# Final project: initial analysis

EDH7916 | Spring 2020

Benjamin Skinner

***NOTE*** *This assignment needs to be completed by the start of the next class. That means everything pushed to your remote GitHub repo before class starts. If you are unsure whether you have successfully pushed your changes, check the online version of your repo at GitHub.com. If you can see your changes there, I can see them too.*

*Remember, I encourage you to save your work, commit smaller changes, and push to your remote GitHub repo often rather than wait until the last minute.*

Now that you've selected a topic and data set for your final reproducible report, it's time to conduct a preliminary analysis. In short, I want you to use the skills you've learned so far to answer the questions you've set for yourself. Specifically, I want you to submit **two items**. Both should be stored in your `final_project` folder.

### (1) Markdown file on how to get data

Submit a Markdown (`.md`) file called `final_project_instructions.md` that explains how to get the data required for your analysis and where it needs to be placed in relation to your R script. The instructions need to be detailed enough that I can find, download, unzip (if necessary), and place the raw data files so that your script will run. Assume I'm me, but am not familar with your data (*i.e.* that I'll figure it out).

This means that it's not enough to say, "Get these files from IPEDS," and then list files. You should include a link to the proper webpage and instructions for any clicking that I need to do to get your data.

### (2) R script

Create an R script called `final_project_analysis.R` that does the following:

1. Reads in your data — if your analysis requires combining multiple data sets (*e.g.* IPEDS data files), then you need to read in and append/join all files as required to construct your analysis data set
2. Selects the variables you need for your analysis (if your raw data contains more columns than you need)
3. Filters rows due to:
    - missing values
    - values that are outside the scope of your analysis (*e.g.*, you are using national data, but only need information about Florida)
4. Mutates (adds) new variables that you need for your analysis
5. Saves your prepared analysis data frame to your `data` directory — you can save it in whichever format makes sense to you: `.csv`, `.rds`, `.rda`, *etc*

Your R script should also have a section that:

1. Generates basic descriptive/summary statistics that are relevant to your analysis (*e.g.* if using IPEDS to show average graduation rates among Florida public colleges and universities, show average graduation rates for each year in your sample). For this part of the assignment, you do not need to produce any formatted output (tables, figures, text, *etc*): we'll work on that after the break. Instead, you can just have results printed to the console.

2. **EXTRA**: If you have familiarity with more advanced inferential techniques such as regression and think you may use one in your report, feel free to fit a model. That said, we have not covered this yet and it is not required. But if you are able to do the other steps and an want to move ahead, then go for it!

Your R script should be well-organized, following our template. It should be well commented throughout, explaining what you are doing, particularly if making subjective data cleaning decisions or writing your own functions. It should also run from top to bottom without error or requiring that the user change anything.

I don't expect that your full analysis will be completed in two weeks. You may end up changing some analyses or adding new ones. But I do expect you are able to put your data together and produce at least a couple of statistics. I fully expect that you will repurpose this code in your final report so the more you do now, the better off you'll be.