Repository Title: Final_Project_Environmental_Data_Analytics

Summary

This repository will house all materials for my final project for Environmental Data Analytics. My project will seek to examine the occurence of water quality contaminants and their potential relationship to county level median household income, size of the community water system (CWS), the state in which the CWS is located, and so on.

Preliminary research questions include: Is median household income correlated with the occurrence drinking water contaminants such as arsenic, PFAS, trihalomethane, and uranium? Does a high concentration of one contaminant correlate with that of another? Does the size of the CWS or the geographic location of the system correlate with the occurennce of these contaminants? Do the concentrations of these contaminants vary in one particular location over time?

Investigator

Rachel Gonsenhauser Duke University Email: rachel.gonsenhauser@duke.edu Role: Data assembler

Keywords

Drinking water, water quality, water systems, arsenic, per- and polyfluoroalkyl substances (PFAS), trihalomethane, uranium, median household income

Database Information

Data was collected from the Centers for Disease Control and Prevention's (CDC) National Environmental Public Health Tracking Network. This tool can be found at: https://ephtracking.cdc.gov/DataExplorer/#/. Data was accessed on February 26, 2020.

Separate datasets from this CDC database were combined by the data assembler into one processed dataset for analyses.

Folder structure, file formats, and naming conventions

The repository includes the following folders of files:

Code: files include coding sessions including cleaning/wrangling, visualization, and analysis of the data.

Data: files include data from the CDC database discussed above. Data file types include Raw and Processed.

In the Raw Data folder, further descriptive subfolders are provided that characterize each dataset contained within (names of subfolders include: Arsenic, MHI, PFAS, Trihalomethane, Uranium). Within each subfolder are three files: the raw dataset, a footnotes file (containing metadata), and a General Information file containing information about the National Environmental Public Health Tracking Network CSV file).

Output: files will include finished output files once code is run and finalized.

Project Files: files include project instructions, a project rubric, and a template for the final project output.

//

File formats for the files contained in the repository include: `.md`: text file, such as this README file `.Rmd`: R markdown file, the format for all coding and output files `.csv`: delimited text file, the format for all data files `.docx`: Word document, the format for the project rubric file `.pdf`: portable document format, the format for General Information documents that accompany each individual CDC dataset `.htm`: HTML webpage, the format of the footnotes (metadata) documents that accompany each individual CDC dataset

//

Downloads from the CDC's Environmental Public Health Tracking Network included the following files named according to the following name convention in parentheses: `A General Information document (General_Information.pdf)` `Data requested (data_HHMMSS.csv)` `Footnotes, or metadata, specific to each query (footnotes_HHMMSS.htm)`, where:

HHMMSS refers to the time of download in hour-hour, minute-minute, second-second format

Processed data files are named according to the following naming convention: `databasename_datatype_details_stage.forma` where:

databasename refers to the database from where the data originated

datatype is a description of data

details are additional descriptive details, particularly important for processed data

stagerefers to the stage in data management pipelines (e.g., raw, cleaned, or processed)

format is a non-proprietary file format (e.g., .csv, .txt)

Coding files are named according to the following naming convention: `Coding Session_stage.Rmd`, where:

stage refers to the type of coding done (e.g. Data Wrangling, Visualization, Analysis)

Metadata

<For each data file in the repository, describe the data contained in each column. Include the column name, a description of the information, the class of data, and any units associated with the data. Create a list or table for each data file.>

Metadata for **raw** datasets is provided below:

1. File name: data_113417.csv Link to footnotes: file:///Users/rachelgonsenhauser/Documents/Final_Project_Environmental_Data_Analytics/Data/Raw/Arsenic/footnotes_113417.htm

`Columns` **stateFIPS**: Federal Information Processing Standard state code of state measurement was taken in **State**: state measurement was taken in **countyFIPS**: Federal Information Processing Standard county code of county measurement was taken in **County**: county measurement was taken in **Year**: year measurement was taken in **Value**: mean concentration of Arsenic (micrograms per liter) by year **Data Comment**: additional information about data **Maximum Contaminant Level**: whether arsenic value collected exceeds the Maximum Contaminant Level (MCL) set by the US EPA (categories include: Greater than MCL (1), Less than or Equal to MCL (2), and Not Detected (3)) **PWS ID**: Public Water System Identification Number **CWS Name**: Community Water System Name **Population Served**: number of people served by CWS **Maximum Contaminant Level**: whether arsenic value collected exceeds the Maximum Contaminant Level (MCL) set by the US EPA (categories include: Greater than MCL, Less than or Equal to MCL, and Not Detected)

2. File name: data_122206.csv Link to footnotes: file:///Users/rachelgonsenhauser/Documents/Final_Project_Environmental_Data_Analytics/Data/Raw/MHI/footnotes_122206.htm

`Columns` **stateFIPS**: Federal Information Processing Standard state code of state measurement was taken in **State**: state measurement was taken in **countyFIPS**: Federal Information Processing Standard county code of county measurement was taken in **County**: county measurement was taken in **Year**: year measurement was taken in **Value**: Median household income ($) **Data Comment**: additional information about data

3. File name: data_112028.csv Link to footnotes: file:///Users/rachelgonsenhauser/Documents/Final_Project_Environmental_Data_Analytics/Data/Raw/PFAS/footnotes_112028.htm

`Columns` **StateFIPS**: Federal Information Processing Standard state code of state measurement was taken in **State**: state measurement was taken in **countyFIPS**: Federal Information Processing Standard county code of county measurement was taken in **County**: county measurement was taken in **Year**: year measurement was taken in (in this case, data was collected only from January 2013 to December 2015) **Value**: concentration of PFOS, PFAS, PFNA, PFBS, PFHxS, PFHpA (ppt/parts per trillion) **Data Comment**: additional information about data **Status**: whether chemical was detected in water system **PWS ID**: Public Water System Identification Number **CWS Name**: Community Water System Name **Population served**: number of people served by CWS **Contaminant**: type of poly- or perfluoroalkyl substance detected (e.g. PFOA, PFOS, etc.)

4. File name: data_122753.csv Link to footnontes: file:///Users/rachelgonsenhauser/Documents/Final_Project_Environmental_Data_Analytics/Data/Raw/Trihalomethane/footnotes_122754.htm

`Columns` **stateFIPS**: Federal Information Processing Standard state code of state measurement was taken in **State**: state measurement was taken in **countyFIPS**: Federal Information Processing Standard county code of county measurement was taken in **County**: county measurement was taken in **Year**: year measurement was taken in **Value**: mean concentration of trihalomethanes (micrograms per liter) by year **Data Comment**: additional information about data **Maximum Contaminant Level**: whetherthe trihalomethane value collected exceeds the Maximum Contaminant Level (MCL) set by the US EPA (categories include: Greater than MCL (1), Less than or Equal to MCL (2), and Not Detected (3)) **PWS ID**: Public Water System Identification Number **CWS Name**: Community Water System Name **Population Served**: number of people served by CWS **Maximum Contaminant Level**: whether arsenic value collected exceeds the Maximum Contaminant Level (MCL) set by the US EPA (categories include: Greater than MCL, Less than or Equal to MCL, and Not Detected)

5. File name: data_113718.csv Link to footnotes:file:///Users/rachelgonsenhauser/Documents/Final_Project_Environmental_Data_Analytics/Data/Raw/Uranium/footnotes_113719.htm

`Columns` **stateFIPS**: Federal Information Processing Standard state code of state measurement was taken in **State**: state measurement was taken in **countyFIPS**: Federal Information Processing Standard county code of county measurement was taken in **County**: county measurement was taken in **Year**: year measurement was taken in **Value**: mean concentration of uranium (micrograms per liter) by year **Data Comment**: additional information about data **Maximum Contaminant Level**: whether arsenic value collected exceeds the Maximum Contaminant Level (MCL) set by the US EPA (categories include: Greater than MCL (1), Less than or Equal to MCL (2), and Not Detected (3)) **PWS ID**: Public Water System Identification Number **CWS Name**: Community Water System Name **Population Served**: number of people served by CWS **Maximum Contaminant Level**: whether uranium value collected exceeds the Maximum Contaminant Level (MCL) set by the US EPA (categories include: Greater than MCL, Less than or Equal to MCL, and Not Detected)

Metadata for **processed** datasets is provided below:

1. File name: CDC_WaterQualityAndIncome_Processed.csv

`Columns` **stateFIPS State countyFIPS County Year Mean arsenic concentration (ug/L) PWS.ID CWS.Name Population.Served Median household income PFAS conncentration (ppt) Mean trihalomethane concentration (ug/L) Mean uranium concentration (ug/L)**

**especially when I decided to go from 2013-2015 to 2014 for PFAS data, to bring to a common unit of analysis, this substantially changed the meaning of the data but my rationale is x

Scripts and code

<list any software scripts/code contained in the repository and a description of their purpose.>

list of the scripts that you've created for the projects and noting what is in these, so names of the files and a description of what each file is doing e.g. "this is where i took raw data and processed it doing x, y, z…", say things like that

Quality assurance/quality control

<describe any relevant QA/QC procedures taken with your data. Some ideas can be found here:> https://www.dataone.org/best-practices/develop-quality-assurance-and-quality-control-plan https://www.dataone.

org/best-practices/ensure-basic-quality-control https://www.dataone.org/best-practices/communicate-data-quality https://www.dataone.org/best-practices/identify-outliers https://www.dataone.org/best-practices/identify-values-are-estimated