AMS 561 Final Project

Sophia Nolas

May 2 2022

1 Project Objectives

In designing this project, I was very curious to see what I could learn about trends in subway ridership in NYC. I was able to find data from 2010-2020, and 2020-2021 on the New York State government website, where they have all kinds of data available to the public.

Initially, I wanted to investigate trends in ridership as it relates to location, demographics, accessibility, and maintenence. Of course, this was a very broad set of questions. And just having the data does not mean it is in a form that is conducive to drawing conclusions. So over the course of the project, I changed my perspective and narrowed my focus. Instead of looking at each area and station in particular, I looked at the system usage as a whole. Instead of looking at demographics of riders (which is not directly available in the dataset), I ended up doing some historical research to see what key dates in the city might correspond to dates of high or low ridership.

I also wanted to see what impact, if any, the introduction of the ridesharing apps Uber and Lyft, as well as the bike renting service Citibike, had on subway usage. I was able to find the dates of introduction of these services, but especially since they were so close to the earliest dates of the dataset, I am not sure whether my hypothesis that these would have an impact was able to be bourne out by analysis.

I was able to find the data, wrangle and sort it into a form that focused in on what I wanted while clearing out any extraneous pieces, represent it graphically in figures, and learn how to design figures that are effective at communicating the information that is in them.

2 Techniques and Tools

The first step was to find the data. There is an unexpectedly vast amount of publicly available data from the New York State government website, so I turned to two datasets in particular. The two which had the broad subway ridership data I was looking for were: ”Fare Card History for Metropolitan Transportation Authority (MTA): Beginning 2010” and ”MTA Daily Ridership Data: Beginning 2020.” In this way I was able to see trends both before and during the COVID-19 pandemic.

During this project, I used several Pandas modules to achieve my goals. I stored my information, which was in the form of a database with columns for the start and end dates of the period, weekly totals of ridership or fares of each type, as well as comparisons to the pre-pandemic ridership, into a Pandas dataframe. I ended up creating newer, smaller dataframes to store just the relevant columns: period start and end dates, and full fare ridership.

That is, the first techniques I used were data wrangling. I examined the data to make sure there weren't any holes; and that all the information was of the right datatype. Ultimately there was no missing data, but there were four rows at the end of the first dataset that were much more spread apart than the date ranges of the rest of the data; so I ended up dropping those out, since that information was going to be expressed in the second dataset anyway (it was the information from the year 2021). I also had to compress the fare information that was taken from every single station in the city, and make it into one value that I could refer to as the systemwide ridership.

I ended up learning a lot about the ”datetime” module in Python. It is a useful way to store dates and times as a datetime object, rather than an integer or string; but over the course of my work, when referencing that column, I occasionally had to switch back to integer or datetime object in order to do Boolean operations and comparisons on years, which are more easily represented sometimes as an integer. However, the datetime module did make visualizing the data very convenient. There are already many ways to use date objects; I did not have to, for instance, switch from a string to an integer in order to give the dates an order, and then back to a string to write the labels. The Pandas modules let me use the datetime objects for the x-axis of the figures, and then use them to refer to entries in the dataframe to find the value at those points.

That is, I used the Pandas plotting function, in order to easily plot the information stored in the dataframe; I was able to directly plot the dates column vs the full fares column which I had calculated. As I used this and refined the visualization, I realized there are some differences in syntax between Pandas plot and the pyplot module in matplotlib. For the purposes of this project, it was simple to directly go from the Pandas dataframe to the Pandas plot; however there are also ways that the pyplot is more straightforward and simple for plotting functions directly or annotating the figure. But because I was using the Pandas plot, it was simpler to stick directly to the resources available in that module.

Finally, I used the data science technique of following up on questions that the data visualizations brings up. Of course it would be interesting to find the maximum and minimum right away; (these occurred during Spring 2015 and during Spring 2019). I was also able to find that the busiest station during the busiest week was the 14th Street Union Square stop. And as for the minimum, I found that the lows in ridership during the start of the COVID-19 pandemic were even lower than those during the week of Hurricane Sandy.

But beyond the intuitive metrics such as maximum and minimum, visualizing the data allowed me to ask more questions that need more investigation. Seeing the trends over the course of 10 years, there are extreme values that made me look at what might have been happening in the city at that time. But there were also periodic fluctuations that raise questions about how people are using the subways, and how seasons or certain periods of time affect that ridership.

3 Results and Conclusions

As I expected, the trends in ridership over the MTA system tracks major events transpiring in the city, as well as demonstrates periodic trends. At first I considered investigating the trends in usage to systemic issues such as repair/maintenance, accessibility, and demographics. While I think there is still a lot to be interpreted from that kind of investigation, in the scope of this project I ended up looking at the system-wide trends; I plotted usage over the course of the entire system for an entire week, against time. Thus I was able to see more ”big picture” trends, which places more emphasis on how disasters like Hurricane Sandy and COVID-19 can throw the system as a whole into complete turmoil.

There are many considerations that must be made in tracking public transit ridership. Take for instance, the many kinds of fare cards. Maybe residents are more likely to purchase weekly or monthly fare cards, or use buses, so by tracking single full fare swipes exclusively on subways maybe I am really more likely looking at tourist usage. Moreover, it is difficult to find data to make any interpretations about lengths of trips, or destinations, or time of day. This is not available from the MTA itself, since it can not be gathered in the units of fare card swipes. So it is important that I was clear about what the data shows, and what it does not. Specifically, the data from 2010-2020 is the sum of all fare card swipes at all stations in the MTA subway system across the whole city, over the course of a week period. The data from the more recent period is the daily estimated ridership of the subway, over the whole subway system. Since the first data set includes every type of fare swiped, over the course of the week, I restricted the data to just the ”Full Fare” swipes and ”Senior/Disabled Fare” swipes. This does not say how many people used the subway that week, but how many people used a single full or senior/disabled ride fare. Meanwhile the more recent data estimates the total ridership. If I were able to look at this in more detail, I would like to further investigate the trends in the different kinds of fare cards over the period; and any affects accessibility initiatives have had.

As it is, I was able to visualize two views of the ridership in the past decade or so. While there are three or four major events that had dramatic affects on subway ridership, overall it seems that there are periodic ups and downs in usage, as well as a slight downward trend in the past decade as a whole in full fare ridership, while slight upward trend in the use of senior/disabled fares.

3.1 Time Period: 2010-2020

When I went into the project, my initial goals were to visualize the data and see what questions came up, and then focus in on structural or demographic information. I was surprised to find that more and more of my questions became "wait what happened that winter?" Even though I have lived in the city since 2013, I found myself travelling through old news articles to try to retrace major events in the city that could explain the data, but I couldn't always find the explanation I was looking for. Of course, there is always the possibility that a certain spike or drop is due to some kind of data gathering error, or simplification error on my end. Or it might be due to broader factors acting together, or a confluence of various factors.

There are a lot of interesting conclusions we can draw from the data, but there is also still a lot left unanswered, not only because of my focusing on broad trends, as well as one method of transit in particular.

Also, during this period ride and bike rental apps were introduced citywide. I compared the dates that these were introduced to the overall trends of ridership, and it seems that there was not a clear connection. That is, there does seem to be an overall downward trend in ridership, but in the period directly after introduction of Uber especially, there was a slight growth, despite periodicity of usage.

Furthermore, as I read more on the topic, I have realized that there is a very diverse and interactive web of transit options in the city. In the period between 2010-2020, the rideshare industry has expanded, as well as accessibility initiatives taken in public transit; there have been drastic variations in gas prices; there have been deals made between ridesharing companies and the Taxi and Limousine Commission; there have been struggles, strikes, and victories for the drivers in the taxi union. That is to say, in the scope of changes in the transportation landscape in New York City in the past decade, there is such a confluence of various factors that compete and collaborate to influence the transit ecosystem in the city, so I don't think I am able to conclude that introduction of Lyft, Uber, or Citibike on its own had a noticeable effect on the ridership over the whole subway system.

3.2 Time Period: COVID-19

This data from 2020-2021 shows how extreme the effect of COVID-19 on public transit was. On a personal note, as someone who used the subway on a nearly daily basis during this period, it was very striking to see the usage trends laid out on a timeline, and compare that to the key historical events. I bring this up because part of studying data is the narratives we create around it. Seeing the systemic usage data in the context of my own memories of being alone on a completely deserted subway train adds a dimension to how I see the ups and downs on a graph. This is very different from how I view those same dramatic impacts in the data during the effects of Hurricane Sandy, which was a year before I moved to the city. I think that is something I really had not considered before about the way that while data can seem so rigid and clear cut when looked at in the abstract, there is always a significant level of interpretation that goes into the analysis of the factual information gathered, as well as potential for errors at every stage of the process.

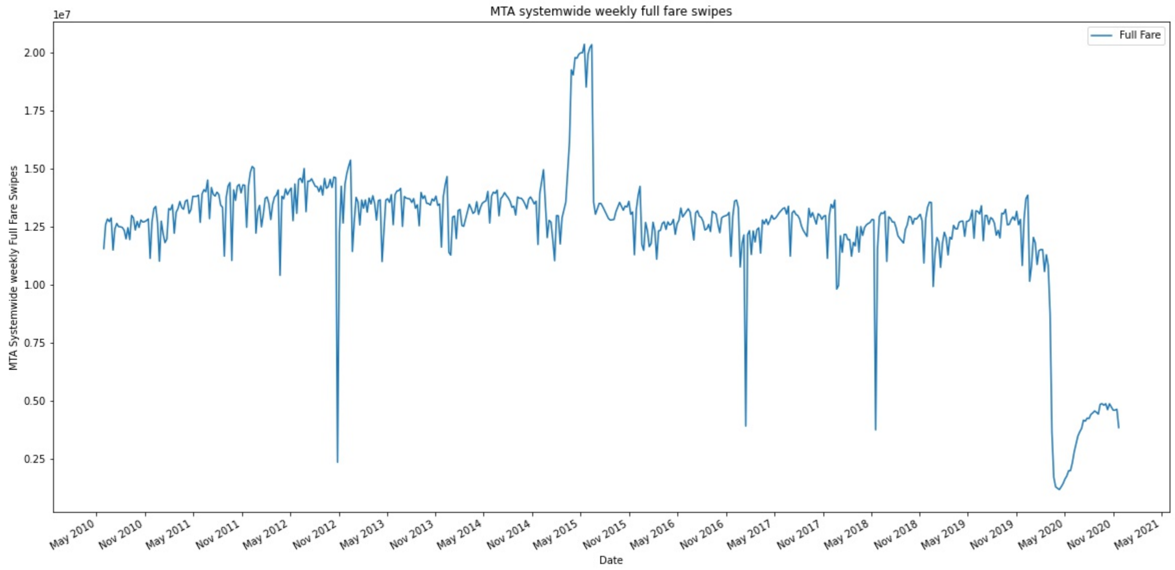
3.3 Significance

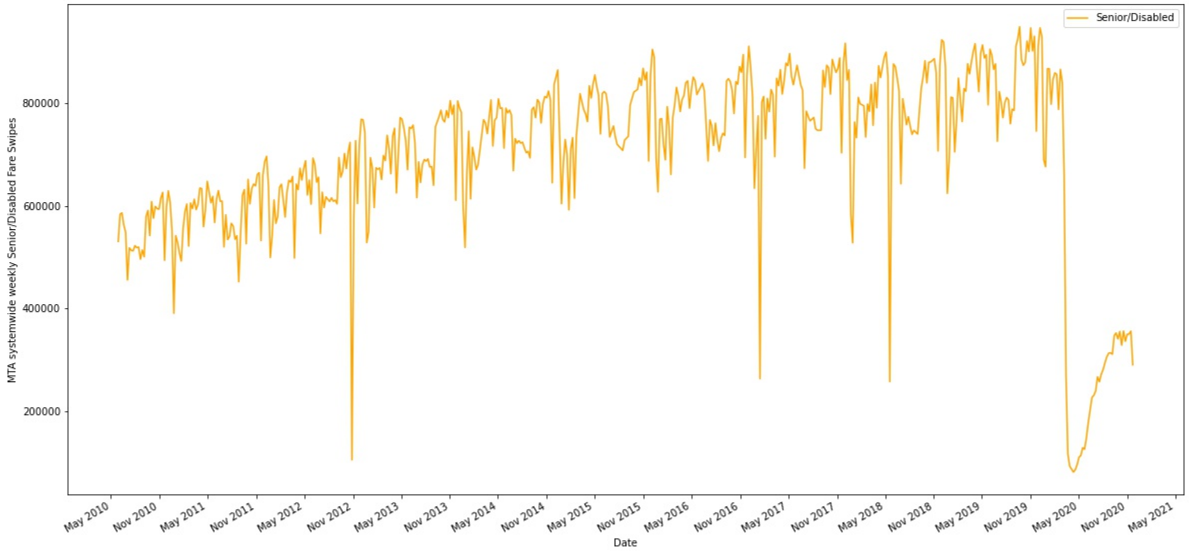
At the end of the day, we can see that though there may be ups and downs in ridership on the subway, even narrowed to single full fare riders, is consistent over the past decade. Furthermore, even during the pandemic when the city was in a state of emergency, 1,164,815 people used a single-ride pass to ride the subway across the city that week when everything was at the worst; during the week Hurricane Sandy hit, 2,341,384 people used a single fare pass. There are seasonal trends and changes, and clearly the impact of tourists on the subway system can use further study. But the usage at these times when tourism is likely not a factor shows what an essential role the MTA serves for the residents of New York City. If I were to take this project further, the next issue I would look at is accessibility. Since 2017 there has been an initiative to increase accessibility in the system, but to this day there is an overwhelming majority of stations that are simply not accessible to people with disabilities. Furthermore, there is available data on station closures which would help to paint a picture of which areas of the city are not accessible to anyone by subway.

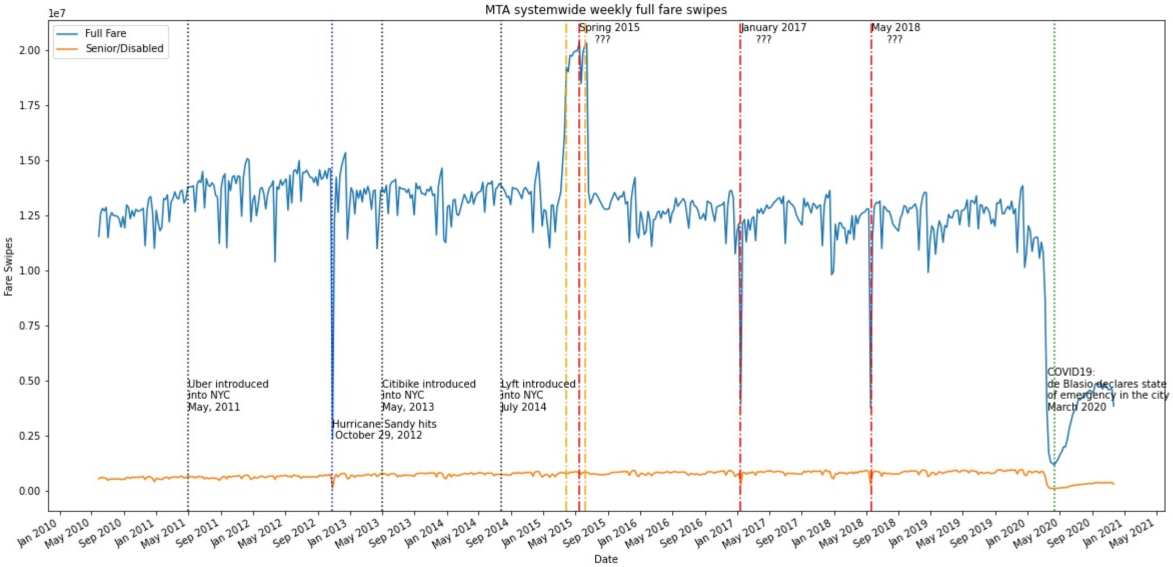
Furthermore, comparing the plots of the full fare and senior/disabled fare data, we can see that while the full fare swipes were gradually declining, the senior/disabled swipes were gradually increasing. While COVID-19 has made future conclusions based on this data difficult, I think it is clear to see that the subway system is essential – not just to tourists or visitors, but to NYC residents, and increasingly, for senior and disabled residents, who currently have the least access to the subway.

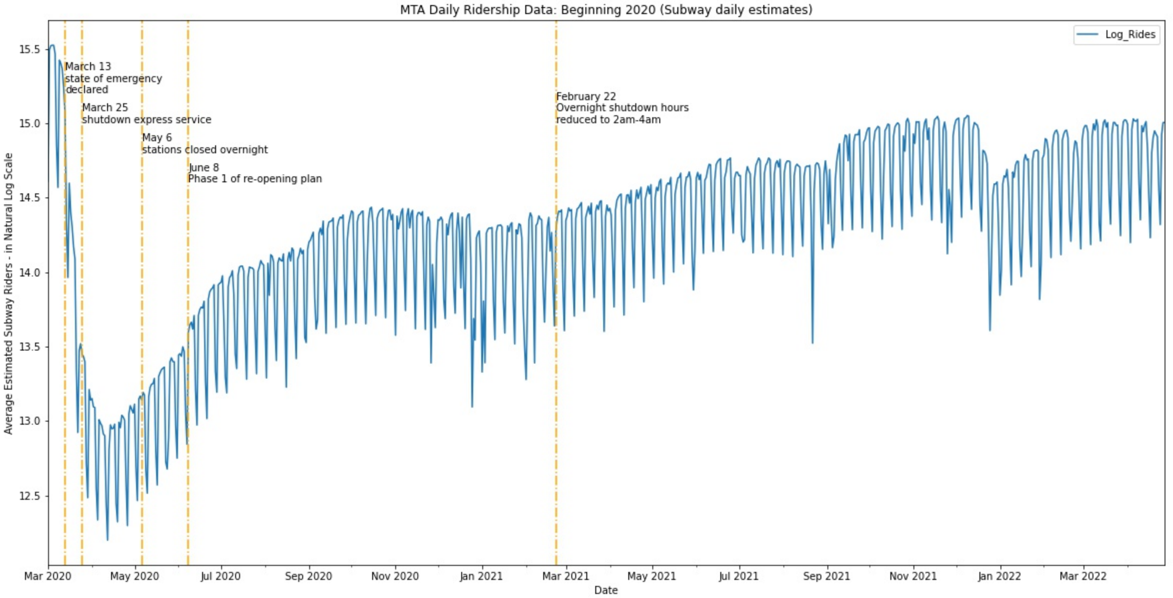
This is the direction I would like to see research like this taken: to inform both policy and public perception of the subway.

4 Graphics


MTA systemwide weekly full fare swipes
1e7
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
MTA Systemwide weekly Full Fare Swipes
Full Fare
May 2010
Nov 2010
May 2011
Nov 2011
May 2012
Nov 2012
May 2013
Nov 2013
May 2014
Nov 2014
May 2015
Nov 2015
May 2016
Nov 2016
May 2017
Nov 2017
May 2018
Nov 2018
May 2019
Nov 2019
May 2020
Nov 2020
May 2021
Date


800000
600000
400000
200000
MTA systemwide weekly Senior/Disabled Fare Swipes
Senior/Disabled
May 2010
Nov 2010
May 2011
Nov 2011
May 2012
Nov 2012
May 2013
Nov 2013
May 2014
Nov 2014
May 2015
Nov 2015
May 2016
Nov 2016
May 2017
Nov 2017
May 2018
Nov 2018
May 2019
Nov 2019
May 2020
Nov 2020
May 2021
Date





References

[] URL: https://www.nytimes.com/2020/04/30/nyregion/subway-close-cuomo-coronavirus.html.

[] URL: https://www.bbc.com/news/business-60864889.

[] URL: https://slate.com/business/2015/11/uber-won-new-york-city-it-only-took-five-years.html.

[] URL: https://ride.citibikenyc.com/about.

[] URL: https://www.cbsnews.com/newyork/news/nyc-top-news-stories-2015/.

[] URL: https://www1.nyc.gov/site/cdbgdr/about/AboutHurricaneSandy.

[] URL: https://www.ny1.com/nyc/all-boroughs/news/2021/03/10/timeline--how-covid-19-changed-nyc.

[] URL: https://www.aapf.org/sayhername.

[] URL: https://www.cbsnews.com/newyork/news/nyc-top-news-stories-2015/.

[] URL: https://www.bbc.com/news/business-60864889.

[] URL: https://slate.com/business/2015/11/uber-won-new-york-city-it-only-took-five-years.html.

[] URL: https://ride.citibikenyc.com/about.

[] URL: https://data.ny.gov/Transportation/MTA-Daily-Ridership-Data-Beginning-2020/vxuj-8kew/data.

[] URL: https://data.ny.gov/Transportation/Fare-Card-History-for-Metropolitan-Transportation-/v7qc-gwpn/data.

[] URL: https://www1.nyc.gov/site/cdbgdr/about/About%20Hurricane%20Sandy.

[AT] Patrick Adcroft and Faraz Toor. *Timeline: How COVID-19 Changed NYC*. URL: https://www.ny1.com/nyc/all-boroughs/news/2021/03/10/timeline--how-covid-19-changed-nyc.

[Gus] Clayton Guse. *Overnight closure of NYC subway to scale back to 2 a.m. to 4 a.m. starting Feb. 22*. URL: https://www.nydailynews.com/coronavirus/ny-covid-cuomo-subway-reopening-mta-overnight-20210215-pb3wpoi7bzfmdizfmloao4kqrq-story.html.

[Lag] Christine Lagorio-Chafkin. *Lyft in New York City: Let's Try This One More Time*. URL: https://www.inc.com/christine-lagorio/lyft-another-nyc-launch-attempt.html.

[] *MTA Slashes Service, NJ Transit on Reduced Schedules.* URL: https://web.archive.org/web/20200326033708/https://www.nbcnewyork.com/news/local/mta-suspends-some-lines-of-service-amid-coronavirus-staff-shortages/2341550/.

[Sif] Andrew Siff. *MTA Resumes Regular Weekday Service; Overnight 4-Hour Closure Stays Both subway and bus services will remain only for essential workers and essential trips during Phase I.* URL: https://www.nbcnewyork.com/news/local/mta-to-resume-regular-weekday-service-when-nyc-enters-phase-1-overnight-4-hour-closure-will-stay/2449009/.