Lab 3 – Class Data Exploration

Assignment Overview

- For this assignment, you'll be looking through the data collected from our student data survey this semester!
- You'll get a small taste of the decisions that analysts make when cleaning data, and you will also complete some basic descriptive tasks using Excel Spreadsheets.

Formatting Instructions

- Your submission for Lab 3 will look a little different from other labs.
    - Questions 1, 3, and 8 will be open response boxes in Gradescope. You may write (or copy) in your answers to these when ready to submit.
    - For questions 2, 4, 5, 6, and 7, you will make some changes to the Class Data in an Excel spreadsheet. To submit this work, you may upload your completed Excel spreadsheet in the Question 2 option for Gradescope. Also please make sure your file saves as an excel workbook (.xlsx) when you save it.
- If working with one or two **partners**, be sure to…
    - Have one person make the submission and then ensure **group members** are **added** in your submission to Gradescope.
    - You can do that by clicking **view/edit group** on the top right of the page once shown your final submission).

Step 0

- For this lab, you **won't** be using RStudio at all. Instead, **you will use Microsoft Excel.**
- Watch the **Excel video playlist** linked here.
- With your Illinois account, Microsoft Office apps are free! https://webstore.illinois.edu/shop/product.aspx?zpid=2816
    - If you don't already, I recommend following this link and following the suggestions to install office on your personal computer. Students typically find it more user friendly than using Office Online
    - If you can't install on your computer (or prefer not to), you may simply go to Excel online and upload our class data sheet. Just keep in mind features may be hidden, and you may need to adjust the zoom size using keyboard shortcuts. Ask Google, or ask one of us at Lab Day!

Question 1: (3pts) **Gradescope Free Response:** The first thing we should do before creating any data visualizations, summaries, or analysis is to clean the data. Look at the raw data file to get a glimpse of why that might be necessary! But before we get into that, let's think big picture for a second:

We all know that people can make subjective choices in how they present results from data to others, but would it be fair to think of the data itself that we plug into our visualizations and analyses as ***objective?***

In what ways might you think of the data we use as objective? Can you think of any reasons or situations that might affect the objectivity of the data we use, even when well-intentioned?

- This question will be graded for a thoughtful attempt, so don't worry about giving the "right answer."

Question 2 (6pts) **Uploaded Excel File:** Open the Class Data in Excel. I have done my best to remove seemingly duplicate entries, but other than this, the data is almost entirely in its raw form. Your job is to first clean the numeric variable columns to *prepare* the data for analysis. Our purpose is to represent the class's responses as accurately as we can while getting the data in a consistent form that is ready to analyze.

Categorical response questions can **remain as they are!** *This includes favorite musician/artist!*

You only need to clean the numeric variables. Make sure these variables only contain numbers in the cells (no hyphens, letters, or other symbols). One exception to this rule:

- For **hourly wage** and **expected income**, dollar signs and commas are ok if using a financial formatting. You can click the $ icon, or use the **currency** or **accounting** format from the dropdown on the Home tab to make entries consistent. Any cells that don't format with the rest likely require some cleaning.

Follow the suggestions from the **data cleaning video** as you make decisions. You are being graded for making *reasonable* choices. You don't need to ask us for permission, but we're happy to talk out any situations you come across. A few more quick tips:

- **Rows that have no data in *some* cells are ok.** Don't delete rows just because not all questions are answered—we can still use the data that was inputted!
- **Unusually high or low values are generally ok to stay.** We can always filter outliers out during later analysis. But **if a value is impossible or just nonsensical** to the question, you *may* choose to remove it.
- Complete Question 3 below as you go!

> **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) **Gradescope Free Response:** Name at least **four** different situations you came across in your data cleaning where you had to make a choice that someone else might have made differently. What did you choose to do, and why might someone else have handled it differently? For full credit, you should document four *different types* of situations. Together, they should reveal a *variety* of judgments you had to make.

Question 4 (4pts) **Uploaded Excel File:** Notice that 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**
- You can use an underscore or hyphen to help make it more readable, but **no other symbols** (when we upload to R for Lab 4, this will be important!)
- We **don't** need fully descriptive column headers. We just need something short, abbreviated, and recognizable that is easy to write and reference with code. Analysts often make a variable key separate!

Question 5 (4pts) **Uploaded Excel File:** Notice that students identified themselves as Freshman, Sophomore, Junior, or Senior. Since this variable is ordinal, we have the option of creating a separate column that represents this information numerically from 1 to 4.

- Create *another* column to the right of the column with this data and give it a sensible column name. *Be sure to keep both the original column and your new column in the data once done.*
- Fill in 1 when the student is a "Freshman," 2 for "Sophomore," 3 for "Junior," and 4 for "Senior" *Check the pre-lab video for a quick way to do this without entering them all manually!*

Question 6 (5pts) **Uploaded Excel File:** Using the sort function shown in the video, sort the data to appear in the following order: **1) Introversion score** with *highest* scores at the *top* and lowest scores at the bottom. For students of the same introversion score, they should be further sorted by 2) **Plan after graduation** (this can be alphabetical). Students with same introversion score and same plan should then be sorted by **3) Heart Rate** in numeric order from smallest to largest.

- When you are done, your spreadsheet should have Introversion 4's at the top, with "A job…" as the highest grad plan, and lower heart rates.
- Be careful to sort your spreadsheet so that all of your rows **remain intact**! We won't be able to use data (or worse, make incorrect inferences) if one row no longer represents one person.

Question 7 (6pts) **Uploaded Excel File:** Apply the AVERAGE(), MEDIAN(), and STDEV.S() functions to the **Studying, Sleeping/Napping, Road trip,** and **Expected Salary** variables, and create these in a neat table. *See the pre-lab video for an example of how they should be formatted. Specifically…*

- Please place your table in the rows directly below the data (with about 1-3 empty rows in between)
- Include your four 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 **cell formulas** to calculate these statistics for each variable. We will check your formulas when grading.
- **Round** these statistics to **2** decimal places (median may be reported as whole number)
- Finally, put filled-in borders throughout this space to make it look like a table.

Question 8 (3pts) **Gradescope Free Response**: Return to your answer for question 1. Now that you have completed the assignment (especially Questions 2 and 3), do you agree with your original response? Or do you have any new thoughts to add to your original response? Briefly explain.