HW2: Clarification for Q1 and Q2

***Updated Feb 5, 5pm PDT***

I think a lot of people have been confused by the wording of the first 2 questions on HW2. To further clarify the questions and what is being asked:

Q1) Either of the two interpretations below will be accepted, but please clarify in your your report which one you chose and any other assumptions you made.

Interpretation A:

* Aggregate percprof for each state by combining the values from each county in that state, and generate a "percprof by state" plot.  
  Note that since percprof is a fraction of popadults, when aggregating it, you will have to use a formula like (s = state, c = county):  
  ∀s,percprofs=(Σc∈spercprofc×popadultsc)/Σc∈spopadultsc

* In the assignment PDF, "total population" was meant to refer to some "total" count of population, not necessarily the "poptotal" column in the dataset.

Interpretation B:

* Plot percprof for each county *grouped by* state, using a suitable type of plot that can help show the properties of the distribution of percprof values found within that state.
* You can describe the relationship by analyzing different statistics that your plot illustrates, such as the mean or median value for each state, the spread of values for each state, states that have outlying counties, and so on.

Q2)

* This question is similar to Question 1, but you are plotting 3 attributes in a single graph, such as by using Facets and GGPairs.

* Percentage of School graduates and College graduates also need to be aggregated, like in Q1, over the total population for each state.

* For this question, you can analyze relationships such as the correlation between the attributes, and states which have the highest and lowest percentages for both attributes.

We'll use an uppercase .R extension for all R code files

Add GT account name at start of each submission files( i.e. in the content of the file , not in the name of the file )

Do I need to explore the effect of varying dataset size N in Q3? No

# Definition of professionals in hw2 part.1

For hw2, I am trying to understand the data set a bit better and I am wondering what professionals actually means in that context? help("midwest") does not offer much details, but it does not seem like it is just the employment numbers.

As I understand, it is the percentage of people who are employed as professionals.

As mentioned below, the percentage of professionals for a given county is located in the midwest$percprof column and is based on the midwest$popadults column (the total # of adults in the county).

Therefore, I believe the 3 columns c('perchsd', 'percollege', 'percprof') are percentages of 'popadults' and not 'poptotal'. This matches the semi-consistent column grouping scheme they have in the dataset (and I agree that it could've been documented better!).

# how do we make a histogram for discrete data?

In Q5 of HW2, it asks to make a histogram for color. However, color is discrete data, which can't be plot using hist or geom\_histogram. Should we just plot it with geom\_bar?

Yes, geom\_bar should do the job

# is there a way to add markers to a plot?

The actual values need to be displayed

you may need the geom\_text function, see for instance §3.3 of the ggplot2 book

HW2 Q1 - What is meant by "describe the relationship"?

For this question, you can think about and discuss different statistics that describe the relationship, such as the median value or professional employment among states, the average spread, outlying counties etc. Your answer will primarily depend on the type of graph you use for the problem.