**Shoe 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. ….........4

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. ...........................................................................................................................................35

Executive Summary

We are an analyst in the IJK shoe company and we 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.

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 sales of shoes for that month mentioned below in the Data Description. Our mission is to build a model that can predict level of sales 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. Shoe\_Sales: 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.

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

Exploratory data analysis

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

Data columns (total 1 columns):

# Column Non-Null Count Dtype

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

0 Shoe\_Sales 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.

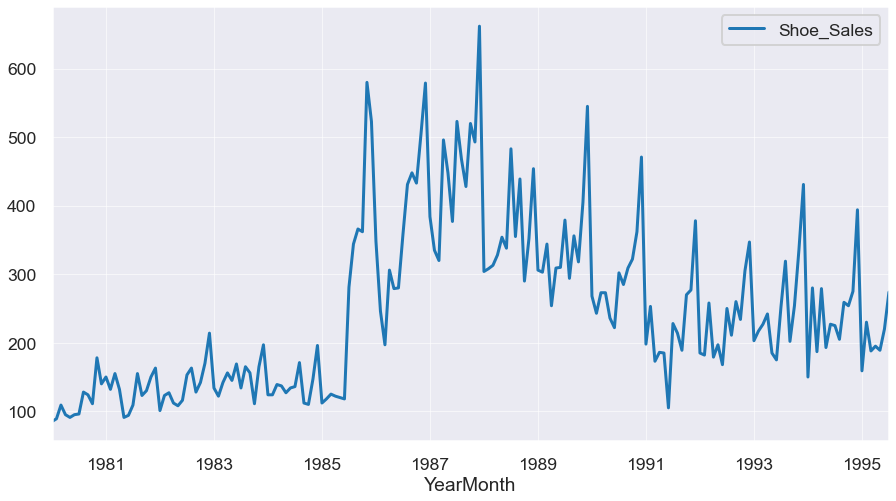
Check for null values

Shoe\_Sales 0

dtype: int64

There are no null values in our data base.

Plotting shoe sales over the years.

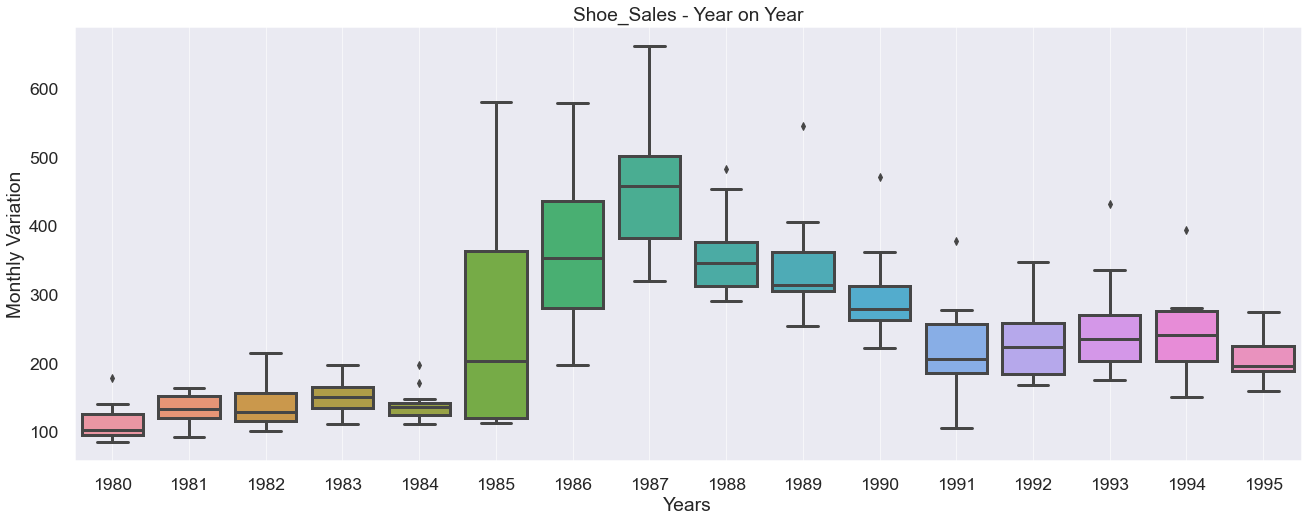


The above graph lets us see the production of shoe sales over the months and years. We can certainly see the rise and fall of production over these years.

|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Shoe\_Sales** | 187.0 | 245.64 | 121.39 | 85.0 | 143.5 | 220.0 | 315.5 | 662.0 |

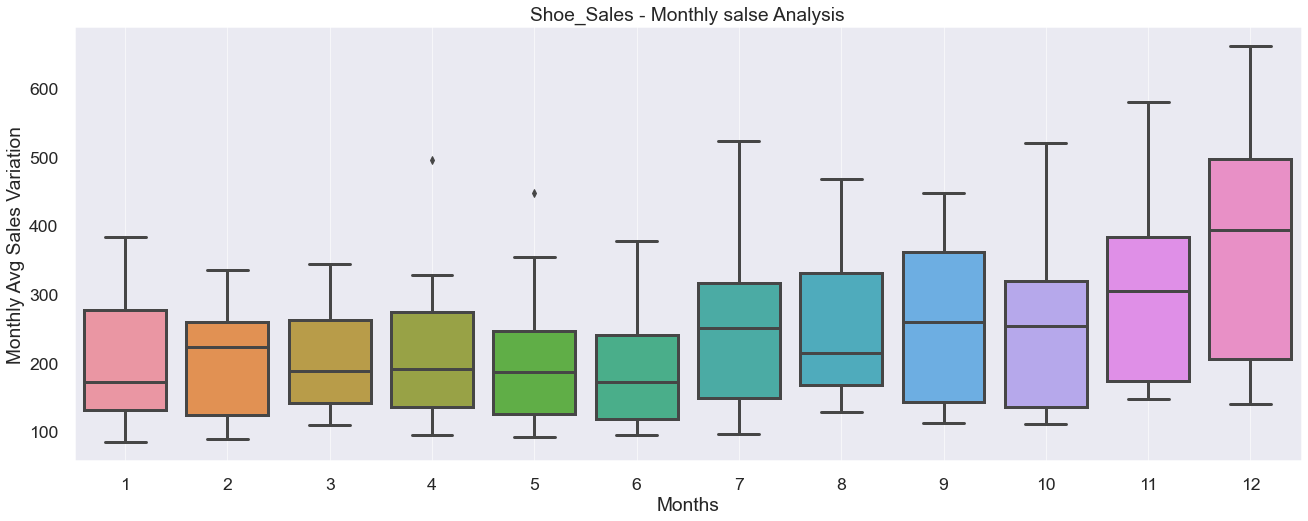
Above we can see the minimum , maximum mean and standard deviation of sales data.

Let us find outliers in the production in the years.



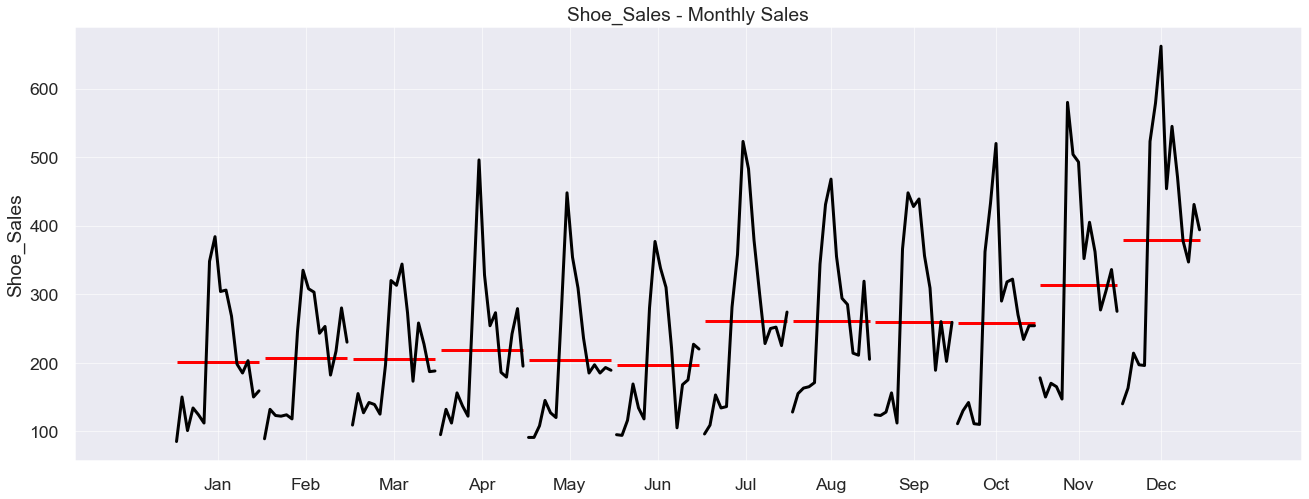
We can see that there are a few months in a few years when the production of shoe sales has been very high. We can also see here that the sale of shoes has been on an increase trend in 1985 and 1986 when we compare the averages production for the years given but there after it came down and now it is quite stable.

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 June. Also, there are only two outliers (April and May). This could be because of winters that the sales are highest in December.

Monthly plot of production for each year:



From the above chart we can clearly see that the sales of shoes for the month of December have remained more same over the years in comparison with the other months. All the months showed upward trend but then the sales went down an d then again went up but still not to the extent of previous years.

| **YearMonth** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **YearMonth** |  |  |  |  |  |  |  |  |  |  |  |  |
| **1980** | 85.0 | 89.0 | 109.0 | 95.0 | 91.0 | 95.0 | 96.0 | 128.0 | 124.0 | 111.0 | 178.0 | 140.0 |
| **1981** | 150.0 | 132.0 | 155.0 | 132.0 | 91.0 | 94.0 | 109.0 | 155.0 | 123.0 | 130.0 | 150.0 | 163.0 |
| **1982** | 101.0 | 123.0 | 127.0 | 112.0 | 108.0 | 116.0 | 153.0 | 163.0 | 128.0 | 142.0 | 170.0 | 214.0 |
| **1983** | 134.0 | 122.0 | 142.0 | 156.0 | 145.0 | 169.0 | 134.0 | 165.0 | 156.0 | 111.0 | 165.0 | 197.0 |
| **1984** | 124.0 | 124.0 | 139.0 | 137.0 | 127.0 | 134.0 | 136.0 | 171.0 | 112.0 | 110.0 | 147.0 | 196.0 |
| **1985** | 112.0 | 118.0 | 125.0 | 122.0 | 120.0 | 118.0 | 281.0 | 344.0 | 366.0 | 362.0 | 580.0 | 523.0 |
| **1986** | 348.0 | 246.0 | 197.0 | 306.0 | 279.0 | 280.0 | 358.0 | 431.0 | 448.0 | 433.0 | 504.0 | 579.0 |
| **1987** | 384.0 | 335.0 | 320.0 | 496.0 | 448.0 | 377.0 | 523.0 | 468.0 | 428.0 | 520.0 | 493.0 | 662.0 |
| **1988** | 304.0 | 308.0 | 313.0 | 328.0 | 354.0 | 338.0 | 483.0 | 355.0 | 439.0 | 290.0 | 352.0 | 454.0 |
| **1989** | 306.0 | 303.0 | 344.0 | 254.0 | 309.0 | 310.0 | 379.0 | 294.0 | 356.0 | 318.0 | 405.0 | 545.0 |
| **1990** | 268.0 | 243.0 | 273.0 | 273.0 | 236.0 | 222.0 | 302.0 | 285.0 | 309.0 | 322.0 | 362.0 | 471.0 |
| **1991** | 198.0 | 253.0 | 173.0 | 186.0 | 185.0 | 105.0 | 228.0 | 214.0 | 189.0 | 270.0 | 277.0 | 378.0 |
| **1992** | 185.0 | 182.0 | 258.0 | 179.0 | 197.0 | 168.0 | 250.0 | 211.0 | 260.0 | 234.0 | 305.0 | 347.0 |
| **1993** | 203.0 | 217.0 | 227.0 | 242.0 | 185.0 | 175.0 | 252.0 | 319.0 | 202.0 | 254.0 | 336.0 | 431.0 |
| **1994** | 150.0 | 280.0 | 187.0 | 279.0 | 193.0 | 227.0 | 225.0 | 205.0 | 259.0 | 254.0 | 275.0 | 394.0 |
| **1995** | 159.0 | 230.0 | 188.0 | 195.0 | 189.0 | 220.0 | 274.0 | NaN | NaN | NaN | NaN | NaN |

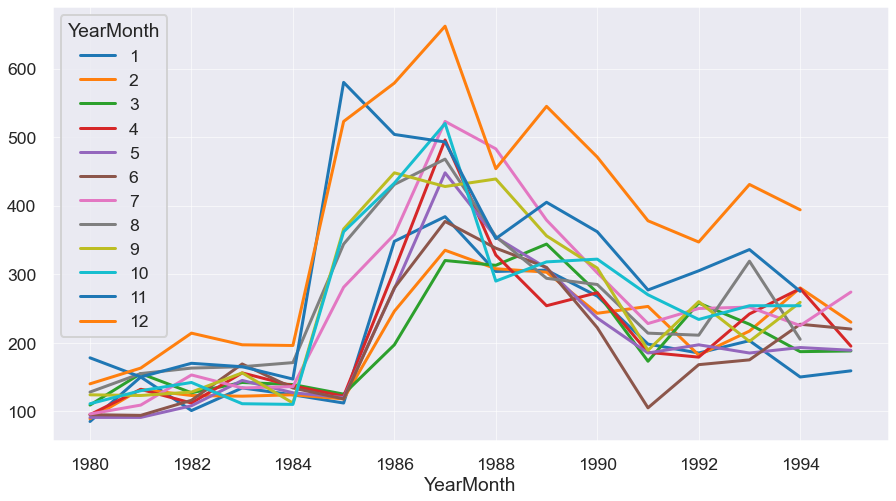
In the above chart we can also see the shoes sold with a calendar format to compare them numerically.

Average monthly production:

Shoe\_Sales - average monthly sales

| **YearMonth** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Shoe\_Sales** | 201.0 | 207.0 | 205.0 | 218.0 | 204.0 | 197.0 | 261.0 | 261.0 | 260.0 | 257.0 | 313.0 | 380.0 |

Plotting monthly production over the years:



As discussed before there is an upward trend year 1987 and thereafter it started to fall and after 1992 have stabilised and have remained in the range of 150 to 280.

Total production for each year:

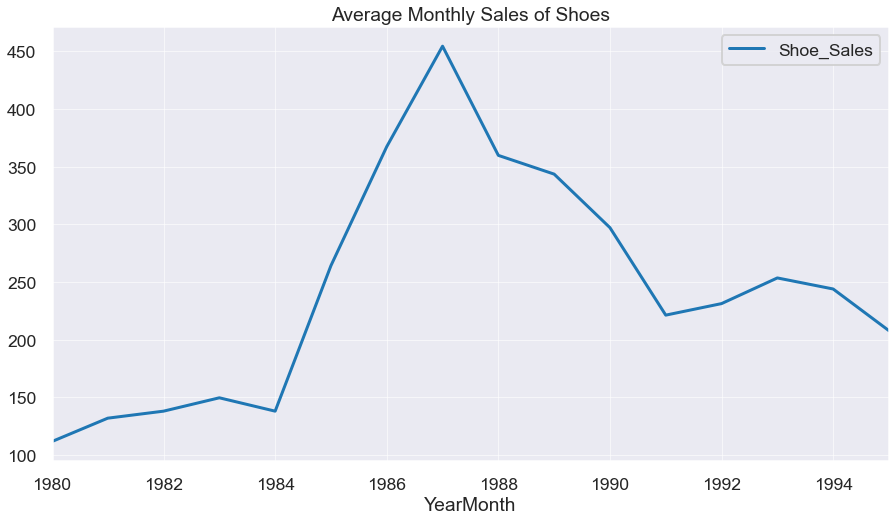
| **YearMonth** | **1980-12-31** | **1981-12-31** | **1982-12-31** | **1983-12-31** | **1984-12-31** | **1985-12-31** | **1986-12-31** | **1987-12-31** | **1988-12-31** | **1989-12-31** | **1990-12-31** | **1991-12-31** | **1992-12-31** | **1993-12-31** | **1994-12-31** | **1995-12-31** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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 the sales of shoes. 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 fallen 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** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Shoe\_Sales** | 111.75 | 132.0 | 138.083333 | 149.666667 | 138.083333 | 264.25 | 367.416667 | 454.5 | 359.833333 | 343.583333 | 297.166667 | 221.333333 | 231.333333 | 253.583333 | 244.0 | 207.857143 |

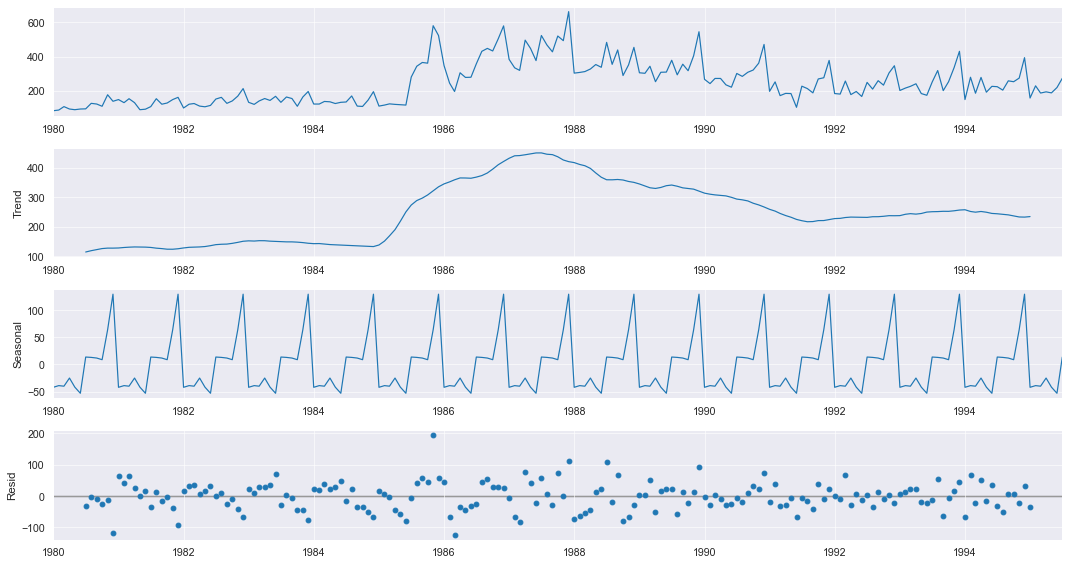


Here we see that we do not have data for the last 5 months of year 1995 and if we had complete data the average sales would have risen some more.

Let us decompose the data for further analysis.

Additive Model:

'Shoe\_Sales -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.

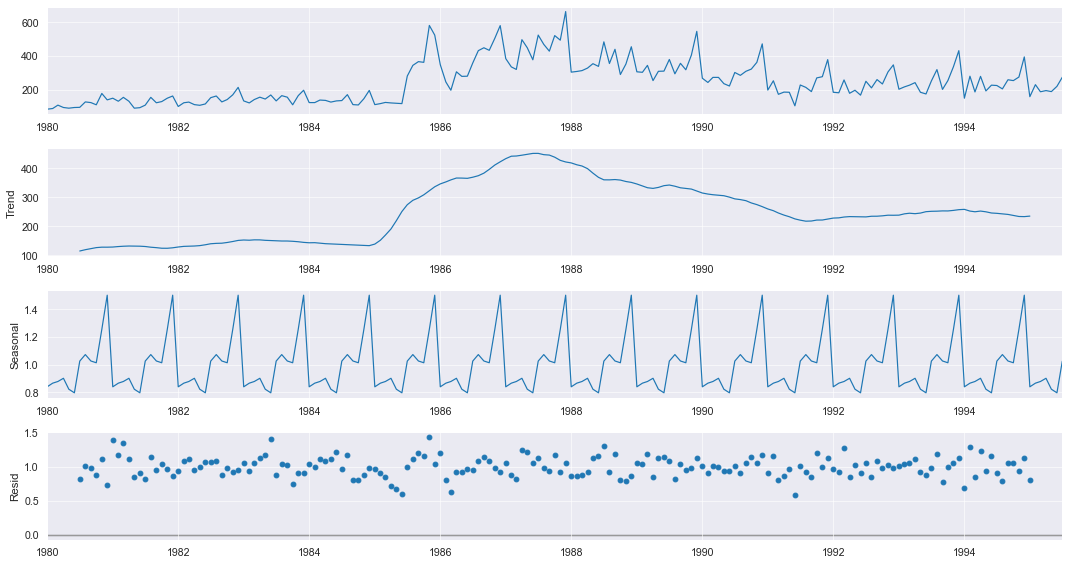
Second graph shows us a trend and we observe that the trend has been rising

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 -200 to 200.

Multiplicative Model:

'Shoe\_Sales -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 Shoe\_Sales Production Train Data

|  | **Shoe\_Sales** |
| --- | --- |
| **YearMonth** |  |
| **1980-01-01** | 85 |
| **1980-02-01** | 89 |
| **1980-03-01** | 109 |
| **1980-04-01** | 95 |
| **1980-05-01** | 91 |
| **1980-06-01** | 95 |
| **1980-07-01** | 96 |
| **1980-08-01** | 128 |
| **1980-09-01** | 124 |
| **1980-10-01** | 111 |
| **1980-11-01** | 178 |
| **1980-12-01** | 140 |

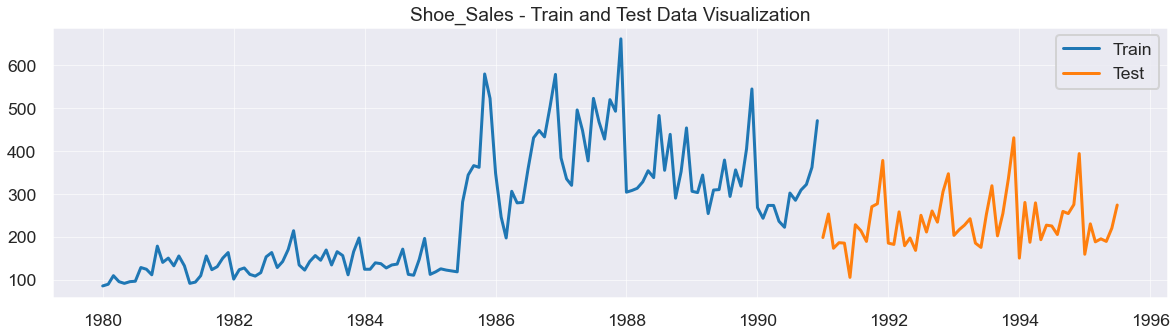
Total number of Observations in Train Set: 132

Total observations in our train set are 132.

First 12 observations of Shoe\_Sales Test Data

|  | **Shoe\_Sales** |
| --- | --- |
| **YearMonth** |  |
| **1991-01-01** | 198 |
| **1991-02-01** | 253 |
| **1991-03-01** | 173 |
| **1991-04-01** | 186 |
| **1991-05-01** | 185 |
| **1991-06-01** | 105 |
| **1991-07-01** | 228 |
| **1991-08-01** | 214 |
| **1991-09-01** | 189 |
| **1991-10-01** | 270 |
| **1991-11-01** | 277 |
| **1991-12-01** | 378 |

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

1991-01-01 421.658686

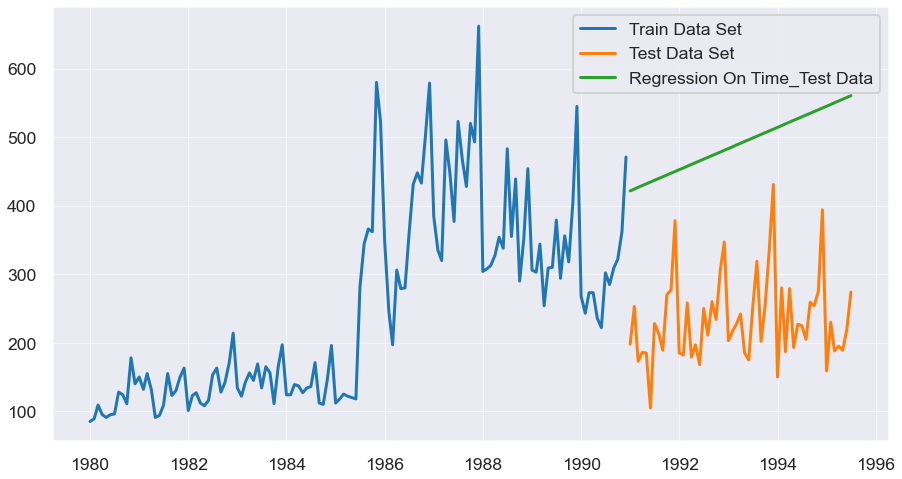
1991-02-01 424.231362

1991-03-01 426.804037

1991-04-01 429.376713

1991-05-01 431.949389

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 266.276 | 123.124 |



Above we have built a linear model of prediction and it has a RMSE of 266.27. After plotting the predictions on a graph and comparing it with the actuals we find it does not give us very good results of predicting sales.

**Naïve approach:**

Predictions:

YearMonth

1991-01-01 471

1991-02-01 471

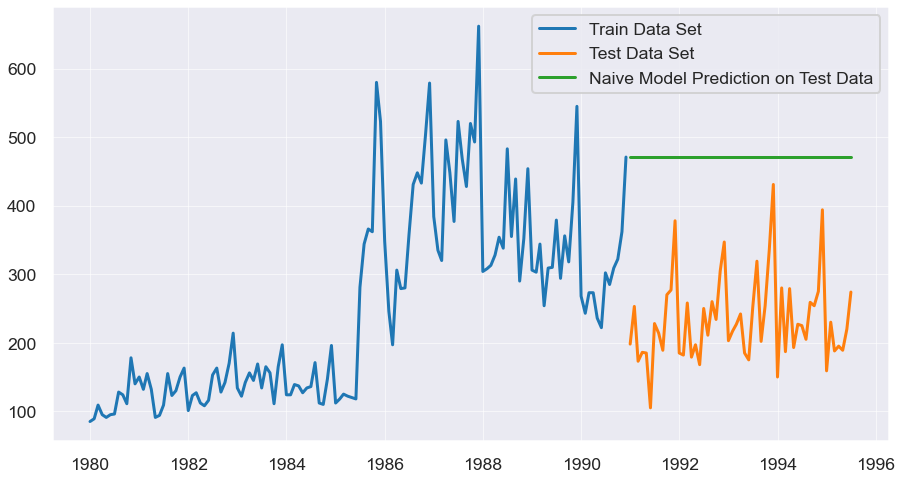
1991-03-01 471

1991-04-01 471

1991-05-01 471

Name: naive, dtype: int64

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **Naive Regression Model** | 245.121 | 114.87 |



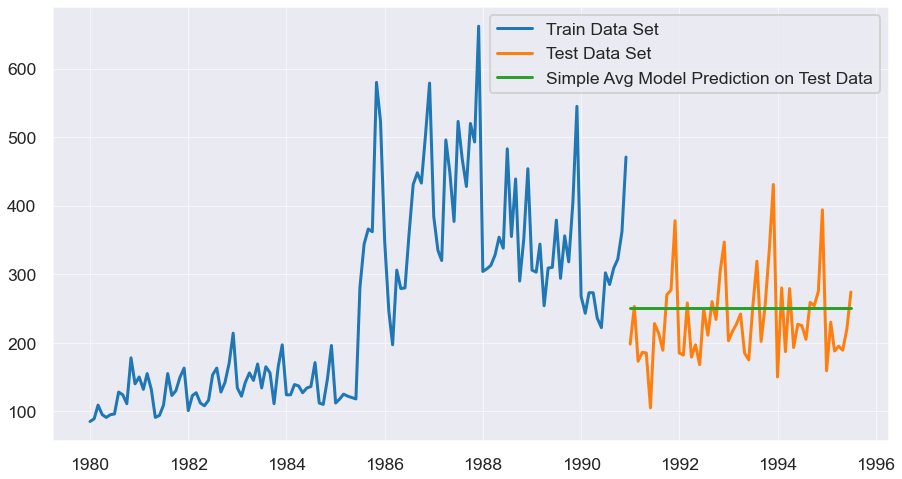
After building a naïve model we observe that the predictions are constant and it has taken the last entry as a constant. Because RMSE of 245.12 we can say that naïve model is also not very effective in predicting the sales in the upcoming months.

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 266.276 | 123.124 |
| **Naive Regression Model** | 245.121 | 114.870 |

**Simple average method:**

Predictions:

|  | **Shoe\_Sales** | | **mean\_forecast** |
| --- | --- | --- | --- |
| YearMonth |  | |  |
| 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** |
|  | | **Shoe\_Sales\_TEST RMSE** | | **Shoe\_Sales\_MAPE** |
| **SIMPLE AVG Model** | | 63.985 | | 24.251 |



Above we built a simple average model and we find a RMSE of 63.98 which means this model is better than the previous two models. 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:

1991-01-01 420.229977

1991-02-01 420.229977

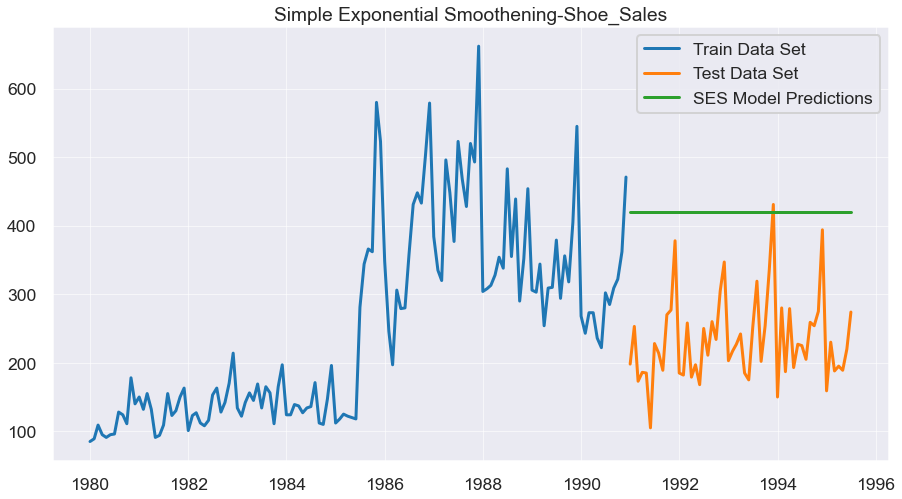
1991-03-01 420.229977

1991-04-01 420.229977

1991-05-01 420.229977

Freq: MS, dtype: float64

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **SES Model, Alpha=0.0987** | 196.405 | 91.799 |



We can see that we after we run our model on the test data, we are getting a RMSE of 196.405 and a constant of 91.799.

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 266.276 | 123.124 |
| **Naive Regression Model** | 245.121 | 114.870 |
| **SIMPLE AVG Model** | 63.985 | 24.251 |
| **SES Model, Alpha=0.0987** | 196.405 | 91.799 |

SES model is best model by far.

**Double exponential smoothening – Holt Model:**

1991-01-01 423.037852

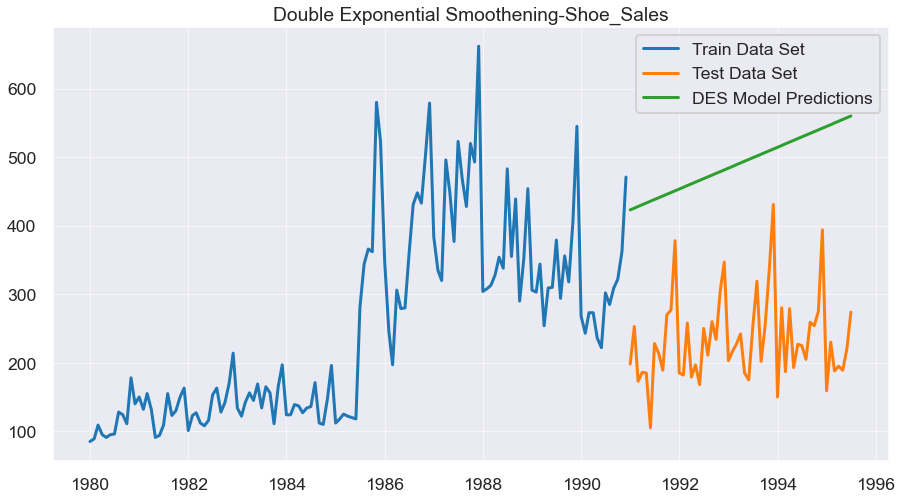
1991-02-01 425.576513

1991-03-01 428.115175

1991-04-01 430.653837

1991-05-01 433.192498

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 266.66 | 123.346 |



This model has a RMSE of 266.66. It gives an upward trend after we plot it in a graph in comparison to the actual.

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 266.276 | 123.124 |
| **Naive Regression Model** | 245.121 | 114.870 |
| **SIMPLE AVG Model** | 63.985 | 24.251 |
| **SES Model, Alpha=0.0987** | 196.405 | 91.799 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 266.660 | 123.346 |

DES model is not such a good performer in predicting the sales.

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

Predictions:

1991-01-01 219.083658

1991-02-01 213.816321

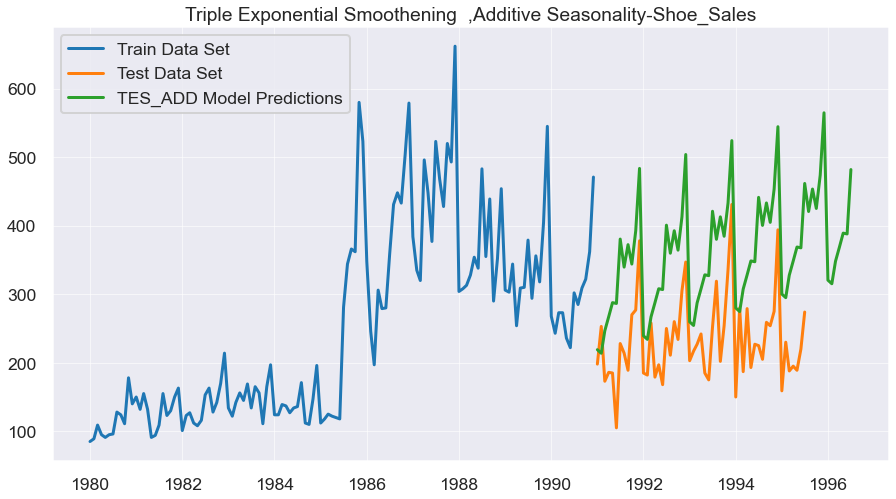
1991-03-01 246.658224

1991-04-01 267.260236

1991-05-01 287.719744

Freq: MS, dtype: float64

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 128.993 | 54.949 |



This model gives us a RMSE of 128.99. And when plotted on a graph we find it include seasonality.

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 266.276 | 123.124 |
| **Naive Regression Model** | 245.121 | 114.870 |
| **SIMPLE AVG Model** | 63.985 | 24.251 |
| **SES Model, Alpha=0.0987** | 196.405 | 91.799 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 266.660 | 123.346 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 128.993 | 54.949 |

So far, this model is only second best to simple average when we compare RMSE with other models.

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

Predictions:

1991-01-01 261.485579

1991-02-01 243.399646

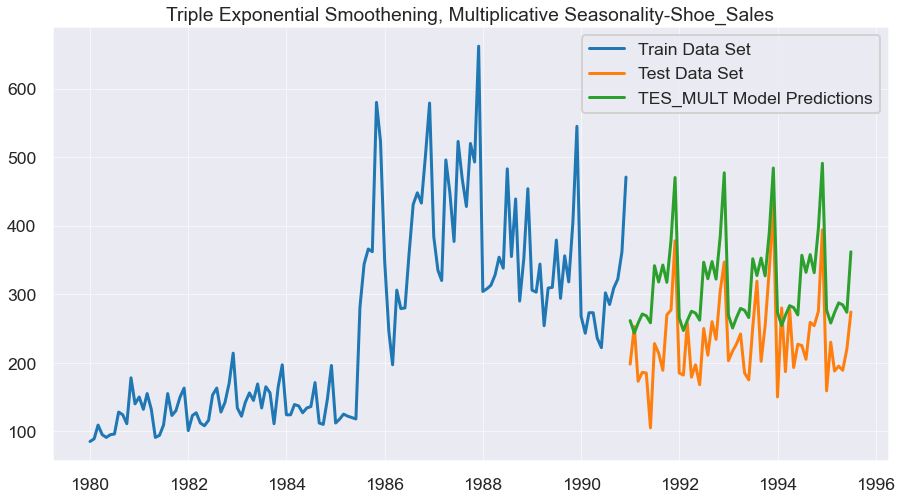
1991-03-01 257.709951

1991-04-01 271.138700

1991-05-01 268.503516

Freq: MS, dtype: float64

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 89.146 | 38.84 |



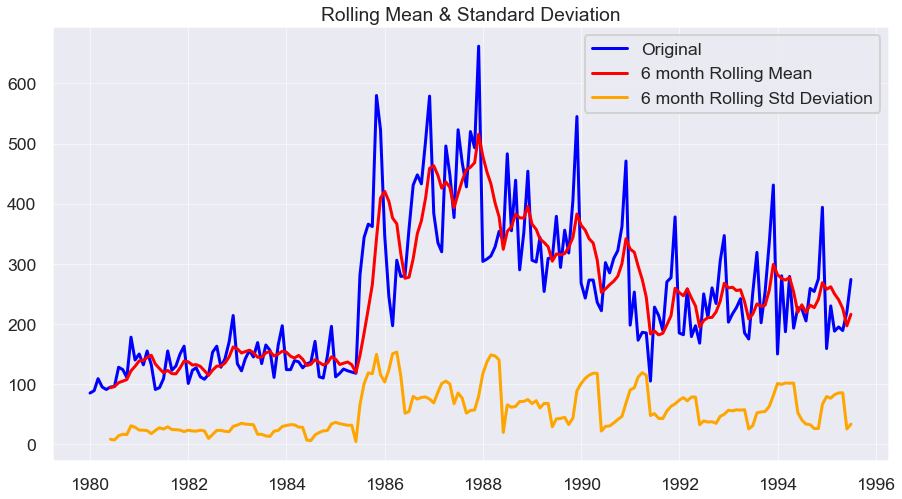
We get a RMSE of 89.146 and MAPE of 38.84. When we plot prediction vs actuals, we see a lot more accuracy in our predictions as it captures seasonality and trend very well.

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 266.276 | 123.124 |
| **Naive Regression Model** | 245.121 | 114.870 |
| **SIMPLE AVG Model** | 63.985 | 24.251 |
| **SES Model, Alpha=0.0987** | 196.405 | 91.799 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 266.660 | 123.346 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 128.993 | 54.949 |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 89.146 | 38.840 |

Thus, after comparing all the models, we have built we observe that Simple average model is the best suited for predictions. It has a RMSE of 63.985 and a MAPE of 24.251. 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 -1.717397

p-value 0.422172

#Lags Used 13.000000

Number of Observations Used 173.000000

Critical Value (1%) -3.468726

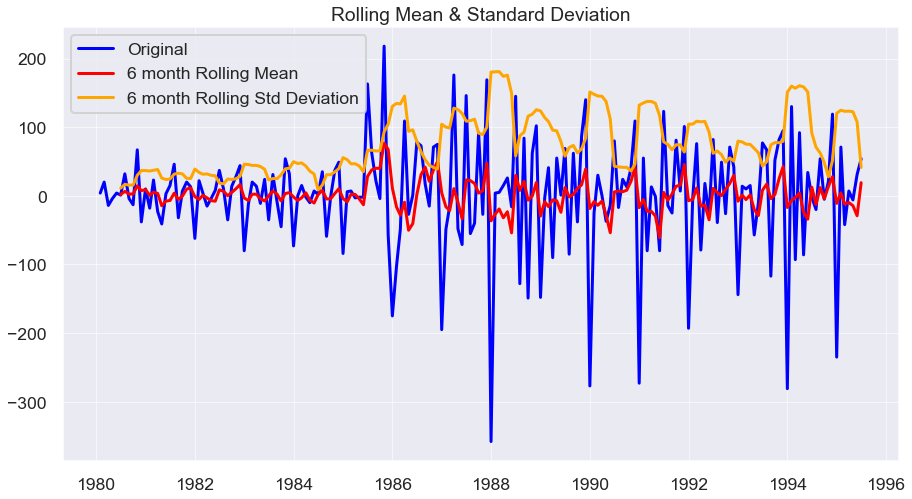
Critical Value (5%) -2.878396

Critical Value (10%) -2.575756

dtype: float64

Here we find that the p value is 0.422 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 -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

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** |
| --- | --- | --- |
| **11** | (2, 1, 3) | 1480.81 |
| **15** | (3, 1, 3) | 1482.57 |
| **5** | (1, 1, 1) | 1492.49 |
| **6** | (1, 1, 2) | 1494.42 |
| **9** | (2, 1, 1) | 1494.43 |

SARIMAX Results

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

Dep. Variable: Shoe\_Sales No. Observations: 132

Model: ARIMA(2, 1, 3) Log Likelihood -734.405

Date: Wed, 25 Jan 2023 AIC 1480.809

Time: 09:04:30 BIC 1498.060

Sample: 01-01-1980 HQIC 1487.819

- 12-01-1990

Covariance Type: opg

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

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

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

ar.L1 0.0142 0.029 0.487 0.626 -0.043 0.071

ar.L2 -0.9971 0.018 -55.476 0.000 -1.032 -0.962

ma.L1 -0.3351 0.090 -3.720 0.000 -0.512 -0.159

ma.L2 0.9874 0.100 9.883 0.000 0.792 1.183

ma.L3 -0.2836 0.084 -3.389 0.001 -0.448 -0.120

sigma2 4273.5973 574.401 7.440 0.000 3147.792 5399.402

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

Ljung-Box (L1) (Q): 0.04 Jarque-Bera (JB): 41.43

Prob(Q): 0.84 Prob(JB): 0.00

Heteroskedasticity (H): 13.28 Skew: -0.59

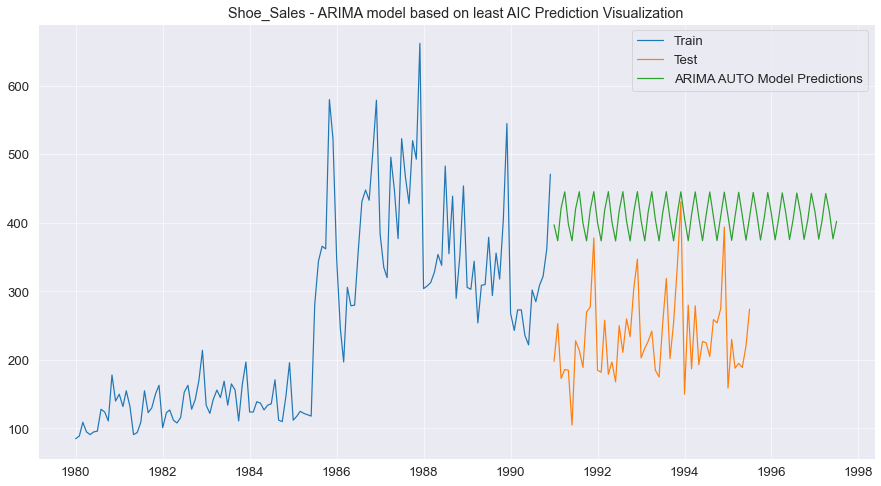
Prob(H) (two-sided): 0.00 Kurtosis: 5.49

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

Warnings:

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

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



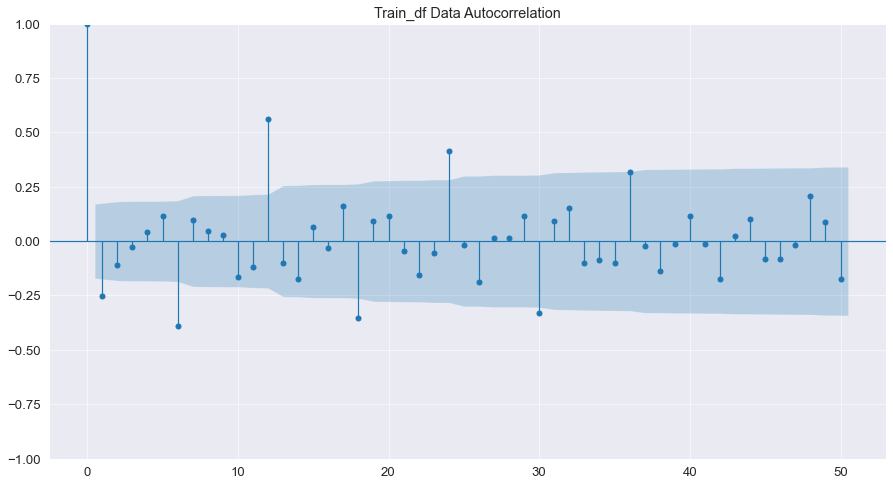
We have plotted our predictions made by ARIMA model on a graph vs actual.

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 266.276000 | 123.124000 |
| **Naive Regression Model** | 245.121000 | 114.870000 |
| **SIMPLE AVG Model** | 63.985000 | 24.251000 |
| **SES Model, Alpha=0.0987** | 196.405000 | 91.799000 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 266.660000 | 123.346000 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 128.993000 | 54.949000 |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 89.146000 | 38.840000 |
| **ARIMA\_AUTO(2,1,3)** | 184.469328 | 85.650194 |

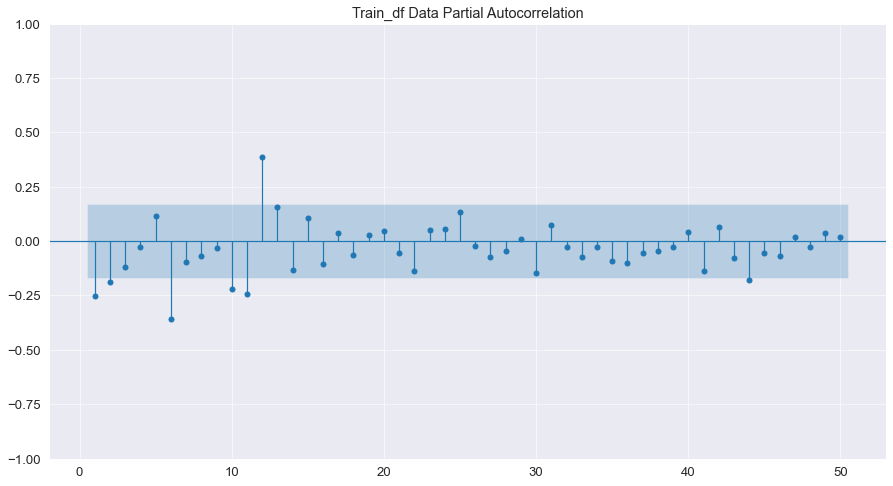
We find ARIMA model has a RMSE of 816 and MAPE of 19.9. Thus, it is not very well suited for predictions.

**SARIMA model based on least AIC:**

ACF:



PACF:



Parameters, seasonality and AIC in ascending order:

|  | **param** | **seasonal** | **AIC** |
| --- | --- | --- | --- |
| **47** | (0, 1, 2) | (3, 0, 3, 12) | 99.847419 |
| **63** | (0, 1, 3) | (3, 0, 3, 12) | 230.534029 |
| **28** | (0, 1, 1) | (3, 0, 0, 12) | 1063.853592 |
| **60** | (0, 1, 3) | (3, 0, 0, 12) | 1064.814556 |
| **29** | (0, 1, 1) | (3, 0, 1, 12) | 1064.988494 |

We have taken seasonality as 12 and d = 1.

SARIMAX Results

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

Dep. Variable: Shoe\_Sales No. Observations: 132

Model: SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12) Log Likelihood -507.955

Date: Wed, 25 Jan 2023 AIC 1035.910

Time: 09:06:35 BIC 1061.128

Sample: 01-01-1980 HQIC 1046.088

- 12-01-1990

Covariance Type: opg

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

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

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

ar.L1 0.3379 0.253 1.336 0.182 -0.158 0.834

ar.L2 0.2426 0.172 1.412 0.158 -0.094 0.580

ar.L3 -0.1175 0.109 -1.075 0.282 -0.332 0.097

ma.L1 -0.7392 0.299 -2.472 0.013 -1.325 -0.153

ar.S.L12 0.8705 1.115 0.781 0.435 -1.314 3.055

ar.S.L24 0.2240 1.461 0.153 0.878 -2.640 3.088

ar.S.L36 -0.0394 0.326 -0.121 0.904 -0.679 0.600

ma.S.L12 -0.6012 1.062 -0.566 0.571 -2.683 1.481

ma.S.L24 -0.3990 1.177 -0.339 0.735 -2.706 1.908

sigma2 2919.7583 0.000 1.4e+07 0.000 2919.758 2919.759

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

Ljung-Box (L1) (Q): 0.02 Jarque-Bera (JB): 5.74

Prob(Q): 0.90 Prob(JB): 0.06

Heteroskedasticity (H): 1.09 Skew: 0.20

Prob(H) (two-sided): 0.81 Kurtosis: 4.16

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

Warnings:

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

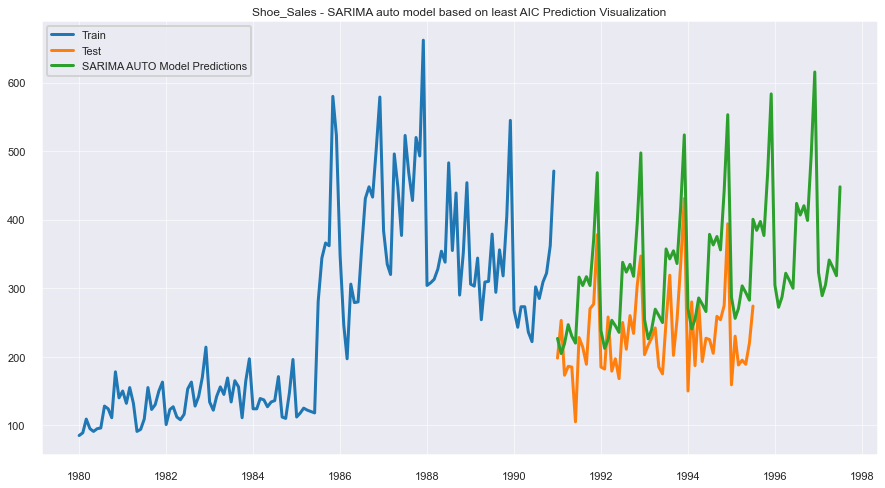
[2] Covariance matrix is singular or near-singular, with condition number 4.97e+24. Standard errors may be unstable.

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

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 266.276000 | 123.124000 |
| **Naive Regression Model** | 245.121000 | 114.870000 |
| **SIMPLE AVG Model** | 63.985000 | 24.251000 |
| **SES Model, Alpha=0.0987** | 196.405000 | 91.799000 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 266.660000 | 123.346000 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 128.993000 | 54.949000 |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 89.146000 | 38.840000 |
| **ARIMA\_AUTO(2,1,3)** | 184.469328 | 85.650194 |
| **SARIMA\_AUTO(3,1,1)(3,0,2,12)** | 90.404287 | 36.645250 |

WE see that SARIMA model has a RMSE of 90.40 and MAPE of 36.64. This model has given us even better results than Holt Winter multiplicative model.

Hence, simple average model is best for predicting future sales of shoes because it has the least RMSE .



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: Shoe\_Sales No. Observations: 132

Model: ARIMA(2, 1, 2) Log Likelihood -743.205

Date: Wed, 25 Jan 2023 AIC 1496.411

Time: 09:06:40 BIC 1510.787

Sample: 01-01-1980 HQIC 1502.252

- 12-01-1990

Covariance Type: opg

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

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

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

ar.L1 0.6739 4.219 0.160 0.873 -7.594 8.942

ar.L2 -0.0695 1.993 -0.035 0.972 -3.976 3.837

ma.L1 -1.0484 4.221 -0.248 0.804 -9.321 7.224

ma.L2 0.1708 3.512 0.049 0.961 -6.713 7.055

sigma2 4941.5633 439.453 11.245 0.000 4080.250 5802.876

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

Ljung-Box (L1) (Q): 0.01 Jarque-Bera (JB): 54.37

Prob(Q): 0.91 Prob(JB): 0.00

Heteroskedasticity (H): 12.71 Skew: 0.00

Prob(H) (two-sided): 0.00 Kurtosis: 6.16

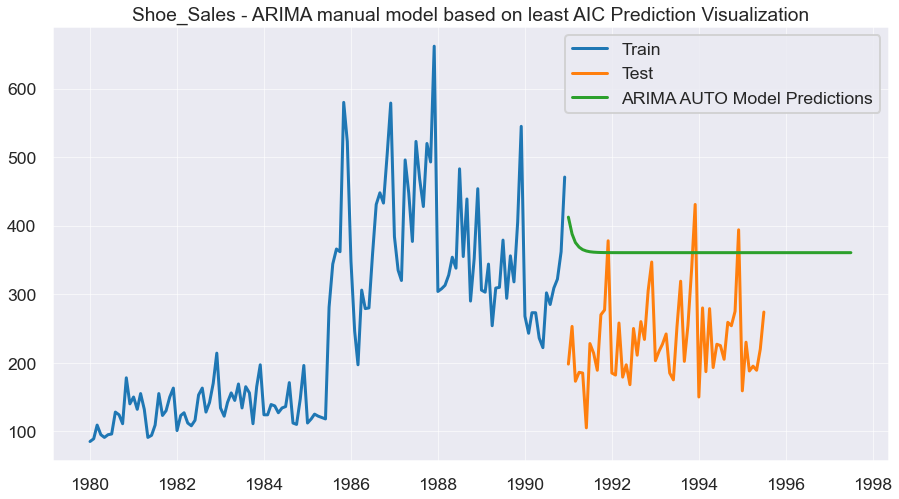
===================================================================================

Warnings:

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

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **Linear Regression Model** | 266.276000 | 123.124000 |
| **Naive Regression Model** | 245.121000 | 114.870000 |
| **SIMPLE AVG Model** | 63.985000 | 24.251000 |
| **SES Model, Alpha=0.0987** | 196.405000 | 91.799000 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 266.660000 | 123.346000 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 128.993000 | 54.949000 |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 89.146000 | 38.840000 |
| **ARIMA\_AUTO(2,1,3)** | 184.469328 | 85.650194 |
| **SARIMA\_AUTO(3,1,1)(3,0,2,12)** | 90.404287 | 36.645250 |
| **ARIMA\_MANUAL(2,1,2)** | 143.558270 | 66.616022 |

With the help of above chart we can see that this model has not worked best as it has a RMSE of 143.55 and MAPE of 66.6.



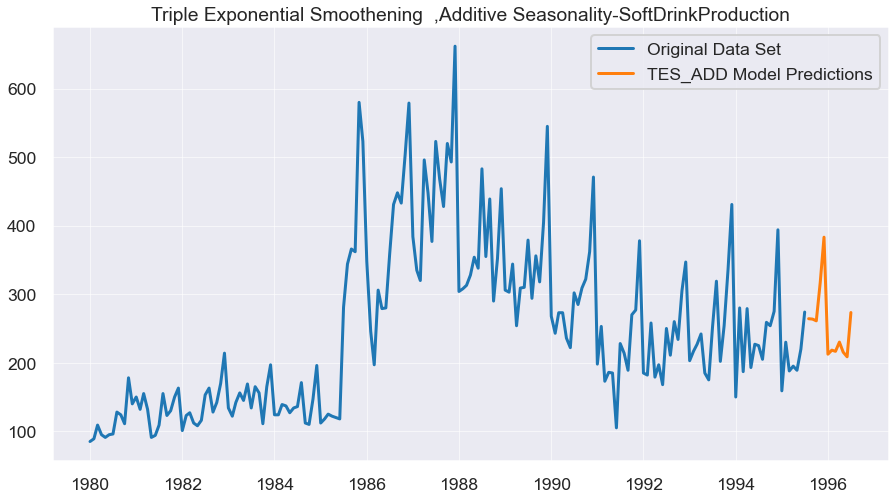
Above is the graph of ARIMA manual model of predictions in comparison of actuals.

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

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **SIMPLE AVG Model** | 64.0 | 24.3 |
| **SARIMA\_MANUAL(4,0,2)(1,1,2,12)** | 74.9 | 32.9 |
| **TES\_MULT Model,Alpha = 0.0713,Beta = 0.0455,Gamma = 8.3567e-07** | 89.1 | 38.8 |
| **SARIMA\_AUTO(3,1,1)(3,0,2,12)** | 90.4 | 36.6 |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 129.0 | 54.9 |
| **ARIMA\_MANUAL(2,1,2)** | 143.6 | 66.6 |
| **ARIMA\_AUTO(2,1,3)** | 184.5 | 85.7 |
| **SES Model, Alpha=0.0987** | 196.4 | 91.8 |
| **Naive Regression Model** | 245.1 | 114.9 |
| **Linear Regression Model** | 266.3 | 123.1 |
| **DES Model,Alpha=1.91e-08,Beta=7.30e-09** | 266.7 | 123.3 |

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** | 264.27 |
| **1995-09-01** | 263.69 |
| **1995-10-01** | 261.15 |
| **1995-11-01** | 316.99 |
| **1995-12-01** | 383.25 |
| **1996-01-01** | 212.44 |
| **1996-02-01** | 218.38 |
| **1996-03-01** | 216.63 |
| **1996-04-01** | 230.10 |
| **1996-05-01** | 215.43 |
| **1996-06-01** | 208.66 |
| **1996-07-01** | 273.31 |

|  | **Shoe\_Sales\_TEST RMSE** | **Shoe\_Sales\_MAPE** |
| --- | --- | --- |
| **TES\_ADD Model,Alpha=0.0883,Beta=6.73e-05,Gamma=0.004455** | 64.4 | 20.49 |

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

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

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