CMSE 381 Final Project

A Perspective on Police Shootings

Andrew Wilt
PID: A55205577

---

1 Introduction

The United States has received a great deal of scrutiny in recent years as both citizens and foreigners have claimed that the country is fundamentally built on racist motivations. Specifically, law enforcement institutions have been put underneath a microscope as they have been called out as being one of the primary offenders and administers of inequalities that can be separated by race. This has led people to go as far as attempting to defund these institutions. This claim continues to persist as media outlets continue to publicize killings that involve white police officers and minority victims, a seemingly archetypal duo at this point. The purpose of this project is to introduce a new perspective of how we view police killings through the investigation of related literature and available data.

This paper includes a detailed K-Means Clustering model that aims to draw insights between variables in this data. With the variables provided in the used datasets, this unsupervised learning option seems to be the most viable route for creating meaningful questions. These questions mostly concern education and income, and how these two features are distributed among observations in the datasets.

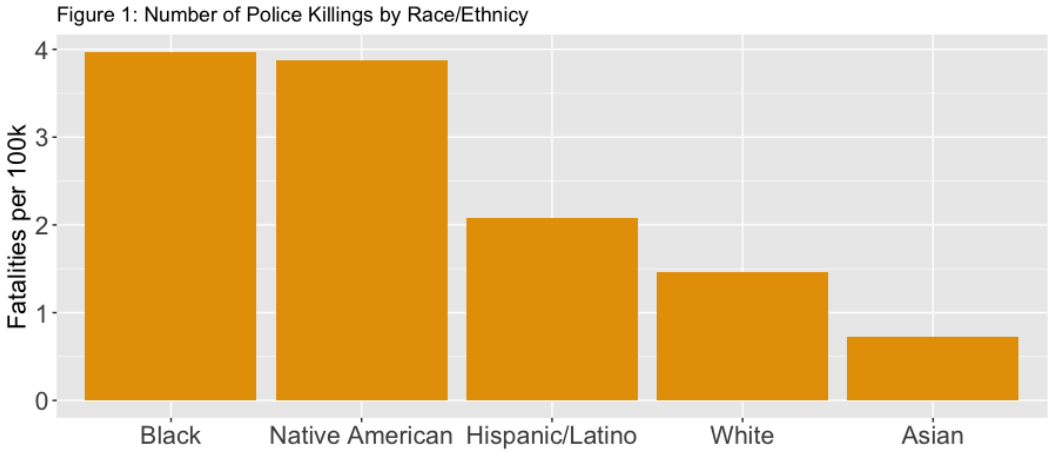
2 Related Work

In a related piece of literature titled “Fatal Force: Exploring Police Shootings With SQL Analytics” [1], publishers Eng and Wenig delve into the same data used in this project. While there is no prediction model involved in this article, these two authors view statistical summaries from various perspectives using SQL. The following subsections will outline their main techniques and briefly explain any meaningful observations and inferences.

2.1 Fatalities by Race

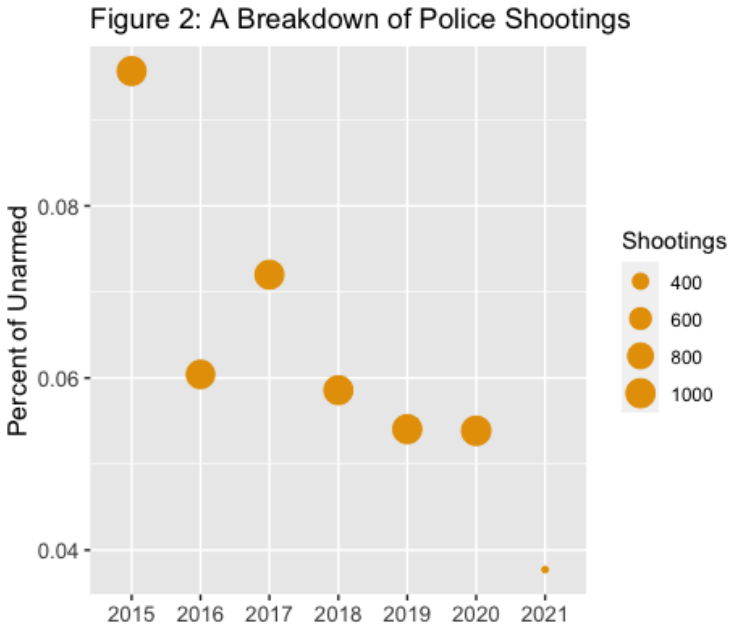
As mentioned earlier, the question of police shootings frequently comes into conversation with some consideration of race. Consequently, the Eng and Wenig article initially look at how the fatality totals are distributed based on race. This is visualized in Figure 1 below. Interestingly, while our graphs compare similar variables, there is a clear difference in that my data is characterizing the Black population to have a higher, per 100k, fatality rate. On the contrary, Eng and Wenig found this leader to be the Native American population. This disparity can most

likely be explained by the evolution of the data, as my results include the newly added data (since the birth of this article in November of 2020).



2.2 Breakdown of Fatal Shootings

The second point of interest in this article showcases all the different combinations of variables in a beautiful SQL plot. The R programming language does not offer this convenience, so Figure 2 below aims to emulate the main characteristics of this plot.



This plot captures the number of killings per year, and the percent of these where the victim of the killing was unarmed. Thankfully, this percentage appears to be trending down. Either way, it is interesting to observe that there were few instances where the victim of the killings was unarmed.

2.3 The Use of Body Cameras

Something that was highlighted Eng and Wenig's article was the claimed "decline" in the use of body cameras in these fatal shooting incidents. This claim was reinforced by a similar table of the one displayed below (Figure 2).

Date	% Body Camera	Number of Incidents
2015	8%	993
2016	15%	960
2017	11%	986
2018	12%	990
2019	14%	999
2020	17%	1021
2021	18%	265

Their claim was a valid interpretation of the apparent decrease in body-cam footage use following the year of 2016. However, it is worth mentioning that apart from this abnormally high year, there has been a steady increase in the use of body camera footage for the recorded shootings in this dataset from 2015 to 2021. In other words, the percentage of body camera uses for these shootings has more than doubled in 5 years.

2.4 State Comparison

The third analysis of interest is one that takes a close look at the total number of fatalities per state. The state data its standardized according to state population. Oddly, using the same data set and the same year, I could not achieve the same output that was depicted in the article. This could be due to this article being published prior to the end of the year that the data represented. This might also be explained by the higher number of total fatalities for each of these states shown below in Figure 4.

Figure 4: Number of Police Killings by State

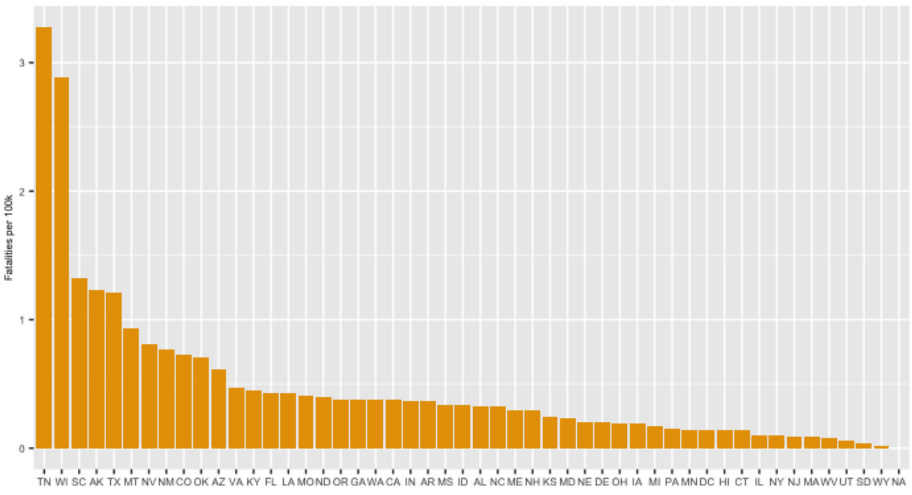


Figure 4 indicates that the state of Tennessee had the highest amount of police shootings during 2020 whereas Eng and Wenig observed it to be Alaska. We would imagine these totals to jump around from year to year. But, it might be interesting to see how they have averaged out over time in respect to each individual state.

3 Dataset

Two datasets were used in this project. The former one, which was used in the article that was discussed above, consists of almost entirely categorical data. Categorical datasets leave us with few options which include creating proportions and totals. The other dataset inherently includes numerical data in the form of percentages, which was eventually used in the K-Means Clustering model for its dynamic ability.

As discussed, this project creates motivation in creating a classification model. Due to all of the observations having a "fatally shot" label, there was no opportunity for classification that I could think of.

4 Methods

Because this dataset does not contain a label or an output value, the main focus in this section is going to be on unsupervised learning. With that, we will choose K-Means Clustering. Because there are a vast number of variables in this dataset, we could just as well go with a supervised learning technique and choose one of the several categorical variables as a label. But because the focal point of this dataset is the fact that these individuals were killed, we are going to attempt to better understand what the data says about these people that are shot and killed. Later I will discuss ways in which we could implement supervised learning for classification in this topic.

This section will briefly describe the intuition behind the K-Means Clustering technique. For the purposes of this project, the implementation of this method will utilize the available K-Means Clustering functions from {base} R.

To put it simply, this technique aims to cluster data that has similar attributes. This can allow us the make sense of data that doesn't have an output or label. Our goal in K-Means Clustering (KMC) is to minimize the within-cluster sum of squares for a set of n observations with dimension d.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

The algorithm for this method typically takes an iterative approach where clusters are assigned and calculated using the least-squared Euclidian distance. This is done continuously until there are no reassignments regarding an observation nearest mean. Unsurprisingly, this algorithm is not guaranteed to find the optimum.

5 Experiment and Discussion

Our initial experiment contains a comparison between Comparable Income (household income divided by population income), and College (the share of the population of 25+ with at least a BA).

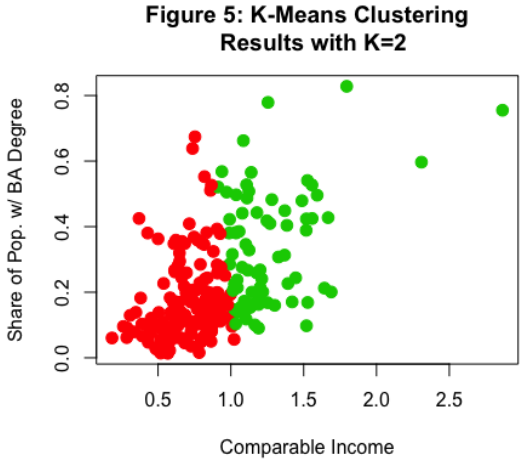
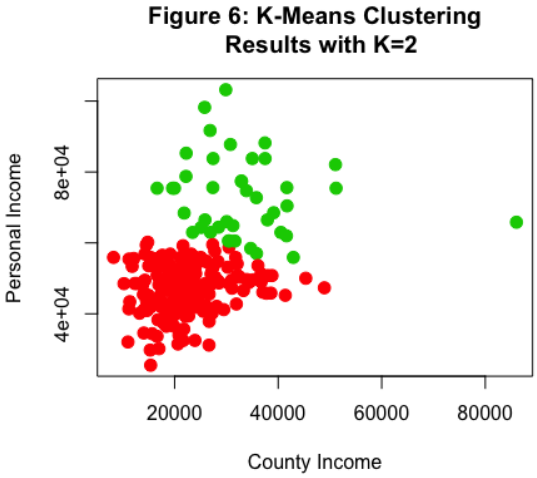


Figure 5 shown above shows the results of running a K = 2 K-Means Clustering model with features that describe the education of the local population and the comparable income, which is calculated by diving the household income by the county income. The results of this model using these features are interesting as there is a clear separator around comparable income equally 1.0. This indicates that having an above average income, or a below average income, is a significant feature in this dataset. Additionally, one might notice that the majority of these observations are in the Red class where the population education rate is relatively low.



Similarly, the model in Figure 6 utilizes the same KMC technique. However, it's features of interest are personal income for a given observation and county income for a given observation. It's interesting to see that there is certainly a group with an income that is about twice the size of

their country's income. Still, this group, holding the majority, seems to measure right around the national income at about $40k. There is also another group, although this one has much more dispersion. This group appears to make a very high personal income, one that is about double the national average. This group, shown in Red, also appears to have a relatively high county income.

The final plot shown in Figure 7 uses this same K = 2 model. Here, we can notice that there are two groups where one live in a population where a high percentage of individuals older than age 25 have a BA degree. In these observations, they also appear to have a very low unemployment rate. Additionally, the other group has a very low percentage of their population that has a BA degree. Something interesting in this group is that their unemployment rate has much more disparity than that of the Red group. This suggests that a number of these observations have abnormally high unemployment rates.

6 Conclusion

One of the great fallacies of statistics is that with correlation, we cannot assume causation. In the article discussed in this report, the main function of the authors' analysis was to use a univariate model to determine causation. I think it's important to perform multivariate analyses for this topic of police killings since there are so many potential significant factors. Throughout implementing this model and recreating the visualizations from the aforementioned model, I often longed for variables that weren't included in either of the used datasets. Without a viable "label" variable, it was difficult to ask a supervised learning question.

An important feature that could answer some of these popular questions involving racial disparities is the outcome of the trial regarding these incidents. Assuming that many of these police shootings ended up in trial, I think having a class label of "justified" and "unjustified" could give some insight into how these racial biases might play out. A model with these described standards could create a classification model that could even be of use in the court of law

References

[1] Eng , Chengyin, and Brooke Wenig. “Analysis of Police Fatal Shootings in the U.S.” *Databricks*, 2 Dec. 2020, databricks.com/blog/2020/11/16/fatal-force-exploring-police-shootings-with-sql-analytics.html.

[2] “List of U.S. States by Population.” *Wikipedia*, Wikimedia Foundation, 1 Mar. 2021, simple.wikipedia.org/wiki/List_of_U.S._states_by_population

[3] “K-Means Clustering.” *Wikipedia*, Wikimedia Foundation, 17 Apr. 2021, en.wikipedia.org/wiki/K-means_clustering.

[4] “Demographics of the United States.” *Wikipedia*, Wikimedia Foundation, 16 Apr. 2021, en.wikipedia.org/wiki/Demographics_of_the_United_States#Race.