Problem Set 3 - 'Taking our first steps' (15%): due 16 October 2020

Extended Due Date: Monday October 19, 2020 at 11:59pm ET.

- Working in teams of one to four people, please use the Canadian General Social Survey (GSS) and a regression model to analyse some aspect of interest.
- Depending on your focus and background, you may like to use a Bayesian hierarchical model, but regardless of the particular model that you use it must be well explained, thoroughly justified, appropriate to the task at hand, and the results must be beautifully described.
- You may focus on any year, aspect, or geography that is reasonable given the focus and constraints of the GSS. As a reminder, the GSS 'program was designed as a series of independent, annual, cross-sectional surveys, each covering one topic in-depth.' So please consider the topic and the year.
- The GSS is available to University of Toronto students via the library. In order to use it you need to clean and prepare it. Code to do this for one year is being distributed alongside this problem set and was discussed in lectures.
- You are welcome to simply use this code and this year, but the topic of that year will constrain your focus. Naturally, you are welcome to adapt the code to other years. If you use the code exactly as is then you must cite it. If you adapt the code then you don't have to cite it, as it has a MIT license, but it would be appropriate to at least mention and acknowledge it, depending on how close your adaption is.
- Using R Markdown, please write a paper about your analysis and compile it into a PDF.
- Your paper must be well-written, draw on relevant literature, and show your statistical skills by explaining all statistical concepts that you draw on.
- Your paper must have the following sections:
    - title, name/s, and date,
    - abstract,
    - introduction,
    - data,
    - model,
    - results,
    - discussion, and
    - references.
- You are welcome to use appendices for supporting, but not critical, material. Your discussion must include sub-sections on weaknesses and next steps.

- In your report you must provide a link to a GitHub repo that fully contains your analysis. Your code must be entirely reproducible, documented, and readable. Your repo must be well-organised and appropriately use folders.
- Your graphs and tables must be of an incredibly high standard. Graphs and tables should be well formatted and report-ready. They should be clean and digestible. Furthermore, you should label and describe each table/figure.
- When you discuss the dataset (in the data section) you should make sure to discuss (at least):
    - Its key features, strengths, and weaknesses generally.
    - A discussion of the questionnaire - what is good and bad about it?
    - A discussion of the methodology including how they find people to take the survey; what their population, frame, and sample were; what sampling approach they took and what some of the trade-offs may be; what they do about non-response; the cost.
    - This is just some of the issues strong submissions will consider. Show off your knowledge. If this becomes too detailed then you should push some of this to footnotes or an appendix.
- When you discuss your model (in the model section), you must be extremely careful to spell out the statistical model that you are using, defining and explaining each aspect and why it is important. (For a Bayesian model, a discussion of priors and regularization is almost always important.) You should mention the software that you used to run the model. You should be clear about model convergence, model checks, and diagnostic issues. How do the sampling and survey aspects that you discussed assert themselves in the modelling decisions that you make? Again, if it becomes too detailed then push some of the details to footnotes or an appendix.
- You should present model results, graphs, figures, etc, in the results section. This section should strictly relay results. Interpretation of these results and conclusions drawn from the results should be left for the discussion section.
- Your discussion should focus on your model results. Interpret them and explain what they mean. Put them in context. What do we learn about the world having understood your model and its results? What caveats could apply? To what extent does your model represent the small world and the large world (to use the language of McElreath, Ch 2)? What are some weaknesses and opportunities for future work?
- Check that you have referenced everything. Strong submissions will draw on related literature in the discussion (and other sections) and would be sure to also reference those. The style of references does not matter, provided it is consistent.

- As a team, via Quercus, submit a PDF of your paper. Again, in your paper you must have a link to the associated GitHub repo in an appendix. And you must include the R Markdown file that produced the PDF in that repo.
- A good way to work as a team would be to split up the work, so that one person is doing each section. The people doing the sections that rely on data (such as the analysis and the graphs) could just simulate it while they are waiting for the person putting together the data to finish.
- It is expected that your submission be well written and able to be understood by the average reader of say 538. This means that you are allowed to use mathematical notation, but you must be able to explain it all in plain English. Similarly, you can (and hint: you should) use survey, sampling, observational, and statistical terminology, but again you need to explain it. Your work should have flow and should be easy to follow and understand. To communicate well, anyone at the university level should be able to read your report once and relay back the methodology, overall results, findings, weaknesses and next steps without confusion.
- It is recommended that you (informally) proofread one another's sections - why not exchange papers with another group?
- Everyone in the team receives the same mark.
- There should be no evidence that this is a class assignment.