Assignment 1: Data Preprocessing and Feature Engineering

Shagufta Perveen

2025-04-155

## R Markdown

## 1. Dataset Selection

We use a **Customer Behavior dataset** from [Kaggle](https://www.kaggle.com/datasets). This dataset contains behavioral information about customers, which aligns with our research topic on customer segmentation and purchase prediction.

# Load the dataset  
data <- read.csv("C:/Users/Buzz Tech/Downloads/Customer\_Behaviour.csv")  
head(data)

## User.ID Gender Age EstimatedSalary Purchased  
## 1 15624510 Male 19 19000 0  
## 2 15810944 Male 35 20000 0  
## 3 15668575 Female 26 43000 0  
## 4 15603246 Female 27 57000 0  
## 5 15804002 Male 19 76000 0  
## 6 15728773 Male 27 58000 0

## 2. Exploratory Data Analysis (EDA)

### Descriptive Statistics

summary(data)

## User.ID Gender Age EstimatedSalary   
## Min. :15566689 Length:400 Min. :18.00 Min. : 15000   
## 1st Qu.:15626764 Class :character 1st Qu.:29.75 1st Qu.: 43000   
## Median :15694342 Mode :character Median :37.00 Median : 70000   
## Mean :15691540 Mean :37.66 Mean : 69743   
## 3rd Qu.:15750363 3rd Qu.:46.00 3rd Qu.: 88000   
## Max. :15815236 Max. :60.00 Max. :150000   
## Purchased   
## Min. :0.0000   
## 1st Qu.:0.0000   
## Median :0.0000   
## Mean :0.3575   
## 3rd Qu.:1.0000   
## Max. :1.0000

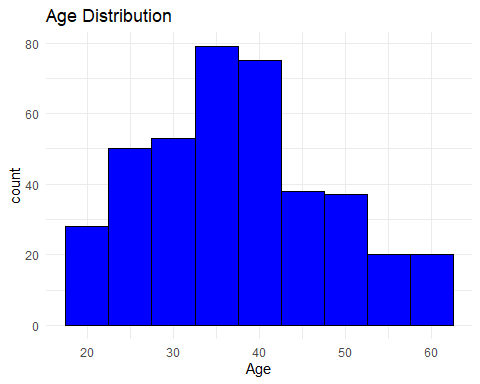
summary(data$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 29.75 37.00 37.66 46.00 60.00

### Visualizations

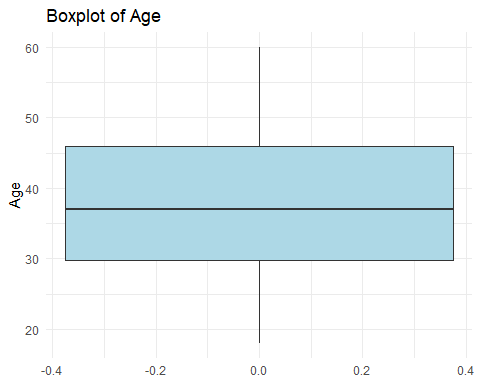
#### Histogram of Age

library(ggplot2)  
ggplot(data, aes(x = Age)) +  
 geom\_histogram(binwidth = 5, fill = "blue", color = "black") +  
 theme\_minimal() +  
 ggtitle("Age Distribution")



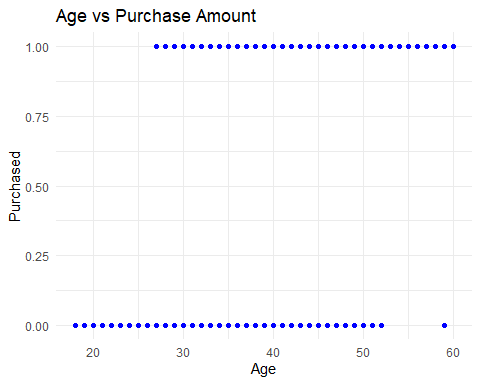
#### Boxplot of Age

ggplot(data, aes(y = Age)) +  
 geom\_boxplot(fill = "lightblue") +  
 theme\_minimal() +  
 ggtitle("Boxplot of Age")



#### Scatter Plot: Age vs Purchase

ggplot(data, aes(x = Age, y = Purchased)) +  
 geom\_point(color = "blue") +  
 theme\_minimal() +  
 ggtitle("Age vs Purchase Amount")



### Variable Descriptions

str(data)

## 'data.frame': 400 obs. of 5 variables:  
## $ User.ID : int 15624510 15810944 15668575 15603246 15804002 15728773 15598044 15694829 15600575 15727311 ...  
## $ Gender : chr "Male" "Male" "Female" "Female" ...  
## $ Age : int 19 35 26 27 19 27 27 32 25 35 ...  
## $ EstimatedSalary: int 19000 20000 43000 57000 76000 58000 84000 150000 33000 65000 ...  
## $ Purchased : int 0 0 0 0 0 0 0 1 0 0 ...

## 3. Data Preprocessing

### Handling Missing Data

# Remove rows with any missing values  
data\_clean <- na.omit(data)  
  
# Impute missing Age values with median  
data$Age[is.na(data$Age)] <- median(data$Age, na.rm = TRUE)

### Outlier Detection and Removal

Q1 <- quantile(data$Age, 0.25)  
Q3 <- quantile(data$Age, 0.75)  
IQR\_value <- Q3 - Q1  
  
# Filter out outliers  
data\_clean <- data[data$Age >= (Q1 - 1.5 \* IQR\_value) & data$Age <= (Q3 + 1.5 \* IQR\_value), ]

### Data Transformation

# Standardization  
data$Age\_standardized <- scale(data$Age)  
  
# Log transformation  
data$Purchased\_log <- log(data$Purchased + 1)

### Data Cleaning

# Standardize text in 'Gender'  
data$Gender <- tolower(data$Gender)

## 4. Feature Engineering

### Customer Lifetime Value (CLV)

# Example: CLV = Frequency \* Average Purchase Amount  
data$Frequency \* data$AvgPurchased

## integer(0)

head(data$CLV)

## NULL

### Recency Feature

# Recency in days from LastPurchaseDate to today  
 as.numeric(difftime(Sys.Date(), as.Date(data$LastPurchaseDate), units = "days"))

## numeric(0)

head(data$Recency)

## NULL

## 5. Conclusion

We have successfully preprocessed the Customer Behavior dataset by handling missing values, detecting and treating outliers, performing transformations, and engineering new features like CLV and Recency. These steps will improve model performance and support deeper insights in our research analysis.