# 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 for examining 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 be having 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 starting point, and then custom scripts were created in

*All authors contributed equally to this research.

Google Collab 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 2019-202. These pollutants include:

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

## 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 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 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 data output 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.