ABC Estate Wines - Sparkling

Business Report 2025

Contents

Table of Figures

Introduction

Problem Statement

As an analyst at ABC Estate Wines, we are presented with historical data encompassing the sales of different types of wines throughout the 20th century. These datasets originate from the same company but represent sales figures for distinct wine varieties. Our objective is to delve into the data, analyze trends, patterns, and factors influencing wine sales over the course of the century.

By leveraging data analytics and forecasting techniques, we aim to gain actionable insights that can inform strategic decision-making and optimize sales strategies for the future.

Objective

The primary objective of this project is to analyze and forecast wine sales trends for the 20th century based on historical data provided by ABC Estate Wines. We aim to equip ABC Estate Wines with the necessary insights and foresight to enhance sales performance, capitalize on emerging market opportunities, and maintain a competitive edge in the wine industry.

Exploratory Data Analysis

About data and the dictionary

The sparkling dataset contains two columns and 187 rows. The dataset explains the sales of sparkling wine from the year of 1980 to 1995 for each month. One column consists of the year and month and the other column signifies the sales of the wine. Below attached is the sample of the dataset.

	YearMonth	Sparkling
0	1980-01	1686
1	1980-02	1591
2	1980-03	2304
3	1980-04	1712
4	1980-05	1471

Figure 1 Sample Data

Treatments to the data:

Indexing of YearMonth

We would index the year month column while uploading the data.

Missing Value

There are no missing values in the sparkling sales column.

Plot the Time Series

We have plotted the time series using the sparkling dataset where the sparkling sales is the y variable and years as the x variable on the graph. Before analysing, we need to understand what trend and seasonality is. In data science, a trend is a pattern that shows how a variable changes over time and seasonality refers to predictable, repetitive, and regular changes in data that occur at specific intervals throughout a year. In this graph we observe that there is no trend since it stays within a range and we can also observe a seasonality factor because of the patterns formed.

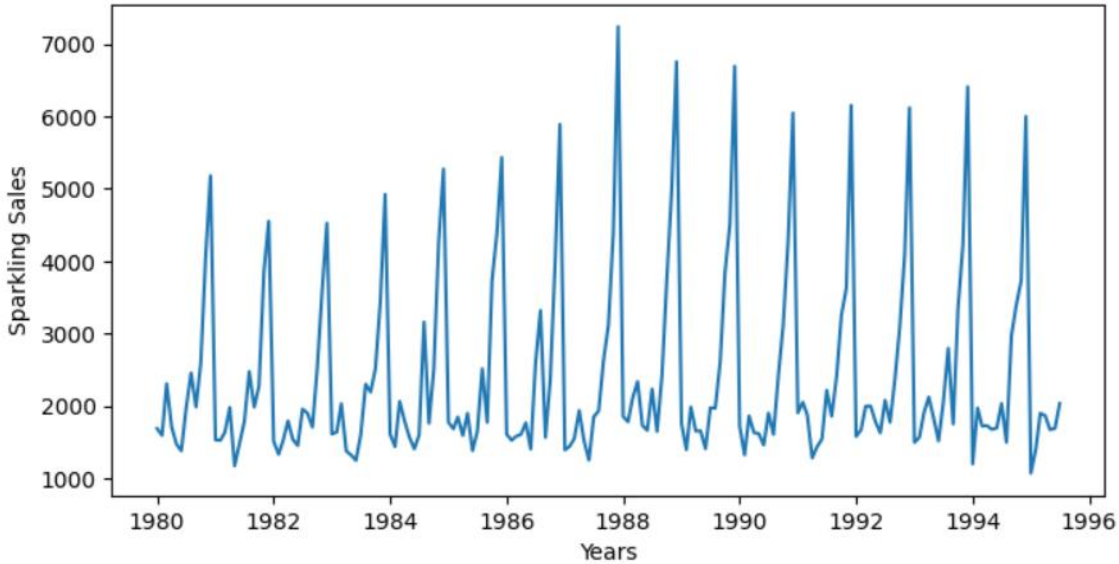

Figure 2 Time Series Plot

Data Analysis

The average of the sparkling sales is 2402.42. The minimum sales of the sparkling is 1070 and the maximum sales is 7242. We have three columns after treating the dataset which is month, year and sales where it is object and integers respectively.

Monthly Analysis

In this scenario, we will plot the data into a monthly box plot to understand the trend and seasonality on a monthly basis throughout the years.

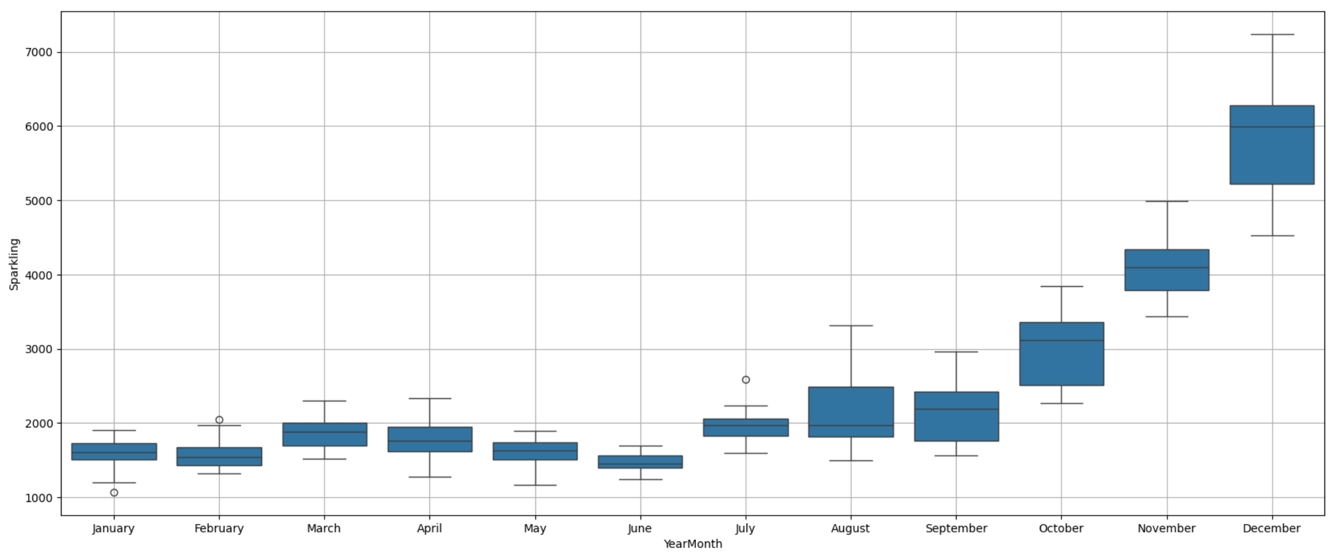

Figure 3 Monthly Analysis of sales

Analysing the graph we understand that throughout January to June the sales of sparkling are at minimum whereas we see a boom in sales of sparkling post June and the highest sales are recorded during the months of december. This might be due to multiple celebrations during the months of november and december which involves christmas and new year.

Yearly Analysis

Here we will be conducting an analysis of the sales of sparkling dataset on a yearly basis.

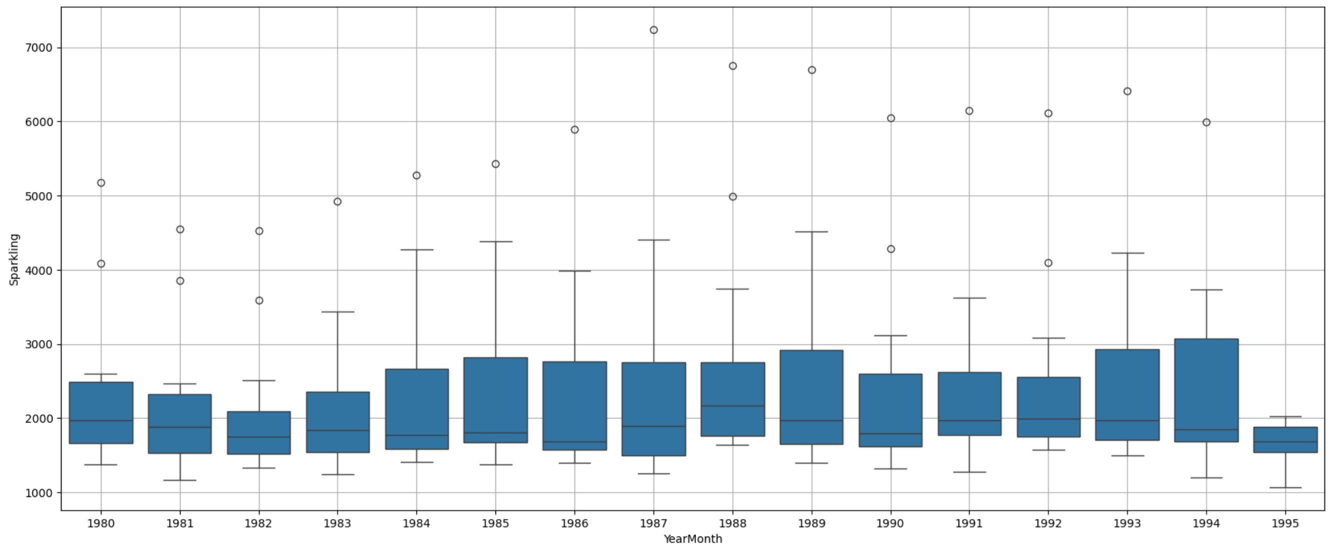

Figure 4 Yearly Analysis of sales

We could see that there is no specific trend in the graph since all years have the same amount of sales.

Time Series Decomposition

Time series decomposition is a statistical process that breaks down a time series into its components to help analyze and understand the data. This process can help identify patterns and variations, and make it easier to model and forecast future data points.

The **three main components** of a time series are:

- **Trend-cycle component:** Combines the trend and cycle into a single component
- **Seasonal component:** Represents seasonal patterns
- **Remainder component:** Contains anything else in the time series

We have seasonality to be in multiplicative or additive manner.The additive model is when the components are added together and the model is often used when the magnitude of the components doesn't change with the level of the time series. The multiplicative model is when the components are multiplied together. This model is often used when the magnitude of the components increases with the level of the time series.

The multiplicative time series decomposition of sparkling dataset is as follows:

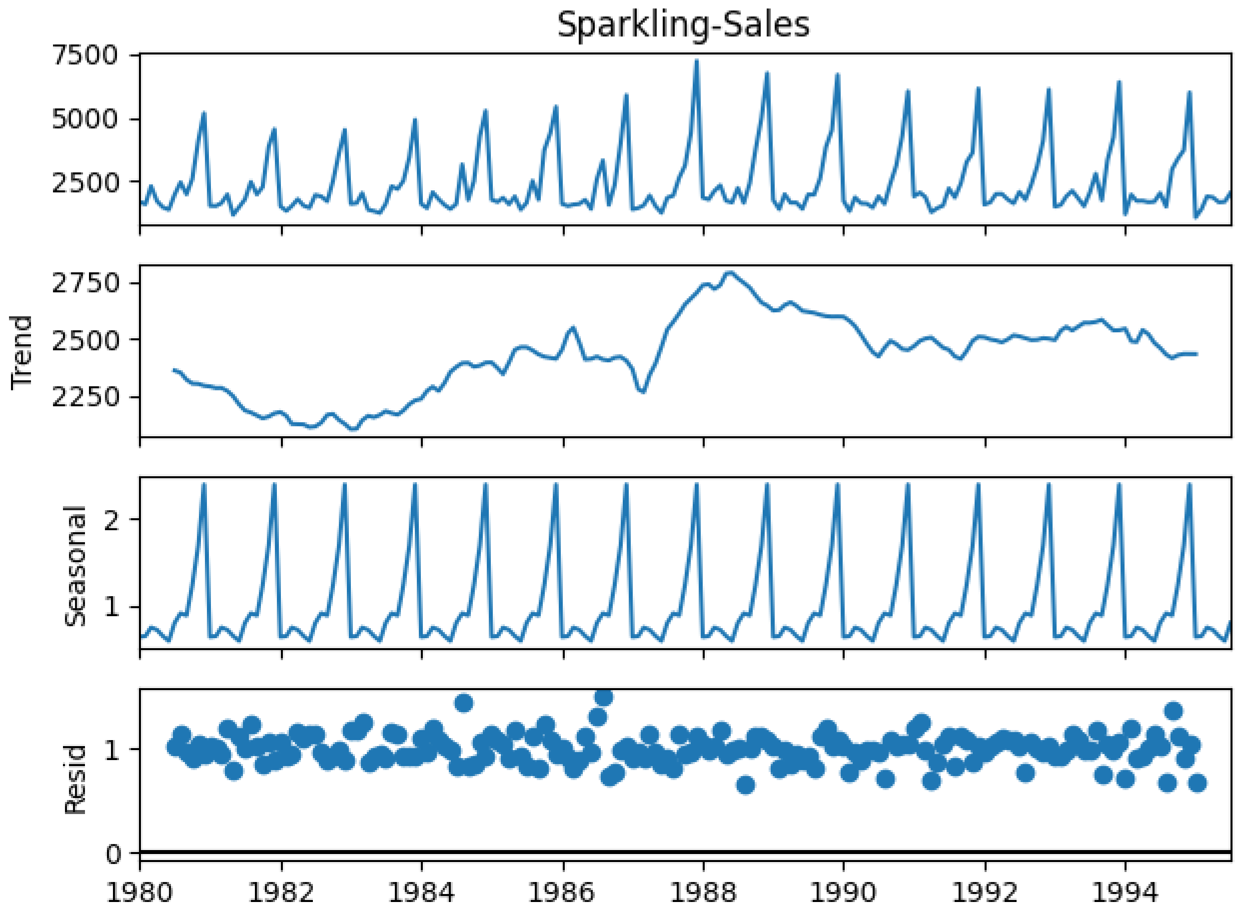

Figure 5 Multiplicative time series decomposition

The additive time series decomposition of sparkling dataset is as follows:

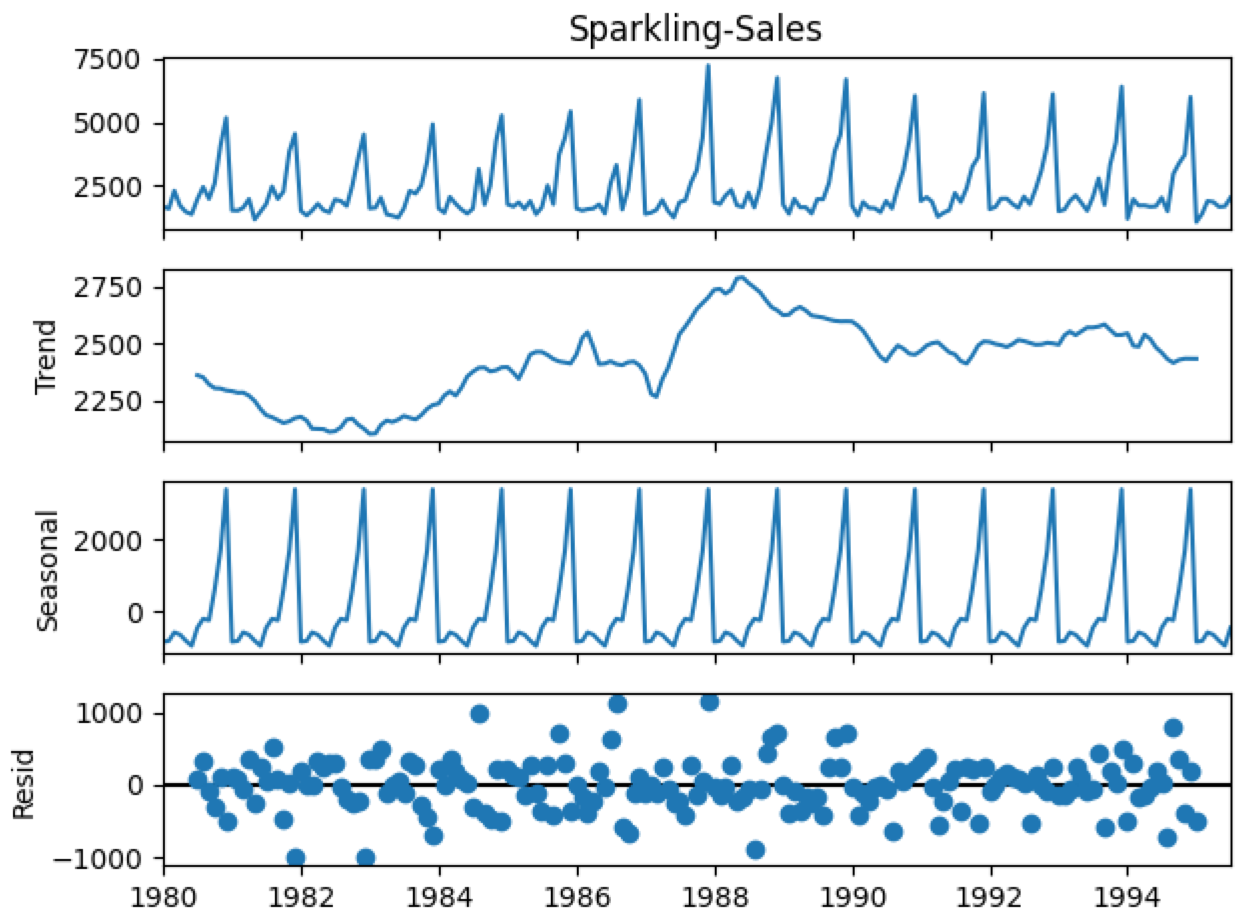

Figure 6 Additive time series decomposition

Both the models do not show any patterns and hence we could go with the additive model.

In the additive model, we can observe that there is no trend in this case since the trend line seems to be in a range and shows an increase or decrease within the range. If we check the seasonality, we could see that there is a regular occurring pattern and the residuals do not form a pattern and are in a very random state. Hence we could say that the dataset contains seasonality and trend.

Summary of Exploratory Data Analysis

Our analysis of sparkling wine sales data from 1980 to 1995 reveals key insights that can inform strategic planning at ABC Estate Wines. The data exhibits strong seasonality, with peak sales occurring during the festive months of November and December, likely driven by holiday celebrations. However, no discernible long-term trend in sales growth or decline was identified.

The time series decomposition confirmed the absence of a clear trend but a trend in range while emphasizing the regular seasonal patterns. This indicates that external factors, such as annual festive periods, play a significant role in driving sales.

These insights suggest that ABC Estate Wines should focus marketing and inventory strategies on maximizing opportunities during high-demand months while exploring ways to sustain sales during off-peak periods. Leveraging this seasonality can enhance operational efficiency and profitability. Future forecasting models can be built to optimize planning and capitalize on these seasonal patterns effectively.

Splitting of dataset

Data splitting is when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other to train the model. We will be splitting the dataset into 70 and 30% where 70% is used for training the dataset and 30% is used for the testing purpose.

In this case we have 130 records for the training dataset and 57 records for the testing purpose.

```
First few rows of Training Data
            Month  Year  Sparkling-Sales
1980-01-01   Jan  1980             1686
1980-02-01   Feb  1980             1591
1980-03-01   Mar  1980             2304
1980-04-01   Apr  1980             1712
1980-05-01   May  1980             1471

Last few rows of Training Data
            Month  Year  Sparkling-Sales
1990-06-01   Jun  1990             1457
1990-07-01   Jul  1990             1899
1990-08-01   Aug  1990             1605
1990-09-01   Sep  1990             2424
1990-10-01   Oct  1990             3116

First few rows of Test Data
            Month  Year  Sparkling-Sales
1990-11-01   Nov  1990             4286
1990-12-01   Dec  1990             6047
1991-01-01   Jan  1991             1902
1991-02-01   Feb  1991             2049
1991-03-01   Mar  1991             1874

Last few rows of Test Data
            Month  Year  Sparkling-Sales
1995-03-01   Mar  1995             1897
1995-04-01   Apr  1995             1862
1995-05-01   May  1995             1670
1995-06-01   Jun  1995             1688
1995-07-01   Jul  1995             2031
```

Figure 7 Train - Test - Split

This is our training and testing dataset. We should not be using randomness while splitting the dataset because that might disrupt the time series and will affect the forecasting.

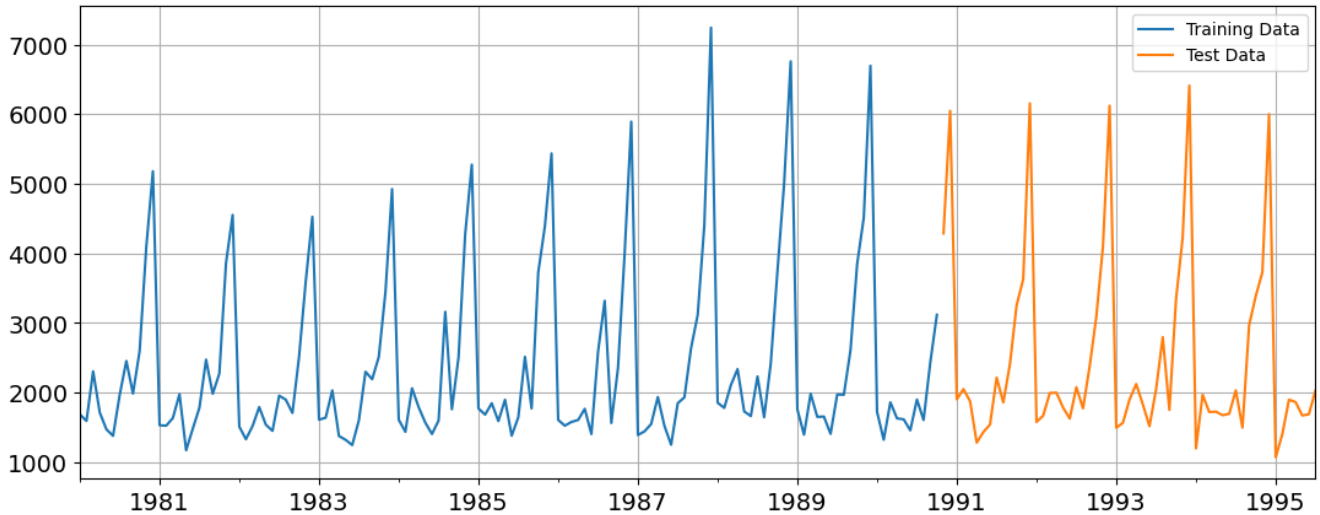

Figure 8 Plot of Train - Test - Split

Above seen is the data splitted into training and testing dataset.

Model Building with original Dataset

In this part we will build models on the training dataset and then we will test the model on the testing dataset and then check the performance of each model.

The models which are being built are:

1. Linear Regression
2. Simple Average
3. Moving Average
4. Simple Exponential Smoothing
5. Double Exponential Smoothing (Holt's Method)
6. Triple Exponential Smoothing (Holt-Winter Method)

Linear Regression

Definition: Linear regression models the relationship between a known independent variable and an unknown dependent variable as a linear equation. It uses the known data to predict the value of the unknown data.

To conduct forecasting for this under linear regression method we need to change the values to instances because this is a supervised learning method.

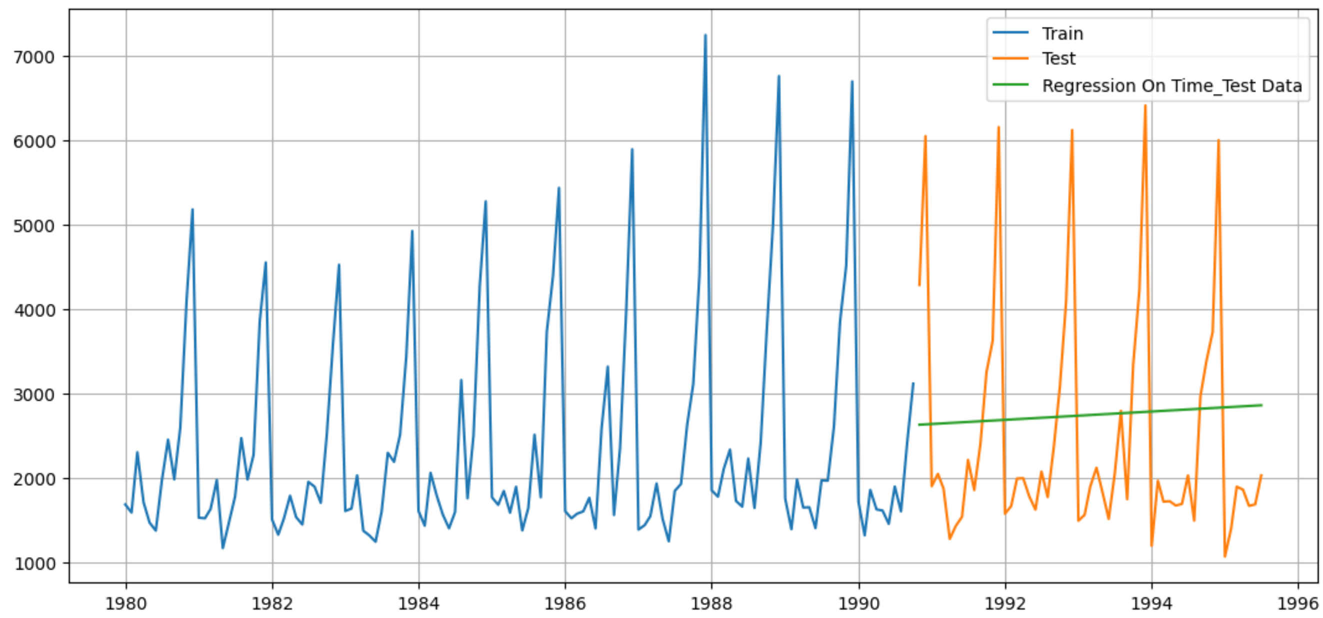

Figure 9 Plot of linear regression forecast

We have fitted the values in the linear regression model and have plotted the values on the graph. We see that the green line is the one which has been forecasted. It is just a straight line and does not capture the seasonality factor.

We will be evaluating with Root Mean Square Error (RMSE). The RMSE for this model is 1392.44. This standalone number does not make sense, so we will move ahead and check for RMSE in others.

Simple Average

Definition: A simple average in time series is a forecasting method that uses the average of past values to predict the future. It's the simplest method for forecasting time series data.

We take the mean of the test data and then plot the value on the graph. Most of the time this can be used as a benchmark for the RMSE.

The graph generated is as below.

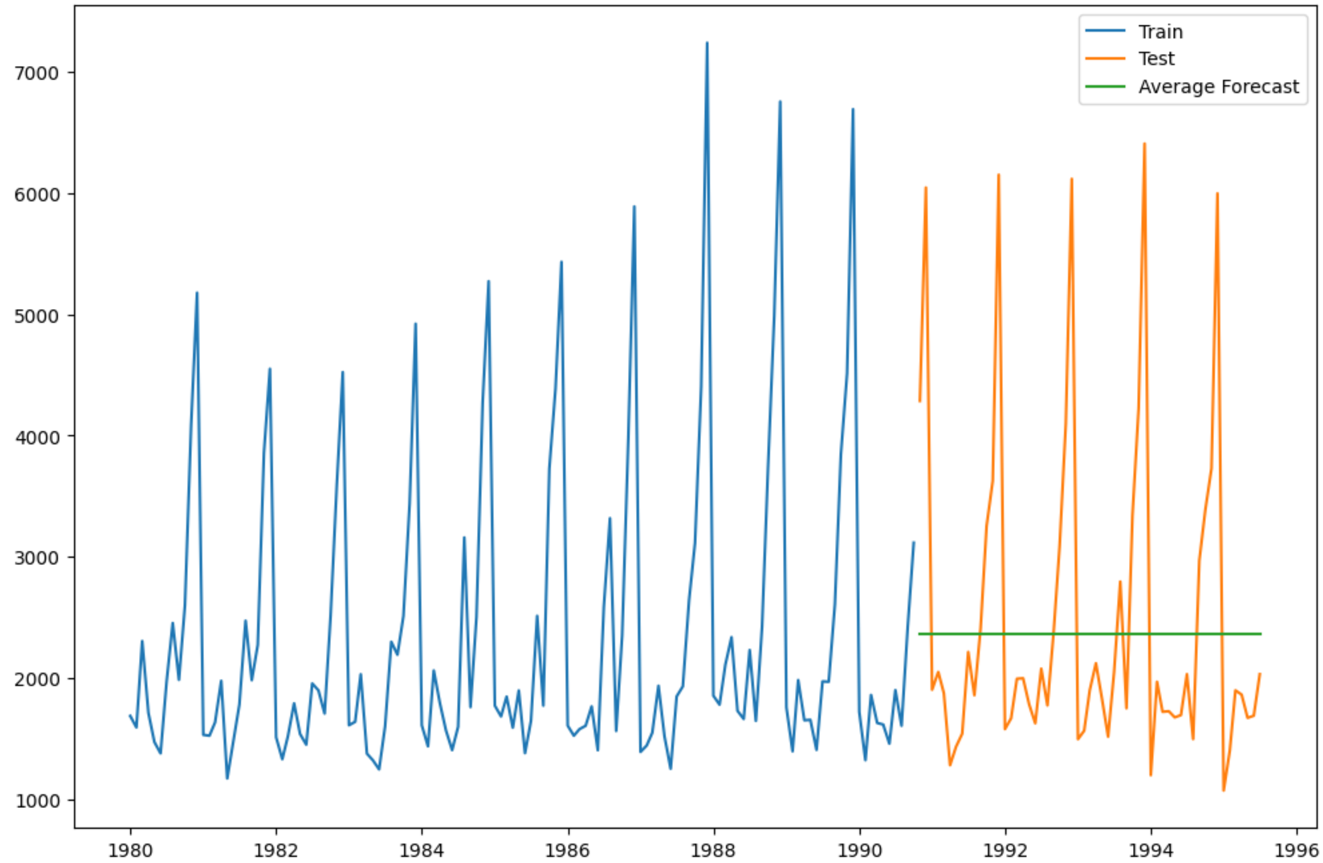

Figure 10 Plot of simple average forecast

The RMSE for this is 1368.75.

Moving Average

Definition: A moving average is a series of averages calculated from subsets of data, with the window of observation shifting across the time series.

For the moving average we are going to calculate the average for various periods and we would choose the best one with maximum accuracy and low error. For this we take 2, 4, 6 and 9 rolling periods. Then we add these columns and make the average values to the dataset.

After the treatment we could see that our dataset would look like as below:

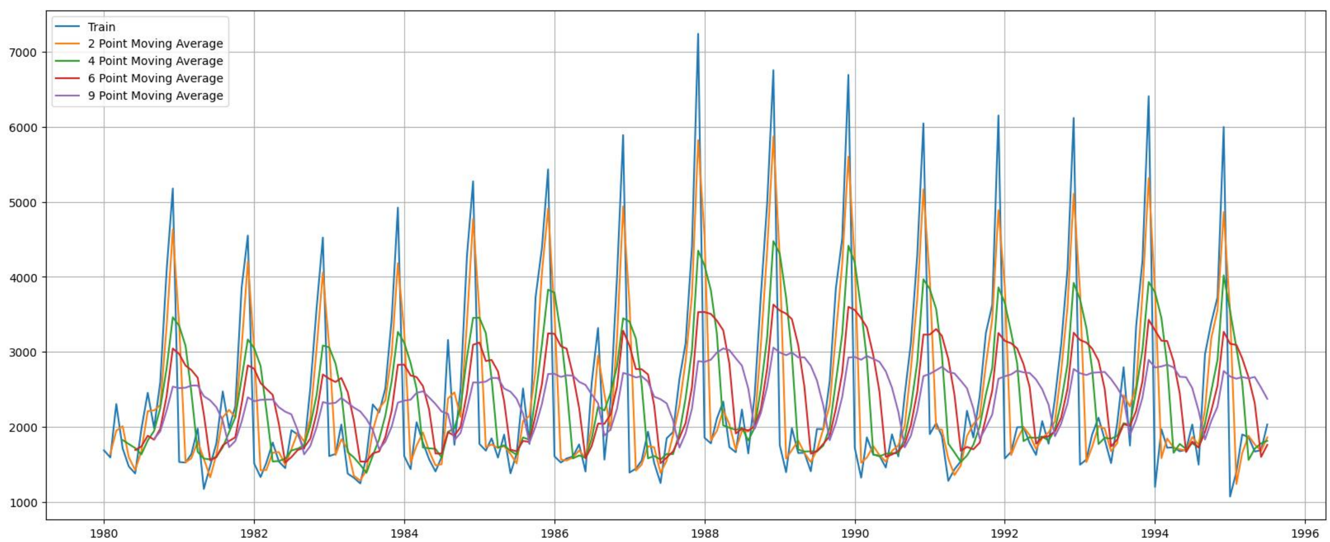

Figure 11 Plot of moving average various windows

Subsequent to this we would divide the dataset into training and test data and then we would forecast. The graph would look like as below:

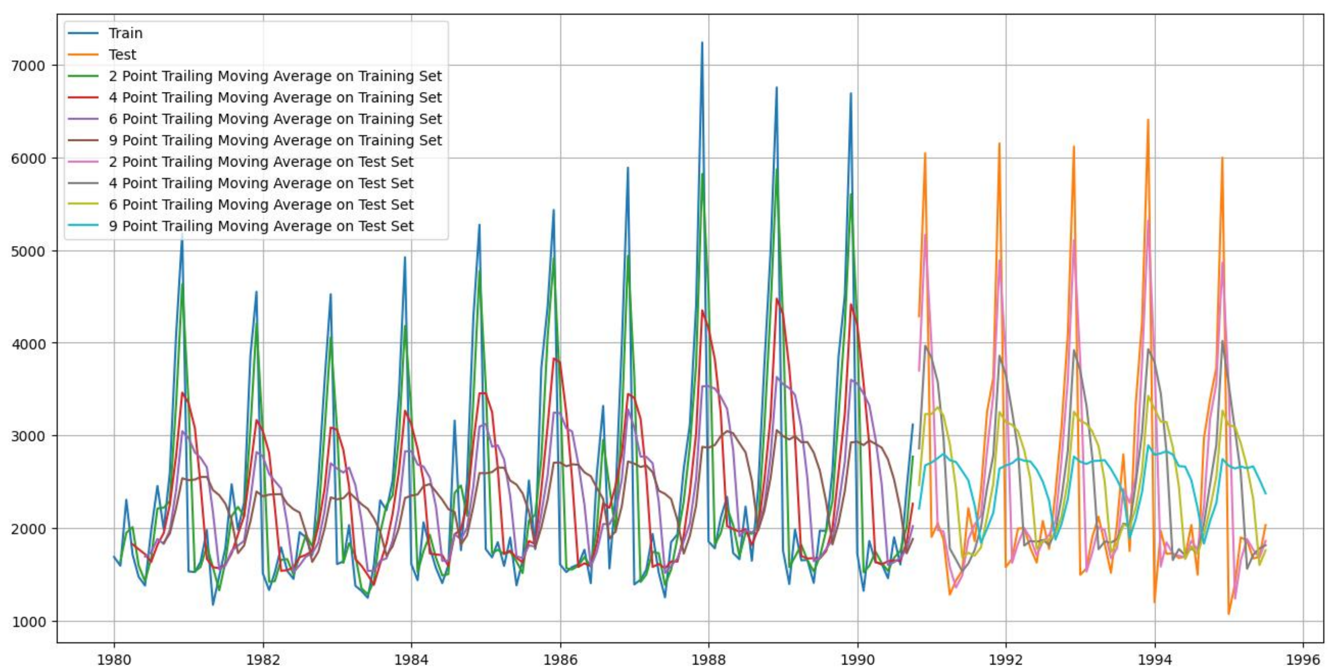

Figure 12 Plot of moving average various windows on test data

However, visually we would not be able to make sense from this lines hence we would compare with the RMSE and then understand which trailing period is good. In this case we have the RMSE as follows:

- For 2 point Moving Average Model forecast on the Training Data, RMSE is **811.179**
- For 4 point Moving Average Model forecast on the Training Data, RMSE is **1184.213**
- For 6 point Moving Average Model forecast on the Training Data, RMSE is **1337.201**
- For 9 point Moving Average Model forecast on the Training Data, RMSE is **1422.653**

We could see that the RMSE is low in case of a 2 point moving average model because the RMSE is low. We can visualise the graph and it is as below:

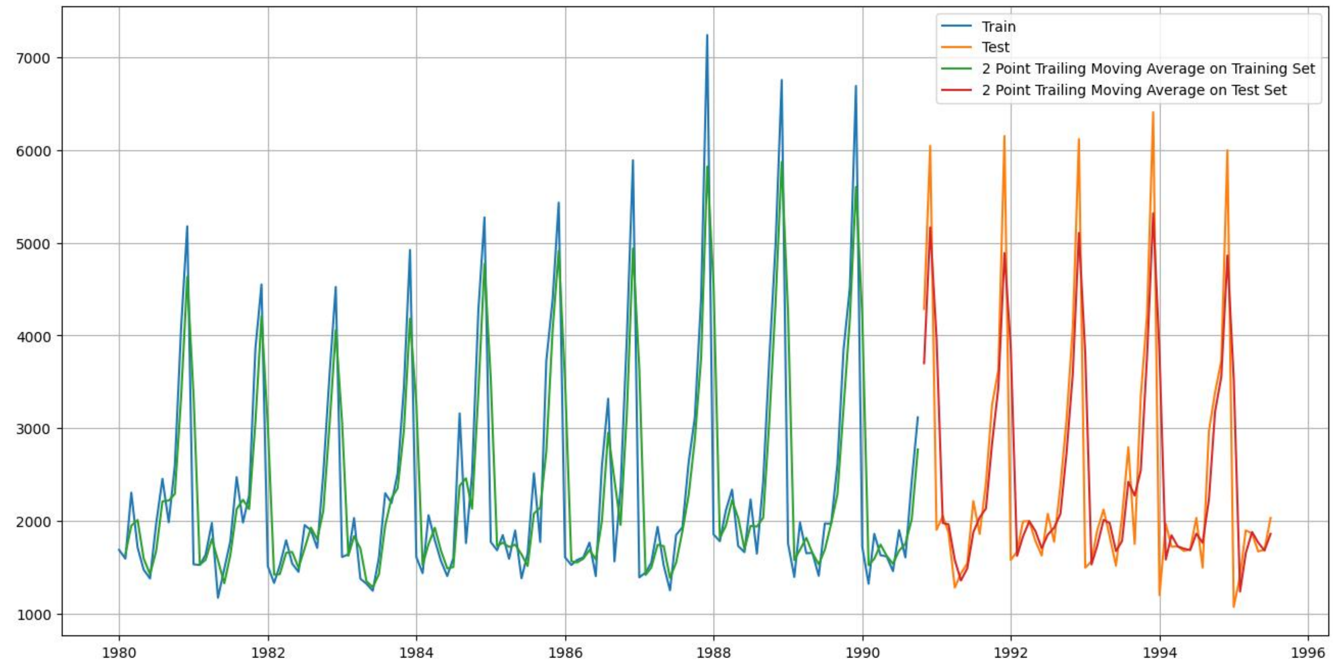

Figure 13 Plot of 2 point trailing moving average

The forecast line for the 2 point trailing moving average is close to the test data and the RMSE is also low. However, let's check out other models in order to get much less RMSE.

Simple Exponential Smoothing

Definition: Simple exponential smoothing (SES) is a time series forecasting method that uses weighted averages to estimate future values based on past observations. It needs a single parameter called alpha (a), also known as the smoothing factor.

In this method we import the necessary libraries and auto fit our train dataset into the simple exponential model. Once we fit, we would use it to predict our test data.

We have got the graph as below:

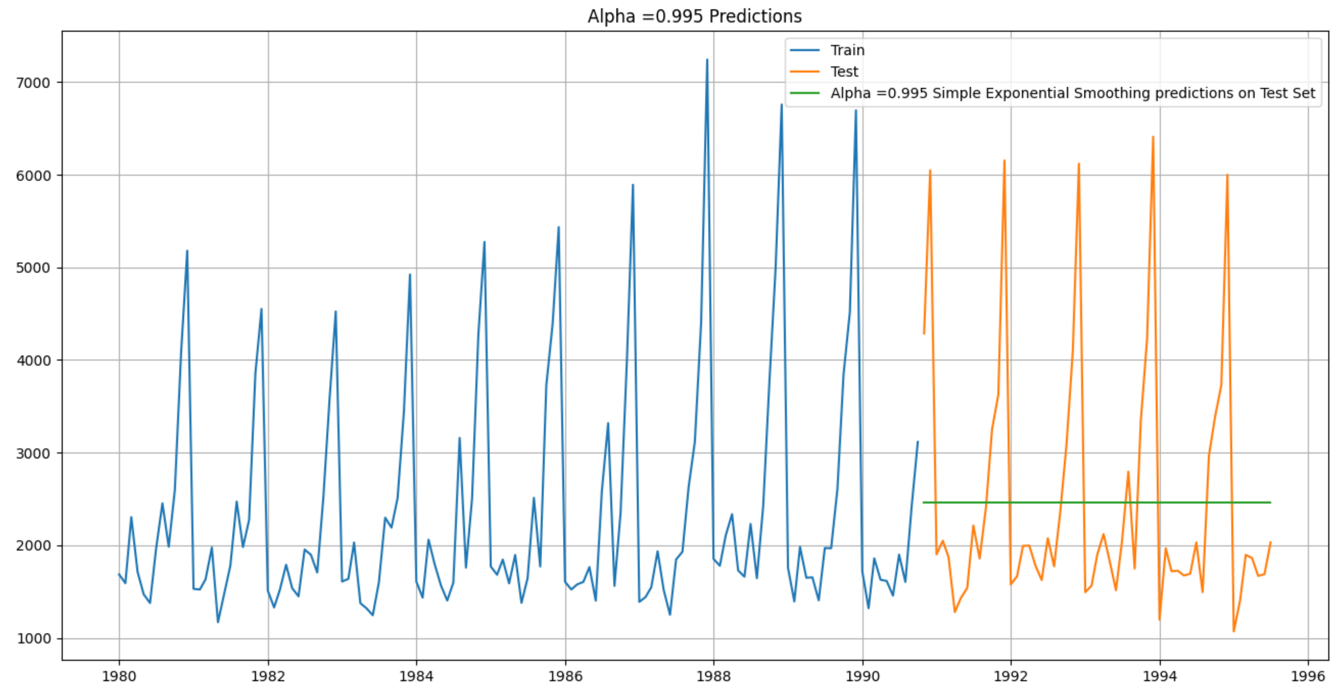

Figure 14 Plot of simple exponential smoothing

The RMSE for this model is 1362.43 which is very low since it does not take seasonality and trend into consideration.

In this scenario we have taken the alpha value as 0.995. We could also check for the best alpha values. The higher alpha value then the higher weightage is given to the recent observations.

We could also analyse the good alpha value which would increase the accuracy and reduce the RMSE. The table would look like below:

	Alpha Values	Train RMSE	Test RMSE
28	0.4	1329.814823	1363.037803
2	0.4	NaN	1363.037803
40	0.4	1329.814823	1363.037803
59	0.4	1329.814823	1363.037803
21	0.4	1329.814823	1363.037803

We could see that the alpha value with 0.4 is giving us the lowest RMSE among the list which is 1363.038. However our prediction of alpha value with 0.995 has the lowest in comparison to this which is 1362.43.

Currently we have the RMSE for all models below:

	Test RMSE
Linear Regression	1392.438305
2pointTrailingMovingAverage	811.178937
4pointTrailingMovingAverage	1184.213295
6pointTrailingMovingAverage	1337.200524
9pointTrailingMovingAverage	1422.653281
Alpha=0.995,SimpleExponentialSmoothing	1362.428949
Alpha=0.4,SimpleExponentialSmoothing	1363.037803

Double Exponential Smoothing (Holt's Method)

Definition: Double exponential smoothing is a statistical method used in time series forecasting to predict data with a linear trend but no seasonal pattern. It's also known as Holt's trend model or second-order exponential smoothing.Double exponential smoothing is different from regression models because it never forgets any part of the past. This means it can take longer to recover if there's a perturbation in the underlying mean.

In this method we import the necessary libraries and auto fit our train dataset into the simple exponential model. Once we fit, we would use it to predict our test data.

We have got the graph as below:

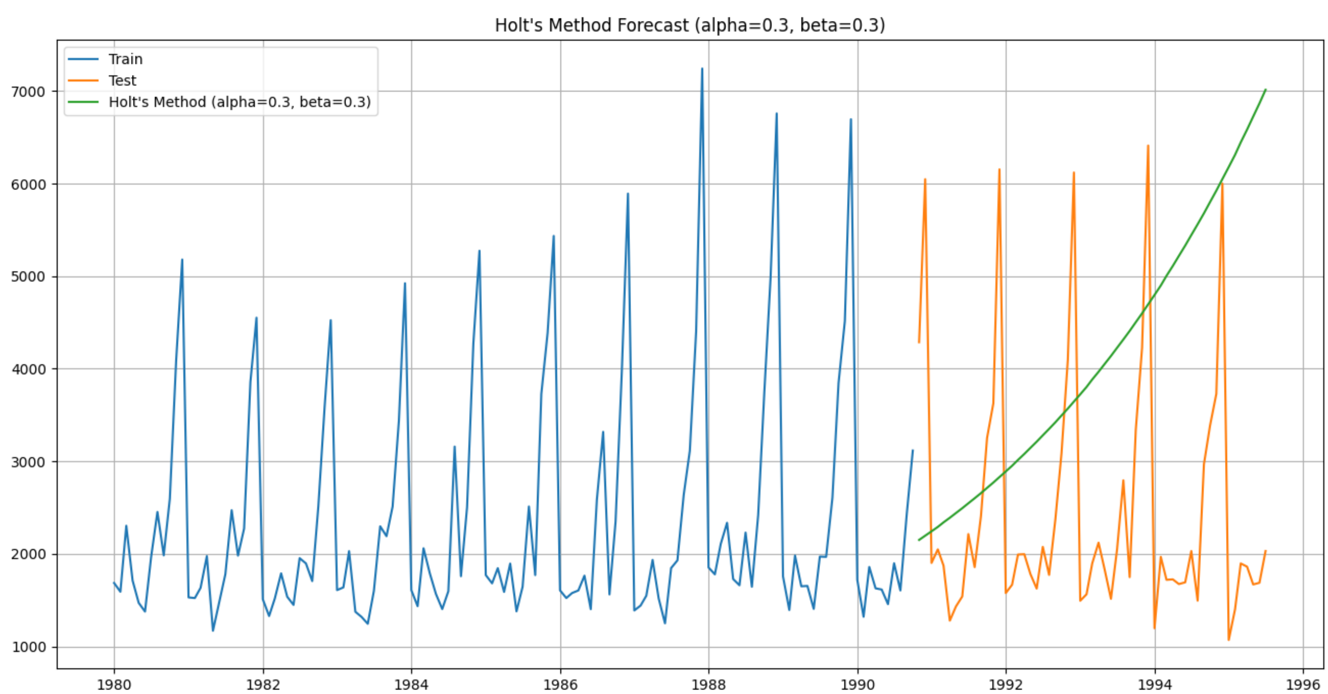

Figure 15 Plot of double exponential smoothing

We took alpha as 0.3 and beta as 0.3 since this combination provide the lowest RMSE at 2619.07.

The top 5 combination of best results are:

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.3	0.3	1712.839849	2619.071230
1	0.3	0.4	1905.151412	18174.142159
9	0.4	0.3	1754.621738	72104.745621
80	1.1	1.1	2284.776385	110704.348239
2	0.3	0.5	2136.358448	362523.847069

However, analyzing the graph we understand that the forecast values and our test data are not the same and the RMSE is also high. Let's move ahead and check other methods.

Triple Exponential Smoothing (Holt-Winter Method)

Definition: Triple exponential smoothing, also known as Holt-Winters exponential smoothing, is a time series forecasting method that uses three smoothing equations to predict values based on a time series' value, trend, and seasonality. The model predicts a current or future value by computing the combined effects of these three influences.

In this method, we import necessary libraries and then check out the best combination of alpha, beta and gamma and then calculate the RMSE and plot the values in the graph.

The graph generated is as below:

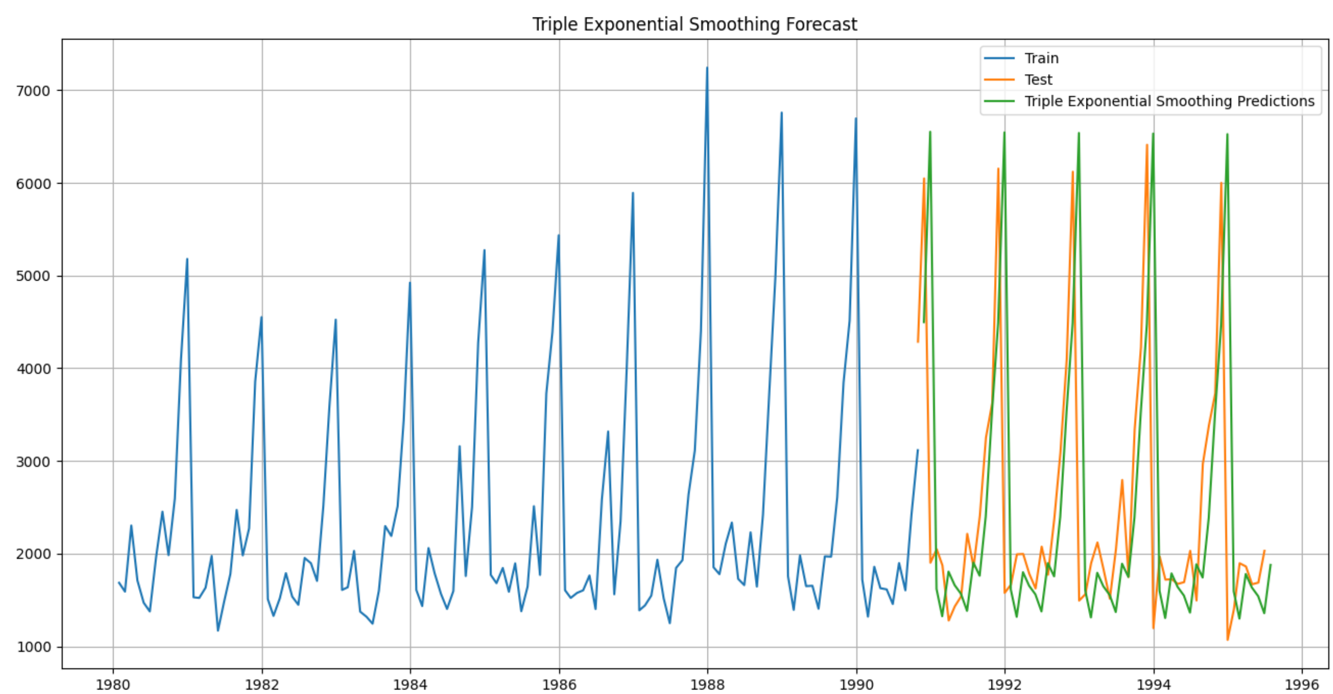

Figure 16 Plot of triple exponential smoothing

We could see that the forecast values are moving closely in relation to the test dataset. However, to confirm, let's find out the RMSE and compare the performance with other models.

The RMSE for this model is 366.859.

Let us look into the best combination of alpha, beta and gamma and then check the RMSE value.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE
32	0.01	0.31	0.21	411.150010	319.794428
102	0.11	0.01	0.21	400.187780	323.803665
33	0.01	0.31	0.31	394.632851	334.235152
101	0.11	0.01	0.11	424.659110	346.287110
103	0.11	0.01	0.31	385.502484	347.812501

These are our top 5 combinations. Lets plot with the first combination where the RMSE is low at 319.79.

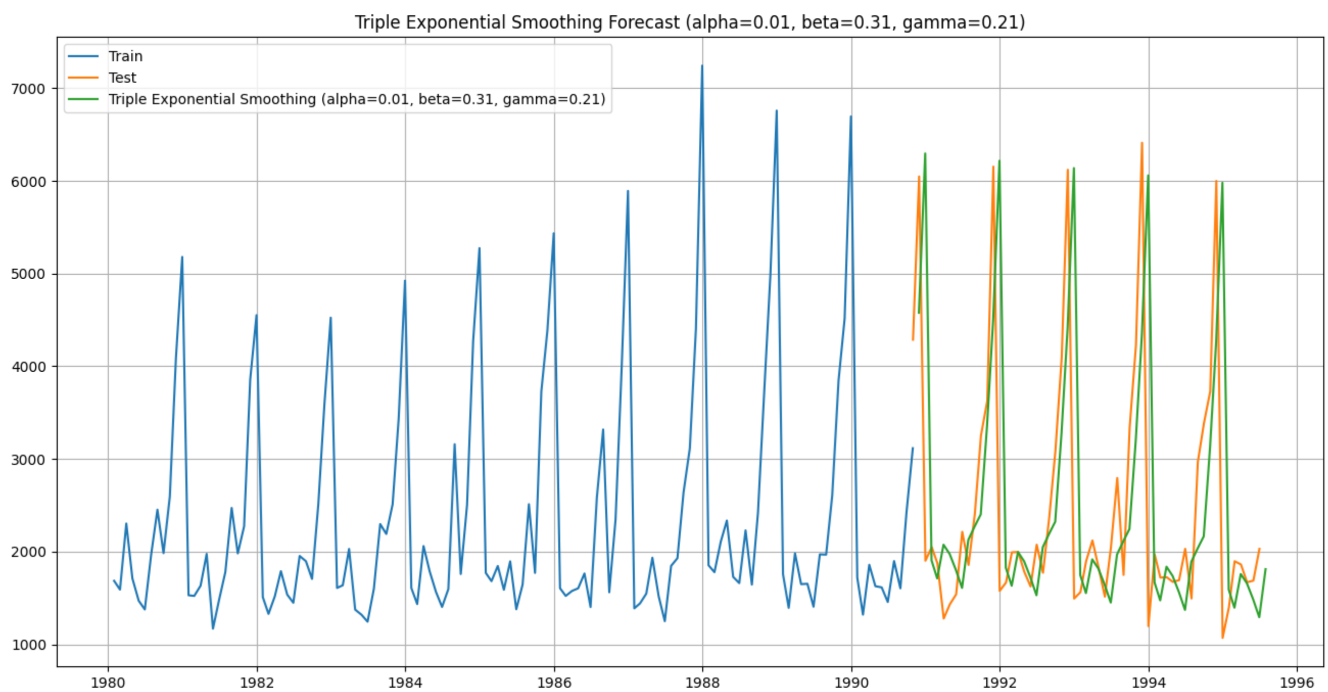

Figure 17 Plot of best triple exponential smoothing

Comparison of models

Comparison of other models is as below:

	Test RMSE
Linear Regression	1392.438305
Simple Average	1368.746717
2pointTrailingMovingAverage	811.178937
4pointTrailingMovingAverage	1184.213295
6pointTrailingMovingAverage	1337.200524
9pointTrailingMovingAverage	1422.653281
Alpha=0.4,SimpleExponentialSmoothing	1363.037803
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	2619.071230
Alpha=0.07,Beta= 0.03,Gamma=0.47,TripleExponentialSmoothing	366.859156
Alpha=0.01,Beta=0.31,Gamma=0.21,TripleExponentialSmoothing	319.794428

With all these models we could observe that the triple exponential smoothing with alpha at 0.01, beta at 0.31 and gamma at 0.21 gives an RMSE of 319.79.

We could go ahead and explore ARIMA and SARIMA models to see if the RMSE could go any further down.

Model Building with Stationary Data

Stationary Data

Stationary time series data is data that has statistical properties that do not change over time. This means that the mean, variance, and autocorrelation structure of the data remain constant.

Dickey-Fuller test

The Dickey-Fuller test is a statistical test that determines if a time series is stationary or has a unit root. The test's null hypothesis is that a unit root is present in an autoregressive (AR) time series model. The alternative hypothesis is usually stationarity or trend-stationarity. If the test shows that the series is not stationary, the data may need adjustments to make it stable for analysis.

Null Hypothesis : Time Series is non-stationary.

Alternate Hypothesis : Time Series is stationary.

While conducting this test we could see that the p-value is greater than 0.05 which is 0.60. Hence the data is not stationary and we fail to reject the null hypothesis and can understand that the time series is non-stationary.

Making Non-Stationary dataset to Stationary dataset

Differencing 'd' is done on a non-stationary time series data one or more times to convert it into stationary. (d=1) 1st order differencing is done where the difference between the current and previous (1 lag before) series is taken and then checked for stationarity using the ADF(Augmented Dicky Fueller) test. If the differenced time series is stationary, we proceed with AR modeling. Else we do (d=2) 2nd order differencing, and this process repeats till we get a stationary time series.

The variance of a time series may also not be the same over time. To remove this kind of non-stationarity, we can transform the data. If the variance is increasing over time, then a log transformation can stabilize the variance.

Non- differenced data plot

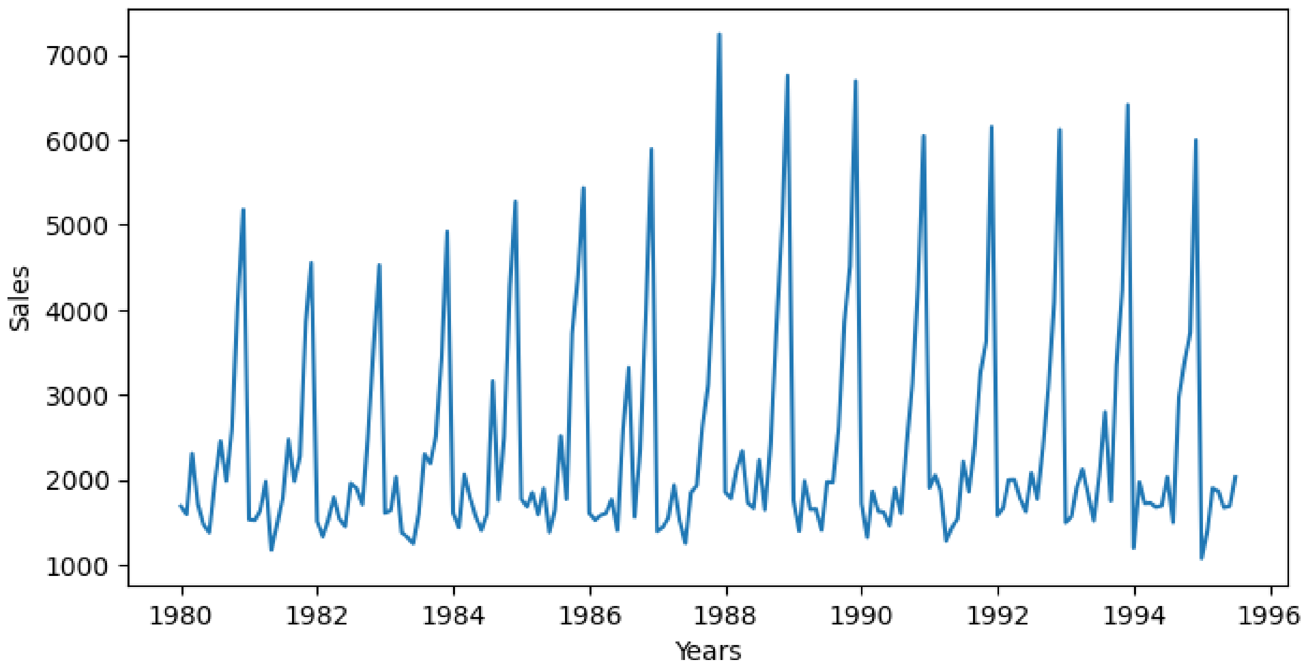

Figure 18 Non-differenced data plot

d=1 differenced data plot

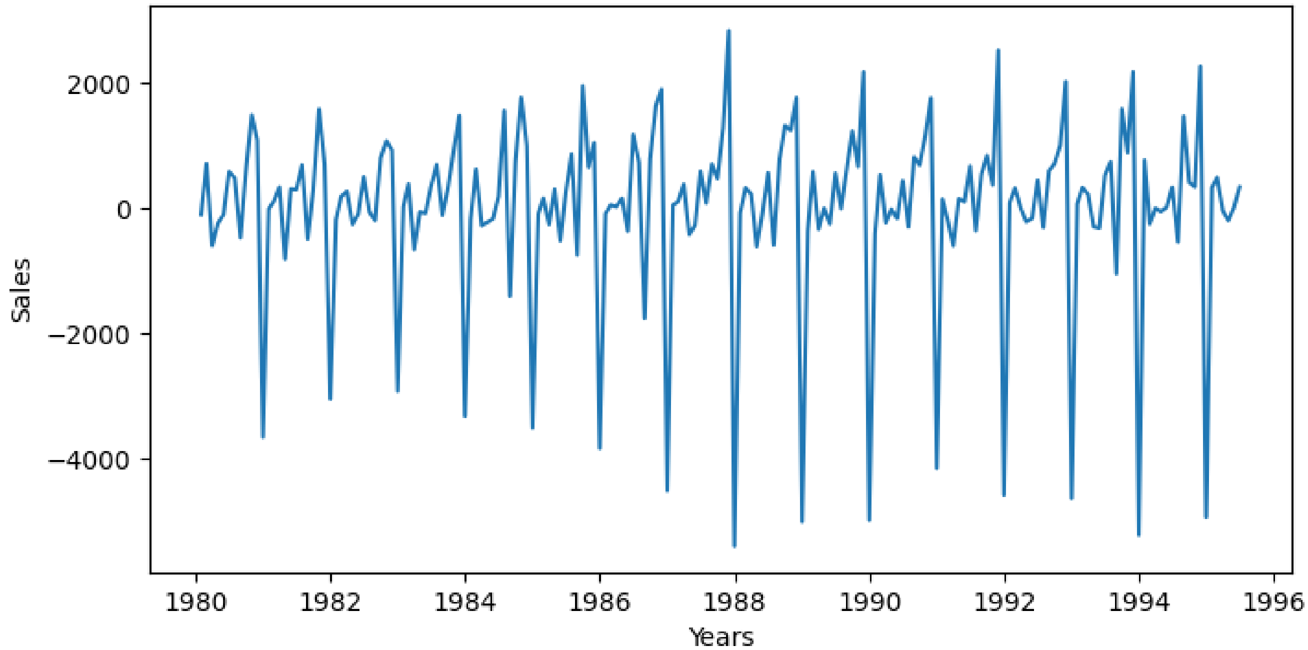

Figure 19 d=1 differenced data plot

We could see the difference in the graphs for non-differenced and differenced where the difference graph is smoothened and is same over the period.

This graph is now stationary and let's move ahead to build the models.

Auto Regressive Integrated Moving Average Model

An autoregressive (AR) model is a machine learning technique that uses past values in a time series to predict future value. AR models use mathematical techniques to determine the correlation between elements in a sequence. They then use this knowledge to predict the next element in an unknown sequence. A moving-average (MA) model is a time series model that uses past forecast errors to predict future values. It's a common way to model univariate time series, and is a key component of the ARMA model.

The partial autocorrelation function (PACF) is a statistical tool used in time series analysis to measure the correlation between a time series and its own lagged values. We look at the Partial Autocorrelations of a stationary Time Series to understand the order of Auto-Regressive models.

We can also use the AIC method in order to get the p for AR model. AIC is a single number score that estimates how well a model fits a given data set. It's calculated using the maximum likelihood estimate and the number of parameters in the model.

Let us use the AIC method to get the p and q value. For q we will take the second lowest AIC because ACF(0) = 1 and ACF(1) = PACF(1).

For AR we would get the combination as :

[(1, 0, 0), (2, 0, 0), (3, 0, 0)]

For ARMA we would get the combinations as :

[(1, 0, 1),
 (1, 0, 2),
 (1, 0, 3),
 (2, 0, 1),
 (2, 0, 2),
 (2, 0, 3),
 (3, 0, 1),
 (3, 0, 2),
 (3, 0, 3)]

For all the values in pdq with differencing factor our combinations will be:

[(1, 0, 1),
 (1, 0, 2),
 (1, 0, 3),
 (1, 1, 1),
 (1, 1, 2),
 (1, 1, 3),
 (2, 0, 1),
 (2, 0, 2),
 (2, 0, 3),
 (2, 1, 1),
 (2, 1, 2),
 (2, 1, 3),
 (3, 0, 1),
 (3, 0, 2),
 (3, 0, 3),
 (3, 1, 1),
 (3, 1, 2),
 (3, 1, 3)]

AR Model

For the AR model we came to find that (1,0,0) has the AIC at the lowest of -95.17. Below is the model generated with ARIMA(1,0,0). The Root Mean Squared Error of our forecasts is 1392.614

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  130
Model:                 ARIMA(1, 0, 0)   Log Likelihood                  50.583
Date:                Sat, 04 Jan 2025   AIC                            -95.166
Time:                        07:50:03   BIC                            -86.563
Sample:                    01-01-1980   HQIC                           -91.670
                         - 10-01-1990
Covariance Type:                  opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.3287      0.037     89.764      0.000       3.256       3.401
ar.L1          0.4505      0.073      6.195      0.000       0.308       0.593
sigma2         0.0268      0.004      6.724      0.000       0.019       0.035
==============================================================================
```

Figure 20 AR model

Below is the graph plotted with this model

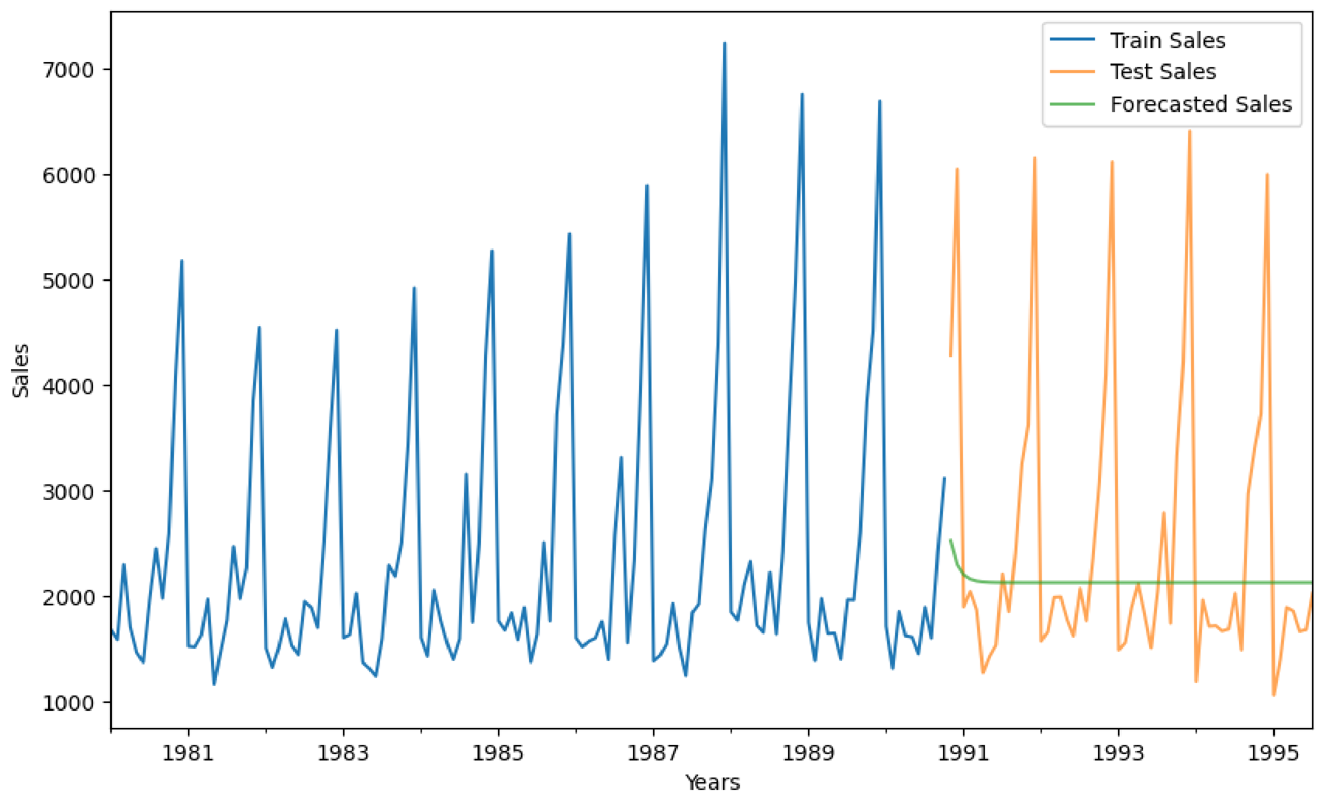

Figure 21 AR model forecast

We could see that the forecasted value is not close to the values of the test dataset.

Autoregressive Moving Average Model

For ARMA model, we take p as 3 and q as 3 as the best combination which is represented as (3,0,3). With this combination we have created the model as below:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  130
Model:                 ARIMA(3, 0, 3)   Log Likelihood                  76.069
Date:                Sat, 04 Jan 2025   AIC                           -136.137
Time:                        10:40:25   BIC                           -113.197
Sample:                    01-01-1980   HQIC                          -126.816
                         - 10-01-1990
Covariance Type:                  opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.3330      0.012    281.966      0.000       3.310       3.356
ar.L1          0.7439      0.162      4.596      0.000       0.427       1.061
ar.L2          0.7132      0.282      2.531      0.011       0.161       1.265
ar.L3         -0.9872      0.163     -6.066      0.000      -1.306      -0.668
ma.L1         -0.7655      0.205     -3.733      0.000      -1.167      -0.364
ma.L2         -0.7360      0.364     -2.021      0.043      -1.450      -0.022
ma.L3          0.9614      0.257      3.745      0.000       0.458       1.465
sigma2         0.0177      0.003      5.138      0.000       0.011       0.024
==============================================================================
```

Figure 22 ARMA Model

The Root Mean Squared Error of our forecasts is 1021.238. The graph is as below for this model:

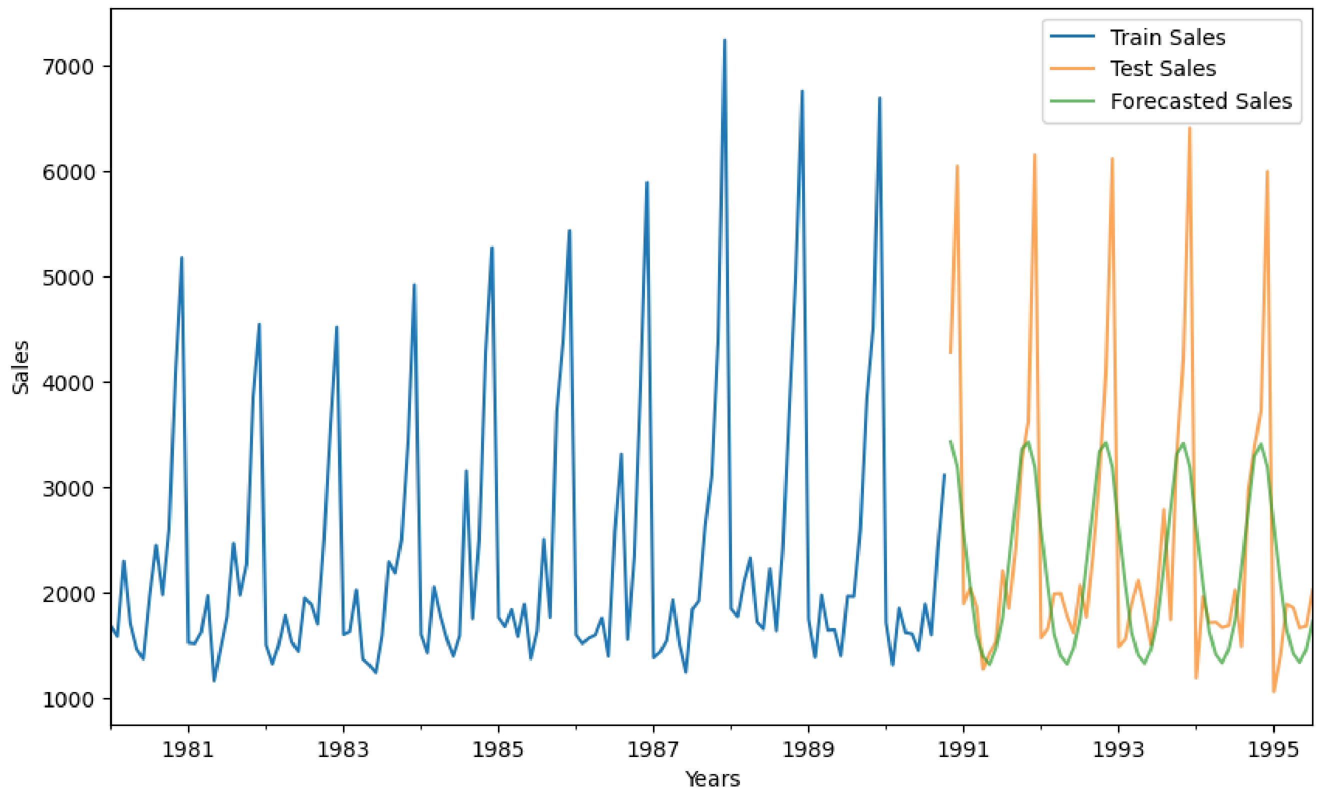

Figure 23 ARMA Model forecast

We can see that the model is performing better than the AR model since this considers the seasonal effect. Let's move ahead and use all the three best combination to achieve a better RMSE.

ARIMA Model

In the ARIMA model, we take p as 3, d as 0 and q as 3 which is represented as (3,0,3). This is the same as the ARMA model since d is 0 which means the model does not require differencing.

```
                              SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  130
Model:                 ARIMA(3, 0, 3)   Log Likelihood                  76.069
Date:                Sat, 04 Jan 2025   AIC                           -136.137
Time:                        10:55:31   BIC                           -113.197
Sample:                    01-01-1980   HQIC                          -126.816
                         - 10-01-1990
Covariance Type:                  opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.3330      0.012    281.966      0.000       3.310       3.356
ar.L1          0.7439      0.162      4.596      0.000       0.427       1.061
ar.L2          0.7132      0.282      2.531      0.011       0.161       1.265
ar.L3         -0.9872      0.163     -6.066      0.000      -1.306      -0.668
ma.L1         -0.7655      0.205     -3.733      0.000      -1.167      -0.364
ma.L2         -0.7360      0.364     -2.021      0.043      -1.450      -0.022
ma.L3          0.9614      0.257      3.745      0.000       0.458       1.465
sigma2         0.0177      0.003      5.138      0.000       0.011       0.024
==============================================================================
```

Figure 24 ARIMA Model

The Root Mean Squared Error of our forecasts is 1021.238. The graph will be as same as ARMA model which is as below:

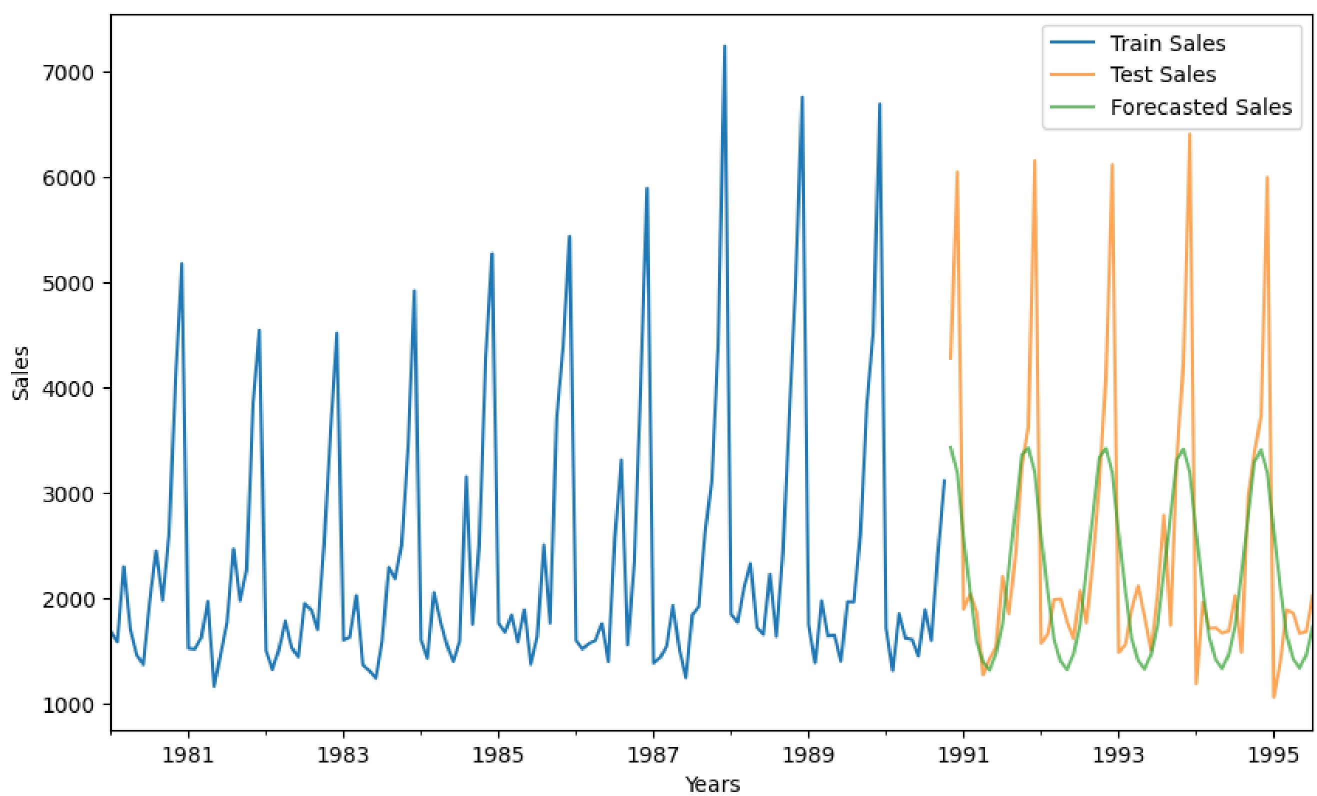

Figure 25 ARIMA Model forecast

The RMSE is low here and we could see the forecast is close, however we should also check other models to come to a conclusion.

SARIMA Model

A Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a statistical analysis model that uses time series data to predict future trends or understand the data set. It's a type of regression analysis that's useful for modeling time series with seasonal patterns, where the mean and other statistics for a given season aren't stationary across years. SARIMA models are an extension of the non-seasonal autoregressive-moving average (ARMA) and autoregressive integrated moving average (ARIMA) models. They combine the concepts of autoregressive (AR), integrated (I), and moving average (MA) models with seasonal components.

The value for the parameters (p,d,q) and (P, D, Q) can be decided by comparing different values for each and taking the lowest AIC value for the model build.

We start by plotting the ACF and PACF plot to check the autocorrelation.

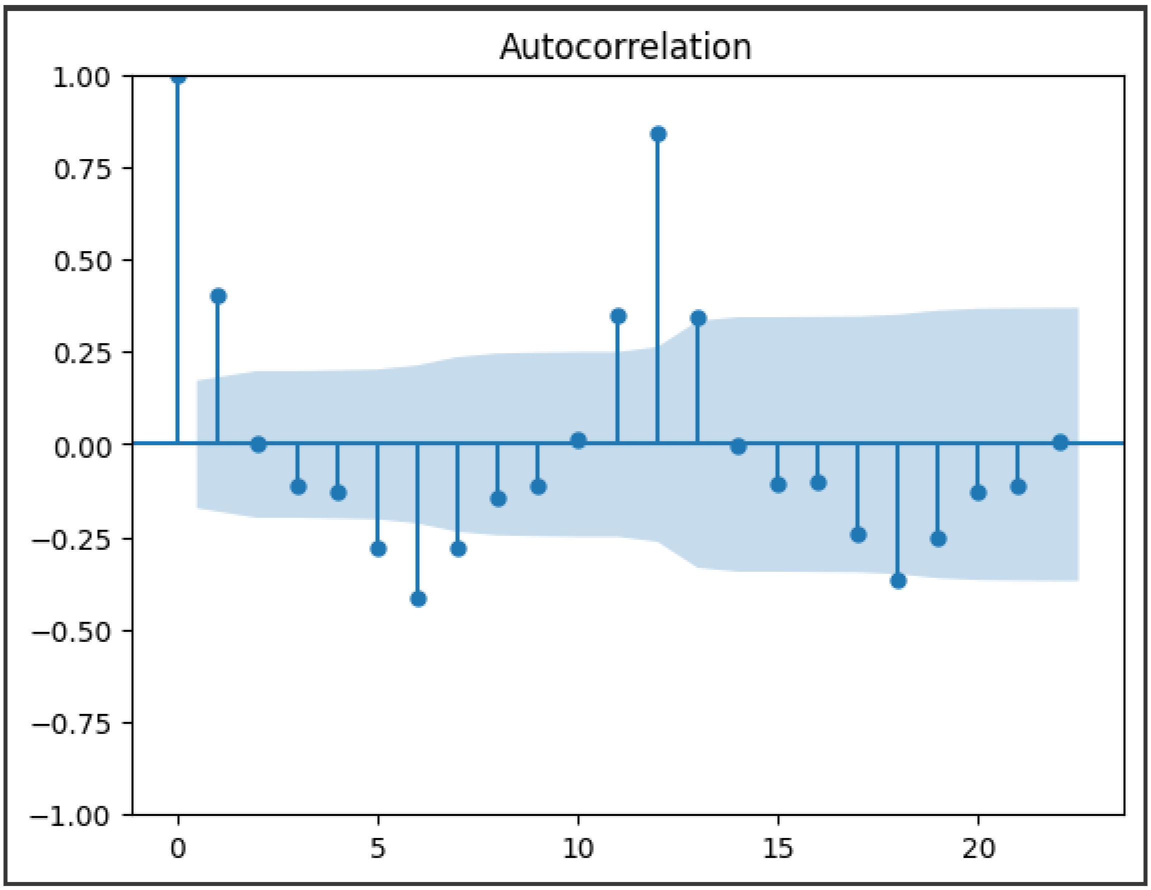

Figure 26 ACF Plot

Lag 1 has a significant positive autocorrelation, followed by rapid decay. There are spikes at higher lags (e.g., around lag 10) that are also outside the confidence interval, indicating the presence of seasonal autocorrelation.

For SARIMA model, the best combination we got is (1, 0, 1) (1, 0, 1, 12) where,

- p = 1 ; non-seasonal AR order
- d = 0 ; non-seasonal differencing
- q = 1 ; non-seasonal MA
- P = 1 ; seasonal AR Order
- D = 0 ; seasonal differencing
- Q = 1 ; seasonal MA
- S = 12 ; time span of repeating seasonal pattern

The model we built is as below:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                            Sparkling   No. Observations:                130
Model:             SARIMAX(1, 0, 1)x(1, 0, 1, 12)   Log Likelihood                146.198
Date:                              Sat, 04 Jan 2025   AIC                          -282.396
Time:                                      11:59:26   BIC                          -268.058
Sample:                                  01-01-1980   HQIC                         -276.570
                                       - 10-01-1990
Covariance Type:                                opg
==========================================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.9991      0.005    209.252      0.000       0.990       1.008
ma.L1         -0.9301      0.041    -22.884      0.000      -1.010      -0.850
ar.S.L12       0.9917      0.006    168.158      0.000       0.980       1.003
ma.S.L12      -0.6329      0.079     -7.985      0.000      -0.788      -0.478
sigma2         0.0045      0.001      8.065      0.000       0.003       0.006
------------------------------------------------------------------------------
```

Figure 27 SARIMA Model

The Root Mean Squared Error of our forecasts is 301.307.

The graph plot is as below:

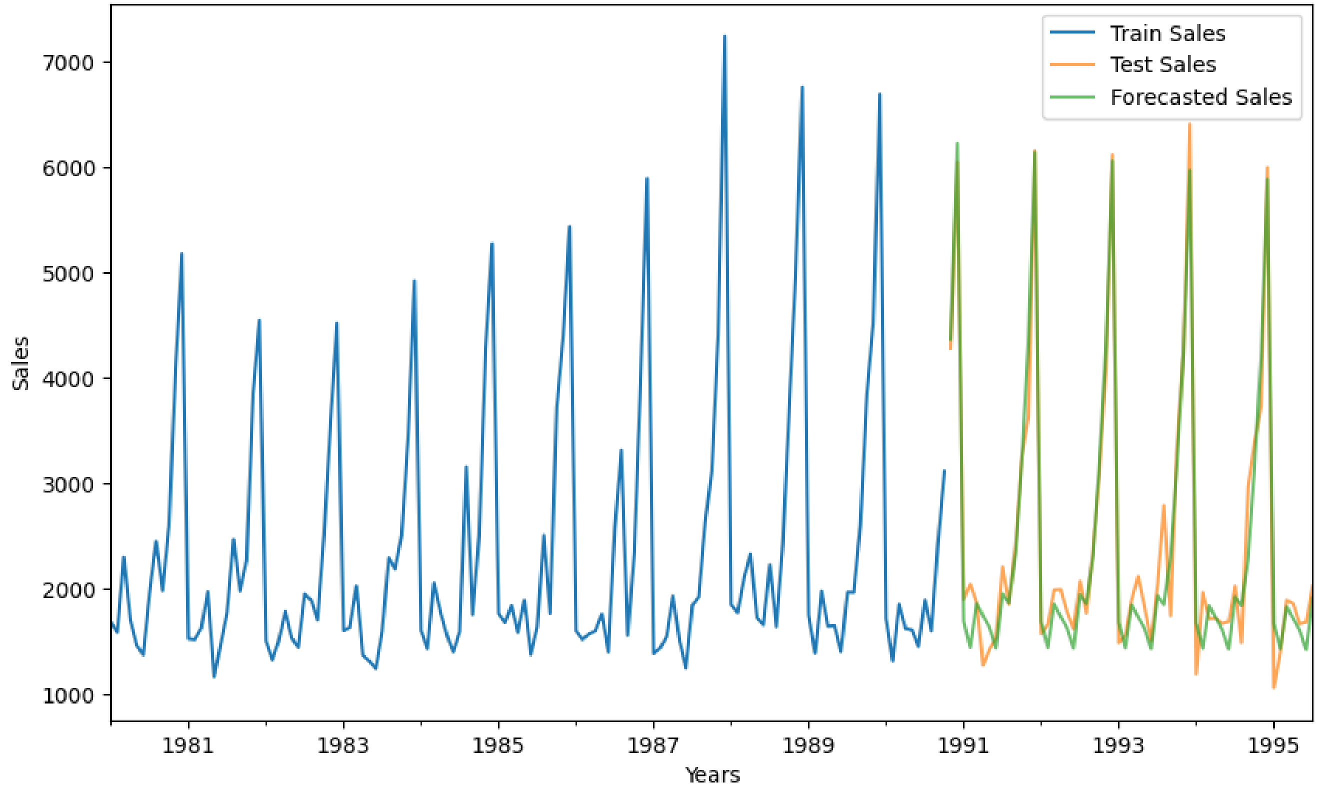

Figure 28 SARIMA Model forecast

We could see how the forecast line is moving in correspondence to the test dataset.

Model Comparison

We have built different models and let us see the RMSE by each model:

	RMSE
Best AR Model : ARIMA(1,0,0)	1392.613697
Best ARMA Model : ARIMA(3,0,3)	1021.237917
Best ARIMA Model : ARIMA(3,0,3)	1021.237917
Best SARIMA Model : SARIMAX(1, 0, 1)x(1, 0, 1, 12)	301.306506

	Test RMSE
Linear Regression	1392.438305
Simple Average	1368.746717
2pointTrailingMovingAverage	811.178937
4pointTrailingMovingAverage	1184.213295
6pointTrailingMovingAverage	1337.200524
9pointTrailingMovingAverage	1422.653281
Alpha=0.4,SimpleExponentialSmoothing	1363.037803
Alpha=0.3,Beta=0.3,DoubleExponentialSmoothing	2619.071230
Alpha=0.07,Beta= 0.03,Gamma=0.47,TripleExponentialSmoothing	366.859156
Alpha=0.01,Beta=0.31,Gamma=0.21,TripleExponentialSmoothing	319.794428

The lowest RMSE is gained through the Best SARIMA model (1,0,1) x (1,0,1,12). Hence we would use this model to forecast for the next 12 months.

Forecasting for next 12 months

We have used to predict the next 12 months which is 01-08-1995 to 31-07-1996. The forecast values are as below:

	forecast	lower_ci_95	upper_ci_95	lower_ci_99	upper_ci_99
1995-08-01	4369.292746	3224.330383	5920.832183	3224.330383	5920.832183
1995-09-01	6225.028206	4590.460796	8441.631001	4590.460796	8441.631001
1995-10-01	1696.337343	1250.016510	2302.017899	1250.016510	2302.017899
1995-11-01	1447.093587	1065.588057	1965.187050	1065.588057	1965.187050
1995-12-01	1866.421717	1373.387211	2536.451479	1373.387211	2536.451479

and the plot for the forecast is as below;

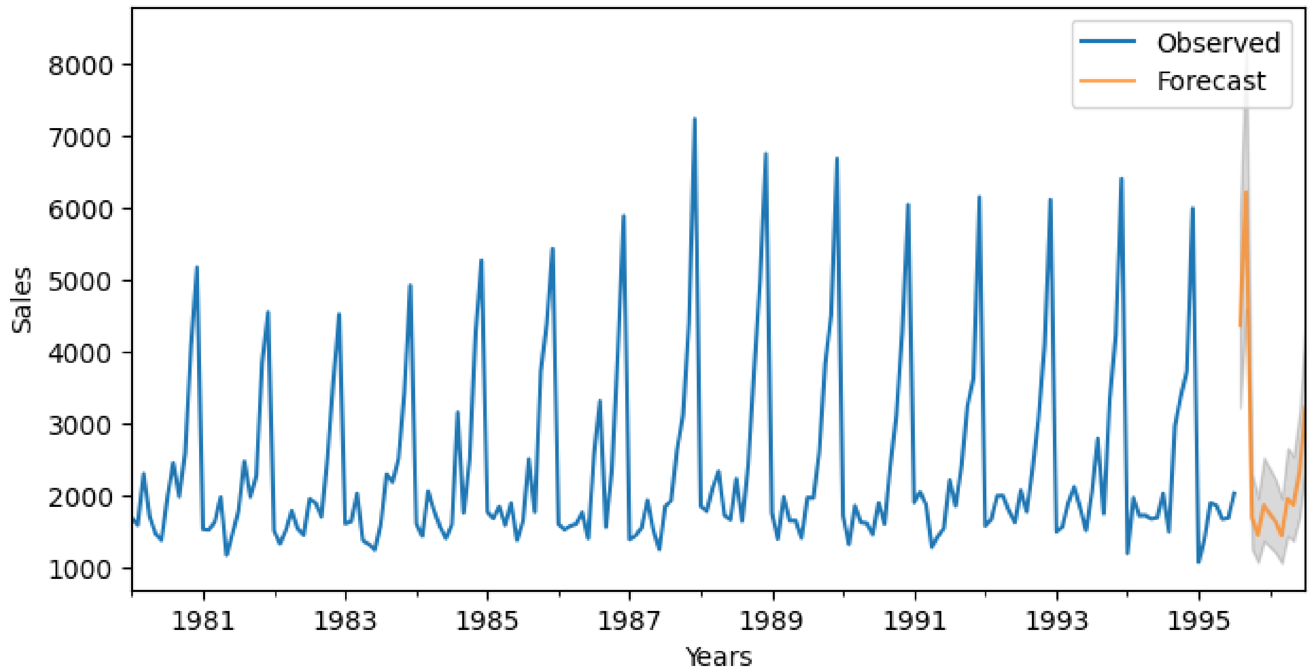

Figure 29 Forecast for next 12 months

Actionable Insights and Recommendations

Insights

1. **Seasonality in Sales:**
   - Sales show consistent peaks during November and December, driven by festive demand.
   - No long-term upward or downward trend in sales observed, though the range shows variability.
2. **Forecasting Model Performance:**
   - **Triple Exponential Smoothing (Holt-Winter Method)** performed best among the simple methods, achieving an RMSE of **319.79**.
   - **SARIMA Model** (1, 0, 1) x (1, 0, 1, 12) had the lowest RMSE at **301.31**, accurately capturing both trend and seasonality.
   - **Linear Regression** and simpler models (e.g., Moving Average, Simple Exponential Smoothing) performed poorly due to their inability to handle seasonality effectively.
3. **Data Characteristics:**
   - The dataset is non-stationary, requiring transformations (e.g., differencing) for advanced modeling.
   - Seasonal patterns (monthly, yearly) are significant, validated by ACF and PACF plots.
4. **Future Sales Forecast:**
   - Forecasts indicate that seasonality will persist, reinforcing the need for seasonal inventory and marketing strategies.

Recommendations

Operational Efficiency:

- Stock up inventories well in advance for the festive months (November-December).
- Use SARIMA model predictions to optimize production and logistics for the next 12 months.

Marketing Strategy:

- Develop promotional campaigns targeting November-December to capitalize on the seasonal peak.
- Explore cross-promotions or bundled offers during off-peak months to maintain steady sales.

Model Refinement:

- Continue exploring ARIMA and SARIMA variations with additional external variables (e.g., holiday events, promotions) to further improve accuracy.
- Incorporate external economic or market factors (e.g., consumer sentiment indices) into the models for a more comprehensive approach.

Automation and Scalability:

- Automate the forecasting process using the SARIMA model to generate monthly sales forecasts.
- Regularly re-train the model with updated data to maintain accuracy.

Scenario Planning:

- Use SARIMA forecasts for various "what-if" scenarios, such as disruptions in supply chains or unexpected demand spikes.
- Plan buffer stocks during anticipated high-demand periods or unforeseen disruptions.

Visualization and Communication:

- Present forecasts visually (e.g., line plots with actual vs. predicted values) to stakeholders for better understanding and decision-making.
- Highlight actionable metrics, such as RMSE improvements and forecast accuracy, to showcase model value.

Expand to Related Products:

- If data is available for other wine categories, replicate this process to understand their seasonality and trends for a holistic business strategy.