Final challenge October 18, 2019

Prepare your project

- File -> New Project -> New Directory -> New Project
- Name it something like NLSY and put it in an appropriate folder on your computer
- Within that folder, make new folders as follows:

```
NLSY/
    NLSY.Rproj
    data/
         raw/
         processed/
    code/
    results/
          tables/
          figures/
```

Prepare the data

- Copy and paste `nlsy.csv` into `data/raw`.
- Create a new file and save it as `clean_data.R`.
- In that file, read in the NLSY data and load any packages you need. Make sure you replace any missing values with `NA`. Hint: there are extra missing values in the `age_bir` variable. Also, the variable names might be useful:

```
colnames_nlsy <- c(
  "glasses", "eyesight", "sleep_wkdy", "sleep_wknd",
  "id", "nsibs", "samp", "race_eth", "sex", "region",
  "income", "res_1980", "res_2002", "age_bir"
)
```

- Add factor labels to `eyesight`, `sex`, `race_eth`, `region`, as in earlier slides. Select those variables plus `income`, `id`, `nsibs`, `age_bir`, and the sleep variables. Then restrict to complete cases and people with incomes $<$ \$30,000. Make a variable for the log of income (replace with `NA` if income $<= 0$).
- Also in that file, save your new dataset as a `.rds` file to the `data/processed` folder.

Do some exploratory analysis

- Create a file called `create_figure.R`. In this file, read in the cleaned dataset. Load any packages you need. Then make a `ggplot` figure of your choosing to show something about the distribution of the data. Save it to the `results/figures` folder as a `.png` file using the `ggsave()` function.
- Create a file called `table_1.R`. In this file, read in the cleaned dataset and use the `tableone` package to create a table 1 with the variables of your choosing. Modify the following code to save it as a `.csv` file. Open it in Excel/Numbers/Google Sheets/etc. to make sure it worked.

```
tab1 <- CreateTableOne(...) %>% print() %>% as_tibble(rownames = "id")
write_csv(tab1, ...)
```

Do some regression analysis

- In another file called `lin_reg.R`, read in the data and run the following linear regression: `lm(log_inc ~ age_bir + sex + race_eth + nsibs, data = nlsy)`. Modify the CI function to produce a table of results for a *linear* regression. Add an argument `digits =`, with a default of 2, to allow you to choose the number of digits you'd like. Save it in a separate file called `functions.R`. Use `source()` to read in the function at the beginning of your script.
- Save a table of your results as a `.csv` file. Make the names of the coefficients nice!
- Using the results, use `ggplot` to make a figure. Use `geom_point()` for the point estimates and `geom_errorbar()` for the confidence intervals. It will look something like this:

```
ggplot(data) +
  geom_point(aes(x = , y = )) +
  geom_errorbar(aes(x = , ymin = , ymax = ))
```

- Save that figure as a `.pdf` using `ggsave()`. You may want to play around with the `height =` and `width =` arguments to make it look like you want.