EXPT NO:2	Implementation of data visualization techniques
DATE: 06.01.2026	

PRE-LAB QUESTIONS (PROVIDE BRIEF ANSWERS TO THE FOLLOWING QUESTIONS)

1. **Why is exploratory data analysis critical before model building?**

   Exploratory Data Analysis (EDA) helps understand the structure, patterns, and quality of data. It identifies missing values, outliers, and relationships between variables, ensuring the data is suitable for building accurate machine learning models.

2. **How do distributions influence algorithm selection in ML?**

   The distribution of data directly impacts how machine learning algorithms perform. Some algorithms assume normally distributed data, while others can handle skewed or non-linear distributions. By analyzing data distributions, appropriate algorithms and preprocessing techniques such as normalization or transformation can be selected for better results.

3. **What insights can outliers provide in business data?**

   Outliers in business data may represent unusual customer behavior, fraudulent transactions, or data entry errors. They can also indicate rare but valuable events, such as high-value purchases. Studying outliers helps businesses detect risks, understand exceptional cases, and make informed strategic decisions.

4. **Why are visual summaries preferred over raw tables?**

   Visual summaries convert large and complex datasets into simple graphical representations. They make it easier to identify trends, patterns, and anomalies quickly, which is difficult when analyzing raw numerical tables. Visualization also improves communication of insights to non-technical stakeholders.

5. **How does visualization improve business intelligence?**

   Visualization enhances business intelligence by presenting data insights in an intuitive and interactive manner. It enables faster decision-making, helps track performance metrics, identifies problem areas, and supports data-driven strategies by turning raw data into actionable insights.

IN-LAB EXERCISE:
OBJECTIVE:

To explore data distribution and variability using advanced visualization techniques.
SCENARIO:
A startup analyzes e-commerce transaction data to understand customer spending behavior and detect abnormal purchase patterns.

IN-LAB TASKS (Using R Language)

- Plot histogram of transaction amounts

- Use boxplot to detect outliers
- Create heatmap of monthly sales intensity

(CODE: CONATAINS STUDENT ROLL NOS
SCREENSHOT OF CODE
SCREENSHOT OF OUTPUT)

CODE WITH OUTPUT:

```
> head(df)
  Transaction_ID Customer_ID Transaction_Date Product_Category Transaction_Amount Payment_
Mode Region
1           5001        2049       2024-01-01            Books                917
Card  North
2           5002        2042       2024-01-02      Electronics               2982
Card   West
3           5003        2006       2024-01-03      Electronics               3777
UPI   East
4           5004        2015       2024-01-04      Electronics                343
UPI   East
5           5005        2043       2024-01-05             Home               3340   NetBan
king   East
6           5006        2037       2024-01-06             Home               4691
Card   East
> View(df)
>
> print("Roll No: 23BAD066")
[1] "Roll No: 23BAD066"
>
> # -------------------- Libraries --------------------
> library(ggplot2)
> library(dplyr)
> library(lubridate)
>
> # -------------------- Load Dataset --------------------
> df <- read.csv("D:/Downloads/2.ecommerce_transactions.csv")
```

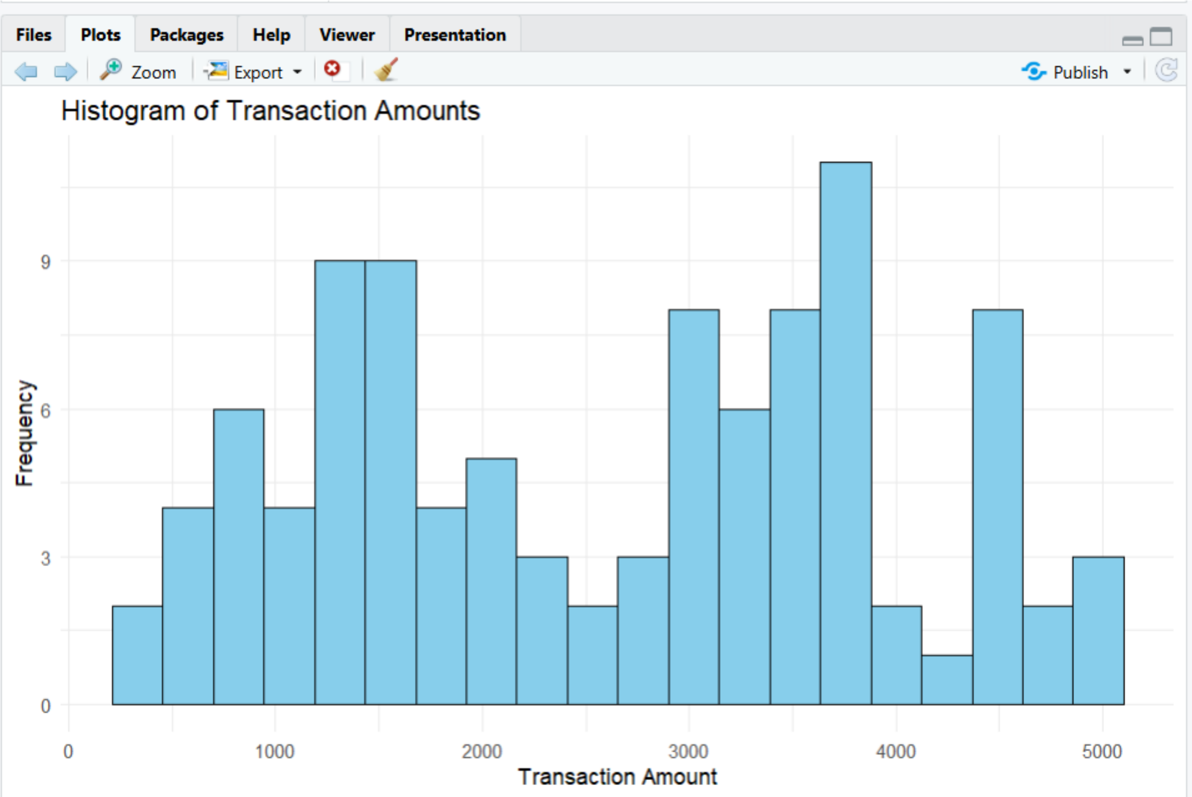
```
> # -------------------- Libraries --------------------
> library(ggplot2)
> library(dplyr)
> library(lubridate)
>
> # df is already loaded (so no need read.csv again)
>
> # -------------------- Convert Date Column --------------------
> df$Transaction_Date <- as.Date(df$Transaction_Date)
>
> # -------------------- Histogram --------------------
> ggplot(df, aes(x = Transaction_Amount)) +
+     geom_histogram(bins = 20, fill = "skyblue", color = "black") +
+     labs(
+         title = "Histogram of Transaction Amounts",
+         x = "Transaction Amount",
+         y = "Frequency"
+     ) +
+     theme_minimal()
>
> # -------------------- Boxplot --------------------
> ggplot(df, aes(y = Transaction_Amount)) +
+     geom_boxplot(fill = "lightgreen", color = "black") +
+     labs(
+         title = "Boxplot of Transaction Amounts",
+         y = "Transaction Amount"
+     ) +
+     theme_minimal()
>
> # -------------------- Heatmap Data --------------------
> heatmap_data <- df %>%
+     mutate(Month = month(Transaction_Date, label = TRUE, abbr = FALSE)) %>%
```

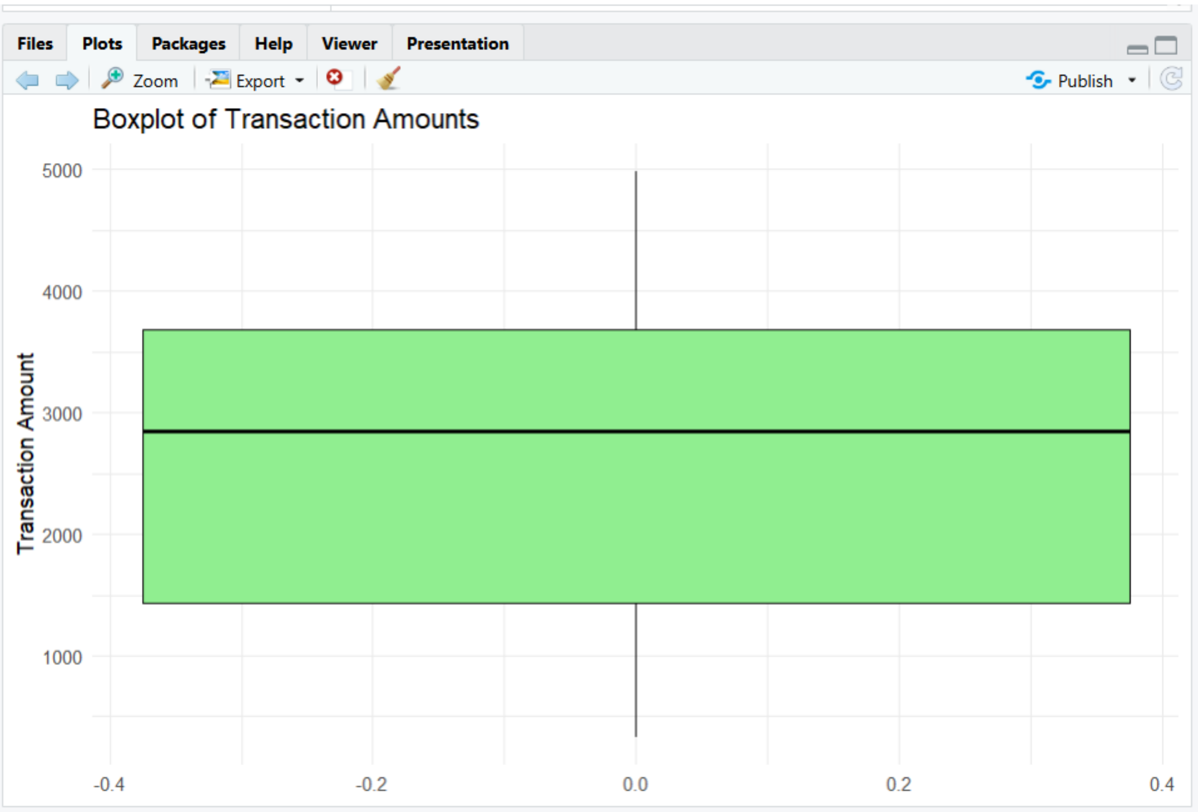
EDA_2.R × | df ×

Filter

	Transaction_ID	Customer_ID	Transaction_Date	Product_Category	Transaction_Amount	Payment_Mode	Region
1	5001	2049	2024-01-01	Books	917	Card	North
2	5002	2042	2024-01-02	Electronics	2982	Card	West
3	5003	2006	2024-01-03	Electronics	3777	UPI	East
4	5004	2015	2024-01-04	Electronics	343	UPI	East
5	5005	2043	2024-01-05	Home	3340	NetBanking	East
6	5006	2037	2024-01-06	Home	4691	Card	East
7	5007	2033	2024-01-07	Electronics	3138	NetBanking	North
8	5008	2008	2024-01-08	Books	3836	UPI	North
9	5009	2044	2024-01-09	Home	3422	UPI	South
10	5010	2044	2024-01-10	Electronics	1367	UPI	South
11	5011	2005	2024-01-11	Electronics	2816	UPI	East
12	5012	2039	2024-01-12	Books	1262	UPI	East
13	5013	2004	2024-01-13	Clothing	2356	UPI	East
14	5014	2006	2024-01-14	Home	3182	UPI	South
15	5015	2045	2024-01-15	Books	3494	Card	West
16	5016	2032	2024-01-16	Electronics	4929	NetBanking	West
17	5017	2030	2024-01-17	Electronics	1356	UPI	South
18	5018	2047	2024-01-18	Books	932	Card	South
19	5019	2035	2024-01-19	Home	4530	Card	South
20	5020	2040	2024-01-20	Clothing	1153	NetBanking	South
21	5021	2016	2024-01-21	Clothing	2739	UPI	South
22	5022	2013	2024-01-22	Electronics	3524	NetBanking	South
23	5023	2042	2024-01-23	Home	3501	UPI	East
24	5024	2030	2024-01-24	Clothing	1742	Card	East
25	5025	2019	2024-01-25	Home	1391	Card	East

```
> print(heatmap_data)
# A tibble: 15 × 3
   Product_Category Month    Total_Sales
   <chr>            <ord>          <int>
 1 Books            January        12181
 2 Books            February       18961
 3 Books            March          19270
 4 Books            April           9788
 5 Clothing         January        11993
 6 Clothing         February       17147
 7 Clothing         March           9282
 8 Electronics      January        30549
 9 Electronics      February       29506
10 Electronics      March          31781
11 Electronics      April           8256
12 Home             January        27250
13 Home             February        7334
14 Home             March          23616
15 Home             April           4564
>
> # -------------------- Heatmap Plot --------------------
> ggplot(heatmap_data, aes(x = Month, y = Product_Category, fill = Total_Sales)) +
+     geom_tile(color = "white") +
+     scale_fill_gradient(low = "lightyellow", high = "darkblue") +
+     labs(
+         title = "Heatmap of Monthly Sales Intensity",
+         x = "Month",
+         y = "Product Category",
+         fill = "Total Sales"
+     ) +
+     theme_minimal()
>
```

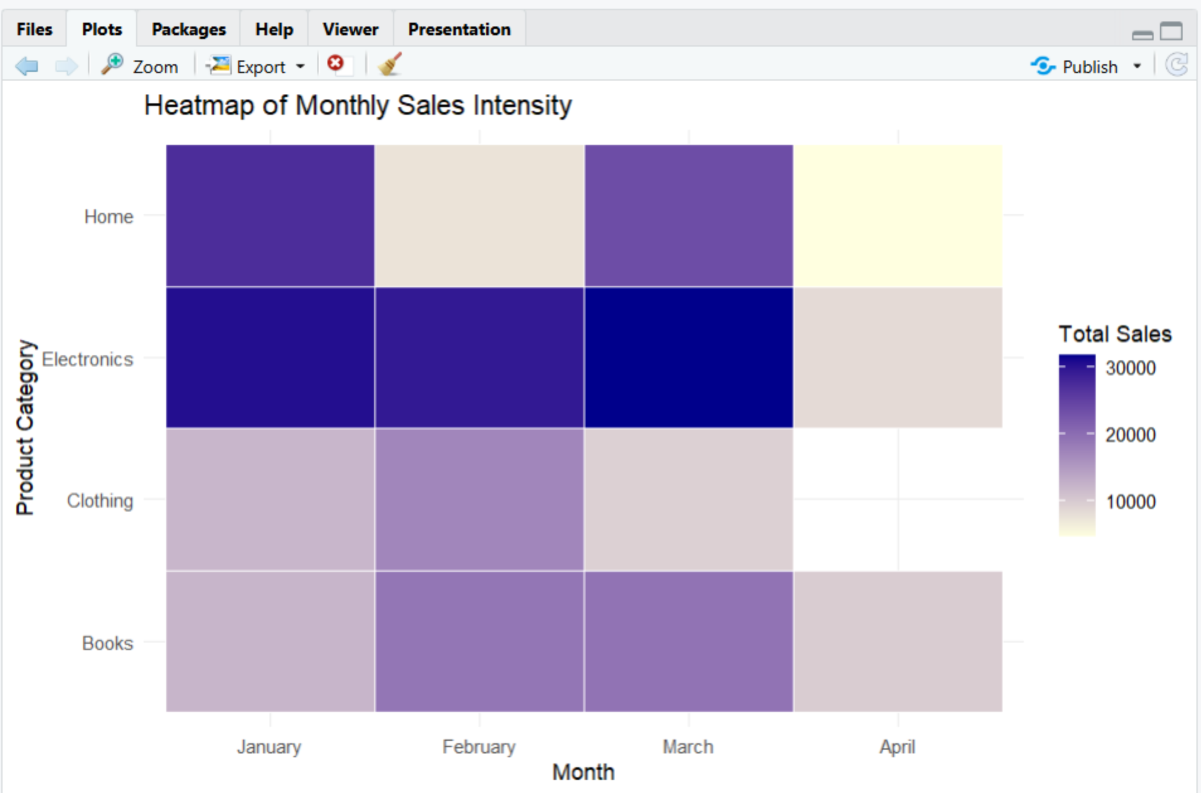




EDA_2.R × df × heatmap_data ×

Filter

	Product_Category	Month	Total_Sales
1	Books	January	12181
2	Books	February	18961
3	Books	March	19270
4	Books	April	9788
5	Clothing	January	11993
6	Clothing	February	17147
7	Clothing	March	9282
8	Electronics	January	30549
9	Electronics	February	29506
10	Electronics	March	31781
11	Electronics	April	8256
12	Home	January	27250
13	Home	February	7334
14	Home	March	23616
15	Home	April	4564



POST-LAB QUESTIONS (PROVIDE BRIEF ANSWERS TO THE FOLLOWING QUESTIONS)

1. **What does right-skewed distribution indicate about customer behavior?**
   A right-skewed distribution indicates that most customers make low to moderate value transactions, while a small number of customers make very high-value purchases. This suggests the presence of occasional high-spending customers.

2. **How can detected outliers impact business decisions?**
   Detected outliers may represent fraudulent transactions, data entry errors, or high-value customers. Identifying them helps businesses improve fraud detection, ensure data accuracy, and create targeted strategies for premium customers.

3. **Which visualization best supports anomaly detection?**

   A boxplot best supports anomaly detection because it visually represents the spread of data using the median and quartiles, making it easy to identify values that lie outside the normal transaction range. Points beyond the whiskers are clearly marked as outliers, helping to quickly detect unusual, extreme, or abnormal transactions that may indicate fraud, errors, or exceptional customer behavior.

4. **How does EDA improve AI model accuracy?**

EDA improves AI model accuracy by identifying data issues such as outliers, skewness, and inconsistencies. Proper cleaning and transformation based on EDA results lead to more reliable and accurate model predictions.

5. **How can visualization guide feature engineering?**
   Visualization helps identify important patterns and trends in data. These insights support the creation of meaningful features such as monthly spending, customer segmentation, and transaction frequency, improving model performance.

ASSESSMENT

Description	Max Marks	Marks Awarded
Pre Lab Exercise	**5**	
In Lab Exercise	**10**	
Post Lab Exercise	**5**	
Viva	**10**	
Total	**30**	
	Faculty Signature	