Problem Set 2

Overview:

One of the goals for this quarter is to get you comfortable using git and GitHub. In this problem set we will be practicing more git/GitHub workflow basics, manipulating data in R, creating GitHub issues, and creating a plot using the ggplot2 library. We are asking you to create a git repository on your local computer which you will later connect to a remote repository on GitHub. This local repository will have an .R file where you will read in data and practice manipulating this data to later create a scatterplot using the ggplot library.

Part I: Command line & Git

1. Using your command line interface (CLI) (e.g. Git Bash, terminal), create a new folder called **lastname_ps2**. Be intentional about where you create this folder (hint: change directories to where you want to save this folder first). Then, change directory into the **lastname_ps2** folder.

   Write the commands you used here (to create the folder and change directory):

2. Turn **lastname_ps2** into a git repository and write the command you used here:

3. Use the `echo` command to output the text `"# YOUR NAME HERE"` and redirect it using `>` to a file called `problemset2.R` (hint: refer to example code in lecture). Write the command you used here:

4. Check the status of your repository. Write the command you used here:

   According to the output, under which heading is `problemset2.R` listed under?

5. What is the git command to check what changes (i.e., differences) were made to `problemset2.R`?

   If you run this command now, do you see an output? Why or why not?

6. Add `problemset2.R` to the staging area and check the status. Write the commands you used here:

   According to the output, under which heading is `problemset2.R` listed under?

7. Use a git command to compute the hash ID for `problemset2.R`. Write the command you used here:

   What is the hash of the blob object?

8. Use a git command to get the content, type, and size of the blob object. Write the commands you used and the outputs you got here:

9. Commit the file and check the commit log. Write the commands you used here:

   According to the output, what is the hash of your commit?

10. Use a git command to get the content, type, and size of the commit object. Write the commands you used and the outputs you got here:

Part II: Manipulating data in R

1. Open `problemset2.R` in RStudio to edit the file and remove the comment containing your name at the top of the file.
2. Load data from off-campus recruiting events by public universities
   `load(url("https://github.com/Rucla-ed/rclass2/raw/master/_data/recruiting/recruit_school_somevars.R`
3. Take some time to investigate the data.
   - How many rows and columns are there?
   - Check missing values
   - What variable(s) uniquely identify the data?
   - Create a 0/1 dummy variable **visited** of whether the high school received a visit or not
   - Filter observations of zero or more in-state recruiting visits by one university of your choice (hint: need variables starting with **visits_by_** and **state_code**).
     - `visits_by_100751` = University of Alabama
     - `visits_by_126614` = University of Colorado Boulder
     - `visits_by_110635` = UC Berkeley
   - Subset your data frame to include the following variables: school_type, ncessch, name, total_students, avgmedian_inc_2564, vists_by_[school]

Part III: GitHub

1. Check the changes (i.e., differences) made to `problemset2.R`. How can you tell if a line has been added or removed?
2. Check the status of your repository. Write the command you used here:

   According to the output, under which heading is `problemset2.R` listed under?

3. Add and commit `problemset2.R`. Write the commands you used here:

4. Log in to your GitHub account online and create a new private repository here: https://github.com/organizations/Rucla-ed/repositories/new

Name it **lastname_ps2** and do NOT initialize it with a `README.md` file. Paste the link to your repository here:

5. Connect your local **lastname_ps1** repository to the remote and push your changes. Write the commands you used here:

Part IV: GitHub issues

1. Navigate to the issues tab for the **rclass2** repository here: https://github.com/Rucla-ed/rclass2/issues

   Create a new issue titled "Problem Set 2 - YOUR NAME" and post any question you have about the class or problem set.

Part V: Plots using ggplot

1. Use the dataframe from part II to create a scatterplot of total enrollment by medican household income.
   - X-axis: `total_students`
   - Y-axis: `avgmedian_inc_2564`
   - Color: `visited`
   - Label your graph

Finally, add and commit this file you are working on (`problemset2.Rmd`) to your repository and push to the remote repository as well.

Part VI: How much time did you spend on this problem set?