MGSC 661

Dec
2023

FINAL PROJECT ON GUN VIOLENCE

TABLE OF CONTENTS

INTRODUCTION

Prelude:
In the intricate web of modern life, the dark strand of gun violence weaves a recurring, somber pattern. Each incident leaves behind a shadow, measured by a mysterious number known as the "violence score". This score, a silent chronicle of loss and pain, blends the tales of lives cut short and wounds inflicted.

Quest:
Our journey is to shed light on this enigmatic "violence score," unraveling the tangled factors that dictate the severity of gun violence incidents. We embark on a voyage through the sea of data, aiming to turn numbers and facts into a lantern that guides us through the fog of uncertainty. The Map: Our treasure map is a dataset, a mosaic of information detailing the when and where of gun violence, the weapons involved, and the echoes of each incident. We navigate through geographical coordinates, trace the lineage of firearms, and interpret the subtle language of each event.

Mysteries to solve:
What hidden currents guide the violence score in the drama of gun violence? How do the choice of weapons and the stage of the incident influence its intensity? Can we craft a compass – a predictive model – to chart the course of future events, foreseeing and perhaps altering their path?

Vision of Hope:
At the end of our quest, we aspire to more than just understanding. We aim to empower the stewards of safety – from policymakers to local heroes – with foresight. This newfound clarity could shape strategies to calm the storm of violence, turning potential tragedies into stories of resilience and hope.

DATA DESCRIPTION

The dataset contains detailed records of gun violence incidents in the United States, spanning from 2013 to 2018. It includes 239,677 incidents with comprehensive information such as the date, location (state and city), number of people killed and injured, sources, notes regarding the incident and details about the participants.

Diving deep into the dataset, we observed some trends and patterns as elaborated below (some distributions are shown in Appendix):

Temporal Trends (Appendix III):
The dataset reveals a disturbing escalation in gun violence incidents from 278 in 2013 to a peak of 61,401 in 2017. This trend highlights an increasing problem over the years, necessitating urgent attention and action. The higher incidence rates in January and July might indicate seasonal influences, possibly linked to social and environmental factors such as holiday-related stress or summer vacation periods. Sundays (37,052 incidents) and Saturdays (36,096) have the highest incidence, possibly due to increased social activities, leisure time, and alcohol consumption, which might escalate conflicts. A significant number of incidents occur on Wednesdays (34,126), Mondays (33,760), and Tuesdays (33,307), indicating that gun violence is a persistent issue throughout the week. Thursdays and Fridays show slightly fewer incidents. This pattern may relate to the culmination of the workweek and the social and economic rhythms associated with it.

Geographical Distribution (Appendix I and Appendix II)
Illinois (17,556 incidents), California (16,306), and Florida (15,029) report the highest number of incidents, revealing regional disparities possibly influenced by state-specific laws, socio-economic conditions, and urbanization. Cities like Chicago (10814), Baltimore (3943), and Washington D.C. (3279), with high numbers of incidents, suggest that urban-specific factors significantly contribute to gun violence, necessitating city-centric strategies.

Demographic Analysis
The significant involvement of individuals in their late teens and twenties indicates a demographic skew towards younger populations in gun violence incidents, pointing to the need for youth-focused interventions. Males are more frequently involved in these incidents, reflecting a gender-specific pattern that requires targeted approaches to address the causes and consequences of male involvement in gun violence. The large number of victims (150,980 cases reflects the extensive human toll of gun violence, encompassing not only physical harm but also psychological and socio-economic impacts on individuals and communities. 60,510 cases involving Subject-Suspects provides insight into the number of individuals likely initiating gun violence incidents, crucial for understanding the perpetrator demographics and for developing prevention and rehabilitation strategies. The high incidence of injuries (84,826 cases) highlights the physical harm caused by gun violence, often leading to long-term health issues and psychological trauma. Each fatality (49,986 cases) represents a life lost and a ripple effect of grief and disruption in families and communities. Unharmed, Arrested (35,144 cases); Unharmed (33,142 cases): These figures indicate the involvement of individuals in incidents without physical harm, highlighting the psychological impact and the complexities of legal consequences in gun violence incidents. The arrests (3,818 cases) reflect the legal repercussions of gun violence, impacting the individuals and the judicial system. The presence of various combinations of being injured, unharmed, killed, and arrested indicates the multifaceted outcomes of these incidents, often involving multiple individuals with different experiences and consequences.

Other characteristics

The prevalence of handguns (16,618) and 9mm weapons (5,434), along with a significant number of incidents involving unspecified gun types (Appendix VI), highlights the need for comprehensive strategies targeting various firearms Top 3 incident characteristics (Appendix VII) reported turned out to *involving "Shot - Wounded/Injured" (*90,236 incidents) *"Shot - Dead (murder, accidental, suicide)"* (49,914 incidents) and *"Non-Shooting Incident"* (43,532 incidents). While going through the top 20 keywords noticed in the notes related to incident, Shot word (49,617 times) and man (24,292 times) turned out to be the most appearing keyword implying common features of an incident involving violence.

Correlation between injuries and fatalities

There is a negative correlation (-0.125) between the number of people killed and the number of people injured in an incident (Appendix X). This could suggest that incidents with higher fatalities might have fewer survivors or witnesses. Considering both fatalities and injuries contribute to violence, we introduced a new variable called violence_score, which is a linear combination of number of people injured and twice of number of people killed. We did twice for number of people killed, because someone's life is very precious and losing it would definitely be deemed as more violent when compared with being injured.

MODEL SELECTION & METHODOLOGY

Preprocessing

In the initial stages of preprocessing, considering the relevance to our problem statement, I decided to exclude several predictors deemed invalid, such as notes, addresses, participant-related, and sources-related features. Upon reviewing the dataset (refer to Appendix XII), it became apparent that there were numerous missing values. Specifically, the latitude and longitude columns had a missing data rate of only 3%; thus, I chose to omit rows with missing values in these columns. Subsequently, I employed the K-nearest neighbor algorithm, utilizing the *FNN* library, with the number of neighbors set to 10. This method utilized latitude and longitude values to impute missing data in the state_house_district, state_senate_district, and congressional_district columns. After imputation, latitude and longitude were discarded from the dataset since I determined that other features, such as the state, sufficiently addressed our problem statement.The date column was broken down into day, month, and year components. Furthermore, I refined the types of guns column based on the top five categories (detailed in the Appendix IX), breaking it down into five distinct levels. A zero in any level indicated the usage of a gun type outside the top five. A similar approach was taken for the incident characteristics column, where the top 10 keywords were identified by segmenting multiple keywords. Consequently, I created 10 levels by deconstructing the incident characteristics column. Considering the objective of addressing gun violence on a state level, I

omitted the city and county columns from the dataset. I encountered numerous missing values in the number of guns. Analysis indicated that the median, which was one, was also the most frequent value. Given the presence of outliers, I opted for the median to impute missing values to minimize the influence of these outliers. All numerical features such as the year, number of guns, violence_score, etc., were converted to a numeric data type. Conversely, categorical features such as day, state_senate_district, Handgun, etc., were cast as factors. This conversion was essential since models like the random forest and gradient boosting machine can utilize features for prediction without requiring explicit conversion into different levels. I also chose to remove the number of guns stolen column as the stolen/not stolen values represented less than 1% of the dataset. To assess skewness, I employed the *e1071* library and observed that among the three numerical predictors, the number of guns exhibited significant skewness, with a value of 67.11. The other variables displayed skewness values of less than 2. To address this, I applied a logarithmic transformation to the number of guns feature, which effectively reduced its skewness to below 10. To mitigate skewness further and enhance prediction accuracy, I applied the interquartile range method to identify and remove outliers, flagging any row with more than one outlier across the numerical columns.
Although the dataset contained only three numerical variables, I diligently ensured the absence of collinearity. I also performed a Principal Component Analysis (PCA) and generated an autoplot (see Appendix XI); however, due to the limited number of numerical variables, PCA did not yield useful insights for our analysis.

Modelling

In the modeling phase, we began by examining the feature importance scores derived from a Random Forest analysis (see Appendix XIII). It was observed that all categories of gun types, such as Handguns and 9mm, received low importance scores with the exception of 'Unknown' values. Consequently, we excluded all specific gun type levels from our analysis. Following this adjustment, we executed both Random Forest and Gradient Boosting Machine models. Additionally, we performed clustering using the K-prototypes method, which is suitable for datasets containing both categorical and numerical features. The purpose of clustering was to uncover patterns that could inform actionable recommendations for relevant authorities and organizations. The findings and subsequent recommendations will be detailed in the forthcoming sections.

RESULTS & DISCUSSION

Feature Importance

Top 5 features shown by feature importance plot (Appendix XIII) were:

Shot Dead (e.g., murder, accidental, suicide): This variable is the most significant predictor, indicating that the incidents involving people shot dead (whether through murder, accidental shootings, or suicide) is a key factor in understanding gun violence. This could suggest that incidents where fatalities occur are those most likely to be reported or recorded, or they could be indicative of the severity of gun violence incidents.
Shot Wounded or Injured: The second most important predictor is the incidents involving individual who were shot but survived with wounds or injuries. This reinforces the significance of the direct physical impact of gun violence on individuals and may relate to the frequency of such incidents.
Shots Fired, No Injuries: This predictor suggests that the mere occurrence of gunfire, even without physical injuries, is a significant aspect of gun violence. It could be associated with the level of fear, the perceived safety of an area, or the likelihood of potential future violence.
Non-Shooting Incident: Interestingly, non-shooting incidents are also a top predictor. This could include incidents where a gun was present or used to threaten, but no shots were actually fired. The importance of this variable might highlight the broader scope of gun violence, which isn't limited to just when a gun is fired.
Year: The year variable's high importance indicates that there are notable trends or changes in gun violence over the years covered in the dataset. This could point to the impact of policy changes, social movements, or other temporal factors that affect gun violence rates.

Random Forest

In this analysis, we employed a Random Forest methodology to optimize the number of trees, with the aim of minimizing the Mean Squared Error (MSE) on out-of-bag values and preventing over-fitting on the training (in-bag) dataset. The corresponding plot of MSE versus the number of trees is included in the Appendix XIV. Our findings indicated that a configuration of 500 trees resulted in the lowest MSE. In constructing the tree models, we chose to randomly sample three predictors at each split. The *Ranger* library was utilized for this purpose due to its efficient performance with high-dimensional and categorical data. The final model, as per the specified parameters, yielded a MSE of 0.314 and an R-squared value of 74.8%. A detailed summary of these results is provided in the Appendix XV.

Gradient Boosting Machine

The gradient boosting machine model was constructed using 500 trees, with each tree having a depth that allows for 4 leaf nodes. This setup was chosen to balance model complexity and computational efficiency. For the process of model training and evaluation, the dataset was divided into training and test subsets. This split was facilitated using the *caret* library, with 20% of the data allocated to the test set and the remaining 80% used for training.

This partitioning strategy aimed to provide a robust assessment of the model's performance on unseen data. Upon completion of the training process, the Gradient Boosting Machine model achieved a Mean Squared Error (MSE) of 0.315. A comprehensive summary of the model's performance and additional details are available in the Appendix XVI.

Clustering

To decide the optimal number of clusters, I created a plot of total within sum of squares v/s number of clusters 2 to 10 (Appendix XVII). This is also called an elbow plot. Once plotted, I saw the number of clusters equal to 9, showed an elbow with the minimum within sum of squares.Therefore, optimal number of clusters turned out to be equal to 9. Then, Kprotoype was computed using library *clustMixType* to create the clusters. Insights drawn from each cluster is described below:

Cluster 1: High-Intensity Incidents in California & Texas

Cluster 1 is characterized by a high frequency of "Shot Wounded or Injured" and "Shot Dead" incidents, suggesting it includes more severe gun-related violence. These incidents are evenly distributed throughout the year, indicating that the underlying causes are persistent and not seasonal. The cluster predominantly affects areas in California and Texas, with specific emphasis on California's 32nd and Texas's 24th State House Districts. This pattern suggests that these regions might be grappling with ongoing issues such as gang violence or other forms of serious criminal activities that consistently result in gun-related injuries and fatalities.

Cluster 2 & 3: Seasonal Violence Peaks in Florida

Clusters 2 and 3 both have a higher proportion of incidents resulting in injuries or death, indicating a higher severity of gun-related incidents in Florida. Particularly notable in Cluster 3 is an increase in such incidents during certain months, possibly suggesting a seasonal pattern. This could be related to factors like holiday-related stress or increased social gatherings during summer months, leading to conflicts that escalate into violence. The consistent presence of Florida's 1st Congressional District in these clusters implies that these patterns are localized and may require targeted interventions during these high-risk periods.

Cluster 4, 5, 6: Lower Fatality Rates with Varied Incidents

Clusters 4, 5, and 6 are marked by very low rates of "Shot Wounded or Injured" and "Shot Dead", suggesting incidents without serious casualties, such as accidental discharges or discovery of weapons. Cluster 6 is particularly noteworthy for its high rate of drug involvement and a significant number of guns involved, hinting at possible links to drug-related activities, especially in Massachusetts's 48th State House District. The low fatality rates in these clusters, combined with specific patterns such as drug involvement in Cluster 6, point to a different set of challenges compared to the more violent clusters.

Cluster 7: Moderate Violence in Texas
Cluster 7 shares similarities with Cluster 1 in terms of the nature of violence but with a slightly lower rate of "Shot Dead" incidents. This cluster, predominantly affecting Texas and specifically its 5th State Senate District, appears to experience a slightly less severe but still concerning level of gun-related violence. The nature of incidents in this cluster might be similar to those in Cluster 1, but the slightly lower fatality rate could indicate differences in the circumstances or contexts of these incidents

Cluster 8: Non-Lethal Aggression in Florida
Cluster 8 stands out for its higher rate of "Brandishing or Flourishing" incidents, indicating a tendency towards less lethal outcomes in gun-related incidents in Florida. This cluster also shows a notable rate of "Armed robbery with injury or death", suggesting a specific pattern of criminal behavior where guns are used more for intimidation than lethal purposes. The prevalence of these types of incidents, particularly in Florida's 18th State House District, points to a unique set of social and criminal issues distinct from other clusters.

Cluster 9: Non-Injurious Gunfire in Florida
Cluster 9 is unique for its high occurrence of "Shots Fired No Injuries", indicating incidents in Florida where firearms are discharged but do not result in physical harm. This pattern, particularly observed in Florida's 92nd State House District, might suggest the use of guns as a means of intimidation or warning, rather than an intent to cause physical injury. The presence of such incidents could reflect a culture of firearm use for communication or display of power within certain communities.

CONCLUSION & RECOMMENDATION

Prediction
Considering that the Mean Squared Error (MSE) for our Random Forest model is smaller than that of the Gradient Boosting Machine model, key responsible authorities in the USA, such as the Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF), the Federal Bureau of Investigation (FBI), and others, can effectively utilize our Random Forest model for predicting violence scores in different states in the coming years. This predictive capability is crucial for guiding targeted actions and shaping policies aimed at reducing violence scores nationwide.

The lower MSE of the Random Forest model signifies its higher accuracy and reliability in capturing the complex dynamics of gun violence. This precision is essential for authorities like the ATF and FBI, enabling them to identify potential hotspots of violence and understand evolving trends. By harnessing this predictive power, these agencies can allocate resources more effectively, design targeted intervention programs, and implement proactive measures in areas most at risk.

Some notable insights using the prediction modelling task included:

Historical Patterns and Societal Evolution: The importance of the 'Year' variable can be interpreted as a mirror reflecting how societal attitudes, legal frameworks, and technological advancements influence gun violence trends. It's like looking at a time-lapse of societal changes, where each frame is a year and each pixel change represents an incident.
Narrative of Gun Presence: The significance of 'Non-Shooting Incident' suggests a hidden narrative in society. It's not just the act of violence that matters, but the mere presence of a gun alters the social fabric, akin to a silent character in a play that changes the scene's tone without saying a word.
Gunfire as a Communication Tool: The 'Shots Fired, No Injuries' variable could be interpreted as gunfire being used as a form of communication – a loud, dangerous one. It's like a morse code of distress or dominance, conveying messages in a language of violence.
Physical vs Psychological Impact: The contrast between 'Shot Dead' and 'Shot Wounded or Injured' variables offers a lens to view the physical versus psychological impact of gun violence. While fatalities are the tragic endpoint, the number of people wounded tells a story of ongoing trauma and resilience in the face of violence.
Gun Violence as an Ecosystem: The interplay of different variables can be viewed as an ecosystem, where various forms of gun violence coexist and interact. Like different species in a rainforest, each type of incident plays a role in the broader ecosystem of urban and societal violence.

Recommendation based on Clustering Insights

1. Storytelling Through Data Visualization
Develop an interactive digital map showing the spread of incidents across California, Florida, Texas, Massachusetts, and Illinois. Each state could have a distinct visual theme representing its cluster's unique characteristics. For instance, a heatmap could represent the concentration of incidents in each congressional and state district, with California showing more intense colors for high-fatality incidents.

2. Cluster Personality Profiles
Assign each cluster a distinctive personality based on its state and the nature of incidents. Cluster 1 (California, high fatalities): "The Intense One" can be marked by its severe outcomes, this cluster might be depicted as a stormy weather graphic, symbolizing the urgent need for intervention.Cluster 5 (California, non-shooting incidents): "The Quiet Threat" can be visualized as a ticking clock, representing the potential danger that lies beneath the surface.

3. Predictive Art Installations
In areas corresponding to Cluster 8 (Florida, low fatalities), install dynamic public art pieces that change color or pattern based on real-time data predictions of gun-related incidents. For instance, a sculpture in a public park could glow peacefully on calm days but change to a stark red on days with higher risk predictions.

4. Cluster-Specific Social Experiments
For Cluster 6 (Massachusetts, possession-related incidents), conduct a mock scenario in a busy public area demonstrating the quick escalation of a situation due to illegal gun possession. Actors and hidden cameras could be used to capture public reactions and to educate on the spot.

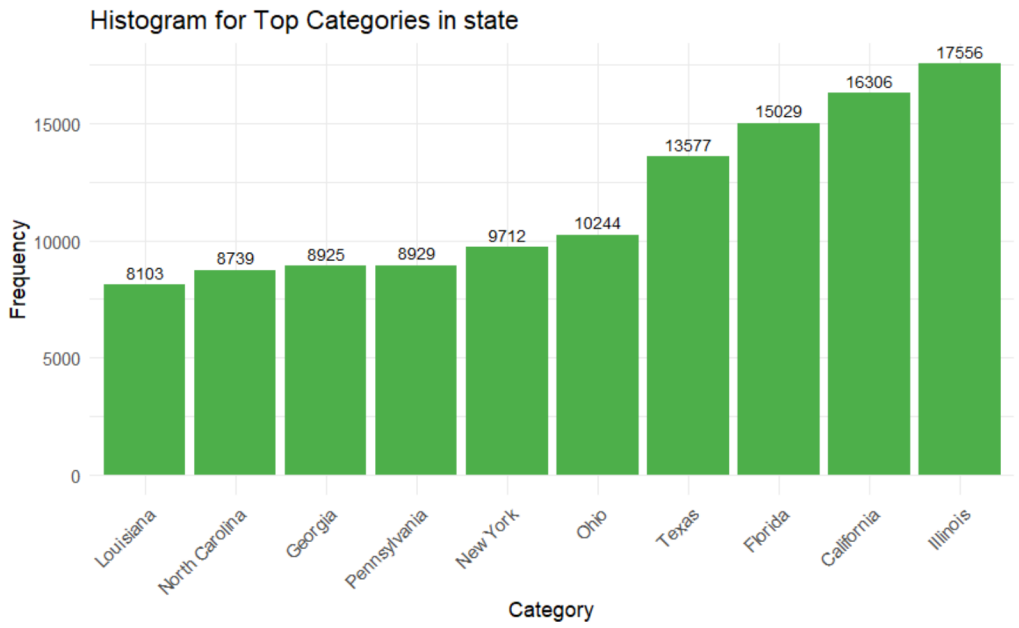
5. Gamification of Gun Safety Education
Develop a mobile app with game scenarios based on clusters 2 and 3 (Florida, high injury rates). Players navigate through digital versions of these districts, making decisions in various conflict scenarios, learning about gun safety and legal consequences.

6. Cluster-Inspired Policy 'Hackathons'
Host a policy hackathon in Texas (Cluster 7) focusing on the challenges of high fatalities and non-shooting incidents. Involve local tech enthusiasts, policymakers, and community leaders to brainstorm technological and social solutions specific to Texas's unique challenges.

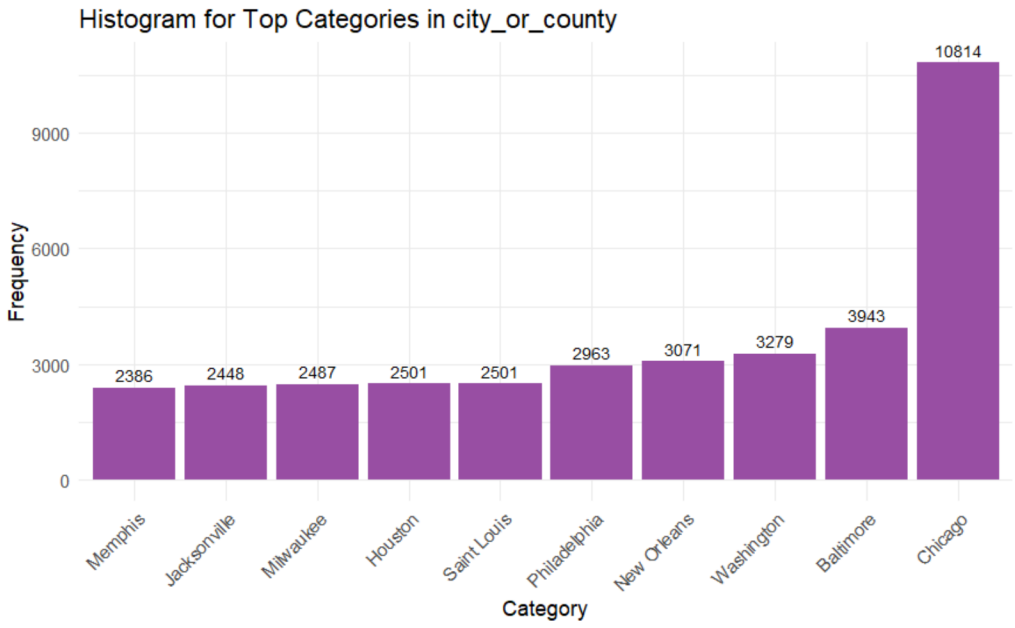
7. Themed Community Engagement Events
In California districts corresponding to Cluster 1, organize community forums themed around the impact of gun violence on families and communities. Include art exhibits from local artists depicting the emotional and social impact of such incidents, combined with workshops on conflict resolution and mental health.

Appendix

Appendix I: Distribution of state

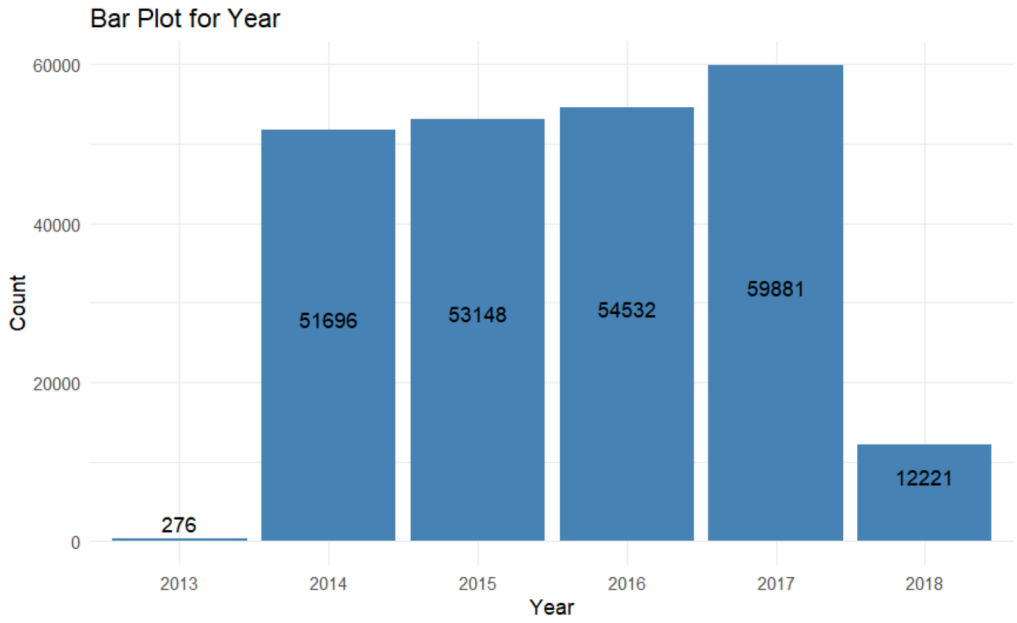


Appendix II: Distribution of city/county

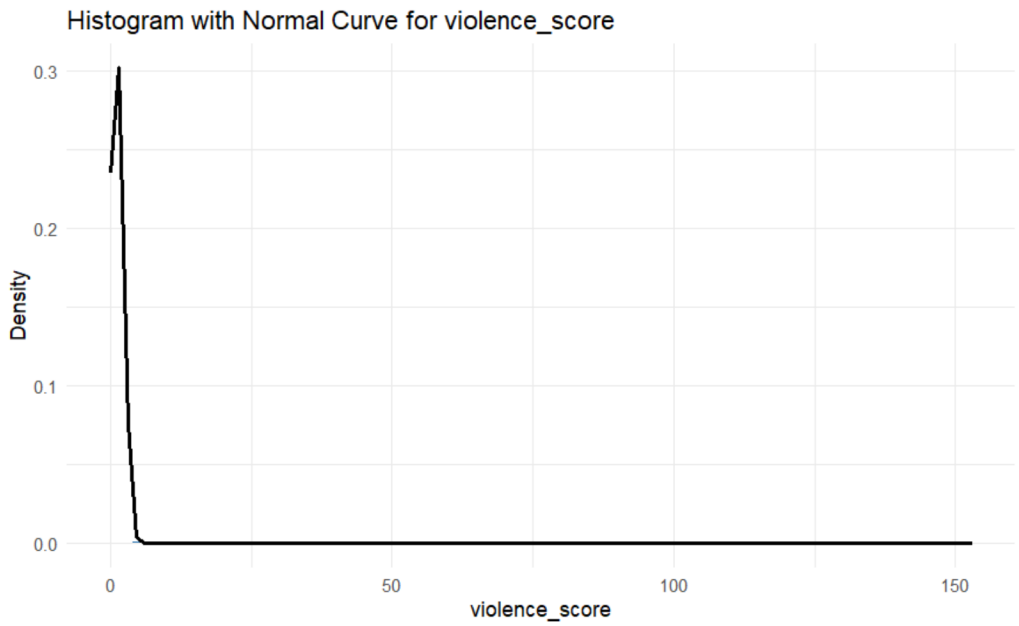


Appendix III: Word cloud showing top 20 keywords in Notes

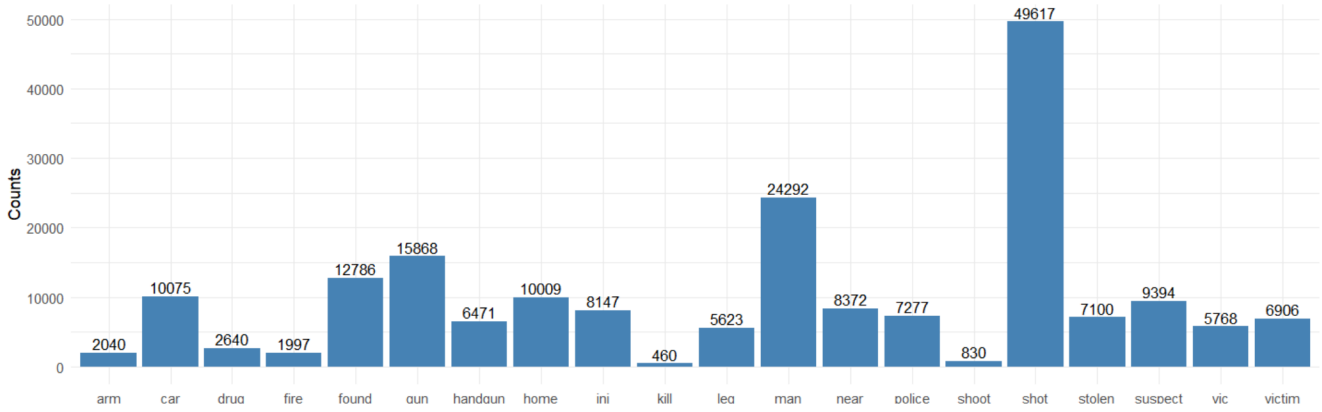
Appendix IV: Distribution of Year



Appendix V: Distribution of violence_score



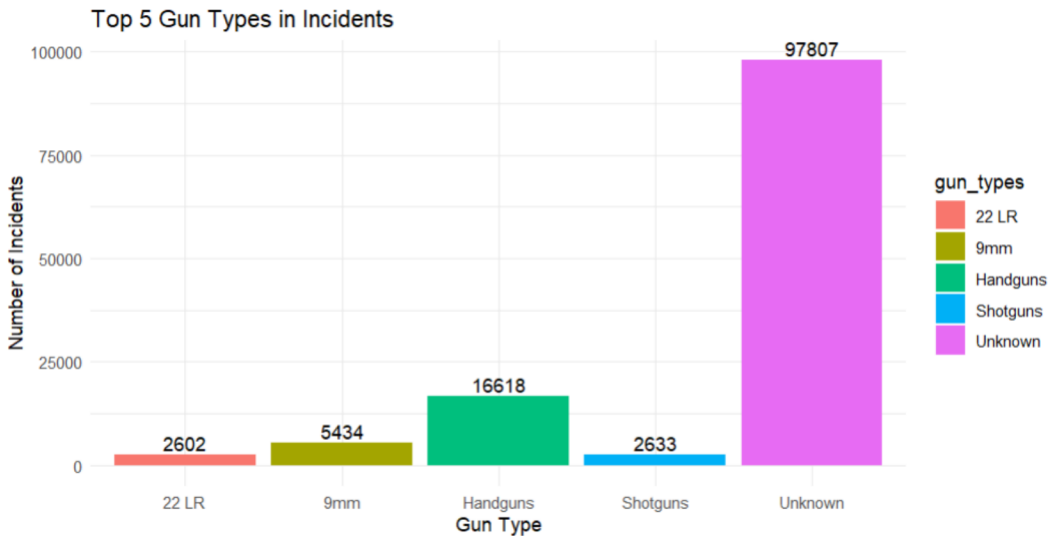
Appendix VI: Distribution of top 20 keywords in notes



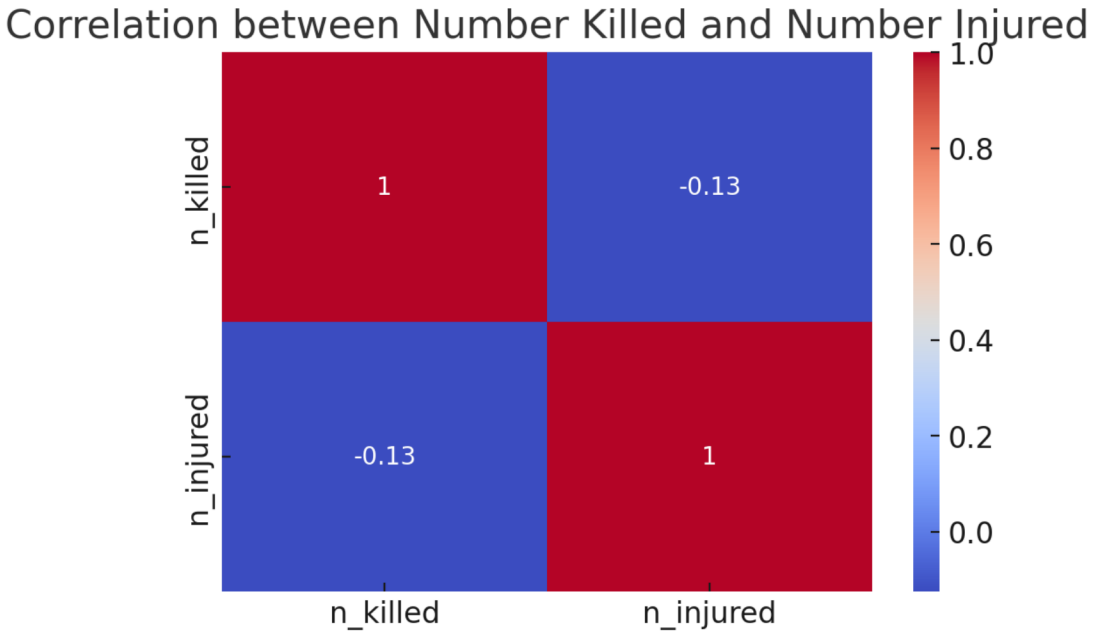
Appendix VII: Distribution of top 10 incident characteristics

Incident_Characteristic	Number of Observations
Shot - Wounded/Injured	90,236
Shot - Dead (murder, accidental, suicide)	49,914
Non-Shooting Incident	43,532
Shots Fired - No Injuries	34,488
Possession (gun(s) found during commission of other crimes)	29,813
Armed robbery with injury/death and/or evidence of DGU (Defensive Gun Use	18,746
Brandishing/flourishing/open carry/lost/foun	17,978
ATF/LE Confiscation/Raid/Arrest	17,249
Officer Involved Incident	16,940
Drug involvement	16,332

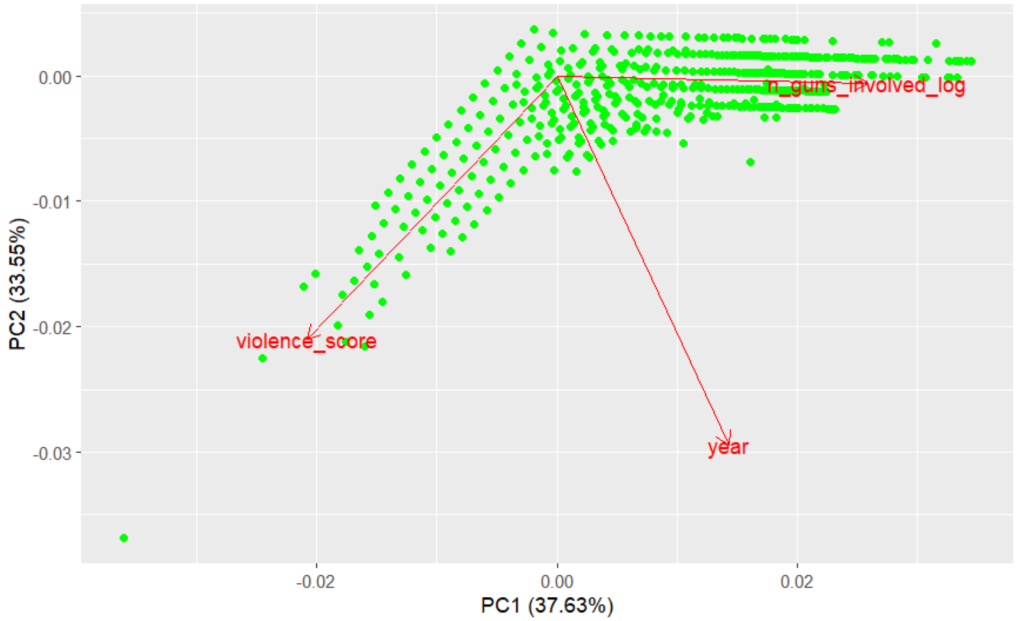
Appendix IX: Distribution of top 5 Gun types



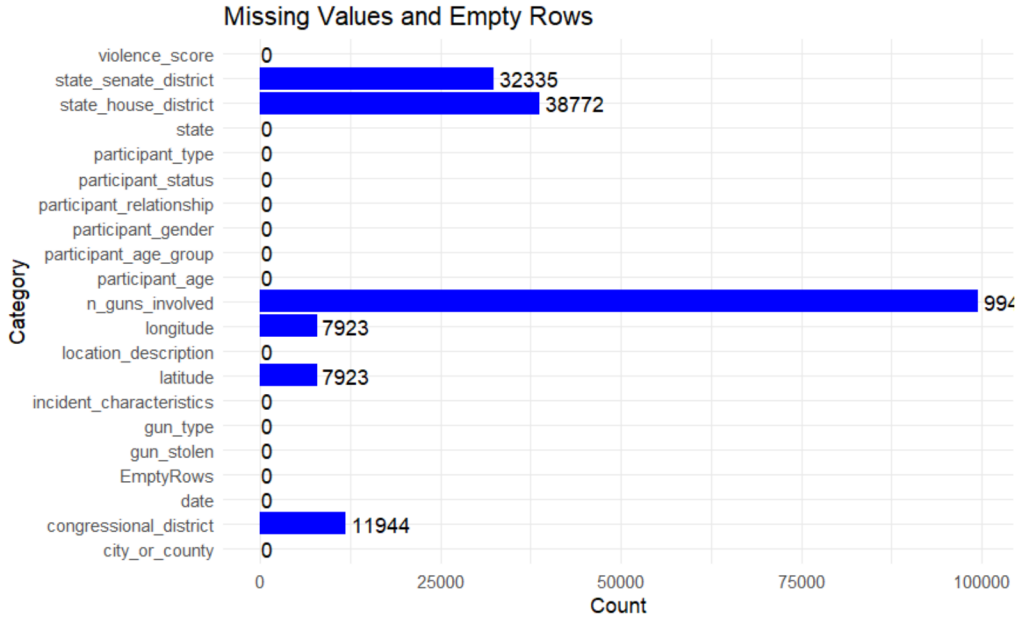
Appendix X: Correlation between people killed and injured



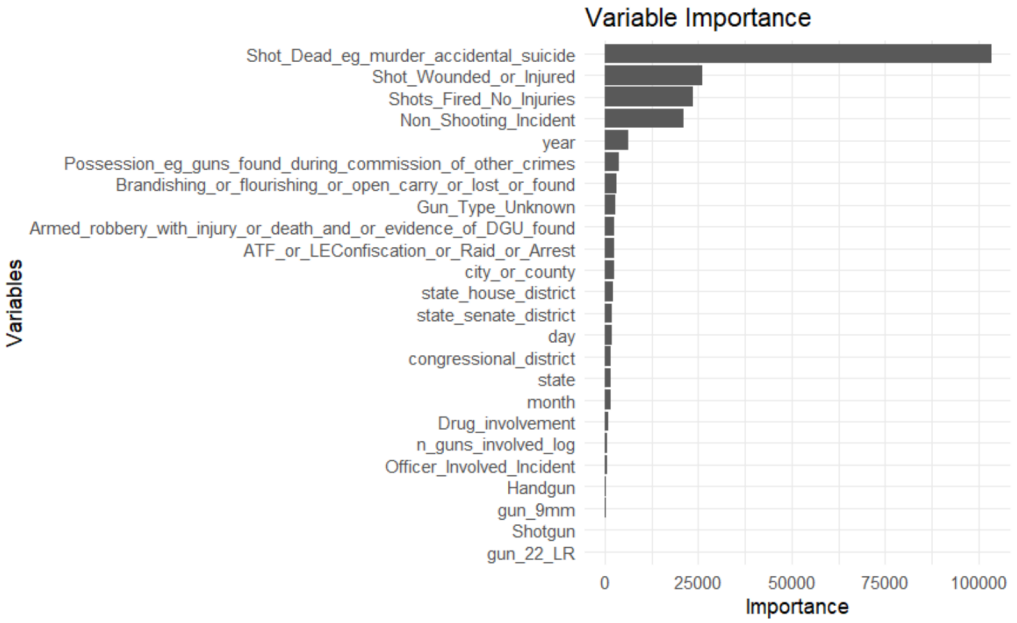
Appendix XI: Principal Component Anlaysis



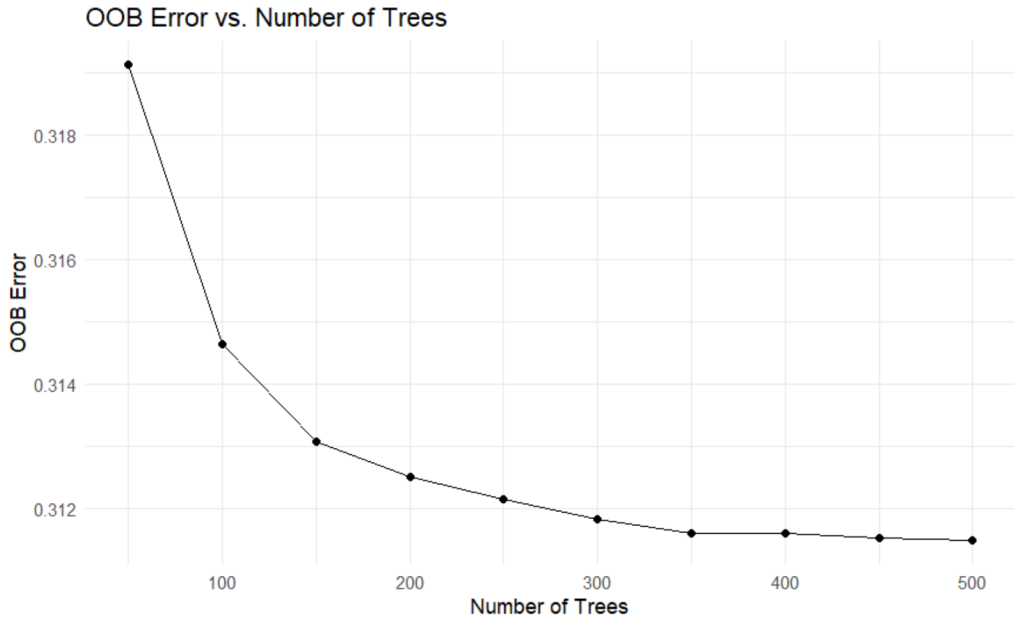
Appendix XII: Missing values in each column



Appendix XIII: Random Forest Feature Importance



Appendix XIV: OOB MSE v/s Number of trees (Random Forest)



Appendix XV: Random Forest Model

```
Call:
 ranger(formula = violence_score ~ ., data = gun_3, num.trees = 500,      mtry = 3, importance = "impurity",
save.memory = TRUE, probability = FALSE)

Type:                             Regression
Number of trees:                  500
Sample size:                      231621
Number of independent variables:  20
Mtry:                             3
Target node size:                 5
Variable importance mode:         impurity
Splitrule:                        variance
OOB prediction error (MSE):       0.3138104
R squared (OOB):                  0.7477329
```

Appendix XVI: Gradient Boosting Machine Model

```
gbm(formula = violence_score ~ ., distribution = "gaussian",
    data = train_data, n.trees = 500, interaction.depth = 3,
    shrinkage = 0.01, cv.folds = 5)
A gradient boosted model with gaussian loss function.
500 iterations were performed.
The best cross-validation iteration was 497.
There were 20 predictors of which 13 had non-zero influence.
```

Appendix XVII: Within Sum of Squares v/s Number of Clusters (Elbow Plot))