**Data Science for Social Scientists**

PSYC 546, Spring 2023

Homework Assignment 3

**Due Date**: February 8th (by 8:15 PM)

**Reminder**: See the assigned Week 2 readings and the lecture slides for a tutorial on how to use R to perform the various functions included in the homework assignment below. **Once completed, you should submit a completed version of this document and your final R script file to the Homework Assignment 3 – Submission Portal on Canvas**.

**R/RStudio**

Your submitted R script file should contain code to answer all of the questions below. Please use comments (e.g., #Question 1) to label the code for each question.

Questions 1 - 6 will use the hw3\_data.sav data set on Canvas.

1. Using the read\_sav() function within the haven package, import the hw3\_data.sav SPSS data file into your R/RStudio environment. If you use the drop-down menu method in RStudio, please still paste the printed code from your console to your saved script file. [1 point]
2. Using any variable recoding method of your choosing (e.g., base-R or tidyverse), create a new variable named **education\_recoded**. Recode the education variable. In the new variable, those without an undergraduate degree should be coded as a 0; those who have completed an undergraduate degree or more should be coded as a 1. Report the frequencies for the two groups below [1 point overall]:
   1. Less than an undergraduate degree (0): 260
   2. Undergraduate degree or more (1): 179
3. In the data set, there are six Likert-type items that measure optimism (op1, op2, op3, op4, op5, op6). We are going to create various mean scale scores to measure participants’ optimism levels.
   1. First, create a mean scale score named **opt\_no\_missing**. This new variable should represent the mean optimism score for a participant on the six items. Specifically, for 3a, participants should only get a scale score calculated if they have data on all six items (i.e., the participant has no missing data). Make sure this new variable is applied to the data frame, and then report the following descriptive statistics on it [1 point overall]:
      1. Number/Sample Size/Count: 384
      2. Mean: 3.71
      3. Standard deviation (SD): 0.727
   2. Similar to 3a, create a mean scale score but name this one **opt\_missing\_ignored**. Now, every participant should get a mean scale score calculated if they have any data at all on the six items (i.e., ignore any missing values). Make sure this new variable is applied to the data frame, and then report the following descriptive statistics on it [1 point overall]:
      1. Number/Sample Size/Count: 439
      2. Mean: 3.67
      3. Standard deviation (SD): 0.757
   3. **Extra challenging one but be brave!** Similar to 3a and 3b, create a mean scale score but name this one **opt\_no\_extreme\_missing** (yes, too long of a variable name but I just wanted to be descriptive here). Now, every participant should get a mean scale score calculated if they have data on at least 3 of the 6 items. That is, if they have missing values on 4 or more of the items, they should not get a mean scale score calculated for them. Make sure this new variable is applied to the data frame, and then report the value below [1 point]:
      1. Number/Sample Size/Count: 432
4. Imagine you find out that there was a data entry error in the data set. Specifically, the individual with an ID value of 215 had their age entered as 23, but you find out their age is actually 32. You do not want to change the value in the actual data file, so create a line of code that recodes the value in the age variable for that particular participant ID with the correct age. [1 point]
5. In Question 6, you are going to export/write this updated data frame to a new data file, because imagine your research teammate requested a copy. To signify your frustration with all the work you had to put into it (especially problem 3c above), you decide to rename all the column names to consist of all uppercase letters (so that they scream at all future users of the data set). This should be able to be accomplished with some quick code coming from a tidyverse renaming function. [1 point]
6. Write the above data frame to a csv file and name it **hw3.csv**. Make sure to upload this file with the rest of your Homework 3 submission. [1 point]
7. For Question 7, download the **race\_recoding.csv** data file from Canvas. In this data file, imagine that data were collected from Qualtrics in a format where each racial category has its own column (with a 1 if a participant self-reported that particular racial group and a blank otherwise). You wish to use race in an ANOVA, which means you need a single column/variable that combines together the information that is currently spread across these multiple columns.

Specifically, using any recoding function of your liking (e.g., base-R or tidyverse), create a new variable called **race**. The coding for this variable should be a 1 if AIAN (for American Indian or Alaska Native), 2 if Asian, 3 if AfricanAmerican, 4 if NHPI (for Native Hawaiian or Other Pacific Islander), and 5 if White. Report the frequencies on the race variable below [2 points overall]:

* 1. AIAN (1): 9
  2. Asian (2): 35
  3. AfricanAmerican (3): 41
  4. NHPI (4): 6
  5. White (5): 109