**Soft Drink Production TS Analysis**

Name: Saurabh Dharmadhikari

PGP-DSBA Online January’ 21

Date: 21/11/2022

Table of Contents:

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

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

Data Description ............................................................................................................................................................3

Sample of the dataset ...................................................................................................................................................4

Exploratory Data Analysis..............................................................................................................................................5

Let us check the types of variables in the data frame. .................................................................................................5

Check for missing values in the dataset .......................................................................................................................5

2.1) Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis. .........................................7

2.2) Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis). ......................................................................................12

2.3) Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Compare both the models and write inferences, which model is best/optimized. …………………………………………………………………………………………………………...................15

2.4) Inference: Based on these predictions, what are the insights and recommendations. .....................................18

Conclusion & Insights ..................................................................................................................................................18

THE END!......................................................................................................................................................................18

Executive Summary

We as an analyst in the RST soft drink company 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.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. The data consists of 187 entries, we have an entry of quantity produced for each month starting from January 1980. We are also provided with units produced over the years till July 1995. Our mission is to build a model that can predict production of soft drinks using a time series analysis. For this we will be developing models such as regression, naïve forecast models, simple average also check the performance on the test data using RMSE. We will also check the stationarity of the data using appropriate statistical test and if found that the data is not stationary we will take appropriate steps to make it stationary.

Data Description

**YearMonth:** It is the month and the year of production.

**SoftDrinkProduction:** Number of units produced for the month and year.

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 |

Above are 5 initial rows of the dataset.

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 |

Above are last 5 observation of the dataset.

Checking null values and number of observations.

Data columns (total 1 columns):

# Column Non-Null Count Dtype

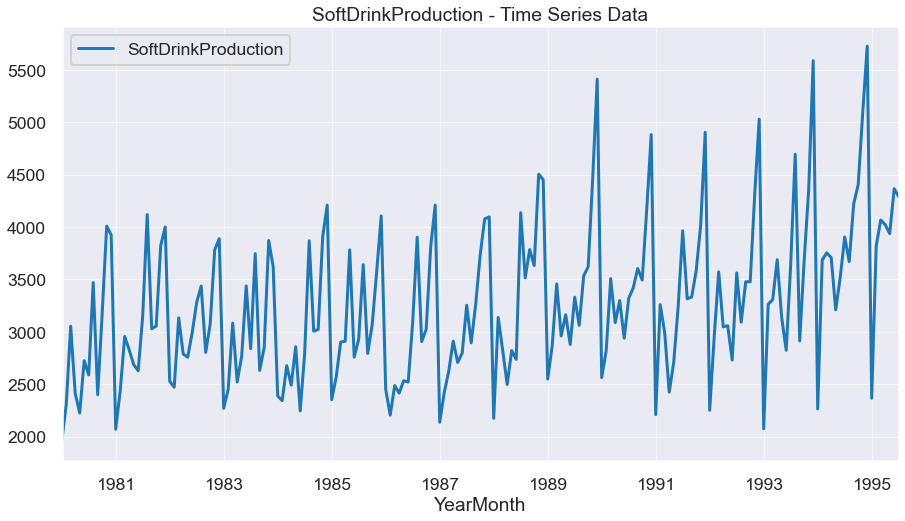
--- ------ -------------- -----

0 SoftDrinkProduction 187 non-null int64

dtypes: int64(1)

memory usage: 2.9 KB

There are no null values and a total of 187 entries are present in our dataset.



The graph above shows us the monthly production of soft drinks over the years starting from January 1980 till July 1995.

We can observe here that there is a seasonality in the production of soft drinks as the production increases during end of each year. Also, we find that the production started to show spikes in increase each year after December 1989.

Basic statistical analysis:

|  | **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 |

We find that average production of soft drink from 1980 to 1985 has been about 3263, from minimum of 1954 to a maximum of 5725.

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

Data columns (total 1 columns):

# Column Non-Null Count Dtype

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

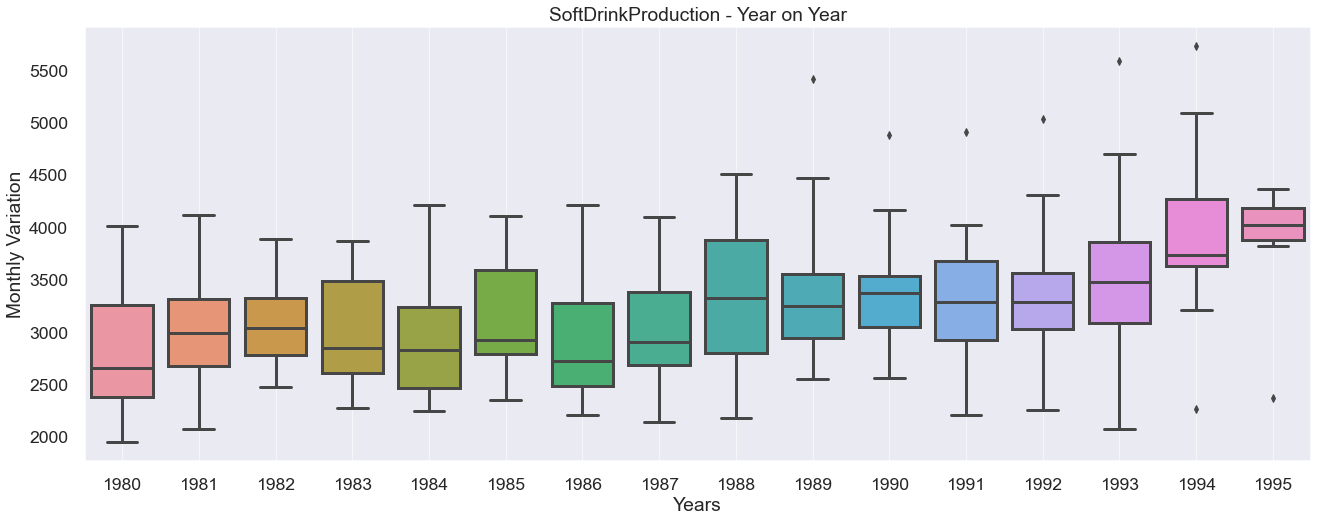
0 SoftDrinkProduction 187 non-null int64

dtypes: int64(1)

memory usage: 2.9 KB

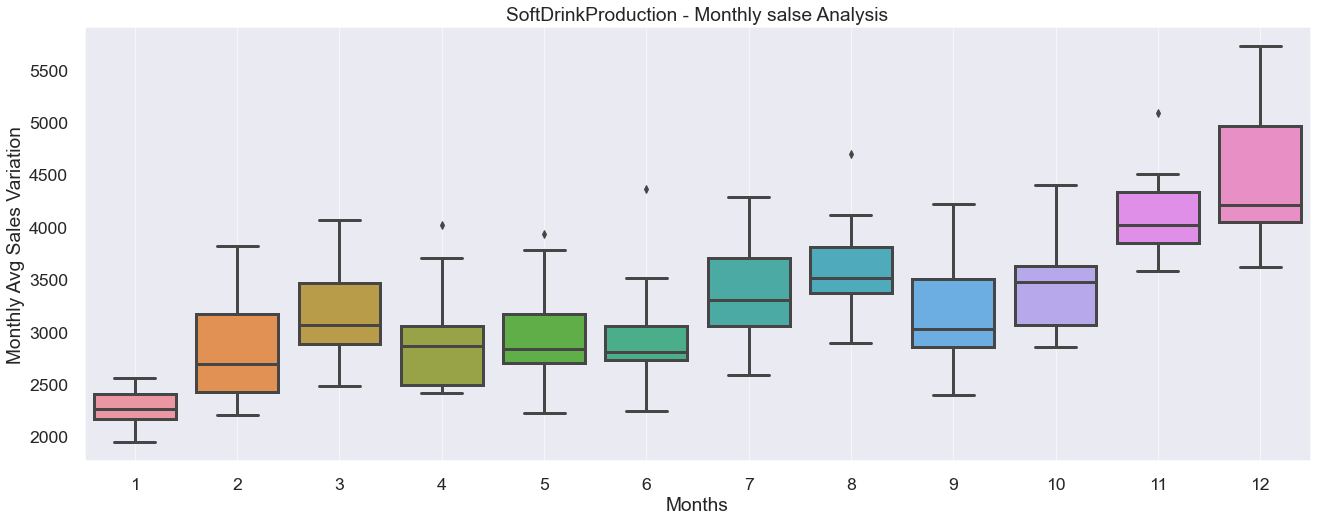
We have only one column because we are treating YearMonth column as index. Column SoftDrinkProduction of integer type data.

Below we have a graph of yearly production of soft drinks with mean of each year and maximum and minimum production in that year.



We can observe that there are a few outliers from year 1989 as we had addressed this earlier that there are spikes in in the production level starting form 1989. But as these outliers are very few and they are true records we will keep this data so that we get much accurate results with our model. We are not treating the outliers.

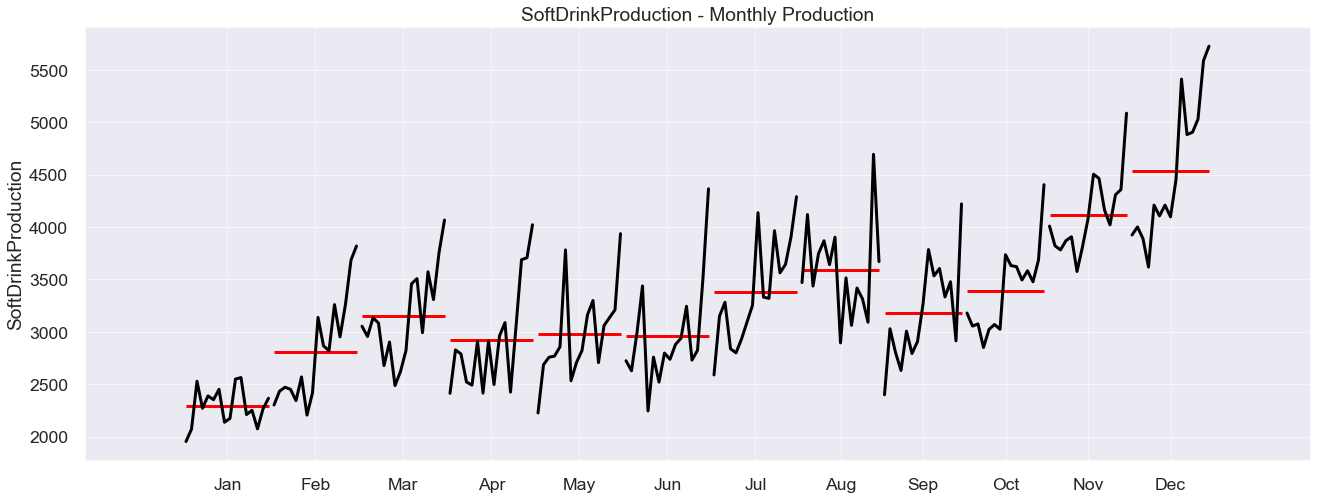
Below is the monthly box plot of the dataset. We can see the mean of each month with max and min production of the month over the years.



With the above graph we can visualize that average monthly production in December has been that highest as we had discussed earlier. Minimum production is in the month of January.

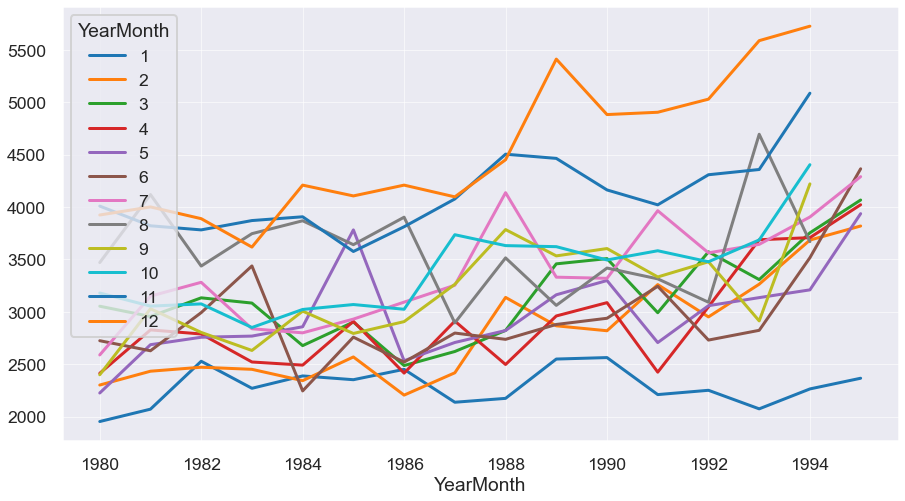
There is a strong case of seasonality.

Below we have a plot of monthly production for each month:



Here we can observe an increasing trend over the years for each month. There were a few drops in trend but then latter it started to rise again.

Below we have a graph for each month:



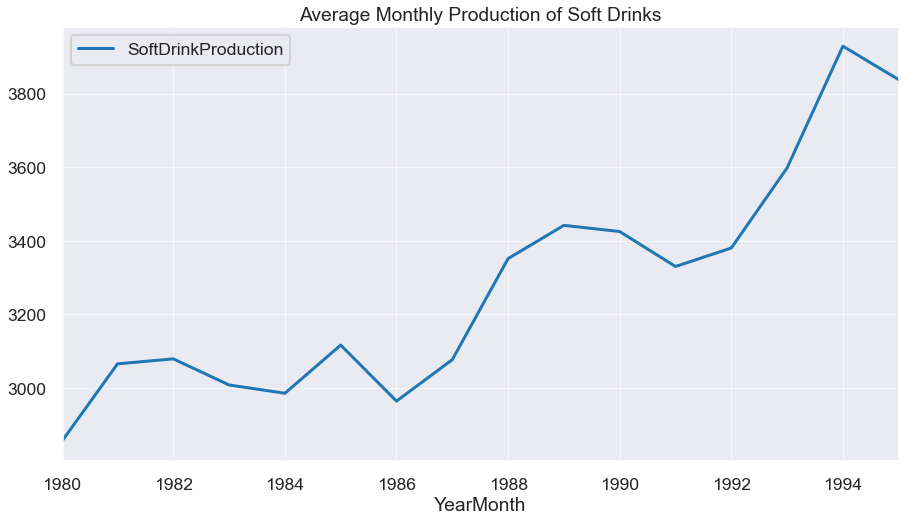
SoftDrinkProduction - Year on Year monthly Production

| **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 |

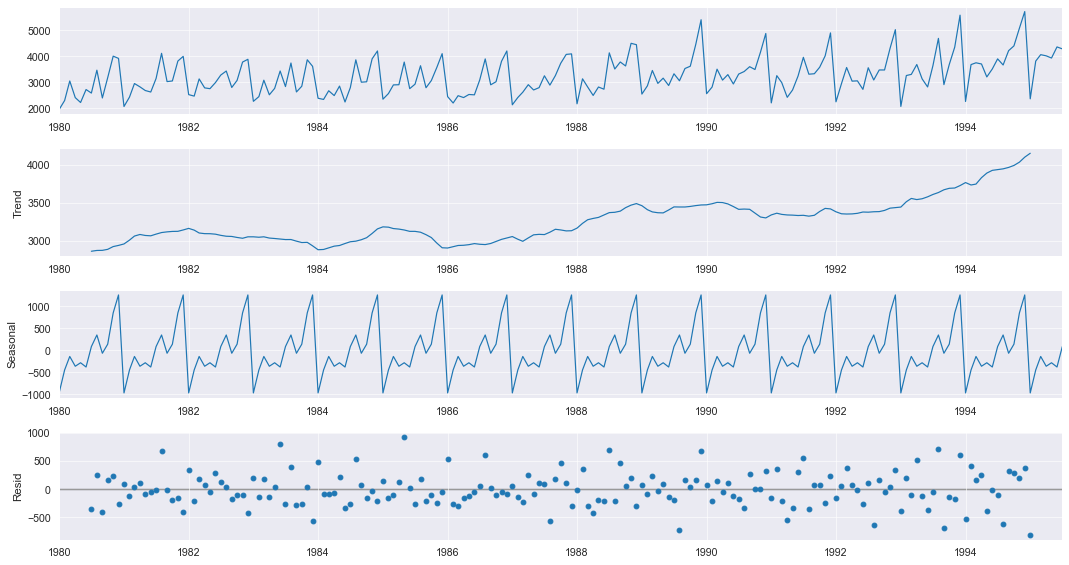
Above we have arranged data for each year in rows and months in column to easily understand the data and compare with exact numbers if required.



Above graph shows us aggregate of yearly production of soft drinks and its trend. As data for year 1995 is not complete we see a sharp decline but otherwise trend has only been increasing.



We can also see the average monthly production each year in the above graph.



**Additive decomposition**

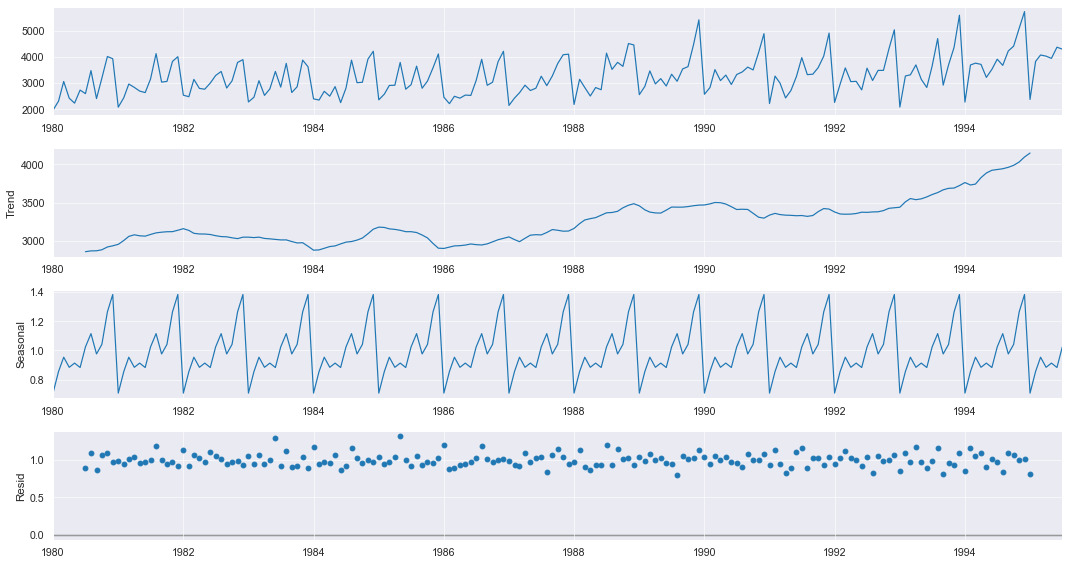
Above is the additive decomposition of the data.

We can see the trend is increasing, also there is a seasonality. Residual is in the range of +/- 1000.

Seasonality has a range of +/- 1000.

Trend has a range of 2500 to 4000.

Seasonality and residual have a similar range in the additive model.



**Multiplicative decomposition**

Above is the multiplicative decomposition of the data.

We can see the trend is increasing, also there is a seasonality. Residual has a variation of 1.5 to 0.5 times.

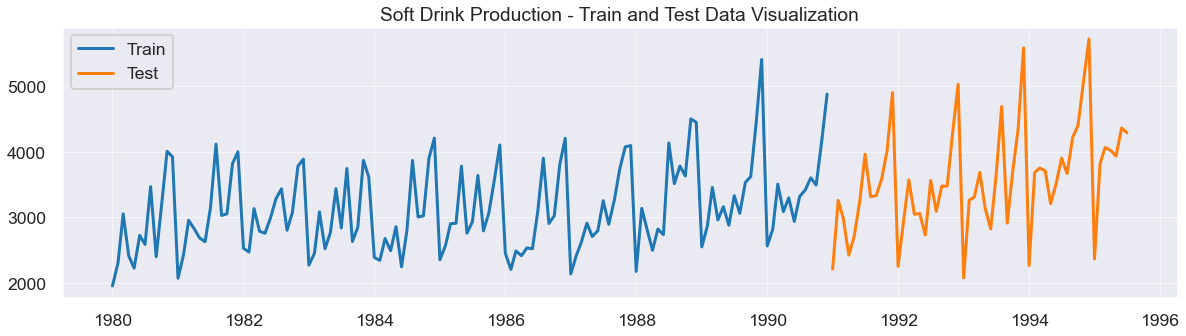
Seasonality has a variation of 0.6 to 1.4 times.

Trend has a range of 2500 to 4000.

Seasonality and trend explain a lot of variations.

Residual in case of multiplicative model is lesser than additive and thus this model is able to explain a lot of variances.

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



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

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

In the graph we can see how the dataset is now divided as train and test with blue live indicating train set and orange line indicating test set. So, it is clear that the data set has been divided at 1991.

Also, the first 12 observations have been noted for both train and test set and we see that the test set starts from January 1991.

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:**

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

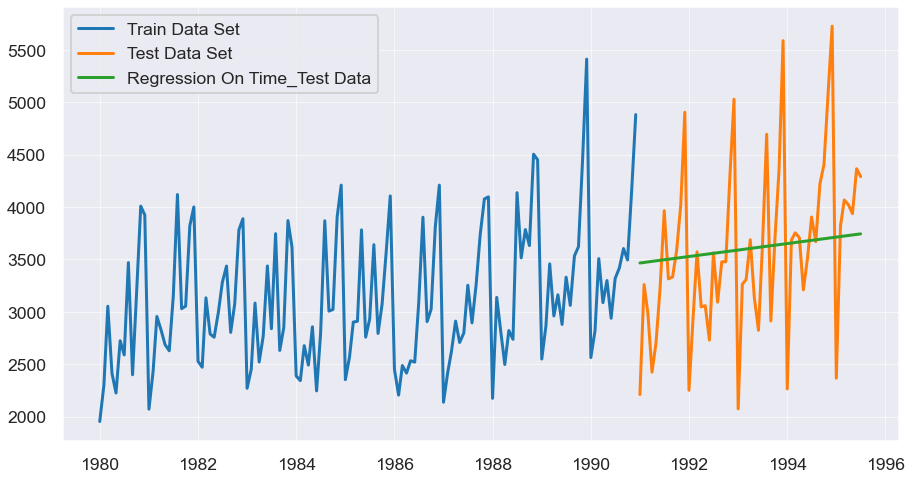
Name: RegOnTime, dtype: float64

When we see the first few prediction we find that the predictions are increasing in values.

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

We see that Root mean squared error is 775.808 and Mean absolute percentage error is 17.417.

Let us see how our predictions will show up in a graph.



Above we have our predictions in in graphical form the green line represents our predictions.   
We can see that there is a lot of variances from predictions and actual values but still we have been able to predict on an average and we have model to forecast.

Naïve Model:

YearMonth

1991-01-01 4882

1991-02-01 4882

1991-03-01 4882

1991-04-01 4882

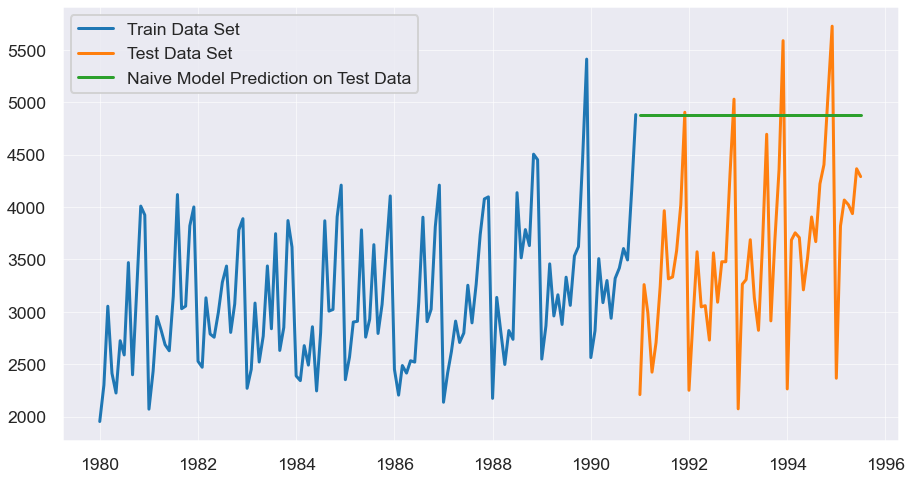
1991-05-01 4882

Name: naive, dtype: int64

We can see here that the predictions are very much constant. Let us explore more.

|  | **SoftDrink\_TEST RMSE** | **SoftDrink\_MAPE** |
| --- | --- | --- |
| **Naive Regression Model** | 1519.259 | 44 |

We see that Root mean squared error is 1519.259 and Mean absolute percentage error is 44.



We can see from the above graph that the predictions are in a straight line that means that predictions are constant for each year and that stands false because each year we have a different actual production.

Let us compare linear regression model with naïve model:

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

We observe that linear regression model scores less in its error and thus is able to predict far better than naïve model.

Simple average Model:

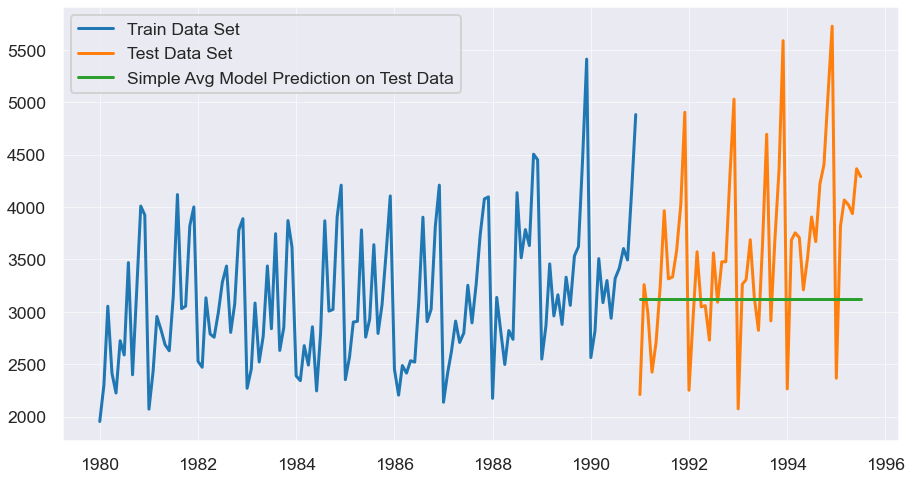
|  | **SoftDrinkProduction** | **mean\_forecast** |
| --- | --- | --- |
| **YearMonth** |  |  |
| **1991-01-01** | 2211 | 3124.166667 |
| **1991-02-01** | 3260 | 3124.166667 |
| **1991-03-01** | 2992 | 3124.166667 |
| **1991-04-01** | 2425 | 3124.166667 |
| **1991-05-01** | 2707 | 3124.166667 |

Above chart gives us the first few predictions made by our model they look to be all constants again.

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

Here we have a RMSE 934.353 and MAPE score of 19.133.

Let us plot our prediction in comparison with actuals.



We see that as this is a simple average of the actuals the predictions are constant in a straight line denoted by a green line.

Let us compare the scores with our other two models:

|  | **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 |

We find that linear regression model that we prepared is still giving us the best predictions as its error score is lesser than the other two.

Simple Exponential Model:

Predictions for test data of SoftDrinkProduction

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

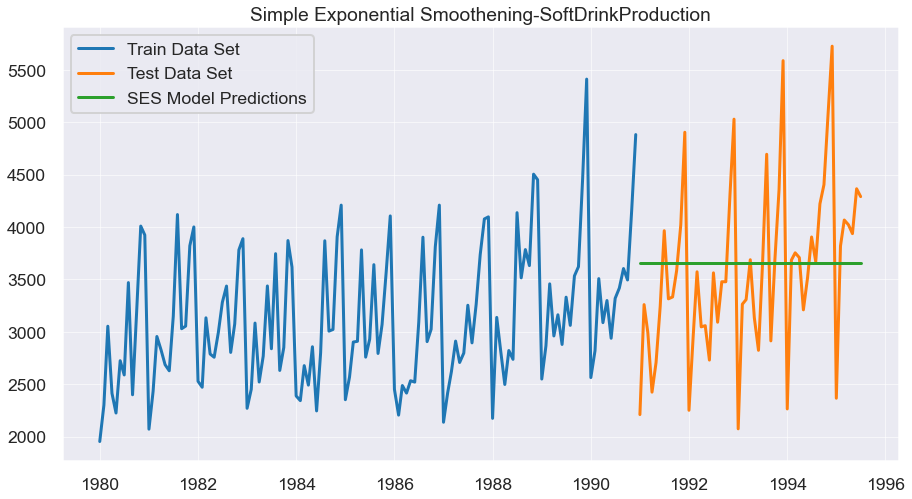
Freq: MS, dtype: float64

We again see that there is a single value that continues as predictions and thus showing a constant.

Let us find more about it.

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

There is a RMSE of 809.502 and a MAPE of 18.997.



We see a straight line as green line here in the graph when predictions are plotted with actuals.

Let’s compare other models in respect of Simple Exponential Model.

|  | **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 |

We find that linear regression is still doing a good job with predicting the production of soft drinks.

Double Exponential Smoothening – Holt Model:

Predictions for test data of SoftDrinkProduction

1991-01-01 3726.462538

1991-02-01 3750.283901

1991-03-01 3774.105264

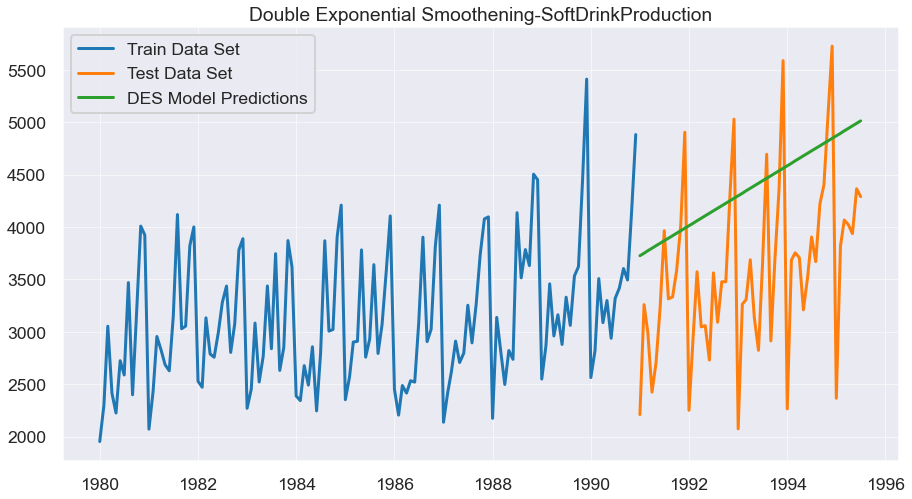
1991-04-01 3797.926627

1991-05-01 3821.747990

Above we have a chart of predictions made by our model.

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

Above are the RMSE and MAPE of this model.



Above we see the predictions plotted on a graph.

Let us compare other models with DES model:

|  | **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 |

Again, linear regression model is better than other models we have built.

Holt Winters Model (With trend and seasonality, additive):

Predictions for test data of SoftDrinkProduction

1991-01-01 2597.216402

1991-02-01 2909.225289

1991-03-01 3306.043928

1991-04-01 3026.503773

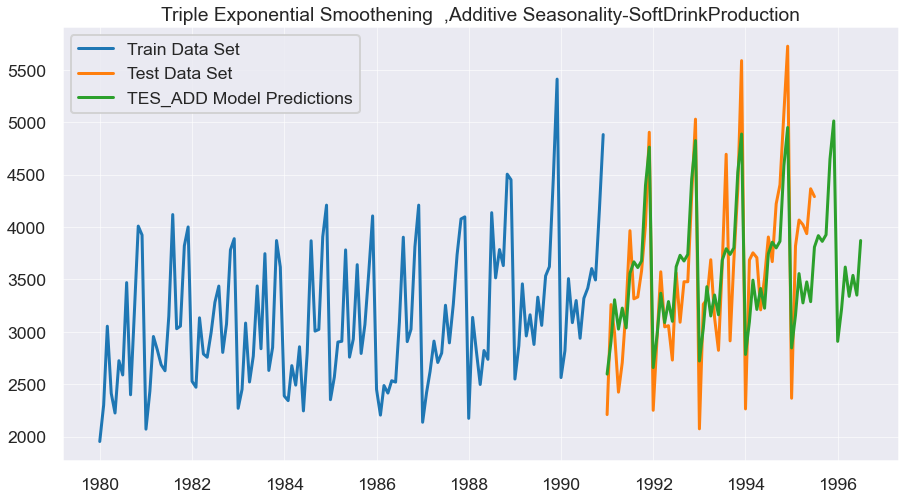
1991-05-01 3226.794337

Freq: MS, dtype: float64

Above are the predictions of our new model.

Let us see the RMSE and MAPE for this model to check its accuracy.

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



We have plotted a graph of our predictions with actuals.

|  | **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 |

Here we can see that Triple Exponential Smoothening(additive) model has the best predictions for our dataset.

We will go ahead to build a model with multiplicative seasonality.

Holt Winters Model (With trend and seasonality, multiplicative):

Predictions for test data of SoftDrinkProduction

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

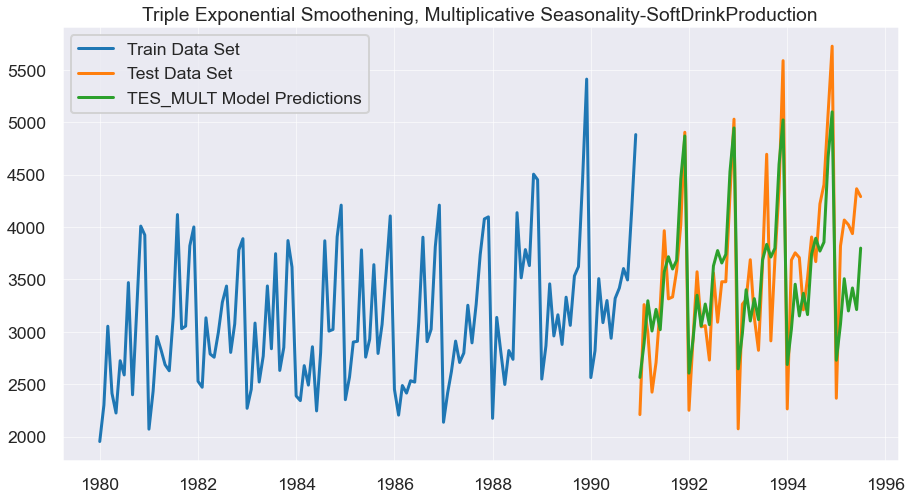
Freq: MS, dtype: float64

Above predictions are for Triple Exponential Smoothening(multiplicative) model.

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

Here we see RMSE and MAPE for our model.

Let us see our predictions on a graph in comparison to the actuals:

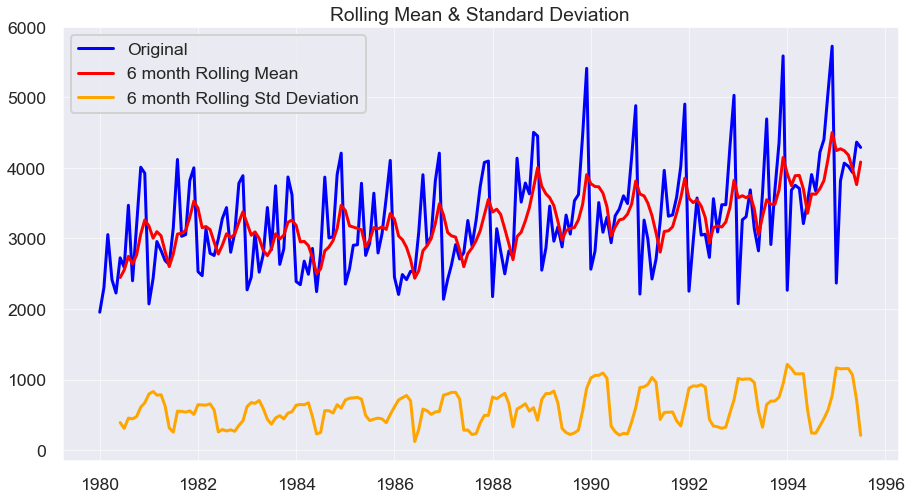


We observe that there are far better results with this model.

|  | **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 |

We can see from the above chart that this model has given us the best predictions and we are able to explain a lot of discrepancies and thus is giving us the best results.

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.



2.4) Inference: Based on these predictions, what are the insights and recommendations.

Using logistic regression model, we can say

For {Passengers who did not survive (Label 0)}:

Precision (90%) – 90% of passengers who did not survive are correctly predicted, out of all passengers who did not survive that are predicted.

Recall (70%) – Out of all the passengers who actually did not survive, 70% of passengers who did not survive have been predicted correctly.

For {Passengers who did survive (Label 1)}:

Precision (97%) – 97% of Passengers who did survive are correctly predicted, out of all passengers who had accident that are predicted.

Recall (99%) – Out of all the passengers who actually did survive, 99% of Customers who did Churn have been correctly predicted.

Accuracy, AUC, Precision and Recall for test data is almost in line with training data. This proves no overfitting or underfitting has happened, and overall, the model is a good model for classification.