NYC Subway Data

Group 11:

- Parker Chea(parker.chea@baruchmail.cuny.edu)
- Huu Nguyen (huuhuyanh.nguyen@baruchmail.cuny.edu)
- Kristen Nanan (kristen.nanan@baruchmail.cuny.edu)
- Tasnia Anika (tasnia.anika@baruchmail.cuny.edu)
- Charles Reynolds (charles.reynolds@baruchmail.cuny.edu)
- Angus Lee (angus.lee@baruchmail.cuny.edu)

Abstract:

This paper used NYPD Complaint Data to identify locations in New York City where crime is significantly higher in relation to subway station locations. The study visualized the volume of riders compared to the volume of complaints to provide insight into the overall security of NYC Subway Stations and its employees, as well as its riders. It also visualized the volume of complaint types at each subway station throughout the day. The proposed solution involved cleaning the data using Python and Alteryx. Tableau is used to create a map of New York City, color-coded to highlight areas where there are more reported crimes. The study broke down the map by the type of crime to identify areas that need additional funding relating to public safety. Lastly, the results were used to allocate more resources to ensure the safety of residents, such as determining the need for more policing or improving current security systems.

Introduction:

Utilizing the NYPD Complaint Data, we determined the locations that require the most assistance and enhancement based on a number of metrics including concentration of ridership in combination with complaint types. Through visualizing the volume of riders compared to the volume of complaints, we were able to create insight into the overall security of NYC Subway Stations and its employees as well as its riders. Furthermore, creating the visuals with geo-spatial data allows us to highlight pain points, which we believe will provide value towards the quality of life as well as focusing on areas of improvement.

Data and Task Abstraction:

- Domain Level: Direct MTA Stakeholders include, in a narrow sense, MTA Riders and MTA Officials/Workers. More broadly, we understand indirect MTA stakeholders to include, all residents of NYC, commercial/private actors, taxpayers whose revenue depends on MTA Functionality, and foreign or domestic visitors to NYC that use its public infrastructure. As we are unable to target all stakeholders our domain for this project focuses more specifically on riders, officials, and employees of the MTA.
- Task: Our task is to analyze existing NYC MTA Felonies, Misdemeanors and Violations as a means of pinpointing pain points for public safety interventions as is related to Subway Stations. Our discoveries are meant to provide a clearer picture of stations where incidents frequently occur. The result will help guide possible interventions for the overall safety and benefits of MTA employees and Riders.
- Evidence of its need:
    - Subway ridership is rebounding since Covid severely decreased ridership, “Before the pandemic, nearly 40 percent of the agency’s operating revenue came from fares. Today, that share has sunk to 23 percent as ridership has slumped” - NYT 02/23.
    - There were 10 killings on the subway last year, compared with an average of two annually in the five years before the pandemic. - NYT 11/22.
- Data: Subway Entrance and Exit Data is currently comprised of 25 attributes mostly strings based on factors and qualities around each station, number of entries and exits, dates, as well as longitude/latitude for 1869 items. This pertains to our topic by setting the volume and outline of stations for further comparison. Our second data set is the NYPD Complaint Data which contains roughly 32 attributes of mostly NYPD system-specific data, as well as longitude/latitude, date, crime, gender, suspect, and police station jurisdiction. This data set contains null and past years that exceed the scope of our Domain. We plan to limit the data to roughly 2017 to address the more recent concerns of our domain stakeholders as well as clean null items. After Cleaning the data for Null items the total item count is 63,505. There are 26,233 felonies, 29,062 misdemeanors, and 7,210 violations across the Burroughs. There are 44 Offense types recorded, and these vary in severity from felonies being the most

severe, Misdemeanors less severe and Violations being the least of all. These complaints take place over time and this too is recorded, by time of day, and by date

- Task Abstraction: To better achieve our direct task of highlighting pain points and analyzing and comparing incidents in the MTA, we were able to isolate the MTA related incidents from the subway stations, using the names as a unique identifier. . For instance, is the longitude/latitude of a public safety incident within 20 Square feet or 5 square feet of a Subway Station? From here we can create a common geospatial relationship to crimes in the subway and directly within the vicinity of a station. This abstraction forms the basis of our ability to further classify and compare incidents from station to station using the station longitude and latitude as the unique identifier or bin between this data set and potentially other data sets.
- Further subsets of abstractions will be explored around summarizing, classifying, and comparing other attributes to add a better and deeper understanding of the pain points discovered in our initial abstraction. Findings from comparing attributes like date/time, distance from the nearest police precinct, and other observable factors from the given attributes.
- The justification for these abstractions centers around our domain stakeholders' perceived experience as they enter and exit subway stations thus criminal activity both in and around a subway station may add to the overall negative sentiment of riders and thus is worth extrapolating.

Solution:

- Our proposed solution is to use this dataset to pinpoint locations across New York City where crime is significantly higher in relation to subway station locations. As mentioned in the previous section, the dataset we will be using requires some cleanup as there are null values as well as data dating back to 2004, which is irrelevant for our purposes. In order to do this, our proposed implementation approach involves cleaning the data so that it is usable through Python. Once our datasets are clean, we will use Tableau to create a map of New York City and color code the areas where there are more reported crimes. To take our visualization further, we also plan to break down the map by type of crime, not just volume. There are some crimes that are similar but the level of severity differs (felony vs. misdemeanor) which we will showcase as well.

- The results of our proposed solution will highlight areas in the city where more resources can be allocated to ensure the safety of residents. Does the area need more policing? What is the condition of current security systems? Trends in our results for certain locations can pinpoint which areas may need additional funding relating to public safety or additional research into why these areas are more susceptible to crimes.
- Our dataset is also broken down by time of day. Through our visualizations, we can see that crime spikes significantly during morning and evening rush hours. A solution for this could be increased police presence in these hotspots in an effort to mitigate crimes from happening.
- Ideal Scenario/Idiom: let's say after creating a heatmap overlapping subway locations and crime rates on a map of NYC, we see that subway locations in Bushwick have higher rates of crime relative to other areas. We take a closer look at Bushwick and through a pie chart breaking down types of crime, we see that a large percentage of those crimes are assaults in the stations. We can use that information and propose that NYC should increase the NYPD manpower at those stations or in the surrounding areas to improve ridership quality.

Milestones:

- To have a clear project plan, we outlined several milestones we accomplished. Our first milestone was to find datasets that we can utilize together based on a similarity, one of the NYC subway ridership and one of the crime incidents in NYC. However, due to differences in the data, our group had trouble founding a way to incorporate the latitude and longitude of both datasets and go off of geo-spatial data. We decided to use the latitude and longitude and begin data cleaning on Python. Two members who are more experienced in Python worked together and made sure proper cleaning was done without losing too much data. This is where most of our issues arose, we hoped we can incorporate the two datasets without losing too much data and without any other complications, however we could not accomplish our initial goals we set during the beginning of the project. Cleaning up the dataset took much longer than anticipated. After we had a clean and concise dataset, we narrowed down the crime incidents of the different train stations. From here, our group members started to use Tableau to visualize the data and information we found. We wanted to create multiple interactive Tableau dashboards and maps along with other Tableau visualizations. We aimed to show, where crime is significantly higher in relation to subway station locations and the type of crimes commited. There are many ways we can expand our findings with more time, research and, better resources. We hoped to attain valuable information about crime in relation to subway stations and we hope that this information is useful to our fellow NYC subway riders.

Discussion, Future Work, and Conclusions:

- The project will utilize subway and complaint data as main sources of information because they are relevant and available datasets that can help answer this project research question. The subway data provides information on the location and ridership of subway stations, which can help to identify areas with high or low transit demand. Additionally, it's possible to identify areas with high or low crime rates with information on the type and location of criminal reports, provided by the complaint data.
- Some of the advantages of these datasets are:
    - They are official and reliable sources of data from government agencies (MTA and NYPD).
    - They cover a large time span (2006-2020) and geographic area (New York City).
    - They have a high level of detail and granularity (e.g., station name, complaint type, precinct, etc.).
    - They can be easily accessed and downloaded online.
- Some of the limitations of these datasets are:
    - They may not capture all the factors that influence crime and transit demand (e.g., socio-economic status, land use, population density, etc.).

- They may contain errors or inconsistencies due to data collection or processing methods (e.g., missing values, duplicates, outliers, etc.).
- They may not reflect the actual experiences or perceptions of residents, subway employees, police officers, etc. who may have different views on safety and security issues.
- They may be affected by external events or changes that are not accounted for in the data (e.g., COVID-19 pandemic, policy reforms, infrastructure improvements, etc.).

No conclusion as of this moment.

Bibliography:

- Zraick, K., Kvetenadze, T., & Paris, F. (2022, November 4). *How safe is the subway? what those who work there have to say.* The New York Times. Retrieved March 17, 2023, from https://www.nytimes.com/2022/11/04/nyregion/new-york-subway-safety.html
- Ley, A. (2023, February 13). *As subway ridership rebounds, some women are reluctant to return*. The New York Times. Retrieved March 17, 2023, from https://www.nytimes.com/2023/02/13/nyregion/nyc-subway-women-crime.html