Formative assessment: data science mini project

Into the Tidyverse

This week we'll carry on working on the mini project applying your new Tidyverse skills. Remember to keep track of all your changes to the code in your git repository with meaningful commit messages, and to add any packages you use to the Conda environment.

Section 1: Load demographics data

Download the NHANES 2005-2006 demographics data from Blackboard. The data (currently in SAS transport format) are described here: https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DEMO_D.htm

Put the new data in the "original" subdirectory of the data directory, and start a new R code file in a subdirectory of the code directory. Work out which Tidyverse package can read .xpt files. Load it and use it to read in the demographics data. Use the data codebook and dplyr code to make a Tibble containing participant ID (respondent sequence number), gender, age and ethnicity.

Section 2: Merge demographics data

Read in the Body Measurement Data (BMX_D.csv). Reformat the variable names of both the Body Measurement Data and the subset of the demographics data to lowercase. Merge the Body Measurement Data with the demographics data, keeping the rows from the Body Measurement Data. Move the demographics variables to the left hand side of the data frame, just to the right of the participant ID.

Section 3: Merge sample information

Read in the derived sample file that records whether accelerometer data are available. Merge it into the body measurement and demographics data, moving the "in sample" variable just to the right of participant ID and renaming it "in_sample".

Section 4: Create an obesity variable

Use the CDC definitions of BMI categories (https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html) to create a new categorical binary variable called "obesity" using dplyr.

Section 5: Summarise the data

Use dplyr verbs to calculate:

- What percentage of the people in the body measurement sample who are considered to have obesity also have activity monitor data? How does this compare to the people who are not considered to have obesity?
- What percentage of males have activity monitor data? How does this compare to females?
- What is the maximum reported height in cm of a child under 16 in this sample?

Section 6: Export the data

Save the combined data frame to the derived data directory as a CSV called body_measurements.csv

Make sure you've documented the new data and the code you've created in the relevant READMEs. And that your Conda environment and git repository are up to date.

Next steps

If you've finished all of this and still have time, have a look at the links in this week's Homework section on Blackboard.