Group Projects: Overview and Timeline

It feels like the semester just started, but now that you've had some time to get acquainted with some of the core command line tools, it's time to start thinking about a project in which to apply your skills.

The first order of business is to establish groups for your projects and pick a data set to work on. Owing to remote instruction, we will form groups through the discussion board on canvas. We have created a thread titled "Forming Groups and Choosing Data" for this purpose. We ask that you form groups of at least three and at most five members. You are free to either form a group and then pick a data set, or find a data set that interests you and post about it on the discussion board to find other interested students to work with. You should feel free to post multiple data sets that you find interesting– posting about a data set does not obligate you to work on that data (but please indicate in your post whether or not you are committed to working with that data set). You should make sure that you form groups with students who will be able to collaborate with you, whether that be over video, chat or email. Toward that end, you may want to include information about what time zone you are in when you post to canvas, but this is not required.

You and your group should find a large data set that interests you. Your data set should be large enough to warrant the use of the compute clusters (either the Statistics Department or CHTC). Your data set should total at least 10GB in size, either consisting of many files or one large file that you can break up easily into chunks of at most 4GB each (preferably much smaller). The data should be in `.csv` format or some other format that you can read and process without difficulty.

Here are some good resources for finding interesting data:

- Kaggle: `https://www.kaggle.com/datasets`
- World Bank Data Catalog `https://datacatalog.worldbank.org`
- U.S. Government's Open Data `http://data.gov`
- HealthData.gov `https://healthdata.gov`
- Open Data Network `https://www.opendatanetwork.com`
- United States Census `https://www.census.gov/data.html`

- Project Gutenberg (thousands of free eBooks) `http://www.gutenberg.org/wiki/Gutenberg:The_CD_and_DVD_Project`
- American National Election Studies `http://electionstudies.org/data-center`
- Purdue datasets list `http://llc.stat.purdue.edu/2018/29000/datasets.html`
- Wikipedia `https://en.wikipedia.org/wiki/Wikipedia:Database_download`
- Forbes list of free data sources `https://www.forbes.com/sites/bernardmarr/2018/02/26/big-data-and-ai-30-amazing-and-free-public-data-sources-for-2018/#4936a3975f8a`

Of course, these are just suggestions. You can also search for your own data source elsewhere on the internet.

1 Timeline and Deadlines

- By **Friday, November 6, 11:59pm**, send an email to the instructors and the TA (`jgillett@wisc.edu`, `kdlevin@wisc.edu`, `bwu62@wisc.edu`), and copying all members of your group, listing the members of your group and their NetIDs, and the data set you are planning to work on. Designate one member of your group to create a github repository for your group, with all group members as well as the instructors and TA (github IDs `jgillett-605`, `kdlevin-uwstat` and `bwu62`) as collaborators.
- By **Friday, November 13, 11:59pm**, add a one-page pdf to your repository titled `project.pdf`, and submit a link to the corresponding commit (i.e., the commit corresponding to a snapshot before the deadline). The summary should include:
  - The names and NetIDs of all group members.
  - A description of your data set and a URL at which to access it.
  - One to three statistical questions that you intend to answer.
  - A short code snippet that reads the data onto your laptop. Of course, if you are working with a really large data set, then this code snippet should only grab a small piece of the data. Please contact the instructors if this is not easy to do (e.g., because the data is packaged in a single large file).
  - A description of the variables available, with more details for the variables that are relevant to your question.
  - A description of the statistical methods that you plan to use Of course, this is just a plan. If you later find that there are better methods, that is okay.
  - A description of the computational tools and the particular computations you expect to perform. Again, this is just a plan. It does not necessarily commit you to having to do precisely these computations.

- By **Friday, November 13, 11:59pm**, post to the discussion board with a description similar to that in your `proposal.pdf`, so that your classmates can see what data you are working with and what you are planning to do with it.

- During the **week of November 18**, read your classmates' project proposals, and post to the discussion board giving feedback.

- By **Friday, November 20, 11:59pm**, you should have given feedback on at least three other projects. To ensure that all projects receive feedback, we ask that you avoid "piling on" to threads until each project has at least three comments already. Your group should also respond to posts on your project as they appear.

- During the **weeks of November 25 and December 2**, schedule time to meet with the instructors and TA to receive feedback, ask questions, etc.

- By **Monday, November 30, 11:59pm**, submit the assignment `Project: First Draft`, consisting of a link to a commit containing a PDF titled `project.pdf`, that includes the names and NetIDs of all members of your group, and updated to consist of three sections:

  - An introduction that summarizes the data, the question that you chose to pursue, your statistical models and computations, and your conclusion. As a rule of thumb, a reader who reads only your introduction and then stops should still have a high-level understanding of what you did.
  - A body that describes your data in more detail (e.g., its source, size, what cleaning steps were performed), your statistical models and computations, and your findings. Please describe any interesting (or not so interesting) findings using plots and tables. Be sure to describe in reasonable detail the tools (both statistical and computational) used in your project, including any difficulties you encountered.
  - A conclusion that revisits your question and conclusion and suggests future work.

  There are no strict length requirements for this document, but consider two to three pages to be a good rule of thumb (a bit longer, if you have a lot of plots and tables). There is, of course, no penalty for writing more, but a document longer than four pages probably indicates that you should try to condense things.

  Once again, if you did most of your writing in an R Markdown file, your PDF may be converted from, say, `report.Rmd`. Please contact the instructors for help with conversion, if needed.

- By **Monday, November 30, 11:59pm**, post your `project.pdf` to the discussion board.

- During the **week of December 2**, read the other groups' projects and give feedback, as with the proposals.

- As with the proposals, by **Friday, December 4, 11:59pm**, you should have given feedback on at least three other projects. To ensure that all projects receive feedback, we ask that you avoid "piling on" to threads until each project has at least three comments already. Your group should also respond to posts on your project as they appear.

- During the **week of December 8**, you and your group should incorporate feedback received on the discussion board and from the instructors and the TA. Schedule time to meet with the instructors and TA for further feedback and questions, as necessary.

- By **Wednesday, December 11, 11:59pm**, submit the assignment `Project: Final Draft`, consisting of a PDF titled `project_final_draft.pdf` that includes the names and NetIDs of all members of your group, and incorporates the feedback from your peers and instructors.