Last updated: March 9, 2025

ds1_nhanes

Authors: Silas Decker, Jeannine Valcour, Liliana Bettolo, Tessa Lawler, Christopher Donovan

Introduction

Herein lies the repository for 'Dietary Patterns in the U.S. and Associated Health and Environmental Impact: A Cluster Analysis', a project for CSYS 5870: Data Science 1.

On Working Directories

We have been only somewhat successfully trying to manage this project such that it works in the Colab environment and also runs locally if it is pulled from GitHub. We are currently using the magic function `%cd` in notebooks to persistenly change the directory to the shared Google Drive. Still not clear whether the mounting of Google Drive will work for everyone, so we will have to chat and troubleshoot if it does not.

Anyhow, libraries in the local version of the project will be managed with in a `venv` and recorded in `requirements.txt`.

File Structure

Folders

- The `data/` directory contains raw data, cleaned data, and other miscellany like FPED crosswalks to link USDA food codes to food groups. Raw data should be only `.xpt` files still warm from the CDC. We should also include cleaned datasets saved as `.csv` files, along with a `.info` file that contains provenance and metadata for datasets in each folder.
- `notebooks/` should have `.ipynb` files and contain the main workflow for the project.
- `ds1_nhanes/` is the package directory which will contain modules with functions to use throughout the project.
- The `outputs/` folder should contain graphs and tables ready to throw into Overleaf.
- The `tests/` folder will contain unit testing for project functions and deployment.
- The `.git/` folder is used to manage version control. There is generally no need to edit it directly.
- Folders should have a `.info` file with provenance for the files contained therein, particularly for modified datasets.

Loose Files

- The `.gitignore` file tells git what not to track.
- The `requirements.txt` file is a log of all the libraries that are used in the project.
- The `README.md` is a markdown that produces the html version shown on the repo.

Reproduction

To reproduce the analysis, follow these steps.

1. Create a Virtual Environment

```
python -m venv .venv
```

2. Activate Environment

For Mac/Linux, activate the environment using:

```
source .venv/bin/activate
```

For Windows Commmand Prompt, activate the environment using:

```
source .venv/Scripts/activate
```

For Windows PowerShell, activate the environment using:

```
.venv\Scripts\Activate.ps1
```

3. Install Packages

Install package versions as specified in the `requirements.txt` file with:

```
pip install -r requirements.txt
```

4. Run a Shell or Something

Might make sense to run a shell that runs the `.ipynb` scripts in order?

5. But Really

The plan is to eventually deploy the analysis in a Docker container unless we suddenly decide it isn't worth the trouble in about a month.

Data and Licensing

The code in this project is licensed under the GNU General Public License v3.

NHANES datasets are made available to the public with attribution by the Centers for Disease Control, but are not covered by any license apparently.

The Food Patterns Equivalents Database (FPED) is made available by the USDA apparently without any license or explanation. Which is fucked up because that means it is basically copyrighted. But this is obviously meant to be available to the public. Can anyone find a license for this?

Carbon emissions and cumulative energy demand for food consumption patterns were derived from the dataFIELD database at the Center for Sustainable Systems at the University of Michigan. It is released with a publication from Heller et al. (2018). However, the data are also lacking any license that I can find.

Changelog

- Major changes go here