How Citi Bike data can revolutionize the future of transportation in New York City

04/15/2023

Kevin Kuc, Nathan Arias, Randy Louie

Overview

Citi Bike, launched in 2013, is a popular bike-sharing system (Ley 2021) in New York City that has become a go-to choice for many commuters and visitors alike. The Citi Bike program was created in response to the growing need for alternative transportation solutions in New York City, which has experienced a lot of traffic congestion and air pollution (Lohr 2019). The program features a physical network of bicycles and docking stations that provide customers with an accessible and affordable transportation option. With Citi Bike, customers can unlock a bike at one docking station and return it at another location, making it a flexible choice for commuting, running errands, or exploring the city (Sisson 2018).

Since the service launched, these bikes have gained popularity and have become the primary bike rental company in New York City (Citi Bike). They offer customers the freedom to do a variety of activities, from leisure strolls to commuting to work, at any time of day or night, 365 days a year (Sisson 2018). Citi Bike has become an integral part of New York City's transportation infrastructure, with thousands of bikes and hundreds of docking stations located throughout the city (Lovelace Jr. 2018).

The program has also expanded to include e-bikes, which offer electric pedal assistance and make it easier to ride uphill or over long distances. Overall, Citi Bike has been a huge success and has helped to promote sustainable transportation and reduce traffic congestion in New York City.

Question

The primary objective of this report is to determine how to expand the network of Citi Bike stations in the future. In doing so, we will need to identify who the current user base is, when and why they use Citi Bike. This will allow us to determine how we can employ the Citi Bike data to enhance ridership in New York City. By breaking down this main question into smaller components, we can explore sub-questions that will deepen our comprehension of the primary issue. Our sub-questions are located in the Compelling Text and Data Stories section of our report.

Data Scraping

The data on the Citi Bike website offers valuable insights into various aspects of a Citi Bike ride. Although the data covers a range of cities, we chose to focus on New York City as it is where the Citi Bike franchise began. Our group's objective was to analyze Citi Bike data from 2013 to 2022. We used Selenium, an automation framework through python that allows for dynamically interacting with web browsers. Using this system we scraped 120 .csv.zip files provided by the NYC Citibike data landing page. In totality, this scraping amounted to 200 million rows of data, one for each instance of usage of Citi Bike. We determined this amount of data was unwieldy for personal computers, so we decided to conduct simple random sampling to create a subset dataset. We decided to take 5000 random rows from each month over the 10 year period. This amounted to about 600,000 rows of data.

It is important to note that Citi Bike excluded certain data from their dataset, including trips taken by staff that service and inspect the systems, trips from "test" stations that occurred mostly in June and July 2013, and trips that lasted less than 60 seconds, as it was uncertain if they were false starts or user errors. Citi Bike believed that this data had the potential to compromise the ability to conduct data analysis.

Data cleaning

The NYC Citi Bike data spanned 10 years since the initial implementation of the system. Over that time, they changed collection software and modified the format of the output data to fit their needs at the time.

- **Problems**: Multiple duplicate columns, inconsistent column naming, columns conveying same information but using different terminology, and data errors.
- **Added and removed columns:** Citi Bike expanded their fleet in 2021 by introducing new bicycle types such as “electric” and “dock anywhere”, which were tracked using a new “bike_type” column. Additionally, they removed certain columns, such as gender and age, from data sets starting in 2016 due to privacy concerns and data regulation changes. As a result, these columns have a large amount of missing data, as not all time periods reported on this information. We could not find data to supplement the missing rows so the remaining data can only be used to analyze those sections of time.

- **Changes in terminology:** In 2017, Citi Bike changed the terminology to describe a user who pays a monthly membership fee to get reduced fees. Before 2017, they were called a ‘Subscriber’ (monthly membership) or ‘Customer’ (no membership). After 2017, they are called a ‘Member’ (monthly membership) or ‘Casual’ (no membership). To solve this issue, we have updated all occurrences of the previous terminology to align with the current terminology.
- **Inconsistent column naming:** Over the years, Citi Bike’s data set changed the naming of columns several times but left the column data the same. For example, there are columns named ‘started_at’, ‘start_time’, and ‘starttime’ that all contain datetime data of when the user began their trip. To solve this problem, we changed the column names to match and merged them together using groupby. Initially, the data set had 46 columns and we reduced it to 19 through merging. Overall, no data was lost due to the removal of columns, as the data was consolidated into other columns.
- **Data errors:** We found errors in the datetime data for start time and stop time columns. There were instances where the datetime data was truncated and only showed minutes and seconds. Datetime data should have the format ‘Month/Day/Year Hour:Minutes:Seconds’ An example error was the values showing “48:14.7” representing ‘Minutes:seconds’. Due to the missing month, day, year, and hours we could not recover this data. We decided to make these all NaN. Another data error we found was that March 29th 2023 has an abnormally high amount of entries compared to the other days.
- **Duplicates:** There were several duplicates and empty rows that were sampled. We dropped these rows from the data set.

After performing data cleaning, the resulting dataset had a shape of (597,388 rows and 18 columns), which was different from the original dataset's shape of (600,000 rows and 46 columns).

Variables and Definitions

Below are the variables along with their definitions that we utilized throughout this project. As stated above, these variables have changed throughout the years so the below variables correspond to the column names listed in our dataset.

- **Bike ID:** A unique identifier for each bicycle in the fleet of Citi Bikes.
- **Birth Year:** The birth year of the bicycle rider.
- **Start station name:** The name of the station the bicycle got checked into after the trip was completed.
- **Start station ID:** A unique identifier that is tied to each Start station.
- **Start station latitude:** The latitude of the starting station.
- **Start station longitude:** The longitude of the starting station.
- **End station name**: The name of the station the bicycle got dropped off at after the trip was completed.
- **End station ID**: The unique identifier that is tied to each End station.
- **End station latitude:** The latitude of the starting station.
- **End station longitude:** The longitude of the ending station.
- **Start time**: A datetime value when the bike was checked out.
- **Stop time**: A datetime value when the bike was checked in.
- **Trip Duration**: The time spent between checking out the bike and returning the bike in seconds.
- **Rideable type:** A categorical variable that consists of classic bike, electric bike, docked bike, and 0 if the type is unknown.

- **User Type:** There are two categories in the categorical variable for Citi Bike users: "customer" for one-time users and "subscriber" for individuals who have subscribed to the service. Additionally, there is a third category labeled as "0" for cases where the user type is unknown.
- **Gender:** A value of ‘1’ represents male, ‘2’ represents female, ‘0’ represents did not specify.
- **Ride ID:** A unique identifier for this ride instance.

Compelling Text and Data Stories

Question 1: Who typically uses Citi Bikes?

The Citi Bike data set provided columns for user birth year and gender from the years 2013-2021. In the year 2022, age and gender columns were omitted. Using this data we were able to create Figure 1, a histogram of ages that is separated by gender. As a group, we decided to drop the ‘unspecified’ gender data from this graph because 86% of the ‘unspecified’ gender were ages 49-51. This seems improbable and most likely means that the user did not fill in either gender or age when they were signing up for the Citi Bike app. Analyzing the remaining data, we found the average age of a Citi Bike user is 38 years old and there are 3.25 times more men than women users. The age representing the highest percentage of users is 30 years old. We found that there was no major difference in the ratio of male/females using Citi Bikes over the data’s 2013-2021 time frame. We believe there is an opportunity to increase marketing targeted towards potential female users.

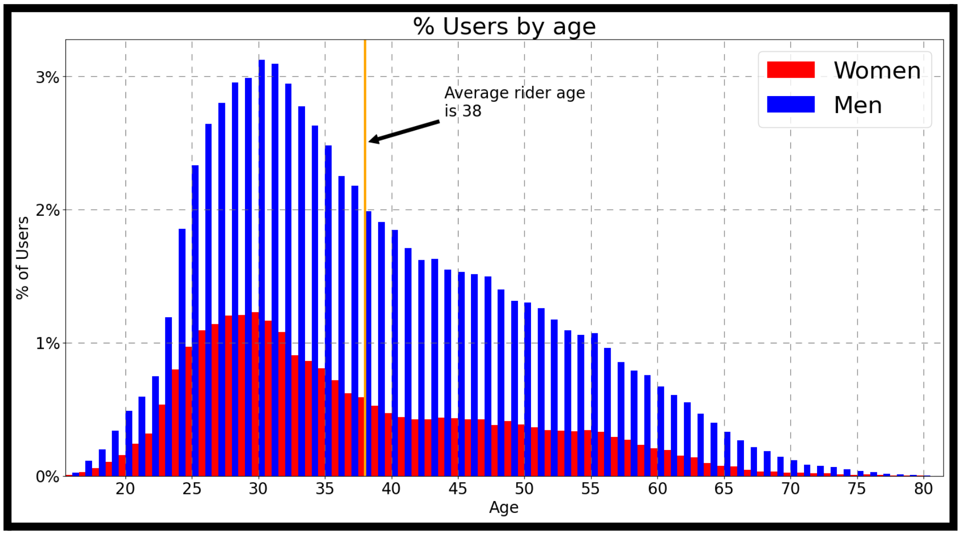


Figure 1: Age and gender demographic

Question 2: When and why do people typically use Citi Bike?

Using datetime data of start and end times of rides, we were able to compile a graph (Figure 2) of when people typically begin and end their rides with Citi Bike. The datetime data is aggregated into 15 minute intervals, displayed as a percentage of total rides, and graphed against a 24 hour x-axis. It is also separated between weekday and weekend rides.

The weekday graph produced two distinct peaks at 7:15am-10am and 4:30pm-8pm. These align with the usual timeframes for morning and after-work commutes. These two peaks represent 40% of total usage during the week. There is a dip between 10am-4:30pm that correlates with the times that people spend at work. This means that a large portion of Citi Bike usage is to commute to and from work. Since the average age of a user is 38 years old which is a typical working age of an American, it makes sense that most users are using the system to commute to and from work on weekdays.

The weekend graph was very different than the week day graph. There is a much shallower slope where peak usage is between 10am and 7pm. There is also less usage on weekends vs weekdays. Weekdays represent about 20% of usage while weekdays accounts for 80% of usage.

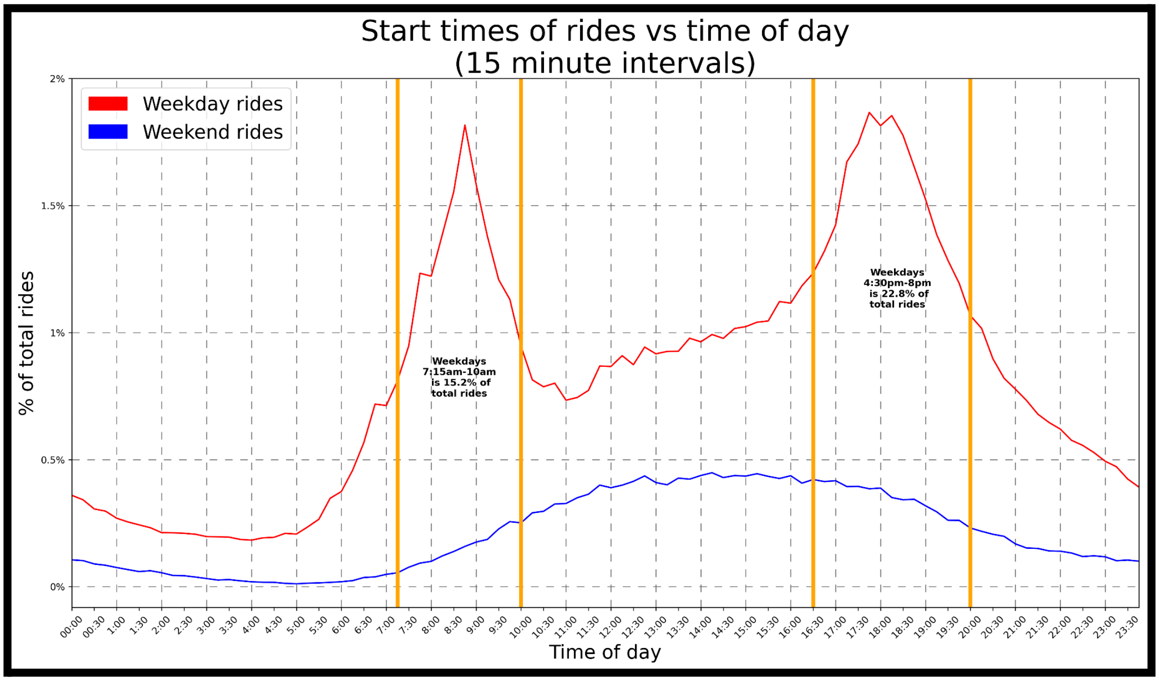


Figure 2: When do users use Citi Bike

Question 3: Where is Citi Bike currently located?

Citi Bikes bikes when not being used are stored at outdoor stations along the street where they are locked. A user can unlock one by paying a fee at the starting station and then returning the bicycle to an ending station. Since the start and ending locations are fixed, the dataset included trip start and ending information. This had station names, station longitude, and station latitude. We were able to plot this data onto a map of the New York City and New Jersey area to find where Citi Bike stations were located.

In 2013 the system launched with 350 stations in lower Manhattan and Northern Brooklyn. In 2022, there were about 1800 stations. It has averaged a 20% increase in stations every year. You can find a Citi Bike station in every zipcode of Manhattan and Harlem. It also has stations in some zipcodes of Bronx, Hoboken, Jersey City, and Brooklyn.

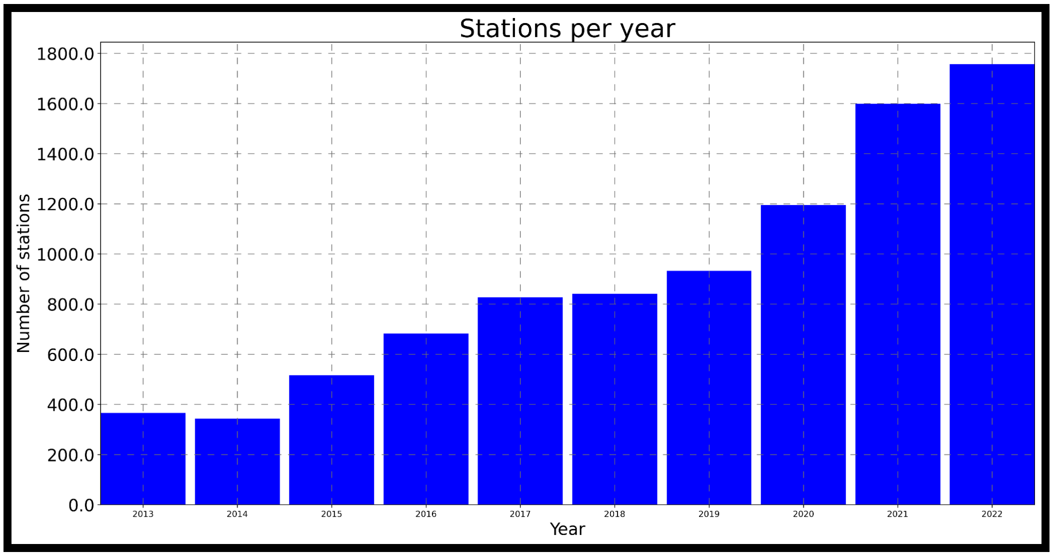
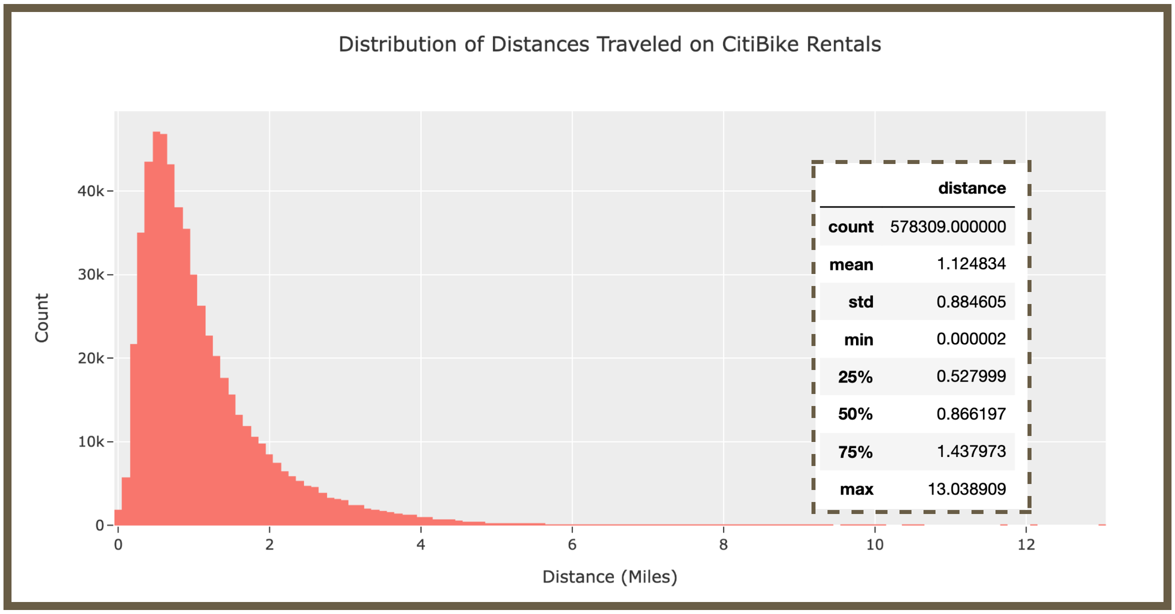


Figure 3: Citi Bike station count per year

Question 4: Distance graph

Figure 4 shows a histogram of ride distances of users. This is a measurement of the Haversine distance of between the start and ending stations using start and stop longitude and latitude data. Haversine distance is slightly different from Euclidean as it measures angular distance between points on a great sphere. It is important to recognize that Haversine distance is not representative of the actual mileage ridden on the bicycle as a cyclist must abide by road rules which typically do not follow straight line paths. One additional consideration was that given the methodology of the distance formula, a cyclist that starts and ends a trip at the same location will have a ride distance of 0 which, after consideration, filtered out for this figure.
There were entries in the data set where the start and ending location were the same. This meant that the user began and ended their trip at the same station. This data was filtered out of this data set. We have no information on how far these users took the bikes.

With these considerations in mind, the average distance traveled by users was 1.12 miles. With this statistic we determined that CitiBike rentals are often a used to supplement public transportation when the last leg of a trip is often too far to walk but easy to bike.



	distance
count	578309.000000
mean	1.124834
std	0.884605
min	0.000002
25%	0.527999
50%	0.866197
75%	1.437973
max	13.038909

Figure 4: Distance of rides

Question 5: What valuable insight can be gained from creating a user trip location density heat map?

In order to examine the major locational hubs for CitiBike users, we opted to generate a heat map of where users end their trips during the previously discussed peak hours of (7am-10am) and (4pm-8pm). Our intuition was that if we analyzed where people ended their trips in the morning, we could see where our users worked and if we looked at where users ended trips in the evening, we would see approximately where our users reside. Our analysis revealed that the majority of bike trips during the evening ended in the major metropolitan zones of New York City. Furthermore, a major trend was observed in the increased usage of Citi Bikes in the southern part of Manhattan and other major hubs such as Brooklyn, Hoboken and Historic Downtown New Jersey. We used these data to inform decisions where future stations should be created to maximize success and utilization. We feel that a station hub would ensure more user involvement.

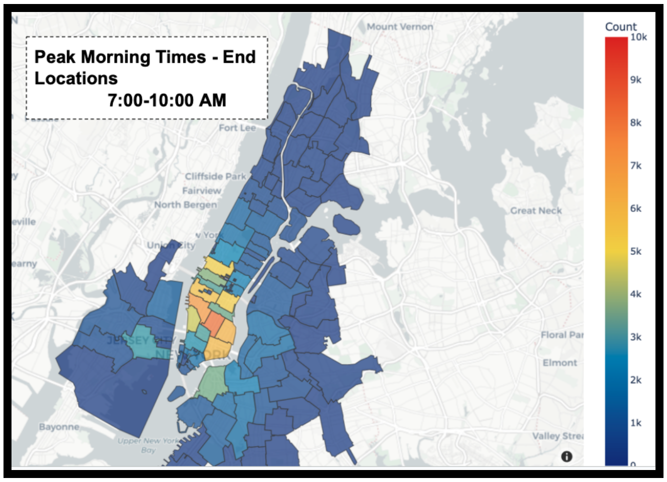


Figure 5: Peak morning time heat map

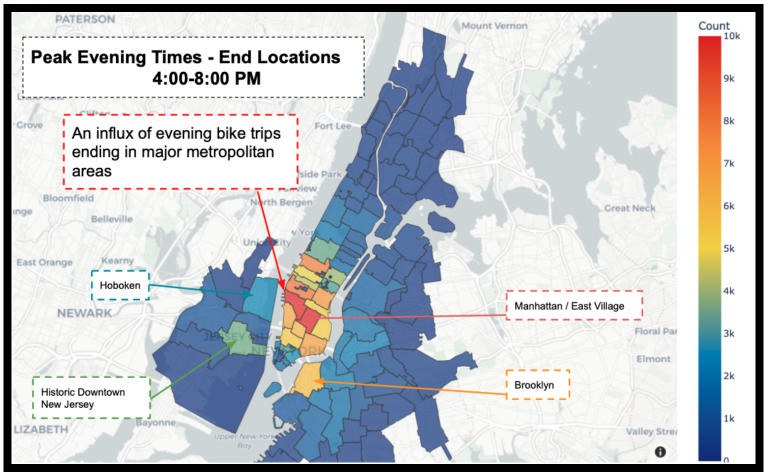


Figure 6: Peak evening time heat map

Question 6: Where does Citi Bike expand next?

To continue the expansion of stations in the New York/New Jersey metropolitan area, Citi Bike must find the next best zipcodes to put stations in. The best zipcodes should have these criteria: 1) Adjacency to zipcodes with Citi Bike stations already. 2) High density of residents or high density of offices/work. Based on these criteria, we were able to show the best zipcodes to expand Citi Bike to using our data set. We created the ‘Density score’ ranking to help easily differentiate the densities of the zipcodes. Figure 4 shows the top five residential zipcodes based on density score that Citi Bike should expand to next. We also verified that these zipcodes are adjacent to others already with Citi Bikes in Figure 5 by plotting the data onto a map using the geopandas library.

Density score is a scale from 0 to 100, 100 being the highest.

$$Residential\ Density\ =\ \frac{zipcode's\ population}{zipcode's\ area\ (m^2)}\quad,\quad Density\ score\ =\ \frac{zipcode's\ residential\ density}{data\ set's\ highest\ residential\ density}$$

The highest density zipcode in the data set has a density score of 1

Figure 7: Top 5 residential zipcodes to expand Citi Bike to

	State	City	Zipcode	Population	Density score
1	New York	Queens	11372	66636	62
2	New York	Kings	11226	101572	53
3	New York	Queens	11373	100820	45
4	New York	Kings	11219	92221	42
5	New York	Bronx	10472	66358	42

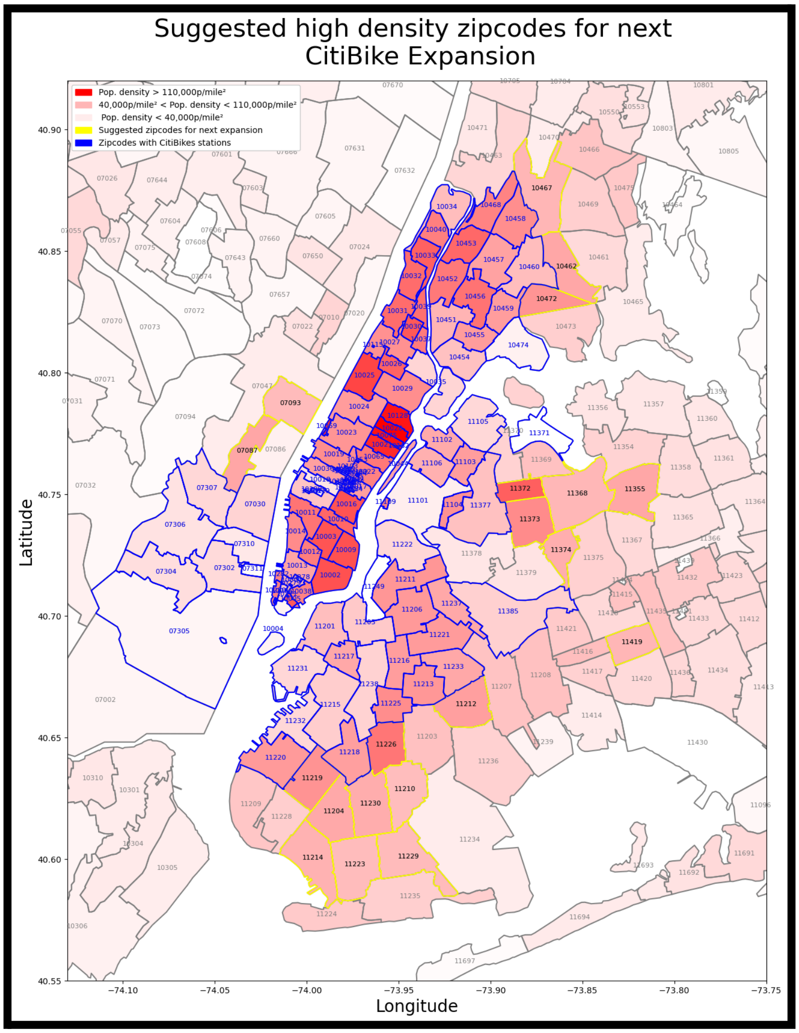


Figure 8: Suggest zipcodes for expansion

Conclusion

Taking the Citi Bike data set from the last 10 years we were successfully able to identify who the user base is, when and why they use the system. We were able to use this information to make data driven decisions about how to expand in the future. We were able to rank all zip codes in the greater New York Metropolitan area and suggest which ones would be best to expand the Citi Bike system into next. Our suggestion is to expand further into southern Brooklyn and into Queens because these zipcodes have both high residential population density and adjacency to existing zipcodes with Citi Bike stations already.

Works Cited

- Ley, Ana. “Citi Bike Struggles to Keep up with New Yorkers' Love of Cycling.” *The New York Times*, The New York Times, 2 Dec. 2021, https://www.nytimes.com/2021/12/02/nyregion/citi-bike-parking-docking-station.html.
- “Citi Bike Is Going to Dramatically Expand!” *Citi Bike NYC*, https://ride.citibikenyc.com/blog/citi-bike-is-going-to-dramatically-expand.
- Lovelace Jr., Berkeley. "New York City's bike-sharing program just got a lot bigger." CNBC, CNBC, 3 Aug. 2018, www.cnbc.com/2018/08/03/new-york-citys-bike-sharing-program-just-got-a-lot-bigger.html.
- Lohr, Steve. "New York City's Bike Share Program Is Finally Growing Up." The New York Times, The New York Times, 8 Aug. 2019, www.nytimes.com/2019/08/08/climate/new-york-city-bike-share-citi-bike.html.
- Sisson, Patrick. "How New York’s Citi Bike Started a Transportation Revolution." Curbed NY, Vox Media, 24 May 2018, ny.curbed.com/2018/5/24/17385236/citi-bike-nyc-history-anniversary.