ARPITA BAYEN

TIME SERIES FORECASTING

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

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

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

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.

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.

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.

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

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.

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

SHOE SALES DATASET

SHOE SALES DATASET

**Problem 1: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.**

1.READ THE DATA AS AN APPROPRIATE TIME SERIES DATA AND PLOT THE DATA.

- The dataset contains 2 variables and 187 data(rows).
- 1 Variable is year month and 1 variable is Sales report of the shoes .
- Year month is object type variable and Shoe-Sales is integer type.
- The year month column is divided to year and month column-now the dataset contains 3 columns.
- There is no null values in the dataset.
- There are no Duplicate values.

Table 1:First Few rows of the Dataset:

	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

Table 2:Last few rows of the dataset

	Shoe_Sales
YearMonth	
1995-03-01	188
1995-04-01	195
1995-05-01	189
1995-06-01	220
1995-07-01	274

Table 3:First few rows of the dataset after adding 2 new column –Month and year

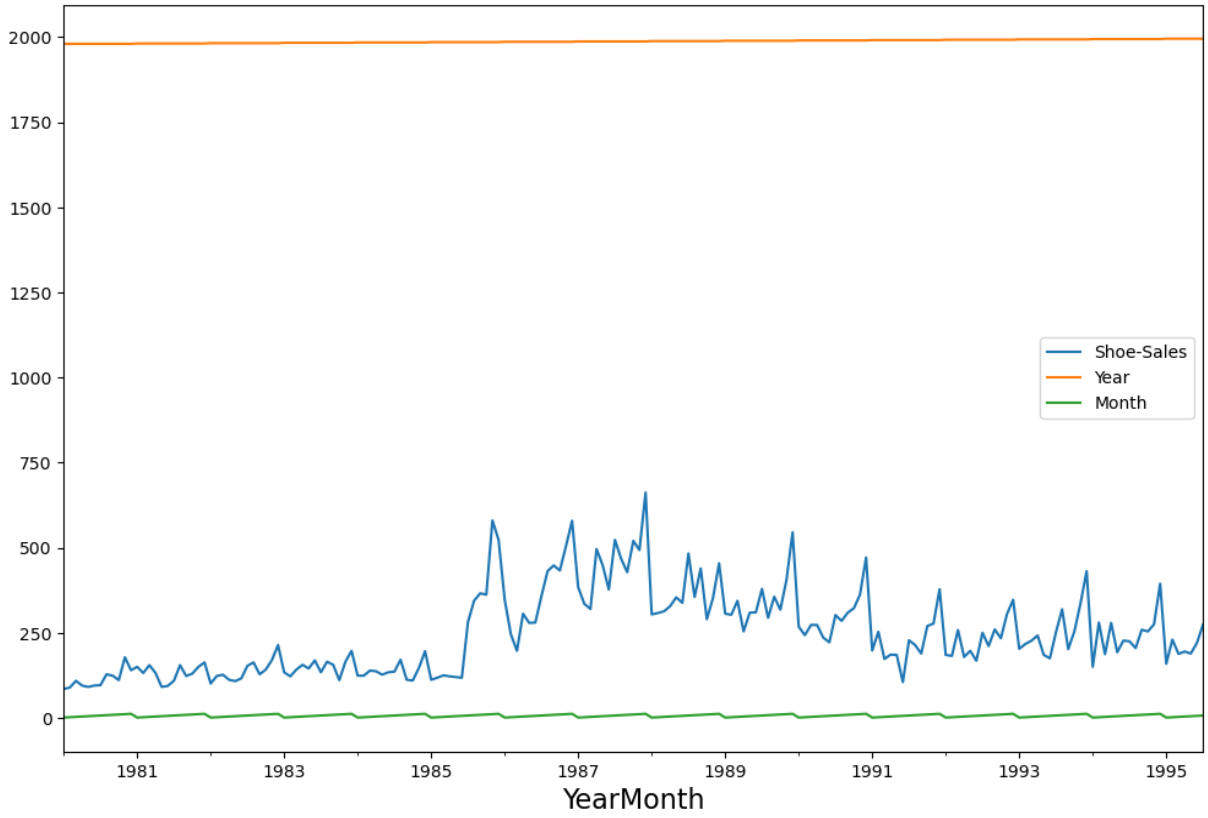
	Shoe_Sales	Year	Month
YearMonth			
1980-01-01	85	1980	1
1980-02-01	89	1980	2
1980-03-01	109	1980	3
1980-04-01	95	1980	4
1980-05-01	91	1980	5

Table 4:Statistical description of the Dataset

	count	mean	std	min	25%	50%	75%	max
Shoe-Sales	187.0	246.0	121.0	85.0	144.0	220.0	316.0	662.0
Year	187.0	1987.0	5.0	1980.0	1983.0	1987.0	1991.0	1995.0
Month	187.0	6.0	3.0	1.0	3.0	6.0	9.0	12.0

EDA

Fig 1:Graph of Shoe-Sales Dataset



This is showing how the shoe sales is varying through out the year and months.

Fig 2:Boxplot of the dataset

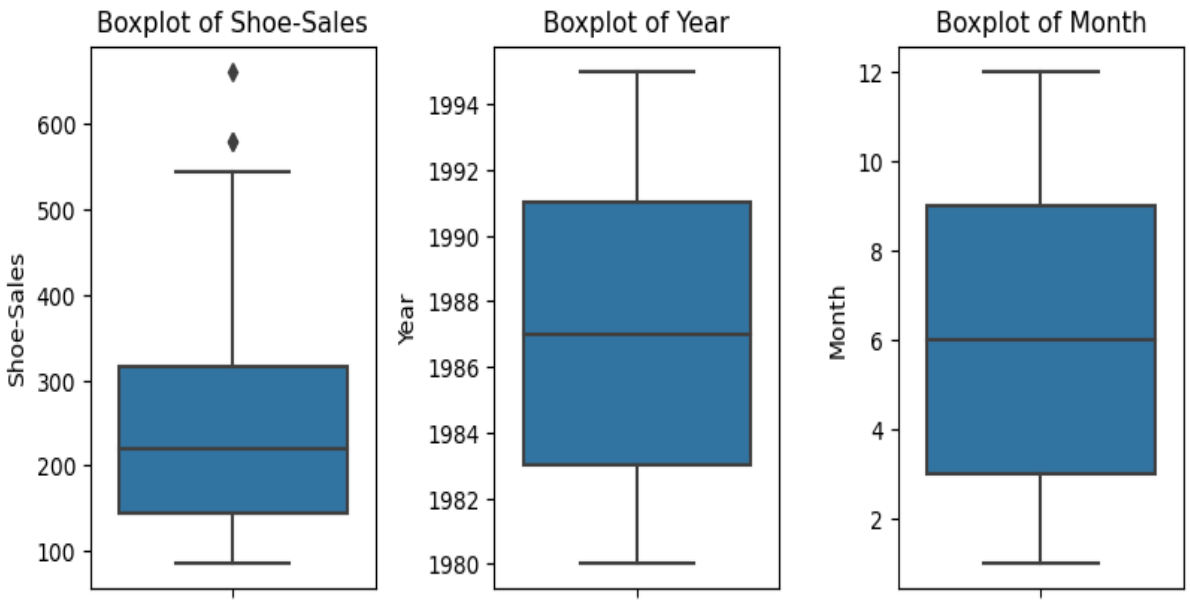
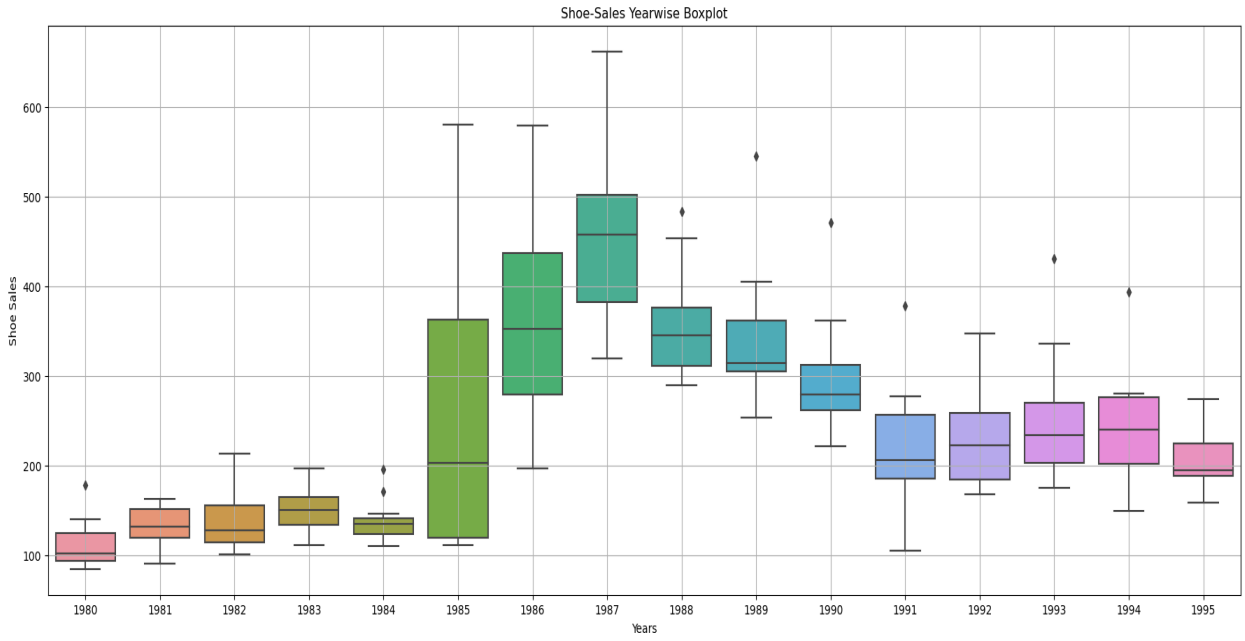
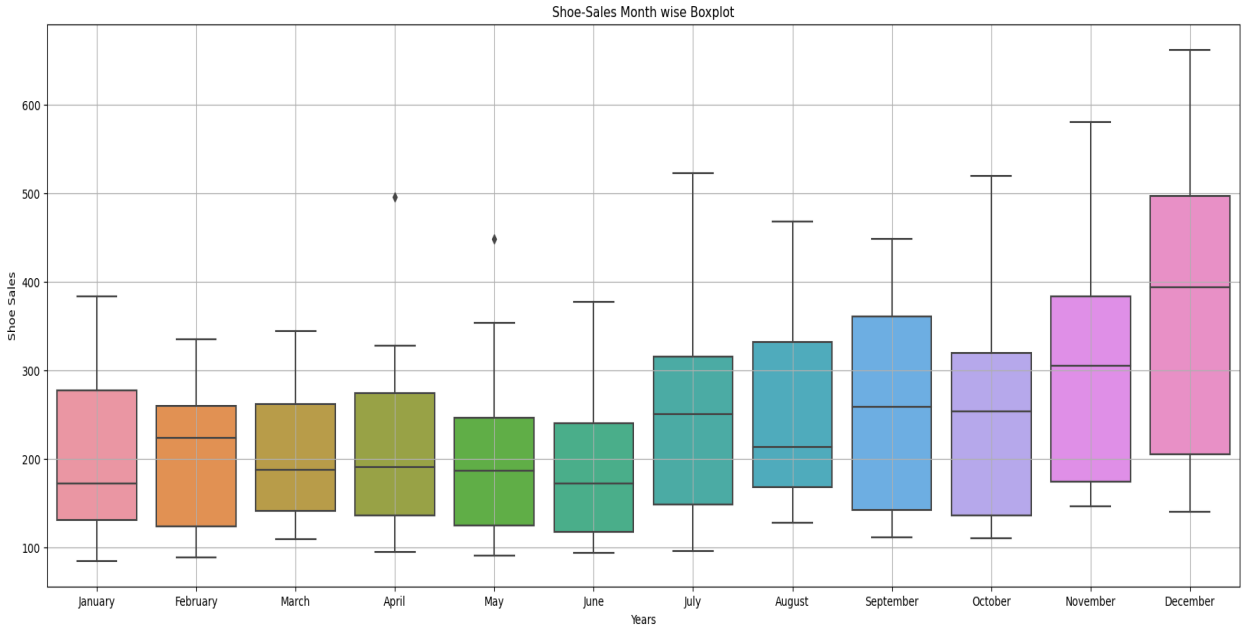


Fig 3:Yearwise boxplot of shoe sales



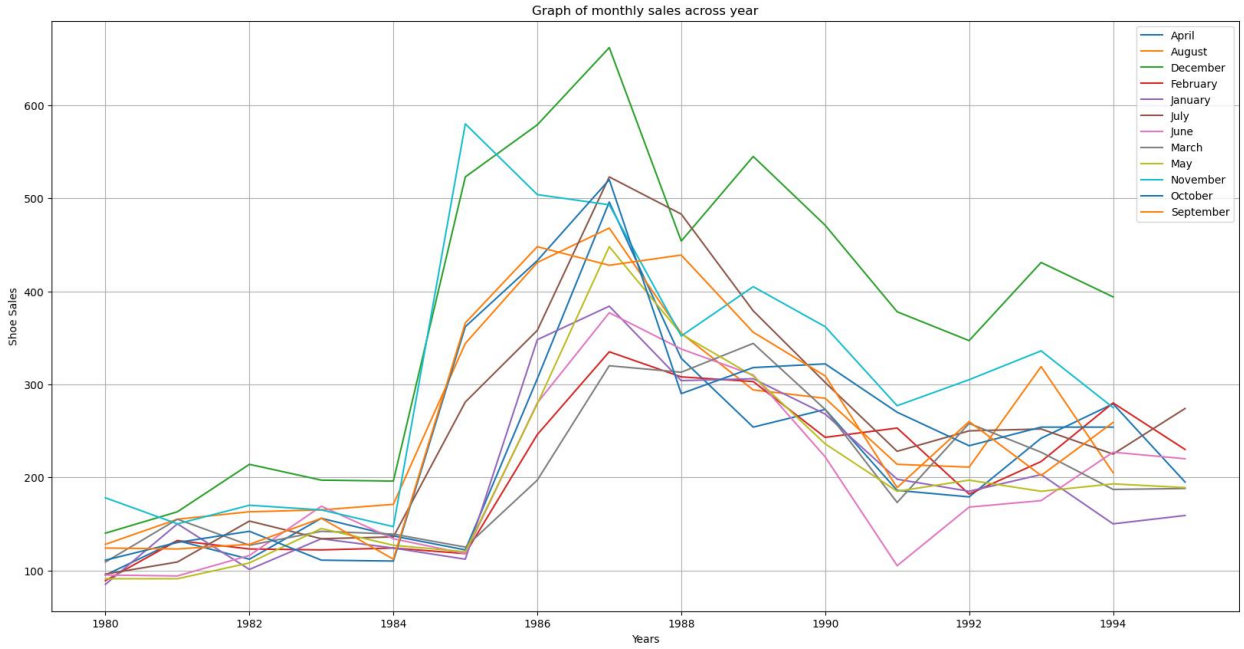
The sales increased in the year 1985,1986,1987.

Fig 4:Month wise boxplot



The sales increased during the last few months of the year.Highest sales is in December month.

Fig 5: Graph of monthly sales across year(Pivot Graph)



DECOMPOSITION:

The dataset is decomposed in to additively and multiplicatively

Fig 6 :Additive decomposition of shoe sales dataset

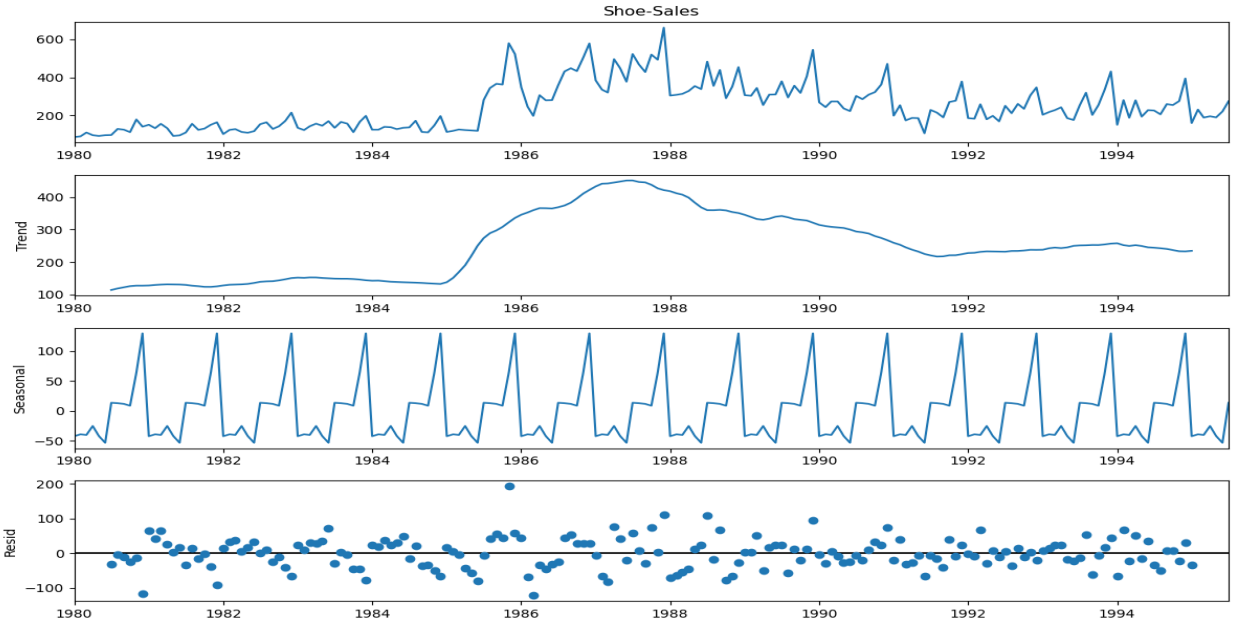
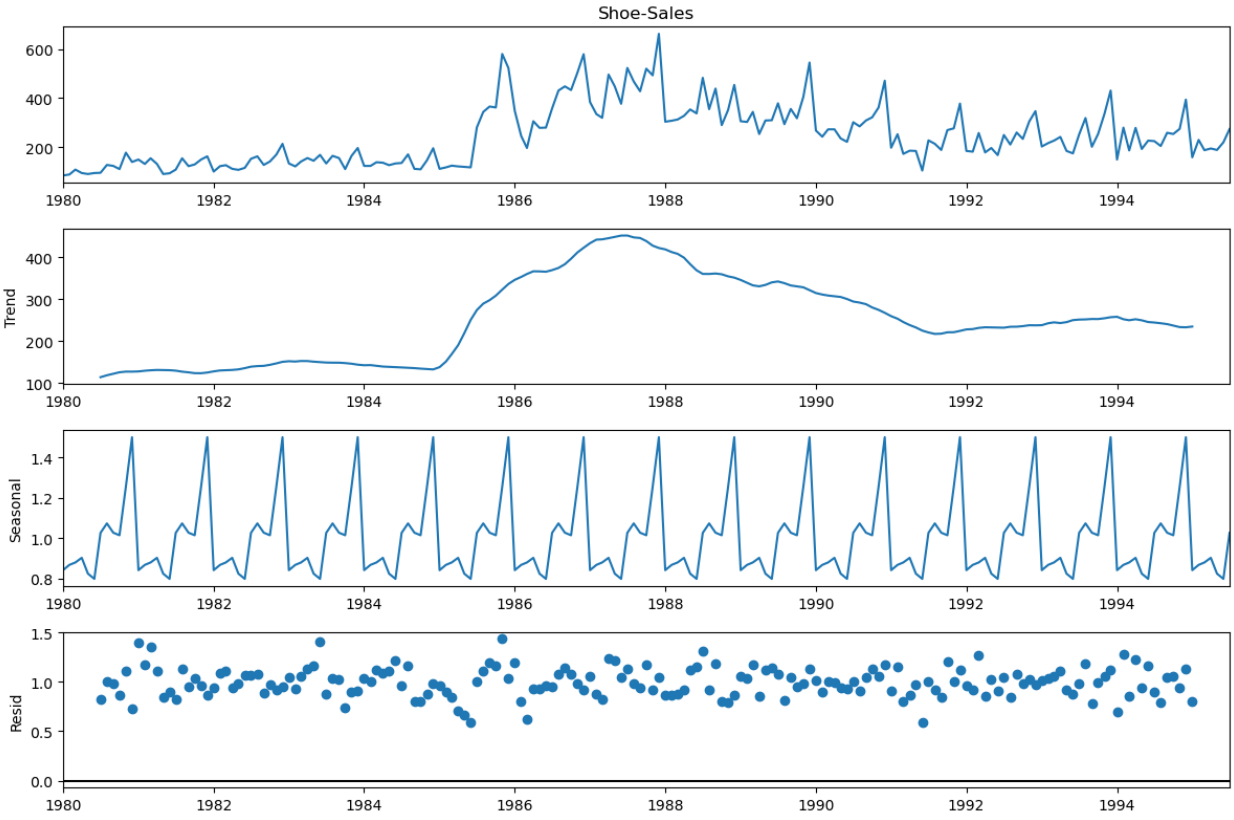


Fig 7:Multiplicative decomposition of shoe sales dataset



Some of the key observations from this analysis:

a) Trend: 12-months MA is not linear which doesnot shows any trend.

b) Seasonality: seasonality of 12 months is clearly visible

c) Irregular Remainder (random): The multiplicative model works as there are no patterns in the residuals

3. SPLIT THE DATA INTO TRAINING AND TEST. THE TEST DATA SHOULD START IN 1991.

The Dataset is splitted into test and train set.The test dataset starts from 1991.

Table 5:First few rows of training dataset

	Shoe-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

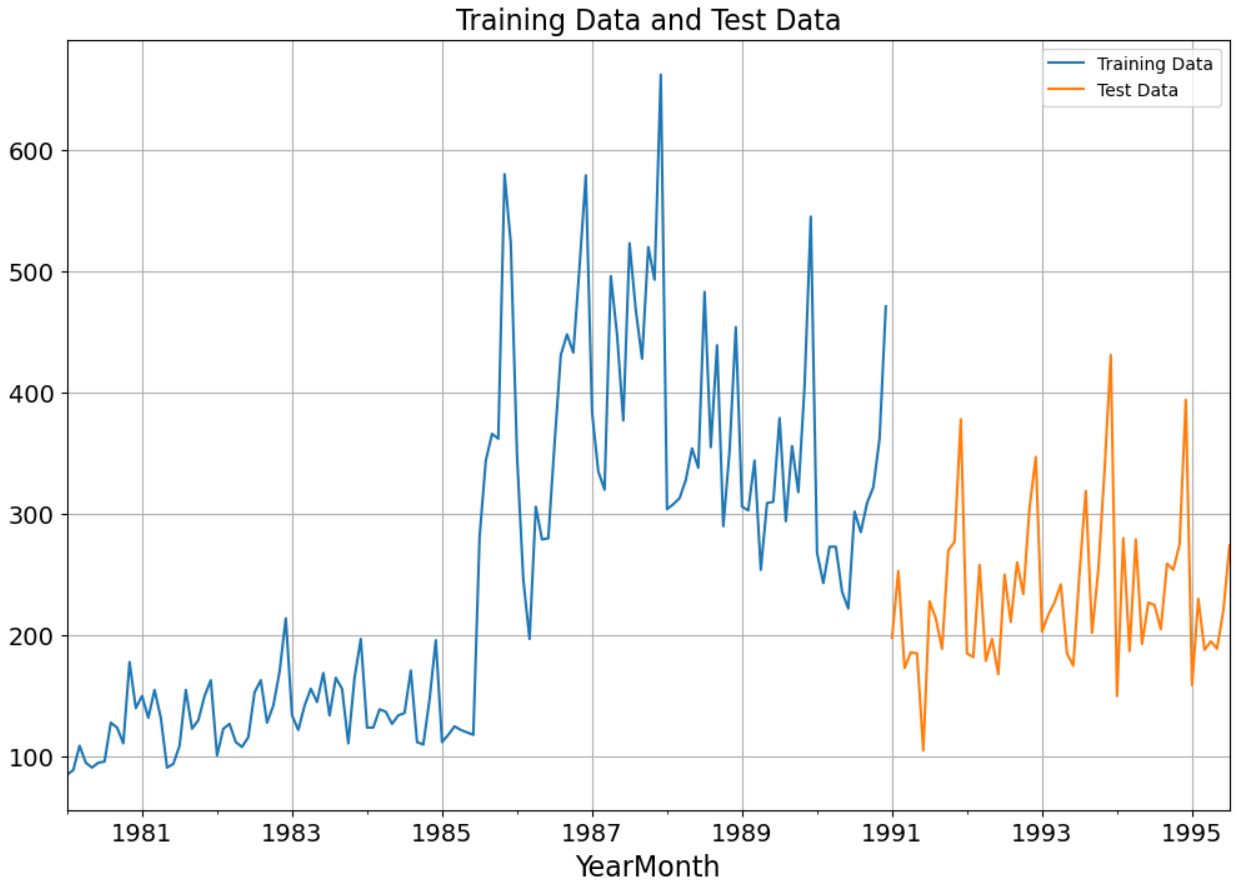
Table 6:First few rows of the test dataset

	Shoe-Sales	Year	Month
YearMonth			
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

The Train dataset contains 132 rows and 3 columns

The test dataset contains 55 rows and 3 columns

Fig 8:Training data and Test Data Graph



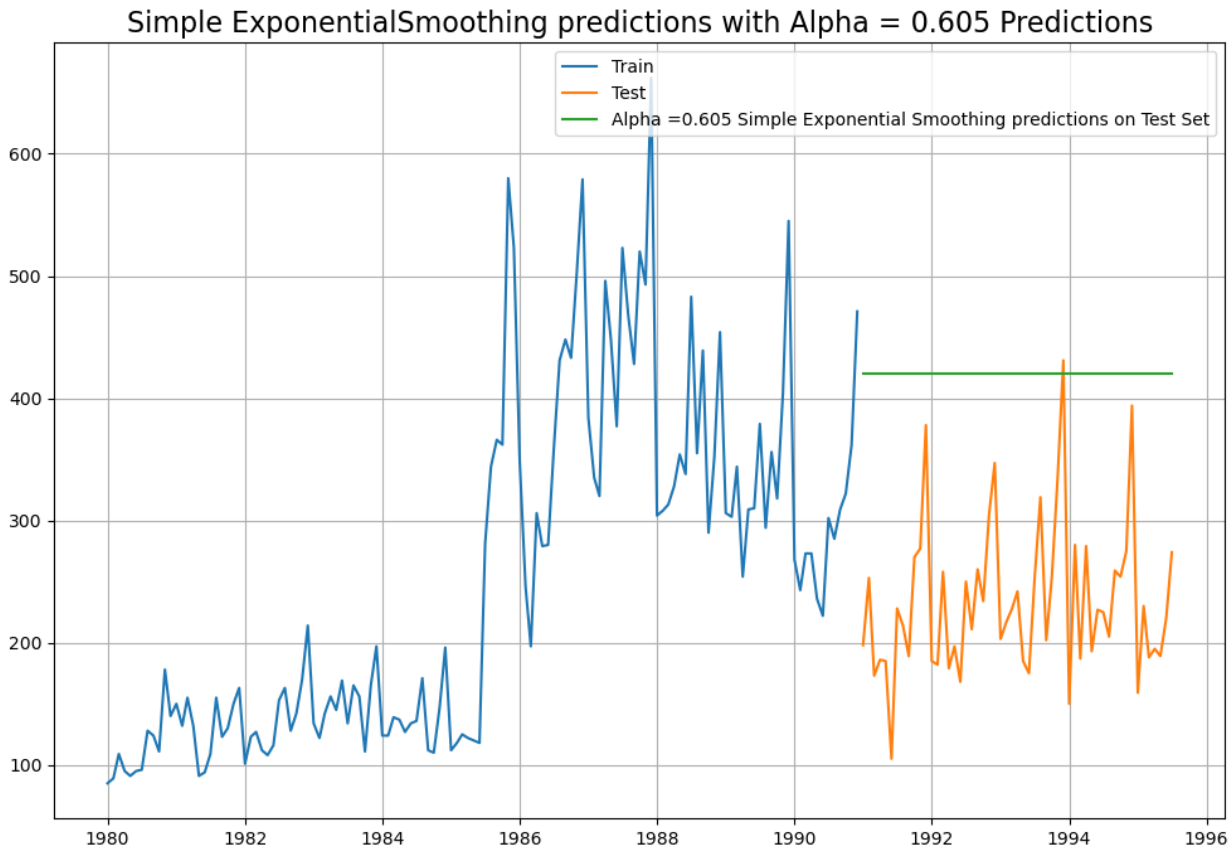
4. Build various exponential smoothing models on the training data and evaluate,the model using rmse on the test data. Other models such as regression,naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using rmse. (please do try to build as many models as possible and as many iterations of models as possible with different parameters.)

1)Simple Exponential Smoothing

Table 7:First few forecast after simple exponential smoothing

1991-01-01	420.229857
1991-02-01	420.229857
1991-03-01	420.229857
1991-04-01	420.229857
1991-05-01	420.229857

Fig 9:Simple Exponential Smoothing Prediction with Alpha= 0.605



RMSE

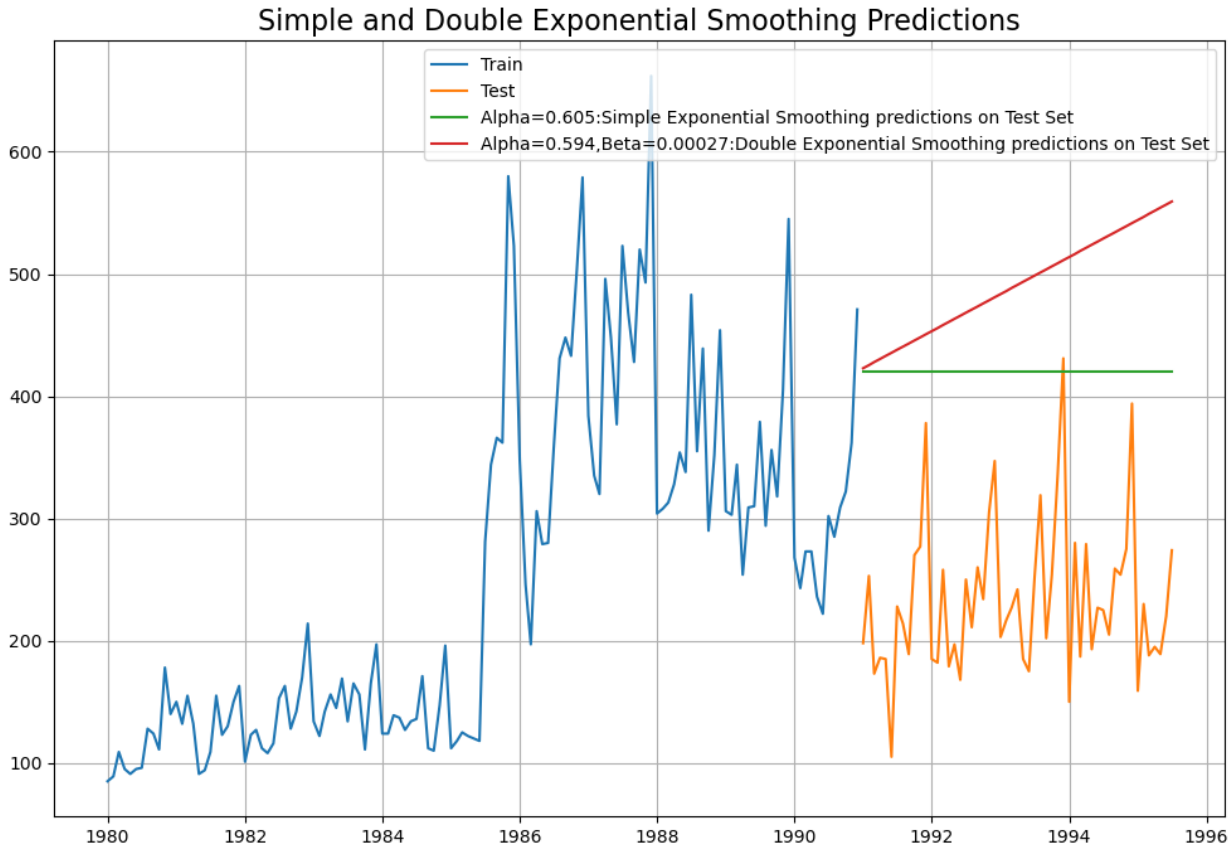
```
SES RMSE: 196.404836419672
SES RMSE (calculated using statsmodels): 196.404836419672
```

2) Double Exponential Smoothing

Table 8:First few rows of the dataset after prediction

```
1991-01-01    422.870987
1991-02-01    425.397576
1991-03-01    427.924166
1991-04-01    430.450755
1991-05-01    432.977344
```

Fig 10: Alpha=0.594,Beta=0.00027:Double Exponential Smoothing predictions on Test Set



RMSE:

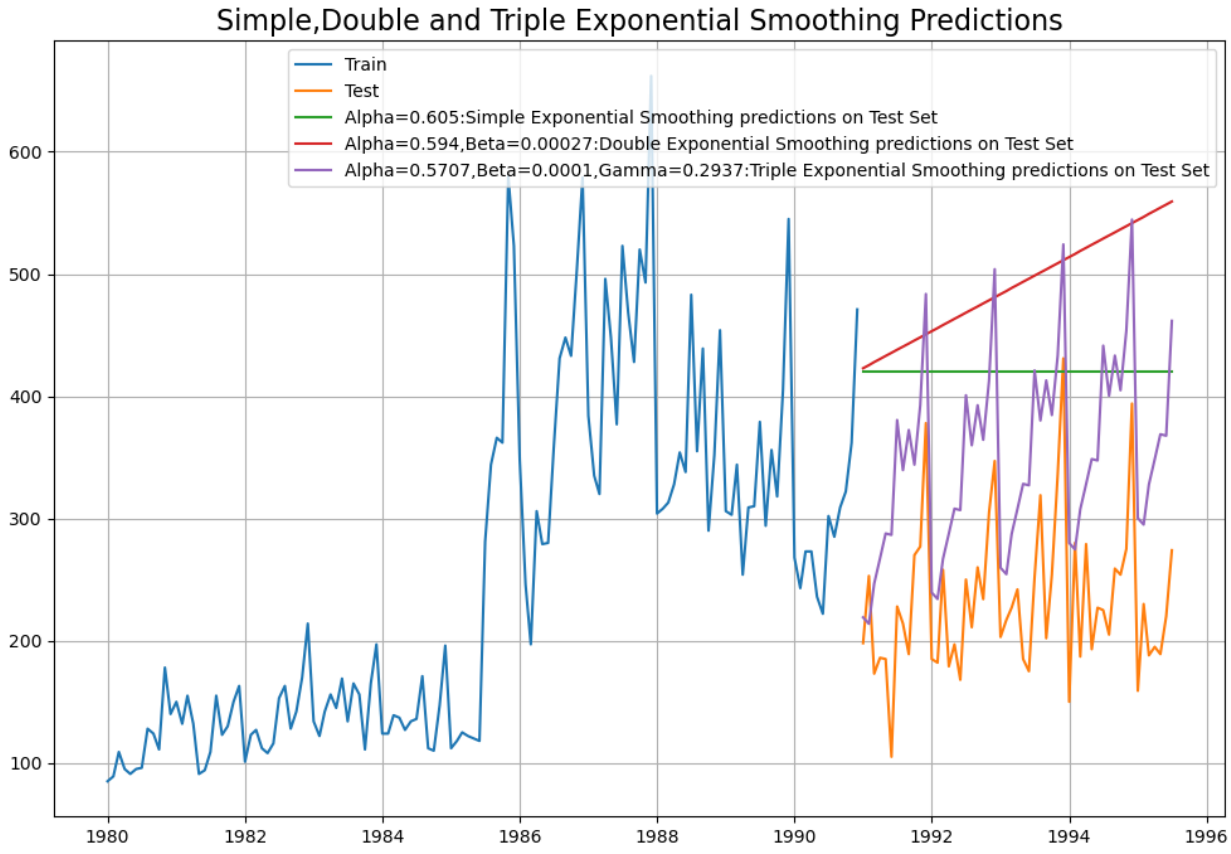
```
DES RMSE: 266.16120808183047
```

3)Triple Exponential Smoothing- Holt-Winters - ETS(A, M, M) - Holt Winter's linear method with additive error and seasonal

Table 9:First few rows of the TES Prediction

```
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
```

Fig 11: Alpha=0.5707,Beta=0.0001,Gamma=0.2937:Triple Exponential Smoothing predictions on Test Set



RMSE

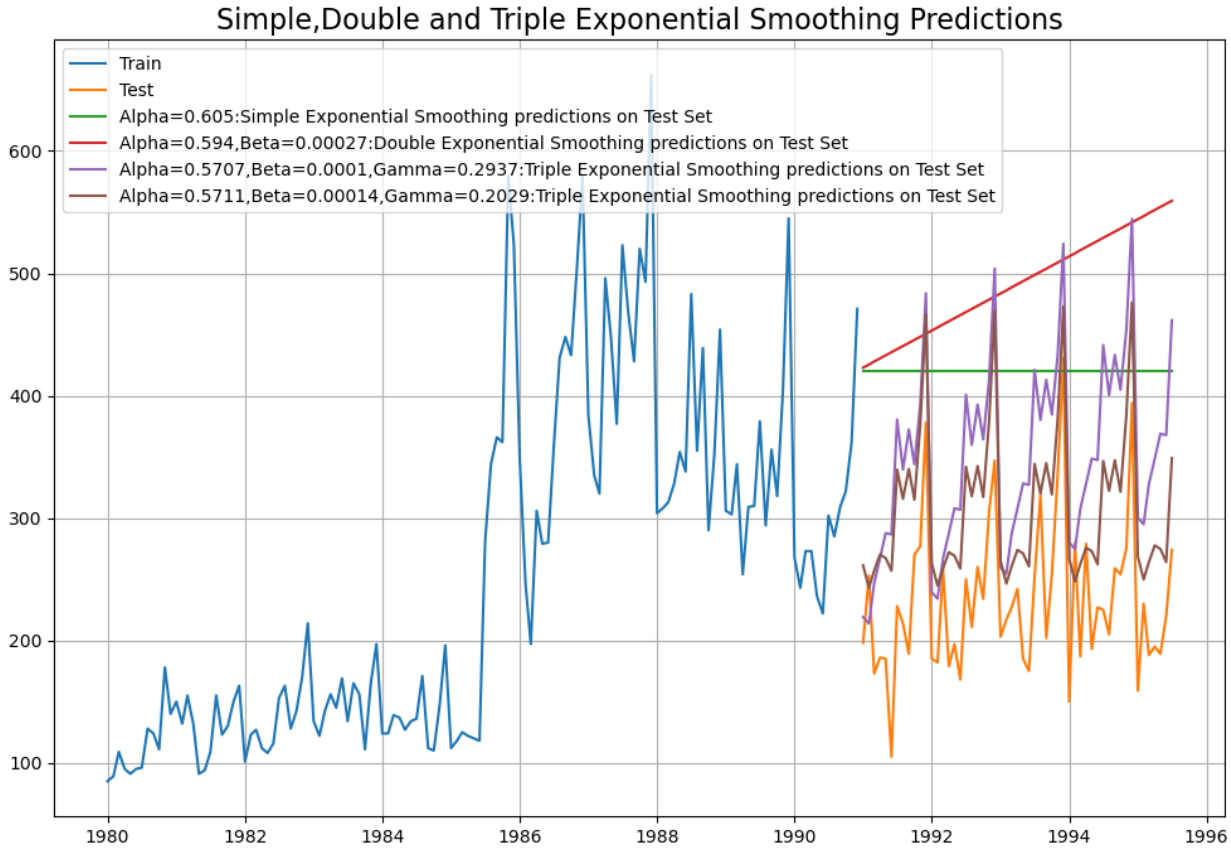
TES RMSE: 128.99252592312354

4)Triple Exponential Smoothing- Holt-Winters - ETS(A, M, M) - Holt Winter's linear method with multiplicative error and seasonal

Table 10:First few prediction

1991-01-01	261.342543
1991-02-01	243.085370
1991-03-01	256.996702
1991-04-01	270.198135
1991-05-01	267.375606

Fig 12: Alpha=0.5711,Beta=0.00014,Gamma=0.2029:Triple Exponential Smoothing predictions on Test Set



RMSE:

```
TES_am RMSE: 83.734048494837
```

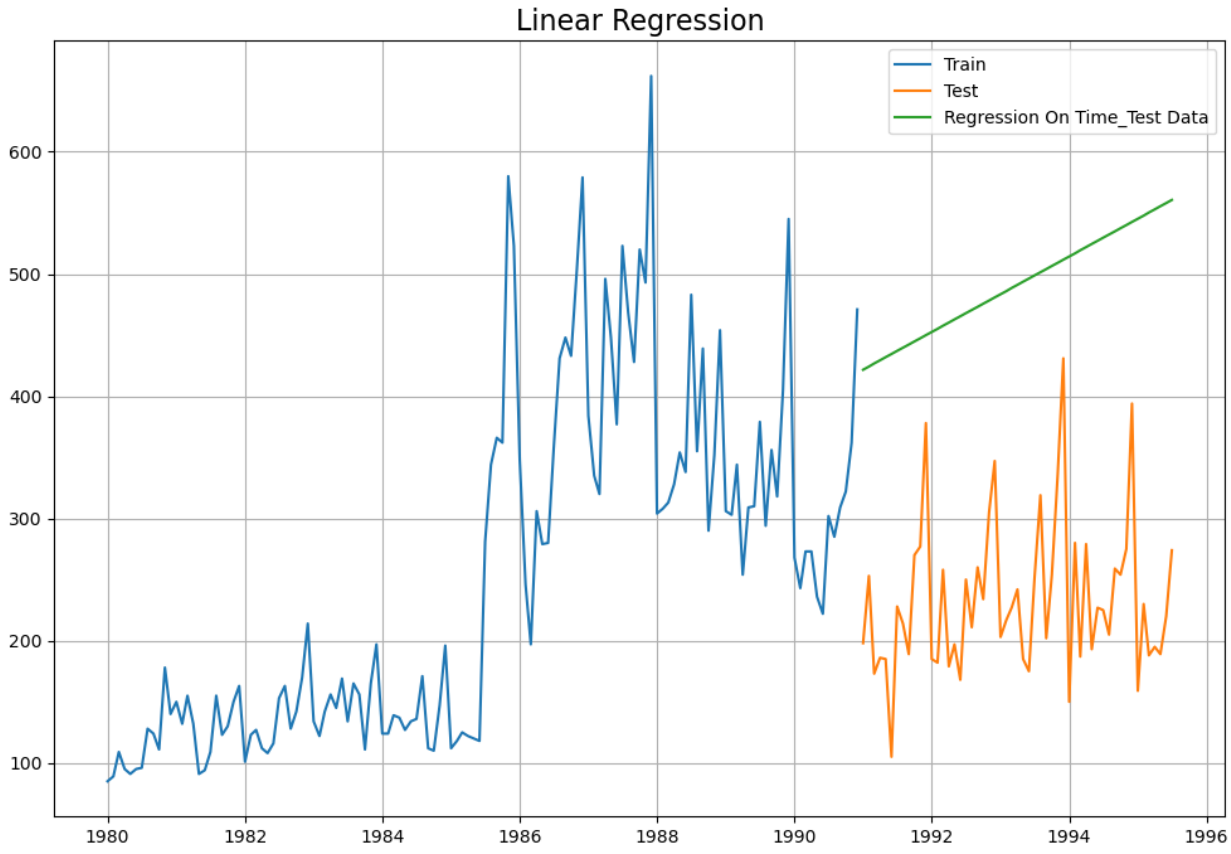
5)Linear Regression

For this particular linear regression, we are going to regress the 'Shoe Sales' variable against the order of the occurrence. For this we need to modify our training data before fitting it into a linear regression.

Table 11:First few predictions

YearMonth	Shoe-Sales	Year	Month	time
1991-01-01	198	1991	1	133
1991-02-01	253	1991	2	134
1991-03-01	173	1991	3	135
1991-04-01	186	1991	4	136
1991-05-01	185	1991	5	137

Fig 13:Linear Regression



RMSE:

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

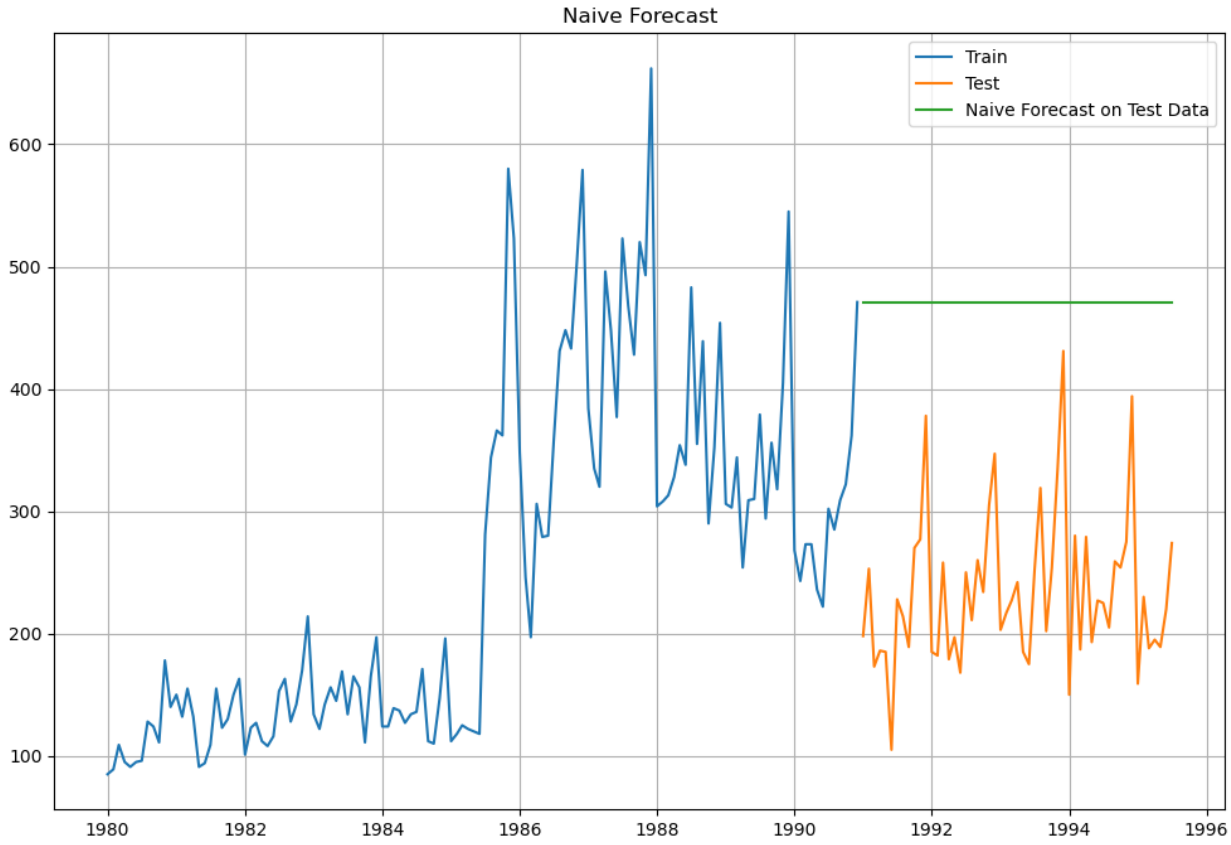
6)Naïve Forecasting

For this particular naive model, we say that the prediction for tomorrow is the same as today and the prediction for day after tomorrow is tomorrow and since the prediction of tomorrow is same as today,therefore the prediction for day after tomorrow is also today.

Table 12:Naïve forecast first few rows

```
YearMonth
1991-01-01    471
1991-02-01    471
1991-03-01    471
1991-04-01    471
1991-05-01    471
```

Fig 14:Naïve forecast



RMSE:

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

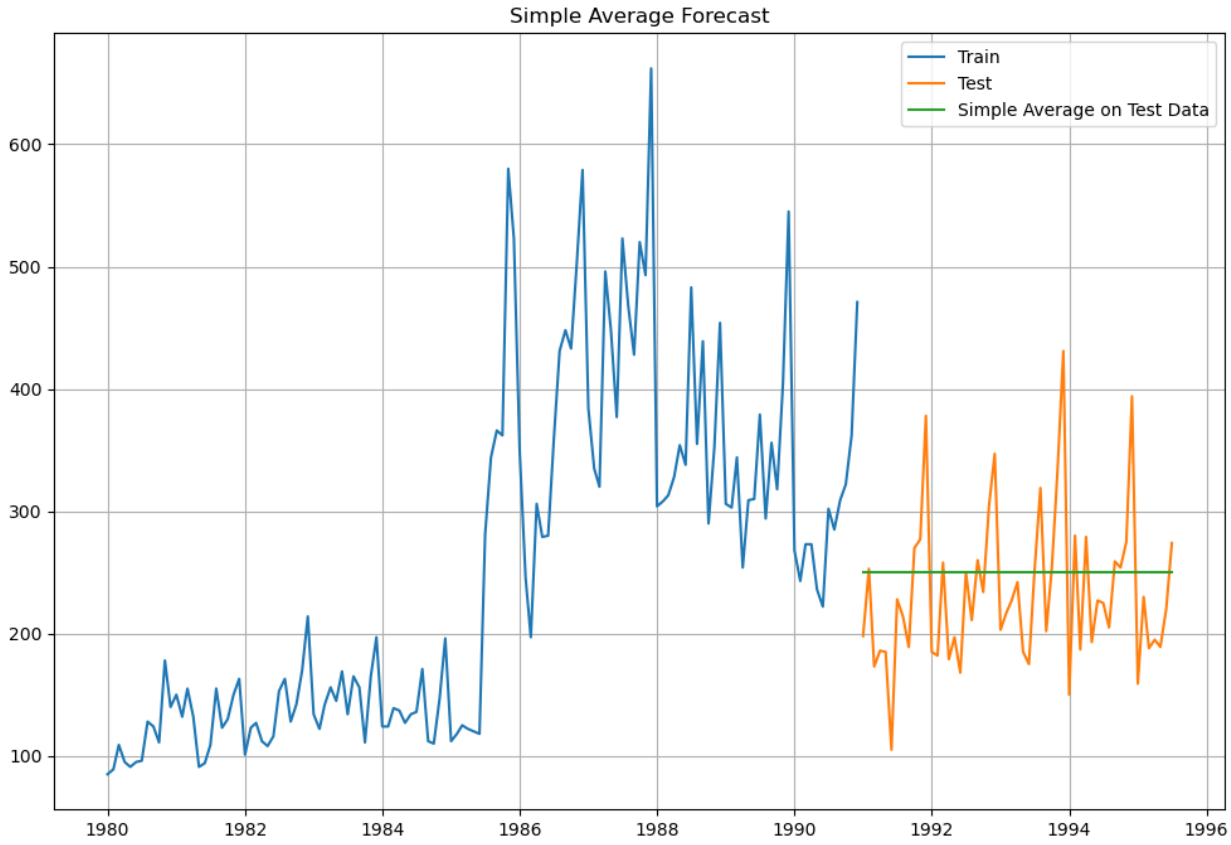
7)Simple Average

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

Table 13:First few rows of Simple Average

	Shoe-Sales	Year	Month	mean_forecast
YearMonth				
1991-01-01	198	1991	1	250.575758
1991-02-01	253	1991	2	250.575758
1991-03-01	173	1991	3	250.575758
1991-04-01	186	1991	4	250.575758
1991-05-01	185	1991	5	250.575758

Fig 15:Simple Average Prediction



RMSE:

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

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.

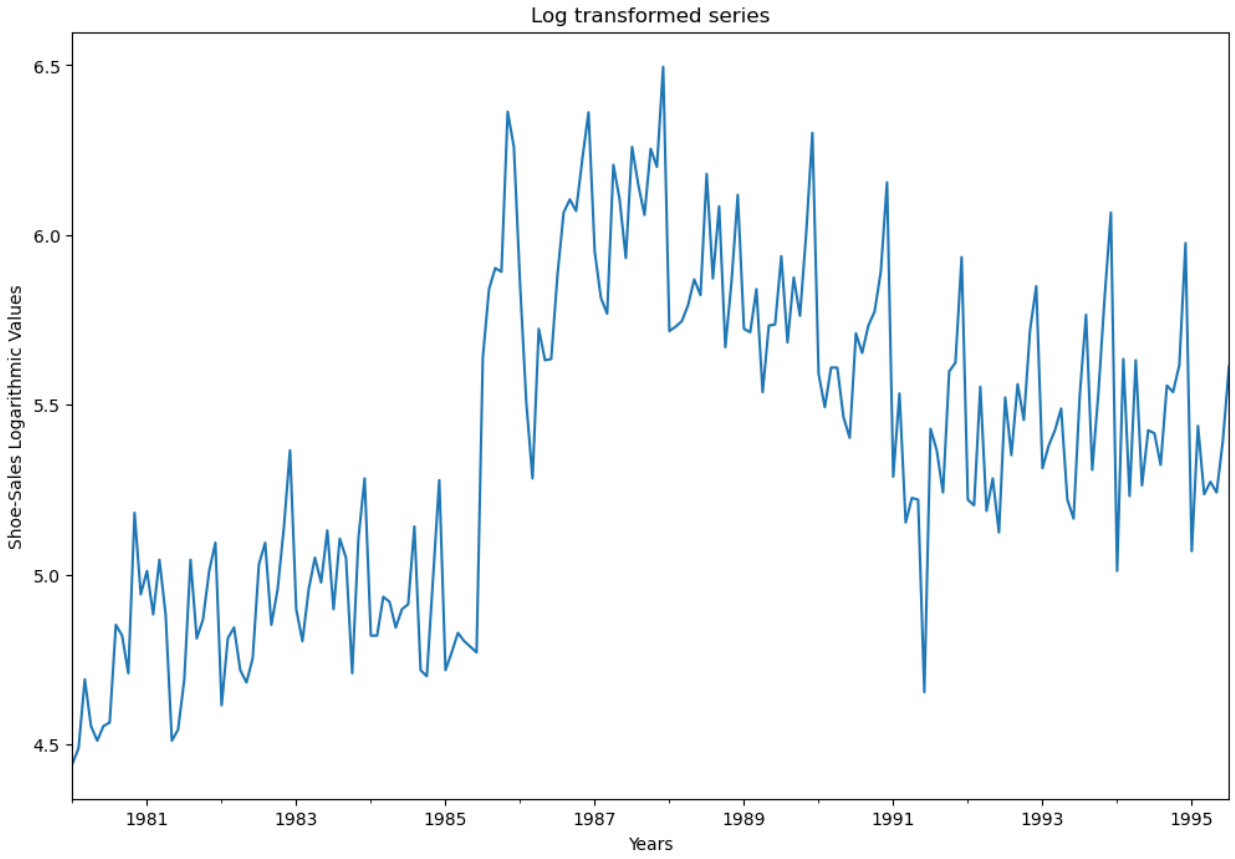
CHECKING FOR STATIONARITY

Checking for Stationarity using Augmented Dickey-Fuller.

```
DF test statistic is -1.717
DF test p-value is 0.4222
```

Here p value is more than 0.05 which means the test series is non stationary.

So logarithmic transformation is done to make the time series stationary.



8)Moving Average Model

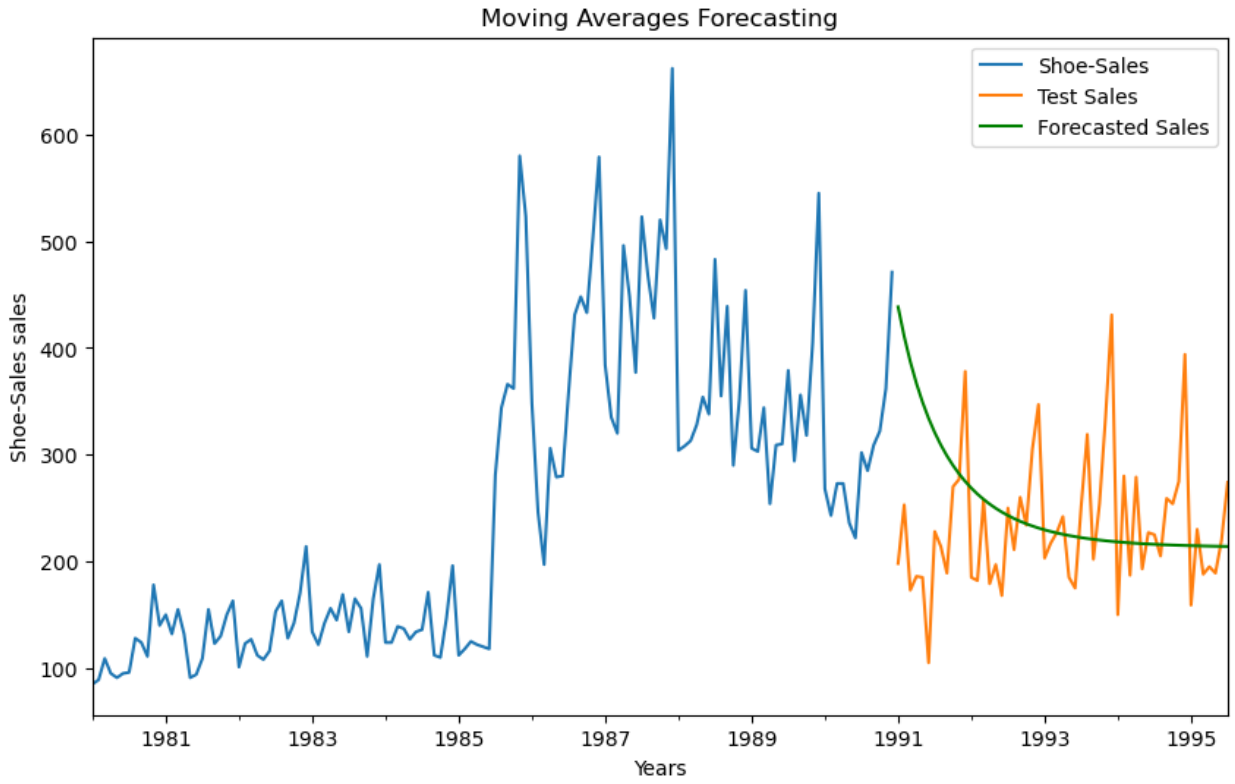
Table 14:Moving Average result

```
                               SARIMAX Results
==============================================================================
=
Dep. Variable:                 Shoe-Sales   No. Observations:
132
Model:                 ARIMA(1, 0, 0)   Log Likelihood
109.609
Date:                Wed, 29 Nov 2023   AIC                              -
213.217
Time:                        19:31:01   BIC                              -
204.569
Sample:                    01-01-1980   HQIC                             -
209.703
                         - 12-01-1990
Covariance Type:                 opg
==============================================================================
=
==============================================================================
=
                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-
```

const	2.3286	0.086	27.006	0.000	2.160 2.498
ar.L1	0.9097	0.037	24.880	0.000	0.838 0.981
sigma2	0.0110	0.001	8.837	0.000	0.009 0.013

RMSE: The Root Mean Squared Error of our forecasts is 92.14

Fig 17:



9)ARMA Forecast

Table 15:ARMA result

SARIMAX Results

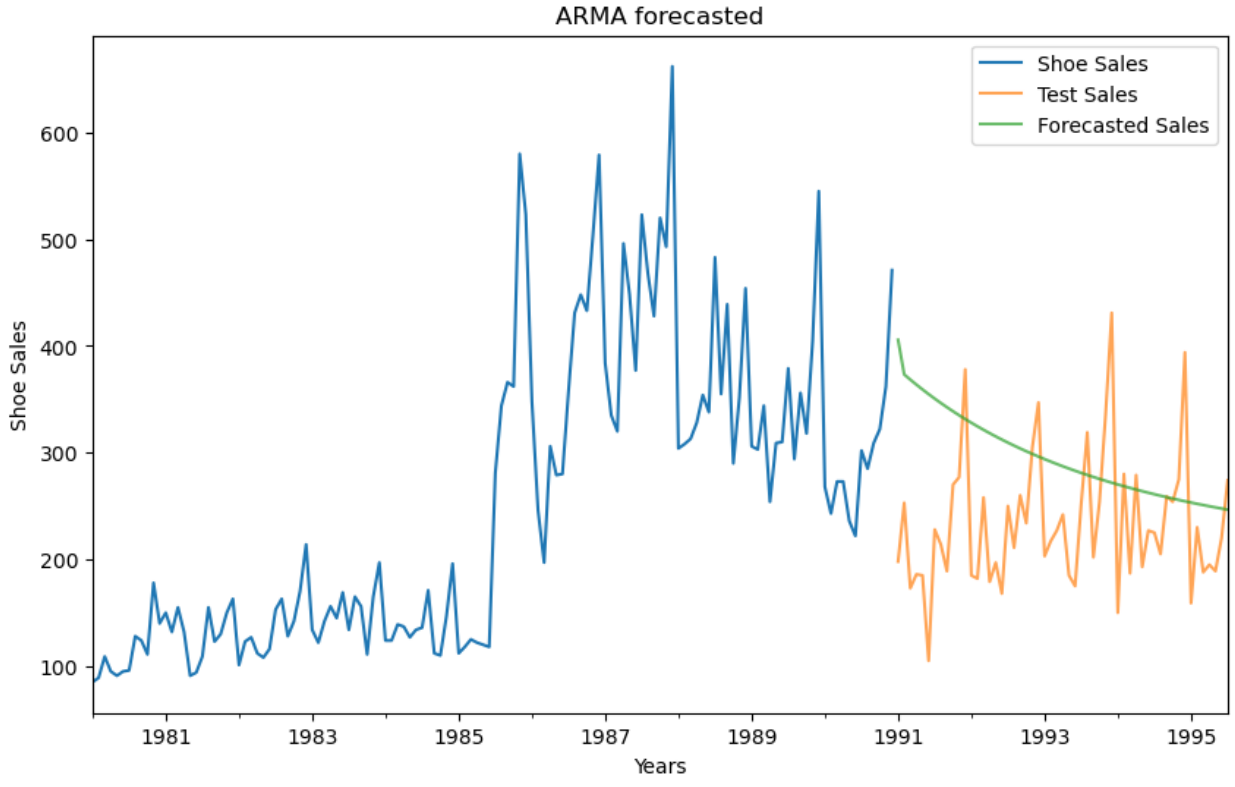
Dep. Variable:	Shoe-Sales	No. Observations:	132
Model:	ARIMA(1, 0, 2)	Log Likelihood	114.838
Date:	Wed, 29 Nov 2023	AIC	-219.677
Time:	19:31:09	BIC	-205.263

```
Sample:                     01-01-1980   HQIC                          -
213.819
                          - 12-01-1990
Covariance Type:                  opg
==============================================================================
=
==============================================================================
=
                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-
const          2.3101      0.149     15.504      0.000       2.018
2.602
ar.L1          0.9783      0.022     44.433      0.000       0.935
1.021
ma.L1         -0.2786      0.088     -3.173      0.002      -0.451      -
0.107
ma.L2         -0.1992      0.095     -2.106      0.035      -0.385      -
0.014
sigma2         0.0101      0.001      9.015      0.000       0.008
0.012
==============================================================================
=
```

RMSE:

```
The Root Mean Squared Error of our forecasts is 100.861
```

Fig 18:ARMA Forecast



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.

10)Automated ARIMA forecast

Table 16:Automated ARIMA result

```
                               SARIMAX Results
==========================================================================
=
Dep. Variable:             Shoe-Sales   No. Observations:
132
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -
743.244
Date:                Wed, 29 Nov 2023   AIC
1492.487
Time:                        19:31:12   BIC
1501.113
Sample:                    01-01-1980   HQIC
1495.992
                         - 12-01-1990
Covariance Type:                  opg
==========================================================================
=
                 coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------
-
ar.L1          0.4699      0.111      4.235      0.000       0.252
0.687
ma.L1         -0.8347      0.068    -12.261      0.000      -0.968      -
0.701
sigma2      4944.0868    405.583     12.190      0.000    4149.158
5739.015
==========================================================================
======
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):
57.30
Prob(Q):                              0.83   Prob(JB):
0.00
Heteroskedasticity (H):              12.81   Skew:
0.01
Prob(H) (two-sided):                  0.00   Kurtosis:
6.24
==========================================================================
======

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

Table 17:Predicted result first few dataset(automated ARIMA)

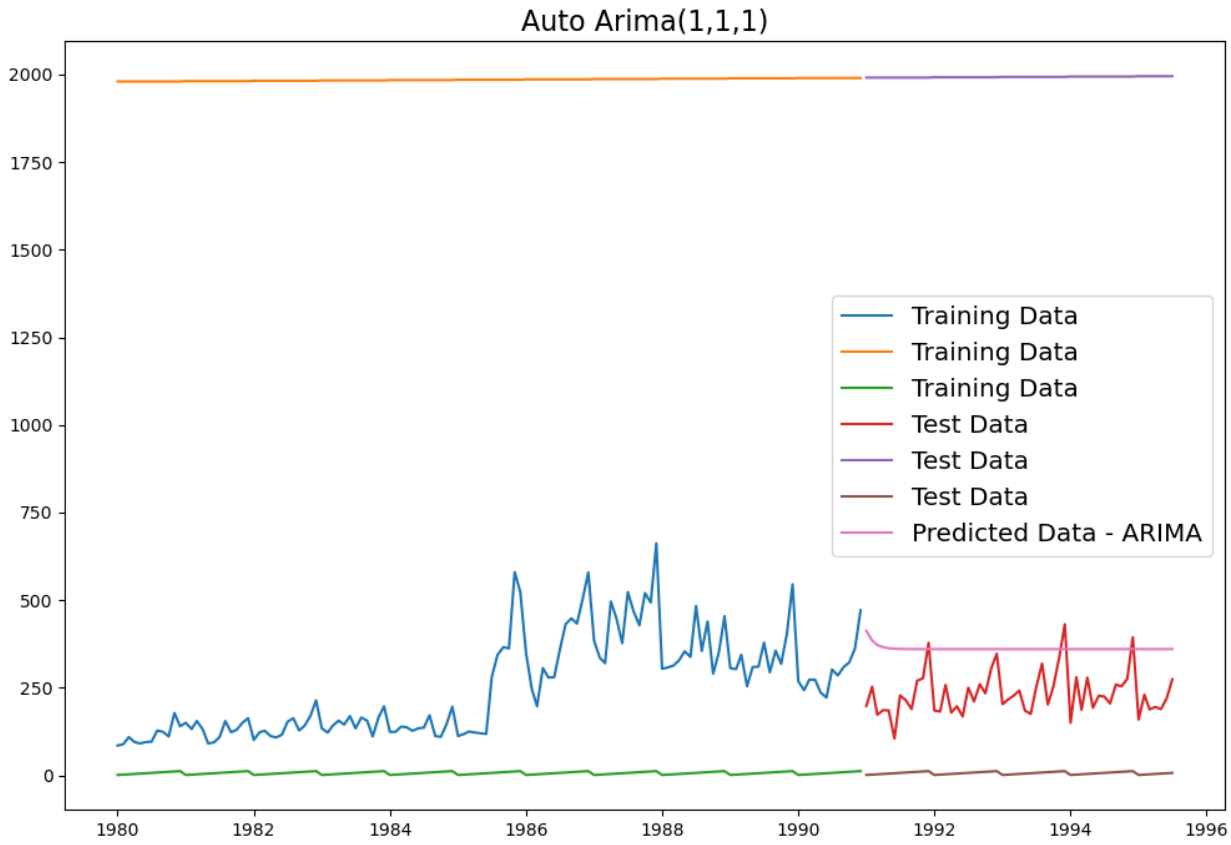
```
1991-01-01    412.252023
```

```
1991-02-01    384.645964
1991-03-01    371.673697
1991-04-01    365.577944
1991-05-01    362.713509
```

RMSE:

```
RMSE for the autofit ARIMA model: 142.82073039239808
MAPE for the autofit ARIMA model: 66.27418450722845
```

Fig 19:Automated ARIMA forecast



11)Automated SARIMA

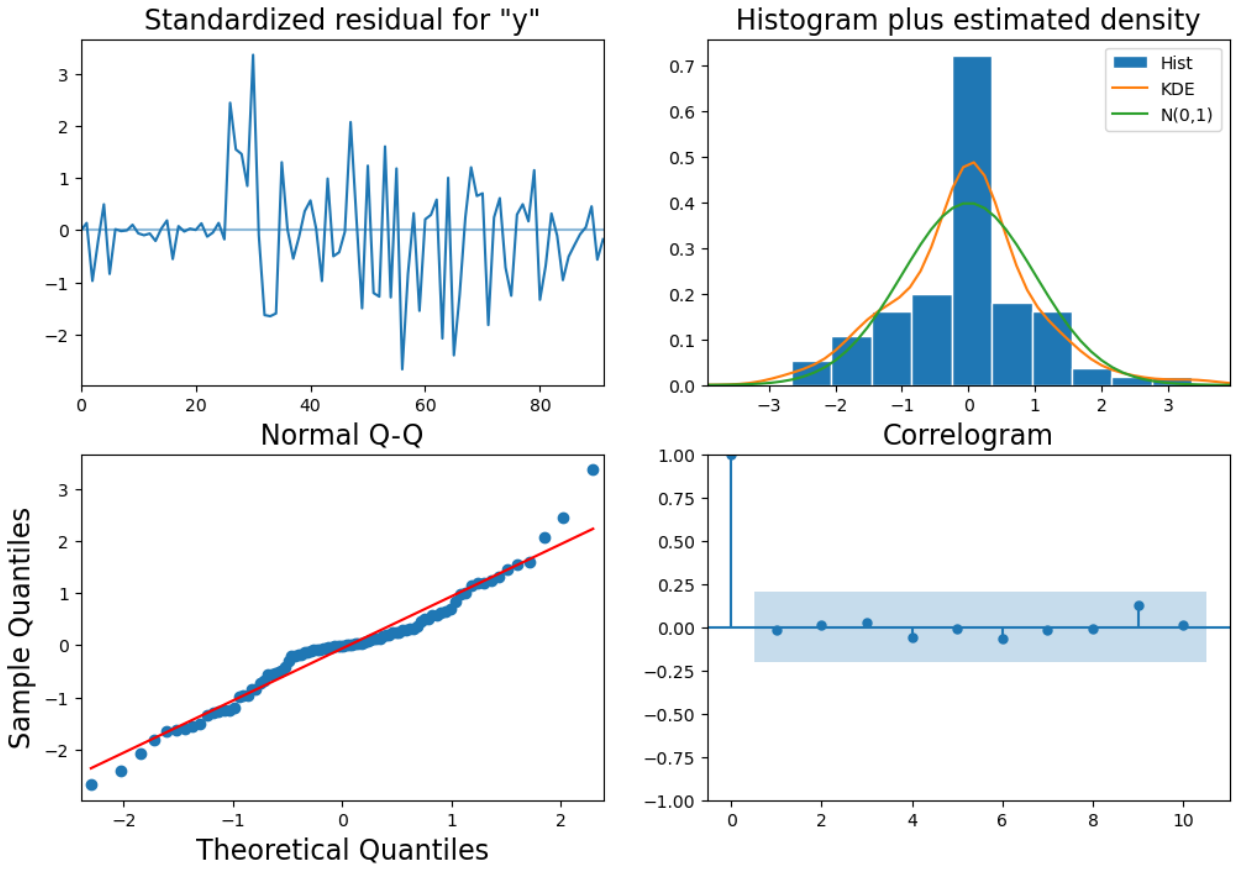
Table 18-Automated SARIMA result

```
                                     SARIMAX Results
==========================================================================================
==================
Dep. Variable:                                   y   No. Observations:
132
Model:             SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12)   Log Likelihood
-507.955
Date:                                 Wed, 29 Nov 2023   AIC
1035.910
Time:                                         19:41:08   BIC
1061.128
```

```
Sample:                                           0   HQIC
1046.088
                                                - 132
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      2919.7516      0.000    1.4e+07      0.000    2919.751
2919.752
==============================================================================
======
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
6.18e+23. Standard errors may be unstable.
```

Fig 20:Automated SARIMA result graph



RMSE:

```
RMSE automated SARIMA is 90.40430459490653
```

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.

	RMSE
Alpha=0.605,SES	196.404836
Alpha=0.594,Beta=0.00027,DES	266.161208
Alpha=0.5707,Beta=0.0001,Gamma=0.2937,TES(additive error)	128.992526
Alpha=0.5711,Beta=0.00014,Gamma=0.2029:TES(multiplicative error)	83.734048
RegressionOnTime	266.276472
Naive Model	245.121306
SimpleAverageModel	63.984570
Best MA Model : AR(1,0,0)	92.140047
(1,1,3),(3,0,3,12),Auto_SARIMA	90.404305

From this consolidated result We can see that Simple Average model has the lowest RMSE value.It is the best model to predict for shoe sales .

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.

Table 19:Prediction using Simple Average Model

	Shoe-Sales	mean_forecast
1994-08-01	NaN	250.575758
1994-09-01	NaN	250.575758
1994-10-01	NaN	250.575758
1994-11-01	NaN	250.575758
1994-12-01	NaN	250.575758
1995-01-01	NaN	250.575758
1995-02-01	NaN	250.575758
1995-03-01	NaN	250.575758
1995-04-01	NaN	250.575758
1995-05-01	NaN	250.575758
1995-06-01	NaN	250.575758
1995-07-01	NaN	250.575758

RECOMMENDATIONS

When the sales is lower then it is off season and When the sales is higher then it is season

Discount should be added in the off season to increase the sales.

More and more new models should be introduced during the season of sales.