MASTERS IN DATA SCIENCE
597FINAL PROJECT
SPRING 2022

1. Project

For the final project for the course, your assignment is essentially to wrangle some data and to show off your skills. You can view 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, analyze and interpret your data, and report on it with tables and figures. However, there are some required elements:

- This will be a group assignment, where the group size is 2-3. You will team up with 1-2 of your classmates for this project. Please let me know if you are unable to find a team. In that case, I will add you to an existing team.
- You must state an overall goal of your data analytics project: the broad goals, what you want to investigate and why, and what your data sources are.
- You must use Git and Github to manage your project. Example:
  - https://github.com/kosukeimai/
- All of your code and the R Markdown file should run in its own directory.
- *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
- The level of statistical sophistication expected in your project will not be high. For example, simply summary and graphical representations, scatter plots, basic analysis like linear regression or time series analysis would suffice as

long as you connect it to the objectives stated in your study clearly. You can, however, do a sophisticated statistical analysis if you wish, but mindless and incorrect applications are likely to be heavily penalized.

- Your report should explain the steps you've taken and why—and not be just a collection of tables and figures. Codes, tables and figures without context or explanation are likely to be penalized. Feel free to describe approaches that didn't work or were more troublesome than expected
- Your team can discuss this project with other teams, 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.

2. Presentation

Each team will be expected to give a 10-15 minute presentation of their projects during our last two classes on April 25 and May 2 (about 10-15 slides should be enough). Besides the presentation, you will turn in your slides, R markdown report and other components required for your project. The last date for turning in your projects is May 6, Friday. Although in-class presentations are encouraged, virtual presentations should be possible.

Your team will be randomly assigned to one of the presentation slots spread over these two days. If you are unable to make a presentation on either of these days, or have a strong preference for a virtual presentation, you should let me know by March 31.

Your presentation should focus on your objectives, why you were interested in the datasets, some of the issues in wrangling it, a few interesting figures or tables and your conclusions. These presentations would provide all of us with an excellent opportunity to learn from each other.

3. Procedures and Dates

Submit (via Canvas) a short description (1/2-1 page) of your data and plans for it by March 31 and let me know if your team has any issues presenting on April 25 or May 2. The description should include links to your data sources. There is no grade associated with this part.

You will submit your final project (via Canvas) by giving the URL to clone your GitHub repository. Also, include any api keys required to access your data.

Your final project will be graded holistically, but these elements will be considered:

- That you have put in some effort to obtain the data
- that you have demonstrated your ability to use R to accomplish your tasks
- that your code is easy to understand
- that your presentation has a logical flow with well-presented tables and figures