Policy Analysis

Ryan Marinelli

30th June 2019

Introduction

The goal of this analysis is to understand spatial relationships with crime. The data consists of arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Thus, we are using panel data for this analysis. The main tool used in this study is cluster analysis. There are several algorithms deployed here to review relationships. The first tool is hierarchical clustering using both agglomerative and divisive modeling. This analysis is followed by secondary clustering using k-means and k-medoids to add further robustness to the analysis.

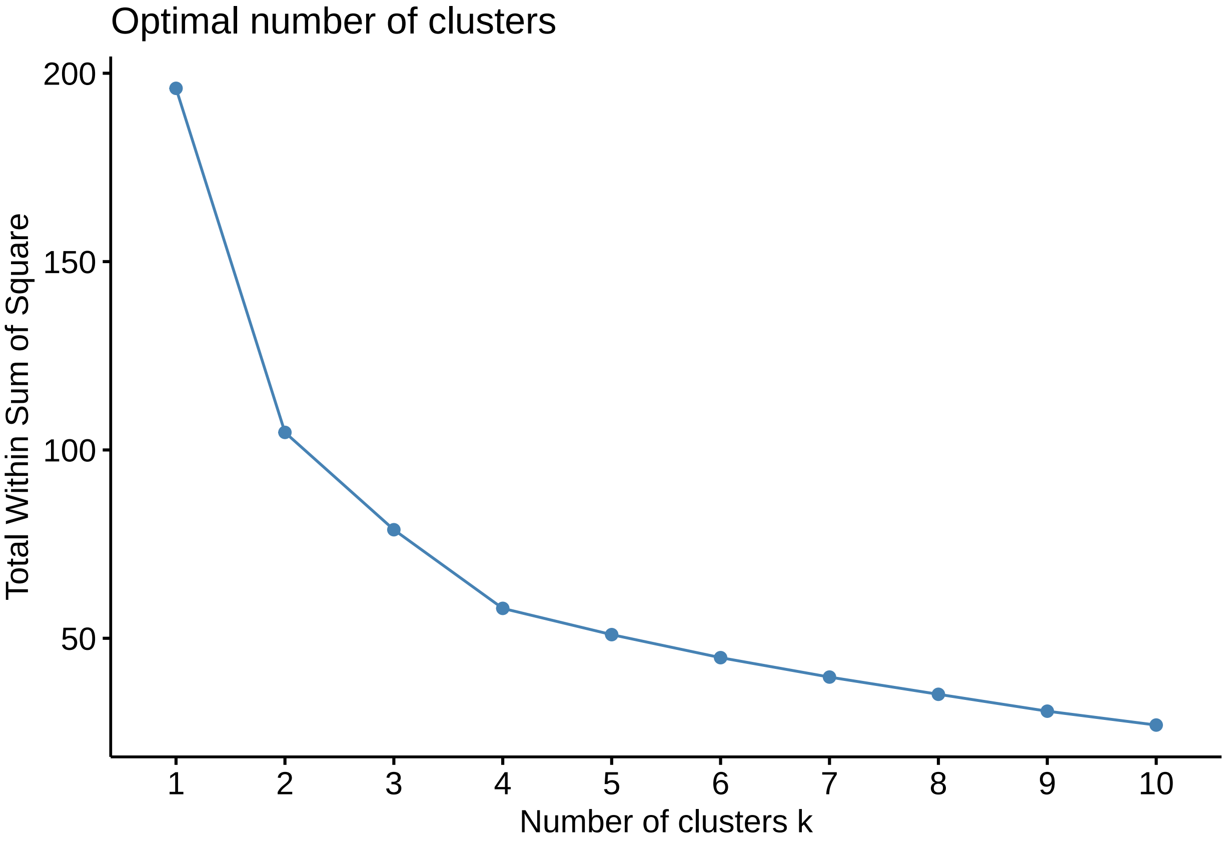
Holistic Analysis

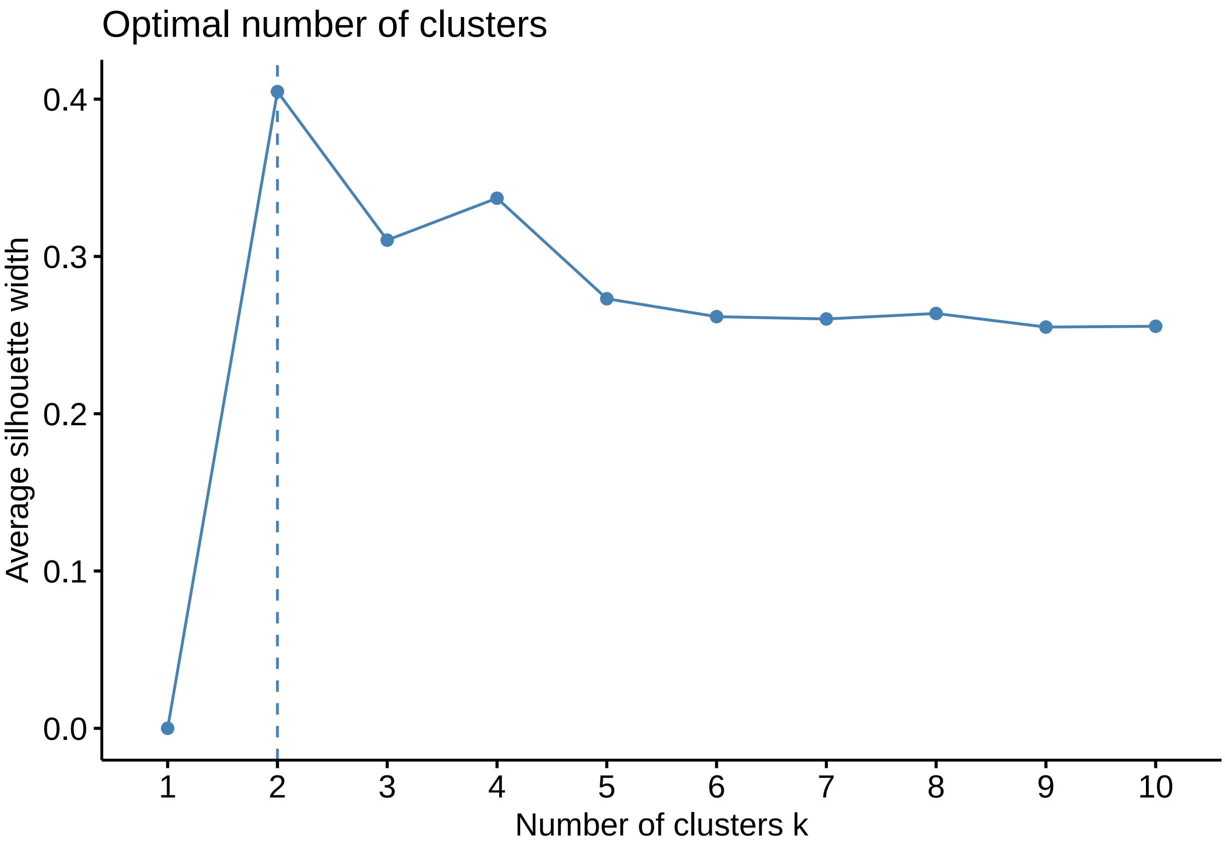
The underlying goal of this analysis is to understand how crime is affecting clusters of states. The way crime affects states is largely generalizable when considering data at the cluster level. It appears the data can be considered in two categories: states that have a high amount of crime and a low amount of crime. For this analysis, three clusters are used with a third cluster acting as a buffer to properly group the less extreme cases. The high crime states appear largely to be from the south and the more populous states. For instance, California is clustered with the high crime states, but it has the population higher than most European countries.The lower crime states are generally the lesser populated. These states are found in the Midwest and Northeast.When considering the hierarchies that found through the clustering process, this trend becomes more pronounced. When considering the divisive hierarchies in particular, it is a rather stark trend. One point that is in interest, which is echoed throughout the clustering analysis, is that the states that are moderately safe are rather heterogenous. In the divisive clustering model, the states that are safer are segmented at a higher height than the states that are considered less safe. The clusters for the safe states are also more partitioned.They are also more diverse geographically speaking. This suggests that some of issues that relate to ameliorating crime are not related to spatial considerations.

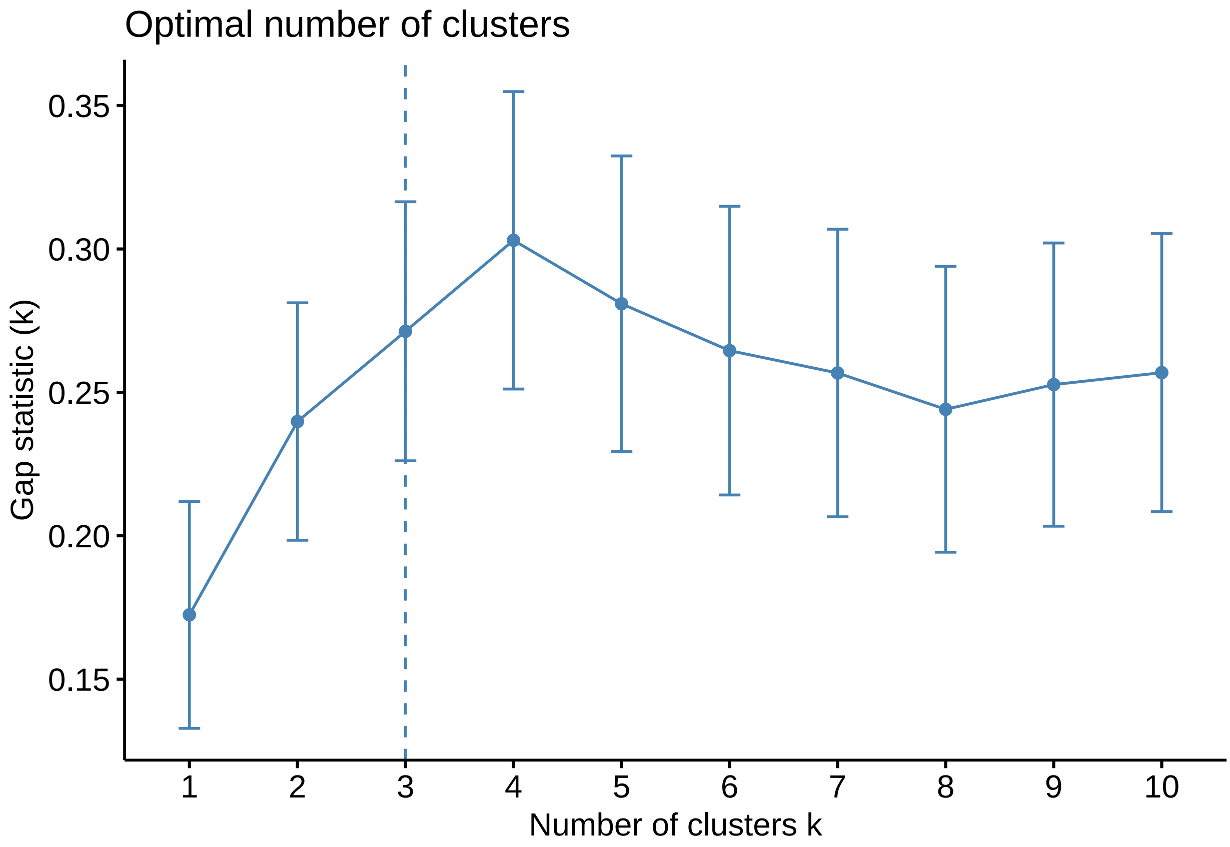
When reviewing K-means, it is possible to see the relationships between states more clearly. The safest states are rather similar, but the moderately safe states vary significantly. The cluster for safe states is more spherical in shape and the cluster is more structured. When considering the sub-clusters, this point is reinforced. Most of the Northwest and Midwest are clustered together with some variation with the less populous western states present. A similar pattern occurs when looking at the sub-cluster of dangerous states as well. The southern states make a cluster that is well pronounced, while the most populated states make up the other cluster. The states that are both southern and large are closest to the middle of the plot. This suggests that they are being pulled to the southern state cluster, while they are still being grouped in the larger state cluster. These states are namely Texas and Florida. The cluster that needs the most consideration is the moderately safe states cluster. They are the most heterogenous cluster. This sub-cluster must be segmented into seven sections and experiences the greatest amount of variation in terms of geography. This is also somewhat expected considering they act as a boundary between states that are considered unsafe and safe.

Cluster Selection

To begin the analysis, there needs to be testing of the data to understand what is considered a reasonable number of clusters to use. The first graph here is using the elbow method. This method minimizes within-cluster variation as the criteria for selection. The second graph uses silhouette analysis to determine clusters. Instead of within-cluster variation, it uses the average distance of each point in a cluster is to the nearest neighboring cluster. This allows for the selection of a well-defined cluster. Lastly, the gap statistic is used. The gap statistic reviews the relationship of the total within intra-cluster variation and its expectation by essentially using hypothesis testing. The goal is to see if the formation of clusters is statistically different from a random uniform distribution and to maximize the GAP statistic. Through reviewing these selection criteria, using three clusters seems reasonable. This is supported by the GAP statistic. As the GAP statistic is a statistical test, it should hold more weight in modeling than more holistic analysis.




Optimal number of clusters
Average silhouette width
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
8
9
10
Number of clusters k



Method Selection

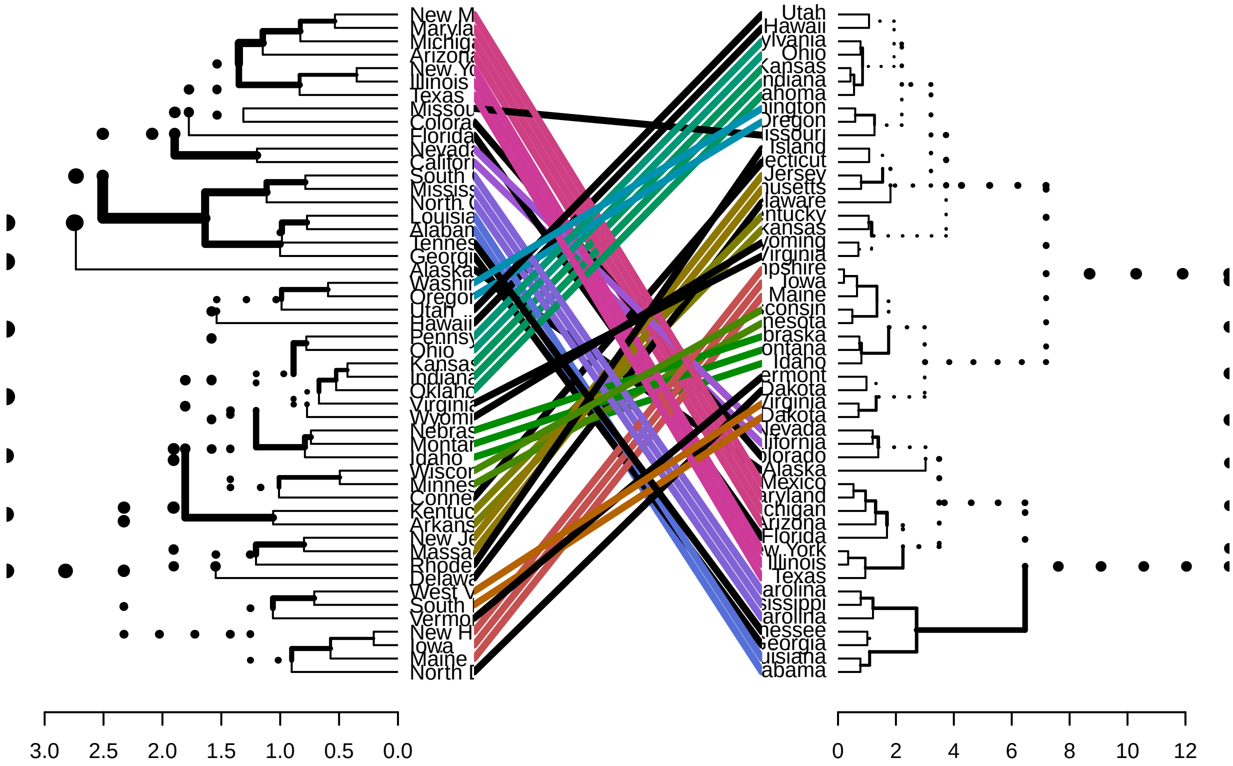
Hierarchical Clustering

The first method being evaluated is aggregate hierarchical clustering. There are several methods considered to implement this clustering procedure. The complete method calculates dissimilarity between all the points of a cluster and all the points of the other cluster and maximizes dissimilarity. Likewise, single link clustering uses the smallest of these values, while average linkage uses the mean dissimilarity. Wald's method minimizes variance of within cluster variation and minimized between-cluster variation with pairs. The agglomerative coefficient is a tool that determines the strength of clusters and was used to make the determination of which method to apply.

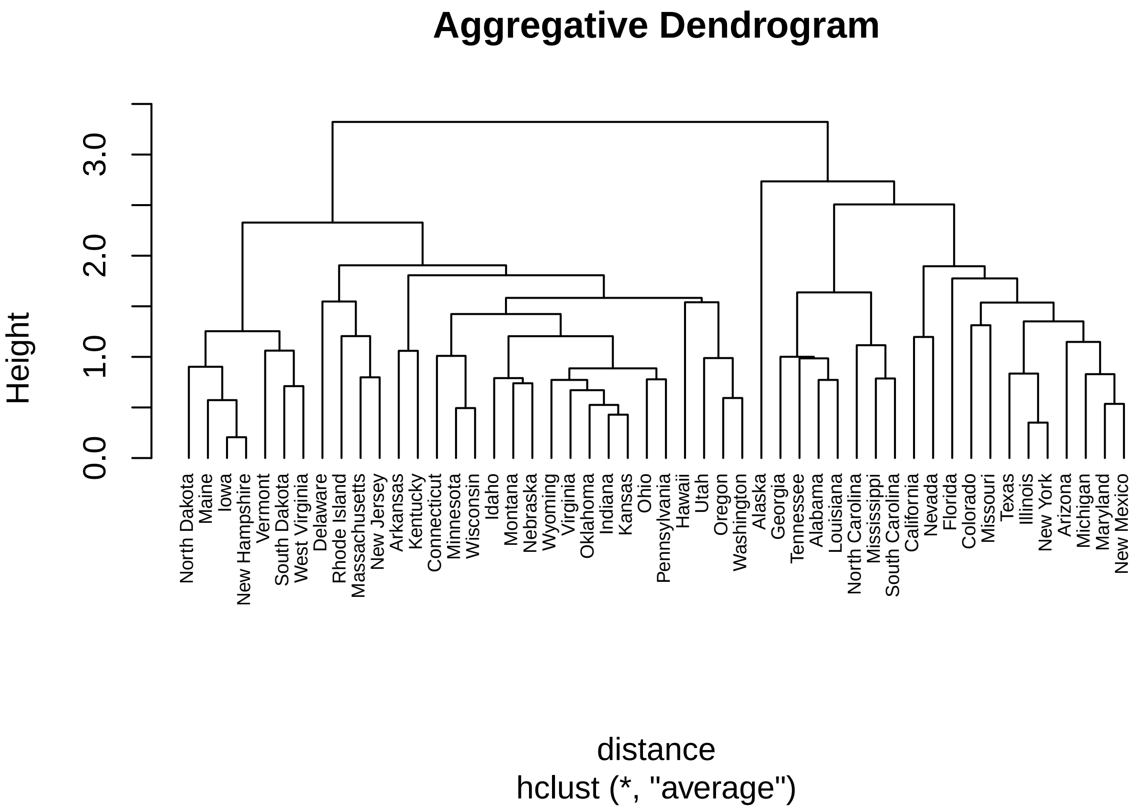
This dendrogram enables us to see the larger picture. Through aggregation, it highlights likes features that are common between sets. It appears that population is one of the more significant factors. For instance, North Dakota, Maine, Iowa, and New Hampshire are clustered together. Those states are sparsely populated and filled with wilderness. The states that are more dangerous and are more populated seem to be more rightward. New York and Texas are part of the same for example.

Aggregation Coefficents	Clustering Methods
.73	Average Linkage
.62	Single Link
.85	Complete Linkage
.93	Ward

In order to ensure robustness of results, a secondary method is used. Through using a tanglegram, it is possible to compare congruency between dendrograms. It appears that the methods yield similar results which suggests that the clusters are rather consistent. The most signficant difference between these methods appears to be the placement of clusters but not the actual composition of them.



```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

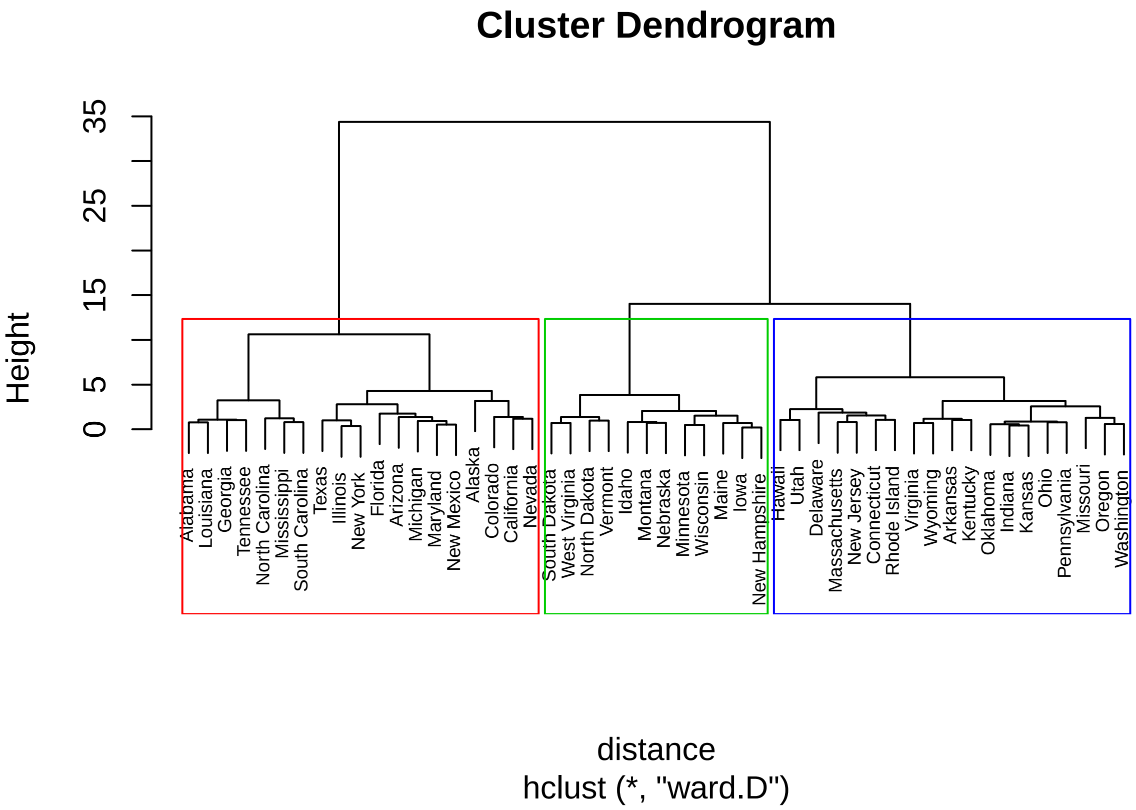


Divisive Clustering

The second hierarchical clustering method applied in this analysis is divisive clustering. This analysis provides more emphasis on factors that differentiate clusters. In this dendrograms, the clusters have had colored boxed transposed upon them. The red cluster appears to be the states that have higher crime. The green cluster appears to have mostly low crime states and the blue cluster appears to have mostly moderate crime states.

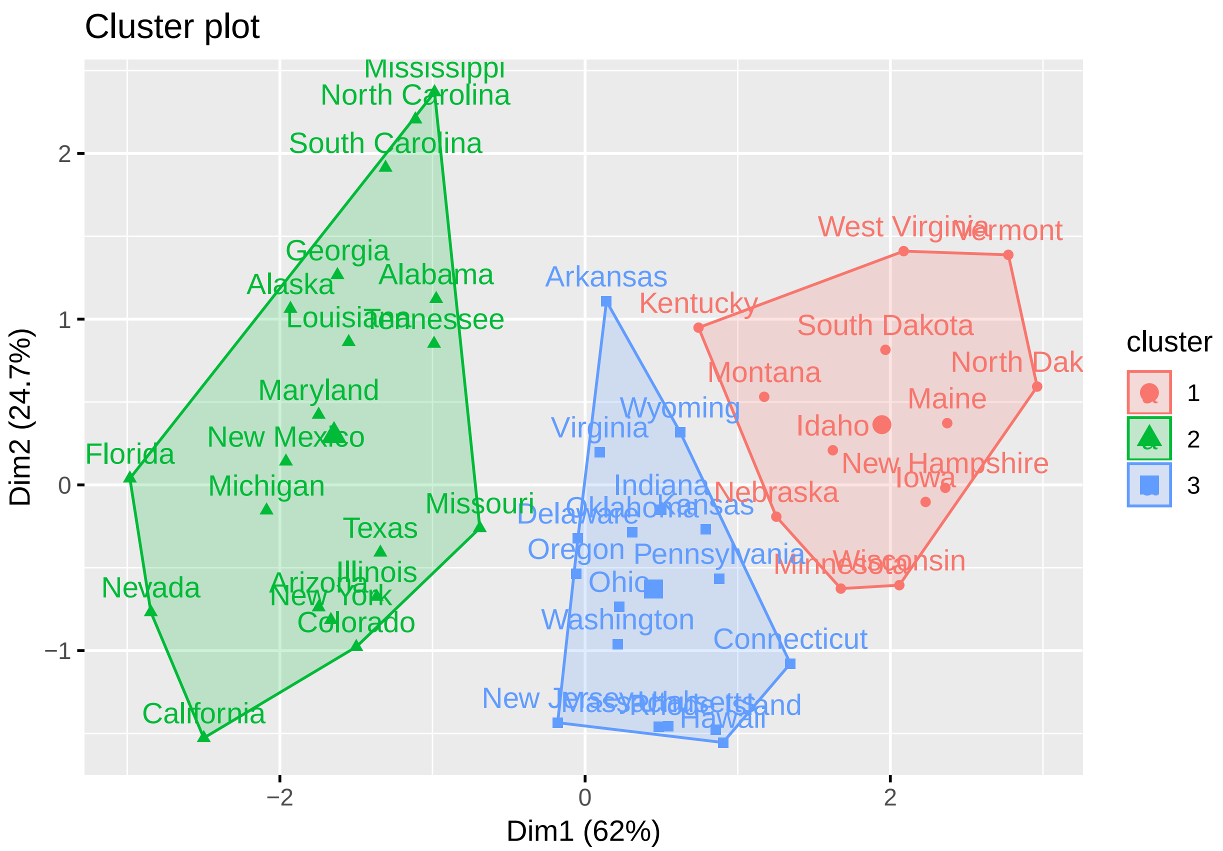
```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
## grouped
##  1  2  3
## 19 19 12
```



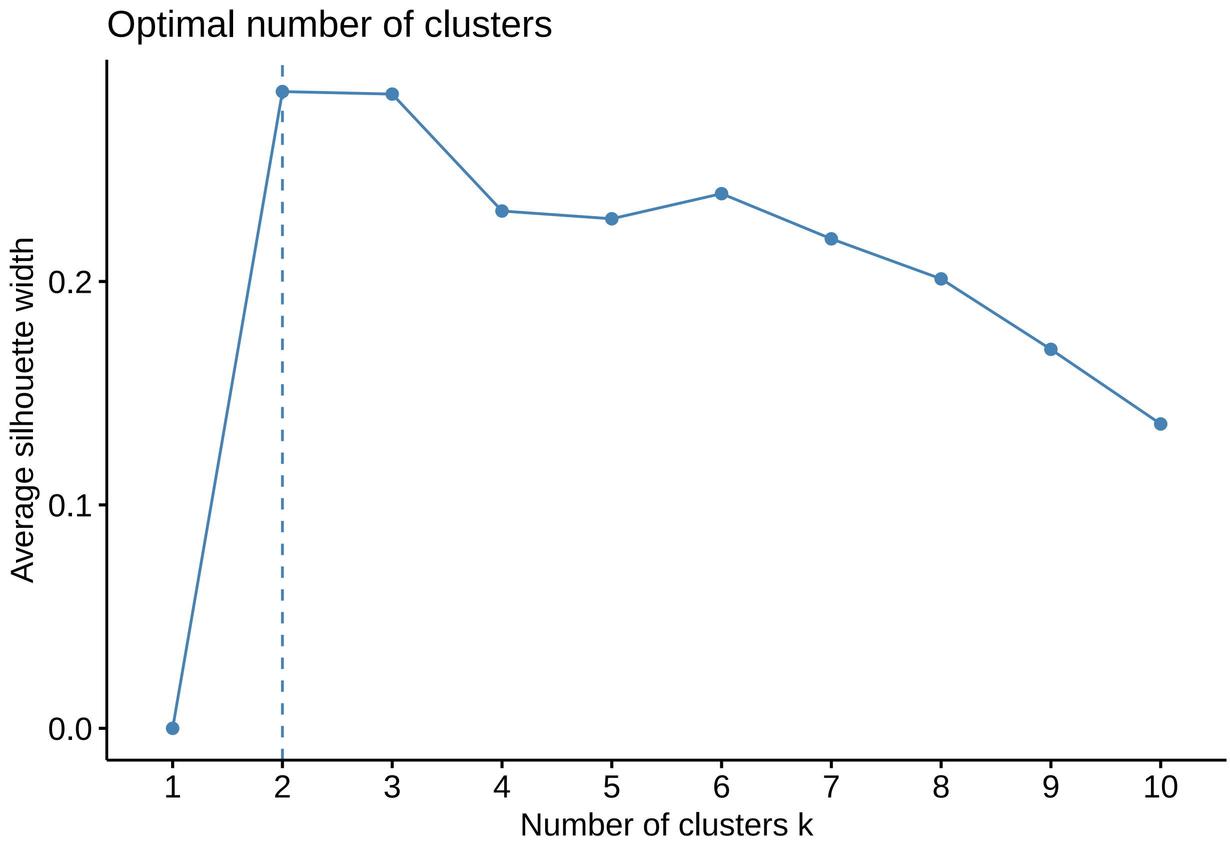
K-Means Analysis

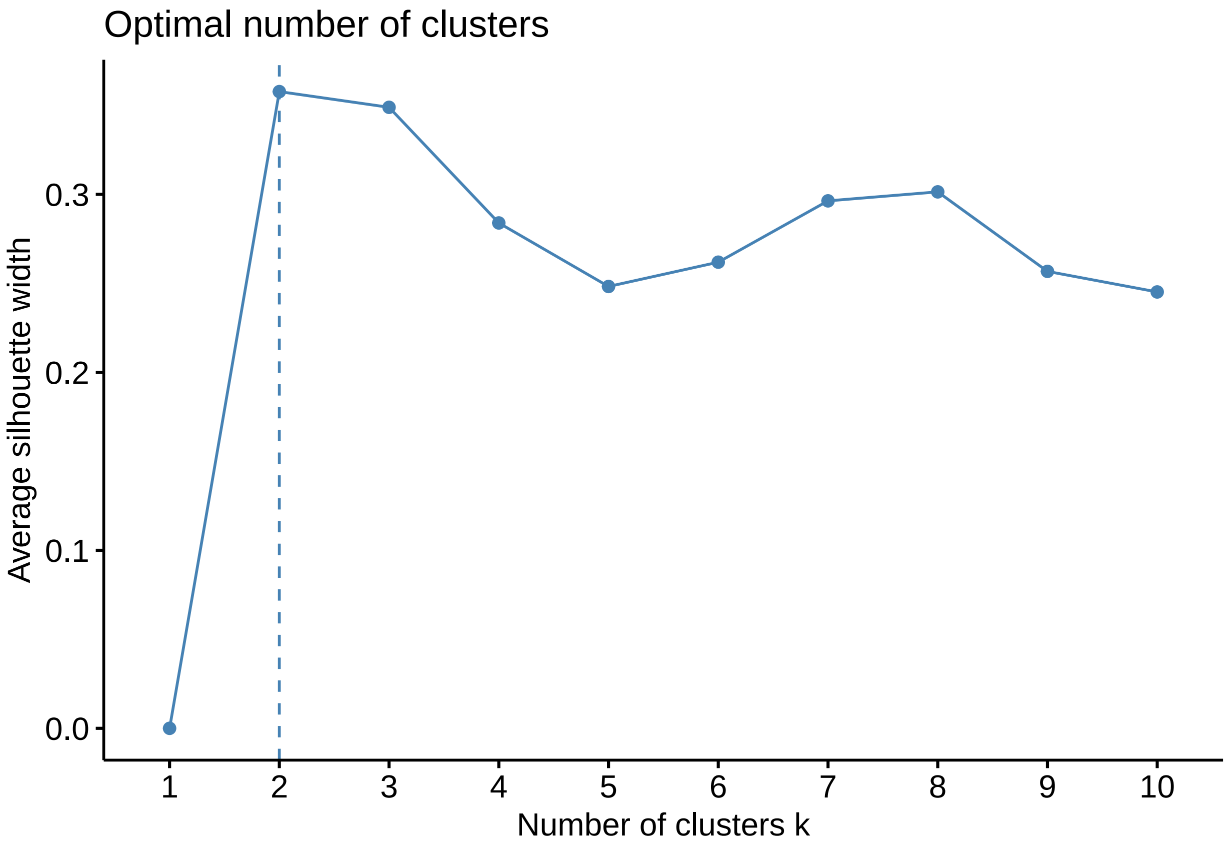
The analysis is continued with using K-means. This algorithm essentially uses a center point and tries to miniminze disance between all the points of a cluster. The paritions in kmeans seem to emulate the other clustering methods. The more dangerous states are in the left most cluster, and the safest states are in the right most.

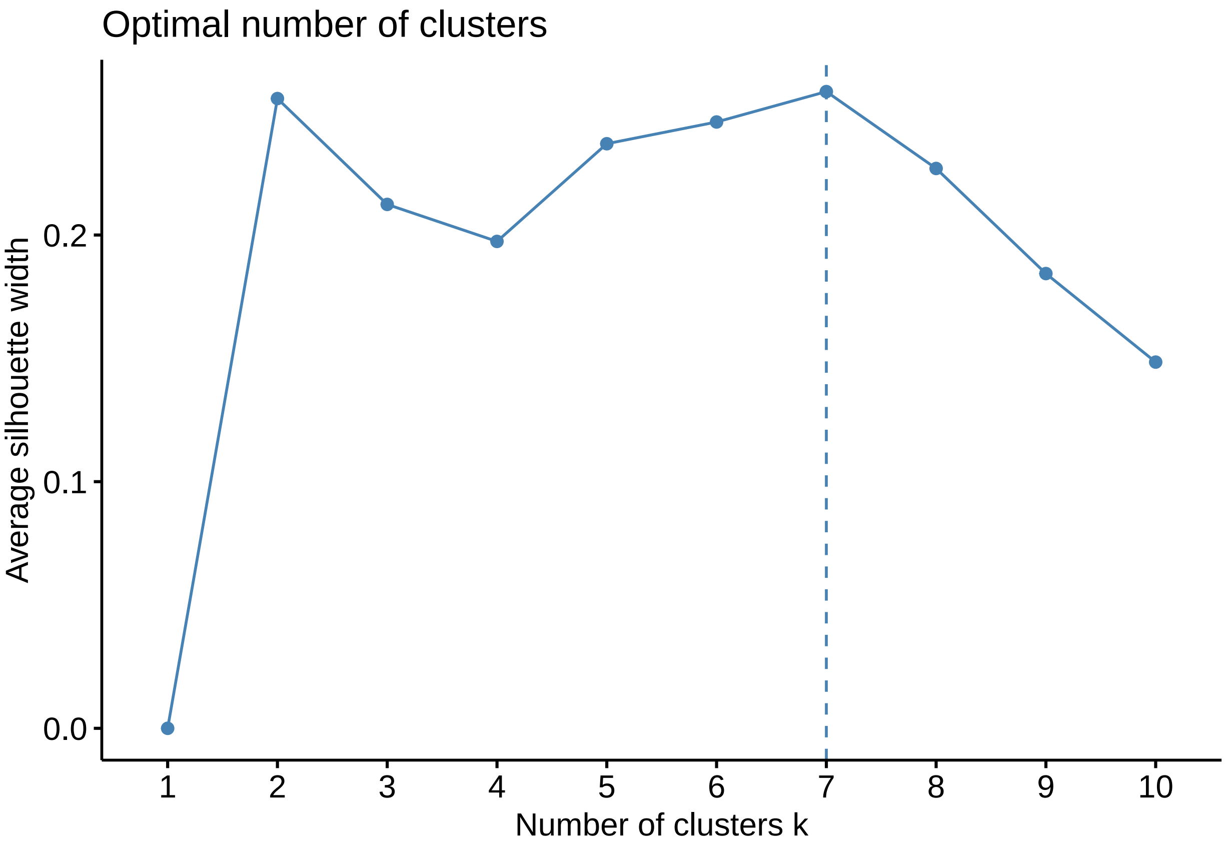


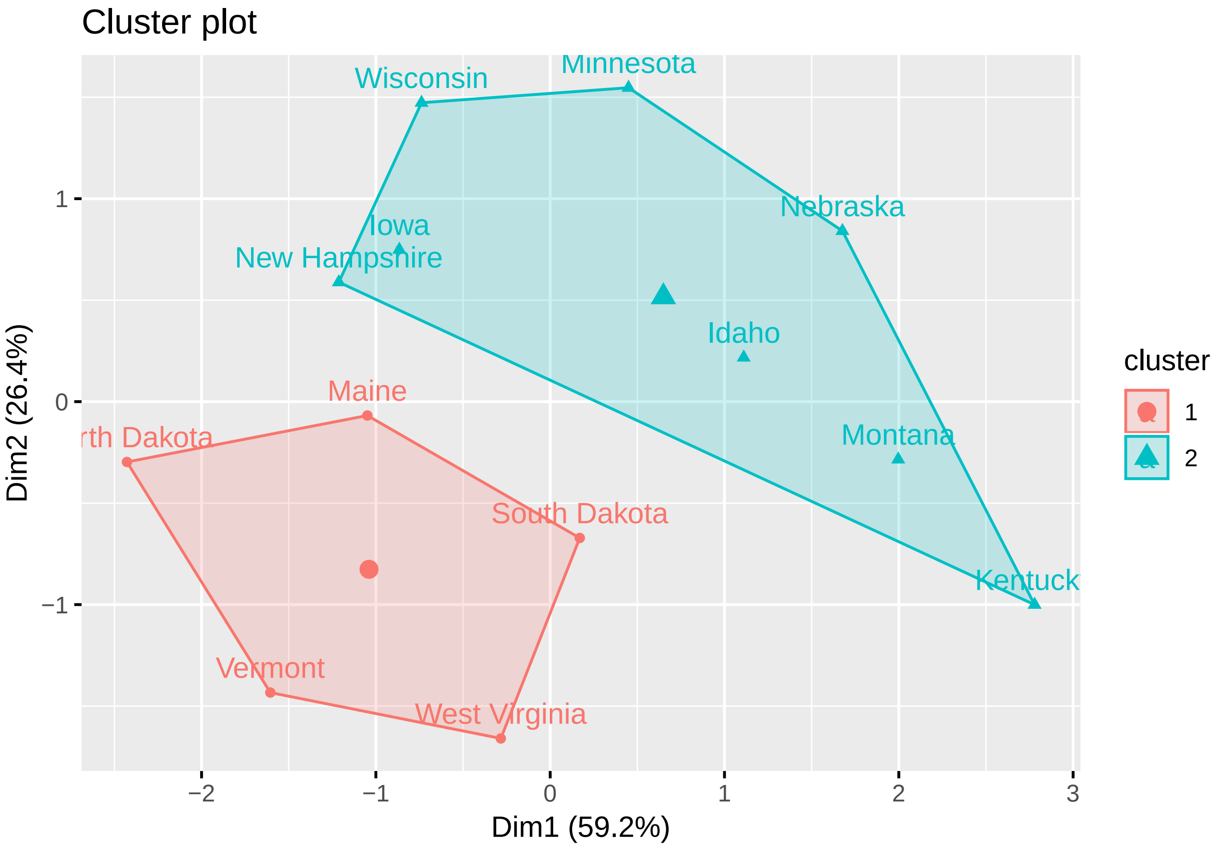
Secondary Analysis of Clusters

Through further analysis of these clusters, we obtain a deeper understanding of the interplay between states. K-means in applied to each cluster to understand the relationship of states within each cluster. The same general methodology is applied as before in attempting to determine the amount of clustered applied to each grouping of states. In this instance, the silhouette method is applied to determine the number of clusters.

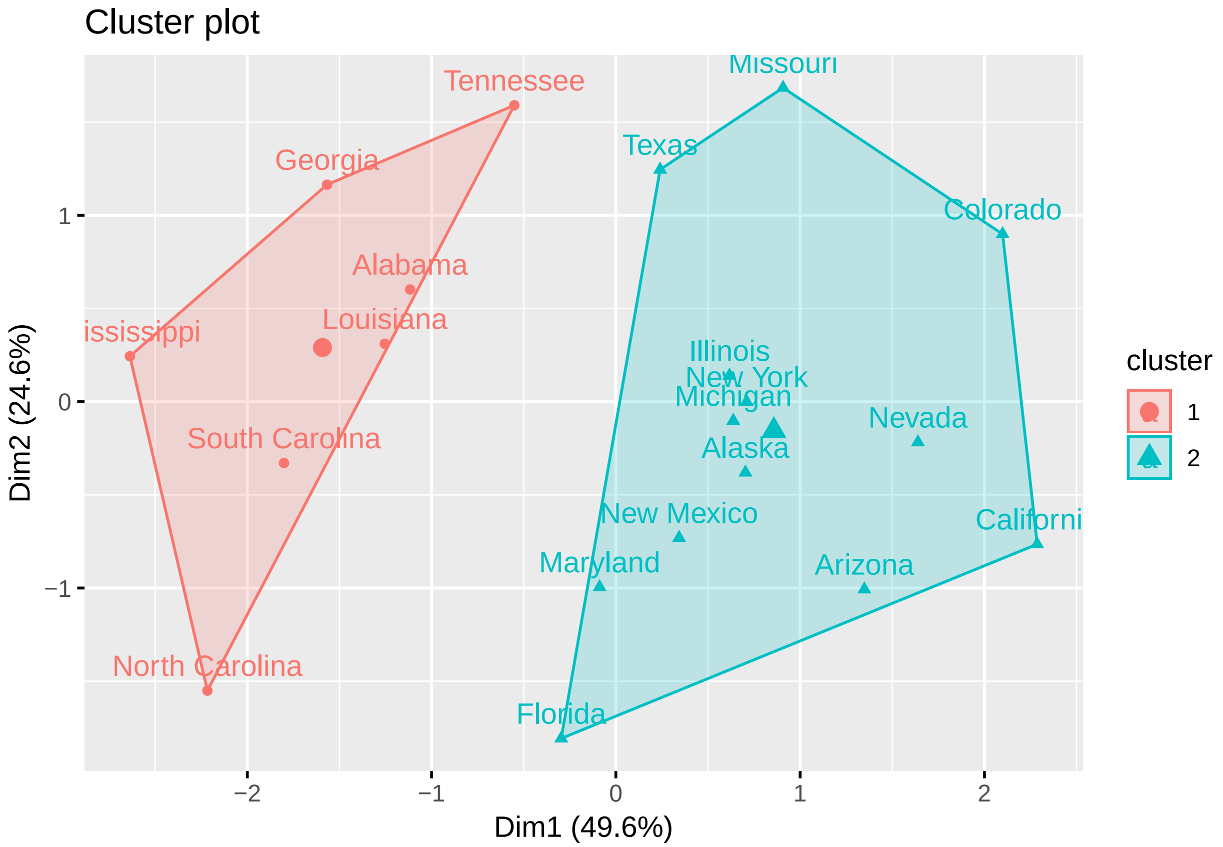

Optimal number of clusters
Average silhouette width
0.0
0.1
0.2
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Optimal number of clusters
Average silhouette width
0.0
0.1
0.2
0.3
1
2
3
4
5
6
7
8
9
10
Number of clusters k

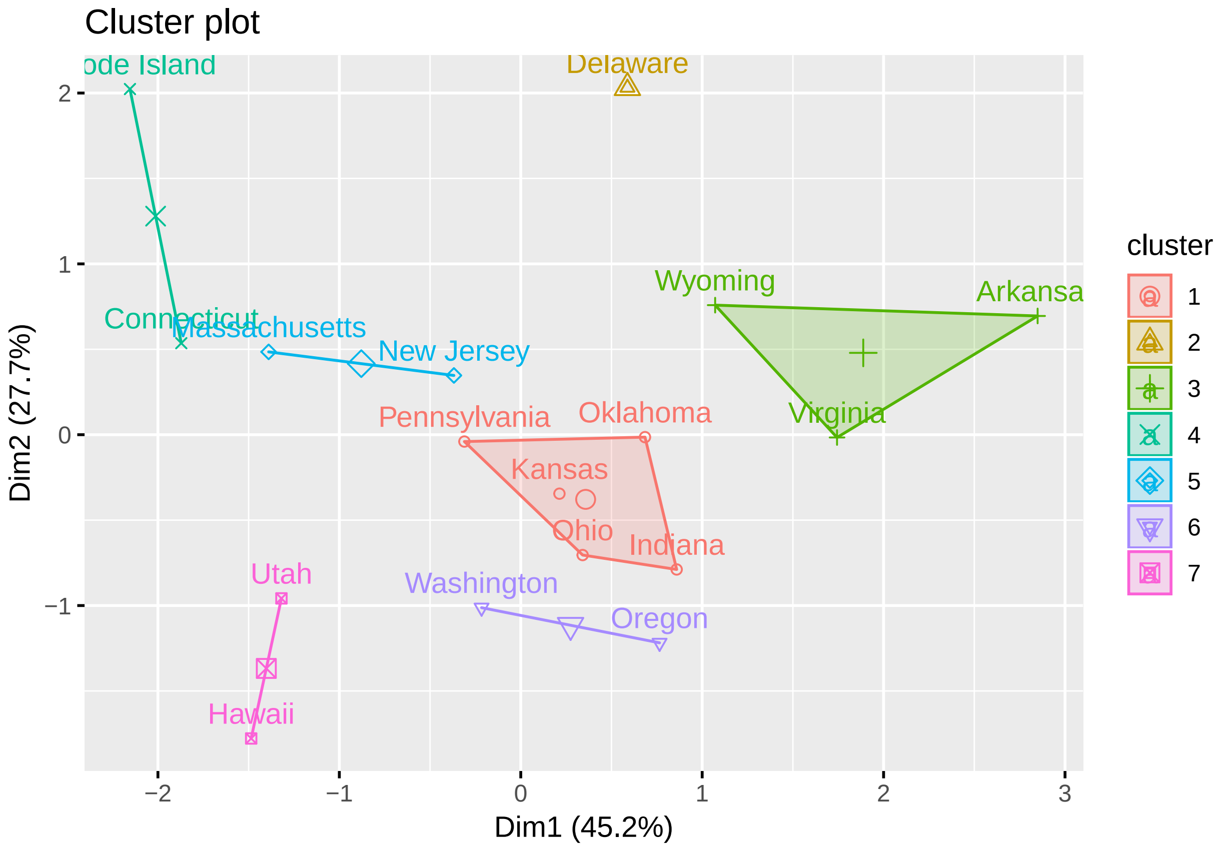

Optimal number of clusters
Average silhouette width
0.0
0.1
0.2
1
2
3
4
5
6
7
8
9
10
Number of clusters k



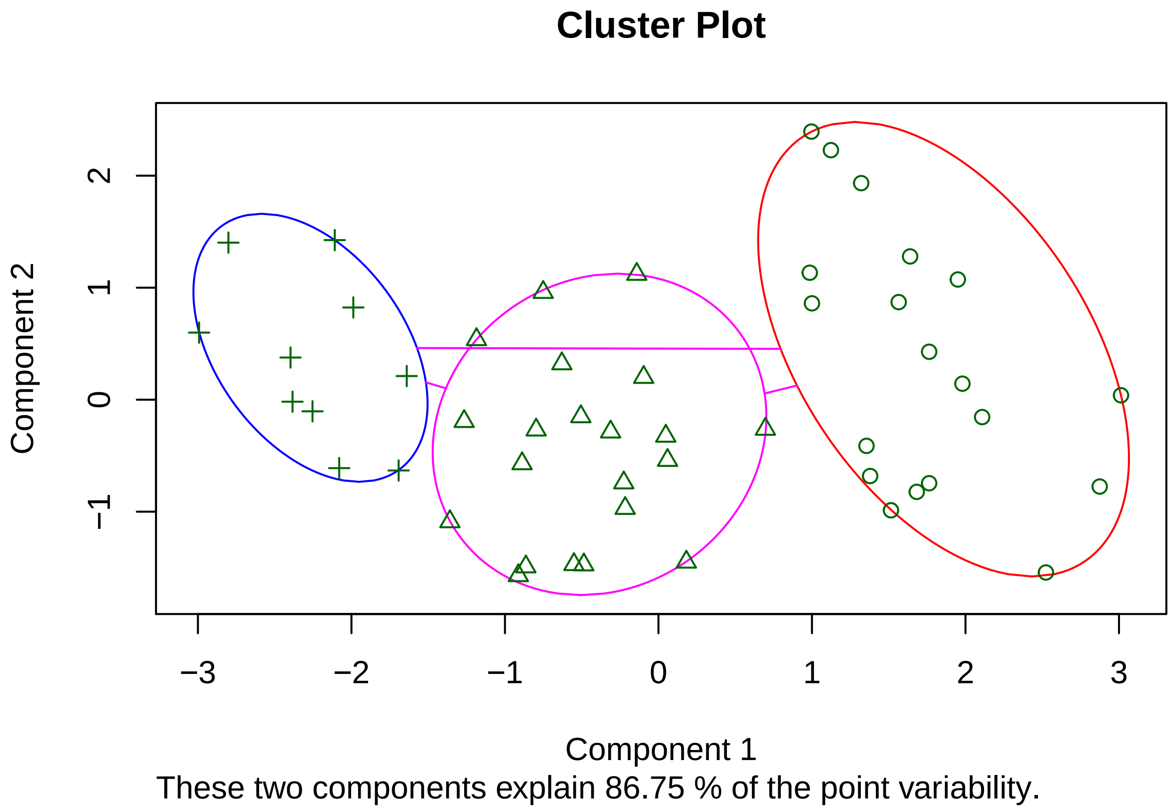
It appears difficult to glean more information in this clustering. The one point of interest is that geography is not an overriding factor. North Dakota is further away than North Dakota than to other points. Likewise, Maine and Vermont are on opposite parts of this cluster. Likewise, New Hampshire is clustered near Iowa. This suggests that there is significant variation in crime neighboring states.



This cluster is rather interesting. The leftward cluster is composed of all southern states. The rightward cluster is more heterogenous. The clustering in the rightward cluster groups Alaska New York and other more populated states. It also suggests that some of southern states on the east coast are more similar to the northern states.



This clustering seems to be the most heterogenous. Silhouette analysis suggested to use seven clusters. Some of these clusters make sense. For instance, Washington and Oregon are very similar. However, the groupings in the middle are interesting. It is surprising that Virginia would be in the same cluster as Wyoming and Arkansas and that Pennsylvania would be in the same cluster as Oklahoma. The most interesting aspect of this cluster is that it the most heterogenous of the 3 clusters used initially.



Lastly, k-medoids were implemented. The main reason for doing so is for a robustness check. The overall structure of the clusters is anaglous to kmeans. Because K-medoids minimizes the summation of pairwise dissimilarities for its criteria, it is less biased due to the presense of outliers in the data.

Policy Reccomendations

When considering policy formation, triage of issues should be at the forefront of thought. Thus, it appears that focusing enforcement efforts in the south would be the most effective method to reduce crime from a national level. Specifically, it would be advisable to promote interoperability between Alabama and Louisiana. As they are neighboring states with similar issues, pooled resources that are tailored to meet their issues should be jointly applicable. Also, as those states are most near the centroid, the solutions that are deployed in those states may be equally suited to address the issues of the states on the fringes of the cluster. In terms of the other sub-cluster for unsafe states, it is a bit more problematic. Most of the states are simply large and do not have a shared culture as the southern states share. One strategy that should be considered would be to promote collaboration in terms of policing strategy for the police departments of Detroit and Chicago. As Detroit is infamous for its crime rate, it is likely one of the main factors increasing the crime rate in Michigan. Likewise, Chicago is an outlier within Illinois causing a spike in the state's crime rate. The two states are also neighboring states as well near the centroid of their cluster. Thus, the solutions that are defined through the collaboration would be equally applicable to the other states in the cluster.

Conclusion

The focus of future policy strategies should focus on the cluster of problematic states. The other two clusters of states either do not have an issue with crime or are too diverse to many a cohesive strategy to target problems. The problematic states demonstrate patterns that can be leveraged to form the basis of an effective strategy to target crime. By creating strategic partnerships with neighboring states near the centroids of each of the sub-cluster, it would be possible to define solutions through collaboration, that will be broadly applicable to the whole cluster of high crime states and will enable relief to reach the states that require it most.