A Causal Analysis on Public Transportation in NYC

Nathaniel del Rosario
Halıcıoğlu Data Science Institute
University of California, San Diego
San Diego, CA 92092
nadelrosario@ucsd.edu

Abstract

Public Transportation within the United States can be considered underdeveloped in many ways with respect to countries in Asia and Europe. However one city that has extensive infrastructure, New York City. NYC is has developed an extensive system of public transport, but where can it improve? More importantly, what variables are influenced by public transport accessibility and vice versa? We aim to develop an understanding of these ideas. In the end we design an accessibility metric, a baseline model and discuss potential improvements to the model.

1 Introduction

The New York City public transportation is arguably one of the best in North America, providing many different methods such as metro, ride share, and bike as the most common. However, it is not a perfect system, possessing its own set of shortcomings. For example, compared to Tokyo's public transportation infrastructure, NYC's system is not as expansive and under serves more areas compared to Tokyo. Considering such context, the question arises, "just how under served are parts of New York City in the scope of public transportation?" and furthermore, are there any effects in other domains due to these under served areas?

I hypothesize that there are in fact different factors whose effects that are correlated with some areas being under served specifically by the NYC metro such as these areas being more likely to experience more ride share and bike usage. Upon witnessing any correlation, the next question becomes “is there causation as well?” Answering such uncertainty is the goal of this project.

This question is important because it involves using population, ridership, geo-spatial, and tract data to help people not only understand their commute as well as identify potential causality between different events and transportation accessibility. On average people will spend at least an hour commuting to and from work and school, and this is a huge chunk of our day (1/16 if you get a full 8 hours of sleep!) Additionally, public transportation companies can benefit greatly from this analysis as they can modify their strategy to appeal more to commuters and plan where to expand service to those who are under served. Lastly, the average citizen would benefit from this information because it could convince them to take public transportation instead of contributing to the increasing problem of traffic congestion in major metropolitan areas.

2 Related Work

In their study, Jin, Kong, and Sui 2019 investigated the distribution of Uber services in New York City and its correlation with urban transportation equity. They found that Uber pickups were disproportionately concentrated in areas with higher incomes, suggesting a potential inequality in service provision. This investigation aligns with our project's aim to explore the correlation between Uber pickups and low-income areas.

Tang et al. 2019 examined the spatio-temporal characteristics of urban travel demand, highlighting the significant influence of various factors such as road density, subway accessibility, and commercial areas. Their findings provide valuable insights into the dynamics of transportation demand, which could inform our analysis of Uber pickup patterns.

While not directly related to our project, Hess and Almeida 2007 studied the impact of proximity to light rail transit stations on property values in Buffalo, New York. Their research underscores the complex interplay between transit accessibility and property values, offering a contrasting perspective to our investigation into the relationship between Uber pickups and income levels.

Additionally, Freeman et al. 2013 explored the association between neighborhood walkability and active travel behavior in New York City. Their findings suggest that increased walkability is linked to higher engagement in active modes of transportation such as walking and cycling. Although focusing on a different aspect of urban mobility, their study provides insights into factors influencing travel behavior, which may inform our analysis.

3 Datasets and Feature Layers

The transportation dataset encompasses crucial infrastructure points vital for urban mobility analysis within New York City. It comprises layers of metro stops and Citi Bike stations distributed across Manhattan, Queens, and Brooklyn, facilitating a comprehensive understanding of transit accessibility. These layers provide essential information such as tract boundaries and point geometry for each station, enabling spatial analysis of transit coverage and its relationship with socioeconomic factors. Moreover, the inclusion of bike stations intersecting with low-income tracts offers insight into equitable access to alternative transportation modes. Additionally, the Uber and Lyft dropoff dataset enriches the analysis by providing detailed pickup and dropoff points along with timestamps, enabling temporal analysis of ride-sharing patterns in conjunction with other transportation modes.

Beyond transportation infrastructure, the dataset encompasses socioeconomic indicators crucial for understanding urban dynamics in New York City. It includes income distribution by tract, providing insights into the spatial distribution of wealth and potential disparities across neighborhoods. Complementary layers such as census tracts and low-income census delineate geographic boundaries and areas characterized by low-income populations, offering a contextual backdrop for analyzing transportation equity. Furthermore, the inclusion of gentrification data in choropleth format allows for the examination of neighborhood transformations over time, complementing the analysis of low-income areas and their relationship with public transportation accessibility. Overall, this diverse dataset facilitates a comprehensive examination of the intersection between urban transportation, socioeconomic factors, and spatial dynamics in New York City.

4 Analysis

To investigate the spatial relationships between rideshare dropoffs, bike and metro stations, and socioeconomic factors in New York City, we employ processes in several key steps aimed at isolating patterns of transportation accessibility and their association with income distribution across different tracts within the city.

4.1 Rideshare Dropoffs By Tract

Firstly, we analyze rideshare dropoffs by tract, utilizing the rideshare dataset to identify the spatial distribution of dropoff points across New York City. This initial step provides insights into the geographic dispersion of rideshare utilization and potential areas of high demand. Next we look metro and bike stations by tract, comparing the distribution to rideshare dropoffs.

4.2 Buffers

Following this, we will create buffers around bike and metro stations to delineate areas of influence for each mode of transportation. By overlaying these buffers with the rideshare dropoffs choropleth, we aim to assess the proximity of rideshare activity to transit hubs, thereby gauging the interconnectivity between different transportation modes.

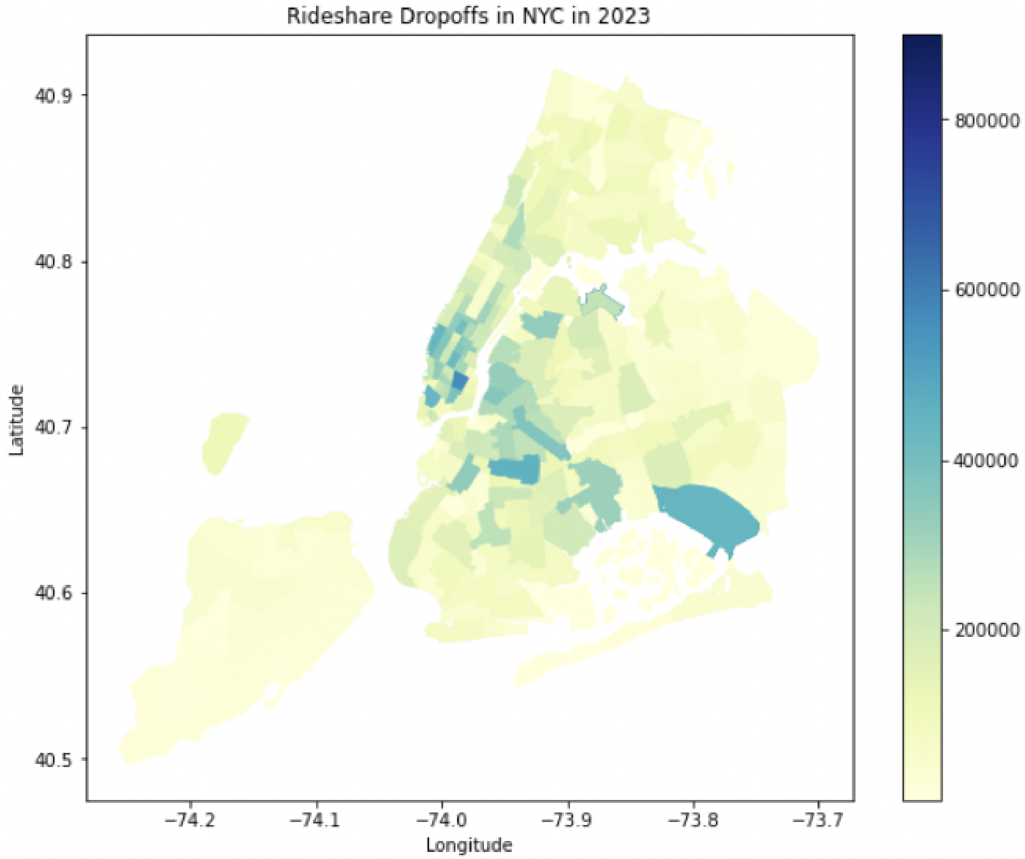

Figure 1: Uber and Lyft Dropoffs, 2023

4.3 Buffers with Median Income

Subsequently, we will overlay the buffers for bike and metro stations with an income choropleth to examine the relationship between transportation accessibility and income levels. This comparative analysis will shed light on potential disparities in access to transportation services based on socioeconomic factors.

4.4 Buffer Choropleths

Additionally, we will aggregate the buffers per tract for bike and metro stations to generate choropleth maps illustrating the concentration of transportation infrastructure within each tract. This step will facilitate a nuanced understanding of transit coverage across different neighborhoods and its implications for urban mobility and equity.

By systematically executing these analytical procedures, we aim to elucidate complex spatial relationships between transportation infrastructure, rideshare activity, and income distribution in New York City. The resulting insights will contribute to a deeper understanding of urban transportation dynamics and inform strategies for enhancing accessibility and equity in the city's transit network.

4.5 Quantifying Accessability

Based on the scales of the number of rideshare dropoffs and metro / bike station buffers, we first normalize the rideshare dropoffs since its scale is in the hundred thousands per tract. Then we define a weighted sum which has tunable weights for each method of public transportation. For simplicity we use the following formula below:

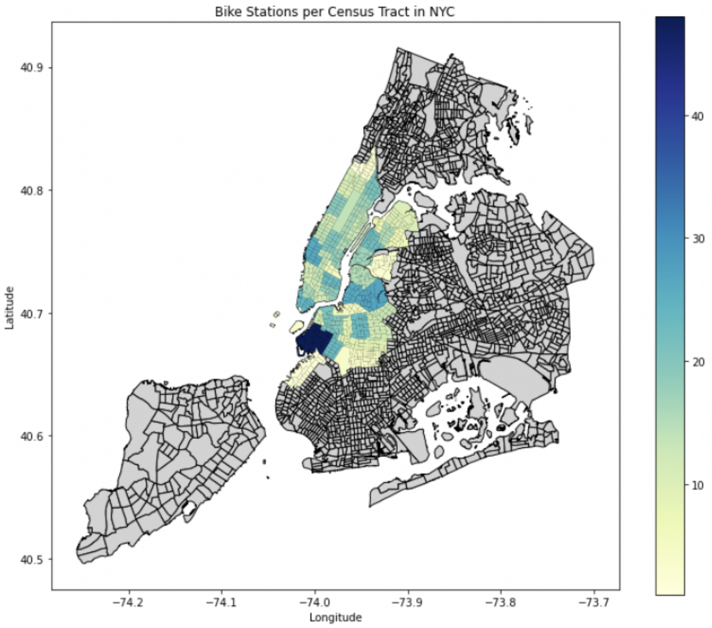

Figure 2: Bike Stations per Tract

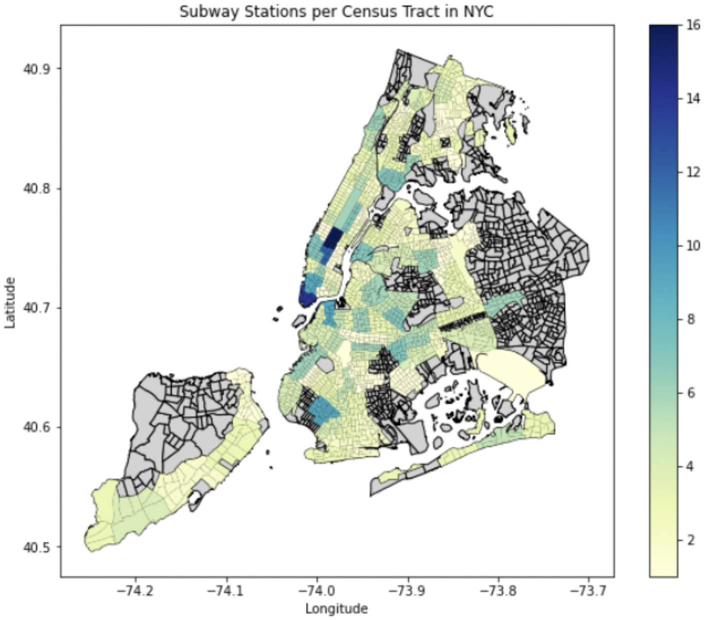

Figure 3: Metro Stations per Tract

$$\text{Accessability} = \log(\sum(.5 * \text{metro buffer}, .25 * \text{bike buffer}, .25 * \text{rideshare}))$$

We take the log of the accessability to get our score to ensure that extreme large values do not skew our score. Doing so ensures the metric is robust to outliers.

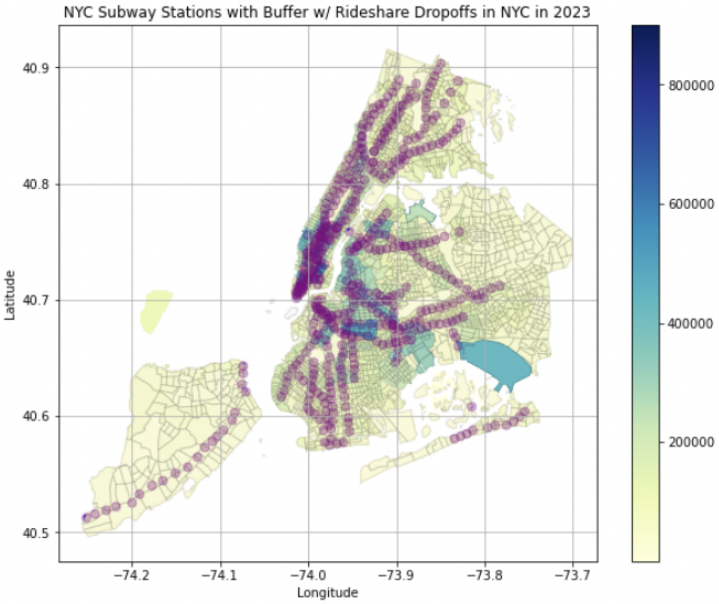

Figure 4: NYC Metro Station Buffers with Uber and Lyft Dropoffs, 2023

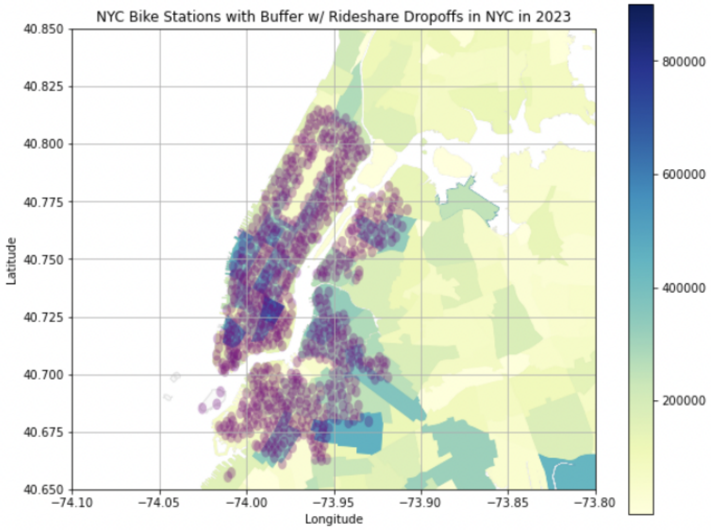

Figure 5: NYC Bike Station Buffers with Uber and Lyft Dropoffs, 2023

4.6 Modeling

Beginning with our EDA, we found that Manhattan metro stations had quite a lot of stops, with more of the underserved areas not being on the island. Additionally, this was more apparent with bike stations, as they were even more concentrated in Manhattan. Overlaying this with the median income choropleth, we saw that a huge number of the stations were also concentrated in Manhattan, with an even larger proportion being in downtown Brooklyn / DUMBO. When we switched to analyzing the counts of metro and bike stations and their buffer ranges being inclusive in counts, we saw a much smoother transition on the choropleths, but still the same overall correlation from the birdseye view.

Overlaying these buffers on income, we used an transparent opacity for each buffer as well as a sequential color spectrum for the individual choropleth to see if the buffer density correlated with income as well. In all 4 plots, (Bike Buffers on Income Choropleth, Station Buffers on Income

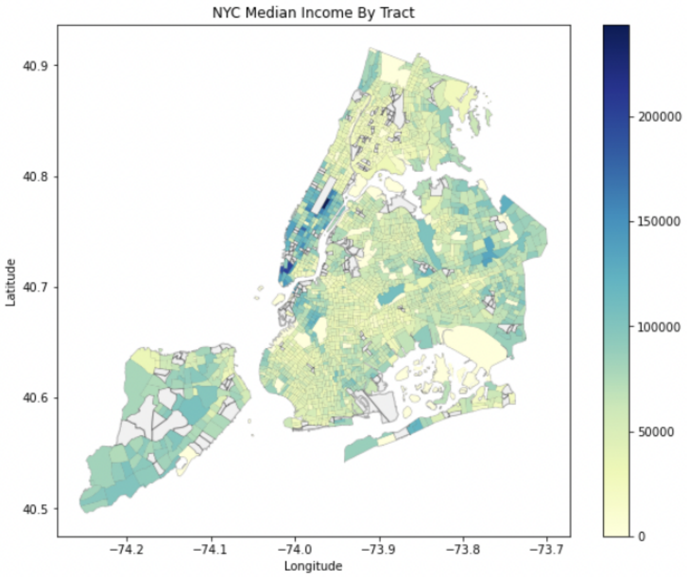

Figure 6: Median Income Tracts

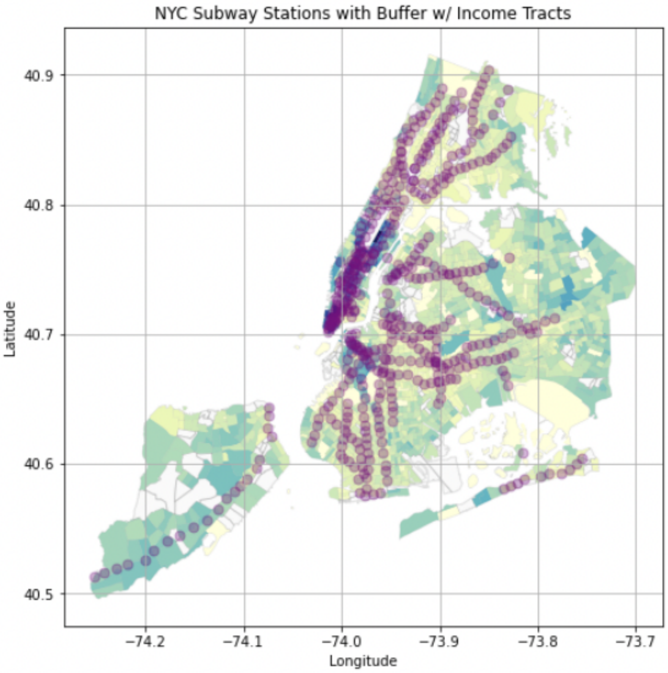

Figure 7: NYC Metro Station Buffers with Median Income Tracts

Choropleth, and Bike Buffer Count / Metro Buffer Count Choropleths) we saw that there still seemed to be a visible correlation, displayed by higher income / darker areas having more purple buffers.

To address the strength / validity of the correlation, we then designed a metric to first combine all 3 methods of transportation's buffers / points and aggregate by census tract, and then log transform it to be robust to outliers. We then used income to predict the accessibility score using a regression model with a tunable log transformation, resulting in an RMSE of .1785, which in our opinion is very admissible.

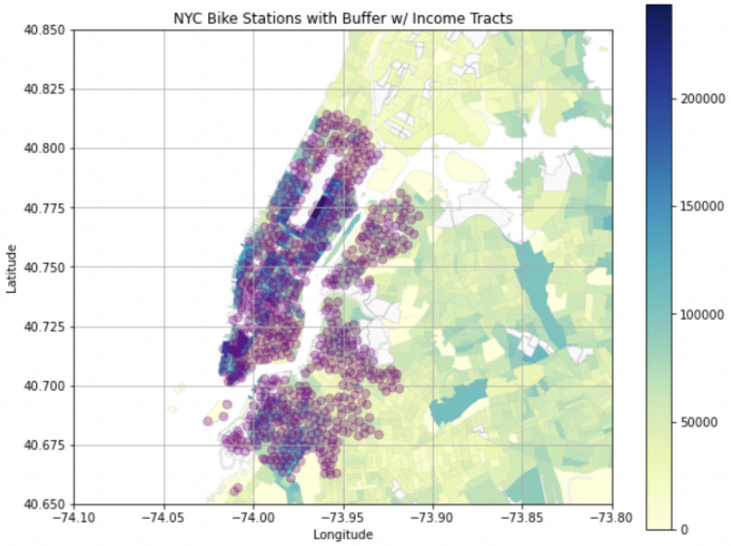

Figure 8: NYC Bike Station Buffers with Median Income Tracts

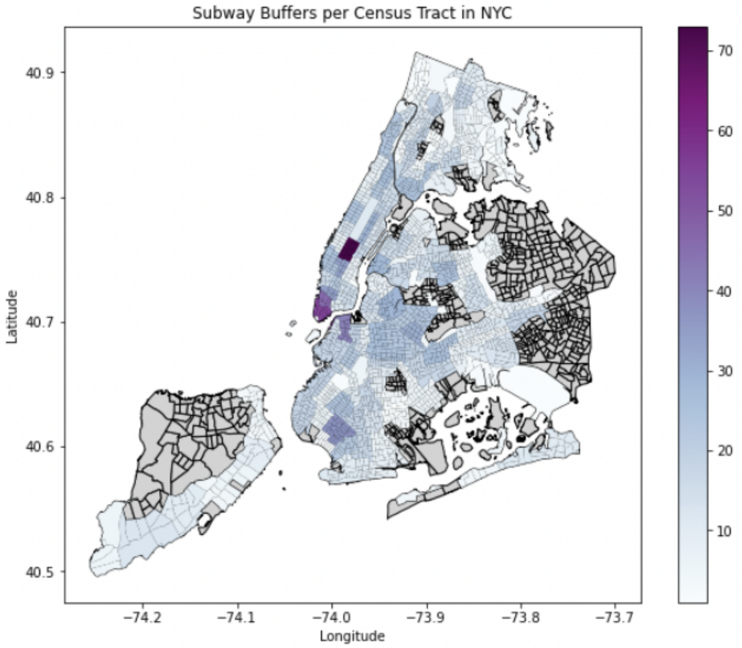

Figure 9: NYC Metro Station Buffers Choropleth

$$\text{Accessability} = w * \text{income}$$

Future models would incorporate possible features such as:

$$\text{Accessability} = w_1 * \text{income} + w_2 * \text{gentrification} + w_3 * \text{population density}$$

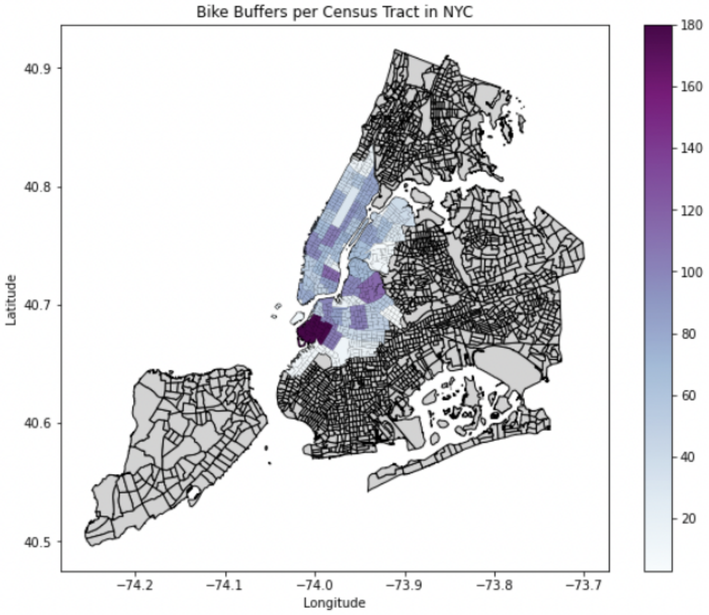

Figure 10: NYC Bike Station Buffers Choropleth

5 Discussion

1. Uber Pickup Distribution and Income: Our findings corroborate the study by Scarlett T. Jin et al. (2019), which highlighted the unequal distribution of Uber services, particularly with fewer pickups in low-income areas. We observed a correlation between buffer density and income, indicating that underserved areas tend to have lower incomes. This supports the notion that transportation equity is closely tied to socioeconomic factors.

2. Spatio-Temporal Characteristics of Urban Travel Demand: Tang et al. (2019) emphasized the significance of various factors, including subway accessibility and point of interests, on travel demand. While our analysis focused on the spatial distribution of transportation services and income, it aligns with the broader understanding that multiple factors influence travel patterns and accessibility in urban areas.

3. Impact of Transit Proximity on Property Values: While not directly related to our analysis, the study by Hess and Almeida (2007) examined the impact of proximity to light rail transit on property values. Our findings complement this literature by highlighting the spatial distribution of transportation services and their potential association with property values and income levels. A potential addition to our analysis would be incorporating property values along with income to predict accessability.

4. Neighborhood Walkability and Active Travel: Freeman et al. (2013) highlighted the association between neighborhood walkability and active travel. While our analysis did not directly focus on active travel, it indirectly relates to the broader concept of transportation equity and accessibility, which can influence active modes of travel.

Our analysis extends the existing literature by providing a spatial perspective on transportation equity and accessibility in New York City. By integrating GIS data with socioeconomic data through a custom tunable accessability score, we identified spatial patterns of transportation services and their relationship with income levels. Additionally, our regression modeling approach provided quantitative insights into the relationship between transportation accessibility and income. The findings contribute to a better understanding of the complex interplay between transportation, socioeconomic factors, and spatial disparities in urban areas.

Throughout the analysis, several trade-offs and decision points were considered:

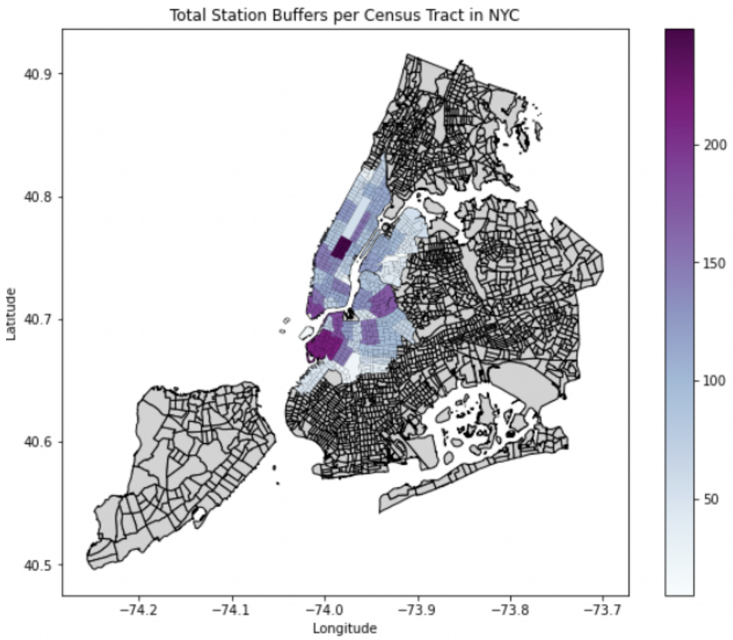

Figure 11: Sum of Metro and Bike Station Buffers Choropleth

1. Buffer Width: The choice of buffer width for subway and bike stations could impact the spatial representation of accessibility. We chose buffer widths based on practical considerations and the spatial extent of transportation services.

2. Spatial Operations: Spatial joins and overlays were utilized to integrate transportation data with census tract boundaries. These operations involved trade-offs in terms of computational complexity and accuracy, particularly when dealing with large datasets.

3. Machine Learning Techniques: The selection of machine learning techniques, such as regression modeling, involved considerations regarding model complexity, interpretability, and predictive performance.

Overall, through navigating these trade-offs we were able to conduct a comprehensive analysis that sparks discussion on transportation equity and accessibility in New York City.

6 Results & Conclusion

From our current analysis, we did not *completely* answer our research question, however we did find results that suggest that our initial hypothesis was in the correct direction. Referring back to what we said, we hypothesized that there are different factors whose effects that are correlated with some areas being under served specifically by the NYC metro such as these areas being more likely to experience more ride share and bike usage.

Income ended up being one of these factors, and although we didn't make regression models involving gentrification and bike usage versus dropoffs, it is realistic and possible to find results from these that support our hypothesis. Additionally, if given more datasets with tracts, geometry, and some specific feature such as housing prices, zoning, and GDP / output, we believe that we could find correlations between these features and accessability as well.

What is great about this apprach we took was that it is replicable for other regions. The only factor stopping someone from performing the same analysis on another dense metropolitan area if the availability of the data, as well as how recent it is. Overall, given the data we were able to access, this project was successful, and at the current stage we were able to conclude a correlation between the different variables, while leaving some room on the table for different future analyses.