Metis Project 1: MTA Data Analysis

Matthew Kwee
July 2021

Introduction

Motivation:

Bower LLC wants to open new retail stores across New York City, and is looking for locations with high foot traffic, preferably places where foot traffic is rapidly increasing as COVID lockdowns end.

Objectives:

Determine the top stations in terms of overall foot traffic.

Of the stations with the highest ridership, find the locations that have recovered the fastest in terms of foot traffic.

Additionally, find locations that have the highest growth rate in terms of foot traffic.

Goals:

Present analysis in an understandable, logical way.

Methodology

Data:

I acquired my data by scraping MTA turnstile data from the MTA website with the 'get_mta.py' provided in the course's resources.

Tools utilized:

I used SQLAlchemy to ingest data into Python Pandas, which in turn was used to clean and organize the data. Finally, I used Matplotlib to present the data in an understandable format, and used a Google Maps API to let Matplotlib plot the data on a map of New York City.

Metrics:

While a station's raw ridership is definitely a factor to be considered, so is its "recovery rate," which indicates how many riders have resumed using the station.

Assumptions

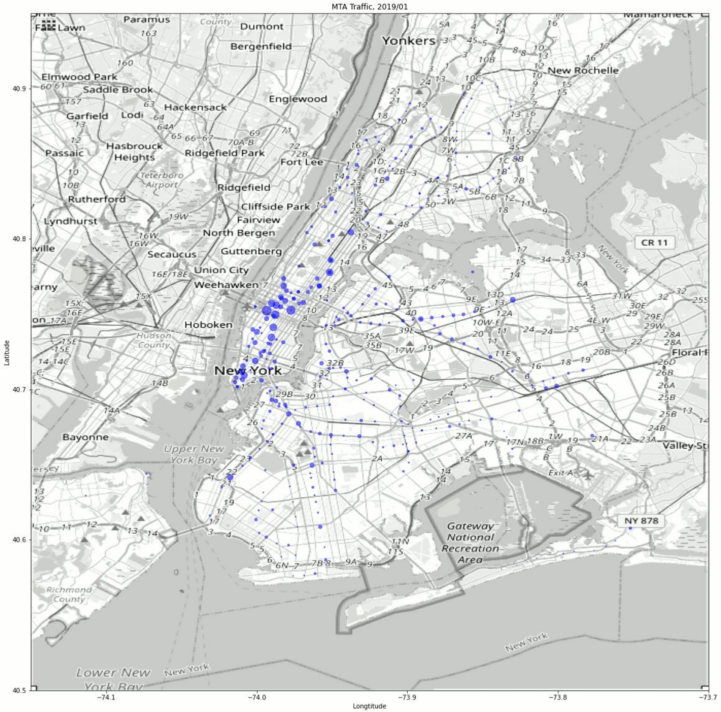
1. Turnstiles are functional.

2. Riders don’t jump the turnstile in order to ride without paying.

3. On average, riders tend to return home the way they came.

Effects of Lockdown

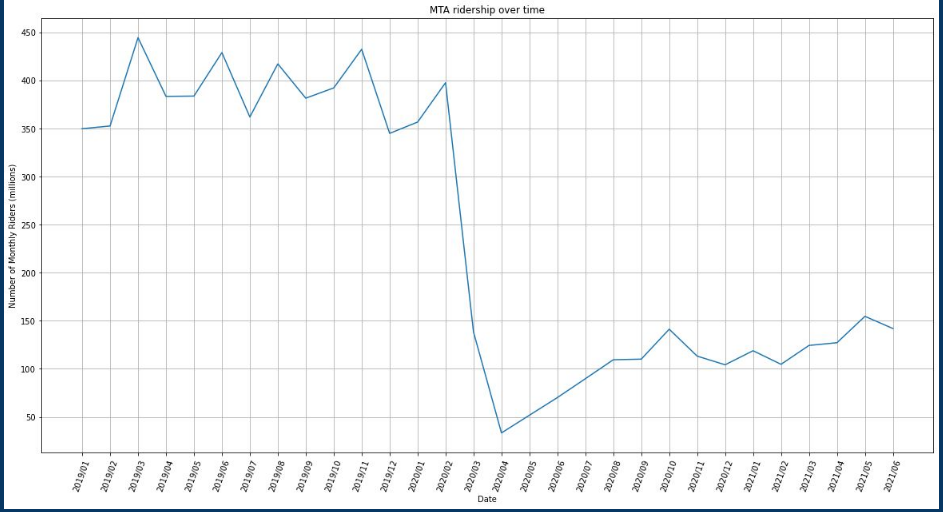
The graphic on the right shows MTA ridership across New York City. MTA ridership disappears between February and April 2020 as COVID lockdowns begin.



The Baseline

Total monthly MTA ridership dropped off dramatically during the COVID-19 lockdowns, but has increased to about 30% of its pre-COVID value.

We'll use this percentage as baseline for determining whether a station's "recovery rate" is "good" or "bad."

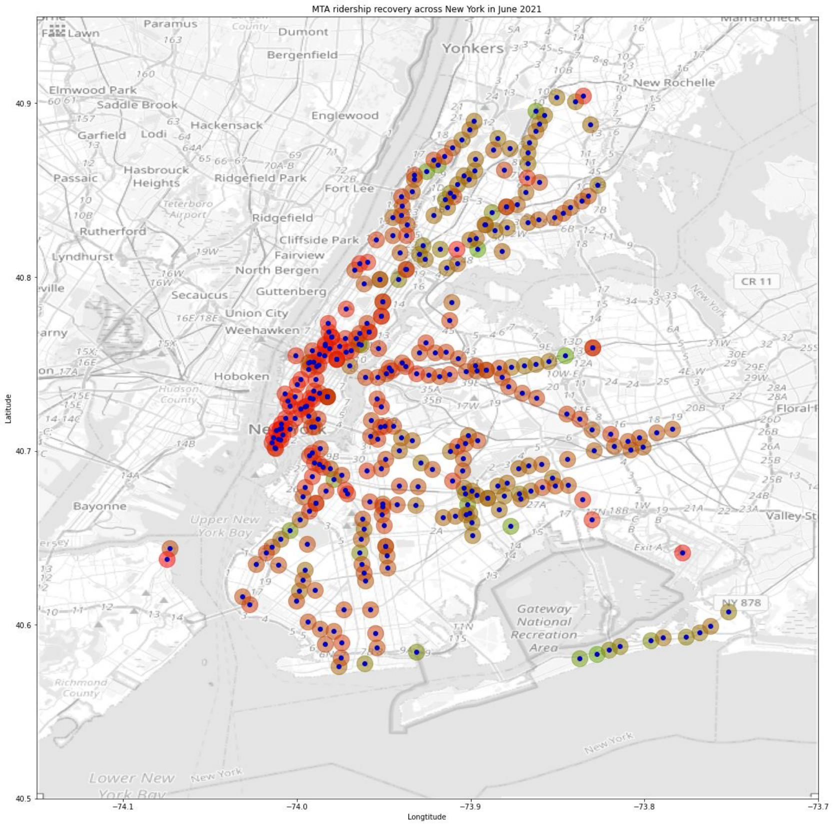


Recovery

A station’s “recovery rate” is calculated by dividing the current number of riders by the number of riders it had before the COVID lockdowns.

Recovery Rate = #Current Riders ÷ #Riders before COVID-19

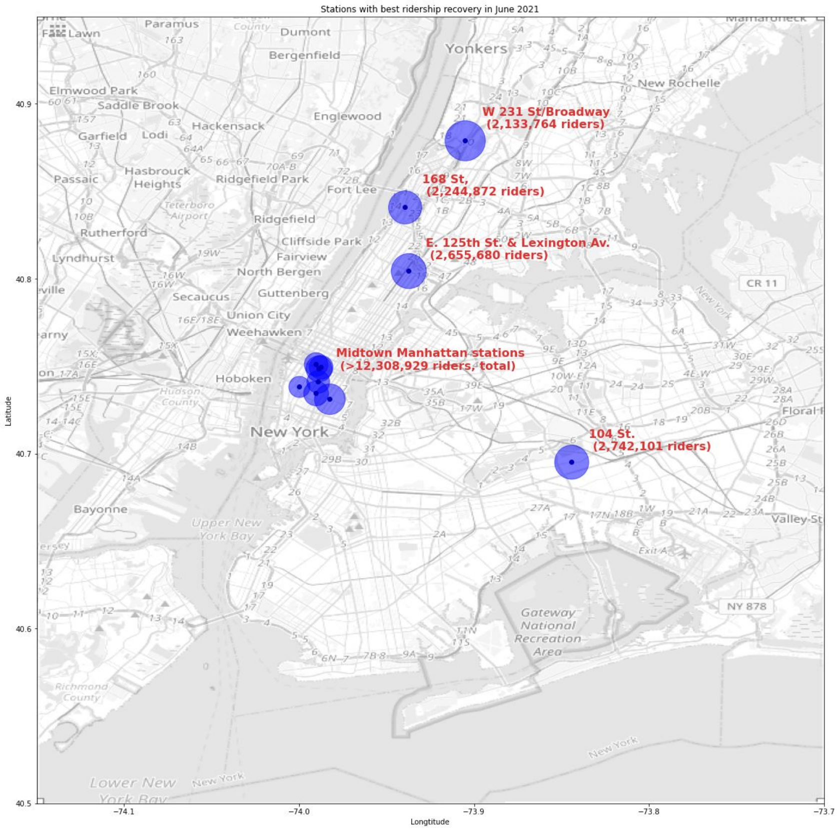
The image on the right shows the “recovery rate” for each station (excluding outliers). Stations surrounded by red have proportionally few riders, while stations outlined by green are returning to full ridership.



Top stations (overall ridership)

The majority of high-traffic MTA stations are in Manhattan Island.

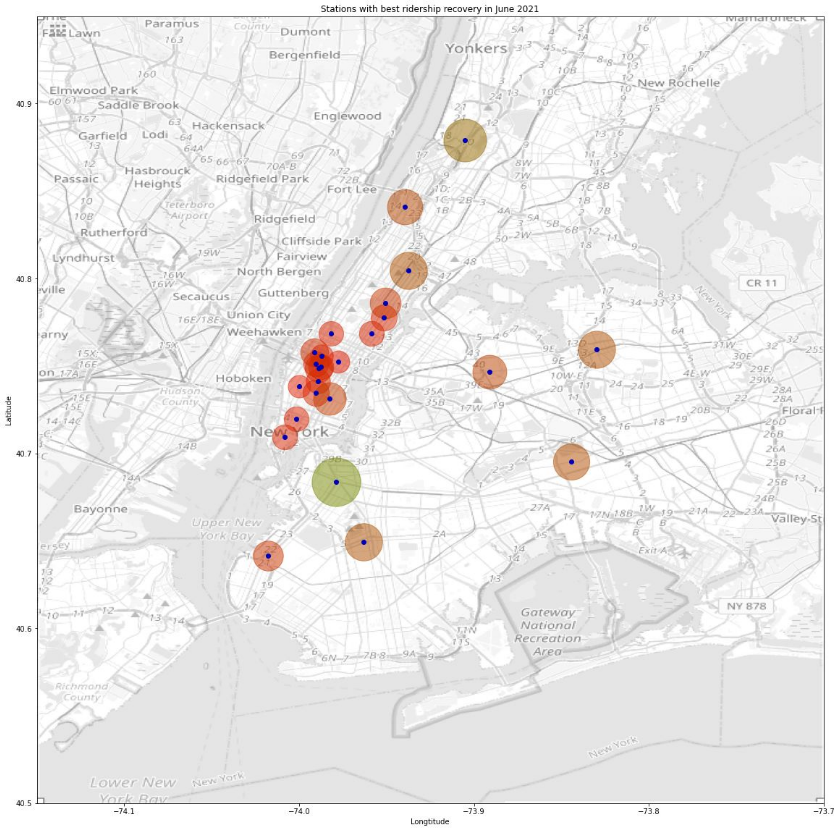
However, one of the top 5 stations in terms of ridership is located in Queens, NY.



Top 25 stations (Recovery)

While Downtown and Midtown Manhattan may have a monopoly on overall foot traffic, they have fairly poor recovery rates compared to the rest of New York City.

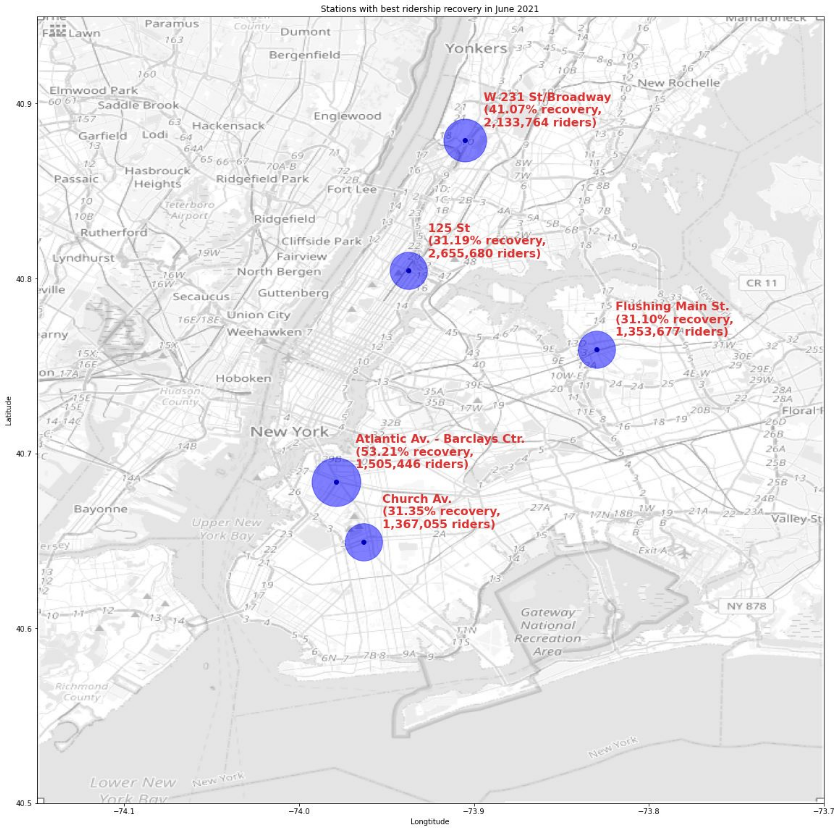
Other boroughs of New York seem to be reopening significantly faster, however.



Top 5 stations (Recovery)

Out of the 25 stations that see the most foot traffic, I highlighted the five with the highest recovery rates.

These may be good places to open a retail store; they have good foot traffic and have recovered as quickly or faster than the average station.



Conclusions

MTA stations in Manhattan see large amounts of foot traffic, but have reopened at a slower rate than other boroughs of New York.

Stations in Brooklyn, Queens, and the Bronx are returning to pre-COVID levels of traffic significantly faster, and should be considered when searching for potential retail store locations.

Future Work

I initially began to build a basic web scraper that would collect all retail listings on websites I linked it to, but I was unable to obtain enough listings in a reasonable time period.

The next step after indexing a sufficient number of listings would be to geocode each listing's address, and assign each listing an MTA station. From there, I would calculate the average price per sqft per unit of foot traffic.

Appendix

Google Maps Geolocation API:

https://console.cloud.google.com/marketplace/product/google/geolocation.googleapis.com