**Data Science for Social Scientists**

PSYC 546, Spring 2023

Homework Assignment 5

**Due Date**: March 1st (by 11:59 PM)

**Reminder**: See the assigned readings and the lecture slides for a tutorial on how to use R to perform the various functions included in the in-class assignment below. **Once completed, you should submit a completed version of this document and your final R script file to the Homework 5 – Submission Portal on Canvas**.

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.

1. Questions 1-4 use the **hw5\_wide.csv** data file from Canvas. Imagine this data involved an experimental treatment condition (0 = control, 1 = treatment) that was designed to boost satisfaction with life (SWL). SWL was measured three times (before the treatment manipulation, immediately following the treatment, and a 1-month delayed post-test). Your task is going to be taking the data that was entered in wide format and reshaping it to be in long format. First, though, rename the three SWL columns in the data frame to have the names “time1”, “time2”, and “time3”. [1 point]
2. Reshape the data frame from wide to long format and assign it to an object named “long\_df”. The key column should be named “time” and the value column should be named “SWL”. [1 point]
3. Order long\_df based on participant id. Make sure this re-ordering gets applied to the data frame. [0.5 points]
4. Group the data by condition and time and summarize the mean SWL values. Put your answers in the table below. [2 points]

|  |  |  |
| --- | --- | --- |
| Condition | Time | Mean SWL |
| Control | 1 | 4.37 |
| Control | 2 | 3.65 |
| Control | 3 | 4.10 |
| Treatment | 1 | 4.13 |
| Treatment | 2 | 4.10 |
| Treatment | 3 | 3.95 |

1. Questions 5-8 use the **hw5\_join.xslx** data file on Canvas. You will first want to import the two sheets included in the data file as separate data frames. [0.5 points]
2. You eventually will be merging these two data frames together. However, you first notice an issue. Apparently, when entering the data, the RA in your lab thought that they were supposed to change the 1 at the beginning of each participant id to a 2 in the wave2 data frame (e.g., participant 1001 is now 2001). As a result, you first need to transform the participant id variable in the wave2 data frame to align with the values included in the wave1 data frame. [1 point]
3. There were some participants who dropped out of the study and did not return for Wave 2. Imagine your advisor wants to know if the participants who returned for Wave 2 or dropped out of the study noticeably differed on Wave 1 variables. First, use filtering join methods to create two data frames: one that has Wave 1 data for the participants who returned in Wave 2, and a second one that has Wave 1 data for the participants who dropped out and did not return for Wave 2. [2 points]
4. Then, perform the following summary statistics on the two groups and report the values in the table below. Without running any type of statistical test, simply state below the table which variable appeared to distinguish the most between participants who dropped out of the study and those that returned for Wave 2. [2 points]

|  |  |  |
| --- | --- | --- |
| Condition | Completers | Dropouts |
| N (sample size) | 220 | 31 |
| Mean age | 36.5 | 22.1 |
| Mean social support | 5.64 | 5.10 |
| Mean ability | 99.2 | 107 |

* The variable that appeared to distinguish the most between participants who dropped out of the study and those that returned for wave 2 is “Age.” The difference between the mean age of completers and dropouts is 14.4 years.