Exploratory Data Analysis of the Fall Quarter 2020 NYPD Crimes and NYC Subway Stations

Abstract: The goal of this project was to gather crime density data within a 0.5 mile buffer around the 25 busiest NYC subway stations between August 8, 2020 and December 26, 2020. The goal was to provide the City of New York with data that may aid in allocating funding towards increasing the safety at busy stations. New York City subway station turnstile data was used to identify the top 25 stations that had the most foot traffic between August 8, 2020 and December 26, 2020. The result was a table and a map of subway stations and corresponding number or crimes within the 0.25 mile radius.

Design: Two datasets were used: 1) NYC turnstile data and 2) New York Police Department (NYPD) complaint data. The NYC turnstile data was obtained from the publicly accessible mta.info website. Both data sets were between August 8, 2020 and December 26, 2020. The NYPD crime data contains all felony, misdemeanor, and violation crimes reported to the New York City Police Department for all complete quarters of 2019 and 2020. The NYPD data has geospatial components. Offenses occurring at intersections are represented at the X Coordinate and Y Coordinate of the intersection. Crimes occurring not at an intersection are geo-located to the middle of the block.

Data & Algorithms: The original NYPD Crime data consisted of 213,412 records for between August 8, 2020 and December 26, 2020. Using SQLalchemy the data was filtered down to only crimes occurring outside, bringing down the dataset to 45,712 records. The sample unit for this study was the number of crimes within a 0.25 mile radius of each subway station. Features in the mta dataset were date, time, entries, exits, and location of the station. Features used in the NYPD dataset will be date, latitude and longitude of the crime.

Tools: Tools used for this project included:

1. Ingesting raw data into a SQL database via web scraping
    a. SQL was used to clean duplicate data
2. SQL was used for cleaning and aggregating both datasets.
3. ArcGIS geospatial analysis toolbox was used to sum the count of crimes within a 0.5 diameter buffer with the busy subway stations at the center.
4. Matplotlib and seaborn were used to provide graphical displays of crime statistics near certain subway stations.