United States Pollution Data Visualization

Arivan Thillaikumaran | a.thillaikumaran@wustl.edu

John Henroid | jhenroid@gmail.com

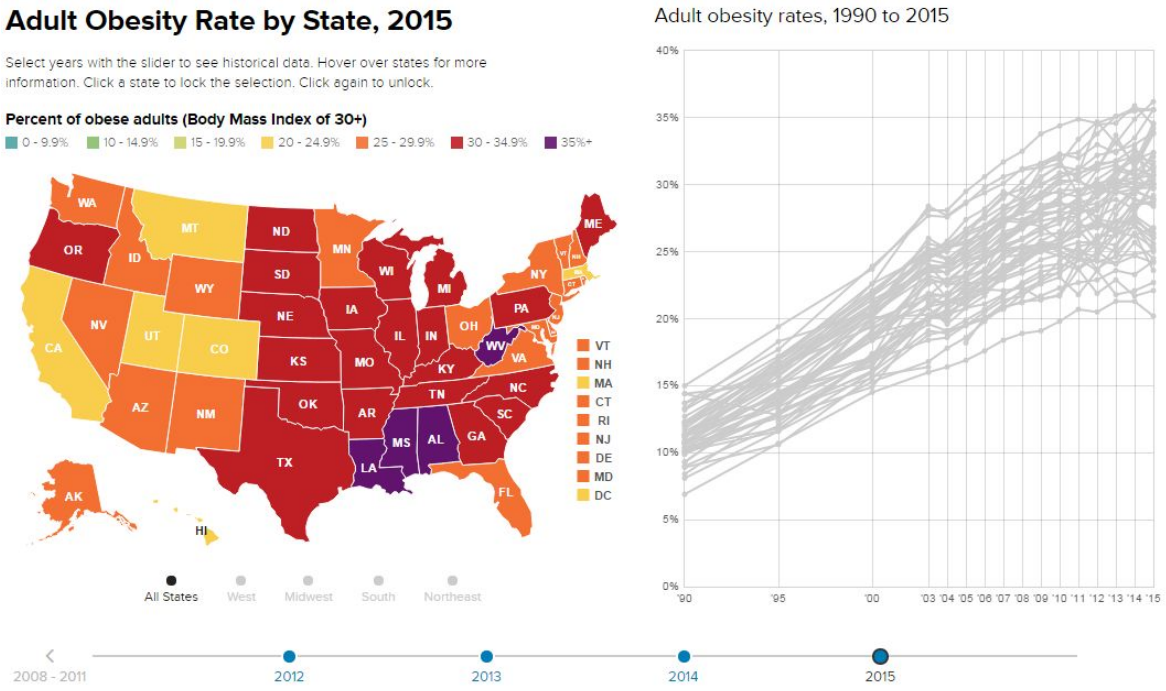
Overview and Motivation:

Pollution is one of the top contributors to climate change. Seeing visualizations of the pollution data may lead to key insights and may help people understand the gravity of the problem easier. Key things we want to identify are the states that contribute the most to overall US pollution, and states that have made the most progress towards decreasing their contribution to pollution.

Also, with our new President elect appointing Myron Ebell and , a climate change skeptic, to lead the transition team for the Environmental Protection Agency, it only seems even more important that people realize the scope of this issue.

Related Work

We were considered initially doing a US state map with graphs of data found from the state of obesity website. It shows a range of data over multiple time series in a clean visualization.



Questions

Which states are guilty of contributing the most to climate change?
Which states have made the most progress in decreasing their yearly pollutants?

Initially we only considered tracking the amount of pollutants measured for each state. We have a large dataset and noticed there are other interesting aspects like Air Quality Indices (AQI) that are being tracked. We decided to use AQI initially for our visualization.

Data

We have pulled our data from https://www.kaggle.com/sogun3/uspollution

The raw data set contains various pollution metrics for different addresses across the country, with measurements taken a few times a day over the span of 16 years. Thus, the raw dataset that we pulled from this website contained over 1.7 million entries. We have used a python script to simplify this data, taking the average of all the entries per year per state. However, we noticed that this dataset only contains a certain range of years for some states; for most of the states we have the data from 2000-2016, but for some we have a smaller range. Initially we chose to fill in the missing years' data with -1's, but this yielded inaccurate line graphs. We then decided to fill in the missing years' data with the average of the years that we did have for the state. There are 6 states for which we do have data; they are marked on our visualization and cannot be interacted with. We decided to take the averages for all months and insert into years that didn't have data.

Exploratory Data Analysis:

We initially started looking at data with Excel and doing to rough bar graphs. We realized quickly that showing all 50 states in a clear way was going to be a challenge. We knew that we had to implement a visualization that changed with user input to clearly represent all different aspects of our dataset.

Implementation

We use the D3 Javascript library to build our visualization. Currently we are working on just the map visualization and color gradients based on our data. We took a sample set for Air Quality Index (AQI) and for each pollutant to see some changes across different states. Some of the states don't have any data and were removed. Those states are appearing in black our final visualization.

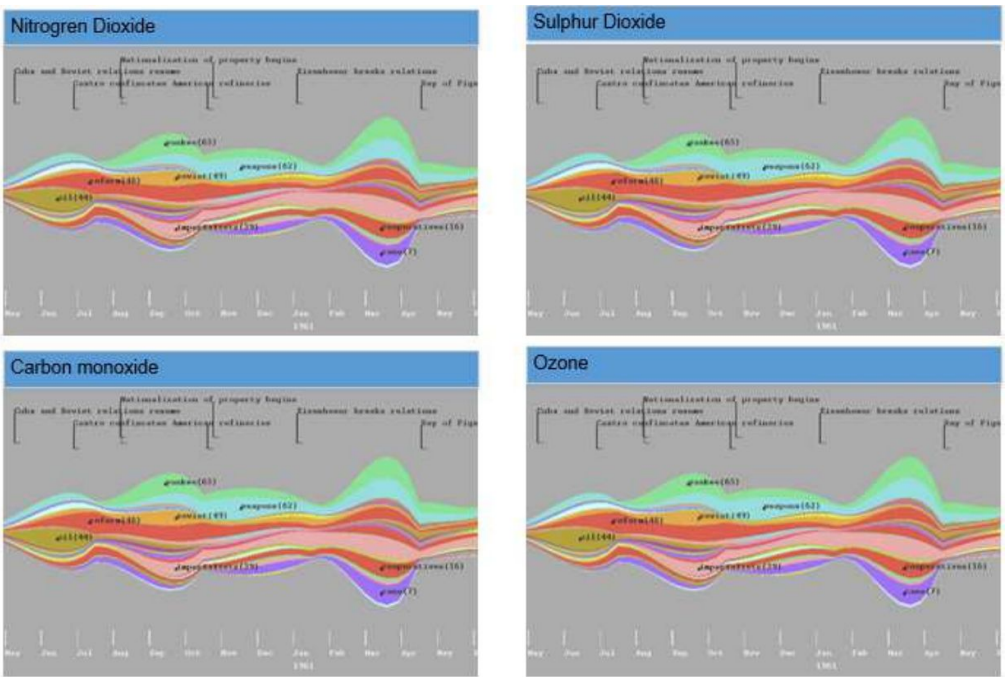
We used a D3 slider control to be able see changes from year to year for a specific pollutant. The user has the ability to look at each pollutant and see a changes over time. If a user were to click on a state then graphs for all pollutants and AQI over time are shown to the right to show an overall view of how well or poorly a state performed.

Design Evolution

We considered many different ways to evaluate and visualize our dataset. We considered bar charts that represent each state's pollution level, but realized that looking at 50 different bars is neither visually appealing nor an efficient way to represent our data. We decided that since the most important aspect of our data is the geographical region, we would represent each states' information using a map visualization. This would turn out to be an effective way of comparing different states pollution levels, and users would be able to use the map to call upon information about that specific state.

We then needed to decide how to present the pollution information for each state. We initially decided to use a time series chart that would show the state's pollution metrics across a 16 year period, but then realized that our dataset is somewhat incomplete and there are a few states for which we do not have the data for all 16 years, rendering a time series chart useless for those states.

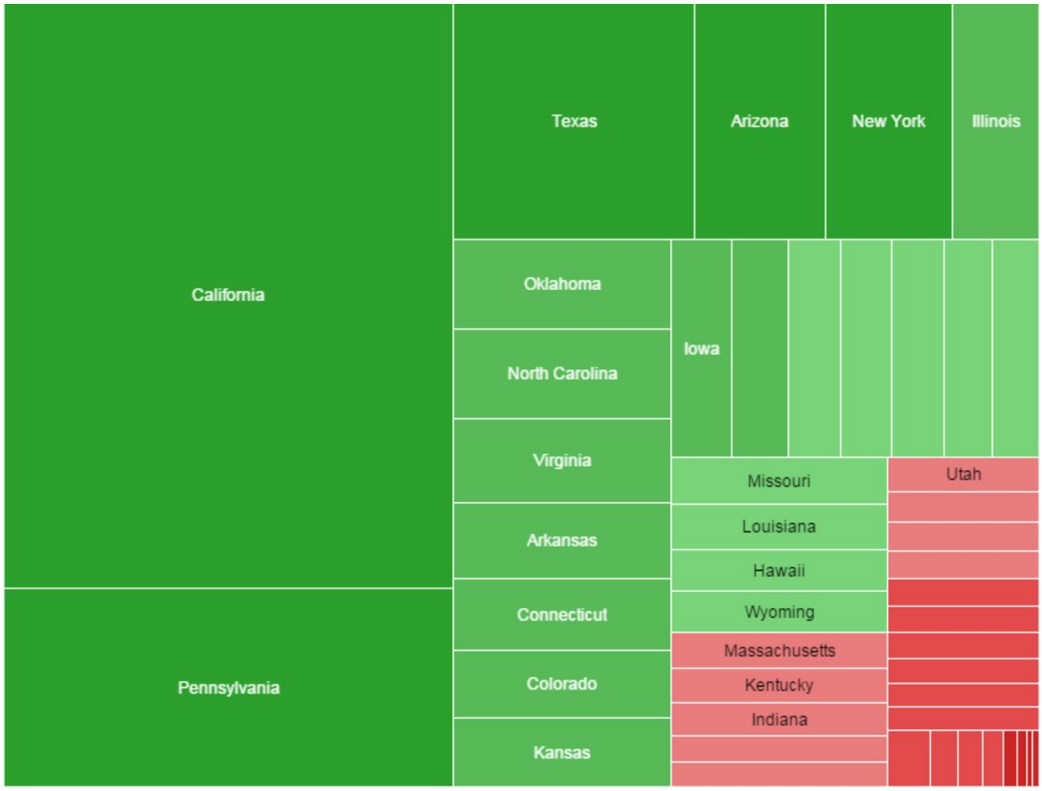
We have considered representing the data in a time series chart by state:



Since we realized that our dataset didn't have some information for different states we realized this would be a bad choice. The main factor for a themeriver is to have time series data and it

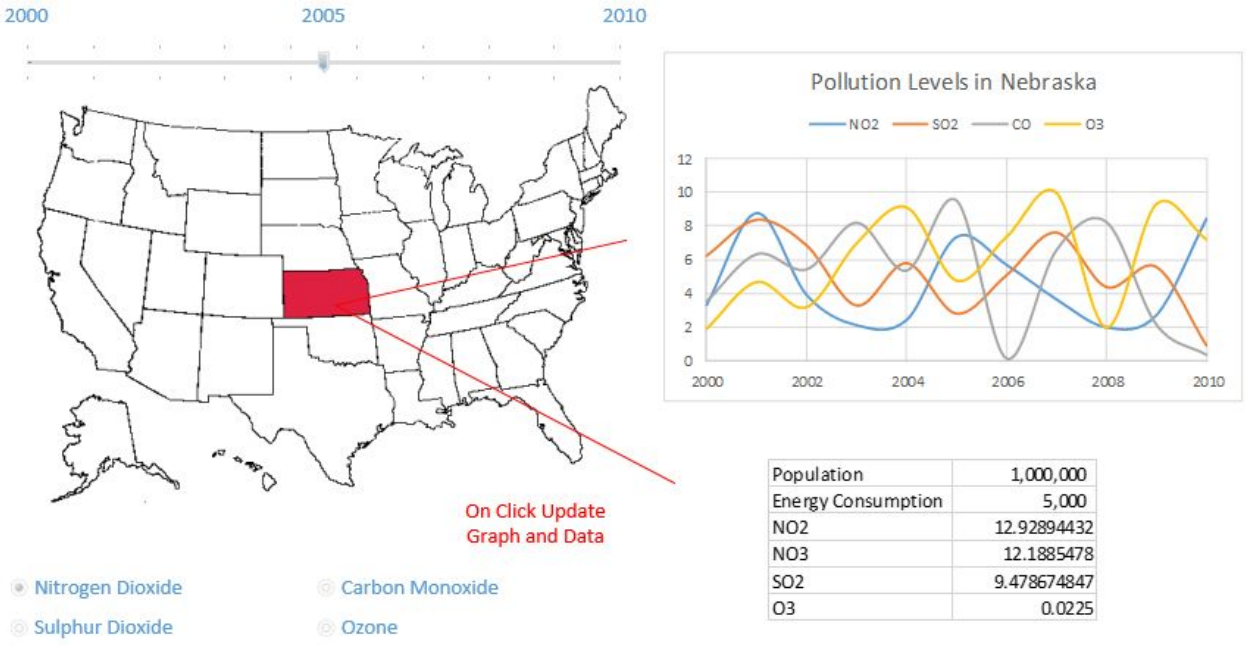
wouldn't be a good representation if we were lacking that information. Missing information couldn't be shown and it would be cluttered with all the states.

And via a treemap:

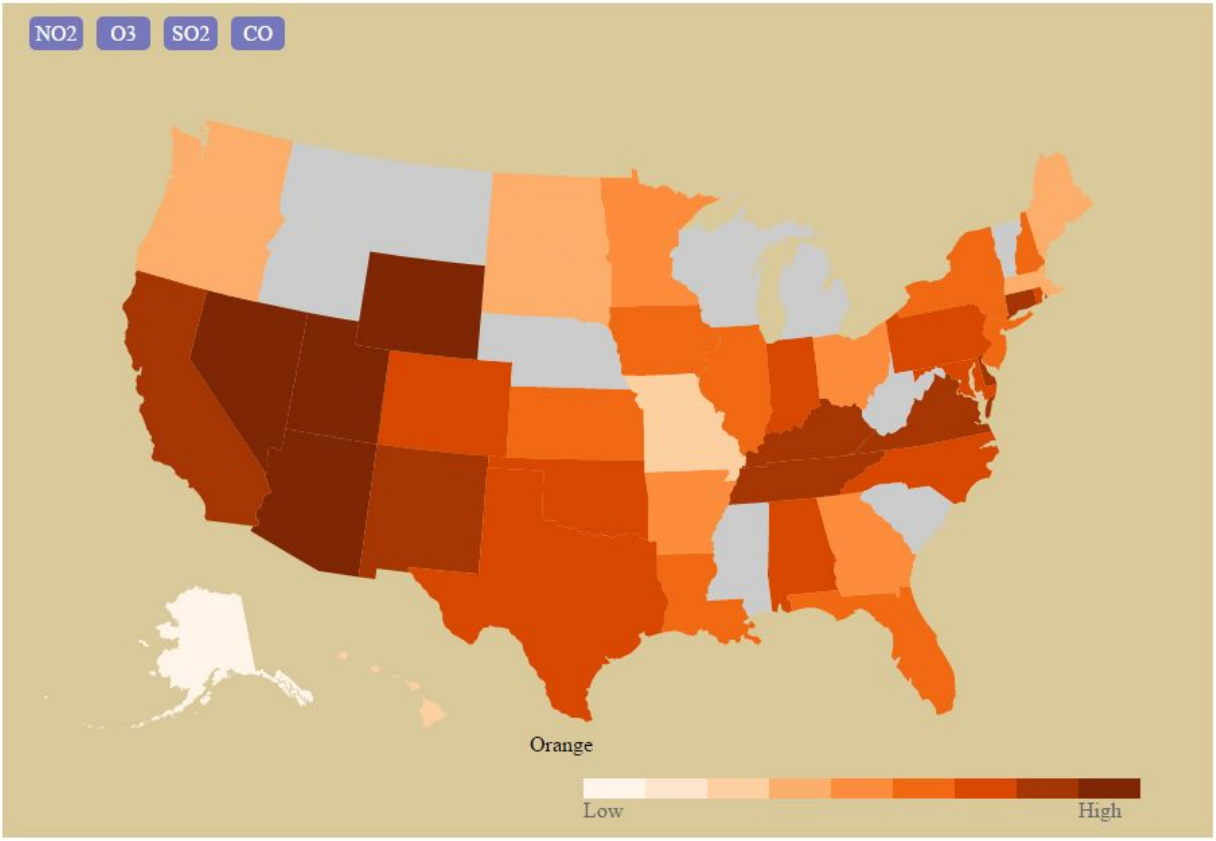


The treemap is still a viable implementation. We felt that the US map was a better representation because it gives us feedback on pollutant levels related to geography.

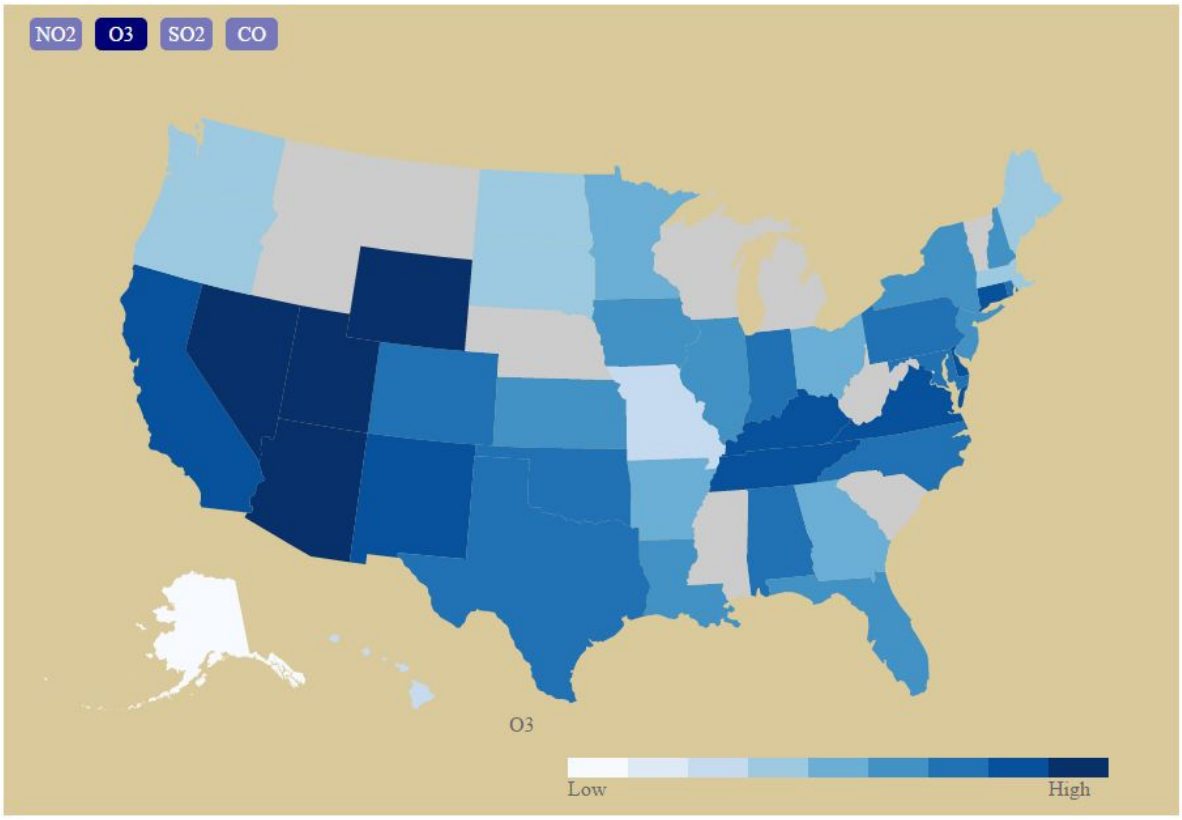
Since we are concerned about the states that have the most and least pollution we are concerned about adjacent states. We feel that we need to represent pollution in a geographical form with the state map and settled on a design shown below.



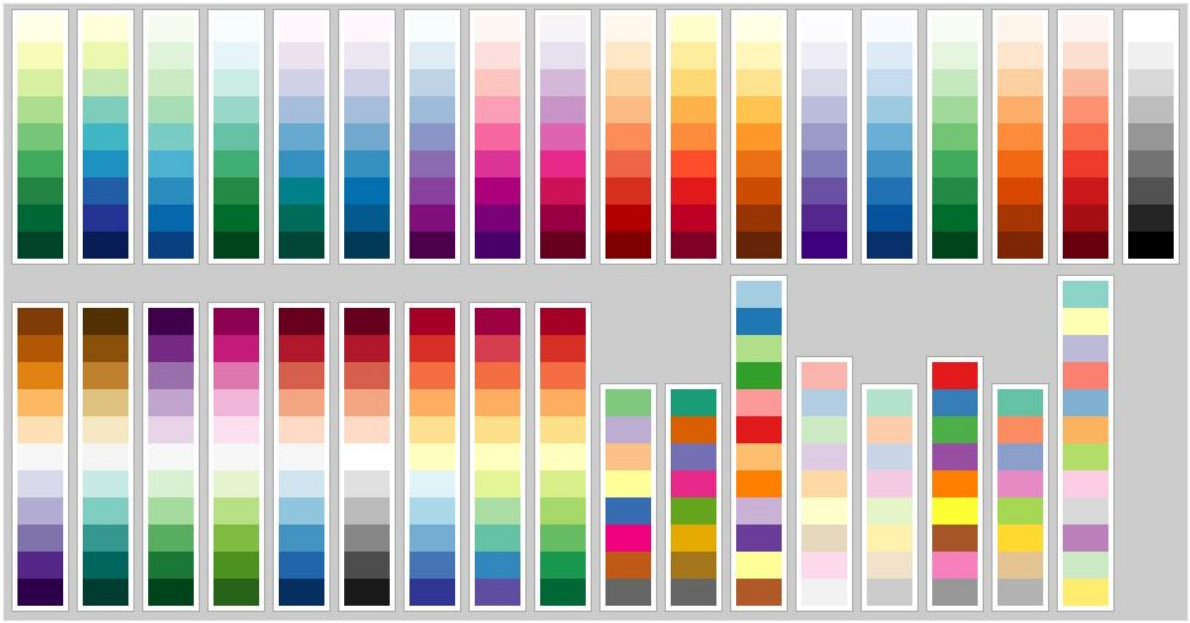
Population	1,000,000
Energy Consumption	5,000
NO2	12.92894432
NO3	12.1885478
SO2	9.478674847
O3	0.0225

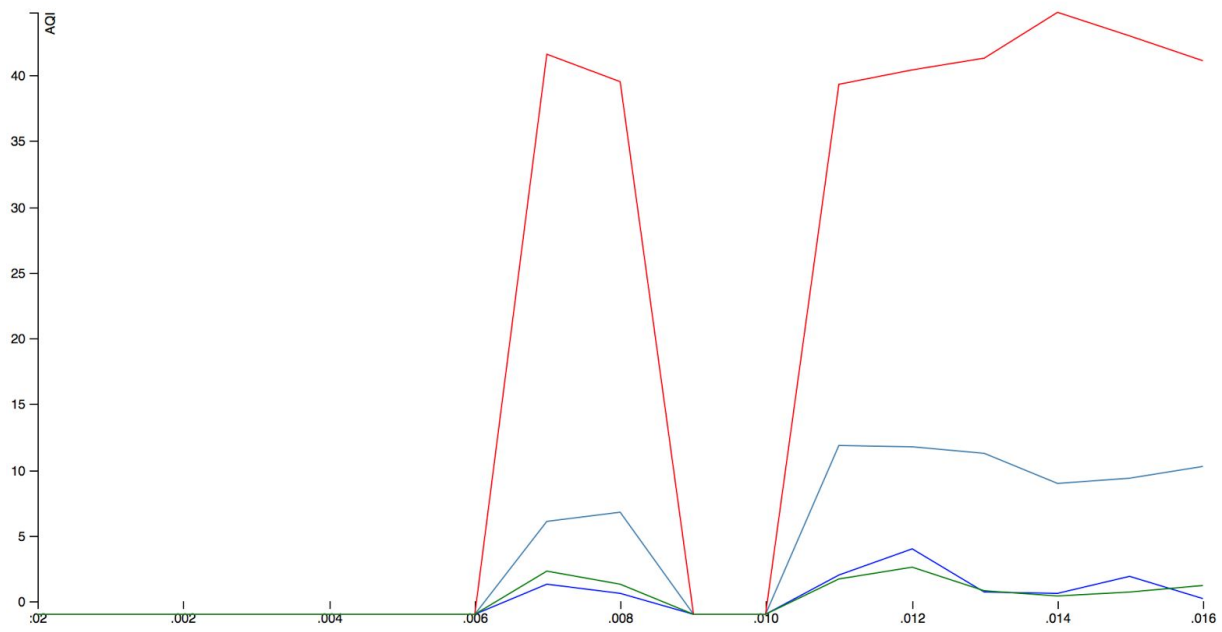


We also have button clicks working to switch between two different sets. A button click will change the color scale gradient as shown in the figure below.

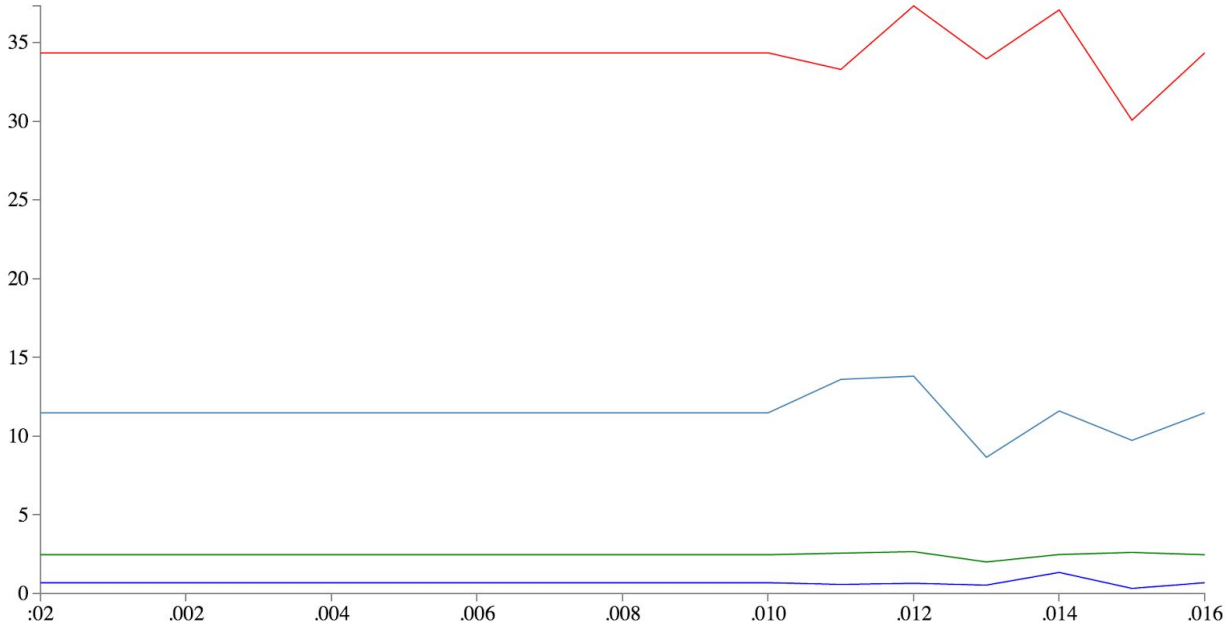


We started using the [built in color brewer] for our gradients. The low end of all of the single colors has very light to almost white colors shown in the figure. We realized that we can't use a white background for the states at the low end of data which forced us to use a non white background.





Our initial line graphs looked like the one above. We chose to fill in the missing values in our dataset with -1's. This worked with the map representation, as we were able to add a simple condition that colored the state a unique color showing the data was missing if it was -1, but as can be seen for the line graph, the imputation causes a misrepresentation of the data. In the example above, it looks like there is a sharp increase from 2006 to 2007. However this is not the case, and there is only data from 2007 onwards.



We decided to instead impute the missing values with the average of the data that we did have. This is what the line graphs looked like after filling in the missing blanks with the average of each state's provided data.

Evaluation

So far, in our implementation we can see some changes between different states, and can see that some states contribute more of one pollutant than the other. There are some issues with our scales as we move from one pollutant to another which we need to fix. We also need to fix highlighting the selected pollutant.

We are in the process of adding charts that can help us depict the data over time. We learned that we need to find a good representation of incomplete data for our visualization in both the map and time series data. We realized that the user can explore the data themselves and determine which are the most polluted states and which states have made the most improvement but that will require them to navigate our visualization. We were thinking that perhaps we should have a summary table that answers our pollution questions but also allows the user to explore the visualization to see the time series data.

We believe our graph conveys our data very well. Because one of the primary attributes of our data is location, users will be able to relate to geographical location when using the visualization, potentially allowing them to apply other information that they may associate with a map when exploring the visualization. The color scheme of the map helps states that are

outliers stand out to the user, and that meets our goal of the user immediate takeaway to be aware that there are states that doing poorly in contributing to the fight against climate change. We believe that users will have some states in mind before really seeing the visualization that they would be interested in, and our visualization very clearly lays out the history and current direction the state is headed in. Users can easily see where on the spectrum the particular state they are studying is. The time of the data point is also an important attribute, and the line charts are the best way to show the trend over time. Users have two perspectives at seeing the measured data, with the average parts per million (or billion in NO2's case) or the Air Quality Index.

We thought we could improve our visualization by having a fifth setting to see the average pollution across all the different pollutants. Currently you can see the change in each particular gas or compare states based on each particular gas, but some users may be only interested in the overall contribution.

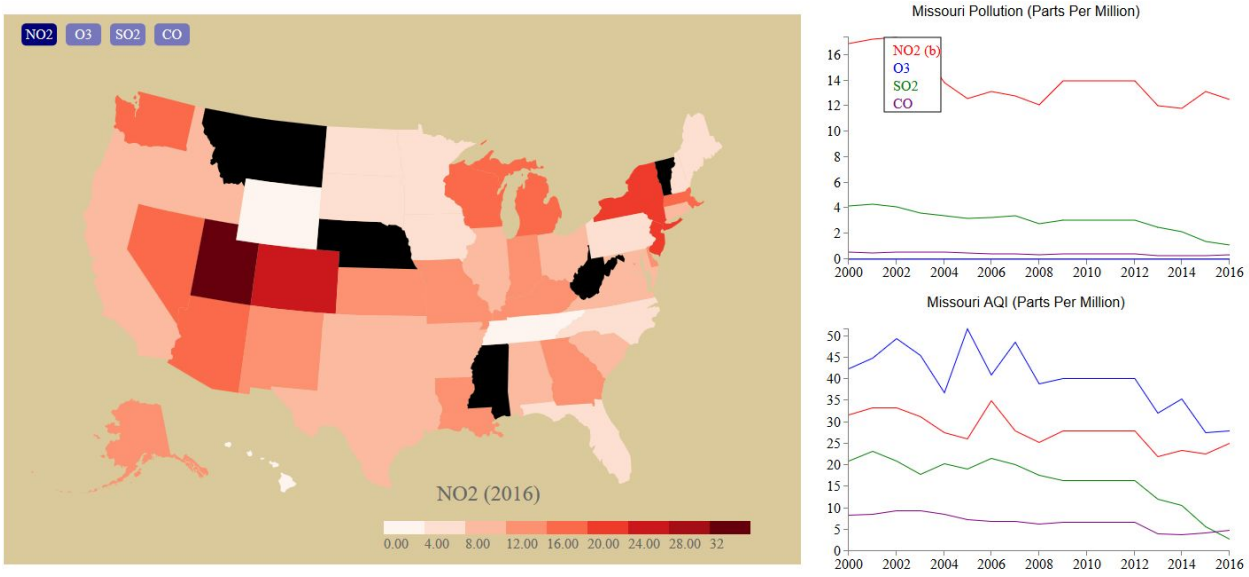
If we wanted to add more information to the visualization, we could list the top two or three industries of each state, which would give more meaning to the analysis of the pollutants separately to the average reader. For example, we can see that the Pacific Northwest consistently has the highest air quality index, and we can show that the main industries are ones that do not contribute heavily to pollution.

States are very generally and slowly towards progress, but for the most part fluctuate between regressing and improving, and overall we see no meaningful trends over the 16 years. However, one can clearly see the history of a particular state they are interested in, and still see the current trend of that state, which still fulfills our goal of wanting users to be aware climate change and the impact that our present day societies have on it.

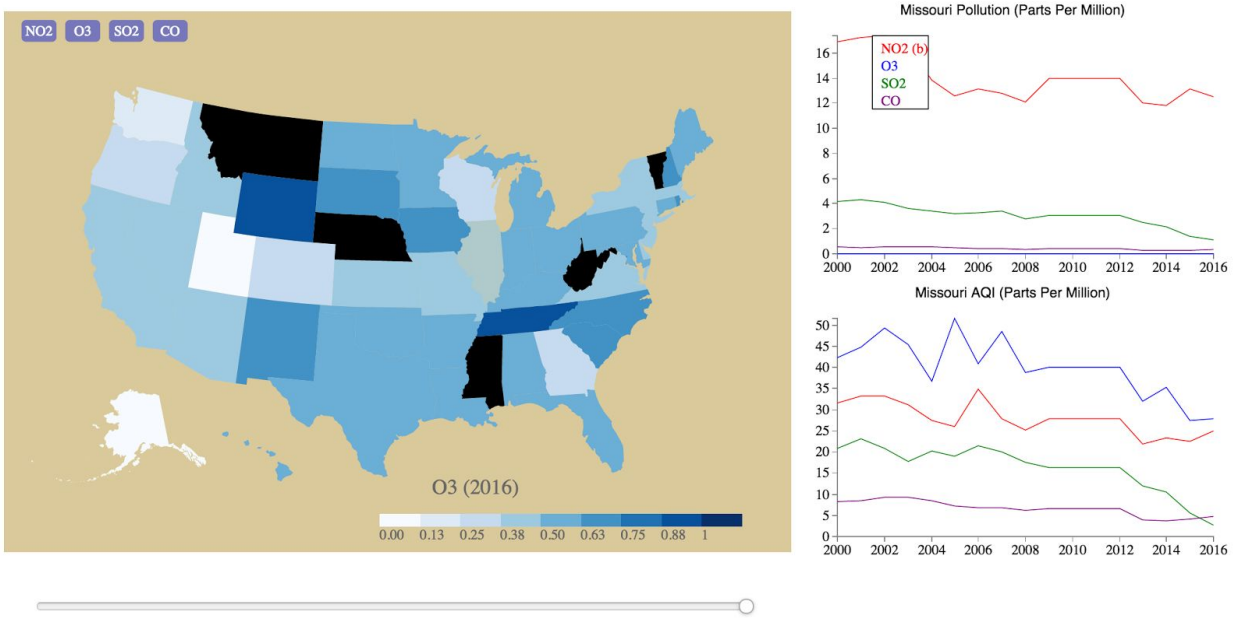
One of the deficiencies in our design was that certain states are harder to see than others. The District of Columbia was an outlier for CO but it is hard to visualize on the map since it's so small compared to other states. A potential fix to this problem is to show smaller states in another visualization that makes them easier to see how they fit in the overall scale with each pollutant.

Final Product:

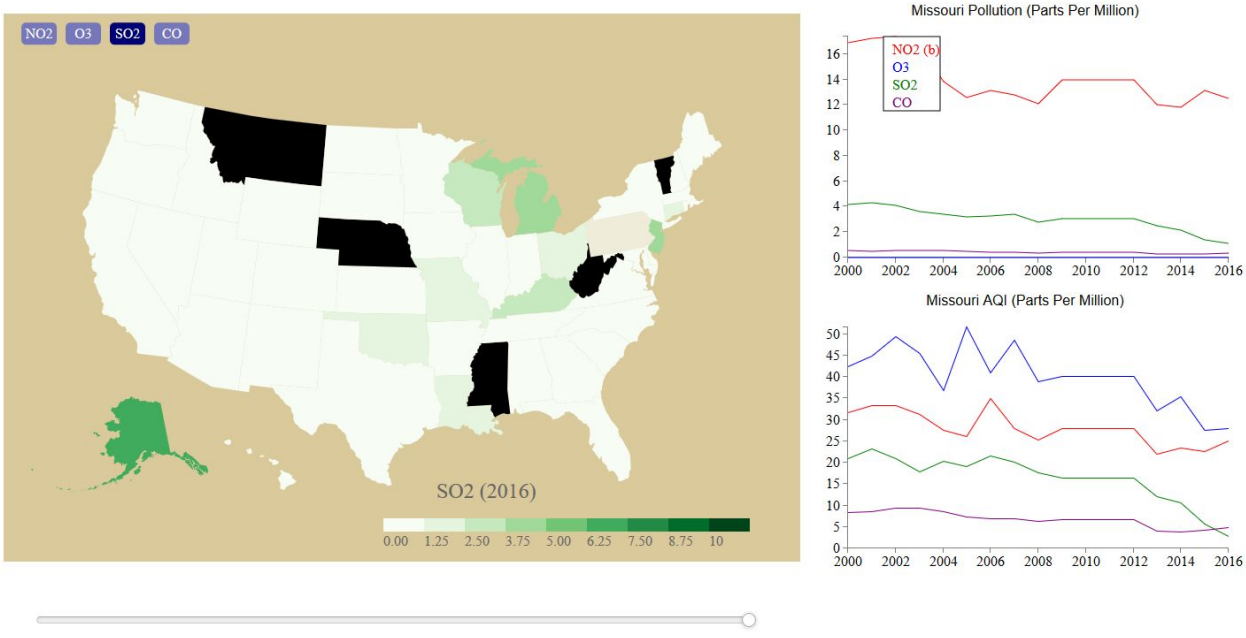
NO_2 Emissions View



O3 Emissions View



SO_2 Emissions View



CO Emissions View