CHRIS SELIG

STATE CLUSTERING PROJECT

Problem Statement

As part of an ongoing effort to reduce homicide rates and to focus discussion between the 50 states, the task is to group related states into a small number of groups to help facilitate discussions.

The Data

The data provided includes the homicide (murder), assault, and rape rates per 100,000 people. The final variable in the data set is the percentage of a states population that lives in an urban area. Below is a sample of the data.

Table 1: Sample of State Data

State	Murder	Assault	Rape	% Urban Population
Alabama	13.2	236	21.2	58
Alaska	10.0	263	44.5	48
Arizona	8.1	294	31.0	80
Arkansas	8.8	190	19.5	50
California	9.0	276	40.6	91
Colorado	7.9	204	38.7	78

A more thorough exploration of the data can be found in the appendix, under “Exploratory Data Analysis.”

Recommendation

The recommended grouping is provided by the hierarchical clustering algorithm. The three cluster solution has the highest silhouette coefficients of the three methods (k-means, hierarchical, and finite mixture models) explored.

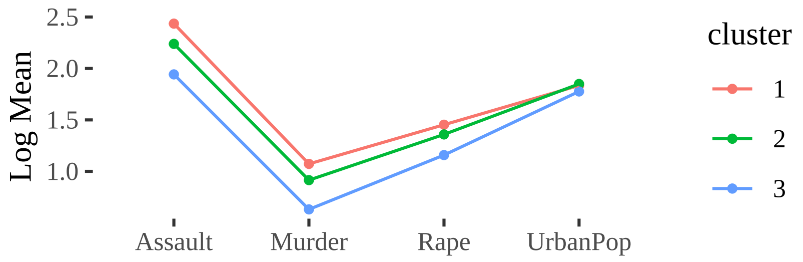


Figure 1: Cluster Classification

Above, the results of the chosen algorithm. The clusters can be characteried as follows:

Cluster 1: Has the highest amount of violent crimes (assault, murder, rape) of the four clusters, but is only the second highest percentage of ubran population. Sixteen states are assigned to this cluster.

Cluster 2: Characterized by the highest percentage of urban population and has the second smallest values for the other variables. Fourteen states are assigned to this cluster.

Cluster 3: Cluster three is the more rural of the three clusters and has the least amount violent crime of all the clusters. This cluster has the largest amount of states (20) of the three clusters.

Here are the states assigned to each cluster:

Cluster 1
Alabama
Alaska
Arizona
California
Delaware
Florida
Illinois
Louisiana
Maryland
Michigan
Mississippi
Nevada
New Mexico
New York
North Carolina
South Carolina

Cluster 2
Arkansas
Colorado
Georgia
Massachusetts
Missouri
New Jersey
Oklahoma
Oregon
Rhode Island
Tennessee
Texas
Virginia
Washington
Wyoming

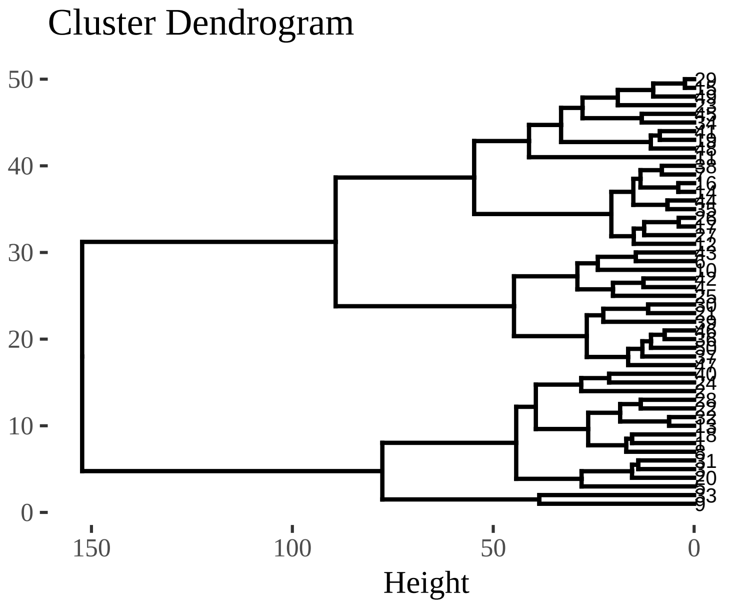
Cluster 3
Connecticut
Hawaii
Idaho
Indiana
Iowa
Kansas
Kentucky
Maine
Minnesota
Montana
Nebraska
New Hampshire
North Dakota
Ohio
Pennsylvania
South Dakota
Utah
Vermont
West Virginia
Wisconsin

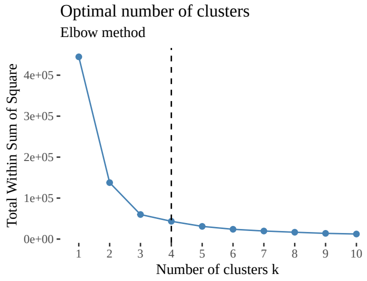
Methodology

Preferred Cluster Solution

Hierarchical Clustering

As a first pass at hierarchical clustering, I calculated the dendrogram below using euclidean distance and the average linkage method. Below is the result.

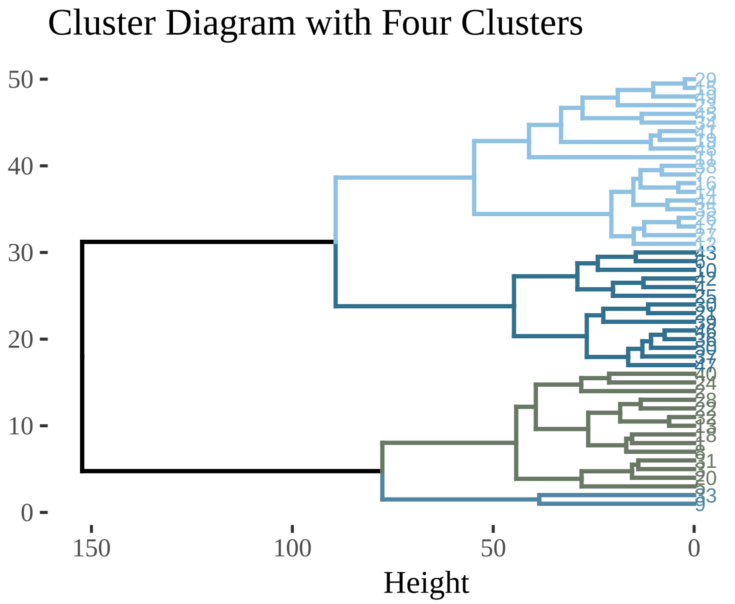




A dendrogram in this format is difficult to interpret with all the clusters, so a decision needs to be made on what the optimal number of clusters are.

At right, a common method to pick the optimal number of clusters is the "elbow method." The elbow method attempts to minimize the total within cluster sum of squares (WSS) so that adding another cluster does not add too much to the WSS. This results in a bend or "elbow" in the plot. You can see a bend at four clusters. Although, it looks like a three cluster or a 5 cluster solutions is also worth exploring.

Below is the refined dendrogram showing the four clusters by color.



Ultimately, the question is how well do the clusters fit the data? One way to answer this is calculate internal validation methods: compactness (how close data points are within a cluster); separation (how far clusters are from each other). The silhouette coefficent will be used to assess the clusters.

In the silhouette plot below, the silhouette coefficents for each data point are shown. There are a few notes of interest. First, the overall average silhouette width is only 0.5. Optimally this number should be as close to 1 as possible. Second, in cluster 1, there are some points with a negative silhouette coefficent. This means that the data points are most likely in the wrong cluster.

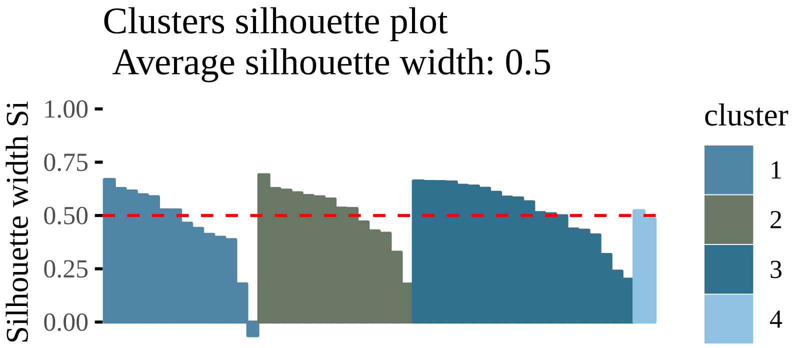


Figure 2: Four Cluster Silhouette Width Plot

Taking a look at the average silhouette for each cluster (below), you can see that all of the average widths are pretty low.

Cluster	Avg Silhouette Width
Cluster 1	0.4544827
Cluster 2	0.5134926
Cluster 3	0.5220297
Cluster 4	0.5027647

Can we do better? Let's try with a three cluster solution instead. Below is the silhouette plot for the 3 cluster solution. With an average silhouette width of 0.53, a better result has been obtained. Also of note, the negative silhouette coefficents have been eliminated.

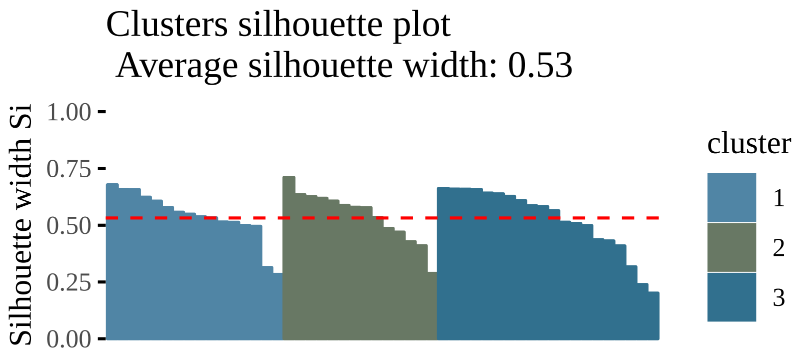


Figure 3: Three Cluster Silhouette Width

Alternative Cluster Models

K-Means Clustering

K-means clustering is an unsupervised learning technique used to group similar data points together. The number of clusters are manually defined and data points are put in the clusters in a way to minimize the in cluster sum of squares. To begin, a couple of decisions need to be made:

1. Whether or not to scale the data: Since the assault variable is quite a bit larger than the others variables, I will scale them.
2. The number of clusters: To do this, another elbow plot will be created using the kmeans algorithm. At right, are the results of the elbow method. After 4 clusters the slope is not very steep so on the first try of the algorithm I will use the four clusters.

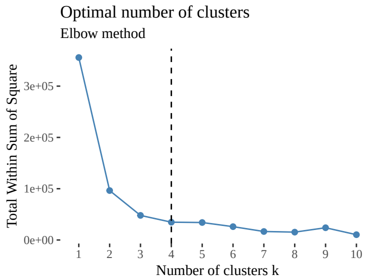


Figure 4: Kmeans Elbow Method

Below is the results of the kmeans clustering algorithm. Remember, the goal is to have well defined clusters with maximized space between them. Clusters 2 and 3, especially, appear to be potentially the same cluster and more generally, there is quite a bit of overlap between the four clusters.

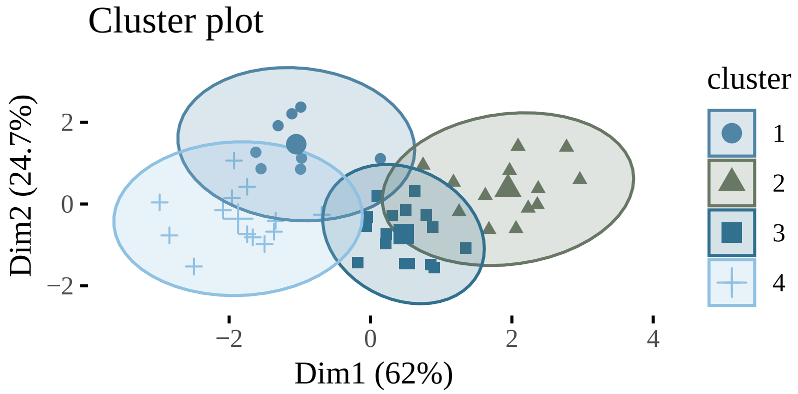


Figure 5: Kmeans Four Cluster Solution

To confirm that the kmeans clustering algorithm did not do a very good job at clustering the variables, below is the silhouette plot. With an average silhouette width of 0.34, the algorithm performed significantly worse than the hierarchical clustering. Also of note, some of the data points in cluster 4 have a negative silhouette index so are probably in the wrong cluster.

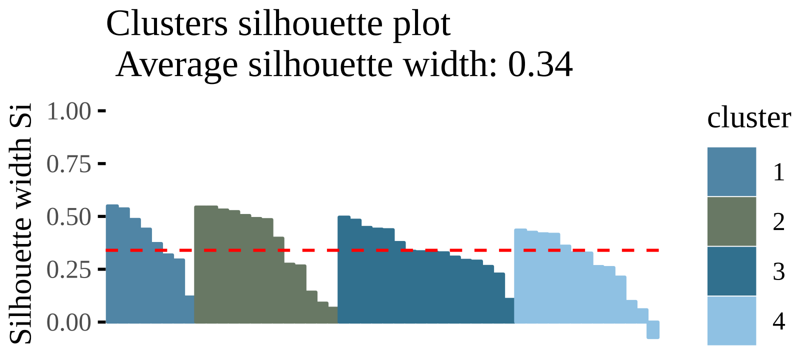


Figure 6: Kmeans Silhouette Width

Finally, I wanted to try the 3 and 5 cluster solutions to see if those are any better. Turns out they are not. At right is the silhouette plots for the 3 and 5 cluster solutions. The average silhouette width for them is even worse.

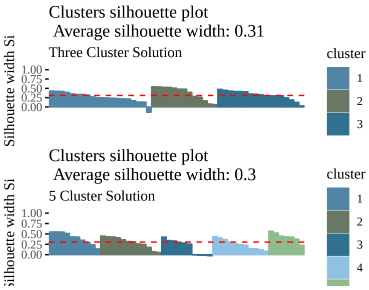


Figure 7: Three and Five Cluster Solutions

Finite Mixture Model

Finite mixture models are a probablistic method of clustering data points. Below, two models are used based on having the lowest Bayesian Information Criterion (BIC).

The best two models are a "EEI" and "VEI" models. The three cluster EEI model is a diagonal model with equal volume and shape where a VEI model is diagonal, with varying volume and equal shape and has four clusters. Below is the silouette coefficents for the VEI model. The average silhousette width is 0.3744, which is better than the k-means clustering, but not has good as the hierarchical clustering.

Table 3: Avg Silhouette Width: 0.3744

cluster	avg sil width
1	0.4562220
2	0.5297521
3	0.2150087

The EEI model performed even worse than the VEI model. The average silhouette width for the clusters is only 0.2279. Even cluster one has a negative silhouette width which implies that the data points in the cluster probably below to a different cluster.

Table 4: Avg Silhouette Width: 0.2279

cluster	avg sil width
1	-0.03617657
2	0.16118898
3	0.53933620
4	0.20581104

Appendix

Exploratory Data Analysis

Summary Statistics

Table 5: Summary Statistics

variable	num	mean	median	variance	std	min	max	iqr	skewness	kurtosis
assault	50	170.760	159.00	6945.16571	83.337661	45.0	337.0	140.000	0.2205325	-1.1454869
murder	50	7.788	7.25	18.97047	4.355510	0.8	17.4	7.175	0.3706342	-0.9492304
rape	50	21.232	20.10	87.72916	9.366385	7.3	46.0	11.100	0.7537694	0.0751026
urbanpop	50	65.540	66.00	209.51878	14.474763	32.0	91.0	23.250	-0.2126297	-0.8719550

Assault

At right, the histogram for the assault variable. With 10 bins, the distribution is multi-modal as it has 3 distinct peaks and appears to have 3 clusters. Median 159 is less than the mean 170.76 so the assault variable is right skewed. A positive skewness confirms this. Finally, with a negative excess kurtosis, this variable is platykurtic, meaning the data distribution is thin-tailed.

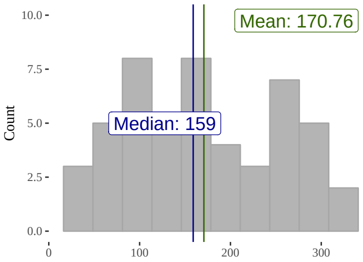


Figure 8: Assault

Murder

The murder histogram (at right) is bi-modal, with a long right tail suggesting at least 2 clusters. The median 7.25 is less than the mean 7.788 so the murder variable is right skewed. A positive skewness confirms this. The murder variable also has a negative excess kurtosis and is thin-tailed.

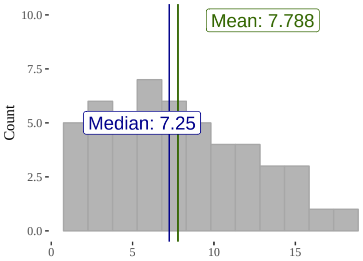


Figure 9: Murder

Rape

Median 20.1 is less than the mean 21.232 so the rape variable is right skewed. A positive skewness value confirms this. The rape variable has a positive kurtosis and is leptokurtic, which means it is a fat tailed distribution. The histogram, at right shows the fat right tail and suggests 2 clusters.

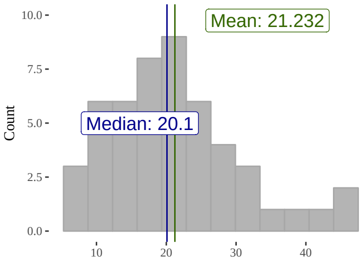


Figure 10: Rape

Urban Population

At right is the histogram for the percentage of population in each state that is urban. As you can see it is bi-modal, with two distinct peaks as well as a a long left tail. The two peaks and long left tail suggest potentially 3 clusters. Median 66 is greater than the mean 65.54 so the percentage of population that is urban variable is slightly left skewed. A negative skewness value confirms the left skew. As with assault and the murder variables, this variable also has a negative excess kurtosis so it is a thin tailed distribution.

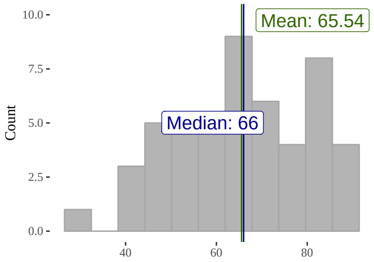
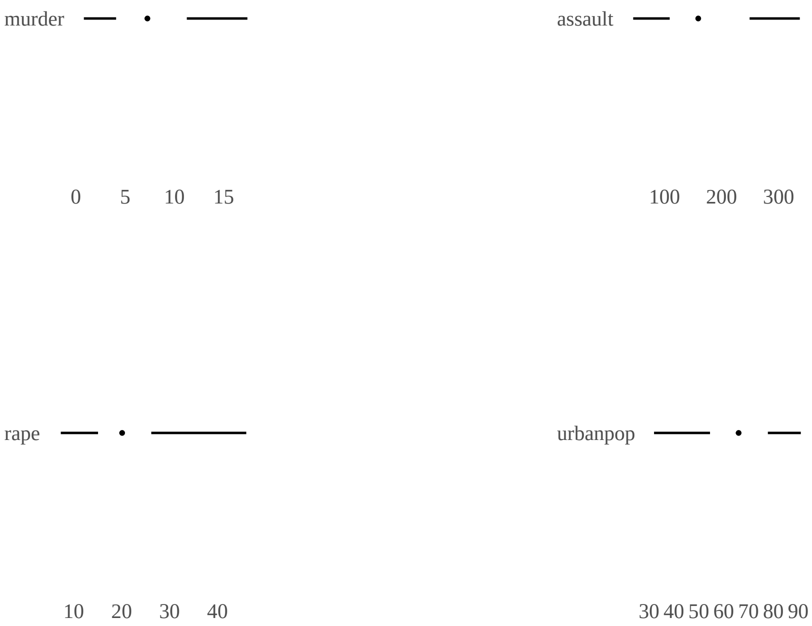


Figure 11: Urban Population Percentage

Outlier Detection

Below, four boxplots are arranged to show a univariate approach to outlier detection. The dot in the middle of them is the median, while the horizontal lines begin at the 25th and 75th percentiles and extend to the end of the 1.5 * the IQR (interquartile range). Any dots outside of this range are tagged as outliers. Since there are no data points outside of the horizontal lines, no outliers are detected.



Individually, none of the variables displayed any outliers, however, collectively, the variables should also be examined. To do this, the Cook's distance will be calculated. Cook's distance estimates the influence a data point using regression analysis.

At right, a multiple linear regression model has been fit with the data. The horizontal line is the Cook's distance, using the traditional 4/n criterion. In this case, everything above 0.08 is an outlier. Three points are flagged: Mississippi, Vermont and North Dakota. Since the goal is to have all states at the discussion, it is inappropriate to remove them from the dataset (and the discussion groups).

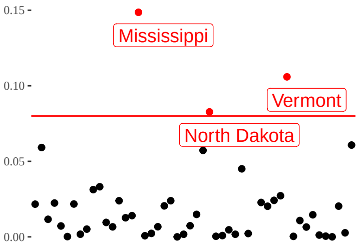


Figure 12: Outlier Dection Using Cooks Distance