Final project

## Overall goal

Your objective is to do an analysis of your choice, on data of your choice, following a few general steps:

* Clean the data
* Analyze the data
* Present results

## Specific objectives

Along the way, you should make sure you hit all these skills, though not necessarily in this order:

* Use mutate() to change/create at least 3 variables.
* Convert at least one of these variables from numeric to factor by creating categories
* Use at least two fct\_ functions to manage your factor variables
* Use select() and at least one select helper (e.g. starts\_with()), at least once
* Use case\_when() at least once
* Apply some eligibility criteria (at least two criteria) to your dataset using filter()
* Use the pipe ( |> or %>% ) to link a few of these steps together
* Create a Table 1 of descriptive statistics about your sample
* Fit a regression and present well-formatted results from the regression
* Create at least 2 ggplots of different types (geoms), these should be reasonably well-formatted for publication (e.g., have appropriate labels, a title, etc.)

and at least one of these:

* Use a \_join() function
* Deal with missing data in a thoughtful way (i.e., more than just using na.rm = TRUE)
* Read in your data and save your results in different directories within an R project using the {here} package

## Presentation

Provide a short description (~ 1 paragraph) of the data and your overall goal for the analysis.

In addition, describe *in words* what you are doing throughout your code, either in code comments or in text between code chunks. For example, when using filter() to restrict your data to meet some criteria, tell us: “Restrict data to older women with high blood pressure.” If you are ever worried that it might not be clear where you are meeting one of the required points, feel free to point it out in your description!

This is not a statistics or epidemiology class, so you don’t have to interpret your findings (they don’t even have to be meaningful, as with most of our examples in class!), but you are welcome to.

Include all of your code and output in a quarto document. Render your document to html, docx, or pdf and submit on Canvas.

## Data

You may choose any dataset you’d like. Options include:

* A dataset you have used/are using for research
* An expanded version of the NLSY dataset using other variables you download at <https://www.nlsinfo.org/investigator/pages/search>
* A [dataset](https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html) that comes with R (load it using data(name\_of\_dataset))
* Any other dataset of your choice

The data you use should have at least 5 meaningful variables (i.e., not id), including some numeric variables that you can turn into factor variables, and it should have at least 100 rows. It should be interesting enough to motivate you to create some interesting tables and plots! If you are not sure whether it will be appropriate for this project, feel free to email us to check.

## Guidelines

Your work must be your own, in that it reflects your decisions about how to analyze the data, what variables to use, what results to present, etc. If you are using the same dataset as a classmate, these choices must be substantially different. When you get an error you just can’t fix, you are welcome to use your classmates, the class materials and resources, and the internet, keeping in mind the requirement for it to be your own work, and the focus on the particular skills and tools we learned in class. When you use an internet resource, please **cite it** (just a link is fine, don’t need formal citations), recalling that you are bound by Harvard University and the Harvard T.H. Chan School of Public Health School’s standards of Academic Integrity.

## Grading rubric

| Criteria | Points |
| --- | --- |
| *Overall goal* |  |
| Analysis objective is clear and well-defined | 2 |
| Dataset is appropriate and has at least 5 meaningful variables with sufficient rows | 1 |
| *Specific objectives* |  |
| mutate() used to change/create at least 3 variables | 1 |
| Conversion of at least one variable from numeric to factor using categories | 1 |
| Use of at least two fct\_ functions to manage factor variables | 1 |
| Use of select() and at least one select helper function | 1 |
| Use of case\_when() at least once | 1 |
| Application of at least two eligibility criteria using filter() | 1 |
| Use of the pipe (|> or %>%) to link multiple steps | 1 |
| Creation of Table 1 with descriptive statistics about the sample | 2 |
| Regression is fitted and well-formatted results are presented | 2 |
| Creation of at least 2 well-formatted ggplots of different types | 2 |
| *Additional objective* |  |
| Use of \_join() function, or dealing with missing data in a thoughtful way, or using the {here} package | 2 |
| *Code description* |  |
| Clear and detailed description of steps taken in the code | 2 |
| Any internet resources used are cited | 1 |