CitiBike: Case for Last-Mile Connectivity in the NYC MTA System

By - Aanvi Goel

ABSTRACT

The goal of this project was to analyze the viability of CitiBike as a solution to last-mile connectivity problem in the MTA system. I worked with data from the MTA website to perform exploratory data analysis and obtain insights into ridership patterns across different stations. I added data available from CitiBike to evaluate the current state of it's network and provide recommendations for improvement based on the bike rack proximity to MTA stations.

DESIGN

Client - Citi Bikes (Lyft)

The Metropolitan Transport Authority runs a massive network of public transportation across the city of New York with the subway serving a daily ridership of more than 5 million and connecting 472 subway stations. While this extensive network covers the breadth of the city, the subway station might still be located substationtially far from the final location for the commuter.

The last-mile problem in the context of commuter travel, is the need for a second mode of transportation to reach a destination and get back. By using CitiBIke to improving the quality of last-mile connectivity in the MTA system, we can increase the CitiBike ridership and enhance the commuter experience.

DATA

- MTA Turnstile Data: Turnstile data files from 2022 till-date are used to get the latest ridership trends this year. The datafiles have a total of 5687284 entries
- MTA Stations: Latitude/Longitude data is pulled from NYC OpenData to get the locations for all the MTA stations
- CitiBike Data: Station information data is pulled to get the locations for all the CitiBike stations in NYC

ALGORITHM

- Exploratory Data Analysis -
    - Using EDA techniques to clean data and drop extra rows/columns
    - Identifying anomalies in the data to fix the incorrect data values and dropping the respective rows in case of small impact
    - Vizualizing Exit data using histograms and rainbow plots
- FuzzyWuzzy -
    - Using the FuzzyWuzzy library to fix the discrepancies between Station names in the MTA Turnstile Data and the MTA Stations Data
- Haversine -
    - Haversine formula is used to calculate the distances between MTA stations and CitiBike stations

TOOLS

- SQL and SQLAlchemy for querying from database
- JSON for reading JSON data files in Python
- Pandas and Numpy for data cleaning and manipulation
- FuzzyWuzzy for data parsing
- Matplotlib for plotting
- Folium for creating map based visualizations

COMMUNICATION

A slidedeck and Jupyter notebook code are included along with this write-up as part of the project