Correlation Between Confirmed Cases, Mask Mandate and Subway Usage in New York County, New York

Chang Xu

Dec 11, 2021

Section 1: Introduction

We have been experiencing this unprecedented global pandemic for about two years now. The pandemic has been tragic and disruptive to many countries and families around the world. Actions have been taken in order to mitigate the influence of the pandemic and to bring our lives closer to the "old normal". Such actions include enterprise-level work-from-home policies and state-wise vaccination policies and mask mandate policies in public. While a lot of COVID-related studies have been conducted, I am interested in using the human-centered data science approach I learned in this class to examine the pandemic from potentially different perspectives to understand how it has changed lives and how it has changed society using the data collected, aggregated and re-represented from many different data sources. Specifically, while the pandemic is very complex and has too many aspects, I want to quantify the influence of the pandemic and the mask mandate policy on the frequency of people using public transportation, especially the subway, in New York County, New York, which is the cultural and economic center of the United States. Understanding this problem is important because it may help policymakers to understand people's action patterns and better allocate resources to accommodate any change that may occur.

Section 2: Background and Related Work

There have been a number of research studies conducted about COVID and public transit as an essential service during the pandemic. The research "The impacts of COVID-19 pandemic on public transit demand in the United States" [1] and the online article "COVID made many of us avoid public transport - what will it take to get us back on the bus?"[2] I read showed that the COVID-19 pandemic and related restrictions led to major transit demand decline for many public transit systems in the United States. "Approximately half of the agencies experienced their decline before the local spread of COVID-19 likely began; most of these are in the US Midwest. Almost no transit systems finished their decline periods before local community spread...The results show substantial departures from typical weekday hourly demand profiles."

While the research focuses on both bus and subway in the entire United States, looking at some detailed metrics such as hourly demand and using data derived from a widely used transit navigation app, and looking into various identity groups such as African American, Hispanic, Female, and people over 45 years old, I am more interested in studying the overall trend and how people's behavior change specifically in New York County, aka Manhattan, for subway alone.

For this project, I use human-centered data science approach to explore the following two questions:

1. How did the mandatory mask-wearing policy in public change the spread of COVID confirmed cases in New York County?

2. How is the frequency of people using the New York City subway system influenced by the pandemic and the county's mask-wearing policy?

My hypotheses are:

1. The mandatory mask-wearing policy reduced the infection rate and slows down the spread of COVID.
2. People use the New York City subway system less when there are more confirmed cases and when mask-wearing is not required, and use the subway system more otherwise.

Section 3: Methodology

Before understanding the methodology I used for this project, it is important to know the datasets I used, each of them is closely related to the questions I explored. The first dataset I used is "COVID-19 data from John Hopkins University", which is a daily updating version of COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). The data covers confirmed cases and deaths on a country level, confirmed cases and deaths by US county, and some metadata that's available in the raw JHU data. The second data I used is "U.S. State and Territorial Public Mask Mandates From April 10, 2020, through August 15, 2021, by County by Day" published on the CDC official site, which contains state and territorial executive orders, administrative orders, resolutions, and proclamations collected from government websites and cataloged and coded using Microsoft Excel.

I used these two datasets to calculate the cumulative confirmed COVID cases, daily confirmed cases, and daily infection rate and to mark the changes in government policy in New York State for mask-wearing in public. Specifically, in calculating the daily infection rate, I calculated a smoothed version of the infection rate by considering a window of 7 days and taking the average to get a smoother curve. I define a concept called "susceptible population" which is the number of confirmed cases subtracted from the total population, and I then calculated the smoothed infection rate by dividing the 7 days averaged cases by this susceptible population and times the ratio by 100. My calculation for cumulative confirmed cases and daily confirmed cases are more straightforward: just doing summation of all confirmed cases till that date and doing subtraction between two cumulative counts for every two consecutive days.

To answer my second question, I introduced another dataset called "NYC_subway_traffic_2017-2021.csv", which is 735.17 MB in size and includes the number of subway station entries and exits, as counted by the number of people passing through the turnstiles located at the station entrances, at 4-hour intervals, for 469 subway stations from February 4th, 2017 to August 13th, 2021. This is an integrated dataset I found on Kaggle (link in section 9). The original turnstile data, which spans from May 5 2010 to November 06, 2021, is published by New York City MTA (Metropolitan Transportation Authority) and can be found on the official MTA website [3]. The Kaggle data is under the CC0: Public Domain license, but the original MTA data does not have a license associated with it, and I assumed it can be used freely by the public. The Kaggle dataset contains fields of Unique_ID (a custom identifier key corresponding to a unique combination of stop name, remote unit, line, and connecting lines), Datetime (Timestamp. The actual interval is from 2 hours before to 2 hours after the given time), Stop_Name, Remote_Unit (higher-level hierarchical identifier), Line (the line to which the stop belongs), Connecting_Lines (lines that pass through this station), Daytime_Routes (daytime lines that pass through this station), North_Direction_Label (destination of north-going lines), South_Direction_Label (destination of south-going lines), and Division (each line belongs to one of three subway system

divisions). The original MTA data contains rows of "C/A, UNIT, SCP, STATION, LINENAME, DIVISION, DATE, TIME, DESC, ENTRIES, EXITS", which are mainly encoded station/line names and timestamps.

I used this dataset to count the daily number of people/times utilizing the New York City subway system and compared whether the number of passengers had changed significantly with the trend of confirmed COVID cases in New York County, New York. I did so by counting all ridership records for every single date. A ridership is defined as a record when a passenger swipes into the subway system and swipes out of the subway system, regardless of the length of time they utilize the system or the mileage they traveled. I also observed if the policy of requiring wearing a mask in public has any influence on the number (do people feel safer as more people are wearing masks and use this type of public transportation more?). In addition, I compared the popularity (defined the same as ridership) of some selected subway stops which are identified as the most popular stops (Times Square, Grand Central, and Herald Square), and some less popular stops (Flushing Av, Alabama Av, and Forest Av) before the COVID outbreak.

The common methodology I chose to conduct my analysis for both questions I asked other than the calculation I mention above is doing time-series analysis over the dataset and using visualization to observe the patterns. The reason that a time-series analysis is the right one to use, is because we are investigating non-stationary data - data over things that are constantly fluctuating over time or affected by time. Using visualization, we can clearly observe the trend and how the variable we are interested in (confirmed COVID cases and population who utilize the New York City subway system) change over time, in accordance with some major events, such as the beginning of the outbreak of COVID-19, the mandatory requirement for wearing masks in public places, the termination of the requirement, and more.

In order to make sure there are no potential ethics problem occurs, I constantly followed the human-centered data science principles while conducting my analysis. None of the datasets I used provide any personally identifiable information. This is especially true for the JHU data and the CDC data as they give aggregated national-level data. By comparison, there could be some possible ethical considerations to using the subway dataset, because although this dataset is highly anonymized, and it is almost impossible to identify who generated those data, it is still recording and tracing people's activity, which, when combined with other data sources, someone might still be able to make some inference about the individuals. This hinders people's privacy to some extend. However, I do not think we should worry too much about this issue, as it is so common that almost everything related to big data has this problem. On the other hand, the data is highly transparent and is published on the official website. However, I do feel that people are not asked for consent when this data is collected. As a traveler to New York City, using the city's subway system, I was never asked for consent for this type of data. The research itself is highly reproducible as the original source code, detailed documentation, and links to relevant data and license are all provided in a GitHub repository.

Section 4: Findings

In order to answer my first analysis question, using the method mentioned above, I plotted three figures about COVID confirmed cases (cumulative confirmed cases, daily confirmed cases, and daily infection rate) with the mask mandate being effective indicated in green in Figure 1, 2, and 3 below respectively. Observing the slope of Figure 1. COVID-19 Cumulative Confirmed Cases in New York County, New York, we can see when mask mandate policy first became effective, the slope of cumulative confirmed cases flattens, although after a long time (from 2020-05 to 2020-11) it starts to rise again, it

flattens again after that. The curve of cumulative confirmed cases stays almost flat until the requirement of wearing masks in public is removed, and then we see a significant increase of the gradient.

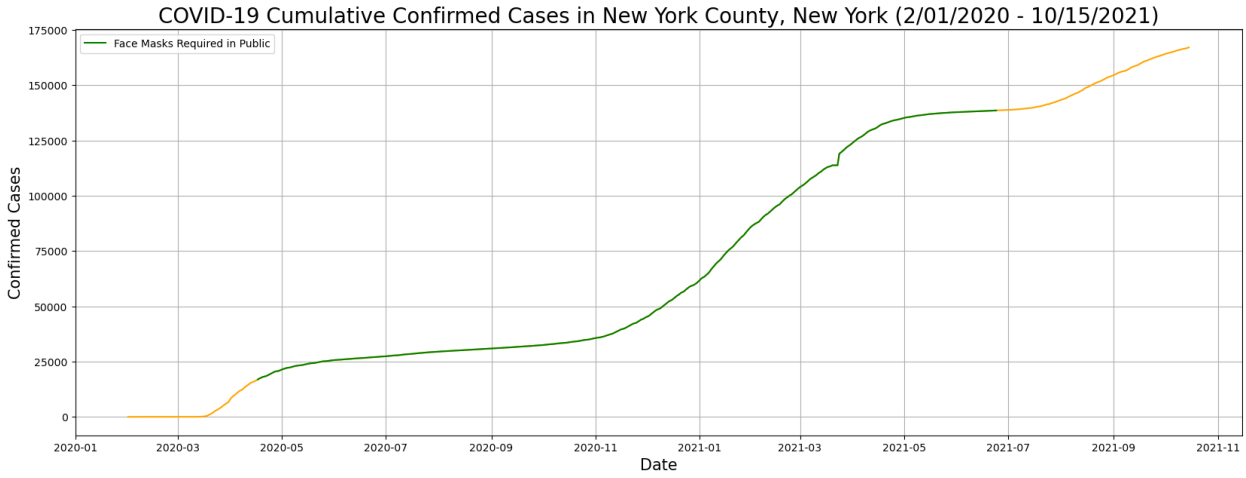

Figure 1. COVID-19 Cumulative Confirmed Cases in New York County, New York

The same trend is more obvious in Figure 2. COVID-19 Daily Confirmed Cases in New York County, New York and Figure 3. COVID-19 Daily Infection Rate in New York County, New York below. As the mask mandate becomes effective, the daily confirmed cases and infection rate drops significantly. The two indicators both stay at the same stable level for a while and start to increase roughly from mid-November, 2020. Regardless of the spike outlier, both indicators reach their peaks roughly around early/mid-January, 2020. Then both indicators drop again and stay relatively stable again from roughly June 2021 to July 2021, when wearing masks in public is no longer required in New York County. And as a response to that change, both the daily confirmed cases and daily infection rates rise again.

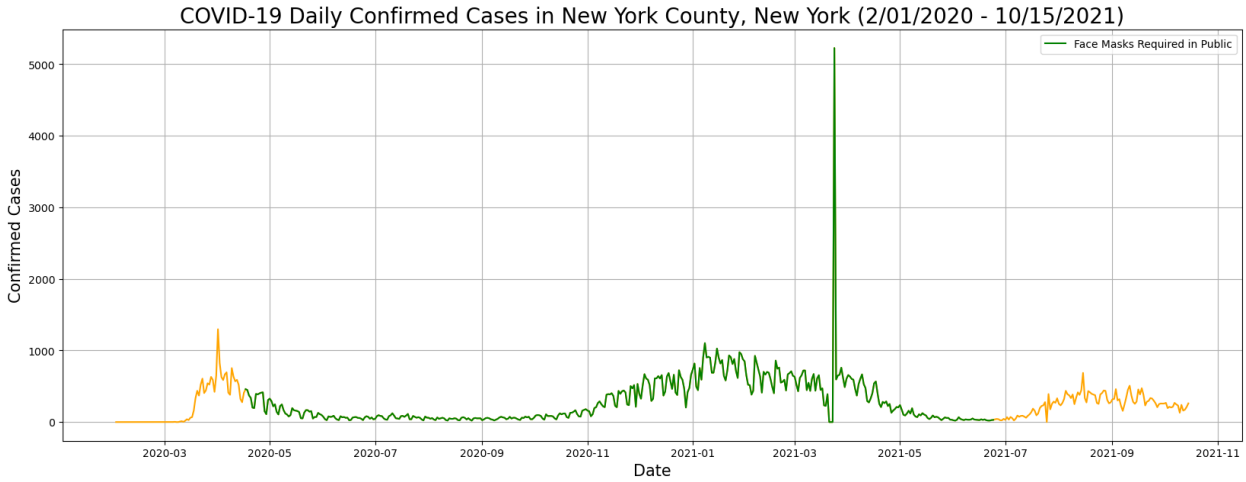

Figure 2. COVID-19 Daily Confirmed Cases in New York County, New York

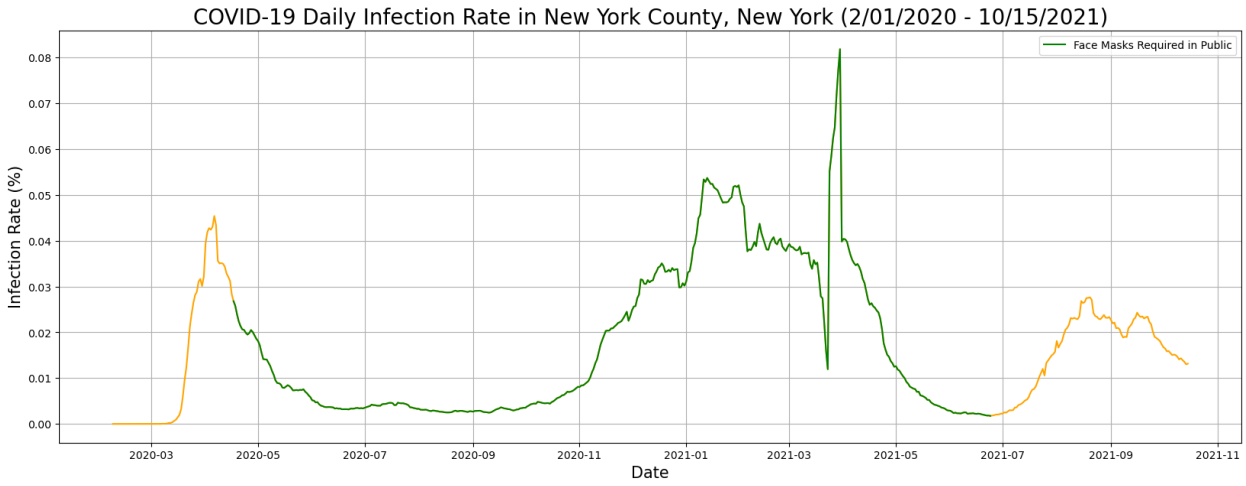

Figure 3. COVID-19 Daily Infection Rate in New York County, New York

The results I got in the 3 figures about confirmed cases and mask mandates show the same trend and highly agree with each other. Hence, we are able to draw the conclusion that the policy of wearing masks in public effectively slows down the infection rate and confirmed cases.

The below Figure 4, Figure 5, and Figure 6 are what I got answering my second question. In Figure 4. New York City Subway Ridership with Daily Confirmed Cases, we see that the ridership of the New York City subway is relatively stable and does not change in accordance with the daily confirmed cases. Therefore it is hard to say that the two things are correlated.

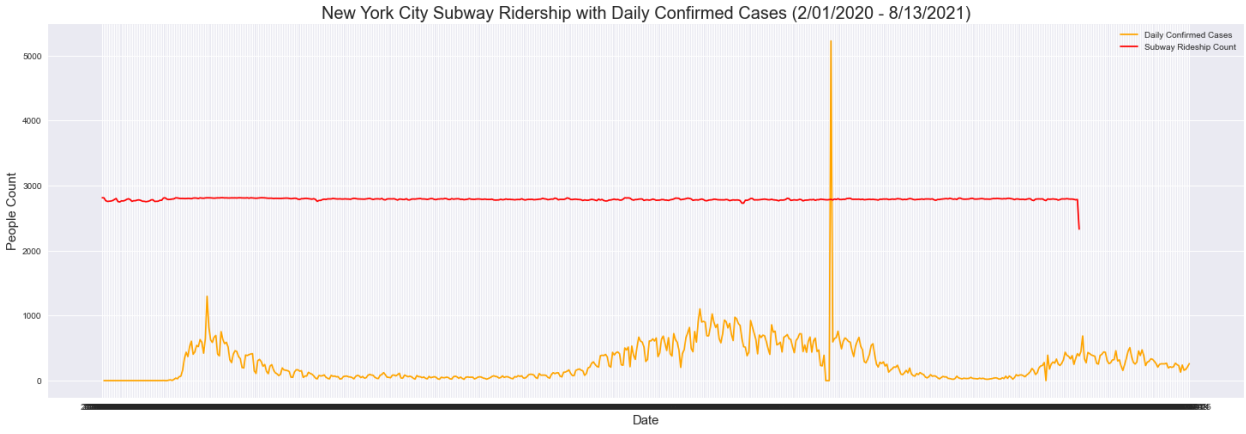

Figure 4. New York City Subway Ridership with Daily Confirmed Cases

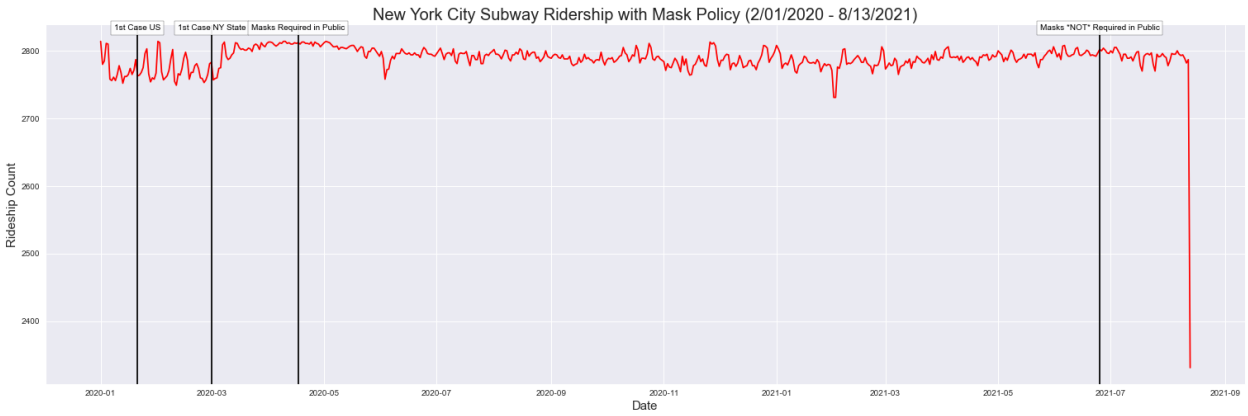

Figure 5. New York City Subway Ridership wit Mask Policy

In Figure 5. New York City Subway Ridership with Mask Policy, while we observe some fluctuations in ridership, it is also hard to relate them with the mask mandate policy, which is the time period between the third and the fourth black lines.

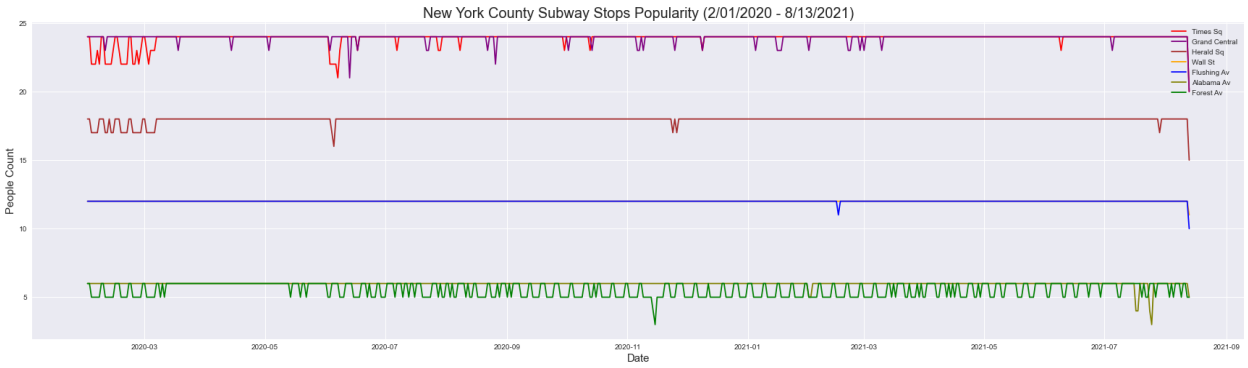

Figure 6. New York City Subway Stops Popularity

Figure 6 compares the popularity of 7 selected subway stops throughout the pandemic. Times Square, Grand Central, and Herald Squares, identified by the MTA official website and an article in 2018 as the “most popular stops”, stay the most popular. Wall Street Stop, which is moderately popular and is a famous stop we all know is in the next level and is relatively stable in popularity too. The least popular stops in this plot, Forest Av and Alabama Av stay the least popular ones. However, we can observe that there are more fluctuations for those least popular stops.

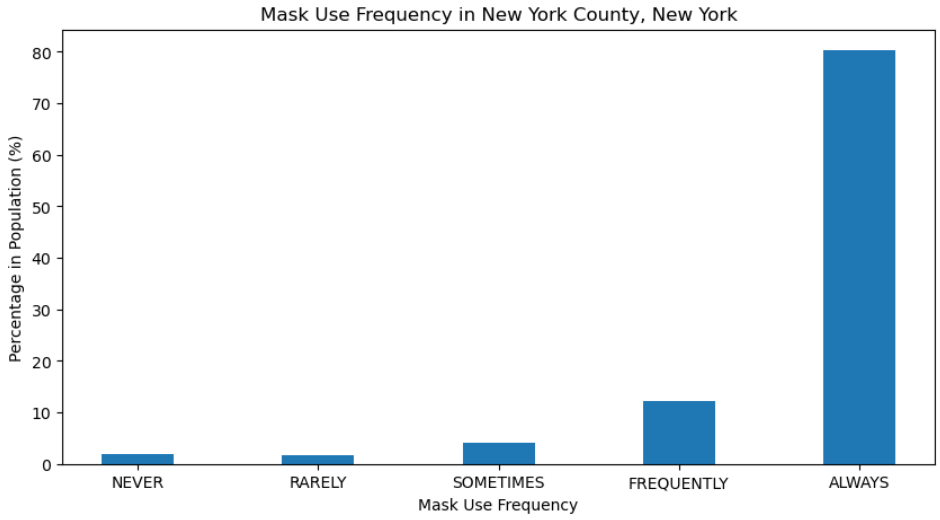

Figure 7. Mask Use Frequency in New York County, New York

An additional finding in my project is the mask use frequency in New York County. As indicated in Figure 7, more than 92% of respondents in the county frequently (12.1%) and always (80.3%) use a mask. We, therefore, can say that people in New York County are very compliant with the policy. This fact will be helpful when we do our discussion in the next section.

Section 5: Discussion and implications

Although the first three figures in the previous section have shown that the mask mandate policy in New York County is effective in lowering the daily confirmed cases and infection rate, we do observe there are some rises in confirmed cases while the mandate policy is still in effect. The timings are interesting. We observe and note that both infection rate and daily confirmed cases start to increase roughly from mid-November, 2020 and reach their peaks roughly around early/mid-January, 2020. Mid-November is when more people start to travel for Thanksgiving, going out to meet their families and friends. More traveling and gathering increases the spread of COVID-19, regardless of the public mask mandate policy. After all, when people have meals with their families and friends, they take their masks off, and when people travel, there is a greater chance that they encounter someone who has COVID. This increases as the holiday season continue and reaches the peak around early/mid-January, 2020, which is when people finish their Christmas and New Year holiday and go back to their life before vocations. Therefore, research can be conducted to see to what extent do holidays increase the spread of COVID. Another interesting thing that I would like to explore is the sudden spike in both daily confirmed cases and the infection rate around April 2021, I looked at “A Historical Timeline of COVID-19 in New York City” [6] but did not find anything accountable.

Moreover, I am interested in the reasons that cause fluctuations in Figure 5 New York City Subway Ridership with Mask Policy. I also did not find anything that is related in “A Historical Timeline of COVID-19 in New York City”, but it is worth exploring. It is also very interesting to see that the ridership I thought would be influenced by the confirmed COVID cases and mask mandate does not seem to be correlated. This can be explained by the large number of commuters in New York City and the city’s special status of being the economic and cultural center of the country, or even the world, but I would like to dive deeper into this if I had the chance.

Section 6: Limitations

There are a few limitations to my project. First, when I calculate the infection rate, I made two assumptions. The first is I assumed the population of the entire time period remains the same, which is as the data provided to us. However, the population of a county is always changing dynamically as some people die and others are born. Another assumption I made is that all people have the same probability of getting COVID, regardless of whether they wear masks or not. This does not reflect the real situation as multiple research papers show that wearing masks reduces the risk of getting infected. In addition, some people are more infectious than others because of their body conditions and their daily activities. Another factor that we cannot overlook but is missing in my analysis is vaccination. Vaccination reduces the risk of getting infected, and throughout the entire period of my study, the number of people getting vaccinated also changes, and not to mention the number of doses they receive, which will complicate the situation even more. Therefore, some work needs to be done to refine the way I calculate the infection rate in order to get a more accurate result.

There is also a limitation to the Kaggle dataset "NYC Subway Traffic 2017-21" I used. First of all, the dataset is too small. The dataset is only 735.17 MB in size and count at 4-hour intervals, which makes the dataset a proportionally distilled version of the real turnstile data. As a result, the analysis results I get are also smaller compared to the real-world case, for example, the daily ridership of the entire subway system is only around 2,000 and the popularity of the stops only ranges from 5 to 25, which is very far from the real situation. In a city like New York City, the real numbers must be much greater. While this dataset may still give the right trend, it is much less accurate than what it could be, and a lot of interesting features of the output pattern may also get omitted in this way. I chose this dataset because it was the only thing relevant I could find that is easy to use. However, if more time is allowed, I would go to the MTA official websites, figure out a way to crawl all the raw data files (which will be enormous), clean them and come up with my own dataset. The fields in the Kaggle dataset are well designed though.

Another limitation is when I explored how the stop popularity changes over the pandemic, I only selected 7 stops to analyze. While the top 3 most popular stops may be representative, the others may not be as symbolic. One fix to this is selecting more stops evenly based on their popularity before COVID, however, data that has the full list of subway stops ranked by their popularity is not available, as people are normally only interested in learning the top X data. In this case, we may need an alternative way to do this, by just calculating the popularity, i.e. ridership, at each stop, but that will be a larger workload, as there are 151 Subway stations in Manhattan. [7]

Section 7: Conclusion

In this project, I used human-centered data science approaches to perform a systematic analysis of COVID-19 in New York County, New York, that mainly addresses two larger questions: first, how the mandatory mask-wearing policy in public changed the spread of COVID confirmed cases in New York County, and second, if the frequency of people using the New York City subway system is influenced by the pandemic and the county's mask-wearing policy. I used time series analysis and visualization to help me find answers to both of my questions. For the first question, after analyzing and plotting the cumulative confirmed cases, daily confirmed cases, and infection rate of New York County from 02/01/2020 to 10/15/2021, I am able to arrive at the conclusion that the mask mandate policy is effective in mitigating the pandemic. However, there is still some increase in confirmed cases and infection rate a long time after the policy is in effect that can make an interesting further study topic. The exploration of my second question is broken into three parts, other than just looking at the relationship between subway

ridership and confirmed cases and the mask mandate policy, I also explored the popularity of 7 selected subway stops. As a result, I was not able to find a correlation between the daily confirmed cases and the mask mandate policy, and the total subway ridership. The popularity of the selected stops stays relatively stable too, with the most popular stops staying the most popular. However, I have observed some larger fluctuations in stop popularity for those less popular stops.

I think overall through this project, I was able to revisit the data processing and combine skills that I learned earlier in this class (A2), which I found very beneficial. From answering the research question posed in this assignment, I learned how to perform time series analysis, which is something really important in data science studies. Another thing I learned and found very valuable is that I used to think we can only draw conclusions from things that already exist, but this research reminds me that we are able to do calculations and extract hidden data/features from what we see on the surface, and then study and explore those that we get. This may sound like common sense, but it was not obvious enough until I finished this research and looked back. Moreover, it is important to always keep in mind the research limitations. In general, I learned to follow the human-centered data science principles in practice, which is very beneficial to the future research and analysis I do. For specific techniques, one thing I learned is rolling average, a simple and common type of smoothing used in time series analysis and time series forecasting. Calculating a moving average involves creating a new series where the values are the average of raw observations in the original time series. In my analysis, I used it to calculate the derivative, infection rate. Another lesson I learned is sometimes the result doesn't work out as we expected, and it turned out some other students also experienced the same thing. Further research can be done to address why this happens.

Section 8: References

[1] "The impacts of COVID-19 pandemic on public transit demand in the United States"
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0242476

[2] "COVID made many of us avoid public transport - what will it take to get us back on the bus?"
https://www.weforum.org/agenda/2021/02/public-transport-covid-data/

[3] "Turnstile Data - Metropolitan Transportation Authority (MTA)"
http://web.mta.info/developers/turnstile.html

[4] "MTA: Top 10 busiest subway stations in 2020"
https://new.mta.info/agency/new-york-city-transit/subway-bus-ridership-2020

[5] "Photo Essay: Inside NYC's Least Visited Subway Stations"
https://untappedcities.com/2018/03/12/photo-essay-inside-nycs-least-visited-subway-stations/

[6] "A Historical Timeline of COVID-19 in New York City"
https://www.investopedia.com/historical-timeline-of-covid-19-in-new-york-city-5071986

[7] "Often Asked: How Many Subway Stations Are There in New York City"
https://lastfiascorun.com/faq/often-asked-how-many-subway-stations-are-there-in-new-york-city.html

Section 9: Data Sources

1. The RAW_us_confirmed_cases.csv file from the Kaggle repository of John Hopkins University COVID-19 data. https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_confirmed_cases.csv
2. The CDC dataset of masking mandates by county. https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i
3. The New York Times mask compliance survey data. https://github.com/nytimes/covid-19-data/tree/master/mask-use
4. The kaggle dataset “Normal and New Normal: NYC Subway Traffic 2017-21 (NYC subway traffic before, during and after the covid lockdown)” https://www.kaggle.com/eddeng/nyc-subway-traffic-data-20172021