Problem 2:

Data Set: SoftDrink.csv

You are an analyst in the RST soft drink company and you are expected to forecast the sales of the production of the soft drink for the upcoming 12 months from where the data ends. The data for the production of soft drinks has 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 was imported in 2 different ways.

Method 1:

The dataset was loaded and the head of the dataset was checked. In this method, the data is read in such a way that it parses the first column (date column) and indicates to the system that this is a one column series through squeeze. This is done by passing the parameters 'parse_dates=True' and 'squeeze=True' while importing the dataset.

Table 1 and Table 2 show the first and last 5 records of the dataset. It is observed that it has 2 columns – YearMonth and the quantity of soft drinks sold in that particular month.

Table 1: Head of the dataset showing the first 5 records

YearMonth	SoftDrinkProduction
1980-01-01	1954
1980-02-01	2302
1980-03-01	3054
1980-04-01	2414
1980-05-01	2226

Table 2: Tail of the dataset showing the last 5 records

YearMonth	SoftDrinkProduction
1995-03-01	4067
1995-04-01	4022
1995-05-01	3937
1995-06-01	4365
1995-07-01	4290

The entries in the YearMonth column are not really a datapoint, but an index for the sales entry.

It can be observed from Table 1 and Table 2 that the dataset has data starting from January 1980 going till July 1995.

There are totally 187 entries in the dataset seen in Figure 1.

```
The no. of entries:  187
```

Figure 1: Number of entries in the dataset

The data is plotted and is shown in Figure 2.

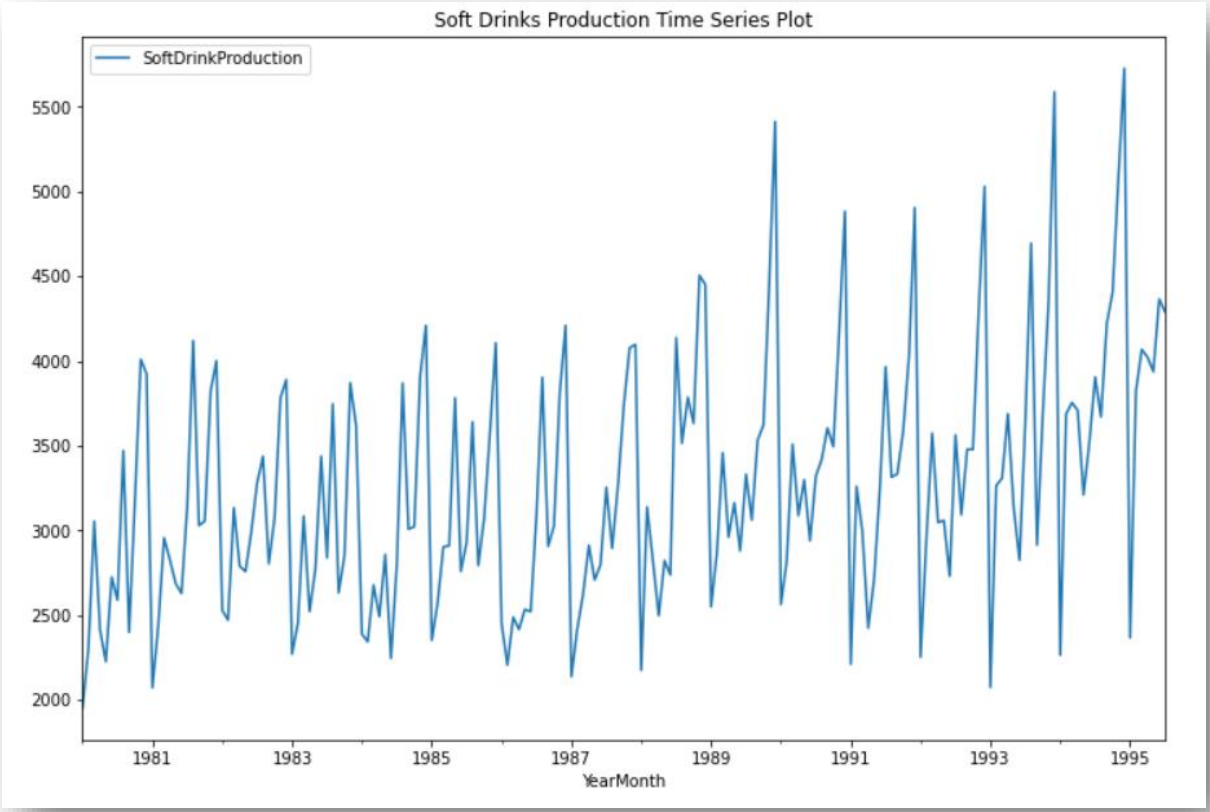

Figure 2: Method 1 – Soft drinks production time series plot

Method 2:

In this method, the dataset is imported and the head of the dataset is checked. The data is further plotted. Table 3 shows the head of the dataset.

Table 3: Head of the dataset showing the first 5 records

	YearMonth	SoftDrinkProduction
0	1980-01	1954
1	1980-02	2302
2	1980-03	3054
3	1980-04	2414
4	1980-05	2226

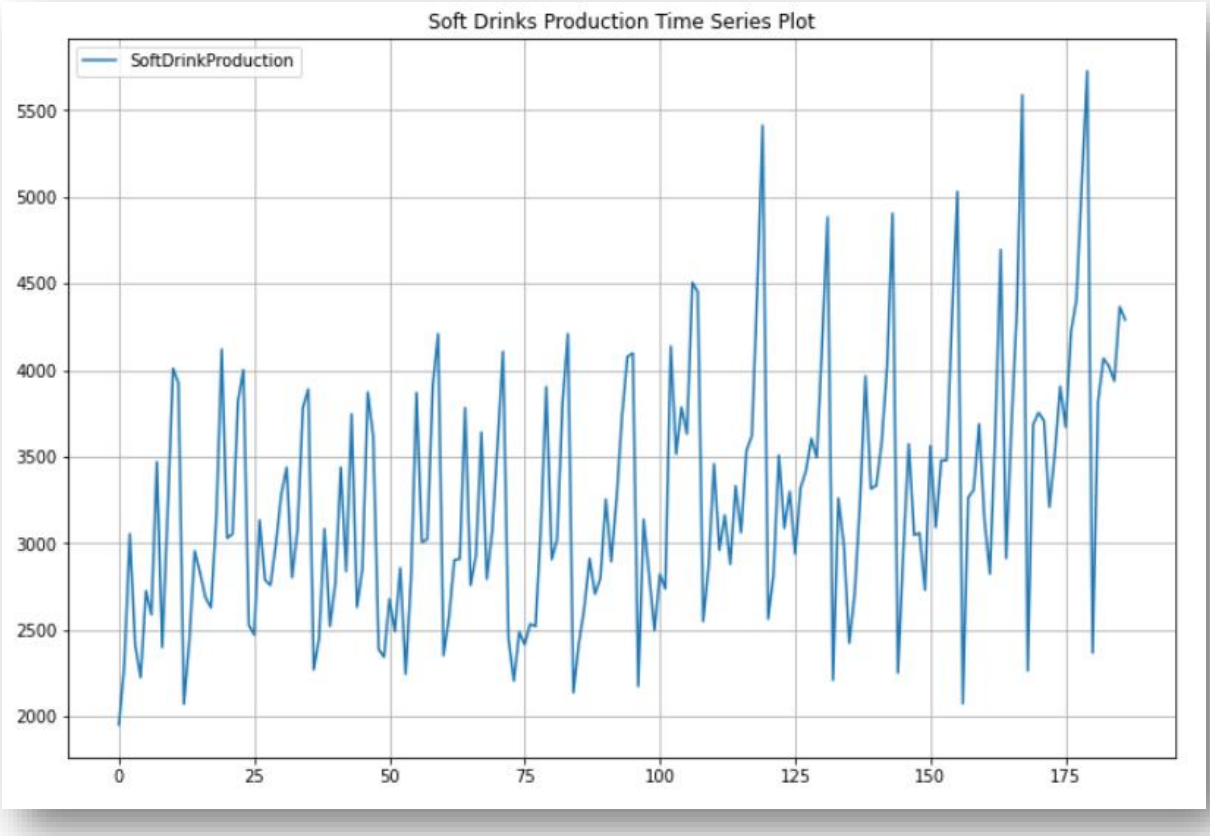

Figure 3: Soft drinks production time series plot

The data is plotted and shown in Figure 3. By observing the x-axis, it can be seen that the data is not read as a time series and it is simply plotted as the x-axis does not have year or month mentioned. In order to make the x-axis as a time series, we need to pass the date range manually through a command in pandas. From Table 3, it is observed that the time series is a monthly time series. Therefore, the time stamp should be defined as a monthly time series. The created time stamp is shown in Figure 4.

```
DatetimeIndex(['1980-01-31', '1980-02-29', '1980-03-31', '1980-04-30',
               '1980-05-31', '1980-06-30', '1980-07-31', '1980-08-31',
               '1980-09-30', '1980-10-31',
               ...
               '1994-10-31', '1994-11-30', '1994-12-31', '1995-01-31',
               '1995-02-28', '1995-03-31', '1995-04-30', '1995-05-31',
               '1995-06-30', '1995-07-31'],
              dtype='datetime64[ns]', length=187, freq='M')
```

Figure 4: Time Stamp

The time stamp is concatenated into the dataframe and the head of the dataset is shown in Table 4.

Table 4: Head of the dataset with time stamp

	YearMonth	SoftDrinkProduction	Time_Stamp
0	1980-01	1954	1980-01-31
1	1980-02	2302	1980-02-29
2	1980-03	3054	1980-03-31
3	1980-04	2414	1980-04-30
4	1980-05	2226	1980-05-31

The column 'YearMonth' is dropped and the data is plotted and checked. Figure 5 shows the time series plot. It can be observed now that the x-axis has data in the form of time series data.

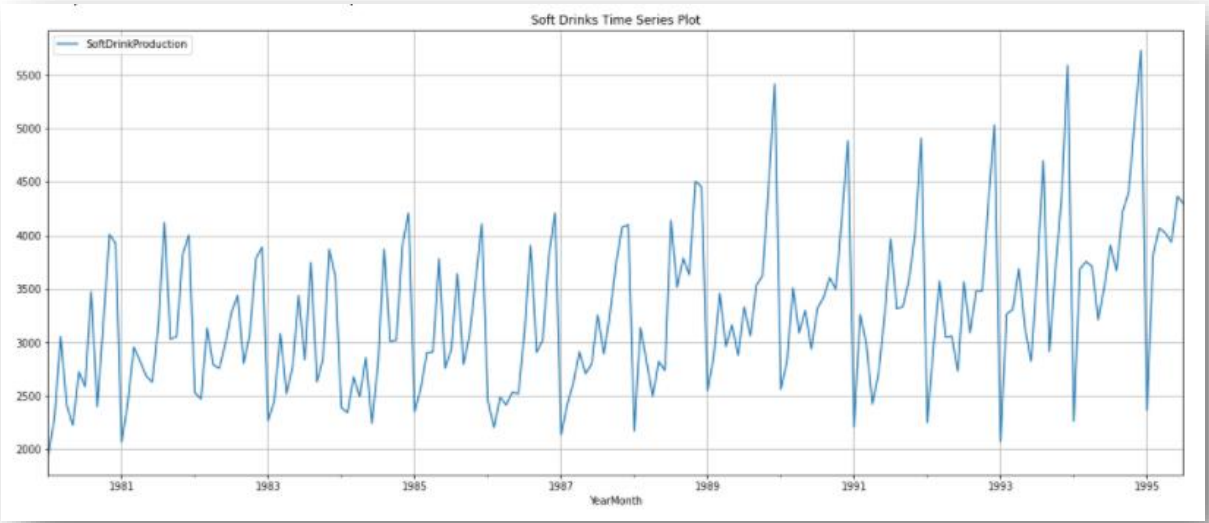

Figure 5: Method 2 – Soft drinks sales time series plot

Inference from the time series plot:

As we can observe from the above Figure 5, the sales for soft drinks are showing a slightly inclining trend in the latter half of the graph. We will explore the trend and seasonality further during decomposition, where we will be able to view a much-detailed report on these two factors. The sales seem to be having a constant shifting or unstable pattern from 1981 to 1990. The increase in the sales of soft drinks can be seen after 1997. The peak sales of soft drinks are observed between the years 1993 to 1995.

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

Table 5: 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   SoftDrinkProduction  187 non-null    int64
dtypes: int64(1)
memory usage: 2.9 KB
```

Table 5 shows that the dataset has 1 column. The column is of different data type integer. There are no null values in the dataset.

Table 6: Descriptive statistics of the dataset

	SoftDrinkProduction
count	187.000000
mean	3262.609626
std	728.357367
min	1954.000000
25%	2748.000000
50%	3134.000000
75%	3741.000000
max	5725.000000

Table 6 shows the description or the summary of the dataset. This helps describe and understand the features of a specific dataset. The most recognized types of descriptive statistics are measures of centre: the mean, median, and mode, which are used at almost all levels of math and statistics.

It can be seen that there are 187 entries meaning there are 187 months in the dataset from January 1980 to July 1995 i.e., 15 years and 7 months. There are no missing values. By looking at Table 6, we are able to deduce that the mean sales are 3262.6 soft drinks. The standard deviation is 728.36. This is probably because the demand for soft drinks varies with different seasons. For example, people tend to drink more soft drinks during hot summer months and less during the winter seasons. This is also because the min and max have significant difference between them. The minimum number of

soft drinks sold in a month is 1954 and maximum number of soft drinks sold in a month is 5725.

The dataset was further checked for missing or null values and it is seen from Figure 6 that there are no missing values in the dataset.

```
Number of null values in the dataset =  SoftDrinkProduction    0
dtype: int64
```

Figure 6: Null values in the dataset

```
Number of duplicate rows = 4
```

Figure 7: Duplicate values in the dataset

The number of duplicate rows in the dataset is found to be 4 shown in Figure 7. However, as each value correspond to a different time index, we will consider that there are no duplicate entries in the dataset. Basically, these are all sales figures for different months.

Yearly boxplot:

Figure 8: Yearly boxplot

Figure 8 shows the yearly boxplot of soft drinks sales. It is seen that soft drinks sales have an unstable upward trend in the first half and then an upward trend in the second half. The highest sales for soft drinks can be observed in 1985 and the lowest sales in 1994. The highest variation in monthly sales for soft drinks seems to be in the year 1993 and on the year 1995 there seems to be the lowest variation in monthly sales. However, the year 1995 has data only till 7 months (till July). Therefore, it is difficult to comment on the sales performance of that year. There are outliers in the yearly sales data, however as it is a time series, we can ignore the outlier data.

Monthly boxplot:

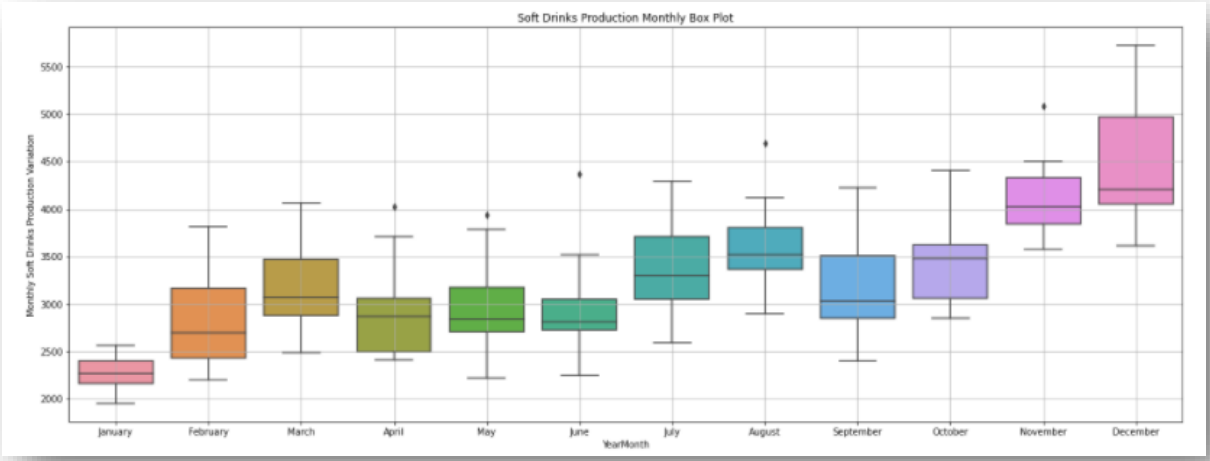

Figure 9: Monthly boxplot

Figure 9 shows the monthly boxplot of soft drinks sales. We can see that there is a slight seasonality element visible in the boxplot. It can be clearly seen that sales have an increasing sales trend in the last quarter of the year, especially in December. The sales for soft drinks seem to pick up from January month and is more or less consistent till August. From September the soft drinks sales have an upward trend till December. The highest sales for soft drinks can be observed in December month.

Monthly sales across years:

The monthly sum of soft drinks sales across years can be seen Table 7.

Table 7: Monthly soft drinks sales across years

Time_Stamp	April	August	December	February	January	July	June	March	May	November	October	September
Time_Stamp												
1980.0	2414.0	3470.0	3924.0	2302.0	1954.0	2589.0	2725.0	3054.0	2226.0	4009.0	3180.0	2400.0
1981.0	2828.0	4119.0	4001.0	2434.0	2072.0	3150.0	2629.0	2956.0	2687.0	3821.0	3055.0	3030.0
1982.0	2789.0	3437.0	3889.0	2472.0	2529.0	3282.0	2993.0	3134.0	2758.0	3782.0	3076.0	2804.0
1983.0	2522.0	3746.0	3618.0	2452.0	2271.0	2839.0	3438.0	3084.0	2769.0	3871.0	2851.0	2632.0
1984.0	2492.0	3869.0	4209.0	2344.0	2389.0	2800.0	2246.0	2678.0	2858.0	3907.0	3023.0	3007.0
1985.0	2910.0	3641.0	4106.0	2570.0	2353.0	2931.0	2759.0	2903.0	3782.0	3576.0	3070.0	2794.0
1986.0	2416.0	3903.0	4209.0	2206.0	2452.0	3093.0	2521.0	2488.0	2534.0	3812.0	3025.0	2907.0
1987.0	2912.0	2895.0	4097.0	2419.0	2138.0	3254.0	2798.0	2622.0	2708.0	4077.0	3736.0	3263.0
1988.0	2498.0	3515.0	4451.0	3138.0	2175.0	4137.0	2738.0	2823.0	2822.0	4504.0	3632.0	3785.0
1989.0	2961.0	3062.0	5411.0	2867.0	2550.0	3331.0	2880.0	3458.0	3163.0	4464.0	3622.0	3534.0
1990.0	3088.0	3418.0	4882.0	2820.0	2564.0	3320.0	2939.0	3508.0	3299.0	4163.0	3495.0	3604.0
1991.0	2425.0	3315.0	4904.0	3260.0	2211.0	3965.0	3244.0	2992.0	2707.0	4021.0	3583.0	3333.0
1992.0	3048.0	3092.0	5029.0	2952.0	2252.0	3563.0	2731.0	3573.0	3059.0	4308.0	3478.0	3478.0
1993.0	3688.0	4694.0	5587.0	3264.0	2075.0	3644.0	2824.0	3308.0	3136.0	4358.0	3686.0	2914.0
1994.0	3708.0	3670.0	5725.0	3685.0	2265.0	3905.0	3517.0	3754.0	3210.0	5086.0	4404.0	4221.0
1995.0	4022.0	NaN	NaN	3819.0	2367.0	NaN	4365.0	4067.0	3937.0	NaN	NaN	NaN

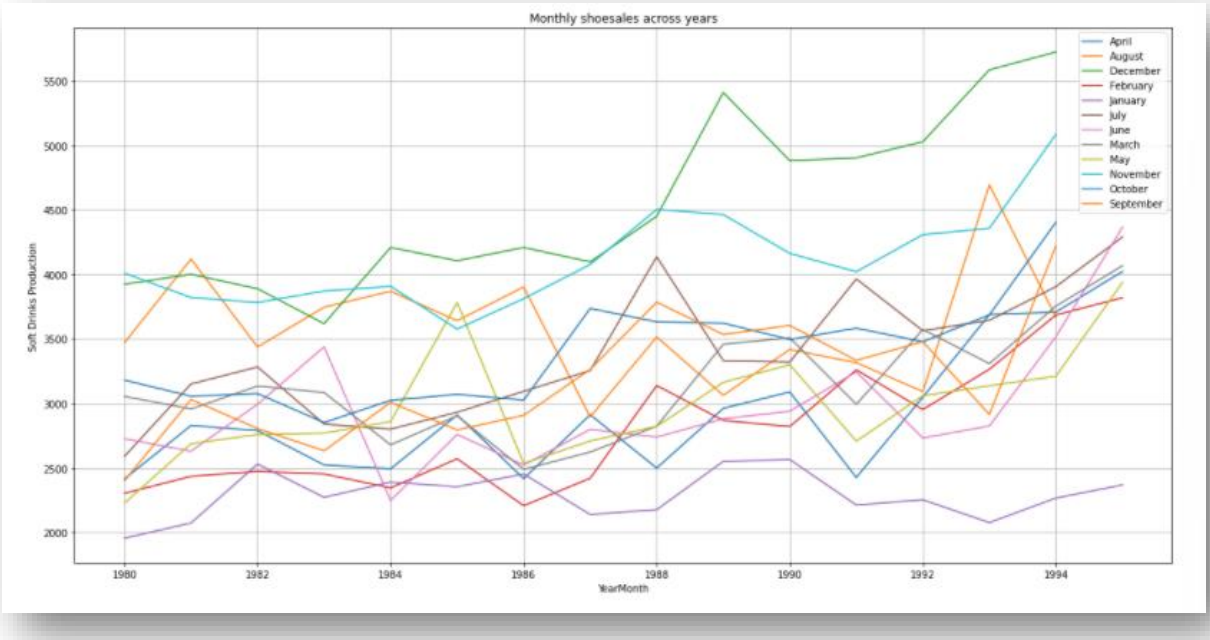

Figure 10: Monthly soft drinks sales across years

From Table 7 and Figure 10, it is observed that the months of December seems to be the month that has the highest sales figures. The second highest sales being made in November month. For certain years, the second highest sales are in August.

Yearly sum of observations:

The sum of soft drinks sales in each year is calculated to understand the data from an annual perspective. The head of sum of soft drinks sales in each year is shown in Table 8.

Table 8: Yearly sum of observations

YearMonth	SoftDrinkProduction
1980-12-31	34247
1981-12-31	36782
1982-12-31	36945
1983-12-31	36093
1984-12-31	35822

The sum of sales in each year is plotted and shown in Figure 11. It is observed from the plotted graph that annual soft drinks sales show a dip initially with sales picking up from the year 1986 right up to the year 1990 and then observing another dip in the sales. The steep drop post 1994 is because of the relatively less (half year data - till July) data available for the year 1995.

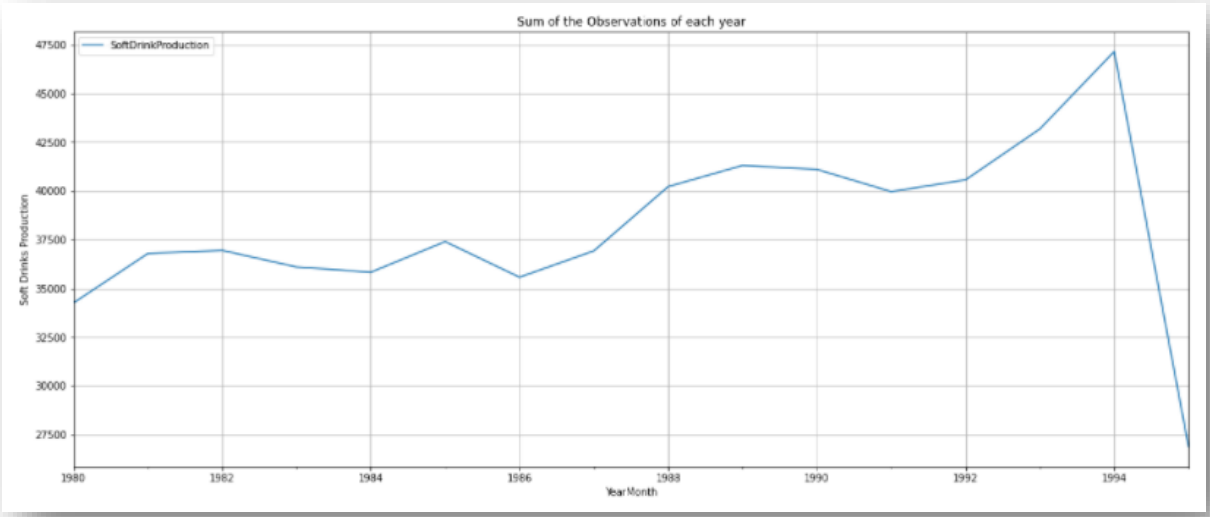

Figure 11: Yearly sum of observations

Mean of observations of each year:

The mean sales of each year are calculated and the head of the data is shown in Table 9.

Table 9: Mean of observations of each year

YearMonth	SoftDrinkProduction
1980-12-31	2853.916667
1981-12-31	3065.166667
1982-12-31	3078.750000
1983-12-31	3007.750000
1984-12-31	2985.166667

The mean sales of each year are plotted and shown in Figure 12. Observations from 'yearly sum of observations' is confirmed by looking at Figure 12 as the pattern is very similar to Figure 11. It is observed from the plotted graph that annual soft drinks sales show a dip initially with sales picking up from the year 1986 right up to the year 1990 and then observing another dip in the sales. The steep drop post 1994 is because of the relatively less (half year data - till July) data available for the year 1995.

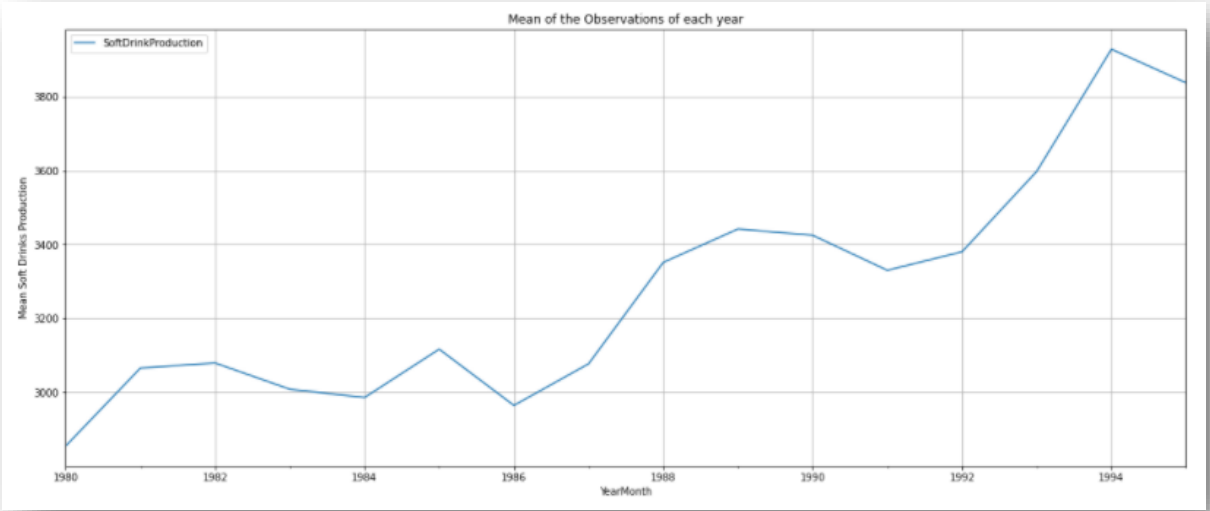

Figure 12: Mean of observations of each year

Sum of observations of each quarter:

The quarterly sum of sales numbers can be observed in Table 10 and Figure 13.

Table 10: Sum of observations of each quarter

YearMonth	SoftDrinkProduction
1980-03-31	7310
1980-06-30	7365
1980-09-30	8459
1980-12-31	11113
1981-03-31	7462

From Figure 13 we can observe that the quarterly sales show an upward trend. The quarterly series is able to catch the seasonality in the data. It can be observed that there is a slight element of seasonality in the time series.

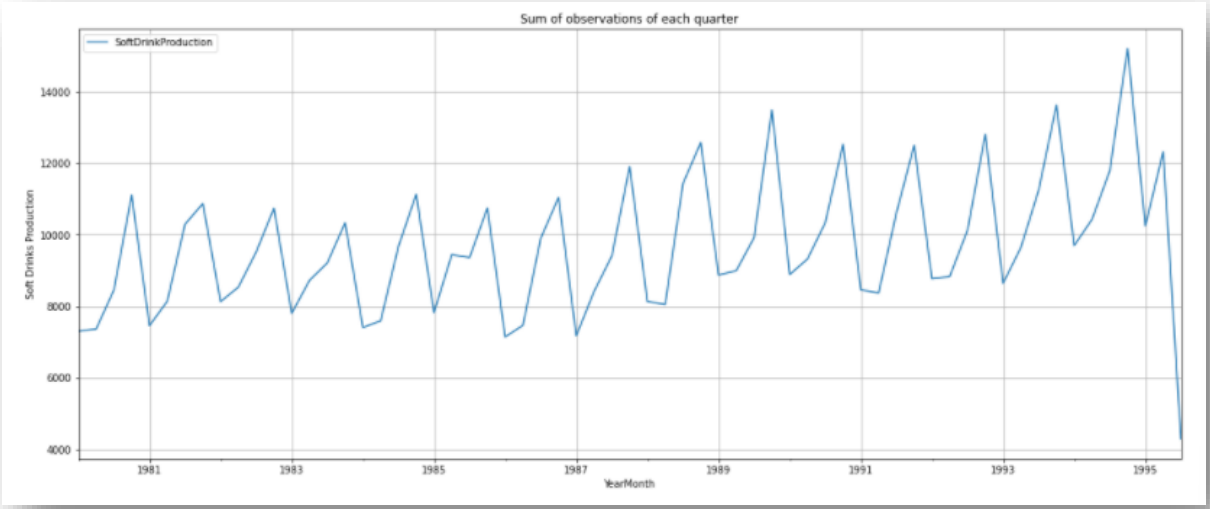

Figure 13: Sum of observations of each quarter

Mean of observations of each quarter:

The mean sales of each quarter are calculated and the head of the data is shown in Table 11.

Table 11: Mean of observations of each quarter

YearMonth	SoftDrinkProduction
1980-03-31	2436.666667
1980-06-30	2455.000000
1980-09-30	2819.666667
1980-12-31	3704.333333
1981-03-31	2487.333333

The mean sales of each quarter are plotted and shown in Figure 14. Observations from 'sum of observations of each quarter' is confirmed by looking at Figure 14 as the pattern is very similar to Figure 13. It is observed from the plotted graph that the quarterly sales show an upward trend. The quarterly series is able to catch the seasonality which is present in the data.

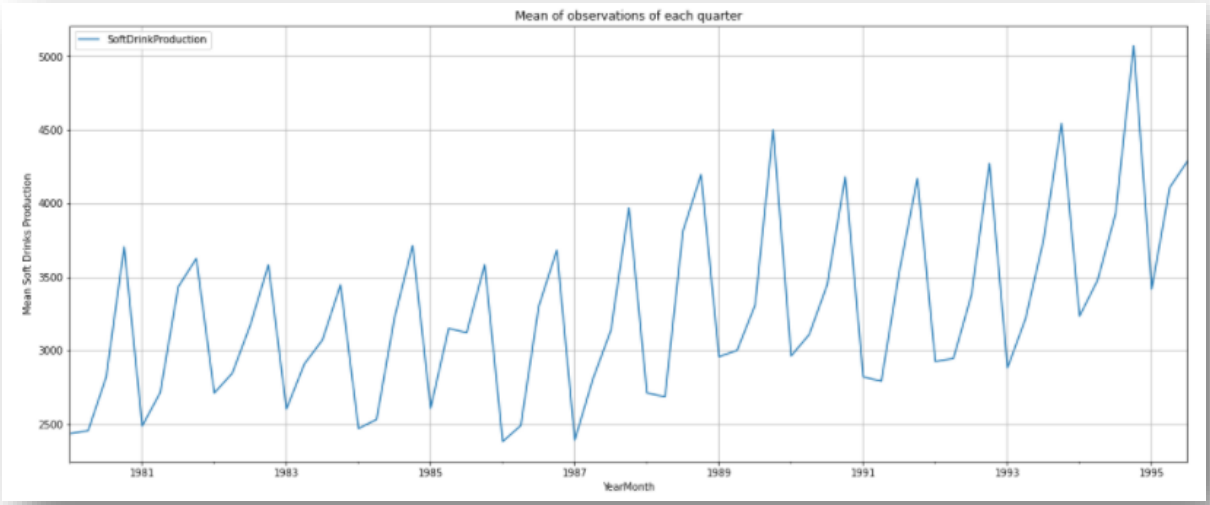

Figure 14: Mean of observations of each quarter

Daily plot:

Daily plot is done to understand the data from a daily perspective. The values which the original series cannot provide is taken as 0 by python if we try to resample the data on a daily basis. This is seen in Table 12.

Table 12: Daily plot data

YearMonth	SoftDrinkProduction
1980-01-01	1954
1980-01-02	0
1980-01-03	0
1980-01-04	0
1980-01-05	0
...	...
1995-06-27	0
1995-06-28	0
1995-06-29	0
1995-06-30	0
1995-07-01	4290

5661 rows × 1 columns

Figure 15 shows the daily plot of the data. However, the graph fails to give us a proper understanding of our data. Thus, resampling the data to intervals where a number of observations are 0 is not a good idea as that does not give us an understanding of the performance of the time series.

To get a very high-level overview of the trend of the Time Series Data (if Trend is present) can be understood by resampling the data keeping the intervals very large.

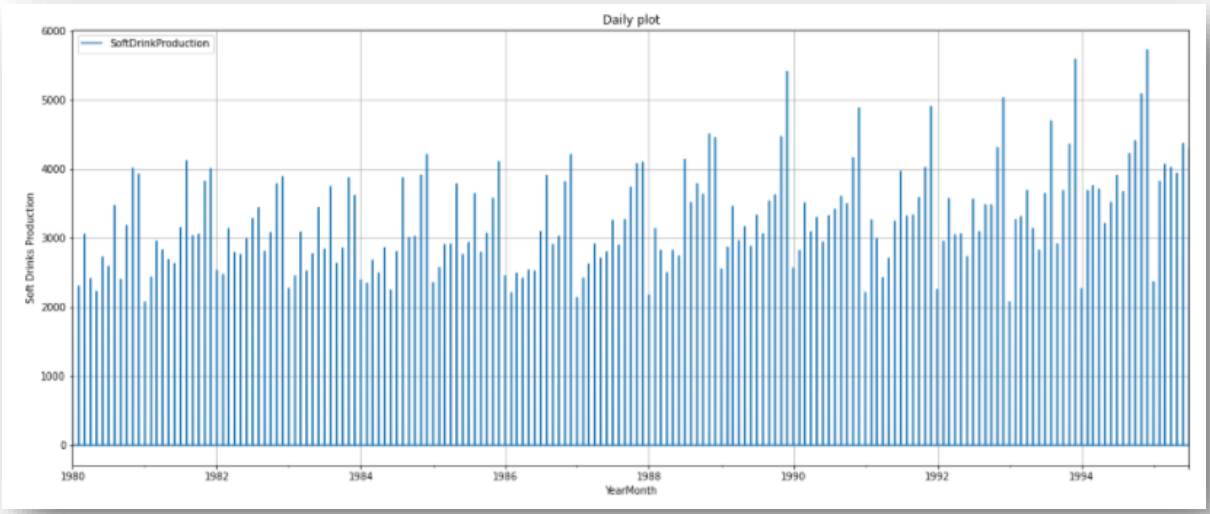

Figure 15: Daily plot

Decade plot:

Decade plot is done to understand the data from a 10-year perspective. The sum of sales in each decade is shown in Table 13.

Table 13: Decade plot data

YearMonth	SoftDrinkProduction
1980-12-31	34247
1990-12-31	378143
2000-12-31	197718

Figure 16 shows the decade plot with the sum of soft drinks sales in each decade. If we take the resampling period to be 10 years or a decade, we see that the seasonality present has been smoothed over and it is only giving an estimate of the trend. The trend first shows an upward trend and then a downward trend.

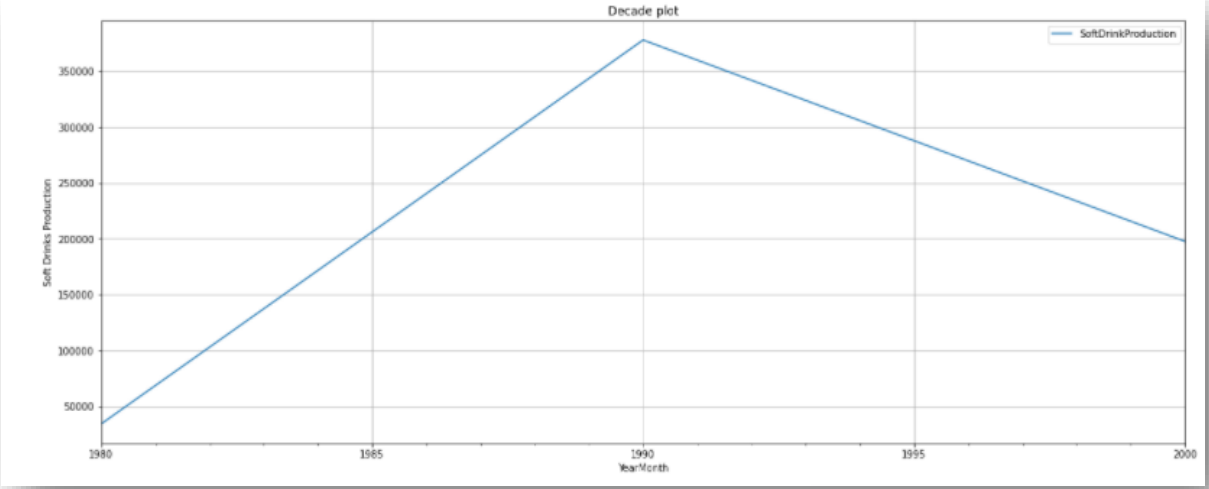

Figure 16: Decade plot

Decomposition:

Seasonalities can be checked decomposing the time series using importing seasonal_decompose function from statsmodel.tsa.seasonal. This function will decompose our time series into trend, seasonality and noise.

Time series decomposition involves thinking of the time series as a combination of level, trend, seasonality and noise components. Decomposition provides an useful abstract model for thinking about time series generally and for better understanding problems during time series analysis and forecasting.

Additive Model:

An additive model is linear where changes over time are consistently made by the same amount. Additive model is preferred when the trend is constant. When trend is erratic, seasonality is same.

y(t) = Level + Trend + Seasonality + Noise

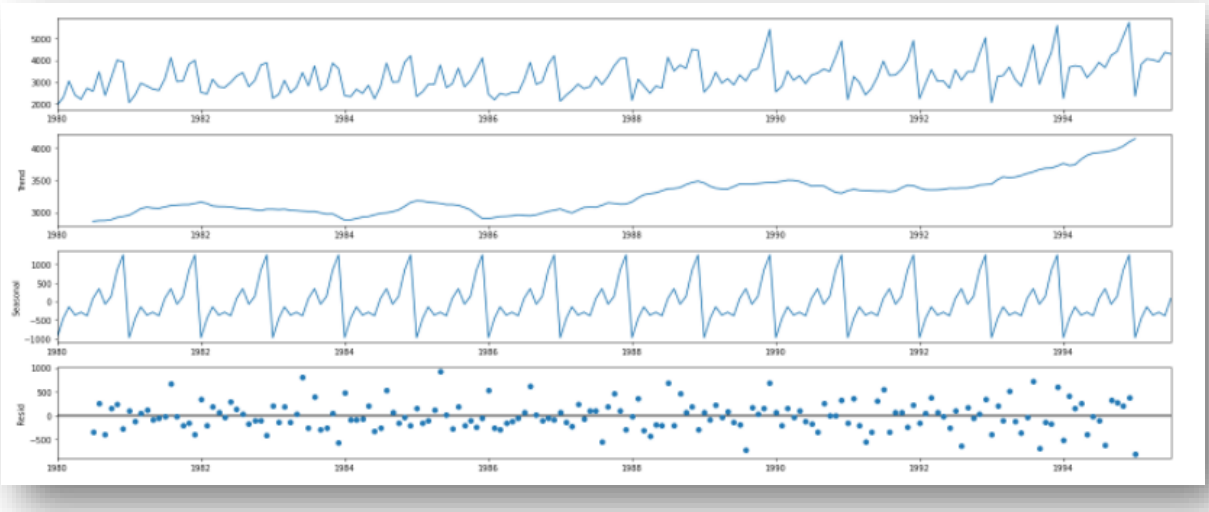

Figure 17: Decomposition - Additive model

From **Error! Reference source not found.**, we can see a clear daily seasonal pattern with a peak f ollowed by two other peaks and then a steep valley. Despite this pattern, there is still a lot of noise that is not explained by our daily seasonality. As per the 'additive' decomposition, we see that there is a pronounced trend in the later years of the data.

Multiplicative Model:

A multiplicative model is nonlinear, such as quadratic or exponential. Changes increase or decrease over time.

y(t) = Level * Trend * Seasonality * Noise

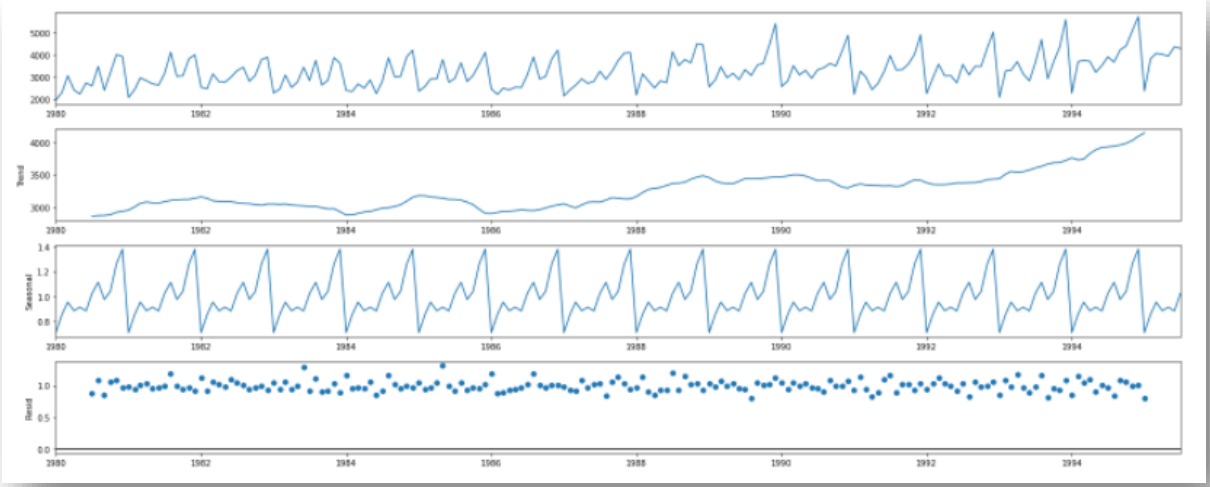

Figure 18: Decomposition - Multiplicative model

From Figure 18, we can see a clear daily seasonal pattern with a peak followed by a valley with another peak in it. The residuals are more or less in the same proximity and form a straight line like pattern in the multiplicative model.

We have decomposed both models, from comparing both the residuals we see a better pattern with the multiplicative decomposition and hence we prefer the multiplicative decomposition.

The trend, seasonality and residual are separated and the trend and residual are combined. A line plot is carried out to analyse the dataset without seasonality and compared with the original dataset shown in Figure 19.

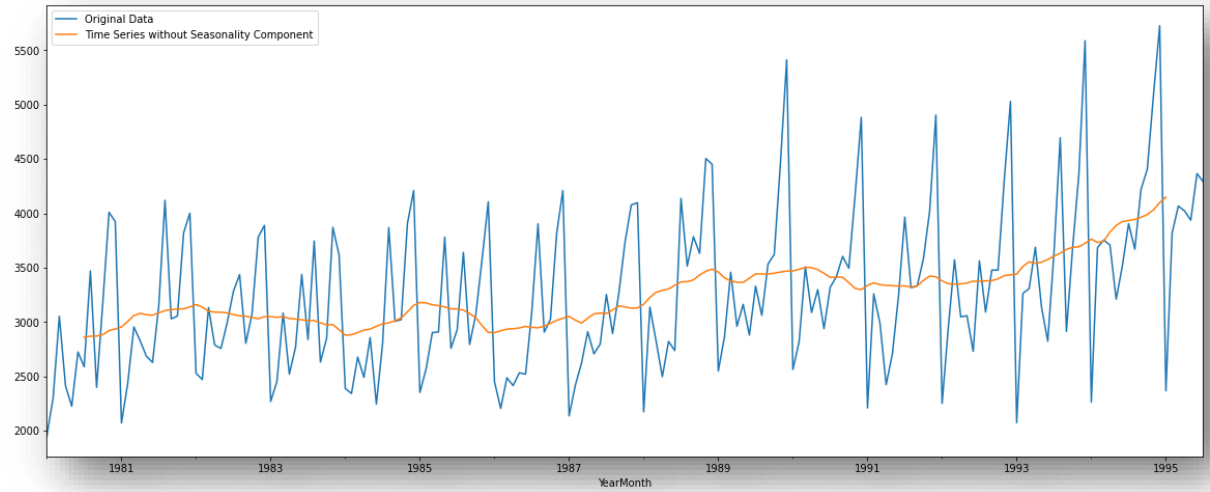

Figure 19: Time series without seasonality component

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

The data is split into train and test datasets. It is mentioned that the test data should start in 1991. Therefore, the train data has data till 1990 and the test data has the rest of the dataset starting from 1991.

Table 14: Head of training data

First few rows of Training Data

Time_Stamp	SoftDrinkProduction
1980-01-31	1954
1980-02-29	2302
1980-03-31	3054
1980-04-30	2414
1980-05-31	2226

Table 15: Tail of training data

Last few rows of Training Data

Time_Stamp	SoftDrinkProduction
1990-08-31	3418
1990-09-30	3604
1990-10-31	3495
1990-11-30	4163
1990-12-31	4882

Table 14 and Table 15 show the first and last 5 records of the training data. It is seen that the training data has data from 1980 till 1990.

Table 16: Head of test data

First few rows of Test Data

Time_Stamp	SoftDrinkProduction
1991-01-31	2211
1991-02-28	3260
1991-03-31	2992
1991-04-30	2425
1991-05-31	2707

Table 17: Tail of test data

Last few rows of Test Data

YearMonth	SoftDrinkProduction
1995-03-01	4067
1995-04-01	4022
1995-05-01	3937
1995-06-01	4365
1995-07-01	4290

Table 16 and Table 17 show the first and last 5 records of the test data. The test data starts from 1991. Figure 20 shows that the train data has 132 observations and the test data has 54 observations.

```
The shape of the training data is (132, 1)
The shape of the test data is (54, 1)
```

Figure 20: Shape of the train and test data

The train and the test data are plotted and shown in **Error! Reference source not found.**. The blue p art of the plot depicts the train data from January 1980 to December 1990. The orange part depicts the test data from January 1991 to July 1995.

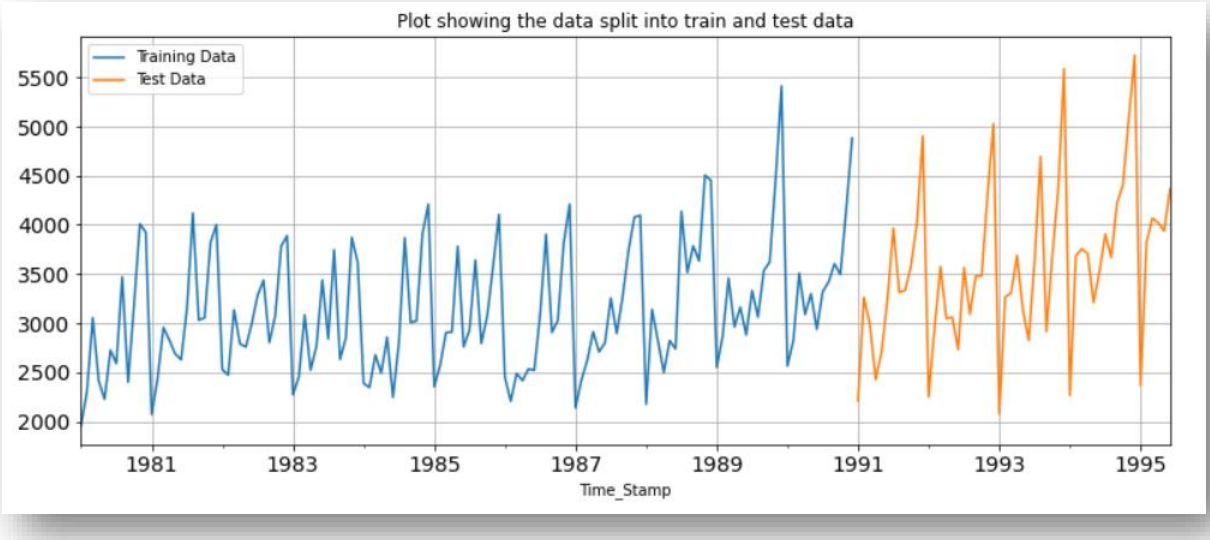

Figure 21: Plot showing the data split into train and test data

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

Models such as linear regression, naïve, simple average models, moving average models, exponential smoothing models like simple exponential smoothing, double exponential smoothing and triple exponential smoothing models were run. The RMSE scores for all these models were calculated and compared against each other to arrive at the optimum model for our dataset.

Model 1: Linear Regression

The 'SoftDrinkProduction' variable is regressed against the order of the occurrence. The training data should be modified before fitting it into a linear regression. Also, for linear regression we should first generate the numerical time instance order for both the training and test set. These values are then added in the training and test set. The model is now run to forecast on the test data and the plot is shown in Figure 22.

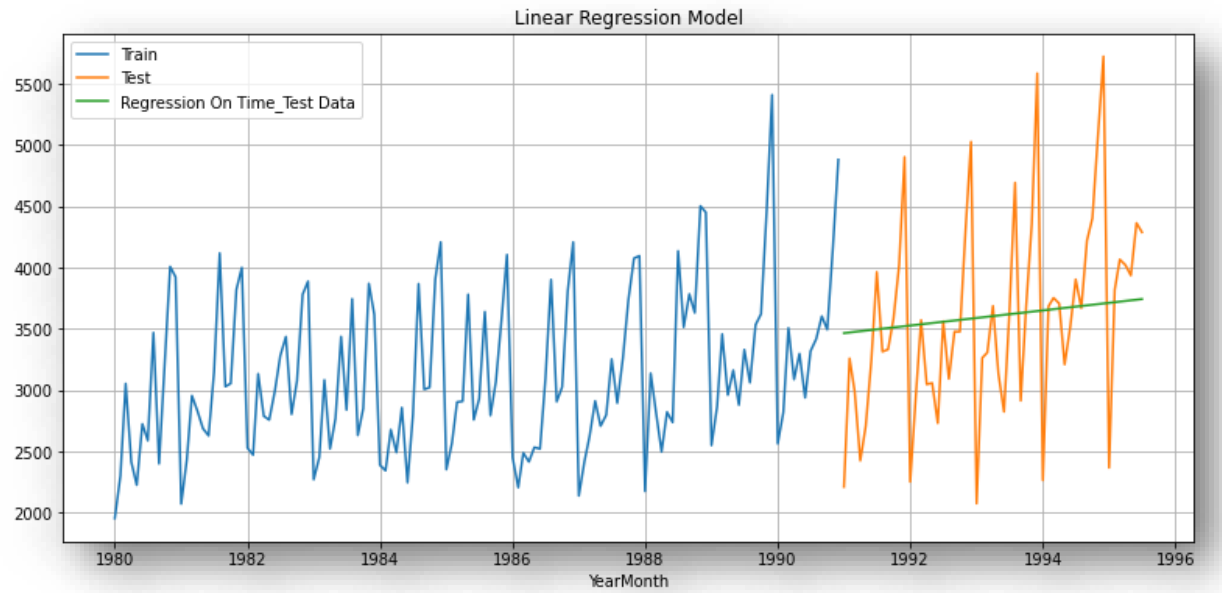

Figure 22: Linear regression forecast on test data

```
For RegressionOnTime forecast on the Test Data,  RMSE is 775.808
```

Figure 23: Linear Regression - RMSE score

The RMSE score for RegressionOnTime forecast on the Test Data is 755.808.

Model 2: Naïve Approach

In this naïve 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. The forecast on the test data using the naïve approach is shown in Figure 24.

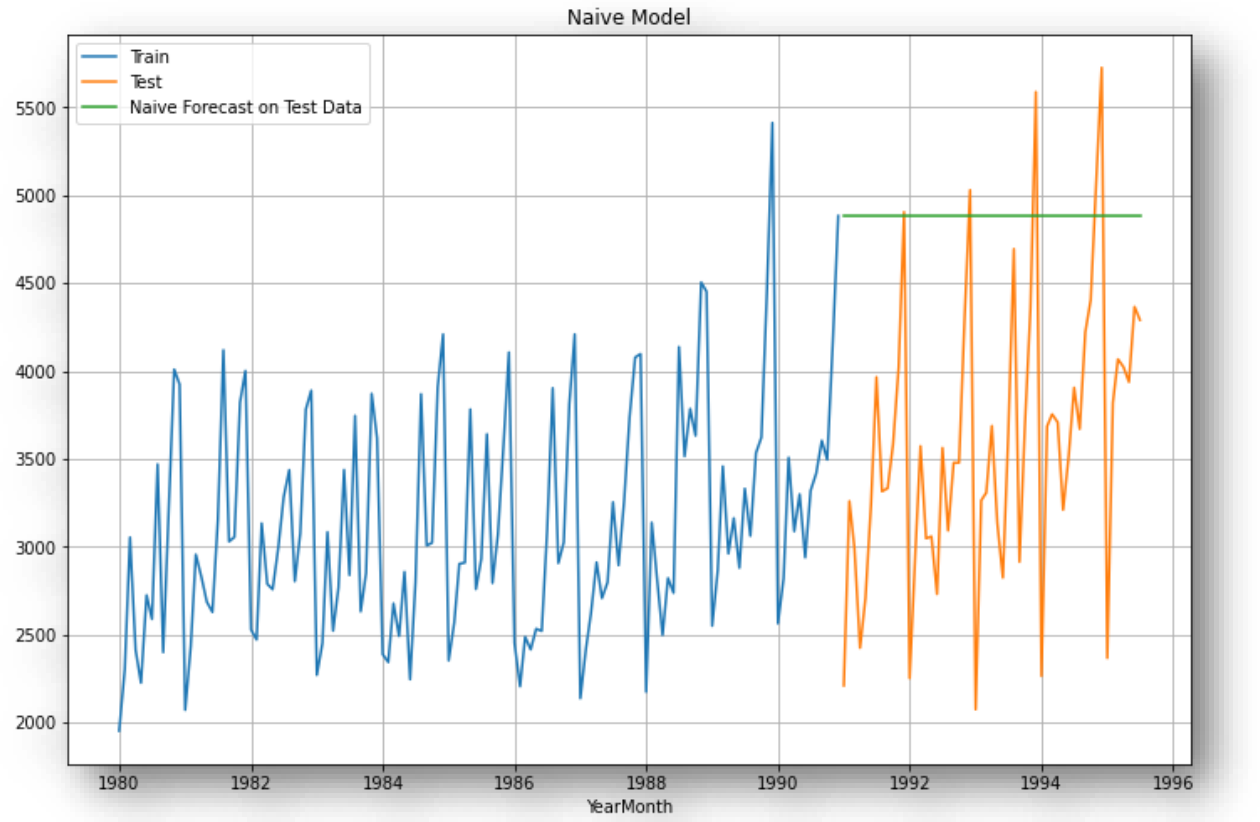

Figure 24: Naïve forecast on test data

```
For Naive forecast on the Test Data,  RMSE is 1519.259
```

Figure 25: Naive Approach - RMSE score

The RMSE score for Naïve model forecast on the Test Data is 1519.259.

Model 3: Simple Average

The model is very simple. The simple average model takes the data by months or quarters or years and then calculates the average for the period. The model then calculates the percentage of it compared to the grand average. The forecast on the test data using the simple average model is shown in Figure 26.

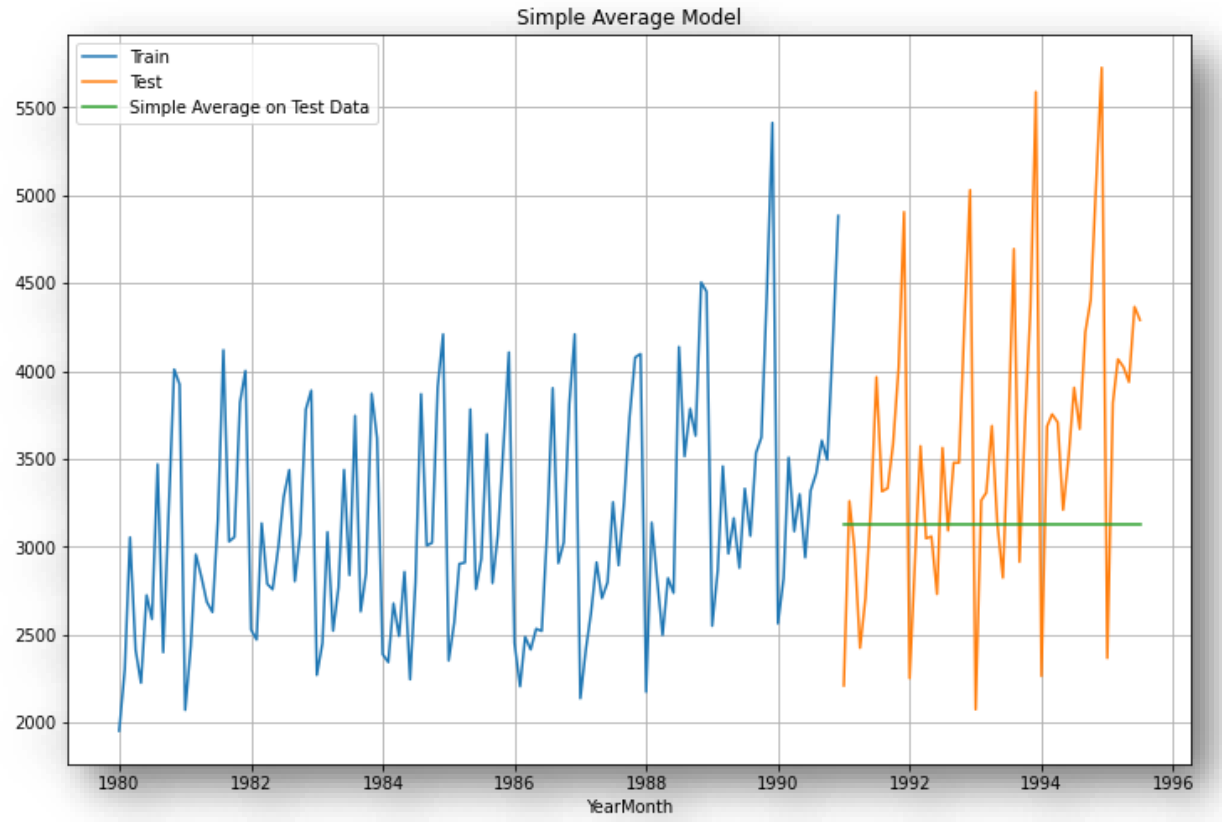

Figure 26: Simple Average forecast on test data

```
For Simple Average forecast on the Test Data,  RMSE is 934.353
```

Figure 27: Simple Average model - RMSE score

The RMSE score for Simple average model forecast on the Test Data is 934.353.

Model 4: Moving Average (MA)

The moving average model is a time series model that accounts for very short-run autocorrelation. The rolling means (or moving averages) are calculated for different intervals. It basically states that the next observation is the mean of every past observation.

The moving average is easy to calculate and, once plotted on a chart, is a powerful visual trend-spotting tool. The moving average of the whole data is shown in Figure 28. This model is one of the most popular and often-used technical indicators to understand stock movement

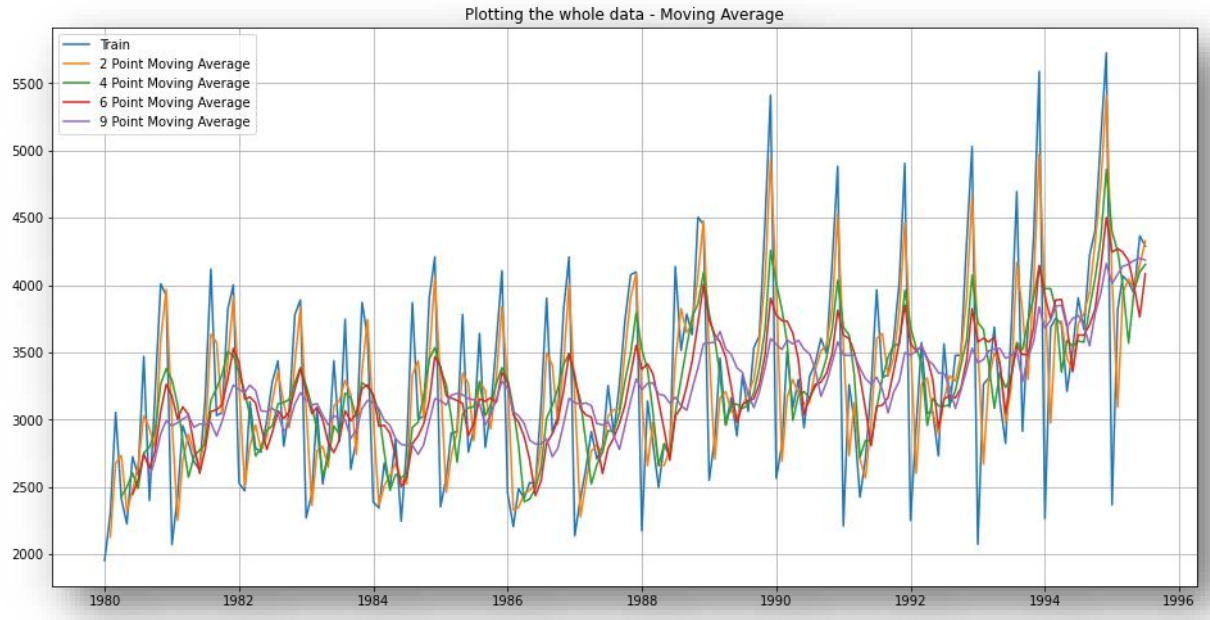

Figure 28: Plotting the whole data - Moving average

The best interval can be determined by the maximum accuracy (or the minimum error). For Moving Average, the entire data is averaged over at different intervals. Moving averages at intervals 2, 4, 6 and 9 are analysed and the forecast on the test data is shown in **Error! Reference source not found.**Figure 29. The RMSE scores of all the four moving averages are calculated and shown in Figure 30.

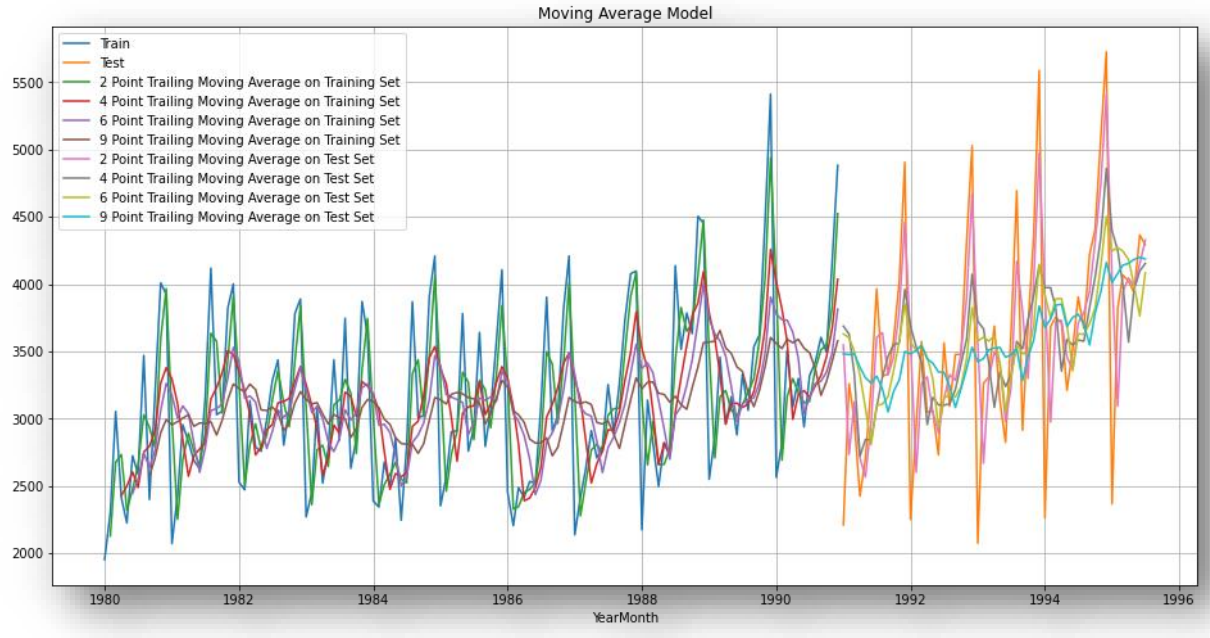

Figure 29: Moving average forecast on test data

```
For 2 point Moving Average Model forecast on the Test Data,  RMSE is 556.725
For 4 point Moving Average Model forecast on the Test Data,  RMSE is 687.182
For 6 point Moving Average Model forecast on the Test Data,  RMSE is 710.514
For 9 point Moving Average Model forecast on the Test Data,  RMSE is 735.890
```

Figure 30: Moving Average model - RMSE score

Comparison of models:

A comparison on the RMSE scores of the models so far were carried out and shown in Table 18.

Table 18: Table comparing the test RMSE scores of the models

	Test RMSE
RegressionOnTime	775.807810
NaiveModel	1519.259233
SimpleAverageModel	934.353358
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827

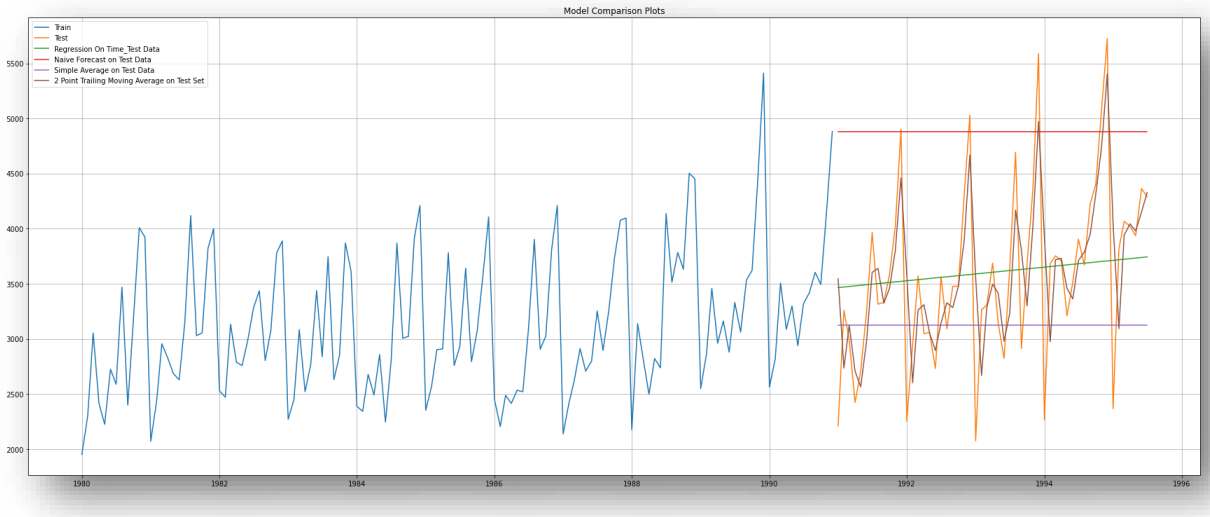

Figure 31: Plot with the forecast on the test data of the models executed so far

Inference from the models:

From Figure 31 we can infer that forecasting on the test data using different models has given us different types of analysis.

The forecast predicted using the linear regression, simple average and naïve approach are a straight line and the prediction is following a trend. The linear regression forecast tells us that the data is on an increasing trend whereas the forecast using the naïve approach and simple average model conveys that there is no variation in the prediction i.e., that is the prediction for today is the same as the prediction on the entire test data as it is a straight line. On comparing the forecast from the three models, linear regression, naïve forecast and simple average with the actual test data we can see that the prediction is not very reliable.

Observing the forecast of the moving average plots, we can see that these are more reliable compared to the other three models. The forecast is the most similar to the test data. The RMSE scores of the models were compared and the 2-point moving average model seems to have the lowest RMSE score.

Exponential Smoothing methods

Exponential smoothing methods consist of flattening time series data. Exponential smoothing averages or exponentially weighted moving averages consist of forecast based on previous periods data with exponentially declining influence on the older observations.

Exponential smoothing methods consist of special case exponential moving with notation ETS (Error, Trend, Seasonality) where each can be none(N), additive (N), additive damped (Ad), Multiplicative (M) or multiplicative damped (Md). One or more parameters control how fast the weights decay and these parameters have values between 0 and 1

There are three types of exponential smoothing models, Simple, Double or Holt's Model and Triple or Holt's-Winter's model

Model 5: Simple Exponential Smoothing

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.

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.

Figure 32 shows the optimum model parameters for the SES model. The alpha or the smoothing level is seen to be 0.216. This alpha value is used to forecast the test data and the plot is shown in Figure 33.

```
{'damping_trend': nan,
 'initial_level': 2297.4228976530508,
 'initial_seasons': array([], dtype=float64),
 'initial_trend': nan,
 'lamda': None,
 'remove_bias': False,
 'smoothing_level': 0.21628856026090063,
 'smoothing_seasonal': nan,
 'smoothing_trend': nan,
 'use_boxcox': False}
```

Figure 32: Simple Exponential Smoothing Model optimum parameters

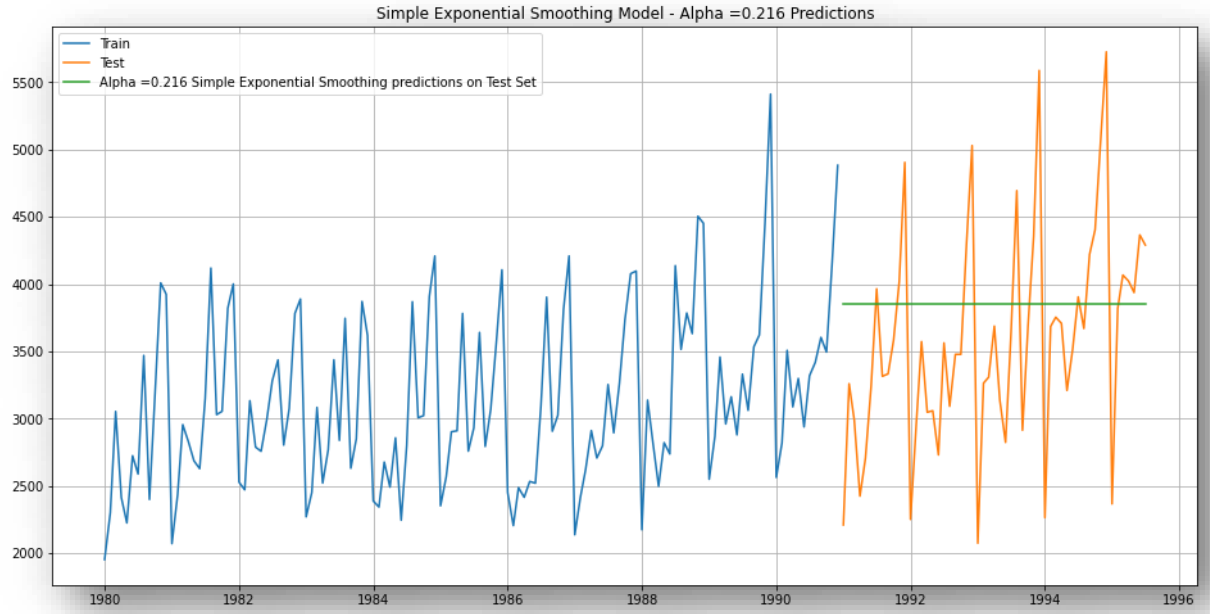

Figure 33: Simple Exponential Smoothing forecast on the test data (alpha = 0.216)

```
For Alpha =0.216 Simple Exponential Smoothing Model forecast on the Test Data, RMSE is 847.635
```

Figure 34: Simple Exponential Smoothing model - RMSE score

The RMSE score for alpha = 0.216 Simple Exponential Smoothing Model forecast on the Test Data 847.635.

The model is tried by setting different alpha values. The higher the alpha value more weightage is given to the more recent observation. That means, what happened recently will happen again. We will run a loop with different alpha values to understand which particular value works best for alpha on the test set. The RMSE scores for the different alpha values are shown in Table 19.

Table 19: Simple Exponential Smoothing model RMSE score for different alpha values

	Alpha Values	Train RMSE	Test RMSE
0	0.3	650.458591	910.187416
1	0.4	656.803409	1005.179377
2	0.5	664.777265	1105.985227
3	0.6	674.988238	1203.565956
4	0.7	687.376817	1294.680933
5	0.8	701.579829	1378.198740
6	0.9	717.287681	1453.359494
7	1.0	734.461852	1519.259233

It is observed from Table 19 that the alpha value 0.3 has the least RMSE score. Therefore, the model is forecasted on the test data with alpha value of 0.3 and the plot is shown in Figure 35.

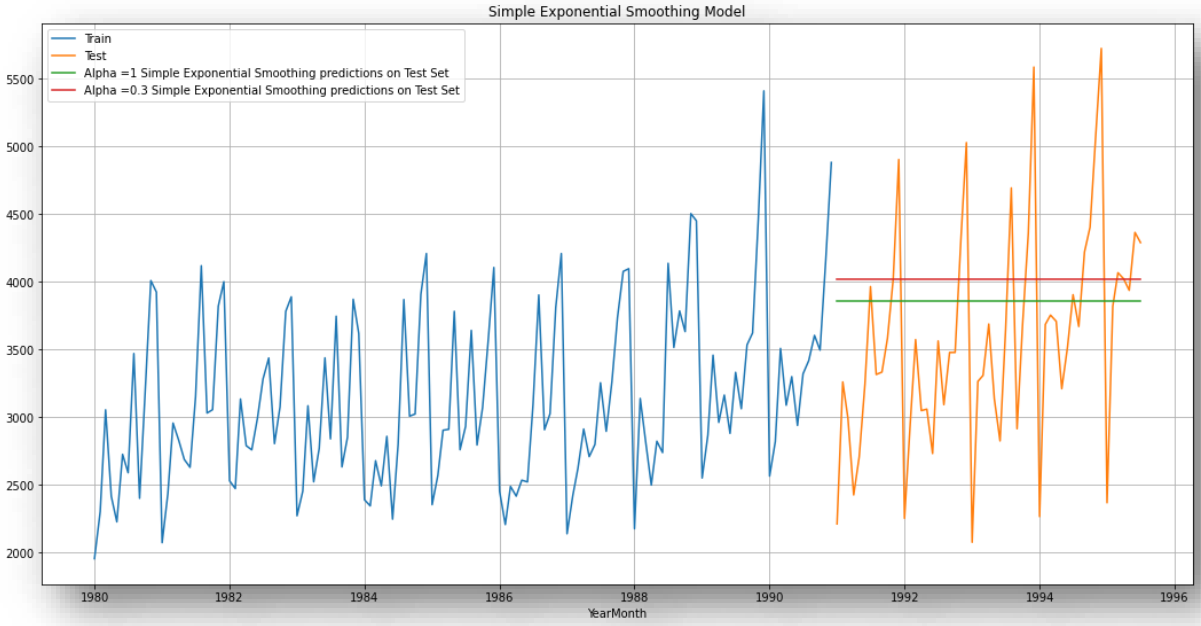

Figure 35: Simple Exponential Smoothing forecast on the test data (alpha = 0.3)

The RMSE score for alpha = 0.3 Simple Exponential Smoothing Model forecast on the Test Data 910.19.

Inference:

Simple Exponential Smoothing model does not take into consideration the trend or seasonality of the data. These models will be a better fit for a dataset without trend or seasonality as prominent parameters. Decomposition of our data has revealed that our data has both trend and seasonality. Therefore, this model would not be very accurate in forecasting.

Model 6: Double Exponential Smoothing (Holt's Model)

One of the drawbacks of the simple exponential smoothing is that the model does not do well in the presence of the trend. This model is an extension of SES known as Double Exponential Smoothing (DES) model which estimates two smoothing parameters.

This model is applicable when data has trend but no seasonality. One smoothing parameter α corresponds to the level series and second smoothing parameter β corresponds to the trend series. 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.

Figure 36 shows the optimum model parameters for the DES model. The alpha or the smoothing level is seen to be 0.438 and the beta or the smoothing trend is seen to be 0.083. This alpha and beta value is used to forecast the test data and the plot is shown in Figure 37.

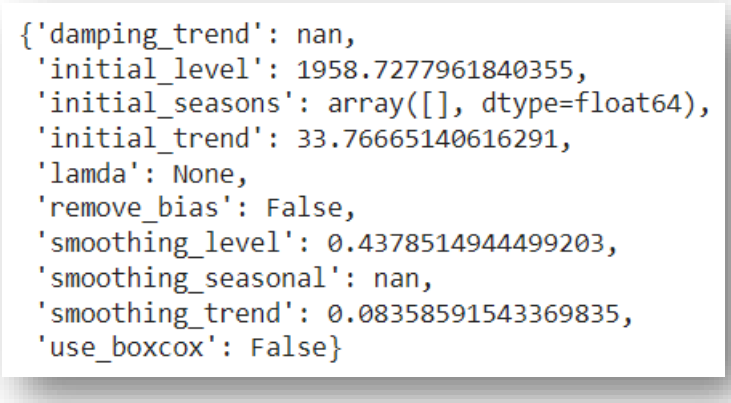

Figure 36: Double Exponential Smoothing Model optimum parameters

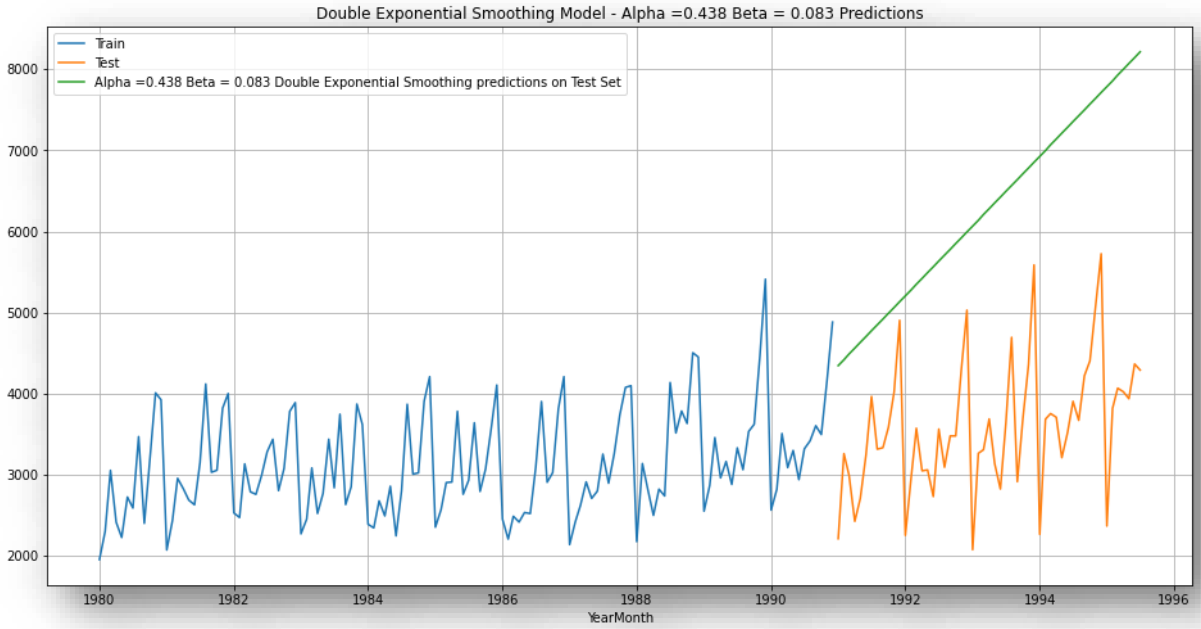

Figure 37: Double Exponential Smoothing forecast on the test data (alpha = 0.438, beta = 0.083)

```
For Alpha =0.438 Beta = 0.083 Double Exponential Smoothing Model forecast on the Test Data, RMSE is 2892.864
```

Figure 38: Double Exponential Smoothing model - RMSE score

The RMSE score for alpha = 0.438 and beta = 0.083 Double Exponential Smoothing Model forecast on the Test Data is 2892.864.

The model is tried by setting different alpha and beta values. The higher the alpha and beta value more weightage is given to the more recent observation. That means, what happened recently will happen again. We will run a loop with different alpha and beta values to understand which particular value works best on the test set. The best RMSE scores for the different combinations of alpha and beta values is shown in Table 20.

Table 20: Double Exponential Smoothing model best RMSE scores for different alpha and beta values

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	734.358128	6574.951181
8	0.4	0.3	738.383045	8561.385420
1	0.3	0.4	764.758634	9273.515812
16	0.5	0.3	741.869941	9764.656865
24	0.6	0.3	752.532546	10614.879977

It is observed from Table 20 that the alpha value 0.3 and beta value 0.3 has the least RMSE score of 6574.95. Therefore, the model is forecasted on the test data with alpha value of 0.3 and beta value of 0.3 and the plot is shown in Figure 39.

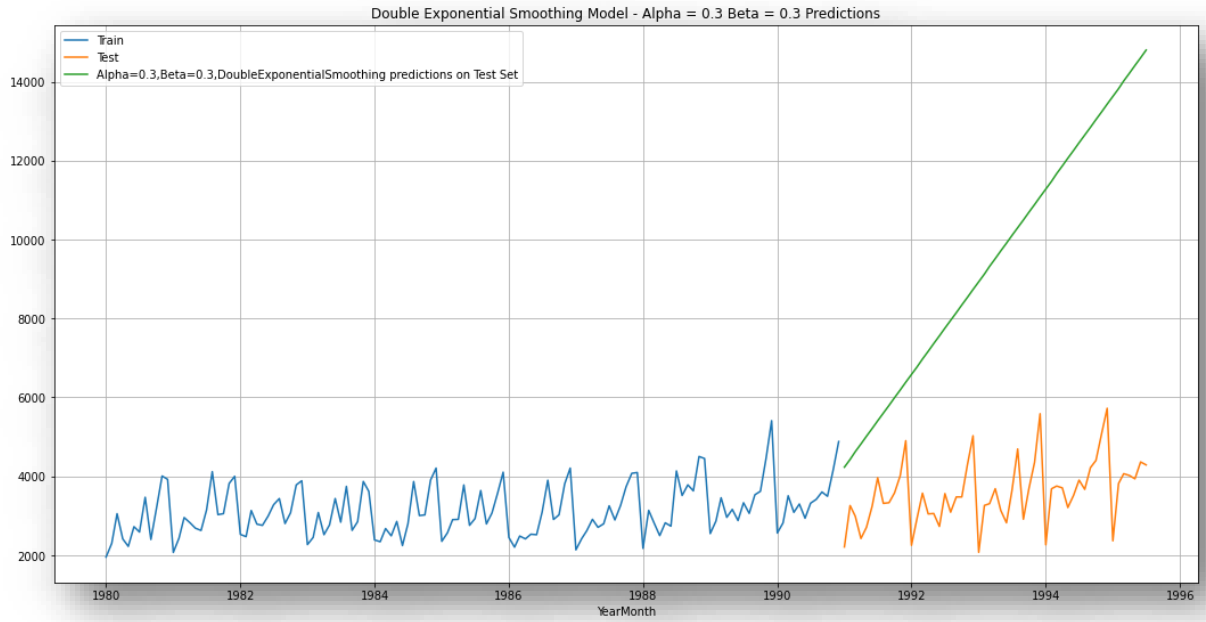

Figure 39: Double Exponential Smoothing forecast on the test data (alpha = 0.3, beta = 0.3)

Inference:

Comparing with the Simple Exponential Smoothing model, we see that the Double Exponential Smoothing has not performed better. This might be because of the fact that the Double Exponential Smoothing model has picked up the trend component as well but not seasonality. Decomposition of our data has revealed that our data has both trend and seasonality. Therefore, this model also would not be very accurate in forecasting.

Model 7: Triple Exponential Smoothing (Holt - winter's Model)

Triple Exponential Smoothing (TES) is an extension of Exponential Smoothing that explicitly adds support for seasonality to the univariate time series. This method is sometimes called Holt-Winters Exponential Smoothing, named for two contributors to the method: Charles Holt and Peter Winters. The TES model estimates three smoothing parameters.

This model is applicable when data has trend and seasonality. One smoothing parameter α corresponds to the level series, second smoothing parameter β corresponds to the trend series and the third smoothing parameter γ corresponds to the seasonality.

Figure 40 shows the optimum model parameters for the TES model. The alpha or the smoothing level is seen to be 0.111, the beta or the smoothing trend is seen to be 0.049 and the gamma or the smoothing seasonal is seen to be 0.230. This alpha, beta and gamma value is used to forecast the test data and the plot is shown in Figure 41.

```
{'damping_trend': nan,
 'initial_level': 2803.2031192879085,
 'initial_seasons': array([0.81675206, 0.85707329, 1.03845496, 0.9260439 , 0.95069866,
        0.97315248, 1.03766339, 1.25338534, 0.99255867, 1.07376893,
        1.35052981, 1.38008798]),
 'initial_trend': 15.090789924689997,
 'lamda': None,
 'remove_bias': False,
 'smoothing_level': 0.11109431519592447,
 'smoothing_seasonal': 0.23045135049306534,
 'smoothing_trend': 0.049376826867578195,
 'use_boxcox': False}
```

Figure 40: Triple Exponential Smoothing Model optimum parameters

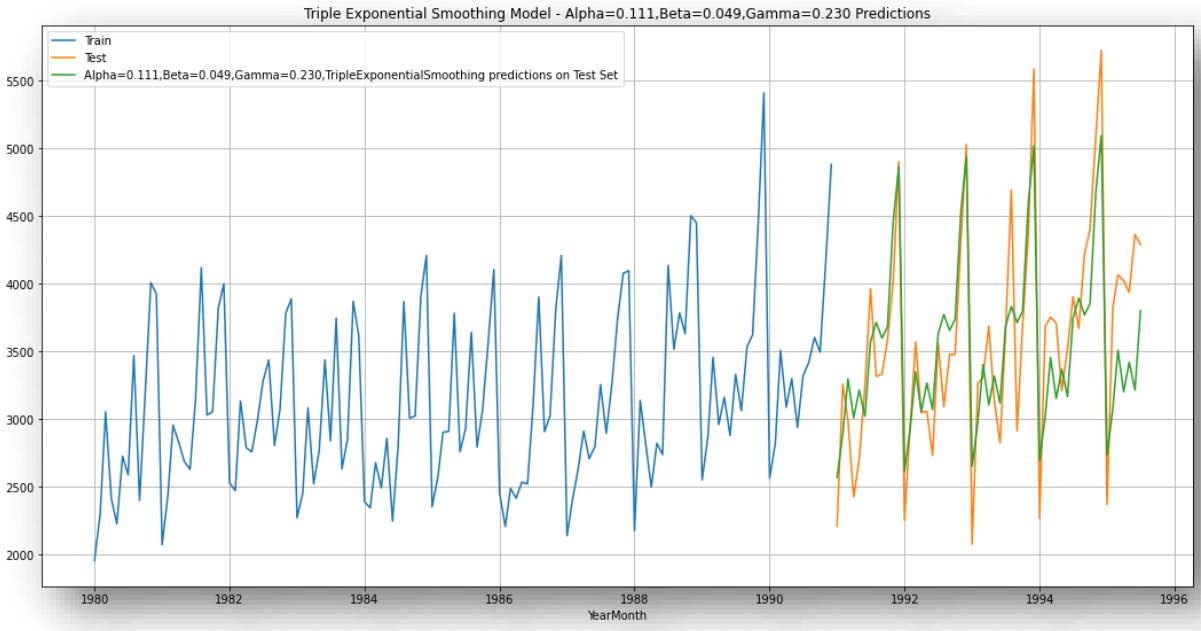

Figure 41: Triple Exponential Smoothing forecast on the test data (alpha = 0.111, beta = 0.049, gamma = 0.230)

```
For Alpha=0.111,Beta=0.049,Gamma=0.230, Triple Exponential Smoothing Model forecast on the Test Data,  RMSE is 447.623
```

Figure 42: Triple Exponential Smoothing model - RMSE score

The RMSE score for alpha = 0.111, beta = 0.049 and gamma = 0.230 Triple Exponential Smoothing Model forecast on the Test Data is 447.623.

The model is tried by setting different alpha, beta and gamma values. The higher the alpha, beta and gamma values more weightage is given to the more recent observation. That means, what happened recently will happen again. We will run a loop with different alpha, beta and gamma values to understand which particular value works best on the test set. The best RMSE scores for the different combinations of alpha, beta and gamma values is shown in Table 21.

Table 21: Triple Exponential Smoothing model best RMSE scores for different alpha and beta values

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
64	0.4	0.3	0.3	373.320057	453.599111
16	0.3	0.5	0.3	413.619443	531.629744
220	0.6	0.6	0.7	555.505445	532.791333
276	0.7	0.5	0.7	611.037427	532.937714
74	0.4	0.4	0.5	455.538886	810.365058

It is observed from Table 21 that the alpha value 0.4, beta value 0.3 and gamma value 0.3 has the least RMSE score of 453.59. Therefore, the model is forecasted on the test data with alpha value of 0.4, beta value 0.3 and gamma value 0.3 and the plot is shown in Figure 43.

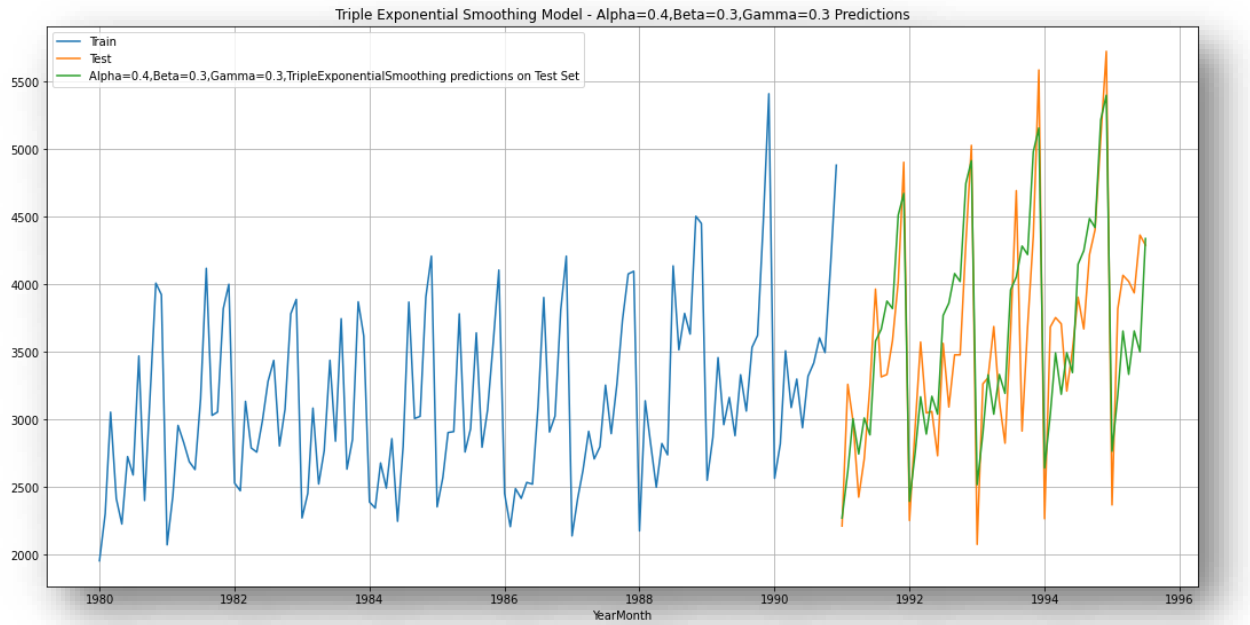

Figure 43: Triple Exponential Smoothing forecast on the test data (alpha = 0.4, beta = 0.3, gamma = 0.3)

Inference:

The Triple Exponential Smoothing has performed the best compared to all the models. This is due to the fact that the Triple Exponential Smoothing is picking up the seasonal component as well as the trend. Decomposition of our data has revealed that our data has both trend and seasonality. Therefore, this model is able to forecast the test data accurately.

Comparison of exponential smoothing models:

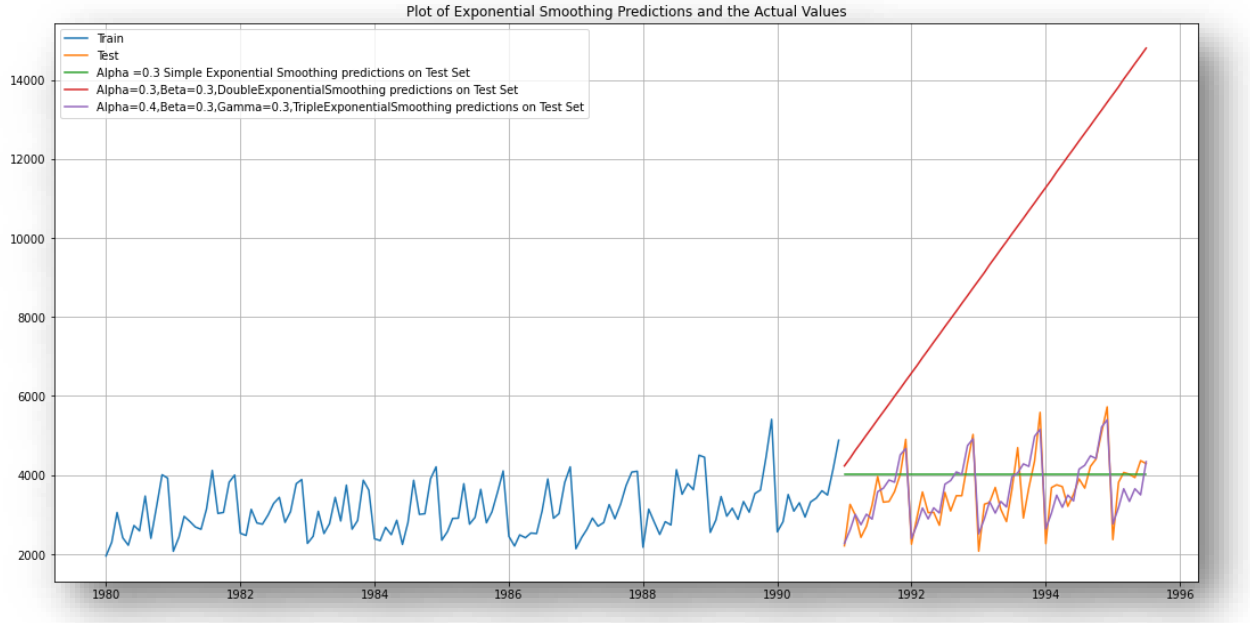

Figure 44: Plot of Exponential Smoothing Predictions and the Actual Values

Inference from the exponential smoothing models:

Decomposition of our data has revealed that our data has both trend and seasonality. Therefore, Triple Exponential Smoothing model is the best to forecast the test data accurately. However, Simple Exponential Smoothing and the Double Exponential Smoothing models are built in this dataset to get an idea of how the three types of models perform.

The differences between the different exponential models are very clear as each of the models have different parameters taken into consideration. The Triple Exponential Smoothing model seems to be the best fit as our data has trend and seasonality. The RMSE scores for the Triple Exponential Smoothing model are also low meaning that it will be a better fit among the other exponential models for this particular dataset.

This is also seen in Figure 44. It shows that the forecast predicted using the SES and DES are a straight line and the prediction is following a trend. The DES model forecast tells us that the data is on an increasing trend whereas the forecast using the SES model conveys that there is no variation in the prediction i.e., that is the prediction for today is the same as the prediction on the entire test data as it is a straight line. On comparing the forecast from the three models with the actual test data we can see that the prediction is the most reliable with TES model as the forecast is very similar to the test data.

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.

Auto-regression (AR) means regression of a variable on itself. One of the fundamental assumptions of an AR model is that the time series is assumed to be a stationary process. When the time series data is not stationary, then we have to convert the non-stationary time-series data to stationary time-series before applying AR models.

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.

We would want the series to be stationary for building ARIMA models and thus we would want the p-value of this test to be less than the α value.

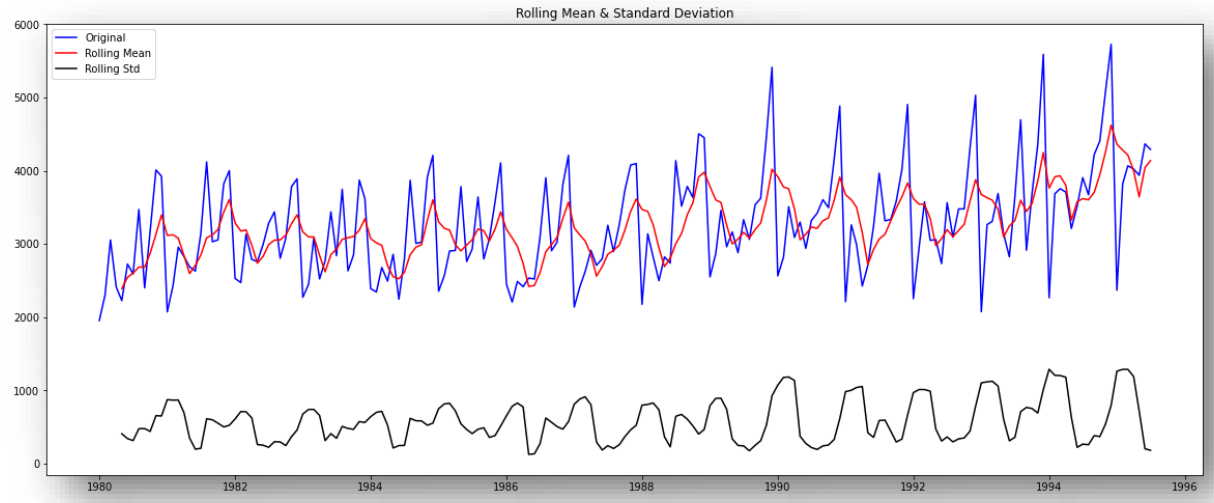

Figure 45: Augmented Dickey-Fuller test – 0 order difference

```
Results of Dickey-Fuller Test:
Test Statistic                  -0.424986
p-value                          0.986102
#Lags Used                      12.000000
Number of Observations Used    174.000000
Critical Value (1%)             -4.011764
Critical Value (5%)             -3.436029
Critical Value (10%)            -3.142044
dtype: float64
```

Figure 46: Augmented Dickey-Fuller test results - 0 order difference

From Figure 46, we can notice that the p-value is 0.986102. This value is not less than 0.05. Hence, we fail to reject the null hypothesis. Therefore, the test is carried out again by taking a difference of order 1 and checking whether the Time Series is stationary or not.

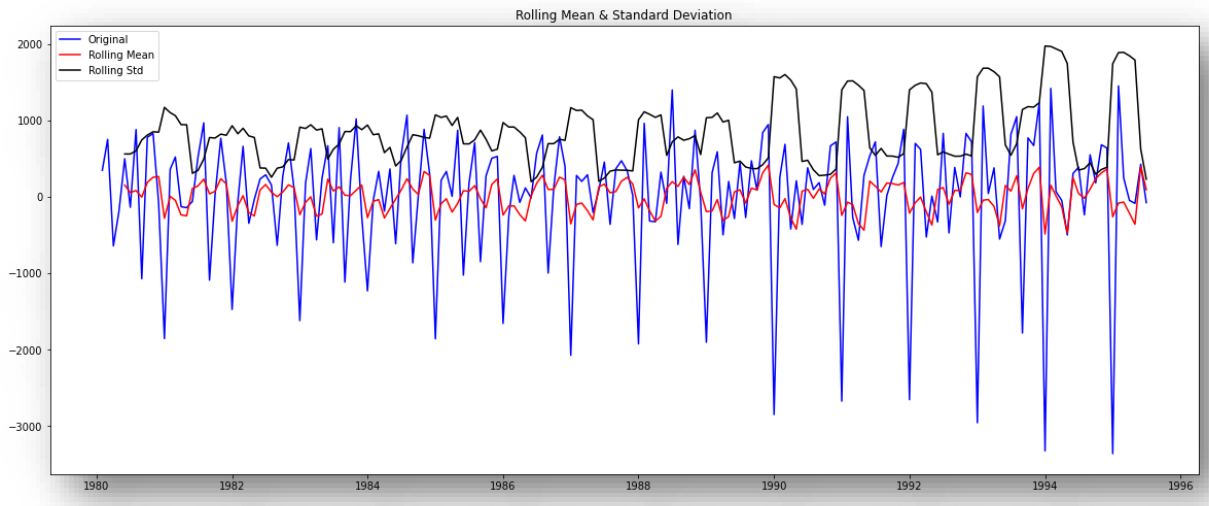

Figure 47: Augmented Dickey-Fuller test – 1 order difference

```
Results of Dickey-Fuller Test:
Test Statistic                -9.481347e+00
p-value                        3.053709e-14
#Lags Used                     1.100000e+01
Number of Observations Used    1.740000e+02
Critical Value (1%)           -4.011764e+00
Critical Value (5%)           -3.436029e+00
Critical Value (10%)          -3.142044e+00
dtype: float64
```

Figure 48: Augmented Dickey-Fuller test results - 1 order difference

From Figure 48, we can notice that the p-value is 3.053709e-14. This value is less than 0.05. Hence, we reject the null hypothesis. We can now conclude that the **time series data is stationary**.

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.

Auto ARIMA Model:

An ARIMA (p,d,q) model consists of the Auto-Regressive (AR) part and the Moving Average (MA) part after we have made the Time Series stationary by taking the correct degree/order of differencing.

The AR order is selected by looking at where the PACF plot cuts-off (for appropriate confidence interval bands) and the MA order is selected by looking at where the ACF plots cuts-off (for appropriate confidence interval bands). The correct degree or order of difference gives us the value of ‘d’ while the ‘p’ value is for the order of the AR model and the ‘q’ value is for the order of the MA model. This is the Box-Jenkins methodology for building the ARIMA models. Figure 49 shows the examples of the parameter combinations for the model.

```
Examples of the parameter combinations for the Model
Model: (0, 1, 0)
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

Figure 49: Examples of the parameter combinations for the ARIMA Model

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 orders respectively which gives us the lowest AIC value. Lower the AIC better is the model. Table 22 shows the parameter combination with the lowest AIC value.

Coding languages tries different orders of 'p' and 'q' to arrive to this conclusion. Remember, even for such a way of choosing the 'p' and 'q' values, we must make sure that the series is stationary.

The formula for calculating the AIC is:

$$\mathbf{2k - 2\ln(L)}$$

where k is the number of parameters to be estimated and L is the likelihood estimation.

Table 22: Best parameter combinations and corresponding AIC Values

	param	AIC
2	(0, 1, 2)	2056.489263
6	(1, 1, 2)	2056.715682
3	(0, 1, 3)	2056.831789
11	(2, 1, 3)	2057.090828
13	(3, 1, 1)	2058.304546

From Table 22, it is observed that the parameter (0,1,2) has the lowest AIC value. Therefore parameter (0,1,2) is used to calculate the summary of ARIMA shown in Figure 50 and plot the diagnostics shown in Figure 51.

SARIMAX Results

Dep. Variable:	SoftDrinkProduction	No. Observations:	132
Model:	ARIMA(0, 1, 2)	Log Likelihood	-1025.245
Date:	Sun, 03 Apr 2022	AIC	2056.489
Time:	12:20:48	BIC	2065.115
Sample:	01-01-1980 - 12-01-1990	HQIC	2059.994
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.5407	0.085	-6.392	0.000	-0.707	-0.375
ma.L2	-0.3913	0.113	-3.475	0.001	-0.612	-0.171
sigma2	3.572e+05	4.62e+04	7.725	0.000	2.67e+05	4.48e+05

Ljung-Box (L1) (Q):	0.61	Jarque-Bera (JB):	0.39
Prob(Q):	0.44	Prob(JB):	0.82
Heteroskedasticity (H):	1.31	Skew:	-0.13
Prob(H) (two-sided):	0.37	Kurtosis:	2.91

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

Figure 50: ARIMA Model Results

From Figure 50, it is observed that the p value of coefficients ma.L1 and ma.L2 are 0.000 and 0.001 which means that these are pretty significant.

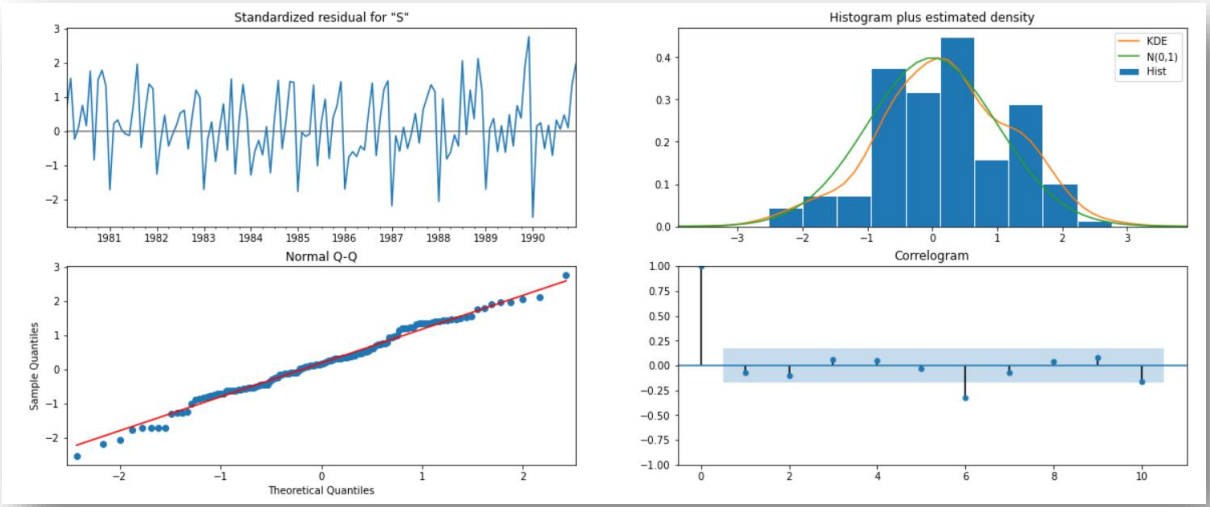

Figure 51: Diagnostic Plot - ARIMA

```
RMSE for ARIMA: 831.6158513483251
MAPE for ARIMA: 18.494207971041433
```

Figure 52: ARIMA RMSE and MAPE results

The RMSE value for the ARIMA is calculated to be 831.62 shown in Figure 52.

Auto SARIMA Model:

For a Seasonal Auto-Regressive Integrated Moving Average (p,d,q)(P,D,Q)F Model or SARIMA model we have to take care of four parameters such as AR (p), MA (q), Seasonal AR (P) and Seasonal MA (Q) with the correct of differencing (d) and seasonal differencing (D). Here, the 'F' parameter indicates the seasonality/seasonal effects over a particular period. We can follow the Box-Jenkins method over here as well to decide the 'p', 'q', 'P' and 'Q' values. For deciding the 'P' and 'Q' values, we need to look at the PACF and the ACF plots respectively at lags which are the multiple of 'F' and see where this cut-off (for appropriate confidence interval bands). Figure 53 shows the examples of the parameter combinations for the SARIMA model.

```
Examples of the parameter combinations for the Model are
Model: (0, 1, 1)(0, 0, 1, 6)
Model: (0, 1, 2)(0, 0, 2, 6)
Model: (0, 1, 3)(0, 0, 3, 6)
Model: (1, 1, 0)(1, 0, 0, 6)
Model: (1, 1, 1)(1, 0, 1, 6)
Model: (1, 1, 2)(1, 0, 2, 6)
Model: (1, 1, 3)(1, 0, 3, 6)
Model: (2, 1, 0)(2, 0, 0, 6)
Model: (2, 1, 1)(2, 0, 1, 6)
Model: (2, 1, 2)(2, 0, 2, 6)
Model: (2, 1, 3)(2, 0, 3, 6)
Model: (3, 1, 0)(3, 0, 0, 6)
Model: (3, 1, 1)(3, 0, 1, 6)
Model: (3, 1, 2)(3, 0, 2, 6)
Model: (3, 1, 3)(3, 0, 3, 6)
```

Figure 53: Examples of the parameter combinations for the SARIMA Model

For the SARIMA models, we can also estimate 'p', 'q', 'P' and 'Q' by looking at the lowest AIC values calculated and shown in Table 23. The seasonal parameter 'F' can be determined by looking at the ACF plot shown in Figure 54. The ACF plot is expected to show a spike at multiples of 'F' thereby indicating a presence of seasonality. In Figure 54, the plot shows a spike at multiples of 6. Therefore, the seasonality is taken to be 6. Also, for seasonal models, the ACF and the PACF plots are going to behave a bit different and they will not always continue to decay as the number of lags increase.

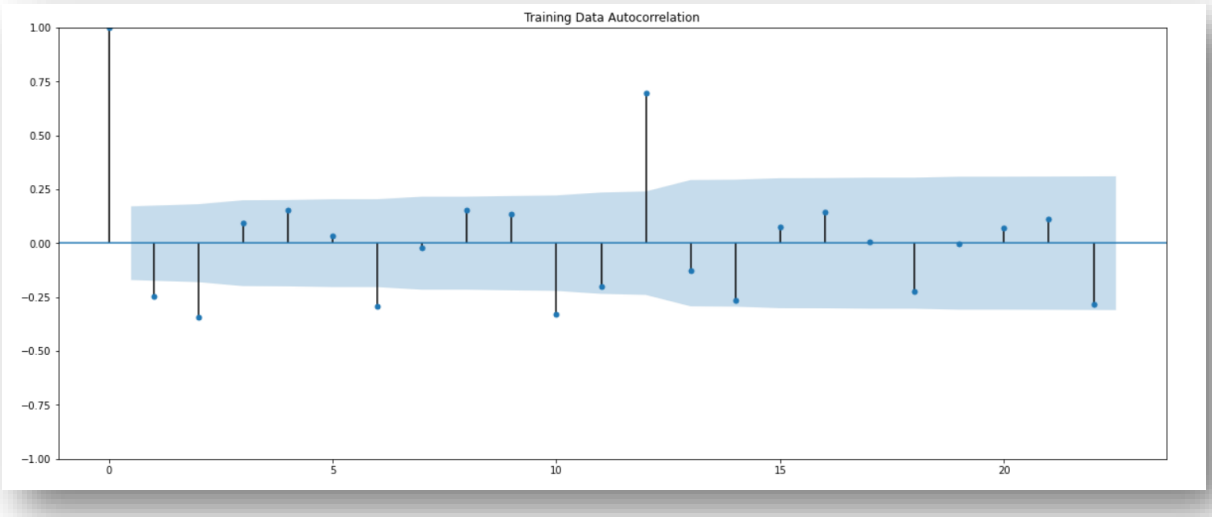

Figure 54: Training Data Autocorrelation (ACF) Plot

Table 23: Best parameter combinations and corresponding AIC Values

	param	seasonal	AIC
255	(3, 1, 3)	(3, 0, 3, 6)	1589.761104
59	(0, 1, 3)	(2, 0, 3, 6)	1590.244407
191	(2, 1, 3)	(3, 0, 3, 6)	1591.011959
123	(1, 1, 3)	(2, 0, 3, 6)	1591.284672
63	(0, 1, 3)	(3, 0, 3, 6)	1592.243834

From Table 23, it is observed that the combination of parameter (3,1,3) and seasonal (3,0,3,6) has the lowest AIC value. Therefore, this combination is used to calculate the summary of SARIMA shown in Figure 55 and plot the diagnostics shown in Figure 56.

SARIMAX Results

Dep. Variable:	SoftDrinkProduction	No. Observations:	132
Model:	SARIMAX(3, 1, 3)x(3, 0, 3, 6)	Log Likelihood	-781.881
Date:	Sun, 03 Apr 2022	AIC	1589.761
Time:	12:24:17	BIC	1624.749
Sample:	01-01-1980 - 12-01-1990	HQIC	1603.950
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.8582	0.139	6.171	0.000	0.586	1.131
ar.L2	-1.0003	0.104	-9.660	0.000	-1.203	-0.797
ar.L3	0.2194	0.130	1.686	0.092	-0.036	0.474
ma.L1	-1.8790	0.060	-31.274	0.000	-1.997	-1.761
ma.L2	1.8614	0.284	6.561	0.000	1.305	2.417
ma.L3	-1.0855	0.187	-5.817	0.000	-1.451	-0.720
ar.S.L6	-0.1188	0.577	-0.206	0.837	-1.249	1.011
ar.S.L12	1.0021	0.035	28.773	0.000	0.934	1.070
ar.S.L18	0.1372	0.586	0.234	0.815	-1.011	1.285
ma.S.L6	0.0535	0.580	0.092	0.926	-1.082	1.189
ma.S.L12	-0.6039	0.138	-4.391	0.000	-0.874	-0.334
ma.S.L18	-0.1253	0.389	-0.322	0.748	-0.888	0.637
sigma2	7.618e+04	8.09e-06	9.42e+09	0.000	7.62e+04	7.62e+04

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	10.59
Prob(Q):	0.98	Prob(JB):	0.01
Heteroskedasticity (H):	1.77	Skew:	0.38
Prob(H) (two-sided):	0.09	Kurtosis:	4.33

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.27e+26. Standard errors may be unstable.

Figure 55: SARIMA Model Results

From Figure 55 the p value of ar.S.L12 and ma.S.L12 is found to be less than 0.05 which makes them pretty significant.

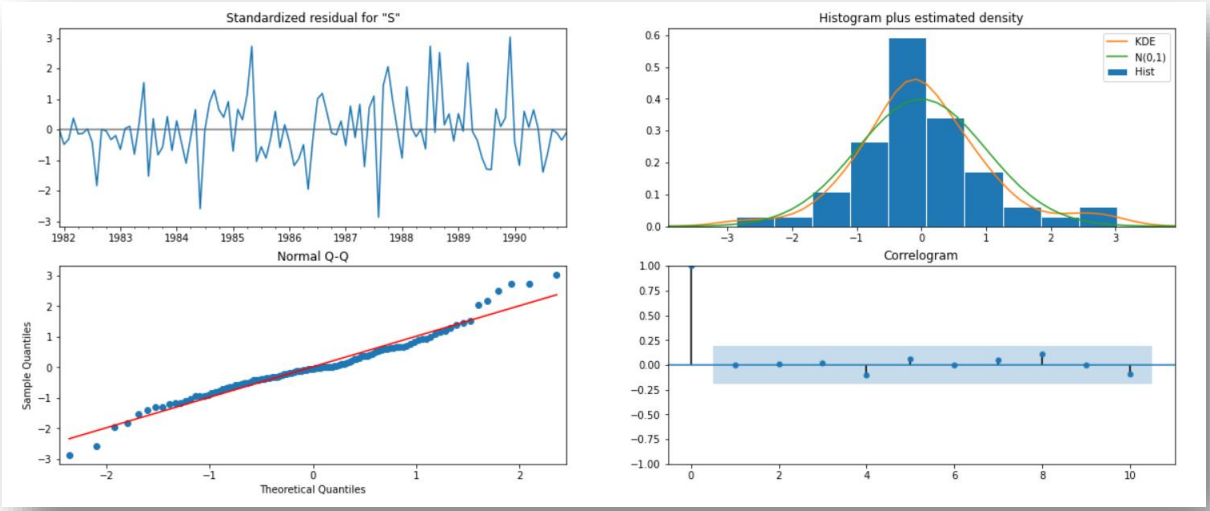

Figure 56: Diagnostic Plot - SARIMA

```
RMSE for SARIMA: 428.6985141635827
MAPE for SARIMA: 10.865842021565404
```

Figure 57: SARIMA RMSE and MAPE results

The RMSE value for the SARIMA model is calculated to be 428.70 shown in Figure 57.

Note: Here, there is both trend and seasonality in the data. Therefore, we should have directly gone for the SARIMA and manual SARIMA model but ARIMA and manual ARIMA models are built over here to get an idea of how the two types of models compare in this case.

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.

Manual ARIMA Model:

In the manual ARIMA (p,d,q) model, the AR order is selected by looking at where the PACF plot cuts-off (for appropriate confidence interval bands) and the MA order is selected by looking at where the ACF plots cuts-off (for appropriate confidence interval bands). The correct degree or order of difference gives us the value of 'd' while the 'p' value is for the order of the AR model and the 'q' value is for the order of the MA model. This is the Box-Jenkins methodology for building the ARIMA models. Figure 58 and Figure 59 is used for selecting the p and q values.

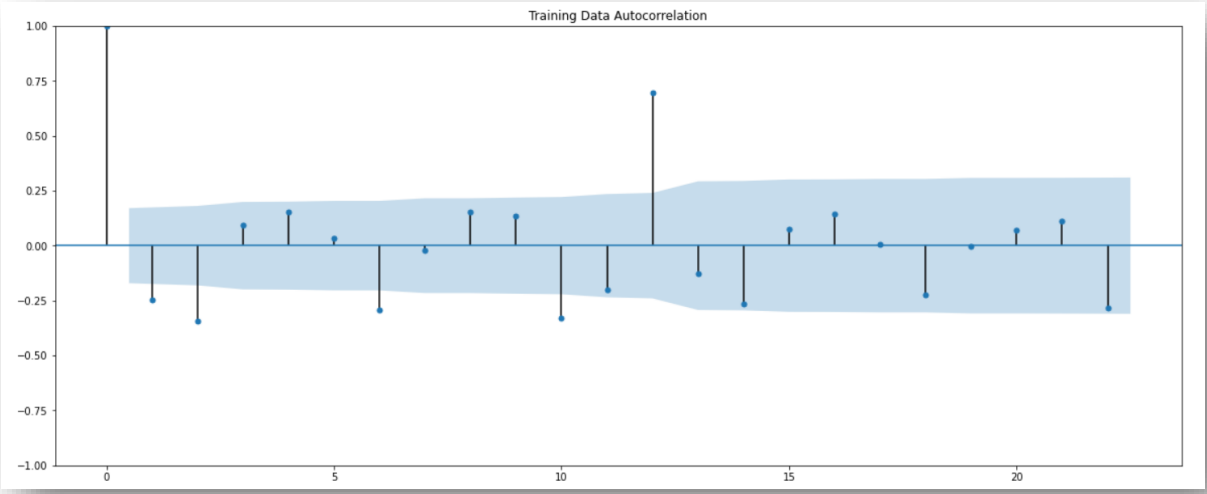

Figure 58: Training Data Autocorrelation (ACF) Plot

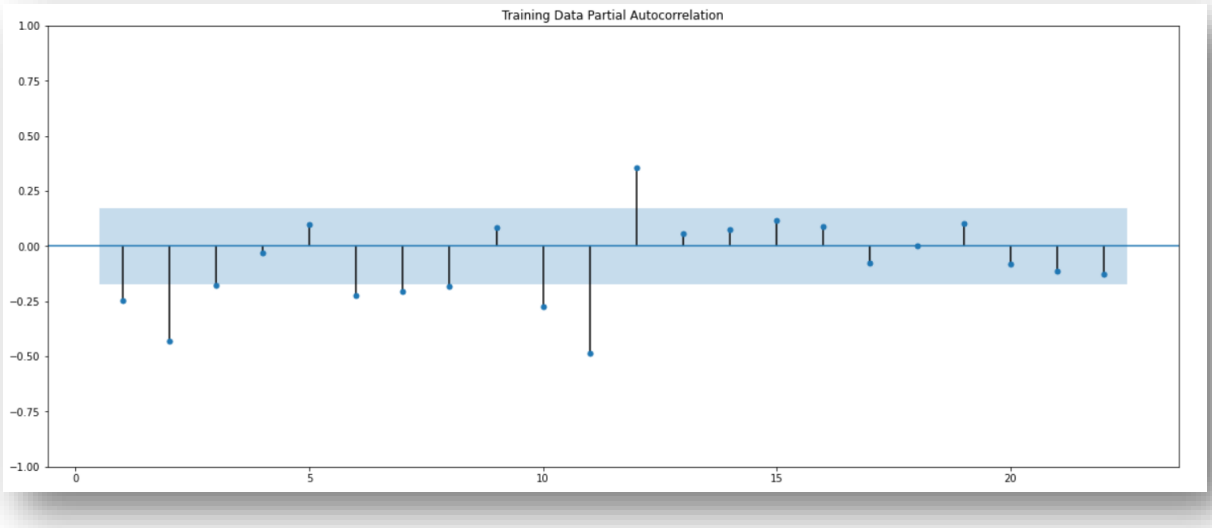

Figure 59: Training Data Partial Autocorrelation (PACF) Plot

Here, we have taken alpha=0.05.

- The Auto-Regressive parameter in an ARIMA model is 'p' which comes from the significant lag before which the PACF plot cuts-off to 3.
- The Moving-Average parameter in an ARIMA model is 'q' which comes from the significant lag before the ACF plot cuts-off to 2.

By looking at Figure 58 and Figure 59, we will take the value of p and q to be 3 and 2 respectively.

Therefore, the parameter of the manual ARIMA model will be (3,1,2). The parameter (3,1,2) is used to calculate the summary of ARIMA shown in Figure 60 and plot the diagnostics shown in Figure 61.

SARIMAX Results

Dep. Variable:	SoftDrinkProduction	No. Observations:	132
Model:	ARIMA(3, 1, 2)	Log Likelihood	-1024.340
Date:	Sun, 03 Apr 2022	AIC	2060.680
Time:	14:44:21	BIC	2077.931
Sample:	01-01-1980 - 12-01-1990	HQIC	2067.690
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3216	0.460	-0.700	0.484	-1.223	0.579
ar.L2	-0.0001	0.189	-0.001	0.999	-0.371	0.370
ar.L3	-0.0227	0.218	-0.104	0.917	-0.449	0.404
ma.L1	-0.2633	0.453	-0.581	0.561	-1.151	0.624
ma.L2	-0.6405	0.400	-1.599	0.110	-1.425	0.144
sigma2	3.514e+05	4.97e+04	7.066	0.000	2.54e+05	4.49e+05

Ljung-Box (L1) (Q):	0.07	Jarque-Bera (JB):	0.43
Prob(Q):	0.80	Prob(JB):	0.81
Heteroskedasticity (H):	1.29	Skew:	0.04
Prob(H) (two-sided):	0.40	Kurtosis:	2.73

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

Figure 60: Manual ARIMA Model Results

From Figure 60, it is observed that the p value of coefficients ma.L1 and ma.L2 are 0.561 and 0.110.

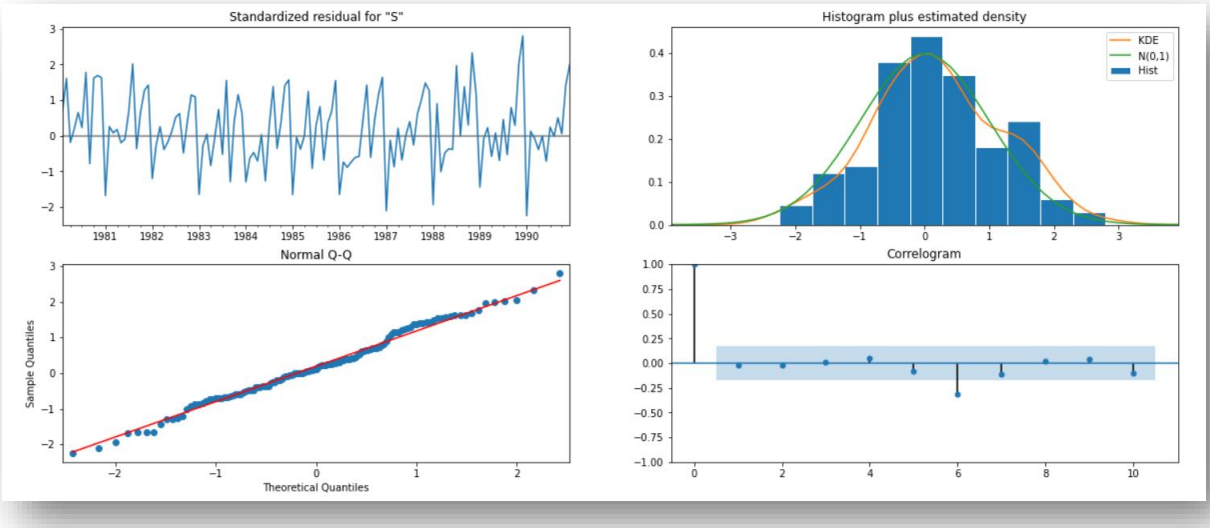

Figure 61: Diagnostic Plot - Manual ARIMA

```
RMSE for manual ARIMA: 822.2174505020024
MAPE for manual ARIMA: 18.363067569135197
```

Figure 62: Manual ARIMA RMSE and MAPE results

The RMSE value for the manual ARIMA model is calculated to be 822.22 shown in Figure 62.

Manual SARIMA Model:

For a Seasonal Auto-Regressive Integrated Moving Average (p,d,q)(P,D,Q)F Model or SARIMA model we have to take care of four parameters such as AR (p), MA (q), Seasonal AR (P) and Seasonal MA (Q) with the correct of differencing (d) and seasonal differencing (D). Here, the 'F' parameter indicates the seasonality/seasonal effects over a particular period. We can follow the Box-Jenkins method over here as well to decide the 'p', 'q', 'P' and 'Q' values. For deciding the 'P' and 'Q' values, we need to look at the PACF and the ACF plots respectively at lags which are the multiple of 'F' and see where this cut-off (for appropriate confidence interval bands).

The seasonal parameter 'F' can be determined by looking at the ACF plot shown in Figure 63. The ACF plot is expected to show a spike at multiples of 'F' thereby indicating a presence of seasonality. In Figure 63, the plot shows significant spikes at multiples of 6. Other spikes in between are not that significant. Therefore, the seasonality is taken to be 6.

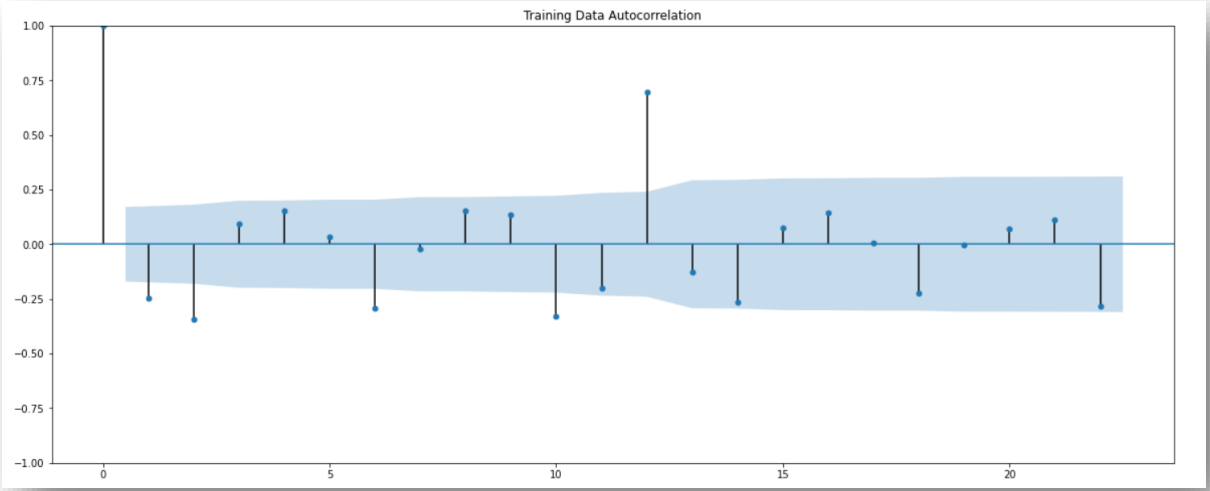

Figure 63: Training Data Autocorrelation (ACF) Plot

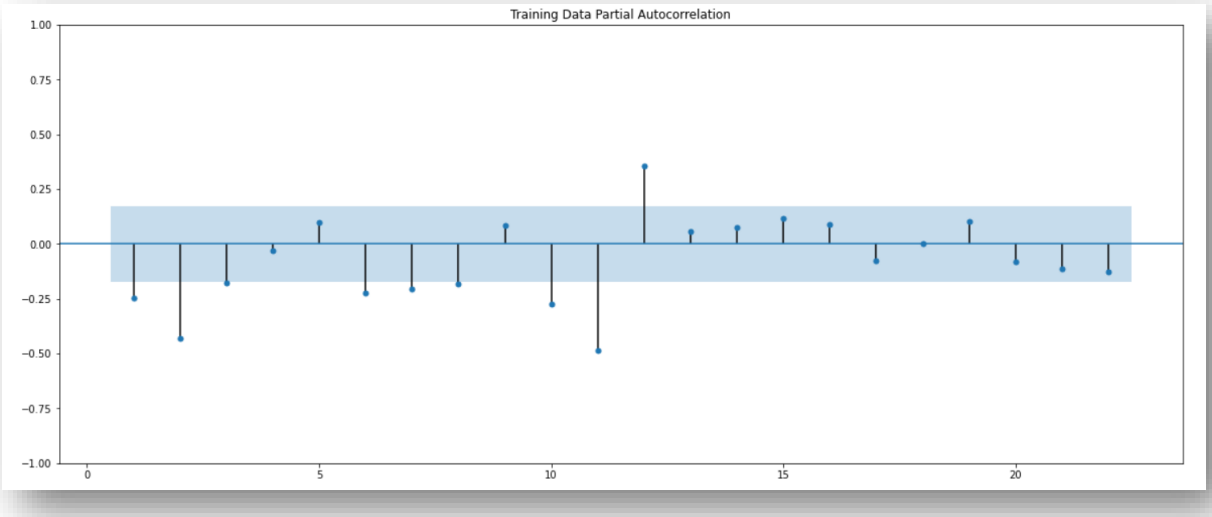

Figure 64: Training Data Partial Autocorrelation (PACF) Plot

Here, we have taken alpha=0.05.

The seasonal period is decided to be 6 by look at the Figure 63. We are taking the p value to be 3 and the q value also to be 2 as the parameters same as the ARIMA model.

- The Auto-Regressive parameter in a SARIMA model is 'P' which comes from the significant lag after which the PACF plot cuts-off to 0.
- The Moving-Average parameter in a SARIMA model is 'Q' which comes from the significant lag after which the ACF plot cuts-off to 0.

Therefore, the parameter of the manual SARIMA model will be (3,1,2) (0,0,0,6). The parameter (3,1,2) (0,0,0,6) is used to calculate the summary of SARIMA shown in Figure 65 and plot the diagnostics shown in Figure 66.

SARIMAX Results

Dep. Variable:	SoftDrinkProduction	No. Observations:	132
Model:	SARIMAX(3, 1, 2)	Log Likelihood	-998.726
Date:	Sun, 03 Apr 2022	AIC	2009.453
Time:	14:44:23	BIC	2026.565
Sample:	01-01-1980 - 12-01-1990	HQIC	2016.406
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.3108	0.665	0.467	0.640	-0.993	1.614
ar.L2	-0.2274	0.260	-0.874	0.382	-0.737	0.283
ar.L3	0.1461	0.171	0.853	0.394	-0.190	0.482
ma.L1	-0.9248	0.673	-1.373	0.170	-2.245	0.395
ma.L2	-0.0296	0.644	-0.046	0.963	-1.291	1.232
sigma2	3.46e+05	4.9e+04	7.061	0.000	2.5e+05	4.42e+05

Ljung-Box (L1) (Q):	0.05	Jarque-Bera (JB):	0.80
Prob(Q):	0.81	Prob(JB):	0.67
Heteroskedasticity (H):	1.62	Skew:	0.15
Prob(H) (two-sided):	0.12	Kurtosis:	2.76

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

Figure 65: Manual SARIMA Model Results

From Figure 65 the p value of ar.L1 and ma.L1 is found to be 0.640 and 0.170.

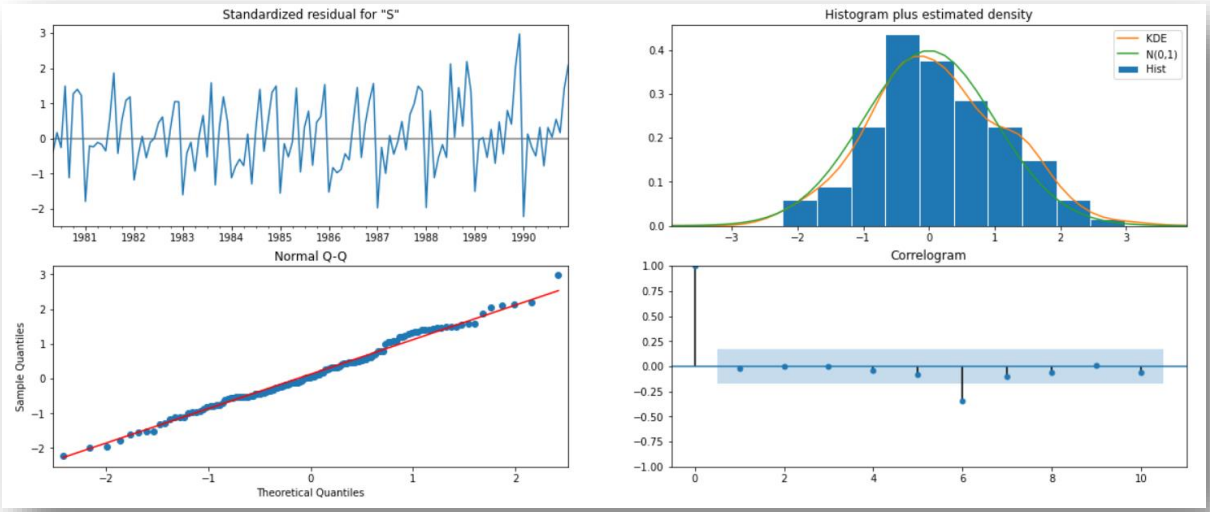

Figure 66: Diagnostic Plot - Manual SARIMA

```
RMSE for manual SARIMA: 833.1223123696125
MAPE for manual SARIMA: 18.427840261880604
```

Figure 67: Manual SARIMA RMSE and MAPE results

The RMSE value for the manual SARIMA model is calculated to be 833.12 shown in Figure 67.

Note: Here, there is both trend and seasonality in the data. Therefore, we should have directly gone for the SARIMA and manual SARIMA model but ARIMA and manual ARIMA models are built over here to get an idea of how the two types of models compare in this case.

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

Table 24: RMSE Values for the various models

	Test RMSE
RegressionOnTime	775.807810
NaiveModel	1519.259233
SimpleAverageModel	934.353358
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827
Alpha=0.216,SimpleExponentialSmoothing	847.635259
Alpha=0.3,SimpleExponentialSmoothing	910.187416
Alpha=0.438 Beta = 0.083 ,DoubleExponentialSmoothing	2892.864115
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	6574.951181
Alpha=0.111,Beta=0.049,Gamma=0.230,TripleExponentialSmoothing	447.622837
Alpha=0.4,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	453.599111

Table 25: RMSE Values for ARIMA and SARIMA Models

	RMSE	MAPE
ARIMA(0,1,2)	831.615851	18.494208
SARIMA(3,1,3)(3,0,3,6)	428.698514	10.865842
ACF&PACF - ARIMA(3,1,2)	822.217451	18.363068
ACF& PACF - SARIMA(3,1,2)(0,0,0,6)	833.122312	18.427840

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.

Table 26: Sorted RMSE Values for the various models

Sorted by RMSE values on the Test Data:

	Test RMSE
Alpha=0.111,Beta=0.049,Gamma=0.230,TripleExponentialSmoothing	447.622837
Alpha=0.4,Beta=0.3,Gamma=0.3,TripleExponentialSmoothing	453.599111
2pointTrailingMovingAverage	556.725418
4pointTrailingMovingAverage	687.181726
6pointTrailingMovingAverage	710.513877
9pointTrailingMovingAverage	735.889827
RegressionOnTime	775.807810
Alpha=0.216,SimpleExponentialSmoothing	847.635259
Alpha=0.3,SimpleExponentialSmoothing	910.187416
SimpleAverageModel	934.353358
NaiveModel	1519.259233
Alpha=0.438 Beta = 0.083 ,DoubleExponentialSmoothing	2892.864115
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	6574.951181

Table 27: Sorted RMSE Values for ARIMA and SARIMA models

Sorted by RMSE values on the Test Data:

	RMSE	MAPE
SARIMA(3,1,3)(3,0,3,6)	428.698514	10.865842
ACF&PACF - ARIMA(3,1,2)	822.217451	18.363068
ARIMA(0,1,2)	831.615851	18.494208
ACF& PACF - SARIMA(3,1,2)(0,0,0,6)	833.122312	18.427840

From Table 26 and Table 27, the optimum models are chosen to be triple exponential smoothing model with alpha = 0.111, beta = 0.049 and gamma = 0.230 and SARIMA model with parameters (3,1,3) (3,0,3,6). There is both trend and seasonality in the data and also the RMSE values for both triple exponential smoothing model and SARIMA model is seen to be the least compared to the other models.

Forecast using Triple Exponential Smoothing Model:

The parameters alpha (smoothing level) = 0.111, beta (smoothing trend) = 0.049 and gamma (smoothing seasonal) = 0.230 are the optimum values that are taken to build the model and forecast the future.

```
RMSE for the full model (Triple Exponential Smoothing): 333.471077684034
```

Figure 68: RMSE for the full model using Triple Exponential Smoothing model

The RMSE value for the full model is calculated to be **333.47** which is seen in Figure 68.

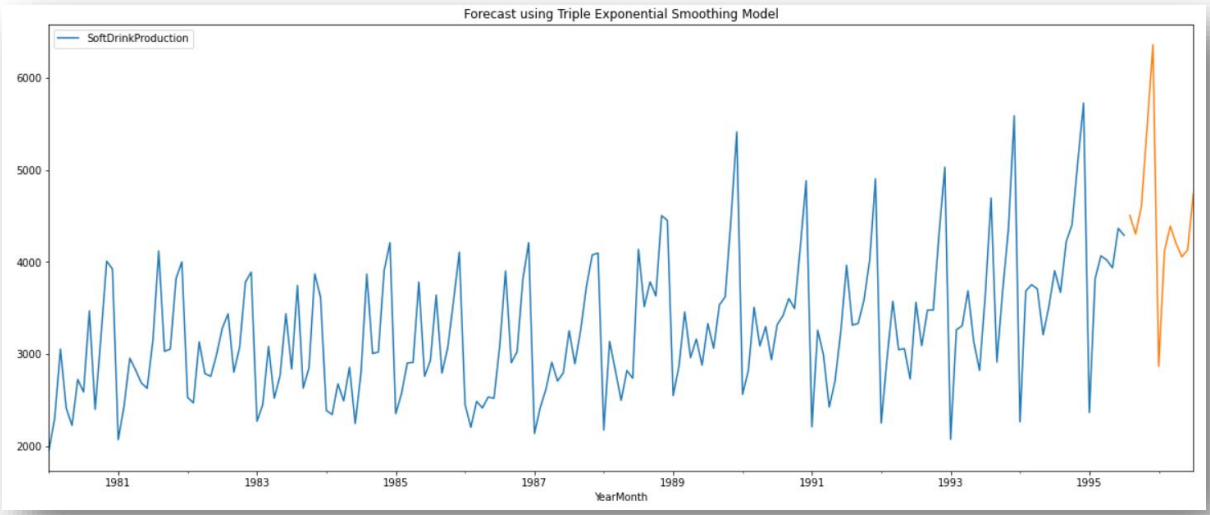

Figure 69: Forecast using the Triple Exponential Smoothing Model

Figure 69 shows the forecast of the sales of soft drink production for the upcoming 12 months. The orange part depicts the forecasted data from August 1995 to July 1996.

Table 28: Data with upper and lower confidence bands

	lower_CI	prediction	upper_ci
1995-08-01	3851.467097	4506.118588	5160.770079
1995-09-01	3649.315240	4303.966731	4958.618222
1995-10-01	3946.991594	4601.643085	5256.294576
1995-11-01	4816.844083	5471.495574	6126.147065
1995-12-01	5703.777992	6358.429483	7013.080974

The upper and lower confidence bands at 95% confidence level is calculated and shown in Table 28. The multiplier is taken to be 1.96 as we want to plot with respect to 95% confidence intervals. The plotted graph is shown in Figure 70.

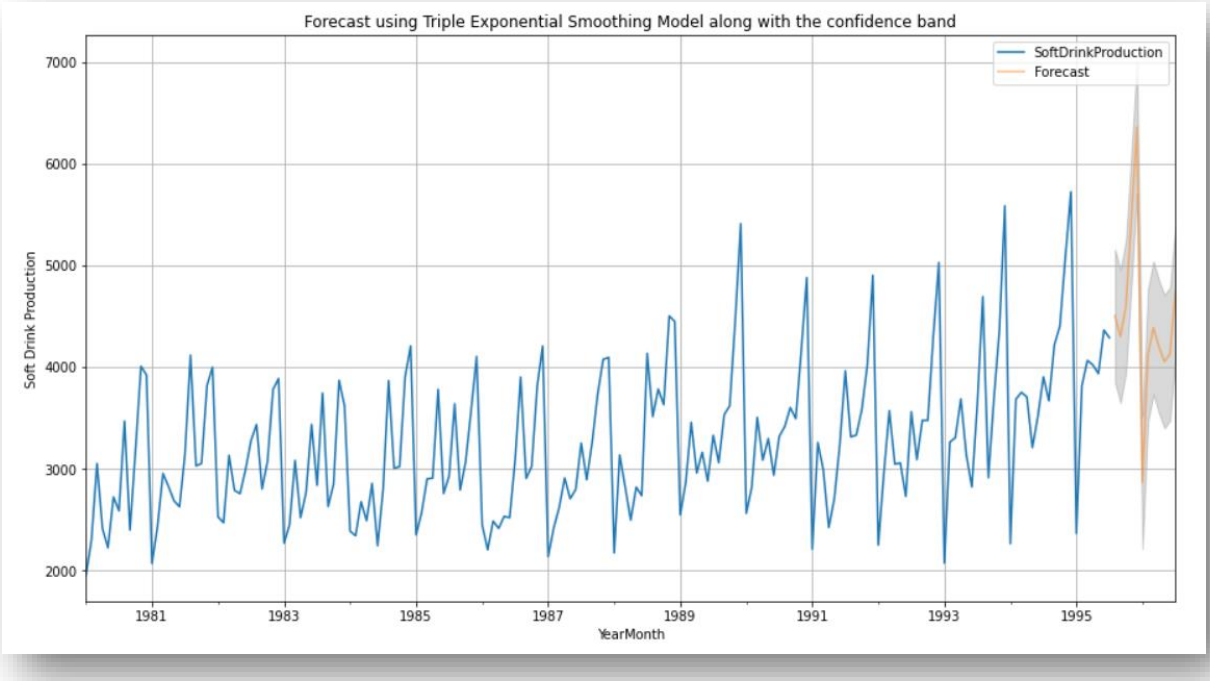

Figure 70: Forecast using the Triple Exponential Smoothing Model along with the confidence band

Figure 70 shows the forecast of the sales of soft drink production for the upcoming 12 months with confidence bands at 95% confidence level. The orange part depicts the forecasted data from August 1995 to July 1996 and the grey shaded region shows the confidence band. This means that even if

there are variations in the data, we can say with 95% confidence that the forecast will fall under the confidence band (the shaded region).

Forecasting using Auto SARIMA Model:

The parameters (3,1,3) (3,0,3,6) are the optimum values that are taken to build the model and forecast the future. The SARIMA model results are shown in Figure 71.

SARIMAX Results

Dep. Variable:	SoftDrinkProduction	No. Observations:	187
Model:	SARIMAX(3, 1, 3)x(3, 0, 3, 6)	Log Likelihood	-1191.191
Date:	Sun, 03 Apr 2022	AIC	2408.382
Time:	18:07:41	BIC	2448.680
Sample:	01-01-1980	HQIC	2424.742
	- 07-01-1995		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.8042	0.662	-1.215	0.225	-2.102	0.494
ar.L2	-0.1702	0.639	-0.266	0.790	-1.424	1.083
ar.L3	-0.1578	0.097	-1.623	0.105	-0.348	0.033
ma.L1	-0.0810	0.673	-0.120	0.904	-1.400	1.238
ma.L2	-0.5683	0.365	-1.557	0.120	-1.284	0.147
ma.L3	0.0073	0.520	0.014	0.989	-1.012	1.027
ar.S.L6	-0.0035	0.711	-0.005	0.996	-1.397	1.390
ar.S.L12	1.0167	0.023	44.367	0.000	0.972	1.062
ar.S.L18	0.0185	0.722	0.026	0.980	-1.397	1.434
ma.S.L6	-0.0813	0.707	-0.115	0.908	-1.468	1.305
ma.S.L12	-0.6802	0.126	-5.391	0.000	-0.927	-0.433
ma.S.L18	-0.0327	0.484	-0.068	0.946	-0.982	0.916
sigma2	1.144e+05	1.11e+04	10.305	0.000	9.26e+04	1.36e+05

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	19.29
Prob(Q):	0.98	Prob(JB):	0.00
Heteroskedasticity (H):	1.60	Skew:	0.50
Prob(H) (two-sided):	0.08	Kurtosis:	4.35

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

Figure 71: SARIMA model results (full model)

```
RMSE for the full model (SARIMA): 445.9190830529295
```

Figure 72: RMSE for the full model using SARIMA model

The RMSE value for the full model is calculated to be **445.91** which is seen in Figure 72.

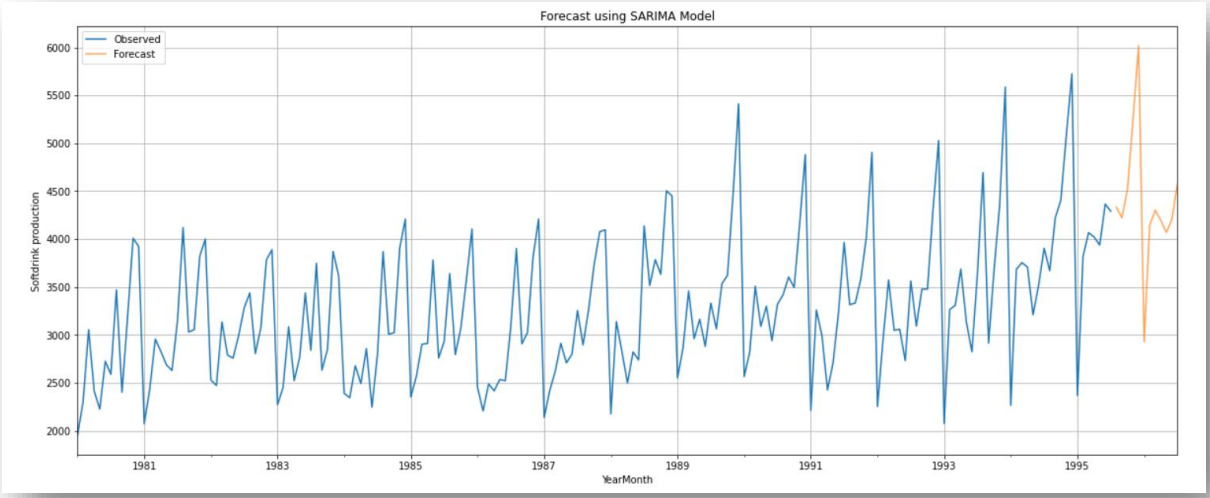

Figure 73: Forecast using the SARIMA Model

Figure 73 shows the forecast of the sales of soft drink production for the upcoming 12 months using the SARIMA model. The orange part depicts the forecasted data from August 1995 to July 1996.

Table 29: Data with upper and lower confidence bands

SoftDrinkProduction	mean	mean_se	mean_ci_lower	mean_ci_upper
1995-08-01	4332.502667	338.261723	3669.521873	4995.483462
1995-09-01	4222.067120	340.484258	3554.730237	4889.404003
1995-10-01	4518.285554	341.785841	3848.397616	5188.173493
1995-11-01	5246.599109	343.792500	4572.778191	5920.420026
1995-12-01	6019.119757	352.983470	5327.284869	6710.954645

The upper and lower confidence bands at 95% confidence level is calculated and shown in Table 29. The multiplier is taken to be 1.96 as we want to plot with respect to 95% confidence intervals. The plotted graph is shown in Figure 74.

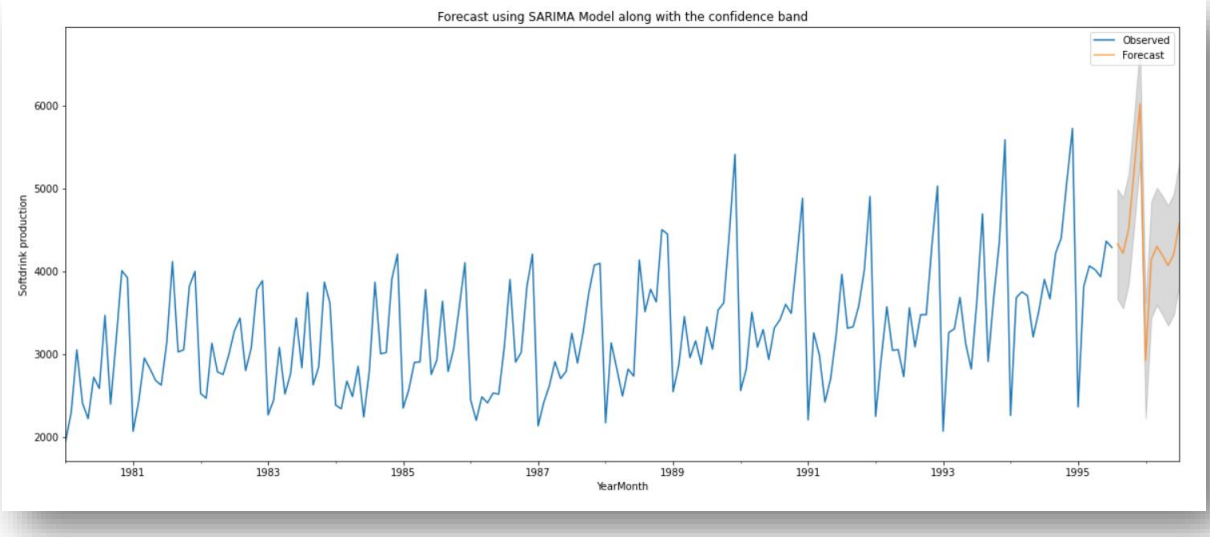

Figure 74: Forecast using the SARIMA Model along with the confidence band

Figure 74 shows the forecast of the sales of soft drink production for the upcoming 12 months with confidence bands at 95% confidence level. The orange part depicts the forecasted data from August 1995 to July 1996 and the grey shaded region shows the confidence band. This means that even if there are variations in the data, we can say with 95% confidence that the forecast will fall under the confidence band (the shaded region).

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

Inference from the models

From Figure 68, we observe that the RMSE score for the full model built using Triple Exponential Smoothing is 333.47. From Figure 72, we observe that the RMSE score for the full model built using SARIMA is 445.91. Both these models take into account the trend and seasonality in the data. The RMSE scores when forecasted on the whole dataset are low in both the models, however, comparatively the RMSE score is lower for the Triple Exponential Smoothing model.

To improve results, we can use data over a larger span of time and try using different combinations of parameters to figure out which model is the best for the data.

Insights and Recommendations

The main objective of this problem is to build a model to forecast the sales of soft drinks production for the next 12 months.

The model was built on a dataset from January 1980 to July 1995. However, if we have data over a larger span of time, we will be able to forecast the sales more accurately. The forecast of soft drinks sales was done with the data available.

Once we forecast the soft drinks sales for the next 12 months, we can make decisions and various measures to improve sales.

We can use this forecast results and form business dealings with the soft drinks companies. We can help them improve the sales while the company can offer us a role or appoint us as their official analyst.

Measures to improve sales

Marketing and promoting the soft drinks can play a major role to increase sales. Paid partnerships and collaborations with social media influencers can have a great impact and also have a good reach to the people. This is a way of advertising the brand which can increase sales.

Collaborations with restaurants and delivery apps can be done to increase the sales to a great extent.

The seasons of the year and the place must also be taken into account. People do not prefer soft drinks during the winters and the rainy seasons. They generally prefer something soothing or hot in such weather conditions. Therefore, during these seasons, different varieties of drinks that is apt for winter seasons like hot chocolate can be introduced and the production of soft drinks can be reduced.

A soft drink dispenser machine or a vendor machine can be installed at different locations where people visit often. It can be installed in locations like railway stations, malls, airports etc., where people can just grab a drink on the move.

The size of the bottles can be varied. Small, medium, large bottles, cans and tetra packs with different capacities can be introduced. This will help improve sales as people can choose the size of the drinks according to their need. This will also reduce wastage and therefore increase sales.

Discounts and offers can be given for a minimum purchase of certain number of soft drinks with a valid time period to make the customers buy again within the validity time. This way we can improve sales and the customers will also be content with the discounted price. Cash back vouchers can also be offered at times to encourage people to buy soft drinks.

Partnerships with theatres and food courts in malls can be useful. People tend to have a soft drink with popcorn while watching a movie. Fast food like pizza and burgers along with soft drinks is also a very popular combination. Combos can be offered in such cases to make people buy soft drinks along with their food.

New flavours can be introduced and feedback from the customers can be collected before launching it in the market. This can help us understand the likes of the consumers and help increase sales of a particular flavour.