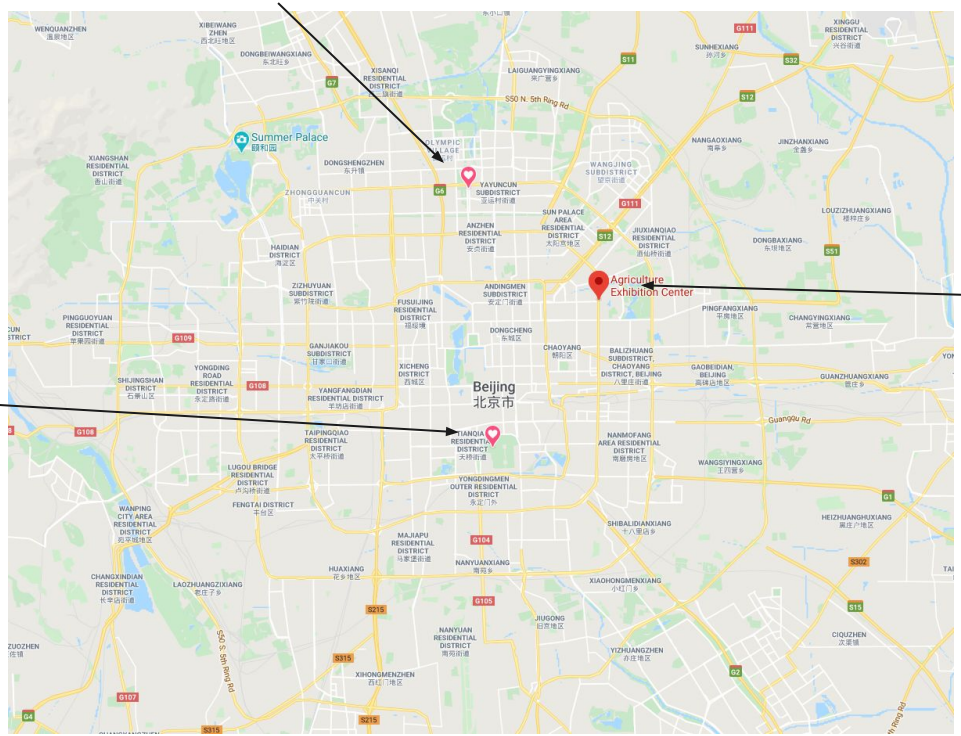
Data Analysis Report on Beijing Greenhouse Gases

By Zhengyang Zhang, Xingzhi Xu

Introduction

- We are interested in relations between greenhouse gases and weather, i.e, would temperature affect the density of greenhouse gases?
- Since Beijing is considered to be one of the cities with the most air pollution, we decided to investigate the relations based on its data. From Kaggle, we obtain a dataset called “Beijing Multi-Site Air-Quality Data Set”, which has the measurement of 6 air pollutants at multiple sites in Beijing.
- Although the original dataset has 12 different sites, many of them are close to each other. Thus, we decided to use only 4 of them:
    - Data from Aotizhongxin (Olympic Sport Center), located at the north of Beijing
    - Data from Huairou, located at the far north of Beijing
    - Data from Tiantan (Temple of Heaven), located at the center of Beijing
    - Data from Nongzhanguan (Agriculture Exhibition Center), located at the east of Beijing

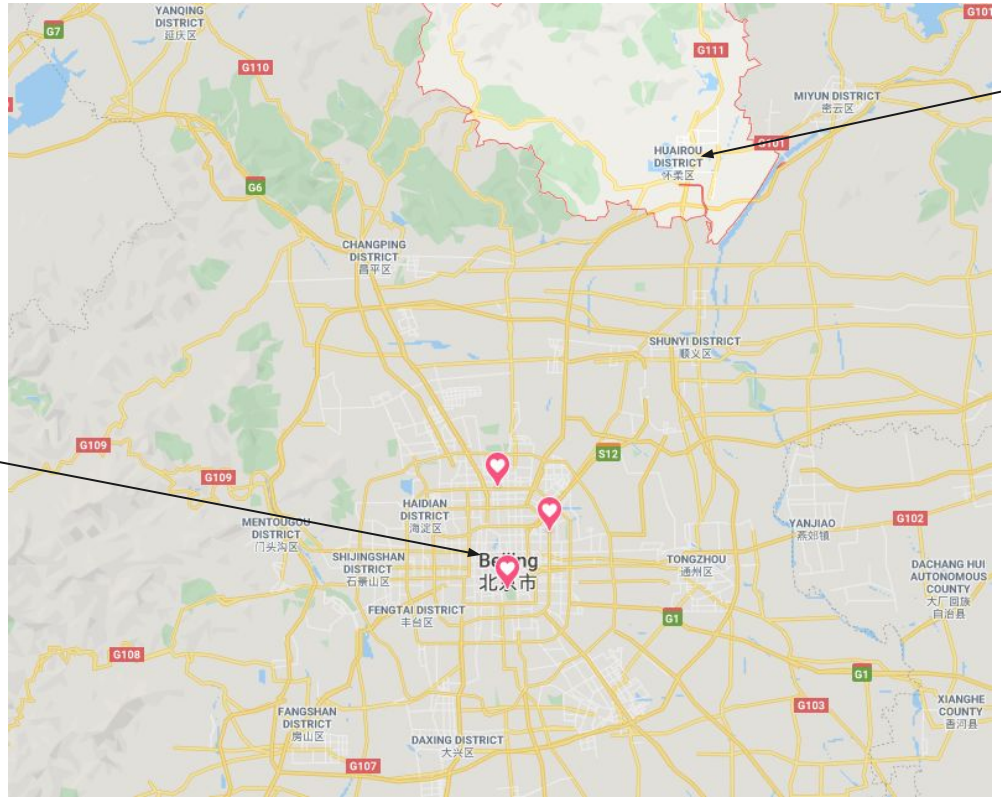
Data from Aotizhongxin, located at the north of Beijing



Data from Nongzhanguan, located at the east of Beijing

Data from Tiantan, located at the center of Beijing

Data from Huairou, located at the far north of Beijing



Urban area of Beijing, with the markers of the other three sources

Data Description

Each site has its own dataframe, and each column has 35064 items (Every hour from March 1st, 2013 to February 28th, 2017)

```
df_atzx.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35064 entries, 0 to 35063
Data columns (total 18 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   No       35064 non-null  int64
 1   year     35064 non-null  int64
 2   month    35064 non-null  int64
 3   day      35064 non-null  int64
 4   hour     35064 non-null  int64
 5   PM2.5    34139 non-null  float64
 6   PM10     34346 non-null  float64
 7   SO2      34129 non-null  float64
 8   NO2      34041 non-null  float64
 9   CO       33288 non-null  float64
 10  O3       33345 non-null  float64
 11  TEMP     35044 non-null  float64
 12  PRES     35044 non-null  float64
 13  DEWP     35044 non-null  float64
 14  RAIN     35044 non-null  float64
 15  wd       34983 non-null  object
 16  WSPM     35050 non-null  float64
 17  station  35064 non-null  object
```

Data from Aotizhongxin

Descriptive Statistics (Part 1)

Here is an description of the dataframe from Aotizhongxin. We have one dataframe from each station.

	No	year	month	day	hour	PM2.5	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	WSPM
count	35064.000000	35064.000000	35064.000000	35064.000000	35064.000000	34139.000000	34346.000000	34129.000000	34041.000000	33288.000000	33345.000000	35044.000000	35044.000000	35044.000000	35044.000000	35050.000000
mean	17532.500000	2014.662560	6.522930	15.729637	11.500000	82.773611	110.060391	17.375901	59.305833	1262.945145	56.353358	13.584607	1011.846920	3.123062	0.067421	1.708496
std	10122.249256	1.177213	3.448752	8.800218	6.922285	82.135694	95.223005	22.823017	37.116200	1221.436236	57.916327	11.399097	10.404047	13.688896	0.910056	1.204071
min	1.000000	2013.000000	1.000000	1.000000	0.000000	3.000000	2.000000	0.285600	2.000000	100.000000	0.214200	-16.800000	985.900000	-35.300000	0.000000	0.000000
25%	8766.750000	2014.000000	4.000000	8.000000	5.750000	22.000000	38.000000	3.000000	30.000000	500.000000	8.000000	3.100000	1003.300000	-8.100000	0.000000	0.900000
50%	17532.500000	2015.000000	7.000000	16.000000	11.500000	58.000000	87.000000	9.000000	53.000000	900.000000	42.000000	14.500000	1011.400000	3.800000	0.000000	1.400000
75%	26298.250000	2016.000000	10.000000	23.000000	17.250000	114.000000	155.000000	21.000000	82.000000	1500.000000	82.000000	23.300000	1020.100000	15.600000	0.000000	2.200000
max	35064.000000	2017.000000	12.000000	31.000000	23.000000	898.000000	984.000000	341.000000	290.000000	10000.000000	423.000000	40.500000	1042.000000	28.500000	72.500000	11.200000

Data from Aotizhongxin

Descriptive Statistics (Part 2)

- “Year”, “month”, “day”, “hour” are the representation of time
- “PM2.5”, “PM10”, “SO2”, “NO2”, “CO”, “O3”, are the six main air pollutants in ug/m^3 (micrometer per meter cubed)
- “Temp” is temperature in degree Celsius, “Pres” is air pressure in hPA, “DEWP” is dew point temperature in degree Celsius, “Rain” is precipitation in milimeter, “WSPM” is wind speed in m/s (meter per second).

	No	year	month	day	hour	PM2.5	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	WSPM
count	35064.000000	35064.000000	35064.000000	35064.000000	35064.000000	34139.000000	34346.000000	34129.000000	34041.000000	33288.000000	33345.000000	35044.000000	35044.000000	35044.000000	35044.000000	35050.000000
mean	17532.500000	2014.662560	6.522930	15.729637	11.500000	82.773611	110.060391	17.375901	59.305833	1262.945145	56.353358	13.584607	1011.846920	3.123062	0.067421	1.708496
std	10122.249256	1.177213	3.448752	8.800218	6.922285	82.135694	95.223005	22.823017	37.116200	1221.436236	57.916327	11.399097	10.404047	13.688896	0.910056	1.204071
min	1.000000	2013.000000	1.000000	1.000000	0.000000	3.000000	2.000000	0.285600	2.000000	100.000000	0.214200	-16.800000	985.900000	-35.300000	0.000000	0.000000
25%	8766.750000	2014.000000	4.000000	8.000000	5.750000	22.000000	38.000000	3.000000	30.000000	500.000000	8.000000	3.100000	1003.300000	-8.100000	0.000000	0.900000
50%	17532.500000	2015.000000	7.000000	16.000000	11.500000	58.000000	87.000000	9.000000	53.000000	900.000000	42.000000	14.500000	1011.400000	3.800000	0.000000	1.400000
75%	26298.250000	2016.000000	10.000000	23.000000	17.250000	114.000000	155.000000	21.000000	82.000000	1500.000000	82.000000	23.300000	1020.100000	15.600000	0.000000	2.200000
max	35064.000000	2017.000000	12.000000	31.000000	23.000000	898.000000	984.000000	341.000000	290.000000	10000.000000	423.000000	40.500000	1042.000000	28.500000	72.500000	11.200000

Data from Aotizhongxin

Goals

Here are some of the relations we would like to investigate:

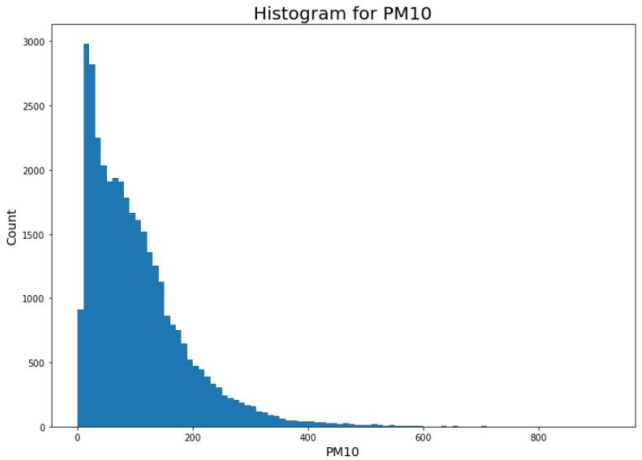
- The relationship between PM2.5 and PM10
- The relationship between O3 and temperature, between O3 and wind speed, and between O3 and rain
- The relationship between CO and burning coal in the winter
- Greenhouse Gases Hourly Changes

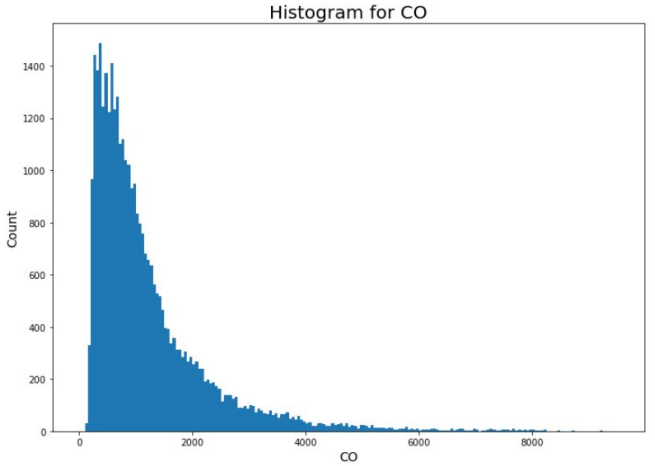
Data Cleaning

- In all four csv files, they all contain the same columns and the same number of rows but with different data. But we find there are missing values in each of the datasets. We want to find the best way we can do fill these missing values and avoid biases.
- We first tried to drop all the rows with missing value, but we ended up with more than 7000 dropped rows. We are not satisfied with this method because 20% of the data would lose forever.
- In the end, we decided to fill the missing value by the average of the same cell from the other three datasets. After we did that, we found out there are still about 200 rows with missing values, and we decide to drop these 200 rows, which is only 0.5% of our data.

Distribution of air pollutants

All six of the air pollutants have right-skewed distribution. Five of them has the average that are below 105 ug/m^3. However, the average for CO is 1227.07 ug/m^3, which is 10 times higher than PM10, the second highest pollutant.





The X-axis for PM10 is at most 800 ug/m^3, but the X-axis for CO is higher than 8000 ug/m^3.

Relationship Between PM2.5 and PM10

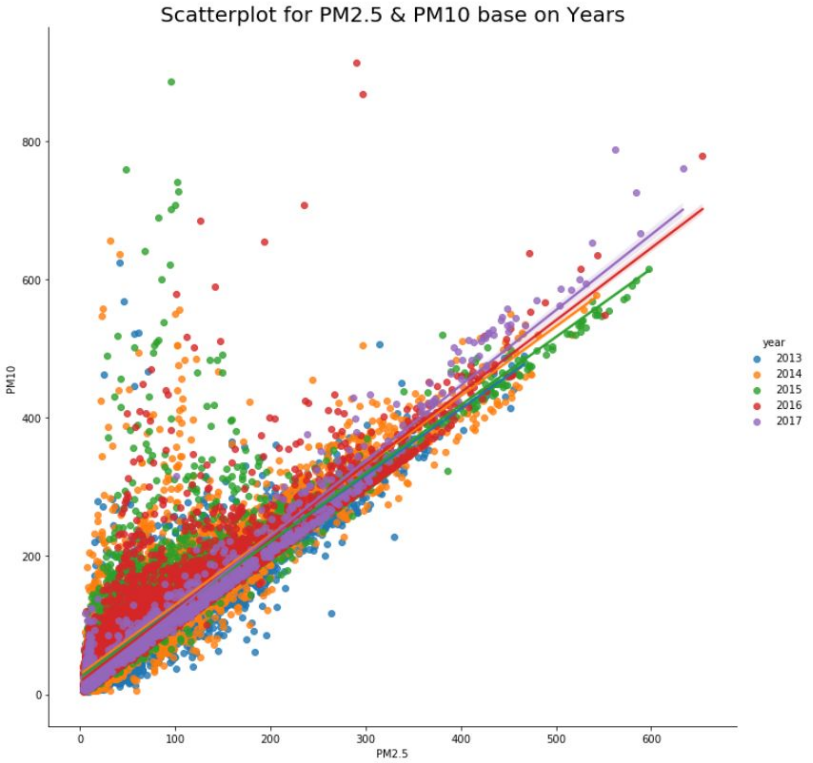
- We are interested in the relationship between PM2.5 and PM10 base on the years. By definition, PM2.5 is particulate matter less than 2.5 micrometers, and PM10 is particulate matter less than 10 micrometers. Thus, we would like to know how many pollutants of PM10 are from PM2.5.

- We decided to do a scatter plot and use the year variable to distinguish the point.

Definitions of PM2.5 and PM10 are from http://www.npi.gov.au/resource/particulate-matter-pm10-and-pm25

Relationship Between PM2.5 and PM10

- As we can see from the scatterplot that every year the PM2.5 and PM10 presented a linear relationship. The linear regression line slightly increases as time goes on.
- Since from the previous parts, we know that the average of PM2.5 is 79.58 ug/m^3 and the average of PM10 is 104.27 ug/m^3. **We conclude that 80% of the PM10 pollutants are less than 2.5 micrometers.**
- However, the proportion of PM2.5 in PM10 is gradually decreasing year by year. Thus, more and more particles larger than 2.5 micrometers have been released.



We do notice there are some outliers where PM2.5 is higher than PM10. We think it is due to errors in the Dataset.

Relationship Between O3 and Temperature

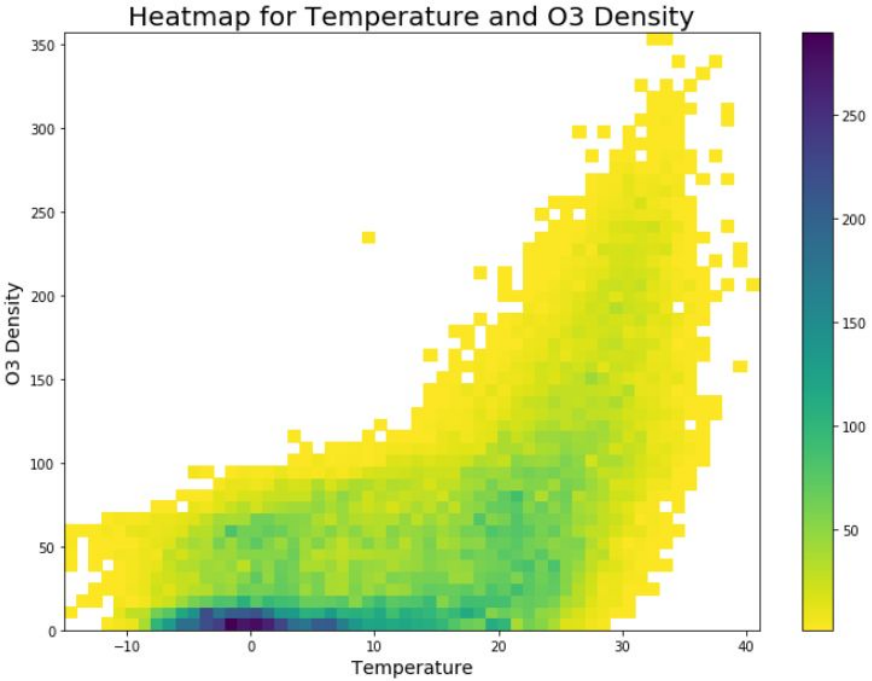
- We are interested in how the temperature could affect the density of greenhouse gases.

- In this case, we did a heat map to visualize the relationship between O3 and greenhouse gases. O3 is the chemical compound of ozone. Most of the ozone are produced by motor vehicle exhaust fumes.

Definition of O3 is from https://www.environment.gov.au/protection/publications/factsheet-ground-level-ozone-o3

Relationship Between O3 and Temperature

- According to the heatmap above, we can see that the density of O3 is usually less than 100 ug/m^3 when the temperature is less than 10 degrees Celsius. Once the temperature rises higher than 10 degrees Celsius, the O3/temperature ratio starts to increase exponentially.
- **We thought that there is an exponential relationship between Temperature and O3.** But we want to know more about other factors.



O3, Wind Speed, and Rain

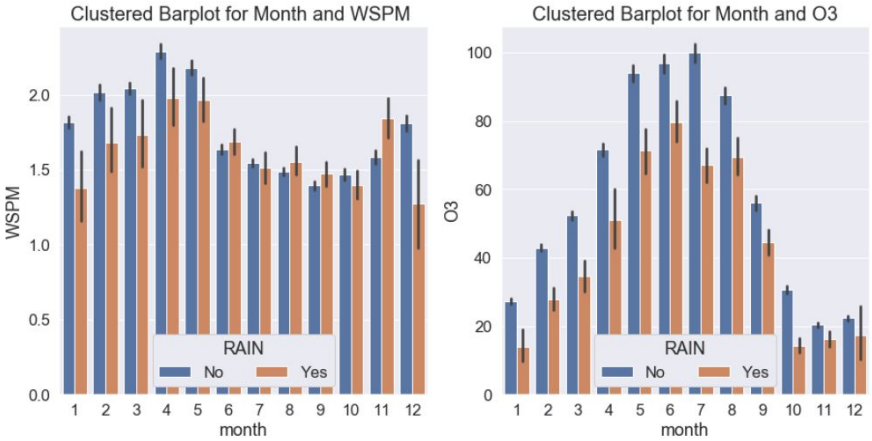
- We also want to find out the relationship between O3 and other factors, such as wind speed and rain.

- O3 is formed when nitrogen oxides (pollutants emitted from motor vehicles) react with sunlight. Initially, we believe that wind would blow the particles away from the urban area in Beijing to other places, which results in less O3 in the urban area. We also believe that rain could be a major factor. We think heavy clouds would block sunlight and halt the formation of ozone, and the rain will wash the particles away.

- We decide to do two cluster barchart to show this relationship.

O3, Wind Speed, and Rain

- From the clustered bar chart, we thought there is a correlation between wind speed and the concentration of O3. Particularly, as the wind speed becomes lower in June, July, and August, the concentration of O3 becomes higher. However, we **cannot conclude that the wind plays a major factor**, because we have high wind speed and high density of O3 at the same time in May.
- We can also see that non-raining day mostly has high O3 density, while raining day has low O3 density. The difference can be as great as 30% in July. Thus, we conclude that **there is a relationship between rain and the concentration of O3.**



The Actual Relationship between O3 and weather

- We want to dig more about the reason behind the high wind speed and the high O3 density in May.

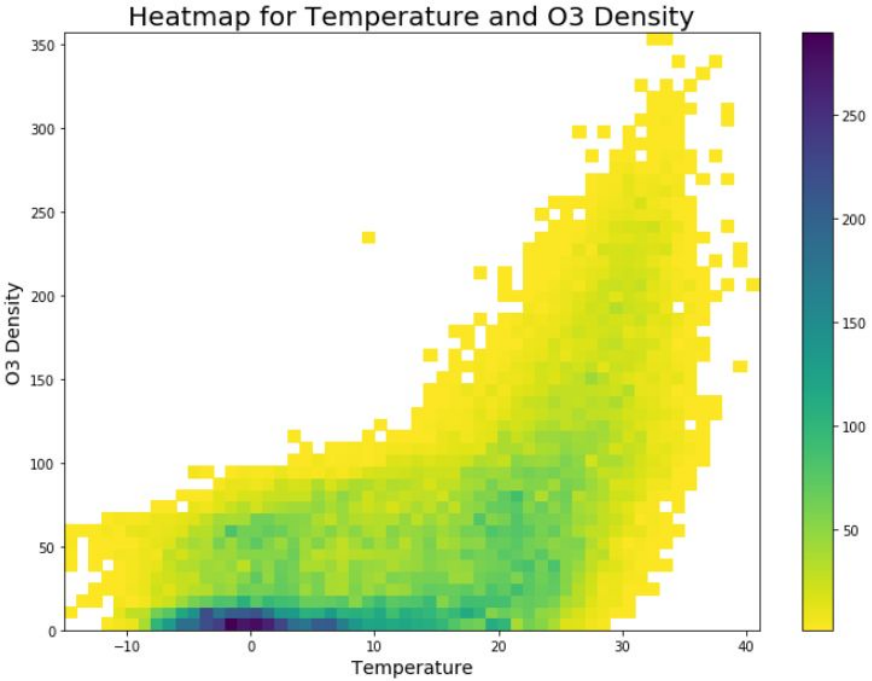
- After some research, we found out that there is actually another major factor to the ozone pollution: air conditioner (which is not in the resource from Department of Agriculture in Australia).

- This finding proves that the relationship between O3 and temperature is **not exponential** but something else.

Source:https://www.baltimoresun.com/news/bs-xpm-1992-06-17-1992169278-story.html

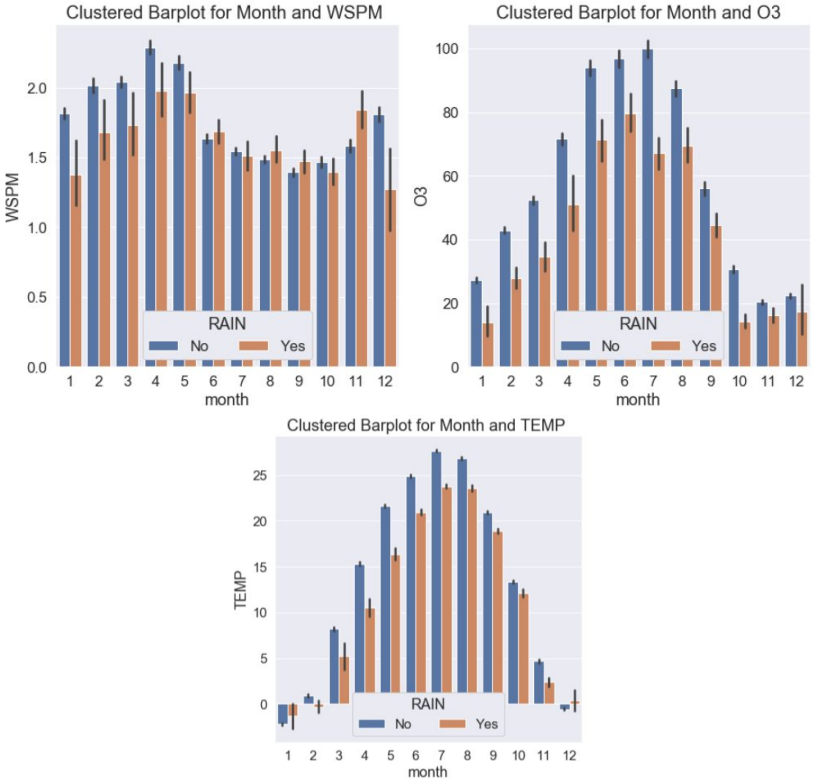
The Actual Relationship between O3 and weather

- We think that the density of O3 is stable below 15 degrees Celsius. It is because there is only one source of pollution: vehicle emission. This source produces constant amount of pollution everyday.
- Once the temperature goes beyond 15 degrees, air conditioner, another source of pollution, starts operating and produces more O3.
- Thus, there is **no relationship between O3 and temperature when temperature is below 15 degrees.** There is a **positive linear relationship when the temperature is above 15 degrees.**



The Actual Relationship between O3 and weather

- At the same time, we discovered the reason behind the high wind speed and the high O3 density in May. It is because the average temperature rises above 15 degrees, and therefore people start to use AC.
- **Our conclusion that “the rain will wash the particles away” still holds**, because the density of O3 is always lower during raining days.



CO and burning coal in the winter

- We're interested in why CO density is much higher compares to other greenhouse gases. We believe it's because Beijing City is burning coal to supply heat to the city during the winter.

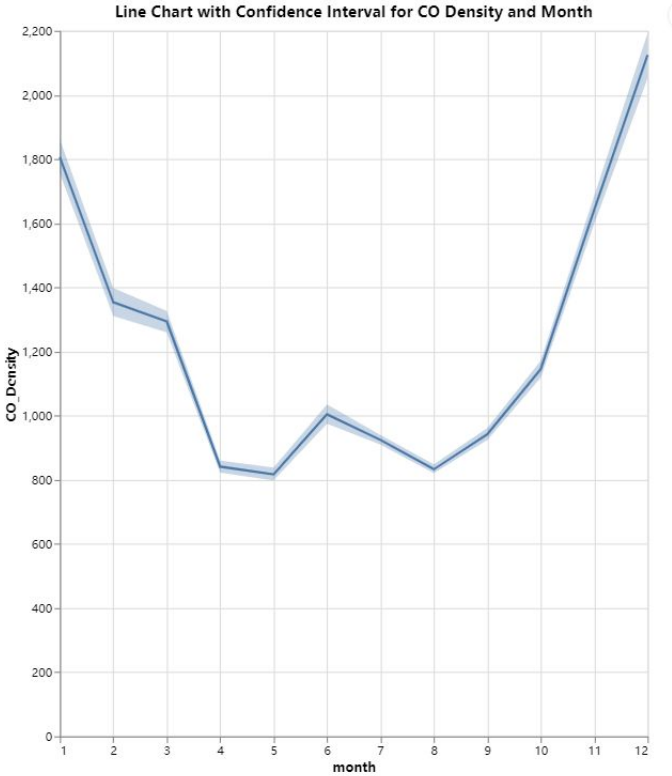
- We would like to do a line chart with confidence interval to see if there is a potential trend in the increase during winter months and observe its confidence interval changes throughout different months.

Definition of CO is from https://www.epa.gov/co-pollution

We learned Chinese Heating System from https://www.scmp.com/news/china/society/article/3037119/chinas-plan-reduce-use-coal-heating-northern-homes-still-facing

CO and burning coal in the winter

- This line chart with confidence interval confirms our theory on the CO increases in the air. We can see there is a one times increase in CO density from October to December, and we believe it's because of burning coal to supply heat. Beijing city didn't change its heat supply source to natural gas until 2017.
- We also find out that the confidence interval for February is quite big compare to other months, we believe it's because the weather is not steady during this month, usage heat varies during the month.
- Thus, we can conclude that **burning coal is the major source contributing to CO pollution.**



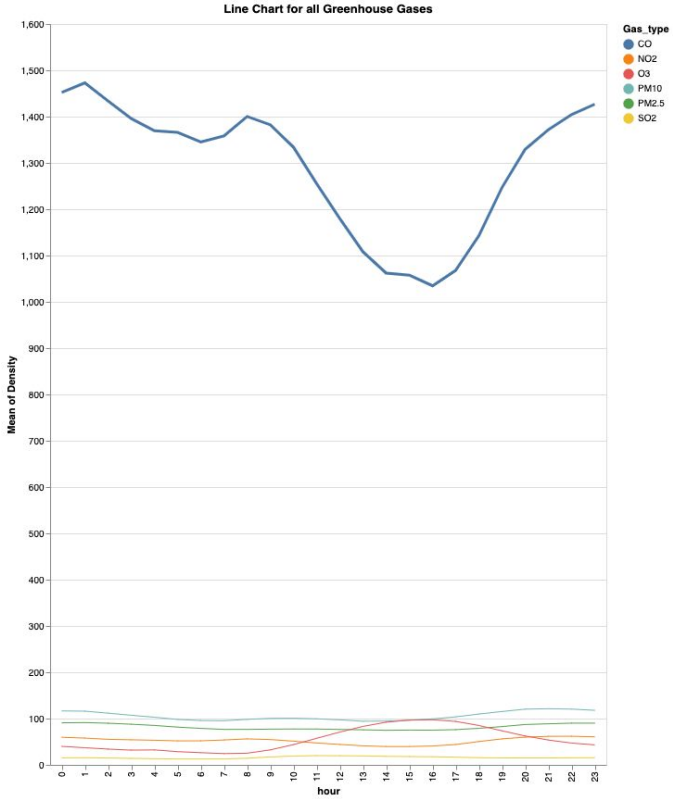
Greenhouse Gases Hourly Changes

- We have investigate several monthly changes for greenhouse gases, but we cannot ignore hourly changes for them as well.
- We decide to make a line chart for all greenhouse gases and see what is going on.

Line Chart for all Greenhouse Gases

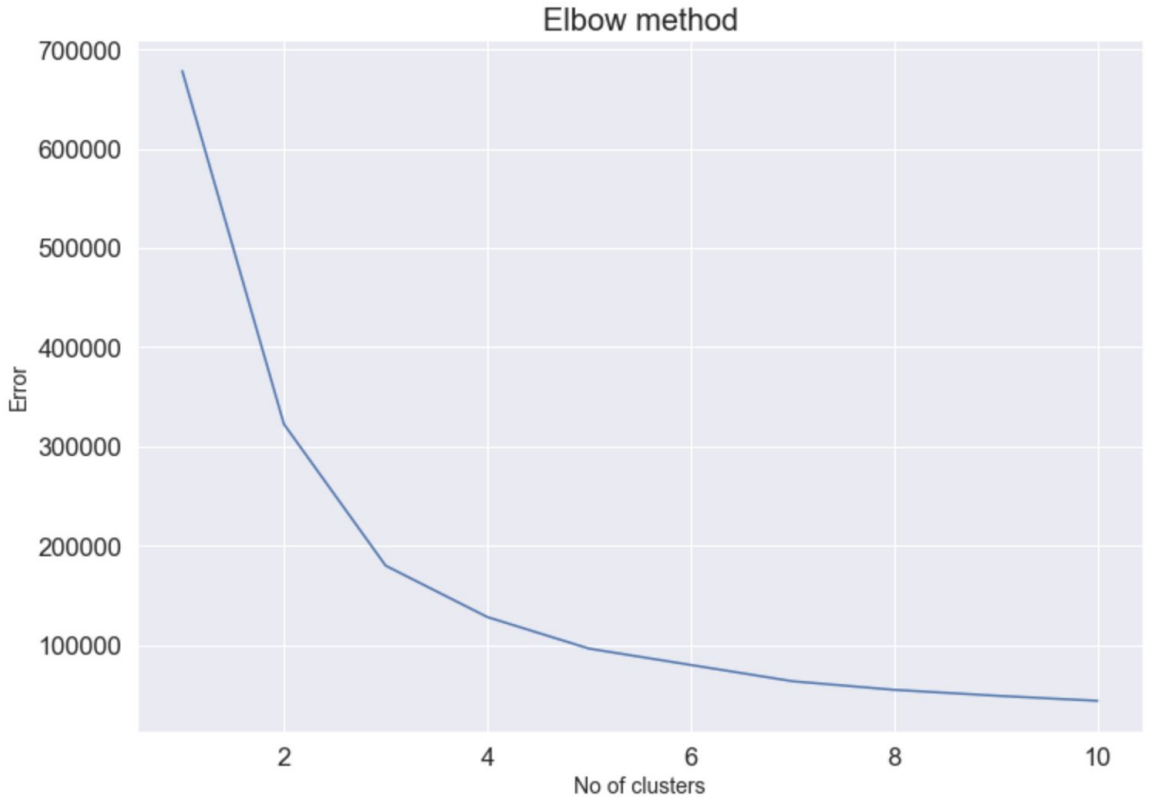
- We're able to see a significant hourly change in CO and O3 compare to other greenhouse gases, which are at the bottom of the graph. Those other gases doesn't seem to change drastically on the scale.

- According to this chart, we can't really say only CO and O3 have effect on the clustering because these two curves are not as flat as the others. We need to further apply machine learning technique to confirm our result.



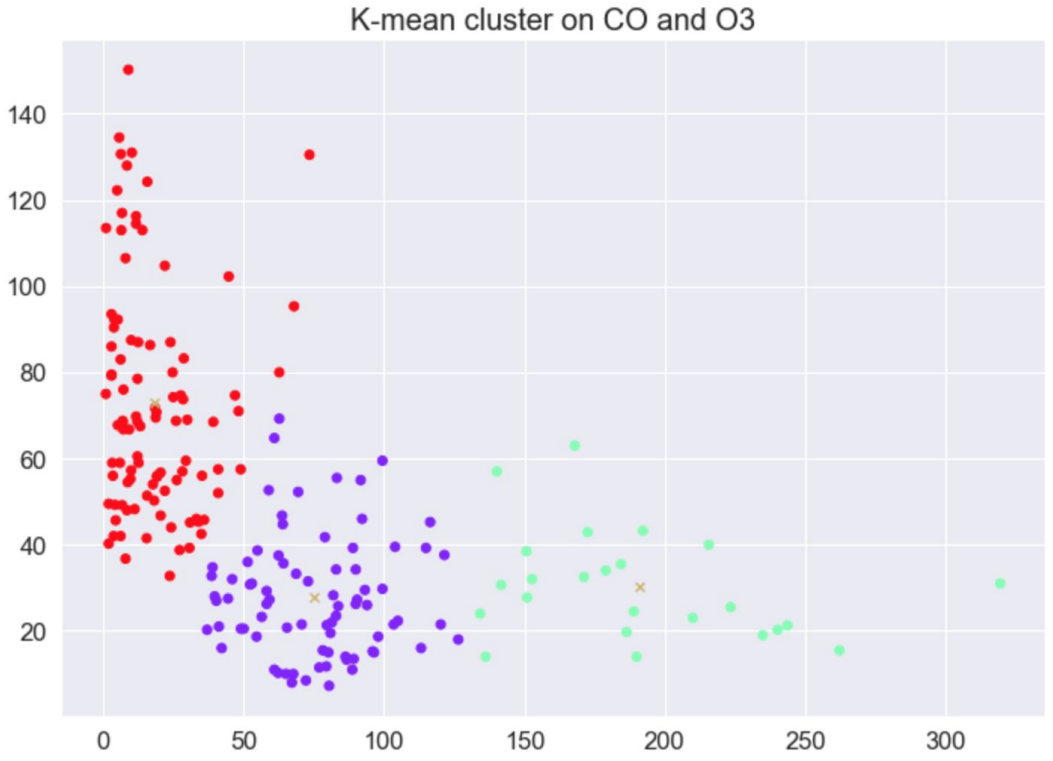
K-Mean Clustering on CO an O3

We first use the the elbow method to find the optimized k, and we find out that k is optimized at k = 3.



K-Mean Clustering on CO and O3

- From the above graph, we're able to easily identify three clusters base on the relationship between CO and O3.
- And also from the previous graph, we can clearly see the **inverse relationship between CO and O3.**
- However, after researching, we did not find any research related to this connection. We are unable to identify the reasoning behind this graph.



Conclusion

- We investigate the relationship between greenhouse gases, we first find out the relationship between PM 2.5 and PM10. We noticed it's a linear relationship because PM2.5 is part of PM10.

- We also do some plots on relationship between O3 and Temperature, Rain and Wind speed as well as some other graph related to greenhouse gases trend and CO confidence interval line chart. We can confirm that turning on heater or AC contributes a significant amount of air pollution.

- Finally we applied K-mean clustering on the O3 and CO and find a clear inverse relationship in clustering. So we believe O3 concentration has negative relationship corresponding to the concentration of CO.

Conclusion

- We would like to apply more Machine Learning Technique in this project investigate more. However, due to the special nature of our dataset, that is, almost all the columns have discrete and quantitative data, the best method we could apply here is k-mean clustering.

- We initially wanted to see how the greenhouse concentration corresponding to the number of people who sent to the hospital due to their respiratory issues and apply the decision tree algorithm. Due to lack of data in this part and its difficulty to acquire the data in this end, we decide not to proceed with the decision tree algorithm on this part, even if we did, we can't guarantee it's a high accuracy decision tree/regression decision tree.

Thank you!

List of Sources

- http://www.npi.gov.au/resource/particulate-matter-pm10-and-pm25
- https://www.environment.gov.au/protection/publications/factsheet-ground-level-ozone-o3
- https://www.baltimoresun.com/news/bs-xpm-1992-06-17-1992169278-story.html
- https://www.epa.gov/co-pollution
- https://www.scmp.com/news/china/society/article/3037119/chinas-plan-reduce-use-coal-heating-northern-homes-still-facing