Analyzing the impact of air pollution on cancer in the U.S.

Introduction

Air pollution is a growing global concern, driven by industrial activities, increasing numbers of vehicles on the road, and other factors. Not only potentially harmful gases but also particulate matter are released into the atmosphere. When inhaled, these particulates can enter the lungs. Especially sensitive groups, including children, the elderly, and individuals with existing respiratory conditions, are vulnerable to poor air quality.

One disease that is often difficult to treat and whose causes are frequently unclear is cancer. Due to the overall unclear effects of air pollution, this project analyzes whether there is a correlation between air pollution and the incidence of cancer. It focuses on the five most populous states in the United States, because air pollution especially affects areas with high population, where industrial activities and the number of vehicles is usually high. By analyzing air quality data (measured using the Air Quality Index, or AQI) and cancer incidence rates, this study aims to find out whether there are patterns or relationships. Specifically, it tries to answer the question: How does air pollution, measured by the AQI, in the five most populous states in the U.S. correlate with the incidence of cancer?

Used Data

Several datasets were utilized to generate the data required for the analysis. For each analyzed year, a dataset containing information about the air pollution in the counties of the U.S. states was used. Each dataset included the number of days in the corresponding year classified into each AQI (Air Quality Index) category. The AQI is a number ranging from 0 to 500 and can be divided into 6 categories: good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy and hazardous. These datasets are available under the U.S. Public Domain license, as no other licencing information is specified. Therefore, I am allowed to use this data if I do not use any associated photos (e.g. in an advertisement) or federal government trademarks/logos.

To address the research question, data on cancer rates was also required. Various parameters could be configured for the chosen dataset and I decided to take the data with a granularity of years and states. The cancer rate is expressed per 100.000 people, indicating the number of new cancer cases recorded during the year. This dataset is available under a specific policy that applies to all CDC WONDER datasets. To download the data, I had to accept the data use restrictions (set the parameter correctly). This policy specifies that the data may 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.

The final dataset used for the analysis is an SQLite database including two tables: one for the AQI data and another for cancer rates. Both tables cover the years 2006 to 2021 and include information only for the five most populous states of the U.S. (California, Florida, New York, Pennsylvania, Texas), as these states probably are the most representative, especially regarding the cancer rates. The *cancer_rates* table contains the crude rate (cancer incidents per 100.000 people) for each year and state. The *aqi* table contains the mean percentage of days of the six AQI categories for each year and state. This data was computed from the original datasets by averaging the data for all the counties of each state. Additionally, the maximum AQI and the median of the maximum AQIs across counties are included.

Data Quality

The data is incomplete due to incompleteness in the original data. The AQI was not measured every day of the year in all the counties. As a result, the percentage for each AQI category was calculated based on the total number of days the AQI was measured (in some cases only 100 days of the year). When computing the mean AQI percentages across the counties of a state, all counties were weighted equally, independently of their area size or population. It is unclear whether this could have influenced

the analysis or if weighting the counties differently would have been more appropriate. However, this is a relatively small issue compared to the advantageous fact that AQI data and cancer rates are available for all selected states and years.

The AQI is a useful way to measure the air quality and air pollution, while the cancer rate per 100.000 is an appropriate way to measure the relative number of cases. That is why the data is highly relevant for answering the question, whether there is a correlation between air pollution and cancer rates. The question limits the analysis to the five most populous states, exactly like the final dataset, which is in a consistent structured format (SQLite).

Both tables provide annual data, which offers a sufficient granularity for this analysis. During the analysis I observed a drop in the cancer rate for each state in 2020. I explain this with the COVID-19 pandemic, as fewer people were allowed to visit the hospital and therefore the cancer could not get diagnosed properly. While the data of the year 2020 is not entirely accurate, the consistent drop across all the states (the pandemic was everywhere) makes the effect on the analysis small, especially because not every cancer is diagnosed no matter if there is a pandemic or not. So the cancer rate will never be entirely accurate but likely is representative.

Analysis

To visualize the cancer rates of the five states over the years, a line chart was used, to better analyze changes over time in comparison to changes in air quality. For visualizing the air quality, a stacked bar chart was chosen and integrated into the same diagram as the cancer rate. Each year is represented by a single bar, which represents 100% of the days on which the AQI was measured (averaged over the counties). The different colours within the bar represent the percentage of days that the AQI was in the corresponding category (e.g. green for good, yellow for moderate, etc, see the full legend in diagram 6). Separate diagrams were created for each state (diagrams 1-5), covering all analyzed years (x-axis). This allows for a clearer visualization of the relationship between air quality (left y-axis) and cancer rates (right y-axis). The cancer rate y-axis uses the same range across all diagrams so that it is easier to compare the differences between the states. The diagrams are sorted from left to right in ascending order of the mean cancer rate.

As shown in the diagrams, air quality became a little better over the years in Texas, New York and Florida. The air quality of California and Pennsylvania is fluctuating. The cancer rates are slightly rising, except for the drop in 2020. It was already mentioned in the section *Used Data* that this was the beginning of the COVID-19 pandemic. A potential explanation may be the shutdown of hospitals over months, so that the cancer couldn't be diagnosed, even though the actual incidence of cancer did not change significantly.

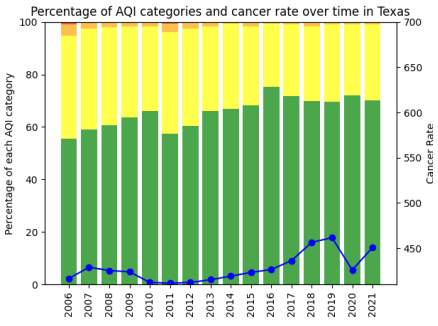

Diagram 1

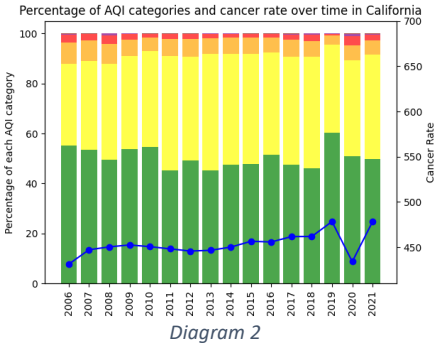

Diagram 2

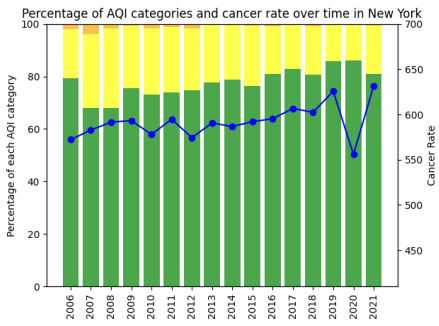

Diagram 3

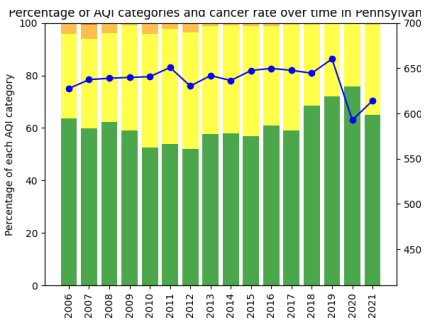

Diagram 4

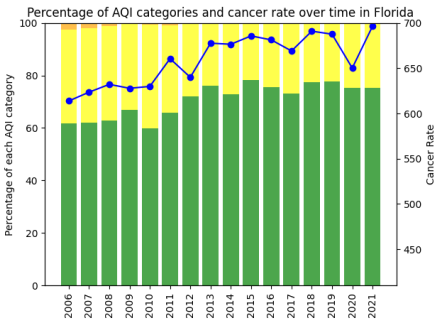

Diagram 5

When examining each state independently over time, no correlation is visible between air quality and the crude cancer rate. For example, even though the air quality of California is the poorest, it has a lower cancer rate than for example Florida or New York. To better compare the slight differences between the states – and because analyzing changes over time provided limited insights – I generated another diagram, that does not have a time axis. It shows the mean crude cancer rate and the average percentages of the AQI categories for each state (diagram 6). In this diagram it is easier to see, that Texas and California have by far the lowest cancer rates, even though California has the worst air quality. Florida and Texas don't have a big difference in the air quality, nevertheless, Texas has the lowest and Florida the highest crude rate. Thus, there is neither a positive nor a negative correlation between air quality and cancer rates. The states are ordered from left to right based on their crude rate, but the AQI data fluctuates without a visible pattern.

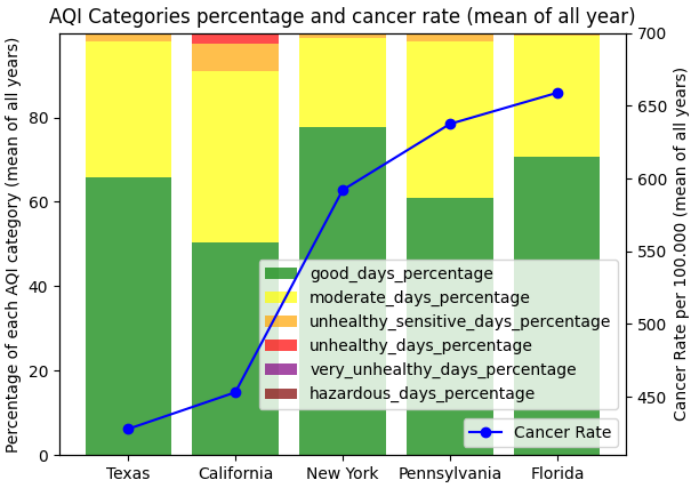

Diagram 6

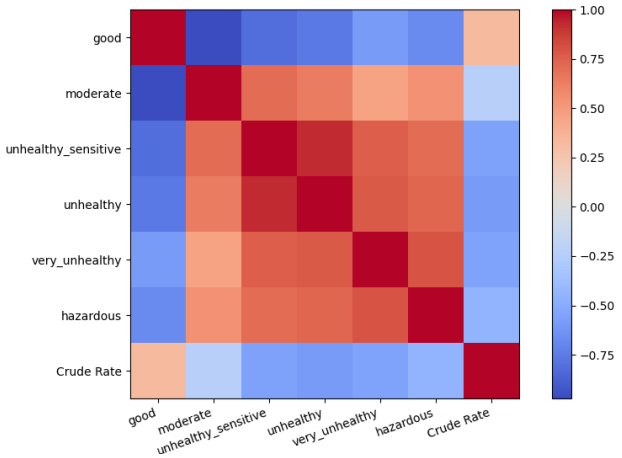

Diagram 7

To further examine the absence of a visible correlation, a heat map of the correlation matrix (computed with the Spearman correlation coefficient) was created. The Spearman correlation coefficient matrix shows the correlation coefficient between pairs of variables (shown on the x-axis and y-axis). A value close to 1 or -1 indicates a strong positive or negative correlation, while a value close to zero suggests no correlation. This is a more reliable analysis of correlations than visual inspection. The heatmap shown in diagram 7 was generated using all data points from diagrams 1-5. The most relevant values are the last row or last column (as the other columns/rows show the correlation between the AQI categories). These show the correlation between the crude cancer rate and the AQI categories. The light colours indicate that if a correlation exists, it is not a strong one.

Interpretation

That no correlation could be detected is probably related to the fact that a poor air quality is not good for the overall health, but especially bad for the lung. The cancer rate however, includes all forms of cancer, where the cause is often unclear. At least lung cancer might correlate with poor, polluted air but the used data lacks the granularity to admit this. Another factor is the unknown number of hours or days, which people diagnosed with cancer, were exposed to outdoor versus indoor air compared to those who did not develop cancer. If there is a correlation between lung cancer and poor air quality that would be a factor which could affect the datasets. The website, that provides the data about the cancer rates, offers the option to decide on the cancer types. But for this analysis, data on all cancer types was used.

The absence of a correlation in the diagrams, showing changes over time, might also be related to the fact that air quality did not change significantly during the analyzed time span. So even if a correlation was present, it would likely lack significance. Additionally, the cancer rate is continuously increasing in each state independently of the air quality. This could be explained by the aging population, as age is a risk factor for cancer. Cancer typically does not develop overnight, so any correlation between cancer

and the air quality is probably difficult to see in the changes over time as it would be time-shifted (how much is unclear, probably years), especially if the change of air quality is minimal.
Also, in the overall comparison of the states without the time factor and only looking at the averaged air quality and cancer rates was no correlation detected. Probably because the values also are relatively similar to each other. The air qualities only differ slightly compared to differences between countries or continents. However, analyzing air quality and cancer rates across countries would introduce additional variables, such as cultural differences (e.g. different diet, habits, environment, …) which all could influence the cancer rates. Although diagram 6 suggests a bigger variation in cancer rates, the cancer rates across the five states are also relatively close, ranging between 450 and 700 cases per 100.000 people. So even if a correlation was detected, this is another reason, why it probably would not be significant.

Conclusion

This report tries to answer, how the air pollution, measured by the AQI, in the five most populous states in the U.S. is correlating with the incidence of cancer. The analysis of the AQI data and crude rates from 2006 to 2021 across these states found no correlation, neither in the time domain for each state individually nor in the overall comparison of the states. One possible explanation for this is that the values – both over time and between the states – are relatively close and especially the cancer rates (because they cover all the cancer types) depend on a lot of factors.
The analysis was limited by the relatively small variations in air qualitiy and the cancer rates. Nevertheless, an analysis across countries would come with other challenges and factors (like diet, habits, environment). Another approach would be to analyze if there are states in the U.S. (or another populous country, so that the cancer rates are representative) with a significant difference in air quality and/or cancer rate and then check if there is a correlation. There are a lot of factors which influence cancer rates. That's why an analysis probably should be limited to the examination of lung cancer and not all cancer types (e.g. breast cancer is mostly hormonally related). These, often unknown, factors make it even more difficult to find out whether there is a correlation between air quality and cancer rates. While the findings suggest no visible or significant correlation, the question cannot be definitively answered, particularly given the limited variability in the data.

Citation:

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

Disclaimer:

ChatGPT was partly used to correct small grammar issues and vocabular with this prompt: "can i give you some text and you correct my grammar and maybe vocabular if something doesn't sound so good?" After that, the original text was pasted into the chat without further instructions. I read everything carefully and only corrected what I found plausible. The original (not corrected) text isn't provided here due to limited space but is available for further questions. ChatGPT did not make any changes to the content and only corrected the order of words or changed some vocabulary that was more appropriate.