CERTIFIED DATA ANALYSTS

ASSIGNMENT 5 :
Customer Purchase Analysis using R

NAZREEN AGOS BIN ABDUL LATIFF

1. Data Import and Cleaning

• The dataset (CSV format) load into R.

df <- read.csv("~/Data Analytic Study UMK/Assignment 5 (R Studio)/Customer Purchase Data.csv")

• Handle any missing values appropriately.

```
> anyNA(df) # Check presence of missing values
[1] FALSE
```

df = na.omit(df) # Remove rows with missing values (if any)

Ensure categorical column as character data type

Character : When you just want to treat text as plain strings (for printing, labeling, or temporary cleaning)
Factor : When the variable has a fixed set of categories (for modeling, grouping, plotting, statistical summaries)

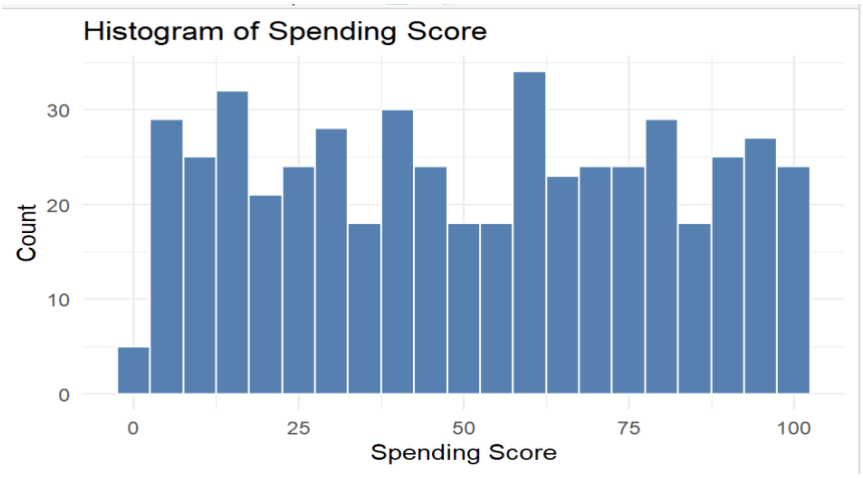
df$Gender <- as.factor(df$Gender)
df$PurchaseCategory <- as.factor(df$PurchaseCategory)

2. Exploratory Data Analysis

The summary statistics for numerical columns.

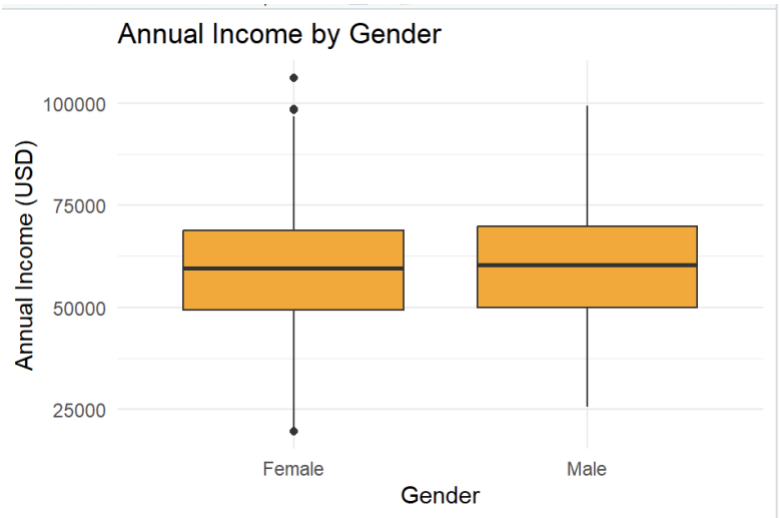
```
> summary(df)
  CustomerID          Gender         Age        AnnualIncome
 Length:500         Female:239   Min.   :18.00   Length:500
 Class :character   Male  :261   1st Qu.:30.00   Class :character
 Mode  :character                Median :42.00   Mode  :character
                                 Mean   :41.28
                                 3rd Qu.:52.00
                                 Max.   :64.00
 SpendingScore          PurchaseCategory TransactionDate
 Min.   :  1.00   Books         :97      Length:500
 1st Qu.: 25.00   Clothing      :74      Class :character
 Median : 50.50   Electronics   :79      Mode  :character
 Mean   : 51.23   Groceries     :89
 3rd Qu.: 77.00   Home & Kitchen:82
 Max.   :100.00   Sports        :79
```

The histogram of SpendingScore.



- The histogram shows that customers are fairly **evenly spread across the score range**.
- There is no single dominant peak (it is **not strongly skewed left or right**).
- Several bins have slightly more customers (scores near **10, 20, 40, 60, 80**).
- This suggests that customer spending behavior is **diverse**, some are low spenders, others moderate or high.

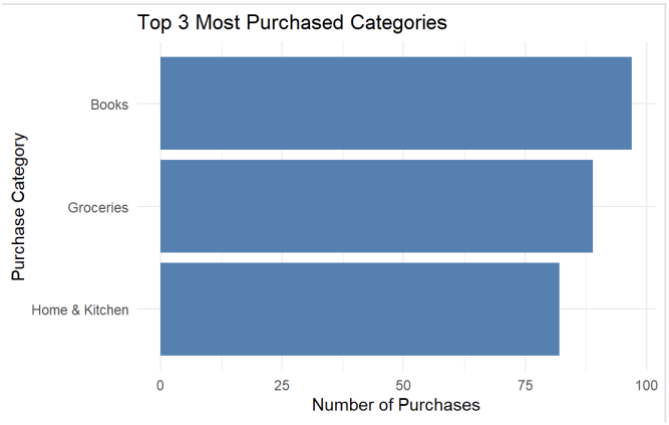
The boxplot of AnnualIncome by Gender.



Boxplot analysis

- **No major difference** is observed in income between Male and Female customers.
- Median and range are **very similar**.
- The outliers in Female group **slightly increase the spread**, but do not significantly shift the center.

Top 3 most purchased categories.



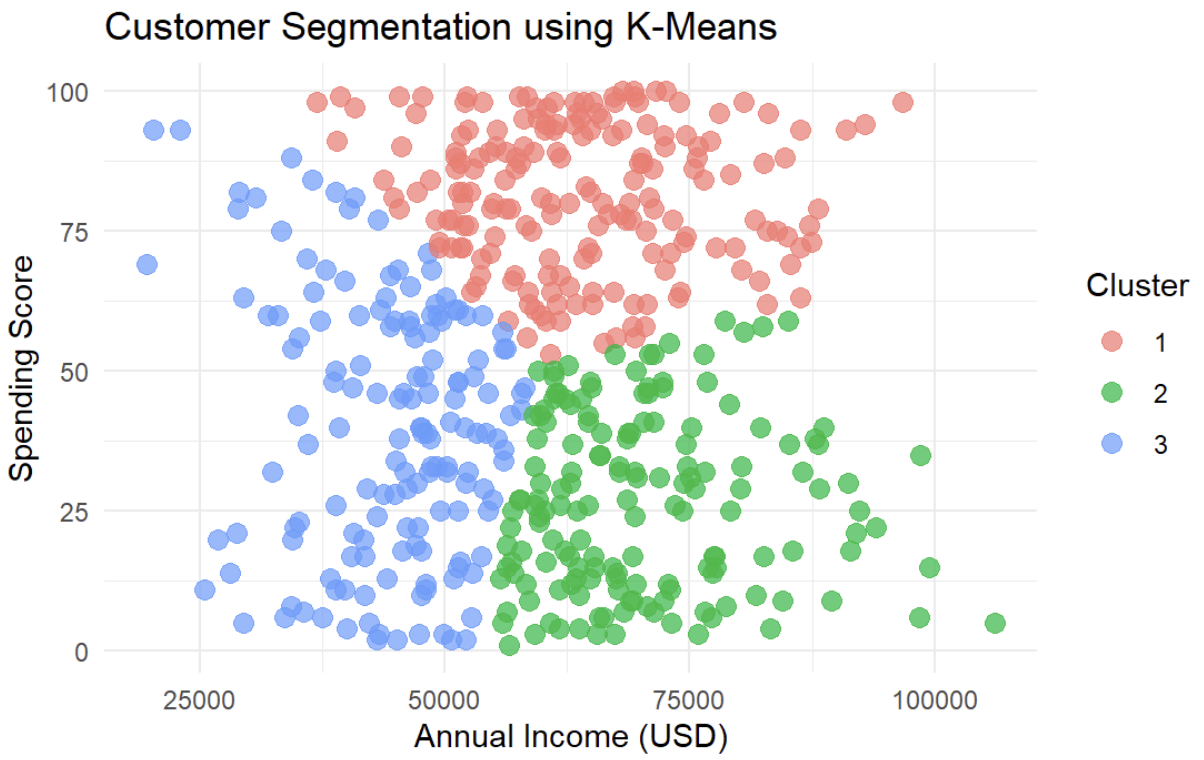
PurchaseCategory	`Number of Purchases`
<fct>	<int>
Books	97
Groceries	89
Home & Kitchen	82

1. **Books lead the list**, meaning customers may prioritize self-improvement, education, or leisure reading. This is a high-engagement category.
2. **Groceries**, a daily necessity, ranks second. Consistent with repeat, utility-driven shopping behavior.
3. **Home & Kitchen** shows solid traction, possibly driven by remote working, home cooking, or domestic upgrades.

3. Customer Segmentation (Clustering)

To segment customers based on their **Annual Income** and **Spending Score**, K-means clustering was applied in R using the following steps:

```
# Ensure AnnualIncome and SpendingScore are numeric
str(df$AnnualIncome)
str(df$SpendingScore)

# Create a new dataframe with relevant columns
cluster_data <- df %>%
  select(AnnualIncome, SpendingScore)

# Remove any missing values (just in case)
cluster_data <- na.omit(cluster_data)

#Scale the data (standardize for fair distance comparison)
cluster_scaled <- scale(cluster_data)

#Apply K-means with 3 clusters
set.seed(123)  # for reproducibility
kmeans_result <- kmeans(cluster_scaled, centers = 3, nstart = 25)

#Add cluster labels back to original data
df$Cluster <- as.factor(kmeans_result$cluster)
```



Cluster 1 (Red) — High Spending, Mid-to-High Income

- These customers spend a lot despite varying income levels.
- They are ideal **target customers**, loyal, profitable, and willing to buy.
- Likely interested in premium offers, loyalty programs, early access.

Cluster 2 (Green) — Low Spending, High Income

- These customers earn a lot but do not spend much.
- May need personalized marketing or more engagement.
- Consider **upselling**, premium offers, or identifying blockers (trust, product relevance).

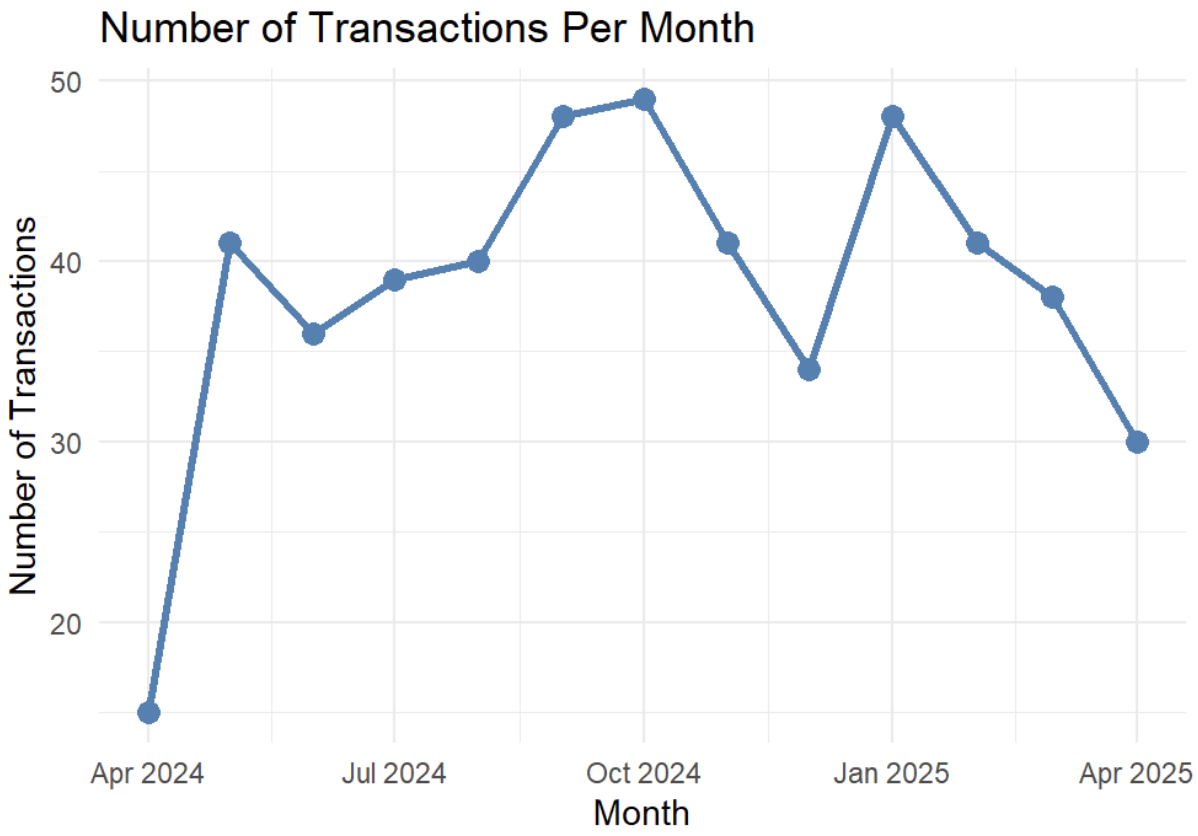
Cluster 3 (Blue) — Mixed Spending, Low to Mid Income

- Includes low-to-mid income shoppers with moderate-to-low spending.
- Possibly price-sensitive.
- Good candidates for **discounts**, **bundles**, or **volume-based promotions**.

4. Time Series Insights

```
#To convert the TransactionDate column into proper Date format,
#I can use the lubridate package, especially if the format is in dd/mm/yyyy,
#which is common in Malaysian datasets.

install.packages("lubridate")  # Run once
library(lubridate)

df$TransactionDate <- dmy(df$TransactionDate)

str(df$TransactionDate) # to confirm it works
```

```
> str(df$TransactionDate)
 Date[1:500], format: "2025-04-21" "2024-04-25" "2024-12-27" "2024-05-27" "2024-08-11" ...
```



The number of customer transactions increased steadily from **April to October 2024**, peaking in **October**, likely due to seasonal campaigns or product launches. A dip followed in **November and December**, before rebounding strongly in **January 2025**. However, a consistent decline occurred from **February to April 2025**, suggesting the need for reactivation strategies or new engagement initiatives during the **early part of the year**.

5. Statistical Analysis

T-Test Interpretation: Spending Score by Gender

A Welch Two-Sample **t-test** was conducted to determine whether there is a significant difference in the **average Spending Score** between **male and female customers**.

```
# Check that Gender is a factor and SpendingScore is numeric

str(df$Gender)
str(df$SpendingScore)

df$SpendingScore <- as.numeric(df$SpendingScore)

t_test_result <- t.test(SpendingScore ~ Gender, data = df)
print(t_test_result)
```

```
        Welch Two Sample t-test

data:  SpendingScore by Gender
t = -0.50778, df = 494.12, p-value = 0.6118
alternative hypothesis: true difference in means between group Female and group Male is not equal to
0
95 percent confidence interval:
 -6.535596  3.851231
sample estimates:
mean in group Female   mean in group Male
            50.53138             51.87356
```

The p-value of **0.6118** is **greater than 0.05**, indicating that the difference in Spending Score between male and female customers is **not statistically significant**. Additionally, the confidence interval includes zero, which further supports the conclusion that the difference may be due to random variation.

Conclusion, there is **no statistically significant difference** in the average Spending Score between male and female customers. Although the mean score for male customers (51.87) is slightly higher than that of female customers (50.53), this difference is not large enough to be considered meaningful from a statistical standpoint (p = 0.6118).