Proposal due: Tuesday, February 11th 2025 11:59 pm on Canvas
Report due: Friday, April 4th 2025 11:59 pm on Canvas

Summary

- For undergraduate students, the minimum group size is two, and the maximum size is three. You will be forming your own groups. Graduate students will submit their own projects.
- Each group will submit a proposal and report that is outlined herein.
- The proposal and report will be 5% and 25% of your final grade, respectively.
- The proposal and project must be completed using RMarkdown or KnitR (LaTeXand R). Failure to do so will result in a 20% deduction of the total possible grade. Overleaf now supports KnitR documents (.Rnw files), so you are able to work collaboratively.
- Generative AI can be used to aid you in writing this report or writing code, however Generative AI writing is poor for technical writing of regression analysis. It is also prone to making egregious errors in coding for statistical models. I suggest you use it to help you outline your writing or support your coding, but not to use it to write sections of your paper or whole chunks of code.
- The dataset to be analyzed should be a "live" dataset (i.e. must not have been previously analyzed in a textbook or paper). One location you could look for data is the UCI Machine Learning Repository. Data from a previous co-op job is a great idea, but make sure you have permission to use it. If you build your own dataset (from, say, the internet), I will be impressed and interested. You must reference where your data comes from if you group doesn't create the dataset, and describe the data collection process as part of your report.
- Any methods beyond class work must be cited, and any R-packages used must be cited. Any facts or opinions that which are not your own must be cited. Any methods from the course do not need to be cited. If you use a method from one of the textbooks, but not used in class, it must be cited. Data source must be cited.

The purpose of this project is for you to:

- apply concepts from class to conduct modelling and analysis on real data,
- learn to write a statistical report, and
- develop your scientific communication and collaboration.

More specific details on proposal and report expectations:

Proposal

- Guidelines: The purpose of the proposal is to find a dataset to analyze and summarize the data, identifying any challenges that may exist within that dataset. The proposal will contain:

  - a statistical description of the dataset, including but not limited to data types, structures, distribution, and so on,
  - who, what, when, where and how the data was collected and any underlying scientific processes that may affect the data,
  - what scientific questions about the data that you will try to answer, and
  - if you plan to collect your own data, it must be constructed or in construction by proposal submission.

- Length: 2 pages max. Font size is 10. If you have to change the font size to condense the proposal to less than two pages, the proposal is too long.
- Submission format: Proposal submissions will be in pdf file format only. You will also submit to me your dataset for quality assurance. I will be giving feedback on your projects—the better you write the proposal, the better the feedback and higher probability of success on the report.
- Post-submission meeting: After submission, each group will schedule a 30-minute meeting with me to discuss your proposal. This is a mandatory meeting and each group member must be present. The meetings will occur between March 8th-14th, inclusive.

Report

- Guidelines: You are expected to write a complete regression analysis of your dataset, and write up the results in a comprehensive report. The report will contain:
  - A short introduction to the dataset; give a detailed description about the data including the number of variables, variables types, summary statistics, graphs of data, etc..
  - A description of the scientific hypotheses you will investigate.
  - A regression analysis that addresses your scientific hypothesis, using regression model building techniques you have learned in this course. Model diagnostics and details on data appropriateness are expected. Plots and tables are highly encouraged, where you need to include the interpretation for each plot/table.
  - Conclusions and recommendations: give your conclusion based on your regression analysis such as important variables identified, the most proper regression model you have discovered, how statistical assumptions may be violated and how they affect your results, difficulties faced when modelling and shortcomings of the current model, interesting findings, etc..
- Graphs, figures, and tables: All graphs must be labelled correctly and readable. Figures and tables must have descriptive titles and captions, and must be referenced in the text with explanation. All graph axes must be labelled, along with reasonable units. All figures must be readable without having to zoom past 100% on a pdf document.
- Length: Maximum 16 pages including figures and tables, with font size 10. A concise and comprehensive analysis is ideal; a 10 page report would be a fine submission. A 16-page report that meanders through too many hypotheses would lose marks due to a lack of focus.

- Submission format: The submission will consist of the following files:
  (1) a RMarkdown (.Rmd) or KnitR (.Rnw) file,
  (2) a compiled pdf,
  (3) the dataset being analyzed,
  (4) a bibliography file (if needed), and
  (5) any other files I need to compile the project.
- Compiling code: I should be able to compile your RMarkdown or KnitR document from my own computer using only the files provided. If I am not able to compile the project with a reasonable amount of effort ($\sim$ 5 minutes of menial debugging), you will be deducted 20% off the report's final mark. If I have to install an `R`-package to run the code or load a large dataset or calculate an analysis using your code, this is no problem. If I have to debug errors in your code, that is a big problem.
- Grading rubric: See the file `data583_projectRubric.pdf` for full details.

If you need guidance or assistance with any parts of the project, see me early and often. I am here to help you learn to write a **good** analysis of a dataset to be read by a working statistician. This is a skill employers (in areas such as data science) are looking for! **Do not wait until the last minute to ask for help!**