Data Project Part II: Demonstrating your data skills

Student name (ID) for each member of this group

Due dates:

- **Part I is due on September 30th at 5pm PST**
- **Part II is due on November 4th at 5pm PST**
- **Part III is due on December 2nd at 5pm 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.

Deliverables:

- Submit a PDF including Part I-II on Gradescope following the instructions below (one PDF per group). Make sure that your code is visible in your submission!

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. 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. Non-submitting group members should confirm that the project submission appears in their Gradescope account.
- 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 page 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.

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 the textbook).

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

10. [1 mark] Include your work for Part I.
11. [2 marks] Calculate a marginal probability based on your outcome variable. Provide an equation (using probability notation) that describes this 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 60 inches. $P(height \geq 60) =?$. This would be a marginal probability. You may need to first add a new variable to your dataset to calculate your probability of interest, such as a binary variable indicating whether height is greater than 60 inches. There is a resource video about how to code such variables that could be helpful!
12. [2 marks] Using any two variables in your dataset (or derived variables), calculate a conditional probability. Provide an equation (using probability notation) that describes this probability and then use R to calculate it.
13. [2 marks] Does your dataset contain a continuous variable? If it does, does the distribution of that variable appear to be normal? Justify your answer using a plot. If your data does not contain a continous variable, give an example related to your dataset of a hypothetical variable that is continuous. That is, imagine what a continuous variable could be in relation to your dataset and topic of interest. For this hypothetical variable, describe what you imagine its shape might be, and how you would check whether or not it is normally distributed.
14. [4 marks] Does your dataset contain a binary variable? If so, does this variable meet the criteria to be considered binomially distributed? If so, describe this variable in terms of n and p. Calculate a probability based on this variable, first write the formula for the probability and then using R to calculate the probability (you do not need to calculate the probability by hand). If your data does not contain a binary variable, you can create one based on an underlying continuous variable or a categorical variable with > 2 levels to answer this question.