A study of Citi Bike utilization across NYC zip codes

Jimmy Zhang

Project goals and approach

Objective: Determine which zip codes Citi Bike should consider adding additional capacity in

I worked with the following public datasets and tables in BigQuery, using SQL queries to conduct exploratory data analysis (EDA), answer key questions about the data, and execute joins and other functions across multiple tables to obtain the final table for analysis:

- new_york_citibike
  - citibike_stations
  - citibike_trips
- geo_us_boundaries
  - zip_codes
- census_bureau_usa
  - population_by_zip_2010

Citi Bike should increase capacity and user acquisition efforts in zip code 10010

Neighborhood: Flatiron District to East River

Number of stations: 7

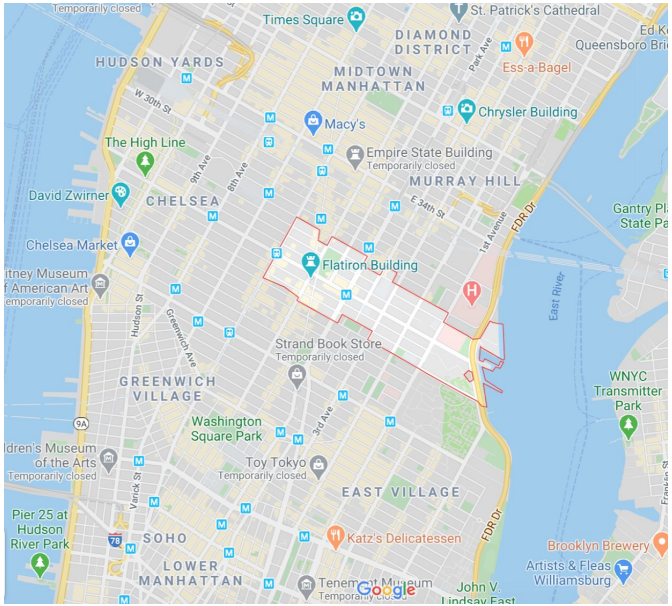
Total capacity: 282

Total population: 95,502

Avg. daily trips taken per docking point: 3.59

Avg. daily trips taken per person: 0.08

Avg. docking points installed per person: 0.02



Key metrics to define success for Citi Bike

1. Daily trips taken per docking point installed *(total trips / total capacity)*
    a. How utilized are the docking points in each zip code? Are there enough docking points to meet the demand in all zips?
2. Daily trips taken per person *(total trips / total population)*
    a. How active are residents in each zip code? Are there zip codes that could be targeted as areas to improve outreach and usage in?
3. Docking points installed per person *(total capacity / total population)*
    a. How available are docking points in each zip code? Are there enough to meet present and future demand?

I considered #1 to be the primary metric for analysis, since we want to make sure that existing demand is being met comfortably before targeting increased future demand

The top 5 zip codes in terms of trips taken per docking point are 10282, 10280, 10010, 10014, and 10012

Zip code	Neighborhood	Avg. daily trips taken per docking point	Avg. daily trips taken per person	Avg. docking points installed per person
10282	Battery Park City - North	5.96	0.06	0.01
10280	Battery Park City - South	3.62	0.02	0.01
10010	Flatiron	3.59	0.08	0.02
10014	West Village	3.46	0.21	0.06
10012	SoHo	3.42	0.16	0.05
Overall		1.58	0.10	0.06

Opportunity: High, Medium, Low

Zip code 10010 is one of the most capacity constrained areas relative to both trips activity and population, while having a user base that is both decently active but has potential to increase in the future

Newly learned and applied SQL skills

- Subqueries
- CTEs
- Different types of joins (inner vs. outer, cross) and across more than 2 tables
- CASE WHEN
- Window functions
- UNION
- Geography functions (ST_GEOGPOINT, ST_WITHIN)
- Statistical functions (AVG, STDDEV)

Other topics covered during mentorship

Mentee: Jimmy Zhang, incoming Product Insights Analyst at Google

Mentor: Storm Hurwitz, Senior Analyst at The New York Times

- SQL interview skills
- A/B testing walkthrough / experimental design questions
- Job search discussions throughout interview process
- Walkthrough of New York Times analytics project examples

Creating the final table

1. Obtained each station's name, capacity, and zip code by joining citibike_stations with zip_codes
2. Obtained the number of trips that started from each station on each date, and each start station's zip code, by joining citibike_trips with zip_codes
3. Obtained the population of each zip code in New York City whose population was between 10,000 and 250,000 by joining population_by_zip_2010 with zip_codes
4. Joined (1) with (2) using station name and with (3) using zip code, and aggregated by the date and zip code, to produce a table with the **total population, total capacity, and total trips taken in each zip code for each date from July 2013 through May 2018**

Sample EDA findings

- 755 out of 968, or 78%, of the stations in citibike_trips exist in citibike_stations
- 54 out of 72, or 75%, of the zip codes for stations in citibike_stations have population data that exists and is considered normal
    - The average and standard deviation of the populations of NYC zip codes, first excluding the 25 zip codes with population 0 in the data, are 131,751 and 81,886, respectively
    - I calculated the upper and lower bounds of the distribution as one-and-a-half standard deviations above and below the mean, took these as goal posts, and chose nearby figures that were more business interpretable
    - Thus, the subsequent analysis considers only the zip codes with population between 10,000 and 250,000