 CS109A: Final Project Report

Group #39 - Police Violence in the US

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

Harvard University - Fall 2020

Table of Contents

[1. Initial Problem Statement 1](#_Toc58008203)

[Motivations 1](#_Toc58008204)

[Problem Statement 2](#_Toc58008205)

[Data Source 2](#_Toc58008206)

[2. Visualizations and Data Summary 2](#_Toc58008207)

[3. A baseline model 4](#_Toc58008208)

[4. Further Modeling Approach 5](#_Toc58008209)

[Modeling Plan 5](#_Toc58008210)

[5. Results 6](#_Toc58008211)

[Model 1: Prediction of Race 6](#_Toc58008212)

[Model 2: Prediction of Mental Illness 7](#_Toc58008213)

[Model 3: Prediction of Judicial System Outcomes: 7](#_Toc58008214)

[6. Conclusions 7](#_Toc58008215)

[High Level Results 7](#_Toc58008216)

[Extending Predictions 8](#_Toc58008217)

[7. Social Impact Study 8](#_Toc58008218)

[Background 8](#_Toc58008219)

[Issue Politicization via Media 9](#_Toc58008220)

[Reflections after Reviewing this Data 9](#_Toc58008221)

[8. References 9](#_Toc58008222)

# Initial Problem Statement

### Motivations

### Problem Statement

The following questions were adjusted after significant EDA efforts:

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

### Data Source

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.

# 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.

![Chart, histogram

Description automatically generated]()

**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.

![Chart, bar chart

Description automatically generated]()

**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.

|  |
| --- |
|  |

Discuss how a larger majority of the “unknown” category is made up of Latino population. Dig into why white is dropping. Possibly join/average 2019/2020

# 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.

* **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.

**Results:** The table below graphically depicts estimates of the evolution in the police killing rate per capita over time, both by racial group and by region. These graph follow the regression specifications under the Modified Question #2. All estimates are calculated relative to the year 2015, which is normalized to 0. 2015 corresponds with the first full year after the heightened national dialogue on policing began (the lack of indictment for the police shooting of Michael Brown happened on Nov 24, 2014). All estimates are in log points. A log point estimate of *a* approximates a (100\**a*)% effect when the magnitude of *a* is small. In short, log points approximate percentage change for small changes in the independent variables. Vertical lines represent two standard errors.

The graph on the left provides evidence that the police killing rate per capita declined most over time for the white racial group, followed by the black racial group. In contrast, the police killing rate per capita *increased* for the “other” racial group. At the aggregate level, there was not a statistically significant shift in the overall shooting rate per capita over this time period. This suggests that the rate of police killings has remained relatively flat over this time period, but that the composition of the victims by racial groups has changed over time—away from white and black victims and toward the “other” racial categories (predominantly Latino). This shift might be explained by changes in police presence across communities, and is consistent with other findings that have found that the “Ferguson Effect” has led to lower police presence in predominantly black communities since the Ferguson uprising in 2014.

The graph on the right provides little evidence of regional differences in the evolution of the police killing rate over time. No geographic region has statistically significant differences from the baseline year of 2015, suggesting that the flat trend in the total levels (illustrated by the graph on the left) is matched by a flat trend at each geographical subregional level (illustrated by the graph on the right).

|  |  |
| --- | --- |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |

# Further Modeling Approach

### Modeling Plan

The basic time dependent relationships with race ended up yielding no obvious correlations, so the analysis shifted to predicting other parameters which were key in determining the likelihood of police shooting occurances.

Models attempted:

* Logistic Regression
* Polynomial Feature Enhanced Logistic Regression
* Regularized Ridge and Lasso Regression
* Random Forest
* Neural Network

Multiple models were attempted based on the broad nature of the data

1. Prediction of race by all other predictors
2. Prediction of Mental Illness
3. Prediction of Outcome of Judicial Process for all Fatal Incidents (was killing Justified?)
   1. Specifically are there any key predictors yielding differences in likelihood of the killing being considered justified by our judicial system.

# Results

### Model 1: Prediction of Race

**Key Predictors from Logistic Regression Model:**

Chart

Description automatically generated

**Model Prediction Accuracy:**

Chart, line chart

Description automatically generated

**Conclusions:**

No models performed much better than a 50% coin flip on being able to differentiate race of the shooting victim based on other predictors. Random forest model with average tree depth generalizes the best with the highest accuracy score using the test data.

### Model 2: Prediction of Mental Illness

Asdfasdf

### Model 3: Prediction of Judicial System Outcomes:

Asdfasdf

# Conclusions

### High Level Results

The project initially intended to show clear trends between police shootings over time, including differences between races. However, the EDA and modeling process yielded no definitive time-based or race-based connection with the occurrence of a deadly police shooting.

Answers to the problem statement questions:

1. *Is the population-weighted incidence rate of police shootings different for white vs. black victims (vs. other races)?*
   1. Yes, there is a higher probability of fatal police shooting occurrence for African Americans than other races. Add statistics
2. *Has the population-weighted incidence rate of police shootings shifted over the period 2015 to 2020?*
   1. No, there is no statistically significant difference over time between the distribution of police violence when controlling for population based on racial group or other key predictors

### Extending Predictions

While the results of the project didn’t align directly with our initial predictions, there were some other very interesting conclusions through modeling different types of predictions.

1. We are able to predict mental illness of a victim based on other data predictors (how could this be used in real-time?)
2. The mapping police violence data can produce an adequate model to predict justification of the police shooting (was the officer charged, acquitted, etc.) based on other raw and derived predictors related to population density.
   1. The two key predictors in these models is signs of mental illness and Year
      1. Expand on mental illness playing a role
      2. While the original baseline model doesn’t show a clear time-based trend in the number of killings, this analysis suggests there is a difference over time on how the cases are handled after the event. (discuss how this could related to change in dash-cams/availability of information)

# Social Impact Study

### Background

Excerpt from the project problem statement:

*The experiences of many Americans - including Breonna Taylor, George Floyd, and Rodney King among, tragically, many others - speak to an epidemic of excessive and unjustified use of force by law enforcement officers whose sworn duty is to protect the communities they serve. Instances of police violence range from unreasonable search and seizure to the excessive application of physical force. Severe abuses of the police’s monopoly on violence, like extrajudicial killings, occur despite a constitutional presumption of innocence.*

Text

Description automatically generated

Black Lives Matter Movement:

Table

Description automatically generated

### Issue Politicization via Media

This issue has become increasingly politicized within the timeframe of the police violence analysis.

asdasdf

### Reflections after Reviewing this Data

asdfasdf

# References

1. Black Lives Matter: <https://blacklivesmatter.com/>
2. Reference data for black lives matter hashtag usage on Twitter: <https://www.pewresearch.org/internet/2018/07/11/an-analysis-of-blacklivesmatter-and-other-twitter-hashtags-related-to-political-or-social-issues/>
3. Population Density Estimates: <https://simplemaps.com/data/us-cities>
4. Urban/Suburban Classifier: <https://medium.com/codait/got-zip-code-data-prep-it-for-analytics-7022b47652d9>
5. Race Classifications:
   1. Original Set: <https://github.com/washingtonpost/data-police-shootings>
   2. Additional Categorical from Wikipedia: <https://en.wikipedia.org/wiki/Race_and_ethnicity_in_the_United_States#Racial_categories>
   3. 2019 Estimates from Census Bureau: <https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>