DSCP Group Project

Timeline

The project is worth 50 points, with intermediate deadlines as listed in the schedule:

- Form a group (0 points) of 4-5 students:
  - To choose your own group, go to our Canvas page's People tab, click on the Project tab, and put your group members into "Project n" (where n is the lowest number from 1 to 30 that is not already in use by another group).
  - Right after this deadline, we will randomly assign students to groups for those who do not choose their own groups.
- Write a one-page proposal of no more than 250 words (4 points) including (the most important one line of) code to read data; descriptions of the variables, statistical methods, and computational steps you will use; and a link to your `github` repository. Turn in a `proposal.html` (knitted from `proposal.Rmd` or from some other source) or a `proposal.pdf`, one per group.

  Write a discussion post to claim your data set by clicking on "Project proposal data set" in the "Discussion" tab of our Canvas page.
- Meet as a whole group with teacher and TA for 5-10 minutes in class to discuss your proposal (3 points).
- Turn in Presentation slides as `presentation.pdf` (or `.html`), once per group.
- Presentation (25 points) to class.
- Attend peers' presentations and give feedback (8 points allocated as 4 points per day).
- Report (10 points) of no more than 600 words with supporting graphics. Turn in a `.pdf` or `.html` file, one per group.

Data

Find a large data set on the internet that interests you.

- Your data set size should be large enough to warrant the use of the compute clusters (either the Statistics HPC or the CHTC)–about 10 to 100 GB.
- It should consist of many smaller files (less than about 4GB each), or one file you can break up easily, or a file or files for which you can run a parallel computation using course tools.

Here are links to many data sources: `www.stat.wisc.edu/~jgillett/DSCP/project/dataLinks.pdf`

Pick a question about one or several variables in your data set about which you are curious. Choose a data set and question different from your peer groups' choices. **Do original work.** Include citations (URLs, etc.) for data, graphs, and methods you get from others.

Statistical computing

Design a statistical computation that includes parallel computing. If possible, automate your computation, from data download to filtering and cleaning, exploring and modeling. (If this is not possible, include a decscription of what you did not automate so you and others can repeat what you did later.)

If you work on CHTC computers, please do not put large files on `learn.chtc.wisc.edu`. Instead, for each parallel job, download the file(s) needed for that job on the remote computer assigned to the job. After processing it, remove it, as otherwise HTCondor will copy it back to `learn` when your job completes.

CHTC staff made the directory `/home/groups/STAT_DSCP`, which is accessible from `learn.chtc.wisc.edu`, with space for making group directories. We can use it on projects, if necessary, to download a large file and break it up for use in parallel jobs.

CHTC has another solution for handling large files involving `/staging`, described at `https://chtc.cs.wisc.edu/uw-research-computing/file-avail-largedata#3-using-staged-files-in-a-job`.

Report

Write a report of no more than 600 words (about three pages of text, possibly extended by graphs) in three sections describing your data, variables, question, and statistical computation.

- Its *introduction* should summarize the data you analyzed, the question you pursued, your statistical computation, and your conclusion. It should outline the body of your report. A reader who quits after your introduction should understand your work broadly.
- Its *body* should describe your data (and its source, size, and cleaning), statistical computation, and results.
  - Include graphical and numeric summaries to efficiently communicate your conclusion.
  - Describe your statistical computation including the number of jobs you ran and the typical job time, memory, and disk space required.
  - Mention weaknesses of your work.
- Its *conclusion* should revisit your question and conclusion in the light of your report's body. It could suggest future work.
- Include a post-conclusion "Contributions" paragraph briefly describing the contributions of each group member. Here is an example (other "Contributions" designs are ok too):

  "

Member	Proposal	Coding	Presentation	Report
Lucy Van Pelt	1	1	1	1
Charlie Brown	1	1	1	1
Linus Van Pelt	0	0.5	0.4	0
Spike	0	0.7	0	0.3

Notes:

- In the table, 1 = full contribution, 0.1-0.9 = partial, and 0 = no contribution.
- Linus attended the presentation without preparation.
- Spike sent a video, but it was unrelated to our presentation slides.

"

Presentations

Make a presentation of 5-7 minutes that summarizes your report.

Use git/github

Use the `git` version control system to track changes, store your code at `github`, and manage collaboration accross members of your group.

- See, for example, `http://pages.stat.wisc.edu/~jgillett/DSCP/git/git.pdf` and `http://pages.stat.wisc.edu/~jgillett/DSCP/git/gitExercise.pdf`.
- Include your TA as a collaborator on your `github` repository.
- Include a line in your report on how to clone your repository, e.g.

  `git clone https://github.com/<ID>/DSCPproject.git`

Grading

Peer feedback on presentations will include voting/ranking that leads to three awards for

- best presentation
- most creative or interesting project
- best visualizations

These are some things will consider when grading your project:

- Does the project demonstrate knowledge of the course? Does the statistical computation make effective use of the HPC or CHTC?
- Is the report no more than 600 words long? (Paste your text into `https://wordcounter.net` to check, as we will stop reading at 600.)
- Is the question engaging?

- Is the analysis correct and persuasive? Where statistical methods are used, are their assumptions discussed?
- Are the graphical and numeric summaries informative? Are their fonts easily legible?
- Is the writing vigorous? (Strunk and White say, "Vigorous writing is concise. A sentence should contain no unnecessary words, a paragraph no unnecessary sentences, ...")
- Are report authors listed at the top?
- Are numbers rounded? "0.3 vs. 4.1" conveys more information faster than "0.337885 vs. 4.078801".
- Did each group member speak in each presentation, with reasonably balanced speaking times?