Exploring New York City Transit

Anurag Prasad & Jarrod Lewis
{anuragp1, jl101995}@bu.edu

Overview

Is there a relationship between subway ridership and CitiBike usage in New York City? This project aims to illustrate patterns between these two major forms of commute. This could potentially help identify smart locations for bike hubs based on proximity to subways. Our hypothesis was that it would be smarter to place more Bike Hubs around overcrowded subway stations. In addition to comparing these forms of transport, this project explores subway and CitiBike usage based on pedestrian traffic and weather in the stations' respective regions. With this information, we can see where it would be worth adding or removing CitiBike stations if there is a level of usage disproportional to the pedestrian traffic, or during which months maintenance of these services is most important.

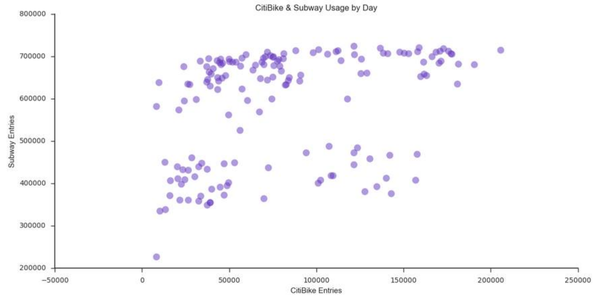
Data

We sourced the following datasets from NYC Open Data. **Subway Stations**: all subway stations with their geolocations. **Subway Turnstile Data**: turnstile entries for each subway station, but with a different naming convention than Subway Stations. **Bi-Annual Pedestrian Counts**: bi-annual pedestrian count of 114 regions, including 100 on-street locations (primarily retail corridors), 13 East River and Harlem River bridge locations, and the Hudson River Greenway. **CitiBike System Data**: a S3 bucket of CitiBike trip histories containing trip duration, start and end station names, and station coordinates. **Central Park Weather Data**: daily weather information for Central Park region, which represents New York City weather as a whole

A challenge of combining Subway Stations and Subway Turnstile Data was that the attribute naming conventions for these datasets varied. This forced us to combine a large part of these data sets manually. We combined these to calculate total entries for each station. We used a 2D-sphere nearest-neighbors approach to classify each subway station and CitiBike Hub by the regions in the Bi-Annual Pedestrian Counts dataset. The Weather dataset contains precipitation, temperature, date, and other descriptors which we chose not to use. We joined Subway Stations and Turnstile Data by date to derive the weather for each trip and day.

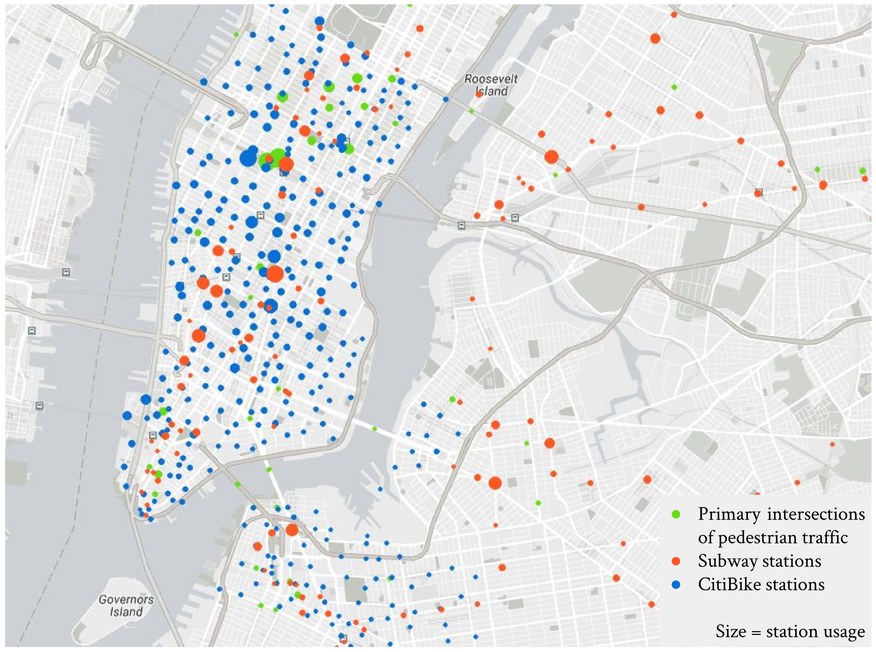
Station Usage Statistics

To see how subway usage varies with CitiBike usage, we ran a correlation on total usage by day for each of the transportation methods. Their correlation is 0.36 with a low p-value of 4.3 e-06. The following scatterplot displays the relationship between daily usage for each station type.



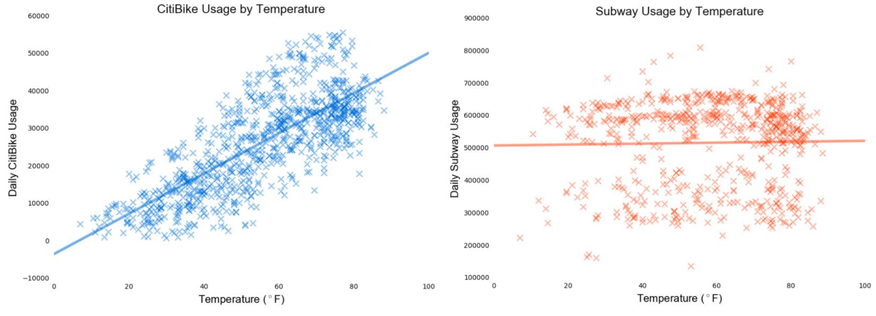
Mapped Stations

Although there appear to be some clusters that support our hypothesis of high correlation between a region's pedestrian traffic and its stations' traffic, when we performed statistical analyses on this data, the expected correlations were not found. However, the 114 pedestrian count regions may have skewed our findings since they are not dispersed evenly throughout the regions in which CitiBike and subway stations exist. Further investigation that would be helpful is finding the correlation between CitiBike stations and subway stations near each other rather than near the same pre-defined regions.



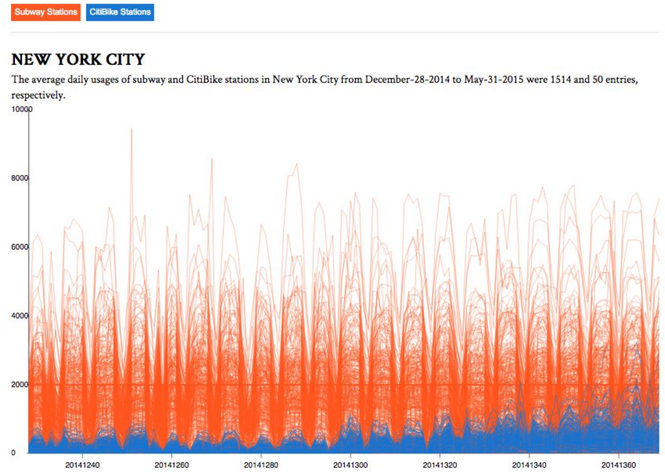
Effect of Weather

The strongest correlation we found in our investigation was between CitiBike usage and temperature. Subway usage did not correlate to any weather events which exemplified how inelastic this subway system is in New York. People will use the subway regardless of external conditions. The Correlation between temperature and CitiBike usage is **0.76** with a very low p-value of **1.1 e-171** (much less than . 05).



Station Usage Time Series

This time series shows a very distinct weekly trend in which there are peaks during the typical work week and valleys on the weekends. The only exception to this trend can be seen in CitiBike usage where usage varies based on the time of year and rises during the summer, when use is more even across the weeks.



Conclusion

Most of the correlations we expected to find were not found. There was a strong relationship between CitiBike usage and temperature, but not precipitation. There was no correlation between Subway Turnstile usage and weather factors. Also, we did not find a correlation between pedestrian counts and station usage. The lack of correlation seems to point toward variables which we did not have the opportunity to investigate. The Subway time series data shows a very strong weekly cycle, under which the usage is much higher during weekdays than during weekends. This suggests that subway usage is controlled by social and economic factors such as commute during working days.

What implications does this investigation have for Boston? The subway system in New York operates on a much larger scale than the T in Boston. Despite having lower complexity than other transportation systems, the MBTA is still viewed as "a system reaching for expansion even as its core deteriorated." (Boston Globe 2015) We hope our findings motivate Boston city planners to explore whether or not such expansion is more valuable than expansion of alternative modes of transport such as bikes, taxis, and ridesharing services. Whereas in New York City the subway is a necessity and CitiBikes are dispensable, in Boston the T is often substituted for alternative modes of transportation. Therefore, we suspect that further investment in the T may not be as beneficial as investment in alternate modes of transportation.