Coding Challenge for Protect Data Team

October 4, 2021

Question 1 - Data Cleaning

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

Please pull the HAM data of subjects 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 subjects from varied sources and they used to have different IDs from what we currently use, i.e. "Master Demo ID". Therefore, we always map IDs onto the "masterdemoid" using the `bsrc.findid(..)` function in the BSRC package in R. You may find the package here: https://github.com/PROTECT-lab/redcap_in_r.

b. Calculate the HAM scores of each subject, then calculate the mean score of each subject and keep only the latest score of each subject.

c. (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: `idmap` is provided in the Data.RData. Please keep only columns "ID" and "masterdemoid", and only our research subjects' data in the final data frame. You are welcome to ask for help regarding installing our package. If you still could not figure out how to use our functions, just use "ID" instead of "masterdemoid" to identify subjects.
3. How to calculate HAM scores: sum all the variables starting with "ham_" except 3a to 3e.
4. "ham_date" indicates when a subject was given HAM. If every every variable except ID, visit time point and date is blank, that means the subject was not given HAM at that visit.

Question 2 - Data Visualization

Our research subjects are recruited from multiple sources. PIs often want to see the effectiveness of each source in order to better utilize grants. You can find a sample data called "recruitment_data". Please use the data to visualize the total number of subjects 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.

You can find graph samples in the folder. You don’t need to make the same graphs as I do, but make sure that you include in your graphs all information that I have. Just for your information, I used `ggplot2` package to make the graphs. And the colors I used for “total” was “#8fcaee”, for the other two graphs were R ColorBrewer palette = 1.