Module 4 Exercise Questions

Samuel Scott

2025-11-17

Question 1 : Without using the multiplication operator (*), write ## a

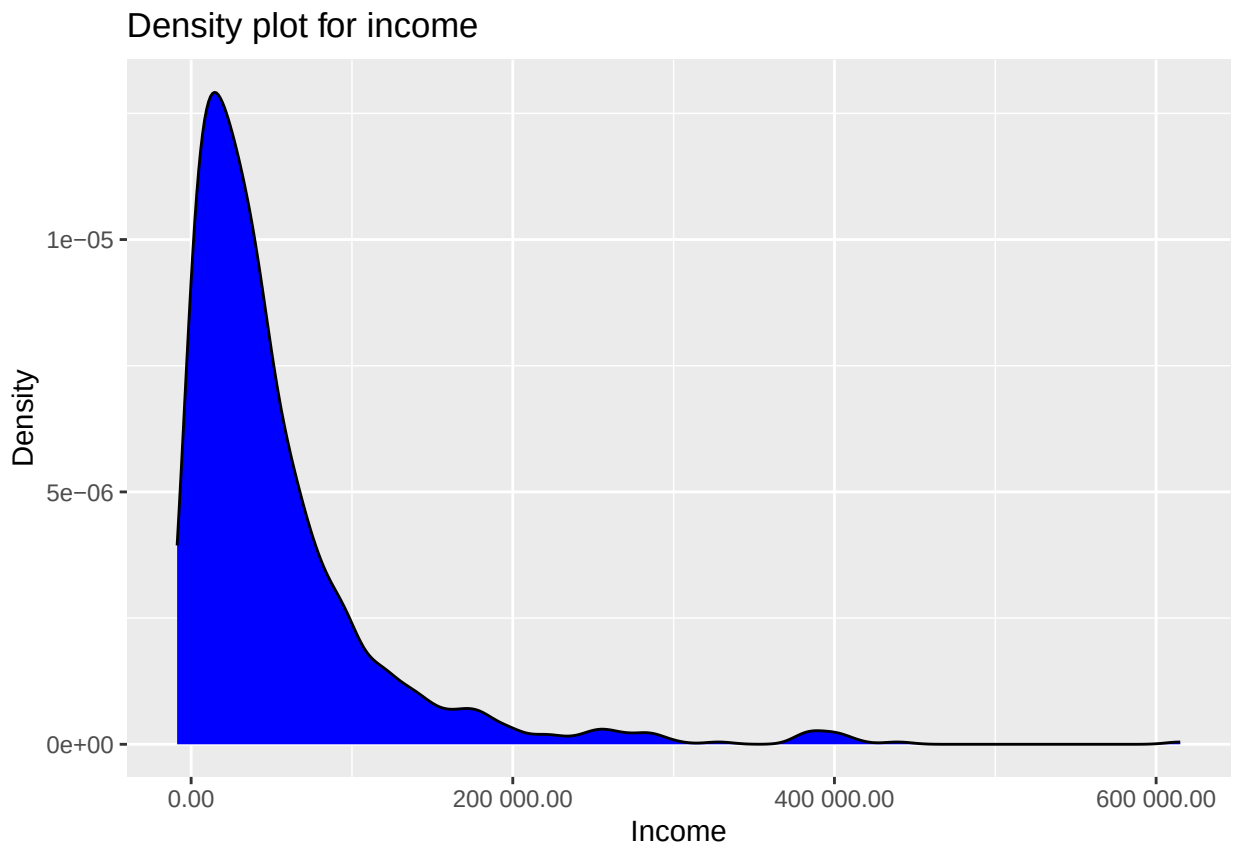
#multiplication script that takes two numbers as input and outputs the product ##using either a "for" loop or a "while" loop. Show your script.

```
mult = function(x,y) {
count = 0
    for (i in 1:x) {
      for (j in 1:y) {
        if(j <= y) {
          count = count + 1
        }
        j = j + 1
      }
     i = i + 1
    }
    return (count)
}
```

##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. (5 points)

```
cust <- read.csv('C:/Users/samsc/Desktop/ADS-500B/Data_and_Code2/custdata.tsv',
header=T, sep='\t')
##cust

library(ggplot2)
ggplot(cust, aes(x=income)) + geom_density(fill='blue') +
scale_x_continuous(labels = scales::label_number(accuracy = 0.01)) +
labs(title = "Density plot for income",
      x ="Income",
      y ="Density")
```



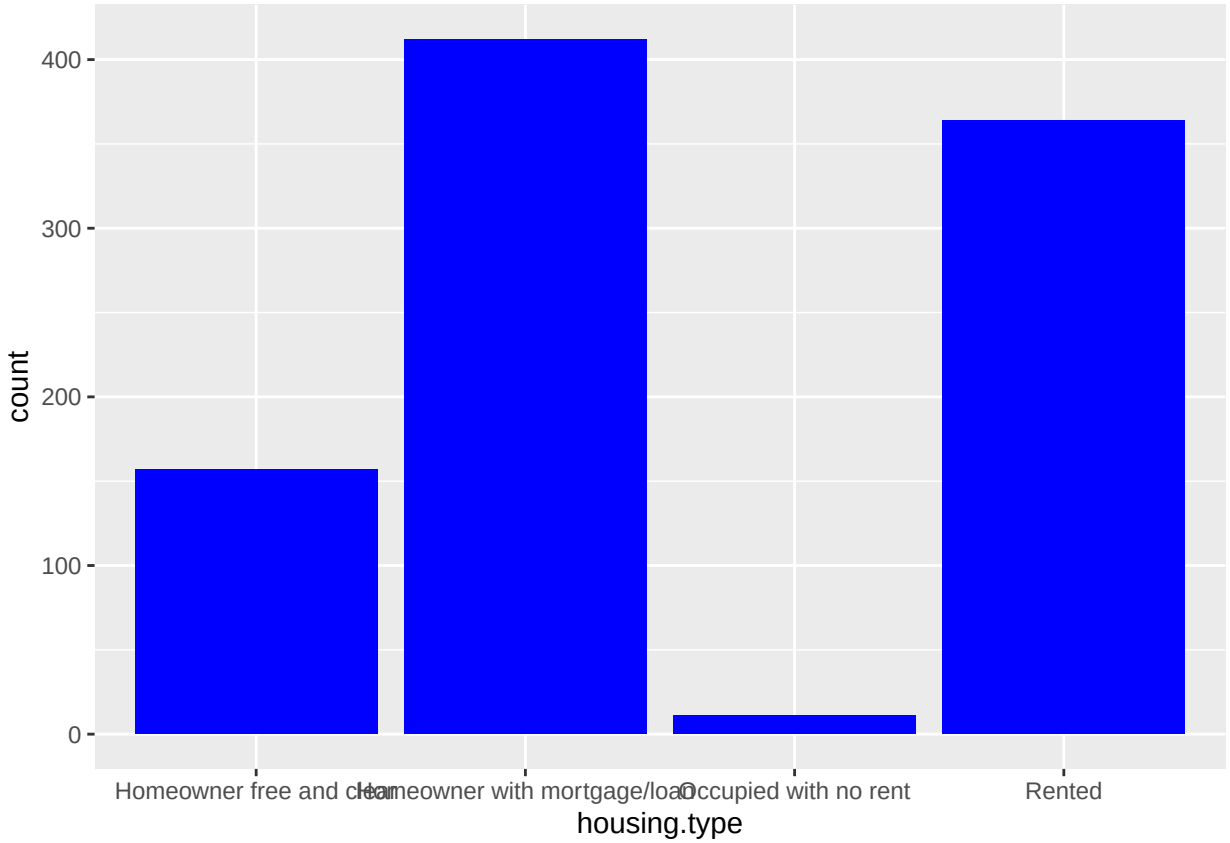

```
#used gemini to help scale x values to readabale using scale_x_continuous
```

###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

```
##The density plot for income shows that most of the customers do not make over
# 100 thousand dollars in income. The density plot peak is near the
#beginning of the data set, which means that there is a positive right skew in
#our dataset. This also means that there are outliers of people that make a lot
#of money and the targer audience is middle class people.I would change the
#x-axis values to be smaller in value for the customer data set to capture more
#of the observations.
```

###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.

```
cust <- subset(cust, !is.na(housing.type)) #(!) means remove I suppose
ggplot(cust) + geom_bar(aes(x=housing.type), fill='blue')
```



###.Using the customers data (custdata.tsv). ## 4.1: Extract a subset of customers that are married and have an income more ##than $50,000.

```
cust_married_fiftyk <- subset(cust, marital.stat == "Married" & income > 50000)
##cust_married_fiftyk
```

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

```
health_ins_col <- cust_married_fiftyk[,4:6]

round(sum(health_ins_col$health.ins) / length(health_ins_col$health.ins), 2)
```

```
## [1] 0.96
```

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

```
health_ins_col_whole <- cust[,4:6]

round(sum(health_ins_col_whole$health.ins) / length(health_ins_col_whole$health.ins), 2)
```

```
## [1] 0.85
```

##5. Using the customers data (custdata.tsv). ###5.1: Based on your own observations of the data, do you think there is any ##correlation between age, income, and number of vehicles? Explain why or why ##not.

```
#There does not seem to be a strong correlation between the 3 variables. Age,
#measure of income, and number of vehicles seem to be all relative and unrelated
#to one another. Job type does not factor age for being qualified for a certain
#job and number of vehicles and is not determine by how old an individual is.
#Although it might make sense for vehicle amount and income to be related or
#correlated, there are a lot of of 0 income observations with multiple vehicle
#counts. Due to these reasons I can conclude that there is not a strong
#correlation between age, number of vehicles, and income.
```

###5.2: Calculate the correlations between age, income, and number of vehicles, #and share your interpretations of the quantitative results. [Hint: Make sure to #2 of 2 remove invalid data points, otherwise you may get incorrect answers!]

```
##cust['num.vehicles']

library(tidyr) #used for the drop_na function i found at geeksforgeeks website

dropped <- drop_na(cust, age, num.vehicles, income)

cor(dropped[,c ("num.vehicles", "age", "income")])
```

```
##              num.vehicles         age     income
## num.vehicles   1.00000000 -0.06366646 0.13919800
## age           -0.06366646  1.00000000 0.01714168
## income         0.13919800  0.01714168 1.00000000
```

###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:

```
dating <- read.csv('C:/Users/samsc/Desktop/ADS-500B/Data_and_Code2/dating.csv')
##head(dating)
```

#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.

```
cor(dating$Games, dating$Icecream) #0.008
```

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

```
#There is not a strong correlation between playing games and eating ice cream.
#It is a very weak correlation meaning they barely effect each other at all.

cor(dating$Games, dating$Miles) #0.466
```

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

```
#Playing games and traveling has a stronger correlation in comparison to playing
#games and eating ice cream. This correlation is positive meaning that when one
#person plays more games they travel more, and vice versa.
```

#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.

```
regr <- ggplot(dating,aes(x=Miles, y=Games)) + geom_point() + stat_smooth(method="lm")
lm(Miles ~ Games,dating)
```

```
##
## Call:
## lm(formula = Miles ~ Games, data = dating)
##
## Coefficients:
## (Intercept)        Games
##       17824         2410
```

```
##Line equation: Games = 17824 + 2410*Miles
```

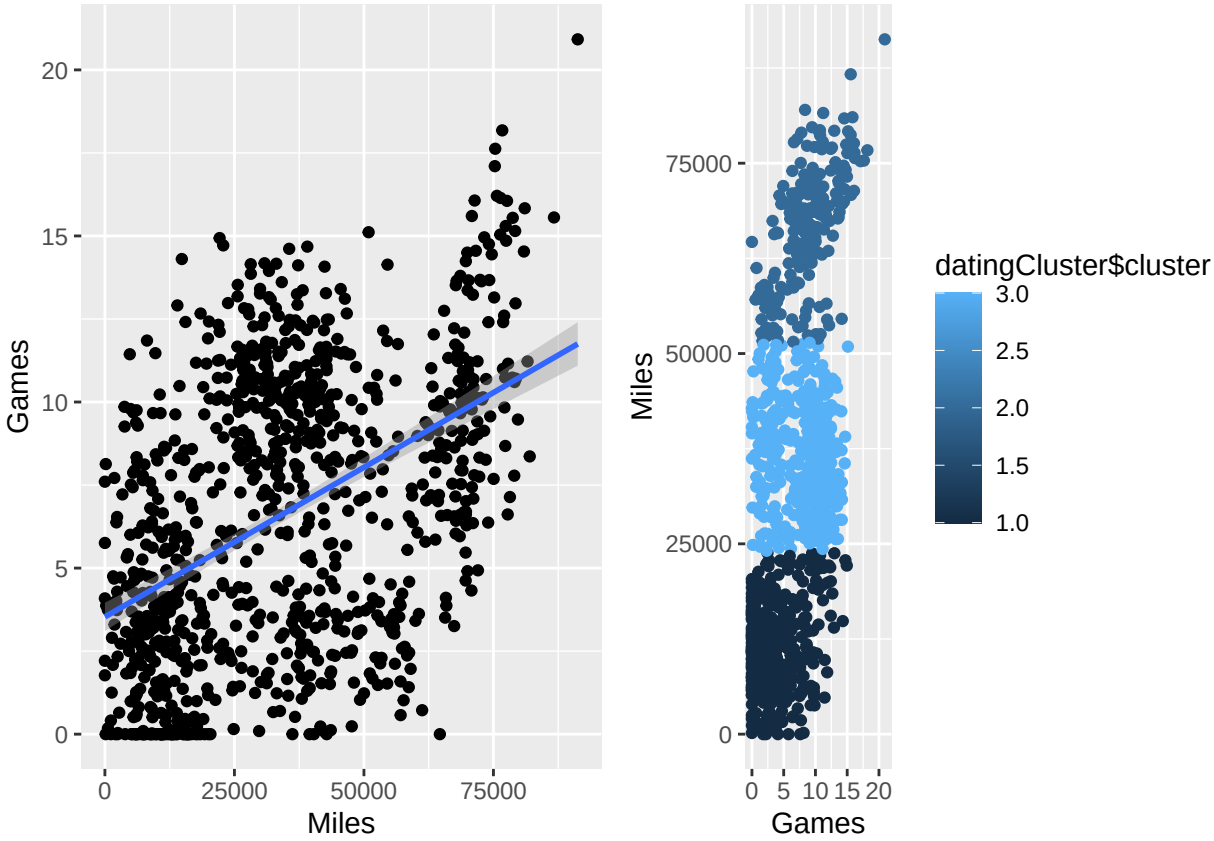
###6.3: Now let us see how well we can cluster the data based on the outcome #(Like). Use Miles and Games to plot the data and color the points using Like. #Now use k-means to create 3 clusters and plot the same data colored by cluster. #Show the plot and compare it with the previous plot. Provide your thoughts #about how well your clustering matched the Like variable in two to #four sentences.

```
set.seed(1)
datingCluster <- kmeans(dating[,1:2], 3)
clusterPlot <- ggplot(dating, aes(Games, Miles, color=datingCluster$cluster)) +
geom_point()

#installed gridextra to have plots show side by side. Credit to stackoverflow
#for showing me this

library(gridExtra)
grid.arrange(regr, clusterPlot, ncol=2)
```

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




```
#The cluster plot shows how well the data clusters around the linear regression
#line in the regression model. The blue line in our regression model gives a
#line of bestfit as to what kind of relationship is represented between Games &
#Miles. The cluster colors change from black to light blue representing
#how well the cluster data clustered around the points where the linear
#regression line is located on the regression plot. Light blue means that the
#data clustered well around those points while black means that the data was not
#clustered around the linear regression line.
```