Module 4 Exercises

Filipp Krasovsky

Question 1: Create a for-loop or while loop for multiplication

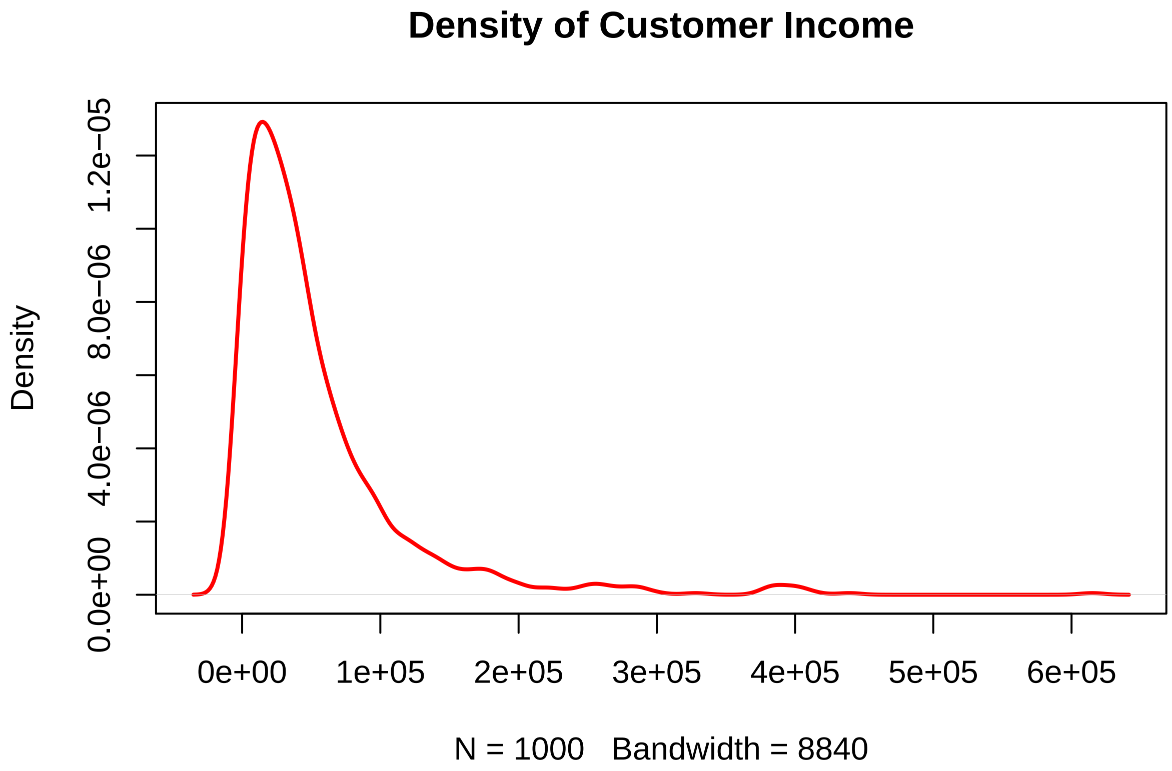
```r
  multiply <- function(a,b){
    out = 0
    for (i in 1:abs(b)){
      out = out + a
    }
    return (out * (b/abs(b)))
  }

  multiply(10,5)
```

```
## [1] 50
```

Question 2.1: Figure out the plot density of income for the customer data.

```r
  #import TSV
  customers <- read.csv('custdata.tsv',sep='\t',header=TRUE)
```

```r
#analyze income variable
#calculate density
income_density <- density(customers$income)
par(mfrow=c(1,1))

plot(income_density, lwd = 2, col = "red",main="Density of Customer Income")
```

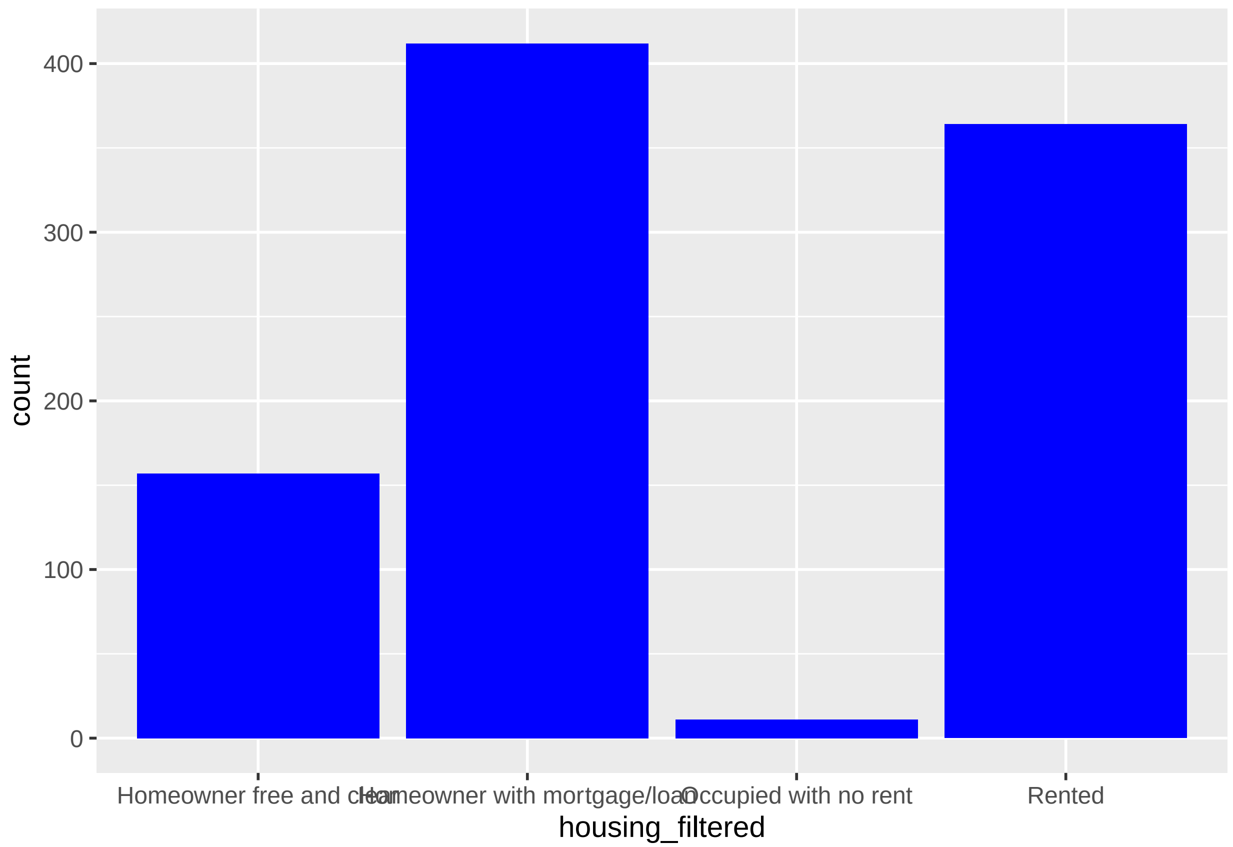


Question 2.2: Interpretation and Analysis

We can interpret this, in a decision-making environment, to mean that we can expect the income of a randomly selected individual from the customer base to have an income that lies between 0 and 100,000. Relying entirely on a visual analysis, we can also assert that the most frequently occurring income is within the 50-60,000 dollar range.

Question 3.1: Create a bar chart of the housing types, removing NAs.

```
library(ggplot2)
housing_filtered <- subset(customers$housing.type,!is.na(customers$housing.type))
housing_filtered <- (as.data.frame(housing_filtered))

ggplot(housing_filtered)+geom_bar(aes(x=housing_filtered),fill="blue")
```



Question 4.1: Extract a subset of customers that are married and have an income above $50,000.

```
sub_cus <-subset(customers,customers$marital.stat=="Married" & customers$marital.stat > 50000)
```

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

```
sub_with_insurance <- nrow(subset(sub_cus,sub_cus$health.ins==TRUE))
sub_ins_ratio      <- round(sub_with_insurance/nrow(sub_cus),2)
print(sub_ins_ratio)
```

```
## [1] 0.88
```

Question 4.3: How does this percentage differ from the whole data set?

```
total_with_insurance <- nrow(subset(customers,customers$health.ins==TRUE))
total_ins_ratio <- round(total_with_insurance/nrow(customers),2)
print(total_ins_ratio)
```

```
## [1] 0.84
```

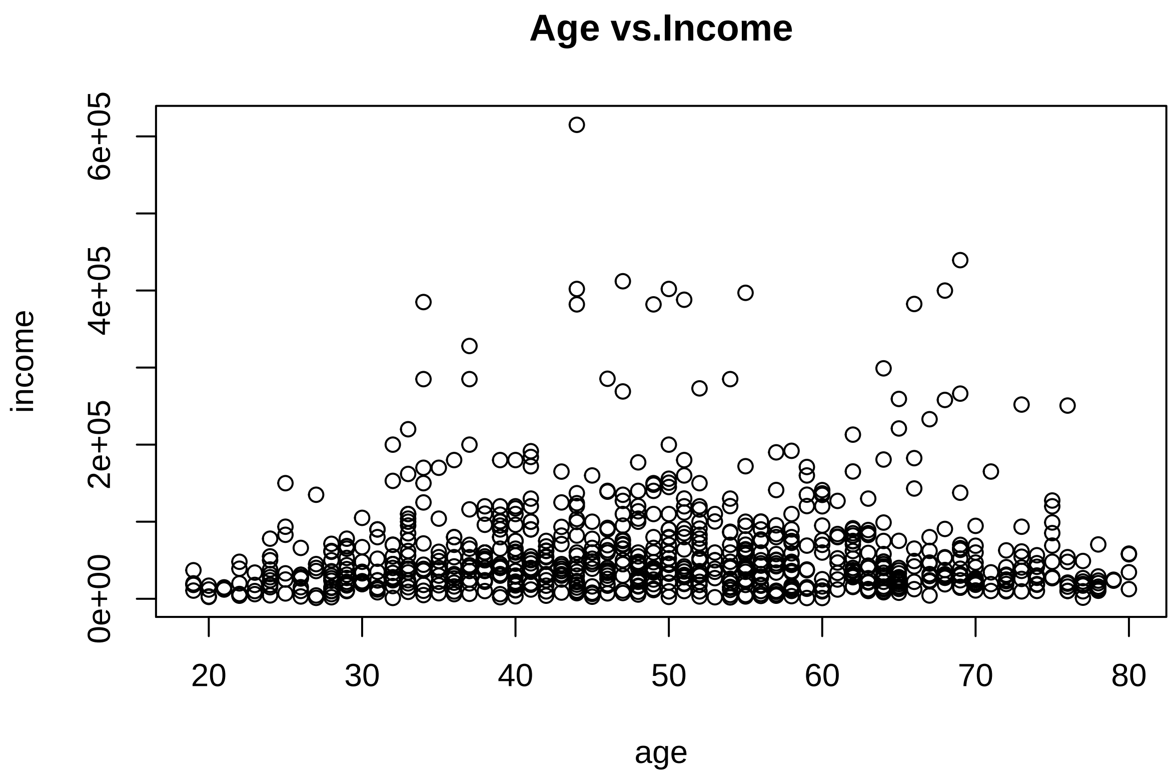
The subset of individuals who are married and have an income above $50k has a ratio of individuals with health insurance that is around 4% higher than the rate of insurance for the entire customer base sampled.

5.1/5.2: Is there any correlation between age, number of vehicles, and income? Reporting correlation findings and interpretations, removing null values first.

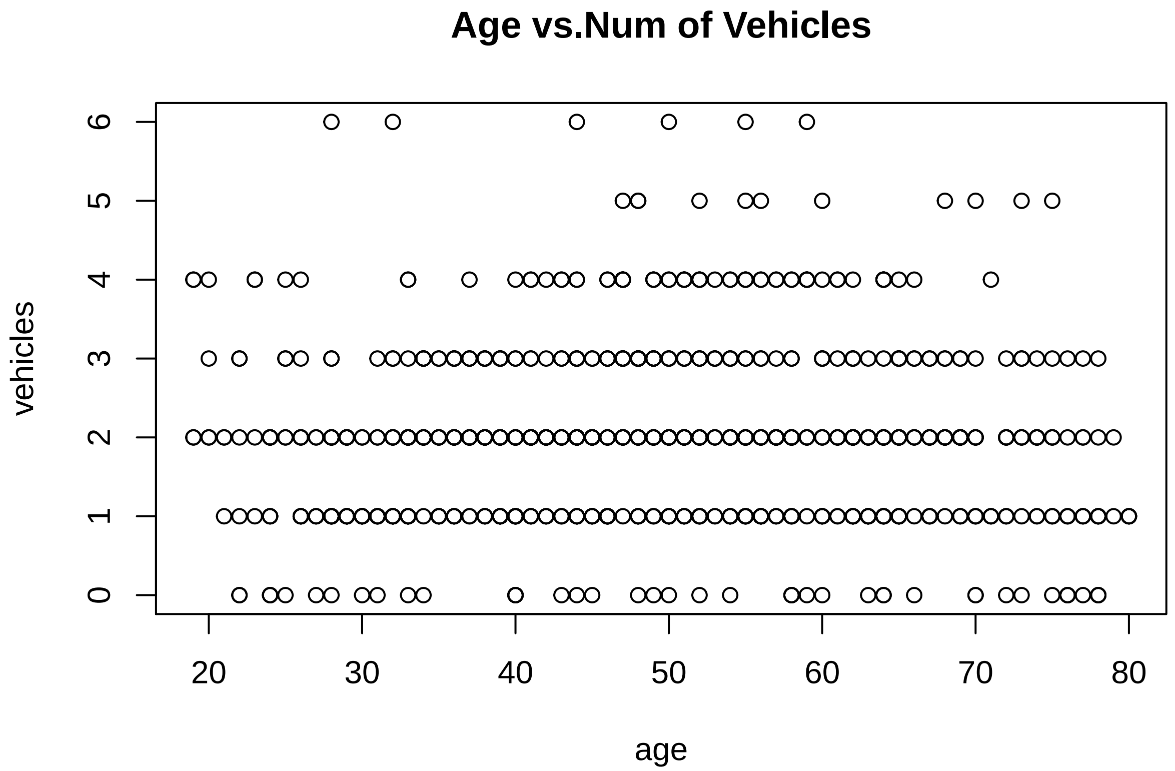
Without looking too closely at data, we can presupose some link between age and income, but this connection is probably nonlinear - people in their twenties tend to make entry level salaries while retired individuals generally tend to have a low income, so mid-career salaries are probably highest for individuals at the ages of 40-50. Similarly, the number of vehicles owned is probably associated with a higher income.

```r
#first,remove any null age, income, or vehicle values
sub <- subset(customers, !is.na(customers$income) & !is.na(customers$age) & !is.na(customers$num.vehicle

#remove all outliers and invalid values. in this context, all negative values. For age in particular,
#we will remove all ages less than 80, based on the average life expectancy in the US.
#we do not filter out zero vals for num.vehicles because it is not irregular for an individual not to d

sub <- subset(sub, sub$age > 0 & sub$age <= 80 & sub$income >= 1000 & sub$num.vehicles >= 0)
```

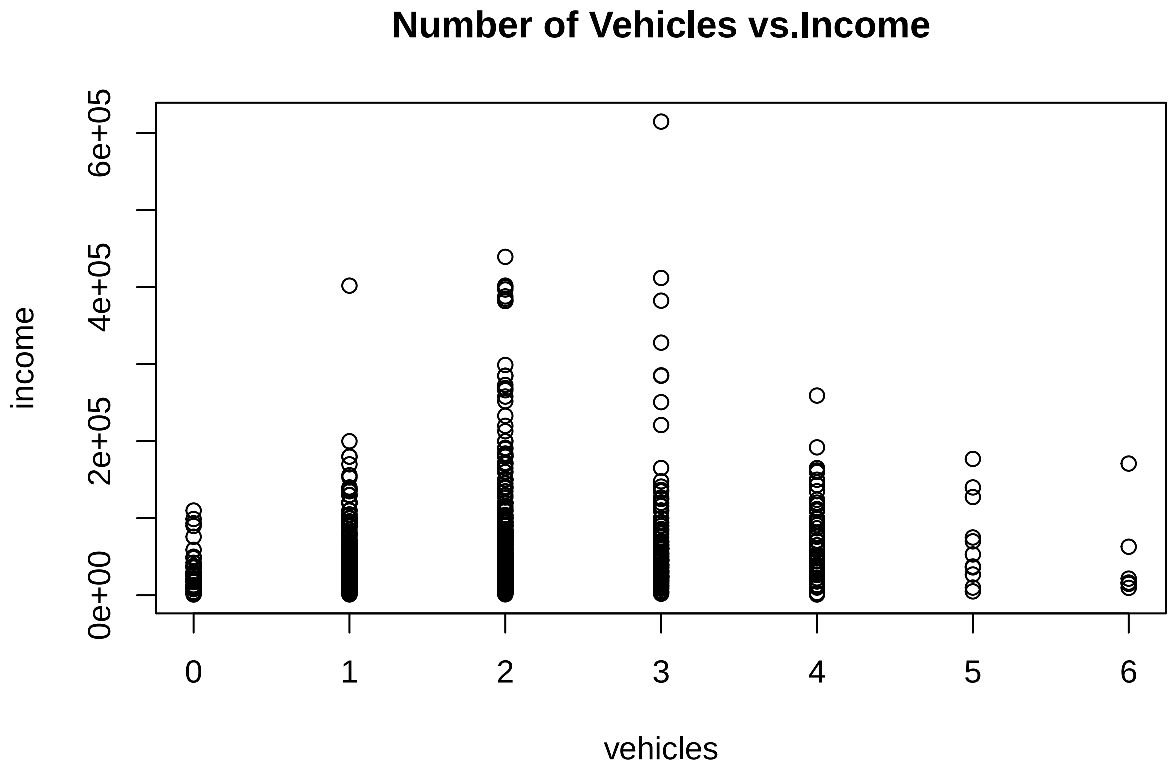
```r
plot(x=sub$age,y=sub$income,main="Age vs.Income",xlab="age",ylab="income" )
```



```r
plot(x=sub$age,y=sub$num.vehicles,main="Age vs.Num of Vehicles",xlab="age",ylab="vehicles")
```



```
plot(x=sub$num.vehicles,y=sub$income,main="Number of Vehicles vs.Income",xlab="vehicles",ylab="income")
```



```r
cor(sub$age,sub$income)
```

```
## [1] 0.02751576
```

```r
cor(sub$age,sub$num.vehicles)
```

```
## [1] 0.005569206
```

```r
cor(sub$income,sub$num.vehicles)
```

```
## [1] 0.1428948
```

Plots of all three possible combinations, combined with the actual correlation results, belie all of the assertions made in 5.1. The closest possible case is the very weak correlation of 14% between income and the number of vehicles. However, the problem with concluding this analysis here is that income does not account for the standard of living that differs state by state.

To offset this noise, we can iterate over each state and find the correlation based on residence.

```r
#get unique state values
all_states <- unique(sub$state.of.res)

#initialize an  nrow(all_states) by 3 matrix filled with NA values.
cor_vals <-  matrix(data=NA,nrow=length(all_states),ncol=3)

#create data frame
state_cors <- data.frame(cbind(all_states,cor_vals))
colnames(state_cors) <- c("State","Income/Vehicles","Income/age","age/Vehicles")
```

```
#iterate over states, if state has >=30 rows, we append r to the output vector and print
for (i in 1:length(all_states)){

  this.state_name <- state_cors[i,1]
  this.state_ref <- subset(sub,sub$state.of.res==this.state_name)

  if(nrow(this.state_ref)>=30){
    state_cors[i,2]<-round(cor(this.state_ref$income,this.state_ref$num.vehicles),2)
    state_cors[i,3]<-round(cor(this.state_ref$income,this.state_ref$age),2)
    state_cors[i,4]<-round(cor(this.state_ref$age,this.state_ref$num.vehicles),2)
  }
}

state_cors <- subset(state_cors,!is.na(state_cors$`Income/Vehicles`))
print(state_cors)
```

```
##           State Income/Vehicles Income/age age/Vehicles
## 1      Michigan            0.26      -0.08         0.03
## 4       Florida            0.25       0.21         0.07
## 5      New York            0.08       0.07        -0.03
## 7      Illinois            0.15        0.1        -0.01
## 12 Pennsylvania            0.26      -0.11        -0.09
## 14   New Jersey            0.23      -0.24         0.02
## 15         Ohio            0.26      -0.22         0.15
## 17   California            0.06       0.13         0.11
## 24        Texas           -0.05       0.11        -0.17
```

Not included in this output are correlations for n < 30, with some datasets as small as two observations - making analysis irrelevant. For the most part, controlling for state of residence greatly increases the correlation coefficient from the otherwise negligent ~1%.

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.

```
#import dataset
dating <- read.csv('dating.csv',header=TRUE)
```

```
print(cor(dating$Icecream,dating$Games))
```

```
## [1] 0.008874313
```

```
print(cor(dating$Miles,dating$Games))
```

```
## [1] 0.4658472
```

Overall, the correlation between traveling and playing games is much higher than the correlation between ice cream and playing games; holistically, there is no correlation between ice cream consumption and playing video games. Travel seems to have a somewhat weak but significant positive association with playing games.

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.

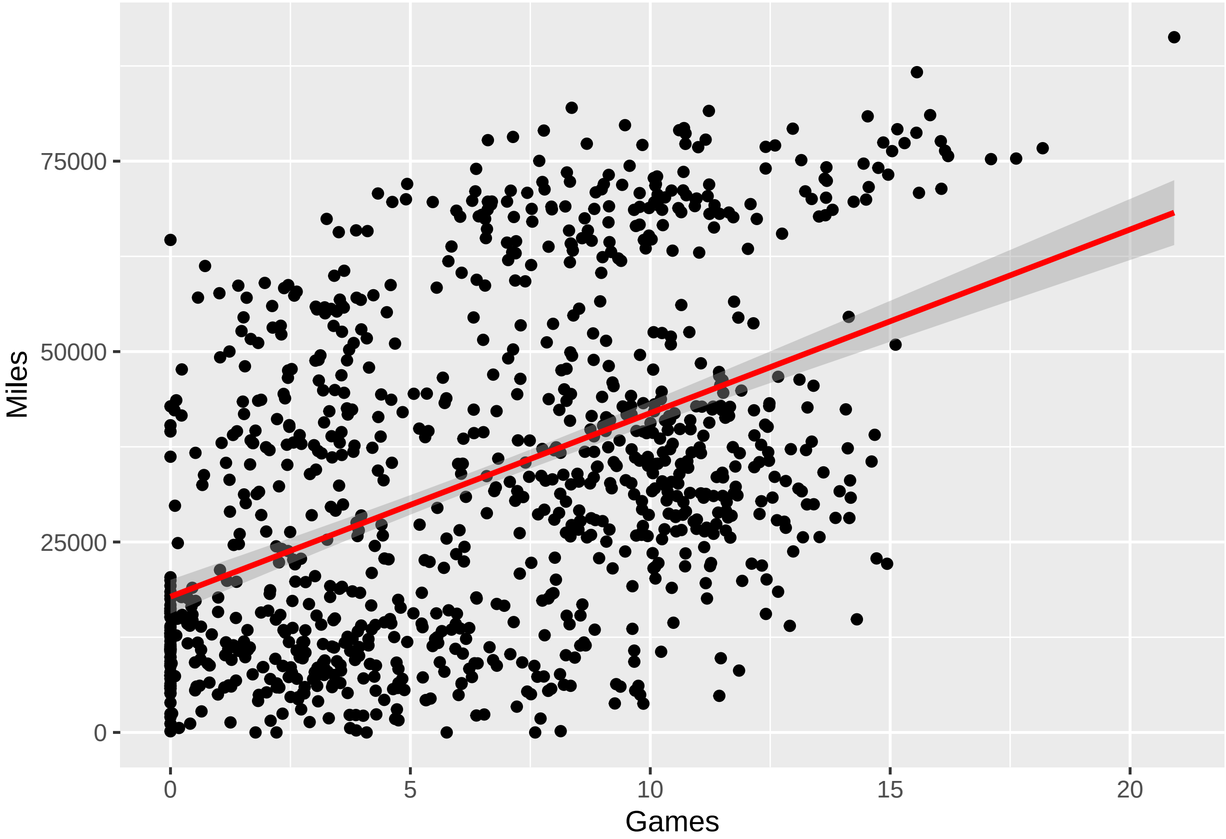
```
lm.fit <- lm(dating$Games ~ dating$Miles)
print(lm.fit$coefficients)
```

```
##  (Intercept) dating$Miles
## 3.531628e+00 9.003403e-05
```

We can therefore conclude that the formula is: Games = 3.5163 + 0.00009 * Miles

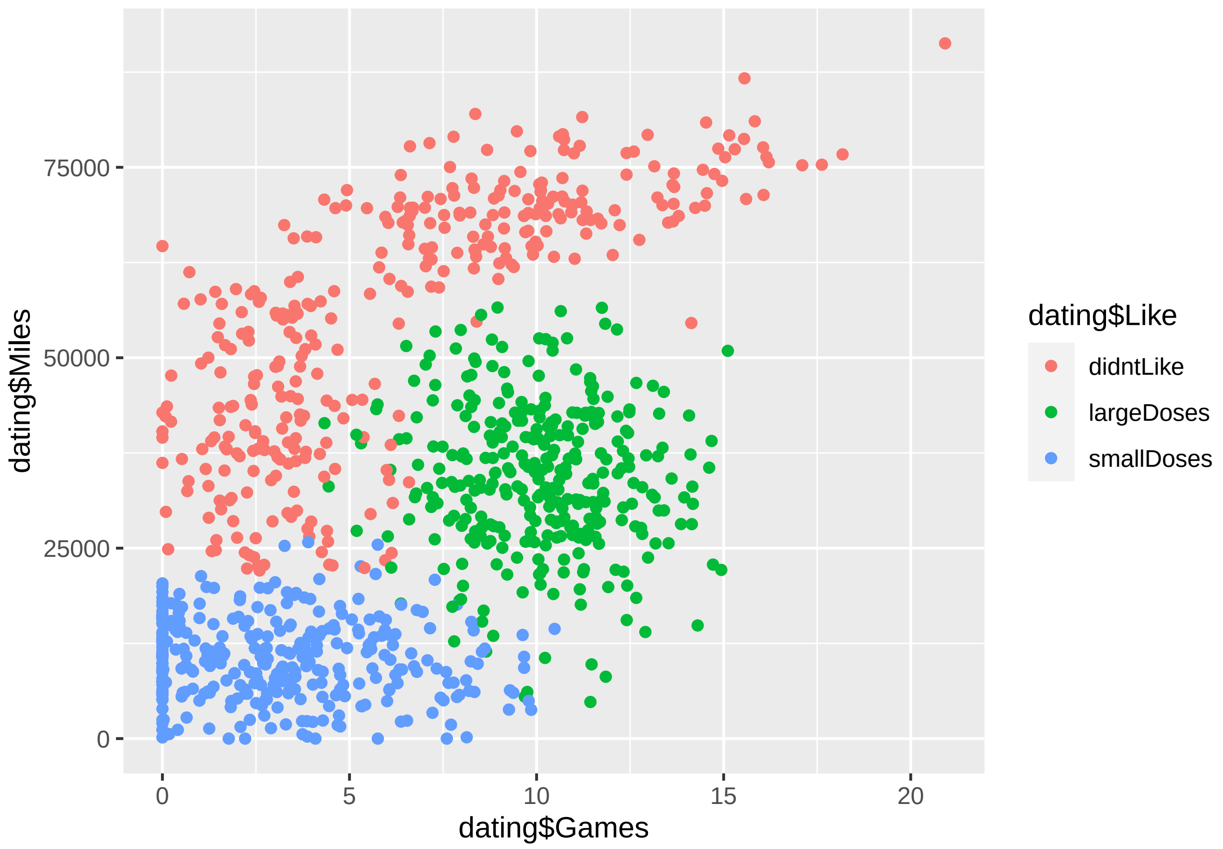
```
#visualization

dates.graph<-ggplot(dating, aes(x=Games, y=Miles))+geom_point()+geom_smooth(method="lm", col="red")
print(dates.graph)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



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. 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. (10 points)

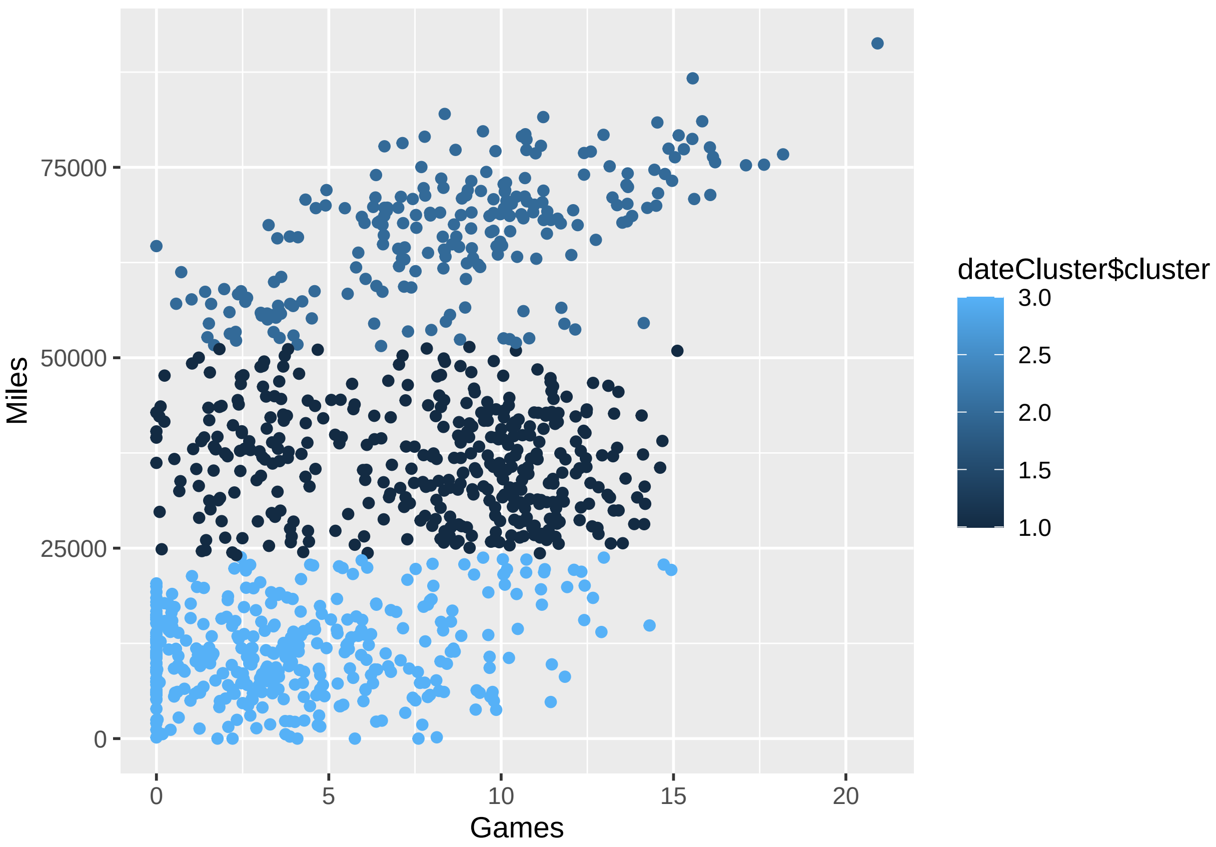
```
library(datasets)
library(ggplot2)
ggplot(dating,aes(dating$Games,dating$Miles,color=dating$Like))+geom_point()
```



```r
set.seed(20)
dateCluster <- kmeans(dating[,1:2],3,nstart=20)
table(dateCluster$cluster, dating$Like)
```

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

```r
ggplot(dating,aes(Games,Miles,color=dateCluster$cluster))+ geom_point()
```



Overall, the kmeans plot was able to capture much of the data accurately for clusters 3 and 2, but saw considerable overlap between clusters 1 and 2; in particular, cluster 2 counterintuitively captures data points that have a value of miles between 25000 and 50000. In conclusion, the clustering information almost matches the actual classes, but not quite.