How Demographic, Temporal and Weather Factors Impact Citi Bike Ridership

Su Park
April 2017

Abstract

Launched in 2013, Citi Bike is the first official public bicycle-sharing system in New York City. Citi Bike has been a success from the start, its annual subscribers surging past 20,000 in its first year and amounting to 60,000 trips per day in 2016. Citi Bike's considerably popularity has increased the interest in analyzing the sociodemographic, holiday and weather factors that affect bicycle-sharing flows and usage. Using publicly available daily and monthly trip data from July 2013 to December 2016, this study takes a regression-based approach in examining the influence of rider characteristics and meteorological data on bike usage, specifically trip distance per user and daily trip counts. The findings allow us to identify the demographics of a frequent Citi Bike rider and segment the potential customers.

1. Introduction

Rising concerns regarding the impact of climate change has brought various kinds of sustainable transportation into the spotlight. Bicycle sharing system, among many others, has been garnering tremendous attention with over 600 major cities globally implementing the program in the last few decades. Bike share's popularity explained by its three major advantages: first, its flexible mobility as a support for multimodal transport connections (Shaheen et al., 2013) to buses or cars during traffic congestion, second, its inexpensive costs and absence of maintenance responsibilities that owning personal vehicles entail, and third, its apparent health benefits.

In New York City, the first official bike sharing initiative, Citi Bike, was launched in May 2013, several years later than other major cities such as Boston or Washington D.C. Citi Bike is a privately-owned system with 10,000 bikes and 603 stations across Manhattan, Brooklyn, Queens and Jersey City. Citi Bike opened to public in May 2013 and by March 31, 2016, the total number of annual subscribers rose to 163,865 with an average of 38,491 rides per day in 2016. As the largest bike sharing system in the United States, Citi Bike is planned to expand to Manhattan neighborhoods up to 160th, Astoria, and inner neighborhoods of Brooklyn and add 12,000 bikes and 750 stations by the end of 2017.

What distinguishes Citi Bike from other bike sharing systems in other cities is its strict limit on trip duration, which helps the system effectively brand itself as a transportation option that is convenient for trips that are too far to walk, but too short for a taxi or the subway. Citi Bike customers are encouraged to use the system's mobile application to simplify and optimize their travel experience. Riders first choose from a selection of Single ride ($4/ride, up to 30 minutes), Day Pass ($12, unlimited 30-minute rides in a 24-hr-period), Annual Membership ($163/year, unlimited 45-miute rides), search the nearest docking station to unlock the bikes, then return the bikes after use. Studies have shown that no time limit for the use of bicycles resulted in excessively long rental periods (Shaheen et. al, 2013). In this sense, Citi Bike's decision to utilize transaction kiosks and mobile applications increases the turnover rate and may partially solve the perpetual rebalancing problem, where the customers can't find an empty station to return their bikes to, a challenge that may potentially be further heightened in the nation's most populated city.

In this study, using daily trip history data across all stations across the Greater New York City from July 2013 to December 2016, I conducted a regression-based study on Citi Bike's ridership in relevance to demographic, weather and holiday factors. The Citi Bike database includes trip data with bike numbers, trip start day & time, trip end day & time, trip start station, trip end stations, and rider types (subscriber or short-term-pass customer) with subscribers' trip data also including the rider's gender and year of birth.

2. Literature Review

Previous academic studies on bicycle sharing-flows and usage have been primarily limited to operations research and optimization models which are more relevant to the initial implementation stage. Dell'amico, et. al (2014) studied methods to optimize resource allocation and discover the best placement of bikes, while Schuijbroek et. al (2017) delved into the question of rebalancing the bikes over time such that the appropriate number of bikes and open docks are available to users. Therefore, Citi Bike in New York City is a comparatively mature program that provides a unique chance for interpreting the factors that influence its flows and usage.

Over the recent years there have been a few studies, although primarily written on European cities, that explored the temporal and sociodemographic factors that impact public bike sharing usage. A study on the Velib' bike share system in Paris, France indicated that ridership positively correlated with proximity to public transit stops (Nair et al., 2012). In addition, it was found that the distance to water, central business districts and parks have negative impact on daily trips, while the presence of food-related businesses has a positive effect on arrivals and trips in the Minneapolis–St. Paul metropolitan area in Minnesota (Wang et al., 2016). Another analysis showed that placing docking stations near neighborhoods with higher job densities results in higher usage of the bicycle sharing system (Wang et al., 2016).

In this study, I used applied linear regression methods to understand the determinants of trip distance and demand for Citi Bike. The objective of this study is similar to the aforementioned studies, yet is meaningful in its intention to focus the analysis on New York

City Citi Bikes, to apply a regression-based approach, and to conduct analysis on both monthly and daily levels. First, the relationship between weather, demographic and temporal factors and travel behavior in New York City may be significantly different from other cities, since New York City has comparatively higher job densities and convenient facilities across neighborhoods and the public transportation system is a preferred method of transportation than cars. In addition, as Citi Bike is about to celebrate its 5th anniversary and pass into the "maturity" stage, an analysis on Citi Bike can provide informative insight in shaping the bike-sharing-systems' future expansions past their implementation stage.

3. Data and Regression Analysis

I decided to run two different regressions, the first regression examining the association of demographic and weather factors with a rider's trip distance, and the second regression investigating the association of weather and temporal factors with daily trip counts.

3.1 Regression I

3.1.1 Regression objective

Are the riders' demographic factors and monthly weather factors associated with their trip distance?

$$\textit{TRIP DISTANCE} = \beta_0 + \beta_1 \textit{AGE} + \beta_2 \textit{GENDER} + \beta_3 \textit{NEIGHBORHOOD} + \beta_4 \textit{USERTYPE} + \beta_5 \textit{PRECIPITATION} + \beta_6 \textit{SNOW DEPTH} + \beta_7 \textit{TEMPERATURE} + \beta_8 \textit{WIND SPEED}$$

The proposed hypothesis is that neighborhood, user type and weather factors except temperature are negatively associated with trip distance, while gender and temperature are positively associated with trip distance. The most commonly believed Citi Bike riders are daily commuters (who are also subscribers) traveling relative short distance to Midtown. In a given time male riders would travel faster and farther than their female riders, while inclement weather conditions will naturally deter people from riding bicycles.

3.1.2 Data

I collected 100 samples from each month from 2013 July to 2016 December and obtained the 1) demographic characteristics including age, gender, user type (subscriber vs customer), destination neighborhood and the 2) weather factors including precipitation, snow depth, temperature and wind speed calculated on average for each month. Weather data had to be collected on a monthly unit because the specific date of each trip was specified only up to November 2015. Out of 4200 samples, 1086 samples with NA for at least one of the factors were deleted to obtain accurate regression results.

Visualized summaries of age, gender, user type and neighborhood are included in the Appendix section after References. The following descriptive summary reveals that male riders are three times more common than female riders. The top six destination neighborhoods were chosen according to frequency, all of which belonged to the Midtown and Midtown South area which includes Times Square, Penn Station, Union Square area. Since Midtown is a neighborhood within Manhattan with one of the high job densities (or business districts) and major transit stops, I posited it meaningful to construct a dummy variable for Midtown based on these top most popular stops. Short-pass customers only constitute around 1% of the riders.

Table 1. Descriptive Summary of Sample Characteristics (I)

Variable	Mean	Std. Dev.	Min	Max	Definition and Unit
TRIP DISTANCE	1712.1	521	0.22	9407.2	Trip distance
AGE	40.67	6	17	118	Self-identified age of each rider
GENDER	0: 704	1: 2410	Self-identified age of each rider. (0: female, 1: male)		
NEIGBORHOOD	0: 2936	1: 178	Destination neighborhood (0: if not top 6 stops, 1: top 6 stops)		
USER TYPE	0: 31	1: 3083	0: customer, 1: subscriber		
PRECIPITATION	0.01	0.05	0.132	0.26	Monthly average, in
SNOW DEPTH	0.9	0.1	0	10	Monthly average, in
TEMPERATURE	55.19	15.8	23.9	79.8	Monthly average, °F
WIND SPEED	5.45	1	4	7	Monthly average, mph

3.1.3 Regression Results

In regression (1), age and temperature are positively associated with trip distance, while gender, neighborhood, user type, precipitation, snow depth and wind speed are negatively associated with trip distance. The signs of all factors except gender are as predicted. Since all of the variables except user type are statistically insignificant, I will refrain from making statistical interpretation on how much a unit change in an independent variable will impact the trip distance.

Nevertheless, it's interesting to notice that the coefficient estimate for gender is negative, indicating that female riders travel farther than male riders. This contradicts my initial hypothesis that male drivers will travel farther within a limited time frame. Perhaps this hints that a significant portion of female riders are short-pass customers, instead of commuters or subscribers, who use bikes for leisurely activities and travel farther distance on average. It's also interesting to notice that age is also positively associated with trip distance.

Snow depth and wind speed are statistically significant in regression (6), where AIC was minimized, and regression (4). However, the adjusted r-squared values for regressions (1) through (6) are extremely low, which means that a minimal portion of change in trip distance can be explained by its linear association with the given independent variables. This led me to think that a more meticulously designed regression with an accurate, daily recorded weather data will more accurately explain the weather's relationship with ridership. Hence I conducted another regression that solely uses a daily average measure of data and sets daily trip counts as the dependent variable.

Table 2. Regression Results (I)

Trip distance	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Intercept	3615.433 *** (931.964)	3682.466 *** (3.05e-05)	1717.87 *** (22.97)	1813.59 *** (55.28)	1562. 901 *** (277.118)	1938.413 *** (133.181)
AGE	2.877 (1.924)	2.520 (1.9224)	-	-	-	2.701 (1.922)
GENDER	-3.315 (53.438)	-12.651 (53.406)	-	-	-	-
NEIGHBORHOOD	-93.183 (96.289)	-	-100.39 (96.09)	-	-	-
USER TYPE	-2019.456 * (879.357)	-2064.410 * (880.497)	-	-	-	-
PRECIPITATION	-588.836 (386.677)	-	-	-615.01 (383.87)	-	-607.768 (384.756)
SNOW DEPTH	-9.338 (12.024)	-	-	-22.07 * (9.80)	-	-
TEMPERATURE	2.594 (2.506)	-	-	-	3.562 (2.171)	-
WIND SPEED	-9.349 (32.196)	-	-	-	-8.691 (31.397)	-46.884 * (19.22)
P-value	0.00075	0.0688	0.2962	0.013	0.0086	0.010
Adjusted R-squared	0.006301	0.001318	2.938e-05	0.0021	0.0024	0.002641

Note: Heteroskedastic-consistent standard errors in parentheses. *** significant at the 1% level, ** significant at the 5% level, * significant at the 10% level

3.2 Regression II

3.2.1 Regression Objective

Are weather and holiday factors associated with the daily number of people using Citi Bike?

$$\boldsymbol{TRIPS} = \beta_0 + \beta_1 \boldsymbol{PRECIPITATION} + \beta_2 \boldsymbol{SNOW\ DEPTH} + \beta_3 \boldsymbol{MAXTEMPERATURE} + \beta_4 \boldsymbol{MINTEMPERATURE} + \beta_5 \boldsymbol{WIND\ SPEED} + \beta_6 \boldsymbol{HOLIDAY} + \beta_7 \boldsymbol{WEEKDAY}$$

The proposed hypothesis is that weather factors except temperature are negatively associated with the number of daily trips. I would posit that both holiday and weekday are positively associated with trips, since this would attract both tourists and daily commuters to ride Citi Bikes.

3.2.2 Data

In Regression II, I decided to use daily average measurements of the weather variables and included the holiday and weekday variables to see how those daily factors may impact trip usage. It must be noted again that this regression only covers trip data from July 2013 to November 2015, since date column was eliminated from December 2015 and all simply categorized by month.

Table 3. Descriptive Summary of Sample Characteristics (II)

4.

Variable	Mean=	Std. Dev.	Min	Max	Definition and Unit
TRIPS	25170	12589	876	14670	Count of daily trips
PRECIPTIATION	0.122	0.039	0	4.969	Daily, in
SNOW DEPTH	1.008	0.004	0	18.90	Daily, in
MAX TEMPERATURE	63.52	17.48	17.24	96.98	Daily maximum, °F
MIN TEMPERATURE	48.91	16.02	2.12	82.04	Daily minimum, °F
WIND SPEED	5.254	1.12	0.671	14.54	Daily average, mph

HOLIDAY	826	23	Dummy variable (0: non-holiday, 1:holiday)
WEEKDAY	241	608	Dummy variable (0: weekend, 1: weekday)

3.2.3 Regression Results

In Regression II (1), the adjusted r-squared value (0.78) has dramatically increased from regression I (1) (0.0063) which had included the rider's demographic characteristics along with monthly average weather data. This means that 78% of the variation in daily trip counts can be explained by temporal and weather variables. This indicates either one or more of the following three mistakes in constructing Regression I: (1) demographic variables have been interfering with the regression in reflecting weather's impact on trip distance, (2) trip distance itself is less strongly associated with weather than daily trip counts, hence an unsuitable dependent variable, and (3) monthly average measurement of weather data cannot accurately reflect the weather's impact on trip usage.

All of the independent variables except minimum temperature are significant at the 1% level, which hints that trip counts are more strongly linearly associated with highest temperature of the day than the lowest temperature. Perhaps most riders tend to bike in the late morning or during day time when the temperature is higher, rather than in early morning or at night when the temperature drops. In this respect, temperature may even have a polynomial or logarithmic relationship with trip counts, since it is natural that people will avoid riding bicycles outside when it gets hotter beyond a certain temperature. As in Regression I, all weather variables except temperature variables are negatively associated with trip counts, with an unit increase in precipitation, snow depth and wind speed respectively dropping the daily trip counts by 8354.73, 522.16 and 372.

Holiday is negatively associated with trip distance, rejecting my initial hypothesis, while weekday is positively associated as predicted. This indicates that because majority of the riders are subscribers who travel short distances for errands or work, they travel less on holidays as opposed to weekdays or non-holidays.

Regressions (2), (3) and (4) show that adjusted r-squared is maximized when only temperature variables are included, indicating that the impact on trip counts is temperature > precipitation/wind speed > holiday/weekday. Overall, weather seems to have more impact on trip counts than temporal factors. This is in line with Gebhart et. al (2014)'s findings which showed a significant reduction in ridership with cold temp, rain, high humidity in Washington D.C., U.S.A.

Table 4. Regression Results (II)

Trips	(1)	(2)	(3)	(4)
Intercept	-1777.96 (1328.52)	38312.1 *** (780.9)	-5950.12 *** (979)	21438.6 *** (744.5)
PRECIPTIATION	-8354.73 *** (549.58)	-6840.4 *** (874.0)	-	-
SNOW DEPTH	-522.16 *** (71.85)	-1552.1 *** (102.2)	-	-
MAX TEMPERATURE	337.49 *** (40.21)	-	414.71 *** (50.76)	-
MIN TEMPERATURE	101.69 * (42.86)	-	97.67 (54.16)	-
WIND SPEED	-372.00 *** (102.73)	-2043.4 *** (140.9)	-	-
HOLIDAY	-10890.70*** (1198.96)	-	-	-13875.1 *** (2456.9)
WEEKDAY	6057.25 *** (430.04)	-	-	5735.7 *** (884.7)
P-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	1.008e-14
Adjusted R-squared	0.7809	0.4387	0.6401	0.07117

Note: Heteroskedastic-consistent standard errors in parentheses. *** significant at the 1% level, ** significant at the 5% level, * significant at the 10% level

4. Limitations

Some of the results are limited due to a lack of availability for hourly and daily ridership. The impact of weather may be more accurately measured by factoring in the weather conditions in the previous three hours or in the morning. If it rains in the morning, it is less likely for a person to decide to bike to a quick errand or work. Fishman et al. (2015) concludes that the evaluation of current performance is crucial for better understanding the effectiveness of bicycle-sharing programs, so Citi Bike should resume recording hourly and daily date if it intends to thoroughly analyze its ridership and accurately predict future bike traffic.

Analyzing ridership by both origin and destination will lend a more fascinating insight into the routes riders frequent and how the pre-existing bike path map can be improved in the future. A more informative neighborhood-level analysis would have been possible if the categorical variable had been constructed to correspond to each of the New York City neighborhoods. This would allow a sociodemographic analysis regarding factors such as median income, education and non-white population and their relationship to trip usage. A station-level analysis based on a docking station's proximity to bike paths or lanes will further shed light on ensuring rider safety and optimizing the location of stations.

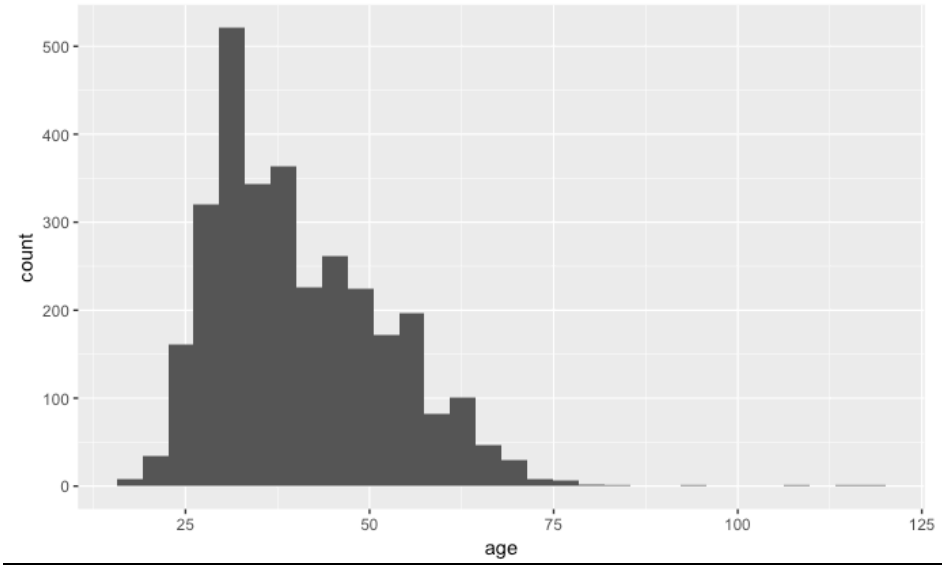
Other weather factors like humidity and fog may have linear association with trip usage, especially for high job density areas with heavy traffic where bikers may be more concerned about safety and have other safer travel alternatives. Since some of the categorical variables were left unknown, starting with eliminating those columns before sampling may have resulted in a slightly more reliable analysis. A polynomial or a log regression may have been more effective in explaining temperature variables, as mentioned in the previous regression analysis.

5. Summary

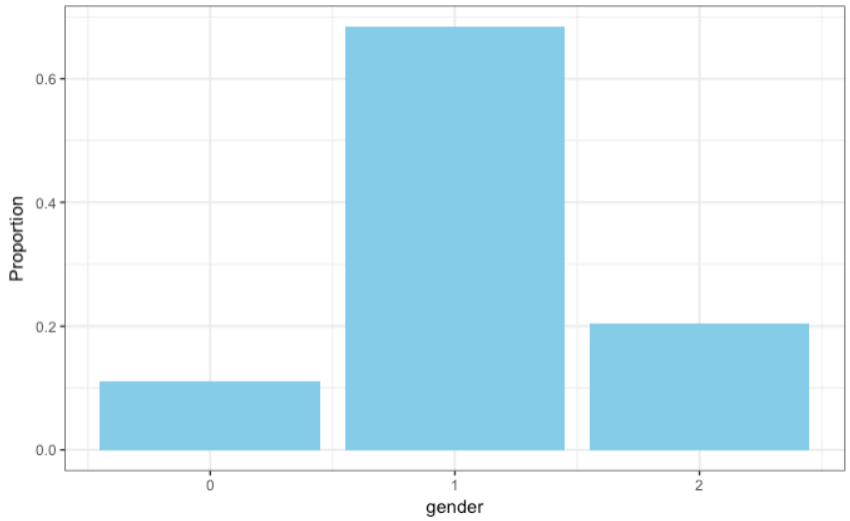
Now in its fifth year since launching, Citi Bike has been a successful a support for multimodal transport connections and a complement to the subway and bus system. A linear regression model (Regression II) helps explain the daily trips' relationship to changes in weather, holiday or weekday factors and the rider's demographic factors. The program is currently most frequented by male subscribers who travel short distances to neighborhoods concentrated in Midtown, so an active marketing approach on Citi Bike's part to expand its target audience is crucial for the system's successful city-wide expansion.

6. Appendix

Graph 1. Age distribution



Graph 2. Gender distribution(0: unknown, 1: male, 2: female)



Graph 3. User type distribution

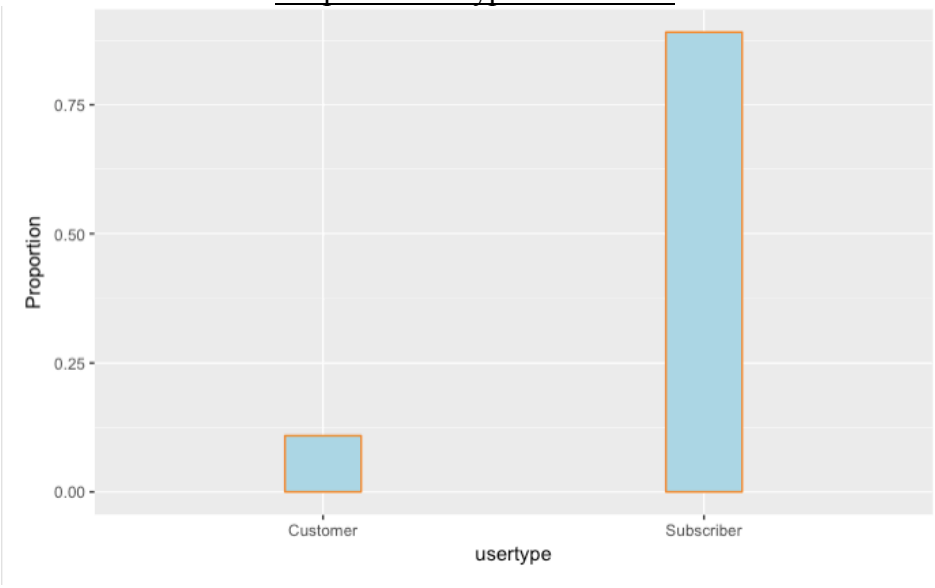


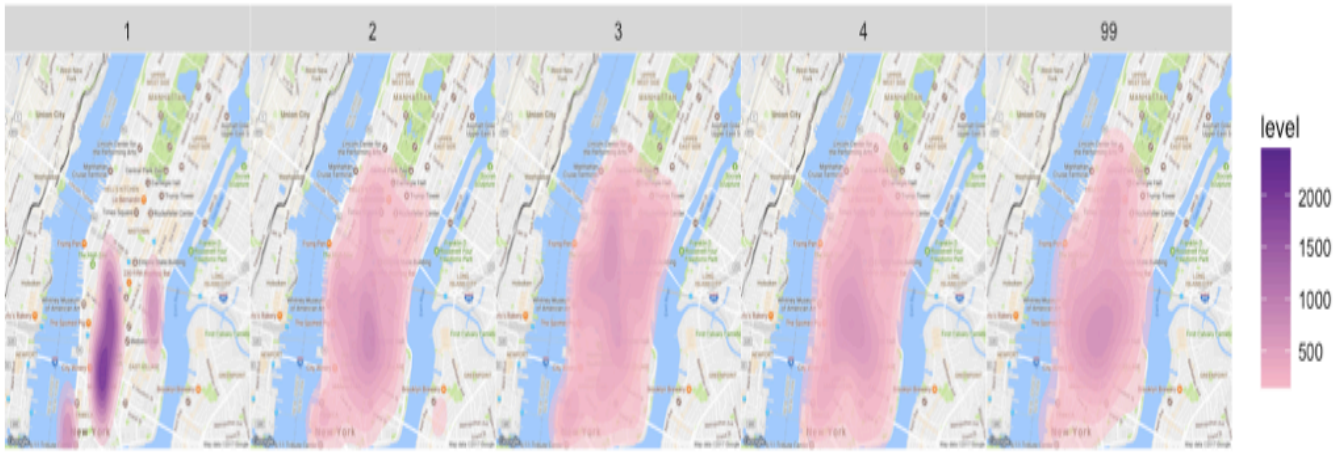
Table 5. Destination Neighborhood

Top 10 Stations: Trip Ends	
Station	Neighborhood
W 21 St & 6 Ave	Midtown South
E 17 St & Broadway	Union Square
9 Ave & W 22 St	Chelsea
8 Ave & W 31 St	Penn Station
Lafayette St & E 8 St	East Village
W 41 St & 8 Ave	Times Square
Broadway & E 14 St	Union Square
E 32 St & Park Ave	Midtown South
W 31 St & 7 Ave	Penn Station
W 25 St & 6 Ave	Midtown South

Table 6. Selected Midtown Stations

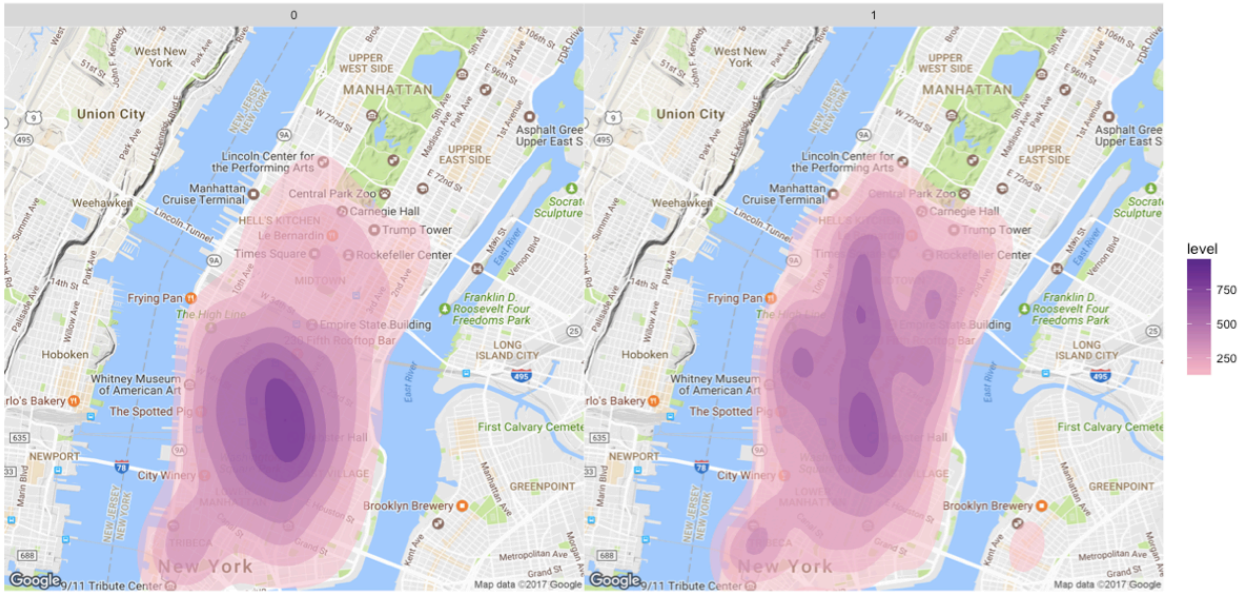
18 Midtown docking stations
W 59 St & 10 Ave
Broadway & W 53 St
Broadway & W 49 St
Broadway & W 41 St
W 44 St & 5 Ave
E 43 St & Vanderbilt Ave
Broadway & W 36 St
5 Ave & E 29 St
W 41 St & 8 Ave
W 33 St & 7 Ave
E 33 St & 5 Ave
W 54 St & 9 Ave
W 53 St & 10 Ave
W 49 St & 8 Ave
W 45 St & 8 Ave
W 43 St & 6 Ave
W 47 St & 10 Ave
W 37 St & 10 Ave

Graph 4. Popular neighborhood by age (1: 0~20, 2: 21-40, 3: 41~60, 4: 60~, 5: unknown)



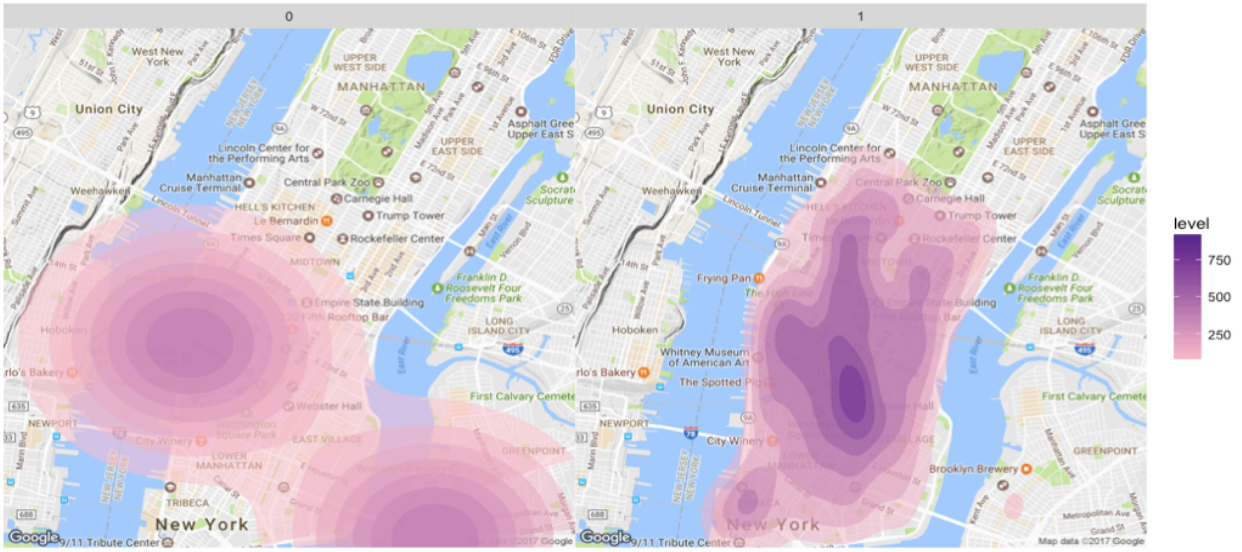
Riders below 20 tend to be limited to the bike lanes along the Hundson River. Riders from age 20 to 40 and over 60 are slightly clustered around midtown south area, while 40~60 show significant ridership in Hell's Kitchen and Upper West Side.

Graph 5. Popular neighborhood by gender (0: female, 1: male)



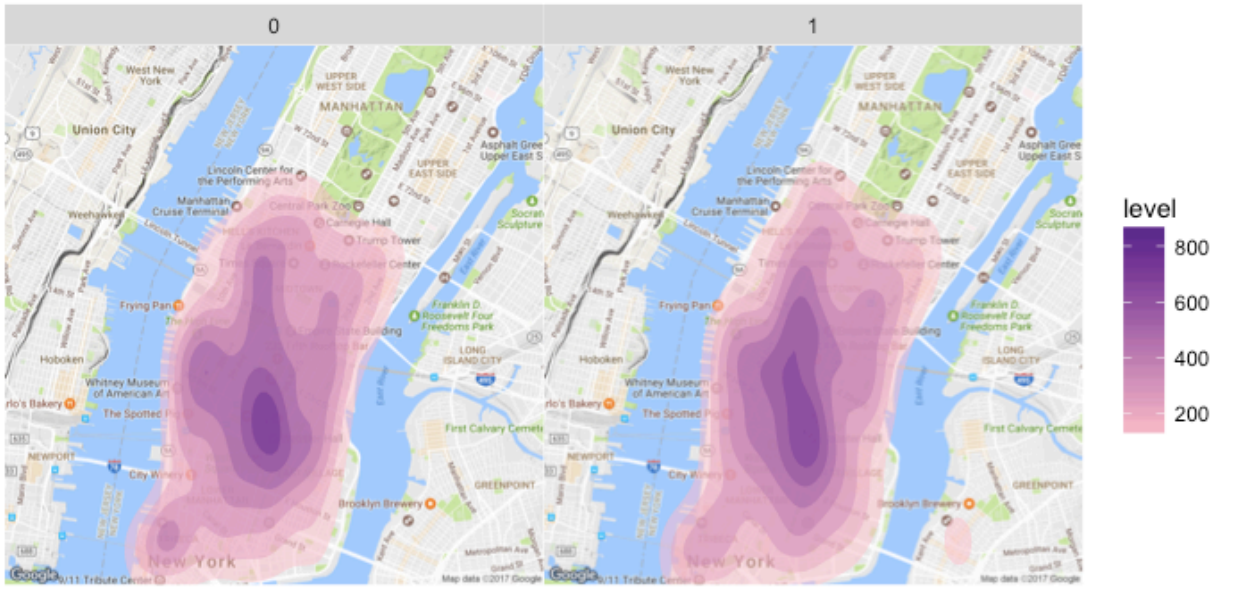
Male riders travel further distance, less concentrated around midtown. Female riders' trips are centered around the Union Square area while male riders' trips are clustered around Union Square, Midtown, Chelsea, Midtown East and Times Square.

Graph 6. Popular neighborhood by user type (0: customer, 1: subscriber)



Subscribers' trips are more concentrated around the Midtown area.

Graph 7. Popular neighborhood by season (0: warmer (March~October), 1: colder (November~February)



There's no significant difference in distribution between colder and warmer seasons, although colder season seems slightly less concentrated in Midtown.

7. References

Dell'amico, M., Hadjicostantinou, E., Iori, M., & Novellani, S. (2014). The bike sharing rebalancing problem: Mathematical formulations and benchmark instances. Omega, 45, 7-19.

Fishman, E., Washington, S., Haworth, N., & Watson, A. (2015). Factors influencing bike share membership: An analysis of Melbourne and Brisbane. Transportation Research Part A: Policy and Practice, 71, 17-30.

Gebhart, K., & Noland, R. B. (2014). The impact of weather conditions on bikeshare trips in Washington, DC. Transportation, 41(6), 1205-1225.

Nair, R., Miller-Hooks, E., Hampshire, R. C., & Bušić, A. (2012). Large-Scale Vehicle Sharing Systems: Analysis of Vélib' International Journal of Sustainable Transportation, 7(1), 85-106.

Schuijbroek, J., Hampshire, R., & Hoeve, W. V. (2017). Inventory rebalancing and vehicle routing in bike sharing systems. European Journal of Operational Research, 257(3), 992-1004.

Shaheen, S., Martin, E., & Cohen, A. (2013). Public Bikesharing and Modal Shift Behavior: A Comparative Study of Early Bikesharing Systems in North America. International Journal of Transportation,1(1), 35-54.

Wang, X., Lindsey, G., Schoner, J. E., & Harrison, A. (2016). Modeling Bike Share Station Activity: Effects of Nearby Businesses and Jobs on Trips to and from Stations. Journal of Urban Planning and Development, 142(1), 04015001.