Last updated: March 14th, 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.

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. They will be imported and called from notebooks.
- 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 `dev/` folder will exist only on Colab and contain a jenky terminal and possibly other tools in development. Generally no need to run anything from here.
- The `.git/` folder is used to manage version control. There is no need to edit it directly or interact with it much at all. Git is managed from the command line.
- 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. This will be rendered and included as a `.pdf` to make it easier to access on Drive as well.
- The text of the license is included in `LICENSE.md`.

Running the Project

This project is set up such that it can be on either Google Colab or cloned and run locally with relative ease.

Colab

The top of each notebook should contain a cell where the Google Drive is mounted to the notebook. Once mounted, this should provide access to the `ds1_nhanes` directory which contains the project and datasets.

Note that this will only work if the `ds1_nhanes` directory is at the top of your MyDrive folder. If it is not, create a shortcut for it there.

The default working directory when opening a notebook is the same directory in which the notebook is located. This is not actually what we want. We want it to be the root of the project directory. So, once the drive is mounted, the next cell should use `os.chdir()` to set the working directory to the `ds1_nhnanes` folder. Then the notebook should be ready to run.

Local

For local use, we will clone the repository from GitHub and reproduce a virtual environment with libraries used in the project. Note that there is currently not a particularly smooth way to deal with different working directories here. As of now, there is a commented out cell at the top of each notebook that the user should un-comment and run to set the proper working directory for local use. Working on a better solution for this.

To run the project locally:

1. Clone Git Repository

```
git clone https://github.com/ChrisDonovan307/ds1_nhanes.git
```

2. Create a Virtual Environment

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

3. 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
```

4. Install Packages

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

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

5. Run a Shell or Something

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

6. 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

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

2025-03-11: Built project structure in Colab, linked Colab with GitHub, and shared all repositories with team members.