Chapter 6

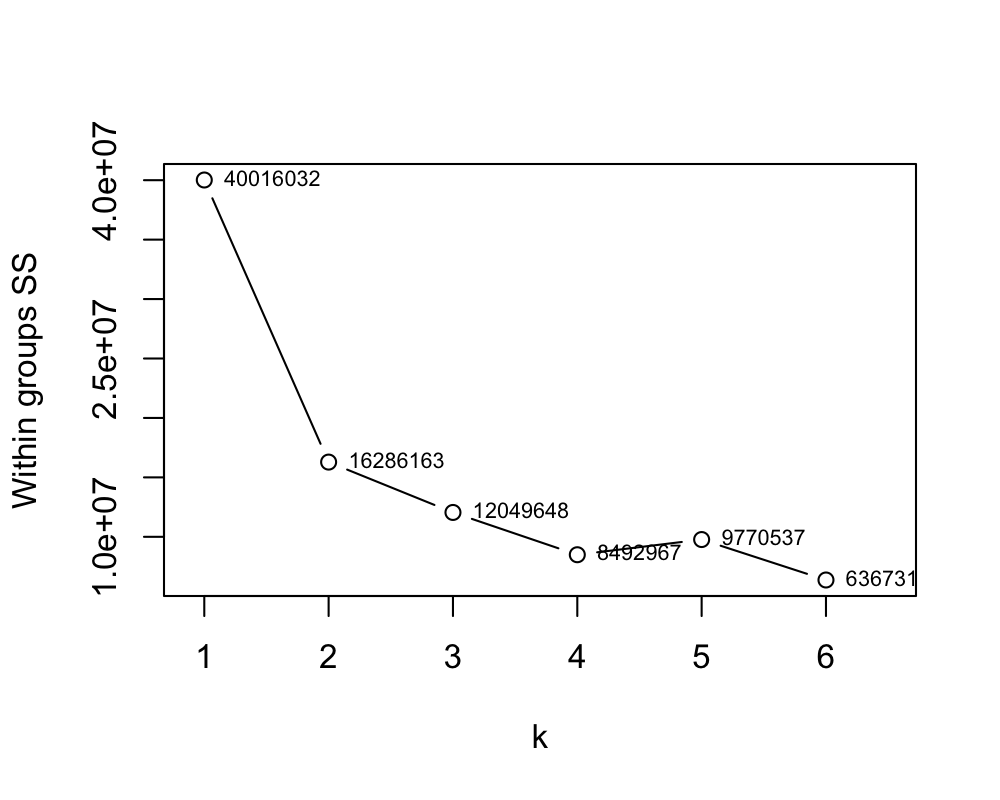
Cody Frisby

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

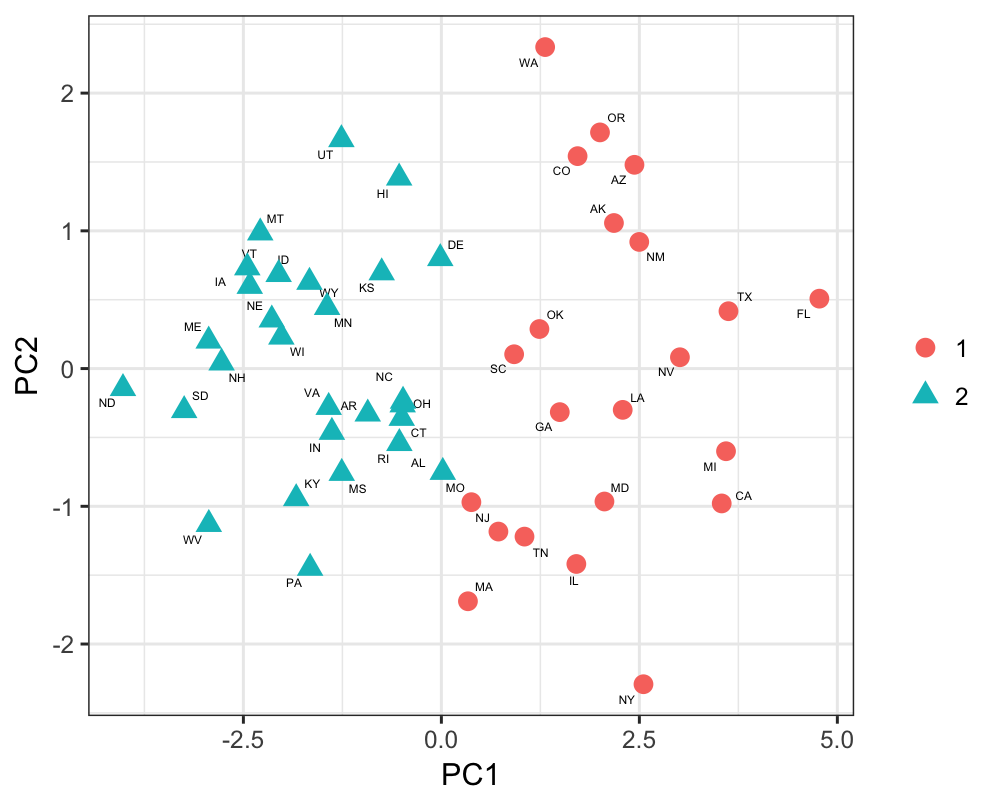


If we fit the model with we get the following centroids.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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