NYC Covid-19 Transportation Analysis Interim Report

Jason Lai, Kewei Zhang

Introduction

Covid-19, unlike all other pandemics in history, shaped our way of life drastically. In the prolonged period of lockdown measures, the society had developed and adjusted to a new lifestyle. In this paper, we examined how lockdown measures affect people's way of life in terms of transportation. To be more specific, we gathered data on public transportation in the New York City area across different years. We included MTA and Citi Bike as measures for public transportation; we included time ranges from pre-covid days(year 2019) to post-covid time(year 2022). We want to find patterns in people's behavior in pre-covid, during-covid, and post-covid period, and we want to see if the "new normal" way of life is any different from the "old normal" way of life.

Related Works

The paper(Wang, 2021) serves as the backbone of our project. It describes the impact of the COVID-19 pandemic on subway and bike usage in New York City. The authors recognized that during the early stages of the pandemic both bikeshare and subway ridership declined significantly, with bikeshare decreasing less rapidly than subway ridership. As the pandemic progressed and lockdown measures were eased, bikeshare ridership began to recover more quickly than subway ridership. The paper also took a unique aspect of bike rides where users are grouped into two categories: subscribed users and casual users, describing the recovery as resilient. It also addresses the changes in combination with several factors, such as demographic and socioeconomic factors, weather conditions, and the availability of outdoor spaces for recreation and exercise. The paper leaves with a note that the potential for alternative modes of transportation such as bikeshare to play a more prominent role in urban mobility in the future.

In addition, we referred to the CDC's COVID-19 timeline as well as the NYC Health department confirmed cases to address the context of the rider data. The timeline records major events and developments of the pandemic, including announcement of on pause of non-essential workers, the vaccine made available to common public, new variant trending, and so on, while confirmed cases number provide a quantity measurement of perceived severeness of pandemic by the general public. Since one of the authors also lived through the pandemic in the NYC area, his observations would also take into consideration when explaining the data.

Problem Statement

We will augment the findings in Wang's research paper and extend aspects that the paper did not cover. For example, the paper specifically mentioned that "it is unknown whether these changes will be sustained post-COVID". Thus, considering the publication year of the paper, we will extend the data from the 2019-2020, which was presented in the paper, to the 2019-2022 comparison and 2020-2022 comparison. We want to see if such a trend continues to exist in "post-covid" time: Do people revert to old normal usage of public transportation after the pandemic, or people are still using public transportation after the pandemic just as during the pandemic? How do subscribed users and casual users grow or decline in the new normal? How about all these questions in terms of subway usage? In this paper we will try to answer these questions.

Approach

Two primary datasets we will be using throughout our study are 2019, 2020, and 2022's MTA turnstile usage and Citi Bike trip data. Although not being explicitly explained in the paper, the methodology and procedures we are going to carry out will be similar to the ones conducted by Wang, Haoyun, *et al* in their original research. Four models and several line graphs will be estimated and created for data analysis purposes. These include daily total subway trips, which were measured by summing every turnstile entry count, daily total bike trips, bike trips taken by subscribers, and bike trips taken by casual users.

MTA turnstile usage data was obtained from New York state open data, while Citi Bike trip data was downloaded from Citi Bike NYC's official website. Below are the data handling procedures that will be applied to each of our acquired datasets:

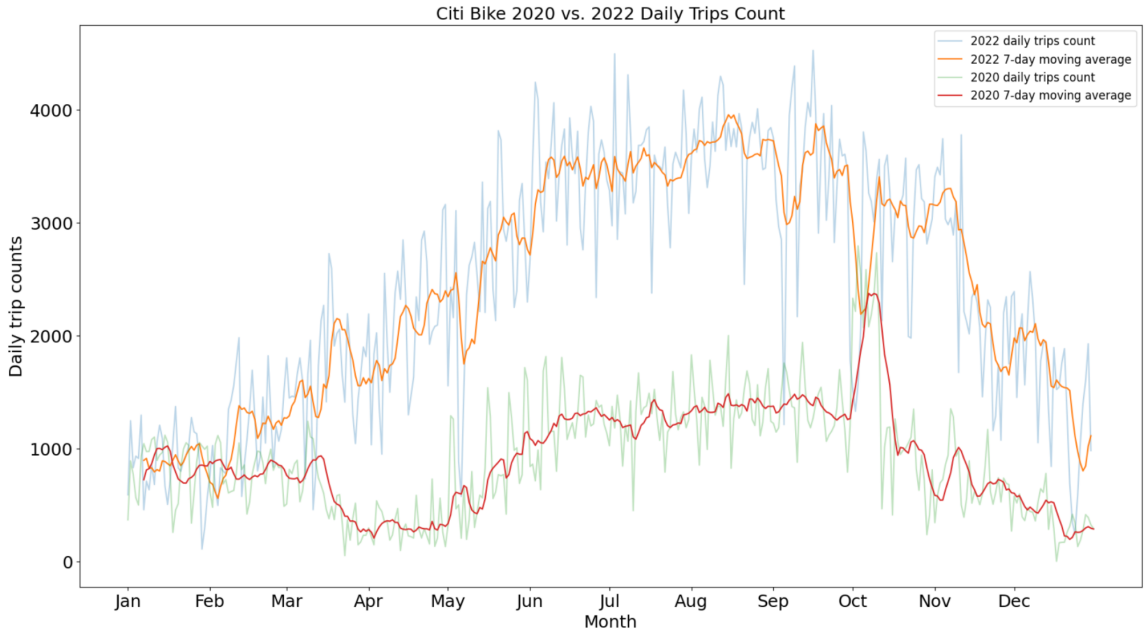
- *Aggregate data*
    - Since each downloadable Citi Bike trip data only detailing the trip histories for a particular month, data aggregation based on year is needed in order to assemble a combined csv file that is ready to be processed by PySpark
    - The turnstile usage data came in handy for 2019 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, 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 those data files, which includes filtering out trips that were taken by staff for testing purposes and trips that were below 60 seconds. However, we will still be checking to see if there exists duplicate records in the dataset, as well as

eliminating any entries that contain null values. Additional columns will also be created for calculation and plotting purposes
    - The MTA turnstile data was quite a mess for us 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 down those issues
- *Calculate statistics*
    - At this stage, we will be using PySpark's summary() function to find the summary statistics (mean, max, min, and standard deviation) for the numeric columns in our DataFrames, which include 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 will create several line graphs (example shown in the preliminary results section) based on those numeric columns mentioned above
- *Explain the graphs*
    - We will examine the generated line graphs, along with our summary statistics, to discover and see if the similarity of trends and any sudden change in usage. Important COVID-19 timeline and dates will also be incorporated into our reasoning for context support

In total, we will be processing more than 30M rows of MTA turnstile usage data and 1.64M rows of Citi Bike's trips data.

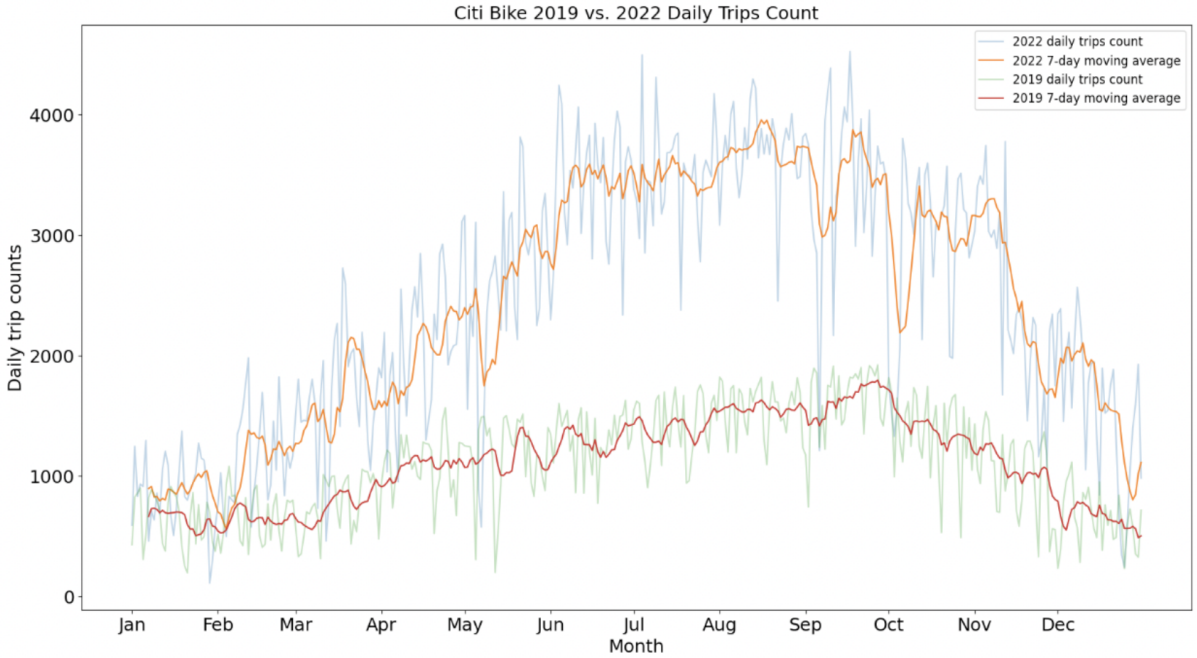
Preliminary Results

So far, six graphs have been generated for Citi Bike trip data. We have two time-category comparisons, 2019-2022 and 2020-2022, crossing three data-categories, daily trips count, subscribed user trips, and casual user trips, with two views: total count and 7-day average count. Below is an example of how one of the comparisons (2020-2022 daily trips count) looks like:



As shown in the graph, we successfully recreated the sudden drop around March 2020, which is the indication of the starting of Covid-19, and our generated pattern also closely resembles the ones created by Wang and Noland in their research paper.

Below is another example of the comparison: 2019-2022 daily trips count



We see a similar trend in bike usage patterns where there are slightly higher number of trip counts in the summertime and diminish quickly as the season approaches winter. Also, it's

surprising to see that the total number of daily trips has grown nearly double in two years, even with the pandemic.

In fact, we've spent most of our time focusing on data gathering, data cleaning, and selecting methodology. Thus, although the results seem limited in the current state, it would surely become promising as our research progresses.

What Is Missing

At the point of writing, the data handling procedures, as explained in the Approach section, are finished for all Citi Bike trips data, yet more work is needed to do for the MTA turnstile data as we've just finished the data cleaning step yesterday. Code-wise, more detailed comments will be added to each cell in the jupyter notebook, so for those who will be interested in reproducing our work, they will have a clear idea of what each cell does. We will also refractor and reorganize our code to make sure it is easy to read and understand.

Moreover, even though graphs have all been generated for Citi Bike trip data, including daily trips count graphs, as well as subgraphs for daily trips count by user type, we haven't yet explained those graphs in the context of events timeline. For example, in the *Citi Bike 2020 vs. 2022 Daily Trips Count* graph, there is a sudden drop for the 2022 7-day moving average around October, while there also exists a sudden spike for the 2020 7-day moving average at around the same period. Therefore, without having those timelines in hand, it would be difficult for us to explain such observations.

Last but not least, we rarely considered the limitations of our paper but one: the weather data. It's a potentially crucial factor that affects Citi Bike usage. It might also explain the sudden drop in October 2022.

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.