Capstone 1 Proposal

Ryan Blauwaert

Motivation

We know that air pollution has negative effects on public health, but it can be difficult to determine the magnitude of these effects. Damage from long term exposure to poor air quality can take decades to manifest in an otherwise healthy individual. Because of this, it is common to look at the data from vulnerable populations (infants and children) when studying environmental health effects. In this analysis, I will look at various air quality metrics to determine trends in overall air quality across the U.S., along with infant mortality rates and childhood asthma data to see if any relationship can be identified.

The Data

For this project, I will be using data from two main sources: the Environmental Protection Agency (EPA), and the Kids Count Data Center (KCDC). The EPA provides annual reports containing air quality statistics for nearly 500 U.S. cities. These reports each contain fourteen different measures of air quality including carbon monoxide, ozone, and sulfur dioxide concentrations. Additionally, the EPA publishes aggregate air quality measures which can be used to more easily assess air pollution trends. These reports span from 1980 to 2020.

The KCDC publishes annual infant mortality statistics broken down by major U.S. city. These data span from 1990 to 2020 and will primarily be used to assess the magnitude of health effects resulting from diminished air quality. Additionally, the Citizen's Committee for Children of New York provides data on the rate of asthma-related hospitalizations from 2000 to 2016. This data may be used to compare trends in asthma prevalence with air quality changes in New York.

MVP

- Use pandas to clean and manipulate data as well as merge and group datasets from different sources.
- Practice using matplotlib and seaborn libraries to create informative visual representations of the data.
- Provide clear and succinct analysis of air quality trends in major U.S. cities.

MVP+

- Break down air quality data into different components to try to determine which pollutants have the most significant effect on health outcomes.

- Create reusable data cleaning scripts which can be used for analysis of future data.

MVP++

- Conduct hypothesis testing to determine if there are any significant effects on infant mortality or asthma prevalence after local/state environmental policies are put in place.