Big Data Project-Group10

Hang Zhang
New York University
hz2447@nyu.edu

Yun Zhang
New York University
yz518@nyu.edu

Vin Liu
New York University
zl1477@nyu.edu

ACM Reference Format:
Hang Zhang, Yun Zhang, and Vin Liu. 2021. Big Data Project-Group10. In *Proceedings of Team Air (Big data).* ACM, New York, NY, USA, 9 pages.

1 INTRODUCTION

We began with the question, "How has COVID-19 impacted our life?". After looking at several potential areas, including but not limited to remote learning, economy, stock market, traffic, and urban migration, we settled on exploring the change of air quality correlated with the COVID-19 pandemics in major cities around the world. We intend to obtain a variety of data, including Air Quality Index, CO, NO2, PM2.5, PM10, and Ozone to measure air quality. We analyze the data and map them to major events during the COVID-19 pandemics, such as major lockdowns and curfews in the cities. Through data gathering, processing, and analysis, we might be able to answer the following questions:

- Is there a significant change in the air quality since the COVID-19 onset?
- Did air quality get better or worse during the pandemic across major cities in the world? What's the possible reason behind such change?
- Does the air quality change in a similar pattern among cities around the world?
- Which areas exhibited a significant reduction of certain pollutants during the lockdown?
- Were there other factors that influenced air quality during the time periods studied?

2 PRIOR KNOWLEDGE

2.1 What is AQI

AQI stands for Air Quality Index, from 0 to 500, the greater the value, the higher level of air pollution and more of health concern. The AQI is divided into six categories. Each category corresponds to a different level of health concern. Each category also has a specific color. In this study, we adopted the first three columns of information: AQI color, Level, and AQI value(0-500)[1].

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

Figure 1: Air Quality Index[2]

2.2 What are the Air Quality Parameters?

- **Particulate Matter (PM10/PM2.5)** - PM10 particles (the fraction of particulates in air of very small size (<10 µm)) and PM2.5 particles (<2.5 µm) are of major current concerns, as they are small enough to penetrate deep into the lungs and so potentially pose significant health risks.
- **O3** - Ozone at ground level is a harmful air pollutant, because of its effects on people and the environment, and it is the main ingredient in "smog"
- **NO2** - nitrogen dioxide causes detrimental effects to the bronchial system.
- **SO2** - Sulfur dioxide irritates the respiratory tract and increases the risk of tract infections. It aggravates conditions such as asthma and chronic bronchitis.
- **CO** - Carbon monoxide is produced by the incomplete burning of materials which contain carbon, including most transport fuels, however even in busy urban centers, CO concentrations rarely exceed health-related standards.

In this project, Air Quality Parameters and Six Major Pollutants are used interchangeably.

3 DATASETS

Air Quality Open Data Platform Worldwide COVID-19 dataset: The dataset contains air quality data for 380 major cities in the world. The data for each major city is based on the average (median) of several stations. The data set provides min, max, median and standard deviation for each of the air pollutant species (PM2.5, PM10, Ozone ...) as well as meteorological data (Wind, Temperature, ...). All air pollutant species are converted to the US EPA standard (i.e. no raw concentrations). All dates are UTC-based. The count column is the number of samples used for calculating the median and standard deviation. Available from https://aqicn.org/data-platform/COVID19/.
EPA Outdoor Air Quality Data: This dataset includes the daily air quality data for many cities and regions in the United States. The data includes reading for Ozone, SO2, CO, NO2, PM2.5, PM10 and AQI. The time range is from 2011 to current. Available from https://www.epa.gov/outdoor-air-quality-data/download-daily-data.
China Air Quality Historical Data: This dataset includes the historical daily air quality data for many cities in China. The data includes reading for Ozone, SO2, CO, NO2, PM2.5, PM10 and AQI. The time range is from 2015 to current.
Available from https://quotsoft.net/air/.
Air Quality Improvement under COVID Visualized: This dataset includes the daily air quality data for Beijing, Hong Kong, London, Paris, Seoul, Shanghai, Singapore and Wuhan. The data includes reading for Ozone, SO2, CO, NO2, PM2.5, and PM10. The time range is from 2017 to Sept 2020. The data has been cleaned by ncydexter on Github. Available from https://github.com/ncydexter/Air-Quality-Improvement-under-COVID-Visualized.
Qualità dell'aria: This dataset includes the daily air quality data for Italy. The data includes readings for Ozone, NOx, CO, NO2, PM2.5, and PM10 in 2020.
Available from https://www.epidata.it/Italia/AQI.html.

COVID Policy Tracker / USA COVID Policy: These datasets are provided by Oxford. They contain systematic information on which governments have taken which measures, and when. The 1st one is worldwide while the 2nd one is specific to the US. The time range is from 1/1/2020 to current.
Available from
https://github.com/OxCGRT/COVID-policy-tracker/blob/master
https://raw.githubusercontent.com/OxCGRT/USA-COVID-policy/master

4 DATA CLEANING AND INTEGRATION

We collect data from three channels mainly:

- Air Quality Improvement under COVID Visualized (AQI under COVID)
- World Air Quality Open Data Platform (WAQI)
- United State Environmental Protection Agency (EPA) Outdoor Air Quality Data
- Oxford COVID-19 Government Response Tracker (Oxford Policy Data)
- China Air Quality Historical Data and Qualità dell'aria

In the preprocessing stage, we performed steps to clean and integrate the data. The raw datasets are stored in city_raw_data folder and the processed data can be found in city_cleaned_data folder. The steps we took are including but not limited to:

- clean irrelevant fields to keep just six pollutants
- compute AQI, append and integrate it into each city's data
- transform and unify datetime
- eliminate low-quality data, replace some of the missing data with substituted values.

For reproducibility, we document the steps performed in a notebook data_cleaning_code.ipynb detailing what we performed to make data reusable.

4.1 Air Quality Improvement under COVID Visualized

Prior to the analysis stage, we kept six city datasets: Beijing, Shanghai, Hong Kong, Wuhan, London, Seoul. Among them, two cities, Singapore and Paris are eliminated due to missing data cells. We used OpenRefine to clean the remaining datasets. In the analysis stage, we decided to completely reject this dataset because data were missing after September 2020. Our original heatmaps and bar charts clearly showed this type of imperfection. Along the process, we gradually replaced this dataset with the World Air Quality (WAQI) dataset. Details of cleaning are noted in the python script.

4.2 World Air Quality Open Data Platform(WAQI)

4.2.1 Prior to the analysis stage. We kept eleven cities or boroughs - Manhattan, Queens, Brooklyn, Bronx, Staten Island for NYC; Los Angeles, Miami, Milan, Rome, Wuhan, Beijing, and Hong Kong. Data from 2019 to current are kept. We exclude waqi-COVID-2015H1.csv, waqi-COVID-2016H1.csv, waqi-COVID-2017H1.csv, waqi-COVID-2018H1.csv since only half-year data is found. The source does not provide ways to find the other half years.
For every other .csv file, we used **OpenRefine** to perform:

(1) Text transform on column Date: value.toDate()
(2) Sort by Date, check the date range
(3) Perform Facet / Filter on blank values, to confirm no missing values
(4) Edit cell -> Cluster and Edit on Specie, to merge miswritten values. i.e. 'wind-gust' and 'wind gust'
(5) Filter on City, to extract data for selected cities

4.2.2 Analysis stage. We added Beijing, Hong Kong, Johannesburg, London, Seoul, Shanghai, and Sydney. The reason is WAQI dataset has better temporal coverage than AQI under COVID dataset. In the .ipynb file, we combined the .csv files for individual years and collected the Date and the measurements for six major pollutants. To handle a small number of missing data cells, we examined them and performed forward-fill. Then we calculated the AQI based on the most significant pollutant for each date. After examining our original heatmaps and bar charts, we added data from the first half-year of 2017 and 2018 since most lockdown periods are in the first half-year. Thus we can analyze and explain our graph in a larger time scope. London data is unique, because it had three lockdowns, and the end of its third lockdown is in 2021. We found this data interesting, therefore rather than excluding it, we chose to include it by fetching 2021's air quality data to update the temporal coverage to March 31st, 2021.

4.2.3 Challenge. Since the dataset has 380 cities around the world, the size is big. Merging Q1-Q4 of each year then feeding them into OpenRefine will cause OpenRefine to halt. We solved this problem by implementing python scripts to merge the .csv files and performed major cleaning steps out of OpenRefine. Although the major cleaning process is done without OpenRefine, the preliminary usage of OpenRefine gave us an overview of the dataset, including its metadata. For example, the clustering function of OpenRefine helped to exclude 'East London' in New Zealand and 'London' in Canada.

4.3 EPA Outdoor Air Quality Data (EPA Data)

4.3.1 Prior to the analysis stage. We selected and downloaded only the data for Greater New York City, Greater Miami, and Greater Los Angeles from EPA website. We excluded Greater Miami at the beginning. The time range is from 2011 to current. EPA data has good quality, so the cleaning process is simple. For each .csv file, we used OpenRefine to perform:

(1) Text transform on column Site Name: value.toTitleCase()
(2) Text transform on column Site Name: value.trim()
(3) Edit cell -> Cluster and Edit on Site Name -> no miswritten values found
(4) Text transform on column Date: value.toDate()
(5) Sort by Date, check the date range

4.3.2 Analysis stage. We found that EPA data has the best quality. Therefore, we used EPA data for three American cities instead. We observed Miami has some interesting data points. Therefore we added it back to the cleaned_data for further discovery. EPA data has each pollutant in individual .csv files. We extracted the dataframe from each .csv file and merged them upon Date, renamed columns, and examined each city to find city-specific issues.

4.3.3 Challenge. We find that data measurements are not in the same frequency for every region. For example, Greater NYC takes measurements of PM10 twice a week, while Greater LA takes measurements of PM10 every day. We forward-filled the missing data cell by assuming the value measured will not change drastically in the future three days, which could lead to a potential discrepancy. In addition, we found the ending dates are not consistent for every region in the dataset. This is because the EPA does not upload the data up to the date when we first started to collect the data. The particular challenge is solved by fetching the newest EPA data and manually setting the end date to Dec 31st, 2020 for all three cities.

4.4 Oxford COVID-19 Government Response Tracker (Oxford Policy Data)

4.4.1 Prior to the analysis stage. Prior to the analysis stage: We focused on policy data in selected countries and regions, China, Italy, Hong Kong, United Kingdom, and California, Florida, New York in the United States. We excluded other city's data, no major quality issues were found.
Steps in OpenRefine:

(1) Pad 'NULL' to blank cells
(2) Filter by CountryName for selected countries
(3) Filter by RegionName for selected states in the US

4.4.2 Analysis stage. Our aim is to use policy data to verify the lockdown periods for selected cities. We added Australia, Brazil, South Africa, and South Korea for a well-rounded global trend. We preserved the country, region, and indicators C6 and C6_flag for containment and closure policies[8]. Later in the process, we included the dataset with notes provided by Oxford to better understand the policy changes.

4.4.3 Challenge. This dataset holds policy data, so it is different from other air quality datasets. Oxford uses a numerical scheme to represent the degrees of containment and closure. We refer to the Oxford Codebook to interpret the policy and find lockdown dates. However, the ambiguity of policies leads to a difficult situation. In most cases, we can identify the lockdown start dates, but not the end dates. Reopening policy usually includes several phases, which is not well documented in the dataset. Also, the dataset only has regional-level data for Brazil, United States, and the United Kingdom. For countries or regions with a smaller size, like South Korea and Hong Kong, country-level data may indicate the correct information. But for larger countries like Australia and China, national policies were not always consistent with city-level policies. Therefore, we did heavy research to manually gather and verify news for cities-wide lockdown periods.

4.5 China Air Quality Historical Data and Qualità dell'aria

Since we found more comprehensive data on selected cities in China and Italy, we decided to exclude these two datasets.

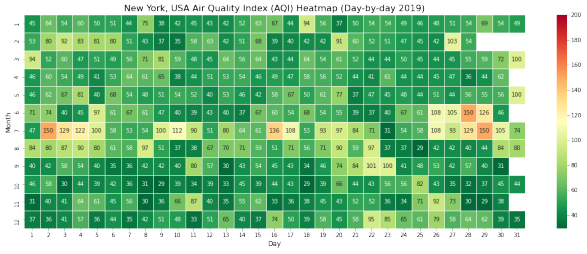
5 GITHUB

https://github.com/vin-lz/cs-gy-6513-big-data-project

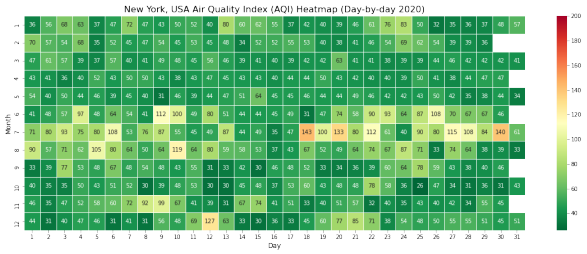
6 DATA ANALYSIS AND VISUALIZATION

6.1 New York City

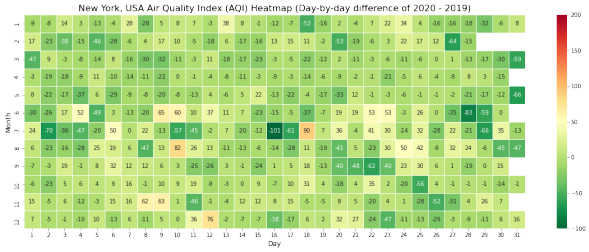
We use Python script to visualize the air quality index(AQI) during various cities' Pandemic lockdown times. We are creating a series of heatmaps, treemaps, and bar charts to try to answer the questions in mind. We take New York City as an example to showcase the visualization and analysis since we live and live through the pandemic here.
Heatmap is a graphical representation of data where the individual values contained in a matrix are represented in colors. It is useful to display a general view of numerical data. In our heatmaps, the X-axis represents the day, the Y-axis represents the months. Each cell represents AQI (0-500) on a specific day. The greater the value, the higher level of air pollution and more health concerns. The health concerns are color-coded, as legend indicates. New York City 2019 and 2020's daily AQI heatmaps are in Fig.2
However, we want to dig deeper to see how lockdown impacts air quality. For that purpose, we create the third heatmap Fig.2(c) to show the AQI difference before and during the lockdown. If one cell has a negative value, it means the value of AQI_2020 is lower than the value of AQI_2019, which further translated to the year 2020 has better quality air compared to the same day in the year 2019, vice versa. In the below heatmap, we can see over 220 cells have negative values. For example on July 16th, 2020, the value -101 means: in 2020, AQI was lowered by 101 (out of 500) compared to what it was in 2019, which is considered as a significant improvement in air quality.
To quantify such change over a year of time, we calculated the AQI change percentile utilizing: (AQI_2019 - AQI_2020) / AQI_2019
In New York City, the change is 17.15%, which is consistent with the previous discovery from heatmaps.



(a) Aqi2019 of NYC



(b) Aqi2020 of NYC

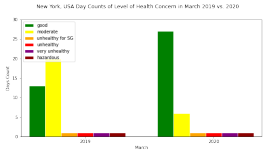


(c) Difference(2020 minus 2019) of NYC

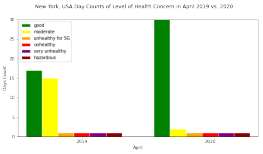
Figure 2: Heatmaps of NYC

Then, we want to explore the AQI monthly change during lockdown periods. Also, we want to use color to enhance the visualization. A bar chart became a natural choice for both purposes because the Bar chart shows the relationship between a numeric and a categorical variable. Each entity of the categorical variable is represented as a bar. The size (in our case, the height) of the bar represents its numeric value.

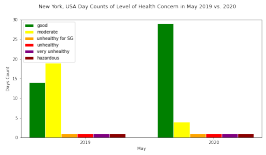
In the bar charts Fig.3, each cell shows the AQI in one particular month during the lockdown period. For New York City, it is March, April, May and June 2020. In each cell, to the left half shows how AQI_2019 be categorized, to the right is for AQI_2020[3]. We see "good" days increase in March, April and May 2020. The reason is obvious: when COVID first hit New York City, all companies announced work-from-home, all schools closed, all businesses shut down, the city went into a serious lockdown situation, with almost no mass traffic, no major industrial production and construction, no major events, the air quality became "good" for 30 days in April and 29 days in May. New York City's phrase1 reopening started June.7. The business started to pick up right after, thus explained the bar chart of June, both green color "good" and yellow color "moderate" return to the same level as for 2019.



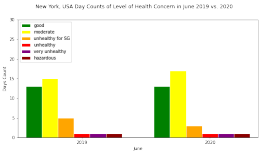
(a) barchart March



(b) barchart April



(c) barchart May



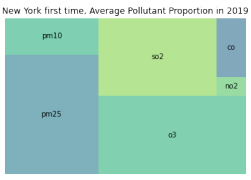
(d) barchart June

Figure 3: Barchart of NYC

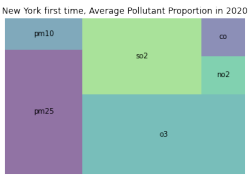
Last but not least, let's take a look at the proportion of air pollutants. Treemap displays hierarchical data as a set of nested rectangles. Each group is represented by a rectangle with an area proportional to its value. Using a treemap here may reveal the correlation between air quality and lockdown life patterns.

From the treemaps Fig.4, we see both PM10 and PM2.5 have a decrease in proportions. As we introduced in the Prior Knowledge part, the activities from construction sites, industries, and automobiles are the main sources contributing to PM10 and PM2.5. Therefore, it's not surprising to see the emission of PM10 and PM2.5 has been significantly reduced during the lockdown period. Besides, we noticed that the proportion of carbon monoxide (CO) dropped more compared to the proportion of nitrogen dioxide (NO2). This can be interpreted as the amount of vehicle's engine idling time drastically reduced during the lockdown since CO pollutant mainly comes from incomplete fuel combustion when vehicle engine idles.[5]

Adopting the same visualization and analytical strategies, we created a series of visuals for each city in the dataset. We will showcase one city in each continent so that we can attempt to answer the question: Does the air quality change in a similar pattern among cities around the world?



(a) treemap of 2019



(b) treemap of 2020

Figure 4: Treemaps of NYC

6.2 The Epicenter - Wuhan, China

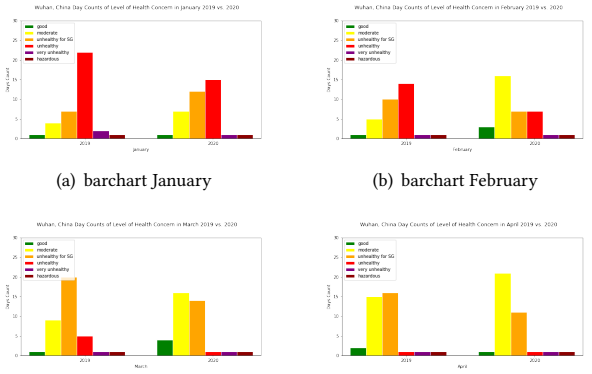
We want to dedicate a part of this report to the epicenter of COVID-19, Wuhan, China. Wuhan city is located in central China, typically experiences worse air quality in winter and spring due to meteorological conditions and central heating. The global outbreak of the coronavirus hit Wuhan drastically at the very beginning of 2020. It has the first major city-mandated lockdown, which lasted from January 23 to April 8, 2020.

From 2019 and 2020's daily AQI heatmaps Fig.5, Jan.24, 2020, and onwards, single-day air quality showed a smaller integer value, which indicates a better and healthier air quality.

The third heatmap Fig.?? shows the AQI difference between 2019 and 2020. Most of the grids had negative values, which means the overall air quality became better in 2020. Also, we can see that 90% of the days with an over 100 AQI decrease concentrate in the lockdown period. Similar to New York City, the strict lockdown has significantly improved the air quality over the period.

To quantify such change over a year of time, we calculated the AQI change percentile utilizing: (AQI_2019 - AQI_2020) / AQI_2019 The change of Wuhan is 22.79%, consistent with the heatmap's discovery.

Furthermore, we create bar charts Fig.6 of the months during the lockdown period to compare the air quality in each particular month in 2019 and 2020. We can see that days with an "unhealthy" label decrease and both "moderate" and "unhealthy for SG" days increase. This trend becomes more obvious in the following months since "moderate" days increase more in the following months than in January. This is because the local government implemented strict lockdown policies and laws to restrict outside travel and closed factories or industry in the Wuhan area and therefore decrease the pollutant emission.



(a) barchart January (b) barchart February

(c) barchart March (d) barchart April

Figure 6: Barchart of Wuhan

From treemap Fig.7, NO2, particulate matter (PM10 and PM2.5) decreased respectively. Ozone proportion's increase is likely due to the decrease in atmospheric NO2 concentrations during this period. The formation of ground-level ozone is not emitted directly into the air but is created by nitrogen oxides and volatile organic compounds under the presence of sunlight.[4] With less particulate matter floating in the air, the strength of sunlight exposure

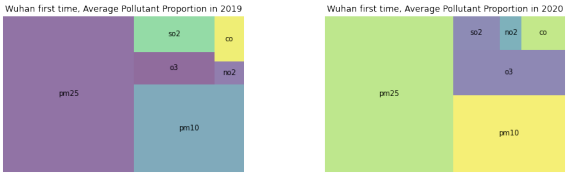
(a) Aqi2019 of Wuhan

(b) Aqi2020 of Wuhan

(c) Difference(2020 minus 2019) of Wuhan

Figure 5: Heatmaps of Wuhan

increases, leading to an increase in proportion. The implementation of strict lockdown measures, such as home quarantining, traffic restrictions, and non-essential enterprise shutdowns, well-aligned and explained the substantial air quality improvement during the COVID lockdown in Wuhan.



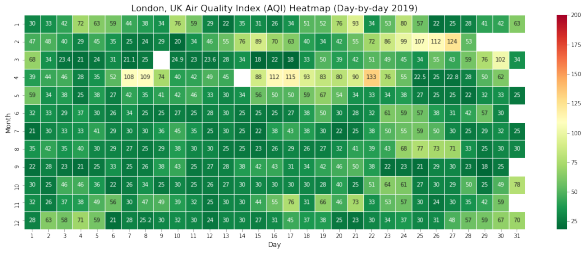
(a) treemap of 2019 (b) treemap of 2020

Figure 7: Treemaps of Wuhan

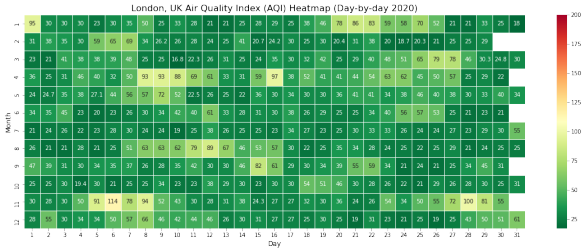
6.3 Los Angeles and London - Multiple Lockdowns

Different from the other two east coast cities in the US, Los Angeles had lockdowns twice. When finding this specific information, we refactored our python code, to make it generally applicable to those multiple lockdown periods. A similar phenomenon exists in London's data, now let's move on to the European continent to explore how multiple lockdowns affect air quality.

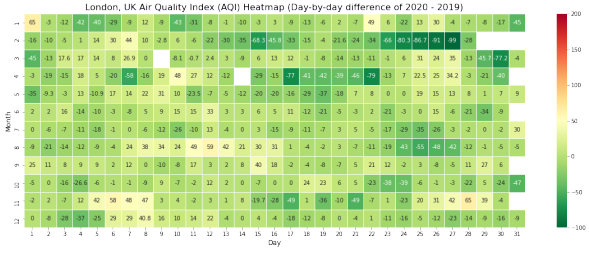
London had three lockdowns over 2020 and 2021. When finding this specific information, we refactored our python script again, to make it applicable to multiple lockdown periods, and it works well. The heatmaps Fig.8 stay almost the same, while we did have three sets of treemaps and bar charts.



(a) Aqi2019 of London



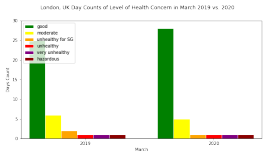
(b) Aqi2020 of London



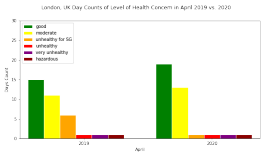
(c) Difference(2020 minus 2019) of London

Figure 8: Heatmaps of London

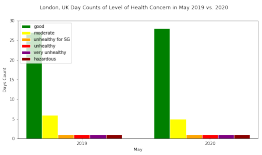
6.3.1 First-time Lockdown. From the bar chart Fig.9 of the first lockdown, we can see the number of "good" days increases, but slightly. The reason behind this is that London always has good air quality compared to other cities. The air quality is good enough on normal days so it has a bottleneck. Therefore, the increment of air quality is subtle but we can still see some improvement.



(a) barchart March



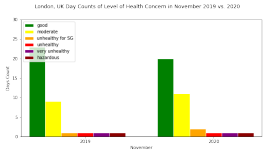
(b) barchart April



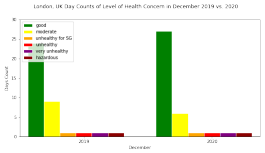
(c) barchart May

Figure 9: Barchart of London, first time

6.3.2 Second-time Lockdown. From the bar charts Fig.10 of the second lockdown, we can see "good" days decreased in November. However, the lockdown has a lagged effect on the air environment. The improvement of air quality would not show immediately when travel restrictions were executed but would show later so we can see the increment of "good" days in December.



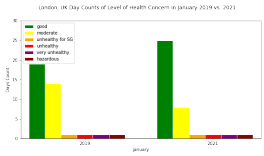
(a) barchart November



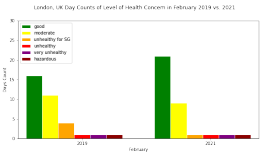
(b) barchart December

Figure 10: Barchart of London, second time

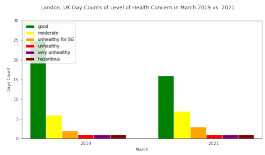
6.3.3 Third-time Lockdown. In terms of the third lockdown Fig.11, the "good" days increase in January and February but decrease in March. That is because the third lockdown has been lifted on 6th, March. There is no lockdown for most days in March 2021. As people start to go out and factories reopen in March, "good" days decrease during that period.



(a) barchart January



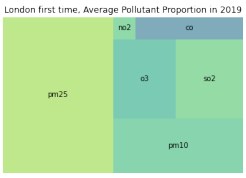
(b) barchart February



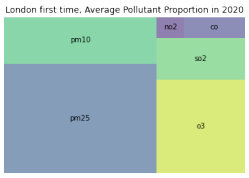
(c) barchart March

Figure 11: Barchart of London, third time

Ozone is an interesting pollutant in London. In the three treemaps Fig.12, Fig.13, Fig.14 we can view the proportional increase of O3, this is because when air quality gets improved, there's less particulate matter (PM10 and PM2.5) in the sky to block sunlight. When the sky clears out, the photochemistry reaction increases the concentration of Ozone. Like in Wuhan, ozone has a trend opposite to proportions of other main pollutants. So with lockdown in effect, fewer social-economical events take place, ozone's increment aligned and explained well of the better air quality.

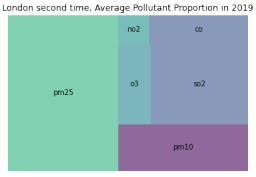


(a) treemap of 2019, first

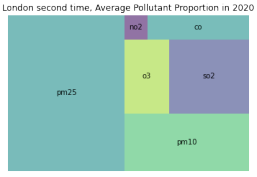


(b) treemap of 2020, first

Figure 12: Treemaps of London, first-time lockdown

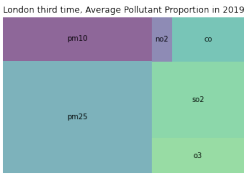


(a) treemap of 2019, second

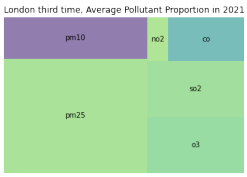


(b) treemap of 2020, second

Figure 13: Treemaps of London, second-time lockdown



(a) treemap of 2019, third

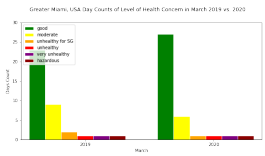


(b) treemap of 2021, third

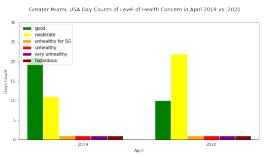
Figure 14: Treemaps of London, third-time lockdown

6.4 Miami - An outlier example

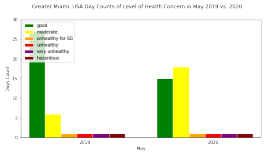
Last but not least, we caught an outlier in our datasets, Miami. The heatmap of 2020 Fig.15(b) has deeper shades of green than 2019 Fig.15(a), which means the air quality of 2020 is better than 2019. To be more precise, the difference between 2020 and 2019 is shown in the heat map below. The air quality of each month turned better but not that obvious in June and July. Miami's situation is unique, located in a no-state-tax state, with a modern metropolitan lifestyle and unbeatable coastal scenic, under republican governor Ron Desantis, Miami became THE place to relocate during the peak time of COVID. According to various news sources, Miami is ranked as the No.4 relocation destination during the pandemic, close by two other Florida cities Tampa, and Orlando. With that said, unlike other major cities in our datasets, with certain safety and sanity measurements, Miami witnessed more traffic, business, home-buying and social events[6][7]. From bar charts, we see in April and May 2020, the green days decreased from 2019. The treemap Fig.17 proportion changes are consistent with such life patterns.



(a) barchart March

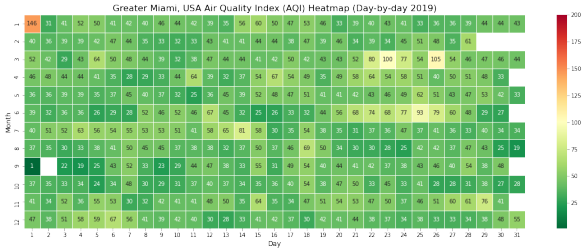


(b) barchart April

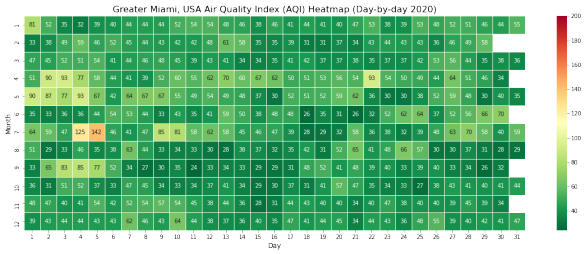


(c) barchart May

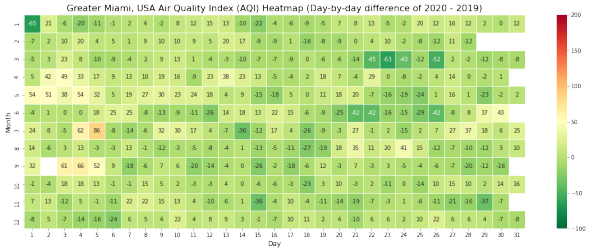
Figure 16: Barchart of Miami



(a) Aqi2019 of Miami

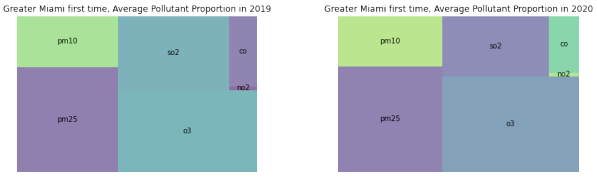


(b) Aqi2020 of Miami



(c) Difference(2020 minus 2019) of Miami

Figure 15: Heatmaps of Miami



(a) treemap of 2019 (b) treemap of 2020

Figure 17: Treemaps of Miami

6.5 Others

Adopting the same visualization and analytical strategies, we created a series of visuals and AQI change percentile for cities in our dataset. Below is an overview of the AQI change:
Recall that AQI change percentile is (AQI_2019 - AQI_2020) / AQI_2019

City	AQI change(%)
Wuhan	22.79
Beijing	7.62
Seoul	10.28
Hong Kong	18.92
New York City	17.15
Miami	-22.52
Los Angeles, first lockdown	-31.33
Los Angeles, second lockdown	-9.65
London, first lockdown	14.10
London, second lockdown	-14.39
London, third lockdown	14.76
Milan	-14.85
Rome	-7.74
Johannesburg	23.07
Sao Paulo	7.41
Sydney	30.80

Due to large quantities of graphs and the limited space, we grouped all graphs according to city and it's lockdown characteristics into the above four groups. See the complete sets of visualization generated here:
https://github.com/vin-lz/cs-gy-6513-big-data-project

7 CONCLUSIONS

(1) In major cities around the world, lockdown has significant impacts on air quality.
(2) There exists a change in major air pollutants components, mainly due to reduced vehicular emissions and limited industrial activities.
(3) Policy has a strong correlation with air quality. Cities imposed strict lockdown measurement had seen better air quality over the lockdown period.
(4) Uncertainty remains when analyzing data from cities that did not have clear lockdown measures, starting and ending dates.
(5) Special events, such as the wildfire in California, relocation wave towards Florida create data outlier in this project.

8 LIMITATIONS AND CHALLENGES

(1) **Data Size** - The amount of data we are handling is pseudo big-data. Since most of the air quality data is reported daily, from 2017 to mid-2020 when most of the lockdowns lifted, the data we are actually collecting is about a few thousand days. For each city, we focused on six major pollutants, concatenated with the Air Quality Index, which brings the data frame to just about seven columns.
(2) **Reopening Date** - Some cities do not impose a strict lockdown policy, for example, said one South Korean Health Administer: "All important actions were taken before the World Health Organization declared the COVID-19 outbreak a public health emergency of international concern on January 30, 2020", Beijing has strict measures to stop the virus spreading, but it was not a lockdown, it's called close-off management which dynamically changed the lockdown area

and time. In cities like Sydney, Sao Paulo and Hong Kong, no clear lockdown start date and end date were discovered.

(3) **Moving factors** - Different cities have different characteristics, therefore have different life patterns and measures to live through Pandemic. In some cities people stay in, in some cities people tend to drive personal vehicles than use public transportation, some cities have multiple lockdowns and reopens, while some never pause their usual life. The whole Pandemic lifestyle is dynamic, all those bring challenges to answer the question: how COVID impacts the air quality.

(4) **Special Events** - Data of Asian country cities in 2019 mid-February could be interfered with by Lunar New Year. Take Beijing as an example, during the Lunar New Year period (2019-02-05 2019-02-15), most of the migrant workers are in their hometown, local residents either stayed in or traveled elsewhere. Surprisingly enough, Lunar New Year becomes a period of Beijing's downtime, with less traffic and limited businesses, therefore could lead to healthier air quality in 2019 compared to 2020. Other special events include the California wildfire in 2020, which brought down the AQI change index by 31.33%.

REFERENCES

[1] Air Quality Index Color Code Guide, webcam.srs.fs.fed.us/test/AQI.shtml.

[2] "AQI Basics." AQI Basics | AirNow.gov, AirNow.gov, U.S. EPA, www.airnow.gov/aqi/aqi-basics/.

[3] "Air Quality Lookup for Children, Seniors and Sensitive Groups." Clean Air Resources, www.cleanairresources.com/tools/aqi_lookup.

[4] "Ground-Level Ozone Basics." EPA, Environmental Protection Agency, 28 Apr. 2021, www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics#wwh.

[5] "Motorists Who Leave Engines Idling Risk Carbon Monoxide Poisoning." TODAYonline, www.todayonline.com/voices/motorists-who-leave-engines-idling-risk-carbon-monoxide-poisoning.

[6] October 20, 2020·112 Comments. "The Number Of People Moving To Florida Is Surging Due To The Pandemic, With Miami Luxury Home Prices Up 42% Last Quarter." The Next Miami, 20 Oct. 2020, www.thenextmiami.com/the-number-of-people-moving-to-florida-appears-to-be-surging-due-to-the-pandemic-with-miami-luxury-home-prices-up-42-last-quarter/.

[7] October 28, 2020 "Everybody is moving to Miami during the pandemic. Honestly, we'd rather you didn't" Miami Herald, 28 Oct. 2020, https://www.miamiherald.com/miami-com/funny-stories/article246774322.html.

[8] OxCGRT. "OxCGRT/COVID-Policy-Tracker." GitHub, 19 Mar. 2021, github.com/OxCGRT/COVID-policy-tracker/blob/master/documentation/codebook.md.