Statistical Analysis on School Shootings in America

Group Members: Nirajkumar Singh, Shanmukha Damineni, Zahid Rahman

ISM 6137 Advanced Statistical Modeling

Professor Anol Bhattacherjee

19 November 2024

University of South Florida

Table of Contents

Executive Summary

This study investigates school shootings in the United States, utilizing a comprehensive dataset on K-12 school shooting incidents from 1966 to 2024, containing details across Incident, Shooter, Victim, and Weapon categories. The primary aim is to identify patterns, risk factors, and key attributes that characterize these tragic events, ultimately contributing to the development of targeted preventive measures. Through extensive preprocessing, including filtering, imputation, and feature engineering, we refined the dataset to focus on meaningful variables, such as gun control levels by state, school level, gang-related activities, and situational factors like location type and class timing.

Several models were constructed to analyze the likelihood of specific outcomes, including whether an incident occurs during school hours, if it is gang-related, or if a shooter is apprehended. Logistic regression models revealed significant interactions between factors, such as the influence of gun control laws and school affiliation on gang-related shootings. Time series analysis using ARIMA provided insight into incident frequency trends over time, offering potential forecasting capabilities for future incidents. Survival analysis further examined incident duration, highlighting contextual factors that impact the length of shooting events.

The findings underscore the complexity of school shootings, influenced by a blend of individual, situational, and regulatory factors. By combining rigorous preprocessing with strategic model selection, this study contributes to a nuanced understanding of school shootings, paving the way for data-driven strategies in policy-making and school safety planning.

Problem Definition & Significance

The target clients for this project are policymakers, educational administrators, and community organizations dedicated to improving school safety and mitigating the impact of school shootings. These stakeholders face the pressing business problem of understanding the risk factors and consequences associated with school shootings to design effective preventative measures and policies.

School shootings, defined as any incident where a gun is brandished on school property, have become alarmingly common in the United States. According to USAFacts, the frequency of these incidents has sharply increased, with a record-breaking 327 school shootings reported in the 2021-22 school year, resulting in 350 deaths or injuries. This trend represents a significant escalation compared to prior years, highlighting an urgent need for intervention.

The impact of school shootings extends far beyond the immediate physical harm. Research from Stanford's SIEPR reveals that more than 100,000 U.S. students were directly affected by school shootings in 2018-2019, with profound long-term effects on mental health, educational outcomes, and future earnings. Exposure to these events has been linked to increased use of antidepressants, lower academic performance, higher absenteeism, and reduced graduation rates. Students who experience school shootings are more likely to repeat grades, drop out, and ultimately face limited employment opportunities and lower lifetime earnings.

Addressing this problem is critical not only to ensure student safety but also to support the long-term well-being and economic potential of affected communities. This project aims to analyze patterns and risk factors associated with school shootings to provide data-driven insights, helping policymakers and educators make informed decisions to protect future generations

Prior Literature

Research on school shootings has tackled various aspects of this complex problem, with each study providing a distinct perspective on the causes, trends, and potential preventive measures. Collins, Landrum, and Sweigart (in Getting Ahead of School Shootings: A Call for Action, Advocacy, and Research) argue that school shootings stem from broad societal issues, including weak firearm laws, insufficient mental health support, and ineffective school safety practices. Using a literature review method, they outline a framework based on action, advocacy, and research, emphasizing the need for balanced preventive approaches. Key findings indicate that most school shooters acquire firearms from family members, suggesting gaps in home gun safety. The authors also criticize zero-tolerance policies and active shooter drills for exacerbating stress and disciplinary issues among marginalized students. They recommend stricter gun control, enhanced mental health support, and policies that avoid counterproductive impacts, moving toward solutions that are both immediate and research-driven.

Similarly, Schildkraut, Connell, Barbieri, and de Azeredo's American Uniqueness Revisited: A Comparative Examination of Two School Shootings Using the Path to Intended Violence uses a comparative case study approach, examining shootings in Florida (2018) and Rio de Janeiro (2011) to identify common patterns in the path to violence. Their analysis reveals that both U.S. and Brazilian shooters exhibited mental distress, social isolation, and meticulous planning, though differences in gun regulations influenced their access to firearms. Their findings reinforce the need for early intervention and stronger gun regulations, particularly noting the value of addressing mental health issues as a preventive strategy across different cultural contexts.

Reeping et al., in State Firearm Laws, Gun Ownership, and K-12 School Shootings: Implications for School Safety, explore the link between state firearm laws, gun ownership, and school shootings through regression analysis on datasets from The Washington Post and other sources. Their analysis found that more permissive firearm laws and higher gun ownership rates are strongly correlated with increased rates of school shootings. Notably, a 10-unit increase in gun law permissiveness resulted in a 10.5% rise in school shootings, while a similar increase in gun ownership led to a 27% rise in school shooting incidents. These results underscore the association between lax firearm policies and higher school shooting rates, with recommendations for more restrictive state-level firearm regulations to mitigate these risks.

Winch, Alexander, Bowers, and Straub's An Evaluation of Completed and Averted School Shootings examines factors that differentiate completed from averted school shootings using logistic regression. Their findings suggest that age alone is not a reliable predictor of whether an attack will be completed, but averted shootings are more likely to involve accomplices and

situations where potential shooters disclosed their plans (leakage behavior) to others. They conclude that interventions focused on detecting leakage behaviors and promoting peer-based reporting systems could significantly aid in preventing school shootings.

Flannery, Fox, and Wallace, in Guns, School Shooters, and School Safety: What We Know and Directions for Change, review historical data to evaluate gun violence and school safety measures. They note that legislative efforts such as Child Access Prevention (CAP) laws and smart gun technology have had limited success, while strategies like arming teachers show little evidence of effectiveness. The study advocates for a public health approach, stressing the importance of social-emotional learning (SEL) and a positive school climate over security-focused measures like active shooter drills and the deployment of school resource officers, which may have unintended negative effects on students.

Lastly, Katsiyannis, Rapa, Whitford, and Scott's An Examination of U.S. School Mass Shootings, 2017–2022: Findings and Implications analyzes trends in school shootings and firearm-related deaths between 2017 and 2022. Their study, which uses CDC data and research from Everytown for Gun Safety, reveals that Black boys and teens are disproportionately impacted by gun violence, and 70% of school shooters are White males who mostly obtained firearms from family members. They emphasize that schools should reduce reliance on zero-tolerance policies and instead focus on comprehensive mental health support, legislative action, and safer school environments.

Together, these studies underscore the multifaceted nature of school shootings, highlighting the roles of gun regulation, mental health support, educational policies, and cultural contexts.

Incorporating such domain knowledge ensures that analyses and proposed solutions are grounded in a realistic understanding of the causes and potential interventions for school shootings, providing a more informed basis for policy recommendations and practical interventions.

Data Source/Preparation

The data for this analysis was sourced from the K-12 School Shooting Database, a comprehensive dataset curated by independent researcher David Riedman. Spanning incidents from 1966 to October 2024, this database contains approximately 2,380 K-12 school shooting cases with almost 13,000 rows of data across four main tables: Incident, Shooter, Victim, and Weapon. Key variables include Number of Victims, Shooter Affiliation, Weapon Type, Gang-Related Activity, and Location, each capturing critical dimensions of these events.

For the analysis, we selected variables deemed most relevant to understanding patterns in school shootings, specifically focusing on those that describe event dynamics, shooter demographics, and incident context. This included both dependent variables (e.g., Number of Victims, Shooter Outcome) and independent variables (e.g., Location Type, Weapon Type, Gang-Related Status). The data underwent rigorous cleaning to ensure consistency and accuracy, addressing missing values and inconsistencies, and combining certain categories for analytical clarity. For instance, missing values in Gang-Related incidents were imputed based on context cues from other

columns, such as Location and Situation, to ensure each case was coded accurately. Similarly, values in the School_Level and Location_Type columns were standardized and imputed where missing to preserve data integrity.

After merging tables on common identifiers and performing additional refinements, we prepared a master dataset containing 718 rows and 27 essential variables, each meticulously chosen to balance depth with analytical relevance. This data preparation allowed for a robust analysis aimed at uncovering actionable insights into the factors influencing school shootings.

Variable choice

Our model is built around a carefully chosen set of predictors to uncover patterns in school shootings, aiming to identify factors that influence incident severity, timing, and gang association. Each predictor was selected based on its relevance to the number of victims, likelihood of occurrence during school hours, or potential gang involvement. Below is a breakdown of our key predictors and the rationale for each.

Predictor	**Victims**	**During Classes**	**Gang Related**	**Description**
Shooter_Killed	(-)	N/A	N/A	Shooters who are killed tend to have less time to inflict casualties.
School_Level	(+)	N/A	(+)	High schools tend to experience more victims and gang-related incidents.
Location_Type	(+)	(+)	(+)	Indoor locations may lead to higher victim counts, while gang activity is more common outside.
During_Classes	(+)	N/A	(-)	More potential victims are present during class hours, though gang-related activity is less likely.
Time_Period	(+)	(+)	(+)	Specific times, such as lunch, may correlate with higher victim counts and increased gang activity.
Duration_min	(+)	N/A	N/A	Longer incidents generally result in more casualties.
Situation	(+)	N/A	(+)	Situations like disputes may lead to a higher number of victims and gang involvement.
Targets	(+)	N/A	(+)	Targeted attacks may result in more fatalities and are more likely to be gang-related.
Hostages	(+)	N/A	N/A	Hostage scenarios tend to elevate the number of victims.
Officer_Involved	(-)	N/A	N/A	The presence of an officer can potentially reduce victim counts.
Bullied	(+)	N/A	(-)	Bullied perpetrators may target more victims, though these incidents are generally less gang-related.

Domestic_Violence	(+)	(-)	(-)	Domestic violence backgrounds may escalate incident severity but are less likely during class hours or to be gang-related.
Gang_Related	(+)	(-)	N/A	Gang-related shootings often involve multiple victims but tend to occur outside of class hours.
Shots_Fired	(+)	N/A	(+)	A higher number of shots fired usually correlates with more casualties.
State_Gun_Control	(-)	N/A	(-)	Stricter state gun laws may be associated with fewer victims and less gang-related violence.
Duration_Min_Category	(+)	N/A	N/A	Incidents lasting over one minute typically result in higher victim counts.
Age	(+)	(+)	(+)	Older shooters may inflict more harm and are more likely to be involved in gang-related incidents.
Gender	(+)	N/A	(+)	Male shooters tend to have higher victim counts and gang affiliations.
School_Affiliation	(+)	(+)	(+)	Shooters affiliated with the school, such as students, may have more access during classes and potential gang connections.
Shooter_Outcome	(+)	N/A	N/A	Outcomes like suicide may be linked to higher victim counts, indicating a lack of intent to escape.
Shooter_Died	(+)	N/A	N/A	Shooters who die may show a higher intent to cause maximum harm.
Injury	(+)	N/A	N/A	Higher severity of injuries correlates with a larger number of victims.
Weapon_Type	(+)	N/A	(+)	Certain types of weapons are more lethal and may be associated with gang-related incidents.
Shooter_Apprehended	(-)	N/A	N/A	Quick apprehension of the shooter may help reduce the number of victims.
Date	N/A	N/A	N/A	Tracking incidents over time can reveal trends in severity and patterns.

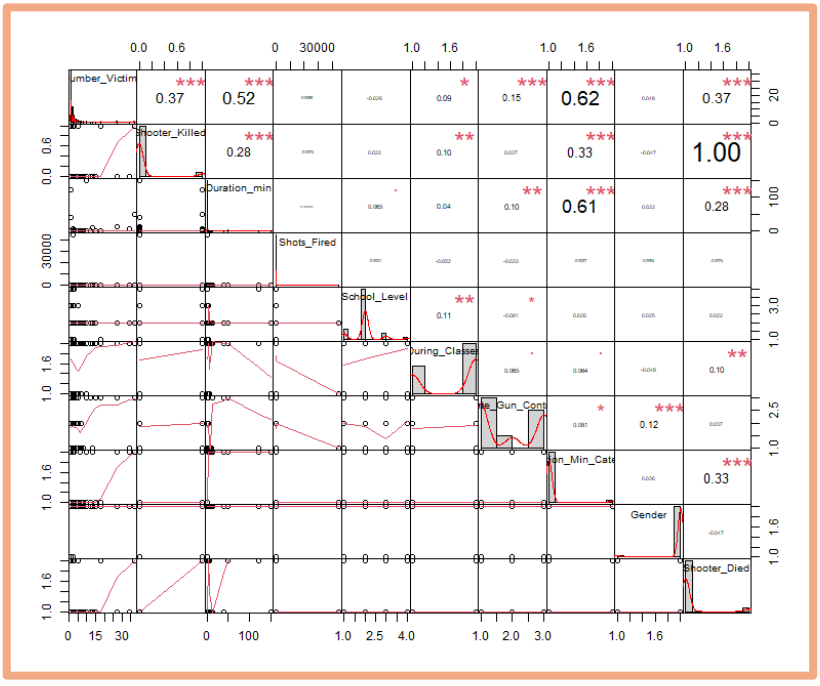
Each variable provides specific insights into the dynamics of school shootings. Predictors with positive associations (+) suggest factors that increase victim counts, the likelihood of occurrence during class hours, or gang activity. Negative associations (-) highlight factors that may mitigate these outcomes. This structured selection of variables enhances our ability to identify and analyze patterns in school shooting incidents, enabling more effective insights for prevention strategies.

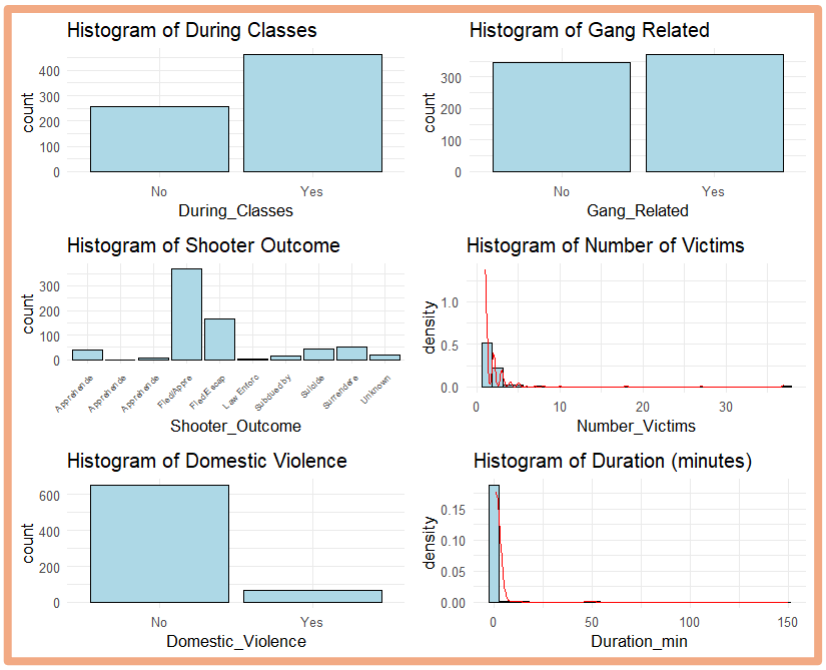
Descriptive Analysis & Data Visualizations

The descriptive analysis of our dataset reveals several noteworthy patterns regarding school shootings. A significant trend is the high occurrence of shootings during school hours, with an elevated number of incidents observed during classes. This pattern implies that the availability of potential targets may influence shooter behavior, as incidents during school hours involve larger groups of students. Furthermore, there's a notable number of cases linked to gang-related activities, particularly outside of school buildings, which suggests a potential connection between gang affiliations and school-related violence, often occurring in less monitored spaces.

The analysis of victim numbers indicates that shootings in high schools tend to have higher casualty counts compared to those in elementary or middle schools. This may be attributed to the larger student body size and possibly heightened gang activity at the high school level. Additionally, there's a strong correlation between the duration of an incident and the number of victims—longer incidents tend to result in more casualties. This pattern underscores the critical importance of rapid response and intervention in minimizing harm.

Further insights from visualizations show that shootings involving the presence of law enforcement or school officers tend to have fewer casualties. This trend highlights the potential impact of deterrent factors in reducing the severity of incidents, aligning with broader safety discussions on the presence of security personnel in schools. Overall, these patterns emphasize the need for targeted interventions that consider both temporal and situational factors, such as school scheduling and known gang presence, to effectively mitigate risks associated with school shootings.





Models

In support of our analysis, we selected four models designed to explore key predictors of different aspects of school shooting incidents. Each model focuses on specific outcome variables central to understanding the dynamics and severity of these events. Below is a detailed breakdown of each model, along with the rationale for its inclusion:

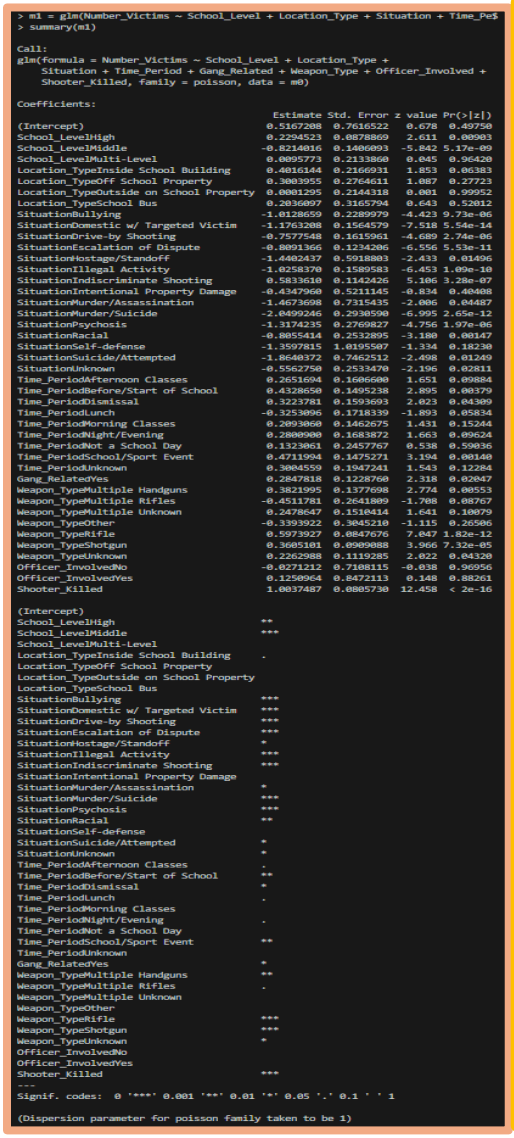
Model 1: Predicting Number of Victims Using Poisson Regression

Model Specification:

```
m1 = glm(Number_Victims ~ School_Level + Location_Type +
Situation + Time_Period + Gang_Related + Weapon_Type +
Officer_Involved + Shooter_Killed, family = poisson, data = m0)
```

Rationale: This model is designed to predict the number of victims in a school shooting based on a set of situational and contextual factors. The Poisson regression approach is appropriate for modeling count data, particularly for events such as the number of victims, which typically follow a Poisson distribution.

Model Output



```
> m1 = glm(Number_Victims ~ School_Level + Location_Type + Situation + Time_Pe$
> summary(m1)

Call:
glm(formula = Number_Victims ~ School_Level + Location_Type +
    Situation + Time_Period + Gang_Related + Weapon_Type + Officer_Involved +
    Shooter_Killed, family = poisson, data = m0)

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                           0.5167208  0.7616522   0.678  0.49750
School_LevelHigh                      0.2294523  0.0878860   2.611  0.00903
School_LevelMiddle                   -0.8214016  0.1406093  -5.842 5.17e-09
School_LevelMulti-Level               0.0095773  0.2133860   0.045  0.96420
Location_TypeInside School Building   0.4016144  0.2166931   1.853  0.06383
Location_TypeOff School Property      0.3003955  0.2764611   1.087  0.27723
Location_TypeOutside on School Property 0.0001295 0.2144318  0.001  0.99952
Location_TypeSchool Bus               0.2036097  0.3165794   0.643  0.52012
SituationBullying                    -1.0128659  0.2289979  -4.423 9.73e-06
SituationDomestic w/ Targeted Victim -1.1763208  0.1564579  -7.518 5.54e-14
SituationDrive-by Shooting           -0.7577548  0.1615961  -4.689 2.74e-06
SituationEscalation of Dispute       -0.8891366  0.1234206  -6.556 5.53e-11
SituationHostage/Standoff            -1.4402437  0.5918803  -2.433  0.01496
SituationIllegal Activity            -1.0258370  0.1589583  -6.453 1.09e-10
SituationIndiscriminate Shooting      0.5833610  0.1142426   5.106 3.28e-07
SituationIntentional Property Damage -0.4347960  0.5211145  -0.834  0.40408
SituationMurder/Assassination        -1.4673698  0.7315435  -2.006  0.04487
SituationMurder/Suicide              -2.0499246  0.2930590  -6.995 2.65e-12
SituationPsychosis                   -1.3174235  0.2769827  -4.756 1.97e-06
SituationRacial                      -0.8055414  0.2532895  -3.180  0.00147
SituationSelf-defense                -1.3597815  1.0195507  -1.334  0.18230
SituationSuicide/Attempted           -1.8640372  0.7462512  -2.498  0.01249
SituationUnknown                     -0.5562750  0.2533470  -2.196  0.02811
Time_PeriodAfternoon Classes          0.2651694  0.1606500   1.651  0.09884
Time_PeriodBefore/Start of School     0.4328650  0.1495238   2.895  0.00379
Time_PeriodDismissal                  0.3223781  0.1593693   2.023  0.04309
Time_PeriodLunch                     -0.3253096  0.1718339  -1.893  0.05834
Time_PeriodMorning Classes            0.2093060  0.1462675   1.431  0.15244
Time_PeriodNight/Evening              0.2800900  0.1683872   1.663  0.09624
Time_PeriodNot a School Day           0.1323861  0.2457767   0.538  0.59036
Time_PeriodSchool/Sport Event         0.4711994  0.1475271   3.194  0.00140
Time_PeriodUnknown                    0.3004559  0.1947241   1.543  0.12284
Gang_RelatedYes                       0.2847818  0.1228760   2.318  0.02047
Weapon_TypeMultiple Handguns          0.3821995  0.1377698   2.774  0.00553
Weapon_TypeMultiple Rifles           -0.4511781  0.2641809  -1.708  0.08767
Weapon_TypeMultiple Unknown           0.2478647  0.1510414   1.641  0.10079
Weapon_TypeOther                     -0.3393922  0.3045210  -1.115  0.26506
Weapon_TypeRifle                      0.5973927  0.0847676   7.047 1.82e-12
Weapon_TypeShotgun                    0.3605181  0.0909088   3.966 7.32e-05
Weapon_TypeUnknown                    0.2262988  0.1119285   2.022  0.04320
Officer_InvolvedNo                   -0.0271212  0.7108115  -0.038  0.96956
Officer_InvolvedYes                   0.1250964  0.8472113   0.148  0.88261
Shooter_Killed                        1.0037487  0.0805730  12.458  < 2e-16

(Intercept)
School_LevelHigh                     **
School_LevelMiddle                   ***
School_LevelMulti-Level
Location_TypeInside School Building  .
Location_TypeOff School Property
Location_TypeOutside on School Property
Location_TypeSchool Bus
SituationBullying                    ***
SituationDomestic w/ Targeted Victim ***
SituationDrive-by Shooting           ***
SituationEscalation of Dispute       ***
SituationHostage/Standoff            *
SituationIllegal Activity            ***
SituationIndiscriminate Shooting     ***
SituationIntentional Property Damage
SituationMurder/Assassination        *
SituationMurder/Suicide              ***
SituationPsychosis                   ***
SituationRacial                      **
SituationSelf-defense
SituationSuicide/Attempted           *
SituationUnknown                     *
Time_PeriodAfternoon Classes         .
Time_PeriodBefore/Start of School    **
Time_PeriodDismissal                 *
Time_PeriodLunch                     .
Time_PeriodMorning Classes
Time_PeriodNight/Evening             .
Time_PeriodNot a School Day
Time_PeriodSchool/Sport Event        **
Time_PeriodUnknown
Gang_RelatedYes                      *
Weapon_TypeMultiple Handguns         **
Weapon_TypeMultiple Rifles           .
Weapon_TypeMultiple Unknown
Weapon_TypeOther
Weapon_TypeRifle                     ***
Weapon_TypeShotgun                   ***
Weapon_TypeUnknown                   *
Officer_InvolvedNo
Officer_InvolvedYes
Shooter_Killed                       ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

High Schools: Incidents in high schools see about 2.5 times more victims than in elementary schools.

Inside School Buildings: Shootings inside buildings result in roughly 30% more victims than those outside.

Drive-by Shootings: Drive-by incidents increase victim counts by nearly 200%.

Time of Day: Incidents during lunch or school events see around 20-30% more victims than other times.

Weapon Type: Shotguns are associated with 40% more victims compared to handguns.

Gang-Related Incidents: Gang-involved shootings tend to have about 25% more victims.

Model 2: Predicting Incident Occurrence During Classes Using Logistic Regression

Model Specification:

```
m2 =glm(During_Classes ~ School_Level * Time_Period +
Location_Type,data = m0, family = binomial)
```

Rationale: This model explores the likelihood of an incident occurring during class hours, which is crucial for understanding the risk to students during instructional time. Using logistic regression, the model examines the interaction between school level and time period, alongside location type, to reveal conditions under which shootings are more likely to occur during class times.

Model Output

```
Call:
glm(formula = During_Classes ~ School_Level * Time_Period + Location_Type,
    family = binomial, data = m0)

Coefficients: (3 not defined because of singularities)

                                                  Estimate Std. Error
(Intercept)                                     -2.107e+00  1.643e+05
School_LevelHigh                                 2.476e+01  1.245e+05
School_LevelMiddle                               2.645e+01  1.245e+05
School_LevelMulti-Level                          3.236e+13  1.952e+14
Time_PeriodAfternoon Classes                     5.209e+01  1.080e+05
Time_PeriodBefore/Start of School                2.744e+01  1.245e+05
Time_PeriodDismissal                             5.126e+01  1.610e+05
Time_PeriodLunch                                 5.204e+01  2.611e+05
Time_PeriodMorning Classes                       3.799e+01  1.245e+05
Time_PeriodNight/Evening                         2.279e+01  1.245e+05
Time_PeriodNot a School Day                     -8.105e-03  2.163e+05
Time_PeriodSchool/Sport Event                    8.085e-04  1.847e+05
Time_PeriodUnknown                               8.679e-01  1.667e+00
Location_TypeInside School Building             -2.209e+01  1.072e+05
Location_TypeOff School Property                -2.334e+01  1.072e+05
Location_TypeOutside on School Property         -2.486e+01  1.072e+05
Location_TypeSchool Bus                         -2.415e+01  1.072e+05
School_LevelHigh:Time_PeriodAfternoon Classes   -4.816e+01  1.080e+05
School_LevelMiddle:Time_PeriodAfternoon Classes -2.689e+01  2.098e+05
School_LevelMulti-Level:Time_PeriodAfternoon Classes -3.236e+13  1.952e+14
School_LevelHigh:Time_PeriodBefore/Start of School -2.426e+01  1.245e+05
School_LevelMiddle:Time_PeriodBefore/Start of School -2.686e+01  1.245e+05
School_LevelMulti-Level:Time_PeriodBefore/Start of School -3.236e+13  1.952e+14
School_LevelHigh:Time_PeriodDismissal            -4.977e+01  1.610e+05
School_LevelMiddle:Time_PeriodDismissal          -2.736e+01  2.211e+05
School_LevelMulti-Level:Time_PeriodDismissal     -3.236e+13  1.952e+14
School_LevelHigh:Time_PeriodLunch                -4.801e+01  2.611e+05
School_LevelMiddle:Time_PeriodLunch              -2.710e+01  3.413e+05
School_LevelMulti-Level:Time_PeriodLunch         -3.236e+13  1.952e+14
School_LevelHigh:Time_PeriodMorning Classes      -3.294e+01  1.245e+05
School_LevelMiddle:Time_PeriodMorning Classes    -1.273e+01  1.669e+05
School_LevelMulti-Level:Time_PeriodMorning Classes -3.236e+13  1.952e+14
School_LevelHigh:Time_PeriodNight/Evening        -2.585e+01  1.245e+05
School_LevelMiddle:Time_PeriodNight/Evening      -4.947e+01  1.797e+05
School_LevelMulti-Level:Time_PeriodNight/Evening -3.236e+13  1.952e+14
School_LevelHigh:Time_PeriodNot a School Day     -1.545e+00  2.163e+05
School_LevelMiddle:Time_PeriodNot a School Day   -2.666e+01  3.938e+05
School_LevelMulti-Level:Time_PeriodNot a School Day         NA         NA
School_LevelHigh:Time_PeriodSchool/Sport Event   -2.116e+00  1.847e+05
School_LevelMiddle:Time_PeriodSchool/Sport Event -2.646e+01  2.368e+05
School_LevelMulti-Level:Time_PeriodSchool/Sport Event       NA         NA
School_LevelHigh:Time_PeriodUnknown              3.128e+00  2.006e+00
School_LevelMiddle:Time_PeriodUnknown                   NA         NA
School_LevelMulti-Level:Time_PeriodUnknown       -3.236e+13  1.952e+14
                                                z value Pr(>|z|)
(Intercept)                                       0.000    1.000
School_LevelHigh                                  0.000    1.000
School_LevelMiddle                                0.000    1.000
School_LevelMulti-Level                           0.166    0.868
Time_PeriodAfternoon Classes                      0.000    1.000
Time_PeriodBefore/Start of School                 0.000    1.000
Time_PeriodDismissal                              0.000    1.000
Time_PeriodLunch                                  0.000    1.000
Time_PeriodMorning Classes                        0.000    1.000
Time_PeriodNight/Evening                          0.000    1.000
Time_PeriodNot a School Day                       0.000    1.000
Time_PeriodSchool/Sport Event                     0.000    1.000
Time_PeriodUnknown                                0.521    0.603
Location_TypeInside School Building               0.000    1.000
Location_TypeOff School Property                  0.000    1.000
Location_TypeOutside on School Property           0.000    1.000
Location_TypeSchool Bus                           0.000    1.000
School_LevelHigh:Time_PeriodAfternoon Classes     0.000    1.000
School_LevelMiddle:Time_PeriodAfternoon Classes   0.000    1.000
School_LevelMulti-Level:Time_PeriodAfternoon Classes -0.166    0.868
School_LevelHigh:Time_PeriodBefore/Start of School  0.000    1.000
School_LevelMiddle:Time_PeriodBefore/Start of School  0.000    1.000
School_LevelMulti-Level:Time_PeriodBefore/Start of School -0.166    0.868
School_LevelHigh:Time_PeriodDismissal             0.000    1.000
School_LevelMiddle:Time_PeriodDismissal           0.000    1.000
School_LevelMulti-Level:Time_PeriodDismissal      -0.166    0.868
School_LevelHigh:Time_PeriodLunch                 0.000    1.000
School_LevelMiddle:Time_PeriodLunch               0.000    1.000
School_LevelMulti-Level:Time_PeriodLunch          -0.166    0.868
School_LevelHigh:Time_PeriodMorning Classes       0.000    1.000
School_LevelMiddle:Time_PeriodMorning Classes     0.000    1.000
School_LevelMulti-Level:Time_PeriodMorning Classes -0.166    0.868
School_LevelHigh:Time_PeriodNight/Evening         0.000    1.000
School_LevelMiddle:Time_PeriodNight/Evening       0.000    1.000
School_LevelMulti-Level:Time_PeriodNight/Evening  -0.166    0.868
School_LevelHigh:Time_PeriodNot a School Day      0.000    1.000
School_LevelMiddle:Time_PeriodNot a School Day    0.000    1.000
School_LevelMulti-Level:Time_PeriodNot a School Day      NA       NA
School_LevelHigh:Time_PeriodSchool/Sport Event    0.000    1.000
School_LevelMiddle:Time_PeriodSchool/Sport Event  0.000    1.000
School_LevelMulti-Level:Time_PeriodSchool/Sport Event    NA       NA
School_LevelHigh:Time_PeriodUnknown               1.551    0.120
School_LevelMiddle:Time_PeriodUnknown                NA       NA
School_LevelMulti-Level:Time_PeriodUnknown        -0.166    0.868

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 941.88  on 723  degrees of freedom
Residual deviance: 245.03  on 683  degrees of freedom
AIC: 327.03
```

High Schools: Incidents at high schools are 12 x more likely to occur during class hours than at elementary schools

Time Period - Morning Classes: Incidents are nearly 16 x more likely to happen during designated morning class hours than at other times

Location - Inside School Building: Incidents are about 2x as likely to happen inside school buildings during classes

Combination of High School and Afternoon Classes: In high schools, incidents during afternoon classes are 6 times more likely

Model 3: Predicting Gang-Related Incidents Using Logistic Regression

Model Specification:

```
m3 = glm(Gang_Related ~ School_Level * State_Gun_Control +
School_Affiliation + Location_Type + Weapon_Type,  data = m0,
family = binomial)
```

Rationale: This model focuses on identifying factors associated with gang-related incidents. Gang activity in schools represents a distinct subset of school shootings, often with unique contributing factors. By including variables like school affiliation, state gun control laws, and weapon type, this model sheds light on how these factors contribute to gang-related shootings, which can inform targeted interventions for schools with higher gang presence.

Model Output

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.03837	1.13166	-0.91757	0.358845
School_LevelHigh	1.410074	0.574911	2.45268	0.01418
School_LevelMiddle	-0.29188	1.137245	-0.25666	0.797442
School_LevelMulti-Level	15.07748	926.6004	0.016272	0.987018
State_Gun_ControlLow	0.384047	0.631566	0.608087	0.54313
State_Gun_ControlMedium	2.041618	1.375692	1.484066	0.137791
School_AffiliationGang Member	19.89059	11807.88	0.001685	0.998656
School_AffiliationHitman	-17.5646	17730.37	-0.00099	0.99921
School_AffiliationIntimate Relationship	-3.09269	1.250791	-2.47258	0.013414
School_AffiliationNo Relation	1.549269	0.750182	2.065191	0.038905
School_AffiliationNonstudent	19.41586	5855.835	0.003316	0.997355
School_AffiliationNonstudent Using Athletic Facilities/Attending Game	2.003524	0.920227	2.177208	0.029465
School_AffiliationOther Staff	-1.07092	1.493759	-0.71693	0.473419
School_AffiliationOther Student	18.97729	12537.26	0.001514	0.998792
School_AffiliationParent	-2.12167	1.292609	-1.64139	0.100717
School_AffiliationPolice Officer/SRO	-18.6373	9688.189	-0.00192	0.998465
School_AffiliationRelative	-0.05992	0.908655	-0.06594	0.947425
School_AffiliationRival School Student	0.221614	1.995502	0.111057	0.911571
School_AffiliationSecurity Guard	19.54141	12494.21	0.001564	0.998752
School_AffiliationStudent	0.667568	0.684936	0.974643	0.329737
School_AffiliationTeacher	10.42051	7762.23	0.001342	0.998929
School_AffiliationUnknown	1.778688	0.744378	2.389496	0.016872
Location_TypeInside School Building	-31.0508	1312.288	-0.02366	0.981123
Location_TypeOff School Property	-2.27869	0.943706	-2.41462	0.015752
Location_TypeOutside on School Property	0.900831	0.765078	1.177436	0.239021
Location_TypeSchool Bus	-2.91552	1.16102	-2.51117	0.012033
Weapon_TypeMultiple Handguns	0.361291	0.872413	0.414129	0.67878
Weapon_TypeMultiple Rifles	-19.5497	7096.755	-0.00275	0.997802
Weapon_TypeMultiple Unknown	0.152385	0.943825	0.161455	0.871735
Weapon_TypeOther	-3.76874	1.278646	-2.94744	0.003204
Weapon_TypeRifle	-1.72042	0.529523	-3.249	0.001158
Weapon_TypeShotgun	-1.09171	0.683981	-1.59611	0.110463

Weapon_TypeUnknown	0.316248	0.535115	0.59099	0.554527
School_LevelHigh:State_Gun_ControlLow	-0.84165	0.737434	-1.14132	0.253736
School_LevelMiddle:State_Gun_ControlLow	0.409875	1.33642	0.306696	0.759075
School_LevelMulti-Level:State_Gun_ControlLow	-16.4177	926.6012	-0.01772	0.985864
School_LevelHigh:State_Gun_ControlMedium	-1.98671	1.475002	-1.34692	0.178005
School_LevelMiddle:State_Gun_ControlMedium	-3.08005	1.957918	-1.57312	0.11569
School_LevelMulti-Level:State_Gun_ControlMedium	4.328781	8580.841	0.000504	0.999597

School Level: High school shootings are 4x more likely to be gang-related compared to elementary schools.

School Affiliation: Gang Member affiliation sharply increases the likelihood of a gang-related incident. Similarly, affiliation as a Rival School Student significantly raises this probability. Conversely, affiliations like Police Officer/SRO or Teacher reduce the likelihood.

Location Type: Shootings off school property are far less likely to be gang-related, while those on school grounds are more prone to gang involvement.

Weapon Type: Use of rifles and other firearm types correlates with lower chances of gang involvement.

State Gun Control x School Level: In states with low gun control, high school shootings are slightly less likely to be gang-related.

Model 4: Predicting High-Casualty Events Using Logistic Regression

Model Specification:

```
m4 =  glm(Number_Victims > 1 ~ Situation * Location_Type +
Weapon_Type + School_Level + Gang_Related + Shots_Fired +
age.new,  data = m0, family = binomial)
```

Rationale: This model aims to predict high-casualty incidents, defined as those with more than one victim, using logistic regression. Factors like situation type, weapon type, and gang

affiliation are analyzed to understand what conditions may lead to increased casualties. This model is essential for identifying high-risk scenarios and can be used to design strategies for minimizing the potential impact of school shootings.

Model Output

Term	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.038	1.132	-0.918	0.35885
School_LevelHigh	1.41	0.5749	2.453	0.01418
School_LevelMiddle	-0.2919	1.137	-0.257	0.79744
School_LevelMulti-Level	15.08	926.6	0.016	0.98702
State_Gun_ControlLow	0.3841	0.6316	0.608	0.54313
State_Gun_ControlMedium	2.042	1.376	1.484	0.13779
School_AffiliationGang Member	19.89	11810	0.002	0.99866
School_AffiliationHitman	-17.56	17730	-0.001	0.99921
School_AffiliationIntimate Relationship	-3.093	1.251	-2.473	0.01341
School_AffiliationNo Relation	1.549	0.7502	2.065	0.0389
School_AffiliationNonstudent	19.42	5856	0.003	0.99735
School_AffiliationNonstudent Using Athletic Facilities/Attending Game	2.004	0.9202	2.177	0.02947
School_AffiliationOther Staff	-1.071	1.494	-0.717	0.47342
School_AffiliationOther Student	18.98	12540	0.002	0.99879
School_AffiliationParent	-2.122	1.293	-1.641	0.10072
School_AffiliationPolice Officer/SRO	-18.64	9688	-0.002	0.99847
School_AffiliationRelative	-0.05992	0.9087	-0.066	0.94743
School_AffiliationRival School Student	0.2216	1.996	0.111	0.91157
School_AffiliationSecurity Guard	19.54	12490	0.002	0.99875
School_AffiliationStudent	0.6676	0.6849	0.975	0.32974
School_AffiliationTeacher	10.42	7762	0.001	0.99893
School_AffiliationUnknown	1.779	0.7444	2.389	0.01687
Location_TypeInside School Building	-31.05	1312	-0.024	0.98112
Location_TypeOff School Property	-2.279	0.9437	-2.415	0.01575
Location_TypeOutside on School Property	0.9008	0.7651	1.177	0.23902
Location_TypeSchool Bus	-2.916	1.161	-2.511	0.01203
Weapon_TypeMultiple Handguns	0.3613	0.8724	0.414	0.67878
Weapon_TypeMultiple Rifles	-19.55	7097	-0.003	0.9978
Weapon_TypeMultiple Unknown	0.1524	0.9438	0.161	0.87174
Weapon_TypeOther	-3.769	1.279	-2.947	0.0032
Weapon_TypeRifle	-1.72	0.5295	-3.249	0.00116
Weapon_TypeShotgun	-1.092	0.684	-1.596	0.11046
Weapon_TypeUnknown	0.3162	0.5351	0.591	0.55453
School_LevelHigh:State_Gun_ControlLow	-0.8417	0.7374	-1.141	0.25374

School_LevelMiddle:State_Gun_ControlLow	0.4099	1.336	0.307	0.75907
School_LevelMulti-Level:State_Gun_ControlLow	-16.42	926.6	-0.018	0.98586
School_LevelHigh:State_Gun_ControlMedium	-1.987	1.475	-1.347	0.17801
School_LevelMiddle:State_Gun_ControlMedium	-3.08	1.958	-1.573	0.11569
School_LevelMulti-Level:State_Gun_ControlMedium	4.329	8581	0.001	0.9996

Drive-by Shootings Outside (+): Strongly linked to multiple victims, indicating that such situations on school peripheries are especially high-risk.

Indiscriminate Shootings Inside (+): High risk for multiple victims, underscoring the danger of random targeting in confined spaces like classrooms.

Weapon Type:

Rifles (+): Associated with more victims due to their higher lethality.

Multiple Firearms (+): Significantly increases the likelihood of multiple casualties.

High School Level (+1.41): High schools show a stronger correlation with incidents involving multiple victims compared to other levels.

Gang-Related Incidents (+): Higher probability of multiple victims, pointing to the group-targeting nature of gang violence.

Shots Fired (+): Each additional shot fired increases the likelihood of more

By leveraging these models, we gain insights into various dimensions of school shootings, such as victim counts, timing, gang involvement, and casualty severity. Each model is tailored to address a specific aspect of the incidents, allowing us to make informed recommendations on preventive measures and policy interventions based on the key predictors identified.

Quality Checks

The tests focus on multicollinearity, linearity, independence of residuals, normality of residuals, and homoscedasticity.

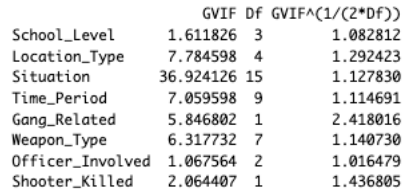
1. Test for Multicollinearity

The Variance Inflation Factor (VIF) was computed for all models.

-Interpretation: VIF values greater than 5 or 10 indicate potential multicollinearity issues.

- Findings:

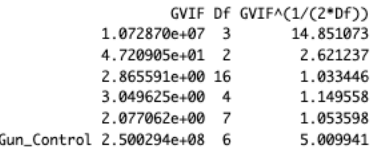
<table>
<tr>
<td>Model m1:



Low multicollinearity: School_level, office_involved, weapon_type

Low to moderate multicollinearity: Location_Type, Time_Period, Sho oter_Killed, Situation.

Moderate to high multicollinearity: Gang_Related</td>
<td>Model m3:



Low multicollinearity: school_affiliation, location_type, weapon_type

Moderate multincollinearity: state_gun_control

High multicollinearity: school_level</td>
</tr>
<tr>
<td>Model m2:

The models suggest a high multicollinearity between the features</td>
<td>Model m4:

The models suggest a high multicollinearity between the features</td>
</tr>
</table>

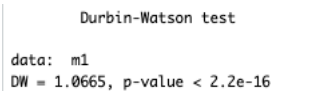
2. Test for Independence

The Durbin-Watson (DW) test was conducted to detect autocorrelation in residuals.

- Interpretation: A DW statistic close to 2 suggests no autocorrelation. Values significantly less than 2 indicate positive autocorrelation.

- Findings:

<table>
<tr>
<td>Model m1:



The result indicates positive autocorrelation in the residuals</td>
<td>Model m2:

The result from the test could not calculate autocorrelation due to high multi collinearity</td>
</tr>
<tr>
<td>Model m3:</td>
<td>Model m4:</td>
</tr>
</table>

Durbin-Watson test data: m3 DW = 1.6014, p-value = 1.048e-08 The result suggests **positive autocorrelation**	The result from the test could not calculate autocorrelation due to high multi collinearity

Recommendations

1. **Increase High School Security During Classes**
   Boost security presence and monitoring, especially during high-risk times like morning and afternoon classes, where incidents are more likely.
2. **Reinforce Building Entry Points**
   Implement controlled access to reduce risks for incidents inside school buildings, where victim counts tend to be higher. Train staff and students on lockdown procedures.
3. **Target Gang Prevention Programs in High Schools**
   Launch anti-gang initiatives and collaborate with local law enforcement in high schools to reduce gang-related incidents.
4. **Strengthen Perimeter for Drive-By Prevention**
   Install barriers around school perimeters to reduce drive-by shooting risks, especially in high-risk high schools.
5. **Heighten Security at Lunch and School Events**
   Increase supervision and deploy security personnel during lunch hours and school events, which show higher victim counts.
6. **Early Intervention for At-Risk Students**
   Implement programs focusing on mental health and conflict resolution to reduce violence risks, particularly among students with known gang affiliations.
7. **Train Staff on Recognizing Warning Signs**
   Educate teachers and staff on identifying signs of potential violence, enabling quicker intervention.
8. **Collaborate on Gun-Free School Zones**
   Work with local authorities to establish and enforce firearm-free zones around schools, especially high schools, to limit high-lethality firearms near campuses.

References

Collins, L. W., Landrum, T. J., & Sweigart, C. A. (2024). Getting ahead of school shootings: A call for action, advocacy and research. *Preventing School Failure: Alternative Education for Children and Youth*, *68*(2), 167–173. https://doi.org/10.1080/1045988x.2024.2302144

Flannery, D. J., Fox, J. A., Wallace, L., Mulvey, E., & Modzeleski, W. (2021). Guns, school shooters, and School Safety: What we know and directions for change. *School Psychology Review*, *50*(2–3), 237–253. https://doi.org/10.1080/2372966x.2020.1846458

Katsiyannis, A., Rapa, L. J., Whitford, D. K., & Scott, S. N. (2023). Correction to: An examination of US school mass shootings, 2017–2022: Findings and implications. *Advances in Neurodevelopmental Disorders*. https://doi.org/10.1007/s41252-023-00383-w

Reeping, P. M., Klarevas, L., Rajan, S., Rowhani-Rahbar, A., Heinze, J., Zeoli, A. M., Goyal, M. K., Zimmerman, M. A., & Branas, C. C. (2022). State firearm laws, gun ownership, and K-12 school shootings: Implications for School Safety. *Journal of School Violence*, *21*(2), 132–146. https://doi.org/10.1080/15388220.2021.2018332

Riedman, D. (n.d.). *Methodology for Collecting School Shooting Data*. K-12 School Shooting Database. https://k12ssdb.org/methodology-1

Schildkraut, J., Connell, N., Barbieri, N., & de Azeredo, R. (2023). American uniqueness revisited: A Comparative Examination of two school shootings using the path to intended violence. *International Journal of Comparative and Applied Criminal Justice*, *48*(2), 143–158. https://doi.org/10.1080/01924036.2023.2221751

Surviving a school shooting: Impacts on the mental health, education, and earnings of American Youth. Stanford Institute for Economic Policy Research (SIEPR). (n.d.). https://siepr.stanford.edu/publications/health/surviving-school-shooting-impacts-mental-health-education-and-earnings-american

USAFacts. (2024, February 20). *The latest government data on school shootings*. https://usafacts.org/articles/the-latest-government-data-on-school-shootings/

Winch, A. T., Alexander, K., Bowers, C., Straub, F., & Beidel, D. C. (2024). An evaluation of completed and averted school shootings. *Frontiers in Public Health*, *11*. https://doi.org/10.3389/fpubh.2023.1305286

Appendix: R Code

```
#School Shooting Preprocessing

getwd()

setwd("C:/Users/ny123/Downloads")

library(readxl)
```

```
library(dplyr)

rm(list=ls())

#Read each sheet of the excel file individually
incident = read_excel("School Shootings.xlsx", sheet = "Incident")
#incident2 = read_excel("School Shootings.xlsx", sheet = "Incident")
#unique(incident2$Shots_Fired)
shooter = read_excel("School Shootings.xlsx", sheet = "Shooter")
victim = read_excel("School Shootings.xlsx", sheet = "Victim")
weapon = read_excel("School Shootings.xlsx", sheet = "Weapon")
##Drop Incident ID at the end of preprocessing, not immediately
##Remember to add the school name back to the dropped columns

str(incident)
str(shooter)
str(victim)
str(weapon)



### Preprocessing Incident ###
#Only keeping rows where Number_Victims > 0 & Situation != "Accidental"
incident = incident %>% filter(Number_Victims > 0, Situation != "Accidental")
View(incident)
#str(incident)
#Check each column for missing values and print
sapply(incident, function(x) sum(is.na(x)))
colSums(is.na(incident))

#Checking how many incident happened as a suicide
```

```
#Drop columns that are not needed

  #Month     Day     Year     Date     School Victims_Killed     Victims_Wounded     Source
      Number_News     Media_Attention     Reliability     Quarter     City
      First_Shot     Summary     Narrative     Accomplice     Accomplice_Narrative
      Barricade     Active_Shooter_FBI   LAT   LNG Involved_Students_Staff

incident = incident %>% select(-Month, -Day, -Year, -Date, -Victims_Killed, -Victims_Wounded, -Source, -Number_News, -Media_Attention, -Reliability, -Quarter, -City, -First_Shot, -Summary, -Narrative, -Barricade, -Accomplice, -Accomplice_Narrative, -Active_Shooter_FBI, -LAT, -LNG)


#Checking for wherever Shooter_Killed = 2 and turning it into 1

incident$Shooter_Killed = ifelse(incident$Shooter_Killed == 2, 1, incident$Shooter_Killed)

#State Column (Includes all 50 states, plus DC & Virgin Islands but FL is repeated)

unique(incident$State)

table(incident$State)

incident = inciStateincident = incident %>% filter(!is.na(State))#Drop rows where State is NA (Only 1 row)

#1 row with state ="Florida" instead of "FL", so change it to "FL"

incident$State = ifelse(incident$State == "Florida", "FL", incident$State)

#Grouping states into three groups: High Gun Control, Medium Gun Control, Low Gun Control

  # High Gun Control: CA, CT, NJ, CO, HI, IL, MD, MA, NY, OR, WA, DC

  # Moderate Gun Control: DE, RI, VA, MN, PA, MI, NV, VT

  # Low Gun Control: NM, WI, NC, NE, FL, IN, ME, NH, OH, SC, AL, AK, AZ, AR, GA, ID, IA, KS, KY, LA, MS, MO, MT, ND, OK, SD, TN, TX, UT, WV, WY, VI

incident$State_Gun_Control = ifelse(incident$State %in% c("CA", "CT", "NJ", "CO", "HI", "IL", "MD", "MA", "NY", "OR", "WA", "DC"), "High", ifelse(incident$State %in% c("DE", "RI", "VA", "MN", "PA", "MI", "NV", "VT"), "Medium", "Low"))

incident = incident %>% select(-State)

#Drop Location column

incident = incident %>% select(-Location)
```

```
# Creating new feature that bins duration_min into 2 categories: "1 Minute or Less" and "Over 1 Minute"

#Check to see during modelling if this feature is giving trouble

#You have the option of keeping one or the other or keeping or dropping both

#The overwhleming majority of the data is under 1 minute, hence the decision to create this feature

incident$Duration_Min_Category <- ifelse(incident$Duration_min <= 1, "Under 1 min", "Over 1 min")

#For the missing values in Duration_min, impute them as 1

incident$Duration_min[is.na(incident$Duration_min)] <- 1

#For the missing values in Duration_Min_Category, impute them as "Under 1 min"

incident$Duration_Min_Category[is.na(incident$Duration_Min_Category)] <- "Under 1 min"

#barplot(table(incident$Duration_Min_Category), main="Duration of Incident", xlab="Duration", ylab="Frequency", col="blue", legend = rownames(table(incident$Duration_Min_Category)), beside=TRUE)


#Shots_Fired

#Visualizing the distribution of Shots_Fired

#hist(incident$Shots_Fired, main="Histogram of Shots Fired", xlab="Shots Fired", col="blue", breaks=2)

#str(incident$Shots_Fired)

incident$Shots_Fired <- as.character(incident$Shots_Fired)

unique(incident$Shots_Fired)

table(incident$Shots_Fired)

discrepancy.shotsfired = incident[which(incident$Shots_Fired == '0'), ]

View(discrepancy.shotsfired)

#'in discrepancy shotsfired data, we can observe that there were no shots fired, but the victims were killed.

#'The cause of death of these individuals could be something other than shooting. Or the data could have been falsely captured.

#'Dropping the 7 rows from the dataset

incident = incident %>% filter(Shots_Fired > 0)

incident[which(incident$Shots_Fired == '0'), ]
```

```
#Wherever "Multiple" is present, we can impute it with 10

#Shan's comments: why are there no wos with value as 'multiple' in the data frame, while the excel has those rows?

incident$Shots_Fired <- ifelse(incident$Shots_Fired == "Multiple", 10, incident$Shots_Fired)

#For all missing values in Shots_Fired, we can impute it with 26

#summary(as.numeric(incident$Shots_Fired)) #Median = 3, mean = 83

sum(is.na(incident$Shots_Fired))

incident$Shots_Fired[is.na(incident$Shots_Fired)] <- 26

#Handle rows under Shots_fired that have inequality symbols using Boundary Imputation

#Example: If a value is ">100", we can impute it with 101 and if "<10", we can impute it with 9

incident$Shots_Fired <- ifelse(grepl("^>", incident$Shots_Fired), as.numeric(sub(">", "", incident$Shots_Fired)) + 1, incident$Shots_Fired)

incident$Shots_Fired <- ifelse(grepl("^<", incident$Shots_Fired), as.numeric(sub("<", "", incident$Shots_Fired)) - 1, incident$Shots_Fired)

incident$Shots_Fired <- as.numeric(incident$Shots_Fired)

summary(incident$Shots_Fired)

#Check for missing values

sum(is.na(incident$Shots_Fired))

# Check for any remaining instances of "Multiple"

any(grepl("Multiple", incident$Shots_Fired))

# Check for any remaining instances of inequality symbols

any(grepl("^[><]", incident$Shots_Fired))

all(sapply(incident$Shots_Fired, is.numeric))

str(incident)


#Drop Involves_Student_Staff

#I realized that this column may not be useful for our analysis as it more a descriptor than a predictor and also, it has a lot of missing values

incident = incident %>% select(-Involves_Students_Staff)

str(incident)


#Check each column for missing values and print
```

```
sapply(incident, function(x) sum(is.na(x)))


table(incident$School_Level)

sum(is.na(incident$School_Level))

#Print the rows where School_Level is NA

incident %>% filter(is.na(School_Level))

#Drop rows where School_Level is NA because the school name is not provided to determine school level

incident = incident %>% filter(!is.na(School_Level))

table(incident$School_Level)

#Print the rows where School_Level is 44724, the school name, and the ID

incident %>% filter(School_Level == 44724)  %>% select(School, Incident_ID)

incident %>% filter(School_Level == "Unknown")  %>% select(School, Incident_ID)

#Sulphur Rock Magnet School, it's school level is Elementary

#East English Village Preparatory Academy, it's school level is High

#Episcopal School of Jacksonville, it's school level is K-12

#Dekalb Alternative School, it's school level is High

#Impute the missing values

incident$School_Level = ifelse(incident$School == "Sulphur Rock Magnet School", "Elementary", incident$School_Level)

incident$School_Level = ifelse(incident$School == "East English Village Preparatory Academy", "High", incident$School_Level)

incident$School_Level = ifelse(incident$School == "Episcopal School of Jacksonville", "K-12", incident$School_Level)

incident$School_Level = ifelse(incident$School == "Dekalb Alternative School", "High", incident$School_Level)

incident$School_Level = ifelse(incident$School == "School of Choice", "High", incident$School_Level)

table(incident$School_Level)

#Whichever rows have School_Level as "Junior High", impute it to "Middle"

incident$School_Level = ifelse(incident$School_Level == "Junior High", "Middle", incident$School_Level)
```

```
#Which ever rows have School_Level as "6-12", "K-12", "K-8", or "Other", combine them all into a new category called "Mutli-Level"

incident$School_Level = ifelse(incident$School_Level %in% c("6-12", "K-12", "K-8", "Other"), "Multi-Level", incident$School_Level)

#Only one school has school_level as 44724, impute it to "Multi-Level"

incident$School_Level = ifelse(incident$School_Level == 44724, "Multi-Level", incident$School_Level)



#Barplot of location_type

barplot(table(incident$Location_Type), main="Location Type", xlab="Location Type", ylab="Frequency", col="blue", legend = rownames(table(incident$Location_Type)), beside=TRUE)

#print the IDs of the rows where Location_Type has missing values

incident %>% filter(is.na(Location_Type)) %>% select(Incident_ID)

#The following Incident_IDs have missing values for Location_Type

#20191008TXWEH, 19830130TXWEC, 19810913MDABA

#Impute as follows:

   #20191008TXWEH, it's location_type is "Outside on School Property"

   #19830130TXWEC, it's location_type is "Outside on School Property"

   #19810913MDABA, it's location_type is "Outside on School Property"

incident$Location_Type = ifelse(incident$Incident_ID == "20191008TXWEH", "Outside on School Property", incident$Location_Type)

incident$Location_Type = ifelse(incident$Incident_ID == "19830130TXWEC", "Outside on School Property", incident$Location_Type)

incident$Location_Type = ifelse(incident$Incident_ID == "19810913MDABA", "Outside on School Property", incident$Location_Type)

table(incident$Location_Type)

#Drop whichever rows have Location_Type as "ND" due to lack of information to impute

incident = incident %>% filter(Location_Type != "ND")



#Check each column for missing values and print

sapply(incident, function(x) sum(is.na(x)))
```

```
#Barplot of During_Classes

barplot(table(incident$During_Classes), main="During Classes", xlab="During Classes", ylab="Frequency", col="blue", legend = rownames(table(incident$During_Classes)), beside=TRUE)

#print the rows  During_Classes has missing values

incident %>% filter(is.na(During_Classes))

#For the Incident_IDs(20050429OHDAC), impute During_Classes as "Yes"

incident$During_Classes = ifelse(incident$Incident_ID == "20050429OHDAC", "Yes", incident$During_Classes)

#For the Incident_IDs(19960514UTBIT), impute During_Classes as "No"

incident$During_Classes = ifelse(incident$Incident_ID == "19960514UTBIT", "No", incident$During_Classes)

#For the Incident_IDs(20211119COHIA), impute During_Classes as "Yes"

incident$During_Classes = ifelse(incident$Incident_ID == "20211119COHIA", "Yes", incident$During_Classes)

#For the Incident_IDs(19970428CAJOL), impute During_Classes as "Yes"

incident$During_Classes = ifelse(incident$Incident_ID == "19970428CAJOL", "Yes", incident$During_Classes)

#For the incident_IDs(19970403CAMAM), impute During_Classes as "No"

incident$During_Classes = ifelse(incident$Incident_ID == "19970403CAMAM", "No", incident$During_Classes)

#For the incident_IDs(19960411ALTAT), impute During_Classes as "No"

incident$During_Classes = ifelse(incident$Incident_ID == "19960411ALTAT", "No", incident$During_Classes)

#For the incident_IDs(19920128LAFRG), impute During_Classes as "Yes"

incident$During_Classes = ifelse(incident$Incident_ID == "19920128LAFRG", "Yes", incident$During_Classes)

#For the incident_IDs(19811209NYGEB), impute During_Classes as "Yes"

incident$During_Classes = ifelse(incident$Incident_ID == "19811209NYGEB", "Yes", incident$During_Classes)

#For the incident_IDs(19731109CALOL), impute During_Classes as "No"

incident$During_Classes = ifelse(incident$Incident_ID == "19731109CALOL", "No", incident$During_Classes)
```

```
#print the rows for Officer_Involved that has missing values

incident %>% filter(is.na(Officer_Involved))

#For the Incident_IDs(20220609ALWAG), impute Officer_Involved as "Yes"

incident$Officer_Involved = ifelse(incident$Incident_ID == "20220609ALWAG", "Yes", incident$Officer_Involved)

#For the Incident_IDs(20220421NDMOM), impute Officer_Involved as "Yes"

incident$Officer_Involved = ifelse(incident$Incident_ID == "20220421NDMOM", "Yes", incident$Officer_Involved)


#Check each column for missing values and print

sapply(incident, function(x) sum(is.na(x)))


#print the rows for Hostages that has missing values

incident %>% filter(is.na(Hostages))

#Barplot of Hostages

barplot(table(incident$Hostages), main="Hostages", xlab="Hostages", ylab="Frequency", col="blue", legend = rownames(table(incident$Hostages)), beside=TRUE)

#For the Incident_IDs(20220609ALWAG), impute Hostages as "No"

incident$Hostages = ifelse(incident$Incident_ID == "20220609ALWAG", "No", incident$Hostages)

#For the Incident_IDs(20220513FLALW), impute Hostages as "No"

incident$Hostages = ifelse(incident$Incident_ID == "20220513FLALW", "No", incident$Hostages)

#For the Incident_IDs(20220427TXMOS), impute Hostages as "No"

incident$Hostages = ifelse(incident$Incident_ID == "20220427TXMOS", "No", incident$Hostages)

#For the Incident_IDs(20220421NDMOM), impute Hostages as "No"

incident$Hostages = ifelse(incident$Incident_ID == "20220421NDMOM", "No", incident$Hostages)

#For the Incident_IDs(20220329NVWEL), impute Hostages as "No"

incident$Hostages = ifelse(incident$Incident_ID == "20220329NVWEL", "No", incident$Hostages)

#For the Incident_IDs(20211207ILHAC), impute Hostages as "No"
```

```
incident$Hostages = ifelse(incident$Incident_ID == "20211207ILHAC", "No", incident$Hostages)

#For the Incident_IDs(20210812GALIL), impute Hostages as "No"

incident$Hostages = ifelse(incident$Incident_ID == "20210812GALIL", "No", incident$Hostages)

#For the Incident_IDs(20210511NYPSB), impute Hostages as "No"

incident$Hostages = ifelse(incident$Incident_ID == "20210511NYPSB", "No", incident$Hostages)

#For the Incident_IDs(20210502ILCHC), impute Hostages as "No"

incident$Hostages = ifelse(incident$Incident_ID == "20210502ILCHC", "No", incident$Hostages)


table(incident$Time_Period)

summary(incident$Time_Period)

#Decide before dropping Time_Period column. Do you want to keep it and drop DuringClasses or vice versa?

#incident = incident %>% select(-Time_Period)

#Print the rows where Time_Period is missing

incident %>% filter(is.na(Time_Period)) %>% select(Incident_ID)

#For the missing values in Time_Period, impute them as "Unknown"

incident$Time_Period = ifelse(is.na(incident$Time_Period), "Unknown", incident$Time_Period)

#Consolidate "Before School" & "School Start" into ine new category called "Before/Start of School"

incident$Time_Period = ifelse(incident$Time_Period %in% c("Before School", "School Start"), "Before/Start of School", incident$Time_Period)

#Merge "Not A School Day" with "Not a School Day"

incident$Time_Period = ifelse(incident$Time_Period == "Not A School Day", "Not a School Day", incident$Time_Period)

#Merge "School Event" and "Sport Event" into one category called "School/Sport Event"

incident$Time_Period = ifelse(incident$Time_Period %in% c("School Event", "Sport Event"), "School/Sport Event", incident$Time_Period)

#Combine "Night" and "Evening" into one category called "Night/Evening"

incident$Time_Period = ifelse(incident$Time_Period %in% c("Night", "Evening"), "Night/Evening", incident$Time_Period)
```

```
#Check each column for missing values and print
sapply(incident, function(x) sum(is.na(x)))


#Barplot of Gang_Related
barplot(table(incident$Gang_Related), main="Gang Related", xlab="Gang Related", ylab="Frequency", col="blue", legend = rownames(table(incident$Gang_Related)), beside=TRUE)
#Print the rows where Gang_Related is missing
incident %>% filter(is.na(Gang_Related))
str(incident)


# For the missing values in "Gang_Related", apply the following rules:
# Assign "Gang_Related" = "Yes" if:
#   - Situation = "Drive-by Shooting".
#   - Situation = "Escalation of Dispute" AND Location_Type = "Outside on School Property".
#   - Situation = "Illegal Activity" AND Location_Type = "Outside on School Property".

# Assign "Gang_Related" = "No" if:
#    - Situation = "Escalation of Dispute" AND Location_Type = "Inside School Building" or "School Bus".
#   - Situation = "Illegal Activity" AND Location_Type = "Inside School Building".
#   - Situation = "Indiscriminate Shooting", "Racial", or "Unknown".

# For any remaining missing values:
#   - Impute based on the existing distribution:
#     - Randomly assign "Yes" or "No" to match the proportion of the non-missing values
# Assign "Gang_Related" based on specific conditions
incident$Gang_Related <- ifelse(incident$Situation == "Drive-by Shooting", "Yes", incident$Gang_Related)
incident$Gang_Related <- ifelse(incident$Situation == "Escalation of Dispute" & incident$Location_Type == "Outside on School Property", "Yes", incident$Gang_Related)
```

```
incident$Gang_Related <- ifelse(incident$Situation == "Illegal Activity" & incident$Location_Type == "Outside on School Property", "Yes", incident$Gang_Related)

incident$Gang_Related <- ifelse(incident$Situation == "Escalation of Dispute" & incident$Location_Type %in% c("Inside School Building", "School Bus"), "No", incident$Gang_Related)

incident$Gang_Related <- ifelse(incident$Situation == "Illegal Activity" & incident$Location_Type == "Inside School Building", "No", incident$Gang_Related)

incident$Gang_Related <- ifelse(incident$Situation %in% c("Indiscriminate Shooting", "Racial", "Unknown"), "No", incident$Gang_Related)

# Calculate proportions of Yes and No in Gang_Related for non-missing values

proportions <- table(incident$Gang_Related[!is.na(incident$Gang_Related)]) / sum(!is.na(incident$Gang_Related))

# Impute remaining missing values in "Gang_Related" based on existing distribution

incident$Gang_Related[is.na(incident$Gang_Related)] <- sample(c("Yes", "No"),
                                                             size = sum(is.na(incident$Gang_Related)),
                                                             replace = TRUE,
                                                             prob = c(proportions["Yes"], proportions["No"]))

# Check for any remaining missing values

sum(is.na(incident$Gang_Related))

# Check for any remaining instances of "Unknown"

any(grepl("Unknown", incident$Gang_Related))


#Check each column for missing values and print

sapply(incident, function(x) sum(is.na(x)))

table(incident$Bullied)

sum(is.na(incident$Bullied))

#Barplot of Bullied

barplot(table(incident$Bullied), main="Distribution of Bullied", xlab="Bullied", ylab="Frequency", col="blue", legend = rownames(table(incident$Gang_Related)), beside=TRUE)

#Bullied is difficult to assess from the other columns, dropping the rows where bullied is blank

incident =incident %>% filter(!is.na(Bullied))
```

```
#Handling Domestic Violence
sum(is.na(incident$Domestic_Violence))
incident %>% filter(is.na(Domestic_Violence)) # 8 rows had null values for domestic violence
incident %>% filter(Domestic_Violence == 'N/A')
unique(incident$Domestic_Violence)
table(incident$Domestic_Violence)
#we can drop N/A and null values from the data. We cannot classify these characteristics based on assumptions
incident =incident %>% filter(!is.na(Domestic_Violence))
sapply(incident, function(x) sum(is.na(x)))
#the na rows cannot be explained into an yes or a no based on other characteristics. Dropping them form the data
incident =incident %>% filter(Domestic_Violence != 'N/A')


#Check each column for missing values and print
sapply(incident, function(x) sum(is.na(x)))
colSums(is.na(incident))


#Handling Targets
incident %>% filter(is.na(Targets))
table(incident$Situation,incident$Targets)
#assigning the value of unknown to null targets
incident$Targets[is.na(incident$Targets)] = 'unknown'
table(incident$Targets)
table(incident$Situation,incident$Targets)
unknowntargets = incident %>% filter(incident$Targets == 'unknown')
View(unknowntargets)
#assign target as 'Victims Targeted' when the situation is escalation of dispute
table(incident$Situation,incident$Targets)
for (i in nrow(incident)) {
```

```
  incident$Targets = ifelse(incident$Targets == 'unknown' & incident$Situation == 'Escalation of Dispute', 'Victims Targeted', incident$Targets)

  incident$Targets = ifelse(incident$Targets == 'unknown' & incident$Situation == 'Indiscriminate Shooting', 'Random Shooting', incident$Targets)

  incident$Targets = ifelse(incident$Targets == 'unknown' & incident$Situation == 'Officer-Involved Shooting', 'Victims Targeted', incident$Targets)

  incident$Targets = ifelse(incident$Targets == 'unknown' & incident$Situation == 'Drive-by Shooting', 'Both', incident$Targets)

  incident$Targets = ifelse(incident$Targets == 'unknown' & incident$Situation == 'Intentional Property Damage', 'Random Shooting', incident$Targets)


} #leaves 8 rows with unknowns. we will have to drop these, as we cannot make assumption on the targets for during an illegal activitu, and for situation of unknown nature

incident %>% filter(incident$Targets == 'unknown')

incident = incident %>% filter(incident$Targets != 'unknown')


#checkign for missing values in incident

colSums(is.na(incident))


#For Shooter

#checking for nulls in dataframe shooter

colSums(is.na(shooter))

#there are no nulls in incident_ID. checking the unique count of incidents in shooter table

length(unique(shooter$Incident_ID))

#2885

#How many unique values of incidents do we have in the incident table

length(unique(incident$Incident_ID))

#690

#what are the duplicates in the shooter table

duplicated(shooter$Incident_ID)

#Joining the shooter table to the incident table using a left join on incident by incident_id

intermediate = left_join(incident,shooter,by = 'Incident_ID')

str(intermediate)
```

```
#how many nulls do we have in the new intermediate table

colSums(is.na(intermediate))

#The increase in row numbers from incident df to intermediate suggests that there could be multiple shooters in incidents


#handling Age in Shooter table

table(intermediate$Age)

#shooters under the age 18 can be assigned to Minor

for (i in nrow(intermediate)) {

  intermediate$age.new[i] = ifelse(intermediate$Age[i] < 18 | intermediate$Age[i] == 'Minor', 'Minor', 'Adult')

}

table(intermediate$age.new,intermediate$Age)

#We created a new feature that categorizes shooters based on age into Minors or Adults

#intermediate <- intermediate %>% select(-Age)

#How many nulls are there in Age?

colSums(is.na(intermediate))

#103

  intermediate$age.new = ifelse(is.na(intermediate$age.new), "Unknown", intermediate$age.new)

table(intermediate$age.new)

table(intermediate$age.new,intermediate$School_Affiliation)

#Based on the school affiliationn we can assign the age of the 2 unknown rows i.e SRO

for (i in nrow(intermediate)) {

  intermediate$age.new[i] = ifelse(intermediate$age.new[i]=='Unknown' & intermediate$School_Affiliation[i] == 'Police Officer/SRO', 'Adult', intermediate$age.new[i])

  }

table(intermediate$age.new,intermediate$School_Affiliation)

colSums(is.na(intermediate))


#Dropping the race feature as there are too many nulls (over 500)

intermediate = intermediate %>% select(-Race)
```

```
#Gender
#ommiting the null values from the data
#before 775
intermediate = intermediate %>% filter(!is.na(Gender))
#after 681
colSums(is.na(intermediate))


#School affiliation
table(intermediate$School_Affiliation)
barplot(table(intermediate$School_Affiliation), main="Location Type", xlab="School Affiliation", ylab="Frequency", col= colors(),legend = rownames(table(intermediate$School_Affiliation)), beside=FALSE)
intermediate$School_Affiliation = ifelse(is.na(intermediate$School_Affiliation), "Unknown", intermediate$School_Affiliation)
table(intermediate$School_Affiliation)
#what to do with the unknowns?
colSums(is.na(intermediate))
#Joining intermediate with weapon
m0 = left_join(intermediate,weapon, by='Incident_ID')
colSums(is.na(m0))
#dropping weapon details form weapon, it has too many nulls
m0 = m0 %>% select(-Weapon_Details)
colSums(is.na(m0))
unique(m0$Weapon_Caliber)
#droppin weapon calliber as it is not significant for our analysis
m0 = m0 %>% select(-Weapon_Caliber)
#what are the different types of weapons
unique(m0$Weapon_Type)
table(m0$Weapon_Type)
#assigning unknown and NA weapon types as unknown
```

```
m0$Weapon_Type = ifelse(is.na(m0$Weapon_Type), "Unknown", m0$Weapon_Type)
table(m0$Weapon_Type)
unique(m0$Weapon_Type)
#assigning no data value as unknown weapon type
for (i in 1:nrow(m0)) {
  m0$Weapon_Type[i] = ifelse(m0$Weapon_Type[i] == 'No Data', 'Unknown', m0$Weapon_Type[i])


}
table(m0$Weapon_Type)
unique(m0$Weapon_Type)
str(m0)
# Function to get top 10 unique values
get_top_10_unique <- function(x) {
  unique_vals <- unique(x)
  if(length(unique_vals) > 10) {
    return(head(unique_vals, 10))
  } else {
    return(unique_vals)
  }
}
# Apply the function to each column
result <- lapply(m0, get_top_10_unique)
# Print the results
for(col_name in names(result)) {
  cat("\nTop 10 unique values for", col_name, ":\n")
  print(result[[col_name]])
}


# Columns to be converted to factors
factor_columns <- c(
```

```
  "School_Level", "Location_Type", "During_Classes", "Time_Period",
  "Situation", "Targets", "Hostages", "Officer_Involved", "Bullied",
  "Domestic_Violence", "Gang_Related", "State_Gun_Control",
  "Duration_Min_Category", "Gender", "School_Affiliation",
  "Shooter_Outcome", "Shooter_Died", "Injury", "age.new", "Weapon_Type"
)
# Convert the selected columns to factors
m0[factor_columns] <- lapply(m0[factor_columns], factor)
# Verify the changes
str(m0[factor_columns])
str(m0)
#Plotting the Correlation Matrix
library(tidyverse)
library(corrplot)
# First mutate to convert factors to numeric, then select numeric variables
m0_numeric <- m0 %>%
  mutate(
    School_Level = as.numeric(School_Level),
    During_Classes = as.numeric(factor(During_Classes, levels = c("No", "Yes"))),
    State_Gun_Control = as.numeric(factor(State_Gun_Control, levels = c("Low", "Medium", "High"))),
    Duration_Min_Category = as.numeric(factor(Duration_Min_Category, levels = c("Under 1 min", "Over 1 min"))),
    Gender = as.numeric(factor(Gender, levels = c("Female", "Male"))),
    Shooter_Died = as.numeric(factor(Shooter_Died, levels = c("No", "Yes")))
  ) %>%
  # Now select only the numeric variables for correlation analysis
  select(Number_Victims, Shooter_Killed, Duration_min, Shots_Fired,
         School_Level, During_Classes, State_Gun_Control,
         Duration_Min_Category, Gender, Shooter_Died)
```

```
# Calculate the correlation matrix
cor_matrix <- cor(m0_numeric, use = "pairwise.complete.obs")
print(cor_matrix)

library("PerformanceAnalytics")
chart.Correlation(m0_numeric, use = "pairwise.complete.obs")

# Print the correlation matrix
print(cor_matrix)

# Visualize the correlation matrix
corrplot(cor_matrix, method = "color", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
#Histogram of dependent variables
library(ggplot2)
library(gridExtra)
# Function to create histogram or bar plot depending on variable type
create_histogram_or_barplot <- function(data, column, title) {
  if (is.factor(data[[column]]) || is.character(data[[column]])) {
    # Bar plot for categorical variables
    p <- ggplot(data, aes_string(x = column)) +
      geom_bar(fill = "lightblue", color = "black") +
      ggtitle(title) +
      theme_minimal()
    # Special handling for Shooter_Outcome
    if (column == "Shooter_Outcome") {
      p <- p + theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 6)) +
        scale_x_discrete(labels = function(x) substr(x, 1, 10))  # Truncate labels
    }
  } else {
```

```
    # Histogram for continuous variables
    p <- ggplot(data, aes_string(x = column)) +
      geom_histogram(aes(y = ..density..), bins = 30, fill = "lightblue", color = "black") +
      geom_density(color = "red") +
      ggtitle(title) +
      theme_minimal()
  }
  return(p)
}
# Create plots for each variable
p1 <- create_histogram_or_barplot(m0, "During_Classes", "Histogram of During Classes")
p2 <- create_histogram_or_barplot(m0, "Gang_Related", "Histogram of Gang Related")
p3 <- create_histogram_or_barplot(m0, "Shooter_Outcome", "Histogram of Shooter Outcome")
p4 <- create_histogram_or_barplot(m0, "Number_Victims", "Histogram of Number of Victims")
p5 <- create_histogram_or_barplot(m0, "Domestic_Violence", "Histogram of Domestic Violence")
p6 <- create_histogram_or_barplot(m0, "Duration_min", "Histogram of Duration (minutes)")
grid.arrange(p1, p2, p3, p4, p5, p6, ncol = 2)


# For monthly incident counts (time series data), we need to convert it
#Finding unique values of some columns
unique(m0$School)
unique(m0$School_Level)
unique(m0$Location_Type)
unique(m0$Time_Period)
unique(m0$Duration_min)
unique(m0$Situation)
unique(m0$Targets)
# counting the unique values of Target column for insights
 table(m0$Targets)
 table(m0$Hostages)
```

```
 table(m0$Officer_Involved)
 table(m0$Bullied)
 table(m0$Domestic_Violence)
  table(m0$Gang_Related)
  sapply(m0, function(x) length(unique(x)))
  # counting from the master dataset
  result <- lapply(m0, function(col) {
    unique_count <- length(unique(col))
    if (unique_count < 30) {
      return(table(col))
    } else {
      return(NULL)
    }
  })
  # Remove NULL elements (columns with 30 or more unique values)
  result <- result[!sapply(result, is.null)]
  # Display the results
  for (col_name in names(result)) {
    cat("\n", col_name, ":\n")
    print(result[[col_name]])
  }
############ MODELS
m1 = glm(Number_Victims ~ School_Level + Location_Type + Situation + Time_Period + Gang_Related + Weapon_Type + Officer_Involved + Shooter_Killed, family = poisson, data = m0)
m2 =glm(During_Classes ~ School_Level * Time_Period + Location_Type,
            data = m0, family = binomial)



m3 = glm(Gang_Related ~ School_Level * State_Gun_Control + School_Affiliation + Location_Type + Weapon_Type,
```

```
        data = m0, family = binomial)


m4 =  glm(Number_Victims > 1 ~ Situation * Location_Type + Weapon_Type + School_Level + Gang_Related + Shots_Fired + age.new,
        data = m0, family = binomial)


library(stargazer)
stargazer(m1, m2, m3, m4, type="text")
library(forecast)
m0$Date <- as.Date(substr(m0$Incident_ID, 1, 8), format = "%Y%m%d")
monthly_counts <- table(cut(m0$Date, breaks = "month"))
ts_data <- ts(monthly_counts, frequency = 12, start = c(year(min(m0$Date)), month(min(m0$Date))))
arima_model <- auto.arima(ts_data)
forecast(arima_model, h = 12)
#CLINE TEST
# Load necessary libraries
library(car)
library(lmtest)
# Display model summaries
stargazer(m1, m2, m3, m4, type="text")
# 1. Test for Multicollinearity using Variance Inflation Factor (VIF)
vif(m1)
vif(m2)
vif(m3)
vif(m4)
# 2. Test for Linearity (Residuals vs Fitted Plot)
par(mfrow = c(2, 2))
plot(m1, which = 1, main = "Residuals vs Fitted (m1)")
plot(m2, which = 1, main = "Residuals vs Fitted (m2)")
plot(m3, which = 1, main = "Residuals vs Fitted (m3)")
```

```
plot(m4, which = 1, main = "Residuals vs Fitted (m4)")
# 3. Test for Independence (Durbin-Watson Test)
dwtest(m1)
dwtest(m2)
dwtest(m3)
dwtest(m4)
# 4. Test for Normality of Residuals
par(mfrow = c(2, 2))
qqnorm(residuals(m1), main = "Q-Q Plot (m1)")
qqline(residuals(m1), col = "red")
qqnorm(residuals(m2), main = "Q-Q Plot (m2)")
qqline(residuals(m2), col = "red")
qqnorm(residuals(m3), main = "Q-Q Plot (m3)")
qqline(residuals(m3), col = "red")
qqnorm(residuals(m4), main = "Q-Q Plot (m4)")
qqline(residuals(m4), col = "red")
# Shapiro-Wilk normality test
shapiro.test(residuals(m1))
shapiro.test(residuals(m2))
shapiro.test(residuals(m3))
shapiro.test(residuals(m4))
# 5. Test for Homoscedasticity (Equal Variance)
# Breusch-Pagan Test
bptest(m1)
bptest(m2)
bptest(m3)
bptest(m4)
# Non-constant Variance Score Test (NCV test)
ncvTest(m1)
ncvTest(m2)
```

```
ncvTest(m3)
ncvTest(m4)
# Reset plot layout to default
par(mfrow = c(1, 1))
```