Progress report

Rethinking PM2.5 Exposure: Chronic Disease Trends in the U.S. (2015 – 2019)

Project scope update

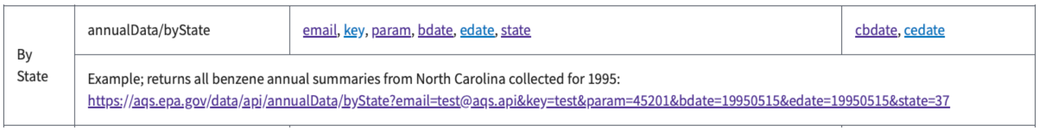
Since statrting the project, I made a small adjustment to the scope. At first, I planned to analyze every U.S. state, but the PM2.5 daily data from the EPA API was extremely large and slow to retrieve. So, I decided to focus on five major states: California, Colorado, Illinois, New York, and Texas. These states represent different geographic regions and still work well for trend comparison.

The global PM2.5 dataset remains part of the project, as planned, to give additional context when discussing environmental patterns.

Data sources

1. **EPA: Air quality system (AQS) API**

- Signed up to get email and API key
- Checked metadata and parameter codes; used "88101" (PM2.5 – local conditions, mass/qa). This can be obtained in the list "parameters classes" and "parameters in a class" services in the website.
- Retrieved annual summary data by state for 2015–2019. API allows only one year per request; state codes are 2-digit.



By State	annualData/byState	email, key, param, bdate, edate, state	cbdate, cedate
	Example; returns all benzene annual summaries from North Carolina collected for 1995: https://aqs.epa.gov/data/api/annualData/byState?email=test@aqs.api&key=test&param=45201&bdate=19950515&edate=19950515&state=37		

- Selected five states (California, Colorado, Illinois, New York, Texas) for geographic diversity and smaller data size. The specific 2-digit code for each state that can obtained via the list states service.
- Extracted arithmetic mean PM2.5 at the county level.
- Resulting data: 14,423 rows (.json).

2. **Data.gov: U.S. Chronic Disease Indicators**

- Retrieve via URL, and filtered for the same five states and 2015–2019.
- Keep only essential variables: disease name, value, value unit, state, year, and geolocation.
- Resulting data: 8,782 rows (.json).

3. **WHO: Air pollution: concentrations of fine particular matter (PM2.5), SDG 11.6.2**

- Downloaded locally and uploaded to Google Drive for access.
- Filtered for 2015–2019, keeping indicator, country, year, and value.
- Resulting data: 4,725 rows (.csv).

Issues / difficulties

- EPA API data is extremely large even in annual summary form (thousands of rows/state), so the analysis is limited to five states.
- WHO data cannot be downloaded directly via URL. It generates link as "blob:null/2cbc4eda-cd3c-4dc0-99d1-ac02442d40f5". So, I downloaded locally, uploaded to Google drive, and retrieve from Google drive link instead.
- Aligning chronic disease and PM2.5 datasets by state and year is challenging, especially since each dataset has different formats and levels of detail.