Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In Project 2, your group did two things: (1) wrote an article, made a poster, or recorded a presentation that was geared towards someone interested in your research topic, and (2) created an RMarkdown report describing how your project meets class learning outcomes to someone like me. The goal of this assessment is to ask you questions from a more technical perspective.

Though you should think of this assessment like homework, **you will only have one attempt.** This means you must make sure to answer each question completely. Per the syllabus, this will be graded as “Pass”/“Not Pass”; you should consider this as “Meets Expectations” vs. “Doesn’t Meet Expectations.” This implies that not all questions must be answered perfectly, but rather that you must demonstrate a sufficient understanding of the technical components of your project. Additionally, your contribution to the group portion of the project will be considered.

This assessment is to be completed by **each student individually, without discussing the question with other classmates or using AI like ChatGPT**. Submit a **hardcopy** of your responses by the due date.

## Basic Information and Data Wrangling

1. What was the main goal or research question of your project?
2. List three variables from your dataset and classify each as categorical/numerical. In the rare case that you did not have three variables, list and describe the ones you did have. Finally, what kinds of R objects are those variables?
3. In your project, you may have performed data wrangling steps to get the data into a format useful for analysis. Suppose you needed to do the following two steps in R: (1) remove some observations, and (2) create a new variable. Fill in the blanks below that would perform these steps based on your group’s code/dataset. Note: if you don’t have appropriate variables (e.g., no categorical variable), then state beforehand a fictional variable that would be relevant to your dataset and then use that variable in your response. Change the number of blanks to fit your response.

library(\_\_\_\_\_\_\_\_\_\_\_\_\_)

\_\_\_\_\_\_\_\_\_\_ |>

\_\_\_\_\_\_\_\_\_\_(\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_) |>

\_\_\_\_\_\_\_\_\_\_(\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_)

## Descriptive Statistics and Graphs

1. Pick a numerical variable from your main dataset and state what that variable is. Based on the data you collected, briefly explain what measures of center and spread would be appropriate for this variable. In the rare event you did not have a numerical variable, think of an example that would be relevant to your study and describe it based on this question. The measures of center and spread should be appropriate to use together as a pair.
2. In your project, you were not required to make a graph of your sample. From the following list, choose a graph type that would be appropriate for your project. Note: this cannot be a graph that you used. State the graph type and how it could be useful for your analysis.

Graph types: bar graph, histogram, boxplot, line graph, scatterplot.

1. For the graph type your chose above (i.e., one you did not create for your project), fill in the blanks to generate a version of the graph based on your main dataset from your R code.

library(\_\_\_\_\_\_\_\_\_\_\_)

ggplot(data = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)

geom\_\_\_\_\_\_\_\_\_(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_) (note: expand the code and number of blanks as needed based on the graph you chose)

## Regression Models

1. Did your group include a regression model in your project? If so, please write out the equation below so someone else could understand what the equation was modeling. If you did not include a regression, write “N/A.”
2. Suppose someone created a regression equation to model the relationship of penguin weight (in lbs, response variable) against the penguin’s height (in cm) and bill length (in mm). The slope with the height variable was 3.5; the slope with the bill length was 6.5; and the intercept was 7.8. Suppose this person said the following: “The weight of a penguin increases by one pound for every 6.5 mm its bill length increases.” There are multiple things wrong with this sentence. State 2 of them and explain why they are wrong.
3. For the regression question in the previous equation, provide a correct interpretation of the slope term for the bill length (6.5).
4. Let’s say you wanted to run a regression with your dataset (if you ran a regression for your project, please consider the situation you used, or a modification to it). However, you have 7 possible explanatory variables to use, and you are unsure which ones to include. Explain in roughly a paragraph how you could use backward selection to choose the variables.

## Confidence Intervals

1. Did your group include a confidence interval in your project? If so, state the interval, what it was measuring, and what kind of interval (e.g., percentile, standard error) it was. If you did not include a confidence interval, write “N/A.”
2. In roughly a paragraph, discuss in your own words the similarities and differences between a bootstrap distribution and a sampling distribution.
3. Suppose someone made a confidence interval for the penguin example mentioned earlier for the difference in penguin heights between male and female penguins. They report a 95% confidence interval of 5.4 to -0.01 inches (taking male – female). They say, “Male penguins are up to 5.4 inches taller than female penguins. Though the interval indicates females may be up to 0.01 inches taller, we can disregard this because it’s such a small amount and conclude males are taller than females.” There are multiple things wrong with this explanation. State 2 of them and explain why they are wrong.
4. Suppose the person above had code that contained the following objects: a dataframe called penguins that contained variables gender and height; a bootstrap distribution called peng\_boot; and a difference in means statistic saved as height\_diff. Fill in the blanks below to create a 99% confidence interval using both the percentile method and the theory-based method. Add blanks/lines as needed. You can assume the infer package is loaded.

Percentile method:

\_\_\_\_\_\_\_\_\_\_\_\_(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_,

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_,

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)

Theory-based method:

\_\_\_\_\_\_\_\_\_\_\_\_(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_,

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_,

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)

## Hypothesis Tests

1. Did your group perform a hypothesis test in your project? If so, what kind of test you ran and the hypotheses. If you did not run a hypothesis test, write “N/A.”
2. In no more than 3 sentences of your own words, explain what a null distribution shows.
3. Suppose someone ran a hypothesis test for the penguin example mentioned earlier to see if the average bill length was less than 5 mm. Write out the hypotheses for this example.
4. For the hypothesis test above, the person ran a simulation-based test (i.e., generated a null distribution) and got a p-value of 0. They say that they must have done something wrong, as p-values cannot be 0. Explain to them in roughly a paragraph why they did not necessarily do something wrong (assume their code is indeed correct).

## Reflection and Peer Evaluation

1. How did this project facilitate your understanding of data analytics and statistics? What did you learn about data analytics and statistics in this project?
2. If you could redo this project, what aspects would you change from the beginning?
3. For each group member, please list their name, what they did well, and what they could improve on. Note: Not all group members will have the same level of knowledge with every aspect of the course content. This is quite typical when working in teams, so please keep this in mind when you are evaluating.