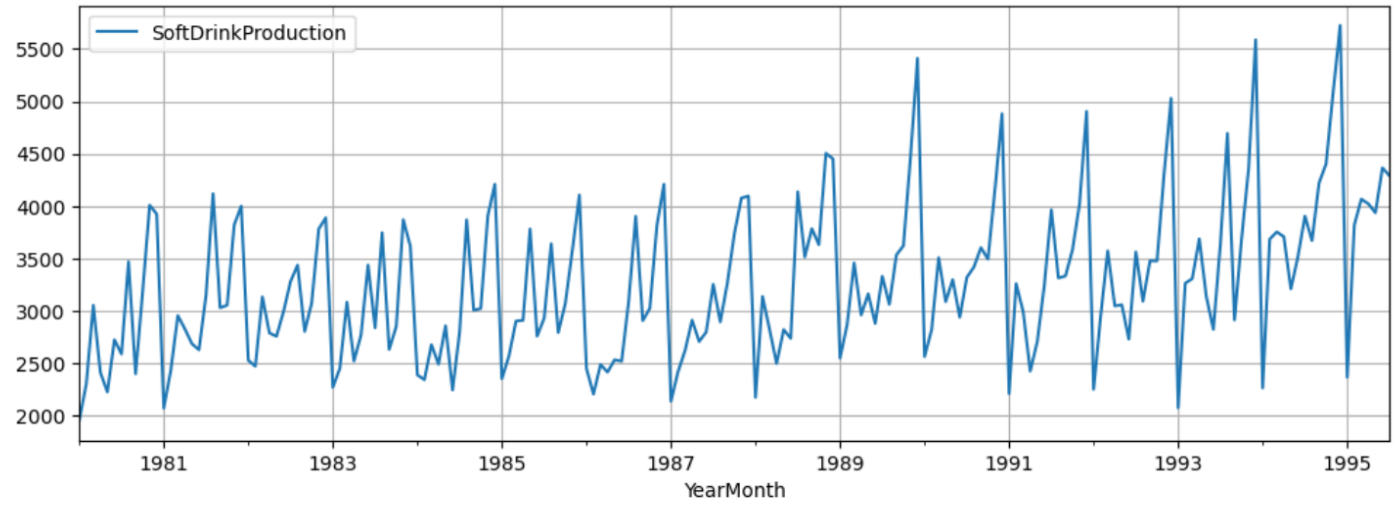
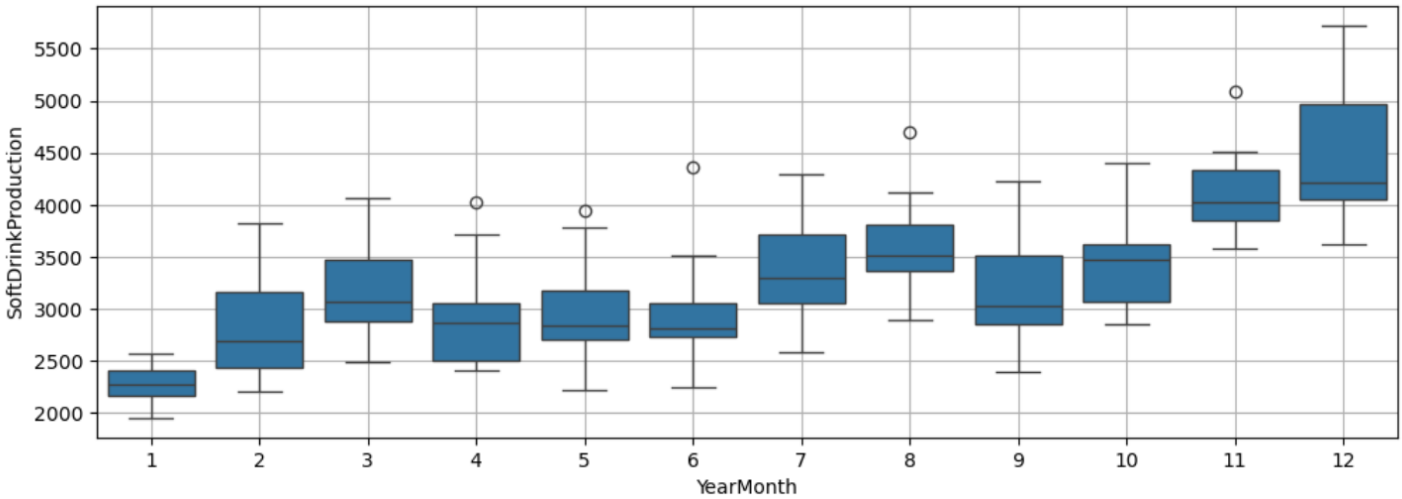
1. Read the data as an appropriate Time Series data and plot the data.

The softdrink production has no impact in trend and slight seasonality(peaks) in the softdrink data.



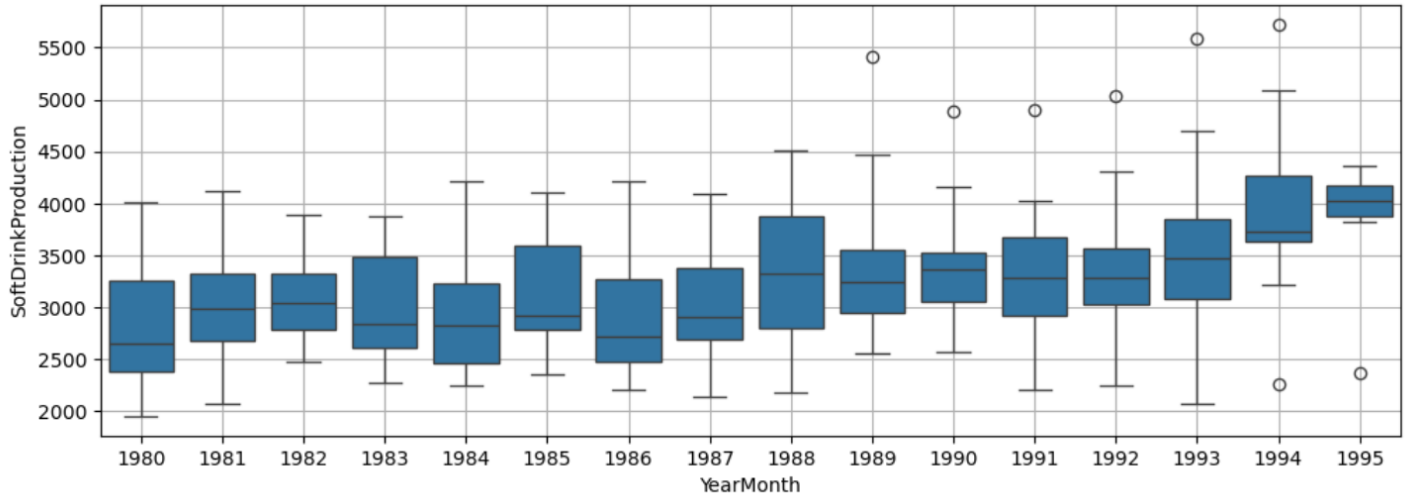
2. Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition.

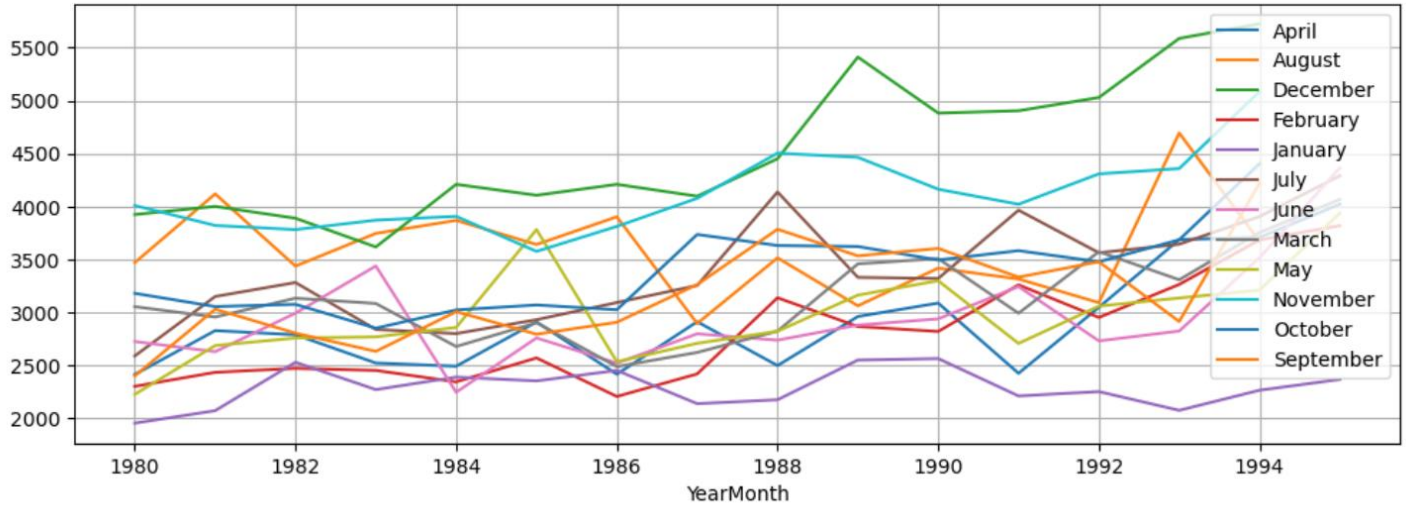
Usually we don't need to treat outliers and the December month was the production happened across years. The boxplot is to understand the overview of production across months. The 2nd highest production was November.



Across Years:

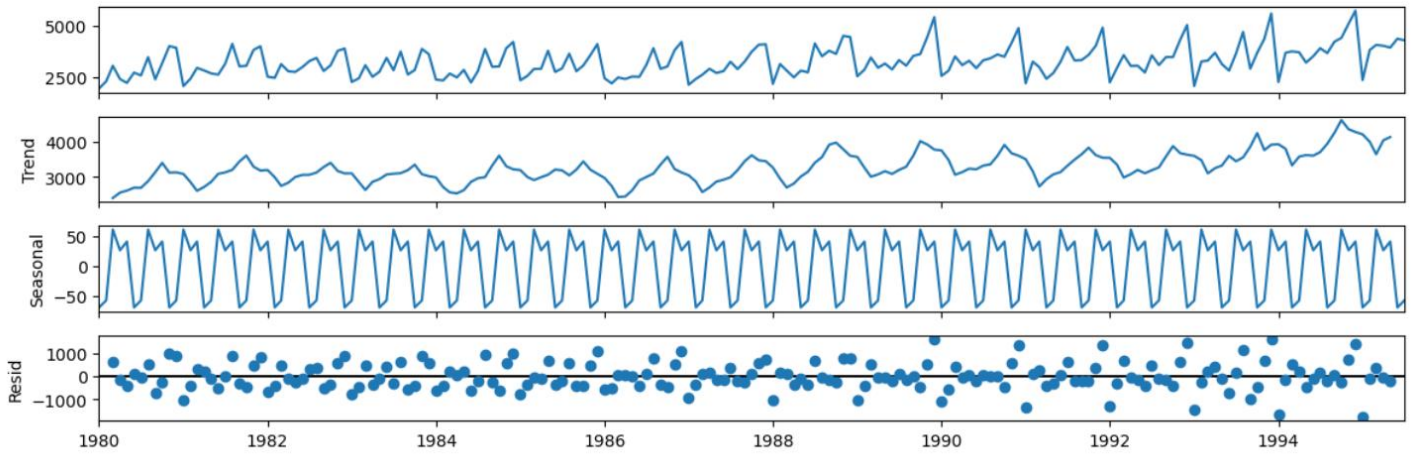
In years, the maximum production was 1994 and 1995.The lowest production was happened at 1980 and totally



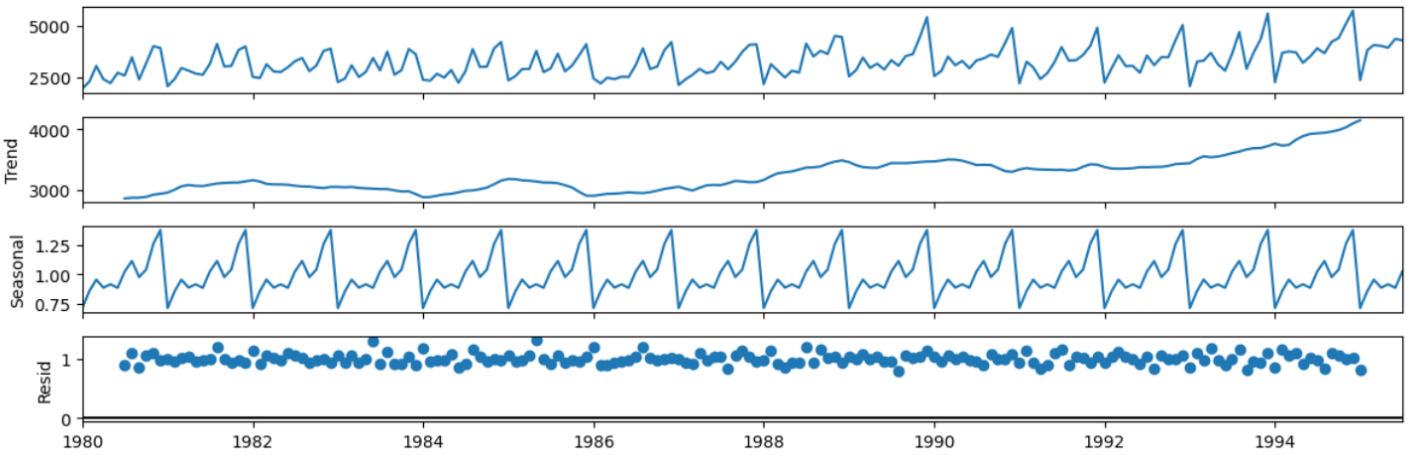


We can clearly say that the december was highest production and November is the 2nd highest production in above flow chart.

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.

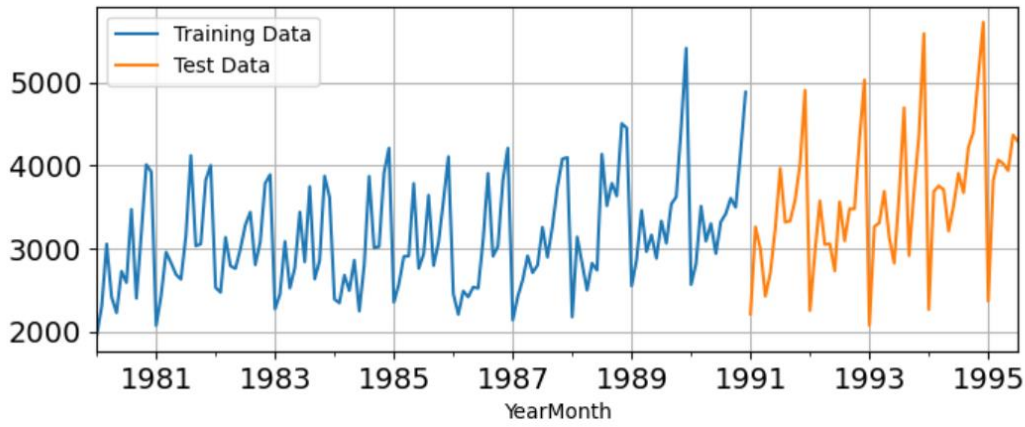
```
Trend
 YearMonth
1980-01-01        NaN
1980-02-01        NaN
1980-03-01        NaN
1980-04-01        NaN
1980-05-01        NaN
1980-06-01        NaN
1980-07-01    2858.83
1980-08-01    2869.25
1980-09-01    2870.67
1980-10-01    2883.83
1980-11-01    2920.29
```

```
Seasonality
 YearMonth
1980-01-01    0.71
1980-02-01    0.86
1980-03-01    0.95
1980-04-01    0.89
1980-05-01    0.91
1980-06-01    0.88
1980-07-01    1.02
1980-08-01    1.11
1980-09-01    0.98
1980-10-01    1.04
1980-11-01    1.26
1980-12-01    1.38
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    0.88
1980-08-01    1.09
1980-09-01    0.86
1980-10-01    1.06
1980-11-01    1.09
1980-12-01    0.97
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.



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 **775.808.** 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
            SoftDrinkProduction  time
YearMonth
1980-01-01                 1954     1
1980-02-01                 2302     2
1980-03-01                 3054     3
1980-04-01                 2414     4
1980-05-01                 2226     5

Last few rows of Training Data
            SoftDrinkProduction  time
YearMonth
1990-08-01                 3418   128
1990-09-01                 3604   129
1990-10-01                 3495   130
1990-11-01                 4163   131
1990-12-01                 4882   132


First few rows of Test Data
            SoftDrinkProduction  time
YearMonth
1991-01-01                 2211   133
1991-02-01                 3260   134
1991-03-01                 2992   135
1991-04-01                 2425   136
1991-05-01                 2707   137

Last few rows of Test Data
            SoftDrinkProduction  time
YearMonth
1995-03-01                 4067   183
1995-04-01                 4022   184
1995-05-01                 3937   185
1995-06-01                 4365   186
1995-07-01                 4290   187
```

	Test RMSE
RegressionOnTime	775.80781

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': 0.15727011341598435,
 'smoothing_trend': nan,
 'smoothing_seasonal': nan,
 'damping_trend': nan,
 'initial_level': 1954.0,
 'initial_trend': nan,
 'initial_seasons': array([], dtype=float64),
 'use_boxcox': False,
 'lamda': None,
 'remove_bias': False}
```

RMSE:

The RMSE is 819.401 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	0.434993	True
smoothing_trend	beta	0.081250	True
initial_level	l.0	1954.000000	False
initial_trend	b.0	348.000000	False

RMSE value is 2820.480

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	775.807810
Alpha=0.995:SimpleExponentialSmoothing	819.401215
Alpha=0.99,Beta=0.0001,Gamma=0.005:DoubleExponentialSmoothing	2820.480441

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.

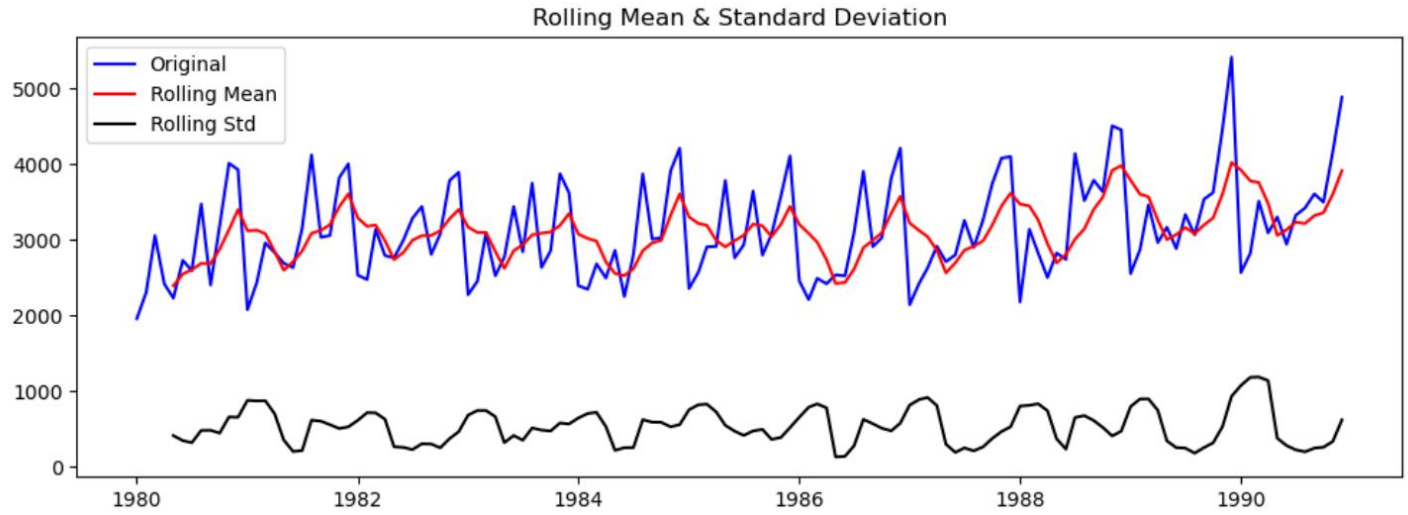
	name	param	optimized
smoothing_level	alpha	0.111284	True
smoothing_trend	beta	0.049473	True
smoothing_seasonal	gamma	0.230372	True
initial_level	l.0	2803.016819	True
initial_trend	b.0	10.486286	True
initial_seasons.0	s.0	0.802840	True
initial_seasons.1	s.1	0.869687	True
initial_seasons.2	s.2	1.082660	True
initial_seasons.3	s.3	0.939548	True
initial_seasons.4	s.4	0.963319	True
initial_seasons.5	s.5	0.988543	True
initial_seasons.6	s.6	1.065419	True
initial_seasons.7	s.7	1.285044	True
initial_seasons.8	s.8	1.008371	True
initial_seasons.9	s.9	1.092992	True
initial_seasons.10	s.10	1.364606	True
initial_seasons.11	s.11	1.417095	True

RMSE is 447.722581

	Test RMSE
RegressionOnTime	775.807810
Alpha=0.995:SimpleExponentialSmoothing	819.401215
Alpha=0.99,Beta=0.0001,Gamma=0.005:DoubleExponentialSmoothing	2820.480441
Alpha=0.99,Beta=0.0001,Gamma=0.005:TripleExponentialSmoothing	447.722581

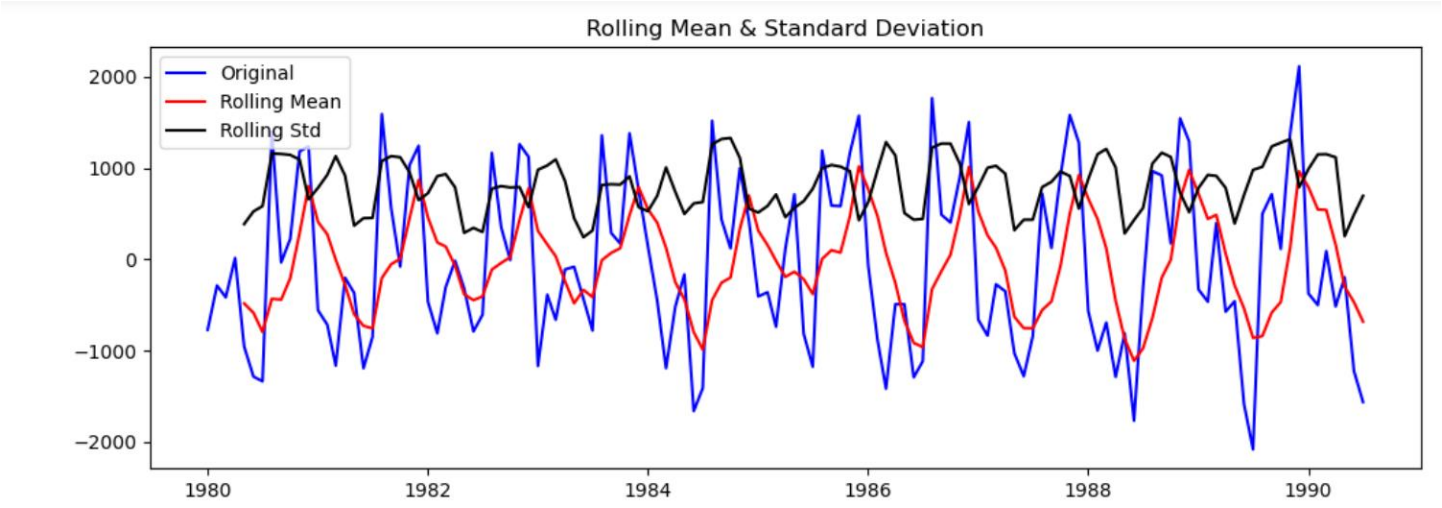
The forecasting reliability is accurate and low RMSE provides greater confidence in model predictor's in (TES) Triple exponential smoothing.

Check the stationarity before differentiation




```
Results of Dickey-Fuller Test:
Test Statistic                  -0.990112
p-value                          0.756854
#Lags Used                      12.000000
Number of Observations Used    119.000000
Critical Value (1%)             -3.486535
Critical Value (5%)             -2.886151
Critical Value (10%)            -2.579896
dtype: float64
```

After Integration:




```
Results of Dickey-Fuller Test:
Test Statistic                  -3.333252
p-value                          0.013456
#Lags Used                      12.000000
Number of Observations Used    114.000000
Critical Value (1%)             -3.489058
Critical Value (5%)             -2.887246
Critical Value (10%)            -2.580481
dtype: float64
```

This test has been done at dickey-fuller test.
It calculates the difference between each observation and the observation 5 time periods ahead(-5).

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 2054.660722

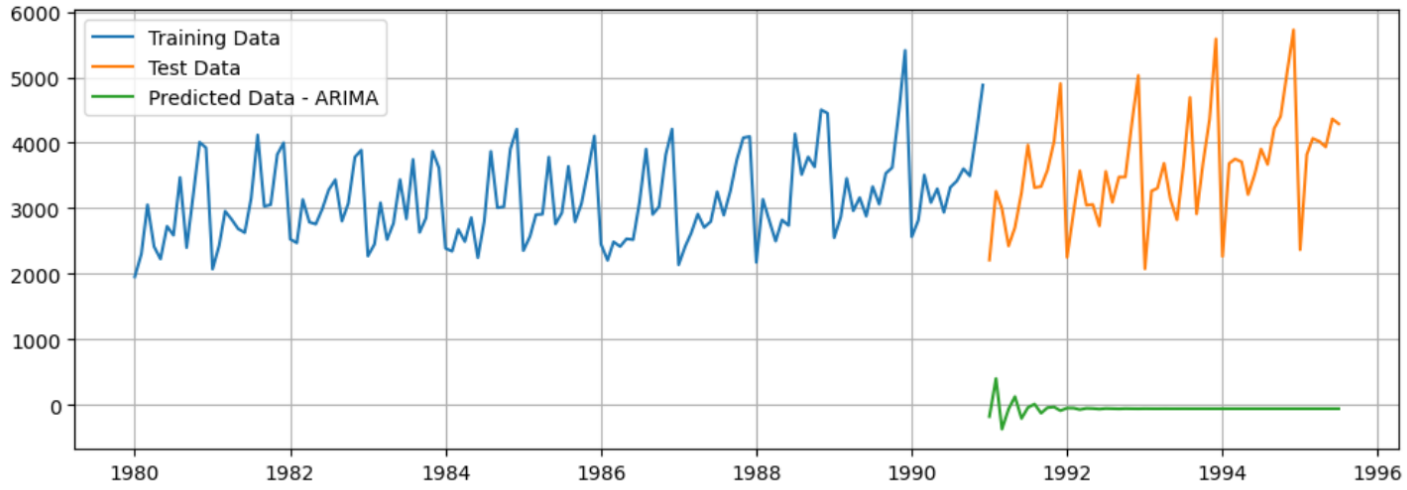
	param	AIC
8	(2, 0, 2)	2054.660722
5	(1, 0, 2)	2063.934513
1	(0, 0, 1)	2074.819171
2	(0, 0, 2)	2076.809093
4	(1, 0, 1)	2076.832074
7	(2, 0, 1)	2078.332056
6	(2, 0, 0)	2079.896704
3	(1, 0, 0)	2082.595906
0	(0, 0, 0)	2099.173011

SARIMAX Results

Dep. Variable:	SoftDrinkProduction	No. Observations:	127
Model:	ARIMA(2, 0, 2)	Log Likelihood	-1021.330
Date:	Thu, 18 Apr 2024	AIC	2054.661
Time:	07:46:27	BIC	2071.726
Sample:	01-01-1980 - 07-01-1990	HQIC	2061.594
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-59.7349	96.231	-0.621	0.535	-248.345	128.875
ar.L1	-0.8265	0.095	-8.691	0.000	-1.013	-0.640
ar.L2	-0.5733	0.108	-5.307	0.000	-0.785	-0.362
ma.L1	1.4237	0.052	27.297	0.000	1.321	1.526
ma.L2	0.9221	0.051	18.122	0.000	0.822	1.022
sigma2	5.542e+05	9.01e+04	6.151	0.000	3.78e+05	7.31e+05

Ljung-Box (L1) (Q):	0.75	Jarque-Bera (JB):	3.71
Prob(Q):	0.39	Prob(JB):	0.16
Heteroskedasticity (H):	1.31	Skew:	0.21
Prob(H) (two-sided):	0.39	Kurtosis:	2.28



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.

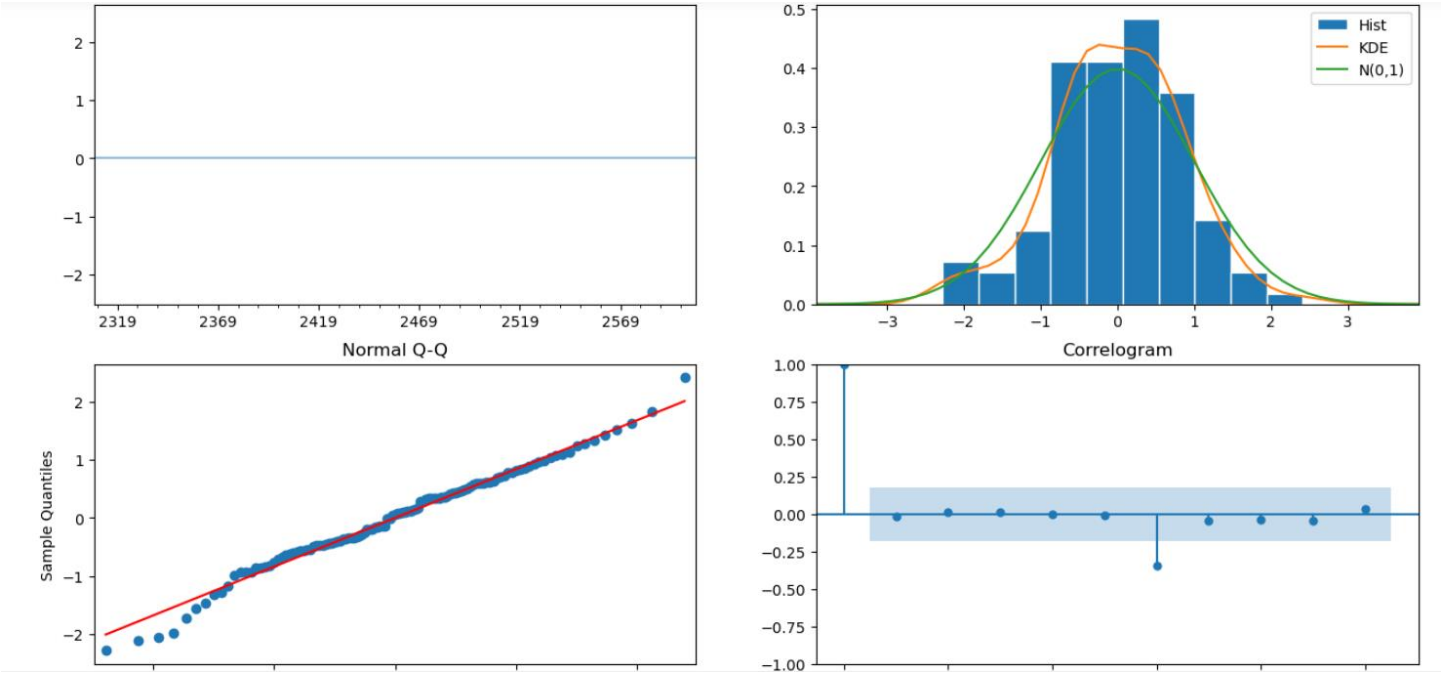
AIC Value is 1865.486341

	param	seasonal	AIC
50	(1, 0, 2)	(1, 0, 2, 5)	1865.486341
47	(1, 0, 2)	(0, 0, 2, 5)	1866.216008
74	(2, 0, 2)	(0, 0, 2, 5)	1867.567417
77	(2, 0, 2)	(1, 0, 2, 5)	1869.279188
53	(1, 0, 2)	(2, 0, 2, 5)	1878.822200
...	...	...	...
10	(0, 0, 1)	(0, 0, 1, 5)	2171.166765
18	(0, 0, 2)	(0, 0, 0, 5)	2235.867577
1	(0, 0, 0)	(0, 0, 1, 5)	2270.027224
9	(0, 0, 1)	(0, 0, 0, 5)	2329.678695
0	(0, 0, 0)	(0, 0, 0, 5)	2488.082615

81 rows × 3 columns

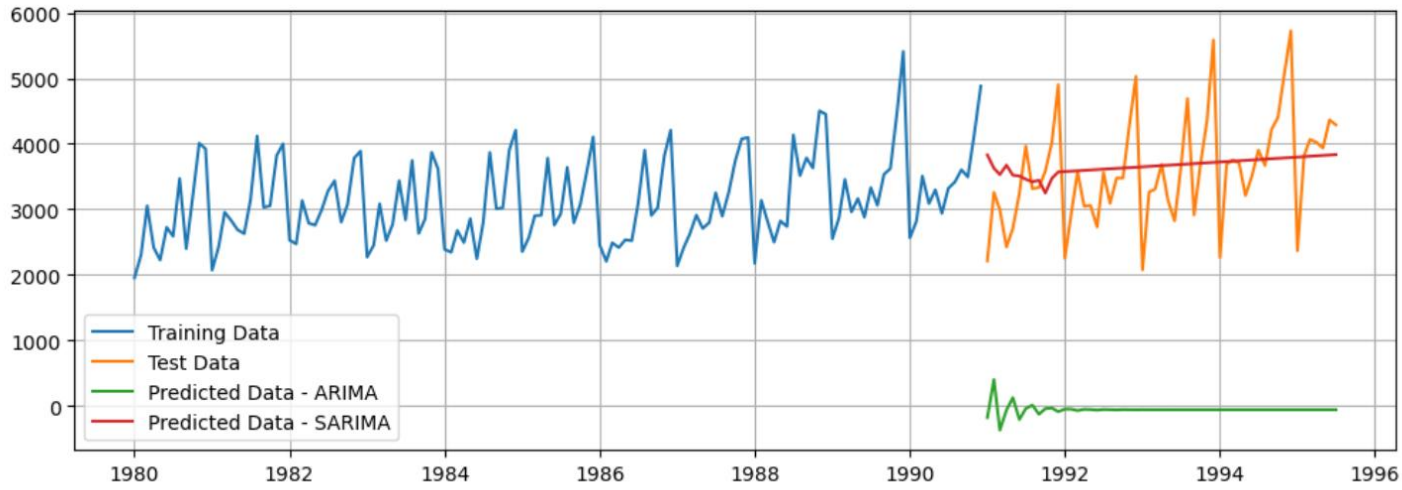
Report :

```
                                SARIMAX Results
==========================================================================================
Dep. Variable:               SoftDrinkProduction   No. Observations:                  132
Model:             SARIMAX(1, 0, 2)x(1, 0, 2, 5)   Log Likelihood                -925.743
Date:                           Thu, 18 Apr 2024   AIC                            1865.486
Time:                                   10:10:01   BIC                            1884.940
Sample:                               01-01-1980   HQIC                           1873.386
                                    - 12-01-1990
Covariance Type:                            opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.0016      0.001    767.465      0.000       0.999       1.004
ma.L1         -0.7138      0.135     -5.292      0.000      -0.978      -0.449
ma.L2         -0.2352      0.153     -1.534      0.125      -0.536       0.065
ar.S.L5       -0.4189      0.316     -1.324      0.185      -1.039       0.201
ma.S.L5        0.3734      0.341      1.094      0.274      -0.296       1.042
ma.S.L10      -0.3426      0.158     -2.168      0.030      -0.652      -0.033
sigma2      4.341e+05   2.54e-07   1.71e+12      0.000    4.34e+05    4.34e+05
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                 1.32
Prob(Q):                              0.84   Prob(JB):                         0.52
Heteroskedasticity (H):               1.96   Skew:                            -0.19
Prob(H) (two-sided):                  0.04   Kurtosis:                         3.34
===================================================================================
```



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

	Test RMSE
ARIMA(2, 0, 2)	3741.600586
SARIMA(1, 0, 2)(1, 0, 2, 5)	790.272035



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
47	(1, 0, 2)	(0, 0, 2, 5)	6872.113320
50	(1, 0, 2)	(1, 0, 2, 5)	6873.443828
74	(2, 0, 2)	(0, 0, 2, 5)	6874.101788
53	(1, 0, 2)	(2, 0, 2, 5)	6875.342430
77	(2, 0, 2)	(1, 0, 2, 5)	6875.432041
...	...	...	...
36	(1, 0, 1)	(0, 0, 0, 5)	7251.074498
54	(2, 0, 0)	(0, 0, 0, 5)	7257.838453
9	(0, 0, 1)	(0, 0, 0, 5)	7266.466892
27	(1, 0, 0)	(0, 0, 0, 5)	7293.182651
0	(0, 0, 0)	(0, 0, 0, 5)	8220.439308

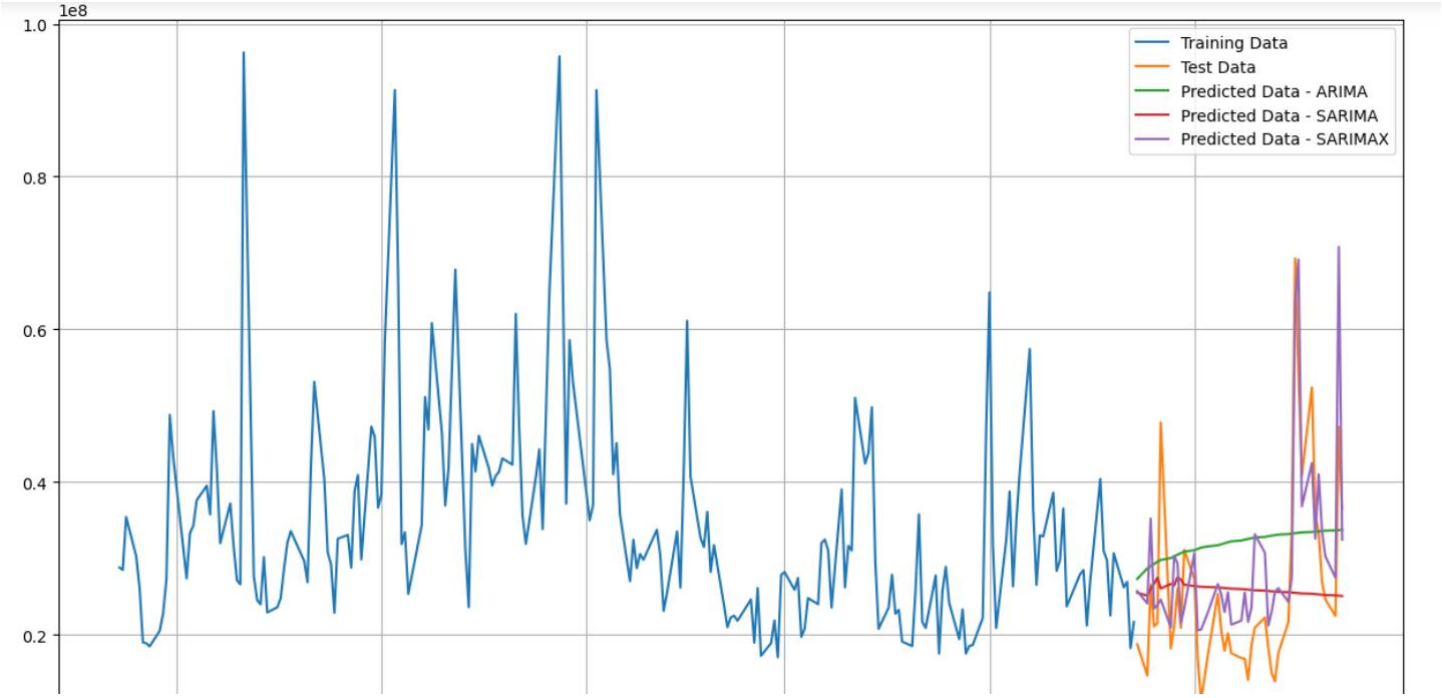
81 rows × 3 columns

6872.11320 is the lowest AIC value

Report:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                              Volume   No. Observations:                  208
Model:             SARIMAX(1, 0, 2)x(0, 0, 2, 5)   Log Likelihood               -3428.057
Date:                           Fri, 10 Feb 2023   AIC                           6872.113
Time:                                   13:59:19   BIC                           6898.297
Sample:                                        0   HQIC                          6882.715
                                             - 208
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
High        4.931e+06   1.12e+04    440.142      0.000    4.91e+06    4.95e+06
Low        -4.861e+06   1.07e+04   -453.761      0.000   -4.88e+06   -4.84e+06
ar.L1          0.9635      0.035     27.540      0.000       0.895       1.032
ma.L1         -0.7040      0.062    -11.424      0.000      -0.825      -0.583
ma.L2         -0.1160      0.077     -1.504      0.132      -0.267       0.035
ma.S.L5       -0.0545      0.119     -0.458      0.647      -0.288       0.179
ma.S.L10      -0.0408      0.082     -0.496      0.620      -0.202       0.121
sigma2      1.116e+14   2.94e-05   3.79e+18      0.000    1.12e+14    1.12e+14
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):              1283.91
Prob(Q):                              0.98   Prob(JB):                         0.00
Heteroskedasticity (H):               0.22   Skew:                             2.40
Prob(H) (two-sided):                  0.00   Kurtosis:                        14.61
===================================================================================
```

	Test RMSE
ARIMA(1, 0, 2)	3663.273022
SARIMA(1, 0, 2)(0, 0, 2)5	3547.744236
SARIMAX(1, 0, 2)(0, 0, 2)5	2892.799281



The model looks good because it's forecasted same like test data.

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 best model among the other models.