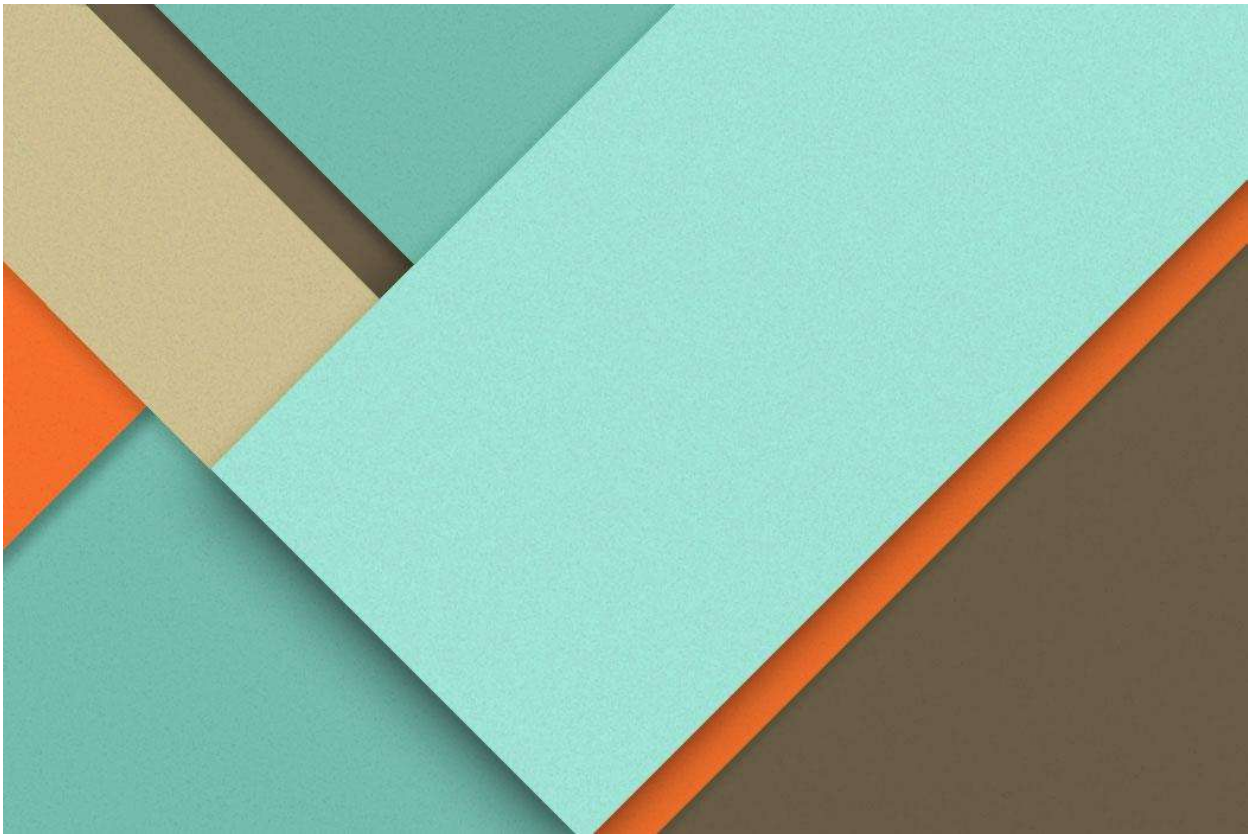

Time Series Forecasting

Shoe Sales Forecasting || SoftDrink Production Forecasting

—

Sourav Maity

PGP - DSBA

Group B5, June 22

Problem 1 : Shoe Sales Forecasting

1.1.1 Objective

The objective of the problem statement is to build an optimum model to forecast the sales of pairs of shoes for the upcoming 12 months from where data currently ends. We additionally have to comment on the models thus build and report our findings and suggest the measure that the company should be taking for future Sales.

1.1.2 Data Dictionary

YearMonth	Month & Year of Shoe Sales
Shoe Sales	The monthly Sales of the Shoes

1.1.3 Descriptive Data Analysis

Top 5 and Bottom 5 Views of the Dataset

	YearMonth	Shoe_Sales
0	1980-01	85
1	1980-02	89
2	1980-03	109
3	1980-04	95
4	1980-05	91

	YearMonth	Shoe_Sales
182	1995-03	188
183	1995-04	195
184	1995-05	189
185	1995-06	220
186	1995-07	274

Information of the dataset

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Shoe_Sales  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

Descriptive Function of the data

	count	mean	std	min	25%	50%	75%	max
Shoe_Sales	187.0	245.636364	121.390804	85.0	143.5	220.0	315.5	662.0

➔ There are no missing values in the dataset.

1.3 Time Series Data : Plot

It is clearly visible from the **Figure** that there is a yearly trend in Shoe Sales with seasonality pattern every year

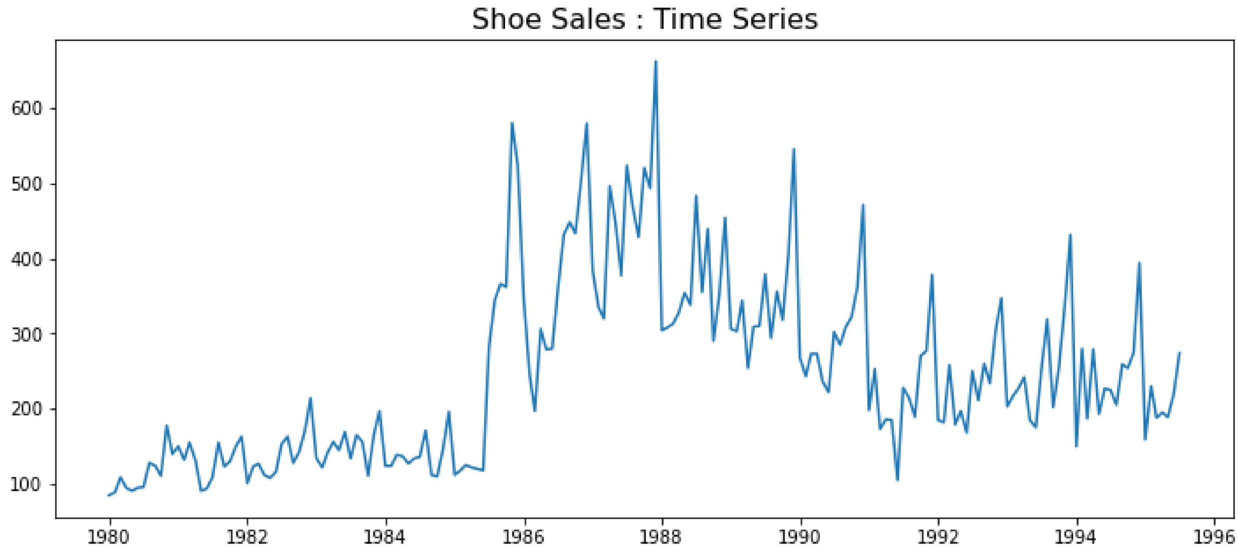

Fig 1. Time Series Plot - Shoe Sales

1.2 Exploratory Data Analysis : Shoe Sales

1.2.1 EDA

This refers to the critical process of initial investigation on the data, to discover patterns, spot anomalies and test hypotheses.

In Figure - **2**, we can see the monthly boxplot of Shoe Sales and this reveals that there were some sales in the month of ‘April’ & ‘May’ which were out of the box. The Sales trend also tends to pick up in the second half of the year and touches highest during December, which may be due to the festive season.

In Figure - 3, we can see that the yearly Sales saw a rapid growth from the year 1985 to 1987. This growth may be due to widespread interest and the result of innovations done to attract the customers.

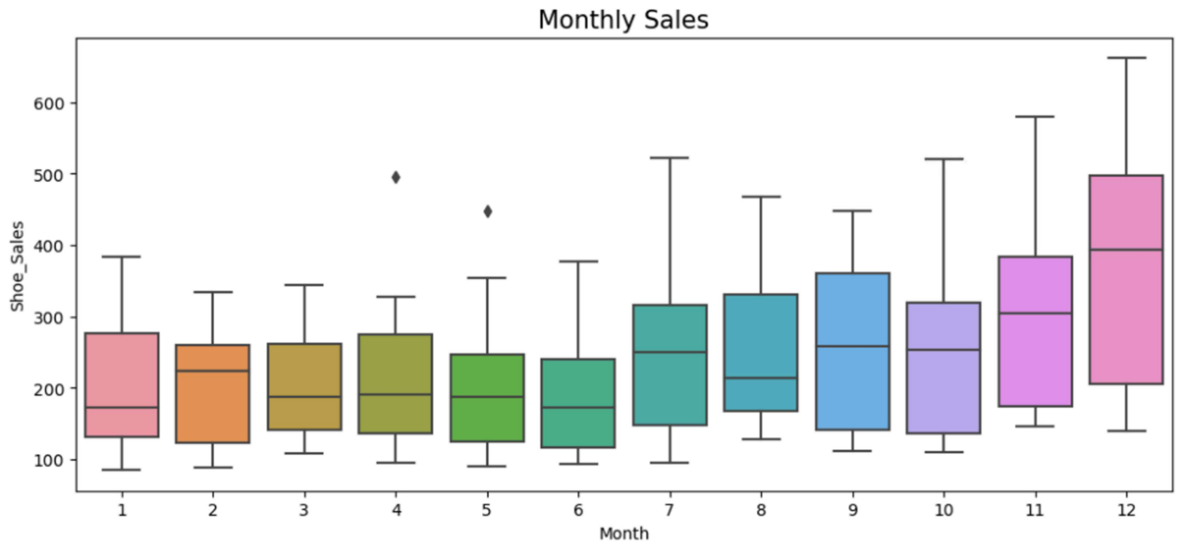

Fig 2. Boxplot - Monthly Shoe Sales

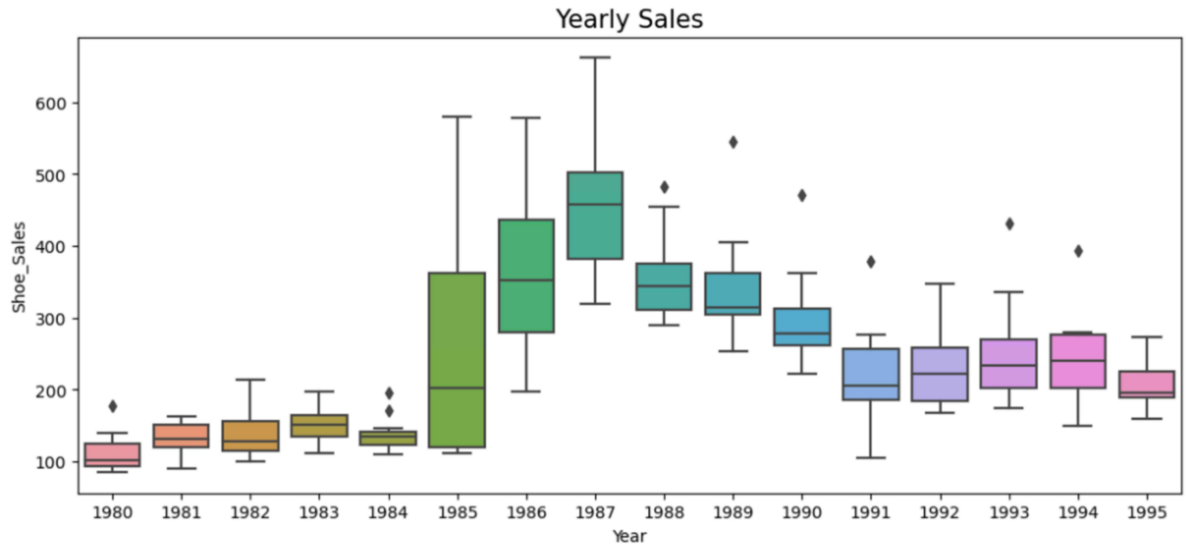

Fig 3. Boxplot - Yearly Shoe Sales

1.2.2 Decomposition

Decomposition of a Time Series consists of a series of combinations of trend, seasonality and residuals i.e noise components. This provides a useful abstract model for thinking about time series and generally gives a better understanding of the problem.

Additive Decomposition is used where change is measured in absolute quantity

Multiplicative Decomposition is used where change is measured in percentage of quantity

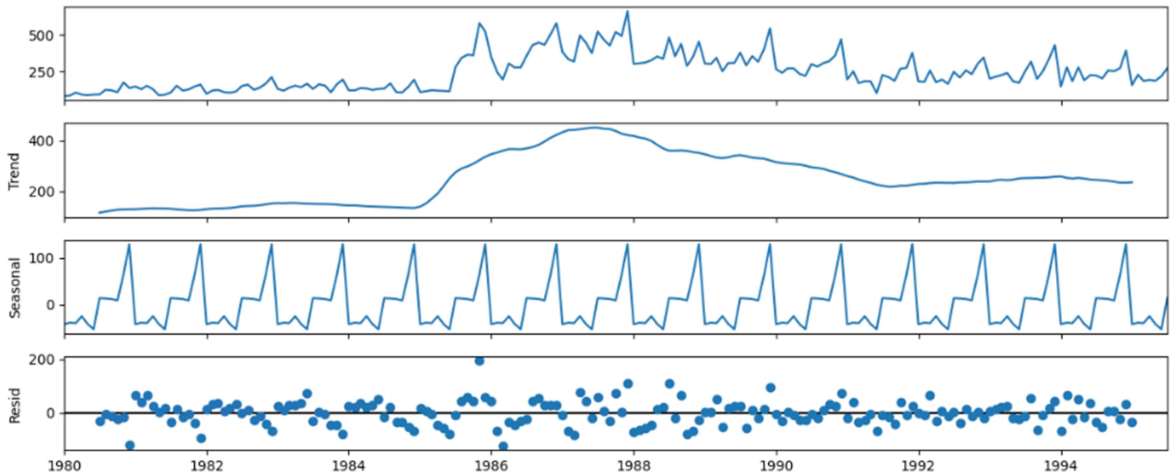

Fig 4. Decomposition - Shoe Sales : Additive

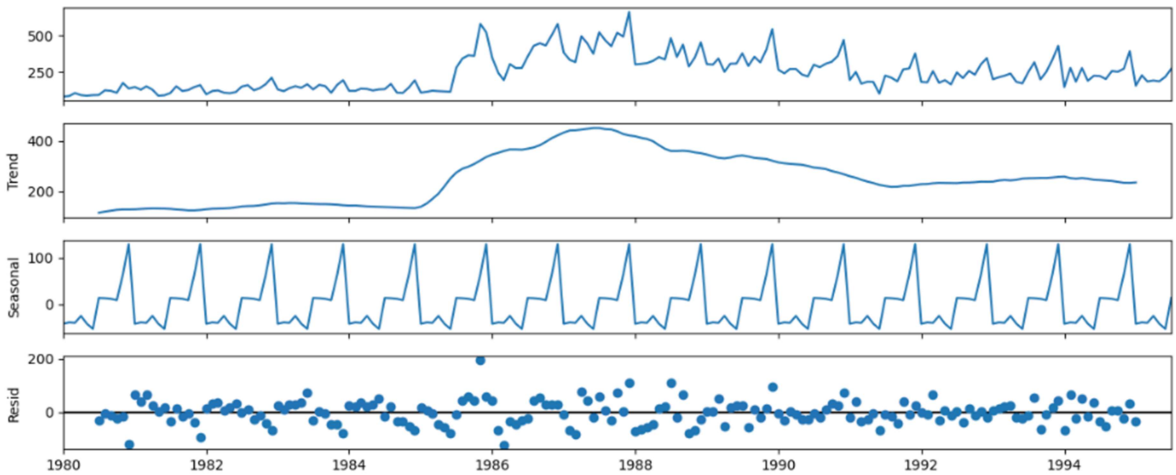

Fig 5. Decomposition - Shoe Sales : Multiplicative

1.3 Splitting of train & test

The train test split is used to estimate the performance of the machine learning algorithm that is used for predictions. The dataset has been splitted between the years 1991. That means test data starts from 1991.

```
Training Data:
             Shoe_Sales
YearMonth
1980-01-01           85
1980-02-01           89
1980-03-01          109
1980-04-01           95
1980-05-01           91
Shape: (132, 1)
```

```
Testing Data:
             Shoe_Sales
YearMonth
1991-02-01          253
1991-03-01          173
1991-04-01          186
1991-05-01          185
1991-06-01          105
Shape: (54, 1)
```

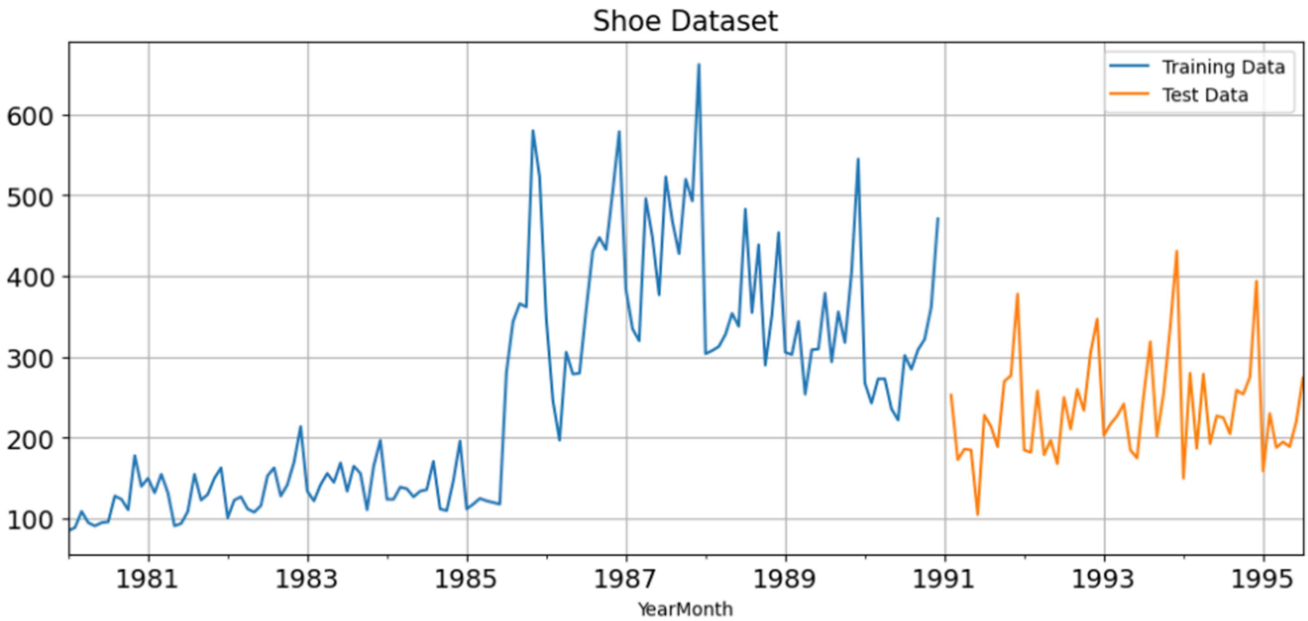

Fig 6. Train- Test Split : Shoe Sales Data

1.4 Building different models and checking RSME

We are building various exponential smoothing models on the training dataset and evaluating the model using RMSE on the test data. The models used are: Linear Regression, NaiveBayes, SimpleAverage, MovingAverage and checking the performance on the test data using RMSE.

The main objective of building multiple models is to pick the optimum model with lowest RMSE and MAPE values.

MAPE - Mean Absolute Percentage Error. It is the average multiplicative effect between the estimated mean and observed outcome. RMSE - Root Mean Square Error i.e standard deviation.

1.4.1 Linear Regression

LinearRegression is a great predictive analytic technique which uses Time as the independent variable and Shoe_Sales as the dependent variable. We can see in Figure 7 that the Linear Regression on time is a constant rise.

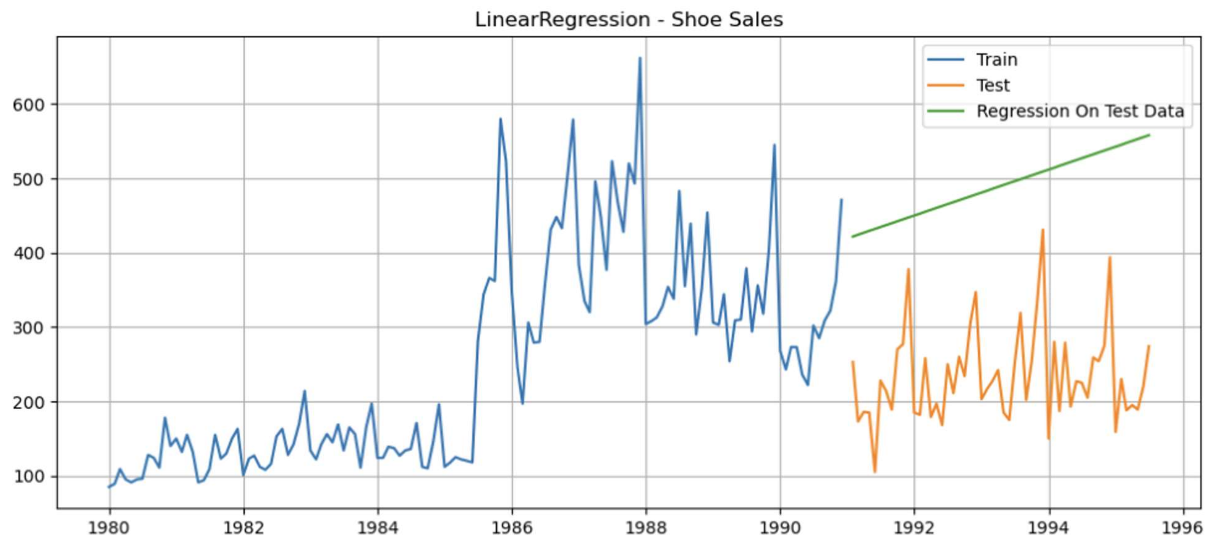

Fig 7. Linear Regression : Shoe Sales Data

The RMSE for LinearRegression is 264.5 approximately with Mean Absolute Percentage Error around 122% which portrays that LinearRegression is not a good model here.

	Test RMSE	MAPE
LinearRegression	264.516794	122.140758

1.4.2 Naive Bayes

Naive Bayes makes use of language models for classifying and making predictions. From Figure 8, it is clear that the Naive Bayes forecast on the test data is constantly the same for the whole data which isn't ideal.

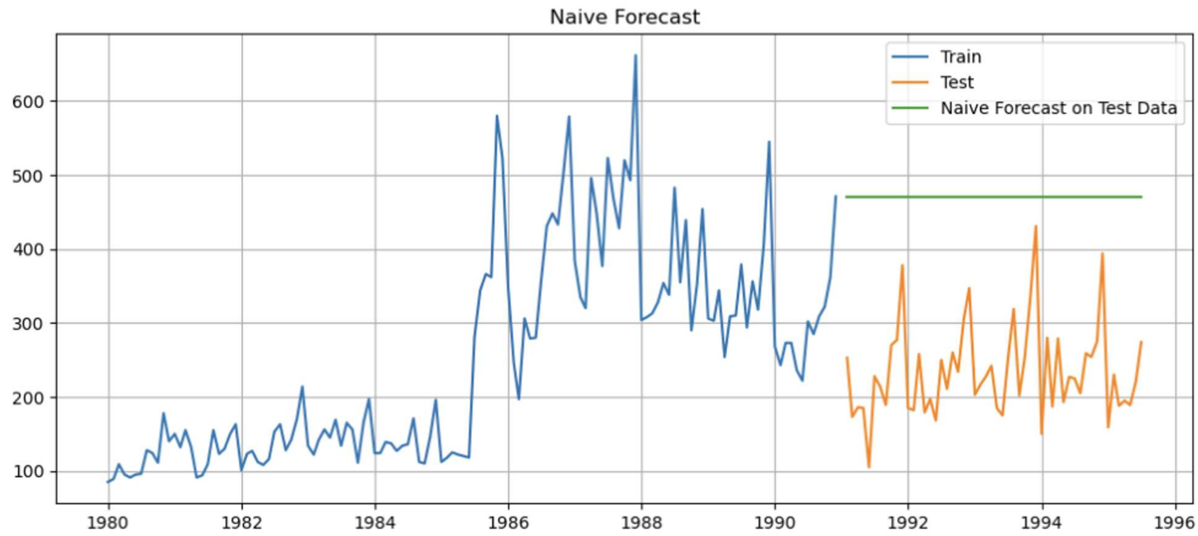

Fig 8 Naive Bayes : Shoe Sales Data

	Test RMSE	MAPE
LinearRegression	264.516794	122.140758
NaiveBayes	244.575066	114.443505

The RMSE for NaiveBayes is 244.5 approximately with Mean Absolute Percentage Error around 114% which also portrays that NaiveBayes is not a good model

1.4.3 SimpleAverage

Simple Average makes use of the mean of the whole data. In Figure 9, it is clear that the forecast on test data using SimpleAverage is a straight line. Although it is close to the actuals, it cannot be a straight line.

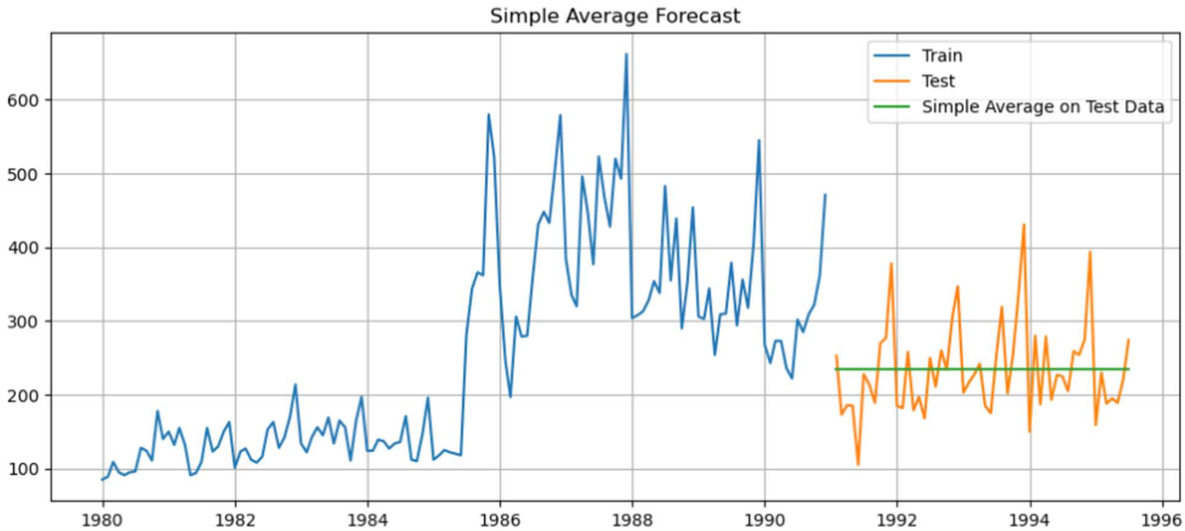

Fig 9: SimpleAverage : Shoe Sales Data

	Test RMSE	MAPE
LinearRegression	264.516794	122.140758
NaiveBayes	244.575066	114.443505
SimpleAverage	62.116279	21.327011

The RMSE for SimpleAverage is 62.11 approximately with Mean Absolute Percentage Error around 21.3% which also portrays that NaiveBayes is not a good model.

1.4.4 MovingAverage

Moving Average makes use of the mean of the rolling data. In Figure 10, it is clear that the forecast on test data using Rolling averages of 2 point, 4 points, 6 points and 9 points and goes very close to the actuals of Test data. Although it is close to the actuals.

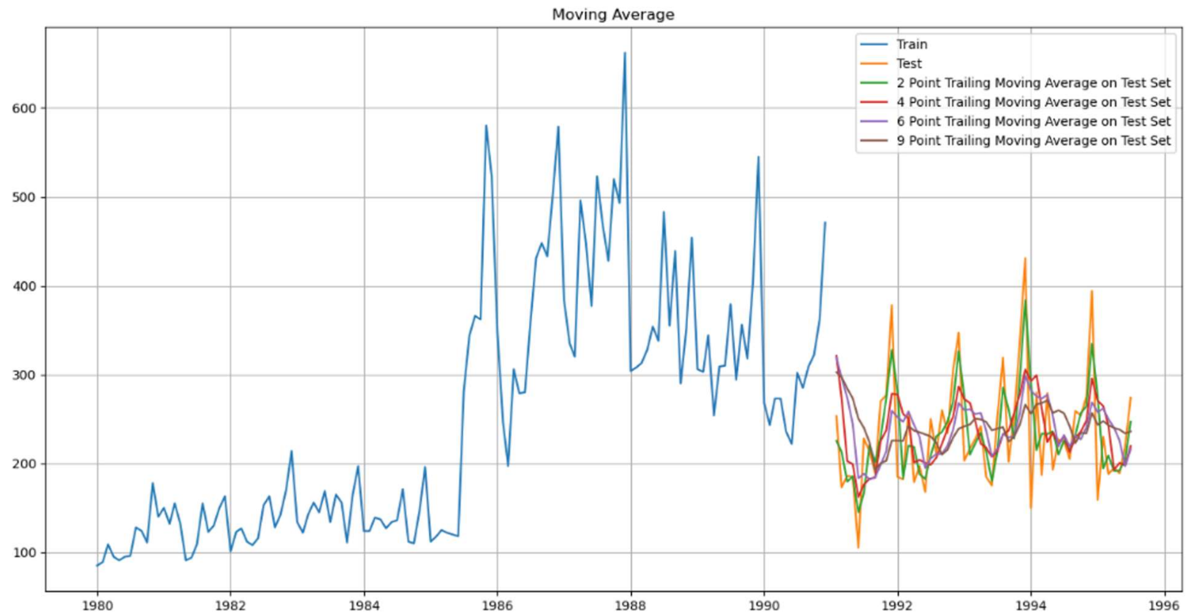

Fig 10: SimpleAverage : Shoe Sales Data

	Test RMSE	MAPE
LinearRegression	264.516794	122.140758
NaiveBayes	244.575066	114.443505
SimpleAverage	62.116279	21.327011
Trailing_2	42.489323	14.787482
Trailing_4	55.199732	19.835338
Trailing_6	61.684757	23.269147
Trailing_9	66.901505	24.726947

The RMSE of 2 point Rolling Moving Average is 42.48 approximately which is lowest of all and very close to the actuals of the Test data.

1.4.5 Simple Exponential Smoothing

Simple Exponential Smoothing is a time series forecasting method for univariate data without trend or seasonality. It requires a single parameter called alpha(α) also called smoothing factor or smoothing coefficient. The alpha value or smoothing coefficient at which graph is plotted id 0.60

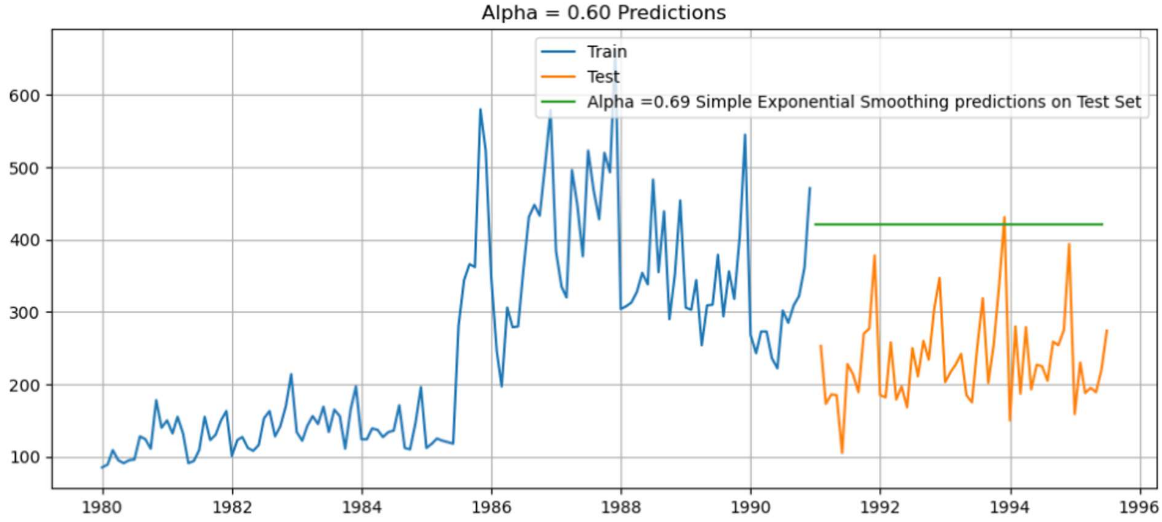

Fig 11: Simple Exponential Smoothing - Shoe Sales Data

1.4.6 Double Exponential Smoothing

Double exponential smoothing employs a level component trend component at each period. The alpha value at which graph is plotted is 0.594 and beta value is 0.0002

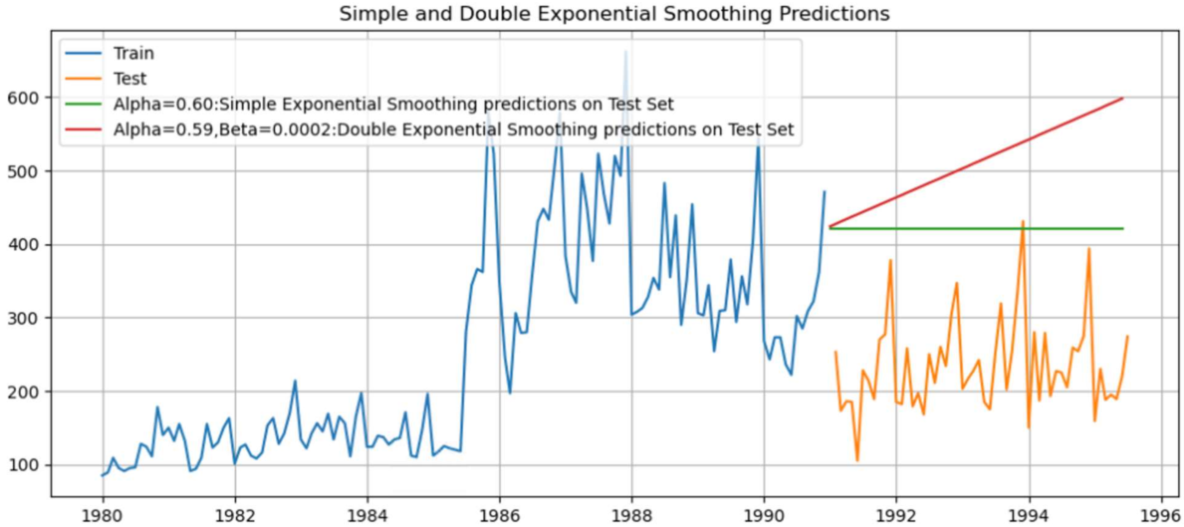

Fig 12: Double Exponential Smoothing - Shoe Sales Data

1.4.7 Triple Exponential Smoothing

This is used in three smoothing equations. Stationary components, trends and seasonal. This is a multiplicative model. The alpha value at which graph is plotted is 0.571, beta value or smoothing trend is 0.001 and gamma or seasonal is 0.202

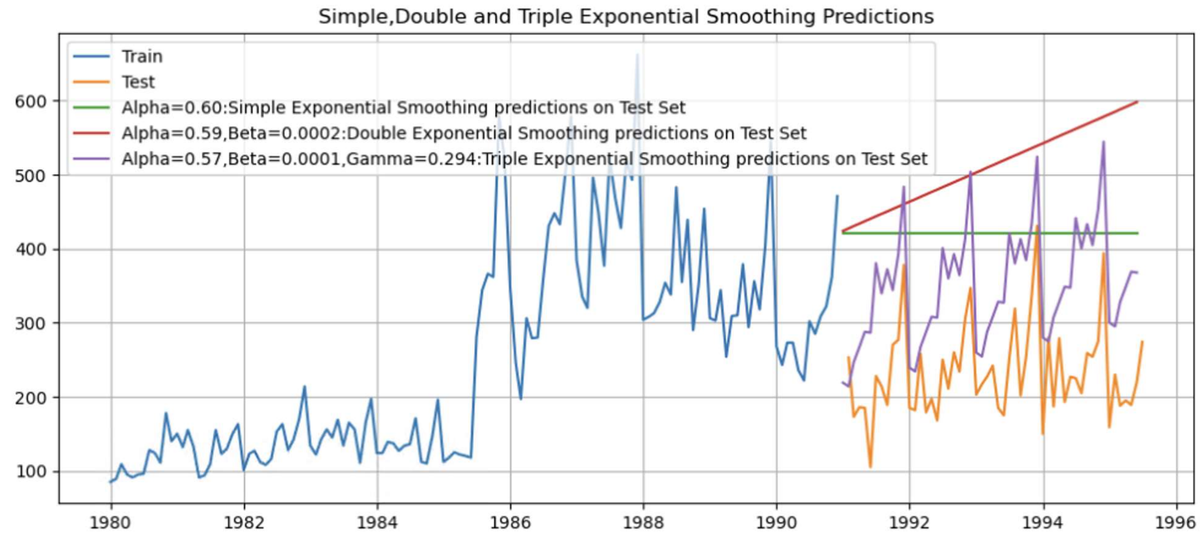

Fig 13: Triple Exponential Smoothing - Shoe Sales Data

	Test RMSE	MAPE
LinearRegression	264.516794	122.140758
NaiveBayes	244.575066	114.443505
SimpleAverage	62.116279	21.327011
Trailing_2	42.489323	14.787482
Trailing_4	55.199732	19.835338
Trailing_6	61.684757	23.269147
Trailing_9	66.901505	24.726947
SES	195.894561	91.420744
DES	286.467010	131.556013
TES	142.698275	57.907264

The RMSE for Moving Average Trailing 2 is lowest and giving closest prediction against the actuals in the test data

1.5 Check for stationarity

- The Augmented Dickey - Fuller Test is a unit root test which determines whether there is a unit root and subsequently whether the series is non-stationary.
- The hypothesis in a simple form for the ADF test is:
  - H0 : The time series has a unit root and is thus non stationary
  - H1 : The time series does not have a unit root and is thus stationary
- When ADF was applied on the model we got a p-value of 0.801 which is higher than 0.05 and we fail to reject the null hypothesis concluding that series is not stationary
- Hence we have done level differencing on the dataset and check for stationarity.
- The p-value after level 1 differencing is 0.036<0.05, hence we now reject the null hypothesis and conclude that the series is stationary with a lag of 1
- Now as the data is stationary we can move to build ARIMA and SARIMA models.

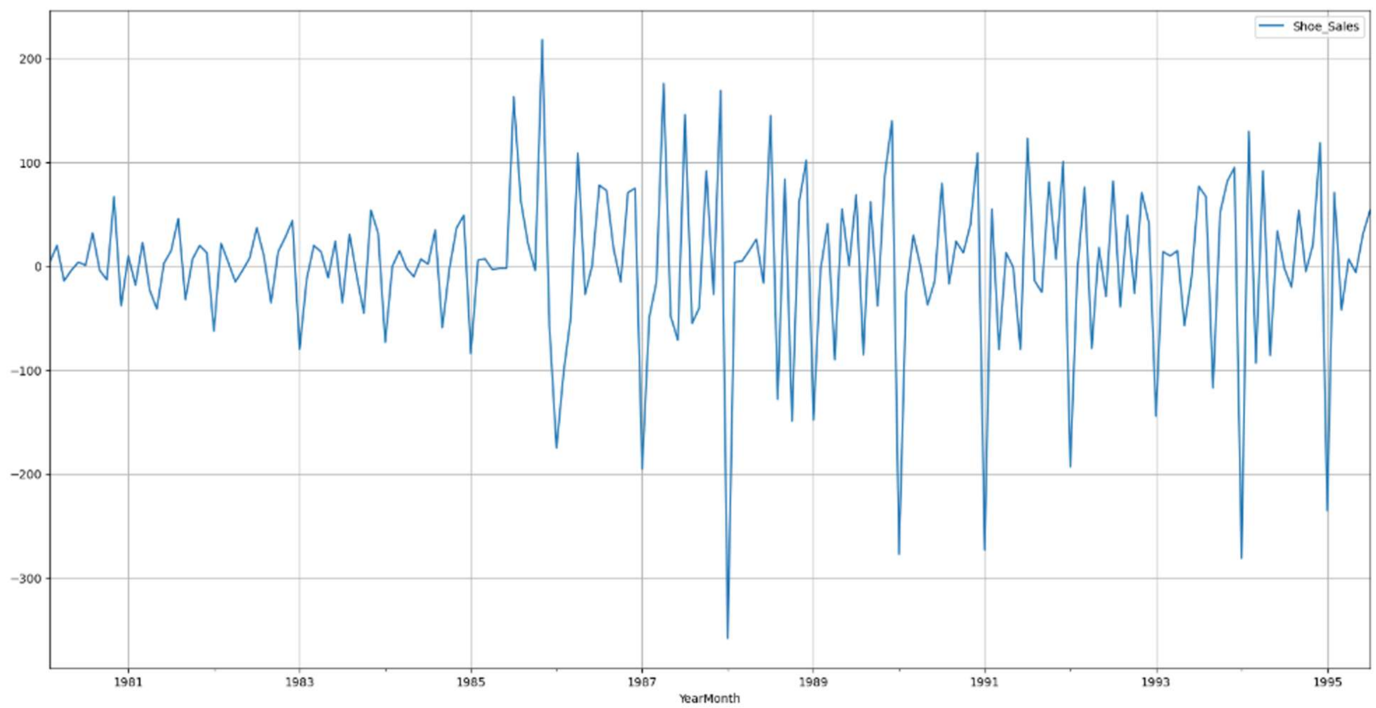

Fig 14: Stationarity of Shoe Sales Data

1.6 ARIMA & SARIMA using lower AIC methods

1.6.1 ARIMA

- An ARIMA model consists of the Auto Regressive part and Moving Average part after we have made the time series stationary by taking the correct degree and order of differencing.
- ARIMA models can be built keeping the Akaike Information Criterion (AIC) in mind as well. In this case we choose the p and q values to determine the AR and MA order respectively which gives us the lowest AIC value. Lower the AIC better is the model
- Coding languages tries different method of 'p' and 'q' to arrive to this conclusion
- The formula for calculating the AIC is 2k-2ln(L) where k is the no. of parameters to be estimated and L is the likelihood estimation.
- ARIMA:
    - We first create a grid of all possible outcomes (p,d,q). The range of 'p' and 'q' being (0,4) and d is constant =1
    - The lowest AIC for ARIMA is clearly (2,1,3) with an AIC of 1479
- The ARIMA summary , graph and diagnostic results are :

	param	AIC
15	(3, 1, 3)	1479.687155
11	(2, 1, 3)	1480.798464
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

```
                               SARIMAX Results
==============================================================================
Dep. Variable:             Shoe_Sales   No. Observations:                  132
Model:                 ARIMA(2, 1, 2)   Log Likelihood                -743.205
Date:                Tue, 31 Oct 2023   AIC                           1496.411
Time:                        13:39:17   BIC                           1510.787
Sample:                    01-01-1980   HQIC                          1502.252
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.6739      4.219      0.160      0.873      -7.594       8.942
ar.L2         -0.0695      1.993     -0.035      0.972      -3.976       3.837
ma.L1         -1.0484      4.221     -0.248      0.804      -9.321       7.224
ma.L2          0.1708      3.512      0.049      0.961      -6.713       7.055
sigma2      4941.5636    439.454     11.245      0.000    4080.250    5802.877
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                54.37
Prob(Q):                              0.91   Prob(JB):                         0.00
Heteroskedasticity (H):              12.71   Skew:                             0.00
Prob(H) (two-sided):                  0.00   Kurtosis:                         6.16
===================================================================================
```

Fig 15.a: Summary of ARIMA

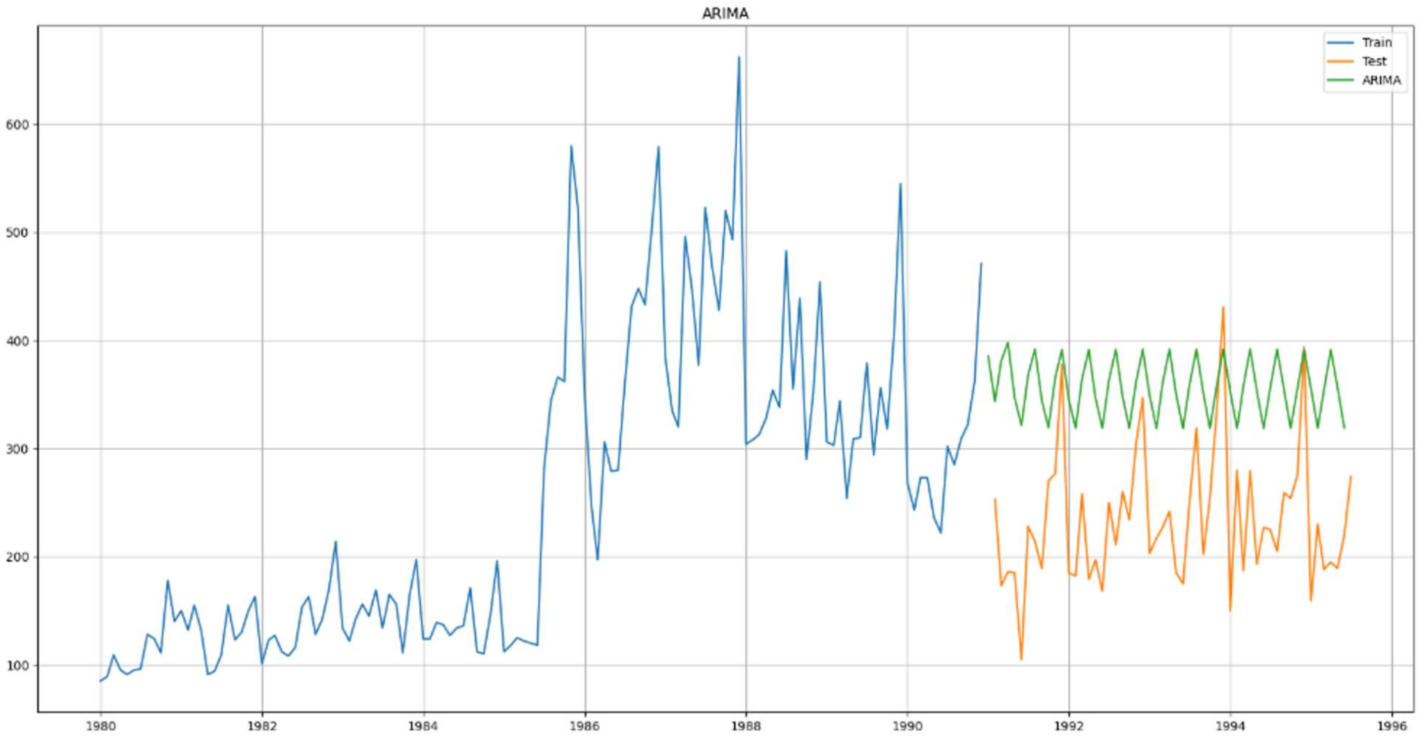

Fig 15.b Graph of ARIMA

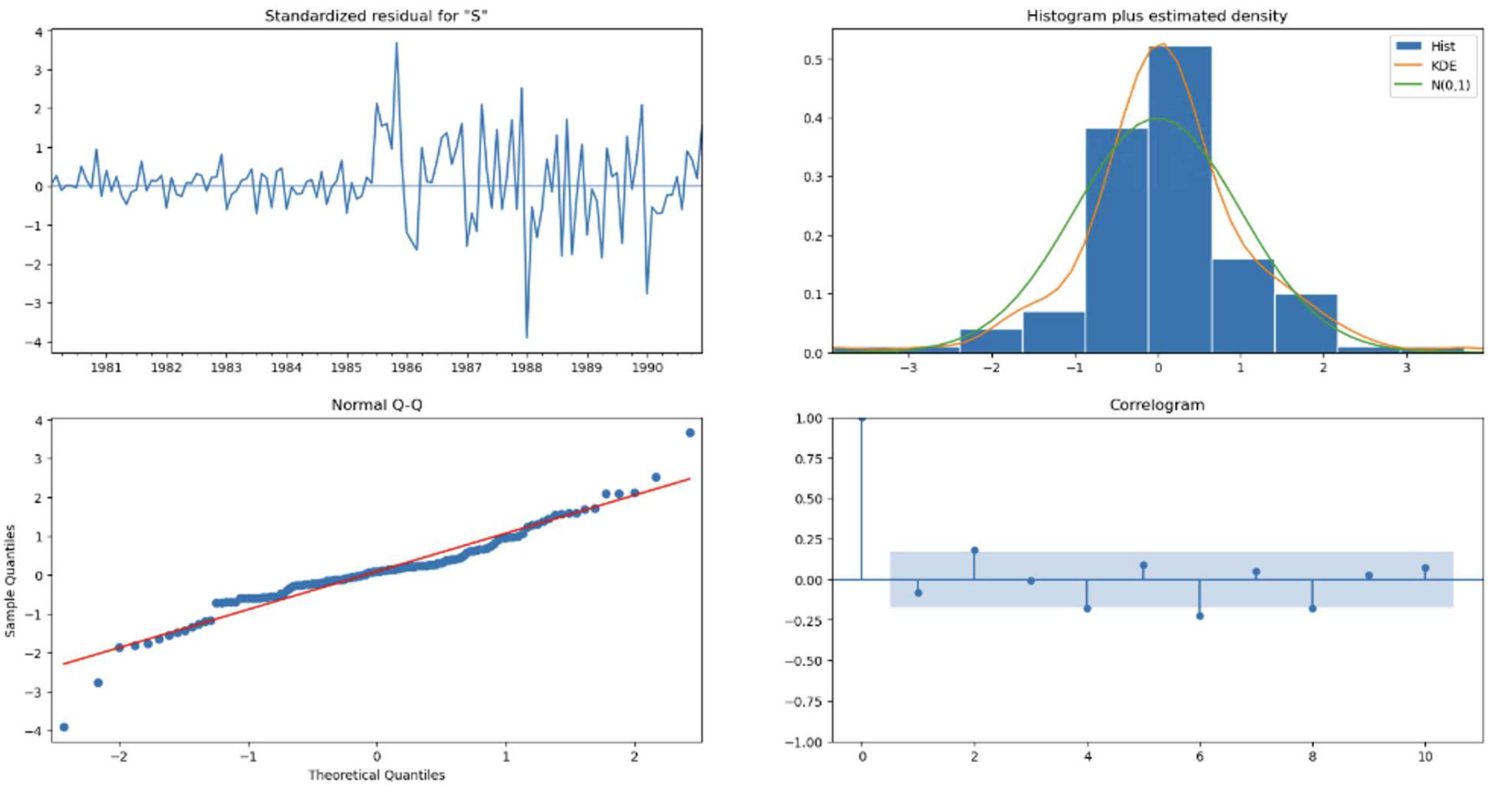

Fig 15.c: Diagnostic result of ARIMA

	RMSE	MAPE
ARIMA(2,1,2)	142.136266	64.286625

The RMSE of ARIMA is 142.13 and MAPE is 64.28

1.6.2 SARIMA

- We create a grid of all possible combinations of (p,d,q) along with seasonal (P,D,Q) and seasonality of 12. The range of 'p' and 'q' being (0,4) and 'd' constant = 1
- The following is the grid of all possible outcomes

```
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 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: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
```

- Going forward we are fitting the SARIMA model into each of the above combinations and end up choosing that one with the least AIC value.
- The lowest AIC for SARIMA is clearly (0,1,2) (1,0,2,12) with an AIC of 1156.1654. Hence we are going to fit the train data and forecast on the test set. And we get the SARIMA summary, graph, diagnostic results.

	param	seasonal	AIC
23	(0, 1, 2)	(1, 0, 2, 12)	1156.165429
50	(1, 1, 2)	(1, 0, 2, 12)	1157.082589
26	(0, 1, 2)	(2, 0, 2, 12)	1157.772313
77	(2, 1, 2)	(1, 0, 2, 12)	1158.491000
80	(2, 1, 2)	(2, 0, 2, 12)	1158.630324
53	(1, 1, 2)	(2, 0, 2, 12)	1158.794178
14	(0, 1, 1)	(1, 0, 2, 12)	1164.297459
41	(1, 1, 1)	(1, 0, 2, 12)	1165.179255
17	(0, 1, 1)	(2, 0, 2, 12)	1165.875706
68	(2, 1, 1)	(1, 0, 2, 12)	1166.103015

```
                                    SARIMAX Results
==========================================================================================
Dep. Variable:                             Shoe_Sales   No. Observations:             132
Model:             SARIMAX(1, 1, 3)x(3, 0, 3, 6)   Log Likelihood             -594.896
Date:                            Wed, 01 Nov 2023   AIC                        1211.792
Time:                                    11:46:06   BIC                        1241.396
Sample:                                01-01-1980   HQIC                       1223.797
                                     - 12-01-1990
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.2295      0.340      0.676      0.499      -0.436       0.895
ma.L1         -0.6392      0.336     -1.903      0.057      -1.298       0.019
ma.L2          0.1782      0.160      1.115      0.265      -0.135       0.492
ma.L3         -0.2425      0.098     -2.487      0.013      -0.434      -0.051
ar.S.L6       -0.4715      0.554     -0.851      0.395      -1.557       0.614
ar.S.L12       0.5741      0.237      2.424      0.015       0.110       1.038
ar.S.L18       0.0006      0.522      0.001      0.999      -1.023       1.025
ma.S.L6        0.4387      0.566      0.775      0.438      -0.671       1.548
ma.S.L12      -0.0999      0.252     -0.397      0.692      -0.593       0.394
ma.S.L18       0.2452      0.319      0.768      0.443      -0.381       0.871
sigma2      3088.8858    483.972      6.382      0.000    2140.319    4037.453
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                22.66
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               7.91   Skew:                             0.27
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.17
===================================================================================

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

Fig 16.a: Summary of SARIMA

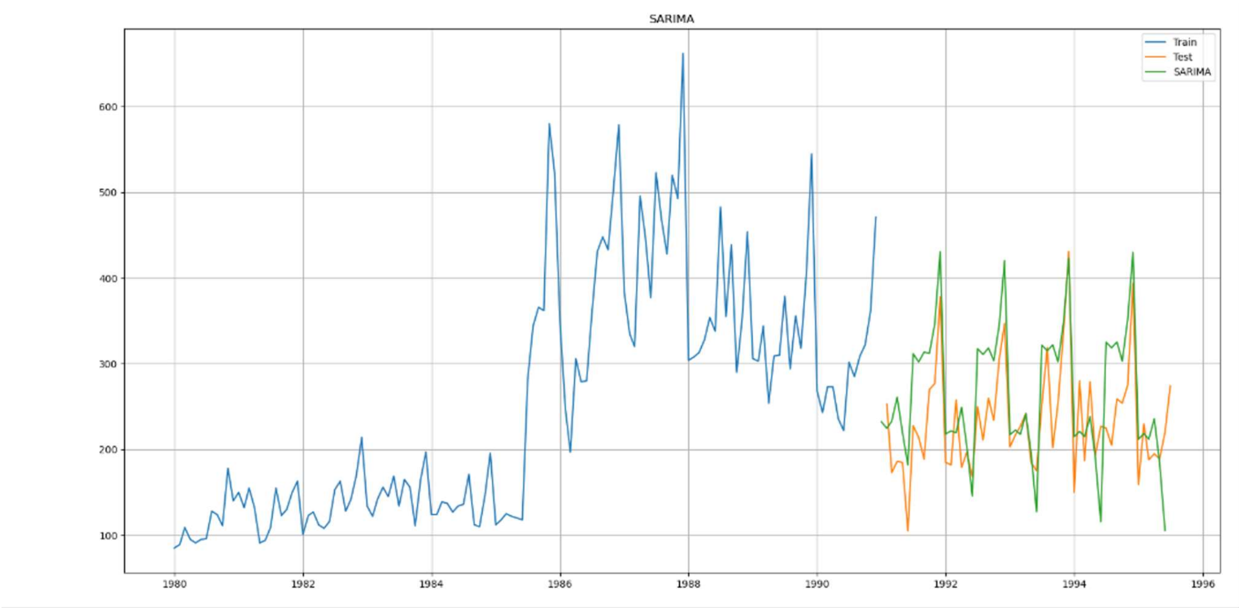

Fig 16.b: Graph of SARIMA

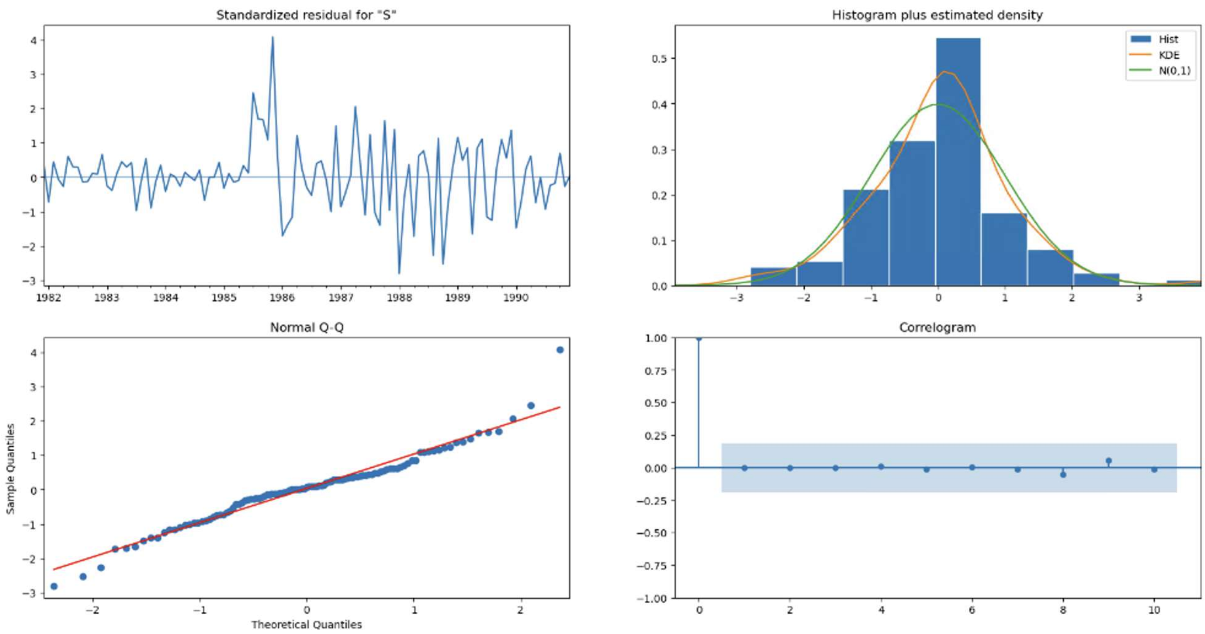

Fig 16.b: Diagnostic of SARIMA

The accuracy of the model SARIMA has the lowest RMSE and MAPE

	RMSE	MAPE
ARIMA(3,1,3)	142.136266	64.286625
SARIMA	94.068822	34.388548

1.7 Table with all the model along with their corresponding parameter and respective RMSE values

- We can see that the best model with least RMSE is 2- point Trailing Moving Average.
- Industry wide exponential smoothing and ARIMA models are more popular. While exponential smoothing technique depends upon the asumption of exponential decrease in weight for past data and ARIMA is employed by transforming a time series to stationary seris and studying the nature of stationary series through ACF and PACF and then accounting auto regressive and moving average effects in a time series.

	Test RMSE	MAPE
Trailing_2	42.489323	14.787482
Trailing_4	55.199732	19.835338
Trailing_6	61.684757	23.269147
SimpleAverage	62.116279	21.327011
Trailing_9	66.901505	24.726947
SARIMA	94.068822	34.388548
ARIMA(3,1,3)	142.136266	64.286625
TES	142.698275	57.907264
SES	195.894561	91.420744
NaiveBayes	244.575066	114.443505
LinearRegression	264.516794	122.140758
DES	286.467010	131.556013

1.8 The most optimum model on the complete data predicting 12 month into the future with confidence intervals

We are going to build the optimum model with AIC -SARIMA as per explanations that are already provided above

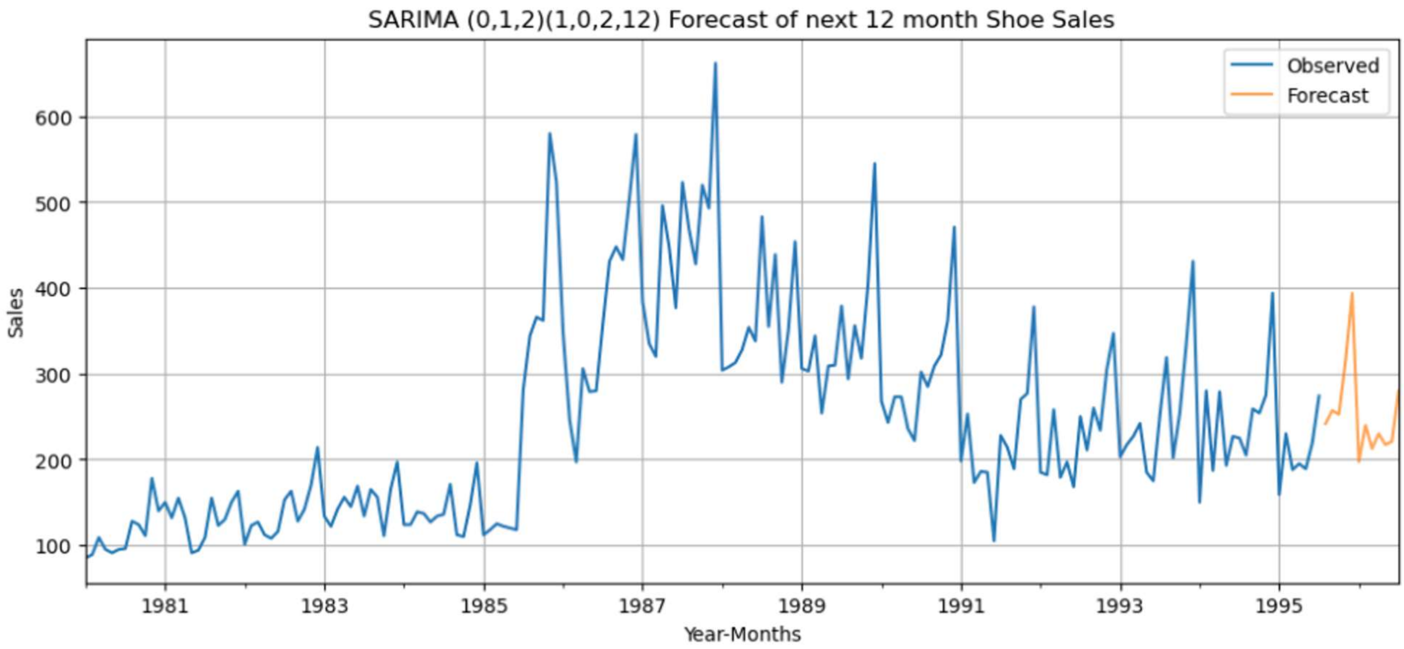

Fig 17: Optimum Model Forecast for next 12 months