DSCI 510: Principles of Programming for Data Science | Project Proposal

Name: Siying Yang USC ID: 4749 4109 11

DSCI 510 Data for Final Project (Homework 4)

Highlight

- Modify the research question in the project idea for the geographical range (in red)
- Totally new on all datasets (in red)

Project Background and Objective

A much more common impact on public health in our daily includes toxic contaminants from anthropogenic sources, for example, traffic-related air pollution, which are often involuntary and constant. However, when researching such issues, it is infeasible and unethical for researchers to assign individuals to different environmental exposure conditions. Hence, to protect public health, we can combine datasets better to draw a correlation between air pollution and respiratory diseases. For example, survey data such as information on air quality; and data on asthma prevalence, hospitalizations, and emergency department visits. All these datasets are legal and ethical, and we can cut pieces of data as a substitution.

To make the project easy to be conducted, I only consider PM2.5 as an air pollutant that may cause asthma as a representative respiratory disease.

(Modify)Research Question: How is the concentration of PM2.5 associated with asthmatic attacks across the United States?

Data Sources

1. Scrap FIPS Code for different states and counties in Wikipedia (change)

website: https://en.wikipedia.org/wiki/Federal_Information_Processing_Standard_state_code

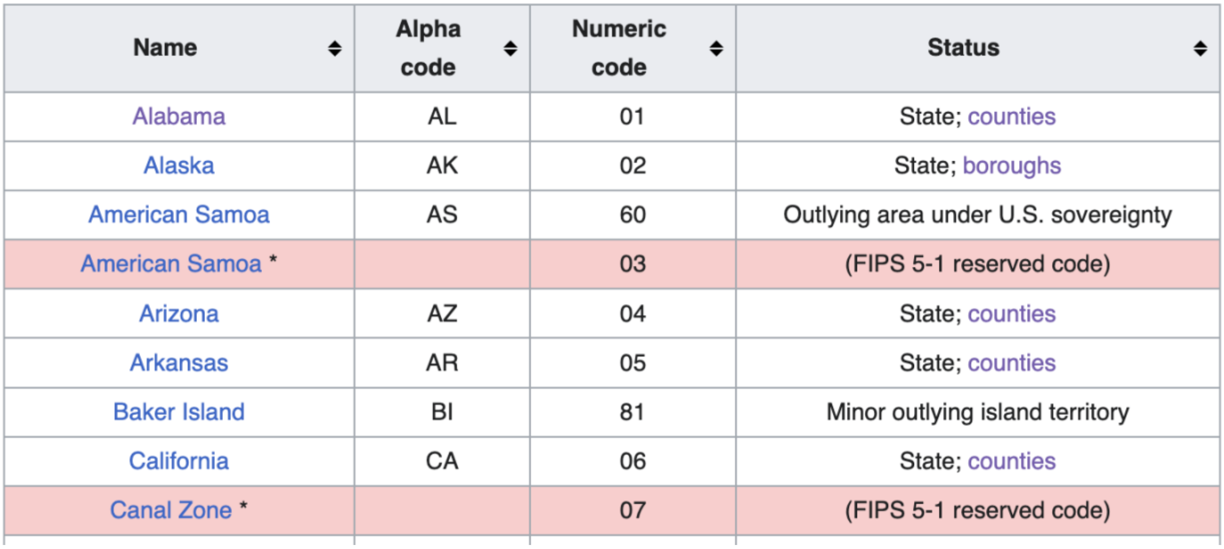
After scraping, the total number of datasets is 3093.

Data will be scraped from the FIPS state codes table for Name of the States, Alpha Code for the States, FIPS Code, and links to the state’s counties page for further scraping.

I accessed the links collected previously for the county FIPS code and scrapped the counties table for County Name and FIPS code.

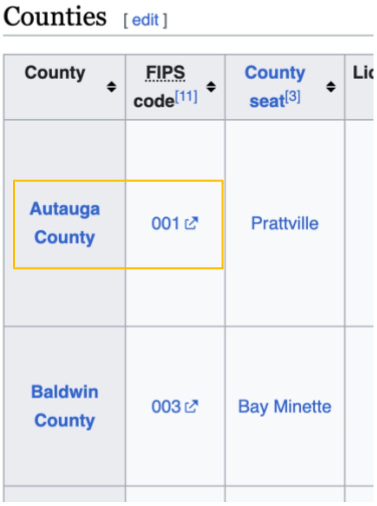
At first, I was unsure about the amount of historical PM 2.5 data available, so I scraped all the states' and counties' FIPS codes.

Then I compared the states listed in data 3 and subtracted those unavailable to reduce the workload of collecting data 2.



Name	Alpha code	Numeric code	Status
Alabama	AL	01	State; counties
Alaska	AK	02	State; boroughs
American Samoa	AS	60	Outlying area under U.S. sovereignty
American Samoa *		03	(FIPS 5-1 reserved code)
Arizona	AZ	04	State; counties
Arkansas	AR	05	State; counties
Baker Island	BI	81	Minor outlying island territory
California	CA	06	State; counties
Canal Zone *		07	(FIPS 5-1 reserved code)

Figure 1. FIPS Code Table (States)



Counties [edit]

County	FIPS code[11]	County seat[3]	Li
Autauga County	001	Prattville	
Baldwin County	003	Bay Minette	

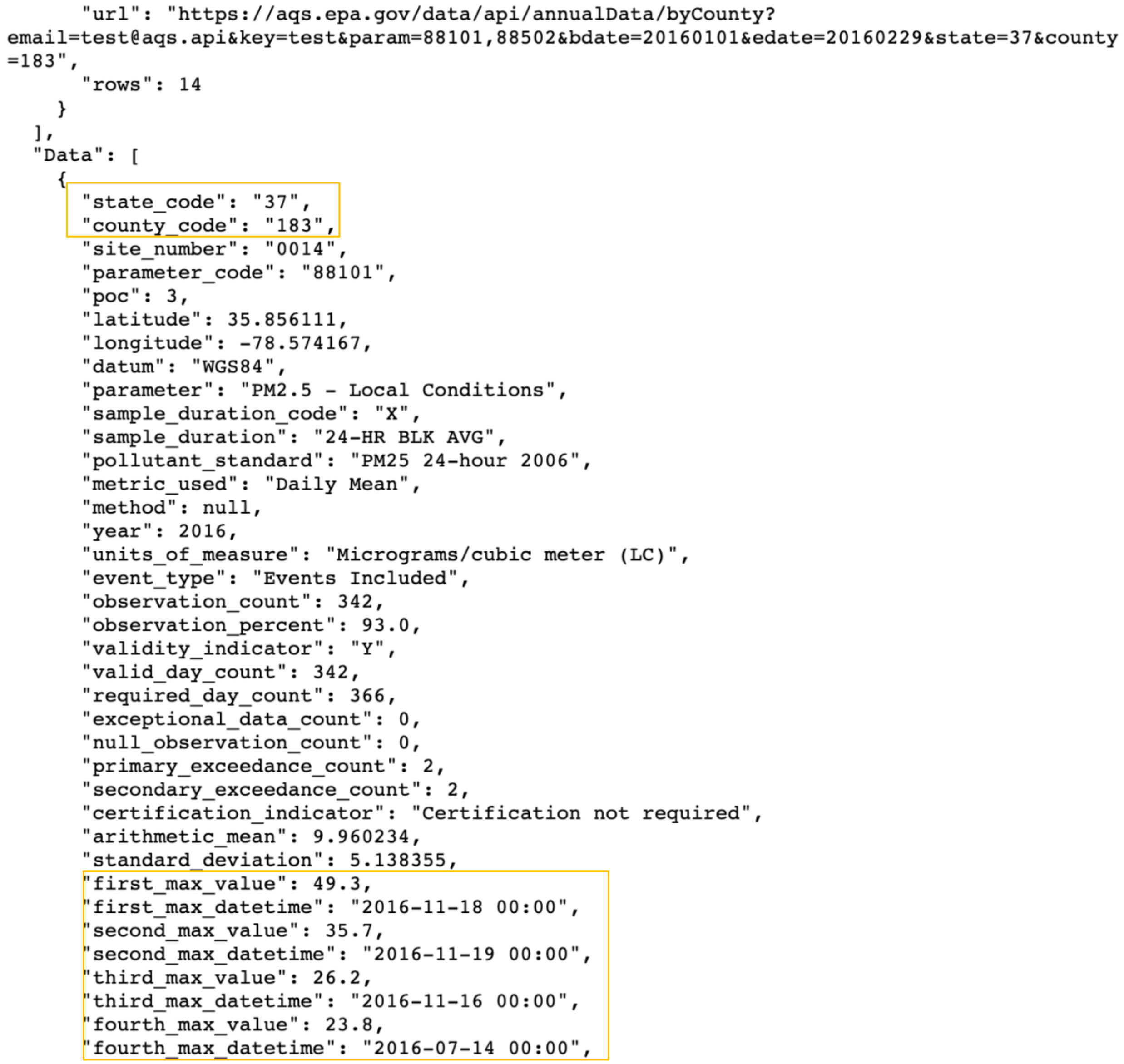
Figure 2. FIPS Code Table (County)

2. Data for air pollution: Air Quality System (AQS) API (change)

website: https://aqs.epa.gov/aqsweb/documents/data_api.html#annual

After scraping by API and pre-processing, the total number of datasets is 2364.

Based on data 1 and 3, I accessed PM 2.5 data by API. FIPS codes for states and counties in data 1 are needed, and I only accessed data from 2011 to 2019, which covered mostly in data 3. Also, the data was collected year by year. The output is four maximum values for PM 2.5.



```
      "url": "https://aqs.epa.gov/data/api/annualData/byCounty?
email=test@aqs.api&key=test&param=88101,88502&bdate=20160101&edate=20160229&state=37&county
=183",
      "rows": 14
    }
  ],
  "Data": [
    {
      "state_code": "37",
      "county_code": "183",
      "site_number": "0014",
      "parameter_code": "88101",
      "poc": 3,
      "latitude": 35.856111,
      "longitude": -78.574167,
      "datum": "WGS84",
      "parameter": "PM2.5 - Local Conditions",
      "sample_duration_code": "X",
      "sample_duration": "24-HR BLK AVG",
      "pollutant_standard": "PM25 24-hour 2006",
      "metric_used": "Daily Mean",
      "method": null,
      "year": 2016,
      "units_of_measure": "Micrograms/cubic meter (LC)",
      "event_type": "Events Included",
      "observation_count": 342,
      "observation_percent": 93.0,
      "validity_indicator": "Y",
      "valid_day_count": 342,
      "required_day_count": 366,
      "exceptional_data_count": 0,
      "null_observation_count": 0,
      "primary_exceedance_count": 2,
      "secondary_exceedance_count": 2,
      "certification_indicator": "Certification not required",
      "arithmetic_mean": 9.960234,
      "standard_deviation": 5.138355,
      "first_max_value": 49.3,
      "first_max_datetime": "2016-11-18 00:00",
      "second_max_value": 35.7,
      "second_max_datetime": "2016-11-19 00:00",
      "third_max_value": 26.2,
      "third_max_datetime": "2016-11-16 00:00",
      "fourth_max_value": 23.8,
      "fourth_max_datetime": "2016-07-14 00:00",
```

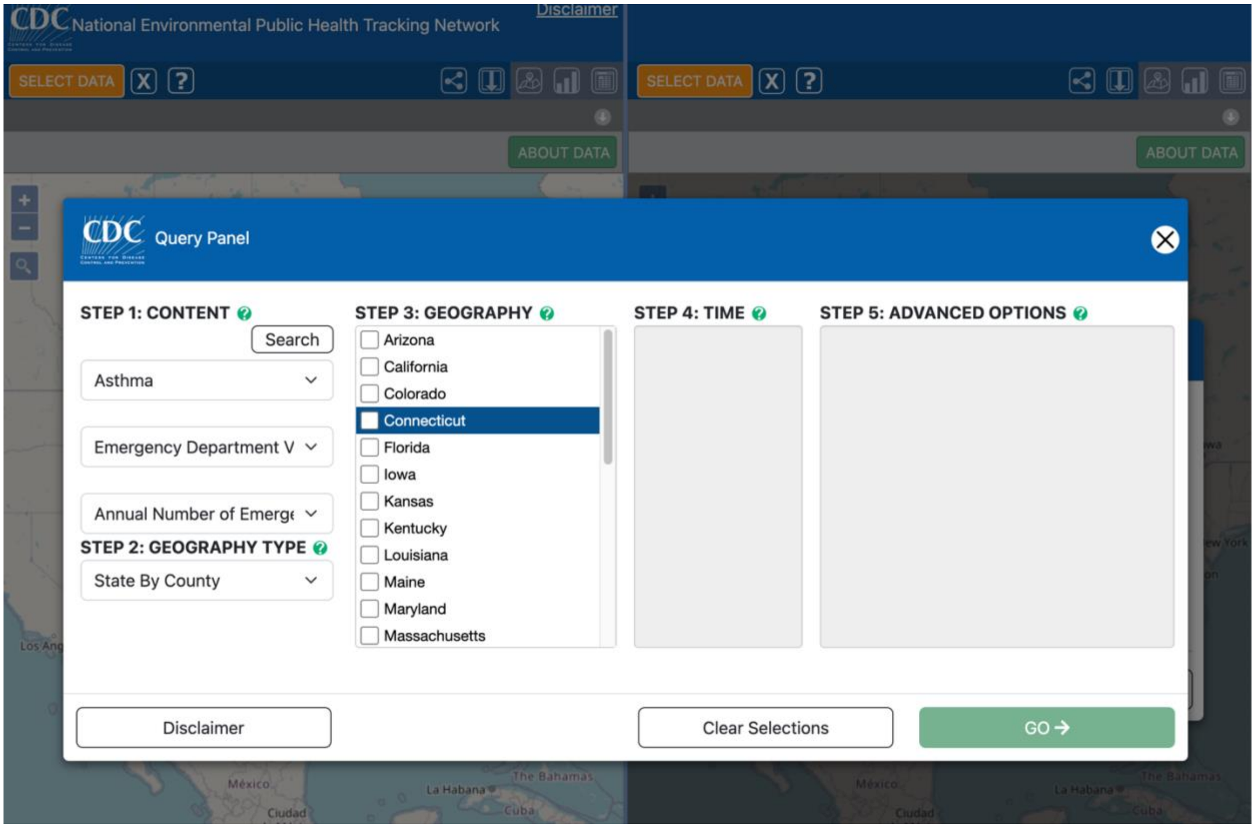
Figure 3. Sample of response

3. Data for asthma emergency department visits by counties (change)

website: https://ephtracking.cdc.gov/DataExplorer/?c=3&i=90&m=-1

After downloading and appending together, the total number of datasets is 10821 (not pre-processed yet), and the final number for analysis should equal to dataset 2.

Download dataset for asthma Emergency Department Visits. As the data is restricted data, no API access is available. I manually download all data from 2011 to 2019 for further combination.



STEP 1: CONTENT

Asthma

Emergency Department V

Annual Number of Emerge

STEP 2: GEOGRAPHY TYPE

State By County

STEP 3: GEOGRAPHY

- □ Arizona
- □ California
- □ Colorado
- □ Connecticut
- □ Florida
- □ Iowa
- □ Kansas
- □ Kentucky
- □ Louisiana
- □ Maine
- □ Maryland
- □ Massachusetts

STEP 4: TIME

STEP 5: ADVANCED OPTIONS

figure 4. Data Explorer Dashboard

Methods in Analysis

- Descriptive statistics (ideally)
  - line charts visualization
    - the trends of annual average PM2.5 in states from 2011 to 2019
    - the trends of annual ED visits across states for asthma from 2011 to 2019
  - a dynamic map visualization
    - the yearly average concentration of PM2.5 across states from 2011 to 2019
    - ED visits for asthma from 2011 to 2019 across states
  - Correlation analysis
    - First, map the concentration of PM2.5 and counts of ED visits by coordinates.

- Then, draw a correlation analysis between the concentration of PM2.5 and ED visits.

Table 1. Sample of a data file for correlation analysis

states, counties	maximum of PM2.5	counts of ED visits
01, 001	15.0131	441