Group Project Part I: Demonstrating your data skills

Student name (ID) for each member of this group

Due dates:

- **Part I is due on March 1st at 10pm PST**
- **Part II is due on April 5th at 10pm PST**
- **Part III is due on April 26th at 10pm PST**

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 marks for assignments submitted thereafter.

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 here. 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 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:

You have 3 choices for this project:

1) If you have a group of students you would like to work with, you may self assign to a group of up to 5 students.

2) If you would like to work in a group but do not know other students in the class, you can fill out the google form (LINK TO BE ADDED) and we will randomly assign students who would like to work in a group into groups and send out the contact information for each group to its members.

3) You can also work alone!

Your task for this project is to find data that is related to health, public health, biology, sociology, demography, justice, or another topic affiliated with public health or biology. These data could be a preexisting 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 (Based on material in the textbook 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.

Because we are asking you to provide some visualizations and use the same dataset for parts II and III of the project, make sure that you choose a dataset with enough observations (row) to have something that you can interpret. You will also need something large enough so that you can run a statistical test in part III. A good general rule here is to choose a dataset with at least 100 observations, and at least 30 in each group if you are comparing across groups. For example, if you are answering a question about maze times in groups of mice exposed to some training program vs. not, you would want to have data on 100 mice, at least 30 of which had undergone the training program and at least 30 of which had not.

Part I

Setup:

You can have one student in your group following these instructions, or have many group members do this and send files back and forth to one another to work on the project together.

1. Create a new folder in your ph142-sp21/ 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 can copy and paste your file into an Excel sheet first to get it into an appropriate format.
4. Copy and paste the questions below into your Rmd file and complete them.
5. Make sure to follow the submission guidelines outlined above when you submit.

Questions:

1. [2 marks] What is the problem your are addressing with these data? State the question you are trying to answer and let us know what type of question this is in terms of the PPDAC framework.

2. [2 marks] What is the target population for your project? Why was this target chosen i.e., what was your rationale for wanting to answer this question in this specific population?

3. [2 marks] 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.

4. [2 marks] 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 (including any options you selected) and the date you accessed the data.

5. [1 mark] Write code below to import your data into R. Assign your dataset to an object.

6. [3 marks] Use code in R 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.

7. [4 marks] 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. This should be a combination of code and written explanation.

Tips

- We anticipate that importing the data into R may be a challenging task for many datasets. The frustration is part of the challenge and a common occurrence if you work with data from the real world. 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 import it 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.