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			

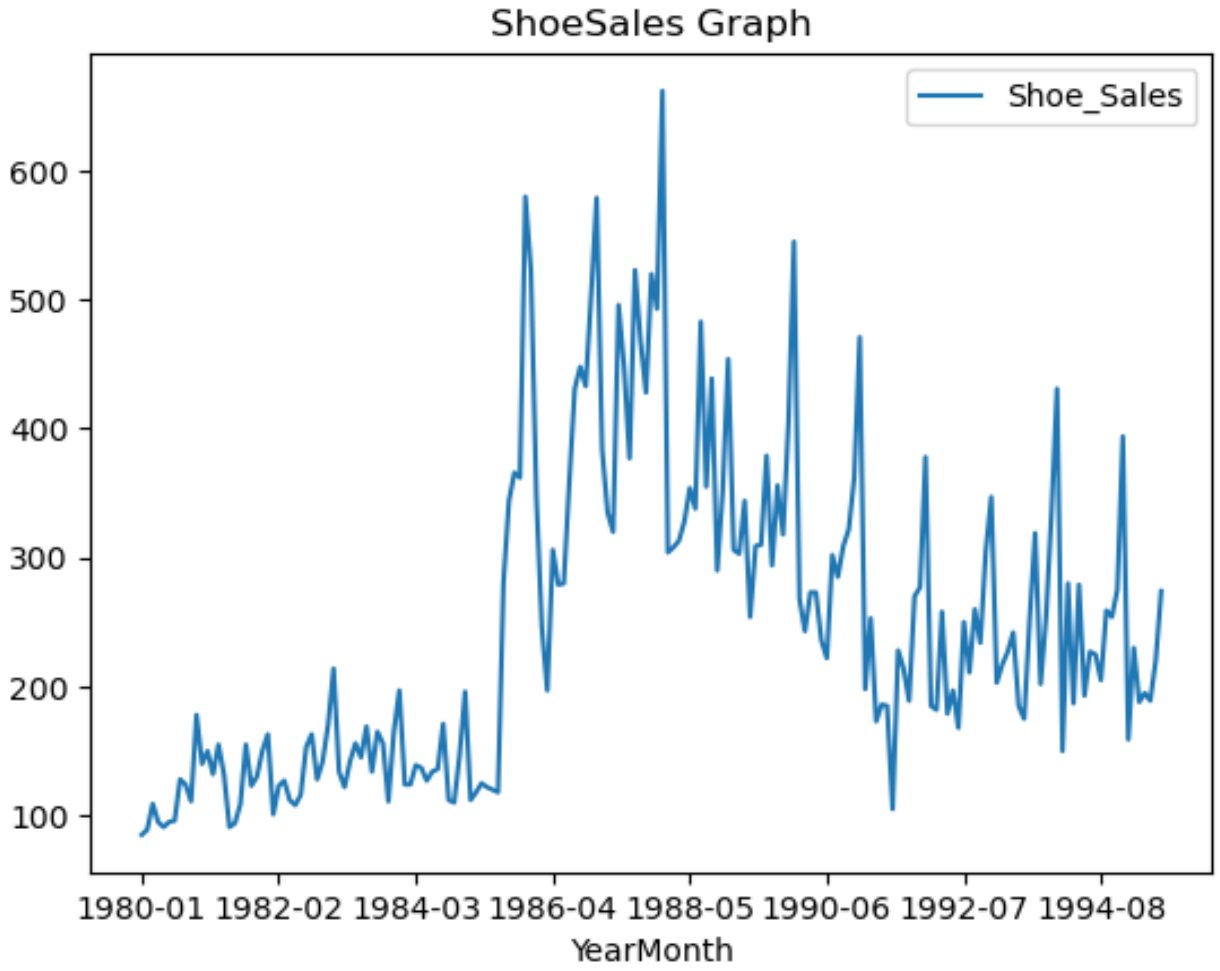
1980-01-01	85	1980	1
1980-02-01	89	1980	2
1980-03-01	109	1980	3
1980-04-01	95	1980	4
1980-05-01	91	1980	5

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

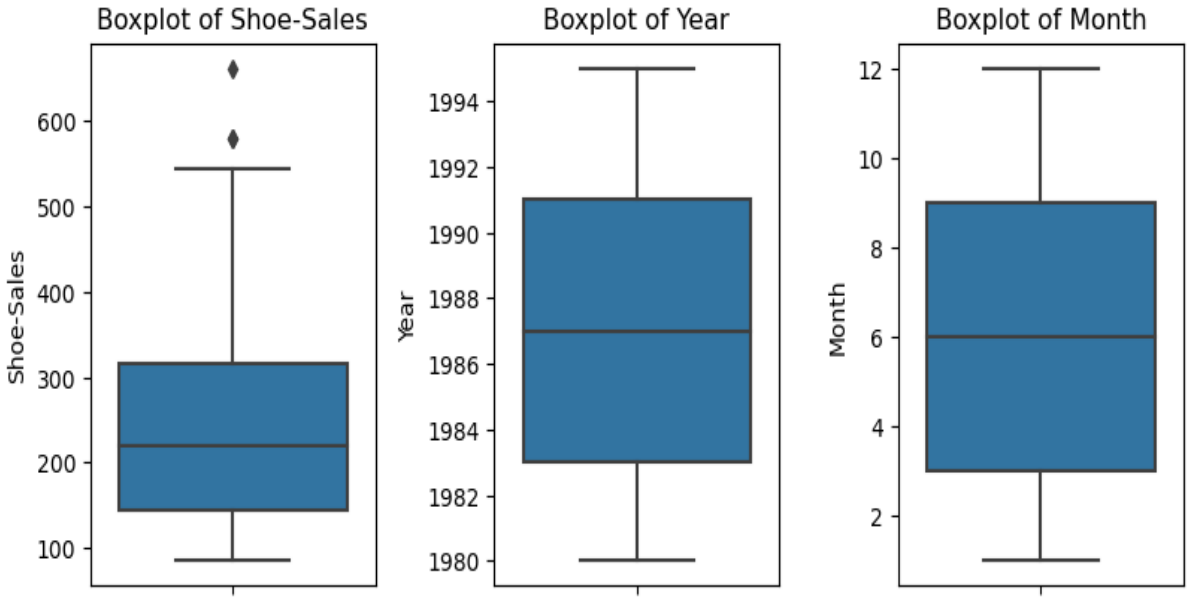
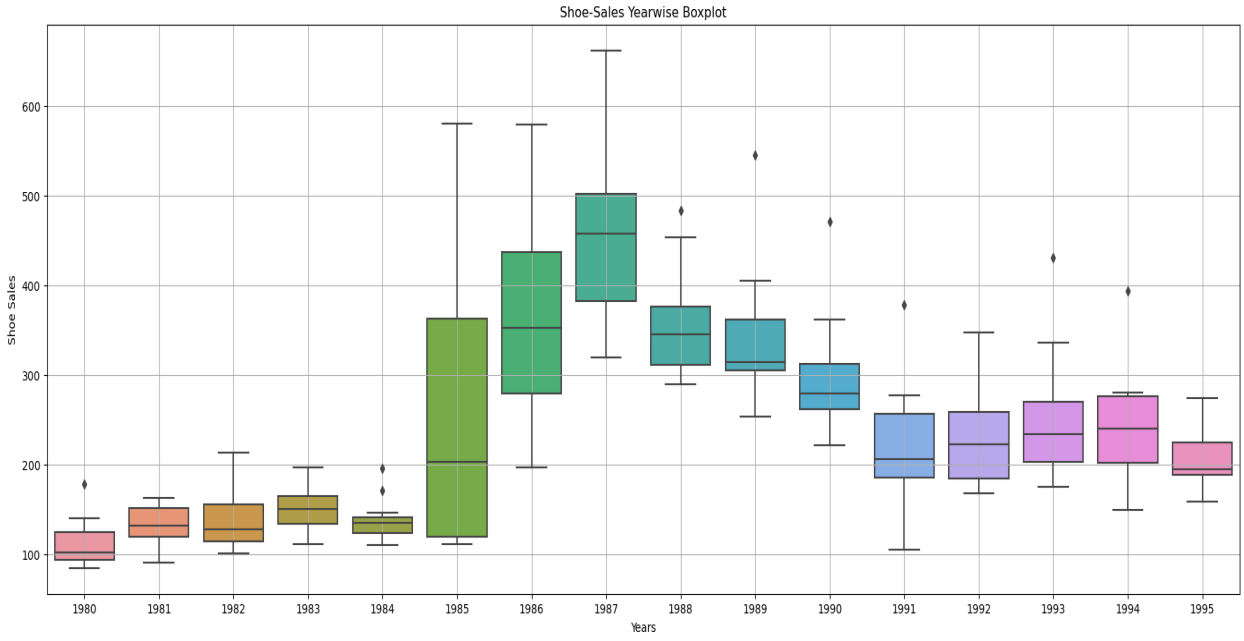
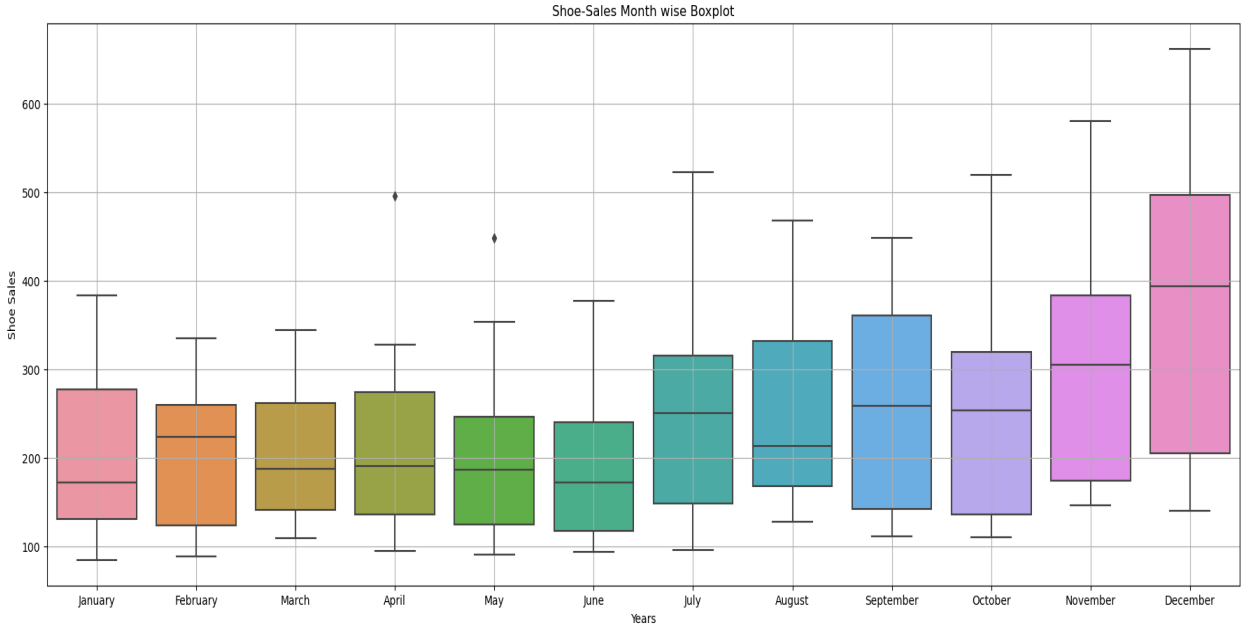


Fig 3:Yearwise boxplot of shoe sales



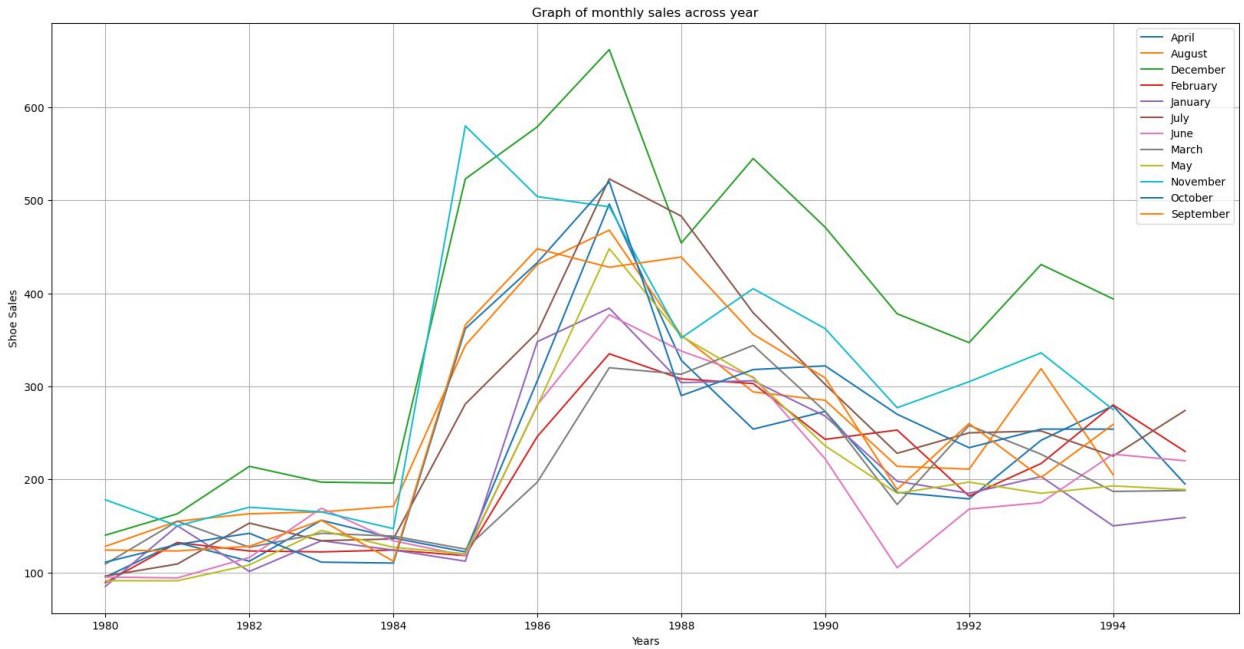
The sales increased in the year 1985,1986,1987.

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

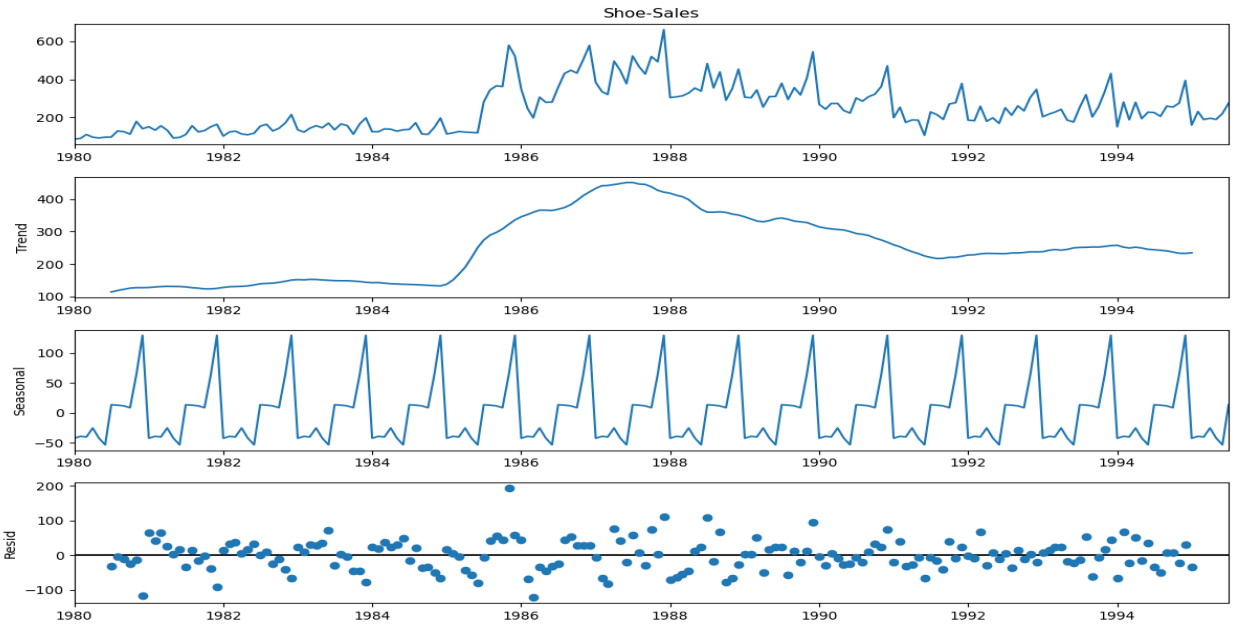
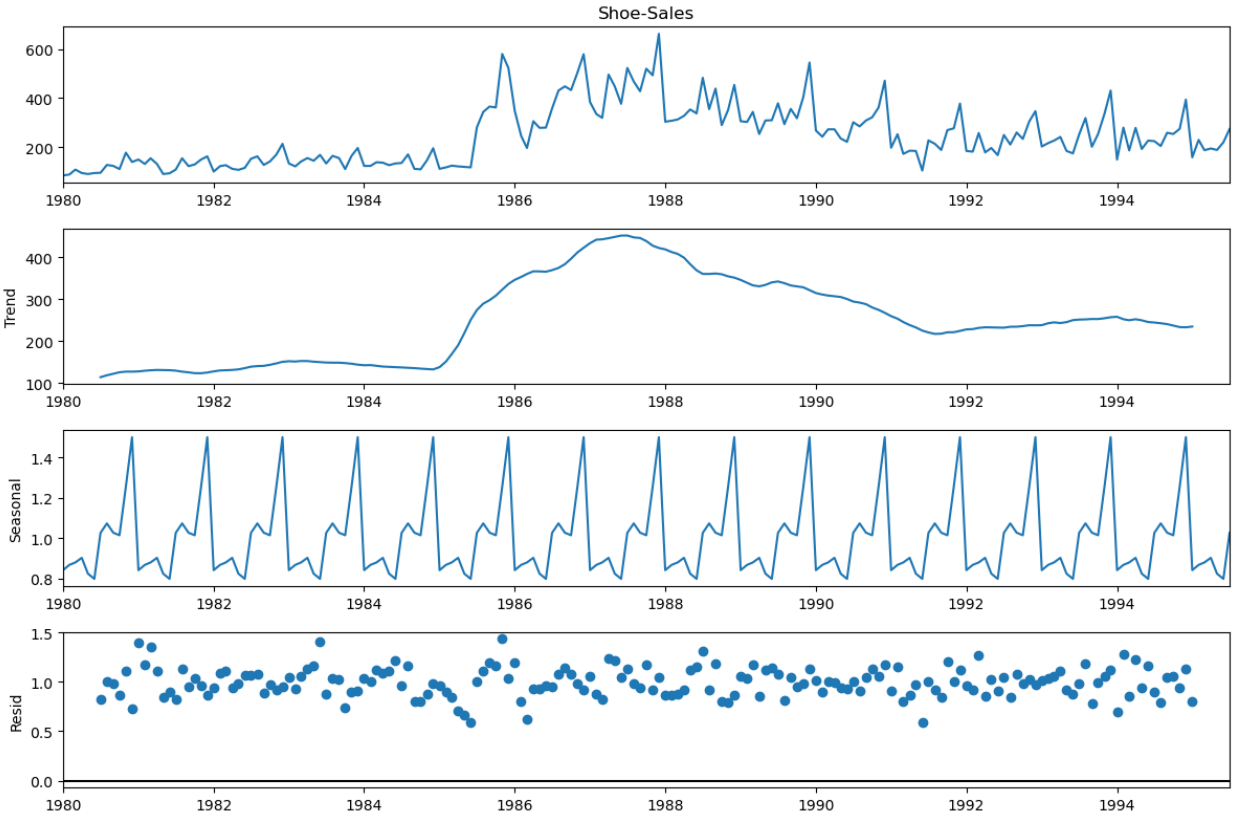


Fig 7:Multiplicative decomposition of shoe sales dataset



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

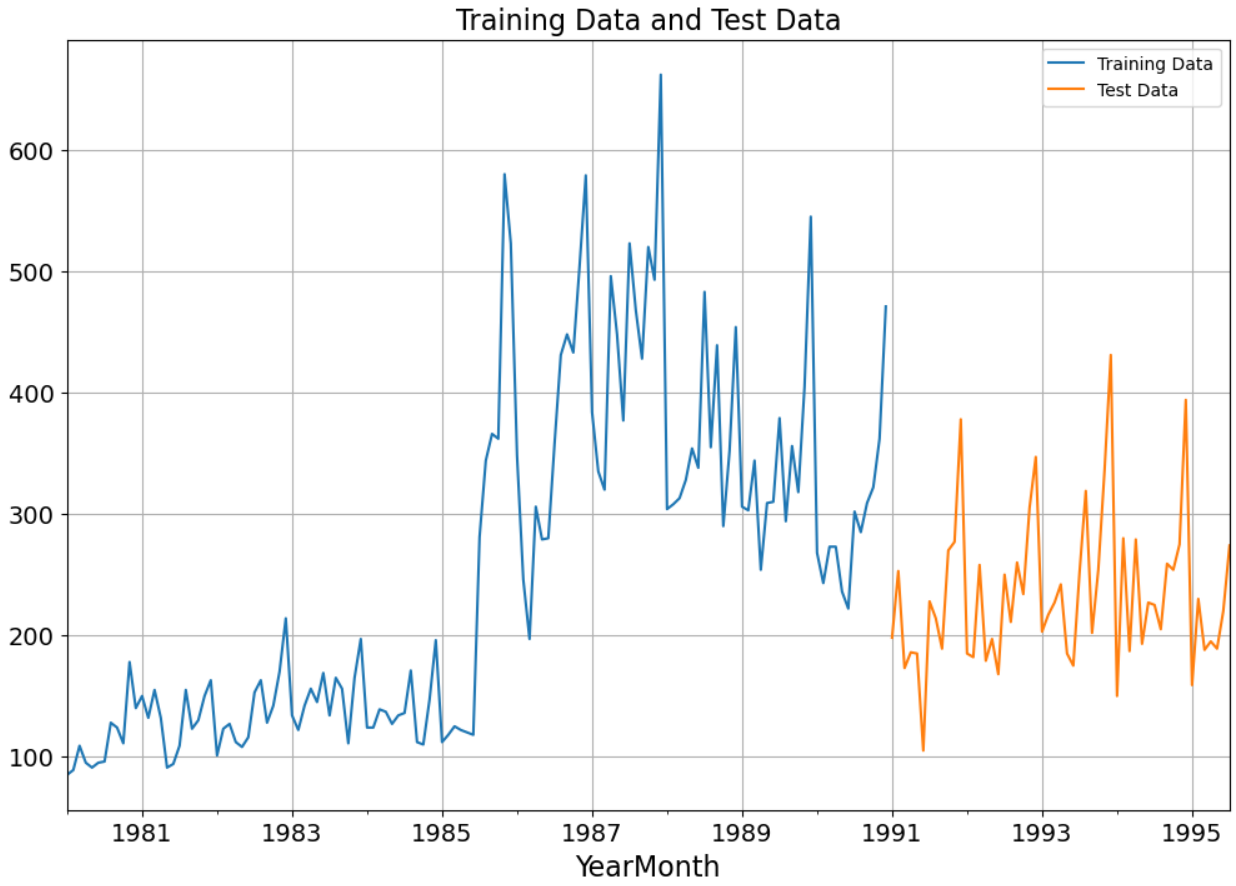
Table 6:First few rows of the test dataset

	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



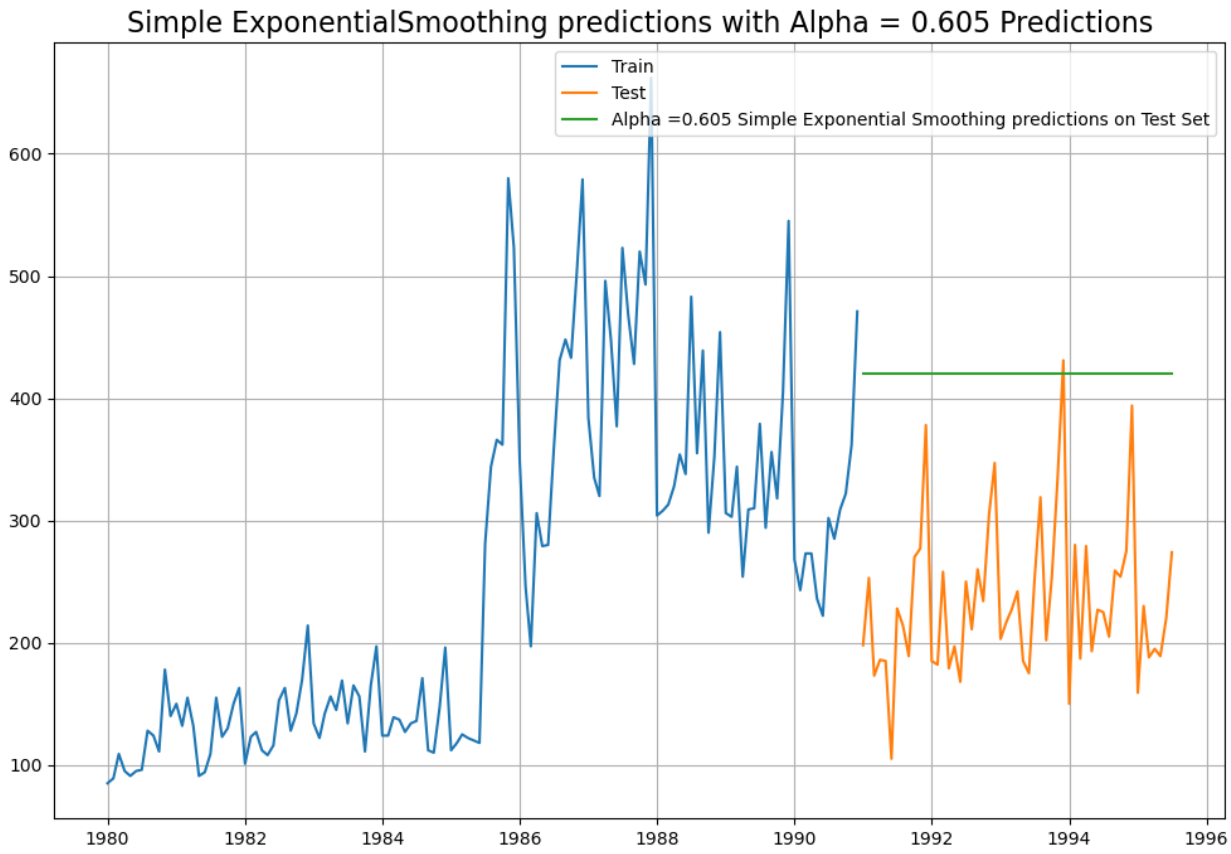
4. Build various exponential smoothing models on the training data and evaluate,the model using RMSE on the test data. Other models such as regression,naïve forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE. (Please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

132	420.229967
133	420.229967
134	420.229967
135	420.229967
136	420.229967

Fig 9:Simple Exponential Smoothing Prediction with Alpha= 0.605



MAPE:

SES MAPE: 91.42076016336813

RMSE

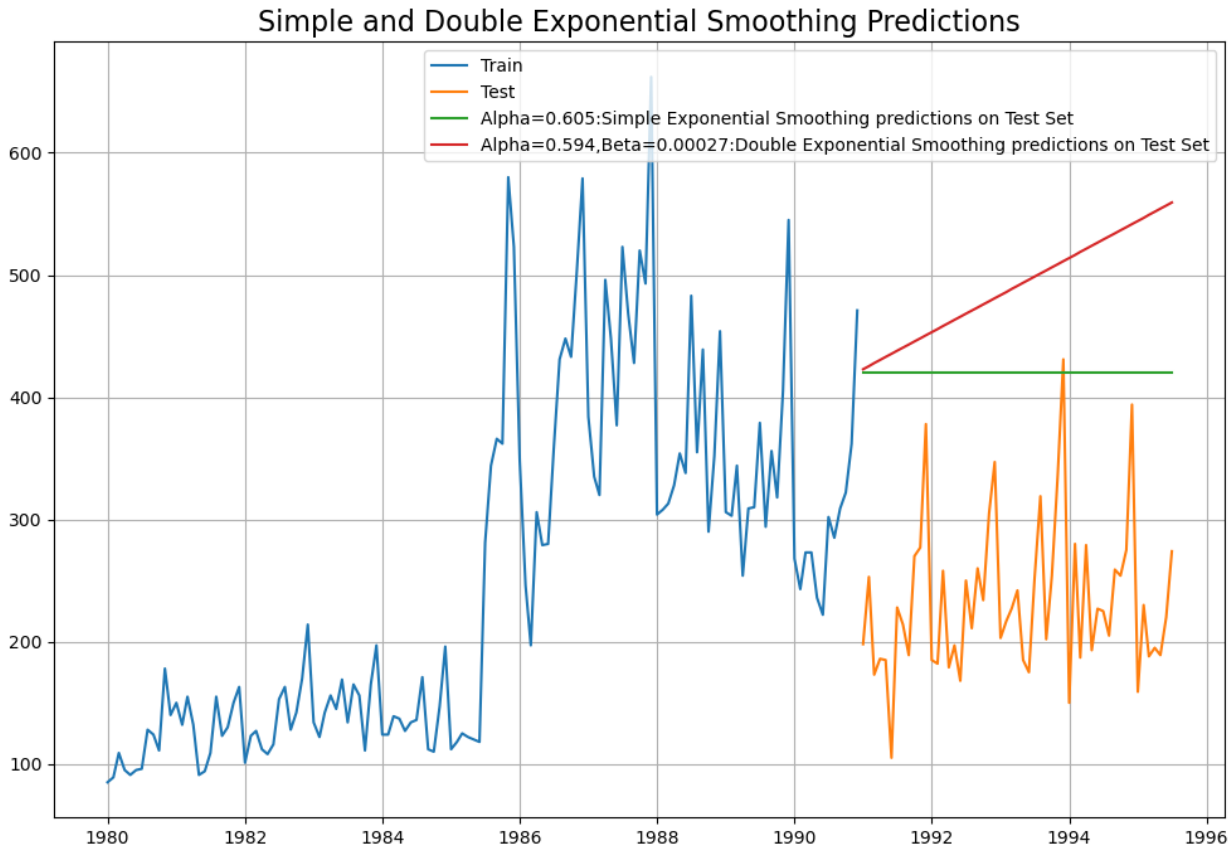
SES RMSE(calculated using sklearn): 195.89459510493793
SES RMSE (calculated using statsmodels): 195.89459510493793

2) Double Exponential Smoothing

Table 8:First few rows of the dataset after prediction

132	422.908695
133	425.448566
134	427.988438
135	430.528309
136	433.068181

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



MAPE:

```
DES MAPE: 122.32528367957886
```

RMSE:

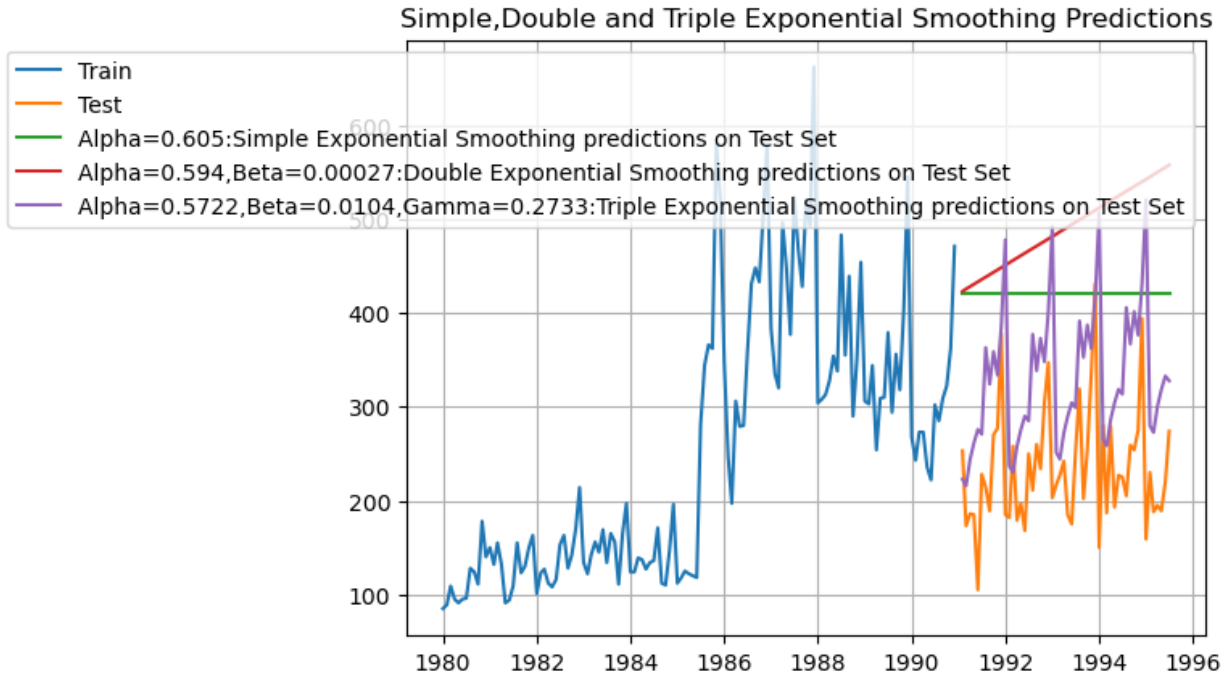
```
DES RMSE(calculated using sklearn): 264.82482731601925
DES RMSE (calculated using statsmodels): 264.82482731601925
```

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

```
132    222.794048
133    215.688235
134    243.726457
135    261.300316
136    275.782177
```

Fig 11: Alpha=0.5722,Beta=0.0104,Gamma=0.2733:Triple Exponential Smoothing predictions on Test Set



MAPE:

```
TES MAPE: 49.90672694891035
```

RMSE

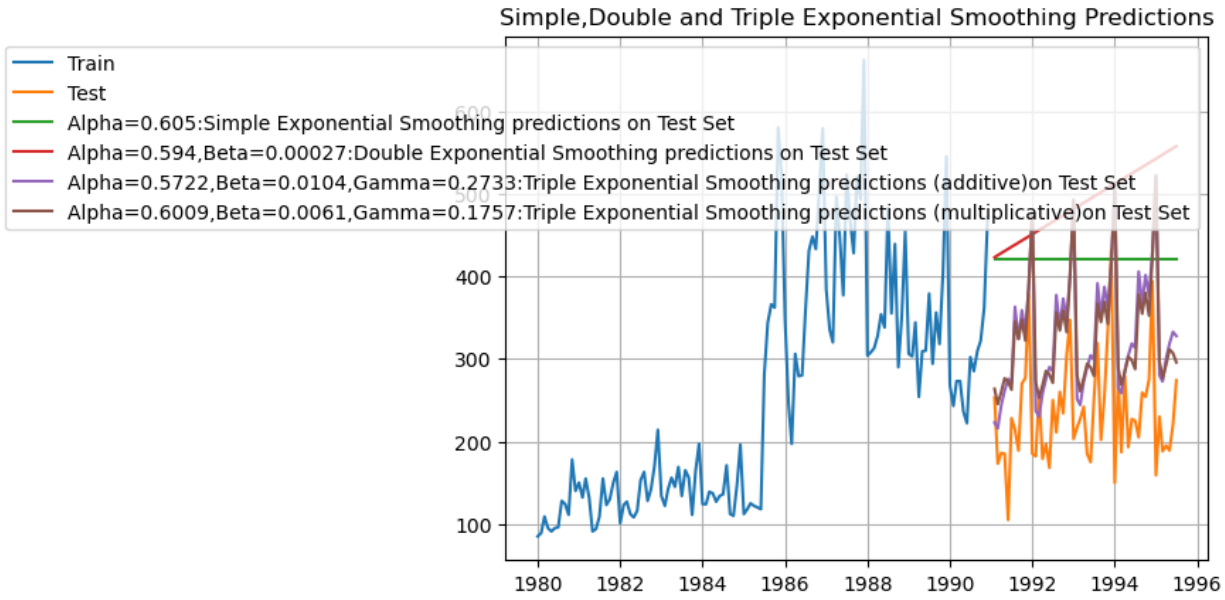
```
TES additive RMSE(calculated using sklearn): 126.38763305091122
TES additive RMSE (calculated using statsmodels): 126.38763305091122
```

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

```
132    263.531912
133    245.139749
134    259.792279
135    276.337606
136    272.442200
```

Fig 12: Alpha=0.6009,Beta=0.0061,Gamma=0.1757:Triple Exponential Smoothing predictions on Test Set



MAPE:

```
TES MAPE: 48.032997724305226
```

RMSE:

```
TES multiplicative RMSE(calculated using sklearn): 122.17854463685187
TES multiplicative RMSE (calculated using statsmodels): 122.17854463685187
```

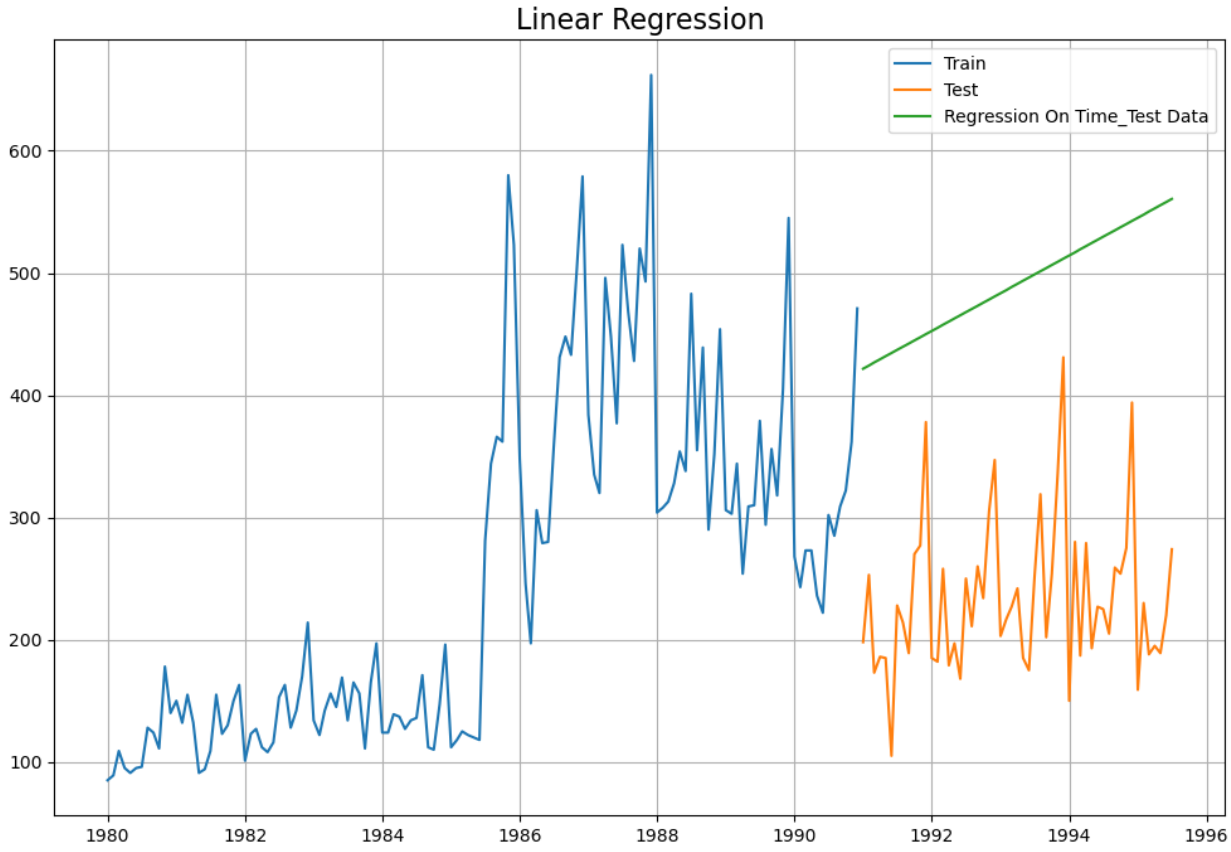
5)Linear Regression

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



MAPE:

```
LR MAPE: 122.14075782294853
```

RMSE:

```
LR RMSE(calculated using sklearn): 264.51679449469304
LR RMSE (calculated using statsmodels): 264.51679449469304
```

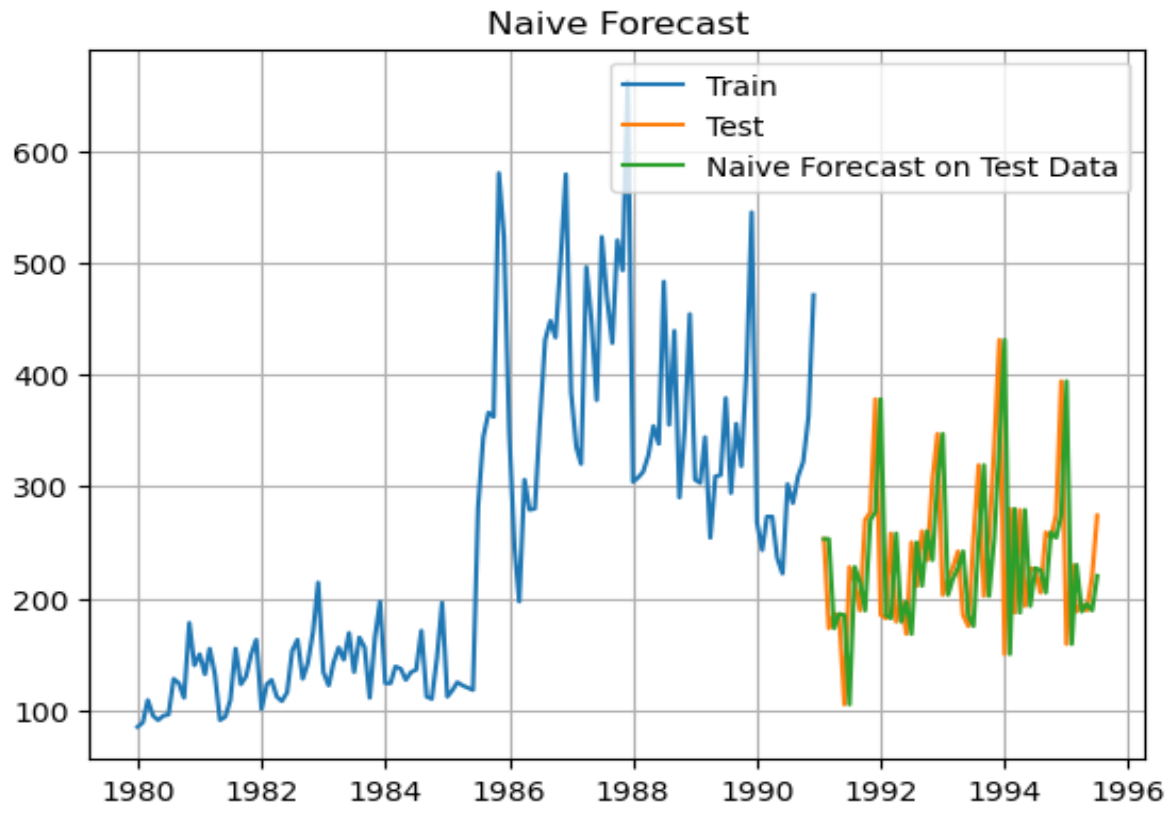
6)Naïve Forecasting

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

```
133    253.0
134    253.0
135    173.0
136    186.0
137    185.0
```

Fig 14:Naïve forecast



MAPE:

```
Naive Approach MAPE: 29.172387135547105
```

RMSE:

```
Naive Approach RMSE(calculated using sklearn): 84.64840135437794
Naive Approach RMSE (calculated using statsmodels): 84.64840135437794
```

7)Simple Average

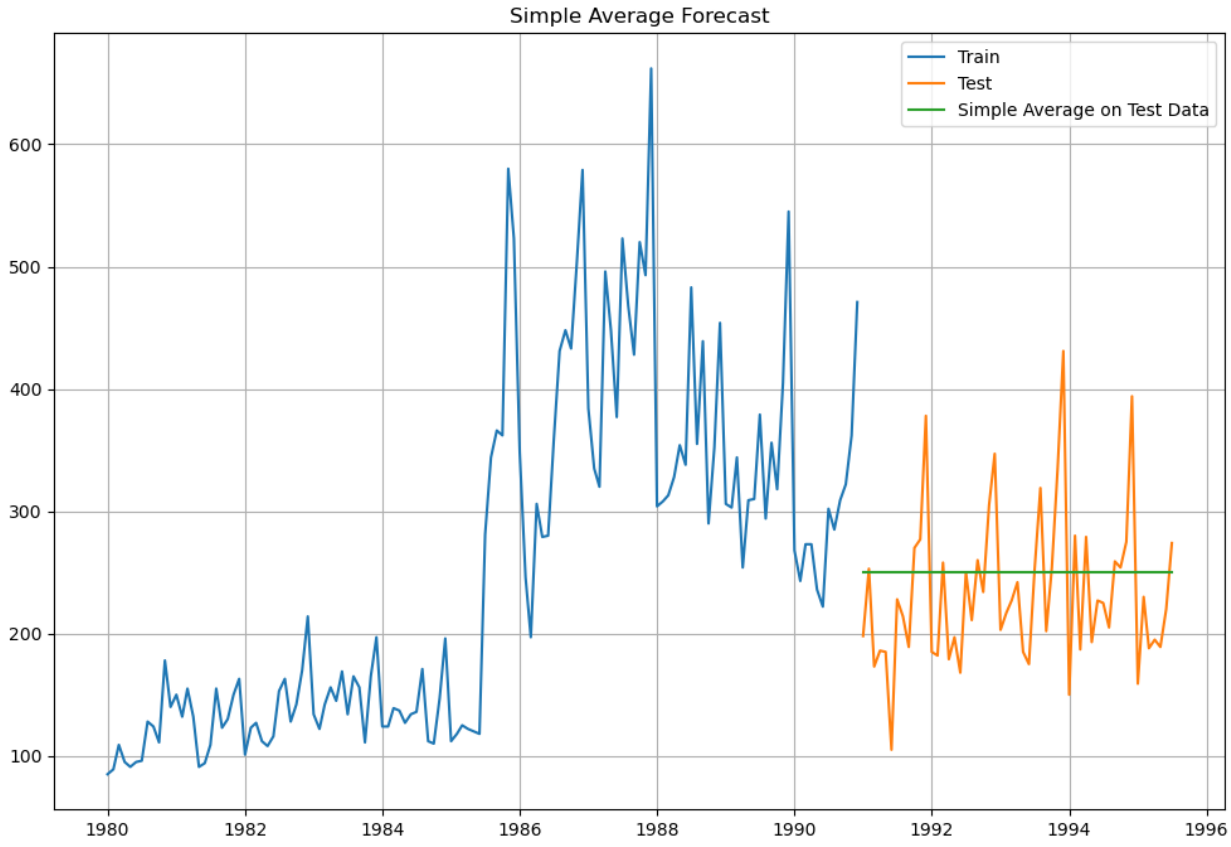
For this particular simple average method, we will forecast by using the average of the training values.

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

	Shoe-Sales	Year	Month	mean_forecast
YearMonth				
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
Critical Values:
        1%: -3.469

        5%: -2.878

        10%: -2.576
```

From the test we can see that ADF statistics value is more than critical values so we fell to reject the null hypothesis. and p value is also greater than 0.05,so we fell to reject the null hypothesis. The dataset is non stationary.

Now we need to change the data to stationary by square root transformation

ADF Test Value after square root transformation:

```
DF test statistic is -3.375
DF test p-value is 0.0119
Critical Values:
        1%: -3.469

        5%: -2.878

        10%: -2.576

From the test we can see that ADF statistic is less than the critical values
which means we can reject the null hypothesis.P value is also less than 0.05
which means we can reject the null hypothesis. The data is stationary.
```

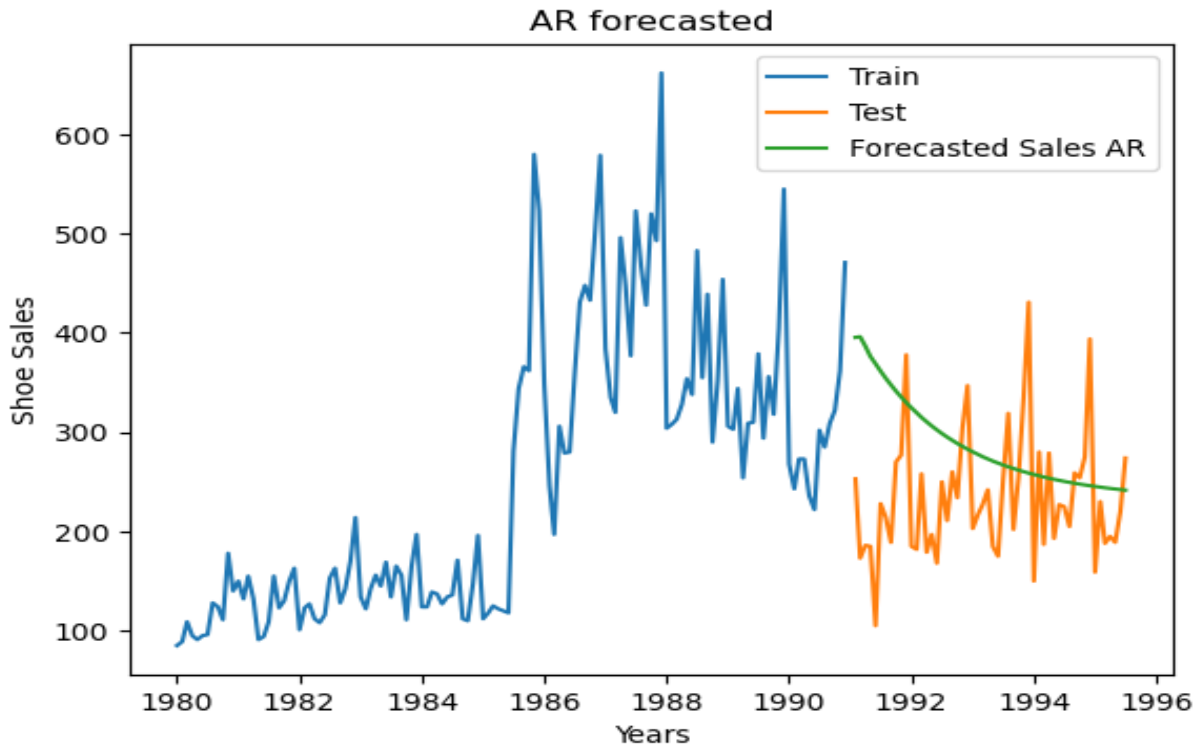
8)Auto Regressive Model

Table 14:Auto Regressive result with AIC (3,0,0)

```
         SARIMAX Results
==========================================================================
=
Dep. Variable:             Shoe_Sales   No. Observations:
132
Model:                 ARIMA(3, 0, 0)   Log Likelihood                -
275.023
Date:                Tue, 14 May 2024   AIC
560.046
Time:                        22:56:37   BIC
574.460
Sample:                             0   HQIC
565.903
                                - 132
Covariance Type:                  opg
==========================================================================
=
```

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         15.2093      2.113      7.197      0.000      11.067      19.351
ar.L1          0.6982      0.080      8.762      0.000       0.542       0.854
ar.L2          0.0555      0.120      0.463      0.643      -0.180       0.291
ar.L3          0.1754      0.091      1.927      0.054      -0.003       0.354
sigma2         3.7274      0.405      9.197      0.000       2.933       4.522
==============================================================================
==============================================================================
Ljung-Box (L1) (Q):                   0.08   Jarque-Bera (JB):
8.67
Prob(Q):                              0.78   Prob(JB):
0.01
Heteroskedasticity (H):               3.92   Skew:
-0.08
Prob(H) (two-sided):                  0.00   Kurtosis:
4.25
==============================================================================
MAPE:
AR MAPE: 39.94486632014559
RMSE:
Auto Regression RMSE(calculated using sklearn): 99.35376352725396
Auto Regression RMSE (calculated using statsmodels): 99.35376352725396
```

Fig 17:



9)ARMA Forecast

Table 15:ARMA result with AIC (3, 0, 3)

SARIMAX Results

Dep. Variable:	Shoe_Sales	No. Observations:	132
Model:	ARIMA(3, 0, 3)	Log Likelihood	-270.771
Date:	Tue, 14 May 2024	AIC	557.541
Time:	23:00:20	BIC	580.604
Sample:	0	HQIC	566.913
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	15.0886	2.279	6.621	0.000	10.622	19.555
ar.L1	-0.3077	0.111	-2.767	0.006	-0.526	-0.090
ar.L2	0.4236	0.058	7.276	0.000	0.309	0.538

```
ar.L3          0.7561      0.086      8.760      0.000       0.587
0.925
ma.L1          1.0145      0.230      4.408      0.000       0.563
1.466
ma.L2          0.4400      0.248      1.775      0.076      -0.046
0.926
ma.L3         -0.3924      0.132     -2.976      0.003      -0.651      -
0.134
sigma2         3.3952      0.829      4.095      0.000       1.770
5.020
==============================================================================
=
==============================================================================
======
Ljung-Box (L1) (Q):                   0.08   Jarque-Bera (JB):
5.92
Prob(Q):                              0.77   Prob(JB):
0.05
Heteroskedasticity (H):               3.56   Skew:
-0.20
Prob(H) (two-sided):                  0.00   Kurtosis:
3.96
==============================================================================
======
```

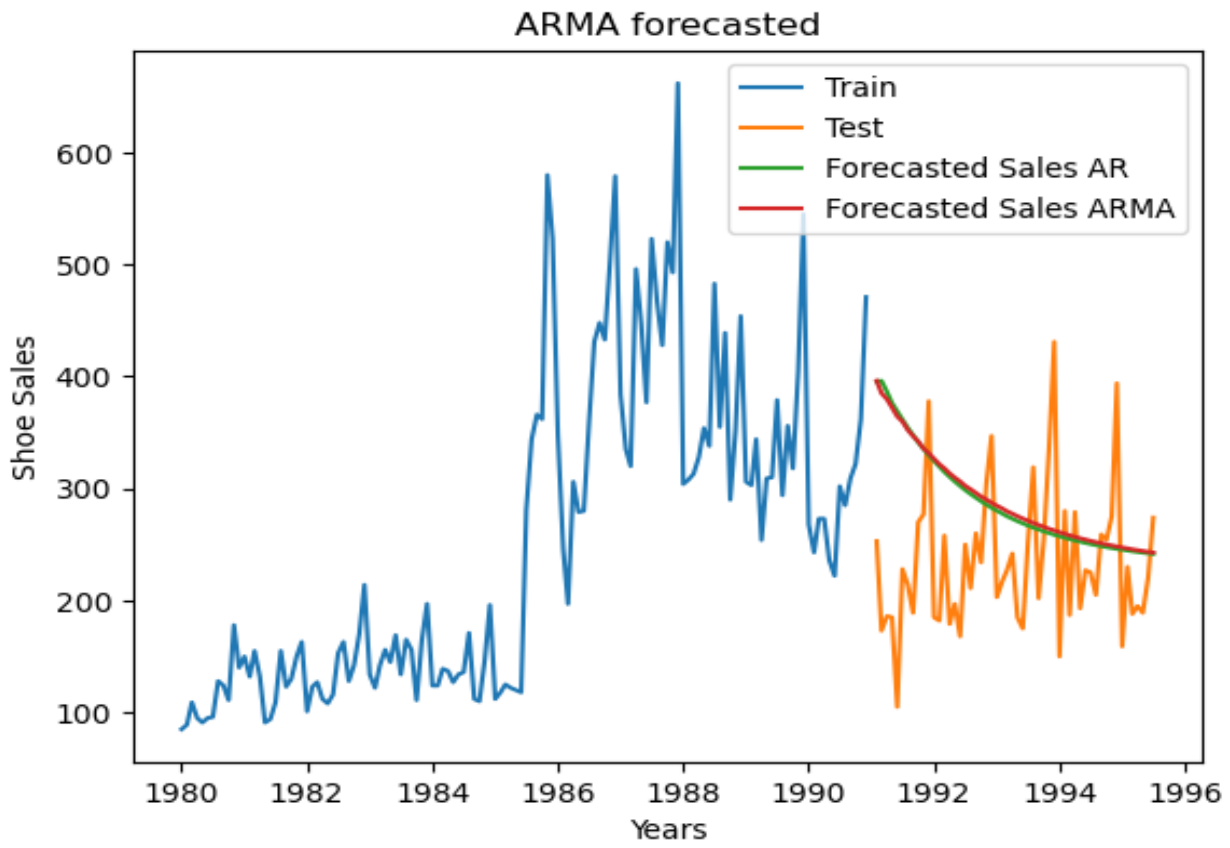
MAPE:

```
ARMA MAPE: 40.2921322267284
```

RMSE:

```
Auto Regression Moving Average RMSE(calculated using sklearn):
99.07071607091686
Auto Regression Moving Average RMSE (calculated using statsmodels):
99.07071607091686
```

Fig 18:ARMA Forecast



6. Build an 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) ARIMA forecast with AIC (3,1,2)

Table 16:ARIMA result SARIMAX Results

Dep. Variable:	Shoe_Sales	No. Observations:	132
Model:	ARIMA(3, 1, 2)	Log Likelihood	-258.419
Date:	Tue, 14 May 2024	AIC	528.838
Time:	23:01:43	BIC	546.089
Sample:	0	HQIC	535.848
	- 132		
Covariance Type:	opg		

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

```
------------------------------------------------------------------------------
-
ar.L1         -0.3453      0.090     -3.857      0.000      -0.521      -
0.170
ar.L2         -0.9996      0.010   -100.110      0.000      -1.019      -
0.980
ar.L3         -0.3422      0.080     -4.283      0.000      -0.499      -
0.186
ma.L1          0.0283      0.068      0.415      0.678      -0.105
0.162
ma.L2          0.9904      0.132      7.497      0.000       0.732
1.249
sigma2         2.9014      0.454      6.384      0.000       2.011
3.792
==============================================================================
=
==============================================================================
======
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):
18.97
Prob(Q):                              0.82   Prob(JB):
0.00
Heteroskedasticity (H):               4.73   Skew:
-0.56
Prob(H) (two-sided):                  0.00   Kurtosis:
4.49
==============================================================================
======


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

Table 17:Predicted result first few dataset(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
```

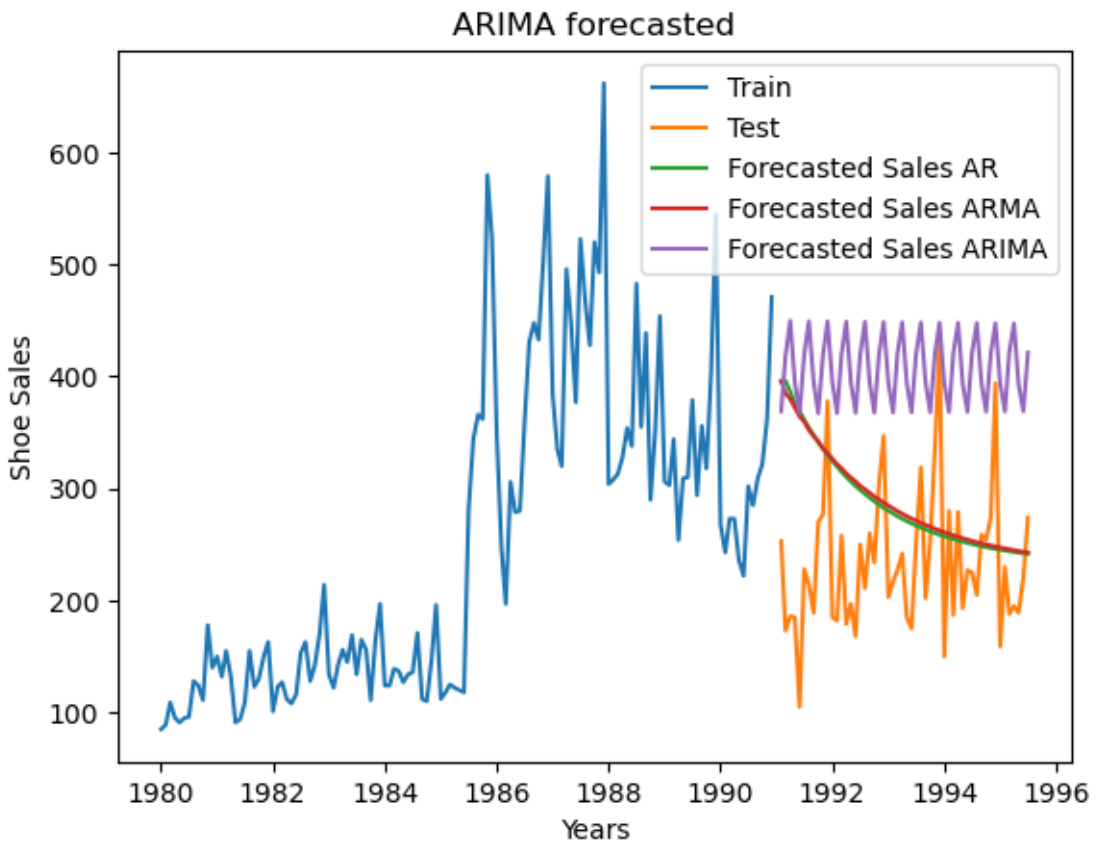
MAPE:

```
84.21653124542891
```

RMSE:

```
ARIMA RMSE(calculated using sklearn): 182.160170240246
ARIMA RMSE (calculated using statsmodels): 182.160170240246
```

Fig 19: ARIMA forecast



11) SARIMA

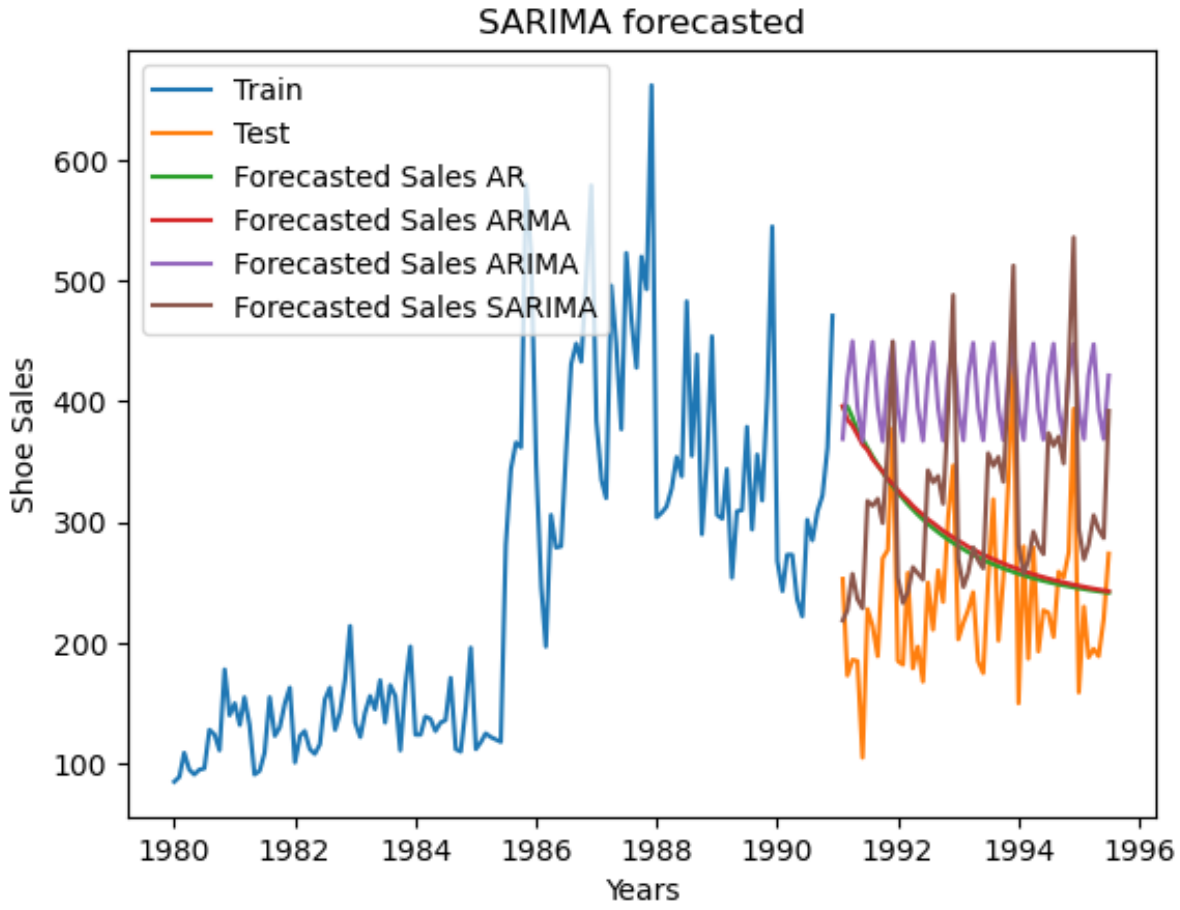
Table 18- SARIMA result

```
         SARIMAX Results
==========================================================================================
==================
Dep. Variable:                                        y   No. Observations:
132
Model:             SARIMAX(3, 1, 0)x(3, 0, [1, 2], 12)   Log Likelihood
-173.183
Date:                                 Tue, 14 May 2024   AIC
364.366
Time:                                         23:08:16   BIC
387.062
Sample:                                              0   HQIC
373.526
                                                 - 132
Covariance Type:                                   opg
==========================================================================================
=
                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------------------
-
ar.L1         -0.3540      0.095     -3.741      0.000      -0.539      -
0.169
ar.L2          0.0366      0.098      0.375      0.707      -0.155
0.228
```

```
ar.L3          0.0320      0.087      0.367      0.714      -0.139
0.203
ar.S.L12       0.8112      0.593      1.367      0.172      -0.352
1.974
ar.S.L24       0.1302      0.737      0.177      0.860      -1.314
1.574
ar.S.L36       0.0928      0.247      0.375      0.707      -0.392
0.577
ma.S.L12      -0.5044      0.716     -0.704      0.481      -1.908
0.899
ma.S.L24      -0.3298      0.620     -0.532      0.595      -1.546
0.886
sigma2         2.2287      0.840      2.652      0.008       0.581
3.876
===================================================================================
======
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):
2.01
Prob(Q):                              0.99   Prob(JB):
0.37
Heteroskedasticity (H):               1.19   Skew:
-0.04
Prob(H) (two-sided):                  0.63   Kurtosis:
3.72
===================================================================================
======

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

Fig 20:SARIMA result graph



MAPE SARIMA:

```
38.983993532156866
```

RMSE:

```
SARIMA RMSE(calculated using sklearn): 92.1066303464584
SARIMA RMSE (calculated using statsmodels): 92.1066303464584
```

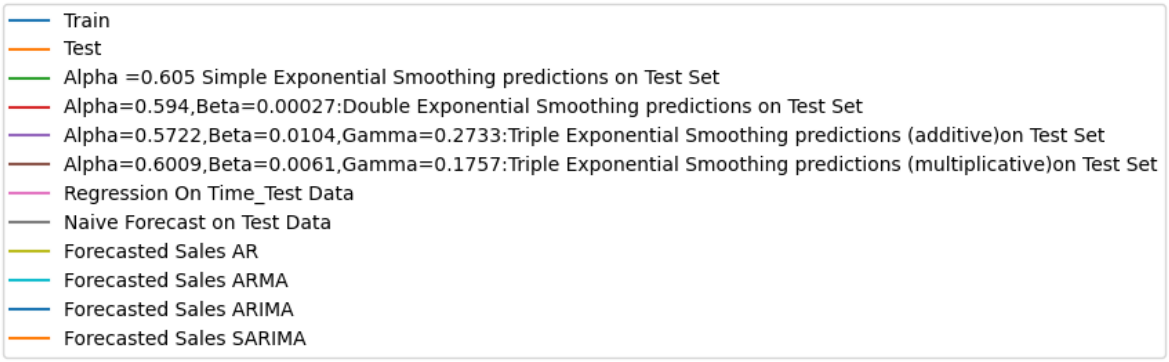
7. Build a table (create a data frame) with all the models built along with their

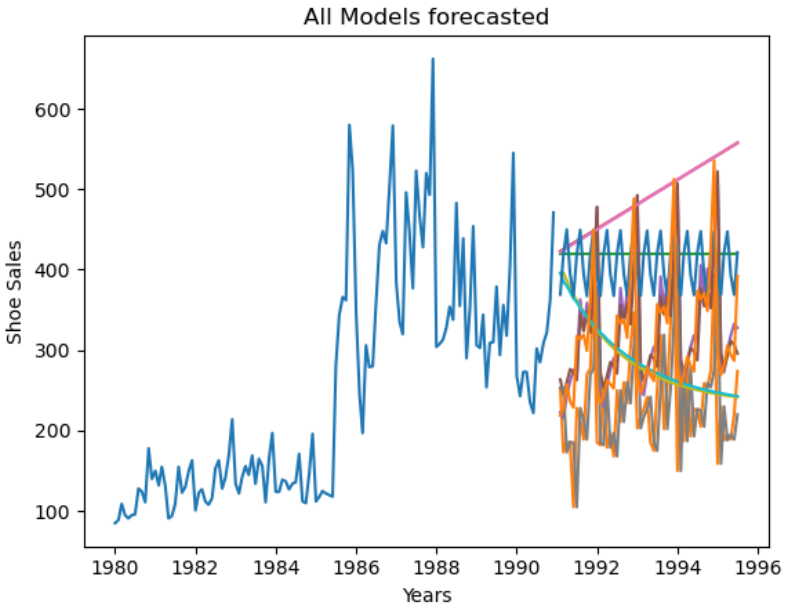
corresponding parameters and the respective RMSE values on the test data.

	RMSE (sklearn)	RMSE (statsmodel)	MAPE
AutoRegression(3,0,0)	99.353764	99.353764	39.944866
AutoRegression Moving Average(ARMA(3,0,3)	99.070716	99.070716	40.292132
ARIMA(3,1,2)	182.160170	182.160170	84.216531

SARIMA(3,1,2)(3,0,2,12)	92.106630	92.106630	38.983994
Alpha=0.605,SES	195.894595	195.894595	91.420760
Alpha=0.594,Beta=0.00027,DES	264.824827	264.824827	122.325284
Alpha=0.5722,Beta=0.0104,Gamma=0.2733,TES Additive	126.387633	126.387633	49.906727
Alpha=0.6009,Beta=0.0061,Gamma=0.1757,TES Multiplicative	122.178545	122.178545	48.032998
Linear RegressionOnTime	264.516794	264.516794	122.140758
Naive Approach	84.648401	84.648401	29.172387

Fig 21:Consolidated graph of all models





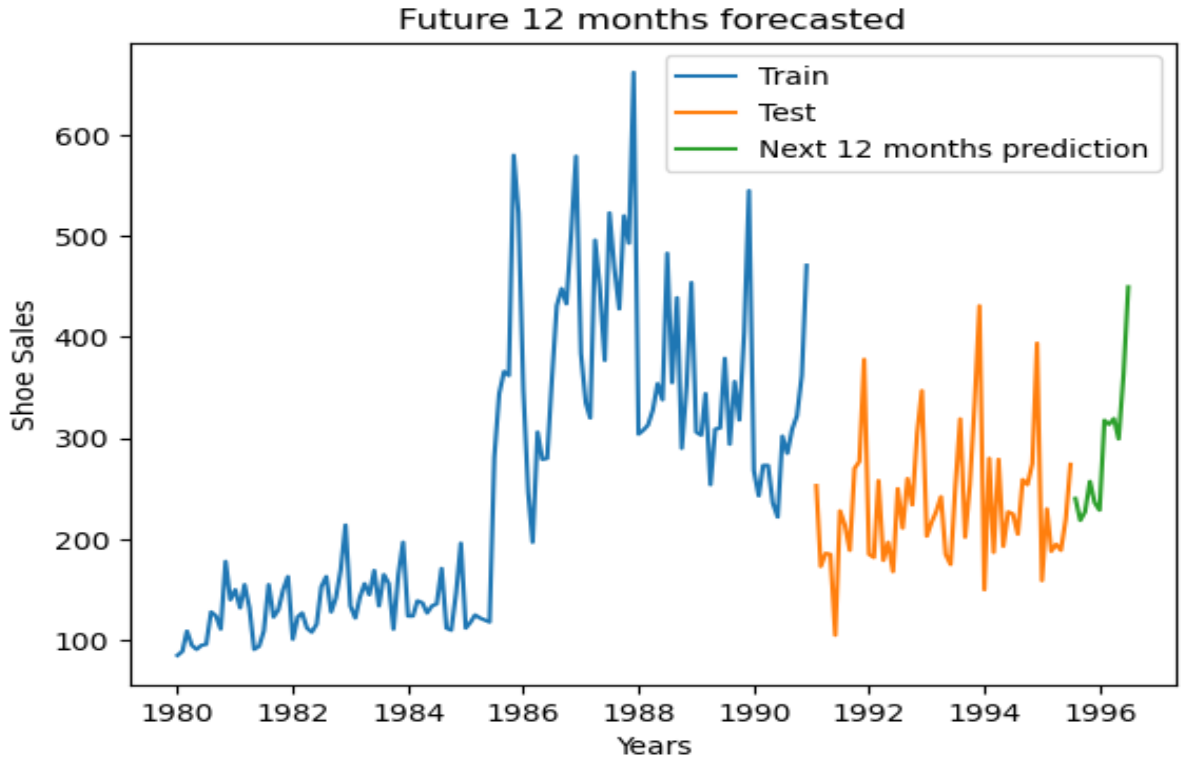
From this consolidated result We can see that Naïve Approach model has the lowest RMSE value and MAPE value.2nd lowest is the SARIMA value.It is the best model to predict for shoe sales .

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 SARIMA Model

	Shoe_Sales
1995-08-01	240.234233
1995-09-01	218.654236
1995-10-01	227.131167
1995-11-01	257.170779
1995-12-01	236.259481
1996-01-01	228.727932
1996-02-01	317.641622
1996-03-01	313.573962
1996-04-01	319.337549
1996-05-01	299.325629
1996-06-01	363.070595
1996-07-01	449.716634

Fig 22:Future 12 months forecasted(using SARIMA)



FINDINGS AND SUGGESTIONS:

1)There are outliers in the month of April and May which tells us that some sales were unsual. 2)In the 2nd half of each year the sales increases.December has the highest sales which may be due to holiday season. 3)The sales can be increased by advertising more and launching new shoes . 4)Discount should be added in the off season(when the sales decreases) to increase the sales.