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.

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)
- Upload your RMD file to the `hw04 Data Management` assignment in Blackboard Learn.
    - I will download this file 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.
- Upload the PDF file to the `04 Data Management` folder in Google Drive.
    - Manually change the file name of your PDF to include your user name before you upload to Google Drive.
    - 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.

You are not expected to have completed data management for every one of your variables under consideration by the submission date. I want to see a VERY good effort has been made (raw data read in, a half-dozen or so variables dealt with, analysis data saved out.)

Questions to ask

Ask yourself and your partner these questions as you go through this exercise:

- Do you need to restrict the data to only one subgroup? (i.e., if you plan to use only participants of a certain age or sex or ethnicity, etc.).
- Do you need to code out missing data? (i.e. was 99 entered as a missing code?)
- Do you need to code out skip patterns? (i.e. are certain codes entered in for questions that won't make sense for a particular group? For example non-smokers may have a response code of `-1: I don't smoke` to a question of "How many cigarettes in the past month have you smoked?")
- Do you need to make response codes more logical?
  - Think about how the "yes" and "no" variables are coded. Does `NO = 0` and `YES = 1`?
  - Think about how the "strongly agree" to "strongly disagree" variables are coded. Do the numbers make sense?
- Consider collapsing a quantitative variable into categories based on percentages of the data you find after examining the frequency table
- Consider collapsing across categories - maybe going from 5 dummy codes for strongly agree, agree, neutral, disagree, strongly disagree to 2 dummy codes that represent strongly agree and agree as one dummy code and disagree and strongly disagree as another code, then neutral coded as missing
- Do you need to create secondary variables? I.e. If you are working with a number of items that represent a single construct, it may be useful to create a composite variable/score.
  - For example, I want to use a list of nicotine dependence symptoms meant to address the presence or absence of nicotine dependence (i.e., tolerance, withdrawal, craving, etc.). Rather than using a dichotomous variable (i.e., nicotine dependence present/absent), I want to examine the construct as a dimensional scale (i.e., number of nicotine dependence symptoms). In this case, I would want to recode each symptom variable so that `YES = 1` and `NO = 0` and then sum the items so that they represent one composite score. In the code below, the `nd_sum` is the new variable I am creating and the variables after of are the variables I am totaling up.
    * `nd_sum=sum (of nd_symptom1 nd_symptom2 nd_symptom3 nd_symptom4)` [THIS IS SAS CODE DO NOT USE]

Quite often you need to peek at the `head()` of the data set or to do a `table()` to check that what you thought you did actually did what you wanted to.

Call out specific variable names. For example

- I am going to use gender in my analysis, and the variable `BIO_SEX` currently is `=1` for males and `=2` for females, and `=6` for unknown. I want an indicator of female so I will create a new variable `female` where `female==1` when `BIO_SEX==2`, `female==0` when `BIO_SEX==1` and `female==NA` when `BIO_SEX==6`.

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

Other resources

- R Cookbook http://www.cookbook-r.com/
- Quick-R http://www.statmethods.net/