Data-analysis of Police Shooting Cases in the US since 2015

I. Introduction

Following the 2014 shooting of Michael Brown in Ferguson, Missouri, and several subsequent controversial police shootings, the deadly burden of police violence has become one of America's urgent public health crises, and its negative effects are still growing. Excessive use of force by the police and unprovoked shootings of civilians have not only claimed the precious lives of innocent people, but also prompted more social groups, especially marginalized groups, to distrust police enforcement, government departments and public health infrastructure. If social issue cannot be solved effectively, it will inevitably lead to the intensification of social conflicts and become one of the key factors leading to social instability.

Public dissatisfaction with the deadly police shooting continues to grow. Many current media and research papers analyze factors such as race, gender, region, and victim community to find widespread inequality, especially color inequality, from 2015 to 2020, Blacks, Indigenous Peoples and People of Color (BIPOC) had significantly higher mortality rates compared to whites across the entire victim group (Native American RR=3.06, Black RR=2.62, Hispanic RR =1.29) and unarmed victims (Black RR=3.18, Hispanic RR=1.45)(Lett et al., 2021). Although policing reform has been gradually pushed forward by public protests and related social movements, data from The Washington Post since it began tracking police deaths in 2015 shows that the number of US police deaths shooting in 2021 will still reach an all-time high of 1,055.

American police are in an unprecedented predicament. Due to the relatively small scale of police shooting deaths in general and the fact that much of the data disclosed is one-sided and incomplete, this has led to some fragmentation of the literature and public misunderstanding of police shooting deaths (Fyfe, 2002; Shane, 2018). Among them, the discussion of racial prejudice also has a strong social incitement, which has released a "crisis" signal for the stable development of society.

To avoid a more serious and broader public health crisis caused by police violence, our group will conduct detailed tracking and comprehensive analysis of each police shooting from more dimensions than other research literature. We not only conducted data analysis on multiple factors such as gender, age, place of residence, income level, education level, and mode of death of victims but also studied the impact of news reports and public opinion on the occurrence of police shootings.

II. Related Work

1. Hemenway, D., Berrigan, J., Azrael, D., Barber, C., & Miller, M., “Fatal police shootings of civilians, by rurality,” *Preventive medicine*, 134, 106046, 2020

David Hemenway et al. examine fatal police shootings across the urban-rural divide. The study used a dataset of police shooting deaths in the United States tracked by The Washington Post from 2015-2017. Since there is currently no uniform national standard to define whether an area is urban, rural

or suburban, and definitions vary widely, five different definitions are used to describe places within the urban-rural boundary to minimize data errors in the analysis results (Hart et al., 2005). Stacked up data from all five classification schemes from The Washington Post over three years and examine the rates of fatal police shootings by levels of urbanicity, using race/ethnicity to stratify the results. The results of the data analysis show that violent crime, gun violence and murder are more of an urban phenomenon in the United States, but not fatal police shootings, which occur in equal proportions in urban and rural areas. In addition, black victimization rates in rural areas were the same or relatively lower than in urban areas, but the opposite was true for whites (Hemenway et al., 2020).

2. Nix, J., & Shjarback, J. A., "Factors associated with police shooting mortality: A focus on race and a plea for more comprehensive data," *PLoS one*, *16*(11), e0259024, 2021.

This paper analyzed fatal and nonfatal shooting data from four states data including Florida, Colorado, Texas, and California to examine possible factors that lead to victim mortality.
They first compare the racial disparities with fatal/injurious police shooting. Then, they used cross tabulate to see the correlation between shooting outcomes and factors like race, gender, age, deadly weapon, country access to trauma care.
Besides, this paper used a logistic regression model to predict victim mortality in each state.

3. Schwartz, G. L., & Jahn, J. L., "Mapping fatal police violence across US metropolitan areas: Overall rates and racial/ethnic inequities, 2013-2017," *PloS one*, *15*(6), e0229686, 2020.

The author used data from Fatal Encounters with an inverse-variance-weighted multilevel model to predict the overall racial inequities in America from 2013 to 2017.
The sample for estimating the overall fatal rate is 5494, among which 94.2% was from a gunshot wound. It compared estimates in a sensitivity analysis using the sample data to those from models which included all reported dead causes. They used ArcGIS version 10.6.1 to geocode the Fatal Encounters data. They combine data on fatal police violence to MSA lever and merged it with race-specific population data from Census Bureau's American Community Survey 5-Year Estimates from 2013 to 2017.
They used multilevel Poisson models to predict the yearly incident rate and incident rate ratio for police killing in each MAS and proposed two estimated models sets.
They conducted a sensitivity analysis comparing the incident rates and standard errors and formally examined whether the MSA-level rates exhibited spatial autocorrelation using the GEoDa, which is a statistical software.
As a result, they found that different geography differs in the incidence of fatal police violence in America.

4. Mentch, L., "On racial disparities in recent fatal police shootings," *Statistics and Public Policy*, *7*(1), 9-18. 2020.

This paper uses five open datasets from Washington Post (WP), United State Census Bureau (DEM), Federal Bureau of Investigation (LEE), United States Department of Justice(ARREST), and ICPSR to analyze the racial disparities in the fatal police shootings between January 2015 to July 2016 in America.

The author doesn't think victims wit\h a firearm or not is a possible factor as it is hard to determine if a suspect is armed or just carries something by accident.

Instead, the author used WP data concentrating on the populations and used a resampling approach to estimate the entire expected victims from each race more comprehensively.

As a result, the author found that there is a distinct racial disparity in police shooting victims, but the disparity became less distinct after the author considered the local arrest demographics.

Based on these literature review and dataset descriptions, we raise the following research questions:

RQ1. What attributes distribution in people suffering fatal killings? like race/gender/age···

RQ 2. What distributed pattern of the number of fatal shooting cases with the poverty rate and education level grouped by state?

RQ 3. What is this killing tendency pattern of fatal shooting in the US by a police officer in the line of duty over time? Did such tragedies issues happen frequently?

RQ 4. what armed level distribution in this fatal killing distribution?

RQ 5. Are there any characteristics of the distribution of shootings in various states in the United States? Which state has the most shootings in the US (most dangerous state)

RQ 6. Is that glee behavior is more likely to trigger polices using terser?

RQ 7. Is there a declare trend of the number of police shooting cases after widely social discussion and reporting? (Combined the prior news report)

RQ 8. What are the characteristics of the top three states with high shooting rates in terms of income, education level, racial distribution, poverty level, etc.?

At the same time, based on our findings on the dataset of available variables, we wanted to predict the race or psychiatric state of the deceased.

RQ 9. Based on this dataset, can we predict whether a victim has signs of mental illness?

RQ 10. Can we predict the likely race of the victim based on the remaining variables?

III. Methodology

According to research questions RQ1-8, these refer to data analysis methods.

In data analysis and visualization, we use histograms, scatter diagrams, heatmaps, box plots, and line charts. The bar chart is mainly used to describe some features distribution, the number

and features of cases at different times, and the consequences of different weapons and escape methods. The heatmap and scatter diagram shows the correlation coefficients between different variables. The heatmap also use to show the features of geographical distribution. A line chart is used to show the trend after public discussion of such incidents. The box plot compares the states with the highest number of such incidents in terms of different measures, such as poverty rates and educational level.

Method	**Advantages**	**Disadvantage**
Bar chart	Showing frequency distribution of each data. Summarize a large data set in visual form. estimate key values briefly. Be easily understood due to its widespread use in business and the media.	Be easily manipulated to yield false impressions.
Heatmap	Giving a direct overview of the result. Providing visual paths to understanding numeric values. Using shapes and colors to convey information is easier to understand.	Color conveys sensory information, and it is not easy to distinguish specific values.
Line chart	Can show the change trend of data, reflect the change of things.	It is not intuitive to see the number and percentage of each part.
Box plot	The overall centralized distribution of data and outliers can be observed.	The skewness and tail weight of data distribution cannot be accurately measured.
Scatter diagram	Uncover relationships between data and find correlations between variables.	Relationships in scatter plots may be affected by a third variable.

According to research questions RQ9-10, these refer to predict the race or psychiatric state of the deceased. We resorted to the algorithm of Random Forest in RQ9 and the algorithm of K nearest neighbor in RQ9.

Random forest is composed of many decision trees, and there is no relationship between different decision trees. When we carry out the classification task, new input samples come in, and each decision tree in the forest will be judged and classified separately. Each decision tree will get its own classification result, which one of the classification results of the decision tree

is classified. At most, then the random forest will treat this result as the final result.

	Advantages	Disadvantages
Random forest	● Decrease data overfit. ● Determine the importance of different features. ● No dimensionality reduction, no need to do feature selection for those many features' data. ● Maintain data accuracy if a large part of the features is missing.	● Attributes with more value divisions will have a greater impact on Random Forests.

Table 1 Random Forest algorithm advantages and disadvantages

For prediction of race of victim, the original data set includes Native American, White, Black, Asian, Hispanic, and other people, that is saying race attribute has multiple values. So, we tried to apply another supervised machine learning to classify different races in victim. Specifically, we resorted to K-nearest-neighbors (KNN) algorithm to predict the victim's race.

In KNN classification, the outcomes will be clusters, and the classification of an object is determined by its nearest neighbors, and the most common classification among the k nearest neighbors (k is a positive integer, usually small) determines the class assigned to the object. In our dataset, we expected six clusters (native American, Asian, Black, Hispanic, White, and other races) could be classified out.

	Advantages	Disadvantages
K-nearest-neighbors (KNN)	● KNN is very easy to implement as it only need calculated the distance between different points based on data of different features. ● New inputted data won't affect the model.	● Sensitive to noisy data and missing data. ● Does not work well with large dataset. ● Data in all the dimension should be normalized and standardized properly.

Table 2 K-nearest-neighbors (KNN) advantages and disadvantages

IV. Data Description

Our datasets are from the **Washington Post which has been compiling a database of every fatal shooting** in the US since 1st January 2015. The main dataset is about victims which have 14

attributes introduce the details about the victim's features of each police killing case, including personal information, death information ID, and information related to the case. The personal information covered ID, name, age, gender, race, living city, and living state, and the death information went through with death date, and manner of death if the victim was armed. Also, some other relevant details about the, if there are signs of mental illness, occurred in the victim, if the victim carried a boy camera if the victim tried to escape, and their threat level. These data can help us to recognize the overall features of the victims and to recall some details during the miserable thing that happened like if the victim were trying to survive or if they were meant to record something.

Apart from that, there still are 2 datasets from US census data that revealed the household income station and the percentage of poverty of each state and each city. And 1 dataset declares the Percentage of a given city's population above 25 that has graduated high school which can tell us a region's education condition and the other dataset told us the share race condition by the city that includes the percentages of white, black, Asian, Native American and Hispanic people respectively in each state and city. If we want to know more or try to figure out the reason for a social event, we always need to go back to the social states, therefore, this kind of relevant information enables us to know more about the fatal killing cases from the side.

V. Experimental Results

RQ1. what kind of people are more likely to be shot? like race/gender/age…

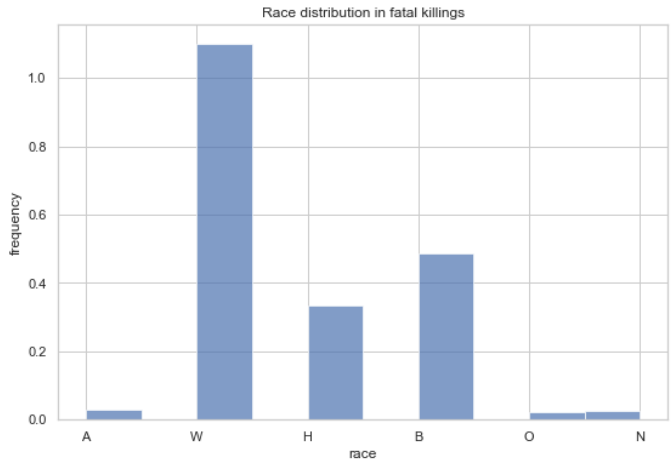


Figure 1 Race distribution in fatal killing

As **Figure 1** illustrated, the white race attributes the highest account of the fatal killing. The following fatal killings happened in the race of black people and Hispanic people sequentially. The race of native Americans, Asians, and others counts comparably lower rates of suffering fatal police Shootings.

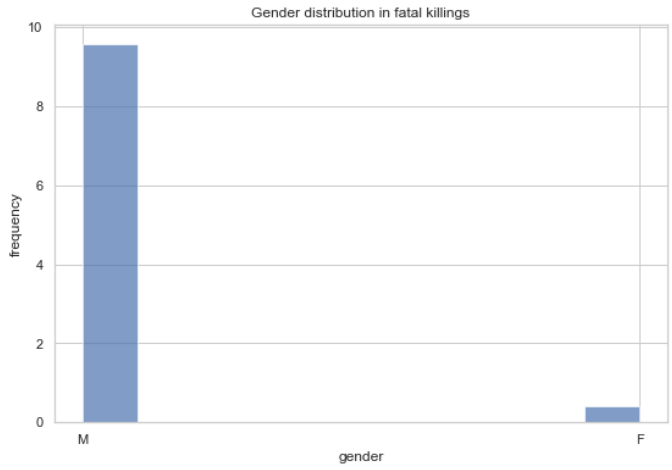


Figure 2 Gender distribution in fatal killings

In gender category, male, as the chart demonstrated, owns quite higher proportion of encountering fatal shooting compared with female. Additionally, the differences in the probability of death between genders is significantly huge **(see Figure 2).**

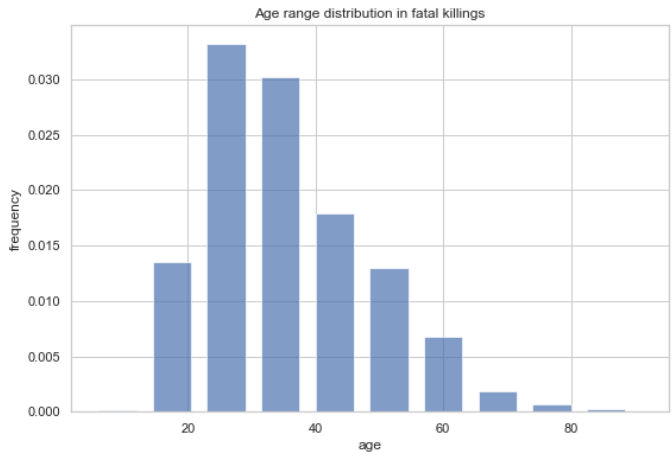


Figure 3 Age range distribution in fatal killings

In fatal police shootings, age range distribution also belongs to our analysis. From this chart, it can be concluded that the younger cluster (between 20 to 40) occupy a considerable proportion in police fatal killings **(see Figure 3).**

RQ2. What distributed pattern of the number of fatal shooting cases with the poverty rate and education level grouped by state?

To investigate whether police killings are related to poverty rates and education levels, we first use two scatter plots to show the number of police killings in each state and their distribution with poverty rates and education levels **(as shown in Figure 4).** Education is expressed as the percentage of people who have completed high school, the higher the percentage, the higher the state's education level.

As you can see from the scatter chart, despite some noise, poverty rates are more concentrated on the lower left, while education levels are more concentrated on the upper left. This means that places

with lower poverty rates have lower rates of police shootings. And places with higher levels of education have lower rates of police shootings.

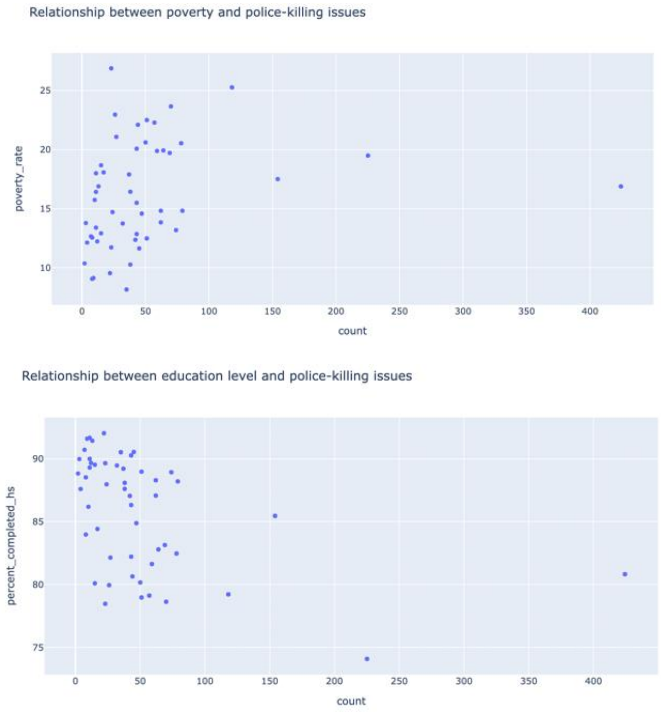


Figure 4 Relationships between poverty and police-killing; Relationships between education level and police-killing.

A comparison of the two charts shows that the scatter plot of education levels is more concentrated. This suggests that the greater the correlation between the level of education in each state and the incidence of police shootings. We used a heatmap (in **Figure 5**) to show the correlation score between variables. As you can see from the absolute value of the score, education is more correlated with the occurrence of a police shooting.

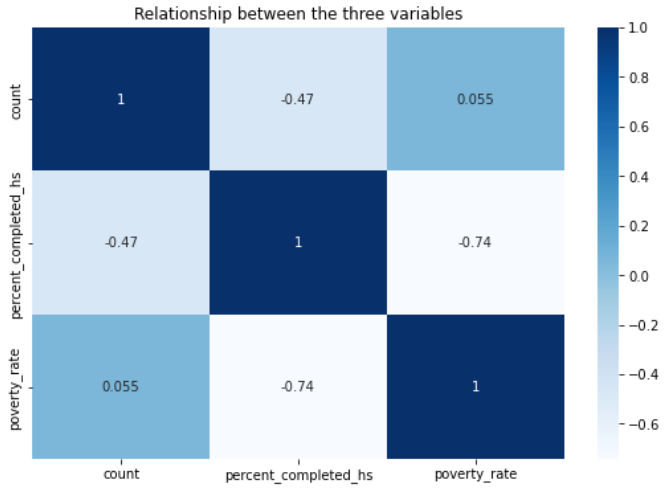


Figure 5 Relationship between the three variables

RQ3.What is this killing tendency pattern of fatal shooting in the US by a police officer in the line of duty over time? Did such tragedies issues happen frequently?

To study the trend of police shootings in the United States over time, we filtered the dataset by different years. The number of police shootings in different years is shown in a bar chart indexed by time. As can be seen from the bar chart, the number of such incidents has decreased year by year from 2015 to 2017 (as shown in **figure 6)**.

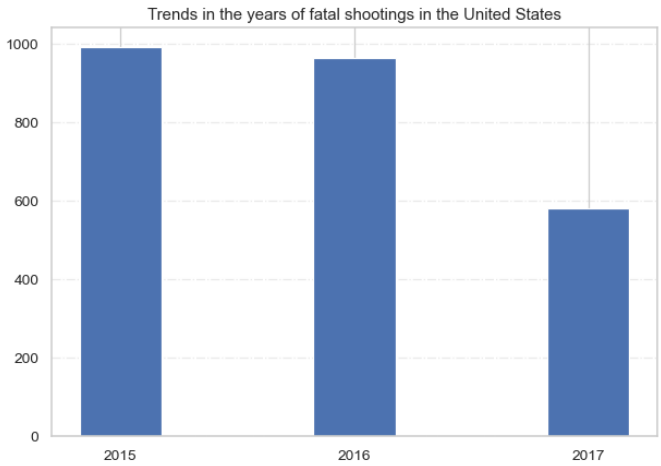


Figure 6 Trends in the years of fatal shootings in USA

When we break down the demographics of police shootings, you see this: Even as police shootings have declined over time, the percentage of white people killed by police has increased. The frequency of such incidents among Asians also rose slightly over time. The percentage of black people killed by police fell more obviously than others.

In terms of the sex ratio, there doesn't seem to have been much change over time. It still happens far more often in men than in women (in figure).

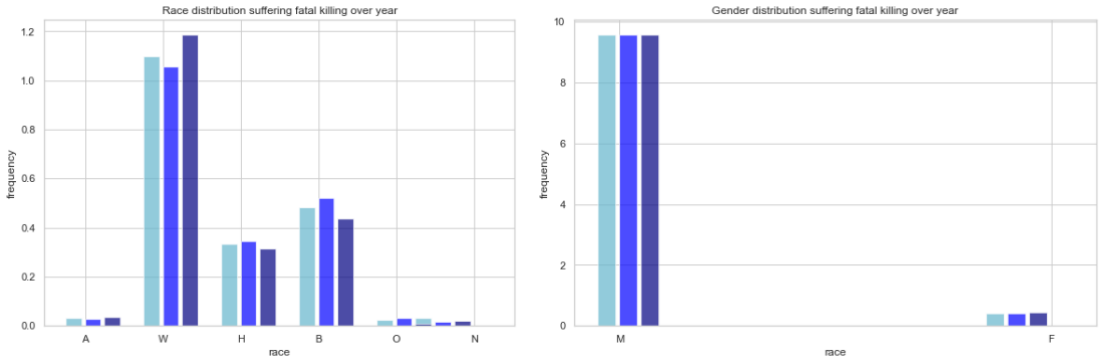


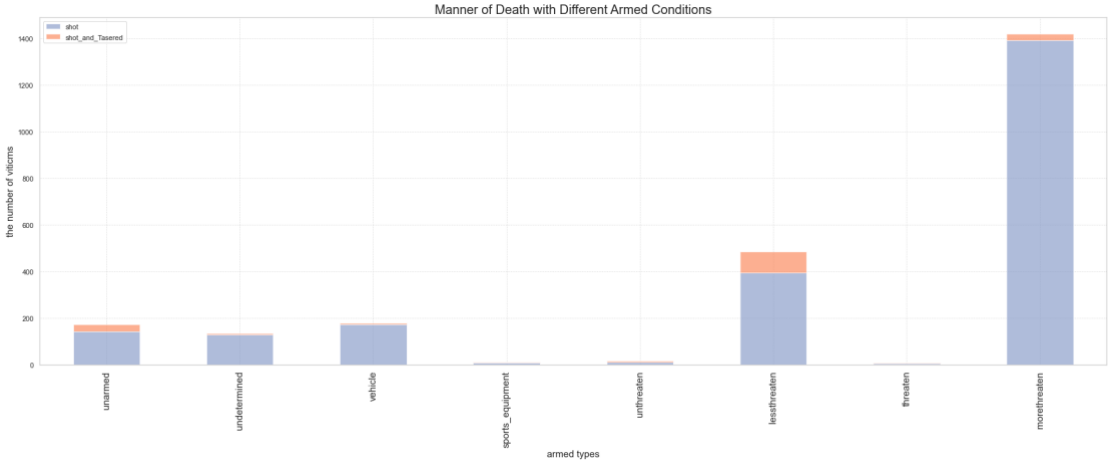
Figure 7 Race distribution over year; Gender distribution over year

RQ4.What armed-level distribution in this fatal killing distribution?

For answering this question well, 68 types of weapons have been divided to 8 types first. Due to the large number of the weapon type, it is hard to find a uniform standard, so that multiple considerations have been involved. And the classification results are shown in the table.

unarmed	unarmed

undetermined	undetermined, unknown weapon
vehicle	vehicle, motorcycle
sports equipment	baseball bat, baseball bat and bottle, baseball bat and fireplace poker
unthreatened	air conditioner, beer bottle, flashlight, garden tool, hand torch, pen, rock, screwdriver, stapler, straight edge razor
Less threatened	ax, baton, bayonet, blunt object, box cutter, brick, carjack, chain, contractor level, cordless drill, crossbow, crowbar, fireworks, flagpole, glass shard, hammer, hatchet, knife, lawn mower blade, machete, meat cleaver, metal hand tool, metal pipe, metal pole, metal rake, metal stick, oar, pickaxe, piece of wood, pipe, pitchfork, pole, pole and knife, scissors, sword, tire iron, nail gun
threatened	bean-bag gun, chain saw, sharp object, shovel, spear
more threatened	hatchet and gun, machete and gun, gun, gun and knife, guns, and explosives



Then, we can immediately see that victims who carried less threatened weapons were the largest group who have been used Tasered. And the second group is the unarmed people. This phenomenon also been shown among victims who carried more threatened weapon. By contrast, undetermined group, vehicle group, sports equipment group, unthreatened group and threatened group are less likely to suffer this kind of hurt.

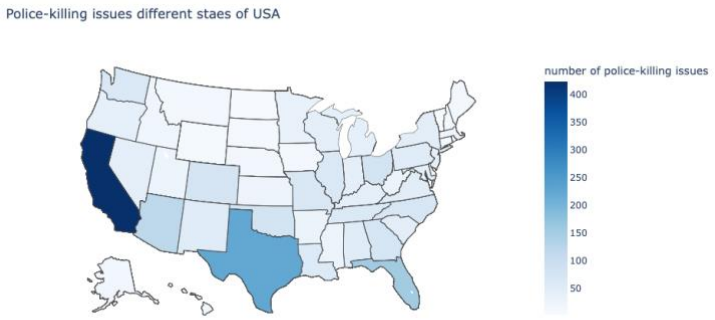
RQ5. Are there any characteristics of the distribution of shootings in various states in the United States? Which state has the most shootings in the US (most dangerous state)

As shown in figure, we used a heatmap of the United States to show the distribution of police shootings and killings in different states. The darker the color, the greater the number of police shootings.

Through the map, we can see that the southern states have far more incidents than the northern states.

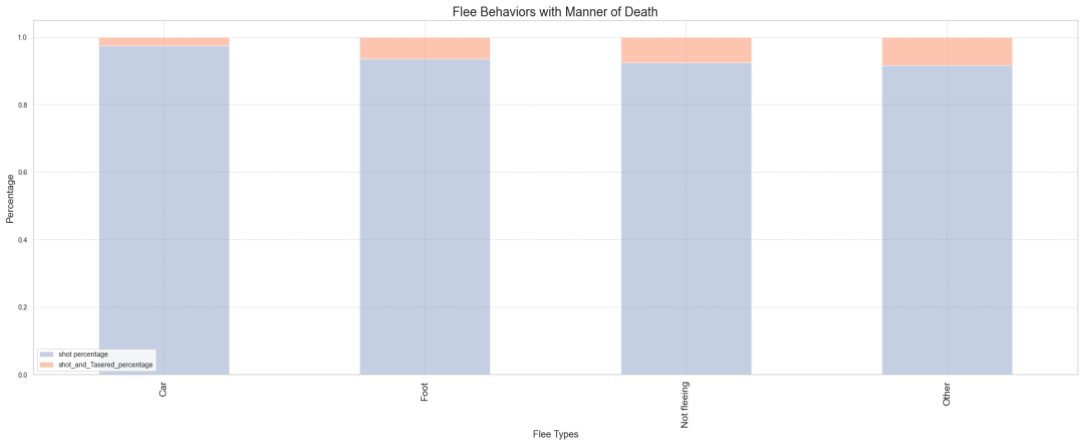
Coastal and border states have more cases than central states.

California (CA) is the state with the highest number of police shootings, followed by Texas (TX). Both states are in the southern PART of the United States. California is a coastal state and Texas is near the border.



RQ6. Is that true that flee behavior is more likely to trigger polices using tasered?

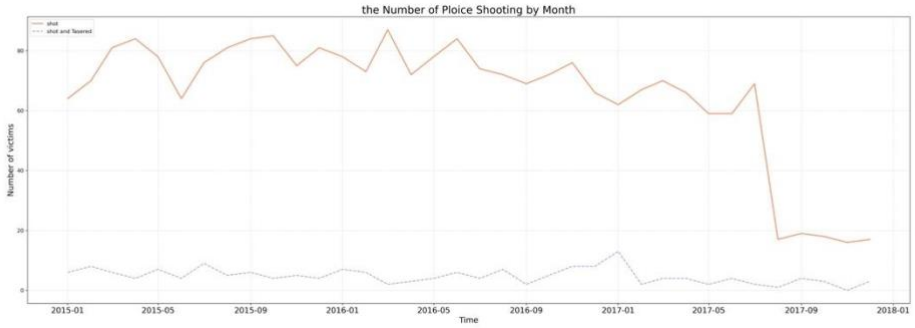
As we can see, there are 3 kinds of way of fleeing, by car, on foot, others, and surprisingly, no fleeing people are most likely to be Tasered by police except other people which are unknown. Next are those people who tried to escape on foot and people fled by car have the smallest chance to be tasered.



RQ7. Is there a declare trend of the number of police shooting cases after widely social discussion and reporting?

From those two pictures, we can know that the reported number of victims were fluctuated between 2015 to 2018. In total, the number of victims maintains a stable number between 60 to 90 per month, and suddenly showing a declare trend after July 2017. To be specific, the number dropped to less than 20 and then keep stable. One of the explanations might be the law. 14"Black Lives Matter" bills have been proposed in many states in U.S.A in 2017. Although most of billed failed, they still have effected a lot, and even they became the truth in some states like Louisiana and

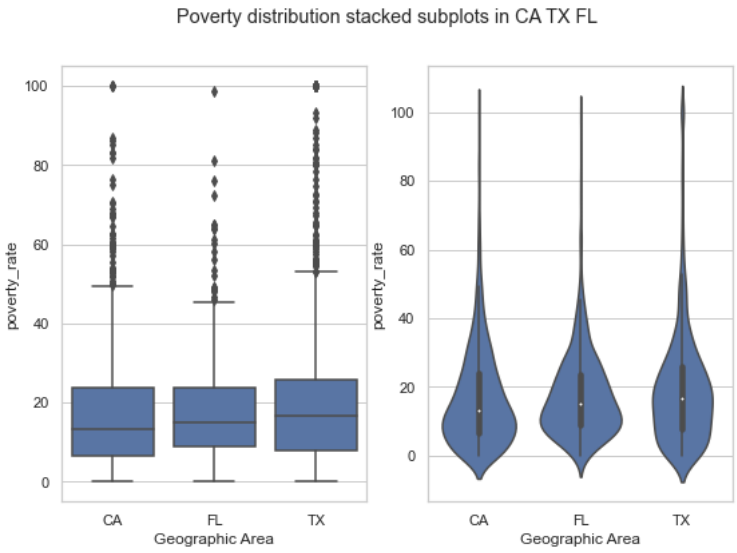
Kentucky.



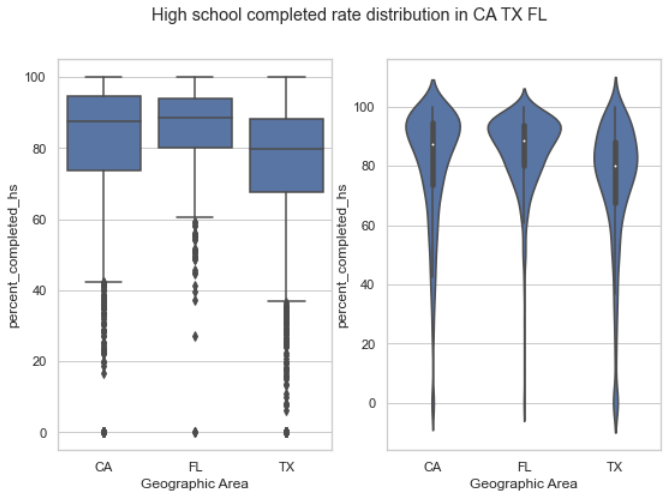
RQ8. What are the characteristics of the states and territories with high shooting rates in terms of income, education level, racial distribution, poverty level, etc.?

As we know, from above chart, "CA" (California),"TX" (Texas),"FL" (Florida) are top 3 that fatal killings.

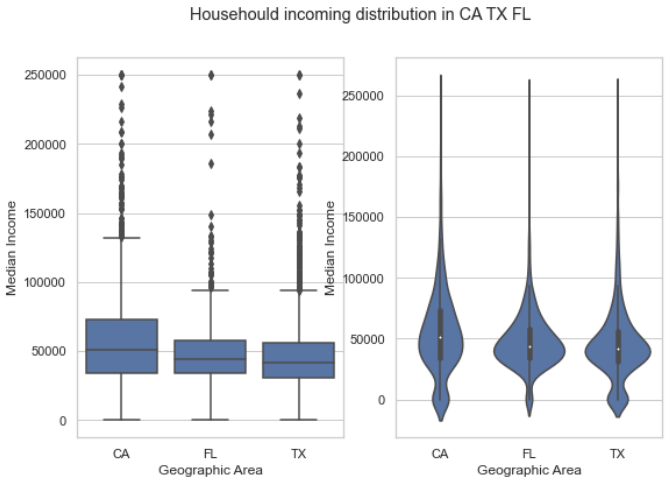
And then, we need to check these states' circumstance about income, education level, racial distribution, poverty level.



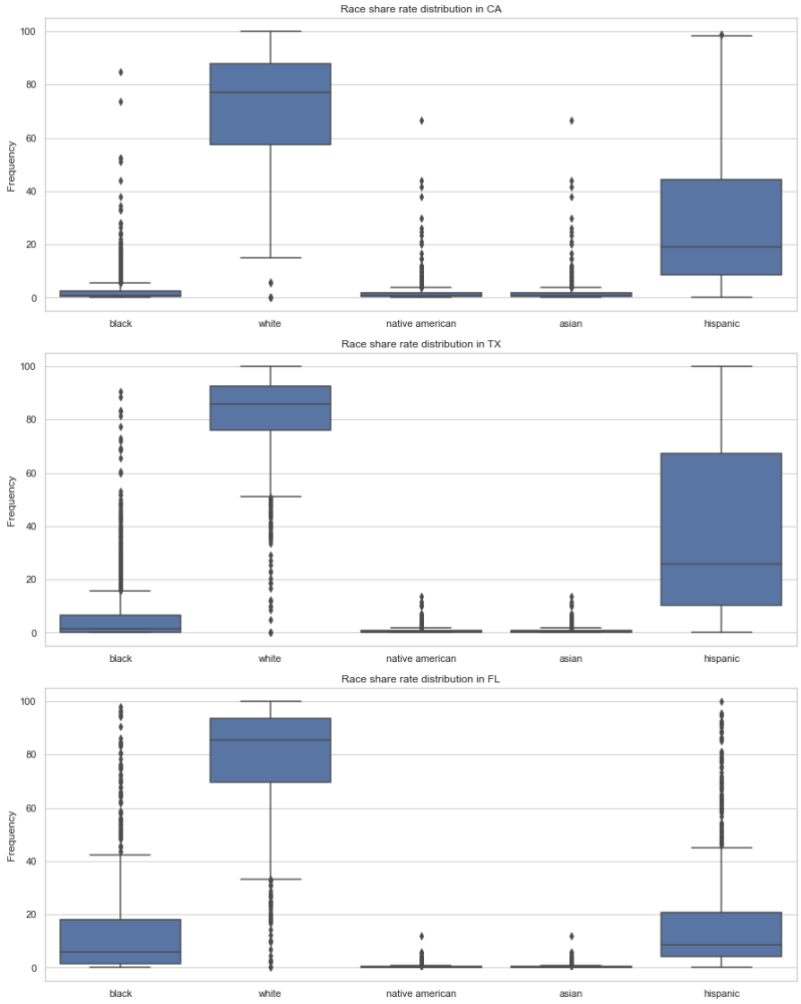
The box plot shows a pattern of poverty rate spread in states with high shooting rates. We can tell these three states' poverty rate fasten between 0% and 20%, their third quartile, Q3, mainly fastened on 20% of poverty rate. While existing lots of outliers, we draw a paralleled violin plot showing distribution pattern of state poverty rate. As right figure demonstrated, the poverty rate patterns in three states are left skewed and show a long tail.



We can tell these three states' high school completed rate fasten between 70% and 90%, their third quartile, Q3, mainly fastened on 80% of high school completed rate. While existing lots of outliers, we draw a paralleled violin plot showing distribution pattern of state high school completed rate. As right figure demonstrated, the high school completed rate patterns in three states are right skewed and show a long tail.



We can tell these three states' household incoming fasten between 25000 and 750000, their third quartile, Q3, mainly fastened on 50000 of household incoming. While existing lots of outliers, we draw a paralleled violin plot showing distribution pattern of household incoming. As right figure demonstrated, the household incoming patterns in three states are left skewed and show a long tail.

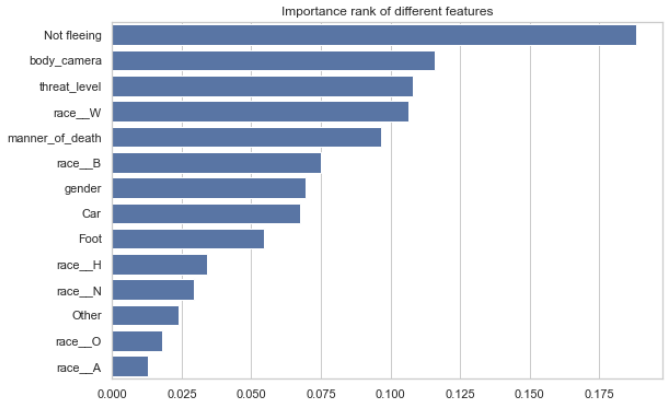


As it shown, white people count higher proportion in both states. While Hispanic race people spread comparably wider. And the native American and Asian people spread uniformed less in CA, TX, FL.

RQ 9. Based on this dataset, can we predict whether a victim has signs of mental illness?

Because most of our features are categorical variables, and because Random Forest algorithms prefer to deal with numbers rather than strings, we had to convert our features to numerical values. We converted the following columns to binary numbers and one-hot encoded data using the pandas "get_dummies" method.dummy columns:

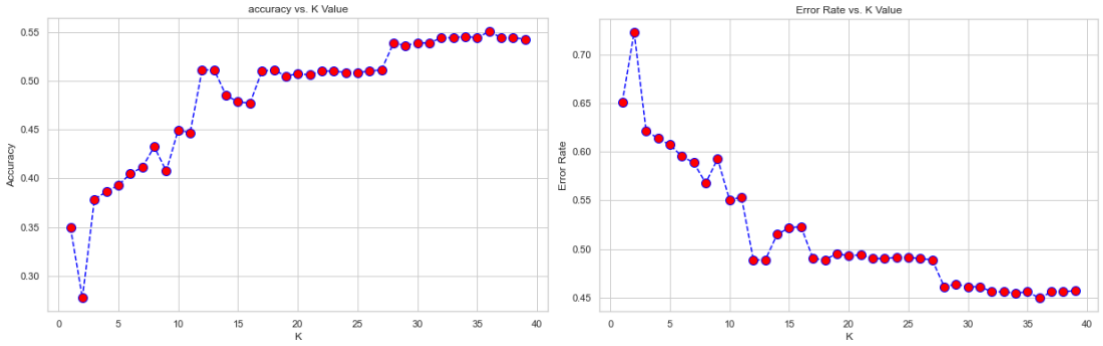
- Race: convert into dummies form
- Gender: M:1; F:0
- manner_of_death: shot and Tasered:1 shot:0
- flee: convert into dummies form
- threat_level: attack:1 other:0 undetermined:0
- signs_of_mental_illness: shot:1 shot and Tasered:0



With the test model, we got an accuracy score of 0.747534516765286. This means that our value from the variable above, is the percentage of correctly predicted labels. There is still some room for improvement regarding improving the model, such as adding more data, getting rid of potential outliers, and changing the number or categorical of features our model takes in. We are satisfied with the result of our model so far.
We also drawn features importance rank, as it illustrated, the category of "flee" attributes the comparably higher counts.

RQ 10. Can we predict the likely race of the victim based on the remaining variables?

Because we cannot get known to a propriate algorithm to predict the category that its attributes with more value divisions yet. And then we wonder what if we testify KNN algorithm's accuracy score with k = 6. We only get the accuracy score of 0.4047306176084, which is not deal. So, we resorted to examine the proper k value by considering accuracy and error rate. And we calculate the following plots. We got known that k=35 would be good choice. We move to testify the accuracy score of k=35, and the result is 0.5440210249671484. Better than k=6, while it is still not ideal for us. So, we drawn a conclusion that the distinguish between victim race is not significant.



VI. References

Fyfe, J. J. (2002). Too many missing cases: Holes in our knowledge about police use of force. *Justice Research and Policy*, *4*(1-2), 87-102.

Hart, L. G., Larson, E. H., & Lishner, D. M. (2005). Rural definitions for health policy and research. *American journal of public health*, *95*(7), 1149-1155.

Hemenway, D., Berrigan, J., Azrael, D., Barber, C., & Miller, M. (2020). Fatal police shootings of civilians, by rurality. *Preventive medicine*, *134*, 106046.

Lett, E., Asabor, E. N., Corbin, T., & Boatright, D. (2021). Racial inequity in fatal US police shootings, 2015–2020. *J Epidemiol Community Health*, *75*(4), 394-397.

Shane, J. M. (2018). Improving police use of force: A policy essay on national data collection. *Criminal justice policy review*, *29*(2), 128-148.