Problem Set 3

Due 14 January

Answer the questions below to the best of your ability. Write clearly, and format your tables and visuals appropriately. You must use `R Markdown` to compose and compile your work. For full credit, `echo` all code chunks, and include your `setup` chunk. Submit your work in hard copy at the beginning of class.

Reverse-engineer a table.

I built the `ssa_water.Rdata` data set as part of a research project with a former graduate student. It includes Demographic and Health Survey data for roughly 624,000 children, aged 6 to 18, in Sub-Saharan Africa. The project was similar to the Pickering and Davis (2012) study linked in this folder.

1. Using the dataset, recreate Table 1 of the Pickering and Davis article to the best of your ability. The numbers will not match exactly because you have different data. However, the table must have the same shape and include the same columns. Note that you may use the abbreviated country/module combo in place of the country name.

Pay my respondents

I conducted two online experiments through a recruitment service called CloudResearch. Respondents who opt to participate follow an external link to our experiment. Upon completion, we generate a unique completion code that respondents report to CloudResearch as proof of their work. Now I need to pay them. Each experiment has a unique data file (`exp10.csv` and `exp11.csv`) that includes the responses to the experiment, completion codes, and the amount of money I owe each respondent as a cash bonus. There is also a CloudResearch id file (`CRid.csv`) that includes the respondent's id number (`rID`) for that survey platform and completion code (`CompCode`).

Your end goal is to combine the three datasets for payment. The dataset must include only the following variables and in this order:

- `AssignmentID`
- `rID`
- `CompCode`
- `Bonus`

The final dataset should include *only* respondents whose `CompCode` in the id file matches an assigned code from the data files.

2. Rather than sending a fully joined, final dataframe, please copy and paste the output from a call to `str()` (e.g., `str(finaldat)`).

Winning Wordle

Do you play NYT Wordle? I do... obviously. It's a simple word game that gives you six tries to guess a 5-letter word. If you're not familiar, go try it out. Naturally, I downloaded a file of the nearly 13,000 5-letter words accepted by Wordle (`WordleDictionary.Rdata`). Now I need your help finding good Wordle words.

The Wordle dictionary data includes the following variables:

- `word`: the 5-letter word
- `let1:let5`: the 1st, 2nd, etc. letter in the word
- `dups`: logical vector equal to TRUE if any letter is repeated within the word

3. Start by identifying the most common letters. Present a table (in descending order) of letters and their frequency of appearance in the dictionary. Hint: think about lengthening the frame to count by letter.
4. Use the frequencies you just calculated to score each word (e.g., I find the word "aahed" gets a score of 22,855). Report the highest and lowest scoring words and their scores.
5. Presumably I don't want to use words with duplicate letters. Among those with no duplicates, what are the best/worst opening words according to our standard?