Citizens for Juvenile Justice

Written by Alex Blumenfeld, Raphael Bruni, and Dan Katz

Final Report

CAS CS 506: Computational Tools for Data Science

Abstract

This report explores police interactions in three Massachusetts cities: New Bedford, Springfield, and Haverhill. Specifically, "incidents" between police officers and civilians were studied: incidents can informally be defined as police interrogations of individuals who were deemed to be engaging in suspicious behavior. Interests in the project were largely derived from attempting to determine if disparities exist among the races of the people involved in the incidents, as well as to gain a sense of familiarity of other key variables, such as which officers were involved in an abundance of incidents, common locations of incidents, and the times of day when a bevy of incidents were occurring. Some of the most notable results included the discovery of large disparities in incidents between white and Black individuals for certain data, as well as findings of how many incidents were concentrated on by a select few officers regarding the New Bedford reports.

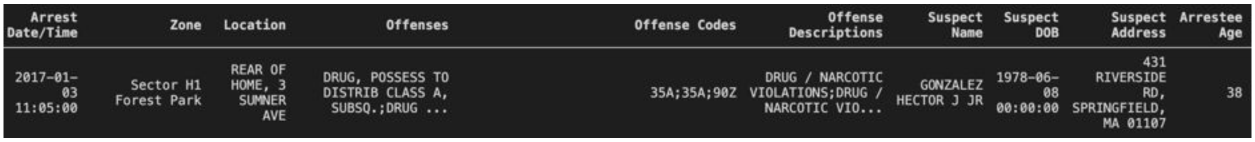
Introduction

The primary objective for the Citizens for Juvenile Justice (CfJJ) project this semester was to understand better the impact of police policy and practice in select cities of Massachusetts on the citizens living in those cities. Much of our analysis was focused on gaining a more complete understanding of incident data by factors such as the race, sex, and age of civilians, as well as the locations of police-civilian interactions. Using data provided by the client that permitted access to incident reports in the cities of New Bedford, Springfield, and Haverhill, as well as additional resources, we found several key insights that could lead to future explorations of the data expanding upon the intuition gained from this project, as well as additional questions that could be asked that may lead to a greater impact in the Massachusetts community. Many of the specific analyses that led to the most meaningful conclusions related to categorizing reports by the race and sex of the involved civilians, creating visualizations to inspect possible clusters of incidents, and examining specific age groups, specifically those involving young adults.

Data Collection, Preparation, and Cleaning

Much of the utilized data came from the Google Document sharing the course project descriptions. While plenty of data was available with which to start our undertaking, additional tasks concerning data collection and cleaning emerged throughout the semester. For example, to best analyze the Haverhill school reports, each team member needed to manually input the key attributes from a folder of PDFs containing approximately 250 different reports. In addition, to create certain maps that showed the relation between the New Bedford incidents and locations of public housing properties and public schools in the city, it was first required that the geographical locations of all of the public institutions were searched for via Google, followed by locating the latitude and longitude points of the establishments through Google Maps. In order to conduct a complete analysis of all of the key New Bedford features, it was necessary to join together two different datasets that contained the bevy of variables pertaining to the incident reports.

The dataset pertaining to Springfield comes from a page on the Springfield Police Department's website which contains logs describing the arrests performed in the city each week, all in PDF format. These PDFs are easily downloadable from the city's website, and their format is standardized enough that we could write one Python script that would download all of the files, extract the key fields from each one, and save the results to a CSV, with only a few incidents where certain blocks of text were assigned to the wrong fields. We had to experiment with several different libraries for reading in each PDF as a string, as each one resulted in somewhat different formatting. Some of the most important information in these reports includes the date of the arrest, the street address where the arrest was made, the date of birth of the suspect (from which we could calculate their age), and the list of crimes that the suspect was charged with upon being arrested. Here is an example of a portion of one arrest record after it has been retrieved from the PDFs:

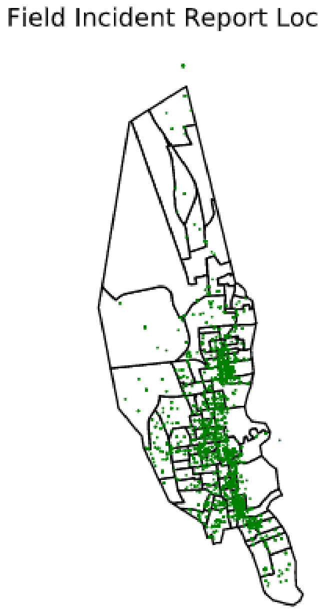


Arrest Date/Time	Zone	Location	Offenses	Offense Codes	Offense Descriptions	Suspect Name	Suspect DOB	Suspect Address	Arrestee Age
2017-01-03 11:05:00	Sector H1 Forest Park	REAR OF HOME, 3 SUMNER AVE	DRUG, POSSESS TO DISTRIB CLASS A, SUBSQ.;DRUG ...	35A;35A;90Z	DRUG / NARCOTIC VIOLATIONS;DRUG / NARCOTIC VIO...	GONZALEZ HECTOR J JR	1978-06-08 00:00:00	431 RIVERSIDE RD, SPRINGFIELD, MA 01107	38

Because these logs do not include information about the race and ethnicity of each suspect, we needed to merge the resulting CSV file with two years' worth of data that our client obtained from NIBRS (the FBI's National Incident-Based Reporting System), which is the national platform where police departments record all arrests that they make, and which lists the race of each suspect. Of the roughly 6,000 arrests in the Springfield logs from 2017 or 2018, only about 4,000 of them could be uniquely matched up with arrests in the NIBRS records -- we could only match them by date of arrest and age of suspect -- meaning we could only get race/ethnicity information for ⅔ of the arrests.

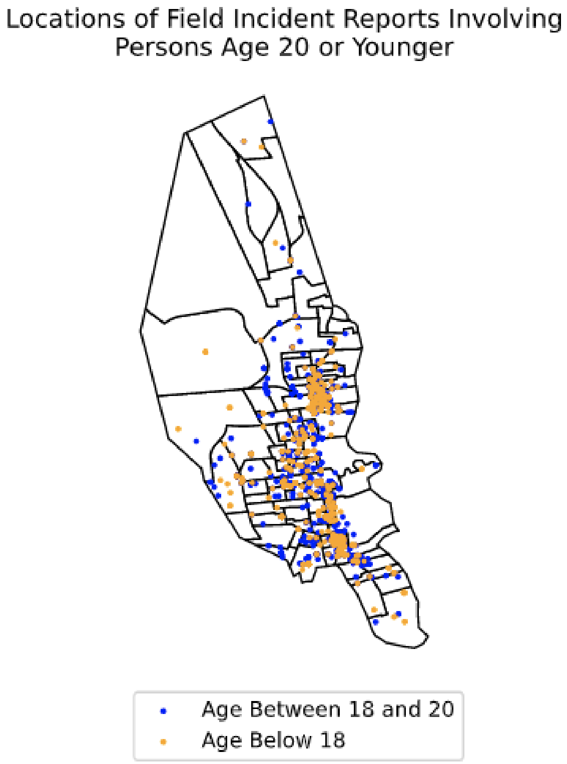
New Bedford Field Incident Reports

The first dataset that was analyzed includes all of the "field incident reports" filed by New Bedford police officers from 2015 up through several months into 2020. These reports include records of when a police officer stops someone on the street because they are suspicious that that person may be committing a crime, often referred to as a stop-and-frisk. We looked for connections between several important variables included in these police reports, such as the location of incidents, age of civilians involved, and proximity to points of interest which we believed might be related to why officers are stopping civilians. Below is a simple map of the city (divided into census block groups) with a plotting of each of the 4,997 police incidents that occurred over the relevant time span:

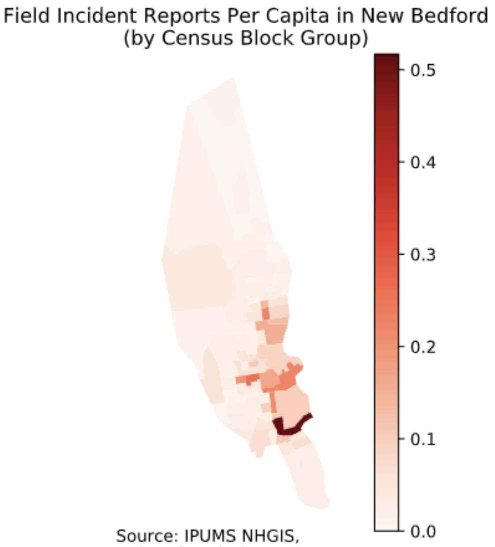


Field Incident Report Locations

There are many permutations of this data which can be used to create different visualizations, and the following maps focus on the age variable to investigate where youths are having the most interactions with police in New Bedford -- of course, this is only measuring impromptu stops initiated by officers, not arrests or reports of actual crimes. In addition, to get an idea of whether people in the 18-20 age range are being stopped in different locations than minors, we overlaid both age groups on the same map, which showed us that the two age groups are having encounters with police in similar areas of the city.


Locations of Field Incident Reports Involving
Persons Age 20 or Younger
● Age Between 18 and 20
● Age Below 18

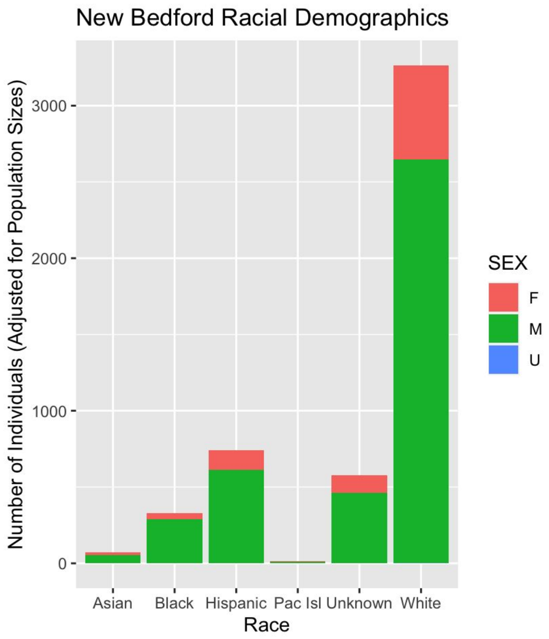
The next map resulted from an attempt to control for the populations of each census block group in New Bedford, showing how frequently field incident reports were filed relative to the total population in that area. The scale on the right represents incidents per capita: so, from 2015 through the pertinent time in 2020, the one neighborhood which is particularly dark red had approximately one field incident report for every two residents. That one ‘block group’ also had by far the most incidents in the entire city in absolute terms, showing that New Bedford police pay particular attention to people passing through that area. That particular neighborhood, Massachusetts Census Tract 6526, had a median household income of $43,405 in 2019 (according to the American Community Survey), slightly below the city’s average of $46, 321.

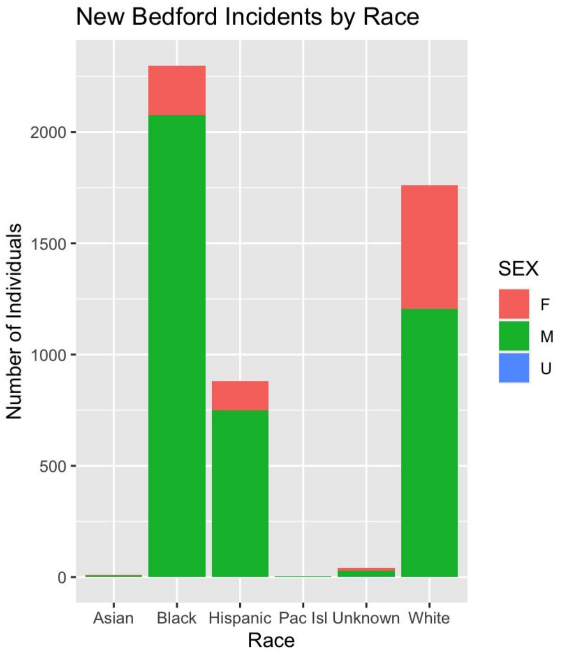


Source: IPUMS NHGIS,
University of Minnesota,
www.nhgis.org.

Race Equity

One of the essential requests by our client concerned the race of the individuals involved in incidents. Our analysis of the New Bedford data found that certain racial/ethnic groups (specifically, Blacks/African Americans and Hispanics) are overrepresented based on their respective population sizes. Despite making up under 7% of the New Bedford population, African-Americans represented greater than 46% of the unique individuals (2,299 different people) involved in police incidents from 2015 through 2020. For context, African Americans are nearly 13 times more likely to be stopped by a police officer than their white counterparts. Hispanics are only slightly overrepresented, making up nearly 18% of incidents (881 people) despite accounting for less than 15% of the population. Hispanics are just over twice as likely to be involved in police-related incidents than whites. To better understand these differences in representation, we produced the two bar charts below, one of which summarizes demographic data in the city of New Bedford, and the other of which shows the incident data by race:





In addition to the overrepresentation of African Americans and Hispanics in the dataset, some noticeable differences exist between the two charts. For one, whites make up a substantially smaller proportion of the data based on their population of the city. In addition, races that are considered “unknown” or “other” are hardly distinguishable in the dataset. This distinction could imply bias on the part of police officers’ to detect certain races of people, or it may simply be a reflection of imperfect data quality on the part of the police department.

Prolific Officers

To gain a deeper sense of specific police officers who engage in many incidents, an analysis of the most prolific officers was conducted. The following table includes statistics concerning the ten most active officers in New Bedford over the past half-decade.

Most Prolific New Bedford Police Officers[1]

Officer	Number of Professional Standards Complaints	Number of Incidents	Salary (2017)	Percentage of Incidents Involving Black Civilians	Percentage of Incidents involving Hispanic Civilians	Percentage of Incidents Involving Black or Hispanic Civilians
Roberto D.	1	459	$89,945	65.58%	18.30%	83.88%
Brian R.	3	305	$96,436	62.30%	16.72%	79.02%
Lorenzo G.	1	299	$117,832	58.19%	24.41%	82.60%
Gene F.	0	266	$86,553	60.15%	16.92%	77.07%
Clint M.	0	241	$27,106	62.66%	21.58%	84.24%
Jorge S.	0	194	$102,620	21.13%	17.53%	38.66%
Pedro M.	0	161	$113,825	50.93%	21.74%	72.67%
Samuel A.	0	148	$61,359	52.03%	26.35%	78.38%
Jason O.	0	120	$105,699	45.83%	13.33%	59.16%
Nathaniel G.	1	96	$36,397	48.96%	14.58%	63.54%

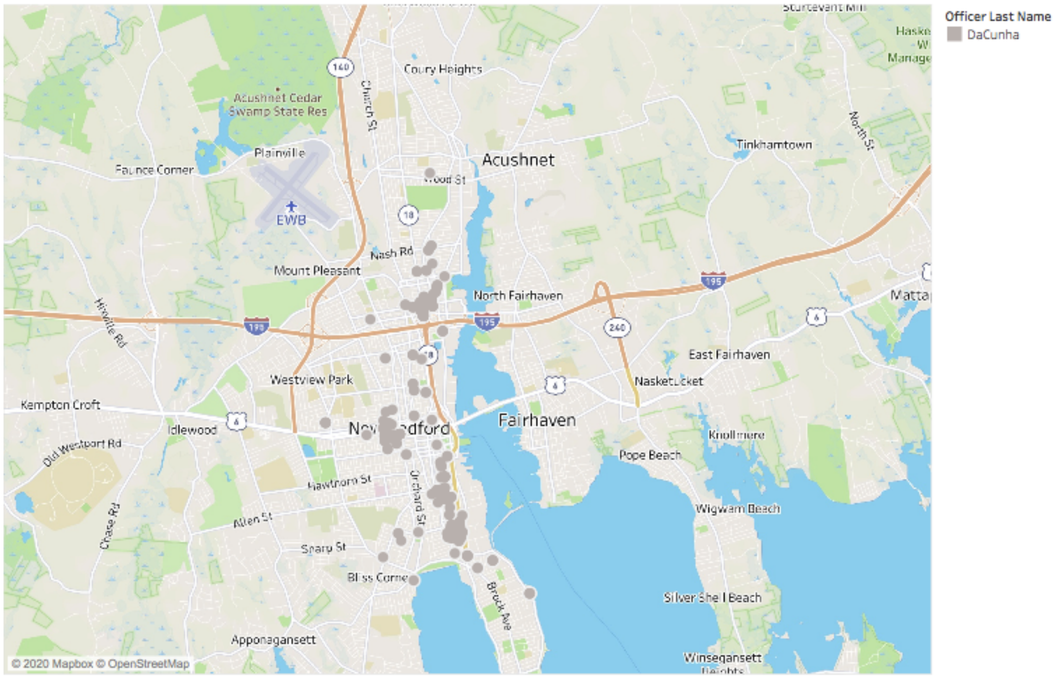
[1] Note that the officer names in the data frame are used solely for the purposes of this project.

The ten officers in the above data frame were involved in nearly 46% of all incidents, a tremendous percentage given that 186 different officers were involved in at least one incident during the time span. Moreover, the fraction of incidents that involved Black civilians was higher than the general African American makeup of the dataset (about 46%) for eight of the ten officers. The data frame also allows for us to compare the salaries of these prolific officers to the average salary of New Bedford police officers and city employees. Indeed, the mean salary for these officers was about $83,777 from the years 2016 through 2018, whereas the typical salary for a city police patrol officer as of September 2020 was $60,458, and the average salary for a city employee in New Bedford was $49,841. The relatively high salaries for these employees compared to other city employees may be an indication of the amount of work that they do (as represented by the number of incidents they are each involved in), their length of time on the force, and possibly the use of overtime. These are all possible areas of follow up research.

Geographic Analysis

It can also be interesting to visualize where many of the most active officers conduct their incidents. For instance, a visualization of incidents by the most prolific officer (Officer DaCunha) is presented below:

Sheet 1

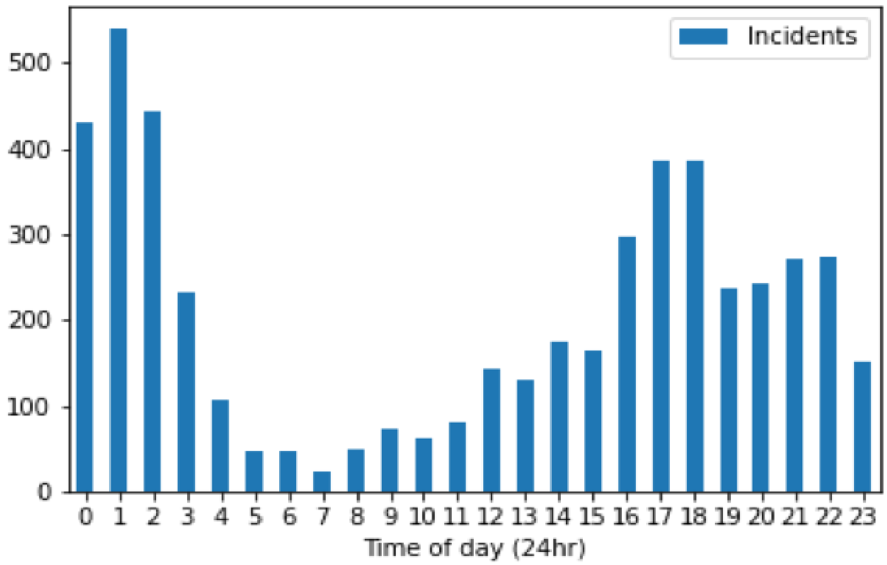


Map based on average of Long and LAT. Color shows details about Officer Last Name. Details are shown for various dimensions. The view is filtered on Officer Last Name, which keeps DaCunha.

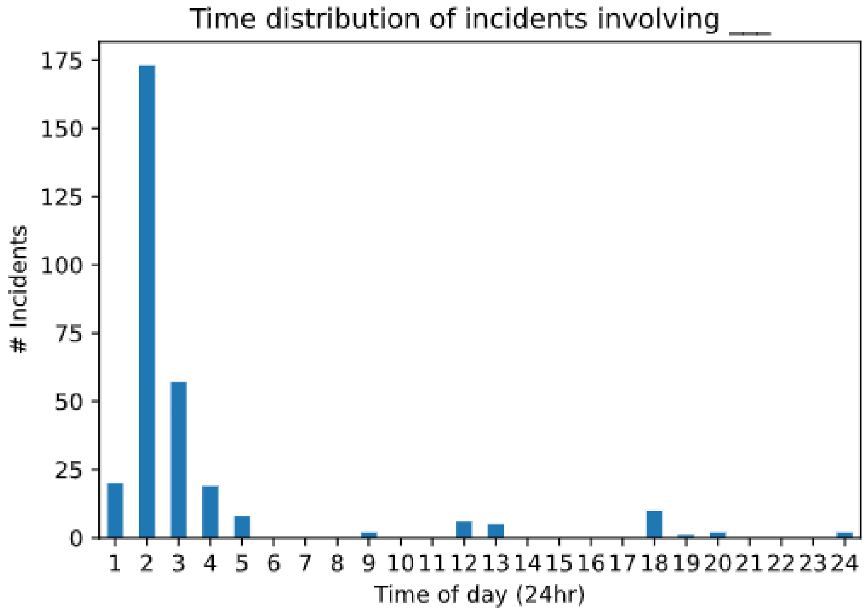
The map for Officer DaCunha is similar to that for many of the other prolific officers in that three noticeable clusters of incidents appear: one near the southern part of the city, another near the eastern section of the city, and another in the north. These regions could be highly populated, or perhaps a disproportionate amount of crime adjusted for population occurs in these areas, and thus police officers are more prone to stop individuals who appear in these neighborhoods.

Time of Day

Our client also asked us to analyze the New Bedford incidents by the time of day in which they occurred. Below is a distribution of the accumulated number of incidents by time of day:



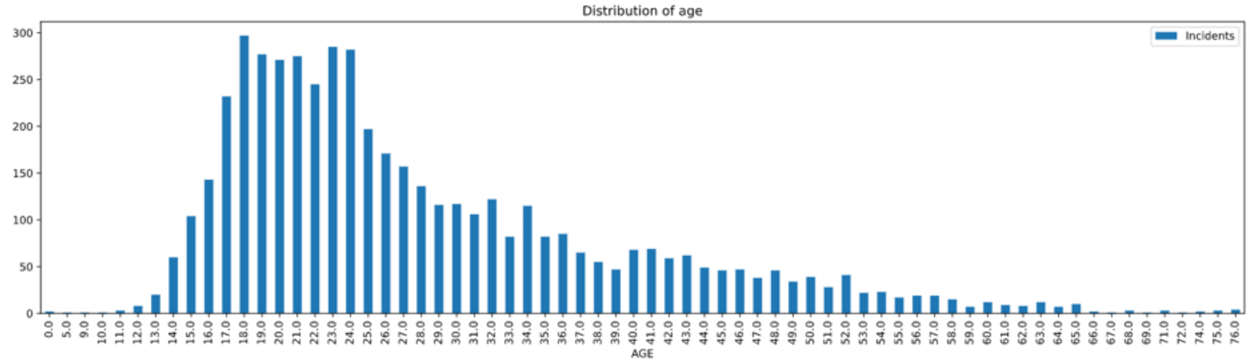
From this display, it is clear that incidents are occurring with more frequency towards the night time, with a peak at around 1 AM and a low at 7 AM. This skew could indicate that more police officers are patrolling at night, which could in turn be a sign that more suspicious behavior is occurring at those times, or perhaps that police frequently stop vehicles out at night. Another interesting observation would be to see what the incident time of day distribution looks like for the most prolific officers. Below is an analysis for one of the officers with the most stops:



As can be seen in the above graph, this particular officer makes many of his stops in the later hours, with a peak occurring at 2 AM. This spread could be an indication of the common times the officer is patrolling or simply that more "suspicious behavior" occurs during those times.

Age

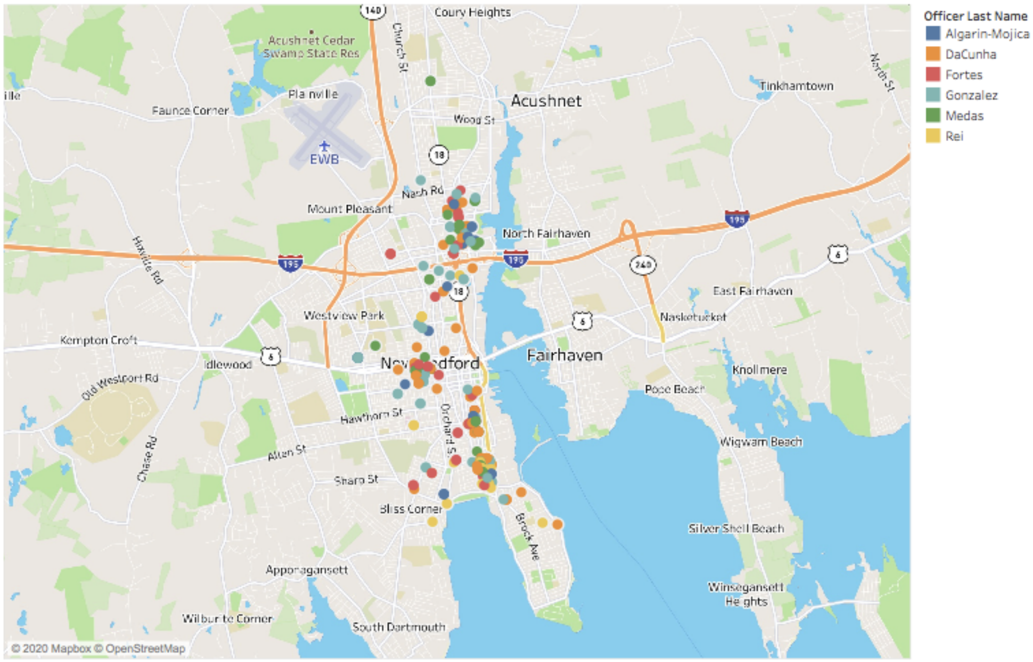
In addition to an analysis of race and time, we also specifically analyzed incidents by certain age groups. Before grouping people by age, below is presented the age distribution for all of the incidents in New Bedford:



It is clear that incidents occur with a much higher frequency from ages 18 to 24. Specifically, CfJJ is interested in examining three separate categories for young people: those aged 10 through 17, those aged 18 through 20, and those aged 21 through 25. After filtering this data even further, we found that people of these three ages (18, 19, and 20) make up a disproportionately high number of incidents in the dataset. Indeed, nearly 17% of total incidents involved individuals in that age range.

CfJJ is also motivated to understand which officers are involved in many incidents that regard young people. An analysis of the data would find that five officers (Officers Algarin-Mojica, DaCunha, Fortes, Gonzalez, Medas, and Rei) reported on more than 47% of all incidents concerning the 18-20 age group. A visualization of these six officers' work with regards to young adults is shown below:

Sheet 1



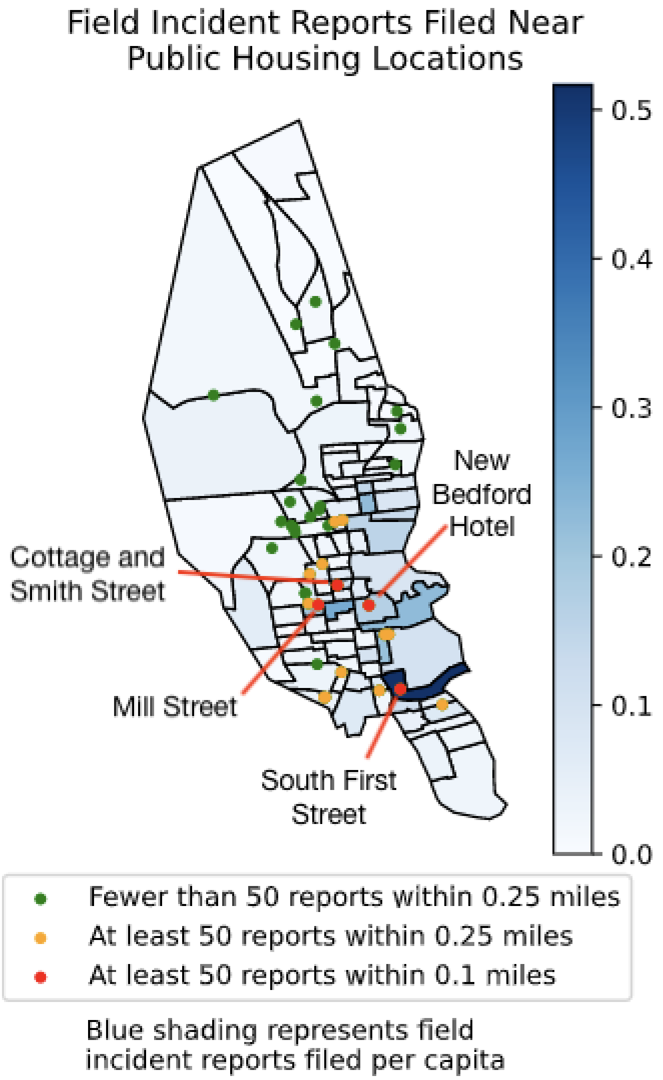
Map based on average of Long and LAT. Color shows details about Officer Last Name. Details are shown for various dimensions. The view is filtered on Officer Last Name, which keeps 6 of 99 members.

As was the case for Officer DaCunha, three noticeable clusters appear in the southern, eastern, and northern parts of the city.

Points of Interest (Public Housing and Public Schools)

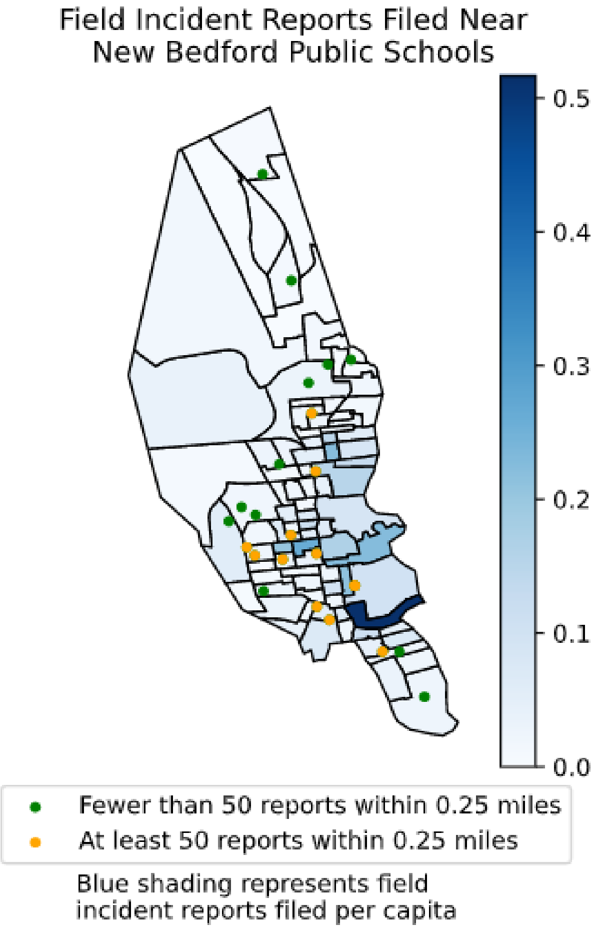
We also investigated the relationship between locations of field incident reports and locations of public housing properties in New Bedford. Plotting the coordinates of the properties in relation to the incidents could provide us with a better understanding of whether interactions between the police and citizens are more likely to occur near particular public housing areas.

Based on the following map, it is apparent that four properties featured particularly high numbers of incidents occurring very close to them, including three locations which are all quite close to one another.


Field Incident Reports Filed Near
Public Housing Locations
0.5
0.4
0.3
0.2
0.1
0.0
New
Bedford
Hotel
Cottage and
Smith Street
Mill Street
South First
Street
● Fewer than 50 reports within 0.25 miles
● At least 50 reports within 0.25 miles
● At least 50 reports within 0.1 miles
Blue shading represents field
incident reports filed per capita

This plot does certainly serve as further proof that most of the police stops performed in New Bedford are occurring in the southern region of the city, but what makes this map so interesting is that over the past five years, four public housing properties each had at least 50 incident reports filed within 0.1 mile of them, a radius of no more than a couple of blocks. Overall, we found that 13.6% of field incident reports transpited within 0.1 mile of a public housing property, and 55.8% occurred within 0.25 miles of public housing. Although none of the project members are particularly well versed in the geography of New Bedford, it seems reasonable to suspect that officers might focus their patrols on public housing properties and perform many stops near public housing as a result, and the above map provides some preliminary evidence that this could be happening.

After looking at whether police might be making lots of stops near public housing locations, we repeated the same exercise with all of New Bedford's K-12 public schools:

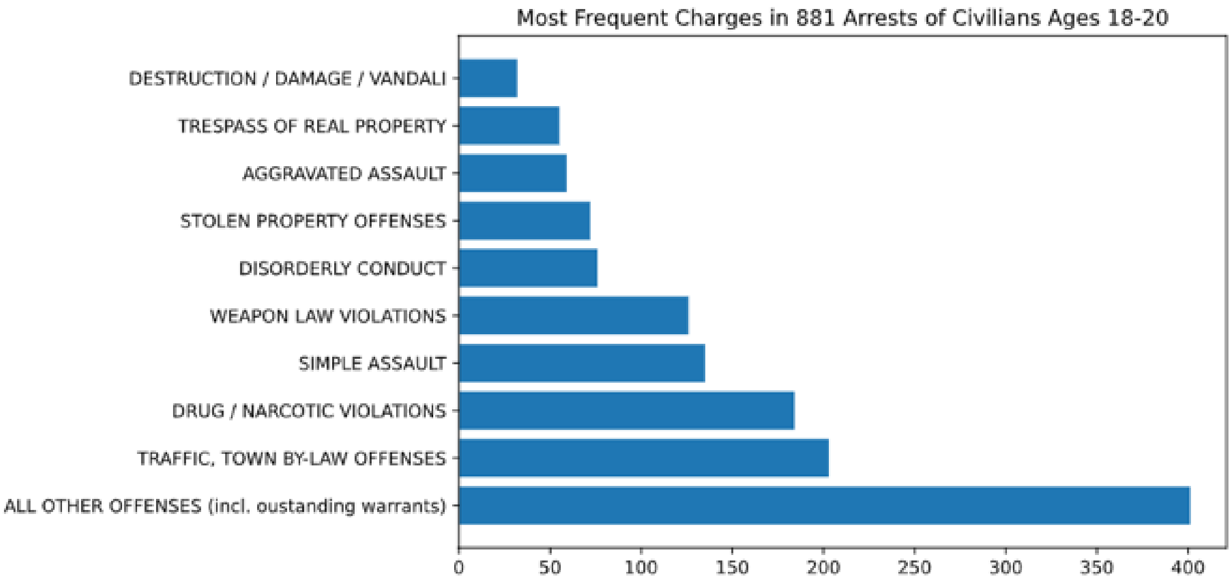

Field Incident Reports Filed Near
New Bedford Public Schools
0.5
0.4
0.3
0.2
0.1
0.0
● Fewer than 50 reports within 0.25 miles
● At least 50 reports within 0.25 miles
Blue shading represents field
incident reports filed per capita

We found that none of the city's public schools had quite as many incidents within a 0.1 mile radius as the top four public housing properties, but there were still several schools which had at least 50 incidents occur less than 0.5 mile away. The public schools followed the same general trend as public housing, in which locations closer to the center of the city had more incidents transpiring nearby than those which were farther away. As for summary statistics, it turned out that only 4.6% of field incident reports occurred within 0.1 mile of a public school, and 36.8% occurred within 0.25 mile of a public school. These results don’t exactly specify to us whether police officers are targeting the area around public schools when filing these incident reports, but they’re still useful for understanding the relationship between this type of law enforcement activity and New Bedford’s education system.

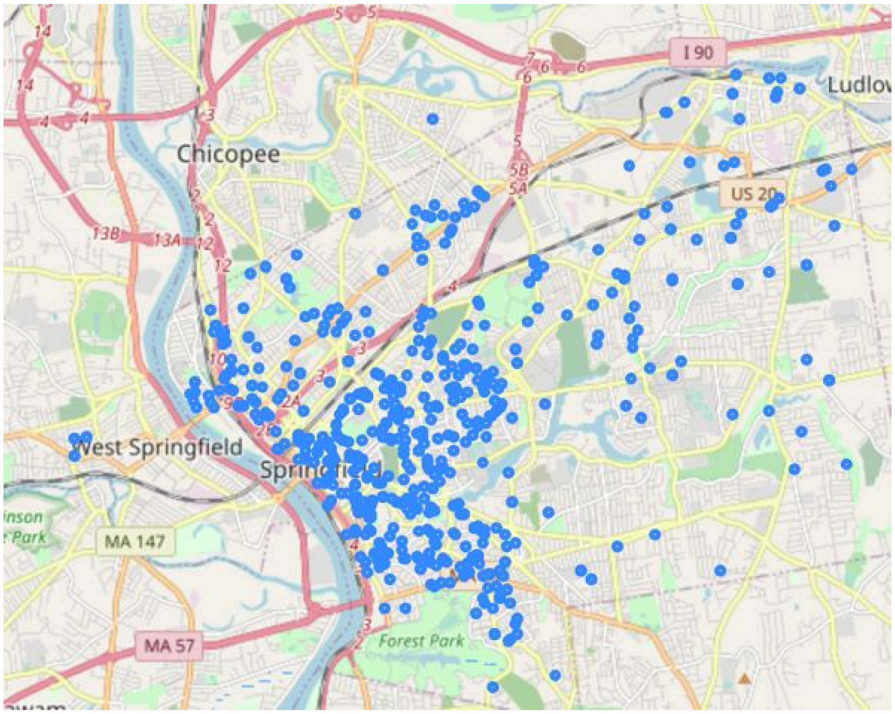
Springfield Arrest Logs

After going through the data cleaning process for the Springfield arrest logs, we focused our efforts on looking for any patterns in arrests of individuals aged 18 to 20. In particular, our client was interested in which offenses young people are being arrested for the most often, as well as the frequency with which minorities are getting arrested relative to their prevalence in Springfield’s population.

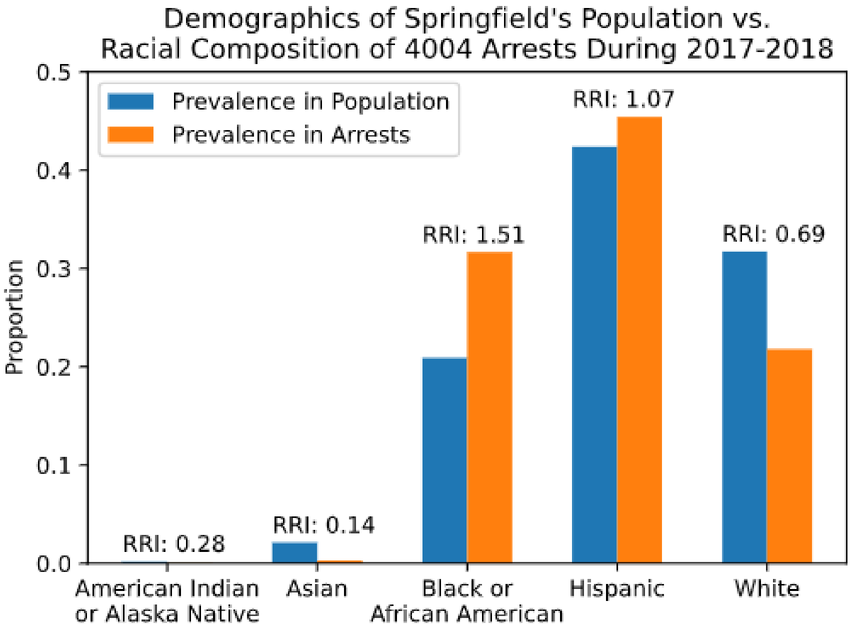
The first important result from the Springfield dataset is the graph below, which depicts the offenses that 18-20 year-olds are being arrested for in the city. Each arrest can have several charges associated with it, so this visualization represents the number of arrests which included a particular charge, not the number of arrests which were exclusively for that one charge. The ‘All Other Offenses’ designation, which is by far the most common, is usually (about ⅔ of the time) meant to denote that a suspect already has a warrant outstanding for their arrest, as the existence of a warrant is listed as a separate charge in the Springfield police department’s public logs.



As for some basic mapping of the Springfield arrests which involved youths, the maps on the next page plots the locations of arrests in Springfield which involved a suspect who was between 18 and 20 years old. Because the street addresses recorded in the reports are imperfect, we were only able to get latitude and longitude information for about 85% of the 881 arrests in question, and it is this portion of them that are displayed below. Just like in New Bedford, most of these arrests are occurring near the downtown area of the city.



As mentioned earlier, once we were able to merge the Springfield arrest logs with the NIBRS data, we found that only about 63% of the arrests from the Springfield PDFs (from the years 2017-2018 in particular) could be uniquely matched to a NIBRS record, but this still gave us a pool of about 4,000 arrests for which we knew the race and ethnicity of the suspect. At this point, the most useful direction was to calculate a “relative rate index” for each main racial group, which captures the extent to which a particular group is over- or under-represented in the arrest figures relative to its frequency in the overall population of Springfield. A display of these calculations is shown below:



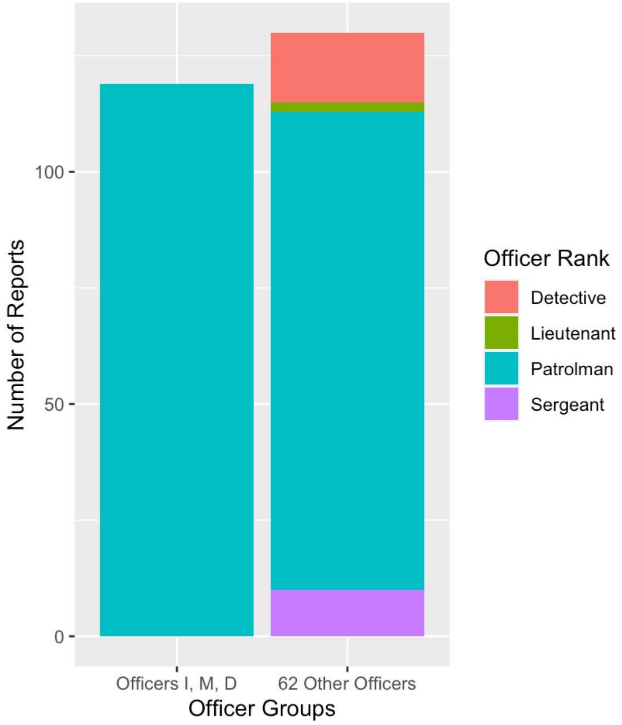
*Proportions do not quite add up to 100% because suspects of unknown race are omitted
**RRI refers to Relative Rate Index

So, we can see that Black/African American residents are the most overrepresented group in this collection of arrests, as they made up only 21% of Springfield's population in 2018, but accounted for roughly 32% of this subset of the arrests performed in 2017 and 2018. As it turned out, Hispanics, who are the single largest ethnic group in Springfield, appeared in the arrest logs at a rate roughly equivalent to their frequency in the population, telling us that most of the overrepresentation of Black suspects is compensated for by a relative underrepresentation of white suspects. This shows a much lower level of overrepresentation of Black residents than what occurred in the field incident reports in New Bedford, where the proportion of people being stopped who were Black was nearly 8 times the proportion of Black residents in the city's population.

Haverhill

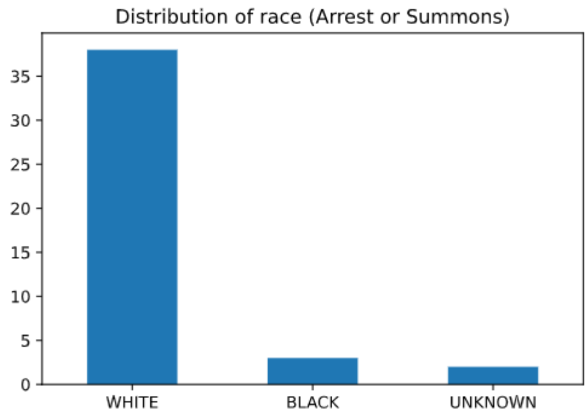
One of the most noticeable aspects of the Haverhill data was the overwhelming amount of missing data for several features. The primary reason why a lot of reports did not include information relating to each variable was because there was no need to include such information. For example, in an incident where a student or teacher reported that an item of theirs had been stolen, the officer writing the report might not yet know who the offender was. In addition, much of the information about the offenders was redacted by the Haverhill police department. Consider, for instance, the "gender" feature: of the 249 incident reports, more than half (approximately 53%) did not declare the gender of a potential offender. Of the reports that did include the feature, a substantial majority of them (greater than 77%) involved male offenders as opposed to female offenders. There was an even larger disparity in the race of offenders: nearly 95% of such reports were associated with white offenders, while the remainder were linked with Black offenders.

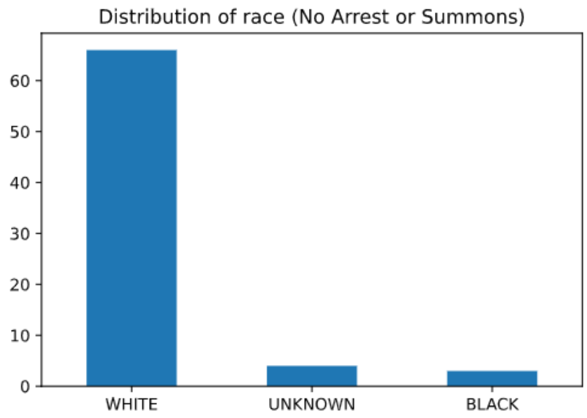
An additional imbalance prevails in the particular officers who report each incident. While the 65 different officers involved the 249 incidents may lead to the viewpoint that a wide variety of officers were working with the Haverhill school, three specific officers accounted for a large number of the reports. Consider the bar chart on the ensuing page.



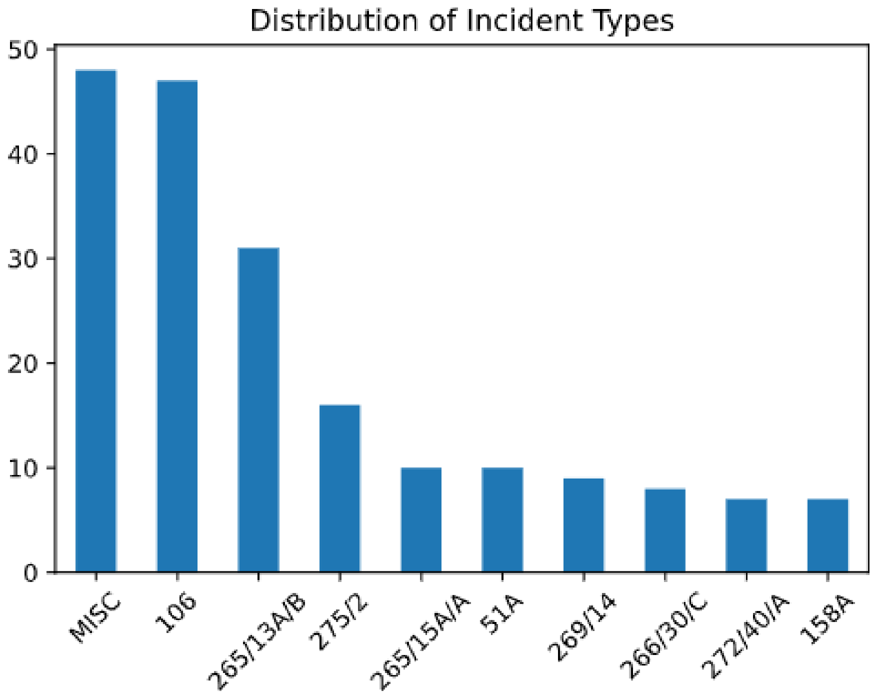
As the left column shows, many of these reports were filed by the "school resource officer" assigned to a particular school, which is why only a few officers account for so many of the reports.

To gain a more contextual insight into the Haverhill dataset, we analyzed the Narrative column of the dataset which describes the officer's account of the incident for key terms. One of these key terms was "summons", and another was "arrest". A summons is a court order to appear in court. We counted the number of incidents which contained either of these keywords, and in total 45 out of 249 contained either of those keywords, which constitutes about 18% of all of the reports. Then, given this more granular context, we conducted an analysis of offender race. Below are two graphs of the distribution of offender race, one for incidents where there was an arrest or summons, and the other where there was neither a arrest nor a summons:





Lastly, to understand which types of incidents were occuring in the Haverhill incidents dataset, we analyzed the frequency of the ten most common incident types. Incident types are standardized by state, and they essentially categorize the incident which occurred. Below is a bar chart distribution of those ten most common incident types:



Ignoring "MISC", which is an incident that fits into no specific category (miscellaneous), the most common incident type was "106", wich is a medical report incident, followed by

“265/13A/B”, which is assault and battery. The third most common incident type was “275/2”, which indicates a threat to commit a crime.

Conclusions

Perhaps the biggest takeaway from the project was the significance of disparities in certain incident reports. Specifically, data from the cities of New Bedford and Springfield proved to include large differences in reports by white civilians and Black civilians. These disparities are especially noteworthy for the New Bedford reports, and while racial differences in the Springfield arrests were less drastic compared to those in New Bedford, they exist nonetheless. Other important conclusions can be derived from examining some of the most prolific officers, the locations of where they typically conduct their work, and the demographics of the individuals they generally involve themselves with. Specifically, for some of the most notable New Bedford officers, a comparatively high number of incidents involved minority citizens.

Limitations and Future Research

Quite a few limitations exist with the Haverhill dataset. The first and most restrictive issue regards the size of the dataset: in total, we received 249 incidents report PDFs from the Haverhill police department. Due to this small sample size, extracting meaningful statistics would be difficult and possibly misleading. So we cautioned our client to interpret the analysis we made with the Haverhill dataset with a grain of salt. In addition, due to the limited number of incidents, a lot of information was missing due to redaction, and made the analysis even more difficult to conduct.

Regarding the Springfield dataset, the main obstacle we faced was that we couldn’t obtain perfect race/ethnicity information for the suspect listed in each arrest. Because only about 4,000 of the roughly 6,000 arrests conducted in 2017 and 2018 could be precisely matched up with race/ethnicity data in the NIBRS reports, we are unsure if the demographic profiles of the merged- and non-merged arrests are the same. This situation makes it tougher to claim that statistics regarding the racial distribution of successfully merged arrests are generalizable to larger populations -- if the omitted arrests don't have the same rough demographic distribution as the ones which could be exactly matched up with records from NIBRS, there could be bias in the "relative rate indices" calculated as part of the race equity analysis for Springfield.

Based on the project in its current form, a number of directions can be taken in the future to expand upon the current analysis. In addition, much qualitative work can be done to gain a better sense of the characteristics of the demographics of interest. Rather than solely collect, clean, and analyze data, the data scientists working with the data should gain a better understanding of the actual people representing numbers in a chart. A more coherent understanding of the data can come in the form of holding discussions and surveys with the residents of cities like New Bedford and Springfield. Grasping how many of these citizens feel about their local police departments and past experiences they've had with specific officers could lead to more accurate hypotheses to be made about the data. In a time when many social justice activists are holding discussions regarding police reform, it could be intriguing to contemplate how changes to police forces at both the state and city levels could affect incident reports. On the quantitative side, it might be ideal to research how many years certain prolific officers have worked in their given departments, as well as the various positions and units they have been assigned to in their employment history. It could also be worthwhile to inspect if any civilians were involved in an inordinate number of incidents, and if any specific incidents led to arrests.