

Exploratory Data Analysis of 2019-2020 NYPD Hate Crimes and NYC Subway Stations

Project proposal prepared for the New York University
Student Diversity and Inclusion Division

Author: Rachel Hausmann

Date: 4/21/2021

Organization: Metis

Question of Interest: Where can we see a correlation between the stations with higher numbers of turnstile passes between 8:00pm - 4:00am and the number of hate crimes reported to the police in the years 2019 and 2020?

Purpose: Exploring any correlation between stations in that time range and crime might aid in informing minorities and other vulnerable student groups where to avoid or practice caution when traveling between 8:00pm and 4:00am.

Datasets: Two datasets will be used: 1) NYC turnstile data and 2) New York Police Department (NYPD) hate crime complaint data. The NYC turnstile data will be obtained from the publicly accessible [mta.info](http://web.mta.info/developers/turnstile.html) website. The NYPD complaint data will be obtained through the NYC OpenData project, which is an online clearinghouse of publicly available data published by various levels of government of New York state. The NYC hate 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.

Mta data source: <http://web.mta.info/developers/turnstile.html>

NYPD hate crime data source:

<https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Map-Year-to-Date-/2fra-mtpn>

The sample unit for this study will be the number of crimes within a specified radius of each subway station. Features in the mta dataset will be date, time, entries, exits, and location of the station. Features used in the NYPD dataset will be date, latitude and longitude of the hate crime.

Tools: Tools used for this project include:

1. Ingesting raw data into a SQL database via web scraping
2. Cleaning and aggregating the database using Python and SQL
3. GeoPanda will be used for the spatial analysis of the two datasets and to create a crime map overlaid with locations of subway stations.

4. Matplotlib will be used to provide graphical displays of crime statistics near certain subway stations.

MVP Goal: A minimum viable product MVP for this project could be a map of the NYC subway stations.