Process Book

November 7, 2019

Overview 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.

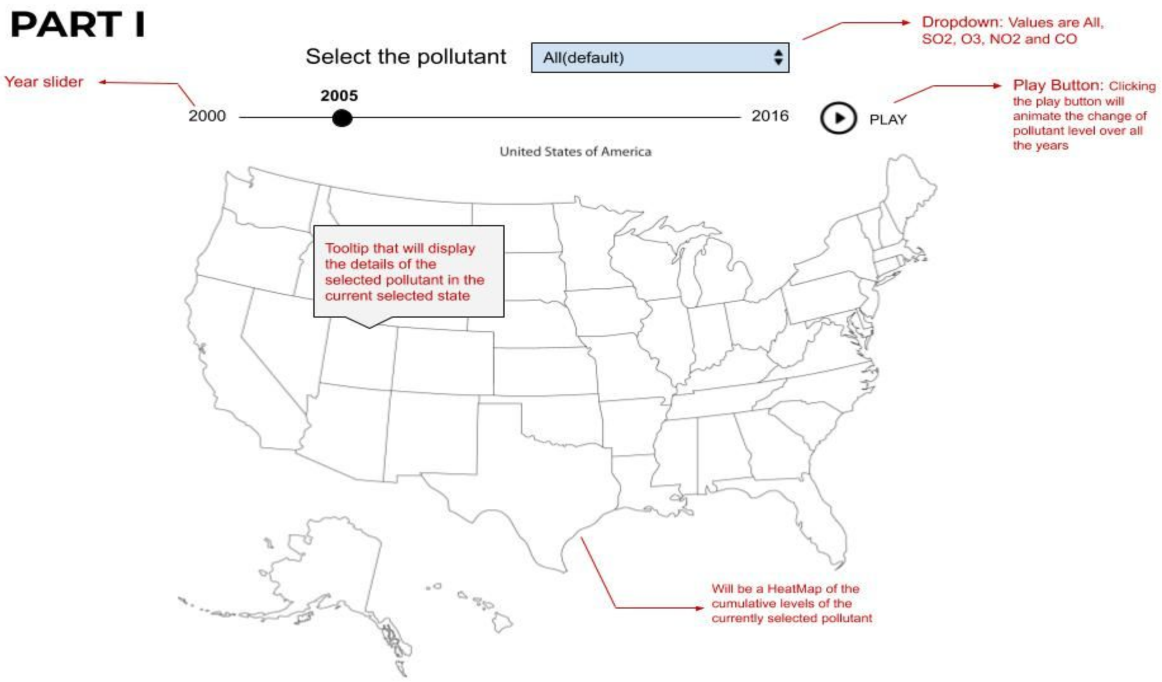
Related Work

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.

<u>Here is what our final selection of the screens look like:</u>



Dropdown: Values are All, SO2, O3, NO2 and CO

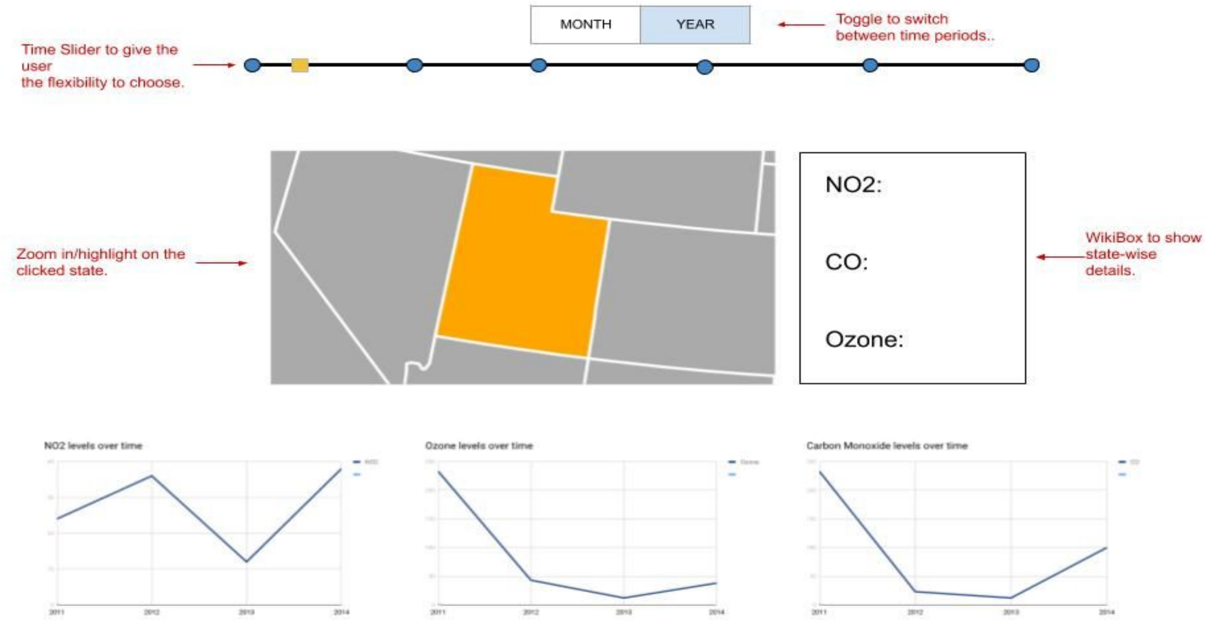
Play Button: Clicking the play button will animate the change of pollutant level over all the years

Tooltip that will display the details of the selected pollutant in the current selected state

Will be a HeatMap of the cumulative levels of the currently selected pollutant

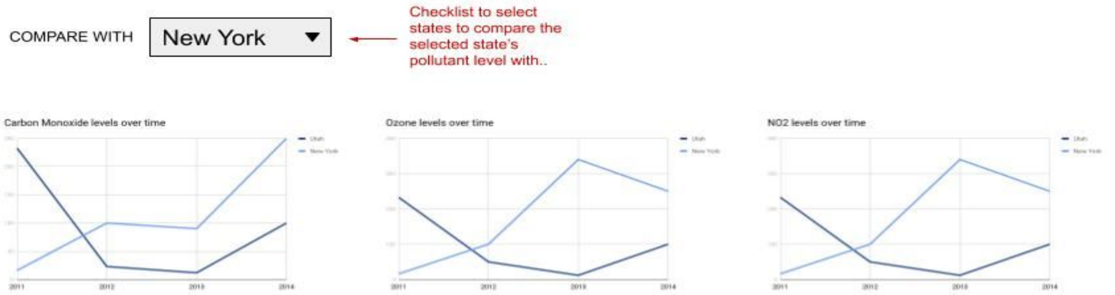
- This design shows the Map view of the pollutant level, as a whole, with a slider which allows the user to change the current year or the current month.
- The user can also toggle between monthly and yearly views.
- The map shows overall pollution level of selected year/month as a Heatmap for each state. 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


MONTH
YEAR
Toggle to switch between time periods..
Time Slider to give the user
the flexibility to choose.
Zoom in/highlight on the clicked state.
NO2:
CO:
Ozone:
WikiBox to show state-wise details.
NO2 levels over time
NO2
2011
2012
2013
2014
Ozone levels over time
Ozone
2011
2012
2013
2014
Carbon Monoxide levels over time
CO
2011
2012
2013
2014

- 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



Checklist to select states to compare the selected state's pollutant level with..

Multiple states can be selected at a time for comparison.

- 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.

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.
- 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.
- The monthly level is the most granular level we need for our visualization.
- For yearly view, we will roll up the monthly data to yearly level using D3 and JS, as and when required.

See data_preprocessing.ipynb for more details.

Peer Review

Peers: **Jess Campbell, Jeremy Thorpe and Lukas Gust.**

We would like to thank Jess, Jeremy and Lukas for taking the time to listen to our ideas, go through our designs and give extremely helpful constructive feedback.

Feedback: What procedure/formula would we use to get accumulated pollution level when we only have data for individual pollutants?

Response: We would use the air-quality index (AQI) parameter in our dataset to map the accumulated pollution level. We also plan to research on how, if at all, the four pollutants in our dataset combine to give a trustworthy and accurate accumulated level of pollution.

Feedback: It's good to show temperature/season of selection. For example, pollution in Utah is higher in winter compared to summer.

Response: We will show temperature/season as part of tooltip/wiki page by doing some computing the selected month and related season, according to the location. Although this remains to be a part of Phase II.

Feedback: They suggested it'd be good to have animation when we click on a state. May be transition? Also have zoom out option to restore back to previous visualization.

Response: We plan to zoom into the state, if D3 allows or zoom into the system-screen itself. We'll add a zoom out button which can be clicked to go back to the initial version of the map.

Feedback: They suggested we include the problems each pollutant is causing or maybe a link to a website which has details about these problems.

Response: The visualization will be cluttered and unreadable by the user if we display all the hazards caused by pollutants along with map. We really like the idea of providing links for more information and we plan to go ahead with some trusted sources.

Feedbacks we're yet to formulate a plan for:

Feedback: A threshold line to show Federally concerned level. If pollutants in that state are above threshold line, then it is a concern. Also show how many pollutants are above this line.

Feedback: Get state-wise data for PM 2.5 level and incorporate that as a pollutant because PM level is a very good indicator of air pollution.

Feedback: Can we include brushing as a part of our tool to select multiple states?