**Data Science for Social Scientists**

PSYC 546, Spring 2023

Week 1 – In-Class Assignment

**Due Date**: January 19th (by 11:59 PM)

**Reminder**: See the Week 1 Lecture Slides for a tutorial on how to use Excel and R to perform the various functions included in the in-class assignment below. **Once completed, you should submit a completed version of this document, your completed Excel file, and your final R script file to the Week 1 – In-Class Assignment – Submission Portal on Canvas**.

**Part A - Excel**

Part A involves you using the software program Excel to compute and report descriptive statistics and other basic commands using the **insurance\_data.xlsx** file. This data set contains information regarding insurance policies for 500 commercial properties from a national carrier.

1. In cells Q2 to Q6, type Minimum:, Maximum:, Mean:, Median:, and Standard Deviation:. Then in cells R2 to R6, execute the functions that return these five values for the **InsuredValue** variable. This variable contains the property values reported in dollars. Provide the answers below [0.25 points for each]:
   1. Minimum: \_\_\_\_\_\_\_ 30000\_\_\_
   2. Maximum: \_\_\_\_\_\_ 53410614\_\_\_\_
   3. Mean: \_\_\_\_\_\_\_ 4964410.962\_\_\_
   4. Median: \_\_\_\_\_\_\_\_ 2593682.5\_\_
   5. Standard Deviation: \_\_\_\_\_\_ 6205833.879\_\_\_\_
2. Instead of calculating all these descriptive statistics one function at a time, use the Data Analysis Add-in to calculate Descriptive Statistics for the **InsuredValue** variable. From the output, please report [0.25 points for each]:
   1. Standard Error of the Mean: \_\_\_\_ 277533.3282\_\_\_\_\_\_
   2. Sample Variance: \_\_\_\_\_ 3.85124E+13\_\_\_\_\_
   3. Skewness: \_\_\_\_ 3.402978591\_\_\_\_\_\_
3. You are concerned about the positive skewness of the InsuredValue variable. As a result, you want to perform a natural log transformation of the variable. In cell K1, label the new variable, **InsuredValue\_nlog**. Then in column K, apply the natural log transformation to the property values contained in column F. [1 point]
4. Columns I and J contain whether the property is insured against earthquakes and floods, respectively. You want to recode these values into 0 and 1 instead of N and Y, respectively. In cell L1 and M1, label the two new variables **Earthquake\_recoded** and **Flood\_recoded**. Then, in column L, use an if() statement to recode the earthquake data into 1s (for Y) and 0s (for N). Finally, do the same thing in M for the flood data (\*Hint: applying the function in column M should not require you to type the function again). [1 point]
5. Instead of the raw values for the property values, you want to transform the property values into *z* scores using the STANDARDIZE() function. In cell N1, label the new variable **InsuredValue\_standardized**. Then, in column N, convert the property values to *z* scores referencing the means and standard deviations in cells R4 and R6, respectively. You should figure out how to use the $ sign in this function so that you do not need to type it out 500 times! [1 point]

**Part B – R/RStudio**

Part B involves you using the software program RStudio to compute and report descriptive statistics and other basic commands using the **survey.csv** file. A description of the data set is as follows:

“This is a real data file, condensed from a study that was conducted by a Graduate Diploma in Educational Psychology students. The study was designed to explore the factors that impact on respondents' **psychological adjustment** and **wellbeing**. The survey contained a variety of validated scales measuring constructs that the extensive literature on stress and coping suggest influence people's experience of stress. The scales measured self-esteem, optimism, perceptions of control, perceived stress, positive and negative affect, and life satisfaction. A scale was also included that measured people's tendency to present themselves in a favorable or socially desirable manner. The survey was distributed to members of the general public in Melbourne, Australia and surrounding districts.”

Your submitted R script file should contain code to answer all of the questions below. Please use comments (e.g., #Question 6) to label the code for each question.

1. The variable **Moptim** contains the mean score on an optimism measure with a 1-5 scale. Use the describe() function from the “psych” package to calculate basic descriptive statistics for this variable. Provide the answers below [0.25 points for each]:
   1. Mean: \_\_\_\_\_ 3.68 \_\_\_\_\_
   2. Standard Deviation: \_\_\_\_ 0.74 \_\_\_\_\_\_
   3. Minimum: \_\_\_\_ 1.17 \_\_\_\_\_\_
   4. Maximum: \_\_\_\_\_ 5 \_\_\_\_\_
   5. Standard Error: \_\_\_\_ 0.04 \_\_\_\_\_\_
2. Using the mean calculated from 6a, create a new variable in the data set that is the mean-centered version of the Moptim variable. Name this new variable **Moptim\_centered**. (\*Hint: beware of missing data on this one). [1 point]
3. Use the class() function to figure out the type of variable for [0.25 points each]:
   1. age: \_\_\_ "integer" \_\_\_\_\_\_\_
   2. Mlifesat: \_\_\_\_ "numeric" \_\_\_\_\_\_
4. You are a teacher with a small class of students. The students recently took a test and received the following scores:

83 91 75 85 85 67 94 78 70 89

Use the c() function to create a vector of these test scores. Assign it to an object with the name **test­\_scores**. [1 point]

1. Using the **test­\_scores** vector created in Question 9, use appropriate direct functions to calculate [0.25 points each]:
   1. The length of the vector (i.e., the number of students): \_\_\_ 9 \_\_\_\_\_\_\_
   2. The average (mean) test score: \_\_\_\_ 81.33333 \_\_\_\_\_\_
   3. The median test score: \_\_\_\_\_ 83 \_\_\_\_\_
2. Remove the **test\_scores** object from your current RStudio environment [0.50 points].