Bellevue University

Fall 2020

Police Shooting in the United States

October 1, 2020

Kevin Angotti

Jennifer Ruiz

Hedyeh Erfani

Executive Summary

Police Shootings are a hot button topic in the United States as their frequency increases year over year. By examining data provided by the FBI, the project team will provide insights into factors contributing to these shootings that may aid in reducing their numbers. An overview of the methods and findings of this project are summarized below.

Methods

The project team utilized a variety of predictive analytics methods to understand relationships within the data and identify factors that create an increased risk of an assailant being shot during an interaction with the police. Key factors examined during the project were race, age, signs of mental illness, whether the suspect was armed, and the gender of the individual.

Findings

The project yielded several interesting insights worth noting. More information regarding the findings can be found down below.

Abstract

Police violence is a pressing issue within the United States. Within the last decade, there has been an increased focus on rates of fatal police shootings involving minorities. Understanding the factors contributing to fatal shootings could provide insight into addressing the issue.

In this project, the authors aim to create a model that can be used for future prediction of fatal police shootings. Using machine learning algorithms and predictive methods to determine locations that may project higher death rates by demographic. Data visualization methods will be applied to examine how race, age, and location data impact shootings.

Keywords: Predictive Analysis, Police Shootings, K-Modes Algorithm

Introduction

Since 2015, there have been more than 5,000 fatal shootings by on-duty police officers (Washington Post), with Black Americans being killed at nearly twice the rate of White Americans. As communities come together in solidarity and awareness of this matter is being raised, the question of which factors are causing these shootings needs to be answered. The FBI has gathered data regarding fatal shootings from 2015 onwards with key variables: name, location, sex, manner of death, and whether persons involved were armed. Given this data, the project team will aim to determine what factors could be causing these fatal shootings.

This project will attempt to create a model that can be used for future prediction of fatal police shootings. The proposed model could also help bring needed attention to locations that may project higher death rates by demographic. We will accompany the cluster analysis model along with other visualizations to help bring focus on this issue.

Background Information

Methods

Data Understanding

When approaching the data for this project, the team began by examining the variables within the dataset. This took time due to the number of categorical variables with complex levels to sift through. To start the process, we used a combination of the R programming language, the Python programming language, and Power BI. In R, we conducted summary statistics of the datasets to see initial trends and what each variable might have to tell us. Followed by a linear regression

model to see what these trends might mean. Within Power BI, we used the combination of our summary statistics and linear model to conduct a series of visuals for the most significant variables. Doing this at the beginning of the project allows us to see what the data looks like before running any type of model on the data. Initial visualizations provided evidence that race may be an ideal target factor. Other areas of interest were body camera usage, gender of the subject, and type of weapon possessed by the subject. The team then used Python to create all of the data models that would be used for analysis and prediction. The team started with a K-Modes clustering model, followed by two linear regression prediction models and finally a decision tree prediction model. These models provided insight into the issue and allowed for future testing and a more in-depth look into why major cities might contain higher shooting towards a single race or why rural areas have higher numbers than others.

Data Preparation

Extensive data preparation was performed on the dataset to transform key categorical predictors into a numeric form. This was done to prepare the data for modeling and further analysis. While creating the new dataset, a numeric value was assigned to each of the levels within the variable to reduce complexity.

Table 1. Dataset Variable Key

Variables	
Sex	male = 0 female = 1
Manner of Death	shot = 0

	shot and tased = 1
Signs of Mental Illness	true = 0 false = 1
Race	Asian = 0 White = 1 Black = 2 Hispanic = 3 Other = 4 Native = 5
Body Camera	true = 0 false = 1
Armed	gun = 0 unarmed = 1 toy = 2 knife = 3 shovel = 4 unknown = 5 nail = 6 box cutter = 7 machete = 8 sword = 9 hammer = 10 metal object = 11

	screwdriver = 12 lawn mower blade = 13 other = 14 ax = 15 vehicle = 16 taser = 17 rock = 18 wood = 19 pipe = 20

Beginning with the armed variable, where the value could be either gun, knife, pipe, unarmed, unknown, etc., values were replaced in the range of 0 to 20. The same was conducted for each of the variables which had similar values reported. There were five variables in total, which were assigned numeric values in this same manner of data cleaning. These variables were 'sex,' 'race,' 'signs of mental illness,' 'body camera,' and 'armed.'

The next process we conducted was to create each of these variables' visuals to get a better look at the data. As you will see in Appendix A, Figures 1 through 4 give a good sense of what the data provides for the number of police shootings in our country and the persons targeted and the manner of their death.

Modeling

The extensive data preparation allowed for a linear regression model to be performed. The linear model conducted was for the variable race as a function of the armed, body camera, and gender variables that returned a good amount of significance with a p-value for each variable all under the 0.05 threshold. The overall P-value of the model yielded 0.0003941 with an F-statistic of 6.088.

To further aid in prediction, the team also utilized a clustering algorithm to understand the variables of the highest importance. The team chose to deploy a k-modes clustering model to better deal with the complex categorical variables contained within the dataset. This model allows for matching clusters based on the number of its matching categories between each data point. This type of clustering model works well with data that has a high amount of categorical data mixed in with some numeric data.

Results

Preliminary results are promising. The primary result has been that this is a learning experience for the team implementing predictive analytics processes. While further analysis is needed to draw definitive conclusions, the results of initial models are shown below.

Linear Model
`lm(formula = race ~ age + armed + body_camera + gender + signs_of_mental_illness,`
`data = shootings_1)`
`Residuals:`
`Min 1Q Median 3Q Max`

-2.1759 -0.6897 -0.2189 0.4117 3.7784
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.166032 0.058867 36.795 < 2e-16 ***
age -0.014805 0.001023 -14.472 < 2e-16 ***
armed 0.006062 0.003002 2.019 0.0435 *
body_camera -0.103938 0.040062 -2.594 0.0095 **
gender -0.120832 0.062137 -1.945 0.0519 .
signs_of_mental_illness 0.213803 0.031106 6.873 7.06e-12 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.9026 on 4889 degrees of freedom
Multiple R-squared: 0.05754, Adjusted R-squared: 0.05658
F-statistic: 59.7 on 5 and 4889 DF, p-value: < 2.2e-16

Table 1. Linear Regression Model

After running the linear regression model in R, which used different summary statistics to determine significance, such as an F-Statistic, P-Value, and R-Squared, these measures only provided some of the information needed to determine if the data could predict on a model. The team also created a few visuals in Power BI (Appendix A) to get a peek at what a model might conclude before we started building our models. The first model was created from a form of K-Means clustering model called K-Modes. This particular model handles categorical data very well. A K-Modes model takes in the entire dataset and returns its prediction on the most common

outcome for each cluster. These results each time the model ran returned 5 clusters; 3 of the 5 were White males ranging 31 to 41 years of age and armed with a weapon; the other two clusters usually produce a combination of Black and Hispanic males also armed.

When building our next model(s), the first linear regression model used race as the target variable and utilized a 60/40 test train split. The result shown in figure 1 returns predictions for each of the race types. Unfortunately, the model was only able to provide an accuracy of 0.05 to 0.11 percent. This indicates that the model was not a good fit for this data. There can be many reasons for this. The linear relationships are not appropriate, or the model itself is not a good fit. Other regression models, like logistic regression, may have been a better fit.

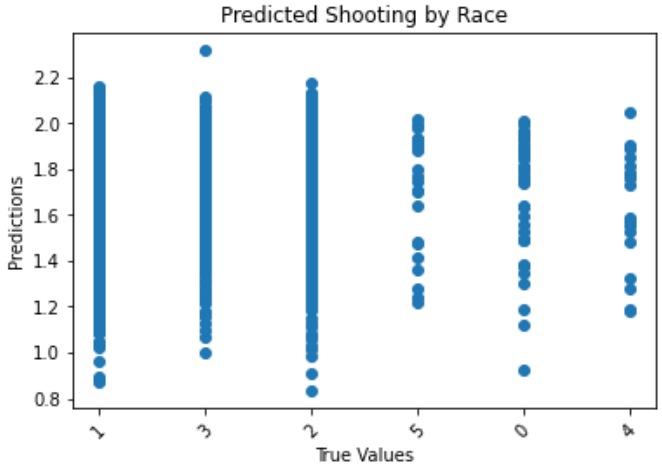


Figure 1: Model 1 predictions

The second linear model also took the race variable as the target variable; however, it looked at the mean squared error (MSE) instead of accuracy. On the dataset as a whole, the predictions when plotted look similar (see figure 2) and return an MSE of 0.81, which is quite close to zero.

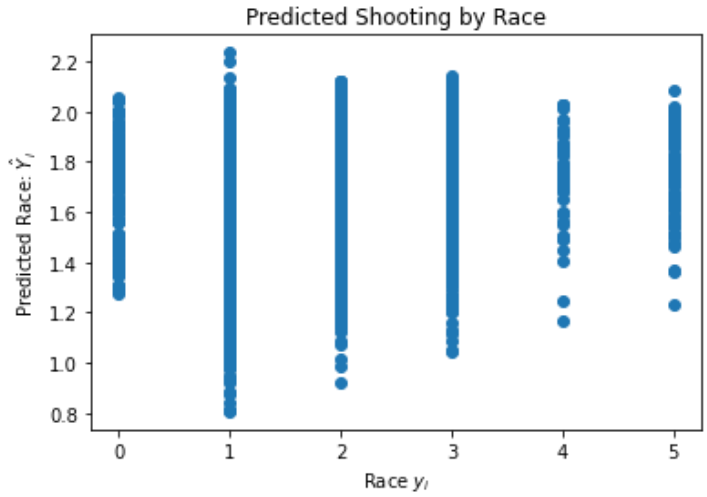


Figure 2: Model 2 Predictions

This model also ran against two variables in a separate run race against armed type. The training and testing data was split 60/40 with a random state of 5 and returned an MSE of 0.86, also close to zero. The residuals were plotted in figure 3; the data indicates that the error only seems to be contained to the White and Black race data.

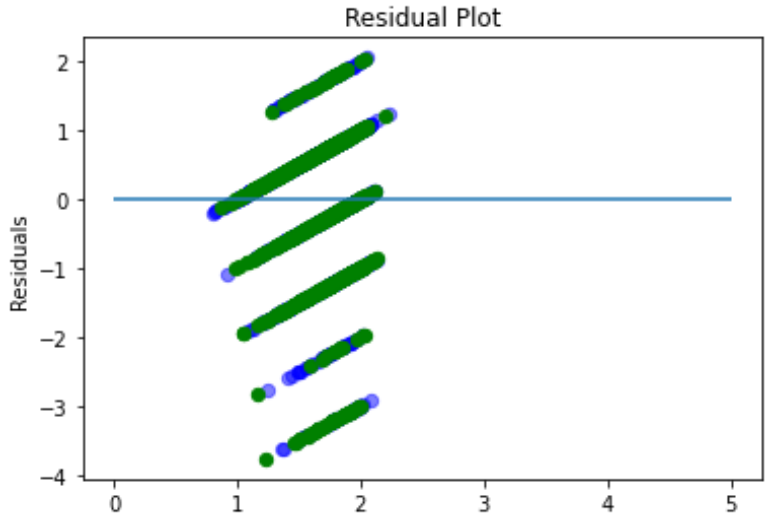


Figure 3: Residuals

The last model we created was a decision tree. This model provides a good prediction of which race will fall victim to a police shooting given specific variables. You can see in figure 4 (below), a person under the age of 42 (top node) and also under the age of 25 (2nd node left) with signs of mental illness (3rd node left) will predict that shooting victims will be as follows, 3 Asian, 82 White, 30 Black, 28 Hispanic, 2 Other, and 1 Native American.

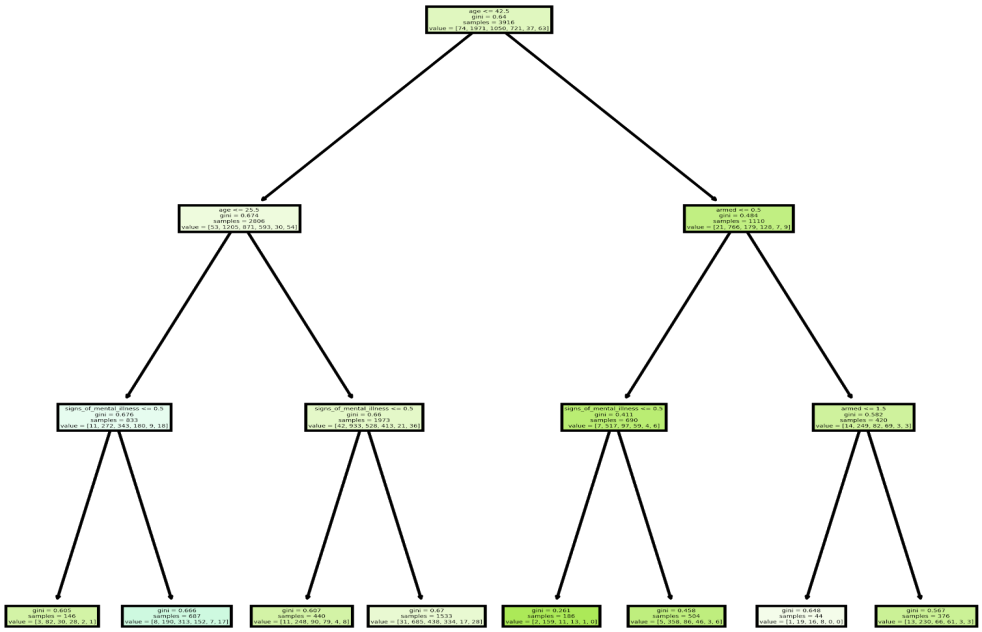


Figure 4: Decision Tree

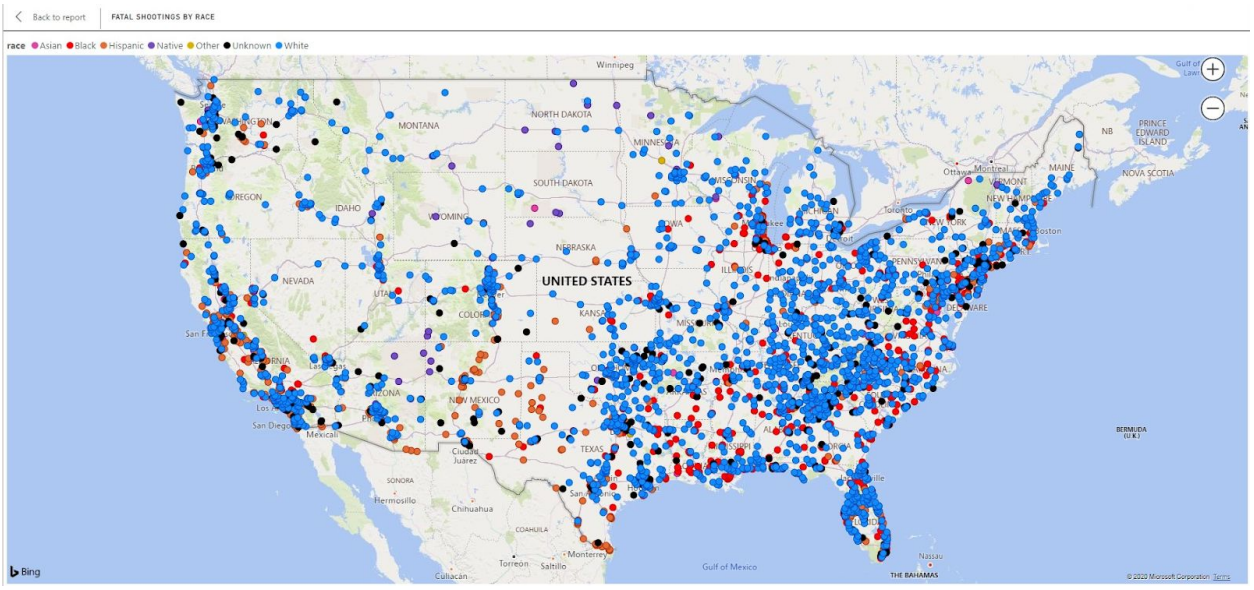
Discussion

Police violence in the United States is an issue of great significance in the country today. Understanding the factors influencing these acts is crucial to predicting their occurrence. This project uncovered trends in age, mental illness, and race that will be useful in future research and

education. The results thus far indicate a strong need for training and education in interacting with those who have mental illness and stricter body camera usage among policy entities. While the preliminary results provide some insight into these events, further investigation is necessary.

Conclusion

Fatal police shootings continue to make prime time news around the country. The predictive methods applied to create the model above begin to shed some light on the factors influencing these events. Predicting fatal police encounters is key for training and prevention. Police departments can take tools like the ones created from our team and use it as a training tool to help understand not only demographics in their area, but also how likely a certain race group might be targeted. If you have an area of high shootings towards one race training can be focused on that department/area. Further work is needed to examine some of these factors and a need to create strategies that will aid in these endeavors. Our project did notice a few factors that can help create or start to create the necessary changes that state and local authorities can focus on. Our team was able to map out from the current data where police shootings have taken place. As you can see in the map below that most police shootings occur in major cities for Blacks and Hispanics while mostly in rural areas for Whites.


Back to report
FATAL SHOOTINGS BY RACE
race Asian Black Hispanic Native Other Unknown White
UNITED STATES
Bing

Acknowledgments

The project team would like to thank Dr. Brett Werner for his continued guidance and support throughout the project. They would also like to thank the FBI for making the project data publicly available.

References

1. Kaggle. (2020). Data police shootings. Retrieved on August 31, 2020, from https://www.kaggle.com/mrmorj/data-police-shootings
2. Washington Post. (2020) Fatal Force. Retrieved on September 1, 2020, from https://www.washingtonpost.com/graphics/investigations/police-shootings-database/
3. Galarnyk, M. (2019). Understanding Decision Trees for Classification (Python). Towards Data Science. Medium. Retrieved from https://towardsdatascience.com/understanding-decision-trees-for-classification-python-9663d683c952
4. Jeevan, M. (2018). How to run Linear regression in Python scikit-Learn. Big Data. Retrieved from https://bigdata-madesimple.com/how-to-run-linear-regression-in-python-scikit-learn/

Appendix A

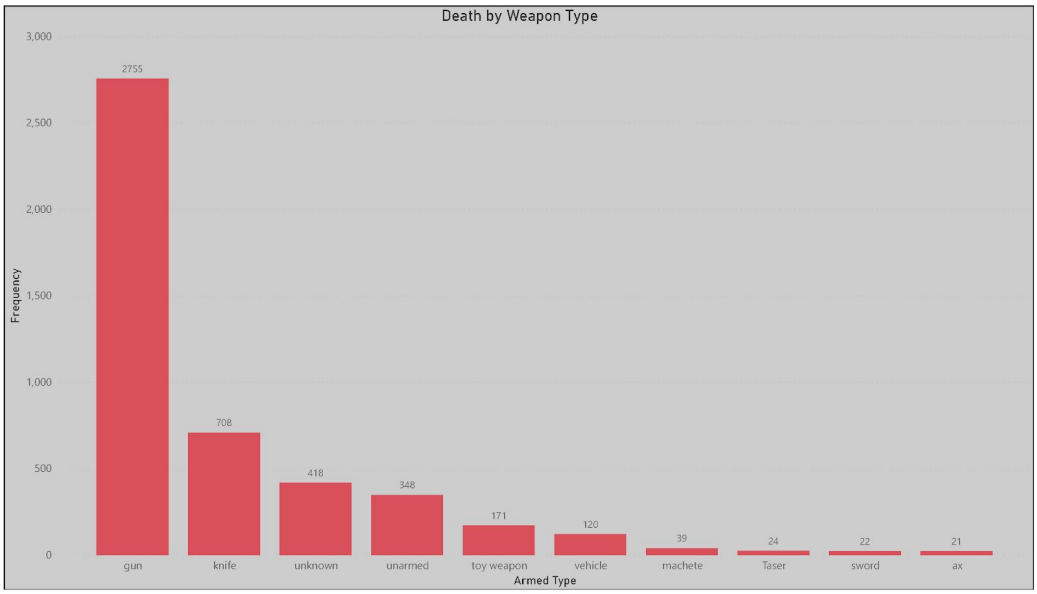


Figure 5: Death by Weapon Type

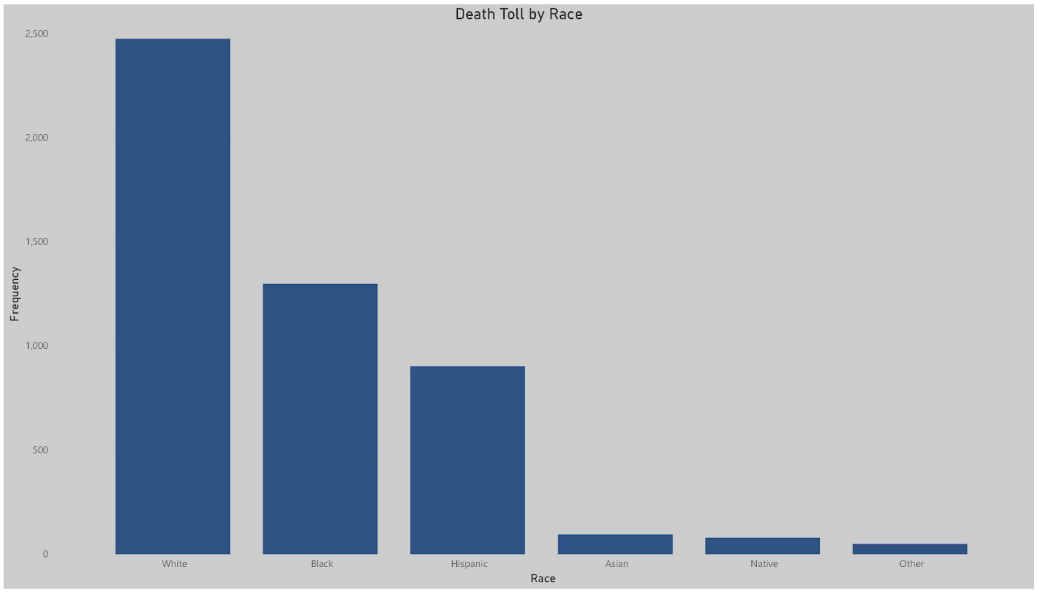


Figure 6: Death by Manner of Race

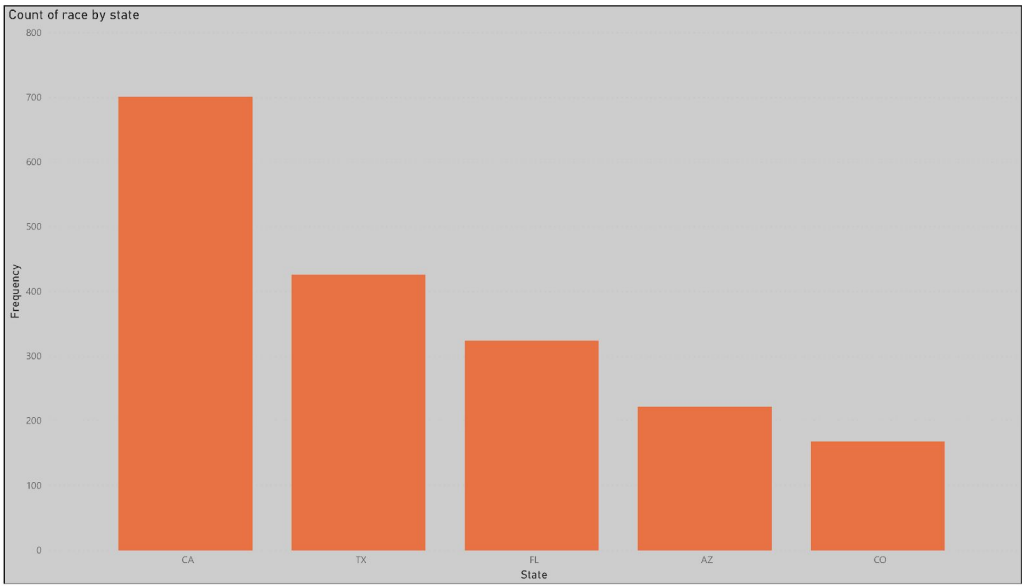


Figure 7: Death by Location

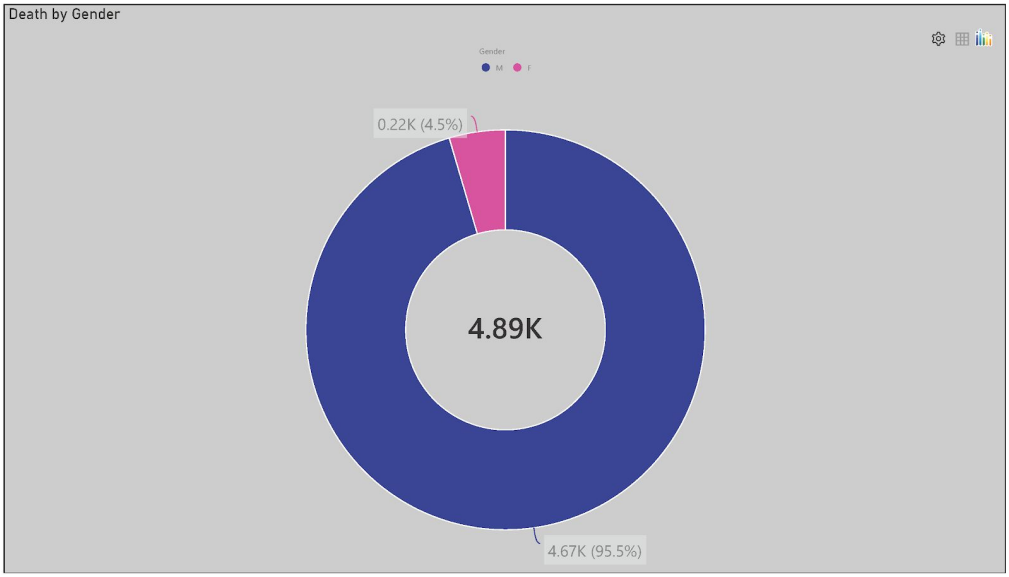


Figure 8: Death by Gender

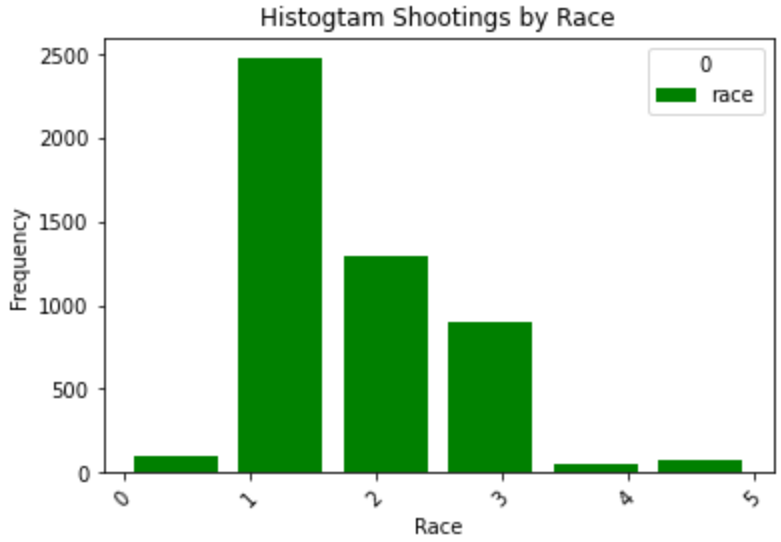


Figure 9: Shooting by Race

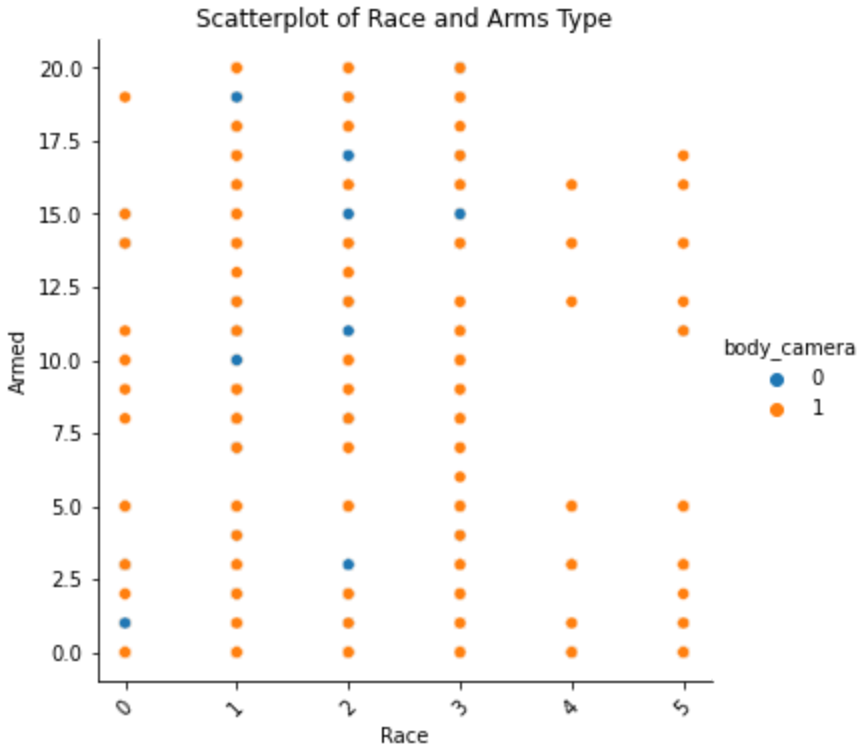


Figure 10: Body Camera Present

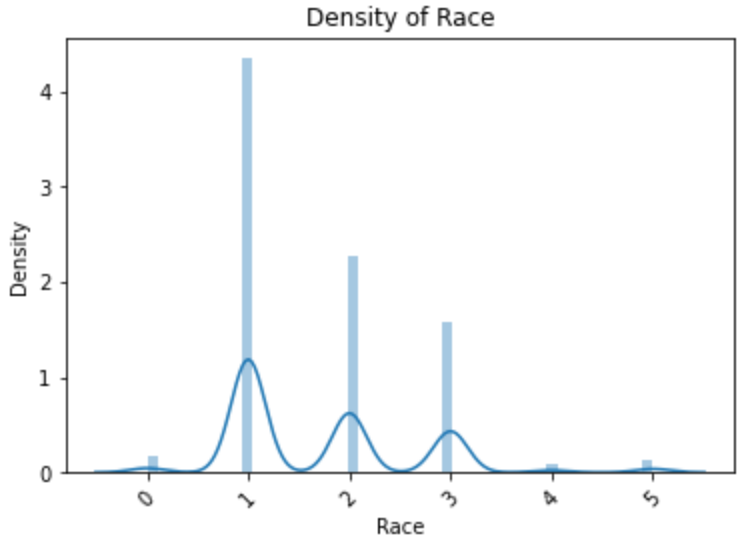


Figure 11: Density of Race

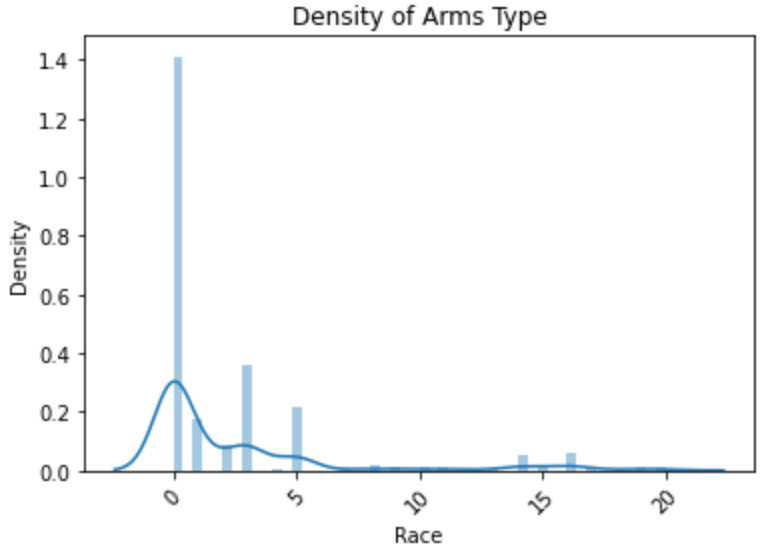


Figure 12: Density of Arms Type