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

Name: **Rachel Gonsenhauser** Institution: 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

stage refers 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

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

Column heading	Description of data
stateFIPS	Federal Information Processing Standard state code
State	state measurement was taken in
countyFIPS	Federal Information Processing Standard county code
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 Contaminnant Level	whether value exceeds MCL set by the US EPA
PWS ID	Public Water System Identification Number
CWS Name	Community Water System Name
Population Served	number of people served by CWS
Maximum Contaminnant Level	whether value exceeds MCL set by the US EPA

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

Column heading	Description of data
stateFIPS	Federal Information Processing Standard state code
State	state measurement was taken in
countyFIPS	Federal Information Processing Standard county code
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

Column Heading	Description of data
stateFIPS	Federal Information Processing Standard state code
State	state measurement was taken in
countyFIPS	Federal Information Processing Standard county code
County	county measurement was taken in
Year	year measurement was taken in
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

Column heading	Description of data
stateFIPS	Federal Information Processing Standard state code
State	state measurement was taken in
countyFIPS	Federal Information Processing Standard county code
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 Contaminnant Level	whether value exceeds MCL set by the US EPA
PWS ID	Public Water System Identification Number
CWS Name	Community Water System Name
Population Served	number of people served by CWS
Maximum Contaminnant Level	whether value exceeds MCL set by the US EPA

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

Column heading	Description of data
stateFIPS	Federal Information Processing Standard state code
State	state measurement was taken in
countyFIPS	Federal Information Processing Standard county code
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 Contaminnant Level	whether value exceeds MCL set by the US EPA
PWS ID	Public Water System Identification Number
CWS Name	Community Water System Name
Population Served	number of people served by CWS
Maximum Contaminnant Level	whether value exceeds MCL set by the US EPA

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

1. File name: CDC_WaterQualityAndIncome_Processed.csv

Column heading	Description of data	Class of data
stateFIPS	Federal Information Processing Standard state code	factor
State	state measurement was taken in	character
countyFIPS	Federal Information Processing Standard county code	factor
County	county measurement was taken in	character
Year	year measurement was taken in	numeric
Arsenic_ugL	mean arsenic concentration (micrograms per liter)	numeric
PWS.ID	Public Water System Identification Number	character
CWS.Name	Community Water System Name	character
Population.Served	number of people served by CWS	integer
MHI	median household income ($)	numeric
PFAS_ppt	PFAS conncentration (parts per trillion)	numeric
TTHM_ugl	mean trihalomethane concentration (micrograms per liter)	numeric
Uranium_ugL	mean uranium concentration (micrograms per liter)	numeric

Metadata note: data on PFAS (originally from `data_112028.csv`) was provided for the years "2013-2015". In order to join datasets for the processed dataset, the year category "2013-2015" was converted to 2014 to bring data to a common annual unit of analysis.

Scripts and code

Coding files for this project all exist in the R Markdown (.Rmd) format in the repository. Project coding files include the following:

File name	Purpose
Coding Session_Data Wrangling.Rmd	Wrangling data in raw datasets and joining into processed dataset
Coding Session_Data Exploration.Rmd	Exploring and wrangling data in processed dataset to prepare for analysis
Coding Session_Visualization and Analysis	Will include statistical tests and plots to analyze data relationships

Quality assurance/quality control

To perform quality assurance and qualitry control, measures will be taken during the data exploration phase to ensure that data ranges will be examined to make sure observations collected make sense given the unit of analysis. For instance, ranges for median household income, population served, and all drinking water contaminants should be above 0.

Additionally, during the data exploration phase, data will be visualized to attempt to detect extreme outliers that may be subject to removal. A "data flagging" column will be added to the dataset which will denote any observations the dataset that seem erroneous.