Name-Sayan Mondal

Date:02/11/2022

Time Series Forecasting.

(Sales of Soft Drinks)

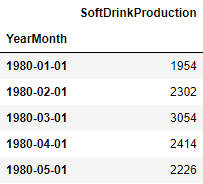


|  |
| --- |
| Contents. Page No |
| 1.Read the data as an appropriate Time Series data and plot the data. 4 |
| 2.Perform appropriate Exploratory Data Analysis to understand the data and also perform decomposition. 4,5,6 |
| 3.Split the data into training and tests. The test data should start in 1991. 6 |
| 4.Build various exponential smoothing models on the training data and evaluate the model using RMSE on the test data. 6,7,8,9,10 |
| 5.Check for the stationarity of the data on which the model is being built. 10,11 |
| 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. 11,12 |
| 7.Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE. 12,13 |
| 8.Build a table with all the models built along with their corresponding parameters and the respective RMSE values on the test data. 13,14 |
| 9.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. 14 |
| 10.Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. 15 |
|  |
|  |
|  |
|  |

|  |
| --- |
| List of Figures Page No. |
| 1. Time Series Plot of Soft Drink Production 4 |
| 1. Additive Decomposition 5 |
| 1. Multiplicative Decomposition 6 |
| 1. Train Test Split 7 |
| 1. Linear Regression 7 |
| 1. Naive Approach 8 |
| 1. Simple Average 8 |
| 1. Simple Exponential Smoothing 9 |
| 1. Double Exponential Smoothing 9 |
| 1. Triple Exponential smoothing(Additive) 10 |
| 1. Triple Exponential Smoothing(Multiplicative) 11 |
| 1. Stationary Time Series Plot 12 |
| 1. PACF Plot 13 |
| 1. ACF Plot 14 |
| 1. Future Plot 15 |

**Sales of Soft Drinks.**

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



The data given is sales of ‘Soft Drink’ for a particular company in 20th century.It is a monthly time series data.

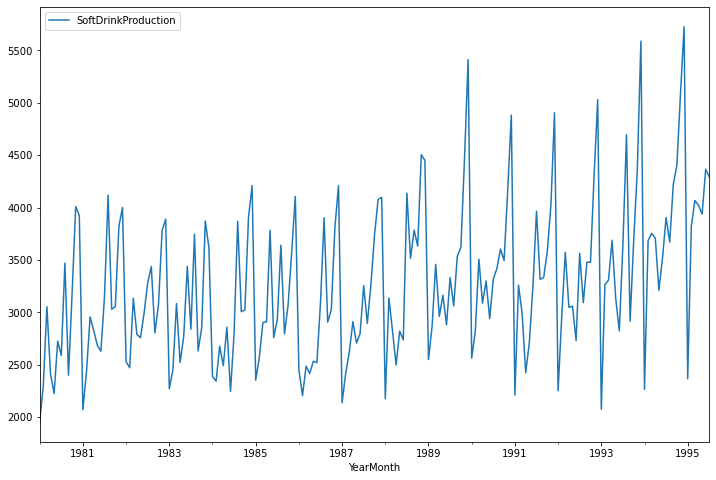
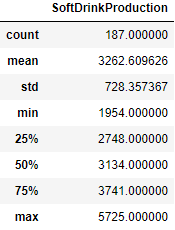


Fig.1

The figure shows time series plot for soft drink sales for a particular company in 20th century.There is seasonality in the data but the trend is not very clear.We will discuss in more detail later.

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

Five point summary of the data:-



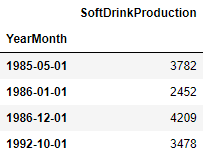
Standard deviation of the data is very high.This may be due to the range(maximum-minimum) of sales being high.There are record for 187 observations.

Missing values:-



There is no missing value in the data.

Duplicate rows:-



There are 4 duplicate rows in the data.This is no anomaly, since different months of corresponding years may have same sales in soft drinks.

Decomposing time series:-

1. Additive Decomposition:-

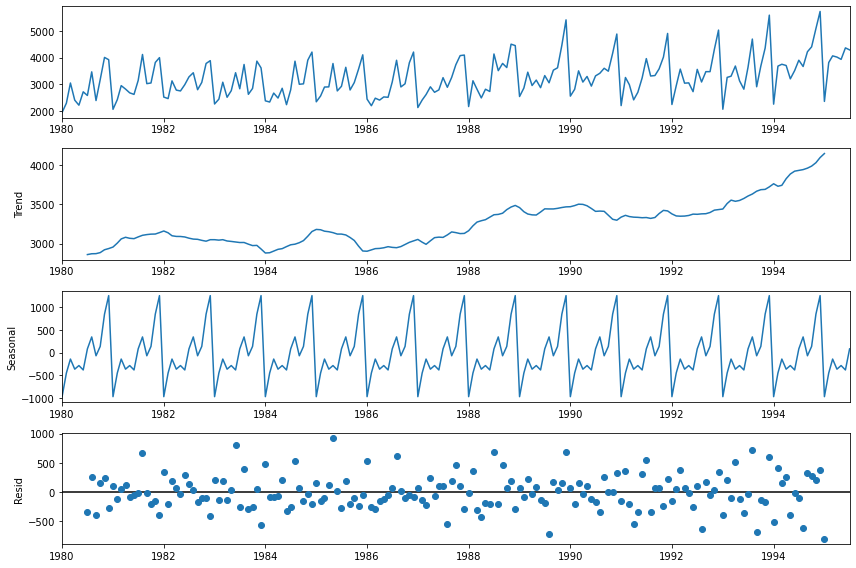


Fig.2

There is a slightly increasing trend in the later years.Seasonality is very much pronounced.The random component has a pattern.So we will further have multiplicative decomposition.

1. Multiplicative Decomposition:-

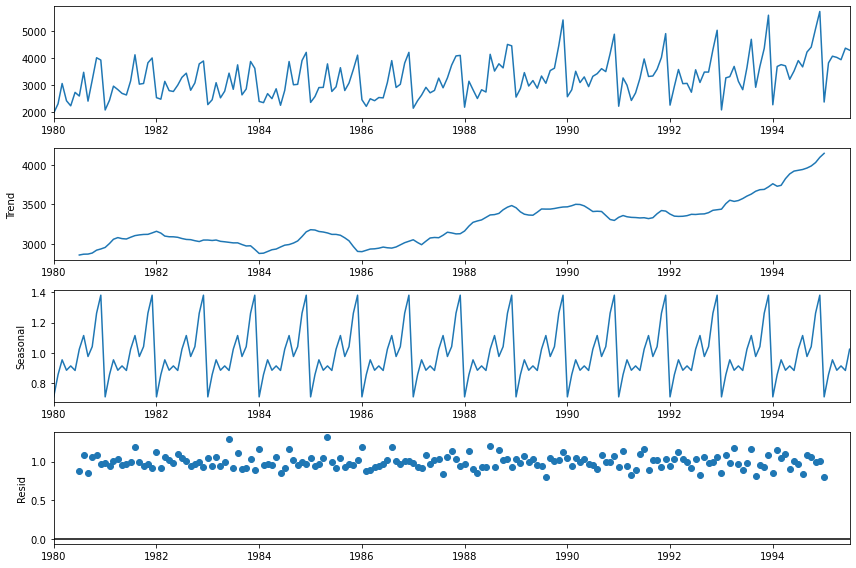


Fig.3

The random component has no pattern as such.Hence the decomposition of the time series is better.

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

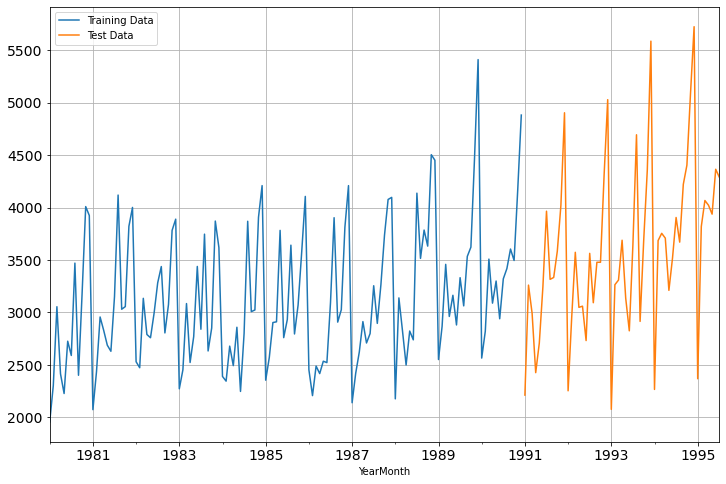


Fig.4

The test data starts at 1991-01-01 and the train data ends at 1990-12-01.Thus there are 132 observations in train data and 55 observations in test data.The data used for testing is almost 30%.

4.Build all the 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 and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

1. Linear Regression.

We take the X-values as consecutive positive integers corresponding to the given time stamps.The y-values are the sales in soft drinks monthly.Then we fit the Linear Regression model on the training set and predict the value on the test set.For evaluation we use RMSE.

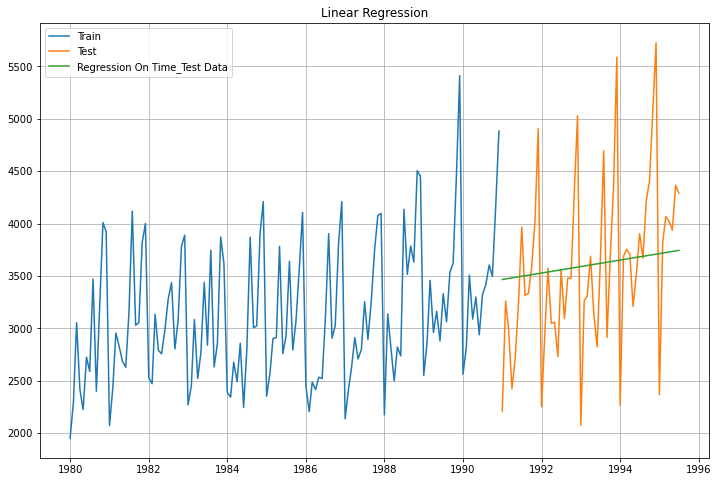


Fig.5

The value of RMSE on test data is 775.81.

1. Naive Approach.

In this model we assume the predicted value at time ‘t’ to be the actual value of the variable at time ‘t-1’.It ignores trend and seasonality in the data.

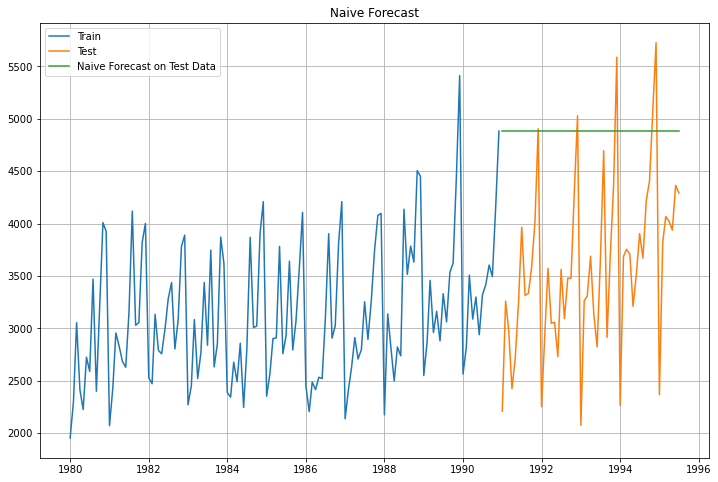


Fig.6

The value of RMSE on test data is 1519.26.It performed poorly in comparison to Linear Regression Model.

1. Simple Average.

In this forecasting model we take the average of the values of the variable to predict the future.

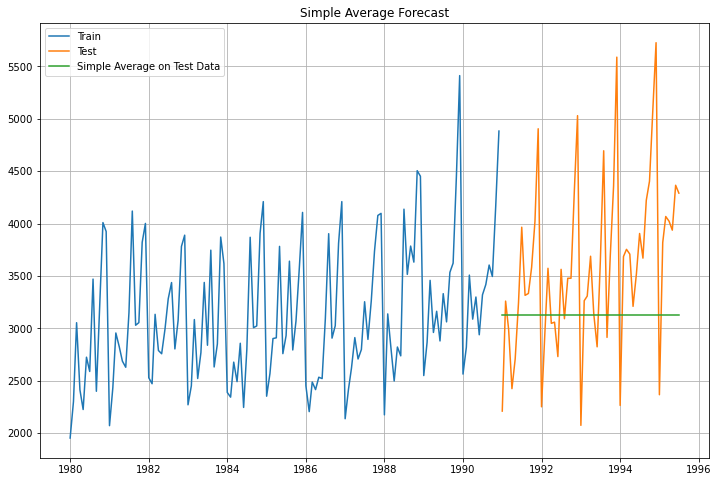


Fig.7

The value of RMSE is 934.35.

1. Simple Exponential Smoothing(SES).

The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing (SES).This method is suitable for forecasting data with no clear trend or seasonal pattern.In Single ES, the forecast at time (t + 1) is given by Winters,1960



Parameter α is called the smoothing constant and its value lies between 0 and 1. Since the model uses only one smoothing constant, it is called Single Exponential Smoothing.

###### Note: Here, there is both trend and seasonality in the data. So, we should have directly gone for the Triple Exponential Smoothing but Simple Exponential Smoothing and the Double Exponential Smoothing models are built over here to get an idea of how the three types of models compare in this case.

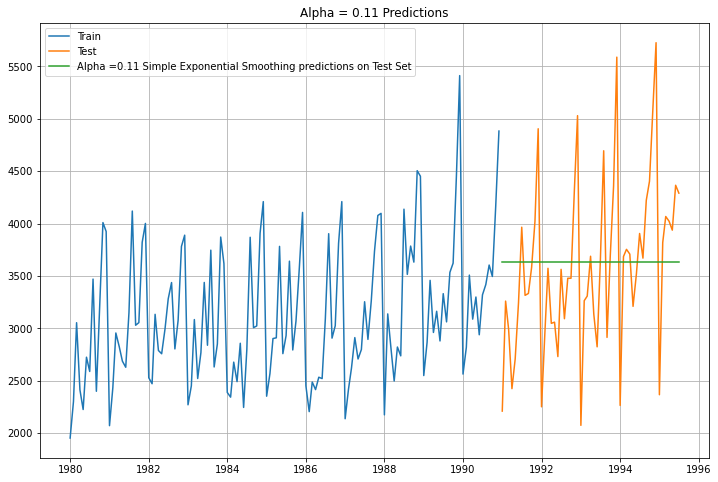


Fig.8

The required value of RMSE is 807.85 .

1. Double Exponential smoothing(DES).

**Double Exponential Smoothing uses two equations to forecast future values of the time series, one for forecasting the short term average value or level and the other for capturing the trend.**



Here, α and β are the smoothing constants for level and trend, respectively,

0 <α < 1 and 0 < β < 1.

The forecast at time t+1 is given by ,



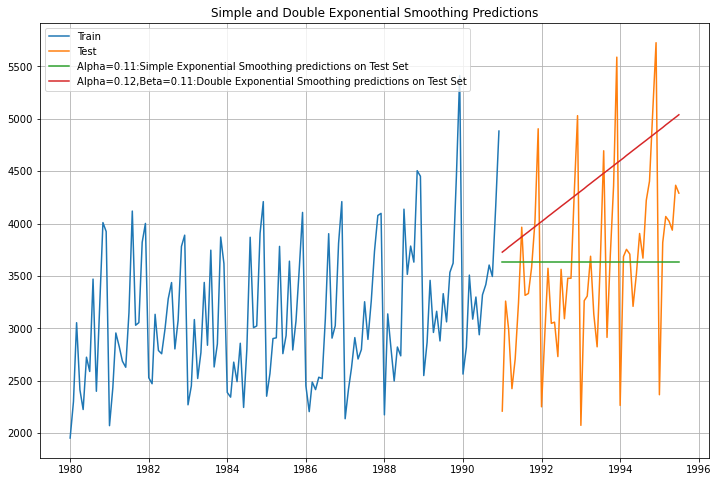


Fig.9

The require value of RMSE is 817.9 .

1. Triple Exponential smoothing(TES).

Additive Seasonality-

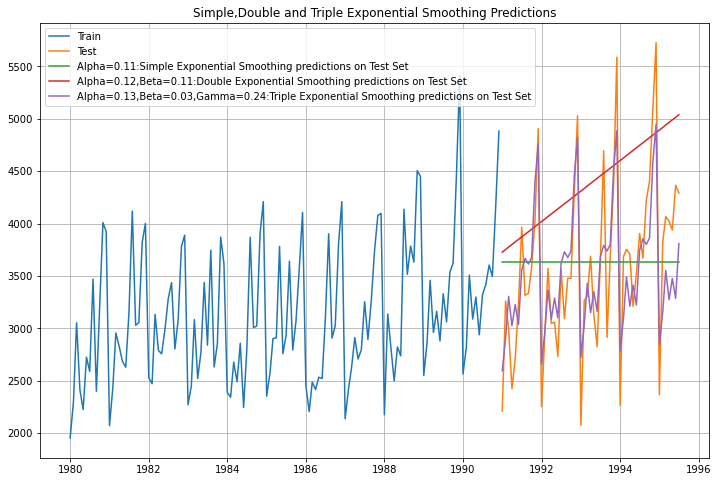


Fig.10

The required value of RMSE is 1283.86 .

Multiplicative Seasonality-

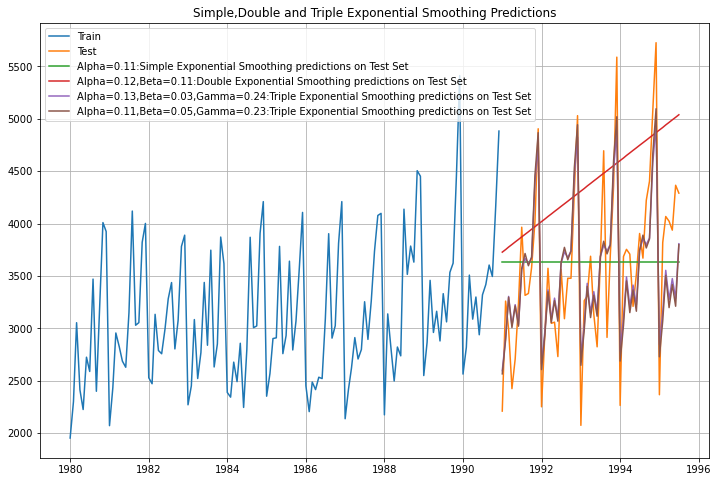


Fig.11

The required value of RMSE is 1307.53 .

Note:We see that the multiplicative seasonality model has not done that well when compared to the additive seasonality Triple Exponential Smoothing model.

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.

The Augmented Dickey-Fuller test is an 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.

After performing the Augmented Dickey-Fuller test the results are ,



Here p-value is greater than alpha=0.05.Thus we fail to reject the null hypothesis.So the time series non-stationary.Hence we again perform the test taking the differencing.Now the results are ,



Here the p-value is much less than 0.05.Hence we reject the null hypothesis.Thus it become stationary.The stationary time series plot is-

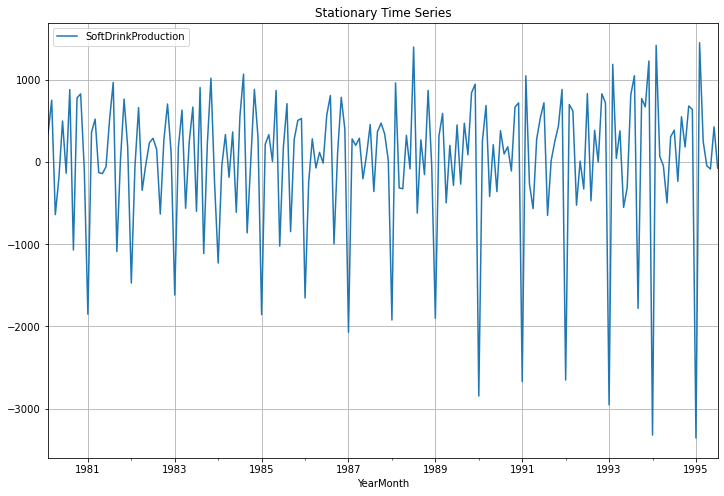
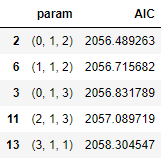


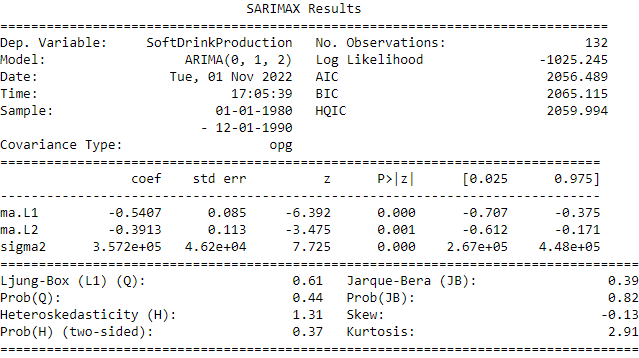
Fig.12

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.

ARIMA model-We are performing an ARIMA model based on the minimum AIC.The model has three parameters (p,d,q).Here d denotes number of differencing.



Thus the minimum AIC value is 2056.48.The corresponding parameters are , p=0 ,d=1 and q=2.We built an ARIMA model with these parameters on train data and check the RMSE value on the test data.



The required RMSE value in test data is 831.62.

7.Build ARIMA/SARIMA models based on the cut-off points of ACF and PACF on the training data and evaluate this model on the test data using RMSE.

PACF plot on the training data ,

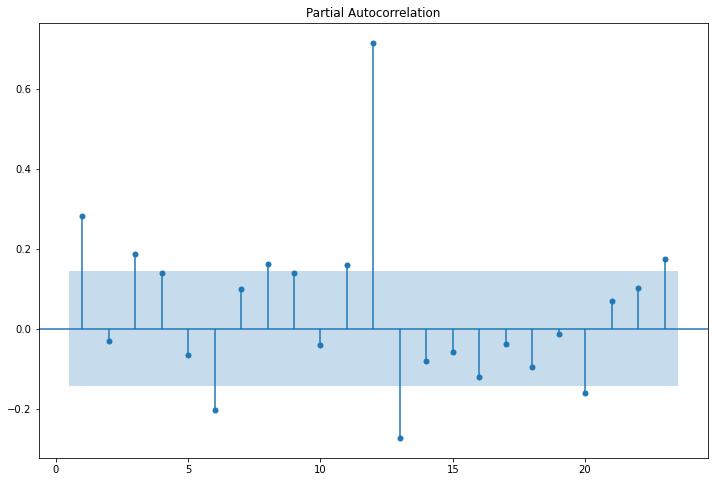


Fig.13

From the PACF plot we can see p=1.

ACF plot on the training data ,

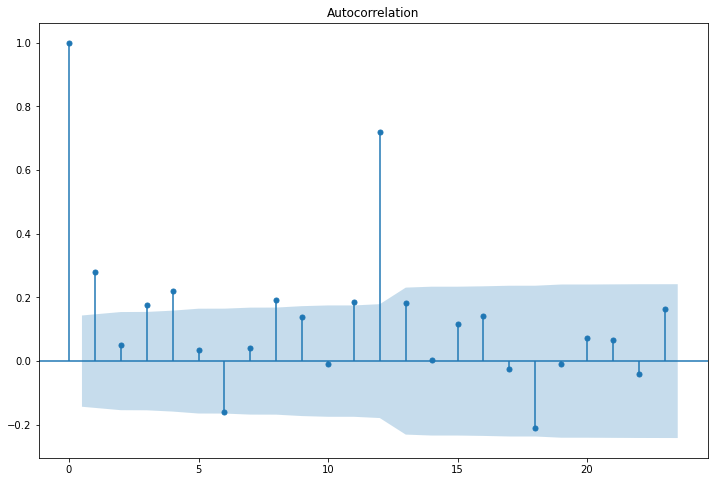
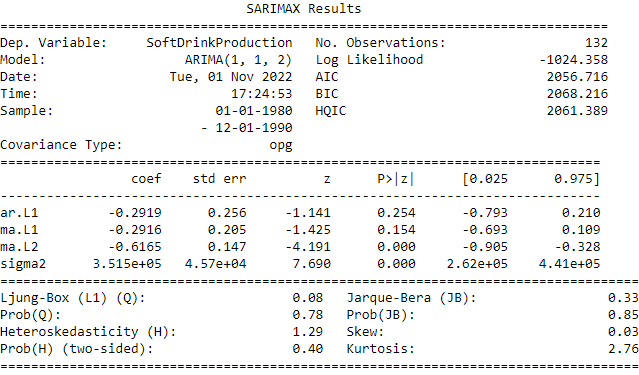


Fig.14

From ACF plot we can see q=2.

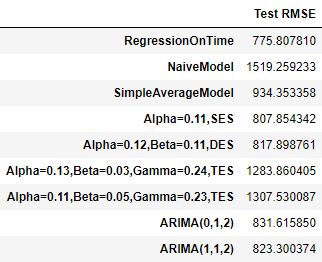
Thus we make an ARIMA model with parameters p=1,d=1 and q=2 on the train data.Here ’d’ is differencing ,hence d=1.



The required RMSE value on the training data is 823.30 .

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

The required table is ,



9.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.

Based on the RMSE value on the test data we observe Linear Regression on time is the best model.Thus predicting 12 months into the future using this model we get ,

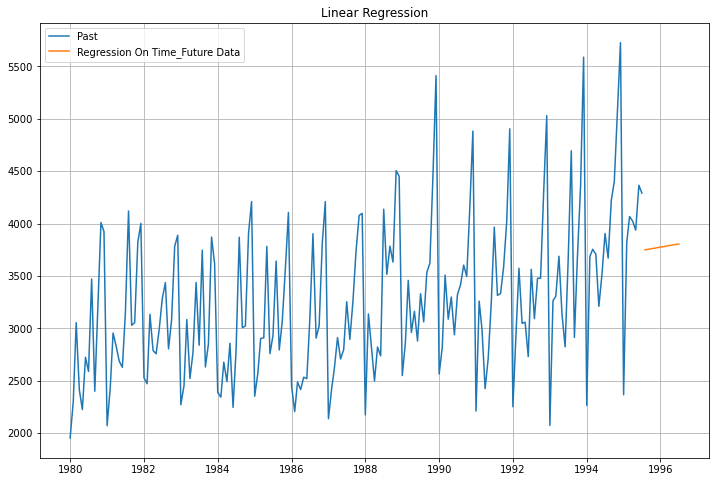


Fig.15

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

The model thus built shows an increasing trend in sales of soft drinks.