Data Project Part II: Demonstrating your data skills

Student name (ID) for each member of this group

Due Dates

- Part II: Monday, July 30th, 10pm PT
- Part III: Monday, August 14th, 12 noon

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.

Deliverables:

- Submit a PDF including Parts I **and** II on Gradescope following the instructions below (one PDF per group).

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 should submit but they MUST identify the group members to gradescope or the other group members will not have the assignment counted.
- 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.

Part II

In Part II of the data project, you will demonstrate a statistical concept from Part II of the course (material on midterm II, Chapters 9-12 of Baldi & Moore).

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

Questions:

10. [1 mark] Include your work for Part I.
11. [3 marks] Describe the type of theoretical distribution that is relevant for your data.

- What type of variable(s) are you investigating (continuous, categorical, ordinal, etc)?
- What theoretical distribution that we have talked about would potentially be appropriate to use with these data (Normal, Binomial, Poisson. . .)
- Why is this an appropriate model for the data you are studying? (HINT what are the assumptions of this distribution)

12.[2 marks] - What are the parameters that define this distribution? - Calculate these parameters for your data.

13. [2 marks] Use your outcome data to calculate a probability. Provide an equation (use fpr,a; probability notation) that describes this probability. Note whether this probability is a conditional or a marginal probability. For example if my outcome variable is height in inches, I might calculate the probability that an individual in the dataset has a height of greater than 5 feet 10 inches. P(height>=70inches) = ? This would be a marginal probability.
14. [2 marks] What type of variable is your primary exposure of interest? If this variable is a demographic variable (age, gender identity, race/ethnic identity) explain how the categories of this variable are defined and what the rationale is for this (for example if gender identity is being used, is the idea to capture something about biology ie using gender identity as a marker for genetic or phenotypic sex, or as a marker of social exposures). If your data is not from a randomized trial where your exposure of interest was randomly assigned, are there important factors that may have affected how this exposure was distributed?
15. [4 marks] Use your data to create a visualization of your data that begins to explore your research question. Include code in R, a visual of some kind and text interpretation. For example, if you outcome is height of children at age 10 and your predictor variable is exposure to food insecurity in the first year of life, you could provide a histogram of height among children exposed to food insecurity and a separate histogram of height among children not exposed to food insecurity. Make sure you describe your interpretation of the results.