Investigating Police Murders
Jonathan Sheehan

## Introduction

The police in the United States have murdered 29,991 people since January 1st, 2000. By the time you are reading this, that number will have blown past 30,000. In 2020, those sworn to protect ended an average of 5.6 lives per day. The aims of this project are to visualize some of the killing trends, as well as optimizing various classifiers to explore the murders.

---

## Data Set

For this project, the primary dataset I used is from fatalencounters.org. In the Journal of Open Health Data, B. Finch et.al call it "the largest collection of PRDs [police related deaths] in the United States and remains as the most likely source for historical trend comparisons and police-department level analyses of the causes of PRDs."

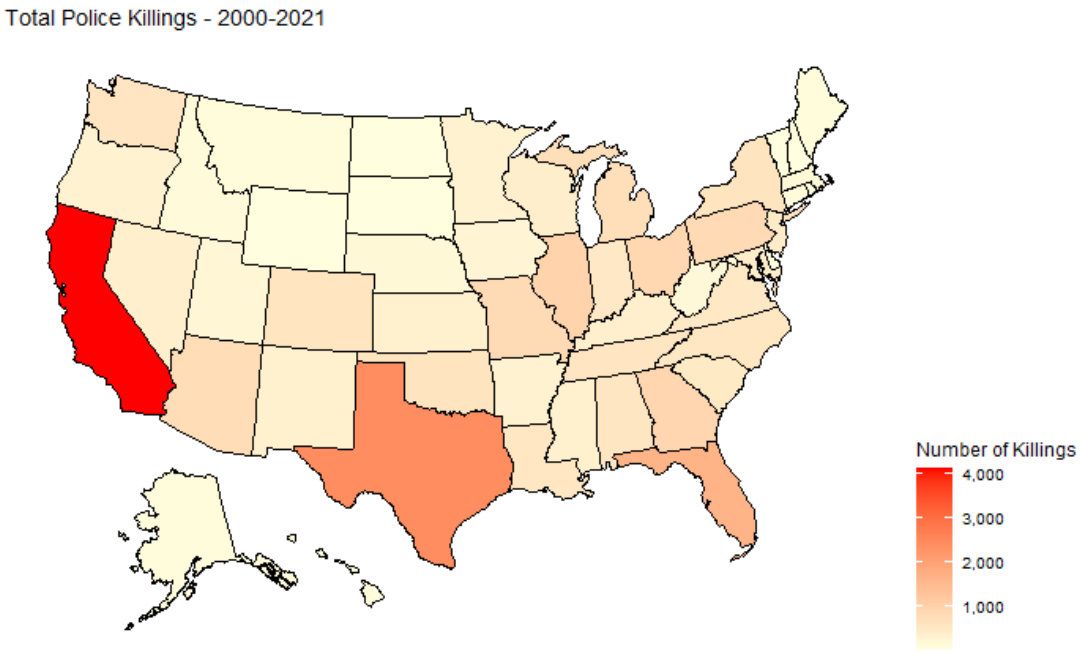


Fig 1: Total Police Killings from Jan. 1 2000 to Mar 31 2021, using Fatal Encounters dataset

This data set was tricky to sort through, as many rows were missing features— an unfortunate side effect of scrounging through news articles and Freedom of Information Act requests for data. I created two data frames: one with a lighter level of cleaning and another where every row was complete. The larger one contained 26,465 rows and 20 columns, and the more polished one contained 10,512 rows and 20 columns. I merged both of these with county level demographic data from the US Dept. of Agriculture's Economic Research Service, so that the models could be fed county-specific features. In addition, I used state-level demographic information to help contextualize my findings. Combining these datasets allowed me to feed information about the county where the killing took place into my models, so that specific demographic information would be used to perform any analysis.

Most of the cleaning and preprocessing that I had to do revolved around feature reduction, factor level reduction, and ensuring that everything was typed properly. This was not an easy thing to do, as part of it included removing some rows from the Killing data. One example of this was my needing to remove

any murders of individuals with a `Race` value of `Middle Eastern`. Removing them wasn't due to the small number of observations, but because no county or state demographic information for Middle Eastern people were in either supplemental dataset.

---

**Related Work**

The inspiration of this project came from 538's github repository on Police Killings. Their article was insightful, but was not filled with reproducible plots and graphs, so I took additional inspiration from the Washington Post's database. On their site, they mention the difficulty in getting a robust database of police killings, so it's commendable that these people and organizations put effort into uncovering the reality of police violence in America. To obtain this data, they had to comb through local news stories, reports from the police, and social media. Although they don't include non-shooting deaths, they mentioned using datasets like Fatal Encounters to supplement their independent searches. Here, I will attempt to recreate some of the visualizations presented from the Washington Post team. They will not always look identical, but this is simply due to my using a different dataset. Unfortunately for me, there wasn't much beyond visualizations that I could explore, so I will not be able to talk about any models just yet.

The first visualization that I recreated was the locations of every police killing after 2014 that I had coordinates for, overlayed on a map of the US with every state shaded according to the killings per 100k residents.

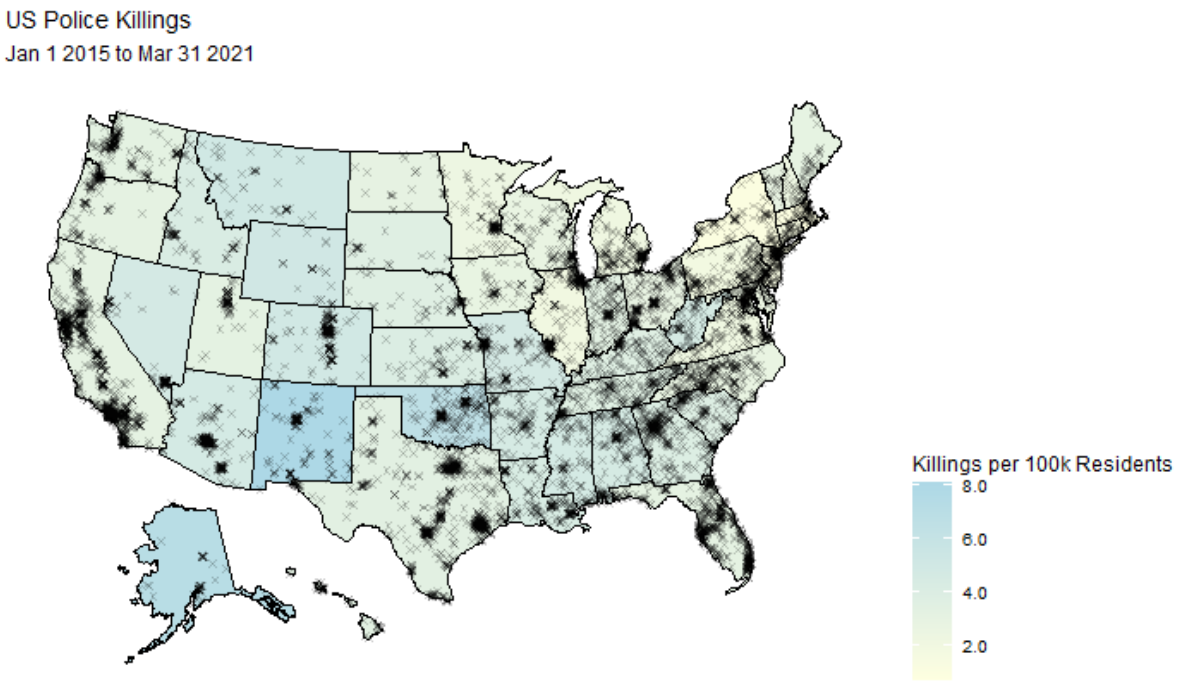


Fig 2: Map of Police Killings in America, with states shaded by Killings per 100k. New Mexico, Oklahoma, and Alaska have the highest incident rates.

This method of visualizing is especially poignant because you can see past the state statistics. Each point on the graph is a life ended, and it is important. It's especially interesting to see where more shootings happen, based on the intensity of the clusters. Most every major US city can be found on this map by killings alone, and many smaller ones stand out as well. That paired with the shading of the states helps to paint a pretty good picture of police killings. We can see that there are a lot of killings in the large population clusters of the county, yet the states with the highest rates of killings aren't around them.

Interestingly, despite a mega-cluster in the Northeast, the 5 states with the lowest rates of police killings are all from the region. Although not included in the Washington Post dashboard the following plot, inspired by the databricks article"Fatal Force: Exploring Police Shootings With SQL Analytics", provides more insight into how states stack up to each other.

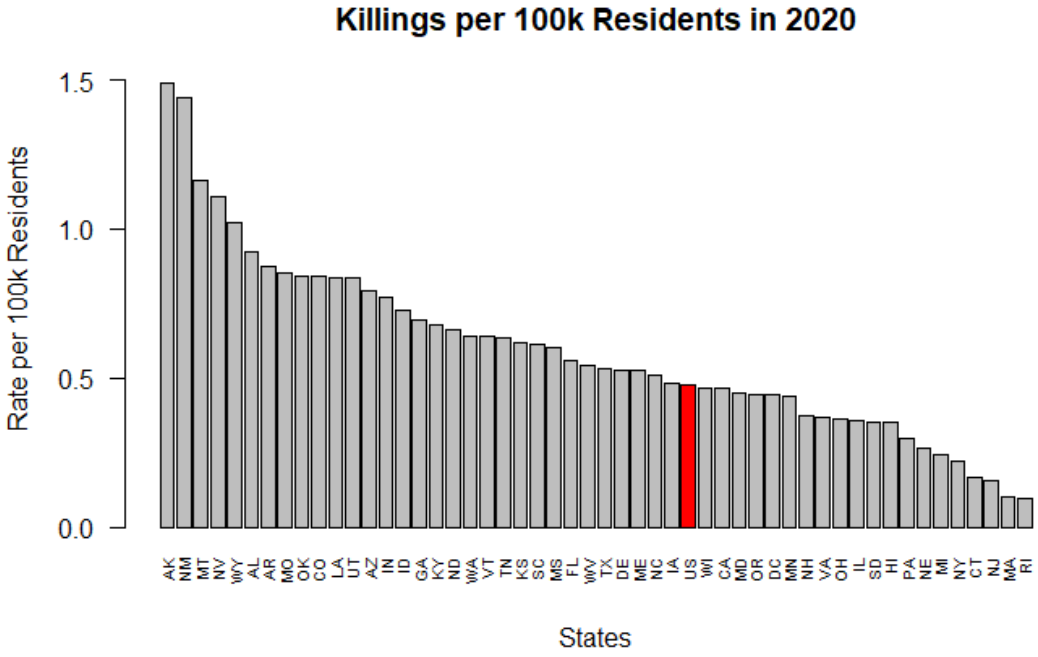


Fig 3: Rates of Police Killings for all US states and Washington D.C. The overall US race is highlighted in red.

In addition to the visualizations for state information, the Washington Post dashboard shows the race, age, and gender breakdowns of the Police killings. The first chart that I reproduced was that of race. There's a lot of information packed into it; the graph contains information on the number of police killings, the rate at which they occurred, and the share of the US population that a particular race makes up.

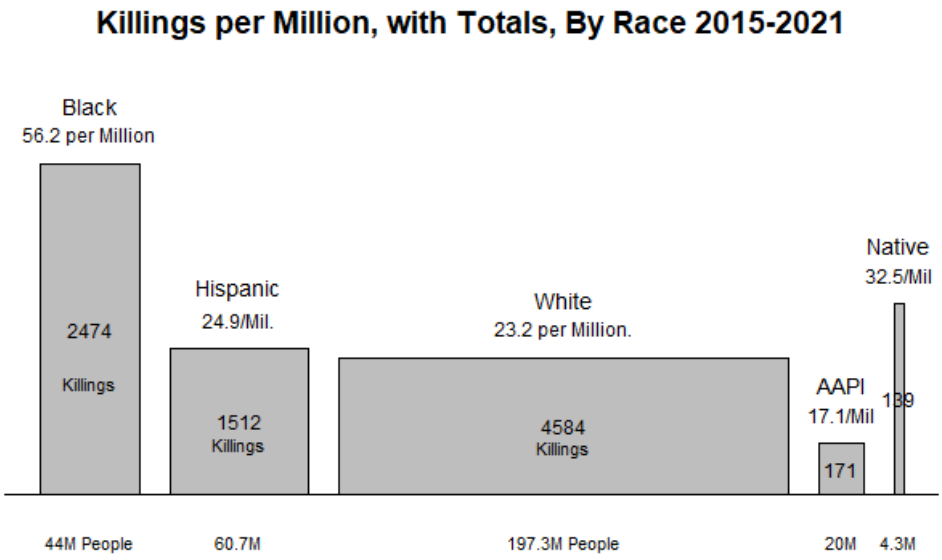

Fig 4: The height of each bar represents the police killing race for each race. The width of each relates to the race's share of the American population. The number of killings is there for context.

This graph clearly shows how race plays a role in how a police encounter turns out. Black Americans are murdered by police at rates unmatched by any racial group. All but one of my numbers are higher than what the Washington Post reported, which is expected due to the robustness of my source. What was most surprising to me was the lack of a drastic increase in the rate for Hispanic Americans. That rate was the only one that I had which was lower than the original number. Part of that may be due to systemic racism in the data collection itself. Many forms, at all levels of government discriminate against Hispanic and Latinx individuals by not including a suitable checkbox for race questions. Many times, there will be a separate question to indicate Hispanic or Latinx Origin, but not always. There are some states where it is thought that most Hispanic or Latinx prisoners get labeled as White due to inadequate forms. This not only lowers incidence rates for Hispanic Americans, but it artificially inflates them for Whites. The plot above looks like White and Hispanic Americans experience police killings at similar rates, but there is strong evidence that this may be misleading. Interestingly, the Washington Post's chart only reports there being 39 Million Hispanic Americans, where my number sat over 60 Million. A quick google search backs up my number, so I'm curious as to where they got this statistic. When using my number for population, their Hispanic murder rate drops by 10 killings per million, and is similarly closer to the White rate. Unfortunately, misleading numbers and inconsistencies like this are fuel for people who want to ignore the large scale societal issues we face surrounding race.

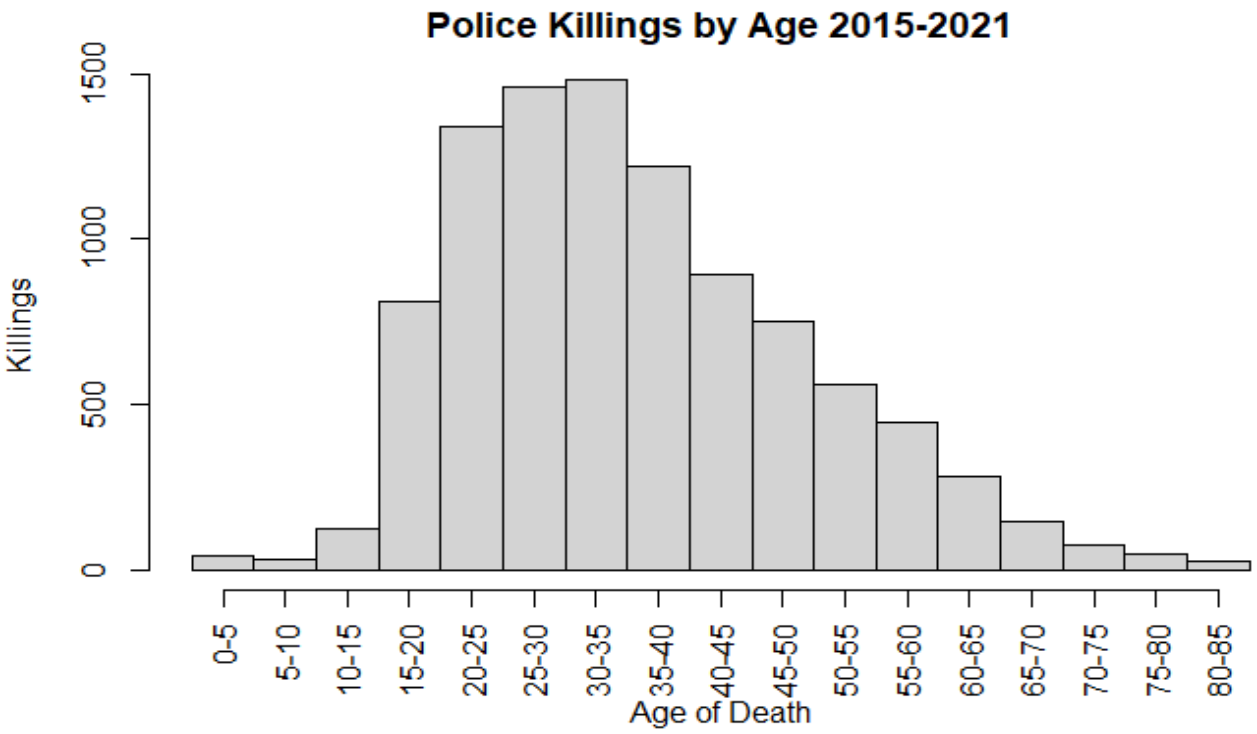


Fig 5: Police Killings by age group, from 2015-2021

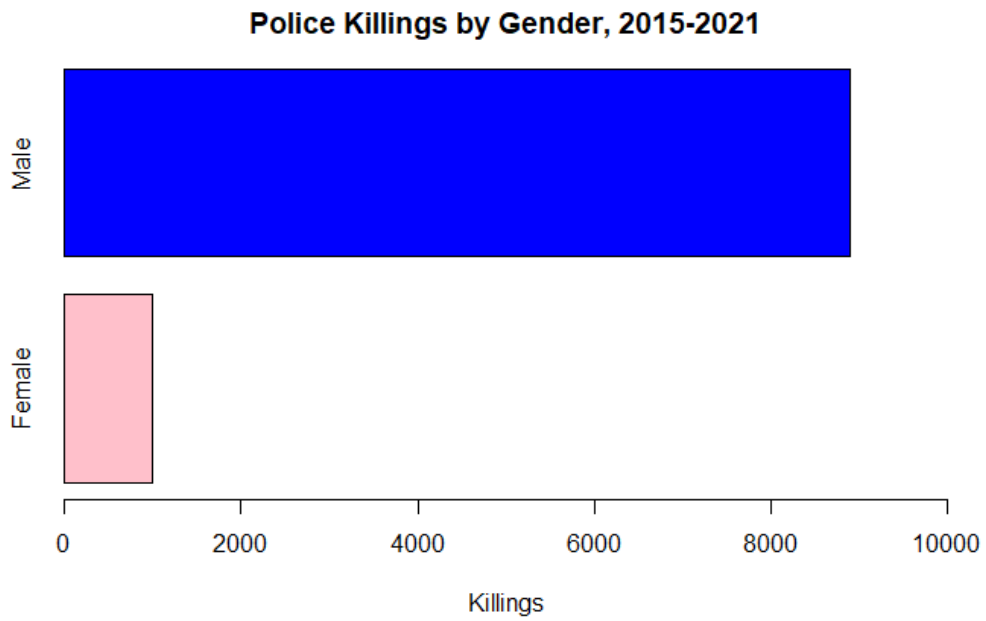


Fig 6. Police Killings by Gender, from 2015-2021

The three demographic plots show that there is a clear association with higher numbers of police killings, and characteristics that would historically classify an individual as threatening. Younger men, especially if they're black, are murdered at disproportionately high rates. Although everyone has some degree of underlying biases, it is up to all of us to recognize where those exist, so we can work to minimize their impact on how we act. Clearly, this growth isn't expected down at the precinct. The gender plot speaks for itself in its disparity. With age, it seems like once a young person hits puberty they are seen as a threat instead of a child. We expect the police force to be held to a higher standard, but they fail to rise above what we expect from non-uniformed members of society. Insead of being a front for justice and inclusion, the American police force falls into the trap of stereotyping and prejudice, which has led to thousands of lives ended, thousands of funerals, and almost no accountability.

---

**Method**

Most of my project was dedicated to finding the optimal model to predict whether or not a murder would happen in a state that had a rate of fatal police encounters higher or lower than the national killings per million people. Most of the killings were classified as being from a state that is worse than the national rate. I chose to spend the most time on this question, as it appeared to be something that would not only show the performance of various model types, but also because it could be a starting point for further investigations.

After splitting my data into testing and training sets, I ran a series of classifiers to try to find the one with the highest rate of correct classifications. To save time, I manually tweaked the parameters for most of the models instead of hypertuning with a grid of parameters. I used the subset containing the killings from 2015 to train and test various logistic regression classifiers, a pruned classification tree, various random forests, and two boosted models, one which had its parameters hypertuned. The hypertuned GBM generated an importance plot of all of the variables, which gave insights into how I could make a better model. I will expand more on this in the results section, but it provided me an opportunity to expand the data I was able to use. I reran all of my models except the pruned tree, and was able to create an even better classifier.

After figuring out the model that performed best at predicting if the police murder rate was higher or lower than the national rate, I attempted to create a model that would be able to predict the race (White, Black, or Hispanic) of the victim, as well as an additional one that would predict the state (California, Texas, or Florida) of the victim.

---

**Results and Discussion**

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.803212   0.087148   9.217  < 2e-16 ***
Age                  -0.002812   0.001897  -1.482    0.138
Armed.Unarmed1       -0.067126   0.054382  -1.234    0.217
Fleeing.Not.fleeing1 -0.006650   0.056340  -0.118    0.906
bool.race1           -0.312414   0.051632  -6.051 1.44e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9168.9  on 6903  degrees of freedom
Residual deviance: 9118.9  on 6899  degrees of freedom
AIC: 9128.9

Number of Fisher Scoring iterations: 4
```

Fig 7: Summary output of the optimal logistic regression classifier

Out of all of the types of models, the logistic regression classifiers performed the worst. As shown above, the only predictor that had significance was `bool.race`, which represented if the victim was white or not. On the testing data, which was made up of 30% of the total data used, it was only able to correctly classify 54.14% of the killings. It only performed marginally better than random, which was expected due to it not being the best model for the dataset. It did however give me the idea to expand upon race.

Out of curiosity, I ran another model where the county level demographic information was included to provide more information for it to work with. The information that was now included was rates of White, Black, Hispanic, Asian, and Native residents, as well as the percentage of foreign born residents and the homeownership rates. Interestingly, all of the predictors added were significant to some degree, and neither `Age`, `Armed.Unarmed`, nor `Fleeing.Not.Fleeing` weren't included. In the best model, 66.89% of the testing data was predicted properly.

Next, I decided to try a classification tree with `Age`, `Armed.Unarmed`, `Fleeing.Not.Fleeing`, as well as the newly added demographic variables. I chose not to include the actual tree here, as it is difficult to read, but it can be found in the accompanying .Rmd file.

```
Variables actually used in tree construction:
[1] "Hispanic"       "Native"         "OwnHomePct"     "ForeignBornPct" "Asian"
"White"          "Black"
Number of terminal nodes:  22
Residual mean deviance:  0.6308 = 656.7 / 1041
Misclassification error rate: 0.1486 = 158 / 1063
          test.data
prune.pred   0   1
         0  86  24
         1  56 290
[1] 0.8245614
```

Fig 8: Results of pruning on the classification tree

The pruned tree performed well, with 82.46% of the killings being correctly classified. It didn't however provide any improvement over the un-pruned tree. When I ran the `cv.tree` function, the appropriate

plots indicated that the ideal size for the tree was 22, which was the size of the original tree. Although the pruning didn't help, the classification tree still performed better than the logistic regression ones. The next natural step was to create random forests and see how they stacked up. After creating two bagged models, with different numbers of trees, and one random forest with `mtry = 6`, the best model ended up being the bagged one with 500 trees. It ended up having a classification rate of 89.47%. The race for first was close, as the random forest classified just 2 more killings incorrectly. Interestingly, the tree did not use any of the case specific parameters it was given. It only needed the county information.

Despite running a cross-validated gbm, and attempting to hypertune 5 parameters, I was unable to create a GBM classifier that performed better than the bagged model. Bagging and random forests aren't as susceptible to overfitting as other models, which likely contributes to its usefulness in this case. Despite their differences, both models highlighted the same things in terms of parameter importance.

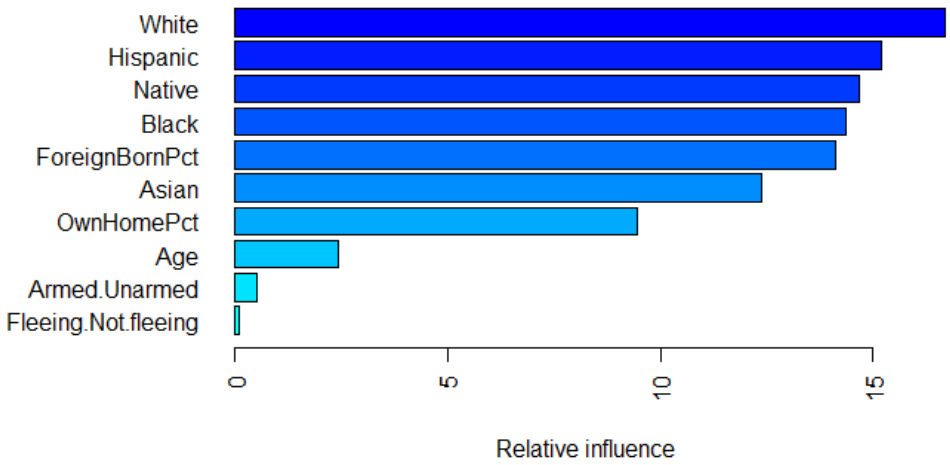


Fig 9: Relative influence of the hypertuned GBM's parameters

At this point, it was clear. My best model was able to outperform all the others, without taking into account whether a victim was armed or if they were fleeing. Seeing that all indications pointed to those two predictors not being important motivated me to drop them from my investigation. Due to the incompleteness of the database I used, not every killing (especially older ones) had information on every column. To include the armed and fleeing information, I had to drop almost 16000 incomplete rows. I originally chose to look just at 2015 as that's when the Washington Post began their database. For the second versions of each, I could now include all of the additional killings in my dataframe, so I chose not to limit the year.

Each model that ran this time around performed better than its counterpart with less data. However instead of the bagged model performing the best, the random forest with `mtry = 6` had the highest correct classification rate at 98.85%. I was very impressed by this model, especially seeing that my first model was half as effective.

From here, I decided that I wanted to quickly try to predict the race of the victim. I trained a few random forests and two GBMs, with the results mirroring those in prior experiments. Although the GBM performed ok, the random forest was best, classifying 63.21% of the victim's races correctly. Similarly, when I tried to predict what state a killing took place in, the random forest outshined the competition. It managed to hit a classification accuracy of 99.40% on the test set.

---

## Conclusion

In my investigation, I was able to train models to determine what state a police murder happened in, as well as if the location was in a state with a killing rate better or worse thsn that of the United States. When I began this project I was not confident that my models would show anything interesting, and boy was I wrong. The fact that I was able to make predictions off of a list of police killing coupled with the fact that nearly every model found that demographic information was most important leads me to one inescapable conclusion: systemic racism is alive and well in the US police force.

When we talk about bad apple cops, it is generally understood to mean those who actively perpetrate violence when none is needed, as well as those who let their biases shape how they do their job. But if the problem with policing in America was just a few bad apples, the patterns that were observed would not be as pronounced. Tens of thousands of lives were ended in a way that allows for the detection of statistically significant patterns. This doesn't happen from a few bad apples; it comes from a culture where we only call out those who have active hate and bias in their actions.

---

**References**


- https://fatalencounters.org/
- https://www.washingtonpost.com/graphics/investigations/police-shootings-database/
- https://www.elonnewsnetwork.com/article/2019/05/ethnicity-race-arrest-forms
- https://apps.urban.org/features/latino-criminal-justice-data/
- https://www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/download-the-data/
- https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D
- https://databricks.com/blog/2020/11/16/fatal-force-exploring-police-shootings-with-sql-analytics.html