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 important environmental issues. Economic development, urbanization, energy consumption, transportation/motorization, and rapid population growth are major driving forces of air pollution in large cities, especially in megacities. Air pollution levels in developed countries have been decreasing dramatically in recent decades. 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, NO2, SO2 and O3. 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 try to use this data find answers about some question. For example, what is the main source that cause air pollution in different city? How does the trend of air pollution during the day? What are the main pollutants in each city? Which cities' pollution problem is getting better? Which cities' is getting worse?

1 INTRODUCTION

With rapid economic development, people start to pay more attention to health issues. Bad 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, it can also damage heart and cardiovascular system. Everyone breathe everyday, therefore it is important 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 NO2, O3, SO2 and CO. NO2 is Nitrogen dioxide, 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 are 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 a main problem to death.

O3 is ozone, a gas compound with three atoms of oxygen element. It exists in both upper atmosphere and ground level. The ozone at 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.

SO2 and CO are like the two above we talked about, because of human activities, these two kind of gases are generated and emitted over the value which it should be. They also threat 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 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 pollutino, 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 MOTIVATION

It is commonly 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, Sulphur Dioxide, Carbon Monoxide and Ozone) for every day from 2000 - 2016. U.S. Since the developed country went through the bad air condition phase, it has experience dealing with air pollution.

By analyzing the U.S. past 16 years air condition data, we can learn from it. We can see many things from this dataset. We can get many questions answered by using this dataset. By analyzing each index according to the timeline, we can see which place has changed drastically from old air condition and which has been stable over the decade. By relating different attributes, we can get answers for questions like if there is a positive or negative relationship exists between them. For example, we relate the mean value of O3 with the mean value of NO2 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 has its influence on the air quality. If there is, what is the effect on air condition? Positive one or negative one?

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. In addition, we can conclude the treating methods for other countries that has already been or will be the victim of the air pollution.

3 PROPOSED WORK

·Data collection

We found the "U.S. Pollution Data" from Kaggle website by Brenda. There are more than millions of data points in this data set. This is the main source we will use. Then, we are going to use some other useful datasets about pollutants to prove our analysis.
The dataset named "Air Quality Annual Summary", this is a summary of air quality from 1987 to 2017 by US Environmental protection Agency. It has more details about the U.S. pollution.

·Data Cleaning

The data sets from U.S. Pollution data filtering algorithm and stored in an .csv file. We need to extract these data to the form that we will use. This step will reduce after manipulation of the attribute subset of the data set.

Especially the most populous states, for example, California, New York, Florida, etc. These states have a large population and good economic development, which may lead to serious environmental pollution. This will lead to a sharp rise in NO2, O3, SO2, and CO.

We will also specifically analyze the state of pollution in Colorado because we are now living in Colorado. We have an obligation to understand and protect the environment in Colorado and to find useful ways to solve environmental pollution problems through analysis.

·Data Reduction

The U.S. Pollution data of the original data set according to time order to show the value of pollution change. Every object has a lot of different properties, some of them is unnecessary for this project. The first step in the data reduction is to reduce the number of attributes. Such as, state code and country code. We clearly know the dataset about the U.S. and the specific name of the state and city, so we don't really need the state code and county code. We can integrate them.

·Transforming

We need to use these data set to make pollution tendency graph that can help us to analysis the pollution problem in The United States and it looks more intuitive and easier to understand.

·Histogram

The stable factors are the arithmetic mean of concentration of pollutants and the calculated air quality index of pollutants from 2000-2016 in different cities of U.S.

4 DATA SET

Our data is coming from The U.S. Pollution Data. It includes State Code, County Code, Site Number, Address, State, County, City, Data Local, NO2 Units, NO2 1st Max, NO2 AQI(The calculated air quality index of NO2 within a given day), O3 Units, O3 Mean, O3 1st Max, O3 AQI(The calculated air quality index of O3 within a given day), SO2 Units, SO2 Mean, SO2 1st Max, SO2 AQI(The calculated air quality index of SO2 within a given day), CO Units, CO Mean, CO 1st Max, and CO AQI(The calculated air quality index of CO within a given day).

Through this dataset, we can have detailed understanding every American city of main pollutants and specific pollution index. Based on the analysis of these data we can get the pollution of 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 take appropriate measures.

5 Evaluation

We have done calculating the results of the correlations between each two kinds of pollutants. We have found that between “O3 Mean” and “SO2 Mean” have no correlations, “SO2 Mean” and “CO Mean” have little correlations, “O3 Mean” and “CO Mean” almost have little correlations, and “SO2 Mean” and “NO2 Mean” have little correlations. Therefore, we do not consider and evaluate the relationships among these pairs.

We have found that “O3 mean” and “NO2 mean” have negative correlation. When we have negative correlation, we can assume that when the O3 decrease, the NO2 increase. The main reason for this result should be that the O3 is a reactant and react with some other kinds of gas, during this process, the NO2 can be produced, or the opposite way, the NO2 produce O3. O3 is a kind of gas in atmosphere which is a good kind, but if it is being at where we live, it becomes harmful. NO2 is also a kind of pollutant, so we want to reduce the amount of both these two pollutants at the same time. Because of the equilibrium of the reaction which means the reactant can produce same amount of the produces. This does not mean we will have NO2 or O3 produces same amount of the other gas. Therefore, we need to find a balance which can have the least amount of both NO2 and O3, or we have all of them become O3 and send all the O3 into the atmosphere.

Through the graph and calculations between “CO mean” and “NO2 mean”, we found that “CO mean” and “NO2 mean” have low positive correlation. That means when one values increase, the other one has little increase. So the only thing we want to do is the reduce the amount of both of these pollutants.

6 TOOLS

The dataset is in csv file, so the mainly tool we have used is excel which contains tables and many programs that can help us to get some main ideas about the whole data. We have look through the whole dataset and decided to get the scatter plot and the correlations between each two kinds of pollutants. We also decided to look deeper after getting the graph of place with pollutant growth trends and the time with pollutant growth trends.

After we got the idea of how to deal with our dataset and what we want to do, we use the Python 3 to calculate the different results and draw the graphs. We have use the “numpy”, an import of Python 3, which can really help us getting different kinds of results such as mean values and correlations. With the results we get, we can start evaluate the reason for those results by thinking and combining with the idea we got from the website or the previous research.

We have also use the Tableau which is a tool for making visualizations. We have used the Tableau to make some similar graph with small amount of data to check and making sure that the graph we got is correct. We also use this small amount of data to make some creative graphs which may help us get some interesting ideas on how to analyze our dataset. For now, we have used the Python 3 to get the correlations between each two kinds of pollutants. And made the scatter plots to have a much more obvious view of the relationship between these pollutants. We have also made a graph about the max value and a graph about the AQI trend.

7 MILESTONES

We will follow below table to do our project:

Project dates	Description	Process
2/27/2018	Project part1 slides	Done
3/6/2018	Project part2 Proposal paper	Done
3/7/2018-4/9/2018	Data cleaning some data and look for some data correlation. Also find some special value whether have relationship.	Done

4/10/2018-4/20/2018	Compare the related data and built graph and analyze the reason.	Keep working
4/20/2018-5/2/2018	Try to find some other interesting data and make some new graph that can get some more interesting information.	Keep working

7.1 What We Have Achieved So Far

After this part of the project, we already finish researching of the past report about the "US Pollution" dataset, and we will get the main idea about the part and ways we want to analyze this dataset. We also developed a series of queries from the data, and transfer it into a new CSV file.
We also did some data analysis:
1. Four main pollutants correlation analysis.
- "O3 mean" and "NO2 mean" have have low negative correlation.
- "CO mean" and "NO2 mean" have low positive correlation.

2. Analyze four main pollutants max value trend over time.
- We found pollutants max value were decrease in recent years. The trend is decrease.

3. Compare some important state AQI change trend.
- Compare California and New York mean AOI for 2000 to 2016
- analysis the reason why some pollutants have big different.

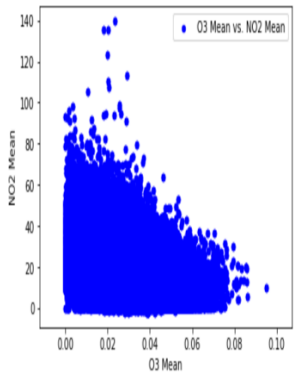
7.2 What Remains to be done

1. We will do more data cleaning on our dataset to find some special data which can be interesting and help us to get some better results or solutions in the future. We will analyze the the reason for the results as well.
2. We will extract some data from important states. With the comparison and the contraction of different states, and making the trend graph for them, we can find the trend and analyze the reason that leads to this trend.
3. We will continue using different graphs to analyze the trends of pollutants and observing how these pollutants change.
4. We will try to use geographical graph to analyze the pollutants.
5. We will analyze how the pollution values change in different time periods.
6. In addition, we will run analysis to determine which pollutants are the most predictive of relationships given the object and the subject information.

8 RESULTS SO FAR

Our result so far tried to find the pollutants whether have correlation. If any two pollutants have correlation, 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()
```

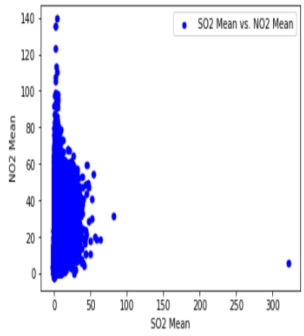



```
corr=dfH['NO2 Mean'].corr(dfH['O3 Mean'])
print(corr)
```

```
-0.432650137437
```

First, we tried to find "O3 mean" and "NO2 mean" whether have correlation. Through the graph and calculations, we found "O3 mean" and "NO2 mean" have low negative correlation. That means when one values increase, the other one has little decrease.

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

plt.show()
```

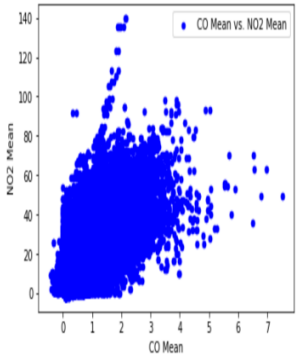



```
corr=dfH['NO2 Mean'].corr(dfH['SO2 Mean'])
print(corr)
```

```
0.348186026785
```

Second, we tried to find “SO2 mean” and “NO2 mean” whether have correlation. Through the graph and calculations, we found “SO2 mean” and “NO2 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()
```

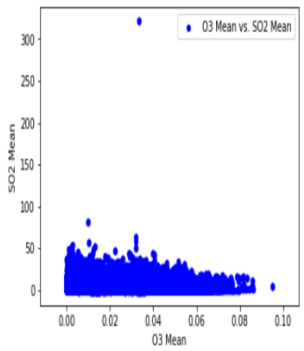



```
corr=dfH['NO2 Mean'].corr(dfH['CO Mean'])
print(corr)
```

```
0.641828104006
```

Third, we tried to find “CO mean” and “NO2 mean” whether have correlation. Through the graph and calculations, we found “CO mean” and “NO2 mean” have 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()
```

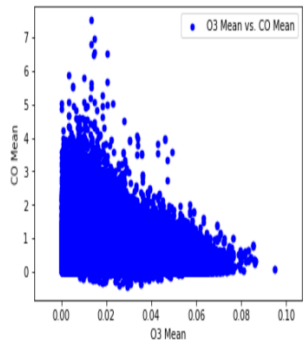



```
corr=dfH['O3 Mean'].corr(dfH['SO2 Mean'])
print(corr)
```

```
-0.110401440276
```

Fourth, we tried to find “O3 mean” and “SO2 mean” whether have correlation. Through the graph and calculations, we found “O3 mean” and “SO2 mean” have no correlation.

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

plt.show()
```

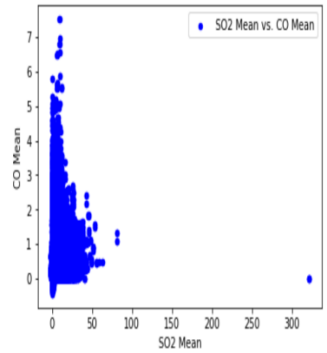



```
corr=dfH['O3 Mean'].corr(dfH['CO Mean'])
print(corr)
```

```
-0.339426430194
```

Fifth, we tried to find “O3 mean” and “CO mean” whether have correlation. Through the graph and calculations, we found “O3 mean” and “CO mean” almost have no correlation.

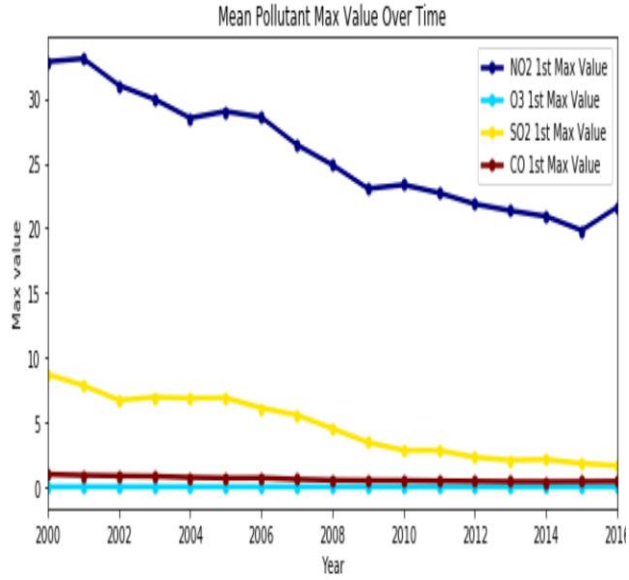
```
dfH.plot(kind="scatter", x="SO2 Mean",y="CO Mean", color="b", label="SO2 Mean vs. CO M

plt.show()
```



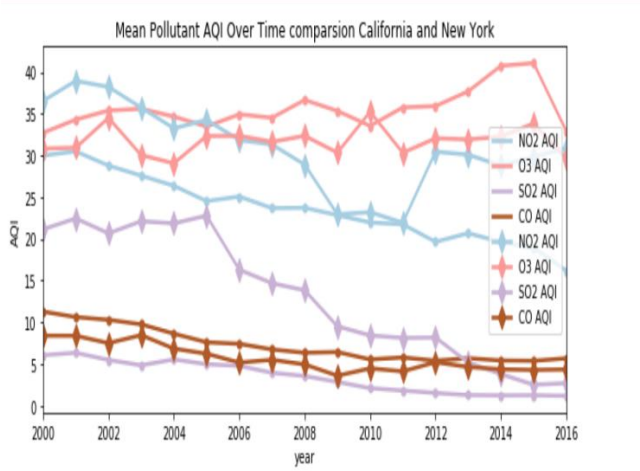

```
corr=dfH['SO2 Mean'].corr(dfH['CO Mean'])
print(corr)
```

```
0.215216375849
```

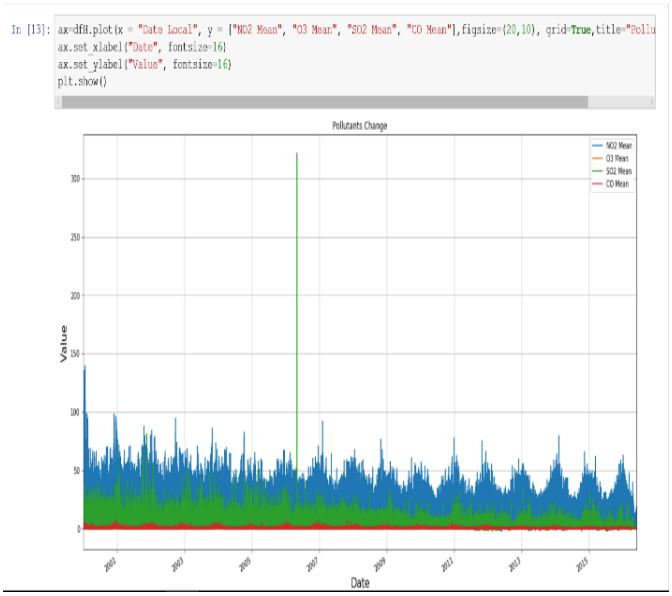
Sixth, we tried to find “SO2 mean” and “CO mean” whether have correlation. Through the graph and calculations, we found “SO2 mean” and “CO mean” almost have no correlation.



Through this graph, we can find some useful information. We can see these four pollutants max value were decrease in recent years. That means the air quality become better.



Compare California and New York mean AOI for 2000 to 2016



This graph we use line chart to show four pollutant mean value from 2000 to 2016.

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. 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

3. Air pollutants measured in Seoul
https://www.kaggle.com/jihyeseo/seoulairreport