NYC Transportation Data Analysis under Covid-19

Github link: https://github.com/shiroro666/Big_Data_Project

Huanyao Ye (hy2084@nyu.edu)
Chenhui Fan (cf2647@nyu.edu)
December, 2020

Abstract

This paper presents the data analysis on datasets which are related to transportation in NYC during COVID-19 pandemic. The two main aspects of transportation we analyzed were subway and CitiBike, which is New York's bike share system. The data was processed by using Spark and the analysis was performed using are visualized by Altair. This report explains and analyzes how COVID-19 has changed transportation like the number of passengers and the distance of the trips.

1. Introduction

With the spread of COVID-19 pandemic, whose first case was found in Wuhan, China, many cities in the US reported infected cases. New York City confirmed its first case of COVID-19 at the beginning of March 2020. Due to the drastically increasing cases, the city of New York proposed a “Stay at Home” policy responding to the severe pandemic situation.

Owing to the closure of part of the business in NYC and the spread of COVID-19 in NYC, people’s way of life changed in many perspectives. Among variety of aspects which are influence by COVID-19 pandemic, we are interested in transportation in New York City and we will focus on how the people changed their ways of mobility to keep safe in our project. There were many data captured to reveal how NYC transportation altered according to the situation of pandemic. We choose two kinds of transportation in the NYC, which are subway and bike. Our goal is to seek how mobility changes during the pandemic by comparing with the statistics during 2019 and combine the difference during the pandemic with the daily COVID-19 cases data to visualize. We analyze how these effects evolve over time, find some of the mobility methods that stay unchanged or functioning during this pandemic, and conclude the potential suggestions for vehicle choices based on analysis.

2. Problem Formulation

The outbreak of Covid-19 significantly reduced the chances of outdoor activities and lessen utilization of public transportation. Following this trend, we predict that the usage of private vehicles will increase exponentially, and utilization of public transportation will decrease dramatically. Accordingly, there will be a considerable amount of people choosing private vehicles as their top priority. Based on this assumption and the datasets we plan to use, we will analyze on several aspects including:

1) The counts of entry/exits data recorded by turnstiles in different boroughs

2) The counts daily riderships of New York City MTA Rail

3) The counts of bicycles that appears on the street

4) The daily counts of CitiBike trips and these trips' average duration

In order to find out how COVID-19 influenced these aspects of transportation, we plan to compare these data during the pandemic with those on previous year, we will calculate the differences and visualized together with COVID-19 cases counts.

3. Related Work

To further analyze the trend and effect we described, we dug into some research on datasets. We obtained some data on subway entries/exists based on NYC subway Turnstile counts, bicycle counts, and NYC CitiBike sharing situation. We also had the dataset of COVID-19 cases counts which has the counts by different boroughs in NYC. Here are the introductions of these datasets we found and how we were going to use them

1) The COVID-19 daily counts of cases include both the total new cases in NYC and in different boroughs including The Bronx, Manhattan, Brooklyn, Queens and Staten Island. We extracted these counts and calculated cumulative counts from them.

2) The NYC subway turnstile daily counts include the entries' and exits' counts in a specific station on a date. Furthermore, it also includes the borough the station belongs to.

3) The daily transit ridership includes riderships for agencies like New York City MTA Rail and San Francisco BART Rail. It also includes with the ridership count on same weekday of same week last year.

4) The bicycle counts record the bicycles conducted by the counters at key location around New York City from 2012. Since the counts were recorded every quarter of an hour, we need to aggregate to see the daily count.

5) The CitiBike datasets record every trips's information like duration and users' data and the trips were put in a same file if they are in a same month. Therefore, we will aggregate tens of files representing trips in different month

4. Data analysis and Result

In this part, we will explain how we processed the data and how we used the processed result to make comparisons by using Altair to visualize. Finally, we will make conclusions on the visualized results.

4.1 Data Cleaning

First, for the datasets we collected, the format of date included in are different. In order to make the analysis process simpler, the format of the date need to be unified. Therefore, we convert the format of dates in our scripts so that the dates are all in format of “yyyy-mm-dd” in the output generated by spark jobs. Furthermore, the transportation records are likely to be influenced by weekday. When we needed to join the date on different years, we first converted date to week's number and weekday and join the dates which has the same weekday.

In addition, there exist some fields in the dataset which does not include any data or include value like “NULL”. We filtered out these records so that they would not cause errors in further calculations.

4.2 Process the dataset of COVID-19 cases

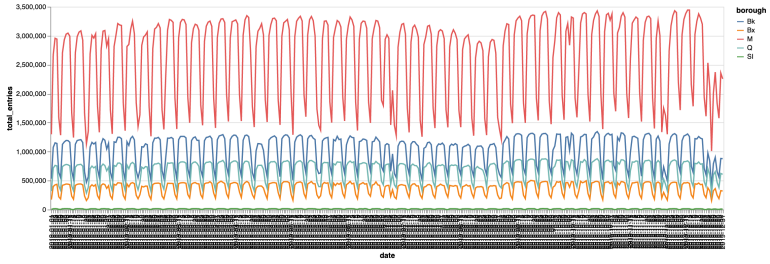
The COVID-19 daily counts of cases not only include the new cases counts but also includes some information we did not need such as the count of hospitalizations and deaths. Furthermore, we need to do aggregation of the daily new cases and new cases in different boroughs on date. Here we wrote two scripts where the first used sparkSQL because it is easier to accumulate values on dates by using sql queries. The second scripts convert the format of dates from “mm/dd/yyyy” to “yyyy-mm-dd”.

	date	new_case_count	cumulative_total	bx_case_count	bx_cumulative_total	bk_case_count	bk_cumulative_total	mn_cas
0	2020-02-29	1	1	0	0	0	0	
1	2020-03-01	0	1	0	0	0	0	
2	2020-03-02	0	1	0	0	0	0	
3	2020-03-03	1	2	0	0	0	0	
4	2020-03-04	5	7	0	0	1	1	
...	...	...	...	...	...	...	...	
273	2020-11-28	1687	297098	310	62710	418	84906	
274	2020-11-29	1596	298694	209	62919	474	85380	
275	2020-11-30	2444	301138	389	63308	777	86157	
276	2020-12-01	2325	303463	416	63724	636	86793	
277	2020-12-02	1399	304862	200	63924	405	87198	

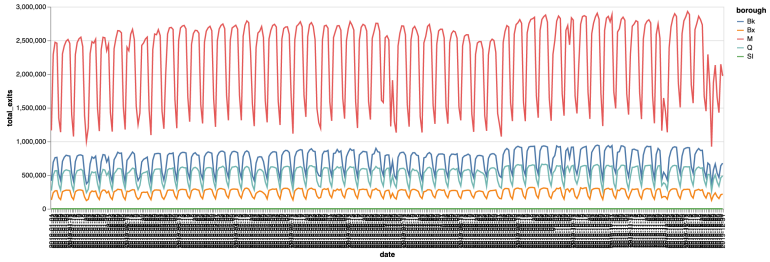
The processed covid-19 daily counts

4.3 Analyze Turnstile Count Data by Borough

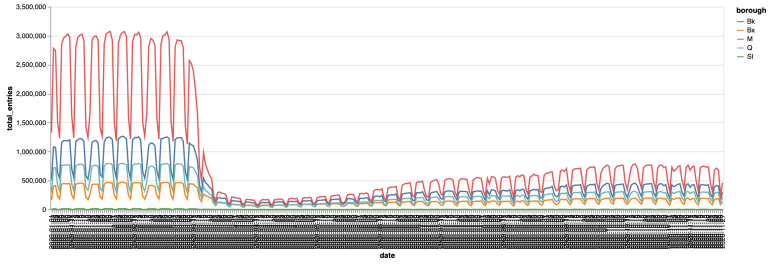
The NYC subway turnstile daily counts include the entries ’and exits ’counts in a station on a date, and we can find the borough a station belongs to in “borough” column. For every records in the dataset, we extracted columns including *(date, borough, entries, exits)*. Since each of the tuple represents the count in a station, we use *(date, borough)* as key and applied reduceBykey operation to aggregate the records which has same date and borough so we can get the results of the counts of total exits and entries in a borough in our output. We calculate the counts for both datasets of 2019 and 2020 and below are visualizations of these counts.



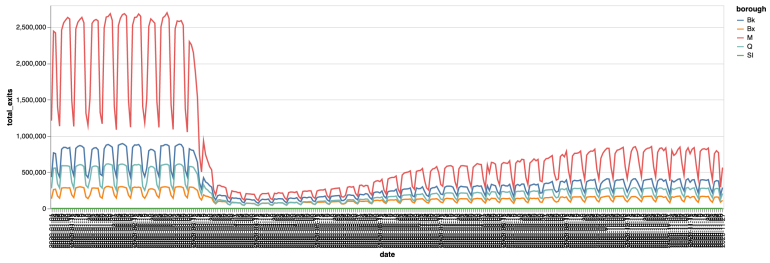
2019 Total entries for different borough



2019 Total exits for different borough



2020 Total entries for different borough



2020 Total exits for different borough

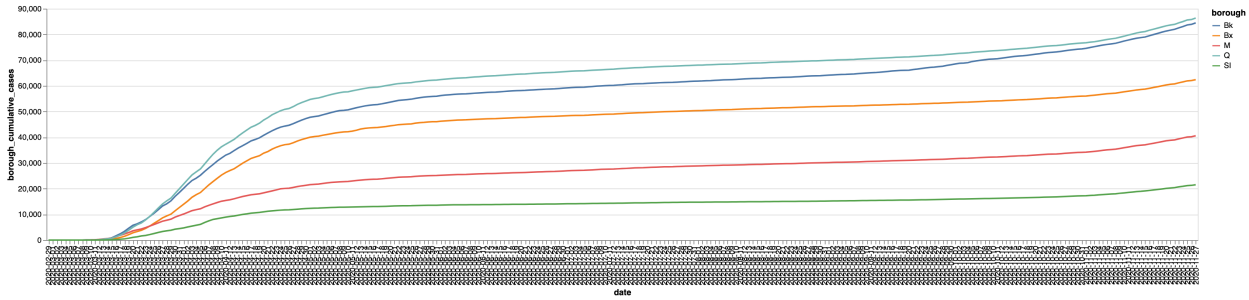
According to the visualized result, we can see that although different boroughs have different counts of entries and exits and Manhattan has the most counts, all boroughs' entries and exits counts falls drastically compare to those in 2019 after the "Stay at home" order. Furthermore, we can see the lines representing counts in different boroughs share same patterns so the COVID-19 pandemic conditions have same effects on passengers in different boroughs.

To find out the effect of COVID-19 cases on the turnstile usage in different boroughs, we performed join operations on the dataset of year 2020 we got from the previous section which includes the columns *(date, borough, entries, exits)* and the COVID-19 cases dataset which include detailed cases in every borough. For the COVID-19 cases dataset, we extract the two columns representing the cases in a borough and add an attribute which indicates the borough name so the processed RDD includes records like *(date, borough, borough_new_cases, borough_new_cases)*. We join the COVID-19 data with the dataset we generated in section 4.3.1 on columns *(date, borough)* and the records in the output include *(date, borough, borough_new_cases, borough_new_cases, total_entries, total_exits)*.

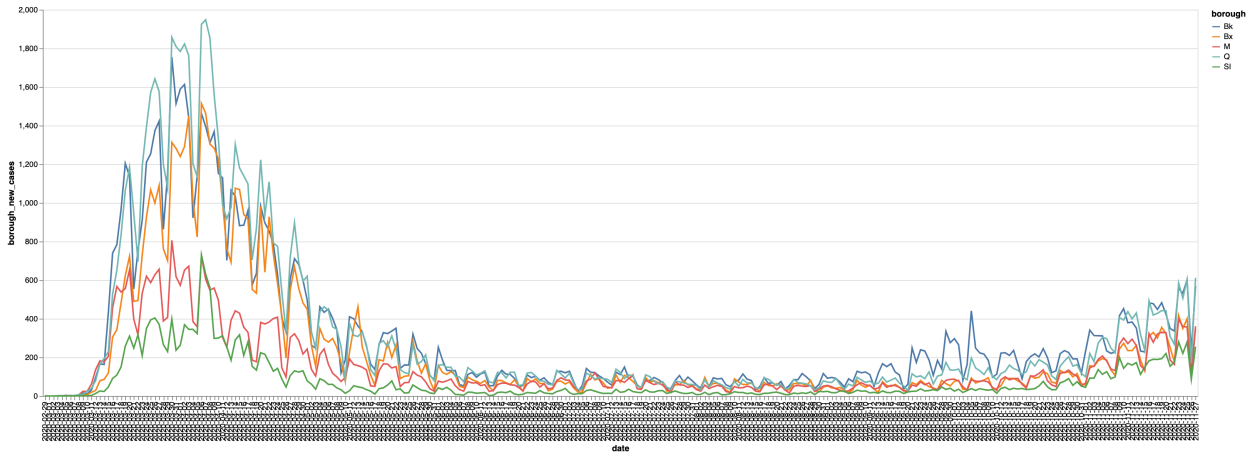
	date	borough	borough_new_cases	borough_cumulative_cases	total_entries	total_exits
0	2020-02-29	Bk	0	0	655107	514521
1	2020-02-29	Bx	0	0	242011	181514
2	2020-02-29	M	1	1	1520724	1428932
3	2020-02-29	Q	0	0	413221	324194
4	2020-02-29	SI	0	0	2448	0

Example of the output

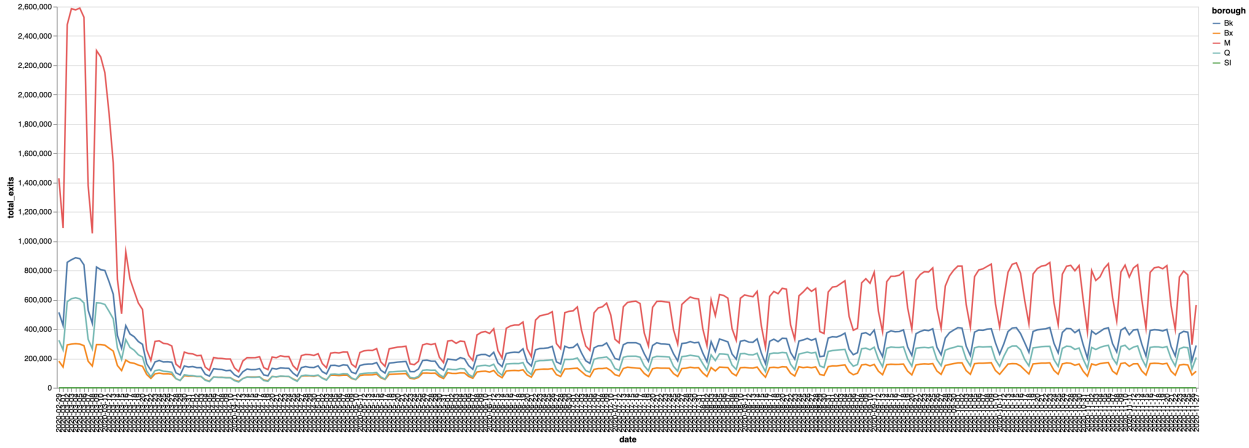
With the new dataset, we visualized on new_cases. cumulative_cases, total_entries and total_exits to see the connection between COVID-19 cases and turnstile usage.



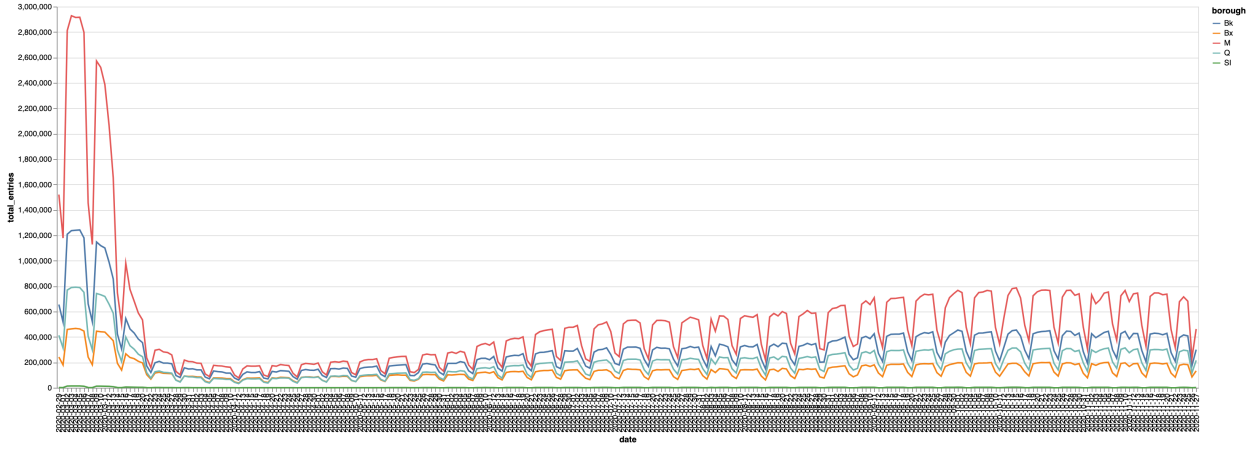
Graph of daily total cases in different borough



Graph of daily new cases in different borough

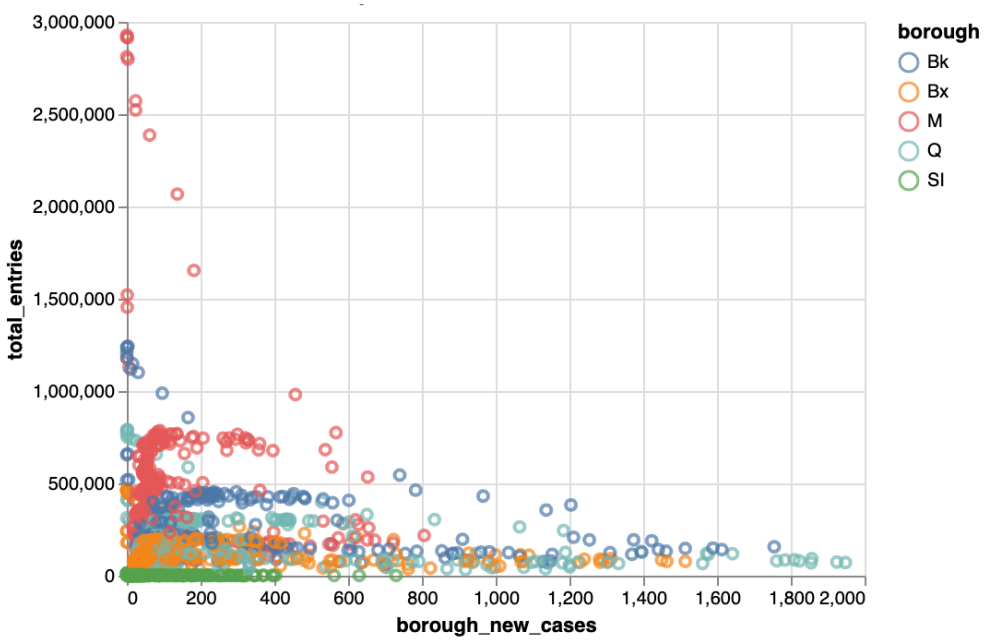


Graph of daily total exits in different borough



Graph of daily total entries in different borough

According to the graphs of the four values in different boroughs, we can get there exists connection between the daily new cases counts in a borough and the turnstile usage data such that if the daily COVID-19 cases counts increased drastically, the entries and exits counts also failed dramatically. However, we were unable to get the conclusion that if a borough had more turnstile entries/exits counts, it would have more daily new cases because although Manhattan had most turnstile usage, Queens had most COVID_19 cases.



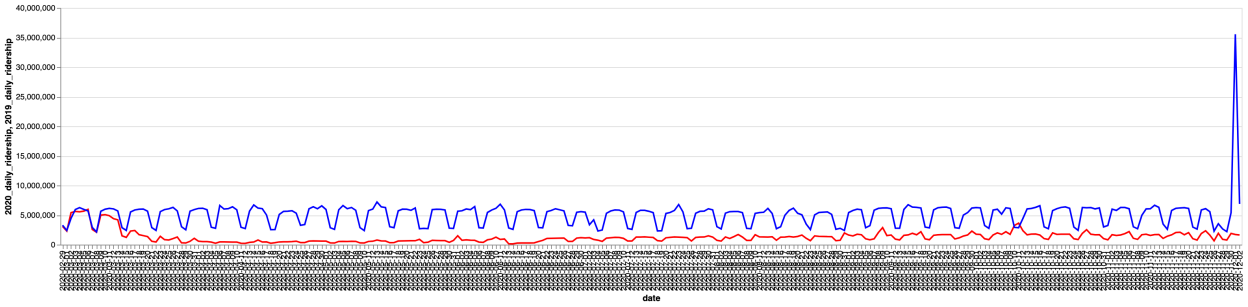
Total Entries vs Daily New Cases by borough

4.4 Analyze riderships (passenger count)

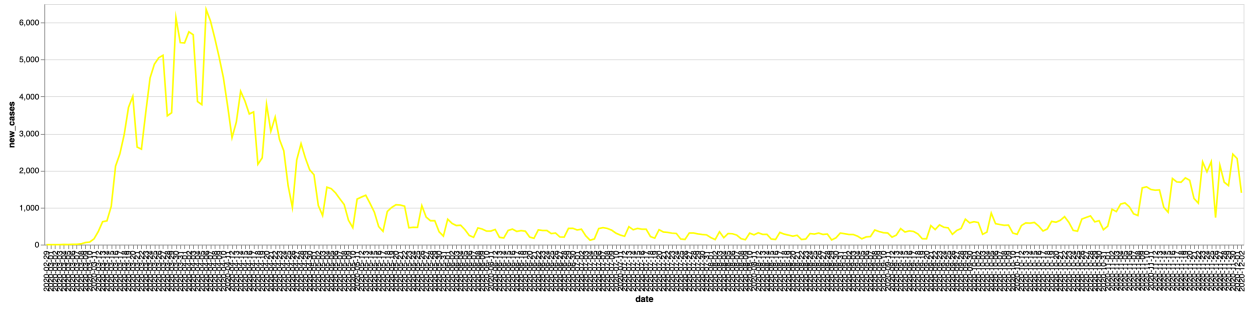
A record in the daily transit ridership dataset include columns *(Agency,Mode,Date,Week Number,Current,Baseline,Lowest)*. The values in agency attribute include “New York City MTA Rail”, “WMATA Bus and Rail” and “San Francisco BART Rail”. The “current” refers current date's ridership while the “baseline” refers to ridership of same weekday of same week in last year. Since we were only analyzing on NYC subway data, we filtered the records which have the agency of “New York City MTA Rail” and include current, baseline columns. Furthermore, we included the differences between a ridership count in an 2020 date with ridership of same weekday of same week in last year. Thus, we computed subtraction(2019_count-2020_count) and ratio(2020_count/2019_count). Finally, we join the above data with the COVID-19 cases counts so the final datasets include columns *(date, 2020_daily_ridership, 2019_daily_ridership, 2019-2020, 2020/2019, new_cases, cumulative cases)*.

	date	2020_daily_ridership	2019_daily_ridership	2019-2020	2020/2019	new_cases	cumulative_cases
0	2020-02-29	3082068	3286628	204560	0.94	1	1
1	2020-03-01	2295977	2400293	104316	0.96	0	1
2	2020-03-02	5411090	4449907	-961183	1.22	0	1
3	2020-03-03	5616179	5932869	316690	0.95	1	2
4	2020-03-04	5546138	6261445	715307	0.89	5	7
...	...	...	...	...	...	...	...
273	2020-11-28	903209	2663054	1759845	0.34	1687	297098
274	2020-11-29	783089	2195114	1412025	0.36	1596	298694
275	2020-11-30	1900762	5397098	3496336	0.35	2444	301138
276	2020-12-01	1722143	35551543	33829400	0.05	2325	303463
277	2020-12-02	1626512	6884032	5257520	0.24	1399	304862

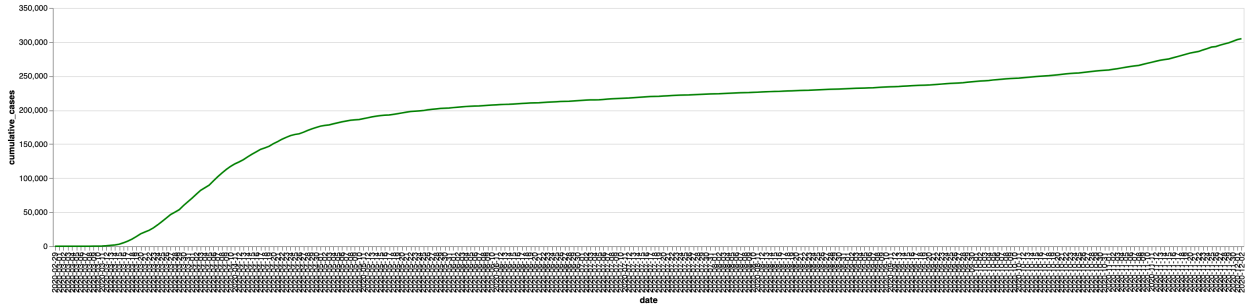
First of all, we draw two lines representing the riderships in 2020 and 2019 to see the difference of the total passengers. Since the red line represents the data in 2020 and the blue line represents the data in 2019, we can obtain the conclusion that the total ridership in a day decreased dramatically compare that in 2019 from this graph.



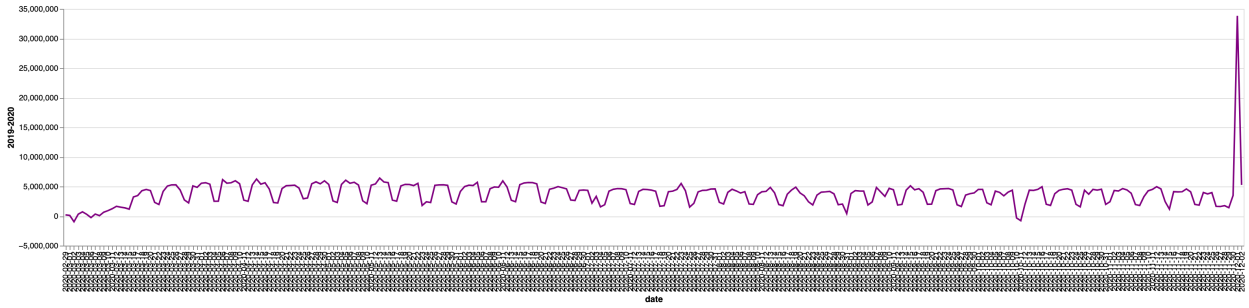
Next, we are going to compared the differences between the 2020 ridership count and 2019 ridership count with COVID-19 cases count. First, we drew the graphs of daily new cases, cumulative cases, 2019_count-2020_count and 2020_count/2019_count to roughly see the connections and then drew scatter plot graph to verify it.



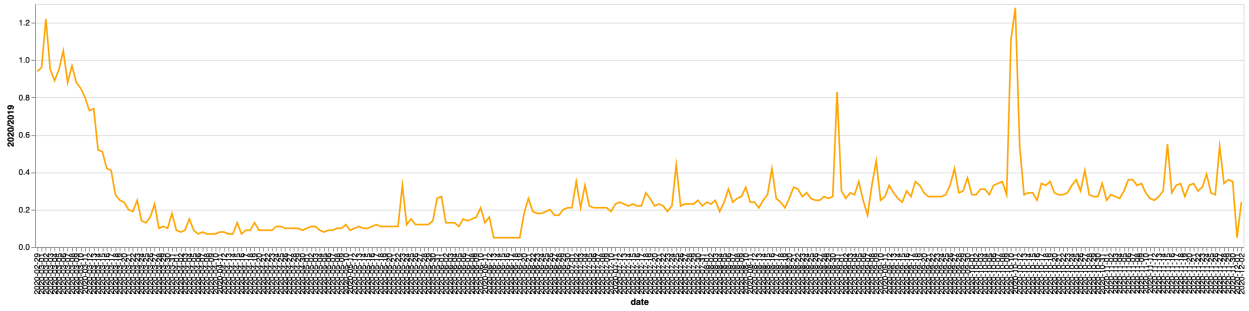
Daily new cases



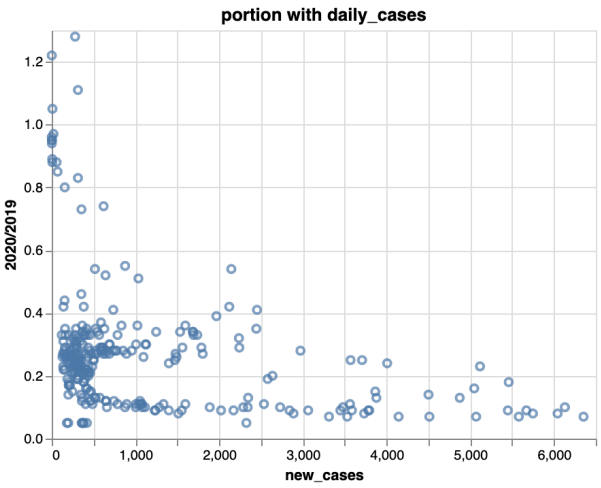
cumulative cases

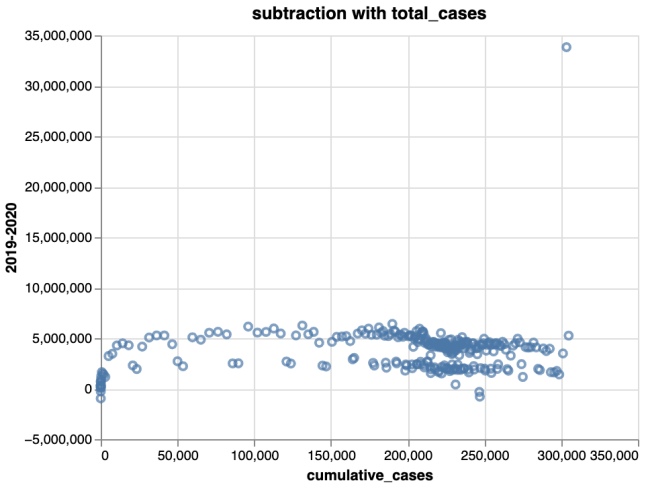


2019_count-2020_count



2020_count/2019_count



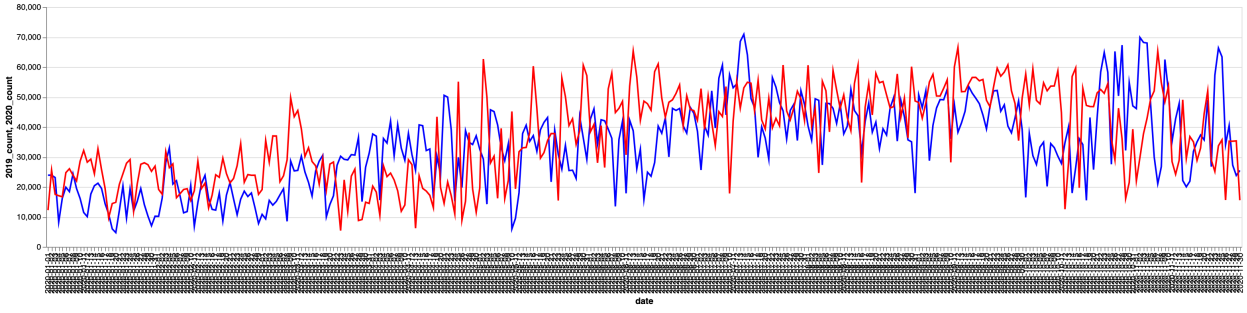


According to the line graph and scatter graph, we are able to get the conclusion that during the COVID-19 pandemic, when the count of new cases in a day is higher, the ridership in that date is less likely to as large as the day in previous year; Also, we can conclude from the second scatter plot that the subtraction of the ridership did not provide much useful information that indicates it was affected by COVID-19 pandemic.

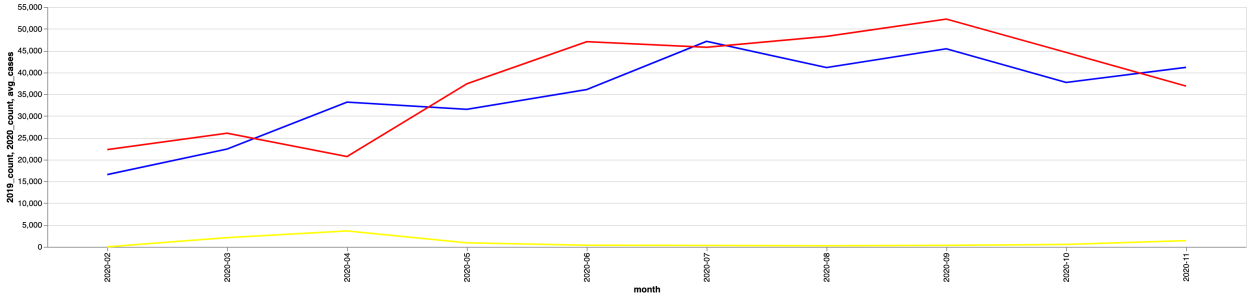
4.5 Analyze the bicycle counts recorded by counters in different locations

The bicycle counts is the count of bicycles that passes by during a period of time recorded by the counters at different location and the counts were recorded every quarter of an hour. We analyzed this dataset in order to approximate the bicycles on the street. We reduced the count on a same day to get the daily total bicycle counts. Furthermore, we also calculate the week number and weekday of the date when joining 2019 counts and 2020 counts in order to reduce the effect brought by different weekday.

	date	2020_count	2019_count
0	2020-01-01	12236	23911
1	2020-01-02	25598	23900
2	2020-01-03	17500	23122
3	2020-01-04	17066	8232
4	2020-01-05	16688	16165
...	...	...	...
330	2020-11-26	15609	34130
331	2020-11-27	35222	40465
332	2020-11-28	35231	27238
333	2020-11-29	35332	23805



According to the visualization of the daily counts in 2020 and 2019, it is hard to tell whether there were less bikes on the street during the COVID-19 pandemic. Therefore, we calculated average daily count each month to compare with the average daily new cases.



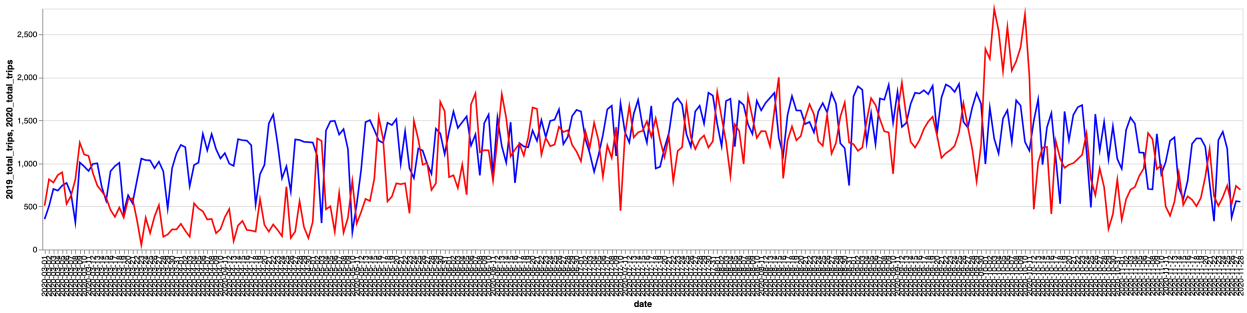
From the visualization of the average daily count on different month (2020 is represented by red line, 2019 is represented by blue line and average new cases is represented by yellow line), although on April, when the COVID-19 pandemic condition was most severe, the count of bicycle was less than that in 2019, it is still insufficient to support the idea that people ride less bikes due to the pandemic because the counts fluctuate a lot.

4.6 Analyze CitiBike trips count and duration

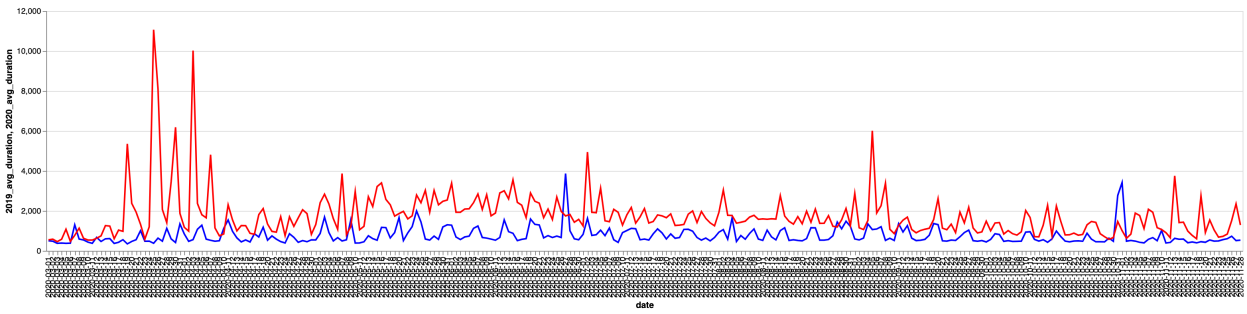
CitiBike is a public bicycle sharing system around NYC and it provides efficiency for people's transportation. We found the data from its official website which includes every trip's information each month. We are interested in analyzing this dataset because although CitiBike seems 'private' because the bike only has one rider, it is a publicly owned facility. Since the each month has a dataset, we need to combine multiple files into a new dataset. For the information recorded in the datasets, we only extract date and duration, which indicate a trip's start date and its duration. After we get the extracted columns, we calculate the count of total trips and average trip duration by date. In order to compare the data of 2020 with 2019, we performed join on the date with same weekday and week number.

	date	2020_total_trips	2020_avg_duration	2019_total_trips	2019_avg_duration
0	2020-03-01	510	552.09	354	504.86
1	2020-03-02	817	584.97	504	479.99
2	2020-03-03	779	459.68	704	375.16
3	2020-03-04	866	578.23	685	388.71
4	2020-03-05	898	1081.00	744	376.60
...	...	...	...	...	...
268	2020-11-24	610	724.56	1370	560.85
269	2020-11-25	743	837.97	1173	606.17
270	2020-11-26	528	1513.05	372	731.59
271	2020-11-27	740	2351.83	563	512.99
272	2020-11-28	694	1293.68	556	534.50

When we get the above dataset above, we drew graphs which shows counts and durations separately(red lines represent 2020, blue line represent 2019).



trips count



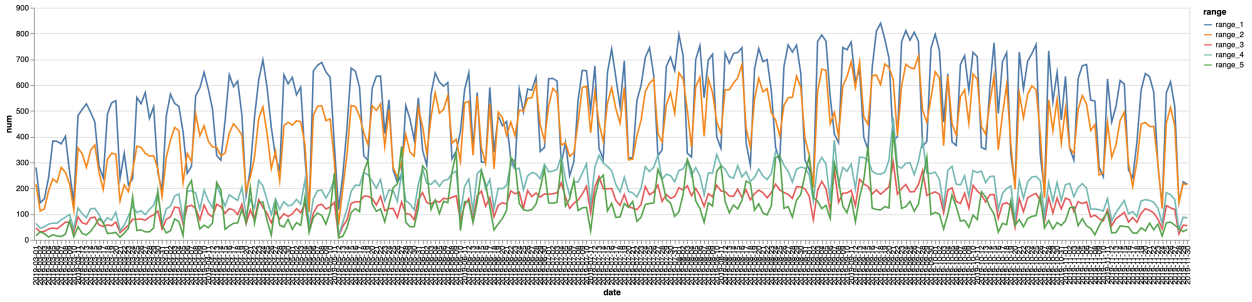
trips duration

According to the graphs, we can infer that the counts of trips during pandemic, especially on March and April, was less comparing to those in 2019. However, the counts were not obviously decreased after May. Also, compare with the difference in counts of trips, the difference of trip duration is more obvious, the average duration of trips during pandemic was longer.

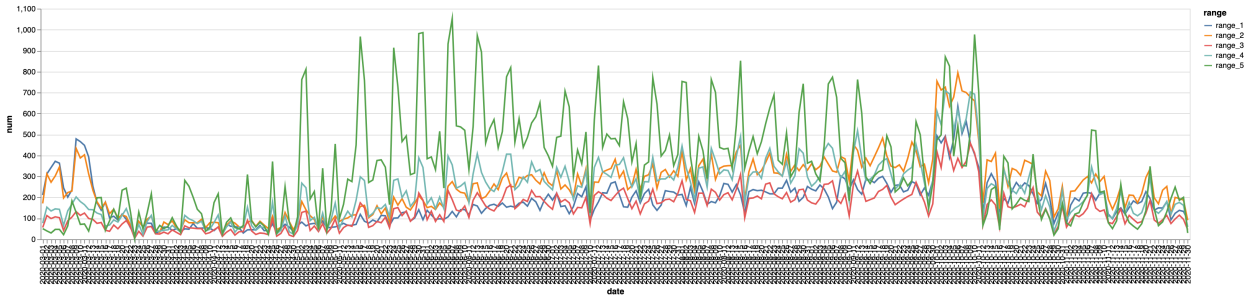
In order to analyze deeper in the duration of trips during pandemic, we divide the duration of trips into five ranges: [0, 300], [300, 600], [600, 900], [900, 1200], [1200,~] (in seconds). For each date, we counted the trips within different ranges and its ratio comparing to the total count.

	date	range	num	total	portion
0	2019-03-01	range_1	281	577	0.49
1	2019-03-01	range_2	216	577	0.37
2	2019-03-01	range_3	46	577	0.08
3	2019-03-01	range_4	64	577	0.11
4	2019-03-01	range_5	16	577	0.03
...	...	...	...	...	...

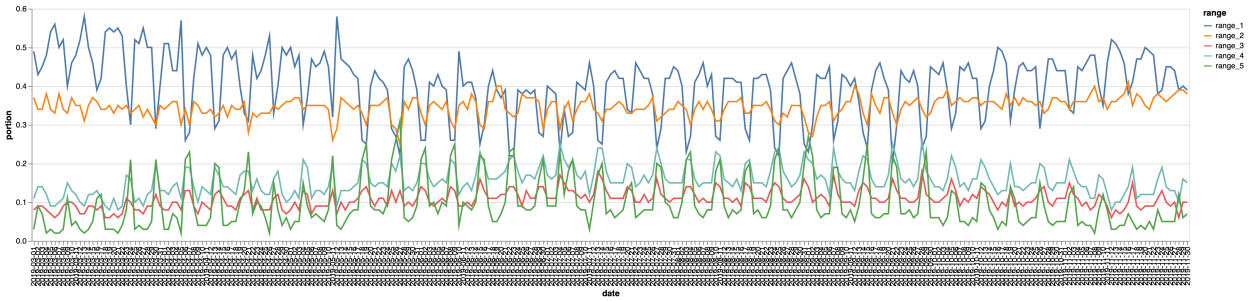
The visualization of 2019 and 2020's number of trips within different range:



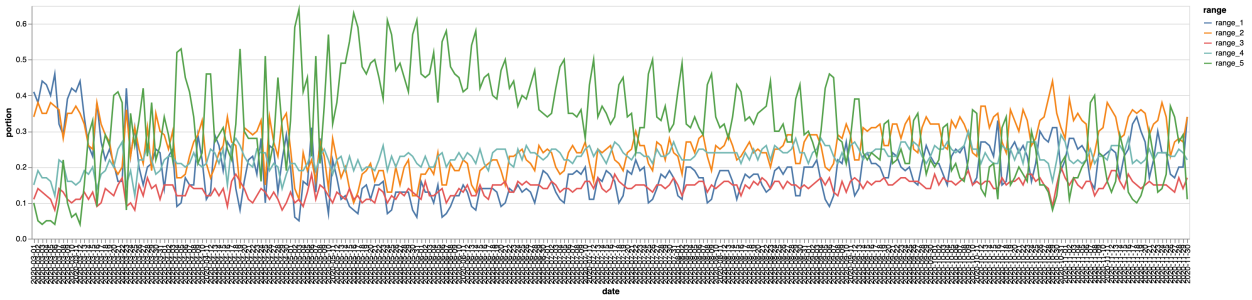
trips count 2019



trips count 2020



trips ratio 2019



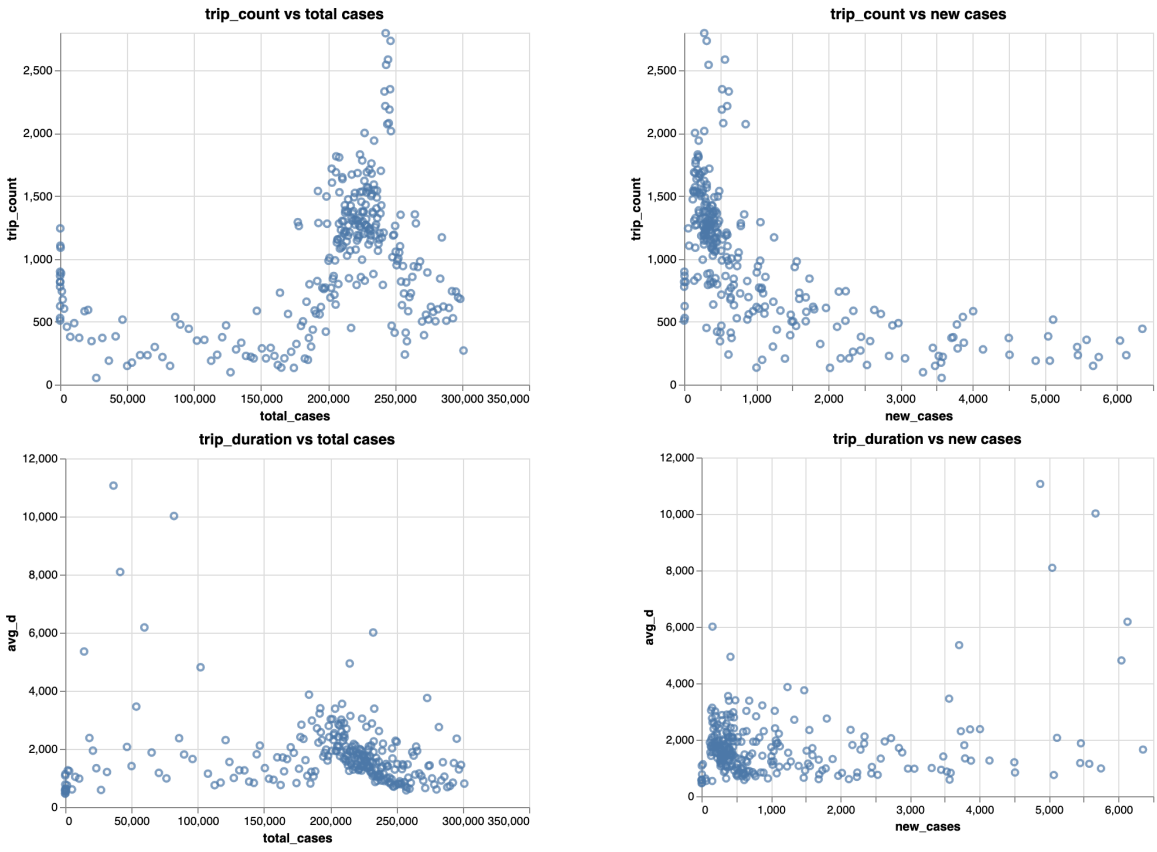
trips ratio 2020

According to the graphs above, we can make the conclusion that during 2019, most trips are within 0~600 seconds while during 2020, especially after May, the most trips are whose duration over 1200 seconds.

In order to analyze the effect of COVID-19 pandemic on CitiBike's counts of trips and trip duration, we join the dataset above with the COVID-19 dataset on date.

	date	trip_count	avg_d	new_cases	total_cases
0	2020-03-01	510	552.09	0	1
1	2020-03-02	817	584.97	0	1
2	2020-03-03	779	459.68	1	2
3	2020-03-04	866	578.23	5	7
4	2020-03-05	898	1081.00	3	10
...	...	...	...	...	...
270	2020-11-26	528	1513.05	730	293266
271	2020-11-27	740	2351.83	2145	295411
272	2020-11-28	694	1293.68	1687	297098
273	2020-11-29	682	1448.71	1596	298694
274	2020-11-30	271	806.05	2444	301138

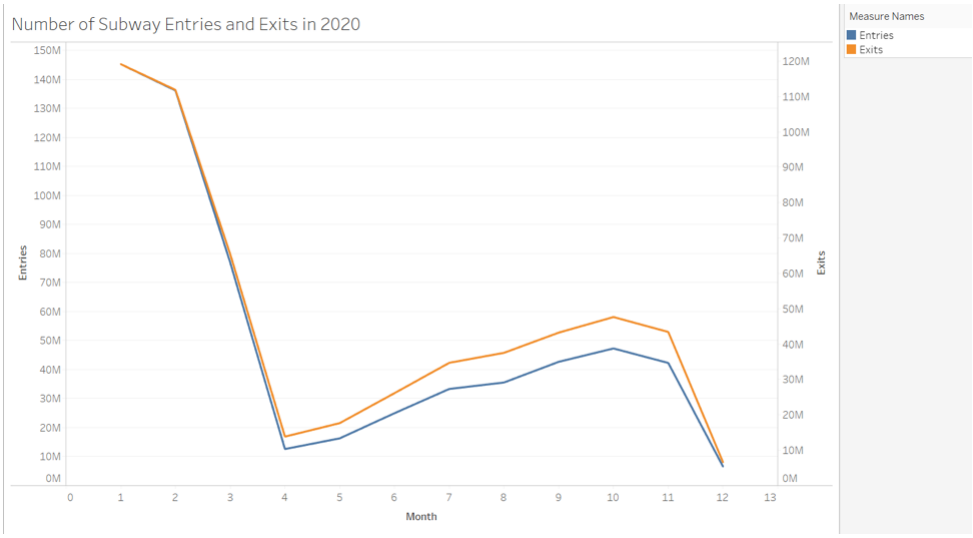
To find out the relationship between CitiBike trips and COVID-19 cases count, we made four scatter point plot:

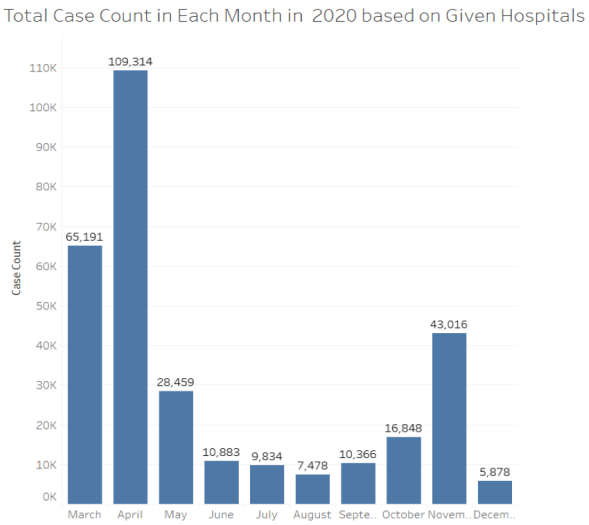


According to the scatter point plots above, we can make the conclusion that when the daily new cases are relatively low, especially on October, there are more people who used CitiBike. In contrast, when the daily new cases were high, there were less users of CitiBike, and the average trip duration tended to be higher.

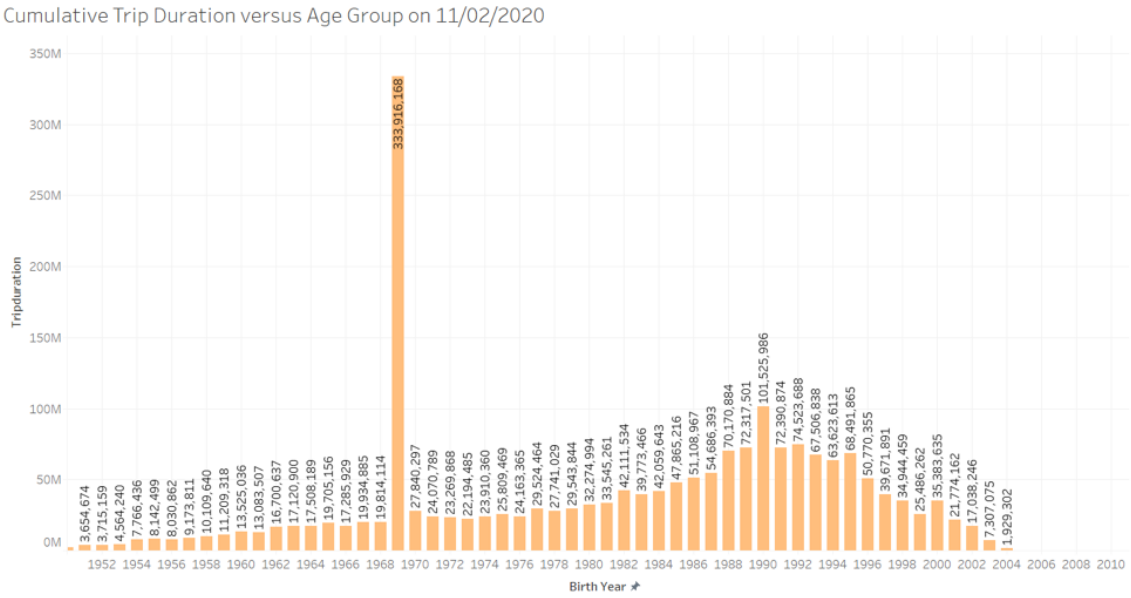
4.7 Key Findings and Challenges

One of the key findings from our analysis was that the number of ridership is directly proportional to the number of pandemic cases. From the line graph, it is notable that subway entries and exits decremented dramatically from March to April and then increased in a stable pattern from May to October. Likewise, revealing the total number of affect patients in each month based on given hospitals, the bar graph shows a similar trend that number of cases reached 109,314 in April but significantly dropped down to 28,459 in May. Thus, it seems that less ridership will lead to less affected residents.





As we have analyzed bikes are most people's choices when there were necessary outdoor events, and the duration of bike trip increased significantly as the pandemic situation evolved; another interesting finding was that the age group that dominated cumulative bike trip duration was around 50 years old according to the bar graph. One of our assumptions was that people in this age group would have to go to grocery store to buy and store food and daily necessities more frequently than any other age groups through biking.



There were many challenges during our analysis. As we looked through many tables and datasets, tables' attributes or titles were written in abbreviate form, and there were no unites for numerical values listed in the table, so it was hard to read and interpret. In addition, there were some similar datasets and tables that did not share common attributes and had different data qualities, so we had to look for additional datasets to compare the statistics.

5. Conclusion

Based on our data analysis, we found that these data proved our hypothesis that the utilization of public transportation will reduce significantly. For example, in section 4.3, the most popular borough Manhattan's total entries reached to 2,255,999 at its peak in December of 2019 but the total entries of Manhattan during the same period in 2020 only had 462,696, which means the total entries of M was reduced by 79.5 percent comparing the previous year. In addition, we extracted the data on ridership in 2019 and 2020 in section 4.4, which revealed huge difference of the daily ridership. For instance, the daily ridership on 12/04/2019 was 8,227,994 while the

daily ridership on the same day in 2020 was 1,609,192 which was nearly one fifth of that ridership in 2019.

Furthermore, we visualized the data about bicycle counts in each day regrading to 2019 and 2020. However, the counts fluctuated and there was no obvious decreases in 2020's counts. Although some months in 2020 had less counts, there were other months that had higher counts than those in 2019.

However, for the CitiBike, which is a public bike system, we can see some changes due to the pandemic. In Section 4.6, we computed total amount of trips and average duration. Also, we divide trips into groups by their trip duration. During the period of Stay at Home Order was declared, which is March and April, the counts of trips in 2020 decreases obviously and the ratio of trips within different duration ranges varied while on May, the counts were similar to those in 2019 and the duration of bicycle usage increased rapidly in terms of our data visualization compare to those in 2019. We inferred from the result that when when people need to ride a long trip, they would be more likely to choose CitiBike.

Thus, it is obvious that people chose private vehicle over public transportation due to the impact of pandemic in terms of duration and counts. This analysis also presents evidence of the possible contribution of bike sharing systems to a more resilient transport system, as it can quickly provide alternative transport options to NYC residents; since users are the ones who define the routes, bikes could reinforce the transport offer to the areas with higher demand. Under the high mobility demands and requires, private vehicle is more flexible and resilient in urban transportation system.

6. Reference

COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths: Dataset Published on qri.cloud. (n.d.). Retrieved December 10, 2020, from https://qri.cloud/nyc-open-data-archive/covid 19daily-counts-of-cases-hospitalizations-and-deaths

NYC Subway Turnstile Counts - 2020: Dataset Published on qri.cloud. (n.d.). Retrieved December 10, 2020, from https://qri.cloud/nyc-transit-data/turnstile_daily_counts_2020

NYC Subway Turnstile Counts - 2019: Dataset Published on qri.cloud. (n.d.). Retrieved December 10, 2020, from https://qri.cloud/nyc-transit-data/turnstile_daily_counts_2019

Daily Transit Ridership | Open Data | Socrata. (2020, December 9). Daily Transit Ridership. https://data.bts.gov/Transit/Daily-Transit-Ridership/dc74-f8qd

Bicycle Counts | NYC Open Data. (2020, December 4). Bicycle Counts. https:// data.cityofnewyork.us/Transportation/Bicycle-Counts/uczf-rk3c

Citi Bike Trip Histories | citibikenyc. (2020, December 4). CitiBike Trip Data. https:// www.citibikenyc.com/system-data

Teixeira, J. F. and Lopes, M. (2020) ‘The link between bike sharing and subway use during the COVID-19 pandemic: The case-study of New York’s Citi Bike’, Transportation Research Interdisciplinary Perspectives, 6. doi: 10.1016/j.trip.2020.100166.