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.

Rose 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

```
DatetimeIndex: 187 entries, 1980-01-01 to 1995-07-01
Data columns (total 1 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rose    185 non-null    float64
dtypes: float64(1)
```

Figure 1: Data description

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

Table 1: Sample data

	count	mean	std	min	25%	50%	75%	max
Rose	185.0	90.394595	39.175344	28.0	63.0	86.0	112.0	267.0

Table 2: Statistical description

- The column Rose lists the sales of Rose 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 90.39
- The median sales is 86.
- There are 2 missing values in the data, which are replaced with the median values.

Exploratory Data Analysis

Plot of the data

- The data shows a strong negative trend with strong seasonality.

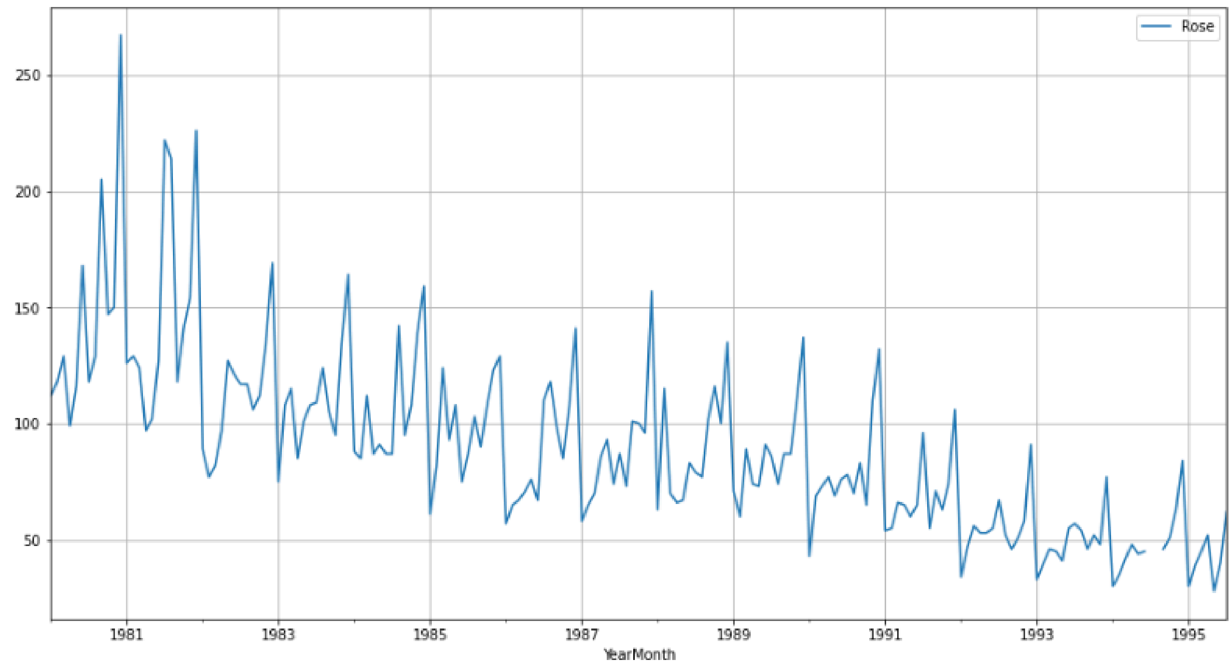

Figure 2: Plot of data

Distribution of data

- The data is right-skewed
- Most commonly sold quantity of sales of Sparkling wine is in the range of 10 to 150.

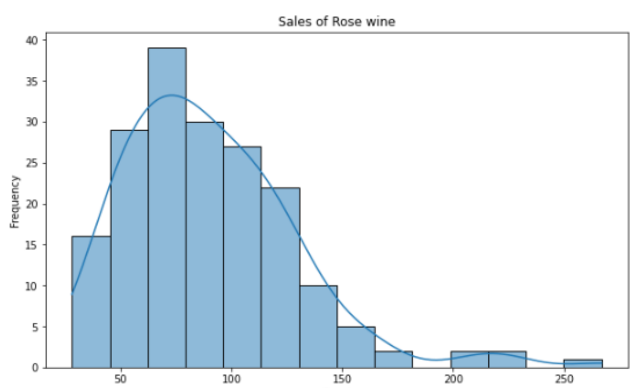

Figure 3: Data distribution

Yearly sales

- There's sharp decline in sales over time.
- There are outlier cases.

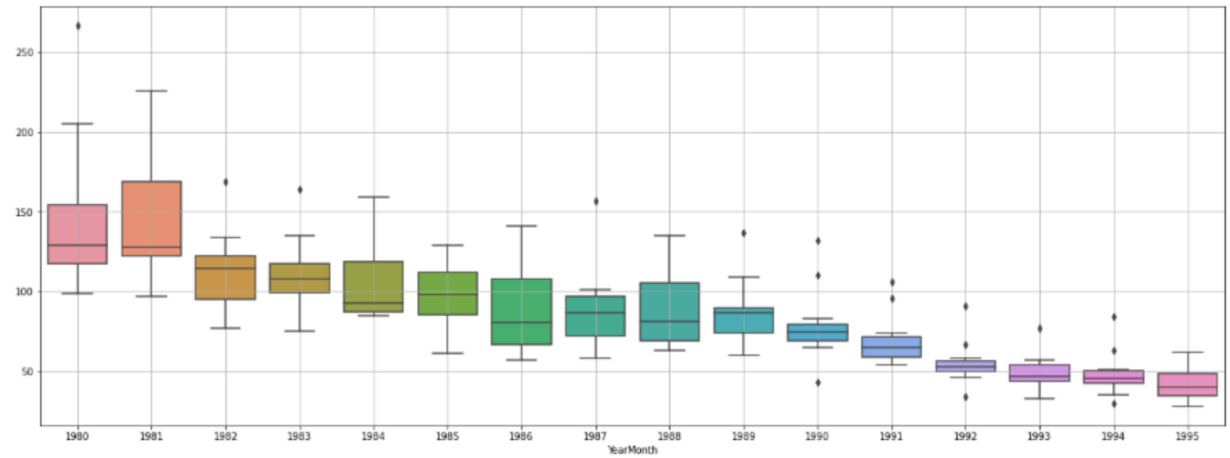

Figure 4: Yearly sales

Monthly sales and trend of months across years

- The sales of Rose wine are higher towards the end of the year.

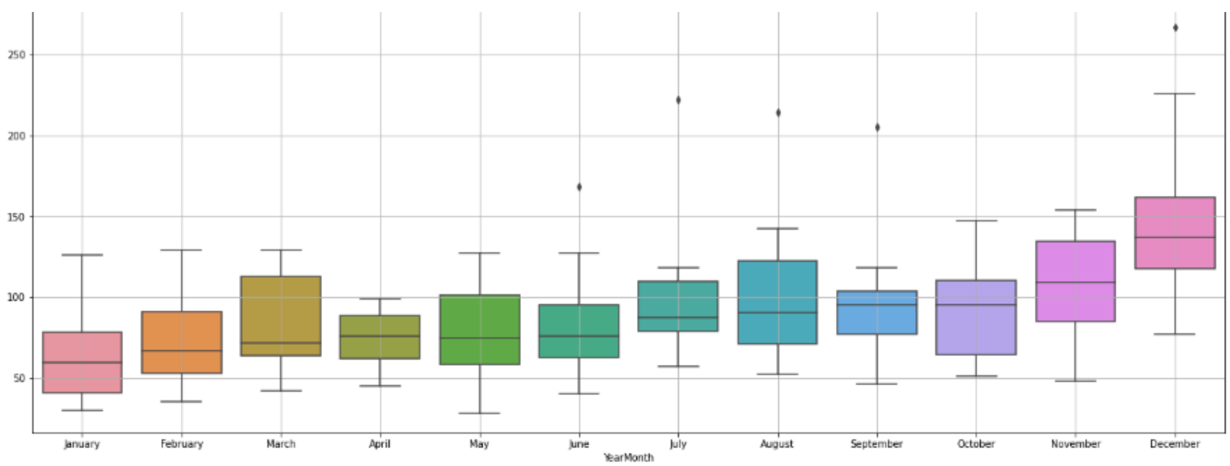

Figure 5: Monthly sales

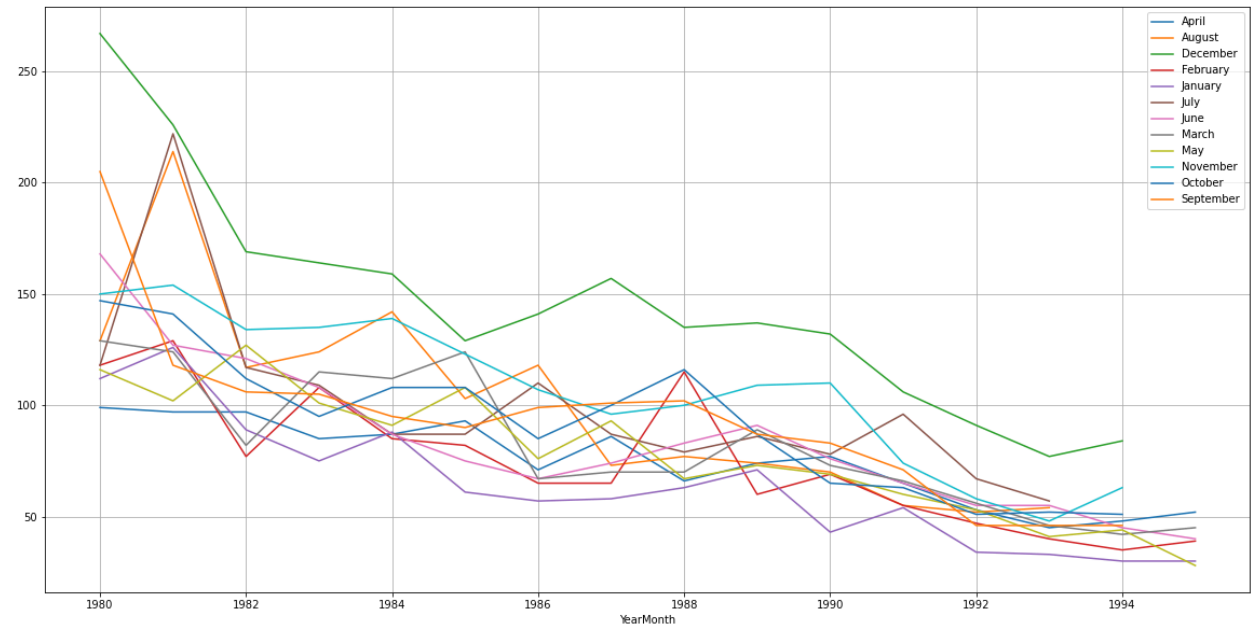

Figure 6: Trend of months across years

Additive decomposition

- Overall negative trend with seasonality varying from -25 to 60 with random residuals

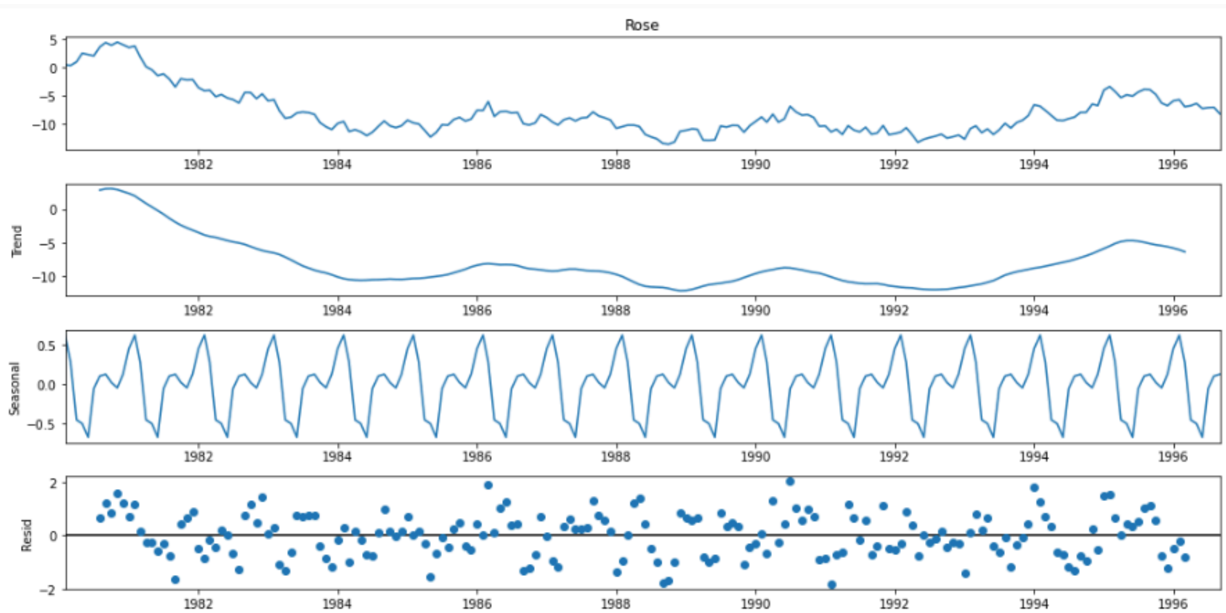

Figure 7: Additive decomposition

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

There are 2 missing values in the data. They are replace with median values.

Splitting the data into train and test set.

Number of rows in training data = 140
Number of rows in testing data = 60

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 double and triple exponential smoothening model has the lowest RMSE score. Therefore, it tracks the level, trend, and seasonality very well.

	RMSE
Linear Regression Model	17.244390
Simple Average Model :	7.866237
Moving Average Model :	0.200754
Single Exponential Smoothening Model :	17.369614
Double Exponential Smoothening Model	7.094577
Triple Exponential Smoothening Model Additive:	7.252973
Triple Exponential Smoothening Model Multiplicative:	7.252973

Table 3: RMSE performance

Plots forecasting the sales based on the models for sparkling wine

- Linear regression model

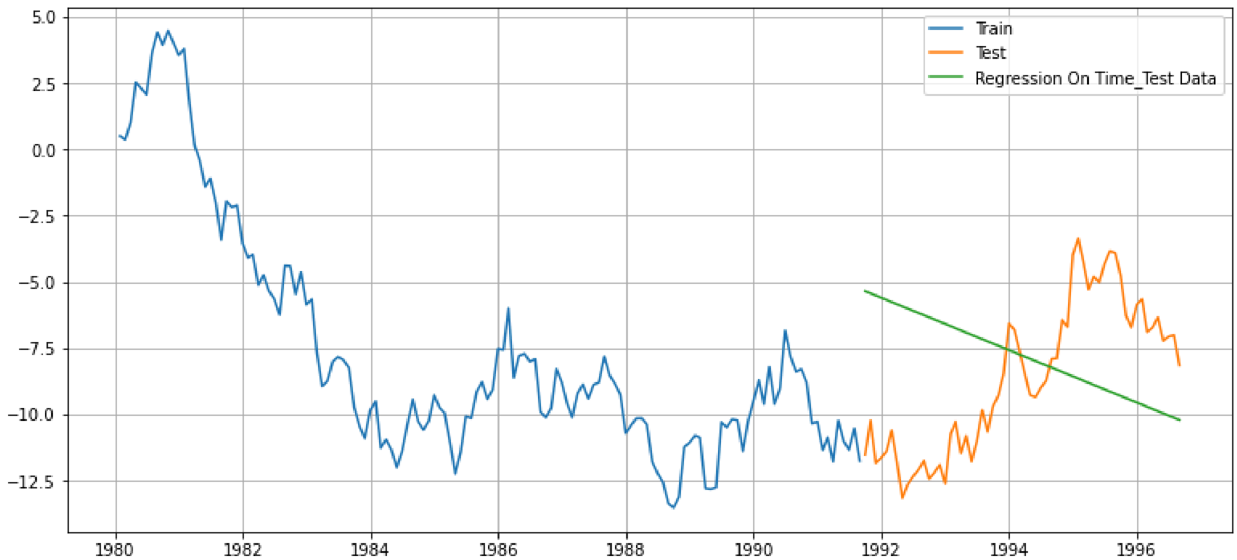

Figure 9: Linear regression model

- Simple average

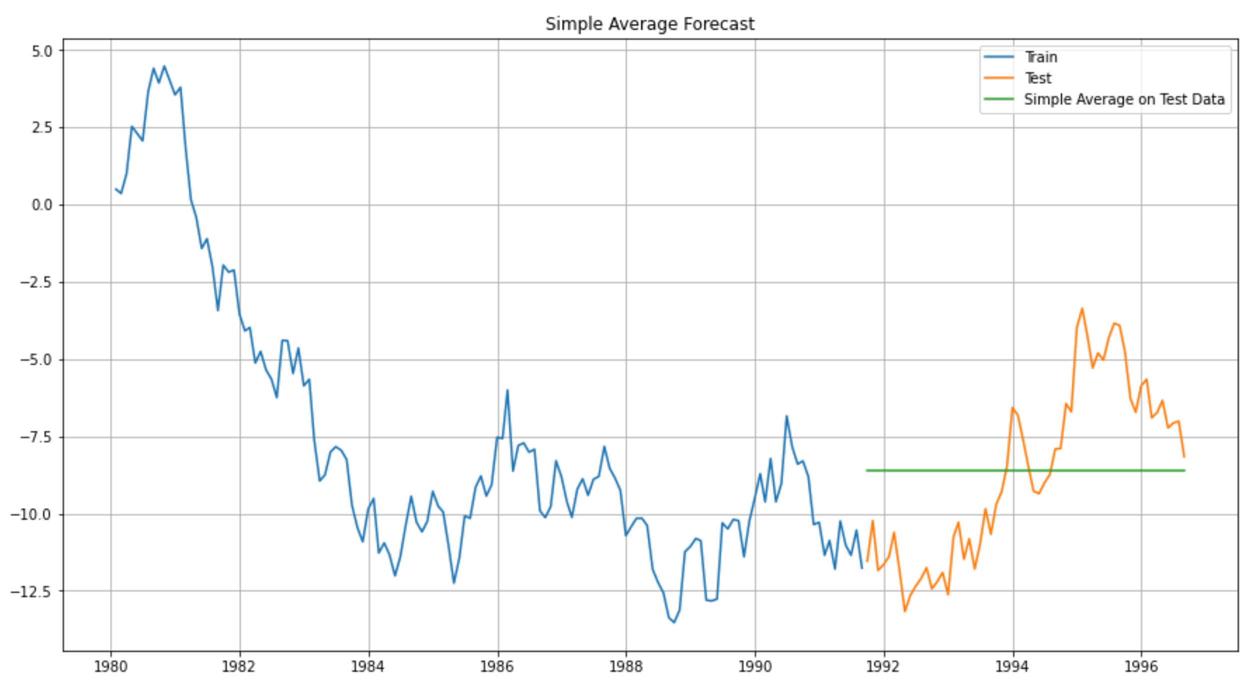

Figure 10: Simple average model

- Moving average

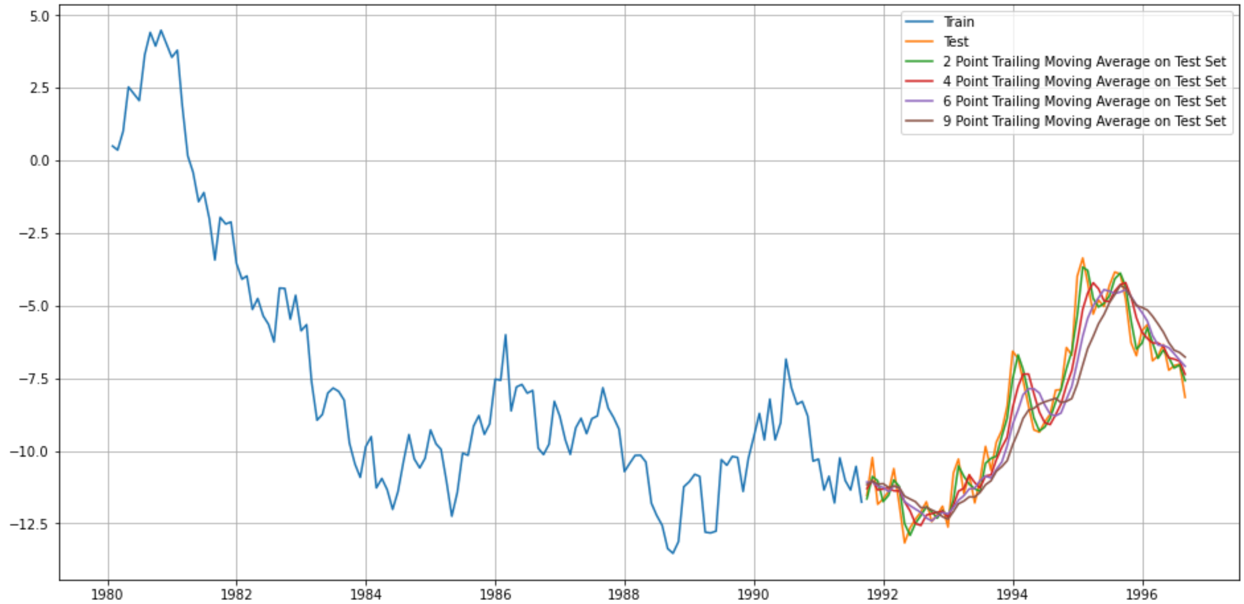

Figure 11: Moving average

- Single exponential smoothening

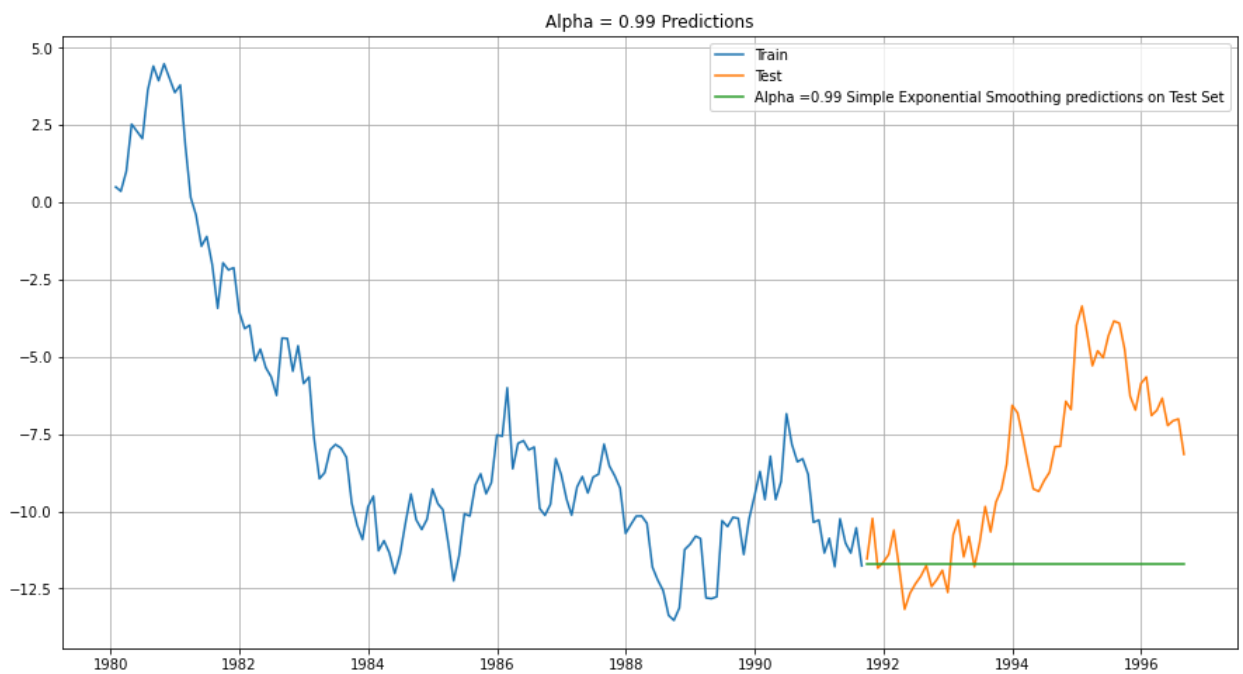

Figure 12: Single exponential smoothening

- Double exponential smoothing

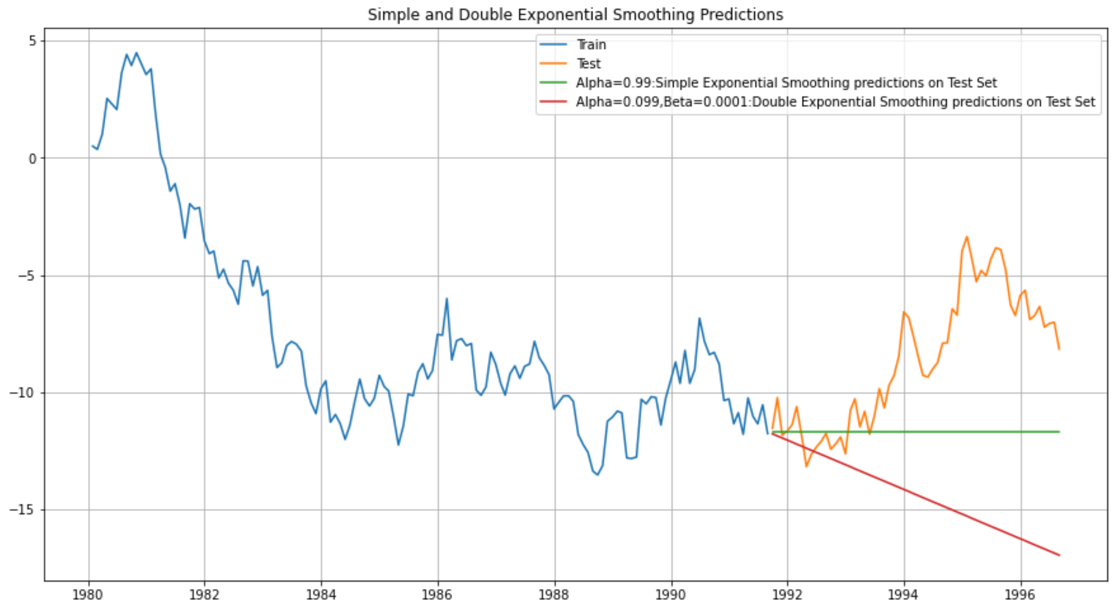

Figure 13: Double exponential smoothing

- Triple exponential smoothing

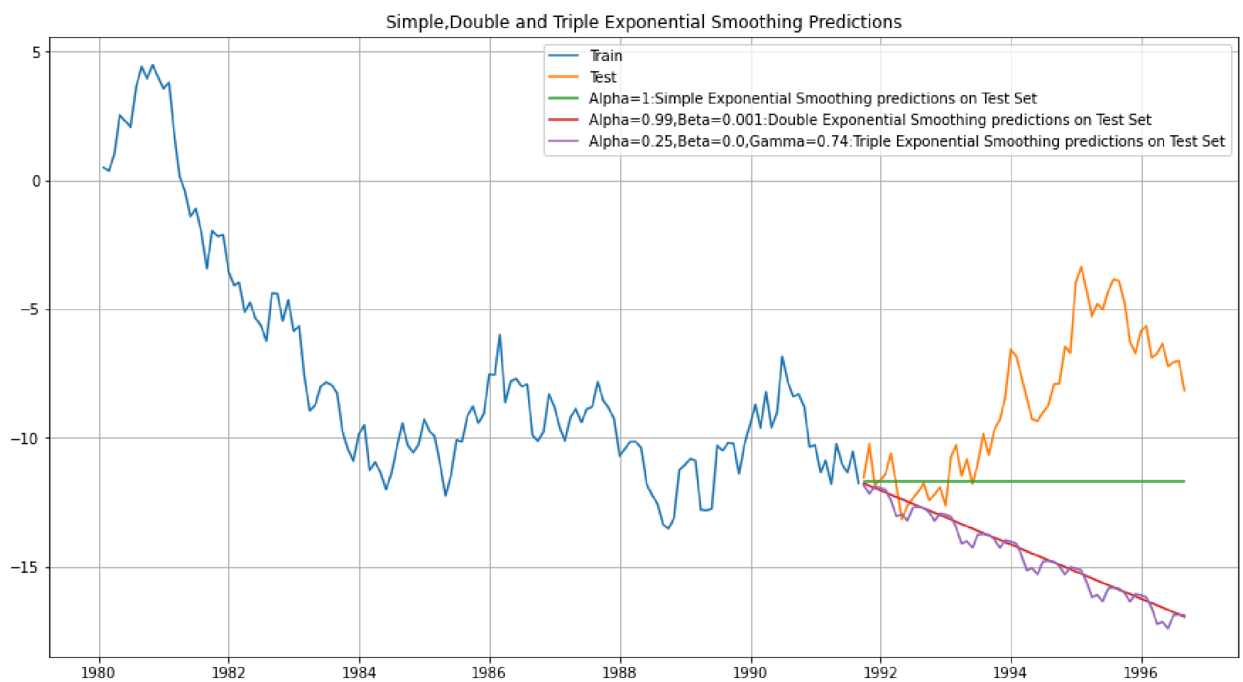

Figure 14: Triple exponential smoothing

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 = 0.1696

0.1696 > 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:

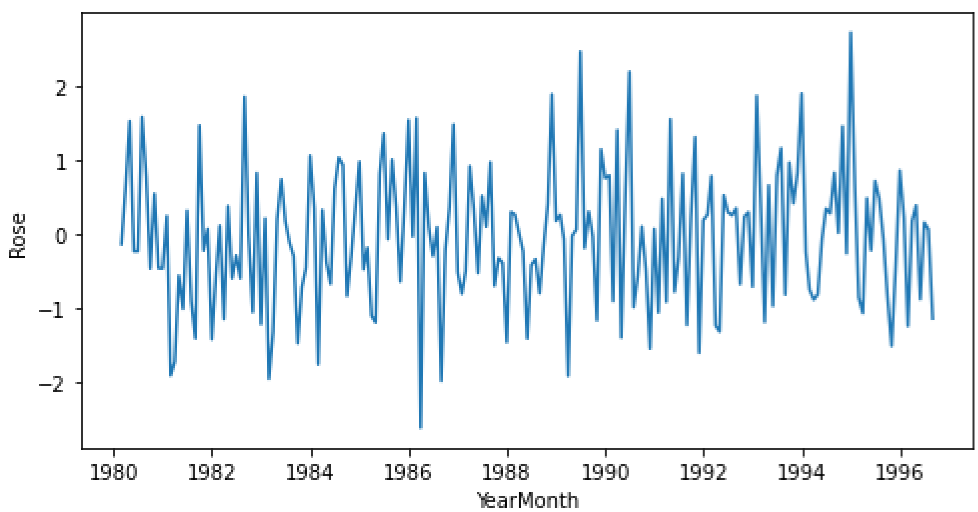

Figure 15: 1st differencing

- 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

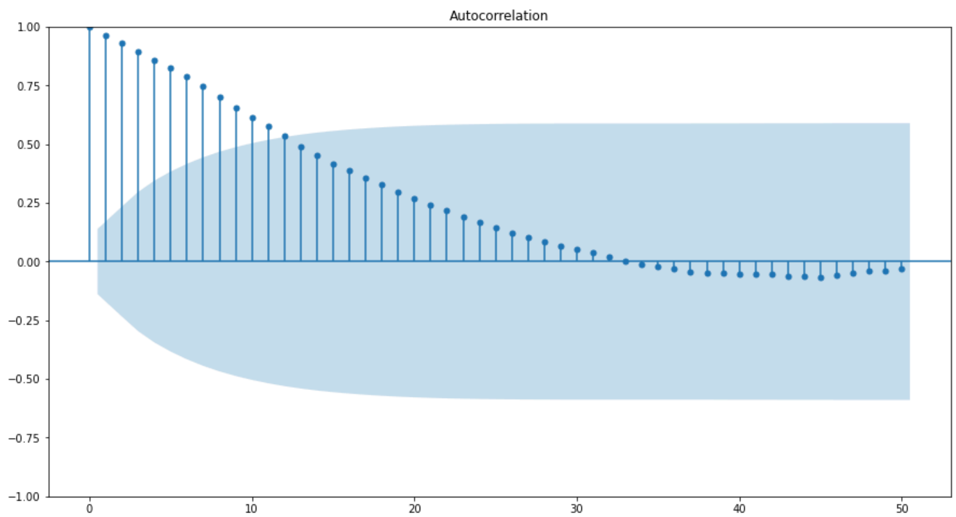

Figure 16: ACF plot

PACF Plot

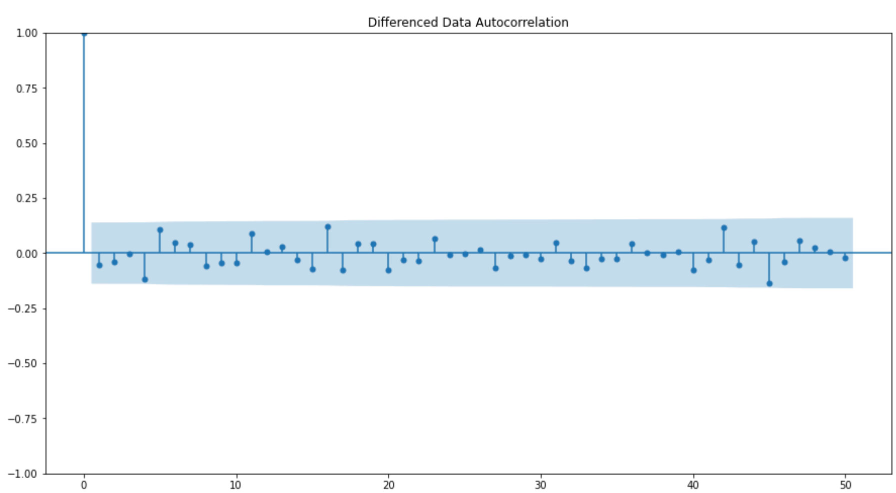

Figure 17: PACF plot

- 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 (1, 0, 0).

	param	AIC
3	(1, 0, 0)	390.153
4	(1, 0, 1)	391.80992
6	(2, 0, 0)	391.810503
8	(2, 0, 2)	392.029717
7	(2, 0, 1)	393.801803
5	(1, 0, 2)	393.809803
2	(0, 0, 2)	583.596142
1	(0, 0, 1)	659.953646
0	(0, 0, 0)	818.315131

Table 4: Auto ARIMA-AIC

Summary of ARIMA (1, 0, 0)

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  140
Model:                 ARIMA(1, 0, 0)   Log Likelihood                -192.077
Date:                Sun, 02 Jun 2024   AIC                            390.153
Time:                        17:04:38   BIC                            398.978
Sample:                    01-31-1980   HQIC                           393.739
                         - 08-31-1991
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -6.7347      2.937     -2.293      0.022     -12.491      -0.979
ar.L1          0.9825      0.017     58.723      0.000       0.950       1.015
sigma2         0.8887      0.111      7.982      0.000       0.671       1.107
===================================================================================
Ljung-Box (L1) (Q):                   0.62   Jarque-Bera (JB):                 0.32
Prob(Q):                              0.43   Prob(JB):                         0.85
Heteroskedasticity (H):               1.07   Skew:                             0.09
Prob(H) (two-sided):                  0.83   Kurtosis:                         2.86
===================================================================================
```

Figure 18: ARIMA(1, 0, 0)

RMSE = 5.8002

Manual ARIMA Model

From ACF and PACF, we get p = 0 and q = 2, with d = 1.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  140
Model:                 ARIMA(0, 1, 2)   Log Likelihood                  -0.223
Date:                Sun, 02 Jun 2024   AIC                              6.447
Time:                        17:04:38   BIC                             15.250
Sample:                    01-31-1980   HQIC                            10.024
                         - 08-31-1991
Covariance Type:                  opg
==============================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1          1.9592     67.507      0.029      0.977    -130.352     134.270
ma.L2          0.9970     68.990      0.014      0.988    -134.221     136.215
sigma2         0.0358      2.458      0.015      0.988      -4.782       4.854
==============================================================================
```

Figure 19: ARIMA (0, 1, 2)

RMSE = 9.058

SARIMA Models

Auto SARIMA

- (1, 0, 2) (0, 0, 2, 12) has the lowest SARIMA value.

	param	seasonal	AIC
47	(1, 0, 2)	(0, 0, 2, 12)	316.549662
50	(1, 0, 2)	(1, 0, 2, 12)	318.119201
53	(1, 0, 2)	(2, 0, 2, 12)	318.232699
65	(2, 0, 1)	(0, 0, 2, 12)	318.69132
38	(1, 0, 1)	(0, 0, 2, 12)	318.808599

Table 5: Auto SARIMA (1, 0, 2) (0, 0, 2, 12)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  140
Model:             SARIMAX(1, 0, 2)x(0, 0, 2, 12)   Log Likelihood                -152.275
Date:                            Sun, 02 Jun 2024   AIC                            316.550
Time:                                    17:05:59   BIC                            332.914
Sample:                                         0   HQIC                           323.190
                                            - 140
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.0026      0.006    159.800      0.000       0.990       1.015
ma.L1         -0.1448      0.105     -1.380      0.167      -0.350       0.061
ma.L2         -0.1102      0.119     -0.928      0.353      -0.343       0.123
ma.S.L12      -0.0695      0.095     -0.729      0.466      -0.256       0.117
ma.S.L24      -0.0567      0.106     -0.536      0.592      -0.264       0.151
sigma2         0.8655      0.121      7.145      0.000       0.628       1.103
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):                 0.40
Prob(Q):                              0.90   Prob(JB):                         0.82
Heteroskedasticity (H):               1.27   Skew:                             0.14
Prob(H) (two-sided):                  0.46   Kurtosis:                         3.07
===================================================================================
```

Figure 20: Auto SARIMA (1, 0, 2) (0, 0, 2, 12)

RMSE = 24.638

Manual SARIMA Model

From ACF and PACF, we get p = 0 and q = 2, with d = 1 and
P = 2, D = 1, Q = 2, S = 12

Summary of SARIMA (0, 1, 2) (2, 1, 2, 12)

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                               Rose   No. Observations:                  140
Model:             SARIMAX(0, 1, 2)x(2, 1, 2, 12)   Log Likelihood                  -2.986
Date:                            Sun, 02 Jun 2024   AIC                             19.971
Time:                                    17:06:00   BIC                             39.881
Sample:                                01-31-1980   HQIC                            28.060
                                     - 08-31-1991
Covariance Type:                              opg
==========================================================================================
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1          0.2131   7447.388   2.86e-05      1.000   -1.46e+04    1.46e+04
ma.L2          0.0343   1618.164   2.12e-05      1.000   -3171.509    3171.578
ar.S.L12   -9.698e-06     37.535  -2.58e-07      1.000     -73.568      73.568
ar.S.L24   -2.366e-05      0.132     -0.000      1.000      -0.258       0.258
ma.S.L12    4.046e-05     36.861    1.1e-06      1.000     -72.247      72.247
ma.S.L24    2.077e-05      0.440   4.72e-05      1.000      -0.862       0.862
sigma2         1.1334   3558.560      0.000      1.000   -6973.517    6975.784
==============================================================================
```

Figure 21: SARIMA (0, 1, 2) (2, 1, 2, 12)

RMSE = 9.055

Best ARIMA and SARIMA models

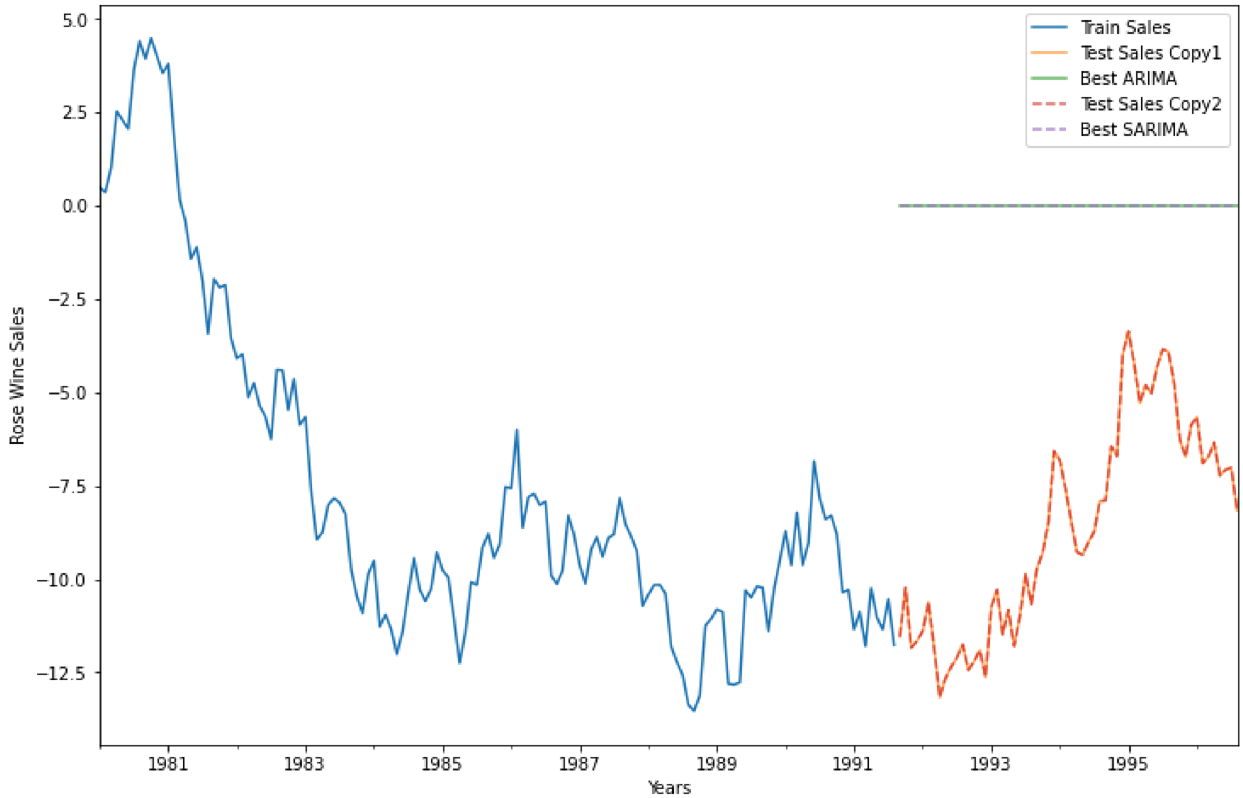

Figure 22: Best ARIMA and SARIMA

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	17.24
Simple Average Model :	7.87
Moving Average Model :	0.20
Single Exponential Smoothening Model :	17.37
Double Exponential Smoothening Model	7.09
Triple Exponential Smoothening Model Additive:	7.25
Triple Exponential Smoothening Model Multiplica...	7.25
Best AR Model :	10.46
Best ARMA Model:	10.00
Best ARIMA Model :	9.06
Best SARIMA Model :	9.05

Table 6: Model performance comparision

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

Forecasting

- The overall forecasted sales are trending downwards.

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.
- There is an overall decline in sales over time. The company needs to take immediate actions to prevent further decline and works towards improving sales.

Recommendations:

- 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.
- Identify the root cause for decline in sales over years and reverse the decline of sales.
- Improve product quality, cost, marketing strategies, etc. can help reverse the decline of sales.