Using MTA Turnstile Data to Optimize Food Truck Locations

Abstract

The goal of this project was to help a food truck business determine the best locations to park their food trucks. I used publicly available MTA turnstile data to determine where the busiest subway stations are located. To best advise my client, I broke out the analysis by looking into the busiest stations over a set time period, busiest stations by time of day, and busiest stations based on people entering or exiting the station.

Design

My client is a food truck business that wants to maximize sales by parking their trucks in the busiest locations. They specialize in offering breakfast food and are a popular choice for commuters looking for something to eat before heading into the office for the day or during a morning break. My client owns 10 food trucks that open at 8am and close at 2pm.

Data

I used MTA turnstile data from February 2022 to April 2022. The dataset contains the cumulative count of entries and exits for every individual turnstile every 4 hours. The raw 3-months of data includes 2,735,970 records and 11 features, all of which are categorical except for the "Entries" and "Exits" count.

Algorithms

I ingested the raw data into a SQL database and queried from that database into Python via SQLAlchemy. After cleaning the data (removing whitespace in columns, removing duplicates, etc.), I created a daily entries and daily exits column to reflect the number of entries/exits each day, since the raw data only included the cumulative count. I also determined the number of entries and exits for each turnstile between 9am and 1pm, since the food truck specializes in breakfast food. I analyzed the following for each station:

- Total daily entries/exits from February to April
- Total morning (9am – 1pm) entries/exits from February to April
- Median daily entries/exits from February to April
- Median morning (9am – 1pm) entries/exits from February to April

Tools

- SQLAlchemy for querying data from database into Python
- NumPy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting

Communication

- Bar charts and box plots presented in the slides