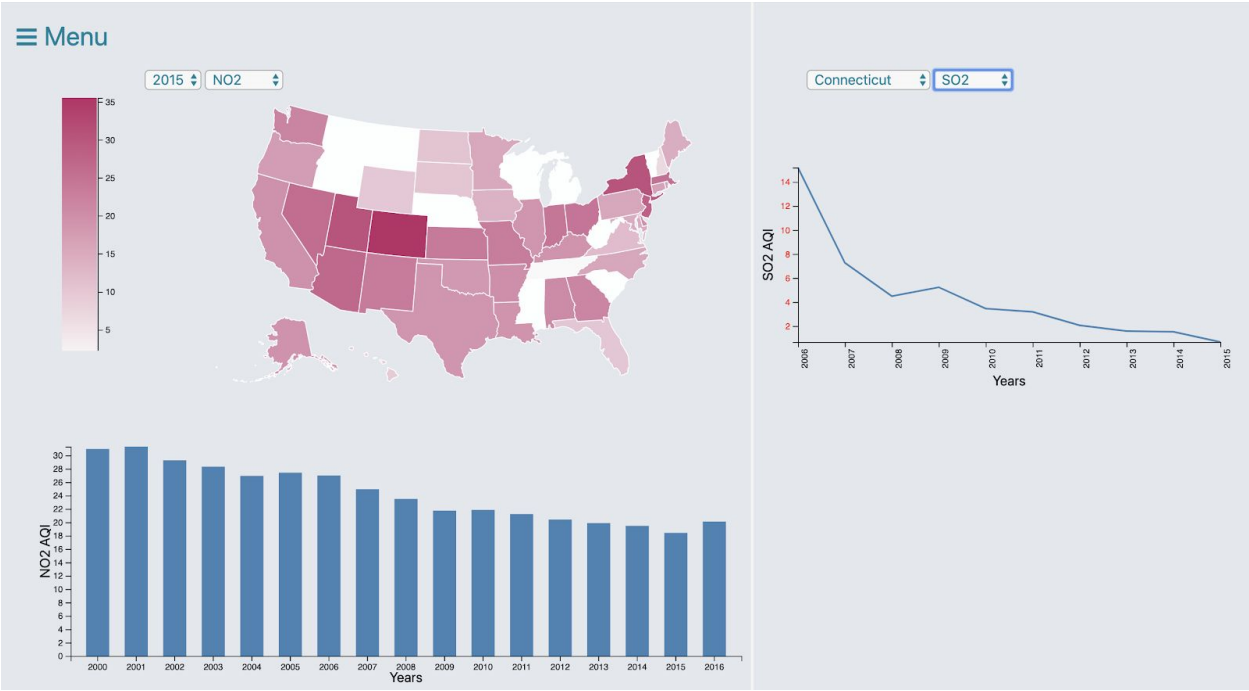
CSE 564 Final Report

Amit Dharmadhikari (112044244)
Yash Makheja (112052696)

Implementation Details

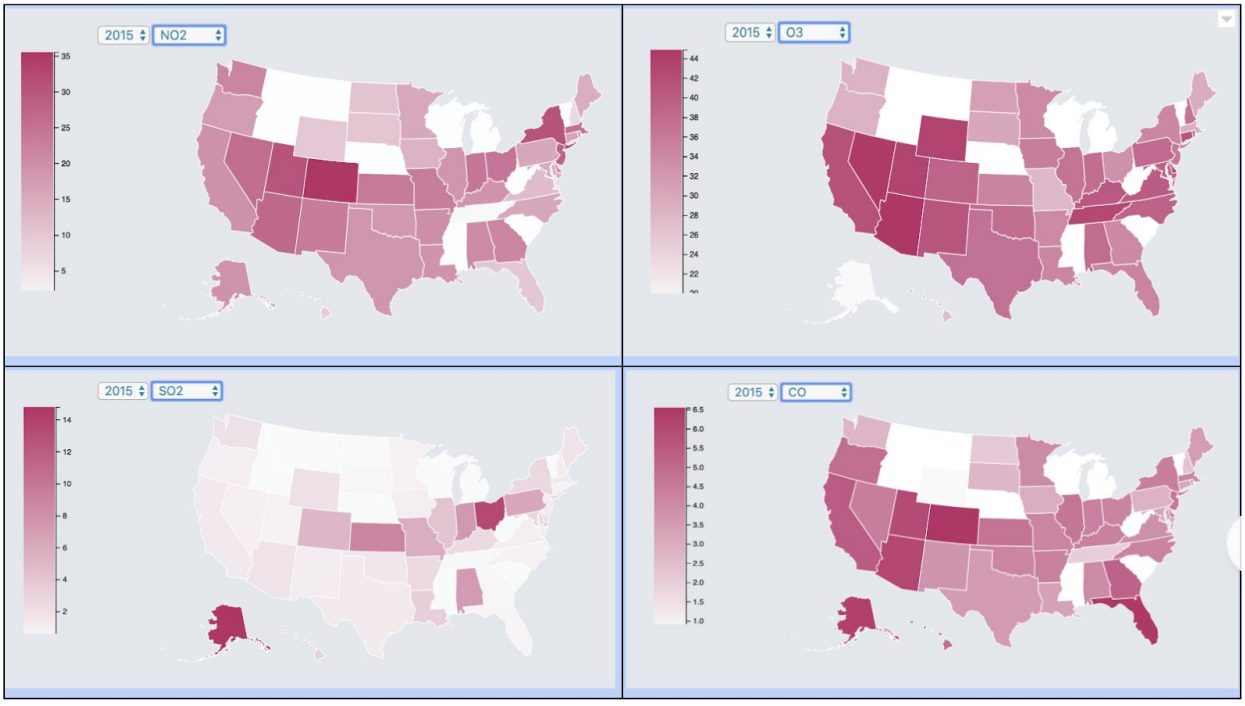
- We designed our dashboard, where a user can visualize several trends by selecting options from the side navigation bar.
- We implemented our back-end in Python using the Flask framework and front-end using HTML, CSS, and Javascript. We used d3.js to make visualizations, AJAX to connect to the back-end and retrieve data, and JQuery positioning our visualizations.
- We also found out the intrinsic dimensionality of the data using principal component analysis and calculated the squared loadings to find the top attributes.
- Using the top two attributes we designed a coordinated view using scatter plot and choropleth map where user can select a part of the scatter-plot and the selected states get highlighted in the choropleth chart.
- Following is the detailed description of each feature and insights we obtained while visualizing our data.

1. Pollutants Analysis



- Under this option, we have shown 3 visualizations - a choropleth map, a bar chart, and a line chart.
- The choropleth map shows the distribution of a particular pollutant across the entire US for a particular year. The pollutant and the year can be selected using drop-down lists on top of the map.
- We have used the AQI (Air Quality Index) to represent the level of pollution. The areas with high AQI, i.e. high pollution, are shaded dark, whereas the areas with low AQI are shaded light.
- The bar chart shows the year-wise series of AQI value for the selected pollutant for the entire US.
- For this, we aggregated the data according to the year and took the mean. The X-axis of the bar chart represents the years while the Y-axis represents the AQI value for the selected pollutant.
- The line chart represents the year-wise series of AQI value for a selected pollutant and a selected state. The pollutant and the state are taken input from the user using drop-down lists on top of the line chart.
- For the line chart, we first filtered the data according to the given state and then aggregated it by year. The X-axis represents the years while the Y-axis represents the AQI value.

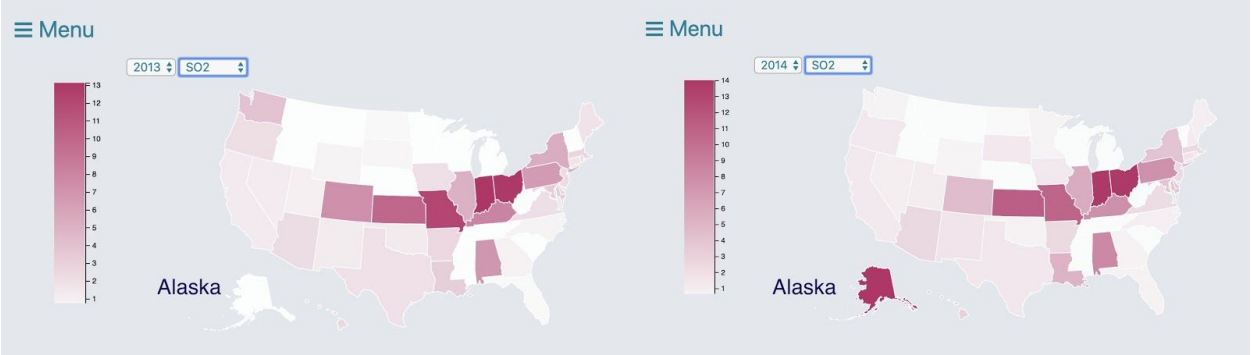
Distribution of Pollutants in terms of AQI levels



- The above grid of charts shows the distribution of all four pollutants across the US in 2015. We see that pollutants like NO2, O3, and SO2 are high in states like Colorado, Utah, Arizona, etc. SO2 follows a starkly different trend. SO2 is high along the belt from

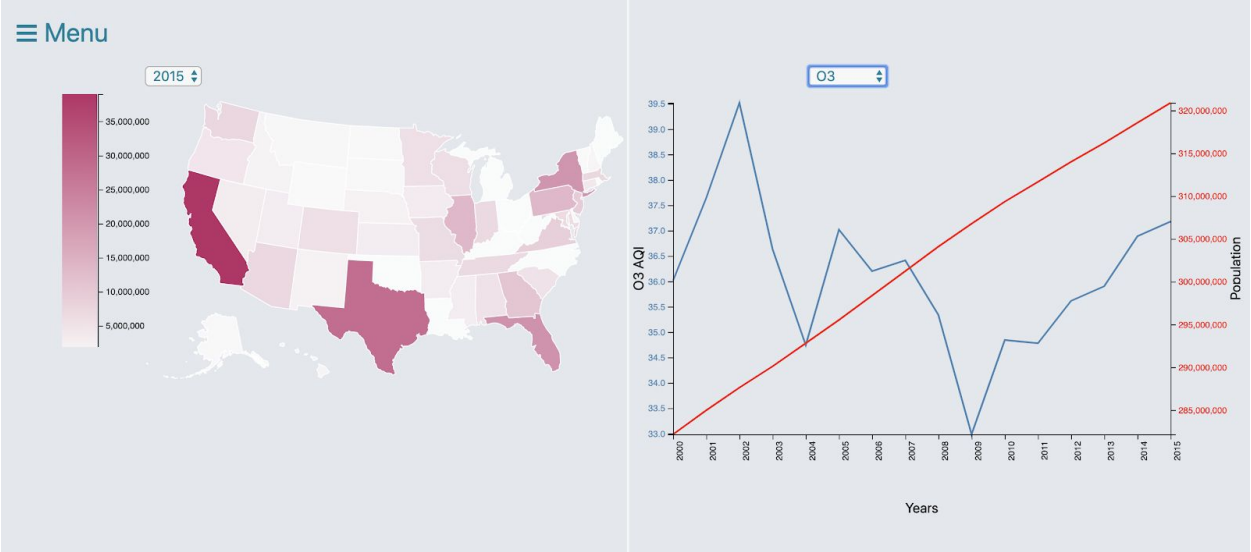
Colorado to Pennsylvania (especially in Ohio), and in Alaska, whereas it is very low in states like California, Texas, and Florida where other pollutants are generally quite high.

- This is because SO2 is mainly produced by industries and power plants and there are a lot of power plants in Ohio and along the entire belt. As for Alaska, it is because of the volcanic eruption of Mt. Pavlof in 2014.



The map to the left above shows the SO2 concentrations in 2013 while the one to the right shows the SO2 concentrations in 2014. We see that there is a sharp increase in the SO2 level in Alaska. This is because of the volcanic eruption of Mt. Pavlof in 2014.

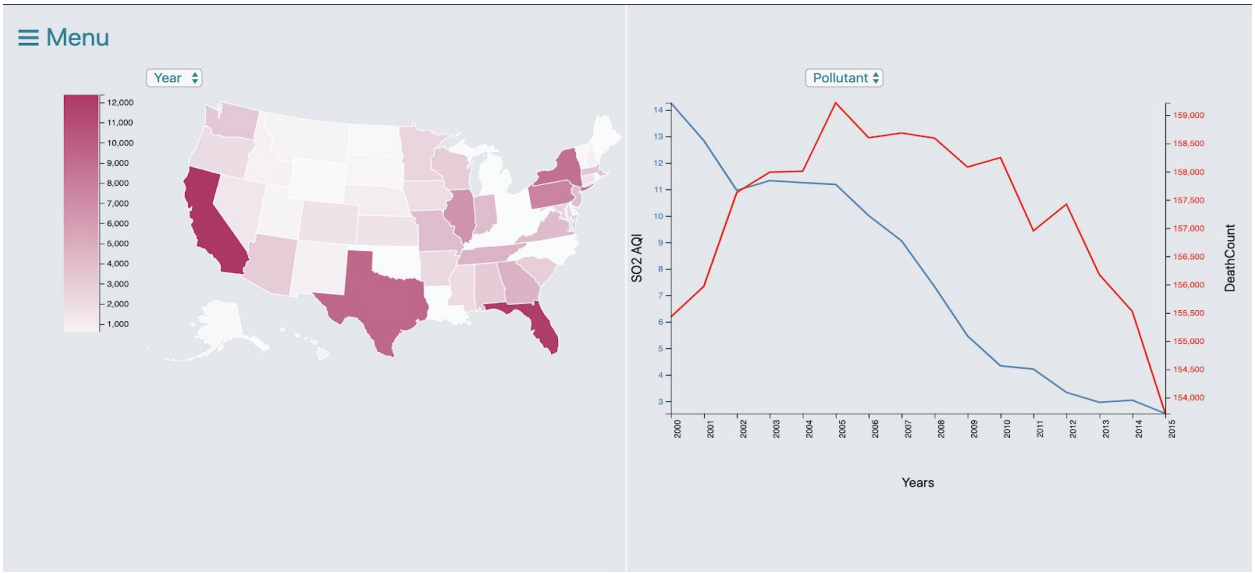
2. Population and pollutants - US



- In this option, the choropleth chart shows the distribution of the population across the country.
- User has the option to select the year from the drop-down list.
- The adjacent dual line chart shows the trend of population and pollutants from 2000-2015 and enables the user to change the pollutant from the drop-down list.
- We see that California, Texas, Florida and New York are the highest populated states in the USA.

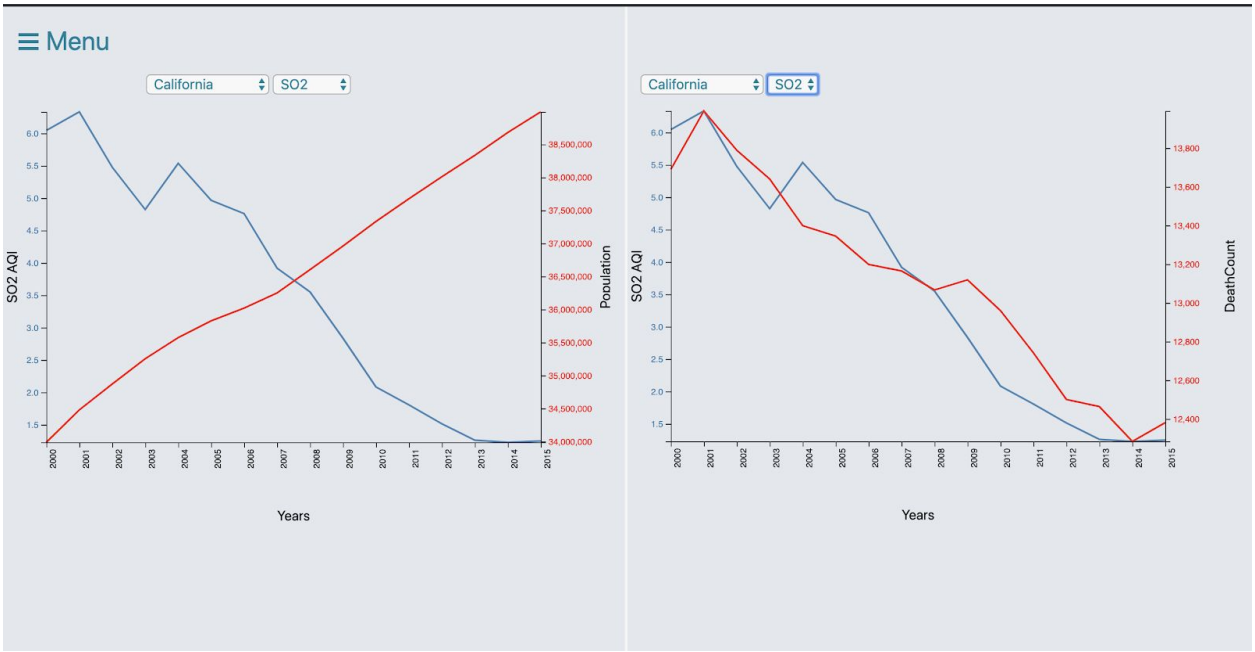
- Also, we see that the population has been following a continuously increasing trend.

3. Deaths and Pollutants - US



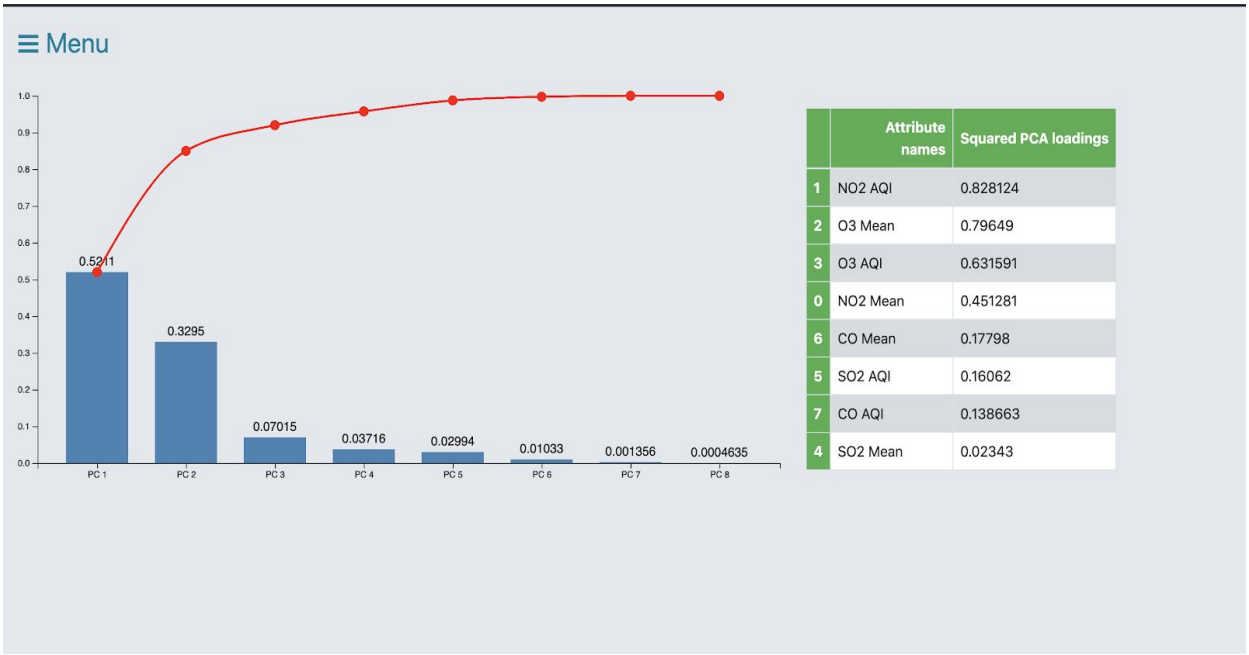
- In this option, the choropleth chart shows the distribution of the death counts caused by lung and bronchi cancer across the country, and year can be chosen from the drop-down list.
- The adjacent dual line chart shows the trend of death count and pollutants from 2000-2015 and enables the user to change the pollutant from the drop-down list.
- We see that California, Texas, Florida and New York are the states with the highest death count in the USA. Thus, we observe that the states with a high population like these four also have high levels of pollutants and a high death count. It could be that a high population means a high number of vehicles, thus causing high pollution and consequently a high death count.
- Also, we see that the death count is following a decreasing trend over the past few years.

4. Population and Deaths vs Pollutants - Statewise



- Here, user can see the compare the trends of pollutants vs death count and population for which state and pollutant can be chosen from the drop-down list.
- We observed that although the population is increasing the pollutant AQI keeps on decreasing for almost every state.
- We also observed that the death count is correlated with the pollutant AQI level and has been plummeting since 2000.

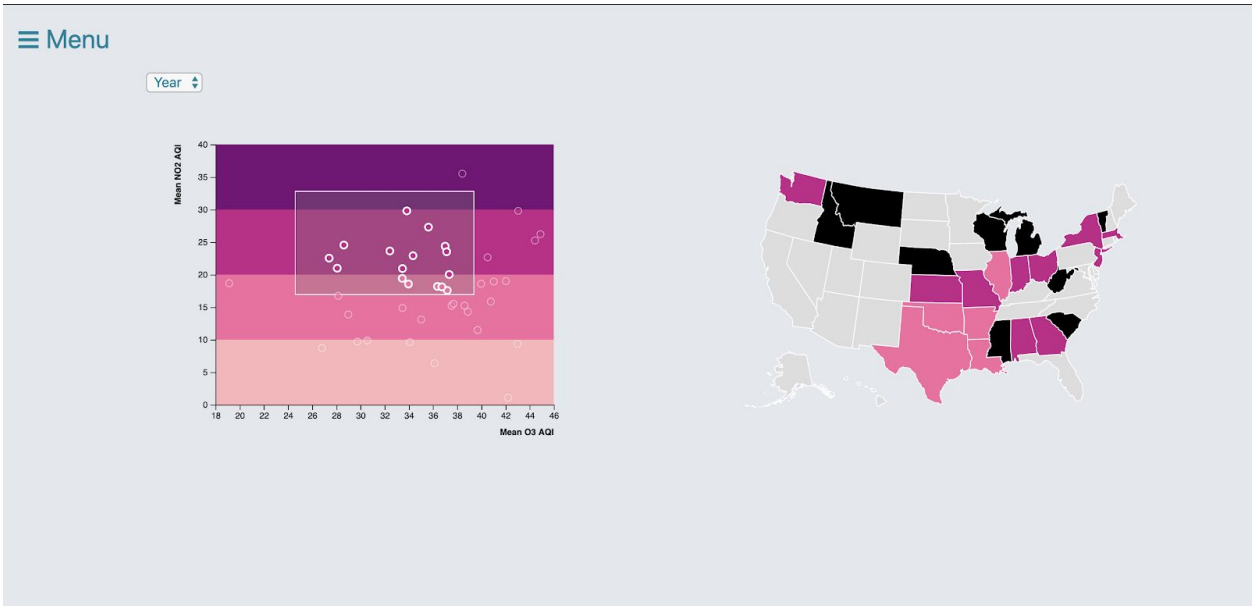
5. Principal Component Analysis



	Attribute names	Squared PCA loadings
1	NO2 AQI	0.828124
2	O3 Mean	0.79649
3	O3 AQI	0.631591
0	NO2 Mean	0.451281
6	CO Mean	0.17798
5	SO2 AQI	0.16062
7	CO AQI	0.138663
4	SO2 Mean	0.02343

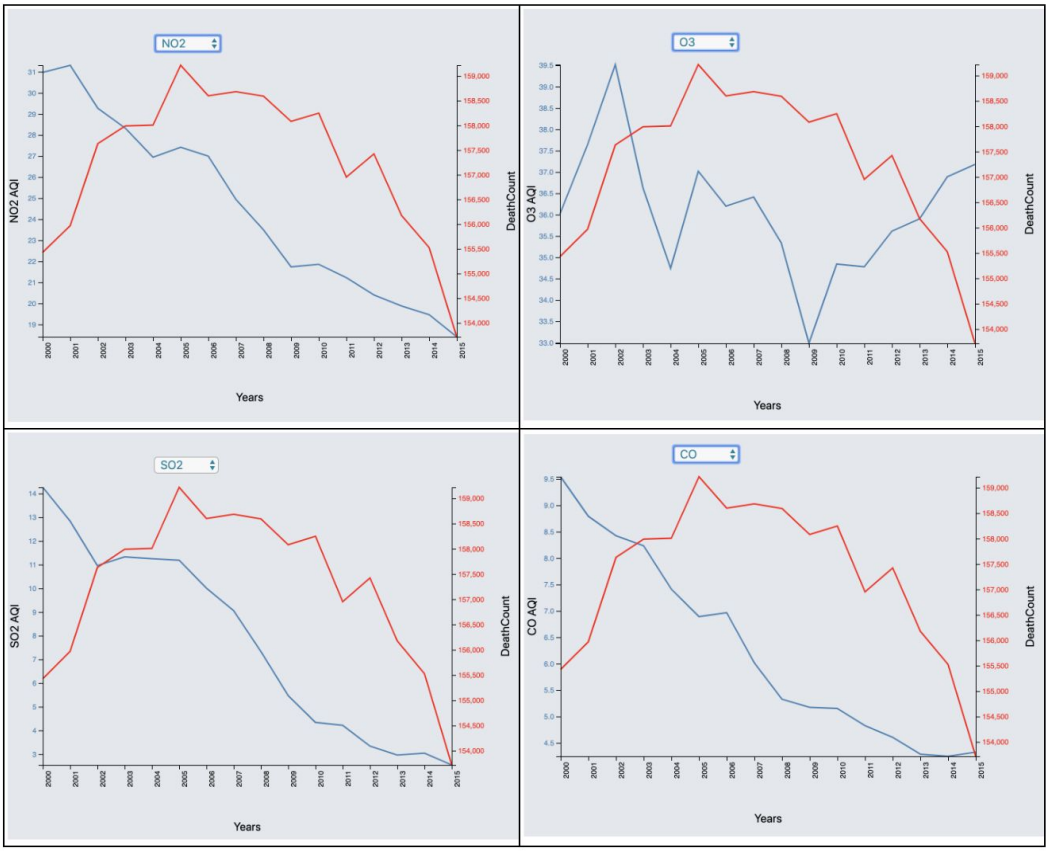
- We performed PCA to find out the intrinsic dimensionality of our data. We passed the mean value and AQI of each of the four pollutants to PCA.
- The chart to the left above is a scree plot of our PCA results. The X-axis represents the number of components while the Y-axis represents the explained variance.
- We see that 92% of the variance in our dataset is explained by the first three components itself. Thus, we decided to use only the first three components for further analysis.
- Further, we calculated the squared loadings for each of the attributes passed to PCA using the first three components. The table to the right shows a sorted list of attributes and their squared loadings.
- We see that NO2 AQI, O3 Mean, O3 AQI, and NO2 Mean are the four most important attributes. This leads us to the conclusion that NO2 and O3 are the most important pollutants in the dataset as they explain the most variance in the data.

6. Coordinated View



- Following option enables a user to visualize the scatter plot of the top two attributes obtained from PCA vis NO2 AQI and O3 AQI.
- The user can select a part of scatter-plot and can see the selected states highlighted over the choropleth map to get a better understanding of the regions.
- User can select a year from the drop-down list to see the distribution over a particular year.
- We see that this scatter plot has an outlier to the left, whose name is Alaska. Again, this could be because of the volcanic eruption of Mt. Pavlof in Alaska, which caused a sudden and drastic change in the levels of particular air pollutants.

Pollutants AQI and Death Count for the entire US from 2000 - 2015



- The above grid shows line charts of each pollutant vs death count for the entire nation year-wise.
- We see that NO2, SO2, and CO are following a clearly decreasing trend while O3 is anomalous.
- We found out that O3 is produced when heat and the ultraviolet part of sunlight interact with the pollutants released from cars. This could be the reason for the anomalous trend of O3 because it is harder to control than the other pollutants.

Conclusion:

Air pollution is one of the most serious problems in the world. Air pollution is the cause of various health effects ranging from short term effects like irritation and headache to long-term effects like lung cancer and heart diseases. Hence, It is important to analyze the data in order to deal with air pollution and devise solutions to counter the effect. By visualizing various trends and learning about insights government can easily develop countermeasures specific to a pollutant or a region. Also, these visualizations help common people visualize the change in the air around them.