Chapter 6

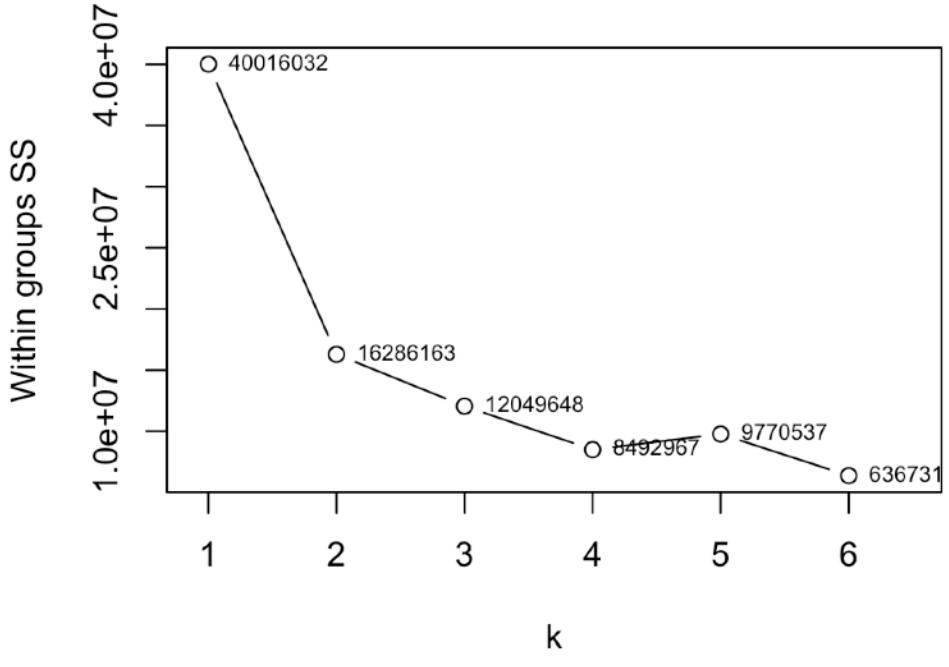
Cody Frisby

10/1/2017

My dataset is one that includes crime data for each state, including the District of Columbia. The crime variables include **murder**, **rape**, **robbery**, **assault**, **burglary**, **theft**, and **vehicle**.

How do we decide the number of clusters to use when we can't visualize the data? Sometimes the "elbow" method can work, except when there isn't a distinct "elbow" in the data. The data here is the within groups sum of squares and we plot that against the number for k. If there is an "elbow" in the plot, we choose that number for k. Here it appears a "good" number for k is 2.

R

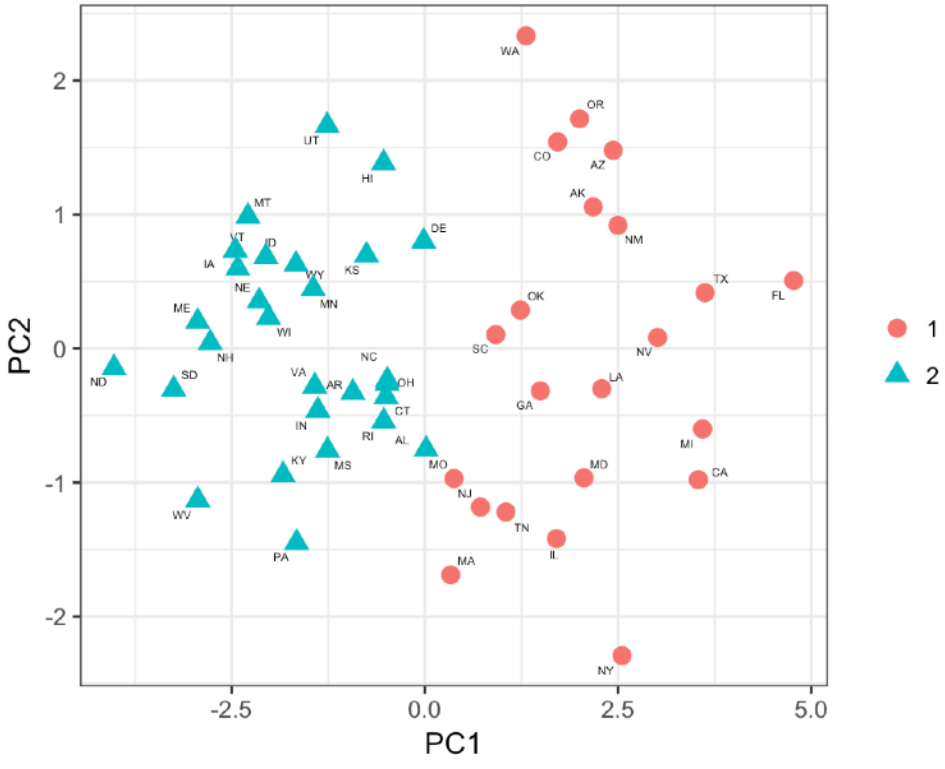


If we fit the model with $k = 2$ we get the following centroids.

Cluster	Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
1	2.71	3.13	2.10	2.84	3.68	4.48	2.64
2	1.37	1.71	0.68	1.31	2.20	3.41	1.18

There are 7 variables, so we can't visualize them all together, but we can use the first two principal components to reduce our dimensions and visualize them by coloring the points by cluster to see how we did and if the result makes sense with our intuition.

Note: I first performed this analysis without excluding DC and without standardizing the variance of each variable. In failing to do both these things, UT was grouped in the cluster with the highest murder rate when it actually is 11th lowest in the dataset.



Rapid Miner

If I do this same analysis in RapidMiner without standardizing the variables, but excluding DC, I get the following centroids.

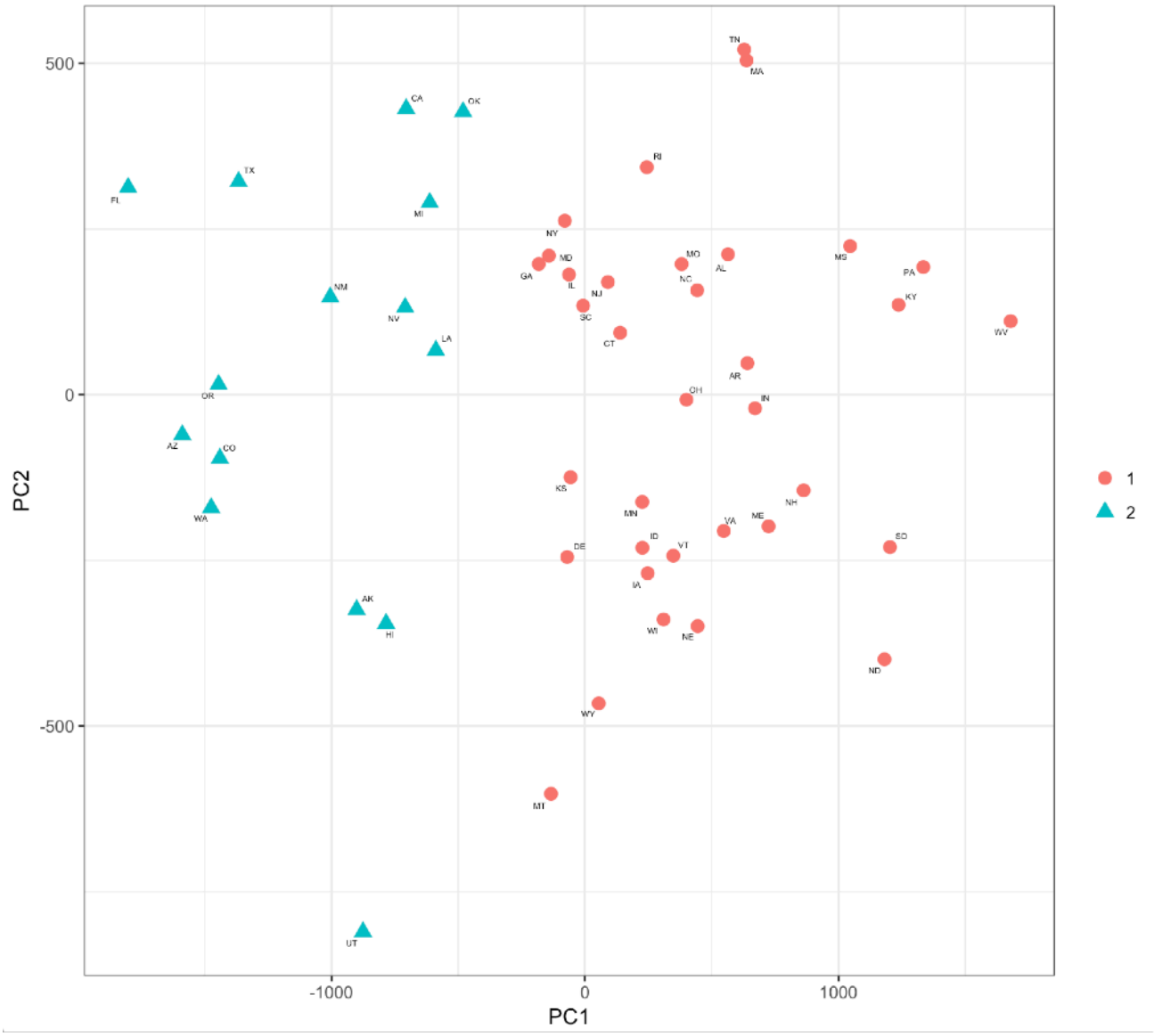
Attribute	cluster_0	cluster_1
Murder	5.757	9.153
Rape	27.686	48.247
Robbery	119.943	193.800
Assault	235.429	369.533
Burglary	991.914	1674.400
Theft	2522.457	3841.533
Vehicle	329.943	504.200

And in R, before standardizing.

	Murder	Rape	Robbery	Assault	Burglary	Theft	Vehicle
1	9.153333	48.24667	193.8000	369.5333	1674.4000	3841.533	504.2000
2	5.757143	27.68571	119.9429	235.4286	991.9143	2522.457	329.9429

As you can see, we get the same result. But, it can be shown that the cluster groupings get it wrong when it comes to individual state cluster groups, such as Utah and DC being on the same group. Intuition, and the data, says that Utah is a much safer place based on violent crime, than DC. This calls to the importance of applying the Kmeans technique when the variance of the variables cannot be assumed to be equal.

By way of illustration, I display the PC plot like about but without standardizing the variance of each variable. (NOTE: DC is excluded from the analysis since it is considered an outlier).



As we can see, UT now belongs to the cluster of more violent crime. If our goal was to try to group more violent crime states vs less violent crime states than we would have failed.