Course Project

United States Police Shootings (2015-2020)

Team 9: Yash Pise, Salvatore Scire, Vidisha Bardekar, Aishwarya Tikare, Dishti Soni

Stevens Institute of Technology

BIA 600-A

Dr. Bei Yan

December 2021

Contents

I. Introduction

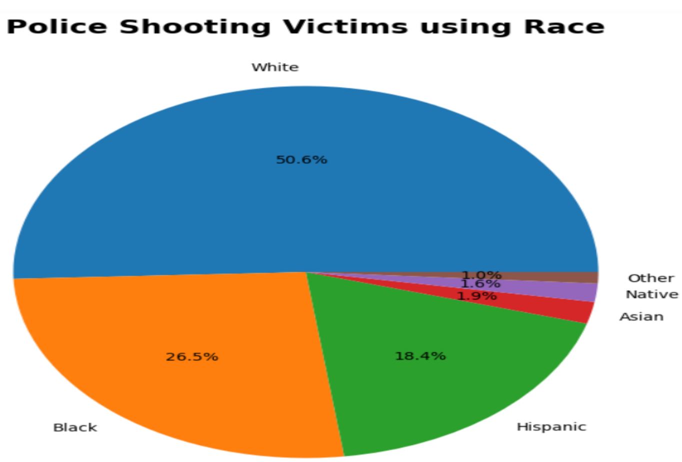
In recent years the number of police shootings in the United States have been relatively stable yet the media has put in the forefront, police shootings based on race. The main purpose of doing this analysis was to primarily discover if race followed by being an armed or unarmed and political party in power was a factor in the variation of police shootings in the United States from 2015 to 2020. The secondary purpose of this analysis was to find out facts on gender, age, weapon type, or mental illness of the victims. We also gathered data on the top 25 cities with the most shootings to see which political party was in power at the time. We then had to find data on total population to run ANOVA and Post-Hoc regressions based on race. Our goal was to simply present statistical findings on a controversial topic and if any important information was discovered, make a suggestion to improve the cause.

II. Data Set Explanation & Mean / Standard Deviation

We downloaded the raw data from Kaggle. The dataset has 4851 rows and 15 columns. It contains basic data about people like their name, age, gender, race, city, state, and armed vs unarmed. Within the shooting/killing dataset, we were able to analyze police shooting victims using race and gender distribution. We were also able to analyze which cities faced the highest rate of shooting, how were they shot, which weapons they were holding, and if they had any mental illness? Additionally, we had to find data from the United States Census on total population by race for the years 2015-2020 with all data being ratio data. This contained only a total 36 data points, 6 total years and 6 total races. The data set for 2020 was not completed from March on, we were unaware of the reason as COVID-19 hit at that time therefore we made our audience aware when presenting total population shootings for 2020. Our main data set was divided into categorical and ratio attributes. Name, gender, city, state and armed vs unarmed are categorical. Age is ratio data. The mean and standard deviation of our data set of people shot by race across 2015-2020 is as follows: Mean: Asian 15, Black 209, Hispanic 146, Native 12, Other 7, and White 397. The standard deviation is: Asian 4.604, Black 66.71, Hispanic 50.73, Native 6.802, Other 4.792, and White 222. In order to break our data down into numerical values, we had one of our team members write code in application python to transfer the categorical data to numerical. We also have a third data set which contains all the majors of the top 25 cities with the most shootings per year breaking the political party in power down by democratic, republican, or independent. This data is nominal.

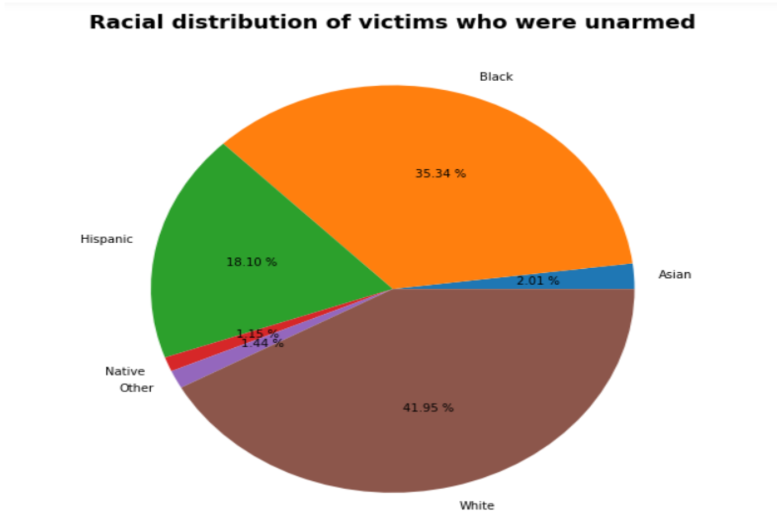
III. Visualizations of Variables

1. Race



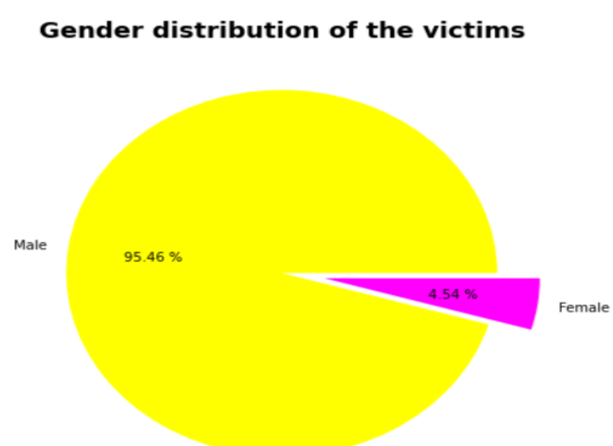
The first way we analyzed our data was breaking the total shootings down by race. As we can see that 50% of the people shot are white and 26% of people are black followed by 18% of people are hispanic. While white people make up half the population of victims shot, there is some disproportions in the percentage of total populations by race.

2. **Racial Distribution of Victims (Unarmed)**



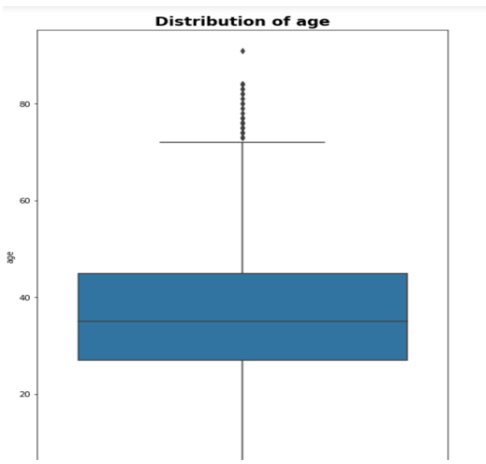
The second visualization we have is victims who are armed vs unarmed. We can see from the chart that 42% of unarmed victims were white, followed by 35% black and 18% hispanic. Considering the population of the United States, the count of unarmed black and hispanic is higher. Out of 100 victims, 6 were unarmed victims for a total 318 unarmed individuals.

3. **Gender Distribution**



Our third visualization shows that 95% of the victims are male and 5% of the victims are female. Males have a higher statistical probability of being in violent crimes than women.

4. Age



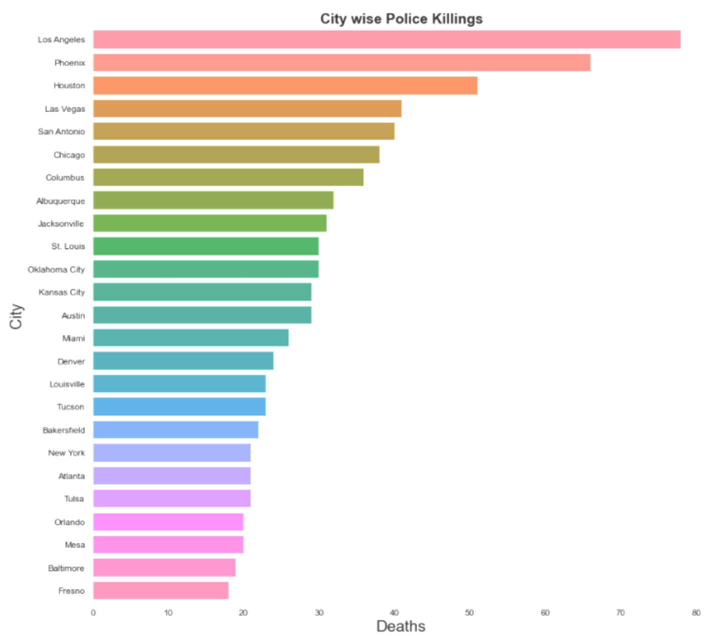
	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_mental_illness	threat_level	flee	body_camera	arms_
811	980	Jeremy Mardis	2015-11-03	shot	unarmed	6.0	M	White	Marksville	LA	False	other	Car	True	
2761	3229	Kameron Prescott	2017-12-21	shot	unarmed	6.0	M	White	Schertz	TX	False	other	Not fleeing	False	

Our fourth visualization for both men and women, and across all racial and ethnic groupings, the risk of being killed by police peaks between the ages of 25 and 45. The minimum age shot was 6 years old; unarmed, which is very unacceptable.

5. Number of Police Shootings (2015-2020)

Our fifth visualization shows the US has an average of 900 shootings per year. The chart is decreasing from 2015-2019, which is a good sign. Data from the second half of 2020 is incomplete.

6. Citywise Police Killings



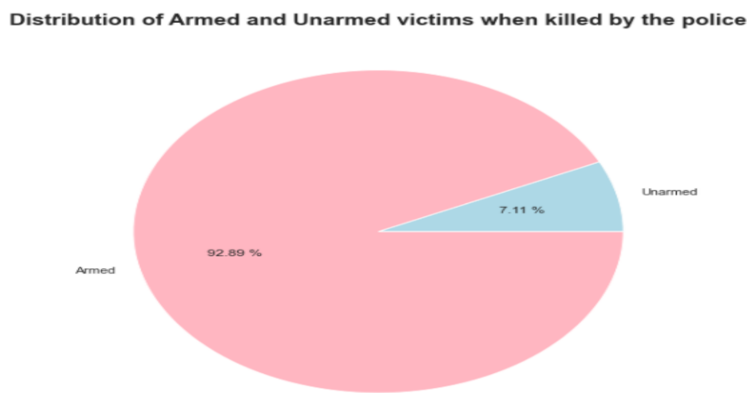
Democratic	63.00%
Republic	31%
Independent	6%

Our next visualization shows that police shootings occurred more frequently in cities where populations are concentrated. From the analysis, cities with the highest rates of shootings are LA, Phoenix and Houston. We could also predict that 63% of the cities had Democratic rule, 31% of the cities had Republic and just 6% of the states were Independent.

7. Most Commonly Used Weapons

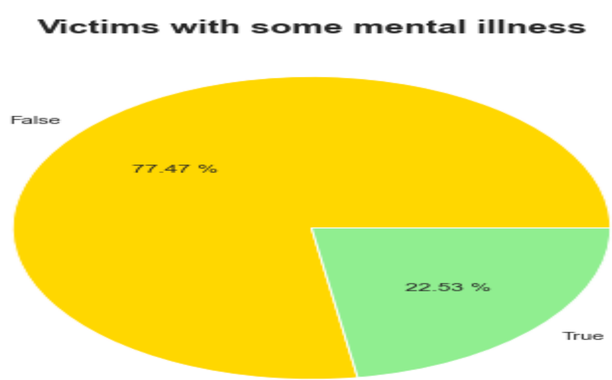
Next we have the most common weapons used by victims in the shootings. Guns are the most commonly used weapon in shootings followed by knives, then unknown, unmamed, vehicle, machete and taser.

8. Distribution of Armed and Unarmed Victims



Next we have a visiulation on armed victims vs unarmed victims. Instances reveal that there were more objectively reasonable police killings, with 93 percent of victims being armed compared to only 7% being unarmed. We cannot ascertain the exact causes for killings based on the limited data we have, but we can deduce that armed victims are more likely to be killed by police than unarmed victims based on the limited data we have.

9. Mental Illness



Our last visualization shows 77% of victims were not suffering from mental illness, whereas 23% were. We can deduce that 23 of the 100 victims suffered from mental illness. Individuals with major mental illness are 11 times more likely than the general public to be victims of violent crime, according to the US Statistics, and women with serious mental illness are more vulnerable than men. Perhaps some of these individuals are undiagnosed.

III. ANOVA + PostHoc Regressions

Before running our regressions, we had studied the data for a while believing we were missing a crucial part of the analysis. Unable to run simple regressions due to our data set, we had to either run Chi-Square test or ANOVA regressions. Upon running both, our answers to the tests were obvious and provided no clear extrapolations. After hours of analysis, it finally clicked that we needed to have the total population by year of each race and then use the total shootings by race to divide the data by total population by race per year. Within our ANOVA regressions, the data entries we used were actually percentages.

The first ANOVA Single- Factor regression we ran included all six races with the shooting percentages over total population. Our Null Hypothesis was are all the averages of total shootings per race the same. Our Alternative Hypothesis was were the averages different. We used a 95% confidence interval. The regression gave us a P value of .00359. Compared to our P value of .05. We concluded that the averages were in fact different proving there is no bias. But, the average of black people shot was a higher number than the others, so we figured that there could be some kind of bias within the regressions when we run the Post Hoc ANOVA regressions. See picture below.

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
A	6	4.7184E-06	7.864E-07	5.67611E-14
B	6	2.90078E-05	4.83463E-06	2.44571E-12
H	6	1.47163E-05	2.45272E-06	7.41555E-13
N	6	1.81477E-05	3.02462E-06	2.69735E-12
O	6	1.82715E-05	3.04524E-06	3.5696E-12
W	6	1.23104E-05	2.05173E-06	4.81132E-13

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	5.3547E-11	5	1.07094E-11	6.430713543	0.000359203	2.53355455
Within Groups	4.99605E-11	30	1.66535E-12			
Total	1.03507E-10	35				

The second ANOVA regression we ran or the first Post-Hoc ANOVA regression assuming unequal variances was between the White and the Black populations. Our Null

Hypothesis was are all the averages the same and our Alternative Hypothesis was the averages are different. With the 95% confidence interval, we had to divide our P Value of .05 by 6 due to the six factors. The P value came to .0083. From the regression, we got a Two Tail P Value of .0053 which is just below the P Value of .0083. Here we will also reject the Null Hypothesis proving that the averages are in fact different. See picture below.

t-Test: Two-Sample Assuming Unequal Variances		
	W	*B*
Mean	2.05173E-06	4.83463E-06
Variance	4.81132E-13	2.44571E-12
Observations	6	6
Hypothesized Mean	0	
df	7	
t Stat	-3.984509203	
P(T<=t) one-tail	0.002646681	
t Critical one-tail	3.130418397	
P(T<=t) two-tail	0.005293362	
t Critical two-tail	3.638827341	

The third ANOVA regression or the second Post-Hoc ANOVA regression assuming unequal variances was between the Black and Hispanic populations. Once again, our Null Hypothesis was are all the averages the same where our Alternative Hypothesis was the averages are different. Our P value of .05 was divided by 6 to get .0083. When we ran the regression, our Two Tail P Value came to be .011 which is greater than the .0083. This is the first case in which we have to fail to reject the null proving that the averages are in fact the same which will prove there is bias here. See picture below.

t-Test: Two-Sample Assuming Unequal Variances		
	B	*H*
Mean	4.8346E-06	2.4527E-06
Variance	2.4457E-12	7.4156E-13
Observations	6	6
Hypothesized M	0	
df	8	
t Stat	3.26807835	
P(T<=t) one-tai	0.00569444	
t Critical one-ta	3.01839452	
P(T<=t) two-tai	0.01138888	
t Critical two-ta	3.48160898	

The fourth ANOVA regression of the third Post-Hoc ANOVA regression assuming unequal variances was comparing the Hispanic and the White populations. Our Null Hypothesis was all the averages are the same where our Alternative Hypothesis was the averages are different. With our P Value of .0083, we ran the regression and received a Two-Tailed P Value of .395. This is a very high P value and we will fail to reject the null. It is safe to assume that there is significant bias here and perhaps more analysis should be run. See picture below.

t-Test: Two-Sample Assuming Unequal Variances		
	H	*W*
Mean	2.4527E-06	2.0517E-06
Variance	7.4156E-13	4.8113E-13
Observations	6	6
Hypothesized M	0	
df	10	
t Stat	0.88829279	
P(T<=t) one-tai	0.19762367	
t Critical one-ta	2.87241124	
P(T<=t) two-tai	0.39524734	
t Critical two-ta	3.27921165	

V. Conclusion

We can draw various conclusions from our findings. The first is Whites make up over 50% of the victims shot in police shootings, followed by Blacks and then Hispanics. When it comes to unarmed victims, Whites make up 42% of the population followed by Blacks at a close 35% then hispanics at 18%. Our gender distribution is 95% male with 5% female given due to males having a higher crime commitment rate. The majority of our population is from ages 25-45. Police shootings peaked in 2015 with around 950 and have decreased to around 850 by 2019. The top 25 cities with the most police shootings have a 63% Democratic government with a 31% Republican government followed by 6% Independent government. There was no surprise to find out that guns were the primary weapon that victims were holding, followed by the knife. Victims who were shot and armed were a total of 93% of the population with unarmed victims being 7%. Victims who were shot had 23% mental illness and 77% of victims did not have mental illness.

The conclusions to be drawn from our ANOVA regressions are the model as a whole, we reject the null meaning the averages are not the same. This case is true when running the Post-Hoc between Whites and Black as well enough though it is very close. When looking at the Post-Hoc for Blacks and Hispanics, the Two-Tail P Value we uncovered is higher making us fail to reject the null or proving the averages to be the same. The biggest bias came in the last Post-Hoc analysis between Whites and Hispanics with a Two-Tail P Value of .395 compared to .0083 in which we would fail to reject the Null Hypothesis proving the averages to be the same. All in all, there was no bias on the regression as whole but when looking at the individual Post-Hoc analysis the number between Black and Whites was very close, and there was bias between Black and Hispanics as well as strong bias between Whites and Hispanics. Implementations to combat these issues should take place within the top cities with the most shootings where the governments create educational programs for officers.