Data-Report: The impact of air pollution on cancer in the U.S.

Question

How does air pollution (measured by the AQI) in the five most populous states in the US correlate with the incidence of cancer?

Data Sources

Data Source 1: Annual Summary of AQI by County (of the US)

The source of this data is a website that contains links to a dataset for each year (from 1980 to 2024). Each dataset provides information about air quality for each county in each state of the U.S. for a specific year. Air quality is measured using the AQI, or Air Quality Index. This index provides an easy way to assess air quality, ranging from 0 to 500, with the following categories:

0-50	good
51-100	moderate
101-150	unhealthy for sensitive groups
151-200	unhealthy
201-300	very unhealthy
301-500	hazardous

The AQI is based on five major air pollutants: ground-level ozone, particle pollution (particulate matter (Feinstaub) including PM2.5 and PM10, which refer to dust particles smaller than 2.5 µm and 10 µm, respectively), carbon monoxide, sulfur dioxide, and nitrogen dioxide. Sensitive groups include individuals with lung or heart diseases, children, and the elderly.

The AQI is measured throughout the year, and each dataset contains the number of days in which air quality was classified as good, moderate, …, hazardous for each county. However, the dataset doesn't include the AQI for every single day, which is why it also specifies the number of days the AQI was actually measured throughout the year. Additionally, the dataset includes the maximum AQI recorded for the year and some more information about air quality and pollutants.

I chose these datasets because the AQI provides an understandable representation of air quality. The datasets are organized by county and state, offering a granular level of detail that is sufficient for analyzing correlations. The website where the data is sourced also includes datasets on a daily or even hourly basis, but I opted for the annual datasets because this level of precision is adequate for correlation analysis. Furthermore, I didn't find any datasets on diseases that were recorded on a daily basis, so using more granular air quality data would not be beneficial. I decided to take the datasets from 2006 to 2021 because Data Source 2 only includes data for these years.

The datasets are provided by the United States Environmental Protection Agency (EPA). They are available under the U.S. Public Domain license, as no other licencing information is specified (EPA data license, website of the datasets, description of the data and formats). Therefore, I am allowed to use this data as long as I do not use any associated photos (e.g. in an advertisement) or federal government trademarks/logos.

In terms of data quality, it is beneficial that the data is separated by year and county for each state in the U.S. It is also advantageous that the total number of days the AQI was measured is specified. Another positive aspect is that the data is very recent, with some data already available for the first half of 2024. The data is consistently structured, following a fixed schema in a common format (CSV), which makes it easy to process. However, the data isn't entirely complete in every aspect. While it includes data for every county, it doesn't always include the AQI for every day. In some cases, the AQI is only recorded for approximately 100 days in a year. I can't comment on the correctness of the data.

Data Source 2: Incidence of Cancer by States in the U.S. per year

This dataset can be downloaded by filling out a form. Another way to automatically download the data is by making a POST request to `https://wonder.cdc.gov/controller/datarequest/[database ID]` with the correct database ID (in this case, D198) and including the parameter `request_xml` with the value of the XML file that can be downloaded after filling out the form.

This data source provides information about cancer rates for each state. For each state, data is available for all the years from 1999 to 2021. The dataset includes the number of people newly diagnosed with cancer in each state for each year (i.e., it represents incidence, not prevalence, which counts how many people currently have cancer). Additionally, the dataset includes the population and the computed crude rate (calculated as `count / population`). A summary for each state is also provided, which aggregates data across all years and includes the total cancer incidence, the total population, and the overall crude rate. The dataset is limited to the five most populous states of the U.S. (California, Florida, New York, Pennsylvania, and Texas) because I chose to focus on these states for my project and specified them when completing the mentioned form. Some metadata is also included within the dataset.

I chose this dataset because it focuses on cancer incidence rather than prevalence. This distinction is important because prevalence also depends on factors such as mortality and recovery rates, which could complicate the analysis. Additionally, the dataset provides data at the state and yearly levels, offering the necessary granularity. A dataset showing only cancer incidence summed up for the last 10 years or aggregating data across all states would not allow a proper analysis of correlations. Moreover, I opted to include data only from 2006 to 2021, excluding 1999-2005, due to issues with data of 2005 caused by Hurricanes Katrina and Rita (a lot of individuals were displaced). These limitations were described in the downloaded dataset.

The dataset is provided by the Centers for Disease Control and Prevention (CDC), which sources the data from selected statewide and metropolitan area cancer registries that meet the data quality criteria. The incidence data are based on diagnoses and populations reported by these registries.

The dataset is available under a specific policy outlined on the CDC WONDER website, which applies to all CDC WONDER datasets. To download the data, the `accept_datause_restrictions` parameter in the POST-request must be set to `true`. This policy specifies that the data is only be used for the purpose of statistical reporting and analysis and prohibits linking it with other datasets or similar attempting to identify an individual. Since I am only using the datasets for my analysis and not violating these terms, I fulfil their obligations. The CDC also provides a suggested citation, which is included at the end of this report.

Regarding data quality, the dataset is consistently structured, following a fixed schema in a common format (XML). There are no missing years or incidences, and the data is relatively recent (only the last two years are missing). Using the mentioned form, many parameters can be adjusted to get the dataset in the desired format. Since the dataset includes the cancer incidence data for each state, it offers sufficient granularity and relevant for answering my question.

Data Pipeline

The data pipeline was implemented in *Python* using *pandas* for data processing and *retry* for error handling. To run the pipeline, both libraries must be installed.

First, the air quality datasets for each year are downloaded and processed. Each dataset is compressed in a zip file, which has to be unzipped. The data from each file is then processed individually and saved to the same table (named `aqi`) in the SQLite database. Since not all states have AQI data for the same number of days, I compute the percentage of good, moderate, … hazardous days for each state and year. This is calculated as `([good, moderate, …, hazardous] days) / (total days with AQI)`. The data is

initially separated by county, and no state-level summaries are provided. To aggregate the data by state, I compute the mean percentages across all counties in each state. For the maximum AQI, I take the highest value recorded across all counties and also calculate the median of all maxima, to avoid the effect of outliers.
Once the air quality is downloaded, processed, and saved, I download the cancer incidence dataset by making a POST-request, as described in the explanation of Data Source 2. The response is an XML file, which requires processing before transforming it with *pandas*. When downloading the dataset via the form, I received a .txt file, but the XML file requires understanding the schema, particularly handling nodes that do not contain the data I need (e.g., the nodes summarizing state data across all years), as these nodes have a different structure and child elements. Additionally, I have to ensure that numbers are correctly formatted.
After processing the XML file (by extracting only the nodes containing the required data and ignoring unnecessary data), I transform the data using *pandas*. I select only the years I need and drop unnecessary columns (Count and Population), focusing instead on the rate per 100.000 people for easier comparisons). Finally, this processed data is saved to the same database but in a separate table (named cancer_rates).
For error handling, I implemented retry logic. If a download fails, it is retried up to three times with a delay of 2 seconds between the attempts. If the download still fails after three attempts or if further processing encounters an error that prevents the data from being processed by the pipeline, one of the following actions are taken:

- Cancer rates data: If an error occurs while downloading/processing this dataset, the pipeline stops, as no analysis can be done without this data.
- AQI datasets: If an error occurs while downloading/processing some of the AQI datasets, the pipeline logs the error and continues. But if fewer than ten AQI datasets are successfully downloaded and processed, the pipeline is stopped as well, as insufficient data would be available.

Result and Limitations

The output data table (SQLite) contains two tables. I chose this format because it is structured and enables further processing, such as generating graphs.
The first table holds data about air quality in the five most populous states of the U.S. for each year from 2006-2021. It includes the mean percentages of the good / moderate / … / hazardous days for each county within the state. Additionally, it contains the maximum AQI and the median AQI values of the counties in each state. A limitation of this aggregated data is that counties vary in area size and populations. This makes it difficult to compute a mean that takes both differences into account (e.g., by weighting counties based on size or population). Consequently, the computed state-level values may not fully reflect these differences between counties and therefore might not bee fully accurate.
The second table contains cancer rate data (per 100.000 people) for the same five states and years. However, the dataset is not entirely accurate because not every cancer case is diagnosed or included in the original dataset. But it probably is representative for analysis purposes.
This project also aimed to analyse the correlation between air pollution and asthma rates. However, after several hours of searching, no suitable dataset could be found. The available datasets had limitations: they aggregated data across years, required creating accounts for access, didn't have state-level data, were a PDF, or were presented on a website without the possibility to download the data. That's why this project only focuses on cancer.

References:

- https://www.airnow.gov/aqi/aqi-basics/
- Data Source 2: United States Cancer Statistics - Incidence: 1999 - 2021, WONDER Online Database. United States Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; 2023 submission; 2024 release. Accessed at http://wonder.cdc.gov/cancer-v2021.html