**Soft Drink Sales**

**TS Analysis**

Name: Saurabh Dharmadhikari

PGP-DSBA Online January’ 21

Table of Contents:

Executive Summary……………………...................................................................................................................................3

Introduction ...................................................................................................................................................................3

Data Description ............................................................................................................................................................3 1) Read the data as an appropriate Time Series data and plot the data. .………………………………………………………….….....3

2)Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition. ….........5

3)Split the data into training and test. The test data should start in 1991. …………………………………………………………....13

4)Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data.  
Other models such as regression,naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE. …………………………………………………………........15

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. .................................................................................................23

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. ..........25

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. ............................................................................................................................30

8)Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data. ..........................................................................................................................................................32

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........................................................................33

10)Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. ...........................................................................................................................................34

Executive Summary

We are analysts in the RST soft drink company and we 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 us from January 1980 to July 1995.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using time series model. The data consists of 187 entries of accidents. We are provided with month and year and total production of soft drinks for that month mentioned below in the Data Description. Our mission is to build a model that can predict level of production for the upcoming months and years. For this we will be developing time series model both additive and multiplicative to see which one suits us best.

Data Description

1. YearMonth: Year and month of production

2. SoftDrinkProduction: Quantity produced

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

Sample of the data set

First 5 observations of the Data set on SoftDrinkProduction

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

Last 5 observations of the Data set on SoftDrinkProduction

|  | **SoftDrinkProduction** |
| --- | --- |
| **YearMonth** |  |
| **1995-03-01** | 4067 |
| **1995-04-01** | 4022 |
| **1995-05-01** | 3937 |
| **1995-06-01** | 4365 |
| **1995-07-01** | 4290 |

We can see the initial 5 rows and last 5 rows of our data set in the above sample. Let us study more about this data set.

Exploratory data analysis

DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01

Data columns (total 1 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 SoftDrinkProduction 187 non-null int64

dtypes: int64(1)

We have a total of 187 rows and 1 column also an index with year and month. The column is int type.

Check for null values

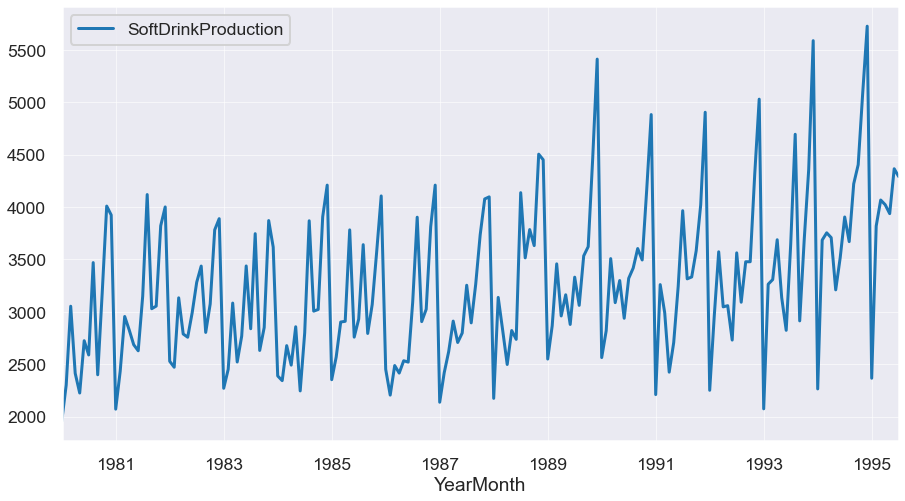
SoftDrinkProduction 0

dtype: int64

There are no null values in our data base.

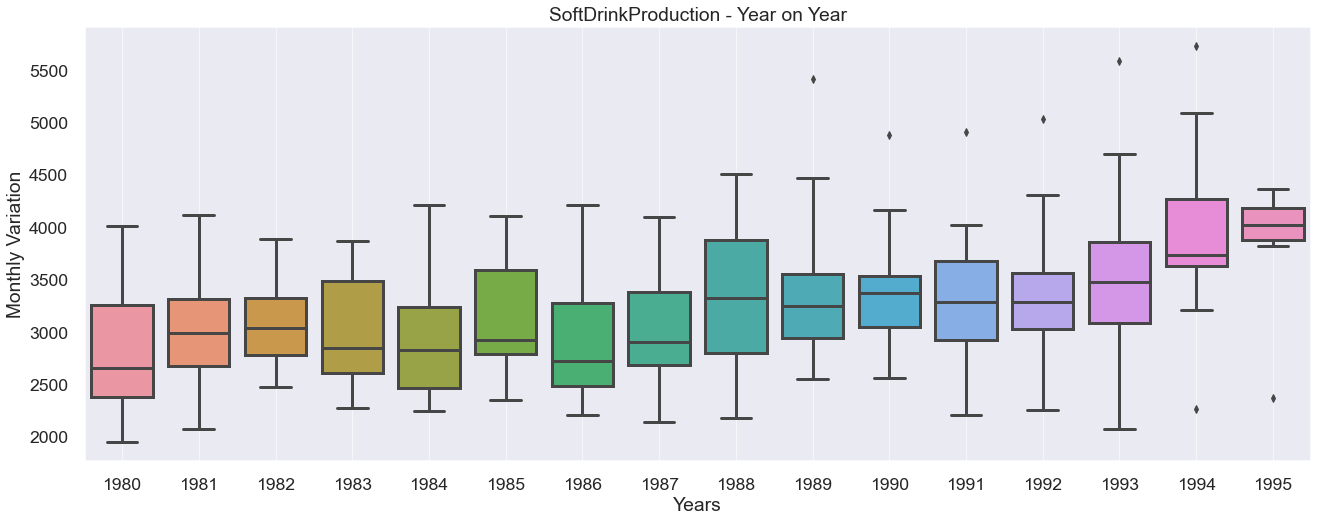
1. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Plotting soft drink production over the years.



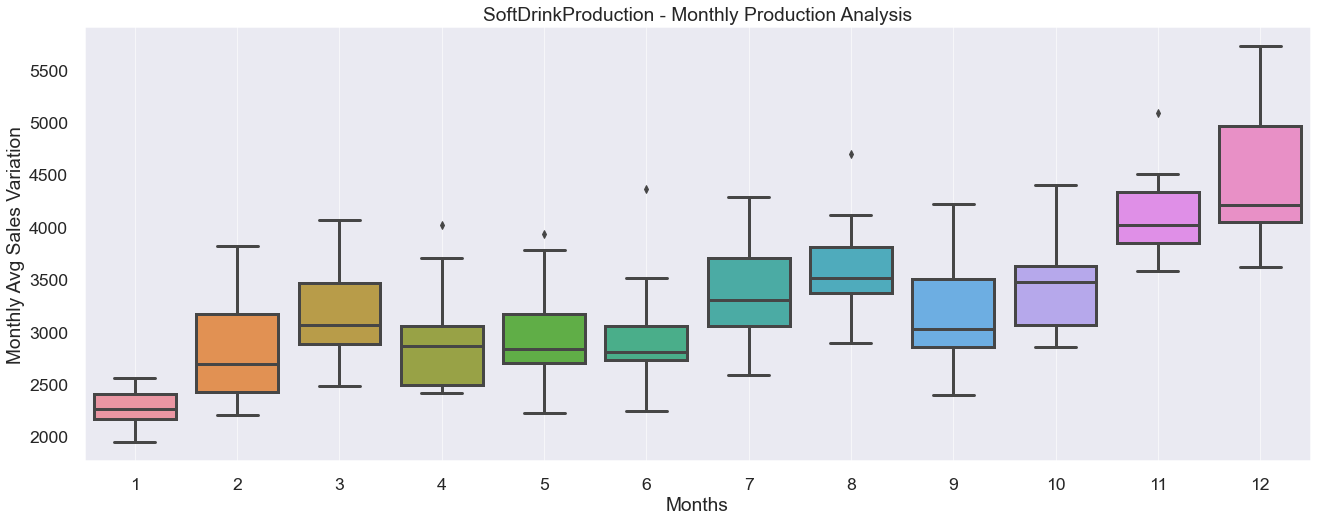
The above graph lets us see the production of soft drinks over the months and years. We can certainly see the rise and fall of production over these years.

Let us find outliers in the production in the years.



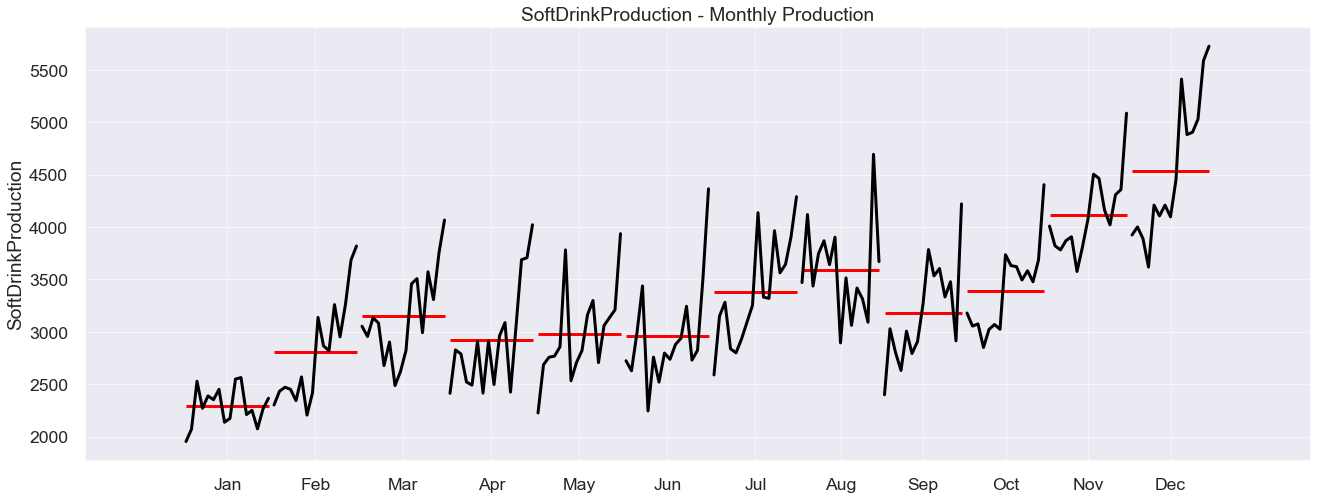
We can see that there are a few months in a few years when the production of soft drinks has been very high and two years (1994 and 1995) when the production has been very low. We can also see here that the production of soft drinks has been on an increase trend when we compare the averages production for the years given.

Let us explore total monthly production for all the years.



Here we can easily see that the average monthly production for December has been the highest and lowest for the month of January. Also, there are a few outliers.

Monthly plot of production for each year:



From the above chart we can clearly see that the production levels of soft drink for the month of January have remained more or less the same over the years but production in other months has risen a lotand show an upward trend.

| **YearMonth** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **YearMonth** |  |  |  |  |  |  |  |  |  |  |  |  |
| **1980** | 1954.0 | 2302.0 | 3054.0 | 2414.0 | 2226.0 | 2725.0 | 2589.0 | 3470.0 | 2400.0 | 3180.0 | 4009.0 | 3924.0 |
| **1981** | 2072.0 | 2434.0 | 2956.0 | 2828.0 | 2687.0 | 2629.0 | 3150.0 | 4119.0 | 3030.0 | 3055.0 | 3821.0 | 4001.0 |
| **1982** | 2529.0 | 2472.0 | 3134.0 | 2789.0 | 2758.0 | 2993.0 | 3282.0 | 3437.0 | 2804.0 | 3076.0 | 3782.0 | 3889.0 |
| **1983** | 2271.0 | 2452.0 | 3084.0 | 2522.0 | 2769.0 | 3438.0 | 2839.0 | 3746.0 | 2632.0 | 2851.0 | 3871.0 | 3618.0 |
| **1984** | 2389.0 | 2344.0 | 2678.0 | 2492.0 | 2858.0 | 2246.0 | 2800.0 | 3869.0 | 3007.0 | 3023.0 | 3907.0 | 4209.0 |
| **1985** | 2353.0 | 2570.0 | 2903.0 | 2910.0 | 3782.0 | 2759.0 | 2931.0 | 3641.0 | 2794.0 | 3070.0 | 3576.0 | 4106.0 |
| **1986** | 2452.0 | 2206.0 | 2488.0 | 2416.0 | 2534.0 | 2521.0 | 3093.0 | 3903.0 | 2907.0 | 3025.0 | 3812.0 | 4209.0 |
| **1987** | 2138.0 | 2419.0 | 2622.0 | 2912.0 | 2708.0 | 2798.0 | 3254.0 | 2895.0 | 3263.0 | 3736.0 | 4077.0 | 4097.0 |
| **1988** | 2175.0 | 3138.0 | 2823.0 | 2498.0 | 2822.0 | 2738.0 | 4137.0 | 3515.0 | 3785.0 | 3632.0 | 4504.0 | 4451.0 |
| **1989** | 2550.0 | 2867.0 | 3458.0 | 2961.0 | 3163.0 | 2880.0 | 3331.0 | 3062.0 | 3534.0 | 3622.0 | 4464.0 | 5411.0 |
| **1990** | 2564.0 | 2820.0 | 3508.0 | 3088.0 | 3299.0 | 2939.0 | 3320.0 | 3418.0 | 3604.0 | 3495.0 | 4163.0 | 4882.0 |
| **1991** | 2211.0 | 3260.0 | 2992.0 | 2425.0 | 2707.0 | 3244.0 | 3965.0 | 3315.0 | 3333.0 | 3583.0 | 4021.0 | 4904.0 |
| **1992** | 2252.0 | 2952.0 | 3573.0 | 3048.0 | 3059.0 | 2731.0 | 3563.0 | 3092.0 | 3478.0 | 3478.0 | 4308.0 | 5029.0 |
| **1993** | 2075.0 | 3264.0 | 3308.0 | 3688.0 | 3136.0 | 2824.0 | 3644.0 | 4694.0 | 2914.0 | 3686.0 | 4358.0 | 5587.0 |
| **1994** | 2265.0 | 3685.0 | 3754.0 | 3708.0 | 3210.0 | 3517.0 | 3905.0 | 3670.0 | 4221.0 | 4404.0 | 5086.0 | 5725.0 |
| **1995** | 2367.0 | 3819.0 | 4067.0 | 4022.0 | 3937.0 | 4365.0 | 4290.0 | NaN | NaN | NaN | NaN | NaN |

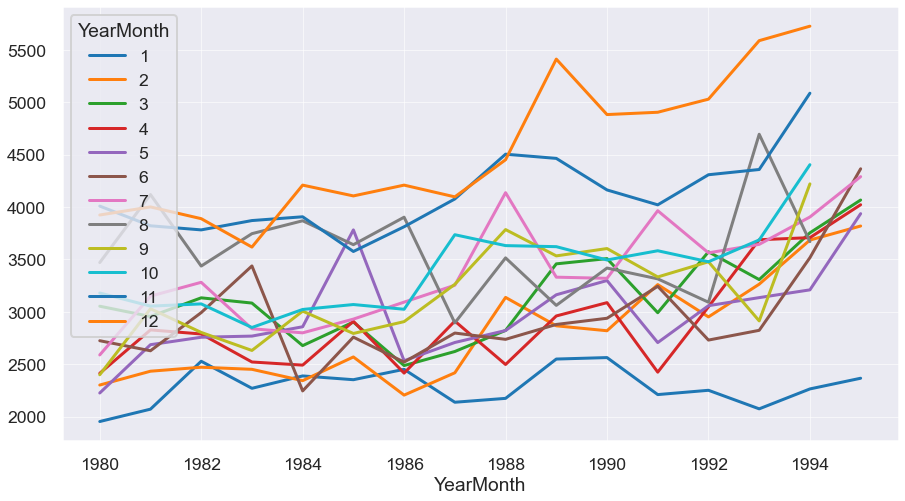
In the above chart we can also see the quantity produced with a calendar format to compare them numerically.

Average monthly production:

SoftDrinkProduction - Average monthly Production

| **YearMonth** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SoftDrinkProduction** | 2289.0 | 2813.0 | 3150.0 | 2920.0 | 2978.0 | 2959.0 | 3381.0 | 3590.0 | 3180.0 | 3394.0 | 4117.0 | 4536.0 |

Plotting monthly production over the years:



With this we easily observe that production for December and November are the highest and for January the lowest.

Total production for each year:

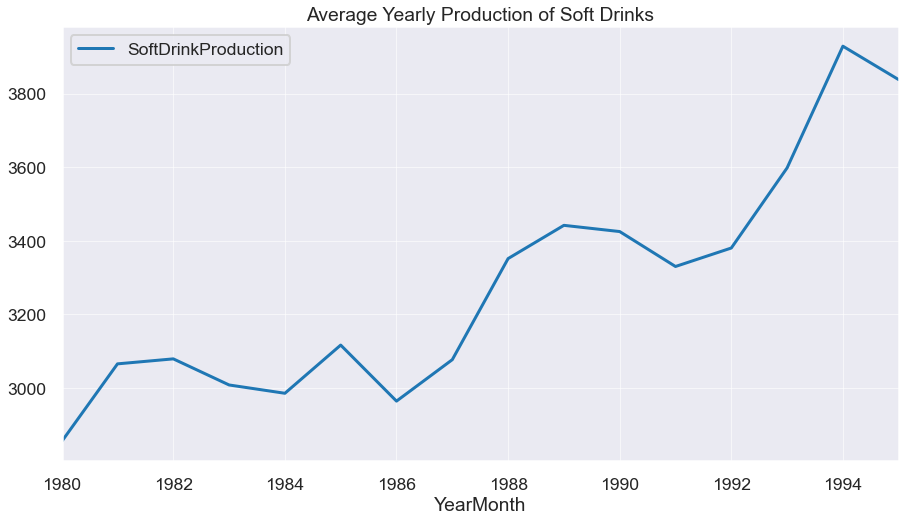
| **YearMonth** | **1980-12-31** | **1981-12-31** | **1982-12-31** | **1983-12-31** | **1984-12-31** | **1985-12-31** | **1986-12-31** | **1987-12-31** | **1988-12-31** | **1989-12-31** | **1990-12-31** | **1991-12-31** | **1992-12-31** | **1993-12-31** | **1994-12-31** | **1995-12-31** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SoftDrinkProduction** | 34247 | 36782 | 36945 | 36093 | 35822 | 37395 | 35566 | 36919 | 40218 | 41303 | 41100 | 39960 | 40563 | 43178 | 47150 | 26867 |



With the help of above graph, we see a steady rise in production of soft drinks. Note: we do not have complete data for the year 1995. Thus, there is a fall in the graph at the end but most probably it would have risen if we had data for the complete year.

Average monthly production of soft drinks:

| **YearMonth** | **1980-12-31** | **1981-12-31** | **1982-12-31** | **1983-12-31** | **1984-12-31** | **1985-12-31** | **1986-12-31** | **1987-12-31** | **1988-12-31** | **1989-12-31** | **1990-12-31** | **1991-12-31** | **1992-12-31** | **1993-12-31** | **1994-12-31** | **1995-12-31** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **SoftDrinkProduction** | 2853.916667 | 3065.166667 | 3078.75 | 3007.75 | 2985.166667 | 3116.25 | 2963.833333 | 3076.583333 | 3351.5 | 3441.916667 | 3425.0 | 3330.0 | 3380.25 | 3598.166667 | 3929.166667 | 3838.142857 |

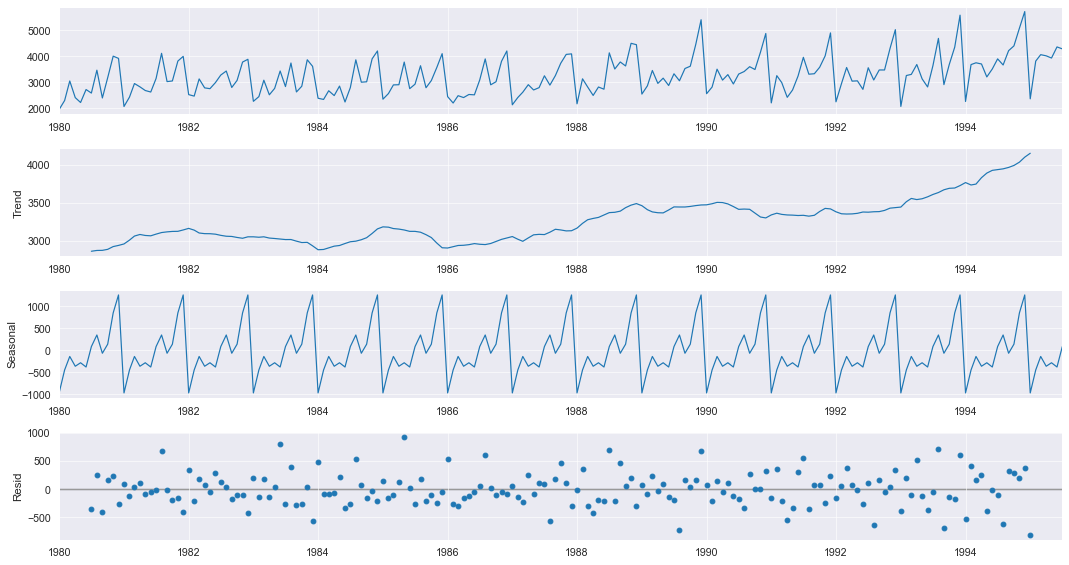


Here we see that we do not have data for the last 5 months of year 1995 and if we had complete data the average sales would have surpassed previous year.

Let us decompose the data for further analysis.

Additive Model:

'Soft Drinks -Decomposition of Production Data - Additive Model'



Above there are four graphs.

First graph shows us the actual plotting of data over the months and years.

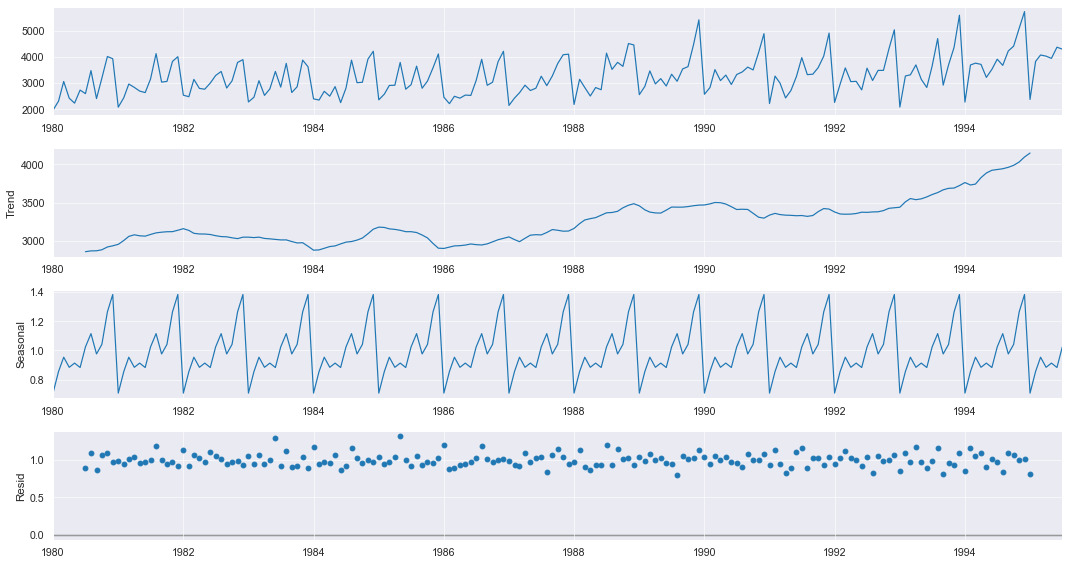
Second graph shows us a trend and we observe that the trend has been rising

Third graph shows us a seasonality and we observe that there is some sort of seasonality.

Fourth graph shows us the residue or data that is un explained. Here we observe that there is a lot of variances ranging from -1000 to 1000.

Multiplicative Model:

'Soft Drink -Decomposition of Production Data - Multiplicative Model'



There are four graphs above.

First graph shows us the actual plotting of data over the months and years.

Second graph shows us a trend and we observe that the trend has been rising

Third graph shows us a seasonality and we observe that there is some sort of seasonality.

Fourth graph shows us the residue or data that is un explained. Here we observe that the magnitude of the residue is very small and thus most of our data is captured by seasonality and trend.

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

Let us split the data set and ready it for predictions.

First 12 observations of Soft Drink Production Train Data

|  | **SoftDrinkProduction** |
| --- | --- |
| **YearMonth** |  |
| **1980-01-01** | 1954 |
| **1980-02-01** | 2302 |
| **1980-03-01** | 3054 |
| **1980-04-01** | 2414 |
| **1980-05-01** | 2226 |
| **1980-06-01** | 2725 |
| **1980-07-01** | 2589 |
| **1980-08-01** | 3470 |
| **1980-09-01** | 2400 |
| **1980-10-01** | 3180 |
| **1980-11-01** | 4009 |
| **1980-12-01** | 3924 |

Total number of Observations in Train Set: 132

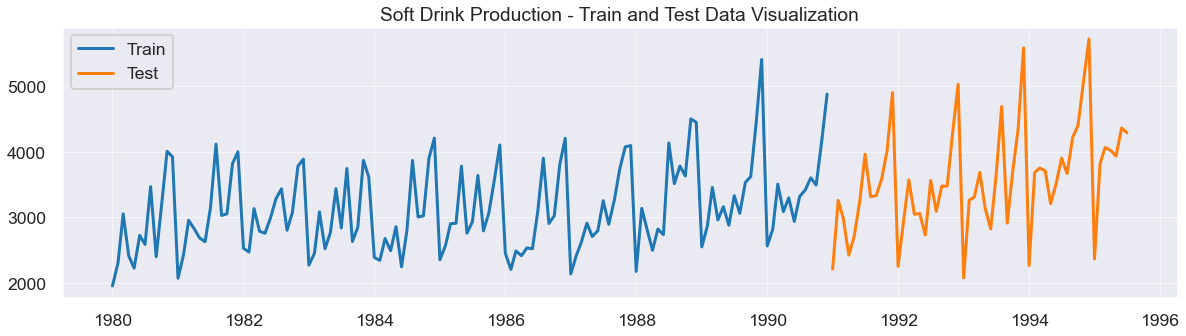
Total observations in our train set are 132.

First 12 observations of Soft Drink Production Test Data

|  | **SoftDrinkProduction** |
| --- | --- |
| **YearMonth** |  |
| **1991-01-01** | 2211 |
| **1991-02-01** | 3260 |
| **1991-03-01** | 2992 |
| **1991-04-01** | 2425 |
| **1991-05-01** | 2707 |
| **1991-06-01** | 3244 |
| **1991-07-01** | 3965 |
| **1991-08-01** | 3315 |
| **1991-09-01** | 3333 |
| **1991-10-01** | 3583 |
| **1991-11-01** | 4021 |
| **1991-12-01** | 4904 |

Total number of Observations in Test Set: 55

Total number of observations in our test set are 55. And we see that our data starts from 1991 January.



Above graph shows us the train and test data.

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

**Linear Regression:**

Predictions:

YearMonth

1991-01-01 3466.213856

1991-02-01 3471.357423

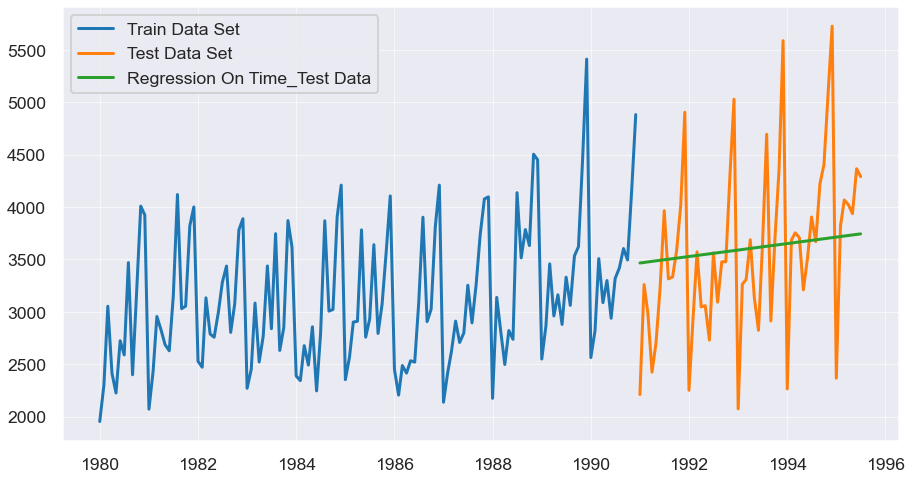
1991-03-01 3476.500990

1991-04-01 3481.644556

1991-05-01 3486.788123

Name: RegOnTime, dtype: float64

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 775.808 | 17.417 |



Above we have built a linear model of prediction and it has a RMSE of 775.808. After plotting the predictions on a graph and comparing it with the actuals we find it gives us a fairly average results towards an increasing trend and we see that they are not very accurate.

**Naïve approach:**

Predictions:

YearMonth

1991-01-01 4882

1991-02-01 4882

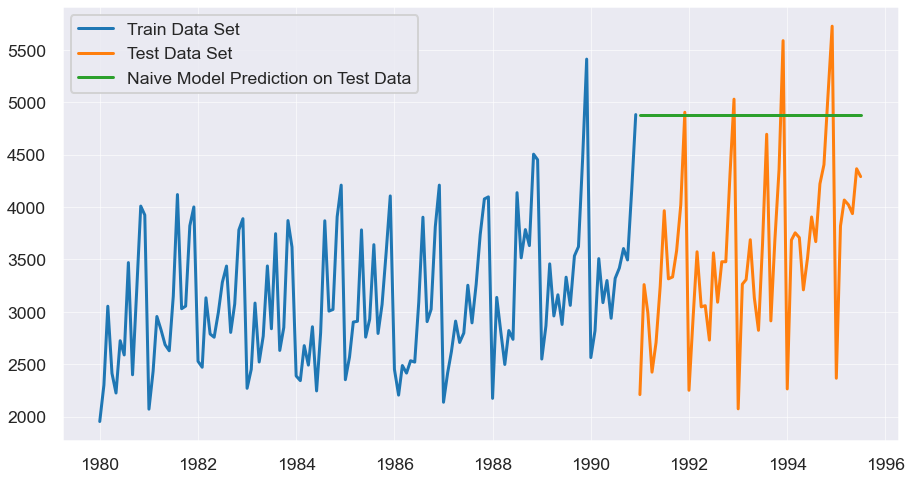
1991-03-01 4882

1991-04-01 4882

1991-05-01 4882

Name: naive, dtype: int64

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **Naive Regression Model** | 1519.259 | 44.275 |



After building a naïve model we observe that the predictions are constant and it has taken the last entry as a constant. Because RMSE of 1519.259 we can say that linear model was much effective than naïve model to predict the production of soft drinks.

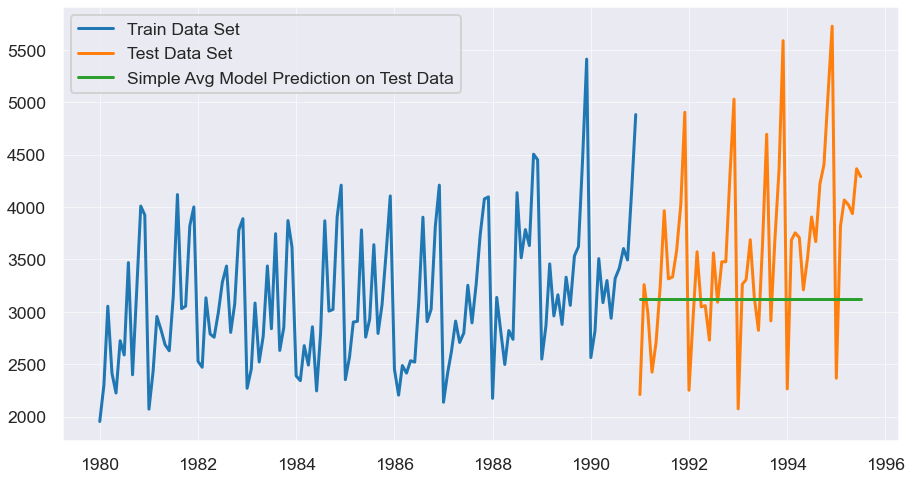
|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 775.808 | 17.417 |
| **Naive Regression Model** | 1519.259 | 44.275 |

**Simple average method:**

Predictions:

|  | **SoftDrinkProduction** | **mean\_forecast** |
| --- | --- | --- |
| **YearMonth** |  |  |
| **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 |

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **SIMPLE AVG Model** | 934.353 | 19.133 |



Above we built a simple average model and we find a RMSE of 934.353 which means linear model is still performing the best. When we plot it on a graph we find that the model has given a constant which is an average of total production. But as there is an increasing trend and seasonality it does not give us the best results.

**Simple exponential model**:

Predictions:

1991-01-01 3656.847753

1991-02-01 3656.847753

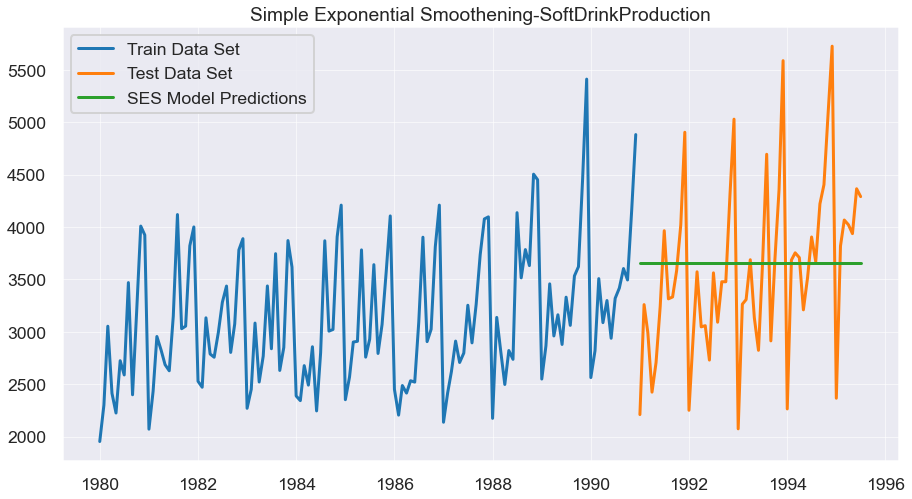
1991-03-01 3656.847753

1991-04-01 3656.847753

1991-05-01 3656.847753

Freq: MS, dtype: float64

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **SES Model, Alpha=0.0987** | 809.502 | 18.997 |



We can see that we after we run our model on the test data we are getting a RMSE of 809.5.2 and a constant of 3656.84.

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 775.808 | 17.417 |
| **Naive Regression Model** | 1519.259 | 44.275 |
| **SIMPLE AVG Model** | 934.353 | 19.133 |
| **SES Model, Alpha=0.0987** | 809.502 | 18.997 |

SES model is second best only to linear regression model.

**Double exponential smoothening – Holt Model:**

1991-01-01 3726.462538

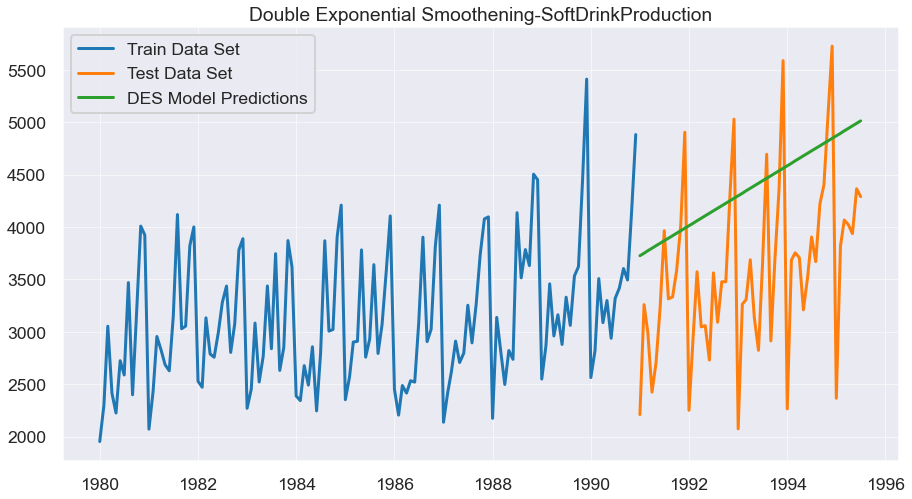
1991-02-01 3750.283901

1991-03-01 3774.105264

1991-04-01 3797.926627

1991-05-01 3821.747990

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 1064.782 | 29.955 |



This model has a RMSE of 1064.782. It give an upward trend after we plot it in a graph in comparison to the actual.

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 775.808 | 17.417 |
| **Naive Regression Model** | 1519.259 | 44.275 |
| **SIMPLE AVG Model** | 934.353 | 19.133 |
| **SES Model, Alpha=0.0987** | 809.502 | 18.997 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 1064.782 | 29.955 |

DES model is not such a good performer in comparison to the other models we have built.

**Holt Winters Model (with seasonality, Trend and Additive):**

Predictions:

1991-01-01 2597.216402

1991-02-01 2909.225289

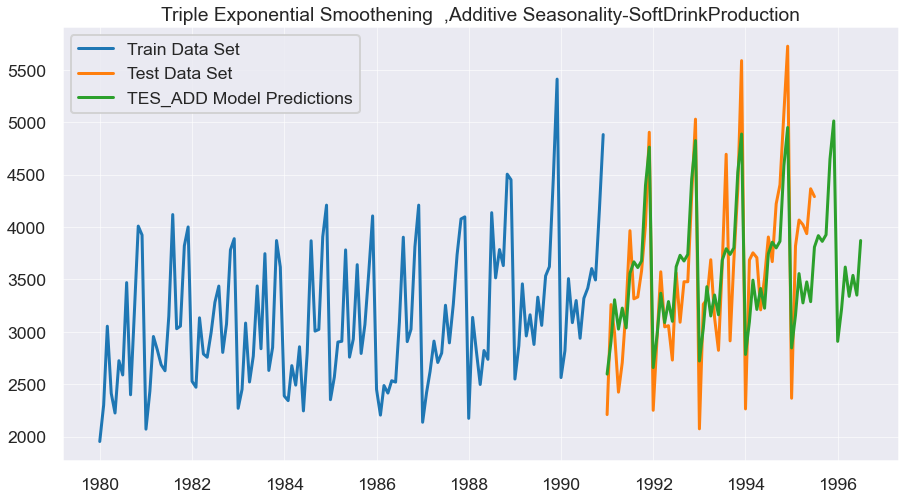
1991-03-01 3306.043928

1991-04-01 3026.503773

1991-05-01 3226.794337

Freq: MS, dtype: float64

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 448.198 | 11.117 |



This model gives us a RMSE of 448.198. And when plotted on a graph we find it include seasonality.

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 775.808 | 17.417 |
| **Naive Regression Model** | 1519.259 | 44.275 |
| **SIMPLE AVG Model** | 934.353 | 19.133 |
| **SES Model, Alpha=0.0987** | 809.502 | 18.997 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 1064.782 | 29.955 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 448.198 | 11.117 |

So far, this model has given us the best results when we compare MAPE with other models.

**Holt Winters Model (with seasonality, Trend and Multiplicative):**

Predictions:

1991-01-01 2565.454769

1991-02-01 2890.069194

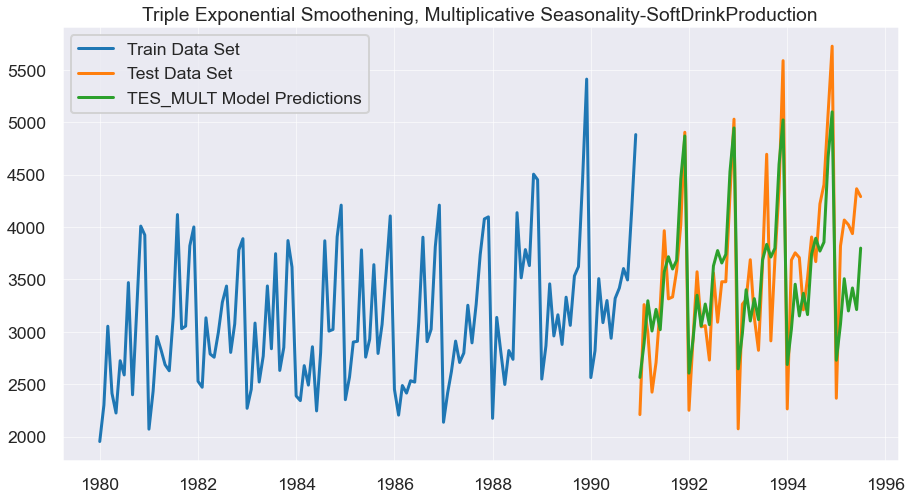
1991-03-01 3296.776543

1991-04-01 3008.139428

1991-05-01 3214.454451

Freq: MS, dtype: float64

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 447.543 | 10.919 |
|  |  |  |



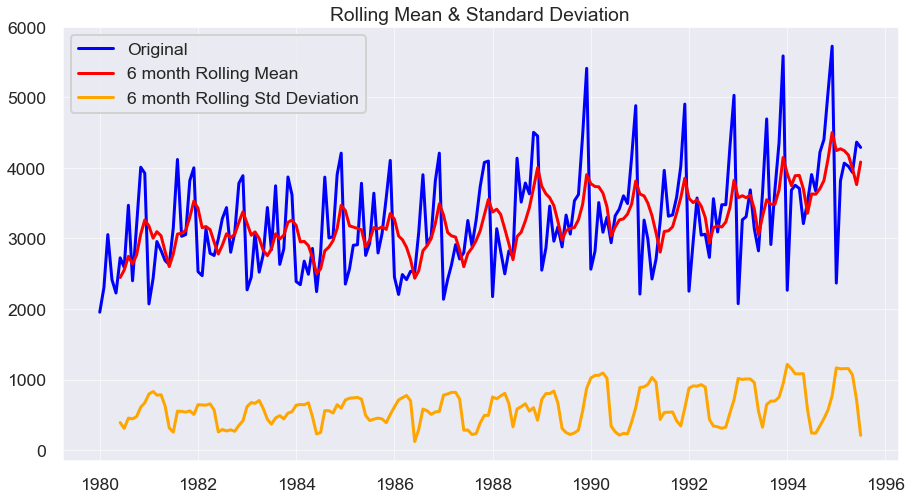
We get a RMSE of 447 and MAPE of 10.9. When we plot prediction vs actuals, we see a lot more accuracy in our predictions as it captures seasonality and trend very well.

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 775.808 | 17.417 |
| **Naive Regression Model** | 1519.259 | 44.275 |
| **SIMPLE AVG Model** | 934.353 | 19.133 |
| **SES Model, Alpha=0.0987** | 809.502 | 18.997 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 1064.782 | 29.955 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 448.198 | 11.117 |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 447.543 | 10.919 |

Thus, after comparing all the models, we have built we observe that Holt Winter model with multiplicative is the best suited for predictions. It has an RMSE of 447.5 and a MAPE of 10.9. Thus, we have least errors and this model is able to explain the seasonality and trend better than any other model.

1. 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.

Let us check for stationarity of the data:



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

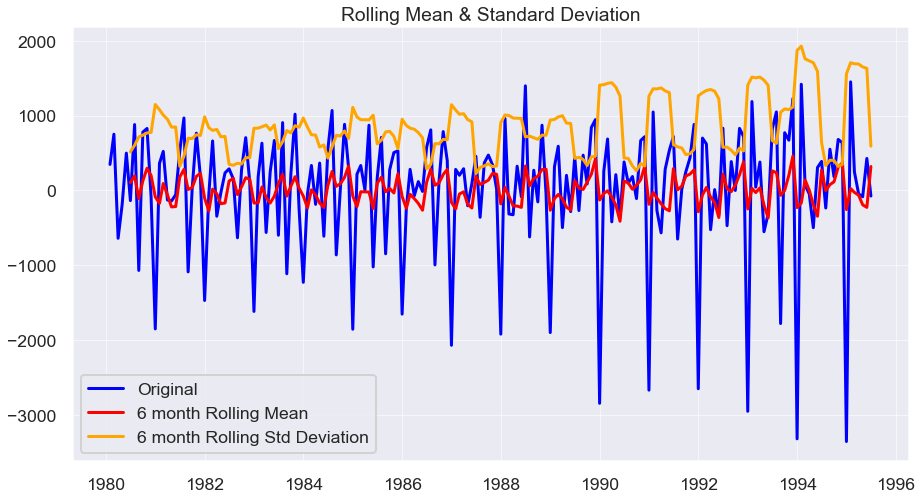
Critical Value (5%) -2.878298

Critical Value (10%) -2.575704

dtype: float64

Here we find that the p value is 0.99 and is thus greater than 0.05 and therefore the data given is not stationary.

Let us try to make this data stationary using one level differencing (d=1).



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

p value is now less than alpha of 0.5. Hence the data now is stationary.

1. 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.

**ARIMA Model:**

|  | **param** | **AIC** |
| --- | --- | --- |
| **2** | (0, 1, 2) | 2056.49 |
| **6** | (1, 1, 2) | 2056.72 |
| **3** | (0, 1, 3) | 2056.83 |
| **11** | (2, 1, 3) | 2057.09 |
| **13** | (3, 1, 1) | 2058.30 |

Above are the parameters and Akaike Information Criteria in the ascending order.

SARIMAX Results

===============================================================================

Dep. Variable: SoftDrinkProduction No. Observations: 132

Model: ARIMA(2, 1, 3) Log Likelihood -1022.545

Date: Wed, 25 Jan 2023 AIC 2057.090

Time: 09:04:55 BIC 2074.341

Sample: 01-01-1980 HQIC 2064.100

- 12-01-1990

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 0.7384 0.193 3.824 0.000 0.360 1.117

ar.L2 0.1606 0.220 0.730 0.465 -0.271 0.592

ma.L1 -1.3910 0.214 -6.500 0.000 -1.810 -0.972

ma.L2 -0.1869 0.338 -0.554 0.580 -0.849 0.475

ma.L3 0.5939 0.180 3.300 0.001 0.241 0.947

sigma2 3.258e+05 8.46e-07 3.85e+11 0.000 3.26e+05 3.26e+05

===================================================================================

Ljung-Box (L1) (Q): 0.00 Jarque-Bera (JB): 1.28

Prob(Q): 0.96 Prob(JB): 0.53

Heteroskedasticity (H): 1.19 Skew: 0.15

Prob(H) (two-sided): 0.56 Kurtosis: 2.62

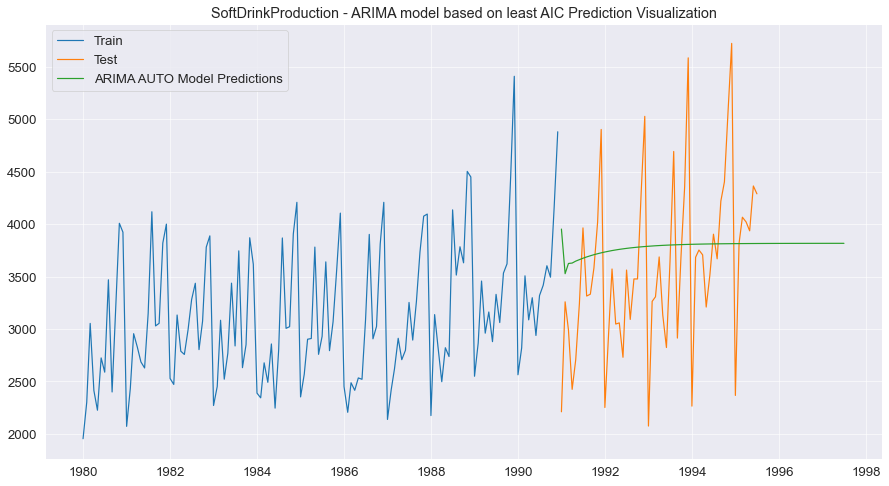
===================================================================================

Warnings:

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

[2] Covariance matrix is singular or near-singular, with condition number 1.23e+28. Standard errors may be unstable.

We have chosen parameter of (2,1,3) on training data to build ARIMA model.



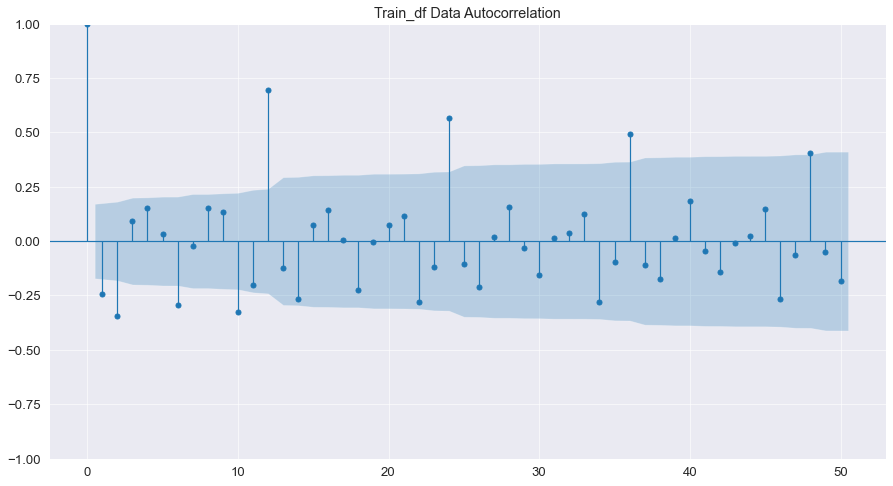
We have plotted our predictions made by ARIMA model on a graph vs actual.

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 775.808000 | 17.417000 |
| **Naive Regression Model** | 1519.259000 | 44.275000 |
| **SIMPLE AVG Model** | 934.353000 | 19.133000 |
| **SES Model, Alpha=0.0987** | 809.502000 | 18.997000 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 1064.782000 | 29.955000 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 448.198000 | 11.117000 |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 447.543000 | 10.919000 |
| **ARIMA\_AUTO(2,1,3)** | 816.455897 | 19.952546 |
|  |  |  |

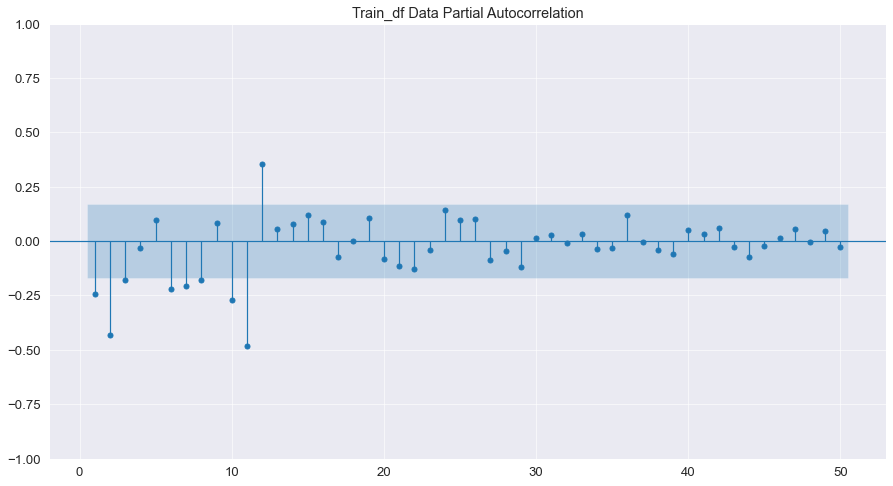
We find ARIMA model has a RMSE of 816 and MAPE of 19.9. Thus it is not very well suited for predictions.

**SARIMA model based on least AIC:**

ACF:



PACF:



Parameters, seasonality and AIC in ascending order:

|  | **param** | **seasonal** | **AIC** |
| --- | --- | --- | --- |
| **28** | (0, 1, 1) | (3, 0, 0, 12) | 1391.515931 |
| **29** | (0, 1, 1) | (3, 0, 1, 12) | 1391.678481 |
| **44** | (0, 1, 2) | (3, 0, 0, 12) | 1392.092390 |
| **45** | (0, 1, 2) | (3, 0, 1, 12) | 1392.889692 |
| **30** | (0, 1, 1) | (3, 0, 2, 12) | 1393.254223 |

We have taken seasonality as 12 and d = 1.

SARIMAX Results

===============================================================================================

Dep. Variable: SoftDrinkProduction No. Observations: 132

Model: SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12) Log Likelihood -668.639

Date: Wed, 25 Jan 2023 AIC 1357.277

Time: 09:06:59 BIC 1382.495

Sample: 01-01-1980 HQIC 1367.456

- 12-01-1990

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 0.1000 0.143 0.701 0.483 -0.180 0.380

ar.L2 -0.0261 0.126 -0.207 0.836 -0.273 0.221

ar.L3 0.0777 0.147 0.529 0.597 -0.210 0.366

ma.L1 -0.9213 0.086 -10.775 0.000 -1.089 -0.754

ar.S.L12 0.5940 0.750 0.792 0.429 -0.877 2.065

ar.S.L24 0.3046 0.700 0.435 0.663 -1.067 1.676

ar.S.L36 0.0913 0.243 0.376 0.707 -0.385 0.568

ma.S.L12 -0.2144 0.764 -0.281 0.779 -1.711 1.283

ma.S.L24 -0.1275 0.476 -0.268 0.789 -1.061 0.806

sigma2 1.175e+05 1.66e+04 7.076 0.000 8.5e+04 1.5e+05

===================================================================================

Ljung-Box (L1) (Q): 0.00 Jarque-Bera (JB): 5.85

Prob(Q): 0.99 Prob(JB): 0.05

Heteroskedasticity (H): 1.24 Skew: 0.41

Prob(H) (two-sided): 0.55 Kurtosis: 3.92

===================================================================================

Warnings:

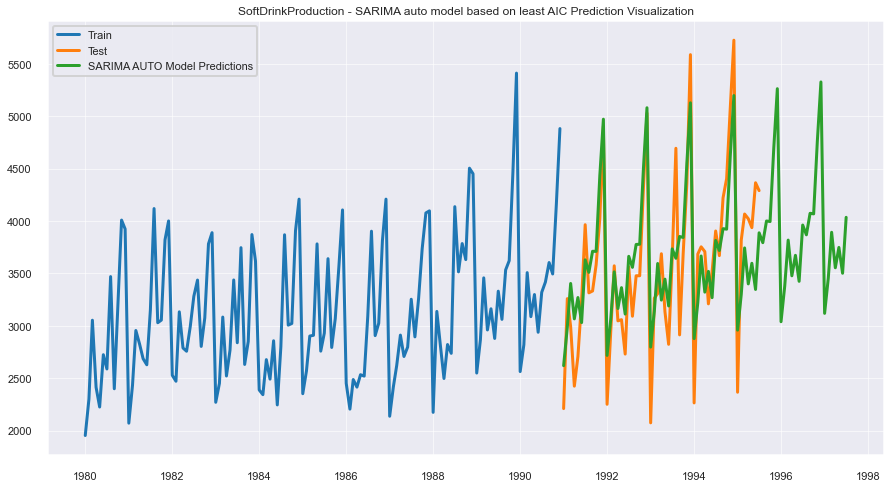
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Let us check the RMSE and MAPE of this model and compare it with others.

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 775.808000 | 17.417000 |
| **Naive Regression Model** | 1519.259000 | 44.275000 |
| **SIMPLE AVG Model** | 934.353000 | 19.133000 |
| **SES Model, Alpha=0.0987** | 809.502000 | 18.997000 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 1064.782000 | 29.955000 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 448.198000 | 11.117000 |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 447.543000 | 10.919000 |
| **ARIMA\_AUTO(2,1,3)** | 816.455897 | 19.952546 |
| **SARIMA\_AUTO(3,1,1)(3,0,2,12)** | 427.633963 | 10.875734 |

WE see that SARIMA model has a RMSE of 427.67 and MAPE of 10.87. This model has given us even better results than Holt Winter multiplicative model.

Hence this model is best for predicting future production of soft drinks.



1. Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

We have used p=2,d=1,q=2 in our model this time.

SARIMAX Results

===============================================================================

Dep. Variable: SoftDrinkProduction No. Observations: 132

Model: ARIMA(2, 1, 2) Log Likelihood -1024.356

Date: Wed, 25 Jan 2023 AIC 2058.713

Time: 09:07:03 BIC 2073.089

Sample: 01-01-1980 HQIC 2064.554

- 12-01-1990

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 -0.2838 0.304 -0.933 0.351 -0.880 0.312

ar.L2 -0.0078 0.179 -0.044 0.965 -0.359 0.343

ma.L1 -0.3008 0.263 -1.145 0.252 -0.816 0.214

ma.L2 -0.6072 0.259 -2.343 0.019 -1.115 -0.099

sigma2 3.515e+05 4.68e+04 7.509 0.000 2.6e+05 4.43e+05

===================================================================================

Ljung-Box (L1) (Q): 0.08 Jarque-Bera (JB): 0.36

Prob(Q): 0.78 Prob(JB): 0.83

Heteroskedasticity (H): 1.29 Skew: 0.04

Prob(H) (two-sided): 0.40 Kurtosis: 2.75

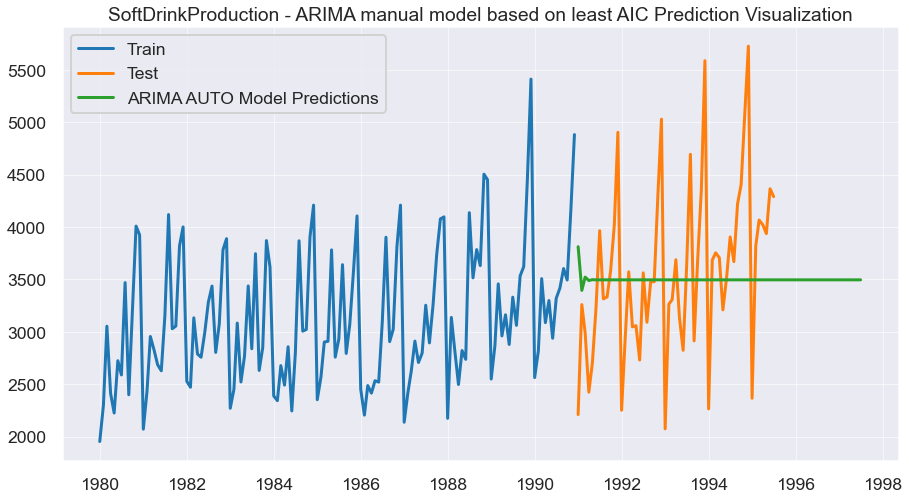
===================================================================================

Warnings:

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

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 775.808000 | 17.417000 |
| **Naive Regression Model** | 1519.259000 | 44.275000 |
| **SIMPLE AVG Model** | 934.353000 | 19.133000 |
| **SES Model, Alpha=0.0987** | 809.502000 | 18.997000 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 1064.782000 | 29.955000 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 448.198000 | 11.117000 |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 447.543000 | 10.919000 |
| **ARIMA\_AUTO(2,1,3)** | 816.455897 | 19.952546 |
| **SARIMA\_AUTO(3,1,1)(3,0,2,12)** | 427.633963 | 10.875734 |
| **ARIMA\_MANUAL(2,1,2)** | 823.030515 | 18.382282 |

With the help of above chart we can see that this model has not worked best as it has a RMSE of 823 and MAPE of 18.3.



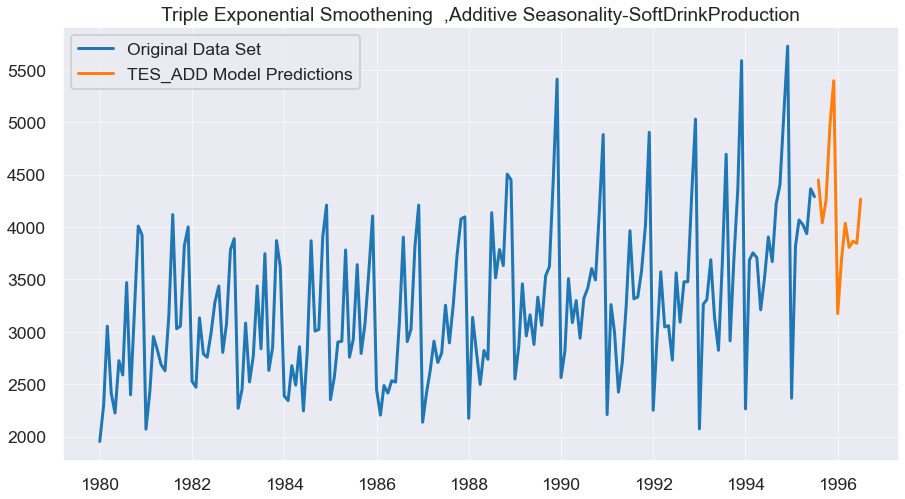
Above is the graph of ARIMA manual model of predictions in comparison of actuals.

1. Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **SARIMA\_AUTO(3,1,1)(3,0,2,12)** | 427.6 | 10.9 |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 447.5 | 10.9 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 448.2 | 11.1 |
| **SARIMA\_MANUAL(4,0,2)(1,1,2,12)** | 582.6 | 11.8 |
| **Linear Regression Model** | 775.8 | 17.4 |
| **SES Model, Alpha=0.0987** | 809.5 | 19.0 |
| **ARIMA\_AUTO(2,1,3)** | 816.5 | 20.0 |
| **ARIMA\_MANUAL(2,1,2)** | 823.0 | 18.4 |
| **SIMPLE AVG Model** | 934.4 | 19.1 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 1064.8 | 30.0 |
| **Naive Regression Model** | 1519.3 | 44.3 |

1. 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.

We will be selecting Triple exponential smoothening with additive seasonality and additive trend as it has the least RMSE.



|  | **Predicted Sales** |
| --- | --- |
| **1995-08-01** | 4448.99 |
| **1995-09-01** | 4040.23 |
| **1995-10-01** | 4254.00 |
| **1995-11-01** | 4976.74 |
| **1995-12-01** | 5395.29 |
| **1996-01-01** | 3174.00 |
| **1996-02-01** | 3698.82 |
| **1996-03-01** | 4035.87 |
| **1996-04-01** | 3805.82 |
| **1996-05-01** | 3864.05 |
| **1996-06-01** | 3845.06 |
| **1996-07-01** | 4266.19 |

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

Thus we see that production will shoot up in the months of November and December.  
They should bolster there logistics for this period of time so that they do not face any issue.

The increase in sales could be because of the holiday and festive season during the year end and hence the production during that period of time also increases a lot to fill in the demand.