Jared Lieberman (jrl2162@columbia.edu)
STAT 5205 Linear Regression Models
22 December 2023

Final Project Report

**Introduction**

The primary goal of our project is to explore relationships between building energy efficiency and sustainable transportation. At a high level we are interested in whether certain types of people take or have the option to choose sustainable transportation, and generally in what contexts they do so. Our case study is New York City, and we utilize inferential procedures, relying on linear regression models, to study more specifically relationships between building energy usage, Citi Bike ride patterns, and neighborhood socioeconomic and housing data. These data not only provide fertile ground for applying statistical techniques from the course, but also lay the groundwork for further research with important policy implications.

**Data Collection**

We built a dataset from multiple data sources that aggregated relevant characteristics across New York City neighborhoods. Our geographic unit of analysis is the City of New York-designated Neighborhood Tabulation Area (NTA). There are 195 NTAs within the city. The technical process of collecting, cleaning, and preparing the data was all conducted in Python and uploaded to a GitHub repository for distributed version control. We prioritized writing clean and accurate code that was computationally efficient.

We pulled Citi Bike data from the company's system data website, which includes downloadable files of Citi Bike Trip Histories from 2013 onwards. We set up an amazon S3 account and pulled this data manually. We pulled just from the year 2021, the same year as our building energy usage dataset, and retrieved 27,661,451 rows of data. The data have information about starting latitude and longitude of each ride, ending latitude and longitude, whether the rider is a member or not, the start and end time of the ride, and the type of bike.

One key subset of our data was pulled from NYC Open Data Neighborhood Tabulation Area (NTA) data. There are 195 NTAs split between the different boroughs. Thus, the NTA datasets each had 195 rows, frequently with a few thousand columns. These data were distinguished by theme between housing, sociological, economic, and demographic characteristics of New York City's population. From these categories, we were able to gather granular data that could provide further insight into different aspects of our research question. For example, we utilized data on household income, industry of occupation, and ratio of income to poverty level from the economic dataset. We also *gathered* important data regarding housing, for example owner- or renter-occupied units, which is particularly useful for creating a more holistic understanding of residential properties. From the demographic dataset, we pulled data for education levels across the NTAs.

For our data on building energy usage, we again utilized NYC Open Data from the "Energy and Water Data Disclosure for Local Law 84 2022" dataset. This included data for over 27,000 buildings across NYC, with 250 columns regarding general and energy-related characteristics of each building. We paid particular attention to the way in which the building was characterized, for example by fields regarding primary building type (i.e. residential, office, etc), and how energy usage was recorded for the building (i.e. water usage, thermal energy usage at the source vs. site, etc.).

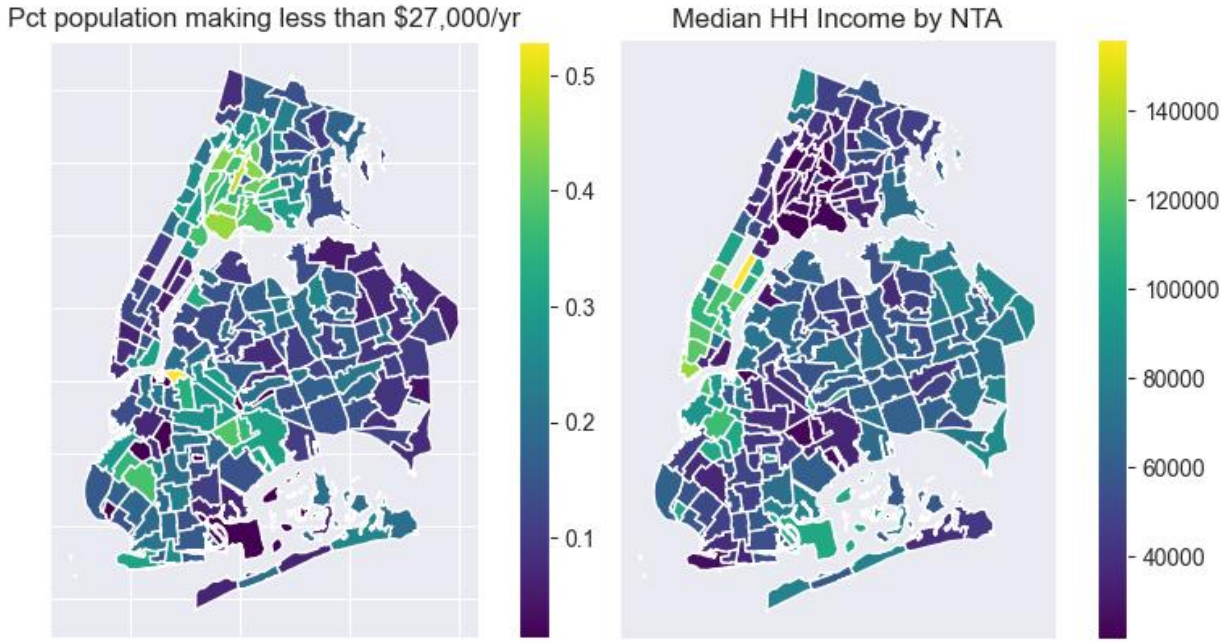
We used geospatial data across each data source not only for our analysis, but also for our exploratory phase of creating maps to better understand the spread and density of Citi Bike rides and building energy usage. For the technical aspects of this task, we used the Python library GeoPandas to merge our initial datasets with respect to the geometry columns. From there, we could make simple calculations in aggregate, for example the number of stations and rides that start in each NTA. In our research process, we spent some time thinking about and coding distance functions to analyze the experiences of riders between different NTAs in the most granular way. However, due to the nature of elevation differences across the city, this was unfortunately not feasible. In order to address that, we considered using the Google Maps Elevation API and even tested a few trips between different stations, but the results necessary to track every trip across every station would not have been computationally or financially feasible. Ultimately, we were able to aggregate fundamental information about Citi Bike stations and rides, building energy usage, and socioeconomic data across each of the 195 NTAs.

For the technical process of data collection, because we pulled data from multiple sources and, due to the nature of the group project working on different computers, it become important to implement a framework that would allow us to extract, transform, and load the data onto a new computer easily and efficiently. We to a data engineering approach to solve this problem by creating Python scripts to automate pulling data from three different APIs, ensuring we had valid credentials, and downloading the data in appropriate formats.

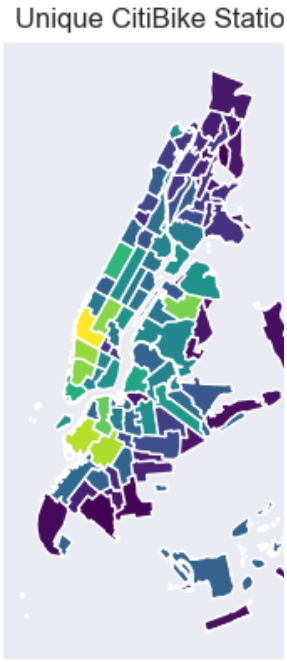
Throughout the process of cleaning the data, we were cognizant of small details and decisions that we knew would impact our analyses later on. In our building energy usage dataset, we conducted data quality check in order to remove data points that contained certain flags. This ensured we were only using reliable data for buildings. For example, if the energy meter was flagged as having gaps in data, we remove this building from the data set. We also removed data points that were missing important data columns, such as latitude and longitude, primary property use type, year built, and energy use metrics. Regarding missing values, we inspected each column in order to decide if the rows or columns needed to be removed or if we could fill the values with zeros or other appropriate values. Most buildings, for example, do not have any green power onsite, and thus we replaced NaN values with a zero for that column. Other decisions we had to make included whether to include Staten Island in our data as well as whether to normalize the unit of analysis for socioeconomic and housing data by an NTA's population. We ultimately created a dataset with 172 rows and 40 columns.

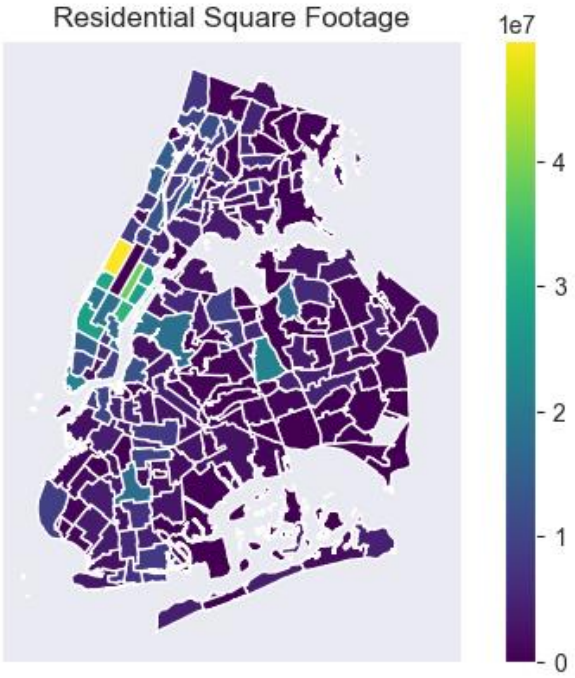
**Exploratory Analysis**

The exploratory analysis phase of our research was of particular interest for us due to the geospatial nature of our data. From our initial dataset, we selected variables of interest that could illustrate important geographic trends, and mapped them across NTA. We provide 5 of the gradient maps below that are representative of how the exploratory analysis helped us better understand next steps in our research. We can see from the income-related maps, Pct population making less than $27,000/yr and Median HH (Head of Household) Income by NTA, that there is significant variation across the city. The population under $27,000, the poverty line, specifically targets displaying pockets of poverty.

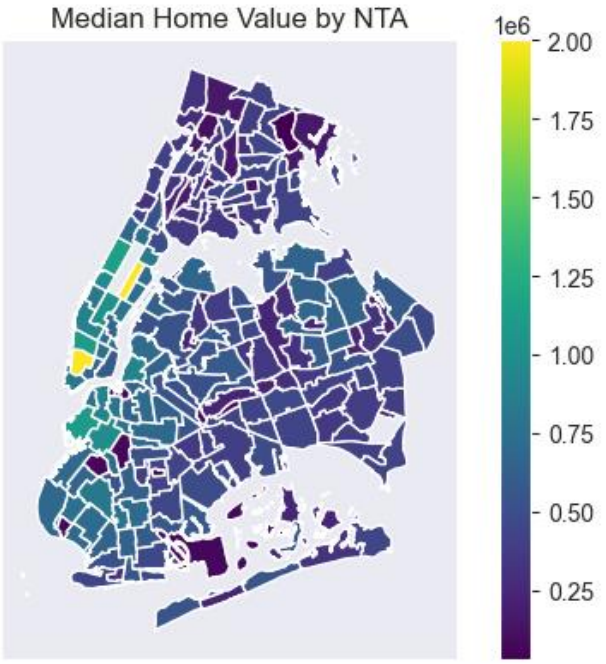


The Unique Citi Bike Stations by NTA map is important for understanding how stations are concentrated in certain areas. Also, of note is how NTAs are missing from the map, and thus do not have any Citi Bike stations. We consider this in our regression analysis.

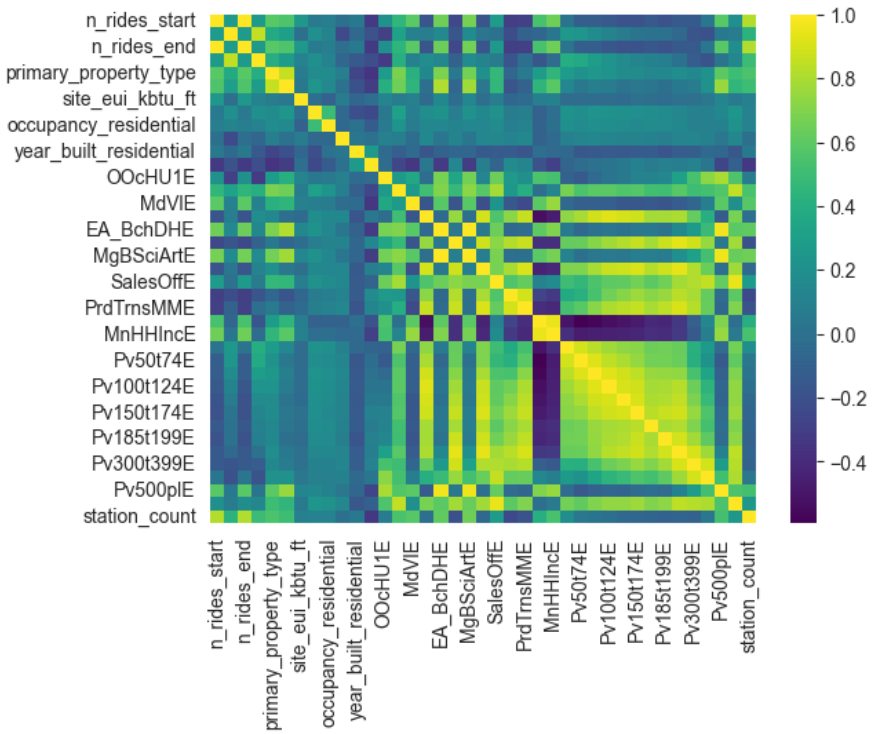



The two maps below, Median Home Value by NTA and Residential Square Footage, are intended to show trends related to buildings. Maps like these helped us parse through property-related data in order to select variables that would be of use to our analysis.




Below, we provide the correlation heatmap that we created in order to make initial findings regarding relationships between our covariates and response variable. The variable descriptions will be discussed in the next section.




**Statistical Models**

We created a number of models by running a full model on all of our variables and then categorizing our variables in order to create reduced models. Our response variable is "site_eui_kbtu_ft" from our buildings energy usage dataset. This pertains to the energy use intensity at the site of the building in units of kilo British thermal units / ft squared. The NTA data tends to be coded, and the overall relevant columns that are not easily defined are: OOcHU1E – Owner occupied units; ROcHU1E – Renter occupied units; MdVlE – median home value (dollars); EA_LTHSGrE – Less than High School; EA_BchDHE - Bachelor's degree or higher; EA_HScGrdE – High school graduates; MgBSciArtE – industry of occupation is in management, business, science, and arts; SrvcE – industry of occupation is in service; NRCnstMntE – industry of occupation is in natural resources, construction, and maintainence; SalesOffE - industry of occupation is in sales; PrdTrnsMME - industry of occupation is in production, transportation, and material moving; MnHHIncE - Mean household income (dollars); Pop_1E – population; and the columns that begin with Pv correspond to intervals of ratios to the poverty level. We also ran models by using our complete dataset as well as only NTAs that had contained Citi Bike data (i.e. rides, stations, etc.). The model output summaries are provided below for each full model, with the complete dataset on the left and the subset on the right.

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 2.083e+02 | 4.681e+02 | 0.445 | 0.656993 | |
| n_rides_start | -7.864e-04 | 3.841e-04 | -2.048 | 0.042548 | * |
| avg_ride_duration_secs_start | -1.084e-02 | 3.828e-03 | -2.833 | 0.005321 | ** |
| n_rides_end | 8.641e-04 | 3.816e-04 | 2.264 | 0.025147 | * |
| avg_ride_duration_secs_end | 1.633e-02 | 4.629e-03 | 3.528 | 0.000573 | *** |
| multifamily_housing_gross | -3.723e-06 | 1.015e-06 | -3.667 | 0.000352 | *** |
| largest_property_use_type_residential | 5.460e+00 | 1.147e+01 | 0.476 | 0.634690 | |
| occupancy_residential | -2.293e-01 | 2.041e-01 | -1.124 | 0.263200 | |
| occupancy_non_residential | -4.102e-02 | 1.798e-01 | -0.228 | 0.819917 | |
| year_built_residential | -1.334e-01 | 2.072e-01 | -0.644 | 0.520821 | |
| year_built_non_residential | 2.378e-02 | 1.785e-01 | 0.133 | 0.894244 | |
| OOcHU1E | 5.967e-03 | 2.554e-03 | 2.336 | 0.020949 | * |
| ROcHU1E | 3.957e-03 | 2.261e-03 | 1.750 | 0.082359 | . |
| MdVlE | -1.325e-05 | 2.696e-05 | -0.491 | 0.624008 | |
| EA_LTHSGrE | 1.077e+02 | 1.257e+02 | 0.857 | 0.392755 | |
| EA_BchDHE | 2.281e+02 | 1.428e+02 | 1.597 | 0.112691 | |
| EA_HScGrdE | -1.800e+02 | 1.658e+02 | -1.086 | 0.279438 | |
| MgBSciArtE | -4.429e+02 | 1.870e+02 | -2.368 | 0.019289 | * |
| SrvcE | -6.405e+01 | 2.069e+02 | -0.310 | 0.757382 | |
| SalesOffE | -6.871e-03 | 4.825e-03 | -1.424 | 0.156741 | |
| NRCnstMntE | -1.397e-02 | 8.032e-03 | -1.739 | 0.084338 | . |
| PrdTrnsMME | -6.170e-03 | 8.325e-03 | -0.741 | 0.459878 | |
| MdHHIncE | 1.651e-03 | 8.928e-04 | 1.849 | 0.066587 | . |
| MnHHIncE | -3.446e-04 | 4.894e-04 | -0.704 | 0.482580 | |
| PvU50E | -3.623e+01 | 1.868e+02 | -0.194 | 0.846477 | |
| Pv50t74E | 9.468e+01 | 2.156e+02 | 0.439 | 0.661288 | |
| Pv75t99E | 1.116e+02 | 2.585e+02 | 0.432 | 0.666744 | |
| Pv100t124E | 9.527e+01 | 1.969e+02 | 0.484 | 0.629330 | |
| Pv125t149E | -3.846e+01 | 2.663e+02 | -0.144 | 0.885372 | |
| Pv150t174E | 3.572e+02 | 2.657e+02 | 1.344 | 0.181155 | |
| Pv175t184E | 1.876e+02 | 4.907e+02 | 0.382 | 0.702749 | |
| Pv185t199E | 4.216e+02 | 4.163e+02 | 1.013 | 0.313000 | |
| Pv200t299E | 3.214e+02 | 1.831e+02 | 1.755 | 0.081495 | . |
| Pv300t399E | 1.132e+02 | 1.730e+02 | 0.654 | 0.514121 | |
| Pv400t499E | -1.048e+02 | 2.489e+02 | -0.421 | 0.674370 | |
| Pop_1E | 2.099e-04 | 1.013e-03 | 0.207 | 0.836188 | |
| station_count | -2.987e-01 | 1.232e-01 | -2.424 | 0.016667 | * |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.57 on 135 degrees of freedom
Multiple R-squared: 0.3733, Adjusted R-squared: 0.2061
F-statistic: 2.233 on 36 and 135 DF, p-value: 0.0005061

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 5.202e+02 | 9.962e+02 | 0.522 | 0.60415 | |
| n_rides_start | -9.670e-04 | 5.631e-04 | -1.717 | 0.09297 | . |
| avg_ride_duration_secs_start | -1.168e-02 | 5.780e-03 | -2.021 | 0.04934 | * |
| n_rides_end | 9.800e-04 | 5.557e-04 | 1.763 | 0.08477 | . |
| avg_ride_duration_secs_end | 3.914e-02 | 8.431e-03 | 4.642 | 3.11e-05 | *** |
| multifamily_housing_gross | -3.752e-06 | 1.628e-06 | -2.304 | 0.02601 | * |
| largest_property_use_type_residential | 3.450e+01 | 3.601e+01 | 0.958 | 0.34327 | |
| occupancy_residential | -6.090e-01 | 1.289e+00 | -0.472 | 0.63894 | |
| occupancy_non_residential | -3.163e-01 | 3.548e-01 | -0.891 | 0.37757 | |
| year_built_residential | -7.342e-01 | 4.486e-01 | -1.637 | 0.10881 | |
| year_built_non_residential | 5.275e-01 | 4.007e-01 | 1.316 | 0.19484 | |
| OOcHU1E | 1.408e-02 | 4.899e-03 | 2.874 | 0.00622 | ** |
| ROcHU1E | 9.130e-03 | 4.136e-03 | 2.208 | 0.03254 | * |
| MdVlE | -3.926e-05 | 5.110e-05 | -0.768 | 0.44637 | |
| EA_LTHSGrE | -2.262e+01 | 2.716e+02 | -0.083 | 0.93400 | |
| EA_BchDHE | 2.367e+02 | 4.191e+02 | 0.565 | 0.57511 | |
| EA_HScGrdE | 4.577e+01 | 4.505e+02 | 0.102 | 0.91954 | |
| MgBSciArtE | 2.517e+02 | 5.200e+02 | 0.484 | 0.63068 | |
| SrvcE | -5.728e+02 | 4.274e+02 | -1.340 | 0.18712 | |
| SalesOffE | -4.100e-03 | 8.489e-03 | -0.483 | 0.63153 | |
| NRCnstMntE | -7.646e-03 | 1.717e-02 | -0.445 | 0.65829 | |
| PrdTrnsMME | 1.747e-02 | 1.617e-02 | 1.080 | 0.28604 | |
| MdHHIncE | 2.112e-03 | 2.084e-03 | 1.013 | 0.31637 | |
| MnHHIncE | 5.730e-04 | 1.057e-03 | 0.542 | 0.59054 | |
| PvU50E | -4.656e+02 | 3.930e+02 | -1.185 | 0.24241 | |
| Pv50t74E | 3.323e+02 | 5.379e+02 | 0.618 | 0.53984 | |
| Pv75t99E | 8.540e+02 | 6.648e+02 | 1.285 | 0.20564 | |
| Pv100t124E | -1.754e+02 | 4.127e+02 | -0.425 | 0.67281 | |
| Pv125t149E | -6.163e+01 | 6.519e+02 | -0.095 | 0.92511 | |
| Pv150t174E | -1.011e+03 | 7.674e+02 | -1.318 | 0.19441 | |
| Pv175t184E | 2.608e+02 | 9.893e+02 | 0.264 | 0.79326 | |
| Pv185t199E | -8.173e+02 | 1.054e+03 | -0.775 | 0.44240 | |
| Pv200t299E | 9.484e+01 | 4.629e+02 | 0.205 | 0.83861 | |
| Pv300t399E | -1.278e+02 | 5.155e+02 | -0.248 | 0.80534 | |
| Pv400t499E | -2.236e+02 | 4.773e+02 | -0.468 | 0.64182 | |
| Pv500plE | -8.978e+02 | 4.013e+02 | -2.237 | 0.03038 | * |
| Pop_1E | -3.153e-03 | 1.849e-03 | -1.706 | 0.09512 | . |
| station_count | 3.428e-02 | 1.827e-01 | 0.188 | 0.85205 | |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.32 on 44 degrees of freedom
Multiple R-squared: 0.6437, Adjusted R-squared: 0.344
F-statistic: 2.148 on 37 and 44 DF, p-value: 0.007856

From the above model on the left, we can see significance at the 0.05 level for the number of rides that start as well as end in the NTA, the average ride length of rides that start as well as end in the NTA, multifamily housing units, owner-occupied units, industry of occupation is in management, business, science, and arts, and the number of stations. From the above model on the right, the subset of NTAs that have Citi Bike data, we can see significance at the 0.05 level for the average ride length of rides that start as well as end in the NTA, multifamily housing units, owner-occupied and renter-occupied units, and income level that is more than 500% above the poverty line.

**Research Questions**

For our research questions, we deconstructed the full model based on categories of interest. Thus, we created reduced models to test hypotheses regarding Citi Bike variables, poverty level variables, buildings variables, and education level variables.

Citi Bike Reduced Model

For this reduced model, we remove five variables from the full model: the start and end ride counts, the start and end ride durations, and the station count within the NTA. Thus, our null hypothesis is that the slopes for each of these variables equals zero and our alternative hypothesis is that at least one slope does not equal zero. The anova output is below.

```
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     49 139142
2     44  82554  5     56588 6.0321 0.0002469 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

With a p-value < 0.05, we reject the null hypothesis, thus noting there is a relationship between our selected Citi Bike ride variables and energy use intensity of buildings in respective NTAs.

Poverty Level Reduced Model

For this reduced model, we remove 12 variables from the full model: median and mean income level as well as the 10 variables for ratio of income to poverty level. Thus, our null hypothesis is that the slopes for each of these variables equals 0 and our alternative hypothesis is that at least one slope does not equal 0. The anova output is below.

```
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     51 3098.2
2     43 2168.0  8    930.18 2.3062 0.03736 *
```

With a p-value < 0.05, we reject the null hypothesis, thus noting there is a relationship between our selected poverty variables and energy use intensity of buildings in respective NTAs.

Buildings Reduced Model

For this reduced model, we remove nine variables from the full model: multifamily housing, largest property use type is residential, occupancy of residential properties, occupancy of non-residential properties, the year built of residential and non-residential properties, owner- and renter-occupied units, and median home value. Thus, our null hypothesis is that the slopes for each of these variables equals 0 and our alternative hypothesis is that at least one slope does not equal 0. The anova output is below.

```
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     52 3989.9
2     43 2168.0  9    1821.9 4.015 0.0008788 ***
```

With a p-value < 0.05, we reject the null hypothesis, thus noting there is a relationship between our selected buildings variables and energy use intensity of buildings in respective NTAs.

Education Level Reduced Model

For this reduced model, we remove three variables from the full model: education level less than high school, education level of high school, and education level of bachelor's degree. Thus, our null hypothesis is that the slopes for each of these variables equals 0 and our alternative hypothesis is that at least one slope does not equal 0. The anova output is below.

```
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     47 83173
2     44 82554  3    619.51 0.1101 0.9538
```

With a p-value > 0.05, we fail to reject the null hypothesis, thus we are unable to state there is a relationship between our selected education level variables and energy use intensity of buildings in respective NTAs.

## Conclusion and Next Steps

This research project was a fruitful opportunity for us to apply the theoretical knowledge of class to data interest. With more time and access to more computational power, we would take this research down further paths and ask more specific questions of our data. In particular, we would be interested in calculating more advanced distance measurements between stations and buildings, thus preventing us from needing to aggregate by NTA. Our initial plan was to calculate ride statistics between each NTA (i.e. average length of ride between Financial District and Morningside Heights), but the number of requests needed to the Google Maps API would have exceeded the allotted amount of free requests. In an ideal scenario, we would prefer to work with as granular data as possible.

This project has the potential to spark further quantitative research in regards to urban planning, architecture, and environmental policy. By highlighting the relationship between building energy ratings and increased utilization of Citi Bikes, our research could stimulate thoughtful discussions and shape policies in two crucial areas: transportation planning and environmental impact. The findings may encourage urban planners to prioritize bike infrastructure and sustainable transportation options, aligning with the goal of reducing car dependency. Additionally, by demonstrating the environmental benefits of promoting bike usage, such as reduced greenhouse gas emissions, this project can influence policies aimed at fostering a greener and more sustainable urban environment. Ultimately, it underscores the interconnectedness of urban sustainability and transportation choices, providing valuable insights for policymakers and city planners to create more eco-friendly and efficient cities.

## Appendix

<u>Model Selection</u>

Our chosen response variable: Site EUI Kbtu/ft

We compared a range of potential response variables manually by graphing them, and discovered that many of them had long right tails and were bounded by zero. Of all of the potential response variables, this one had just a few clear outliers.

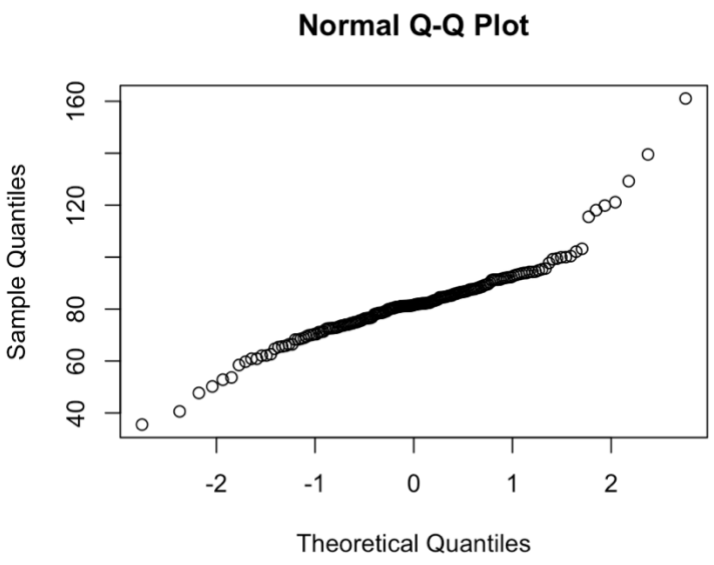
We chose not to do any transformation to our Y variable because we did not want to lose interpretation of the regression coefficients. Perhaps we could have done a log transformation, and not lost all interpretability and gotten better results.

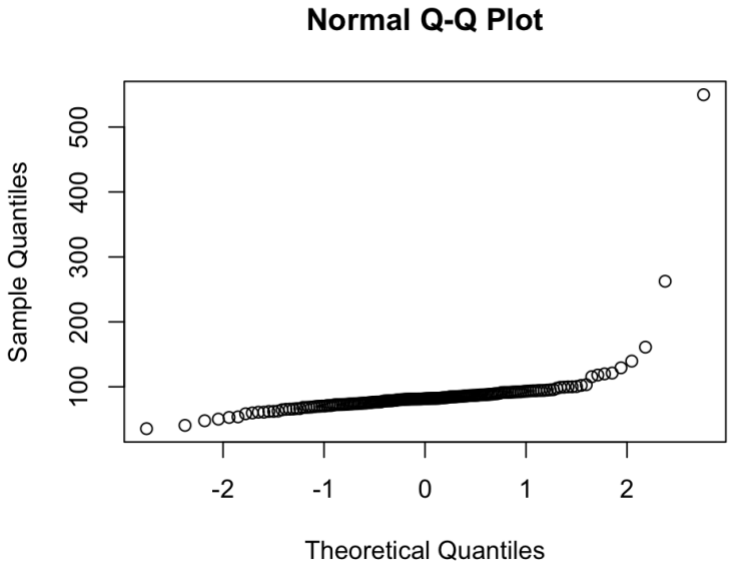
The input variables for our model were discussed earlier in the paper.

Diagnostics and Model Validation

We constructed plots in order to assess any potential violation of regression assumptions. The plots below help illustrate any potential violation of normality, linearity, constant variance, and independence.

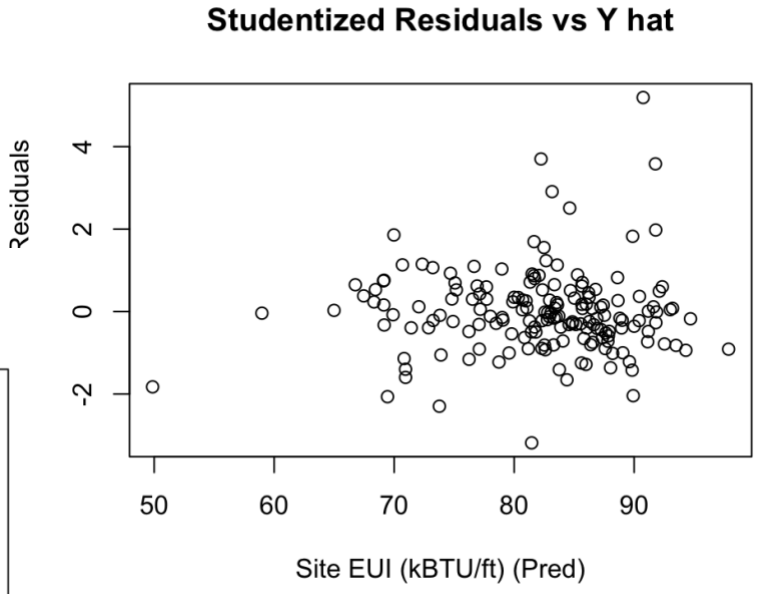
Provided below are two QQ plots. For diagnostic purposes and to understand the data in full, we reconstructed the QQ plot on the right by removing a few outlying data points. For example, we removed the NTA for the West Village, which had dramatically larger values for our energy usage response variable. In both cases, it looks clearly heavy tailed, but less so in the plot with the removed outliers.

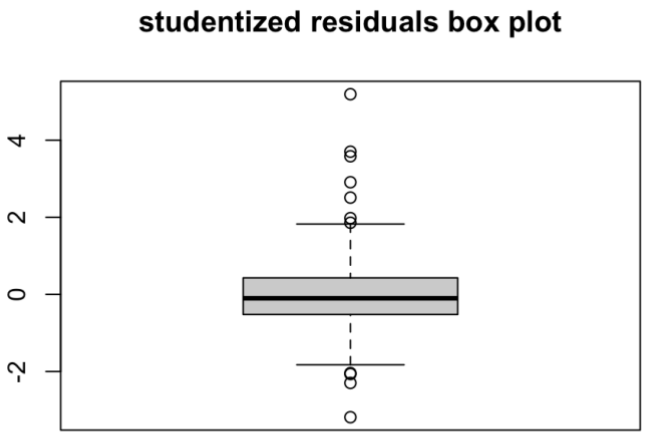


We also include a plot of studentized residuals against our predictor variable, and a box plot (outliers removed).

We plot the residuals against our y hat variable and look for randomness.




The box plot once again displays some fairly significant outliers on the tails of the distribution. Ideally, we would have preferred to see only one or two outliers on either side of the box plot.