TIME SERIES FORECASTING

PROJECT REPORT

Shoesales.csv
&
SoftDrink.csv

By
HARI HARAN

25th Feb, 2024

`

`

`

`

Problem 1 for the Data Set :: Shoesales.csv

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

In this report, we will focus on analyzing the shoe sales data from January 1980 to July 1995. The task in hand is to review the given data over a period of time to identify patterns, trends, seasonality changes. This report aims to draw inferences and insights about the product sales.

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

This data set has 187 rows and 1 column :

1. Rows have the Yearly sales (Months and Dates) - YearMonth
2. Columns has the sales value - Sales

YearMonth	Shoe_Sales
1980-01-01	85
1980-02-01	89

YearMonth	Shoe_Sales
1995-06-01	220
1995-07-01	274

Table 1.1. - Above table shows Head(Left) and Tail (right) of the data set

The dataset is further divided by extraction of month and year columns from the Year-Month column and renamed as Sales, Year and Month for better analysis of the given data set.

YearMonth	Shoe_Sales	Year	Month
1980-01-01	85	1980	1
1980-02-01	89	1980	2
1980-03-01	109	1980	3
1980-04-01	95	1980	4
1980-05-01	91	1980	5

```
**Head of the given Dataset**
             Sales  Year  Month
YearMonth
1980-01-01      85  1980      1
1980-02-01      89  1980      2
1980-03-01     109  1980      3
1980-04-01      95  1980      4
1980-05-01      91  1980      5
**Tail of the given Dataset**
             Sales  Year  Month
YearMonth
1995-03-01     188  1995      3
1995-04-01     195  1995      4
1995-05-01     189  1995      5
1995-06-01     220  1995      6
1995-07-01     274  1995      7
```

Table 1.2. - New rows in the data set

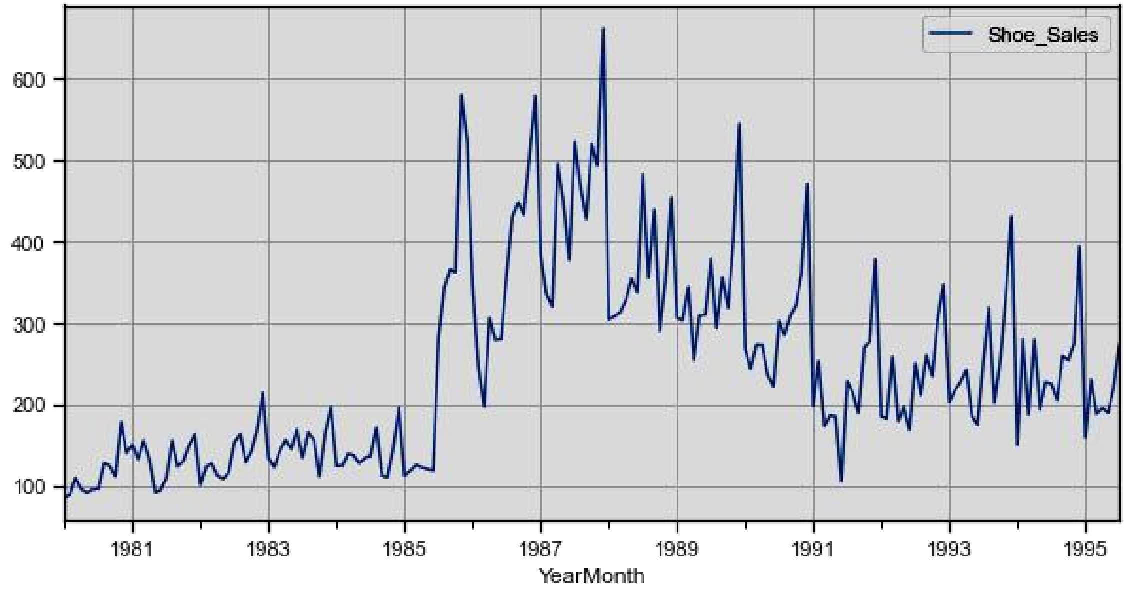

Fig. 1.1. - Time series Plot

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

Performing EDA,

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Sales   187 non-null    int64
 1   Year    187 non-null    int64
 2   Month   187 non-null    int64
dtypes: int64(3)
memory usage: 5.8 KB
```

Table 1.3. - Data set info

	count	mean	std	min	25%	50%	75%	max
Sales	187.0	245.636364	121.390804	85.0	143.5	220.0	315.5	662.0
Year	187.0	1987.299465	4.514749	1980.0	1983.0	1987.0	1991.0	1995.0
Month	187.0	6.406417	3.450972	1.0	3.0	6.0	9.0	12.0

Table 1.4. - Data set description

No Null Values found,

```
Sales      0
Year       0
Month      0
dtype: int64
```

Boxplot,

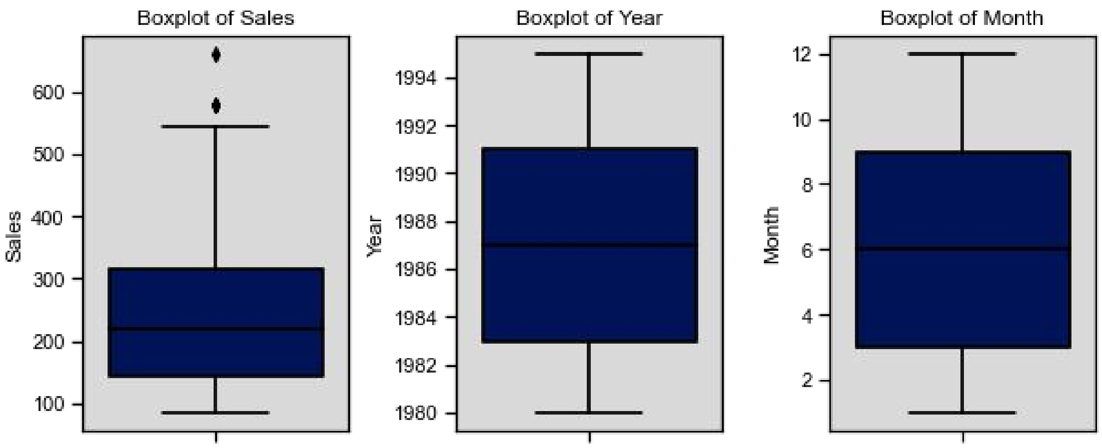

Fig. 1.2 Boxplot of the data

The box plot shows:

- Sales boxplot has outliers but we are choosing not to treat, as it will not have much impact on the time series analysis.

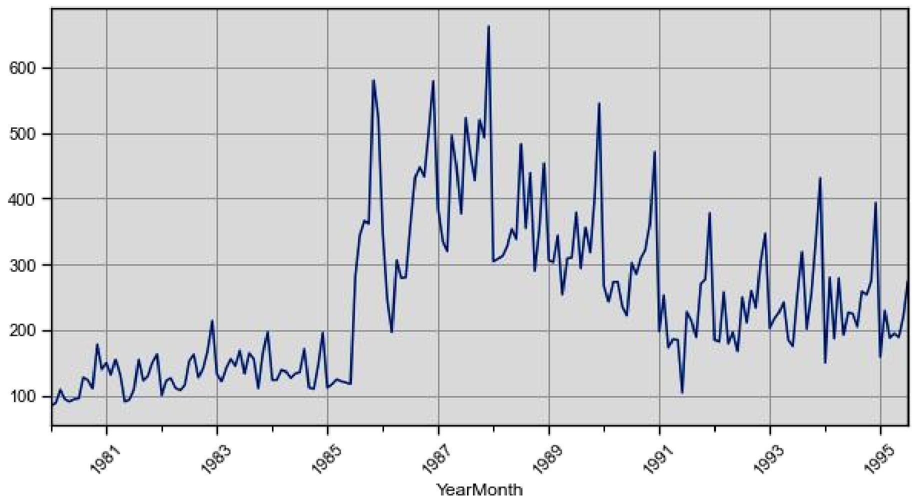

Fig. 1.3. - Time Series Plot of the Sales (with new columns)

The above plot shows the trend and seasonality pattern. It also shows that there was a peak of sales in the year 1986-1989. After 1991, there is a slow decline in the pattern.

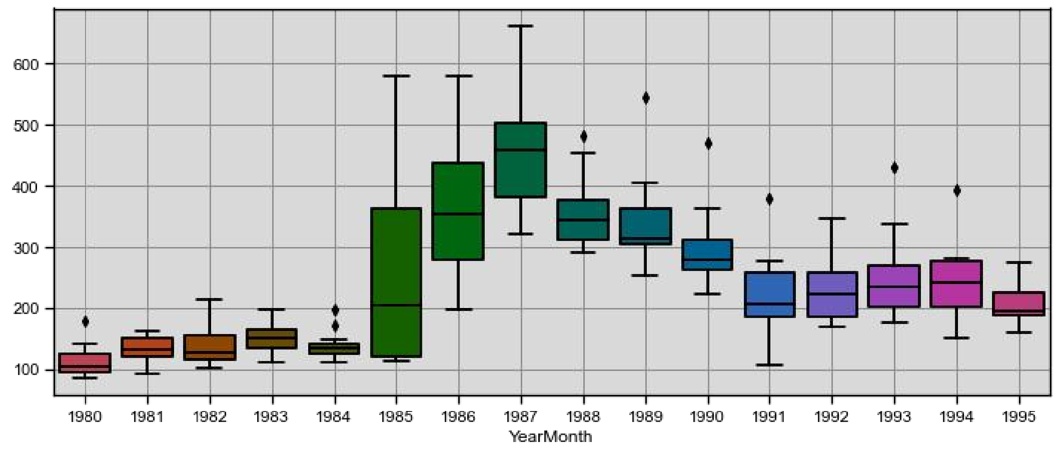

Fig. 1.4. - Boxplot of the Sales- Yearly

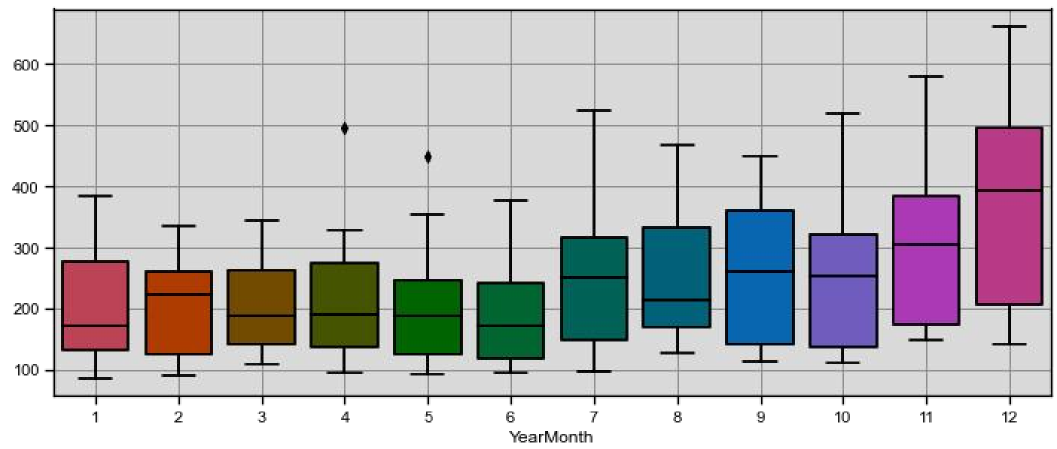

Fig. 1.5. - Boxplot of the Sales- Monthly

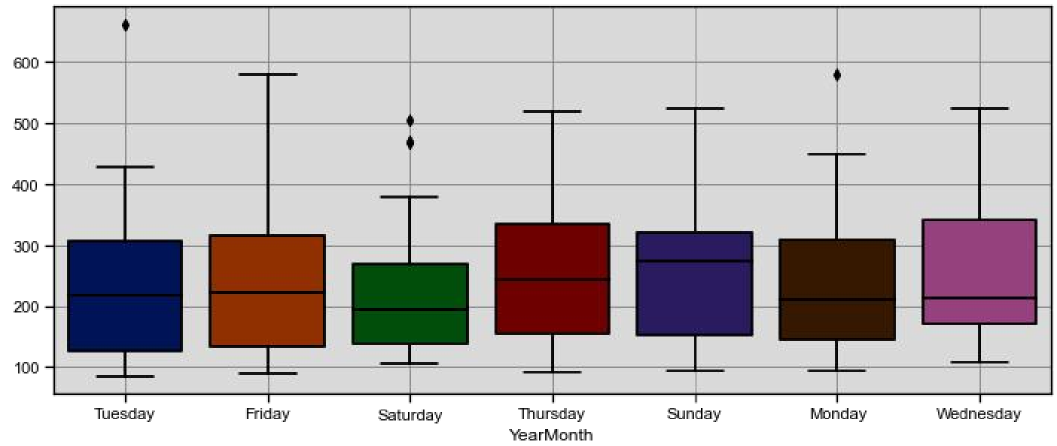

Fig. 1.6. - Boxplot of the Sales- Weekly (Weekdays Sales)

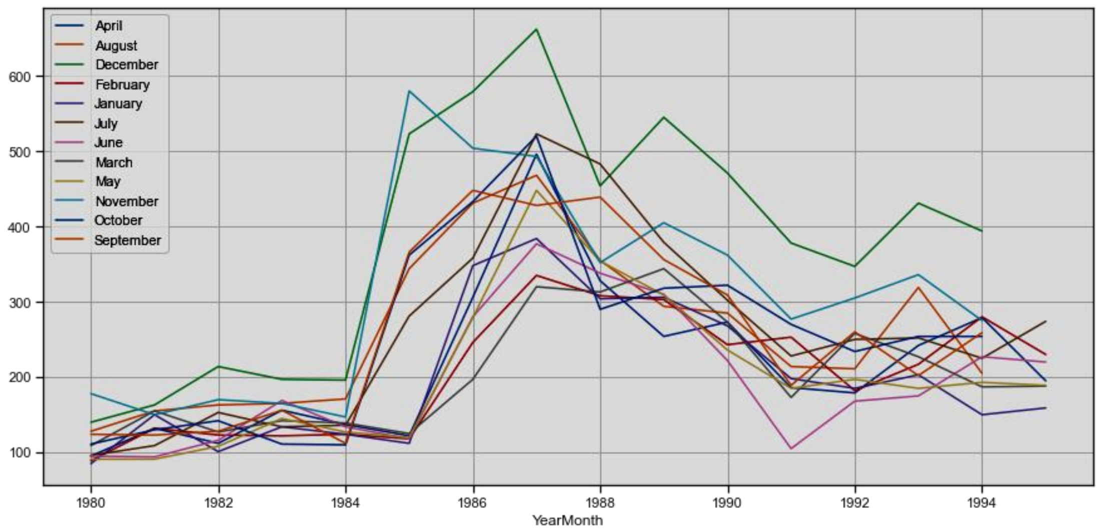

Fig. 1.7. - Monthly Sales over the Years

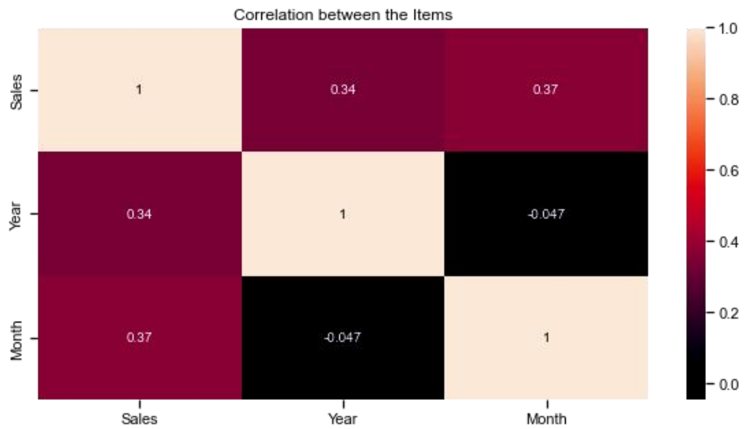

Fig. 1.8. - Correlation Heat Map

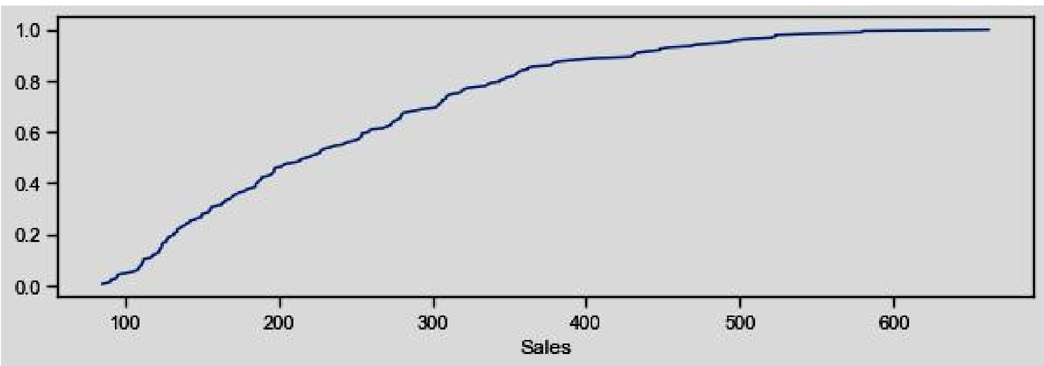

Fig. 1.9. - Empirical Cumulative Distribution Function Curve Plot

This plot shows the distribution of sales over the years in the data :

- Around 65% of the sales has been less than 400.
- 600 is the highest sales value achieved over the years.

Decomposition of the Time series,

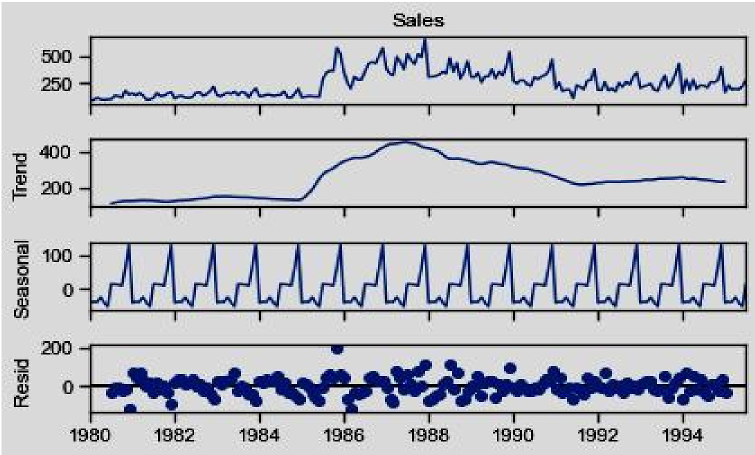

Fig.1.10. - Time Decomposition - Additive

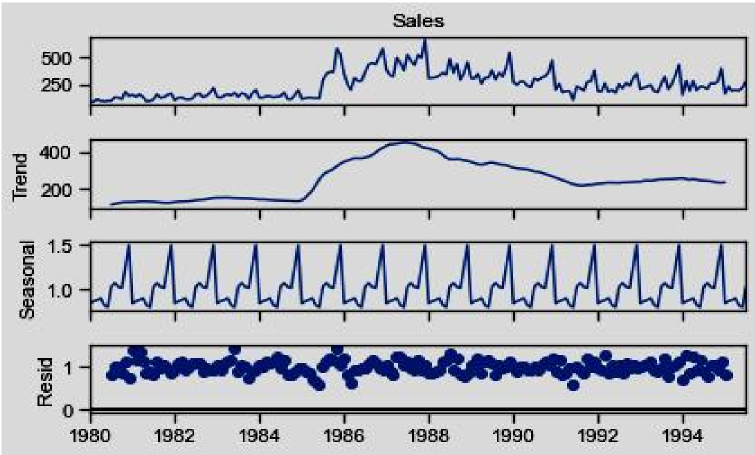

Fig.1.11. - Time Decomposition - Multiplicative

The above plots show:

- Peak year of sales is between the years 1986 to 1990.
- Both trend and seasonality are present.
- The trend has declined over the years after 1990 and then it declined.
- Residue is spread and is not in a straight line for **additive** and the residue is almost in a straight line for **multiplicative.**
- Residue for multiplicative is between 0 to 1, while for additive it is between 0 to 200.

So the multiplicative model has more stable residual plot and lower range of residuals. This means that the observations are properly captured by the model and not too much spread.

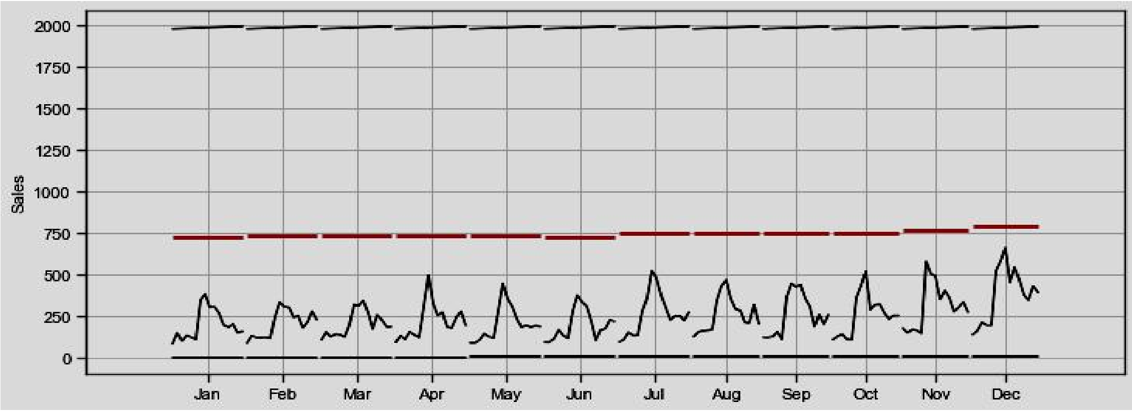

Fig.1.12. - Monthly Plot

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

Now , we have to split the data set into training and test set for model building analysis.

As per instructions, data is split from 1980-1990 is train data set, then 1991 to 1995 is the test data set.
Rows & Columns for train-test date set:

- Train dataset has 132 rows and 3 columns.
- Test dataset has 55 and 3 columns.

First few rows of Training Data

YearMonth	Sales	Year	Month
1980-01-01	85	1980	1
1980-02-01	89	1980	2
1980-03-01	109	1980	3
1980-04-01	95	1980	4
1980-05-01	91	1980	5

Last few rows of Training Data

YearMonth	Sales	Year	Month
1990-08-01	285	1990	8
1990-09-01	309	1990	9
1990-10-01	322	1990	10
1990-11-01	362	1990	11
1990-12-01	471	1990	12

First few rows of Test Data

YearMonth	Sales	Year	Month
1991-01-01	198	1991	1
1991-02-01	253	1991	2
1991-03-01	173	1991	3
1991-04-01	186	1991	4
1991-05-01	185	1991	5

Last few rows of Test Data

YearMonth	Sales	Year	Month
1995-03-01	188	1995	3
1995-04-01	195	1995	4
1995-05-01	189	1995	5
1995-06-01	220	1995	6
1995-07-01	274	1995	7

Table 1.5. Train-Test data set (head & tail)

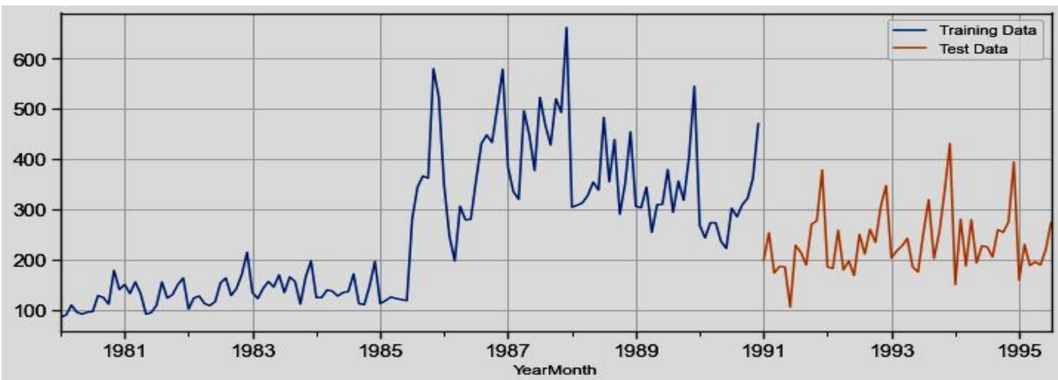

Fig.1.13. - Train-Test Plot

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

Model 1 : Linear Regression Model

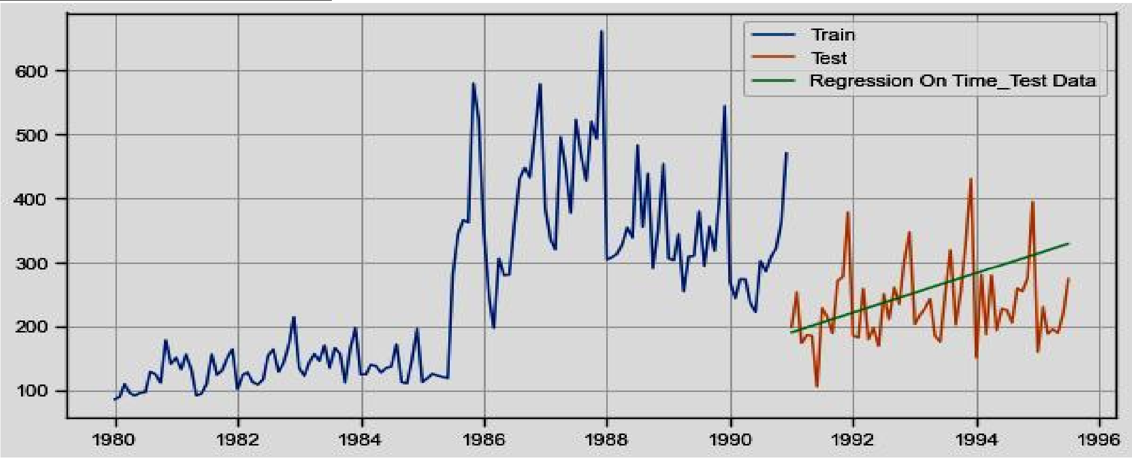

Fig. 1.14. - Linear Regression Plot

From the above plot,
The green line shows the prediction made by the model and the orange lines shows the actual test values. The predicted values are really far from the actual test values with an upward trend.
The RSME value for the Linear Reg Model = 73.11

Model 2 : Naïve forecast Model

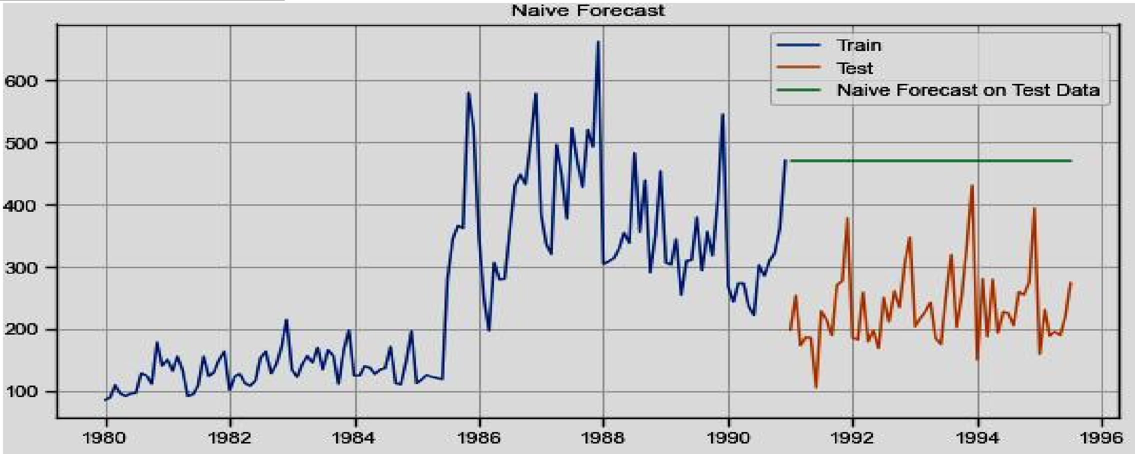

Fig. 1.15. - Naïve forecast Model Plot

From the above plot,
The green line shows the prediction made by the model and the orange lines shows the actual test values. The predicted values are really far from the actual test values.
The RSME value for the Naïve Model = 245.12

Model 3: Simple Average

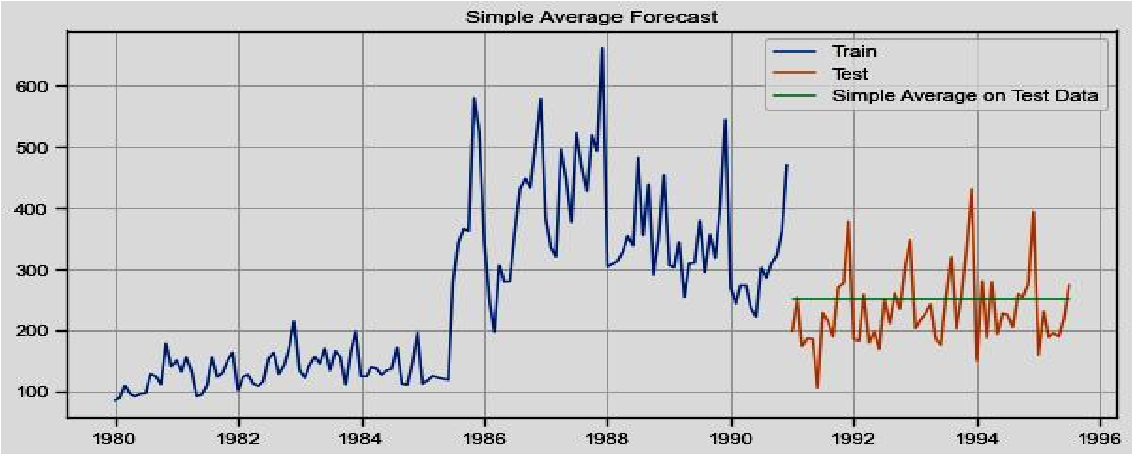

Fig. 1.16. - Simple Average Model Plot

From the above plot,
The green line shows the prediction made by the model and the orange lines shows the actual test values. The predicted values are really far from the actual test values.

The RSME value for the Simple Average Model = 63.98

Model 4: Simple Exponential Smoothing (SES) Model

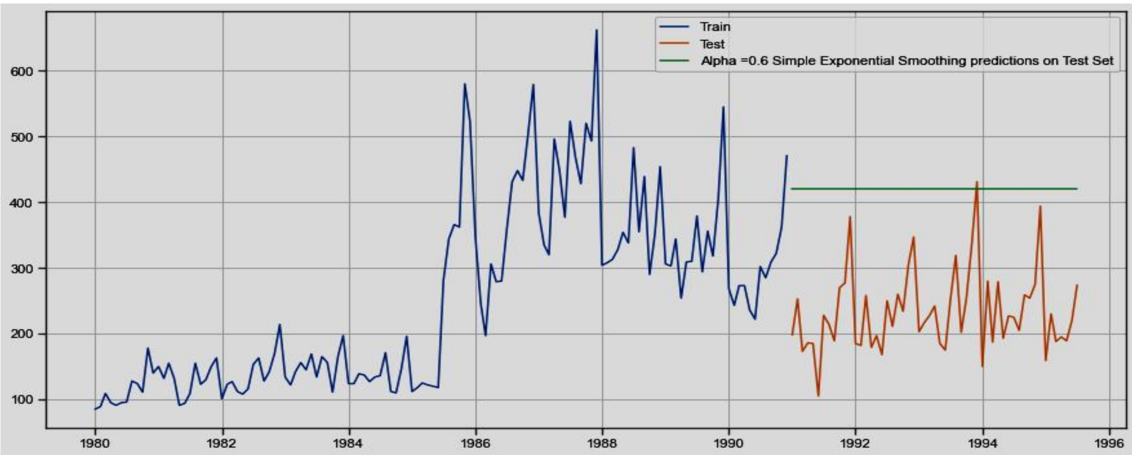

Fig. 1.17. - SES Model Plot

From the above plot,
Using auto-fit and finding the best parameters for this model, green line shows the prediction made by the model at alpha = 0.6 and the green line at alpha = 0.6 is considered the best among all the other alpha.

RSME value for SES Model (at alpha - 0.6) = 115.874445

Model 5: Double Exponential Smoothing - Holt's Model (DES)

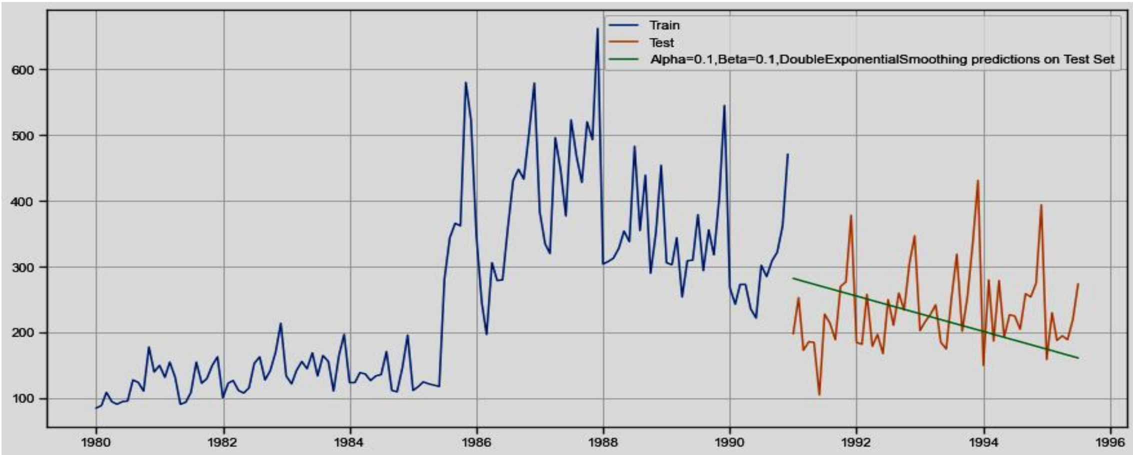

Fig. 1.18. - DES Model Plot

From the above plot,
The green line shows the prediction made by the model and the orange lines shows the actual test values. The predicted values are really far from the actual test values.

For this plot **alpha = 0.1** and **Beta = 0.1** is considered, the RSME value for the Holt's Model = 76.918569

Model 7: Triple Exponential Smoothing a.k.a Holt - Winter's Model(TES)

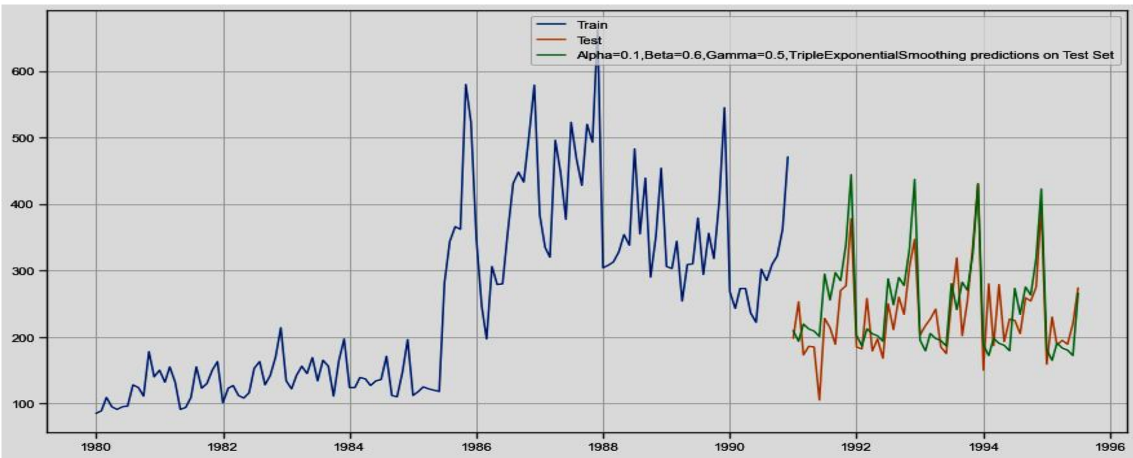

Fig. 1.19. - TES Model Plot

From the above plot,
Here we have considered **Alpha = 0.1**, **Beta = 0.6** and **Gamma = 0.5**,

The **RSME** value for the TES Model = 45.832046.

This is the best model which has both additive trend as well as seasonality.

1.5. Check for the stationary 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 stationary and comment. Note: Stationary should be checked at alpha = 0.05.

Applying Augmented Dickey-Fuller test whether the series has unit and whether it is stationary or non-stationary. Hypothesis for the ADF test:

- *H0 : The Time Series has a unit root and is thus non-stationary (Null Hypothesis)*
- *H1 : The Time Series does not have a unit root and is thus stationary. (Alternate Hypothesis)*

We see that for α = 5% (significance), the Time Series is non-stationary.

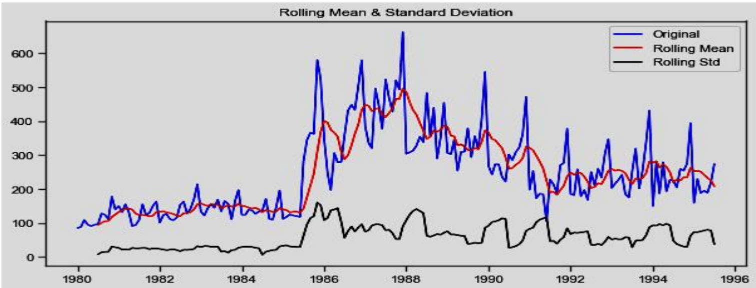

Fig. 1.20. - Augmented Dickey-Fuller Test Plot

Test Output ,

```
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
```

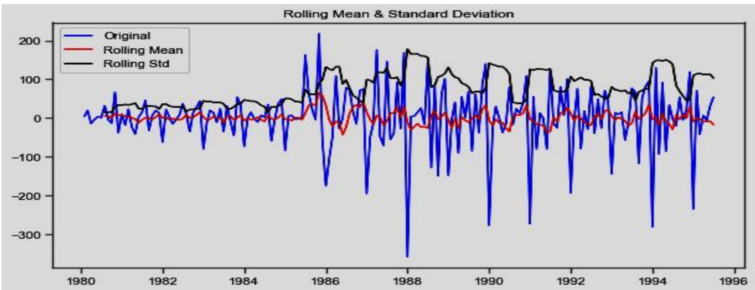

Fig. 1.21 Augmented Dickey-Fuller Test Plot - .diff() Method

Test Output ,

```
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
```

From the above output values, we can **reject** the **null hypothesis** that the series **not stationary**.

We can now **build ARIMA/ SARIMA models**, as we have proven that **the series is stationary**.

1.6. 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.

ARIMA MODEL:

Some parameter combinations for the Model :

Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)

	Params	AIC
15	(3, 1, 3)	1479.686833
11	(2, 1, 3)	1480.804807
5	(1, 1, 1)	1492.487187
6	(1, 1, 2)	1494.423859
9	(2, 1, 1)	1494.431498
2	(0, 1, 2)	1494.964605
3	(0, 1, 3)	1495.148474
14	(3, 1, 2)	1495.655855
13	(3, 1, 1)	1496.346864
7	(1, 1, 3)	1496.385878
10	(2, 1, 2)	1496.410739
1	(0, 1, 1)	1497.050322
12	(3, 1, 0)	1498.930309
8	(2, 1, 0)	1498.950483
4	(1, 1, 0)	1501.643124
0	(0, 1, 0)	1508.283772

Applying for loop for determining the optimum values of ***p***,***d***,***q*** where ***p*** is the order of the ***AR*** **(Auto-Regressive)** part of the model, while ***q*** is the order of the ***MA*** **(Moving Average)** part of the model and ***d*** is the **difference that is required** to make the series stationary.

p and ***q*** values are in the range of (0,4) were given to the for loop, while a fixed value of 1 was given for ***d***, since we had already determined ***d*** to be 1, while checking for stationary using the ADF test.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Sales   No. Observations:                  132
Model:                 ARIMA(3, 1, 3)   Log Likelihood                -732.843
Date:                Sun, 25 Feb 2024   AIC                           1479.687
Time:                        14:52:12   BIC                           1499.813
Sample:                    01-01-1980   HQIC                          1487.865
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.5588      0.117      4.767      0.000       0.329       0.789
ar.L2         -1.0067      0.022    -44.874      0.000      -1.051      -0.963
ar.L3          0.5423      0.119      4.540      0.000       0.308       0.776
ma.L1         -0.8858      0.105     -8.405      0.000      -1.092      -0.679
ma.L2          1.0383      1.621      0.640      0.522      -2.139       4.216
ma.L3         -0.8377      1.288     -0.650      0.515      -3.362       1.687
sigma2      4155.6928   6405.968      0.649      0.517   -8399.774    1.67e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.85   Jarque-Bera (JB):                44.05
Prob(Q):                              0.36   Prob(JB):                         0.00
Heteroskedasticity (H):              13.71   Skew:                            -0.14
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.83
===================================================================================

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

Table 1.6. - ARIMA MODEL SUMMARY

The **RSME** value for the ARIMA Model = 135.906753 .

SARIMA MODEL:

Some parameter combinations for Model…
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (0, 1, 3)(0, 0, 3, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (1, 1, 3)(1, 0, 3, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
Model: (2, 1, 3)(2, 0, 3, 12)
Model: (3, 1, 0)(3, 0, 0, 12)
Model: (3, 1, 1)(3, 0, 1, 12)
Model: (3, 1, 2)(3, 0, 2, 12)
Model: (3, 1, 3)(3, 0, 3, 12)

	param	seasonal	AIC
206	(3, 1, 0)	(3, 0, 2, 12)	1035.710703
220	(3, 1, 1)	(3, 0, 0, 12)	1035.907113
222	(3, 1, 1)	(3, 0, 2, 12)	1035.909874
204	(3, 1, 0)	(3, 0, 0, 12)	1036.067322
205	(3, 1, 0)	(3, 0, 1, 12)	1036.545417

For SARIMA Model,

p = q = range(0, 4) d= range(0,2) D = range(0,2) pdq = list(itertools.product(p, d, q)) model_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, D, q))]

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  132
Model:             SARIMAX(3, 1, 0)x(3, 0, [1, 2], 12)   Log Likelihood             -508.855
Date:                            Sun, 25 Feb 2024   AIC                           1035.711
Time:                                    15:00:55   BIC                           1058.407
Sample:                                         0   HQIC                          1044.871
                                            - 132
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.3834      0.093     -4.125      0.000      -0.566      -0.201
ar.L2         -0.0143      0.094     -0.151      0.880      -0.199       0.171
ar.L3         -0.0529      0.086     -0.617      0.537      -0.221       0.115
ar.S.L12       0.8857      1.092      0.811      0.417      -1.254       3.026
ar.S.L24       0.2010      1.426      0.141      0.888      -2.594       2.995
ar.S.L36      -0.0254      0.314     -0.081      0.935      -0.641       0.590
ma.S.L12      -0.6244      1.041     -0.600      0.548      -2.664       1.415
ma.S.L24      -0.3756      1.155     -0.325      0.745      -2.640       1.889
sigma2      2986.2185      0.000   1.44e+07      0.000    2986.218    2986.219
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 1.67
Prob(Q):                              0.94   Prob(JB):                         0.43
Heteroskedasticity (H):               1.31   Skew:                            -0.05
Prob(H) (two-sided):                  0.46   Kurtosis:                         3.65
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 3.75e+24. Standard errors may be unstable.
```

Table 1.7 SARIMA MODEL SUMMARY

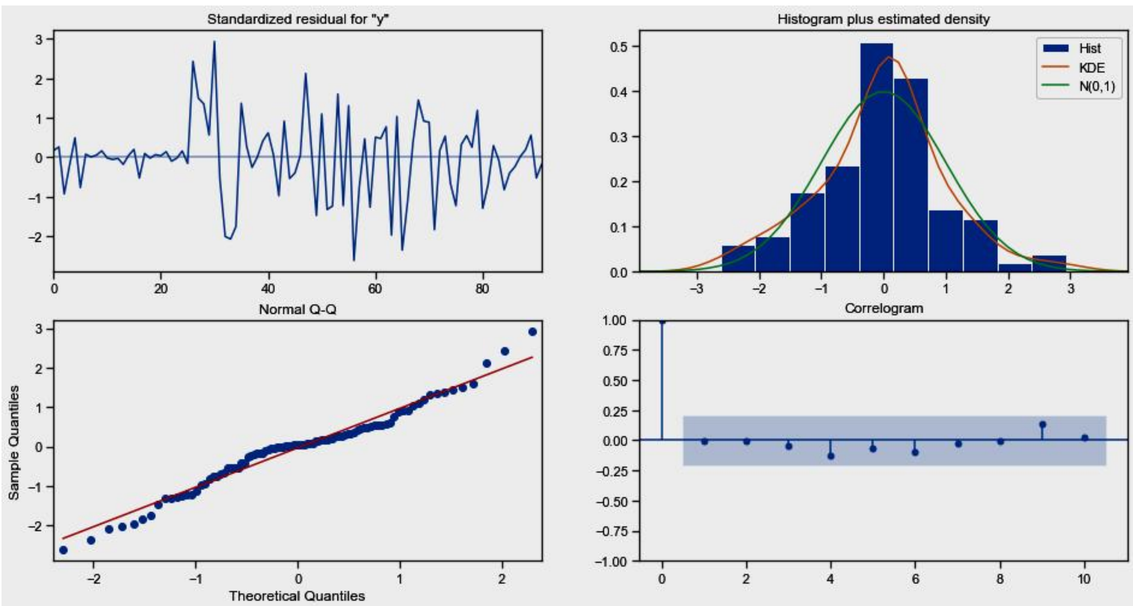

Fig. 1.22. - SARIMA Diagnostic Plot

The **RSME** value for (3,1,0),(3,0,2,12) SARIMA Model = 73.012517

1.7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

MODELS	TEST RSME
Alpha=0.1,Beta=0.6,Gamma=0.5,TripleExponentialSmoothing	45.832046
Simple Average Model	63.984570
(3,1,0),(3,0,2,12),Auto_SARIMA	73.012517
Linear Regression	73.111522
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	76.918569
Alpha=0.6,SimpleExponentialSmoothing	115.874445
Auto_ARIMA	135.906753
Alpha =0.5709, Beta = 0.000168, Gamma = 0.2031 Tripple Exponential Smoothing Model (Trend = Additive, Seasonality = Multiplicative) forecast on the Test Data	196.425507
Naive Model	245.121306

Table 1.8. - ALL MODELS SUMMARY - RSME Values

In the above table **lowest RSME value** model (**highlighted**) is in the top and best suited model to for time series prediction.

1.8. 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.

Year-Month	Sales_Predictions
1995-08-01	215.993616
1995-09-01	230.676643
1995-10-01	249.729067
1995-11-01	301.894467
1995-12-01	414.504032
1996-01-01	189.401997
1996-02-01	270.887545
1996-03-01	237.137077
1996-04-01	269.920268
1996-05-01	246.657350
1996-06-01	270.599996
1996-07-01	282.050536

Table 1.9. - Sale Prediction

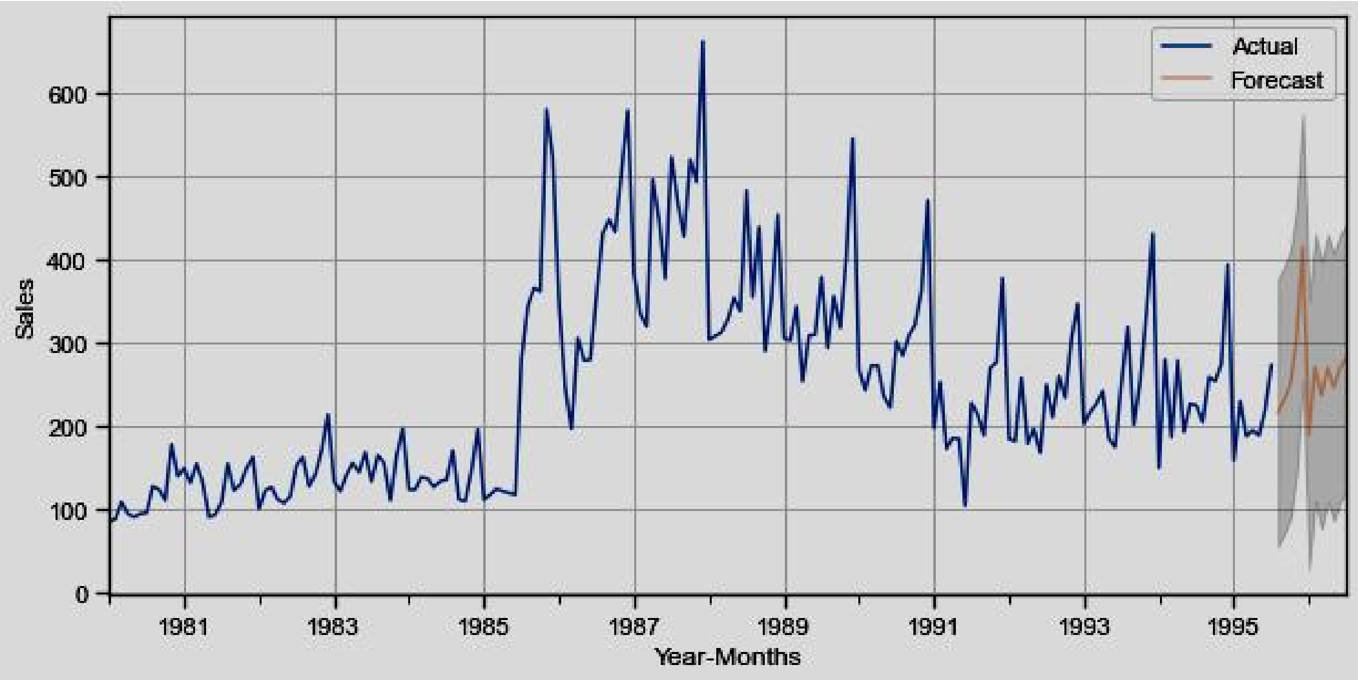

Fig 1.23 Sale Prediction Plot

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

- The shoe sales prediction looks similar to last year at a steady trend.
- The product didn't perform well at the early years after coming into the market,i.e., 1981 to 1985.
- The company profited with highest record above 600 sales between 1986 to 1990. The sales however declined over the years after 1991.
- Company hasn't sold more than 600 -700 units of their products as the data shows. Company should keep this in view and can work on this point to increase their sales above these numbers.
- There's in much of seasonal pattern but for the next 12 months from August 1995 to June 1996 it will go steady same as the previous years.
- Company should invest sin some new ad-campaigns for the first half of the year to increase its sale.
- Company can conduct survey on it's product qualities, latest market trend or fashion.
- Company can collect data based on gender and age, as which groups are purchasing more of their products, which can also throw some lights where their products lack compared to their competitors.
- Constant market trend and quality of the products is recommended company's future business plan, to stay ahead of the curve.

Problem 2 for the Data Set : SoftDrink.csv

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

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

In this report, we will focus on analyzing the soft-drink production data from January 1980 to July 1995. The task in hand is to review the given data over a period of time to identify patterns, trends, seasonality changes. This report aims to draw inferences and insights about the product production and their sales.

Note : *Here we are considering* ***production quantity*** *because as the* ***sales increases*** *production will also increase to meet the demand and vice-versa.*

This data set has 187 rows and 1 column :

1. Rows have the Yearly sales (Months and Dates) - YearMonth
2. Columns has the Production value - SoftDrinkProduction

YearMonth	SoftDrinkProduction
1980-01-01	1954
1980-02-01	2302

YearMonth	SoftDrinkProduction
1995-06-01	4365
1995-07-01	4290

Table 2.1. - Above table shows Head(Left) and Tail (right) of the data set

The dataset is further divided by extraction of month and year columns from the Year-Month column and renamed as Sales, Year and Month for better analysis of the given data set.

YearMonth	SoftDrinkProduction	Year	Month
1980-01-01	1954	1980	1
1980-02-01	2302	1980	2
1980-03-01	3054	1980	3
1980-04-01	2414	1980	4
1980-05-01	2226	1980	5

```
**Head of the given Dataset**
            Production  Year  Month
YearMonth
1980-01-01        1954  1980      1
1980-02-01        2302  1980      2
1980-03-01        3054  1980      3
1980-04-01        2414  1980      4
1980-05-01        2226  1980      5
**Tail of the given Dataset**
            Production  Year  Month
YearMonth
1995-03-01        4067  1995      3
1995-04-01        4022  1995      4
1995-05-01        3937  1995      5
1995-06-01        4365  1995      6
1995-07-01        4290  1995      7
```

Table 2.2. - New rows in the data set

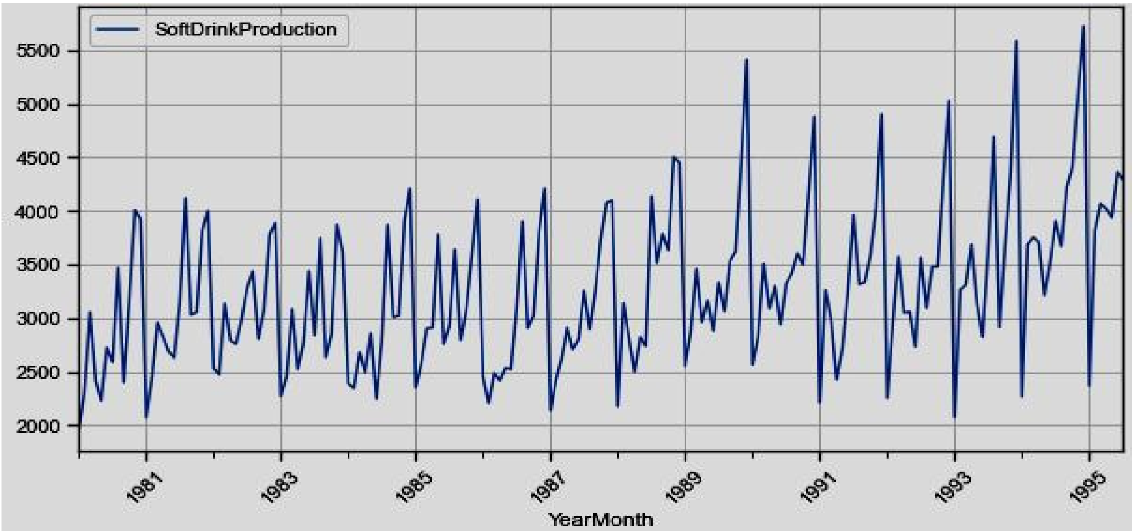

Fig. 2.1. - Time series Plot

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

Performing EDA,

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Production  187 non-null    int64
 1   Year        187 non-null    int64
 2   Month       187 non-null    int64
dtypes: int64(3)
memory usage: 5.8 KB
```

Table 2.3. - Data set info

	count	mean	std	min	25%	50%	75%	max
Production	187.0	3262.609626	728.357367	1954.0	2748.0	3134.0	3741.0	5725.0
Year	187.0	1987.299465	4.514749	1980.0	1983.0	1987.0	1991.0	1995.0
Month	187.0	6.406417	3.450972	1.0	3.0	6.0	9.0	12.0

Table 2.4. - Data set description

No Null Values found,

```
Production    0
Year          0
Month         0
dtype: int64
```

Boxplot,

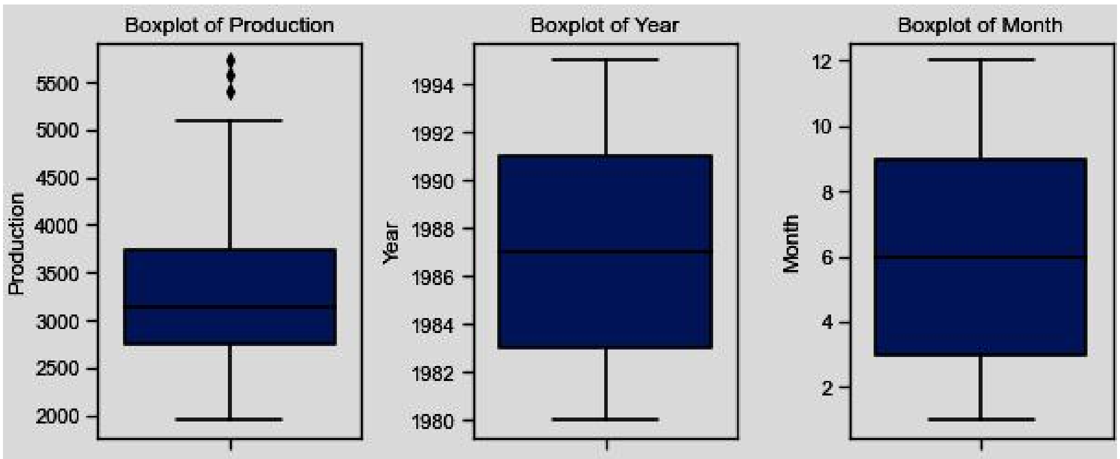

Fig. 2.2 Boxplot of the data

The box plot shows:

- Sales boxplot has outliers but we are choosing not to treat, as it will not have much impact on the time series analysis.

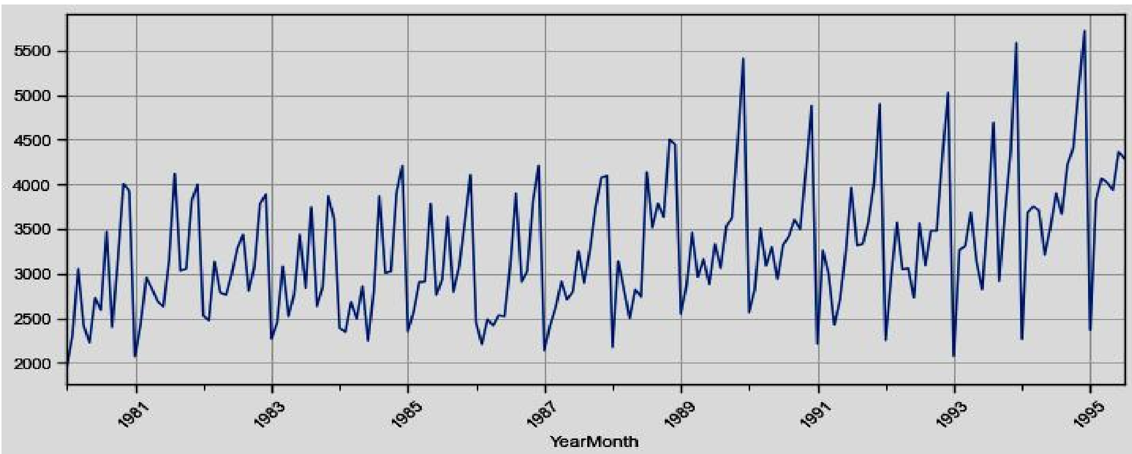

Fig. 2.3. - Time Series Plot of the Sales (with new columns)

The above plot shows the trend and seasonality pattern. It also shows that the peak of production, due to increase in demand and hence the sales, in the year 1990. After 1991, there is a slight decline in the pattern, but rise can be observed after 1994. It's on rise after 1994 with a similar patter. Increase of production is observed most during the second-half of the year and decrease during the first-half.

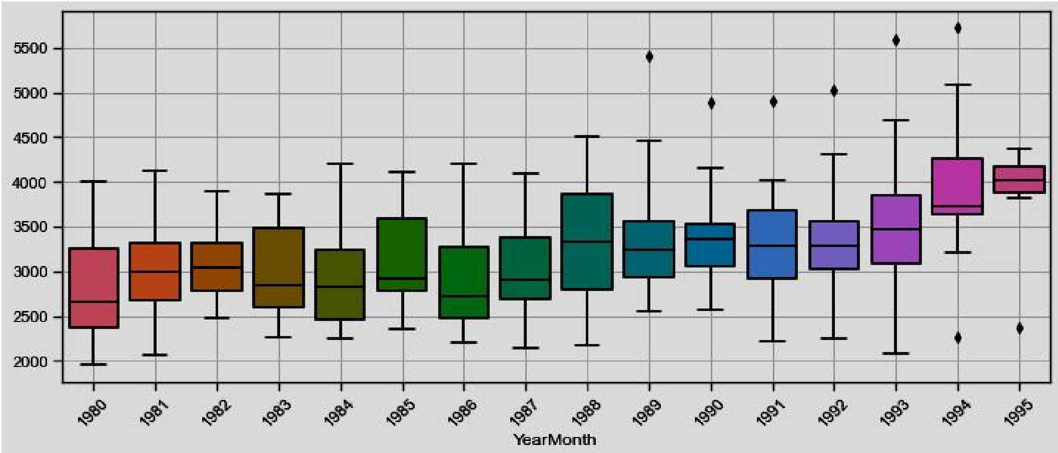

Fig. 2.4. - Boxplot of the Production - Yearly

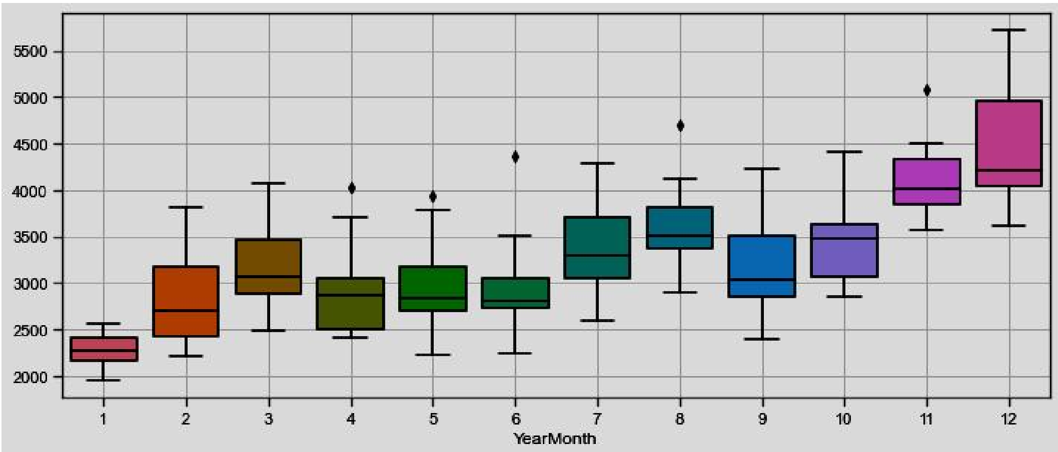

Fig. 2.5. - Boxplot of the Production - Monthly

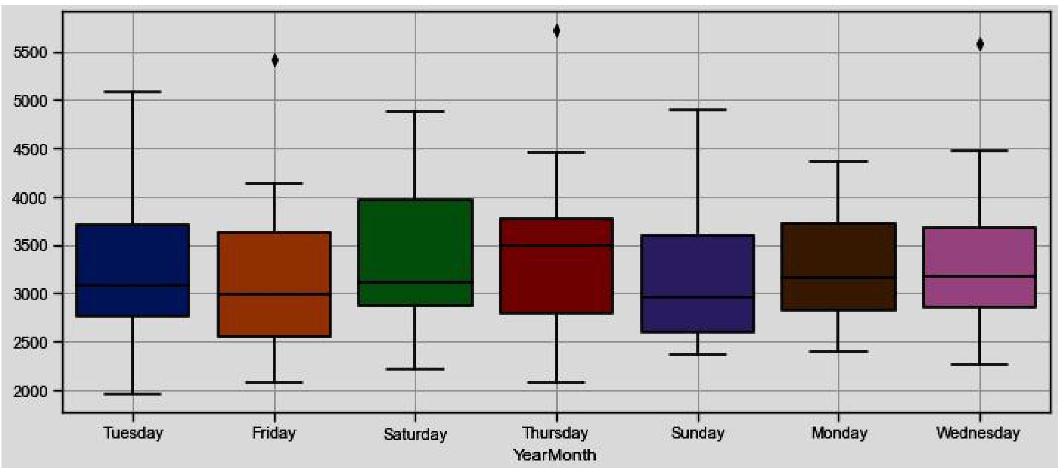

Fig. 2.6. - Boxplot of the Production - Weekly

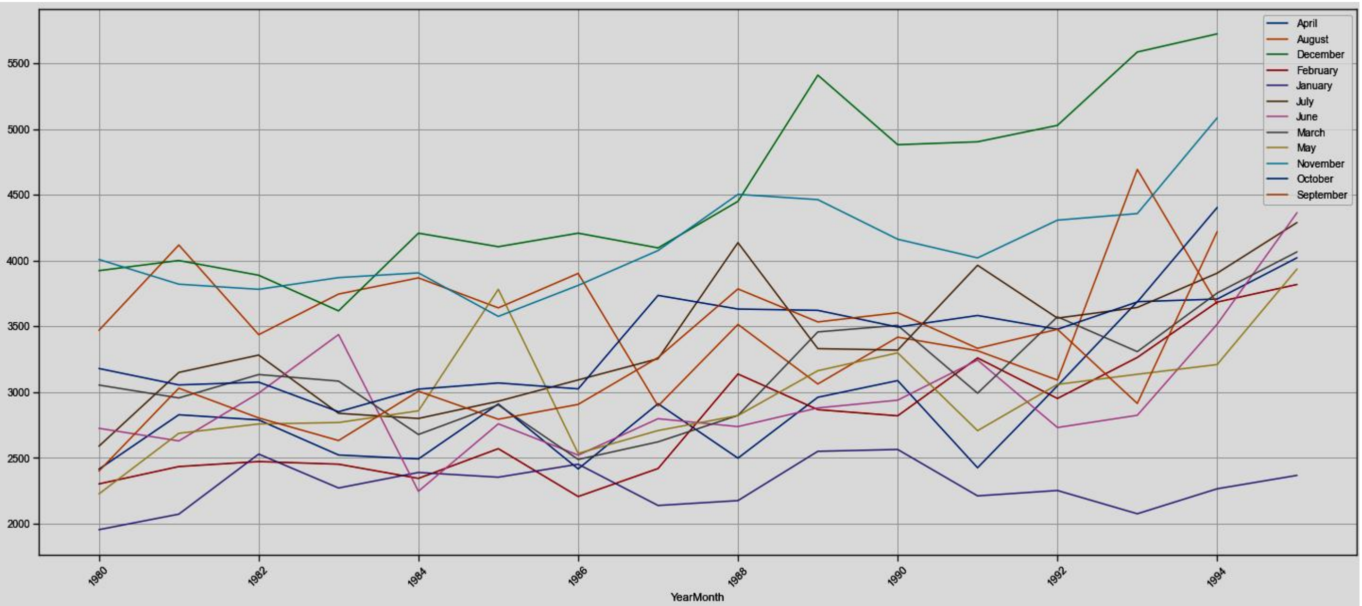

Fig. 2.7. - Monthly Sales over the Years

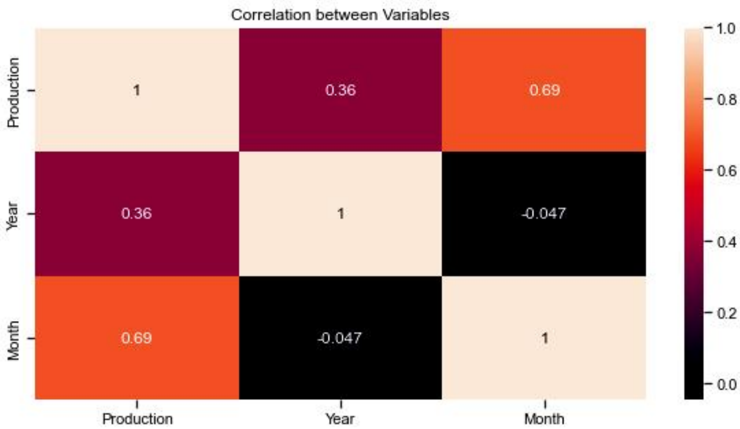

Fig. 2.8. - Correlation Heat Map

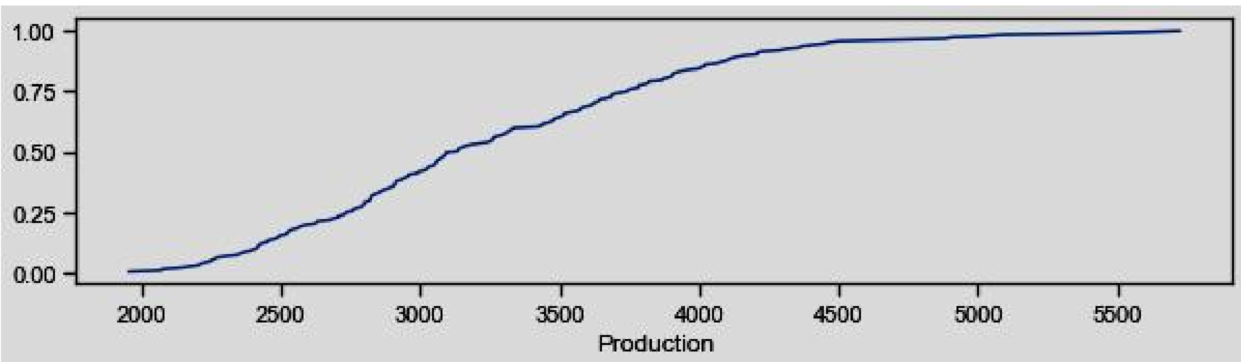

Fig. 2.9. - Empirical Cumulative Distribution Function Curve Plot

This plot shows the distribution of production over the years in the data :

- Around 73% of the production has been less than 4000.
- 5500 is the highest production value achieved over the years.

Decomposition of the Time series,

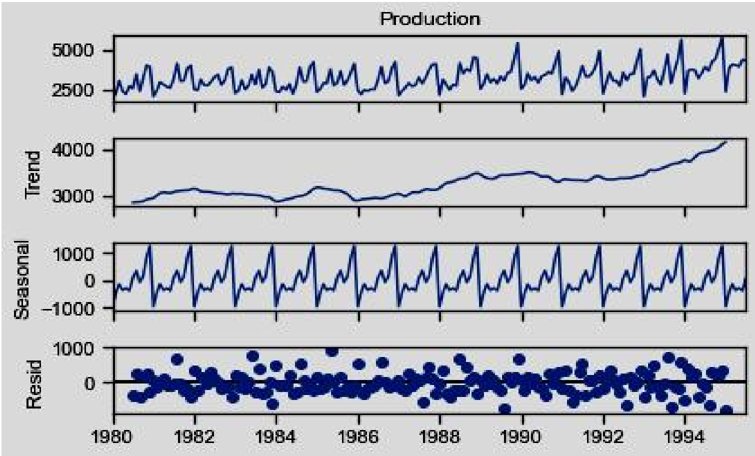

Fig.2.10. - Time Decomposition - Additive

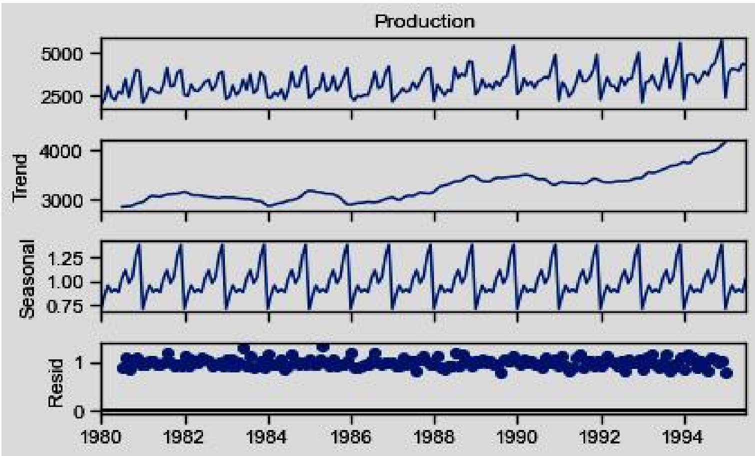

Fig.2.11. - Time Decomposition - Multiplicative

The above plots show:

- Peak year of sales is between the years 1992 to 1995.
- Both trend and seasonality are present.
- The trend gradually on the rise from the years after 1988.
- Residue is spread and is not in a straight line for **additive** and the residue is almost in a straight line for **multiplicative.**
- Residue for multiplicative is between 0 to 1, while for additive it is between 0 to 1000.

So the multiplicative model has more stable residual plot and lower range of residuals. This means that most of the observations are properly captured by the model as it's almost linear and not too much spread.

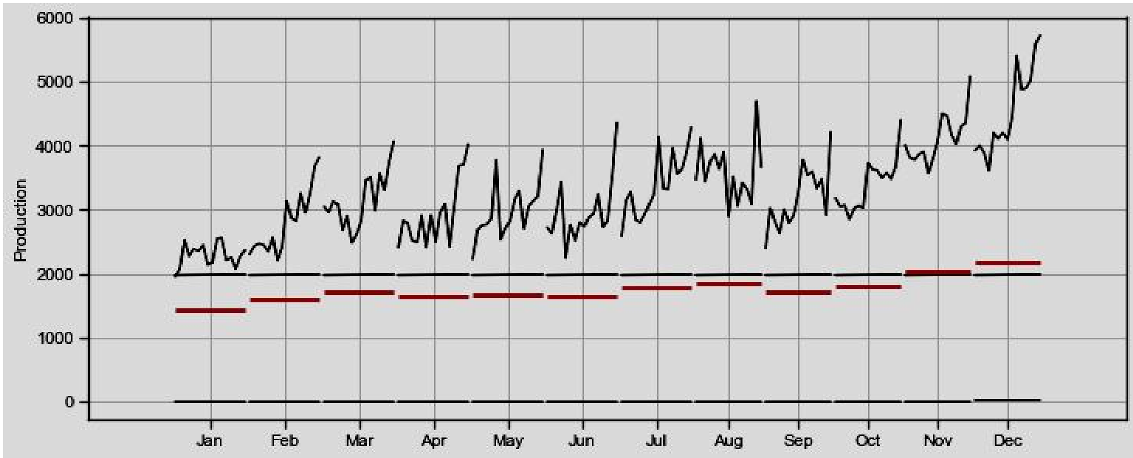

Fig.2.12. - Monthly Plot

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

Now , we have to split the data set into training and test set for model building analysis.

As per instructions, data is split from 1980-1990 is train data set, then 1991 to 1995 is the test data set.
Rows & Columns for train-test date set:

- Train dataset has 132 rows and 3 columns.
- Test dataset has 55 and 3 columns.

First few rows of Training Data

YearMonth	Production	Year	Month
1980-01-01	1954	1980	1
1980-02-01	2302	1980	2
1980-03-01	3054	1980	3
1980-04-01	2414	1980	4
1980-05-01	2226	1980	5

Last few rows of Training Data

YearMonth	Production	Year	Month
1990-08-01	3418	1990	8
1990-09-01	3604	1990	9
1990-10-01	3495	1990	10
1990-11-01	4163	1990	11
1990-12-01	4882	1990	12

First few rows of Test Data

YearMonth	Production	Year	Month
1991-01-01	2211	1991	1
1991-02-01	3260	1991	2
1991-03-01	2992	1991	3
1991-04-01	2425	1991	4
1991-05-01	2707	1991	5

Last few rows of Test Data

YearMonth	Production	Year	Month
1995-03-01	4067	1995	3
1995-04-01	4022	1995	4
1995-05-01	3937	1995	5
1995-06-01	4365	1995	6
1995-07-01	4290	1995	7

Table 2.5. Train-Test data set (head & tail)

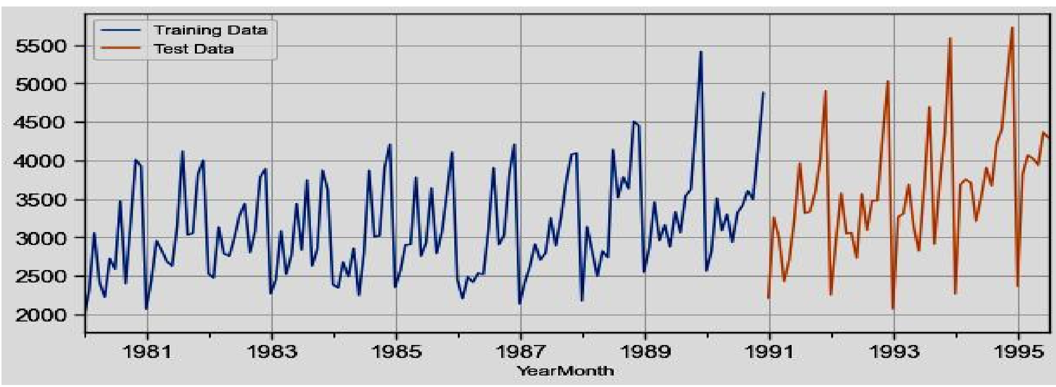

Fig.2.13. - Train-Test Plot

2.4. **Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression,Naïve forecast models, and simple average models. should also be built on the training data and check the performance on the test data using RMSE.**

<u>Model 1 : Linear Regression Model</u>

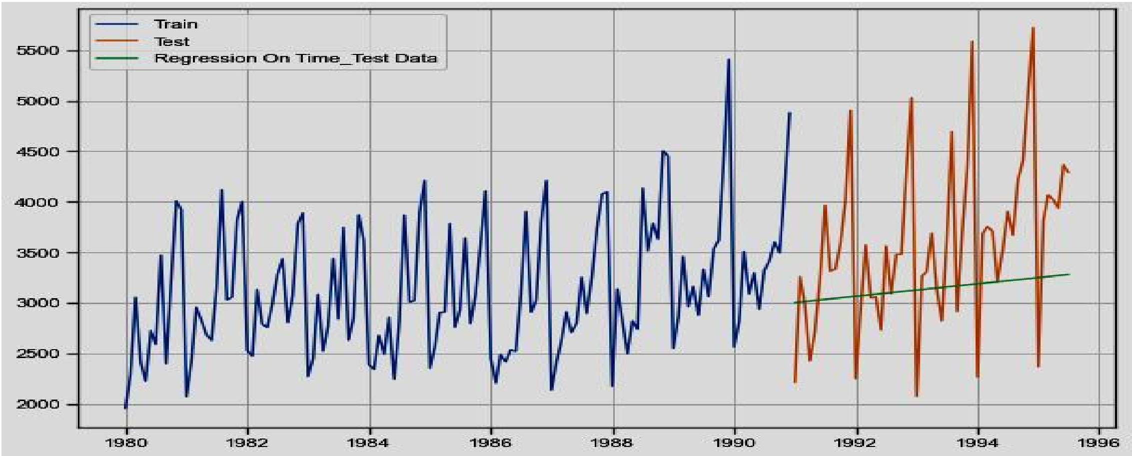

Fig. 2.14. - Linear Regression Plot

From the above plot,
The green line shows the prediction made by the model and the orange lines shows the actual test values.
The predicted values are really far from the actual test values with an upward trend.
The RSME value for the Linear Reg Model = 898.17

<u>Model 2 : Naïve forecast Model</u>

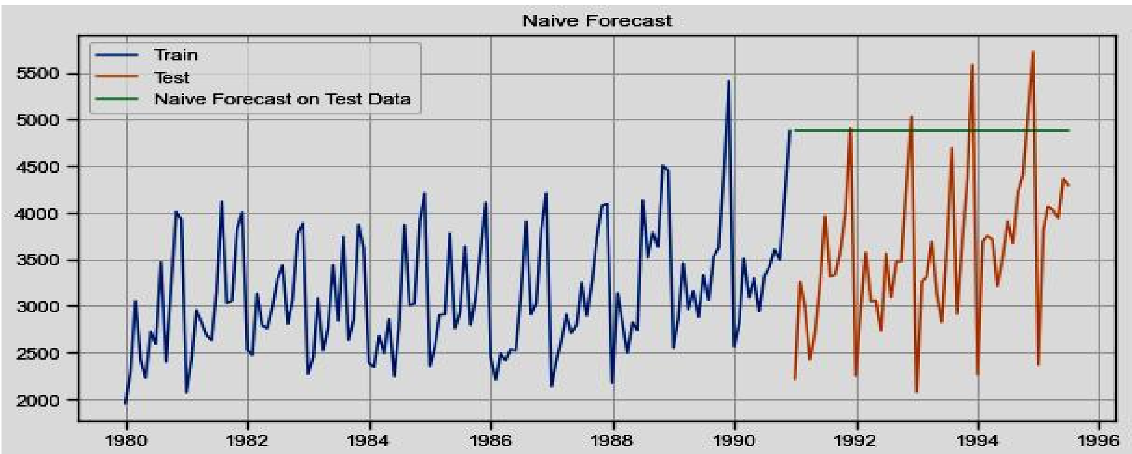

Fig. 2.15. - Naïve forecast Model Plot

From the above plot,
The green line shows the prediction made by the model and the orange lines shows the actual test values.
The predicted values are really far from the actual test values.
The RSME value for the Naïve Model = 1519.26

Model 3: Simple Average

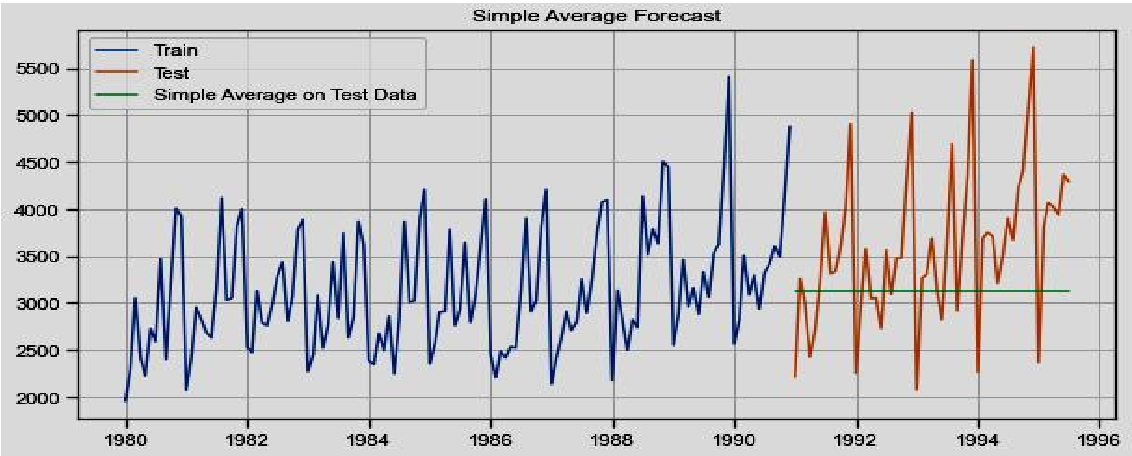

Fig. 2.16. - Simple Average Model Plot

From the above plot,
The green line shows the prediction made by the model and the orange lines shows the actual test values. The predicted values are really far from the actual test values.

The RSME value for the Simple Average Model = 934.35

Model 4: Simple Exponential Smoothing (SES) Model

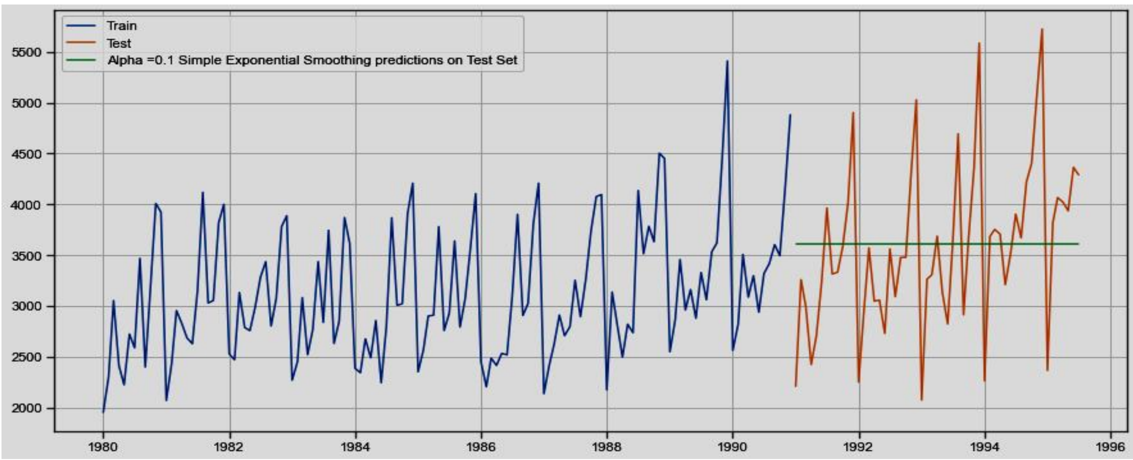

Fig. 2.17. - SES Model Plot

From the above plot,
Using auto-fit and finding the best parameters for this model, green line shows the prediction made by the model at alpha = 0.1 and the green line at alpha = 0.1 is considered the best among all the other alpha values.

RSME value for SES Model (at alpha - 0.1) = 807.35

Model 5: Double Exponential Smoothing - Holt's Model (DES)

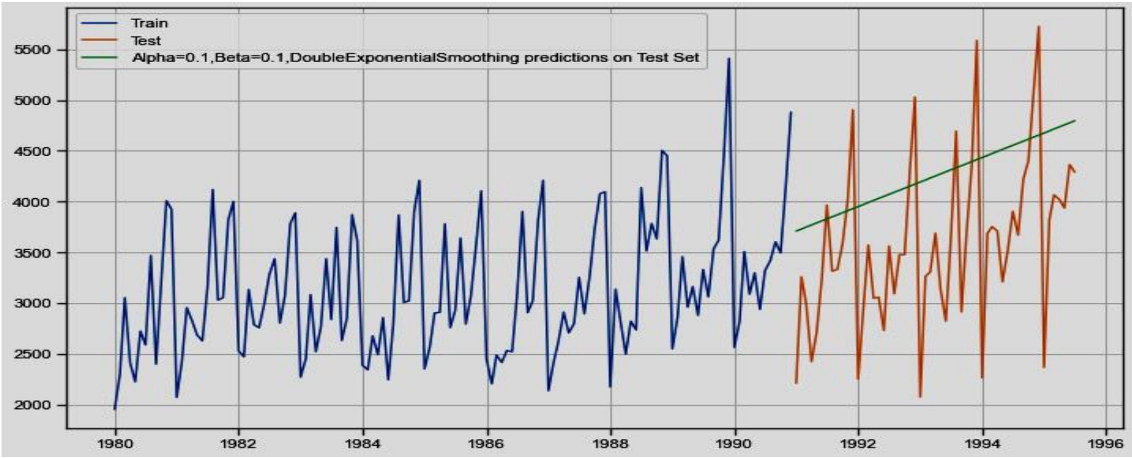

Fig. 2.18. - DES Model Plot

From the above plot,
The green line shows the prediction made by the model and the orange lines shows the actual test values. The predicted values are really far from the actual test values.

For this plot **alpha = 0.1** and **Beta = 0.1** is considered, the RSME value for the Holt's Model = 982.94

Model 7: Triple Exponential Smoothing a.k.a Holt - Winter's Model(TES)

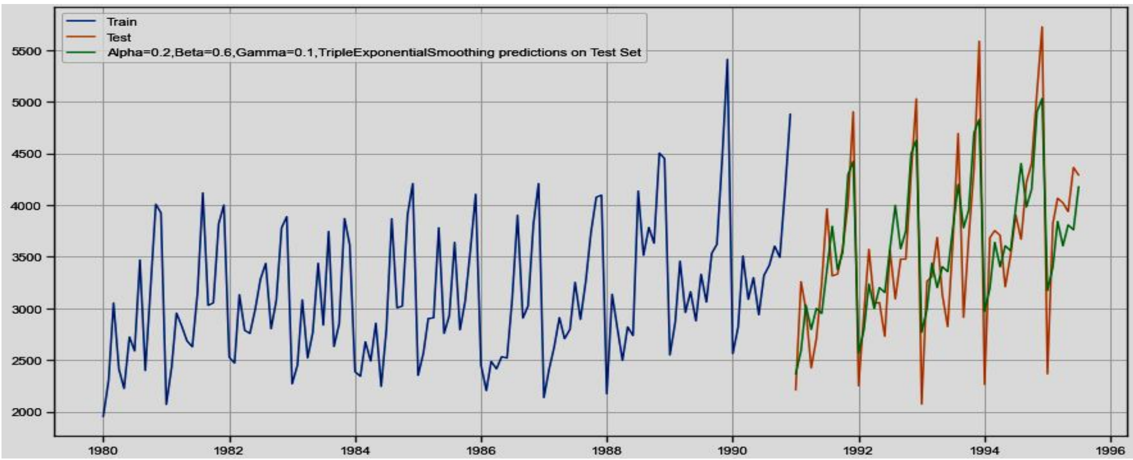

Fig. 2.19. - TES Model Plot

From the above plot (only best Parameters considered),
Here we have considered **Alpha = 0.2**, **Beta = 0.6** and **Gamma = 0.1**,

The **RSME** value for the TES Model = 421.581209

This is the best model which has multiplicative trend and additive seasonality.

2.5. Check for the stationary 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 stationary and comment. Note: Stationary should be checked at alpha = 0.05.

Applying Augmented Dickey-Fuller test whether the series has unit and whether it is stationary or non-stationary. Hypothesis for the ADF test:

- ***H0 : The Time Series has a unit root and is thus non-stationary (Null Hypothesis)***
- ***H1 : The Time Series does not have a unit root and is thus stationary. (Alternate Hypothesis)***

We see that for α = 5% (significance), the Time Series is non-stationary.

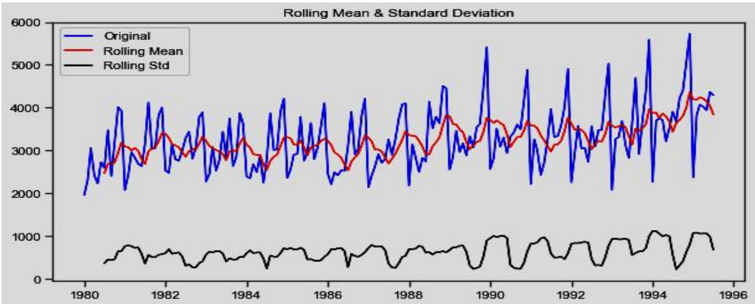

Fig. 2.20. - Augmented Dickey-Fuller Test Plot

Test Output ,

```
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
```

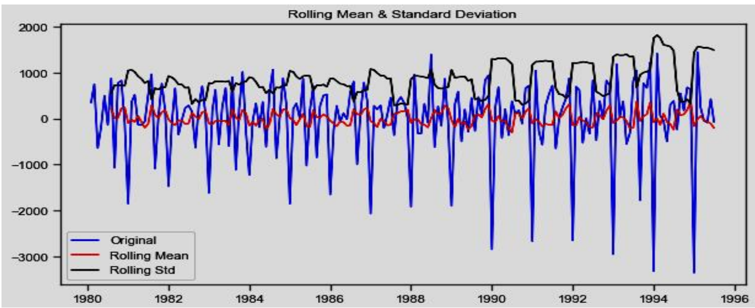

Fig. 2.21 Augmented Dickey-Fuller Test Plot - .diff() Method

Test Output ,

```
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
```

From the above output values, we can **reject** the **null hypothesis** that the series **not stationary**.

We can now **build ARIMA/ SARIMA models**, as we have proven that **the series is stationary**.

2.6. 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.

ARIMA MODEL:

Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)

	Param	AIC
2	(0, 1, 2)	2056.489263
6	(1, 1, 2)	2056.715682
3	(0, 1, 3)	2056.831789
11	(2, 1, 3)	2057.088851
13	(3, 1, 1)	2058.304546
7	(1, 1, 3)	2058.712159
10	(2, 1, 2)	2058.712702
9	(2, 1, 1)	2059.100672
15	(3, 1, 3)	2059.590769
14	(3, 1, 2)	2060.679966
5	(1, 1, 1)	2061.523084
1	(0, 1, 1)	2069.59963
12	(3, 1, 0)	2070.365367
8	(2, 1, 0)	2073.234861
4	(1, 1, 0)	2097.872122
0	(0, 1, 0)	2103.733834

Applying for loop for determining the optimum values of ***p***,***d***,***q*** where ***p*** is the order of the ***AR*** **(Auto-Regressive)** part of the model, while ***q*** is the order of the ***MA*** **(Moving Average)** part of the model and ***d*** is the **difference that is required** to make the series stationary.

p and ***q*** values are in the range of (0,4) were given to the for loop, while a fixed value of 1 was given for ***d***, since we had already determined ***d*** to be 1, while checking for stationary using the ADF test.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:             Production   No. Observations:                  132
Model:                 ARIMA(0, 1, 2)   Log Likelihood               -1025.245
Date:                Sun, 25 Feb 2024   AIC                           2056.489
Time:                        17:20:55   BIC                           2065.115
Sample:                    01-01-1980   HQIC                          2059.994
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.5407      0.085     -6.392      0.000      -0.707      -0.375
ma.L2         -0.3913      0.113     -3.475      0.001      -0.612      -0.171
sigma2      3.572e+05   4.62e+04      7.725      0.000    2.67e+05    4.48e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.61   Jarque-Bera (JB):                 0.39
Prob(Q):                              0.44   Prob(JB):                         0.82
Heteroskedasticity (H):               1.31   Skew:                            -0.13
Prob(H) (two-sided):                  0.37   Kurtosis:                         2.91
===================================================================================

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

Table 2.6. - ARIMA MODEL SUMMARY

The **RSME** value for the ARIMA Model = 831.62

SARIMA MODEL:

Some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (0, 1, 3)(0, 0, 3, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (1, 1, 3)(1, 0, 3, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
Model: (2, 1, 3)(2, 0, 3, 12)
Model: (3, 1, 0)(3, 0, 0, 12)
Model: (3, 1, 1)(3, 0, 1, 12)
Model: (3, 1, 2)(3, 0, 2, 12)
Model: (3, 1, 3)(3, 0, 3, 12)

	Param	Seasonal	AIC
252	(3, 1, 3)	(3, 0, 0, 12)	1350.316002
220	(3, 1, 1)	(3, 0, 0, 12)	1354.246924
221	(3, 1, 1)	(3, 0, 1, 12)	1355.642079
236	(3, 1, 2)	(3, 0, 0, 12)	1355.766143
254	(3, 1, 3)	(3, 0, 2, 12)	1356.662371

For SARIMA Model,

p = q = range(0, 4) d= range(0,2) D = range(0,2) pdq = list(itertools.product(p, d, q)) model_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, D, q))]

```
                                   SARIMAX Results
==========================================================================================
Dep. Variable:                                y   No. Observations:                  132
Model:             SARIMAX(3, 1, 3)x(3, 0, [], 12)   Log Likelihood                -665.158
Date:                          Sun, 25 Feb 2024   AIC                           1350.316
Time:                                  17:33:00   BIC                           1375.534
Sample:                                       0   HQIC                          1360.494
                                          - 132
Covariance Type:                            opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9032      0.152      5.933      0.000       0.605       1.202
ar.L2         -0.9168      0.132     -6.964      0.000      -1.175      -0.659
ar.L3          0.2060      0.142      1.452      0.146      -0.072       0.484
ma.L1         -1.8094      0.111    -16.344      0.000      -2.026      -1.592
ma.L2          1.8167      0.220      8.260      0.000       1.386       2.248
ma.L3         -0.9482      0.155     -6.128      0.000      -1.251      -0.645
ar.S.L12       0.4318      0.117      3.680      0.000       0.202       0.662
ar.S.L24       0.3456      0.125      2.773      0.006       0.101       0.590
ar.S.L36       0.2215      0.123      1.797      0.072      -0.020       0.463
sigma2       1.03e+05    4.8e-06   2.15e+10      0.000    1.03e+05    1.03e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                 3.40
Prob(Q):                              0.84   Prob(JB):                         0.18
Heteroskedasticity (H):               1.34   Skew:                             0.30
Prob(H) (two-sided):                  0.42   Kurtosis:                         3.72
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.27e+26. Standard errors may be unstable.
```

Table 2.7 SARIMA MODEL SUMMARY

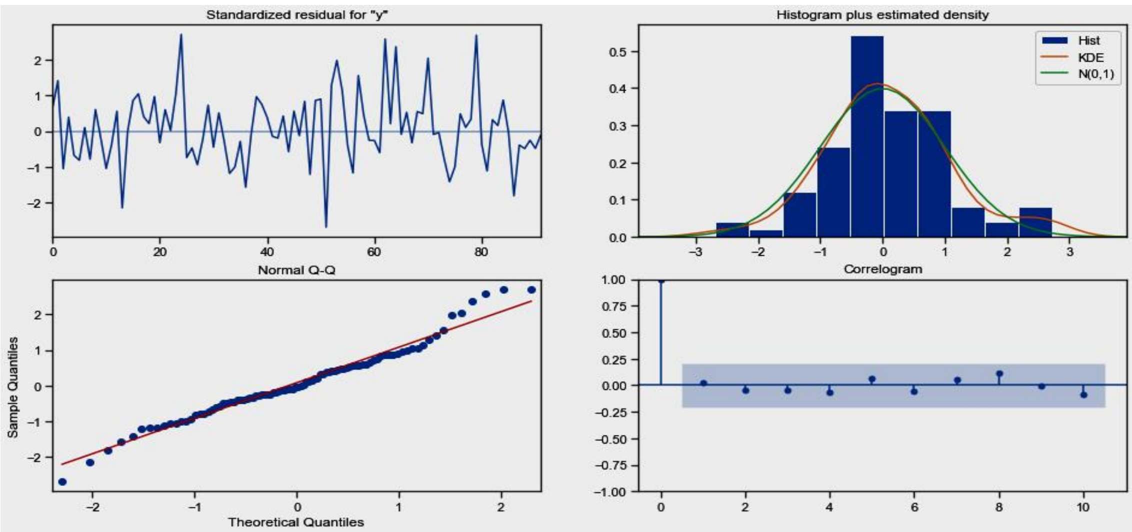

Fig. 2.22. - SARIMA Diagnostic Plot

The **RSME** value for (3,1,3),(3,0,0,12) SARIMA Model = 429.436551

2.7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

MODELS	TEST RSME
Alpha=0.2,Beta=0.6,Gamma=0.1,TripleExponentialSmoothing	**421.581209**
(3,1,3),(3,0,0,12),Auto_SARIMA	429.436551
Alpha=0.1,SimpleExponentialSmoothing	807.346865
Alpha = 0.1464, Beta = 0.0399, Gamma = 0.2626 Tripple Exponential Smoothing Model (Trend = Multiplicative, Seasonality = Additive) forecast on the Test Data	819.401216
Auto_ARIMA	831.615849
Linear Regression	898.172528
Simple Average Model	934.353358
Naive Model	1519.259233

Table 2.8. - ALL MODELS SUMMARY - RSME Values

In the above table **lowest RSME value** model (**highlighted**) is in the top and best suited model to for time series prediction.

2.8. 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.

Year-Month	Production_Preds
1995-08-01	4894.949815
1995-09-01	4650.259062
1995-10-01	4958.306778
1995-11-01	5750.734987
1995-12-01	6261.721837
1996-01-01	3921.951564
1996-02-01	4659.263043
1996-03-01	5000.732030
1996-04-01	4843.913920
1996-05-01	4907.651876
1996-06-01	5005.094416
1996-07-01	5549.053625

Table 2.9. - Production Prediction

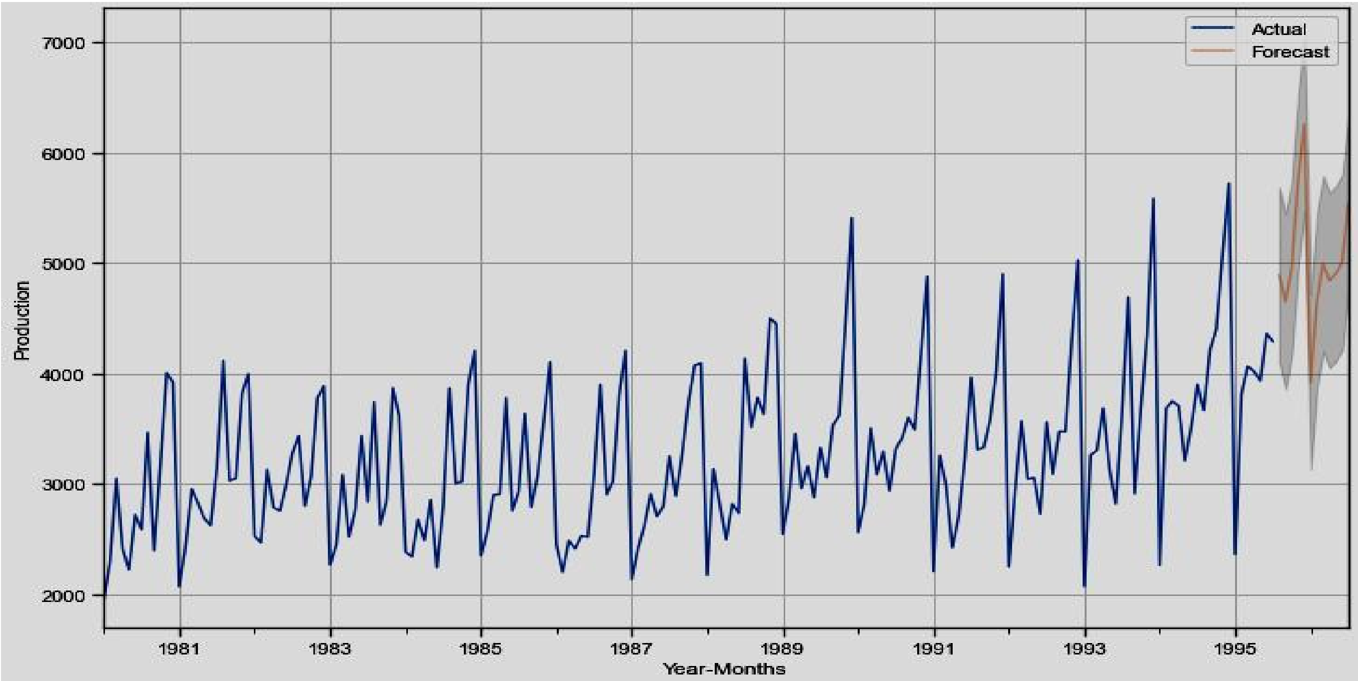

Fig 2.23 Production Prediction Plot

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

- The prediction plot indicates as steady growth in production of soft-drinks over the years and highest between 1989 to 1991.

- The data suggest that the sales will increase over the next 12 months after August 1997 and company should maintain their quality and work on improving their products.

- The trend suggest that the sales are higher in the second half of the year than the first half.

- Business should consider running ad-campaigns for first half of the years and may offer attractive deals or offs during the festivities in the first of the year and take advantage of these periods. This may boost their sales in the first half than later

- There is high growth in production showing after 1995 and company must prepare to meet these demands for better sale and increase their profit margins by making slight changes in their prices.

- Company can introduce a new products during this period and can take customer reviews on their current products as well.