**Extra Credit 1**

PSYC 1090

Instructions: Let’s pretend you have been working as a senior data scientist for an exercise science company the last 10 years. Your boss asks you to tidy/wrangle two datasets. Buuuuuuutttt there’s a catch ☺ You also need to teach a new hire how to do this so you need to carefully document each step so the new hire can learn because they didn’t have Shelby to teach them how to do it in college like you did.

**Step One:** Read the below text and answer a few questions about the data.

Your boss conducted a study that wanted to test whether individuals’ degree of openness to new experiences as either low or high (see [here](https://www.psychologytoday.com/us/basics/openness) for a brief review of openness and what the concept means) and the number of times they try a new innovative workout class (out of 3 classes) influences their ratings about how much they like the new class on a scale from 1 *(did not like the class at all)* to 10 *(loved the class)*.

|  |
| --- |
| 1. What is your first independent variable?   **Degree of openness to new experiences**.   * 1. Is IV1 within- or between-subjects?   **IV1 is a between-subjects variable because each participant is categorized as either having a low or high degree of openness to new experiences, and this categorization does not change within the participants**   * 1. Is IV1 categorical or continuous? If it’s categorical, how many levels are there?   **IV1 is categorical since there are two levels, that is low openness and high openness.**   1. What is your second independent variable?   **The number of times an individual tries a new innovative workout class (out of 3 classes).**   * 1. Is IV2 within- or between-subjects?   **IV2 is a within-subjects variable due to the fact that the same participants can try different numbers of workout classes, and their experiences may vary depending on the number of classes attended.**   * 1. Is IV2 categorical or continuous? If it’s categorical, how many levels are there?   **IV2 is categorical since there are three levels.**   1. What is your dependent variable?   **It is the participants’ ratings about how much they like the new class on a scale from 1 *(did not like the class at all)* to 10 *(loved the class)*.**   * 1. Is your DV categorical or continuous? If it’s categorical, how many levels are there?   **My DV is continuous.**   1. What type of statistical test would you need to run to answer the research question your boss is curious about? **I would run a **two-way ANOVA** (Analysis of Variance).** |
|  |
|  |
|  |

**Step Two:** Please download and examine the two CSV files attached to this assignment in Canvas. I have included short snippets of what these datasets look like here, but these are NOT the full datasets. Below are Tables 1 and 2; this is the [messy] data as you received it. You are to turn that messy data into the Table 3 by tidying it.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 1 | | | |  | Table 2 | |
| **PID** | **Class 1** | **Class 2** | **Class 3** | **+** | **PID** | **Openness** |
| 1 | 2.3 | 4.5 | NA | 1 | high |
| 2 | 1.3 | 2.7 | 2.6 | 2 | low |
| 3 | 3.5 | 2.9 | 4.1 | 3 | low |
| 4 | 4.8 | 7.4 | 9.5 | 4 | high |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Table 3 | | | |
| **=** | **PID** | **Class** | **Openness** | **Class Ratings** |
| 1 | 1 | high | 2.3 |
| 1 | 2 | high | 4.5 |
| 1 | 3 | high | 5.4 |
| 2 | 1 | low | 1.3 |

Be sure to document all your steps in the boxes below. There are three computations you need to do in a specific order to turn this messy data into the tidy Table 3 above. Here’s what you need to provide in this document for *every single* computation:

1. Screenshots of your written code in R
2. Screenshots of your data matrix after you ran the code using View()
3. A few sentences describing WHAT you did (to teach the new hire)
4. A few sentences describing WHY you decided to do what you did (to teach the new hire)

|  |
| --- |
| **Computation One**  HINT: Some people skipped class.  **Handle Missing Values for the people who skipped class3 by performing a mean imputation for Class3**  **Screenshot (105)** |

|  |
| --- |
| **Computation Two**  HINT: Don’t think too *\*long\** about this one.  **Merge the tables based on the PID column.**  Screenshot (106) |

|  |
| --- |
| **Computation Three**  HINT: Get it together!  **Select and arrange the appropriate columns needed.**  Screenshot (106) |
| **I took the new hire through a way to handle missing values in a dataset by performing a mean imputation for Class3. This was done to ensure consistency.**  **I took the new hire through a way of merging tables based on a given column. This was done to ensure we work with data in a single table.**  **I took the new hire through a way to select and arrange the required columns of data so that it can be processed and viewed as a table using the View() function.** |

**EXTRA *Extra* Credit**

Note: This is NOT required to get up to 10% added onto your exam grade (that part is entirely above). However, if you want to challenge yourself and answer the below questions, I will add up to 15% onto your exam grade.

Instructions: Your boss is concerned about the missing data from class three. So, you create an entirely new data matrix with data from all variables but only observations from classes one and two. Make that new data matrix.

|  |
| --- |
| **Computation Four**  HINT: Use dplyr.  Screenshot (109) |

|  |
| --- |
| **Computation Five**  HINT: Get it together but in a different way!  Screenshot (110) |