Vijay Tripathi

iViz 2

Introduction

The New York City subway system is the largest subway system in the world by number of stations; over 460. What began as three independent private companies in the early 1900s moves over 2 million people per day. All kinds of New Yorkers use the subway to commute across the city; Brooklyn residents near Coney Island to Wall St bankers and executives. Despite it's vastness, it is the least on-time system in the world. Only 76.4% of trains operate on schedule, compared to a global average of over 90%[1]. It has been estimated that train schedule deviation costs New Yorkers nearly 1 million hours of lost time; almost 35,000 each day[2]. The MTA (Metropolitan Transporation Authority) attributes most of these tardies due to aged traffic control and signaling systems that regularly fail[3]. The pace of construction and renovation for MTA systems is known for being notoriously slow and over budget. Thankfully, the MTA is working towards transparency and posts much of its data on New York State's OpenData portal[4].

Objective

The scope of this visual analysis is aimed towards ridership analytics. Although the MTA produces their own metrics for passengers, consider this visualization the beginning of many analyses of the New York subway; both visually and through machine learning. OpenData NY has many datasets concerning train delays and other performance data; this will be used in future analyses.

The objective of this visualization is to view ridership analytics of every station.

---

[1] Griffin, Anthony. "Clearing Clearing Up Misconceptions about Train Punctuality", *Kokoro Media*, 12 Jun. 2020, https://kokoro-jp.com/unfiltered/103/

[2] Walker, Ameena (2017). "NYC commuters collectively spend 35,000 hours each day waiting on delayed trains", *Curbed New York,* 25 Oct. 2017, https://ny.curbed.com/2017/10/25/16548750/mta-subway-delays-average-wait-time-report

[3] Walker, Ameena (2017). "MTA data reveals the most delayed subway lines", *Curbed New York,* 17 Jun. 2017, https://ny.curbed.com/2017/6/17/15824804/mta-most-delayed-subway-lines-data-nyc

[4] "Agencies & Authorities > Metropolitan Transportation Authority", *New York State,* https://data.ny.gov/browse?Dataset-Information_Agency=Metropolitan%20Transportation%20Authority&limitTo=datasets

Dataset

The first dataset is entrance and exit geospatial data for every station. This table includes the lines available at every station, ADA accessibility, division, staffing, cross streets, and crossover (transfer) ability. This was primarily used to get geographic information for every station. It's shape is 1869 x 32.
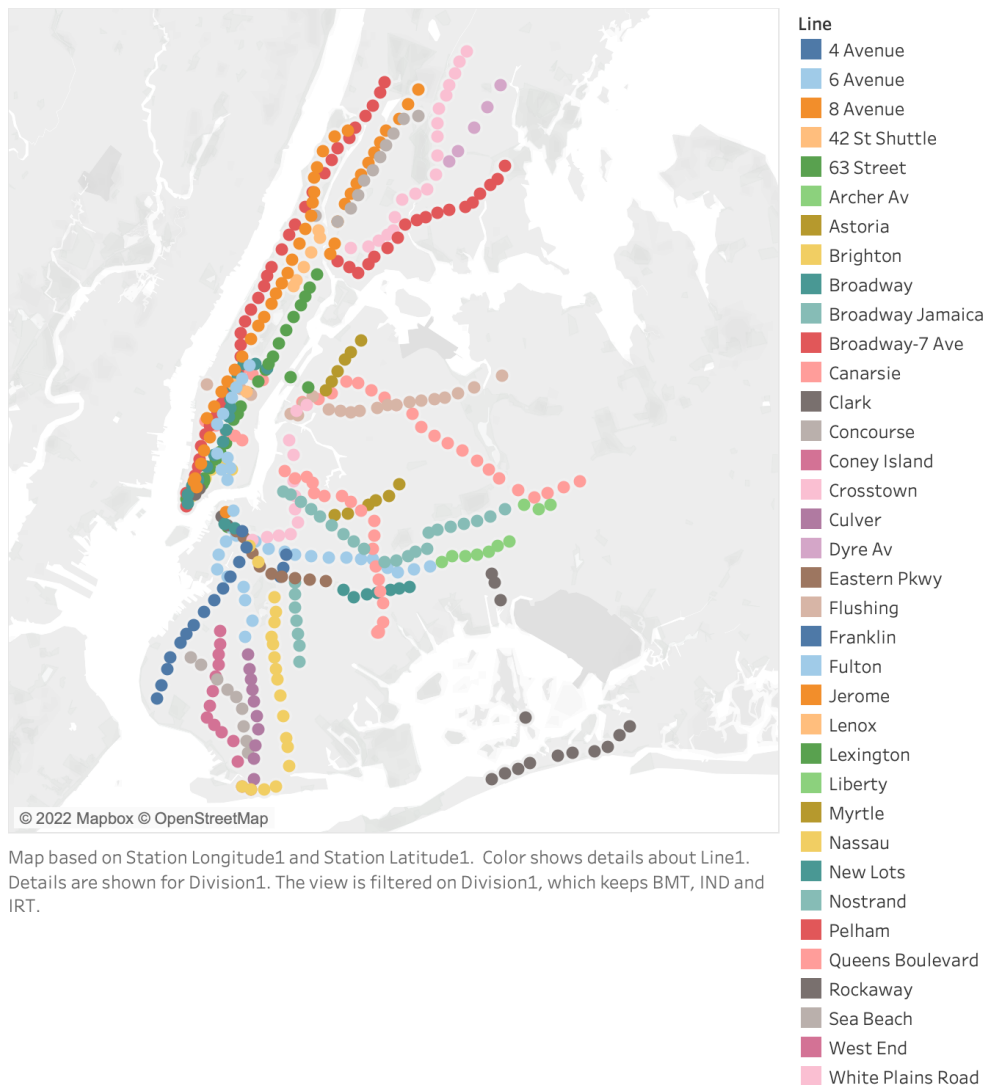
The second table is turnstile information for each station; number of entries and exits, aggregated every 4 hours. Eight of these are used, one for each year from 2014-2021. Each table's shape is 60,000+ x 9 and has a file size of at least 1 GB, which added significant time importing data and making worksheet changes (even using an extract). Due to large file size, I had the chance to learn about Tableau's hyperd, the back-end mechanics of data importing.

Every turnstile year was unioned together, and this union was related to the entrance/exit dataset. The station naming scheme was different for each dataset. The turnstile dataset named each station based off of how MetroCards and bank payment cards present the transaction on their ledgers while the entrance/exit used the name presented on station signage. Even after using Data Interpreter, this required lengthy preprocessing.

Visualization Plan

The MTA published an official subway map in 1979, and maps since then have been iterations on that design. Presenting rapid transit maps accurately but in an easily digestible form is a topic of debate among global transit experts. A map was constructed in Tableau where each station is represented as a dot.

MTA Stations

Map based on Station Longitude1 and Station Latitude1. Color shows details about Line1.
Details are shown for Division1. The view is filtered on Division1, which keeps BMT, IND and
IRT.

Line
- 4 Avenue
- 6 Avenue
- 8 Avenue
- 42 St Shuttle
- 63 Street
- Archer Av
- Astoria
- Brighton
- Broadway
- Broadway Jamaica
- Broadway-7 Ave
- Canarsie
- Clark
- Concourse
- Coney Island
- Crosstown
- Culver
- Dyre Av
- Eastern Pkwy
- Flushing
- Franklin
- Fulton
- Jerome
- Lenox
- Lexington
- Liberty
- Myrtle
- Nassau
- New Lots
- Nostrand
- Pelham
- Queens Boulevard
- Rockaway
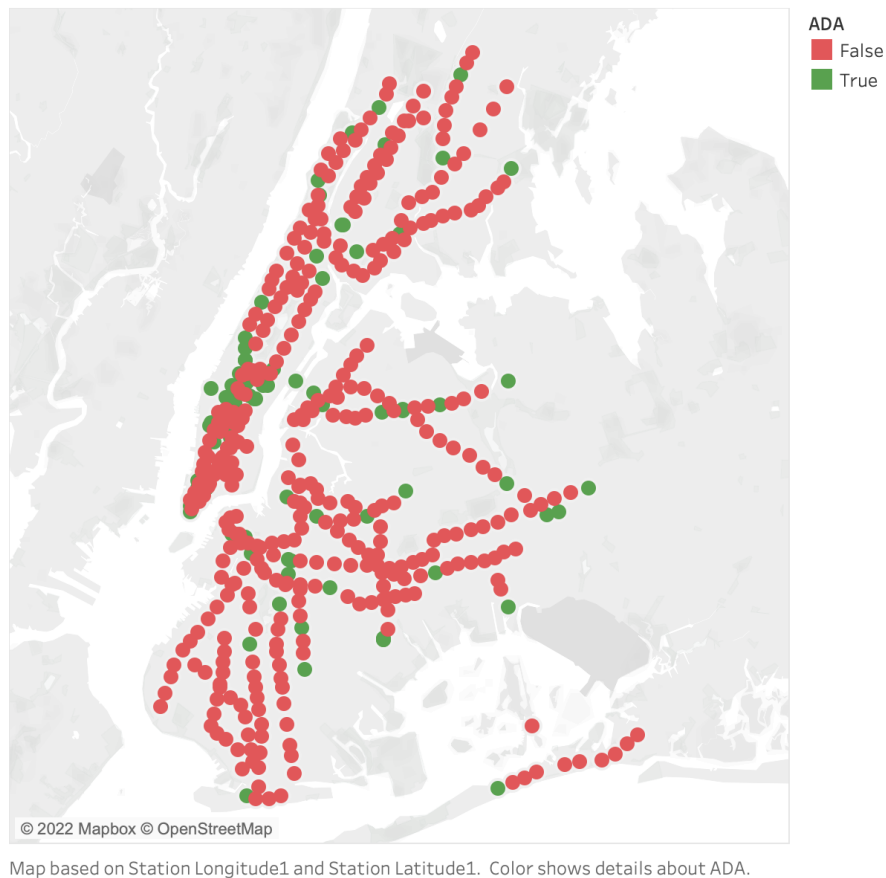- Sea Beach
- West End
- White Plains Road

Each line is indicated by a color. Although there are a lot of colors, each line can still be differentiated. Significant overlapping occurs in downtown Manhattan, so zooming in to individual stations would be necessary. The size of the dots were also tuned to manage overlapping, especially throughout Manhattan. A line graph was attempted to connect the dots, but this resulted in jagged lines that deviated from the path; this data is not suitable for the line graph in Tableau.

Next, each station's ADA compliance was examined. ADA stations are equipped with an elevator/escalator for station entry and an "autogate"[5], a retrofitted emergency exit gate that opens upon the swipe of a "Reduced-Fare AutoGate MetroCard". It is well known that stations aren't accessible friendly, but the degree of such was underestimated.

[5] "Accessible by Subway", *MTA,* https://new.mta.info/accessibility/travel/subway

ADA Compliant Stations

Map based on Station Longitude1 and Station Latitude1. Color shows details about ADA.

Of the 472 service stations, only 90 are accessible, with most of them being in mid-lower Manhattan. Arguably, residents of the outer boroughs are the ones who would most benefit from ADA services, but negligence of the outer boroughs is an underlying theme of the MTA. Since publishing the entrance/exit dataset, which contains station ADA compliance, the MTA has made 40 additional stations accessible.

Next, average yearly ridership was plotted per station. A density graph was used in combination with color to show stations that have the most entries through the turnstile. Green stations have lower average yearly entires compared to red. Years were used in pages for the animation.

Avg Entries Density -
Turnstile_Usage_Data__2015.csv
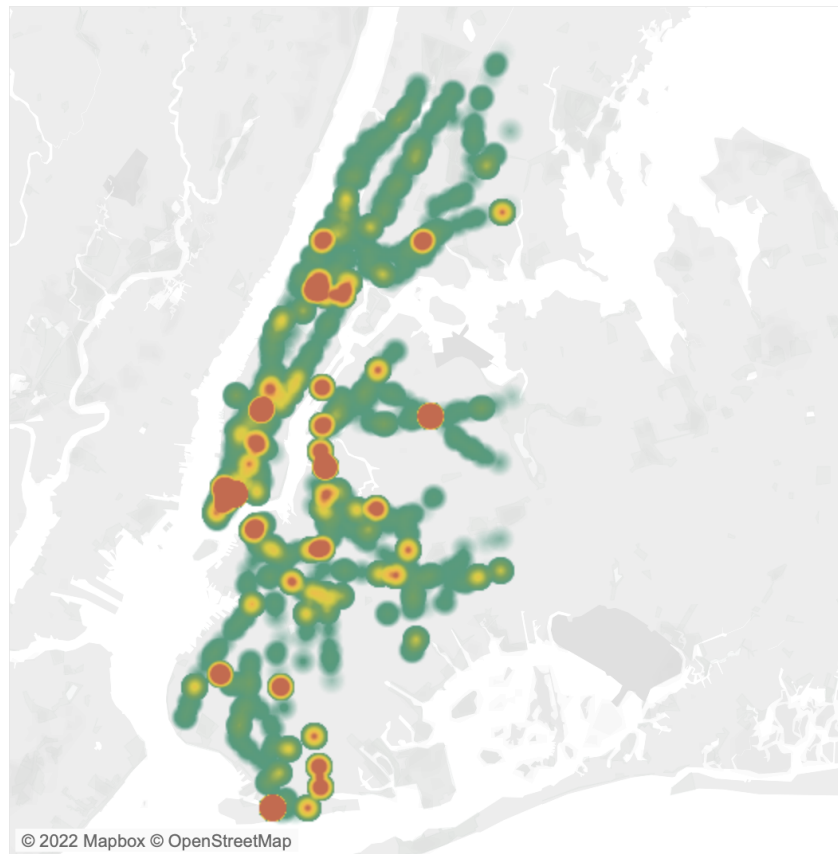
Map based on Station Longitude1 and Station Latitude1. Color shows average of Entries. The
view is filtered on Station Longitude1, which keeps non-Null values only.

This specific visualization is for 2015. The initial goal was to plot every 4-hour
aggregation period, but this was not possible due to processing power limitations and
there was little difference in day-to-day entries. This visualization was attempted to be
published on Tableau Public but the turnstile data alone exceeds the 10 GB storage
limit. Interestingly, uptown Manhattan and the Bronx saw a steady increase in average
entries, with the White Plains Road line have a similar appearance to mid-town.
Dumbo and the surrounding areas saw similar growth. The Stuyvesant and Medical
City neighborhoods saw a sharp decrease after 2018, as well as Meatpacking,
Chelsea, and the West Village. Why is the area between midtown and downtown
Manhattan experiencing lower average yearly entries?

Finally, a bar graph was created to illustrate monthly year-over-year change in entries. Color was used to supplement the percent difference and years were again used in the animation pages.
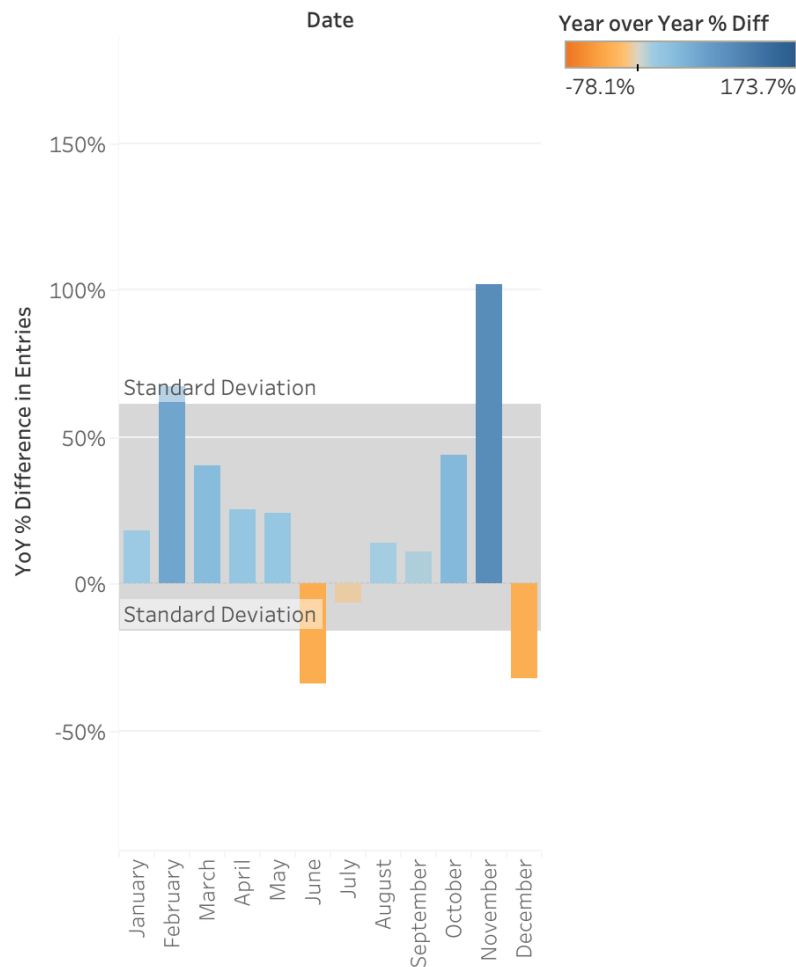
The first major COVID lockdown is clearly visible in 2020's YoY graph. Overall, 2020 ridership was 8% of 2019. In fact, ridership took longer to reach "pre-pandemic levels" than the MTA expected and a task force was established. 2017 experienced little deviation from 2016 ridership. Other than these insights, all other monthly YoY information have no clear explanation. No month experiences consistent increased growth or decline for any year. Each month fluctuates its average entries every year and does not seem to be isolated to a season/quarter. Even holiday entries seems to grow and shrink outside of 2020.