Abstract

The NYC MTA subways system is an important part of NY's infrastructure system. The subway is the primary choice of travel for many NYC residents and for neighboring states, such as NJ and Connecticut. As a frequent commuter, I personally use the subway system on a daily basis and one frustration I have experienced many times is waiting in line to either enter or leave a turnstile. The goal of this project is to use MTA data to view the foot-traffic to see the activity of each turnstiles and see if there are ways to improve the flow of traffic and to lessen foot traffic to certain "popular" turnstiles.

Design

The main goal is to understand the utilization rate for turnstiles in order to see if there are turnstiles that are under-utilized that can help improve the flow of traffic within a certain station.
By looking at the foot-traffic, it will allow better understanding of the utilization rate for each turnstiles and investigate whether additional turnstiles need to be built out or direct foot-traffic to under utilized turnstiles. Directing traffic to other less utilized turnstiles will help with the flow of traffic and led to happier commuters.

Data

3 months of data were taken from the MTA website. The data-frame originally contained 2.9M rows of data across 13 different categories. Initial analysis showed that the data needed to be cleaned up before any analytical work can be done. The heart of the analysis was understanding the relationship between the enter and exit columns of the dataset. A wide range of functions and methods were used to help first clean the data before analyzing. Some initial datapoint that were tackled first were the following:

1. Duplicate entries
2. Outliers (foot traffic in the negatives and/or foot traffic in the high millions/billions)
3. Turnstile counter resetting and/or freezing

After cleaning the data, initial findings showed that there were indeed turnstiles that were under-utilized and turnstiles that were way over utilized.

Algorithms

1. Combine the following columns to create unique turnstiles for every station:
2. Used the .groupby function along with the .diff() to calculate the foot-traffic for every turnstile across all the stations
3. Utilized the .describe function to take a closer look at the foot-traffic and to see if there are any outliers
4. Selected the top 10 stations with the most foot-traffic and from the top 10 selected the busiest station to further investigate the turnstile usage

Tools

NumPy and Pandas for data manipulation
Matplotlib and Seaborn for plotting