ADS-500B Assignment 4

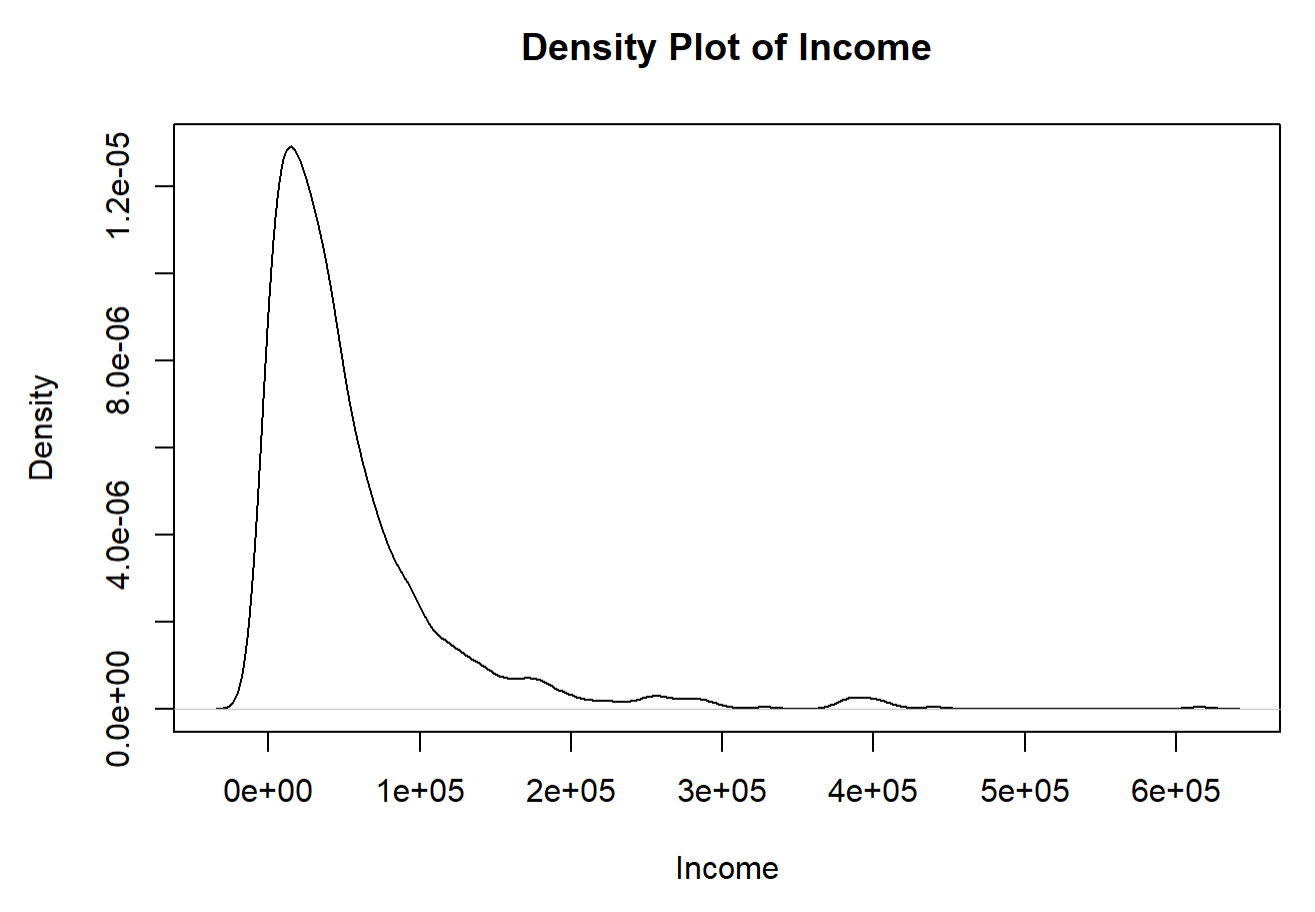
Gabriel Duffy

2024-03-25

```
# 1.) Define the numbers for multiplication
num1 <- 5
num2 <- 10

# Initialize the result variable
result <- 0

# Using a for loop for multiplication
for (i in 1:num2) {
    result <- result + num1
}

# Print the result
cat("Result of", num1, "x", num2, "is:", result, "\n")
```

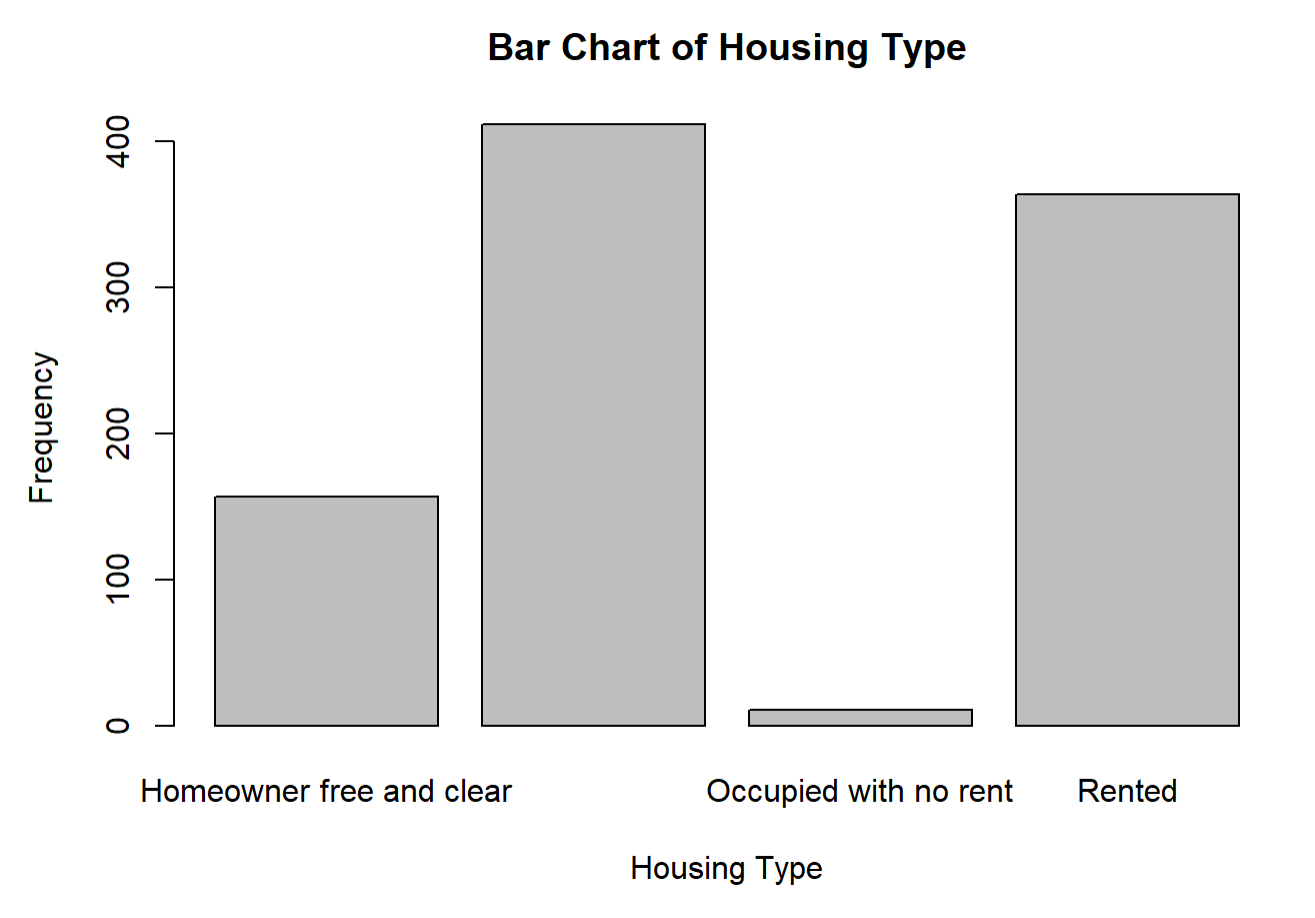
```
## Result of 5 x 10 is: 50
```

```
# 2. Figure out how to plot density of income
# Set the file path
file_path <- "C:/Users/gabed/OneDrive/Documents/R/custdata.tsv"

# Load the data
custdata <- read.table(file_path, header = TRUE, sep = "\t")

# Extract the income variable
income <- custdata$income

# Calculate density estimation of income
income_density <- density(income)

# Plot the density estimation
plot(income_density, main = "Density Plot of Income", xlab = "Income", ylab = "Density")
```




```
# 2.2 Provide a couple of sentences of description along with the plot. Imagine you are explaining this to your m
anager or a senior leader
# This density plot shows the distribution of income among our clientele.The x-axis represents income levels, whi
le the y-axis indicates the frequency of occurrence for each income value. The plot reveals a peak around a low i
ncome level, indicating the most common income bracket is low income among our customers.
```

```
# 3.1
# Set the file path
file_path <- "C:/Users/gabed/OneDrive/Documents/R/custdata.tsv"

# Load the data
custdata <- read.table(file_path, header = TRUE, sep = "\t")

# Filter out rows with "NA" values in the housing type field
filtered_custdata <- subset(custdata, !is.na(housing.type))

# Create a bar chart for housing type
barplot(table(filtered_custdata$housing.type),
        main = "Bar Chart of Housing Type",
        xlab = "Housing Type",
        ylab = "Frequency")
```




```
#4.1
# Load the knitr package
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.3.3
```

```
# Read the data from custdata.tsv
df <- read.delim("C:/Users/gabed/OneDrive/Documents/R/custdata.tsv")

# Filter the data to include only married customers with income > $50,000
subset_df <- subset(df, marital.stat == "Married" & income > 50000)

# Display the first 25 rows of the subset of customers as a nicely formatted table
kable(head(subset_df, 25))
```

	custid	sex	is.employed	income	marital.stat	health.ins	housing.type	recent.move	num.vehicles	age	state.of.res
12	17134	M	TRUE	220000	Married	TRUE	Homeowner free and clear	TRUE	2	33	Indiana
24	30768	M	TRUE	80000	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	50	New Jersey
41	52197	M	NA	65100	Married	TRUE	Homeowner free and clear	FALSE	2	66	Florida
44	52436	F	TRUE	139000	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	46	Pennsylvania
46	53214	M	TRUE	84010	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	57	New York
48	54177	M	FALSE	51500	Married	TRUE	Homeowner free and clear	FALSE	2	86	New York
52	62999	M	TRUE	91000	Married	TRUE	Homeowner with mortgage/loan	FALSE	4	51	New York
55	67776	M	TRUE	52000	Married	TRUE	Homeowner with mortgage/loan	FALSE	1	38	Florida
57	68221	M	TRUE	78000	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	29	California
58	69062	M	TRUE	120300	Married	TRUE	Homeowner free and clear	FALSE	4	59	California
60	74447	M	TRUE	162000	Married	TRUE	Homeowner with mortgage/loan	FALSE	4	33	Maryland
63	78476	M	TRUE	76000	Married	TRUE	Homeowner with mortgage/loan	FALSE	4	60	Illinois
66	80549	M	NA	85200	Married	TRUE	Homeowner free and clear	FALSE	2	75	Arkansas
67	82503	M	TRUE	70000	Married	TRUE	Homeowner free and clear	FALSE	2	69	California
74	90863	M	TRUE	285020	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	37	Washington
76	94743	M	TRUE	299000	Married	TRUE	Homeowner free and clear	FALSE	2	64	Missouri
77	96964	M	TRUE	266200	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	69	Texas
79	98086	M	NA	52100	Married	TRUE	Homeowner with mortgage/loan	TRUE	2	69	Florida
89	107458	M	TRUE	182500	Married	TRUE	Homeowner with mortgage/loan	TRUE	2	66	Florida
94	112116	M	TRUE	221000	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	65	Minnesota
100	117491	M	TRUE	110000	Married	TRUE	Homeowner with mortgage/loan	FALSE	4	58	New Jersey
101	117900	M	TRUE	60000	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	40	Massachusetts
108	126507	M	TRUE	112000	Married	TRUE	Homeowner with mortgage/loan	FALSE	4	51	Florida
111	133268	M	TRUE	140000	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	49	Wisconsin
121	150055	F	TRUE	78000	Married	TRUE	Rented	FALSE	2	24	Virginia

```
#4.2
# Read the data from custdata.tsv
df <- read.delim("C:/Users/gabed/OneDrive/Documents/R/custdata.tsv")

# Filter the data to include only married customers with income > $50,000
married_over_50k <- subset(df, marital.stat == "Married" & income > 50000)

# Count the number of married customers over $50,000 with health insurance
married_over_50k_with_insurance <- sum(married_over_50k$health.ins, na.rm = TRUE)

# Count the total number of married customers over $50,000
total_married_over_50k <- nrow(married_over_50k)

# Calculate the percentage
percentage_insurance <- (married_over_50k_with_insurance / total_married_over_50k) * 100

# Print the percentage
percentage_insurance
```

```
## [1] 96.2963
```

```
#4.3 To show how my hypothesis that married couples who make > 50k will have a much higher chance of being insure
d I compared there insurance percentage to the opposite set of data the group of people who are never married and
make < 50k see results below: # Read the data from custdata.tsv
df <- read.delim("C:/Users/gabed/OneDrive/Documents/R/custdata.tsv")

# Filter the data to include only "Never Married" customers with income <= $50,000
never_married_under_50k <- subset(df, marital.stat == "Never Married" & income <= 50000)

# Count the number of "Never Married" customers <= $50,000 with health insurance
never_married_under_50k_with_insurance <- sum(never_married_under_50k$health.ins, na.rm = TRUE)

# Count the total number of "Never Married" customers <= $50,000
total_never_married_under_50k <- nrow(never_married_under_50k)

# Calculate the percentage
percentage_insurance_never_married_under_50k <- (never_married_under_50k_with_insurance / total_never_married_und
er_50k) * 100

# Print the percentage
percentage_insurance_never_married_under_50k
```

```
## [1] 59.64912
```

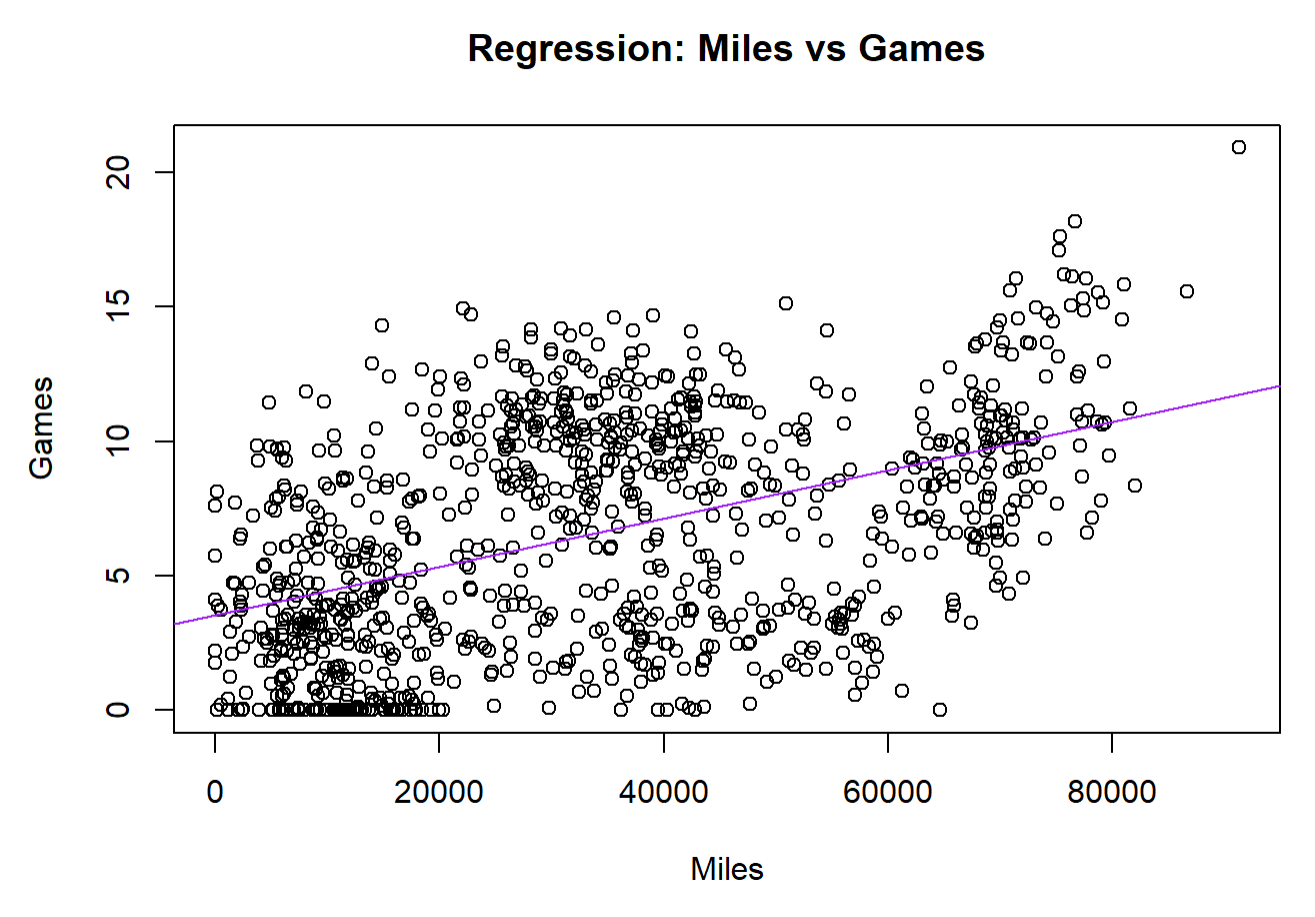
```
#5.1 My hypothesis is that there is going to be a minimal correlation between age,income,and number of vehicles.
Personally, I see all the time low income groups of people over leveraged financially with a lot of vehicles. I f
eel like to see a strong correlation we would have to bring a column that has data points surrounding financial l
iteracy.
```

```
#5.2 After running the correlation matrix I've came to the conclusion there is very small postie correlation betw
een a persons age, income and number of vehicles. I've came to this conclusion because the correlation coefficien
ts are all barley above zero. I came to this conclusion finally by analyzing the correlation coefficients that yo
u will see the below command will output.
# Read the data from custdata.tsv
df <- read.delim("C:/Users/gabed/OneDrive/Documents/R/custdata.tsv")

# Remove invalid data points
df <- na.omit(df)

# Subset the relevant columns
subset_df <- df[, c("age", "income", "num.vehicles")]

# Compute the correlation matrix
correlation_matrix <- cor(subset_df)

# Print the correlation matrix
print(correlation_matrix)
```

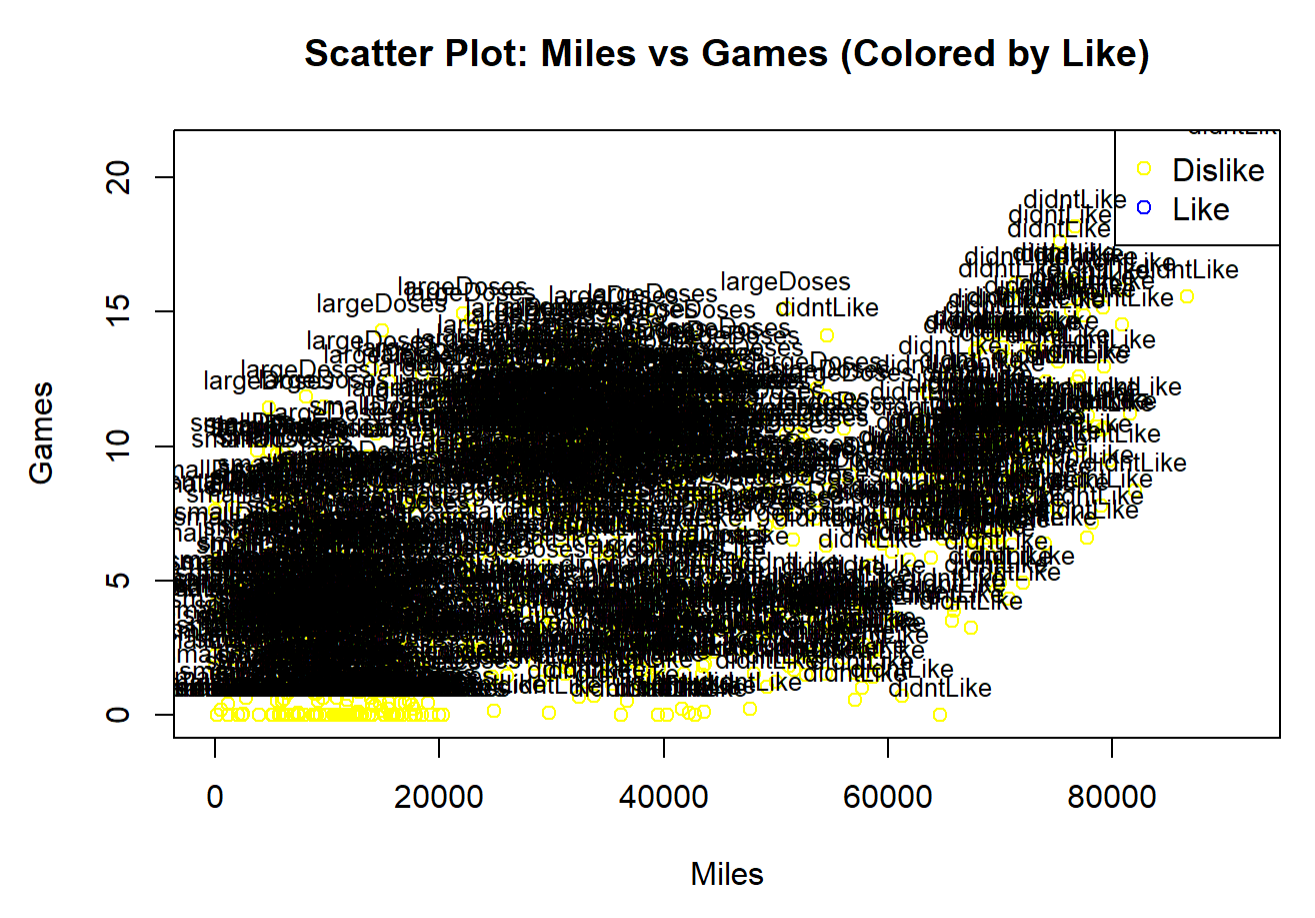
```
##                    age    income num.vehicles
## age          1.0000000 0.1271190    0.1052154
## income       0.1271190 1.0000000    0.1137691
## num.vehicles 0.1052154 0.1137691    1.0000000
```

```
#6.1 The relationship between eating ice cream and playing games is very minimal which the correlation percentage
below will represent. The relationship between traveling and playing games is stronger than the previous relation
ship which will be reflected by a correlation coefficient closer to .5 represented below:
# Read the data from the provided file
data <- read.csv("C:/Users/gabed/OneDrive/Documents/R/dating.csv")

# Compute correlation between Icecream and Games
cor_icecream_games <- cor(data$Icecream, data$Games)

# Compute correlation between Miles and Games
cor_miles_games <- cor(data$Miles, data$Games)
```

```
#6.2
# Perform linear regression
model <- lm(Games ~ Miles, data = data)

# Plot the regression line
plot(data$Miles, data$Games, main = "Regression: Miles vs Games", xlab = "Miles", ylab = "Games")
abline(model, col = "purple")

# Write the equation of the regression line
intercept <- coef(model)[1]
slope <- coef(model)[2]
eq <- paste("Games =", round(intercept, 2), "+", round(slope, 2), "* Miles")
text(225, 450, eq, pos = 3, col = "blue")
```




```
#6.3 Now let us see how well we can cluster the data based on the outcome (Like). Use Miles and Games to plot the
data and color the points using Like
# Plot the data with colors based on the Like variable
plot(data$Miles, data$Games, col = ifelse(data$Like == 1, "blue", "yellow"),
     main = "Scatter Plot: Miles vs Games (Colored by Like)",
     xlab = "Miles", ylab = "Games")
legend("topright", legend = c("Dislike", "Like"), col = c("yellow", "blue"), pch = 1)

# Add labels for clarity
text(data$Miles, data$Games, labels = data$Like, pos = 3, cex = 0.8, col = "black")
```




```
#6.3 Now cluster the data using k-means and plot the same data using clustering information. Show the plot and co
mpare it with the previous plot. Provide your thoughts about how well your clustering worked in two to four sente
nces? I personally thought the clustering worked well. I actually like the out put of using the clustering method
better than previous methods. One drawback about clustering that I experienced in my clustering of Miles and Game
s visualization is that the data points are overlapping each other which make it hard to distiguish data points.
# Perform k-means clustering
set.seed(123)  # For reproducibility
k <- 2  # Number of clusters
kmeans_result <- kmeans(data[, c("Miles", "Games")], centers = k)

# Plot the data with colors based on the cluster assignments
plot(data$Miles, data$Games, col = kmeans_result$cluster,
     main = "Scatter Plot: Miles vs Games (Colored by Cluster)",
     xlab = "Miles", ylab = "Games")
legend("topright", legend = paste("Cluster", 1:k), col = 1:k, pch = 1)

# Add labels for clarity
text(data$Miles, data$Games, labels = kmeans_result$cluster, pos = 3, cex = 0.8, col = "black")
```