IMAAD IMRAN HAJWANE

202101132 / 23

DMPM LAB

ASSIGNMENT 01

STATEMENT:

Data Exploration & Visualization

1. Read the dataset file that is supplied to you.

2. Identify the variables in the file and determine whether any variable has any missing values.

3. Input some of the variables that have missing values using their corresponding mean values. Verify whether your task has been correctly done.

4. Determine the "summary" information for the numerical variables.

5. Identify the "distributions" of the numerical variables and plot the distributions.

6. Transform the numeric variables into their natural log values and scale [0 - 1] values.

7. Check whether the numeric variables follow normality conditions.

8. Find the correlation matrix for all the variables in the dataset and plot the graph of the correlation matrix.

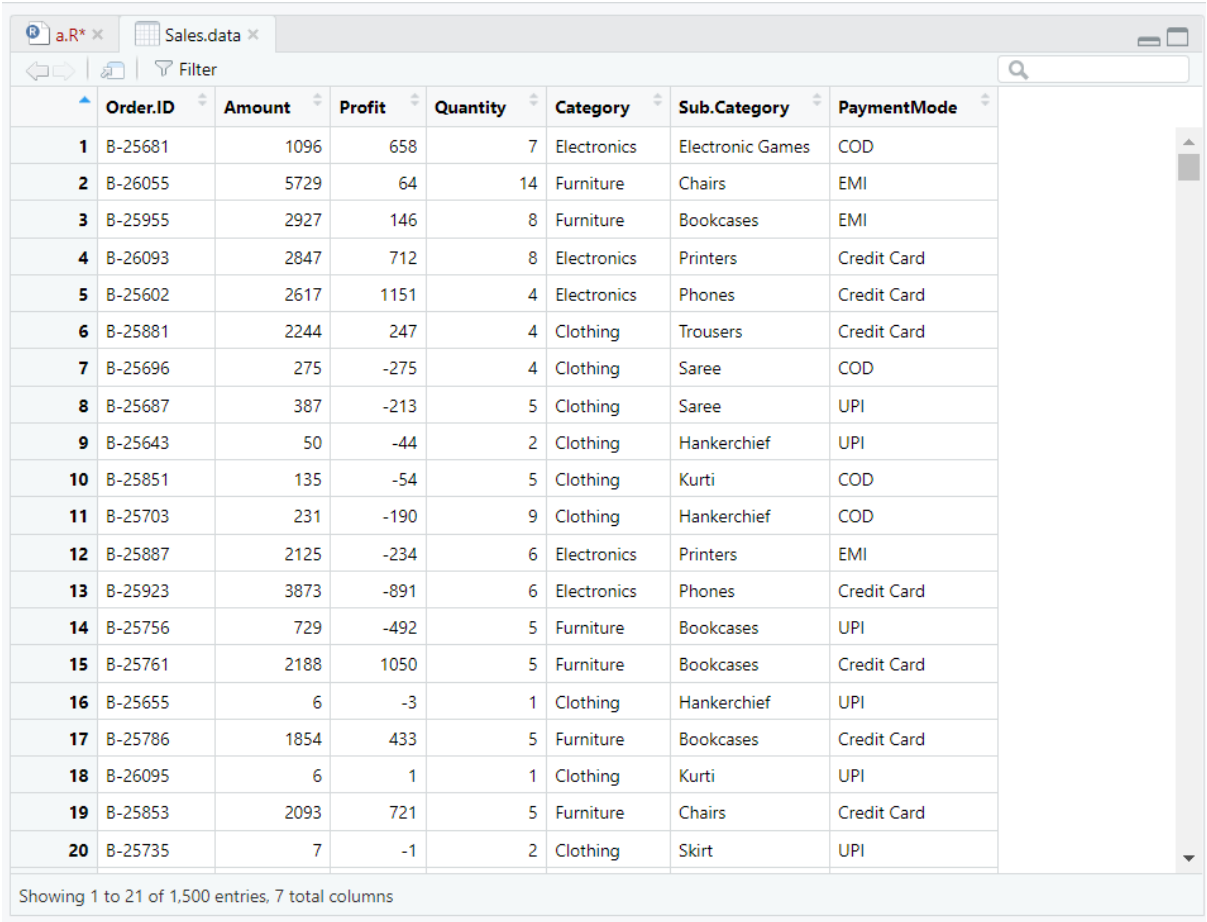
9. Any additional ways of Data Exploration & Visualization will be highly appreciated.

SOLUTIONS:

1.

Sales.data <- read.csv ("C:/Users/iamim/OneDrive/Desktop/Sixth Semester/DMPM_LAB/A1/Sales data.csv")

View (Sales.data)



a.R* × Sales.data ×

Filter

	Order.ID	Amount	Profit	Quantity	Category	Sub.Category	PaymentMode
1	B-25681	1096	658	7	Electronics	Electronic Games	COD
2	B-26055	5729	64	14	Furniture	Chairs	EMI
3	B-25955	2927	146	8	Furniture	Bookcases	EMI
4	B-26093	2847	712	8	Electronics	Printers	Credit Card
5	B-25602	2617	1151	4	Electronics	Phones	Credit Card
6	B-25881	2244	247	4	Clothing	Trousers	Credit Card
7	B-25696	275	-275	4	Clothing	Saree	COD
8	B-25687	387	-213	5	Clothing	Saree	UPI
9	B-25643	50	-44	2	Clothing	Hankerchief	UPI
10	B-25851	135	-54	5	Clothing	Kurti	COD
11	B-25703	231	-190	9	Clothing	Hankerchief	COD
12	B-25887	2125	-234	6	Electronics	Printers	EMI
13	B-25923	3873	-891	6	Electronics	Phones	Credit Card
14	B-25756	729	-492	5	Furniture	Bookcases	UPI
15	B-25761	2188	1050	5	Furniture	Bookcases	Credit Card
16	B-25655	6	-3	1	Clothing	Hankerchief	UPI
17	B-25786	1854	433	5	Furniture	Bookcases	Credit Card
18	B-26095	6	1	1	Clothing	Kurti	UPI
19	B-25853	2093	721	5	Furniture	Chairs	Credit Card
20	B-25735	7	-1	2	Clothing	Skirt	UPI

Showing 1 to 21 of 1,500 entries, 7 total columns

2.

str (Sales.data) # Display the structure of the dataset

summary (Sales.data) # Display summary statistics of the dataset

```
> # Step 2: Overview of the dataset
> str(Sales.data)    # Display the structure of the dataset
'data.frame':   1500 obs. of  7 variables:
 $ Order.ID    : chr  "B-25681" "B-26055" "B-25955" "B-26093" ...
 $ Amount      : int  1096 5729 2927 2847 2617 2244 275 387 50 135 ...
 $ Profit      : int  658 64 146 712 1151 247 -275 -213 -44 -54 ...
 $ Quantity    : int  7 14 8 8 4 4 4 5 2 5 ...
 $ Category    : chr  "Electronics" "Furniture" "Furniture" "Electronics" ...
 $ Sub.Category: chr  "Electronic Games" "Chairs" "Bookcases" "Printers" ...
 $ PaymentMode : chr  "COD" "EMI" "EMI" "Credit Card" ...
> summary(Sales.data) # Display summary statistics of the dataset
   Order.ID             Amount           Profit            Quantity
 Length:1500        Min.   :   4.00   Min.   :-1981.00   Min.   : 1.000
 Class :character   1st Qu.:  47.75   1st Qu.:  -12.00   1st Qu.: 2.000
 Mode  :character   Median : 122.00   Median :    8.00   Median : 3.000
                    Mean   : 291.85   Mean   :   24.64   Mean   : 3.743
                    3rd Qu.: 326.25   3rd Qu.:   38.00   3rd Qu.: 5.000
                    Max.   :5729.00   Max.   : 1864.00   Max.   :14.000
   Category         Sub.Category       PaymentMode
 Length:1500        Length:1500        Length:1500
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

#Get variable names

variables <- colnames (Sales.data)

cat ("Variables in the dataset:", variables, "\n")

```
> variables <- colnames(Sales.data)
> cat("Variables in the dataset:", variables, "\n")
Variables in the dataset: Order.ID Amount Profit Quantity Category Sub.Category PaymentMode
>
```

```
Values
  variables          chr [1:7] "Order.ID" "Amount" "Profit" "Quant…
```

Check for missing values

missing_values <- colSums(is.na (Sales.data))

cat ("Missing values for each variable:\n", missing_values, "\n")

```
> # Step 4: Check for missing values
> missing_values <- colSums(is.na(Sales.data))
> cat("Missing values for each variable:\n", missing_values, "\n")
Missing values for each variable:
 0 0 0 0 0 0 0
```

Display percentage of missing values for each variable

percentage_missing <- (missing_values / nrow(Sales.data)) * 100

cat ("Percentage of missing values for each variable:\n", percentage_missing, "\n")

```
> # Optional: Display percentage of missing values for each variable
> percentage_missing <- (missing_values / nrow(Sales.data)) * 100
> cat("Percentage of missing values for each variable:\n", percentage_missing, "\n")
Percentage of missing values for each variable:
 0 0 0 0 0 0 0
```

3.

```
> Sales.data$Profit[is.na(Sales.data$Profit)]=mean(Sales.data$Profit)
> View(Sales.data)
```

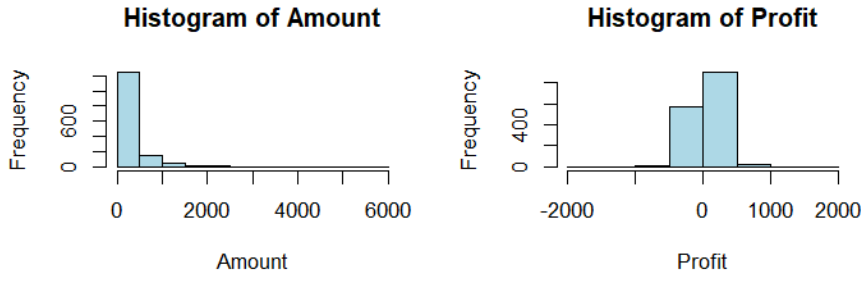
4.

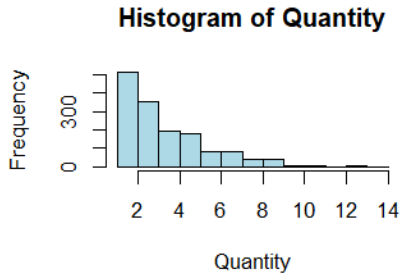
Select only numerical variables
> numerical_variables <- Sales.data[sapply(Sales.data, is.numeric)]
> # Obtain summary information for numerical variables
> numerical_summary <- summary(numerical_variables)
> # Print the summary
> print(numerical_summary)

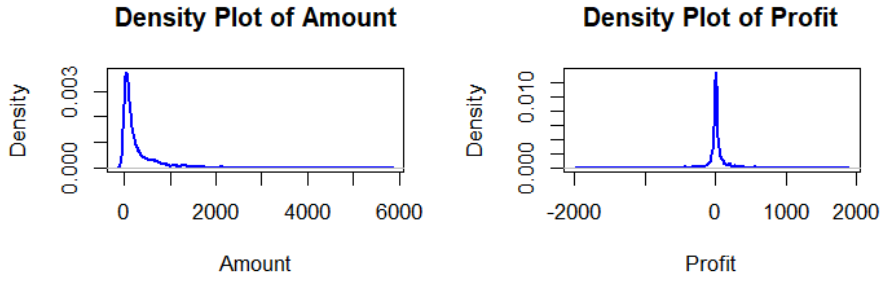
```
> # Select only numerical variables
> numerical_variables <- Sales.data[sapply(Sales.data, is.numeric)]
> # Obtain summary information for numerical variables
> numerical_summary <- summary(numerical_variables)
> # Print the summary
> print(numerical_summary)
     Amount            Profit            Quantity
 Min.   :   4.00   Min.   :-1981.00   Min.   : 1.000
 1st Qu.:  47.75   1st Qu.:  -12.00   1st Qu.: 2.000
 Median : 122.00   Median :    8.00   Median : 3.000
 Mean   : 291.85   Mean   :   24.64   Mean   : 3.743
 3rd Qu.: 326.25   3rd Qu.:   38.00   3rd Qu.: 5.000
 Max.   :5729.00   Max.   : 1864.00   Max.   :14.000
>
```

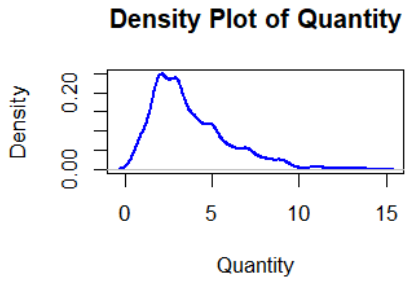
5.

```
# Select only numerical variables

numerical_variables <- Sales.data[sapply(Sales.data, is.numeric)]

# Plot histograms for numerical variables

par(mfrow = c(2, 2))  # Setting up a 2x2 grid for subplots

for (variable in colnames(numerical_variables)) {

  hist(numerical_variables[[variable]], main = paste("Histogram of", variable), col = "lightblue", border = "black", xlab = variable)

}

# Plot density plots for numerical variables

par(mfrow = c(2, 2))  # Resetting the layout

for (variable in colnames(numerical_variables)) {

  plot(density(numerical_variables[[variable]]), main = paste("Density Plot of", variable), col = "blue", lwd = 2, xlab = variable)

}
```









6.

Assuming 'Sales.data' is your dataset

Replace 'YourNumericVariable1', 'YourNumericVariable2', etc. with the actual numerical variable names in your dataset

Select only numerical variables

numerical_variables <- Sales.data[sapply(Sales.data, is.numeric)]

Log transformation

log_transformed_data <- log(numerical_variables + 1) # Adding 1 to avoid log(0) issues

Scaling to the range [0, 1]

scaled_data <- scale(log_transformed_data, center = FALSE, scale = apply(log_transformed_data, 2, max))

Print the first few rows of the scaled data for verification

print(head(scaled_data))

```
> # Scaling to the range [0, 1]
> scaled_data <- scale(log_transformed_data, center = FALSE, scale = apply(log_transformed_data, 2, max))
> # Print the first few rows of the scaled data for verification
> print(head(scaled_data))
        Amount Profit  Quantity
[1,] 0.8089626    NaN 0.7678741
[2,] 1.0000000    NaN 1.0000000
[3,] 0.9224131    NaN 0.8113677
[4,] 0.9192118    NaN 0.8113677
[5,] 0.9094808    NaN 0.5943161
[6,] 0.8917186    NaN 0.5943161
>
```

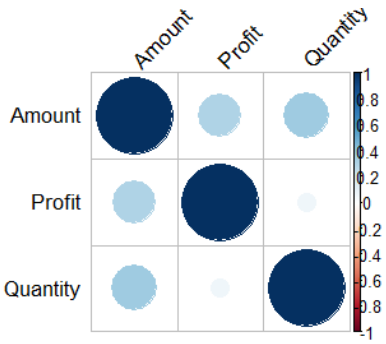
7.

```
# Select only transformed and scaled numeric variables
scaled_numerical_variables <- scaled_data
# Shapiro-Wilk Test
shapiro_test_results <- sapply(scaled_numerical_variables, function(x)
shapiro.test(x)$p.value)
# Q-Q Plots
par(mfrow = c(2, 2))  # Setting up a 2x2 grid for subplots
for (variable in colnames(scaled_numerical_variables)) {
  qqnorm(scaled_numerical_variables[[variable]], main = paste("Q-Q Plot of", variable))
  qqline(scaled_numerical_variables[[variable]], col = 2)
}
# Print the results of the Shapiro-Wilk Test
print(shapiro_test_results)
```

8.

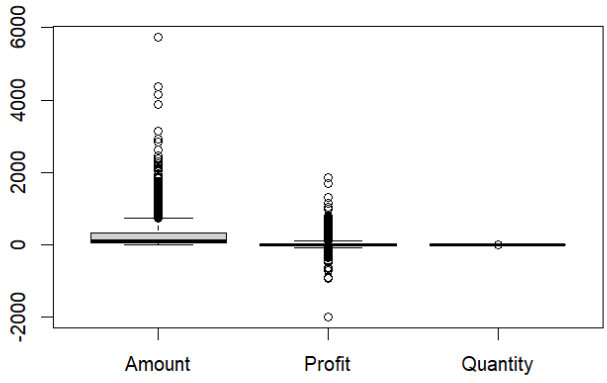
```
# Select only numeric variables
numeric_variables <- Sales.data[sapply(Sales.data, is.numeric)]
# Calculate the correlation matrix
correlation_matrix <- cor(numeric_variables, use = "complete.obs")
# Print the correlation matrix
print(correlation_matrix)
# Plot the graph of the correlation matrix using corrplot
# Install the corrplot package if not already installed
# install.packages("corrplot")
library(corrplot)
# Plotting the correlation matrix
corrplot(correlation_matrix, method = "circle", type = "full", tl.col = "black", tl.srt = 45)
```

```
> # Calculate the correlation matrix
> correlation_matrix <- cor(numeric_variables, use = "complete.obs")
> # Print the correlation matrix
> print(correlation_matrix)
             Amount     Profit   Quantity
Amount   1.0000000 0.30924224 0.35248611
Profit   0.3092422 1.00000000 0.06309573
Quantity 0.3524861 0.06309573 1.00000000
> # Plot the graph of the correlation matrix using corrplot
> # Install the corrplot package if not already installed
> # install.packages("corrplot")
> library(corrplot)
corrplot 0.92 loaded
> # Plotting the correlation matrix
> corrplot(correlation_matrix, method = "circle", type = "full", tl.col = "black", tl.srt = 45)
>
```



9.

boxplot(numeric_variables)



```
# Assuming 'Category' is a categorical variable
barplot(table(Sales.data$Category), col = "lightblue", main = "Bar Plot of Category")
```