**Statistical Thinking and Programming Assignment**

Choose **one** of the prompts below and answer each question as best you can. Please reach out for assistance sooner rather than later. I am available Wednesday from 11-1 in the Data Transformation Lab for walk-in office hours. I also try to respond promptly to email and am glad to set appointments to talk. Plus, if a number of people ask similar questions, I’ll know to try and clarify to your classmates.

50% of your grade will be based on the planning and reasoning stage and 50% will be based on running and attempting to interpret the models. All code can be adapted directly from class examples, which are available with annotations in the [github repository](https://github.com/ndporter/pythonDiSC).

You may complete a second prompt as well. If you do, your higher score will be used for your initial assignment grade and the second prompt will count toward any missed points *on this assignment* at 50% of your score, to a maximum of 100% for the assignment (e.g. no extra credit beyond the assignment’s value).

For example, if you scored 60% on one prompt and 50% on another, your final score would be 60% + (50%\*50%) = 60% + 25% = 85%. At the end of this assignment there will also be an opportunity to improve your grade on the first assignment.

**Prompts (choose one or two and answer the questions below the prompts):**

1. You want to know if students who thought the Backstreet Boys and N Sync made music the same length of time were less likely to choose BSB as the better band than students who correctly determined that BSB made music for longer. Use the *bestBand* and *bsbLonger* variables to answer the question as best you can. Notice that *bsbLonger* is a recoded version of the one we used in class; its value is 1 if they found that BSB made music longer than N Sync and 0 otherwise. The preference for BSB or NS should be the outcome.
2. It seems like reading the data into Pandas (downloading from Drive and then pointing to its location) was a problem for many people. To find out, you want to see if there’s a difference in how long the assignment took (*minSpent*) for people who scored perfectly on that question (*readXlsx*) and those who didn’t. Use *minSpent* together with *perfectReadExcel*, the recoded version of *readXlsx* (so 1 is a perfect score and everything else is 0) to investigate this question. Time spent should be the outcome.
3. As a teacher, my goal with this assignment was that students would typically spend between 1 and 4 hours on the first assignment. To determine if this was true, I’d like to be 95% certain that the average time spent on the assignment would be in that range if I were to teach it again. Investigate the average time spent on the assignment using both a summary number for *minSpent* and some kind of statistical estimation.

**For each prompt you choose, answer the following questions by adding your answer following the question :**

1. *What prompt did you choose?*
2. *What type of variable (Nominal or Interval) is each variable you are investigating (there may be one or two)? Note: yes/no variables are coded as 1=yes and 0=no and are 2-category nominal variables. (10 points)*
3. *Which variable is the dependent variable (the outcome or thing you are primarily interested in) and which is the independent variable (the cause or thing that helps explain the dependent variable)? Justify your answer. If you are doing prompt three, explain which type of variable the single variable you are considering is (dependent or independent). (10 points)*
4. *What is one appropriate statistical procedure that can help you answer your research question and decide whether the finding is likely to be due to random chance. Use the* [*online version*](https://github.com/ndporter/pythonDiSC/blob/master/VariableTypesTable.JPG) *of the chart we made on the board to help you decide. (15 points)*
5. *Use the annotated class notes to help you describe what information (results) you will get from the statistical procedure in (4) and how you will know the answer to the research question in the prompt. (15 points)*
6. *Try to perform the procedure in Python using the datafile we used in class. Copy and paste both the code and the output below even if you don’t successfully run the procedure. You will get at least partial credit for any steps you attempt, as long as I can see what you tried. Complete each of the following steps:*
   1. *Import any packages necessary (5 points)*
   2. *Read the datafile into pandas from github (5 points)*
   3. *Remove any cases that are missing on any variables you’re using (10 points)*
   4. *Run the procedure (10 points)*
   5. *Write a brief explanation of what the results tell you about the answer to the research question (20 points)*

**If you would like to earn additional points on the first Python assignment (again to a max of 100% plus the original bonus), please use the same data as above (the scores) and follow the prompts below, pasting any code and output with your answer.**

*You want to understand similarities and differences in how students explain why they think one band is better than another. The responses have been classified into 5 categories: hotttnesss, quality (of music), productivity/longevity, specific members and “other” (everything else) and each person’s response has been coded 1 for a category if it was mentioned and 0 otherwise.*

1. Use a for loop to find and print the proportion of students that cited each reason (regardless of whether they chose BSB or N Sync). Hint: you can make a list of column names to use as input to the for loop.

Because the variables are coded 1/0, this proportion is equivalent to the mean of the variable. Use the mean function in pandas, not describe or other packages. You can look at the key to the first assignment if you’re unsure how to use pandas to get the mean.

1. Choose one of the reasons (not “other”) and do the following:
   1. Find which camp (BSB or NS) gave that reason more often and describe it in numbers (a difference of means) and words. You do not need to use a statistical test.
   2. Brainstorm at least one reason why that difference (assuming it’s not random) might exist. List the possible reason and what additional data or information (if any) you would need to test that hypothesis (reason), even if you don’t know what tests you might need.