Last updated: March 31st, 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 `graveyard/` folder is a place to store trash that might turn out to be treasure one day.
- 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.

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 and the working directory is set to the root of the `ds1_nhanes` folder. 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.

Local

For local use, we will clone the repository from GitHub and reproduce a virtual environment with libraries used in the project. The first code chunk should set the working drive to the root directory with a local path.

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 Shell Script

The `run.sh` script in the root directory will run each of the notebooks in the analysis to refresh the results of the project. You can run it from the command line using:

```
source run.sh
```

Should probably not be saving over our own notebooks like this. Perhaps outputs should end up going somewhere else so we can compare the original results to the reproduced results.

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

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?

Changelog

2025-03-31: Giving scripts number scheme, adding `run.sh` to run all notebooks and refresh outputs.

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