|  |  |
|---|---|
| **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 is critical because it helps us to identify missing values, detect anomalies, and understand the underlying structure of the data before we feed it into an algorithm. If EDA is skipped, a model learns from noise rather than actual patterns.

2. How do distributions influence algorithm selection in ML?
   The shape of the data tells what algorithm should be used. Many algorithms like Linear Regression or Gaussian Naive Bayes assume data follows a normal (bell curve) distribution. If the data is highly skewed or non-linear, we might need to choose non-parametric algorithms (like Decision Trees) or apply transformations to the data first.

3. What insights can outliers provide in business data?
   Outliers highlight deviations from normal behavior, which helps businesses detect critical anomalies like fraudulent transactions, operational errors, or unique opportunities like high-value customer spending that require special attention.

4. Why are visual summaries preferred over raw tables?
   Visual summaries are preferred because they allow humans to instantly recognize trends, patterns, and relationships like seasonality or correlations that are difficult and time-consuming to spot when scanning rows of raw numbers and tables.

5. How does visualization improve business intelligence?
   Visualization turns static metrics into actionable stories. It improves business intelligence by allowing non-technical decision-makers to grasp complex insights quickly. This leads to faster decision-making, such as identifying a sudden drop in sales or spotting a rising product category in real-time.

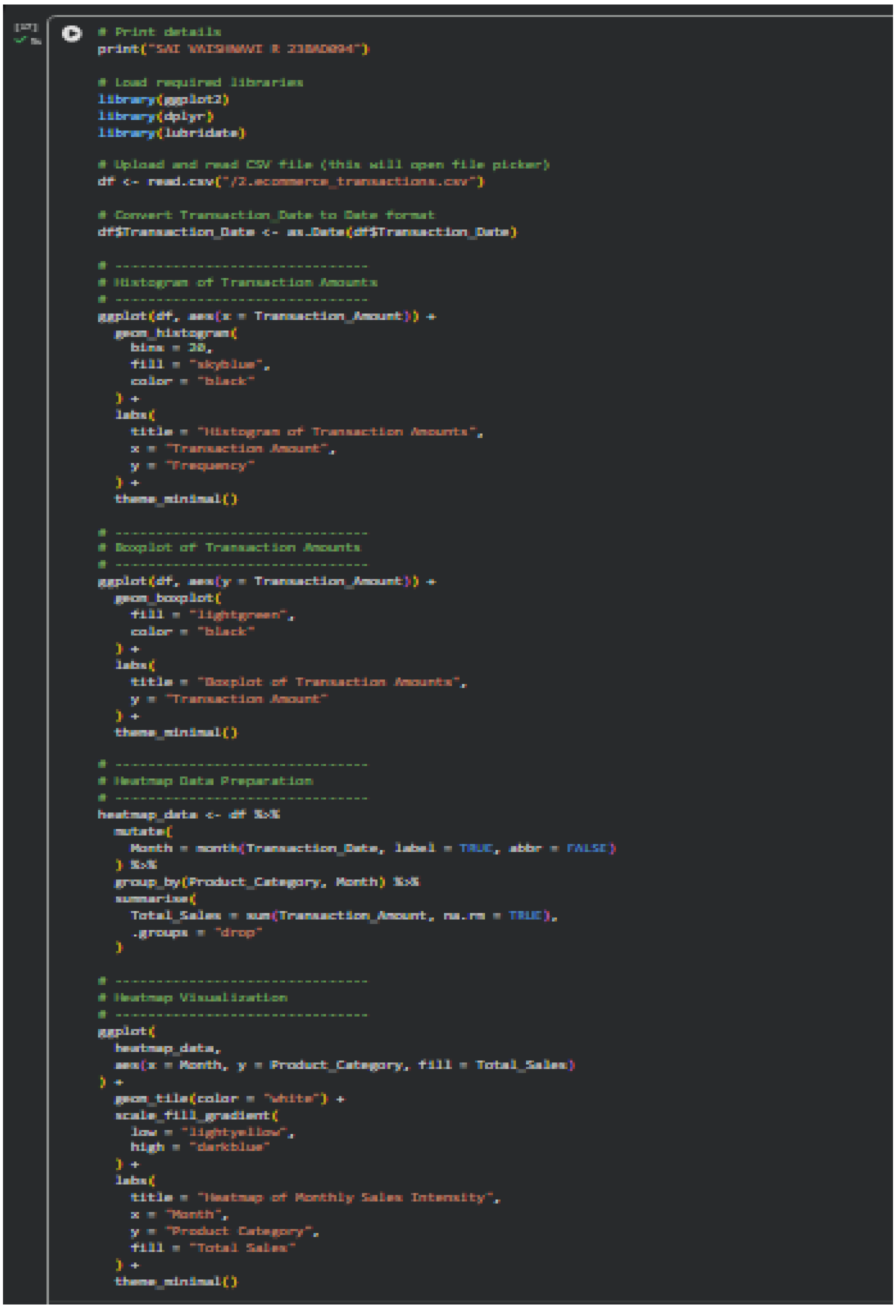
**IN-LAB EXERCISE:**
**OBJECTIVE:**

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)**

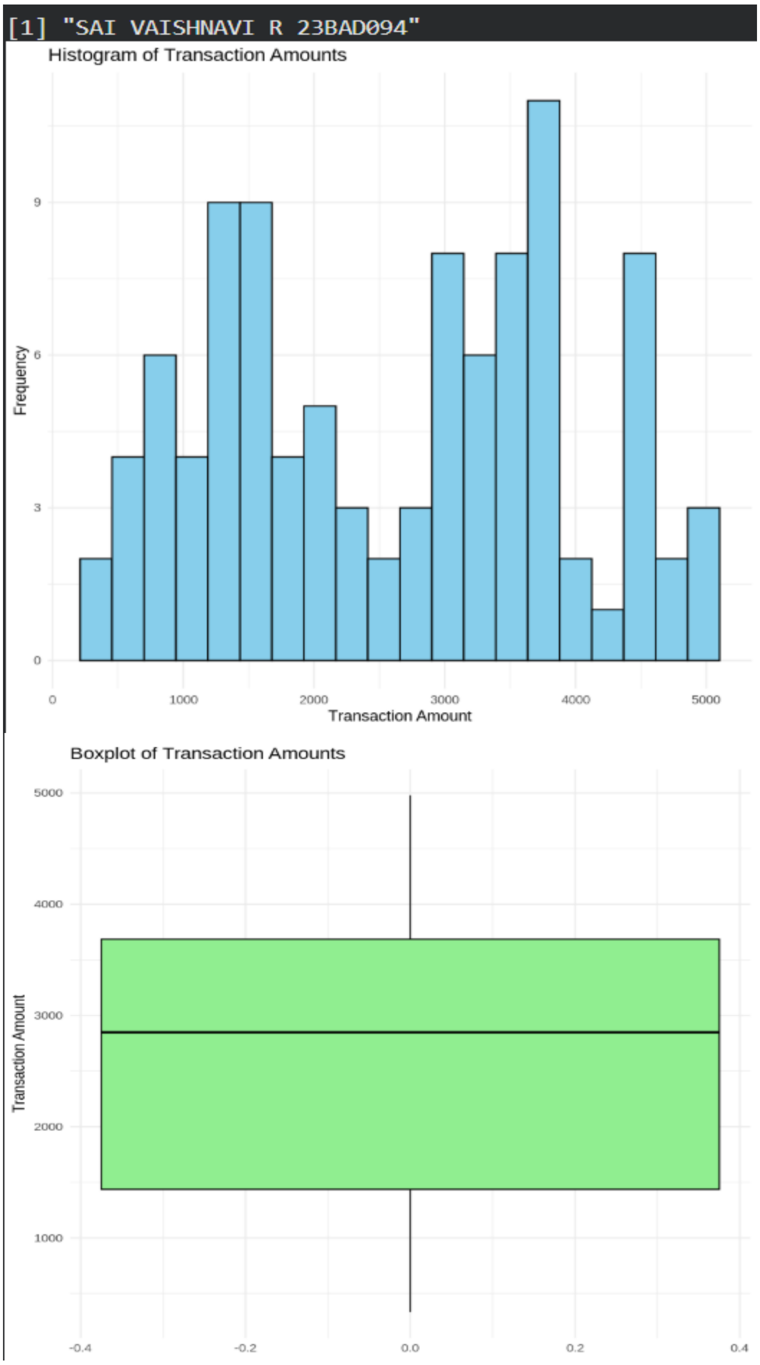
- Plot histogram of transaction amounts
- Use boxplot to detect outliers
- Create heatmap of monthly sales intensity

**CODE:**

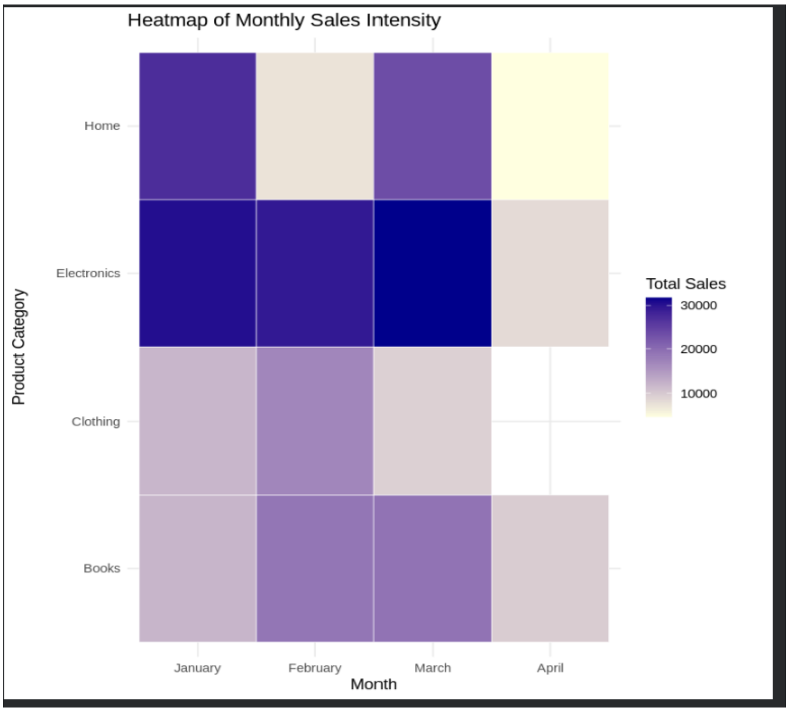


```r
# Print details
print("SAI VAISHNAVI R 23BAD094")

# Load required libraries
library(ggplot2)
library(dplyr)
library(lubridate)

# Upload and read CSV file (this will open file picker)
df <- read.csv("/2.ecommerce_transactions.csv")

# Convert Transaction_Date to Date format
df$Transaction_Date <- as.Date(df$Transaction_Date)

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

# --------------------------------
# Boxplot of Transaction Amounts
# --------------------------------
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 Preparation
# --------------------------------
heatmap_data <- df %>%
  mutate(
    Month = month(Transaction_Date, label = TRUE, abbr = FALSE)
  ) %>%
  group_by(Product_Category, Month) %>%
  summarise(
    Total_Sales = sum(Transaction_Amount, na.rm = TRUE),
    .groups = "drop"
  )

# --------------------------------
# Heatmap Visualization
# --------------------------------
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()
```

**OUTPUT:**



```
[1] "SAI VAISHNAVI R 23BAD094"
```


Heatmap of Monthly Sales Intensity
Product Category
Home
Electronics
Clothing
Books
January
February
March
April
Month
Total Sales
30000
20000
10000

## 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 the majority of customers make small to medium-sized purchases, while only a very small number of customers make extremely high-value purchases.

2. How can detected outliers impact business decisions?
   Outliers force businesses to decide whether to exclude or exploit extreme data. Ignoring them prevents skewed forecasts like average sales, while analyzing them can uncover high-revenue markets or identify operational failures (like fraud) that require immediate policy changes.

3. Which visualization best supports anomaly detection?
   The Boxplot is the best tool for anomaly detection. It statistically defines "normal" data within the Interquartile Range (IQR) and visually isolates anomalies as individual dots outside the whiskers, making them impossible to miss.
4. How does EDA improve AI model accuracy?
   EDA ensures that the model is not trained on misleading data. By revealing issues like class imbalance or multicollinearity beforehand, analysts can choose the correct algorithms and clean the data, preventing the model from learning false patterns.

5. How can visualization guide feature engineering?
   Visualizations highlight complex shapes and trends that raw numbers hide. For example, seeing a curved relationship in a scatter plot suggests creating a polynomial feature, or seeing distinct clusters suggests creating a new categorical label, which makes the data easier for the model to understand. This is how visualization guide feature engineering.

### ASSESSMENT

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