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. **(5 points)**

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)**
   - 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. **(5 points)**

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. **(5 points)**

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. **(5 points)**
   - 4.2: What percentage of these customers have health insurance? **(5 points)**
   - 4.3: How does this percentage differ from that for the whole data set? **(5 points)**

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. **(5 points)**

- 5.2: Report your correlation numbers and interpretations. [Hint: Make sure to remove invalid data points, otherwise you may get incorrect answers!] **(10 points)**

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. **(10 points)**
   - 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. **(10 points)**
   - 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)**