ΕΦΑΡΜΟΓΗ ΑΝΑΛΥΣΗΣ ΣΥΣΤΑΔΩΝ ΜΕ ΧΡΗΣΗ ΤΗΣ R

```
# Clustering wines
# https://rpubs.com/gabrielmartos/ClusterAnalysis
> install.packages("rattle.data")
> library(rattle.data)
> data(wine,package="rattle.data")
> head(wine)
```

	Type	Alcohol	Malic	Ash	Alcalinity	Magnesium	Phenols	Flavanoids	Nonflavanoids	Proanthocyanins	Color	Hue	Dilution	Proline
1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
3	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
4	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
5	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735
6	1	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450

```
> wine.stand<-scale(wine[-1]) #To standarize the variables
```

Προετοιμασία δεδομένων

Πριν από τη δημιουργία συστάδων, προτείνεται η απομάκρυνση (ή εκτίμηση) των ελλείπουσων τιμών καθώς και η τυποποίηση (rescale) των μεταβλητών (για ευκολότερη σύγκριση).

```
# Prepare Data
mydata <- na.omit(mydata) # listwise deletion of missing
mydata <- scale(mydata) # standardize variables
```

Περιγραφή Συνόλου Δεδομένων

```
> str(wine)
'data.frame':    178 obs. of  14 variables:
 $ Type            : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ Alcohol         : num  14.2 13.2 13.2 14.4 13.2 ...
 $ Malic           : num  1.71 1.78 2.36 1.95 2.59 1.76 1.87 2.15 1.64 1.35 ...
 $ Ash             : num  2.43 2.14 2.67 2.5 2.87 2.45 2.45 2.61 2.17 2.27 ...
 $ Alcalinity      : num  15.6 11.2 18.6 16.8 21 15.2 14.6 17.6 14 16 ...
 $ Magnesium       : int  127 100 101 113 118 112 96 121 97 98 ...
 $ Phenols         : num  2.8 2.65 2.8 3.85 2.8 3.27 2.5 2.6 2.8 2.98 ...
 $ Flavanoids      : num  3.06 2.76 3.24 3.49 2.69 3.39 2.52 2.51 2.98 3.15 ...
 $ Nonflavanoids   : num  0.28 0.26 0.3 0.24 0.39 0.34 0.3 0.31 0.29 0.22 ...
 $ Proanthocyanins: num  2.29 1.28 2.81 2.18 1.82 1.97 1.98 1.25 1.98 1.85 ...
 $ Color           : num  5.64 4.38 5.68 7.8 4.32 6.75 5.25 5.05 5.2 7.22 ...
 $ Hue             : num  1.04 1.05 1.03 0.86 1.04 1.05 1.02 1.06 1.08 1.01 ...
 $ Dilution        : num  3.92 3.4 3.17 3.45 2.93 2.85 3.58 3.58 2.85 3.55 ...
 $ Proline         : int  1065 1050 1185 1480 735 1450 1290 1295 1045 1045 ...
```

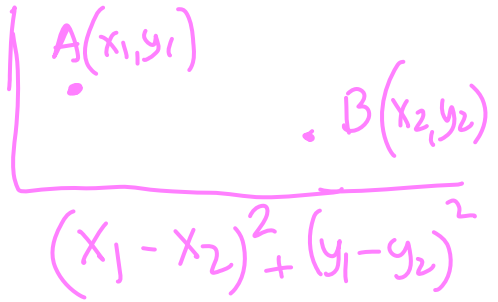
ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

- Οι Ιεραρχικές μέθοδοι ταξινόμησης χρησιμοποιούν τον πίνακα αποστάσεων (ομοιότητας) για την εφαρμογή του αλγορίθμου συσταδοποίησης. Η επιλογή του μέτρου της απόστασης επηρεάζει το σχήμα των συστάδων, αφού κάποιες παρατηρήσεις που θα είναι κοντά σύμφωνα με ένα μέτρο απόστασης, ενδέχεται να είναι μακριά σύμφωνα με κάποιο άλλο.

```
> d<-dist(wine.stand,method="euclidean")
```

- Χρησιμοποιούμε το dataset wine.stand και εφαρμόζουμε την Ευκλείδια απόσταση για να δημιουργήσουμε τον πίνακα αποστάσεων. Άλλες επιλογές που μπορούμε να χρησιμοποιήσουμε είναι οι "maximum", "manhattan", "canberra" (*weighted Manhattan distance*), "binary" και "minkowski".

"distance" = "Απόσταση"



ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

```
> H.fit <- hclust(d, method="ward.D2")
```

- Στη συνέχεια εφαρμόζουμε ιεραρχική ταξινόμηση (συνάρτηση hclust(…)) χρησιμοποιώντας τη μέθοδο του Ward.

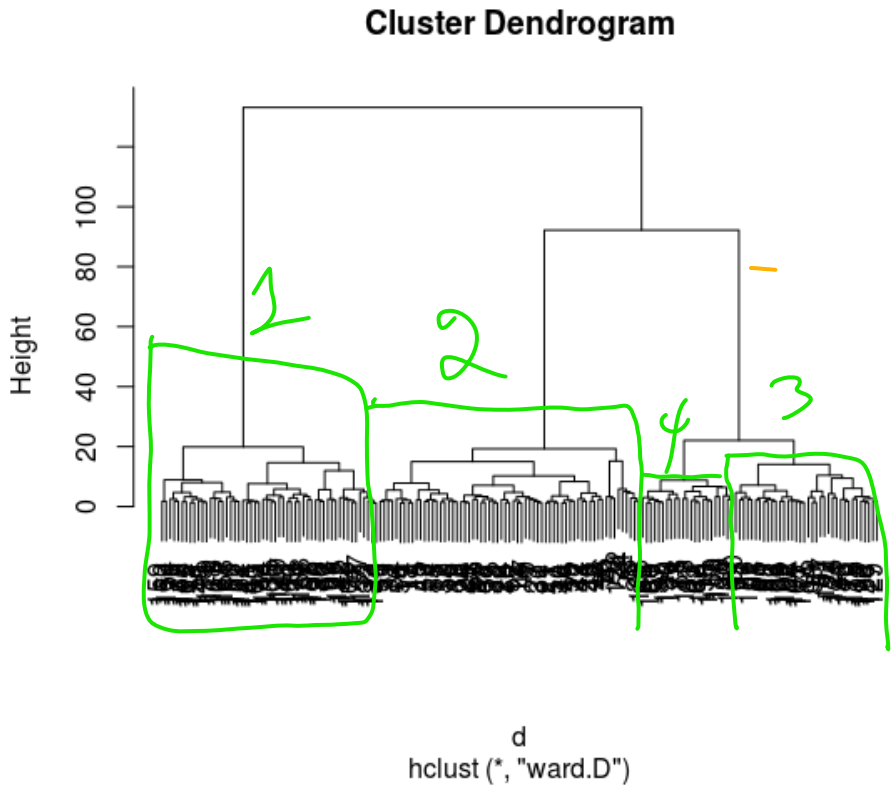
- Η μέθοδος του Ward οδηγεί σε συστάδες, τέτοιες ώστε να ελαχιστοποιείται η συνολική εντός-συστάδων διακύμανση (*total within-cluster variance*) δημιουργία συστάδων.

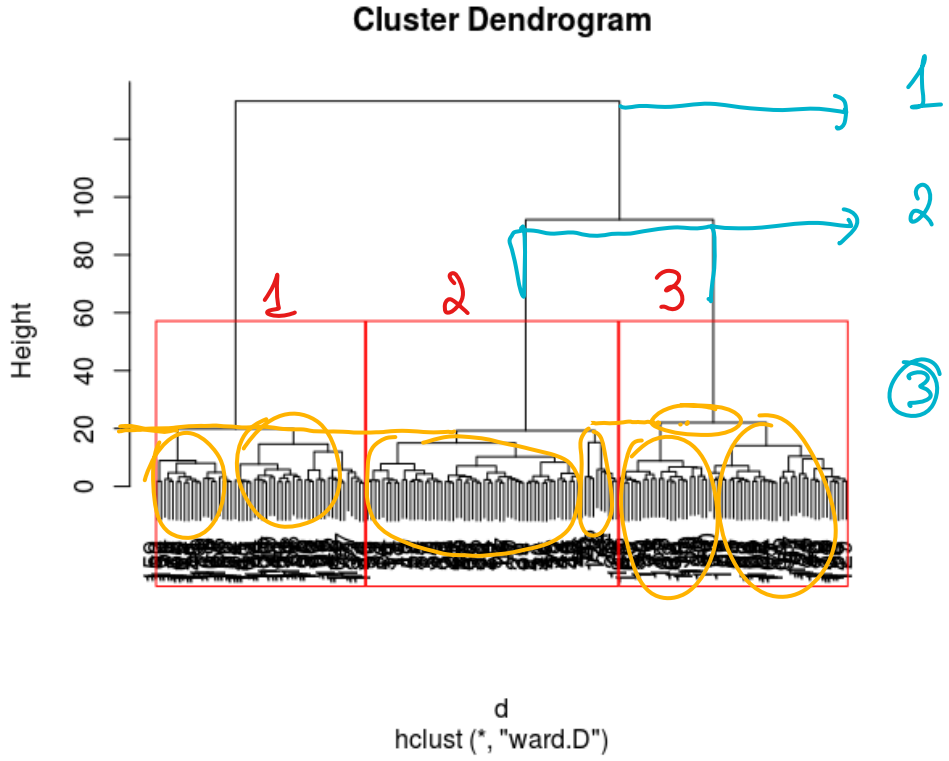
- Στην R, η αρχική μέθοδος που προτάθηκε από τον Ward (1963) υλοποιείται με το όρισμα ward.D2. Η επιλογή ward.D δε χρησιμοποιεί τα τετράγωνα των αποστάσεων για τη δημιουργία του νέου πίνακα σε κάθε βήμα του αλγορίθμου.

- Το αποτέλεσμα της Ιεραρχικής Ταξινόμησης απεικονίζεται σε ένα δενδρόγραμμα. Μπορούμε να ζητήσουμε να εμφανιστεί συγκεκριμένος αριθμός ομάδων σε αυτό ενώ μπούρουμε να βάλουμε και περιγράμματα ώστε να ξεχωρίζουν μεταξύ τους οι συστάδες

```
> plot(H.fit) # display dendogram
> groups <- cutree(H.fit, k=3)
# cut tree into 3 clusters,
# draw dendogram with red borders around the 3 clusters
> rect.hclust(H.fit, k=3, border="red")
```

ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ





ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

```
> H.fit <- hclust(d, method="ward.D2")
```

- Εκτός των ward.D και ward.D2, άλλες επιλογές για το method είναι οι "single", "complete", "average" (= UPGMA), "mcquitty" (= WPGMA), "median" (= WPGMC) και "centroid" (= UPGMC).
- Η μέθοδος του Ward βρίσκει συμπαγείς & σφαιρικές συστάδες
- Η μέθοδος complete linkage βρίσκει όμοιες συστάδες. Η μέθοδος single linkage βασίζεται στη φιλοσοφία '*friends of friends*' για τη δημιουργία συστάδων.
- Οι άλλες επιλογές είναι κάτι ενδιάμεσο μεταξύ των μεθόδων single και complete linkage.
- Σημειώνεται επίσης ότι οι μέθοδοι "median" και "centroid" δεν αποτελούν μονότονα μέτρα απόστασης και τα δενδρογράμμα που δίνουν είναι δύσκολο να ερμηνευτούν

rattle package - rattle data set about wines

https://cran.r-hub.io/web/packages/rattle.data/rattle.data.pdf#page7

more data for data science.

Murtagh, Fionn and Legendre, Pierre (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? Journal of Classification, 31, 274-295. doi:10.1007/s00357-014-9161-z.

- Ward criterion that squares the dissimilarities before clustering takes place.Original by Ward 1963.

- **package fcp** "flexible procedures for clustering. various co clustering methods & various methods εκτιμησης του αριθμου των clusters

Cristian Henning https://www.unibo.it/sitoweb/

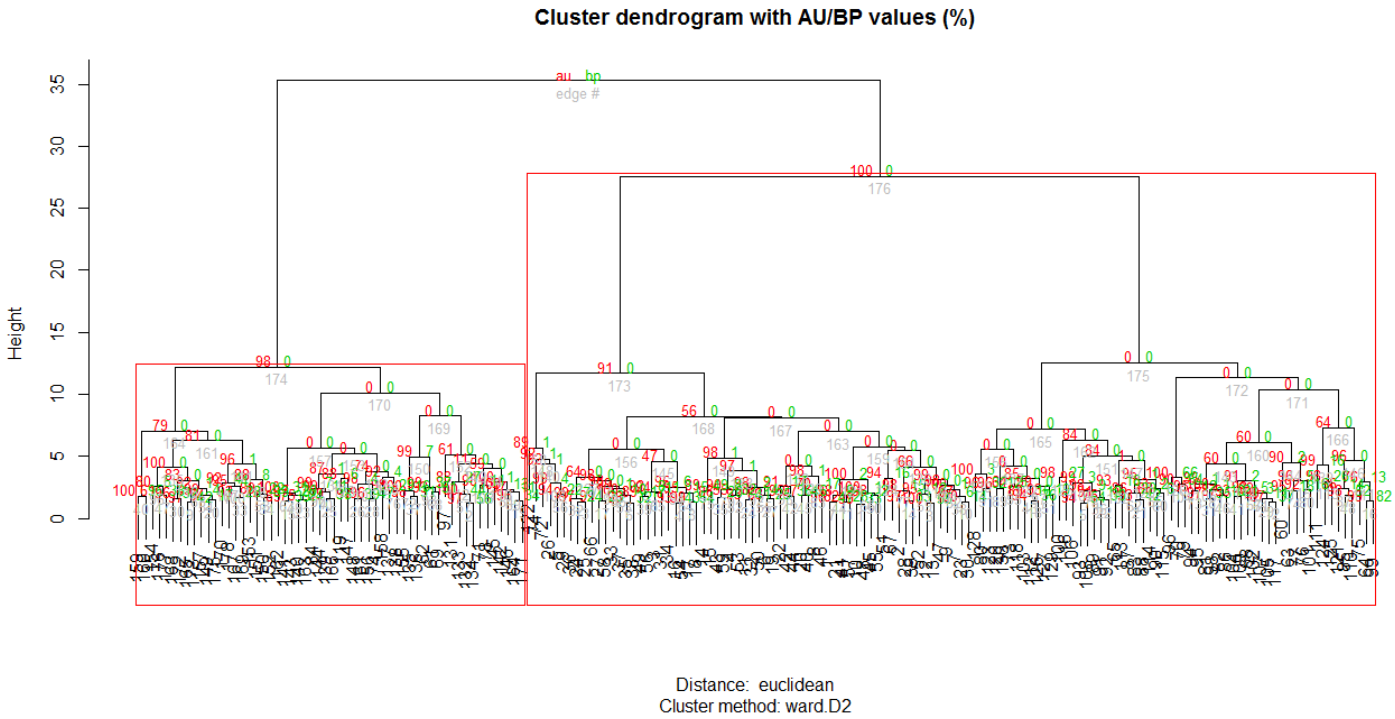
Ward Hierarchical Clustering with Bootstrapped p values

pv = p-values

Efron 1980

```
> library(pvclust)
> fit.pvclust <- pvclust(t(wine.stand),
method.hclust="ward",method.dist="euclidean")
plot(fit.pvclust) # dendogram with p values
# add rectangles around groups highly supported by the
data
pvrect(fit.pvclust, alpha=.95)
```

japanese 2019-2022



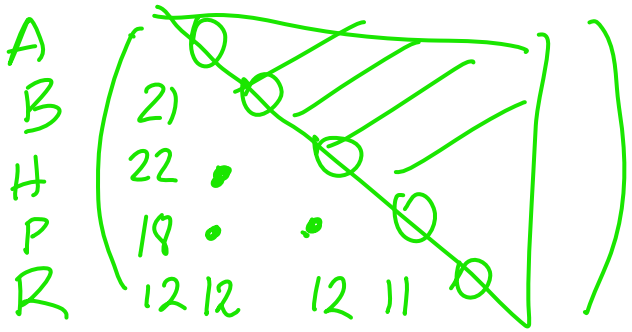
ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

Εφαρμογή (Καρλής 2005, 9.5.3)

```
> datacl1<-
cbind(c(24.7,12.5,11.6,14.3,13.6),c(5.7,11.9,13.4,10.2,10.7)
,c(30.8,14.4,14.8,16,26.9))
> #datacl1
> X<-as.data.frame(datacl1)
> rownames(X)<-
c("Albania","Bulgaria","Hungary","Poland","Romania")
> colnames(X)<-c("Births","Deaths","InfantD")
> X
         Births Deaths InfantD
Albania    24.7    5.7    30.8
Bulgaria   12.5   11.9    14.4
Hungary    11.6   13.4    14.8
Poland     14.3   10.2    16.0
Romania    13.6   10.7    26.9
```

ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

Εφαρμογή (Καρλής 2005, 9.5.3)




```
> #### similarity matrix, Euclidean Distance
> d<-dist(X,method="euclidean")
> d
             Albania  Bulgaria   Hungary    Poland
Bulgaria 21.359775
Hungary  22.065811  1.794436
Poland   18.640011  2.947881  4.355456
Romania  12.783583 12.605554 12.557866 10.933892

> fit<-hclust(d,method="single")
> plot(fit) # display dendrogram
> groups<-cutree(fit,k=2) # cut tree into 2 clusters
> groups
 Albania Bulgaria  Hungary   Poland  Romania
       1        2        2        2        2
> # draw dendrogram with red borders around the k=2 clusters
> rect.hclust(fit,k=2,border="red")
```

ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

Εφαρμογή (Καρλής 2005, 9.5.3)

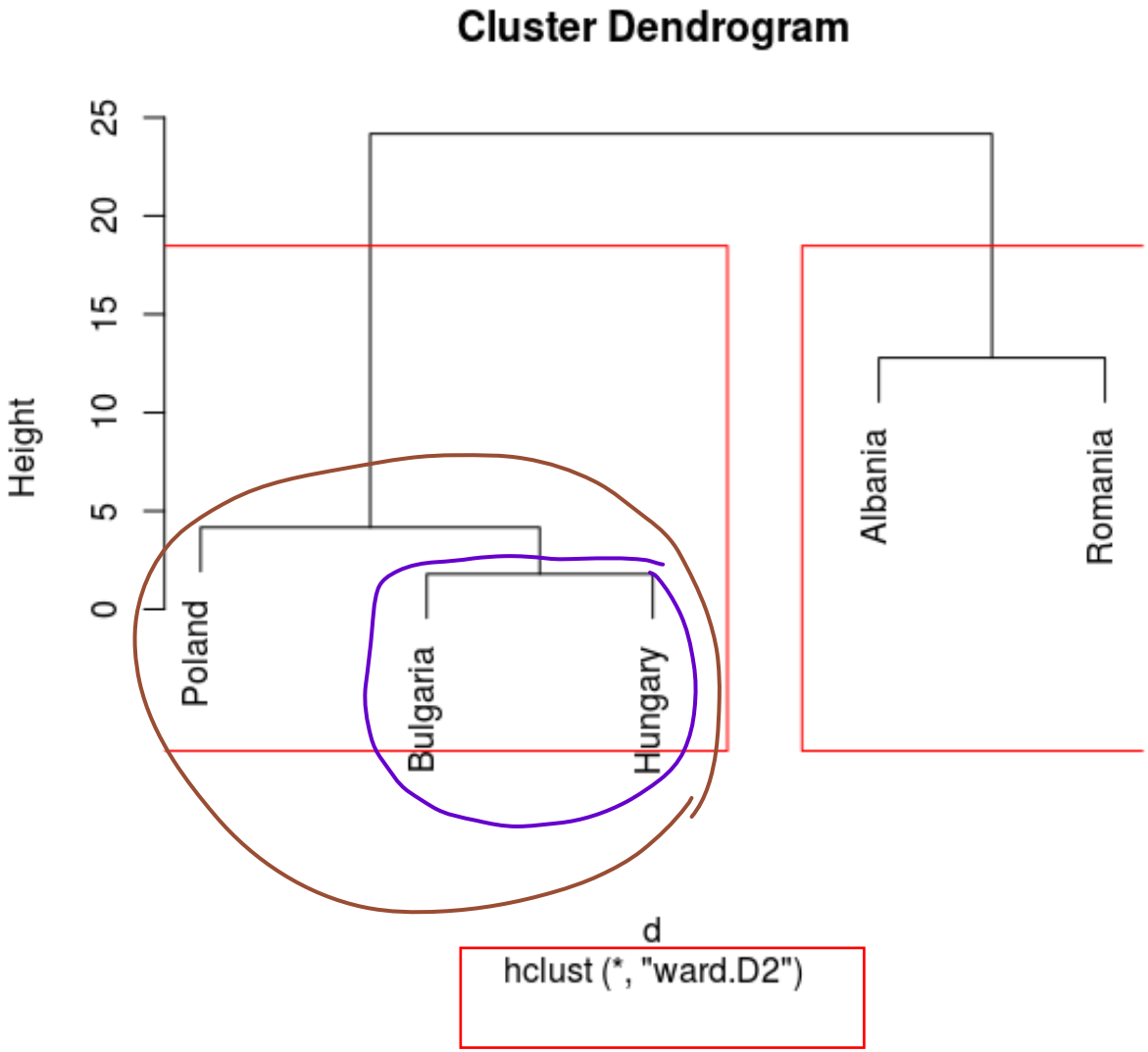
ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

Εφαρμογή (Καρλής 2005, 9.5.3)

```
> fit<-hclust(d,method="ward.D2")
> plot(fit) # display dendrogram
>  groups<-cutree(fit,k=2)  #  cut  tree  into  2
clusters
> groups
 Albania Bulgaria  Hungary   Poland  Romania
       1         2         2         2         1
# draw dendrogram with red borders around the k=2
clusters
> rect.hclust(fit,k=2,border="red")
```

ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ

Εφαρμογή (Καρλής 2005, 9.5.3)



ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΣΤΟΝ ΠΙΝΑΚΑ ΑΠΟΣΤΑΣΕΩΝ

- Έστω ότι έχουμε στη διάθεσή μας τον παρακάτω πίνακα αποστάσεων για 10 παρατηρήσεις

0	9	3	6	8	10	11
9	0	7	5	6	9	10
3	7	0	9	3	5	2
6	5	9	0	7	8	8
8	6	3	7	0	4	1
10	9	5	8	4	0	6
11	10	2	8	1	6	0

ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΣΤΟΝ ΠΙΝΑΚΑ ΑΠΟΣΤΑΣΕΩΝ

- Για να εφαρμόσουμε ιεραρχική ταξινόμηση στην R, θα πρέπει πρώτα να περάσουμε τον πίνακα και στη συνέχεια να τον δηλώσουμε ως πίνακα απόστασης d.

```
> dmat<-
matrix(c(0,9,3,6,8,10,11,9,0,7,5,6,9,10,3,7,0,9,3,5,2,
6,5,9,0,7,8,8,8,6,3,7,0,4,1,10,9,5,8,4,0,6,11,10,2,8,1
,6,0),ncol=7)
> dmat
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    0    9    3    6    8   10   11
[2,]    9    0    7    5    6    9   10
[3,]    3    7    0    9    3    5    2
[4,]    6    5    9    0    7    8    8
[5,]    8    6    3    7    0    4    1
[6,]   10    9    5    8    4    0    6
[7,]   11   10    2    8    1    6    0
```

ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΣΤΟΝ ΠΙΝΑΚΑ ΑΠΟΣΤΑΣΕΩΝ

- Στη συνέχεια εφαρμόζουμε τις παρακάτω εντολές

```
> d <-as.dist(dmat) # to as.dist xreiazetai na perasei o
dmat ws pinakas apostasewn
> fit<-hclust(d,method="ward.D2")
> plot(fit) # display dendrogram
> d<-as.dist(dmat) # to as.dist xreiazetai na perasei o dmat
ws pinakas apostasewn
> fit<-hclust(d,method="ward.D2")
> plot(fit) # display dendrogram
> groups<-cutree(fit,k=3) # cut tree into 3 clusters
> groups
[1] 1 2 3 2 3 3 3
> # draw dendrogram with red borders around the k=3 clusters
> rect.hclust(fit,k=3,border="blue")
```

ΙΕΡΑΡΧΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΣΤΟΝ ΠΙΝΑΚΑ ΑΠΟΣΤΑΣΕΩΝ