

**DATA WRANGLING AND MANAGEMENT WITH R**  
**FINAL PROJECT**  
**SPRING 2023**

1. PROJECT

For the final project for the course, your assignment is essentially to wrangle some data and to show off your skills. The final project is worth 15% of the final grade, or about two and a half weeks worth of regular assignments—that should give you a good idea of how much effort to plan to put into it. For the project you will want to bring data into R, clean it, tidy it, perhaps create new variables, perhaps summarize your data, and report on it with tables and figures. Furthermore, 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 (requiring scraping, cleaning, or an API). It is not strictly required that the datasets be related, but your grade will certainly be better if they are.
- Your R Markdown file, your code, and any data files should be self-contained. That is, it should run on my computer or Posit Cloud account without modification (remember that you do not want Jenny Bryan to set your computer on fire).
- Your report should be in the form of an R Markdown file, not a notebook, script, or some other format. You should see `output: html_document` near the top of the document.
- 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 show your code (that is, use “`echo = TRUE`”) other than when 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.
- It is very important that your report should explain the steps you’ve taken and why—I do not want to see just a collection of tables and figures. Of course, that text should be outside of the code chunks, and should not simply be section headers. Feel free to describe approaches that didn’t work

or were more troublesome than expected. Only code and the occasional comment should be inside code chunks.

- 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 contact me about it.
- I suggest that you not use data from Kaggle. For most of Kaggle's datasets, the data is already in very good shape, plus the datasets tend to be used very widely.

## 2. PROCEDURES AND DATES

Submit via the Canvas link a short description (one or two paragraphs) of your data and plans for it by **April 15 at 10 am**. The description should include links to your likely data sources. There is no grade explicitly associated with this part, but it will harm your project grade if you somehow omit this step. If you do not hear from me you can assume that your description is approved.

The project itself is due at our scheduled final exam date, **May 4 at 9:00 am**.

You can submit your final project via Canvas, as a GitHub repository, or by asking me to duplicate your Posit Cloud project. Via Canvas is best if you have just an Rmd file and possibly a few small data files. I will have more details about these options later. *Be sure to include any required API keys.*

Your final project will be graded holistically, but the TAs and I will be looking at these elements

- that you have demonstrated your ability to use R or SQL, and in particular the methods discussed in this course, to accomplish your tasks
- that your code is easy to understand
- that your report is well written with well-presented tables and figures.

## 3. A FINAL NOTE

You will both enjoy the project more and you will do better if you pick something that you are genuinely interested in. You may also find yourself expanding the project somewhat so that you can brag about it in a job interview—it's hard to make someone interested in a project that you are not interested in yourself.