Damini Rijhwani STAT Final Report

Air Quality Data Analysis in US

Problem:

The problem I have attempted to solve, is the abundance of data on air quality in United States of America and the methods on analyzing the data. In this project I attempt to create a simple tool to analyze data provided by the Centers for Disease Control and Prevention and provide hypothesis of certain results.

The motivation of this project stems from the rapid changes in climate change due to the underlying cause of emissions of pollutants in the atmosphere. The ecological changes faced in the recent decades and the health problems that affect all demographics are largely contributed to the change in air quality. The quality of air is a symptom from exploited natural resources such as combusting fossil fuels, deasil emissions and secondary partial formations. Analysis should therefore focus on the quality of air to estimate the origins and patterns of harmful emissions.

Approach to Problem:

My approach to solving this problem was to firstly gather appropriate dataset that represent all states and counties within US. There is a public dataset[1] available for public use. The dataset contains annual information on the average 2.5 PM for counties and also the exceeded levels of NAAQs permitted levels of gas emissions for several gases qualified as pollutants. 2.5 PM is the particulate matter that have a diameter of less than 2.5 micrometers.

I wanted to initially try and link the air quality dataset with another dataset such as climate change and then compare the spatial effects and patterns of air quality and climate. However multiple datasets had different time stamps for data, and I was unsure of how to align them together. So I decided to simply analyze one dataset.

I have three graphs for visualization that have several versions of the graph for different controlled variable, such as year, and state. The first graph is a spatial heatmap of the United States. The dataset has particulate matter (PM2.5) and National Ambient Quality Standards information per county for all states. So a visual representation of density of air quality measures based on geolocation was a useful graph.

I used ggplot to generate the plot. I had to mutate a column from the air quality dataset to a county maps dataset since I had to overlay a county map dataset on a ggplot and then fill with corresponding measurement value for each county. The vectorization here was simple to match states with the state of the air quality, and using the indexed matching, add corresponding state measured values to a new column in the county data frame. The spatial graphs were for individual years of data. The fill for the ggplot was then the new column on measured values.

This then made it harder to analyze change of particular matter and how much National Ambient Air Quality Standards (NAAQ) were met *over time* since I would need to produce several spatial maps for each year to analyze the air quality. So then an additional graph of heat maps representing only States with changes in % of day exceeding NAAQs standards with years on the x-axis, helped showcase data better. The dataset was altered to have an average of all the counties for each State to represent a single measure for each year. The result of the data was visually appealing yet less informative. The reason was that, one particular state, "California", had longer periods of invoking the NAAQs standards (will be analyzed later) and so the bar was adjusted to fit a larger range of data solely because of one state. In order to obtain some useful information, I adjusted the dataset to filter for percentages of exceeding PMs by a smaller range and eliminating that observation beyond the range. This surfaced much more valuable set of information(will be shown in results section).

The challenge remained that, the average of countries has caused data to be lost for each county, and I wanted to see if I could show the variety of measured values for multiple years and several counties. This directed me toward interactive ggvis plots. So ggvis plots was a challenging task.

I was not very familiar with implementing ggvis, and so suffered some basic problems. I initially had problems with having an older R version which was not updated and so I kept getting several errors, only to realize that the documentation I was looking at was also very old. Searching for tutorial for ggvis, had tutorials from older version as the top list of searches, so I had to take some time to figure around making a simple plot. I was able to plot using ggvis, however, apparently I assumed interactive plots needed another interactive tool. The most popular one with R is Shiny. Shiny allows interactive web apps to be built from R.

So I built a Shiny web app and realize I need to convert it into markdown. The app contains to R files; server.R and the ui.R file. The server.R returns server functions used by ui.R to return a UI objects. The format of shiny was not to difficult and so I had to set up and start the services on RStudio from Shiny. They had an extensive tutorial on the basics.

I then decided to have a slider and a drop down menu for the choices of manipulating the graph. A ggvis plot is generated with the input from the user on the interactive tool that captures a year as an input from the slider and a value to measure and plot using the dropdown menu. The data frame is changed in the server.R to filter for the required inputs. The slider input, the plot and the drop down were implemented in UI.r with some formatting to position the features appropriately. The server.R is where the ggvis plot is implemented.

I then rechanged my code and added it to the markdown file instead, and had a ggvis interactive session. The runtime for my markdown file is “shiny”. This was needed to perform interactive plots.

A ggvis plot layers using a boxplot is made to represent the range of measured value in a State with respect to the County. I did run into several error while trying to generate this data. Mainly because not each County for certain year of measure had a value, in fact, an observation for the specific state would be missing instead of a “NA’ this resulted in me not being able to further add features into the ggvis and rather stick with data that had complete county information. The ggvis plot had some difference from ggplot in the context I used it. I had x as StateName and y as measure Values, and then processed the graph to look presentable. I added several if condition to check for the three measured values of input from the drop down. For which ever the condition is true, the associated measurement type for the measure value is used to filter the data. The default filter starts with “Micograms per cubic meter”.

Tasks Accomplished:

1. I was able to generate spatial maps to show locations where air quality was more affected.
2. I was able to generate a heat map to show the change in air quality factors with progression of years.
3. I was able to create a simple interactive tool to analyze a few of the measurements.
4. I was unable to add more features to the interactive tool to further narrow options and give more results.

Results and Analysis:

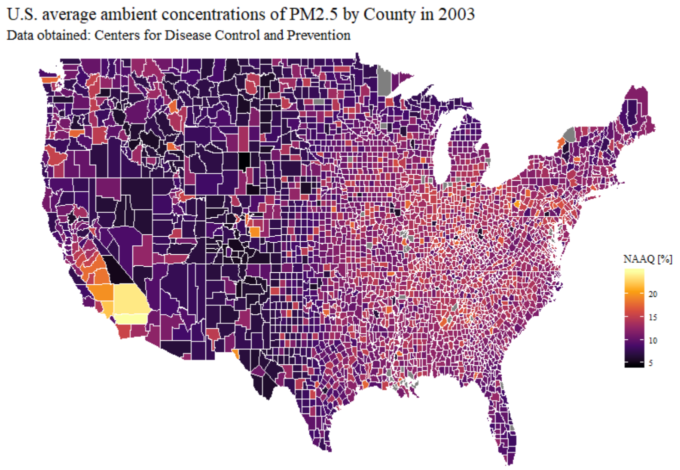

U.S. average ambient concentrations of PM2.5 by County in 2003
Data obtained: Centers for Disease Control and Prevention

Figure 1

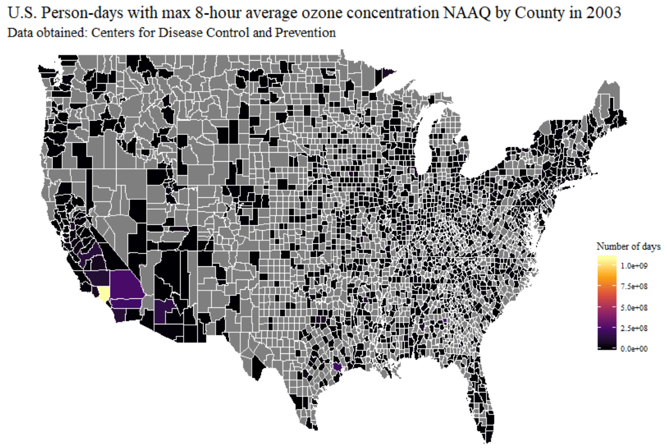

U.S. Person-days with max 8-hour average ozone concentration NAAQ by County in 2003
Data obtained: Centers for Disease Control and Prevention

Figure 2

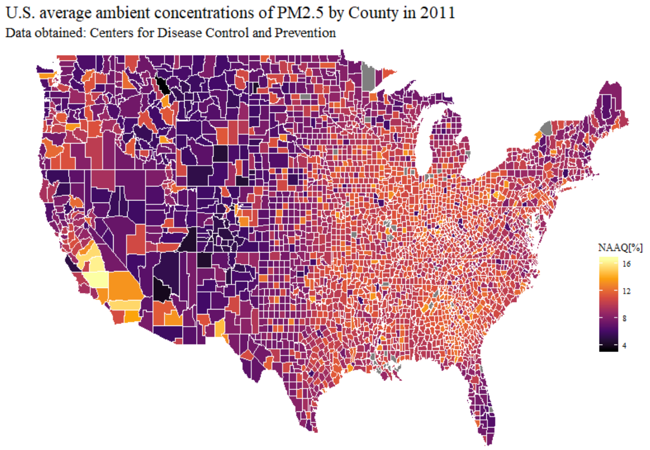

U.S. average ambient concentrations of PM2.5 by County in 2011
Data obtained: Centers for Disease Control and Prevention

Figure 3

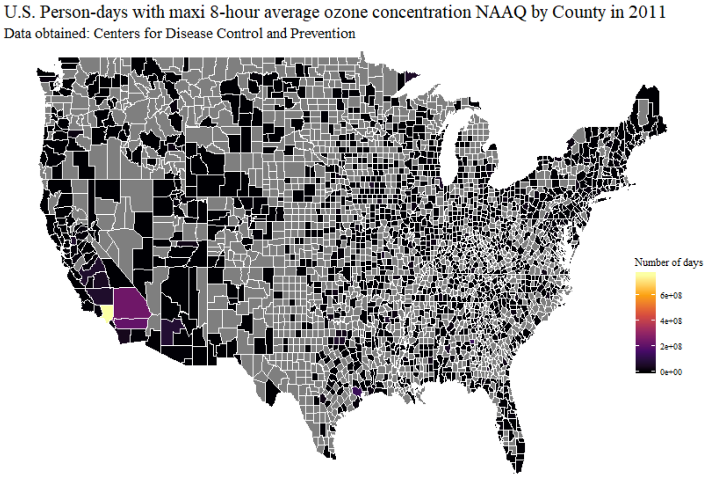

U.S. Person-days with maxi 8-hour average ozone concentration NAAQ by County in 2011
Data obtained: Centers for Disease Control and Prevention

Figure 4

As the spatial heatmaps show, California has significantly longer periods of days that it has invoked the NAAQs standards and also had longer periods of above NAAQs standards of the levels of ozone(Figure 3, 4) emissions. In context of these results, California has mountains and terrains that traps pollutant much more easily and also a warm climate that traps ozone much more. However, California also has a 1:2 ratio of people to vehicle which greatly increases carbon emissions.

Also having had a look at the spatial graphs, it is evident form the bars that from 2003(Figure 1) to 2011(Figure 3, there has been a decreased in the percentage of periods of exceeding NAAQs standards. This indicates, overtime some actions and legislations may have helped reduce the level of pollution all around the United States most probably a legislative law. In order to understand better, heatmap of states with changes with time is shown below.

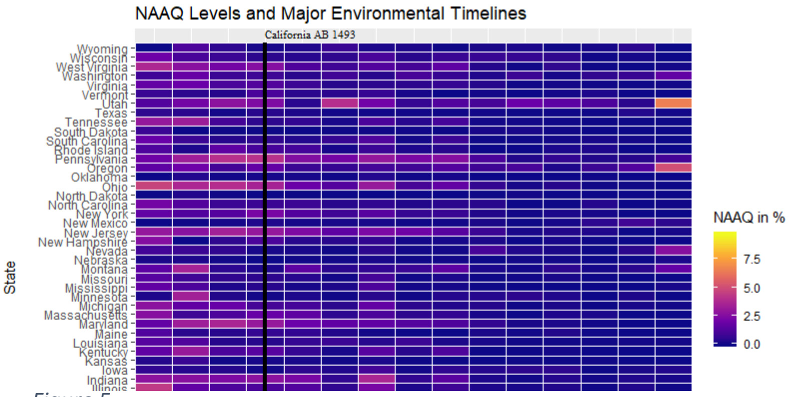


Figure 5

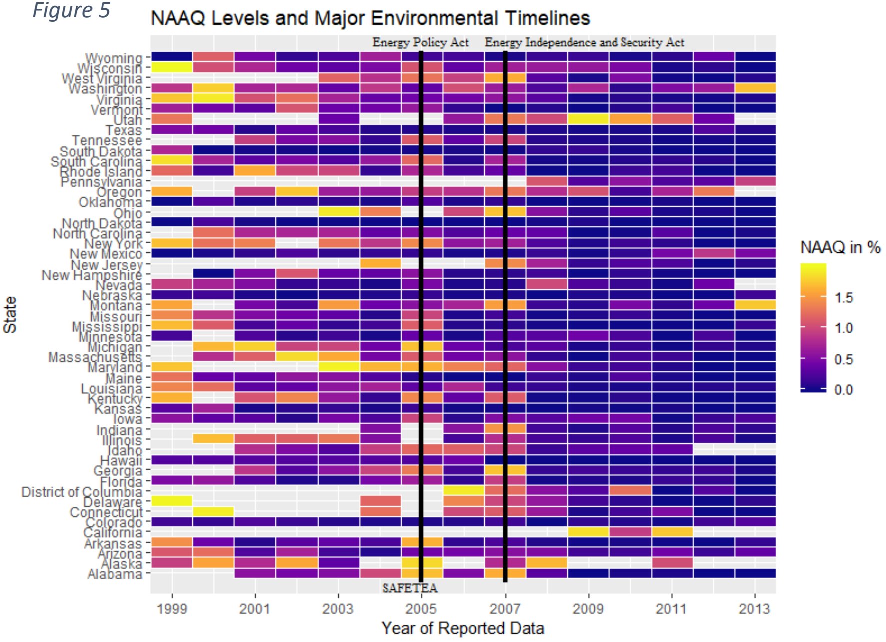


Figure 6

Figure 5 depicts the average period of NAAQs standard invoked for all years recorded in the dataset. The data shows a significant emission of pollutants from California, however interestedly, there is rapid decrease to the next year 2006 (Figure 5). It was imperative to analyze that data and match important legislative acts to see if any laws had an effect. And as you can see there was a law passed specifically for California amended to California AB 1493. The changes made in the California AB 1493 puts some preventative measures on regulating low emission options of vehicles and also passed a bill for reviewing regulations of non-personal vehicles and feasible reduction of greenhouse gasses emitted by passenger vehicles. Therefore, the significant drop is an effect of the changes in legislations for California.

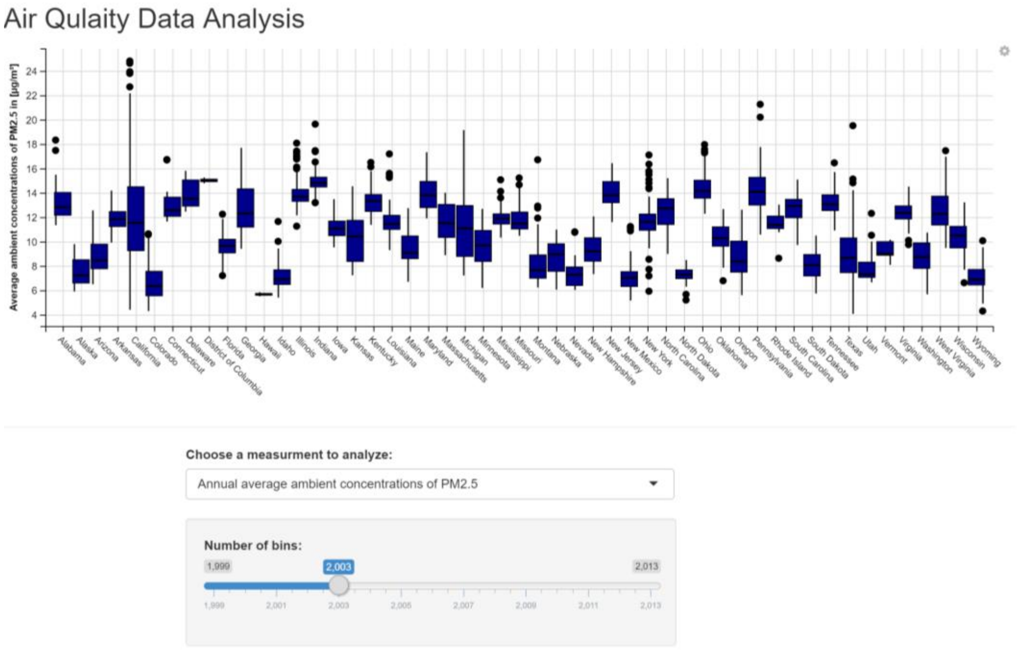
However, no other pair of data seems useful from Figure 5, the same problem exists as with the spatial maps; the data is skewed towards normalizing the range to include California. Figure 6, is an altered heat map with % of period of days exceeding NAAQs between 0 - 2 %. As shown by the year 2007 there was a significant decrease in Air quality measure, and the three other acts passed in 2005 and 2007 are shown.

- Energy Policy Act,
- Safe, Accountable, Flexible, Efficient Transportation Equity Act
- Energy Independence and Security Act,

The hypothesis is, that due to the three acts shown, either of them, or combined, could have had an impact on the decrease.

The interactive ggvis with shiny web app contain three features. A plot, which is altered based on the slider input of year (shown in Figure 7) and the value of measurement to plot. There are three values used to plot graphs, the annual average ambient concertation of PM2.5 and the percentage of days with PM2.5 levels over NAAQs and the number of days with maximum 8-hour average ozone concertation over the NAAQs.

Using the interactive tool, the annual ambient concentrations and the percentage of days with the PM2.5 levels over NAAQs both have a significant drop over the years and is evident in the ggvis plots. The range of ambient concentrations also seems to have narrowed for most states, converging to low concentrations. So overall, there has been a decrease over the years for most counties as well.

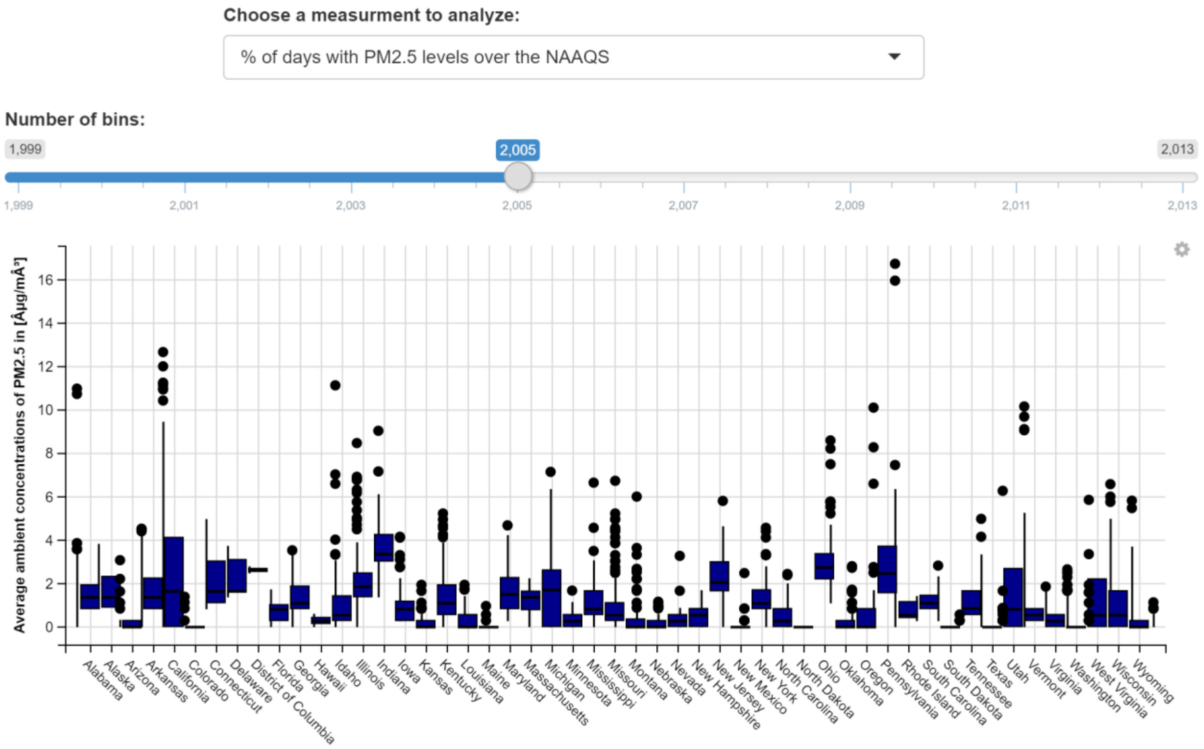

Air Qulaity Data Analysis

Choose a measurment to analyze:

Annual average ambient concentrations of PM2.5

Number of bins: 2,003

Figure 7



Choose a measurment to analyze:

% of days with PM2.5 levels over the NAAQS

Number of bins: 2,005

Figure 8

The Figure 7 above is from running server.R and ui.R(attached as supplementary material), but is very similar to the markdown version shown in Figure 8. The code has been reconfigured into markdown file.

The tools here engineered have simplified and provided quick analysis of the air quality data in the United states. From the spatial graphs, it can be seen that California has higher air quality concerns and that over years there was a decrease for the Unites States. The heat map plotted with ggplot depicts the changing air-quality with time for each state as an average of the counties. The visual information from the ggplot was helpful to then add in the environmental acts that could have possible affected the changes in air quality. The ggvis plot was able to provide a hands-on tool, to study the data, and further analysis the variety of measurements from each county for all the years in the dataset.

Note: run markdown as shiny markdown

References

[1] Data.cdc.gov. (2019). [online] Available at: https://data.cdc.gov/Environmental-Health-Toxicology/Air-Quality-Measures-on-the-National-Environmental/cjae-szjv/data [Accessed 4 May 2019].