NHANES Project

NHANES Data

The assignment link is here: https://classroom.github.com/a/wwXfoMO4

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.

Project

We want to demonstrate the skills to connect to the cluster, download data, and perform an analysis. The expected output is a report to be given to a collaborator. Code should be hidden unless otherwise specified. This indicates we should have a statement of the purpose of the analysis/introduction, a description of the data (**not** using names from code, unless relevant), a description of the methods used (including any model formulations if models are used), the results of the analysis. The results section should **reference** figures or tables for the analysis and provide **insight** to what you are seeing. Tell the reader what you are seeing. Then a conclusion/discussion section should provide some distillation of the analysis and what was learned.

The Data

The data in this project is from Wave I (2015-2016) data from the NHANES. Each data set has an online associated codebook. You will likely need to reference these codebooks to understand the data. For example, https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&CycleBeginYear=2015 will list out some of the questionnaires.

Purpose

We would like to show a few concepts:

1. Connecting to the Cluster
2. Interacting with a Database using SQL, dplyr + dbplyr, or Python equivalents
3. Perform an analysis on real-world data, with missingness patterns

Analysis

We would like find associations of smoking, alcohol use, diabetes, and oral health.

We define:

- Alcohol use (ALQ101): as Had at least 12 alcohol drinks/1 year.
- Diabetes status (DIQ010): Doctor told you have diabetes (group borderline as yes)
- Ever smoker (SMQ020): Smoked at least 100 cigarettes in life?
- Oral health issue: (OHAREC) if you are recommended care other than "Continue your regular routine care"

Tasks

1. Name your analysis code file as index.Rmd/index.qmd. The output should be `index.pdf`. The PDF should be generated from the Rmd/qmd.
2. (code should be demonstrated here). Automatically download the `nhanes_wave_i.sqlite` database from the cluster at `/users/bcaffo/dataPublic/nhanes` on the JHPCE computing cluster.

    1. Include the sqlite file **in** your GitHub (it is publicly available data)
    2. Put the file in a `data/` subfolder (this will be checked).

3. (code demonstrated here) Connect to the SQLITE database using the DBI/dplyr/dbplyr packages (in R) or equivalent in Python. Create counts of missing values for diabetes status and alcohol use using SQL commands.
4. Create a "table 1" that describes the population, including, age, gender, education level, smoking status, alcohol use, diabetes status, and oral health issues. Decide how to present missing data.

    1. Make sure you label variables appropriately for a collaborator.

5. Create a table of smoking status versus whether the person was told benefit giving up cigarettes (from (OHQ_I). Discuss missingness and the relationship of these 2 variables with a hypothesis test.

6. We want to determine the effect of smoking and alcohol use on good oral health (outcome). We would adjust for any demographics you think relevant/significant. Discuss how a model was chosen and any performance metrics. Decide whether to adjust for diabetes and discuss why.
7. Create one figure that demonstrates the effect of age on good oral health.

Notes

Do not use `setwd` in your scripts.

Make sure the figure aspect ratios are appropriate (get someone else to look it over!).

Any questions on this project should be posted as an issue on https://github.com/jh-adv-data-sci/adv_data_sci_2023/issues.