NYC COVID-19 Transportation Analysis Final Report

Jason Lai, Kewei Zhang

Introduction

The COVID-19 pandemic has had an unprecedented impact on our way of life, particularly in terms of transportation. In this study, we sought to explore the effects of lockdown measures on public transportation in the New York City area. Specifically, we analyzed data from two primary sources of transportation, namely the MTA subway turnstile usage and Citi Bike trip data, spanning from year 2019 to 2022. By examining patterns in public transportation usage before, during, and after the pandemic, we aimed to identify any significant changes in people's transportation behavior, as well as to explore whether a "new normal" has emerged in the post-COVID period.

Problem Statement

Our study aims to build upon the findings of Wang and Noland's research and extend the investigation to aspects that were not covered in their original study. For instance, the authors stated that "it is unknown whether these [transportation usage behavior] changes will be sustained post-COVID." Therefore, to address this uncertainty, we will expand the dataset used in the study from 2019 vs. 2020 to 2019 vs. 2022 and 2020 vs. 2022, in order to assess whether the trends observed before and during the pandemic persist into the post-COVID period. Specifically, we seek to understand whether there has been a return to pre-pandemic levels of public transportation usage, or whether usage patterns have remained unchanged. We will also examine the growth or decline of Citi Bike's subscribed and casual users trips in the "new normal," as well as analyze the patterns in the context of subway usage. Throughout our analysis, we aim to provide insight into these questions and contribute to a deeper understanding of the post-pandemic public transportation landscape.

To provide further context for our data analysis, we consulted multiple official sources, including the CDC's COVID-19 timeline. These sources helped us to better understand the progression of the pandemic and how it may have impacted the behavior of public transportation riders in NYC. Our team members who lived through the pandemic in the NYC area also provided valuable observations to further inform our analysis. By incorporating these various sources of information, we aim to provide a comprehensive and nuanced understanding of the trends and patterns in public transportation usage before, during, and after the pandemic.

Related Work

The research paper *Bikeshare and subway ridership changes during the COVID-19 pandemic in New York City* by Wang and Noland is a valuable source of information that forms the foundation of our study. It provides a comprehensive analysis of the impact brought by the COVID-19 pandemic on subway and Citi Bike usage in New York City. The authors note that both subway and bike share ridership experienced a significant decline during the early stages of the pandemic, but bikeshare usage showed a more resilient recovery as lockdown measures were eased. The paper also highlights the different patterns of bike usage by subscribed and casual users, and examines the various factors that influenced these trends, including demographic and socioeconomic factors, weather conditions, and the availability of outdoor spaces for recreation and exercise. The authors conclude by suggesting that alternative modes of transportation, such as bikeshare, may become more popular in urban areas in the future, as people continue to prioritize health and safety concerns in their daily lives. Overall, the paper provides valuable insights into the impact of the pandemic on transportation patterns in New York City, and serves as an important reference for our own analysis of the data.

Methods, Architecture, and Design

The MTA turnstile usage and Citi Bike trip data from 2019, 2020, and 2022 served as our two primary datasets. The MTA turnstile usage data was obtained from the New York State open data platform, while the Citi Bike trip data was sourced from the official website of Citi Bike NYC. It is worth noting that the methodology and procedures employed in our study closely mirror those adopted by Wang and Noland, although not explicitly outlined by them in their paper. To be more specific, the authors use a 7-day moving average as a measurement for their data, which provides benefits such as smooth out fluctuation, enhances comparability, and increases accuracy. Smoothing out fluctuation makes it easier to identify trends and patterns in the data, as outliers and anomalies can be minimized, thus resulting in a more accurate representation of the overall data across different time periods. As a result, for the purposes of our analysis, we estimated four columns of interest and generated eight line graphs using the 7-day average measurement. These columns are the daily total bike trips, the categorization of bike trips taken by subscribers and casual users, and daily total subway trips, calculated by summing the entry count of every turnstile across all stations.

Reason we have chosen 2022 as the “new normal” is that the COVID-19 vaccine is a crucial tool for combating pandemics, yet it was not widely available until late spring and early summer of 2021. As a result, many individuals did not receive their shots until later in the year, while the city was still pushing the vaccine schedule in the summer when the Delta variant surged. In contrast, vaccines and boosters were more widely available and being administered to more individuals in 2022. Although the Omicron variant was still dominant early in the year, vaccination requirements for indoor activities throughout the city were ended in March. Mask

mandates were also lifted in September, which indicated a return to some semblance of normalcy. Therefore, we believe that 2022 is a more appropriate year for comparison and analysis when assessing the pandemic's impact on public transportation usage.

Below is the architecture of the data handling procedures that we applied to each of the datasets:

- *Aggregate data*
    - Since each downloadable Citi Bike trip data only detailing the trip histories for a particular month, data aggregation based on year is needed in order to assemble a combined CSV file that is ready to be processed by PySpark.
    - The turnstile usage data came in handy for 2019 and 2020, as each CSV file already contains the data for the whole year. However, for some unknown reasons, the 2022 data did not come intact, as each CSV file only records a period of data. Therefore, a similar data aggregation procedure, as mentioned above, is needed to combine all separate CSV files into one combined usable file.
- *Clean data*
    - Luckily, the Citi Bike staff have already performed some data cleaning before uploading the CSV files, which includes filtering out trips that were taken by staff for testing purposes and trips that were below 60 seconds. However, we still checked to see if there exists duplicate records in the dataset, as well as filtered out entries containing inappropriate year. Additional columns were also created for summary statistics calculation and graph plotting purposes.
    - The MTA turnstile data was quite difficult to clean, as it often contains inconsistent, duplicate, and outlier values. However, with the help of PySpark's numerous built-in functions, we were eventually able to tackle those issues. Detailed explanation of our cleaning process can be found in the next section.
- *Calculate statistics*
    - At this stage, we used custom PySpark functions to find the summary statistics (mean, max, min, and standard deviation) for the numeric columns in our DataFrames, which are daily total subway trips count, daily total bike trips count, daily bike trips taken by subscribers and casual users, as well as bike trips length.
- *Plot graphs*
    - By using Matplotlib, we created eight line graphs based on the numeric columns mentioned above, in order to demonstrate the transportation usage behavior pattern throughout a year. Key COVID-19 timelines were also incorporated into the 2020 vs. 2022 graphs for additional context support.
- *Explain the graphs*
    - We examined our generated graphs, along with summary statistics, to discover and see the similarity of trends and any sudden change in usage.

In summary, we processed more than 30M rows of MTA turnstile usage data and 1.64M rows of Citi Bike trip data.

Design of Cleaning Process

Data cleaning is an essential step in the process of big data analysis, as it aims to eliminate inconsistencies, errors, missing values, and other discrepancies that can significantly impact the accuracy and reliability of the analysis results. In the case of the Citi Bike dataset, although the data had already undergone preliminary cleaning, it was grouped by month rather than by year per dataset. Therefore, to prepare the data for analysis, we utilized the terminal command *cat *.csv > citibike2019.csv* to aggregate all the CSV files for a particular year into a single combined file, rendering it readily available to be further processed by PySpark.

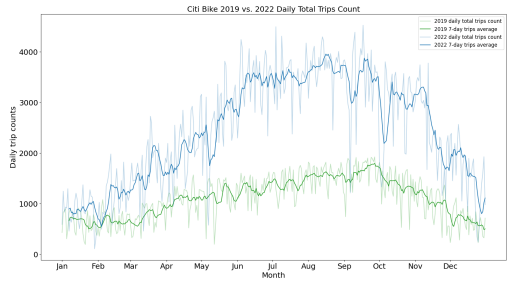
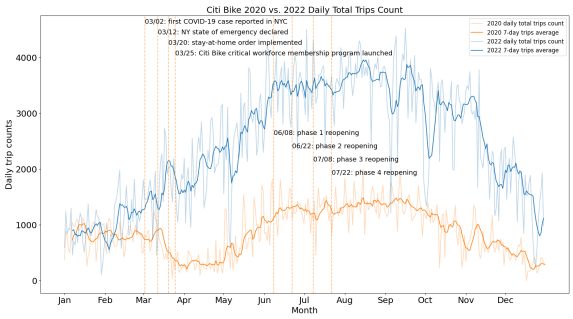
The MTA turnstile data, on the contrary, presented us with significant challenges, as a considerable amount of dirty data was included in the datasets. For example, C/A, Unit, and SCP are the three data columns that can uniquely identify a turnstile across all NYC subway stations when combined. These columns each stand for the control area, remote unit ID of a station, and the subunit/channel/position of a given device. However, the combination of these three columns was insufficient to form a primary key field that satisfied the uniqueness constraint of the dataset. This is due to the presence of six data points per 24 hours, recorded at 4-hour intervals, that record the cumulative entry value since the initialization of a given turnstile. In addition, the cumulative entry value would occasionally reset to zero due to the rollover of the counter, resulting in a memory reset. What's more, the data points timestamps are also not consistent across stations, as the need for staggering prevents flooding the system with audit readings all at once. Therefore, the approach for adding timestamp as the fourth column, along with C/A, Unit, and SCP, to form a primary key field, was not feasible.

To tackle these challenges, we developed a method that begins by filling the dataset with empty values. Since the entry timestamp is not appropriate for selection, we decided to aggregate the entry count by day. We only considered the difference between the maximum and the minimum entry value for a turnstile on a same date, which also takes into account the possibility of random resets causing later entry count to be smaller than the previous count. Moreover, in accordance with the article *Taming the MTA's Unruly Turnstile Data* written by Chris Wong, we learned that the difference between the maximum and minimum entry value shall be limited to 10,000. This limitation is necessary as turnstile readings are similar to odometers, which means they have a count limit. If the numbers exceed this limit, the net values become excessively large and meaningless. Thus, a workaround to avoid this issue is to discard any differences that are beyond 10,000. The author claims that 10,000 is a reasonable threshold, as it equates to 2,500 people per hour or 41 people per minute passing through a turnstile.

However, this method alone is still insufficient to address the problem of missing data points in the last chunk of time, as the 24-hour period is denoted using two different dates. Therefore, to address this underestimation, we decided to fill the last chunk of time using the mean value of the other five chunks. This filling procedure may cause overestimation, as the last data point is typically taken at midnight, when fewer people use public transportation. Nevertheless, we deemed this overestimation acceptable as our objective is to compare the "old normal" and the "new normal," while focusing on the relative behavior of the data rather than its absolute count. Although applying this cleaning approach to the entire dataset would cause overestimation, the relativity between dataset characteristics is preserved, and the margin of error is deemed negligible.

Results - Citi Bike

We generated six line graphs to analyze the Citi Bike trip data, comparing two time categories (2019 vs. 2022 and 2020 vs. 2022) across three data categories: daily total trips count, subscribed user trips, and casual user trips. We analyzed the data from two perspectives: per-day count and 7-day average count. The results are presented in a table format below, with each row representing a data category and each column representing a time comparison. The line color for 2019, 2020, and 2022 is green, orange, and blue, respectively. In addition, to provide better contextual understanding of the 2020 vs. 2022 graphs, we added vertical dashed lines to denote important COVID-19 timelines. Due to spacing, the detailed summary statistics can be found in our GitHub repository.

	2019 vs. 2022	2020 vs. 2022
Daily Total Trips Count		

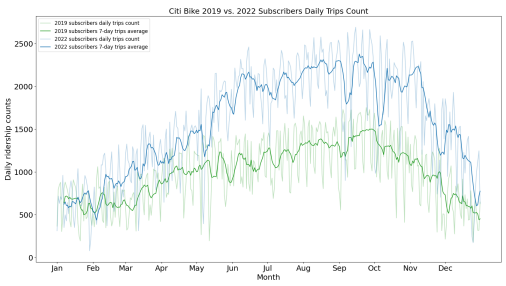
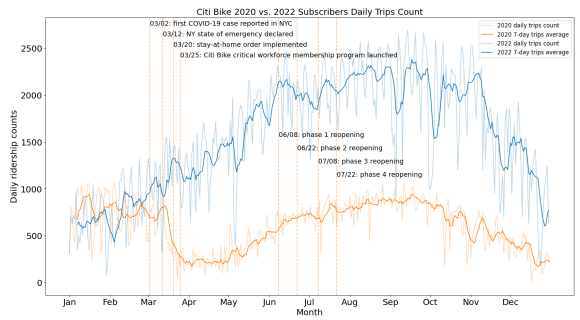
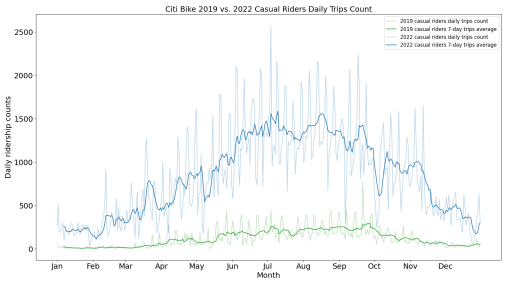
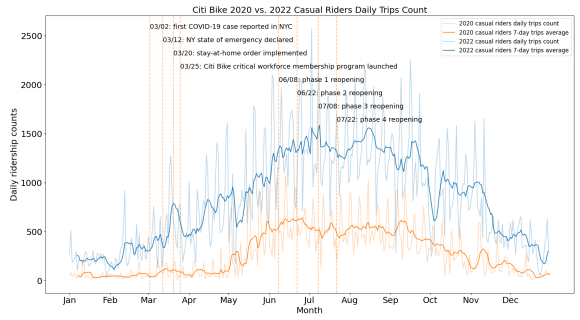
Subscribers Daily Total Trips Count		
Casual Users Daily Total Trips Count		

Figure 1: Citi Bike 2019 vs. 2022 and 2020 vs. 2022 daily trips comparison

Results - Subway

We created two line graphs to visualize the daily entries count of the Metropolitan Transportation Authority (MTA) subway. Figure 2 compares the total daily entries count of 2019 and 2022, demonstrating the difference between the pre-pandemic “old normal” and the post-pandemic “new normal.” Figure 3 presents the total daily entries count of 2020 and 2022, representing the during-pandemic and after-pandemic periods. Similar dashed lines were added for the 2020 vs. 2022 graph in order to provide better contextual information.

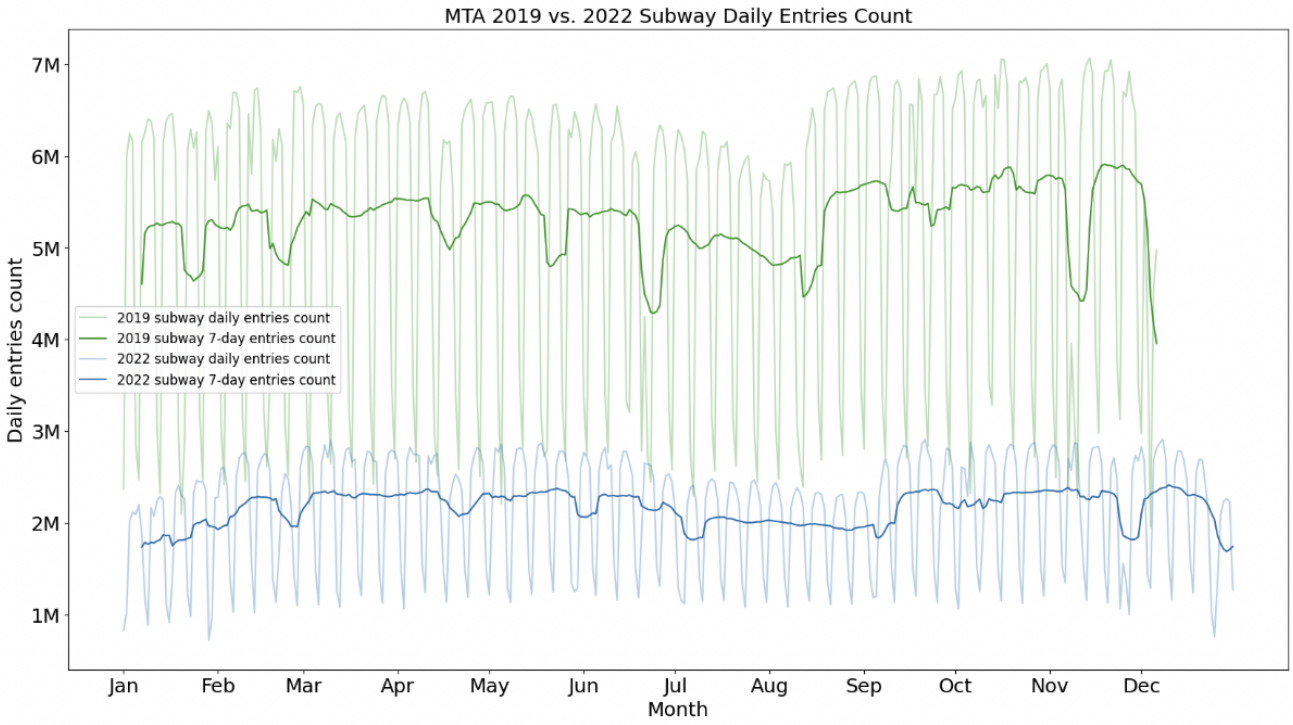


Figure 2: MTA 2019 vs. 2022 Subway Daily Entries Count

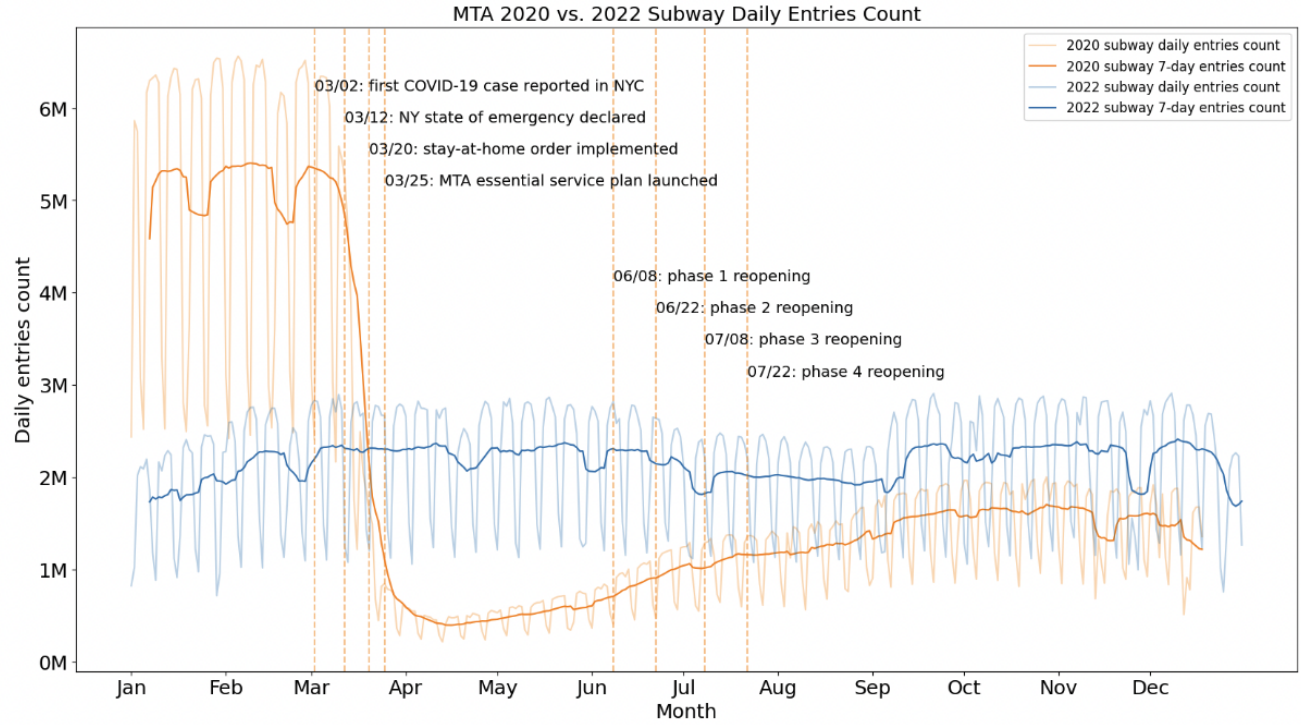


Figure 3: MTA 2020 vs. 2022 Subway Daily Entries Count

	year	total_entries	daily_entries_mean	daily_entries_std_dev	daily_entries_min	daily_entries_max
0	2019	1.798135e+09	5288632.28	1634452.13	1950547.0	7069657.0
1	2020	6.919683e+08	1965819.17	1831803.92	216717.0	6562732.0
2	2022	7.853337e+08	2151599.17	634793.56	716793.0	2911307.0

Figure 4: MTA 2019, 2020, and 2022 subway ridership summary statistics

Observation and Explanation - General Observations

In all eight graphs, we were able to successfully recreate the sudden drop that occurred around March 2020, which coincides with the onset of the COVID-19 pandemic. As expected, both subway and Citi Bike followed a similar pattern, with a sudden drop in ridership around early March that remained relatively stable until slowly recovering starting around early May.

One interesting observation is that both the Citi Bike and MTA subway graphs exhibit random "dents" in the 7-day average measurement throughout the year, which seem to coincide with each other. We believe that these dents may be attributed to bad weather conditions, as we can see a similar pattern when we examine the weather data from World-Weather.info. This suggests that weather conditions play a significant role in transportation usage patterns and should be taken into account when developing transportation policies and strategies.

Observation and Explanation - Citi Bike

All six graphs show a consistent trend in bike usage patterns, with peaks occurring around the middle of the x-axis. This suggests that trip counts are higher during the summer months, which steadily decline as the season approaches winter. This finding is consistent with our understanding of seasonal variations in transportation usage, as people tend to be more active and spend more time outdoors during the warmer months. It also highlights the importance of considering seasonal factors when analyzing bike usage data.

Taking a closer look at Figure 5, there is a visible drop in the daily total trips around the middle of March, which coincides with the declaration of New York State emergency. Looking at the subscribed (Figure 6) and casual trips (Figure 7), we can conclude that the drop in subscribed trips is responsible for the major drop in the daily total trips, this is because subscriber trips exhibit a similar pattern around March, while the casual trips remain relatively flat at the same period. This means that people who use Citi Bike service on a regular basis decided to stop their routine. Additionally, the drop slowed down when Citi Bike announced the critical workforce membership program, which aimed at providing affordable access to bike-sharing for essential workers during the COVID-19 pandemic. The program offered a 50% discount on the annual membership fee to all employees of healthcare facilities, grocery stores, pharmacies, and other essential businesses, but it seemed ineffective in boosting the subscribed ridership, as trips count

remained relatively low until early May, when Governor Andrew Cuomo announced that a four-phase reopening plan for businesses will be coming in early June.

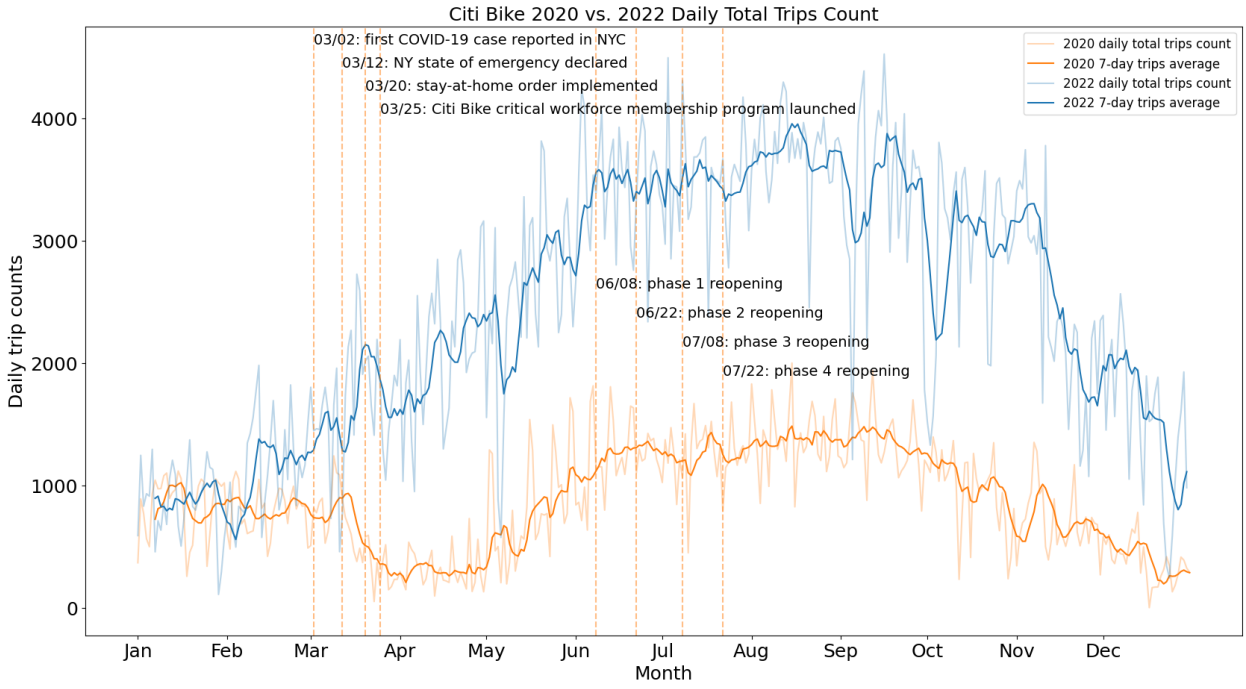


Figure 5: Citi Bike 2020 vs. 2022 Daily Total Trips Count

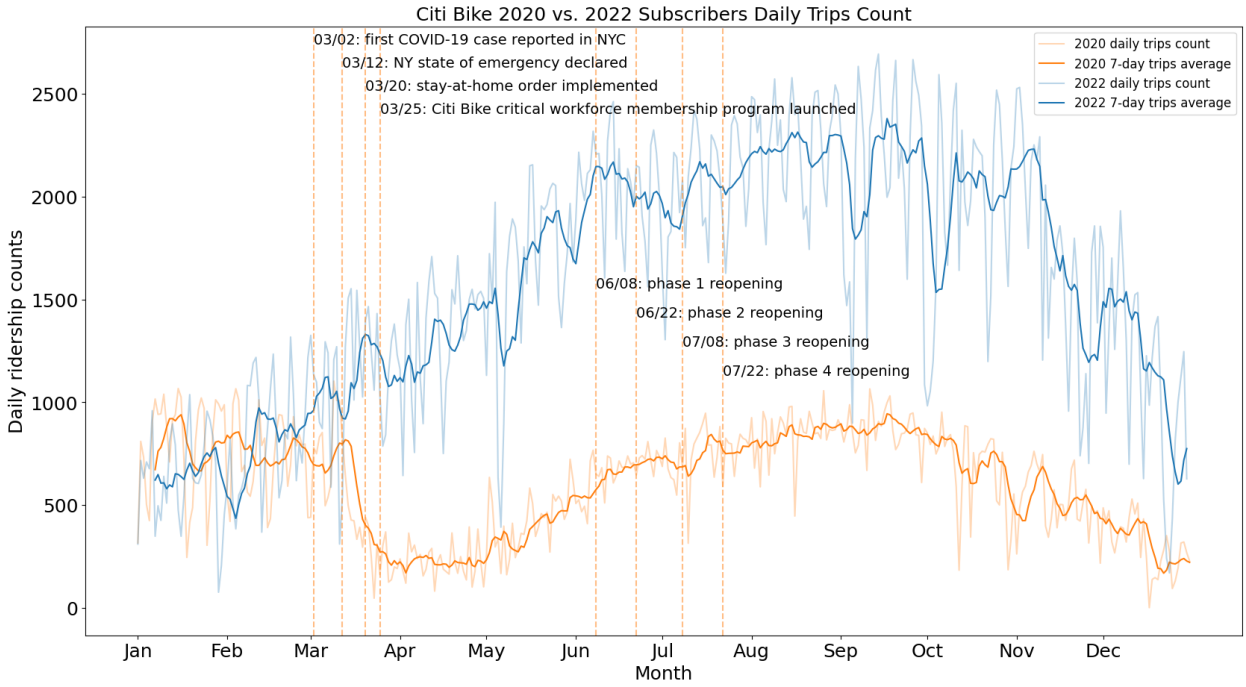


Figure 6: Citi Bike 2020 vs. 2022 Subscribers Daily Trips Count

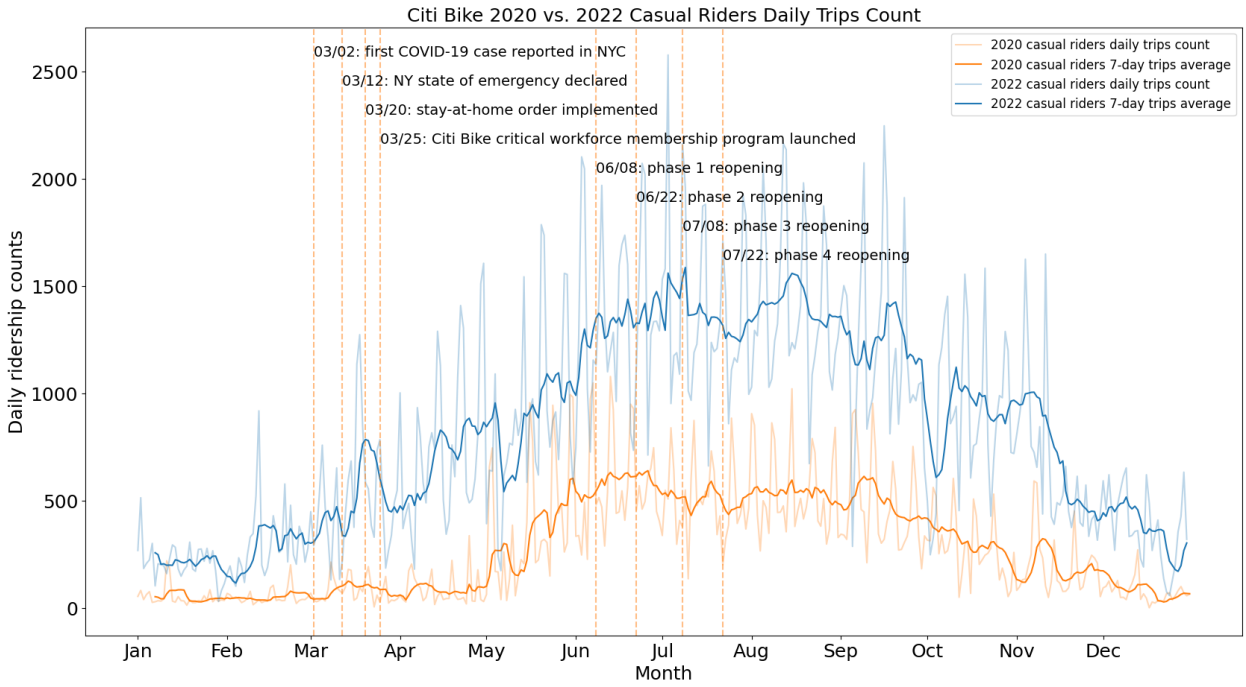


Figure 7: Citi Bike 2020 vs. 2022 Casual Riders Daily Trips Count

To analyze the changes between the "old normal" and the "new normal," we can focus on Figures 8, 9, and 10. It is surprising to note that the total number of daily trips has nearly doubled from 2019 to 2022. This significant increase can be attributed to the rise in casual trips during the summer of 2022. In 2022, the mean count of daily casual trips was 829, which is about seven times higher than that of in 2019, which was only 117.

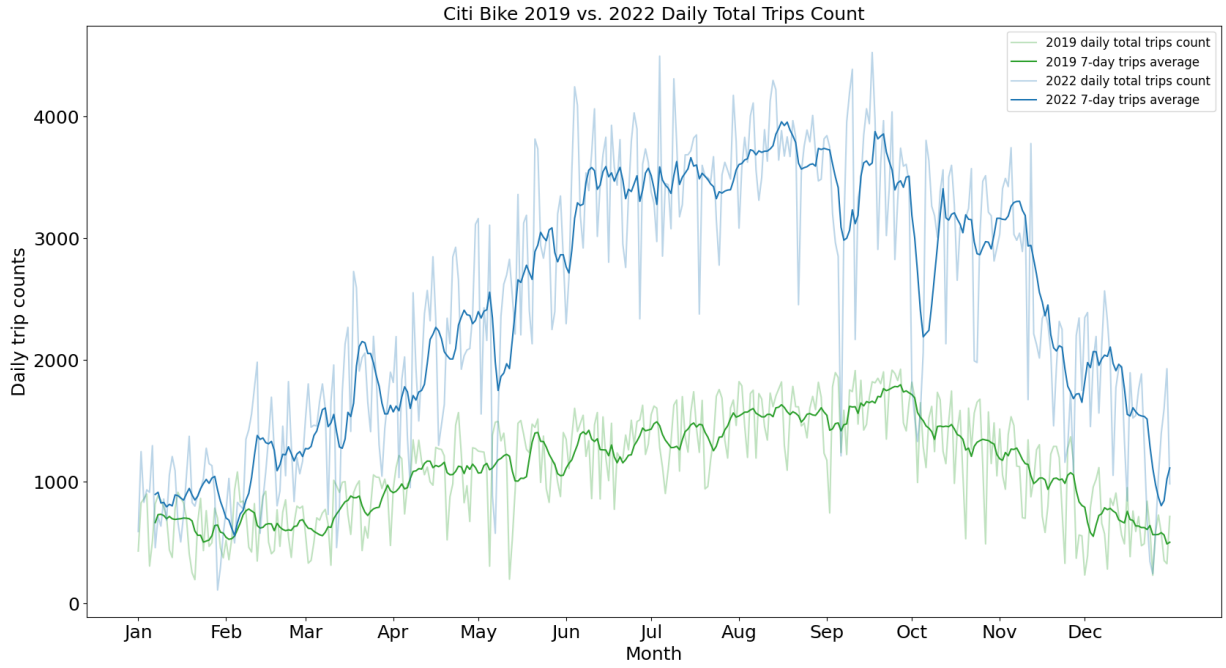


Figure 8: Citi Bike 2019 vs. 2022 Daily Total Trips Count

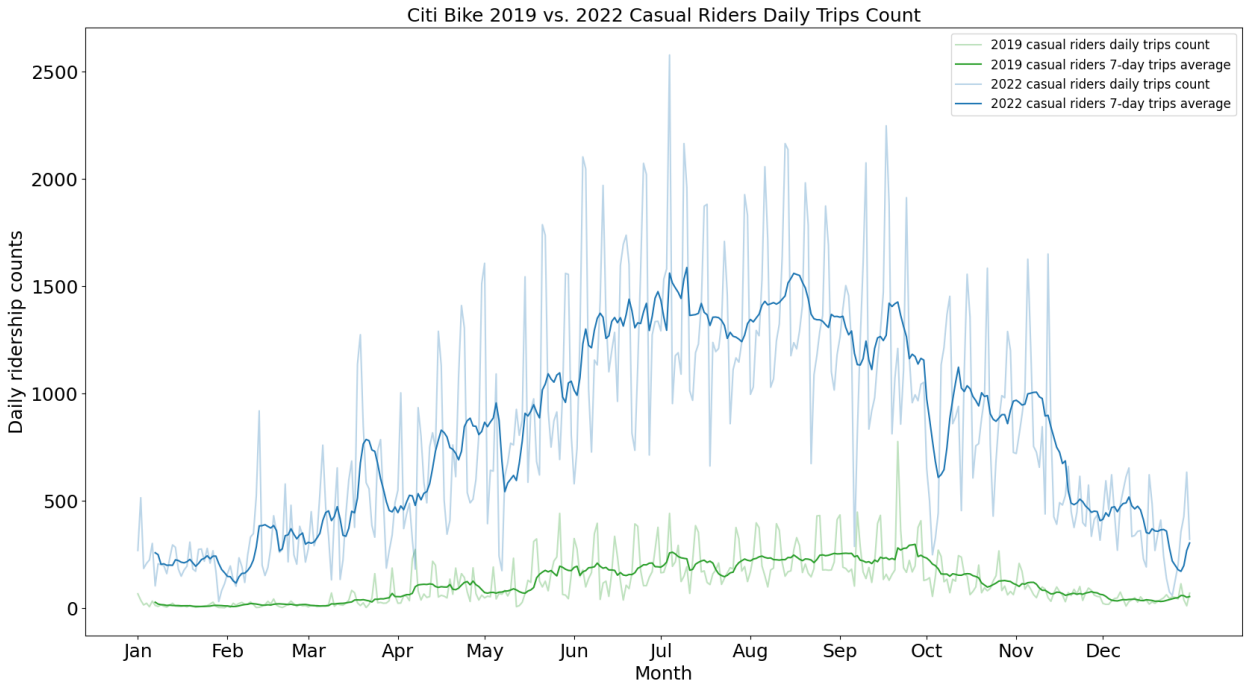


Figure 9: Citi Bike 2019 vs. 2022 Casual Riders Daily Trips Count

In comparison to the casual trips, the trips made by subscribed users have seen a less dramatic difference. The daily "old normal" trips average for subscribed users was 991, whereas it was 1573 for the "new normal," representing around a 50 percent increase. This suggests that the bike program is becoming increasingly popular among casual users, perhaps due to the pandemic introducing them to new possible ways of public transportation or as a form of recreation.

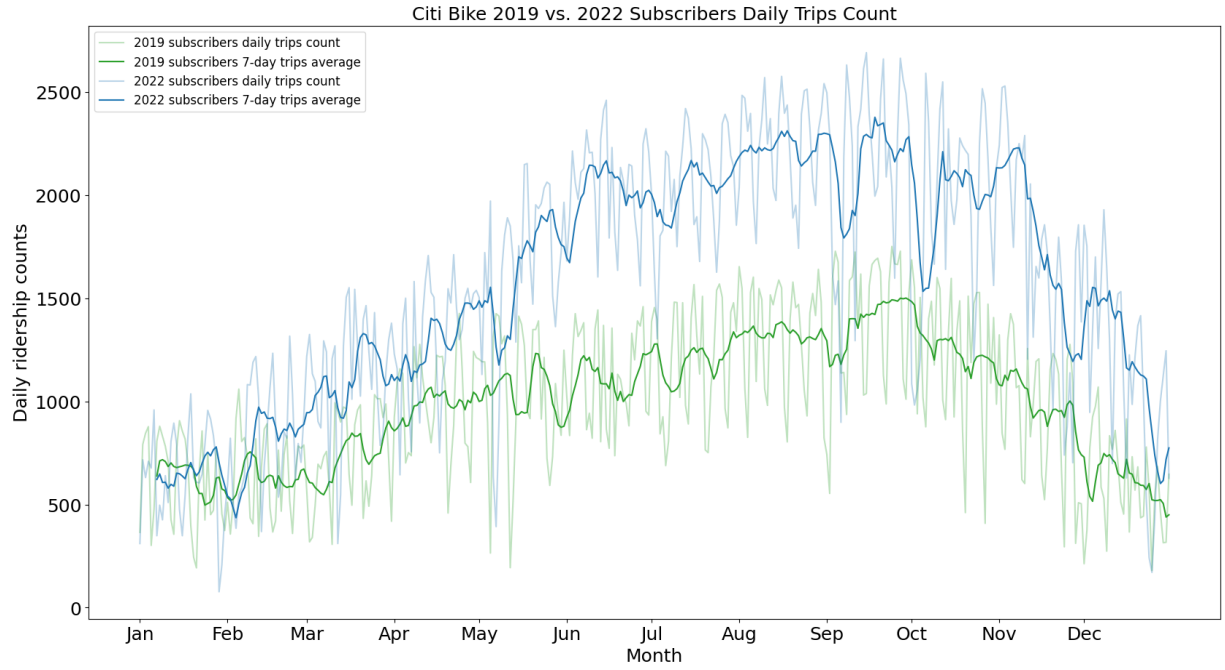


Figure 10: Citi Bike 2019 vs. 2022 Subscribers Daily Trips Count

Overall, our Citi Bike data analysis highlights the value of the bike program as an alternative mode of transportation, particularly for casual users. As we continue to analyze the data, we can gain additional insights into the factors driving these changes and use these insights to inform policy decisions and further improve the bikeshare program.

Observation and Explanation - Subway

The pattern and shape of our graphs closely resemble those created by Wang and Noland in their research paper, indicating that our data cleaning method is indeed effective and non-destructive. However, we did observe a slight overestimation of approximately 0.2 million in comparison to their results. This overestimation was expected and serves as further confirmation of the robustness of our data cleaning process.

In contrast to the increase in Citi Bike ridership, the New York City subway system has experienced a significant decline in daily entries. Between 2019 and 2022, the mean entries count fell from approximately 5.28 million to 2.15 million (Figure 4), a reduction of nearly 60%. This trend can be attributed to the emergence and growth of remote work as a viable and mature working style. According to a 2022 survey conducted by The Partnership for New York City, 26% of its members had returned to the office as of March 2022, with a projected increase to 45% by September 2022 and 54% by the end of the year. Additionally, the survey found that 80% of respondents require their employees to be vaccinated before returning to the office, and 87% plan to adopt a hybrid work model that allows for both in-person and remote work. These results suggest that the return to the office may be slow but is steadily progressing. The shift in work style from traditional office-based work to hybrid and remote work models is likely a contributing factor to the decrease in subway ridership and changes in public transit patterns observed in New York City.

Limitations

One limitation of our study is that we relied solely on qualitative information to explain our results. We did not include external quantitative data to support our observations, which could have strengthened our analysis. For instance, we could have used confirmed positive cases and death numbers to better understand the impact of the pandemic on our findings. Additionally, we could have considered weather data such as temperature, precipitation, and wind speed to further investigate the relationship between the “dent” in the 7-day average and weather patterns. Incorporating quantitative data could have provided a more robust analysis of the factors that influenced our findings, and improved the overall quality of our study, making it more relevant to policymakers and stakeholders. For future research, we should consider including both qualitative and quantitative data to provide a more comprehensive understanding of our research topic.

References

Citigroup Inc. “Citi Bike System Data: Citi Bike NYC.” *Citi Bike: NYC's Official Bike Sharing System*, 2019, https://citibikenyc.com/system-data.

MTA Headquarters, New York City Transit. “Turnstile Usage Data: 2019: State of New York.” *Turnstile Usage Data: 2019 | State of New York*, 30 Dec. 2019, https://data.ny.gov/Transportation/Turnstile-Usage-Data-2019/xfn5-qji9.

MTA Headquarters, New York City Transit. “Turnstile Usage Data: 2020: State of New York.” *Turnstile Usage Data: 2020 | State of New York*, 8 Mar. 2021, https://data.ny.gov/Transportation/Turnstile-Usage-Data-2020/py8k-a8wg.

New York City Department of Health and Mental Hygiene. “COVID-19.” *COVID-19 - NYC Health*, https://www.nyc.gov/site/doh/covid/covid-19-main.page.

Ng, Sunny. “NYC Subway Ridership - Based on Turnstile Usage Data, Updated Weekly.” *NYC Subway Ridership - Based on Turnstile Usage Data, Updated Weekly*, 18 June 2020, https://www.subwayridership.nyc/.

U.S. Department of Health & Human Services. “CDC Museum Covid-19 Timeline.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 15 Mar. 2023, https://www.cdc.gov/museum/timeline/covid19.html.

Wang, Ding, et al. “Impact of COVID-19 Behavioral Inertia on Reopening Strategies for New York City Transit.” *International Journal of Transportation Science and Technology*, Elsevier, 14 Apr. 2021, https://www.sciencedirect.com/science/article/pii/S2046043021000046.

Wang, Haoyun, and Robert B. Noland. “Bikeshare and Subway Ridership Changes during the COVID-19 Pandemic in New York City.” *Transport Policy*, Pergamon, 13 Apr. 2021, https://www.sciencedirect.com/science/article/pii/S0967070X21000974.

Whong, Chris. “Taming the MTA's Unruly Turnstile Data.” *Medium*, Qri.io, 31 Mar. 2020, https://medium.com/qri-io/taming-the-mtas-unruly-turnstile-data-c945f5f96ba0.

Partnership Survey Finds Slow but Steady Increase in Workers Returning to the Office, Topping 50% by the End of the Year. *Partnership for New York City*, 15 Sept. 2022, pfnyc.org/news/partnership-survey-finds-slow-but-steady-increase-in-workers-returning-to-the-office-topping-50-by-the-end-of-the-year

World-Weather.info. “New York October 2022 Weather Forecast.” World-Weather.info, n.d. Web. 7 May 2023. https://world-weather.info/forecast/usa/new_york/october-2022/.