Predictive Analytics in Gun Violence: Enhancing Police Preparedness and Response

Wenya Cai

McGill University

Course Number: MGSC 661

Author Note

All visualizations in this report are color-blind friendly, emphasizing my commitment to accessibility in data analytics for inclusive understanding. For more information, please refer to the Appendix - Color Accessibility in Data Analytics

Predictive Analytics in Gun Violence: Enhancing Police Preparedness and Response

Overview

This report encapsulates a detailed analysis of over 260,000 gun violence incidents in the United States, spanning from 2013 to 2018. The primary objective is to harness machine learning techniques to predict potential outcomes of these incidents, particularly focusing on the number of victims and the likelihood of police officers being shot or fatally wounded. The predictive models developed in this study aim to provide police forces with critical foresight, enhancing their preparedness before arriving at the scene. Furthermore, the findings can offer valuable insights to local hospitals, enabling them to mobilize resources efficiently in anticipation of potential casualties.

Data Exploration and Description

Overview of the Dataset

The dataset under analysis originates from the Gun Violence Archive (GVA), a non-profit organization dedicated to providing comprehensive and accurate data about gun-related violence in the United States. This dataset encompasses recorded incidents of gun violence from January 2013 to March 2018. An initial examination of the data reveals a concerning upward trend in gun violence incidents. While the data for 2013 and 2018 are incomplete, a clear increase is observed from 2014 to 2017. Notably, there were approximately 51,000 incidents in 2014, escalating to 61,000 in 2017. This represents a significant increase of 10,000 incidents over three years. (Figure 2)

Dissecting Seasonal and Cultural Influences on Gun Violence

Weekend Prevalence

A notably higher occurrence of gun violence incidents is seen on weekends, with around 6,000 incidents typically occurring on Saturdays and Sundays. (Figure 3)

Holiday Peaks

The data shows pronounced peaks around January 1st, and the Independence Day period (July 4th and 5th). (Figure 4) This could be linked to the tradition of discharging firearms during celebrations, despite legal prohibitions and associated risks. Research indicates that bullets fired into the air can climb up to two miles and stay airborne for over a minute. Upon descent, these bullets can attain a velocity sufficient to cause fatal injuries, emphasizing the risks of celebratory gunfire.[1]

Geographic Distribution of Incidents

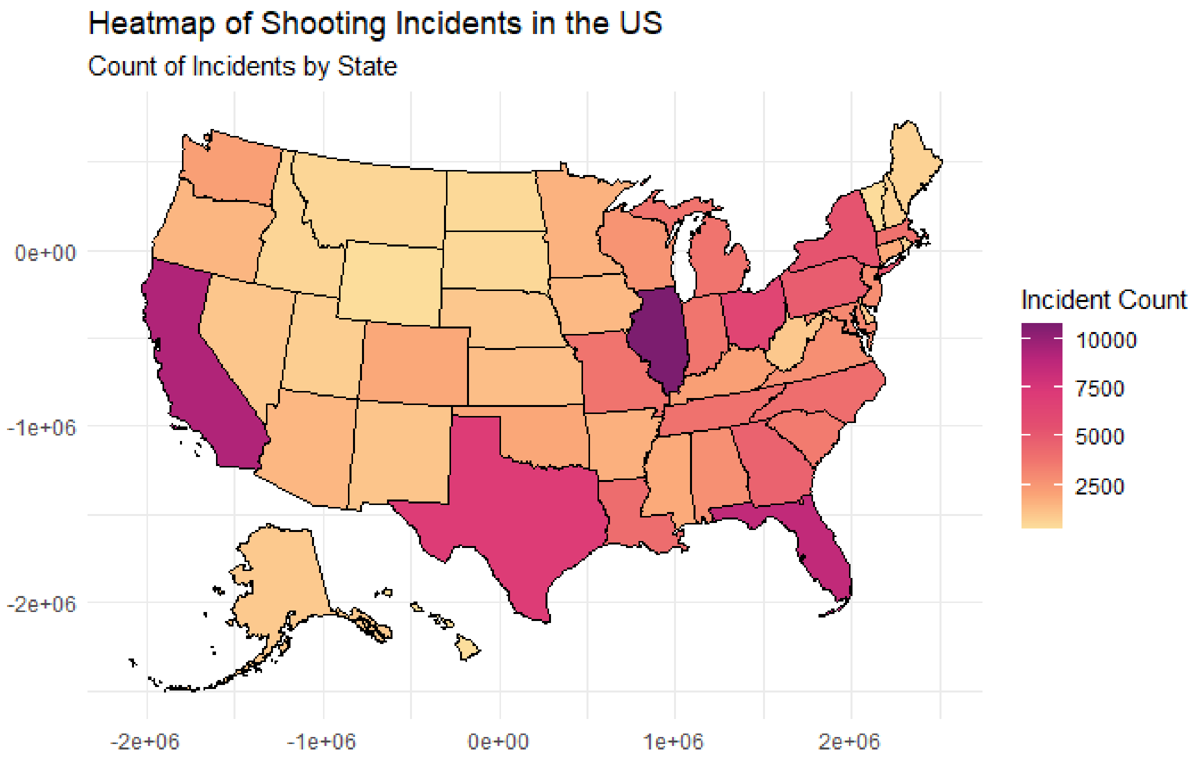


Figure 1

The dataset reveals a varied geographic distribution of gun violence incidents, with certain states and locations emerging as hotspots. Notably, urban areas and specific locales such as retail chains and public spaces feature prominently (Table 2). Figure 1 offers a visualization of the state-wise distribution and intensity of gun violence incidents across the United States. The heatmap serves to illustrate the disproportionate concentration of incidents in specific regions, including Illinois, California, and Florida.

[1] New Years Eve Gunfire Reduction Program. LAPD Online. (2021, May 22)

Data Selection and Preprocessing Strategy

For effective analysis, the dataset has been narrowed down to cover the period from 2014 to 2017. This period provides complete annual data, enabling a comprehensive understanding of trends. The focus is on incidents involving a single suspect with complete data across all variables, which still leaves a substantial dataset of more than 50,000 cases for analysis. Given the limitations of R in manipulating large datasets efficiently, Python is utilized for feature engineering. This approach enables the extraction of relevant features and the preparation of the dataset for advanced analytical models.

Key Variables for Analysis

Date and Time Variables

1. *Date*: The specific date of the incident, converted into a numerical format for model compatibility.
2. *Year, Month, Weekday*: Numerical variables. Extracted from the date to capture specific temporal patterns.
3. *Days_since_start*: Numerical variable. Represents the number of days since the beginning of 2014, useful for trend analysis over time.

Incident Specifics

4. *Is_weekend, on_jan1, on_jul4*: Binary variables, indicating if the incident occurred during identified peak times.
5. *State, City_or_Coun*ty: Categorical variables, with City/County further refined due to high category counts. Indicating where the incident took place.
6. *Congressional_district, State_senate_district, State_house_district*: Numerical variables. Geographical identifiers provide additional location context.

Impact Assessment

7. *N_injured, N_killed:* Numerical variables. Quantitative measures of the incident's impact.
8. *Victims:* Numerical variables. A combined count of injured and killed, offering a broader impact perspective.

Suspect Information

9. *Suspect_age_group:* Binary variable. Distinguishes between adults (18+) and teens (12-17).
10. *Suspect_exact_age:* Numerical variables. Provides specific age data.
11. *Suspect_gender*: Binary classification of the suspect's gender.

Officer Involvement

12. *Officer_shot, Officer_killed, Officer_involved:* Binary variables indicating any law enforcement involvement and consequences.

Model Selection and Methodology

Random Forest for Feature Selection

In the endeavor to analyze gun violence incidents, Random Forest method was employed to evaluate the importance of various factors. The analysis was conducted using the *randomForest* function in R, focusing on two metrics: *%IncMSE* and *IncNodePurity*. These metrics assess the impact of excluding a variable on the Mean Squared Error and the effect of a variable on node impurity reduction, respectively. Our findings revealed that the *state_house_district* variable was crucial in predicting the number of victims, as indicated by a high *%IncMSE* value. Conversely, the state variable was prominent in *IncNodePurity,* suggesting its significant influence in reducing node impurity in the model. Interestingly, while variables like *suspect_age_group* and *suspect_gender* showed lower importance in *%IncMSE, suspect_exact_age* exhibited higher importance in *IncNodePurity*. This highlighted that while some variables might not worsen the Mean Squared Error significantly when excluded, they still provide substantial information for making splits in the tree. Temporal variables such as date, year, and month were found to be consistently significant, implying that the timing of an incident is a key predictor for the number of victims. However, specific dates like *on_jul4* and *on_jan1*, as well as the *is_weekend* variable, showed limited predictive power, leading to their exclusion from further analysis like logistics regression.

Victims Prediction with Tree-based Models

In the development of our predictive models, we constructed both a Random Forest and a Gradient Boosting model, utilizing pre-incident data such as date and location. Given the likelihood of these details being observed by witnesses, we opted for broader categories like suspect age group and gender, instead of more specific data like exact age. Each model was configured with 1000 trees and an interaction depth of 4, operating under the assumption of a Gaussian distribution for the target variable. The data was partitioned into training (75%) and testing (25%) sets, and the models' performances were evaluated based on the Mean Squared Error (MSE) by comparing the predicted number of victims to the actual figures in the test set. Additionally, a comprehensive Random Forest model was developed, incorporating extensive post-incident data, including exact age and officer involvement. This model was intended to yield deeper insights, particularly regarding the effects of various factors post-incident.

Officer Involvement Prediction with Logistics Regression

In our statistical analysis aimed at predicting police involvement and casualties in gun violence incidents prior to their arrival, we developed two logistic regression models. Acknowledging the nuanced care logistic regression requires compared to tree-based models, we tailored our approach to enhance model accuracy and relevance. For both models, the dataset was divided into training and testing sets based on the binary outcomes *'officer_involved'* and *'officer_killed'*. The first model was designed to predict the likelihood of police involvement in an incident, while the second focused on the probability of officer casualties.

To optimize the logistic regression models, we streamlined the variable selection. We removed variables with lesser predictive power such as *is_weekend, on_jan1,* and *on_jul4*. This decision was informed by their limited influence on the outcome in our preliminary analyses. Additionally, we chose to exclude the specific date variable, instead utilizing *days_since_start, year,* and *month.* This approach allowed us to capture essential time trends without overcomplicating the model. Geographical information was refined by focusing on district-level data rather than state-level to prevent multicollinearity, thus ensuring a more reliable and interpretable model.

Model Performance and Results

Comparative Analysis of Random Forest and Gradient Boosting Models

Analysis on pre-arrival prediction revealed that the Random Forest model achieved a Mean Squared Error (MSE) of 0.618, outperforming the Gradient Boosting model, which recorded an MSE of 0.640. This result is noteworthy considering the widespread acclaim of Gradient Boosting as one of the most potent predictive algorithms currently available. The superior performance of the Random Forest model in this context can be attributed to its robustness against overfitting and noise. Random Forest, by averaging results across numerous decision trees, demonstrates resilience to extreme values, making it less susceptible to being influenced by outliers. In contrast, Gradient Boosting, which iteratively refines its predictions based on previous errors, can be more sensitive to noise and outliers. This sensitivity often limits its performance, especially in scenarios where the signal-to-noise ratio is not optimal.

Enhanced Performance with Comprehensive Post-Incident Data

A notable improvement was observed when the Random Forest model was applied with a complete set of post-incident data. The MSE in this case dropped to 0.595, reinforcing the importance of comprehensive data in enhancing predictive accuracy. Factors considered in the post-incident model, as identified in the earlier feature importance analysis, played a crucial role in reducing the MSE and refining the model's predictive capabilities.

Tree-based Model Predictions Visualization

To further our understanding of the models' performance, we employed *ggplot2* to create a multi-lines chart comparing actual and predicted numbers of victims (Figure 5). This visualization confirmed our initial hypotheses, illustrating that while both models captured the overall trend, there were notable differences in their behavior. The Gradient Boosting model exhibited more fluctuations, potentially making it more adept at predicting unusual and large-scale incidents, such as the Charleston church shooting in 2015, where the patterns of violence deviate markedly from the norm. On the other hand, the Random Forest model displayed greater stability, aligning with the fact that most gun violence incidents are of a smaller scale, contributing to its lower MSE.

Logistic Regression for Officer Involvement Evaluations

In assessing police dynamics, the logistic regression model yielded impressive accuracy rates: 0.989 for predicting officer involvement and 0.997 for officer fatality rates in test sets. These encouraging results suggest significant potential in using such models to anticipate and, potentially, mitigate future incidents involving officer casualties.

Conclusions

The significance of this study is underscored by the broader context of rising gun violence in the United States, affecting both civilians and police officers. Recent trends, as reported by sources like the

FBI and the Fraternal Order of Police, indicate an increase in violent crimes, including those targeting law enforcement personnel[2]. This environment amplifies the need for predictive tools that can provide law enforcement agencies with timely and accurate information about potential threats and risks.

By harnessing the power of data analytics, our analysis seeks to bridge the gap between the onset of a gun violence incident and the arrival of first responders. The insights gleaned from this study are intended not only to bolster police safety and effectiveness but also to aid local medical facilities in preparing for potential emergencies. Our approach combines statistical sophistication with practical application, aiming to make a significant contribution to managing and mitigating the impacts of gun violence in American communities.

[2] BBC. (2022, April 25). Number of US police officers murdered up by 59% - FBI.

Appendix - Color Accessibility in Data Analytics

In the realm of data analytics, the importance of accessibility cannot be overstated, particularly in the use of color schemes in visual representations. Statistics indicate that about 1 in 20 individuals experience some form of colorblindness. This condition arises from variations in cone cells within the eyes, leading to a diminished or altered perception of color. Key to understanding this phenomenon is the concept of "confusion lines" in color space, which represent hues that appear indistinguishable to those with color vision deficiencies.

To address this challenge, my report adopts a color-blind friendly approach in all visualizations. This method involves mathematically replicating the confusion line phenomenon to modify colors in a way that two hues appearing distinct to those with normal vision may merge into a singular hue for someone with colorblindness. By doing so, I aim to provide an equitable visual experience, allowing individuals with varying degrees of color vision to perceive and interpret data effectively.

I wish to extend my special thanks to resources like the website dedicated to simulating colorblindness for people with normal or near-normal color vision[3], and to the author of the *rcartocolor* library in R, which offers the option to display only colorblind-friendly palettes. The color schemes used in this report are chosen from these sources, ensuring that the visualizations are accessible and inclusive. This commitment to accessibility underscores the ethos of my data analytics approach, where clarity, comprehension, and inclusivity are paramount.

[3] https://davidmathlogic.com/colorblind/#%23648FFF-%23785EF0-%23DC267F-%23FE6100-%23FFB000

Appendix - Graphs

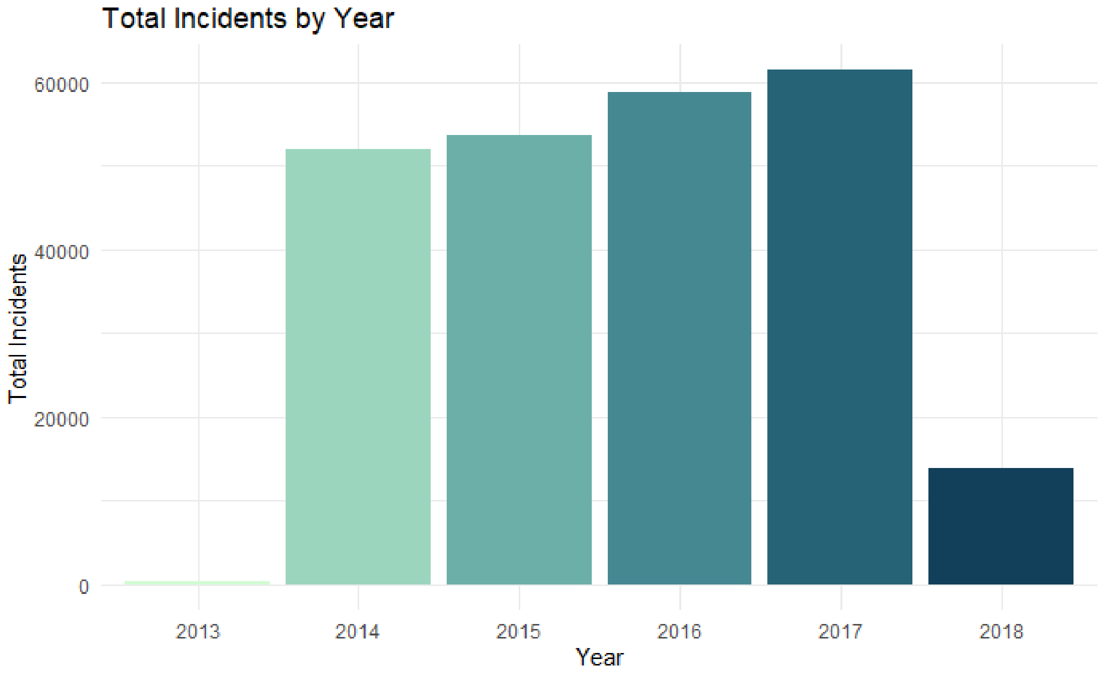


Figure 2

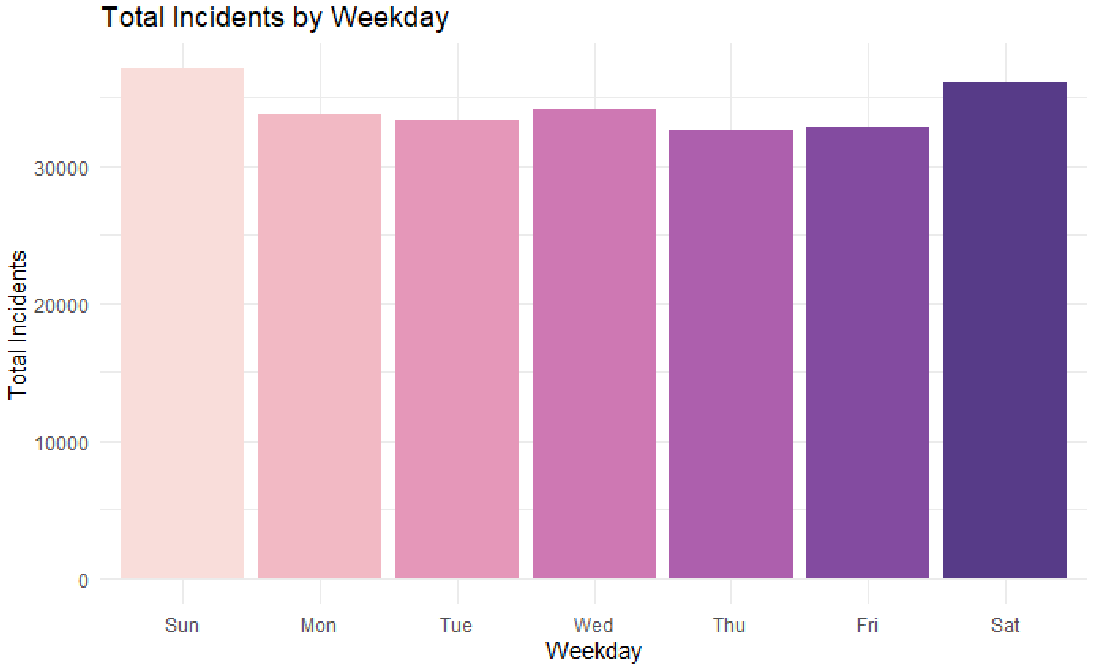


Figure 3

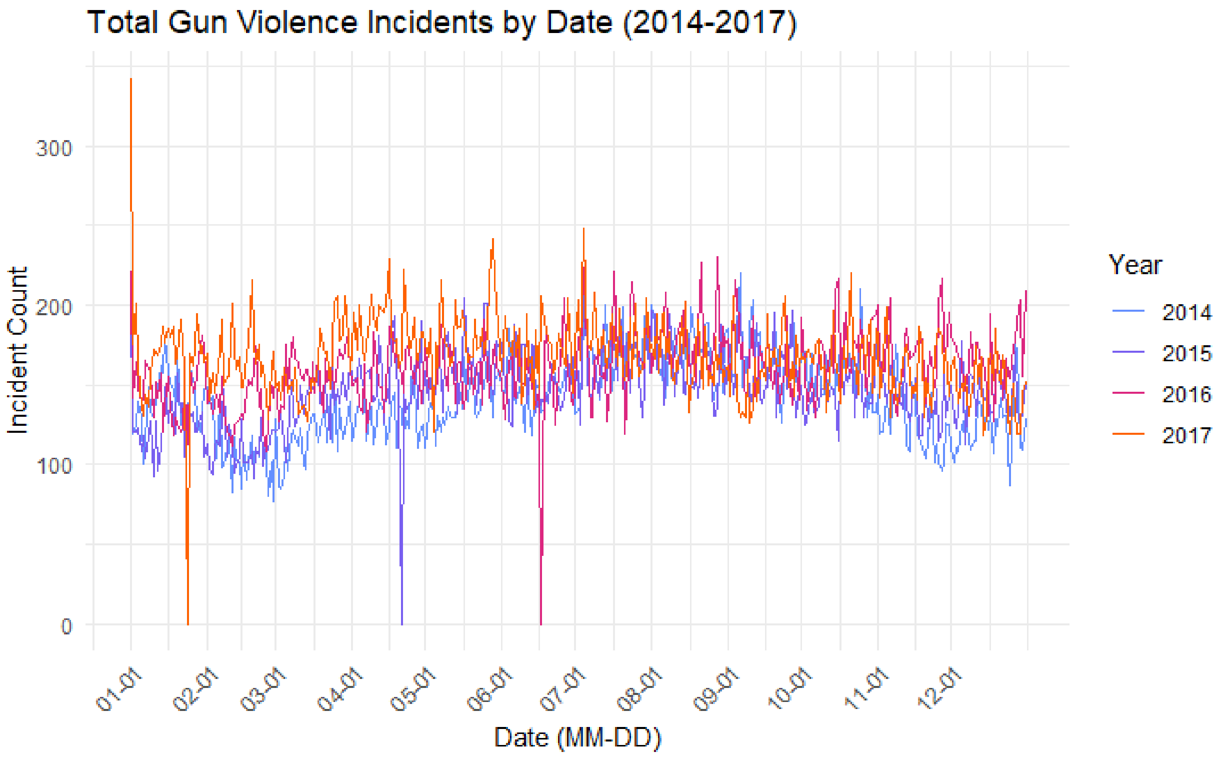


Figure 4

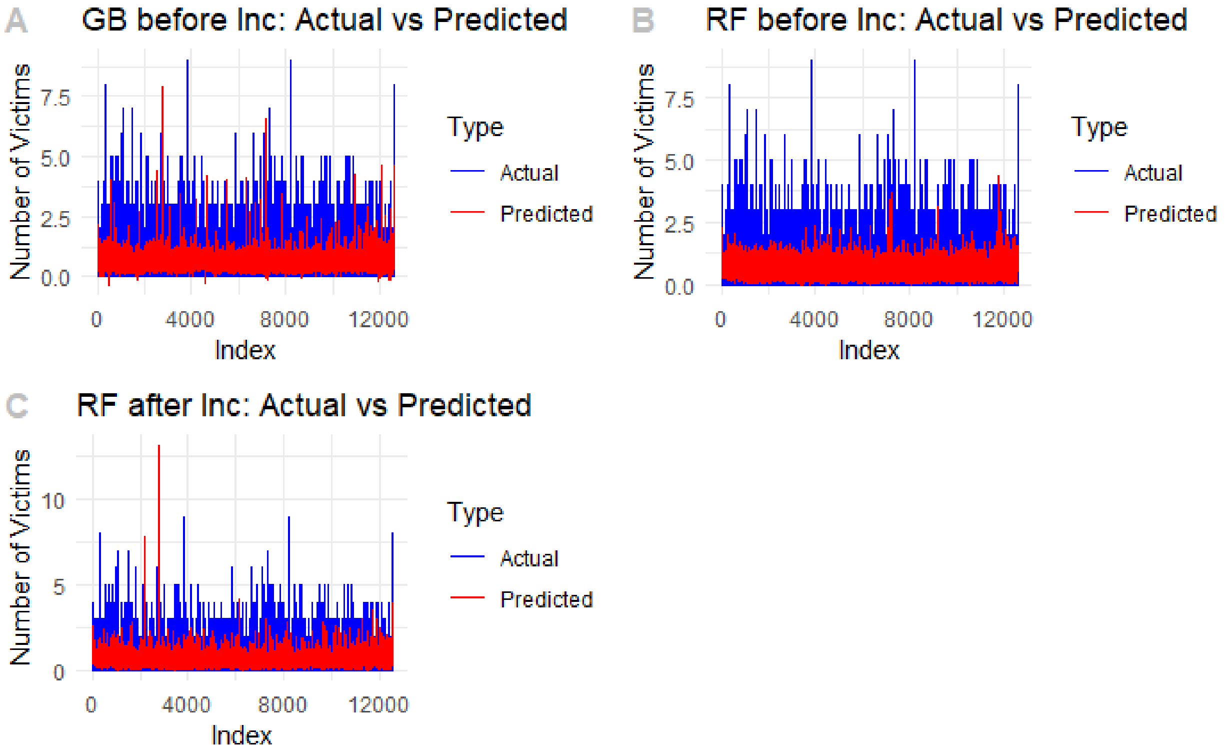


Figure 5

Appendix - Tables

Top 5 most dangerous states

State	Victims
Illinois	10767
California	9273
Florida	8613
Texas	7022
Ohio	6347

Table 1

Top 5 most dangerous locations

Locations	Frequency
Walmart	351
Austin	240
7-Eleven	166
Motel 6	152
McDonald’s	151

Table 2

References

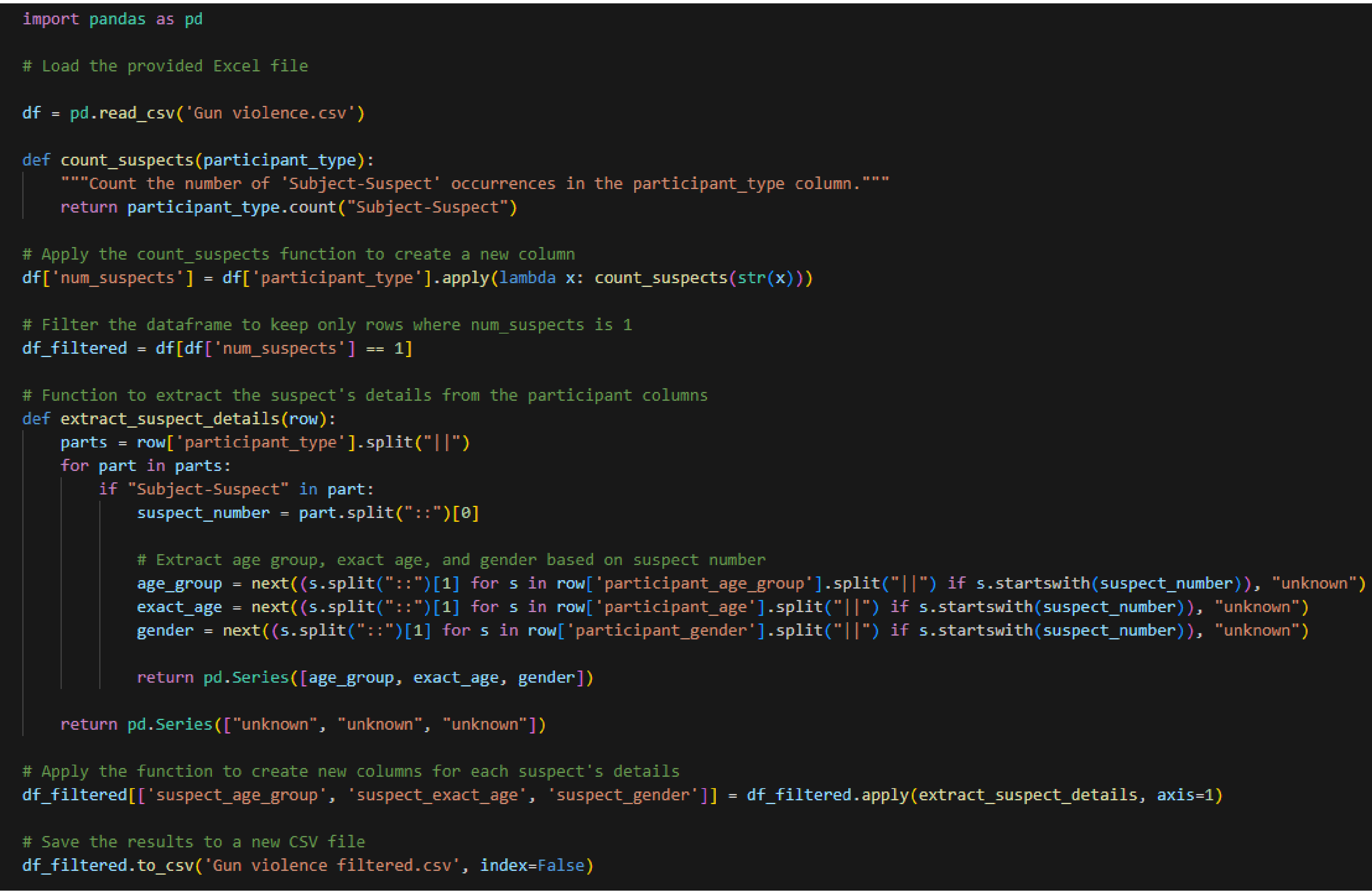
BBC. (2022, April 25). *Number of US police officers murdered up by 59% - FBI*. BBC News. https://www.bbc.com/news/world-us-canada-61218611

New Years Eve Gunfire Reduction Program. LAPD Online. (2021, May 22). https://www.lapdonline.org/new-years-eve-gunfire-reduction-program/

R Code

Please refer to the *MGSC661_FinalProject.R* file submitted along this report.

Python Code for Data Extraction



```python
import pandas as pd

# Load the provided Excel file

df = pd.read_csv('Gun violence.csv')

def count_suspects(participant_type):
    """Count the number of 'Subject-Suspect' occurrences in the participant_type column."""
    return participant_type.count("Subject-Suspect")

# Apply the count_suspects function to create a new column
df['num_suspects'] = df['participant_type'].apply(lambda x: count_suspects(str(x)))

# Filter the dataframe to keep only rows where num_suspects is 1
df_filtered = df[df['num_suspects'] == 1]

# Function to extract the suspect's details from the participant columns
def extract_suspect_details(row):
    parts = row['participant_type'].split("||")
    for part in parts:
        if "Subject-Suspect" in part:
            suspect_number = part.split("::")[0]

            # Extract age group, exact age, and gender based on suspect number
            age_group = next((s.split("::")[1] for s in row['participant_age_group'].split("||") if s.startswith(suspect_number)), "unknown")
            exact_age = next((s.split("::")[1] for s in row['participant_age'].split("||") if s.startswith(suspect_number)), "unknown")
            gender = next((s.split("::")[1] for s in row['participant_gender'].split("||") if s.startswith(suspect_number)), "unknown")

            return pd.Series([age_group, exact_age, gender])

    return pd.Series(["unknown", "unknown", "unknown"])

# Apply the function to create new columns for each suspect's details
df_filtered[['suspect_age_group', 'suspect_exact_age', 'suspect_gender']] = df_filtered.apply(extract_suspect_details, axis=1)

# Save the results to a new CSV file
df_filtered.to_csv('Gun violence filtered.csv', index=False)
```