The Citi Bike Usage Analysis: Clustering to Address Rebalancing

Team members: Xurui Chen(xc1454), Yushi Chen (yc3763), Yunhe Cui (yc3420), Yuchen Ding(yd1402), Manrique Vargas (mv1742)

Introduction:

New York City launched the largest bike-sharing system in North America in 2013 and today the system has 12,000 bikes and 750 stations in 60 city neighborhoods. The system has been a success with ridership approaching 40000 trips per day. There is a set of management and optimization problems coming along with the success, however. In New York City, depending on locations of Citi Bike stations, the bike usages are varied because of differences in the geographical location, the demography, and traffic pattern. The problem arises when the trips are generated in an unbalanced fashion among different stations. This makes some stations full of bikes (no place for end rides) or empty stations (no trip generation). Using a clustering method to cluster all the Citi Bike stations by analyzing end and start trips throughout the day, we can help to solve this problem. In this research, we use 2015 Citi Bike data for all the bike stations to build a prediction model on Citi Bike usage. Then we further analyze the socio-economic features to find special patterns. Moreover, based on the model we built, the future usage of Citi bike could be predicted. In the end, some suggestions on optimization for each bike station will be delivered.

Objective

Use clustering technique to divide the groups of station having similar imbalances at different time of the year. This will allow to create a rebalancing strategy for the bike sharing operator.

Data

In our research, we used the Citi Bike trip data from 2015 which has the trip duration, start/end time and location as well as some user information (gender, age and user type). We combined the individual trip data to a Citi Bike station dataset, which contains the basic information of 818 stations. Through grouping the station ID and calculating the start/end Citi Bike trip times in the trip dataset, we summarized the dataset of total hourly trip times which start from and end at a particular station. For each Citi Bike station, we had 48 sets of characteristic features attached. By clustering those 818 Citi Bike stations, we summarized that those stations fell into 6 different types of clusters.

Some other data that is included in our research to further help to understand the characteristics of each cluster. The data includes: 1) The real-time data for each Citi Bike station: We gathered the real-time Citi Bike data every 10 minutes to collect the empty station number. Then we calculated the sum of the hourly empty times in two weeks. As each station was record same times in two weeks, it is reasonable to use empty station times to represent the empty station rate.

2) Demographic data (by taxi zone) which may have an impact on the usage difference(time and amount) in each Citi Bike cluster: The demographic features include residents' gender, age, the working population distribution, and the residential population distribution. 3) Metro data: We measured the number of nearby metro stations for each Citi Bike station. We used each Citi Bike station as the centroid and a radius of 20km as basic value to visualize and identify the density of the subway station near each Citi Bike station. (Figure 1. Histogram of the distance of Subway station and Citi Bike Station). Identifying from the Figure 1, the value of 1.25 km will be used as the final radius value. 4) Other: We included traffic data into our research to identify the potential influences of the vehicle traffic condition and the usage of the Citi Bike.

Methodology and modeling

The k-means clustering method was used as our methodology for our exploratory research and built out the prediction model. K-means clustering method is a type of unsupervised machine learning algorithm that used to draw an inference from the dataset that is consisting of input data without labeling the responses. The method calculates the cluster means after randomly assigning the cluster's center. Furthermore, we used Elbow method (Figure.2) to find out the optimized number of Citi Bike clustering. Elbow method is used to find the appropriate number of cluster analysis.

For the dataset that we prepared before, we didn't consider the capacities for each Citi Bike station when we are using k-means clustering method for the CitiBike data, because the capacity could be unified or randomly decided, which in some level is influenced by humans. Moreover, we assumed that the information about capacity can be ignored when the dataset for clustering and analyzing is huge. As we use hourly usage situation to cluster the type of CitiBike station, which based on the demand from a customer per hour, the capacity cannot reflect the demand from the customers. Thus, we will only use the hourly start and end trip in each station. Based on the Elbow plotting for K-means clustering, we identified that six clusters are the optimized number of clustering for the Citi Bike stations in the New York City.

We measured the correlation between demographic factors and the 6 different types of Citi Bike stations, and also explore the potential relationship between traffic data and the subway data. Our model can be used to make a prediction on Citi Bike future usage in the 6 different clusters.

Result & Data Analysis

From the correlation coefficient of the external factors with the hourly Citi Bike usage, we will know which time interval is affected most by those external factors. Then will visualizate the external factors and citibike stations in the map to see the geography relationship. Then we will identify the different characteristics for the 6 clusters and observe the spatial visualization map. We analyzed the result from 4 following perspectives:

Demographical: The comparison between residential and working population

First, we tried to measure the correlation between residential/working population and the hourly starting/ending trips in each Citi Bike stations. The feature we measured for each population group includes gender(male/female) and age(under 29, between 30 to 54 and over 55). Based on the correlation coefficient results for two different population groups, we found that the Citi Bike usage has a negative correlation with the residential population(Figure.3), but has a positive correlation with the working population. It indicated that the working population contributed more on the Citi bike usage than the residential population, and the bike rides are strongly influenced by the commuting patterns of the working population (most commonly day jobs). For both working and residential population, female riders have a slightly stronger correlation with the Citi Bike rides usage(based on the correlation scores), and according to the data and statistical score, the trend of the Citi Bike usage started to increase from morning peak hour and reached a maximum point around off work peaking hour. The age factor appeared similar influence on the residential and working population, which shows that riders under 29 contribute more on the Citi Bike usage and population that over 55 years old has the lowest contribution on the Citi bike usage. This indicates that the young population, no matter in the residential or working population, rides more citibikes than other two age groups.

A relationship between subway and CitiBike

We calculate that the density of the subway stations within the circle where each Citi Bike stations as a centroid. We found that the distance from the nearest subway station

To the Citi Bike station has a negative correlation with the hourly usage of the Citi Bike. This reveals that when Citi Bike and subway station is close enough, the Citi Bike usage is heavily influenced by the subway; however, the influence was merely detected from the morning rush hours. Moreover, the density of subway station shown a positive correlation with Citi Bike usage. Especially, during night times (from 9 pm to 12 am), the result showed that the higher the density inside the circle leads to the higher Citi Bike usage; however, during the morning rush hour, the two factors shown relatively lower correlations. Furthermore, regarding the ending trip, similar to the starting trip, the nearest station has a negative correlation with the Citi Bike usage, whereas the density of the subway has a positive correlation. Different from the starting trip, the ending trip feature showed that the more subway stations nearby the Citi Bike stations, the more Citi Bike usage it was. (Figure 4)

Trip Duration types per six clusters

In order to further identify the characteristics of the 6 Citi bike clusters, we used the trip duration information to estimate the unique characters for the different clusters. By comparing the mean trip duration for each cluster, we can identify that cluster 0, 2 and 6 have longer average trip duration compared to other three clusters. For cluster 0, the trip numbers and trip durations are the highest among the clusters. By comparing with the clusters locations, stations belong to cluster 0 are generally allocated near Central Park, Downtown Brooklyn, and Chinatown, which indicates that the demand in that neighborhood is relatively higher than others. (Figure 8)

Empty dockers and Usage correlation with Citi Bike

From figure 5 and figure 6, we can observe that the Citi bike trips curves have more dramatic changes than vehicle traffic - because the empty/full stations don't allow it to be continuously smoothed. So the empty stations will have a huge effect on the Citi bike usage. From different types of clusters, we can identify the unique characters for the 6 different clusters:

- Cluster 0 "from stations - balanced, empty at night".

Cluster 0 shows more ‘start’ trips in the morning and more ‘end’ trips in the afternoon. This probably corresponds to the mixed-used surrounding area near the subway that receives bikes in the morning.

- Cluster 1 "from stations - critical morning, empty noon".

Cluster 1 shows much more start trips than end trips in the morning and more end trips than start trips in the afternoon. This suggests a residential area away from the subway station where people take bikes in the morning and commute back in the afternoon.

- Cluster 2 "from stations - residential, empty noon"

Cluster 2 shows much more ‘start’ trips than ‘end’ in the morning and more ‘end’ than ‘start’ in the afternoon. Again, this probably a residential area far away from the subway station where people need to commute extensively by bike.

- Cluster 3 "to stations - stations, empty at night, full during the day "

‘End’ trips number are much larger than ‘start’ trips number in the morning but similar to it in the afternoon - a probable commercial area near a subway station.

- Cluster 4 "to stations - stations, empty at night, full during the day "

These stations are similar to cluster 3 with the difference that there is a different empty station pattern at night. Empty at night would mean that these bikes were ridden at the end of the day, probably to commute to the subway.

- Cluster 5 "to stations - stations, empty at night"

These stations are similar to cluster 3 and 4 with the difference that it is not empty during the day, so probably there is a commercial area.

Discussion

This project used the K-means method to divide 818 Citi Bike stations in New York City into 6 clusters. Based on the result, we clustered all the bike stations into 6 different clusters. The characteristics for those stations indicated that different methods are needed for solving the problems. For those stations, which have higher demands in the morning peak hour(such as cluster 0 and 1), might need transfer more bikes in the bike dockers before the rush hour. For other clusters, such as cluster 2, has more bike trips that end from the station. Those stations have more residential usage than other clusters, which requires more empty bike dockers before the evening peaking hour.

Furthermore, From our research, it can be observed that the 'empty hours' per station factor has a large influence on the Citi Bike usage. The limitation of this feature is that the data available for empty stations could only be accessed in real time. We aggregated the real-time data for two weeks which can give us some information for Citi Bike usage during this particular time of the year. However, gathering real-time data is time and energy consuming. In future research, we can gather a larger set of real-time data, and use this factor 'hourly empty times' per station to cluster the Citi Bike station**.** Figure 7 shows a clustering done by 'empty hours' variable which represents the total stations without any bikes. It can be observed that the clustering separates the stations more evenly by 'empty' and not 'empty' at different times. This cluster will be better for a rebalancing solution in the future.

Conclusion

This applied data science study can help future bike sharing systems to design the location and capacity of bike stations. We found that using clustering can classify the data to make groups that are useful for a potential rebalancing strategy. This project consisted of a descriptive data analysis. Further work should put this into action by making prescriptive work and design the redistribution strategy including the optimization of redistributing routes and time of the day.

Reference:

Contardo, C.; Morency, C.; and L.-M. Rousseau. 2012. Balancing a dynamic public bike-sharing system. Technical Report CIRRELT-2012-09, Universite de Montreal, Canada

Kaltenbrunner, A.; Meza, R.; Grivolla, J.; Codina, J.; and Banchs, R. 2010. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. Pervasive and Mobile Computing 6(4):455 –466. Human Behavior in Ubiquitous Environments: Modeling of Human Mobility Patterns.

Nair, R.; Miller-Hooks, E.; Hampshire, R. C.; and Busiˇ c,´ A. 2013. Large-scale vehicle sharing systems: Analysis of Velib. International Journal of Sustainable Transportation 7(1):85–106.

O'Mahony, E., & Shmoys, D. B. 2015.. Data Analysis and Optimization for (Citi) Bike Sharing. In *AAAI* (pp. 687-694).

Raviv, T.; Tzur, M.; and Forma, I. 2013. Static repositioning in a bike-sharing system: models and solution approaches. EURO Journal on Transportation and Logistics 2(3):187– 229.

Rixey, R., 2013. Station-level forecasting of bikesharing ridership: Station network effects in three U.S. systems. Transportation Research Record: Journal of the Transportation Research Board, No. 2387, pp. 46-55.

Shaheen, S. A.; Guzman, S.; and Zhang, H. 2010. Bikesharing in europe, the americas, and asia. Transportation Research Record: Journal of the Transportation Research Board 2143(1):159–167.

Shu, J.; Chou, M. C.; Liu, Q.; Teo, C.-P.; and Wang, I.-L. 2013. Models for effective deployment and redistribution of bicycles within public bicycle-sharing systems. Operations Research 61(6):1346–1359.

Appendix:

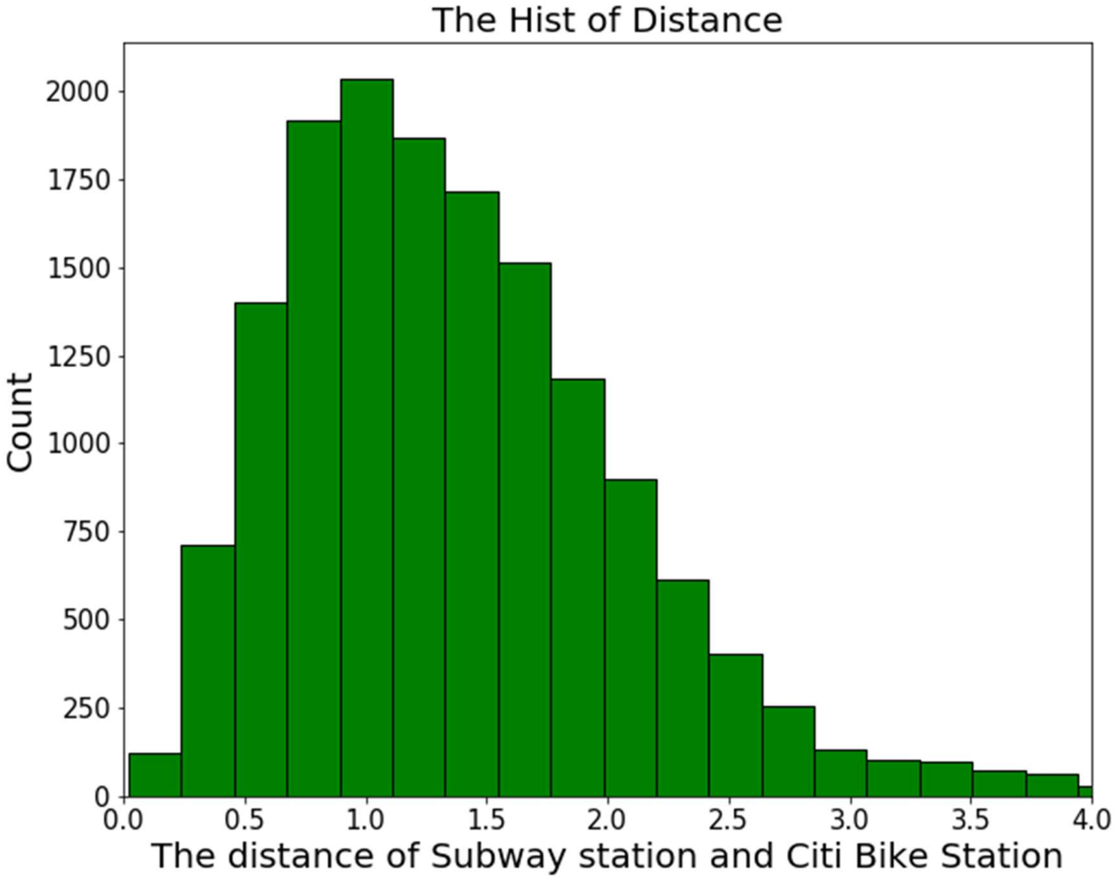


Figure 1: Histogram of the distance of Subway station and Citi Bike Station

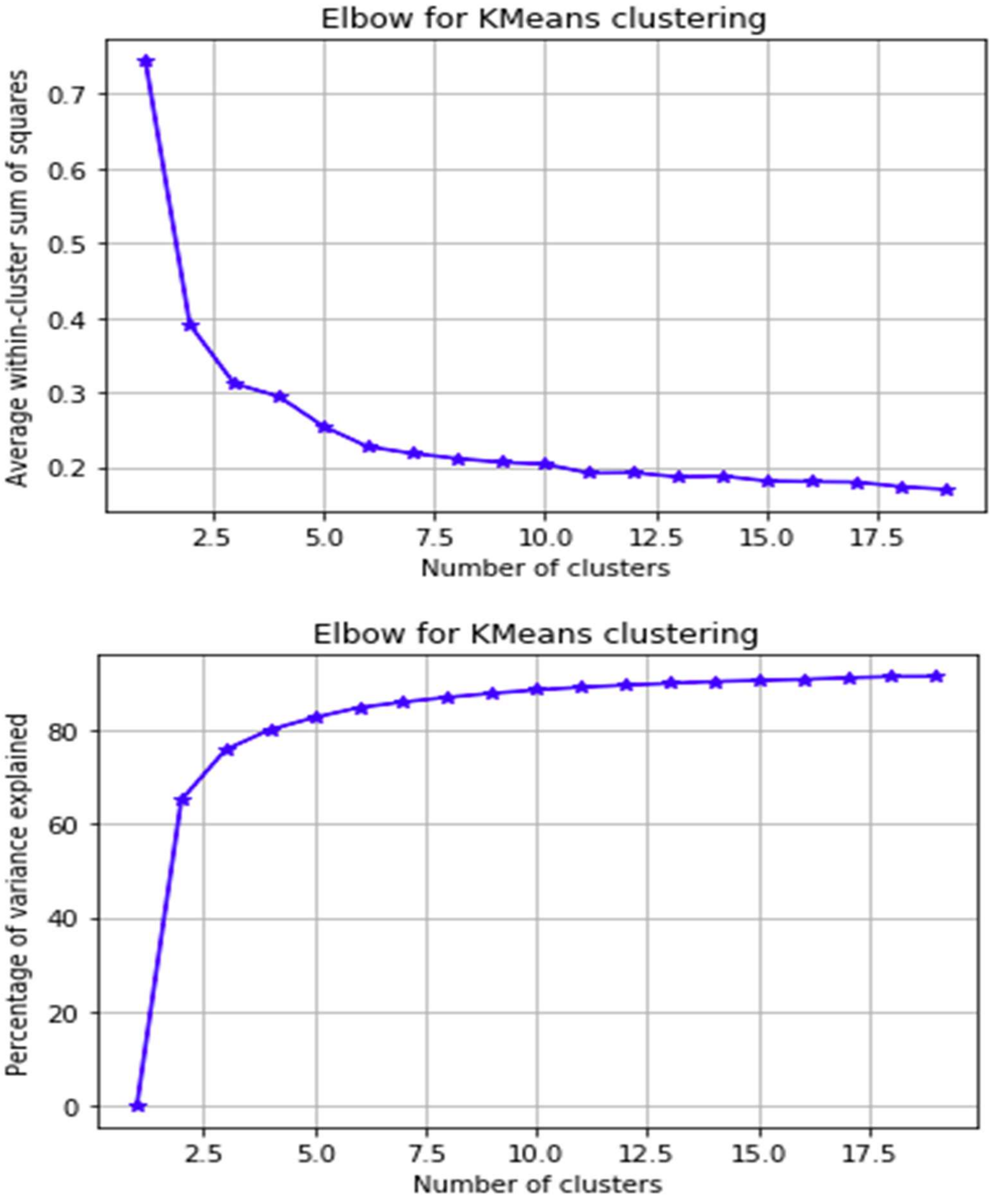


Figure 2: The Elbow for K-means clustering method for Citi Bike Station in New York

\

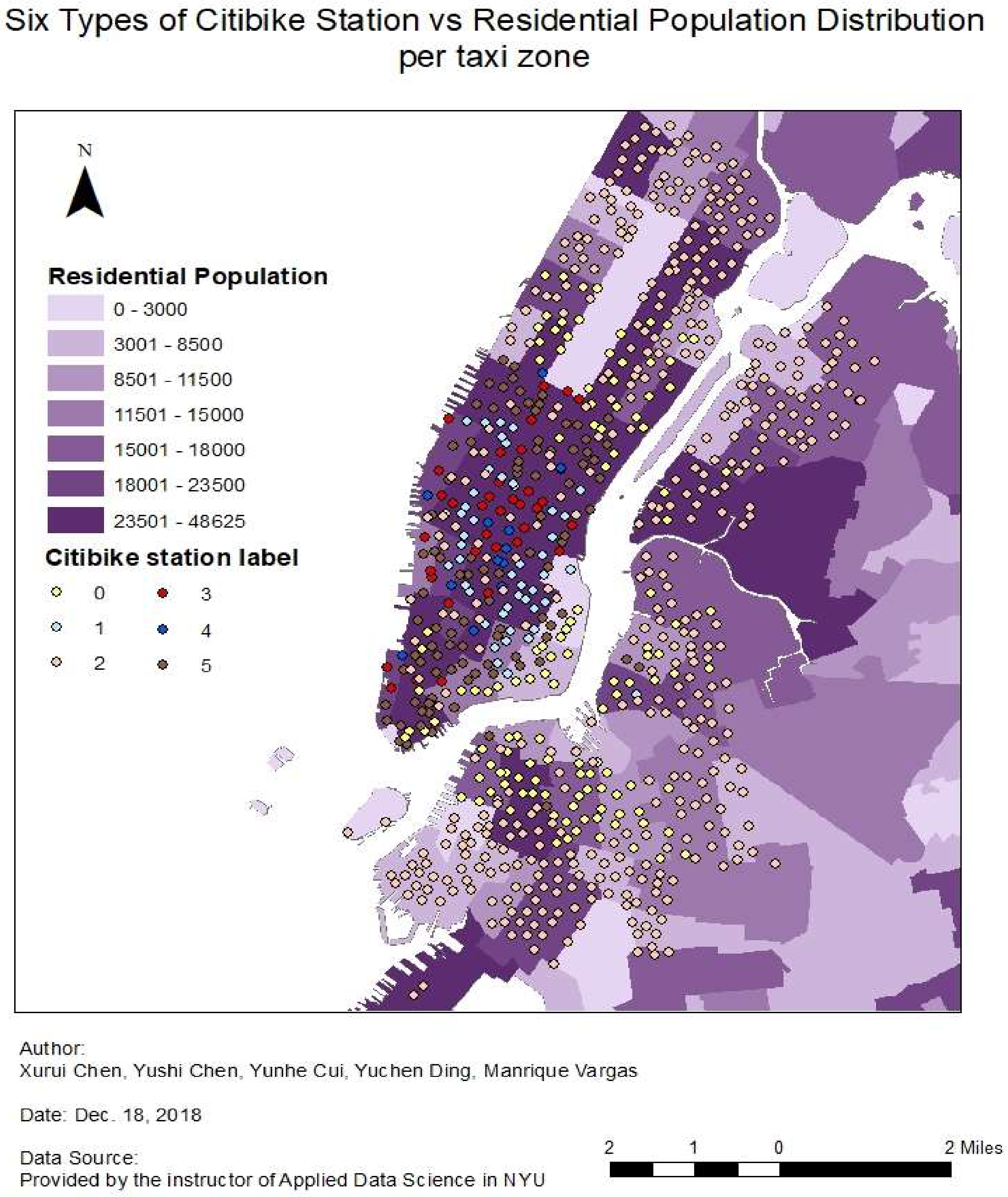
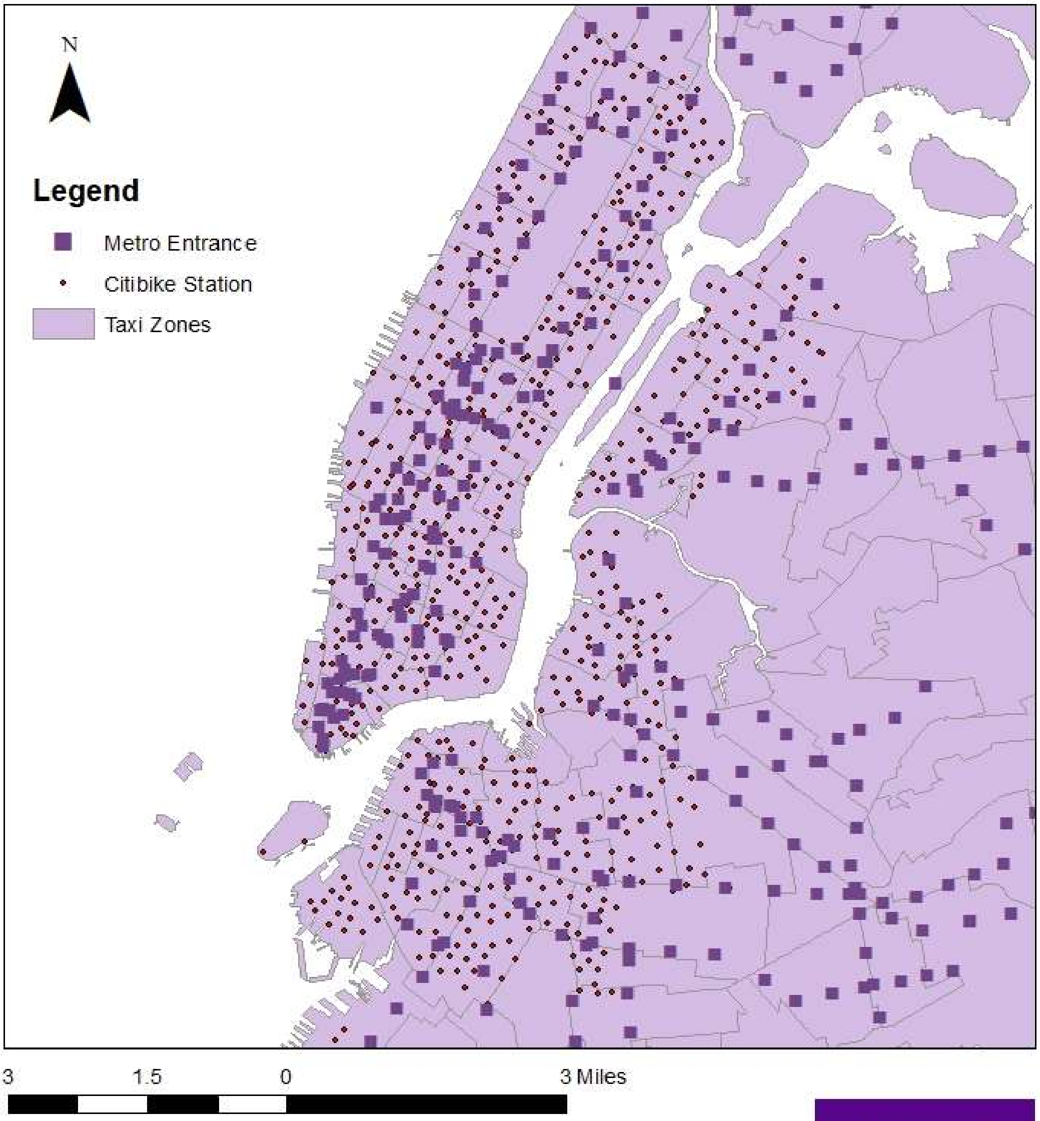

Six Types of Citibike Station vs Residential Population Distribution per taxi zone
N
Residential Population
0 - 3000
3001 - 8500
8501 - 11500
11501 - 15000
15001 - 18000
18001 - 23500
23501 - 48625
Citibike station label
0
1
2
3
4
5
Author:
Xurui Chen, Yushi Chen, Yunhe Cui, Yuchen Ding, Manrique Vargas
Date: Dec. 18, 2018
Data Source:
Provided by the instructor of Applied Data Science in NYU
2
1
0
2 Miles

Figure 3: Six types of Citi Bike stations Vs. Residential population

Citibike Stations and Metro Entrances



Author:
Xurui Chen, Yushi Chen, Yunhe Cui, Yuchen Ding, Manrique Vargas

Date: Dec. 18, 2018

Data Source:
Citibike: Provided by the instructor of Applied Data Science in NYU
Taxi zone: Provided by the instructor of Applied Data Science in NYU
Metro: NYC open data

NYU

Figure 4: Citi bike station and the metro entrance

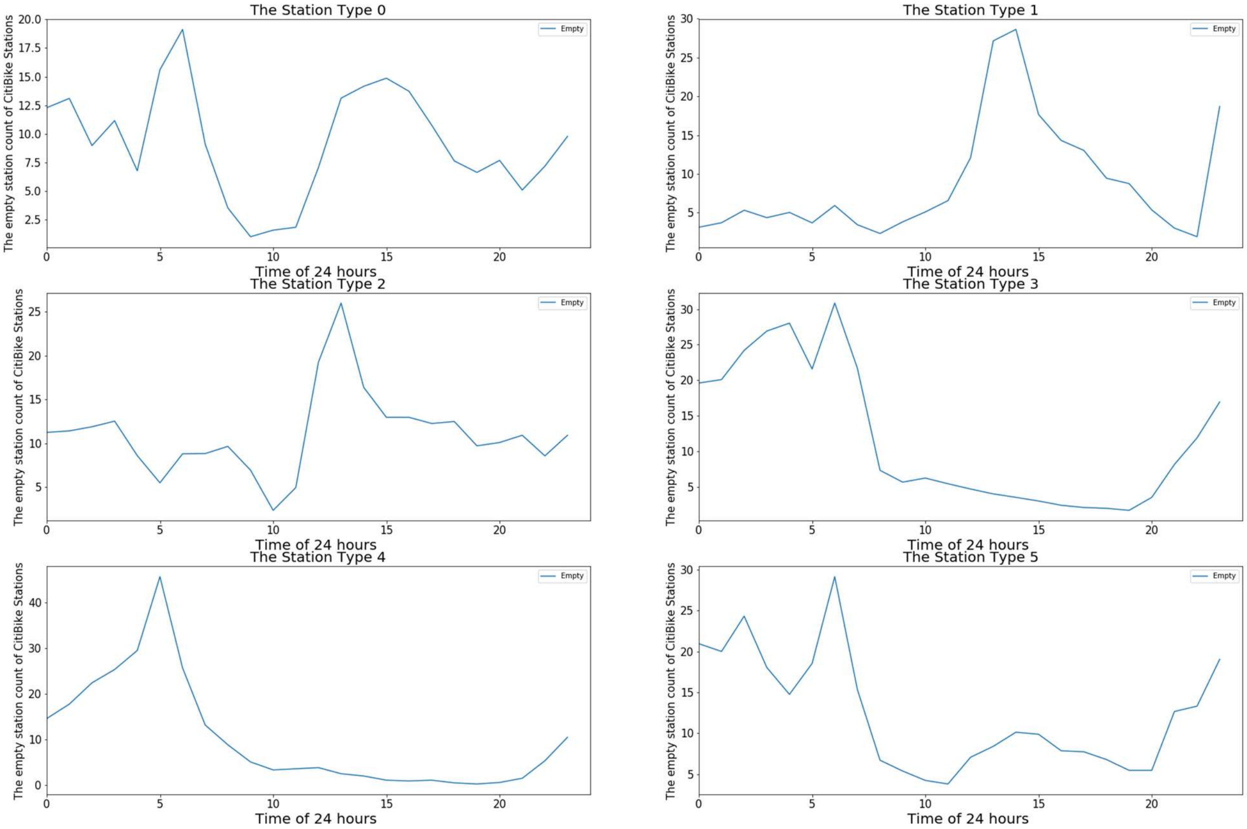


Figure 5: The empty station count of 6 Types of Stations in 24 hours

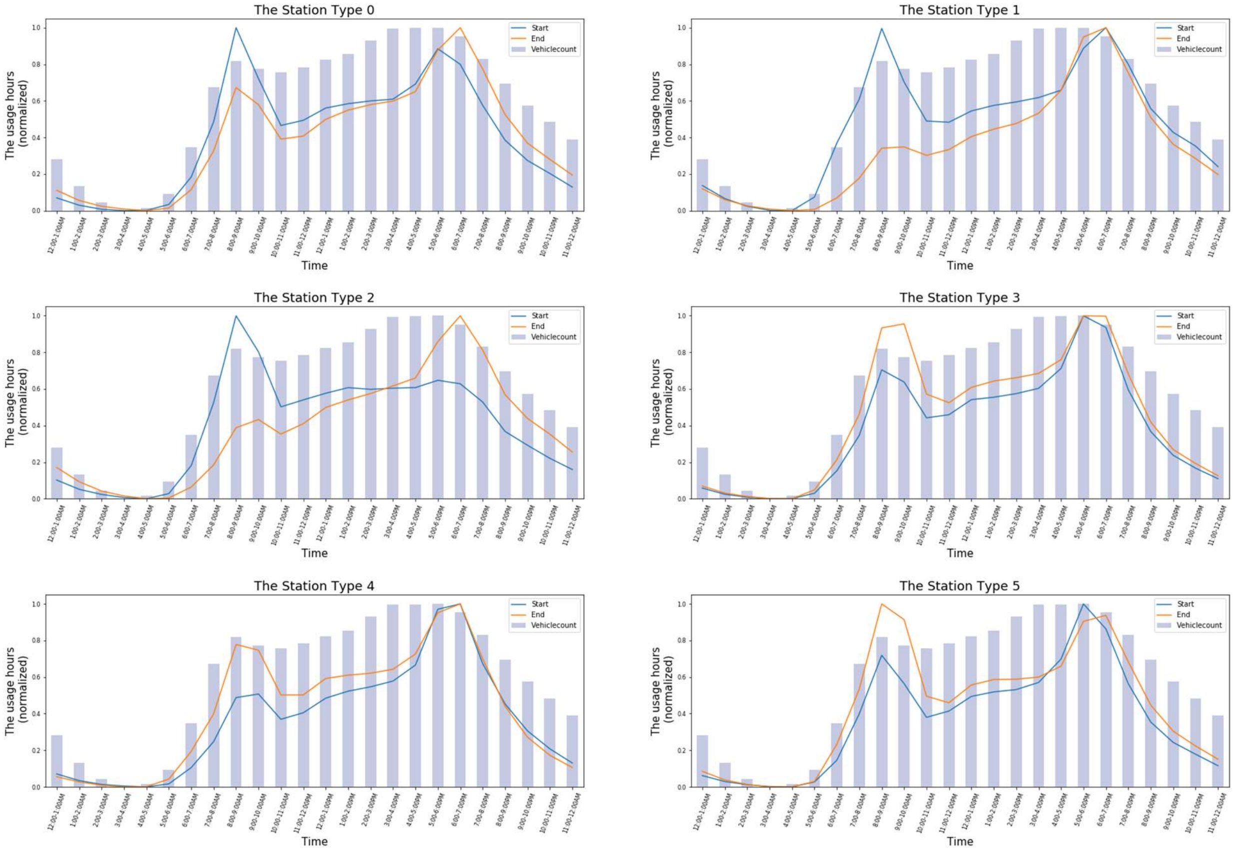


Figure 6. ‘Start’ and ‘End’ Trips and traffic ‘vehicle count’ normalized distribution per hour of the day for clusters using hourly usage

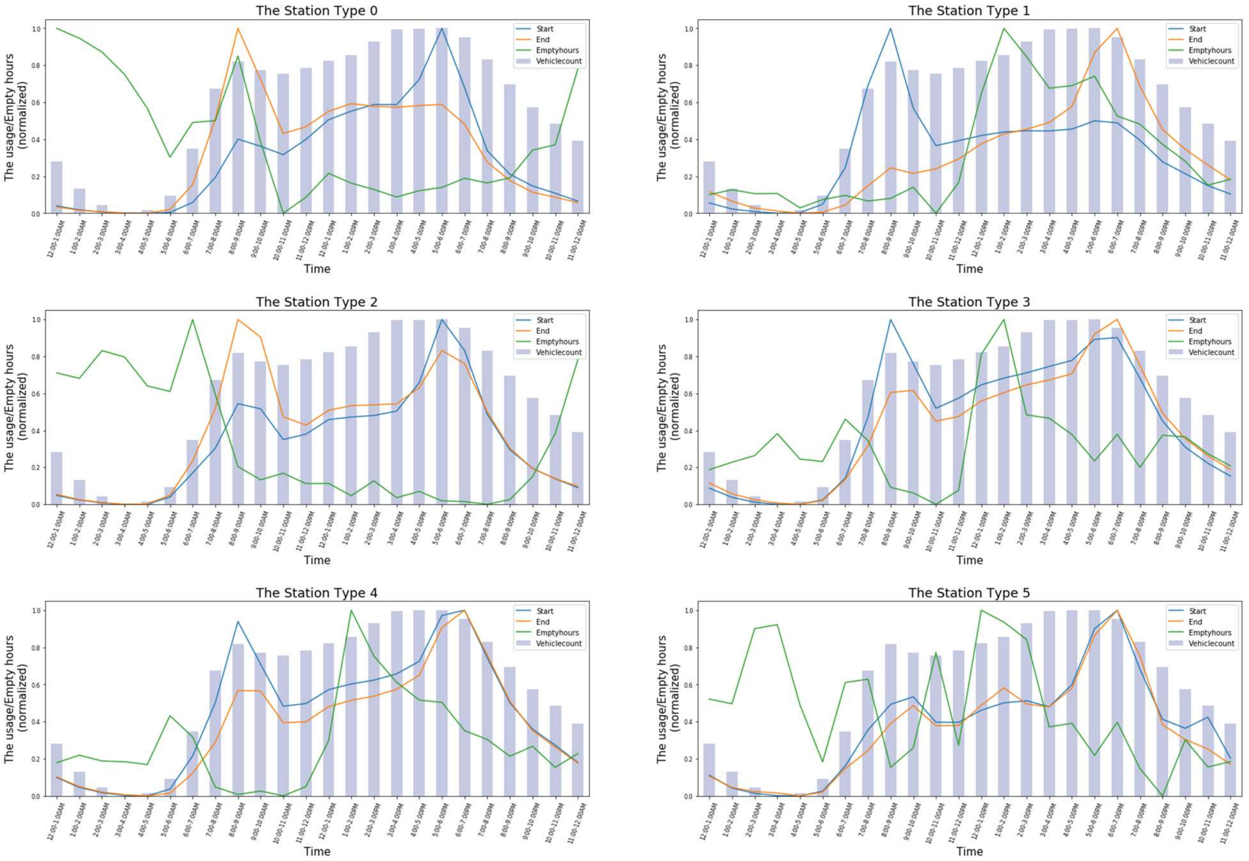


Figure 7. ‘Start’ and ‘End’ Trips, ‘Empty-hours’, and traffic ‘vehicle count’ normalized distribution per hour of the day for clusters using hourly empty stations

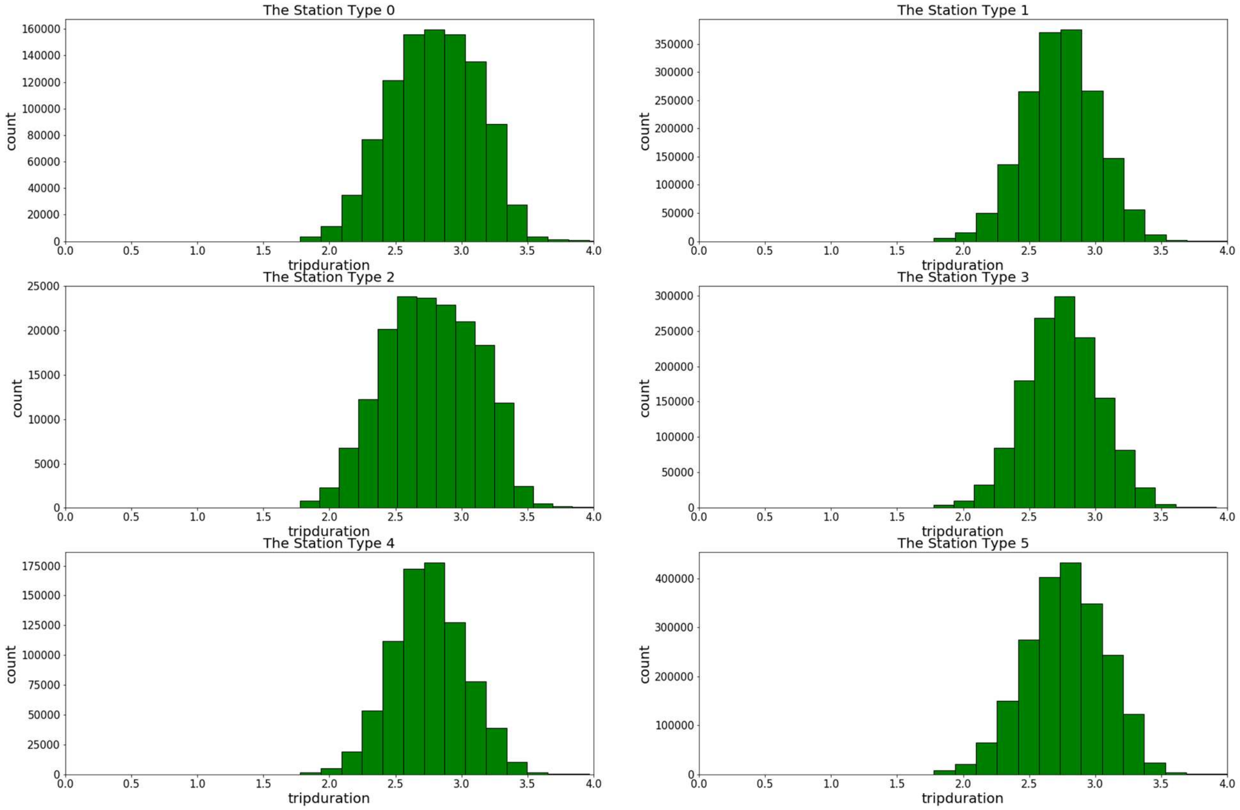
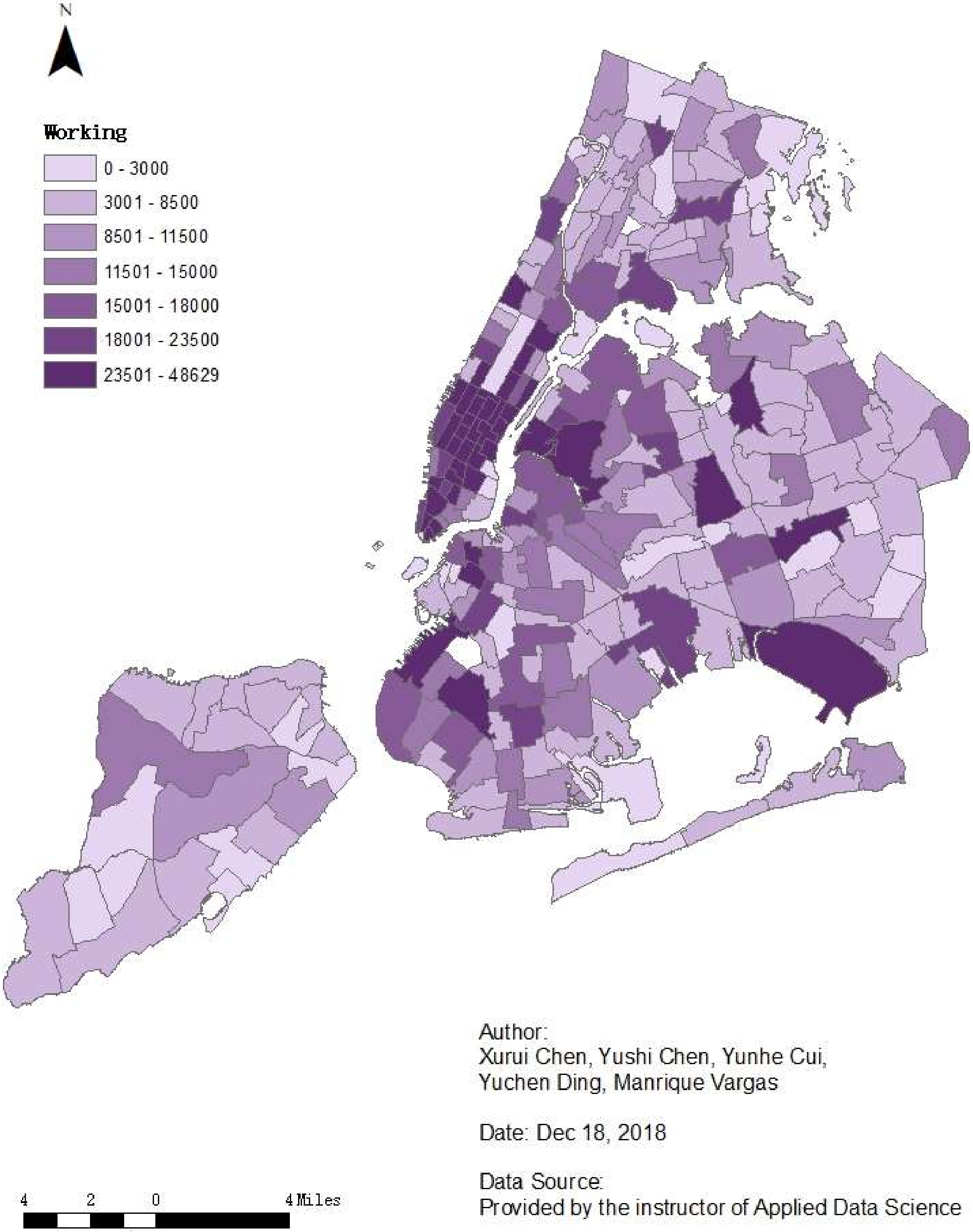


Figure 8. Trip duration per hour of the day for clusters

Working Population Distribution in NYC per taxi zone

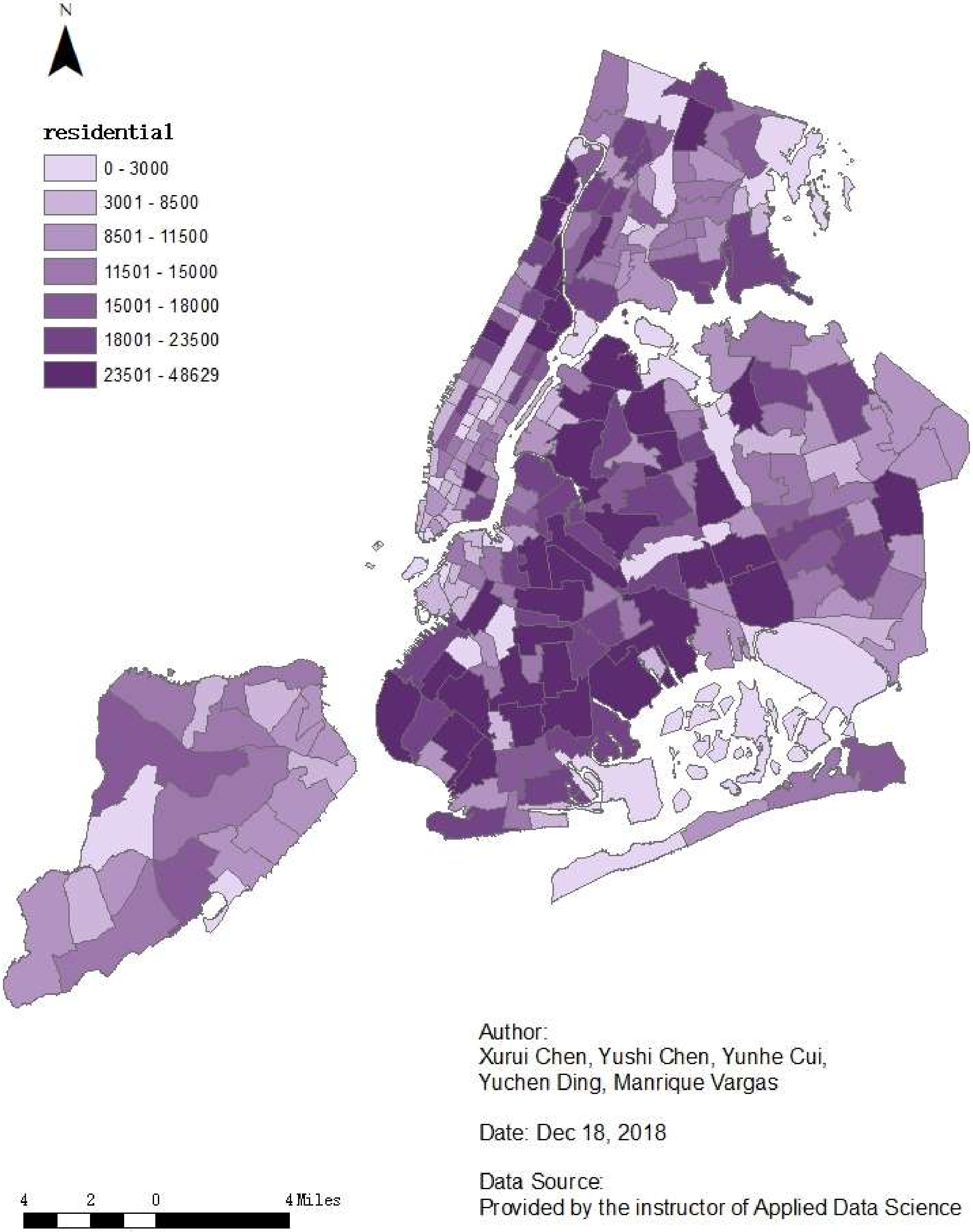


Author:
Xurui Chen, Yushi Chen, Yunhe Cui,
Yuchen Ding, Manrique Vargas

Date: Dec 18, 2018

Data Source:
Provided by the instructor of Applied Data Science

Residential Population Distribution in NYC per taxi zone



Author:
Xurui Chen, Yushi Chen, Yunhe Cui,
Yuchen Ding, Manrique Vargas

Date: Dec 18, 2018

Data Source:
Provided by the instructor of Applied Data Science

For retrieving the code for the project:

Github: https://github.com/Sherryairui/CitiBike_Usage_Prediction_Model