Exploratory Data Analysis of the Metropolitan Transportation Authority Turnstile Data

By: Sarah Alabdulwahab

Abstract

The goal of this project was to explore the Metropolitan Transportation Authority (MTA) turnstiles data to determine the busiest stations in order to reduce the traffic by adding more turnstiles in those stations. I worked on four months of data provided by the MTA, which had features such as the station, date, cumulative number of entries and exits, etc. I performed feature engineering on the previously listed features along with some group by operations to produce the daily entries per station, this was the focus of my analysis. Finally, I was able to reach a few conclusions after performing EDA and the visualizations.

Design

This project originates from the SDAIA "T5" Data Science Bootcamp. The data is provided by the MTA, and presents the data of all turnstiles in all station in New York City through Long Island, southeastern New York State, and Connecticut. Finding the busiest stations would enable the MTA to take action to reduce traffic in that station by adding more turnstiles.

Data

The dataset contains 3,351,726 turnstile data entries with 11 features for each, this data was collected from the past four months. A few feature highlights include the name of the station, the data and time of the entered data, the cumulative entries and exits per turnstile. The data includes nearly 400 unique stations, therefore grouping the data based on the stations allowed me to produce in-depth analysis of the top 10 of them.

Algorithms

Data Cleaning & Feature Engineering

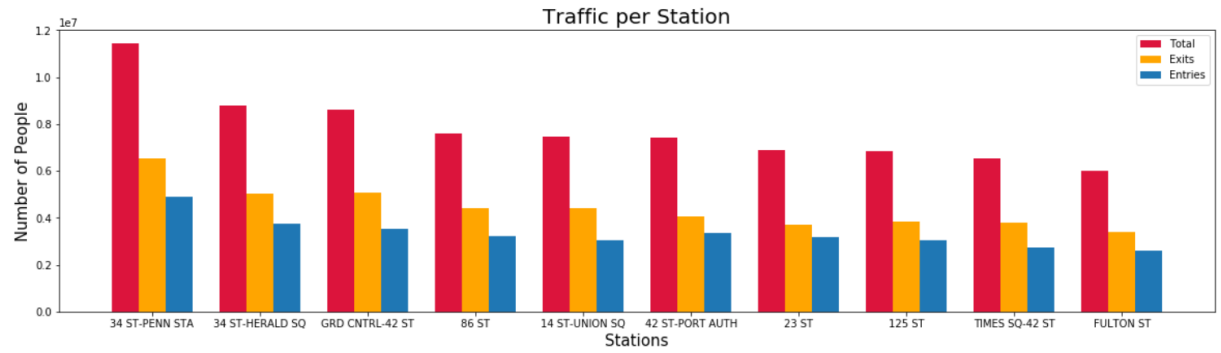
- Converting the date feature from objects to date-time
- Combining the C/A, UNIT, SCP, and STATION features to produce the unique turnstile ID.
- Extracting the daily number of entries and exits per turnstile.
- Adding the daily number of entries and exits to produce the total daily traffic.

Tools

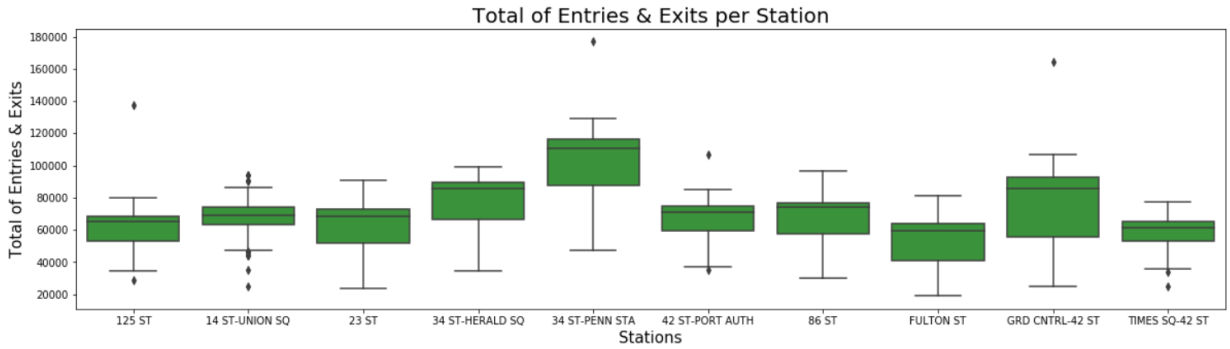
- SQLAlchemy for database creation.
- Pandas and Numpy for data manipulation.
- Matplotlib and Seaborn for plotting.
- Tableau for interactive visualizations.

Communication

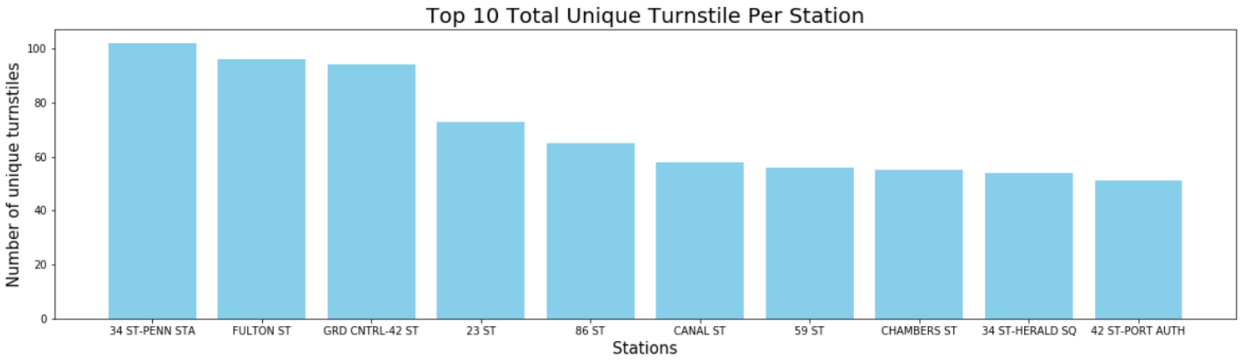
Top 3 busiest stations are 34 ST-PENN STA, 34 ST-HERALD SQ, and GRD CNTRL-42 ST.



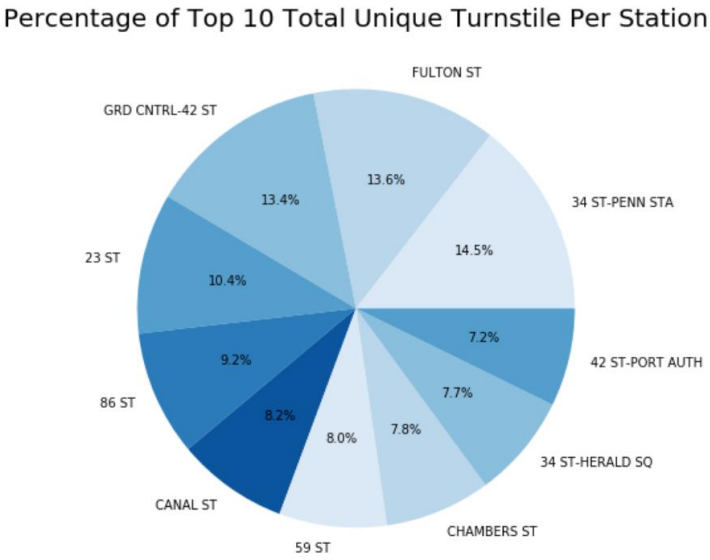
Range of daily entries per station.



The number of turnstiles in the top 10 stations.



The percentage of turnstiles in the top 10 stations.



Showing the positive correlation between daily entries and exits.