Biostat 625 Final Group Project

The final project will be completed in groups. The groups will be self-organized, with three to four students in each group. Each group will identify a big public dataset of interest, propose a question of interest related to the dataset of choice, perform analysis on the dataset of choice to address the question of interest. It is expected that there will be some computational challenges involved in the project, which require the use of some computational techniques learned in (or outside) the lectures of Biostat 625.

Grading: The following grading criteria will be used (6 points each item × 5 items = total 30 point. For each item, 1: fail, 2: poor, 3: satisfactory, 4: good, 5: very good, 6: excellent):

1. The dataset consists of real-world data, and the question of interest is of real-world relevance.
2. The computational challenge is real, and the approach that was taken is appropriate and satisfactory.
3. The R (or other) computer code is clearly written and well-organized. The coding style is neat and consistent. The comments are clearly written and helpful.
4. The report is clearly written and sufficiently introduces the dataset of choice, the question of interest, the approach taken, and the results.
5. The amount of work is not trivial.

Furthermore the following requirements should also be satisfied if possible (otherwise a penalty may be applied):

- R Markdown should be used to write the report for the final project.
- The report should be no more than 5 pages.
- The contributions from each of the group member should be clearly stated (although all the students in a group will receive equal points for the final group project).
- A GitHub repository should be used for the group members to collaborate on the final project.
- Graphics (ideally with ggplot2) should be used to present the results.

Submissions (no late submission allowed):

- Final group project proposal (due date given on Canvas): one page proposal (not graded).
- Final group project proposal presentation (in class): (not graded).
- Final group project report and computer code (due date given on Canvas): link to GitHub repository and report up to five pages (will be graded).

Sample public datasets (other self-selected datasets are also highly recommended):

- CDC: `https://www.cdc.gov/DataStatistics/`.

- Census: https://www.census.gov/data.html.
- EPA: https://www.epa.gov/aqs.
- NCI: https://portal.gdc.cancer.gov/.
- Insurance (synthetic): https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.
- MovieLens: https://grouplens.org/datasets/movielens/.
- Airlines: http://stat-computing.org/dataexpo/2009/.
- Kaggle: https://www.kaggle.com/datasets.