Project proposal
Group Members: Zijia Tang (zt81), Peilin He (ph155), Xingyu Bai (xb30), Carter Davis (cjd94)
Date: Friday, Feb. 22nd, 2025
CS 216: Everything Data

Part 1: Introduction and Research Questions

Air pollution remains a major environmental and public health challenge in the U.S., shaped by policy interventions, urbanization, population growth, and natural events like wildfires. While environmental policies aim to reduce pollution, their effectiveness across regions remains uncertain. Meanwhile, urban expansion and extreme weather continue to influence air quality, making it essential to understand these factors for better policy decisions.

To systematically analyze air pollution trends and their driving factors, we aim to address the following key research questions:

1. How has air pollution changed in major U.S. cities since 1980, and what factors contribute to these trends?
2. To what extent do population density and urbanization correlate with variations in air pollution levels?
3. How have major environmental policies influenced air pollution at the state and city levels?
4. What is the impact of wildfires on air pollution levels in different regions, and how do weather conditions modulate this effect?

This research is **substantial** because it goes beyond basic statistical analysis, employing machine learning models such as XGBoost and causal inference techniques to uncover nonlinear relationships and potential threshold effects in air pollution trends. By integrating spatiotemporal analysis and model interpretation, we seek to provide a deeper understanding of the dynamic correlation between urbanization, policy interventions, and air quality. Additionally, our approach is **feasible** within a six-week timeframe, leveraging well-documented datasets from the EPA, NASA, and NOAA, alongside existing computational tools such as Python's sklearn and shap libraries. This research is particularly **relevant** in the context of increasing urbanization, climate change, and policy debates on air quality regulation. As governments worldwide seek to implement data-driven policies to mitigate pollution, our findings will provide actionable insights for sustainable urban planning and environmental governance. Additionally, by assessing policy effectiveness at regional levels, our study can help address disparities in air quality improvements across different states, ensuring equitable and targeted interventions.

Part 2: Data Sources

We plan to use the following datasets for this project:

1. EPA Air Quality Data

- **Source**: EPA Outdoor Air Quality Data (EPA AQS Air Data)

- **Description**: This dataset includes daily measurements of pollutants such as PM2.5, NO_2, and SO_2 across U.S. cities, providing a basis for evaluating long-term air quality trends and assessing the effectiveness of policy interventions.

2. Urbanization Data

- **Source**: World Bank Urbanization Data
- **Description**: This dataset records annual urban population percentages, enabling an analysis of the relationship between urbanization and air pollution levels over time.

3. Environmental Policy Data

We will examine major U.S. environmental policies and their impact on air pollution, using official documents and research reports. These include:

- **1990 Clean Air Act Amendments**
  - Legislation | EPA Overview
  - **Effect**: Reduced SO_2 emissions by 88% and NO_x by 50% (1990–2020); PM2.5 levels dropped by 41% (2000–2020).
- **1997 National Ambient Air Quality Standards (NAAQS) for PM2.5**
  - EPA Report
  - **Effect**: PM2.5 concentrations declined by 39% in major U.S. cities by 2019.
- **2005 Clean Air Interstate Rule (CAIR)**
  - Federal Register
  - **Effect**: SO_2 emissions from power plants in eastern U.S. states fell by 68% by 2015.
- **2015 Clean Power Plan**
  - **Effect**: Though paused by the Supreme Court in 2016, the policy accelerated coal plant retirements, leading to a 28% decline in coal-generated emissions by 2020.
- **California's AB 1493 (2002) and Zero-Emission Vehicle (ZEV) Mandates**
  - California Air Resources Board
  - **Effect**: Catalyzed nationwide EV adoption; ZEV states experienced a 15% faster reduction in NO_2 levels than non-ZEV states.

4. Weather and Wildfire Data

- **Source**: NOAA Integrated Surface Database (ISD)
- **Description**: This dataset includes historical hourly weather data (temperature, wind speed, and precipitation) for major U.S. cities since 1980.
- **Relevance**: By merging ISD data with EPA pollution metrics, we can isolate the impact of wildfires and extreme weather events on air pollution trends.

These datasets provide a comprehensive foundation for evaluating the interplay between environmental policy, urbanization, and pollution. With structured CSV formats and robust historical coverage, they enable effective data integration and machine learning analysis to uncover meaningful trends and policy implications.

Part 3: What Modules Are You Using?

Module 3: Visualization

We will use this module to create visual representations of our findings to better understand trends in air pollution, urbanization, and policy effectiveness. We will do this by generating time trend plots using Matplotlib to track changes in PM2.5 and NO_2 levels from 2000 to 2023, correlating them with major policy interventions like the Clean Air Act amendments. Our **justification** is that clear and interpretable visualizations are essential for identifying patterns and communicating key insights effectively. The **concepts** we will primarily use include geospatial heatmaps with Geopandas to analyze pollution disparities across different states, as well as scatter plots and correlation heatmaps with Seaborn to explore relationships between urbanization, wildfires, and pollution levels. We plan to utilize this module throughout the analysis and final report **stage**.

Module 4: Data Wrangling

We will use this module in the early phases of our project to clean, structure, and prepare datasets for analysis. We will do this by handling missing data in EPA air quality records through interpolation techniques and applying regex-based text processing to extract key policy information, such as enactment dates and regulatory terms. Our **justification** is that consistent and well-structured data is critical for ensuring accurate and reliable analysis. The **concepts** we will focus on include time synchronization of NOAA weather timestamps with pollution data and standardizing variables across multiple sources. We will use this module during the data preparation **stage**.

Module 6: Combining Data

We will use this module to merge multiple datasets to investigate how different environmental and policy factors contribute to air pollution trends. We will do this by using Pandas' merge function to integrate EPA pollution data with NOAA weather variables like temperature and wind speed at city and date levels, allowing us to analyze meteorological influences on pollution fluctuations. Our **justification** is that combining data from multiple sources enables a comprehensive analysis of air quality dynamics. The **concepts** we will apply include integrating urbanization data from the World Bank with NASA FIRMS wildfire records using geographic coordinates and incorporating policy enactment years as dummy variables to assess regulation impacts. We will use this module throughout the data integration **stage**.

Module 7: Statistical Inference

We will use this module to analyze the statistical relationships between air pollution, policy interventions, and environmental factors. We will do this by applying hypothesis testing, such as t-tests, to compare pollution levels before and after regulatory changes like the implementation of the 1997 NAAQS for PM2.5. Our **justification** is that statistical inference techniques allow us to validate findings and ensure the robustness of our conclusions. The **concepts** we will use include bootstrapping methods to assess the reliability of correlations between wildfires and pollution spikes, as well as regression models to determine the significance of urbanization and population density on air quality trends. We will incorporate this module in the statistical analysis **stage**.