Assignment 1

Due 02/01/2019 by beginning of class

Instructions:
You may work with others on these problems but you must turn in your own work. You may type your solutions in any way you like (LaTeX, Markdown, Office, etc...) as long as you present your work clearly and in an organized way.

Unless otherwise specified, you must hand in a printed copy of your work at the beginning of class.

Problems:

1 Complete the course "Introduction to R" from Datacamp (`https://www.datacamp.com/courses/free-introduction-to-r`). You are only expected to complete the free part of their course, and is designed to give you practice with writing basic R commands and familiarity with R types especially dataframes. Download and print the statement of accomplishment and hand this in for credit.

2 Download the toy data from the website: `https://cgrudz.github.io/teaching/stat_757_2019/assignments/19_02_01/data_19_02_01.zip`

The comma delimited file data.csv contains information from the event data recorder in a car driven by two drivers, Alan and Barbara. Records are sampled every 30 seconds during a trip, and include information on the location, speed, and heading of the vehicle. Each record also includes a field called "event", which denotes whether the driver exceeded a pre-defined acceleration threshold at any point during that 30-second period. We assume that these events are correlated with risky driving behavior.

The file data-dictionary.txt contains a brief description of each variable included in the data set.

The data is too long to analyze by hand, but luckily we can use R to compute the following:

2.1 Identify how many total "events" Alan had. Explain how you extracted and calculated the total number of Alan's events data from the dataframe.

There are many ways to solve this, but because we will need to perform some analysis on each of Alan and Barbara in the following steps, I recommend storing these datum as separate variables, e.g.,

$$alan \leftarrow data[data\$driver == \text{`Alan'},] \tag{1}$$

$$barbara \leftarrow data[data\$driver == \text{`Barbara'},] \tag{2}$$

so that each is the sub-dataframe that contains only rows that correspond to either Alan or Barbara. From here we can find the number of 'Y' entries with a similar command, extracting only the vector of events with vector entries associated to 'Y' and finding the length,

$$a\_event \leftarrow alan\$event[alan\$event == \text{`Y'}] \tag{3}$$

$$a\_event \leftarrow length(a\_event) \tag{4}$$

The variable a_event is the total number of 'Y' events and is equal to 40

2.2 Identify how many events Alan had relative to the *total number of measurements* of their trips. Explain how you extracted the total number of measurements of Allan's trips from the dataframe.

Given the dataframe *alan* from the previous problem, we only need to enter the command,

$$a\_total\_measurements \leftarrow length(alan\$event), \tag{5}$$

to store the total number of measurements of Alan in the variable $a\_total\_measurements$. The relative number is computed from

$$a\_event/a\_total\_measurements \tag{6}$$

and equals approximately 0.1793722.

2.3 Identify how many events Alan had relative to the *total number of trips* they took. Explain how you extracted the total number of trips Allan took from the dataframe.

There are several ways to find the number of unique Trip ID numbers — we will look at two. The simplest way is to use the "unique" function as follows,

$$length(unique(alan\$trip\_id)) \tag{7}$$

with the output equal to 14. We can also treat the Trip IDs as categorical labels, and find the answer as follows:

$$alan\_trips \leftarrow levels(as.factors(alan\$trips)) \tag{8}$$

$$length(alan\_trips) \tag{9}$$

In this case it's actually easiest to work with the data in its original type, but sometimes you can make your analysis easier by changing type... The number of risky events for Alan relative to the number of unique trips is around 2.857143.

2.4 Find the mean and median speed of all of Allan's trips. Explain how you computed these summary statistics.

The easiest way to solve this is with the *summary* function, as you should usually try on any new data set. There are also built-in *mean* and *median* functions which you will often use, but my preference is using the command,

$$summary(alan) \tag{10}$$

The mean speed is around 27.940 and the median is around 26.250.

2.5 Repeat 2.1 but for Barbara.

15

2.6 Repeat 2.2 but for Barbara.

$15/625 = 0.024$

2.7 Repeat 2.3 but for Barbara.

$15/20 = 0.75$

2.8 Repeat 2.4 but for Barbara.

Mean=29.810, Median=29.375

2.9 Does Allan or Barbara seem to be a more risky driver? Why? Extra credit if you provide additional quantitative analysis of the data.

In total number of risky events, and the normalized relative number of risky events per trip and per length of time driving, Alan has a higher number. This is sufficient to conclude that Alan is riskier. Barbara has a slightly higher speed in some cases, but nothing alarming if you look at the locations and see that Barbara is driving on the highway in the San Jose, CA area.

Hint 1: There are *many ways* to solve problem 2 above, but I haven't shown you exactly how to do this yet. It is your job to do some research on how to access different sub-vectors of dataframes. R has abundant documentation available — try some resources on my website or search for your own.

Hint 2: Did you know that the type "LOGICAL" can be used to index/ extract slices of vectors? Try setting up a statement based on what you want to be "TRUE".