Data Project Part II+III: Demonstrating Your Data Skills

Group number

Due dates:

- **Part I:** February 28th, 10:00pm
- ~~**Part II:** April 3rd, 10:00pm~~
- **Part II+III:** April 29th, 10:00pm

Make sure to provide enough time for Gradescope submission to be uploaded if you include large visualizations.

Deliverables:

- Submit a PDF including Part I-III to Gradescope here (https://www.gradescope.com/courses/79434/assignments/456151) following the instructions below. (one PDF per team)
- Complete the peer evaluation form here (https://docs.google.com/forms/d/e/1FAIpQLSeok6y7Htj8KR086pqOtq2MGmuI (one per person) **At the end of the Google Form, make sure to follow the link to a second form to submit your name and SID**. If you do not complete the second form, we have no way to give you credit for your evaluation.

Submission Process (READ CAREFULLY):

- **DO NOT INCLUDE YOUR NAMES ON YOUR SUBMISSION. PUT YOUR GROUP NUMBER INSTEAD.** We will be facilitating a blinded peer review of the report after submission, so we need each report to be non-identifiable.
- Download your PDF from Datahub using the File Viewer on the bottom right panel of RStudio.
- Please submit a PDF of your group project to Gradescope **here**. When turning in each part, please submit all questions through the current part. *For example when turning in Part II please 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 you can't see it properly on Gradescope, open the PDF in a PDF viewer at the same time so you can make the 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 II:

In Part II of the data project we are asking that you demonstrate a statistical concept from Part II of the course (material on Midterm II).

Problem 1 [2 points]: If you are using the same dataset as Part I, just write "See Part I" for this question (your final submission should include Parts I-III). If you have changed your dataset, for this question, you will need to create a new summary of the new dataset you are working with, but you do not need to fully re-do Part I. This summary should include:

- The problem you are addressing with these data expressed in terms of the PPDAC framework

- Target population for your project? Why was this target chosen?
- Sampling frame used to collect the data you are using
- The source and description of the contents of your dataset
- URL to the original data source if applicable. If not (e.g., the data came from your research), provide 1-2 sentences saying where the data came from. If you completed a web form to access the data and selected a subset, describe these steps.

Problem 2 [2 points]: Describe a quantity you will estimate in your problem using probability notation. Are you planning to calculate marginal probabilities? Conditional probabilities?

Problem 3 [2 points]: Describe the type of theoretical distribution that is relevant for your data. - What type of variable is your outcome? - What theoretical distribution that we have talked about would potentially be appropriate to use with these data (Normal, Binomial, Poisson. . .)

Problem 4 [4 points]: Use the data you have to demonstrate a statistical concept from Part II of the course. Describe the concept that you are demonstrating and interpret the findings.

Part III

In Part III of the data project we are asking that you demonstrate a statistical concept from Part III of the course (statistical testing).

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

Problem 5a [2 points]: Identify the statistical test that you applied to your data (must be a concept we covered in Part III of the course).

Problem 5b [2 points]: What assumptions are required by the method you chose in **5a**? Describe how you assessed whether these assumptions are met by your dataset.

Problem 5c [2 points]: Explain why this test is appropriate for the data you have and the question you are trying to answer. Use at least one visualization technique and include both the output and the R code that generated it.

Problem 5d [2 points]: Clearly state the null and alternative hypotheses for this test.

Problem 6 [2 points]: Include the R code you used to generate your results - annotate your code to help us follow your reasoning.

Problem 7 [4 points]: Present your results in a clear summary. This should include both a text summary and a table or figure with appropriate labelling.

Problem 8 [4 points]: Interpret your findings. Include a statement about the strength of this testing, your conclusions and the generalizability of your findings.

Problem 9 [1 point]: Complete the peer evaluation form for group work here. This will be counted as a completion point. **Each individual in the group should fill out their own peer evaluation form.** This is also due by April 29th at 10:00pm. Make sure to follow the link at the end of the form to the second form where you can fill out your name and SID to get credit for this problem.