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

Name: Siying Yang USC ID: 4749 4109 11

DSCI 510 Final Project Description (Homework 3)

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.

Research Question: How is the concentration of PM2.5 in urban environments associated with

asthmatic attacks in New York City?

Data Sources

1. Data for air pollution: eLichens Air Quality API

website: https://lab.elichens.com/doc/api#tag/History-API

Needed historical data example from API:

• location: latitude and longitude

• dates of a particular period

• PM 2.5 values and categories

Figure 1. Sample of response

2. Data for asthma emergency department visits by districts:

website: https://data.cccnewyork.org/data/table/6/asthma-emergency-department-visits#6/9/3/abbr/u

Data will be scraped from 2005 to 2016, considering age from 0 - 17 in case there is insufficient historical air quality data in any particular year. And the UHF districts will be a key for searching coordinates.

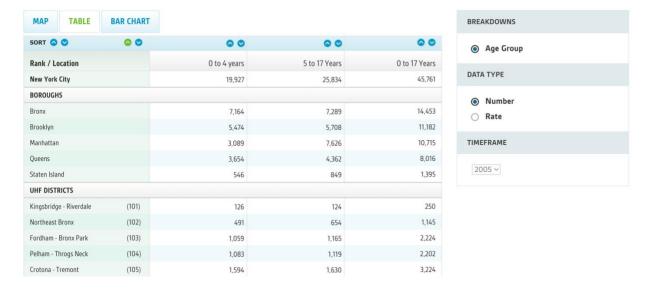


Figure 2. Asthma ED Visit Rates by districts

3. Data for coordinates and districts transformation: Geopy

website: https://github.com/geopy/geopy

The API can identify a specific district's coordinates. The geographic data in the PM2.5 dataset is recorded as coordinates, and the one in the ED visits dataset is recorded as a concrete district so that this API can bridge the above two datasets.

BMW of Bayside (Douglaston Campus), 247–21, Northern Boulevard, Dougla (40.7668833, -73.7398197)

figure 3. a sample for UHF districts to coordinate

Methods in Analysis

- Descriptive statistics (ideally)
 - line charts visualization
 - o the trends of annual average PM2.5 in New York City from 2005 to 2016
 - the trends of annual ED visits in New York City for asthma from 2005 to
 2016
 - a dynamic map visualization
 - the yearly average concentration of PM2.5 across New York City from 2005 to 2016
 - o ED visits for asthma from 2005 to 2016 across New York City
 - Correlation analysis

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

- o 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

Coordinate	concentration of PM2.5	counts of ED visits
(40.7668833, -73.7398197)	15.0131	441