The Impacts of the COVID-19 on Public Safety in NYC

Shuhong Cai, Huixuan Wang, Shizhan Li
Courant Institute of Mathematical Sciences
New York University
{sc8540,hw2544,sl8774}@nyu.edu

Abstract

In early 2020, coronavirus broke out and exponentially spread across the world. Under such a situation, individuals are more likely to engage in accidents due to various types of stress, poor health conditions, and financial crisis. Our study aims to explore the impacts of COVID-19 on New York City's public safety among different regions over time, including crime rates and traffic accidents. We want to answer several questions: Does COVID-19 reduce the number of accidents due to the disappearance of potential victims in public? Does COVID-19 spur more crimes due to the unemployment and discrimination? Do different areas present various tendency? The analysis result shows that there is a strong correlations between motor collisions and NYPD arrests. The outbreak of COVID-19 is accompanied by the increase of arrests. Motor collisions have negatively correlations with COVID-19 cases. The patterns we found can help government better administer the public transit and social safety.

1. Introduction

In 2020, New York City experienced an outbreak of COVID-19, a highly contagious and potentially deadly virus. A state of emergency was declared on March 12 and by March 22 the city enacted "shelter-in-place"[1]. Through the lockdown phase of the crisis, stay-at-home mandates might increase both the mental and financial burden and negative discrimination among citizens, leading to more crimes. At the same time, COVID-19 crisis may also have profound impacts on many public safety organizations that typically responds to those accidents. In this project, we explored the relationship between the COVID-19 and public safety with Coronavirus Data, NYPD Arrest Data and Motor Vehicle Collision Data. Our research adopted big data framework like HDFS, MapReduce and Hive and provided some meaningful insights based on analysis results.

2. Motivation

The scope, size, and severity of this global public health disaster is unlike any seen in the last 100 years. New York City is the epicenter of the outbreak in the United States. Thus the research focuses on NYC is typical. The users of our research are public Safety Government and citizens. Policy makers can get lessons and respond in time in the large-scale public safety events. Citizens can adjust their daily activities to better protect themselves. For both local authorities and individuals, the capacity to recognize the impacts and take actions is necessary. They can respond to such emergencies timely in the future, reducing the costs to preserve order.

3. Goodness

Michelle and Anna investigated the relationship between crime type and COVID-19 quarantine in 2021[2]. They found that all crimes types except for murder and burglary, exhibited a statistically significant difference during COVID-19 conditions compared to the same time the previous year. Marshall found that social distancing on highways undermines compliance with social norms and poses potential long-term increases in non-compliance and dangerous driving[3]. Other scholars also explored the relationship between COVID-19 and hate crimes, elder mistreatment and other public related issues. They all proved that COVID-19 has effects on public safety.

Crimes and traffic accidents are two major sources that threaten public safety. Understanding the relationship between COVID-19 and public safety events can help provide lessons to better respond to such emergencies in the future and reduce governance costs. According to these, we obtained crimes and traffic accidents data from NYC Open Data; COVID-19 data from NYC Department of Health and Mental Hygiene (DOHMH). After using MapReduce for data cleaning and Ingestion, Hive for data joining and processing, and Tableau for data visualization, we believe that our research results and conclusions are correct and can be trusted.

4. Data Sources

The analyses conducted are based on the data supplied by three different NYC government agencies, i.e. *New York City Police Department*, *New York City Department of Transportation*, and *New York City Department of Health and Mental Hygiene*.

4.1. NYPD Arrest Data

As a key indicator of the public safety situation in NYC communities, the arrest data from New York City Police Department (NYPD) is used. Our data source is a combination of two tables with the same schema, which are NYPD Arrest Data (Historic)[4] , and NYPD Arrest Data (Year to Date)[5].

The Historic dataset contains a list of every arrests in NYC going back to 2006 through the end of the previous calendar year (2021). This dataset, in `CSV` format, has over 5.3 million pieces of records and a size of 1.1 GB. The Year to Date dataset contains the record from the beginning of this year (2022) to the latest update as we downloaded the data. As it is manually extracted reviewed every quarter. The latest record was created on September 30, 2022. With around 14 thousand records stored in `CSV` format, its size is around 25 MB.

In both datasets, each record represents an arrest and includes information about the type of crime, the location, time of enforcement and suspect demographics.

4.2. NYC Coronavirus Data[6]

This dataset contains information about Coronavirus in New York City. It includes three major measurements: average rate of cases per 100,000 people, rate of molecular testing per 100,000 people, the percentage of people tested with a molecular test who tested positive per 100,000 people. The data collection spans from August 2020 to October 2022 and updates every week. These attributes are stratified by week and three different geographies: citywide, borough, and modified ZIP Code Tabulation Area (MODZCTA). Note that the Health Department uses ZCTAs which solidify ZIP codes into units of area. The modified ZCTA (MODZCTA) geography combines census blocks with smaller populations to allow more stable estimates of population size for rate calculation[7]. The MODZCTA reflects people's zip code of residence, not testing or hospitalization at the time of reporting. With MODZCTA, we can do more explicit spatial analysis. The data is in CSV format. There are 366 pieces of records in total and the size is about 300KB.

4.3. Motor Vehicle Collision Data [8]

This dataset contains information from all police-reported motor vehicle collisions in NYC where someone is injured or killed, or where there is at least $1000 worth of damage. We decided to use it since travel restriction policies can fundamentally change urban travel patterns. Each row represents a crash event, including crash date and time, location, zip code, number of persons injured, etc. This dataset provides us with a brief overview of the level of development, disruption, and public safety in different regions over time. The data is in CSV format. There are 1.95 million pieces of records in total and the size is about 396MB.

5. Approach

In this section, we first introduced the experimental environment, then illustrated the data processing pipeline that we followed. After that, we explained the approaches we took in each dataset and how we combined different datasets for further analyses.

5.1. Experiment Setup

In our project, both MapReduce programs written in Java, and *Apache Hive*, which are both backed by the *Hadoop MapReduce* engine on the *NYU Dataproc* cluster, are used for big data processing. And *Tableau* is used for visualization and some of the analytics job.

Apache Hive is a data warehousing framework, which converts SQL-like queries into one or more MapReduce jobs. And Tableau is a popular interactive data visualization software, which can connect to Hive using the ODBC Driver from *Cloudera*. They are both helpful for the analyses and visualization of big datasets.

5.2. Analytics Pipeline

As shown in Figure 1, data ingestion was done in each category of dataset. This step includes data cleaning and profiling, where we dropped unused columns, detect and remove badly formatted records, and count the number of records based on dates, locations, or other important fields.

After processing the datasets, they are used as sources of external Hive tables. The tables are joined together according to the date or location information. And further analysis and visualization were completed based on the joined Hive table.

5.3. NYPD Arrest Data

To carry out the relation analysis between the arrest data and other factors, we first need to extract the necessary fields. We are interested in the crime location, occurrence date of the reported event, and the offense level whose value would be *felony* (F), *misdemeanor* (M), or *violation* (V). During the analysis, we group the crime records to find out the count of the arrests made in given location each week, which is implemented as a MapReduce program.

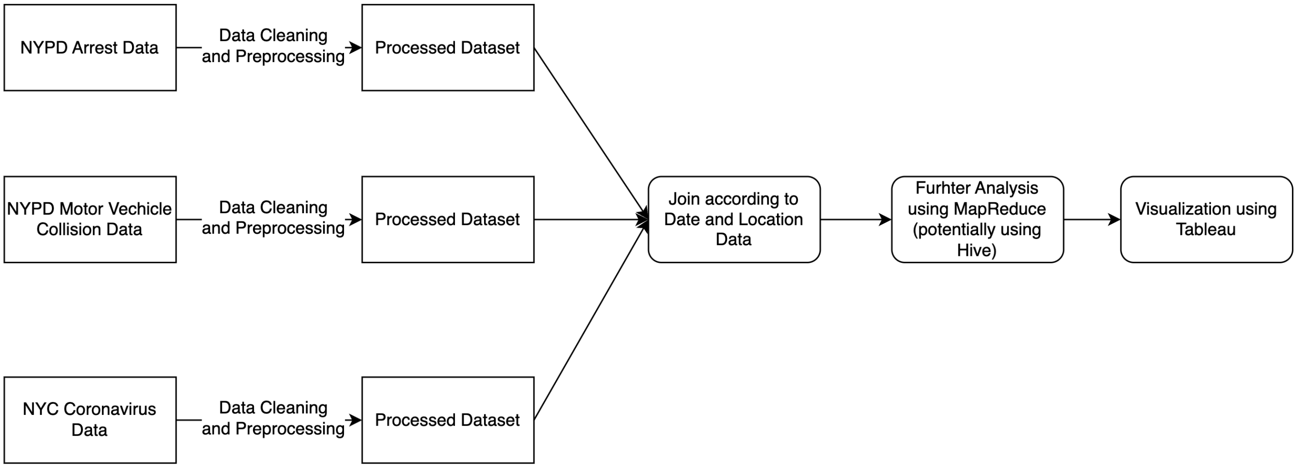

NYPD Arrest Data
Data Cleaning and Preprocessing
Processed Dataset
NYPD Motor Vechicle Collision Data
Data Cleaning and Preprocessing
Processed Dataset
NYC Coronavirus Data
Data Cleaning and Preprocessing
Processed Dataset
Join according to Date and Location Data
Furhter Analysis using MapReduce (potentially using Hive)
Visualization using Tableau

Figure 1: Design Diagram

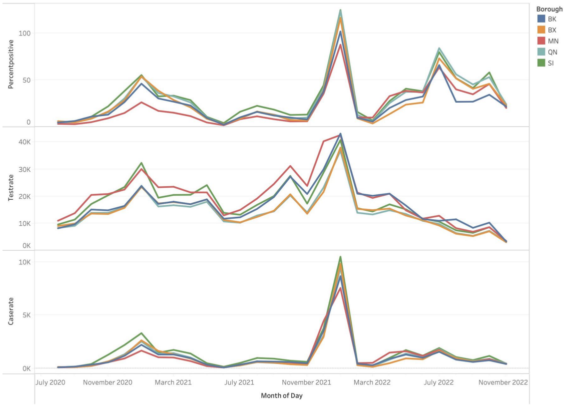


Figure 2: Trends of COVID-19 in boroughs

5.4. NYC Coronavirus Data

In original data, each row presents the aggregated sum by full-weeks starting each Sunday and ending on Saturday. The key, week_ending, is the date of the ending Saturday. The columns are geographic information, including citywide, five boroughs and zip codes. We used MapReduce to group the data by both date and geographic information. Then we created the COVID table in hive and joined it with other tables. For the three measurements: `testrate`, `caserate`, `percentpositive`, we plotted their changing curves in Figure 2 and found that they shows the similar tendency in 5 boroughs. They all reflect the consistent public health situation and exclude the influence of covid test sample distinction between different areas.

5.5. Motor Vehicle Collision Data

Each row represents a crash event, including crash date and time, location, zip code, number of persons injured, etc. We use two MapReduce programs to process this dataset. To get data among different zip-code areas, we set the key of Map to be the date and zip-code; to get data among latitude and longitude, we set the key of Map to be the date, latitude, and longitude. What we get from MapReduce is how many accidents happen and how many people are injured or dead within one particular day and location.

6. Code Challenges

Some coding challenges were faced when parsing the data during data cleaning and profiling.

6.1. Handling Date Information

Although the date information is recorded in the same format (`MM/dd/yyyy` in `String`), the date given in the Coronavirus dataset stands for the week ending at the date instead of the exact date. There are two issues with it. Firstly, to join our indicator (arrests and collisions) Hive table with the Coronavirus data, we need to sum up the daily statistics together by week, and represent it by the week ending date. Additionally, we want to make sure that the Hive, and Tableau can process the date information properly, instead of treating it like a `String`, so that we can easily adjust the range and granularity of the date for our analyses.

To resolve this, in the data ingestion stage, the `Calendar` class in Java is used to calculate the week ending date, and the `SimpleDateFormat` class is used to format the date in Hive default format (i.e. `yyyy-MM-dd`).

6.2. Handling Location Information

As shown in Table 1, the location data in each datasets is provided in different formats. Two approaches are used to group the location data for further analysis.

6.2.1 Normalizing location data to MODZCTA

One of the approaches we chose is to convert the ZIP Code and precinct data to MODZCTA. For the ZIP Code data, the NYC DOHMH provides the mapping data from regular ZIP Code (ZCTA) to the modified one (MODZCTA) [9]. It is a many-to-one relationship, where each ZIP Code has only one MODZCTA, which makes the mapping possible.

For the precinct, there is no simple mapping to either ZIP Code or MODZCTA. The contact information[10] for each precinct is used as their representing ZIP Code. This solution, however, only maps the arrest data to around half of the ZIP Code. The results revealed using this method were not ideal down to community level. So they eventually are only used for results by borough.

6.2.2 Normalizing location data to grids

As the latitude and longitude information is given in Arrest and Motor Vehicle Collision Datasets, it can be used as a more precise source of location data. Although this kind of data can be directly consumed and analysed, the amount of records is too large. To reduce the workload of our system, both latitudes and longitudes are rounded to the nearest 0.01 (`~0.7 miles`), which enables us to group the data points nearby together, and compare it with the coronovirus data.

6.3. Handling COVID Case Information

The original datasets of coronavirus are separated into three tables. Each table has 183 geographic information columns. That causes the trouble when we want to load the data into hive since we need to provide the schema by ourselves. According to the characteristics of other dataset, we want to make the data of one zip code in one sampling time becomes an independent row. This can help us reduce the number of columns and facilitate the future join operations. From observing our data and metadata, we noticed that the level of geography is indicated following the underscore (_) in each column heading and the same index of one column corresponds to the same geographic area. Therefore we can use the column index to stand for the same zip code information. For `testrate`, `caserate` and `percentpositive`, we created three mappers. For each mapper, we extracted the date and the index of column as the key and inserted different character mark before record data in three mappers. This can help us assemble the columns in a fixed order in reducer. Besides, in `TestRateMapper`, we parsed the header information in setup function so that we can know the exact zip code the column index corresponds to before mapping. Then we attached the `zipcode` with the `testrate` and passed it to reducer. In the reducer, we collected three columns for the same time and area information and put them in the fixed order, finally replaced the column index with true zip code. Another problem we need to deal with is the null value. At first, we wanted to replace the null value with average values of its neighbors in the time series. But it was hard to achieve in MapReduce. Since the columns for coronavirus are all rate values and there are few null values, we decided to replace the null values with the citywide rate in the same period.

7. Results

7.1. NYPD Arrests and Motor Vehicle Collisions

As shown in Figure 3, the police arrest data and traffic collision data, with respect to either location only, time only, and both location and time, have a very strong positive correlation, with *Pearson correlation coefficients* larger than 0.9. The results imply that both data can be a good indicator of the public safety situation in the city.

7.2. COVID Cases and Public Safety by Time

We find that traffic collision numbers and COVID-19 positive test rates are negatively correlated, and NYPD arrest numbers and COVID-19 positive rates are negatively correlated. We plotted a graph of the rate of positive tests for COVID-19, the number of traffic accidents, and the number of NYPD arrests over three years, as shown in Figure 4. We can find that there are three outbreaks of COVID-19: winter 2022, winter 2021, and summer 2022. During all these three periods, both traffic accidents and police arrest numbers have significant rises.

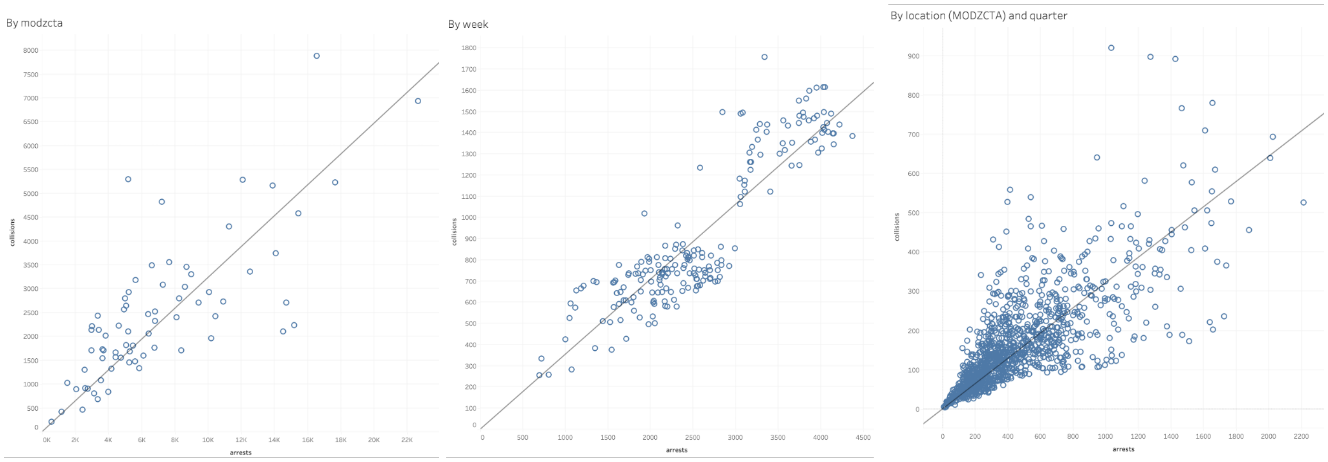
Putting it in another way, we can discover that after each outbreak has subsided, there is a period of terrible public safety situation. This finding is very much in line with common sense. People may tend to reduce social activities when the epidemic is severe; however, mental and property damage and other instability brought by this will erupt following epidemic subsides.

To be more specific, we classified our arrests data into three types according to the level of offense:felony, misdemeanor, violation. We then plotted the graphs of their amounts and percent positive rate. 5, 6, 7 show the similar pattern: With the outbreaks of covid-19, the amounts of three types of crimes all increase. When the positive rate declines, the amounts of three types of crimes all decrease.

This finding reminds the government and the population to be more aware of public safety events when the epidemic outbreaks. Lock-down policy does not reduce the occurrence of accidents and crimes. Negative emotion and finan-

Dataset	Location-related Data		
NYPD Arrest	Borough	Precinct	Latitude and Longitude
Coronavirus	Borough	MODZCTA (*Modified ZIP Code Tabulation Area*)	
Motor Vehicle Collisions	Borough	ZIP Code	Latitude and Longitude

Table 1: Location-related data supplied by each dataset



(a) by location ($r \approx 0.934$)

(b) by week ($r \approx 0.987$)

(c) by quarter and location ($r \approx 0.911$)

Figure 3: Arrests and Motor Vehicle Collisions

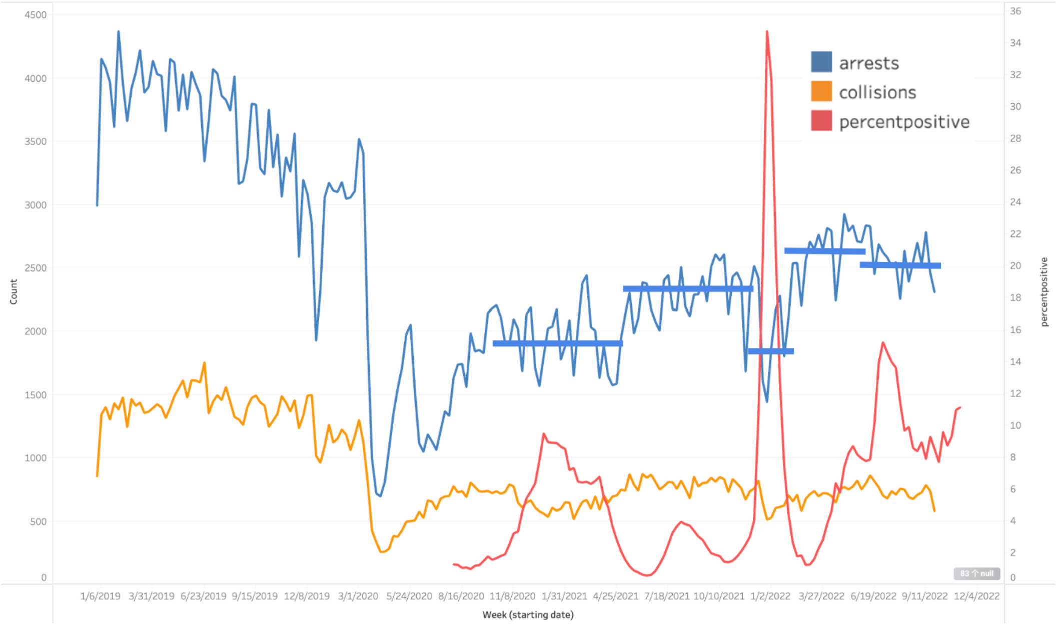


Figure 4: Correlation between COVID-19 cases and public safety among time

cial burden, racial discrimination may increase the risk of crimes.

7.3. COVID Cases and Public Safety by Location

Using the location information in our datasets, we tried to find some insight into public safety and COVID-19

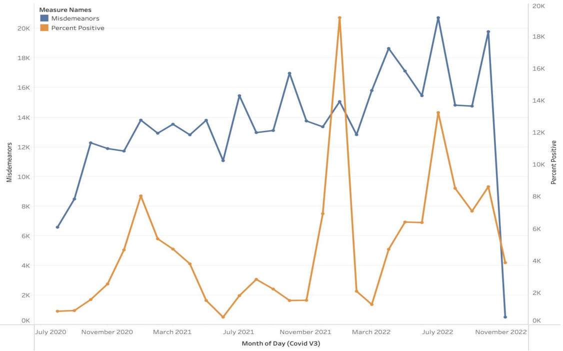


Figure 5: Correlation between COVID-19 cases and misdemeanor among time

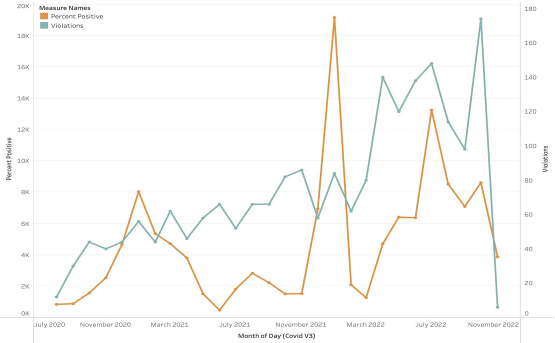


Figure 6: Correlation between COVID-19 cases and violation among time

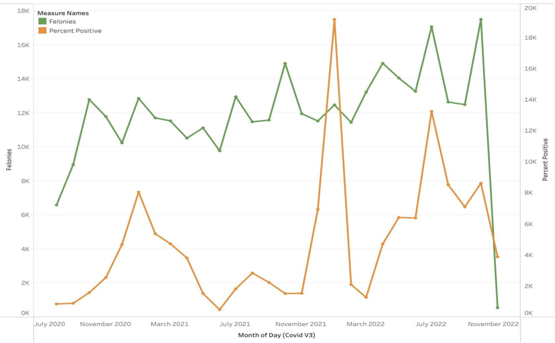


Figure 7: Correlation between COVID-19 cases and felonies among time

among different locations. We have selected two representative time points: the week of 01/16/2021, and the week of 04/16/2022. In Figure 8, we can see that the positive rate is high in 2021, especially in rural areas. However, in 2022, the positive rate is relatively low among rural areas but high in Manhattan. In Figure 9, we can find that most areas remain a low motor Vehicle collision rate compared to pre-epidemic, except the Bronx, Upper Manhattan, and some other locations. Besides, the total motor Vehicle collision rate raised from 2021 to 2022. From Figure 10, we can see how the NYPD arrest number has changed since the epidemic began. There is a significant rise in police arrest numbers in Downtown Brooklyn, Financial District, and Forest Hills. Other places remain the same or have a lower rate of police arrest cases.

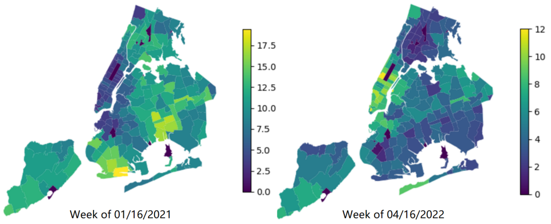


Figure 8: COVID-19 positive rate among different locations of two weeks

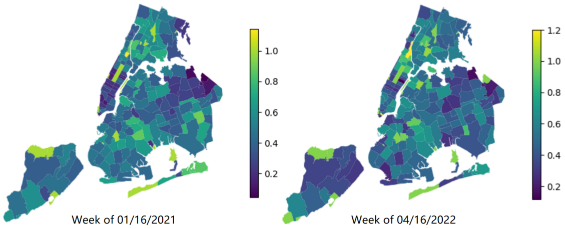


Figure 9: Motor collisions number change rate compared to pre-epidemic

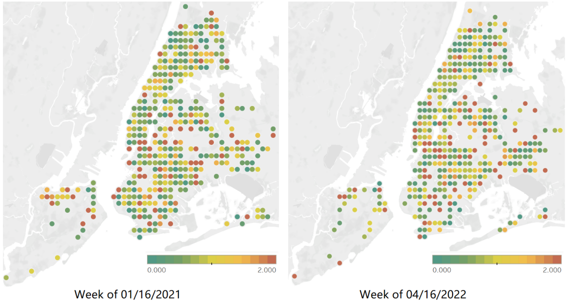


Figure 10: NYPD arrests number change rate compared to pre-epidemic

We find it hard to generalize a model for the impact of COVID-19 on public safety among regions. We have analyzed some of the following reasons: COVID-19 and motor Vehicle collisions might lack intrinsic connection since many crashes happen on highways and have little impact on the local epidemics; NYPD arrests change rate on latitude and longitude scale may be challenging to summarize the pattern because the scale is too small and should be combined to a larger geographical scale.

8. Obstacles

Several obstacles were encountered in developing our analysis.

At first, Hive CLI (*Beeline*) was majorly used for our analysis. As we got familiar with Tableau, we found that the Tableau is a much powerful tool for both analytics and visualization. Different scale, range, and key-value pairs can be easily configured in the application. It is more intuitive to use than the CLI. Moreover, as the cluster became busier towards the end of the semester, the CLI was significantly slower than weeks ago. So we decided to move most of our job to Tableau.

Additionally, we tried to use Tableau to draw the heat-map, but it was hard to draw the result we wanted, and we couldn't combine it with external datasets. So we exported data from Tableau and used python's GeoPandas package to draw the heat-map. GeoPandas has a large amount of external geographic data, we just need to simply put data of all the zip code areas into an array in order.

When analyzing the relationships between public safety and the COVID-19 positive test rate among regions, the correlation was insignificant statistically. We sample many different time slices to find some patterns and have to conclude that the relationship between them is unclear.

9. Summary

Based on the analysis before, we have these findings:

- Strong Correlation was found between the number of motor collisions and the number of NYPD arrests.
- With the outbreaks of covid-19, the number of NYPD arrests increase. With the recession of COVID-19, the number of NYPD arrests decrease.
- Negatively correlation between number of motor collisions and COVID-19 cases.

Those patterns we found could enable the government to better administer the public transit and social safety, and give businesses an idea of how to adjust the management and operation in accordance with the pandemic. The impact across regions needs further analysis and cannot be generalized by a simple model. The public safety may be affected by other geographic factors, such as population, economic level, education level. In the future research, we can consider those elements with COVID-19 data to explore the region pattern.

Acknowledgment

We would like to express our deep gratitude to Professor *Yang Tang*, *NYU High Performance Computing Team*, and *Google Cloud Platform*. This project is not possible without their support. Also, the datasets generously provided by *New York City Open Data*, and *NYC Department of Health and Mental Hygiene* are very much appreciated. Additionally, we would like to thank *Tableau* for sponsoring the license, which is of great use for the better visualization of our results.

References

[1] J. Rosenbaum, N. Lucas, G. Zandrow, W. A. Satz, D. Isenberg, J. D'Orazio, N. T. Gentile, and K. E. Schreyer, "Impact of a shelter-in-place order during the covid-19 pandemic on the incidence of opioid overdoses," *The American journal of emergency medicine*, vol. 41, pp. 51–54, 2021. 1

[2] M. M. Esposito and A. King, "New york city: Covid-19 quarantine and crime," *Journal of criminal psychology*, vol. 11, no. 3, pp. 203–221, 2021. 1

[3] M. W. Meyer, "Covid lockdowns, social distancing, and fatal car crashes: more deaths on hobbesian highways?" *Cambridge Journal of Evidence-Based Policing*, vol. 4, no. 3, pp. 238–259, 2020. 1

[4] NYPD Arrests Data (Historic). New York City Police Department (NYPD). Accessed Nov 6, 2022. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u 2

[5] NYPD Arrest Data (Year to Date). New York City Police Department (NYPD). Accessed Nov 6, 2022. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc 2

[6] COVID-19 Historic data. Github. Accessed Nov 6, 2022. [Online]. Available: https://github.com/nychealth/coronavirus-data/tree/master/trends 2

[7] COVID-19 ZIP Code metadata. Github. Accessed Nov 6, 2022. [Online]. Available: https://github.com/nychealth/coronavirus-data#geography-zip-codes-and-zctas 2

[8] Motor Vehicle Collisions - Crashes. Police Department (NYPD). Accessed Nov 6, 2022. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95 2

[9] Modified Zip Code Tabulation Areas (MODZCTA). New York City Department of Health and Mental Hygiene (DOHMH). Accessed Nov 6, 2022. [Online]. Available: https://data.cityofnewyork.us/Health/Modified-Zip-Code-Tabulation-Areas-MODZCTA-/pri4-ifjk 4

[10] Precincts - NYPD. New York City Police Department (NYPD). Accessed Nov 6, 2022. [Online]. Available: https://www.nyc.gov/site/nypd/bureaus/patrol/precincts-landing.page 4

Appendix

Attached with this report are source codes, scripts, and some miscellaneous data (e.g. mappings for location) that were produced for this project.

- `arrests`: Assets related to NYPD Arrest datasets:
  - `by_grid`: MapReduce code for grouping arrest data by latitude and longitude;
  - `by_modzcta`: MapReduce code for grouping arrest data by precinct;
  - `compileAndRun.sh`: Script for compiling and run the MapReduce programs;
  - `modzcta.csv`: Mapping between regular ZIP code and MODZCTA;
  - `precinct_zip.sh`: Mapping between precinct and ZIP code.
- `covid`: Assets related to NYC Coronavirus dataset:
  - `Data_ingestion`: Commands and codes used for processing COVID data;
  - `hive_commands`: Hive commands used to create the COVID table.
- `collision`: Assets related to Motor Vehicle Collision dataset:
  - `code_collision`: MapReduce program used to process motor collision data;
  - `code_python`: Python script used to draw the heat-maps.