Hierarchical Clustering using Average Linkage

Fang Zhou

November 22, 2018

load data and omit the

```
df=USArrests
```

View the first six lines of the dataframe

```
head(df)

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

Make a summary of the data

```
summary(df)

##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

Standardize different variables and view the new data

```
df <- scale(df)
head(df)

##                Murder   Assault   UrbanPop         Rape
## Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473
## Alaska     0.50786248 1.1068225 -1.2117642  2.484202941
## Arizona    0.07163341 1.4788032  0.9989801  1.042878388
## Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602
## California 0.27826823 1.2628144  1.7589234  2.067820292
## Colorado   0.02571456 0.3988593  0.8608085  1.864967207
```

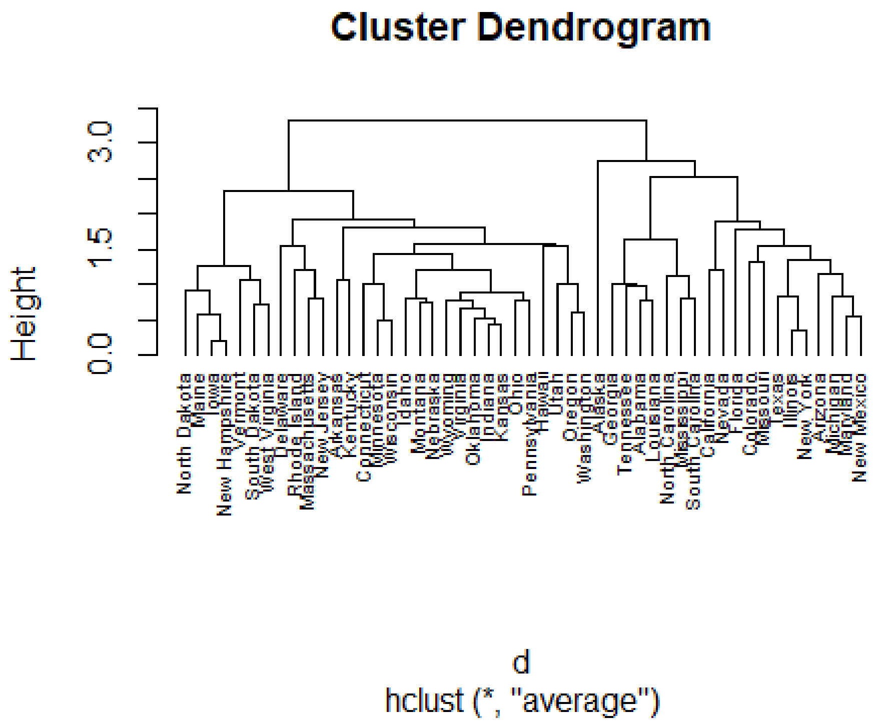
Great! Now we can come to the clustering!

In R, we use package ‘cluster’ to do agglomerative hierarchical clustering.

```
### compute the dissimilarity values
d <- dist(df, method = "euclidean")

### Hierarchical clustering using Average Linkage
library('cluster')
hc1 <- hclust(d, method = "average" )

# Plot the obtained dendrogram
plot(hc1, cex = 0.6, hang = -1)
```



The height of the cut to the dendrogram controls the number of clusters obtained. It plays the same role as the k in k-means clustering. Thus, we need to decide the value of k first.

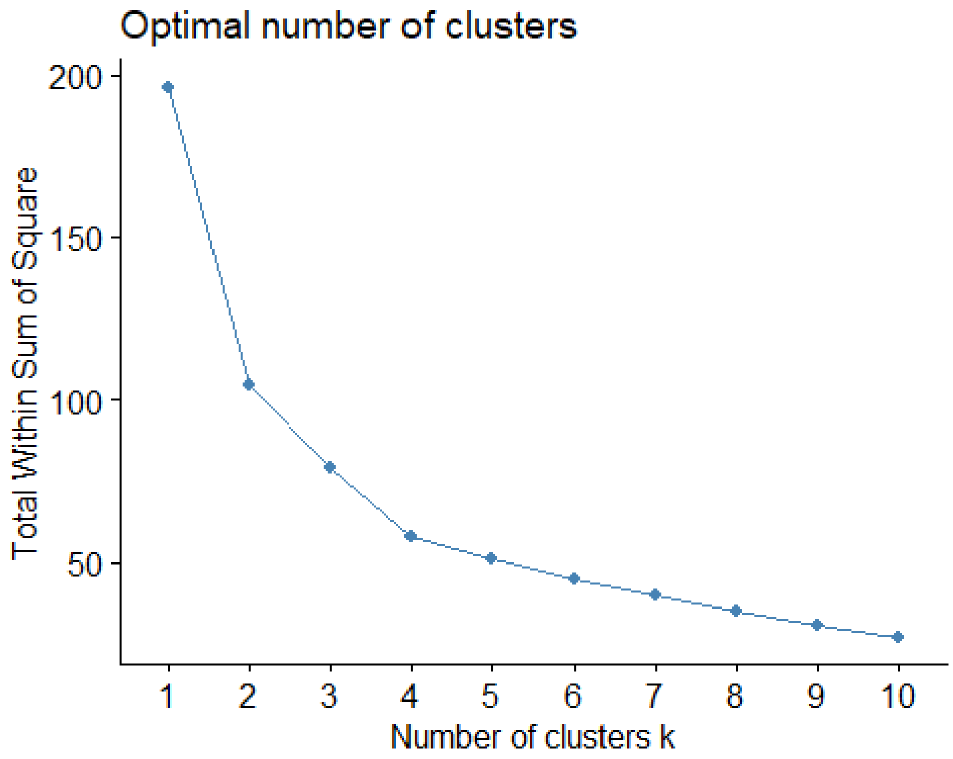
We use Elbow Method to determine the number of clusters obtained.

```
### use package 'factoextra' to do elbow method
library('factoextra')

## Loading required package: ggplot2

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at http
s://goo.gl/13EFCZ
```

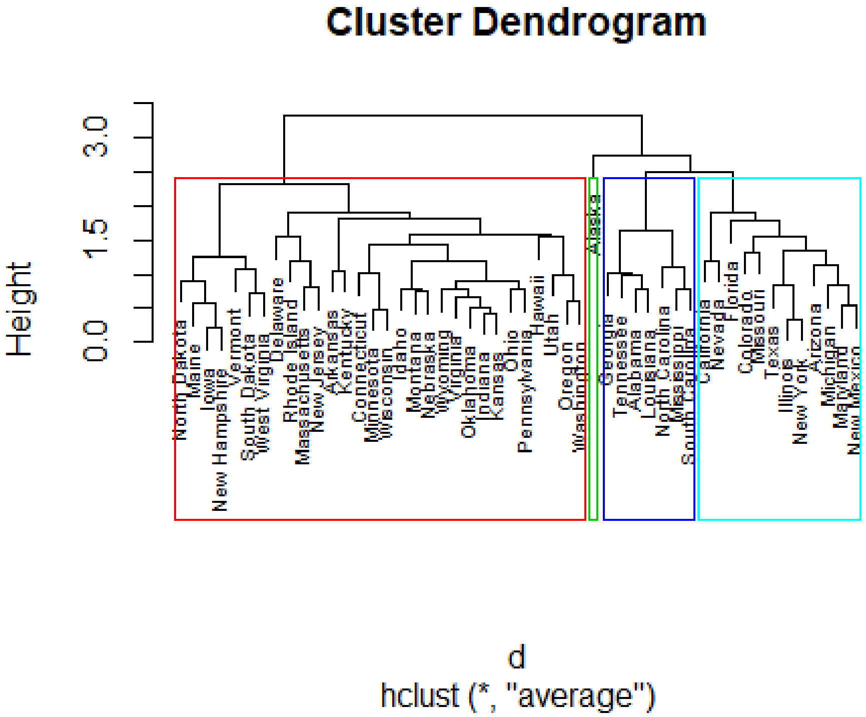
```r
###  plot within-cluster sum of squares(wss) against k
fviz_nbclust(df, FUN = hcut, method = "wss")
```



From the plot above we can see that if k < 4, the change of WSS is very fast; While k > 4, the change of WSS becomes slow. Thus, we can determine the number of clusters id 4.

Now we divide the states into 4 clusters based on the outcomes of agglomerative hierarchical clustering.

```r
plot(hc1, cex = 0.6)
rect.hclust(hc1, k = 4, border = 2:5)
```


Cluster Dendrogram
Height
0.0
1.5
3.0
North Dakota
Maine
Iowa
New Hampshire
Vermont
South Dakota
West Virginia
Delaware
Rhode Island
Massachusetts
New Jersey
Arkansas
Kentucky
Connecticut
Minnesota
Wisconsin
Idaho
Montana
Nebraska
Wyoming
Virginia
Oklahoma
Indiana
Kansas
Ohio
Pennsylvania
Hawaii
Utah
Oregon
Washington
Alaska
Georgia
Tennessee
Alabama
Louisiana
North Carolina
Mississippi
South Carolina
California
Nevada
Florida
Colorado
Missouri
Texas
Illinois
New York
Arizona
Michigan
Maryland
New Mexico
d
hclust (*, "average")