US Fatal Police Shooting Analysis and Prediction

Yuan Wang
University of Rochester
yuan.wang1@simon.rochester.edu

Yangxin Fan
University of Rochester
yfan24@ur.rochester.edu

Abstract

We believe that "all men are created equal" [11]. With the rise of the police shootings reported by media, more people in the U.S. think that police use excessive force during law enforcement, especially to a specific group of people. We want to add our two cents point of view by multi-dimensional analysis to reveal more facts than the monotone mainstream media. The more facets we understand the problem, the better solution that the whole society could approach.

1. Introduction

Our report has three parts: First, we analyzed and quantified fatal police shooting news reporting deviation of mainstream media. Second, we used FP-growth to mine frequent patterns, clustered hotspots of fatal police shootings, and brought multi-attributes (social economics, demographics, political tendency, education, gun ownership rate, police training hours, etc.) to reveal connections under the iceberg. Third, we built regression models based on correlation analysis for numeric variables selection to predict police shooting rates at the state level. We also built classification models based on Chi-square testing for categorical variables selection to predict the victims' race of fatal police shootings. The main datasets we choose for our analysis include: 1. Washington Post Fatal Police Shooting Dataset (WP data) [13] covers fatal police shooting from 01-01-2015 to 12-02-2020. 2. KilledByPolice (KBP): Fatal police shooting reported in KilledByPolice website [6] from 01-01-2015 to 11-04-2020.

2. Related work

Several studies have been conducted based on utilizing local crime data to explain racial disparities and differences in fatal police shootings. Mentch (2020) [8] implemented resampling procedures to take factors like local arrest demography and law enforcement density into account. He found substantially less racial disparity after accounting for local arrest demographics. On the contrary, Ross (2015) [9] built a multi-level Bayesian model to investigate the extent of racial bias in the recent shooting of civilians by police. He concluded that racial discrimination observed in police shootings is not explainable due to local-level race-specific crime rates. Noticeably, Mentch and Ross had reached contradictory conclusions. But they inspired us to use data mining and machine learning techniques to incorporate more factors rather than only crime data to understand fatal police shootings in the US better.

3. Methodology

We defined **reporting deviation rate** and **total absolute reporting deviation rate** to evaluate the media's reporting bias.

In WP dataset analysis, we used **FP-growth** and **word cloud** to reveal the frequent patterns and **DBSCAN clustering** to find fatal shooting hotspots. We also implemented **correlation analysis** to analyze correlation between multiple numeric attributes and fatal police shooting rate and tested the significance of their correlations. We used **T-test/ANOVA** to measure the significance of fatal police shooting rate by categorical attributes.

In fatal police shooting rate prediction, we used results of correlation analysis to select numeric predictors. We constructed a series of regression models, including **Kstar**, **K-Nearest-Neighbor**, **Random Forest**, and **Linear Regression**, to predict state level's fatal police shooting rate. We measured their performance by **ten-fold cross validation** scores. In victims' race prediction, we used **Chi-square testing** to do **variables selection**. We built a series of classification models, including **Gradient Boosting Machine**, **Multi-class Classifier**, **Logistic Regression**, and **Naïve Bayes Classifier**, to predict the race of fatal police shooting victims. We measured their performance by **stratified five-fold cross validation** scores.

4. Media reporting analysis

Since 2015, The Washington Post (WP) has created a database cataloging every fatal shooting nationwide by a

police officer in the line of duty. There have been less than 1,000 people killed by police every year. The killed rate of African American people is disproportionally higher than any other race (use Black or B to distinguish with Asian or A in the following). The Figure-1 show the number of people killed by police shooting by year in national wide. Figure-2 shows the average proportion rate of each race killed by police shooting from WP's website.

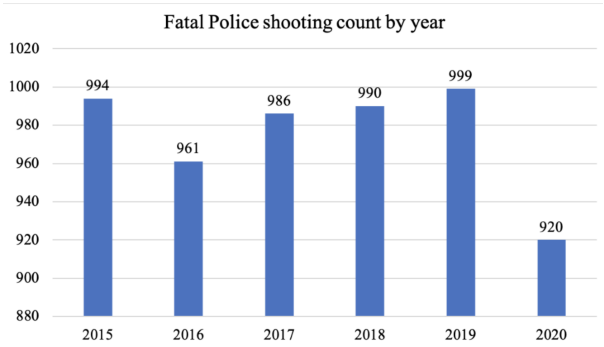

Figure 1. Number of people killed by police shooting by year till 02/12/2020

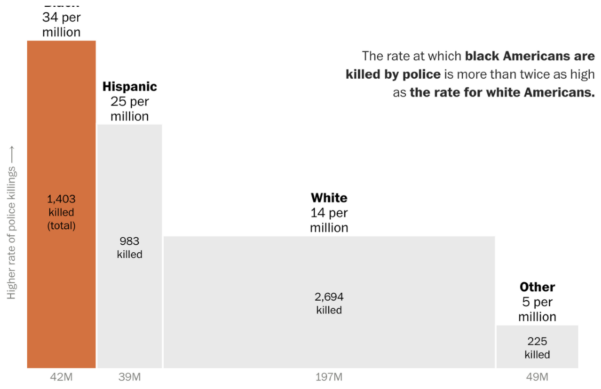

Figure 2. The average proportion rate of each race killed by police shooting, [14]

Admittedly, there is no doubt that Black people's rate is high than any other group of people if we compare it with the population proportion. However, once we add the proportion of violent incidents offenders [12]to each racial group, we see the ratios have matched each other accordingly. See Figure-3 below.

We collected 2472 police shooting victims with known reported media and race from 2016 till now from KBP. We hold our null hypothesis that media reported news by each race should follow the real happened case distribution. We use the racial proportion of victims from WP Data as the ground truth. We selected media with over 100 fatal police shooting news reporting, which include one conservative media, FOX (318) and three liberal media: ABC (244), CBS (227), and NBC (135). The media's political inclination is showed by Figure-4: Political bias of selected media. We excluded media with less than 100 reporting since most of them are local media whose news may be impacted by the local demographics. Figure-5 shows the comparison results: all four media have different deviations on reporting the truth. In general, Black victims were over-reported among all the media.

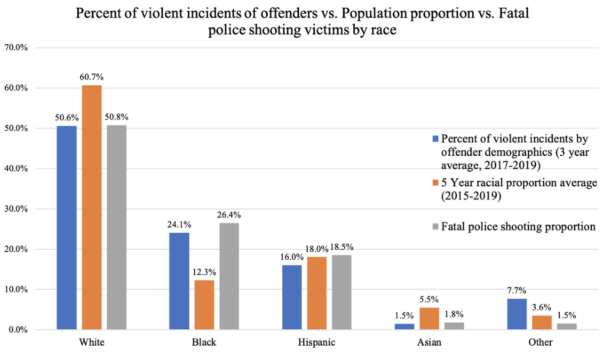

Figure 3. Percent of violent incidents of offenders (3-year average) VS. 5-year average population proportion VS. Fatal police shooting victims by race

Figure 4. Political bias of selected media

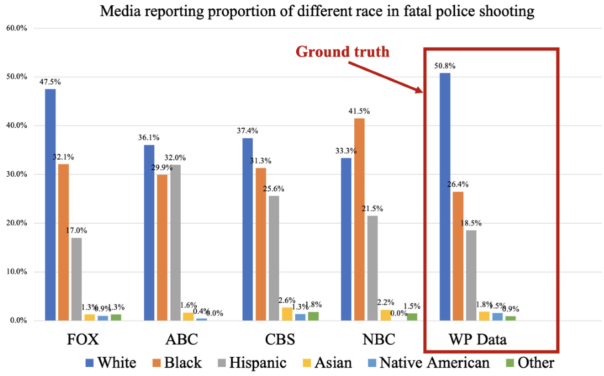

Figure 5. Media reporting proportion of police shooting by different race

To exam the difference of deviation between four media, we defined the measurements and calculation methods:

1. Reporting deviation rate of media B regarding race

A = $R(B, A)$ = reported proportion of race A by media B – real proportion of race A in WP Data. If $R(B, A) < 0$, media B underreports race A victims. Else $R(B, A) > 0$, media B overreports race A victims.

2. Total absolute reporting deviation rate of media B = $= \sum_{i=1}^{N} |R(B, A_i)|$, Ai is the i-th race and N is the number of races.

We then get Figure-6: Four media reporting deviations. FOX has the least deviation rate from the WP Data, and there are only –3.3% deviation for White, +5.6% for Black, -1.5% for Hispanic, -0.6% for Asian, and –0.6% for Native American, and +0.4% for Other. Nevertheless, ABC, CBS, and NBC have larger reporting deviations, which underreported 10% White victims while overreported Hispanic and African American victims. Specifically, NBC underreported White victims' proportion by 17.4% and overreported Black victims' proportion by 15.0%. Furthermore, it even reported more Black victims (41.5%) than whites (33.3%). ABC overreported Hispanic victims' proportion by 13.4%. It reported Hispanic victims (32.0%) at the same level as White victims (36.1%). The Figure-7 shows four media total absolute deviation rate.

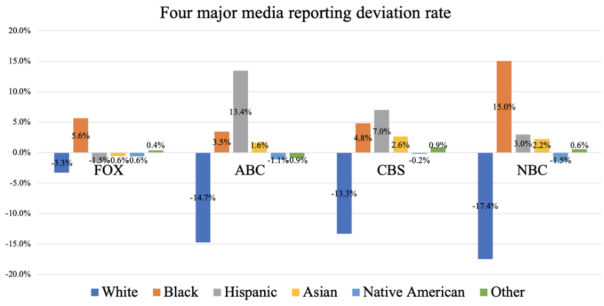

Figure 6. Four major media reporting deviation rate

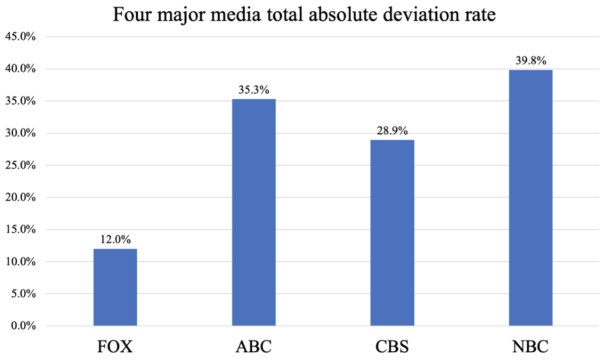

Figure 7. Four major media total absolute deviation rate

In terms of total absolute reporting proportion error, NBC has the largest reporting deviation rate (39.8%), followed by ABC (35.3%), and CBS (28.9%), while FOX has the least rate (12.0%) shown above.

5. WP fatal police shooting dataset insight

In this part, we use FP-growth and word cloud to reveal the frequent pattern behind the WP dataset. We use location data from the WP dataset to cluster police shooting incidents and find shooting hotspots. We also tried multi-attributes such as social economics, demographics, political tendency, education, gun ownership rate, police training hours, etc., to verify the possible reason for the police shooting.

5.1. Frequent Pattern Mining

From the frequent pattern mining, we can conclude a typical victim shot by police: **a "man" (96%) "without mental illness" (77%) uses "gun" (57%) "attack" (65%) police then get "shot" (95%) by police who does not wear "body camera" (88%)**. see below Figure-8 and Figure-9. **"California," "Texas," "Florida"** are the top three states were happened more frequently in total number, see Figure-10.

Therefore, our subsequent analysis considers gun ownership rate, crime rate, Marijuana legality, and governor's party by state level. The frequent pattern uses FP-growth [HPY00], and the threshold of minimum support is 50% of the total transactions of the WP dataset.

We also apply DBSCAN [5] to the longitude and latitude of fatal police shooting locations to identify hotspot clusters. Set parameters eps=0.5 and min_sample=50, we find the dense areas of fatal police shootings, see below Figure-11. We discover that Los Angles and Atlanta metropolitan areas have two of the largest hotspots. Generally, all the fatal police shooting hotspots are in the top population cities in the country.

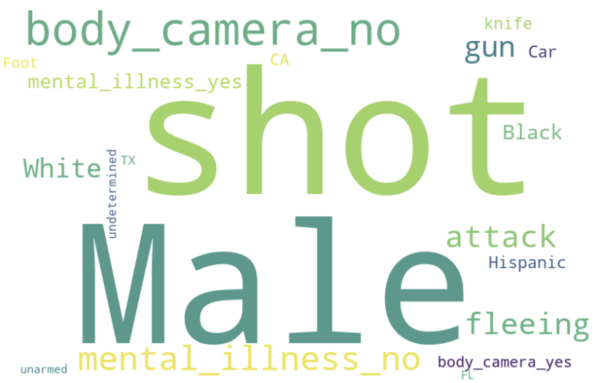

Figure 8. Word cloud of police shooting

5.2. Correlated variables analysis

5.2.1 Quantitative Variable analysis

To avoid the population distorting the analysis, we normalized the number to the yearly average fatal police shooting

Items	Frequency	Percentage
Male,	5563	96%
shot,	5525	95%
shot, Male,	5278	91%
body_camera_no,	5099	88%
body_camera_no, Male,	4875	84%
shot, body_camera_no,	4856	83%
shot, body_camera_no, Male,	4639	80%
mental_illness_no,	4469	77%
mental_illness_no, Male,	4301	74%
shot, mental_illness_no,	4274	73%
shot, mental_illness_no, Male,	4109	71%
mental_illness_no, body_camera_no,	3959	68%
mental_illness_no, body_camera_no, Male,	3804	65%
shot, mental_illness_no, body_camera_no,	3797	65%
attack,	3762	65%
shot, mental_illness_no, body_camera_no, Male,	3645	63%
shot, attack,	3614	62%
attack, Male,	3614	62%
Not fleeing,	3609	62%
shot, attack, Male,	3467	60%
Not fleeing, Male,	3431	59%
hot, Not fleeing,	3399	58%
body_camera_no, attack,	3334	57%
gun	3322	57%
gun, shot,	3279	56%
shot, Not fleeing, Male,	3228	55%
gun, Male,	3209	55%
shot, body_camera_no, attack,	3208	55%
body_camera_no, attack, Male,	3198	55%
gun, Male, shot,	3166	54%
Not fleeing, body_camera_no,	3155	54%
shot, body_camera_no, attack, Male,	3073	53%
Not fleeing, body_camera_no, Male,	3001	52%
shot, Not fleeing, body_camera_no,	2982	51%
gun, body_camera_no,	2955	51%
mental_illness_no, attack,	2923	50%
gun, body_camera_no, shot,	2920	50%

Figure 9. Frequent pattern of police shooting

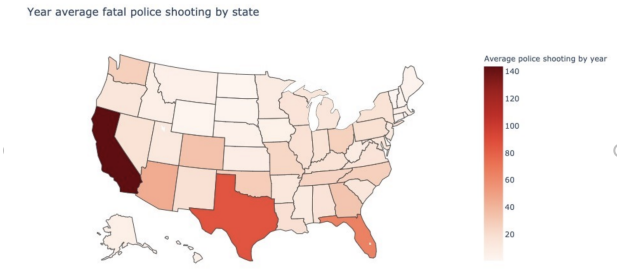

Figure 10. Yearly average Fatal Police shooting per 1m by State

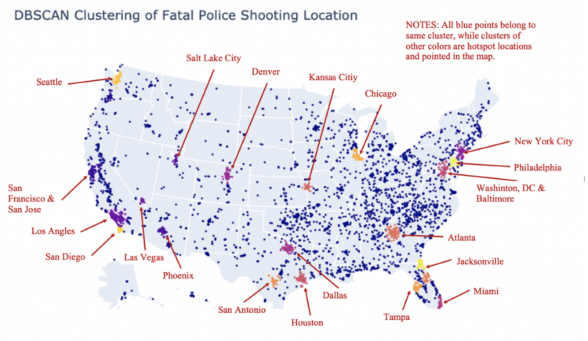

Figure 11. Fatal police shooting hotspots distribution

per one million people (**fatal police shooting rate**). We use this density-kind value for the analysis afterwards. Figure-12 shows that every year on average, how many people were shot by police. **New Mexico** and **Alaska** where have relatively less population, become the top state. The color is getting darker from east to west except for large population states such as California, Washington. Doesn't it look like the U.S. history of territory expansion?

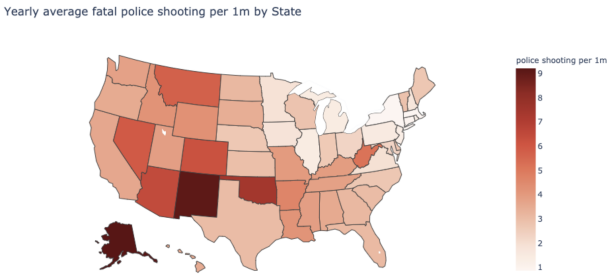

Figure 12. Yearly average Fatal Police shooting per 1m by State

It looks the longer the state joined the U.S., the lower the fatal police shooting rate in that state. The correlation coefficient is 68% between the fatal police shooting rate and the U.S. history of territory expansion. Our interpretation is: the reason that U.S. police use excessive violence may root from the westward expansion when handling the violent criminals, see Figure-13.

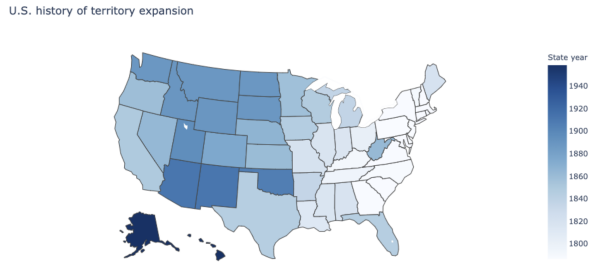

Figure 13. US history of territory expansion

The correlation coefficient is 64% between gun ownership rate [17] and the fatal police shooting rate. 57% of victims hold guns (not including other weapons), and 65% of victims chose to attack police. This hold gun rate doubles than the average gun ownership rate among the country, which is 30% according to Pew's report [7], see Figure-14.

The third high correlation variable is the land area [16], 59%, followed by violence rate [3], 48%, poverty rate 37%, unemployment rate 29%, see Figure-16. Surprisingly, police basic training hours negatively correlate with the fatal police shooting rate, see Figure-15. Although TrainingReform [10] appeals appeal to increase police training hours, the current data shows the opposite result. It may suggest reviewing and improving the training itself rather than a single slogan for more hours.

We also tested the correlation coefficient's significance to guarantee the association, which are all proved with rel-

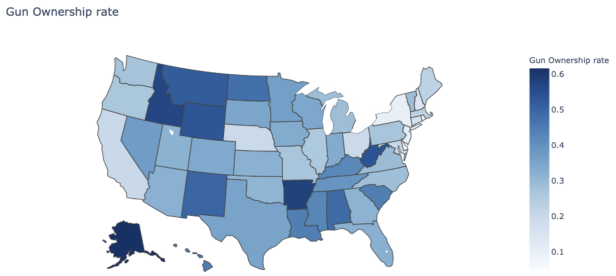

Figure 14. Gun ownership rate by state

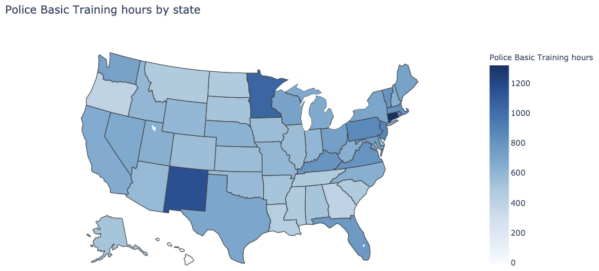

Figure 15. Police Basic Training hours by state

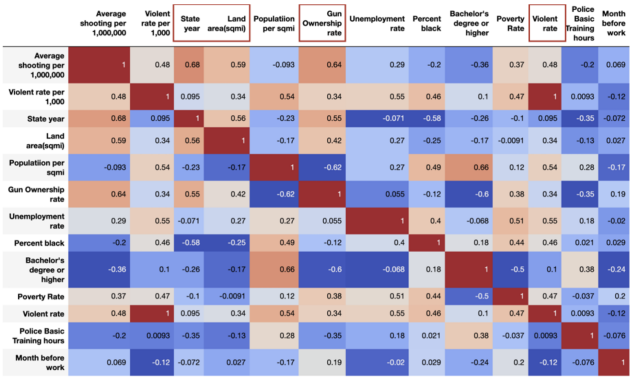

Figure 16. Correlation table

atively small p_value, see Table-1.

$$t = r\sqrt{\frac{n-2}{1-r^2}}, \alpha = 0.01$$

Variable	r	n	t - statistic	p_value
State year	0.68	51	6.4920	4.06E-08
Land area(sqmi)	0.59	51	5.1152	5.21E-06
Violent rate	0.48	51	3.8301	0.0004
Gun ownership rate	0.64	50	5.7707	5.27E-07

Table 1. Correlation coefficient test

5.2.2 Categorical variables analysis

In this part, we tested the significance of the fatal police shooting rate by state-level Governor's party [15] and Marijuana Legality [2]. We failed to reject the null hypothesis, and we can conclude that there is no difference between those states on the fatal police shooting.

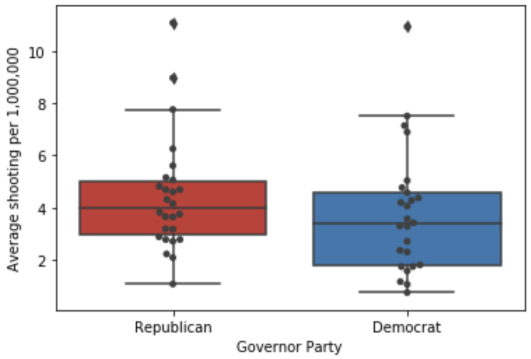

Figure 17. Boxplot of fatal police shooting in Republican and Democrat states

T-test results:

H_0: $\mu_{GOP} = \mu_{Dems}$

H_A: the average fatal police shooting rate are not equal between Republican and Democrat governor states

Test result: t = 0.9936, Pvalue = 0.3254, fail to reject the null hypothesis. The average fatal police shooting rate are equal between Republican and Democrat governor states

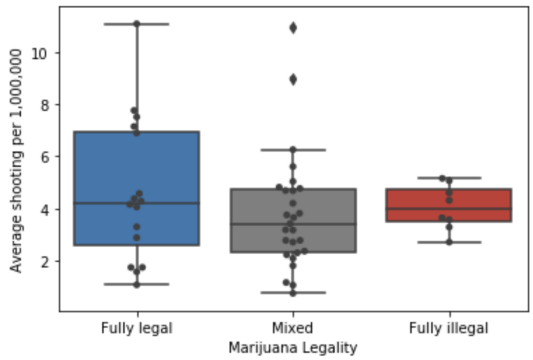

Figure 18. Boxplot of fatal police shooting among different marijuana legality states

One-way ANOVA:

H_0: $\mu_{FL} = \mu_{MML} = \mu_{FI}$

H_A: at least one of the average rates differs from one of the others

Test result: F = 0.6492, P_{value} = 0.527, fail to reject the null hypothesis. The average fatal police shooting rate are equal among different marijuana legality states

6. Fatal police shooting rate and victims race prediction

In this part, we used the insights we draw from WP data and multi-attributes correlation analysis to build predictive models. We constructed a series of regression models to predict fatal police shooting rates on the state level and a series of classification models to predict fatal police shooting victims' race.

6.1. Fatal police shooting rate prediction on state level

According to above correlation analysis, we chose **the violent crime rate**, **land area**, and **gun ownership rate**, **state_joined_year** based on their highest correlation coefficient with the fatal police shooting rate. We acquired more data points by looking at each state every year from 2015 to 2019 separately.

In the Weka machine learning software, we tried all models and chose three of the best-performed models based on ten-fold cross-validation performance. The best one is Kstar [1]. It achieved 28.04% cross-validation relative absolute error and explained 88.53% variance, followed by K-Nearest-Neighbor Regression and Random Forest. These three models all performed much better than the baseline linear regression model, see Table-2.

Algorithm	Correlation coefficient	Mean absolute error	Relative absolute error
Kstar	**0.8853**	**0.4731**	**28.04%**
K-Nearest-Neighbor	0.8517	0.5396	31.98%
Random Forest	0.8594	0.6205	36.78%
Linear Regression	0.4489	1.5238	90.33%

Table 2. Ten-fold cross validation results

Figure-19 displays the cross-validation prediction error of each data point in the Kstar model (each data point represents the fatal police shooting rate of a state in a particular year). The X-axis is the real police shooting rate, while the Y-axis is the predicted police shooting rate. The large cross means a higher error rate.

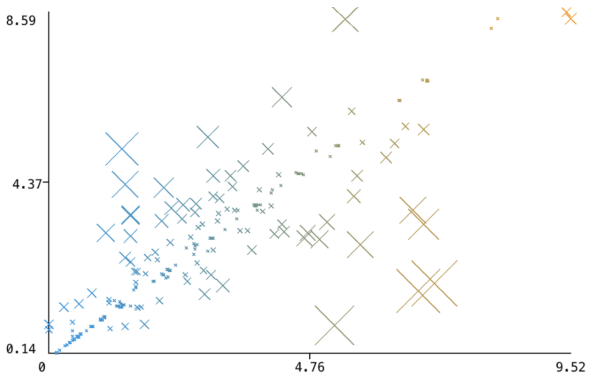

Figure 19. Predicted fatal police shooting rate vs. Real fatal police shooting rate

The prediction model tells us that the reason for fatal police shootings could be complex. It is related to the state joined year, state land area, gun ownership rate, and violent crime rate. It suggests us to understand this problem from multi-dimensional aspects.

6.2. Predict victims' race in fatal police shooting

This prediction intends to test whether or not there is racial discrimination during the fatal police shooting. The null hypothesis is that the model cannot predict the victim's race (No racial discrimination). The alternative hypothesis is that the model can predict the victim's race (racial discrimination). We use WP data from 01/01/2015 to 02/12/2020 and excluded the data missing the race information. The total records are 4518. Since "age" is the only numeric variable, we applied the chi-square test to select the predictor for the rest of the variables.

6.2.1 Chi-square testing

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \alpha = 0.05$$

where χ^2 = chi squared, O_i = observed value, E_i = expected value

	Asian	Black	Hispanic	Native American	Other	White	Total
No BodCam	63 (66.8)	993 (1044.6)	717 (719.3)	56 (59.0)	38 (36.4)	2053 (1993.8)	3920
Bodcam	14 (10.2)	211(159.4)	112 (109.7)	12 (9.0)	4 (5.6)	245 (304.2)	598
Total	77	1204	829	68	42	2298	4518

Table 3. The chi_square contingency table for body_ camera

Variables	Chi-square score	DF	P-Value
signs_of_ mental illness	109.05	5	< .00001
flee	90.00	15	< .00001
armed	41.12	5	< .00001
body_camera	35.90	5	< .00001
threat_level	25.21	5	0.0001
gender	18.85	5	0.0020
manner_of_death	8.80	5	0.1171
is_gencoding_exact	8.31	5	0.1398

Table 4. Chi-square testing for categorical variables

After applying chi-square testing to the above categorical variables, we find that threat_level, signs_of_mental_illness, armed, flee, body_camera, and gender are not independent of the race at 0.05 statistically significant level, see Table 4. On the other hand, manner_of_death and is_gencoding_exact are independent of the race at 0.05 statistically significant level. For city and state, the degree of freedoms (DF) is too large to apply chi-square testing. Finally, we chose **armed**, **age**, **gender**, **signs_of_mental_illness**, **threat_level**, **flee**, and **body_camera** as predictors and city, age as back-up predictors for the racial classification model.

6.2.2 Classification model

In the Weka machine learning software and Python AutoML package, we tried all models and chosen the top three

best-performed models based on stratified five-fold cross-validation performance. see Table-5 below.

Algorithm	Precision	Recall	F1-score
Gradient Boosting Machine	**0.589**	**0.611**	**0.600**
Multi-Class Classifier	0.591	0.598	0.594
Logistic Regression	0.588	0.597	0.592
Naïve Bayes	0.577	0.588	0.582

Table 5. Stratified cross validation results

We find that adding city and state attributes could boost model performance. Gradient Boosting Machine [4] performs best, having 0.589 precision and 0.611 recall, slightly better than predicting all victims to be white (about 50% precision and recall). GBM algorithm gives us an idea of the importance of attributes we selected for prediction. **City**, **state**, **armed**, and **age** attributes play essential roles in racial prediction. See Figure-20 below. We failed to reject the null hypothesis since even the best-performed model cannot predict victims' race well, proving that there is no racial discrimination for observed fatal police shootings in WP data.

7. Conclusion

In conclusion, we found that mainstream media disproportional reporting fatal police shooting by the race, which may instigate hostile sentiments between police and the public. We suggest mainstream media report all news according to the realistic. Second, we found that the police shooting rate depends on many variables. The top four significant attributes were **state joined year**, **state land area**, **gun ownership rate**, and violent crime rate. Choosing these four attributes as predictors, our best-performed regression model could predict the fatal police shooting rate with about 88.53% correlation coefficient. Admittedly, we cannot find all the influence factors. It indicates that the fatal police shooting is a **complex multi-dimensional** problem. We also found two variables (police basic training hour, months police can work before basic training) appealed by CNBC negatively and weakly correlated with the fatal police shooting. Third, based on the WP dataset, we tried to depict a typical scenario when a police shooting happened and remark the hotspots among the country. Last, our three best performance models show no significant evidence to conclude that racial discrimination happened during fatal police shootings recorded by the WP dataset.

References

[1] John G Cleary and Leonard E Trigg. K*: An instance-based learner using an entropic distance measure. In *Machine Learning Proceedings 1995*, pages 108–114. Elsevier, 1995.

[2] DISA Global Solutions. Map of Marijuana Legality by State. https://disa.com/map-of-marijuana-legality-by-state.

[3] FBI. Crime in the United States, by Region, Geographic Division, and State. https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/topic-pages/tables/table-4.

[4] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[5] Kamran Khan, Saif Ur Rehman, Kamran Aziz, Simon Fong, and Sababady Sarasvady. Dbscan: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*, pages 232–238. IEEE, 2014.

[6] KilledByPolice. Police Shootings Database – Killed By Police. https://killedbypolice.net/.

[7] Kim Parker, Juliana Menasce Horowitz, Ruth Igielnik, J. Baxter Olophant and Anna Brown. The demographics of gun ownership. https://www.pewsocialtrends.org/2017/06/22/the-demographics-of-gun-ownership/.

[8] Lucas Mentch. On racial disparities in recent fatal police shootings. *Statistics and Public Policy*, 7(1):9–18, 2020.

[9] Cody T Ross. A multi-level bayesian analysis of racial bias in police shootings at the county-level in the united states, 2011–2014. *PloS one*, 10(11):e0141854, 2015.

[10] The Institute for Criminal Justice Training Reform. State Law Enforcement Training Requirements. https://www.trainingreform.org/.

[11] Thomas Jefferson. United States Declaration of Independence. https://en.wikipedia.org/wiki/United_States_Declaration_of_Independence.

[12] U.S. Department of Justice. Criminal Victimization. https://www.bjs.gov/content/pub/pdf/cv19.pdf.

[13] Washington Post. Fatal Police Shooting Dataset. https://github.com/washingtonpost/data-police-shootings.

[14] Washington Post. Police Shootin Database. https://www.washingtonpost.com/graphics/investigations/police-shootings-database/.

[15] Wikipedia. List of current United States governors. https://en.wikipedia.org/wiki/List_of_current_United_States_governors.

[16] Wikipedia. List of U.S. states and territories by area. https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_area.

[17] World Population Review. Gun Ownership by State 2020. https://worldpopulationreview.com/state-rankings/gun-ownership-by-state.