Effects of the COVID-19 Pandemic on the Environment

Team Name: Reduced to the Home
Alex Biehl*
Max Christ*
Divya Pathai*
alex.biehl@nyu.edu
mmc639@nyu.edu
dp3170@nyu.edu
New York University
New York City, NY, USA

ABSTRACT

The health of our environment is of the utmost importance for the future of all living things on this planet. The COVID-19 pandemic changed many aspects of our society, and has offered a unique opportunity to examine how these changes may have impacted the natural world. Using the power of Big Data frameworks like MapReduce and Spark, this paper will explore environmental related datasets and try to examine what changes might have occurred during the course of lockdowns and work-from-home shifts during 2020. Furthermore, it will attempt to evaluate the longevity of any of these changes. Due to availability of data, and in the interest of limiting scope, the project will focus on the environmental impacts in the New York City metro area.

KEYWORDS

datasets, COVID-19, work from home, pandemic, environment, climate

1 INTRODUCTION

The pandemic has caused a huge shift in working habits for millions of people. Specifically, many people began to work from home on a semi-permanent to permanent basis. This has wide ranging impacts on society, and we want to investigate what impacts it may have on the environment. We expect there may be interesting/surprising conclusions to draw from this data, along with obvious positive impacts, for example, reduced numbers of cars on the road. There will also be a balance of negative impacts, such as increased waste, increased home energy consumption, and an increase in the number of shipped goods. We hope to find more interesting relationships hidden within the data. Some questions we hope to explore include:

- Will the number of people who are no longer commuting regularly offset the some of the other negative impacts of the pandemic?
- Has there been a significant improvement in air quality, and do we expect any changes to be long lasting?
- Has there been any significant reduction in commercial energy costs due to empty offices?
- Has the shift to remote work had any lasting effects on the volume of traffic going in and out of the city?
- Have there been any lasting geographic population shifts that may ease the environmental stress of cities?

The goal of this paper is to investigate these questions by analysing large datasets that are publicly available on the web. The process is broken down into three steps: dataset discovery, data cleaning and wrangling, and data analysis/visualization. During the first step, an initial list of potential datasets was created, and then iteratively modified and paired down until a compact list of relevant datasets was reached. During the second step, the team used the OpenRefine application to analyze the quality of the data. This was used as a

*All authors contributed equally to this research.

starting point, and then custom scripts were created in Google Collab or Jupyter notebooks to clean the data with OpenClean. All of the code and original datasets can be found at the github repository: **https://github.com/divyap2706/Big-Data-Project-X**.

2 FINAL LIST OF DATASETS

After an iterative dataset search, the final list of datasets was settled on and is described below.

2.1 Vehicle Travel in and out of NYC

The "Vehicle Travel in and out of NYC" dataset is from the NYC OpenData site, and provides information about the volume of vehicle traffic on the city's bridges and tunnels. Specifically, it provides hourly vehicle totals for each bridge/tunnel for each direction (incoming and outgoing). The dataset spans starts in January 2010 and ends in December 2020. The dataset is located at the following link:

https://catalog.data.gov/dataset/hourly-traffic-on-metropolitan-transportation-authority-mta-bridges-and-tunnels-beginning-

2.2 National Energy Usage

The "National Energy Usage" dataset is produced by the US Energy Information Administration. It describes the energy usage of various industries and describes the energy sources in each case. The data can be found at the following link:

https://www.eia.gov/opendata/

2.3 Sidewalk Cafe Licenses and Applications

The "Sidewalk Cafe Licenses and Applications" dataset is distrubuted by NYC OpenData and lists all applications for outdoor, sidewalk cafe licenses in NYC. Specifically it provides information on approval status, number of chairs/tables, square footage, and length of time before final decision. The data spans 2018-2021. The data can be found at the following link:

https://data.cityofnewyork.us/Business/Sidewalk-Caf-Licenses-and-Applications/qcdj-rwhu

2.4 Real Estate Listings and 'Hotness' by Area

The "Real Estate Listings and 'Hotness' by Area" is created by Realtor.com, and provides real estate data by zipcode on a national level. For each zipcode, the number of listings, the average time a property is listed, and the relative "hotness" (or buying popularity) rating of the zipcode is listed on a monthly basis. The data encompasses the past few years up until 2021. The data can be accessed at:

https://www.realtor.com/research/data/

2.5 EPA Air Quality System (AQS)

The EPA Air Quality System (AQS) (https://www.epa.gov/aqs) contains ambient air pollution data collected by EPA, state, local, and tribal air pollution control agencies from over thousands of monitors. AQS also contains meteorological data, descriptive information about each monitoring station (including its geographic location and its operator), and data quality assurance/quality control information. AQS data is used to:

- assess air quality
- assist in attainment/non-attainment designations
- evaluate State Implementation Plans for non-attainment areas
- perform modeling for permit review analysis
- prepare reports for Congress as mandated by the Clean Air Act

Our dataset comprises of rows received from the Daily Summary Data API (https://aqs.epa.gov/aqsweb/documents/data_api.html#daily) for all pollutants defined in the Air Quality Index (AQI) Pollutant Parameter class, for New York county, NY; years 2019-2020. These pollutants include:

- Carbon Monoxide
- Ozone
- PM10
- PM2.5 - Local Conditions
- Acceptable PM2.5 AQI & Speciation Mass

2.6 EPA Water Quality Portal (WQP)

The Water Quality Portal (WQP) is a cooperative service sponsored by the United States Geological Survey (USGS), the Environmental Protection Agency (EPA), and the National Water Quality Monitoring Council (NWQMC). It serves data collected by over 400 state, federal, tribal, and local agencies.

The portal allows users to filter the data based on many criteria, and then download the filtered data directly to a .csv file. Since this project is focused on the environmental impacts on NYC, the country was limited to the US, the state was limited to NY, and the counties were limited to the 5 boroughs of NYC (US:NY:005, US:NY:047, US:NY:061, US:NY:081, and US:NY:085). We decided not to focus on measurements from rivers or ground water due to the limited number of sites. Therefore, the data was limited to ocean and estuary measurement sites. Since we were looking for water related measurements, we selected water as the only sample media. We chose a date range of 01/01/2018-4/11/2021 so that there was sufficient context. Then for the characteristics field, we selected values that were of interest (Chlorophyll a, Chlorophyll a (probe), Chlorophyll a corrected for pheophytin, Dissolved oxygen saturation, Dissolved oxygen (DO), Enterococcus, Nitrate, and Nitrite). Finally, for "data to download," we selected "Sample results (physical/chemical metadata)," and comma separated as the format. A PDF was added to Big-Data-Project-X/data/raw/ which is a screenshot of the parameters used on the website to retrieve the data. https://www.waterqualitydata.us/

3 DATA CLEANING AND INTEGRATION PROCEDURE

After selecting the final datasets, the team used OpenRefine to examine the quality of the data. Specifically, for each column in each dataset, a facet was created to see if there were anomalies in categorical data, like malformed values. Additionally, for each column, the cells were clustered to see if there were any values that should be merged. These steps were used as a starting point to gather quality information, and then custom data cleaning scripts were written for each dataset using the OpenClean Python library. All of these scripts were written as Google Collab or Jupyter notebooks under normal conditions. All of the necessary libraries are !pip

installed to ensure that the notebooks can be reproduced by others. Often OpenClean's stream function was used to further extend the notebook's reproducibility to users who may not have much computing power to run the notebooks. Additionally, OpenClean's own profiling tools were used extensively in addition to the initial profiling done in OpenRefine. Exact data cleaning challenges and procedures will be outlined below in detail.

3.1 Vehicle Travel in and out of NYC

One of the main challenges of this dataset was the fact that the Plaza ID's for all of the bridges and tunnels changed in 2017. This was discovered by looking at the individual max and min of the date column for each of the Plaza IDs. In order to fix this, research into the database was needed. After looking at the data definitions on the original website, a mapping function was created to map the original Plaza IDs to the new Plaza IDs. This mapping function was applied as an argument to the update function in OpenClean.

The team also uncovered another quality issue: two Plaza ID's only had data for one direction (either inbound or outbound). In order to prevent potential misunderstandings in the data, Plaza IDs that only had a single direction worth of data were filtered out of the dataset.

3.2 National Energy Usage

This dataset had a few data quality issues. All the values in the month column were decimal. For example January was 1.0, February was 2.0 and so on. To fix this, replace transformation was used in OpenRefine. Another issue was that the values in consumption column had trailing zeros after the decimal point. Regex was used to remove the zeros. The year column had some blank values which are now excluded from the dataset.The instructions to clean the data set using OpenRefine have been mentioned in the text file Instructions uploaded on the github repository.

3.3 Sidewalk Cafe Licenses and Applications

There were a couple data quality issues in this dataset. The first was that the City column had many different capitalization's of the same string. To solve this, all of the values were changed to uppercase. This eliminated all of the clusters, except for one. The one cluster that remained, was there were two spellings of "Long Island City." This was manually fixed by using the update function in OpenClean. The last data quality issue was: the Business Name column had clusters for the same value with different capitalization's. This was fixed by using the upper function in conjunction with OpenClean's update function. There were also clusters in the Street column, but the Street column is not of interest to this project. Therefore, we left the Street column as is. Lastly, the Building column had multiple datatypes. This was verified as correct, because the Building can be any of the following formats: "1234," "100A," "200-300," etc.

3.4 Real Estate Listings and 'Hotness' by Area

In both of the sister datasets from Realtor.com, there were similar problems. The first, was that there was one row in each dataset that contained information related to the data. In other words, the row provided extra annotations of the data. This was discovered when analyzing the datatypes of each column in the OpenClean default profiler. OpenRefine was used to discover the exact value of the key, and then the row with that key was filtered out using the filter function in OpenClean. The second issue in both datasets, was that there were a few rows that contained both floats and integers. This was fixed by creating a custom function to use as an argument to the 'update' function, to cast the values to floats if need be.

One additional challenge, was that there were two clusters on the city column. In each case, it was a city name that had two words. For example, "park forest, il." The other value it was clustered with was the same string, but with the two words of the city reversed. For example, "forest park, il." A google search verified that these are both indeed valid cities. Both of the datasets also had negative numbers in certain columns, and this was verified as correct after researching the origins of the data columns on Realtor.com.

3.5 EPA Air Quality System (AQS)

The dataset retrieved from AQS was profiled and reviewed using a combination of the OpenRefine tool for initial insight gathering and OpenClean for more in-depth analysis. Profiling was done initially with the DefaultColumnProfiler to get a basic idea of the dataset's statistics. All columns except for two, aqi and pollutant_standard, had no empty or null values.

An output of the distinct values of both the aqi and pollutant_standard columns showed that aside from the empty cells, the data appeared to be regular and consistent. The aqi (Air Quality Index) column contained integers between 0 and 136, and the pollutant_standard column contained values such as Ozone 1-hour 1979, Ozone 8-Hour 2008, and PM25 24-hour 2006. According to the AQS data dictionary (https://aqs.epa.gov/aqsweb/documents/AQS_Data_Dictionary.html), AQI values are a unitless measure of the amount of pollutant that can be used to relate the pollutant to the healthy levels and indicate possible health concerns with elevated levels. Values of 0 through 50 are considered good, 50 through 100 are considered moderate, etc. With no significant outliers in the range of values present, the team created two datasets from the original–one with all null aqi values and one with no null values–and retrieved the set of distinct parameter values from the associated rows of each dataset. Comparing the two lists of distinct parameter values, we found overlaps on Carbon monoxide, Ozone, and Acceptable PM2.5. While initially concerning, checking the distinct Sample Durations for each filtered set showed that all the rows with an empty AQI were set to 1 Hour, while the Sample Duration of all rows with non-null AQI's were either 8-hour or 24-hour averages.

Next the team checked for functional violations between the parameter, parameter code, and pollutant standard rows. Three sets of dependencies were discovered, one for Carbon monoxide, one for Ozone, and one for PM2.5 - Local Conditions. After reviewing the rows found in each dependency, the team determined that there appeared to be no violations between parameters, parameter codes, and pollutant standards.

Lastly, the team checked for uniqueness and clusters in address, state, county, and city columns; all of which appeared to be correct and regular. At this point the team determined that the data was clean enough for use without further wrangling.

3.6 EPA Water Quality Portal (WQP)

The dataset retrieved from the WQP was profiled and reviewed using the OpenRefine tool for insight gathering. While the dataset contained a wealth of useful attributes that we were interested in exploring, we discovered to our dismay that after filtering for the contaminants that we were interested in (Chlorophyll, Dissolved Oxygen, Enterococcus, Nitrate/Nitrite) that there just wasn't sufficient data to be useful. Ideally we would have liked to have data points that spanned 2019, 2020, and into 2021. Unfortunately what we discovered that the dataset only contained measurements dating from Feb 2018 to early 2020. We would have loved to make use of the EPA's water quality data in our analysis, but given the lack of coverage of the time period we were examining, we decided to drop the WQP data from our final analysis.

4 DATA ANALYSIS AND VISUALIZATION

The main strategy for analysis and visualization was to aggregate the data to a manageable time scale (i.e. plot monthly or weekly data over the course of one year) and then compare the data from 2020 to previous years, as well as the other datasets, to look for interesting correlations. PySpark was the main method used for aggregating the data, and gathering quick metrics on the data. Jupyter Notebooks proved useful for this purpose, because it allowed for simultaneous plotting and visualization in the same code. For visualization in Jupyter Notebooks, the team used Altair. The team found Altair to be very powerful, and it worked well with pyspark because it plots dataframes very intuitively. In addition to Altair, we wanted to be able to visualize the Realtor data spatially. Due to prior work experience, we chose to use Tableau to visualize this spatial data. However, in order to do this successfully, two additional datasets were required for representing the Realtor data by US County. Finally, D3 was also used to perform some of the visualization. The exact methodologies and findings for each dataset will be examined further below.

4.1 Vehicle Travel in and out of NYC

The Vehicle Travel dataset had a very fine temporal granularity; it was recorded hourly for 10 years. In order to make this more manageable to examine and visualize, the data was aggregated by week. In other words, the inbound vehicle counts were summed together for each week, and outbound vehicle counts were also summed together for each week. This was done using a spark job (*Bridges_analysis.py*) which took the cleaned dataset *New_Bridges_Tunnels.csv* as input. After more analysis was done on other datasets which had a minimum temporal granularity of month, we realized that it would be useful to be able to compare the Vehicle Travel dataset on the same axis/scale, and therefore aggregated monthly. This was done later alongside the visualization code in a Jupyter Notebook using Spark.

To visualize the data, we used Altair in a Jupyter Notebook. To get a better sense of what kind of yearly trends are typical, we plotted every year on the same plot where the x axis represented the weeks of the year. The resulting plot is included in figure 1. After getting a sense of what is typical, we plotted just 2019 vs 2020. These two plots are seen in figures 2 and 3. The data from 2019 is typical when compared with past history. However, the

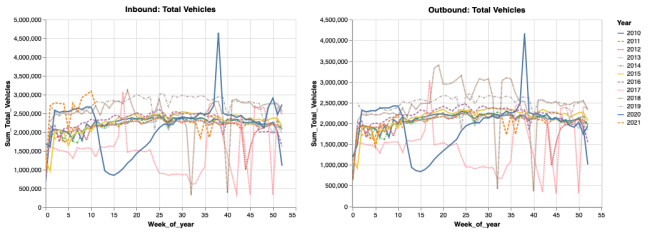

Figure 1: Weekly inbound and outbound traffic over NYC bridges and tunnels.

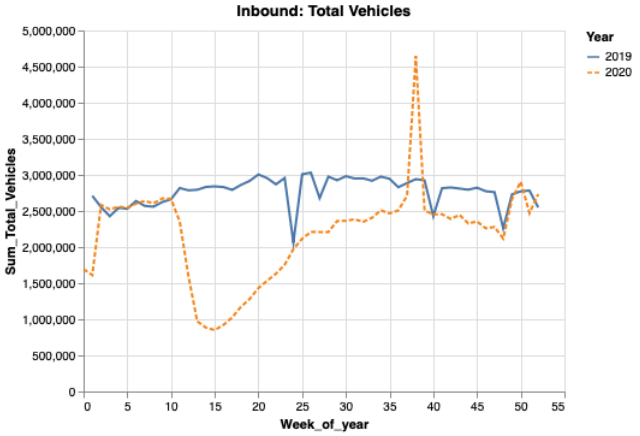

Figure 2: Weekly inbound traffic over NYC bridges and tunnels.

data for 2020 (the year of the pandemic) is clearly very different. There is a huge decrease in both inbound and outbound traffic starting around March, when the pandemic officially began and lock-downs/quarantines were put in place. It took many weeks, but the number of vehicles entering and leaving the city eventually returned to historical norms. Additionally, there was a huge spike in inbound/outbound vehicle traffic during September 2020. We hypothesized that this might be related to a mass move-in/move-out due to expiring leases, and this theory will be explored in further sections.

4.2 National Energy Usage

To analyze the National Energy Usage dataset we used d3 and observable notebook to create the visualization. The dataset had monthly energy consumption of each state over the years. The energy consumption for year 2019 and 2020 has been visualized to identify trends in consumption during the pandemic. We have plotted a diverging bar graph with states on the Y-axis and difference in energy consumption on X-axis. Using a diverging bar graph gives us a clear idea of which states saw an increase in energy consumption and which states saw a decrease in energy consumption. The length of the bar graph encodes the difference in energy consumption in Megawatt hours. The energy consumption for every state in the dataset comes from different energy sources. It has been aggregated using the rollup function in d3 for both years. Then we

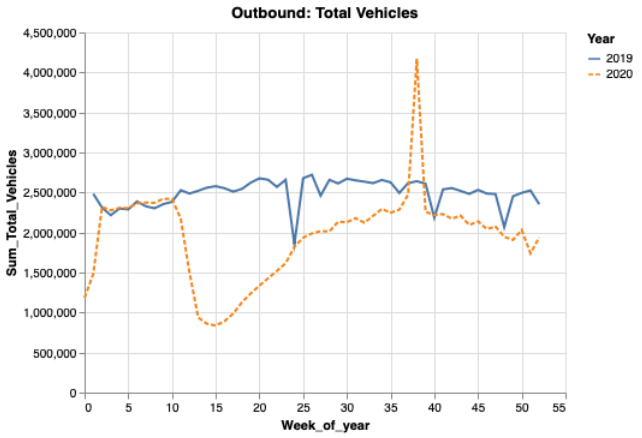

Figure 3: Weekly outbound traffic over NYC bridges and tunnels.

calculate the difference by subtracting the consumption in 2019 from consumption in 2020 for each state. Link to the observable notebook: https://observablehq.com/d/e5d1f5906bfbcde2

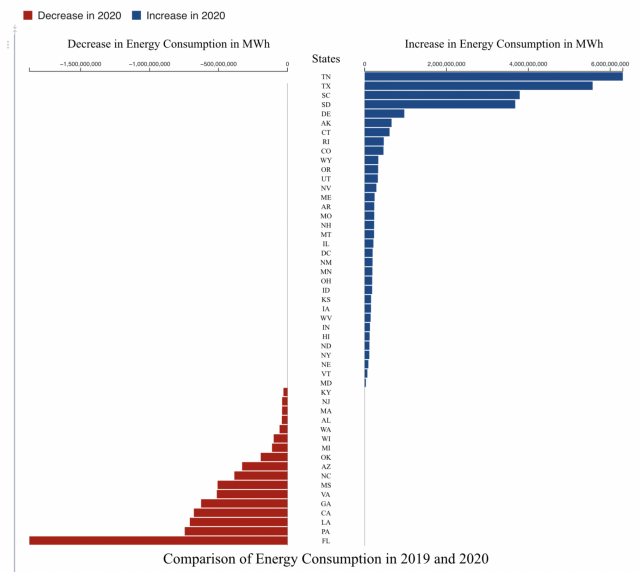

Figure 4: Comparison of Energy Consumption in 2019 and 2020.

4.3 Sidewalk Cafe Licenses and Applications

The strategy for analyzing the Sidewalk Cafe application data centered on evaluating if there was a significant increase in number of applications in 2020 during the pandemic, and then evaluate whether or not this had a significant impact on the amount of available space and amount of waste produced in the city. To examine these ideas, we created a Jupyter Notebook, and used PySpark to gather statistics. To make visualization and analysis easier, we decided to aggregate the data by month, and filter out all data from years other than 2019 and 2020. Specifically, the date used for aggregation and filtering was the date which the application was submitted. 2019 was used as a baseline to compare with 2020.

Unfortunately, after comparing the total number of applications submitted in 2019 vs 2020, we discovered that the data stopped after June 19, 2020. Additionally, the number of applications significantly decreased after the pandemic began in March (see figures 5 and 6). Therefore, there was not enough data to properly take an average of monthly data without a significant bias in the last few months of available data. We decided that this dataset was therefore not viable for much more inspection because of it's limited date range and small amount of data. In retrospect, this is not very surprising, because the first few months of the pandemic (March-June) there were still many restrictions, and most restaurants were still closed. However, for good measure we decided to try looking at a couple other statistics. For example, we calculated the average proposed square footage of sidewalk cafe applications for each year. There was only a −2.423% difference between years, so it was not significant. The full analysis that was done on this dataset can be found at Big-Data-Project-X/src/vis_src/Sidewalk_analysis.ipynb.

4.4 Real Estate Listings and 'Hotness' by Area

There were a few main metrics that we were interested in looking at in the Realtor data to see if there were any big shifts demographics during the pandemic. Specifically, those metrics include supply score, demand score, hotness score, listing price, various listing count metrics, and average days on market. Our initial hypothesis was that many people moved out of the densely populated NYC and relocated to the less populated suburbs to avoid the affects of the pandemic. This hypothesis was based on anecdotal evidence, experience, and logical reasoning. In order to investigate this further, we needed to determine what US Counties make up NYC, and which counties make up the suburbs of the NYC Metro area. In order to do this, we looked at maps of typical definitions of the NYC Metro Area. For example, see Figure 7 that was found on the internet.

intake_month	count
01	43
02	41
03	65
04	110
05	60
06	15
07	16
08	39
09	80
10	21
11	31
12	62

Figure 5: Number of sidwewalk cafe applications submitted during 2019.

intake_month	count
01	33
02	58
03	13
04	3
05	1
06	6

Figure 6: Number of sidwewalk cafe applications submitted during 2020.

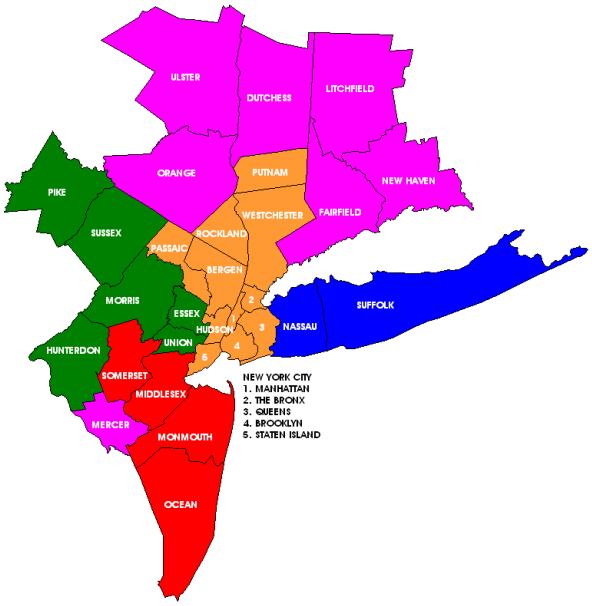

Figure 7: Example from the internet of counties defined to be part of the NYC Metro Area.

Next, we needed to determine which zipcodes were part of each of those counties, because the Realtor data is recorded by zipcode. To accomplish this, we went to https://www.unitedstateszipcodes.org and searched for each of the respective counties, and copied the zip codes that were part of the 5 boroughs of NYC into a file named *Zips_In_NYC.csv*, and the zip codes that were part of any of the other counties into a file called *Zips_In_Metro_Area.csv*. This was subject to being slightly messy data, so we create d a spark job that would look for duplicates in the list (*Zip_Code_Intersect.py*). The program found 3 duplicates, and after searching google, we found them all to be part of Nassau County, NY, so we wrote another spark job to remove them from the NYC zipcodes data (*Clean_Zip_Codes.py*).

The next step was to get an idea of the metrics mentioned earlier, by aggregating them by "Area," (either NYC or NYC Metro Area) and "*month_date_yyyymm*". Once again, we used pyspark in a Jupyter Notebook to do this. Then we plotted the aggregated data by year to get a sense of yearly trends. These plots can be seen in figures 8-15.

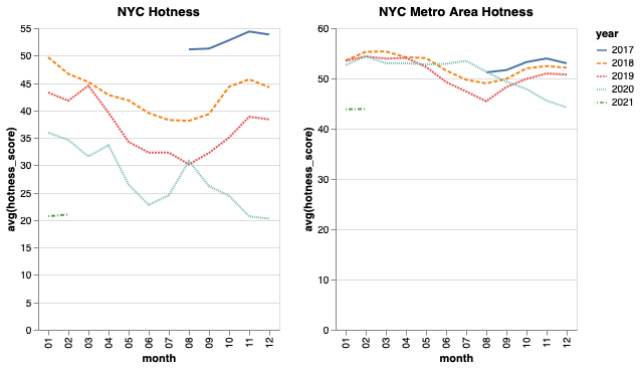

Figure 8: Comparison of Hotness Score of the NYC and NYC Metro Area plotted for the past few years.

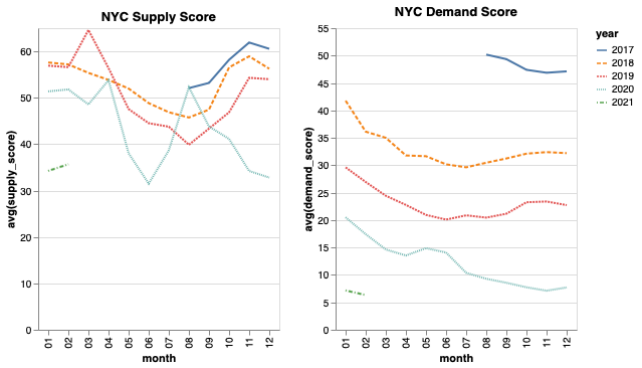

Figure 9: Comparison of NYC Supply and Demand Scores plotted for the past few years.

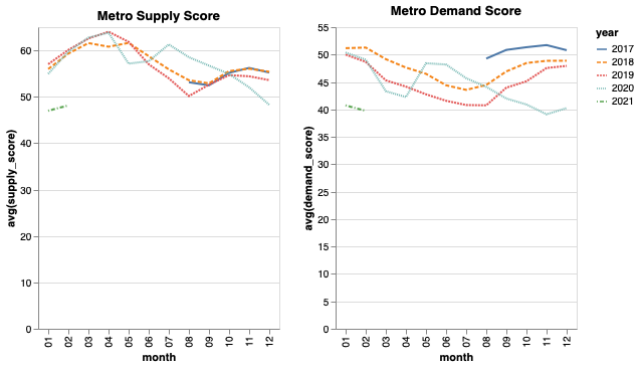

Figure 10: Comparison of NYC Metro Area Supply and Demand Scores plotted for the past few years.

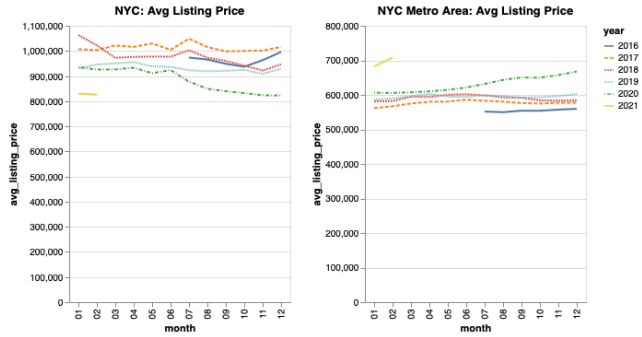

Figure 11: Comparison of average listing price of the NYC and NYC Metro Area plotted for the past few years.

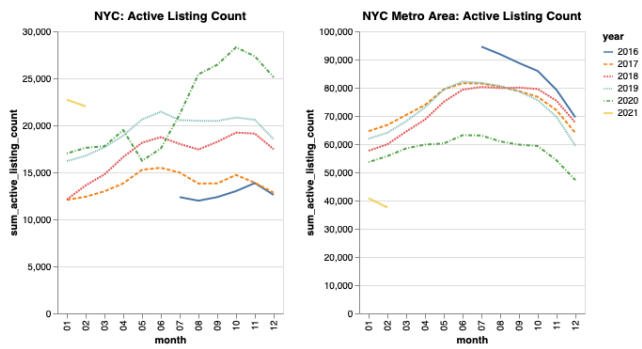

Figure 12: Comparison of the number of active listings in the NYC and NYC Metro Area plotted for the past few years.

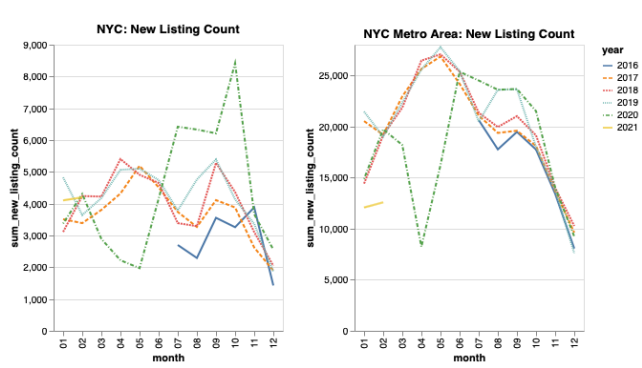

Figure 13: Comparison of the number of new listings in the NYC and NYC Metro Area plotted for the past few years.

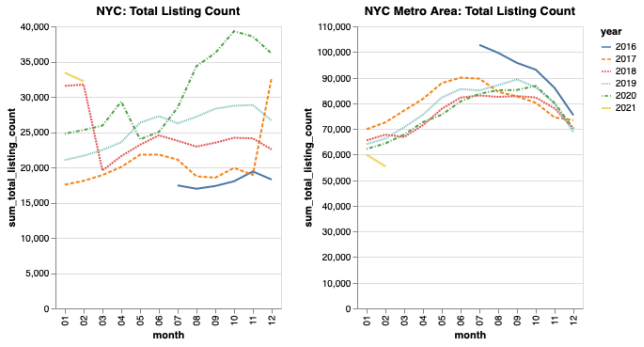

Figure 14: Comparison of the number of total listings in the NYC and NYC Metro Area plotted for the past few years.

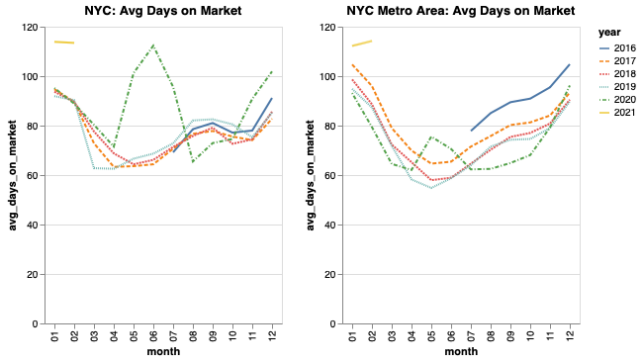

Figure 15: Comparison of the average number of days on the market in the NYC and NYC Metro Area plotted for the past few years.

There were a few useful insights gathered from these plots. The first, which is demonstrated in Figure 8, is that overall, the Hotness Score in both NYC and the NYC Metro Area decreased significantly in 2020. Typically there is a dip in hotness in the middle of the year, but it increases by the end of the year. However, in 2020 the score did not recover by the end of the year for either area. Interestingly, there was a significant spike in hotness in NYC around September, and this corresponded to a large increase in supply at the same time (see Figure 9). This information, coupled with the fact that there was not a corresponding increase in demand, leads us to the conclusion that there was a major 'move-out' event when many people's leases expired in September. This is also supported by the fact that the number of new, active, and total listings in NYC increased significantly during the months after the pandemic (see Figures 12, 13, 14). This is further supported by Figure 11, which demonstrates that the average listing price of properties in NYC fell considerably during 2020.

In our original hypothesis, we also were inclined to believe that many of the people living in the city moved out to the surrounding suburbs of the Metro Area. This is supported by Figure 10, which shows that there was a large increase in demand in the Metro Area shortly after the pandemic began, during a time of the year when there is typically a dip in demand. Naturally, it follows that the average listing price significantly increased during 2020 when compared with previous years due to this increase in demand as seen in Figure 11. Furthermore, the average time on the market for metro area properties was the lowest it has been for the the past 3 years except for a small increase shortly after the beginning of the pandemic when there were still strict lock-downs (see Figure 14).

Naturally, the team wanted to be able to visualize some of these insights spatially on a map of the Tristate Area. One of the team members had prior experience completing a very similar task in Tableau, so the spatial visualization was done in Tableau. The first step in this process, was to get the shape files for each of the US Counties that are part of the NYC Metro Area. The US Census Bureau provides this data, but on a national scale at https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html. Specifically, we used the *cb_2018_us_county_500k.zip* shape file. Next, since the Realtor data was organized by zipcode, and not county, we needed a dataset that could translate zipcodes into geographical coordinates to plot inside of the county shapes in Tableau. The data at https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/ was used for this purpose. Next, in order to represent the change in Realtor data between 2020 and previous years on a map, we decided to write a spark notebook at Big-Data-Project-X/src/analysis_src/Hotness_percent_change.ipynb to generate the percent change between 2019 and 2020 for three of the more interesting metrics (hotness, supply, and demand).

Finally, in Tableau, we used an inner join where the geometry field of the shape file intersects $MAKEPOINT([Latitude], [Longitude])$ in the zipcode dataset. Then we were able to create a relationship between the two joined datasets, and the Realtor data by matching on zipcodes. With the data source fully constructed, we then created three maps in Tableau to visualize the average percent change by county in Hotness Score, the Supply Score, and the Demand Score between 2019 and 2020. The resulting dashboard can be viewed at https://public.tableau.com/profile/max.christ4104#!/vizhome/shared/FTWSHZR2P and in Figure 16.

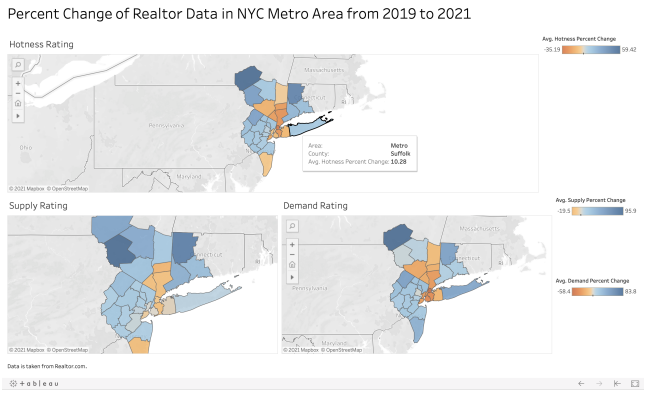

Figure 16: A screenshot of the Realtor data spatially represented in Tableau Public.

4.5 EPA Air Quality System (AQS)

The EPA AQS dataset, being rather verbose, contained multiple air quality parameters over a variety of sample durations, as well as many more attributes than we were interested in. Some of these attributes–such as the site information, latitude, longitude, etc.–were not as useful to us due to the fact that we only retrieved data for the New York City site (160 Convent Ave, New York, NY 35620). Other attributes such as the methods of collection were also not very useful to us due to the fact that our goal was to seek insights in how this data correlated with other datasets. In addition to this, the daily nature of entries was much too fine granularity to be effectively compared against the other datasets. Thus we needed to not only extract the attributes we were most interested in, but also aggregate the rows down to a more manageable monthly frequency.

Analysis and preparation of the data for visualization was performed using PySpark and was run on the NYU Peel High Performance Computing cluster. Since we were using Altair-viz to create the visualizations of the data, we generated pared down datasets for each of the AQI pollutant parameters:

- Carbon Monoxide
- Sulfur Dioxide
- Nitrogen Dioxide (NO2)
- Ozone
- PM10 (Particulate Matter under 10 micrometers)
- PM2.5 (Particulate Matter under 2.5 micrometers)

The main attribute that we were interested for each parameter was its Air Quality Index (AQI) score. Per the AQS Data Dictionary[1], AQI is a unit less measure of the amount of pollutant that can be used to relate the pollutant to the healthy levels and indicate possible health concerns with elevated levels. To isolate this attribute for each parameter, we split the data for each parameter by year, and then aggregated each by the month. We then created new attributes for the parameter name, month number, and year in each aggregated dataset to facilitate visualization. One dataset was created for each parameter, for each year in the data, and with two additional datasets for each year with all of the parameters.

Unfortunately the parameters Sulfur Dioxide and Nitrogen Dioxide did not have AQI values present, and thus were left out of analysis.

Lastly, since there were two different parameter names that corresponded to PM2.5 (PM2.5 - Local Conditions and Acceptable PM2.5 AQI & Speciation Mass) we created a final dataset that contained both parameters.

As can been seen in figures 17 and 21 for CO and PM10, unfortunately there wasn't complete data for all parameters for both 2019 and 2020. While not much can be said about Carbon monoxide since there is no 2020 data, PM10 presents a slightly more interesting story. Despite the fact that we only had data for the first three months of 2020, we can see almost immediately there is quite a large difference between the 2019 and 2020 levels in the beginning of the year. While we cannot glean much in the way of insights from such partial data, it would certainly merit additional exploration if and when additional data is made available.

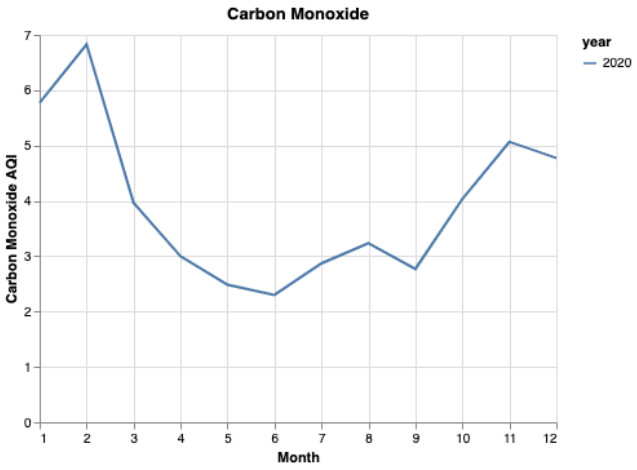

Figure 17: Carbon Monoxide over 2020

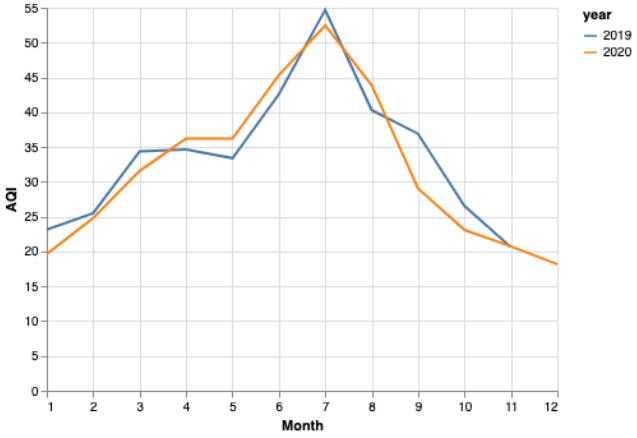

Figure 18: Ozone over 2019 and 2020

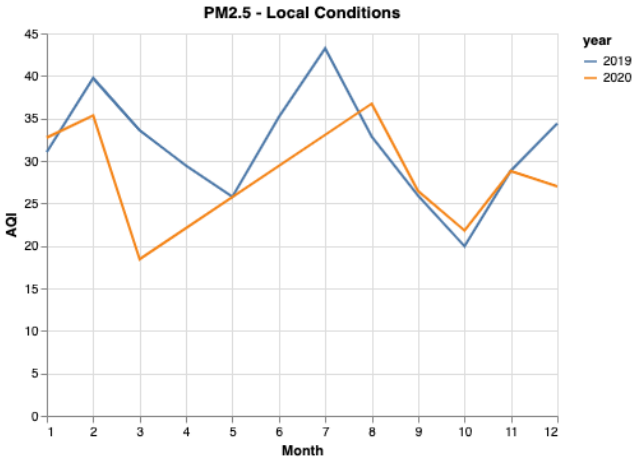

Figure 19: PM25 - Local Conditions over 2019 and 2020

[1] https://aqs.epa.gov/aqsweb/documents/AQS_Data_Dictionary.html

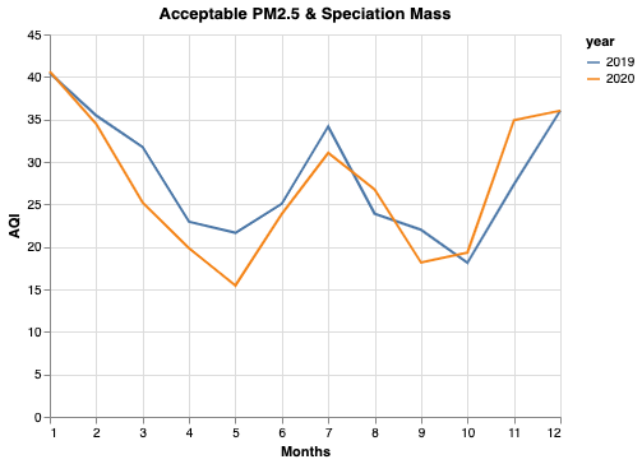

Figure 20: Acceptable PM2.5 & Speciation Mass over 2019 and 2020

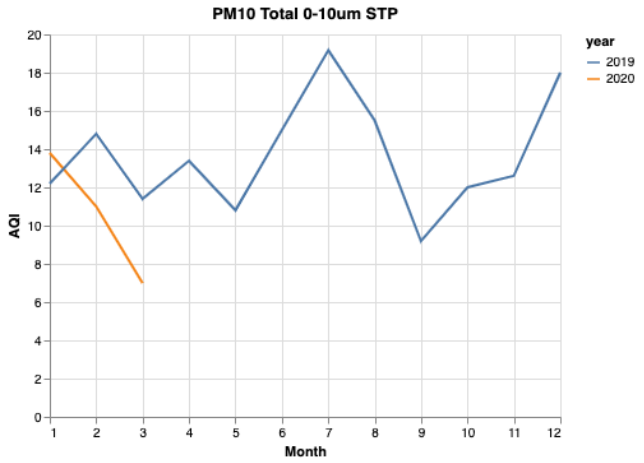

Figure 21: PM10 Total 0-10um STP

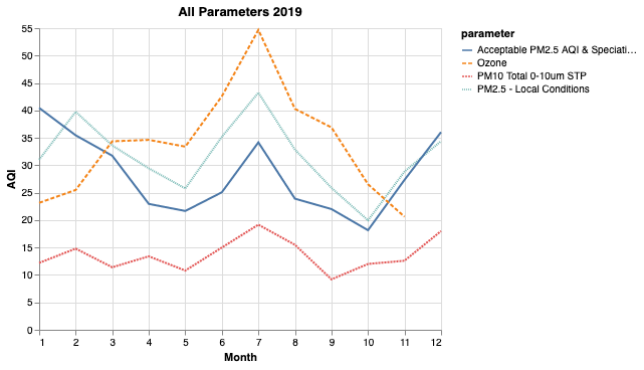

Figure 22: All parameters available in 2019

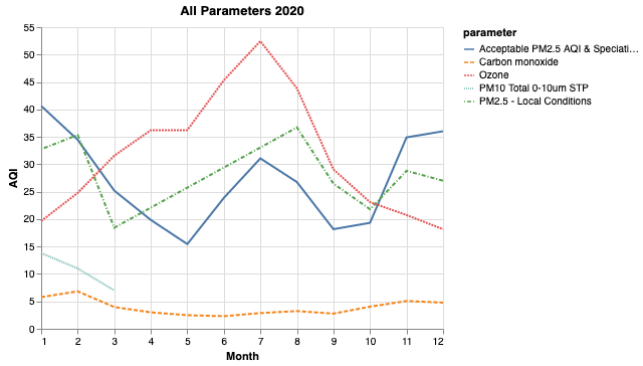

Figure 23: All parameters available in 2020

Ozone on the other hand had data spanning both 2019 and 2020, but unfortunately as can be seen in figure 18, it appears that there is not a clear difference between ozone levels in 2019 and 2020. This in itself is somewhat surprising, as we would have assumed that there would have been at least a consistent, moderate reduction in ground-level ozone pollution during 2020 due to decreased car use, but it seems that this is not the case. It is possible that other factors such as increased generation by other sources such as power plants could have made up for the reduction in car traffic, but we are unable to draw any further conclusions from the data available to us.

Out of all of the parameters, it is the PM2.5 data that presents the most interesting and complete story. Before we dive in, a note on nomenclature. There are two plots for PM2.5 data; PM2.5 - Local Conditions and Acceptable PM2.5 & Speciation Mass. Both of these parameters represent valid measures of PM2.5 data, with the main difference being the PM2.5 - Local Conditions has been evaluated and considered eligible for use in making NAAQS (Nation Ambient Air Quality Standards) decisions, while Acceptable PM2.5 & Speciation Mass still reasonably matches quality standards, but should not be used for NAAQS decisions[2]. For our purposes we considered both to still be valid measures of PM2.5 data, but we kept them as separate plots and did not average them together. As can be seen in figures 19 and 20 there is a clear, if slight, reduction in PM2.5 concentrations over the course of the first seven months of 2020 compared to 2019. This lined up with our assumption that due to the large numbers of people working from home and not commuting nearly as often would cause a decrease in air pollutants. What surprised us though was that the data showed a slight increase in PM2.5 levels in the later part of 2020, compared to 2019 levels. To attempt to glean insight into why this was, we compared the PM2.5 data for 2020 with our other datasets, the results of which we will discuss further in the Conclusions and Insights section.

The code used to perform this analysis and aggregation can be found in the Python file *epa_air_analysis.py*, and can be located in the project directory {*PROJECT_DIR*}/*src*/*analysis_src*/.

[2]https://www.epa.gov/aqs/aqs-memos-technical-note-reporting-pm25-continuous-monitoring-and-speciation-data-air-quality

All plots generated from these datasets were created using Altair-Viz and Pandas, and can be found in the notebook *EPA_Air_vis_altair.ipynb* and can be located in the project directory {*PROJECT_DIR*}/*src*/*vis_src*/.

5 CONCLUSIONS AND INSIGHTS

The final phase of this project was to compare the datasets we examined, and look for interesting correlations and relationships between them. The goal of this task was to learn novel information, and attempt to answer the original questions which this project was created to answer. Since the EPA WQP water quality and OpenNYC sidewalk cafe license application datasets were deemed unhelpful due to unforeseen problems in the data outlined in the above sections, they were not compared with any of the other datasets. The following sections will explore the original research questions and the methods we used to evaluate them.

5.1 Will the number of people who are no longer commuting regularly offset the sum of the other negative impacts of the pandemic?

Based on the data this project looked at, we were unable to come up with a conclusive answer to this question. In fact, we will explore in a section below how vehicle traffic in and out of the city returned to fairly normal levels midway through the pandemic, so fewer commuters might not have had the effect as previously thought.

5.2 Has there been a significant improvement in air quality, and do we expect any changes to be long lasting?

Based on the data available to us, it appears that there were mild improvements in some of the parameters that are used to measure air pollution, such as in the case of PM2.5 and partially in PM10 (figures 21, 20, 19), but they were not consistent throughout the year, or were affected by external factors. Other pollutants, such as ground-level ozone (figure 18) did not show any marked improvement. While it is heartening to see at least some environmental improvement, the data appears to show that any improvement may be temporary.

Figure 25 appears to indicate that while pollutants such as PM2.5 may decrease during periods of reduced vehicle traffic, they will only increase again once traffic spikes or resumes. Based on this evidence, we must conclude that it is unlikely that any positive environmental changes will persist for long without sustained changes to how we work and commute.

5.3 Has there been any significant reduction in energy consumption due to staying indoors and remote work practice?

Referring to the visualization(Figure 4) of the energy dataset, it appears that working from home has caused an increase in energy consumption in majority of the states. States like Tennessee, Texas and South Carolina had the highest increase in energy consumption. However, a decrease in energy consumption can also be observed in few states like New Jersey, Kentucky with Florida reducing the consumption by most.

5.4 Has the shift to remote work had any lasting effects on the volume of traffic going in and out of the city?

Figures 1, 2, and 3 show the vehicle traffic in and out of the city in 2020 compared with previous years. It is clear that there was a significant dip in traffic entering and exiting the city during the first few months of quarantine. However, by midway through the year, other than a spike in September, the number of vehicles returned to roughly pre-pandemic levels and remained that way. Therefore, we do not believe that the shift to remote work led to any significant difference in the number of vehicles entering and exiting the city. A potential theory might be that most commuters use public transport such as the LIRR, Metro North, etc, rather than personal vehicles, meaning that commuter traffic does not contribute much to the overall number of vehicles going in and out of the city. On a separate note, interestingly, overall more vehicles entered the city than left. However, this could be due to the absence of one of the bridges in the datasource.

5.5 Have there been any lasting geographic population shifts that may ease the environmental stress of cities?

As described in detail in section 4.4, there certainly appears to have been a major demographic shift in NYC during 2020. To briefly recap, there seems to have been a major 'move-out' event during September 2020. Evidence of this event can be found in the large spike in supply in NYC, a steadily declining demand, a large increase in listings, and steep decline in prices. The large spike in supply in August happened to correspond to a large spike in the number of vehicles entering and exiting the city in September, when many leases expire. This can be seen in Figure 24. Note, since the Realtor data was generated monthly, and the vehicle traffic was aggregated weekly, the vehicle data needed to be re-aggregated to be monthly. This way, the two datasets could be compared on equivalent axes.

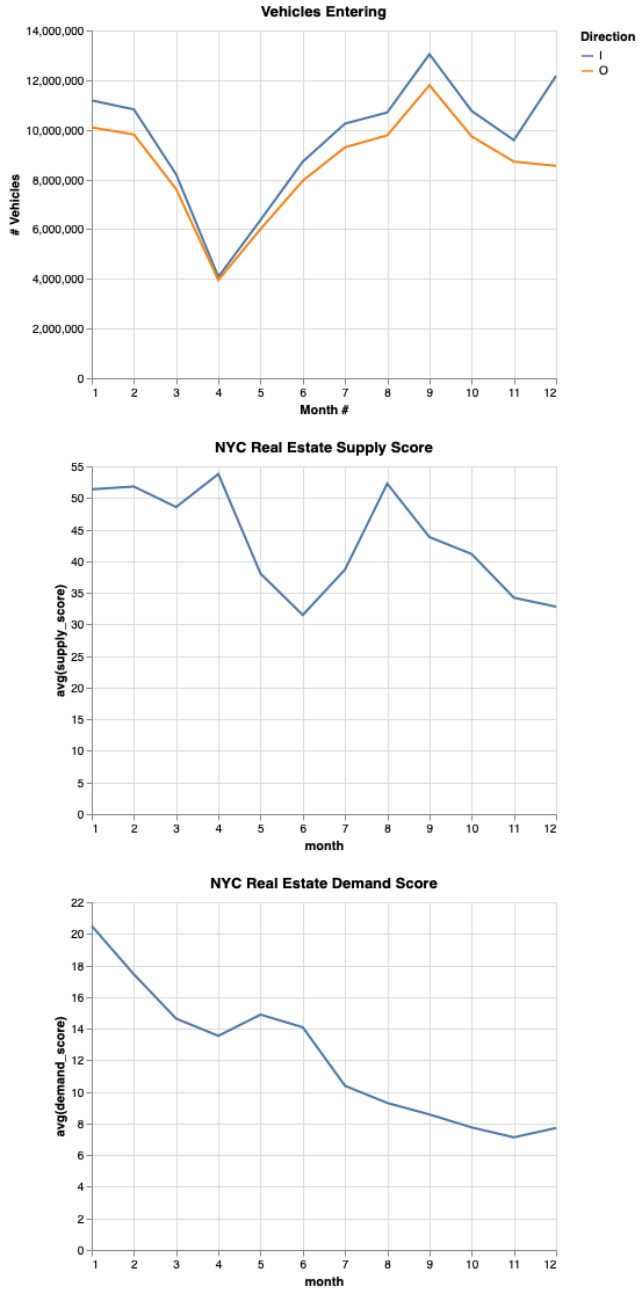

Figure 24: Comparison between Realtor Supply and Demand of NYC, and vehicle travel in and out of the city.

We believe this large spike in vehicle traffic can be attributed to many people moving out of their apartments when their leases expired. Additionally, it appears that many of the people moved to the surrounding metro area, because the supply in the NYC Metro Area steadily declined after September, when it typically increases. Also, the demand surged during the middle of the pandemic, the prices in the Metro Area significantly increased, and the number of active listings was the lowest it has been in years. Therefore, based

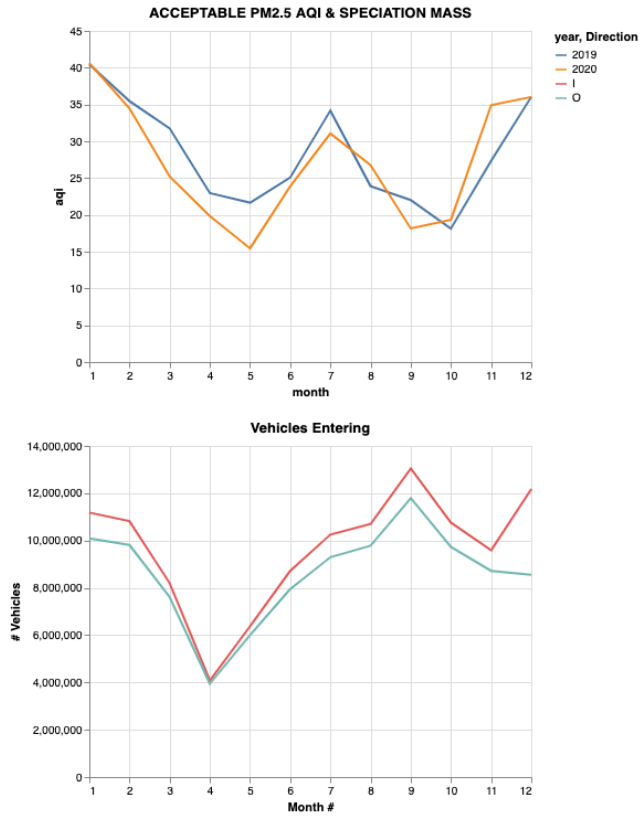

Figure 25: Traffic compared with PM2.5 Levels

on the available data, we surmise that there will be a lasting impact on the population of NYC.

The next plausible step was to investigate whether this had any observable impacts on the environment. Therefore, we compared the mass 'move-out' event seen in the vehicle traffic data, with the air quality data. Figures 25-27 were created for this purpose. Some interesting observations that can be drawn from 25 include that during the time that vehicle travel was greatly reduced during the beginning of the pandemic, the Acceptable PM2.5 AQI and Speciation Mass levels were significantly lower than they were in the previous year. This tends to suggest that the lower number of vehicles entering and exiting the city had a positive impact on air quality by reducing the Air Quality Index (AQI). More in-depth historical analysis of this air quality metric is needed to be completely sure. Additionally, the spike in vehicle traffic in September corresponded directly with what seems like a premature rise in the AQI based on the trend of the previous year. This suggests that the large 'move out' scenario discussed earlier prematurely degraded air quality for that time of the year. Lastly, figures 26 and 27 show that the sharp spikes in NYC supply and hotness during the month of August correspond to shifted PM2.5 Local peaks. This might be coincidental, since the other peaks do not necessarily seem to match. More data is needed to investigate if this is a meaningful correlation or not. However, it is plausible that this increased housing supply spike was related to the uptick

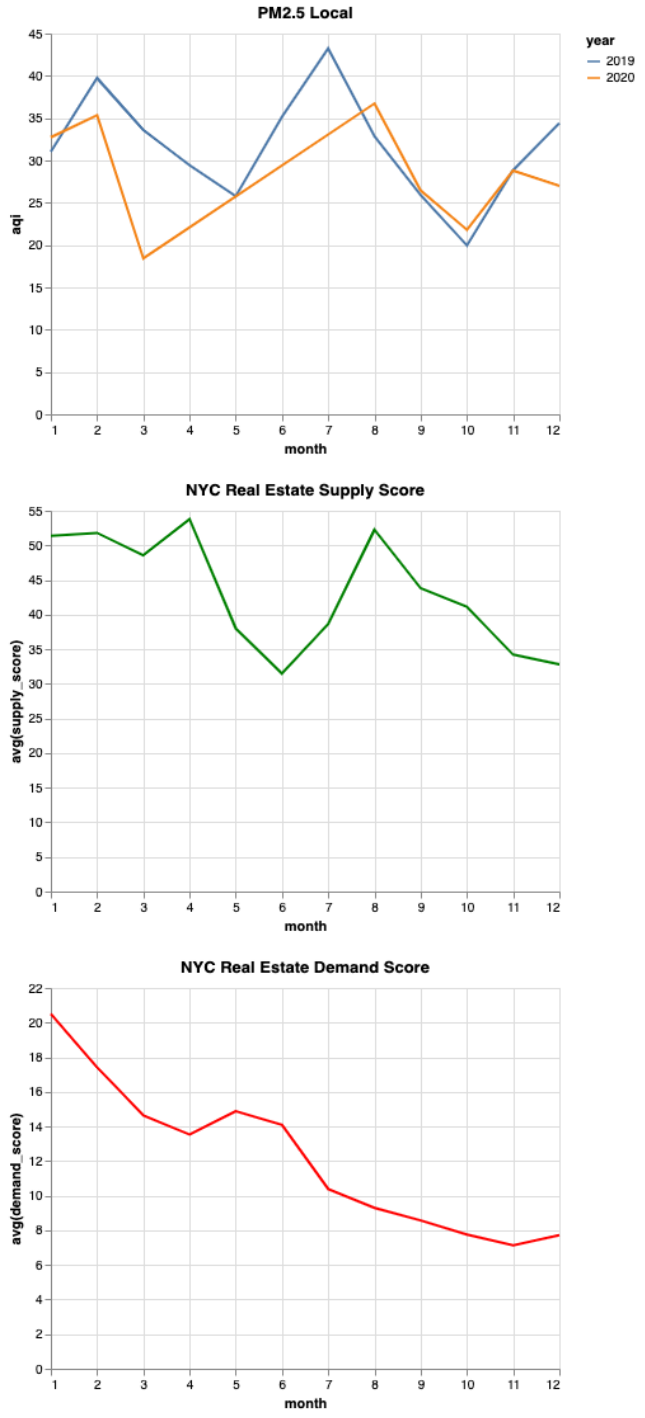

Figure 26: NYC Supply/Demand compared with PM2.5 Levels

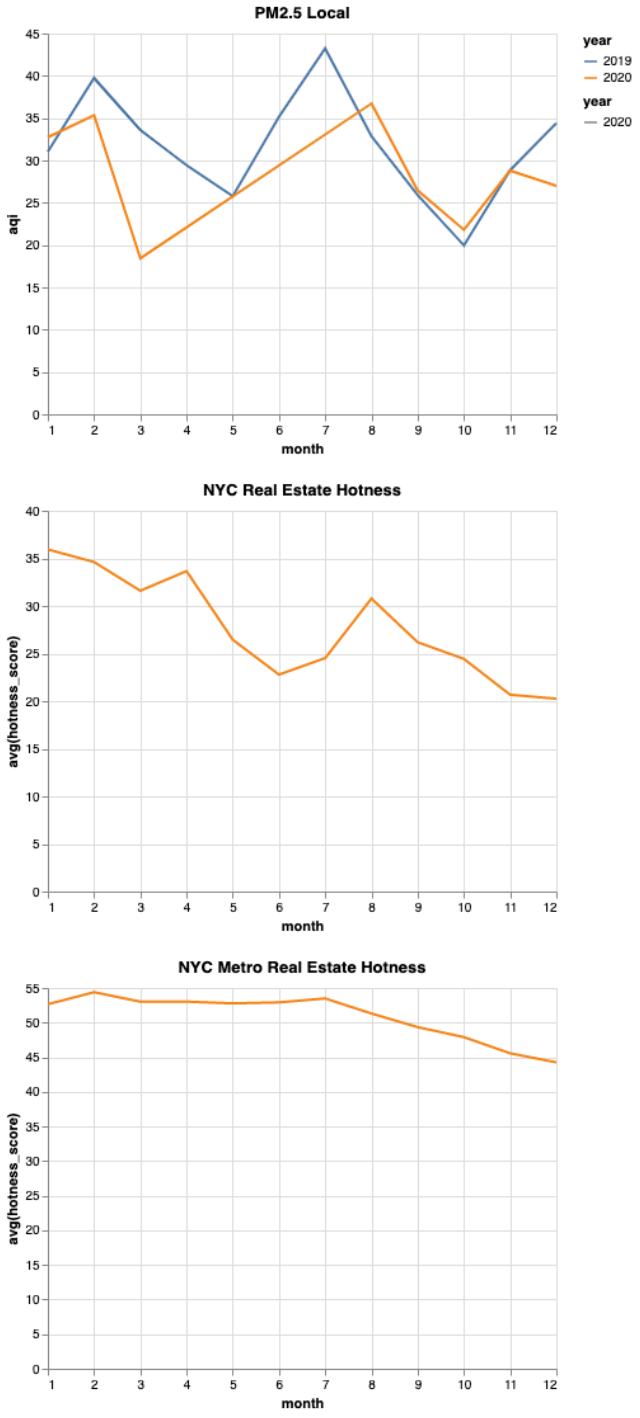

Figure 27: NYC and Metro "Hotness" compared with PM2.5 Levels

in traffic seen earlier, and therefore affected the air quality. Like previously mentioned, more data would be needed to evaluate this hypotheses.

5.6 Final Thoughts

This project uncovered some interesting data, and unforeseen relationships, especially with regards to Real Estate data, vehicle traffic, and air quality. However, we did not uncover as many environmental insights as we would have liked. Part of this is due to the difficulty in finding sufficiently recent and available environmental datasets. Another reason is how significant time and analysis was required before discovering that a given dataset was not viable or helpful. Therefore, in the future as more data is released from the 2020 period, more analysis can be done to further investigate this space. Furthermore, more data should be gathered from other domains to verify the conclusions and theories laid out in this report. However, this report has shown that there were certainly interesting events and trends related to the data examined during the period containing the COVID-19 pandemic.