Coding Challenge for Protect Data Team

November 20, 2024

Question 1 - Data Cleaning

In our lab, we use the Hamilton Depression Rating Scale (HAM) to measure our research participants' level of depression. Our clinicians typically score participants at every visit (baseline, 3-month, 1-year, etc). You can find sample data in two data frames, "HAM_protect" and "HAM_sleep", in *Data.RData* from the folder.

Please pull the HAM data of participants of interest and score them. You can find the list of IDs in the folder. Name the cleaned data frame as "final_df".

Instructions

a. Map ID: Our study contains participants from varied sources and they used to have different IDs from what we currently use, i.e. "new_id.
b. Calculate the HAM scores of each participant, then calculate the mean score of each participant and keep only the latest score of each participant.
c. Our study is longitudinal so participants are re-consented routinely throughout their participation as we implement new protocols. Some participants also participate in supplement studies and their data for these are stored elsewhere.
d. For this challenge, get the HAM score that is closest to 1 year after their first consent date. These HAM scores could either be in the main data frame "HAM_protect" or the supplement data frame "HAM_sleep".
e. (Bonus point, optional) Set up a repository on your GitHub with all output files stored in this repository.

Tips

1. Imagine you are the data manager who pulls data for collaborators in the lab. The cleaner your deliverables, the better. You can do what's beyond my instructions.
2. ID mapping: `id_map` is provided in the Data.RData. Please keep only column "new_id" and only our research participants' data in the final data frame.
3. How to calculate HAM scores: sum all the variables starting with "ham_" except 3a to 3e.
4. The variables "bq_date" and "fug_date" indicate when a participant was given HAM. If every variable except ID, visit time point and date is blank, that means the participant was not given HAM at that visit.

Question 2 - Data Visualization

Our research participants are recruited from multiple sources. PIs often want to see the effectiveness of each source in order to better utilize grants. You can find sample data called "recruitment_data". Please use the data to visualize the total number of participants from each source as well as the number by Age, Gender, and Group.

Tips

1. Again, imagine you are the data manager in the lab. The clearer your deliverables, the better. You can do what's beyond my instructions.

Please include your R script in your submission.