NYC Subway Crime

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 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. The data was filtered for crimes committed in subway stations and buckets were created to normalize longitude and latitude of subway stations. 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. The limitations of the dataset are acknowledged – through cleaner geographic data and a larger temporal scope, we can enhance subway safety further with more detailed strategies.

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: Our 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 boroughs (Exhibit 1). 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 the time of day, and by date. There are further 45 Offense Types that exist across levels of severity (Exhibit 2).

Exhibit 1:

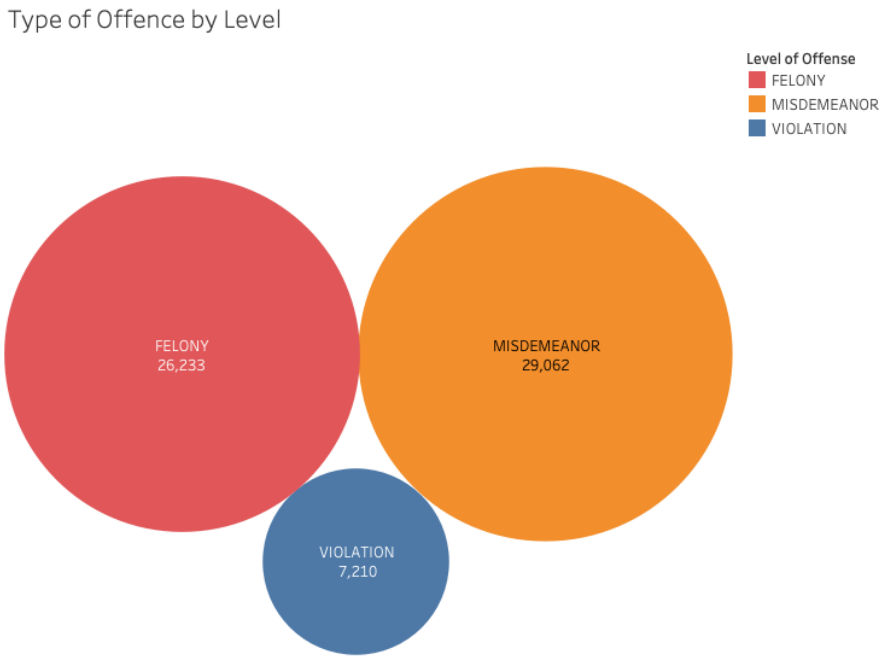
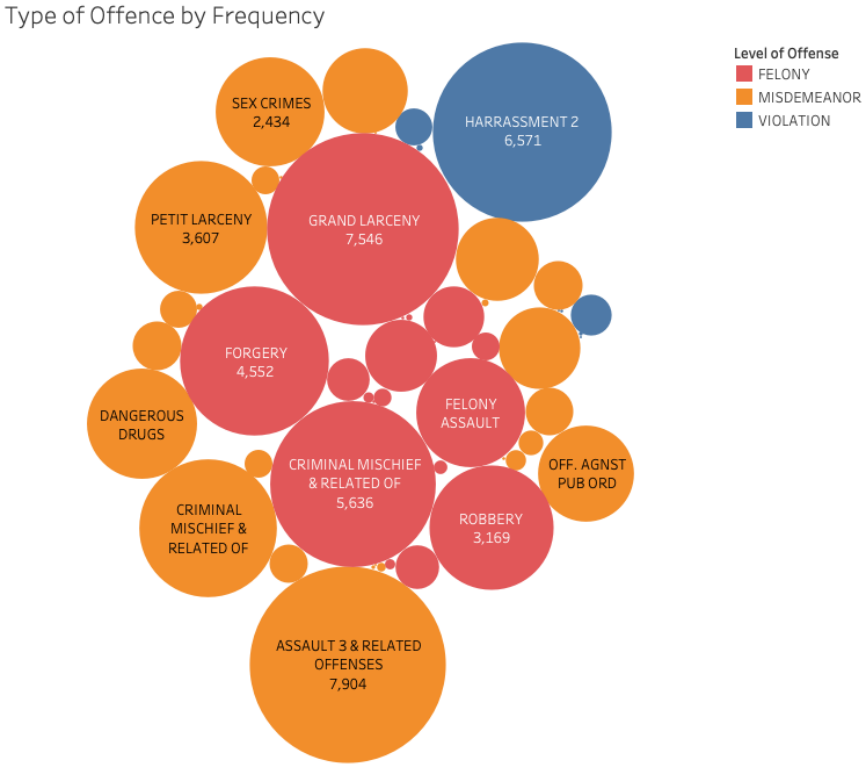


Exhibit 2:



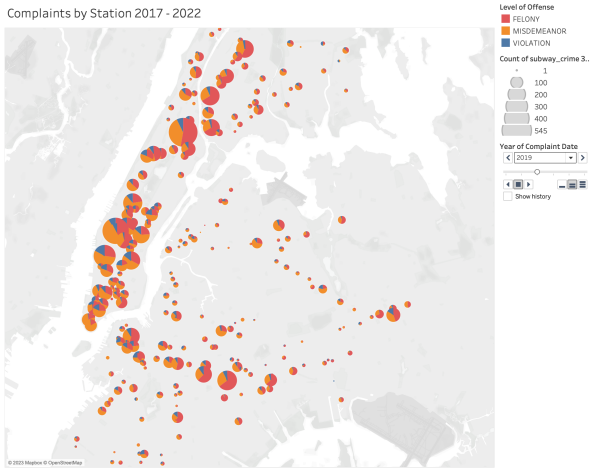
- Task Abstraction: To better achieve our direct task of highlighting pain points and analyzing and comparing incidents in the MTA, we were able to isolate the MTA-related incidents from the subway stations, using the station names as a unique identifier. For instance, the longitude/latitude of an MTA Crime incident varied from incident to incident regardless of having matching station names. So we had to isolate all instances for every station and take a median coordinate to identify a single set of coordinates for every unique subway station geo spatially. From here we can create a common geospatial relationship to crimes in the subway and directly within each station. This abstraction forms the basis of our ability to further classify and compare incidents from station to station using the station name to identify the frequency and create an in-depth analysis to be covered in our solution.
- Further subsets of abstractions were explored around summarizing, classifying, and comparing time attributes to add a better and deeper understanding of the pain points.
- 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:

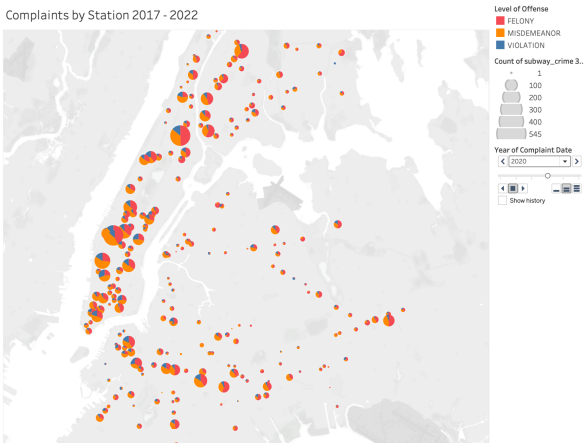
- Our solution is to use this dataset to pinpoint locations across New York City where crime is significantly higher in relation to subway station locations. In order to do this, we utilized Tableau to create a map of New York City and color code the areas where there are more reported crimes. The results of our visualizations will highlight areas in the city where more resources can be allocated to ensure the safety of commuters. 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 first visualization is a map of New York City, showing the breakdown of crime at each subway station by level of offense. The size of each mark indicates the volume of crime at each station. The mark is also a pie chart showing the different levels of offense: felony, misdemeanor and violation. We are able to see which types of crime

occur most at each station. We also included a time lapse to show how crime has changed over the years since 2017. This provided us with a clear picture of how crime dropped in 2020 (due to COVID-19) and how it climbed back up to similar numbers post-pandemic, as shown in Exhibit 3. This visualization gives a quick glimpse into where the most problematic stations are. These results can help pinpoint which stations may need additional assistance with surveillance or police patrolling

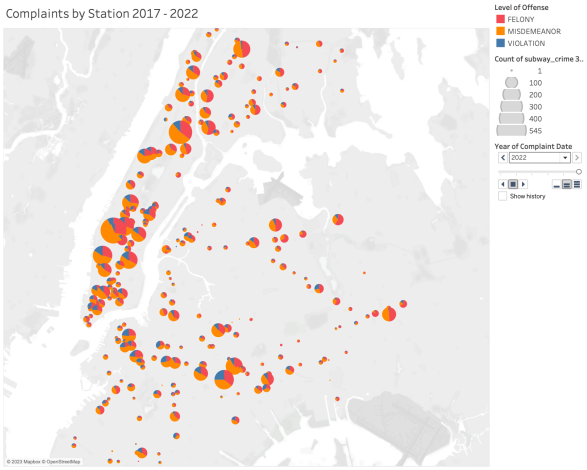
Exhibit 3:



Crime by Station: 2019



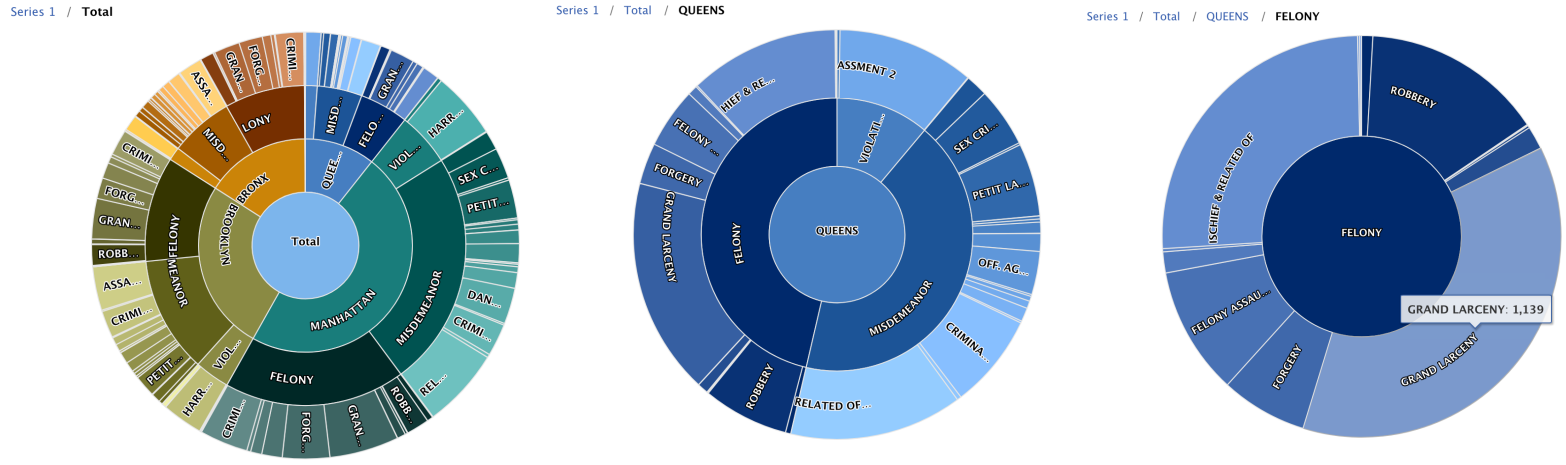
Crime by Station: 2020



Crime by Station: 2022

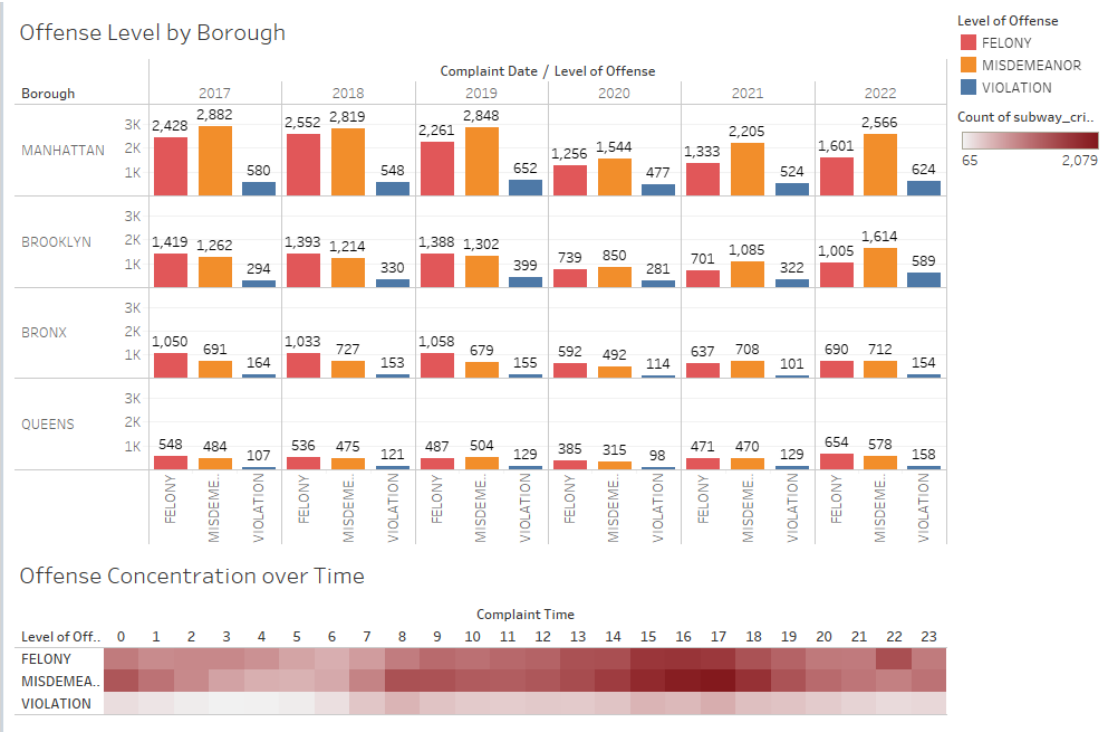
- In addition to seeing our data by level of offense, we wanted to see a breakdown of the individual crime. Since there were over 45 unique crimes, we created a sunburst chart to break down the data by each borough and level of offense to see the exact crime and the number of complaints. By using this data, we can see if there are any particular crimes affecting a borough. For example, we can see that grand larceny accounts for the most felonies in Queens, outlined in Exhibit 4. From there, resources can be allocated to combat that specific crime.

Exhibit 4:



- Exhibit 5 shows a bar chart displaying the number of offenses by borough and severity over time. It gives us an idea of the health of each borough over the years, allowing us to perform deep dives into select locations and points in time.
- The heat map shows the concentration of offenses on a 24 hour scale. Looking at the time heatmap, we believe that the peaks in incident occurrence corresponds with rush hour. This would help in identifying which times of day might need additional resources, such as increased police presence.

Exhibit 5:

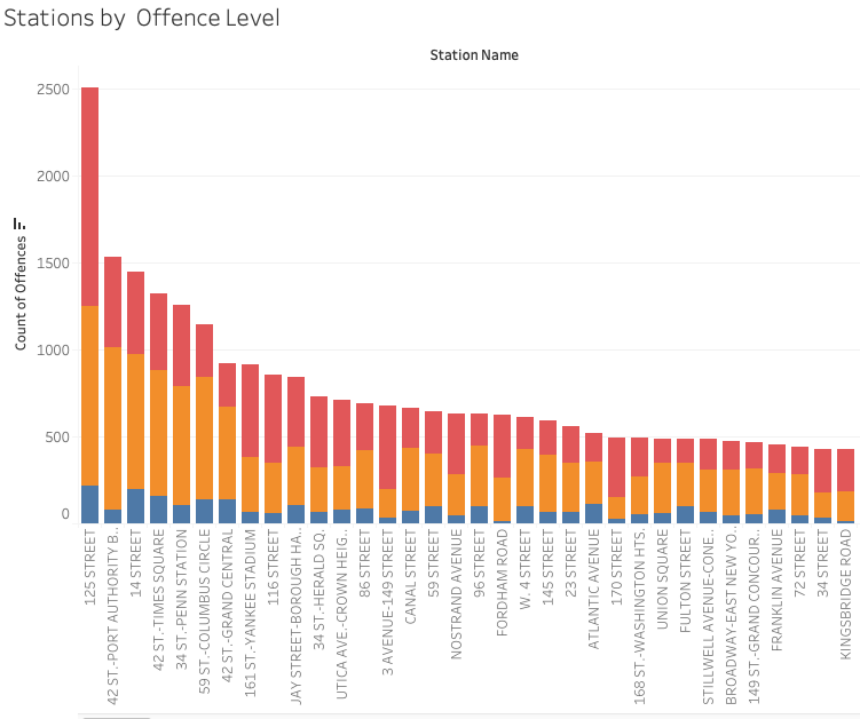


Additional Visuals:

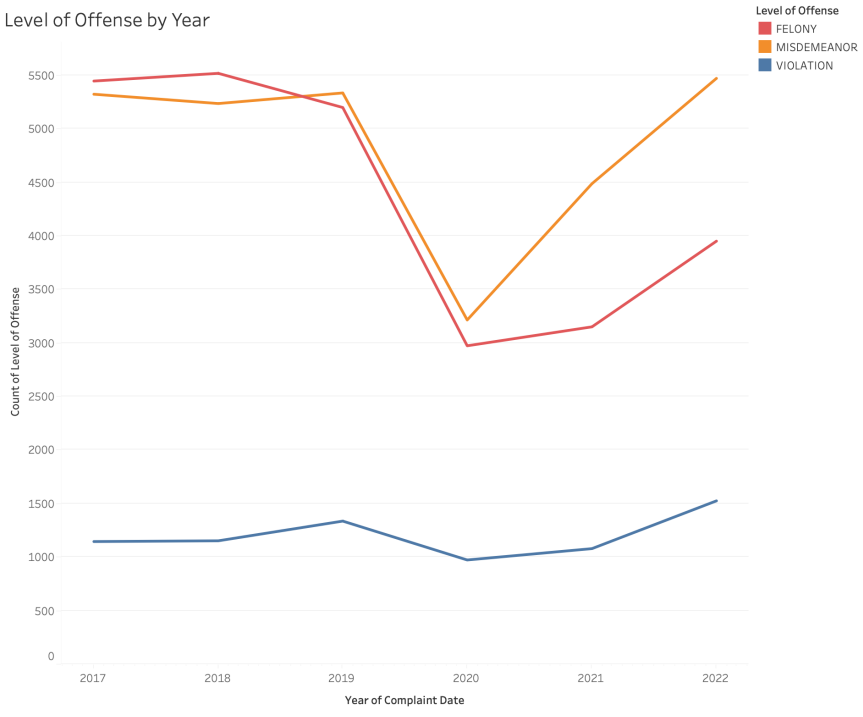
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62505 entries, 0 to 62504
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Complaint_Date        62505 non-null  object
 1   Complaint_Time        62505 non-null  object
 2   Precinct              62505 non-null  int64
 3   Offense_Type          62505 non-null  object
 4   Crime_Completion      62505 non-null  object
 5   Level_of_Offense      62505 non-null  object
 6   Dummy_Violation_Code  62505 non-null  int64
 7   Borough               62446 non-null  object
 8   STATION_NAME          62505 non-null  object
 9   Latitude              62505 non-null  float64
 10  Longitude             62505 non-null  float64
dtypes: float64(2), int64(2), object(7)
memory usage: 5.2+ MB
```

This is a Python Screenshot of the Data

- This is a bar graph showing the stations by offense level. The red being felonies, orange as misdemeanors, and blue as violations. There are over 400 stations, however, this graph captures the stations with the highest offenses.



- This is a line graph showing the Level of Offense by Year. The years shown on the graph are between 2017-2022. There is a dip in offenses in 2020 due to COVID-19. However, in 2022 crime is back to how it was pre-covid.



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 finding 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 of the crime dataset and begin data cleansing on Python and alteryx. Two members who are more experienced in Python worked together and made sure proper cleansing was done without losing too much data. This is where most of our issues arose, we hoped we could incorporate the two datasets without losing too much data and without any other complications, however, we could not fully 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 types of crimes committed. We were able to obtain several conclusions and solutions from our dataset. There are many ways we can expand our findings with more time, research and better resources. We hoped to obtain valuable information about crime in relation to subway stations and we hope that this information is useful to our fellow NYC subway riders.

Discussion:

- Advantages of the Dataset:
  - The analysis of crime data in relation to New York City subway stations benefits from several advantages offered by the NYPD Complaint Data. The dataset provides a comprehensive and detailed record of crime incidents, including information about crime types, dates, locations, suspects, and police jurisdiction.

This extensive range of data allows for thorough analysis and provides valuable insights into crime patterns and trends. Moreover, the inclusion of geospatial data in the form of longitude and latitude coordinates enables spatial analysis, facilitating the identification of crime hotspots in proximity to subway stations. Additionally, the dataset includes timestamps, enabling the examination of temporal patterns and trends in criminal activity. This temporal dimension is crucial for understanding how crime evolves over time and identifying periods of heightened criminal activity around subway stations.

- Limitations of the Dataset:
  - Despite its numerous advantages, the dataset has certain limitations that need to be acknowledged in the analysis. One notable limitation is the inconsistency of information regarding subway stations. The dataset may lack uniformity in how stations are identified or categorized, making it challenging to group and analyze crime reports effectively. This limitation hampers the creation of a more detailed and accurate map of crime patterns in relation to specific subway station locations. Furthermore, the dataset may contain instances of incomplete data entries, including null values, which can affect the reliability and completeness of the analysis. The presence of such incomplete data can introduce potential biases and undermine the accuracy of crime pattern identification and interpretation. It is important to exercise caution and consider the impact of these limitations when drawing conclusions or making policy recommendations based on the dataset.

Future Work:

- Future research should consider addressing the limitations of the dataset to further enhance the analysis of crime in relation to subway stations in New York City. Efforts should be made by the NYPD to standardize and consolidate the information regarding subway stations, improving the accuracy and consistency of crime mapping. This standardization process may involve creating a standardized naming or categorization system for subway stations to ensure consistent and uniform identification across all records. Additionally, incorporating additional datasets that provide contextual information, such as demographic and socioeconomic indicators, would enrich the analysis and provide a more comprehensive understanding of crime

patterns and their underlying causes. By integrating these contextual factors, researchers can better explore the socio-economic dynamics that contribute to crime incidents and identify potential risk factors associated with specific subway stations.

- Expanding the time range of the dataset beyond the currently available data, which ends in 2022, is another important avenue for future work. This expansion could include incorporating real-time data to capture ongoing trends and facilitate proactive crime prevention measures. Real-time data would allow for more immediate identification of emerging crime patterns, enabling law enforcement agencies and relevant stakeholders to respond swiftly and implement targeted interventions to address potential security threats. Furthermore, the inclusion of real-time data would facilitate the evaluation of the effectiveness of implemented security measures and enable timely adjustments or modifications to enhance their impact.

Conclusions:

- In conclusion, the analysis of NYPD Complaint Data in relation to New York City subway stations has provided valuable insights into crime patterns and their implications for subway security. The dataset's rich crime information, geospatial data, and temporal analysis have contributed to a better understanding of crime dynamics. Through this project, we have gained valuable experience in leveraging tools like Tableau for data visualization and employing data cleaning techniques to ensure data quality. This hands-on approach has allowed us to create meaningful visualizations that provide actionable insights. However, it is important to recognize the limitations of the dataset, such as the inconsistency of station information and the lack of contextual details. Future research should address these limitations and explore the integration of additional datasets to improve the accuracy and comprehensiveness of crime analysis. By doing so, evidence-based strategies for enhancing the safety of MTA employees and riders can be developed, contributing to a more secure and enjoyable subway experience for all.

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