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

In Project 1, your group did two things: (1) wrote an article 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 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

1. In no more than 3 sentences, describe the dataset you used.
2. What was the main goal or research question of your project?
3. 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?
4. List the name of one data frame object from your code.

## Data Importing and Wrangling

1. Suppose I give you an R script that contains the following code to load a dataset called wf\_lfp.csv.

lfp <- read.csv(file = "/Users/nbussberg/STS 2300/wf-lfp.csv")

Explain in 1-2 sentences why this code is problematic despite it working on my laptop. Then, in another 1-3 sentences, explain how I should modify this code so that it will work for you.

1. In your project, you performed relevant data wrangling steps to get the data into a format useful for analysis. Suppose you needed to do the following steps in R: (1) remove some observations, (2) remove some variables, (3) create a new variable, and (4) sort the data based on values in a categorical variable. Fill in the blanks below that would perform these steps based on your group’s code and 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.
2. Pick a categorical variable from your dataset and state what that variable is. Based on the data you collected, briefly explain what descriptive statistic(s) would be appropriate for this variable. In the rare event you did not have a categorical variable, think of an example that would be relevant to your study and describe it based on this question.
3. Consider the second graph your group made that appears in your article or presentation. What variable(s) was (were) being plotted? Briefly explain why this graph type is appropriate for that variable(s). For example, if you made a histogram, why was a histogram used and not other types of graphs?
4. For the second graph that appears in your article/presentation (from the question above), explain how you customized it away from default settings. Do not simply paste your R code; I want you to explain in words what you did.
5. From the following list, choose a graph type that you did not use in your project. State the graph type and why you did not make/use this graph. This could include why the graph is inappropriate or why one of the graphs you did use should be preferred.

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

## R/RMarkdown Code

1. Pick an R chunk in your RMarkdown file that shows results after you run it. When you knit the document, it likely shows both the code and the results. How would you modify the code to hide the results? Be specific and include example code below.
2. List three R packages you used in your RMarkdown document (if you had less than three, then just list the ones you had. Briefly explain what each of those packages were used for (1-2 sentences max for each).
3. Suppose you chose to summarize your numerical variable from Question 9 with summary statistics made by the package dplyr’s summarize() function. Let’s also say that you wanted to split the summary statistics output table based on values in the categorical variable in Question 10. Fill in the blanks that follow on how you would do this. Use the names of datasets and variables from **your code**. In the even that you did not have a numerical or categorical variable, state briefly a name for one that could be relevant and use this fictional variable in your response below. Do not add more lines to the code below than I provide.

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

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

\_\_\_\_\_\_\_\_\_\_\_\_ = \_\_\_\_\_\_\_\_\_\_\_\_,

\_\_\_\_\_\_\_\_\_\_\_\_ = \_\_\_\_\_\_\_\_\_\_\_\_)

1. Suppose we stored the results of the summarize() function above in an object called my\_results. What kind of object is my\_results?
2. For the object my\_results above, write template code below that shows how you would use the kableExtra package to format the table in a nicer format.

library(kableExtra)

\_\_\_\_\_\_\_\_\_(\_\_\_\_\_\_\_\_\_\_\_) (note: expand the code and number of blanks as needed to complete this question)

1. For the graph type your chose in Question 13 (the 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. Note: I understand that this graph may not produce something useful/informative (which is likely why you didn’t use it) – this question is to assess your R knowledge.

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

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

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

1. For the graph you made in the question above, explain how you would change the color of the points, lines, or bars. I strongly recommend including template R code in your explanation.
2. For the graph you made in the question above, suppose you saved this graph as an R object called my\_awesome\_graph. Fill in the blanks below to create a title, update the x-axis label, and update the y-axis label. The titles and labels do not have to be perfect; I am assessing your ability to code in R for this question.

my\_awesome\_graph +

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

\_\_\_\_\_\_\_ = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_,

\_\_\_\_\_\_\_ = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)

## 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.