Crime Data Clustering Analysis Report

October 2025

1 Problem Statement

The objective was to perform clustering (K-means and Hierarchical) on a crime dataset to group 50 U.S. states based on crime rates (Murder, Assault, Rape) and urban population (UrbanPop), identify the optimal number of clusters, and draw inferences about crime patterns.

2 Dataset Description

The dataset contains 50 U.S. states with four features:

- Murder: Murder rates per 100,000 people (range: 0.8–17.4).
- Assault: Assault rates per 100,000 people (range: 45–337).
- UrbanPop: Percentage of urban population (range: 32–91).
- Rape: Rape rates per 100,000 people (range: 7.3–46.0).

The data was standardized to ensure fair clustering.

3 Clustering Results

Both K-means and Hierarchical clustering (with Ward's method and Euclidean distance) were applied, resulting in 4 clusters. The optimal number of clusters ($k = 4$) was assumed based on prior outputs (elbow plot and dendrogram).

3.1 K-means Clustering

Number of states per cluster:

- Cluster 0 (Low-crime, rural): 13 states
- Cluster 1 (High-crime, semi-urban): 8 states
- Cluster 2 (High-crime, urban): 12 states
- Cluster 3 (Moderate-crime, urban): 17 states

3.2 Hierarchical Clustering

Number of states per cluster:

- Cluster 0 (Moderate-crime, urban): 19 states
- Cluster 1 (High-crime, urban): 12 states
- Cluster 2 (Low-crime, rural): 12 states
- Cluster 3 (High-crime, semi-urban): 7 states

4 Cluster Characteristics and Inferences

4.1 Low-crime, rural states (K-means Cluster 0, Hierarchical Cluster 2)

States: Idaho, Iowa, Maine, Minnesota, Montana, Nebraska, New Hampshire, North Dakota, South Dakota, Vermont, West Virginia, Wisconsin
Characteristics: Low crime rates (Murder: $\sim$3.1–3.6, Assault: $\sim$76–78.5, Rape: $\sim$11.8–12.2), low urban population ($\sim$52).
Inference: Safe, rural states with minimal crime, likely due to low population density.

4.2 High-crime, urban states (K-means Cluster 2, Hierarchical Cluster 1)

States: Alaska, Arizona, California, Colorado, Florida, Illinois, Maryland, Michigan, Nevada, New Mexico, New York, Texas
Characteristics: High crime rates (Murder: $\sim$11.0, Assault: $\sim$264, Rape: $\sim$33.6), high urban population ($\sim$76.5).
Inference: Urban, metropolitan areas with significant crime challenges.

4.3 Moderate-crime, urban states (K-means Cluster 3, Hierarchical Cluster 0)

States: Connecticut, Delaware, Hawaii, Indiana, Kansas, Massachusetts, Missouri, New Jersey, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, Utah, Virginia, Washington, Wyoming (plus Kentucky and Arkansas in Hierarchical)
Characteristics: Moderate crime rates (Murder: $\sim$5.9–6.2, Assault: $\sim$141–142, Rape: $\sim$19.2), high urban population ($\sim$71–73.6).
Inference: Urban states with controlled crime levels, likely due to effective policing.

4.4 High-crime, semi-urban states (K-means Cluster 1, Hierarchical Cluster 3)

States: Alabama, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee
Characteristics: Very high crime rates (Murder: $\sim$13.9–14.7, Assault: $\sim$243–251, Rape: $\sim$21.4–21.7), moderate urban population ($\sim$53–54).
Inference: Semi-urban states with significant crime issues, possibly due to socioeconomic challenges.

5 Comparison of Clustering Methods

The Adjusted Rand Index (ARI = 0.8849) indicates strong agreement between K-means and Hierarchical clustering. Key correspondences:

- K-means Cluster 0 $\leftrightarrow$ Hierarchical Cluster 2 (Low-crime, rural)
- K-means Cluster 1 $\leftrightarrow$ Hierarchical Cluster 3 (High-crime, semi-urban)
- K-means Cluster 2 $\leftrightarrow$ Hierarchical Cluster 1 (High-crime, urban)
- K-means Cluster 3 $\leftrightarrow$ Hierarchical Cluster 0 (Moderate-crime, urban)

Mismatches (2 states):

- Arkansas: K-means Cluster 1 (High-crime, semi-urban) vs. Hierarchical Cluster 0 (Moderate-crime, urban).
- Kentucky: K-means Cluster 0 (Low-crime, rural) vs. Hierarchical Cluster 0 (Moderate-crime, urban).

These mismatches reflect boundary cases due to intermediate crime rates or urban population values.

6 Mismatch Analysis

Table 1 shows the feature values for Arkansas and Kentucky compared to cluster means.

Analysis:

- **Arkansas:** K-means assigns it to Cluster 1 (High-crime, semi-urban) due to its high assault rate (190) and moderate urban population (50), closer to Cluster 1 means (Assault: 243.625, UrbanPop: 53.750). Hierarchical assigns it to Cluster 0 (Moderate-crime, urban) due to its murder rate (8.8) and rape rate (19.5), aligning with Cluster 0 means (Murder: 6.211, Rape: 19.184).
- **Kentucky:** K-means assigns it to Cluster 0 (Low-crime, rural) due to its low urban population (52) and assault rate (109), closer to Cluster 0 means (UrbanPop: 52.077, Assault: 78.538). Hierarchical assigns it to Cluster 0 (Moderate-crime, urban) due to its higher murder rate (9.7), closer to Cluster 0 mean (Murder: 6.211).

Table 1: Feature Values for Mismatched States and Cluster Means

	Murder	Assault	UrbanPop	Rape
Arkansas	8.8	190	50	19.5
Kentucky	9.7	109	52	16.3
K-means Cluster Means				
Cluster 0	3.600	78.538	52.077	12.177
Cluster 1	13.938	243.625	53.750	21.412
Cluster 2	10.967	264.000	76.500	33.608
Cluster 3	5.853	141.176	73.647	19.335
Hierarchical Cluster Means				
Cluster 0	6.211	142.053	71.263	19.184
Cluster 1	10.967	264.000	76.500	33.608
Cluster 2	3.092	76.000	52.083	11.833
Cluster 3	14.671	251.286	54.286	21.686

7 Visualizations

7.1 Cluster Size Visualization

A bar chart (`cluster_sizes.png`) compares the number of states in each cluster:

- Low-crime, rural: 13 (K-means) vs. 12 (Hierarchical)
- High-crime, semi-urban: 8 (K-means) vs. 7 (Hierarchical)
- High-crime, urban: 12 (K-means) vs. 12 (Hierarchical)
- Moderate-crime, urban: 17 (K-means) vs. 19 (Hierarchical)

The chart highlights the larger size of the moderate-crime, urban cluster in Hierarchical clustering due to Arkansas and Kentucky.

7.2 Scatter Plots

Scatter plots visualize cluster separation:

- `cluster_scatter_plots.png`: Murder vs. Assault for K-means and Hierarchical clusters.
- `urbanpop_rape_scatter.png`: UrbanPop vs. Rape for K-means and Hierarchical clusters.

These plots illustrate how clusters differ across key features, with UrbanPop vs. Rape highlighting the role of urbanization in rape rates.

8 Key Findings

- Crime patterns are strongly tied to urbanization: high-crime clusters are urban or semi-urban, while low-crime clusters are rural.
- The high-crime, semi-urban cluster (e.g., Alabama, Georgia) suggests regional socio-economic challenges, particularly in Southern states.
- The consistency between K-means and Hierarchical clustering (ARI = 0.8849) validates the robustness of the identified patterns.
- Mismatches (Arkansas, Kentucky) highlight states with ambiguous characteristics.

9 Limitations and Next Steps

- The optimal number of clusters ($k = 4$) was assumed; elbow plot and dendrogram descriptions could confirm this choice.
- Additional visualizations (e.g., scatter plots of Murder vs. Rape) could provide deeper insights.
- Further analysis of mismatched states could explore socio-economic or regional factors.
- The dataset is limited to four features; additional variables (e.g., income, education) could enhance clustering.