Metis Exploratory Data Analysis
Project Write Up
Tara Koplik

Abstract:

The NYPD Transit Bureau is interested in leveraging data to boost the effectiveness of their scheduling practices, in order to place more officers in subway stations where crimes are likely occurring. NYC Subway Harassment Data and MTA Turnstile Data for 2019 (the most recent year that Harassment Reporting data was available for) was analyzed to get an initial idea of what the data could tell us. Foot traffic and the number of reports of harassment at a given station were shown to have a strong positive correlation. High priority stations were identified as well as low priority stations where less officers are needed.

Design:

MTA turnstile data was aggregated at the station level to obtain monthly total foot traffic and overall foot traffic for 2019. The Harassment Report data was aggregated the same way and the datasets were merged on station.

Data:

The NYC Subway Harassment Data reflected stations that had one or more reports of harassment for the year. This dataset originally included 32 columns most of which were outside of the scope of this analysis and excluded. Since it is noted in the footnotes of the raw data that a crime listed as occurring at a subway station could have either occurred on the train where that particular station was the next stop or at that station, overall foot traffic was examined rather than the entrances and exits alone.

Algorithms:

- The correlation coefficient for total foot traffic and total number of harassment reports per station for 2019 was calculated.
- Exploratory data analysis in pandas FuzzyWuzzy library was used for fuzzy string matching to be able to combine station names that varied between the datasets
- Top priority stations were identified by using pandas rank function.

Tools:

Ingesting the raw data into a SQL database - Python via SQLAlchemy. Pandas data frames were used to view, clean, and aggregate the data. Python visualization libraries Matplotlib and Seaborn were used to create the charts. Google maps was used to create a custom map to display key data points.

Communication:

In addition to the slides below is the link for the map created using Google Maps:

https://www.google.com/maps/d/u/0/edit?mid=1U9TVzWkrCYUZ78rXWT2mpZ22JgippT-E&ll=40.7630730195453%2C-73.97431315000001&z=11