Subway Ridership in New York City during the Coronavirus Pandemic

Team Members: Kathryn Buckwalter, Julissa Guzman, Patrick Saitta, Eddie Xu, Mackenzie Baucum & Genevieve Sloup

Motivation & Summary

Mass Transit ridership in New York City was severely impacted by Governor Cuomo's March 20 "NY on PAUSE" order, which mandated all non-essential workers stay home to contain the spread of the Corona Virus Disease.

The scope of this project is to analyze and quantify how the COVID-19 pandemic affected usage of subways in New York City.

Questions & Data

We sought to answer the following questions:

How was total ridership affected by comparing Ridership in 2020 to ridership in 2019?

Was there more of an impact on ridership at “Transit Hubs” such as Penn Station compared to stations located in more residential areas such as the Upper East Side?

Did areas established as having higher rates of positive COVID-19 cases see a bigger change in ridership?

What conclusion can be drawn about the relationship between the rate of positive COVID-19 cases and ridership during the last week of March?

Questions & Data (cont.)

We obtained Data from the following sources:

New York City Health Department

NYC Coronavirus Disease 2019 (COVID-19) obtain by modified ZIP code tabulation areas (ZCTA)

Data containing daily count of NYC residents who tested positive for SARS-CoV-2

The Metropolitan Transportation Authority:

Turnstile Entries & Exit Data for All Subway Stations from 2019 and 2020

Data Cleanup & Exploration

Problems with the data

- Turnstile entry and exit data was cumulative from previous years and by individual turnstile ID

SCP	STATION	LINENAM	DIVISION	DATE	DAY	TIME	DESC	ENTRIES	EXITS
00-00-00	34 ST-PENN STA	123ACE	IRT	3/23/19	SUNDAY	2:00:00	REGULAR	9465056	0005416433
00-00-00	34 ST-PENN STA	123ACE	IRT	3/23/19	SUNDAY	6:00:00	REGULAR	9465117	0005416435
00-00-00	34 ST-PENN STA	123ACE	IRT	3/23/19	SUNDAY	10:00:00	REGULAR	9465718	0005416483
00-00-00	34 ST-PENN STA	123ACE	IRT	3/23/19	SUNDAY	14:00:00	REGULAR	9466554	0005416550
00-00-00	34 ST-PENN STA	123ACE	IRT	3/23/19	SUNDAY	18:00:00	REGULAR	9467406	0005416643
00-00-00	34 ST-PENN STA	123ACE	IRT	3/23/19	SUNDAY	22:00:00	REGULAR	9468117	0005416711
00-00-00	34 ST-PENN STA	123ACE	IRT	3/24/19	MONDAY	2:00:00	REGULAR	9468597	0005416757
00-00-00	34 ST-PENN STA	123ACE	IRT	3/24/19	MONDAY	6:00:00	REGULAR	9468691	0005416764
00-00-00	34 ST-PENN STA	123ACE	IRT	3/24/19	MONDAY	10:00:00	REGULAR	9469077	0005416805
00-00-00	34 ST-PENN STA	123ACE	IRT	3/24/19	MONDAY	14:00:00	REGULAR	9469883	0005416872
00-00-00	34 ST-PENN STA	123ACE	IRT	3/24/19	MONDAY	18:00:00	REGULAR	9470931	0005416976

Data Cleanup & Exploration

We had to calculate total entries per station by finding the first and last count for each turnstile, using groupby, min and max.

```
#found the lowest cumulative entries count based on station, line name, SCP for March 2019
march_2019_min = newdf_2019.groupby(["STATION", "LINENAME", "SCP"]).min()

march_2019_min
```

STATION	LINENAME	SCP	DATE	DAY	TIME	DESC	ENTRIES	EXIT
86 ST-2 AVE	Q	00-00-00	3/23/19	FRIDAY	13:00:00	REGULAR	3619381.0	0002134693 ...
		00-00-01	3/23/19	FRIDAY	13:00:00	REGULAR	2648355.0	0001167294 ...
		00-00-02	3/23/19	FRIDAY	13:00:00	REGULAR	2122526.0	0001074938 ...
		00-00-03	3/23/19	FRIDAY	13:00:00	REGULAR	1309084.0	0000941007 ...
		00-05-00	3/23/19	FRIDAY	13:00:00	REGULAR	36570002.0	0050331652 ...

Data Cleanup & Exploration (cont.)

- Some of the data collected was incorrectly captured - possible turnstile malfunction.
    - We selected several stations we identified as having viable data to conduct our analysis

```
#created a new DF with targeted neighborhoods in 2020
newdf_2020 = march_2020_df.loc[(march_2020_df["STATION"] == "E 180 ST") |
                               (march_2020_df["STATION"] == "86 ST-2 AVE") |
                               (march_2020_df["STATION"] == "96 ST") |
                               (march_2020_df["STATION"] == "CITY HALL") |
                               (march_2020_df["STATION"] == "ATL AV-BARCLAY") |
                               (march_2020_df["STATION"] == "W 4 ST-WASH SQ") |
                               (march_2020_df["STATION"] == "BROADWAY JCT") |
                               (march_2020_df["STATION"] == "JKSN HT-ROOSVLT") |
                               (march_2020_df["STATION"] == "WORLD TRADE CTR") |
                               (march_2020_df["STATION"] == "QUEENSBORO PLZ")]
newdf_2020
```

	SCP	STATION	LINENAME	DATE	DAY	TIME	DESC	ENTRIES	EXIT
8032	00-00-00	CITY HALL	NRW	3/21/20	SUNDAY	0:00:00	REGULAR	5524.0	5348.0
8033	00-00-00	CITY HALL	NRW	3/21/20	SUNDAY	4:00:00	REGULAR	5524.0	5350.0
8034	00-00-00	CITY HALL	NRW	3/21/20	SUNDAY	8:00:00	REGULAR	5524.0	5352.0
8035	00-00-00	CITY HALL	NRW	3/21/20	SUNDAY	12:00:00	REGULAR	5527.0	5357.0
8036	00-00-00	CITY HALL	NRW	3/21/20	SUNDAY	16:00:00	REGULAR	5536.0	5365.0

Data Cleanup - Total Ridership YOY

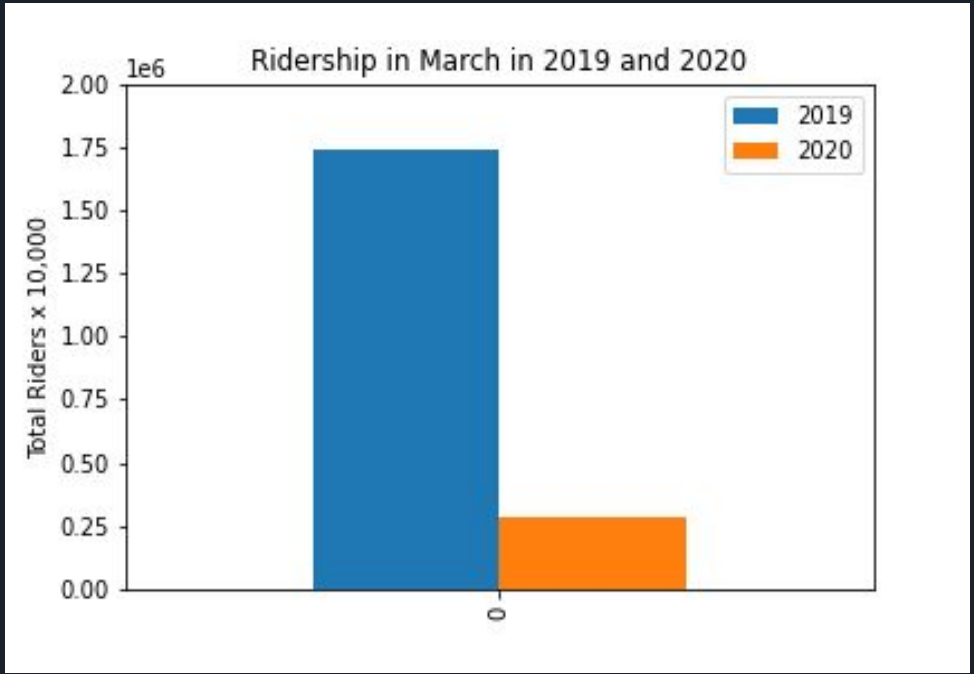
We merged 2019 and 2020 dataframes and added a column to calculate change.

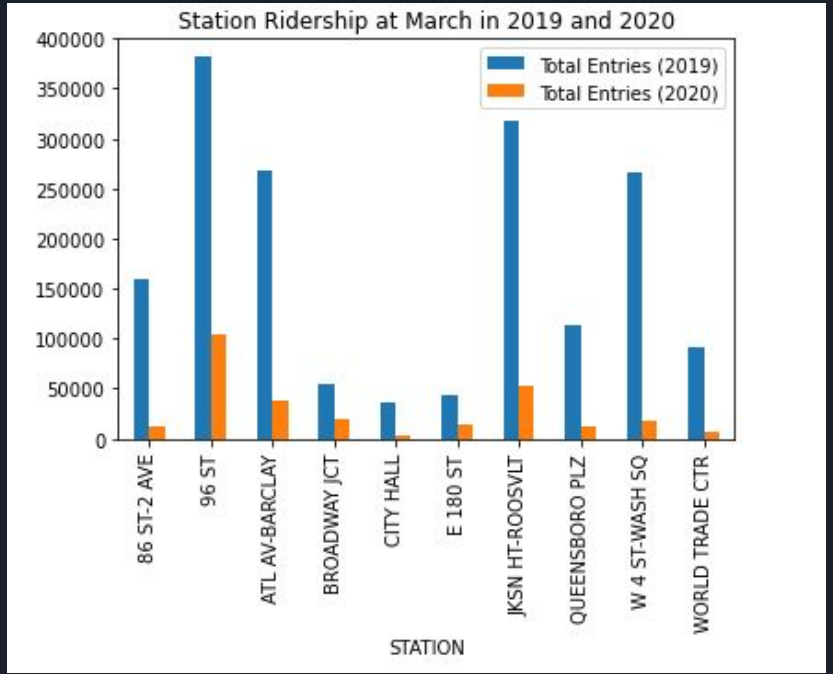
```
#added a new column with the entry count per day for March 2020
march_2020_merge["ENTRIES CHANGE"] = march_2020_merge["ENTRIES_x"] - march_2020_merge["ENTRIES_y"]
march_2020_merge = march_2020_merge.reset_index()
march_2020_merge.head()
```

	STATION	SCP	DATE_x	DAY_x	TIME_x	DESC_x	ENTRIES_x	EXIT_x	DATE_y	DAY_y	TIME_y	DESC_y	ENTRIES_y	EXIT_y	ENTRIES CHANGE
0	86 ST-2 AVE	00-00-00	3/27/20	WEDNESDAY	9:00:00	REGULAR	5302061.0	2988082.0	3/21/20	FRIDAY	13:00:00	REGULAR	5298209.0	2984796.0	3852.0
1	86 ST-2 AVE	00-00-01	3/27/20	WEDNESDAY	9:00:00	REGULAR	3876825.0	1629249.0	3/21/20	FRIDAY	13:00:00	REGULAR	3875060.0	1627996.0	1765.0
2	86 ST-2 AVE	00-00-02	3/27/20	WEDNESDAY	9:00:00	REGULAR	3073461.0	1544023.0	3/21/20	FRIDAY	13:00:00	REGULAR	3072294.0	1542951.0	1167.0

Total Ridership Year Over Year

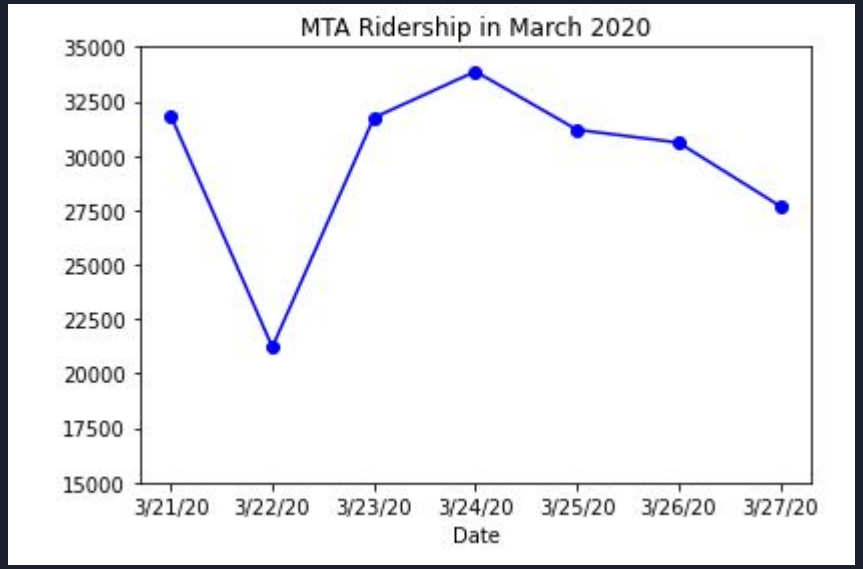
The following graphs demonstrate the overall change in turnstile entries in 2019 vs 2020 of all subway stations and selected stations identified as a transit hubs and residential area stations.

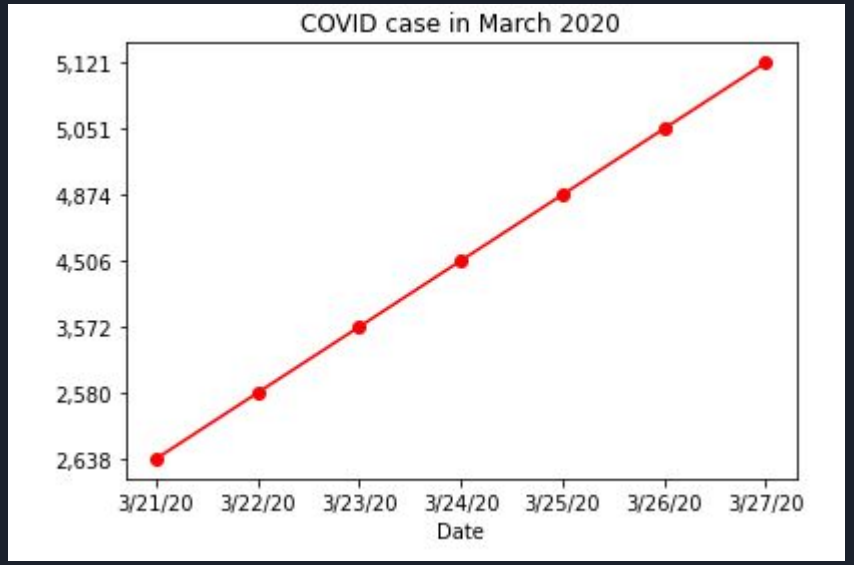




Ridership and COVID Cases (3/21/20 - 3/27/20)

The following graphs depict the decrease in subway ridership, and increase in rate of positive COVID-19 during the last week of March 2020.





“Hotspots” vs. Areas with Lowest COVID-19 Rates

We identified the neighborhoods having the highest and lowest COVID rates in New York City, and to examine if there was any relationship between COVID rates and change in subway ridership..

We extracted the necessary columns from the NYC Coronavirus Data and sorted by ascending and descending. Then cross referenced that with subway maps and selected our stations to analyze.

```
#sorted the covid case rate from the highest to lowest
covid_case_rate = covid_case_rate.sort_values(by=["COVID_CASE_RATE"], ascending = False)

covid_case_rate.head()
```

	Modified_ZCTA	NEIGHBORHOOD_NAME	BOROUGH_GROUP	COVID_CASE_RATE
140	11369	Airport/East Elmhurst	Queens	4837.88
125	11239	East New York	Brooklyn	4688.49
139	11368	Corona/North Corona	Queens	4627.48
74	10469	Allerton/Baychester/Pelham Gardens/Williamsbridge	Bronx	4602.27
142	11372	Jackson Heights	Queens	4561.81

```
#pulled 4 neighborhoods with the highest covid cases
covid_case_max = covid_case_rate.nlargest(4, "COVID_CASE_RATE")
print(covid_case_max.shape)

covid_case_max.head()
```

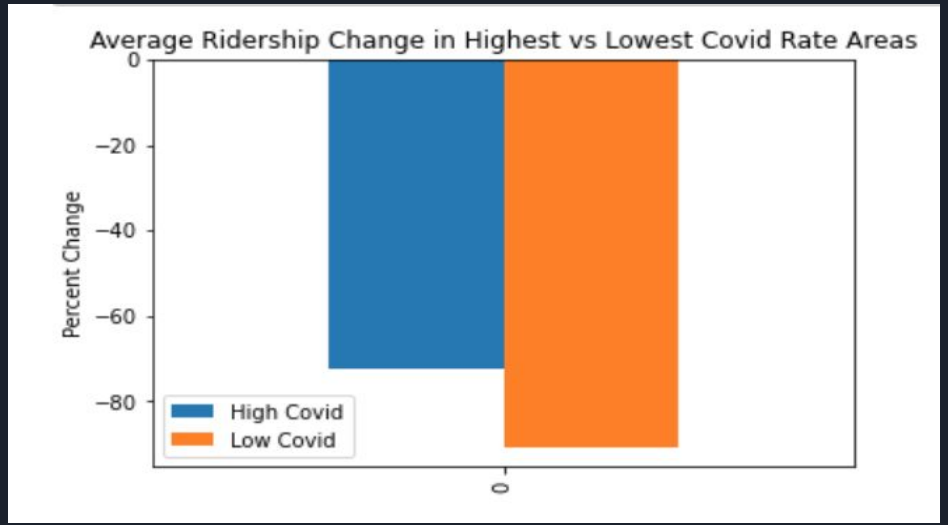
```
(4, 4)
```

	Modified_ZCTA	NEIGHBORHOOD_NAME	BOROUGH_GROUP	COVID_CASE_RATE
140	11369	Airport/East Elmhurst	Queens	4837.88
125	11239	East New York	Brooklyn	4688.49
139	11368	Corona/North Corona	Queens	4627.48
74	10469	Allerton/Baychester/Pelham Gardens/Williamsbridge	Bronx	4602.27

“Hotspots” vs. Areas with Lowest COVID-19 Rates

We calculated the percent change at each subway station, and then the average percent change in turnstile entries from 2019 to 2020.

Of the areas sampled, we found that areas with the highest rates of covid saw less of a reduction in ridership than areas with lowest, at about 72% and 91% respectively.



“Transit Hubs” vs Residential Stations

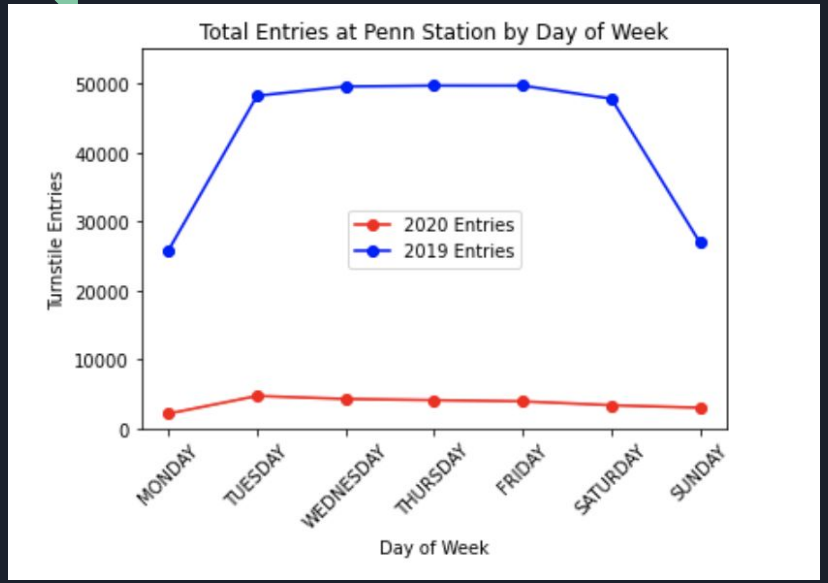
We selected the turnstiles at Penn Station and 86th Street/2nd Avenue on the Upper East side to analyze if there was a difference in ridership in a transit hub vs a residential neighborhood. After calculating the total change at each turnstile, we grouped by Station and Day to begin building our chart.

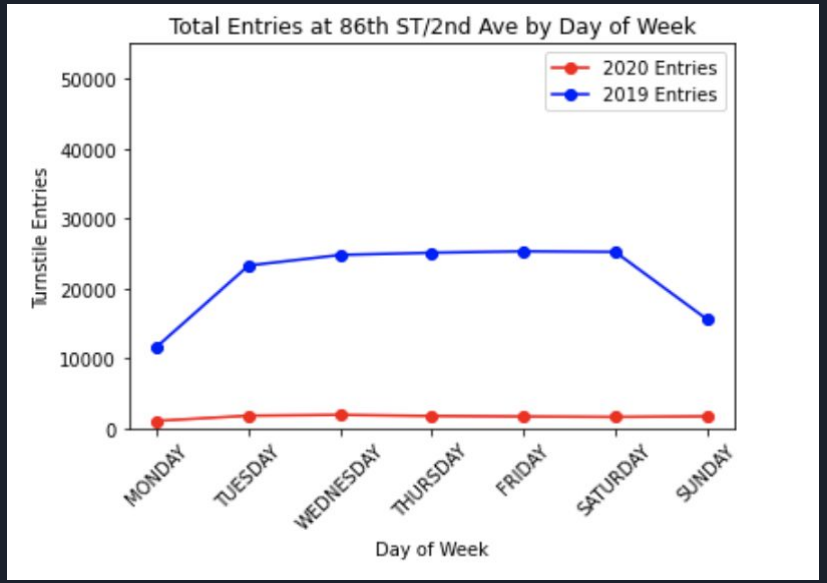
```
# Create df calculating the total change in entries per day of week per station
DOW2019_change = DOW2019_merge.groupby(["STATION","DAY"]).sum()["ENTRIES CHANGE"].rename("TOTAL ENTRIES")
DOW2019_total = pd.DataFrame(DOW2019_change)
DOW2019_total
```

		TOTAL ENTRIES
STATION	DAY	
34 ST-PENN STA	FRIDAY	49652.0
	MONDAY	25805.0
	SATURDAY	47757.0
	SUNDAY	26911.0
	THURSDAY	49662.0
	TUESDAY	48187.0
	WEDNESDAY	49532.0

“Transit Hubs” vs Residential Stations

The following graphs compare 2019 and 2020 for two stations: Penn Station and 86th - 2nd Ave Station. It demonstrates that weekday ridership was impacted more than weekday, with greater impact at Penn.






```
# Use loc to fix the order of days of week for x axis
field = "DAY"
day_order = ["MONDAY", "TUESDAY", "WEDNESDAY", "THURSDAY", "FRIDAY", "SATURDAY", "SUNDAY"]
Penn_sort_df = Penn_DOW_df.set_index(field).loc[day_order]
Penn_sort_df = Penn_sort_df.reset_index()
Penn_sort_df
```

Discussion/Summary of Analysis

Overall ridership of MTA subways decreased in all stations.

Stations in neighborhoods with lower rates of positive COVID-19 rates had significantly higher decrease in ridership

Stations deemed to be “Transit Hubs” were determined to have a higher percent change in ridership compared to stations in residential areas.

As the rate of positive COVID-19 cases continued to rise in the last week of March, subway ridership decreased.

Post Mortem (Future Analysis)

Due to time and data constraints, we were only able to conduct a limited analysis. Ideally, in the future, we would like to expand our analysis to include:

- Expand our analysis to include all stations, and account for misrepresented turnstile entry and exit data
- How ridership was affected during different times of day - Peak vs Off Peak Hours?
- Examine more weeks of data
- Assess MTA revenue and ridership
- Assess a relationship between subway ridership and e-hail taxi services like Lyft and Uber during the COVID-19 pandemic.

Questions?