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 in understanding the overall structure and quality of the dataset before applying any machine learning model. It reveals missing values, outliers, noise, data imbalance, and relationships between variables. By identifying these issues early, appropriate data cleaning, transformation, and feature selection can be performed, which prevents poor model performance and incorrect predictions.
2. **How do distributions influence algorithm selection in ML?**
   Data distribution affects how well an algorithm performs because many ML algorithms rely on statistical assumptions. For example, linear regression works best when data follows a normal distribution, while skewed or non-linear data may reduce its accuracy. In such cases, data transformations (log, normalization) or algorithms like decision trees and random forests, which do not assume any distribution, are preferred.
3. **What insights can outliers provide in business data?**
   Outliers can represent unusual but important events in business data. They may indicate fraudulent transactions, system errors, or exceptional customer behavior such as high-value purchases. Analyzing outliers helps businesses detect risks, improve security, understand premium customers, and make better strategic decisions rather than blindly removing them.
4. **Why are visual summaries preferred over raw tables?**
   Visual summaries present large volumes of data in an easily understandable format. Patterns, trends, and anomalies that are hard to notice in raw tables become immediately visible through charts and plots. Visualization reduces cognitive load, saves time, and helps both technical and non-technical users interpret data effectively.
5. **How does visualization improve business intelligence?**
   Visualization transforms complex data into meaningful insights that support informed decision-making. It helps businesses monitor performance, identify trends, compare metrics, and detect problems quickly. By providing clear and interactive insights, visualization enhances strategic planning, operational efficiency, and overall business intelligence.

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)

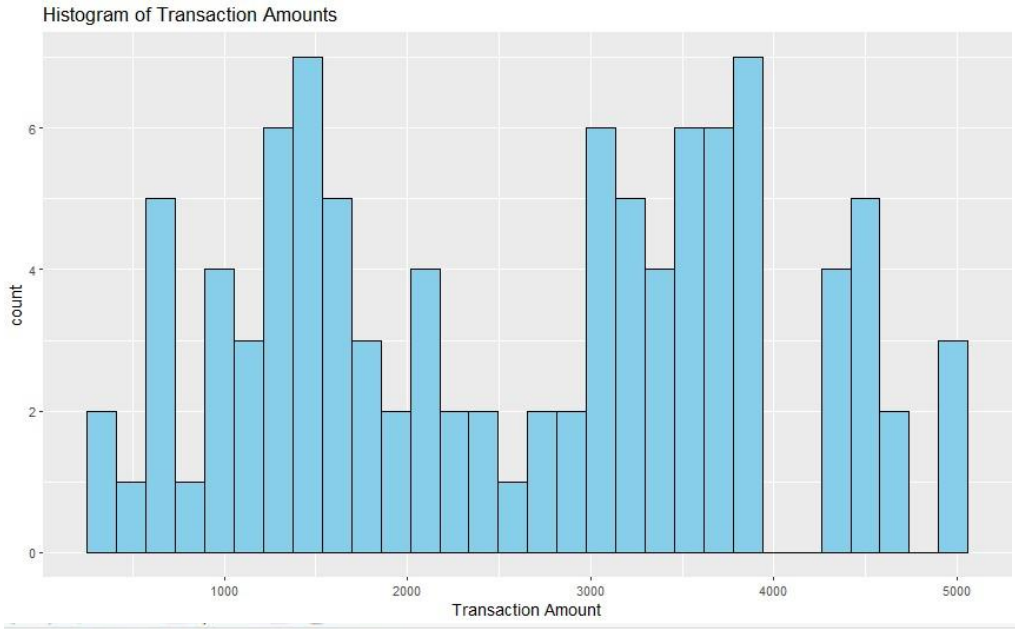
- Plot histogram of transaction amounts
- Use boxplot to detect outliers
- Create heatmap of monthly sales intensity

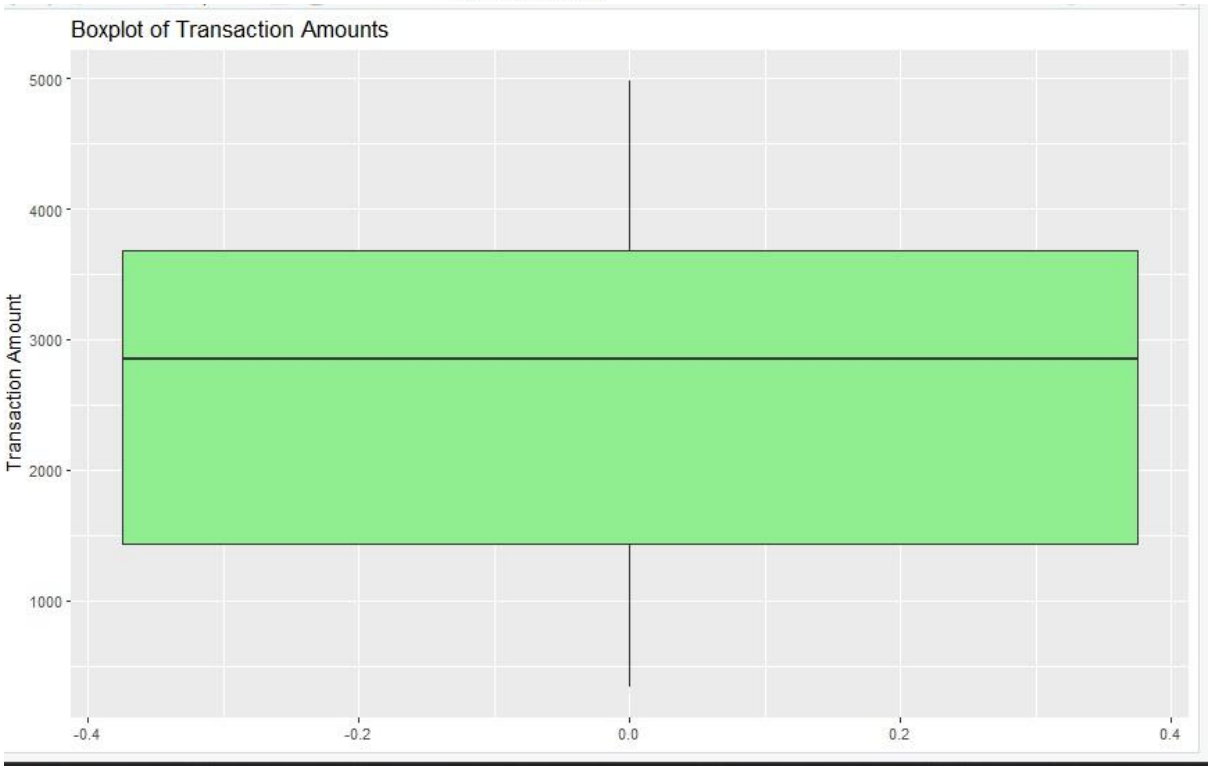
CODE:

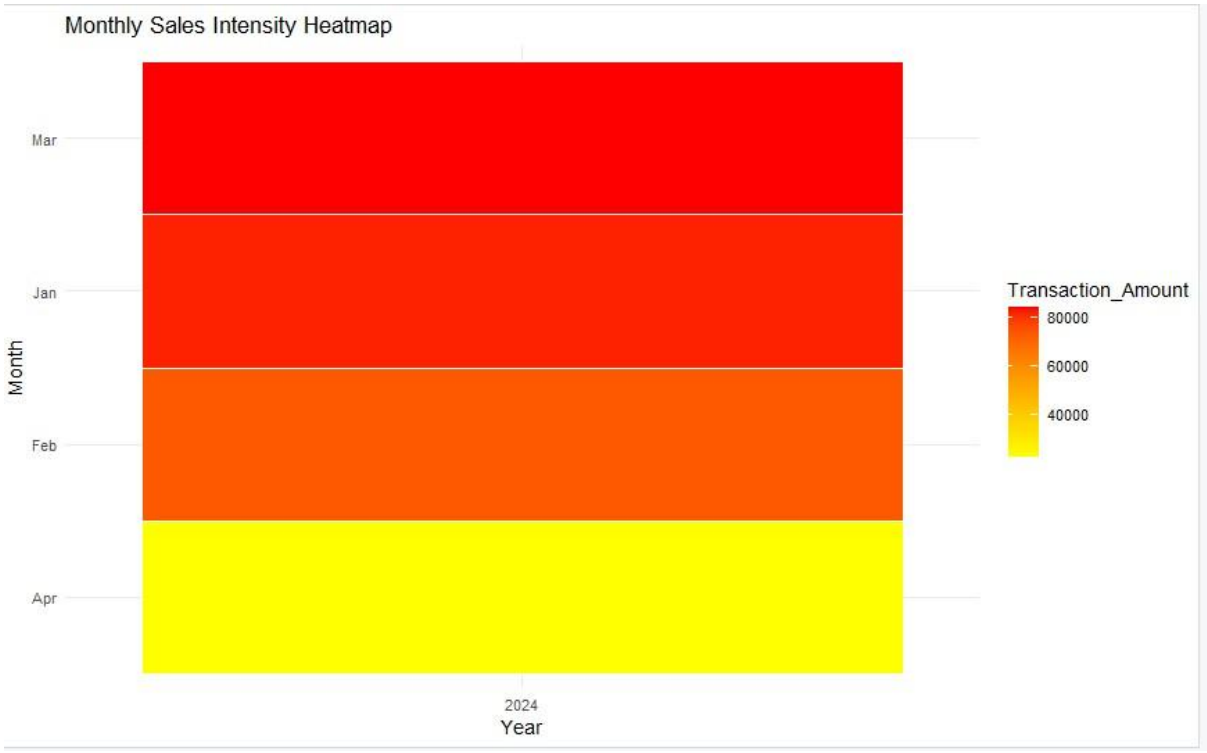
Untitled1* | Source on Save | Run

```r
# ============================================
# Roll No : 23BAD101
# ============================================

library(ggplot2)

# Load data
data <- x2_ecommerce_transactions

# Convert date
data$Transaction_Date <- as.Date(data$Transaction_Date)

# Extract month & year
data$month <- format(data$Transaction_Date, "%b")  # Jan, Feb...
data$year <- format(data$Transaction_Date, "%Y")

# ----------------------------
# Histogram
# ----------------------------
ggplot(data, aes(x = Transaction_Amount)) +
  geom_histogram(fill = "skyblue", color = "black", bins = 30) +
  labs(title = "Histogram of Transaction Amounts",
       x = "Transaction Amount")

# Boxplot for detecting outliers in transaction amount
boxplot(data$Transaction_Amount,
        main = "Boxplot of Transaction Amounts",
        ylab = "Transaction Amount",
        col = "lightgreen",
        outcol = "red",
        pch = 19)

# ----------------------------
# Heatmap (WORKS EVEN FOR 1 YEAR)
# ----------------------------
monthly_sales <- aggregate(Transaction_Amount ~ month + year,
                           data = data,
                           sum)

ggplot(monthly_sales,
       aes(x = year, y = month, fill = Transaction_Amount)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "yellow", high = "red") +
  labs(title = "Monthly Sales Intensity Heatmap",
       x = "Year",
       y = "Month") +
  theme_minimal()
```

OUTPUT:


Histogram of Transaction Amounts
count
Transaction Amount
0
2
4
6
1000
2000
3000
4000
5000


Boxplot of Transaction Amounts
Transaction Amount
1000
2000
3000
4000
5000
-0.4
-0.2
0.0
0.2
0.4



POST-LAB QUESTIONS (PROVIDE BRIEF ANSWERS TO THE FOLLOWING QUESTIONS) 1
What does a right-skewed distribution indicate about customer behavior?
Most customers spend low amounts, while a few customers make very high-value purchases.

2 How can detected outliers impact business decisions?
They help identify fraud, exceptional customers, pricing issues, or data errors, influencing risk control and strategy.

3 Which visualization best supports anomaly detection?
Boxplot is most effective for detecting anomalies and outliers.

4 How does EDA improve AI model accuracy?
It reveals data issues, feature relationships, and patterns, leading to better preprocessing and model selection.

5 How can visualization guide feature engineering?
Visuals show correlations, trends, and distributions, helping decide feature transformations, scaling, or removal.

ASSESSMENT

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