Exploratory Data Analysis Project:
Connecting Citibike to the MTA Network

Abstract

The aim of my project was to use NYC metro data to inform an exploratory analysis of potential locations for new Citibike stations. I identified 40 subway stations located along bike routes outside of the current Citibike service area, and explored the weekday evening subway traffic patterns to find the stations with the most foot traffic. These results could be combined with analysis of Citibike trip data and users' comments to select and prioritize sites for new bikeshare docks.

Design

Citibike is partway through a 5-year expansion project, during which they solicited user suggestions on new bike station locations throughout Queens, the Bronx and Brooklyn. One advantage of these new transportation resources could be to help alleviate the "first/last mile problem," where transit riders spend disproportionate time and effort reaching the first point of their transit journey, or connecting from the final transit stop to their destination.

By introducing new bicycle docks near busy subway stations, Citibike could build membership among transit commuters and facilitate more frequent trips within their network. Using the available MTA traffic data, I focused on subway turnstile exits and entries within evening peak hours, and selected stations with more weeknight than weekend traffic to promote sites with the most potential weekday commuters. In order to recommend locations that could safely accommodate new Citibike riders, I excluded train stations that were not within a reasonable distance of any bike lanes, and limited the scope to stations outside of the current Citibike service area.

The results highlight the busiest selected MTA stations, as focal points for regions that could be profitably served by the bikeshare system as an additional commuting option.

Data

I used MTA turnstile data from May - July 2019, to analyze typical commute patterns before the influence of the pandemic, during summer months when travelers might be more likely to try biking. I included only turnstile counts that reflected evening peak hours, leaving over 400,000 records. I also used a reference file from MTA with geographical and route information about all of the stations in the network.

To find the GPS coordinates of current Citibike stations, I downloaded a file describing the most recent month of Citibike trips and collected the unique stations. Bike route mapping files were also available through NYC Open Data, and I used the Open Data API to approximate the number of bike routes within a short distance of each station.

Algorithms

Feature Engineering:

The main feature I added and analyzed was the difference in cumulative turnstile exits or entries at a given station between two selected times, to approximate the volume of rush hour foot traffic at each station.

I also looked into the difference between exits and entries within peak hours, and used the ratio of weekday to weekend station exits as a proxy for the rate of commuter traffic.

Visualizations:

The top ten recommended sites and their average weekday traffic are shown in a bar graph, but I found geographical plots more informative to connect information between the bikeshare network and MTA stations. I plotted maps based on the bike lane infrastructure and borough boundaries to visualize the process of selecting MTA station locations outside of the Citibike network and near bike lanes, and included a heatmap in one result to show the volume of rush hour station traffic.

Tools:

- SQL data ingestion, querying the database with SQLAlchemy
- Data cleaning and visualization with Pandas
- Plotting with Matplotlib and Geopandas
- Querying an API to add a geographical feature