Data Project Part III: Demonstrating your data skills

Student name (ID) for each member of this group

Due dates:

- **Part I is due on October 7th at 5pm PST**
- **Part II is due on November 4th at 5pm PST**
- **Part III is due on December 2nd at 5pm PST**

Make sure to provide enough time for Gradescope to process your submission if you are including large visualizations.

- Late penalty: 50% late penalty if submitted within 24 hours of due date, no marks for assignments submitted thereafter.

Deliverables:

- Submit a PDF including Parts I-III to Gradescope, following the instructions below (one PDF per team).

Submission Process (READ CAREFULLY):

- Download your PDF from Datahub using the File Viewer on the bottom right panel of RStudio. (More -> Export)
- Please submit a PDF of your group project to Gradescope. When turning in each part, please submit all questions through the current part. For example, when turning in Part II, include all questions from Part I.
- Make sure to add all of your group members to the submission. Only one group member has to submit.
- Please answer each problem on a new page. You can specify a pagebreak in Rmd using`\\newpage`.
- You must indicate on Gradescope which questions are on which pages. If the page thumbnails make it difficult to see on Gradescope, open the PDF in a PDF viewer at the same time so you can make the page selections accurately.
- If the submission guidelines are not followed, we may deduct points, as this creates a logistic burden on our end to have to resolve individual cases.

Part III

In Part III of the data project, you will demonstrate a statistical concept from Part III of the course (chapters 13-24 and non-parametrics).

You should be using the same dataset for Part III that you used in part II.

16. [1 mark] Include parts I and II of your project.

17. [2 marks] Identify a statistical test to apply to your data. This must be a statistical test that we cover in part III of the course. Name the statistical test you have chosen and explain why this is the appropriate test for these data. For example, if I have pre- and post-intervention measurements of morning sleepiness recorded as a quantitative variable, I might choose a paired t test, because the paired t-test is appropriate for continuous outcome data in 2 groups that are inherently related.

18. [2 marks] What assumptions are required by the testing method you chose? Are these assumptions met by your data? How did you assess this? For example, one of the assumptions of the t-test is that the data are normally distributed, so you might choose to assess this with a histogram, or a q-q plot.

19. [2 marks] Clearly state the null and alternative hypotheses for your test.

20. [2 marks] Conduct the statistical test. Include the R code you used to generate your results. Annotate your code to help us follow your reasoning.

21. [4 marks] Present your results in a clear summary. This should include both a text summary and a table or figure with appropriate labeling. For example, if your outcome and predictor/exposure variables are both binary, this might be a 2x2 table. If your method was regression, you might present your regression line graphically. Include your code and annotations.

22. [4 marks] Interpret your findings. Include a statement about the evidence, your conclusions, and the generalizability of your findings. Our analyses and conclusions depend on the quality of our study design and the methods of data collection. Any missteps or oversights during the data collection process could potentially change the outcome of what we are trying to find. Consider the methods used to collect the data you analyzed. Was there any potential issue in how the participants were selected/recruited, retained, or assessed that may have impacted the outcome of your analysis/visualization? Were there any potential biases that you might be concerned about? Were there factors that were not measured or considered that you think could be important to the interpretation of these data?

23. [1 mark] Create a statement of contribution. This is now common in journal articles. For example, the American Journal of Epidemiology provides the following instructions to authors: "Authorship credit should be based on criteria developed by the International Committee for Medical Journal Editors (ICMJE): 1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the article or reviewing it and, if appropriate, revising it critically for important intellectual content; 3) final approval of the version to be published. Authors should meet all conditions. In addition, each author must certify that he or she has participated sufficiently in the work to believe in its overall validity and to take public responsibility for appropriate portions of its content. Author names should be listed in ScholarOne and author contributions should be detailed in the cover letter (e.g., "Author A designed the study and directed its implementation, including quality assurance and control. Author B helped supervise the field activities and designed the study's analytic strategy. Author C helped conduct the literature review and prepare the Methods and the Discussion sections of the text.").

An example from a recent issue of the BMJ (Woolf, Masters, and Aron BMJ 2021;373:n1343):

"Contributors: SHW led the production of this manuscript and had primary responsibility for the composition. He is guarantor. RKM contributed revisions and had primary responsibility for data acquisition and analysis, the modeling results that form the basis for this study, and production of the supplementary

material. LYA contributed revisions and had primary responsibility for dealing with the study's policy implications in the discussion section. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted."

For your project, please craft a statement indicating the contributions of each group member. If your group divided the assignment responsibilities by question you may use question numbers to indicate which member had primary responsibility for each question (for example: Member XX had primary responsibility for questions x,x,x...).