Exploratory Data Analysis Project

Subway vs. Citi Bike Ridership Before and After Covid Hit NYC

By William (Wei-Lun) Mai

Abstract

The goal of this project is to analyze the impact of Covid on public transportation usage mainly on the subway and Citi Bike in New York City. After analyzing the impact of Covid when it struck the city in March 2020, I will drill down to see if there is a new Citi Bike usage pattern by analyzing the Citi Bike trips data around several selected subway stations, which had a ridership drop in 2021 compared to the pre-Covid level.

Design

I first compare the ridership between the subway and Citi Bike during the month between March and May 2020 to identify the impact of Covid on these transportation methods. Later I analyze the ridership drop of each subway station by comparing maximum daily entries between 2020 and 2021. Next, I pick a few subway stations from different districts for further analysis of whether the surrounding Citi Bike stations have experienced any substantial change in the usage pattern. Final results could be presented to Citi Bike to adjust the location and the number of Citi Bike stations based on the new pattern discovered.

Data

Two datasets are used in this study. The MTA turnstile dataset contains the count of each turnstile entry every four hours for all subway stations. This dataset is mainly used to calculate the overall ridership of the subway and each station for a certain period. The Citi Bike dataset contains detailed information on each Citi Bike trip taken. This dataset is used to calculate the numbers of overall Citi Bike trips and trips originated from a specific station for a certain period. Both datasets contain the data for the period from March to May and all these following years: 2019, 2020, and 2021.

Algorithms

Daily subway entries and daily Citi Bike trips: All numbers of daily subway entries and daily Citi Bike trips are the moving 7-day average to remove the seasonality.

Maximum daily entry: Maximum daily entry is derived from taking the maximum daily entry between March and May for the same year. In this case, the Covid impact on 2020 data will be removed i.e. the maximum happened on the date before Covid struck in 2020.

Subway ridership drop: Subway ridership drop is calculated by taking the difference of Maximum daily entry between 2021 and 2020. A negative number means the ridership decreased and a positive number means ridership increased.

Tools

- DB Browser for SQLite for importing and accessing both datasets
- Numpy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting

Communications

Presentation, write-up, and Jupyter Notebook files can be found on the following GitHub page:
https://github.com/zyzzyva1423/EDA_Project