Final project

Your name here

Due 12/7/2022 at 5pm

Note1: Start with the file `final_project_2022.Rmd` (available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F22/tree/main/assignments). Modify that file to include your answers. Make sure you can "knit" the file (e.g. in RStudio by clicking on the `Knit` button). Submit both the Rmd file and the knitted PDF via Canvas

1. Data description, part 1

1a) Introduce your data in its raw form (i.e. before you do any data wrangling or create any variables). Where does it come from? What does the data represent? What are the most important variables? (This answer should be words only – no code or math necessary.)

1b) Briefly tell us about some research questions that this data could help address. (Again, no code or math necessary.)

2. Data wrangling

Because everyone has different data, the data wrangling skills you can display will differ. Use this section to show any data wrangling skills you can apply to your data.

Ideally these tasks will be necessary for subsequent things you want to do with the data, so you might skip this section to begin with and see which data wrangling tasks are necessary for your data analysis. It's okay to do some data wrangling here that is not actually necessary for your project just to show us what you've learned.

Try to include the following:

- create a new variable, and/or recode an existing variable (i.e. convert the values of the variable to something more useful for your project)
- assign new names to variables
- merge one dataset with another (may not be possible in your data, e.g. if you have data from a survey experiment and there is no contextual information provided about the subjects)
- reshape your dataset (`pivot_longer`, `pivot_wider`)

3. Data description, part 2

3a) Present a table of summary statistics for the important variables in your dataset. This should include all of the variables you include in your analysis below.

The code below shows how to use `vtable::sumtable()` (i.e. the `sumtable()` function in the `vtable` package) to make a simple table of summary statistics. The code below uses the `mtcars` dataset that is loaded along with tidyverse; you should make a table for your own dataset.

Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
mpg	32	20.091	6.027	10.4	15.425	22.8	33.9
cyl	32	6.188	1.786	4	4	8	8
disp	32	230.722	123.939	71.1	120.825	326	472
hp	32	146.688	68.563	52	96.5	180	335

```r
### make sure to install the package once on your machine
### via [install.packages(vtable)] to make this code run,
### but don't put install.packages() in your final code.
mtcars |>
  vtable::sumtable(vars = c("mpg", "cyl", "disp", "hp"))
```

Comment on any notable or surprising features of the table.

3b) Use `group_by()` and `summarize()` to present a table that summarizes one variable grouped by levels of another variable.

4. Visualization

Make two figures with `ggplot` that illustrate something interesting about your data. Try to include at least one figure that includes three or more aesthetics (e.g. `x`, `y`, and `color`) and a figure that shows more than one `geom` (e.g. `geom_point()` and `geom_smooth()`).

In words, describe what you think are some interesting take-aways from your visualizations.

5. Inference with one variable

Using an important variable from the dataset (e.g. the dependent variable from the regressions below), and assuming that your dataset is an iid random sample from a large population, construct a confidence interval for the mean of the variable

- using the normal approximation approach
- using the bootstrap

Interpret and compare the resulting confidence intervals.

6. Regression

Here you should present regressions that further explore one of the relationships that you visualized in question 4.

Specifically,

- explain why it might be useful to use regression to explore the relationship in addition to the figure you showed
- run regressions for the same dependent variable with more than one model (e.g. different predictors, transformations of the predictors) and present the results in a single regression table (e.g. using `huxtable::huxreg()`)
- interpret some of the more important results in words being sure to pay attention to units and avoiding causal language unless it can be justified

- compare what the regression says about the relationship to what the visualization said about the relationship
- in at least one model, include an interaction term and interpret it
- interpret the standard error, confidence interval, and p-value of a coefficient from one of the regression models