MASTERS IN DATA SCIENCE 597
FINAL PROJECT
SPRING 2018

1. Project

For the final project for the course, your assignment is essentially to wrangle some data and to show off your skills. I am intentionally not specific about how you do so, but you have the weekly assignments as models. Think of the project as the equivalent of chaining together multiple weeks of assignments: you should bring data into R, clean it, tidy it, perhaps create new variables, perhaps summarize your data, and report on it with tables and figures. However, there are some required elements:

- You must get your data from at least two distinct sources, at least one of which must be at least somewhat difficult to work with (requires scraping or cleaning).
- You must use Git and Github to manage your project.
- All of your code and the R Markdown file should run in its own directory, without any additional files or code.
- *Every* code chunk must be labeled.
- You must include a step where you save a tidy version of (perhaps just some of) your data as a csv file. The idea is that the csv file would be an easy place for someone else to start from.
- Your report, generated from an R Markdown file, should be as good looking and well formatted as you can make it—that includes tables and figures. Do not use "echo = TRUE" except as truly needed.
- We have not done statistical analyses more sophisticated than correlation and linear regression in this course and there is no need for it in your report. You can do so if you wish, however.
- If some parts of your project are relatively easy, you should balance that out by going into more depth in other aspects.
- Your report should explain the steps you've taken and why—I do not want to see just a collection of tables and figures. Feel free to describe approaches that didn't work or were more troublesome than expected.
- I expect that you will discuss this project with others, but please avoid using datasets in common (I realize that might still happen by coincidence). All of the work submitted must be your own. Be sure to credit the sources of your data and any other material—it is better to over-credit than to under-credit. If you have any questions about properly crediting others' work just see me about it.

2. Presentation or Written Project

About half of the class will give a 5-10 minute presentation of their projects during our last class on April 30 (think 5-10 slides). Besides the presentation, those students will turn in their slides and other components required for their project. The other half of the class will not give a presentation but will produce a formal report. Students who present will have until the end of that week to turn in their project. Students who hand in a formal report will turn in their project at the time of the last class.

In any case, focus on why you were interested in the datasets, some of the issues in wrangling it, and a few interesting figures or tables. While keeping in mind that what was time-consuming for you may not be interesting for others, remember that the course has emphasized mechanics and that your classmates may very well be interested in, say, what regular expression you used to reformat a particular column.

3. Procedures and Dates

Submit (via Sakai) a short description of your data and plans for it by April 10 at 2 pm. I will also ask for your preference as to presentation versus report as part of that "assignment." The description should include links to your data sources. There is no grade associated with this part.

You will submit your final project (via Sakai) by giving the URL to clone your GitHub repository. Also submit any api keys required using the format

```
api.key.<your last name>.<anything else> <- "abcdefg".
```

Your final project will be graded wholistically, but I will be looking at these elements

- that you have demonstrated your ability to use R to accomplish your tasks
- that your code is easy to understand
- that your report is well written with well-presented tables and figures