hw04 - Data Management Assignment

Assignment Overview

By now you should know what variables you want to use, and you've looked over the codebook enough now that you have an idea of some potential problems that you will encounter. This assignment uses your chosen research data, and the variables that you chose in the last assignment when you created a personal research codebook. You will thoughtfully review the variables you are interested in, document which ones will need changed and how.

All raw data needs to stay raw, and all changes need to be documented. You will create a script/code file that will make all changes to the data in a programatically and reproducible way. You will create a single code file that imports your raw data, performs some data cleaning steps, and saves out an analysis ready data set that you will use throughout the semester.

You are **not** expected to have completed data management for every one of your variables under consideration by the submission date. At minimum you must have cleaned the 4 variables you are going to visualize in the next graphing assignment.

Instructions

> For all coding assignments: Any reference of "Math 315 folder" or "class folder" means either a folder on your laptop where you keep all files for this class, OR your `stats_project` in R Studio cloud.

1. Create a new Rmarkdown file named `dm_dataname.Rmd`. (e.g. `dm_addhealth.Rmd` or `dm_caterpillar.Rmd`).
    - Save this into your class folder (i.e. or upload to R Studio cloud).
2. Download your chosen analysis data set from Google Drive.
    - Put this file in a folder named `data`.
3. Read in raw data into a data frame named `raw` in the first code chunk.

```r
raw <- read.csv(path to data)
```

4. Do some data cleaning.
    - You must explain at each step what you are doing and why you are doing it.
    - Write these explanations outside the R code chunks as normal text.
    - Don't forget to confirm that any changes you make actually work.
5. Restrict the variables to only the ones you are investigating.
    - Suggested to use the `select` statement found in the `dplyr` package
    - Example, if you recoded the variable `SEX` from "m" and "f" to be an indicator of `female` (1/0), you want to analyze the new variable `female` and not use the old variable named `SEX`.

```r
clean_ah <- raw %>% select( list the variables to keep (or drop) here)
```

6. Save the resulting data set to your `data` folder as `datasetname_clean.Rdata` e.g. `addhealth_clean.Rdata`.
    - This will serve as your analysis data set to do all your subsequent assignments on.

```r
write(clean_ah, "data/addhealth_clean.Rdata")
```

Submission instructions

- This is an individual project assignment

- You will compile your RMD code to create a PDF file (knit to PDF)
- Rename your PDF `dm_datasetname_username` e.g. `dm_addhealth_rdonatello`
- Upload your PDF file to the **hw04 Data Management** folder in Google Drive.

Example

Here are a few examples of data management / data cleaning files on other data sets.

- Any of the Data Management files listed here: https://norcalbiostat.netlify.com/data/raw_data/
- Data set on depression https://norcalbiostat.github.io/AppliedStatistics_notes/import-data.html