Data Project Part I: Demonstrating Your Data Skills

Due dates:

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

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

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

Submission Process (READ CAREFULLY):

- **DO NOT INCLUDE YOUR NAMES ON YOUR SUBMISSION.** We will be facilitating a blind 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.

Instructions:

Groups of 4-5 students have been assigned randomly to facilitate teamwork across group members coming from varying backgrounds and skill levels and develop a professional work ethic.

Peer reviews will be distributed shortly after each due date. Each group will provide a peer review for two other groups. Peer reviews will be due one week from the due date for each portion of the project. The peer review process will be blinded.

Your task for this project is to find data that is loosely related to health, public health, biology, sociology, demography, justice, or another topic affiliated with public health or biology. These data could be a pre-existing data set from the Internet, data you have access to (and permission to use) from your lab or internship, or, less frequently, something you create from a hard copy.

You will then import your data into R and use it to demonstrate three statistical concepts covered in class, one from each section of the class:

- Part I: Collecting, Exploring, and Visualizing Data (Edition 4 Chapters 1-8 and early lectures on dplyr and ggplot2)
- Part II: From Chance to Inference (Edition 4 Chapters 9-16)

- Part III: Statistical Inference (Edition 4 Chapters 17-25 and lectures on bootstrapping and permutation tests)

For example, for Part I you could create a data visualization using `ggplot2`. For Part II, you could demonstrate how the data could be used to calculate a conditional probability of interest.

The objectives of this assignment are to:

- Gain competence finding public health data and reading it into R to perform your own analyses.
- Apply the PPDAC framework to a question of your choosing.
- Demonstrate your newly-acquired statistical skills.
- Create a report on your dataset that summarizes your findings in a clear way.

Part I:

Setup:

1. Create a new folder in your `ph142-sp20/` directory called `project/`.
2. In this `project/` folder, create an .Rmd for your project.
3. Find a dataset you're interested in a upload it into this `project/` folder. *You can click "Upload" in the File Viewer to upload your data onto Datahub. Make sure to use a data format you know how to read into R, such as csv, xlsx, etc. You may need to copy and paste your file into an Excel sheet first to get it into an appropriate format.
4. Complete the following questions in your .Rmd,
5. Make sure to follow the submission guidelines outlined above when you submit.

Answer the following questions:

Problem 1 [2 points]: What is the problem your are addressing with these data? Express this in terms of the PPDAC framework.

Problem 2 [2 points]: What is the target population for your project? Why was this target chosen?

Problem 3 [2 points]: What is the sampling frame used to collect the data you are using. Describe why you think this sampling strategy is appropriate for your question. To what group(s) would you feel comfortable generalizing the findings of your study and why?

Problem 4 [2 points]: Write a brief description (1-4 sentences) of the source and contents of your dataset. Provide a URL to the original data source if applicable. If not (e.g., the data came from your internship), 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 5 [1 point]: Import your data into R. Assign your dataset to an object.

Problem 6 [3 points]: Use code to answer the following questions:

i) What are the dimensions of the dataset?
ii) Provide a list of variable names.
iii) Print the first six rows of the dataset.

Problem 7 [4 points]: Use the data to demonstrate a statistical concept from Part I of the course. Describe the concept that you are demonstrating and interpret the findings. You should use a combination of code and written explanation.

Tips

- We anticipate that importing the data into R may be a challenging task for many datasets. We want you to experience this because it is important to know how difficult it can be (in some cases) to read data into a statistical software. To make this easier on yourself, choose data that has a "rectangle" format with no merged headings. For example, it should contain variable headings where each variable

has its own row of data. There should be no summary information at the end of the data, or any information outside the "rectangle" that makes up your dataset.

- The data will be easiest to use in R if the variable names do not contain spaces or unusual characters. If you need to, you can rename variables in Excel to be of the format: "my_variable_name" rather than "my variable" or "my variable * 100", as examples.
- If you are having trouble importing the data, try making a much smaller data set and get that to import first. This can help you isolate the problem. Some datasets you find will be thousands or even millions of rows. Given that this may be your first time importing data, we recommend you choose something smaller!
- To make your report look presentable, check out this **cheat sheet style guide on .Rmd.**