# Visualizing Data of Changes in Air Pollution in Seoul, South Korea

**CSC-475: Seminar in Computer Science**
**Fall 2023**

**Ting Chen**
Department of Computer Science
Furman University
ting.chen@furman.edu

## Abstract

A series of plots and graphs are made to establish a basic understanding of air pollution data from Seoul, Korea from 2017 to 2019. The air quality changes was measured by the hourly average of six air pollutant volumes. With minor data cleaning, data aggregation, and exploratory data analysis, the data set was ready to be processed using python library to visualize the trends of changes as well as the story behind the numbers. From summary plots for different time blocks of the year, we can see the major changes occurs at the similar time throughout the three years time span. In summary, air pollution gets worst later in the afternoon until late night from the average data calculated for 24 hours. The monthly average exhibit the trend of worst pollution during winter season. Furthermore, it more be helpful to investigate the peaks of the data to ensure the data validity and explain the abnormality, whether it is caused by human activities.

## 1 Introduction

As climate change has become a serious issue for the humanity, we should understand the different aspect of the affects and how the environment may change our living styles. For example, air pollution is closely related to our daily life. The first step of the project is to visualize the trend of air quality based on air pollution data. The next step is to attempt to answer the abnormality of the plots, whether it is caused by data anomalies or human imposed effect on the environment.

The desire outcome for this project is a series of visualization comprised of different interactive plots and maps to display the changes of air pollution. The goal is to interpret the large amount of data by making plots based on the volume of the air pollutants. We collect large amount of data constantly, it is important to see the story behind the data. We care about the environmental effected brought by climate change, one of the obvious effect is the air quality. Using previous air pollution data can help us understand how air quality had change in a small area and these small changes may add up to large effect as seen in a global scale. We need a tool to help us keep track and visualize the changes to advocate more environmental policies as well as controls for human behaviors that directly contribution to air pollution.

As for the next step of the visualization, it would be more helpful to create an interactive dashboard publicly available for any potential audience. The target audience are students who have some basic knowledge of interpreting graphs. The dashboard should be fully functional for different aspect of the air pollution, such as periodic changes, locational changes, etc. The dashboard is a collective pocket for all the plots and graphs that are meaningful.

The data set used for studying the periodic trend of air pollution is from kaggle [1], it records hourly changes of six air pollutants range from 2017 to 2019. The minor data cleaning removed unreasonable values as well as a basic exploratory analysis has been conducted to understand the data set. A series of plot scaled from hourly average, daily averages, to monthly averages were plotted in the results session. The discussion of obvious abnormality of each figure was outline under each visualization. The limitation includes lack of recent data will be address in future work as well as some other brainstorm ideas for further research will be mentioned.

## 2 Related Works

As we all known that Climate Change is a serious problem globally, raising awareness for younger generation is important for collective actions to promote positive changes to combat negative effects of environmental damages. For the generation with extensive access to online sources, it is convenient to educate them for promoting changes of actions. One of the best way to raise awareness and spread concerns with others around is through stimulate climate-related behavior [2]. The target audience is high school students who have access to the Internet. The ultimate goal is to develop a program that will educate the public about consequence of climate changes to our daily lives though understanding of large amounts of valuable data collected constantly. Mebane, Benedetti, Barni, and Francescato proposed "a new psychological environmental intervention program seeking to increase students' awareness of climate change and related emotions and to promote students' empowerment" [2]. The similar intention of raising awareness of environmental issues is shared with this seminar project. The researchers develop two modules for the program. First, they will educate the students with the effects of climate change on environment and then laboratory participation to learn protection measures for the environment. Therefore, for the first part of education purpose, visualizing data for the key to deliver information gain from large data sets which will be very useful for students to be aware of the changes that is happening. A factor that can be noticed by the public is air quality in daily lives. The sudden raise of air pollution level can be sensed by people outdoor. Due to lack of time and expertise, the project is focuses on air pollution data of a small region to promote self learning.

The quicker way to understand severity of air pollution in certain area is through air quality index specified the location. With the mobile technology, the access to real time air quality is easy and credible. Especially in urban area, any type of new events that cause changes in air quality can be quickly detected and recorded to spread to the public. Chen [3] studies the urban air quality using Google Earth Keyhole Markup Language (KML) [4] for visualizations. The research highlights the power of interactive visualization. The researcher used real-time monitoring data of the capital city, Beijing, one of the most industrial cities in the world. It is important to get the most recent updates of air quality. Chen [3] used existing technology to create an user-friendly tool for visualizing air quality which is something the project can aim to accomplish for a deliverable product.

Taking air pollution data for next step analysis, Yang, Peng, and Zheng have develop an analysis scheme to explore the data from multiple dimensions through various displays [5]. The paper incorporated heterogeneous data for analysis to accomplish the goal of identifying the cause, transmission, and evolution of data. The system is closely resembles how visualizations can benefits the professionals even more compared to the public for in-depth analysis. Another similar project that focus on developing visualizations for urban area is from Lu, Ai, Zhang and He who also study air quality in big cities of China. [6]. Instead of using existing technology, the researchers develop two novel visualization tools for specific purpose of analyzing air quality. This sparks an inspiration for another approach to pursuit the project. But, due to lack of technical skills and time permitted, using simple visualization tools will work better to accomplish the goal of visualization. However, the results from the research is important to guide the process of this project. The researchers used the results of significant findings to guide their further studies. One of the goal for this project is also to use the current findings of past data to guide further research.

## 3 Data

The data used for the first part of the project is the air pollution data from Seoul, Korea, obtained from kaggle. The data set contains 647511 rows and 11 columns. The variables includes the time and date that data was recorded, the location data, and the 6 air pollutants. The six air pollutants in the data set are sulfur dioxide($SO_2$), nitrogen dioxide($NO_2$), ozone($O_3$), carbon monoxide($CO$), particles are 10 micrometers or less in diameter($PM_{10}$), and particles that are 2.5 micrometers or less in diameters($PM_{2.5}$). The location data consist of the station code for the 25 stations for data as well the physical address and geographical data (latitude and longitude). The measurement date ranges from January 1st, 2017 to December 31st, 2019. The hourly average volume of each air pollutant was recorded as decimal values.

### 3.1 Location Data

| Row # | Station Code | Address |
|---|---|---|
| 0 | 101 | 19, Jong-ro 35ga-gil, Jongno-gu, Seoul, Republic of Korea |
| 25905 | 102 | 15, Deoksugung-gil, Jung-gu, Seoul, Republic of Korea |
| 51810 | 103 | 136, Hannam-daero, Yongsan-gu, Seoul, Republic of Korea |
| 77714 | 104 | 215, Jinheung-ro, Eunpyeong-gu, Seoul, Republic of Korea |
| 103610 | 105 | 32, Segeomjeong-ro 4-gil, Seodaemun-gu, Seoul, Republic of Korea |
| 129500 | 106 | 10, Poeun-ro 6-gil, Mapo-gu, Seoul, Republic of Korea |
| 155405 | 107 | 18, Ttukseom-ro 3-gil, Seongdong-gu, Seoul, Republic of Korea |
| 181296 | 108 | 571, Gwangnaru-ro, Gwangjin-gu, Seoul, Republic of Korea |
| 207190 | 109 | 43, Cheonho-daero 13-gil, Dongdaemun-gu, Seoul, Republic of Korea |
| 233094 | 110 | 369, Yongmasan-ro, Jungnang-gu, Seoul, Republic of Korea |
| 259000 | 111 | 70, Samyang-ro 2-gil, Seongbuk-gu, Seoul, Republic of Korea |
| 284905 | 112 | 49, Samyang-ro 139-gil, Gangbuk-gu, Seoul, Republic of Korea |
| 310797 | 113 | 34, Sirubong-ro 2-gil, Dobong-gu, Seoul, Republic of Korea |
| 336689 | 114 | 17, Sanggye-ro 23-gil, Nowon-gu, Seoul, Republic of Korea |
| 362594 | 115 | 56, Jungang-ro 52-gil, Yangcheon-gu, Seoul, Republic of Korea |
| 388498 | 116 | 71, Gangseo-ro 45da-gil, Gangseo-gu, Seoul, Republic of Korea |
| 414404 | 117 | 45, Gamasan-ro 27-gil, Guro-gu, Seoul, Republic of Korea |
| 440296 | 118 | 20, Geumha-ro 21-gil, Geumcheon-gu, Seoul, Republic of Korea |
| 466200 | 119 | 11, Yangsan-ro 23-gil, Yeongdeungpo-gu, Seoul, Republic of Korea |
| 492094 | 120 | 6, Sadang-ro 16a-gil, Dongjak-gu, Seoul, Republic of Korea |
| 517988 | 121 | 14, Sillimdong-gil, Gwanak-gu, Seoul, Republic of Korea |
| 543893 | 122 | 16, Sinbanpo-ro 15-gil, Seocho-gu, Seoul, Republic of Korea |
| 569798 | 123 | 426, Hakdong-ro, Gangnam-gu, Seoul, Republic of Korea |
| 595702 | 124 | 236, Baekjegobun-ro, Songpa-gu, Seoul, Republic of Korea |
| 621607 | 125 | 59, Gucheonmyeon-ro 42-gil, Gangdong-gu, Seoul, Republic of Korea |

Table 1: A table of 25 locations that the air pollution data was collected in Seoul, South Korea. The 25 station codes represents 25 gu(districts) of the city of Seoul similar to boroughs in New York.

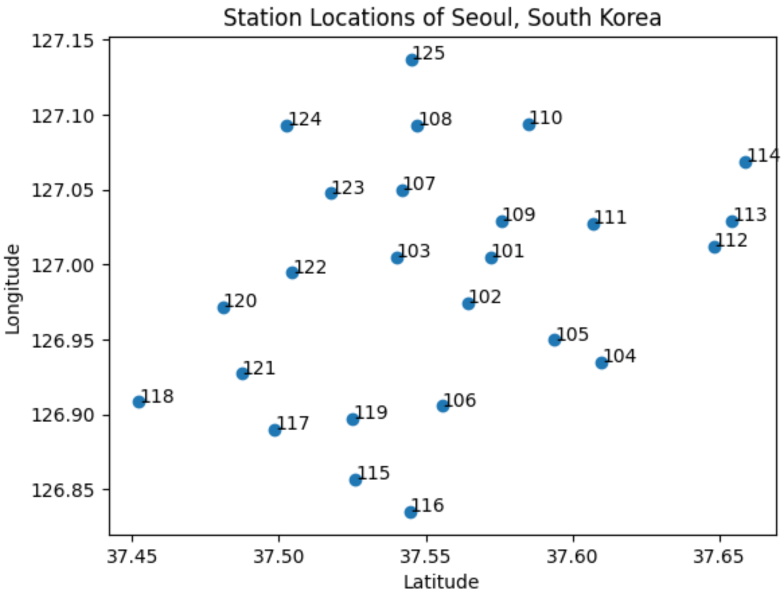

Figure 1: Locations of data in Seoul, South Korea by latitude and longitude.

### 3.2 Periodic Data

After data cleaning, the rows with unreasonable values (-1 for air pollutant volumes) were removed from the data set. Data aggregation for time was applied to the data set to limit the number of data points on the plots. First, we expand the date time attribute to Date, Year, Month, Day, and Hour columns for better categorize the data. Then, the daily average, monthly averages for each station was calculated by grouping the desire time period. These averages were calculated on the fly of creating graphs.

| Pollutant | NA (-1) | Min | Max |
|---|---|---|---|
| $SO_2$ | 3976 | 0.0 | 3.736 |
| $NO_2$ | 3834 | 0.0 | 38.445 |
| $O_3$ | 4059 | 0.0 | 33.6 |
| $CO$ | 4036 | 0.0 | 71.7 |
| $PM_{10}$ | 3962 | 0.0 | 3586.0 |
| $PM_{2.5}$ | 3973 | 0.0 | 6256.0 |

Table 2: A table of summary of missing values and ranges of the six air pollutants.

### 3.3 Limitation of Data

The data set was not updated constantly that includes the recent records, but kaggle did specified the original source which is from official Korean government. The next step could be acquiring more data and transform them into similar format as the current data set to plot further trends of air pollution changes in Seoul, Korea. Another limitation is that we do not know validity of all the records, although the data was the the Korean government website, I did not find the exact data set from the original source.

## 4 Methods

The air pollution data was collected from kaggle [1]. The data set was collected as a csv file containing data from the 2017 to 2019 in English. A basic data exploratory analysis was conducted to understand the size, variables, and missing data from the data set. A dot plot was made to summarized the location data. The unreasonable values, negative values for volumes are being treated as NA, these values will be ignored when calculating summary data and creating visualizations. Then data was aggregated by stations and time, the averages of periodic volumes each air pollutant was calculated to plot the line graph of daily averages. After plotting the data, the visualization was been studies and analyzed to proposed further research question and spark new ideas for more plots that can be visualize to answer fundamental research questions regarding the changes of air pollution over a period of time.

# 5 Results

The visualizations are categorized by the date and time for a better interpretation of data changes over time of three years. The following plots are aggregated for all 25 districts of Seoul, South Korea. Since the locations are close to each other on the global scale, the standardized average was calculated for easier data visualization. Note that the y axis of each plots contain negative values because the data has being standardized. From different time blocks, we can see the trends of each air pollutant. The results of each plot will be discussed below.

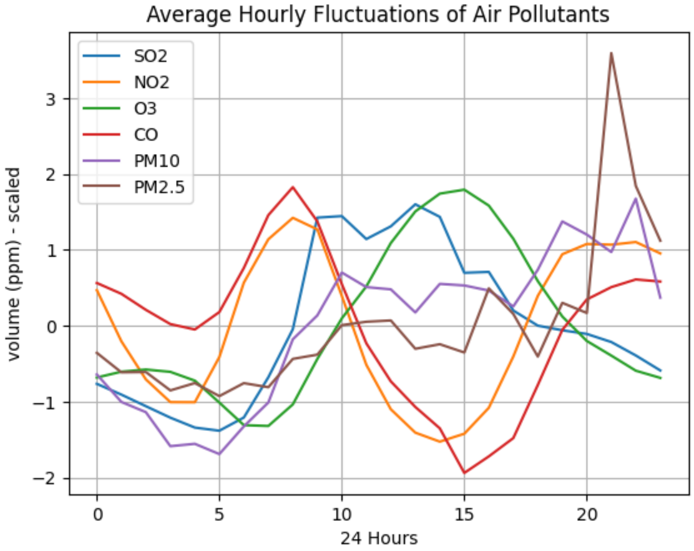

Figure 2: A line plot of 24 hours standardized average for each of the six air pollutants.

From Figure 2, we can see that carbon dioxide($CO$) and nitrogen dioxide ($NO_2$) follow a similar trend throughout the day. Since hour 0, carbon dioxide has higher volume until hour 14 (2:00pm). From there, nitrogen dioxide ($NO_2$) takes over and the two pollutants increase at the same rate. Compared to $CO$ and $NO_2$, ozone($O_3$) exhibits an inverse trend. Because ozone forms when nitrogen oxides and organic compounds react with each other in sunlight and hot temperatures, it makes sense that levels of ozone increases during the day and then declines after 3pm for slower human activities. Another mentionable point is the particles with diameters less than 2.5 micrometers ($PM_{2.5}$) at roughly hour 21 (9:00pm) has peak for daily average. It would be nice to re-investigate the data set to make sure the data was consistent. Then, more research can be done to determine the cause of the peak.

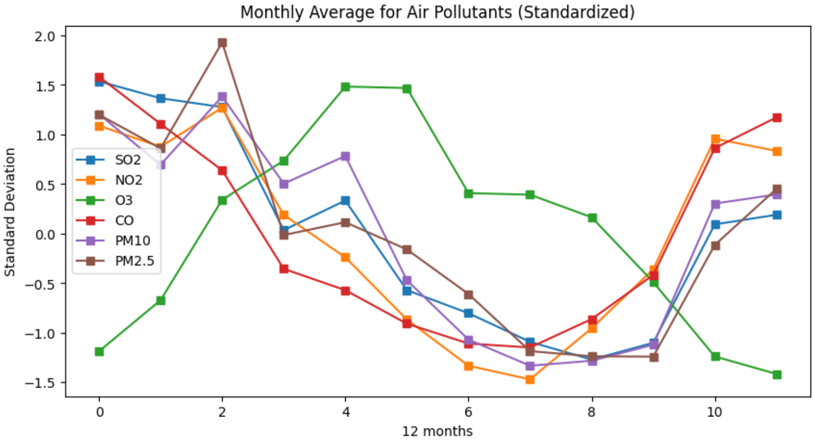

Figure 3: A line plot of monthly average for each of the six air pollutants.

From monthly summary, we clearly see the inverse relationship between ozone ($O_3$) to other five pollutants. As of January (month 0), ozone starts off low and gradually increase until May (month 4), while other pollutants fluctuate but consistently follow a decreasing trend until October (month 9). The air pollution is worse during winter season (November, December, January, and February) and gets to worst point during March in Seoul, South Korea. We can further investigate whether this peak is caused by environmental factor such as pollen or other human activities. From Figure3, we can investigate the changes of air pollution over a year to better advocate for policies to control air pollution.

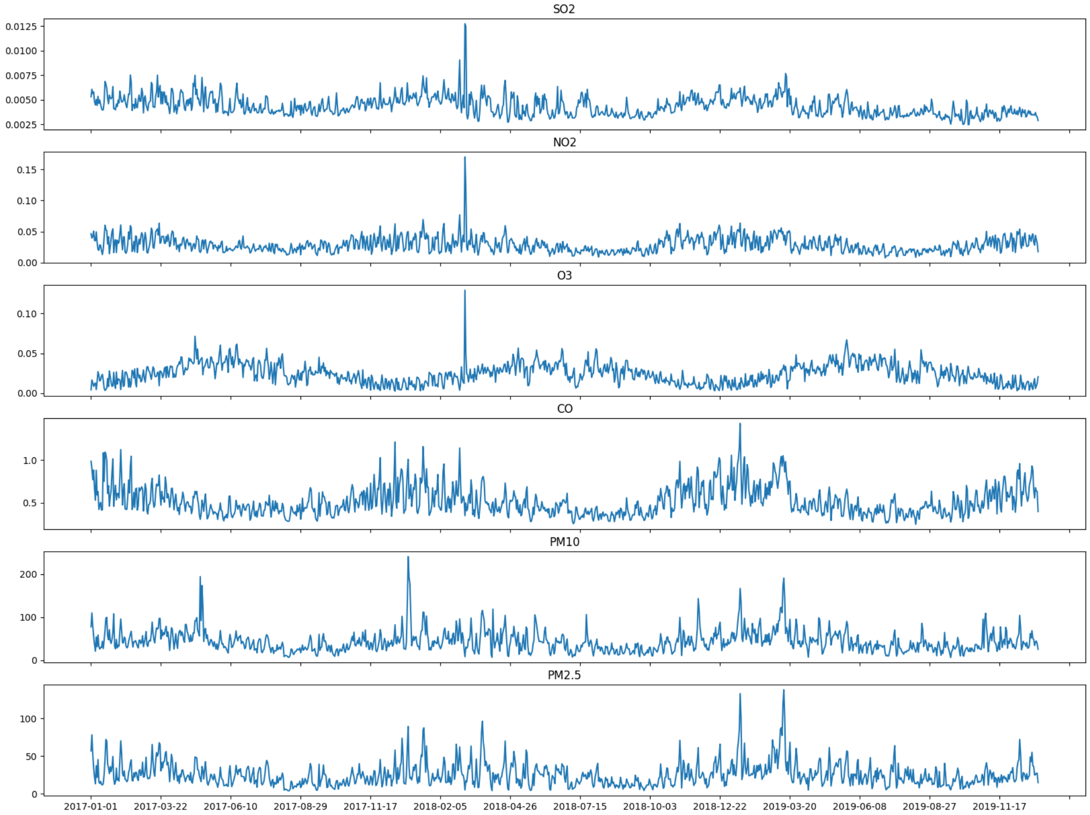

Figure 4: A line plot of daily average for each of the six air pollutants from 1/1/2017 to 12/31/2019.

The daily averages for aggregated station data can be seen in Figure 4. We can better compare the changes of each air pollutant across the same period of time from 2017 to 2019. An interesting period to mention is early March of 2018, the volume of $SO_2$, $NO_2$, and $O_3$ all have peaks at the exact same time. Further research is required to see whether this is caused by human activity or natural disaster. Overall, the trend is consistent across 3 years of data from Seoul, South Korea. Air pollution did not get dramatically worst over 2017 to 2019. However, there are more small particles ($PM_{10}$ and $PM_{2.5}$) from end of to 2018 to beginning of 2019. This could be another interesting research topic to explain the data.

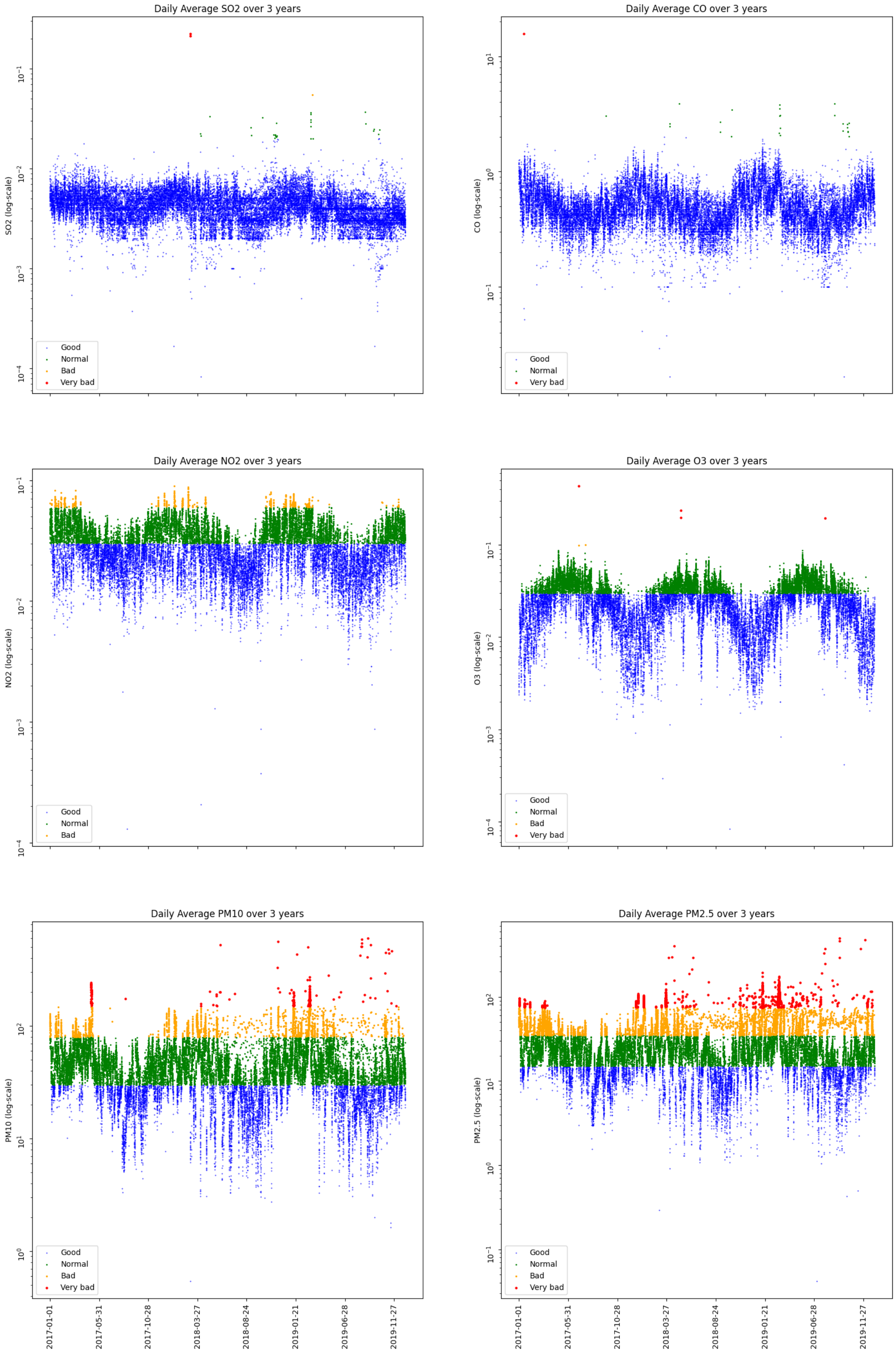

Figure 5: A multi-plots for air pollution across 3 years for each air pollutant.

From the log scale plot from 2017 to 2019, the most concerned pollutant is obviously PM10 and PM2.5. The daily averages fluctuates through the year, it is clear that air pollution becomes worse at the beginning of the year. Generally, Seoul has normal air quality according to the guidelines provided by the data set. Another important aspect is the trend of increase small particles (PM10 and PM2.5) as time progresses. In the plot, yellow and red marks unhealthy levels of air quality. There are considerably more yellow and red marks in 2019 than in 2017. This shows air pollution is a serious problem since only in two years of time and in small regions, the changes can be detected through visualization. Since the plot includes aggregated data from all 25 locations, the individual daily average air pollution is plotted in Figure 6 to see the distribution for each location.

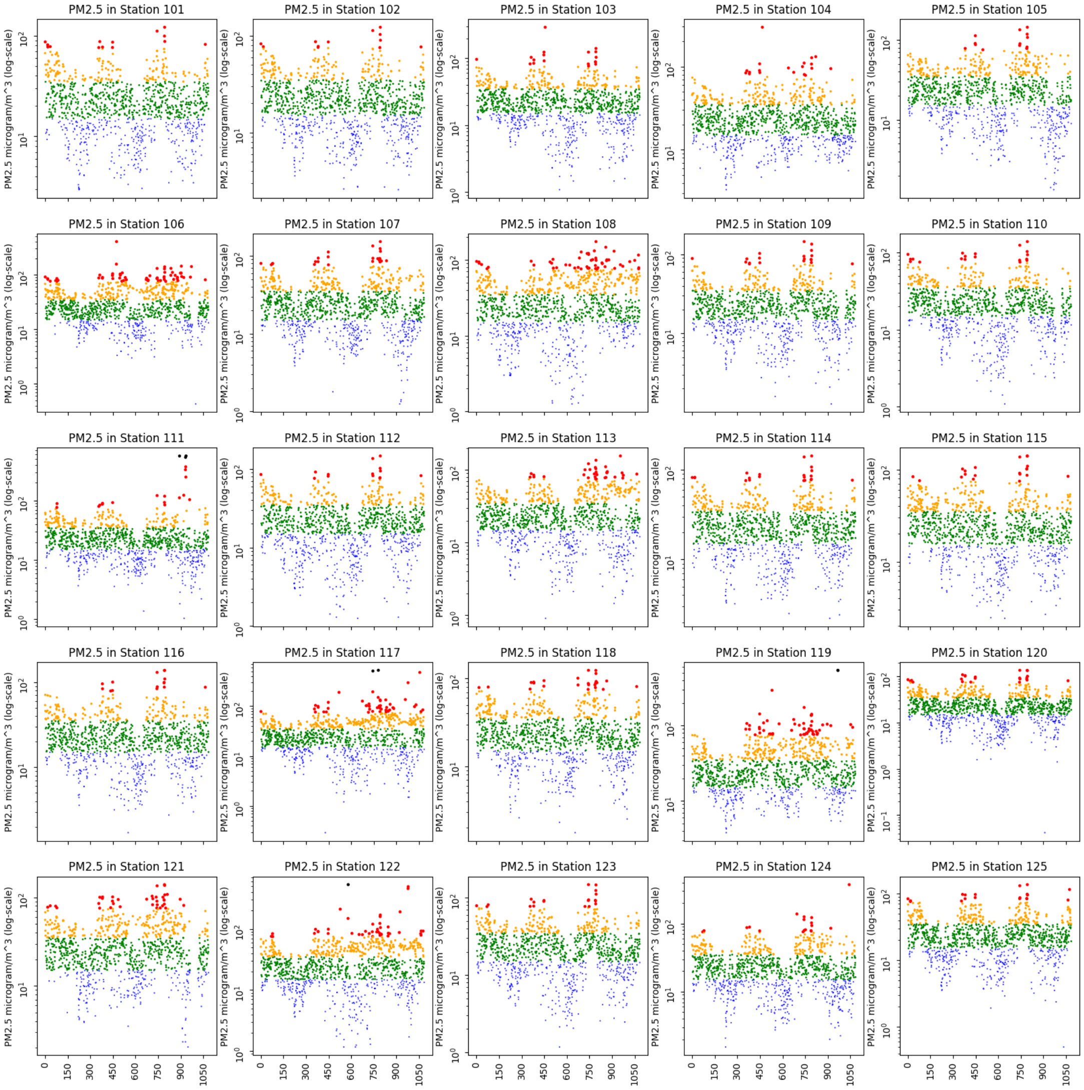

Figure 6: The average volume of PM2.5 in all 25 locations.

The distribution of PM2.5 across 25 locations seems very similar to each other which makes sense since according to latitude and longitude, these locations are very close to each other. But, station 120 and station 125 have relatively high volume of PM2.5. According to the location plot in Figure 1, station 120 is located in the western region of Seoul ad station 125 is located in the northern part of Seoul. The reasonable next step is to see whether these locations have factories or other environmentally hazardous production, then the officials can propose better regulation to mitigate air pollution in the future.

## 6 Future Work

The major limitation of the data set was it does not contain most updated data for recent years. As we known, by the end of 2019 until beginning of 2022, COVID-19 has reduce interactions and industrial human activities by a great amount, these new data might be interesting to visualize as a continuation of the existing data. The locations of the data recorded was compact within a small region, it might be interesting to compare the result to capitals of other countries and further compare whether COVID-19 had affect air pollution the same way between multiple major cities. As minor data cleaning and validation had been perform on the data set, we assume the data set was correctly translated from Korean government website to kaggle [1] and majority of data was kept by the author [1] of kaggle community. Another aspect of the visualization is to bring it downscale for individual. We tend cared more about our surrounding than the world, small changes may be ignored, but these small changes accumulates and caused irreversible damage to our only habitat, Earth. It would be helpful to educate each one of us that changes does exist, we could monitor our behavior to limit the damage of environment.

## References

[1] bappekim. Air pollution in seoul.

[2] et al. Mebane, Minou Ella. Promoting climate change awareness with high school students for a sustainable community. *Sustainability*, 15(14), 2023.

[3] Pengyu Chen. Visualization of real-time monitoring datagraphic of urban environmental quality. *Springer Open*, Feb 2019.

[4] Google Developers.

[5] Xudong Yang, Hui Peng, and Qingming Zhang. Visual analysis of heterogenous air pollution data. page 300–306, 2021.

[6] Wei Lu, Tinghua Ai, Xiang Zhang, and Yakun He. An interactive web mapping visualization of urban air quality monitoring data of china. *Atmosphere*, 8(8), 2017.