Citi Bike Expansion Capstone I
Denzel S. Williams
Springboard DS Career Track
September '20 Cohort

Problem Statement

How can Citi Bike use the socio-economic properties of a specific region along with historical ride data to predict how many bike stations should be allocated to it.

Extension: Recommendations on where in the region should the bike stations be placed.

Context

Citi Bike is a privately owned public bicycle sharing system serving the New York City boroughs of the Bronx, Brooklyn, Manhattan, and Queens, as well as Jersey City, New Jersey. It officially opened in May 2013 with 332 stations and 6,000 bikes. Annual expansions have brought the totals to 706 stations and 12,000 bikes as of October 2017, making the service the largest bike sharing program in the United States. Further expansions for Citi Bike are planned to extend its service area across the Bronx, Brooklyn, Manhattan, and Queens, and increase the number of bikes to 40,000.

At this point in time, CitiBike doesn't offer dockless biking, where the user can pick up and drop off a bike wherever they please, therefore bike stations remain a critical component to their expansion plans. Where should bike stations be placed? How many bike stations should be in a given radius? How many bikes should a station be able to hold? Answers to these questions are crucial to a successful expansion.

Criteria for Success

- Develop a model that can predict the number of stations that should be placed in a region of New York City.
- *Output recommendations on approximate locations in the region where bike stations could be placed.*

Scope of Solution Space

Constraints

Citi Bike can't simply pop a bike station anywhere they please and they have to follow the rules and regulations of the region. Additionally, a region may not have the physical attributes to support Citi Bike such as too little roads to ride on. Due to complexity of the problem and the nature of the data that we have, it is impossible to output accurate locations on potential bike stations.

Data Sources

New York City Neighborhood Profiles

https://furmancenter.org/neighborhoods
https://data.cityofnewyork.us/City-Government/Demographics-and-profiles-at-the-Neighborhood-Tabu/hyuz-tij8
https://data.cityofnewyork.us/City-Government/Community-Districts/yfnk-k7r4

Citi Bike Data (Trip Data + Daily Data)

https://www.citibikenyc.com/system-data

Key Stakeholders

Bill de Blasio - Mayor of NYC
Laura Anglin - Deputy Mayor
Polly Trottenberg - Department of Transportation Commissioner
John Zimmer - Lyft President
Ed Skyer - Citi's Head of Public Affairs
Ken Podziba - CEO of Bike New York
Joe Cutrufo - Transportation Alternatives Spokesperson
Council Members of New York City
Laura Fox - General Manager for Citi Bike

Approaching the Problem

Data Wrangling + Cleaning - New York can be grouped into boroughs, community districts, and neighborhoods. The socio-economic data that I found was on the community district level.

1. Scrap the trip data from the Citi Bike S3 bucket
2. Scrap the daily ride data summaries from the Citi Bike website
3. Collect the community district data for the 59 community districts in NYC
4. I might have to (want to) create a relational database for all this data

Data Exploring - After the data is cleaning there are some preliminary questions that I would like to answer. Additionally if I could create interactive maps that would be awesome.

1. Distribution of stations amongst the 59 communities
2. Average trip duration
3. Average trip distance
4. Which age group using the service the most
5. Most popular station overall, start station, stop station
6. Which community is the most popular overall, starting, and stoping
7. Inter community travel vs intra community travel
8. Bike station development over the years (map)
9. City bike station start/end station over the years (map)

10. Bike station coverage by quarter mile (map)
11. Is the city bike an effective replacement for the subway (nyc subway data required + map)
12. City bikes in a given radius
13. Average distance to nearest station (advanced)
14. Station usage over the day (advanced). Do people normally stop at this station or start at this station? (Can I get max number of bikes a station can hold data)
15. General Station usage over the years (map)