**Lab 3 – Class Data Exploration**

**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 (click view/edit group on the top right of the page once shown your final submission after matching pages).

**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.

**Step 0**

* For this lab, you **won’t** be using RStudio at all. Instead, **you will use Microsoft Excel!**
* 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 aren’t able to 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!
* Watch the **[Excel video playlist](https://www.youtube.com/watch?v=iQWC3Q4D5VM&list=PLTE0IJCCTM9LdpwKNRLjoio__qLIrYZDC&index=1)** linked here.

**Question 1:** (3pts) **Gradescope Free Response:** Lots of people know it’s easy to manipulate or cherry-pick data points to convey a message. But is data itself “objective”? Share your answer to this question, and briefly explain your answer.

* Answer this before proceeding to other questions in the lab! This question will be graded for a thoughtful attempt, so don’t worry about giving the “right answer.”
* If working with partners, each partner should briefly write their own response!

**I think data is objective. Because data is something that we directly collect from the real world and represents a character of real things. Data is based on observations and experiments. These processes are usually accurate and objective.**

**Question 2** (6pts) **Uploaded Excel File:** Open the Class Data in Excel. You’ll notice that there are *formatting issues* scattered throughout the spreadsheet. Specifically, we need to clean the numeric variable columns to prepare the data for analysis. Follow the advice from the **data cleaning video** as you make decisions. Specifically, make sure numeric 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 though likely still require some cleaning.

Note that you might have to use your best judgment in some cases! You are being graded for making *reasonable* choices, so don’t feel that you need to ask about every decision. A few more quick tips:

* Entries that have no data in some cells are ok! We don’t have to delete rows just because not all questions are answered
* Unusually high or low values are also ok to stay! We can always filter outliers out during later analysis.

**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 handle it differently? Choose different types of changes for full credit.

1. 你住的多远那一格：I live in Lincoln, 3-hours, a little over 100 miles, 其余都是删除miles，“around 200 miles”改为200，“119-124 miles” 改为中间值
2. 断了几个骨头，B258那列，我删了
3. C100，D148删了，e61-2,G61,g101,g177,range类比如g197，g214,J196,j353
4. G53删除 can i get full credit if i write down the reason why i choose this kind of change and others do not change like that

But what if most people actually have a same choice (they do the data cleaning in the same way?)

还有我表格的空格怎么办

1.Situation: G53 in the original Excel, the annual salary is a range($150-200,000). I choose to delete the cell because I can not choose a number in this range, other people may choose the middle salary $175000 because middle value sometimes represents the person’s thoughts on the question. 2.Situation: A28, “a little over 100 miles”. I choose to delete the cell because I can’t decide a number which is a little over 100, but other people may choose to use 100 or a number slightly over 100 because they might estimate the distance. Deleting will lose this information. 3.Situation: G106: “n/a”. I choose to keep the n/a, because the specific question does not apply to this person and it will be different from an empty cell. But other people may delete it because it is not a number. 4.Situation: E61: “3-Feb” This is actually a range 2-3, but excel changes it to 3-Feb. I choose to keep a middle value 2.5 in the cell, because the question asks “hours of exercises(round to 1 decimal place)” and I think if the person is not sure, I can pick 2.5 to represent the duration of exercise. But others may choose to use 2 or 3 because they may treat the “3-Feb” as a binary choice, or they just delete the cell.

**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**
* 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/grad*.* 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
* Fill in 1 when the student is a “Freshman,” 2 for “Sophomore,” 3 for “Junior,” and 4 for “Senior/Grad student.” *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 by **1) Class section** (11am, 12pm) that students are in. Students in the same section should be further sorted by 2) **Academic Level** (Freshmen, Sophomore, Junior, Senior/Grad). And students in the same section of the same academic level should be sorted by **3) Miles from Champaign** (least to most).

* When you are done, your spreadsheet should have all STAT 212 at 11am 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**! We won’t be able to use data (or worse, make incorrect inferences) if one row no longer represents one person.
* ***Hint:*** If you’re struggling to sort by academic level, here’s a hint: There’s a reason you completed Question 5’s task before completing this task. :)

**Question 7** (6pts) **Uploaded Excel File:** Apply the AVERAGE(), MEDIAN(), and STDEV.S() functions to the **Broken Bones**, **Hourly Wage,** and **Sleep** variables, and create these in a neat table. *See the pre-lab video for an example of how they should be formatted.*

* Please place your table in the rows directly below the data (with about 1-3 empty rows in between)
* Include your three 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.

**需要在正下方吗那个表格列，然后sleep是平均时长还是quality**

**Median的sleep是整数吗？**

**In question 7 , I need to apply functions to the “Sleep” variable, but there are two columns related to the sleep variable**

**Question 8** (3pts) **Gradescope Free Response**: Return to your answer for question 1. Has your answer changed, or remained the same, after completing this assignment? Briefly explain.

Changed. Because when the data is not cleaned, I can see some ambiguous data. Maybe that’s because someone is not sure about what he wants to fill in the form. For example, in the “expected future salary” column, some people will give a range, which will bring some difficulty and inaccuracy to our data analysis. So some types of data are not entirely objective.