OMNY vs MetroCard Use

2024 MTA Open Data Challenge Submission

Evan Powell

Introduction

Currently, the OMNY (One Metro New York) system, a contactless fare payment system, is being implemented in public transit in the New York Metropolitan Area. OMNY is available in all New York City Transit rail stations, and is being rolled out for other systems, with plans to phase out its predecessor, the MetroCard.

This investigation focuses on the rates at which New York City Subway and Staten Island Railroad riders use OMNY vs. MetroCard, referred to in this report as the OMNY use rate. Over the course of the investigation, OMNY use rates were analyzed at the individual station level, the zip code level, and the system aggregate level. Throughout the investigation, OMNY use rate is score between 0 and 1, representing the proportion of riders who paying with OMNY.

Data Sources

The Data from this investigation was sourced from several agencies, most importantly the MTA. The primary dataset used in this investigation is named MTA Subway Hourly Ridership: Beginning July 2020[1]. At the start of the investigation, the dataset began in February 2022, and only data between Feb 2022 and Sept 2024 is considered in this report. In addition to hourly ridership data, the investigation used datasets 'MTA Stations and Complexes' and 'MTA Colors' from the MTA.

Data used in this investigation is also sourced from the US Census Bureau and the City of New York. US Census Data consists of Income, Age, Population, and Area information at the Zip Code Tabulation Area level, as well as TIGER/Line shapefiles. Data from the City of New York was used for borough outlines.

Systemwide Analysis

Prior to examining the OMNY use rate at a more granular level, we can get an overview of how the payment method is used across the system. The data presented in this section, along with station-level analysis, can be found in section 4 of the attached IPython notebook. Ridership data was retrieved for the month of September 2024 for these sections of the investigation. The systemwide rate of OMNY use for this period was **59.46%**.

---

[1] Name as of Oct. 25, 2024. For the duration of the investigation, up until the writing of this report, the dataset was named MTA Subway Hourly Ridership: Beginning February 2022.

Borough Level

We can also provide a borough-level overview of OMNY use rates. This data is shown in Table 1 and Figure 1.

	Borough	OMNY Use Rate
1.	Manhattan	0.631
-	System Overall	0.595
2.	Brooklyn	0.591
3.	Queens	0.526
4.	Bronx	0.477
5.	Staten Island	0.437

Table 1: OMNY use by Borough

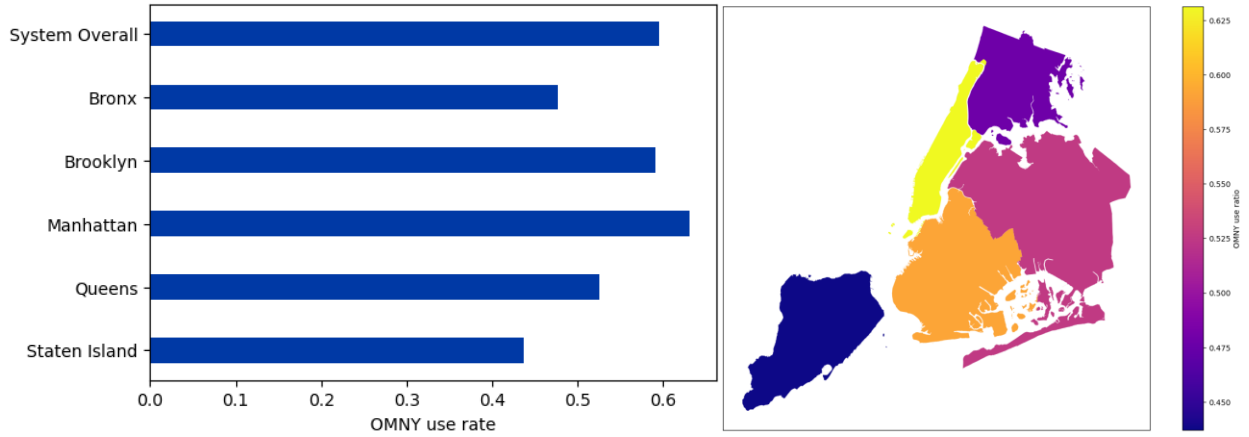

Figure 2: OMNY use by Borough

We can observe that there is a moderate difference in OMNY use rates between boroughs. Manhattan is the only borough with a higher OMNY use rate than the system aggregate, while the OMNY use rate for the outer boroughs follows the same order as their population. Later in this investigation, we will examine geographic differences with greater precision.

Fare Class Category

We can also analyze the rates at which different fare classes use OMNY and MetroCard. Several fare classes recorded in the data were not considered because they only apply to MetroCard. This data is shown below in Table and Figure 2. Note: 'Other' fare class category contains a considerably different set of specific codes for OMNY and MetroCard, and this rate may not be useful when drawing conclusions.

Fare Class Category	OMNY use rate
System Overall	0.595
Full Fare	0.801
Students	1.000
Seniors & Disability	0.099
Fair Fare	0.000
Other	0.121

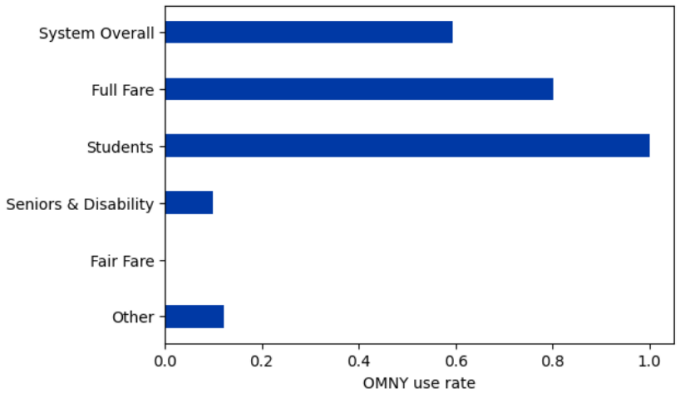

Table & Figure 2: OMNY use by Fare Class Category

Some conclusions can be drawn:

- As Student MetroCards have been replaced with Student OMNY cards, this fare class has a near total adoption of OMNY
- Full fare-paying riders have an OMNY adoption rate of about 80%, significantly higher than the system overall adoption rate of 59%
- Only about 10% of riders with cards categorized as Senior & Disability have switched to OMNY
- Fair fare riders have negligible rates of OMNY adoption

For the Student and Fair Fare classes, these can be attributed to current distribution and expiration patterns, as set by the MTA.

Station-Level Analysis

The next series of Analyses undertaken in this report draws conclusions about OMNY use at the level of individual stations (or station complexes, where applicable). This section also corresponds to section 4 in the IPython notebook. Ridership data for September 2024 was aggregated by subway station, so the current state of OMNY vs MetroCard use can be examined station-by-station. The data is mapped in Figure 3. Table 3 contains the 10 stations with the highest and lowest OMNY use rates. An interactive map of this data can be found in section 4.2 of the notebook.

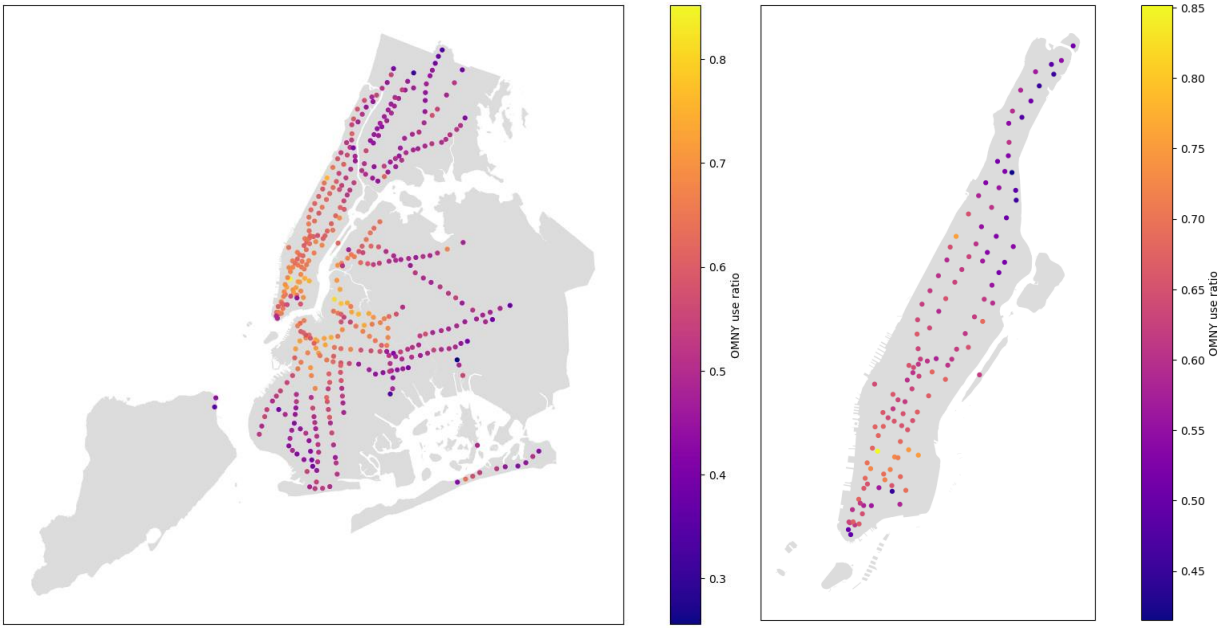

Figure 3: OMNY use rate at New York City Transit rail stations

	Station Name	OMNY Use Rate
1.	West 4 St – Wash Sq (A, C, E, B, D, F, M)	0.852
2.	Bedford Av (L)	0.812
3.	Lorimer St (L) / Metropolitan Av (G)	0.776
4.	116 St-Columbia University (1)	0.755
5.	Morgan Av (L)	0.753
6.	1 Av (L)	0.750
7.	3 Av (L)	0.745
8.	Classon Av (G)	0.741
9.	Graham Av (L)	0.740
10.	Central Av (M)	0.734
…		
417.	233 St (2, 5)	0.397
418.	Jamaica Center – Parsons / Archer (E, J, Z)	0.393
419.	Eastchester-Dyre Av (5)	0.391
420.	Nereid Av (2, 5)	0.377
421.	Jamaica-179 St (F)	0.375
422.	Tompkinsville (SIR)	0.362
423.	Wakefield-241 St (2)	0.345
424.	Canarsie-Rockaway Pkwy (L)	0.344
425.	Woodlawn (4)	0.312
426.	Aqueduct Racetrack (A)	0.256

Table 3: Top and Bottom stations by OMNY Use

We can make several observations about the OMNY use rates when broken down by station:

- Stations closer to CBD have higher OMNY adoption rates. This is especially true for neighborhoods between Midtown and Downtown.
- Williamsburg and Greenpoint have higher OMNY adoption rates than other areas of Brooklyn that are a similar distance from lower Manhattan. In fact, many of the stations in the top 10 (by OMNY use rate) are along the BMT Canarsie Line in Williamsburg.
- 116 St (1) - Columbia University has a higher OMNY adoption rate than the surrounding stations. This could suggest Columbia students have adopted OMNY faster than the average riders.
- W 4 St - Washington Sq (A, C, E, B, D, F, M) has the highest OMNY use rate in the system (~85%). This, along with the previous observation, could indicate college students use OMNY at a higher rate (this is one of several stations close to NYU buildings), however this pattern does not hold as strong for 8th St NYU (R, W) or Astor Pl (6). Other stations named after colleges and universities do not appear to exhibit this pattern, except for maybe Bedford Park Blvd - Lehman College (4), which has a slightly higher OMNY use rate than surrounding stations.
- Grand St (B, D) has a much lower OMNY use rate than other stations in the Lower East Side.
- Bowling Green (4, 5) and South Ferry (1) / Whitehall St (R / W) have lower OMNY use than the rest of the Financial District. One possible explanation is that Staten Island Railroad riders, transferring from the ferry, are responsible for this. At the two fare-collecting stations in Staten Island, OMNY use rates are much lower than those of Financial District stations.
- Mets-Willets Point (7) has a higher OMNY use rate (~67%) than nearby stations (46-52%) on the 7 train. This is likely due to riders going to Mets games, and not residents of Corona or Flushing. 161 St - Yankee Stadium (B, D, 4) does not exhibit a similar pattern: this station has an OMNY use rate only marginally higher than surrounding stations (though Yankee Stadium is far less isolated than Citi Field).
- Aqueduct Racetrack (A) has the lowest OMNY adoption rate in the system, significantly lower than other nearby stations. This is likely due to riders going to Aqueduct Racetrack, and not residents of Ozone Park.

Zip Code-Level Analysis

Several of the patterns that emerged in the station-level analysis, especially the trend of increased OMNY usage closer to Lower Manhattan, can be generalized to a neighborhood level. To determine if there are any demographic factors that correlate with OMNY use, we can aggregate use across Zip Code Tabulation Areas (ZCTAs). These are US Census statistical entities that are used in place of zip codes by the US Census Bureau. They are not the same as zip codes, and USPS realignment, coupled with Census Bureau methodology, means there a number of differences in these boundaries. Nevertheless, significant demographic information is aggregated over ZCTAs, and they provide a good scale for neighborhood level analysis.

In this investigation, each ZCTA with at least one New York City Transit rail station was included. ZCTAs with multiple stations saw this data aggregated into a single data point. This corresponds to section 5 of the IPython notebook. Figure 4 shows the OMNY use rate for each ZCTA[2].

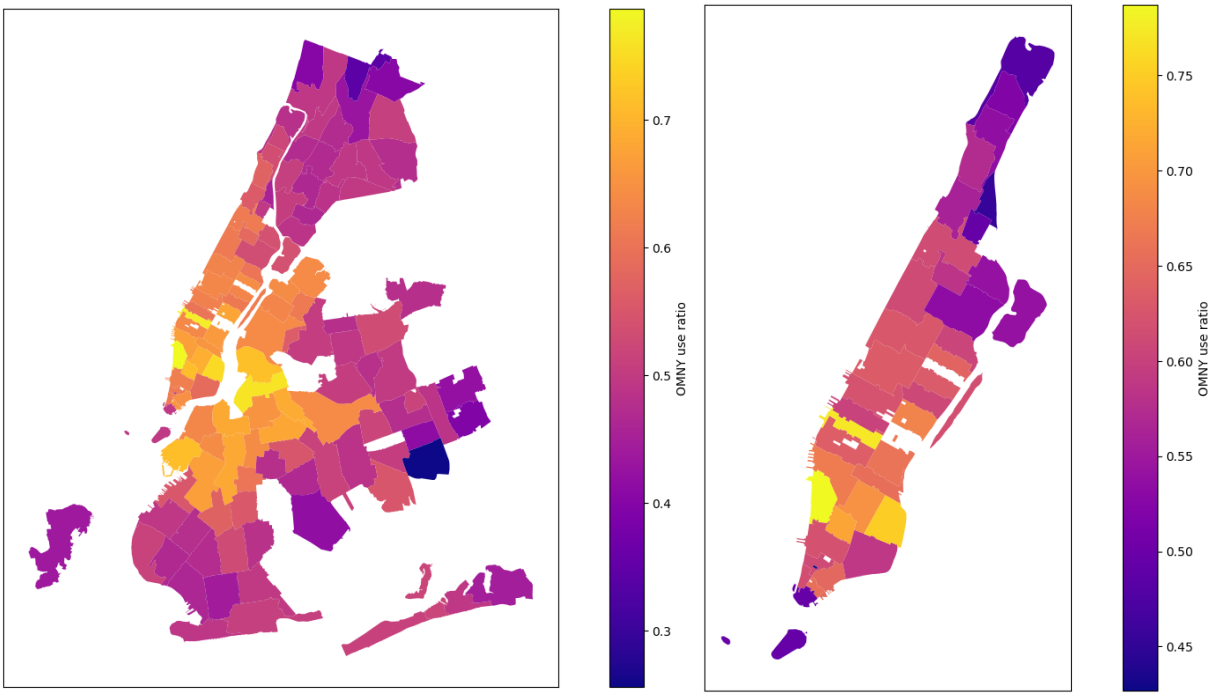

Figure 4: OMNY use rate by Zip Code Tabulation Area[2]

[2] The color spectrum for the highlight of Manhattan is scaled to show the full range in this subset. Grand Central Station has its own ZCTA, leading to a gap in Midtown East. A few Times Square and Bryant Park data points are labelled in ZCTA 10018, which mostly lays between 34th and 42nd streets. This is accounts for the brighter yellow region in Midtown West.

The broad pattern of greater OMNY use in the CBD is also evident from the ZCTA level maps, however a few key differences emerge.

- The East and West Villages are shown to have even higher OMNY use rates than surrounding neighborhoods, with similar OMNY use exhibited in Williamsburg.
- Other areas of Brooklyn, and to a lesser extent Astoria and Long Island City, have OMNY use rates similar to much of Manhattan.
- The Upper East Side and East Harlem have noticeably lower OMNY use rates than the Upper West Side, which becomes more evident at the ZCTA level.
- The OMNY use rates of Bowling Green (4, 5) and South Ferry (1) / Whitehall St (R / W) are evident in ZCTA 10004 (representing the tip of Manhattan and some of the Financial District). ZCTA 1004 has a much lower OMNY use rate than adjacent regions.
- Aside from Aqueduct Racetrack (A), which is the only NYC Transit rail station in ZCTA 11420, all the single-station observations from the prior section, such as patterns around colleges or baseball stadiums, cannot be seen at this magnification.

Comparison to Demographic Features

With OMNY use data tabulated at the ZCTA level, comparisons can be made to demographic features. This investigation examined the relationship between OMNY use and income, age, and population.

INCOME: A great number of behavioral patterns are correlated with income, and the first demographic factor examined was median income. Figure 5 plots the OMNY use rate against Median Household Income in 2022 inflation-adjusted dollars for each ZCTA.

A linear regression was performed on the data to determine if these factors were correlated. This was done both over the whole dataset and at the borough level (Staten Island has one relevant ZCTA and was excluded). The coefficient of determination values (R^2 scores) are in Table 4. As is evident from these scores, there is little correlation between ZCTA income and OMNY use.

Region	R^2 Score
System Overall	0.177
Manhattan	0.024
Bronx	0.254
Brooklyn	0.368
Queens	0.084

Table 4: R^2 scores: OMNY use vs. Income

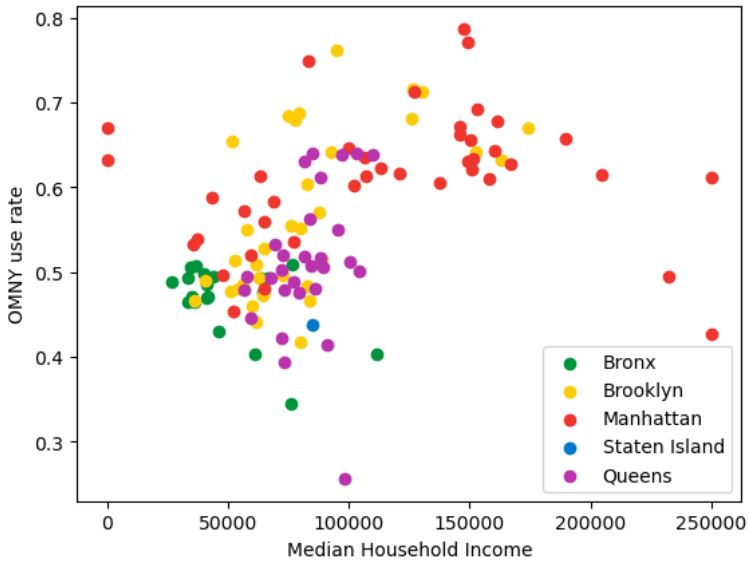

Figure 5: OMNY Use Rate vs. Median Household Income of ZCTA[3]

AGE: Because a pattern emerged where students appeared to be using OMNY at a higher rate, age was examined next. US Census data for median age and ages 18-24 as a percentage of the population was used. They are shown in Figures 6 and 7.

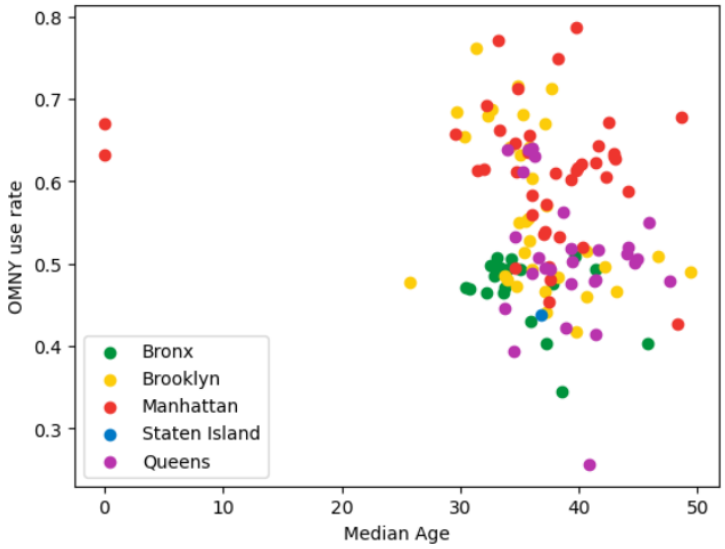

Figure 6: OMNY Use Rate vs. Median Age of ZCTA[3]

[3] Several small ZCTAs in Manhattan have a NYC Subway station but lack income or age data.

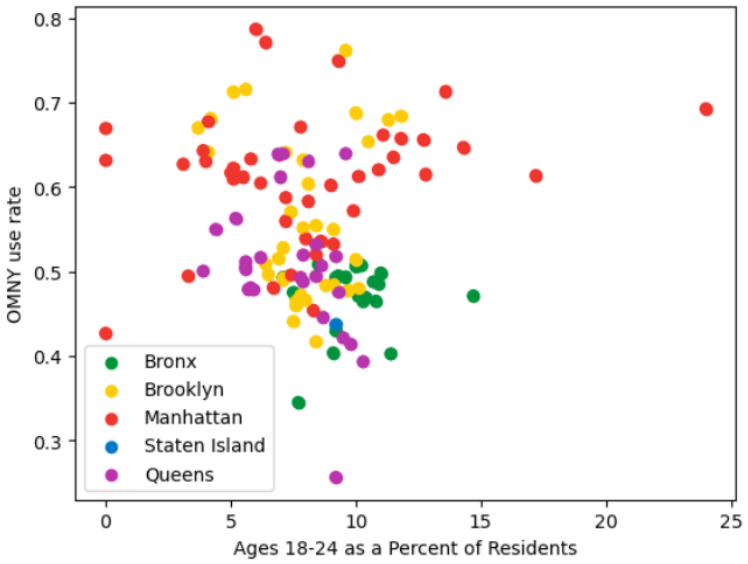

Figure 7: OMNY Use Rate vs. Ages 18-24 (%) by ZCTA[3]

These plots indicate a lack of correlation between either age variable and OMNY use rates. One observation of note is that there are two Manhattan ZCTAs with a high percentage of people between the ages of 18 and 24. In Table 5 are the ZCTAs with the highest percentage of people in this age range. Looking at this data, we can see that these two data points contain NYU and Columbia University. The next several ZCTAs also contain or are adjacent to colleges and universities: Fordham University, Pace University, and NYU, among others. These two data points have moderately high OMNY usage, but this alone does not point to a broader pattern.

	ZCTA	NYC Transit Rail Stations	OMNY use
1.	10003	8 St-NYU (R,W), 3 Av (L), 2 Av (F), Astor Pl (6), 14 St-Union Sq (L,N,Q,R,W,4,5,6)	0.693
2.	10027	135 St (C,B), 125 St (A,C,B,D), 125 St (1), 116 St-Columbia University (1)	0.614
3.	10458	Bedford Park Blvd (B,D), Kingsbridge Rd (B,D), Fordham Rd (B,D)	0.471
4.	10038	Fulton St (A,C,J,Z,2,3,4,5)	0.647
5.	10012	Prince St (R,W), Spring St (6), Broadway-Lafayette St (B,D,F,M)/Bleecker St (6)	0.713

Table 5: OMNY Use Rate and NYC Transit Rail Stations for ZCTAs with Greatest Percentage Age 18-24

POPULATION DENSITY: As OMNY use rates are higher for Manhattan than other boroughs, the relationship between OMNY use and population density was also investigated. Relevant data was sourced from the US Census Bureau. This is shown below in Figure 8.

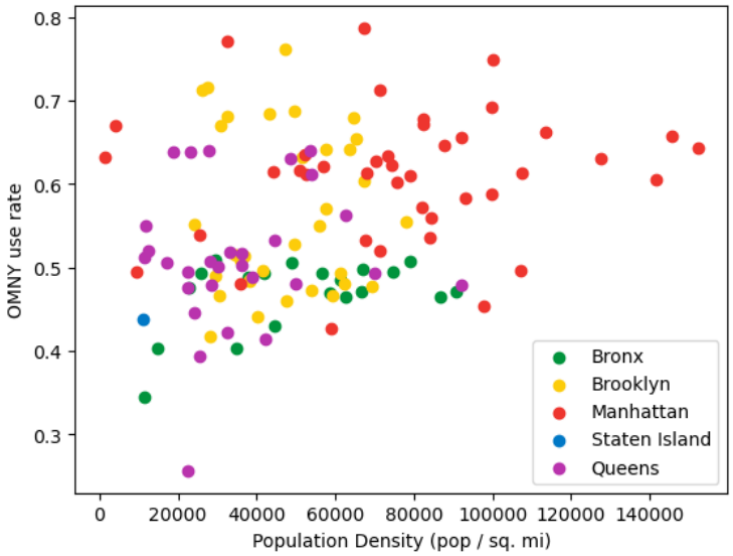

Figure 8: OMNY Use Rate vs. Population Density by ZCTA

A linear regression was also performed on this relationship, and the coefficient of determination was 0.094, suggesting these are not correlated. One possible explanation is that many ZCTAs with high OMNY use have significant commercial presences and may be less dense some residential neighborhoods. We can plot population density (Figure 9) and see that this is true in several areas:

- Population-dense regions in the Upper West and especially Upper East Sides have similar or lower OMNY use rates to more commercial areas in lower Manhattan
- Areas in Downtown Brooklyn that have higher OMNY use rates do not have noticeably higher population densities that ZCTAs in the borough. Some areas, such as Flatbush, Crowne Heights, Bedford-Stuyvesant, and Borough Park, appear denser than Downtown Brooklyn ZCTAs.
- Long Island City and Astoria, whose OMNY use rates are among the highest in Queens, do not have far higher population densities than other neighborhoods (and appear less dense than Jackson Heights).

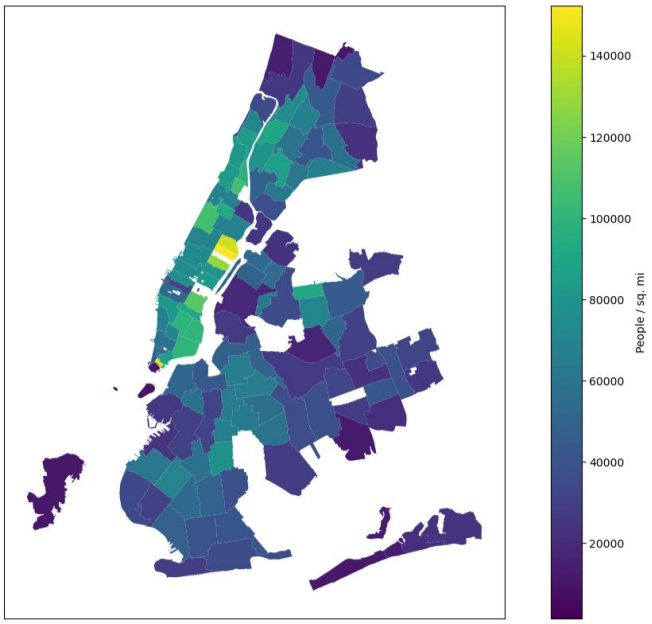

Figure 9: Population Density by Zip Code Tabulation Area

Multi-Factor Comparison: Machine Learning Modelling

While no significant pattern emerged from any one of the demographic factors investigated alone, this does not mean that a more complex relationship cannot be analyzed. In section 5.6 of the IPython notebook, linear and random forest models are created to analyze the relationship between all demographic variables and OMNY use. Borough information is included in the domain using one-hot encoding.

The linear model achieved a coefficient of determination score of 0.450, a slight improvement on prior scores. The random forest scores generally fell between 0.450 and 0.480 when using separate training and test sets (80/20 split chosen at random).

This means these models are better than random guessing, though there remains significant variance in the target (OMNY use rates) that the models cannot account for. There are a few reasons why this unimpressive score might still suggest some correlation:

- Because we are modelling human behavior, the target is inherently noisy, and is likely influenced by a large number of factors that are very difficult to quantify or observe
- Using census data reduced the granularity of the dataset, as data is aggregated (in this case, over Zip Code). This makes it difficult to reflect the individual choices between OMNY and MetroCard that we are trying to analyze.

Temporal Trend Analysis

The prior sections have largely focused on different OMNY use rates with respect to geographic and demographic factors. Because OMNY was introduced relatively recently, these rates continue to change, and worthwhile insight can also come from examining changes in OMNY use over time. This section, which corresponds to section 6 in the IPython notebook, uses data from the first day of every month between February 2022 and September 2024. The use of only one day each month is because of resource limitations. The data is aggregated at the station level, where each data point represents one day at one station. In this section, the upward trend in the data is modelled both at the station and system level, and predictions are made about future OMNY use.

Station-Level Temporal Analysis & Future Prediction

First, the change in OMNY use over time for a single station is analyzed. This can be done for any fare-collecting NYC Transit rail station (see section 6.3 of the IPython notebook), but Brooklyn Bridge-City Hall (4,5,6) / Chambers St (J,Z) was chosen for this section of the investigation. At this step, a linear model was created to represent the relationship between time and OMNY use rate. Figure 10 illustrates the rate over time, with the line of best fit from the model superimposed. The coefficient of determination of the model is 0.675. As this is human behavior data, which tends to be noisy and lead to lower R^2 scores, this indicates a clear correlation that is moderately well captured by the model.

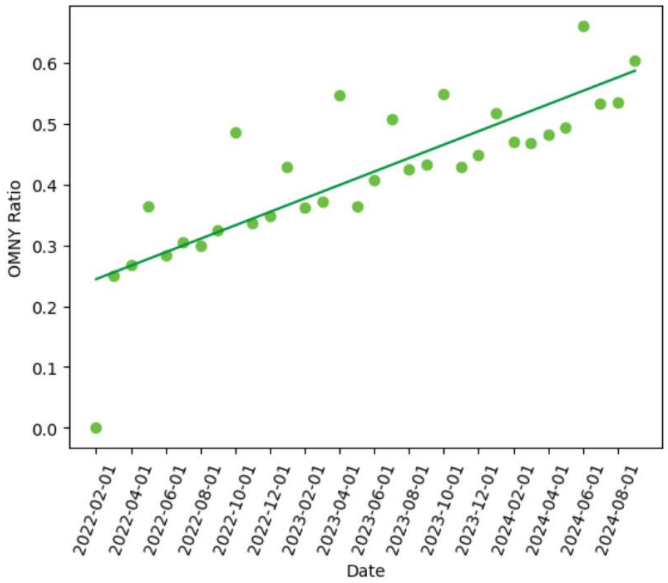

Figure 10: OMNY use rate over time at Brooklyn Bridge -City Hall / Chambers St.

One observation that can be made from the data is a recurring pattern of outliers where OMNY use is higher. These are times when the first of the month lands on a weekend. This suggests that at this station, weekday commuters are less likely to use OMNY than other riders (such as tourists going to the Brooklyn Bridge). The difference between weekend and weekday rates likely accounts for at least some of the noise in data, and an investigation over all days may have a higher coefficient of determination.

We can use this model to make predictions about future OMNY use at this station. For example, the model estimates that OMNY use rates will reach 85% at Brooklyn Bridge – City Hall / Chambers St in August of 2026, and will reach 95% in May of 2027. These predictions don't account for changes in MTA policy or fare class differences, but do provide a fun way to model potential future OMNY adoption if no policy changes are made. Other target OMNY rates can be tested in the IPython notebook.

System-Level Temporal Analysis & Future Prediction

Next, we can aggregate each day's data over all subway stations and make predictions for the system as a whole. Again, a linear model is created from this data to represent the relationship between OMNY use and time. Figure 11 represents the total OMNY rates over time, with the model's line of best fit included. This model has a coefficient of determination of 0.699, suggesting that the model has a good fit for representing this data (as it is noisy behavior data).

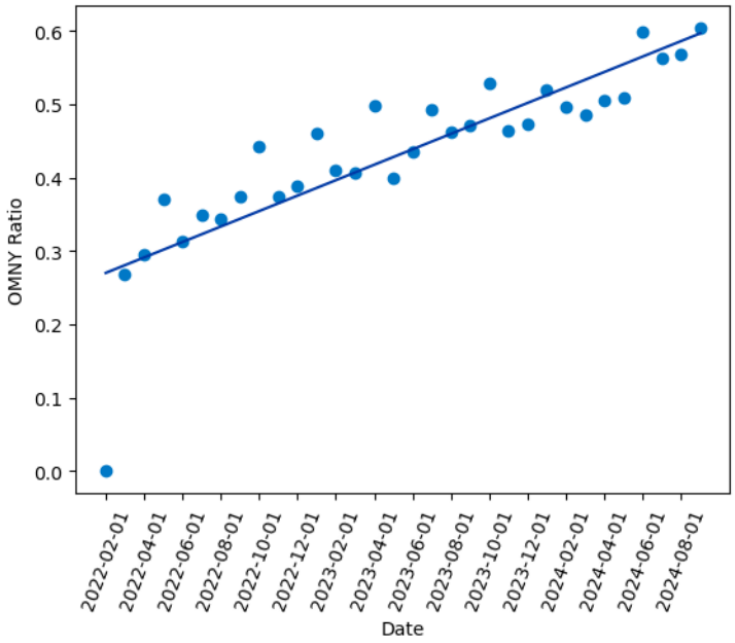

Figure 11: OMNY use rate over time for all NYC Transit Rail Stations

One observation that can be made is that the effects of the weekend are less pronounced on the system as a whole. This means that the higher coefficient of determination might not indicate a far better correlation between OMNY use rates and time since introduction, but rather that the model performs better when there is a smaller difference between weekday and weekend data.

We can again use this model to make predictions about future OMNY use, with the same caveats that the model is not sensitive to MTA policy changes or fare class differences. The model estimates that NYC Subway and SIR riders will reach 85% OMNY use in September of 2026 and will reach 95% use in June of 2027.

Conclusions

This investigation examined the rates of OMNY use with respect to a large number of variables, with differing strengths of correlation. The systemwide analysis highlighted that OMNY use differs greatly across fare class category. This means that MTA policy and distribution of OMNY Cards vs MetroCards has a significant impact on OMNY use rates. While this is not surprising, one less trivial insight to make is that fare classes other than full fare are responsible for much of the remaining MetroCard use: full fare riders use OMNY 80% of the time, compared to a system average of 60%.

Several sections focused on geographic and demographic factors, with varying results. A strong geographic pattern emerged wherein stations closer to the Lower Manhattan CBD saw a greater OMNY use rate, with the highest rates being reported in Lower Manhattan and Williamsburg. A number of smaller geographic and demographic quirks emerged, such as heightened OMNY use around major universities and Mets-Willets Point, and lowered OMNY use at Aqueduct Racetrack, Grand St, and The Battery. These patterns largely disappeared when analyzed at levels higher than individual stations, and no larger demographic correlations emerged from investigating OMNY use against US Census data.

The final section highlighted change over time, and illustrated a clear trend of increased OMNY use among NYC Subway and SIR riders. The correlation between these factors was strong, and the model representing this correlation was used to make quick predictions about future OMNY use.

Datasets Used

MTA Datasets

- MTA Subway Hourly Ridership: Beginning July 2020, *Metropolitan Transportation Authority*, data.ny.gov, https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-Beginning-July-2020/wujg-7c2s/about_data
- MTA Subway Stations and Complexes, *Metropolitan Transportation Authority*, data.ny.gov, https://data.ny.gov/Transportation/MTA-Subway-Stations-and-Complexes/5f5g-n3cz/about_data
- MTA Colors, *Metropolitan Transportation Authority,* data.ny.gov, https://data.ny.gov/Transportation/MTA-Colors/3uhz-sej2/about_data

US Census Datasets

- U.S. Census Bureau. "TOTAL POPULATION." *Decennial Census, DEC Demographic and Housing Characteristics, Table P1*, 2020, https://data.census.gov/table/DECENNIALDHC2020.P1?g=040XX00US36$8600000. Accessed on October 25, 2024.
- U.S. Census Bureau, U.S. Department of Commerce. "Annual Geographic Information Table." Geography, GEO Geography Information, Table GEOINFO, 2023, https://data.census.gov/table/GEOINFO2023.GEOINFO?q=area&g=040XX00US36$8600000. Accessed on October 25, 2024.
- U.S. Census Bureau. "Income in the Past 12 Months (in 2022 Inflation-Adjusted Dollars)." American Community Survey, ACS 5-Year Estimates Subject Tables, Table S1901, 2022, https://data.census.gov/table/ACSST5Y2022.S1901?q=income&g=040XX00US36$8600000. Accessed on October 25, 2024.
- U.S. Census Bureau. "Age and Sex." American Community Survey, ACS 5-Year Estimates Subject Tables, Table S0101, 2022, https://data.census.gov/table/ACSST5Y2022.S0101?q=age&g=040XX00US36$8600000. Accessed on October 25, 2024.
- 2020 TIGER/Line Shapefile: Zip Code Tabulation Areas, *U.S. Census Bureau*, https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

City of New York Datasets

- Borough Boundaries, *City of New York*, data.cityofnewyork.us, https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm