A person holding an object

Description automatically generated

*“****A Dichotomy***

***of Gun Violence***

***Incidents in***

***North America****”*

MGSC

Multivariate Statistics

Final Project

Submitted By:

Arham Anwar

McGill ID xxxxxxxx

**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 within the dataset. The philosophy of this project is to make the best out of available the project is answering the following key questions –

*“What type of crime was each incident?”*

*“What are the trends of each type of crime scene?”*

* 1. **Summary of Project**

Data source and cleaning -> feature engineering ->

This research leverages a dataset containing information on gun violence incidents with multiple guns involved. By applying data preprocessing and clustering techniques, we aim to identify inherent structures and patterns within the data. Our goal is to provide a nuanced understanding of the diverse factors associated with these incidents and shed light on potential avenues for targeted interventions.

* 1. **The Goals**

Cluster Analysis: Apply PCA and k-means clustering to identify inherent structures and groupings in the dataset.

Descriptive Statistics: Generate summary statistics for each cluster, offering insights into the characteristics of incidents within each group.

Visualizations: Utilize visualizations, such as scatter plots, box plots, and heatmaps, to effectively communicate patterns and relationships in the data.

**Section 2 – Data Description**

Data source, feature engineering, initial data exploration

**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 next subsection we discuss the import features engineered from the original dataset.

**2.2. Feature Engineering:**

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. These engineered features aim to capture essential nuances within the dataset.

**2.2.1. Lethality – Weapon of Choice:**

First of all, quantity and quality of weapon lethality was extracted by calculating number of guns by gun type, i.e number of handguns, rifles, 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 possibly 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 compliment 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 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, 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 among the perpetrator and the victims is the final piece of the puzzle and gives clear insight on 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:**

An initial exploration of the dataset was performed. An exploration of the distribution of key variables, such as the number of injured and killed, stolen guns, and participant demographics, offers a preliminary understanding of the dataset's characteristics. Investigating relationships between variables, such as the correlation between the number of victims and the type of guns involved, provides insights into potential causal links and dependencies crucial for developing the model.

Profile of the dataset available indicates –

Based on the various steps in this section, the final selection of features ensured no variables with repeated information were taken. Additionally, variables with no correlation to crime scene were eliminated, such as the ‘source-link’. Variable like the ‘source-link’ are realised after the crime scene takes place and thus offer no valuable insight in light of the objective of this project.

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. Many candidate algorithms were tested, and the final model selection was K-Means Clustering in combination with PCA.

Multicollinearity: After feature creation multicollinearity was tested and highly correlated variables were combined using PCA

**3.1.** **The decision of ‘K’**

The optimal number of clusters was determined through careful consideration of the data and iterative testing. Figure X, representing a plot of total weighted sum of squares vs ‘K’, in combination with Figure X, 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. Achilies heal to the curse of dimensionality:**

PCA was employed to club engineered features of the same type. For e.g.,

3.3. Visualization: To enhance interpretability, we complemented our analysis with various visualizations, including scatter plots for PCA representation, box plots for cluster comparisons, and a heatmap to illustrate variable relationships.

Descriptive Statistics: We employed summary statistics and stargazer tables to provide an overview of each cluster's characteristics. This step facilitated a comprehensive understanding of the distinct features associated with different incident groupings.

**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. The cluster identified are as follows:

**Cluster 1: The Urban Turbulence**

Cluster 1, characterized by a size of 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 endeavors. 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 in nature. 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 conflicts among co-workers. 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**

4.3. Cluster Trends from the lens of an analyst

Stargazer tables were generated to present summary statistics for each cluster. This included mean values and other relevant metrics for key variables, offering a detailed profile of incidents within each grouping.

4.4. Implications and Recommendations

The results carry implications for policymakers, law enforcement, and community stakeholders. By identifying patterns and characteristics associated with different incident clusters, targeted interventions can be devised to address specific challenges and mitigate risks.

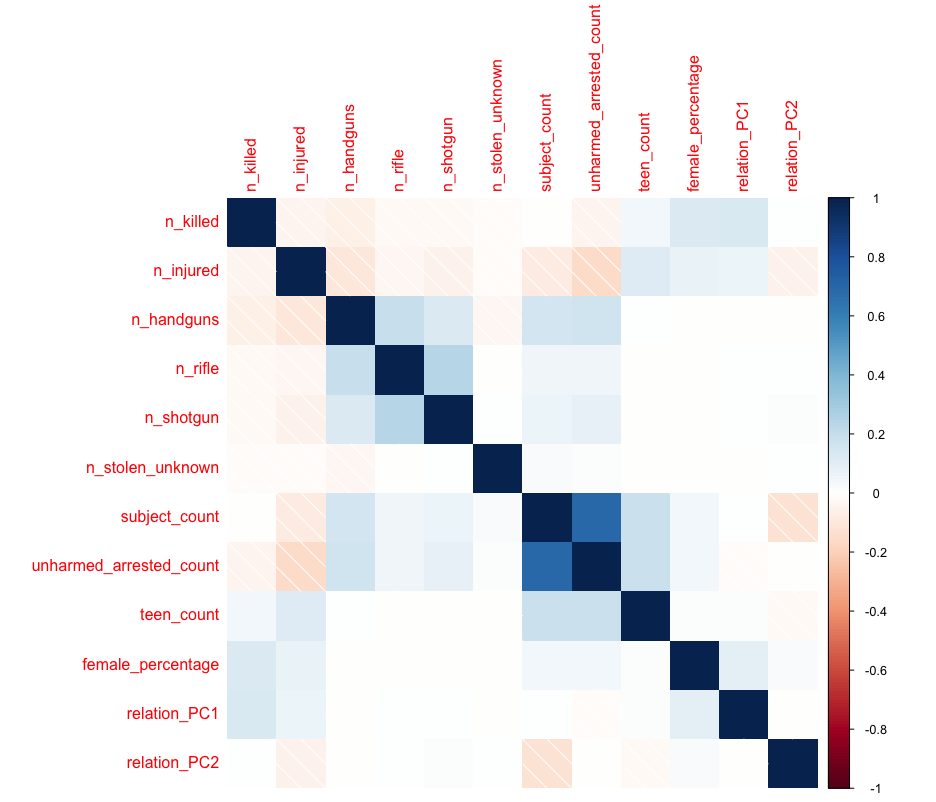
Section 5 – Classifications/Predictions and Conclusions [1-2 Pages]

**Section 7 – Future Scope:**

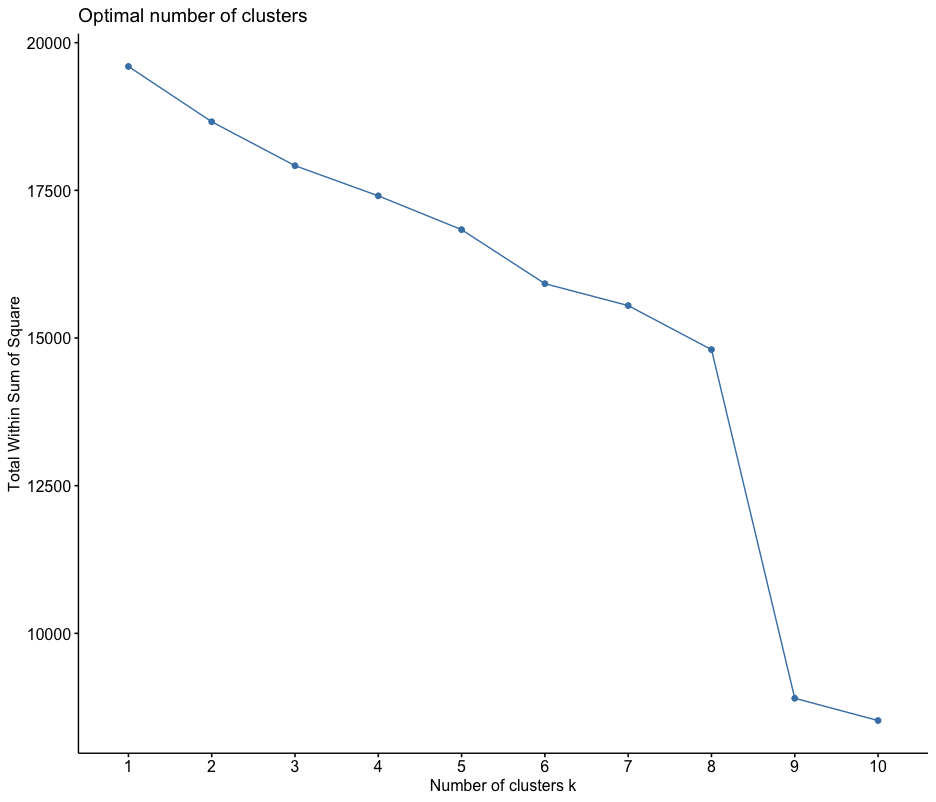
**Section 8 – Appendices**

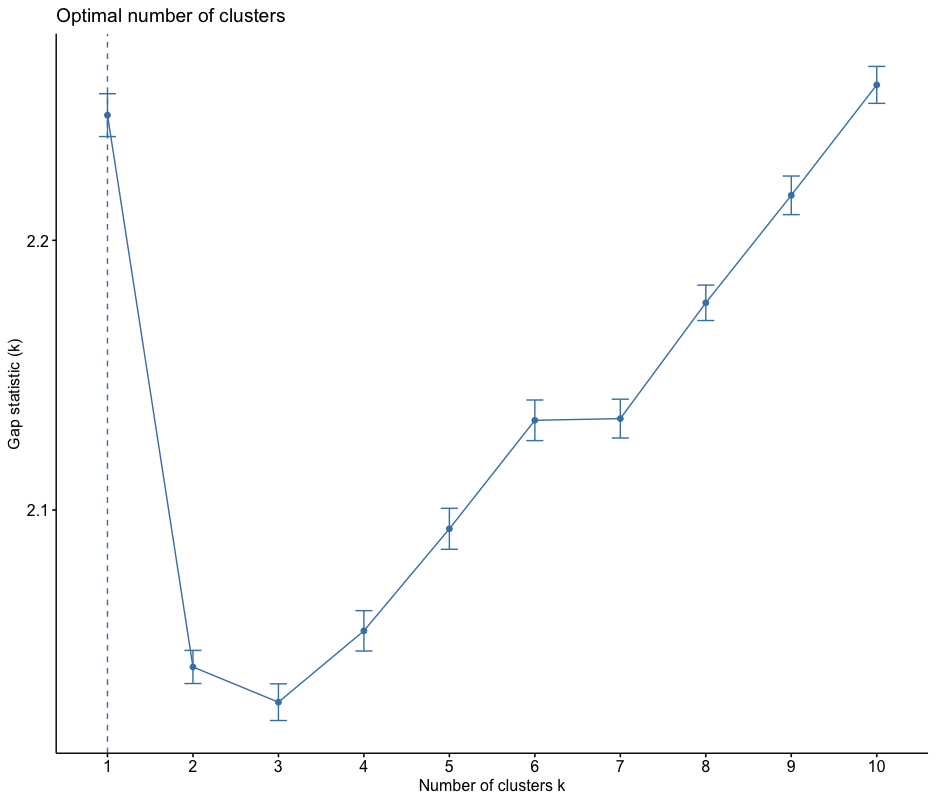
**8.1 Figures**

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



*Fig 2 – Within Sum of Squares vs cluster choice ‘k’*

**



*Fig 4 – Visual Representation of Clusters with respect to Eigen Vectors*

*A diagram of a cluster of clusters

Description automatically generated with medium confidence*

*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.*

Table 1

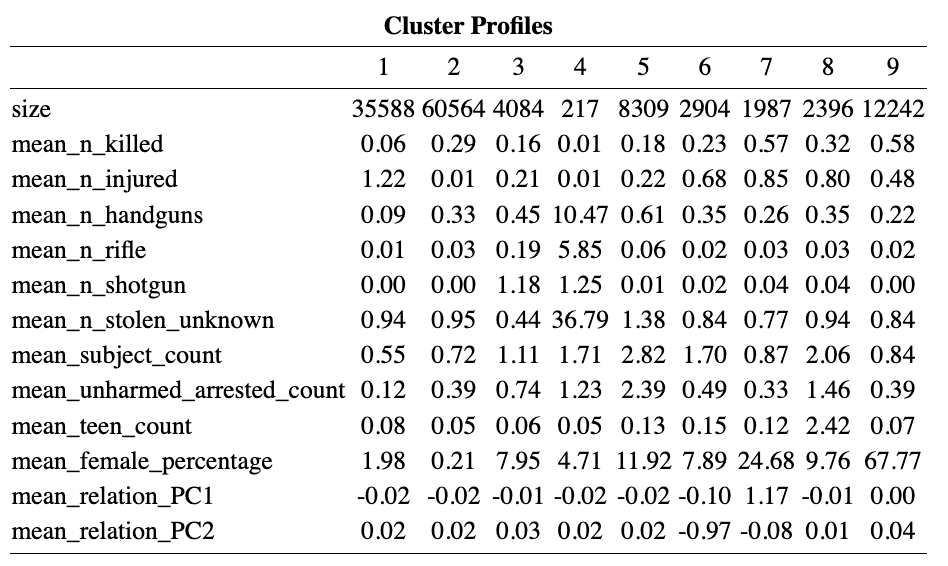
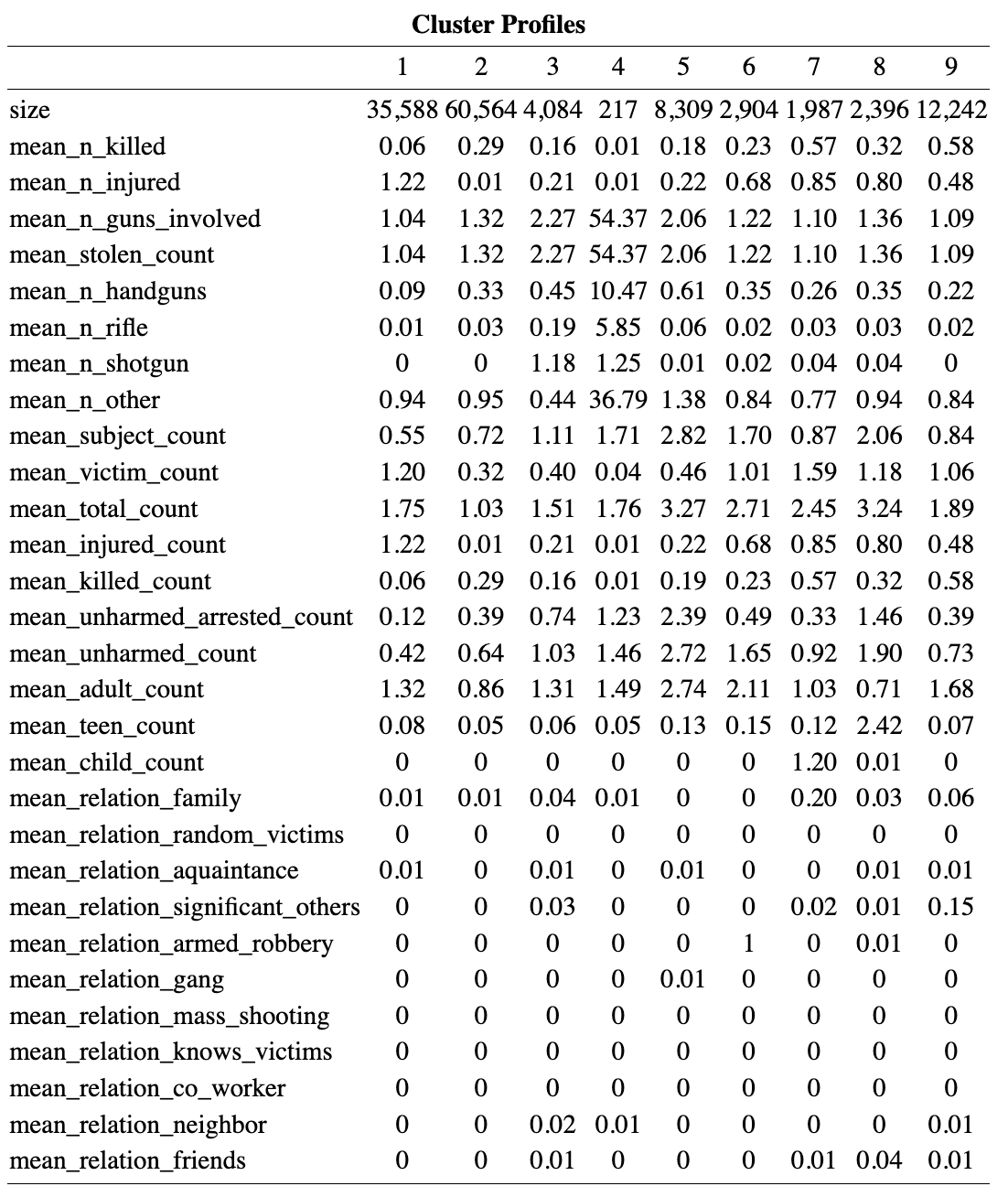


Table 2



Section 7 – Code

Section 8 - Citations