CITADEL | CITADEL | Securities

TheDataOpen

Team 15 Where to Build the Next Bike Share Station?

A Recommendation System Based On Demographics And Transportation

Raymond Ji **Yili Wang** **Ying Jin** **Weipeng Shao**

January 25, 2020

1 Topic question

Over the last few years, public bicycle sharing systems have expanded across major US metropolitan cities, including New York City, Boston and San Francisco. A promoter of eco-friendly transportation, bike sharing emerged as a potential complement and alternative to traditional transport solutions. However, not all bike sharing systems are proved to be successful as demonstrated by Pronto's failure in Seattle in 2017.

Citibike, a leader in the bike sharing operator market, officially entered Manhattan, New York City in 2013 and gradually expanded its services to Brooklyn. With another ambitious expansion covering the Bronx, Queens, and the distant neighborhoods of Brooklyn expected in 2023, Citibike must have a strong conviction that there will be a market for bike sharing systems in these not yet covered neighbourhoods.

Topic Question: Where to Build the Next Bike-sharing Station?

In a role of data scientist, we would like to propose the bike-sharing company a data-driven solution to answer the question of which areas to prioritize in order to maximize the number of future users in these target expansion areas, under limited time and budget.

In this report, we will explore provided datasets in depth and conduct statistical hypothesis testings with informative visualizations of covered neighbourhoods' ridership data. After a thorough data analysis, we will fit a predictive regression model of ridership using carefully selected variables. This fitted model can then predict which uncovered areas could have the highest number of potential new riders for the operator company to prioritize. To be more specific, the following aspects will be studied as part of our research:

1. Demographic features: Ridership and demand could be driven by demographic features in target areas. Distinct demographic groups might have very different riding behaviour.
2. Public transportation: Presence of subway stations in an area could affect bike-sharing ridership since people could choose Citibikes to solve their "last mile" problems.
3. Ride-sharing solutions: As an alternative to bikeshare for short-distance commutes, popularity of ride-sharing solutions could affect bikeshare ridership.

After drawing conclusions from the extensive data exploration and statistical testing following the above perspectives, our study will provide insights for Citibike executives about geographical bike-sharing demand and assist in the making of their expansion plan. Hopefully, our research will also help people obtain a better understanding of the driving factors behind the broader trends in NYC transportation industries, which could be eventually extrapolated to other metropolitan cities such as Boston and SF.

2 Executive Summary

An initial exploratory data analysis allowed us to discover strong time patterns in ridership data, visible differences in ride popularity with regards to geographical location, and allowed us to define typical demographic features of a bike-share user.

These observations lead us towards digging more in depth into demographic features and competition with other transportation solutions. After testing a wide spectrum of hypotheses using statistically robust methods, we obtained the following key findings:

- Bike-sharing systems as a mean of transportation are more preferred in areas with older and high-income population
- The closer and the more popular the nearest subway station is, higher the bike-share ridership. This highlights the "last mile" complementary effect of bike-sharing services
- Even though bike-sharing and ride-sharing solutions have both experienced strong dynamics, they are substitutes for short-distance commutes

These findings prompted us to consider these factors as building blocks of a bike-sharing ridership prediction model. The resulting regression model highlighted the following findings:

- Demographics and alternative transportation features both have strong significance as individual regressors to the ridership data
- However, they have strong multi-collinearity which prompted us to perform a principal component analysis in order to improve prediction performance

- Together, the above regressors achieved an adjusted R^2 of 0.8832 which indicates strong predictive power

Finally, our model provided the below recommendation and expansion strategy for target areas in the Bronx, Brooklyn and Queens:

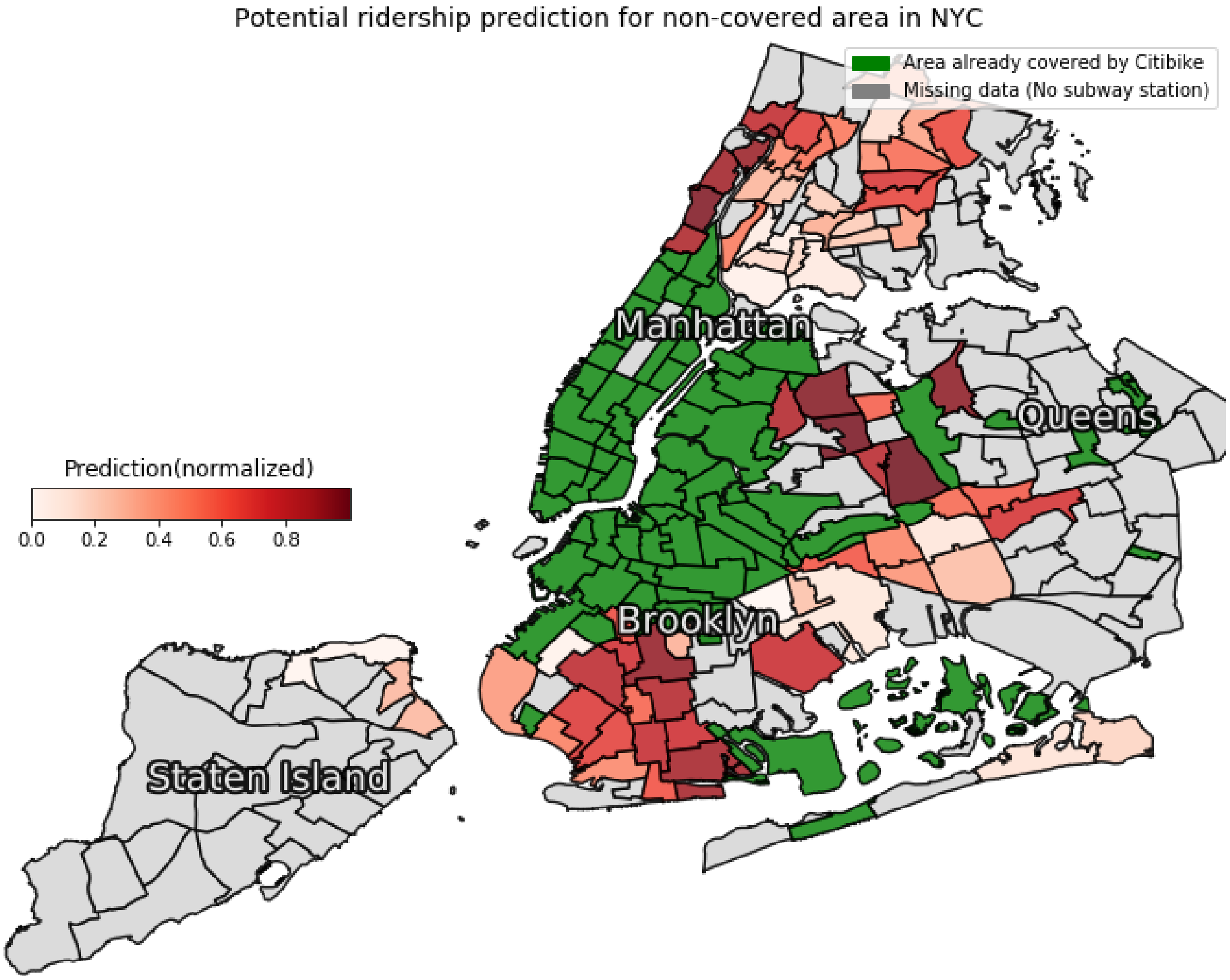

Potential ridership prediction for non-covered area in NYC
🟩 Area already covered by Citibike
▨ Missing data (No subway station)
Prediction(normalized)
0.0
0.2
0.4
0.6
0.8
Manhattan
Queens
Brooklyn
Staten Island

Figure 1: Ridership prediction in non-covered areas

Thus, from a purely ridership-driven point of view, executives at Citibike could consider prioritizing the south side of Brooklyn over the Bronx, since those uncovered areas in Brooklyn have characteristics that seem to indicate a higher immediate demand than in the Bronx.

3 Technical Exposition

3.1 Data Wrangling and Feature Engineering

For our analysis, we used the following datasets.

- *nyc_bikeshare.csv*: New York City bike-share ridership data with ride-specific information
  Because we only have access to geographic coordinates of each station, an additional mapping had to be performed. We constructed neighbourhood boundaries using polygon information from *geographic.csv* then checked the membership of each station to these polygons to find out which Neighbourhood Tabulation Area it belongs to.

- *nyc_rideshare.csv*, *nyc_yellow_taxi.csv*, *nyc_green_taxi.csv*: Trip data from NYC Uber and taxi services that we used to study the relationship between bike-share and ride-share usage.

- *mta_trips.csv*, *mta_key.csv*: Trip data from NYC public subways from 2013-2019 that we used to examine the relationship between bike-share and public transportation. The geographical coordinates (longitude, latitude) of *mta_key.csv* are switched over which necessitated addtionnal debugging.

- *demographics.csv*: Demographic data for each Neighbourhood Tabulation Area

We also used the two following external datasets to add additional insights into the defining characteristics of each NTA:

- *real_estate.csv*: Real estate home value (USD) of NYC by precincts. The data is then mapped from precinct level to NTA level.

- *shooting.csv*: Crime (shooting) cases with case location's longitude and altitude. The data is mapped from longitude and latitude to NTA level.

3.2 Exploratory Data Analysis

3.2.1 Observation 1: Increasing trend and strong annual seasonality in Citibike ridership

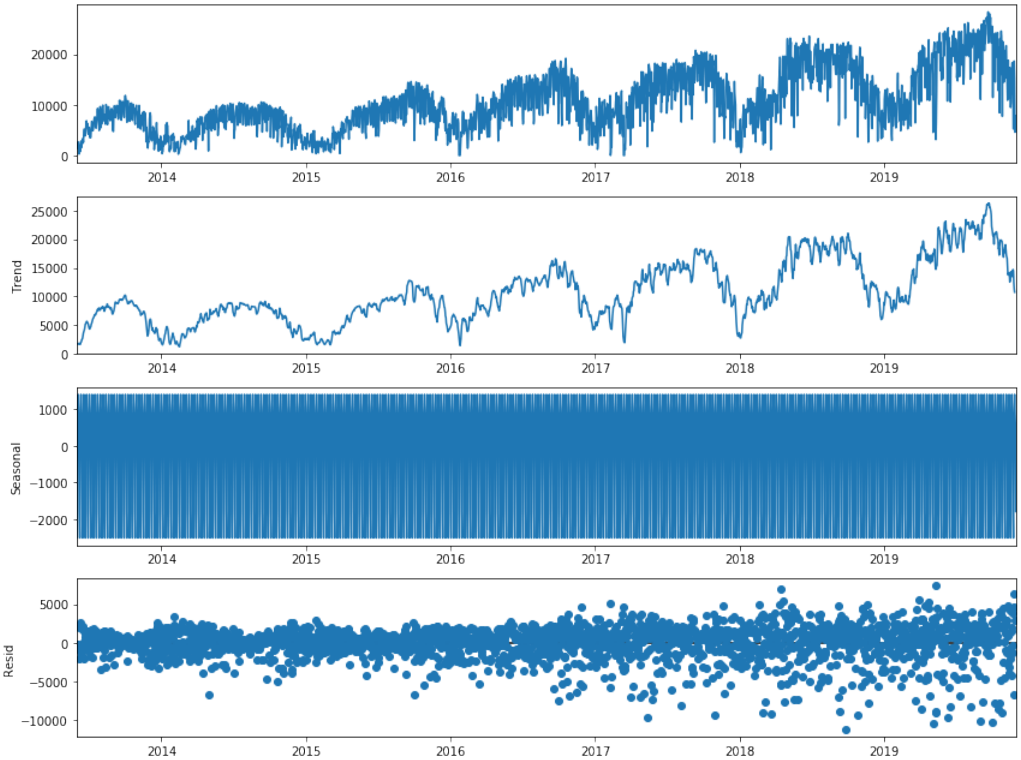


Figure 2: Time series decomposition of bike-share ridership.

From the decomposition plot of Citibikes daily ridership time series, we observed an upward trend over the last five years. More specifically, there is an obvious annual pattern and the variances of residuals grow as time goes. The latter might be explained by the fact that Citibike added more stations as time went, introducing additional data points.

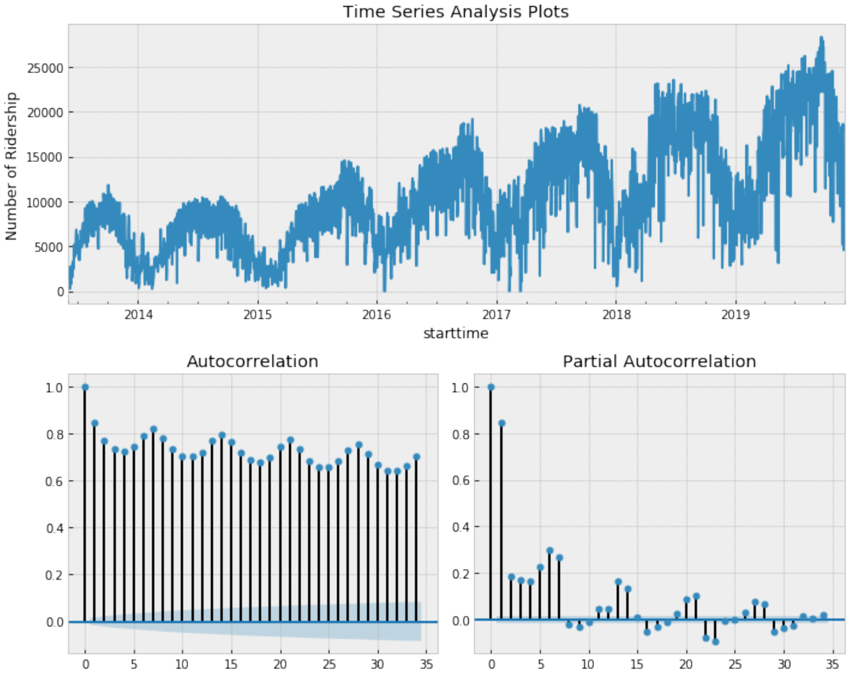


Figure 3: ACF-PACF plot of ridership time series.

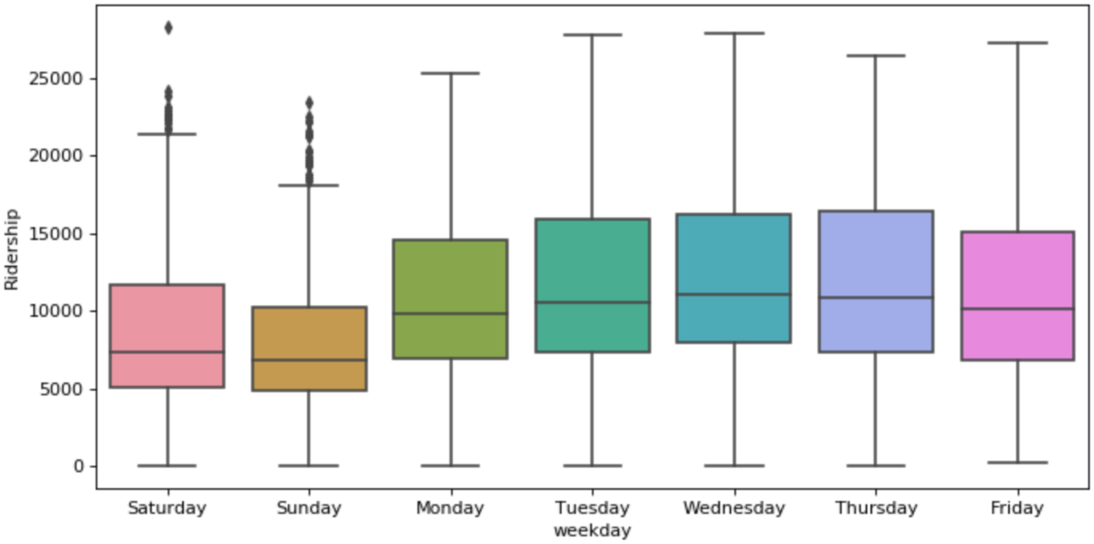


Figure 4: Bike-share ridership by weekday.

According to the ACF-PACF plots above, it is apparent that the seasonal pattern shows every 7 days. Based on these strong evidences completed by the Figure 4 box plot, we can conclude that the ridership on business days are higher than ridership on weekends.

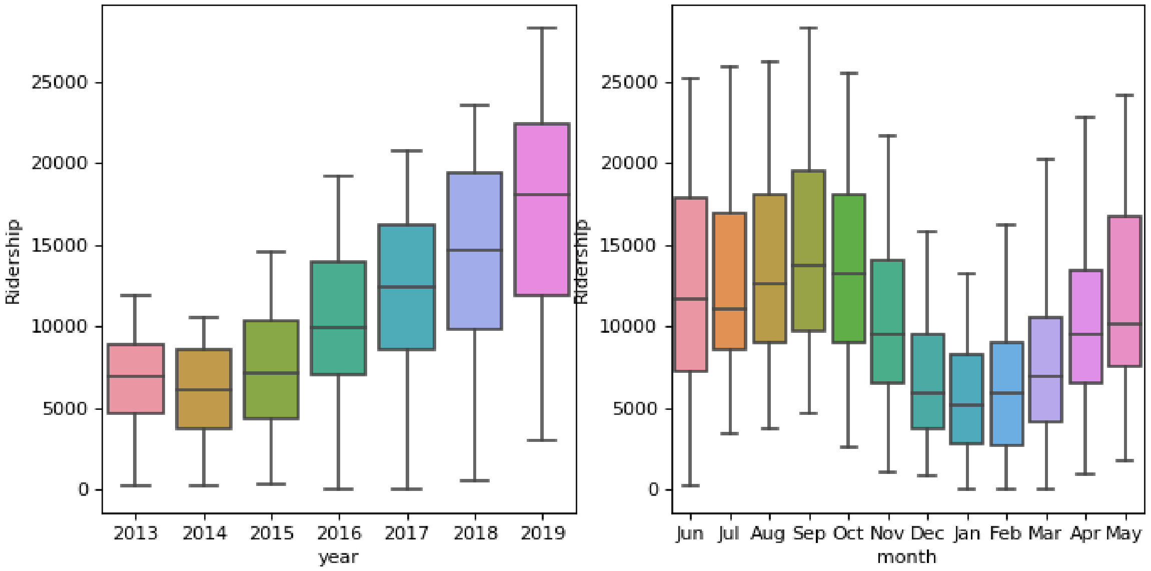


Figure 5: Decomposition of Ridership Over Years

Moreover, the annual pattern shows that besides the growth of CitiBikes in NYC, people use CitiBikes more during summer time (June to Oct) than during winter time and spring time (Nov to Apr). This could be explained by the tourism peak period or pleasant weather condition during summer.

3.2.2 Observation 2: CitiBike rides are mostly short-duration commutes and riding time shows obvious peak-hour patterns

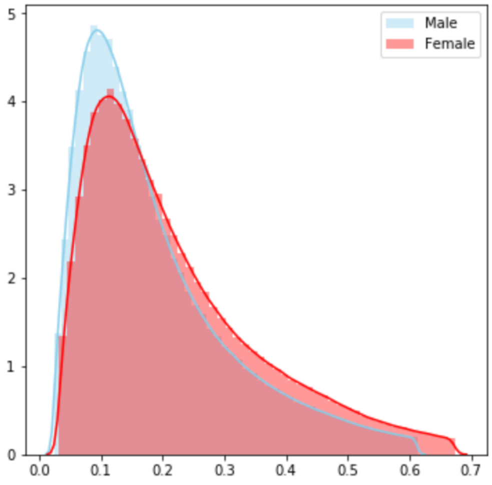


Figure 6: CitiBikes' ride duration (hour) grouped by gender

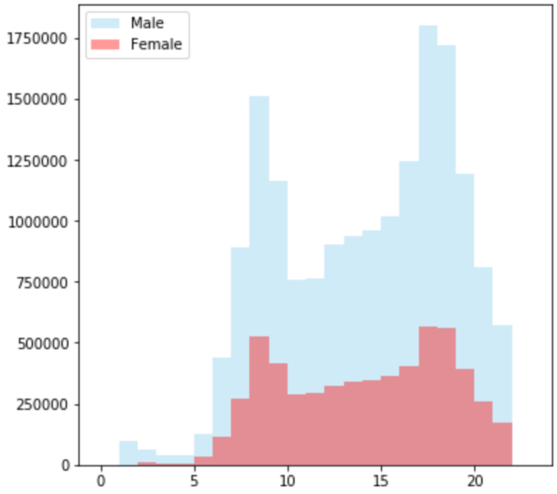


Figure 7: CitiBikes' riding start time grouped by gender

From the histogram of ridership duration (in hours) grouped by gender, besides the fact that there is substantially more male bikers than female bikers, we can observe that most of the CitiBike ride durations are around 0.15 hours (10 mins).

In addition, from the right histogram, two ridership peaks at morning peak hours (8-9 am) and afternoon peak hours (5-6 pm). This common pattern in both genders indicates that many CitiBike riders are biking to commute to work and from work. This may suggest the relationship of Bikeshare and other commuting options, which we will rigorously explore with statistical testings in later sections.

3.2.3 Observation 3: Bike-sharing usage is affected by closeness to famous business and commercial districts in Manhattan

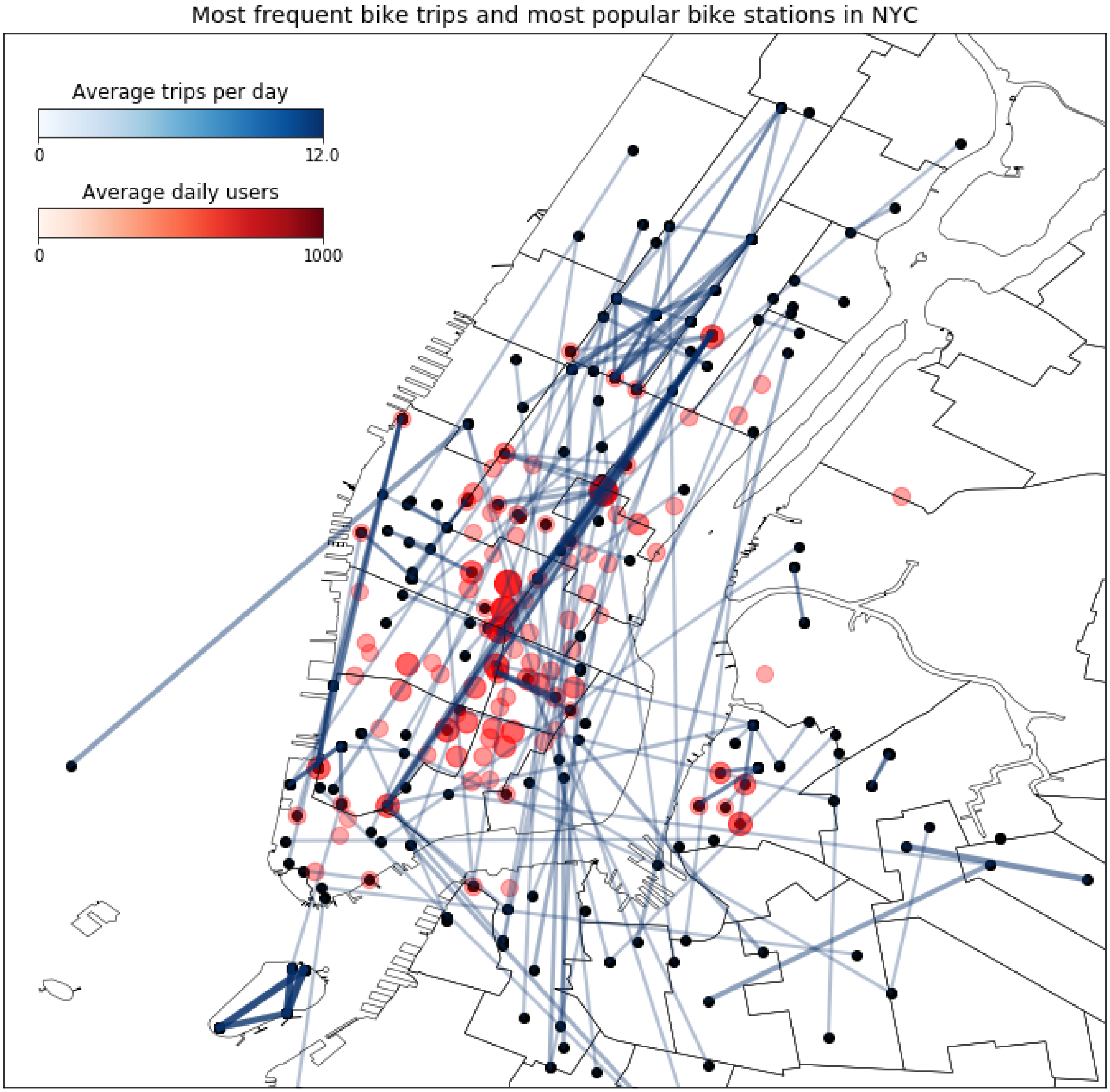

Most frequent bike trips and most popular bike stations in NYC
Average trips per day
0
12.0
Average daily users
0
1000

Figure 8: Most frequent bike trips and popular bike stations in NYC.

On the figure above, we laid out on a map of New York City the top 200 most frequent bike-sharing routes (entry-exit station pair) with blue lines and top 100 most popular bike stations with red circles. The color intensity is proportional to respectively the number of trips alongside this route and the average number of daily users.

We have a few interesting observations:

- Most biking activities are concentrated around Manhattan with a few popular stations in Brooklyn. This could be explained by the longer historical presence of Citibike in the borough as well as denser population in Manhattan.
- There is significant activity alongside landmarks associated with biking activity: Governor's Island tour, Central Park tour and the famous bike lane on Park Avenue.

We see that bike-sharing popularity within an neighbourhood is highly affected by local geographical features and human activities. This lead us towards examining more in depth the different demographic features of each neighbourhood.

4 Hypothesis Testing and Illustrations

To better understand what factors may guide the new construction plan, we consider its relationship to human society on the NTA-scale, including the demographic characteristics related to Bikeshare ridership patters, the role that Bikeshare plays in providing commuting options for people. More insightlyful findings require quantitative analysis in addition to intuitive observations, we further back-up our findings with statistical hypothesis testings in depth.

4.1 Finding 1: Areas with older and higher-income population tend to use CitiBikes more, compared to areas with younger and lower-income population

Since which neighborhood the bike station belongs to seems to have an effect on ridership demand, we classified New York City neighborhoods at NTA level into geographical clusters based on demographic features such as income and age, as well as crime rate and real estate price that we obtained from external datasets. Then, we examined the relationship between each of these clusters and CitiBikes' ridership data to find out what could be the main driving factors that potentially have impact on bike-sharing ridership.

First of all, we applied K-means classifications to classify NTAs by population characteristics. To obtain the optimal K (number of clusters), the elbow method is used. From the graph below, we choose K = 4.

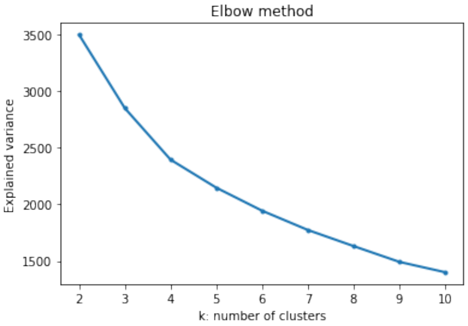


Figure 9: Elbow Method of K-Means Classification

Besides household characteristics from provided datasets, some external datasets are used in this section. We use shooting case numbers within an area to approximate the neighborhood safety and average home prices as a possible demographic indicator.

After classification, we perform Principal Component Analysis (PCA) to evaluate the importance of new clusters. As shown in the 3D Cluster graph, a promising cluster pattern could be observed. We can tell that cluster 1, cluster 2, and cluster 3 are efficiently separated along PC1 which explains over 60% variance. Cluster 4 is separated from the other ones along PC2. The first three PCs explain more than 85% variance in the data.

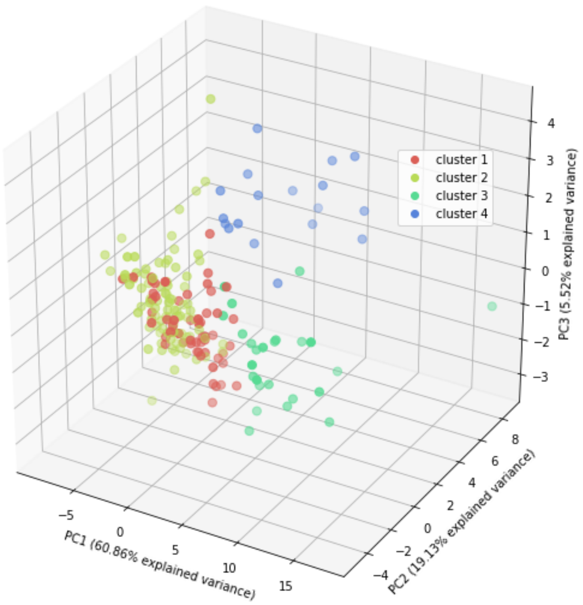


Figure 10: 3D Clusters with Explained Variance

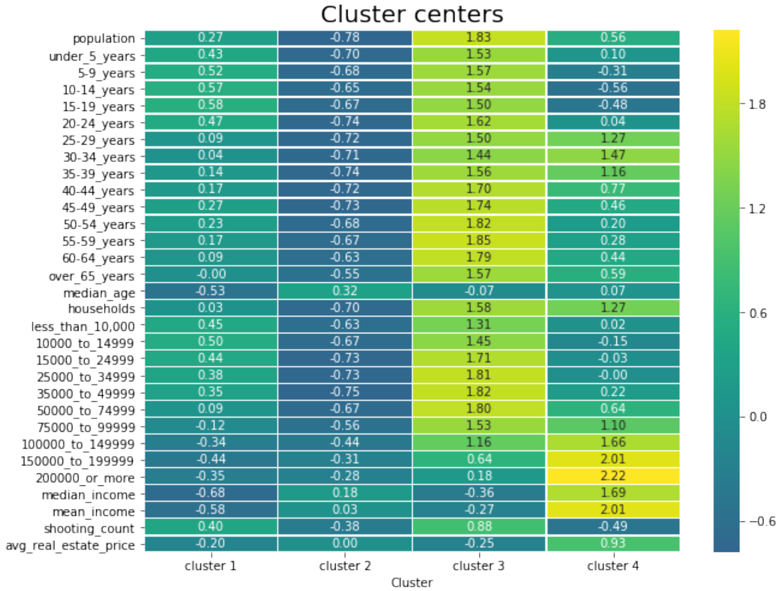


Figure 11: Cluster Centers with Demographic Features by NTA

According to the loading factor heatmap as shown above, the driving forces of each cluster are sharply unveiled. Cluster 1 membership is contributed by young-age group and relatively low-income group. Cluster 2 membership is associated with features of low population number and higher-age group. Cluster 3 membership is characterized by features like high population number, high income, high age population, and low safety level. Cluster 4 membership is mainly determined by group with high safety level and group with high real estate price.

The clusters could be visualized in the NYC map as below.

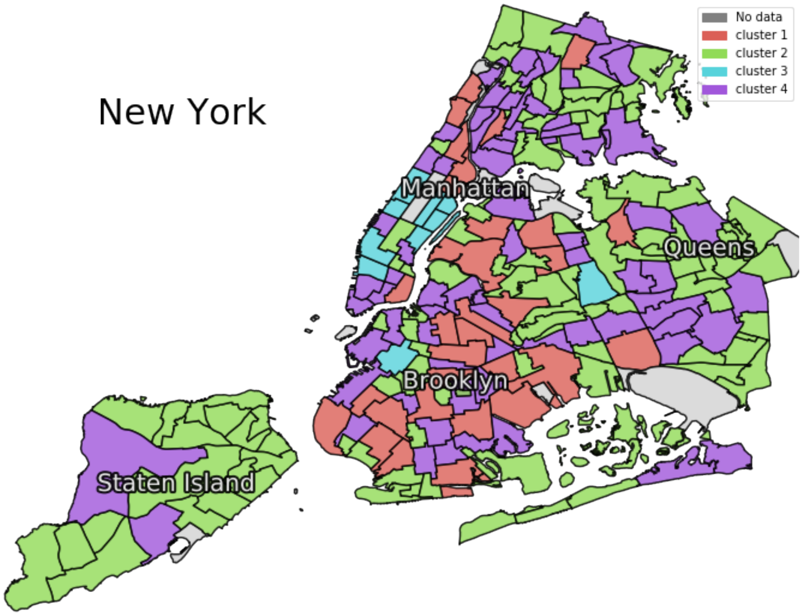


Figure 12: Heatmap of Clusters

When we associate the geographical clusters with Citibikes' ridership, the riderships within four clusters show very different distributions. To test the difference in cluster means, we firstly apply F-test to assess the inequality of four clusters' means.

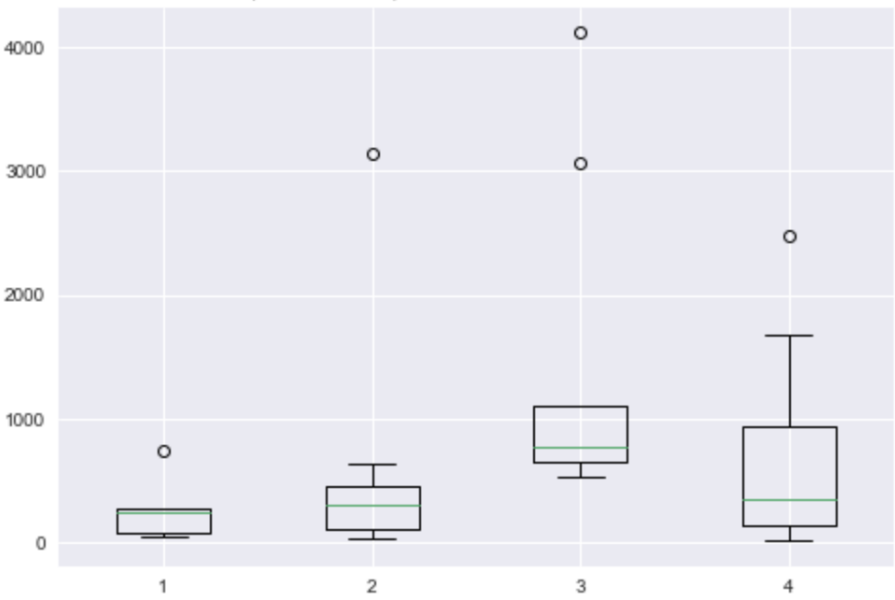


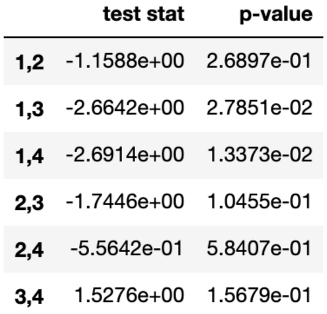
Figure 13: Boxplot of Ridership in Different Clusters

We have the null hypothesis for four-group F test, which is relatively robust to its assumptions of normality:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

We get F-statistics = 3.7866 and p-value = 0.01618, where we reject the null hypothesis at 95% confidence level and conclude that four clusters have significantly different means.

Furthermore, we look at pairwise differences of these clustered NTAs, which helps us identify what are driving demographic factors that tells apart the ridership modes. As the ridership is always non-negative, it is hard to be strictly normal, so we conduct Welch's t-tests to improve robustness of our hypothesis testings.



	test stat	p-value
1,2	-1.1588e+00	2.6897e-01
1,3	-2.6642e+00	2.7851e-02
1,4	-2.6914e+00	1.3373e-02
2,3	-1.7446e+00	1.0455e-01
2,4	-5.5642e-01	5.8407e-01
3,4	1.5276e+00	1.5679e-01

Figure 14: Welch's T-Test Result Table

From the result above, we can tell (cluster 1, cluster 3) and (cluster 1, cluster 4) have statistically significant means in ridership. Cluster 1 corresponds to young-age and low-income population, cluster 3 is residence of high income and high age population, cluster 4 is the area of high safety level and high real estate price. The test result states that geographical cluster (cluster 3) with wealthy and old population has higher ridership than cluster (cluster 1) with young and low-income population. The cluster with safe and highliving standard (cluster 4) results in higher ridership than poor-living-standard cluster (cluster 4).

This seems counter-intuitive at first, since many people assume that younger people are more open to such new things, and they may use Bikeshare to reduce commuting costs. However, as National Association of City Transportation Officials mentioned, the reason behind this phenomenon is that, "stations tend to be especially spread out in low-income neighborhoods, which contributes to lower usage by lower-income groups overall."

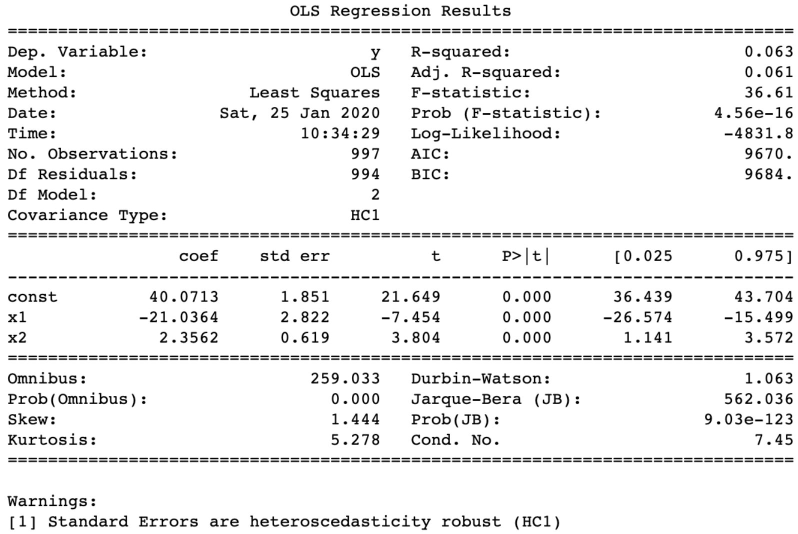
4.2 Finding 2: Bikeshare ridership is highly associated with its relationship with the nearest subway station, indicating the complementary role of bikeshare to public transportation

There is no doubt that the success of CitiBikes is contributed to the fact that it effectively helps people to commute "the last mile" to their destinations. Subway certainly functions as a great complementary of CitiBikes, because commuters usually combine subway and bikes to complete their travels.

To examine the complementary relationship bewteen bikeshare and subway, we fit a linear regression model by regressing the CitiBike ridership in NYC on 1) the distance of a bike station to its nearest subway station 2) the ridership of the subway station.

We aim to fit the following model:

Bike Ridership = β_0+β_1∗Distance to Nearest Subway Station+β_2∗Ridership of the Nearest Subway Station



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.063
Model:                            OLS   Adj. R-squared:                  0.061
Method:                 Least Squares   F-statistic:                     36.61
Date:                Sat, 25 Jan 2020   Prob (F-statistic):           4.56e-16
Time:                        10:34:29   Log-Likelihood:                -4831.8
No. Observations:                 997   AIC:                             9670.
Df Residuals:                     994   BIC:                             9684.
Df Model:                           2
Covariance Type:                  HC1
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         40.0713      1.851     21.649      0.000      36.439      43.704
x1           -21.0364      2.822     -7.454      0.000     -26.574     -15.499
x2             2.3562      0.619      3.804      0.000       1.141       3.572
==============================================================================
Omnibus:                      259.033   Durbin-Watson:                   1.063
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              562.036
Skew:                           1.444   Prob(JB):                    9.03e-123
Kurtosis:                       5.278   Cond. No.                         7.45
==============================================================================

Warnings:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

Figure 15: OLS Regression Summary Table

From the summary table, t-statistics and p values show that the bike ridership is negatively related to distance to nearest subway station, positively related to the ridership of nearest subway station.

Specifically, every 1 kilometer of drop in the distance from a certain bike station to the nearest subway station brings about -21.04 ridership to and from the bike station, while every 1 ridership of the nearest subway station increases ridership to and from the bike station by 2.37. This value is larger than 1, which may due to the trend of people commuting in all ways, indicated by ridership of subway.

With results above, we validate our assumption that the CitiBikes, as a complementary of subway that helps solve the 'last-mile' problem, tend to have higher ridership if its station is close to subway stations. Meanwhile, the popularity of the nearest subway station plays a role in bike station popularity as well.

4.3 Finding 3: Even though riderships of Bikeshare and rideshare are both increasing over time, they are substitutes for short-distance commutes, especially Uber/Lyft rides

The riderships of both bikeshare and rideshare grow over time. We compare the weekly bike ridership time series to that of rideshare and TLC, where we aggregate the overall ridership of all stations within each week. We consider the weekly pattern because it's appropriate in the sense of illustration while keeping the trending and seasonality information. We can see from the stacked percentage time series graph that, the bikeshare ridership series is volatile (resulted by its seasonality effect) while the TLC series are more stable. Bikeshare and Uber/Lyft share have increasing market shares over time. More importantly, bikeshare and rideshare complement each other as the decrease of one usually comes with the decrease of the other.

We also decompose the TLC trips into short trips and long trips with a threshold of 3 miles. Both kinds of trips decrease overtime giving the rise of app rideshare and bikeshare.

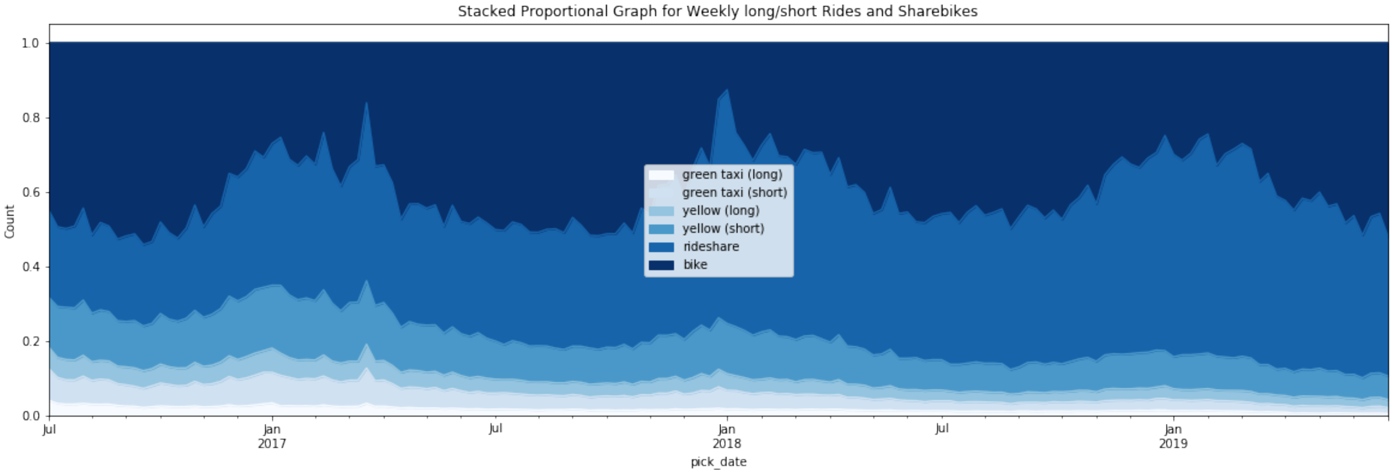
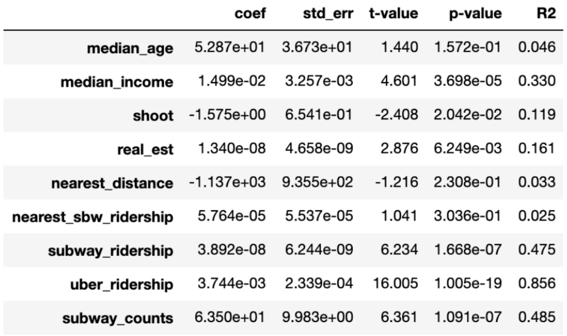


Figure 16: Stacked Proportional Graph Between Different Tranportation Methods

5 Factor Analysis Using Regression Model

As discussed in the previous sections, we find that the following features could be potentially driving forces of CitiBikes ridership demand:

1) Demographic factors: *median_age*, *median_income* 2) Neighborhood factors: *shoot_count*, *real_est* 3) Subway complementary factors: *nearest_distance*, *nearest_sbw_distance*, *subway_ridership*[1], *subway_counts*[2]



	coef	std_err	t-value	p-value	R2
median_age	5.287e+01	3.673e+01	1.440	1.572e-01	0.046
median_income	1.499e-02	3.257e-03	4.601	3.698e-05	0.330
shoot	-1.575e+00	6.541e-01	-2.408	2.042e-02	0.119
real_est	1.340e-08	4.658e-09	2.876	6.249e-03	0.161
nearest_distance	-1.137e+03	9.355e+02	-1.216	2.308e-01	0.033
nearest_sbw_ridership	5.764e-05	5.537e-05	1.041	3.036e-01	0.025
subway_ridership	3.892e-08	6.244e-09	6.234	1.668e-07	0.475
uber_ridership	3.744e-03	2.339e-04	16.005	1.005e-19	0.856
subway_counts	6.350e+01	9.983e+00	6.361	1.091e-07	0.485

Figure 17: Single feature screening result table by regression model.

By regressing the CitiBike ridership on the individual regressor, we obtain the table as shown above.

[1] *subway_ridership* the daily sum of subway stations within a NTA

[2] *subway_counts*: the total num of subway stations within a NTA, an indicator of subway station density

To ensure significance of regressors, we want R-squared to be as large as possible and p-value to be lower than 5%.

Thus, we select the following regressors to include in the sub-model: median_income, shoot, real_est, uber_ridership,subway_ridership,subway_counts.

To further improve the robustness of prediction model, we compared the submodel with full model where all the regressors are considered. Applying F test, we tested the following hypothesis:

$$H_0: \text{ Submodel is true}$$

$$H_1: \text{ Full model is true}$$

We obtained F-test statistic = 1.0561 with degree of freedom (3,41). Since p-value is 0.3782 which is higher than significance level 0.05, we fail to reject null hypothesis and have evidence to exclude the spurious features, and choose the sub-model as final prediction model.

However, the submodel still suffers from severe multi-collinearity within regressors. To fix this, we use Principal Component Regression (PCR) to find best linear combination of our raw features (normalized). It's shown that the first four components could explain 95.08% variance, so we pick the first four principal components as new features in PCR. In this way, we improve the robustness and goodness of fit of our regression dand prediction model, yielding an adjusted R-squared of 0.8839. The first and the third components are most significant with a p-value smaller than 0.01. The first component is consisted of positive combination of most variables except for the crime rate.

6 Prediction for the Optimal Station Sites of Undeveloped Area

We used available explanatory features of NTAs that are not yet covered by CitiBike to predict the total potential daily bikeshare ridership within these areas. Since we included the number of subway stations inside NTAs as a feature, the areas without subway lines cannot be fitted into our prediction model. Therefore we left these districts outside of our test set, with the rationale that these places are less-crowded since subway stations are not constructed there.

The (normalized) predicted ridership within these available NTAs are visualized in Figure 18, with green-colored areas already covered by Citibike and grey-colored areas not considered by our model. The predictions are also normalized to reveal the suggested amount of attentions to different areas.

Interesting patterns and suggestions can be drawn from our predictions. The darkest areas, corresponding to the direction of Brooklyn and NTAs connected to Manhattan, are most strongly suggested to extend CitiBike to, while other more isolated or faraway places may have less potential demands for Bikeshare.

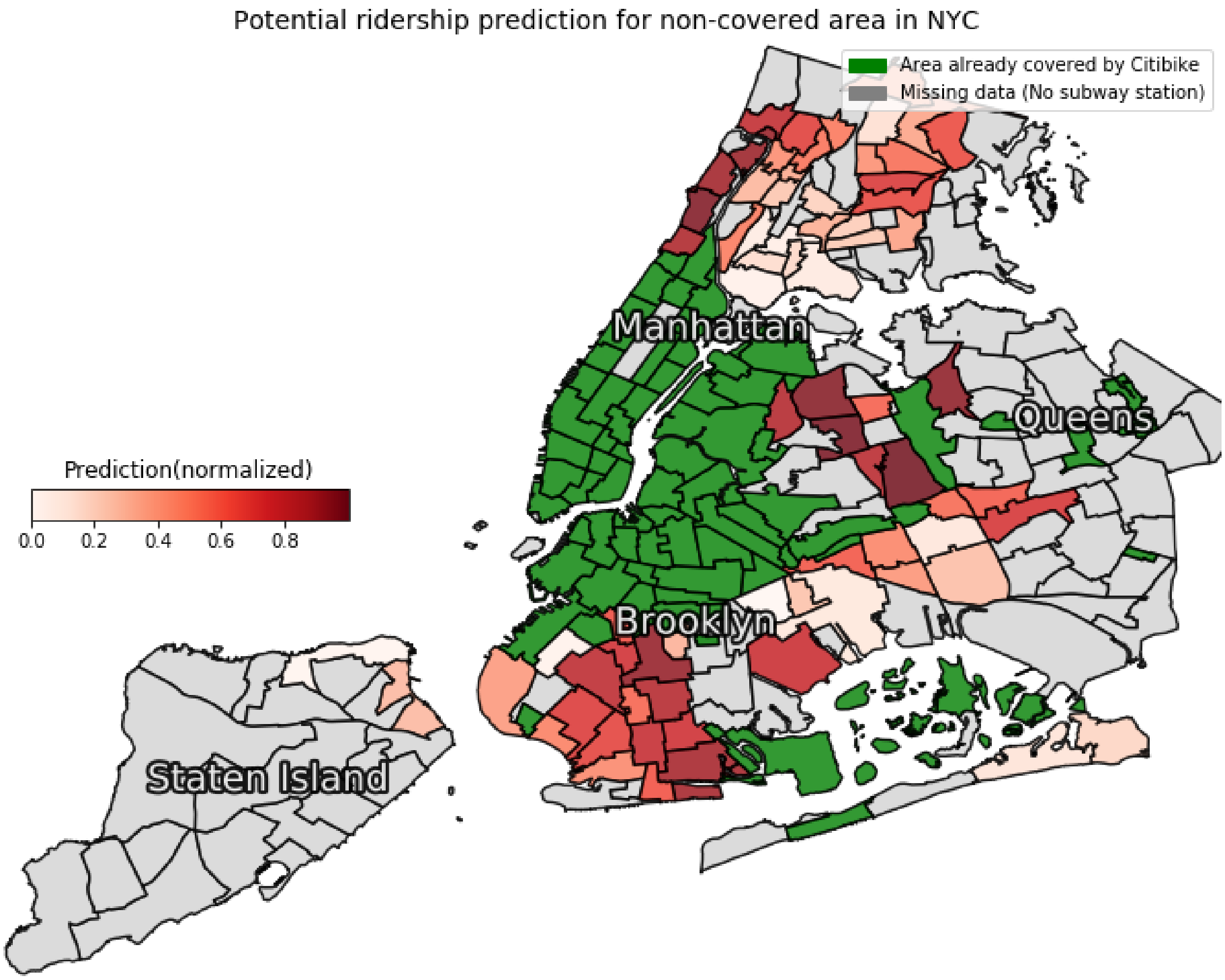

Potential ridership prediction for non-covered area in NYC
🟩 Area already covered by Citibike
▨ Missing data (No subway station)
Prediction(normalized)
0.0
0.2
0.4
0.6
0.8
Manhattan
Queens
Brooklyn
Staten Island

Figure 18: Ridership prediction in non-covered areas

7 Conclusion and Further Research

The same BikeShare station recommendation system could also be applied to other metropolitan cities in the U.S such as SF and Boston, because the ridership demand are majorly contributed by demographic data, neighborhood features, and nearest complementary public transportation.

However, for other cities the driving factors might not be the same and additional factors need to be considered. For example, in San Francisco where there are many steep slopes, we can predict that slopeness of the target will have a big impact on bike ridership.

References

[1] Blei, David M. Probabilistic topic models. Communications of the ACM, 2012, vol. 55, no 4, p. 77-84.

[2] New York Neighborhoods Crime Data (shooting): https://www.neighborhoodscout.com/ny/new-york/crime

[3] New York City Real Estate Data: https://www.census.gov/topics/housing/data/datasets.html