Lab 1 – Class Data Exploration

Name – NetID

Name – NetID [if applicable]

Name – NetID [if applicable]

Assignment Overview:

- You will get a small taste of what it is like to clean data and complete some basic descriptive tasks using Excel Spreadsheets!

What are you submitting?

- For this assignment, you will turn in **two** separate files
  - Your Edited excel file (as a .xlsx file). *Do **not** convert your excel to a pdf or csv. We need to see your formulas.*
  - A word or pdf file with your answers to the written questions (Q1, Q3, and Q9).
- These should both be submitted directly to **Canvas.**

Submission Notes

- **Working Solo?**
  - Great! Please still list your name and netID at the top of your word document.
- **Working in a group of 2-3?**
  - Awesome! Choose **ONE** person to make a submission. Put all of your names and netIDs at the top of the word doc. When we grade your assignment, we will log grades for everyone in your group and upload to Canvas manually! Your partners will receive their scores by approximately September 21st.
- **Ignore your file preview** after submitting into Canvas! **Your Excel file may look strange, and that is ok!** We will download your file to grade it, so this is not a problem.

Question 1: (1pt) Consider this statement: "Data is objective."

Do you agree with that statement? Why or why not? *Answer this before proceeding to question 2! This question will be graded for a thoughtful attempt, so don't worry about having the "right answer" here.*

Question 2 (5pts) *In your Excel Sheet.* You'll notice that there are *formatting issues* scattered throughout the spreadsheet. Complete these changes in your **Excel sheet** following the advice from the data cleaning video.

- **Leave Alone:**
  - Blank cells are ok as they are!
  - Unusually high or low values are also ok—even if it says 0…or 999,999,999…We can filter those out as needed any time in analysis.
  - You do **not** need to clean multiple choice question columns since they are just categorical options.
- **Change:**
  - Cells that ask for **money** amounts can include $ and , symbols. But make it **consistent** for all values. *Suggestion: Highlight the column, then go to the formatting dropdown and choose "Currency"*
  - All other columns containing numeric data should only include numbers and nothing else. No symbols, no words, no units…just like in the data cleaning video.
  - If someone wrote their answer in words, it either needs to be converted to the numeric equivalent if the meaning is understood, or emptied to a blank cell if not understood.
  - If someone listed a range or ambiguous descriptive answer, use your judgment to choose the data point or simply clear the value. Be as reasonable and consistent as you can! We won't deduct for your subjective choices on this.

- If a cell appears as a *date*, you may either change the format and fix it if understood, or clear the cell if not understood. *I would suggest just clearing those.*

Did you know…many data scientists report that cleaning and organizing datasets is more than half of the work they do? https://www.projectpro.io/article/why-data-preparation-is-an-important-part-of-data-science/242

Question 3 (4pts) *In your Word Document.* Name at least **three** different situations you came across in question 2 where you had to make a subjective choice. What was subjective about that choice and why might someone else make a different choice than you? *Choose different types of changes for full credit.*

Question 4 (3pts) *In your Excel Sheet.* Notice that all of the column names are rather lengthy. Re-name each of these column names such that the title has no more than **12 characters** in length**.** There should be **no spaces or symbols** besides an underscore or hyphen**.** *We **don't** need fully descriptive column headers. We just need something short, abbreviated, and recognizable that we could use in a program like RStudio later on!*

Question 5 (4pts) *In your Excel Sheet.* Notice that the **Academic level** variable lists categorical options of Freshman, Sophomore, Junior, Senior. Since these are ordered entries, we have the option of converting these entries to numbers from 1 to 4. Create *another* column to the right of the academic level column, give it a column name that makes sense, and fill in 1 when academic level is "Freshman," 2 for "Sophomore," 3 for "Junior," and 4 for "Senior/Grad student." *Hint: I did an example of this in the pre-lab videos using sorting and cell dragging!*

Question 6 (5pts) *In your Excel Sheet.* Using the sort function shown in the video, sort the data by **1) Class/section** that students are in (STAT 212 10am first, then 9am following), followed by **2) Academic Level** (Freshmen to Senior), and lastly by **3) Miles from Champaign** (least to most). When you are done, your spreadsheet should have all STAT 200 students at the top, listed in order by Academic level, and further listed by Miles from Champaign. Be careful to sort your spreadsheet so that all of your rows remain intact!

Question 7 (6pts) *In your Excel Sheet.* Apply the AVERAGE(), MEDIAN(), and STDEV.S() functions to the **Sleep, Heart Rate,** and **Shower Time** variables.

- Place these and format them the *same way* you see it in the pre-lab video
    - Directly below the data, with about 1-3 blank rows between the last row of data and your first row of headers for your table.
    - Include your two variable names as a header row for your table and bold these labels.
    - Write Mean, Median, and Standard Deviation on the far left column of your table, and then bold these labels.
    - Use formulas to calculate these statistics for each variable. Be sure the formulas are used—don't just type in an answer.
    - **Round** these statistics to **2** decimal places.
    - Finally, put filled-in borders throughout this space to make it look like a table.

Question 8 (5pts) *In your Excel Sheet.* Make a table to record the percentage of students who answered "A", "B", "C", or "D" to the question about **Choose a letter as randomly as you can.**

- Place these and format them the *same way* you see it in the pre-lab video. It will just be 4 rows instead of 2 rows of percentages.
    - Use the COUNTIF function in your calculation.
    - Convert these two values to percentages using the % option from the menu.
    - Add the labels A, B, C, and D to the left of the percentages, and use borders and bolding to format the table the same way as the previous question. *Note: No need for variable names above.*

Question 9 (2pts) *In your Word Document.* Return to your answer for question 1. Has your view remained the same or changed after completing this assignment? Briefly explain.