FDSII Final Project Guidelines

Work in groups of 2 or 3. Please self-select these groups ASAP. Once you have your group, create a public GitHub repository for it (one of the team members will take the lead in doing this using their GitHub account). This way you can collaborate on writing the code via GitHub pull requests (PR). Having a GitHub repo to which everybody has contributed is a requirement for the project.

Basic project types

- Analyze a publically available dataset to answer an interesting question or questions.
- Do a simulation of some phenomenon.

Analyzing an existing dataset (or sets)

The internet is full of datasets! As we have seen, we can copy and paste tables directly from, e.g., Wikipedia. There are also internet repositories of data. Kaggle is a great one (you need to sign up for an account but it's free). The open-access journal PLOS has a list of recommended repositories. And there are many others.

As we've already seen in class, getting a dataset in shape for analysis can take a bit of work, and could easily take over half of the total time to do the project. (It's like painting a car or a house: the majority of time is spent cleaning and scraping and masking off the windows, etc., and then the actual painting goes really fast.) So deciding on a dataset and getting going on the data wrangling should get started early!

Simulation

Another route to go is to do a simulation project. Earlier in the class (and in FDS I), we did some simulations of elections. It might be interesting to, say, do a realistic simulation of the upcoming presidential election based on real polling data and the real electoral college system. Or you could simulate the spread of a disease in a population depending on various behaviors of the population (rates of masking, social distancing, etc.). Or you could simulate a sports tournament, a war-game, etc. If you have an idea for a simulation, but have no idea how to get started, let's chat!

Deliverables

A Written Report

The written report should should be co-authored by all of the members of the group. This collaborative effort can take several forms. Each person can right a draft of a section or sections, and then you can edit each others sections, or one person can write a rough draft of the whole thing, and the others can edit and polish the final project.

The format of the written report is somewhat flexible. But all of them should have

- a title page

- a general introduction section: this should describe the project or question and motivate the reader to be interested in it.

- a methods or analysis section that describes what you did

- a results section that describes what your found (emphase the figures)

- a conclusion section: this should summarize the findings and convey the take-home message.

- a references section for any literature on the topic you cite.

Many of you will recognize the above as being APA format, but your figures should be in the document (with captions) rather than tacked on at the end.

There should be no code in the written report. It should be text and figures only.

A Jupyter Notebook (or Notebooks)

The notebook(s), that you used to do your analyses. These don't need to "submitted", they just need to be in your GitHub repo. Where appropriate, text can be common to both the notebook at the report.

A Presentation

The presentation will be a ~10 minute (so 5 to 10 slides) description of your project given to the class during the last week of classes. It should be basically a presentation of the written report, which means that the slides should consist of the figures from the written report and perhaps some sparse bullet-point slides to assist with the narrative.

The presentation will also be a team effort. You can divide it up however you wish, but everyone on the team should describe some component of the project to the class. For example, one person could give the introduction, another person could describe the findings, and the third person could wrap things up with the conclusions.