Data Management Assignment

Purpose

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 05 graphing assignment.

Coding instructions

1. Use your `dm_dataname.Rmd` code file created last week.
2. Read in raw data into a data frame named `raw` in the first code chunk.
3. Restrict the variables to only the ones you are investigating.
    - Suggested to use the `select` statement found in the `dplyr` package
4. Do some data cleaning.
    - You must explain at each step what you are doing and why you are doing it.
    - Don't forget to confirm that any changes you make actually work.
5. Write out 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.

Submission instructions

- This is a peer-reviewed team assignment
- You will compile your RMD code to create a PDF file (knit to PDF)

Draft - Google Drive

- Knit to PDF
- Manually change the file name of your PDF to include your user name
    - E.g. `dm_addhealth.pdf` becomes `dm_addhealth_DonatelloCoia.pdf` before upload.
    - If you forget to do this your file WILL be overwritten in Google Drive by the next person who forgets.
    - Don't forget that if you knit the code file again (i.e. you forgot to add something) you will have to manually rename your file before uploading again.
- Upload the PDF file to the `04 Data Management` folder in Google Drive.

Peer Review

If there is not 4+/4- things to comment on, write 1 thing you learned and provide help on 1 thing. (renaming a variable, using headers, turning variables to lower case, organization, coding for a new variable etc.)

Final - Blackboard

- Upload your RMD file to the `hw04 Data Management` assignment in Blackboard Learn.
- I will download this file, change the path and run it on my computer.
- It must run for credit. If it compiles to PDF for you then this should be no problem for me.

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