Data Management

Purpose

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. 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 create a single code file that imports your raw data, performs some data cleaning steps, and saves out an analysis ready data set.

1. Read in raw data.
2. Restrict the variables to only the ones you are investigating.
    - R users use the `select` statement, SPSS users use `KEEP`.
3. Do some data cleaning. You don't have to have all variables cleaned by the submission date but you will be coming back to this code file very often during the semester.
    - 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.
    - SPSS users may perform tasks using point & click, but the code must be pasted into a `.sps` file that can be run on command.
4. Save the resulting data set to your `data` folder as `datasetname_clean.Rdata`, or `datasetname_clean.sav`. e.g. `addhealth_clean.Rdata`.

This code file must run successfully on my computer to earn credit. You can keep submitting files until it works.

See the DM file for Add Health from last fall: https://norcalbiostat.netlify.com/data/dm_addhlth.html as an example. *Warning, do not copy this code from the HTML file directly. It will contain special characters that will prevent your code from working. Plus the sleep variable didn't quite work the way we intended.*

Submission

- Upload your data management file to the appropriate assignment folder in Google Drive.
    - Use this filename: `dm_userid.Rmd` for R users, or `dm_userid.sps` for SPSS users.
    - I would advise **not** changing the copy of the `dm` file that you will be using later on. Only the one you upload for me to review.