Group Project

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 4GB each) or one file you can break up easily.

Here are links to many data sources: `www.stat.wisc.edu/~jgillett/605/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.

Statistical computing

Design a statistical computation that answers your question. Parallelize your computation, if possible, to make it run reasonably quickly (less than 30 minutes, possibly on many CPUs). Automate your collection of data from the internet and your analysis.

- If you work with Slurm/HPC, put your code (but not your data) in a `projectHPC` directory. Make a `projectHPC.tar` file of that directory. We should be able to run your analysis by running `tar xvf projectHPC.tar; projectHPC/submit.sh`.
- If you work with HTCondor/CHTC, put your code (but not your data) in a `projectCHTC` directory. Make a `projectCHTC.tar` file of that directory. We should be able to run your analysis by running `tar xvf projectCHTC.tar; projectCHTC/submit.sh`.

  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.

  Christina made the directory `/home/groups/STAT_DSCP`, which is accessible from `learn.chtc.wisc.edu`, with a higher disk quota than our home directories. We can use it on projects, if necessary, to download a large file and break it up for use in parallel jobs.

Report

Write a report of no more than 750 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 chart above, 1 = full contribution, 0.1-0.9 = partial contribution, 0 = no contribution.
  - Linus attended the presentation without preparation.
  - Spike sent a video, but it was unrelated to our presentation slides.

Write your report in an R Markdown file, `report.Rmd`. Knit it to `report.html`.

Presentations

Make two presentations of about 4 minutes each related to your project:

- a proposal consisting of data, a question, and a suggested analysis
- a data analysis summarizing 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/605/git/git.pdf` and `http://pages.stat.wisc.edu/~jgillett/605/git/gitExercise.pdf`.
- Include your teacher and TA as collaborators on your `github` repository.

- Include a `README` in your `.tar` with how to clone: `git clone https://github.com/<ID>/605project.git`
- Test `git clone https://github.com/<ID>/605project.git` before turning in your project.

Timeline

The project is worth 40 points, with intermediate deadlines as follows:

- Fr 11/11/22: Form a group (1 points) of 4-5 students:
  - To choose your own group, 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) at `https://canvas.wisc.edu/courses/324552/groups#tab-37786`.
  - Right after this deadline, we will randomly assign students to groups for those who do not choose their own groups.

  Choose a data set of about 10 to 100 GB. Turn in a `group.txt` file containing the names and NetIDs of your group members and a link to your data set.
- Fr 11/18/22: Write a one-page proposal (5 points) including 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`). Present your proposal in class on We 11/23/22.
- Fr 12/2/22: First draft (10 points) of report. Turn in `draft.html` (knitted from `draft.Rmd`).
- Fr 12/9/21: Presentation (10 points) to class lasting four minutes. Turn in slides as `presentation.html` (or `.pdf`) by Fr 12/9/21. Presentations are in class on Mo 12/12/22 or We 12/14/22).
- Mo 12/12/22 and We 12/14/22: Attend peers' presentations (1 point allocated as $\frac{1}{2}$ point per day).
- Fr 12/16/22: Final report (10 points) of no more than 750 words with supporting graphics. Turn in `report.html` (knitted from `report.Rmd`).
- Peer feedback (3 points) on proposals and first drafts presentations, due three days after the respective deadlines.

Grading

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 750 words long? (Paste your text into `https://wordcounter.net` to check, as we will stop reading at 750.)

- 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".

- Does the `github` repository include commits indicating balanced contributions from all group members?

- Did each group member speak in each presentation, with reasonably balanced speaking times?