**MGMT 6962 Final Project**

**Name**: Yihui Yang **RIN**: 661979332

# Abstract and Introduction

Air quality has a strong correlation with people’s health. If people live in a place with bad air quality, he or she might suffer from breathing system damage, lung problems, even cancer. To help ordinary people to prevent these tragedies from happening, both government and third-party organizations have established a series of policies to regulate the related industries and set up metrics to monitor the air quality.

Here is the website provided by EPA and partners to show the current air quality data. The raw data as indicated at the right-bottom of the graph below is due to the courtesy of the New York Dept. of Environmental Conservation.

Graphical user interface

Description automatically generated

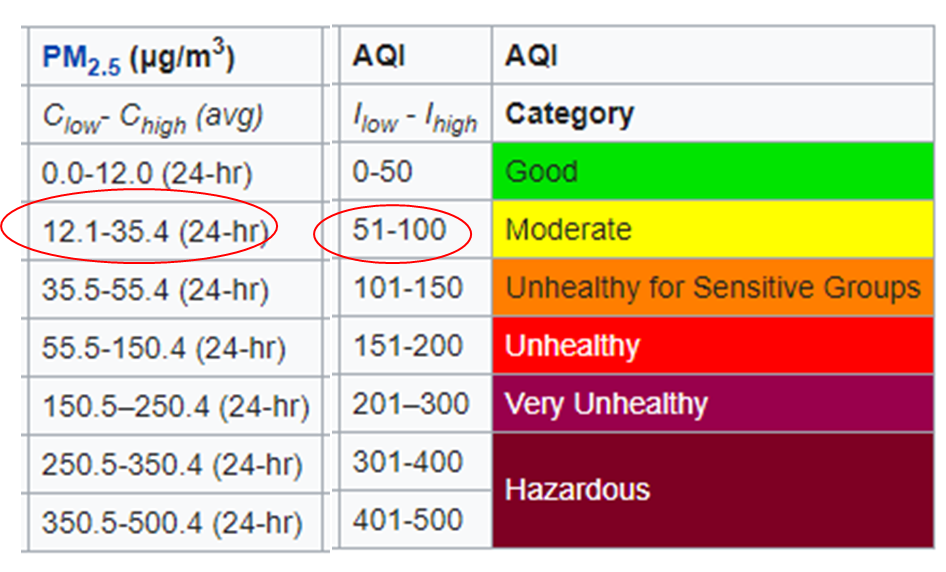
A widely-used indicator is AQI – Air Quality Index. AQI is based on the five main pollutants -- ground-level ozone(O3), particle pollution (PM2.5 and PM10), carbon monoxide(CO), sulfur dioxide(CO2) and nitrogen dioxide(NO2). When calculating the AQI, scientists will calculate the IAQI number of each pollutant, and set the AQI with the highest AQI data. Here is how the AQI is calculated.

A picture containing text, letter

Description automatically generated

Text

Description automatically generated



We can use PM2.5 as an example. If the current PM2.5 concentration is 26.4 μg/m3. Based on the formula and charts above, here is the calculating process:

Text

Description automatically generated

The primary goal is to output a feasible and accurate model that can predict future AQI data based on time series prediction.

Secondly, I want to conduct a classification model to find whether there are some similar patterns within the data and I can group them. Meanwhile, I want to find which pollutant, among those five, has the strongest correlation with AQI.

Thirdly, I will use several hypothesis tests to exam whether air quality has improved after years’ and years’ debate and control.

Last but not least, a graph from the Economist triggered my interest in this topic.

Chart, line chart

Description automatically generated

Covid-19 is undoubtedly a worldwide issue this year. Meanwhile, due to the lockdown, the air quality of some cities getting better after March. I want to find the data and use a hypothesis test to make sure that the drop in AQI is significant.

# Data Description and Exploratory Data Analytics

To accomplish the goals above, I pick the data from aqistudy.cn. The timeframe I picked is from 10/01/2015 to 10/10/2020.



The first column is the DateTime.

API: official Air Quality Index of that day

PM2.5: official PM2.5 concentration(μg/m3).

PM10: official PM10 concentration(μg/m3).

SO2: official SO2 concentration(μg/m3).

CO: official CO concentration(mg/m3).

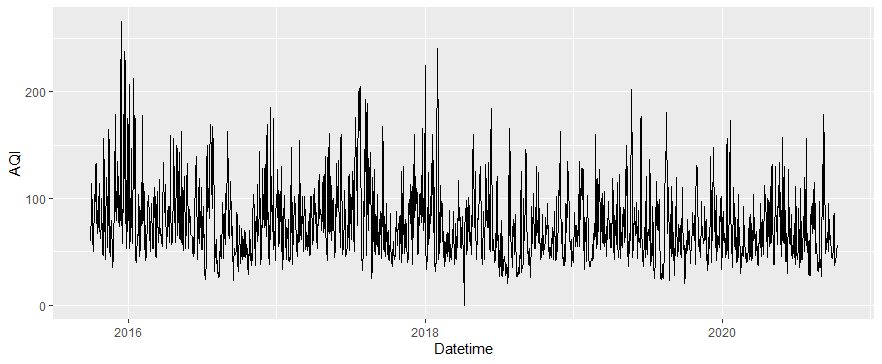
NO2: official NO2 concentration(μg/m3).

O3\_8h: official ozone concentration(μg/m3) in 8 hours.

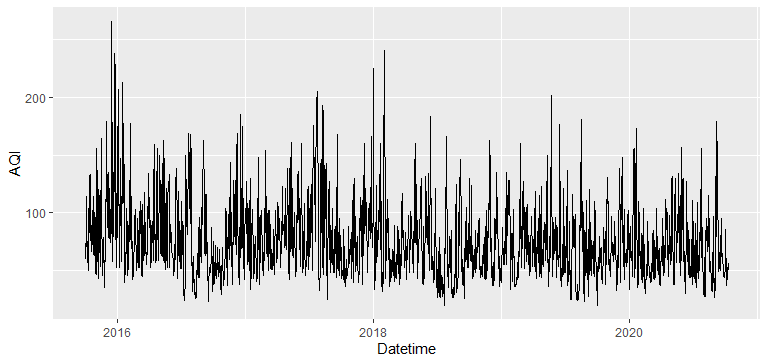
Based on the data above, I believe I can conduct Kmeans to classify the data into different groups so that different policies can be made to better regulate the industries. Meanwhile, time series prediction can be an alert to warn people. If possible, I wish my prediction can help people schedule their outdoor activities more wisely.

# Analysis

First of all, I conduct a line chart regarding all AQI data.



As mentioned above, there will be null values due to the technique issue. Before moving ahead, I should replace 0 with the mean of that column.



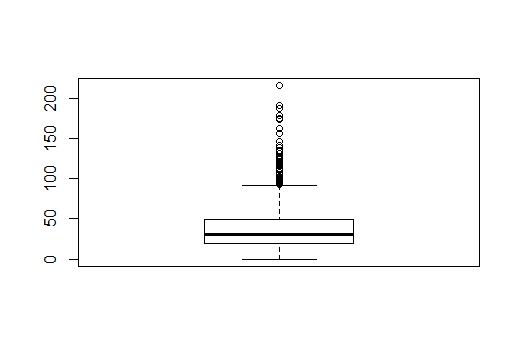
From the trend above, I can draw some brief conclusions. For example, AQI is a seasonal index. The peaks usually happened at the end of each year and the beginning of the next year. It also seems that AQI is getting low, which means the air quality is getting better recently. However, it still needs several hypothesis tests to confirm it.

A picture containing calendar

Description automatically generated

The summary of each column can provide very limited information. Basically, what I can tell is that CO, a fatal gas, is really low in terms of the density in our air. Perhaps I should move on to different methods.

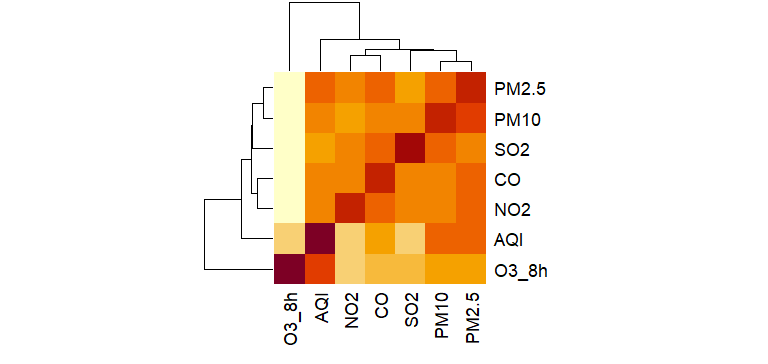
Then, take a look at the PM2.5 boxplot.



It looks like there are no many insights that we can drive from boxplot.

For the next step, I will conduct a correlation analysis with heatmap to find the numeric relationship within the data.

Heatmap can visualize the correlation within the data. After filtering out the DateTime column, here is the heatmap.

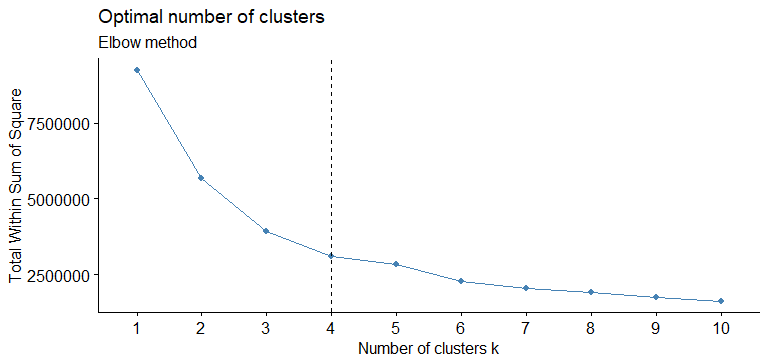


Surprisingly, ground-level ozone has a strong correlation with AQI. The second strongest indicator is PM2.5. PM2.5 and PM10 are strongly correlated with each other.

# Model Development and Application of model(s)

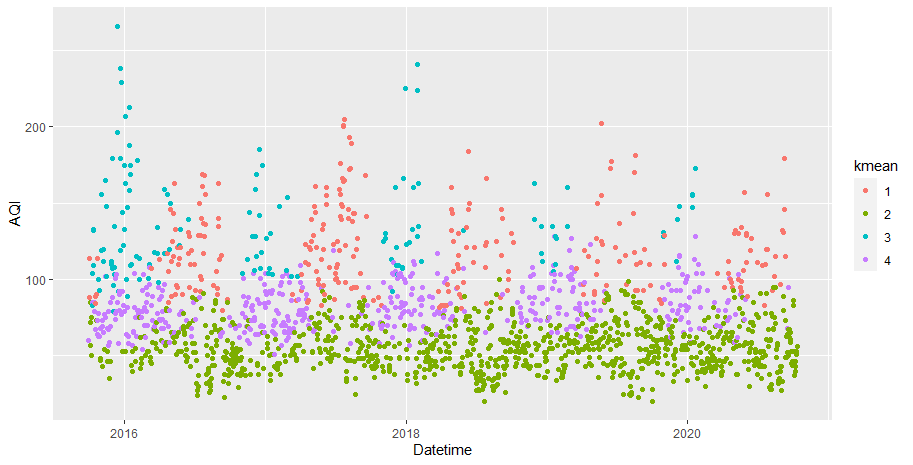
Since the majority of the data are numeric, I can use kmean to classify the data and find the pattern accordingly.

Firstly, I used the elbow method to find the optimal number of clusters.



So, the number of clusters is 4.

Then, choose 4 as the k number, I conduct kmean function in R. I plotted the results with ggplot. The graph below clearly indicates that there are some patterns regarding DateTime.



It will provide a better look when aggregating the data by pollutants within each group.



For group 1, from the graph, it seems that this phenomenon usually happens in the mid of the year. Commonly, during those days, ozone is quite high compared with the rest of the year.

For group 2, in terms of AQI data, the green dots which are located below the 100-line indicate that people can enjoy really good air quality during those days.

For group 3, the days are heavily polluted. AQI, PM2.5, PM10, SO2, CO, and NO2 are all high. These dots usually appear from the end of a year to the beginning of the next year.

For group 4, these days are not too bad, but people still need to take care of themselves when going out.

Overall, the green dots are becoming more and more frequent these years. The relevant parties, on the other hand, can spend more time to figure out what leads to group 1 caused by ground-level ozone and group 3 caused by AQI, PM2.5, PM10, SO2, CO, and NO2.

Hypothesis Test

There are a few questions that I am interested in:

1) Are air quality getting better?

To break down this question, I will generate 3 new questions:

* Is the air quality of 2016 greater than in 2017?
* Is the air quality of 2017 greater than in 2018?
* Is the air quality of 2018 greater than in 2019?

2) Did the reduction of human movement help to improve the air quality?

To answer this question, we can compare the AQI data during the lockdown and in the same period last year.

Firstly, Question 1:

* Is the air quality of 2016 greater than in 2017?

Hypothesis 0: AQI2016 <= AQI2017

Hypothesis 1: AQI2016 > AQI2017

Text, letter

Description automatically generated

P-value is way greater than 0.05, so I can’t reject the hypothesis that AQI2016 <= AQI2017. Perhaps, the AQI from 2016 to 2017 doesn’t change too much.

* Is the air quality of 2017 greater than in 2018?

Hypothesis 0: AQI2017 <= AQI2018

Hypothesis 1: AQI2017 > AQI2018

Text

Description automatically generated

P-value is way smaller than 0.05, so I reject the hypothesis and conclude that AQI2017 > AQI2018. The air quality from 2017 to 2018 improved a lot.

* Is the air quality of 2018 greater than in 2019?

Hypothesis 0: AQI2018 <= AQI2019

Hypothesis 1: AQI2018 > AQI2019

Text

Description automatically generated

P-value is greater than 0.05, so I can’t reject the hypothesis that AQI2018 <= AQI2019. Perhaps, the AQI from 2018 to 2019 doesn’t change too much.

Question 2

I filter out two data set, one called df\_2019\_during\_cov\_period and the other called df\_2020\_during\_cov. I use these two data to conduct the following hypothesis test.

Hypothesis 0: AQI2019\_during\_cov\_period<= AQI2019\_during\_cov

Hypothesis 1: AQI2019\_during\_cov\_period> AQI2019\_during\_cov

Text

Description automatically generated

P-value is smaller than 0.05, so I reject the hypothesis and conclude that AQI2019\_during\_cov\_period> AQI2019\_during\_cov. In terms of AQI, air quality improves a bit during the same period.Perhaps, limiting human movement can improve air quality.

Time series prediction

To conduct time series prediction, I firstly filter out irrelevant data and only keep datetime and AQI columns. Meanwhile, I only keep the data from 2016 to 2019 to eliminate the effect of COVID-19. Then, I set frequency as 30(month level) and the number of validation as 5 to fit the linear trend model. Here is the result:

Text, letter

Description automatically generated

After that, I tried to fit a model with seasonality:

Text

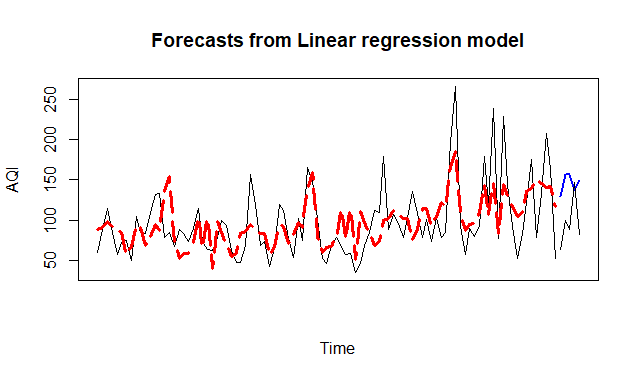
Description automatically generated

Last but not least, I fitted the model with seasonality and trend:

A close up of a screen

Description automatically generated

Overall, the model with seasonality and trend outperforms the rest of the models. So, I plot the train-test-valid data to show the prediction.



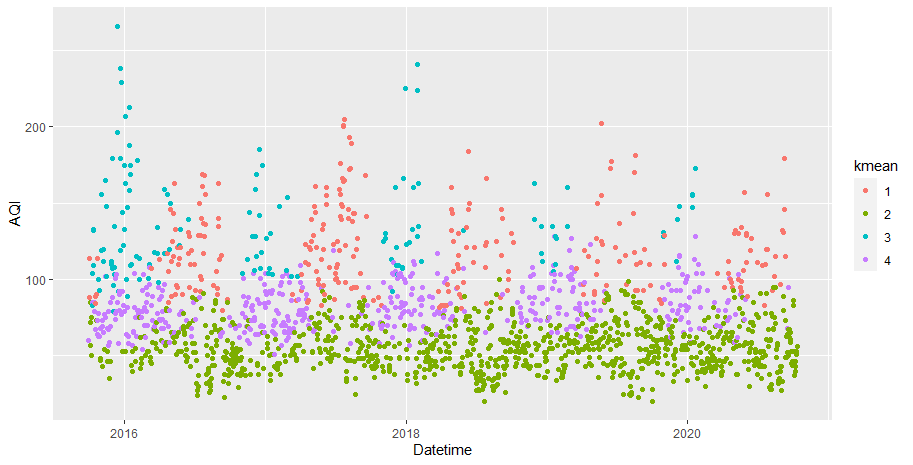
Blackline is the actual data. The Red and dash line is the train forecasts and the blue line is the validation forecasts. Apparently, the prediction results have some overlap with the actual data. Even if sometimes, there is a huge gap between the prediction and the actual results, the model can still predict the trend.

# Conclusions and Discussion

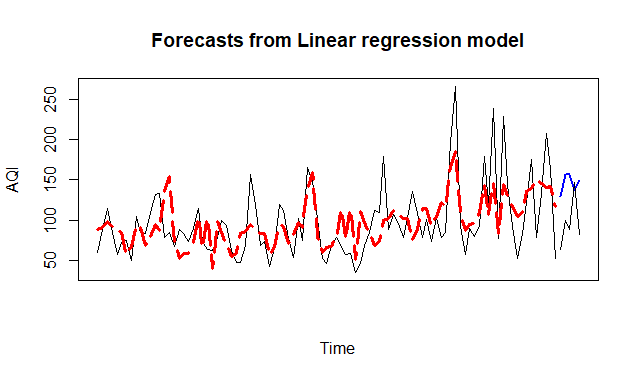
Respectively, the average AQI in 2016 is 80.82. The average AQI in 2017 is 83.63. The average AQI in 2018 is 70.13. The average AQI in 2019 is 72.62. From the model and analysis above, the air quality is not going worse at least. From 2017 to 2018, air quality actually improved significantly.

Ground-level ozone has a strong statistical relationship with AQI data, and the PM2.5 is the second strongest factor.

Comparing with the same time in 2019, the limitation of human moving contributes to air quality improvement. Last year, from January to April, the average AQI was 74, but this year the number is down to 67.5. Meanwhile, the difference is significant statistically.



The air pollution during summertime(red dots) is usually caused by ground-level ozone. Ground-level ozone is created by NOX and VOC. This process happens much easier and faster in a hot atmosphere, which explains why red dots appear more frequently during the mid of the year. In water, however, the weather is cold. The transmission of pollutants is slower than usual, so the majority of heavily polluted days are caused by particles and gas emitted by vehicles.



The model I choose is based on a linear model which considers seasonal and trend effect. The R2 is more than 0.4. For the model with real data, the R2 value is acceptable.

Reference

<https://www.airnow.gov/?city=Troy&state=NY&country=USA>

[https://en.wikipedia.org/wiki/Air\_quality\_index#:~:text=The%20index%20is%20based%20on,than%202.5%20%CE%BCm)%20and%20PM10.](https://en.wikipedia.org/wiki/Air_quality_index)

<https://www.economist.com/graphic-detail/2020/09/05/air-pollution-is-returning-to-pre-covid-levels>

<https://www.aqistudy.cn/historydata/daydata.php?city=%E4%B8%8A%E6%B5%B7&month=2020-09>

<https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics#:~:text=Tropospheric%2C%20or%20ground%20level%20ozone,volatile%20organic%20compounds%20(VOC).&text=Ozone%20can%20also%20be%20transported,can%20experience%20high%20ozone%20levels>