The Most Optimal NYPD Subway Station Patrol Placements

Asad Ranavaya, Samuel Slobada, Dhruv Shetty

ar7182@nyu.edu, ss16249@nyu.edu, dks7920@nyu.edu

Abstract

It is a well-known fact that the New York City subway system can be quite dangerous in a city of over 8 million people. The city as well as the transit governing body, Metropolitan Transportation Authority (MTA), has taken numerous steps to curb both violent and non-violent crime throughout NYC. Some prominent measures are the random inspection of containers, NYPD patrolling of various subway cars and stations, as well as making it a felony punishable up to 7 years in prison for assaulting a transit official. Still, crime continues to be an issue and the question remains what the best use of NYPD and government resources is to best tackle this problem. Using a collection of data from NYPD's stop and frisk set, complaints, as well as arrests, we have formulated certain hotspot stations that would benefit the most from police attention. We will also speculate different timing of police officers so as to retain the highest potential of patrol in an attempt to lower the estimated violent crimes in an area.

1 INTRODUCTION

Crime in New York City is not a new phenomenon, neither is the attempt to use data to curb potential criminal offenses. The goal of this research and data analysis is to find the subway stations with the most potential to have a violent crime committed in or around it's close proximity. Many studies have been performed on the NYC subway transit system; one prominent study examines the feasibility of safety gates to prevent passengers who may jump, fall, or be pushed onto open tracks [1]. However, while NYPD is ramping up patrols on different subway stations, likely based on their own information and heuristics, we could not find specific public reports on the basis of their patrols. This is likely, though, because should this information become publicly available, its efficacy would drastically decrease as their future locations would be known.

1.1 Motivations

Anyone who has used the NYC subway system can clearly tell that it presents a great opportunity for criminal conduct. For a system used so widely by so many people, it is dirty, often unpleasant to the nostrils, and can feel very unsafe. We chose to analyze this system for betterment opportunities to both increase the safety in the public transit system but also improve the public confidence of those who travel on it. Recently there have been numerous accounts of straphangers being accosted or assaulted with little to no NYPD intervention not because of lack of care but simply because there was no police present at the time of the crime [2]. Should the subways with the highest potential for violent crime be known ahead of time, NYPD can more easily target specific stations with the resources available to them and increase overall safety.

2 GOODNESS

The data sourced to derive this analysis was supplied directly from the New York City website and aggregated by police officers in the field. We have collected and analyzed three different facets of a crime and chosen specific offenses to be filtered upon. A criminal opportunity does not exist in a vacuum, there are different stages to this process. The pretense stage can be highlighted as the potential for a crime and the past tense is observed as a crime committed. While it may make sense to derive our entire statistic from the past tense analysis, i.e., the NYPD criminal arrest data, this is ill advised. Firstly, if a station has a high complaint number of violent crimes but a low arrest number this may indicate either a red herring or that NYPD response to the complaints have been successful. Likewise, we approached the stop and frisk information in the same manner; if resources are pulled to a higher crime area, then instead of fixing the problem, resources may simply be shifted in a different direction. To maximize patrol points, we aggregated all three data sources and analyzed them as parts of a whole.

The second distinction we made was between violent and non-violent crimes. We made this distinction based on severity alone. While non-violent crimes should gain police attention, the goal of the analysis was to find the most optimal police placement where we defined optimality as reducing violent crime as much as possible. As the data sets used contain both, we decided to filter based on felonies, weapon possession, and physical force required in either stop or arrest. We also filtered out non-violent felonies from the final list by going through the remaining unique crime descriptions and removing what we deemed to be non-violent, such as the category "Tax Law". We also kept felonies without a crime description as we could not ascertain whether they were violent or non-violent.

3 DATA SOURCES

3.1 NYPD Crime Arrest Data

This dataset was sourced from the NYC government website. It contains data points collected from arresting officers that detail criminal arrests from 2006 up to June of 2022. The cleanliness of the data is ensured by the Office of Management Analysis and Planning before becoming publicly available. Each observation contains the borough of arrest, arrest classification code, identification key, arrest classification description, latitude, longitude, as well as the subway station it has been mapped to. The original dataset contains 5.3 million rows and was over 1 gigabyte in size. Figure 1 shows a sample of the cleaned dataset.

3.1.1 Data Sample

arrests (ar7182_nyu_edu.arres... borough	arrests (ar7182_nyu_edu.arrests) code_description	arrests (ar7182_nyu_edu.arres... dateof	arrests (ar7182_nyu_edu.arres... classification_code	arrests (ar7182_nyu_edu.arres... identification_key	arrests (ar7182_nyu_edu.arres... latitude	arrests (ar7182_nyu_edu.arres... longitutde	arrests (ar7182_nyu_edu.arres... subway
K	RAPE 1	09/12/2018	157	187,490,113	40.576157	-73.975984	198
K	RAPE 1	06/15/2020	157	214,204,908	40.576157	-73.975984	198
K	WEAPONS POSSESSION 1 & 2	11/21/2020	792	220,803,428	40.662563	-73.908921	218
M	ASSAULT POLICE/PEACE OFFICER	11/18/2020	106	220,672,343	40.826439	-73.950452	156
M	ASSAULT POLICE/PEACE OFFICER	11/27/2020	106	221,024,262	40.749780	-73.987781	366
B	ASSAULT POLICE/PEACE OFFICER	12/13/2020	106	221,767,153	40.903489	-73.850342	280
M	CRIMINAL MIS 2 & 3	11/25/2020	268	220,970,485	40.749157	-73.988272	145
M	ASSAULT POLICE/PEACE OFFICER	11/27/2020	106	221,024,260	40.749780	-73.987781	366
M	CRIMINAL MIS 2 & 3	12/21/2020	268	222,099,711	40.720637	-74.005211	410
B	WEAPONS POSSESSION 3	12/21/2020	793	222,107,788	40.835925	-73.921831	291
M	CRIMINAL MIS 2 & 3	12/09/2020	268	221,569,415	40.750664	-73.990867	358
K	CRIMINAL MIS 2 & 3	12/14/2020	268	221,795,661	40.682666	-73.910026	35
K	WEAPONS POSSESSION 1 & 2	11/07/2020	792	220,190,966	40.661440	-73.916434	221
B	CRIMINAL MIS 2 & 3	12/10/2020	268	221,671,327	40.827820	-73.925931	48

Figure 1: Data sample of cleaned NYPD Historical Arrest Information

3.2 NYPD Crime Complaint Data

This dataset contains NYC crime complaint data from 2006 up to June of 2022. Each observation in this set is a complaint made to the New York City Police Department. Relevant fields in the data sample figure 2 are complaint type, identification key, relevant complaint code description, date and time, latitude, longitude, as well as the subway station each record has been mapped to. The original dataset has around 7.83 million observation points and was over 2 gigabytes in size. Data points that were not kept described various other portions of a complaint and internal record keeping information, they can be found on the NYC government website for the historic complaint dataset [4].

3.2.1 Data Sample

complaints (ar7182_nyu_edu.complai... code_description	complaints (ar7182_nyu_edu.complai... dateof	complaints (ar7182_nyu_edu.complai... location	complaints (ar7182_nyu_edu.complai... timeof	complaints (ar7182_nyu_edu.complai... type	complaints (ar7182_nyu_edu.complai... classification_code	complaints (ar7182_nyu_edu.complai... identification_key	complaints (ar7182_nyu_edu.complai... latitude	complaints (ar7182_nyu_edu.complai... longitude	complaints (ar7182_nyu_edu.complai... subway
DANGEROUS DRUGS	01/08/2010	STREET	01:41:00	FELONY	117	100,009,209	40.810773	-73.952592	399
DANGEROUS WEAPONS	09/26/2007	TRANSIT - NYC SUBWAY	20:30:00	FELONY	118	100,026,305	40.576995	-73.981524	469
DANGEROUS DRUGS	01/04/2019	STREET	08:00:00	FELONY	117	100,073,193	40.697415	-73.936007	373
FELONY ASSAULT	05/08/2021	STREET	17:10:00	FELONY	106	100,170,113	40.867702	-73.919487	270
ROBBERY	11/07/2015	CHAIN STORE	14:40:00	FELONY	105	100,248,559	40.678807	-73.920349	38
DANGEROUS DRUGS	09/19/2006	STREET	23:05:00	FELONY	117	100,265,989	40.844352	-73.914744	262
ROBBERY	04/11/2008	RESIDENCE - APT. HOUSE	08:46:00	FELONY	105	100,266,488	40.706232	-73.932710	306
ROBBERY	05/27/2013	STREET	10:55:00	FELONY	105	100,281,286	40.750664	-73.990867	358
ARSON	06/06/2015	RESIDENCE - APT. HOUSE	11:01:00	FELONY	114	100,319,501	40.679337	-73.930375	298
FELONY ASSAULT	07/01/2007	STREET	23:15:00	FELONY	106	100,338,471	40.887811	-73.860453	248
ROBBERY	06/09/2018	DEPARTMENT STORE	19:16:00	FELONY	105	100,343,427	40.678807	-73.920349	38
ROBBERY	08/25/2007	STREET	23:30:00	FELONY	105	100,380,084	40.790322	-73.947687	458
ROBBERY	12/28/2020	STREET	06:05:00	FELONY	105	100,381,541	40.878933	-73.904788	267
FELONY ASSAULT	01/09/2014	STREET	20:15:00	FELONY	106	100,389,046	40.680443	-73.950359	127
ROBBERY	09/04/2018	CHAIN STORE	15:30:00	FELONY	105	100,469,302	40.840266	-73.842508	202

Figure 2: Data sample of cleaned NYPD Historical Complaint Information

3.3 NYPD Stop and Frisk Data

This dataset contains information from NYPD stop and frisk reports. Stop, Question and Frisk is an initiative by NYPD to curb criminal activity before it can occur. It is based on the legal principle of probable cause and allows an officer to confront potential criminals and ascertain if a crime was about to be committed. Information from 2003 until 2021 is provided by NYC.gov, however, most years have a different data format and data for each year is posted separately. As a result, we used data combined from 2013 until 2016 because they all had the same data format. This combined dataset had around 280 thousand observations and was over 100 megabytes in size. Figure 3 depicts a current inspection station that passengers may be subject to on their subway journey. Relevant fields kept were the stop description, date, time, latitude, longitude, and subway it was mapped to. The full gamut of information stored in these data sets is available on NYC.gov under the database file specifications [5].

Figure 3: NYPD Subway Inspection Stop. Photograph by Asad Ranavaya, 2022

3.3.1 Data Sample

Abc stopandfrisk (ar7182_nyu_edu.stopandfr... description	# stopandfrisk (ar7182_nyu_edu.stopandfr... dateof	# stopandfrisk (ar7182_nyu_edu.stopandfr... latitude	# stopandfrisk (ar7182_nyu_edu.stopandfr... longitude	# stopandfrisk (ar7182_nyu_edu.stopandfr... subway	# stopandfrisk (ar7182_nyu_edu.stopandfr... timeof
PHYSICAL_FORCE_USED	12,012,016	40.747655	-73.883800	240	315
ASSAULT	12,172,016	40.766895	-73.921391	304	210
CPW	12,192,016	40.753999	-73.942430	346	1,635
MENACING	10,222,016	40.746812	-73.891752	438	1,825
PHYSICAL_FORCE_USED	12,062,016	40.759570	-73.830139	25	1,545
PHYSICAL_FORCE_USED	12,062,016	40.759570	-73.830139	25	1,545
PHYSICAL_FORCE_USED	11,282,016	40.759570	-73.830139	25	1,334
PHYSICAL_FORCE_USED	11,232,016	40.759570	-73.830139	25	1,430
VIOLENT_CRIME_SUSPECTED	10,302,016	40.759570	-73.830139	25	647
CPW	12,222,016	40.745596	-73.903389	54	1,745
ASSAULT	11,302,016	40.748924	-73.937375	307	1,443
ASSAULT	11,302,016	40.748924	-73.937375	307	1,443
FORCIBLE TOUCHING	11,122,016	40.745596	-73.903389	54	1,000
CPW	10,272,016	40.710710	-73.793039	188	2,156
CPW	10,272,016	40.710710	-73.793039	188	2,156

Figure 4: Data sample from NYPD Stop and Frisk Reports for 2013-2016

3.4 NYC Subway Geo Data

This dataset was provided by the City of New York and contains the geographic information of each public subway station [6]. It also contained the names of the stations with service schedules and available times. We kept the names, ID's, latitudes, and longitudes. This set is main mapping point of this analytical project. There are 473 stations in total throughout the city of New York.

3.4.1 Data Sample

Abc subways (ar7182_nyu_edu.subwa... name	# subways (ar7182_nyu_edu.subwa... id	# subways (ar7182_nyu_edu.subwa... latitude	# subways (ar7182_nyu_edu.subwa... longitude
Astor Pl	1	40.730054	-73.991070
Canal St	2	40.718803	-74.000193
50th St	3	40.761728	-73.983849
Bergen St	4	40.680862	-73.974999
Pennsylvania Ave	5	40.664714	-73.894886
238th St	6	40.884667	-73.900870
Cathedral Pkwy (110th St)	7	40.800582	-73.958067
Kingston - Throop Aves	8	40.679919	-73.940859
65th St	9	40.749720	-73.898788
36th St	10	40.751960	-73.929018
Delancey St - Essex St	11	40.718306	-73.987409
Van Siclen Ave	12	40.678028	-73.891658
Norwood Ave	13	40.681520	-73.879626

Figure 5: Data sample of NYC subway locations

4 DESIGN DIAGRAM

We used the design diagram shown in Figure 6 for data profiling, cleaning, and analysis. Each data set had a map reduce job to filer out violent crimes as well as clean the information for easy use in hive, for example the latitudes and longitudes of both arrest and complaint data sets need to be broken up from a single string value. The first round of map reduce jobs also helped to gain an understanding of each data set such as count of crimes per borough. The second round of map jobs involved a replicated side join with the cached subway geography information to map each observation from each data set to a related subway. We used hive to create tables from the resulting map reduce comma separated results to be used for further analysis in Tableau.

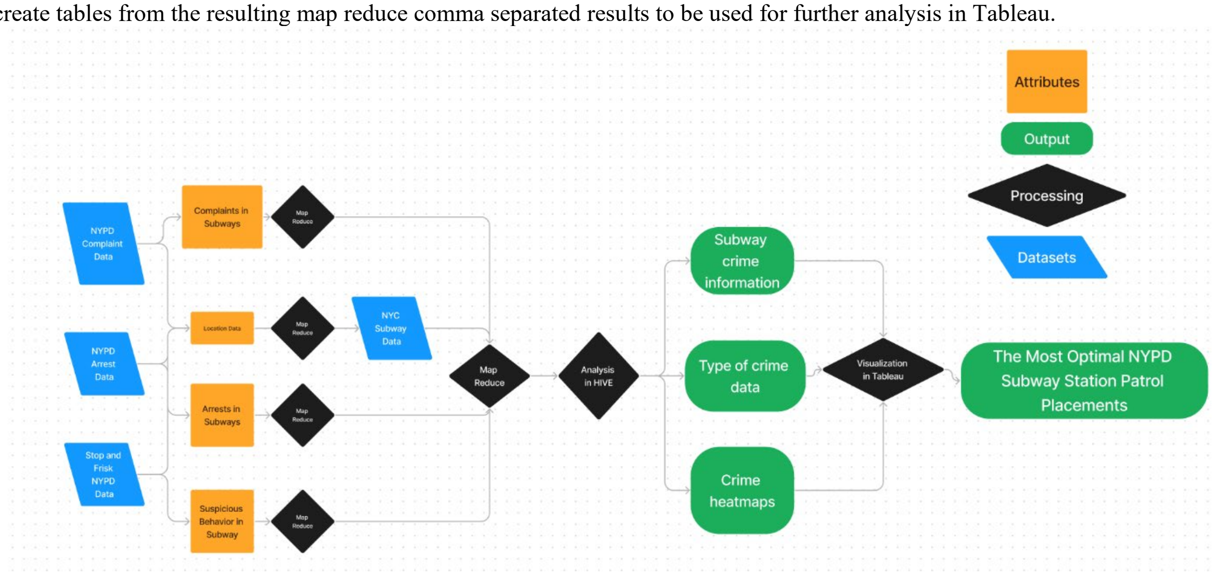


Figure 6: Design diagram of analytical process

5 CODE CHALLENGES

5.1 Mapping Coordinates

The biggest initial challenge was how to relate the datasets in a meaningful way. NYPD crime complaints, arrest, and stop and frisk datasets do not have a in or near subway field. All of the datasets had some form of relation to subway stations, but nothing was as definitive as geographical coordinates. For example, the arrest data set had a jurisdiction code field with 0 being patrol, 1 being transit, and 2 representing housing authority. While we could filter on transit, we may miss a patrol arrest near a subway station. This also still didn't map the arrest, complaint, or stop to a particular station. Therefore, we decided to map a data point to a subway based on latitude and longitude within a given fault tolerance. This was based on the idea that a reasonable patrol officer could influence or respond to a crime within one block or so. A latitude or longitude degree up to four decimal places is accurate to 11.1 meters. Based on this, we created an interval with a radius of around 50 meters, or 0.0005 degrees fault tolerance in either direction of the subway. If an observation point occurred within two different subways, it was only mapped to the first of the two subways that came up in the HashMap search.

5.2 Changing EPSG2263 to EPSG 4368

While the NYPD crime arrest and complaint dataset-maintained coordinates in EPSG 4368 (GPS), the NYPD stop and frisk data set maintained coordinates in EPSG 2263, a much more accurate coordinate system for the New York State Plane Long Island Zone. Converting this system required first identifying what EPSG 2263 [7] was, and then how to project these coordinates onto the same system (GPS/EPSG 4368) that the subway system dataset [6] used. With the use of the java library proj4j [8], available from maven, we easily overcame this issue and were able to successfully map the dataset.

5.3 Challenging Data Format and Inaccurate Documentation

While the NYPD Arrest Data and NYPD Complaint Data had just one huge dataset for all the years each, the Stop and Frisk data was split into separate files each year from 2003 to 2021. The most challenging part about this was that the column structure and data structure differed a lot from year to year. For example, from 2003 to 2016, the NYC.gov used .csv file extension for the datasets and .xlsx after year 2016. The only uniformity in these datasets was for the data between 2013 and 2016. When we tried to write another Map filtering job for the newer datasets (2017 - present), we have encountered the problem with different naming of the columns, different positions of the columns in the dataset, different format of the date, inaccurate position of the columns between the NYC.gov column description and actual dataset. Also, the description codes were very different from the dataset for the years 2013 to 2016. After experiencing all of the difficulties with the cleaning and conversation with our professor, we have decided to use only the data from the years 2013, 2014, 2015, and 2016.

6 DATA CLEANING

6.1 NYPD Crime Arrest Data

The cleaning process for this data set was quite straightforward. The MapReduce job was a mapper only filter job where we tokenized the dataset by record and then filtered based on a string split by comma values. We extracted the previously discussed key fields, discarded non-felonies, and saved the output to hdfs. The mapper output null keys and the full string line, separated by commas, as the value.

The second map reduce job was based on a separate MapReduce that found unique crime descriptions for each code. Again, this was a map only job that further filtered out non-violent felonies for the dataset. This job was also a map only job that output a null key and a value of the filtered records.

6.2 NYPD Crime Complaint Data

The cleaning process for this data involved removing incorrect data, filtering and then some profiling. The relevant columns for this analysis included the complaint number, date and time of occurrence, offense code and description, level of offense, and location information. The first step in the process was to download and store the dataset in HDFS using curl. This was followed by a MapReduce job to filter the relevant columns and remove any empty or incorrect rows. The dataset was then filtered to include only felony crimes, resulting in a reduction to 2.4 million rows. The next step was to count the unique offenses within the filtered dataset using a MapReduce job. This resulted in a count of 42 different offense codes. The dataset was then further filtered to include only violent crimes, as identified by a specific set of offense codes. This resulted in a reduction to 0.8 million rows. Finally, a MapReduce job was run to count the number of violent crimes within the filtered dataset by hour of the day. The final output included the count of violent crimes per hour, with a total of 831442 rows.

6.3 NYPD Stop and Frisk Data

Firstly, we had to concatenate data between years 2013 – 2016 (one dataset corresponded to one year). We chose this range of the years of Stop and Frisk data because the datasets had the same column structure. On the other hand, the column structure and data types for other years were very different, and the column structure description provided by NYC.gov was inaccurate. We used map only jobs to filter the following columns: date of stop, time of stop, crime description, latitude, and longitude. We kept all of the rows, which contained full date in the correct format. For the time of stop, we checked if the time was 3 or 4 digits long. If it was 3-digits long, we concatenated 0 to the beginning of time string. All of the rows with different time formats were dropped. Only the non-empty latitude and longitude rows were kept. Violent crime column filtering was one of the most challenging, when it came to the Stop and Frisk data cleaning. We wanted to keep only rows, which contained violent crimes. We considered crime as violent, if and only if: there was a weapon found on the frisked person, there was a need for physical force by NYPD officer, the column "REASON FOR FRISK – VIOLENT CRIME SUSPECTED" from the original dataset was marked as "Y", or the "CRIME DESCRIPTION" column was empty (we assumed that it is considered violent). The crime description column was determined using HashMap of violent crimes, which we have previously agreed on within ourselves.

6.4 NYC Subway Geo Data

This dataset only needed to have one MapReduce job ran to remove some unnecessary columns. This was a map only job that removed the subway station schedules and service information, keeping only the id, name, and latitude and longitude. The mapper output null keys and the station record as Text for values.

7 DATA PROFILING

7.1 NYPD Crime Arrest Data

The first statistic we were interested in was the borough with the highest violent crime count. Brooklyn came in with the highest number of arrests recorded at 74,470 arrests, with Queens at 50,787, the Bronx at 46,187, Manhattan at 45,661, and finally Staten Island with 9,130. It is interesting to note that this follows expectation as the order of boroughs by population count are Brooklyn, Queens, Manhattan, the Bronx, and Staten Island with only Manhattan and the Bronx switching places in our findings.

Another important data point was the highest arrest count by crime description. From figure 7, we can see that Weapons Possession 3 had the highest incidence at almost 3500 cases, with Weapons Possession 1&2 being the second largest category while Arson was almost non-existent with 1 recorded arrest. Another thing to note is that many crime descriptions may fall under a single code, an example being “Terrorist threat” and “Making a terrorist threat” both being considered as code 665.

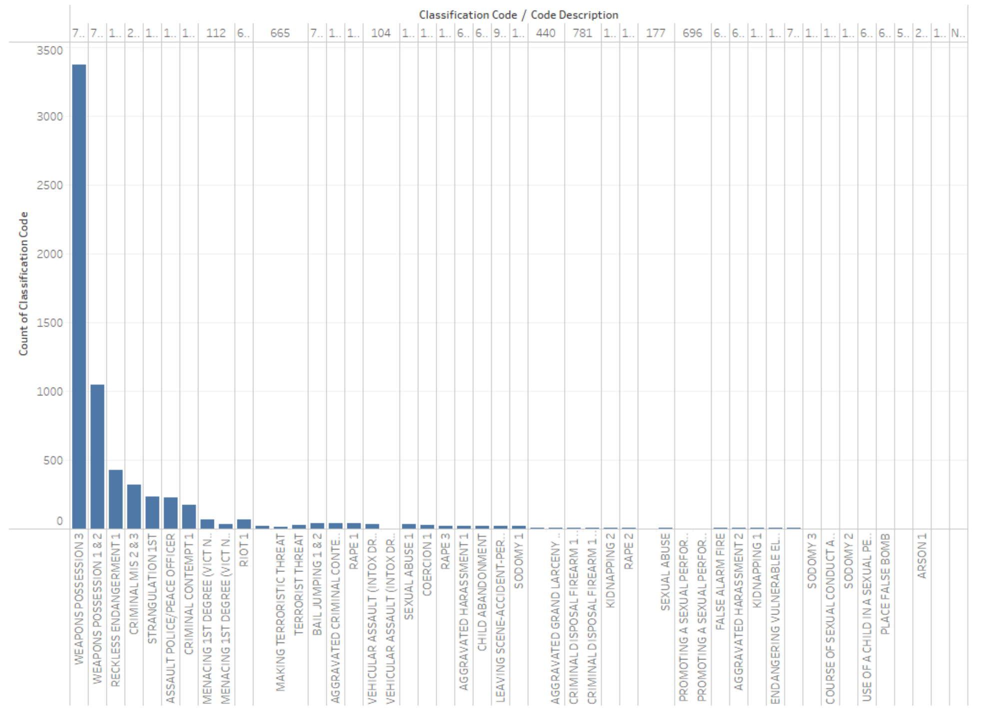


Figure 7: Arrest count by code and description

Another preliminary profiling point was to look at which subway station had the highest arrest count, indicating a future point of observation in our analysis. As shown in figure 8, Broadway station in Brooklyn is most clearly an outlying data point. While each subway has its own unique identifier, they may be grouped together in the dataset as they are often referring to a station in a central location vicinity. This is important because we mapped crime data to the first station that satisfied the distance interval, this can be seen where only one of the two Broadway station's (the G train) gets all of the data mapped to it, while the station 79 (the N and W) get's zero crimes mapped to it. This is sufficient for our analysis as the area interval is still within 50 meters in radius.

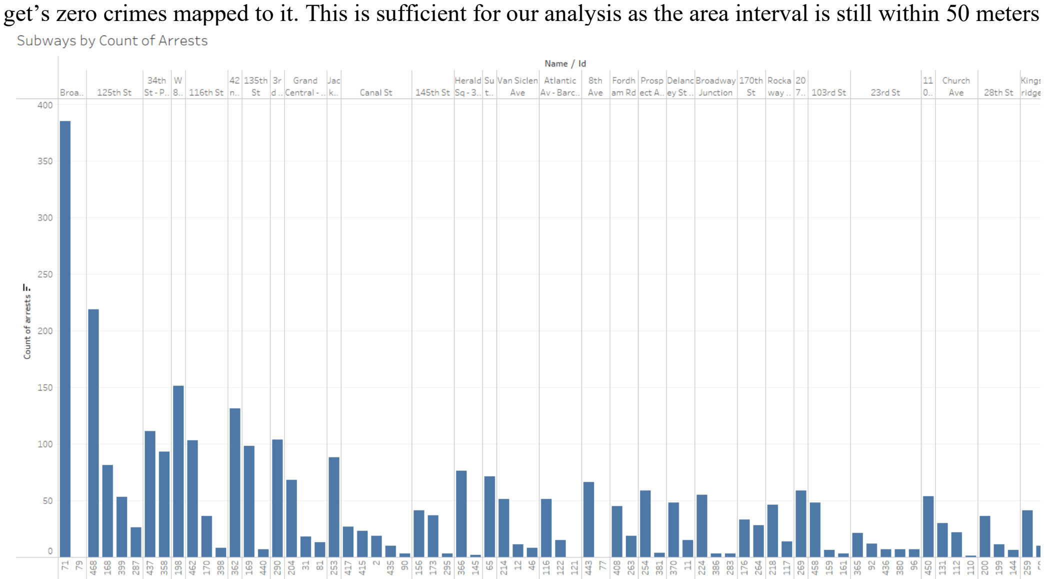


Figure 8: Top Subway Stations by Arrest Count

7.2 NYPD Crime Complaint Data

An analysis of the NYPD Crime Complaint data revealed that the stations with the highest number of complaints were those located on 125th Street and W 8th Street. These findings align with arrest data and stop and frisk data, suggesting that these stations may be areas of higher crime. To address this issue, it may be beneficial to allocate a proportional number of NYPD officers to these stations based on the perceived level of danger (as indicated by the size bubble). This strategy could potentially mitigate crime rates in these areas.

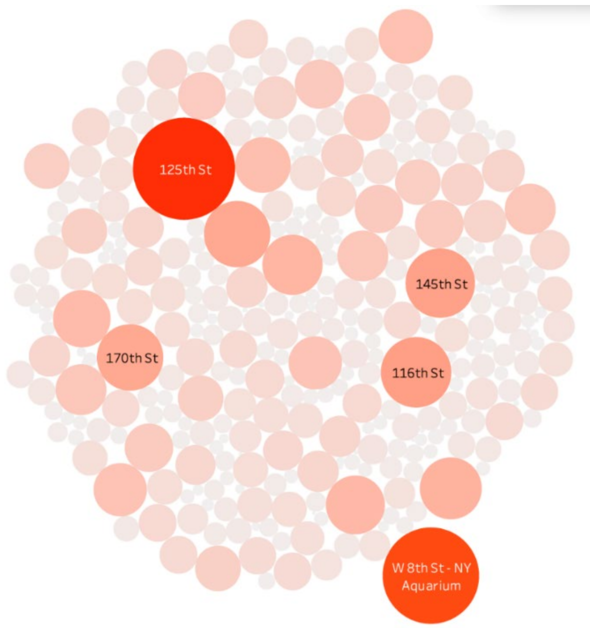


Figure 9: Stations with the highest complaint count

Further analysis of the NYPD Crime Complaint data revealed that the most frequently reported complaints were related to robbery and felony assault. These findings may be explained by the fact that individuals are more likely to report incidents of robbery in order to retrieve stolen property and incidents of assault due to the severity of the offense. Additionally, the data showed a significant number of complaints related to drug and weapon offenses. These results provide valuable insights into the types of crimes being committed in the area and can inform the development of targeted prevention and intervention strategies.

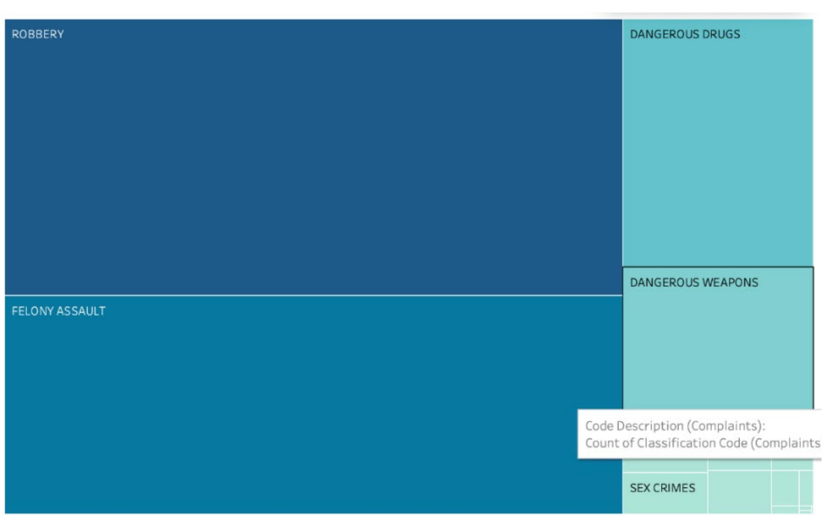


Figure 10: Highest Crime Complaint Category

To further examine patterns in crime data, the data was split by time of day, with the time periods of 12AM to 6AM classified as late night/early morning, 6AM to 6PM classified as daytime, and 6PM to 12AM classified as nighttime. The analysis, shown in Figure 11, indicated that robbery was more likely to occur during the nighttime, while assault was more likely to occur during the daytime. This information could be utilized by the NYPD in order to better allocate police resources and anticipate the types of crimes that are more likely to occur at specific times of day.

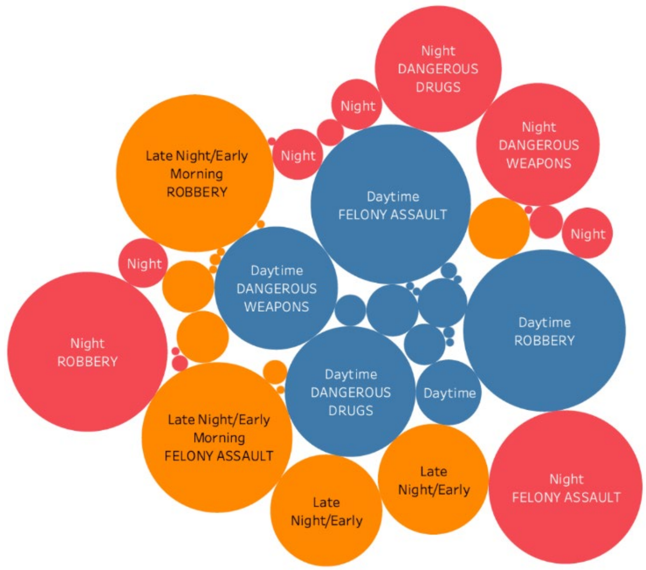


Figure 11: Complaint Count by Time of Day

7.3 NYPD Stop and Frisk Data

The two interesting outcomes of the Stop and Frisk dataset were: discrete count of violent crimes by the hour, and discrete count of the violent crimes mapped to the subway stations. Looking at the first part of the profiling, we can see that the least frequent time for the violent crimes on the subway station was early morning from 6:00 AM to 7:59 AM. The highest number of violent crimes reported on the subway stations were in the afternoon from 3:00 PM to 4:59 PM. Another spike in the violent crimes per hour are between 8:00 PM to 10:59 PM. We found this very interesting, and it confirmed our hypothesis with the evening (night) violent crimes, but we didn't predict the spike in violent crimes during the afternoon period.

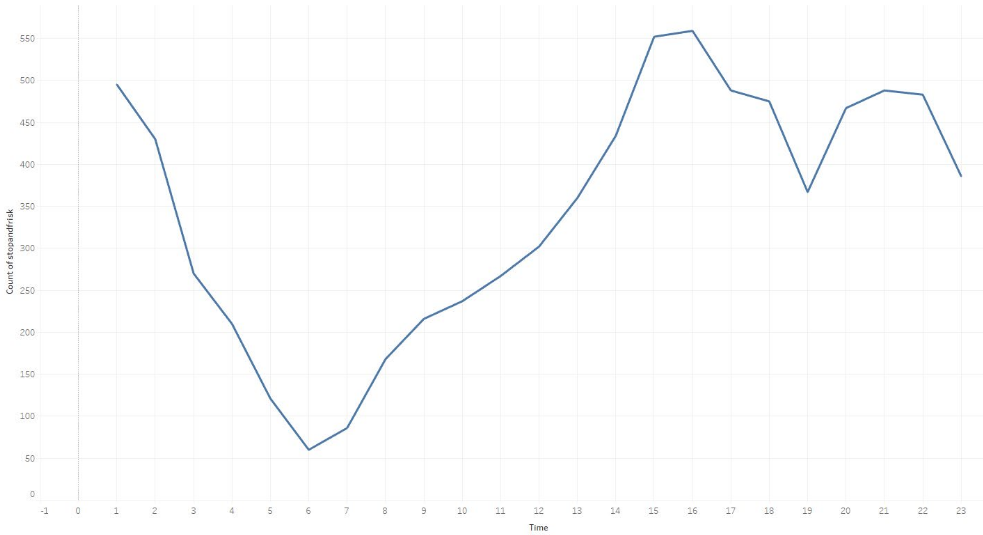


Figure 12: Discrete count of crimes by hour

The second interesting profiling point was achieved when we mapped crimes to the subway stations. We could therefore look at the most dangerous subway stations in New York City, according to the NYPD Stop and Frisk dataset. The top 5 most dangerous subway stations are 125th Street, Grand Central – 42nd Street, 42nd Street - Port Authority Bus Terminal, 34th Street – Penn Station, and 86th Street. The number below the station name in Figure 10 below is the count of the crimes corresponding to the subway station. For example, the subway station 125th Street has recorded 432 violent crimes between the years 2013 to 2016.

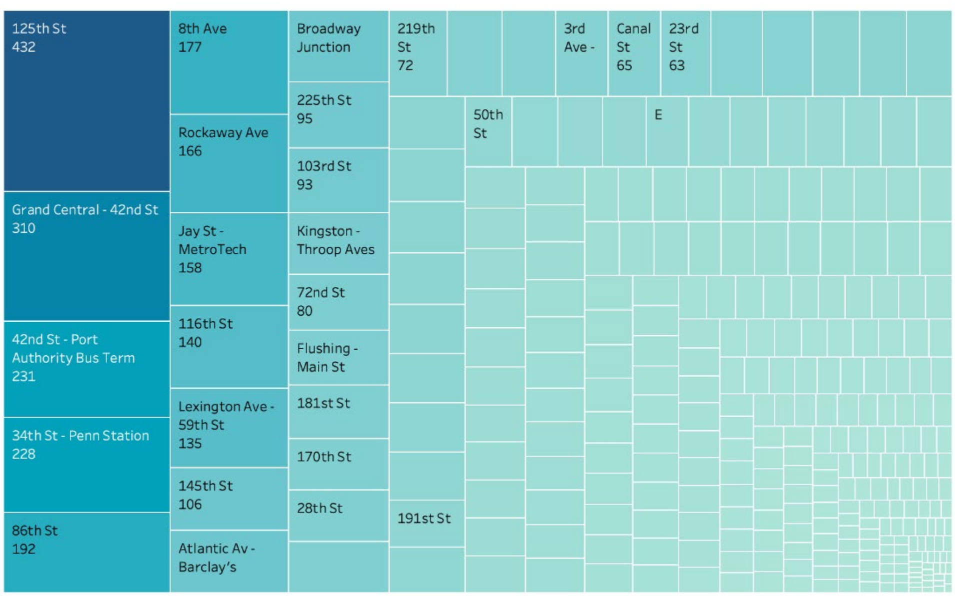


Figure 13: Discrete count of suspected violent crimes mapped to the subway stations

8 RESULTS

8.1 Metric Used for Determining Patrol

We found that, on average, the arrest data per subway station trended downward when compared with the same subway stations with higher stop and frisk searches of passengers. This metric is crucial for future results as it allows us to determine where to forecast NYPD patrol spread. We also found with this same data visualization in figure 14 that complaint information had little effect on the trend of actual arrests. This may be due to the fact that in figure 10 we can see that the highest complaint is due to robbery, where police reports are filed after the scene, this indicating that complaints do not often precede criminal activity.

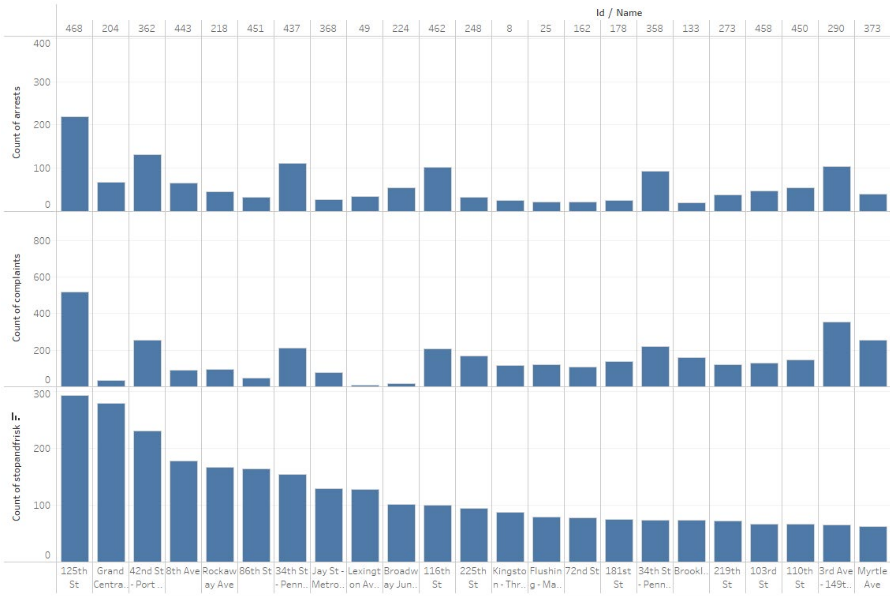


Figure 14: Combination of Arrests, Complaints, and Searching datasets sorted descending by Stop and Frisk.

8.1.1 Which Metric to Use

We initially struggled to understand how to connect the datasets given the numerous possible amounts of metrics available. We finally settled on count of criminal activity as a baseline for which we would further filter and drill down on in specific stations.

8.2 Stations With Highest Patrol Necessity

For each subway we aggregated the counts of each arrest, complaint, and stop by NYPD and visualized them in figure 15. While this provides a quick glance at the prominent stations, we highlighted the problem stations in figure 16. These are the top 10 stations by criminal arrest activity contrasted with number of complaints and people stopped and searched. As depicted in figure 16, Broadway station in Brooklyn has the highest number of arrests, with virtually no-one being searched. The rest of the arrest data trend line is almost linear after the tenth station, as a result, we recommend that the division of NYPD resources for these stations invest heavily into stop and frisk search stations to curb future criminal activity. Also, West 8th Street has numerous complaints but not search stations, which points to a possible opportunity for NYPD patrolling.

Figure 15: Count of data points by subway. From Left to Right, Arrest, Complaint, and Stop and Frisk. Darker red indicates a higher density.

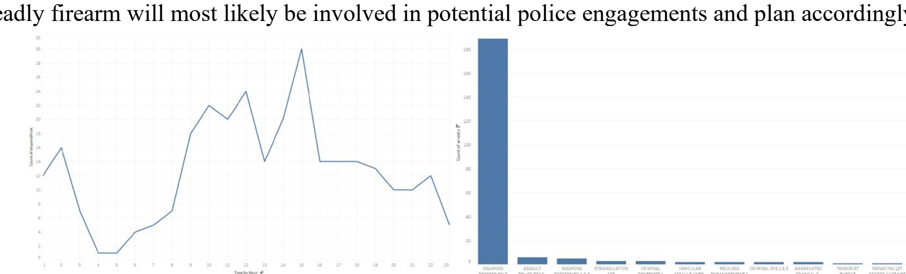
Figure 16: Categories of Arrest, Complaint, and Stop and Frisk Counts mapped together grouped by Arrests

8.2.1 Lack of Good Data for Conclusion of this Result

The problem we faced with this result is that the stop and frisk dataset from NYPD may be lacking in both data input and cleanliness. As the format changed numerous times from 2003 to 2021, there was no standard, and we fear that some stations with little to no data points may point to an issue of insufficient data. However, we chose to continue with the analysis as if the data was correct and change accordingly if new information becomes available.

8.3 Estimated Time of Patrol

We found that there are certain peak times of patrol for a given station as well as what crimes to expect for that station. In the following figure 17, we investigated the top 3 stations, 125th Street, Grand Central, and Port Authority. Based on likely searches combined with proven arrests, the predicted time of criminal activity varies by location, however, by far, the most recorded crime is weapons possession. This is important because NYPD can tailor the time of patrol based on station and reasonably expect that possession of a deadly firearm will most likely be involved in potential police engagements and plan accordingly.



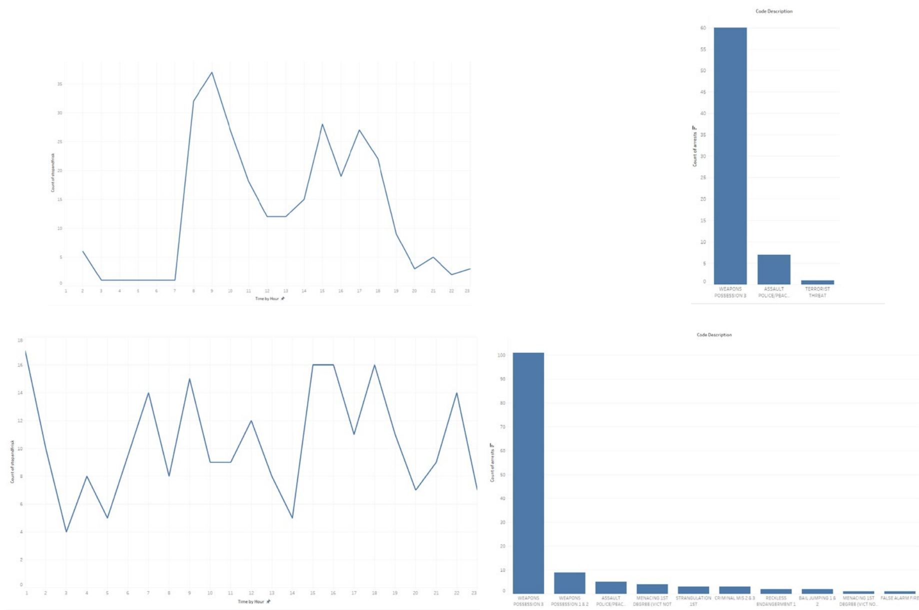


Figure 17: Suspected Crime time by hour contrasted with actual arrest data. From top to bottom: 125ths Street, Grand Central, Port Authority

8.3.1 Tableau As a Data Tool

The biggest obstacle we encountered with this result was the use of Tableau to analyze our filtered data. While Tableau is an extremely powerful to comprehend data trends, its cloudera Hadoop connection is quite slow with almost any interaction requiring a call to the Hadoop file system server.

9 SUMMARY

To identify subway stations that may require additional attention from the NYPD, we analyzed criminal arrest data, complaint data, and suspicious activity data in relation to subway locations. The analysis revealed a weak correlation between stop and frisk activity and arrest activity. Based on these findings, it may be advisable for the NYPD to increase their efforts to search for suspicious activity at stations with higher arrest counts in order to potentially reduce criminal activity in these areas.

9.1 Future Work

Potential future opportunities for this work involve adding more stop and frisk data to gain a more holistic understanding of the visualizations. Another important metric would be to use regression analysis to determine the exact extent of how much each data set influences the others.

10 ACKNOWLEDGEMENTS

We would like to acknowledge Tableau for providing us with a student license. Their software was crucial for our research as well as creating powerful visualizations. We would also like to thank OSGeo for providing and maintaining the free proj4j open-source java library to easily convert between coordinate systems.

11 REFERENCES

[1] SYSTEM-WIDE PLATFORM SCREEN DOOR FEASIBILITY STUDY (2019). NYCT, New York City, NY.

[2] Trotta, D. (2022, May 9). New York subway shooting suspect indicted on terrorism charge. Reuters. Retrieved December 13, 2022, from https://www.reuters.com/world/us/new-york-subway-shooting-suspect-indicted-terrorism-charge-2022-05-08/

[3] (NYPD), P. D. (2022, June 9). *NYPD arrests data (historic): NYC Open Data*. NYPD Arrests Data (Historic) | NYC Open Data. Retrieved December 13, 2022, from https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u

[4] NYC.gov. (2022, June 9). *NYPD Complaint Data Historic: NYC Open Data*. NYPD Complaint Data Historic | NYC Open Data. Retrieved December 13, 2022, from https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

[5] NYC.gov. (n.d.). *Stop, question and Frisk Data*. Publications, Reports - NYPD. Retrieved December 13, 2022, from https://www.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

[6] NYC.gov. (n.d.). *Subway stations*. NYC Open Data. Retrieved December 13, 2022, from https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49

[7] NYC.gov. (2022, June 9). *NYPD Complaint Data Historic: NYC Open Data*. NYPD Complaint Data Historic | NYC Open Data. Retrieved December 13, 2022, from https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

[8] OSGeo. (n.d.). *Welcome to PROJ4J*. Proj4J. Retrieved December 13, 2022, from https://trac.osgeo.org/proj4j/