 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. The Post conducted additional reporting in many cases.

**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. We've also done extensive original research to further improve the quality and completeness of the data; searching social media, obituaries, criminal records databases, police reports and other sources to identify the race of 90 percent of all victims in the database.

**Notes:** Washing Post sources their data also from the mapping police violence dataset, but they cleaned the data. Meaning that the mapping police violence dataset has more entries because it also has some NaN values, which could still be of interest for this analysis.

Key predictor and response variables:

|  |  |  |
| --- | --- | --- |
|  | Washington Post | Mapping Police Violence |
| Response | * Count of violent incidents | * Count of violent incidents |
| 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
  + Finalize dataset most capable of answering initial project questions
* Dig in further to selected datasets with initial look at trends
  + Total count of incidents (by race, time)
    - Table, Incident by race, by year (Bar plot)
  + Sample means, standard deviations
    - Aggregate, By Year, By Geography
  + % of incidents by other predictors:
    - by threat level, allegedly armed vs. unarmed, by race, by fraction of US population
* Take a deeper look at distribution of each predictor variable
  + Make sure the data makes sense, and is complete (clean as necessary)
  + Make sure the variables are each informative (i.e., appropriate distribution of observations across values)
* Review Correlations
  + By race, gender, state, age, and other predictors like:
    - With and without threat level, flee, signs of mental illness, unarmed
    - Allegedly armed/unarmed and threat level
* Perform visualizations to start gaining deeper insights about the predictors
  + Sns.pairplot, Scatterplots, Violinplots, histograms
  + Outcomes, by race (criminal charges, official disposition)

# Visualizations and Data Summary

This analysis is broken down into initial views and then some more informative visualizations

|  |  |  |
| --- | --- | --- |
| *Fig 1: Incidents over time* | *Fig 2: Incident by race and location* | *Fig 3: WaPo predictor pairplot* |
| *Fig 4: Correlation of predictors* | *Fig 5: Incident by race and location* | *Fig 6: WaPo predictor pairplot* |

# Revised project question based on the insights you 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, after controlling for observable characteristics?
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, including: (1) state median income level; (2) state population; and (3) state violent crime rate.

* **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 with the total shootings per capita in state *s* during year *t* for race *r*. This expression will be evaluated for when race is white and when race is black.
  + denotes with state 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.