**Note**: You may use generative AI (e.g., ChatGPT) to help with this assignment, so long as you run the code in R/Stata yourself. You are also welcome to work in groups to complete this, but just be sure to create your own document yourself that contains your records of completing the homework.

**Overview**: Please upload a script file (file ending in .R or .do) **and** a Word document in Canvas with the key output you generated. The main point of this homework is for you to explore the program on your own, but please create a polished script file to share with me as well as some of your key output.

**Task 1: Find a dataset on your own and load it into R or Stata**

* Please try to find a dataset that has at least one quantitative variable (can be ordinal).
* If you don’t know where to look for a dataset, ICPSR has a great repository: <https://www.icpsr.umich.edu/web/pages/ICPSR/index.html>. You can also look for some additional ideas here (though this is now outdated and some links are broken): <https://nathanfavero.com/wp-content/uploads/Potential-Data-Sources.pdf>.

**Task 2: Pick 2-5 variables and do some basic description of those variables**

* For each variable, generate descriptive statistics and create a graph or two.

**Task 3: Modify some variables** (note: feel free to skip this if tasks 1 and 2 took you several hours to complete)

* For one of your quantitative variables, convert it to a z-score (standardize it). Then, create descriptive statistics and graphs so that you can compare the variable before/after the transformation.
* Now do the same thing, except make a log transformation. Compare to the original variable.
* Reverse the coding of a variable, so that high and low values are switched. You can do this by multiplying the variable by -1. Sometimes, it will make more sense to instead take the maximum value (or max value + 1) of the current variable and then subtract the variable from this (e.g., reversed\_var = 8 - original\_var).
* Make a new dummy variable (out of a continuous variable or a categorical variable with more than 2 categories)

**Tips for completing this in R:**

**Task 1: Loading data into R**

* Plain text files: see Ch. 8 of <https://r4ds.hadley.nz/>
* Excel files: first convert to plain text (e.g., CSV) **or** see Ch. 21 of <https://r4ds.hadley.nz/>
* Stata, SPSS, SAS files: use the package haven <https://haven.tidyverse.org/>

**Task 2: Descriptive stats**

* Overview: <https://www.statmethods.net/stats/descriptives.html>
* See also Section 14.6 Numeric summaries in Ch. 14 of <https://r4ds.hadley.nz/>

Examples:

Brief summary stats for quantitative variables: summary(data$democrat)

More complete summary stats with the psych package (first run library(psych)): describe(data$democrat)

Tables for qualitative variables (#s): table(data$region)

Tables for qualitative variables (%s): prop.table(table(data$region))

**Task 3: Modifying variables**

* Use mutate(), as describe in Ch. 4 of <https://r4ds.hadley.nz/>

Examples (replace highlighted sections to match your data):

Standardizing (converting to z-scores): data <- data |> mutate(new\_var = scale(old\_var))

Log transformation: data <- data |> mutate(new\_var = log(old\_var))

Log transformation method that avoids converting 0s to missing:

data <- data |> mutate(new\_var = log(1 + old\_var))

Create dummy variable:

data <- data |> mutate(new\_var = if\_else(string\_var == "High", 1, 0))

Create dummy var.: data <- data |> mutate(new\_var = if\_else(old\_var > 3, 1, 0))

**Some helpful Stata commands**

**Reading in data**

Open a Stata file: use "covid.dta"

Open a CSV file: insheet using "hw3.csv"

**Basic data description & visualization**

Histogram: hist *variable*

Boxplot: graph box *variable*

Stripplot (first time, need to install package: ssc install stripplot): stripplot *variable*

Frequency table: tab *variable*

Descriptive stats for quantitative variables: sum *variable(s)*

* To add median, percentiles: sum *variable(s), detail*

**Data management**

Clear dataset so you can read in another dataset: clear

Delete extra data: keep if *variable* == *value*

Bring variables (columns) to the front of the dataset: order *variable(s)*

Create a new variable: gen new\_variable\_name = [expression]

* Example: gen age = 2019 - birth

Recode a variable: recode old\_variable ( old\_value(s) = new\_value ) ( old\_value(s) = new\_value ), gen( new\_variable\_name )

* Example: recode race (2/5 = 0) (1 = 1), gen(white)
* See details on syntax here: <http://wlm.userweb.mwn.de/Stata/wstatrec.htm>