Bike-Share Program in NYC: How Time and Region Affect Supply-Demand Imbalance

Stat 139 Final Project
JY & TW

Introduction

The recent expansion of bike-share programs has been accompanied by regional imbalance of bike supply and demand. Since bikes are continuously transferred between stations by users, problems arise when gaps exist between the volume of local incoming and outgoing trips [1]. Shortage of available bikes, even temporarily, will influence user experience and impede the sustainable growth of the program. Therefore, bikes need to be systematically transferred between regions by the bike-sharing platform to offset the imbalance. The designing and execution of bike transfer requires accurately identifying the neighborhood regions and time period in which supply-demand imbalance exists.

In this project, we applied statistical methods to analyze the data of Citi Bike – a bike-share program in New York City. Different from some previous studies that build model to facilitate night transportation of the bikes [2] or incentivize customers to take the routes that can balance the trips [1], this project focuses on two aims:

(1) Determining important spatial and temporal features associated with the supply-demand imbalance, such as neighborhood regions and time of the day. The identified features are visualized to reveal details of the dynamics of the bike-share program in New York City.

(2) Building multiple linear regression models to predict the number of trips at rush hour and facilitate day-time bike relocation.

Methods

Data

The analysis is performed on Citi Bike System Data of July – September 2016 [3]. The neighborhood regions where each trip starts and ends were identified based on the latitude and longitude information. The total number of docks in each neighborhood region was calculated using Citi Bike Station Data [4].

Data Cleaning and Transformation

i. Normalized trip imbalance:

Normalized trip imbalance was calculated as the difference between the outgoing and incoming trips divided by total number of docks in the region. This value measures the unmet demand of bikes normalized by the supply capacity of the region, and serves as the response variable in the analysis of spatial and temporal effect on the supply-demand imbalance.

$$Normalized\ imbalance = \frac{Outgoing\ trip\ number - incoming\ trip\ number}{total\ dock\ number\ in\ the\ region}$$

Since the distribution of normalized trip imbalance is leptokurtic with heavy tails, we transformed the variable by taking square root of the absolute values and assigning the original signs. Data from five neighborhood regions with few observations were dropped from the analysis,

including "Sunset Park", "Hudson", "Highlands", "Canarsie and Flatlands" and "Central Harlem". The remaining regions include roughly equal numbers of observations (**Supplementary Figure 1**). The distribution of the square-root transformed trip imbalance in each neighborhood were checked. Although the transformed data from some neighborhood regions exhibit bimodal shapes, the distributions are mostly symmetric.

ii. Trip numbers at rush hour:

In order to identify the rush hour, we computed the total number of incoming trips in each hour of day. The result exhibits two clear peaks at 8-9 am (defined as morning rush hour) and 5-6 pm (defined as afternoon rush hour) (**Supplementary Figure 2**). The numbers of incoming and outgoing trips at morning and afternoon rush hours were calculated and used in building prediction models for Aim 2. Since the distribution of hourly trip numbers is highly right-skewed, we transformed these variables by taking the 5^{th} root. The transformation leads to a largely symmetric distribution of the incoming trip numbers at afternoon rush hour, which serves as the response variable in the models, while maintains a linear relationship between the trip numbers at morning and afternoon rush hours. Six neighborhood regions were excluded from the analysis due to few trips at rush hour, which include the five regions excluded in the analysis of the normalized trip imbalance and one additional region "East Harlem".

Aim 1. Investigation of Spatial and Temporal Factors

In order to determine whether neighborhood region, hour-of-day (as a categorical variable) and interaction of these two variables significantly influence the normalized trip imbalance, linear regression models with different numbers of predictors were fitted. Extra-Sum-of-Square (ESS) tests were performed between the nested models to determine whether the extra variable in the larger model plays a significant role in explaining the variability of the normalized trip imbalance. The *p*-values of these tests were adjusted for multiple comparison using Bonferroni correction. Due to the unequal variance between neighborhood regions, we multiplied each observation by a weight, which was computed as the inverse of the variance of the observations from the same region. Compared to the result of unweighted regression, the residuals of the weighted regression conform better to the normal distribution (**Supplementary Figure 3**).

Aim 2. Construction and Selection of Prediction Models

Models were constructed to predict the numbers of incoming trips at the afternoon rush hour. The numbers of incoming and outgoing trips at morning rush hour of the same neighborhood region in the same day were used as predictor variables. Additional predictors include day-of-month, week-of-day, neighborhood regions, average duration and percent of subscribers of the incoming/outgoing trips at morning rush hour.

We employed linear regression models with different sets of predictors to predict the response variable. Each model was trained with data of July and August separately, and tested on data of August and September respectively. The cross-validation (CV) R^2 was calculated for each test on the result transformed back to the original scale. Since the cross validation R^2 is calculated on the test data, it directly estimates the error of out-of-sample prediction and is not affected by the number of predictors used in the model. Automatic model selection methods based on AIC were employed to search for models with better predictive ability.

Due to potential correlation of observations across days, we also converted the data into time series and applied autoregressive integrated moving average (ARIMA) models [5]. The result of this time-series method is presented solely because its comparison with the linear models led to the finding of important predictor variables used in the best linear regression model. The details of this

method are not presented in this report. Instead, a brief explanation can be found in the supplementary material.

Results

Time and Region Affect the Supply-Demand Imbalance

The F-scores of the Extra-Sum-of-Square (ESS) tests suggest that hour-of-day, neighborhood region and the interaction of these two variables significantly affect the trip imbalance of the bike-share program (**Table 1**). Including day-of-week does not explain additional variability of the trip imbalance. Therefore we conclude that the trip imbalance exhibits different patterns within a day in different neighborhood regions. Month also significantly affects the trip imbalance (data not shown), suggesting potential long-term dynamics.

Table 1. Result summary of Extra-Sum-of-squares Tests

Additional predictor	Smaller Model	F-score	*p*-value
Neighborhood	No predictor	7.0235	< 0.001
Hour-of-Day	No predictor	657.01	< 0.001
Neighborhood	Hour-of-Day	91.83	< 0.001
Hour-of-Day	Neighborhood	163.35	< 0.001
Interaction	Neighborhood+ Hour-of-Day	85.807	< 0.001
Day-of-week	Neighborhood × Hour-of-Day	1.4322	0.2314

** The Bonferroni-adjusted critical p-value is 0.083.*

The interactive effect between neighborhood regions and hour-of-day are further illustrated in **Figure 1**. The imbalance reaches the highest point around 8:00 in the morning and 5:00 – 6:00 in the afternoon in most regions. Within regions, the morning and afternoon peaks are of opposite directions, indicating that regions with high incoming (outgoing) volume at morning rush hour exhibit high outgoing (incoming) volume at afternoon rush hour. The timing of the peaks suggests that the bike-share program is mostly used by commuters.

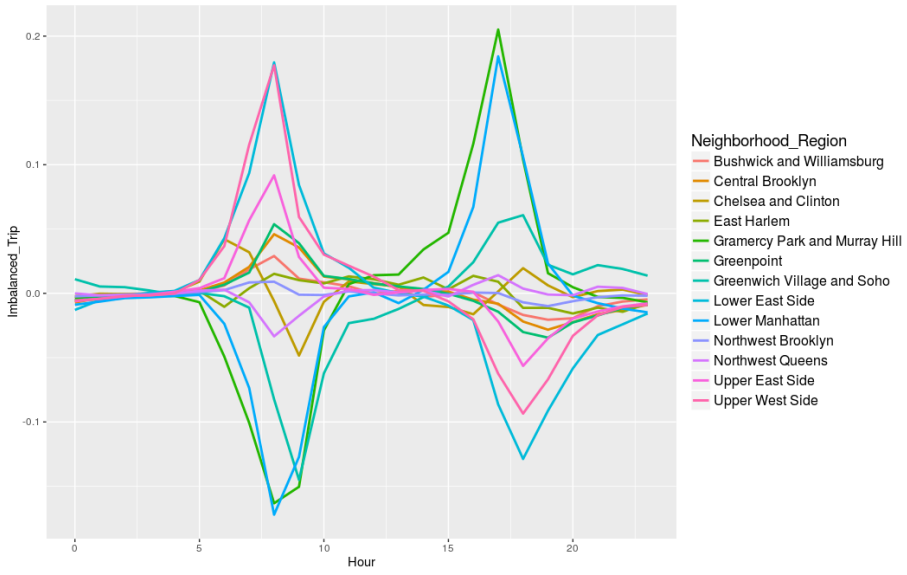


Figure 1. Daily pattern of normalized trip imbalance by neighborhood region.

Several neighborhood regions exhibit large daily volatility of the trip imbalance and significant peaks at rush hour. The regions with positive peaks (more outgoing trips than incoming trips) in the morning, such as Upper West Side and Lower East Side, are likely to function primarily as residential areas. The regions with negative peaks in the morning (more incoming trips than outgoing trips), such as Lower Manhattan and Greenwich Village and Soho, are likely to function as business districts. The peaks indicate high volume of trips between the main residential and working areas at morning and afternoon rush hours. In the next section, we present models that predict the incoming trip numbers at afternoon rush hour to facilitate the daily estimation of bike demand in each region.

Models to Predict Trip Numbers at Afternoon Rush Hour

In order to build the models, we first looked for predictors that are highly correlated with the number of trips at afternoon rush hour. Based on the hypothesis that commuters act as the major user group at rush hour, we examined the relationship between the numbers of outgoing trips at morning rush hour and incoming trip numbers at afternoon rush hour of the same day. The result demonstrates a strong linear relationship between the morning and afternoon trip numbers, and the relationship is similar across months (**Figure 2**). Since our purpose is to predict future trip numbers, we excluded month from the models and trained models on one month to predict the outcome in the following month. By fitting a simple linear regression model using the number of outgoing trips at morning rush hour (Model 1 in **Table 2**), we obtained a cross-validation R^2 of 0.85. Including the number of incoming trips at morning rush hour as another predictor (Model 2) does not improve the testing R^2, although both incoming and outgoing trips at morning rush hour are significantly associated with the response variable Model 2 (p-value < 0.001 for both predictors). Residuals from both Model 1 and Model 2 are normally distributed and roughly equally spread around zero along the fitted values, indicating that the model assumptions are satisfied. Including additional predictors or interaction terms to the model and applying AIC-based model selection algorithms boosted the cross-validation R^2 to above 0.9 (Model 3-6 in **Table 2**). However, the residuals deviate from normal distributions in all these models.

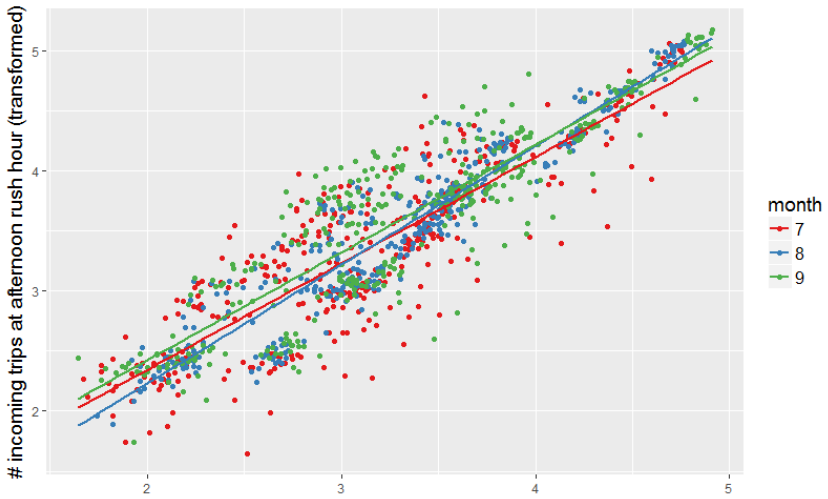


Figure 2. Relationship between trip numbers at morning and afternoon rush hours by month.
(The lines represent the linear regression best-fitting lines of the data from individual months.)

We also applied time series methods to predict trip numbers at afternoon rush hour for each

region. Since both the time series method and Model 1 include the same variables, the results of this two methods were compared. Although the time series method underperforms the simple linear regression model in most regions, we noticed three regions that the time series method generates significantly lower mean square of error (MSE): Central Brookline, Greenpoint, and Bushwick and Williamsburg (**Supplementary Table 1**). We discovered that the numbers of morning outgoing trips and afternoon incoming trips in these regions do not exhibit as strong linear relationship as in the other regions (**Figure 3**). This finding suggests that neighborhood regions and its interaction with the number of morning outgoing trips are important variables for predicting the afternoon trip numbers. By adding neighborhood regions and its interaction with the morning trip number into Model 1, we obtained the best model with out-of-sample prediction R^2 of 0.916 (Model 8). The assumptions of linear regression are generally satisfied in this model (**Supplementary Figure 4**).

Table 2. Summary of linear regression models

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Predictors	Morning outgoing trips	Morning outgoing + incoming trips	Morning outgoing × incoming trips	Forward selection	Backward selection	Stepwise selection	Morning outgoing trips + neighborhood	Morning outgoing trips × neighborhood
CV R^2(Aug)	0.886	0.881	0.876	0.873	0.883	0.875	0.889	0.926
CV R^2(Sep)	0.820	0.814	0.816	0.953	0.768	0.776	0.906	0.906
Average CV R^2	0.853	0.847	0.846	0.913	0.825	0.826	0.898	**0.916**

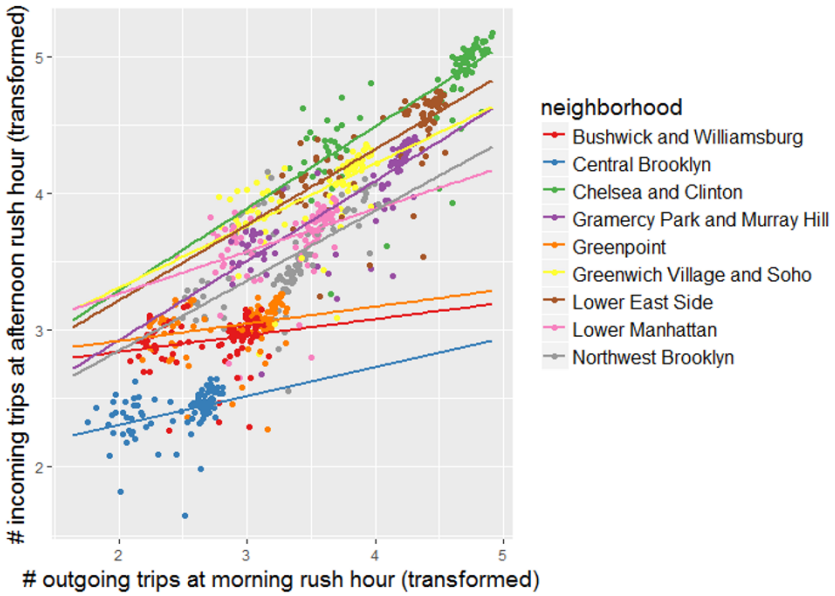


Figure 3. Relationship between trip numbers at morning and afternoon rush hours by region. (The lines represent the linear regression best-fitting lines of the data from individual regions.)

Discussion

Our analysis suggests that neighborhood region and time of the day are two important factors that influence the bike-share trip imbalance in New York City. To reach this conclusion, we handled the heteroschedasticity issue of the data by adjusting the weight of observations and applied the Extra-Sum-of-Squares test with correction for multiple comparison. The results indicates that the neighborhood regions exhibit different daily dynamics of the trip imbalance.

We visualized the trip imbalance by hour and neighborhood region to further investigate into the dynamic patterns. We noticed concentrated trip flows at morning and afternoon rush hour. Specifically, Greenwich Village and Soho, Lower Manhattan, Gramercy Park and Murray Hill regions exhibit heavy incoming trip volume at morning rush hour and outgoing trip volume at afternoon rush hour. These regions are consistent with the central business districts at Wall Street and Midtown Manhattan. The bike-share program provides a good choice for people to commute to work in these regions to avoid traffic jam. On the contrary, Upper West Side and Upper East Side are primary residential areas, which exhibit opposite trip imbalance patterns from the business regions. Further investigating into details of the patterns may assist the design of the bike-share system, such as optimizing routes and frequency of bike relocation and choosing the location of bike stations.

However, it is important to consider potential assumption violation of the tests. The trips from the same users over time are likely not independent. Future study may include new variables to control for additional confounding variables, such as weather of the day and traffic control. In addition, month effects should be incorporated due to the expansion of the bike-share program. In future studies, pairwise tests can be performed on regional groups of interest. Conducting more in-depth study such as investigating the association between starting and ending stations of trips at specific time period of a day would also be helpful for improving the design and efficiency of the bike-share system.

In the second half of the project, models were built to predict trip numbers at afternoon rush hour. The simple linear regression model with morning outgoing trips as the predictor is able to predict out-of-sample data with R^2 of 0.85. This result demonstrates that the model is able to explain 85% of the variability of the number of trips at afternoon rush hour solely using the number of trips at morning rush hour. It is consistent with the observation of symmetric and opposite trip imbalance at morning and afternoon rush hour from the previous analysis. The association is likely due to similar groups of users commuting to work and back home using the bike-share program. Adding neighborhood regions and interactions boosted the cross validation R^2 to above 0.9, and resulted in the best model in this study. The outgoing trips at afternoon rush hour can be determined using a similar model with morning incoming trips as a predictor. These two models will assist the bike-share program to predict the supply and demand in each region at afternoon rush hour with high accuracy. It is worth mentioning that this best model was not picked up by automatic model selection algorithms, which demonstrates that selecting predictors according to thorough data exploration or model comparison may perform better than automatic algorithms.

We also considered the time effect of the data, including the difference across months and potential dependency of data across days. By examining the data by month, we did not notice obvious temporal clustering. Although there is a slight interaction between month and the number of trips at morning rush hour, the slopes are largely similar across months. We excluded month from the models because the models were designed to predict outcome in future months. The results indicate that the models perform consistently across months. However, the models may be sensitive to future difference between months, and we suggest to regularly check and adjust the models in practice.

By comparing a time series method with the linear regression models, we noticed three regions with weak linear relationship between the numbers of morning outgoing trips and afternoon

incoming trips. The regions are Central Brooklyn, Greenpoint, and Bushwick and Williamsburg. Different from all the other regions that locate in Manhattan, these three regions locate in Brooklyn (map in **Supplementary Figure 5**). Brooklyn is connected with Manhattan through subways and bridges, which may prevent people who work in Manhattan and live in Brooklyn from using bikes as a main commute option. Instead, the bike-share program in Brooklyn may be primarily used for local trips within the area, and therefore does not exhibit similar rush hour patterns as the other regions do. This finding may help adjust the design of the bike-share program to accommodate the distinct needs in Brooklyn and Manhattan. Detailed difference between these two regions can be further investigated by tracking individual trips.

In summary, we identified that neighborhood region, hour-of-day and their interaction significantly influence the supply-demand imbalance of Citi-Bike at New York City. Inspired by the patterns of the trip imbalance, we constructed models to predict trip numbers at afternoon rush hour. Our results also revealed distinct functions of neighborhood regions and differentiated local needs of the bike-share program. These findings serve as the beginning of understanding the dynamics of bike-share programs in New York City. Further investigation may offer great help for better design and daily maintenance of the bike-share program.

References

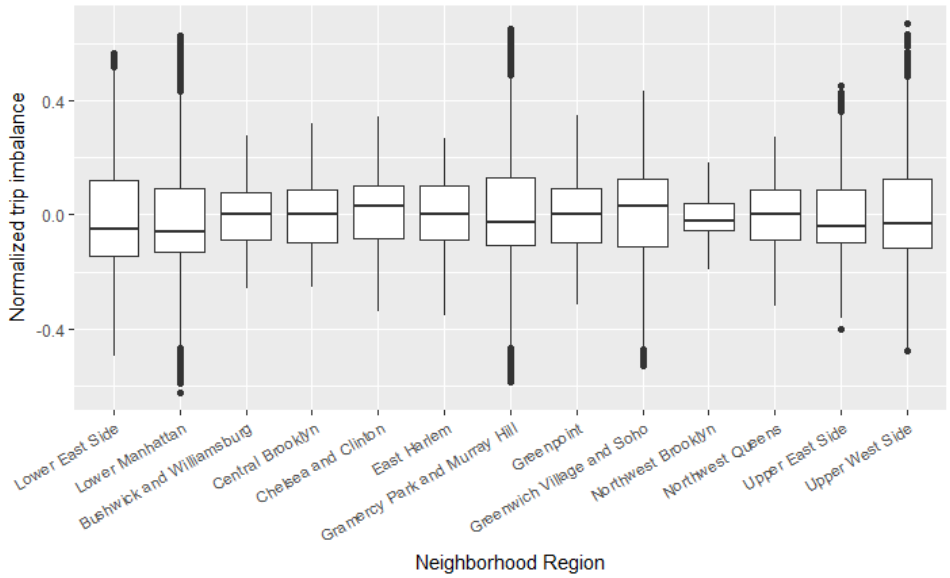
1. A. Singla et al. 2015. Incentivizing users for balancing bike sharing systems. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press 723-729.
2. E. O'Mahony and D. B. Shmoys. 2015. Data Analysis and Optimization for (Citi) Bike Sharing. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press 687-694.
3. Citi Bike System Data: https://s3.amazonaws.com/tripdata/index.html
4. Citi Bike Station Information: https://feeds.citibikenyc.com/stations/stations.json
5. R.H. Shumway and D. S. Stoffer. Time Series Analysis and its Applications. New York: Springer Press, 2011, pp. 84

Supplementary Material

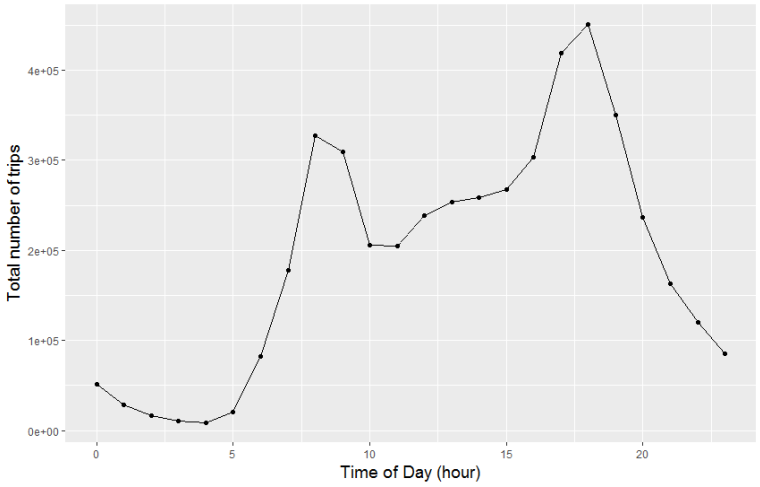
Time series method in predicting trip numbers at afternoon rush hour

In the time series method, the numbers of morning outgoing trips and afternoon incoming trips for each neighborhood region are ordered together by time, so the time series contains alternate numbers of morning outgoing trips and afternoon incoming trips. The resulted time series are stationary and can be fitted into ARIMA models. The models were selected using *auto.arima()* function from the R package "forecast". Since the prediction interval of ARIMA model increases each step forward, the model is trained by a moving window of 30-day historical data (including data from the past 30 days and the outgoing trip numbers at morning rush hour of the same day) and predict on the afternoon incoming trip number of each day in the test data.

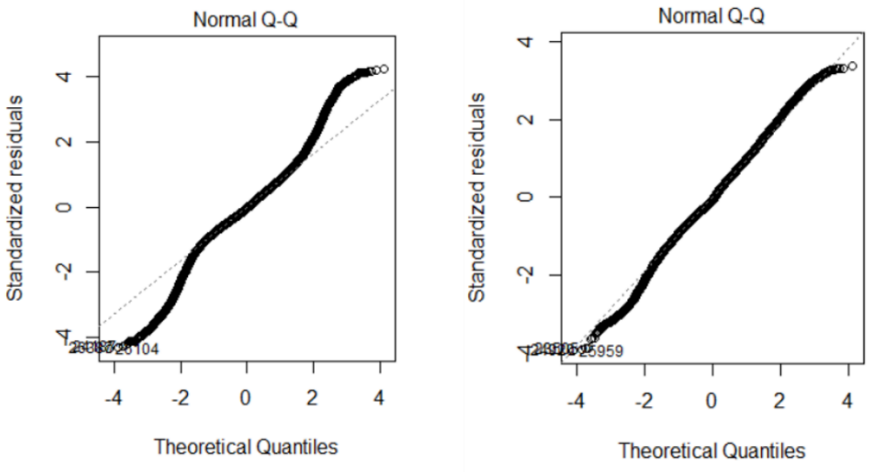
Supplementary Figure 1. Normalized trip imbalance by neighborhood region



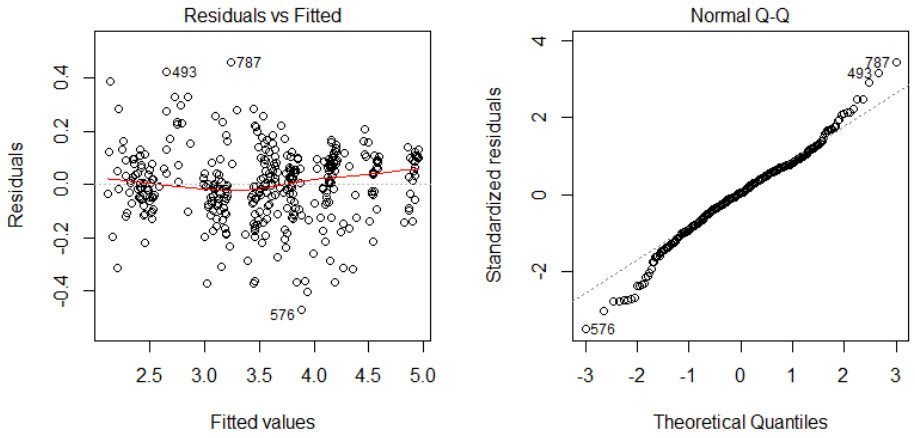
<u>Supplementary Figure 2</u>. Total number of trips grouped by hour-of-day in July –Sep 2016



<u>Supplementary Figure 3</u>. Representative Q-Q plot of residuals from unweighted (left) and weighted (right) linear regression. (Predictors in this model are neighborhood regions and hour-of-day.)



<u>Supplementary Figure 4</u>. Residual plots of Model 8



<u>Supplementary Figure 5</u>. Map of New York City

1 Lower Manhattan
2 Upper West Side
3 Greenwich Village and Soho
4 Northwest Brooklyn
5 Upper East Side
6 Chelsea and Clinton
7 Lower East Side
8 Gramercy Park and Murray Hill
9 Greenpoint
10 Bushwick and Williamsburg
11 Central Brooklyn
12 Northwest Queens
13 East Harlem

<u>Supplementary Table 1</u>. Comparison between time series method and Model 1

Neighborhood regions	**MSE (Time series)**	**MSE (Model 1)**	$(\frac{MSE_{Model1}-MSE_{ts}}{MSE_{ts}})$
Central Brooklyn	789	11597	1371%
Greenpoint	4995	16163	224%
Bushwick and Williamsburg	3435	10637	210%
Northwest Queens	559	374	-33%
Greenwich Village and Soho	179079	111036	-38%
Upper West Side	41368	24796	-40%
Lower Manhattan	48999	25841	-47%
Chelsea and Clinton	1656986	430457	-74%
Lower East Side	720043	132774	-82%
Upper East Side	61338	11020	-82%
Northwest Brooklyn	165899	16731	-90%
Gramercy Park and Murray Hill	382238	24263	-94%