Quantifying the Effect of Quality Public Space on Citi Bike Dock Usage

Aaron D'Souza, Daniel Fay, Kristi Korsberg, Jonathan Pichot, Ziman Zhou

Introduction and Hypothesis

In the mid-twentieth century, Jane Jacobs, the famed urban activist, journalist, and author, understood that "'we must look at how people use cities to understand how to shape them.'"[1] Jan Gehl, an architectural pioneer in the same field, has spent the past 40 years pursuing a similar goal; 'Making Cities for People' is the trademark phrase on the website for Gehl People, his international architectural firm.

Concurrent to research and activism about public space was the rise of public bike shares. The first appeared in Amsterdam in the 1960s, and since, more technologically advanced versions have flourished in cities across the world, New York City included.[2] As demand for such transportation grew, so too did the volume and spread of associated infrastructural changes. As of October, 2016, New York City's Citi Bike boasted 590 active stations in Manhattan, Queens, Brooklyn, and Jersey City.[3] Public bike shares are one of the most obvious manifestations of the ways in which cities are scaling for better human use.

It is the aim of this study to quantify the value of public space by evaluating how ridership volumes at Citi Bike stations change based on attributes associated with the built environment. It is the expectation that docking stations within high quality public space will see higher ridership volumes than those in areas with poor quality built environments.

Literature Review

Much research has been conducted on the factors that impact bike share ridership, such as weather, time, and population demographics. While all are significant in explaining ridership volumes, of particular relevance for this study are papers which evaluate the impacts of the built environment on public bike share ridership. "The built environment includes all of the physical parts of where we live and work."[4]

---

[1] Jared Green, "Jan Gehl: The City Is Big," The Dirt, February 10, 2014, accessed December 01, 2016, https://dirt.asla.org/2014/02/10/jan-gehl-the-city-is-big/.
[2] Wafic El-Assi, Mohamed Salah Mahmoud, and Khandker Nurul Habib, "Effects of Built Environment and Weather on Bike Sharing Demand: A Station Level Analysis of Commercial Bike Sharing in Toronto," Transportation, December 16, 2015, Page 2, doi:10.1007/s11116-015-9669-z.
[3] Inc. Motivate International, "Citi Bike: Unlock a Bike, Unlock New York | Citi Bike NYC," Citi Bike NYC, 2013, accessed December 01, 2016, https://member.citibikenyc.com/.
[4] "Impact of the Built Environment on Health," June 2011.

Among others, researchers in Lyon, France published a thorough analysis of the built environment's effect on public bike share ridership. Lyon is a city in east-central France and home to about 500,000 people with a bike share program called Vélo'v. Researchers surmised that bike station usage would be dependent upon the quality of the built environment around the station. They tested this hypothesis by quantifying built environment attributes within a 300m buffer around each bike station and evaluating the significance and strength of the variables with a linear regression model.[5] They considered the proximity to student housing or a university campus, the presence of other public transportation options, the capacity of each bike station, bicycle infrastructure, and points of interest to be variables that capture the built environment.[6] The study found that proximity to students, bike station capacity, and railway stations explained public bike share ridership flows.[7]

This study uses many of the same built environment features outlined in the above-mentioned analysis to capture the quality of public space. It is unique because it evaluates the built environment in New York City, where the impact of public space on bike share ridership has yet to be evaluated. It is also unique in the assumption that Citi Bike ridership volumes can be used to quantify the value of public space.

Methodology & Data Collection

The basic project methodology consisted of analyzing features which reflect the quality of public space within a given radius of each Citi Bike station. We also collected demographic data, which was included in order to build the most accurate model possible. Data collection efforts consisted of finding data sources, creating public space scores for each variable, and associating those scores to Citi Bike stations.

We downloaded the list of stations from Citi Bike's official website. The corresponding zip code for each start station was identified with the Google Maps Geocoding API.

To quantify the built environment, we used five variables to define the quality of public space: the quality of the street, the presence of nearby parks, the quality and kind of bike lanes, the number and quality of nearby trees, and the amount of traffic volume. These explanatory variables were augmented by other features we envisioned might have an effect on ridership,

[5] Tien Dung Tran, Nicolas Ovtracht, and Bruno Faivre D'Arcier, "Modeling Bike Sharing System Using Built Environment Factors," Procedia CIRP 30 (2015): Page 293, doi:10.1016/j.procir.2015.02.156.
[6] Tien Dung Tran, Nicolas Ovtracht, and Bruno Faivre D'Arcier, "Modeling Bike Sharing System Using Built Environment Factors," Page 295.
[7] IBID.

including nearby subway entrances, median household income, and population density. A list of all data sources can be found in Table 1 of Appendix A.

The variable all these features are meant to explain is ridership. We understand ridership in this project as a demonstrated preference. Ridership counts were collected for each month in 2015 by summing the number of trip start points. With this data, we were then able to test various scenarios depending on day of the week and time of the year.

We considered features within 185 feet of a station as having an effect on usage. Parks and subway entrances within that radius were counted as existing (1) or not existing (0). Street quality, traffic volume, and bike lane quality scores are the average of all scores within that radius. The tree canopy score was calculated as the sum total of all trees scores within the radius. Median household income was found by station zip code, and the population density value was determined by each station's respective census tract.

Regression Analysis

We chose to run several different statistical tests on our dataset including Ordinary Least Squares Regression, Lasso Regression, Ridge Regression, and K-Means Clustering. Before running the regressions, we used Quantile-Quantile plots to test the distribution of each regressor. The R^2 for all but two of the regressors indicated that they were normally distributed, with Parks and Subway Entrance not normally distributed. Nevertheless, we believed it was reasonable to include all the regressors in the analyses because together, they better quantify the effect of public space. We used the original explanatory variable values to run the OLS Multivariate, Ridge, and Lasso regressions so that we could visualize the actual coefficients corresponding to each regressor. September ridership was chosen for this analysis based on the assumption that it represents the most normal riding conditions, i.e., relatively mild weather conditions and typical commuting volumes. Significant variables are defined as those which have a p-value of less than .05.

Bi-Variate Ordinary Least Squares Regression

The results of the OLS Bivariate regression can be found in Table 2. The bi-variate linear regression was run on each of the eight regressors using the formula, $y = coefficient_i \times + intercept_i$. The coefficients for bike lane score, subway entrance, median household income, and population density are all less than .05, indicating they are significant. The R-squared values are very small for each though, meaning that no singular public space

attribute explains much of the variance in Citi Bike ridership. OLS regression plots for each variable demonstrate this pattern as well. Figure 2A, found in Appendix B, is a series plot of each standardized regressor with the fitted OLS line. Trends in the ridership data are difficult to identify, and the fitted regression lines hardly capture the spread of data points associated with each model.

Table 2: Coefficients, Intercepts, P-Values and R^2

	Coefficient	Coefficient P-Value	Intercept	Intercept P-Value	R-Squared
Bike Lane	251.3414	.0014	2299.7840	$5.25*10^{-27}$	.0034
Park	-41.0444	.8418	2771.7704	$5.01*10^{-49}$	.0001
Street Quality	-53.2608	.7699	3151.0147	.0208	.0003
Subway Entrance	1474.0188	.0004	2588.9812	$7.14*10^{-52}$	.0410
Trees	-5.6788	.3386	2921.8696	$2.91*10^{-32}$	.0031
Traffic Volume	.0105	.3594	2605.5088	$1.79*10^{-28}$	.0028
Median Household Income	0.0258	$1.05*10^{-11}$	554.7822	.0984	.1439
Population Density	1520493.6210	.0008	1905.2877	$8.74*10^{-11}$	.0371

Multi-variate Ordinary Least Squares Regression

The results of the OLS Multivariate linear regression can be found in Table 3. The regression was run on all of the eight regressors, and the result suggested the following model:

$$ridership = 239.510 \times bike\_lane\_score + 75.332 \times park + 15.600 \times street\_quality\_score + 1142.860 \times subway\_entrance - 9.956 \times tree\_score - 0.002 \times traffic\_volume + 0.022 \times median\_hh\_income + 1462000 \times pop\_density - 311.896$$

By evaluating the p-values, the feature selection conclusion is similar to that of the bi-variate regression: the significant coefficients are bike lane score, subway entrance, median household income, and population density. The R^2 and the adjusted R^2 values of the regression model are both very small, being 0.225 and 0.204 respectively. This suggests that the linear model of public space attributes can not explain the variance in Citi Bike ridership. If

cross-validation is applied here, the in-sample and out-of-sample R^2 values are 0.225 and 0.208 respectively, with a large residual sum of squares. The residuals of both in and out of sample tests are not evenly distributed along zero, indicating that the OLS multivariate linear model does not fit the data well. This might be due to undetected outliers or a weak correlation between the ridership and the public space factors. Figures 2B-1 and 2B-2 depict residuals and can be found in Appendix B.

Table 3: Coefficients, Intercepts, P-Values

	Coefficient	Standard Error	T	P-Value
Intercept	-311.8957	1304.8530	-.2390	.8110
Bike Lane	239.5098	72.8370	3.2880	.0010
Park	75.3324	187.3400	.4020	.6880
Street Quality	15.6001	165.7230	.0940	.9250
Subway Entrance	1142.8595	398.1520	2.8700	.0040
Trees	-9.9558	5.7610	-1.7280	.0850
Traffic Volume	-.0015	.0100	-.1470	.8830
Median Household Income	.0217	.0040	5.8420	.0000
Population Density	$1.462*10^6$	$4.29*10^5$	3.4060	.0010

Ridge Linear Regression

We performed the Ridge regression on the training sets and then tested the resulting model on the remaining data. We used an efficient Leave-One-Out cross-validation method (set 4-fold) to identify the optimal lambda value, which is 10. The result of the Ridge Regression suggested the following model:

$$ridership = 224.611 \times bike\_lane\_score + 88.091 \times park + 33.153 \times street\_quality\_score + 791.865 \times subway\_entrance - 5.779 \times tree\_score - 0.001 \times traffic\_volume + 0.024 \times median\_hh\_income + 3.341 \times pop\_density + 108.468$$

According to the in-sample and out-of-sample tests, the R^2 of the former is around 0.193 and that of the latter is around 0.144. Unexpectedly, the residual sum of squares is also very large. The residuals of both in and out of sample tests are not evenly distributed along zero,

indicating that the Ridge model does not fit the data as well as expected. This might be due to undetected outliers or a weak correlation between the ridership and the public space factors. Figures 2C-1 and 2C-2 depict residuals and can be found in Appendix B.

Lasso Linear Regression

We also performed the Lasso regression on the training sets and then tested the resulting model on the test sets. The best model was selected by 4-fold cross-validation. The result suggested the model:

$$ridership = 202.037 \times bike\_lane\_score + 1151.815 \times subway\_entrance + 0.020 \times median\_hh\_income + 1107879.01 \times pop\_density - 107.020$$

with the auto-selected optimal lambda of 2.75775729747.

The Lasso regression further confirmed patterns identified by OLS and Ridge analyses. It kept bike lane score, subway entrance, median household income and population density in the model, and eliminated all other regressors. The survived regressors in Lasso were those found to be significant in OLS. The in-sample R^2 of the test was about 0.216 and the out-of-sample was around 0.206. The residual sum of squares was very large. Similar to the Ridge regression, residuals of both in and out sample tests are not evenly distributed along zero. Therefore, our Lasso model did not fit the data as well as expected, which may be due to a weak correlation between the ridership and the public space. Figures 2D-1 and 2D-2 depict the residuals and can be found in Appendix B.

K-Means Cluster

We also conducted a K-Means test on the significant regressors and the average ridership for September, 2015. The results of a Silhouette Score indicated that the data should be divided into six clusters. While this method did not provide a perfect explanation for the clusters, specific patterns were observed within each one. First, a cluster of high average ridership with high median household income and a high bike lane score is observed along the Citi Bike Stations along 1st Avenue and 2nd Avenue in Manhattan. Second, a cluster of high average ridership with high median household income following a cluster of low average ridership with low median household income is observed along Broadway in Manhattan. Last, Brooklyn and Queens consist of clusters of low average ridership for average median household income which is independent of the street rating score. These three patterns are depicted in Figures 3B - 3D in Appendix C. Observing the above clusters, it can be inferred that median household income is a strong predictor of average ridership.

Conclusion

The results of the analysis did not show a strong signal between Citi Bike ridership and public space, however the quality of the bike lanes and the proximity to a subway station were found to be significant regressors in the model. Intuitively this makes sense, the higher the quality of bike lanes in the area the more bikes in the area. Additionally, proximity to a subway station may be significant because Citi Bike users complete a majority of their journey on the subway and only use Citi Bike for the first or last mile of their journey.

The results prove that our assumption of Citi Bike ridership being a proxy for effective public space was false. Nonetheless, the methodology used to develop the model given a different dependent variable, such as urban sensing data, could prove to be very useful for both public agencies and private companies. With more data collected, other methodologies such as polynomial regression and Box-cox transformation can be applied to improve the models by identifying the forms of the independent and dependent variables(polynomial or logarithmic). It is also necessary to perform spatial analyses to identify if there exists a spatial correlation between citi-bike ridership in one location and its neighbour locations. If so, we would obtain a more accurate model by running a spatial regression. If an effective model were developed, planners and companies could evaluate public space over time and space to identify successful implementations and facilities. Quantifying public space allows us to understand historical mistakes and successes so that moving forward we can build a smarter city.

Bibliography

El-Assi, Wafic, Mohamed Salah Mahmoud, and Khandker Nurul Habib. "Effects of Built Environment and Weather on Bike Sharing Demand: A Station Level Analysis of Commercial Bike Sharing in Toronto." Transportation, December 16, 2015, 1-25. doi:10.1007/s11116-015-9669-z.

Green, Jared. "Jan Gehl: The City Is Big." The Dirt. February 10, 2014. Accessed December 01, 2016. https://dirt.asla.org/2014/02/10/jan-gehl-the-city-is-big/.

"Impact of the Built Environment on Health." June 2011. Accessed December 1, 2016. https://www.cdc.gov/nceh/publications/factsheets/impactofthebuiltenvironmentonhealth.pdf.

International, Inc. Motivate. "Citi Bike: Unlock a Bike, Unlock New York | Citi Bike NYC." Citi Bike NYC. 2013. Accessed December 01, 2016. https://member.citibikenyc.com/.

Tran, Tien Dung, Nicolas Ovtracht, and Bruno Faivre D'Arcier. "Modeling Bike Sharing System Using Built Environment Factors." Procedia CIRP 30 (2015): 293-98. doi:10.1016/j.procir.2015.02.156.

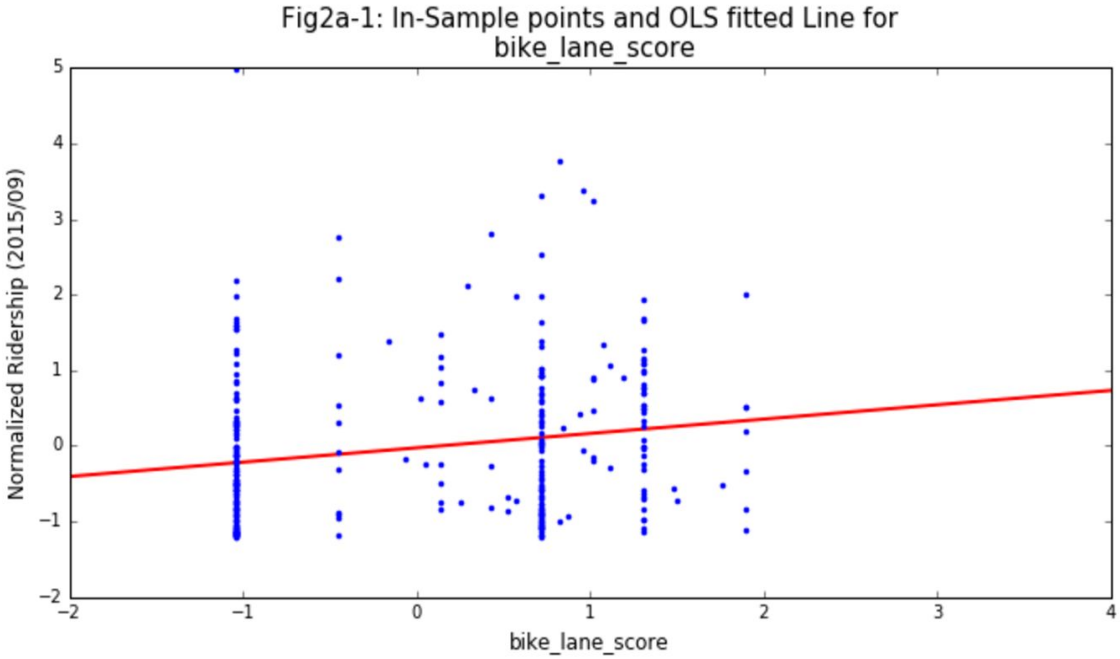
Appendix A

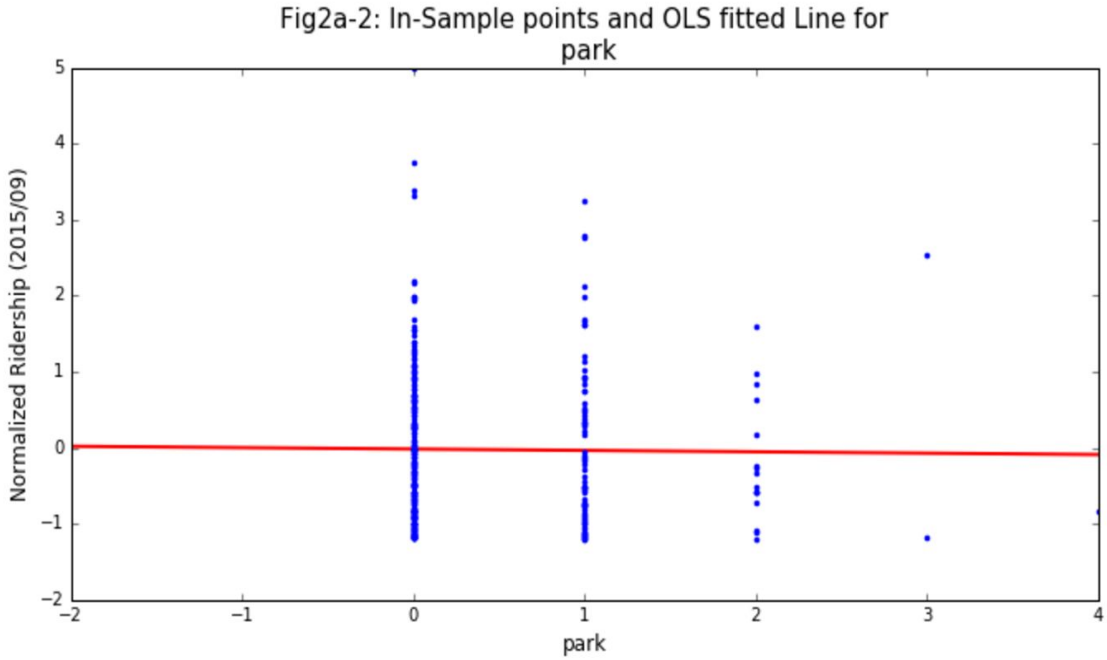
Table 1: Data sources

Variable Name	Data Source
Bike Lane	http://www.nyc.gov/html/dot/downloads/misc/nyc-bike-routes.zip
Median Household Income	http://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t 1. Filter 'Geographies' to '5 digit zip code tabulation area - -860' 2. Filter to NY State 3. Add 'all 5-digit zip code tabulation areas fully/partially within New York' 4. Filter 'Topics' to 'People' -> 'Income/Earnings (Households)' 5. Select table 'B19013' entitled 'MEDIAN HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2014 INFLATION-ADJUSTED DOLLARS)'
Park	https://data.cityofnewyork.us/api/geospatial/rjaj-zgq7?method=export&format=GeoJSON
Population Density	https://data.cityofnewyork.us/api/geospatial/fxpq-c8ku?method=export&format=GeoJSON
Street Quality	http://www.nyc.gov/html/dot/html/about/datafeeds.shtml 1. Select 'Download the Street Assessment Rating Shapefile (zip)'
Subway Entrance	https://data.cityofnewyork.us/api/geospatial/drex-xx56?method=export&format=GeoJSON
Traffic Volume	https://www.dot.ny.gov/tdv 1. Select 'TDV_Shapefile_AADT_2015.zip'
Trees	https://data.cityofnewyork.us/Environment/NYC-Urban-Tree-Canopy-Assessment-Metrics-2010/hnxz-kkn5

Appendix B

Figure 2A-1 - 2A-8: plot series of OLS bi-variate regression results using standardized values except for Park and Subway Entrance.





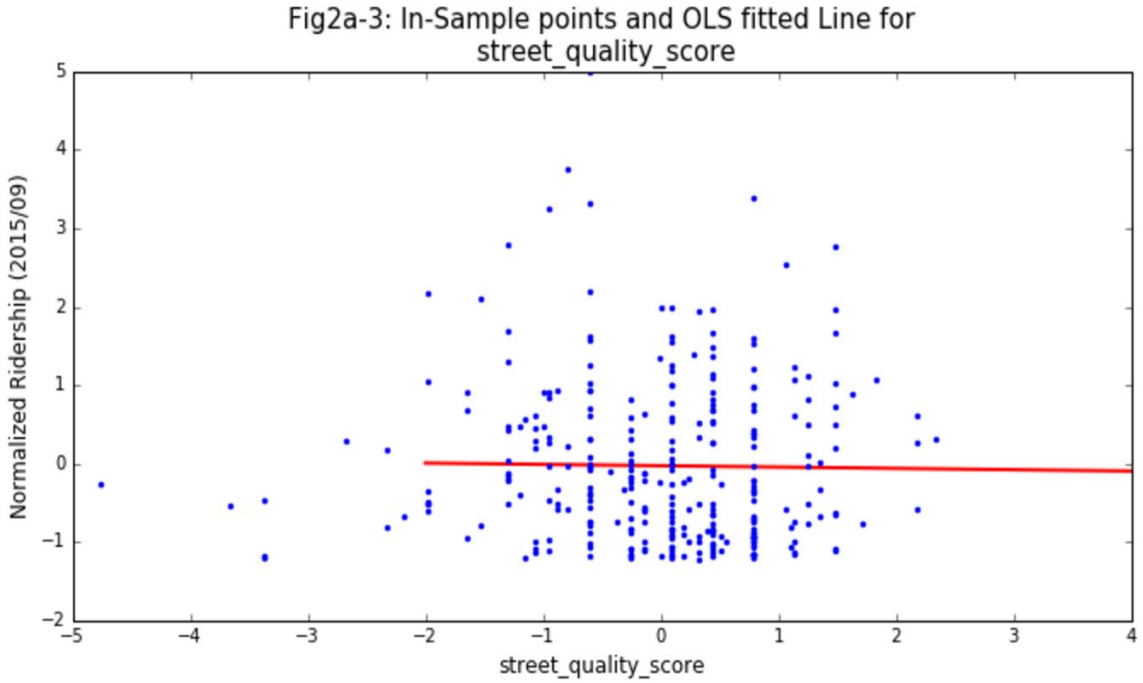


Fig2a-3: In-Sample points and OLS fitted Line for
street_quality_score
Normalized Ridership (2015/09)
street_quality_score
5
4
3
2
1
0
−1
−2
−5
−4
−3
−2
−1
0
1
2
3
4

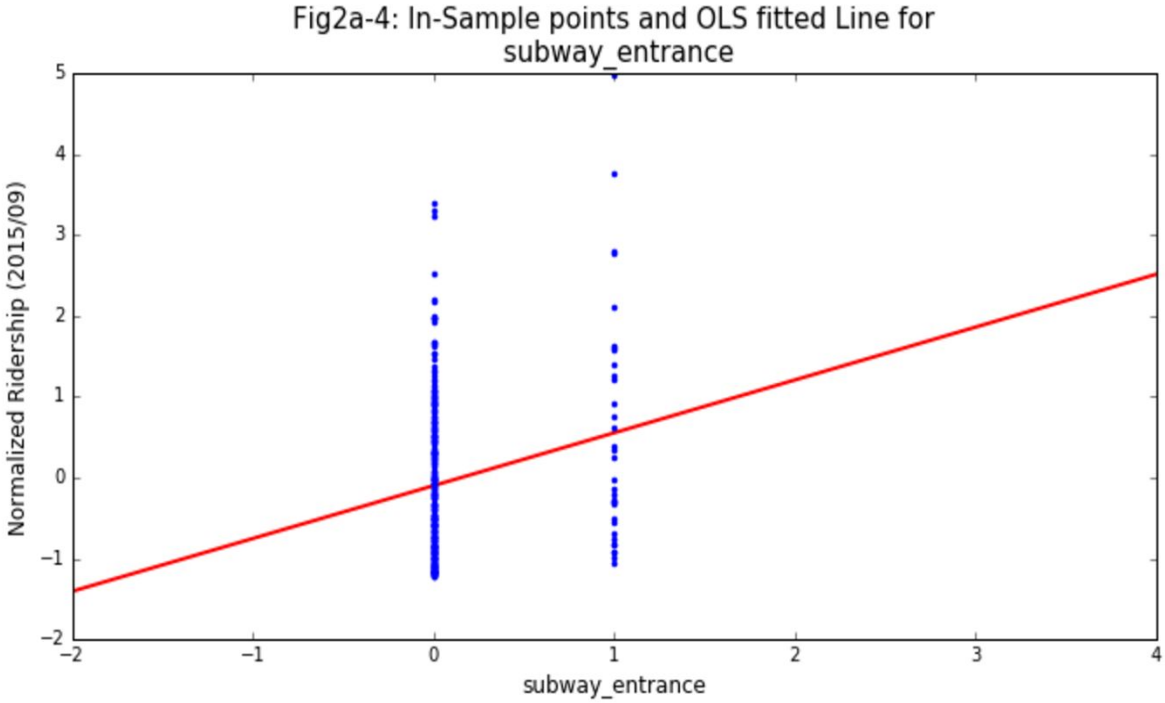


Fig2a-4: In-Sample points and OLS fitted Line for
subway_entrance
Normalized Ridership (2015/09)
subway_entrance
5
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4

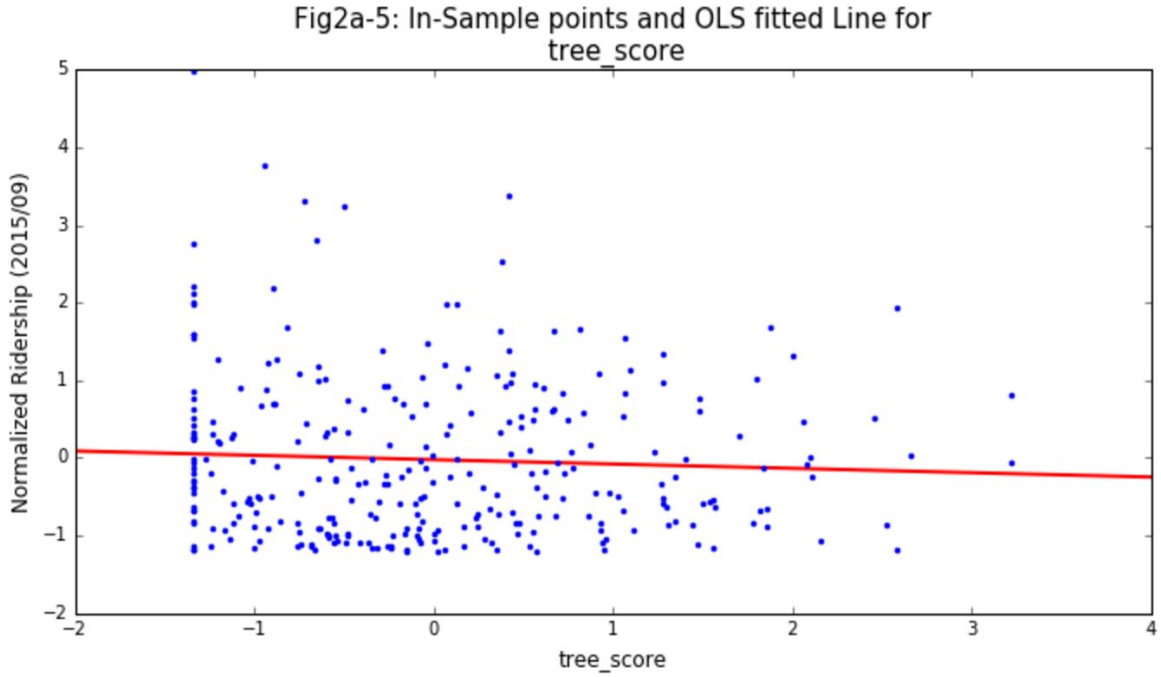


Fig2a-5: In-Sample points and OLS fitted Line for
tree_score
Normalized Ridership (2015/09)
tree_score
5
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4

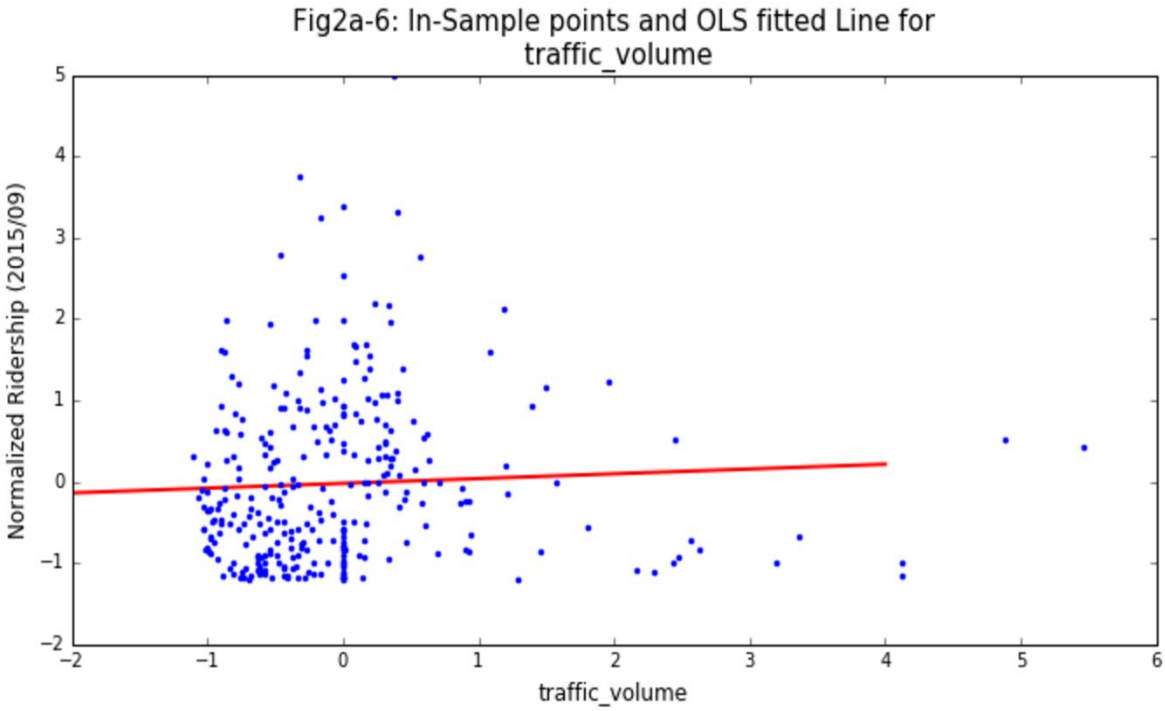


Fig2a-6: In-Sample points and OLS fitted Line for
traffic_volume
Normalized Ridership (2015/09)
traffic_volume
5
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4
5
6

Fig2a-7: In-Sample points and OLS fitted Line for median_hh_income

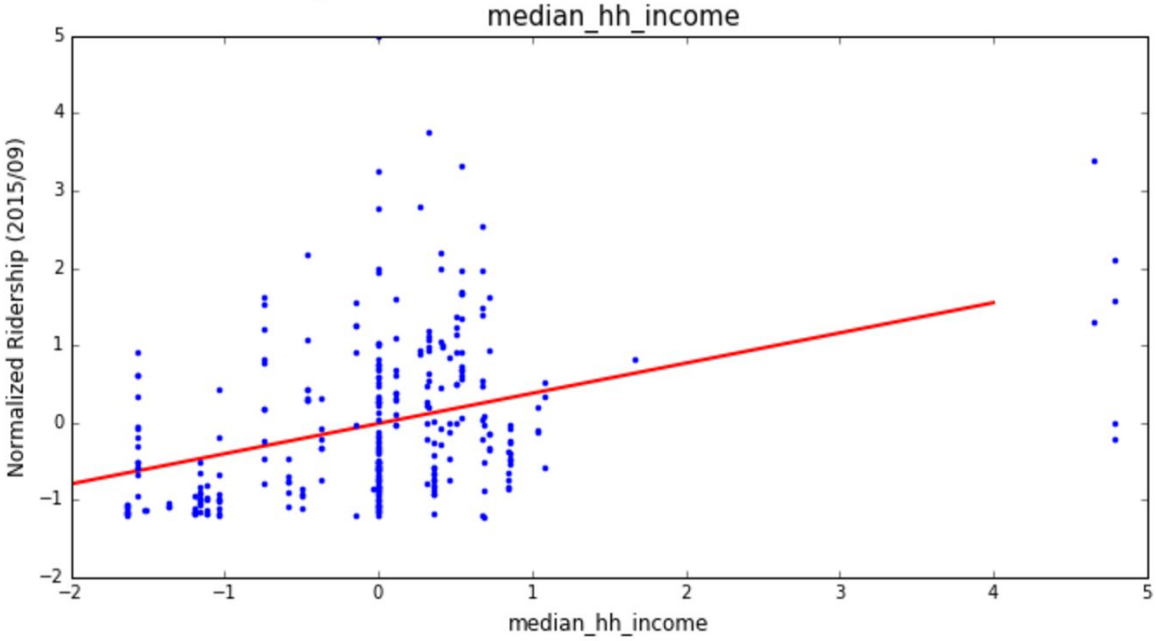
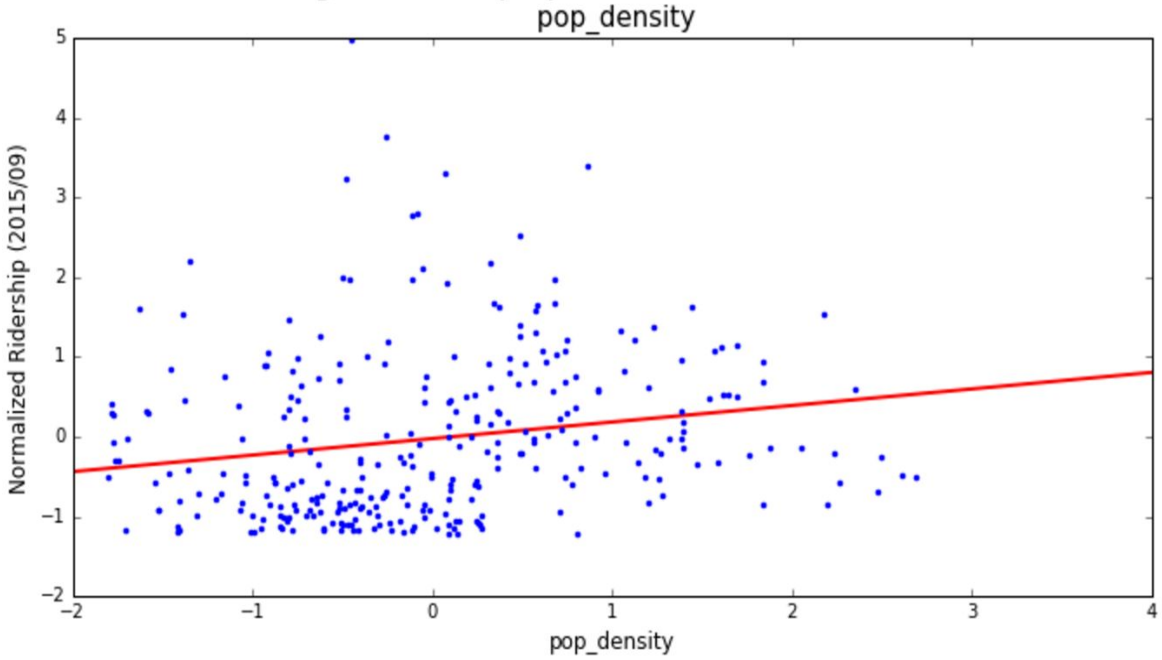
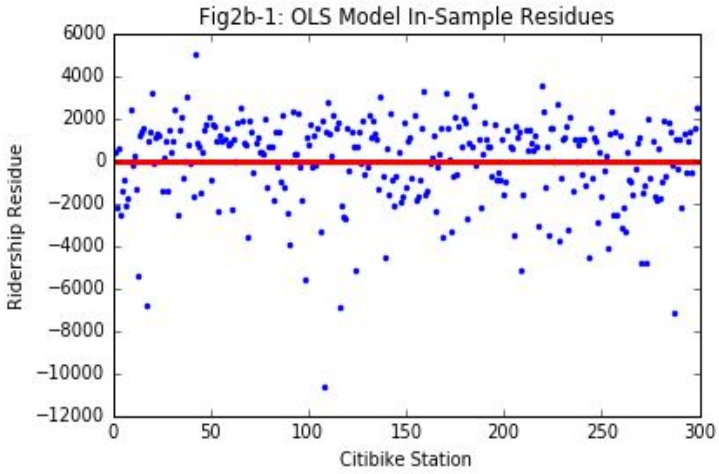
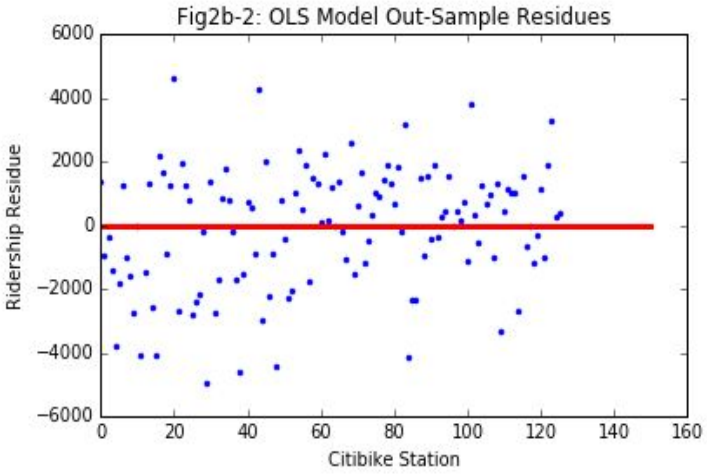


Fig2a-8: In-Sample points and OLS fitted Line for pop_density

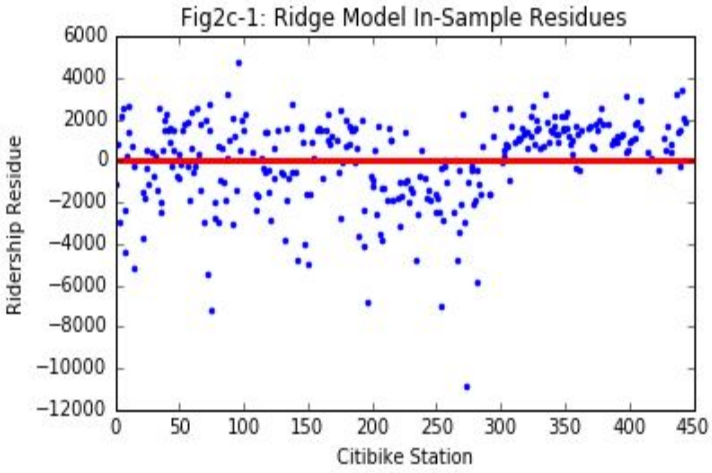


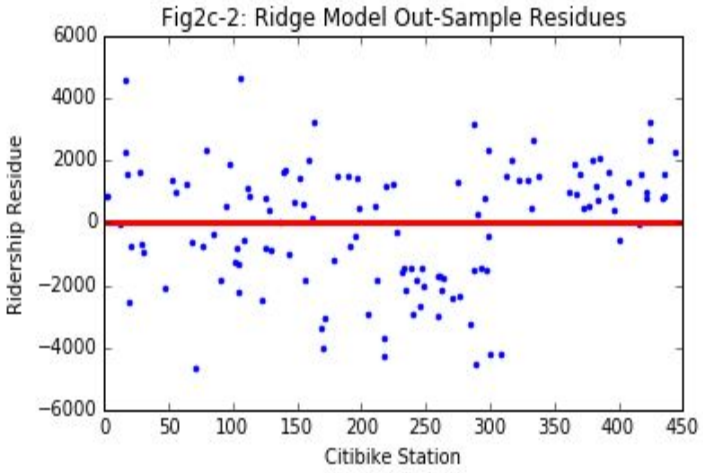
Figures 2B-1 and 2B-2: plot of sum of squared residuals for OLS multi-variate regression model.



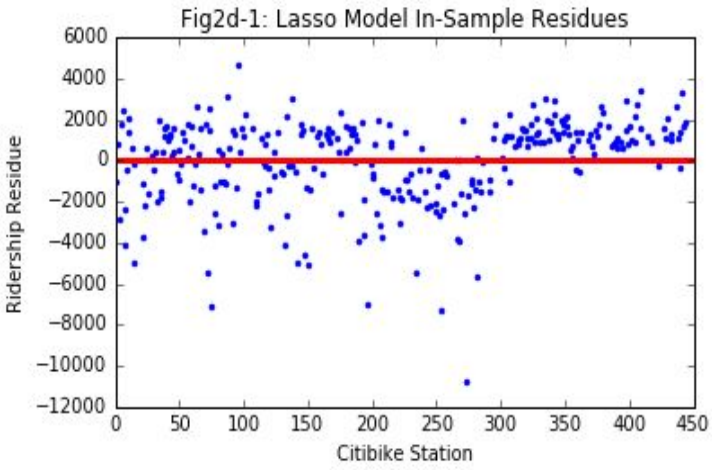


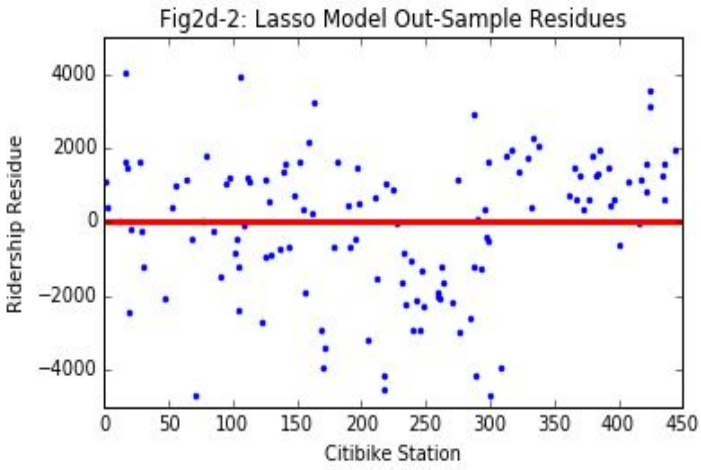
Figures 2C-1 and 2C-2: plot of sum of squared residuals for Ridge regression model.





Figures 2D-1 and 2D-2: plot of sum of squared residuals for Lasso regression model.





Appendix C

Figure 3A: Map of K-Means clusters of Citi Bike stations in New York City.

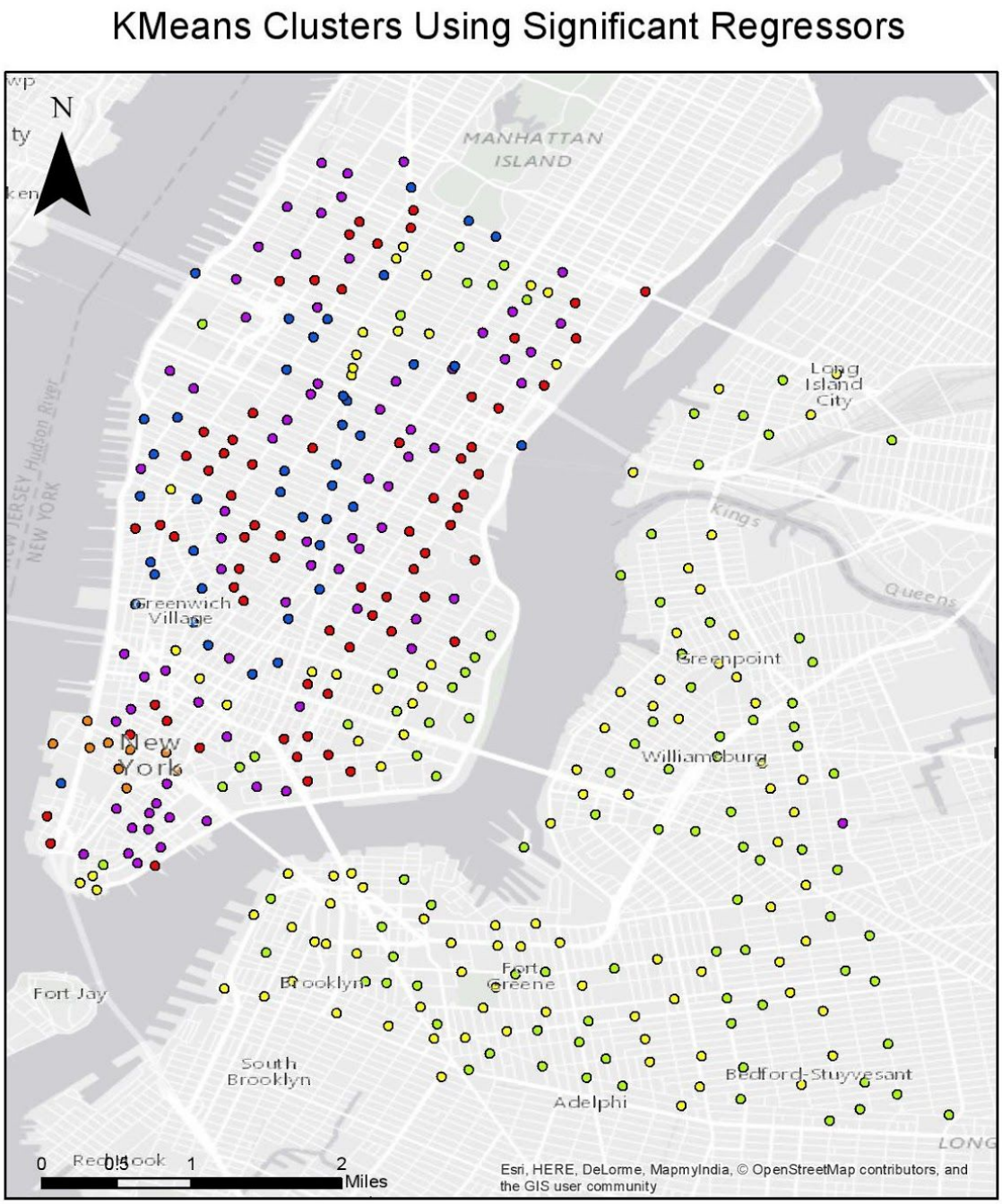


Figure 3B: Map representing clusters of Citi Bike Stations along 1st Avenue and 2nd Avenue in Manhattan.

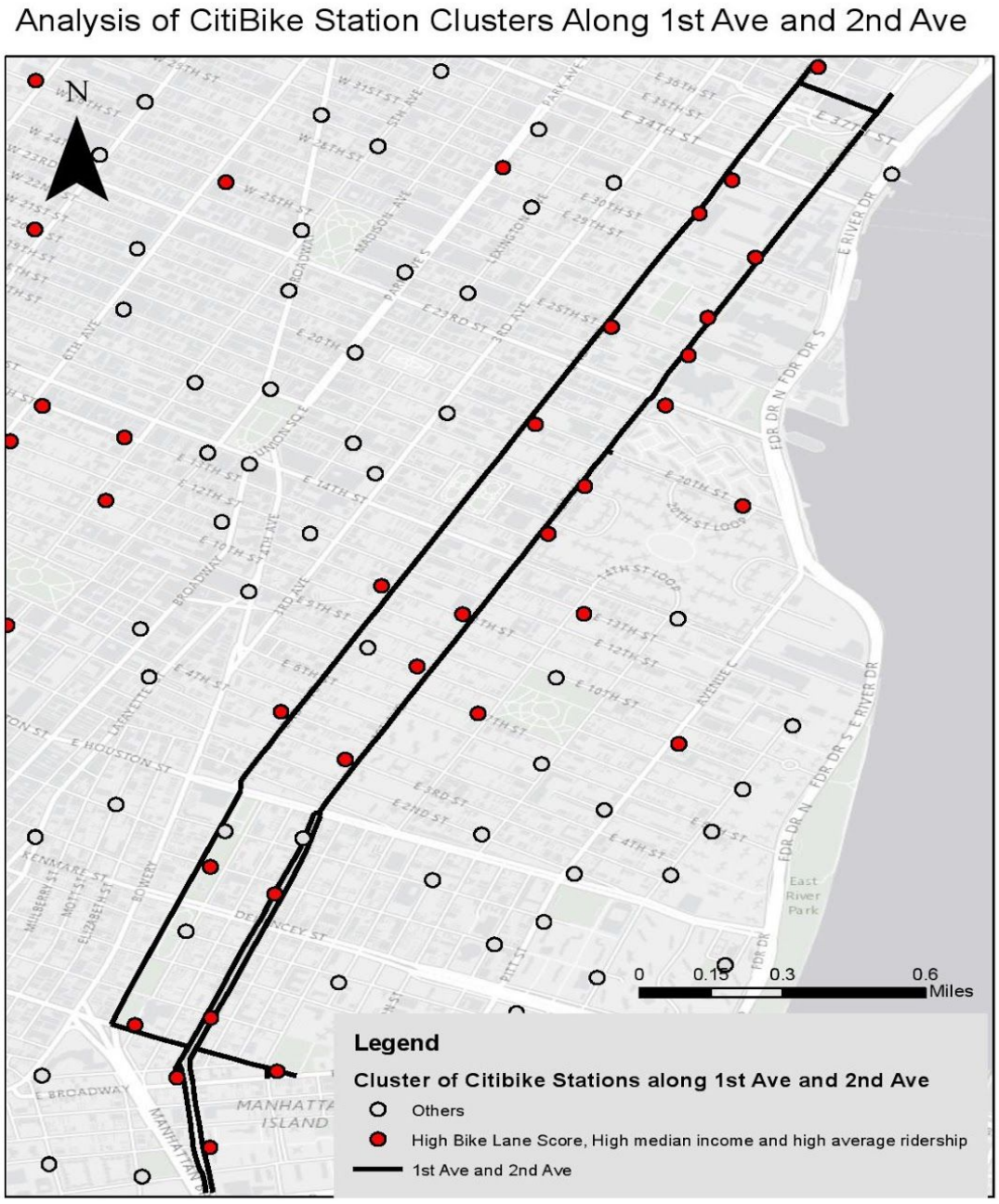


Figure 3C: Map representing clusters of Citi Bike Stations along Broadway in Manhattan.

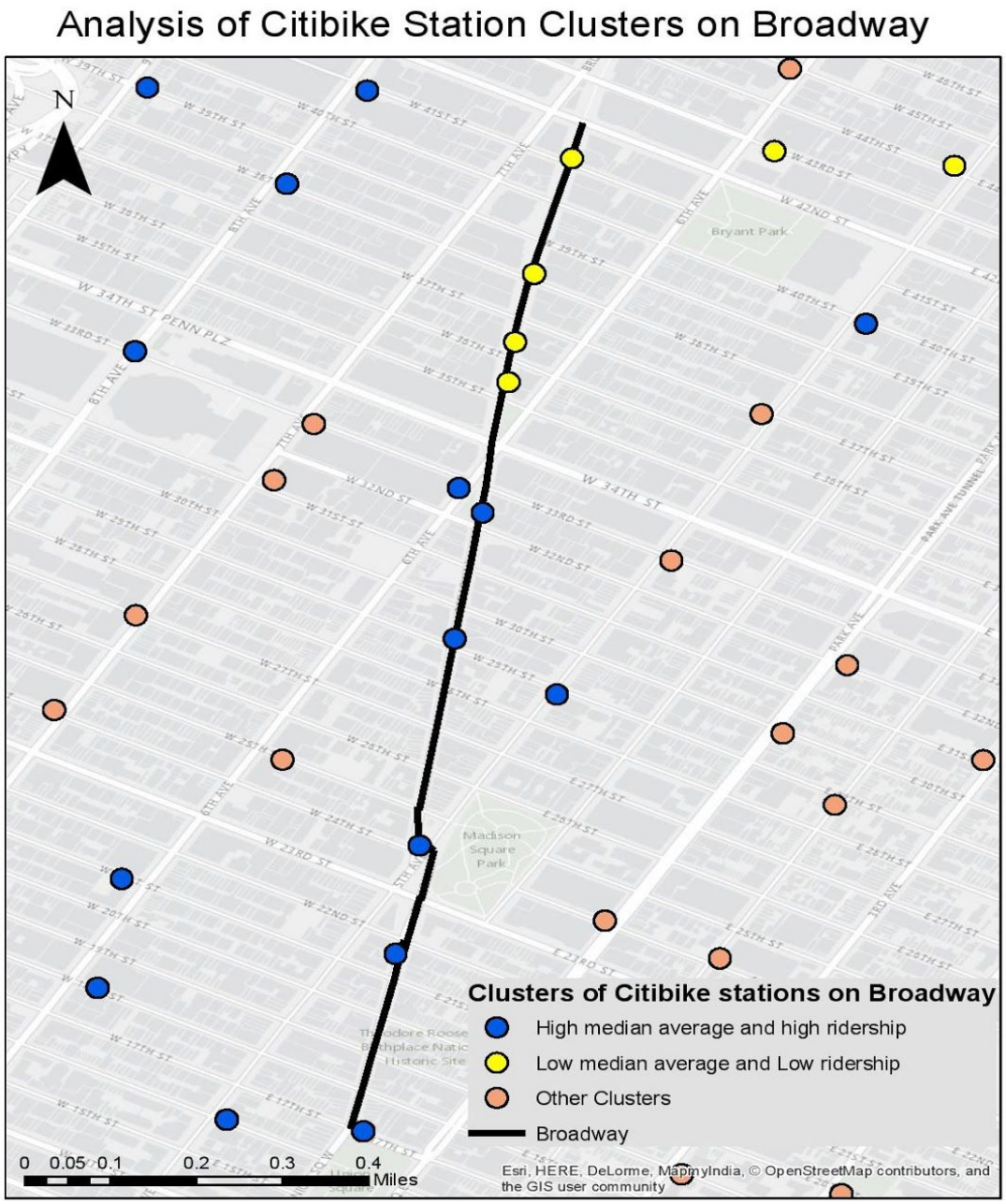
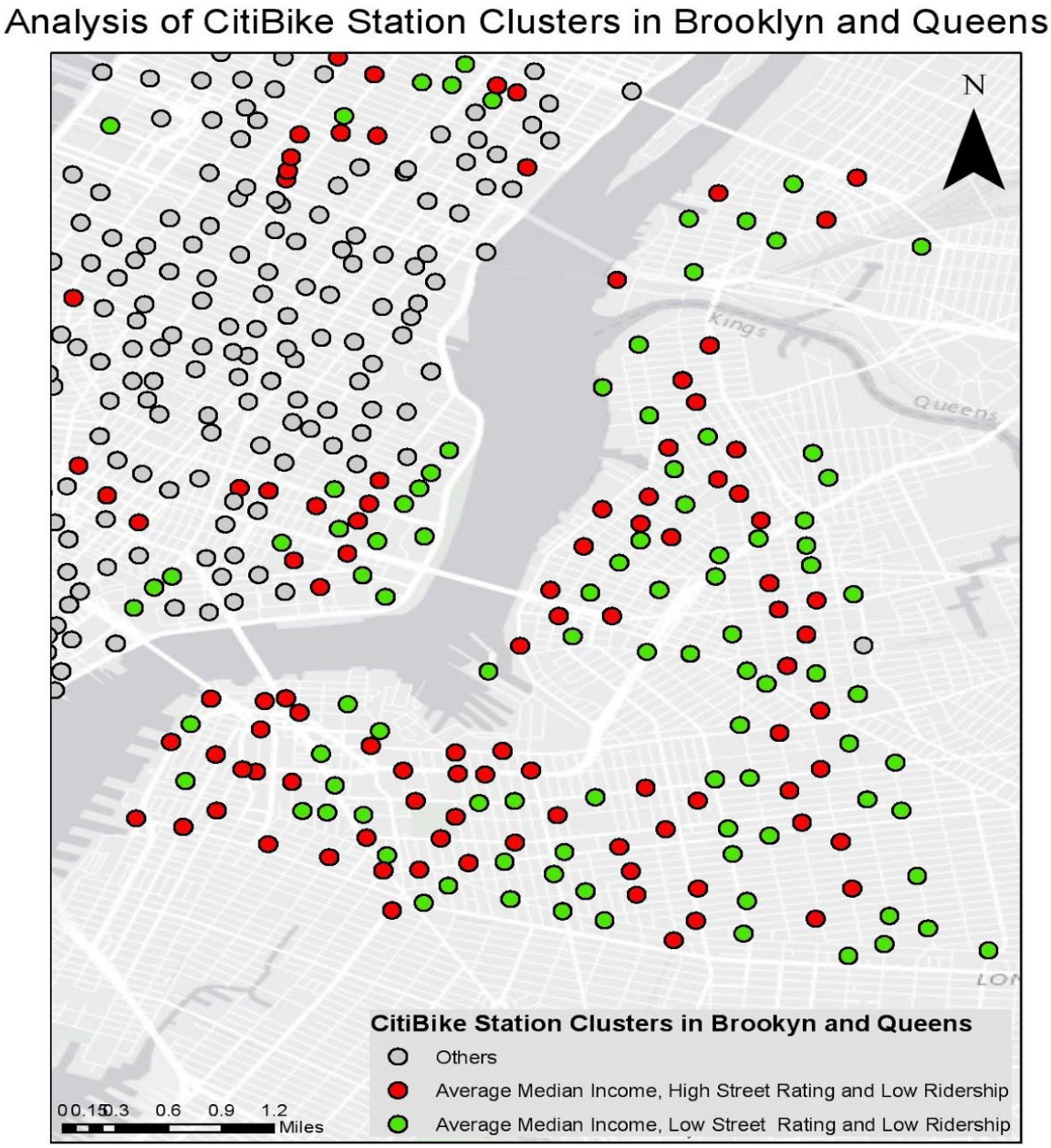


Figure 3D: Map representing clusters of Citi Bike Stations in Brooklyn and Queens.



Appendix D

Table 4: Team member contributions

Data Collection & Processing

Street Assessments	Daniel
Parks	Aaron
Subway Entrances	Aaron
Bike Lanes	Kristi
Tree Canopy	Jonathan
Traffic Volume	Daniel
Ridership	Kristi
Income	Kristi
Population Density	Aaron
Merge Features	Daniel
Standardize Data	Daniel
Google Geocode	Kristi

Analysis

OLS, Ridge, and Lasso regression	Kay(Ziman)
OLS regression comparing months	Daniel
OLS regression same month	Kristi

Paper Writing

Introduction	Kristi
Literature Review	Kristi
Methodology & Data Collection	Jonathan
OLS	Daniel, Kristi, Kay(Ziman)
Ridge and Lasso	Kay(Ziman)
Conclusion	Daniel

Visualizations	Aaron, Kay(Ziman)

Project Details

Project pattern & Github maintenance	Jonathan
Data import scripting	Jonathan, Kay(Ziman)
Website & Carto Map	Jonathan

Appendix E

Project website and code can be found at http://pichot.github.io/citibike-publicspace/
Interactive map of data sources exists here
https://pichot.github.io/citibike-publicspace/map.html

Figure 4: Screenshot of Carto map of Citi Bike Docking stations. Source.