Subway Insights: Demographics, COVID-19, and the Future of NYC Transit

Jeanie Liu, Kazuma Parkinson, and Jonathan Gotian

Preface

We are a consulting firm whose client is New York City Metropolitan Transportation Authority (MTA). They have three areas of interest: (1) What are the current subway traffic patterns in NYC as a whole and across neighborhoods? How does demographics affect subway demand? Has demand been impacted by COVID? (2) What will the future demand for the subway be? Based on future demand, which stations and subway carts should New York City invest in for repairs? (3) New York City is also interested in building city-owned small convenient stores outside of subway stations as a way of generating revenue. They wonder: What type of stores should they build outside of different subway stations?

Executive Summary

In order to best understand subway patterns, we must analyzed the subway turnstile data from New York City subway stations which capture the total amount of entrances and exits through the subway stations. We analyzed this data to understand the frequency to which people use New York City subway stations and how the MTA should best respond to this demand. Specifically, prior to COVID-19, subway use is found to be higher during Fall and Spring months than in Winter and Summer months. In the beginning months of the pandemic, subway use dramatically fell and has since has a slow but upwards trajectory. Lastly, it was found that on a station by station basis, the same stops (Union Sq, Columbus Circle, Herald Sq and Penn Station) all yielded among the top seven of highest entrances over the past five years.

We then narrowed down our analysis to the neighborhood level. We identified the top 10 neighborhoods for different demographic patterns to inform New York City about where to invest in convenient stores and what offerings should the stores provide to satisfy consumer tastes.

We also tried to predict the median subway entries of each neighborhood with different demographic variables like household income, median rent, and neighborhood population. We found that only neighborhood population is a significant predictor, so New York City Subway should focus on neighborhoods with high populations like Flushing and Washington Heights in adding subway stops or increasing maintenance.

Lastly, we focused on the impact of COVID-19 on subway traffic. We found that the COVID lockdown in March, 2020 significantly decreased the subway traffic, but it has been increasing steadily since the drop. To predict future subway traffic, New York City Subway should seek variables besides covid cases as it is a relatively poor predictor of subway entries.

Data

The main dataset we are analyzing is "NYC_subway_traffic_2017-2021.csv ("https://www.kaggle.com/datasets/eddeng/nyc-subway-traffic-data-20172021?resource=download). This dataset provides data on the number of people who enter and exit each subway station through turnstiles at 4-hour intervals between

2017 and 2021. Within the above link, it also has NYC neighborhood census data of 2020. Examples of relevant metrics include percentage of car-free commute, neighborhood population, poverty rate, and percentages of different races. We could utilize this dataset to see how demographics affect subway traffic. The "Stations.csv" file further informs us about all the subway stations' locations, IDs, lines they belong to, and other related information. In addition, we also found a dataset on COVID: COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths (https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3). This dataset provides daily counts of COVID cases, hospitalizations, and deaths since 2020/02/29 to now. At the time we downloaded the data, the last date with data was 2022/05/10.

Data Cleaning Process

We merged our neighborhood census data with our subway traffic data. When doing this, our merged dataset kept 51 of our 59 neighborhoods. This is likely because 8 neighborhoods do not have subway stations. The subway traffic dataset did not contain Staten Island, but the subway stations dataset did. Due to this inconsistency, we removed Staten Island from our analyses. When examining the subway entry and exit data, we discovered significant differences between entries and exits. This was likely due to multiple ways of being able to exit the subway. We thus focused our analysis on factors which predict subway entries. One factor we considered was COVID. To analyze the potential impact, we merged our COVID dataset with the subway traffic data, joining by borough. This merged dataset only contained 4 of the 5 boroughs (not including Staten Island).

Findings

Our neighborhood analyses ranked neighborhoods by variables in the NYC neighborhood census data. These tables found in our report can provide NYC with information on how to optimally create potential convenience stores outside of subway stations. For example, if a neighiborhood has a larger elderly population, a potential convenience store in that neighborhood outside of a subway station may want to sell more newspapers. However, more research would have to be done on consumer behavior patterns by demographic characteristics for NYC to optimize potential economic gains from potential convenience stores.

We further found that neighborhood population is a strong predictor of median subway entries and exits in each neighborhood. Thus, New York City Subway could consider adding stops to neighborhoods with high population, such as Flushing/Whitestone, Jamaica/Hollis, and Washington Heights/Inwood.

The COVID-19 pandemic did significantly impact daily subway entries as entries ranged from 2 to 6 million daily pre-COVID, dropped to below 1 million by March of 2020, and never reached above 2 million entries even by July of 2021. This could be because COVID was still a concern for people at the time, and many international students and tourists remained out of the country.

Despite COVID's influence, we found that covid cases is not a significant predictor of total subway entries in NYC. While covid cases is a significant predictor of total subway entries at the borough level, it can only explain very small variation (less than 3%) of total subway entries. Thus, New York City Subway should look beyond Covid to predict future demands, such as tourism and immigration policies as the pandemic dies down.

We mapped the subway stations by line.

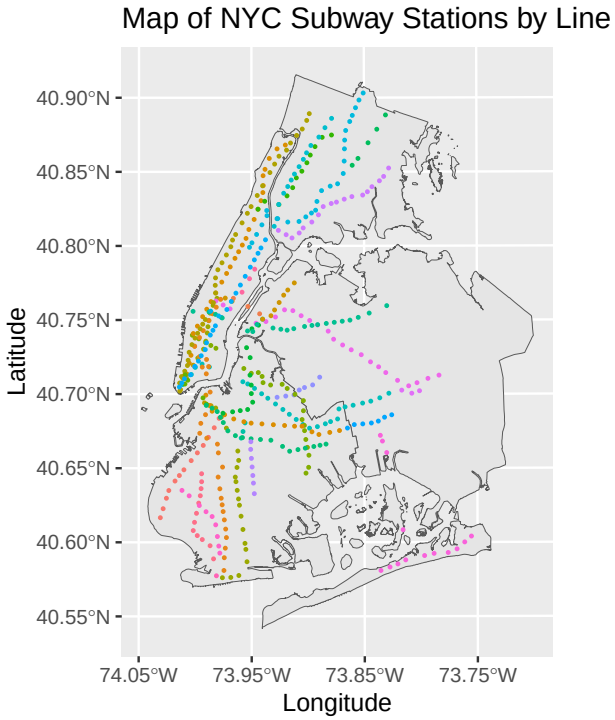
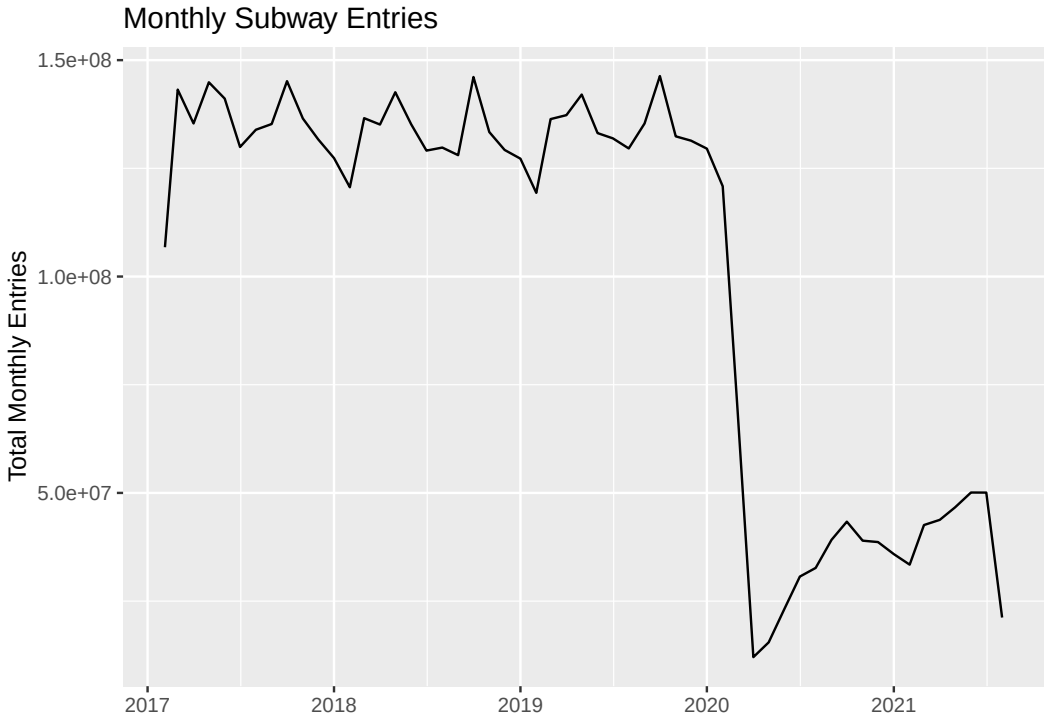


Table 1: Total Number of Subway Stops by Borough

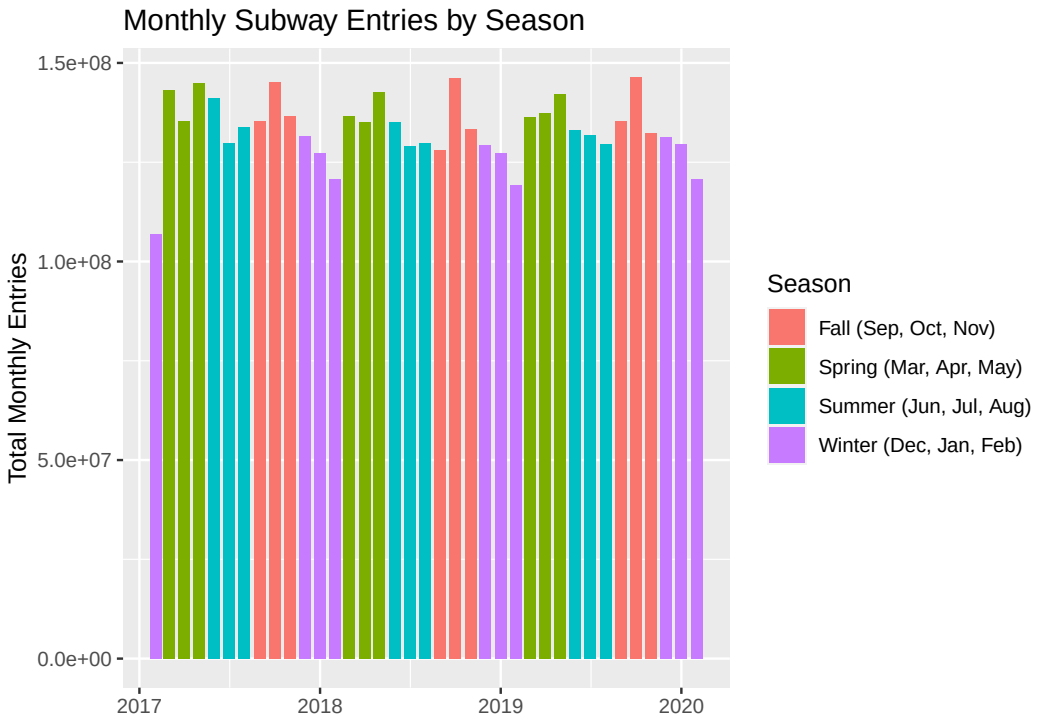
Borough	Total Number of Stops	Percentage of Total Number of Stops
Brooklyn	169	35.58
Manhattan	153	32.21
Queens	83	17.47
Bronx	70	14.74

This is a map of all the subway stops in New York City, colored based the line that runs through each respective stop. Based on the attached table, is it apparent that a majority of stops are within Brooklyn and Manhattan.

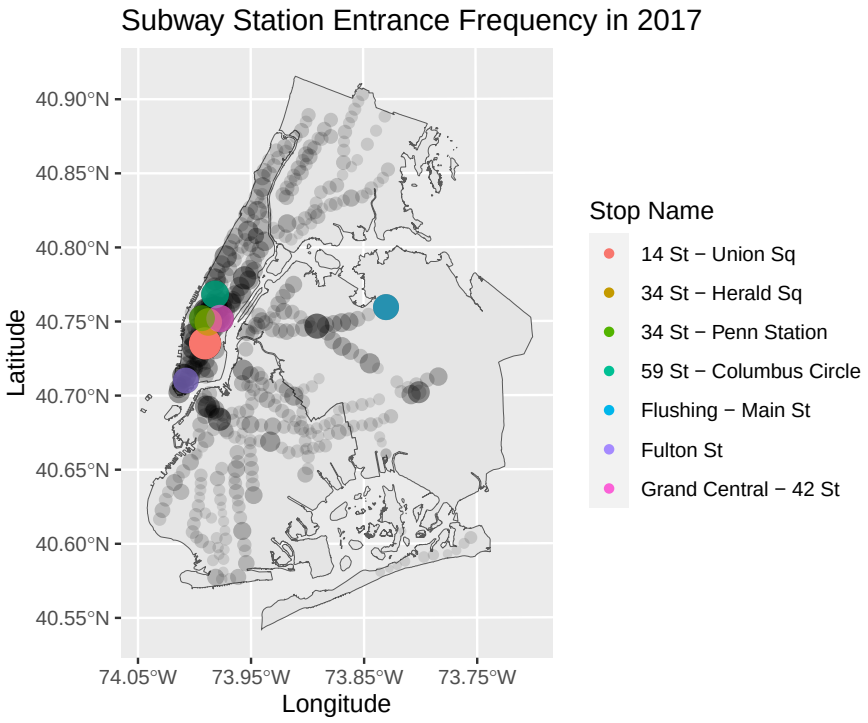
We portrayed monthly subway entries and analyzed for trends.

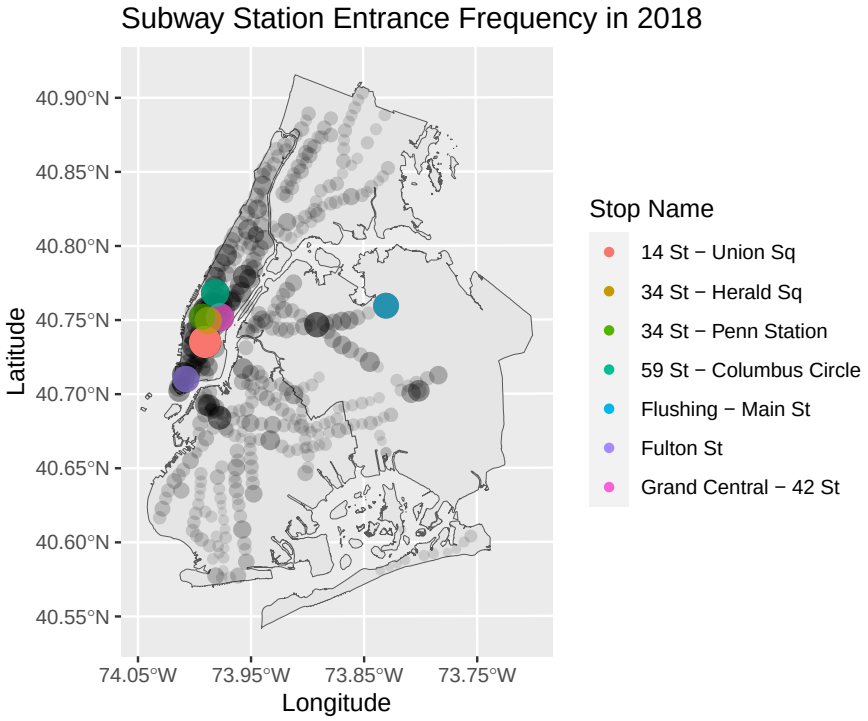


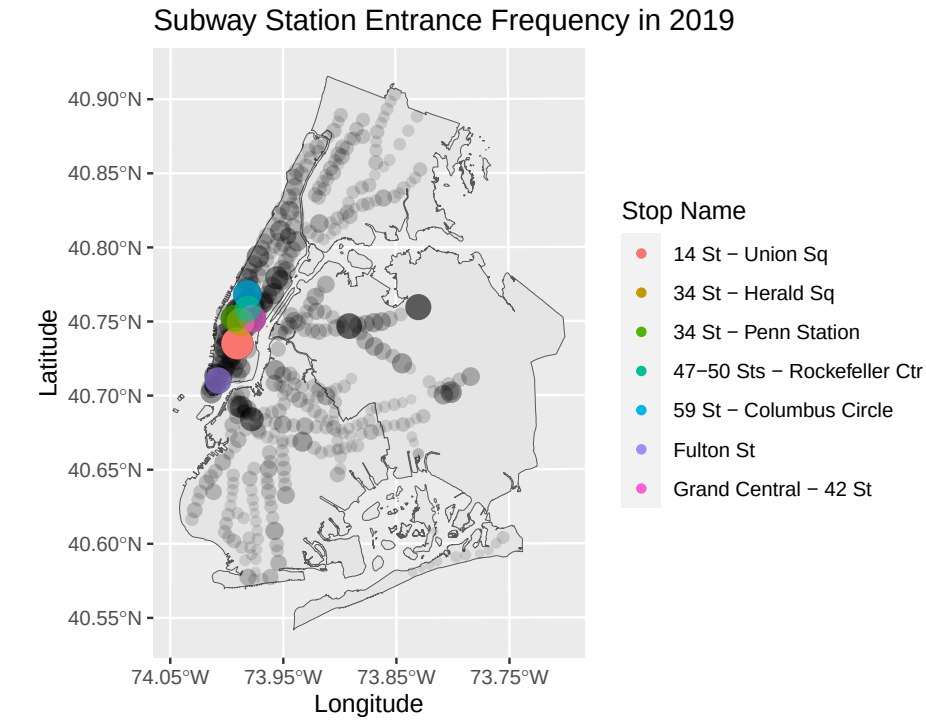
This graph depicts the total entries of each subway station per month. This graph is especially important due to the steep decline of subway use in NYC during the COVID-19 pandemic. Lockdowns and social gathering restrictions required residents and commuters to remain at home. As a result, there was minimal subway use during this time period. However, prior to COVID-19,(approximately March 2020), there seems to indicate a pattern based on subway per month.

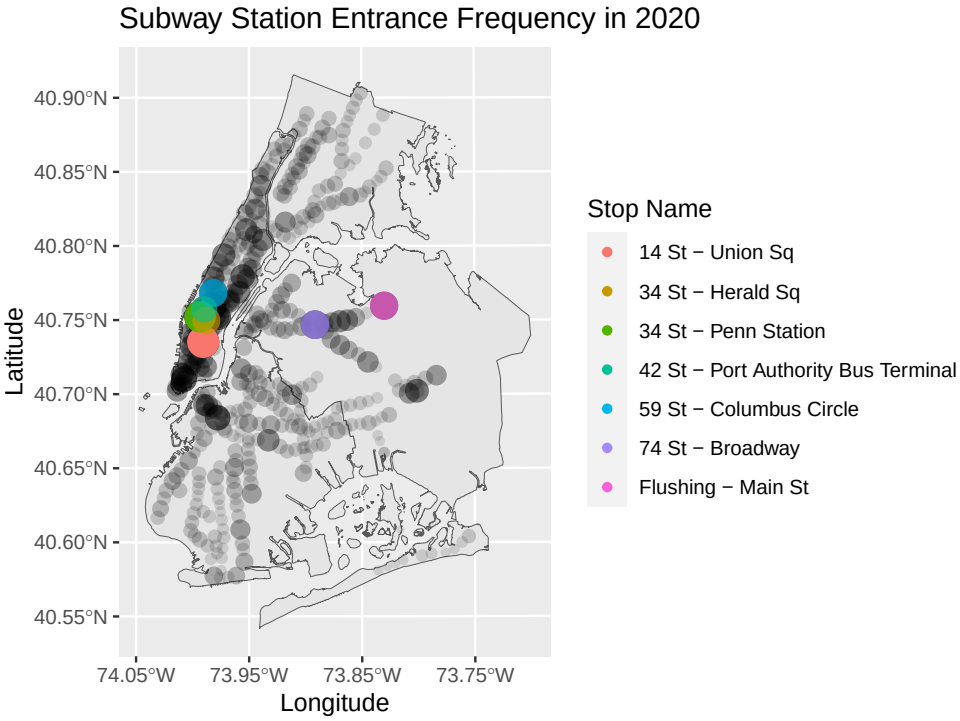


To understand if there is a trend of subway use based on season, we created a plot that depicted the total number of monthly subway station entries colored by season. This graph indicates subway patterns prior to the start of the COVID-19 pandemic to analyze pre-existing trends before the change to the 'new normal'. Based on this graph, it is evident that monthly subway entries are highest during the Fall and Spring months and lower during the Summer and Winter months. This reduced use of the subway during the Summer months could be indicative of better weather prompting people to walk or bike in the city as well as subway route closure due to inclement weather during the Winter months. However, more analysis would need to be done to determine the route cause of this trend. Furthermore, COVID-19 has created a new environment for residents and commuters on how they interact with the subway so it is evident that NYC takes the necessary steps to bring the city back to these trend levels and increases current use of the subway.


Subway Station Entrance Frequency in 2017
40.90°N
40.85°N
40.80°N
40.75°N
40.70°N
40.65°N
40.60°N
40.55°N
Latitude
74.05°W
73.95°W
73.85°W
73.75°W
Longitude
Stop Name
14 St − Union Sq
34 St − Herald Sq
34 St − Penn Station
59 St − Columbus Circle
Flushing − Main St
Fulton St
Grand Central − 42 St


Subway Station Entrance Frequency in 2018
40.90°N
40.85°N
40.80°N
40.75°N
40.70°N
40.65°N
40.60°N
40.55°N
Latitude
74.05°W
73.95°W
73.85°W
73.75°W
Longitude
Stop Name
14 St − Union Sq
34 St − Herald Sq
34 St − Penn Station
59 St − Columbus Circle
Flushing − Main St
Fulton St
Grand Central − 42 St


Subway Station Entrance Frequency in 2019
40.90°N
40.85°N
40.80°N
40.75°N
40.70°N
40.65°N
40.60°N
40.55°N
Latitude
74.05°W
73.95°W
73.85°W
73.75°W
Longitude
Stop Name
14 St − Union Sq
34 St − Herald Sq
34 St − Penn Station
47−50 Sts − Rockefeller Ctr
59 St − Columbus Circle
Fulton St
Grand Central − 42 St


Subway Station Entrance Frequency in 2020
40.90°N
40.85°N
40.80°N
40.75°N
40.70°N
40.65°N
40.60°N
40.55°N
Latitude
74.05°W
73.95°W
73.85°W
73.75°W
Longitude
Stop Name
14 St − Union Sq
34 St − Herald Sq
34 St − Penn Station
42 St − Port Authority Bus Terminal
59 St − Columbus Circle
74 St − Broadway
Flushing − Main St

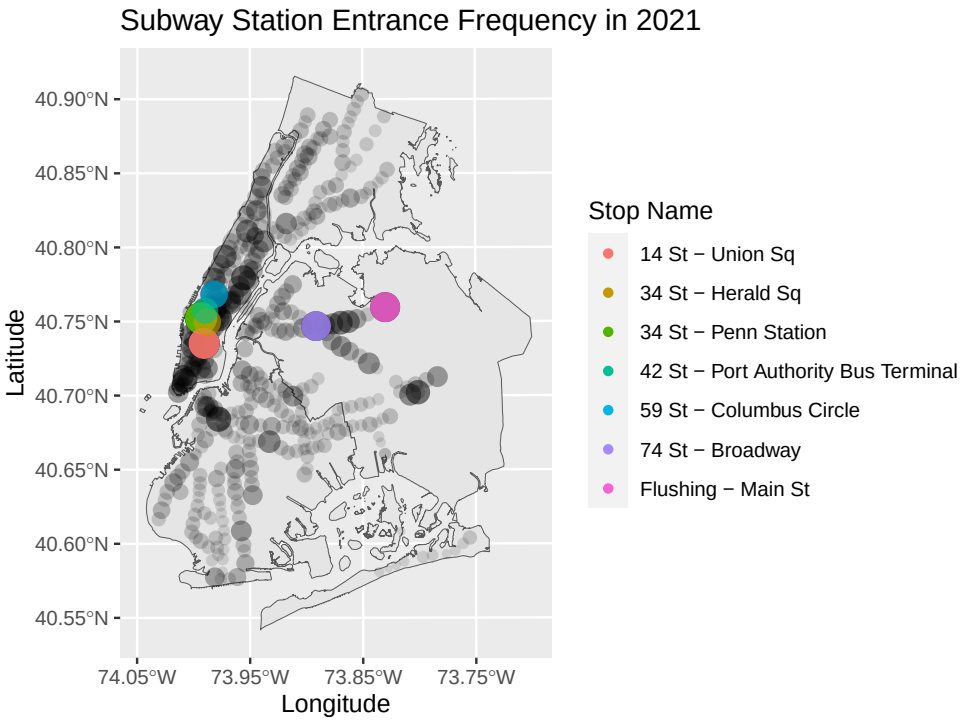


Table 2: Most Entered Subway Stations in 2017

Stop Name	Total Entries
14 St - Union Sq	31459678
59 St - Columbus Circle	21017321
34 St - Herald Sq	20675901
Grand Central - 42 St	20383824
Flushing - Main St	17290978
34 St - Penn Station	17182059
Fulton St	17104479

Table 3: Most Entered Subway Stations in 2018

Stop Name	Total Entries
14 St - Union Sq	33223321
Grand Central - 42 St	23191133
59 St - Columbus Circle	23122675
34 St - Herald Sq	21708657
Fulton St	19680506
34 St - Penn Station	19051527
Flushing - Main St	18056193

Table 4: Most Entered Subway Stations in 2019

Stop Name	Total Entries
14 St - Union Sq	32015015
59 St - Columbus Circle	23179407
34 St - Herald Sq	22051787
Grand Central - 42 St	21063952
34 St - Penn Station	19565441
Fulton St	18860558
47-50 Sts - Rockefeller Ctr	18561252

Table 5: Most Entered Subway Stations in 2020

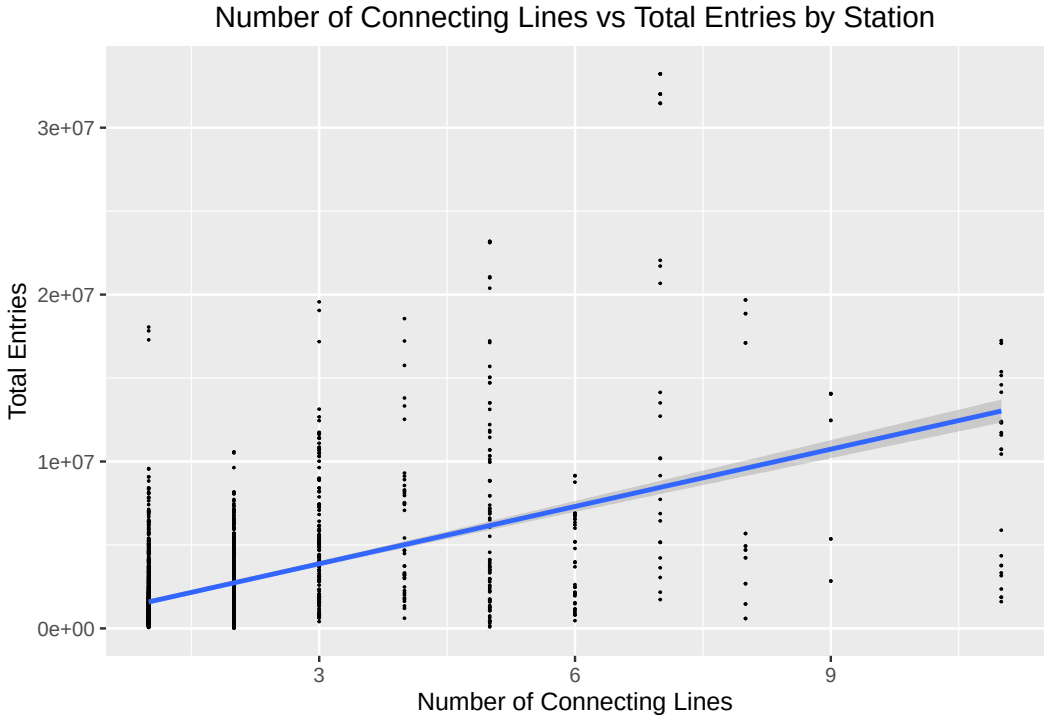
Stop Name	Total Entries
14 St - Union Sq	10187851
34 St - Penn Station	9635966
74 St - Broadway	7385189
59 St - Columbus Circle	7183507
Flushing - Main St	7011959
34 St - Herald Sq	6877181
42 St - Port Authority Bus Terminal	5882872

Table 6: Most Entered Subway Stations in 2021

Stop Name	Total Entries
34 St - Penn Station	5633753
14 St - Union Sq	5157958
Flushing - Main St	4618427
74 St - Broadway	4560190
59 St - Columbus Circle	3665609
34 St - Herald Sq	3631992
42 St - Port Authority Bus Terminal	3157749

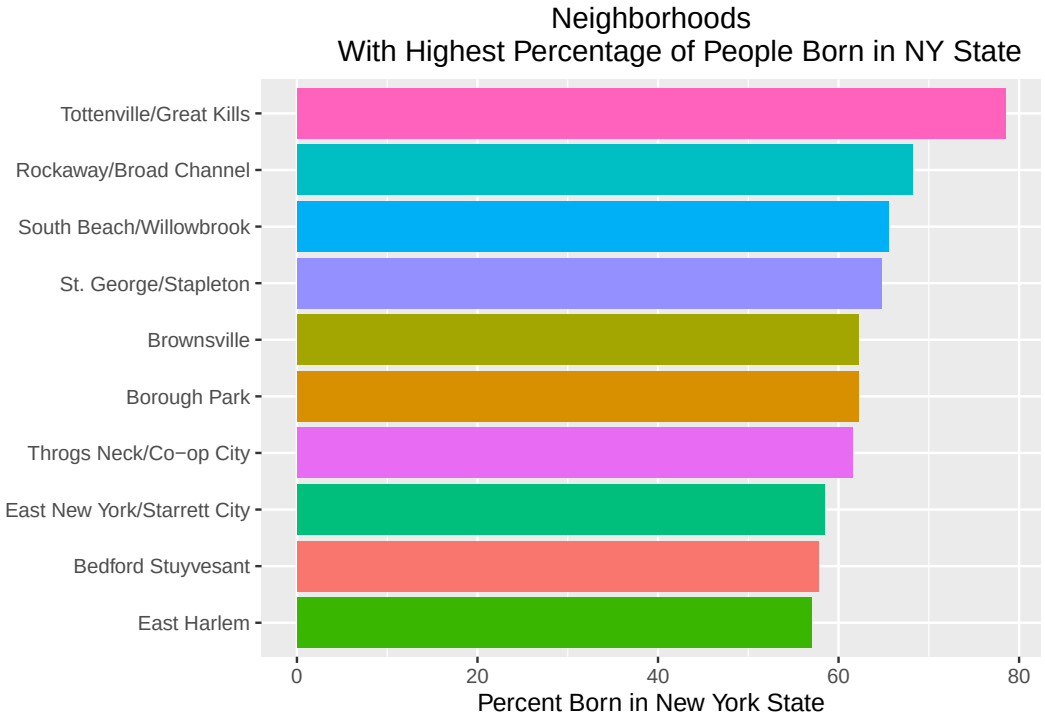
This map indicates the number of annual entries per stop. The amount of annual entries per stop is indicated equally by the size and transparency of the point, i.e. the less transparent and larger the point is, the greater the number of annual stops are. To that, the map highlights the seven most entered stops for each year with a table below indicating the exact amounts. Based on the graph, it is evident that Union Sq is the most entered subway station in the first four years. To that, additional interesting patterns emerge that speak to how residents and commuters interact with the subway. Specifically, a majority of the most entered subway spots were located in southern Manhattan and Midtown. In addition, Penn Station is one of the biggest hubs for commuters to enter the city to it is interesting to see that it was the second most entered subway spot in 2020 and most entered in 2021. Penn Station only reached its peak of entrances after the COVID-19 pandemic. Unlike Penn Station, Grand Central, the other large commuter hub, remained in the top seven prior to the COVID-19 pandemic and has since not seen as much demand. Overall, Union Sq, Columbus Circle, Herald Sq and Penn Station have all remained in the top seven most entered subway stations throughout the past five years.

Next, we try to see if there is a correlation between number of connecting lines and total entries in a year (by stop). Perhaps stops with more connecting lines will have more entries.

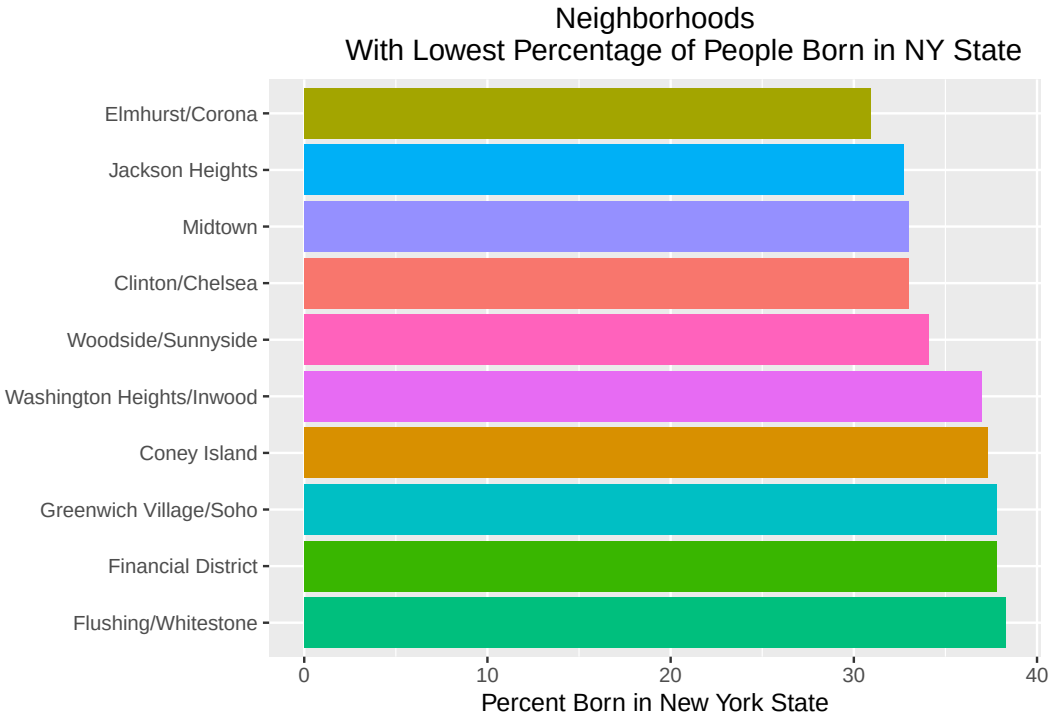


Here, our scatterplot and line of best fit indicate a positive relationship between number of connecting lines at an individual subway station and total entries at that station. Number of connecting lines is a statistically significant predictor, with significance of $<$ 2e-16. Furthermore we obtain an R^2 value of 0.2837, which is fairly large. This means that 28.37% of the variation in total number of entries at a given subway stop can be explained by the linear relationship with number of connecting lines at an individual subway stop. Our point estimate for slope is 1143879, meaning that on average for each additional connecting line, entries in a year at an individual subway station increase by 1,143,879.

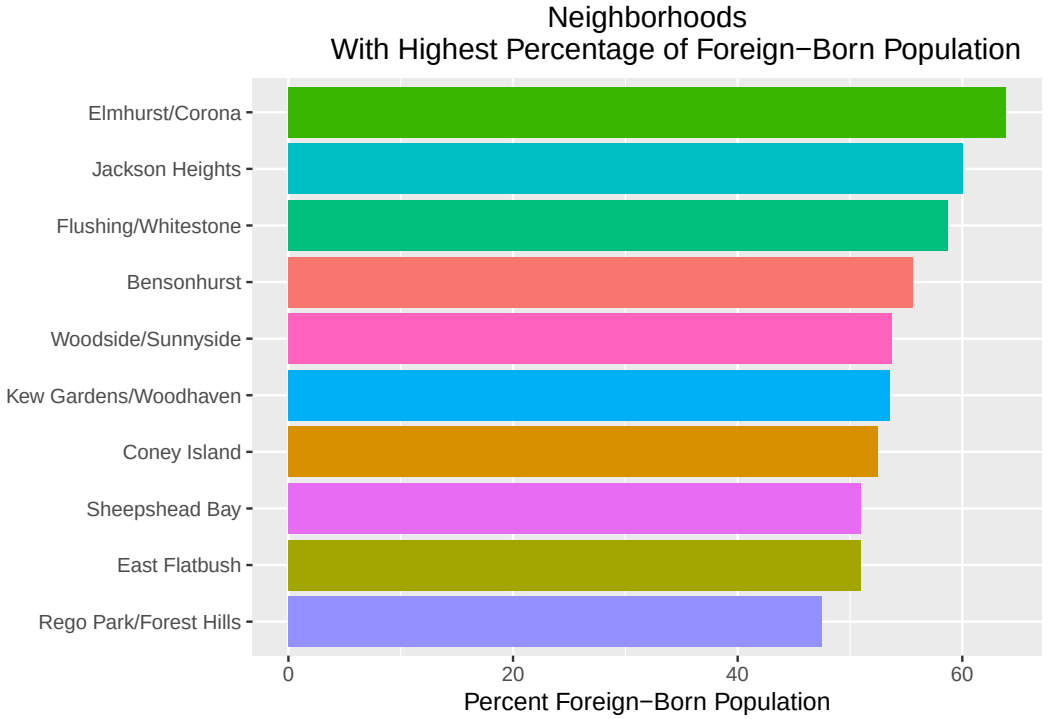
We identified neighborhoods with specific demographic and behavioral patterns.



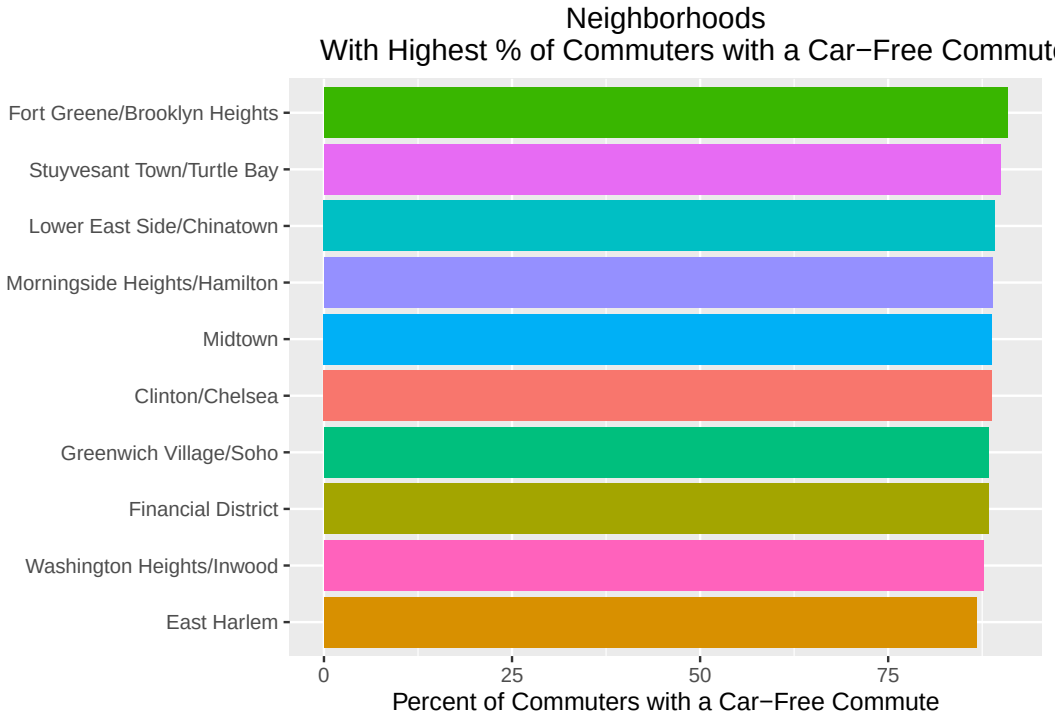
These are the top 10 neighborhoods in NYC with the greatest percentage of their populations being born in NY State. If NYC wants to open convenience stores outside of subway stations, they could try to cater the convenience stores to more "local" people in neighborhoods with a larger percentage of people born in New York State.



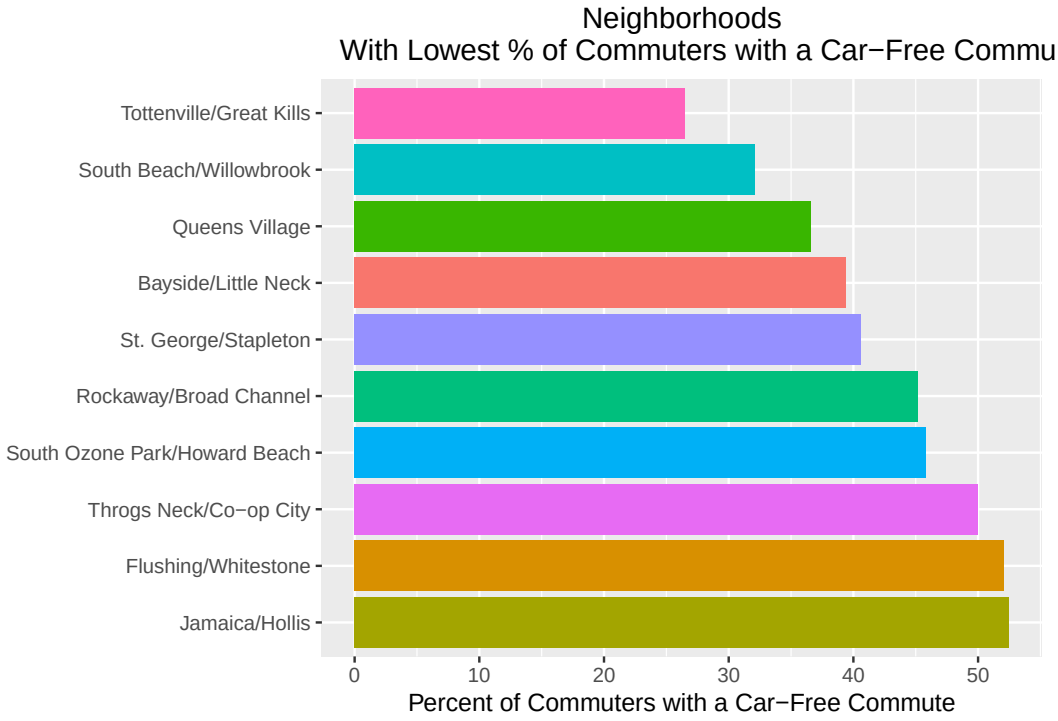
These are the top 10 neighborhoods in NYC with the smallest percentage of their populations being born in NY State. In neighborhoods with small proportions of the population being born in NY State, NYC could try to stock convenience stores outside of these subway stations with a range of items that are popular in other regions.



These are the top 10 neighborhoods with the larges foreign-born populations. In neighborhoods with large proportions of the population being foreign-born, NYC could try to stock convenience stores outside of these subway stations with items that are more international.



These are the ten neighborhoods with the largest percentage of car-free commutes. Residents of these neighborhoods may be more likely to utilize the subway. NYC should conduct further research to determine what these residents would be interested in purchasing in potential convenience stores outside of subway stations.



These are the ten neighborhoods with the smallest percentage of car-free commutes. Perhaps NYC should

not place convenience stores in subway stations located in these neighborhoods since those who commute to work in a car are less likely to utilize the subway.

While we manually identified the top 10 neighborhoods based on different demographic patterns, we created a function to increase efficiency.

Table 7: Top Ten Neighborhoods in Population

Neighborhood	Population
Flushing/Whitestone	260282
Jamaica/Hollis	249331
Washington Heights/Inwood	219998
Flatlands/Canarsie	215637
Upper East Side	214219
Queens Village	208786
Upper West Side	207134
Bensonhurst	205850
Ridgewood/Maspeth	199043
Morrisania/Crotona	188075

Table 8: Top Ten Neighborhoods in Population aged 65+

Neighborhood	`Population aged 65+`
Coney Island	23.7
Bayside/Little Neck	22.7
Throgs Neck/Co-op City	20.8
Upper West Side	20.1
Rego Park/Forest Hills	19.7
Flushing/Whitestone	19.6
Lower East Side/Chinatown	19.5
Riverdale/Fieldston	19.2
Upper East Side	19.2
Tottenville/Great Kills	18.6

Table 9: Top Ten Neighborhoods in Poverty rate

Neighborhood	`Poverty rate`
Mott Haven/Melrose	44.2
Hunts Point/Longwood	44.2
Brownsville	39.9
Highbridge/Concourse	36.4
Fordham/University Heights	36.4
Morrisania/Crotona	35.8
Belmont/East Tremont	35.8
Kingsbridge Heights/Bedford	32.9
East Harlem	31.5
Borough Park	30.8

Table 10: Top Ten Neighborhoods in Population density (1,000 persons per square mile)

Neighborhood	Population density (1,000 persons per square mile)
Upper East Side	107.8
Central Harlem	103.1
Fordham/University Heights	98.2
Stuyvesant Town/Turtle Bay	92.3
Lower East Side/Chinatown	92.0
Kingsbridge Heights/Bedford	91.5
Morningside Heights/Hamilton	80.7
Washington Heights/Inwood	75.1
Highbridge/Concourse	75.0
Upper West Side	69.9

Table 11: Top Ten Neighborhoods in Median household income (2018$)

Neighborhood	Median household income (2018$)
Financial District	147640
Greenwich Village/Soho	147640
Park Slope/Carroll Gardens	137370
Upper East Side	133850
Upper West Side	126260
Stuyvesant Town/Turtle Bay	114530
Clinton/Chelsea	103930
Midtown	103930
Tottenville/Great Kills	98640
Fort Greene/Brooklyn Heights	94330

Table 12: Top Ten Neighborhoods in Mean travel time to work (minutes)

Neighborhood	Mean travel time to work (minutes)
Rockaway/Broad Channel	51.4
Riverdale/Fieldston	49.8
Sheepshead Bay	49.0
Flatlands/Canarsie	49.0
Jamaica/Hollis	48.9
East Flatbush	48.4
Throgs Neck/Co-op City	48.4
Tottenville/Great Kills	47.9
Kew Gardens/Woodhaven	47.5
Bensonhurst	47.4

Table 13: Top Ten Neighborhoods in Percent Asian

Neighborhood	Percent Asian
Flushing/Whitestone	54.5
Bayside/Little Neck	46.0
Bensonhurst	38.9
Woodside/Sunnyside	36.3
Lower East Side/Chinatown	36.1
Hillcrest/Fresh Meadows	34.4
Sunset Park	32.7
Elmhurst/Corona	32.7
Rego Park/Forest Hills	29.5
Bay Ridge/Dyker Heights	28.2

Table 14: Top Ten Neighborhoods in Percent Hispanic

Neighborhood	Percent Hispanic
Kingsbridge Heights/Bedford	74.1
Fordham/University Heights	71.1
Highbridge/Concourse	68.8
Mott Haven/Melrose	67.2
Hunts Point/Longwood	67.2
Washington Heights/Inwood	66.9
Jackson Heights	62.5
Morrisania/Crotona	61.8
Belmont/East Tremont	61.8
Elmhurst/Corona	55.8

Table 15: Top Ten Neighborhoods in Percent white

Neighborhood	Percent white
Tottenville/Great Kills	82.5
Upper East Side	74.6
Financial District	72.2
Greenwich Village/Soho	72.2
Upper West Side	69.1
Borough Park	68.9
Stuyvesant Town/Turtle Bay	68.5
Greenpoint/Williamsburg	64.5
South Beach/Willowbrook	64.1
Park Slope/Carroll Gardens	63.9

Table 16: Top Ten Neighborhoods in Percent black

Neighborhood	Percent black
East Flatbush	87.1
Brownsville	71.3
Williamsbridge/Baychester	67.2

Neighborhood	Percent black
Flatlands/Canarsie	62.2
South Crown Heights/Lefferts Gardens	60.2
Jamaica/Hollis	59.2
Crown Heights/Prospect Heights	55.7
Queens Village	53.9
Central Harlem	53.0
East New York/Starrett City	52.8

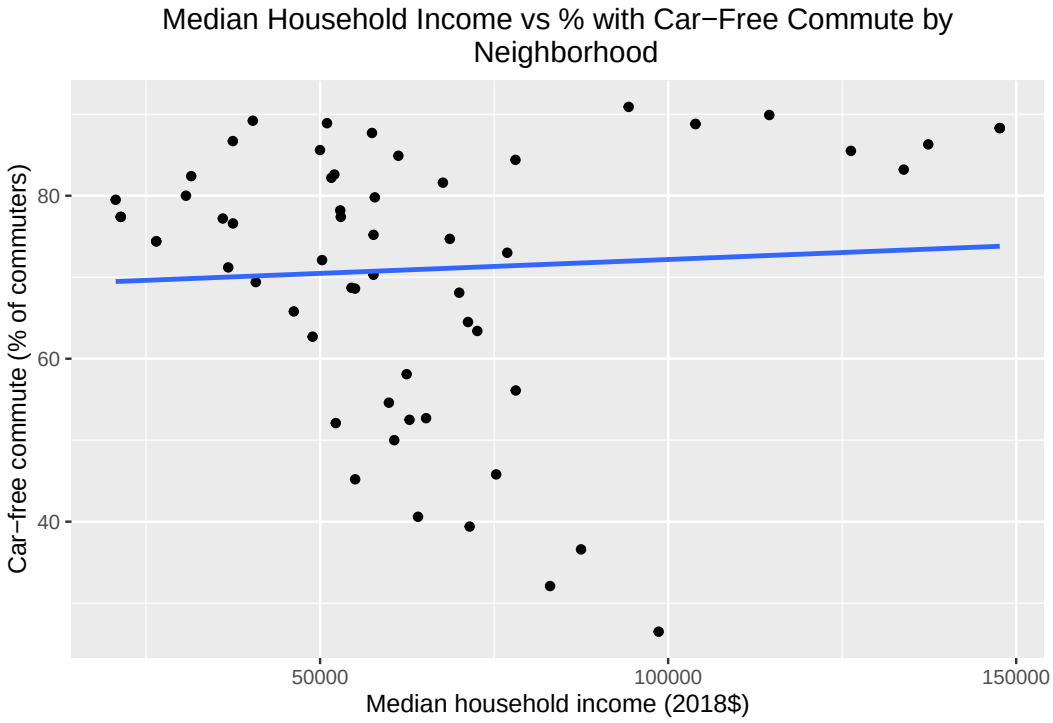
Table 17: Top Ten Neighborhoods in Disabled population

Neighborhood	Disabled population
Mott Haven/Melrose	20.9
Hunts Point/Longwood	20.9
Morrisania/Crotona	19.5
Belmont/East Tremont	19.5
Morris Park/Bronxdale	15.5
Parkchester/Soundview	14.9
Kingsbridge Heights/Bedford	14.2
Highbridge/Concourse	14.0
Fordham/University Heights	13.8
East Harlem	13.6

The above tables display some of the information which may help the city determine what types of convenience stores should be placed outside of different subway stations.

In the following section, we looked at how different factors impact percentage of neighborhood residents with a car-free commute (likely related) to subway usage.

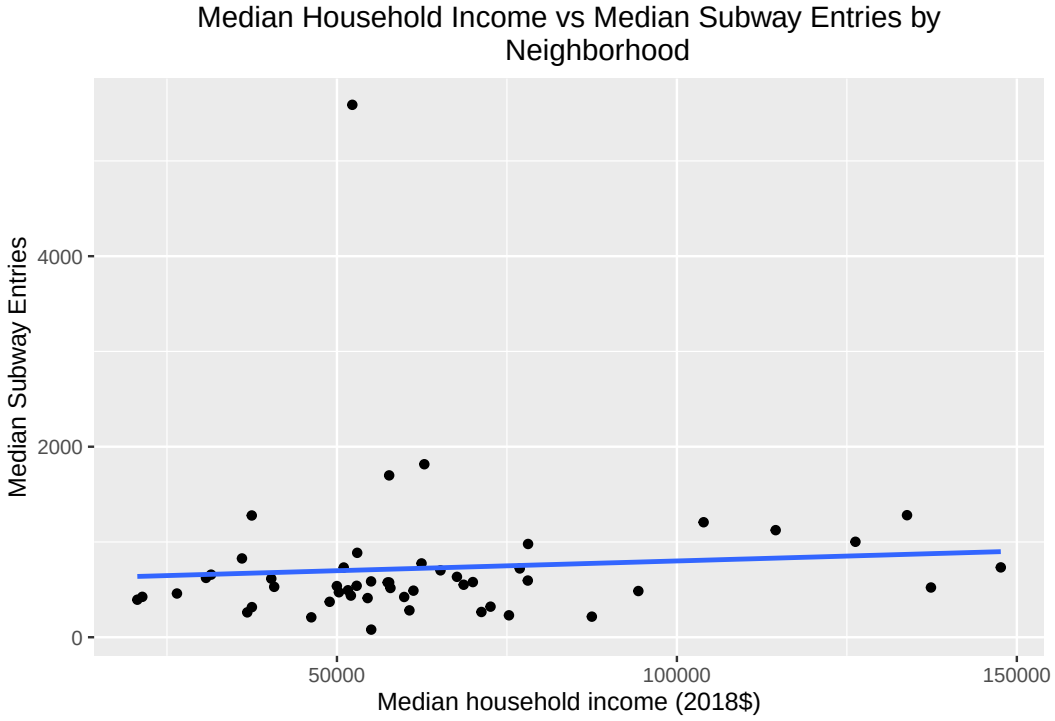
Does median household income of a neighborhood predict percentage with a car-free commute?



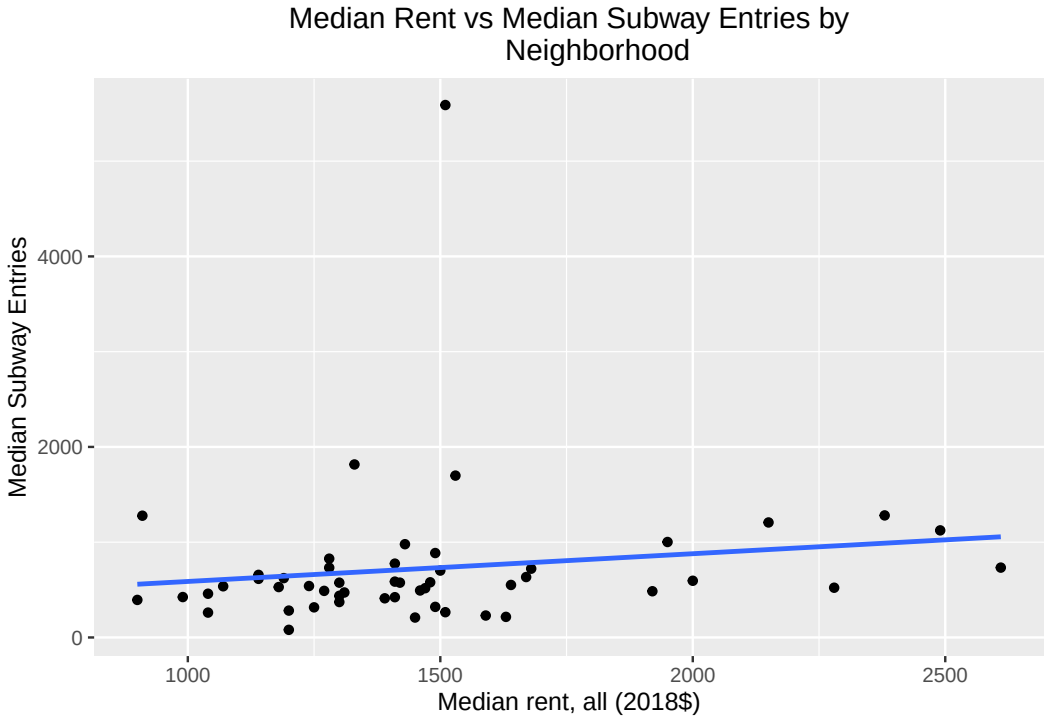
The scatterplot shows that median household income is not a good predictor of percentage of commuters with a car-free commute. When looking at the details of the model, median household income is not a statistically significant predictor using a significance level of .05 (.632 $>$ .05).

We found the median subway entries and exits, difference between entries and exits, and the ratio of entries over exits of each neighborhood to see whether some neighborhoods have higher subway traffic as well as different entries to exits ratios.

Here, we try to predict median subway entries with demographic data, including household income, median rent, both household income and median rent, neighborhood population, and all three variables.

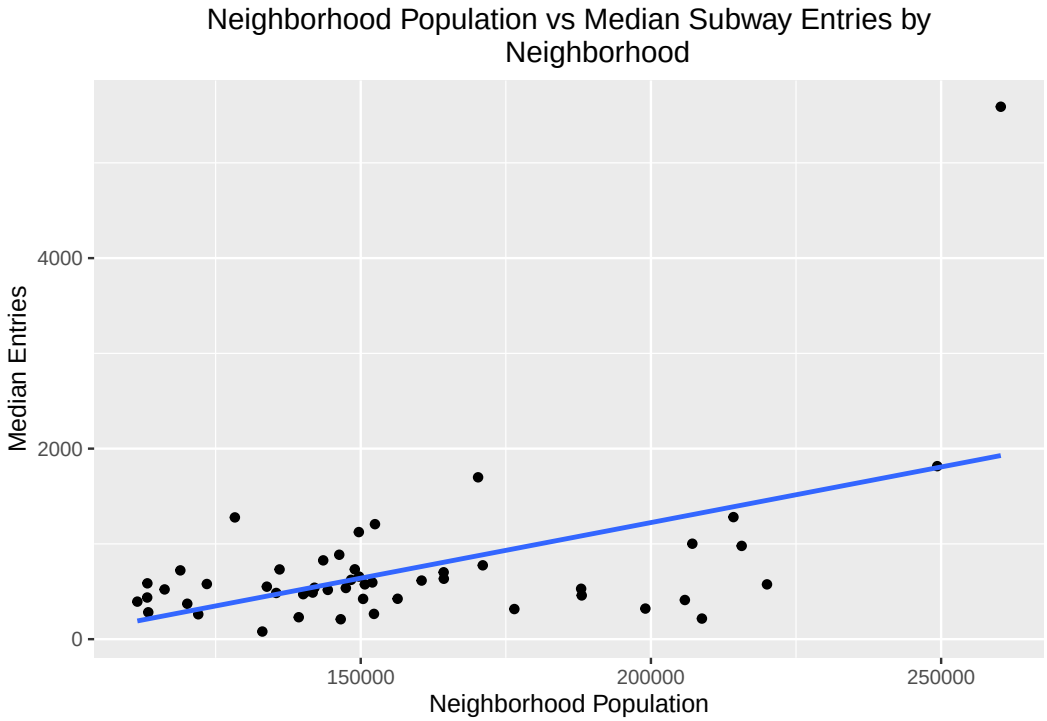


Household income by neighborhood is not a good predictor of median subway entries by neighborhood as seen by the scatterplot. The linear model shows an R^2 of .00568 and an insignificant predictor (p = .5992).



Median rent by neighborhood is not a good predictor of median subway entries by neighborhood as seen by the scatterplot. The linear model shows an R^2 of .020279 and an insignificant predictor (p = .3127).

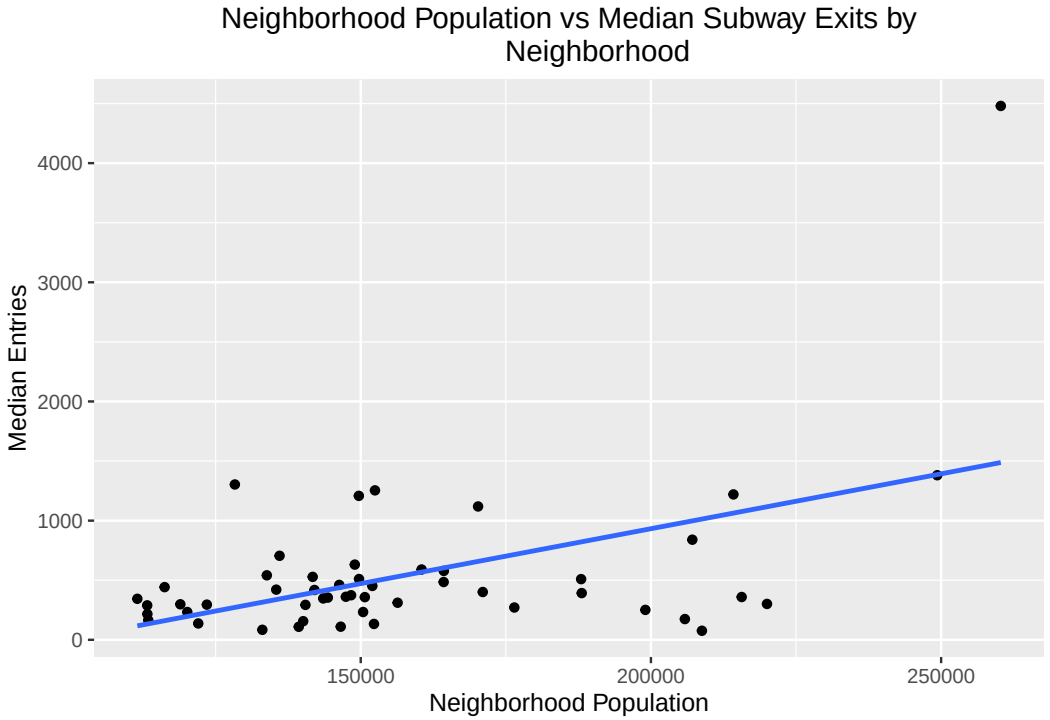
When running a multiple regression with both median rent and median household income by neighborhood as predictors, our model is still insignificant. The p-value for a F-Test is .3207.



Lastly, we then tried using neighborhood population as a predictor for median subway entries by neighborhood. The scatterplot shows a more consistent relationship. As neighborhood population increases, so do median entries. We created a linear model and got a fairly large R-Squared value of 0.2773, and neighborhood population was a significant predictor (p = 7.2e-05). We got a point estimate for the slope of .01167. On average, for each additional person in the neighborhood, median entries increases by .01167.

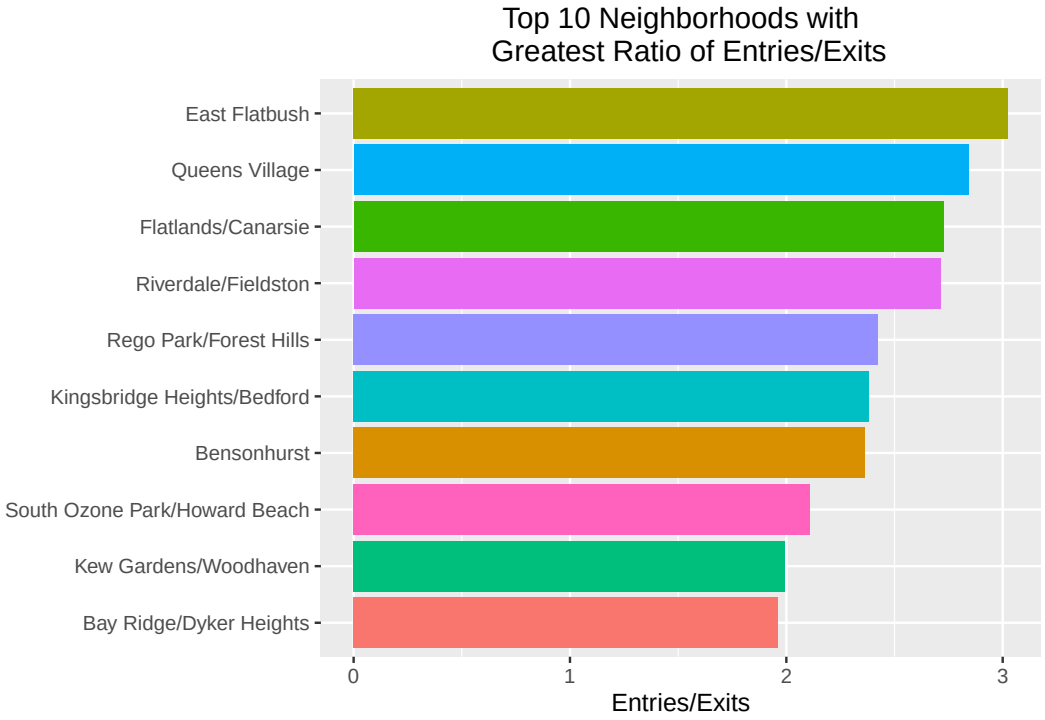
Out of curiosity, we then tried doing a multiple regression with all three predictors to see if median rent or median income would become significant predictors. At an alpha level of .05, they remained insignificant.

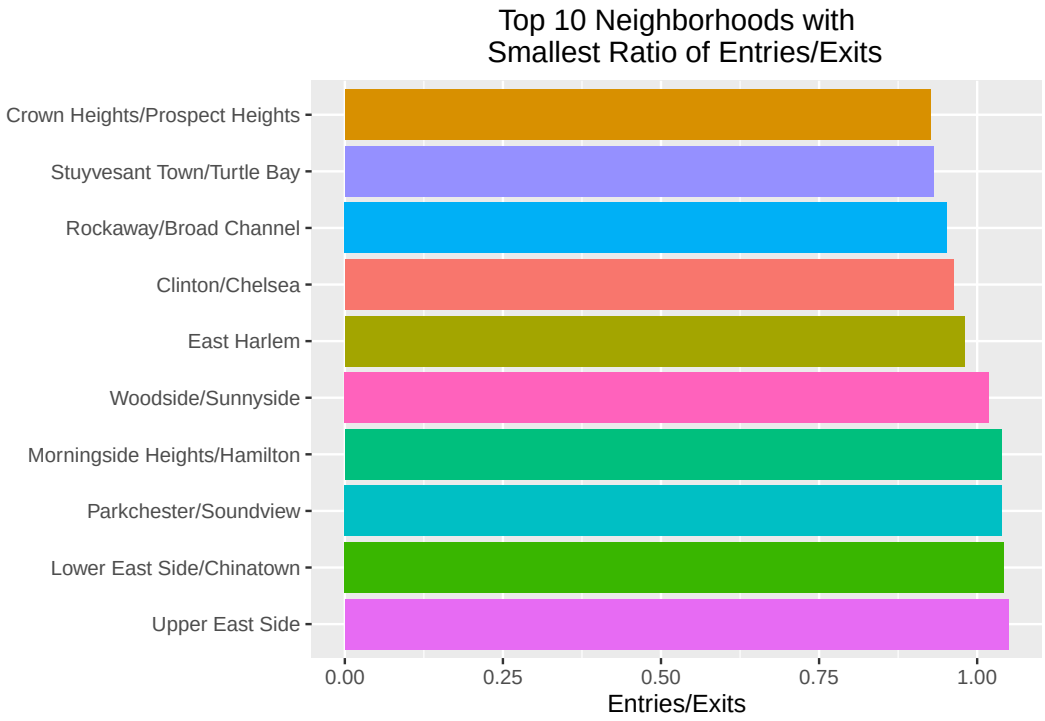
Instead of predicting for subway entries, we predicted for subway exits. We sticked with neighborhood population as the predictor since it is a significant predictor for entries.



Again, we found that neighborhood population is a very significant predictor for exits. As neighborhood population increases, median number of exits by neighborhood increases. We got R^2 = .2456 and a point estimate for the slope of .009213. On average, for each additional person in the neighborhood, median exits increases by .009213.

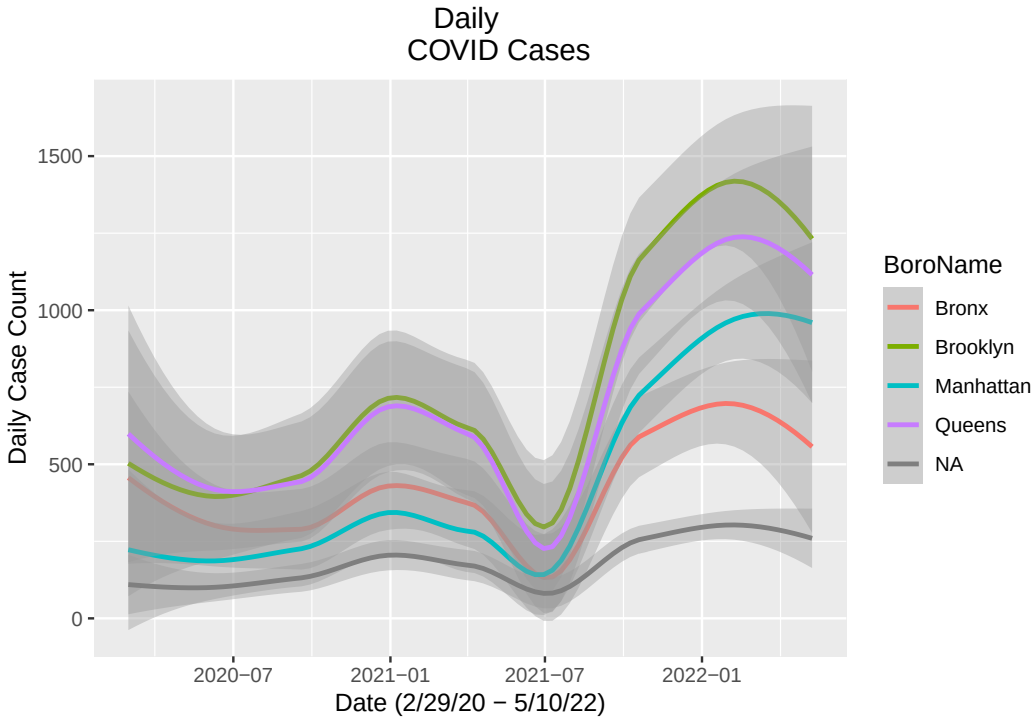
Interestingly, some neighborhoods have a much different number of entries than exits. We wonder what could be contributing to this. We hypothesize that this is because there are two ways to exit the subway. One can exit via the turnstile or the emergency door, which is sometimes used as a normal exit. Our data only accounts for the turnstiles. There are approximately 5.60 billion entries, compared to 4.46 billion exits, showing about a 25% discrepancy. Since neighborhoods with larger populations would naturally have a greater difference between the median entries and exits, we focused on ratio of entries/exits.



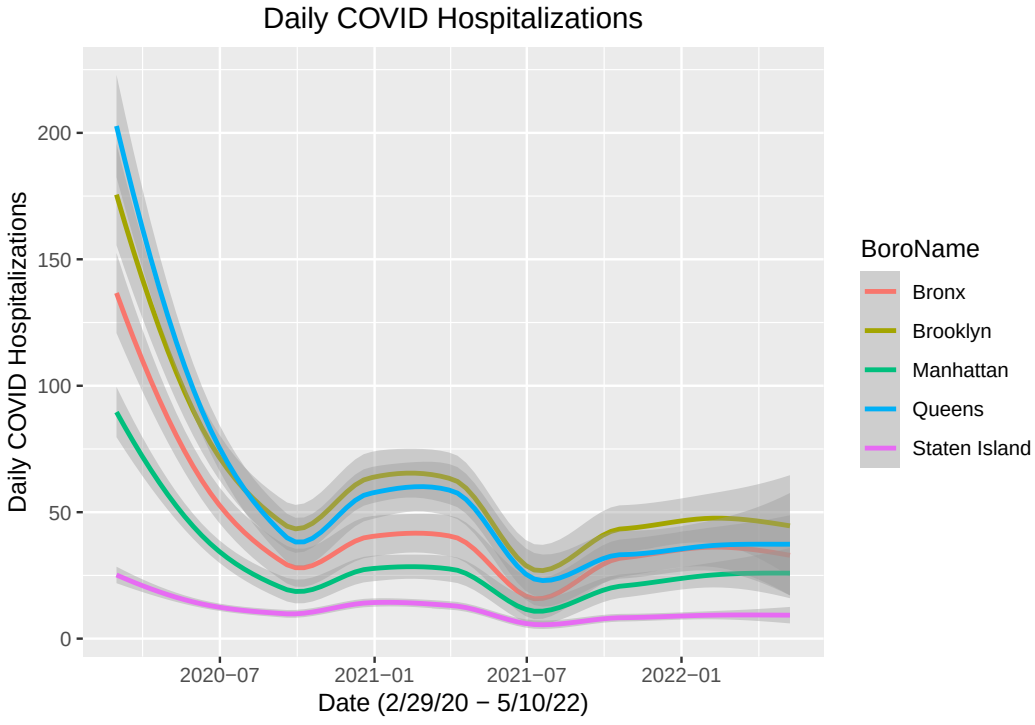


Here are the ten neighborhoods with the largest ratios of entries/exits and the ten neighborhoods with the smallest ratios of entries/exits. As seen in the second horizontal bar chart (smallest ratios), only five neighborhoods have a ratio less than 1. Due to the discrepancy in entries and exits, we will focus the rest of our subway analyses only on entries.

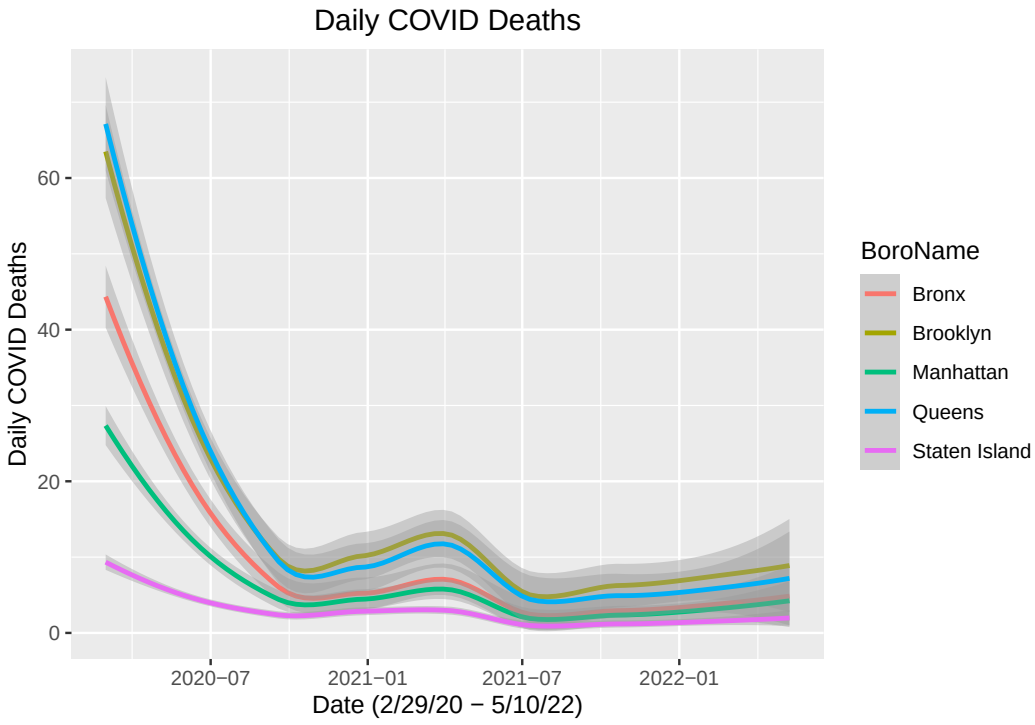
Due to the pandemic, subway traffic has drastically declined. In this section, we present COVID metrics patterns by borough first then see how COVID has impacted subway traffic.



The graph above displays trends in number of daily COVID cases by borough.

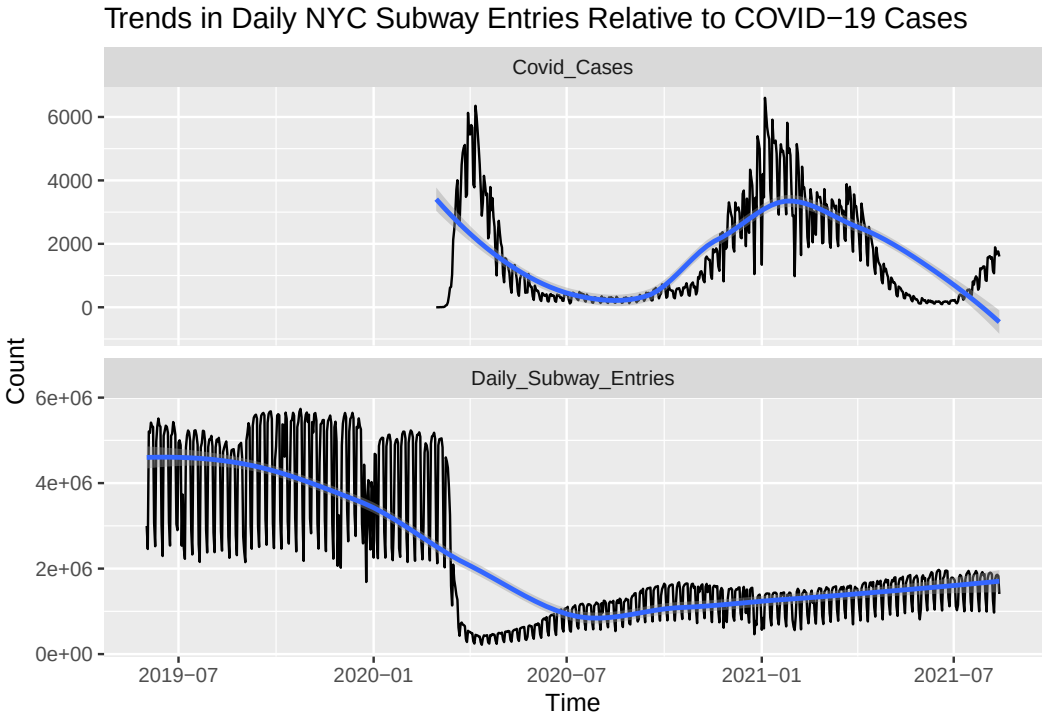


The graph above displays trends in number of daily hospitalizations due to COVID by borough.



The graph above displays trends in number of daily COVID deaths by borough.

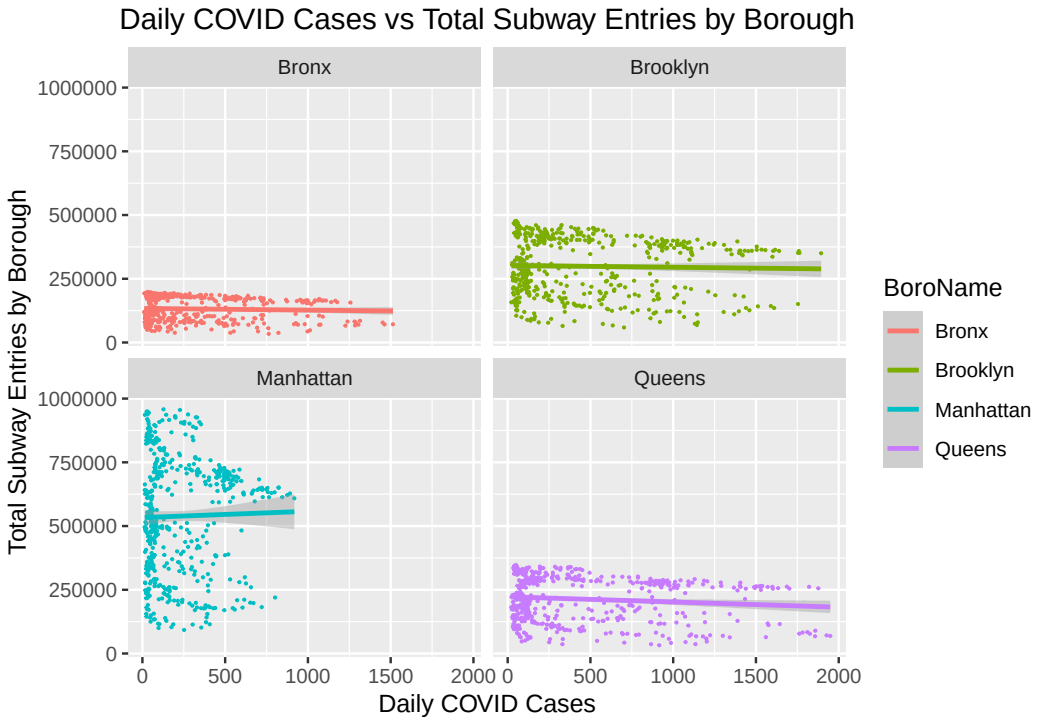
Here, we visualized the change in subway station entries before and after COVID.



Before the COVID pandemic, we could see that total daily subway entries across all stations mostly remained above 2 million entries and could reach nearly 6 million entries on some days. As COVID cases began to rise in March of 2020, we saw a sharp decline in subway entries to below 1 million entries. Since the pandemic till July of 2021, subway entries have been increasing but not anywhere close to pre-pandemic levels.

Next, we try to see if there is a relationship between COVID cases and subway entries by borough.

Note: There are no subway stations in Staten Island.



The scatterplots above show negative trends for the Bronx, Brooklyn, and Queens, indicating that as Daily COVID Cases in a borough increased, subway entries decreased. However, this relationship does not appear very strong. Perhaps this could be due to a lag time between COVID cases and impact on peoples' behaviors. Interestingly, the line of best fit for Manhattan has a positive slope, showing that on average, as daily COVID cases increased, total subway entries in Manhattan also increased. We are not sure why this could be.

Using linear models

As we saw a drastic decline in subway entries since the COVID pandemic, we wondered whether we could use COVID cases to predict subway daily entries.

We found that daily covid cases is a very poor predictor of daily subway entries as its p-value of 0.108 is greater than 0.05. Its R-suqared value is also very small at 0.005067. This means that only 0.5% of the variation in daily subway entries can be explained by the linear relationship with daily covid cases.

We speculated that maybe covid prevalence differed across boroughs, so we wrote a function that would predict subway entries with covid cases for each borough.

Table 18: Linear Models for Each Borough in Predicting Subway Entries with Covid Cases

Borough	Coefficients	P-Values	R-Squared
Manhattan	-141.540	4.42e-02	0.00762
Brooklyn	-40.970	5.88e-03	0.01423
Bronx	-20.342	8.55e-03	0.01297
Queens	-38.721	8.67e-05	0.02868

As shown in the table, all four linear models have p-values that are less than 0.05, indicating that COVID cases is a significant predictor of subway traffic in each borough. However, it is important to note that the R-squared values are all very small, ranging from 0.0076 (Manhattan) to 0.0287 (Queens). Take Manhattan as an example. This means that only 0.76% of the variation in the daily subway entries can be explained by the linear relationship with daily covid cases. Even the best fitted model, Queens, can only explain 2.87% of the variation in the daily subway entries with its linear model. Thus, there are probably many other factors that affect subway traffic than COVID ever since the pandemic started.