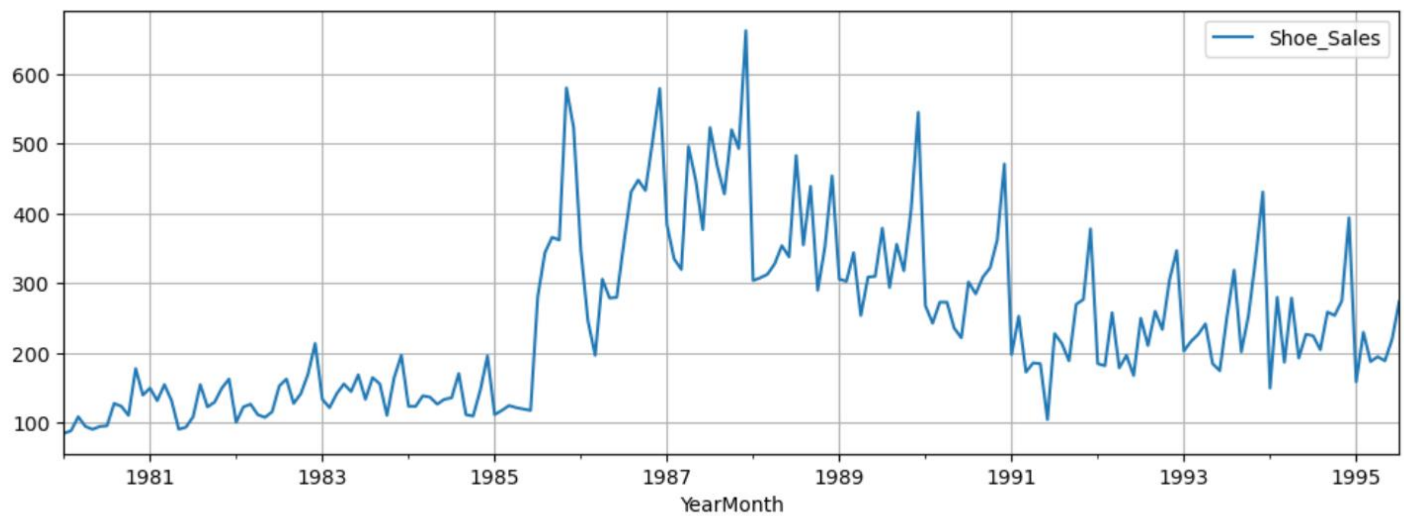
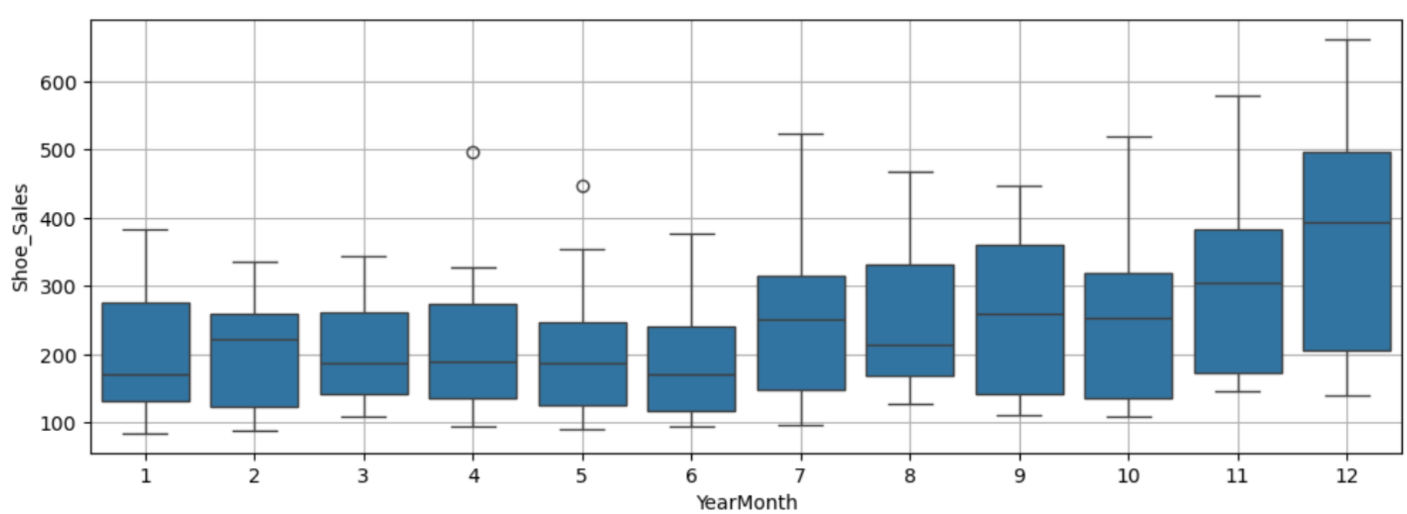
1. Read the data as an appropriate Time Series data and plot the data.

The shoesales has impact in trend of 1985 to 1991 and huge seasonality(peaks) in after 1985 the softdrink data.

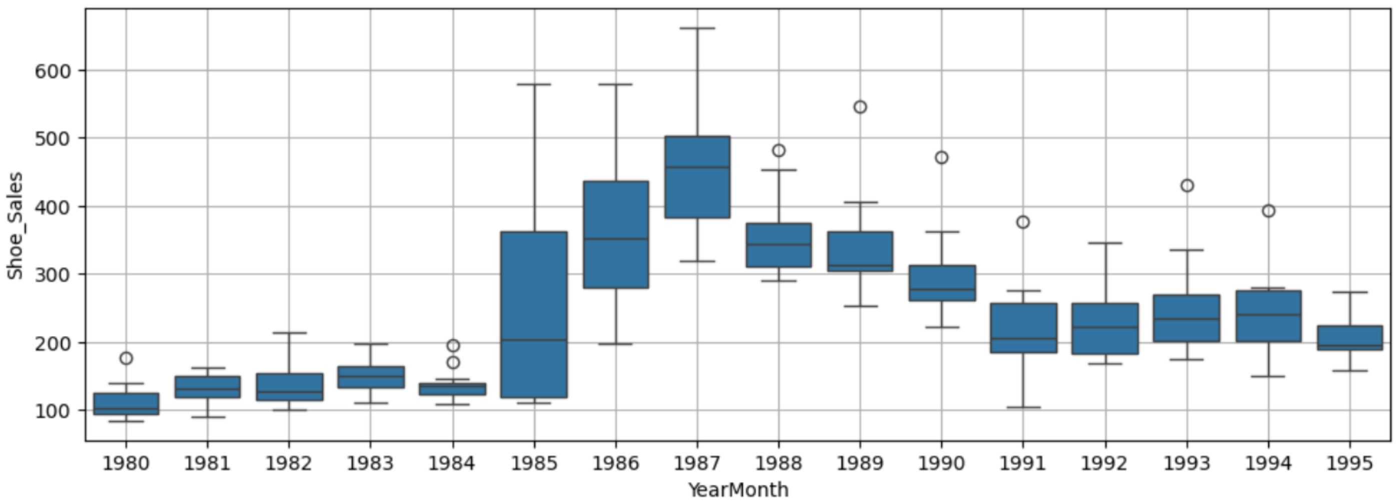


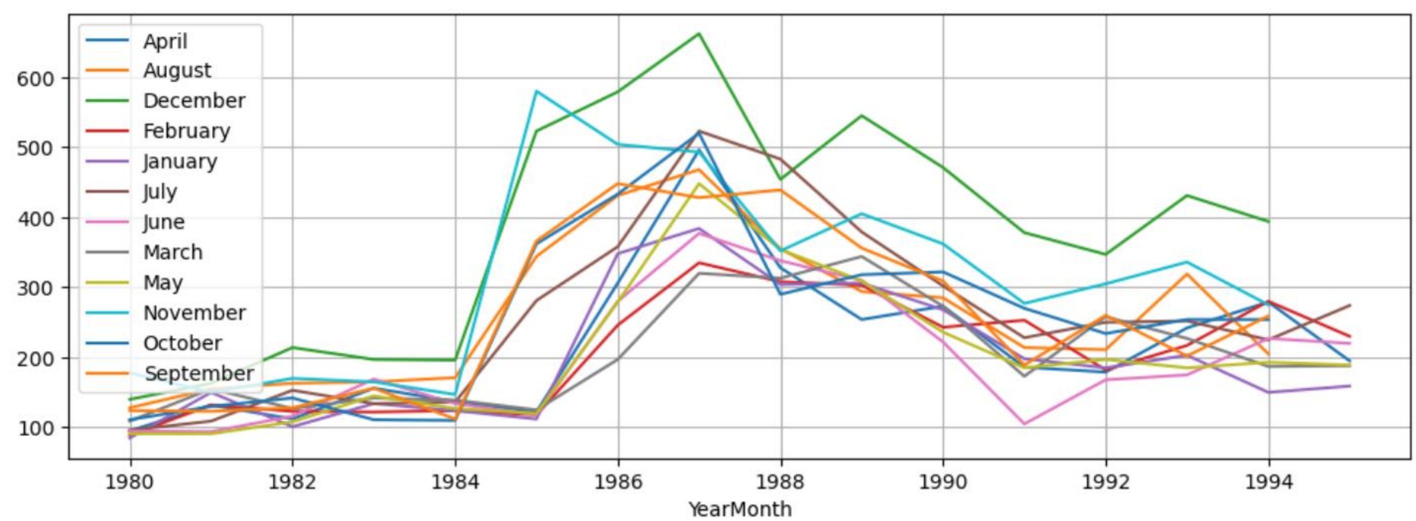
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

Usually we don’t need to treat outliers and the December month was the production happened across years. The boxplot is to understand the overview of production across months. The 2nd highest sales was November.



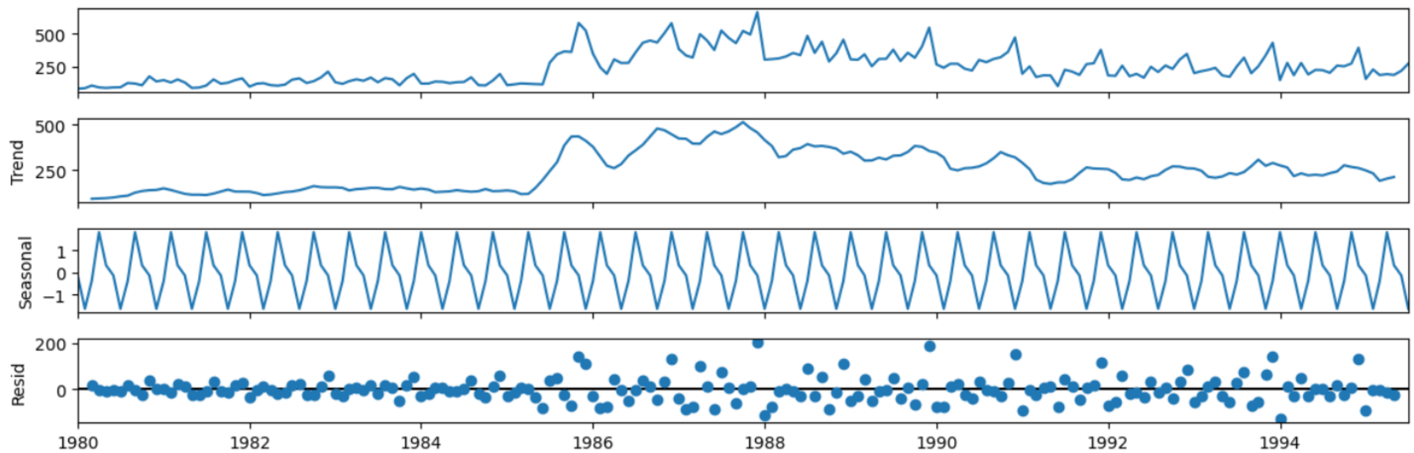
Across years:



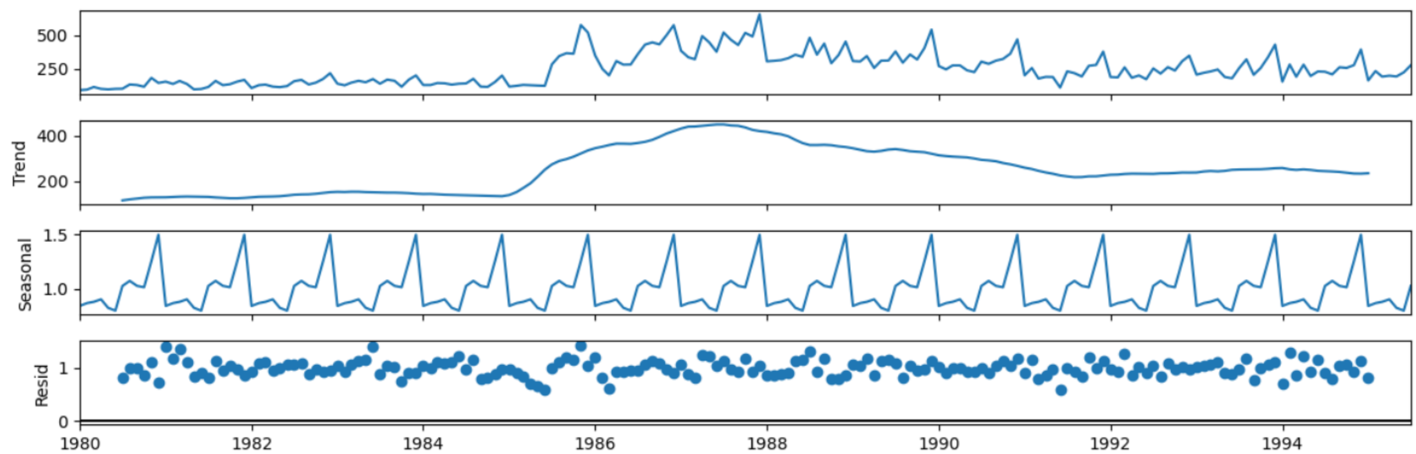


The december was highest sales and November is the 2nd highest sales in above flow chart.

Additive Series



Multiplicative Series:



In above 2 cases, multiplicative series is gives a pattern and we can take it as a consideration

We can see the trend, season and residual in below images.

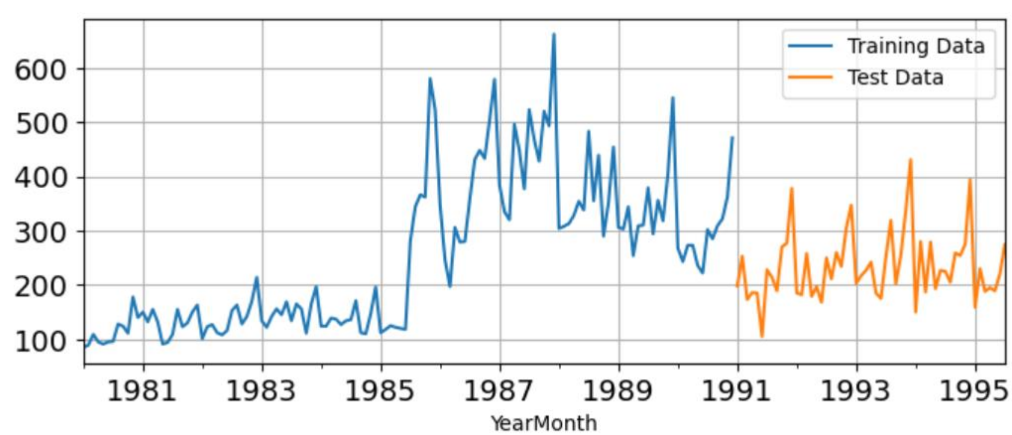
```
Trend
 YearMonth
1980-01-01       NaN
1980-02-01       NaN
1980-03-01       NaN
1980-04-01       NaN
1980-05-01       NaN
1980-06-01       NaN
1980-07-01    114.46
1980-08-01    118.96
1980-09-01    122.67
1980-10-01    126.12
1980-11-01    127.67
1980-12-01    127.62
Name: trend, dtype: float64
```

```
Seasonality
 YearMonth
1980-01-01    0.84
1980-02-01    0.87
1980-03-01    0.88
1980-04-01    0.90
1980-05-01    0.82
1980-06-01    0.80
1980-07-01    1.03
1980-08-01    1.07
1980-09-01    1.03
1980-10-01    1.01
1980-11-01    1.25
1980-12-01    1.50
Name: seasonal, dtype: float64
```

```
Residual
 YearMonth
1980-01-01     NaN
1980-02-01     NaN
1980-03-01     NaN
1980-04-01     NaN
1980-05-01     NaN
1980-06-01     NaN
1980-07-01    0.82
1980-08-01    1.00
1980-09-01    0.98
1980-10-01    0.87
1980-11-01    1.11
1980-12-01    0.73
Name: resid, dtype: float64
```

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

The train shape is 132 and test is 55 for this dataset and we need atleast one month test data for one year in further evaluation but the test data is starts from 1991. Here Is the below plotted chart we can see the chart.



The test data is started from 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, naive forecast models, simple average models etc. should also be built on the training data and check the performance on the test data using RMSE.

Let's we look into the model building.

Linear regression:

The linear regression of RMSE is **266.276** The RMSE is high indicates that high noise or significant amount of variability of this model. It's impact the decision-making process.

```
First few rows of Training Data
            Shoe_Sales  time
YearMonth
1980-01-01          85     1
1980-02-01          89     2
1980-03-01         109     3
1980-04-01          95     4
1980-05-01          91     5

Last few rows of Training Data
            Shoe_Sales  time
YearMonth
1990-08-01         285   128
1990-09-01         309   129
1990-10-01         322   130
1990-11-01         362   131
1990-12-01         471   132
```

```
First few rows of Test Data
            Shoe_Sales  time
YearMonth
1991-01-01         198   133
1991-02-01         253   134
1991-03-01         173   135
1991-04-01         186   136
1991-05-01         185   137

Last few rows of Test Data
            Shoe_Sales  time
YearMonth
1995-03-01         188   183
1995-04-01         195   184
1995-05-01         189   185
1995-06-01         220   186
1995-07-01         274   187
```

	Test RMSE
RegressionOnTime	266.276472

Simple Exponential Model:

The simple exponential smoothing is the method of current forecast weighted for past observation. It's also known as holt's method.

Formula: **Ft+1 = alpha.Yt+(1-alpha).Ft+1 .**

Ft+1 is the actual observation of next period.

Alpha- Smoothing average(0<alpha<1)
Yt is the actual observation of current period.

Ft is the actual observation of current period.

```
{'smoothing_level': 0.6051903749099211,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 85.0,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

	Test RMSE
RegressionOnTime	266.276472
Alpha=0.995:SimpleExponentialSmoothing	196.425508

RMSE:

The RMSE is 196.426 and high RMSE value is implies that have a larger deviation from actual values. It's less reliable forecast and It's not capture all underlying patterns.

Double exponential Smoothing:

It's a extended method SES to capture trend and seasonality

Parameters:

	name	param	optimized
smoothing_level	alpha	0.603381	True
smoothing_trend	beta	0.000099	True
initial_level	l.0	85.000000	False
initial_trend	b.0	4.000000	False

RMSE value is 311.020

It’s a less reliable and model forecast deviate from actual values. The large Rmse value may not capture the underlying pattern.

	Test RMSE
RegressionOnTime	266.276472
Alpha=0.995:SimpleExponentialSmoothing	196.425508
Alpha=0.99,Beta=0.0001,Gamma=0.005:DoubleExponentialSmoothing	311.020473

Triple Exponential Smoothing:

Holt’s winter method it’s an extension of double exponential smoothing(Holt’s method) it incorporates the seasonality in addition to the level and trend components.

The level captures the underlying pattern and it represents the average value of the seasonality over time.
The Trend represent the rate of change the series over time.
The Seasonal represents the periodic fluctuations.

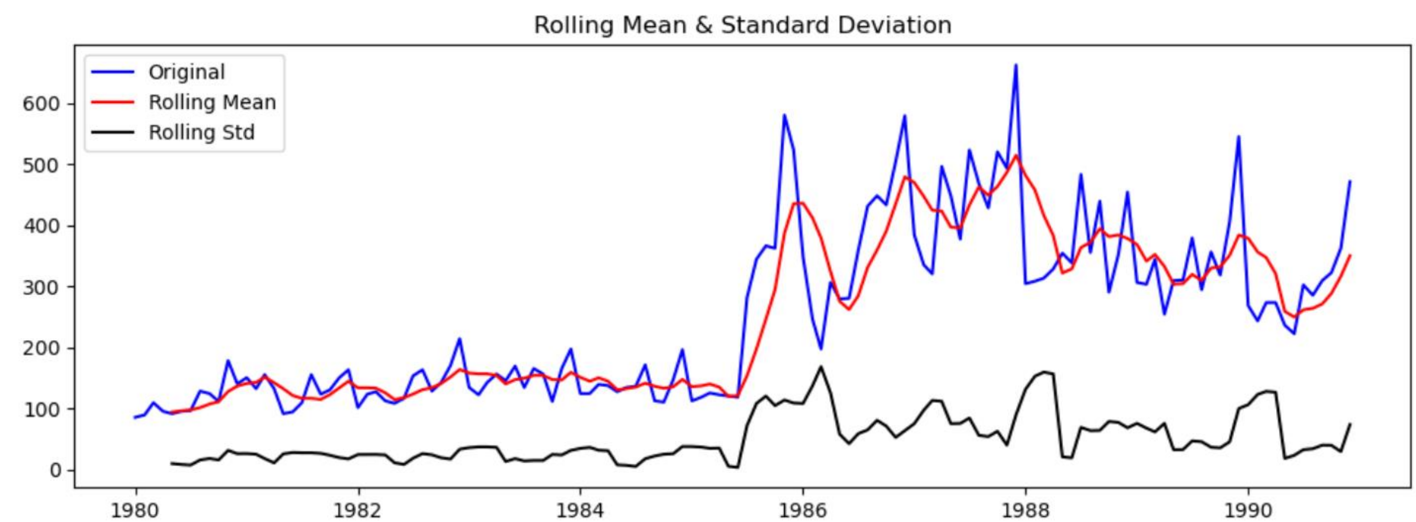
	name	param	optimized
smoothing_level	alpha	0.571129	True
smoothing_trend	beta	0.000148	True
smoothing_seasonal	gamma	0.202947	True
initial_level	l.0	116.355292	True
initial_trend	b.0	0.112199	True
initial_seasons.0	s.0	1.056793	True
initial_seasons.1	s.1	1.011303	True
initial_seasons.2	s.2	1.233747	True
initial_seasons.3	s.3	1.406631	True
initial_seasons.4	s.4	1.321627	True
initial_seasons.5	s.5	1.079369	True
initial_seasons.6	s.6	1.180182	True
initial_seasons.7	s.7	1.501831	True
initial_seasons.8	s.8	1.723691	True
initial_seasons.9	s.9	1.470413	True
initial_seasons.10	s.10	1.754853	True
initial_seasons.11	s.11	1.921014	True

RMSE is 83.734

	Test RMSE
RegressionOnTime	266.276472
Alpha=0.995:SimpleExponentialSmoothing	196.425508
Alpha=0.99,Beta=0.0001,Gamma=0.005:DoubleExponentialSmoothing	311.020473
Alpha=0.99,Beta=0.0001,Gamma=0.005:TripleExponentialSmoothing	83.734048

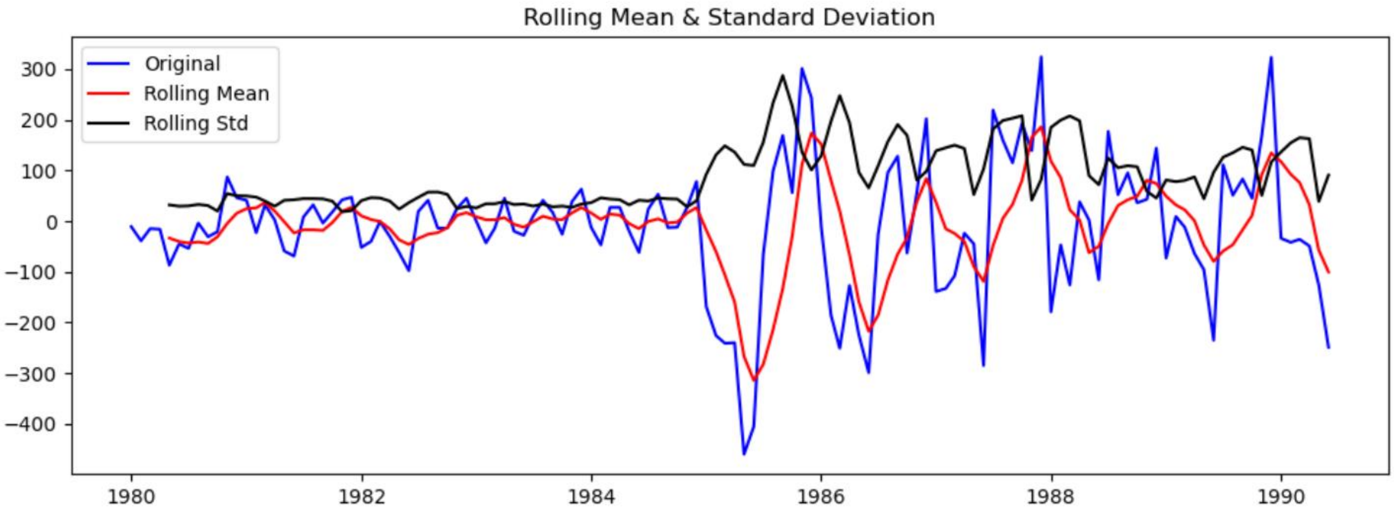
The forecasting reliability is accurate and low RMSE provides greater confidence in model predictor's in (TES) Triple exponential smoothing.

Check the stationarity before differentiation




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.361129
p-value                          0.600763
#Lags Used                      13.000000
Number of Observations Used    118.000000
Critical Value (1%)             -3.487022
Critical Value (5%)             -2.886363
Critical Value (10%)            -2.580009
dtype: float64
```

After Integration:




```
Results of Dickey-Fuller Test:
Test Statistic                  -3.002191
p-value                          0.034690
#Lags Used                       8.000000
Number of Observations Used    117.000000
Critical Value (1%)             -3.487517
Critical Value (5%)             -2.886578
Critical Value (10%)            -2.580124
dtype: float64
```

This test has been done at dickey-fuller test.
It calculates the difference between each observation and the observation 6 time periods ahead(-6).

Null Hypothesis is Non- stationarity and alternate Hypothesis is stationarity

In this time series forecasting, p-value is less than 0.05 is to reject the null hypothesis and go with Stationarity

5.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.

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

7.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

(Note:I gave a Parameters, RMSE values fro 5,6 and 7)

ARIMA:

Auto regressive, integrated and moving averages are used to find the complex pattern in time series data.

p,d,q is denoted as a parameter of ARIMA model
p is the AR
d is the I

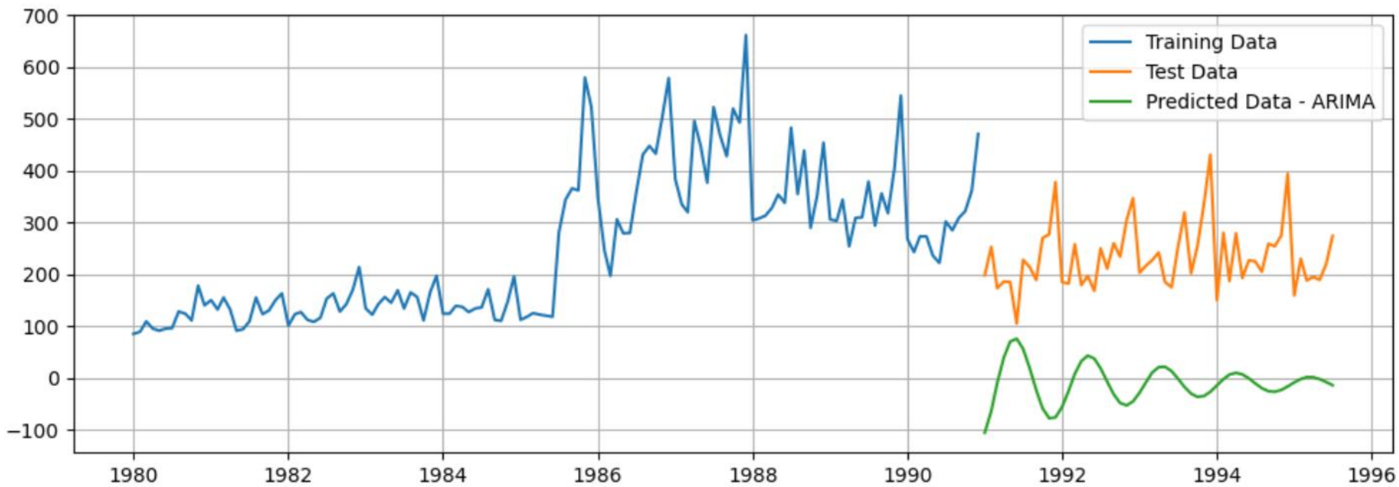
q is the moving average component(MA)

The model performance is calculated by lowest AIC value

AIC value is 1531.218983

	param	AIC
8	(2, 0, 2)	1531.218983
2	(0, 0, 2)	1536.732162
7	(2, 0, 1)	1541.142821
3	(1, 0, 0)	1550.605708
6	(2, 0, 0)	1552.605613
4	(1, 0, 1)	1552.605691
1	(0, 0, 1)	1555.902381
5	(1, 0, 2)	1557.842640
0	(0, 0, 0)	1580.453045

```
                               SARIMAX Results
==============================================================================
Dep. Variable:             Shoe_Sales   No. Observations:                  126
Model:                 ARIMA(2, 0, 2)   Log Likelihood                -759.609
Date:                Wed, 17 Apr 2024   AIC                           1531.219
Time:                        19:36:32   BIC                           1548.237
Sample:                    01-01-1980   HQIC                          1538.133
                         - 06-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -11.0157     12.908     -0.853      0.393     -36.315      14.283
ar.L1          1.6350      0.049     33.120      0.000       1.538       1.732
ar.L2         -0.9188      0.041    -22.398      0.000      -0.999      -0.838
ma.L1         -1.3827      0.055    -25.040      0.000      -1.491      -1.275
ma.L2          0.7817      0.054     14.442      0.000       0.676       0.888
sigma2      9978.4047   1070.617      9.320      0.000    7880.034    1.21e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.06   Jarque-Bera (JB):                10.05
Prob(Q):                              0.80   Prob(JB):                         0.01
Heteroskedasticity (H):              11.01   Skew:                            -0.34
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.20
===================================================================================
```



The RMSE is 257.891705 and let's we look into the next model
This is not a good model because it's predicted far way from the test data.

SARIMA:

Seasonal AR and MA components into a single model to capture trend and seasonality.

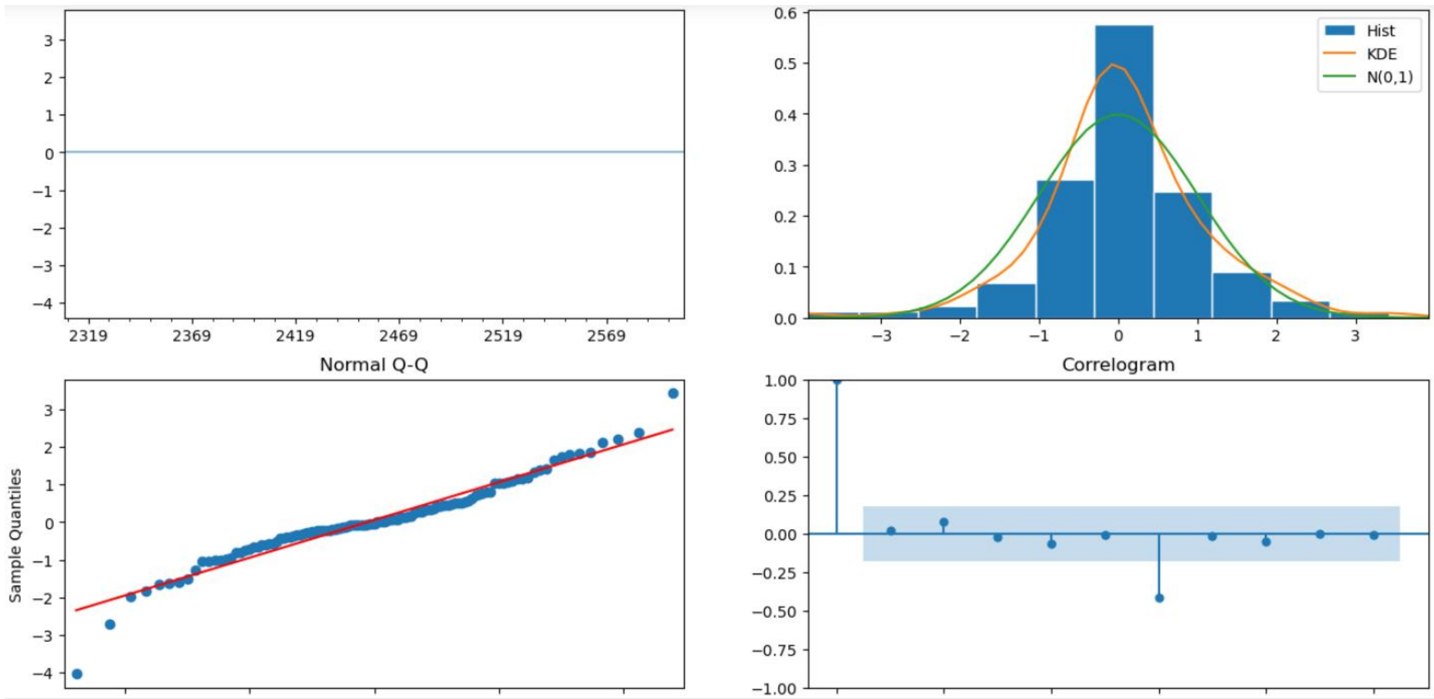
AIC Value is 1372.889434

	param	seasonal	AIC
74	(2, 0, 2)	(0, 0, 2, 5)	1372.889434
53	(1, 0, 2)	(2, 0, 2, 5)	1374.005737
47	(1, 0, 2)	(0, 0, 2, 5)	1374.311948
77	(2, 0, 2)	(1, 0, 2, 5)	1374.808643
50	(1, 0, 2)	(1, 0, 2, 5)	1376.259389
...	...	...	...
10	(0, 0, 1)	(0, 0, 1, 5)	1591.322117
18	(0, 0, 2)	(0, 0, 0, 5)	1632.020706
1	(0, 0, 0)	(0, 0, 1, 5)	1698.864603
9	(0, 0, 1)	(0, 0, 0, 5)	1711.881319
0	(0, 0, 0)	(0, 0, 0, 5)	1856.676269

81 rows × 3 columns

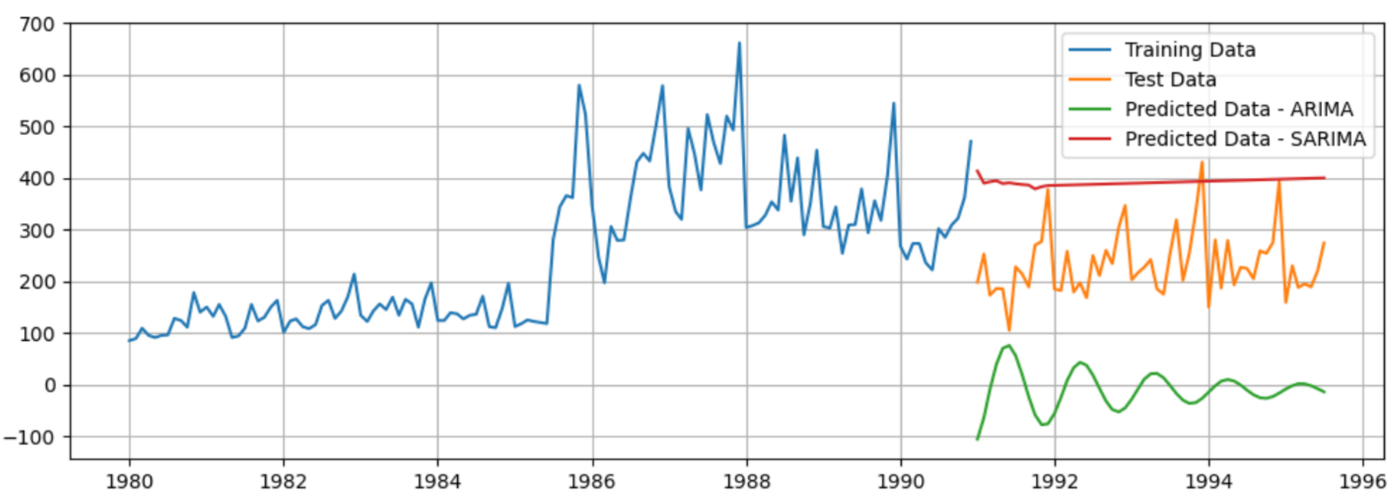
Report :

```
                                      SARIMAX Results                                      
===========================================================================================
Dep. Variable:                        Shoe_Sales   No. Observations:                   132
Model:             SARIMAX(1, 0, 2)x(0, 0, 2, 5)   Log Likelihood                 -681.156
Date:                           Thu, 18 Apr 2024   AIC                            1374.312
Time:                                   09:55:11   BIC                            1390.987
Sample:                               01-01-1980   HQIC                           1381.083
                                    - 12-01-1990                                         
Covariance Type:                             opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.0009      0.010    103.024      0.000       0.982       1.020
ma.L1         -0.3722      0.087     -4.288      0.000      -0.542      -0.202
ma.L2         -0.1889      0.086     -2.191      0.028      -0.358      -0.020
ma.S.L5        0.0045      0.110      0.041      0.967      -0.210       0.219
ma.S.L10      -0.0743      0.104     -0.715      0.474      -0.278       0.129
sigma2      5425.7814    525.554     10.324      0.000    4395.714    6455.849
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                33.63
Prob(Q):                              0.85   Prob(JB):                         0.00
Heteroskedasticity (H):              12.06   Skew:                            -0.19
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.58
===================================================================================
```



The above 4th diagram represents the error are dependent, homoskedasticity and normally distributed curve

	Test RMSE
ARIMA(2, 0, 2)	257.891705
SARIMA(1, 0, 2)(0, 0, 2, 5)	170.325838



The above chart represents this is also not a good model because it’s straight line occures in predicted sarima.

SARIMAX:

We can include the exogenous variable. It can include the other relevant variables that may affect the behaviour of the times series.

AIC Values:

	param	seasonal	AIC
0	(0, 0, 0)	(0, 0, 0, 5)	-2771.624576
27	(1, 0, 0)	(0, 0, 0, 5)	-2769.624576
9	(0, 0, 1)	(0, 0, 0, 5)	-2748.436602
36	(1, 0, 1)	(0, 0, 0, 5)	-2746.436602
54	(2, 0, 0)	(0, 0, 0, 5)	-2746.436602
...	...	...	...
74	(2, 0, 2)	(0, 0, 2, 5)	-2505.368890
26	(0, 0, 2)	(2, 0, 2, 5)	-2505.368890
53	(1, 0, 2)	(2, 0, 2, 5)	-2503.368890
77	(2, 0, 2)	(1, 0, 2, 5)	-2503.368890
80	(2, 0, 2)	(2, 0, 2, 5)	-2501.368890

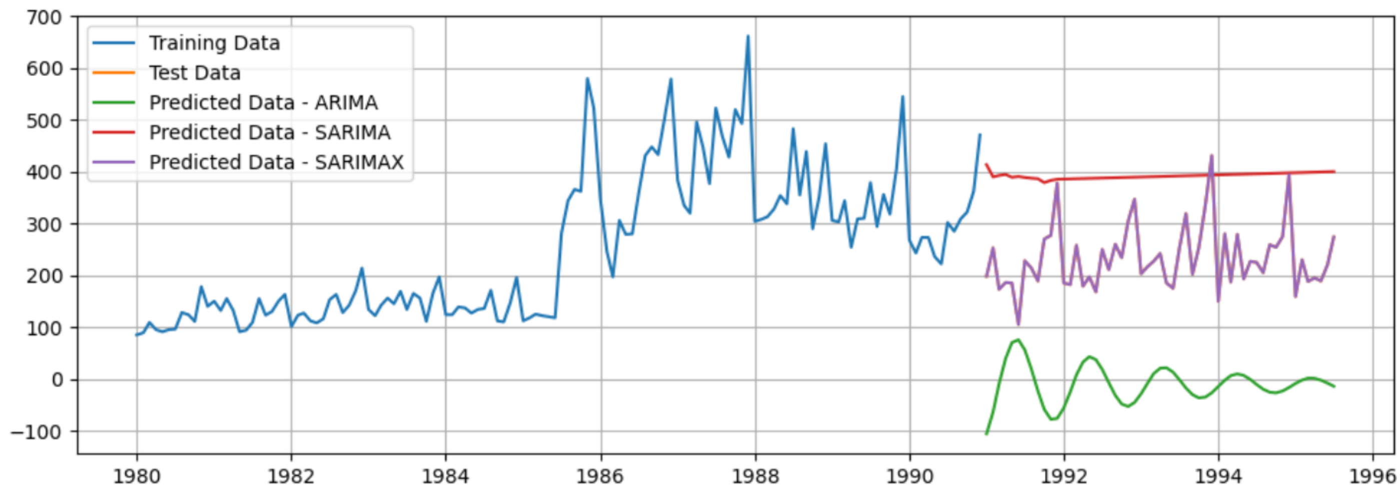
81 rows × 3 columns

-2771.624576 is the lowest AIC value

Report:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:             Shoe_Sales   No. Observations:                  132
Model:                        SARIMAX   Log Likelihood                1387.812
Date:                Thu, 18 Apr 2024   AIC                          -2771.625
Time:                        10:06:17   BIC                          -2765.874
Sample:                    01-01-1980   HQIC                         -2769.288
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Shoe_Sales     1.0000         -0       -inf      0.000       1.000       1.000
sigma2          1e-10   1.73e-10      0.578      0.564   -2.39e-10    4.39e-10
===================================================================================
Ljung-Box (L1) (Q):                   nan   Jarque-Bera (JB):                   nan
Prob(Q):                              nan   Prob(JB):                           nan
Heteroskedasticity (H):               nan   Skew:                               nan
Prob(H) (two-sided):                  nan   Kurtosis:                           nan
===================================================================================
```

	Test RMSE
ARIMA(2, 0, 2)	257.891705
SARIMA(1, 0, 2)(0, 0, 2, 5)	170.325838
SARIMAX(0, 0, 0)(0, 0, 0, 5)	0.000000



For this parameter, Most of us are nan because of negative AIC values

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

Based on the above analysis,we can take the SARIMAX model.

SARIMAX outperforms other SARIMA and ARIMA models in terms of forecast accuracy, as evidenced by lower error metrics such as Root Mean Squared Error (RMSE).

The inclusion of exogenous variables in SARIMAX allows the model to capture additional information not present in the time series.

It is recommended to prioritize its use for forecasting and decision-making purposes.

So SARIMAX is the appropriate model for the above charts.