Dingyi Duan

Module 4 Assignment

ADS-500B

03/28/2021

Module 4 Assignment Questions

Note that the answers to each of these questions should be the direct result of running appropriate commands and not involve any further processing, including manual work. Answers without the commands used to achieve them will not get any grade.

Datasets (located in your assignment prompt in Blackboard) contain two files.

- The first, is customer data related to health insurance. The data set file name is “custdata.tsv”. You will use this data set to answer questions in sections 2-5. Field names (in order): custid, sex, is.employed, income, marital.stat, health.ins, housing.type, recent.move, num.vihicles.
- The second file contains observations related to dating. The data set file name is “dating.csv”. You will use this data set to answer questions in section 6. Field names (in order): Miles, Games, Icecream, Like

1. Write a multiplication script using either a “for” loop or a “while” loop. Show your script.

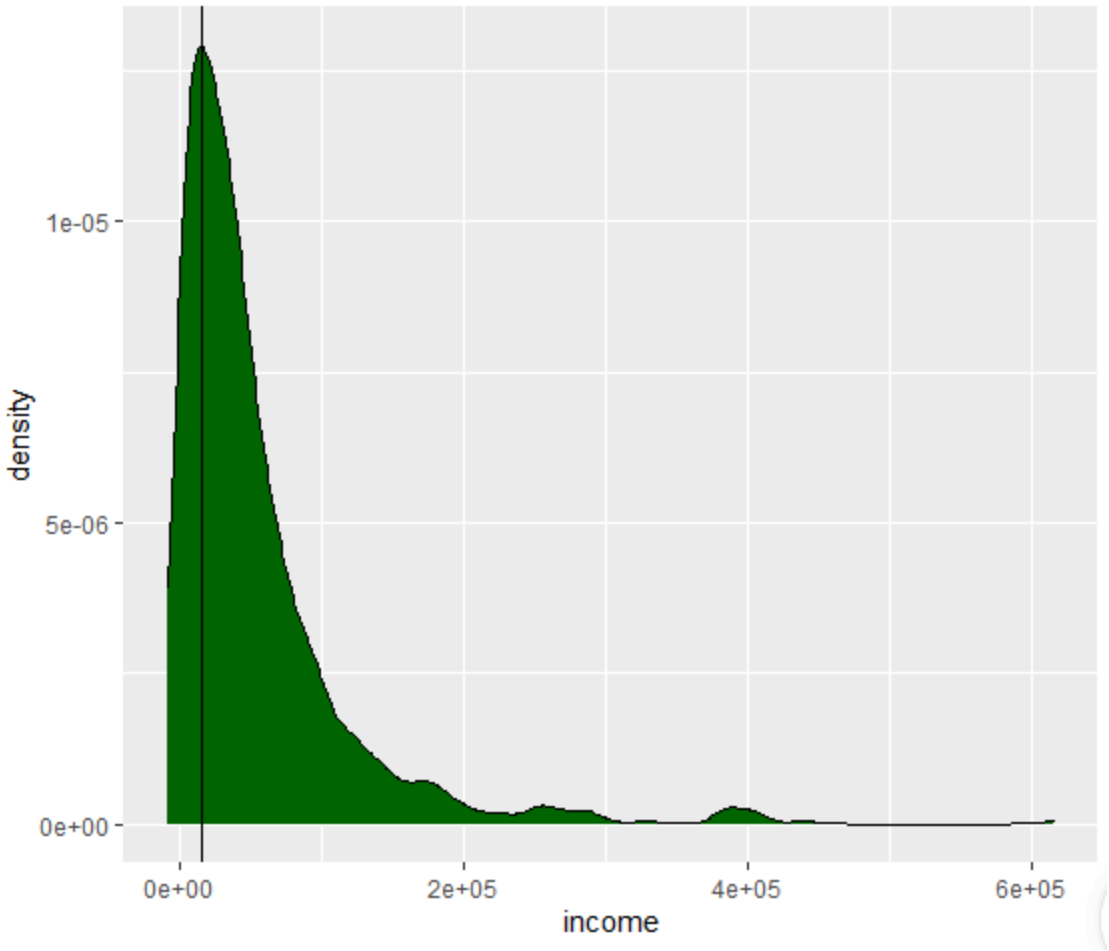
```
#Q1

#Generate 2 vectors with random numbers range from 0-100
pool=c(1,2,3,4,5,6,7,8,9,10)
num = 5

#For loop for calculating product, positive if number in "pool" greater
#than 5, otherwise negative
for (count in 1:10){
cat ("position in the vector is: ",count)
if (pool[count]>num){
print(pool[count]*num)
} else {
print(pool[count]*num*(-1))
}
}
```

```
position in the vector is:  1[1] -5
position in the vector is:  2[1] -10
position in the vector is:  3[1] -15
position in the vector is:  4[1] -20
position in the vector is:  5[1] -25
position in the vector is:  6[1] 30
position in the vector is:  7[1] 35
position in the vector is:  8[1] 40
position in the vector is:  9[1] 45
position in the vector is:  10[1] 50
```

2. Using the customers data (custdata.tsv). Like histogram, you can also plot the density of a variable.

- 2.1: Figure out how to plot density of income.

```
library(ggplot2)
library(tidyverse)

custdata = read.table(file.choose(),header=T,sep='\t')
print(summary(custdata))


#find highest y value which is 39
which.max(density(custdata$income)$y)

#use ggplot to plot density
ggplot(custdata,aes(x=income))+geom_density(fill="darkgreen")+
  geom_vline(xintercept = density(custdata$income)$x[39])
```



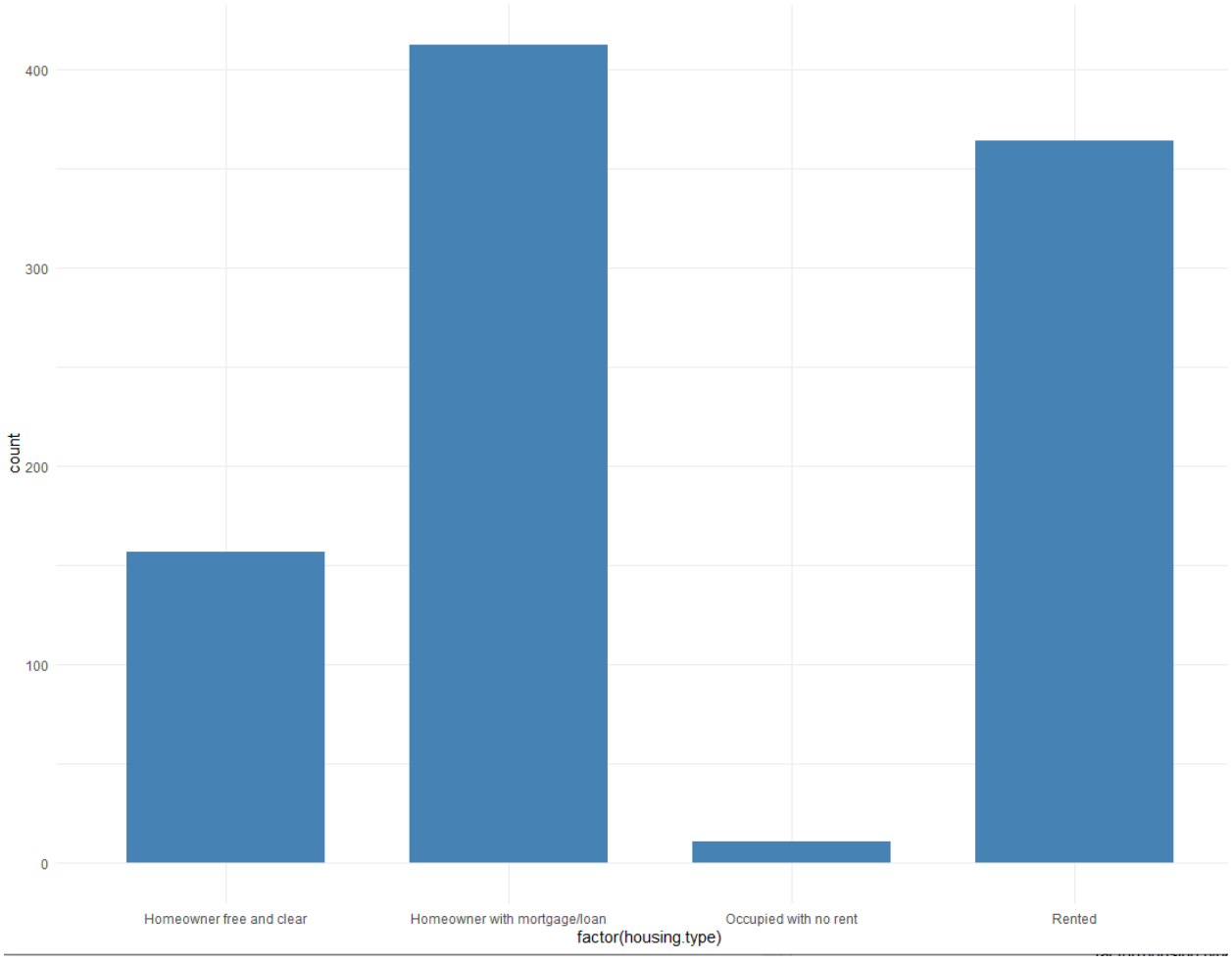
- 2.2: Provide a couple of sentences of description along with the plot. Imagine you are explaining this to your manager or a senior leader.

Answer: The income ranges from below zero to $600,000 which shows a great diversity in wealth. The black vertical line indicates the peak of the density. We can see that the amount of people whose income is below zero is still more than those who have really high income. This states that the majority of the customers are mid to below-mid class.

3. Using the customers data (custdata.tsv).

- 3.1: Create a bar chart for housing type using the customers data. Make sure to remove the “NA” type. [Hint: You can use subset function with an appropriate condition on housing type field.] Provide your commands and the plot.

```r
#subset custdata to remove "NA" from "housing.type'
new_custdata = subset(custdata,housing.type != "NA")

#bar plot for "housing.type"
ggplot(new_custdata, aes(x=factor(housing.type)))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal()
```



This histogram chart indicates that most of the customers have either mortgage/ loan or rent.

4. Using the customers data (custdata.tsv).

- 4.1: Extract a subset of customers that are married and have an income more than $50,000.

```
#Q4

#subset the original to married and income of > $50000
Q4_custdata = subset(custdata,marital.stat == "Married" & income >50000)
```

- 4.2: What percentage of these customers have health insurance?

Answer: 96.3%

```
#count the married and income of > $50000 who have insurance
ins_count_Q4 = sum(Q4_custdata$health.ins == "TRUE")
#count all the married and income of > $50000
all_ins_count_Q4 = sum(Q4_custdata$health.ins == "TRUE") +
                   sum(Q4_custdata$health.ins == "FALSE")
#calculate %
ins_pct_Q4 = (round(ins_count_Q4*100/ all_ins_count_Q4,2))
cat(ins_pct_Q4,"% of the customers who are married and have an income more
      than $50,000 have insurance coverage.")
```

```
96.3 % of the customers who are married and have an income more
      than $50,000 have insurance coverage.
```

- 4.3: How does this percentage differ from that for the whole data set?\

Answer:The whole data set has 84.1% of customers that have insurance. This indicates that richer couples are more likely to have insurance coverage.

```
#count all who have insurance
ins_count_custdata = sum(custdata$health.ins == "TRUE")
#count all customers
all_ins_count_custdata = sum(custdata$health.ins == "TRUE") +
                         sum(custdata$health.ins == "FALSE")
#calculate the %
ins_pct_count = round(ins_count_custdata*100 / all_ins_count_custdata,2)


cat(ins_pct_count,"% of all customers have insurance coverage.")
```

```
84.1 % of all customers have insurance coverage.
```

5. Using the customers data (custdata.tsv).

- 5.1: In the customers data, do you think there is any correlation between age, income, and number of vehicles? Explain why or why not.

Answer: From common sense, I think that there should be a strong positive correlation between age and income; income and vehicles, as age increases, income increases, thus, it is possible that there is also a positive correlation between income and number of vehicles.

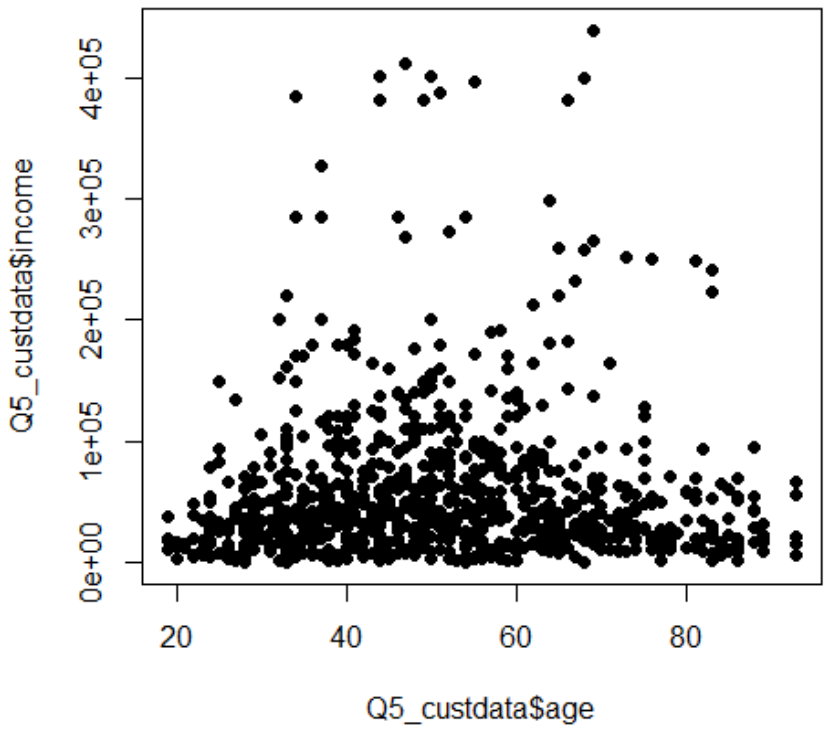
- 5.2:Report your correlation numbers and interpretations. [Hint: Make sure to remove invalid data points, otherwise you may get incorrect answers!]

```
#Q5

#subset the original database to exclude the outliers and NA which are
#age < 0 and > 100; income < 0 and > 60000, NA for number of cars owned.
Q5_custdata = subset(custdata, subset=(custdata$age > 0) & (custdata$age < 100) &
  (custdata$income > 0) & (custdata$income < 60000) & (custdata$num.vehicles != "NA"))

#plot the corresponding scatter plot for all three variables using color-blind-friendly colors
plot(Q5_custdata$age,Q5_custdata$income,pch=19)
plot(Q5_custdata$age, Q5_custdata$num.vehicles, col = "#E69F00", pch = 19)
plot(Q5_custdata$income, Q5_custdata$num.vehicles, col = "#56B4E9", pch = 19)

#calculate their correlations and round to two decimal place
round(cor(Q5_custdata$age,Q5_custdata$income, method = c("pearson", "kendall", "spearman")),2)
round(cor(Q5_custdata$age,Q5_custdata$num.vehicles, method = c("pearson", "kendall", "spearman")),2)
round(cor(Q5_custdata$income,Q5_custdata$num.vehicles, method = c("pearson", "kendall", "spearman")),2)
```

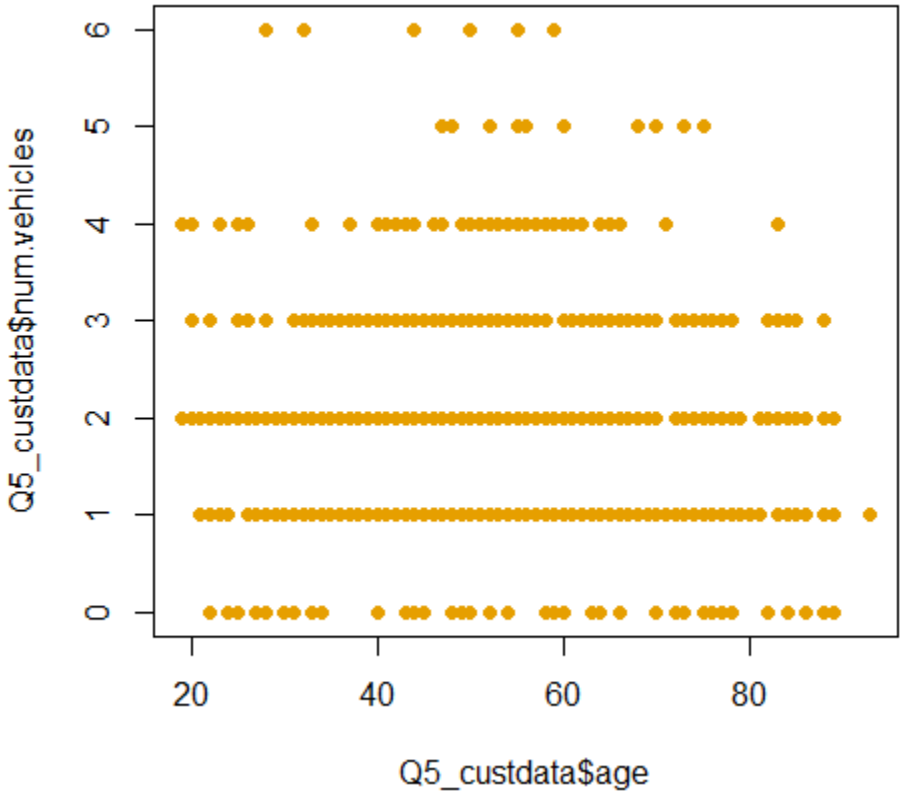
```
> round(cor(Q5_custdata$age,Q5_custdata$income, method = c("pearson", "kendall", "spearman")),2)
[1] -0.02
> round(cor(Q5_custdata$age,Q5_custdata$num.vehicles, method = c("pearson", "kendall", "spearman")),2)
[1] -0.07
> round(cor(Q5_custdata$income,Q5_custdata$num.vehicles, method = c("pearson", "kendall", "spearman")),2)
[1] 0.14
```

After removing the outliers for age (age should be between 0 and 100), income(income should be between 0 and 60000), number of vehicles (exclude “NA”).

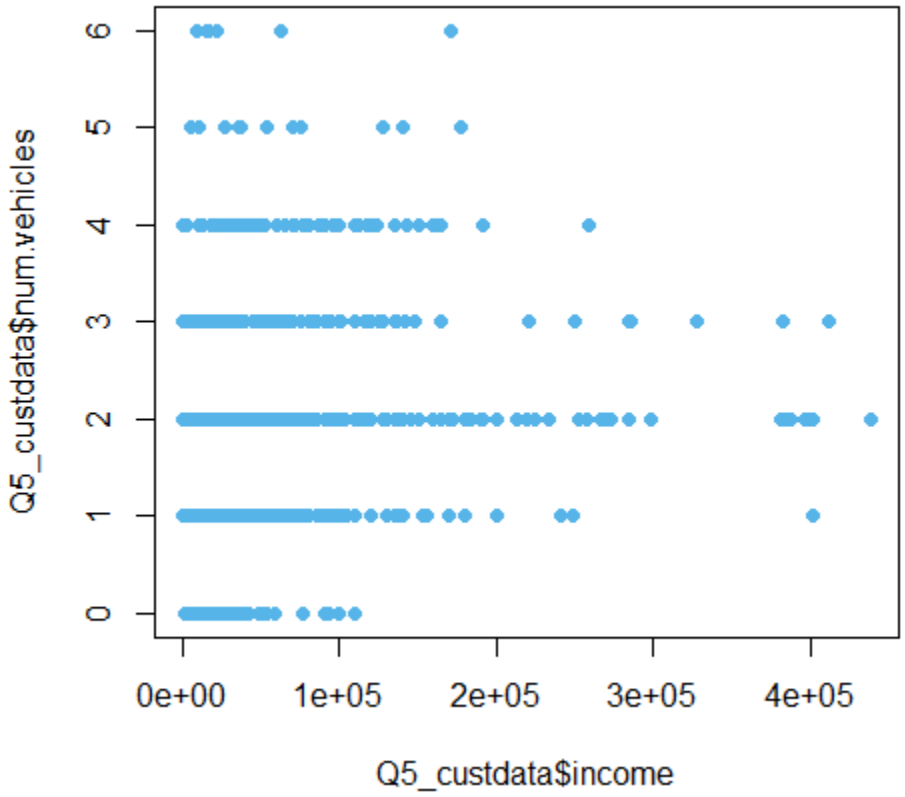
Correlation between age and income: -0.02



Correlation between age and number of vehicles: -0.07



Correlation between income and number of vehicles: 0.14



Conclusion: There is no strong correlation between age, income and number of vehicles owned.

6. You are given a data file containing observations for dating. Someone who dated 1000 people (!) recorded data about how much that person travels (Miles), plays games (Games, and eats ice cream (Icecream). With this, the decision about that person (Like) is also noted. Use this data to answer the following questions using R:

- 6.1: Is there a relationship between eating ice cream and playing games? What about traveling and playing games? Report correlation values for these and comment on them.

Answer: There is no obvious relation between eating ice cream and playing games; However, there's a positive correlation between traveling and playing games.

Correlation:

- **Ice cream vs Games: 0.01**
- **Miles vs Games: 0.47**

```
library(ggplot2)
library(tidyverse)
library(ggpubr)
library(datasets)
library(factoextra)

#Q6

#read the file
dating = read.table(file.choose(),header=T,sep=',')


#calculate the correlation coefficients
round(cor(dating$Icecream,dating$Games, method = c("pearson", "kendall", "spearman")),2)
round(cor(dating$Miles, dating$Games, method = c("pearson", "kendall", "spearman")),2)
```

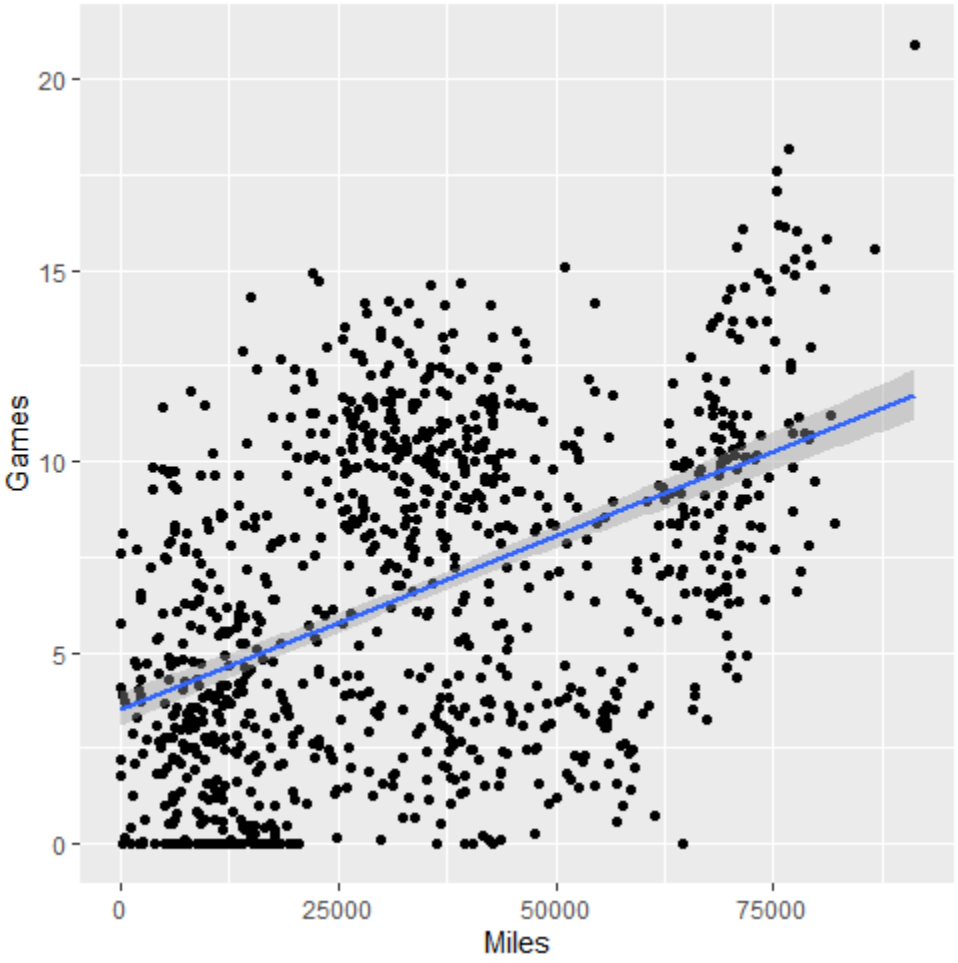
```
> round(cor(dating$Icecream,dating$Games, method = c("pearson", "kendall", "spearman")),2)
[1] 0.01
> round(cor(dating$Miles, dating$Games, method = c("pearson", "kendall", "spearman")),2)
[1] 0.47
```

- 6.2: Let us use Miles to predict Games. Perform regression using Miles as the predictor and Games as the response variable. Show the regression graph with the regression line. Write the line equation.

Equation: Games = 0.00009003 * Miles + 3.532

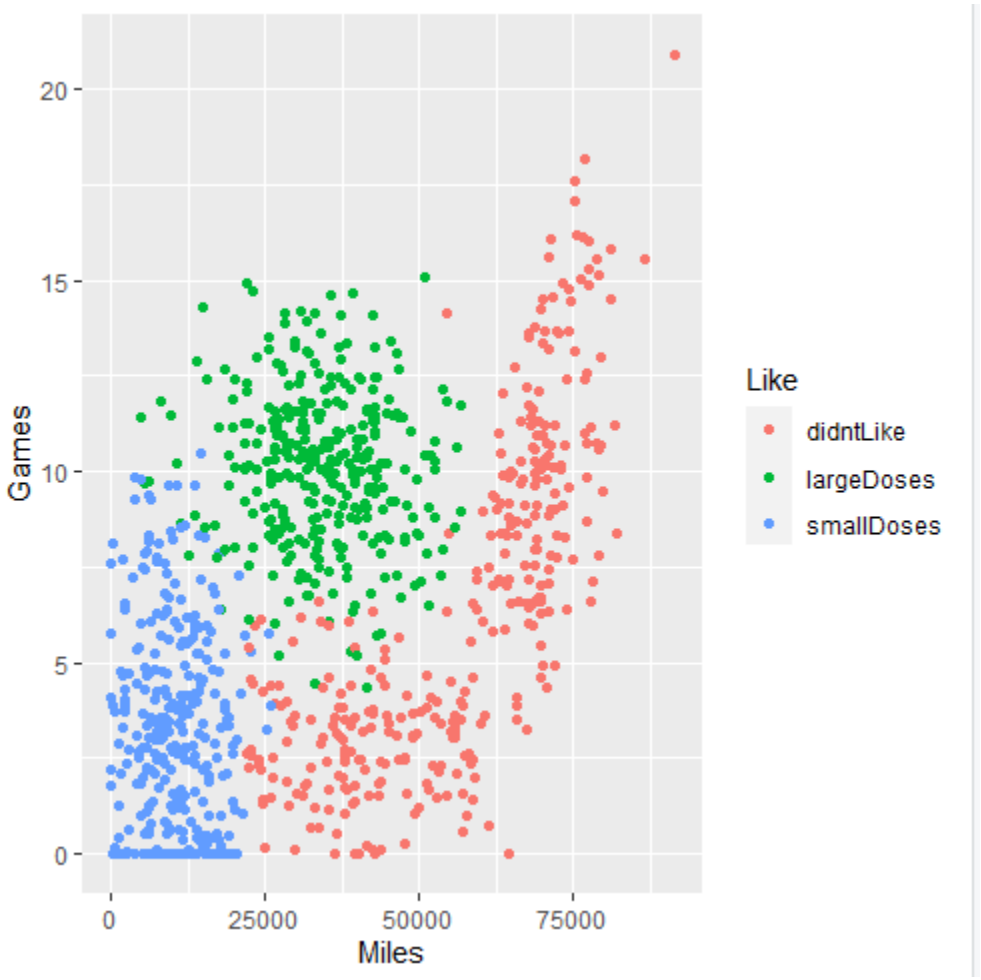
```
#plot the variables and show regression line
ggplot(dating, aes(x=Miles,y=Games)) + geom_point() + stat_smooth(method='lm')

#calculate the coefficient for the predictor(Miles) and intercept.
lm(Games ~ Miles, dating)

#linear equation: Games = 0.00009003 * Miles + 3.532
```

```
Coefficients:
(Intercept)        Miles
  3.532e+00    9.003e-05
```



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 variables and color points using "Like"
ggplot(dating, aes(x=Miles,y=Games,color=Like)) + geom_point()
```



Now cluster the data using k-means and plot the same data using clustering information. Show the plot and compare it with the previous plot. Provide your thoughts about how well your clustering worked in two to four sentences.

Answer: From the two plots we can see that the cluster information matches the actual classes for the data points quite well in general. The difference is that in cluster mode each cluster is more distinguishable horizontally. Some points from "largeDoses" below 25000 miles are categorized into the first cluster; many points from "didntLike" class that are below 5 "Games" belong to the second cluster.

```r
# Plot variables and color points using "Like"
ggplot(dating, aes(x=Miles,y=Games,color=Like)) + geom_point()


#generate a random number seed and then use the kmeans function.
set.seed(100)
datingCluster = kmeans(dating[, 1:2], 3, nstart = 50)

table(datingCluster$cluster, dating$Like)

#plot the new cluster data
ggplot(dating, aes(x=Miles,y=Games,color=datingCluster$cluster)) + geom_point()
```

```
  didntLike largeDoses smallDoses
1       211          14          0
2       122         267          3
3         9          46        328
```