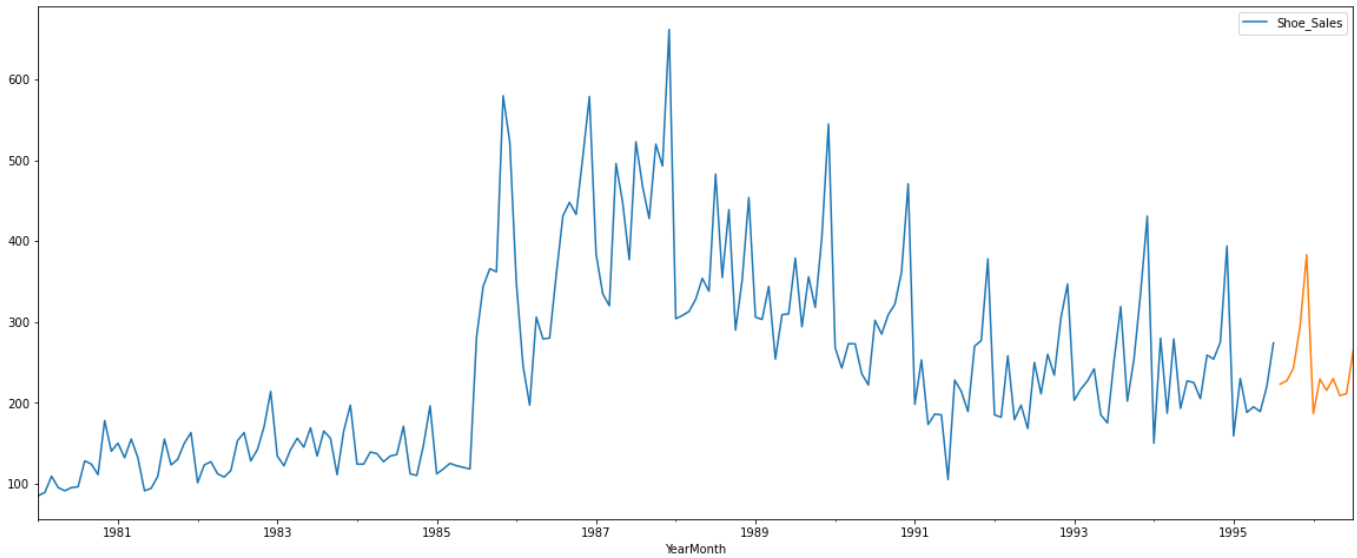
9. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

ShoeSales Data

Building the most optimum model on the Full Data.

RMSE : 65.72

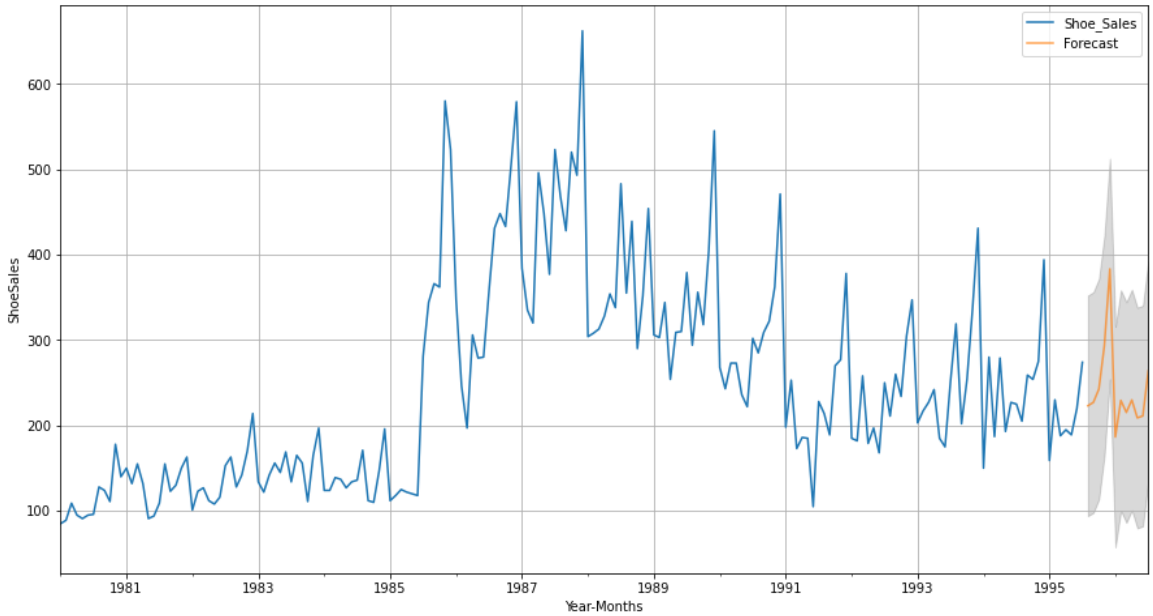
Getting the predictions for 12 months into future



We have calculated the upper and lower confidence bands at 95% confidence level

	lower_CI	prediction	upper_ci
1995-08-01	93.929879	223.070867	352.211856
1995-09-01	97.890414	227.031403	356.172392
1995-10-01	113.444210	242.585199	371.726188
1995-11-01	164.320143	293.461131	422.602120
1995-12-01	254.106116	383.247105	512.388093

Plot the forecast along with the confidence band



The above picture shows an increasing decreasing trend of sales for 12 months into future.

Soft Drink Data

Building the most optimum model on the Full Data

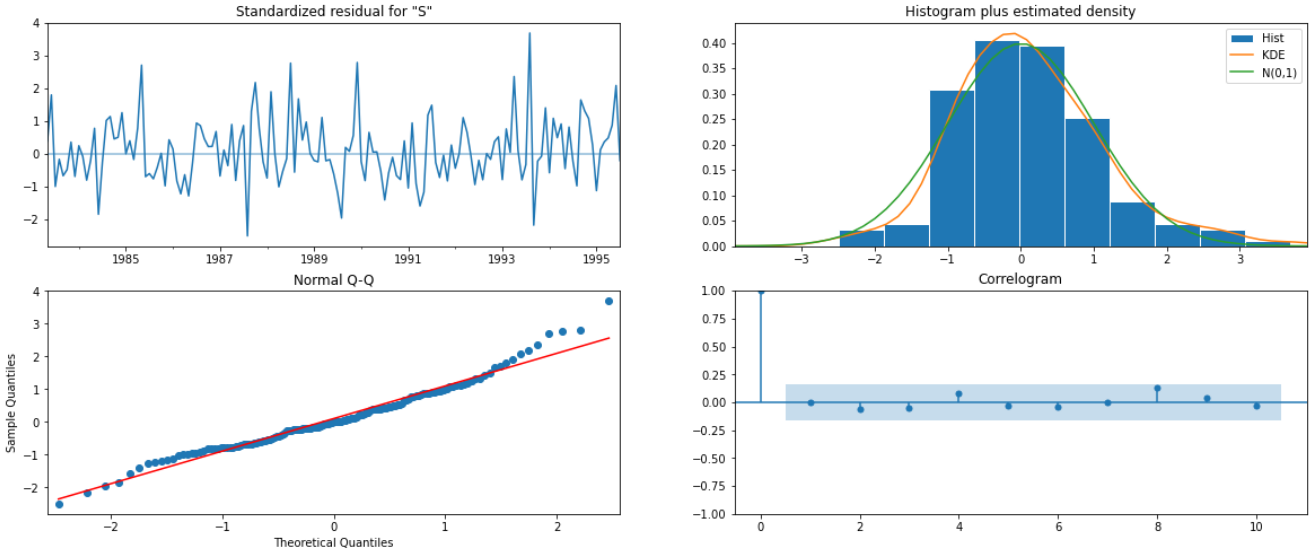
SARIMAX Results

Dep. Variable:	SoftDrinkProduction	No. Observations:	187
Model:	SARIMAX(0, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-1073.053
Date:	Sat, 01 Apr 2023	AIC	2156.106
Time:	09:19:07	BIC	2171.058
Sample:	01-01-1980 - 07-01-1995	HQIC	2162.181
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.9061	0.080	-11.329	0.000	-1.063	-0.749
ma.L2	0.0553	0.085	0.652	0.515	-0.111	0.222
ma.S.L12	-0.7159	0.088	-8.093	0.000	-0.889	-0.543
ma.S.L24	0.0924	0.098	0.945	0.345	-0.099	0.284
sigma2	1.268e+05	1.23e+04	10.343	0.000	1.03e+05	1.51e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	15.99
Prob(Q):	0.95	Prob(JB):	0.00
Heteroskedasticity (H):	1.52	Skew:	0.60
Prob(H) (two-sided):	0.15	Kurtosis:	4.08

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



Evaluate the model on the whole and predict 12 months into the future

SoftDrinkProduction	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	4474.357367	356.051563	3776.509127	5172.205607
1995-09-01	4197.505465	357.618152	3496.586767	4898.424162
1995-10-01	4504.320943	361.544807	3795.706142	5212.935743
1995-11-01	5190.166618	365.429272	4473.938406	5906.394829
1995-12-01	6037.629324	369.272877	5313.867784	6761.390863

We have calculated the upper and lower confidence bands at 95% confidence level

Table 25

```
RMSE of the Full Model 459.7380945255836
```

Plot the forecast along with the confidence band

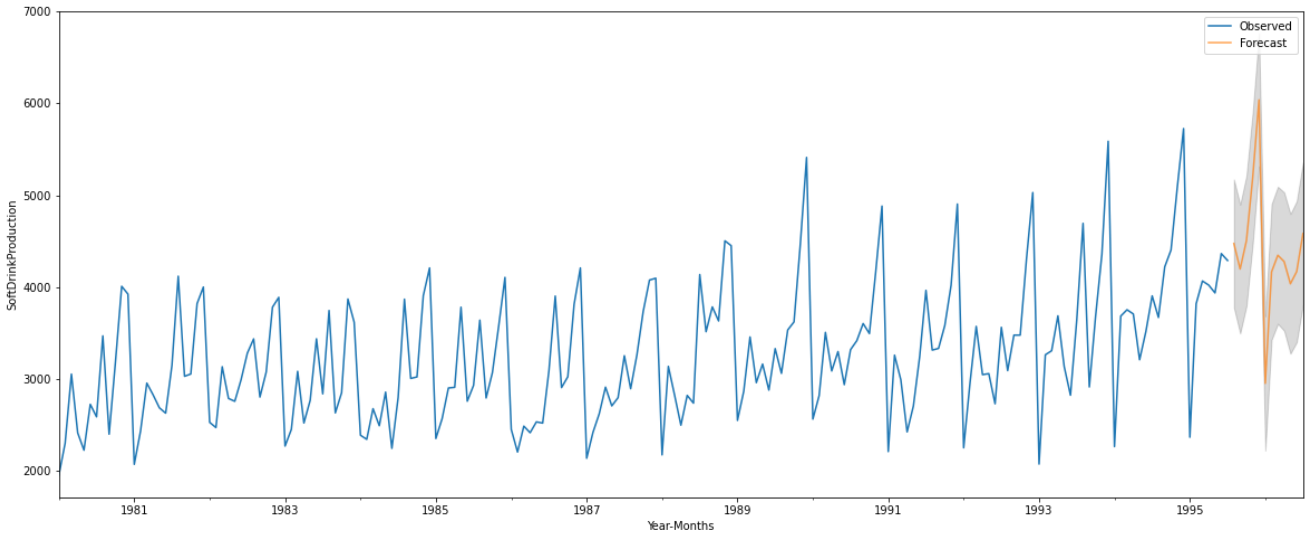


Fig 25

The above picture shows an increasing decreasing trend of sales for 12 months into the future.

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

Shoe Sales Data

Comments/Insights

- December month has the highest sales in a year
- Model plot was build based on the trend and seasonality. We see the future prediction is in line with the previous year predictions. Model is actually a best fit model.
- Alpha=0.1,Beta=0.6,Gamma=0.2,TripleExponentialSmoothing has the least Test RMSE 41.23. Hence best model/method for complete data to predict 12 months into future.

Recommendations

- Shoe sales are seasonal.
- Company should plan a head and keep enough stock from September till December to capitalize on the demand.
- In order to increase the sales company should plan some promotional offers(Ex- 15% off) from January till June so that there will be steady sales throughout the year.
- Production can be kept at minimal in order to keep the product alive, with maybe maximum production happening in the months of March and October when the sales of shoes are at an upper level compared to the rest. Some promotional offers during the dull months, especially during the festive season of the last quarter might help with the sales.
- For the last 15 years a downward trend in the sales of Shoes is observed. Hence the reason behind such a decreasing trend has to be found out. This could because of external factors like competitors.
- There is need for analysis of price of shoes of competitors to gain market attention. Further the decreasing trend could be because of internal factors like Shoe quality, Brand name, Distribution chain mismanagement.There is a peak in sales especially for the month of December. This opportunity of higher sales in this month needs to be exploited by providing sales related incentives to the Distributors.

Soft Drink Data

Comments/Insights

- December month has the highest sales in a year
- Model plot was build based on the trend and seasonality. We see the future prediction is in line with the previous year predictions. Model is actually a best fit model.
- SARIMA(0,1,2)(0,1,2,12) has the least Test RMSE 434. Hence best model/method for complete data to predict 12 months into future.

Recommendations

- Soft Drink sales are seasonal.
- Company should plan a head and keep enough stock from September till December to capitalize on the demand.
- In order to increase the sales company should plan some promotional offers from January till June so that there will be steady sales throughout the year.
- Offering extra 10% quantity in soft drink bottle can also attract customers.
- Production can be kept at minimal in order to keep the product alive, with maybe maximum production happening in the months of March and October when the sales of Soft Drink are at an upper level compared to the rest. Some promotional offers during the dull months, especially during the festive season of the last quarter might help with the sales.
- For the last 15 years a downward trend in the sales of Soft Drink is observed. Hence the reason behind such a decreasing trend has to be found out. This could because of external factors like competitors.
- There is need for analysis of price of soft drinks of competitors to gain market attention.Further the decreasing trend could be because of internal factors like Soft Drink quality, Brand name, Distribution chain mismanagement, Advertisement quality etc. There is a peak in sales especially for the month of December. This opportunity of higher sales in this month needs to be exploited by providing sales related incentives to the Distributors.

THANK YOU