Preparing Data for Analysis

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. I want to see a VERY good effort has been made (raw data read in, at least 2 quant and 2 cat variables dealt with, analysis data saved out.) You can always check the rubric in Blackboard learn for more grade specific details.

Instructions

1. Create a `dm_dataname.Rmd` or `dm_dataname.sps` script file.
   - Rename this to include your username. I.e. `dm_dataname_rdonatello.Rmd`.
   - Make sure this is your **math615/script** folder, and your data is in your **math615/data** folder
2. If you have not done so yet, restrict the variables to only the ones you are investigating.
   - R users - use the `select` statement found in the `dplyr` package
   - SPSS users use `KEEP`
3. One by one, check each variable for necessary adjustments. Complete each of the following steps for each variable.
   - First explain in English what the variable name is and what it measures.
   - Then examine the variable using the `freq` function in the `descr` package.
   - Identify the data type of the variable using `class`. Does this match with the intended data type?
   - Recode the data as necessary (See ASCN Ch 3 and the Collaborative course notes for help)
   - Always confirm your recodes worked as intended by creating another table or summary. 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`.

Submission instructions

Draft

Upload your PDF file to the `03 Data Management` folder in Google Drive.

- R users compile your RMD file to create a PDF (knit to PDF)
- SPSS users will have to present additional output to demonstrate that the recodes were successful.
  - Include the `FREQUENCIES` code in your `dm` file to create tables or summaries of the modified variables, so that it shows up in your output window.
  - After you are done, close & restart SPSS by opening your `dm` file. Click "Run all"
  - Export your **output** file as PDF and upload it to Google Drive

Peer Review instructions

As a reviewer, this is what you're checking for:

- Check that they have 2 quantitative and 2 categorical variables present
- Did they check for missing, out of range values?
- Did they present some sort of proof that a certain recode worked as intended?
- Does their work appear to be reproducible?

Final

Upload your code file to the `03 Data Management/code` folder in Google Drive.

- I will download this file and run it on my computer.
- The file it creates will be the item that is graded.
- You can keep submitting files until it works.

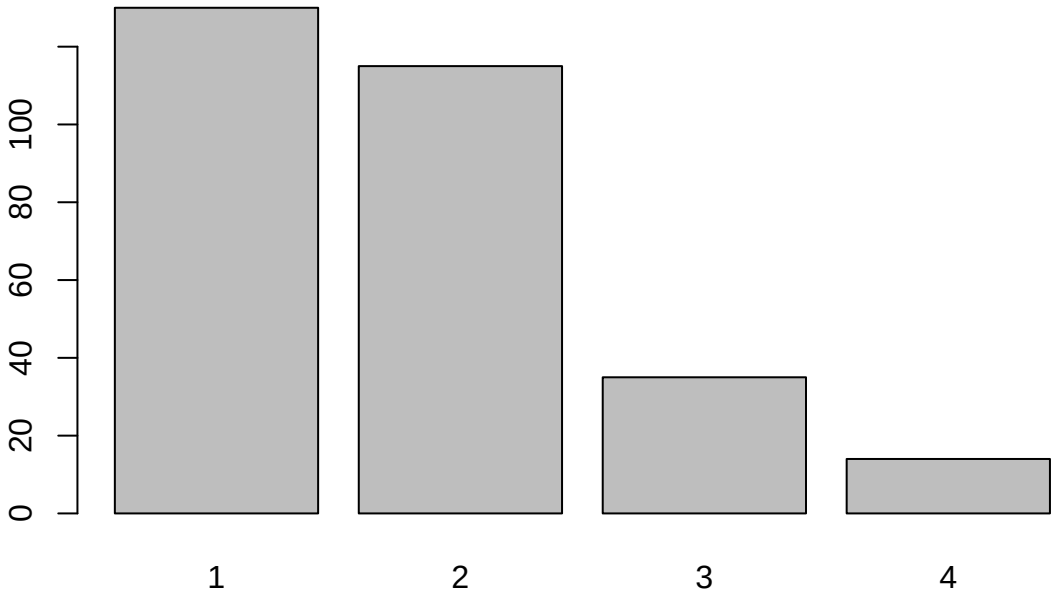
Examples

Example 1: General Health

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(descr)
raw <- read.delim("https://norcalbiostat.netlify.com/data/depress_081217.txt",
                  sep="\t",  header=TRUE)
mydata  <- raw %>% select(age, marital, cesd, health)
```

The variable `health` records a persons perceived general health as being either Excellent, Good, Fair or Poor. This is considered an ordinal categorical variable.

```
freq(mydata$health)
```




```
## mydata$health
##       Frequency Percent
## 1           130  44.218
## 2           115  39.116
## 3            35  11.905
## 4            14   4.762
## Total       294 100.000
```

```
class(mydata$health)
```

```
## [1] "integer"
```

The variable `health` currently is an integer with numeric values 1-4, but the codebook states that this is a categorical variable where 1=Excellent, 2=Good, 3=Fair, 4=Poor. So I need to convert this numeric variable to a factor variable. There are no values outside the 1-4 range, such as a -9 that codes for missing data so I do not need to make any further adjustments (You want to code out missing before you convert variables to factors)

```
mydata$health_cat <- factor(mydata$health, labels=c("Excellent", "Good", "Fair", "Poor"))
```

I will confirm that the recode worked by making a two-way `table`

```
table(mydata$health, mydata$health_cat, useNA="always")
```

```
##
##        Excellent Good Fair Poor <NA>
##   1          130    0    0    0    0
##   2            0  115    0    0    0
##   3            0    0   35    0    0
##   4            0    0    0   14    0
##   <NA>         0    0    0    0    0
```

This shows that all 1's are now 'excellent', 4's are now 'poor' and so forth.

keep only selected variables and save to an external file

```
clean <- mydata %>% select(age, marital, cesd, health_cat)
save(clean, file="depression_clean.Rdata")
```

Example 2: Example DM files on AddHealth

R

The R example is posted at https://norcalbiostat.netlify.com/data/dm_addhlth.html. *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.*

SPSS

You can download and view a copy of a DM file in SPSS for AddHealth using [this link].