Unveiling Dichotomies in North American Gun Violence through Multivariate Insights

Submitted By

Arham Anwar

McGill ID 261137773

MGSC661

Multivariate Statistics

Final Project

Contents

Section 1 – Introduction

Gun violence is a critical societal issue, necessitating a comprehensive understanding of incidents to inform preventive strategies. This study employs a dataset of gun-related incidents to extract valuable insights. Through advanced statistical techniques and visualizations, this analysis aims to unravel patterns, characteristics, and relationships hidden behind recorded history and extend previous efforts in curating this dataset. The project aims to answer the following key questions –

"What type of crime occurred in each incident?"

"What are the trends of these types of gun violence?"

"What can we, as a society, do about it?"

1.1. Summary of Project

This project delves into the comprehensive analysis of over 260,000 US gun violence incidents from 2013 to 2018, utilizing Kaggle data aggregated from the Gun Violence Archive. Through meticulous feature engineering, the study extracts crucial insights, including lethality based on weapon type, relationships among victims and suspects, and age profiles. The subsequent application of K-Means Clustering and PCA reveals nine distinct clusters, ranging from smaller-scale urban conflicts to organized crime and extreme outlier 'terrorism' events. Yearly trends expose alarming increases in specific clusters, necessitating targeted interventions. The findings advocate for focus areas and conclude with possible model extensions.

1.2. The Goals

Through meticulous feature engineering and clustering, it aims to identify patterns, trends, and clusters, providing insights to guide targeted interventions. The ultimate objective is to contribute to informed decision-making for mitigating the complex factors associated with gun violence in the United States.

Section 2 – Data Description

The Data Description section encompasses the origin of the dataset from Kaggle, feature engineering processes, and initial explorations.

2.1. Data Source:

The data source was taken from Kaggle, titled "*Gun Violence Data Comprehensive record of over 260k US gun violence incidents from 2013-2018*" The owner aggregate data from Gun Violence Archive (GVA), which is a not-for-profit corporation formed in 2013 to provide free online public access to accurate information about gun-related violence in the United States. While the dataset was comprehensive, it had many important pieces of information hidden in its original features. In the next subsection we discuss the import features engineered from the original dataset.

2.2. Feature Engineering:

The dataset demanded extensive cleansing, slicing, and dicing. The analysis involves preprocessing steps, including the calculation of counts for various aspects such as stolen guns, the number of subjects, victims, and total individuals involved, as well as categorization based on age groups and genders. Engineered features listed below aim to capture essential nuances within the dataset.

2.2.1. Lethality – Weapon of Choice:

First of all, quantity and quality of weapon lethality were extracted by calculating the number of guns by gun type, i.e number of handguns, rifles, and shotguns used in each crime incident. Some portions of guns were stolen guns or unknown build which were captured through the feature '*number of unknown and other*' guns. The lethality can give interesting insights, for instance, AK-47, an assault rifle, is highly unlikely to be used in lower-level unorganized crimes.

2.2.2. Victims vs Suspects:

Secondly, number of victims and suspects was calculated which gives us a picture of how organized the crime was. For instance, a crime incident with more than 20 victims could imply mass shooting/professional criminals, or on the other hand, a crime scene with just one victim and one suspect could very likely be a conflict rising from a personal backstory. The chances of it being a random serial shooting is not negated, which is why we use it in complement with the whole dataset.

2.2.3. Kill-Death-Assist:

Thirdly number of injured, killed, unharmed arrested, and unharmed participants was calculated. This information is crucial to understand whether the incident was driven by individuals who have done gun shooting before. For instance, the goal of a low level crime such as theft, is usually not to kill people, therefore cases related to theft and robbery are less like to have any killed people and have more unharmed or injured participants.

2.2.4. Age Profile:

Fourth, participants were classified by age group and the count of each age group was calculated for every crime scene. The age brackets are child, teen, and adult. Shootings involving only ‘child’ and ‘teen’ participants as victims possibly implies kidnapping and shooting.

2.2.5. Female Percentage:

Similar to age group, the female percentage of participants was also calculated which gives insight into various types of crimes such as the perpetrator may be specifically targeting women for personal reasons. This could be due to a personal grudge, domestic violence, or other motives related to the victims' gender.

2.2.6. Relationship Status:

Relationship status between the perpetrator and the victims is the final piece of the puzzle and gives clear insight on the driving forces of the perpetrator. To quantify relationships, we have binary value columns for each relationship which were extracted from one of the original columns. For instance, ‘Gang vs Gang’ clearly tells us about the type of incident, however, the source of truth for these features was not fully complete and thus we don’t have information on many cases.

2.3. Feature Selection for Model:

The final selection of features ensured no variables with repeated information were taken. This was done by analyzing the multicollinearity matrix and using PCA to reduce dimensionality. Additionally, variables with no correlation to the crime scene were eliminated, such as the ‘source link’. Variable like the ‘source-link’ are realized after the crime scene takes place and thus offer no valuable insight in light of the objective of this project. The finale set of features employed for clustering can be seen in Figure-1. After feature engineering, the next step was to find the right model for grouping crime incidents.

Section 3 – Model Selection & Methodology:

In line with the objective of this project, it was known that we needed to apply a clustering algorithm to club criminal activities. Three candidate algorithms were tested, such DBScan, K-Median Clustering, Hierarchical clustering and K-Means clustering. The final model selection was K-Means Clustering in combination with PCA.

3.1. The decision of 'K':

The optimal number of clusters was determined through careful consideration of the data and iterative testing. Figure 2A, representing a plot of the total weighted sum of squares vs 'K', in combination with Figure 2B, representing a plot of gap statistic vs 'K' gives the best k for this objective in combination with case considerations to be 9.

3.2. Achilles heal to the curse of dimensionality:

PCA was employed to club engineered features of the same type. For instance, we had 8 features just for the relationship of the participants. It was only natural that some of these categories had some overlapping information. PCA was employed just among sets of similar variables to to mitigate the curse of dimensionality.

Section 4 – Results

The application of PCA and k-means clustering revealed distinct groupings within the gun violence dataset. Clusters were identified based on shared characteristics, allowing for a more granular understanding of the incidents. A snapshot of the results can be found in Figure 3 and table -1. The cluster interpretations are as follows:

Cluster 1: The Urban Turbulence

Cluster 1, characterized by 35,588 incidents, appears to encapsulate a series of smaller-scale urban incidents within the United States. The moderate involvement of firearms and lower overall violence levels suggest that these incidents may involve familial disputes or localized conflicts. The prevalence of family-oriented relationships among victims and perpetrators implies that these incidents might be rooted in personal matters within the community.

Cluster 2: The Ruthless Warfare

With a staggering size of 60,564 incidents, Cluster 2 paints a vivid picture of a more extensive and intense set of events. The high involvement of guns and elevated violence levels point toward a possible connection with organized crime or gang-related activities. This cluster might be indicative of larger-scale conflicts, resembling a form of ruthless urban warfare that unfortunately has become a notable feature within certain regions of the United States.

Cluster 3: The Stealthy Offenders

Cluster 3, consisting of 4,084 incidents, suggests a different facet of criminal activity. The high involvement of firearms and moderate violence levels, coupled with a variety of victim-perpetrator relationships, implies a more strategic and planned approach to criminal endeavours. This cluster may represent incidents involving organized groups engaging in activities such as armed robbery, indicating a certain level of sophistication in their operations.

Cluster 4: Definition of Terrorism

Comprising only 217 incidents, Cluster 4 stands out as a set of extreme outliers. The exceptionally high number of guns involved and elevated violence levels suggest incidents that are unique or particularly extreme. These may be indicative of rare but highly impactful events such as mass shootings or terrorist activities, underscoring the need for special attention and analysis.

Cluster 5: The Strained Relationships

Cluster 5, with 8,309 incidents, appears to revolve around conflicts arising from personal relationships. The moderate involvement of firearms and high violence levels hint at the intensity of disputes among acquaintances and friends. This cluster might mirror the challenges posed by strained personal relationships, escalating to dangerous levels that impact public safety in various communities across the United States.

Cluster 6: The Domestic Disturbance

With 2,904 incidents, Cluster 6 seems to focus on incidents within families. The moderate violence levels and lower number of guns involved suggest a less extreme form of domestic disturbance. This cluster sheds light on the unfortunate reality of familial conflicts that, while impactful, may not escalate to the levels seen in larger-scale urban incidents.

Cluster 7: The Isolated Incident

Cluster 7, comprised of 1,987 incidents, seems to encapsulate relatively isolated events with a moderate level of violence. The low involvement of firearms and the presence of mass shootings in some cases highlight incidents that, while not frequent, pose potential public safety concerns. These isolated incidents may underscore the challenges faced by law enforcement in preventing such occurrences.

Cluster 8: The Tense Workplace

With 2,396 incidents, Cluster 8 suggests incidents occurring within workplace settings, potentially involving workplace conflicts. The moderate involvement of firearms and high violence levels imply that these are not mere disagreements but rather serious altercations within the professional sphere. This cluster underscores the challenges of maintaining a safe working environment and the potential for workplace conflicts to escalate.

Cluster 9: The Teen Turmoil

The largest cluster with 12,242 incidents, Cluster 9 represents events where teenagers play a significant role. The low to moderate involvement of firearms and a moderate level of violence suggest a mix of conflicts within this age group. This cluster reflects the challenges of addressing teenage turmoil and conflicts, underscoring the need for targeted interventions and support mechanisms to ensure the well-being of the younger population in the United States.

Section 5 – Cluster Trends from the lens of an Analyst

The aftermath of the clustering model was to analyze crucial trends and find learnings to mitigate gun violence.

5.1. Yearly Trends

As seen in Figure-4, the two key clusters which have shown an exponential incline are cluster 1 and cluster 2. Clusters demanding focused effort due to alarming rates of increase: Urban Turbulence (smaller-scale urban conflicts), Ruthless Warfare (high gun involvement in organized crime); and those with moderate rates: Teen Turmoil (conflicts involving teenagers with moderate violence), Strained Relationships (conflicts among acquaintances with guns). Please see Figure 5 showcasing an oversimplified regression to predict the number of cluster 4 'Mass Shooting' activities to occur in 2024.

5.2. What this means for us:

In the contemporary societal landscape, a gigantic wave of discontent has given rise to noteworthy trends, notably an uptick in issues such as gang warfare and an unsettling surge in teenage involvement in gun-related incidents. Within Cluster 1, denoted as Urban Turbulence, localized urban conflicts underscore the imperative for strategic interventions to rectify underlying urban challenges and fortify community safety measures. Concurrently, the manifestation of Ruthless Warfare in Cluster 2, marked by heightened gun engagement within organized crime, underscores the necessity for a comprehensive strategy involving intensified law enforcement, community engagement, and targeted initiatives to dismantle the foundations of organized criminal activities.

In the midst of this societal upheaval, Clusters 9 and 5 reveal moderate escalations in Teen Turmoil and Strained Relationships, respectively. Addressing these issues requires proactive measures, including the implementation of youth-centric outreach programs to provide constructive alternatives for teenagers navigating turbulent circumstances. Also, fostering conflict resolution initiatives is paramount in alleviating the repercussions of strained relationships among acquaintances possessing firearms. As we confront these intricate challenges, a holistic approach that encompasses rigorous root cause analysis, community empowerment, and the implementation of comprehensive gun control measures emerges as a collective imperative to foster a safer and more resilient society.

Section 7 – Conclusion & Future Scope:

In conclusion, this comprehensive analysis of gun violence incidents in the United States has successfully utilized advanced statistical techniques, visualizations, and clustering methodologies to extract meaningful insights from a dataset spanning from 2013 to 2018. The identified clusters, ranging from Urban Turbulence and Ruthless Warfare to Teen Turmoil and Strained Relationships, shed light on the diverse nature of gun-related incidents and emphasize the urgent need for targeted interventions. The year-over-year trends underscore the alarming rise in specific clusters, necessitating focused efforts to address escalating issues such as gang warfare, teenage involvement, and conflicts among acquaintances. As society grapples with these challenges, the findings advocate for a multifaceted approach that includes community engagement, strategic law enforcement, youth-centric programs, and comprehensive gun control measures to foster a safer and more resilient societal landscape. This project serves as a foundation for informed decision-making and proactive measures aimed at mitigating the complex factors contributing to gun violence in the United States.

Section 8 – Appendices

8.1 Figures

Fig 1 - Correlation Plot of features finally selected for clustering analysis

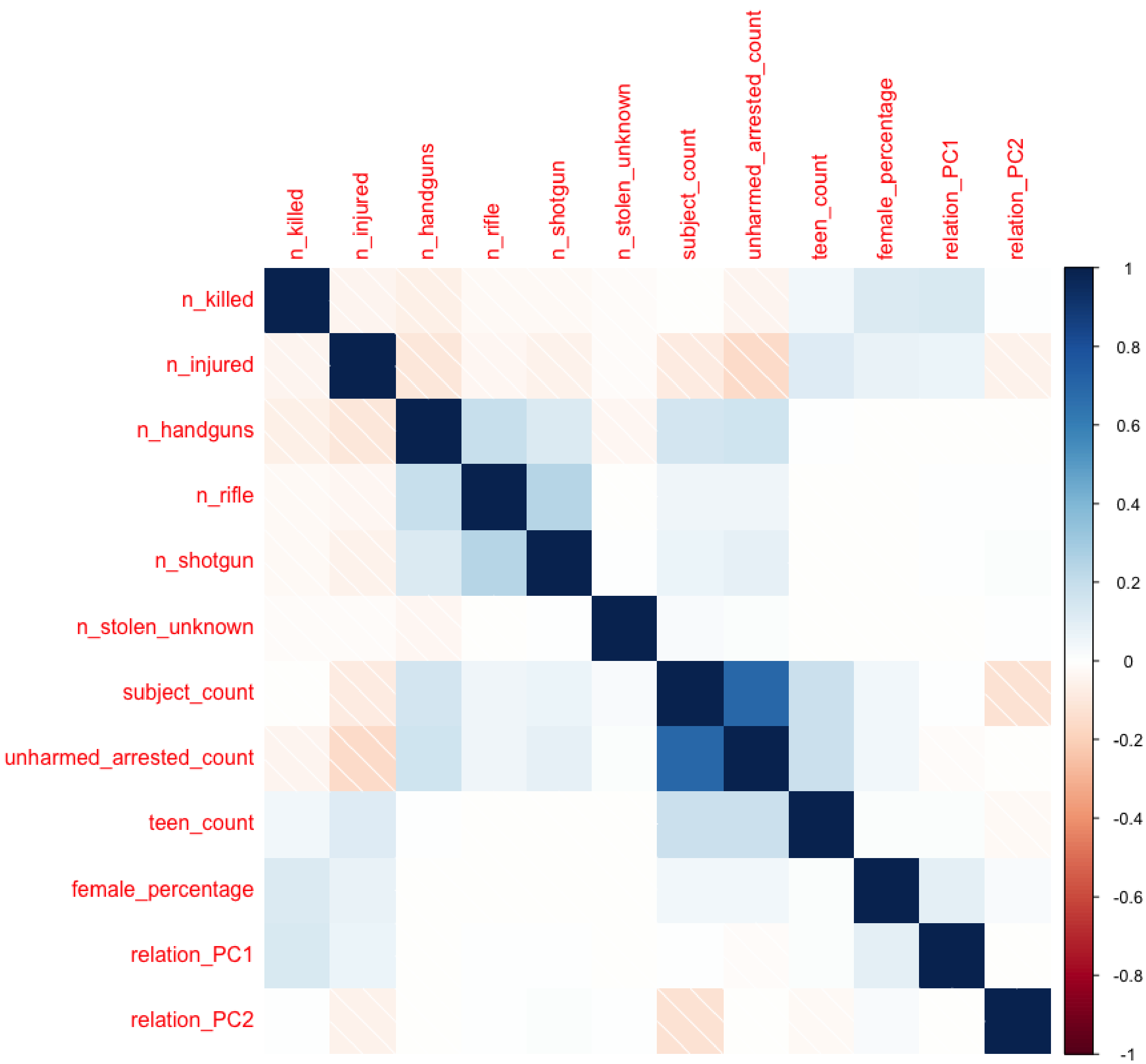


Fig 2A – Within Sum of Squares vs cluster choice 'k'

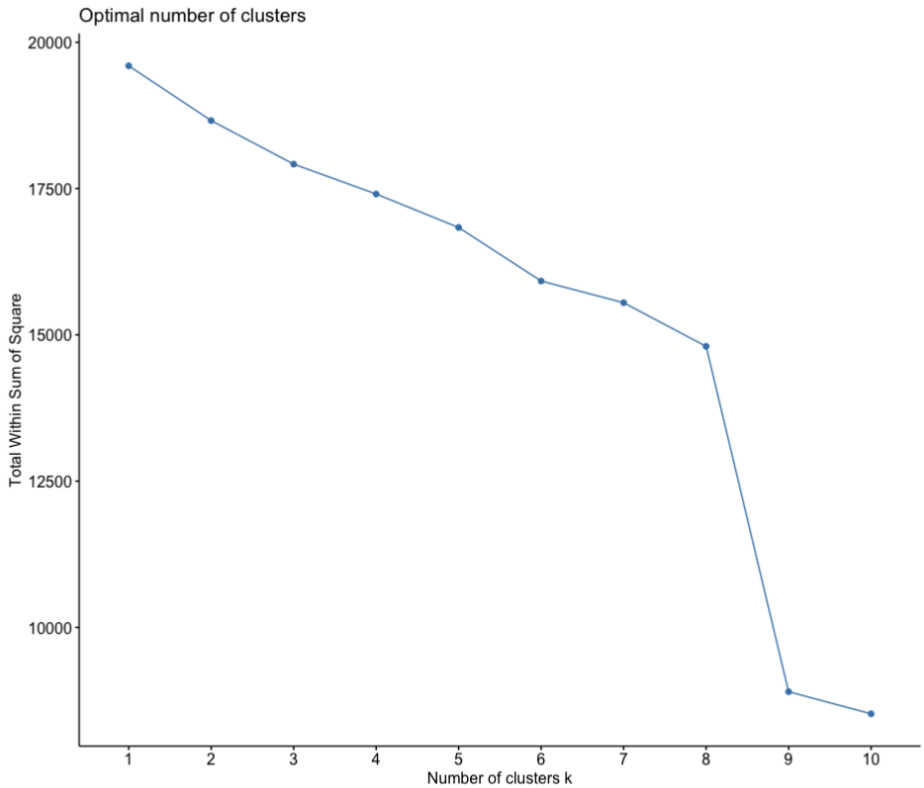


Fig 2B – Gap Statistic vs cluster choice 'k'

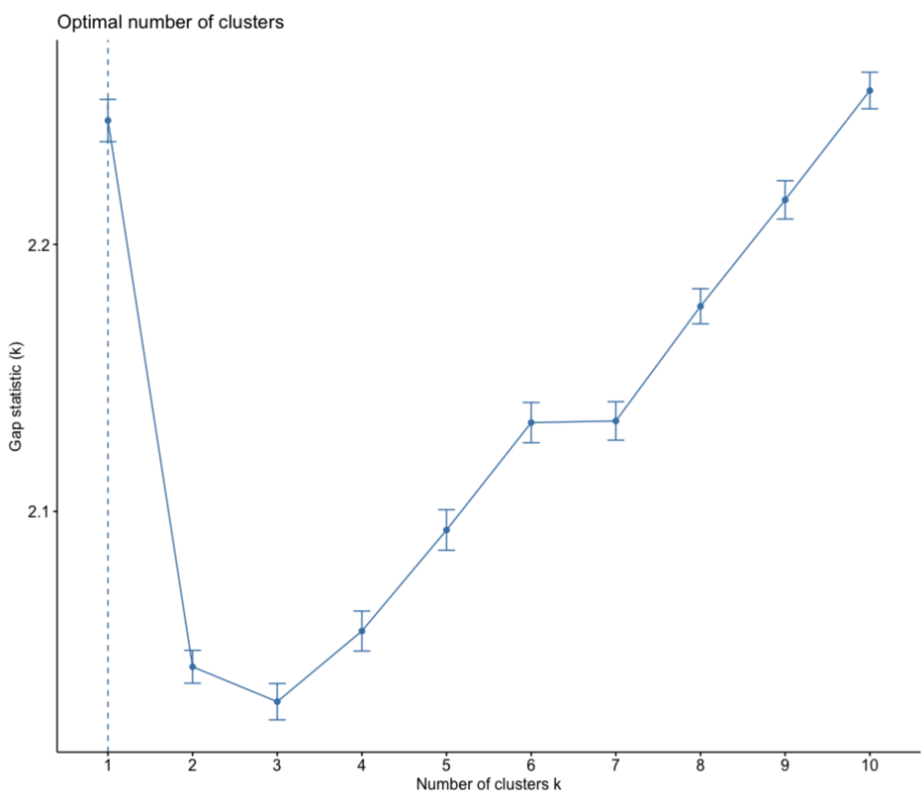
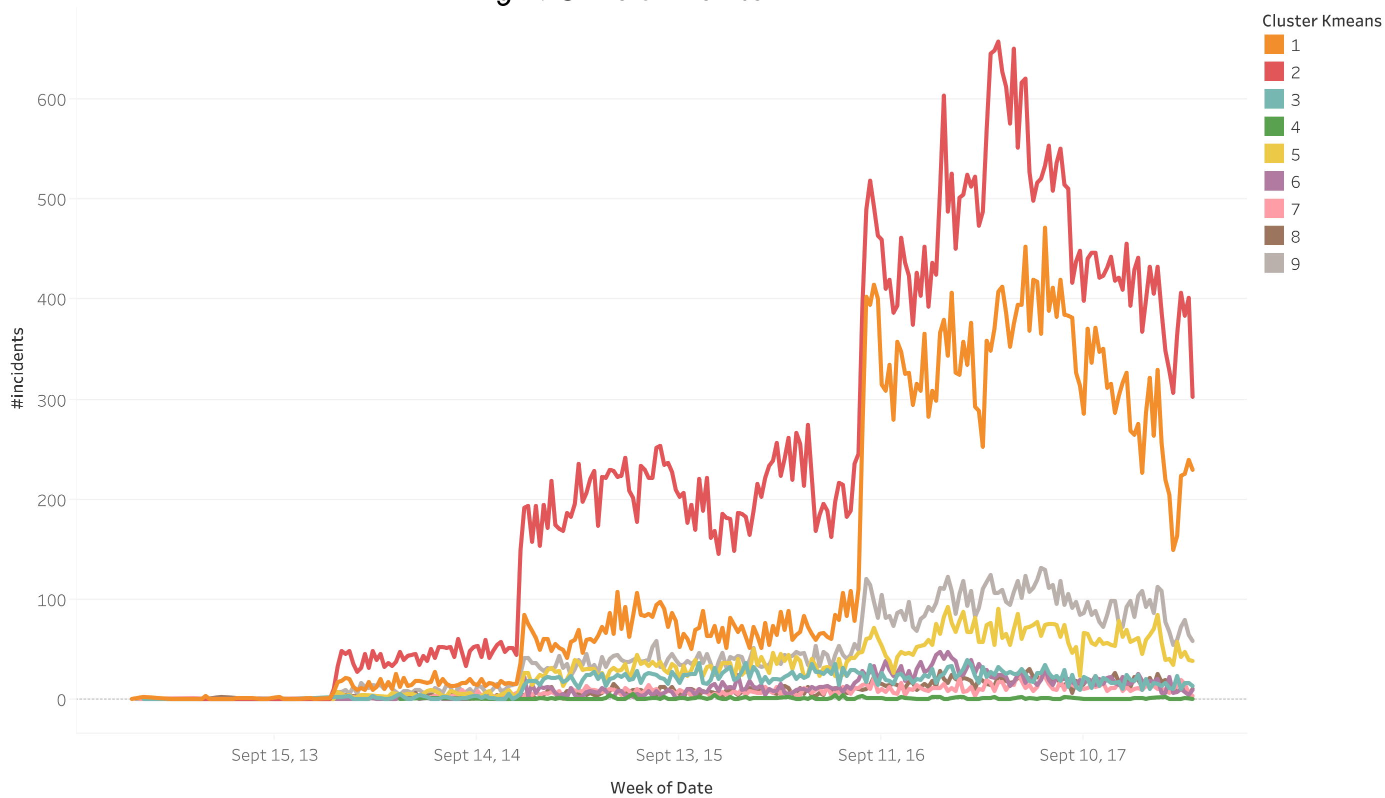


Fig 3 – Visual Representation of Clusters with respect to Eigen Vectors

Cluster 1: Urban Turbulence, smaller-scale urban conflicts.
Cluster 2: Ruthless Warfare, high gun involvement in organized crime.
Cluster 3: Stealthy Offenders, strategic criminal activities with firearms.
Cluster 4: Extreme Outliers, rare incidents with exceptionally high gun use.
Cluster 5: Strained Relationships, conflicts among acquaintances with guns.
Cluster 6: Domestic Disturbance, incidents within families.
Cluster 7: Isolated Incident, infrequent events with moderate violence.
Cluster 8: Tense Workplace, conflicts in professional settings with guns.
Cluster 9: Teen Turmoil, conflicts involving teenagers with moderate violence.

Fig 4: Crime on the rise



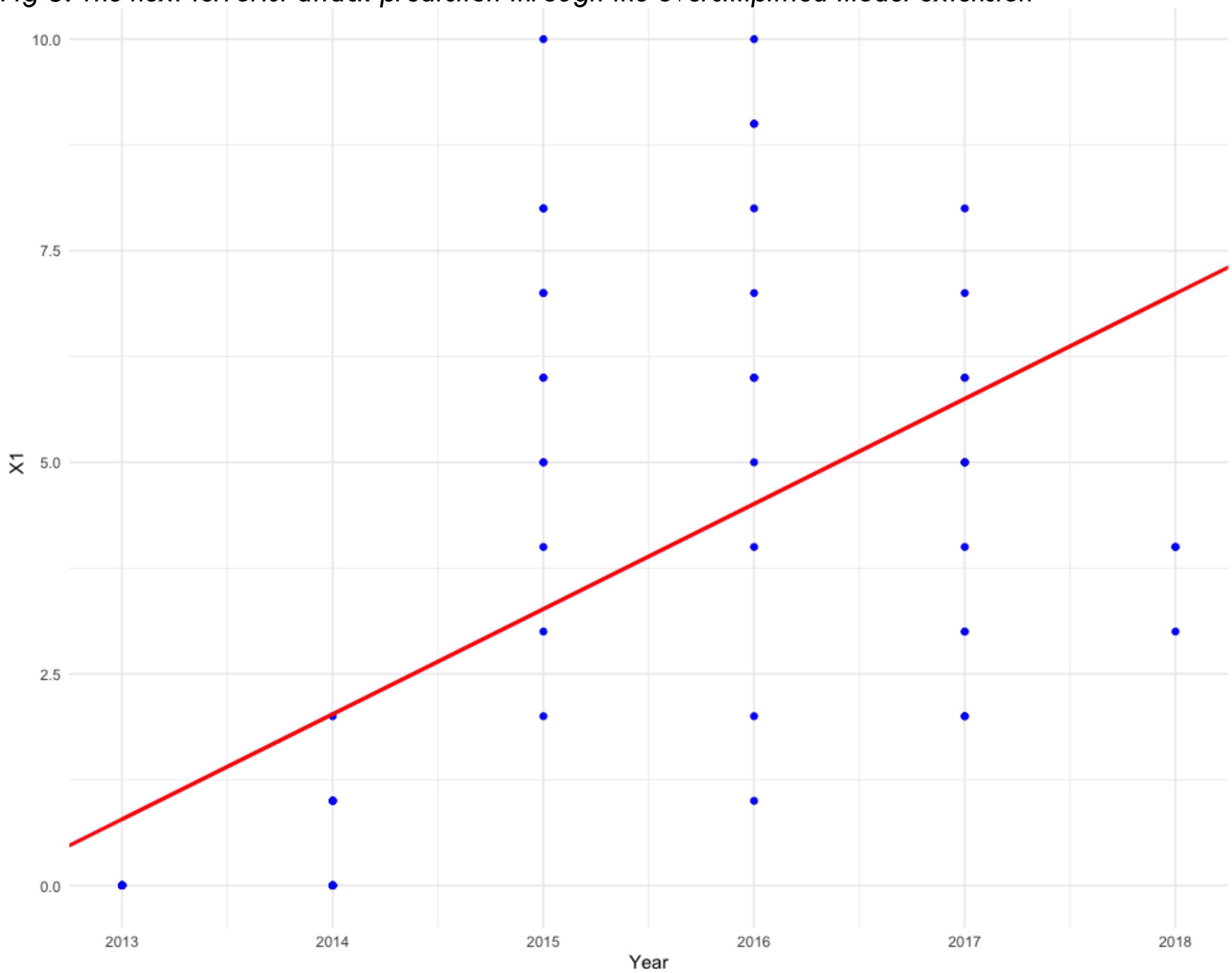
Clusters showing alarming rates of increase:

- *Cluster 1: Urban Turbulence, smaller-scale urban conflicts.*
- *Cluster 2: Ruthless Warfare, high gun involvement in organized crime.*

Clusters showing moderate rates of increase:

- *Cluster 9: Teen Turmoil, conflicts involving teenagers with moderate violence.*
- *Cluster 5: Strained Relationships, conflicts among acquaintances with guns.*

Fig 5. The next terrorist attack prediction through the oversimplified model extension



The X axis represents year and Y axis represents cases of cluster 4, classified as a terrorist or mass shooting. According to this simplified regression, the prediction for 2024 is

$$= \quad \text{Intercept} + \text{slope} * 2024$$
$$= \quad 12$$

which is in common sense too high of a prediction, but this figure serves as a resource for future projects which can be extended from the clustering work done in this project to help governments prepare for what is to come.

Table 1
The table represents the mean of features for each cluster except the first row which represents the size of the cluster, or in other words the number of instances.

Cluster Profiles

	1	2	3	4	5	6	7	8	9
size	35,588	60,564	4,084	217	8,309	2,904	1,987	2,396	12,242
mean_n_killed	0.06	0.29	0.16	0.01	0.18	0.23	0.57	0.32	0.58
mean_n_injured	1.22	0.01	0.21	0.01	0.22	0.68	0.85	0.80	0.48
mean_n_guns_involved	1.04	1.32	2.27	54.37	2.06	1.22	1.10	1.36	1.09
mean_stolen_count	1.04	1.32	2.27	54.37	2.06	1.22	1.10	1.36	1.09
mean_n_handguns	0.09	0.33	0.45	10.47	0.61	0.35	0.26	0.35	0.22
mean_n_rifle	0.01	0.03	0.19	5.85	0.06	0.02	0.03	0.03	0.02
mean_n_shotgun	0	0	1.18	1.25	0.01	0.02	0.04	0.04	0
mean_n_other	0.94	0.95	0.44	36.79	1.38	0.84	0.77	0.94	0.84
mean_subject_count	0.55	0.72	1.11	1.71	2.82	1.70	0.87	2.06	0.84
mean_victim_count	1.20	0.32	0.40	0.04	0.46	1.01	1.59	1.18	1.06
mean_total_count	1.75	1.03	1.51	1.76	3.27	2.71	2.45	3.24	1.89
mean_injured_count	1.22	0.01	0.21	0.01	0.22	0.68	0.85	0.80	0.48
mean_killed_count	0.06	0.29	0.16	0.01	0.19	0.23	0.57	0.32	0.58
mean_unharmed_arrested_count	0.12	0.39	0.74	1.23	2.39	0.49	0.33	1.46	0.39
mean_unharmed_count	0.42	0.64	1.03	1.46	2.72	1.65	0.92	1.90	0.73
mean_adult_count	1.32	0.86	1.31	1.49	2.74	2.11	1.03	0.71	1.68
mean_teen_count	0.08	0.05	0.06	0.05	0.13	0.15	0.12	2.42	0.07
mean_child_count	0	0	0	0	0	0	1.20	0.01	0
mean_relation_family	0.01	0.01	0.04	0.01	0	0	0.20	0.03	0.06
mean_relation_random_victims	0	0	0	0	0	0	0	0	0
mean_relation_aquaintance	0.01	0	0.01	0	0.01	0	0	0.01	0.01
mean_relation_significant_others	0	0	0.03	0	0	0	0.02	0.01	0.15
mean_relation_armed_robbery	0	0	0	0	0	1	0	0.01	0
mean_relation_gang	0	0	0	0	0.01	0	0	0	0
mean_relation_mass_shooting	0	0	0	0	0	0	0	0	0
mean_relation_knows_victims	0	0	0	0	0	0	0	0	0
mean_relation_co_worker	0	0	0	0	0	0	0	0	0
mean_relation_neighbor	0	0	0.02	0.01	0	0	0	0	0.01
mean_relation_friends	0	0	0.01	0	0	0	0.01	0.04	0.01

Section 7 – Code

```
#==========================================================================#
# libraries
#==========================================================================#

# libary
#install.packages("cluster")
#install.packages("factoextra")

# Load necessary libraries
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(factoextra)

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://g
oo.gl/ve3WBa

library(purrr)
library(ggplot2)
library(reshape2)
library(stargazer)

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary St
atistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

library(cluster)
library(factoextra)
library(cluster)
library(dplyr)
library(stargazer)
library(ggplot2)
library(ggfortify)
```

```r
#==========================================================================#
# read and set project filters
#==========================================================================#

file_path <- "/Users/arham/Downloads/02. MVS/Final Project/Dataset 1 — Gun vi
olence.csv"
gun <- read.csv(file_path)

gun <- subset(gun, n_guns_involved != 0) # filter to gun incidents ,i.e., n_g
uns_involved >1
# filter to relevant columns
gun <- gun[, c("incident_id", "date", "state", "city_or_county", "latitude",
"longitude", "n_killed", "n_injured", "congressional_district", "gun_stolen",
"gun_type", "incident_characteristics",
               "n_guns_involved", "notes", "participant_age",
               "participant_age_group", "participant_gender",
               "participant_relationship","participant_status", "participant_
type")]

#### Prelimary filter to columns which are not repeated or non-redundant
gun <- na.omit(gun)

#==========================================================================#
# read and set project filters
#==========================================================================#

##============================ 1. Stolen Guns


# count number of occurences of stolen of gun_stolen and record in new column
(a cell has values like :0::Unknown||1::Unknown )
gun$stolen_count <- sapply(gun$gun_stolen, function(x) length(strsplit(x, spl
it = "\\|\\|")[[1]]))
# remove gun_stolen column
gun <- gun[, !(names(gun) %in% "gun_stolen")]

##============================ 2. gun_types



# First, convert all the text in the "gun_type" column to lowercase
gun$gun_type <- tolower(gun$gun_type)
# Count the number of occurrences of specific keywords in the "gun_type" colu
mn, split using "||"
gun$n_handguns <- sapply(gun$gun_type, function(x) length(grep("handgun", str
split(x, split = "\\|\\|")[[1]])))
gun$n_auto <- sapply(gun$gun_type, function(x) length(grep("auto", strsplit(x
, split = "\\|\\|")[[1]])))
```

```
gun$n_mm <- sapply(gun$gun_type, function(x) length(grep("mm", strsplit(x, split = "\\|\\|")[[1]])))
gun$n_spl <- sapply(gun$gun_type, function(x) length(grep("spl", strsplit(x, split = "\\|\\|")[[1]])))
gun$n_mag <- sapply(gun$gun_type, function(x) length(grep("mag", strsplit(x, split = "\\|\\|")[[1]])))
# Sum up counts and remove unnecessary columns
gun$n_handguns <- gun$n_handguns + gun$n_auto + gun$n_mm + gun$n_spl + gun$n_mag
gun <- gun[, !(names(gun) %in% c("n_auto", "n_mm", "n_spl", "n_mag", "n_win"))]
# Count occurrences of 'win' and 'rifle', adjust counts, sum them up, and remove unnecessary columns
gun$n_win <- sapply(gun$gun_type, function(x) length(grep("win", strsplit(x, split = "\\|\\|")[[1]])))

gun$n_rifle <- sapply(gun$gun_type, function(x) length(grep("rifle", strsplit(x, split = "\\|\\|")[[1]])))
gun$n_rifle <- gun$n_rifle + gun$n_win
gun <- gun[, !(names(gun) %in% c("n_win"))]
# Count occurrences of 'gauge' and 'shotgun', adjust counts, sum them up, and remove unnecessary columns
gun$n_gauge <- sapply(gun$gun_type, function(x) length(grep("gauge", strsplit(x, split = "\\|\\|")[[1]])))
gun$n_shotgun <- sapply(gun$gun_type, function(x) length(grep("shotgun", strsplit(x, split = "\\|\\|")[[1]])))
gun$n_shotgun <- gun$n_shotgun + gun$n_gauge
gun <- gun[, !(names(gun) %in% c("n_gauge"))]
# Count the number of occurrences of "||", add 1 to it, and subtract counts of shotguns, rifles, and handguns
gun$n_other <- sapply(gun$gun_type, function(x) sum(gregexpr("\\|\\|", x)[[1]] > 0))
gun$n_other <- gun$n_other - gun$n_shotgun - gun$n_rifle - gun$n_handguns + 1
# remove gun_type column
gun <- gun[, !(names(gun) %in% "gun_type")]


##============================ 3.  number of suspects, victims, and total people involved
# count number of Subject-Suspect, Victim, and Total people involved
gun$subject_count <- sapply(gun$participant_type, function(x) length(grep("Subject-Suspect", strsplit(x, split = "\\|\\|")[[1]])))
gun$victim_count <- sapply(gun$participant_type, function(x) length(grep("Victim", strsplit(x, split = "\\|\\|")[[1]])))
gun$total_count <- sapply(gun$participant_type, function(x) length(strsplit(x, split = "\\|\\|")[[1]]))
#remove participant_type column
gun <- gun[, !(names(gun) %in% "participant_type")]
```

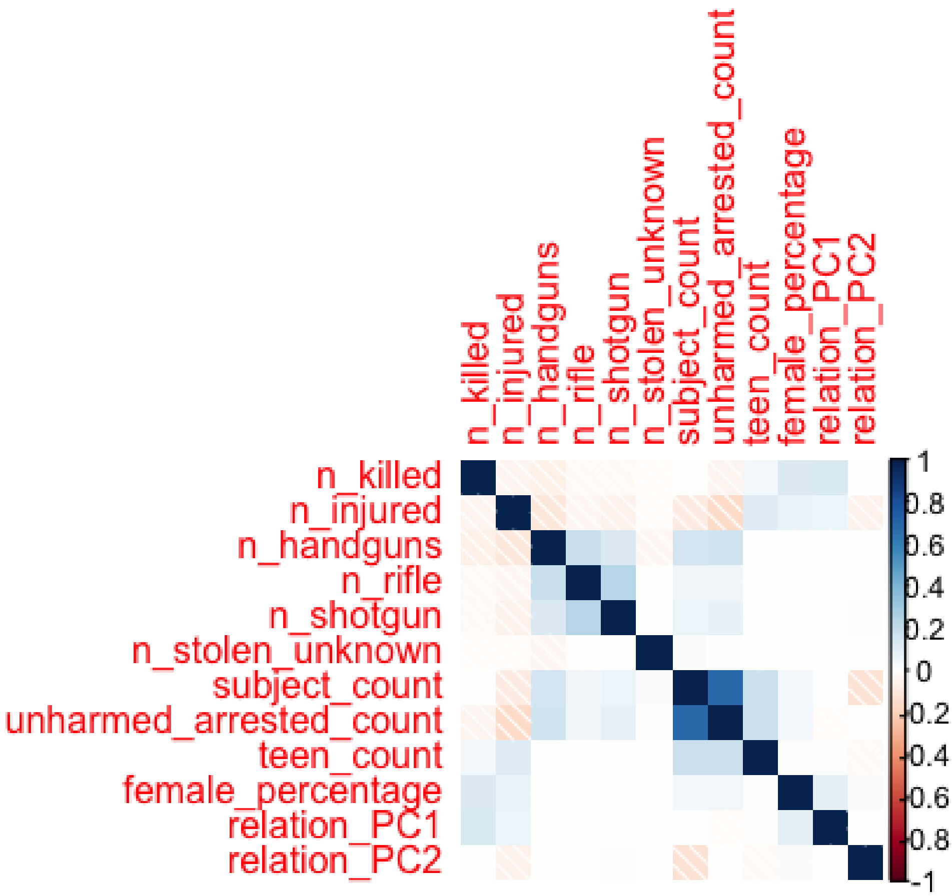
```r
##============================ 4. number of injured, killed, unharmed arreste
d, and unharmed
# from participant_status column count number of Injured, Killed, and 'Unharm
ed, Arrested', and 'Unharmed'
gun$injured_count <- sapply(gun$participant_status, function(x) length(grep("
Injured", strsplit(x, split = "\\|\\|")[[1]])))
gun$killed_count <- sapply(gun$participant_status, function(x) length(grep("K
illed", strsplit(x, split = "\\|\\|")[[1]])))
gun$unharmed_arrested_count <- sapply(gun$participant_status, function(x) len
gth(grep("Unharmed, Arrested", strsplit(x, split = "\\|\\|")[[1]])))
gun$unharmed_count <- sapply(gun$participant_status, function(x) length(grep(
"Unharmed", strsplit(x, split = "\\|\\|")[[1]])))


# from participant age group count number of Adult 18+ and Teen 12-17, child
0-11
gun$adult_count <- sapply(gun$participant_age_group, function(x) length(grep(
"Adult 18+", strsplit(x, split = "\\|\\|")[[1]])))
gun$teen_count <- sapply(gun$participant_age_group, function(x) length(grep("
Teen 12-17", strsplit(x, split = "\\|\\|")[[1]])))
gun$child_count <- sapply(gun$participant_age_group, function(x) length(grep(
"Child 0-11", strsplit(x, split = "\\|\\|")[[1]])))
# remove participant_age_group column
gun <- gun[, !(names(gun) %in% "participant_age_group")]


##============================ 5. gender ratio

calculate_female_percentage <- function(participant_gender) {
  genders <- strsplit(participant_gender, "\\|\\|")[[1]]
  total_participants <- length(genders)
  female_count <- sum(grepl("Female", genders))

  if (total_participants > 0) {
    return((female_count / total_participants) * 100)
  } else {
    return(NA)
  }
}

##============================ 6. relations between participants
gun$participant_relationship <- tolower(gun$participant_relationship)
# Family
# Random victims
# Aquaintance
# Significant Others
# Armed Robbery
# Gang
# Mass Shooting
# Knows victims
```

```r
# Co-worker
# Neighbor
# Friends
# Home Invasion
# Does Not Know Victim

# New column relation_family =  if family is present in participant_relations
hip column, then 1 else 0
gun <- gun %>% mutate(relation_family = ifelse(grepl("family", participant_re
lationship, ignore.case = TRUE), 1, 0))
# New column relation_random_victims =  if random victims is present in parti
cipant_relationship column, then 1 else 0
gun <- gun %>% mutate(relation_random_victims = ifelse(grepl("random victims"
, participant_relationship, ignore.case = TRUE), 1, 0))
# New column relation_aquaintance =  if aquaintance is present in participant
_relationship column, then 1 else 0
gun <- gun %>% mutate(relation_aquaintance = ifelse(grepl("aquaintance", part
icipant_relationship, ignore.case = TRUE), 1, 0))
# New column relation_significant_others =  if significant others is present
in participant_relationship column, then 1 else 0
gun <- gun %>% mutate(relation_significant_others = ifelse(grepl("significant
others", participant_relationship, ignore.case = TRUE), 1, 0))
# New column relation_armed_robbery =  if armed robbery is present in partici
pant_relationship column, then 1 else 0
gun <- gun %>% mutate(relation_armed_robbery = ifelse(grepl("armed robbery",
participant_relationship, ignore.case = TRUE), 1, 0))
# New column relation_gang =  if gang is present in participant_relationship
column, then 1 else 0
gun <- gun %>% mutate(relation_gang = ifelse(grepl("gang", participant_relati
onship, ignore.case = TRUE), 1, 0))
# New column relation_mass_shooting =  if mass shooting is present in partici
pant_relationship column, then 1 else 0
gun <- gun %>% mutate(relation_mass_shooting = ifelse(grepl("mass shooting",
participant_relationship, ignore.case = TRUE), 1, 0))
# New column relation_knows_victims =  if knows victims is present in partici
pant_relationship column, then 1 else 0
gun <- gun %>% mutate(relation_knows_victims = ifelse(grepl("knows victims",
participant_relationship, ignore.case = TRUE), 1, 0))
# New column relation_co_worker =  if co-worker is present in participant_rel
ationship column, then 1 else 0
gun <- gun %>% mutate(relation_co_worker = ifelse(grepl("co-worker", particip
ant_relationship, ignore.case = TRUE), 1, 0))
# New column relation_neighbor =  if neighbor is present in participant_relat
ionship column, then 1 else 0
gun <- gun %>% mutate(relation_neighbor = ifelse(grepl("neighbor", participan
t_relationship, ignore.case = TRUE), 1, 0))
# New column relation_friends =  if friends is present in participant_relatio
nship column, then 1 else 0
gun <- gun %>% mutate(relation_friends = ifelse(grepl("friends", participant_
relationship, ignore.case = TRUE), 1, 0))
```

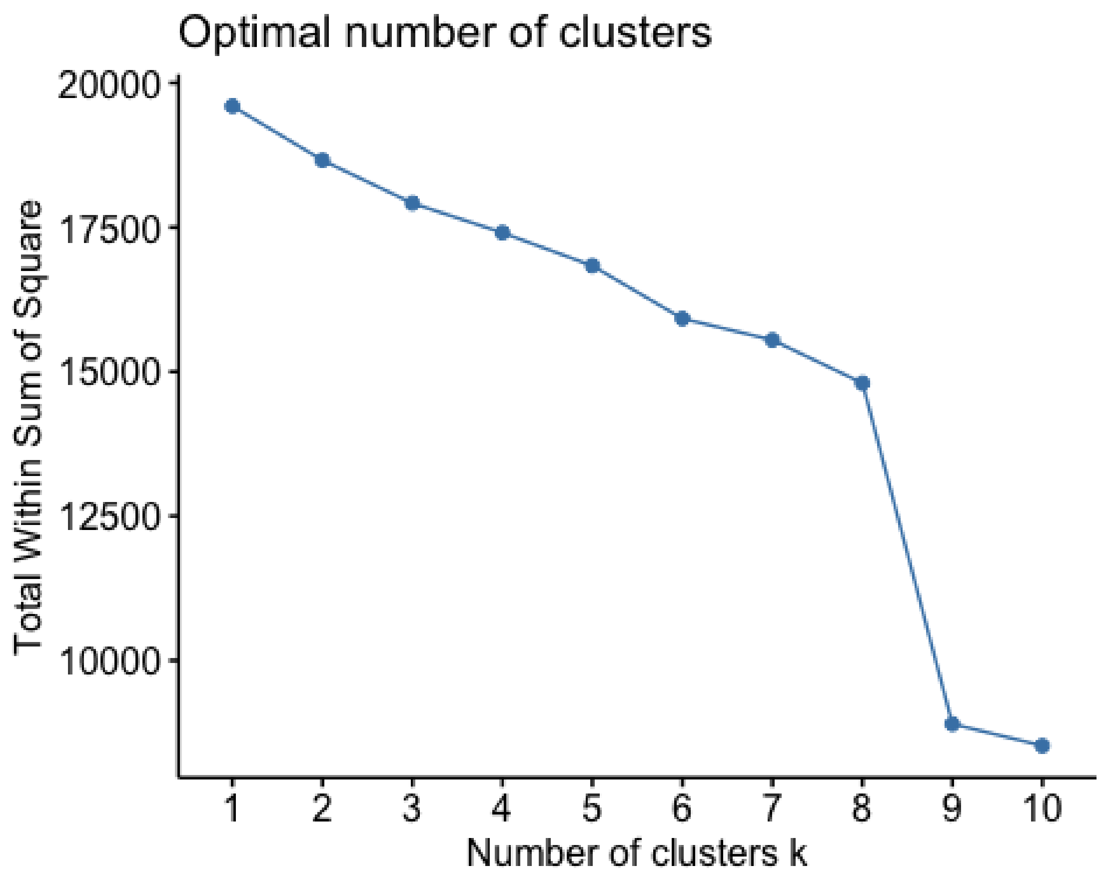
```r
# New column relation_home_invasion =  if home invasion is present in partici
pant_relationship column, then 1 else 0
gun <- gun %>% mutate(relation_home_invasion = ifelse(grepl("home invasion",
participant_relationship, ignore.case = TRUE), 1, 0))
# New column relation_does_not_know_victim =  if does not know victim is pres
ent in participant_relationship column, then 1 else 0
gun <- gun %>% mutate(relation_does_not_know_victim = ifelse(grepl("does not
know victim", participant_relationship, ignore.case = TRUE), 1, 0))
# remove participant_relationship column
gun <- gun[, !(names(gun) %in% "participant_relationship")]










#==========================================================================#
# More Feature Engineering
#==========================================================================#

# Apply the function to create a new column for female percentage
gun <- gun %>%
  mutate(female_percentage = sapply(participant_gender, calculate_female_perc
entage))
# remove participant_status column
gun <- gun[, !(names(gun) %in% "participant_status")]


# save as gun_preprocessed_v1.csv
write.csv(gun, file = "gun_preprocessed_vF.csv", row.names = FALSE)
#gun <- read.csv("gun_preprocessed_vF.csv")

tempo = gun

selected_columns <- c(
  "child_count",
  "relation_family",
  "relation_random_victims",
  "relation_aquaintance",
  "relation_significant_others",
  "relation_armed_robbery",
  "relation_gang",
  "relation_mass_shooting",
  "relation_knows_victims",
  "relation_co_worker",
  "relation_neighbor",
```

```r
  "relation_friends",
  "relation_home_invasion",
  "relation_does_not_know_victim"
)

gun_pca_result <- prcomp(gun[, selected_columns])
principal_components <- as.data.frame(gun_pca_result$x[, 1:2])
names(principal_components) <- c("relation_PC1", "relation_PC2")
gun <- cbind(gun[, -which(names(gun) %in% selected_columns)], principal_compo
nents)
gun <- gun[, !(names(gun) %in% selected_columns)]

gun <- gun[, !(names(gun) %in% c("participant_age","congressional_district",
"state", "city_or_county", "latitude", "longitude"))]
exclude_columns <- c('incident_id', 'date', 'notes', 'incident_characteristic
s', 'gun_type', 'participant_relationship', 'location_description',
                     'participant_gender', 'adult_count', 'gun_stolen','unhar
med_count','stolen_count', 'victim_count', 'total_count', 'injured_count', 'k
illed_count','n_guns_involved')
X <- gun[, !(names(gun) %in% exclude_columns)]


names(gun)

##  [1] "incident_id"                "date"
##  [3] "n_killed"                   "n_injured"
##  [5] "incident_characteristics" "n_guns_involved"
##  [7] "notes"                      "participant_gender"
##  [9] "stolen_count"               "n_handguns"
## [11] "n_rifle"                    "n_shotgun"
## [13] "n_other"                    "subject_count"
## [15] "victim_count"               "total_count"
## [17] "injured_count"              "killed_count"
## [19] "unharmed_arrested_count"  "unharmed_count"
## [21] "adult_count"                "teen_count"
## [23] "female_percentage"          "relation_PC1"
## [25] "relation_PC2"

#rename n_other to n_stolen_unkown
names(X)[names(X) == 'n_other'] <- 'n_stolen_unknown'

# X[is.infinite(X)] <- 0
X[is.na(X)] <- 0
OG <-X
X <- scale(X)
X <- as.data.frame(X)
W <- X
Z <- X
```

```
######### correlation
library(corrplot)

## corrplot 0.92 loaded

corr <- cor(X)
corrplot(corr, method = "shade")
```



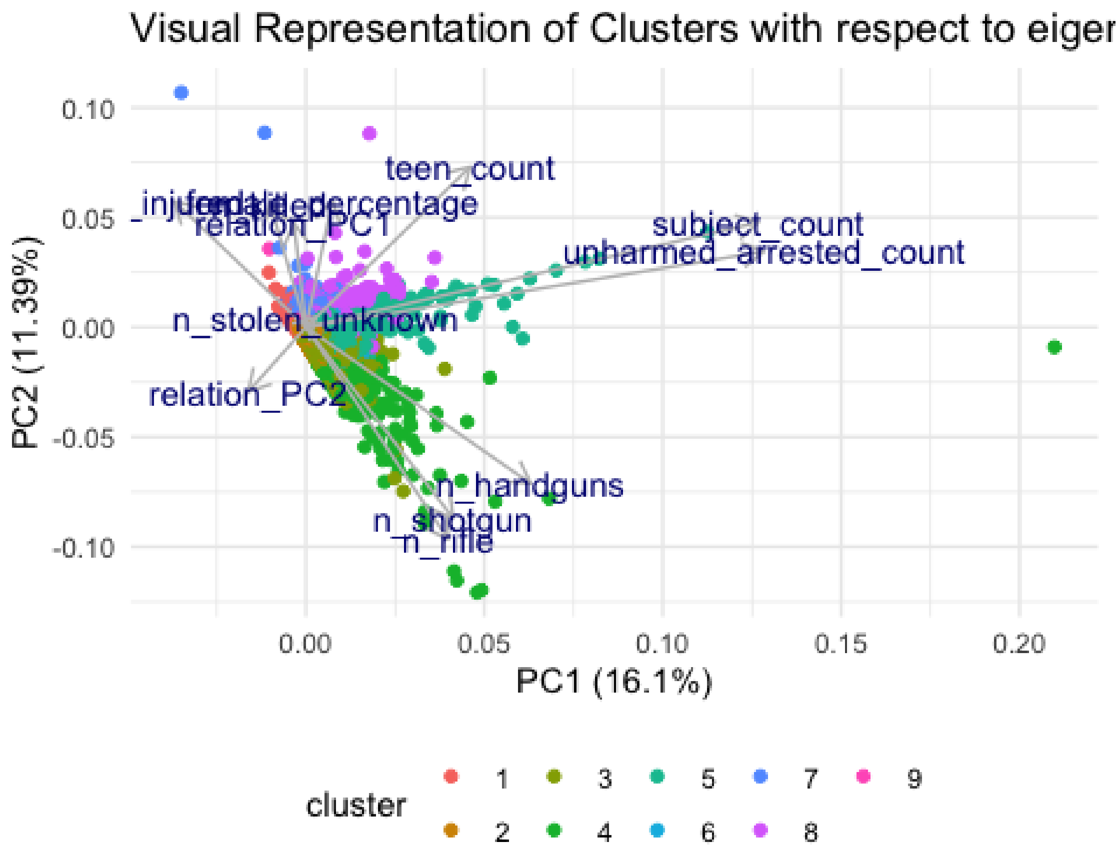

```
# ==========================================================================#
## K Means Clustering
# ==========================================================================#

W <- X

#================= What should be the k?

## support 1
set.seed(123)  # for reproducibility
sample_indices <- sample(nrow(W), 1000)  # adjust the size as needed
subset_W <- W[sample_indices, ]
par(pty = "m")
fviz_nbclust(subset_W, pam, method = "wss")
```




```
## support 2
#calculate gap statistic based on number of clusters
##gap_stat <- clusGap(subset_W,
                    #FUN = pam,
                    #K.max = 10, #max clusters to consider
                    #B = 50) #total bootstrapped iterations

###plot number of clusters vs. gap statistic
##viz_gap_stat(gap_stat)

# k = 9


#================= apply k means for 9 clusters


set.seed(123)
kmeans_W <- kmeans(W, centers = 9, nstart = 25)
kmeans_W$cluster <- as.factor(kmeans_W$cluster)
# add cluster column to W
W$cluster <- as.factor(kmeans_W$cluster)
OG$cluster <- as.factor(kmeans_W$cluster)
```

```
summary_by_cluster <- OG %>%
  group_by(cluster) %>%
  summarise(
    size = n(),
    mean_n_killed = round(mean(n_killed),2),
    mean_n_injured = round(mean(n_injured),2),
    mean_n_handguns = round(mean(n_handguns),2),
    mean_n_rifle = round(mean(n_rifle),2),
    mean_n_shotgun = round(mean(n_shotgun),2),
    mean_n_stolen_unknown = round(mean(n_stolen_unknown),2),
    mean_subject_count = round(mean(subject_count),2),
    mean_unharmed_arrested_count = round(mean(unharmed_arrested_count),2),
    mean_teen_count = round(mean(teen_count),2),
    mean_female_percentage = round(mean(female_percentage),2),
    mean_relation_PC1 = round(mean(as.numeric(relation_PC1)),2),
    mean_relation_PC2 = round(mean(as.numeric(relation_PC2)),2)
  )



summary_by_cluster = data.frame(summary_by_cluster)
summary_by_cluster = t(summary_by_cluster)
colnames(summary_by_cluster) <- summary_by_cluster[1, ]
summary_by_cluster <- summary_by_cluster[-1, ]


stargazer(summary_by_cluster, title = "Cluster Profiles",type = "html", digit
s = 2)

##
## <table style="text-align:center"><caption><strong>Cluster Profiles</strong
></caption>
## <tr><td colspan="10" style="border-bottom: 1px solid black"></td></tr><tr>
<td style="text-align:left"></td><td>1</td><td>2</td><td>3</td><td>4</td><td>
5</td><td>6</td><td>7</td><td>8</td><td>9</td></tr>
## <tr><td colspan="10" style="border-bottom: 1px solid black"></td></tr><tr>
<td style="text-align:left">size</td><td>35588</td><td>60564</td><td>4084</td
><td>217</td><td>8309</td><td>2904</td><td>1987</td><td>2396</td><td>12242</t
d></tr>
## <tr><td style="text-align:left">mean_n_killed</td><td>0.06</td><td>0.29</t
d><td>0.16</td><td>0.01</td><td>0.18</td><td>0.23</td><td>0.57</td><td>0.32</
td><td>0.58</td></tr>
## <tr><td style="text-align:left">mean_n_injured</td><td>1.22</td><td>0.01</
td><td>0.21</td><td>0.01</td><td>0.22</td><td>0.68</td><td>0.85</td><td>0.80<
/td><td>0.48</td></tr>
## <tr><td style="text-align:left">mean_n_handguns</td><td>0.09</td><td>0.33<
/td><td>0.45</td><td>10.47</td><td>0.61</td><td>0.35</td><td>0.26</td><td>0.3
5</td><td>0.22</td></tr>
## <tr><td style="text-align:left">mean_n_rifle</td><td>0.01</td><td>0.03</td
><td>0.19</td><td>5.85</td><td>0.06</td><td>0.02</td><td>0.03</td><td>0.03</t
```

```
d><td>0.02</td></tr>
## <tr><td style="text-align:left">mean_n_shotgun</td><td>0.00</td><td>0.00</
td><td>1.18</td><td>1.25</td><td>0.01</td><td>0.02</td><td>0.04</td><td>0.04<
/td><td>0.00</td></tr>
## <tr><td style="text-align:left">mean_n_stolen_unknown</td><td>0.94</td><td
>0.95</td><td>0.44</td><td>36.79</td><td>1.38</td><td>0.84</td><td>0.77</td><
td>0.94</td><td>0.84</td></tr>
## <tr><td style="text-align:left">mean_subject_count</td><td>0.55</td><td>0.
72</td><td>1.11</td><td>1.71</td><td>2.82</td><td>1.70</td><td>0.87</td><td>2
.06</td><td>0.84</td></tr>
## <tr><td style="text-align:left">mean_unharmed_arrested_count</td><td>0.12<
/td><td>0.39</td><td>0.74</td><td>1.23</td><td>2.39</td><td>0.49</td><td>0.33
</td><td>1.46</td><td>0.39</td></tr>
## <tr><td style="text-align:left">mean_teen_count</td><td>0.08</td><td>0.05<
/td><td>0.06</td><td>0.05</td><td>0.13</td><td>0.15</td><td>0.12</td><td>2.42
</td><td>0.07</td></tr>
## <tr><td style="text-align:left">mean_female_percentage</td><td>1.98</td><t
d>0.21</td><td>7.95</td><td>4.71</td><td>11.92</td><td>7.89</td><td>24.68</td
><td>9.76</td><td>67.77</td></tr>
## <tr><td style="text-align:left">mean_relation_PC1</td><td>-0.02</td><td>-0
.02</td><td>-0.01</td><td>-0.02</td><td>-0.02</td><td>-0.10</td><td>1.17</td>
<td>-0.01</td><td>0.00</td></tr>
## <tr><td style="text-align:left">mean_relation_PC2</td><td>0.02</td><td>0.0
2</td><td>0.03</td><td>0.02</td><td>0.02</td><td>-0.97</td><td>-0.08</td><td>
0.01</td><td>0.04</td></tr>
## <tr><td colspan="10" style="border-bottom: 1px solid black"></td></tr></ta
ble>

# pca on W excluding cluster column
pca <- prcomp(W[, -ncol(W)], scale = TRUE)
# autoplot(pca, data = W[, -1], colour = 'cluster', loadings = TRUE, loadings
.label = TRUE) +
#   theme_minimal() +
#   theme(legend.position = 'bottom') +
#   ggtitle("PCA Colored by Cluster")


# autoplot clusters
autoplot(pca, data = W[, -1], colour = 'cluster', loadings = TRUE,loadings.la
bel = TRUE,loadings.color = 'grey', loadings.label.color = 'navyblue') +
  theme_minimal() +
  theme(legend.position = 'bottom') +
  ggtitle("Visual Representation of Clusters with respect to eigen vectors")
```




```
# # autoplot clusters
# autoplot(pca, data = W[, -1], colour = 'cluster', loadings = TRUE, loadings
.label = TRUE,loadings.color = 'black', loadings.label.color = 'black') +
#   theme_minimal() +
#   theme(legend.position = 'bottom') +
#   ggtitle("PCA Colored by Cluster")
#
# #autoplot(pca, data = W[, -1], loadings = TRUE, loadings.label = TRUE) +
#   theme_minimal() +
#   theme(legend.position = 'bottom') +
#   ggtitle("Crime Scene Incident Split")


# add cluster names to OG by incident id
tempo$cluster_kmeans <- W$cluster

guns_f = tempo

centroid_col_list <- c(
  "n_killed",
  "n_injured",
  "n_guns_involved",
  "stolen_count",
```

```r
  "n_handguns",
  "n_rifle",
  "n_shotgun",
  "n_other",
  "subject_count",
  "victim_count",
  "total_count",
  "injured_count",
  "killed_count",
  "unharmed_arrested_count",
  "unharmed_count",
  "adult_count",
  "teen_count",
  "child_count",
  "relation_family",
  "relation_random_victims",
  "relation_aquaintance",
  "relation_significant_others",
  "relation_armed_robbery",
  "relation_gang",
  "relation_mass_shooting",
  "relation_knows_victims",
  "relation_co_worker",
  "relation_neighbor",
  "relation_friends",
  "relation_home_invasion",
  "relation_does_not_know_victim",
  "female_percentage",
  "cluster_kmeans")

# filter guns_f to only columns in centroid_col_list
guns_f1 <- guns_f[, centroid_col_list]

guns_f1[is.na(guns_f1)] <- 0

# convert all columns to numeric
guns_f1 <- sapply(guns_f1, as.numeric)

guns_f1 = data.frame(guns_f1)

summary_by_cluster_f1 <- guns_f1 %>% group_by(cluster_kmeans) %>% summarise(
  size = n(),
  mean_n_killed = round(mean(n_killed), 2),
  mean_n_injured = round(mean(n_injured), 2),
  mean_n_guns_involved = round(mean(n_guns_involved), 2),
  mean_stolen_count = round(mean(stolen_count), 2),
  mean_n_handguns = round(mean(n_handguns), 2),
  mean_n_rifle = round(mean(n_rifle), 2),
  mean_n_shotgun = round(mean(n_shotgun), 2),
  mean_n_other = round(mean(n_other), 2),
```

```
  mean_subject_count = round(mean(subject_count), 2),
  mean_victim_count = round(mean(victim_count), 2),
  mean_total_count = round(mean(total_count), 2),
  mean_injured_count = round(mean(injured_count), 2),
  mean_killed_count = round(mean(killed_count), 2),
  mean_unharmed_arrested_count = round(mean(unharmed_arrested_count), 2),
  mean_unharmed_count = round(mean(unharmed_count), 2),
  mean_adult_count = round(mean(adult_count), 2),
  mean_teen_count = round(mean(teen_count), 2),
  mean_child_count = round(mean(child_count), 2),
  mean_relation_family = round(mean(relation_family), 2),
  mean_relation_random_victims = round(mean(relation_random_victims), 2),
  mean_relation_aquaintance = round(mean(relation_aquaintance), 2),
  mean_relation_significant_others = round(mean(relation_significant_others),
2),
  mean_relation_armed_robbery = round(mean(relation_armed_robbery), 2),
  mean_relation_gang = round(mean(relation_gang), 2),
  mean_relation_mass_shooting = round(mean(relation_mass_shooting), 2),
  mean_relation_knows_victims = round(mean(relation_knows_victims), 2),
  mean_relation_co_worker = round(mean(relation_co_worker), 2),
  mean_relation_neighbor = round(mean(relation_neighbor), 2),
  mean_relation_friends = round(mean(relation_friends), 2),
  mean_relation_home_invasion = round(mean(relation_home_invasion), 2)
)
summary_by_cluster_f1 = t(summary_by_cluster_f1)

colnames(summary_by_cluster_f1) <- summary_by_cluster_f1[1, ]
summary_by_cluster_f1 <- summary_by_cluster_f1[-1, ]
summary_by_cluster_f1 <- summary_by_cluster_f1[-nrow(summary_by_cluster_f1),
]
summary_by_cluster_f1 <- round(summary_by_cluster_f1, 2)

# stargazer summary_by_cluster_f1
stargazer(summary_by_cluster_f1, title = "Cluster Profiles",type = "text",col
umn.sep.width = "5pt", digits = 2)

##
## Cluster Profiles
## ============================================================================
===============
##                                    1      2      3     4     5     6     7
8     9
## ----------------------------------------------------------------------------
---------------
## size                             35,588 60,564 4,084  217  8,309 2,904 1,9
87 2,396 12,242
## mean_n_killed                     0.06   0.29  0.16  0.01  0.18  0.23  0.5
7  0.32   0.58
## mean_n_injured                    1.22   0.01  0.21  0.01  0.22  0.68  0.8
5  0.80   0.48
```

```
## mean_n_guns_involved               1.04   1.32  2.27  54.37 2.06  1.22  1.1
0  1.36   1.09
## mean_stolen_count                  1.04   1.32  2.27  54.37 2.06  1.22  1.1
0  1.36   1.09
## mean_n_handguns                    0.09   0.33  0.45  10.47 0.61  0.35  0.2
6  0.35   0.22
## mean_n_rifle                       0.01   0.03  0.19  5.85  0.06  0.02  0.0
3  0.03   0.02
## mean_n_shotgun                      0      0    1.18  1.25  0.01  0.02  0.0
4  0.04    0
## mean_n_other                       0.94   0.95  0.44  36.79 1.38  0.84  0.7
7  0.94   0.84
## mean_subject_count                 0.55   0.72  1.11  1.71  2.82  1.70  0.8
7  2.06   0.84
## mean_victim_count                  1.20   0.32  0.40  0.04  0.46  1.01  1.5
9  1.18   1.06
## mean_total_count                   1.75   1.03  1.51  1.76  3.27  2.71  2.4
5  3.24   1.89
## mean_injured_count                 1.22   0.01  0.21  0.01  0.22  0.68  0.8
5  0.80   0.48
## mean_killed_count                  0.06   0.29  0.16  0.01  0.19  0.23  0.5
7  0.32   0.58
## mean_unharmed_arrested_count       0.12   0.39  0.74  1.23  2.39  0.49  0.3
3  1.46   0.39
## mean_unharmed_count                0.42   0.64  1.03  1.46  2.72  1.65  0.9
2  1.90   0.73
## mean_adult_count                   1.32   0.86  1.31  1.49  2.74  2.11  1.0
3  0.71   1.68
## mean_teen_count                    0.08   0.05  0.06  0.05  0.13  0.15  0.1
2  2.42   0.07
## mean_child_count                    0      0     0     0     0     0    1.2
0  0.01    0
## mean_relation_family               0.01   0.01  0.04  0.01    0     0    0.2
0  0.03   0.06
## mean_relation_random_victims        0      0     0     0     0     0     0
0     0
## mean_relation_aquaintance          0.01    0    0.01    0   0.01    0     0
0.01   0.01
## mean_relation_significant_others   0      0    0.03    0     0     0    0.0
2  0.01   0.15
## mean_relation_armed_robbery         0      0     0     0     0     1     0
0.01    0
## mean_relation_gang                  0      0     0     0   0.01    0     0
0     0
## mean_relation_mass_shooting         0      0     0     0     0     0     0
0     0
## mean_relation_knows_victims         0      0     0     0     0     0     0
0     0
## mean_relation_co_worker             0      0     0     0     0     0     0
0     0
```

```
## mean_relation_neighbor             0      0    0.02  0.01    0     0     0
0    0.01
## mean_relation_friends              0      0    0.01    0     0     0   0.0
1  0.04   0.01
## -----------------------------------------------------------------------------
---------------

# save tempo to csv
#write.csv(tempo, file = "Final_Clustering.csv", row.names = FALSE)

# # read tempo
# tempo <- read.csv("Final_Clustering.csv")

guns_clusters = tempo
# replace blank with 0
guns_clusters[is.na(guns_clusters)] <- 0

## Section 5 - Appendix and the lens of an analyst

## 1 cluster Centroids
cols  = list(names(guns_clusters))

View(cols)

centroid_col_list <- c(
  "n_killed",
  "n_injured",
  "n_guns_involved",
  "stolen_count",
  "n_handguns",
  "n_rifle",
  "n_shotgun",
  "n_other",
  "subject_count",
  "victim_count",
  "total_count",
  "injured_count",
  "killed_count",
  "unharmed_arrested_count",
  "unharmed_count",
  "adult_count",
  "teen_count",
  "child_count",
  "relation_family",
  "relation_random_victims",
  "relation_aquaintance",
  "relation_significant_others",
  "relation_armed_robbery",
  "relation_gang",
  "relation_mass_shooting",
```

```r
  "relation_knows_victims",
  "relation_co_worker",
  "relation_neighbor",
  "relation_friends",
  "relation_home_invasion",
  "relation_does_not_know_victim",
  "female_percentage",
  "cluster_kmeans")


guns_clusters_subset = guns_clusters[, centroid_col_list]

View(guns_clusters_subset)
# show column data types
sapply(guns_clusters_subset, class)
```

```
##                      n_killed                      n_injured
##                     "integer"                      "integer"
##               n_guns_involved                   stolen_count
##                     "integer"                      "integer"
##                    n_handguns                        n_rifle
##                     "integer"                      "integer"
##                     n_shotgun                        n_other
##                     "integer"                      "numeric"
##                 subject_count                   victim_count
##                     "integer"                      "integer"
##                   total_count                  injured_count
##                     "integer"                      "integer"
##                  killed_count        unharmed_arrested_count
##                     "integer"                      "integer"
##                unharmed_count                    adult_count
##                     "integer"                      "integer"
##                    teen_count                    child_count
##                     "integer"                      "integer"
##               relation_family        relation_random_victims
##                     "numeric"                      "numeric"
##          relation_aquaintance   relation_significant_others
##                     "numeric"                      "numeric"
##         relation_armed_robbery                  relation_gang
##                     "numeric"                      "numeric"
##         relation_mass_shooting         relation_knows_victims
##                     "numeric"                      "numeric"
##            relation_co_worker              relation_neighbor
##                     "numeric"                      "numeric"
##              relation_friends         relation_home_invasion
##                     "numeric"                      "numeric"
## relation_does_not_know_victim              female_percentage
##                     "numeric"                      "numeric"
##                cluster_kmeans
##                      "factor"
```

```
# convert all columns to numeric
guns_clusters_subset <- sapply(guns_clusters_subset, as.numeric)
guns_clusters_subset = data.frame(guns_clusters_subset)
#create median table with cluster_kmeans in columns, and all rest variables i
n rows
median_table = aggregate(guns_clusters_subset, list(guns_clusters_subset$clus
ter_kmeans), mean)

#Transpose table(median_table)
median_table = t(median_table)
View(median_table)


# use cluster_kmeans as column names
colnames(median_table) <- median_table[1,]
# remove first row
median_table <- median_table[-1,]
# remove last row
median_table <- median_table[-nrow(median_table),]
# remove killed_count row
median_table <- median_table[-which(rownames(median_table) == "killed_count")
,]
# Round the median_table to 2 decimal places
rounded_median_table <- round(median_table, 2)



###########

library(tidyr)

##
## Attaching package: 'tidyr'

## The following object is masked from 'package:reshape2':
##
##     smiths

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.
0.0 ──
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ lubridate 1.9.3     ✔ tibble    3.2.1
## ✔ readr     2.1.4

## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```r
# convert date to date format
guns_f$date <- as.Date(guns_f$date)
# create a new column year-month column
guns_f$year_month <- format(guns_f$date, "%Y-%m")


# Create a table of the number of incidents per cluster per year-month
incidents_per_cluster_per_year_month <- guns_f %>%
  group_by(cluster_kmeans, year_month) %>%
  summarise(n = n())
```

```
## `summarise()` has grouped output by 'cluster_kmeans'. You can override usi
ng
## the `.groups` argument.
```

```r
print(incidents_per_cluster_per_year_month)
```

```
## # A tibble: 499 × 3
## # Groups:   cluster_kmeans [9]
##    cluster_kmeans year_month     n
##    <fct>          <chr>      <int>
##  1 1              2013-01        4
##  2 1              2013-03        2
##  3 1              2013-04        1
##  4 1              2013-05        7
##  5 1              2013-06        3
##  6 1              2013-07        6
##  7 1              2013-08        4
##  8 1              2013-09        6
##  9 1              2013-10        3
## 10 1              2013-11        3
## # ℹ 489 more rows
```

```r
library(tidyverse)

df_pivoted <- incidents_per_cluster_per_year_month %>%
  pivot_wider(names_from = cluster_kmeans, values_from = n)

# If you want to fill missing values with 0
df_pivoted[is.na(df_pivoted)] <- 0

# Print the pivoted data frame
print(df_pivoted)
```

```
## # A tibble: 63 × 10
##    year_month   `1`   `2`   `3`   `4`   `5`   `6`   `7`   `8`   `9`
##    <chr>      <int> <int> <int> <int> <int> <int> <int> <int> <int>
##  1 2013-01        4     0     1     0     0     0     1     0     1
```

```
##   2 2013-03        2     0     1     0     1     0     0     2     1
##   3 2013-04        1     0     1     0     1     0     2     1     1
##   4 2013-05        7     1     0     0     0     1     1     1     2
##   5 2013-06        3     0     0     0     1     0     0     3     1
##   6 2013-07        6     2     0     0     0     0     0     2     2
##   7 2013-08        4     2     1     0     0     0     1     0     3
##   8 2013-09        6     0     0     0     0     0     1     1     2
##   9 2013-10        3     1     1     0     0     0     0     0     1
##  10 2013-11        3     0     0     0     0     0     0     0     2
## # ℹ 53 more rows

df_pivoted = data.frame(df_pivoted)
```

Section 8 – Citations

1. https://www.kaggle.com/datasets/jameslko/gun-violence-data

Thank You