The Air We Breathe: Air Quality and Health Outcomes in Kentucky

Jianing Gong[a], **Madhvi Malhotra**[b], **Maggie Sullivan**[c], **and Sanha Tahir**[d]

[a]All authors are M.S. candidates in Georgetown University's Data Science for Public Policy Program

This manuscript was compiled on December 6, 2022

Much research has been conducted on the relationship between air pollution and health outcomes. For our final project, we dived deeper into this relationship in a single state in the United States: Kentucky. With 2019 data from the U.S. Centers for Disease Control and Prevention (CDC) and the U.S. Environmental Protection Agency (EPA) , we conducted an exploratory analysis with data visualization and linear regression on overall, respiratory, and mental health outcomes. Due to a very limited sample size (potential political reasons discussed), our results did not show a statistically significant relationship, but we provided recommendations for further research.

Air Quality | Health Outcomes | Sulfur Dioxide | Policy

Air!, an invisible friend we need for survival and breathe it every second of our everyday lives. However, we are living in times where we are also breathing various invisible particles with it. Sometimes this can be harmful and would result in many unknown diseases. As per the survey by WHO (2021) *, 58 % of premature deaths related to outdoor air pollution were due to ischemic heart disease and stroke, while 18 % were due to chronic obstructive pulmonary disease and acute lower respiratory infection, respectively, and 6 % cause of death was due to lung cancer. Not only that, according to the American Psychological Association, polluted air is harmful to people's brains as well. When children are exposed to highly polluted air, their cognitive abilities may be damaged. Furthermore, there are studies that have concluded that exposure to certain PM can lead to oxidative stress-mediated inflammation in the brain.†. Other studies have found supporting evidence that these pollutants are connected to negative mental health outcomes.‡

Common air pollutants with known health impacts were first regulated as "Criteria Pollutants" by the 1970 Clean Air Act, which established health-based National Ambient Air Quality Standards (NAAQS). The six criteria pollutants are carbon monoxide(CO), ground-level ozone, lead, nitrogen dioxide(NO2), particulate matter(PM), and sulfur dioxide(SO2).

Out of these, Sulfur dioxide(SO2) has a particularly significant presence in Kentucky, which is one the largest coal-producing states in the United States. In fact, in 2016, Kentucky ranked as the fourth-highest coal producer in the United States, producing 42.9 million tons§. SO2 is a direct product of the combustion of coal and the smelting of sulfur-containing ores. It is a colorless, soluble gas with a characteristic pungent smell that forms sulphuric acid when combined with water. SO2 can also be directly credited for the creation of respiratory symptoms in healthy patients and those with underlying lung disease. Sulfur dioxide can cause respiratory problems such as bronchitis and can irritate your nose, throat, and lungs. It may cause coughing, wheezing, phlegm, and asthma attacks. The effects are worse when you are exercising. Sulfur dioxide has been linked to cardiovascular disease. Groups that are most sensitive to sulfur dioxide include children, adults with lung diseases, and asthmatics.

However, studies have not demonstrated dose-dependent health risk responses to increased exposure to these pollutants, except at high concentrations. Furthermore, many studies examining the effects of ambient-level exposure to NO2, SO2, and CO have failed to find associations with adverse health outcomes. This research aimed to find out the correlation between air quality and health status and is based on the 2019 data collected in Kentucky. The research analyzes 10 counties and over 13,000 data, running a regression model and showing results in a map visualization.

* Ambient (outdoor) air pollution, WHO,https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

† Particulate matter, oxidative stress, and neurotoxicity, Sheba MohanKumara, Arezoo Campbell, Michelle Block, Bellina Veronesi, 2007

‡ Growing evidence for the impact of air pollution on depression, Ali, Naureen Akber and Khoja, Adeel, 2019

§ https://eec.ky.gov/Energy/Coal%20Facts%20%20Annual%20Editions/Kentucky%20Coal%20Facts%20-%2017th%20Edition%20(2017).pdf

Data

Data Sources. Our data consisted of three main datasets:

1. Annual_aqi_by_county_2019.csv (referred to as "AQI" through the remainder of this paper),
2. Daily_42401_2019_SO2.csv (referred to as "Sulfur"), and
3. PLACES_Local_Data_for_Better_Health_Census_Tract_Data_2021_release.csv (referred to as "Health")

We downloaded the AQI and the Sulfur data from the U.S. Environmental Protection Agency (EPA) and the Health data from the Centers for Disease Control and Prevention (CDC). These two U.S. Government agencies regularly collect and publish these official datasets as part of their mandate. As a result, this data is freely available, relatively free of errors, and consistently available across many years.

We subset all datasets to 2019 data for the state of Kentucky. The 2021 Health data release is the most up-to-date data available on the CDC's website and contains data collected in 2018 and 2019. More recent data AQI and Sulfur data are available on the EPA website; however, we downloaded 2019 data to be able to compare to the Health data in the same year. According to the CDC[¶] website, this Health data is based on "the Behavioral Risk Factor Surveillance System (BRFSS), Census 2010 population counts, annual Census county population estimates, and the Census American Community Survey (ACS) estimates." According to the EPA website[‖], the AQI and Sulfur datasets contain both daily and annual data because of the Clean Air Act, which "requires that state, local, and tribal air pollution control agencies monitor the air for ambient levels of certain pollutants. In addition to the required monitoring, many agencies perform additional and/or voluntary monitoring of substances and meteorological parameters."

AQI. For the Air Quality data, we have used the Annual and Daily AQI data. The annual AQI data has the pollutant summary; all metrics (mean, maxes, percentiles) are based on the daily maximum 8-hour average, not the individual sample values. We have data for days with Carbon Dioxide (C O), Nitrogen (NO2), Ozone, PM2.5, and PM 10. The daily summary files show the AQI by county and have one record per day per county. For our visualization, we have calculated the monthly mean of the AQI data.

Sulfur. For the Sulfur data, we have used the Daily AQI Summary data. The file contains SO2 readings collected from different sites and associated metadata. The readings themselves include units of measure, and arithmetic mean/maximum/minimum values of observed counts. The associated metadata includes site numbers, site geolocations, method collection and local dates. These latter meta details were critical in helping us identify feasible methods to rollup data for appropriate analysis. The daily file shows the AQI by county, and may have multiple records based on different sample collection methods per day, per site.

Health. For the Health data, in line with our research questions, we chose to focus on respiratory and mental health-related health measures for each of the four health categories in the dataset. Our final list of health categories and the corresponding measures (and measure ids) are below:

- Health Outcomes: current asthma (CASTHMA), Chronic obstructive pulmonary disease among adults aged >=18 years' (COPD), Cancer (excluding skin cancer) among adults aged >=18 years (CANCER), and Depression among adults aged >=18 years (DEPRESSION)
- Prevention: Visits to doctor for a routine checkup within the past year among adults aged (CHECKUP), Current lack of health insurance among adults aged 18-64 years (ACCESS2), and Taking medicine for high blood pressure control among adults aged >=18 years with high blood pressure (BPMED)
- Risk Behaviors: Current smoking among adults aged >=18 years (CSMOKING) and No leisure-time physical activity among adults aged >=18 years (LPA)
- Health Status: Physical health not good for >=14 days among adults aged >=18 years (PHLTH), Mental health not good for >=14 days among adults aged >=18 years (MHLTH), Fair or poor self-rated health status among adults aged >=18 years' (GHLTH)

[¶] Methodology, CDC,https://www.cdc.gov/places/methodology/index.html,2022

[‖] About AQS Data, EPA,https://aqs.epa.gov/aqsweb/documents/about_aqs_data.html,2022

We excluded all other health measures that were not relevant to our current study, such as "Arthritis among adults aged >=18 years" and "Chronic kidney disease among adults aged >=18 years." All 120 Kentucky counties were represented in the data, with each country containing anywhere from 1 to 190 prevalence values for the focus health measure. For our analysis, we took each county's mean and median prevalence values for each of our focus measures, so there was one corresponding crude prevalence value for each county. In addition to the county, the dataset included location data by CountyFIPs, Geolocation, and LocationId.

We focused on the county name and the latitude and longitude values in the Geolocation column. For our final analytic sample, we dropped columns with duplicate information and other less-important to only keep the following column values: 'index', 'Year', 'CountyName', 'CountyFIPS', 'Category', 'Measure', 'Data_Value_Unit', 'Data_Value_Type', 'Data_Value', 'TotalPopulation', 'Geolocation', 'LocationID', 'MeasureId'. See the image below for more information about the original dataset post subset to 2019 and Kentucky.

Limitations. Our data sources and research approach do have limitations that may affect the results and conclusion. For example, the health data only includes results from adults (age 18+). Therefore, we cannot accurately isolate whether people at certain ages (especially children) experience the health impacts of air pollution differently. Also, the health data on cancer does not disaggregate on the type (only noting that skin cancer is excluded), so we cannot isolate specific cancers like lung cancer. We also lack detailed demographic data (such as race or gender) to better understand the population in the health surveys. With only one year of data, it is also difficult to accurately identify trends in any of the datasets.

Finally, our sample size is extremely limited in Kentucky. Although we have health data for all Kentucky counties, the AQI data was only collected in 27 counties, while Sulfur data was collected in even fewer counties, at 10. The Kentucky state government website provides some reason for this: "The locations of the monitoring stations are selected in accordance with U.S. EPA regulatory requirements and are generally established near populous areas or pollutant sources. Each year the station locations are reviewed to ensure that adequate coverage is being provided."

Although coal production (and consequently coal mining employment) in Kentucky has been declining for years, the coal industry is still integral to Kentucky culture and politics both in rhetoric and campaign contributions. As one Kentucky professor shared, "Like bourbon and horse racing, coal is an integral part of Kentucky's identity. If you're going to win an election, you need to be supportive of the industry."[**] For example, in a heated senate race in 2016, a GOP-associated non-profit reportedly helped incumbent Mitch McConnel conduct a 12-week, 750,000 digital campaign which explicitly aimed "to educate Kentuckians on the disastrous policies of the Obama Administration when it comes to the Commonwealth's coal-based economy, and on legislation aimed at stopping those policies." Coal is still king in Kentucky.[††]

Methods

Data Cleaning. The data acquired from the sources was in excellent condition and did not require extensive cleaning at the elemental level.

For the purposes of our explorations, analysis, and visualizations, however, certain modifications needed to be made and are detailed below:

- **Health**: We subset the data using a custom function to our focus measures:
  - Current asthma (CASTHMA),
  - Chronic obstructive pulmonary disease among adults aged >=18 years' (COPD),
  - Cancer (excluding skin cancer) among adults aged >=18 years (CANCER),
  - Depression among adults aged >=18 years (DEPRESSION)
  - Visits to doctor for a routine checkup within the past year among adults aged (CHECKUP)

[**] Coal-Fired Politics: In Kentucky Senate Race, Bitter Rivals Woo a Dying Industry, Katherine Bagley,https://insideclimatenews.org/news/08102014/coal-fired-politics-kentucky-senate-race-bitter-rivals-woo-dying-industry/,2014

[††] Trump, McConnell promised to bring back coal. But Kentucky lost 2,700 jobs on their watch,https://www.courier-journal.com/in-depth/news/politics/elections/kentucky/2020/10/14/election-2020-kentucky-coals-influence-nowhere-found/5927720002/

  - Current lack of health insurance among adults aged 18-64 years (ACCESS2)
  - Taking medicine for high blood pressure control among adults aged >=18 years with high blood pressure (BPMED)

We did this by aggregating on the basis of category and Measure.

- **Air Quality**: This was a relatively clean and aggregated data, by virtue of it being an annual file data.
- **Sulfur Data**: While most counties only had one site for sulfur data collection, the counties of Jefferson and Henderson had two and three sites. Each site would then have records for all days of the year and for different data collection methods. Using thresholds set by the EPA, the AQI values were then used to categorize days as:
  - Good days (if AQI < 50)
  - Moderate Days ($50 <$ AQI < 100)
  - Unhealthy Days for Sensitive Groups ($100 <$ AQI < 150)
  - Unhealthy Days ($150 <$ AQI)

  Our analysis then led us to group by county name and site number, aggregating for general, maximum, and minimum AQI values as well as aggregating to find the total days for each type in the entire year of 2019.

See 0_Final Data Cleaning.ipynb for all data cleaning code.

Data Merging. When we began our data merging and before we had condensed our datasets, our final air-health-sulfur dataset was confusingly large, with over a million rows.

Upon deeper analysis, we realized that it was because of the duplication of County Names in both data sets; each county would have multiple records of health measures and multiple records of sulfur data reports. In this instance of our journey, we realized the need for condensing data. By this, we mean that each county should only have 1 record.

While this would have made our coding exercises much less arduous, we understand that it would have been an oversimplification. It would have condensed different measures of health into a single number and readings from different sites into one.

As such, we decided to group the SO2 using County Name and Site Number, and similar methods for the health dataset. These methods maintained the integrity of the individual readings while generalizing just enough for us to gain valuable insights.

After this, it was not challenging to merge the three datasets on the name of the Kentucky county. As this was government-issued, there was a great deal of standardization across the County Names.

This post-condensation merging left us with 6,900 rows; a massive reduction in the volume of the merged dataset and allowing for more efficient analysis down the line.

However, a county is indeed a large area for us to make any conclusive analyses. With this motivation, we set out to calculate the distance between regex-cleaned geolocations to calculate the distance between the sites of data collection, using the following Haversine formula:

$$d = 2 * r * arcsin * \sqrt{(sin^2((x_2 - x_1)/2)) + (cos(x_1) * cos(x_2)) * (sin^2((y_2 - y_1)/2))}$$

where:
x_i = latitude of point i
y_i = longitude of point i

It is pertinent to note that while both health and sulfur data had geolocations, these were stored in a 'point', string format. For further analysis, longitude and latitude were extracted using regex methods. We then refined the location information to precisely locate collection points within 15 miles of each other.

Prior to regression, we also identified that each county in the health data set might have up to multiple different data values for each measure. Therefore we grouped by county and took the mean measure of the data values for each focus measure. We pivoted this data to a wide dataframe where each column had the average data value. This allowed us to have one unique value for each county for each focus measure. See 1_Final Data Merging. ipynb for all data cleaning codes.

Regression. When running a simple linear regression on the data, created two new dataframes: one by merging the AQI and Health data alone and the other by merging Sulfur and Health data alone. This allowed us to maintain clean dataframes for two different sets of potential independent variables and test multiple dependent variables.

To do so more easily, we created four functions (two for each dataset). These functions created a custom scatterplot with a regression line and then ran a linear regression model and printed the summary. See 4_Regression.ipynb for these functions. We then used these to test seven dependent variables related to health (all focus measures and Health Outcomes and Health Status).

Results. We explored many possible questions with our data. In the regression analysis, likely due to our small sample size of Kentucky counties, we found common trends in our scatterplots but no statistically significant relationships. For example, fewer people reported depression as the percentage of moderate+ AQI days increased. However, the air quality data in most cases actually resulted in an opposite trend of what we would expect (i.e., health outcomes improved in relation to a higher number of moderate+ AQI days). In one example, the average percentage of people reporting poor health actually decreased as the percentage of moderate+ AQI days increased. In contrast, sulfur data generally, but this was primarily due to an outlier county. These relationships, again, were not statistically significant but should be explored with a larger sample size. See 4_Regression.ipynb for all regression plots and model summaries and below, we record some of our more significant investigations:

- **Figure 1** shows the relationship between the number of non-good AQI days and depression. We found out that as the percentage of moderate or worse air quality days increases, fewer people report cancer. More precisely, with each 0.1 increase in the proportion of moderate+ AQI days, the portion of people who report depression goes by 7 percent.

- **Figure 2** on the other hand, shows a positive relationship between mean SO2 observations and the percentage of people reporting poor health; an increase of 1 point in the mean SO2 observations increases the portion of people reporting poor health with 0.15

- **Figure 3** show that for every 0.1 percentage increase in the portion of days with moderate or unhealthy AQI, there is a 9 percent decrease in people reporting poor health.

As noted above, we failed to reject the null hypothesis for these regression results i.e. this result does not hold statistical significance, which is generally counter-intuitive to our expectations. While our suggested reasons for this insignificance have already been discussed, they might be worth recapping:

- Small data sample, both in terms of the number of counties and in terms of the time for which this analysis was being done

- Political context may discourage data collection of sensitive measures

We also went on to investigate the variation in reported monthly average AQI across the year:

- **Figure 4** shows a heat map of the Monthly Average AQI for all the county's in Kentucky. The higher the AQI value, the greater the level of air pollution and the greater the health concern. We can see that the AQI is lower during colder months except for Jefferson, where the AQI is consistently higher as compared to other counties. The grey boxes show the missing AQI for the County.

In order to understand the relationship between the sites for sulfur data collection and health measures data collection, we went on to use the folium package to map both datasets. The interactive visualizations provided here, will allow for deeper insight into each data point, but here are a few highlights:

- **Figure 5** shows the locations of health data collection points with associated measures as hover points. Superimposed on this map are the sites for sulfur data collection, surrounded by a 15-mile, crimson circle. The figure illustrates that generally, health measures are being collected within 15 miles of the where sulfur data is being collected. Having said that, there does not seem to be a relation between sulfur collection sites and the health measures being collected in that area.

- **Figure 6** shows a magnification of this map, in the busy city of Louisville, which has three different sites for sulfur data collection. On the other hand, there is a vast number of sites for health data collection, which are almost entirely clustered towards the eastern side of the city. The mean AQI values for these sites are between 1 and 3, which fails to make any clear association between the sites that are selected for sulfur data collection and health measure data collection. However, it is interesting to note that there are more health measures being recorded for this urban city (a total of 6,840) than in other areas e.g. Lexington, another urban area that has 2 sulfur collection sites and 1092 records of health measures, or Paducah, a smaller area with only one sulfur collection point and 204 records of health measures.

- **Figure 7** shows a similar magnification of Mammoth National Park. While we only have one site for sulfur data collection, there are 3 sites for health measure collections (with 12 measures being recorded in each) in the busy city of Louisville, which has three different sites for sulfur data collection. The sites for health data collection points are clustered towards the eastern side of the city. The mean AQI values for these sites are between 1 and 3, which fails to make any clear association between the sites that are selected for sulfur data collection and health measure data collection.

Summary. Our findings do not support a statistically significant relationship between air quality and health outcomes in Kentucky. However, as previously discussed, there are many limitations to this research exercise. The primary limitations of our analysis include the lack of AQI and Sulfur data for all 120 counties in Kentucky and the study of data from only one year. For future analysis, we recommend a number of ways to expand on our research. First, this could be adapted to a longitudinal study to better understand trends over time in health outcomes and air quality levels. As many health effects are not seen for years, this longitudinal study might demonstrate a stronger impact than our study of a sample from one year. For another example, one could conduct a national analysis of AQI, sulfur, and health data, with dummy variables for coal-producing states and non-coal-producing states. This would help isolate the influence of coal production on health outcomes.

In addition to providing a basis for future research, our work has policy implications as it highlights a major data gap in air quality across Kentucky counties. This research could be used as an argument to expand air quality measures across additional Kentucky counties. However, due to the political influence of coal mining and power companies in Kentucky, this is likely to face strong pushback from the private sector as well as government officials with strong ties (particularly financial ties from campaign donations) from these private sector companies.

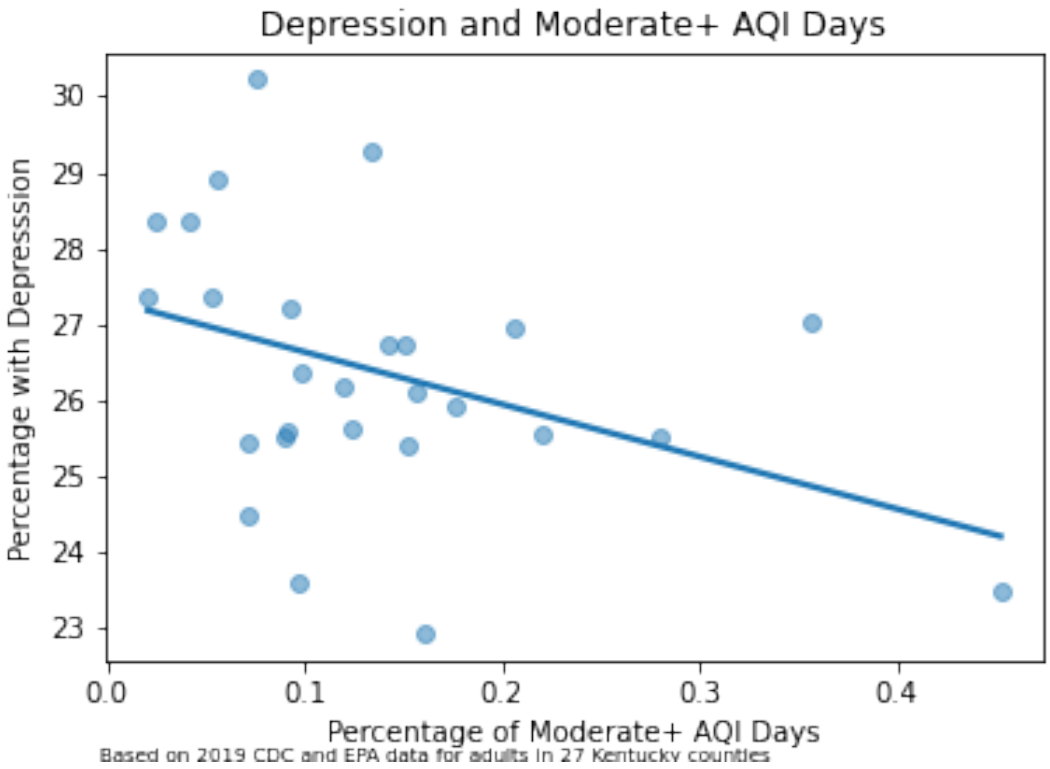

Fig. 1. Regression result for Depression and Moderate+ AQI Days in Kentucky. Further statistical modeling did not find this relationship significant.

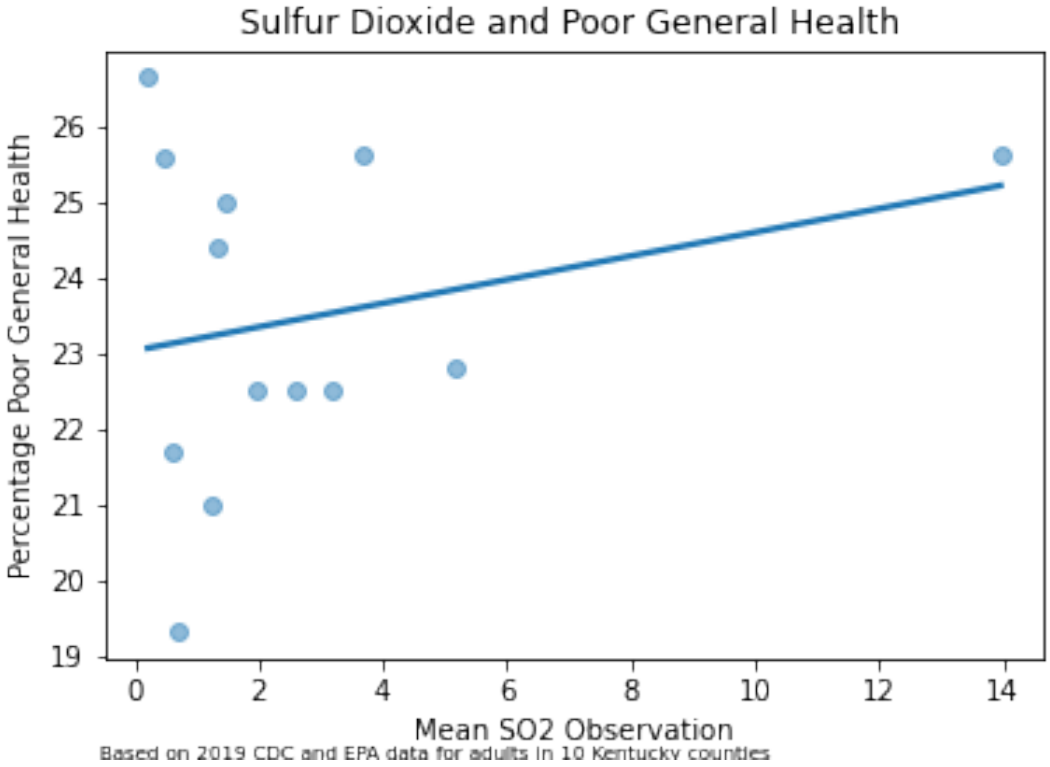

Fig. 2. Regression result for Sulfur Dioxide and Poor General Health in Kentucky. Further statistical modeling did not find this relationship significant.

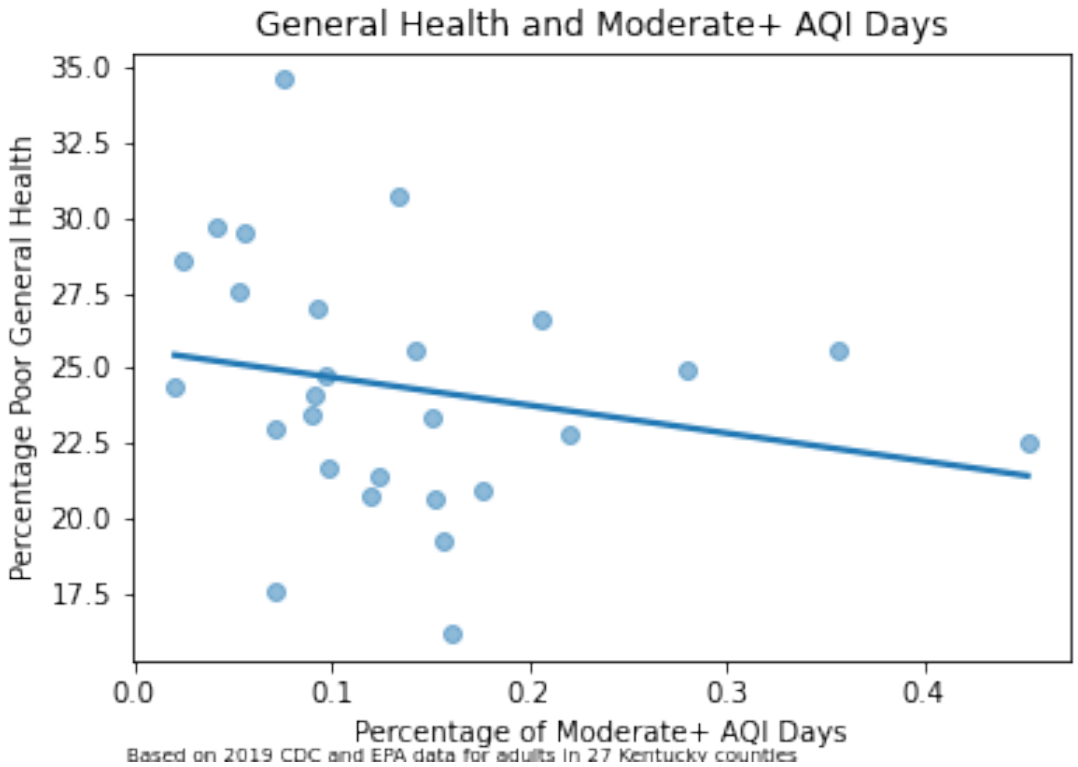

Fig. 3. Regression result for Poor General Health and Moderate+ AQI Days in Kentucky. Further statistical modeling did not find this relationship significant.

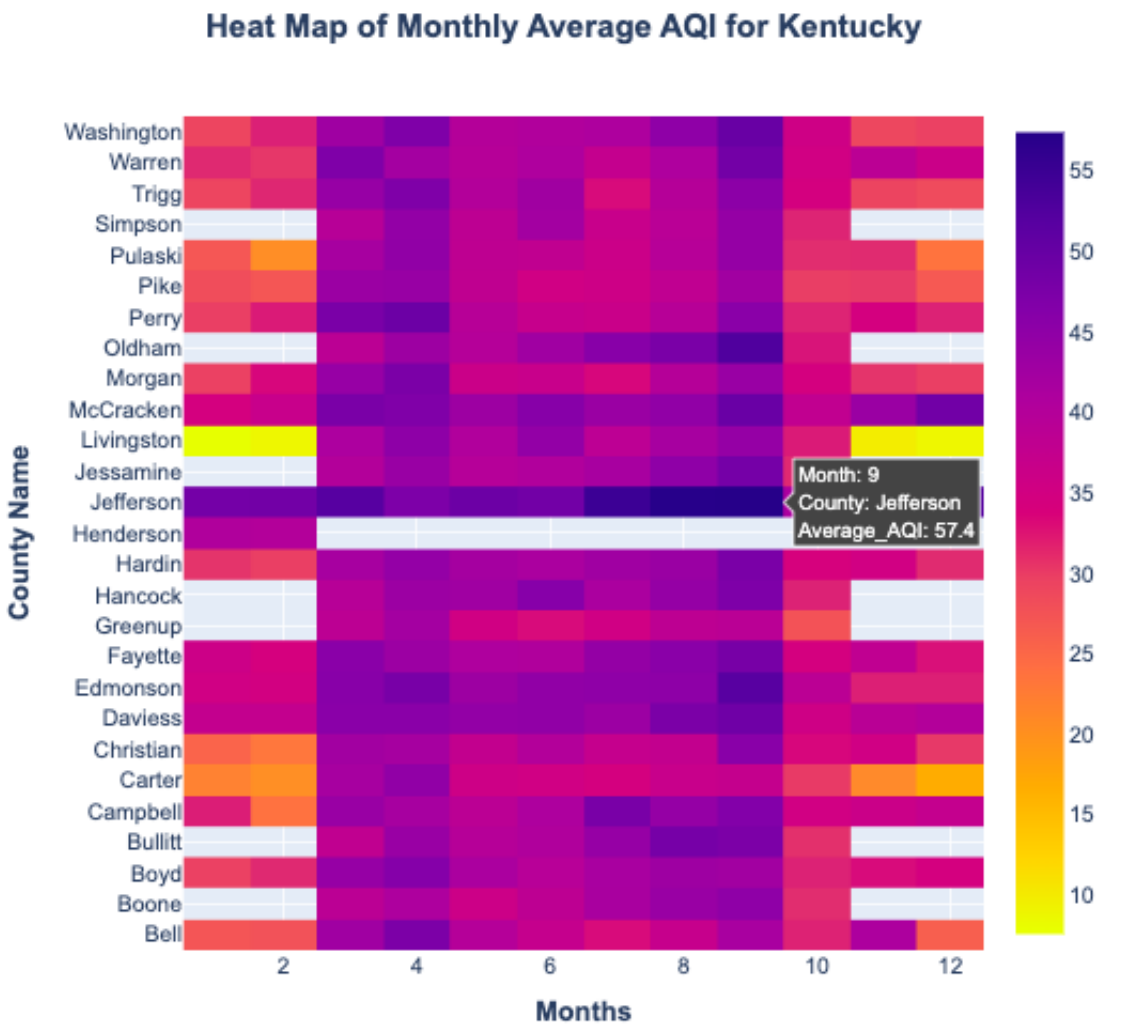

Fig. 4. Heat Map of Monthly Average AQI for Kentucky

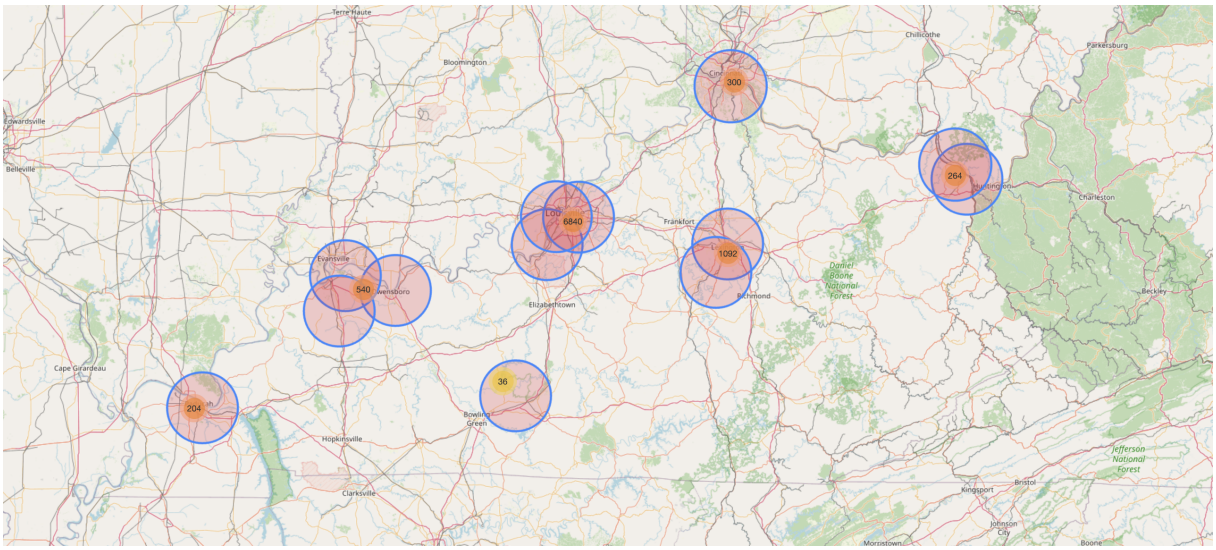

Fig. 5. Comparing data collection points to sulfur and health

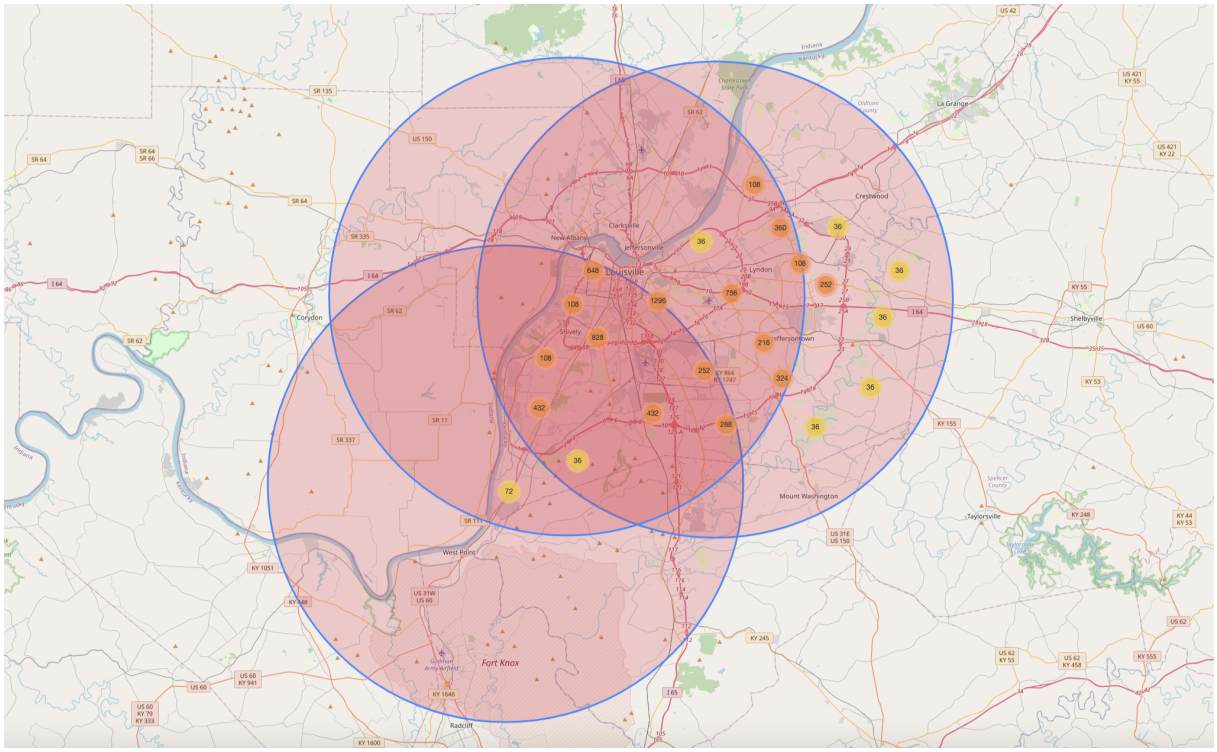

Fig. 6. Comparing data collection points to sulfur and health: zoomed-in view of Louisville, the largest city in Kentucky.

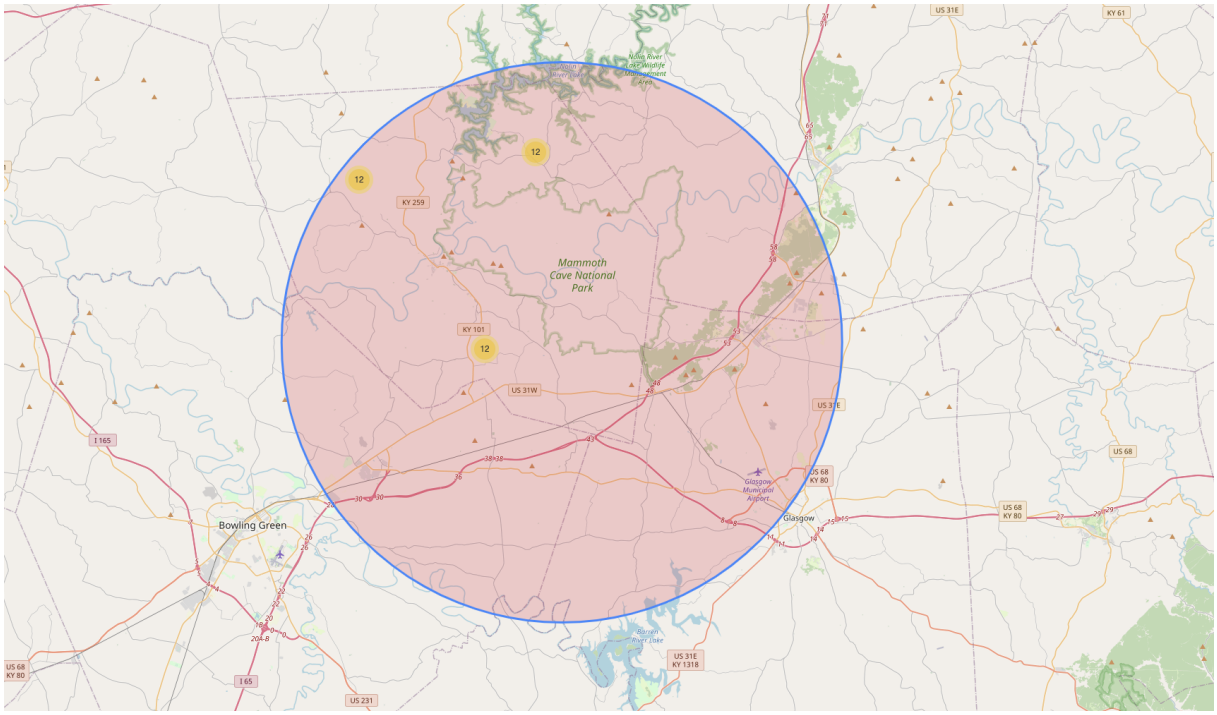

Fig. 7. Comparing data collection points to sulfur and health: zoomed-in view of Mammoth National Park