CS 6630: Process Book - Now You See Me

Team Members: Ambuj Arora, Shaurya Sahai, Sushmitha S N

Basic Info

Project Title

Now You See me

Team Members

1. Ambuj Arora, u1265867, ambuj.arora@utah.edu
2. Shaurya Sahai, u1266148, u1266148@utah.edu,
3. Sushmitha S N, u1265043, sushmitha.sn@utah.edu

Project Repository:

https://github.com/ar-ambuj23/dataviz-project2019

Overview:

Now You See Me is a data visualization project that would allow the user to visualize trends of major US pollutants such as Ozone, Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide over a period of time. It would also show a comparison between pollution levels of 2 different states. This project would include 3 main visualizations:

1. Showing the map visualization of various pollutant levels over a period of time.
2. Tabular view of pollutants to see multiple state data at once.
3. A comparative visualization between 2 states.

Background and Motivation

Our motivation for the project comes from the fact that we wanted to build something which could be used by the general public and/or experts to analyse trends in something that affects all of us.

Air pollution, as most of us are aware, is one of the most serious problems in this age and time. It refers to the contamination of the atmosphere by toxic chemicals or organic materials. Polluted air has an adverse effect on the ecological system. It's important to study the statistics of air pollution because it shows how the quality of air is changing over time. Generally, the statistics reflect the levels of different pollutants such as ozone, nitrogen dioxide, sulfur dioxide, carbon monoxide, etc. There is no denying the fact that reducing the pollutants in the air is crucial for human health and environment. Therefore, the study of air pollution is very important.

We have taken US pollution dataset from 2000 to 2016. Our tool displays trends of major air pollutants such as Ozone, Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide in the United States. We are providing both temporal and spatial views for clear visualization of the data. Using our tool, we aim to let users observe the trend in air pollution over a period of time in various states. Also, by studying the existing pattern we plan to extrapolate and make future predictions.

Related Work

Our initial source of motivation was the fact that all of the team members wanted an environment-centric topic to explore with the visualization concepts along with the tools that D3 offers. We collected our dataset from https://www.kaggle.com/ and also had a look at the *kernels* which other people had put there. We got inspired from the variety of visualizations people had put up and wanted to extend the idea including multiple functionalities.

We were also quite amazed with the way D3 handles maps and makes it extremely easy for the developers. We wanted to build our visualization around that and hence, we came up with our visualization designs.

One particular visualization that inspired us was the 3D visualisation for debt across years and the story telling aspect of it. It focussed on major events and took care to give the user enough context about what was happening while he/she interacted with the visualisation. We wish to take our tool easy-to-use and interactive for the user.

Project Objectives

We aim to show how pollution trends across various states over a period of time. The main questions that we are trying to answer are

1. Which is the highest and the lowest polluted state for a given year for a given pollutant?
2. What is the trend in pollution for any given pollutant over a period of time?
3. Can a user have an animated view of the pollution trend? Can he pause the animation?
4. What is the trend in pollution change with respect to a certain geographic area in the US?
5. Can a user see multiple state's pollution data at once?
6. Can a user see which state has the least Carbon monoxide or Sulphur Dioxide or Nitrous Oxide or Ozone?
7. Can a user see which state has the highest Carbon monoxide or Sulphur Dioxide or Nitrous Oxide or Ozone?
8. Can a user get a sorted view of states?
9. What is the trend in various pollutants for a selected state over a period of time?
10. Can a user compare the pollution trend between any 2 states?

We would like to use the concepts we learned in class and assignments to visualize the US air quality across various states. Also, we have included a few other d3 based visualization to build an effective user interface in interesting visualization of the US air pollution.

Data

The dataset deals with pollution in the U.S. Pollution in the U.S. has been well documented by the U.S. EPA. We aim at visualizing the distribution and trends of pollutant levels across the whole US. There are four pollutants that are visualised, namely:

- NO2
- SO2
- O3
- CO

Source of the data: https://www.kaggle.com/sogun3/uspollution

The data was in csv format originally. For the first release, we have converted the csv to json using Python to concentrate more on the visualisations. In the final release, the csv format of the data will be used as input to the code, wherein it will be converted into json for the visualization.

Data Cleanup and Pre-Processing

Cleanup

- The data is on a daily-level having the pollutant levels across different states, cities and counties in the US.
- The data, in the raw format, was actually a lot noisy than we had initially expected.
- We used Pandas and Numpy for cleaning the data using Python.
- As an initial step, we removed the unnecessary columns which didn't have any information relevant to our visualizations.
- The data had around 1 million rows initially. So, we had originally planned to cut down on some part of the data to make it possible to be visualised using D3.
- But after further analysis, we found that the data had redundant rows corresponding to the combination of a particular state, city, county and date.
- So, we removed the redundant rows after taking care of the NaN values.

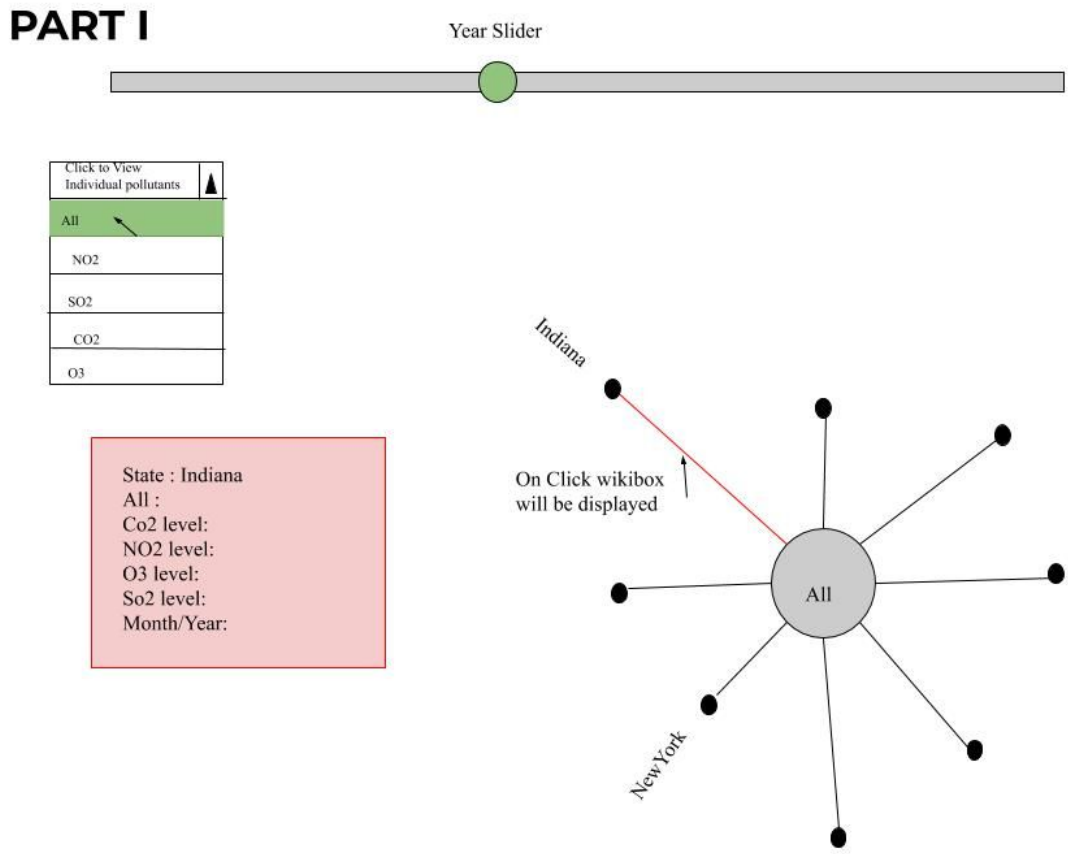
Pre-Processing

- The most important pre-processing that we needed to do on the data, was pre-processing.
- In our visualization designs, we had planned on giving a flexibility to the user to toggle between monthly and yearly view. But, some of the states do not have data for a couple of months. Hence we decided to discard it.
- As the data had each row corresponding to each day, we had to roll it up.
- The rolling was again done using Pandas by having the Date column as the index for rolling.
- We rolled up the data for a particular state, city and county combination, with date as rolling index, from daily level to monthly level and then to yearly level.

See data_preprocessing.ipynb for more details.

Exploratory Data Analysis and Design Evolution

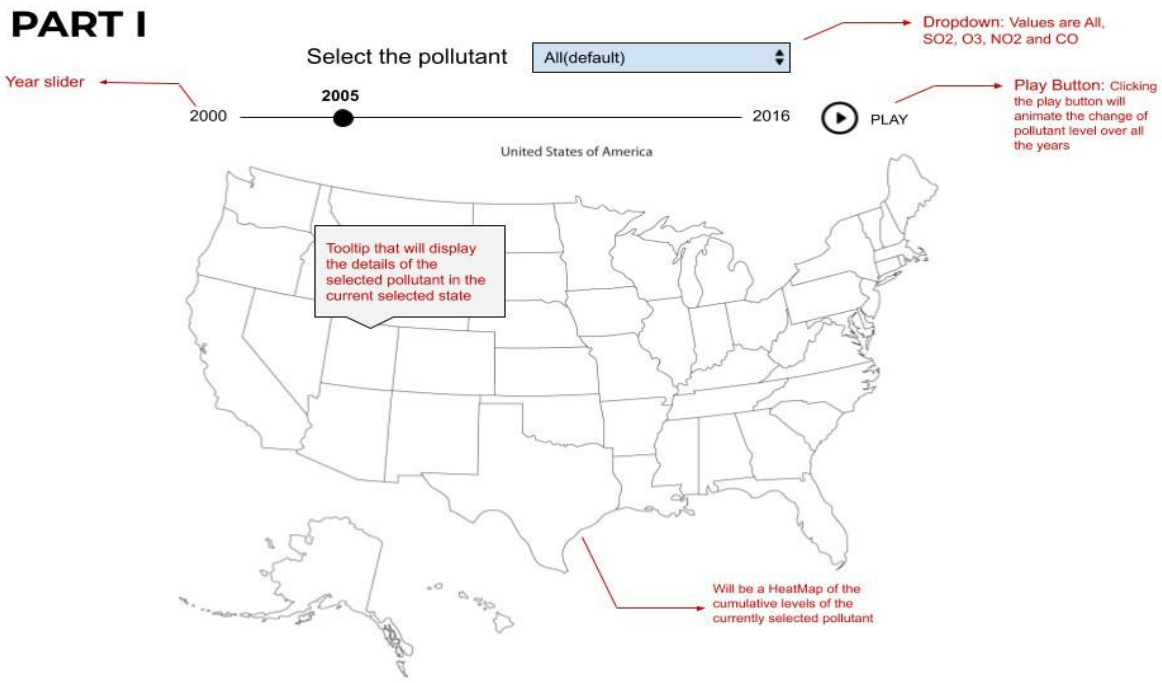
Below is the initial designs design that could have been the alternate visualization to our final visualization.



- This visualization focuses on the distribution of the selected pollutant level for all the states.
- A dropdown menu allows the users to visualise data about a particular or all pollutants.
- This has been achieved by using a radar display which has all the states as its edges.
- The length of an edge signifies the pollution level in that state.
- When an edge is selected, an infobox is displayed having all the details for that particular state such as individual pollutant levels, and month/year of selected time period
- This variant also gives user the flexibility to change the current year being displayed.

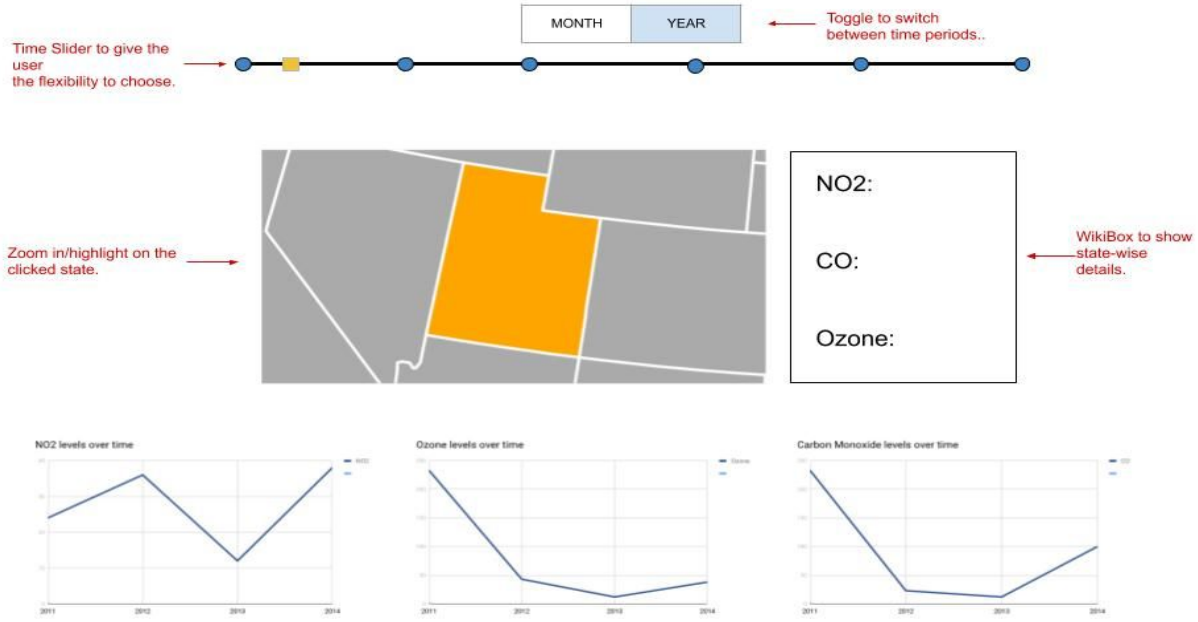
While the above visualization depicts a lot of interesting correlations, we came to the conclusion that it will not answer many specific questions nor does it give any specific insight to the data. After the peer review session and discussion with TA we came up with below alternate design

Modified Design and Implementation



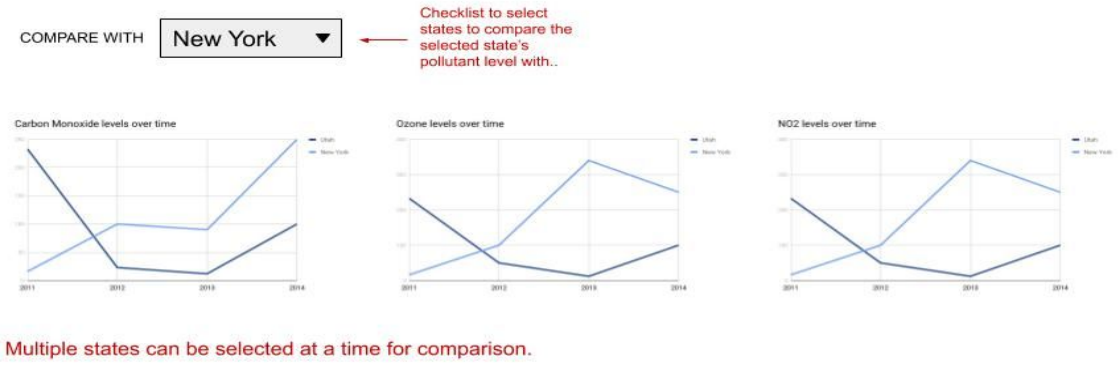
- This design shows the Map view of the pollutant level, as a whole, with a slider which allows the user to change the current year.
- The map shows overall pollution level of selected year as a Heatmap for each state. We plan on extending this, if time allows, to city level.
- When the user clicks on a particular state, the visualisation is changed to Part 2 which is a zoomed view of the selected state with all the analysis graphs.

PART II



- This design view is a zoomed view of the state selected in part 1.
- As a starter, we plan on highlighting the selected state with all the analysis line graphs but later we will change this functionality to be a zoomed view of the state.
- Wikibox will appear to show pollutants levels of selected state.
- Line charts to display trend in each pollutant.
- Each time a new state is selected, the old selection will be cleared. We will use animation to have better visual effects.

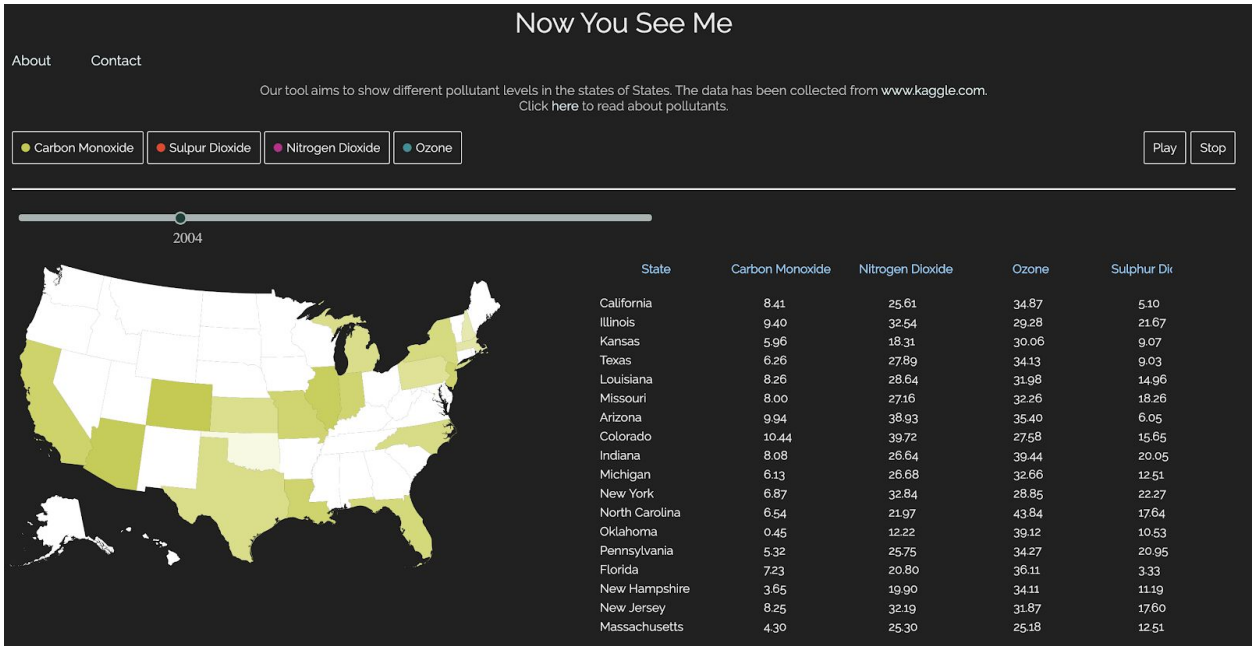
PART III



- This view is an extension of the 2nd design and allows the users to compare the pollution trend between any 2 (possibly many) states using the “compare with” dropdown menu.
- As a must have feature, we will allow the user to compare two states but later we will increase this capability to 5 states.

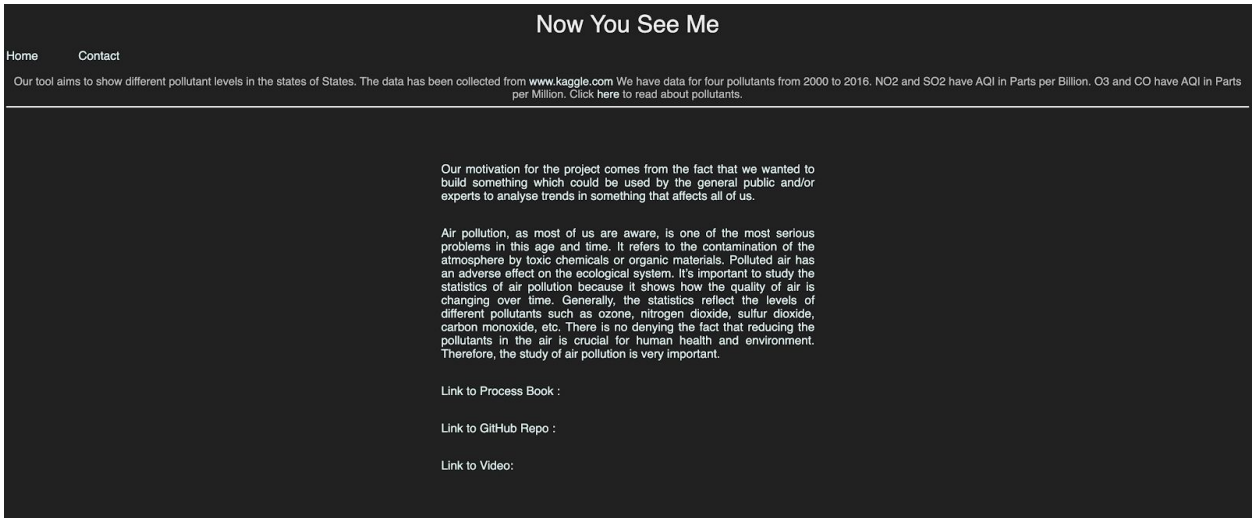
Modified Design and implementation

Below is the home page of our visualization

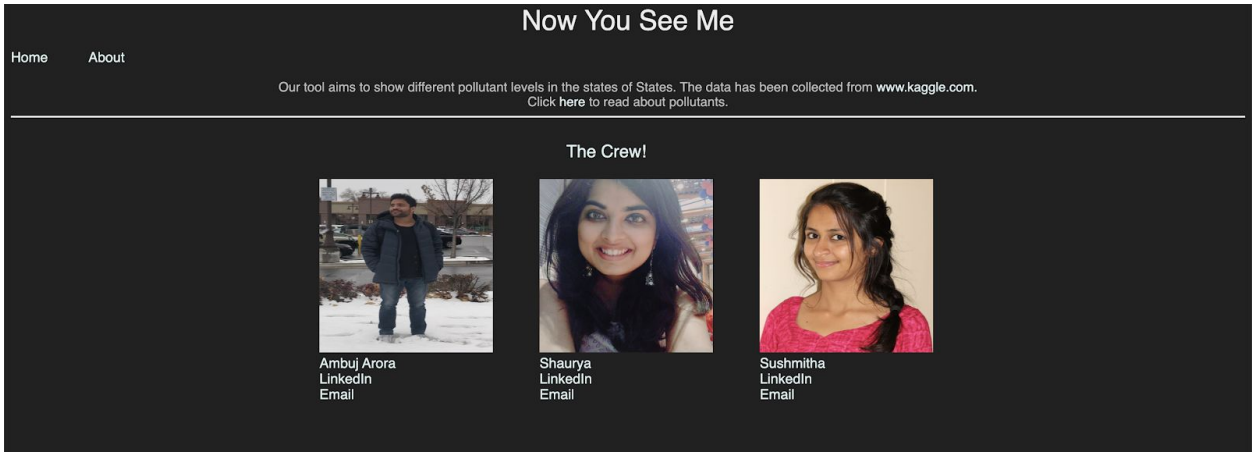


About tab: This tab tells the motivation behind selecting this topic.
This tab has references to

- The Process Book
- Link to the GitHub Repo
- Link to the Video



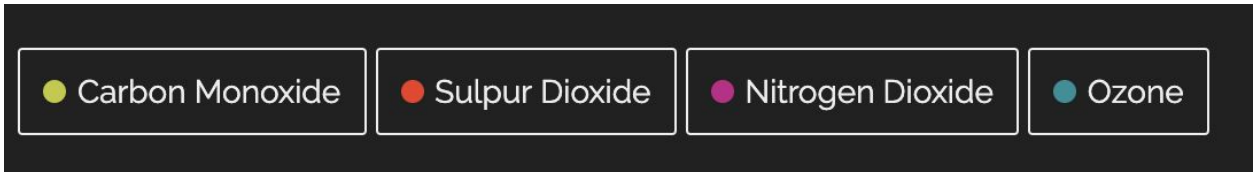
Contact tab: This tab has information about team members.



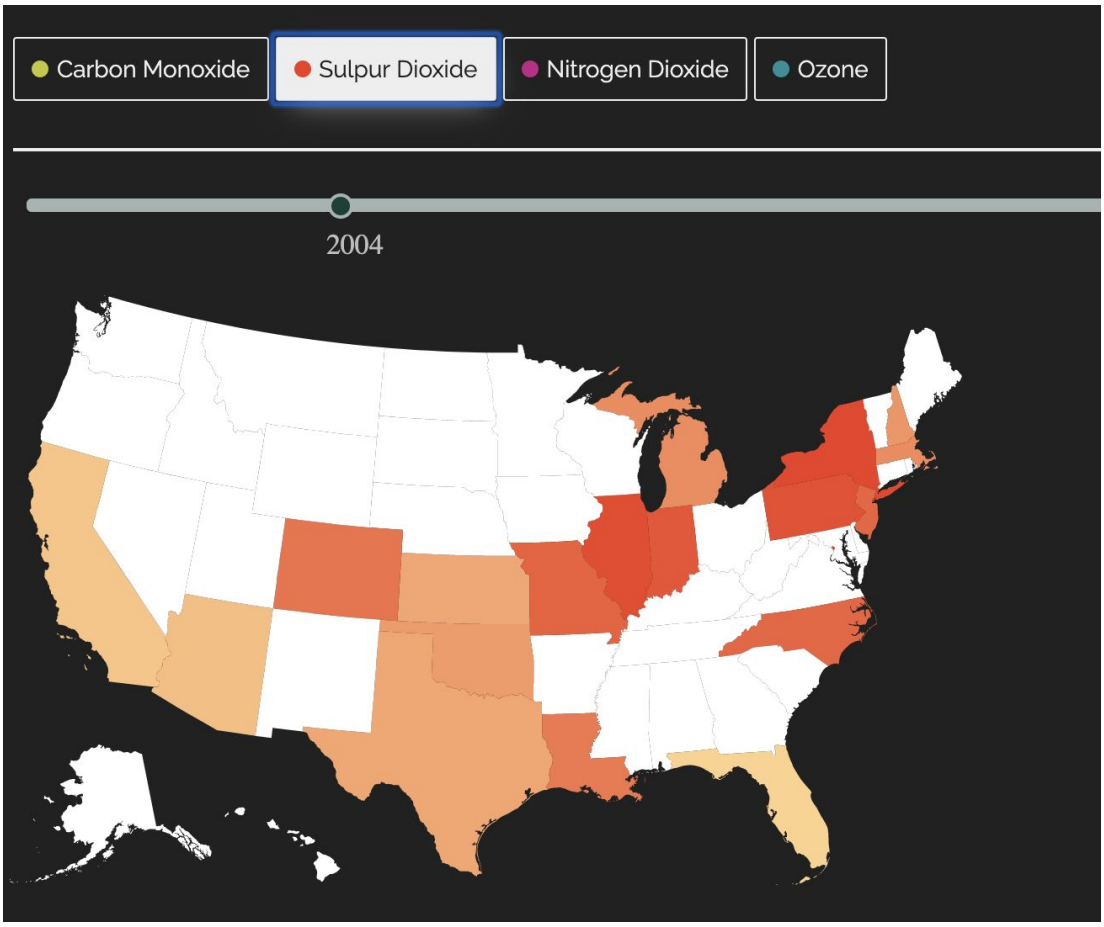
Not every user will know the hazards caused by pollutants. To address this issue, we have also given a link to a webpage that has information about health effects caused by pollutants.

In addition to this, we have also given a link to the website from where we pulled pollution data.

Buttons :



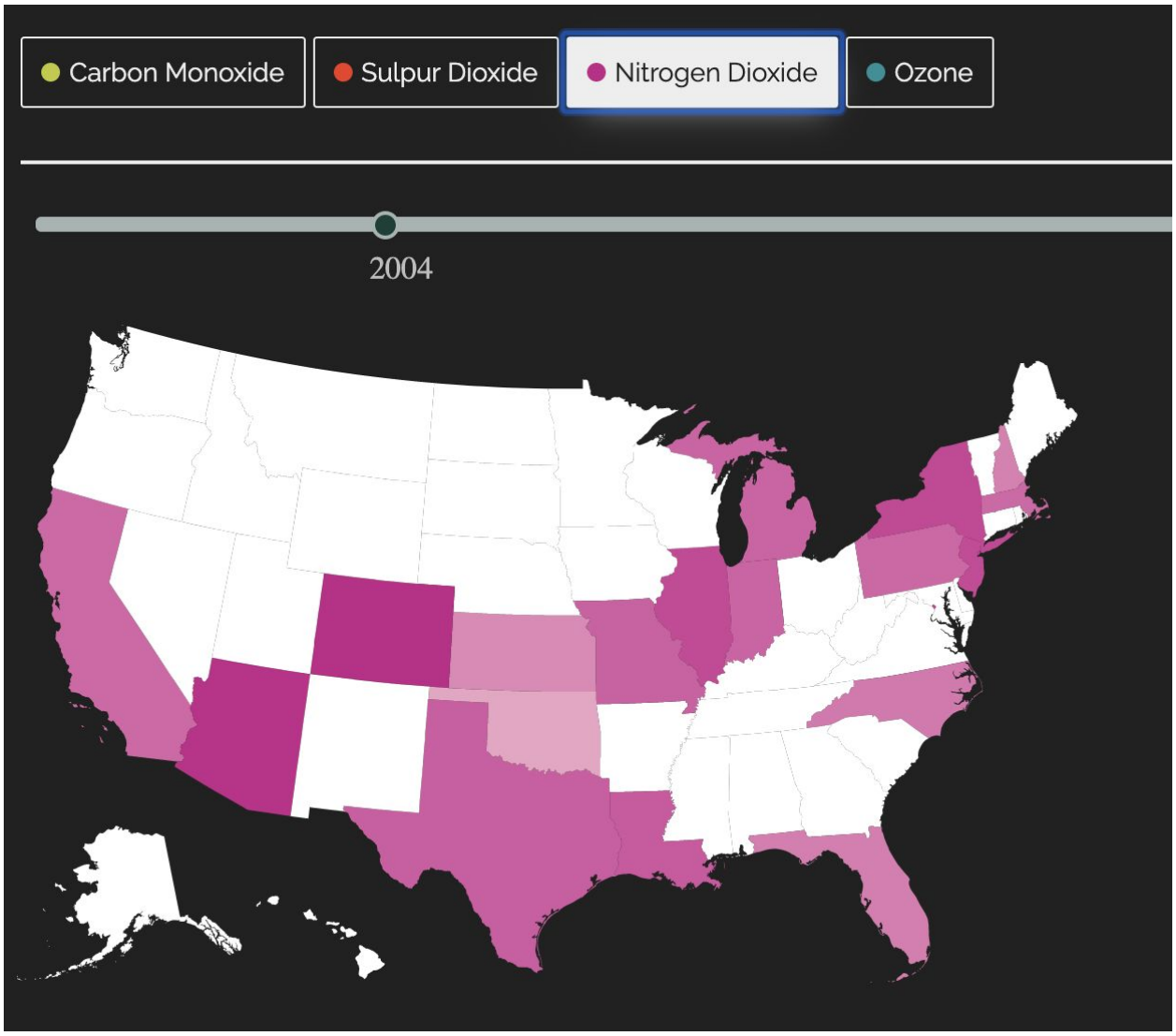
On click of each button, heatmap gets updated according to the button clicked. For example below image shows the sulphur Dioxide level across various states of the US for the year 2004.



Functionality remains the same for the other buttons as well.
Play and Stop Buttons and the Year Slider

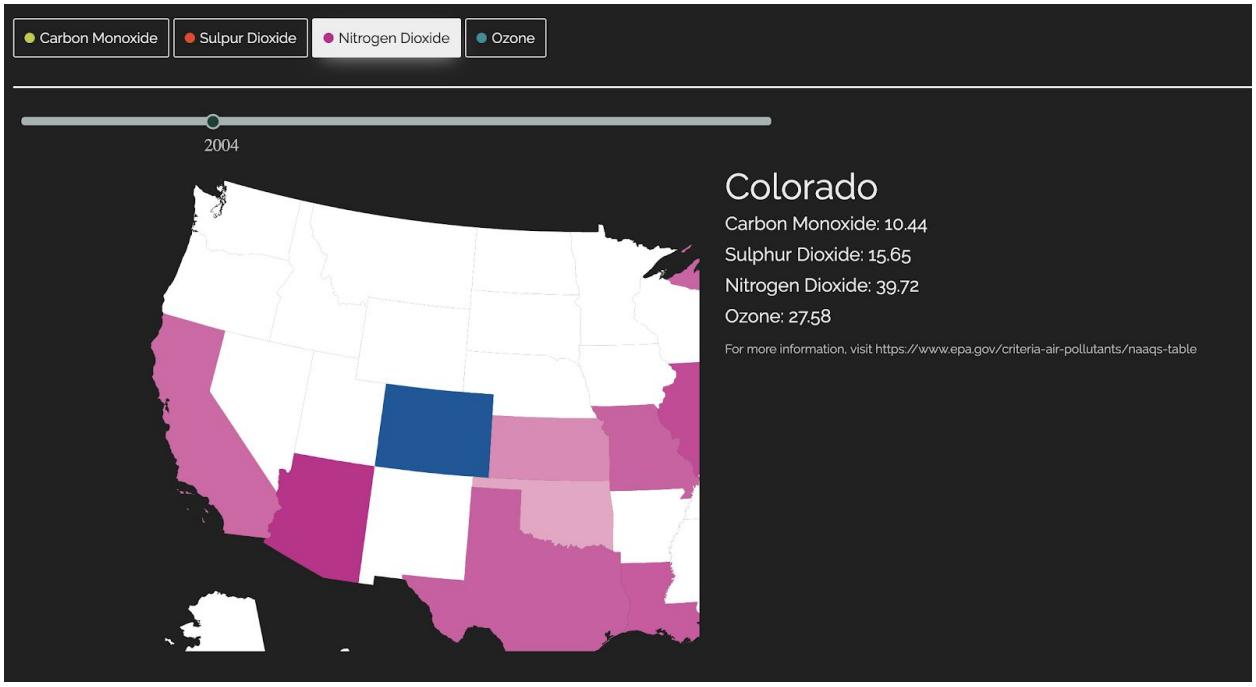
When play button is clicked, year slider moves from current time to the end that is 2016. Heatmap and table gets updated according to the time. Stop button is to stop playing at any point of time. Text in the slider button helps to identify which year data we are currently visualizing.
Heatmap:

Heatmap below shows the trend in the selected pollutant for any given time.

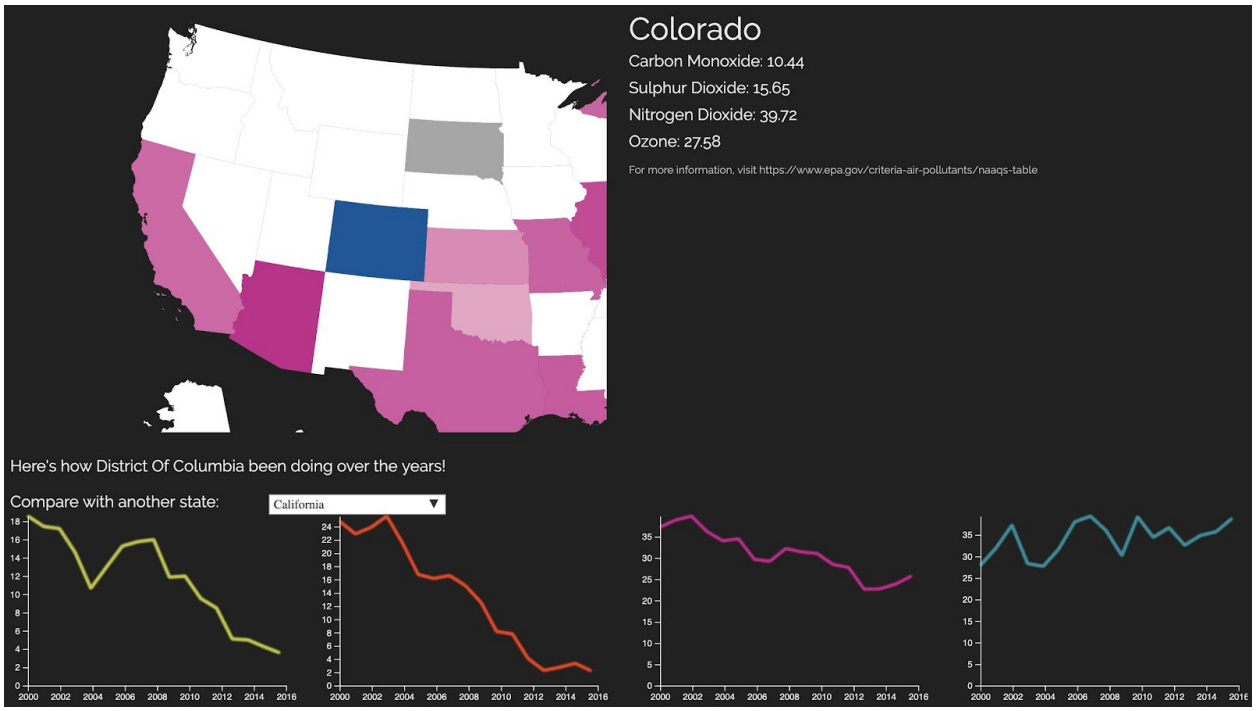


The above is the heatmap for Nitrogen Dioxide for the year 2004. State with the highest color intensity has the highest Nitrogen Dioxide levels for the year 2004.
Functionality of the heatmap remains the same across all the years and the pollutants.

Zoom In and Zoom Out Features.
Below is the image that shows that user has selected the Colorado State for the year 2004. A wikibox next to the heatmap helps user to find all the pollutant levels of the Colorado for the year 2004. Also the highlighted feature of the state is helpful to zoom out and to visualize the better visualization of the state. Also, on click of a state, the table will disappear and wikibox will appear to provide better visual aid for the user. Similarly, on zoom out wikibox wiull disappear and table will reappear.

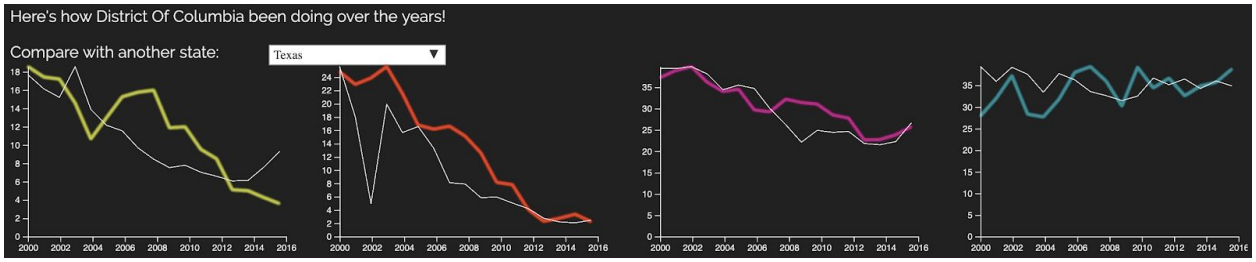


Line charts



The above is the trend in the Carbon Monoxide, Sulphur Dioxide, Nitrogen Dioxide, and Ozone for the state Colorado from year 2000 to 2016.
On Zoom Out line charts will disappear.

Comparison between trend in air quality between 2 states



Drop down helps us to select a state to compare with already selected state. On click of a particular state in the dropdown, a new line will appear. From above feature we can easily visualize which state among the two has the highest pollution.

Table

State	Carbon Monoxide	Nitrogen Dioxide	Ozone	Sulphur Dioxide
California	8.41	25.61	34.87	5.10
Illinois	9.40	32.54	29.28	21.67
Kansas	5.96	18.31	30.06	9.07
Texas	6.26	27.89	34.13	9.03
Louisiana	8.26	28.64	31.98	14.96
Missouri	8.00	27.16	32.26	18.26
Arizona	9.94	38.93	35.40	6.05
Colorado	10.44	39.72	27.58	15.65
Indiana	8.08	26.64	39.44	20.05
Michigan	6.13	26.68	32.66	12.51
New York	6.87	32.84	28.85	22.27
North Carolina	6.54	21.97	43.84	17.64
Oklahoma	0.45	12.22	39.12	10.53
Pennsylvania	5.32	25.75	34.27	20.95
Florida	7.23	20.80	36.11	3.33
New Hampshire	3.65	19.90	34.11	11.19
New Jersey	8.25	32.19	31.87	17.60
Massachusetts	4.30	25.30	25.18	12.51

Above is the table which has the information about the pollutants of all the states for the year selected in time slider. This table can be changed dynamically using time slider and play/stop button. Also table can be sorted based on column headers. Above image has the table sorted according to State. Similarly, by sorting it according to the pollutant, a user can see the highest/lowest polluted state Main goal of above visualization is to provide the user with an option to analyze multiple state/pollutant data at one place.

Design Issues

The data pre-processing was a bit tricky! We had to make a lot of decisions regarding the data roll-up. While rolling up the data, we found that there are some states which have data only for few months in a given year. Hence, we decided on not having the monthly view. Another major roadblock for us was the dropdown. It was not possible for us to have a dropdown inside a svg using the 'select'. Hence, we had to make the dropdown using 'rect' elements and appending text onto that. Other than that, everything else pretty smooth.

Demo

About tab has the link to the video which shows the project demo.

Evaluation

Our main aim of the project was to build a simple yet precise and elegant tool to view interesting insights about US air pollution. Initially, we thought that we had all the essential data to develop our visualization. Eventually, when we started working, we realized we do not have the data for multiple states. Moreover, for some states, there was no data for a few months, which was making the HeatMap for that state disappear entirely. Hence, we decided not to implement month-wise visualization. Also, data processing was a bit tricky because we needed to make various decisions while processing the data. The data processing decisions can be found at the GitHub project repo.

We wanted our visualization to answer a few specific questions while we came up with the project proposal. Later on, we decided to answer a couple more questions in the visualization.

Below are the set of questions and the corresponding visualization that answer them.

1. Which is the highest and the lowest polluted state for a given year for a given pollutant? Consider the year 2004 and the pollutant as Carbon Monoxide.
    - The answer to the above question can be found in the heatmap. The heatmap is color-coded according to the pollutant level. The more intense the map color, the more polluted that state is concerning that pollutant.
2. What is the trend in pollution for any given pollutant over a period of time?
    - On click of the heatmap, the line charts for all the pollutants from the year 2004 to 2016 are displayed, which shows the trend in pollutants over a period of time.
3. Can a user have an animated view of the pollution trend? Can he pause the animation?
    - This question can be done using the "play" and "stop" button.
4. What is the trend in pollution change for a specific geographic area in the US?

- This question can also be answered by the play and stop button. We can observe the trend in a specific pollutant in a particular geographic area over the period of time.

5. Can a user see multiple state's pollution data at once?
   - The table answers this question. Using the table, it would be easy for the user to analyze multiple state data at once. Also, a user can view all the pollutant data at once.
6. Can a user see which state has the least Carbon monoxide or Sulphur Dioxide or Nitrous Oxide or Ozone?
7. Can a user see which state has the highest Carbon monoxide or Sulphur Dioxide or Nitrous Oxide or Ozone?
   - Sort according to the header feature in the table, answers the above questions.
8. Can a user get a sorted view of states?
   - Sort the table according to the state column.
9. What is the trend in various pollutants for a selected state over a period of time?
   - Users can manually select the time using the time slider, or the play button can be used to observe the trend.
10. Can a user compare the pollution trend between any two states?
    - A drop down to select state allows a user to compare two given states.

We have implemented three out of four optional features in the project, along with the must-have features. A future enhancement to this project would visualize significant events that happened that impacted change in pollution for a specific time.

References:

Data: https://www.kaggle.com/sogun3/uspollution
Data Processing:
https://github.com/ar-ambuj23/dataviz-project2019/blob/master/code/data_preprocessing.ipynb