CS 3300 Project 2 Writeup

Mapping Pollution Patterns Across U.S. Cities

For this project, we utilized two primary datasets:

- U.S. Cities Data (uscities.csv): This dataset includes comprehensive information about cities in the United States. Key variables include city names, state IDs, latitude, longitude, and population. These variables are necessary for us to map the locations.

Attribute	**Description**
city	Name of the city.
city_ascii	ASCII-compatible name of the city.
state_id	Two-letter abbreviation of the state.
state_name	Full name of the state.
county_fips	FIPS code for the county.
county_name	Name of the county.
lat	Latitude of the city.
lng	Longitude of the city.
population	Population of the city.
density	Population density per square kilometer.
timezone	Timezone of the city.
zips	ZIP codes associated with the city.
id	Unique identifier for the city.

- Air Quality and Water Pollution Data (cities_air_quality_water_pollution.csv): This dataset provides air quality and water pollution indexes for various cities globally. For our project, we focused on cities within the United States, trying to create mappings from uscities.csv.

Attribute	**Description**
City	Name of the city.
Region	Region or state in which the city is located.
Country	Country of the city.
AirQuality	Air quality index of the city.
WaterPollution	Water pollution index of the city.

Both datasets were sourced from public data repositories. The U.S. Cities data was obtained from SimpleMaps, and the pollution data was sourced from Kaggle.

For our data preparation, we went through a few steps:

- Filtering: We filtered the pollution data to include only cities from the United States.
- Matching and Merging: We matched cities from both datasets based on city names and integrated them, ensuring the same format when working with both datasets.
- Subset Selection: We selected a subset of cities with a population greater than 10,000 for better clarity in the visualization of our data spread across the country.
- Data Cleaning: Removed and normalized the formats of the data so we can parse them easily.
- Coherence: We want higher pollution values to mean worse pollution. For air quality, 100 represents no pollution and vice versa, so we used (100 - air quality index) instead. This step is not needed for water pollution because 100 means very polluted water.

b) Visual Design Rationale

Our visualization aims to present a clear and intuitive understanding of air and water pollution levels across various U.S. cities. We employed the following design decisions:

- Marks and Channels:
    - Cities are represented as circles (marks) on a U.S. map.
    - Circle size is proportional to the city population.

        - Circle color indicates pollution levels, using a sequential color scale where darker shades represent higher pollution levels with different color scales for Air Pollution and Water Pollution.

- Color Scales:
    - d3.interpolateReds for air pollution.
    - d3.interpolateBlues for water pollution.
        - These color choices are intuitive (red often signifies danger or caution, blue is associated with water) and aid quick comprehension.
    - d3.interpolatePurples for population.
- Transformations:
    - Used d3.geoAlbersUsa projection for mapping geographic coordinates to the SVG canvas.

c) Interactive Elements Design

Our visualization includes interactive elements to enhance user engagement and exploration:

- Toggle Buttons:
    - Users can switch between air pollution and water pollution visualizations.
    - This interactivity allows users to compare different types of pollution across cities.
- Mouseover Tooltip:
    - Hovering over a city circle displays the city name allowing the removal of clutter and letting the circles shine as the main data visualization.
    - In order for the audience to identify the city more clearly when they hover on a circle, the currently selected circle will remain in the same opacity while the other circles will have their opacities reduced

d) The Story and Insights

This project's objective was to examine the relationship between pollution levels and population sizes in U.S. cities. Initial efforts involved using a population-based slider to explore whether increases in population corresponded with heightened pollution levels. However, this approach did not reveal a distinct correlation between these two factors. Instead, a more pronounced association was observed between pollution levels and geographic locations.

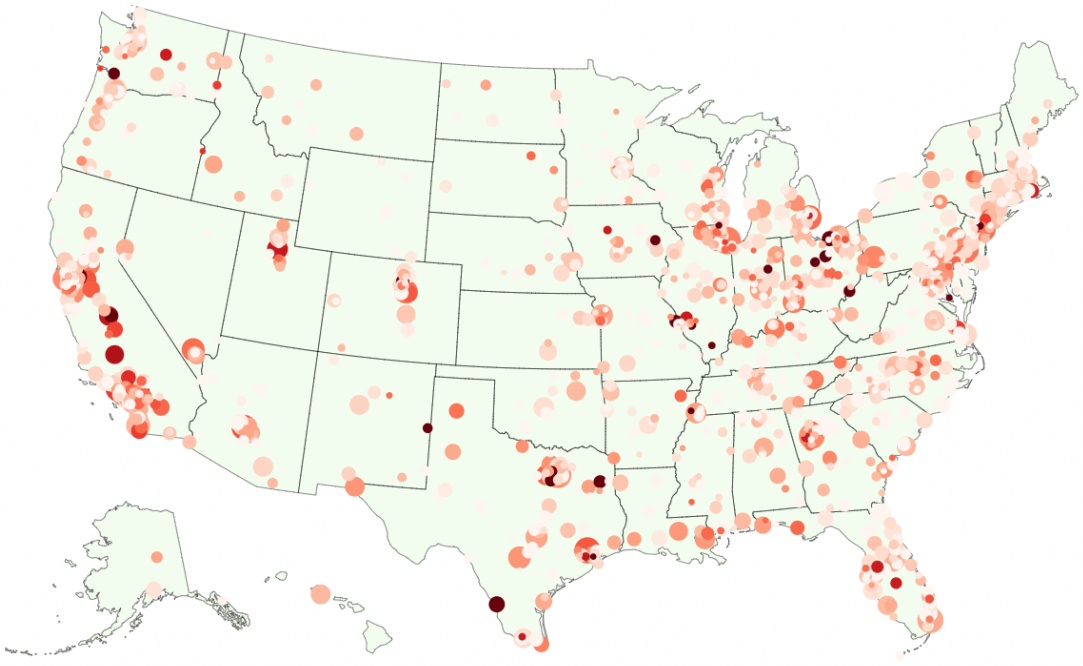
To delve deeper into this observation, we shifted focus, replacing the slider with three distinct buttons to separately visualize total pollution (a combination of air and water pollution), air pollution alone, and water pollution alone. This change in approach allowed for a more nuanced exploration of pollution patterns.

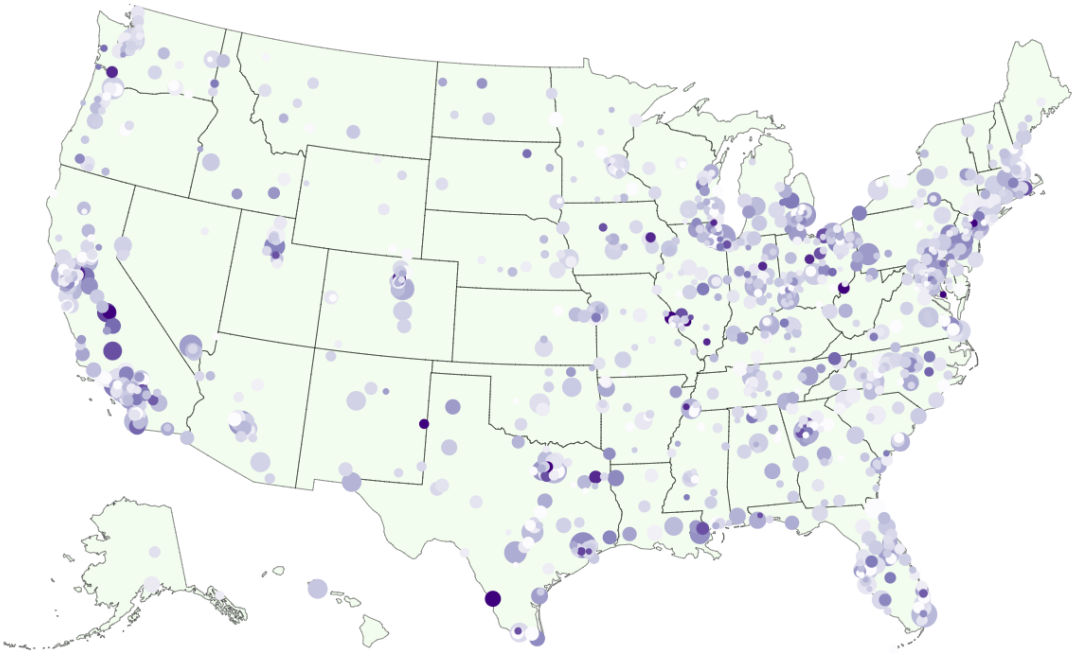
Key Findings:

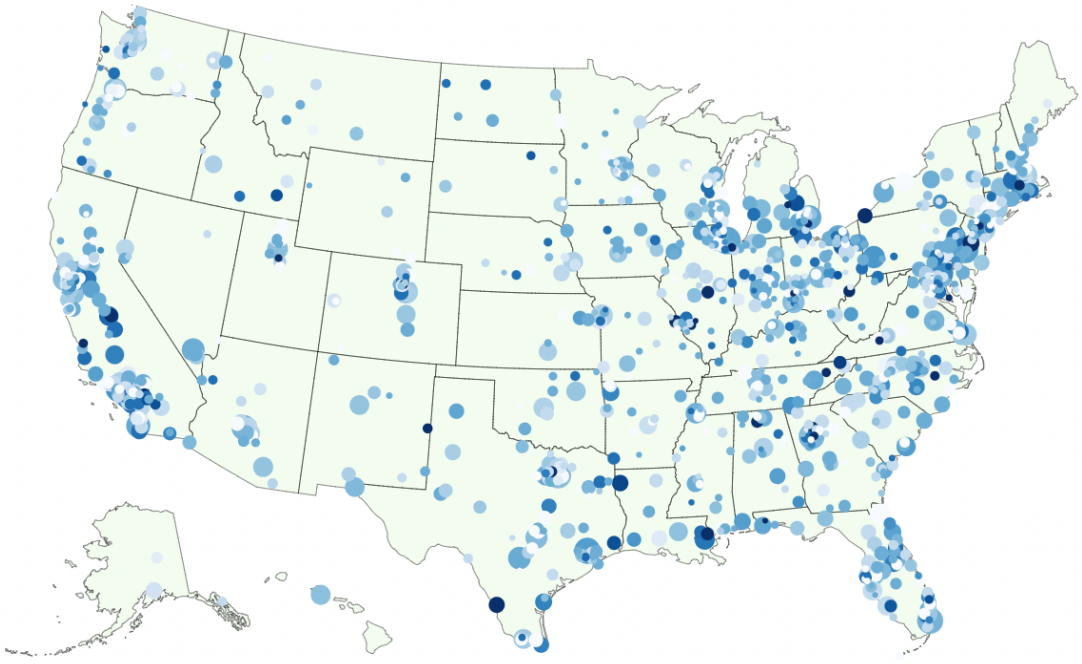
- **Geographic Correlation Over Population Size:** The analysis revealed that pollution distribution is more strongly correlated with geographic location rather than population size. Notably, higher levels of both air and water pollution were found along coastlines, particularly along the California coast and the upper east coast of the United States.

- **Unexpected Lack of Correlation with Population**: Contrary to the initial hypothesis, there was no clear indication that larger populations directly contribute to higher pollution levels. This finding challenges common perceptions and suggests the influence of other factors.
- **Potential Explanatory Factors:** The lack of a direct correlation between population size and pollution levels could be attributed to various reasons:
    - Stricter Environmental Regulations: Larger populations might prompt tighter government policies and regulations aimed at controlling pollution.
    - Economic Structure Shifts: Densely populated areas may experience a shift away from heavy industries, known for higher pollution levels, towards service-based or technology-driven industries.
    - Infrastructure and Planning: Urban areas with larger populations often have more developed infrastructure and urban planning, which could include effective waste management systems and pollution control measures.
    - Regional Variations: The distinct patterns observed along coastlines suggest regional variations in pollution sources. These could be linked to factors such as industrial activities, shipping, tourism, and natural features influencing pollution dispersion.
- **Limitations and Further Research:** While the study provides valuable insights, it also highlights the need for further research to understand the complex dynamics between urban population, industrial activity, geographical factors, and pollution levels.

From this, we were able to visualize three different visualizations of our map:







e) Work Distribution

- Jenny: preprocessed the datasets, combined information from the two files to be used together, created scales for the three pollution categories
- Jessie: Created map & upload dataset, part of the interactive mouseover tooltip, change of opacities during the interactive process
- Peter: Created the buttons to toggle between the visualizations and implemented part of the mouseover feature where it displays the pollution level of the city
- David: Worked on the design structure and implementation, finalization, and the writeup.