Analyzing the Socio-Economic Impacts of COVID-19 on New York City

Hayden Edelson, Jin Gyu, and Andy Juan
The COVID Guys in NYC
Tandon School of Engineering, New York University

Abstract

In this document, we present a comprehensive analysis of the socio-economic impacts of COVID-19 on New York City. As NYU students, citizens of New York, we are passionate about the city's recovery. We analyze publicly available data to determine how the City has changed as a result of the pandemic, in what directions various socioeconomic trends are moving, and what the future may look like. Specifically, we analyze trends in crime, mobility, and economic activity using datasets from New York City's OpenData store, the New York MTA, and investing.com, a financial market data provider. We list our datasets in Section 2. We present our data profiling, cleaning, and wrangling methodology for all of our datasets in Section 3. We utilize PySpark, OpenRefine, OpenClean, and Pandas to conduct our analyses, and we present our code in Jupyter notebooks, stored in the Github repository referenced in Section 6.

1 Introduction

New York City is the most vibrant and diverse city in the world. The energy of this great metropolis derives in large part from its thriving economy, its cultural inclusivity, and its never-ending selection of places to go and things to do. COVID-19 has had a devastating impact on New York. For a time, it seemed that COVID had shut the lights off in the city that never sleeps—and it wasn't always clear that they'd ever turn back on.

But today, New York appears to be roaring back to life, so we set out to analyze the socio-economic impacts of COVID-19 on the City. Specifically, we analyze trends in crime, mobility, and economic activity in an attempt to develop a comprehensive understanding of the social disruptions wrought by COVID-19 and their consequences on New Yorkers' way of life. We use datasets from New York City's OpenData store, the New York MTA, and investing.com, a financial market data provider. We believe all three of these sources provide trustworthy provenance and deep insight into the phenomena of interest.

In analyzing crime, we focus on hate crime. The past year in the United States has been marked by racial unrest and a dramatic spike in anti-Asian hate crime. Therefore, we felt it important to develop better insights into recent hate crime data—and to consider how New York City's rich cultural makeup has been impacted in the wake of COVID-19.

In mobility, we focus on subway ridership. The New York City subway is a pillar of the city's social, cultural, and economic constitution. By analyzing subway ridership, we aim to understand broader trends in traffic flows and social activity.

In economic activity, we focus on investigating the impact of COVID-19 on general economic indicators as well as on the NYC small business community. We conduct time series analyses of multiple stock market indicators and compare them to the daily COVID positive rate in NYC. In addition, we compare local business formation in 2019 and 2020 by analyzing data from the NYC Department of Consumer Affairs.

Overall, the data tell a cautionary tale. The impacts of COVID-19 have been dramatic, and the data show that many of its effects continue to linger. While those of us who live here know that the city, its people, and its institutions are adjusting to the best of their ability, daily life remains far from "normal" in many ways.

2 Datasets

Briefly, we list and describe the datasets used in this report.

2.1 COVID-19

COVID-19 Cases in New York City: `https://github.com/nychealth/coronavirus-data/blob/master/trends/tests.csv`

This dataset, provided by the New York City Department of Health, shows daily coronavirus tests and test results along with short-term moving averages.

2.2 Crime

NYC Total Crime: `https://data.cityofnewyork.us/Public-Safety/NYPD-Hate-Crimes/bqiq-cu78`

This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to 2020.

NYC Hate Crime 2019-2021: `https://data.cityofnewyork.us/Public-Safety/NYPD-Hate-Crimes/bqiq-cu78`
This dataset, provided by the New York Police Department, contains details of confirmed hate crime incidents in New York City from 2019 to 2021.

NYC Hate Crime 2018: `https://www1.nyc.gov/site/nypd/stats/reports-analysis/hate-crimes-archive-2018.page`
This dataset, provided by the New York Police Department, contains details of confirmed hate crime incidents in New York City in 2018.

NYC Hate Crime 2017: `https://www1.nyc.gov/site/nypd/stats/reports-analysis/hate-crimes-archive-2017.page`
This dataset, provided by the New York Police Department, contains details of confirmed hate crime incidents in New York City in 2017.

2.3 Mobility

MTA Turnstile Data: `http://web.mta.info/developers/turnstile.html`
This data warehouse, managed by the New York MTA, contains CSV files showing daily, turnstile-level subway ridership data. Files are added on a weekly basis. We used files dating back to the beginning of 2020, coinciding with the onset of the pandemic.

2.4 Economy

Legal Operating Businesses in NYC: `https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh`
This dataset, provided by the New York City Department of Consumer Affairs, features information about individuals and businesses that hold DCA licenses, allowing them to legally operate in New York City.

Stock Market Data: `https://www.investing.com/indices/`
This website contains historical data for numerous stocks and stock market indices. We an-

alyze the 2020 performance of the S&P 500, the Dow Jones Industrial Average, the Nasdaq 100, the Russell 1000, the Russell 2000, and the New York Stock Exchange Composite.

3 Data Cleaning and Integration

3.1 Crime

For the crime data, our big data analysis tools included PySpark, OpenRefine, and Pandas. We used PySpark and distributed computing to analyze data files that were too large to manage on our local drives. We used OpenRefine to facilitate data cleaning, and we used Pandas to conduct additional exploratory data analysis and generate visualizations.

The NYC Total Crime file (described above) is approximately 2GB large. For a file this large, PySpark proves to be a much more efficient solution than simple analytical libraries on local drives. Using PySpark, we found that this dataset had missing values and duplicate values in its ”Full Complaint ID” and ”Arrest ID” columns. Based on our investigation of the data, we concluded that rows with a ”Full Complaint ID” but no ”Arrest ID” represented complaints that did not result in arrests. Rows with duplicate ”Full Complaint ID” values but different ”Arrest ID” values represented different arrests for different people involved in the same crime incident. And finally, rows with duplicate ”Arrest ID” values but different ”Full Complaint ID” values represented multiple complaint calls for the same incident. Subsequently, we decided to drop rows with duplicate values, since we were primarily interested in incidents of crime, not necessarily arrests or complaints. After using PySpark to eliminate duplicates, we used OpenRefine to eliminate typographical errors and to standardize the data formatting. Finally, we used Pandas in a Jupyter notebook to store the cleansed data and continue our data profiling.

Another data quality issue that we encountered with this dataset was that there were two rows where the ”Complaint Year Number” was 2021, but the “Month Number” was 12, indicating that these complaints were from the future. In both cases, we changed the “Month Number” to 1 based on the “Record Create Date”. These data cleaning and integration steps prepared our data for the analysis presented in Section 4.

3.2 Mobility

For mobility data, our big data analysis tools included OpenClean and Pandas, using SQL-like merge and group-by queries to adjust the geographical and temporal resolution of our data. This data was generally high-quality. There were no missing values, but through extensive wrangling and feature engineering, we did discover some inconsistencies and impossible values. First, we read in all the data files provided by the MTA from 1/2/2020 to 5/1/2021. The data provides the cumulative counts of the number of people who enter and exit the subway at turnstile-level specificity. Data is collected at regular time intervals throughout the day, and new data files are released on a weekly basis.

For our purposes, turnstile-level specificity was too granular. We focused our analysis on overall daily ridership and daily ridership by station. To compute these values, we first compute daily entries and exits per turnstile (rather than cumulative), and then we group by station, linename, and date to compute daily entries and exits per station. Later, we analyze total daily ridership for the entire subway system as well as average daily ridership per station, computed on a quarterly basis, from the start of 2020.

We use Python's OpenClean library to identify value outliers and clusters. Using OpenClean's DBSCANOutliers algorithm, we see that one station reports over 3.6 billion entries on a single day. This is an impossibly large value. To eliminate extreme outliers and nonsensical values, we assume that no more than 8 million people (approximately the total population of NYC) can enter or exit a single subway station on a single day. We believe it is reasonable to assume that the number of daily entries or exits for a single subway station should be less than the population of New York City, even taking into account tourists and multiple entries per individual. Removing these values eliminates outliers when the DBSCANOutliers algorithm is run with an epsilon value of 0.5.

We attempt to use OpenClean's clustering algorithm to identify value clusters and typographical errors, but the algorithm fails to identify any. Through manual searching, we identified a number of inconsistencies in the dataset's ”linename” column, which lists the subway lines that stop at a given station. For example, for Times Square-42nd st, the ”linename” column includes both ”1237ACENQRSW” and ”ACENQRS1237W”—the same lines but in different orders. These kinds of duplicate values interfere with group-by queries that group by train line, so we employ a lookup table to correct all such data quality issues.

3.3 Economy

For economic data, our primary big data analysis tool was Pandas. We used Pandas to structure the data, identify data quality issues, and prepare the datasets for visualization.

The dataset we used to analyze legal operating businesses in New York City was obtained from NYC OpenData. We imported the dataset into a Pandas DataFrame, and used the Pandas library to filter it down to the relevant time period. The dataset contained records dating back to 2001, but for the purposes of our analysis, we focused on 2019 and 2020.

Using Pandas, we identified numerous null values and incomplete records. In addition, the ”License Expiration Date” and ”License Creation Date” columns had several formatting issues. For these columns, we used Pandas to drop rows that had invalid values.

We retrieved our financial market data from investing.com. We chose investing.com because it is a well respected data provider with user-friendly access mechanisms. Our datasets span the entire year of 2020. The data included daily stock market prices, formatted as CSV files. We used Pandas to read the data into DataFrames and make formatting adjustments, such as removing the commas from the daily closing prices, to prepare the data for visualization.

4 Analysis and Findings

4.1 Crime

The overall crime data show that COVID-19 has, somewhat surprisingly, had a benign impact on the overall crime rate in New York City. The total number of crimes in NYC decreased from approximately 46 thousand in 2019 to approximately 40 thousand in 2020 (an 11.5% decrease). From the start of the pandemic in March 2020 to the reopening in June, the crime rate was lower than it had been in the previous two years over the same time period. Notably, April 2020 experienced a record drop in total crime. From June to November 2020, the number steadily increased, but as NYC retreated back into lockdown—restricting restaurants and public schools—the numbers again dwindled. This trend did not differ among felony, misdemeanor, or violation charges. Figure 1 shows these crime rates over the period from January 2018 to December 2020. These findings indicate that NYC's restrictions on the movement of its citizens had a considerable effect on the number of crimes that could occur.

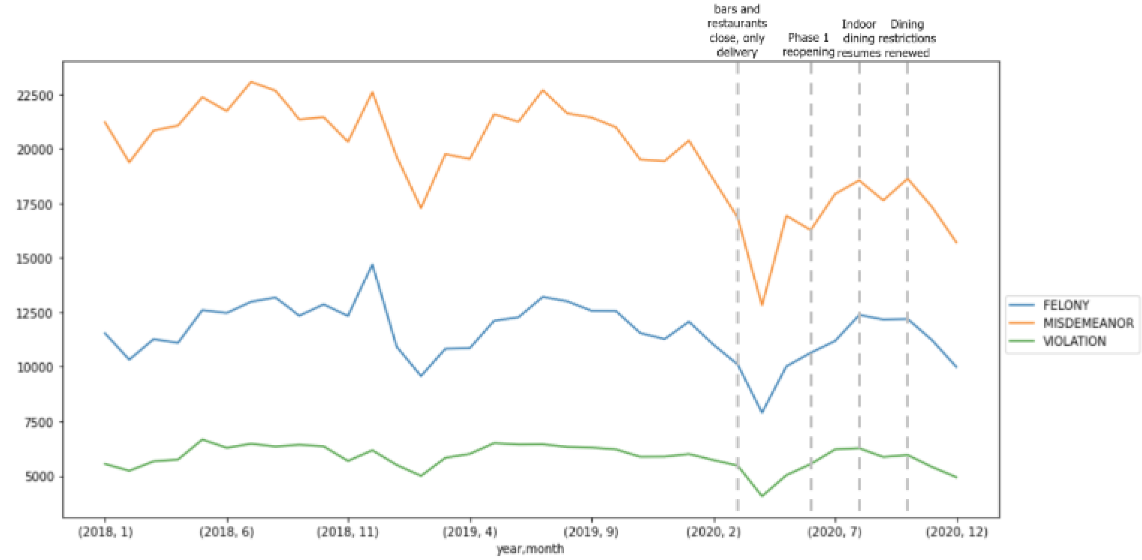


Figure 1: Felony, misdemeanor, and violation crimes

Some of the most interesting, and perhaps telling, trends in New York City crime data pertain to hate crime. Overall, the number of hate crimes in New York City dropped from 421 in 2019 to 264 in 2020. However, the trends differed based on the kind of hate crime and the victim group. There was a significant decrease in the number of anti-Jewish crimes—almost a 50% drop. Astonishingly, the decrease in anti-Jewish hate crime accounted for almost the entire delta between hate crime in 2019 and hate crime in 2020.

Although the city experienced a profound movement of Black Lives Matter activism in 2020, the number of anti-Black crimes remained the same. In fact, census data reports that the Black population in New York City has decreased by approximately 3% since 2010, so stagnation in the number of anti-Black hate crimes could actually imply an increase in the number of crimes per capita of the Black population. The broader social implications of this data are difficult to intuit. They could indicate that BLM has led to an increase in racial conflict, or they could exemplify why our city needs movements like BLM.

The most apparent increase from 2019 to 2020 was in anti-Asian hate crimes. While in 2019, there was only 1 case of anti-Asian hate crime, that number grew to 28 cases in 2020. And in the first quarter of 2021 alone, that number has reached 42—making Asians the most targeted demographic in the City. Figure 2 shows some of the most significant trends in New York City hate crimes from January 2019 to Q1 2021. Appendix Exhibit 1 shows 2019, 2020, and Q1 2021 hate crime data organized by the targeted demographic group.

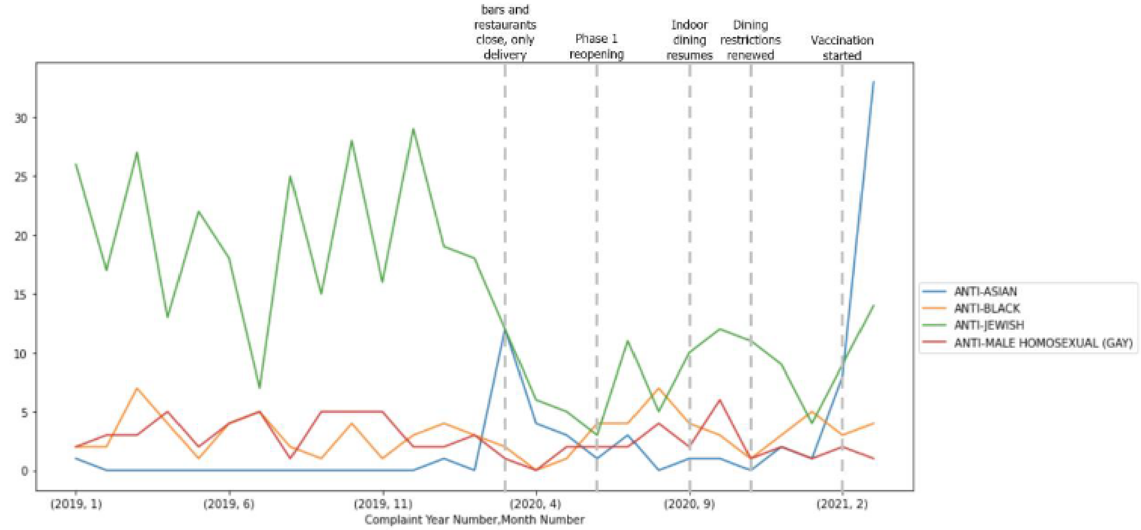


Figure 2: Hate crimes by targeted demographic

This trend is deeply worrying. While the overall crime rate decreased in 2020, theoretically making the city a "safer" place in terms of crime, anti-Asian hate crime spiked dramatically, leaving many fearful and incensed. Obviously, these numbers are tightly correlated to the spread of COVID-19. The surge in hate crime can be traced back to March 2020, suggesting that COVID-19 is a significant factor explaining this increase.

As an increasing proportion of New York City's population becomes vaccinated, and the city continues to reopen, anti-Asian hate crime becomes an ever greater threat. The spike in hate crimes has inspired a national "Stop Asian Hate" movement to protect Asian Americans in the streets and persecute their attackers.

4.2 Mobility

COVID-19 has had, and continues to have, a significant effect on New York City subway ridership. Subway cars that were once packed with morning commuters and weekend revelers now ride empty, as companies have enabled their employees to work from home and much of our social activity remains restricted. What was once perceived as the preferred mode of transit for many New Yorkers is now generally avoided, as our aversion to crowded spaces with poor ventilation has made the subway an inviable option for some.

Figure 3 shows daily subway entries from January 2020 to May 2021. As is shown in the graph, daily subway ridership experienced a steep drop-off in March and April of 2020, and the number of subway riders remains deeply depressed compared to pre-COVID levels.

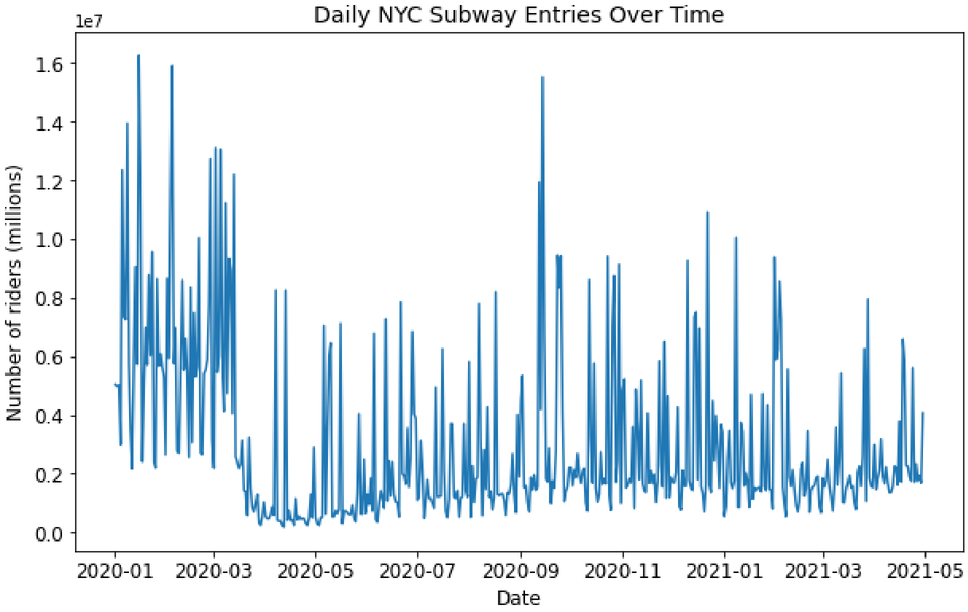


Figure 3: Daily NYC subway ridership

Given this pattern in total daily ridership, we sought to discover how movement patterns had changed as well. We wanted to answer questions like: is this drop-off primarily attributable to work from home? Or is it due to a decrease in social activity and events? We analyzed how people's movement patterns changed by tracking which subway stations were the most visited over the course of the pandemic. We hypothesized that if ridership to and from Midtown Manhattan had rebounded to a substantial degree, then work from home was likely not the main culprit. Moreover, if stations in Brooklyn and Queens appeared more active, perhaps these data reflect New York's shifting population centers.

Figure 4 shows the top 20 subway stations for Q1 2020 and Q3 2020. In Q1, the top 2 subway stations were, unsurprisingly, 34th St-Herald Square and Grand Central Station. In Q3, which was perhaps just after the first wave of the pandemic and leading into the second, many of these core Manhattan transit hubs were conspicuously absent from the top of the list. Ridership to certain stations remained high, but the character and locations of these stations shifted away from major transit hubs in lower Midtown Manhattan to much less dense and much less central.

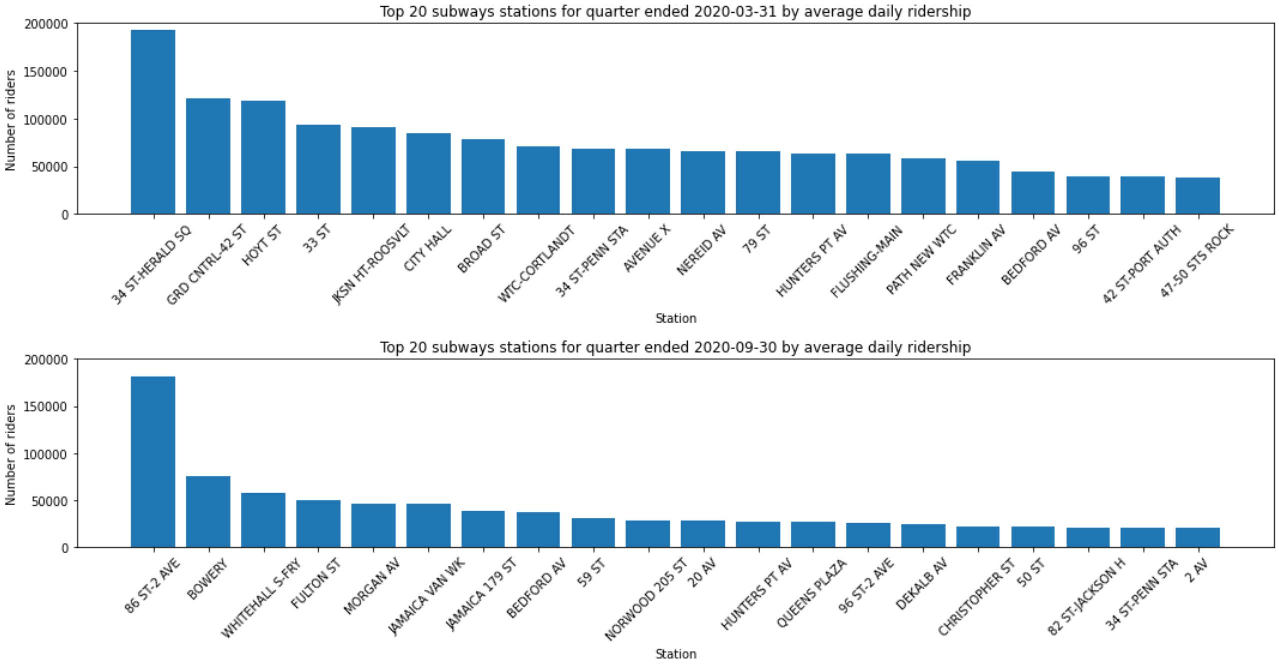


Figure 4: Top 20 subway stations, Q1 Q3 2020

And subway ridership continues to decline. Figure 5 shows the top 20 subway stations in Q1 2021. While the makeup of this top 20 seems to more closely resemble that of prepandemic New York, the volume of passengers is even lower than it was in Q3. Perhaps these trends merely reflect changing population centers: New York City's population continues to dwindle to some extent as city dwellers move to suburbs and lower cost states. These data may suggest that while economic and social life are returning to "normal," as people return to their offices and the city's center, they do not suggest that New York is returning to the city it was prior to the pandemic. These depressed ridership numbers may forewarn serious issues for the City's "belove-hated" subway system, which already suffers from chronic financial distress and service interruptions. It may be a fairly long time before people feel comfortable riding the subway again.

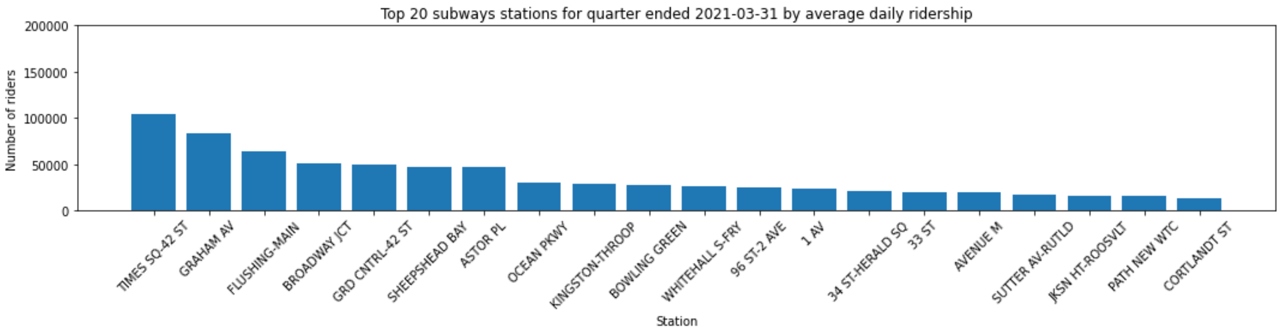
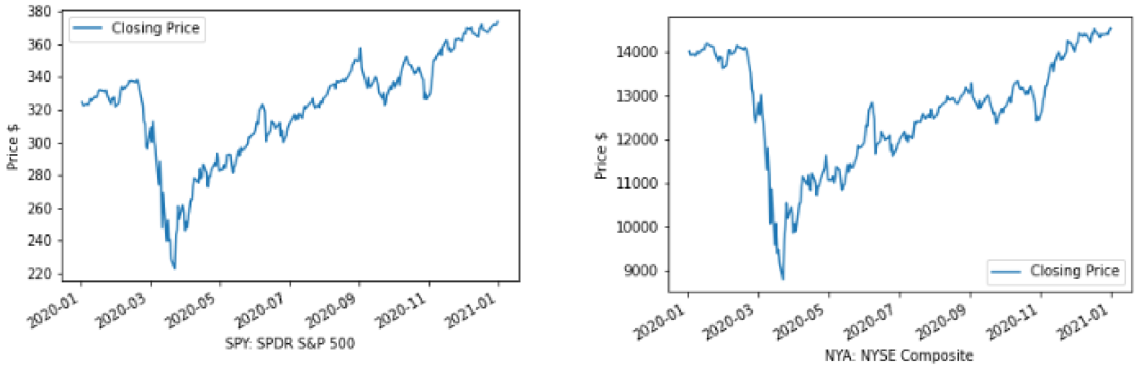


Figure 5: Top 20 subway stations, Q1 2021

4.3 Economy

U.S. stock markets have several tradeable financial products that track the performance of specific industries and market sectors. Figure 6 shows the performance of several stock market indices over the course of 2020. The S&P 500 index is a free-float, weighted-measurement index of 500 of the largest publicly traded companies in the United States. It is one of the most commonly followed equity indices. The S&P 500's performance is shown in the top-left plot of Figure 6. We also show the NYSE Composite, the Nasdaq-100, the Russell 2000, the Russell 1000, and the Dow Jones Industrial Average. In these charts, we see a significant dip across all market indicators between the months of March and April 2020. In the months following this crash, we observe a strong recovery that even surpasses pre-COVID prices.



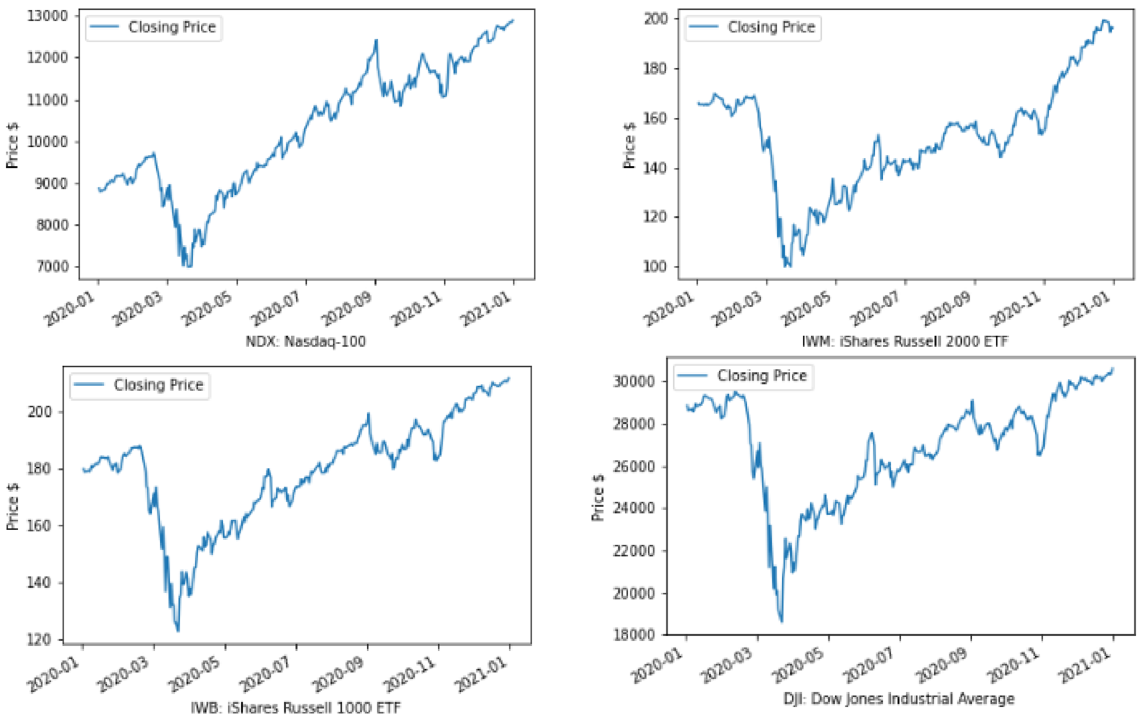


Figure 6: Financial market indicators

These financial market indicators are highly inversely correlated to COVID cases in New York City. Figure 7 shows confirmed daily COVID cases in NYC. We can see that the dramatic spike in COVID cases very closely corresponds to the drop in the financial markets, and the subsequent reduction in daily COVID cases correlates to the financial markets' rise. These data show how the phenomena that unfold in New York City have systemic implications for broader society. As one of the earliest and most hardly hit cities, New York's resilience and robust recovery inspired confidence in the financial markets' outlook for the future.

For many small businesses, however, economic recovery has not been quite so swift. From 2019 to 2020, there was a 60% drop in the number of new business licenses issued by New York City's Department of Consumer Affairs. The drop was more severe in some industries than in others, but the overall 60% drop is a fearsome decline. Figure 8 shows the 10 industries with the largest declines in the number of new licenses issued.

Small businesses in New York City are integral components of the City's cultural and economic backbone. They not only generate revenue and economic activity, but also form

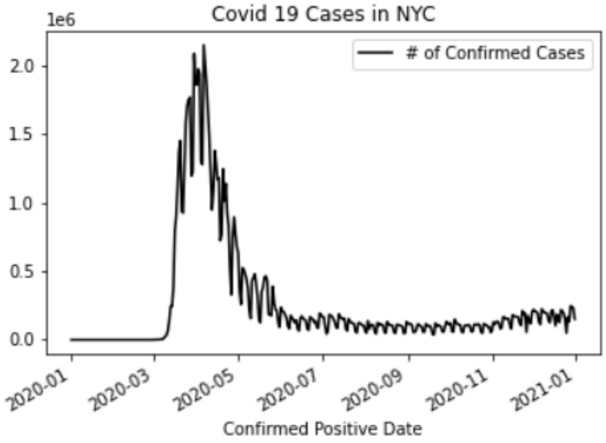


Figure 7: Daily COVID-19 cases in NYC

the basis of many families and communities. They represent the heritage of New York City's entrepreneurial, artistic, and immigrant history. A long-term depression in small business formation could have deeply negative consequences for New York's economic and social fabric. We sincerely hope that the City's small business economy is able to rebound forcefully in 2021.

5 Conclusion

In this project, we attempted to present a comprehensive analysis of the socio-economic impacts of COVID-19 on New York City. As New York City emerges from the COVID-19 pandemic, much of life appears to be returning to normal, but we sought to determine if the data support that perception. We analyzed trends in crime, mobility, and economic activity using datasets from New York City's OpenData store, the New York MTA, and investing.com. We used PySpark, OpenRefine, OpenClean, and Pandas to integrate several large datasets and conduct extensive data profiling, cleaning, and wrangling. We used Jupyter notebooks to generate visualizations to inform our analysis, and we store these notebooks in the Github repository referenced in Section 6. Our analysis shows that while life may appear to be getting back to normal, the data remind us that there is much work to be done to ensure a widespread and robust recovery—for both our city and its inhabitants.

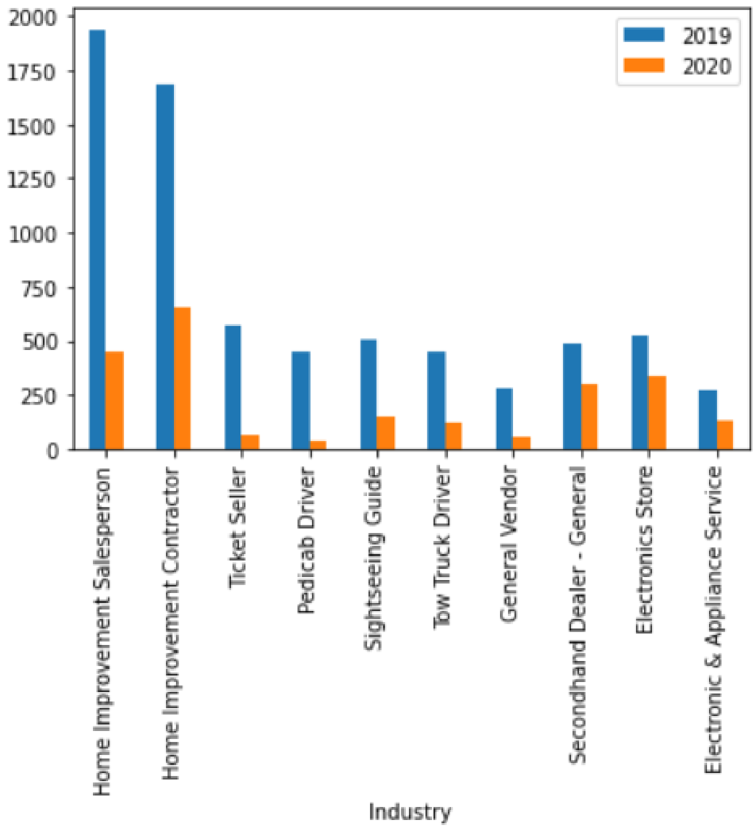


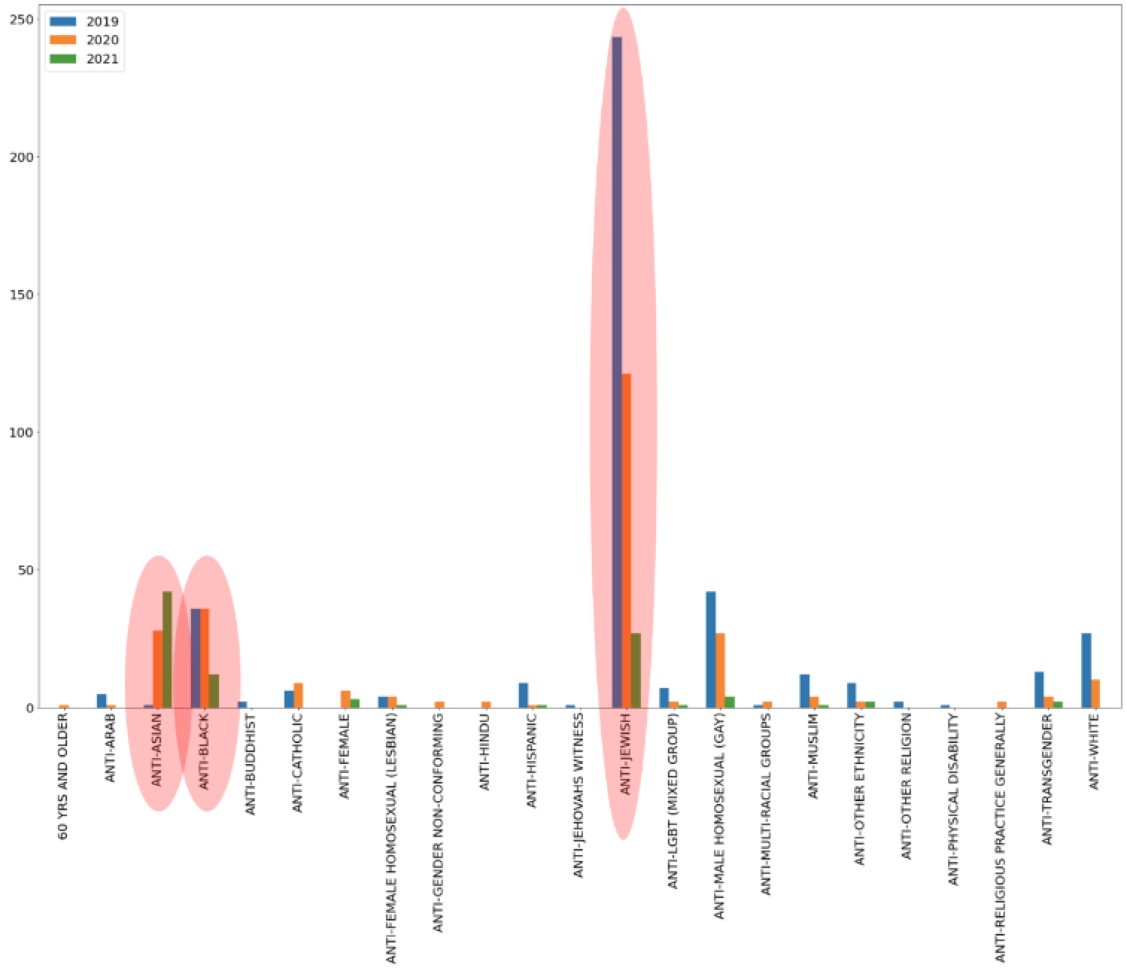
Figure 8: 10 largest drops in new business license issuances

6 Github

Link to GitHub repo: `https://github.com/jgingh7/NYC-COVID-BigData`

Appendix

Exhibit 1: Hate crimes in 2019, 2020, and 2021 by targeted demographic


60 YRS AND OLDER
ANTI-ARAB
ANTI-ASIAN
ANTI-BLACK
ANTI-BUDDHIST
ANTI-CATHOLIC
ANTI-FEMALE
ANTI-FEMALE HOMOSEXUAL (LESBIAN)
ANTI-GENDER NON-CONFORMING
ANTI-HINDU
ANTI-HISPANIC
ANTI-JEHOVAHS WITNESS
ANTI-JEWISH
ANTI-LGBT (MIXED GROUP)
ANTI-MALE HOMOSEXUAL (GAY)
ANTI-MULTI-RACIAL GROUPS
ANTI-MUSLIM
ANTI-OTHER ETHNICITY
ANTI-OTHER RELIGION
ANTI-PHYSICAL DISABILITY
ANTI-RELIGIOUS PRACTICE GENERALLY
ANTI-TRANSGENDER
ANTI-WHITE