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.

Dataset (located in your assignment prompt in Canvas).

- The dataset is customer data related to health insurance. The data set file name is “custdata.tsv”. You will use this data set to answer the questions in the assignment. Field names (in order): custid, sex, is.employed, income, marital.stat, health.ins, housing.type, recent.move, num.vehicles.

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

- 1.1: Figure out how to plot density of income. **(10points)**
- 1.2: Provide a couple of sentences of description along with the plot. Imagine you are explaining this to your manager or a senior leader. **(10points)**

2. Create a bar chart for housing type (x axis) and income (y axis) using the customers data. Make sure to remove the “NA” type. [Hint: Remove “NA” type from the “housing.type” and “income” columns by using the subset function with an appropriate condition on the “income” and “housing type” field. Use ggplot and add “+geom_bar(stat=’identity’)” after the initial ggplot command.] Provide your commands and theplot. **(10 points)**

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

- 3.1: Extract a subset of customers that are married and have an income more than $50,000. **(5 points)**
- 3.2: What percentage of these customers have health insurance? **(10 points)**
- 3.3: How does this percentage differ from that for the whole data set? **(10 points)**

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

- 4.1: In the customers data, do you think there is any correlation between age, income, and number of vehicles? Explain why or why not. To determine this:
  - Remove any age outliers. Assume age > 100 is an outlier.
  - Then plot a scatterplot between 2 of the 3 variables mentioned where the color of each point is the 3rd variable. [Hint: use ggplot for this].

    (10 points)
- 4.2: Report the best correlation coefficient between "income” and

“num.vehicles”, and “num.vehicles” and “age”. To determine this:

- Remove “NA” type from the “housing.type” and “income” columns by using the subset function
- Then determine the correlation coefficient between the columns "income” and “num.vehicles”, and “num.vehicles” and “age”

(15 points)