Dataset being using:
https://www.kaggle.com/datasets/hasaanrana/online-shopping-dataset

Objectives and Questions

1. Customer Demographics

Objective: Understand customer demographics and segment users to target the right audience with tailored strategies.

Questions

- What is the gender distribution of our customers?
- How does the frequency of online shopping vary by gender or other demographics?
- Are there any demographic groups (e.g., by gender, age, or location) that prefer certain online retailers or products more than others?
- How can we segment our customers based on their shopping behavior (e.g., frequency, preferred products)?

2. Customer Shopping Behavior

Objective: Analyze how often customers shop, what they buy, and the factors influencing their decisions.

Questions

- What motivates customers to shop online (e.g., convenience, discounts, product variety)?
- What are the most frequent reasons for customers to make online purchases? (e.g., discounts, ease of shopping, product availability)
- Are customers shopping online more frequently, and what could increase the frequency of their purchases?
- What products or categories are customers buying the most (e.g., fashion, electronics, books)?

3. Customer Experience Improvement

Objective: Enhance the overall customer experience across all touchpoints.

Questions:

- What are the main pain points that need to be addressed in the online shopping journey (e.g., product availability, delivery issues)?
- How can we improve the customer experience to increase customer satisfaction and reduce complaints or returns?
- What can we do to improve customer service and ensure smooth communication with shoppers?

Cleaning

```
> # Checking the unique values in the columns
> unique(Online_Shopping_Data$Gender)
[1] "Male"            "Female"          "12/3/2023 21:50" "Other"
> unique(Online_Shopping_Data$Online_Shopping_Freq)
[1] "Rarely or Never"         "Once in a month"
[3] "Once in a week"          "Multiple times per week"
[5] "12/4/2023 21:50" # These highlighted date values don't belong under "Gender" and
"Online_Shopping_Freq" columns
> # Changing incorrect values to NA
> library(dplyr)
> library(stringr)
> Online_Shopping_Data <- Online_Shopping_Data %>%
+    mutate(Gender = ifelse(str_trim(Gender) == "12/3/2023 21:50", NA, Gender),
> > Online_Shopping_Data <- Online_Shopping_Data %>% mutate(Online_Shopping_Freq =
ifelse(str_trim(Online_Shopping_Freq) == "12/4/2023 21:50", NA, Online_Shopping_Freq))
Online_Shopping_Data$Gender <- trimws(Online_Shopping_Data$Gender)
> # Removing leading and trailing white spaces
> Online_Shopping_Data[] <- lapply(Online_Shopping_Data, function(x) {
+    if (is.character(x)) trimws(x) else x
+ })
> # Apply function to capitalize the first letter of each string in character columns
> Online_Shopping_Data[] <- lapply(Online_Shopping_Data, function(x) {
+    if (is.character(x)) {
+        # Capitalize first letter of each value, keep the rest in lower case
+        return(sapply(x, function(val) paste0(toupper(substr(val, 1, 1)),
tolower(substr(val, 2, nchar(val))))))
+    } else {
+        return(x)
```

```
+     }
+ })
```

Exploratory Data Analysis

```
# View Summary of data table
> summary(Online_Shopping_Data)# See summary data below
# View frequency counts for each column
> lapply(Online_Shopping_Data, function(x) table(x))
```

Summary Data

Column Name	Length	Class	Mode
Gender	201	character	character
Online_Shopping_Freq	201	character	character
Online_Purchase_Proportion	201	character	character
Review_Check_Freq	201	character	character
Attraction_Factor	201	character	character
Retailer_Choice_Factors	201	character	character
Preferred_Payment	201	character	character
Local_vs_Intl_Retailers	201	character	character
Preferred_Marketplace	201	character	character
Security_Concern_Level	201	character	character
Promo_Participation	201	character	character
Price_Sensitivity	201	character	character
Comfortable_Price_Range	201	character	character
Frequent_Products	201	character	character
Major_Drawback	201	character	character
Authenticity_Concern	201	character	character
Desired_Improvements	201	character	character

Frequency Counts For Each Column

Most frequent value
Second most frequent value

Gender	Female: 125 (62%) Male: 74 (37%) Other: 1 (1%)
Online Shopping Frequency	Multiple times per week: 11 Once in a month: 99 (49%) Once in a week: 18 Rarely or never: 72 (36%)

Online Purchase Proportion	0-20%: 92 (46%) 21-40%: 39 41-60%: 40 (20%) 61-80%: 24 81-100%: 6
Review Check Frequency	Always: 123 (61%) Frequently: 37 (18%) Occasionally: 24 Rarely or never: 17
Attraction Factor	Attractive discounts and promotions: 36 Ease and comfort of shopping from home: 98 (49%) Time-saving: 22 User-friendly website/app interface: 7 Wide variety of products: 38 (19%)
Retailer Choice Factors	Brand reputation and trustworthiness: 53 (26%) Customer service and support: 14 Price and discounts: 49 Product reviews and ratings: 85 (83%)
Preferred Payment	Bank transfers or direct deposits: 3 Cash on delivery (if available): 167 (83%) Credit or debit card: 25 (12%) Other digital wallets: 6
Local vs Intl Retailers	No preference, depends on the product: 150(75%) Prefer international retailers: 19 (16%) Prefer local retailers: 32
Preferred Marketplace	Aliexpress: 13 Amazon: 19 Daraz: 99 (49%) Others: 70 (35%)
Security Concern Level	Neutral: 38 Not concerned at all: 7 Somewhat concerned: 52 (26%) Very concerned: 104 (52%)
Promo Participation	Always take advantage of promotions and discounts: 28 Frequently participate: 42 Occasionally participate: 67 (33%) Rarely or never participate: 64 (32%)
Price Sensitivity	Extremely price-sensitive: 43 (21%) Moderately price-sensitive: 114 (57%) Not price-sensitive at all: 10 Slightly price-sensitive: 34

Comfortable Price Range	1k to 5k: 124 (62%) 5k to 10k: 42 (21%) Less than 1k: 25 More than 10k: 10
Frequent Products	Beauty and personal care: 22 Books and media (e.g., movies, music): 16 Clothing and fashion accessories: 114 (57%) Electronics and gadgets: 39 (19%) Foods: 1 Fridge magnets: 1 Gadgets and clothing: 1 Home appliances and furniture: 2 Jewellery and stationery items: 1 Jo bhi dil mei ayega wahi lunga: 1 Shoes: 1 Sports: 1
Major Drawback	Beats ke headphone ki jagha "the bluetooth device is ready to pair" walay headphone de diya: 1 No, I have not experienced any drawbacks: 47 (23%) Shoes: 1 Yes, delayed or problematic deliveries: 28 Yes, difficulty with returns or refunds: 23 Yes, issues with product quality: 101 (50%)
Authenticity Concern	Mostly confident, with occasional concerns: 82 (41%) Neutral or unsure: 63 (31%) No, often concerned about product authenticity and quality: 19 Yes, always confident in product authenticity and quality: 37
Desired Improvements	Better product descriptions and images: 92 Enhanced customer reviews and ratings system: 33 (16%) Improved search and filtering options: 11 More personalized recommendations: 14 (Additional other values as required)

Frequency Count Insights:

- A majority of respondent are female (125)
- Most respondents shop online once a month (99), followed by those who shop rarely or never (72). Only a few people shop online multiple time per week (11)
- Most respondents purchase 0-20% of their items online (92), followed by those purchasing in the 21-40% range

- The majority of respondents always check reviews, which suggests that product reviews play a significant role in decision-making. A smaller number check reviews less frequently or not at all
- The biggest factor for online shopping is the ease and comfort of shopping from home, followed by attractive discounts and promotions. However, a majority of respondents rarely or never participate in promotions. This suggests that there is a strong interest in promotions but participation is not consistent.
- The most influential factors for choosing a retailer are product reviews (we saw under review check frequency that a majority of respondents always check reviews). Based on this information, brands should focus on building trust through positive product reviews and ratings.
- Cash on delivery is the most popular payment method – explore why cash on delivery is preferred and whether there is potential to push digital wallet option
- Most people do not have a clear preference and purchase from any retailer depending on the product. This shows flexibility in consumer preferences but also suggests that you don't necessarily need to focus on one over the other unless you have specific product lines that could benefit from either international or local offerings.
- Daraz is the most popular marketplace followed by Amazon and Aliexpress
- Most users are concerned about security. Invest in enhancing security measures and reassuring consumers about the safety of their transactions
- The majority of respondents are moderately price sensitive (price is an import factor in decision making, but not the only one)
- Clothing and fashion accessories are the most frequently purchased products
- The most common drawback is related to product quality (101), which suggests that improving the quality of your products may help reduce customer dissatisfaction
- The majority of users are either mostly confident or neutral regarding the authenticity of products
- Most users want better product descriptions and images

Running Analysis for Objective 1

```
> install.packages("tidyverse")
> # Filtering out rows where Gender is "NANA"
> cleaned_data <- Online_Shopping_Data %>%
+     filter(Gender != "NANA")
>
> # Gender distribution bar plot
> ggplot(cleaned_data, aes(x = Gender)) +
+     geom_bar(fill = "skyblue", color = "black") +
+     labs(title = "Gender Distribution of Online Shoppers", x = "Gender", y = "Count")
+
+     theme_minimal()
> # Filtering out rows where Gender is "NANA", "Other", or NA in Preferred_Marketplace
```

```
> cleaned_data <- Online_Shopping_Data %>%
+     filter(Gender != "NANA", Gender != "Other", !is.na(Preferred_Marketplace))
>
> # Creating a bar plot for Gender and Preferred Marketplace
> ggplot(cleaned_data, aes(x = Gender, fill = Preferred_Marketplace)) +
+     geom_bar(position = "fill") +  # "fill" makes it a stacked bar chart with
percentages
+     labs(title = "Gender vs Preferred Marketplace",
+          x = "Gender",
+          y = "Proportion",
+          fill = "Preferred Marketplace") +
+     theme_minimal() +
+     scale_y_continuous(labels = scales::percent)  # Show percentages on y-axis
```

OUTPUT: