Assignment 4

EDH7916 | Spring 2020

Benjamin Skinner

NOTE *This assignment needs to be completed by the start of the next class. That means everything pushed to your remote GitHub repo before class starts.*

Remember, I encourage you to save your work, commit smaller changes, and push to your remote GitHub repo often rather than wait until the last minute.

Task 1

Using the `hsls_small.csv` data set and the online codebook, answer the following questions. You **do not** need to save the final output as a data file: just having the final result print to the console is fine. For each question, I would like you to try to pipe all the commands together. Throughout, you **should** account for missing values by dropping them.

For each question, show your data work and then answer the question in a short (1-2 sentence(s)) comment.

Questions

1. Compute the average test score by region and join back into the full data frame. Next, compute the difference between each student's test score and that of the region. Finally, return the mean of these differences by region.
2. Compute the average test score by region and family income level. Join back to the full data frame. **HINT** You can join on more than one key.
3. Select the following variables from the full data set:
   - `stu_id`
   - `x1stuedexpct`
   - `x1paredexpct`
   - `x4evratndclg`

   From this reduced data frame, reshape the data frame so that it is long in educational expectations, meaning that each observation should have two rows, one for each educational expectation type.

Task 2

If you haven't already, download the raw data you will use for your final project. If this requires point-and-click steps, be sure to save them in a markdown file in your repo so that you can reproduce them later. Please save the data in your `data` directory (it should not push to your remote repo, but that's okay).