Data Analysis using Time Series Forecasting for ABC Estate Wines

Name: Aishwariya Hariharan
PGP-DSBA Online September’ 23
Date: 02 June 2024

Contents

List of Figures

List of Tables

Wine Sales Forecasting

Context

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 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.

Sparkling wine

Q1. Define the problem and perform Exploratory Data Analysis - - Read the data as an appropriate time series data - Plot the data - Perform EDA - Perform Decomposition

Data Description

Data Dictionary

```
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Sparkling  187 non-null    int64
dtypes: int64(1)
```

Figure 1: Data description

	Sparkling
YearMonth	
1980-01-01	1686
1980-02-01	1591
1980-03-01	2304
1980-04-01	1712
1980-05-01	1471

Table 1: Sample data

	count	mean	std	min	25%	50%	75%	max
Sparkling	187.0	2402.417112	1295.11154	1070.0	1605.0	1874.0	2549.0	7242.0

Table 2: Statistical description

- The column Sparkling lists the sales of Sparkling wine (integer values) from 1980 to 1995, i.e. 15 years.
- There are 187 rows in the data.
- The mean of the data, i.e. the average sales is 2402.41
- The median sales is 1874.
- There are no missing values in the data.

Exploratory Data Analysis

Plot of the data

- The data shows a slight positive trend with strong seasonality.

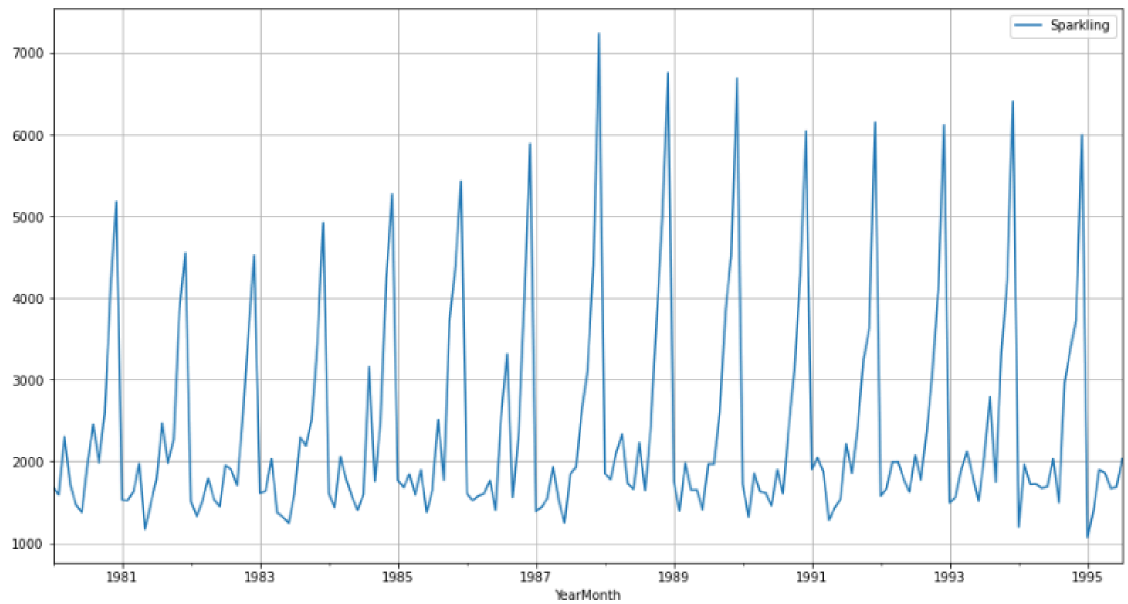

Figure 2: Plot of data

Distribution of data

- The data is right-skewed
- Most commonly sold quantity of sales of Sparkling wine is in the range of 1000 to 3000.

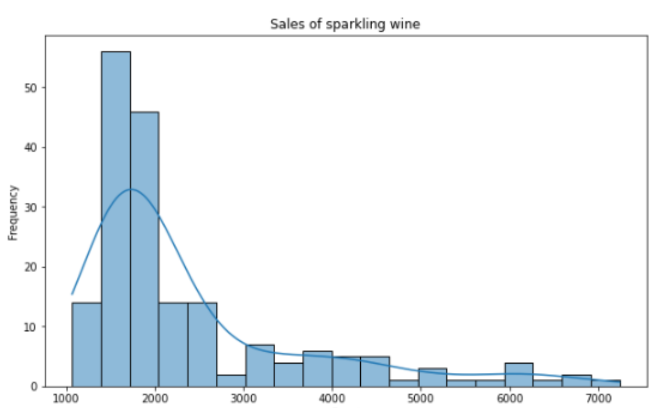

Figure 3: Data distribution

Yearly sales

- The median sales across years are almost the same.
- There are outlier cases.

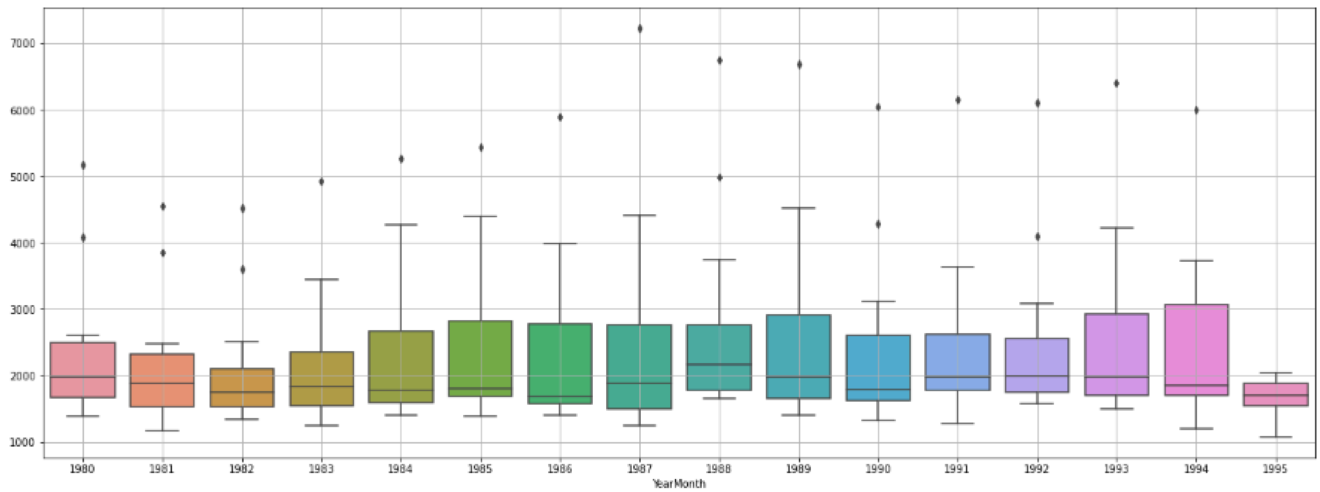

Figure 4: Yearly sales

Monthly sales and trend of months across years

- The sales of Sparkling wine are higher towards the end of the year, post August.

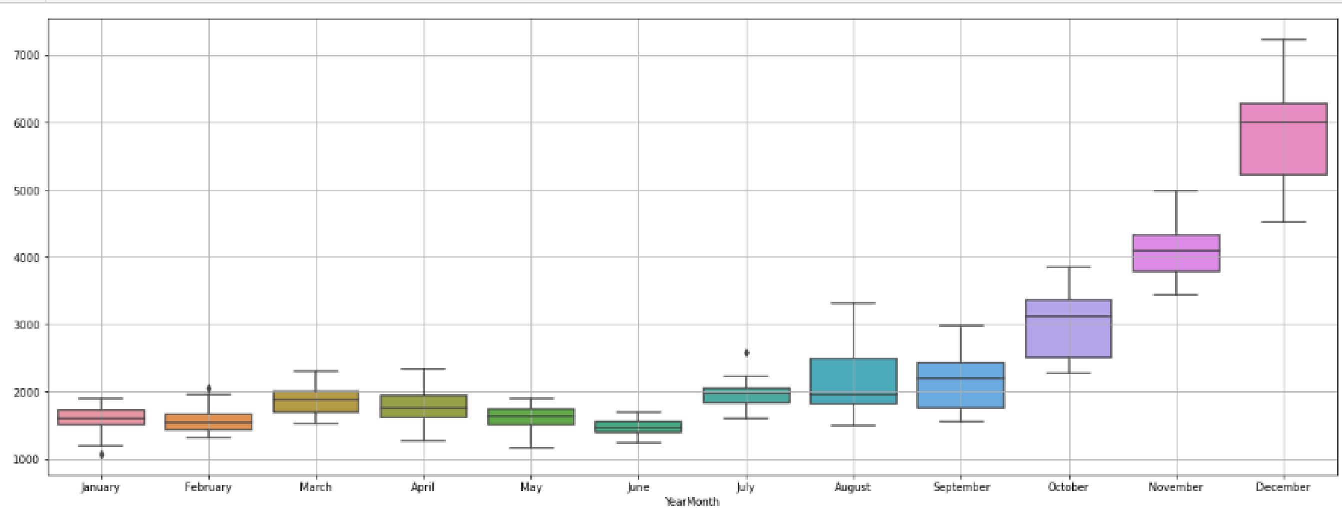

Figure 5: Monthly sales

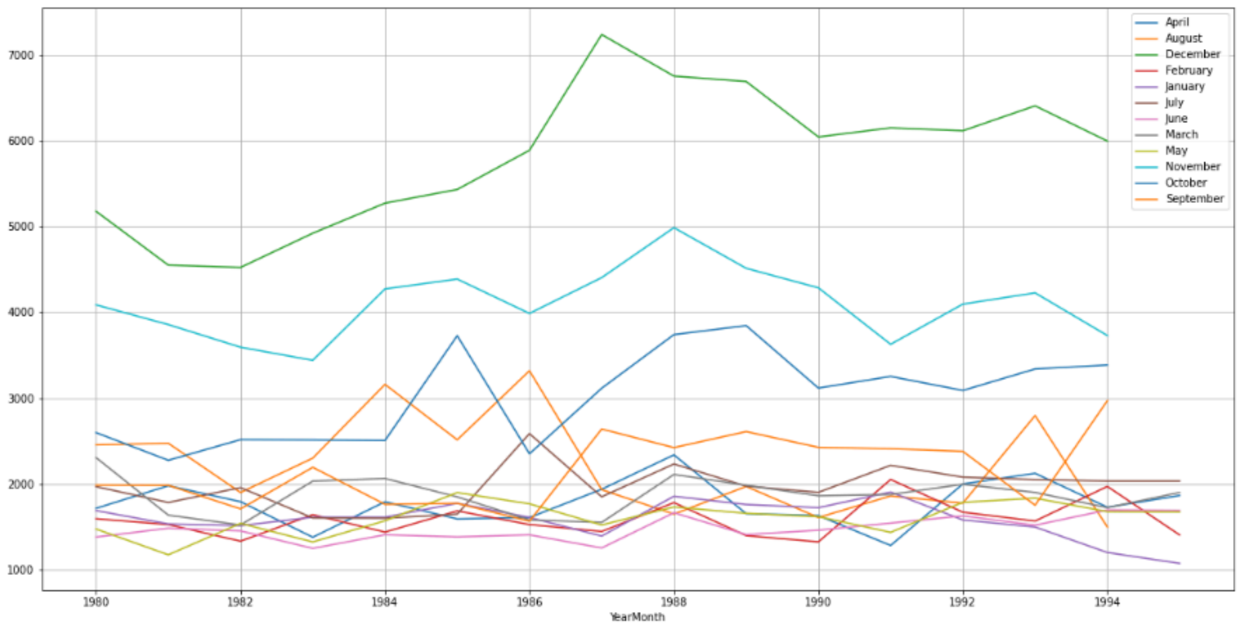

Figure 6: Trend of months across years

Additive decomposition

- The data shows an overall positive trend.
- The seasonality varies between -1000 to 3000
- The residuals are random.

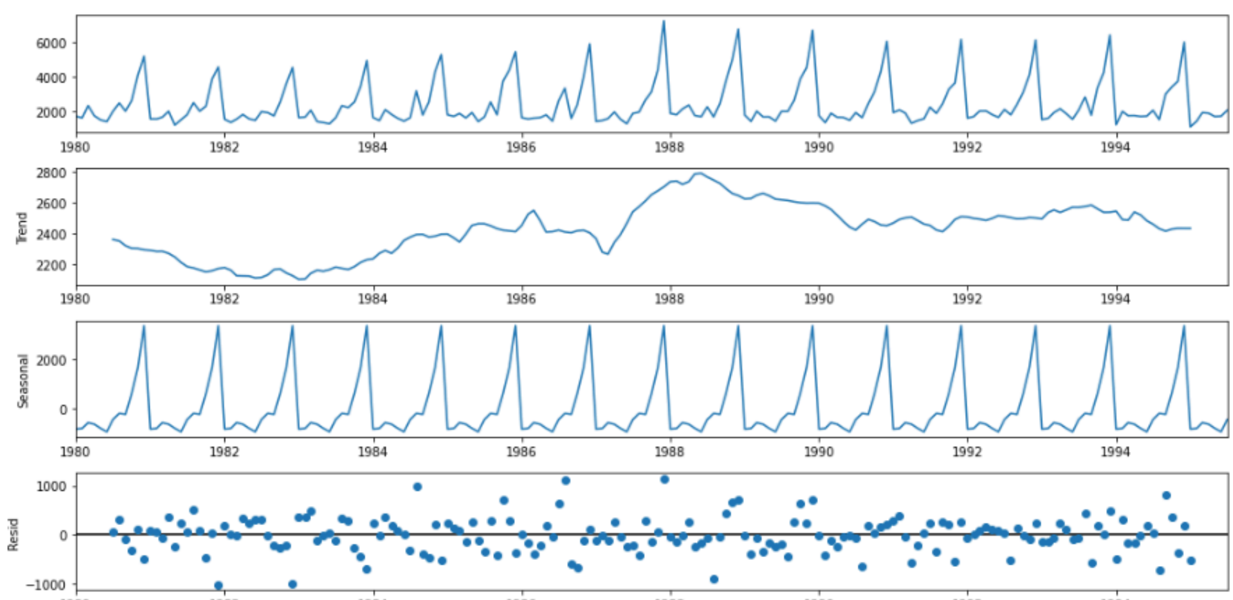

Figure 7: Additive decomposition

Q2. Data Pre-processing - Missing value treatment - Visualize the processed data - Train-test split

There are no missing values in the data.

Splitting the data into train and test set.

Number of rows in training data = 130
Number of rows in testing data = 57

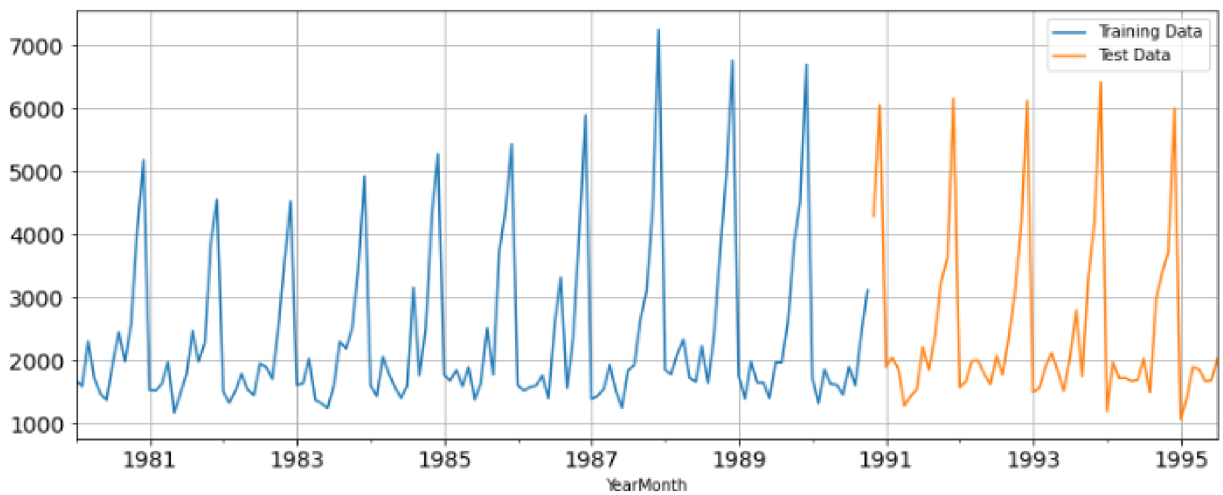

Figure 8: Splitting the data into train and test

Q3. Model Building - Original Data - - Build forecasting models - Linear regression - Simple Average - Moving Average - Exponential Models (Single, Double, Triple) - Check the performance of the models built

Performance of the models based on RMSE scores

- The RMSE score measures the prediction accuracy. The lower the better.
- The triple exponential smoothening model has the lowest RMSE score. Therefore, it tracks the level, trend, and seasonality very well.

	RMSE
Linear Regression Model	1374.550202
Simple Average Model :	1368.746717
Moving Average Model :	811.178937
Single Exponential Smoothening Model :	1363.702251
Double Exponential Smoothening Model	1472.253632
Triple Exponential Smoothening Model Additive:	366.859156

Table 3: RMSE performance

Plots forecasting the sales based on the models for sparkling wine

- Linear regression model

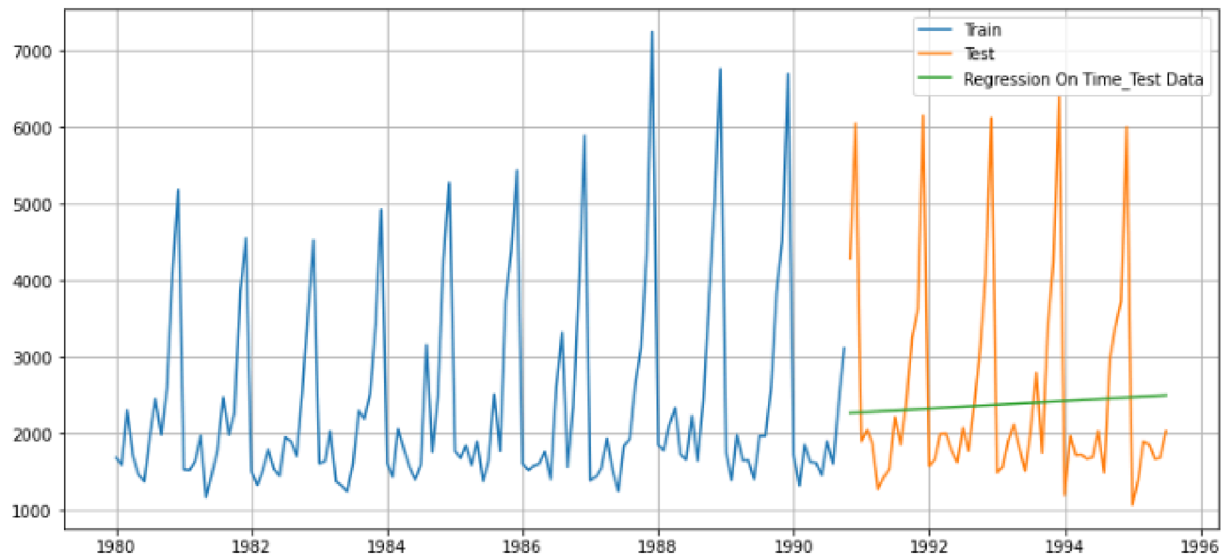

Figure 9: Linear regression model

- Simple average

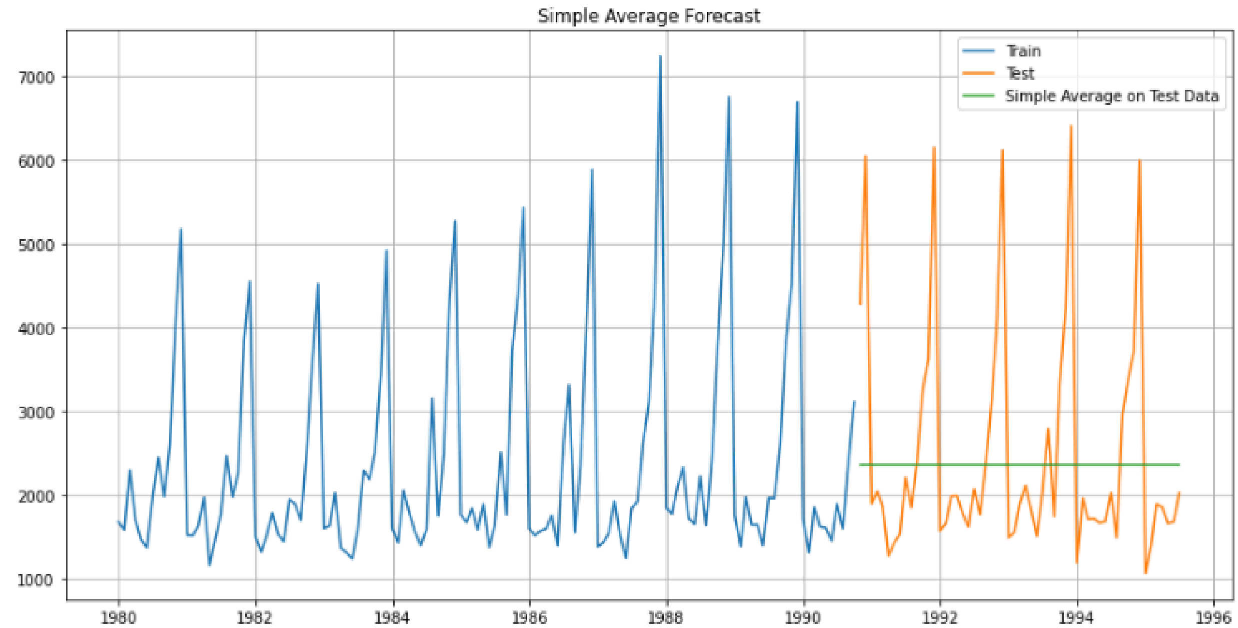

Figure 10: Simple average model

- Moving average

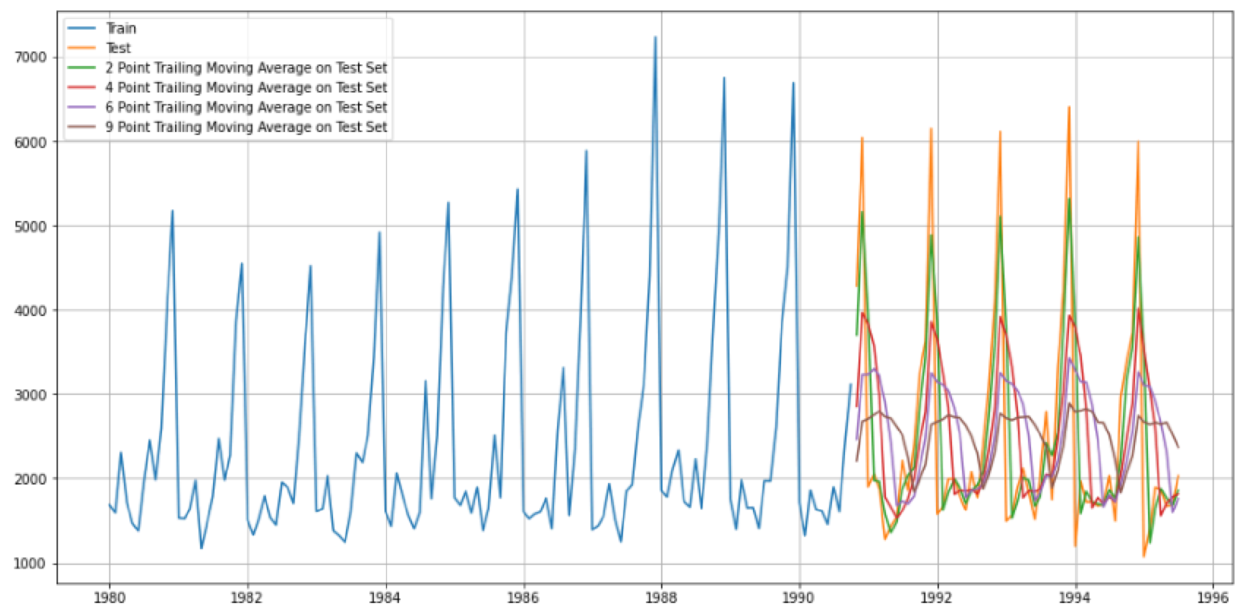

Figure 11: Moving average

- Single exponential smoothening

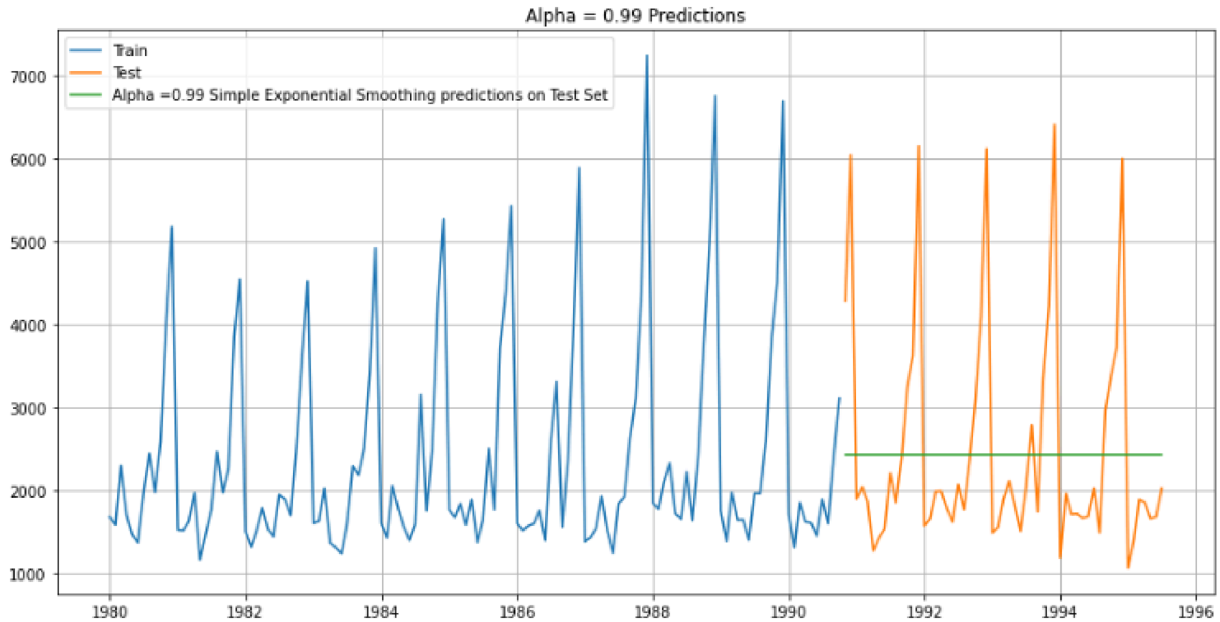

Figure 12: Single exponential smoothening

- Double exponential smoothening

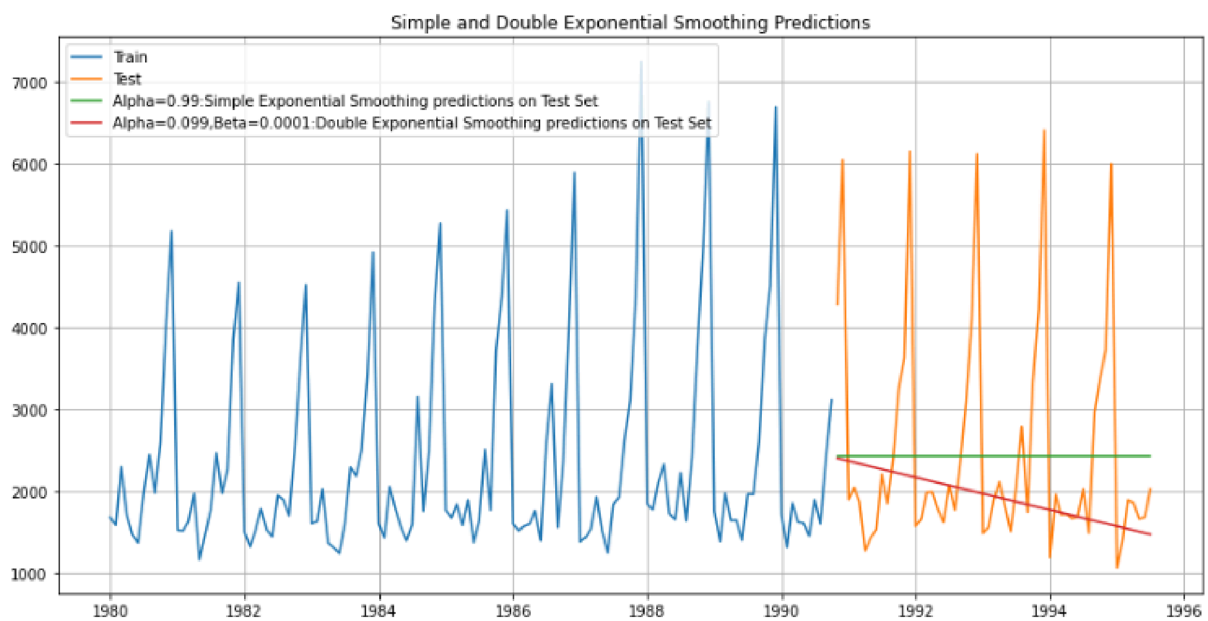

Figure 13: Double exponential smoothening

- Triple exponential smoothening

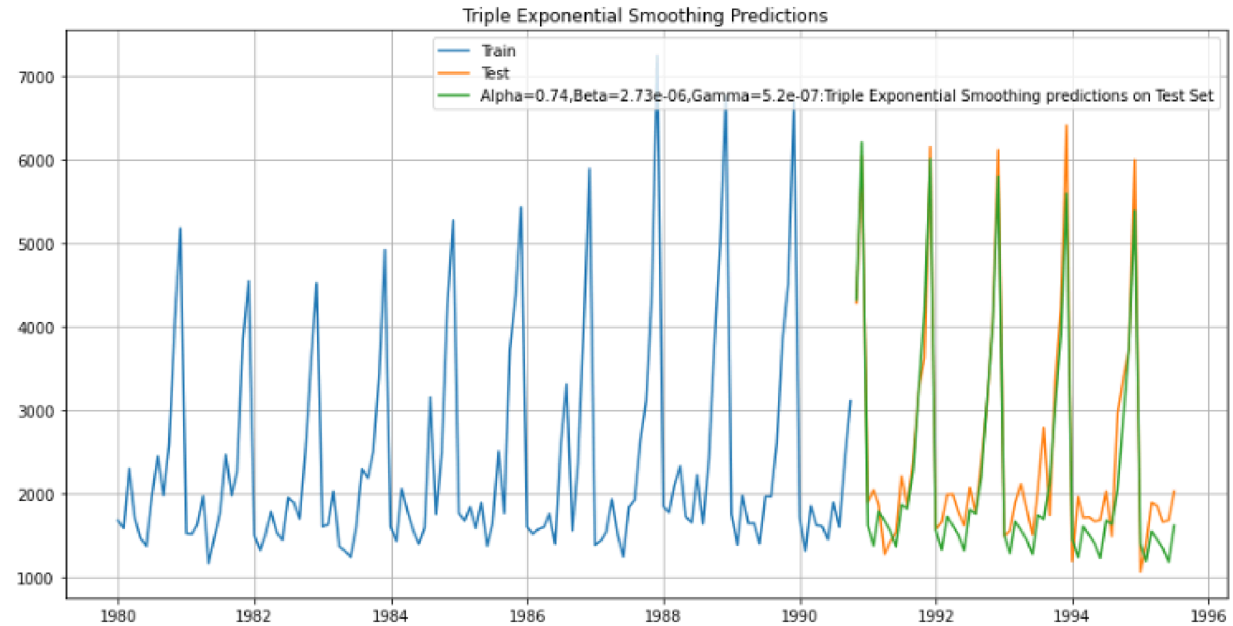

Figure 14: Triple exponential smoothening

Q4. Check for Stationarity - Check for stationarity - Make the data stationary (if needed)

Using the Augmented Dickey-Fuller Test:

Null Hypothesis: Time series is not stationary
Alternative Hypothesis: Time series is stationary

P-value = 0.6011

0.6011 > 0.05

Therefore, the time series is not stationary.

We take the difference of the time series with period = 1 and re-run the test:

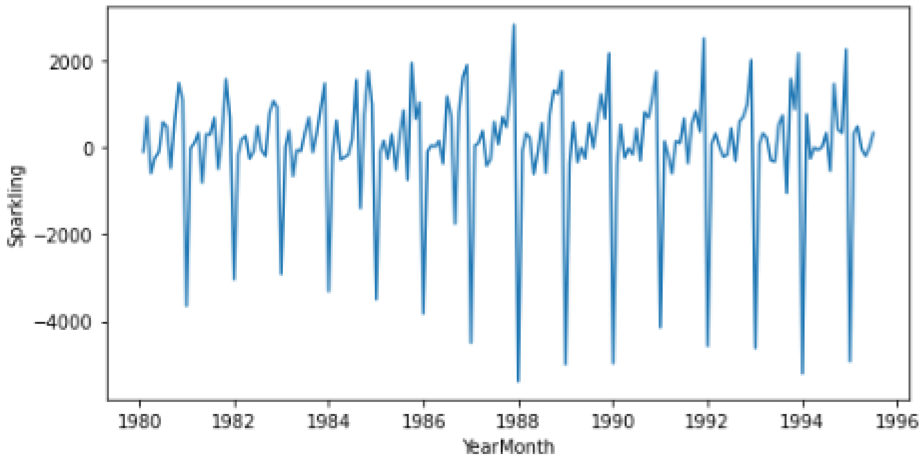

Figure 15: 1st differencing

- We observe seasonality even after differencing. This could imply that the variance in the data is increasing.
- We stop differencing at 1 since the data now looks stationary.

Q5. Model Building - Stationary Data

Model building

ACF plot

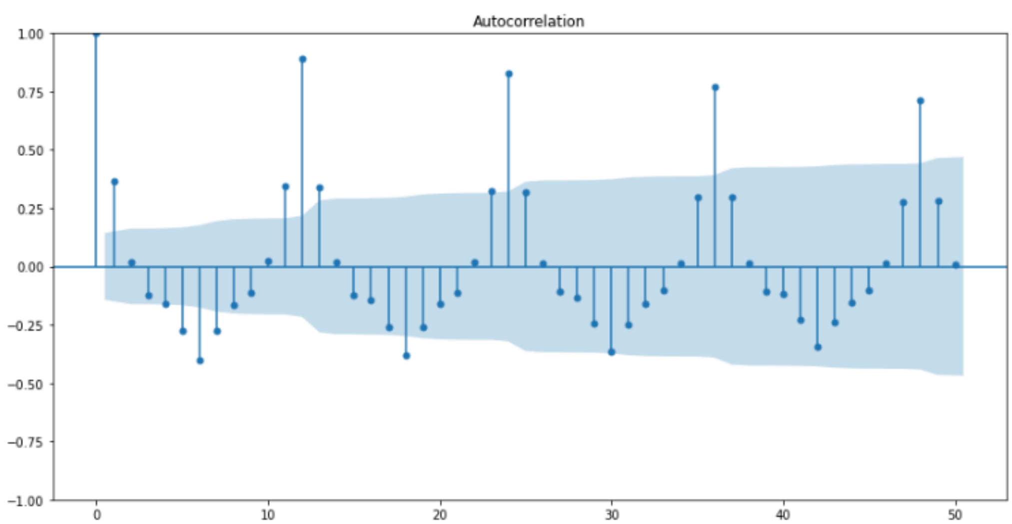

Figure 16: ACF plot

PACF Plot

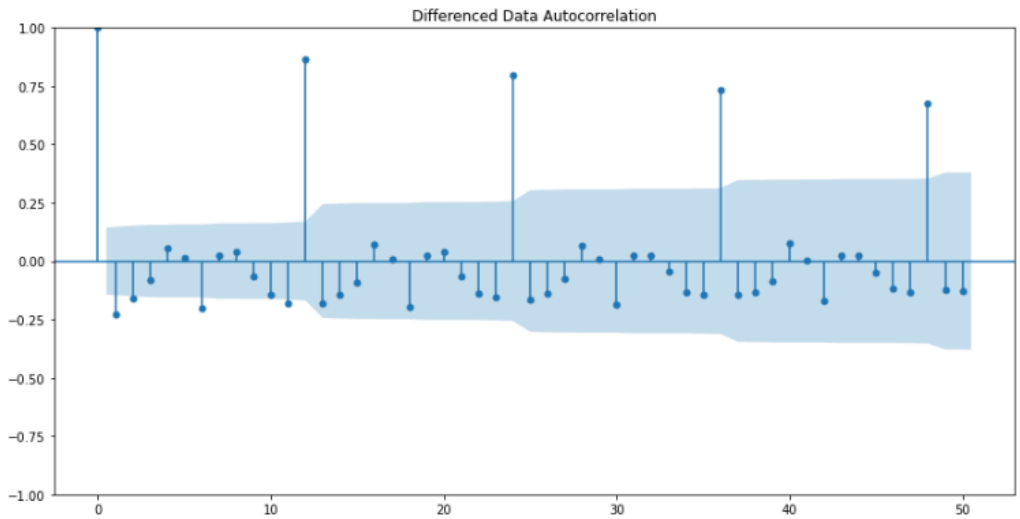

Figure 17: PACF plot

- p(AR) has 2 significant lags

ARIMA Models

Auto ARIMA

- For an Auto-ARIMA, we calculate the best p and q parameters by looking at the lowest corresponding Akaike Information Criterion (AIC) values.
- Here, the lowest AIC is for (2, 0, 1).

	param	AIC
7	(2, 0, 1)	2197.084442
1	(0, 0, 1)	2204.869799
6	(2, 0, 0)	2204.880722
2	(0, 0, 2)	2206.111207
4	(1, 0, 1)	2206.142158
5	(1, 0, 2)	2207.163048
3	(1, 0, 0)	2207.502101
8	(2, 0, 2)	2208.120889
0	(0, 0, 0)	2228.48366

Table 4: Auto ARIMA-AIC

Summary of ARIMA (2, 0, 1)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  130
Model:                 ARIMA(2, 0, 1)   Log Likelihood               -1093.542
Date:                Sun, 02 Jun 2024   AIC                           2197.084
Time:                        08:21:36   BIC                           2211.422
Sample:                    01-01-1980   HQIC                          2202.910
                         - 10-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       2379.9379    112.866     21.086      0.000    2158.725    2601.151
ar.L1          1.2114      0.135      8.991      0.000       0.947       1.475
ar.L2         -0.4998      0.124     -4.046      0.000      -0.742      -0.258
ma.L1         -0.8128      0.152     -5.349      0.000      -1.111      -0.515
sigma2      1.182e+06   1.28e+05      9.214      0.000    9.31e+05    1.43e+06
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                39.43
Prob(Q):                              0.90   Prob(JB):                         0.00
Heteroskedasticity (H):               2.19   Skew:                             0.96
Prob(H) (two-sided):                  0.01   Kurtosis:                         4.90
===================================================================================
```

Figure 18: ARIMA(2, 0, 1)

RMSE = 1338.13987

Manual ARIMA Model

From ACF and PACF, we get p = 2 and q = 2, with d = 1.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:              Sparkling   No. Observations:                  130
Model:                 ARIMA(2, 1, 2)   Log Likelihood                  52.483
Date:                Sun, 02 Jun 2024   AIC                            -94.966
Time:                        08:21:37   BIC                            -80.667
Sample:                    01-01-1980   HQIC                           -89.156
                         - 10-01-1990
Covariance Type:                  opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0013      0.010     -0.132      0.895      -0.021       0.018
ar.L2         -1.0000      0.005   -196.015      0.000      -1.010      -0.990
ma.L1          0.0062      0.112      0.056      0.956      -0.213       0.226
ma.L2          0.9999     22.870      0.044      0.965     -43.824      45.824
sigma2         0.0246      0.562      0.044      0.965      -1.078       1.127
==============================================================================
```

Figure 19: ARIMA (2, 1, 2)

RMSE = 1843.421

SARIMA Models

Auto SARIMA

- (0, 0, 2) (2, 0, 2, 12) has the lowest SARIMA value.

	param	seasonal	AIC
26	(0, 0, 2)	(2, 0, 2, 12)	1534.072513
53	(1, 0, 2)	(2, 0, 2, 12)	1534.4277
23	(0, 0, 2)	(1, 0, 2, 12)	1535.130924
80	(2, 0, 2)	(2, 0, 2, 12)	1536.287014
77	(2, 0, 2)	(1, 0, 2, 12)	1536.302588

Table 5: Auto SARIMA (0, 0, 2) (2, 0, 2, 12)

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  130
Model:             SARIMAX(0, 0, 2)x(2, 0, 2, 12)   Log Likelihood                -760.036
Date:                            Sun, 02 Jun 2024   AIC                           1534.073
Time:                                    08:51:57   BIC                           1552.516
Sample:                                         0   HQIC                          1541.543
                                            - 130
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1          0.1863      0.101      1.840      0.066      -0.012       0.385
ma.L2         -0.1082      0.120     -0.902      0.367      -0.343       0.127
ar.S.L12       0.6572      0.720      0.913      0.361      -0.753       2.067
ar.S.L24       0.3848      0.744      0.517      0.605      -1.073       1.843
ma.S.L12       0.5441      3.347      0.163      0.871      -6.016       7.104
ma.S.L24      -3.6923      5.366     -0.688      0.491     -14.209       6.824
sigma2      1.094e+04    3.2e+04      0.341      0.733   -5.19e+04    7.37e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):                20.03
Prob(Q):                              0.83   Prob(JB):                         0.00
Heteroskedasticity (H):               1.34   Skew:                             0.50
Prob(H) (two-sided):                  0.39   Kurtosis:                         4.91
===================================================================================
```

Figure 20: Auto SARIMA (0, 0, 2) (2, 0, 2, 12)

RMSE = 641.08

Manual SARIMA Model

From ACF and PACF, we get p = 1 and q = 2, with d = 1 and
P = 1, D = 1, Q = 2, S = 12

Summary of SARIMA (1, 2, 2) (1, 1, 2, 12)

```
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                              Sparkling   No. Observations:              130
Model:             SARIMAX(1, 1, 2)x(1, 1, [1], 12)   Log Likelihood             145.652
Date:                               Sun, 02 Jun 2024   AIC                       -279.305
Time:                                       10:47:35   BIC                       -262.732
Sample:                                   01-01-1980   HQIC                      -272.576
                                        - 10-01-1990
Covariance Type:                                 opg
==========================================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7265      0.393     -1.851      0.064      -1.496       0.043
ma.L1         -0.1068      0.362     -0.295      0.768      -0.816       0.602
ma.L2         -0.7328      0.326     -2.249      0.025      -1.371      -0.094
ar.S.L12       0.1464      0.168      0.869      0.385      -0.184       0.477
ma.S.L12      -0.7265      0.172     -4.220      0.000      -1.064      -0.389
sigma2         0.0045      0.001      6.966      0.000       0.003       0.006
==============================================================================
```

Figure 21: SARIMA (1, 2, 2) (1, 1, 2, 12)

RMSE = 303.552

The Best SARIMA RMSE value 303.552 is close to the Triple Exponential Smoothening RMSE value, which is 366.85.

Best ARIMA and SARIMA models

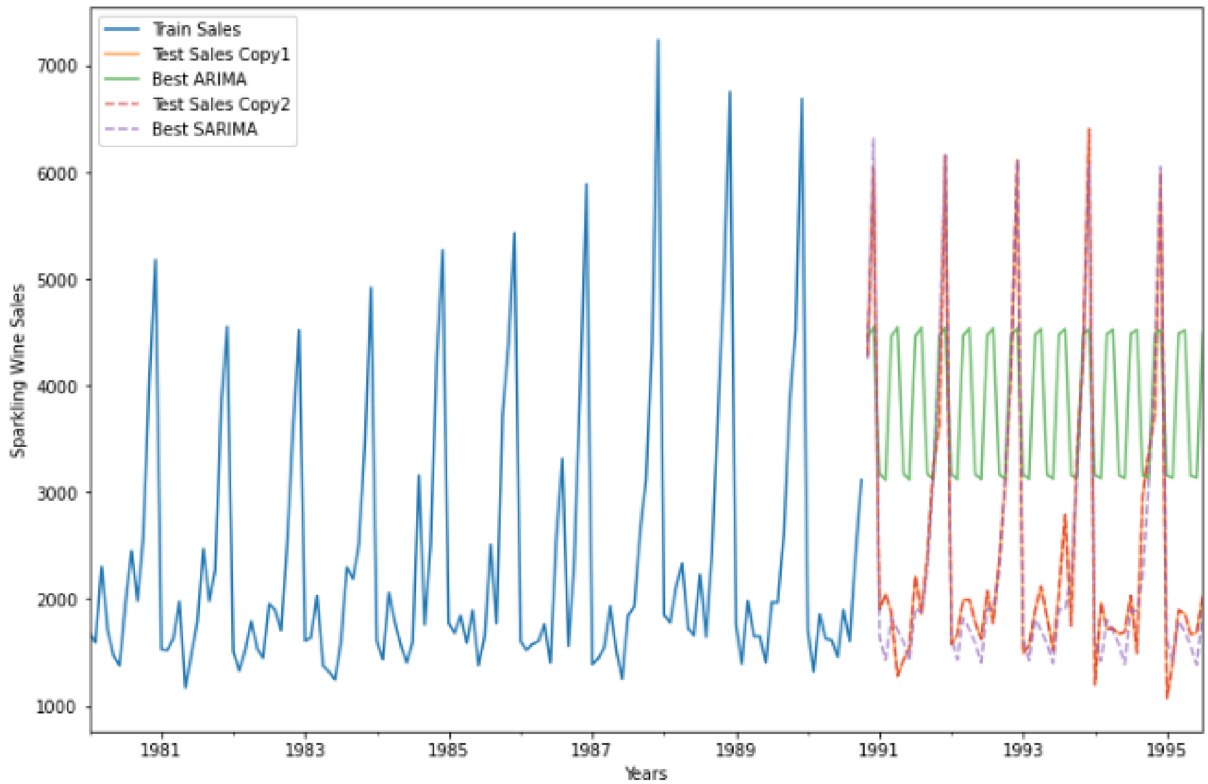

Figure 22: Best ARIMA and SARIMA

Q6. Compare the performance of models - - Compare the performance of all the models built - Choose the best model with proper rationale - Rebuild the best model using the entire data - Make a forecast for the next 12 months

Comparison of model performance

	RMSE
Linear Regression Model	1374.55
Simple Average Model :	1368.75
Moving Average Model :	811.18
Single Exponential Smoothening Model :	1363.70
Double Exponential Smoothening Model	1472.25
Triple Exponential Smoothening Model Additive:	366.86
Triple Exponential Smoothening Model Multiplica...	381.66
Best AR Model :	1398.06
Best ARMA Model:	1021.27
Best ARIMA Model :	1843.42
Best SARIMA Model :	303.55

Table 6: Model performance comparison

- The triple exponential and SARIMA models have the lowest scores.

Forecasting

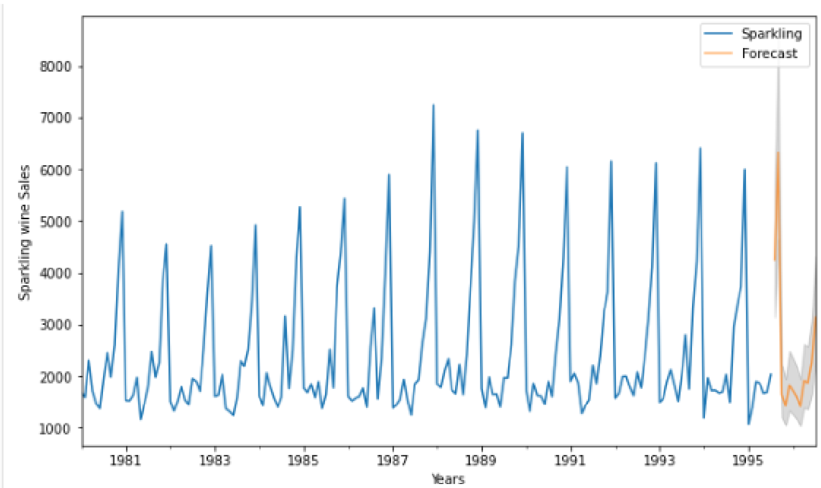

Figure 23: Forecasting

- The forecasted sales of Sparkling from August'95 to July'96 look like the given chart.
- We can observe that the overall forecasted sales are similar to the duration of 1993-94 and higher than August'94 to July'95.
- The confidence interval band increases as the duration of forecast increases.

Insights and recommendations

Insights

- The SARIMA model's parameters reveal a significant seasonal dependency on the sales.
- The 12th order lag residual in the 1st differential series hints the importance of monitoring sales of corresponding months of previous years.

Recommendation

- Continuously monitor sales trends, seasonal highs and lows.
- Adjust inventory levels and modify marketing strategy accordingly.
- Inventory during year ends must be carefully and strategically managed due to high sales during year end and also increase profitability during this period.