Group Project Part I: Demonstrating your data skills

Student name (ID) for each member of this group

Due Dates:

- Part I: Sep 26th 10pm
- Part II: Oct 24th 10pm
- Part III: Dec 5th **5pm**

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)
- Make sure your code does not run off the page!
- 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 on Gradescope.** Only one group member has to submit.
- Please answer each problem on a new page. You can specify a page break in your 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 having to resolve individual cases creates a logistic burden on our end.

Instructions:

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 statistical concepts covered in class, the project will be divided into three parts, representing skills and concepts 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 (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 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. Do not choose a dataset that is a time series - we do not cover methods for those types of data in PH142 and you will not be able to use the data to complete the assignment. Also be careful with datasets that present aggregate data. Data that have been aggregated are often not suitable for the types of distributions and tests we want you to demonstrate for this project. Meeting with your supporting GSI will be the best way to determine if the dataset you have found is appropriate for completing the project.

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-fa25/ directory called project/.
2. In this project/ folder, create an .Rmd for your project. Edit the header of this .rmd to include the names of each of your group members.
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 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. A question statement should be as specific as possible, for example rather than;"how is sleep related to health" a specific question would be "Do students who regularly get 8 hours of sleep have fewer visits to the health center?"
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 this question fills a gap in the scientific literature, how the findings might make an impact, and how the findings might be used.
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. If you downloaded publicly available data from the internet, provide information on how and when the data were accessed, including 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. Include information any ethical approval you obtained to work with these data and/or who granted you permission to use the dataset. 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.
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 your 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, why you chose this visualization 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. 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.