Exploratory Data Analysis Project

Subway Station Analysis (Mentor App)
by Mariya Graff

ABSTRACT

The goal of the project was to identify the MTA subway stations in NYC best positioned to reach the target audience for my hypothetical client: "Mentor App" – a start-up that provides a platform for career mentorship from active professionals in the field. I utilized two main data sources: the MTA's turnstile data that provides entry and exit counts for all turnstiles in NYC, and the rankings from CollegeRaptor with student enrollment per college for 2022.

I leveraged the data to first identify NYC's top 3 colleges by student body in 2022. Then mapped out the locations of the main campuses, as well as the subway stations around each campus. I then narrowed it down to top 3 stations per campus by average daily traffic, resulting in 12 target stations around 4 target campuses. I calculated the estimate weekly and monthly projected total traffic, and provided further estimates for impressions and conversions for ad placement.

DESIGN

I focused my analysis on the student body in NYC as the best target market to launch the app. As a start-up, the Mentor App can leverage the networks of these top NYC colleges by population for greater reach, and advertise to a student body from broad fields of study, which will help the company expand its userbase and get further learnings for its offering.

DATA

My main dataset, the MTA turnstile data, contains data for 5025 turnstiles in 478 stations with records of entries and exits per 4-hour periods from 9/25/2021 to 10/29/21. I chose to utilize the latest month available, as I believe, it will be a more accurate estimate for the upcoming ridership and traffic in January 2022. I also created a database using SQLite where I inputted the data from the CollegeRaptor ranking, as well as the addresses of the college campuses. I converted the addresses to latitude/longitude coordinates for each campus of interest for visualization and analysis in Tableau.

ALGORITHMS

Cleaning:

1. Removed duplicate records that resulted from double audits (Regular and Recovered)
2. Removed duplicate station records that resulted from the subway lines included in different orders
3. Identified and addressed abnormal data
    a. Corrected negative audit counts reported in reverse for specific turnstiles
    b. Replaced outliers in 4-hour audits that were higher than the reasonable threshold of 14400 (or 1 person a second) with averages for each turnstile for that day.
4. Aggregated and analyzed data from stations per campus to identify the top stations per daily traffic, the average traffic per day of the week, aggregate average traffic per campus.

TOOLS

- NumPy, Pandas, SQLite for data manipulation and summary statistics
- Matplotlib, Seaborn and Tableau for visualization

COMMUNICATION

A presentation with a description of the top level process, key findings and results with visualizations.