

## Data Description & Descriptive Statistics

### Goal

The goal of this data assignment is to understand the variables in your dataset and their connections with each other. Your task is to collect and describe a set of data of your choice and to perform some descriptive statistical analyses. The hardest part will be finding an appropriate dataset to use. Additionally, you will want to think carefully about the observational units (rows) in the dataset, they must be independent.

The report should include:

- The source of your data and a description of all relevant variables. What is the observational unit (i.e., row)?
- Appropriate summary statistics with adequate explanation / interpretation (Use R. You do not need to give definitions of the statistics, but you should indicate what the statistics say about your data. Try to examine binary or other relationships as well as describing individual distributions.) You might think of making an appropriate table of some kind. Also, if you want, the package `skimr` has a function called `skim` which is great for summarizing data succinctly.
- Graphical displays with adequate explanation / interpretation (These should effectively summarize your data and point out any interesting features. You do not need a picture or table for every variable. Be careful with the word *normal*. Unless you actually check that the data are normal, stick with words like symmetric or bell-shaped.)
- A comment on anything of interest that occurred in doing the project. Were the data approximately what you expected or did some of the results surprise you? How did the sampling go? Do you think you got a representative sample of your population?

### Data Limitations

- The dataset must include at least 10 variables, with at least 4 independent quantitative and at least 4 independent categorical variables. (Label/ID/name does not count as a variable because it cannot be used in a model.) Please be sure to use full variable names / descriptions in your sentences (or make your abbreviation clear to your reader).
- The dataset should have at least 100 **independent** cases / observations (ideal number of observations is 200-400). [Be wary of missing observations!!]

- If you happen to want to use a dataset which is too big to fit on your computer, I can help you set up using the data on a server with SQL commands (but you'd need to know/learn SQL to access it).
- Be very careful with time (e.g., year) as a variable because it can be an indication that your observational units are not independent.
- Because we will be doing hypothesis testing as the next step, you need to indicate what population your data describes. If it is a census, then maybe it is representative of an even larger population? (For example, a census of state information from 2015 might be somewhat representative of 2016? Is it?) Also, discuss the limitations of describing a larger population.
- Ideas of data sources: <https://hardin47.netlify.app/courses/data/>.

## Format

DO:

- use captions for every plot; e.g., in the chunk command give the caption:  

```
““{r fig.cap = "here is the caption"}””
```
- use complete sentences.
- annotate everything that the reader sees.
- keep the file to 4 or fewer pages (mostly graphics).
- turn in both the .Rmd and .pdf file (via email to me). Make sure the .Rmd file can knit!

DON'T:

- do not print any warning or error messages. Only print code that is interesting and relevant to the reader (e.g., use `echo=FALSE`).
- do not print lists of data.
- no overplotting (use boxplots instead of scatterplots when appropriate; use `alpha=0.1` for transparent plotting symbols).
- **no linear models** for this assignment; **no hypothesis tests or inference** of any kind is expected. If you are curious about relationships in your data, it is possible you could run a t-test or a chi-squared test, but most people will not have a hypothesis test of any kind.
- do not include any tables, output, or graphs which are unannotated.
- do not be tempted to turn in everything you do. Only turn in the interesting parts of the analysis. One of the hardest parts of being a consultant is figuring out what to tell the client.

## Pairs

I strongly encourage you to work in pairs if you would like to. The assignment at hand is a semester-long project, so your pairing would be the same for the entire semester. Pairing has learning and communication benefits, working individually has other benefits. You may choose to work together or alone.