Problem Set 4 - 'Diving into the deep end' (20%): due end of 2 November 2020

- The primary goal of this paper is to predict the overall popular vote of the 2020 American presidential election using multilevel regression with post-stratification.
- As the point is to forecast the election before it happens, no late submissions can be accepted.
- We expect you to work as part of a group of 4 people, but groups of size 1-4 are fine. We have suggested a split of the work based on a 4-person group, but these are just suggestions.
- (Person 1) Individual-level survey data (please see video: https://web.microsoftstream.com/video/80e25a4e-f33e-428e-94c9-b599ec374f43):
    - Request access to the Democracy Fund + UCLA Nationscape 'Full Data Set': https://www.voterstudygroup.org/publication/nationscape-data-set. This could take a day or two. Please start early.
    - Given the expense of collecting this data, and the privilege of having access to it, if you don't properly cite this dataset then you will get zero for this problem set.
    - Once you have access then pick a survey of interest. We will use "ns20200102.dta" in the example (your number may be different).
    - This will be a large file and is not yours to share. Do not push it to GitHub (use the .gitignore file - see here: https://carpentries-incubator.github.io/git-Rstudio-course/02-ignore/index.html).
    - Use the example R code to get started preparing this dataset, and then go on cleaning and preparing it based on what you need.
    - Make graphs and tables about the survey data and write beautiful sentences and paragraphs explaining everything.
- (Person 2) Post-stratification data (please see video: https://web.microsoftstream.com/video/4e0770a4-89ef-403b-8480-cad62eaecd0a):
    - We will use the American Community Surveys (ACS).
    - Please create an account with IPUMS: https://usa.ipums.org/usa/index.shtml
    - You want the 2018 1-year ACS. Then you need to select some variables. This will depend on what you want to model and the survey data, but some options include: REGION, STATEICP, AGE, SEX, MARST, RACE, HISPAN, BPL, CITIZEN, EDUC, LABFORCE,

INCTOT. Have a look around and see what you are interested in, remembering that you will need to establish a correspondence to the survey.
    - Download the relevant post-stratification data (it's probably easiest to change the data format to .dta). Again, this can take some time. Please start this early.
    - This will be a large file and is not yours to share. Do not push it to GitHub (use the .gitignore file - see here: https://carpentries-incubator.github.io/git-Rstudio-course/02-ignore/index.html).
    - Given the expense of collecting this data, and the privilege of having access to it, if you don't properly cite this dataset then you will get zero for this problem set.
    - Clean and prepare the post-stratification dataset.
    - Remember that you need cell counts for the sub-populations in your model. See examples in the readings.
- (Person 3 - start with simulated data while waiting for the real data) Modelling.
    - You will want to explain vote intention based on a variety of explanatory variables. Construct the vote intention variable so that it is binary (either 'supports Trump' or 'supports Biden').
    - You are welcome to use `lm()` but you would need to explain the nuances of this decision in the model section (Hint: start here: https://statmodeling.stat.columbia.edu/2020/01/10/linear-or-logistic-regression-with-binary-outcomes/).
    - That said, you should probably use logistic regression if it is at all possible for you. If you don't know where to start then look at (in increasing levels of complexity) `glm()`, `lme4::glmer()`, or `brms::brm()`. There are examples of each in the readings.
    - Think very deeply about model fit, diagnostics, and other similar things that you need in order to convince someone that your model is appropriate.
    - You have flexibility of the model that you use, (and hence the cells that you'll need to create next). In general, the more cells the better, but you may want fewer cells for simplicity in the writing process and to ensure a decent sample in each cell.
    - Apply your trained model to the post-stratification dataset to make the best estimate of the election result that you can. The specifics will depend on your modelling approach but will likely involve `predict()`, `add_predicted_draws()`, or similar. See the examples in the readings. We are primarily interested in the distribution of your forecast of the overall Presidential popular vote,

and how the explanatory variables affect this. But great submissions would go beyond that. Also, you're taking a statistics course, so if you just gave a central estimate and nothing else, then that would not be great.
    - Create beautiful graphs and tables of your model and results.
    - Create wonderful paragraphs talking about and explaining everything.
- (Person 4 - start with simulated data/results while waiting) Write up.
    - Using R Markdown, please write a very thorough paper about your analysis and compile it into a PDF.
    - The paper must be well-written, draw on relevant literature, and show your statistical skills by explaining all statistical concepts that you draw on.
    - The paper must have the following sections:
        - title, name/s, and date,
        - abstract and keywords,
        - introduction,
        - data,
        - model,
        - results,
        - discussion, and
        - references.
    - The paper may use appendices for supporting, but not critical, material.
    - The discussion needs to be substantial. For instance, if the paper were 10 pages long then a discussion should be at least 2.5 pages. In the discussion, the paper must include subsections on weaknesses and next steps - but these must be in proportion.
- The report must provide a link to a GitHub repo that contains everything (apart from the raw data that you git ignored because it is not yours to share). The code must be entirely reproducible, documented, and readable. The repo must be well-organised and appropriately use folders and README files.
- The graphs and tables must be of an incredibly high standard, well formatted, and report-ready. They should be clean and digestible. Furthermore, you should label and describe each table/figure.
- When you discuss the datasets (in the data section) (remember there will be at least two datasets to discuss) you should make sure to discuss (at least):
    - Their key features, strengths, and weaknesses generally.
    - The survey 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.
- The dataset section is probably an appropriate place to include an explanation of what post-stratification is (in non-statistical language) and the strengths and weaknesses of it, although this discussion may fit more naturally in another section. Regardless, be sure to justify the inclusion of each explanatory variable.
- 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, although you may push the details of this to an appendix depending on how detailed you get. How do the sampling and survey aspects that you discussed assert themselves in the modelling decisions that you make? How can you convince a reader that you've neither overfit nor underfit the data? 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. It must include text explaining all of these and summary statistics and similar. However, 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, but this time interpreting them, and explaining 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? Who is going to win the election? How confident are you in that forecast? Do you have a small or large distribution? What could that mean? Are you more confident in certain states? Do certain explanatory variables carry more weight than others? Etc.
- 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, but it must be consistent.

- If you don't cite R then you will get zero for this problem set.
- 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. And you must include the R Markdown file that produced the PDF in that repo.
- The R Markdown file must exactly produce the PDF. Don't edit it manually *ex post* - that isn't reproducible.
- 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. We have recommended a split above, but you do what works for you.
- 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. The average person doesn't know what a p-value is nor what a confidence interval is. You need to explain all of this in plain language the first time you use 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 work - 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.
- **Again, no extensions are possible, for obvious reasons. The submission portal will close soon after midnight.**