Identifying the Busiest Days and Times at the Busiest NYC Subway Stations

By Nat Goodby

Abstract

The goal of this project was to understand which New York City subway stations are busiest and when to inform signature collection efforts by a non-profit organization called Transportation Alternatives. I worked with NYC Metropolitan Transportation Authority data to identify the ten busiest stations, as well as target days of the week and times that Transportation Authority can use for planning their outreach efforts.

Design

The mission of Transportation Alternatives is to *reclaim New York City's streets from the automobile and advocate for better walking, biking, and public transit for all New Yorkers.* Transportation Alternatives plans to collect signatures in support of their NYC 25x25 plan, which will be delivered to NYC's city council at the end of 2022. They want to identify priority locations, dates, and times for collecting signatures, and subway stations are one of their identified targets. They are interested in identifying the busiest stations and understanding what days and times traffic at those stations tends to be highest.

Data

NYC Metropolitan Transportation Authority Turnstile Data:
http://web.mta.info/developers/turnstile.html

This analysis began with acquiring all available MTA turnstile data from 2021. This data includes entry and exit counts for the turnstiles across all NYC subway stations. The initial dataset contained 10,717,660 observations. After cleaning the data, 8,646,820 observations remained and were used for analysis and visualization.

Algorithms

Thorough Exploratory Data Analysis was performed on the data. The primary metric used in this analysis was station traffic, which was created by adding together entry and exit counts for all turnstiles at each station for a given time period. After this metric was created, outliers and suspicious values were identified and discarded to ensure accurate results. Data was then aggregated by station across various time periods and visualized.

Tools

- SQLAlchemy to query data from SQL database.
- Pandas for data cleaning and manipulation.
- Matplotlib and Seaborn for data visualization.

Communication

Presentation slides used to communicate findings can be found in the project's repository on my GitHub page here. There you will also find JPGs of visualizations created for the analysis, along with the code used to clean and visualize the data.