Data Acquisition, Extraction, and Storage
Project Global Air Pollution with Economics Data

Zhe HUANG, Linghao Zeng, Vivien Conti

January 9, 2024

1 Introduction

The primary objective of this project is to develop an expansive dataset that meticulously chronicles global air pollution metrics over a three-year period, spanning from January 1, 2021, to December 31, 2023. This dataset is poised to serve as a critical resource for a variety of applications, including but not limited to, in-depth environmental research, innovative data visualization tools, and the development of predictive machine learning models aimed at forecasting future pollution trends. Through this endeavor, we aim to provide a versatile dataset that can significantly contribute to the understanding and analysis of air quality on a global scale.

For complete access to all project resources, including datasets, code, and documentation, please visit our shared Google Drive folder. The folder can be accessed through the following link: Project Resources on Google Drive.

2 Data Sources

In this section, we will introduce the data sources we used and the technical details of how the data were collected.

2.1 World Air Quality Index Project

World Air Quality Index Project (https://waqi.info/) provides real-time air quality data from thousands of monitoring stations worldwide. The website offers downloadable historical data in CSV format. However, it restricts bulk downloads of global data and requires verification for each download. Additionally, the provided API does not support retrieving historical data. To overcome these limitations, we adopted the following methodology. The complete process and code used for data extraction, decryption, and analysis are documented in a Jupyter Notebook file named 'waqi.ipynb'.

2.1.1 Website Analysis (Packet Sniffing)

Using Chrome's Developer Tools, we captured response packets by refreshing and monitoring browser requests. This process helped us locate packets corresponding to historical data and URLs for different regions.

1. **Packet for Historical Data (GET Request)**:

   Example URLs:

   - Chongqing Ankang: https://api.waqi.info/api/attsse/9239/yd.json
   - Guiyang Ma'anshan: https://api.waqi.info/api/attsse/1368/yd.json

   Upon comparison, we noticed that the 'idx' number in the URL varies. Accessing these URLs in Chrome or Postman revealed that the data is structured on a monthly basis in JSON format, with each 'data' entry corresponding to a specific month. However, the contained data is encoded (encrypted).

2. **Packet for Global Region Indices (POST Request)**:

   URL: https://api.waqi.info/mapq2/bounds

   Similar as before, the 'data' key in the packet in the JSON packet corresponds to the 'idx' of different regions.

3. **Packet for Decryption Function**:

   URL: https://aqicn.org/webapp/dist/historic-module-dyn.2b2626b6ef49374f9dcd.js

   Through breakpoint debugging of the website, we observed that encrypted data is parsed into daily values for various indicators after being processed through functions contained in this package.

2.1.2 Data Scraping, Decryption, Cleaning, and Writing to CSV Files

We retrieved all region indices ('idx') using POST requests via the requests library. By looping through the indices, we injected them into the URL. Using the GET method, we obtained raw data in JSON format for each region. We then located the decryption function in the JS file, adapted it into Python, and used it to decode the raw data. The decoded data, a nested dictionary list, was processed to extract relevant information. We combined the data in a time-ordered sequence [time, pm2.5, pm10, O3, NO2, SO2, CO].

The city names were split into country/region formats, and through regular expressions, characters unsuitable for system filenames were replaced. We used the OS library to create corresponding directory paths (e.g., data/FR/Paris/station_name.csv). Finally, the data was written to CSV files located in the respective paths using pandas.

2.2 OpenAQ Platform

OpenAQ (https://openaq.org/) is a nonprofit organization in the environmental technology sector. It collects and standardizes open air quality data from stations around the world onto a freely available open-source data platform. This ensures that individuals concerned about air quality have unrestricted access to the data necessary for analysis. By enabling universal access to air quality data, OpenAQ empowers a global community of change-makers to address air inequality—the unequal access to clean air. As far we know, OpenAQ was the first to consolidate diverse ground-level ambient air quality data on a centralized open-source platform and remains the world's largest open-source air quality data platform.

2.2.1 Get the data from the platform

Our initial step was to collect data from various stations. The OpenAQ platform provides an API for data retrieval. Through a GET request, we were able to extract and store the following information for each station: station_id, country, city, first_date_of_measure. Now, we aimed to extract all measurements made by each station since their first measurement date. However, with a very high number of stations (48000) and the API allowing only 300 requests per 5-minute interval, it was impossible to extract our data within a reasonable time frame. Therefore, we chose to reduce the data extraction by selecting only one station per city. Unfortunately, even with this reduction in data, the API retrieval time was still too long. Ultimately, we discovered that OpenAQ data can be directly accessed as gzipped CSV files from the Open Data on AWS Program's S3 bucket (the bucket root url is https://openaq-data-archive.s3.amazonaws.com/). To use this resource, we employed and configured AWS CLI to send requests to the S3 bucket. Iterating over the station IDs we selected earlier, we retrieved all measurements from the S3 bucket for each station since 2021. The downloaded files were directly saved in a data folder with the following structure: year/month/station_id_day_of_measure.csv. Then, for a given month, the number of files equaled (the number of active stations for that month multiplied by the number of days in the month). Each CSV file contains the following information: location_id (equivalent to the station's ID), sensor_id (sensor ID used for measurement), location, date time, latitude, longitude, parameter, units, and value. Indeed, stations measure multiple parameters, so each measurement is characterized by a date time and a parameter.

2.2.2 Structuring the data

As mentioned earlier, we found it sensible to organize the data following this schema: Country/City/station_name.csv. To achieve this, we created a dictionary and iterated through all the files in the year and month folders to concatenate all the dataframes from CSV files of the same station. Thus, the final dictionary had station IDs as keys, and for each of these IDs, the associated value was the concatenation of all dataframes from that ID. For each station, we could retrieve the country and city names in which it is located based on its ID. Finally, we were able to save all the files with the desired structure (Country/City/station_name.csv).

2.3 World Bank Open Data

The World Bank Open Data is an invaluable resource for global data, offering a wide range of socio-economic indicators. This platform provides free and open access to global development data, including detailed information on various aspects like economic growth, population, and environmental conditions.

2.3.1 Auxiliary Data Selection

For our study, we focused on acquiring auxiliary data related to air pollution from the World Bank Open Data. Given the vast array of datasets available, we selectively extracted six specific CSV files, which were deemed most relevant to our study. These files include critical indicators such as GDP, Forest Area, Industry Value Added, Urban

Population, and others, chosen after a thorough review of all the available data on the site. While the World Bank Open Data offers an extensive collection of files, these selected datasets provide the most direct relevance to our study of air pollution.

2.3.2 Data Processing and Refinement

The raw data obtained was quite extensive and complex, containing a lot of unnecessary information. For instance, certain keys present in the raw CSV files were not pertinent to our study. To refine this data into a usable format, we employed a combination of the Pandas library for data manipulation and manual editing.

Our primary focus was on the data corresponding to the years 2021 and onwards, aligning with our air pollution data spanning from 2021 to 2023. However, it is important to note that the data for 2023 was not fully available at the time of our research. Typically, annual data like this is released in the middle of the year or in spring, which means our auxiliary data for air pollution is currently limited to the years 2021 and 2022.

3 Data Integration Methodology

3.1 Standardization of Country Codes

In the process of merging our data and constructing the comprehensive database, we prioritized the establishment of a standardized key across all sources' data. Our chosen key was the alpha-2 country code, commonly recognized and used as a global standard for identifying nations.

For example, the data from the World Bank Open Data initially utilized the alpha-3 country code system. To maintain uniformity, we implemented a mapping system when inputting this data into our database, converting alpha-3 codes to their alpha-2 counterparts.

3.2 Streamlining Air Pollution Data

In dealing with the air pollution data sets, it was imperative to curate the information meticulously to uphold coherence across the various tables within our database. During the data entry phase, we strategically omitted extraneous details that did not contribute to the overall narrative of our analysis. For instance, while the original records included precise latitude and longitude coordinates, we decided to retain only country and city information.

3.3 Creation of a Country Attribute Table

Furthermore, we sought to enrich our database by categorizing countries based on specific attributes. Income level, sourced from the World Bank Open Data, was utilized as a distinguishing attribute. To this end, we created a separate country attribute table within our database. This table serves as a repository for such attributes, enabling us to conduct more nuanced analyses that consider the socio-economic backdrop of each nation.

4 Database Structure

According to the methodology introduced before, we created the database containing the following tables:

4.1 Countries Table

- Serves as the foundation of our database.
- Contains 'country_code' and 'country_name'.
- 'country_code' is a unique identifier and acts as the primary key.

4.2 AirQuality Table

- Dedicated to storing air quality records.
- Fields include 'id', 'country_code', 'city', 'datetime', 'parameter', 'units', and 'value'.
- 'id' is an auto-incrementing primary key.
- Links to 'Countries' table via 'country_code'.

4.3 EconomicIndicators Table

- Captures economic metrics for each country.
- Consists of ‘country_code‘, ‘indicator_name‘, ‘year‘, and ‘value‘.
- Uses a composite primary key of ‘country_code‘, ‘indicator_name‘, and ‘year‘.
- Linked to ‘Countries‘ table.

4.4 CountryAttributes Table

- Stores additional country attributes.
- Includes ‘country_code‘, ‘attribute_name‘, and ‘value‘.
- ‘country_code‘ is the primary key and links to ‘Countries‘ table.

5 Results and Use cases

In our comprehensive database, we have successfully aggregated a substantial volume of global air pollution metrics. This dataset encompasses detailed air quality records from 1,278 cities (regions) across 57 countries. The collection comprises a total of 41,320,964 individual air quality data entries and 2,660 records pertaining to various economic indicators. The total size of this database stands at 2.65 GB.

Here are some use cases of this database:

1. Calculates the average PM2.5 levels in a specific country (e.g., "US") for a given time period (the year 2021):

```
SELECT AVG(value) as average_pm25
FROM AirQuality
WHERE country_code = 'US'
AND parameter = 'pm25'
AND datetime BETWEEN '2021-01-01' AND '2021-12-31';
```

Result:

```
    average_pm25
0       8.781307
```

8.781307

2. Identifiy the country with the highest average PM10 levels:

```
SELECT country_code, AVG(value) as average_pm10
FROM AirQuality
WHERE parameter = 'pm10'
GROUP BY country_code
ORDER BY average_pm10 DESC
LIMIT 1;
```

Result:

```
    country_code average_pm10
0             AE    81.220956
```

3. Compares the GDP of countries with their PM2.5 pollution levels:

```
SELECT a.country_code, AVG(a.value) as average_pm25, e.value as gdp
FROM AirQuality a
INNER JOIN EconomicIndicators e ON a.country_code = e.country_code
WHERE a.parameter = 'pm25'
AND e.indicator_name = 'GDP (current US$)'
AND e.year = 2021
GROUP BY a.country_code, e.value;
```

Result (first 3 lines):

	country_code	average_pm25	gdp
0	AE	-89.357486	4.151788e+11
1	AT	9.329183	4.792954e+11
2	AU	-29.390119	1.559034e+12

6 Conclusion

In this project, we successfully compiled a comprehensive database of global air pollution and economic data, utilizing advanced data acquisition techniques like website analysis, packet sniffing, data scraping, and decryption. Our approach streamlined the integration of diverse data sets into a cohesive structure, encompassing over 41 million air quality records and economic indicators from multiple countries.

The techniques we employed allowed us to overcome significant data accessibility challenges, enabling us to standardize and simplify complex data for more efficient analysis. The resulting database is not only a testament to our technical proficiency in handling large-scale data but also serves as a valuable resource for future environmental and economic studies.

With this database, we have laid a foundation for deeper insights into the relationship between air quality and economic factors, opening avenues for informed research and policy-making aimed at addressing global air pollution challenges.