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

Introduction

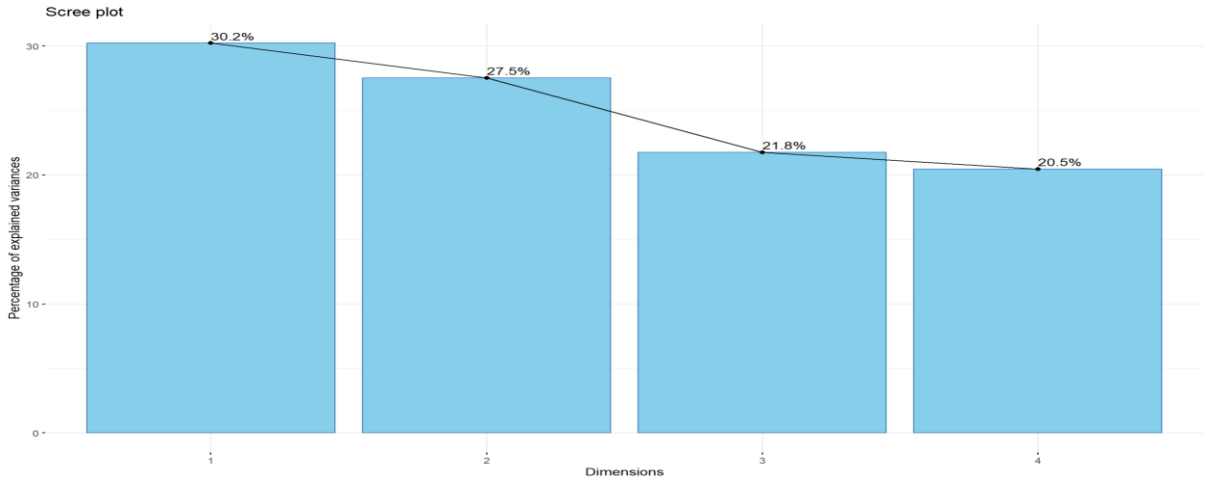
This report presents a comprehensive multivariate analysis of a U.S. homicide dataset using principal component analysis (PCA), clustering, factor analysis, and classification techniques. The dataset contains detailed information on homicide incidents, including victim and perpetrator characteristics,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

1. Data Cleaning

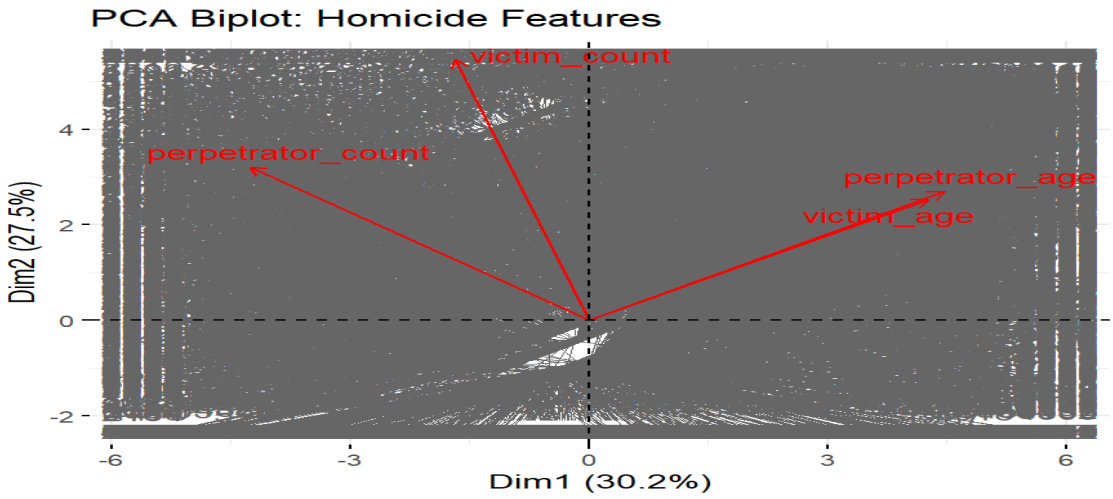
Data were loaded from "database.csv" and using the following steps:

- 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.

2. Principal Component Analysis (PCA)

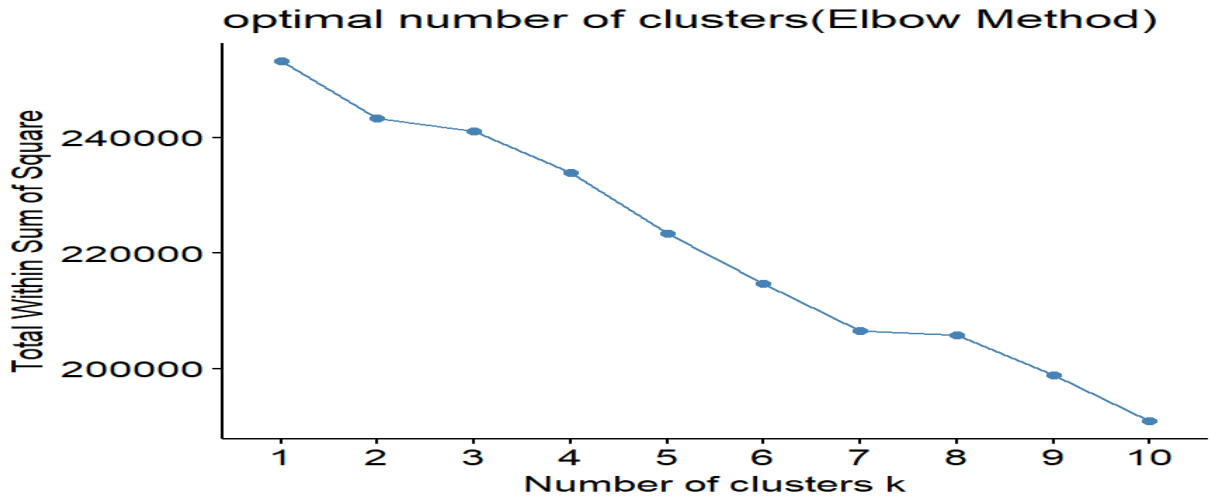


The scree plot shows that the first two components capture most of the information—PC1 explains ~30% and PC2 ~27%, totaling 57%—after which each additional component adds relatively little. This "elbow" at PC2 suggests keeping two components for a compact yet informative representation.

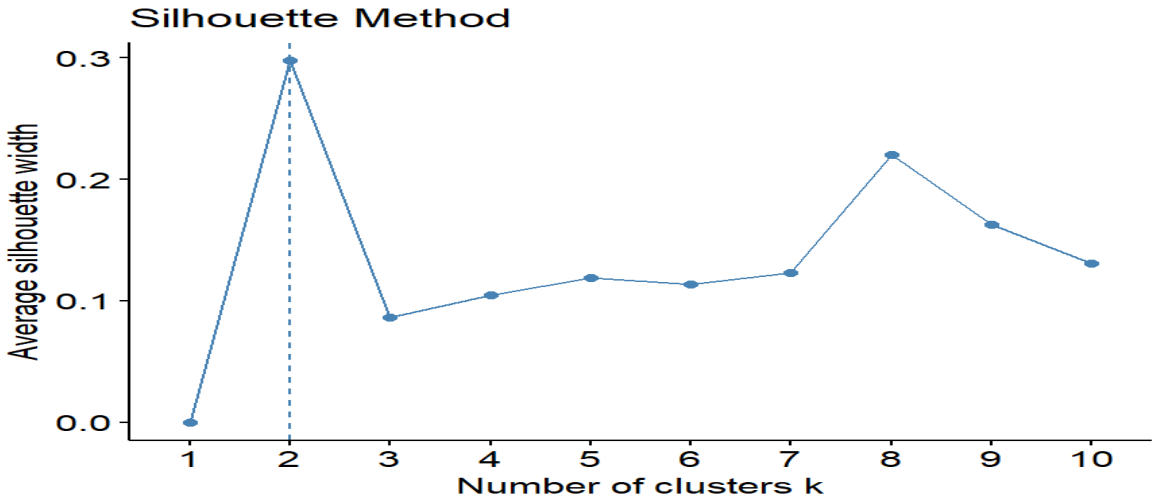


- Both victim_age and perpetrator_age load significantly in the direction of age variation, which is captured by PC1 (horizontal axis).
- The victim_count and perpetrator_count loadings are strong, indicating that PC2 (vertical axis) represents incident scale.
- Cases with older victims and more perpetrators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

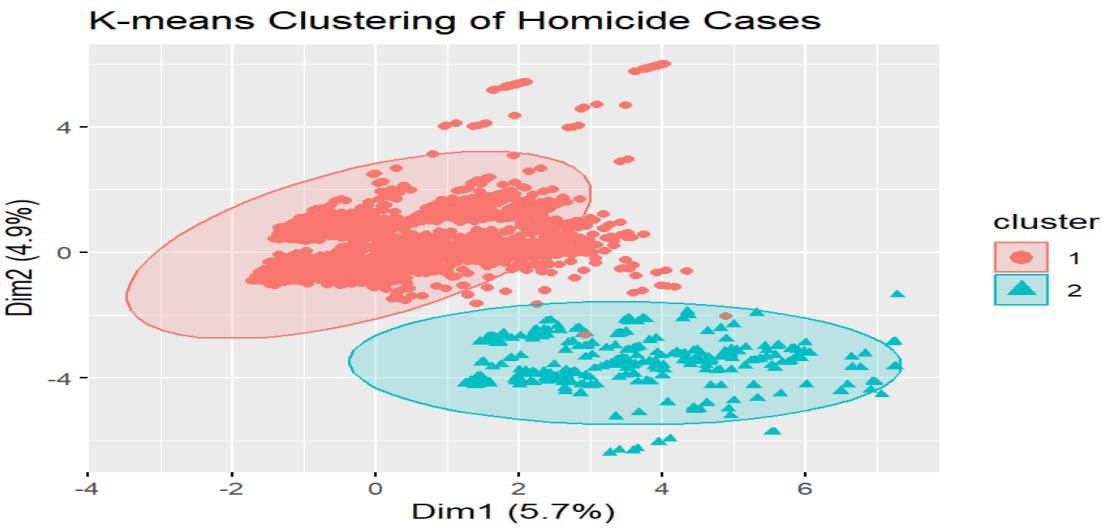
3. K-Means Clustering



This elbow at k = 2 indicates that two clusters capture the majority of the structure without needless complexity. The within-cluster sum of squares drops the steepest between k = 1 and k = 2, after which the curve flattens.

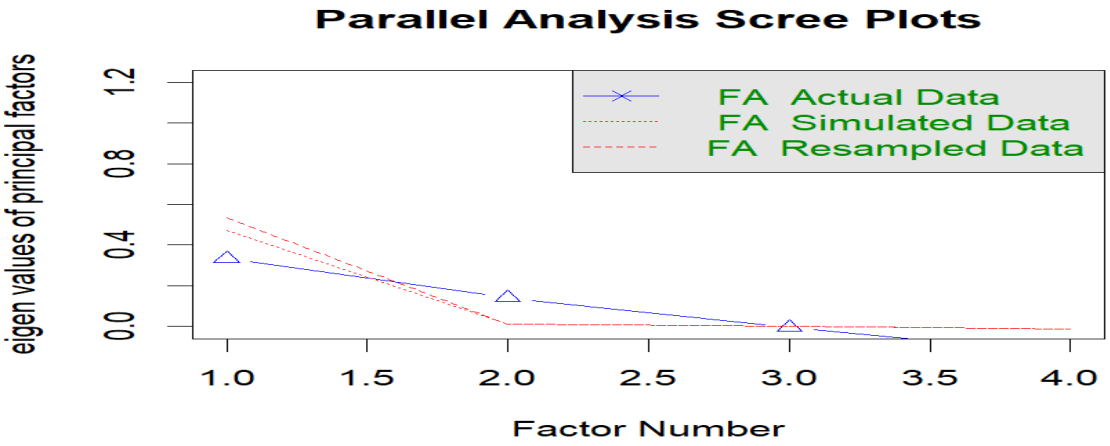


Two clusters produce the most cohesive and well-separated grouping, according to the silhouette plot, which peaks at k = 2 with the maximum average silhouette width; additional clusters result in a lower average silhouette and diminishing separation.



On the positive side of Dim1, Cluster 1 (red circles) represents incidences with higher numbers and combined scores for age (older victims/perpetrators, more persons involved). Younger ages and fewer participants are represented by Cluster 2 (teal triangles), which is located on the negative side of Dim1. While Cluster 2 is more compact, the broader circle surrounding Cluster 1 on Dim2 indicates higher incident scale variability.

4. Factor Analysis



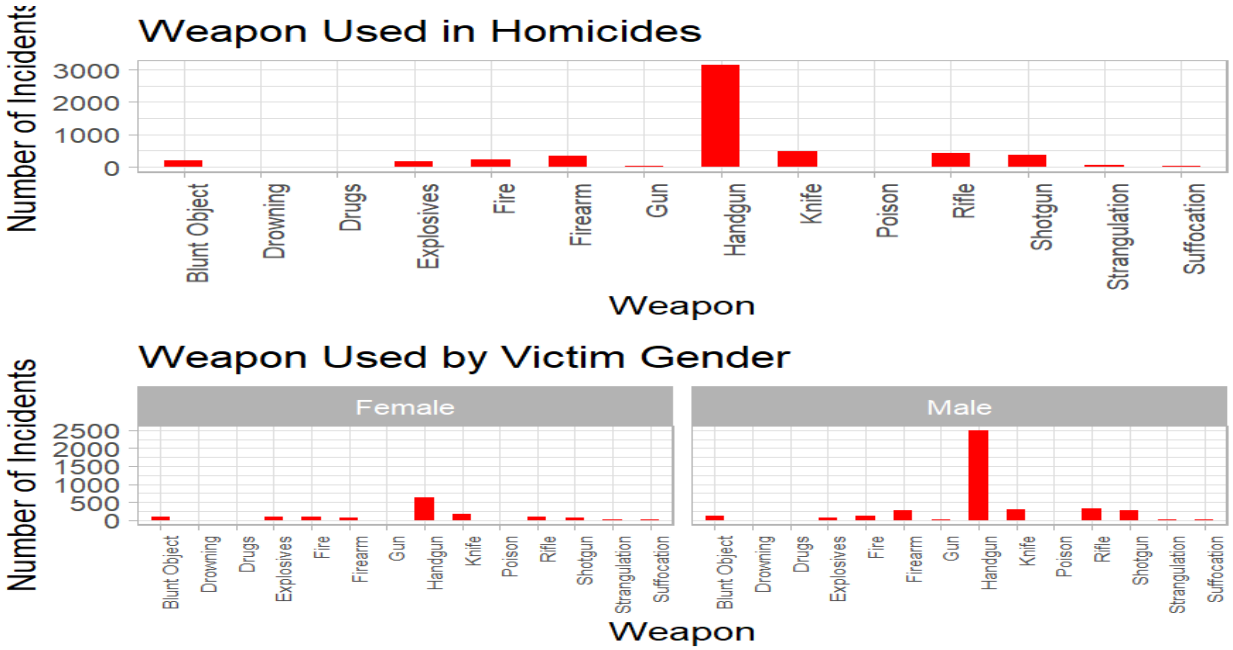
Factor Loadings Matrix

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

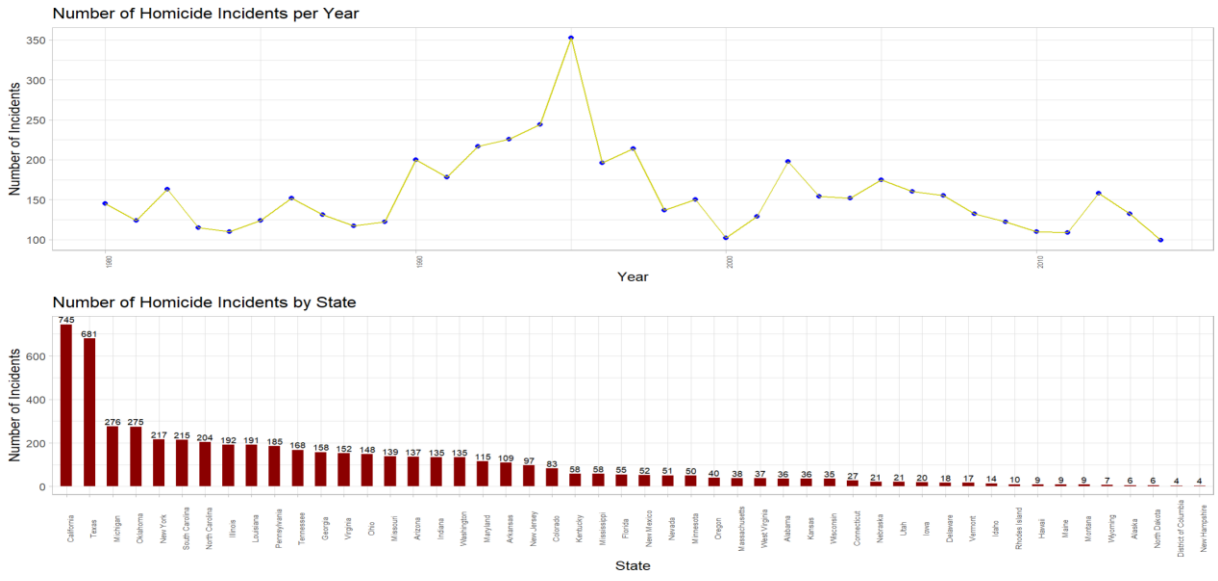
Variable	**ML1**	**ML2**
victim_age	0.297	
perpetrator_age	0.486	
victim_count		0.399
perpetrator_count	-0.179	0.368
SS loadings	0.362	0.296
Proportion Var	0.091	0.074
Cumulative Var	0.091	0.165

Both components were found to explain 16.5% of the overall variation: ML1 loads 0.297 on victim_age and 0.486 on perpetrator_age, explaining 9.1% of the variance, while ML2 loads 0.399 on victim_count and 0.368 on perpetrator_count, explaining 7.4%.

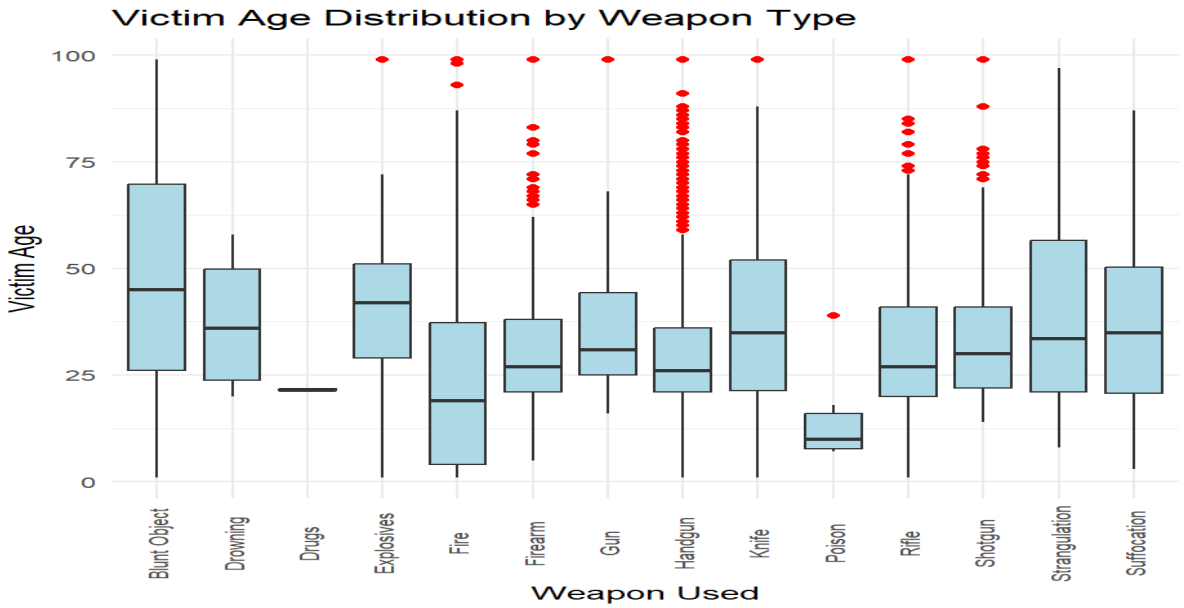
5. Exploratory Visualizations



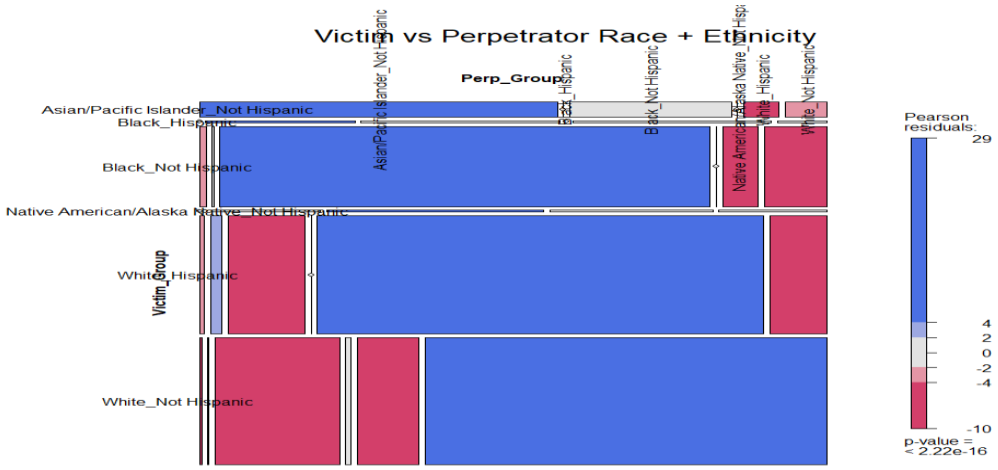
Handguns are the most common weapon overall. Female victims are more often harmed by close-range tools like knives and strangulation, while male victims face a broader mix of firearms (rifles, shotguns) as secondary weapons



Homicide incidents peaked in the mid-1990s before declining toward 2010.
California and Texas have the highest incident counts, far exceeding other states.



Victim ages vary by weapon: blunt objects and strangulation skew older (medians around 50), poison and drugs skew younger (medians around 25–30), and firearms/knives lie in between with wide age spreads.



Same-race pairings dominate: Black non-Hispanic and White non-Hispanic victim–perpetrator cells are highlighted in blue (more cases than expected), while cross-race combinations are bathed in red (fewer than expected), showing a strong non-random association between victim and perpetrator race/ethnicity.

Conclusion

- Two core dimensions—age and incident scale—explain 16.5 % of variance and split cases into “younger/small-scale” versus “older/large-scale” clusters.
- Homicides peaked in the mid-1990s and then declined; California and Texas show the highest counts.
- Handguns dominate overall; women face more close-range weapons (knives, strangulation), men face a broader mix of firearms.
- Victim ages differ by weapon: strangulation/blunt objects skew older, poisoning skews younger.
- Strong same-race victim–perpetrator pairing far outweighs cross-race incidents.