The Link Between Air Quality and Health

https://github.com/brho7443/DataMineProject/tree/main/Data_Mine_Project

Anish Timalsina
anti5381@colorado.edu
Student id: 110277849
CSCI 4502-001

Adam Prieto
adpr7816@colorado.edu
Student id: 106640190
CSCI 5502-001

Brock Hoos
brho7443@colorado.edu
Student id:
CSCI 4502-001

Scott Silverstein
scsi3286@colorado.edu
Student id:
CSCI 5502-001

ABSTRACT

This project investigates the critical relationship between air quality and public health outcomes in the United States. By analyzing trends in respiratory diseases, life expectancy, mortality rates, and CO2 emissions from various sources, this study highlights the impact of poor air quality on public health. Leveraging datasets such as emissions data, asthma rates, and life expectancy statistics, a comprehensive analysis was conducted to understand historical trends and identify patterns. The results reveal alarming increases in respiratory conditions and mortality associated with worsening air quality, emphasizing the need for actionable insights for policymakers to mitigate these effects.

INTRODUCTION

Poor air quality poses a significant threat to both human health and the environment, with severe impacts on respiratory health and an increased risk of mortality. Research has shown that pollutants like ozone and particulate matter (PM) are directly linked to respiratory conditions, including asthma, as well as cardiovascular issues, contributing to higher rates of hospitalizations and premature deaths (EPA, 2024). Vulnerable populations—such as children, the elderly, and individuals living in high-pollution areas—face heightened risks, as these groups are more susceptible to the detrimental effects of air pollution. This problem is particularly pressing due to its widespread effects and the challenges in mitigating its sources, including industrial emissions, vehicular pollutants, and reliance on fossil fuels. Climate change and extreme weather further exacerbate the situation, making air quality management increasingly complex. Additionally, understanding the role of medical history is crucial, as individuals with pre-existing respiratory or cardiovascular conditions are at greater risk from poor air quality. Key ideas include leveraging the abundance of data available from government sources, satellite imagery, and medical records to develop sophisticated models that anticipate air quality fluctuations in real time. These models integrate diverse datasets to identify trends, account for individual medical histories, inform policy decisions, and ultimately protect vulnerable populations. By focusing

on historical patterns, medical impacts, and current challenges, this project seeks to contribute actionable insights to mitigate the effects of poor air quality and improve public health outcomes.

RELATED WORK

(what has been done in the literature? limitations? How is your work different?

Real-time air quality data has become accessible through networks like AirNow and the EPA, which manage numerous monitoring stations across the United States. Machine learning models—such as Random Forest, Gradient Boosting, and Neural Networks—are widely used to predict air quality levels based on this data, with applications from companies like Plume Labs and IQAir offering consumer-focused forecasts.

While past studies have explored the impact of traffic and industrial emissions on air quality in cities like Beijing and Los Angeles, they often rely on single-source data. This approach can limit the models' ability to capture complex interactions between multiple pollutants and environmental factors. The EPA emphasizes that multi-pollutant modeling is crucial for assessing cumulative effects, as factors like temperature and humidity can alter pollutant formation and intensify health risks, particularly for asthma and other respiratory conditions (EPA, 2024; Tiotiu et al., 2020). Such findings underscore the need for comprehensive models that integrate diverse data sources and account for pollutant interactions.

Our work is different because (.....)

METHODOLOGY

Dataset

This study utilized multiple datasets to analyze the relationship between air quality and health outcomes. Each dataset was carefully preprocessed to ensure consistency and reliability for analysis.

We used these datasets:

AsthmaTotals.csv

- Fields: State, Total, Year
- Description: This dataset provides asthma case counts by state and year, enabling analysis of how asthma prevalence is influenced by air quality. It helps highlight areas with high asthma rates and correlate them with pollutant levels or other environmental factors.

ChildhoodMortality.csv

- Fields: Year, Age Group, Death Rate
- Description: This dataset contains childhood mortality rates from 1900 to 2018, segmented by age groups ranging from 1–4 years in 1900 to 15–19 years in 2018. This data helps evaluate long-term health trends influenced by air quality and other factors.

LEAB.csv

- Fields: Year, Race, Sex, Average Life Expectancy, Age-Adjusted Death Rate
- Description: This dataset provides life expectancy trends and death rates adjusted for age, helping assess whether air quality has had a significant impact on mortality and life expectancy over time.

LeadingCauseOfDeathInUSA.csv

- Fields: Year, 113 Cause Name, Cause Name, State, Deaths, Age-Adjusted Death Rate
- Description: This dataset tracks leading causes of death in the United States, providing insights into health trends influenced by environmental and social factors.

Mortality-Rates-by-State.csv

- Fields: State, Year, Total Mortality Rate
- Description: This dataset records mortality rates across states and years, offering a metric to correlate air pollution levels with broader public health impacts.

RespiratoryConditionsPerMedicaid.csv

- Fields: State, Year, Month, Condition, Beneficiary Count, Percentage of Beneficiaries, Data Quality
- Description: This dataset tracks respiratory conditions such as bronchitis, pneumonia, and influenza. By analyzing this data, the study highlights disease trends over time and their potential link to air quality.

Emissions.csv

- Fields: Year, State, Sector, Fuel, Emissions Value
- Description: This dataset categorizes CO2 emissions by state, sector, and fuel type (coal, petroleum, natural gas). It helps pinpoint sources of pollution and track emission trends over time.

DiseaseData.csv

- Fields: Year, Disease Name, Number of Diagnoses
- Description: This dataset tracks the number of diagnoses for various diseases such as the cancer cases, across different years.

Data Preprocessing

Data Cleaning:

Data Transformation:

Tools And techniques

Used python and its libraries like pandas, numpy, and matplotlib .

preprocessed the data by handling missing values, standardizing formats, and removing outliers.

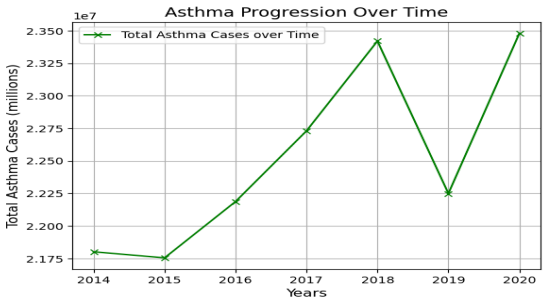
statistical analysis to correlate emissions with health outcomes

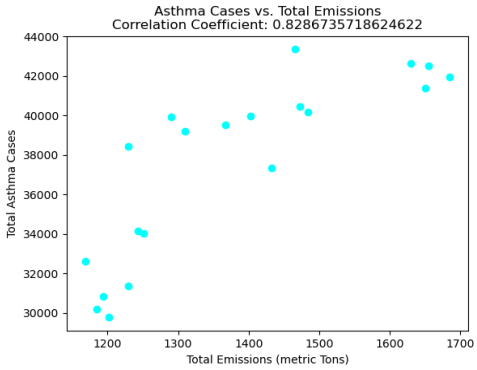
visualization of identifiable data trends and patterns.

Key analysis:

Asthma Trends

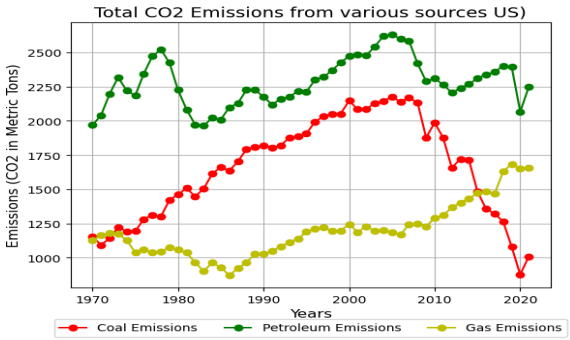
Asthma cases in the United States increased steadily, peaking at approximately 2.3 million cases. A dip in 2019, where cases were recorded at 2.225 million, may be attributed to disruptions caused by the COVID-19 pandemic

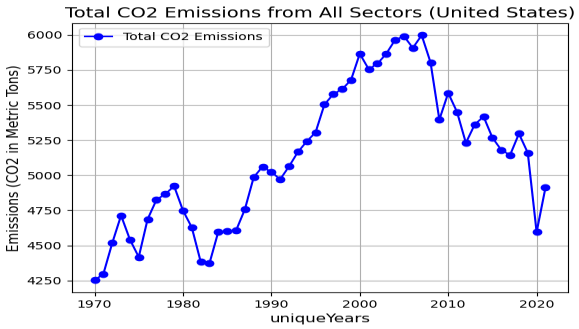




CO2 Emission Trends

Emission from Petroleum were found to dominate total co2 contributions in recent years

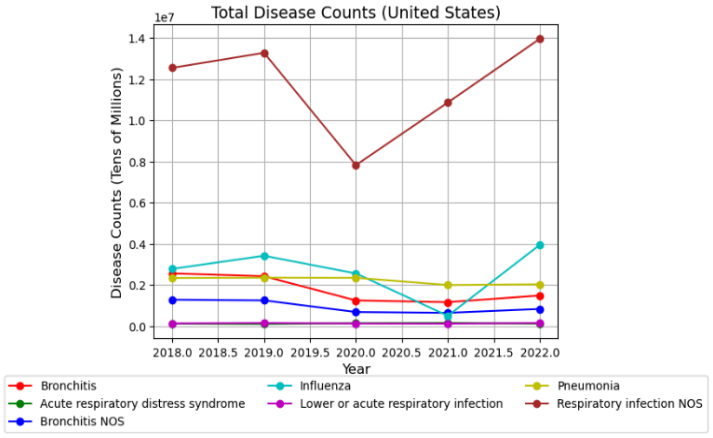




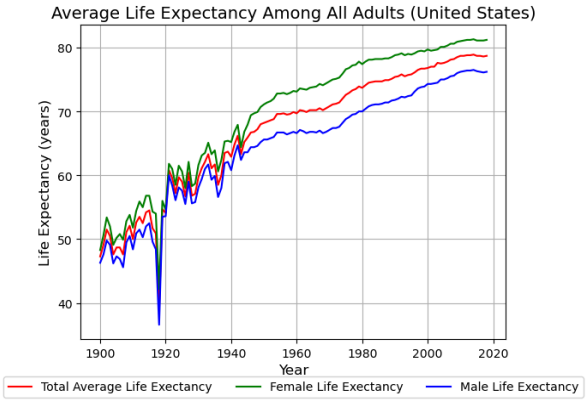
The reduction in coal emissions reflects shifts in energy policies. With a transition toward cleaner energy sources like natural gas, there has been a substantial reduction in CO2 emissions, declining from approximately 6050 metric tons in 2008 to 4850 metric tons in 2021. Petroleum continues to dominate total CO2 emissions, underscoring the ongoing reliance on fossil fuels

Respiratory Disease Trends

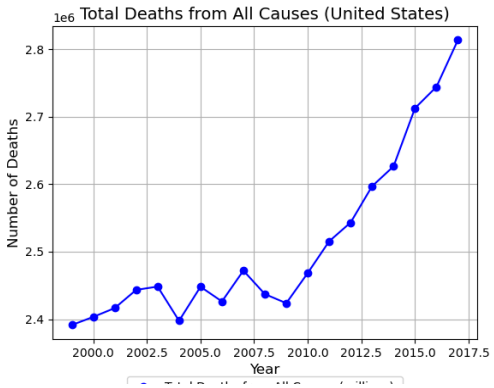
Respiratory diseases such as respiratory infection NOS and influenza fluctuated between 2018 and 2022, with notable spikes during and after the COVID-19 pandemic. Respiratory infection NOS still remained the most common condition throughout this period.



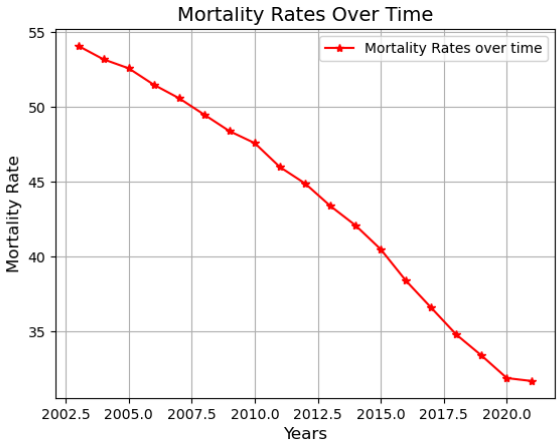
Mortality and Life Expectancy Trends

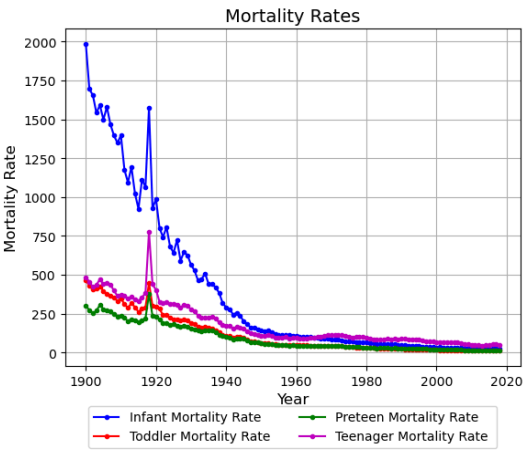


The data shows the steady increase in life expectancy for all adults in the U.S. from 1900 to 2020. Both male and female life expectancies exhibit upward trends, with women consistently outliving men.



This graph displays the rising total number of deaths in the U.S. from 2000 to 2017. Deaths increased steadily, reflecting factors such as population growth and an aging demographic





There is also a dramatic decline in mortality rates across all age groups from 1900 to 2020.

Life expectancy for adults in the U.S. increased steadily from 1900 to 2020, reflecting advancements in healthcare and living standards. Total deaths in the U.S. rose consistently between 2000 and 2017, largely due to demographic factors such as population growth and aging.

EVALUATION

The success of our project will be evaluated using both quantitative and qualitative metrics:

Metrics

Quantitative Metrics:

1. Root Mean Squared Error (RMSE): Used to evaluate the accuracy of predictive models for respiratory disease cases based on emissions and environmental factors.
2. R^2 (Coefficient of Determination): Assessed the strength of correlations between CO2 emissions, asthma cases, and respiratory diseases.
3. Change Ratios: Quantified the rate of change in emissions, mortality rates, and respiratory disease counts over time.

Qualitative Metrics: (not sure if needed)

1. Stakeholder Feedback: Evaluated engagement with interactive tools, as well as the perceived utility of visualizations for policymaking.
2. Real-World Impact: Success is measured through positive feedback from stakeholders and implementation of insights in public policies or health-related actions.

Evaluation Setup

1. Datasets:

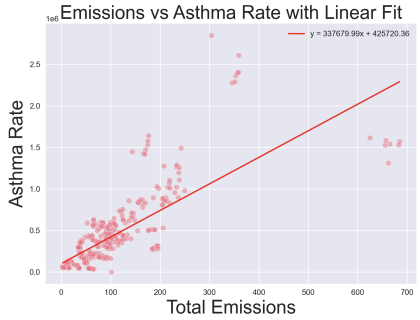
2. Data Partitioning:

3. Baseline Methods: Historical averages of disease counts and static emissions levels served as baselines for predictions

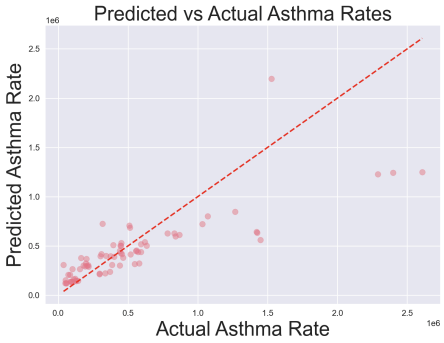
Results:

Asthma case:

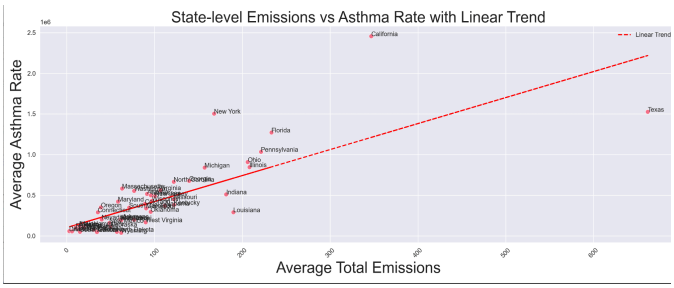
The relationship between total CO2 emissions and asthma rates was explored using linear regression. The data highlights the correlation between emissions and asthma cases, supported by a correlation coefficient of 0.788. The model achieved a performance score of $R^2 = 0.602$ and RMSE = 337,946.275.



This graph shows the direct relationship between total emissions and asthma rates. The linear fit suggests a strong positive correlation.



This graph visualizes the model's accuracy by comparing predicted asthma rates against actual rates. The tight clustering along the diagonal indicates good model performance.

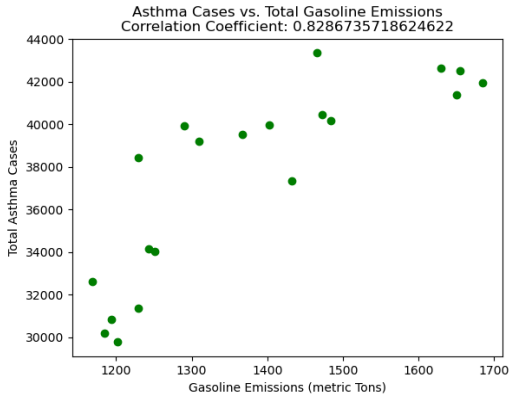


This graph plots state-level emissions against asthma rates, providing insights into regional variations and trends. States like California and Texas stand out with higher emissions and asthma rates.

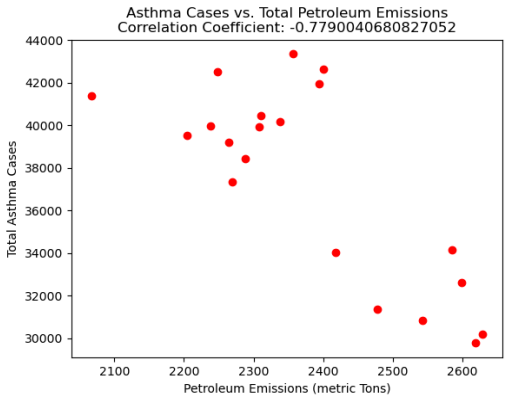
Asthma vs Types of Emissions

Asthma cases have shown a strong correlation with total emissions, but the relationship is highly dependent on the type of emission source. Each type of emission: gasoline, petroleum, and coal affects asthma cases differently

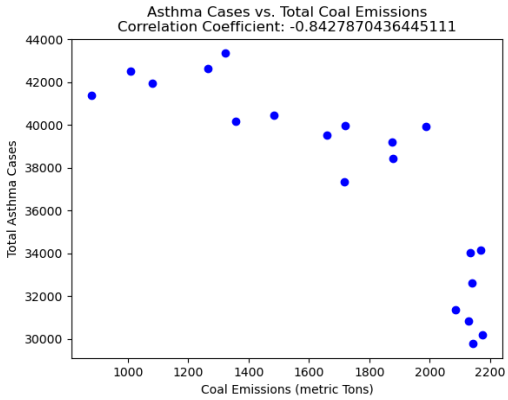
Gasoline Emissions: A positive correlation coefficient of 0.828 indicates that higher gasoline emissions strongly correlate with increased asthma cases. This suggests that areas with higher gasoline emissions experience more significant respiratory health impacts, underlining the importance of targeting vehicle emissions in public health policies.



Petroleum Emissions: The correlation coefficient of -0.779 suggests an inverse relationship between asthma cases and petroleum emissions. As petroleum emissions increase, asthma cases tend to decrease slightly. This may be due to improvements in refining processes or regulatory interventions targeting petroleum-related emissions.



Coal Emissions: A strong negative correlation coefficient of -0.842 demonstrates that reductions in coal emissions are closely associated with a decline in asthma cases. This result also highlights the effectiveness of policies aimed at reducing coal usage and emissions, with a direct positive impact on public health outcomes.

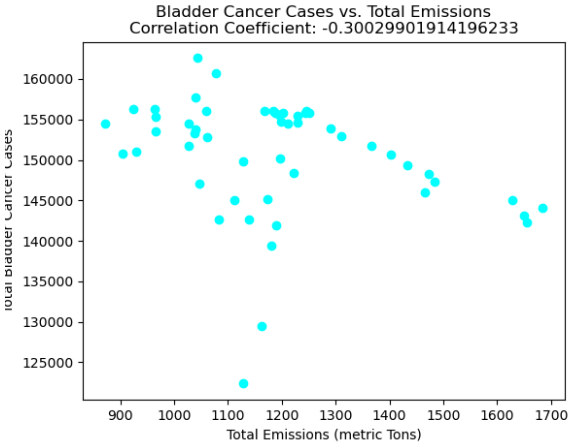


Cancer Cases:

Bladder cancer case:

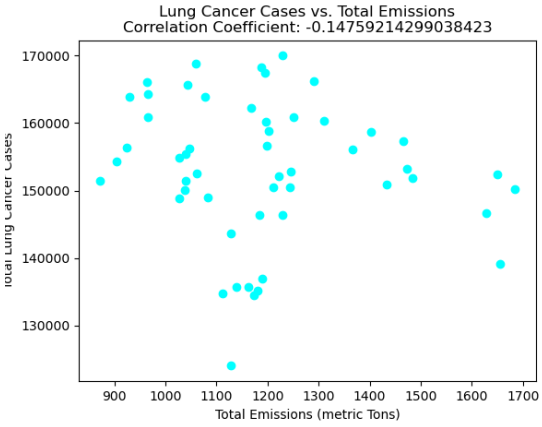
In bladder cancer cases, coal and petroleum emissions exhibit moderate positive correlations (0.454 and 0.472, respectively), suggesting a potential contribution to increased bladder cancer prevalence. Conversely, gasoline

emissions show a weak negative correlation (-0.300), indicating minimal impact from this source. Total emissions, representing the combined contributions of coal, petroleum, and gasoline, also show a weak negative correlation (-0.300), suggesting that while coal and petroleum emissions individually contribute to increased bladder cancer cases, the overall impact of total emissions is less significant.



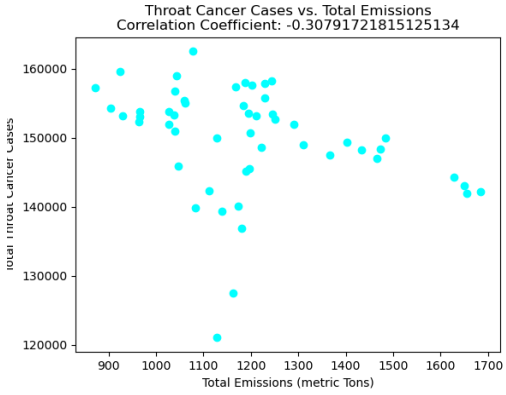
Lung Cancer case:

In lung cancer cases, petroleum emissions show a moderate positive correlation (0.304), suggesting a potential contribution to increased lung cancer. Coal emissions also exhibit a moderate positive correlation (0.344), further highlighting their potential impact on lung cancer cases. However, gasoline emissions show weak negative correlations (-0.148).



Throat Cancer Case:

For throat cancer cases, coal and petroleum emissions display moderate positive correlations (0.428 and 0.432, respectively), suggesting that higher emissions from these sources are associated with an increase in throat cancer prevalence. In contrast, gasoline emissions show a weak negative correlation (-0.307), indicating minimal impact. Total emissions, which combine contributions from all sources, also reflect this weak negative trend, highlighting that coal and petroleum emissions are the primary contributors to throat cancer cases, while the overall impact of all emissions combined is less pronounced.



Interpretation

1.) Strong Correlations:

We found a strong correlation between asthma and gas/total emissions This does not mean that gas/total emissions cause asthma because family heritage and many other factors can be the cause of this, but it does mean that there is a significant pattern between these totals that needs to be further explored. We found a slight correlation between throat cancer and coal emissions This does not mean that throat cancer is caused by coal emissions, but rather there seems to be an interesting pattern that calls for further

exploration.

2.) Impact of policy:

The reduction in coal emissions reflects successful policy interventions aimed at transitioning to cleaner energy sources. For example, the decline from 6050 metric tons of CO2 in 2008 to 4850 metric tons in 2021 highlights the impact of adopting alternative fuels such as natural gas.

These changes likely contributed to improvements in overall public health, as evidenced by declining mortality rates and steady increases in life expectancy.

3.) Health Disparities:

4.) Challenges:

Disruptions like the COVID-19 pandemic can obscure trends in disease progression, making it difficult to interpret short-term dips or spikes.

A significant challenge encountered during this study was finding complete, consistent datasets. For example, gaps in historical asthma data (1970–2000) and disparities in respiratory disease reporting limited the depth of some analyses.

DISCUSSION

(lesson learned, what worked well, what didn't directions for future work)

ETHICAL CONSIDERATIONS

When researching available datasets for this project, it became apparent that there was a major lack of data regarding different diseases and case numbers in the United States. Upon researching *why* this was the case, we learned that American health regulations often prevent direct reporting of disease numbers to protect patient information and liberties. Since patient confidentiality is a core tenet in healthcare, our research needed to focus on data that respected existing safeguards and protocols. In addition, using patient health data to study the effects of air quality on health must also take other biases into account. These include — but are not limited to — race, gender, household income, and location, which further impact availability of data and patient privacy protections. In conclusion, while existing health regulations are crucial for safeguarding patient health, they present significant research challenges that forced us to reframe our goals and project structure, but laws should alway be followed to maintain patient protection and keep healthcare safe for all.

CONCLUSION

(summary, reiterate key tasks and finding)

We didn't find a very strong correlations with lung cancer or any emission

We assume we didn't find any strong correlations because of the advancements in technology and medicine or other factors

REFERENCES

"Research on Health Effects from Air Pollution." *EPA*, Environmental Protection Agency, www.epa.gov/air-research/research-health-effects-air-pollution

Tiotiu, Angelica I., et al. "Impact of Air Pollution on Asthma Outcomes." *MDPI*, Multidisciplinary Digital Publishing Institute, 27 Aug. 2020, www.mdpi.com/1660-4601/17/17/6212.

Association, American Lung. "Lung Cancer Trends Brief: Mortality." *Lung Cancer Mortality - Lung Cancer Trends Brief | American Lung Association*, www.lung.org/research/trends-in-lung-disease/lung-cancer-trends-brief/lung-cancer-mortality-(1)

"Archived 'most Recent' Asthma Data (2018-2014)." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 11 May 2023, www.cdc.gov/asthma/archivedata/index.html

"Population Per State 2010-2020." *Https://Www2.Census.Gov/Programs-Surveys/Popest/Datasets/2010-2020/State/Totals/Nst-Est2020-Popchg2010-2020.Csv*.

"AirData Website File Download Page." *EPA*, Environmental Protection Agency, aqs.epa.gov/aqsweb/airdata/download_files.html#Meta

Air Pollution and Lung Cancer: A Review by International ..., www.jto.org/article/S1556-0864(23)00601-9/fulltext.

Wilkinson, Alexander J K, et al. "Greenhouse Gas Emissions Associated with Suboptimal Asthma Care in the UK: The Sabina Healthcare-Based Environmental Cost of Treatment (Carbon) Study." *Thorax*, U.S. National Library of Medicine, 27 Feb. 2024, pmc.ncbi.nlm.nih.gov/articles/PMC11041603/#:~:text=Poorly%20controlled%20asthma%20was%20associated,10%20000%20person%2Dyears

APPENDIX

On my honor, as a University of Colorado at Boulder student, I have neither given nor received unauthorized assistance on this work.

~ Anish Timalsina, Adam Prieto, Brock Hoos, Scott Silverstein

Work done:

Anish Timalsina: Course Project Proposal Report (abstract, Introduction, Related work, Evaluation, Data sets, references)

Gathering dataset

Adam Prieto: Data Analysis, manipulation, preprocessing, visualization and graphics, ethical considerations.

Brock Hoos:

Scott Silverstein: