Report: Air Pollution in the United States of America

Data Selection

The data of various years, spaced over two decades, was selected to provide an overall perspective on the changes in pollution levels.

Five data sets, of years 2000, 2004, 2008, 2012 and 2016 were chosen.

Pollutant Selection

The following five pollutants were selected, widely accepted as **"criteria pollutants"** by the EPA.

Ozone, PM2.5, Carbon Monoxide, Lead, Sulphur Dioxide

Data Fields

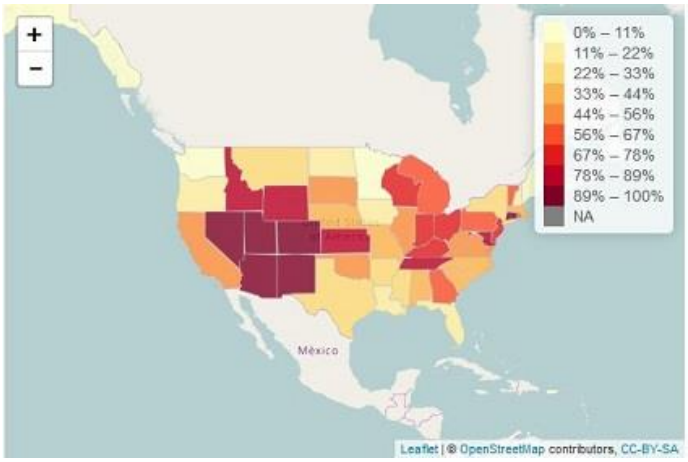
The following data fields were selected for the corresponding reasons:

- Latitude/Longitude - They provide spatial data which is easy to understand, and mappable as well.
- Parameter: Required to identify which pollutant is present in the area.
- Parameter Code: Used to uniquely identify parameter names in the data.
- Metric Used: Gives an idea of how big or small a reading is compared to the others.
- Arithmetic Mean: The mean of all readings allow us to have a single representative reading for a particular pollutant in a particular area. This does not deal with outlier cases though.
- Arithmetic Standard Deviation: Mean brings in outlier bias; hence we use Standard Deviation to deal with that.
- State Name: States are first level in the visualization of the whole map, before further increase in granularity.
- County Name: The next level of granularity, also required for better visualization.
- City Name: Further granularity, this gives the highest granularity for the visualization that we can offer

Visualizations

1. State wise Comparison

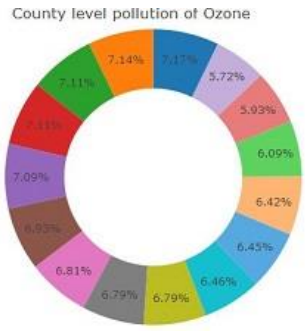
- **Description** - A Choropleth Map providing a top level view of the entire country
- **Purpose** - To visualize state wise comparisons for particular criteria pollutants.
- **User Inputs** - Year and Pollutant
- **How to read the map** - Red gradient. Deeper shades represent higher degrees of the particular pollutant's presence.
- **Analysis** - The presence of PM2.5 in coastal regions is high. PM2.5 concentration in Hawaii is very high. The reason we found for this, is that a high number of ships travel there.



---

2. County Contribution

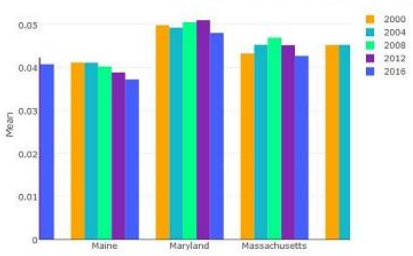
- **Description** - Donut chart to visualize how much a county in a state contributes to a particular pollutant.
- **Purpose** - To identify which counties are areas of high pollution in a particular county.
- **User Inputs** - State, Year and Parameter.
- **How to read the map** - Hover on a chunk of the donut chart to find which county contributes how much to a particular pollutant.



---

3. Year wise Change in Pollutant Concentration

- **Description** - A bar chart to describe how the concentration of a particular pollutant has changed in every state over a gap of 4 years (from 2000 to 2016).
- **Purpose** - To identify trends in changing pollutant levels in a particular state over a long time.
- **User Inputs** - Pollutant
- **How to read the chart** - Select the state of interest and hover over it to see the values.
- **Analysis** - On the South East Coast in 2000, Ozone levels were high. Following a lawsuit in the South East, the ozone levels in the regions were reduced significantly over the years. Over the years, the levels of PM2.5 has consistently reduced across almost all states.



Data Peculiarities

- ### FIPS Codes Problems with the Data

  As we lacked domain specific knowledge at the beginning of the assignment, we found it difficult to map the states precisely in the visualization. At this point, we came to know that even if latitudinal and longitudinal boundaries change, the FIPS code remains constant, and hence is a better way to map the counties for the required visualization. Also, the dataset did not contain two-letter abbreviations for the states, which made the FIPS code more appropriate for our purpose.

- ### Lead

  Data related to lead for some states is sparse.

- ### Counties

  The number of counties given in the dataset are rather small compared to the total number of counties in the USA. Data for 800 counties is available in the dataset, while there exist more than 3000 counties.

Libraries and Tools used

Shiny, Maptools, Rgeos, Leaflet, Maps, Plotly

Scalability Issues in R

Due to the single threaded nature of R, when we are running the application using Shiny, no other processes are allowed to run. Also, as all the data being processed needs to be fit into RAM,it becomes very difficult to analyse data as the size of datasets become exceedingly greater, which makes the system not scalable.