U.S. Pollution Dataset Analysis

Peng Yan
Crew
University of Colorado Boulder
peya8291@colorado.edu

Zhixing Chang
Crew
University of Colorado Boulder
chzh5137@colorado.edu

Tianlun Zhao
Crew
University of Colorado Boulder
tizh6070@colorado.edu

Xiaoyang Dou
Crew
University of Colorado Boulder
xido3947@colorado.edu

ABSTRACT

Air pollution is one of the most critical environmental issues. Economic development, urbanization, energy consumption, transportation/motorization, and rapid population growth are major driving forces of air pollution in large cities. Air pollution levels in developed countries have been decreasing dramatically in recent years. However, in developing countries, air pollution levels are still at relatively high levels, though the levels have been gradually decreasing or have remained stable during rapid economic development. The World Health Organization (WHO) published the "WHO Air Quality Guidelines (AQGs), Global Update" in 2006. These updated AQGs provide much stricter guidelines for PM, Nitrogen Dioxide, Sulfur Dioxide and Ozone. Considering that current air pollution levels are much higher than the WHO-recommended AQGs, interim targets for these four air pollutants are also recommended for member states, especially for developing countries in setting their country-specific air quality standards [1]. This dataset deals with pollution in the U.S. Pollution in the U.S. has been well documented by the U.S. EPA. We are trying to resolve several questions about air pollution by using this dataset. For example, what is the main source that causes air pollution in the different city? How does the trend of air pollution during the day? What are the main pollutants in each city? How is the air pollution problem in different cities? Are those cities having better air quality, or worse?

1. INTRODUCTION

With rapid economic development, people start to pay more attention to health issues. Lousy air condition can result in uncomfortable feelings in areas of eyes, nose, throat, and lung. It can not only cause disease like aggravate asthma and other respiratory conditions, but it can also damage the heart and cardiovascular system. Everyone breathe every day, therefore it is essential for people to monitor air quality so that people can protect themselves when necessary.

The air pollution contains four major kinds of gases, which are Nitrogen Dioxide, Ozone, Sulfur Dioxide and Carbon Monoxide. Nitrogen dioxide is a nasty-smelling gas. It is naturally formed by lighting, and some are produced by plants, soil, and water. However, not all nitrogen dioxide in our country's air is formed in this way. The major part of the nitrogen dioxide in our atmosphere is generated by burning fossil fuels like coal, oil, and gas. Since our cars run on fuels, motor vehicle emission is one of the major causes of excess nitrogen dioxide gas. Some other source is manufacturing industries' everyday production activities, such as food processing, electricity, and heat generation from coal-fired stations. This gas irritates respiratory tube. Long time of exposing to it can increase the risks of having lung infections, which can cause the problem of respiratory, higher response to allergens and even heart disease which can be the main problem to death.

Ozone is a gas compound with three atoms of oxygen element. It exists in both upper atmosphere and ground level. The ozone in the atmosphere is harmless but helpful for the whole earth. However, the ozone at ground level is a kind of harmful air pollutant. It is blamed for the ozone pollution. This kind of air pollution is like the nitrogen dioxide, that generated by cars, power plants, industrial activities and some other chemical reactions that happen under the sunlight. Breathing ozone can also trigger various health problems just like what nitrogen dioxide does, with symptoms like coughing, scratchy throat and chest pain.

Sulfur Dioxide and Carbon Monoxide are like the two above we talked about, because of human activities, this two kind of gases are generated and emitted over the value which it should be. They also threaten people's health.

Due to how harmful these air pollutants are to human beings, we are interested in monitoring these gas indexes in our air. The dataset we choose to process is about these four pollutants' everyday condition in the U.S. from 2000 to 2016. Each pollutant's max value and max hour is documented. Other attribute types in our dataset include monitoring site (address, city, state, country), and monitoring date code. By relating the max value with monitoring site, we can see the trend and pattern of these pollutants condition in some particular address, city, and state. We can also relate max hour with monitoring date to each pollutant or relate one pollutant value to another one. Within this dataset, there can be many possibilities we can discover and evaluate. Depending on the goal we have which is trying to have some kinds of solutions for the pollution, we can make visualizations to help achieve it. And furthermore, by visualizing information, we can think deeper like what the reason of the increase or decrease of certain pollutant is, what the cause is, and what we learn from this experience [2].

2. MOTIVATION

It is common for developing countries to suffer from air pollution. London used to be known for its London Fog. California also deals with air quality issue for many years. Our data includes 28 fields (data type include numeric and nominal/string), four major pollutants (Nitrogen Dioxide, Sulfur Dioxide, Carbon Monoxide and Ozone) for every day from 2000 - 2016 in the U.S. Since the developed country went through the bad air condition phase, it has experience and knowledge on dealing with air pollution.

By looking at the dataset and analyzing the U.S. previous 16 years' data about air pollution, we can see many things, and get many questions answered from this dataset. By analyzing each index according to the timeline, we can see the place where air condition has changed drastically and the states where air condition has been stabled over the decades. By relating different attributes, we can get answers to questions like if there is a positive or negative relationship exists between them. For example, we related the mean value of Ozone with the mean value of Nitrogen Dioxide from each year to find the correlation between them. Once we know there is a negative correlation, by only knowing the moving trend of one element, we can easily predict the general trend of the other.

We also have other questions like whether the huge population with rapid growth influences the air quality. If there is, how does it affect on air condition? Is it positive effect or negative?

By making more effort and doing more research, we want to know what kind of methods or regulations government used help better the air quality, or what kind of activities worsens the circumstance. Also, we can conclude the treating methods for other countries that have already been or will be the victim of the air pollution.

3. RELATED WORK

There are many data research about air pollutants in recent years. For example, “openair — an R package for air quality data analysis”. Since 2011, frequent occurrences of haze in China have become a cause for panic and routinely appear as a major topic in the media and on climate websites. Visual exploration of air pollution with spatio-temporal data is a solution that makes complex data understandable because graphical representation is relatively intuitive. However, there are several problems that prevent the widespread use of more insightful analysis. For example, a coherent set of data analysis tools for air pollution purposes does not exist. and, many users are not aware of the tools available or how to apply them. If these problems can be overcome, there are many potential benefits including: a more comprehensive evidence base to support decision making, identification of the factors controlling pollutant concentrations [3].

Through another research paper “Air pollutants measured in Seoul”, we can find the researcher did a lot of data clean work, filling missing values with 0s, filling missing values with mean, filling missing values with interpolation. We also can visualize the data set that could help us do some analysis work [4]. Compare the previous research our data analysis focus on how to help user in different place of U.S. quickly and accurately to find what are the major pollutants and what are the relationship between of them. Then, it can guide people or government formulate corresponding policies.

4. DATA SET

We are using "The U.S. Pollution Data" as our dataset. It includes State Code, County Code, Site Number, Address, State, County, City, Data Local, Nitrogen Dioxide Units, Nitrogen Dioxide 1st Max, Nitrogen Dioxide AQI(The calculated air quality index of Nitrogen Dioxide within a given day), Ozone Units, Ozone Mean, Ozone 1st Max, Ozone AQI(The calculated air quality index of Ozone within a given day), Sulfur Dioxide Units, Sulfur Dioxide Mean, Sulfur Dioxide 1st Max, Sulfur Dioxide AQI(The calculated air quality index of Sulfur Dioxide within a given day), Carbon Monoxide Units, Carbon Monoxide Mean, Carbon Monoxide 1st Max, and Carbon Monoxide AQI(The calculated air quality index of Carbon Monoxide within a given day).

Through this dataset, we can understand deeply about main pollutants and specific pollution index of every American city. Based on the analysis of this dataset we can get the pollution of a heavy trend of a city from 2000-2016. Through our team's textual data analysis and chart trends, decide whether this city should focus on pollution sources and how much efforts they need to take on appropriate measures.

5. MAIN TECHNIQUES APPLIED

This project required us to follow all the aspects of data mining. So we began with cleaning and preprocessing the data to get the million data entries into a more manageable and usable form. After we got many of the valid and complete data organizing that we wanted, we began to focus on the analytical side of data mining. We focused on aspects such as correlations, patterns, and a variety of other techniques.

5.1 Data Cleaning and Preprocessing

The initial U.S. Pollution Data was fairly messy and had a lot of missing data and other various issues. There was over 1.4 million dataset that gathered four major pollutants (Nitrogen Dioxide, Sulfur Dioxide, Carbon Monoxide and Ozone) for every day from 2000 - 2016.

In the U.S. Pollution data, there is a total of 28 fields. The four pollutants (Nitrogen Dioxide, Ozone, Sulfur Dioxide and Ozone) each has five specific columns. It needed a lot of work. Through use the box plot (Figure 1), we found there were many missing values and outliers recorded by the million data of four major pollutants (Nitrogen Dioxide, Sulfur Dioxide, Carbon Monoxide and Ozone). Most of the data were probably automatically recorded daily by data instruments which maybe make mistakes occasionally. Some of the missing values and errors were easily removed/ignored because the data usually included has significant outliers. However, there were many datasets that appeared normal at a glance, but actually, they were outliers that could severely alter our analysis. In some instances, the data had values like N/A indicating no data. In other instances, which were much more difficult to detect, was when a data value for something like the maximum daily Sulfur Dioxide mean value equaled something like above 300 (Figure 2). Comparing with the other data, this data is a bad value and should be ignored, but when having to work with all the various data types, it became increasingly difficult to have to detect these noisy data values. It was not as simple as checking if the value was null, but instead, we had to use our real-world understanding of the pollutants to account for any bad data.

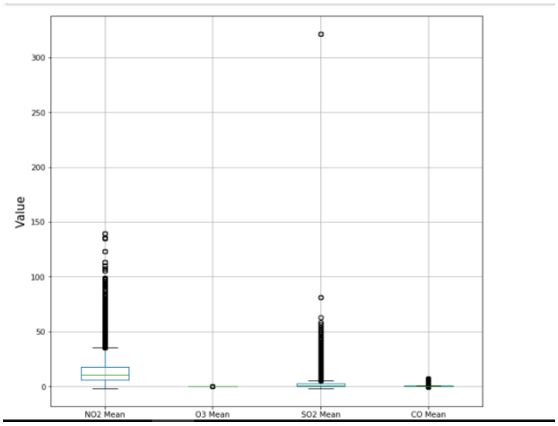


Figure 1

Figure 2

If we wanted to be as precise as possible without bringing in so much of the human error in analysis., if we could have used an outlier technique, the cleaning data would be acceptable. Instead of simply saying 300 was too high for Sulfur Dioxide, we could instead run an algorithm against all other Sulfur Dioxide values found that a value like 300 was too high. If we were to implement this in a more professional and wanted to ensure our results were accurate, detecting outliers would be a more mathematically sound using some algorithm. After cleaning up a lot of the bad data were completely removed from our data sets as. However, there is no need to keep all attributes that would not be used for some analysis.

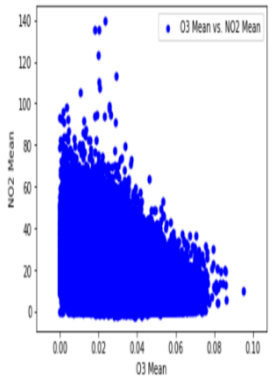
After we finished a lot of the data cleaning, we began to integrate the data. For much of the pollutant data, there were some different values at a different time at one day. In most instances, for things like daily max value or min value. We decided to average the values for any data on the same day. This obviously solved the issue of data redundancy in

terms of having multiple values for a single day, but it also gave us a more accurate data set.

5.2 Correlation

We tried to find the pollutants whether have a correlation. If any two pollutants correlate, there should have some reason lead to this result. We also try to find some special value in some special state. For example, pollutants in California and New York. Because these two states have too many people. The value should be more special than the state has few people.

```
#dfH['Date Local'] = dfH['Date Local'].dt.str.split('-').str[0].astype(int)
ax=dfH.plot(kind="scatter", x="O3 Mean",y="NO2 Mean", color="b", label="O3 Mean vs. NO2 Mean")
plt.show()
```

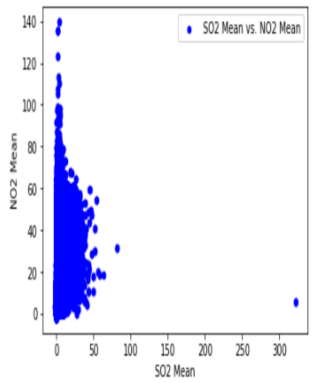



```
corr=dfH['NO2 Mean'].corr(dfH['O3 Mean'])
print(corr)
```

```
-0.432650137437
```

First, we tried to find "Ozone mean" and "Nitrogen Dioxide mean" whether have a correlation. Through the graph and calculations, we found "Ozone mean" and "Nitrogen Dioxide mean" have a low negative correlation. That means when one values increase, the other one has a little decrease.

```
ax=dfH.plot(kind="scatter", x="SO2 Mean",y="NO2 Mean", color="b", label="SO2 Mean vs. NO2 Mean")
plt.show()
```

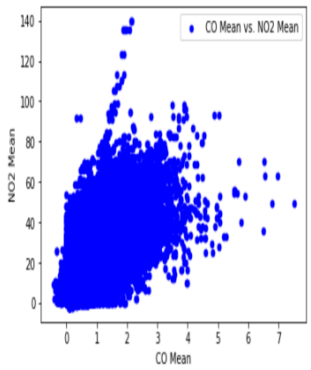



```
corr=dfH['NO2 Mean'].corr(dfH['SO2 Mean'])
print(corr)
```

```
0.348186026785
```

Second, we tried to find "Sulfur Dioxide mean" and "Nitrogen Dioxide mean" whether have a correlation. Through the graph and calculations, we found "Sulfur Dioxide mean" and "Nitrogen Dioxide mean" almost do not have any correlation.

```
dfH.plot(kind="scatter", x="CO Mean",y="NO2 Mean", color="b", label="CO Mean vs. NO2 Mean")
plt.show()
```




```
corr=dfH['NO2 Mean'].corr(dfH['CO Mean'])
print(corr)
```

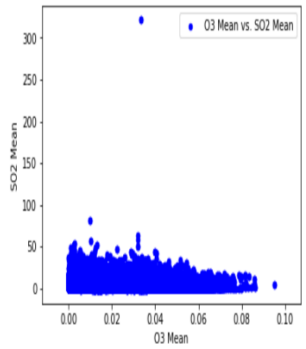
```
0.641828104006
```

Third, we tried to find "Carbon Monoxide mean" and "Nitrogen Dioxide mean" whether have a correlation. Through the graph and calculations, we found "Carbon Monoxide mean" and "Nitrogen Dioxide mean" have a low positive correlation. That

means when one values increase, the other one has little increase.

```
dfH.plot(kind="scatter", x="O3 Mean",y="SO2 Mean", color="b", label="O3 Mean vs. SO2 Mean")

plt.show()
```



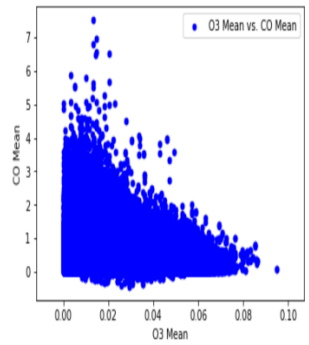

```
corr=dfH['O3 Mean'].corr(dfH['SO2 Mean'])
print(corr)
```

```
-0.110401440276
```

Fourth, we tried to find "Ozone mean" and "Sulfur Dioxide mean" whether have a correlation. Through the graph and calculations, we found "Ozone mean" and "Sulfur Dioxide mean" have no correlation.

```
dfH.plot(kind="scatter", x="O3 Mean",y="CO Mean", color="b", label="O3 Mean vs. CO Mean")

plt.show()
```



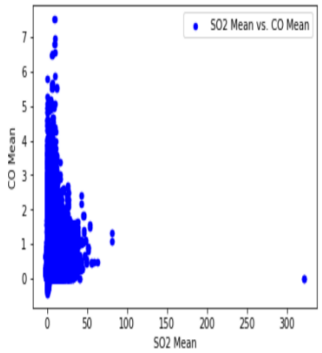

```
corr=dfH['O3 Mean'].corr(dfH['CO Mean'])
print(corr)
```

```
-0.339426430194
```

Fifth, we tried to find "Ozone mean" and "Carbon Monoxide mean" whether have a correlation. Through the graph and calculations, we found "Ozone mean" and "Carbon Monoxide mean" almost have no correlation.

```
dfH.plot(kind="scatter", x="SO2 Mean",y="CO Mean", color="b", label="SO2 Mean vs. CO Mean")

plt.show()
```




```
corr=dfH['SO2 Mean'].corr(dfH['CO Mean'])
print(corr)
```

```
0.215216375849
```

Sixth, we tried to find "Sulfur Dioxide mean" and "Carbon Monoxide mean" whether have a correlation. Through the graph and calculations, we found "Sulfur Dioxide mean" and "Carbon Monoxide mean" almost have no correlation.

5.3 Data Transformation

Four major pollutants max value:

```
State                    Arizona
Date Local    2000-01-19 00:00:00
NO2 Mean                 139.542
Name: 1468, dtype: object
State                          Arizona
Date Local          2000-01-10 00:00:00
NO2 1st Max Value                   267
Name: 1432, dtype: object
State                    Arizona
Date Local    2000-01-10 00:00:00
NO2 AQI                      132
Name: 1432, dtype: object
```

```
State               Pennsylvania
Date Local    2000-06-10 00:00:00
O3 Mean                 0.095083
Name: 77152, dtype: object
State                 Country Of Mexico
Date Local          2007-07-04 00:00:00
O3 1st Max Value                 0.141
Name: 744381, dtype: object
State                  California
Date Local    2013-06-29 00:00:00
O3 AQI                       218
Name: 1364788, dtype: object
```

```
State                  California
Date Local    2000-12-20 00:00:00
CO Mean                   7.50833
Name: 10505, dtype: object
State                        California
Date Local          2000-12-20 00:00:00
CO 1st Max Value                   19.9
Name: 10504, dtype: object
State                  California
Date Local    2000-12-20 00:00:00
CO AQI                        201
Name: 10505, dtype: object
```

```
State                   Oklahoma
Date Local    2006-05-04 00:00:00
SO2 Mean                  321.625
Name: 596159, dtype: object
State                        Oklahoma
Date Local          2006-05-04 00:00:00
SO2 1st Max Value                   351
Name: 596159, dtype: object
State                   Illinois
Date Local    2002-02-14 00:00:00
SO2 AQI                      200
Name: 229463, dtype: object
```

Four major pollutants min value:

```
State                    Kansas
Date Local    2015-11-02 00:00:00
SO2 Mean                0.591667
Name: 1663965, dtype: object
State                    Arizona
Date Local    2000-03-17 00:00:00
SO2 AQI                      NaN
Name: 306, dtype: object
```

```
State                   Colorado
Date Local    2014-03-06 00:00:00
CO Mean                 0.145833
Name: 1513982, dtype: object
State                    Arizona
Date Local    2000-10-09 00:00:00
CO AQI                         1
Name: 1081, dtype: object
```

```
State                  California
Date Local    2007-02-05 00:00:00
O3 Mean                  0.013792
Name: 636115, dtype: object
State                  California
Date Local    2000-12-07 00:00:00
O3 AQI                         8
Name: 12030, dtype: object
```

```
State                Pennsylvania
Date Local    2012-12-11 00:00:00
NO2 Mean                  9.1375
Name: 1299522, dtype: object
State                     Arizona
Date Local    2000-05-16 00:00:00
NO2 AQI                         1
Name: 528, dtype: object
```

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Through calculate Min-max normalization formula we can get each state the Min-max normalization value. This approach, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers. We also can easily compare each state air condition.

6. KEY RESULTS

Most of our results show a very basic idea from our data and analysis. We try to find and analyze the data in a variety of situations. Some things show very clear and concise results. On the other hand, some results were very volatile and showed very little. It would have helped to have more accurate pollutants data instead of just daily values. However, we were still able to find some insightful results with the data sets we had.

6.1 Mean Pollutant Max Value Over Time

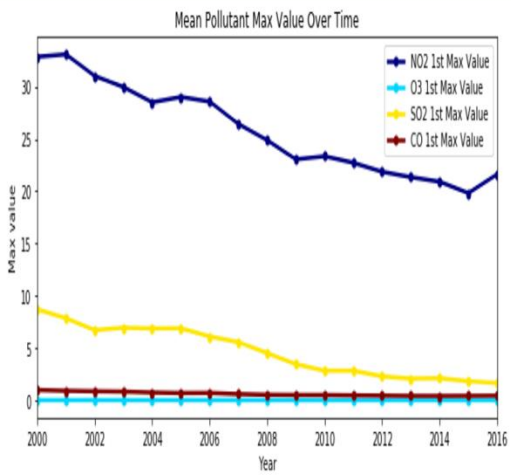


Figure3

First, we want to focus on four major pollutants (Carbon Monoxide, Sulfur Dioxide, Ozone, Nitrogen Dioxide) max value over time (from 2000 to 2016), 1st Max Value is the maximum value obtained for four major pollutants concentration in a given day. The monitoring and analysis of ambient air pollutant concentrations are important aspects of the issue of air quality management. The evaluation of air quality is important both in terms of the intrinsic interest in the levels and variations of pollutant concentrations in the ambient air and in determining compliance with air quality standards. The ambient air quality standards for several pollutants are stated in terms of maximum concentrations not to be exceeded more than once per year. And, in the case of air contaminants, adverse effects depend on the concentration level, as well as the length of time elevated concentrations may persist. Thus, the extreme values of air quality, e.g. the maximum concentration, are of considerable interest[5].

Through the plot (Figure 3), we can see the four major pollutants (Carbon Monoxide, Sulfur Dioxide, Ozone, Nitrogen Dioxide) max value decrease from 2000 to 2016. That means in the recent years U.S. air much better than before. This information makes complete sense logically. Air pollution has decreased even though population and the number of cars on the roads have increased. The shift is the result of regulations, technology improvements, and economic changes. This result also can answer our one question in this project we sought. We can easily predict the general trend in the future years the U.S. air condition will be better year by year.

6.2 Mean Pollutant AQI Over Time Comparison California and New York

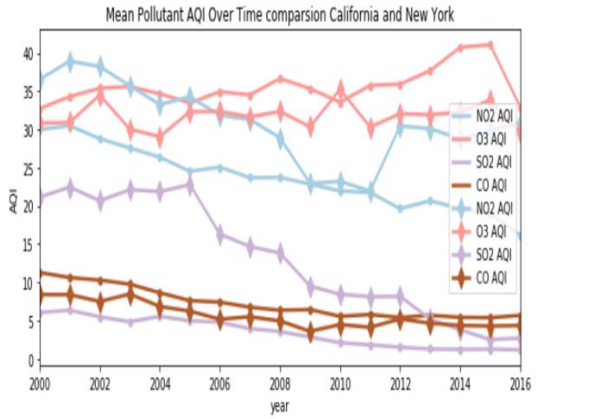


Figure 4

Second, we focus on four major pollutants (Carbon Monoxide, Sulfur Dioxide, Ozone, Nitrogen Dioxide) AOI over time (from 2000 to 2016) at some important state. Because California and New York are important two states and there are a lot of elements could affect the air condition. The AQI values of these two states are very representative.

The AQI is an index for reporting daily air quality. It tells you how clean or polluted your air is, and what associated health effects might be a concern for you. The AQI focuses on health effects you may experience within a few hours or days after breathing polluted air. EPA calculates the AQI for five major air pollutants regulated by the Clean Air Act: ground-level ozone, particle pollution (also known as particulate matter), carbon monoxide, sulfur dioxide, and nitrogen dioxide. For each of these pollutants, EPA has established national air quality standards to protect public health. Ground-level ozone and airborne particles are the two pollutants that pose the greatest threat to human health in this country.

"Good" AQI is 0 to 50. Air quality is considered satisfactory, and air pollution poses little or no risk. "Moderate" AQI is 51 to 100. Air quality is acceptable; however, for some pollutants, there may be a moderate health concern for a very small number of people. For example, people who are unusually sensitive to ozone may experience respiratory symptoms[6].

Through the data plot (Figure 4), we can see four major pollutants values under 50 and decrease in recent years. However, we also can see the pollutants Ozone and Nitrogen Dioxide a little higher than the other two pollutants (Sulfur Dioxide and Carbon Monoxide). Comparing these two states, we can find New York AQI value lower than California. That

means New York's air better than California. There are many reasons for this result. For example, California's population, cars number and manufactory number higher than New York. We can easily predict the general trend in the future years California and New York's air condition will be better year by year.

6.3 Pollutants AQI Over Time Comparison California, New York, Colorado, Florida

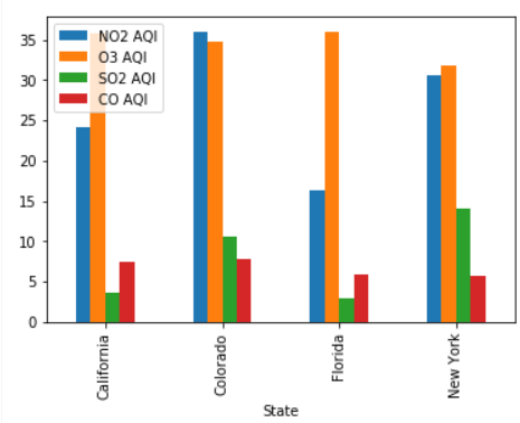


Figure 5

First, we put the four major pollutants (Nitrogen Dioxide, Sulfur Dioxide, Nitrogen Dioxide, Carbon Monoxide) and four states (California, New York, Colorado, Florida) in one graph (figure 5). We can transverse compare the four major in these four states. And we found Nitrogen Dioxide and Ozone are the major pollutants in these four states and Colorado air condition is the worst in these four states. Then we do the longitudinal comparison.

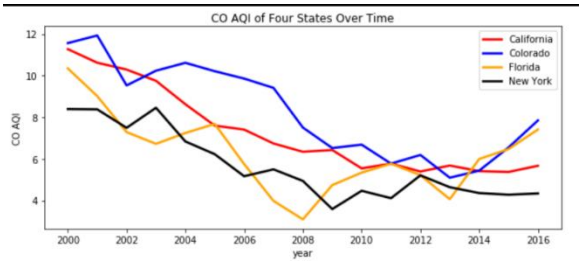


Figure 6

This graph (Figure 6) shows the Carbon Monoxide AQI of four states, which are California, Colorado, Florida, and New York, from 2000 to 2016. Each line has a different color from the others to represent the Carbon Monoxide AQI of one state. As we can see from the legend that the blue line is for Colorado, the red one is for California, the yellow one is for Florida and the black one is for New York. The general trend of these four lines is that they all mainly slope downwards at first but then they hit a minimum point at some time and they bounce back from there.

For Colorado, the Carbon Monoxide AQI shows decreasing trend from the year 2000 to 2013. There are minor fluctuations during the period from the year 2001 to 2004, and the period from the year 2009 to 2013, but the main trend is decreasing. The reason we say that the main trend is decreasing is that each time the fluctuation is over, the ending AQI is lower than the AQI before this fluctuation. That is, as we can see from the graph, the Colorado Carbon Monoxide AQI at the year 2004, which is after the fluctuation, is lower than that at the year 2001, that is the time before the fluctuation. The turning point for Colorado is the year 2013. After the year 2013, its Carbon Monoxide AQI goes back up. Till 2016, the AQI has gone back to the same value as the year 2008. That is the pollution that takes five years (2008-2013) to go down only uses three years (2013-2016) to go back up.

New York successfully lowers Carbon Monoxide pollution from the year 2000 to 2009 with a fluctuation during the period from 2002 and 2004 and the period from 2006 to 2009. Still, the main direction of the black line is going downwards. New York has its lowest Carbon Monoxide pollution in the year 2009. After 2009, Carbon Monoxide AQI has increased for three years until 2012. The Carbon Monoxide pollution gets better again, but at a slower rate than before.

California has the smoothest line. Its Carbon Monoxide AQI has been decreasing from 2000 to 2010 and it approaches its steady value six at 2010. And from there, the value changes in small amount around the value 6.

Contrary to California, Florida is super volatile. But still, we can see a general trend, which is that 2008 is the division year of the decrease and increase.

California and Colorado have overall higher Carbon Monoxide AQI compared to Florida and New York.

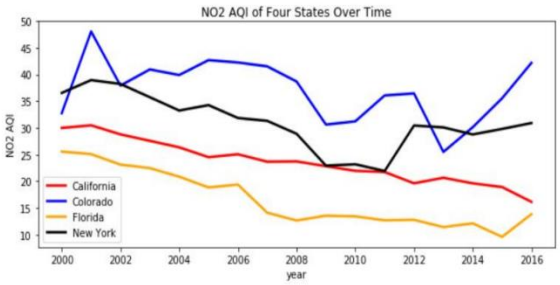


Figure 7

This graph (Figure 7) shows the Nitrogen Dioxide AQI of four states, which are California, Colorado, Florida, and New York, from 2000 to 2016. Each state is assigned a line with specific color. Unlike the Carbon Monoxide AQI graph we analyzed above, we know which state intends to have the worst Nitrogen Dioxide condition over time based on this graph. As we observed, the blue line which represents Colorado has higher Nitrogen Dioxide AQI compared to other three states except there is two short period that it gets lower than New York. Florida has been the lowest for the entire documented period, which means that it has best air condition if we only consider the quantity of Nitrogen Dioxide in the air.

The overall trend for the Nitrogen Dioxide AQI of both Florida and California is going downwards during the period, which means the air condition has been consistently improving. However, Colorado and New York are more volatile. Colorado's Nitrogen Dioxide AQI has its peak at the year 2001 and then fluctuates between the 35 and 45 until the year 2013, at which the index goes down even more and reaches its lowest value. But this value only the last a year, it goes back up the next year at a faster rate. The AQI at the year 2016 is close to the highest value of this documented period as we can see from the graph.

Colorado and New York have overall higher Nitrogen Dioxide AQI compared to California and Florida.

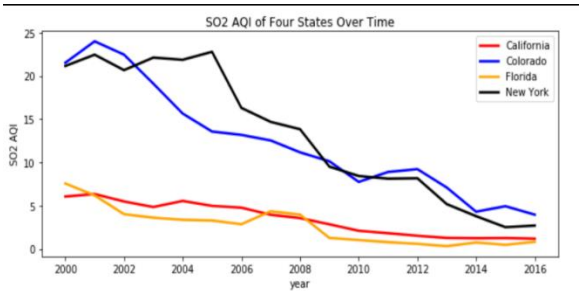


Figure 8

This graph (Figure 8) shows the Sulfur Dioxide AQI of four states, which are California, Colorado, Florida, and New York, from 2000 to 2016. Each colored line corresponds to a state. It is quite obvious in this graph that Colorado and New York have overall higher Sulfur Dioxide AQI compared to Florida and California. The difference between them gradually becomes smaller over the years. And the reason for this change is that Colorado and New York have decreased their Sulfur Dioxide quantity at a faster speed than other two states. Colorado and New York have their value at the beginning several years between 20 and 25, which is much higher than Florida and California, which have values between 5 to 10. And at the end of this record, Colorado has dropped their Sulfur Dioxide AQI to values between 5 and 10, that is 10 to 15unit drop compared to 16 years ago. For Florida and California, they have dropped less than 5 unit. And therefore, all four states have similar Sulfur Dioxide AQI at the end of this record.

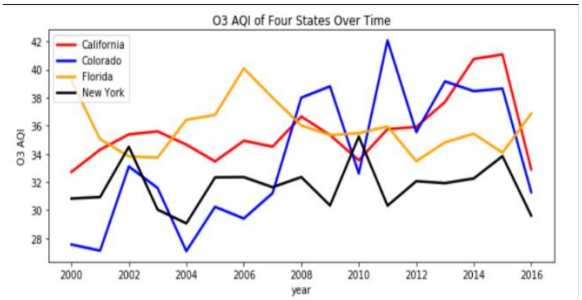


Figure 9

This graph (Figure 9) shows the Ozone AQI of four states, which are California, Colorado, Florida, and New York, from 2000 to 2016. All lines that represent different state are mixed. They all fluctuate in a range. Florida has the worst Ozone condition overall before the year 2008 and then Colorado takes this place after. Colorado has the most volatile Ozone AQI as we can tell from the graph. It has been inconsistently increasing from the year 2000 and 2015, which means the overall trend is going up but there are some backward. It has the best Ozone condition at the beginning of this record but then it becomes highest after the year 2008. And between the year 2015 and year 2016, it dramatically drops from 39 to 31, which is close to the beginning point.

New York has its Ozone value fluctuate between 29 and 35. California has most its values between 32 and 38. However, it has values outside this range after the year 2013 and goes back to normal at the year 2015. And Florida has its immoral high values after the year 2004 and goes back to the normal range in the year 2008.

6.3 Mean Pollutant AQI for The United States from 2000-2016

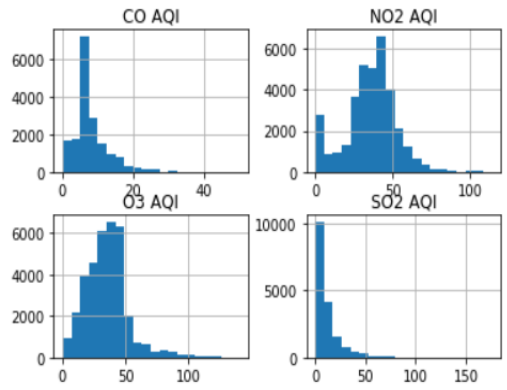


Figure 10

According to the above plot (Figure 10), it shows the main pollutants (Carbon Monoxide, Sulfur Dioxide, Ozone, Nitrogen Dioxide) AQI over time (from 2000 to 2016) at the United States. The horizontal axis represents the value of contaminant AQI, and the vertical axis represents the number of occurrences from 2000-2016.

In general, we found that the Carbon Monoxide AQI value is relatively low, it is mainly distributed in 5-10. This proves that the United States has better control over Carbon Monoxide. The value of Nitrogen Dioxide AQI seems to be high; it is mainly distributed in 25-50. This proves that the United States needs to take further control of Nitrogen Dioxide. The value of Ozone AQI seems to be relatively high; it is mainly distributed in 12.5-50, which reflects that Ozone has a great impact on the environmental pollution in the United States. We need to pay attention to the control of Ozone. The value of Sulfur Dioxide AQI seems to be relatively low; it is mainly distributed in 0-8.5, which reflects the United States' control of Sulfur Dioxide is very effective.

Through the above analysis, Ozone has become the main pollutant that the United States and even the world should pay attention to. Ozone pollution near the ground mainly comes from human production and life. Ozone produces a photochemical reaction between nitrogen oxides and volatile organics that requires sufficient light and temperature. As long as air is involved in the combustion, nitrogen oxides are produced. The volatile organics are derived from industrial emissions, motor vehicle exhaust, decoration, paint, etc. A certain concentration of Ozone in the air will first affect the health of the human respiratory system. In addition to the respiratory system, Ozone can also cause damage to the body's nervous system and skin. High concentrations of ozone can cause dizziness, headache, irritation of the eyes, decreased vision, and memory loss. Therefore, we should pay attention to control of Ozone and call on the government and people to work out an effective plan to reduce Ozone AQI in the future.

7. APPLICATIONS

7.1 Benefits for Our Life

Through air pollution analysis, we can get a lot of benefits. These benefits include:

Healthier Living: these data analysis could help people plan activities and routes around current air quality conditions. While this is of benefit to everyone, it also has some special value to someone who has asthma, allergies or other respiratory conditions. And it can improve transportation-related air quality for all citizens: if enough travelers alter their routes to avoid pollution hot spots, those hot spots will cool down.

Citizen Awareness & Engagement: A second major benefit is citizen awareness of conditions in their community. The more the population is aware of air quality problems, the more likely they are to translate their concerns into modified behaviors, and political engagement. They can also be of help in identifying particular trouble spots or pollution sources that may never surface without 'boots on the ground' making discoveries [7].

7.2 Ozone Reuse

However, these pollutants can also be used valuable. For example, Ozone reuse, Ozone has stronger sterilization, oxidation, decolorization, and deodorization. It is precisely because of these basic properties that Ozone has been increasingly used in the production and life of human beings. Ozone has entered the production and life of human beings extensively. The sterilization and sterilization function of Ozone has many advantages such as broad spectrum, high efficiency, high cleanliness, convenience, economy, environmental protection, etc. It has been widely used for sterilization in workshops, warehouses and public places; fruits and vegetables, food preservation; fish, meat, Decomposition of residual poisons such as eggs, vegetables, and fruits is eliminated; sterilization and preservation of domestic water; deodorization and deodorization of toilets and dumps; chemical waste treatment and many other fields.

In addition, with the development of the economy, the wastewater generated in industrial production has increased year by year, and the direct discharge of wastewater will cause environmental pollution, which has constituted a serious ecological and life threat to many regions of the world. Ozone is one of the strongest oxidants in natural materials, and its oxidizing power is second only to highly corrosive fluorine. It can oxidize and degrade organic substances

such as pesticides and dyes into non-toxic carbon dioxide and water, eliminate hazardous substances in wastewater, and does not generate new secondary pollutants. While decomposing pollutants, ozone also has the functions of decolorization, deodorization, and sterilization, and it really counts. At present, the ozone oxidation method has been applied in the treatment of papermaking wastewater, printing and dyeing wastewater, refinery wastewater, and coking wastewater[8].

Therefore, we can make good use of regions and cities with high Ozone AQI values to achieve Ozone reuse, turning pollutants into important resources for improving environmental pollution.

7.3 Future Air Condition Detection

By analyzing AQI of four pollutants in four states in the United States, we covered important facts like which state have the best air condition, and for each state, when it has the worst pollution and when it recovers from those conditions. There are applications we can get from these facts. During analysis, we notice that for some pollutants in the certain state, the AQI fluctuates in some pattern or some certain range. We can use this fact to detect air condition in the future. For example, if we already know the range this pollutant in California normally stays in, then if the AQI gets out of this range, we should be aware and figure out what causes the problem, human activities or the Earth itself. Something we could do next step is to use the timeline of those AQI and relate them to what happened in these states, such as regulations about environmental protection issued by the state. The benefit of this is to know what helps with improving air condition and what does not. Therefore, we can learn from the experience and encourage the effective regulations and protests.

8. Visualization

link to a video demonstrating the interactive site：
https://screencast.com/t/qbGOfZSv

9. REFERENCES

[1] Environ Health Prev Med. 2008 Mar; 13(2): 94–101. Air pollution and population health: a global challenge Published online 2008 Feb 28. doi: 10.1007/s12199-007-0018-5 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2698272/

[2] U.S. Pollution Data Pollution in the U.S. since 2000 https://www.kaggle.com/sogun3/uspollution

[3] openair — an R package for air quality data analysis David C. Carslaw,∗, Karl Ropkins King's College London, Environmental Research Group, Franklin Wilkins Building, 150 Stamford Street, London SE1 9NH, UK Institute for Transport Studies, University of Leeds, LS2 9JT, U http://www.openair-project.org/pdf/openair_paper_preprint.pdf

[4] Air pollutants measured in Seoul https://www.kaggle.com/jihyeseo/seoulairreport

[5] Estimating the maximum value of autocorrelated air quality measurements Dept. of Chemical and Environmental Engineering, Rensselaer Polytechnic Institute, Troy, NY 12181, U.S.A. https://ac.elscdn.com/0004698182903419/1s2.0-0004698182903419- main.pdf?_tid=504e1efe-a95a-4fe7-b7a2-5aa4b2ebab77&acdnat=1524776109_93b6e0ed4a3a5af25b6a918a1196a232

[6] Air Quality Index (AQI) Basics https://airnow.gov/index.cfm?action=aqibasics.aqi

[7] Open Data & Air Pollution: A Powerful Tool in the Struggle for Cleaner Air, Laura | March 14, 2017 https://www.opendatasoft.com/2017/03/14/open-data-air-pollution-a-powerful-tool-in-the-struggle-for-cleaner-air/

[8] Technical Institute of physics and Chemistry, Chinese Academy of Sciences http://www.ipc.cas.cn/kxcb/kpwz/201503/t20150304_4317094.html