MGSC661 Final Project: Investigation of Gun Violence in the USA

Tyler Nagano

December 15th, 2022

1 Introduction

One of the most controversial issues facing the United States of America today is the Second Amendment, which is the right for all people to keep and bear arms. This issue has received increased media attention due to a rise in the number of gun-related deaths and some of the most deadly mass shootings in history.

Individuals against the right to bear arms advocate for more comprehensive restrictions because they believe that increased regulation, such as background checks and bans on assault weapons, will lead to fewer deaths involving guns. In addition, the pro-regulation side claims that rarely are guns used in self-defense, and the ease of accessibility fuels criminal activity. Advocates for the right to bear arms typically agree that some regulations such as not allowing convicted felons to own a gun are good, but more regulations would infringe on the right to self-defense. The proponents of reduced regulations state that increased regulations give too much power to government institutions and instead better education on gun safety and use need to be implemented.

Gun violence in the US is the leading cause of death among children and young adults over other causes such as car accidents and drug use. The US also has the highest number of guns per capita at 135 civilian guns per 100 people. Canada has the second highest number of guns per capita but has higher restrictions which some believe is the reason for significantly less gun violence.

This project will investigate publicly available data on gun violence incidents in the United States. The goal of the project will be to identify characteristics of cities in the United States and their differences in the number of gun violence occurrences.

2 Data Description

2.1 Overview of Dataset

The dataset was collected from the gun violence archive which is a non-profit organization which systematically verifies the accuracy of gun-related violence in the US. The dataset contains about 240k incidents from all 50 states and Washington DC. There are 29 columns consisting of location information (state, city/county, address, congressional district, latitude, longitude, location description, state house district, state senate district), the sources of information (incident URL, source URL, participant sources), gun characteristics (gun type and gun stolen), participant information (participant age, name, gender, status, type, age group), incident notes, and incident characteristics such as if police officers were involved.

The number of people killed, number of people injured and number of guns involved are also in the dataset.

2.2 Data Pre-Processing

Many of the columns in the dataset do not provide any meaningful information for statistical analysis. The incident ID, all sources of information, and incident notes were removed. Of the remaining 24 columns, the location information with missing values or blank entries were removed (congressional district, latitude, longitude, state house district, state senate district, address, location description). For the gun characteristics, approximately 41.5% of the entries had empty values so these columns were removed as well. Finally, for the participant characteristics 51% of the names and 93.42% of the relationships have empty values. In addition, 38.51% of the participant ages are empty so these were removed because there is another age group column which can provide similar information. Finally, 41.49% of the number of guns involved entries were missing, but the empty values were replaced by the median number of guns of other entries. The median value is 1 which is a reasonable assumption that each act of gun violence has at least 1 gun.

After removing columns with a significant number of empty or missing values, 5 columns (incident characteristics, participant age group, status, gender and type) still contained empty or missing values. All rows with at least 1 empty value from the 5 columns were removed. All of the data processing steps resulted in about 192k records to remain in the dataset.

2.3 Feature Engineering

The incident characteristics has 109 categories of characteristics of each incident. Some categories which are highly correlated (>0.8) we're merged together if the categories had similar meanings. This removed three columns resulting in 106 characteristics. This column was converted into dummy variables because many incidents have more than 1 category. For the participant age group, status, gender and type columns contain information on both victims and suspects. Each of the columns was split into dummy variables separately for victims and suspects by using the participant type. This created six categories of dummy variables (victim status, victim age group, victim gender, suspect status, suspect age group, suspect gender). An incident was given a one in the column if at least one victim or suspect in the incident had the characteristic. In addition, the number of victims and the number of suspects were calculated from the participant type column and the number of unharmed participants were calculated from the participant status column.

The city column was used to separate cities by which ones had less than 1 gun violence act a week which would be roughly 300 gun violence acts because the data spans 6 years. The year was extracted from the date column resulting in a factor with six categories (2013-2018)

2.4 Distribution of Selected Variables

2.4.1 Continuous Variables

For the number of people killed, 97.2% of the incidents had zero or one people killed. Four or more people were killed in 207 of the incidents. For the number of people injured, 93.3% of the incidents had zero or one people injured. Only 1014 of the incidents had four or more people killed. For the number of guns involved, 94.5% of the incidents involved only one gun. Only 2379 of the incidents involved more than four guns. For the number of unharmed individuals, 80.9% of the incidents involved zero or one gun. Only 5381 of the incidents had more than 5381 people who were unharmed. For the number of victims, 87.9% of the incidents had zero or one people who were victims. Only 2065 of the incidents had four or more victims. All of the continuous variables are highly positively skewed which in this means that the majority of the incidents fall between zero and four with a small number of incidents having a value greater than four.

Correlations between continuous variables were calculated. The majority of the columns are not highly related to each other. The number of people injured and the number of victims are positively correlated which means that as the number of people injured increase so does the number of victims. In addition, the number of suspects and the number of unharmed individuals also increase together too.

Bar plots of binned categories (0, 1, 2-3, >=4) for all continuous variables and a correlation matrix of categorical variables can be seen in the Appendix.

2.4.2 Categorical Variables

For the incident characteristics, 65 of the characteristics are present in less than 1% of all the incidents in the dataset. Only 4 of the characteristics are present in greater than 10% of the incidents. These characteristics are shot wounded/injured), shot dead (murder, accidental, suicide), possession of guns found during the commission of other crimes and non-shooting incident. In addition, 20 of the categories had greater than 5000 incidents with this category.

For the suspects age group, 38.2% of incidents were committed by adults older than eighteen, 4.4% of incidents were committed by teenagers between age twelve and seventeen and 0.58% of the incidents were committed by children under the age of twelve. For the

victims age group, 12.4% of incidents had victims older than eighteen, 1.92% of incidents had victims between age twelve and seventeen and 0.78% of victims were younger than 12.

For the suspects gender, 60.4% of incidents had a male as the suspect whereas only 5.32% of incidents had a female as the suspect. For the victims gender, 62.9% of incidents had a male as the victim whereas only 14.2% of the victims were female.

For the suspects status, 16.6% of the incidents had an injured suspect, 11.6% of incidents had a suspect who was killed, 35.1% of incidents had a suspect who was unharmed and 18.6% of incidents had a suspect that was arrested. For the victims status, 9.4% of the incidents had an injured victim, 5.2% of incidents had a victim who was killed and 6.8% of the incidents had victims who we're unharmed.

Bar charts of all categorical variables can be seen in the Appendix.

3 Model Selection and Methodology

3.1 Determining Response Variable

My goal for this project was to find characteristics that distinguish different geographical regions in the US. In the exploratory phase, different classifications of regions were explored such as the state, census bureau regions and divisions as well as cities. The distribution of gun violence events we're unequal across the different geographical regions due to many factors not considered in the dataset. To make the response variable balanced, any city with greater than one act of gun violence per week was considered to be high and any city with less than one gun violence was considered to be low.

3.2 Collinearity and Principal Component Analysis

To test for collinearity in the data set, a correlation matrix was created on all variables and principal component analysis was performed on continuous variables. All continuous variables with >0.8 correlation we're removed from the rest of the analysis. Highly correlated categorical variables (have the same category prediction) were combined into a single column if the meaning of the category is similar. Otherwise the columns were removed. Principal component analysis attempts to maximize the amount of variance explained while creating new columns that are completely uncorrelated. Highly related variables determined from principal component analysis will be removed from the dataset

3.3 Random Forest Model Selection

The random forest algorithm was selected to model the problem because it is highly resilient to biases due to multicollinearity and creates accurate predictions by training many models on the data which reduce the variance in predictions. This model is also a good choice for my dataset due to the high dimensionality (123 predictors). The random forest model uses bootstrap aggregation and trains many decision trees. A decision tree model classifies the data based on all the predictors. A random forest model does takes bootstraps of the data which means that it takes a sample of the data by iteratively taking one observation at random then putting the observation back in the data. This is repeated for the number of observations of the dataset. For each of the bootstrapped samples a decision tree using the square root of the number of predictors is trained. This process is repeated a number of times and the average prediction of all trained trees are used to predict the outcome.

First, to determine the best complexity control parameter to use for the Random Forest model a Decision Tree model. The data was split into a training set and testing set with 90% of the observation used in training and 10% in testing. After determining the optimal complexity control parameter using the decision tree, the number of predictors to be used for each tree in the random forest model was tuned. The optimal complexity control parameter and number of predictors to use on each split were used in training the random forest model on the testing set. The model was evaluated using the Out-of-Bag error and predictions were made using the testing dataset.

4 Results

4.1 Correlation Matrix and Principal Component Analysis

The correlation matrix and principal component analysis identified that the number of unharmed victims and number of suspects we're highly correlated. The number of suspects was removed from the analysis. In addition, the correlation matrix identified that gun at elementary/primary school and school incident were highly related so these were combined into a single column called school. The accidental shooting injury, accidential negligent discharge and accidental shooting were also highly related so they were combined into a single column called accident.

4.2 Final Model Parameters and Predictions

4.2.1 Model Parameters, Out-of-Bag Error and Feature Importances

The final random forest model uses 123 predictors to predict the response variable of cities with high or low number of gun violence incidents. The best hyperparameters for the random forest model was to use a complexity parameter of 0.000351 and 11 predictors for each decision tree built. In addition, the total number of trees built was set to be 100. The Out-of-Bag error on my training data set was 34.07%. Using the confusion matrix, the model had an error rate of 37.27% on cities with a high number of gun violence incidents and the model had an error rate of 30.93% on cities with a low number of gun violence incidents. Out of the variables used in the model, the top five predictors which when removed from the model decrease the accuracy are male suspects, arrested suspects, the category shot wounded/injured, the number of victims and the number of unharmed suspects. For the high gun violence areas, the category of if the officer was involved in the shooting replaces arrested suspects in the top five of mean decrease in accuracy if removed from the model. For the low gun violence areas, the number of male suspects is replaced by the number of injured individuals in the top five of mean decrease in accuracy. For the Mean Decrease in Gini Index, the most important predictors which separate the two classes in order are the number of victims, if the suspects are male, the number of unharmed individuals, the number of injured individuals and if the suspects are arrested or not.

4.2.2 Testing Data Predictions

On the testing data set the overall accuracy of predictions was 65.76%. The sensitivity which is the%age of cities that are high gun violence cities are predicted to be high was found to be 62.59%. The specificity which is the%age of cities that are low gun violence cities are predicted to be low was found to be 68.88%. The positive predictive value which is the percentage of cities that when predicted to be high violence are actually high gun violence cities is 66.34%. The negative predictive value which is the percentage of cities predicted to be low violence cities are actually low gun violence cities is 65.26%. The base rate accuracy of the model is 50.51% so this model does actually have some predictive power. In addition, the confusion matrix metrics listed before demonstrate that this is a balanced model and is not biased to predict high or low gun violence due to overfitting on certain characteristics.

5 Business Insights and Conclusions

Overall, this report provides government institutions some insights into the differences between cities with more than one gun violence incident per week and cities that have less than one gun violence incident. Based on the results of the model, the variables that differentiated between high and low gun incident cities are the number of victims, the number unharmed, and the number of injured individuals. This makes sense because cities with a high number of gun violence incidents likely need to be more densely populated than cities with a low number of gun violence incidents. One interesting category that helped to differentiate high number of gun violence cities from low number of gun violence cities is if the officer was involved in the incident. The high number of gun violence cities may contain more individuals and therefore more police officers leading to more chances that a police officer will get into an altercation with gun violence. Another category that had one of the largest differences between high and low gun violence cities is domestic violence. This category made low gun violence cities less accurate, but high gun violence cities more accurate. This may be the case based on combining the fact that police officers are more likely to be involved in high gun violence cities. In low gun violence cities, there may be a smaller police force and less people report cases of domestic violence in general.

The limitation of the model is that the number of gun violence incidents in cities is likely highly related to the population size of the city itself. Many of the predictors that are contributing heavily to the accuracy of the results such as the number of people killed will likely only occur if the population density of the city is high. Similarly, it is unlikely that a mass gun violence event will occur in a city with low population density because there are fewer locations that many people will all be in the same location.

In conclusion, gun violence in the US is a growing issue due to many factors that may not be captured in this dataset. This investigation helps to highlight some of the differences in high and low gun violence cities such as increased police involvement in high gun violence areas and demonstrates the complexity of the issue due to the likely high correlation with number of gun violence incidents and the population of a city. Despite the limitations of the study, increased regulations on guns, with more limitations on Second Amendment rights, will likely help to reduce the problem of gun violence by helping cities with larger populations and increased number of gun violence incidents to more effectively manage this growing issue.

6 Appendix

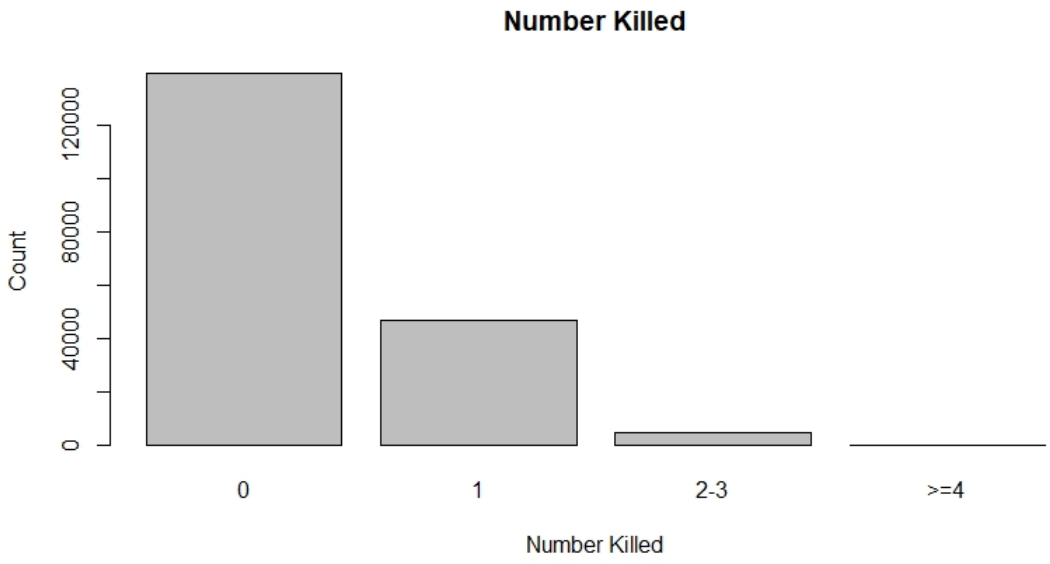

Figure 1: Counts of Number of People Killed in Incidents

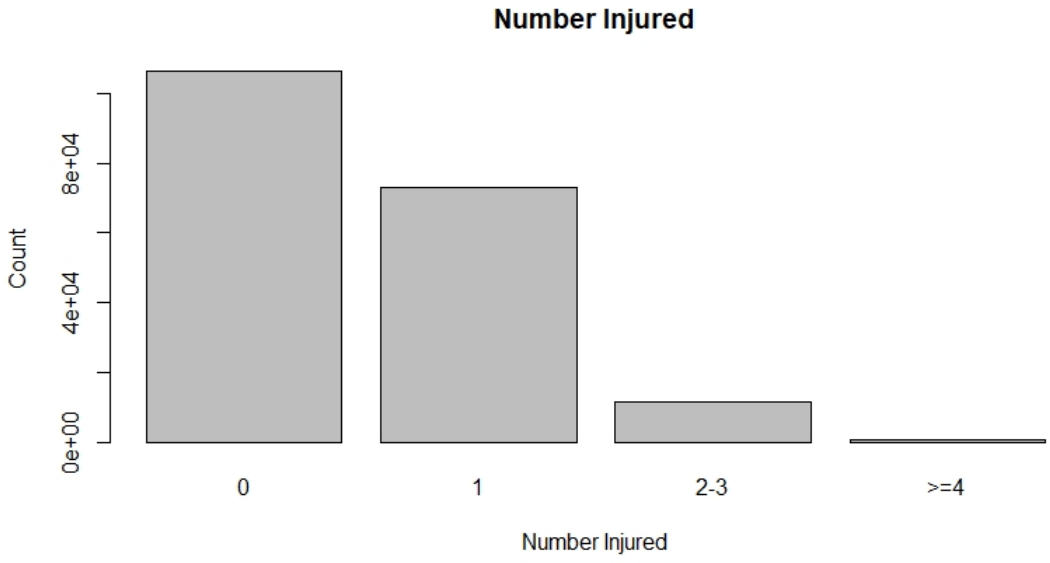

Figure 2: Counts of Number of People Injured in Incidents

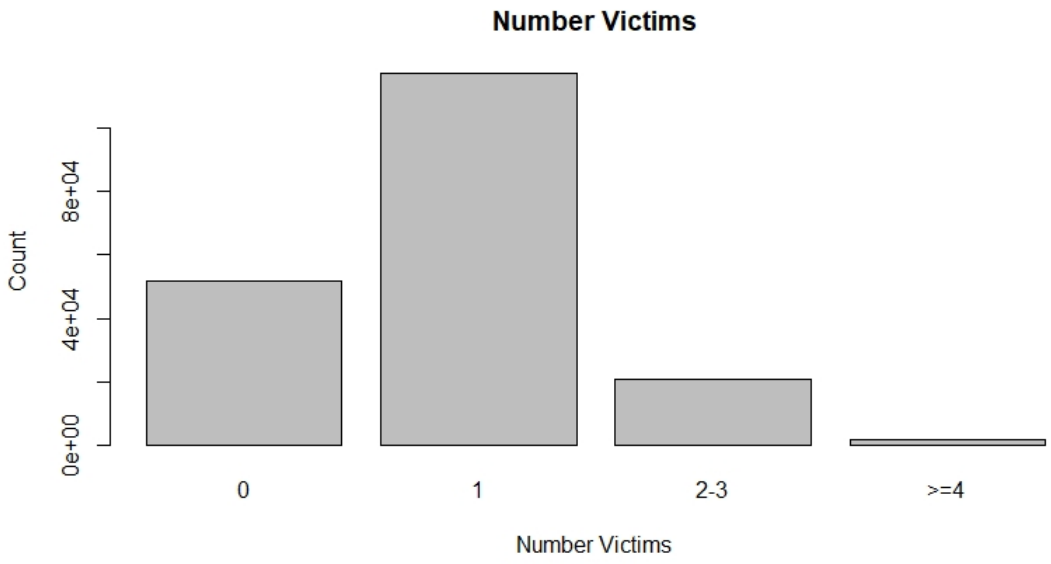

Figure 3: Counts of Number of Victims in Incidents

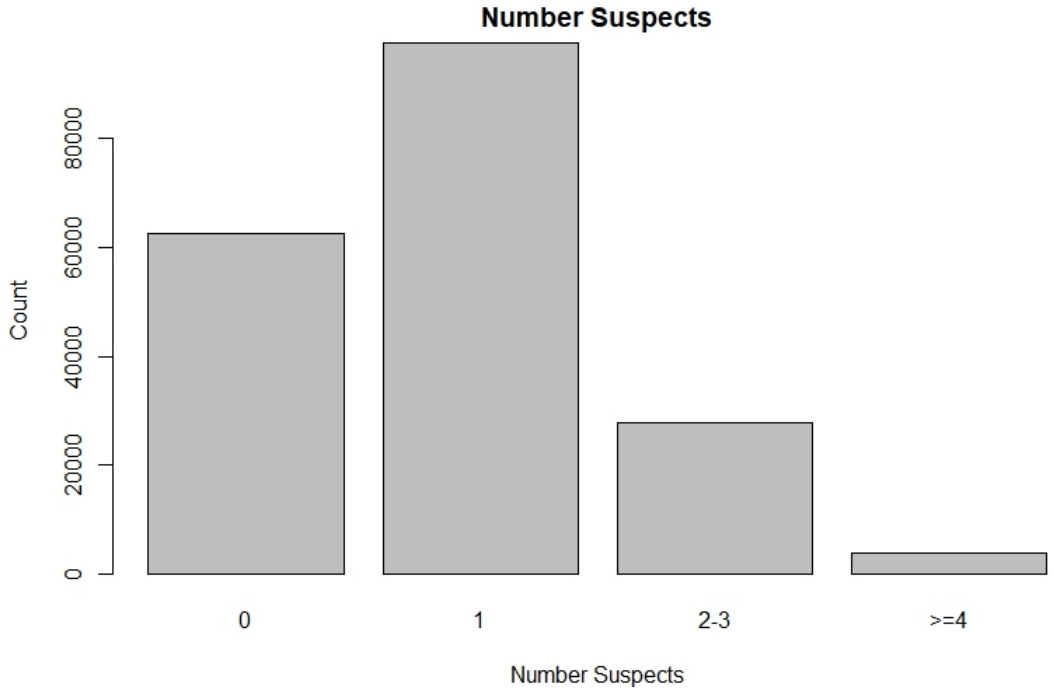

Figure 4: Counts of Number of Suspects in Incidents

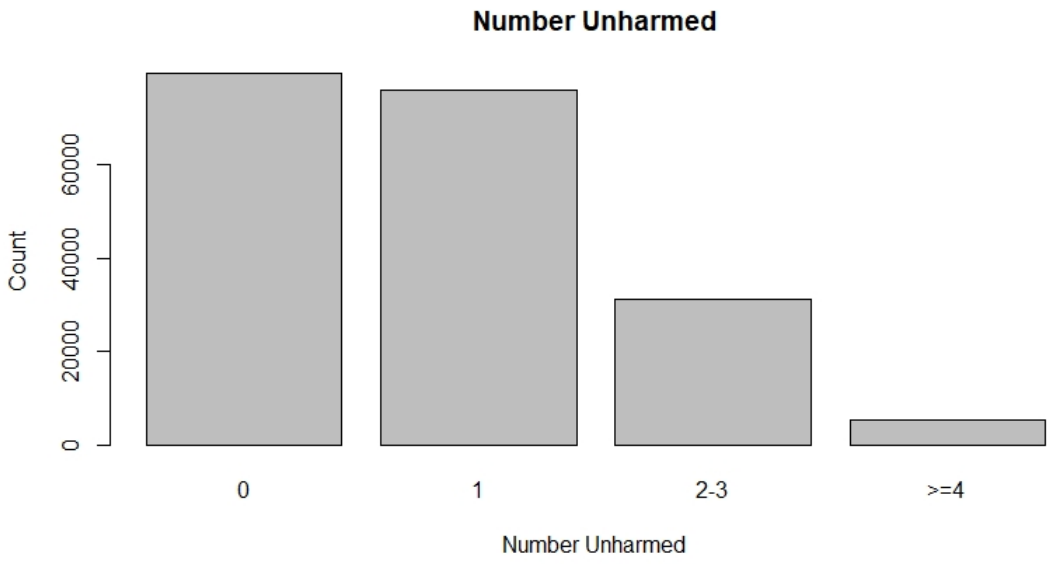

Figure 5: Counts of Number of People Unharmed in Incidents

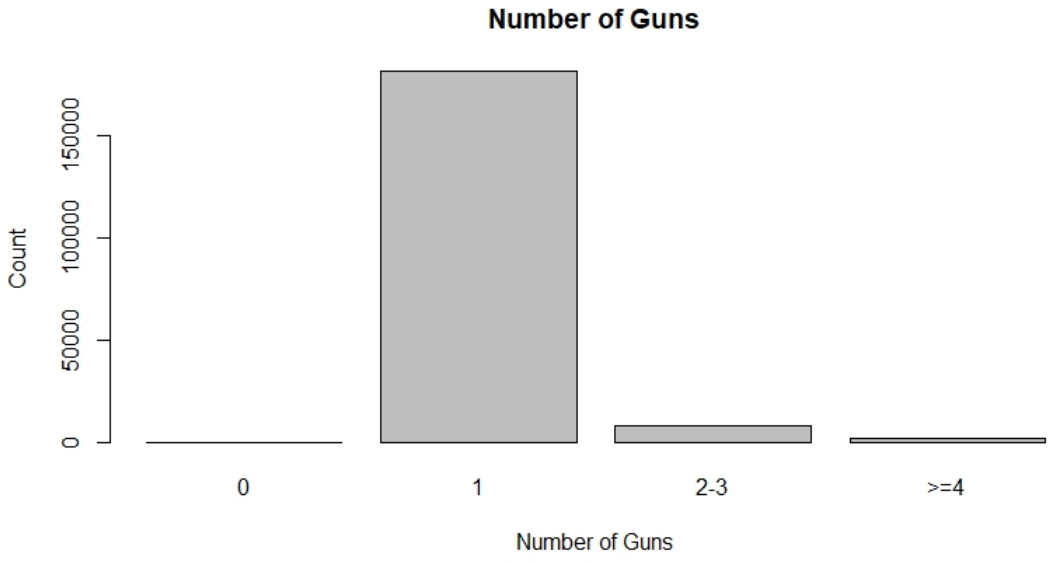

Figure 6: Counts of Number of Guns in Incidents

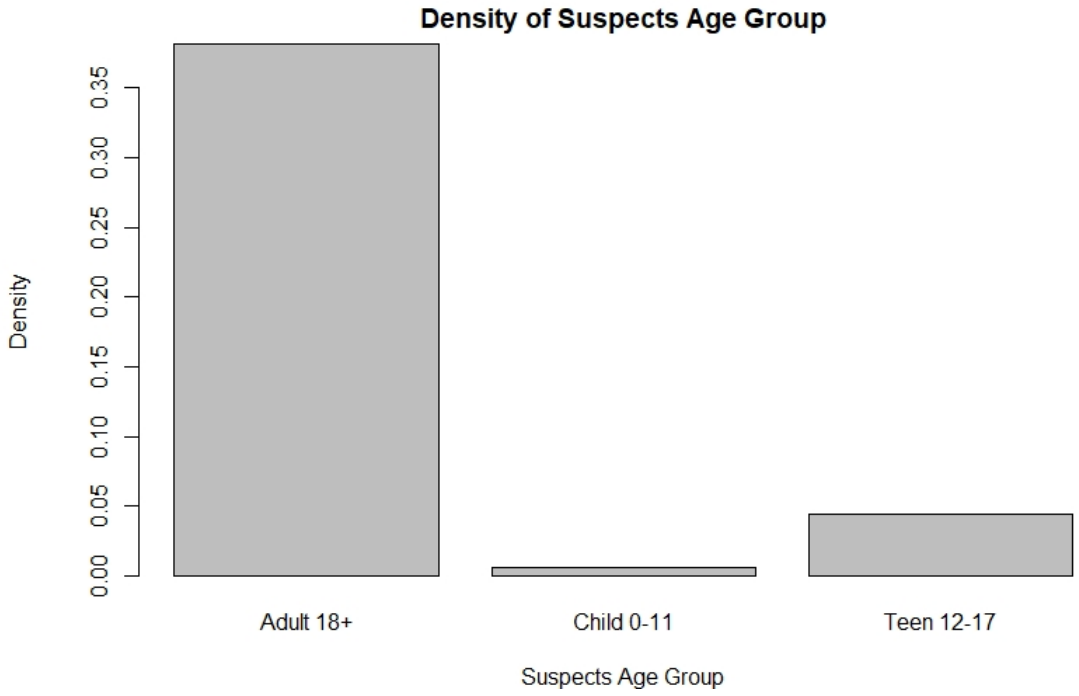

Figure 7: Density of Suspects Age Group

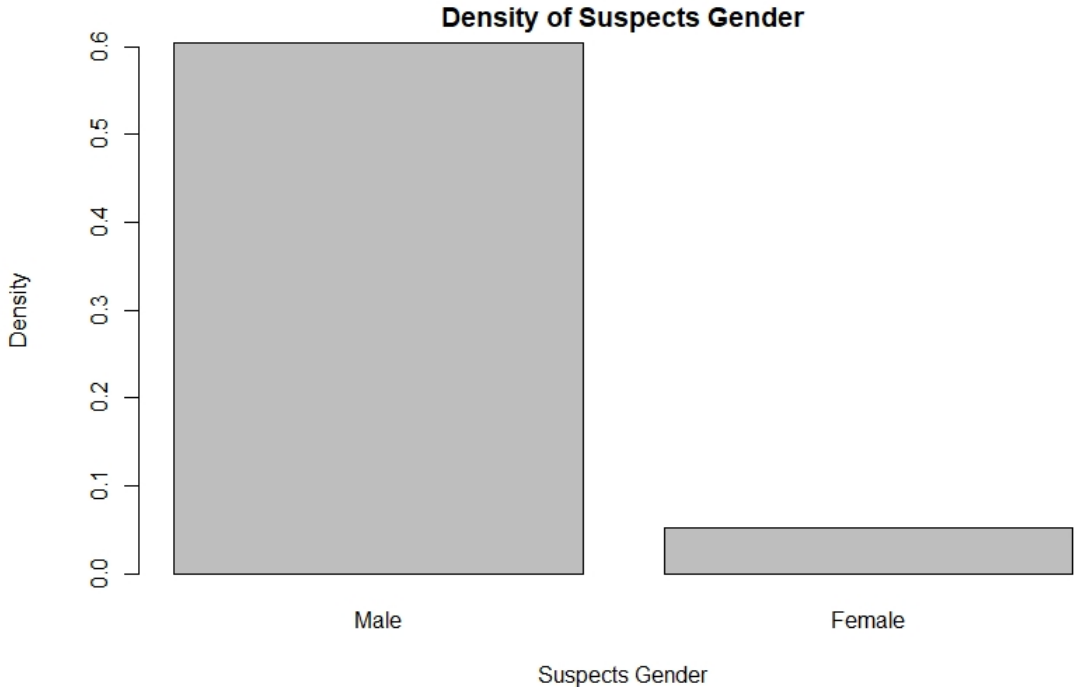

Figure 8: Density of Suspects Gender

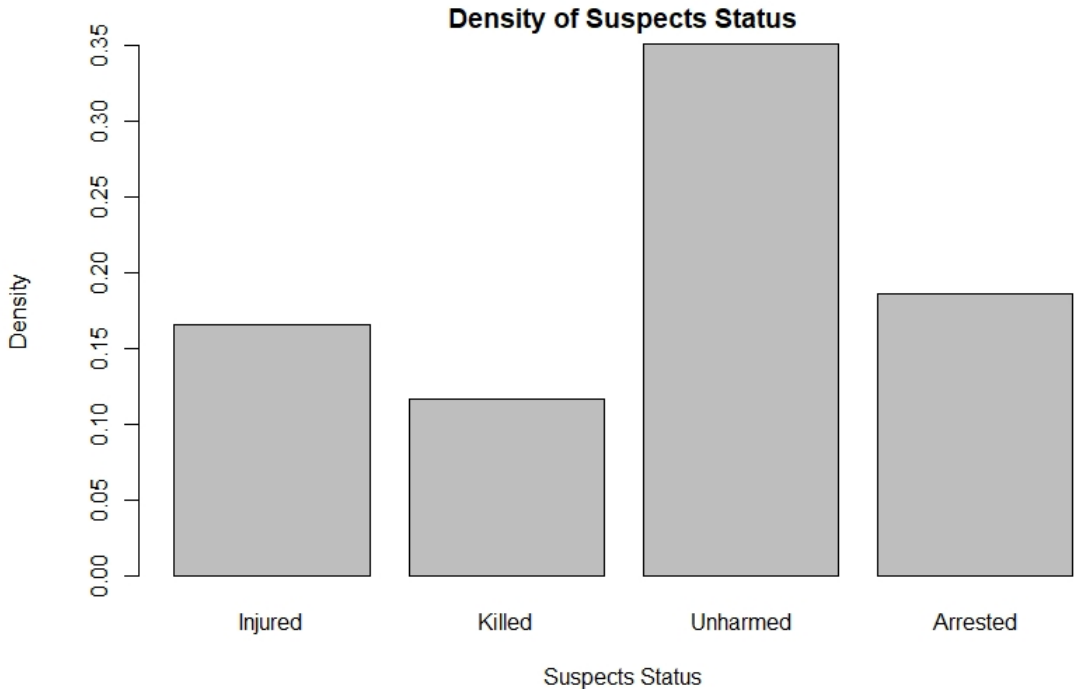

Figure 9: Density of Suspects Status

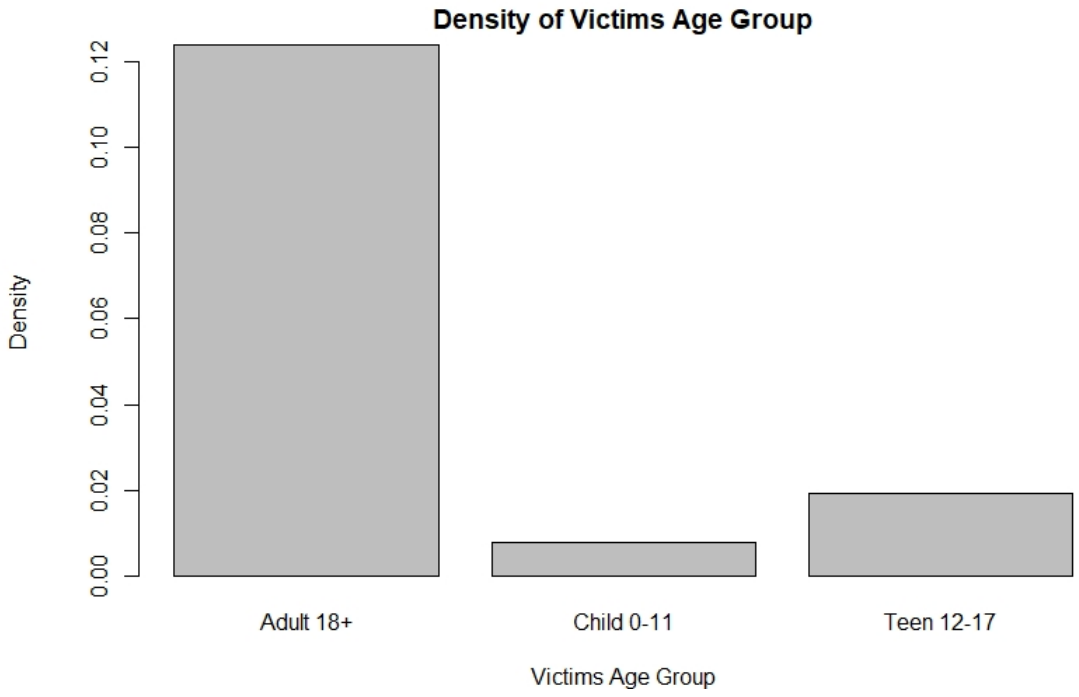

Figure 10: Density of Victims Age Group

Figure 11: Density of Victims Gender

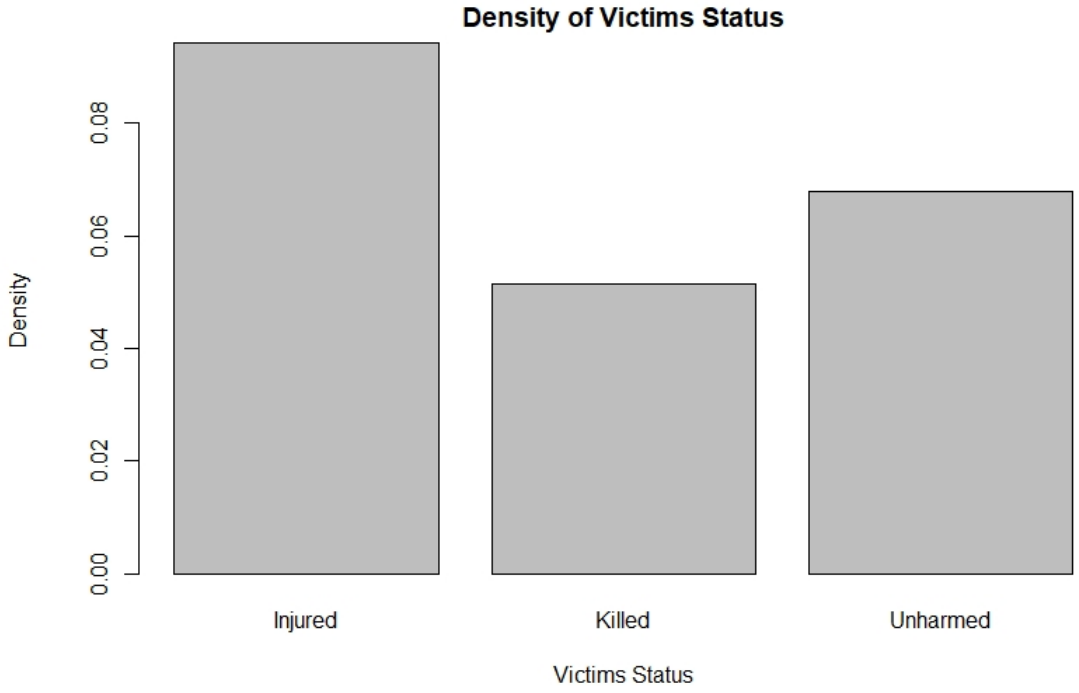

Figure 12: Density of Victims Status

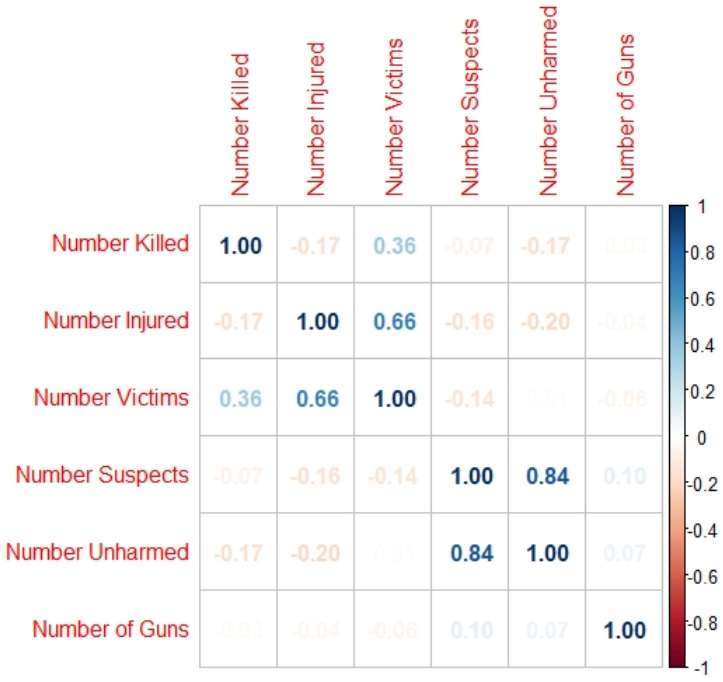

Figure 13: Correlation Matrix Continuous Variables

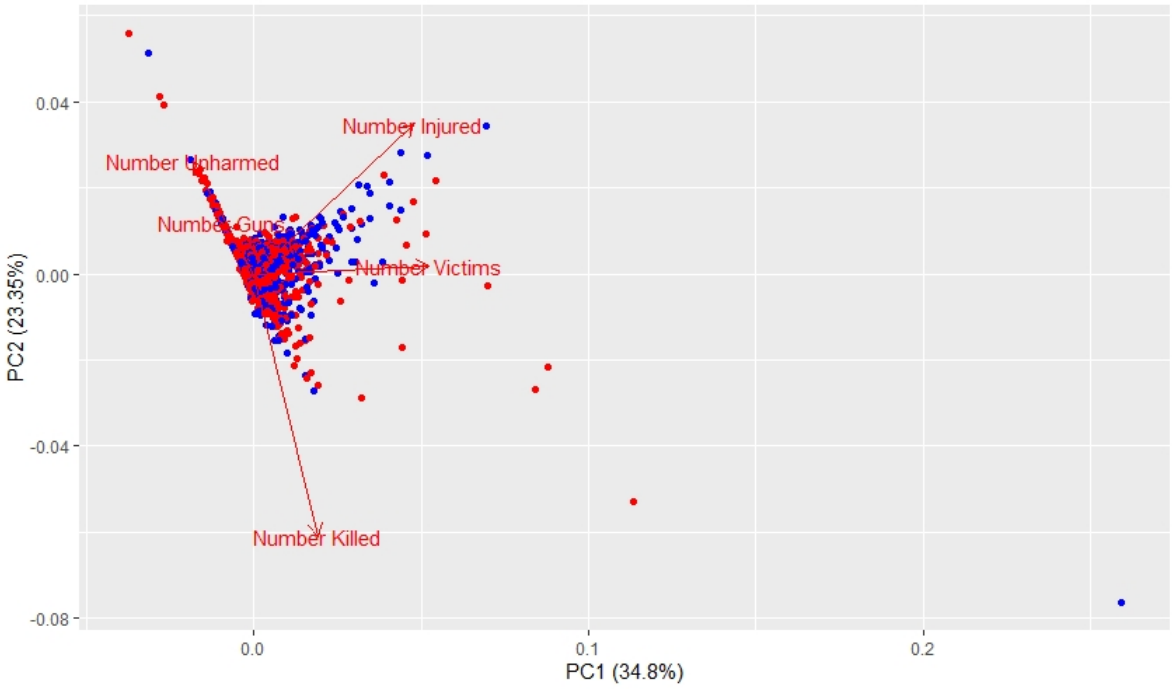

Figure 14: Principal Component Analysis Continuous Variables

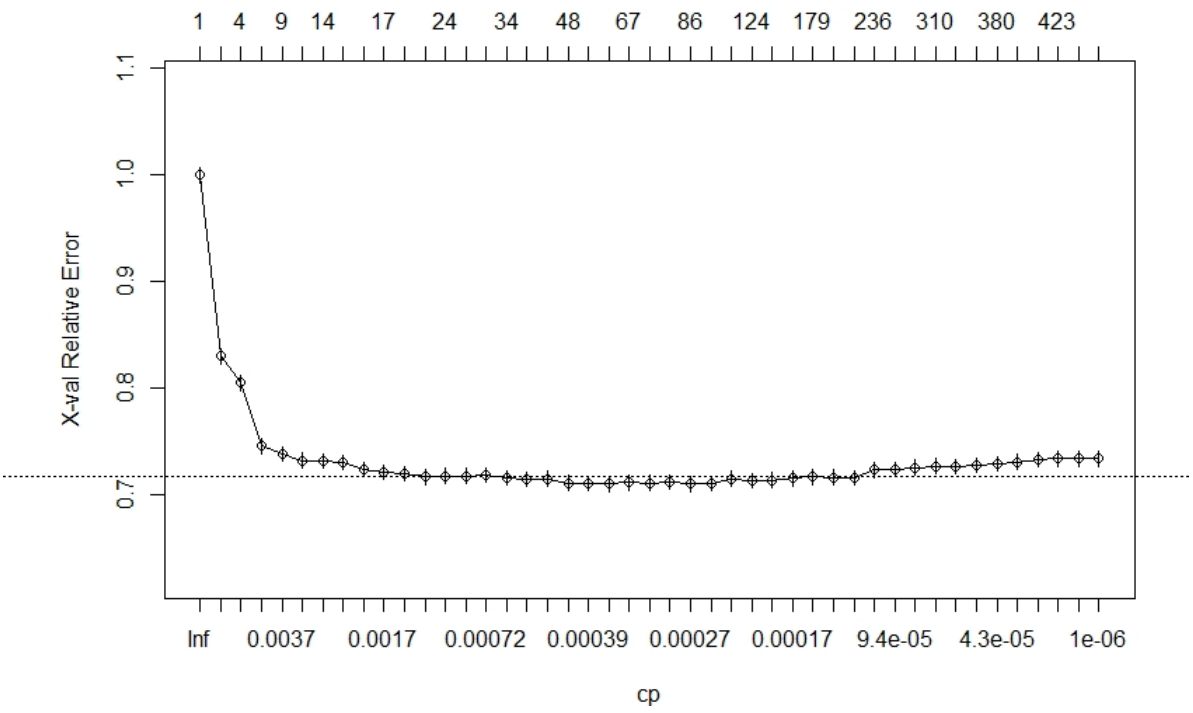

Figure 15: Tuning Complexity Parameter using Decision Tree

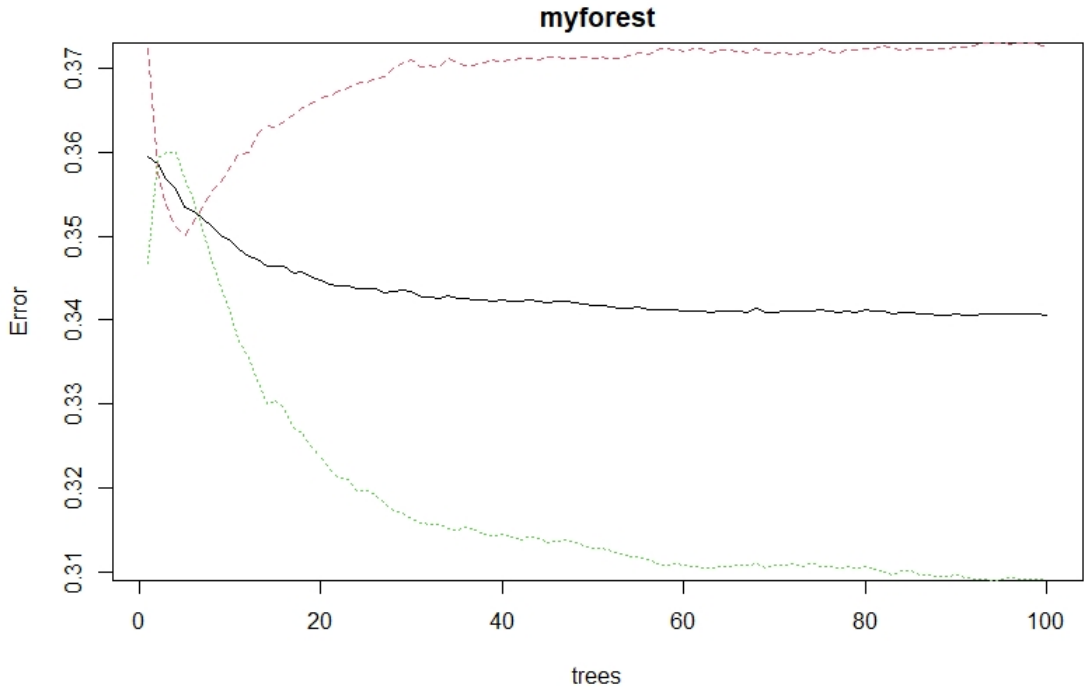

Figure 16: Out-of-Bag Error (Green: High, Red: Low)

Variable	High	Low	Accuracy Decrease	Gini Decrease
Number Victims	0.0346	0.0145	0.0245	1111.6789
Male Suspects	0.0760	-0.0176	0.0287	1045.7257
Number Unharmed	0.0274	0.0178	0.0225	939.9615
Number Injured	0.0084	0.0264	0.0175	906.1083
Arrested Suspects	0.0102	0.0424	0.0265	896.0722
Shot (Wounded/Injured)	0.0059	0.0431	0.0247	769.0275
Accident	0.0078	0.0117	0.0098	716.1097
Male Victims	0.0136	0.0141	0.0139	692.7774
Suicide	0.0029	0.0096	0.0063	532.7490
Officer Involved	0.0202	0.0009	0.0105	434.2127
Domestic Violence	0.0130	-0.0021	0.0053	369.4317
Number Guns	0.0019	0.0006	0.0012	327.1758
Number Killed	0.0190	0.0107	0.0148	324.0799
Possession of Guns	-0.0039	0.0107	0.0035	316.8322
Suspects Unharmed	0.024	0.012	0.018	314.946

Table 1: Top 15 Variables by Mean Decrease Gini

	Reference	
Prediction	High	Low
High	53602	31846
Low	26965	60225

Table 2: Confusion Matrix from Training

	Reference	
Prediction	High	Low
High	5944	3017
Low	3550	6671

Table 3: Confusion Matrix for Predictions