CSC110 Project Report:
Investigating the Impact of COVID-19 on City Air Quality Using AQHI

Umair Hussain, Pete Chen, Roy Mao, Jazli Muhammad Khairi Leong

Friday, November 5, 2021

Introduction

Climate change is a problem that has been increasingly prevalent for the last century. In the 21st century, climate change is a problem that is recognized internationally. The issue is so critical that it has been placed as one of the UN's global issues. Climate change concerns how the rising temperatures of the planet contribute to total planet deterioration in many facets. Combining these problems leads to our question of **how has the pandemic impacted the air quality within several cities across the world?** Our motivations stem from our desire to save the planet. In addition, to preventing a potential catastrophe. We hope that finding the answer to this matter will allow us to develop new ideas which enable us to execute more efficient and effective solutions to climate change. The query we have selected investigates how the COVID-19 pandemic has affected climate change and, in particular, air quality in cities. Air quality has a big impact on climate change. Our goal is to tackle one of the leading issues regarding climate change. To analyze the aftereffect, we will require measurements to assess and compare with statistics before the pandemic. The Canadian government developed an index called the AQHI, which measures "what the air quality around you means to your health" ("About the Air Quality Health Index", n.d.). The scale ranges from 1 to 10 where the higher the number is, the greater the risk is to one's health. COVID-19 changed our lifestyle and, in turn, the way we consume non-renewable resources. One of the many changes that occurred was the move to remote work and learning. This shift had omitted the use of transportation and office spaces. The global step away from these facilities and general transportation may have consequently limited the number of greenhouse gasses emitted. On the contrary, this same shift to stay at home could have also contributed to producing more greenhouse gasses than before, as residential heating and electricity consumption may have increased during the pandemic. Our questions stem from the curiosity of understanding if whether the possible lower carbon footprint that arose from the shift away from transportation and remote work outweighed the greenhouse emissions that resulted from heightened residential electric and heating expenditures.

Datasets

One of the datasets we used for this project was sourced from a website by the name of "Kaggle global-air-quality data" and is in a csv format. This is the 10cityaq_data.csv file, and it contains the air quality data for ten different cities for our range of "Post-Covid" Times (late 2020 - 2021). Within this CSV dataset, specific information such as the date, country, city, the specie of pollutant, the minimum, maximum, median amount of the specie on that specific day, and the variance. Parts of the dataset we used for our project include the date, country, city, specie of pollutant and median, but we excluded the other data for our calculations. Then, the other datasets we included, which are in the format ("city" - air - quality) are also in CSV format and were sourced from the website "Air Quality Historical Data Platform". These datasets include information such as the date and the levels of pm25, pm10, o3, no2, so2, co. For these datasets, the information we used within our program only span from the years of 2018 - 2020, and we only used ground-level ozone (O_3), fine particulate matter, ($PM_{2.5}$) and nitrogen dioxide (NO_2) as the substances.

Computational Overview

Our project combines data filtering with implementing algorithms as these our our two main computational components of the project. The data set we will be using needs to be cleaned and rearranged differently to allow us the ability to calculate AQHI values for each day. We imported the python module 'pandas' to help us with mutating and manipulating csv files. Our first step was to read and filter the "10cityaq_data.csv" file using pd.read. This file

includes data from January 2021 to May 2021 as it is the most recently available data that is publicly accessible on the internet that we could find. We dub this time range to be the post-COVID period. We achieved this through our filter.py program. We searched through the 'Date' column of the csv file and converted each of the strings representing the data to a datetime format. Next we used the .loc method to separated or filtered the air quality data based on the city and saved it to a variable that corresponds to the name of the city. Information regarding 'NO_3' (nitrogen dioxide), '0_3' (ozone) and '$PM_{2.5}$' (fine particulate matter) were then extracted or 'filtered' from the dataset. Afterwards each dataframe was converted to a new csv file using the .to_csv method naming each in the form of "AQ of ___.csv" where the underscore represents the city. Afterwards our aqhi_calculation.py file reads each city's AQ (air quality) file. Pandas then gathered the data for NO_3, 0_3 and $PM_{2.5}$ from each AQ file and then extracted the median as a substitute for the mean, as only the minimum, maximum and median were given. The median performs theoretically performs better than the average as oftentimes the average may be skewed by an outlier value that lies outside several standard deviations of the mean. Then an assert statement is used to ensure that no data is missing. The data is then passed into a function named "___aqhi_cal" as the underscore corresponds an abbreviated form of the city that returns a list of the calculated AQHI values for each date. Using this formula ("Air Quality Health Index (Canada)", 2021):

$$AQHI = (\frac{1000}{10.4}) \times [(e^{0.000537 \times O_3} - 1) + (e^{0.000871 \times NO_2} - 1) + (e^{0.000487 \times PM_{2.5}} - 1)]$$

The list is then sorted and then converted to a csv file through pandas. This process is repeated for each of the ten cities which can be ran through a function called "create_csv_for_all()" when called.

Next the same processes will be done but for data from before COVID-19 as it takes data from each earlier than 2021. We can then compare the data and assess the impact of COVID-19 on air quality using the AQHI values. Files that have the same naming convention as "beijing-air-quality.csv" contain the data from the pre-COVID era. The website that we downloaded the information provided the information in separate files. The filter_pre_pandemic_and_AQHI_calculation.py file uses each city's pre-COVID data csv files. Each csv file is then read using and sorted by date using pandas. This data is then filtered for dates between January 2018 to January 2020. The data is then filtered again for ground-level ozone (O_3), fine particulate matter, ($PM_{2.5}$) and nitrogen dioxide (NO_2). However, much of the data was missing as many whitespaces or wholes appeared in the dataset. This was solved using pandas .fillna with input "method=ffill" as it uses the previous days' data to replace the missing data. This data was then passed into the same algorithm or function that calculates AQHI levels for each date named "pre____aqhi_cal" as the underscore corresponds an abbreviated form of the city. Then it was each was passed into another function to convert the list into a dataframe and then into a csv file using pandas.

The next phase of the project was to present the data in a visual manner and to achieve this we used the python library 'plotly'. csv_to_plots.py imports pandas and plotly's graph_objects and plotly subplots to help achieve this vision. The program then reads each post-COVID AQHI csv files and defines a function that creates a graph comprised of ten different subplots. Each of these subplots are a go.Scatter line plot mapping the date for the x-axis to the AQHI value on the y-axis. Each of these graphs are semi-interactive as plotly allows one to hover their mouse over important parts of the graph and see the date with its AQHI value.

The final program called precovid_data_to_graph.py is similar to the csv_to_plots.py file but instead graphs pre-COVID AQHI data. The file reads each of the 2018-2020 AQHI data of each city through the use of pandas. A line plot is then made for each of the cities that uses the go.Scatter module from the plotly library mapping the date of the x axis to the AQHI value for that date on the y-axis. Ten different graphs are then produced through plotly because the datasets are larger thus a larger graph would visually represent the data better. Similar to the post-COVID subplots, each of these graphs are semi-interactive as plotly allows one to hover their mouse over important parts of the graph and see the date with its AQHI value.

Our program uses new libraries such as plotly and pandas to help us manipulate and create csv files. This allowed us to accurately store and arrange data in a specific manner that was organized. Using pandas specifically granted us the ability to extract data from csv files and filter through them to grab the needed data that was required to perform AQHI calculations. The plotly library allows us to visually represent the data in an effective manner that allows the reader or user of the program can understand. Through the application of these tools, we can extract all data for a particular city and, additionally, all of the data for a certain air pollutant specie. To expand this, we can further draw out precise values of that same pollutant specie that corresponds to a specific date. Pandas also provides the ability to write data into another CSV file, which allows us to add our final calculated AQHI data to the new CSV file to then convert them into graphs (pandas also has a plotting functionality, but we prefer using Plotly because it is more detailed) specific.

Obtaining the Data Sets and Running The Program

Check requirements.txt for all required libraries.

For our datasets our post-COVID data was obtained and downloaded through the following website: https://www.kaggle.com/sohelranaccselab/global-air-quality-data. This website contained the data related to our research from late 2020 - 2021. However the data set was quite large involving hundreds cities and locations and was too much data to work with. So we decided to use 10 major cities for our analysis and created a smaller data set from their large one (accessed through the link below). For our pre-covid data it was obtained through this website https://aqicn.org/data-platform/register/. This website contained information from 2020 - 2015 (depending on the city). However there were white spaces throughout the csv file making it hard for pandas to navigate. We decided to manually remove these white spaces and recreate the data set. We have included a link to a UTSend pickup:

Claim ID: GvpKCmpKKxHvSyJb, Claim Passcode: htFwioymKwRzhZ3z.

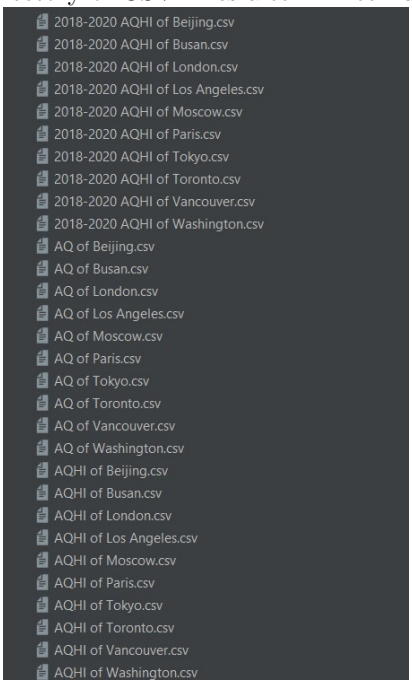
And just in case the UTSend service does not work here is a Google Drive link with all of the csv files that were preprocessed:

https://drive.google.com/drive/folders/1vTzi5nNSrW5x8H7UPzSXnwRXWvCyBHEv. These are the csv files required to run our program correctly.

HOW TO RUN:

When in main.py is ran, the users should run the code in a specific sequence. The file should be run with a python console as the user needs to import files in a certain order as the functions depend on files that were created in the previous function. The user should type "import filter as filter" and run "filter.create_all_new_csv()". After this is ran the user should see ten new csv files that were created named 'AQ of ____.csv' pertaining to the city. Then the user should run import AQHI_calculation as aqhi_calc and aqhi_calc.create_aqhi_csv_for_all(). After this function is ran there should be ten csv files that are named 'AQHI of ____.csv' pertaining to the city. Next the user needs to run "import filter_pre_pandemic_and_aqhi_calculation as prefilter" and run "prefilter.create_18_20_data_csv()" creating ten more csv files named "2018-2020 AQHI of ____.csv" files. One's directory should look something like this after each of the functions are ran:

Figure 1: Directory of CSV Files after Three Functions Ran

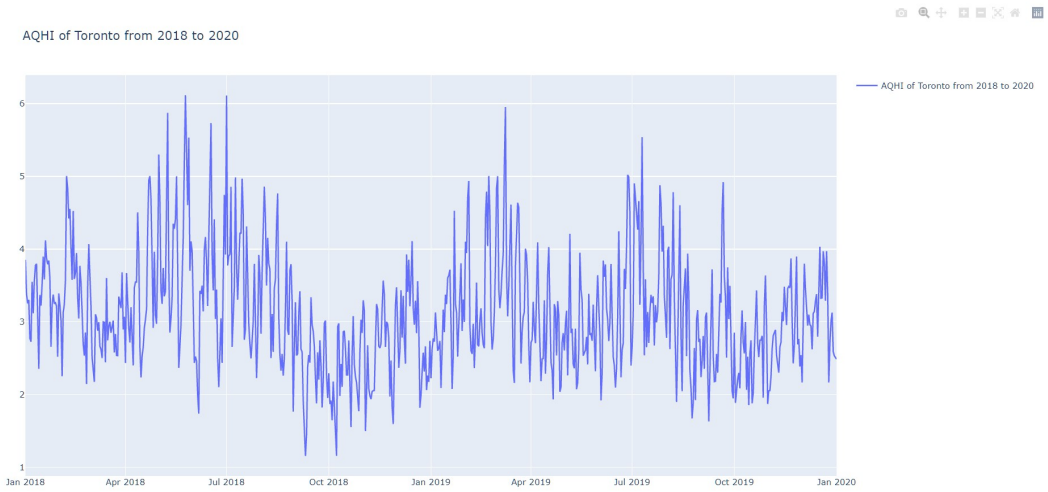


Then to create the graphs of post-COVID data the user needs to run "import csv_to_plots as plot" and "plot.create_graph()", this will open a new tab of plotly that displays ten different supblots that graph the date to AQHI of that day. This data should look like this:

Figure 2: AQHI of Ten Cities from January 2021 to May 2021

Finally to see pre-COVID data represented visually the user runs "import precovid_data_to_graph as pregraph" and run "pregraph.graph_all_18_20()" after this ten tabs will open that are represent graphs each city's pre-COVID-19 era air quality data mapping the date to the AQHI value. Several tabs should be open similar to this example of a graph which look like this

Figure 3: AQHI of Toronto from 2018 to 2020



Changes to Project Plan

Between the project proposal and now, changes we made to our project include getting a brand new data set for pre-covid figures to compare and contrast the effect of the pandemic on the air quality. Something else we had to incorporate into our project after our original proposal was creating two separate filters for post-COVID and pre-COVID, because the two sets had different formatting and so it wasn't a good idea to have a "one-filter-fits-all" approach we had when we were initially planning the project. Overall, there were a lot of adjustments we made from our original proposal as we progressed through the project.

Discussion

The results of our computational exploration produced twenty graphs that represented and plotted AQHI values from January 2018 to January 2020 to represent pre-COVID data and from January 2021 to May 2021 portray post-COVID data. The first set of graphs that were produced were the post-COVID data graphs. For Tokyo, the AQHI ranged from a value of 1.28 to 5.21 and pass March seemed to rarely exceed an AQHI level of 4. London seemed to portray a similar story ranging from a AQHI value of 1.51 to 4.57, and looked like the average was around 2 to 3. Los Angeles ranged from AQHI values of 1.97 to 5.31 and averages around 3.2. Washington's AQHI data ranges from 0.96 to 2.81 as this gives a low average of AQHI of around 1.8. Beijing produces a much higher range of values from 1.65 to 8.10 as the average is around is eyeballed to be around 4.2, this would make sense because of the high population leading to the amount of greenhouse gasses being emitted. Busan has a range an AQHI value of 1.53 to 6.10 and has an average of around 3.5. Toronto has a range AQHI values of 1.07 to 3.69 and an average of roughly 2.3. Vancouver has one of the lowest AQHI values and is therefore has the best air quality out of all of the ten cities. Vancouver has an average AQHI value of 0.9 and ranges from 0.54 to a maximum of 1.33 proving it to be the lowest AQHI city. Paris presented a case where the AQHI values decreased significantly as the values ranged from 1.75 to 5.50. Paris seemed to have an average of 3.3 before May but afterwards decreased significantly to its lowest value of 1.75. The last city of Moscow had AQHI values of 1.83 to 5.40 and had an eyeballed mean of around 3.1. When examining the pre-COVID data it can be seen that generally the AQHI values are a wider range of variance. This larger range makes sense as a wider range of time would lead to a higher minimum and maximum for the range. Though a significant difference lies in the actual AQHI values of the graph as most of the graphs present a higher average AQHI value than its post-COVID counterparts. Two graphs that have incomplete data are Tokyo and Moscow as the original downloaded dataset only had data from May 2018 and July 2019 respectively. By choosing a diverse set of cities from across the world that all represented similar data, it showed that a universal trend can be found. This trend can be identified to be that AQHI values from before COVID-19 are higher than AQHI values during COVID or post-COVID times. Thus our conclusion can be solidified that AQHI values are by COVID-19 in some particular way, and therefore the COVID-19 pandemic has affected air quality where it has reduced greenhouse gas emissions because of the lower AQHI values. Our computational exploration has helped us solve the find and answer to our questions to confidently answer our problem.

Some limitations we encountered include dealing with incomplete and insufficient datasets. Originally our idea was to use more than ten cities and analyze the results of each of them, although much data regarding air quality was could not be found. Many of our datasets for each city had missing data entries and we had to improvise by using the previous day's data and insert them into those missing data entries. Future improvements and next steps for further exploration regards the expansion of the city and additional detail. Gathering data from more cities and creating graphs for each one would allow us to find a more definitive answer. Delving into greater detail on the analysis of each graph rather than just rough estimates would be to calculate the real mean, population mean, median, range, standard deviation would also grant us the ability to present greater detail of our calculated data. Using real statistical analysis would allow for the results and analysis to be more scientific and more grounded in truth. Gathering more complete data that was from February 2020 to December 2020 would help us fill the gap in the timeline as we did not possess that data to create a complete continuous graph. It would assist in creating a better idea of what the effects of the pandemic were on air quality and generally give more data to be assessed and analyzed as more data leads to more answers.

Overall our problem question was answered through our computational explorations. We concluded that air quality has been positively affected because of COVID-19 and we can prove this through the data that was created. Limitations that we faced concern mainly missing data within the datasets. General improvements would be to increase our level of detail in analysis and to gather more data for us to put into the algorithm and return graphs.

References

"About the Air Quality Health Index." *Government of Canada.*
https://www.canada.ca/en/environment-climate-change/services/air-quality-health-index/about.html.

"Air Quality Health Index (Canada)." *Wikipedia.* Wikipedia Foundation. 3 April 2021,
https://en.wikipedia.org/wiki/Air_Quality_Health_Index_(Canada)

"Air Quality Historical Data Platform." Aqicn.org, 2021,
https://aqicn.org/data-platform/register/.

IEA, Global CO2 emissions in transport by mode in the Sustainable Development Scenario, 2000-2070, IEA, Paris, 25 May 2021,
https://www.iea.org/data-and-statistics/charts/global-co2-emissions-in-transport-by-mode-in-the-sustainable-

development-scenario-2000-2070
Rana, Sohel. “Covid-19 Air Quality Worldwide -2021.” Kaggle, 27 May 2021, https://www.kaggle.com/sohelranaccselab/global-air-quality-data.