Group #14

Final Project – Milestone 3

Data and Cleaning

We sourced data from:

- Washington Post's Police Shootings repository (https://github.com/washingtonpost/data-police-shootings)
- US census data (census.gov, https://www.census.gov/quickfacts/fact/table/US/PST045219)
- Police use of force dataset for the largest 100 US police departments from an activist group named Campaign Zero (http://useofforceproject.org/#review)
- The most recent 2016 release of Law Enforcement Management and Administrative Statistics (LEMAS) dataset from the Bureau of Justice (https://www.icpsr.umich.edu/web/ICPSR/studies/37323/summary#)
- List of victims of fatal encounters with the police, from mappingpoliceviolence.org

In general, for data cleaning, we replaced place holders for missing or N/A data entries with NaNs (if any), and dropped observations with NaNs. We also merged many datasets together, ensuring that they are joined on the same index (i.e. matching city AND state name). Additionally, for LEMAS & census data, we used FIDS to match county-wise demographic information with police departments based in those counties. Victim statistics compiled from the "mapping police violence" (MPV) dataset for each police department were also merged with LEMAS using unique ORI9 IDs corresponding to individual police departments. For initial EDA, MPV entries with multiple or unknown police departments responsible were ignored. State-level police departments were excluded from comparative analysis of police/victim/county racial composition.

Exploratory Data Analysis

Victim demographics & geographic features

We wanted to get an idea of the demographics of the shooting victims. First, we used data from the Washington Post, which included the following categories: victim name, gender, age, race, date of shooting, city and state where shooting took place (and latitude/longitude if available), manner of death, whether the victim was armed, whether the victim showed signs of mental illness, threat level to the officer, if the victim fled, and whether the office was wearing a body camera. We examined the distributions of many of these variables by making barplots. We also looked at the top 10 weapons that victims were armed with, if any. Next, we were interested in the distribution of shootings by age, and we also split this up by race. Because there appeared to be systematic differences in this analysis, we used US census data for race as a fraction of the total population to plot people killed as a proportion of their race. Although we found datasets with median household income, education level, and percentage of people below the poverty line, we were unable to analyze them prior to this milestone submission, but these would also be interesting factors to examine.

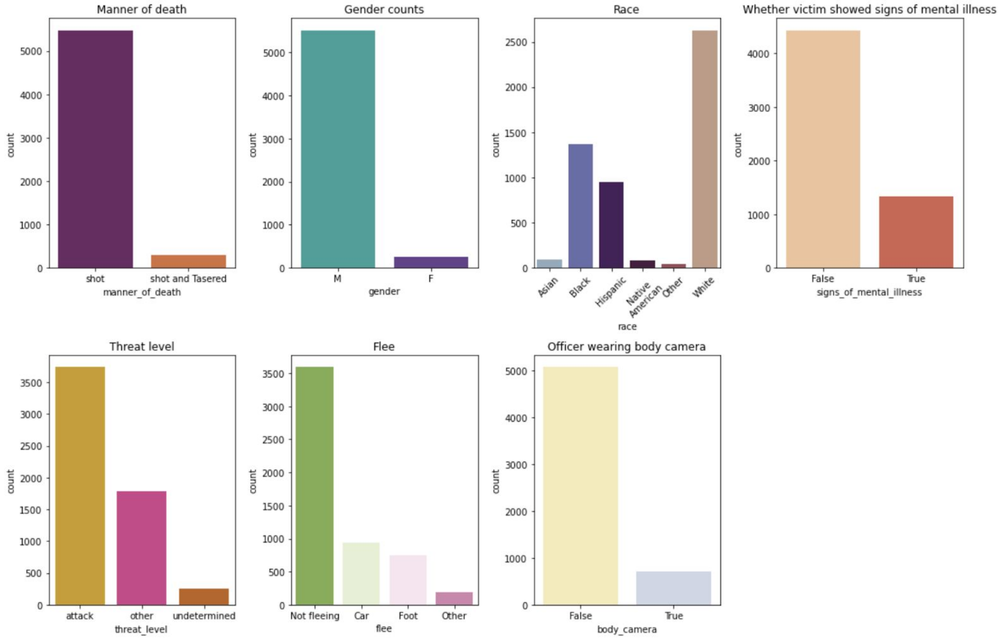
After focusing on information about the victims, we transitioned to looking at the time and geographic distribution of shootings. We checked whether shootings were more likely to occur on a given day of the week, month, or year. We next looked at shootings per city (top 10) and ranked by state, first without normalizing by population sizes, and then normalizing using US census data. To visualize this, we created a choropleth map of the number of shootings per 1 million people, and overlaid coordinates of each shooting.

Police policies on the use of force, body cameras, and training requirement

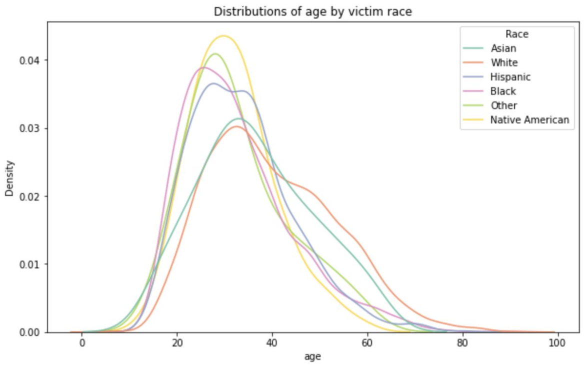
We also wanted to look at how police use of force and other policies relate to police violence. We collapsed the Washington Post dataset to obtain the number of police shootings per city and combined this with the use of force policy dataset for the largest 100 US city police departments. Regardless of whether a police department has adopted a use of force policy or not, the rate of fatal police shootings is relatively the same. This is interesting as it does not align with what the use of force project reports. This may be due to the fact that their analyses were performed in 2016 when a small number of departments had implemented at least one of these policies. We are aggregating the number of police shootings from 2015 to present day, and a significant portion of these police departments have implemented use of force policies since the original report was published.

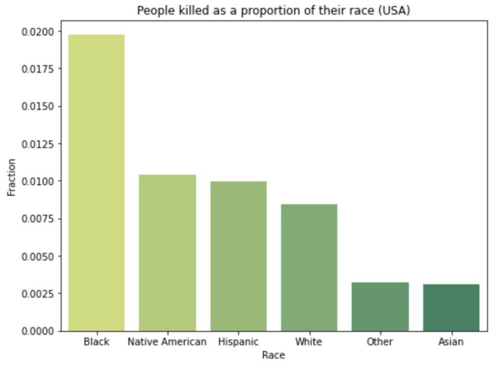
We also analyzed the percentage of fatal shootings in the Washington Post dataset with body camera usage at the state and city level (we only looked into cities with the 100 largest police departments from the use of force dataset, since there are over 2500 cities in the Washington Post dataset and most cities have very few fatal police shootings), and compared this with the rate of fatal shootings per 1 million population calculated using the census data. In addition, we looked into how police training requirements (both in academy and in field) relate to the rate of fatal police shootings in states and cities using the LEMAS data. The LEMAS dataset contains more information on police policies that could be further explored.

Preliminary Findings and Visualizations

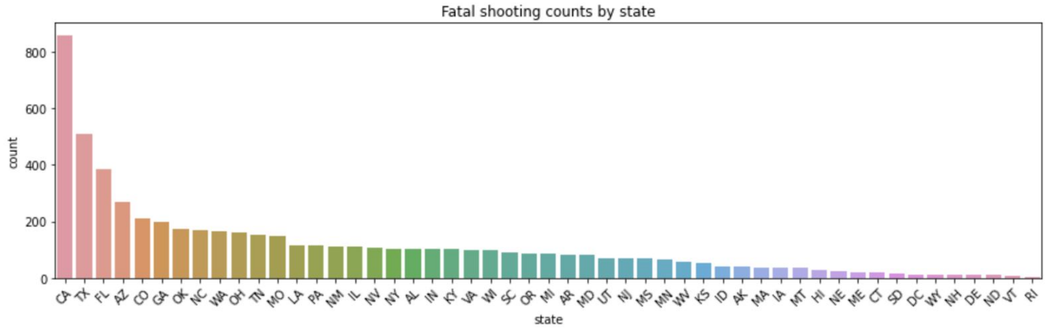


Examining many of the categories individually, we see that a majority of police violence victims were shot, with a small fraction both shot and tasered. Almost all of the victims were male and predominantly white, followed by black, hispanic, and a much smaller fraction were Asian, Native American, or other. Most victims did not show signs of mental illness. For threat level, a majority of the victims were perceived as attacking, although one caveat is that this metric is likely biased towards whatever the police officer reported (and thus is unlikely to say that a victim was not attacking, even if this was actually the case). Most victims were not fleeing at the time of the shooting. Finally, officers were generally not wearing body cameras. When we examined the most common weapons for victims to be armed with, the top 4 were gun, knife, unarmed, and toy weapon (plot not shown here), which would suggest that often the victim was not a threat; this could be interesting to examine in further detail.

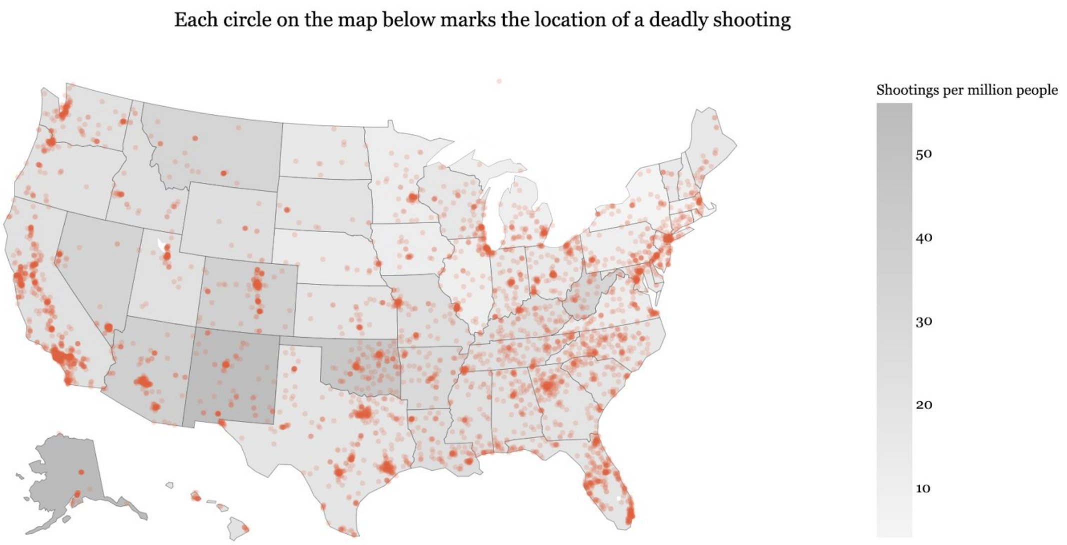




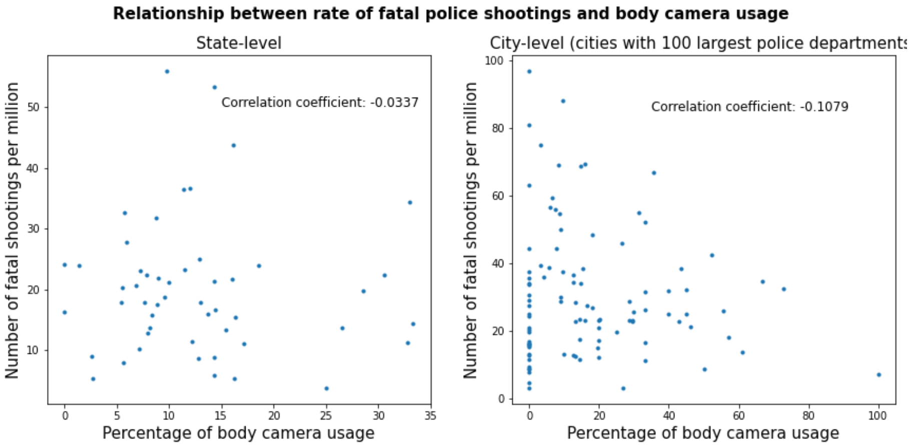
The median age of shooting victims was ~30 years old, with the distribution skewed right. When splitting up the distributions by race, it became clear that some distributions were skewed further left (i.e., younger victims) than others. Black victims tended to be the youngest, with Hispanic, other, and Native American also tending to be younger. The ages of Asian and white victims were skewed furthest rightwards. This suggested that shootings disproportionately affect victims of different races. When using US census data for the number of people of each race and using those values to normalize the number of victims, we found that Black victims were killed at a much higher proportion of their race than others.



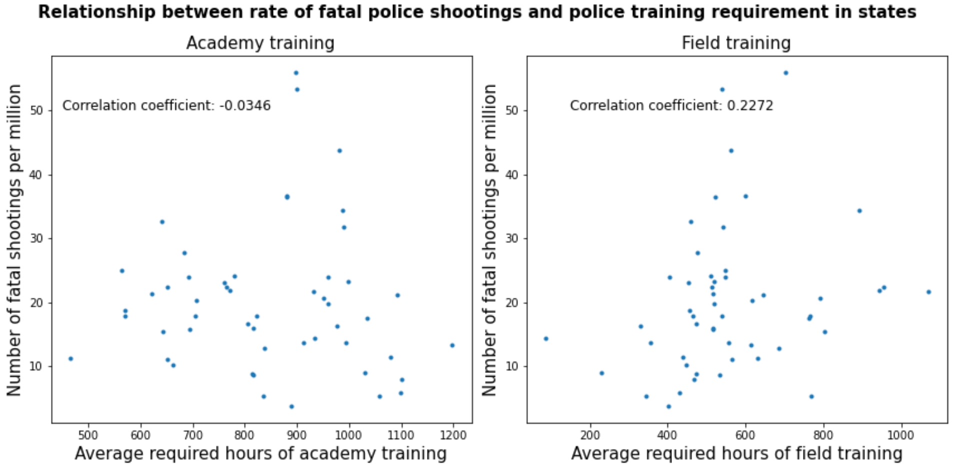
When examining number of shootings by state, we found that large states were often overrepresented (some of which were consistent with the cities plot above; e.g., Los Angeles is in CA, Phoenix is in AZ, and Houston is in TX, which are all states in the top 4), although this was not always the case (e.g., NY is ~halfway down the list, and MA is in the bottom third).

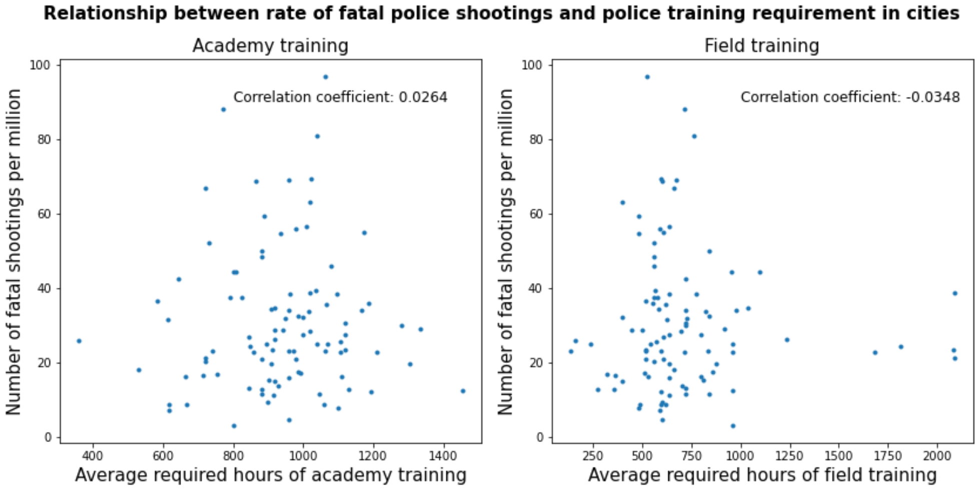

Each circle on the map below marks the location of a deadly shooting

A simple recreation of the Washington Post Police shootings map. More opaque clusters indicate hotspots of police shootings. States that are darker have more shootings per million people.

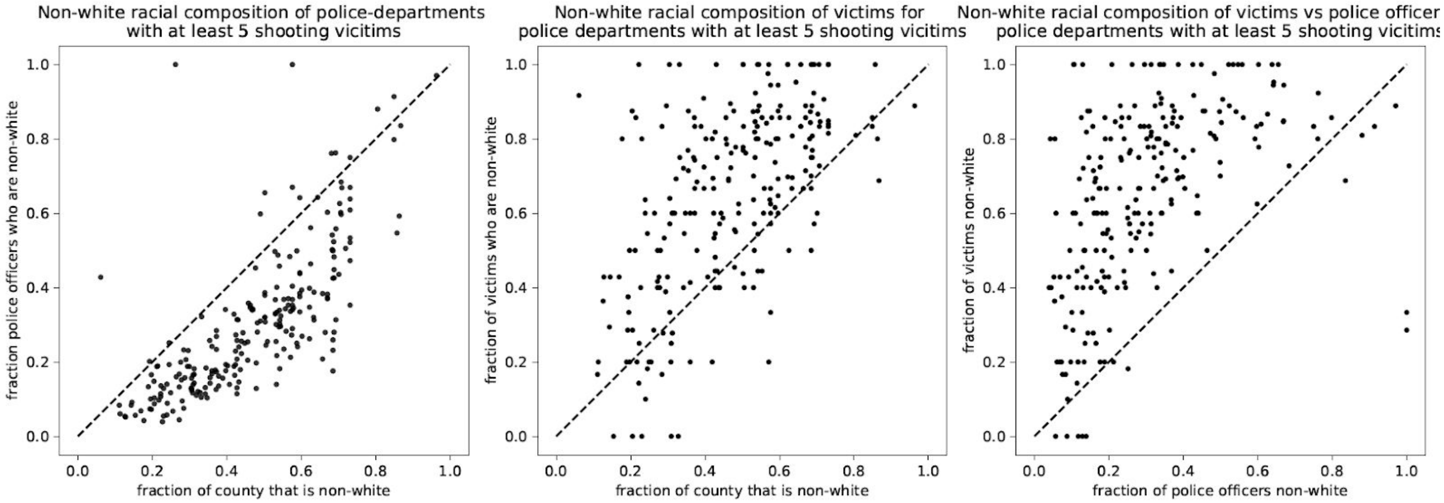

Relationship between rate of fatal police shootings and body camera usage

In the scatter plots above, we compared body camera usage and rate of fatal police shootings. We see a weak negative correlation between the rate of fatal police shootings and body camera usage at the state level, and a stronger negative correlation at the city level. These correlations are inconclusive but suggest that: 1) cities may be a more appropriate level of analysis, since different departments within a state may adopt different policies and the effect is washed out when aggregating at the state level; 2) states and cities where fatal police shootings occur frequently may lack policies/enforcement on body camera usage and appropriate use of force; 3) relatively low rates of fatal police shootings are associated with a range of values for the body camera usage, and this could be a result of high body camera deployment deterring police violence, or places lack the incentive to deploy body cameras due to few incidents of police shootings.


Relationship between rate of fatal police shootings and police training requirement in states

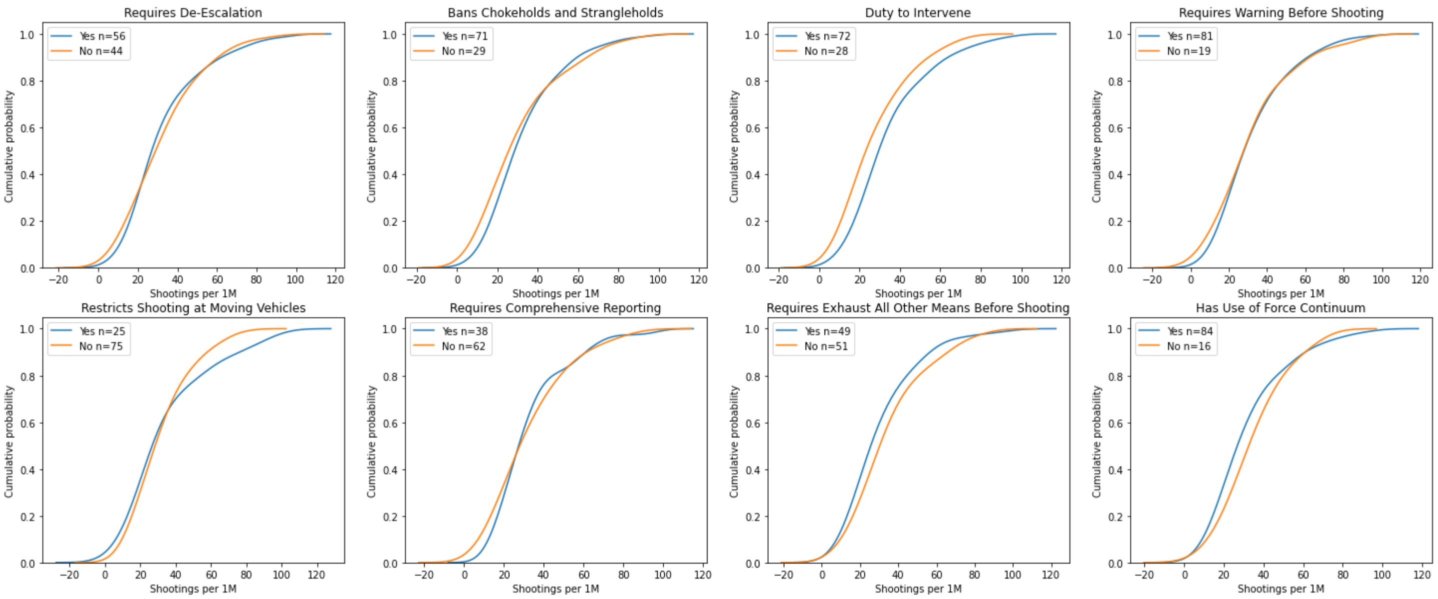


At the state level, we see that academy training hours are weakly negatively correlated with the rate of fatal police shootings, while field training hours are positively correlated with the rate of fatal police shooting. However, at the city level, we see the opposite trend with very weak correlations. This is quite shocking, and suggests that training hours may not be a good/reliable predictor for the rate of police shootings (at least at the city level). Alternatively, it's possible that since the LEMAS data is from 2016, it does not capture the changes since then and it would be more fair to look at the fatal police shootings that occured in 2016 only instead.



At the level of individual police departments, we analyzed the racial composition of both the police officers and the victims relative to the county in which the department is based. Plotted above are scatters containing points for each police department with at least 5 shootings victims. Comparing the fraction of non-white police officers to the fraction of non-white individuals in the county (left-most plot), the vast majority of points fall below the unity line, indicating that police departments tend to be much whiter (69.5% white mean across the above included departments) than the communities they serve (53.2% white (note this may be lower than national average due to geographic biases in where departments are localized)). In stark contrast, victims of police shootings are overwhelmingly less white than the racial composition of the county (middle plot). In comparing the racial composition of the police officers and the victims (right-most plot), as expected from the first two plots we see an even starker contrast: for nearly every police department analyzed (94.4%), the department is more non-white than the victims shot by that department.

Cumulative probabilities of shootings for police departments that implement each policy or not



Regardless of whether a police department has adopted a use of force policy or not, the rate of fatal police shootings is relatively the same. This is interesting as it does not align with what the use of force project reports. This may be due to the face that their analysis was performed in 2016 when less departments had implemented at least one of these policies. We are aggregating the number of police shootings from 2015-present, and a significant portion of these police departments have implemented use of force policies since the original report was published.

Revised Project Question

Based on our individual research and EDA, we want to determine if people with certain demographics (e.g., age, race) are more likely to be killed than others which could provide evidence for discrimination, and how this may vary at the city vs. state level. Further, does the relationship between the demographics of the police departments and the jurisdictions they preside over predict police violence.

Revised question: *Do police demographics and policies predict the total number of fatal police shootings and the demographics of victims at the city and state level?*

Baseline Model

If we were to guess race based on the variables in the Washington Post dataset, we could predict based on the majority class for race (W = white, B = Black, H = Hispanic, A = Asian, N = Native American, O = other):

W	0.509707
B	0.266817
H	0.181716
A	0.017156
N	0.015124
O	0.009481

Thus, our prediction would be that every victim is white, and a baseline model would perform at ~51% accuracy. To improve upon this, we could use logistic regression or other ensemble methods (e.g., random forest), and we could then also look at feature importances.