**Data Science for Social Scientists**

PSYC 546, Spring 2023

Homework Assignment 1

**Due Date**: January 25th (by 8:15 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 homework 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 Homework Assignment 1 – 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 **workplace\_safety\_data** file. This data set contains information regarding workplace injuries for a national construction company from 2020-2022. Imagine you work in in the Risk and Compliance Department of the organization and are tasked with analyzing the workplace incidents. Here is a codebook for the variables:

* **Date**: Date of safety incident
* **Injury Location**: location on body where injury occurred
* **Gender**: employee gender
* **Age Group**: employee age group (18-24, 25-34, 35-49, 50+)
* **Incident Type**: Type of safety incident, e.g. Burn, Fall, Vehicle
* **Days Lost**: number of days employee off work
* **Plant**: manufacturing location where incident occurred
* **Report Type**: Type of safety report filed - First Aid, Lost Time, Medical Claim, Near Miss
* **Shift**: work shift when incident occurred - Afternoon, Day, Night
* **Department**: work department where incident occurred
* **Incident Cost**: total cost of safety incident
* **WkDay**: 3-letter code for day of the week
* **Month**: month number
* **Year**: year number

1. In cell S2 type Incidents:. Then in cell T2 use the count() function to calculate the total number of incidents. You can provide the dates in Column A as the range for this function. Provide the answer below [0.50 points]:
   1. Incidents: \_\_\_\_ 514 \_\_\_\_\_\_
2. In cell S3 type Males:. Then in cell T3 use the countif() function to calculate the number of males that experienced a workplace injury for the company. Provide the answer below [0.50 points]:
   1. Males: \_\_\_\_\_ 458 \_\_\_\_\_
3. In cell S4, type Proportion Males:. Then in cell T4 reference cells T2 and T3 to calculate the proportion of total number of incidents that were experienced by males. Provide the answer below [0.50 points]:
   1. Proportion Males: \_\_\_\_ 0.891050584 \_\_\_\_\_\_
4. Imagine that if the cost of an injury exceeds $1,000, it is categorized as a major workplace incident by the company’s insurance provider. In cell O1, type in a new variable name called **Major Incident**. Then in column O, use an if() statement to categorize the incident costs contained in column K as major incidents or not. If the incident cost exceeded $1,000, have it coded in column O as a 1 (if not have it coded as a 0). [1 point]
5. In cell S5, type Proportion Major Incidents:. Then in cell T5 calculate the mean of column O to figure out the proportion of workplace injuries that were categorized as major incidents. Provide the answer below [0.50 points]:
   1. Proportion Major Incidents: \_\_\_\_ 0.43385214 \_\_\_\_\_
6. Your supervisor wants the relative Incident Cost compared to the average cost calculated for each workplace injury. First in cell S6 type Average Incident Cost:. Then in cell T6 calculate the mean of column K to get the average incident cost. Then, in cell P1 type in a new variable name called **Relative Cost**. Finally, in column P calculate the difference between each incident in column K and the average incident cost contained in cell T6. (\*Hint: the use of a symbol should make it that you only need to type this formula once and not hundreds of times!). [1 point]
7. Your supervisor would also want some descriptive statistics on the lost employee days due to the injuries. Instead of calculating a bunch of descriptive statistics one function at a time, use the Data Analysis Add-in to calculate Descriptive Statistics for the **Days Lost** variable. From the output, please report [0.25 points for each]:
   1. The modal number of lost days: 0
   2. The mean number of lost days: 0.736381323
   3. The maximum number of lost days: 5
   4. The total sum of lost days for the company: 378.5

**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. Your submitted R script file should contain code to answer all of the questions below. Please use comments (e.g., #Question 8) to label the code for each question.

1. The variable **smoke** contains whether the participant self-reported being a smoker (1) or not (0). Run a categorical descriptive analysis (i.e., a frequency analysis) to figure out how many participants reported being a smoker or not. Provide your answers below [0.50 points each]:
   1. Number of smokers: \_\_\_\_\_85\_\_\_\_\_
   2. Number of non-smokers: \_\_\_351\_\_\_\_\_\_
2. The variable **Mslfest** contains the mean scale score on a self-esteem measure with a 1-4 response 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: \_\_\_ 3.35 \_\_\_\_\_\_\_
   2. Standard Deviation: \_\_\_ 0.55 \_\_\_\_\_\_\_
   3. Range: \_\_\_\_\_ 2.6 \_\_\_\_\_
   4. Standard error: \_\_\_\_ 0.03 \_\_\_\_\_\_
3. You wish to do some analysis with the **age** variable in the data set, however you are concerned about skewness in the distribution of ages. As a result, create a new variable in the data set called **age\_squareroot** that consists of performing a square root transformation of the age variable. [1 point]
4. There is a variable called **Mnegaff** in the data set that contains the mean scale score on a negative affect measure with a 1-5 response scale. Imagine that you wish to have all the variables in your report coded in a way that higher scores imply more positive outcomes. As a result, you wish to reverse code the negative affect scale scores. Create a new variable in the data set called **Mnegaff\_reversed** that contains the reverse coded values, so that a score of 1 on the original Mnegaff variable would be a 5 on the new reversed variable (\*Hint: look at the Week 1 lecture slides for a similar example, this should be accomplished through a brief single line of code). [1 point]
5. To double-check that you performed the reverse scoring in Question 11 correct, perform mean and standard deviations of the original **Mnegaff** variable and the created **Mnegaff\_reversed** variable. Use the mean() and sd() functions on both variables. If done correctly, the means should be inverted on a 1-5 scale and the standard deviations should be the same. Provide your answers below [0.25 points each]:
   1. Mean of Mnegaff variable: \_\_\_\_ 1.94317 \_\_\_\_\_\_
   2. Standard Deviation of Mnegaff variable: \_\_\_ 0.7205311 \_\_\_\_\_\_\_
   3. Mean of Mnegaff\_reversed variable: \_\_\_\_ 4.05683 \_\_\_\_\_\_
   4. Standard Deviation of Mnegaff\_reversed variable: \_\_\_\_ 0.7205311 \_\_\_\_\_\_