

Preliminary Project Submission

Due Date: Monday March 23, 11:59 PM

Introduction

The project is designed to give you hands-on experience answering research questions using data. You will engage with all aspects of analysis: from selecting a dataset and variables to conducting statistical inference. Along the way, we will assess you via two submissions: a preliminary submission and a final submission.

This assignment is subject to the honor code. You will submit your own work. Although it is fine to discuss the project with other members of your lab, plagiarism and code sharing are not acceptable. Group work is not mandatory. If you choose to work in a group, you may have a common response variable, but all (or most, if not possible) explanatory variables must be different for each group member.

Preliminary Submission: Organization

You will need to submit two file on Canvas: a Microsoft Word (.docx) or PDF, and your R script. The document should contain all of the materials described below, and an Appendix with the properly commented R Code you used to complete the assignment. You should **also submit** your full R script file (everything you used to complete the assignment) on Canvas. This code should run without errors in RStudio.

You should select one dataset from the several project datasets available on your Canvas page. Note that different lab sections will have different project datasets. When choosing your data, look at both the data file and the accompanying codebook, which will define all of the variables and introduce you to the dataset. Think about questions you would enjoy exploring. What are some research questions you could imagine? Hypotheses?

Steps to Complete the Assignment

1. After selecting your dataset, choose **two response variables**: one numerical and one dichotomous. Remember that a “response” variable is the variable we seek to *explain*. For example, we may want to explain variation in income, weight, or whether someone died.
 - Your two variables may be completely different OR you can choose to dichotomize your numerical variable. If you choose to dichotomize, you **MUST** justify this choice using external research and strong logic. Recall that a dichotomous variable sorts all of your observations into two mutually exclusive groups.
2. Next, you should select **three explanatory variables**. One should be numerical, one should be categorical (3 or more levels) and one should be dichotomous. Explanatory variables should help explain the variation in your response variables. For example, genetic makeup (explanatory variable) can explain likelihood of an Alzheimer’s diagnosis (response variable). Think carefully about these variables! You will use them to frame research questions. We highly advise that you consider your research questions (described in Q5) **BEFORE** committing to these explanatory variables. After selecting all variables for this project, you should have:
 - one numerical **response** variable
 - one dichotomous **response** variable
 - one numerical **explanatory** variable,
 - one dichotomous **explanatory** variable,

- one categorical (between 3 and ~8 levels) **explanatory** variable.
3. Report the variables you have chosen. Provide their exact R name, an intuitive meaning, whether they are response or explanatory, and their type (categorical, numerical, etc.).
 4. In order to receive full credit on this assignment, **you must recode at least one variable**. Recoding is NOT just assigning NAs to unreasonable values or naming variable levels: you must create a new categorical variable, using either an existing categorical variable or a numerical variable. After you have labeled, cleaned, and recoded your variables, write a brief paragraph explaining your procedures. Common procedures include:
 - Constructing a categorical variable with the appropriate number of categories (REQUIRED)
 - Assigning NAs to any unreasonable values
 - Cleaning any typos/incorrect categories
 5. Formulate four research questions. It can be as simple as “What is the association between X and Y?”. The first three research questions will be about your numerical response variable and each of the explanatory variables. The fourth question will be about your dichotomous response variable and one of the categorical explanatory variables. Give compelling rationales to why the explanatory variables might be associated with the response variable.
 6. Create six plots to visualize the relationships between:
 - your numerical response variable and each of the explanatory variables (three plots);
 - your dichotomous response variable and each of the explanatory variables (three plots).

Each plot must have the appropriate title and labels. Be sure to include the plots at the end of your docx or pdf document. Use your judgment on how many plots per page to include such that they can be easily understood (six is too many).

Tips

- It is ideal to recode a variable based on existing research and standards. Using outside sources to justify your dichotomization or recoding decisions is the best way to earn full credit. Google Scholar or the library’s DiscoverE tool are both great places to start.
- Consider your results and whether they make sense. If your results, especially the plots, seem strange, check your code to make sure you have done the appropriate cleaning or recoding procedures.
- Before submitting your R script, clean your RStudio environment, and make sure that the entire script runs without errors.
- In the case of bar plots, it is often better to use **marginal proportions** with appropriate margins, especially if there is a significant difference between counts across groups. Remember that, in order to do this, we use the following code logic:

```
#Create two-way contingency table
table <- (responsevar,explanatoryvar)

#Create column proportions table
proptable <- prop.table(table, margin=2)

#Create bar plot using proportions with side-by-side bars
barplot(proptable, beside=TRUE)
```

- Your final project will be based on this assignment! Therefore, it is in your best interest to be **very meticulous** about this assignment. This is your opportunity to get detailed, helpful feedback that you can incorporate into the final project. You are welcome to change your dataset, variables, or research questions for the final project. However, this will force you to double your work, and so isn’t recommended.