Week 4 Assignment

Student Name

2024-02-04

# Introduction

In this assignment I choose to work with consumer behavior and shopping habits data set. The consumer behavior and shopping dataset consist of wide range variables including including demographic information, purchase history, product preferences, shopping frequency, and online/offline shopping behavior. This dataset is reach of collection of data that is can be used to provide comprehensive insights into consumers’ preferences, tendencies, and patterns during their shopping experiences.

# Variables Selection

The following qualitative and quantitative variables were selected for analysis

## Qualitative Variables

### • Gender (Male)

The customer's gender identity is a crucial demographic factor that affects their product choices and buying habits.

### • Subscription Status (Yes/No)

Provides information about the customer's degree of loyalty and possibilities for recurring revenue by indicating whether they have chosen to use a subscription service.

## Quantitative Variables

### • Age

The customer's age, which provides demographic data for marketing techniques that are targeted and segmented.

### • Purchase Amount (USD)

The cost of the item(s) purchased indicated by the transaction's monetary value, which is expressed in United States Dollars (USD).

# Analysis

#load libraries  
library("ggplot2")  
  
#read the dataset  
data<-read.csv('/home/addis/Downloads/shopping\_behavior\_updated.csv')  
#view the dataset  
head(data)

## Customer.ID Age Gender Item.Purchased Category Purchase.Amount..USD.  
## 1 1 55 Male Blouse Clothing 53  
## 2 2 19 Male Sweater Clothing 64  
## 3 3 50 Male Jeans Clothing 73  
## 4 4 21 Male Sandals Footwear 90  
## 5 5 45 Male Blouse Clothing 49  
## 6 6 46 Male Sneakers Footwear 20  
## Location Size Color Season Review.Rating Subscription.Status  
## 1 Kentucky L Gray Winter 3.1 Yes  
## 2 Maine L Maroon Winter 3.1 Yes  
## 3 Massachusetts S Maroon Spring 3.1 Yes  
## 4 Rhode Island M Maroon Spring 3.5 Yes  
## 5 Oregon M Turquoise Spring 2.7 Yes  
## 6 Wyoming M White Summer 2.9 Yes  
## Shipping.Type Discount.Applied Promo.Code.Used Previous.Purchases  
## 1 Express Yes Yes 14  
## 2 Express Yes Yes 2  
## 3 Free Shipping Yes Yes 23  
## 4 Next Day Air Yes Yes 49  
## 5 Free Shipping Yes Yes 31  
## 6 Standard Yes Yes 14  
## Payment.Method Frequency.of.Purchases  
## 1 Venmo Fortnightly  
## 2 Cash Fortnightly  
## 3 Credit Card Weekly  
## 4 PayPal Weekly  
## 5 PayPal Annually  
## 6 Venmo Weekly

### Create a contingency table for Gender and Subscription Status

contingency\_table <- table(data$Gender, data$Subscription.Status)  
# Print the contingency table  
print(contingency\_table)

##   
## No Yes  
## Female 1248 0  
## Male 1599 1053

### Calculate chi-square test

chi\_square\_result <- chisq.test(contingency\_table)  
# Print chi-square test results  
print(chi\_square\_result)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: contingency\_table  
## X-squared = 676.79, df = 1, p-value < 2.2e-16

### Calculate the correlation coefficient between Age and Purchase Amount

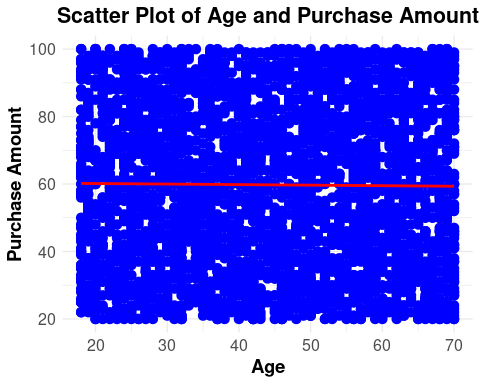
correlation\_coefficient <- cor(data$Age, data$Purchase.Amount..USD.)  
print(paste("Correlation Coefficient:", correlation\_coefficient))

## [1] "Correlation Coefficient: -0.0104236473786865"

### Create a scatter plot

ggplot(data, aes(x = Age, y = Purchase.Amount..USD.)) +  
 geom\_point(color = "blue", size = 3) +  
 labs(title = "Scatter Plot of Age and Purchase Amount",  
 x = "Age",  
 y = "Purchase Amount") +  
 theme\_minimal() +  
 theme(axis.text = element\_text(size = 12),  
 axis.title = element\_text(size = 14, face = "bold"),  
 plot.title = element\_text(size = 16, face = "bold", hjust = 0.5)) +  
 geom\_smooth(method = "lm", se = FALSE, color = "red") # Add a linear regression line for better insight

## `geom\_smooth()` using formula = 'y ~ x'



# Interpretation

## Qualitative Variables

(Gender and Subscription Status) The contingency table and chi-square test were conducted to explore the association between the qualitative variables - Gender and Subscription Status. The contingency table indicates that all females in the dataset have a subscription status of “No,” while there is a mix of subscription statuses among males, with both “Yes” and “No” categories present. The chi-square test yielded a highly significant p-value (p < 2.2e-16), suggesting a strong association between Gender and Subscription Status. This shows relationship implies that subscription preferences differ significantly between genders.

## Quantitative Variables

The scatter plot, correlation coefficient, and linear regression line were used to examine the relationship between the quantitative variables - Age and Purchase Amount. The scatter plot shows points scattered without a clear trend, and the correlation coefficient is close to zero (approximately -0.0104), indicating a very weak negative correlation. This implies that age and purchase amount are not strongly related; as one variable changes, the other does not show a consistent pattern of change

# Reflection

This reflection emphasizes the significance of understanding associations and correlations in data analysis, as these can guide further investigations and provide valuable insights into variable relationships. The analysis of qualitative variables, specifically Gender and Subscription Status reveals a significant association between Gender and Subscription Status, which is crucial for businesses and researchers to tailor strategies to specific demographic groups. This knowledge can be used for targeted marketing and personalized customer engagement, enabling informed decision-making.

The correlation between age and purchase amount is weak, suggesting age alone isn’t a strong predictor. This suggests a nuanced approach to customer segmentation, considering multiple factors influencing purchasing behavior. Understanding correlations helps avoid superficial relationships and prompts deeper investigation into the complex interplay of variables, revealing which factors, beyond age, significantly contribute to purchase amounts.