Using data for optimal train usage
By: Jing Lee

Abstract
In this project, I am analyzing subway station entry data to recommend a train schedule that will optimize passenger experience by sending the most trains at the times of the most entries. By calculating an optimal train schedule based on the usage data, trains can be more efficiently used throughout the day. As a proof of concept, I focused on portions of the train 1 mta turnstile data only and compared it to the train 1 schedule. Based on my analysis, the peak usage of train 1 subways is slightly shifted from the peak hours of ridership. Instead, trains are arriving more frequently after peak ridership. For a train schedule that will more efficiently use trains, train arrivals should be shifted by 1 hour earlier allowing riders to get on them during peak hours. The same analysis can be applied to other trains to identify any need for an updated train schedule.

Design
In the project, the targeted client would be the MTA itself. By optimizing the use of trains, the MTA will aim to reduce the cost of running trains while aiming to keep customer satisfaction high by keeping wait times low.

Data
The MTA turnstile dataset for the investigation was from 09/25/2021 to 12/24/2021 (3 months). A row of data includes the station details (booth ID, name), trains passing the station (can be more than one train), division of the train, date and time, and cumulative entries and exit data at 4-hour intervals. For the analysis, the daily entries and exits were calculated. Then, the data was grouped by station. On the train schedule, only a few stations were provided with arrival information. Instead of choosing all trains on line 1, I focused on the stations that were on the train 1 schedule, so I could get a good sample of the station ridership with its corresponding train arrival data. Besides the entries data, the exits data was analyzed as well to get a good understanding of train ridership and to identify when trains should be more heavily utilized.

Algorithms
Exploratory data analysis was performed on the data sets:

- Removing duplicates entries since each unique station's turnstile should have 1 entry only per datetime.
- Creating a datetime object from the date and time columns for date and hourly analysis.
- Calculating the actual entries and exits data by looking at the difference between daily counter values.
- Recalculating the daily entries and exits since impossibly high values were observed. Some of these are likely due to the counter being reset.
- For the train schedule, entries in the timetable (.pdf) file were uploaded into excel, then time was converted to 24-hour format and then uploaded to pandas.

Tools

- Ingest raw data into a SQL database and use SQLAlchemy to query database in Python
- Pandas for exploratory data analysis and Matplotlib for data visualization
- MS Excel for copying train schedule (.pdf) into an excel file to upload into pandas

Communication

Below is a figure showing the number of entries at select stations on train 1 and with the number of trains arriving at each hour of the day.