Group Project Part I: Demonstrating your data skills

Student name (ID) for each member of this group

Due dates:

- Part I is due on February 24th at 10pm PST
- Part II is due on March 31st at 10pm PST
- Part III is due on May 12th at 12pm noon 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.

Submission Process (READ CAREFULLY):

1. Download your PDF from Datahub using the File Viewer on the bottom right panel of RStudio. (More -> Export)
2. 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.
3. Make sure to add all of your group members to the submission. Only one group member has to submit. Non-submitting group members should confirm that the project submission appears in their Gradescope account.
4. Please answer each problem on a new page. You can specify a pagebreak in Rmd using `\newpage`.
5. 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.
6. 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:

Think carefully about how you will approach working in a group together. You will be asked to include a statement of contribution in part III - more detailed instructions are included in the part III assignment.

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. The data should have humans as the unit of analysis, or be based on aggregates of human data (for example, rates of mortality) for multiple units of analysis (multiple hospitals/clinics, etc.).

These data could be a preexisting data set that has been made publicly available on the internet, data you have access to (and permission to use) from your lab or internship, or, less frequently, something you create (with appropriate permission) from a hard copy. You will then import your data into R and use this dataset to demonstrate concepts covered in class in three submissions, each focused on one of the three sections of the class:

- Part I: Collecting, Exploring, and Visualizing Data (based on material in edition 4 of the textbook, chapters 1-8, and early lectures on `dplyr` and `ggplot2`)
- Part II: Demonstrating your data skills (edition 4, chapters 9-16)
- Part III: Statistical Inference (edition 4, chapters 17-25 and lectures on bootstrapping and permutation tests)

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.
- Develop your ability to choose statistical methods appropriately.
- Demonstrate your newly-acquired statistical skills.
- Demonstrate your ability to communicate statistical analyses and findings.

Because we are asking you to provide some visualizations and use the **same dataset** for parts I, II, and III of the project, make sure that you choose a dataset with enough observations (rows) to have something that you can interpret. You will also need a dataset large enough so that you can run a statistical test in part III. A good general guide here is to choose a dataset with around 100 observations, with approximately 30 observations in each group if you decide to compare across groups in the third part. For example, if you are answering a question about mean days in the ICU between groups of patients exposed to some intervention procedure vs. not, you would want to have data on about 100 patients, about 30 of whom had undergone the intervention and about 30 of whom had not. Your outcome of interest should be something that can be defined as a continuous, discrete or binary variable.

GSI Check-in

To double check that your dataset is valid, you are required to meet with your assigned GSI at least once before the Part I due date (Feb. 24th). Your assigned GSI should be reaching out to you shortly after groups are assigned and will provide instructions on how to set up a check-in meeting. If you have any questions and would like to reach out to your assigned GSI, please follow the guidelines on the “Data Project” page on the course website: https://ph142-ucb.github.io/sp25/data-proj/

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. In the past, students have shared their code with group members using a shared Google document.

1. Create a new folder in your ph142-sp25/ directory called project/.
2. In this project/ folder, create an .Rmd for your project (see lab 1 for how to create a new .Rmd!).
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] The first part of our PPDAC framework is to identify the problem you 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. A question statement should be as specific as possible. For example: Do students who regularly get 8 hours of sleep have fewer visits to the health center? This question is an example of an etiologic or causal question.

2. [2 marks] Why is this question interesting or important? You could talk here about how existing data/studies suggest this might be important, how the findings might make an impact, how the findings might be used, or why you are personally interested in this question.

3. [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?)

4. [2 marks] What is the sampling frame used to collect the data you are using? It may be helpful here to read any protocol papers, trial registration records, '.Readme' files or documentation that are associated with your dataset. If you have trouble identifying how the records/individuals were sampled, confirm with your supporting GSI that your dataset will be usable for the purposes of the class. 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?

5. [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.

6. [1 mark] Write code below to import your data into R. Assign your dataset to an object. Make sure to include and annotate this code in your submission (you can use a # to comment out regular text within code chunks to annotate).

7. [3 marks] Write code in R (included in your submission with annotation) to answer the following questions:

i) What are the dimensions of the dataset?

ii) What are the variable names of the variables in your dataset?

iii) Print the first six rows of the dataset.

8. [2 marks] Use the data to demonstrate a data visualization skill we have covered during Part I of the course. Choose a visualization relevant to your stated problem. Include your code in your submission. For example, you could visualize the distribution of our outcome with a histogram, or use a bar graph to represent the distribution of your exposure variable.

9. [2 marks] Describe the skill that you are demonstrating and interpret your findings. For example, if you have created a histogram, describe the central tendency, shape of the distribution, etc.

Tips

- We anticipate that importing the data into R may be a challenging task for many datasets. This 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" in 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". This is important because R will not recognize variable names with spaces between the words, for example.
- 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.