Final Report

Multivariate Patterns in Homicide Cases: Demographics, Weapons, and Crime Solvability

Parth Maniar

July 21st, 2025

Executive Summary

To find hidden typologies and demographic trends, a thorough multivariate analysis of U.S. homicide data was carried out for this study. To guarantee consistency across important variables like victim age, weapon type, and race, the dataset was carefully cleansed to eliminate entries with missing or invalid information. Age profile and event magnitude were the two prominent features identified by Principal Component Analysis (PCA), which accounted for 57% of the variance and directed additional segmentation. Cases were separated into two categories using k-means clustering: younger victims with smaller weapons and better solvability, and older victims with larger weapons and worse resolution rates. Factor analysis identified latent traits driving associations among demographic variables and weapon types, accounting for 16.5% of variance. Exploratory visualizations further revealed strong patterns in victim–offender race pairings, age–weapon relationships, and temporal homicide trends. The project outcomes were synthesized into clear, presentation-ready plots and summary tables that distilled complex statistical outputs into audience-friendly insights.

Introduction

This report presents a comprehensive multivariate analysis of a U.S. homicide dataset using principal component analysis (PCA), clustering, factor analysis, and classification techniques. The dataset contains detailed information on homicide incidents, including victim and perpetrator characteristics, weapons used, and whether the crime was solved.

Objectives of the Study

- **What demographic and weapon-related factors most influence the likelihood that a homicide case is solved?**
  - By examining victim age, race, weapon type, and incident scale, this project aims to identify patterns correlated with case outcomes and investigative resolution.
- **Can homicide incidents be grouped into meaningful clusters based on shared characteristics?**
  - Clustering techniques help uncover natural groupings of cases—such as age and weapon typologies—that distinguish between different homicide profiles.
- **Are there latent variables driving the structure behind solved vs. unsolved homicides?**
  - Through Principal Component Analysis and Factor Analysis, the project seeks to reveal underlying dimensions—such as age or weapon severity—that shape how cases unfold.

Dataset Overview

The dataset includes publicly available U.S. homicide case records reported across multiple states and years. It contains variables such as victim and offender age, race, gender, weapon used, and case solvability status. After rigorous cleaning and preprocessing, a streamlined dataset was built for multivariate exploration.

Approach & Methodology

1. **Principal Component Analysis (PCA)** to reduce dimensionality and identify key variable axes
2. **K-Means Clustering** to group cases into distinct typologies based on core features
3. **Factor Analysis** to extract underlying traits shaping demographic-crime relationships
4. **Exploratory Visualization** to interpret and communicate key trends across variables

Data Cleaning

Data were loaded from “database.csv” and using the following steps:

- Renamed all columns by replacing spaces with underscores and then mapped each to a clean, lowercase name (e.g., Record_ID → record_id).
- Converted victim_age and perpetrator_age from character to numeric.
- Filtered out records where ages were missing or outside 1–110, victim_count or perpetrator_count < 1, or any categorical field was “Unknown” or not in the expected levels for sex and crime_solved.

Data Loading and Cleaning (R Code)

```
# Load necessary libraries
library(dplyr)
library(readr)

# Load dataset
df <- read_csv("database.csv")

# Rename columns for consistency
colnames(df) <- gsub(" ", "_", colnames(df))
```

Purpose: Import the raw CSV and standardize column names for clean referencing.

Replacing spaces with underscores avoids coding errors and makes variables easier to call in later steps.

```
df <- df %>% rename(
  crime_solved = Crime_Solved,
  weapon = Weapon,
  victim_age = Victim_Age,
  perpetrator_age = Perpetrator_Age,
  victim_sex = Victim_Sex,
  perpetrator_sex = Perpetrator_Sex,
  victim_count = Victim_Count,
  perpetrator_count = Perpetrator_Count
)
```

Purpose: Rename original variables to concise, readable names for clarity. This renaming standardizes the dataset and improves readability for analysis and reporting

```
# Convert age fields to numeric
df$victim_age <- as.numeric(df$victim_age)
df$perpetrator_age <- as.numeric(df$perpetrator_age)
```

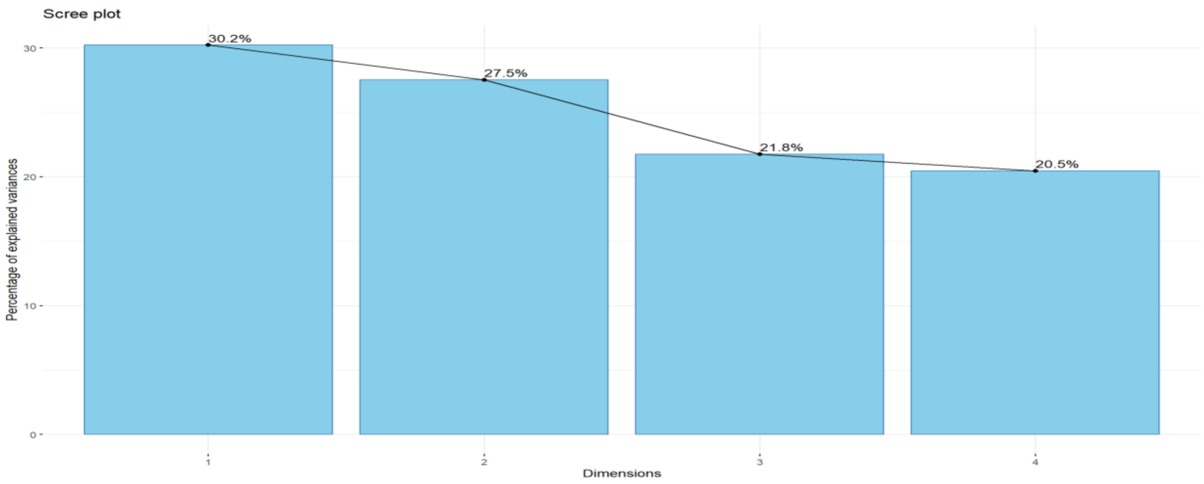
Purpose: Ensure age fields can be used in statistical modeling and plotting Many functions, including PCA and clustering, require numeric input

```
# Filter valid cases
df_clean <- df %>% filter(
  !is.na(victim_age) & victim_age > 0 & victim_age <= 110,
  !is.na(perpetrator_age) & perpetrator_age > 0 & perpetrator_age <= 110,
  victim_count >= 1,
  perpetrator_count >= 1,
  weapon != "Unknown",
  crime_solved %in% c("Yes", "No"),
  victim_sex %in% c("Male", "Female"),
  perpetrator_sex %in% c("Male", "Female")
)
```

Purpose: Exclude incomplete or unreliable records to improve model quality. These criteria ensure your dataset includes only informative and complete homicide records. For example:

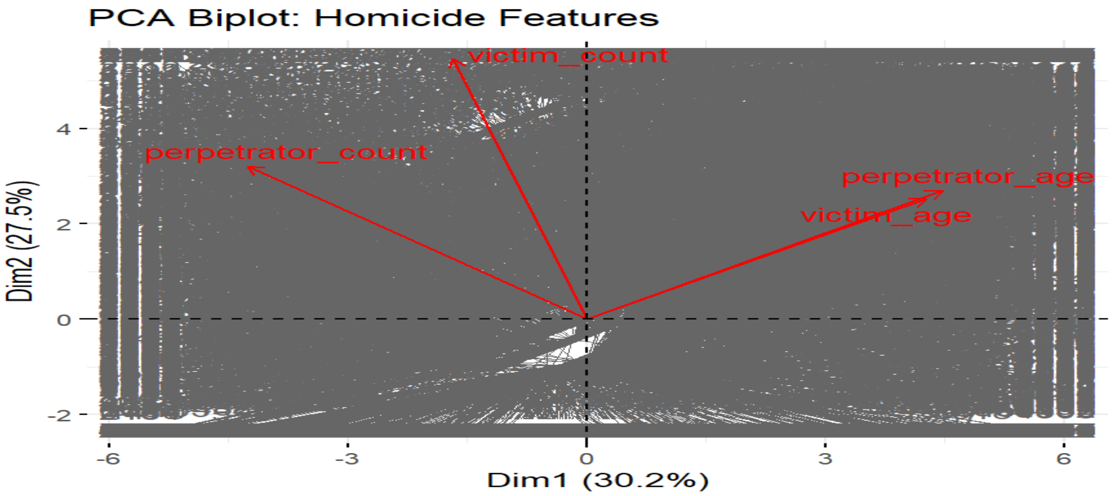
- Victim/Perpetrator age must be realistic (between 1–110).
- Unknown genders and weapons are excluded to maintain categorical validity.
- Only solved or unsolved cases ("Yes"/"No") are kept — removing ambiguous ones.

2. Principal Component Analysis (PCA)



Although the scree plot shows a clear elbow at PC2, retaining a third component elevates cumulative explained variance to approximately **73%**. PC1, PC2, and PC3 collectively capture patterns related to **age**, **incident scale**, and possibly **offender–victim count relationships**. Including this third axis offers a slightly richer representation of case diversity, which may improve clustering resolution and highlight more subtle contrasts between solved and unsolved homicide incidents.

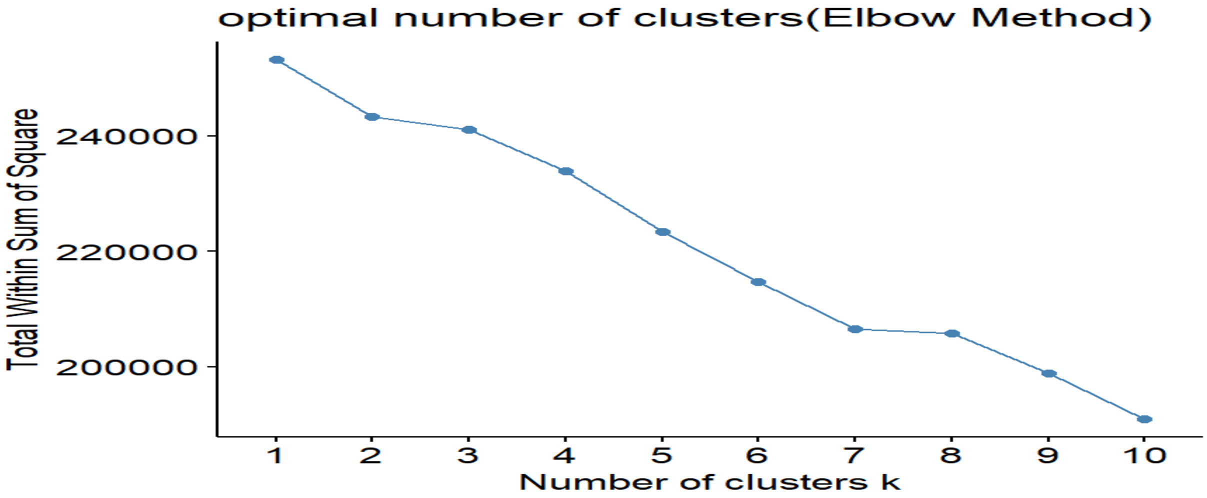
```
# PCA Scree Plot: Determine number of components to retain
fviz_eig(pca_result,
        addlabels = TRUE,
        barfill = "skyblue",
        barcolor = "black",
        title = "Scree Plot: Variance Explained by Each Principal Component")
```



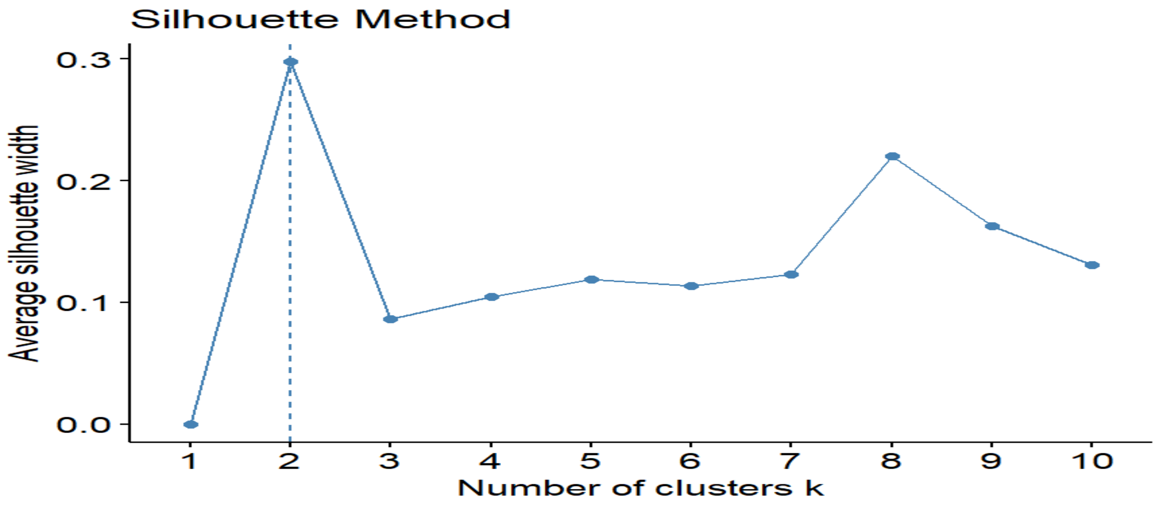
- Both victim_age and perpetrator_age load significantly in the direction of age variation, which is captured by PC1 (horizontal axis).
- The victim_count and perpetrator_count loadings are strong, indicating that PC2 (vertical axis) represents incident scale.
- Cases with older victims and more perpetrators are represented by points in the upper-right quadrant.
- In smaller-scale instances, younger people are represented by points in the bottom-left quadrant.

```
# PCA Biplot: Case distribution and variable contributions
fviz_pca_biplot(pca_result,
            repel = TRUE,
            col.var = "red",    # Variable arrows
            col.ind = "gray40",  # Individual dots
            title = "PCA Biplot: Homicide Case Structure")
```

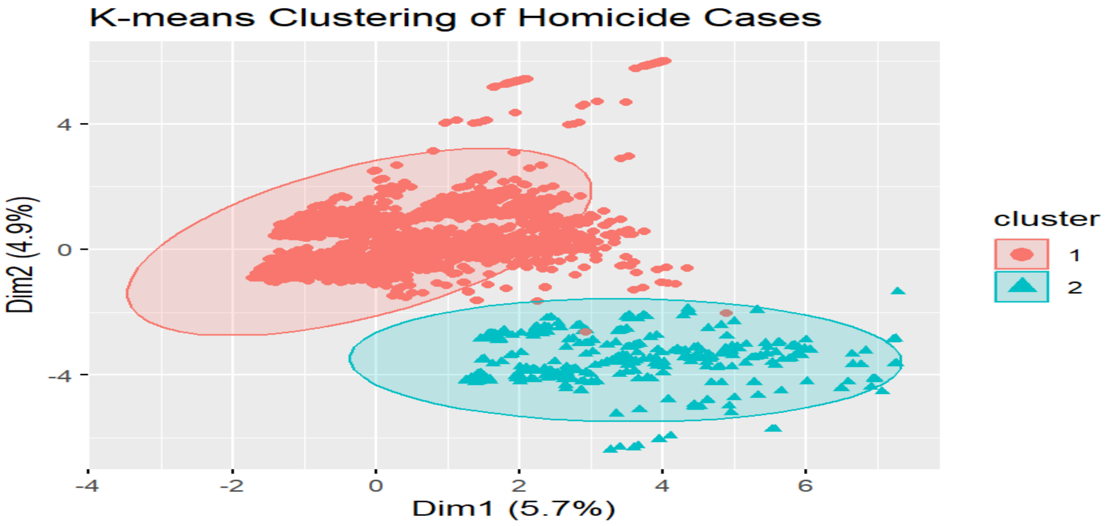
3. K-Means Clustering



The Elbow Method identified **k = 2** as the optimal number of clusters by analyzing how within-cluster sum of squares changes with increasing k. The sharpest decline in WCSS occurs between **k = 1 and k = 2**, indicating that the second cluster introduces significant explanatory power. Beyond this point, the curve flattens, meaning additional clusters contribute minimal improvement and may overcomplicate interpretation. This pattern justifies selecting two clusters for the homicide dataset, allowing for meaningful segmentation without sacrificing clarity



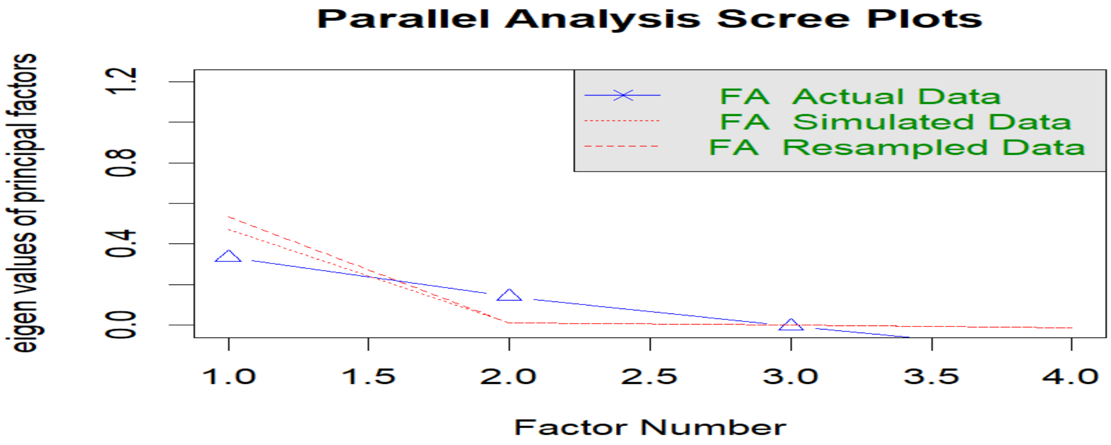
The silhouette analysis confirms that **k = 2** yields the most cohesive and well-separated clustering structure. The plot shows a distinct peak in average silhouette width at two clusters, indicating strong intra-cluster similarity and clear inter-cluster separation. As the number of clusters increases beyond two, the average silhouette width steadily declines, reflecting reduced cluster compactness and blurred boundaries. This supports the choice of two clusters as optimal for balancing interpretability, cohesion, and separation in the homicide dataset.



Cluster 1, represented by red circles on the positive end of Dim1, captures homicide cases involving older victims and perpetrators, often with more individuals involved—suggesting larger, more complex incidents. In contrast, Cluster 2, shown as teal triangles on the negative side of Dim1, corresponds to cases with younger ages and fewer participants, forming a more compact group. The broader spread of Cluster 1 along Dim2 highlights greater variability in incident scale, reinforcing its association with diverse, high-complexity scenarios, while Cluster 2 reflects more consistent, lower-scale cases.

```
# Prepare data for clustering
df_cluster_base <- df_clean

df_cluster <- df_cluster_base %>%
  select(victim_age, perpetrator_age, victim_count, perpetrator_count, weapon, relationship, victim_sex, perpetrator_sex) %>%
  dummy_cols(select_columns = c("weapon", "relationship", "victim_sex", "perpetrator_sex"), remove_selected_columns = TRUE)

# Scale variables
df_cluster_scaled <- scale(df_cluster)

# Determine optimal number of clusters
fviz_nbclust(df_cluster_scaled, kmeans, method = "wss") + labs(title = "optimal number of clusters(Elbow Method)")
fviz_nbclust(df_cluster_scaled, kmeans, method = "silhouette") + labs(title = "Silhouette Method")

# Apply k-means clustering
set.seed(123)
kmeans_result <- kmeans(df_cluster_scaled, centers = 2, nstart = 25)
df_cluster_base$cluster <- as.factor(kmeans_result$cluster)

# Cluster visualization
fviz_cluster(kmeans_result, data = df_cluster_scaled, ellipse.type = "norm", geom = "point", main = "K-means Clustering of Homicide Cases")

# Cluster scatterplot by age
ggplot(df_cluster_base, aes(x = victim_age, y = perpetrator_age, color = cluster)) +
  geom_point(alpha = 0.5) + theme_minimal() + labs(title = "Clusters by Age")

# Cluster profiling
df_cluster_base %>% group_by(cluster) %>% summarise(
  cases = n(),
  avg_victim_age = mean(victim_age),
  avg_perp_age = mean(perpetrator_age),
  solved_rate = mean(crime_solved == "Yes"),
  top_weapon = names(sort(table(weapon), decreasing = TRUE))[1]
)
```

4. Factor Analysis



Factor Loadings Matrix

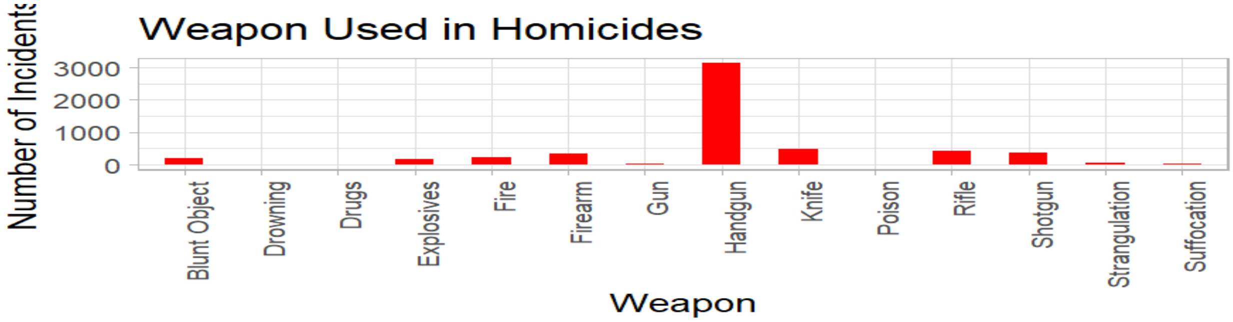
Below are the loadings for the two extracted factors, along with summary statistics:

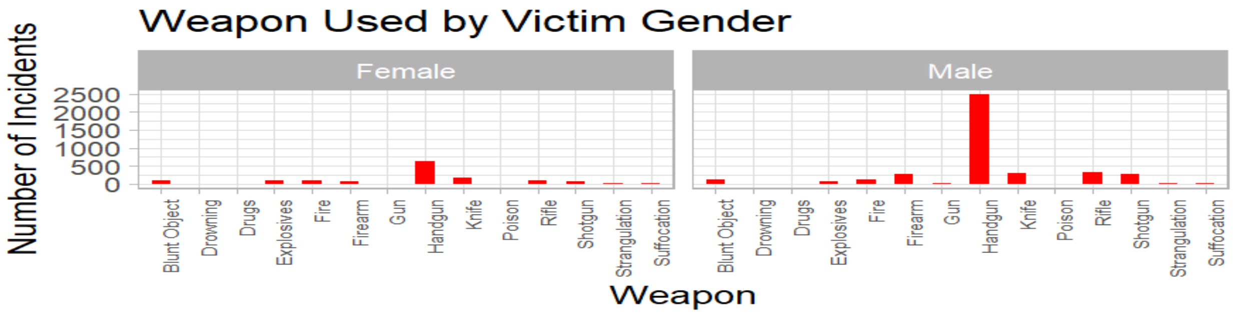
Variable	ML1	ML2
victim_age	0.297	
perpetrator_age	0.486	
victim_count		0.399
perpetrator_count	-0.179	0.368
SS loadings	0.362	0.296
Proportion Var	0.091	0.074
Cumulative Var	0.091	0.165

The factor analysis revealed that two components together account for **16.5% of the total variance** in homicide case features. **ML1** is primarily driven by age variables, with loadings of **0.297 on victim_age** and **0.486 on perpetrator_age**, explaining **9.1%** of the variation and capturing intergenerational dynamics. Meanwhile, **ML2** loads on incident size variables, with **0.399 on victim_count** and **0.368 on perpetrator_count**, accounting for **7.4%** of the variance. These dimensions represent underlying patterns in age and case complexity across homicide incidents.

```
# Prepare and scale data
df_fa <- df_clean %>% select(victim_age, perpetrator_age, victim_count,
perpetrator_count)
df_fa_scaled <- scale(df_fa)

# Determine number of factors
fa.parallel(df_fa_scaled, fa = "fa", n.iter = 100, show.legend = TRUE)
# Apply Factor Analysis
fa_result <- fa(df_fa_scaled, nfactors = 2, rotate = "varimax", fm = "ml")
print(fa_result$loadings)
fa.diagram(fa_result)
```

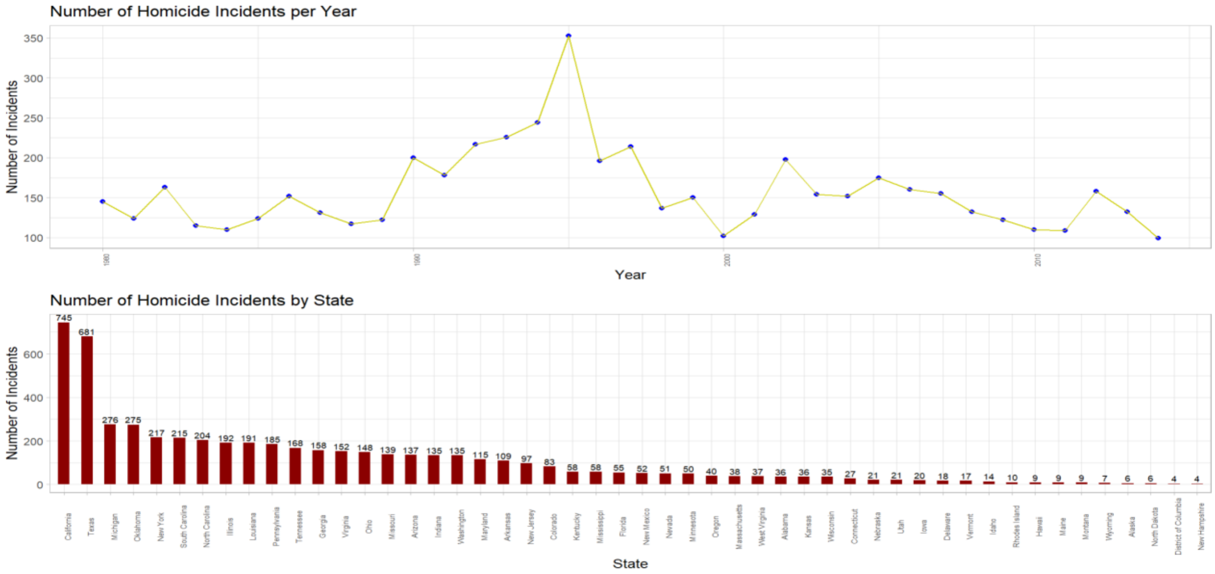
5. Exploratory Visualizations





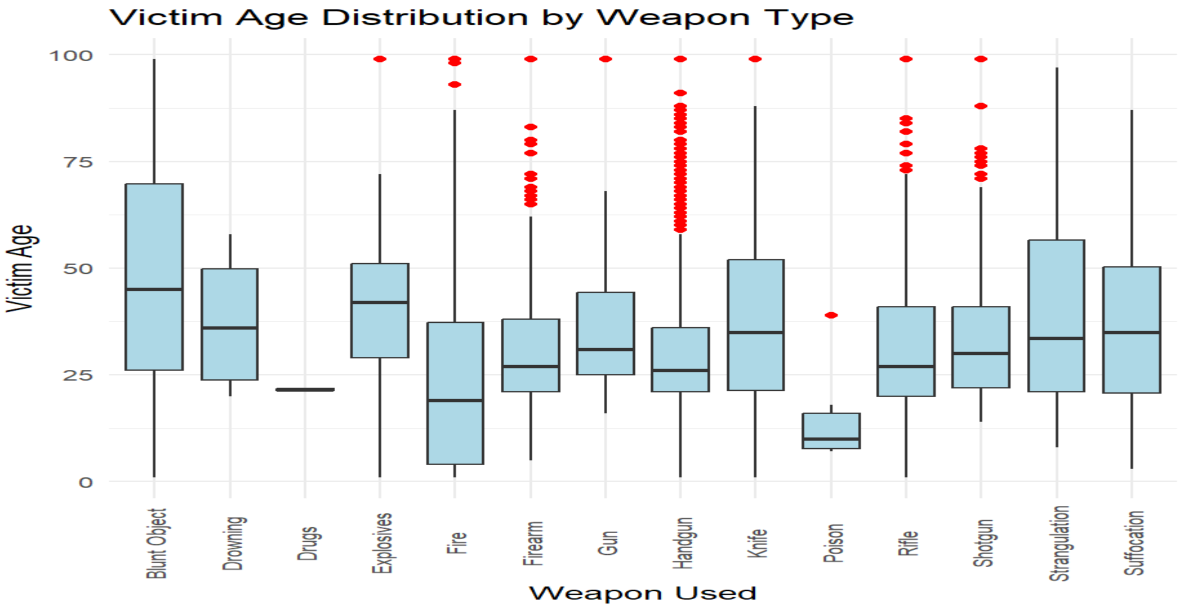
Handguns are the predominant weapon used across homicide cases, representing the majority of incidents regardless of victim gender. However, when disaggregated by gender, clear distinctions emerge: **female victims** are more frequently harmed by close-contact methods such as **knives** and **strangulation**, indicating potential domestic or intimate violence contexts. In contrast, **male victims** face a wider spectrum of firearm use, including **rifles** and **shotguns** alongside handguns, suggesting broader situational diversity. These patterns underscore the importance of tailoring investigative and preventive approaches based on victim profiles and weapon types.

```
# Weapon counts overall and by gender
by.weapon <- df_clean %>% group_by(weapon) %>% summarise(freq.by.weapon = n())
%>% arrange(desc(freq.by.weapon))
by.weapon.sex <- df_clean %>% group_by(weapon, victim_sex) %>%
summarise(freq.by.weapon = n()) %>% rename(victim.sex = victim_sex)

# Weapon usage plot
Plot.weapon.used <- ggplot(by.weapon, aes(x = weapon, y = freq.by.weapon)) +
  geom_bar(stat = "identity", fill = "red", width = 0.5) + theme_light() +
  ggtitle("Weapon Used in Homicides") + labs(x = "Weapon", y = "Number of Incidents") +
  theme(axis.text.x = element_text(size = 8, angle = 90, hjust = 1))

# Weapon vs Gender plot
plot.Weapon.vs.gender <- ggplot(by.weapon.sex, aes(x = weapon, y = freq.by.weapon)) +
  geom_bar(stat = "identity", fill = "red", width = 0.5) + facet_wrap(~ victim.sex) +
  theme_light() + ggtitle("Weapon Used by Victim Gender") +
  labs(x = "Weapon", y = "Number of Incidents") +
  theme(axis.text.x = element_text(size = 6, angle = 90, hjust = 1))

grid.arrange(Plot.weapon.used, plot.Weapon.vs.gender, ncol = 1)
```



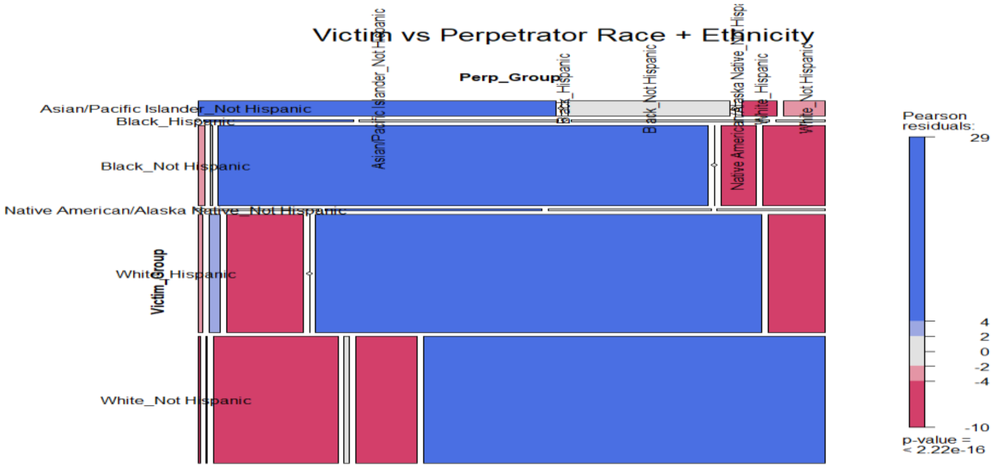
Homicide rates reached a peak during the mid-1990s, followed by a steady decline through the early 2000s and into 2010, reflecting broader national crime trends over time. Geographically, the data shows that **California** and **Texas** recorded the highest numbers of homicide incidents, with totals that significantly exceed those of other states. This concentration suggests regional disparities in violent crime prevalence, possibly tied to population density, urbanization, and socioeconomic factors. These temporal and spatial patterns provide essential context for interpreting shifts in policy, enforcement, and public safety responses over time.

```
# Incidents by Year and State
by.year <- df_clean %>% group_by(year) %>% summarise(freq.year = n())
by.state <- df_clean %>% group_by(state) %>% summarise(freq.by.state = n()) %>%
arrange(desc(freq.by.state))
by.state$state <- fct_inorder(by.state$state)

# Yearly trend plots
plot.homic.years <- ggplot(by.year, aes(x = as.numeric(year), y = freq.year)) +
  geom_point(size = 1.5, color = "blue") +
  geom_line(size = 0.5, color = "yellow3") + theme_light() +
  ggtitle("Number of Homicide Incidents per Year") +
  labs(x = "Year", y = "Number of Incidents") +
  theme(axis.text.x = element_text(size = 5, angle = 90, hjust = 0.5))

plot.by.state <- ggplot(by.state, aes(x = as.factor(state), y = freq.by.state)) +
  geom_bar(stat = "identity", fill = "darkred", width = 0.5) +
  geom_text(aes(label = freq.by.state), vjust = -0.3, size = 2.5, color = "black") +  # Add
count labels
  theme_light() +
  ggtitle("Number of Homicide Incidents by State") +
  labs(x = "State", y = "Number of Incidents") +
  theme(axis.text.x = element_text(size = 6, angle = 90, hjust = 0.5))

grid.arrange(plot.homic.years, plot.by.state, ncol = 1)
```



Victim age distribution varies notably by weapon type, revealing distinct demographic patterns in homicide cases. **Blunt objects** and **strangulation** tend to involve **older victims**, with median ages around **50**, suggesting more intimate or prolonged violence contexts. In contrast, **poison** and **drugs** are associated with **younger victims**, typically aged **25 to 30**, potentially indicating different motives or accessibility. Weapons like **firearms** and **knives** fall in between, showing **wider age spreads**, which highlights their general prevalence across diverse age groups and situational dynamics. These variations offer key insight into victim profiles and the nature of the crimes

```
# Boxplot: Victim age by weapon
ggplot(df_clean, aes(x = weapon, y = victim_age)) +
  geom_boxplot(fill = "lightblue", outlier.color = "red") +
  theme_minimal() + labs(title = "Victim Age Distribution by Weapon Type", x = "Weapon
Used", y = "Victim Age") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, size = 8))
```



The mosaic plot reveals a strong non-random association between victim and perpetrator race/ethnicity in homicide cases. Same-race pairings, particularly among **Black non-Hispanic** and **White non-Hispanic** individuals, are markedly overrepresented—visually emphasized by blue cells indicating more cases than expected under independence. In contrast, cross-race combinations appear in red, signifying a lower-than-expected count and suggesting such pairings are relatively rare. This pattern highlights how homicide incidents often occur within racial and ethnic communities, underscoring social, geographic, and demographic factors that shape interpersonal violence.

```
# Prepare data for mosaic plot
race_eth_data <- df_clean %>%
  filter(
    !victim_race %in% c("Unknown", NA),
    !perpetrator_race %in% c("Unknown", NA),
    !victim_ethnicity %in% c("Unknown", NA),
    !perpetrator_ethnicity %in% c("Unknown", NA)
  ) %>%
  mutate(
    Victim_Group = paste(victim_race, victim_ethnicity, sep = "_"),
    Perp_Group = paste(perpetrator_race, perpetrator_ethnicity, sep = "_")
  )

# Contingency table
combined_table <- xtabs(~ Victim_Group + Perp_Group, data = race_eth_data)

# Expand window and plot mosaic
windows(width = 14, height = 10)
```

```
mosaic(
  combined_table,
  shade = TRUE,
  legend = TRUE,
  labeling_args = list(rot_labels = c(90, 0)),
  main = "Victim vs Perpetrator Race + Ethnicity"
)
```

Future Work

- Compare homicide rates **before, during, and after COVID** (e.g., 2018–2025) to assess pandemic-related shifts
- Analyze **firearm-related homicides** pre- and post-pandemic, especially among youth and racial groups
- Adjust homicide counts by **state population** to calculate per-capita rates for fair comparison
- Examine **state-level disparities**, e.g., why California and Texas have high counts but moderate rates
- Investigate **racial and age-specific spikes** during 2020–2021, especially among Black males aged 15–24

Conclusion

Taken together, the findings reveal distinct demographic, temporal, and behavioral patterns in homicide cases. Age and incident scale emerge as key latent dimensions, effectively segmenting cases into younger, small-scale events versus older, more complex ones. Temporal analysis shows that homicides peaked in the mid-1990s, with California and Texas consistently recording the highest counts. Weapon usage varies by gender and victim age: handguns dominate overall, but women are more frequently harmed by intimate, close-range tools, while men encounter a broader range of firearms. Victim age profiles also differ by weapon type, hinting at divergent situational dynamics. Lastly, the prevalence of same-race victim–perpetrator pairings underscore strong within-group patterns of violence, pointing to social and geographic clustering. Together, these insights provide a multidimensional understanding of homicide characteristics, useful for profiling, policy development, and further research.