 CS109A: Final Project Milestone 3

Group #39 - Police Violence in the US

Team Members: Chika Okafor, Vasco Meerman, Matthew Parker, and David Koupaei

Harvard University - Fall 2020

# A description of the data

After researching many available crime datasets, the team quickly realized that almost all projects referenced one of two key datasets: [Mapping Police Violence](https://mappingpoliceviolence.org/) and the [Washinpost's repo](https://github.com/washingtonpost/data-police-shootings). The team is confident these are the most thorough and trustworthy datasets publicly available.

**Washington Post:** In 2015, The Post began tracking more than a dozen details about each killing — including the race of the deceased, the circumstances of the shooting, whether the person was armed and whether the person was experiencing a mental-health crisis — by culling local news reports, law enforcement websites and social media, and by monitoring independent databases such as Killed by Police and Fatal Encounters.

**Mapping Police Violence:** This information has been meticulously sourced from the three largest, most comprehensive and impartial crowdsourced databases on police killings in the country: FatalEncounters.org, the U.S. Police Shootings Database and KilledbyPolice.net.

**Selection:** The mapping police violence dataset has more predictors and data entries which could still be of interest for this analysis, but it also has some NaN values and incomplete data. Washington Post sources their data from the mapping police violence dataset, but they cleaned the data more completely making it more useful for many required comparisons in this investigation. Therefore, the best data source to answer the proposed analysis questions is the Washington Post dataset.

Key predictor and response variables:

|  |  |  |
| --- | --- | --- |
|  | Washington Post | Mapping Police Violence |
| Size | * 17 columns * 5716 rows | * 27 columns * 8507 rows |
| Response | * Count of violent incidents | * Count of violent incidents |
| Useful Predictors | * Date * Manner of death * Armed? * Age * Gender * Race * State * Signs of mental illness * Threat level * Flee * Longitude/Latitude | * Victim's age * Victim's gender * Victim's race * Date of Incident (m/d/y) * State * Cause of death * Death justified? * Criminal Charges? * Symptoms of mental illness? * Unarmed? * Alleged Weapon * Alleged Threat Level * Fleeing * Geography |

*Table 1: Predictor and Response Variables Reviewed*

# EDA Methods

The EDA section was performed by all team members to ensure a diverse set of observations and informed conversations moving into the modeling and results section of the project.

The general flow of the investigation was:

* Review and Describe data
  + Starting with many available sets, whittle down to only the most useful data (*research*)
  + Finalize dataset most capable of answering initial project questions (*Washington Post set*)
* Take a deeper look at distribution of each predictor variable
  + Make sure data is sensible and complete, clean as needed (*info(), value\_counts, histograms, pairplots*)
  + Make sure variables are informative (appropriate distribution of observations across values with *hist*)
* Dig in further to selected datasets with initial look at trends
  + Total count of incidents by race, time, and other predictors (*bar plots and histograms*)
  + Sample means, standard deviations, aggregate by year, by Geography (*barplot, try map plot*)
  + % of incidents by other predictors:
    - by threat level, allegedly armed, by race, by fraction of US population, and mental health
* Review Correlations
  + By race, gender, state, age, and other predictors like: (*heatmap plot*)
    - With and without threat level, flee, signs of mental illness, unarmed (*various plots*)
    - Allegedly armed/unarmed and threat level (*various plots*)
* Perform additional visualizations gaining deeper insights of predictors (*Scatterplots, swarm plots, histograms*)
  + Outcomes, by race (criminal charges, official disposition, etc.)

# Visualizations and Data Summary

This analysis is broken down into various investigations to answer the problem statement question. A lot of effort was given to normalizing the data by population and by racial diversity using external data sources to help understand the bias in the incidents more clearly.

**3.1: Location Based Analysis**

Initially it appears there is a large disparity with incidents in particular states. However, after correcting for state populations, there is a much more standard distribution of violent incidents.

|  |  |
| --- | --- |
| *Fig 1: Total Incident by State* | Chart, bar chart, histogram  Description automatically generated  *Fig 2: Incident by State, Normalized by Population* |

**3.2: Race Based Distributions**

Race has a similar issue where the initial estimates revealed that whites were most frequently involved in violent incidents, but after correcting for racial distributions, there is more bias toward minorities, particularly black victims.

Whites are overwhelmingly involved based on total incident count, but they are the 2nd least likely to have a violent incident by percentage when normalizing the population.

|  |
| --- |
| *Fig 3: Total Incidents by Race* |
| *Fig 4: Incident by Race Normalized for Population Demographics* |

**3.3: Time Based Distributions**

The other key item of interest for our model predictions is whether the trends have shifted over time. Considering 2020 is a partial year, there is a very even number of total incidents over the last 5 years, but some trends arise when filtering by race, particularly a decrease for whites and an increase in unknown data, generally flat trend otherwise.

|  |
| --- |
| *Fig 5: Incident by Year filtered by Race* |

**3.4: Other**

The team started investigating other correlations in the data using heatmaps, pairplots, and value\_counts investigations. There was also some initial work into plotting geographically, to help with final visualizations to substantiate the geographical differences that we expect the model to predict.

|  |  |
| --- | --- |
| *A picture containing chart  Description automatically generatedFig 6: Correlation of predictors from Wapo Data* | *Chart, scatter chart  Description automatically generated*  *Fig 7: Geographic distribution of incidents* |

# Revised project question based on insights gained through EDA

Initial problem questions from Milestone 2:

1. How overwhelming is the evidence of discrimination in police violence?
2. Are there geographic differences in these injustices?
3. Has the frequency of police violence towards the citizens it tries to protect increased over the last 4 years?

EDA was very informative to shaping our thoughts and ideas about what questions would result in the most impactful analysis. The team compressed into two updated questions.

Modified questions:

1. Is the population-weighted incidence rate of police shootings different for white vs. black victims (vs. other races)?
2. How has the population-weighted incidence rate of police shootings shifted over the period 2015 to 2020?
   1. By race of victim
   2. By geographic region

# A baseline model

For both of our modified questions, we will restructure the Washington Post dataset, so that each row corresponds with a unique state and year.  For each state-year, we will construct two outcome variables: (1) the total black civilian shootings per capita and (2) the total white civilian shootings per capita.  We will also add various control variables, potentially including: (1) state median income level; (2) total population; and (3) black share of the population.

* **To answer Modified Question #1**, we will use a Paired T-Test, comparing the means for total black shootings per capita and total white shootings per capita
* **To answer Modified Question #2**, we will fit the following Linear Regression Model:
  + where denotes the total shootings per capita in state *s* during year *t* for race *r*. This expression will be evaluated at least for when race is white and when race is black
  + denotes the total shootings per capita in state *s* during year *t* for geographic region *r*
  + denotes fixed effects
  + denotes with time fixed effects
  + denotes the year. As 2015 is omitted from the regression specification, the coefficients are normalized such that the 2015 values are set to 0
  + are a series of state and year controls
  + The coefficient of interest is , which demonstrates the relative impact on the outcome variable for each subsequent corresponding year, compared to the normalized year of 2015.