CMSE 381 Project

Hattie Pimentel

April 18, 2021

Contents

Problem Statement

Among the 467 Americans killed by police in 2015 is a particularly conspicuous name: Freddie Gray. Freddie Gray's death as a result of a spinal cord injury inflicted by Baltimore police through unrevealed circumstances sparked widespread concern and anger over police killings in the United States, particularly police killings of people of color (McDonell-Parry and Barron). With awareness of these circumstances and public anger, analyzing trends in police killings is important to identity areas of particular concern and increase clarity of the problem.

In this vein, this report first analyzes and reproduces claims made by the FiveThirtyEight journalist Casselman. Next, this report creates visualizations to help get a better understanding of the data behind police killings. Finally, this project sought to develop models of its own to answer three questions: Can the included parameters in the police killings FiveThirtyEight dataset be used to predict whether the killed individual was female? Can included parameters be used to determine if the deceased was a person of color? After removing elements of the data related to location, can the included parameters be used to determine if the deceased was a person of color? In pursuit of answers to these questions, this report uses logarithmic regression, K-nearest neighbors, and support vector machines classifications methods.

Past Scholarship

A FiveThirtyEight report by Ben Casselman called "Where Police Have Killed Americans in 2015" sought to quantify who had been killed by police in 2015. The report relies on "verified crowdsourced" data —that is, data that the Guardian collected from open-source projects and then verified. In the article, Casselman finds that "[p]olice killings tend to take place in neighborhoods that are poorer and blacker than the U.S. as a whole." Casselman also notes that police killings are not uniform; despite the trends, another example of a police killing in 2015 was Vincent Cordaro, a white man killed in an census tract where the median household income was greater than $140,000.

As noted by Brown and Ray, women killed by police, particularly Black women, are more likely to be unarmed and killed as "collateral damage," when they were never a focus of the police in the first place. Iati, Jenkins, and Brugal wrote that women are killed by police at a much lower rate than men, and often under different circumstances. For instance, women are more likely to be accidentally shot or suffering a mental health crisis when they are killed. All five of these authors agree that the subject deserves greater attention and analysis.

Model Design

This projected used the "police-killings" dataset from FiveThirtyEight, which includes information about all individuals killed by police in 2015 in the United States. Four hundred, sixty-seven lines in the dataset represent one police killing each. The data was loaded, the columns transformed into fitting variable types, and NA values removed when appropriate.

First, this project sought to validate the results of Casselman's FiveThirtyEight article, "Where Police Have Killed Americans in 2015." Casselman's quantitative findings are simple statistics, and they were replicated by masking the dataset, counting the number of entries that matched the mask, and dividing by the number of total entries in the dataset. To visualize these statistics, histograms were created.

Next, since the original article did not include general analysis of the dataset, additional visualizations were created. These include a map of the rates of police killings in each state, which was realized by grouping the data by state, dividing the total number of killings in each state by the state population (with state populations drawn from the "usmap" library's statepop dataframe), and using the "usmap" library to create a map. Additionally, bar charts of police killings by tract median household income quintile and race/ethnicity were created. These were drawn by grouping the killings by the relevant x-values, and then drawing them with the "ggplot2" library. The inside of each bar was colored to reflect what types of arms, if any, the killed individuals carried at the time of their deaths.

Lastly, this report sought to answer questions of its own. First, the report sought to fit a logistic regression model to determine whether the deceased was female or male. The data was split into training and testing sets. Initially, to get a feel for how the parameters interacted with the dependent variable, all parameters were included, except for ones that were plainly meaningless to the regression model, such as name, city, latitude, and similar. Given the results detailed below, the parameter selection was later modified to include parameters with lower p-values. Given the very limited data on female killing victims (the entire dataset contains only 22 women), the classifier presented a risk of having high accuracy rates while providing limited useful information. Therefore, the threshold was modified to allow for greater misclassification of deceased males in an attempt to correctly classify a greater number of deceased females.

Next, a KNN was used to attempt to predict gender. Given the limited number of women in the dataset, a small K was used. Multiple values of K were tested manually, and K was set equal to two with all non-trivial continuous variables included.[1] Again, due to the possibility for great accuracy with a small true positive rate of deceased females, every possible subset of parameters was searched for one that would provide for the greatest true positive rate on the training data using a function that generates combinations in the utils library. Given the quickness of testing the various subsets, this was a computationally reasonable method that was simple to implement.

For the next question, an SVM was created to detect whether a victim was a person of color. Non-white individuals were defined as persons of color, and individuals of unknown race/ethnicity were removed from the dataset. For both a linear and radial kernel, SVMs were created with all non-trivial parameters. Tuning was performed for the cost variable for the linear variety and cost and gamma for radial. Afterwards, this process was repeated after removing all variables related to location and including only variables indicating the cause of death and the arms, if any, carried by the deceased at the time of their killing.

Results

This project was able to validate many of the findings in Casselman's FiveThirtyEight article using the police-killings data.[2]

[1] The output of the KNN model with K=2 was identical to the output at K=1 and K=3.

[2] A data table like the one included in Casselman's article is regenerated in the accompanying code. It is not included here because it is quite long.

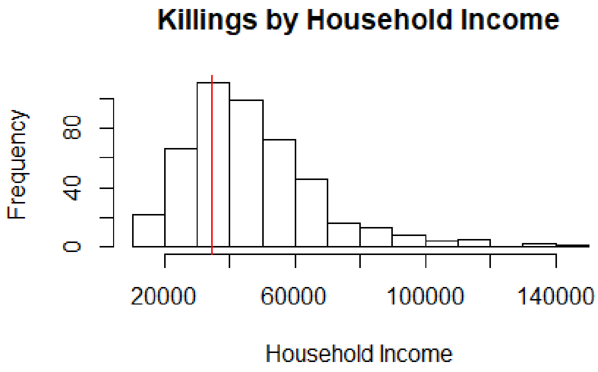


Figure 1: Histogram of killings by median tract household income, with the red line indicating the bottom quintile of median tract household income.

First, Casselman writes "[a]bout 30 percent of the killings — 139 of the 467 — took place in census tracts that are in the bottom 20 percent nationally in terms of household income." Likewise, my analysis showed that 29.76445% of killings, or 139 killings, took place in areas in the bottom quintile of national household income, as is shown in Figure 1.

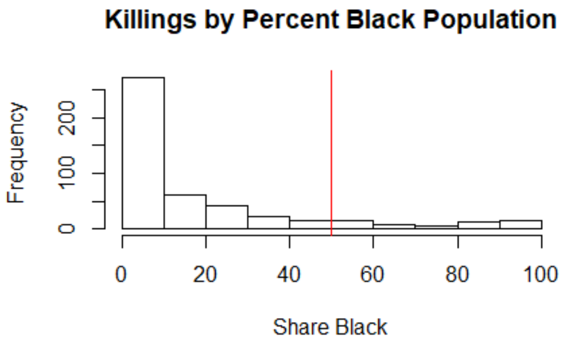


Figure 2: Histogram of killings by percentage of the population that is black, with the red line indicating the boundary between a majority and minority black population.

Next, Casselman notes that "[a] quarter of those killed by police died in tracts with majority-black populations." When "majority-black populations" were defined as those populations with greater than 50% of the population being Black, I obtained an analogous result of 11.7773%. When this term was instead defined as populations where the Black population was greater than the Hispanic or white populations (the only races included in the dataset as percentages), as in a majority-minority population being Black, I obtained a result of 14.3469%. Regardless, this claim could not be validated, as is shown in Figure 2.

Third, the article argues that "[o]f the 136 African-Americans killed by police who are in the Guardian's database, 56 — more than 40 percent — died in tracts in the poorest 20 percent nationally" (Casselman). My project indicated that there were 135 African Americans killed by police, not 136. Despite this minor difference, I find that 56 of the African Americans killed encountered police in tracts in the lowest quintile of median household income, which is 41.48148%.

Next, this project sought to add some visualizations to facilitate understanding of the dataset and spark relevant questions.

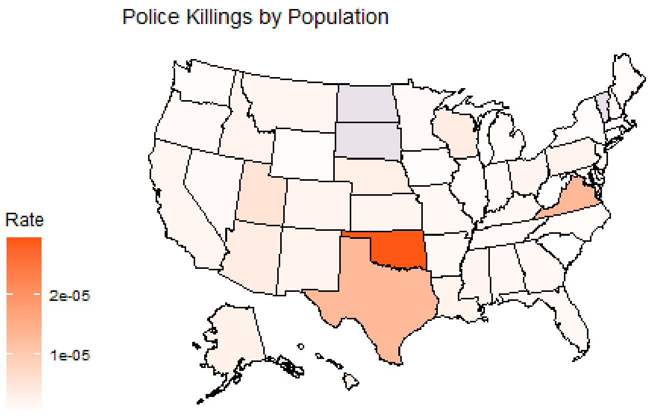
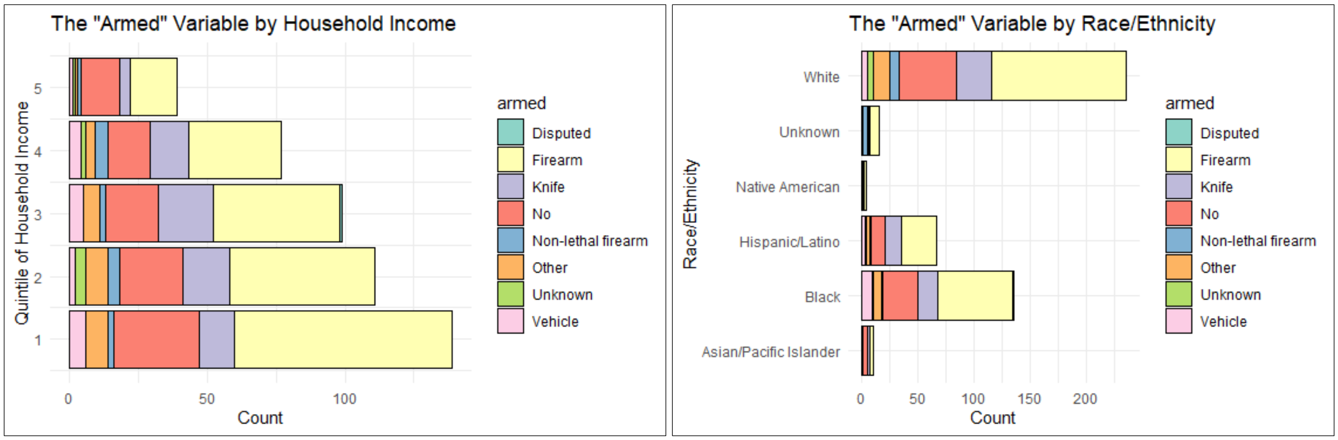


Figure 3: Map of the rates of police killings in the U.S. in 2015.

Figure 3 indicates that Oklahoma, Virginia, and Texas had abnormally high rates of police killings in 2015, while North and South Dakota both had zero. Since the rate of police killings per person is relatively small, this graph would likely vary significantly from year-to-year.



Figures 4, 5: Breakdown of the "armed" variable by quintile of tract median household income and race/ethnicity.

Figures 4 and 5 break down the “armed” variable —what weapon, if any, the deceased was carrying when killed— by household income quintile and race/ethnicity. They reveal that the largest percentage of the deceased were carrying a firearm, and a fairly large percentage were carrying no weapon at all. Persons of unknown race were much more likely to be carrying a non-lethal firearm. This is curious, and likely indicates some peculiarity in how the data was collected that should be explored further.

Next, this report analyzed whether gender could be determined from other included parameters with logistic regression. None of the parameters included (which initially was all parameters that were non-trivial) were statistically significant. The parameters with the highest p-values were removed in an attempt to minimize their disorientating effects and obtain a more effective model, and the model was recreated with those that remained. Still, the smallest p-value was 0.1791 (associated with the percent of population of 25 years of age or older with at least four-year college degree), so there was no clear evidence of an association between the included parameters and gender. However, the results could still warrant future analysis; the intercept for age was negative, and the intercept for the unemployment rate was positive, sparking questions about how these variables impact behavior by gender and policing in different circumstances. At a threshold of 0.5, the model testing accuracy was 95.13%, but this was merely a result of always guessing "male." After lowering the threshold, an accuracy of 59.72% could be obtained, with a true positive rate of 57.14%.

Next, a KNN was fit to the data to predict gender, as described above. As the result of the KNN being fitted to all non-trivial continuous variables with K=2, the accuracy was 92.76%. This output occurred with a true positive rate for guessing females at zero, but with four misclassifications of men as women. After all subsets of included variables were tested, no models with improved true positive rates for identifying women were found. In sum, this analysis had limited success separating and predicting female police killings from male police killings with the parameters included and could not conclusively conclude that there was a large difference in their circumstances.

Lastly, SVMs were used to predict whether victims were people of color. First, SVMs with both linear and radial kernels were built with all non-trivial parameters in the dataset and tuned. The linear SVM had an accuracy rate of 73.57%, and the radial SVM had an accuracy of 75.71%. This indicates that the circumstances in which people of color are killed look different than circumstances where white individuals are killed; deaths of the two groups occur in neighborhoods with different socioeconomic characteristics, different causes of death, and different types of arms carried by the deceased. These results are relevant even though they include economic factors and neighborhood descriptors because part of the discussion about police killings includes consideration of how police actions affect poorer communities, especially ones that are primarily communities of color. When SVMs were built and tuned only using cause of death and information on how or if the deceased was armed, the accuracy rate for identifying people of color was 55% for the linear model and 54% for the radial kernel. This is relevant to the question, as it demonstrates a difference in police interactions with the groups.

Conclusion

Future analysis could seek to analyze female versus male police killings with a dataset that included more variables, such as whether the individual was accidentally killed or whether the individual was suffering from a mental health crisis. More data on female police killings could further improve results, though this would come at the cost of having a specific time frame to analyze given the sparsity of data. Regarding killings and race/ethnicity, future analysis should obtain and analyze more detailed information about differences between the causes of death and state of arms between these groups, perhaps particularly when the deceased was unarmed. Including and analyzing more data for smaller groups, like Native Americans, could reveal trends that were not evident in this dataset.

The important topic of trends in police killings deserves greater attention. Quantitative analysis should be supplemented with qualitative analysis to make sure outliers and edge cases are not being missed, and to gain greater understanding of the topic and how it affects trust and other interactions with policing.

References

andrewflowers, dmil. (9 Feb. 2018). Police Killings [online dataset]. GitHub. https://github.com/fivethirtyeight/data/blob/master/police-killings/README.md

Brown, M. & Ray, R. (25 Sept. 2020). *Breonna Taylor, police brutality, and the importance of #SayHerName*. Brookings Institution. https://www.brookings.edu/blog/how-we-rise/2020/09/25/breonna-taylor-police-brutality-and-the-importance-of-sayhername/

Casselman, B. (3 Jun. 2015). *Where Police Have Killed Americans in 2015*. FiveThirtyEight.com. https://fivethirtyeight.com/features/where-police-have-killed-americans-in-2015/

Iati, M., Jenkins, J., and Brugal, S. (4 Sept. 2020). *Nearly 250 women have been fatally shot by police since 2015*. The Washington Post. https://www.washingtonpost.com/graphics/2020/investigations/police-shootings-women/

McDonell-Parry, A., & Barron, J. (12 Apr. 2017). *Death of Freddie Gray: 5 Things You Didn't Know.* Rolling Stone. https://www.rollingstone.com/culture/culture-features/death-of-freddie-gray-5-things-you-didnt-know-129327/