SIMILARITY, NEIGHBORS AND CLUSTER

Jose A. Mawyin
DATA 607

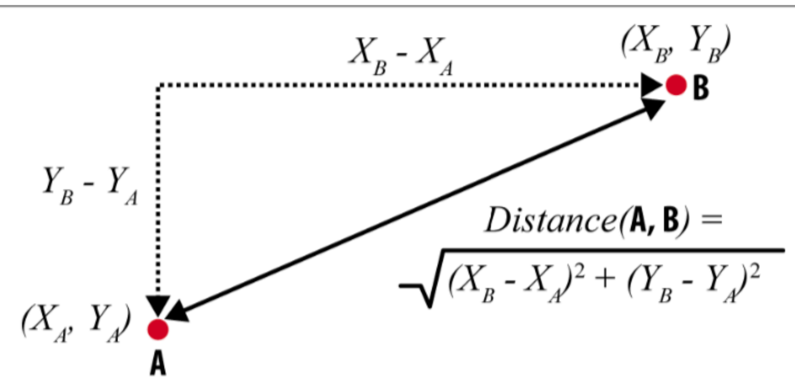
WHAT IS IT?

- **Similarity & Object Distance:**
  - Find similarity between objects (vectors) by quantifying the distance of vector attributes.
- **Neighbors:**
  - Finding similar objects based on their distance (*nearest neighbors*).
- **Cluster:**
  - Groups of objects with similar characteristics.

HOW TO FIND THE DISTANCE BETWEEN OBJECTS?

Examples of Euclidean distance

Case	Attributes				Decision
	Length	Height	Width	Weight	Quality
1	4.7	1.8	1.7	1.7	high
2	4.5	1.4	1.8	0.9	high
3	4.7	1.8	1.9	1.3	high
4	4.5	1.8	1.7	1.3	medium
5	4.3	1.6	1.9	1.7	medium
6	4.3	1.4	1.7	0.9	low
7	4.5	1.6	1.9	0.9	very-low
8	4.5	1.4	1.8	1.3	very-low



$$\sqrt{(d_{1,A} - d_{1,B})^2 + (d_{2,A} - d_{2,B})^2 + \ \dots + (d_{n,A} - d_{n,B})^2}$$

NOT THE ONLY WAY TO FIND OBJECT DISTANCE….

$$d_{\text{Euclidean}}(\mathbf{X}, \mathbf{Y}) = \| \mathbf{X} - \mathbf{Y} \|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots}$$

$$d_{\text{Manhattan}}(\mathbf{X}, \mathbf{Y}) = \| \mathbf{X} - \mathbf{Y} \|_1 = | x_1 - y_1 | + | x_2 - y_2 | + \cdots$$

$$d_{\text{Jaccard}}(X, Y) = 1 - \frac{| X \cap Y |}{| X \cup Y |}$$

$$d_{cosine}(\mathbf{X}, \mathbf{Y}) = 1 - \frac{\mathbf{X} \cdot \mathbf{Y}}{\| \mathbf{X} \|_2 \cdot \| \mathbf{Y} \|_2}$$

CENTROID CLUSTERING: K-MEANS ALGORITHM

The standard algorithm is the **Hartigan-Wong** algorithm, which defines the total within-cluster variation as the sum of squared distances Euclidean distances between items (**x**) and the corresponding centroid (**μ**: mean value of points in cluster):

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

The *total within-cluster sum of square* measures the "**compactness**" of the clustering and we want it to be as small as possible.

$$tot.withiness = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

CLUSTER ANALYSIS OF USARRESTS DATASET

```r
library(tidyverse)  # data manipulation
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
```

```r
df <- USArrests
df <- na.omit(df)
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

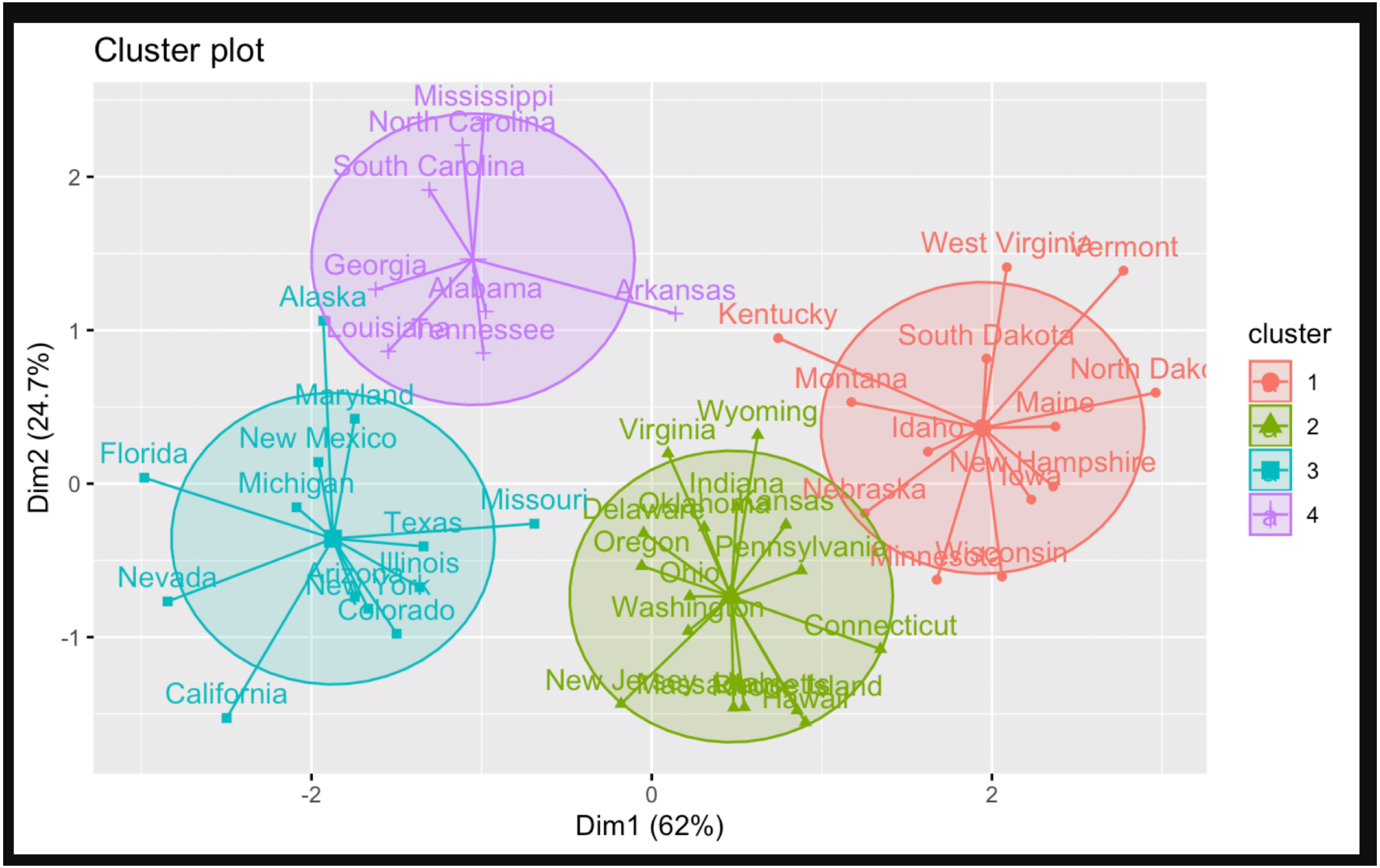
```r
df <- scale(df)
head(df)
```

```
              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
```

```r
k4 <- kmeans(df, centers = 4, nstart = 25)
k4
```

```
K-means clustering with 4 clusters of sizes 10, 14, 16, 10

Cluster means:
     Murder  Assault UrbanPop     Rape
1  2.950000  62.7000 53.90000 11.51000
2  8.214286 173.2857 70.64286 22.84286
3 11.812500 272.5625 68.31250 28.37500
4  5.590000 112.4000 65.60000 17.27000

Clustering vector:
       Alabama         Alaska        Arizona       Arkansas     California       Colorado    Connecticut 
             3              3              3              2              3              2              4 
      Delaware        Florida        Georgia         Hawaii          Idaho       Illinois        Indiana 
             3              3              2              1              4              3              4 
          Iowa         Kansas       Kentucky      Louisiana          Maine       Maryland  Massachusetts 
             1              4              4              3              1              3              2 
      Michigan      Minnesota    Mississippi       Missouri        Montana       Nebraska         Nevada 
             3              1              3              2              4              4              3 
 New Hampshire     New Jersey     New Mexico       New York North Carolina   North Dakota           Ohio 
             1              2              3              3              3              1              4 
      Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina   South Dakota      Tennessee 
             2              2              4              2              3              1              2 
         Texas           Utah        Vermont       Virginia     Washington  West Virginia      Wisconsin 
             2              4              1              2              2              1              1 
       Wyoming 
             2 

Within cluster sum of squares by cluster:
[1]  4547.914  9136.643 19563.863  1480.210
 (between_SS / total_SS =  90.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"        
[8] "iter"         "ifault"      
```

```
fviz_cluster(k4, data = df, show.clust.cent = TRUE, star.plot = TRUE, ellipse.type = "euclid")
```

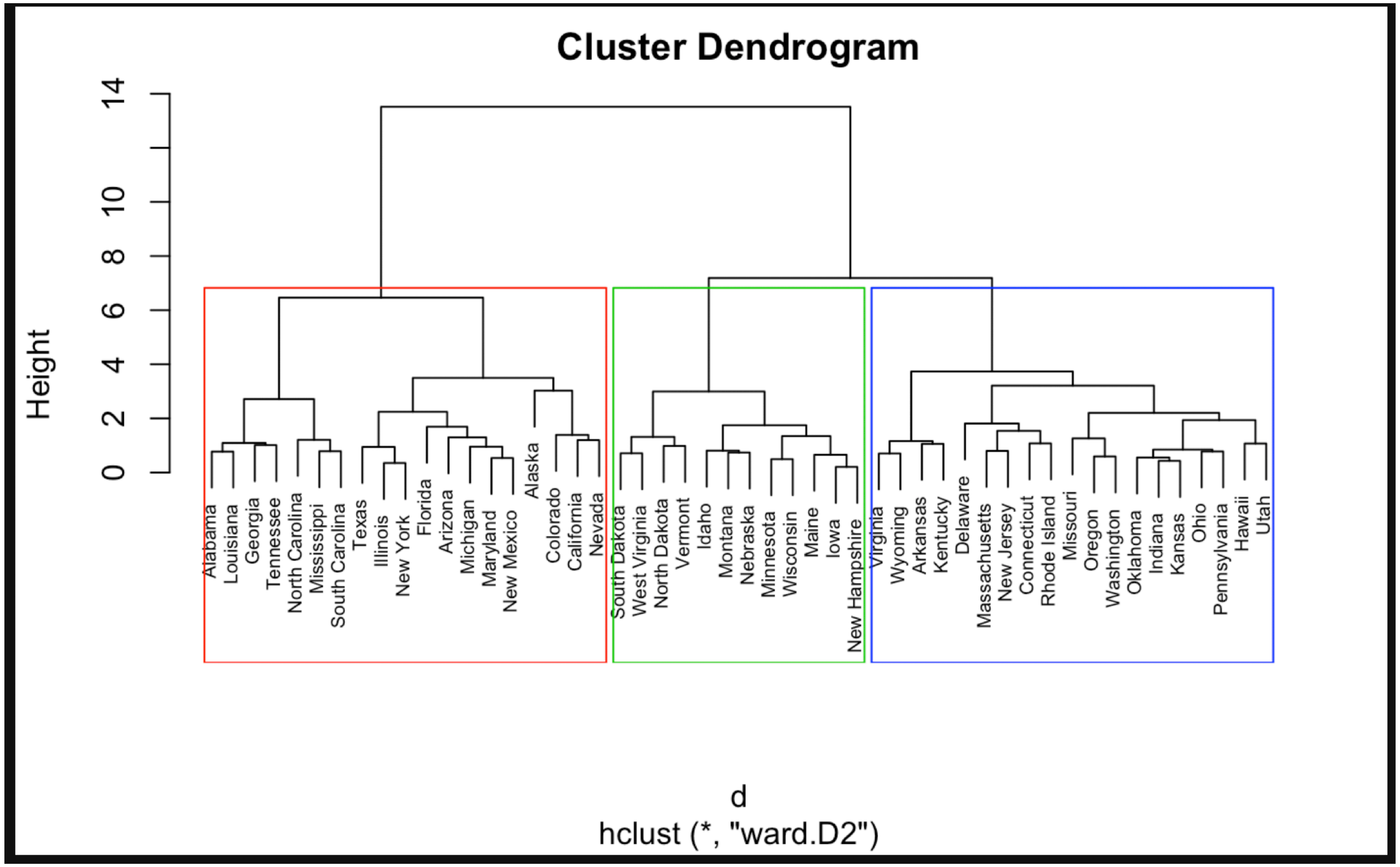


DIFFERENT K-MEANS ALGORITHMS

Algorithm	Advantages	Disadvantages
Lloyd	- For large data sets - Discrete data distribution - Optimize total sum of squares	- Slower convergence - Possible to create empty clusters
Forgy's	- For large data sets - Continuous data distribution - Optimize total sum of squares	- Slower convergence - Possible to create empty clusters
McQueen	- Fast initial convergence - Optimize total sum of squares	- Need to store the two nearest-cluster computations for each case - Sensitive to the order the algorithm is applied to the cases
Hartigan	- Fast initial convergence - Optimize within-cluster sum of squares	- Need to store the two nearest-cluster computations for each case - Sensitive to the order the algorithm is applied to the cases

HIERARCHICAL CLUSTERING

```
# Dissimilarity matrix
d <- dist(df, method = "euclidean")
# Hierarchical Clustering using Ward's method
Dendro <- hclust(d, method = "ward.D2" )
#Plot Dendrogram
plot(Dendro, cex = 0.6)
#Plot resctangles around "k"" groups
rect.hclust(Dendro, k = 3, border = 2:5)
```



WHAT HAVE WE LEARNED?

- *There are different measures of "distance"*
- *The appropriate metric will depend on what we are studying*
- *Experience, qualitative/quantitive and visualization will guide us to the best distance measure.*