13 February 2022

Time Series Project

PGP-DSBA June-Batch

SOOREJ KUNNOOL BALAN

2022

Table of Contents

[Problem 5](#_Toc95680556)

[1. Read the data as an appropriate Time Series data and plot the data. 5](#_Toc95680557)

[Sparkling Dataset: 5](#_Toc95680558)

[Rose Dataset: 6](#_Toc95680559)

[2. Perform appropriate Exploratory Data Analysis to understand the data and perform decomposition. 7](#_Toc95680560)

[Sparkling Dataset: 7](#_Toc95680561)

[Rose Dataset: 12](#_Toc95680562)

[3. Split the data into training and test. The test data should start in 1991. 16](#_Toc95680563)

[Sparkling Dataset: 16](#_Toc95680564)

[Rose Dataset: 17](#_Toc95680565)

[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. 18](#_Toc95680566)

[Simple Exponential Smoothing(SES): 18](#_Toc95680567)

[Double Exponential Smoothing(DES): 22](#_Toc95680568)

[Triple Exponential Smoothing(TES): 26](#_Toc95680569)

[Linear Regression (LR): 31](#_Toc95680570)

[Naïve Forecast: 34](#_Toc95680571)

[Simple Average Forecast: 36](#_Toc95680572)

[Moving Average Forecast: 39](#_Toc95680573)

[Inference: 42](#_Toc95680574)

[5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and 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. 43](#_Toc95680575)

[Sparkling Dataset: 43](#_Toc95680576)

[Rose Dataset: 44](#_Toc95680577)

[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. 46](#_Toc95680578)

[ARIMA model: 46](#_Toc95680579)

[SARIMA model: 53](#_Toc95680580)

[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. 62](#_Toc95680581)

[ARIMA model: 62](#_Toc95680582)

[SARIMA model: 68](#_Toc95680583)

[8. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data. 74](#_Toc95680584)

[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. 75](#_Toc95680585)

[10. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales. 77](#_Toc95680586)

[THE END 78](#_Toc95680587)

List of Tables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table-1.1 Sparkling Dataset Head and Tail | 5 |  | Table-1.47 Sparkling: LR Test Dataset | 31 |
| Table-1.2 Sparkling: Added Time Series | 5 |  | Table-1.48 Sparkling: LR Test/Predict | 32 |
| Table-1.3 Sparkling: Timeseries dataset | 5 |  | Table-1.49 Sparkling: Model RMSE | 32 |
| Table-1.4 Rose Dataset Head and Tail | 6 |  | Table-1.50 Rose: LR Train Dataset | 33 |
| Table-1.5 Rose: Added Time Series | 6 |  | Table-1.51 Rose: LR Test Dataset | 33 |
| Table-1.6 Rose: Timeseries dataset | 7 |  | Table-1.52 Rose: LR Test/Predict | 33 |
| Table-1.7 Sparkling: Concise data summary | 7 |  | Table-1.53 Rose: Model RMSE | 34 |
| Table-1.8 Sparkling: Data Summary of numeric columns | 8 |  | Table-1.54 Sparkling: Train Dataset-Tail | 34 |
| Table-1.9 Sparkling: Pivot table of monthly sales across years | 10 |  | Table-1.55 Sparkling: Test Dataset-Head | 34 |
| Table-1.10 Rose: Concise data summary | 12 |  | Table-1.56 Sparkling: Model RMSE | 35 |
| Table-1.11 Rose: Null data | 12 |  | Table-1.57 Rose: Train Dataset-Tail | 35 |
| Table-1.12 Rose: 1994 data before & after interpolation | 12 |  | Table-1.58 Rose: Test Dataset-Head | 35 |
| Table-1.13 Rose: Data Summary of numeric columns | 13 |  | Table-1.59 Rose: Model RMSE | 36 |
| Table-1.14 Rose: Pivot table of monthly sales across years | 15 |  | Table-1.60 Sparkling: Simple Average Test Dataset | 37 |
| Table-1.15 Sparkling: Train dataset head/tail | 17 |  | Table-1.61 Sparkling: Model RMSE | 37 |
| Table-1.16 Sparkling: Test dataset head/tail | 17 |  | Table-1.62 Rose: Simple Average Test Dataset | 38 |
| Table-1.17 Rose: Train dataset head/tail | 18 |  | Table-1.63 Rose: Model RMSE | 38 |
| Table-1.18 Rose: Test dataset head/tail | 18 |  | Table-1.64 Sparkling: Moving Average dataset | 39 |
| Table-1.19 Sparkling: SES – auto parameters | 19 |  | Table-1.65 Sparkling: Moving Average - RMSE | 39 |
| Table-1.20 Sparkling: Test/Predict dataset head | 19 |  | Table-1.66 Sparkling: Model RMSE | 40 |
| Table-1.21 Sparkling: SES Models RMSE | 20 |  | Table-1.67 Rose: Moving Average dataset | 40 |
| Table-1.22 Sparkling: Model RMSE | 20 |  | Table-1.68 Rose: Moving Average - RMSE | 41 |
| Table-1.23 Rose: SES – auto parameters | 20 |  | Table-1.69 Rose: Model RMSE | 42 |
| Table-1.24 Rose: Test/Predict dataset head | 21 |  | Table-1.70 Sparkling: Model RMSE | 42 |
| Table-1.25 Rose: SES Models RMSE | 21 |  | Table-1.71 Rose: Model RMSE | 42 |
| Table-1.26 Rose: Model RMSE | 22 |  | Table-1.72 Sparkling: ADF on Train dataset | 44 |
| Table-1.27 Sparkling: DES – auto parameters | 22 |  | Table-1.73 Sparkling: ADF on Train dataset with first order of differencing | 44 |
| Table-1.28 Sparkling: Test/Predict dataset head | 22 |  | Table-1.74 Rose: ADF on Train dataset | 45 |
| Table-1.29 Sparkling: DES Models RMSE | 23 |  | Table-1.75 Rose: ADF on Train dataset with first order of differencing | 46 |
| Table-1.30 Sparkling: Test/Predict dataset head | 23 |  | Table-1.76 Sparkling: Auto ARIMA model parameters | 47 |
| Table-1.31 Sparkling: Model RMSE | 24 |  | Table-1.77 Sparkling: Auto ARIMA AIC values | 47 |
| Table-1.32 Rose: DES – auto parameters | 24 |  | Table-1.78 Sparkling: Auto ARIMA Summary | 48 |
| Table-1.33 Rose: Test/Predict dataset head | 25 |  | Table-1.79 Sparkling: Auto ARIMA Test/Prediction Head | 49 |
| Table-1.34 Rose: DES Models RMSE | 25 |  | Table-1.80 Sparkling: Model - RMSE | 49 |
| Table-1.35 Rose: Model RMSE | 26 |  | Table-1.81 Rose: Auto ARIMA model parameters | 51 |
| Table-1.36 Sparkling: TES – auto parameters | 26 |  | Table-1.82 Rose: Auto ARIMA AIC values | 51 |
| Table-1.37 Sparkling: Test/Predict dataset head | 26 |  | Table-1.83 Rose: Auto ARIMA Summary | 52 |
| Table-1.38 Sparkling: TES Models RMSE | 27 |  | Table-1.84 Rose: Auto ARIMA Test/Prediction Head | 53 |
| Table-1.39 Sparkling: Test/Predict dataset head | 28 |  | Table-1.85 Rose: Model - RMSE | 53 |
| Table-1.40 Sparkling: Model RMSE | 28 |  | Table-1.86 Sparkling: Auto SARIMA model parameters | 55 |
| Table-1.41 Rose: TES – auto parameters | 29 |  | Table-1.87 Sparkling: Auto SARIMA AIC values | 55 |
| Table-1.42 Rose: Test/Predict dataset head | 29 |  | Table-1.88 Sparkling: Auto SARIMA model summary | 56 |
| Table-1.43 Rose: TES Models RMSE | 30 |  | Table-1.89 Sparkling: Auto SARIMA model prediction head | 57 |
| Table-1.44 Rose: Test/Predict dataset head | 30 |  | Table-1.90 Sparkling: Model - RMSE | 57 |
| Table-1.45 Rose: Model RMSE | 31 |  | Table-1.91 Rose: Auto SARIMA model parameters | 59 |
| Table-1.46 Sparkling: LR Train Dataset | 31 |  | Table-1.92 Rose: Auto SARIMA AIC values | 59 |
| Table-1.93 Rose: Auto SARIMA model summary | 60 |  | Table-1.103 Sparkling: SARIMA model prediction head | 70 |
| Table-1.94 Rose: Auto SARIMA model prediction head | 61 |  | Table-1.104 Sparkling: Models - RMSE | 70 |
| Table-1.95 Rose: Model - RMSE | 61 |  | Table-1.105 Rose: SARIMA Model summary | 72 |
| Table-1.96 Sparkling: ARIMA Model summary | 63 |  | Table-1.106 Rose: SARIMA model prediction head | 73 |
| Table-1.97 Sparkling: ARIMA model prediction head | 64 |  | Table-1.107 Rose: Models - RMSE | 73 |
| Table-1.98 Sparkling: Models - RMSE | 64 |  | Table-1.108 Sparkling: Models – RMSE | 74 |
| Table-1.99 Rose: ARIMA Model summary | 66 |  | Table-1.109 Rose: Models – RMSE | 74 |
| Table-1.100 Rose: ARIMA model prediction head | 67 |  | Table-1.110 Sparkling: 12 months forecast | 75 |
| Table-1.101 Rose: Models - RMSE | 67 |  | Table-1.111 Rose: 12 months forecast | 76 |
| Table-1.102 Sparkling: SARIMA Model summary | 69 |  |  |  |

List of Figures

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Figure-1.1 Sparkling: Sales vs Years | 6 |  | Figure-1.36 Rose: Original vs Trailing Moving Averages | | 41 |
| Figure-1.2 Rose: Sales vs Years | 7 |  | Figure-1.37 Rose: Train/Test/Predict- 2 point Moving Average | | 41 |
| Figure-1.3 Sparkling: Boxplot | 8 |  | Figure-1.38 Sparkling: Train dataset | | 43 |
| Figure-1.4 Boxplot: Sparkling sales across years | 9 |  | Figure-1.39 Sparkling: Train dataset with first order of differencing | | 44 |
| Figure-1.5 Boxplot: Sparkling sales across months | 9 |  | Figure-1.40 Rose: Train dataset | | 45 |
| Figure-1.6 Sparkling monthly sales across months | 10 |  | Figure-1.41 Rose: Train dataset with first order of differencing | | 45 |
| Figure-1.7 Sparkling: Average & % change in sales | 11 |  | Figure-1.42 Sparkling: ACF plot | | 46 |
| Figure-1.8 Sparkling: Additive decompose | 11 |  | Figure-1.43 Sparkling: PACF plot | | 47 |
| Figure-1.9 Sparkling: Multiplicative decompose | 11 |  | Figure-1.44 Sparkling: Auto ARIMA diagnostic plot | | 48 |
| Figure-1.10 Rose: Boxplot | 13 |  | Figure-1.45 Sparkling: Train, Test, Auto ARIMA prediction | | 49 |
| Figure-1.11 Boxplot: Rose sales across years | 14 |  | Figure-1.46 Rose: ACF plot | | 50 |
| Figure-1.12 Boxplot: Rose sales across months | 15 |  | Figure-1.47 Rose: PACF plot | | 50 |
| Figure-1.13 Rose monthly sales across months | 15 |  | Figure-1.48 Rose: Auto ARIMA diagnostic plot | | 52 |
| Figure-1.14 Rose: Average & % change in sales | 16 |  | Figure-1.49 Rose: Train, Test, Auto ARIMA prediction | | 53 |
| Figure-1.15 Rose: Additive decompose | 16 |  | Figure-1.50 Sparkling: ACF – autocorrelation plot | | 54 |
| Figure-1.16 Rose: Multiplicative decompose | 16 |  | Figure-1.51 Sparkling: PACF – partial autocorrelation plot | | 54 |
| Figure-1.17 Sparkling: Train/Test Dataset | 17 |  | Figure-1.52 Sparkling: Auto SARIMA model diagnostics | | 56 |
| Figure-1.18 Rose: Train/Test Dataset | 18 |  | Figure-1.53 Sparkling: Train, Test Auto SARIMA model prediction | | 57 |
| Figure-1.19 Sparkling: SES Train/Test/Predict plot | 19 |  | Figure-1.54 Rose: ACF – autocorrelation plot | | 58 |
| Figure-1.20 Rose: SES Train/Test/Predict plot | 21 |  | Figure-1.55 Rose: PACF – partial autocorrelation plot | | 58 |
| Figure-1.21 Sparkling: DES Train/Test/Predict plot | 23 |  | Figure-1.56 Rose: Auto SARIMA model diagnostics | | 60 |
| Figure-1.22 Sparkling: DES Train/Test/Predict plot | 24 |  | Figure-1.57 Rose: Auto SARIMA model diagnostics | | 61 |
| Figure-1.23 Rose: DES Train/Test/Predict plot | 25 |  | Figure-1.58 Sparkling: ACF – autocorrelation plot | | 62 |
| Figure-1.24 Sparkling: TES Train/Test/Predict plot | 27 |  | Figure-1.59 Sparkling: PACF – partial autocorrelation plot | | 62 |
| Figure-1.25 Sparkling: TES Train/Test/Predict plot | 28 |  | Figure-1.60 Sparkling: ARIMA plot diagnostics | | 63 |
| Figure-1.26 Rose: TES Train/Test/Predict plot | 29 |  | Figure-1.61 Sparkling: Train, Test and ARIMA prediction | 64 | |
| Figure-1.27 Rose: TES Train/Test/Predict plot | 31 |  | Figure-1.62 Rose: ACF – autocorrelation plot | 65 | |
| Figure-1.28 Sparkling: LR Train/Test/Predict plot | 32 |  | Figure-1.63 Rose: PACF – partial autocorrelation plot | 65 | |
| Figure-1.29 Rose: LR Train/Test/Predict plot | 34 |  | Figure-1.64 Rose: ARIMA plot diagnostics | 66 | |
| Figure-1.30 Sparkling: Naive Train/Test/Predict plot | 35 |  | Figure-1.65 Rose: Train, Test and ARIMA prediction | 67 | |
| Figure-1.31 Rose: Naive Train/Test/Predict plot | 36 |  | Figure-1.66 Sparkling: ACF – autocorrelation plot | 68 | |
| Figure-1.32 Sparkling: Simple Average model Train/Test/Predict plot | 37 |  | Figure-1.67 Sparkling: PACF – partial autocorrelation plot | 68 | |
| Figure-1.33 Rose: Simple Average model Train/Test/Predict plot | 38 |  | Figure-1.68 Sparkling: SARIMA plot diagnostics | 69 | |
| Figure-1.34 Sparkling: Original vs Trailing Moving Averages | 39 |  | Figure-1.69 Sparkling: Train, Test and SARIMA prediction | 70 | |
| Figure-1.35 Sparkling: Train/Test/Predict- 2 point Moving Average | 40 |  | Figure-1.70 Rose: ACF – autocorrelation plot | 71 | |
| Figure-1.71 Rose: PACF – partial autocorrelation plot | 71 |  |  |  | |
| Figure-1.72 Rose: SARIMA plot diagnostics | 72 |  |  |  | |
| Figure-1.73 Rose: Train, Test and SARIMA prediction | 73 |  |  |  | |
| Figure-1.74 Sparkling: Sales & 12 months Forecast | 75 |  |  |  | |
| Figure-1.75 Rose: Sales & 12 months Forecast | 76 |  |  |  | |
|  |  |  |  |  | |

# Problem

For this assignment, the data of different types of wine sales in the 20th century is to be analyzed. Both data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

 

Please do perform the following questions on each of these two data sets separately.

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

### Sparkling Dataset:

Let’s look at the head and tail of the dataset.

Table

Description automatically generated with medium confidence Table

Description automatically generated

Table-1.1 Sparkling Dataset Head and Tail

We can see that we have sales data of Sparkling wine from January 1980 to July 1995. First we will create a monthly date series and add to the dataframe as shown below:

A screenshot of a computer

Description automatically generated with medium confidence

Table-1.2 Sparkling: Added Time Series

Now we can drop the YearMonth column and make the Time\_Series column as the dataframe index.

Graphical user interface, table

Description automatically generated with medium confidence

Table-1.3 Sparkling: Timeseries dataset

Let’s plot the above timeseries data and check the graph:

Chart

Description automatically generated

Figure-1.1 Sparkling: Sales vs Years

From the above graph we do not see much of a trend, sales show a slight increase from 1980 to 1988 and then on shows a slight downward trend.

The data does show a seasonal trend in the sale across years.

### Rose Dataset:

Let’s look at the head and tail of the dataset.

Graphical user interface

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with medium confidence

Table-1.4 Rose Dataset Head and Tail

We can see that we have sales data of Rose wine from January 1980 to July 1995. First we will create a monthly date series and add to the dataframe as shown below:

A screenshot of a computer

Description automatically generated with medium confidence

Table-1.5 Rose: Added Time Series

Now we can drop the YearMonth column and make the Time\_Series column as the dataframe index.

Table

Description automatically generated with medium confidence

Table-1.6 Rose: Timeseries dataset

Let’s plot the above timeseries data and check the graph:

Graphical user interface, chart

Description automatically generated

Figure-1.2 Rose: Sales vs Years

From the above graph we can see a clear downward trend in the sale of Rose wine from 1980 to 1995.

The data does show a seasonal trend in the sale across years.

We seem to have missing data in 1994, which we will impute later.

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

### Sparkling Dataset:

#### Concise Data Summary:

Let’s look at the concise data summary of the dataset.

Text

Description automatically generated

Table-1.7 Sparkling: Concise data summary

#### Let us check the type of variables in the data frame

There are a total of 187 observations and 1 column in the dataset. We have 1 column of int64 type.

#### Check for missing values in the dataset

From Table-1.7 we can see that all the columns have 187 non-null values. Hence there are no missing/null values in the dataset.

#### Data summary



Table-1.8 Sparkling: Data Summary of numeric columns

#### Univariate Data Analysis

Let’s check the central measures of tendency, quartiles, histogram, and boxplot of the numerical column.

Sparkling, which is sales per month, is a continuous variable with the below stats (refer Table 1.8):

Mean = 2402.42

Standard Deviation = 1295.11

Min value in dataset = 1070

Max value in dataset = 7242

Range = Min – Max = 6172

Q1(1st Quartile) = 1605

Q2(2nd Quartile)/Median = 1874

Q3(3rd Quartile) = 2549

IQR(Inter-Quartile Range) = Q3- Q1 = 944

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 1070

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 3965

##### *Sparkling: Boxplot*

Let’s look at the boxplot of Sparkling sales:

Chart, box and whisker chart

Description automatically generated

Figure-1.3 Sparkling: Boxplot

We can see some outliers in the boxplot and the boxplot shows the data has a positive skew.

##### *Sparkling: Yearly Sales*

Let’s look at the boxplot of yearly sales.

Chart, box and whisker chart

Description automatically generated

Figure-1.4 Boxplot: Sparkling sales across years

From the above boxplot we can see that outliers are present in almost every year. The average sales across the years have remained between 1750 to 2250.

There is no apparent trend towards one direction, the average sales have increased and decreased slightly and periodically.

##### *Sparkling: Monthly Sales*

Let’s look at the boxplot of sales across the months.

Chart, histogram

Description automatically generated

Figure-1.5 Boxplot: Sparkling sales across months

The above boxplot indicates seasonality in the sales. The first 6 months of the year shows average sales around 1750, but then on the sale of Sparkling wine increases almost exponentially from July to December.

Hence the business will have to look at better ways of marketing and campaigning between July to December to have maximum impact on sales.

##### *Sparkling: Monthly Sales across Years*

Let’s look at the pivot table of monthly sales data across years.

Table

Description automatically generated

Table-1.9 Sparkling: Pivot table of monthly sales across years

Let’s plot the above data and see the graph:

Chart, line chart

Description automatically generated

Figure-1.6 Sparkling monthly sales across months

The monthly sales for December are highest among other months across the years, followed by November.

The sales from January to June is considerably lesser than the sales from July to December across the years.

##### *Sparkling: Average and % change in sales*

Let’s look at the average and % change in sales plots.

Graphical user interface

Description automatically generated

Figure-1.7 Sparkling: Average & % change in sales

The average sales plot does not show a clear trend. The sales were highest in 1998 and lowest in 1983.

The % change in sales also corroborates the fact that the sales have remained constant across the years with slight fluctuations year on year.

#### Decomposition

Let’s perform both additive and multiplicative seasonal decompose and check the graphs for trend, seasonal and residue.

Graphical user interface, histogram

Description automatically generated Graphical user interface, histogram

Description automatically generated

Figure-1.8 Sparkling: Additive decompose Figure-1.9 Sparkling: Multiplicative decompose

In Figure-1.8, we can see the residuals in additive decompose displays a hint of trend in the data, whereas in figure 1.9 we can see that residuals are grouped near 1.0

Hence multiplicative decomposition is more apt in this scenario. We can see a slight positive trend. There is seasonality in the data.

### Rose Dataset:

#### Concise Data Summary:

Let’s look at the concise data summary of the dataset.

Text

Description automatically generated

Table-1.10 Rose: Concise data summary

#### Let us check the type of variables in the data frame

There are a total of 187 observations and 1 column in the dataset. We have 1 column of int64 type.

#### Check for missing values in the dataset

From Table-1.10 we can see that we have 187 observations but only 185 entries. Hence there are 2 missing/null values in the dataset.

Let’s look at the null data entries:

Graphical user interface, text, application

Description automatically generated

Table-1.11 Rose: Null data

Both the nulls are present in 1994. We will do linear interpolation to fill the null values. Let’s look at the 1994-year data before and after the impute.

A picture containing table

Description automatically generated Table

Description automatically generated

Table-1.12 Rose: 1994 data before & after interpolation

From Table 1.12 we can see that both the null values have been imputed via linear interpolation.

#### Data summary



Table-1.13 Rose: Data Summary of numeric columns

#### Univariate Data Analysis

Let’s check the central measures of tendency, quartiles, histogram, and boxplot of the numerical column.

Rose, which is sales per month, is a continuous variable with the below stats (refer Table 1.13):

Mean =89.91

Standard Deviation = 39.24

Min value in dataset = 28

Max value in dataset = 267

Range = Min – Max = 239

Q1(1st Quartile) = 62.5

Q2(2nd Quartile)/Median = 85

Q3(3rd Quartile) = 111

IQR(Inter-Quartile Range) = Q3- Q1 = 48.5

Quartile Min value = max(Qmin,Q1 – 1.5 \* IQR) = 28

Quartile Max value = min(Qmax,Q3 + 1.5 \* IQR) = 183.75

##### *Rose: Boxplot*

Let’s look at the boxplot of Rose sales:

Chart, box and whisker chart

Description automatically generated

Figure-1.10 Rose: Boxplot

We can see some outliers in the boxplot and the boxplot shows the data has a slight positive skew.

##### *Rose: Yearly Sales*

Let’s look at the boxplot of yearly sales.

Chart, box and whisker chart

Description automatically generated

Figure-1.11 Boxplot: Rose sales across years

From the above boxplot we can see that outliers are present in almost every year. The average sales across the years clearly shows a downward trend with highest average sale in 1980 and least in 1995.

##### *Rose: Monthly Sales*

Let’s look at the boxplot of sales across the months.

Chart, box and whisker chart

Description automatically generated

Figure-1.12 Boxplot: Rose sales across months

The above boxplot indicates seasonality in the sales. The sales marginally increase from January to August and from September to December displays a clear uptick. The boxplot shows outliers in some months.

##### *Rose: Monthly Sales across Years*

Let’s look at the pivot table of monthly sales data across years.

Table

Description automatically generated with low confidence

Table-1.14 Rose: Pivot table of monthly sales across years

Let’s plot the above data and see the graph:

Chart, line chart

Description automatically generated

Figure-1.13 Rose monthly sales across months

The monthly sales for December are highest among other months across the years.

We can see a clear downward trend in the sales of each month.

##### *Rose: Average and % change in sales*

Let’s look at the average and % change in sales plots.

Graphical user interface

Description automatically generated

Figure-1.14 Rose: Average & % change in sales

The average sales plot shows a clear downward trend. The sales were highest in 1998 and lowest in 1983.

The % change in sales shows more negative ticks than positive indicating the presence of decreased sales across years.

#### Decomposition

Let’s perform both additive and multiplicative seasonal decompose and check the graphs for trend, seasonal and residue.

Graphical user interface, application

Description automatically generated Graphical user interface

Description automatically generated

Figure-1.15 Rose: Additive decompose Figure-1.16 Rose: Multiplicative decompose

In Figure-1.15, we can see the residuals in additive decompose displays a hint of trend in the data, whereas in figure 1.16 we can see that residuals are grouped near 1.0

Hence multiplicative decomposition is more apt in this scenario. We can see trend is negative. There is seasonality in the data.

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

### Sparkling Dataset:

Let’s split the Sparkling dataset into train and test. Let’s check the shape of the train/test datasets.

Shape of train dataset for Sparkling is (132, 1) i.e., 132 of 187 (70.6%) observations are in train set.

Shape of test dataset for Sparkling is (55, 1) i.e., 55 of 187 (29.4%) observations are in test.

Let’s check the head and tail of the train/test datasets.

Text, table

Description automatically generated Text, table

Description automatically generated

Table-1.15 Sparkling: Train dataset head/tail Table-1.16 Sparkling: Test dataset head/tail

From table 1.15 we can see that train dataset has sales value from Jan 1980 to Dec 1990.

The test dataset, based on Table 1.16 shows the data from Jan 1991 to Jul 1995.

Let’s plot the graph of train/test datasets:

Chart, histogram

Description automatically generated

Figure-1.17 Sparkling: Train/Test Dataset

### Rose Dataset:

Let’s split the Rose dataset into train and test. Let’s check the shape of the train/test datasets.

Shape of train dataset for Rose is (132, 1) i.e., 132 of 187 (70.6%) observations are in train set.

Shape of test dataset for Rose is (55, 1) i.e., 55 of 187 (29.4%) observations are in test.

Let’s check the head and tail of the train/test datasets.

Text

Description automatically generated Text

Description automatically generated

Table-1.17 Rose: Train dataset head/tail Table-1.18 Rose: Test dataset head/tail

From table 1.17 we can see that train dataset has sales value from Jan 1980 to Dec 1990.

The test dataset, based on Table 1.18 shows the data from Jan 1991 to Jul 1995.

Let’s plot the graph of train/test datasets:

Chart

Description automatically generated

Figure-1.18 Rose: Train/Test Dataset

## 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.

We will build the models on both the Sparkling and Rose training datasets and check the performance.

### Simple Exponential Smoothing(SES):

#### Sparkling Dataset

Let’s perform Simple Exponential Smoothing model on the Sparkling train dataset.

First we will run the simple exponential smoothing model in optimized mode and let the model determine the smoothing factor for the level. Let us check the parameters thus determined:

Text

Description automatically generated

Table-1.19 Sparkling: SES – auto parameters

From table 1.19 we can see that SES in auto mode has determined the smoothing factor level as 0.0496.

Let’s forecast for the test dataset based on the above model and check the predictions (first 5).

Table

Description automatically generated with medium confidence

Table-1.20 Sparkling: Test/Predict dataset head

Let’s calculate the RMSE of the SES in auto mode model:

RMSE value on test dataset for SES autofit model is 1316.0354872762928.

Let’s plot the train/test and forecast data.

Chart

Description automatically generated

Figure-1.19 Sparkling: SES Train/Test/Predict plot

From Figure 1.19 we can see that the forecast for SES is a constant with value 2724.932624

Next we will create multiple SES models by iterating smoothing\_level values from 0.1,0.2 so on till 0.9 and check the RMSE values of all the models thus created.

Text

Description automatically generated

Table-1.21 Sparkling: SES Models RMSE

From table 1.21 we can see that the RMSE value is least for the alpha 0.0496 as determined by SES auto and hence is the optimal model for Simple exponential smoothing on the Sparkling train dataset. Let’s store the model parameter, model and RMSE into a results dataset for comparison between models.

Shape, rectangle

Description automatically generated

Table-1.22 Sparkling: Model RMSE

#### Rose Dataset

Let’s perform Simple Exponential Smoothing model on the Rose train dataset.

First we will run the simple exponential smoothing model in optimized mode and let the model determine the smoothing factor for the level. Let us check the parameters thus determined:

Text

Description automatically generated

Table-1.23 Rose: SES – auto parameters

From table 1.23 we can see that SES in auto mode has determined the smoothing factor level as 0.0987.

Let’s forecast for the test dataset based on the above model and check the predictions (first 5).

Graphical user interface, text

Description automatically generated

Table-1.24 Rose: Test/Predict dataset head

Let’s calculate the RMSE of the SES in auto mode model:

RMSE value on test dataset for SES autofit model is 36.79623342215522.

Let’s plot the train/test and forecast data.

Graphical user interface, chart

Description automatically generated

Figure-1.20 Rose: SES Train/Test/Predict plot

From Figure 1.19 we can see that the forecast for SES is a constant with value 87.104983.

Next we will create multiple SES models by iterating smoothing\_level values from 0.1,0.2 so on till 0.9 and check the RMSE values of all the models thus created.

A picture containing application

Description automatically generated

Table-1.25 Rose: SES Models RMSE

From table 1.25 we can see that the RMSE value is least for the alpha 0.0987 as determined by SES auto and hence is the optimal model for Simple exponential smoothing on the Rose train dataset. Let’s store the model parameter, model and RMSE into a results dataset for comparison between models.



Table-1.26 Rose: Model RMSE

### Double Exponential Smoothing(DES):

#### Sparkling Dataset

Let’s perform Double Exponential Smoothing model on the Sparkling train dataset.

First we will run the double exponential smoothing model in optimized mode and let the model determine the smoothing factor for the level and trend. Let us check the parameters thus determined:

Text, letter

Description automatically generated

Table-1.27 Sparkling: DES – auto parameters

From table 1.27 we can see that DES in auto mode has determined the smoothing factor level as 0.6886 and smoothing trend as 0.00001

Let’s forecast for the test dataset based on the above model and check the predictions (first 5).

Text

Description automatically generated with medium confidence

Table-1.28 Sparkling: Test/Predict dataset head

Let’s calculate the RMSE of the DES in auto mode model:

RMSE value on test dataset for DES autofit model is 2007.238525758568.

Let’s plot the train/test and forecast data.

Chart, line chart

Description automatically generated

Figure-1.21 Sparkling: DES Train/Test/Predict plot

From Figure 1.21 we can see that the forecast for DES is a straight line with a downward slope.

Next we will create multiple DES models by iterating both smoothing\_level and smoothing\_trend values from 0.1,0.2 so on till 0.9 and check the RMSE values of all the models thus created.

Graphical user interface, text, application

Description automatically generated

Table-1.29 Sparkling: DES Models RMSE

From table 1.29 we can see that the RMSE value is least for the alpha 0.1 and beta 0.1

Let’s create this model and check the predictions (first 5).

Table

Description automatically generated with medium confidence

Table-1.30 Sparkling: Test/Predict dataset head

Let’s plot the graph of train, test and forecast for DES with alpha 0.1 and beta 0.1

Chart, histogram

Description automatically generated

Figure-1.22 Sparkling: DES Train/Test/Predict plot

Let’s store the model, model parameter and RMSE into the results dataset.

Graphical user interface, text

Description automatically generated

Table-1.31 Sparkling: Model RMSE

#### Rose Dataset

Let’s perform Double Exponential Smoothing model on the Rose train dataset.

First we will run the double exponential smoothing model in optimized mode and let the model determine the smoothing factor for the level and trend. Let us check the parameters thus determined:

Text, letter

Description automatically generated

Table-1.32 Rose: DES – auto parameters

From table 1.32 we can see that DES in auto mode has determined the smoothing factor level as 0.0175 and smoothing trend as 0.00003

Let’s forecast for the test dataset based on the above model and check the predictions (first 5).

Text, table

Description automatically generated

Table-1.33 Rose: Test/Predict dataset head

Let’s calculate the RMSE of the DES in auto mode model:

RMSE value on test dataset for DES autofit model is 15.707084805421891.

Let’s plot the train/test and forecast data.

Chart

Description automatically generated

Figure-1.23 Rose: DES Train/Test/Predict plot

From Figure 1.23 we can see that the forecast for DES is a straight line with a downward slope.

Next we will create multiple DES models by iterating both smoothing\_level and smoothing\_trend values from 0.1,0.2 so on till 0.9 and check the RMSE values of all the models thus created.

Graphical user interface, text, application

Description automatically generated

Table-1.34 Rose: DES Models RMSE

From table 1.34 we can see that the RMSE value is least for the DES auto model we have seen earlier.

Let’s store the model, model parameter and RMSE into the results dataset.

Graphical user interface, text

Description automatically generated

Table-1.35 Rose: Model RMSE

### Triple Exponential Smoothing(TES):

#### Sparkling Dataset

Let’s perform Triple Exponential Smoothing model on the Sparkling train dataset.

First we will run the triple exponential smoothing model in optimized mode and let the model determine the smoothing factor for the level, trend and seasonal. Let us check the parameters thus determined:

Text, letter

Description automatically generated

Table-1.36 Sparkling: TES – auto parameters

From table 1.36 we can see that TES in auto mode has determined the smoothing factor level as 0.1113, smoothing trend as 0.0495 and seasonal trend as 0.3620

Let’s forecast for the test dataset based on the above model and check the predictions (first 5).

Table

Description automatically generated

Table-1.37 Sparkling: Test/Predict dataset head

Let’s calculate the RMSE of the TES in auto mode model:

RMSE value on test dataset for TES autofit model is 404.286809456071.

Let’s plot the train/test and forecast data.

Chart, histogram

Description automatically generated

Figure-1.24 Sparkling: TES Train/Test/Predict plot

From Figure 1.24 we can see that the forecast for TES is more aligned to the test data than SES/DES.

Next we will create multiple TES models by iterating level, trend, and seasonal values from 0.1,0.2 so on till 0.9 and check the RMSE values of all the models thus created.

Text

Description automatically generated

Table-1.38 Sparkling: TES Models RMSE

We can see from Table 1.38 that for alpha = 0.4, beta = 0.1 and gamma = 0.2 (level, trend, seasonal smoothing constants respectively) the model has a lower RMSE (317.4343) than the RMSE value of TES autofit model (404.2868).

Let’s create this model and check the predictions:

Table

Description automatically generated

Table-1.39 Sparkling: Test/Predict dataset head

Let’s plot the train/test and forecast data.

Chart, histogram

Description automatically generated

Figure-1.25 Sparkling: TES Train/Test/Predict plot

Let’s store the model, model parameter and RMSE into the results dataset.

Text

Description automatically generated with low confidence

Table-1.40 Sparkling: Model RMSE

#### Rose Dataset

Let’s perform Triple Exponential Smoothing model on the Rose train dataset.

First we will run the triple exponential smoothing model in optimized mode and let the model determine the smoothing factor for the level, trend and seasonal. Let us check the parameters thus determined:

Text, letter

Description automatically generated

Table-1.41 Rose: TES – auto parameters

From table 1.41 we can see that TES in auto mode has determined the smoothing factor level as 0.0715, smoothing trend as 0.0453 and seasonal trend as 0.00007.

Let’s forecast for the test dataset based on the above model and check the predictions (first 5).

Graphical user interface, text

Description automatically generated

Table-1.42 Rose: Test/Predict dataset head

Let’s calculate the RMSE of the TES in auto mode model:

RMSE value on test dataset for TES autofit model is 20.1566441820518.

Let’s plot the train/test and forecast data.

Graphical user interface, chart

Description automatically generated

Figure-1.26 Rose: TES Train/Test/Predict plot

From Figure 1.26 we can see that the forecast for TES is more aligned to the test data than SES/DES.

Next we will create multiple TES models by iterating level, trend, and seasonal values from 0.1,0.2 so on till 0.9 and check the RMSE values of all the models thus created.

Text

Description automatically generated

Table-1.43 Rose: TES Models RMSE

We can see from Table 1.43 that for alpha = 0.1, beta = 0.2 and gamma = 0.1 (level, trend, seasonal smoothing constants respectively) the model has a lower RMSE (9.223453) than the RMSE value of TES autofit model (20.156644).

Let’s create this model and check the predictions:

Text

Description automatically generated with medium confidence

Table-1.44 Rose: Test/Predict dataset head

Let’s plot the train/test and forecast data.

Chart

Description automatically generated

Figure-1.27 Rose: TES Train/Test/Predict plot

Let’s store the model, model parameter and RMSE into the results dataset.

Text

Description automatically generated

Table-1.45 Rose: Model RMSE

### Linear Regression (LR):

In Linear Regression on time series data, we will create a time ordered sequence and regress the sales value against the sequence

#### Sparkling Dataset

Let’s first create a sequential column and add the same to train and test datasets.

Let’s look at the head and tail of train and test datasets.

Table

Description automatically generated with medium confidenceTable

Description automatically generated with medium confidence Graphical user interface

Description automatically generated with medium confidence

Table-1.46 Sparkling: LR Train Dataset Table-1.47 Sparkling: LR Test Dataset

Let’s build linear regression model on the train dataset with ‘Time’ as the predictor and ‘Sparkling’ (sales as the target variable.

Using this model lets see the predictions on the test dataset (first 5):

Table

Description automatically generated

Table-1.48 Sparkling: LR Test/Predict

Let’s calculate the RMSE value of the linear regression model on test dataset, which is 1389.135174897992.

Let’s plot the train, test, and predict graph:

Chart, histogram

Description automatically generated

Figure-1.28 Sparkling: LR Train/Test/Predict plot

Let’s store the model, model parameter and RMSE into the results dataset.

Text

Description automatically generated with medium confidence

Table-1.49 Sparkling: Model RMSE

#### Rose Dataset

Let’s first create a sequential column and add the same to train and test datasets.

A screenshot of a computer

Description automatically generated with low confidenceLet’s look at the head and tail of train and test datasets.

Graphical user interface, table

Description automatically generated with medium confidence Graphical user interface, application, chat or text message

Description automatically generated

Table-1.50 Rose: LR Train Dataset Table-1.51 Rose: LR Test Dataset

Let’s build linear regression model on the train dataset with ‘Time’ as the predictor and ‘Rose’ (sales as the target variable.

Using this model let’s see the predictions on the test dataset (first 5):

Table

Description automatically generated

Table-1.52 Rose: LR Test/Predict

Let’s calculate the RMSE value of the linear regression model on test dataset, which is 15.2689893739829.

Let’s plot the train, test, and predict graph:

Chart

Description automatically generated

Figure-1.29 Rose: LR Train/Test/Predict plot

Let’s store the model, model parameter and RMSE into the results dataset.

Text

Description automatically generated with low confidence

Table-1.53 Rose: Model RMSE

### Naïve Forecast:

In Naive forecast the prediction for tomorrow is the same as today and the prediction for day after tomorrow is the same as tomorrow, which in turn is the same as today. Hence the prediction value on entire test dataset will be the last observation value of the train dataset.

#### Sparkling Dataset

In Naïve model, the test dataset will have prediction values as the last value of train dataset.

Let’s add the last observation value of train dataset as prediction for test dataset.

Let’s look at the tail of train dataset and head of test dataset.

Table

Description automatically generated with medium confidence Table

Description automatically generated

Table-1.54 Sparkling: Train Dataset-Tail Table-1.55 Sparkling: Test Dataset-Head

In table 1.54 we can see that the last observation of train dataset has value 6047, which is the ‘Predict’ value across the test dataset.

The RMSE value for this Naïve Forecast model is 3864.2793518443914.

Let’s plot the train, test, and Naïve forecast as shown below:

Chart, histogram

Description automatically generated

Figure-1.30 Sparkling: Naive Train/Test/Predict plot

Let’s store the model, model parameter and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.56 Sparkling: Model RMSE

#### Rose Dataset

In Naïve model, the test dataset will have prediction values as the last value of train dataset.

Let’s add the last observation value of train dataset as prediction for test dataset.

Let’s look at the tail of train dataset and head of test dataset.

Table

Description automatically generated Table

Description automatically generated

Table-1.57 Rose: Train Dataset-Tail Table-1.58 Rose: Test Dataset-Head

In table 1.57 we can see that the last observation of train dataset has value 132, which is the ‘Predict’ value across the test dataset.

The RMSE value for this Naïve Forecast model is 79.71877616175045.

Let’s plot the train, test, and Naïve forecast as shown below:

Graphical user interface, chart

Description automatically generated

Figure-1.31 Rose: Naive Train/Test/Predict plot

Let’s store the model, model parameter and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.59 Rose: Model RMSE

### Simple Average Forecast:

In Simple average model the prediction for test is going to be the mean of the train dataset over the time series data.

#### Sparkling Dataset

Let’s build predict as the mean of the train dataset and apply it over the test dataset.

Table

Description automatically generated

Table-1.60 Sparkling: Simple Average Test Dataset

From Table 1.60 we can see that the test dataset has the ‘Predict’ variable assigned with 2403.80303 which is the mean of the train data set sales value.

The RMSE for this Simple Average model is 1275.0818036965309.

Let’s plot the train/test and predict for Simple Average model.

Chart

Description automatically generated

Figure-1.32 Sparkling: Simple Average model Train/Test/Predict plot

Let’s store the model, model parameter and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.61 Sparkling: Model RMSE

#### Rose Dataset

Let’s build predict as the mean of the train dataset and apply it over the test dataset.

Text

Description automatically generated

Table-1.62 Rose: Simple Average Test Dataset

From Table 1.62 we can see that the test dataset has the ‘Predict’ variable assigned with 104.939394 which is the mean of the train data set sales value.

The RMSE for this Simple Average model is 53.46057380286427.

Let’s plot the train/test and predict for Simple Average model.

Graphical user interface, chart

Description automatically generated

Figure-1.33 Rose: Simple Average model Train/Test/Predict plot

Let’s store the model, model parameter and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.63 Rose: Model RMSE

### Moving Average Forecast:

In the trailing moving average forecast model, we will consider multiple trailing windows, and create moving averages across the dataset. This dataset is then split into test and train based on the condition all data prior to 1991 to be considered as train and from 1991 as test.

#### Sparkling Dataset

Let’s consider the original Sparkling dataset and create different rolling windows from 2 to 9.

Table

Description automatically generated

Table-1.64 Sparkling: Moving Average dataset

Let’s plot the above data in a graph.

Chart

Description automatically generated

Figure-1.34 Sparkling: Original vs Trailing Moving Averages

Let us split the above dataset into train and test based on year 1991 and calculate the RMSE of each moving average.

Text

Description automatically generated

Table-1.65 Sparkling: Moving Average - RMSE

From Table 1.65 we can see that RMSE is least for moving average with rolling window of 2.

Let’s plot train, test and moving average with rolling window of 2.

Chart

Description automatically generated

Figure-1.35 Sparkling: Train/Test/Predict- 2 point Moving Average

Let’s store the model, model parameter and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.66 Sparkling: Model RMSE

#### Rose Dataset

Let’s consider the original Rose dataset and create different rolling windows from 2 to 9.

A screenshot of a computer

Description automatically generated

Table-1.67 Rose: Moving Average dataset

Let’s plot the above data in a graph.

Chart, line chart, histogram

Description automatically generated

Figure-1.36 Rose: Original vs Trailing Moving Averages

Let us split the above dataset into train and test based on year 1991 and calculate the RMSE of each moving average.

Graphical user interface, text

Description automatically generated

Table-1.68 Rose: Moving Average - RMSE

From Table 1.68 we can see that RMSE is least for moving average with rolling window of 2.

Let’s plot train, test and moving average with rolling window of 2.

Chart

Description automatically generated

Figure-1.37 Rose: Train/Test/Predict- 2 point Moving Average

Let’s store the model, model parameter and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.69 Rose: Model RMSE

### Inference:

Let’s look at the model comparison for Sparkling & Rose datasets based on RMSE values.

Graphical user interface, text, application

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

Table-1.70 Sparkling: Model RMSE Table-1.71 Rose: Model RMSE

From Table 1.70 we can see that for Sparkling dataset, Triple Exponential Smoothing (alpha=0.4, beta = 0.1 & gamma = 0.2) has the least RMSE values of all the other models evaluated. This is not surprising as the Sparkling dataset has level, trend, and seasonality as we had seen in Decompose section of question 2.

Similarly, we can see from Table 1.71 that for Rose dataset Triple Exponential Smoothing (alpha=0.1, beta = 0.2 & gamma = 0.1) has the least RMSE values of all the other models evaluated. This is because Triple Exponential smoothing model considers level, trend, and seasonality, which are present in the Rose dataset.

In Rose dataset we can see the 2 point Moving Average and Linear Regression shows better results than that of Sparkling dataset. This is because the trend is prominent in Rose dataset, whereas in Sparkling dataset trend is not so apparent.

## Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and 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.

For testing stationarity of a time series, we can run the Augmented Dickey-Fuller(ADF) test. This 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:

𝐻o : The Time Series has a unit root and is thus non-stationary.

𝐻a : The Time Series does not have a unit root and is thus stationary.

If a given time series is non-stationary, we can take one order of differencing and then check stationarity on that time series and in case its still non-stationary then 2 order of differencing and so on.

Stationarity is to be checked at alpha = 0.05, which means that for the data to be stationary the p-value of the ADF test must be lesser than 0.05.

The models are to be built on the train datasets of Sparkling and Rose. Hence we will test for stationarity for these train datasets.

### Sparkling Dataset:

Let’s consider the train dataset of Sparkling. The plot of training dataset with rolling mean and standard deviation with trailing window of 12 is given below:

Chart, histogram

Description automatically generated

Figure-1.38 Sparkling: Train dataset

Let’s perform the ADF test on the training dataset and check the stats:

Text

Description automatically generated with low confidence

Table-1.72 Sparkling: ADF on Train dataset

From table 1.72 we can see that p-value is 0.56741 which is greater than 0.05. Hence we don’t have enough evidence to reject the null hypothesis, the time series has a unit root and is thus non-stationary.

Let’s consider the first order of differencing of the training dataset. The plot of the training dataset with first order of differencing and corresponding rolling mean and standard deviation with trailing window of 12 is given below:

Chart

Description automatically generated

Figure-1.39 Sparkling: Train dataset with first order of differencing

Let’s perform the ADF test on the training dataset with first order of differencing and check the stats:

A picture containing text

Description automatically generated

Table-1.73 Sparkling: ADF on Train dataset with first order of differencing

From table 1.73 we can see that p-value is 8.4792e-11 which is lesser than 0.05. Hence we have enough evidence to reject the null hypothesis, hence the alternate hypothesis holds true. So, the training dataset of Sparkling with first order of differencing does not have a unit root and is thus stationary.

### Rose Dataset:

Let’s consider the train dataset of Rose. The plot of training dataset with rolling mean and standard deviation with trailing window of 12 is given below:

Graphical user interface, chart, line chart

Description automatically generated

Figure-1.40 Rose: Train dataset

Let’s perform the ADF test on the training dataset and check the stats:

A picture containing text

Description automatically generated

Table-1.74 Rose: ADF on Train dataset

From table 1.74 we can see that p-value is 0.7569 which is greater than 0.05. Hence we don’t have enough evidence to reject the null hypothesis, the time series has a unit root and is thus non-stationary.

Let’s consider the first order of differencing of the training dataset. The plot of the training dataset with first order of differencing and corresponding rolling mean and standard deviation with trailing window of 12 is given below:

Graphical user interface, chart

Description automatically generated

Figure-1.41 Rose: Train dataset with first order of differencing

Let’s perform the ADF test on the training dataset with first order of differencing and check the stats:

A picture containing text

Description automatically generated

Table-1.75 Rose: ADF on Train dataset with first order of differencing

From table 1.75 we can see that p-value is 3.8948e-8 which is lesser than 0.05. Hence we have enough evidence to reject the null hypothesis, hence the alternate hypothesis holds true. So, the training dataset of Sparkling with first order of differencing does not have a unit root and is thus stationary.

## Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

### ARIMA model:

#### Sparkling Dataset

Let’s first look at the ACF and PACF plots of first order difference of training dataset ( we have established that in question 5 that first order difference of training dataset is stationary).

Timeline, box and whisker chart

Description automatically generated

Figure-1.42 Sparkling: ACF plot

Timeline

Description automatically generated

Figure-1.43 Sparkling: PACF plot

From Figure 1.42 we can see that the ACF plot cuts off below 95% confidence interval after lag of 0, hence visually we can say that q = 0.

From Figure 1.43 we can see that PACF plot cuts off below 95% confidence interval after lag of 0, hence visually we can say that p =0.

Since the training dataset is stationary at first order of differencing, value of d=1.

For the auto ARIMA model, we will evaluate models where d=1 and p, q ranges from 0 to 2.

We can see few examples of the model parameter below:

Text

Description automatically generated

Table-1.76 Sparkling: Auto ARIMA model parameters

We will create ARIMA models on the training dataset, based on the above p,d,q parameters and calculate the AIC value. Let’s look at the results:

Text

Description automatically generated

Table-1.77 Sparkling: Auto ARIMA AIC values

From table 1.77 we can see that AIC model is least for model parameters(p=2,d=1,q=2).

Let’s build this ARIMA model (p=2,d=1,q=2), and let’s look at the model summary and plot diagnostics.

Table

Description automatically generated

Table-1.78 Sparkling: Auto ARIMA Summary

Chart, histogram

Description automatically generated

Figure-1.44 Sparkling: Auto ARIMA diagnostic plot

Standardized residual does not display any obvious seasonality. The KDE plot of residuals is like normal distribution hence we can say that the model residuals are normally distributed. The Q-Q plot shows the samples were mostly taken from a normal distribution. The residuals have very low correlation with lagged versions of itself.

Let’s predict the values for test dataset using this auto ARIMA model.

Table

Description automatically generated

Table-1.79 Sparkling: Auto ARIMA Test/Prediction Head

Let’s plot the train/test and auto ARIMA prediction.

Chart

Description automatically generated

Figure-1.45 Sparkling: Train, Test, Auto ARIMA prediction

The RMSE value for the Auto ARIMA model on test dataset is 1299.9795689481477.

Let’s store the model parameter, model and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.80 Sparkling: Model - RMSE

#### Rose Dataset

Let’s first look at the ACF and PACF plots of first order difference of training dataset ( we have established that in question 5 that first order difference of training dataset is stationary).

Chart, timeline

Description automatically generated

Figure-1.46 Rose: ACF plot

Timeline

Description automatically generated

Figure-1.47 Rose: PACF plot

From Figure 1.46 we can see that the ACF plot cuts off below 95% confidence interval after lag of 2, hence visually we can say that q = 2.

From Figure 1.43 we can see that PACF plot cuts off below 95% confidence interval after lag of 2, hence visually we can say that p =2.

Since the training dataset is stationary at first order of differencing, value of d=1.

For the auto ARIMA model, we will evaluate models where d=1 and p,q ranges from 0 to 2.

We can see few examples of the model parameter below:

A picture containing text

Description automatically generated

Table-1.81 Rose: Auto ARIMA model parameters

We will create ARIMA models on the training dataset, based on the above p,d,q parameters and calculate the AIC value. Let’s look at the results:

Text

Description automatically generated

Table-1.82 Rose: Auto ARIMA AIC values

From table 1.82 we can see that AIC model is least for model parameters(p=2,d=1,q=3).

Let’s build this ARIMA model (p=2,d=1,q=3), and let’s look at the model summary and plot diagnostics.

Table

Description automatically generated

Table-1.83 Rose: Auto ARIMA Summary

Chart, line chart, histogram

Description automatically generated

Figure-1.48 Rose: Auto ARIMA diagnostic plot

Standardized residual does not display any obvious seasonality. The KDE plot of residuals is like normal distribution hence we can say that the model residuals are normally distributed. The Q-Q plot shows the samples were mostly taken from a normal distribution. The residuals have very low correlation with lagged versions of itself.

Let’s predict the values for test dataset using this auto ARIMA model.

A screenshot of a computer

Description automatically generated with low confidence

Table-1.84 Rose: Auto ARIMA Test/Prediction Head

Let’s plot the train/test and auto ARIMA prediction.

Chart

Description automatically generated

Figure-1.49 Rose: Train, Test, Auto ARIMA prediction

The RMSE value for the Auto ARIMA model on test dataset is 36.81742866084295.

Let’s store the model parameter, model and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.85 Rose: Model - RMSE

### SARIMA model:

#### Sparkling Dataset

Let’s first look at the ACF and PACF plots of first order difference of training dataset ( we have established that in question 5 that first order difference of training dataset is stationary).

Timeline

Description automatically generated

Figure-1.50 Sparkling: ACF – autocorrelation plot

Timeline

Description automatically generated

Figure-1.51 Sparkling: PACF – partial autocorrelation plot

We have seen earlier that p=0, q=0 and since stationarity is at first order of difference, d= 1. From Figure 1.50 we can see that there is a seasonality after every 12 lags. Based on this, from Figure 1.50, if we only look at multiples of 12, we can see that after lag 48, the ACF cuts below 95% confidence interval, hence Q =4. From Figure 1.51, if we only look at multiples of 12, we can see that after lag 12, the PCAF cuts below 95% CI, hence P = 1.

For the auto SARIMA model, we will evaluate models where d=1, p and q ranges from 0 to 2, P and Q ranges from 0 to 4 and D = 0.

We can see few examples of the model parameter below:

Text

Description automatically generated

Table-1.86 Sparkling: Auto SARIMA model parameters

We will create SARIMA models on the training dataset, based on the above (p,d,q)(P,D,Q,12) parameters and calculate the AIC value. Let’s look at the results:

Text, chat or text message

Description automatically generated

Table-1.87 Sparkling: Auto SARIMA AIC values

From table 1.87 we can see that AIC model is least for model parameters (p=0,d=1,q=2)(P=2,D=0,Q=4,Seasonality=12).

Let’s build this SARIMA model (p=0,d=1,q=2)(P=2,D=0,Q=4,Seasonality=12), and let’s look at the model summary and plot diagnostics.

Table

Description automatically generated

Table-1.88 Sparkling: Auto SARIMA model summary

Chart, histogram

Description automatically generated

Figure-1.52 Sparkling: Auto SARIMA model diagnostics

Standardized residual does not display any obvious seasonality. The KDE plot of residuals is like normal distribution hence we can say that the model residuals are normally distributed. The Q-Q plot shows the samples were mostly taken from a normal distribution. The residuals have very low correlation with lagged versions of itself.

Let’s predict the values for test dataset using this auto SARIMA model.

Table

Description automatically generated

Table-1.89 Sparkling: Auto SARIMA model prediction head

Let’s plot the train/test and auto SARIMA prediction.

Chart

Description automatically generated

Figure-1.53 Sparkling: Train, Test Auto SARIMA model prediction

The RMSE value for the Auto SARIMA model on test dataset is 487.38855106632866.

Let’s store the model parameter, model and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.90 Sparkling: Model - RMSE

#### Rose Dataset

Let’s first look at the ACF and PACF plots of first order difference of training dataset ( we have established that in question 5 that first order difference of training dataset is stationary).

Chart, timeline, box and whisker chart

Description automatically generated

Figure-1.54 Rose: ACF – autocorrelation plot

Timeline

Description automatically generated

Figure-1.55 Rose: PACF – partial autocorrelation plot

We have seen earlier that p=2, q=2 and since stationarity is at first order of difference, d= 1. From Figure 1.54 we can see that there is a seasonality after every 12 lags. Based on this, from Figure 1.54, if we only look at multiples of 12, we can see that after lag 36, the ACF cuts below 95% confidence interval, hence Q =3. From Figure 1.55, if we only look at multiples of 12, we can see that after lag 12, the PCAF cuts below 95% CI, hence P = 1.

For the auto SARIMA model, we will evaluate models where d=1, p and q ranges from 0 to 2, P and Q ranges from 0 to 4 and D = 0.

We can see few examples of the model parameter below:

Text

Description automatically generated

Table-1.91 Rose: Auto SARIMA model parameters

We will create SARIMA models on the training dataset, based on the above (p,d,q)(P,D,Q,12) parameters and calculate the AIC value. Let’s look at the results:

Text, chat or text message

Description automatically generated

Table-1.92 Rose: Auto SARIMA AIC values

From table 1.92 we can see that AIC model is least for model parameters (p=0,d=1,q=2)(P=3,D=0,Q=4,Seasonality=12).

Let’s build this SARIMA model (p=0,d=1,q=2)(P=3,D=0,Q=4,Seasonality=12), and let’s look at the model summary and plot diagnostics.

Table

Description automatically generated

Table-1.93 Rose: Auto SARIMA model summary

Chart, histogram

Description automatically generated

Figure-1.56 Rose: Auto SARIMA model diagnostics

Standardized residual does not display any obvious seasonality. The KDE plot of residuals is like normal distribution hence we can say that the model residuals are normally distributed. The Q-Q plot shows the samples were mostly taken from a normal distribution. The residuals have very low correlation with lagged versions of itself.

Let’s predict the values for test dataset using this auto SARIMA model.

Text

Description automatically generated

Table-1.94 Rose: Auto SARIMA model prediction head

Let’s plot the train/test and auto SARIMA prediction.

Chart

Description automatically generated

Figure-1.57 Rose: Auto SARIMA model diagnostics

The RMSE value for the Auto SARIMA model on test dataset is 18.479344474798747.

Let’s store the model parameter, model and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.95 Rose: Model - RMSE

## 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.

### ARIMA model:

#### Sparkling Dataset

Let’s first look at the ACF and PACF plots of first order difference of training dataset ( we have established that in question 5 that first order difference of training dataset is stationary).

Timeline

Description automatically generated

Figure-1.58 Sparkling: ACF – autocorrelation plot

Timeline

Description automatically generated

Figure-1.59 Sparkling: PACF – partial autocorrelation plot

From Figure 1.58 we can see that the ACF plot cuts off below 95% confidence interval after lag of 0, hence visually we can say that q = 0.

From Figure 1.59 we can see that PACF plot cuts off below 95% confidence interval after lag of 0, hence visually we can say that p =0.

Since the training dataset is stationary at first order of differencing, value of d=1.

Let’s create ARIMA model (p=0,d=1,q=0) on the Sparkling train dataset and check model summary and diagnostic plots.

Table

Description automatically generated

Table-1.96 Sparkling: ARIMA Model summary

Chart, histogram

Description automatically generated

Figure-1.60 Sparkling: ARIMA plot diagnostics

Standardized residual does not display any obvious seasonality. The KDE plot of residuals is like normal distribution hence we can say that the model residuals are normally distributed. The Q-Q plot shows the samples were mostly taken from a normal distribution. The residuals have very low correlation with lagged versions of itself.

Let’s predict the values for test dataset using this ARIMA model.

Table

Description automatically generated

Table-1.97 Sparkling: ARIMA model prediction head

Let’s plot the train/test and ARIMA prediction.

Chart, line chart

Description automatically generated

Figure-1.61 Sparkling: Train, Test and ARIMA prediction

The RMSE for the test dataset with this ARIMA model is 3864.2793518443914.

Let’s store the model parameter, model and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.98 Sparkling: Models - RMSE

#### Rose Dataset

Let’s first look at the ACF and PACF plots of first order difference of training dataset ( we have established that in question 5 that first order difference of training dataset is stationary).

Chart, timeline, box and whisker chart

Description automatically generated

Figure-1.62 Rose: ACF – autocorrelation plot

Timeline

Description automatically generated

Figure-1.63 Rose: PACF – partial autocorrelation plot

From Figure 1.62 we can see that the ACF plot cuts off below 95% confidence interval after lag of 2, hence visually we can say that q = 2.

From Figure 1.63 we can see that PACF plot cuts off below 95% confidence interval after lag of 2, hence visually we can say that p =2.

Since the training dataset is stationary at first order of differencing, value of d=1.

Let’s create ARIMA model (p=2,d=1,q=2) on the Rose train dataset and check model summary and diagnostic plots.

Table

Description automatically generated

Table-1.99 Rose: ARIMA Model summary

Chart, histogram

Description automatically generated

Figure-1.64 Rose: ARIMA plot diagnostics

Standardized residual does not display any obvious seasonality. The KDE plot of residuals is like normal distribution hence we can say that the model residuals are normally distributed. The Q-Q plot shows the samples were mostly taken from a normal distribution. The residuals have very low correlation with lagged versions of itself.

Let’s predict the values for test dataset using this ARIMA model.

Graphical user interface, text, application

Description automatically generated

Table-1.100 Rose: ARIMA model prediction head

Let’s plot the train/test and ARIMA prediction.

Chart

Description automatically generated

Figure-1.65 Rose: Train, Test and ARIMA prediction

The RMSE for the test dataset with this ARIMA model is 36.87120264924714.

Let’s store the model parameter, model and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.101 Rose: Models - RMSE

### SARIMA model:

#### Sparkling Dataset

Let’s first look at the ACF and PACF plots of first order difference of training dataset ( we have established that in question 5 that first order difference of training dataset is stationary).

Timeline

Description automatically generated

Figure-1.66 Sparkling: ACF – autocorrelation plot

Timeline

Description automatically generated

Figure-1.67 Sparkling: PACF – partial autocorrelation plot

From Figure 1.66 we can see that the ACF plot cuts off below 95% confidence interval after lag of 0, hence visually we can say that q = 0. We can see that data has seasonality of 12 lags. If we only look at multiples of 12, after 48, the ACF plot cuts off below 95% CI, hence Q = 4.

From Figure 1.67 we can see that PACF plot cuts off below 95% confidence interval after lag of 0, hence visually we can say that p =0. If we only look at multiples of 12, after 12, the ACF plot cuts off below 95% CI, hence P = 1.

Since the training dataset is stationary at first order of differencing, value of d=1.

Let’s create SARIMA model (p=0,d=1,q=0,P=1,D=0,Q=4,Seasonality=12) on the Sparkling train dataset and check model summary and diagnostic plots.

Table

Description automatically generated

Table-1.102 Sparkling: SARIMA Model summary

Chart, histogram

Description automatically generated

Figure-1.68 Sparkling: SARIMA plot diagnostics

Standardized residual does not display any obvious seasonality. The KDE plot of residuals is like normal distribution hence we can say that the model residuals are normally distributed. The Q-Q plot shows the samples were mostly taken from a normal distribution. The residuals have very low correlation with lagged versions of itself.

Let’s predict the values for test dataset using this SARIMA model.

Table

Description automatically generated

Table-1.103 Sparkling: SARIMA model prediction head

Let’s plot the train/test and SARIMA prediction.

Chart

Description automatically generated

Figure-1.69 Sparkling: Train, Test and SARIMA prediction

The RMSE for the test dataset with this SARIMA model is 1159.0083426709882.

Let’s store the model parameter, model and RMSE into the results dataset.

Graphical user interface, text, application

Description automatically generated

Table-1.104 Sparkling: Models - RMSE

#### Rose Dataset

Let’s first look at the ACF and PACF plots of first order difference of training dataset ( we have established that in question 5 that first order difference of training dataset is stationary).

Chart, timeline, box and whisker chart

Description automatically generated

Figure-1.70 Rose: ACF – autocorrelation plot

Timeline

Description automatically generated

Figure-1.71 Rose: PACF – partial autocorrelation plot

From Figure 1.70 we can see that the ACF plot cuts off below 95% confidence interval after lag of 2, hence visually we can say that q = 2. We can see that data has seasonality of 12 lags. If we only look at multiples of 12, after 36, the ACF plot cuts off below 95% CI, hence Q = 3.

From Figure 1.71 we can see that PACF plot cuts off below 95% confidence interval after lag of 2, hence visually we can say that p =2. If we only look at multiples of 12, after 12, the ACF plot cuts off below 95% CI, hence P = 1.

Since the training dataset is stationary at first order of differencing, value of d=1.

Let’s create SARIMA model (p=2,d=1,q=2,P=1,D=0,Q=3,Seasonality=12) on the Rose train dataset and check model summary and diagnostic plots.

Table

Description automatically generated

Table-1.105 Rose: SARIMA Model summary

Chart, histogram

Description automatically generated

Figure-1.72 Rose: SARIMA plot diagnostics

Standardized residual does not display any obvious seasonality. The KDE plot of residuals is like normal distribution hence we can say that the model residuals are normally distributed. The Q-Q plot shows the samples were mostly taken from a normal distribution. The residuals have very low correlation with lagged versions of itself.

Let’s predict the values for test dataset using this SARIMA model.

Text, table

Description automatically generated

Table-1.106 Rose: SARIMA model prediction head

Let’s plot the train/test and ARIMA prediction.

Chart

Description automatically generated

Figure-1.73 Rose: Train, Test and SARIMA prediction

The RMSE for the test dataset with this SARIMA model is 13.660427723818085.

Let’s store the model parameter, model and RMSE into the results dataset.

Graphical user interface, text, application, email

Description automatically generated

Table-1.107 Rose: Models - RMSE

## Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

#### Sparkling Dataset

Let’s look at the table with the model, parameters and RMSE values (for the models on test trainset).

Graphical user interface, text, application

Description automatically generated

Table-1.108 Sparkling: Models – RMSE

From table 1.108 we can see that Triple Exponential Smoothing with alpha = 0.4, beta = 0.1 and

gamma = 0.2 has the least RMSE value of all the models. The next best model is auto SARIMA model with parameters(2,1,2) and seasonal parameter(2,0,4,12).

#### Rose Dataset

Let’s look at the table with the model, parameters and RMSE values (for the models on test trainset).

Graphical user interface, text, application

Description automatically generated

Table-1.109 Rose: Models – RMSE

From table 1.109 we can see that Triple Exponential Smoothing with alpha = 0.1, beta = 0.2 and

gamma = 0.1 has the least RMSE value of all the models. The next best model is the 2 point Moving Average model.

## 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.

#### Sparkling Dataset

In question 8 we have determined that the best forecast model is Triple Exponential Smoothing (Holt-Winter's forecasting) with model parameters - alpha = 0.4, beta = 0.1 and gamma = 0.2. Let’s build the model on the complete Sparkling dataset.

The RMSE value of the model on the Sparkling dataset is 386.4215.

Let’s forecast 12 months into the future & also create upper and lower confidence bands at 95%.

The forecast data looks as shown below:

Table

Description automatically generated

Table-1.110 Sparkling: 12 months forecast

Let’s plot the Sparkling dataset and the 12 month’s forecast together with 95% confidence bands.

Chart

Description automatically generated

Figure-1.74 Sparkling: Sales & 12 months Forecast

#### Rose Dataset

In question 8 we have determined that the best forecast model is Triple Exponential Smoothing (Holt-Winter's forecasting) with model parameters - alpha = 0.1, beta = 0.2 and gamma = 0.1. Let’s build the model on the complete Rose dataset.

The RMSE value of the model on the Sparkling dataset is 17.1155.

Let’s forecast 12 months into the future & also create upper and lower confidence bands at 95%.

The forecast data looks as shown below:

Table

Description automatically generated

Table-1.111 Rose: 12 months forecast

Let’s plot the Sparkling dataset and the 12 month’s forecast together with 95% confidence bands.

Chart, line chart

Description automatically generated

Figure-1.75 Rose: Sales & 12 months Forecast

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

#### Model building Summary

For both the datasets – Sparkling and Rose, we have run the datasets against the following models.

* Simple Exponential Smoothing
* Double Exponential Smoothing
* Triple Exponential Smoothing
* Linear Regression
* Naïve Forecast
* Simple Average Forecast
* Moving Average Forecast
* Auto ARIMA
* Auto SARIMA
* ARIMA – based on ACF/PACF cutoff
* SARIMA – Based on ACF/PACF cutoff

The modeling technique with best performance based on the least RMSE value against test dataset has been Triple Exponential Smoothing. This is a good model in this scenario as the level, trend and seasonality are all considered. By looking at the forecast of this model (refer section 9), we can see that trend and seasonality are in line with the original sales dataset respectively. On most models we have run against various model parameters and hence the RMSE results and the choice of appropriate model provides a good amount of confidence.

#### Observations

* We can see that average monthly and yearly sales of Sparkling wine exceeds Rose wine by a huge margin.
* Overall, the sales trend for Sparkling Wine is stable, we can see that the trend is on a downslide from 1988 to 1995.
* The sales trend for Rose Wine is on a decline.
* Average monthly sales for Sparkling (across the years) in the first 6 months of the year (January to June) is around 1750, but in the next 6 months the sales increase exponentially to 6000 in December.
* Average monthly sales for Rose (across the years) have a much gradual increase from January (around 60) to December (around 125).
* Least sales for both types of wine occur in January whereas maximum sales for both types occur in December.
* No reason is provided as to why the sales are decreasing, is it because the source materials are rare, as in the wines are a luxury or collector’s item. Both the wines have existed for more than 15 years, hence, will need more information to provide meaningful action items.

#### Suggestions

* January to June is lean period for sale of both types of wine. Business will have to run a discount campaign to increase retail sales.
* Business will also have to liaise with hotels/restaurants to ensure Sparkling/Rose wine is prominently listed in their offering. This could boost the sales in lean period (January-December).
* June to December is the peak season for sales, hence business will need to forecast all source materials accordingly that demand in July-December period is met.
* Discounts and campaigns need to be run throughout the year, customized for the lean and high season.
* Business needs to promote their wine offering as part of loyalty programs offered by retailers.
* Sales opportunities through e-commerce channels also needs to be pursued.
* It will be great marketing if the business could arrange wine tasting tours in their estate.
* Wine tasting events in localities where wine sales/consumption is high could also be a great marketing opportunity.
* Discounts for loyal customers based on purchase, could be a good way to ensure a healthy customer base.

## THE END