Assignment 1

Name Student no.

Q1

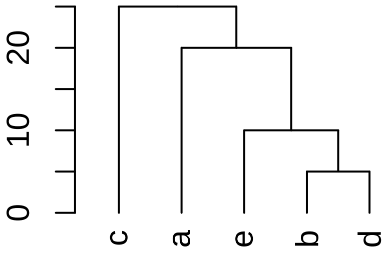
```
set.seed(123)
x <- matrix(sample(-5:5, 10), nrow=5)
rownames(x)<- letters[1:5]
colnames(x)<- c("U", "V")
x
```

```
##    U  V
## a -2 -5
## b  2 -3
## c  5  3
## d  4 -4
## e  1  0
```

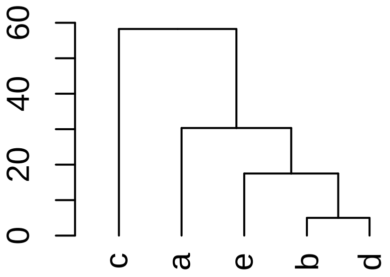
```
dx <- dist(x)^2
dx
```

```
##     a   b   c   d
## b  20            
## c 113  45        
## d  37   5  50    
## e  34  10  25  25
```

```
hs <- hclust(dx, "single")
ha <- hclust(dx, "average")
plot(as.dendrogram(hs))
```



```
plot(as.dendrogram(ha))
```

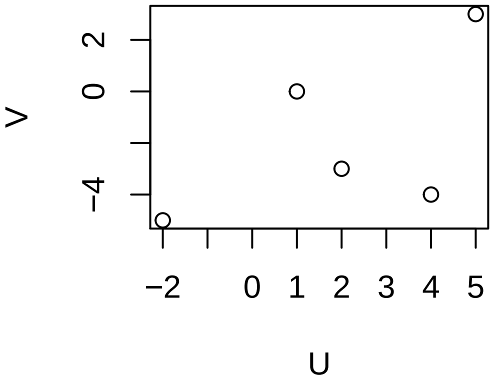


```
c1 <- apply(x[1:3,], 2, mean)
c2 <- apply(x[4:5,], 2, mean)
```

```
kmeans(x, centers= rbind(c1, c2), algorithm ="Lloyd")
```

```
## K-means clustering with 2 clusters of sizes 2, 3
##
## Cluster means:
##           U         V
## 1 -0.500000 -2.500000
## 2  3.666667 -1.333333
##
## Clustering vector:
## a b c d e
## 1 2 2 2 1
##
## Within cluster sum of squares by cluster:
## [1] 17.00000 33.33333
##  (between_SS / total_SS =  30.9 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

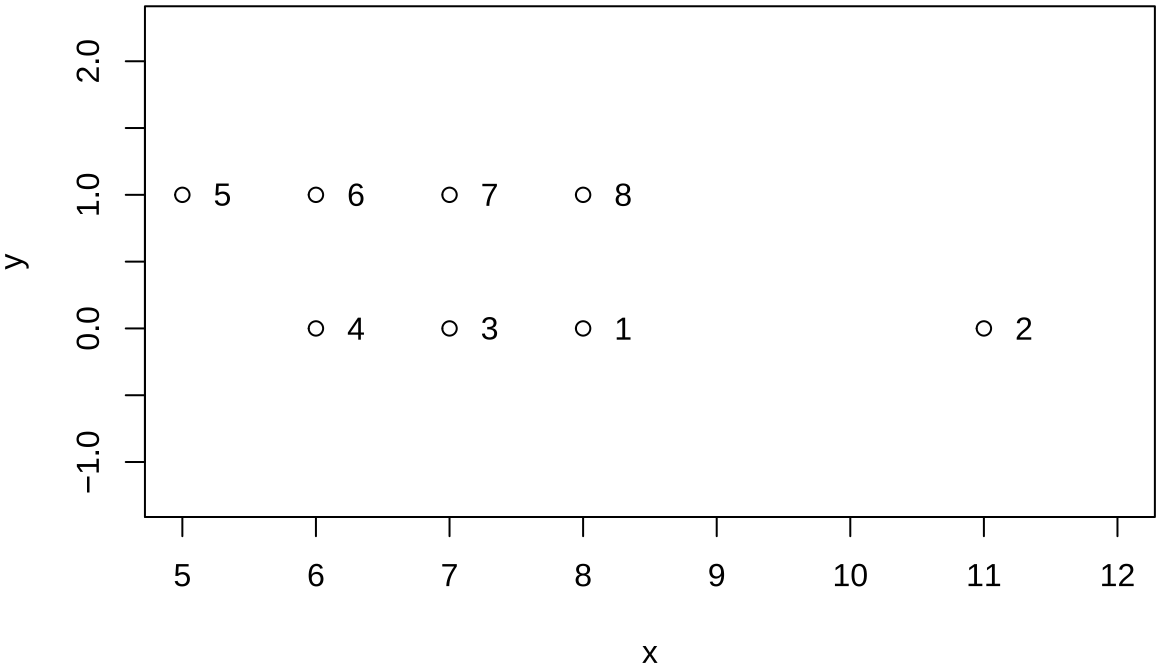
```
plot(x)
```



Q2

a)

```
x <- c(8,11,7,6,5,6,7,8)
y <- c(0,0,0,0,1,1,1,1)
plot(x,y, xlim = c(5, 12), asp=1)
text(x+.3,y, 1:8 )
```



```
d <- data.frame(x=x,y=y)
kmeans(d,2)$cluster
```

```
## [1] 1 1 2 2 2 2 2 1
```

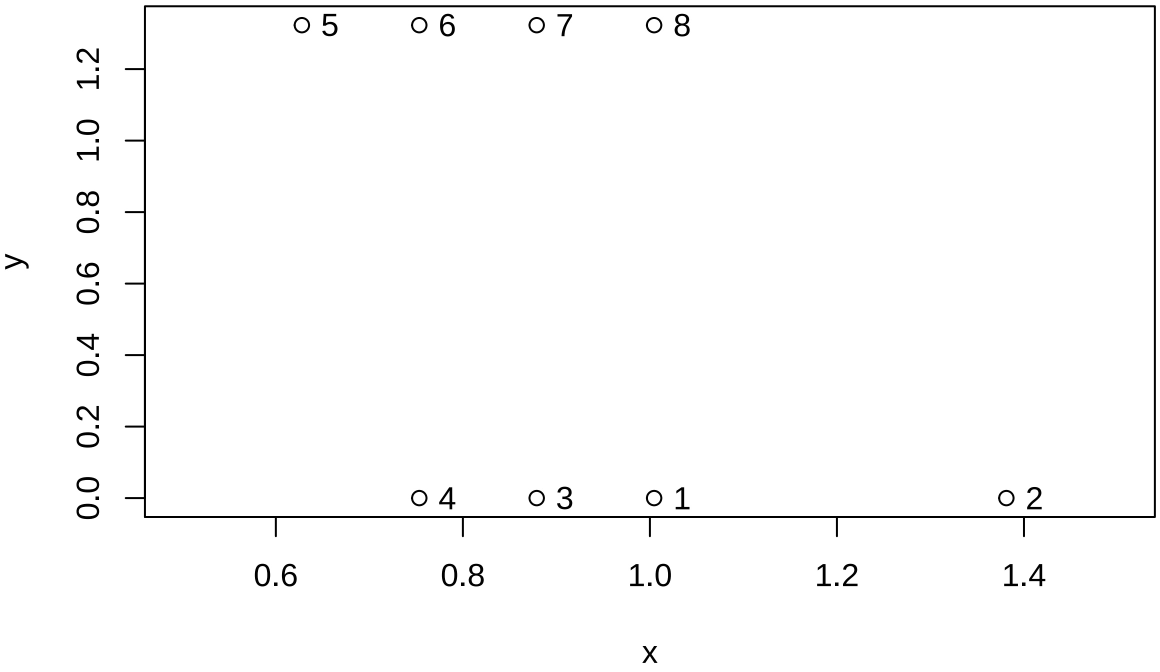
```
kmeans(d,2, nstart=10)$cluster
```

```
## [1] 2 1 2 2 2 2 2 2
```

NOTE: use nstart to run the algorithm from 10 random starts. Better convergence. Point 2 is in a cluster of its own

b)

```
d1 <- scale(d, center=F)
plot(d1,  xlim = c(0.5,1.5))
text(d1[,1]+0.03,d1[,2], 1:8)
```



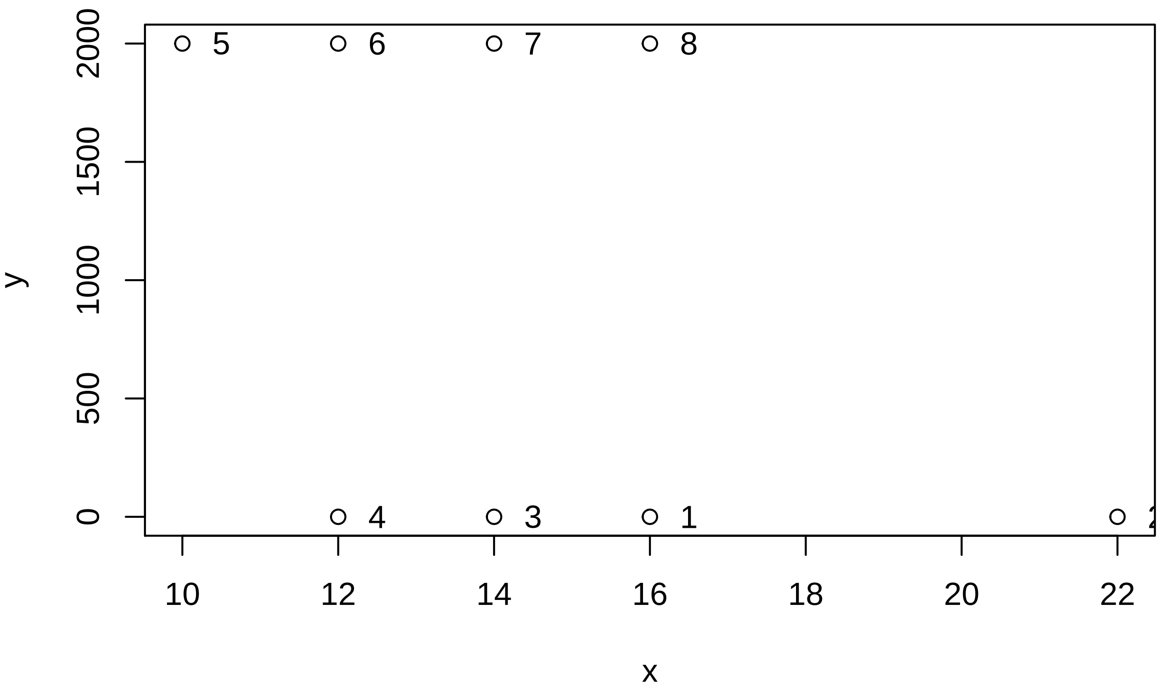
```
kmeans(d1,2, nstart=10)
```

```
## K-means clustering with 2 clusters of sizes 4, 4
##
## Cluster means:
##           x        y
## 1 0.8161517 1.322876
## 2 1.0044944 0.000000
##
## Clustering vector:
## [1] 2 2 2 2 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 0.07882883 0.22072072
##  (between_SS / total_SS =  92.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Points 1-4 in 1 cluster, 5-8 in the other

c)

```
d <- data.frame(x=2*x,y=2000*y)
plot(d)
text(d[,1]+ 0.5,d[,2], 1:8)
```



```
kmeans(d,2, nstart=10)
```

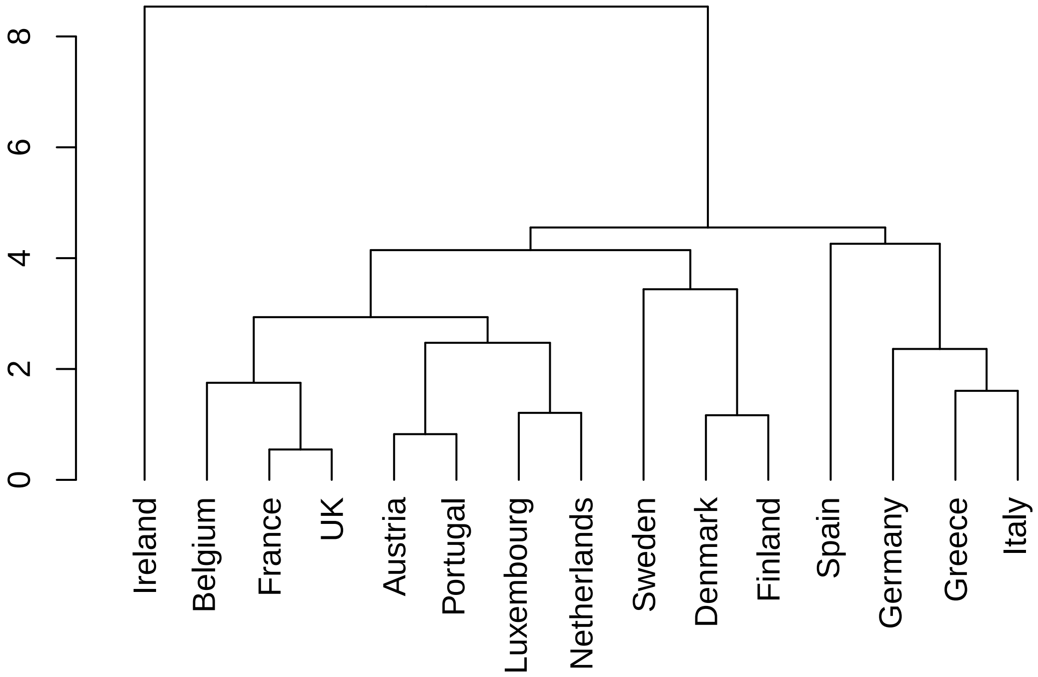
```
## K-means clustering with 2 clusters of sizes 4, 4
##
## Cluster means:
##    x    y
## 1 13 2000
## 2 16    0
##
## Clustering vector:
## [1] 2 2 2 2 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 20 56
##  (between_SS / total_SS = 100.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Points 1-4 in 1 cluster, 5-8 in the other

Q3

a)

```
eupop <- read.table("data/eupop.txt", header=T, row.names=1)
eupop <- eupop[,-5]
d <- dist(eupop)
h<- hclust(d, "average")
plot(as.dendrogram(h))
```



Ireland is outlier.

b)

```
source('code/h1code.R')
sumPartition(eupop, cutree(h,3))
```

```
## Final Partition
##
## Number of clusters  3
##
##           N.obs Within.clus.SS Ave.dist..Centroid Max.dist.centroid
## Cluster 1    10         55.305           2.211730          4.030199
## Cluster 2     4         17.535           1.831295          3.117090
## Cluster 3     1          0.000           0.000000          0.000000
##
##
## Cluster centroids
##
##      Cluster 1 Cluster 2 Cluster 3 Grand centrd
## p014 18.23     15.25     22.2      17.7
## p1544 42.52     43.775    46.2      43.1
## p4564 23.94     24.25     20.3      23.78
## p65.  15.36     16.725    11.3      15.45333
##
##
## Distances between Cluster centroids
##
##           Cluster 1 Cluster 2 Cluster 3
## Cluster 1  0.000000  3.523457  7.683521
## Cluster 2  3.523457  0.000000  9.960735
## Cluster 3  7.683521  9.960735  0.000000
```

Ireland is in Cluster 3. Germany Greece, Italy, Spain are in Cluster 2. Everyone else is in Cluster 1. Cluster 3: highest proportion of children(under 15), lowest percentage of over 65. Cluster 2: below average for under 15s and above average proportion of over 65s. Cluster 2 and 3 are furtherest apart, cluster 1 and 2 are closest. Cluster 2 is more compact than cluster 1.

c)

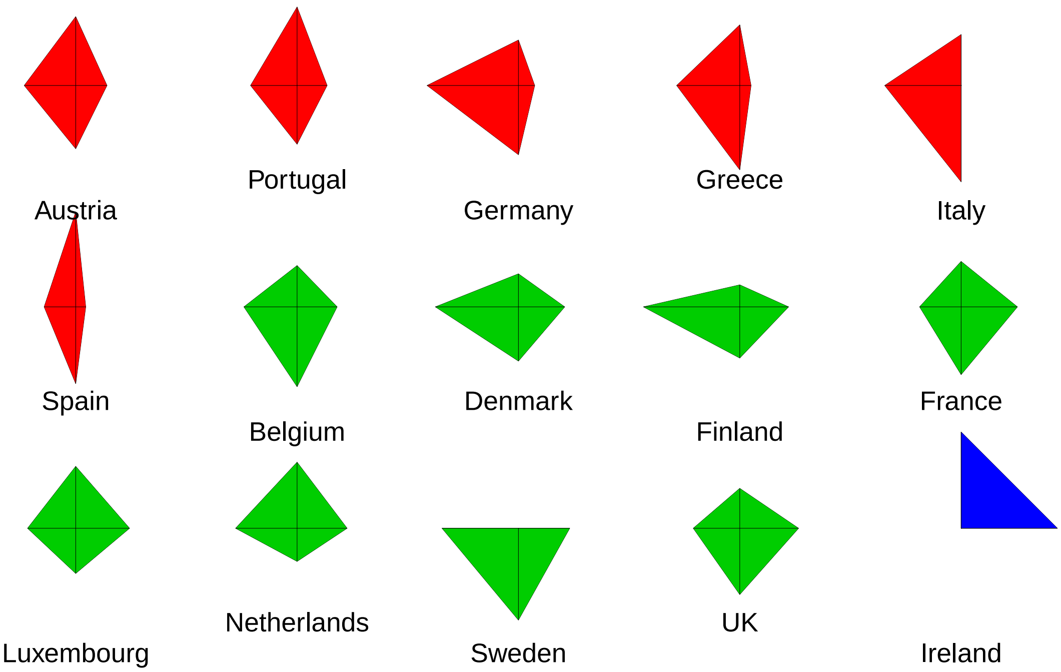
```r
km <- kmeans(eupop, 3,nstart=10)
km
```

```
## K-means clustering with 3 clusters of sizes 6, 8, 1
##
## Cluster means:
##       p014    p1544    p4564     p65.
## 1 15.81667 44.03333 23.91667 16.26667
## 2 18.55000 42.01250 24.11250 15.36250
## 3 22.20000 46.20000 20.30000 11.30000
##
## Clustering vector:
##     Austria     Belgium     Denmark     Finland      France  Luxembourg
##           1           2           2           2           2           2
## Netherlands    Portugal      Sweden          UK     Germany      Greece
##           2           1           2           2           1           1
##       Italy       Spain     Ireland
##           1           1           3
##
## Within cluster sum of squares by cluster:
## [1] 26.38333 39.37625  0.00000
##  (between_SS / total_SS =  61.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

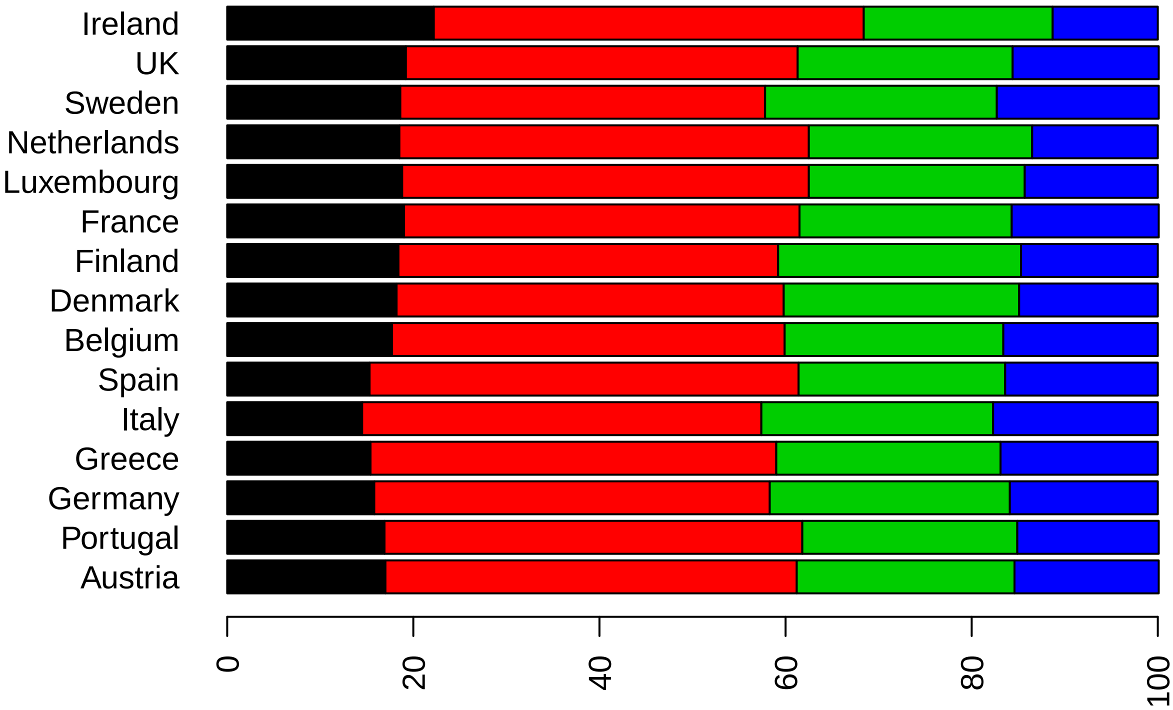
Cluster agreement with 3 cluster solution of hclust, except, Portugal and Austria are clustered with Greece, Italy, Germany and Spain. This cluster still has lower proportion of children and above average porportion of over 65s.

d)

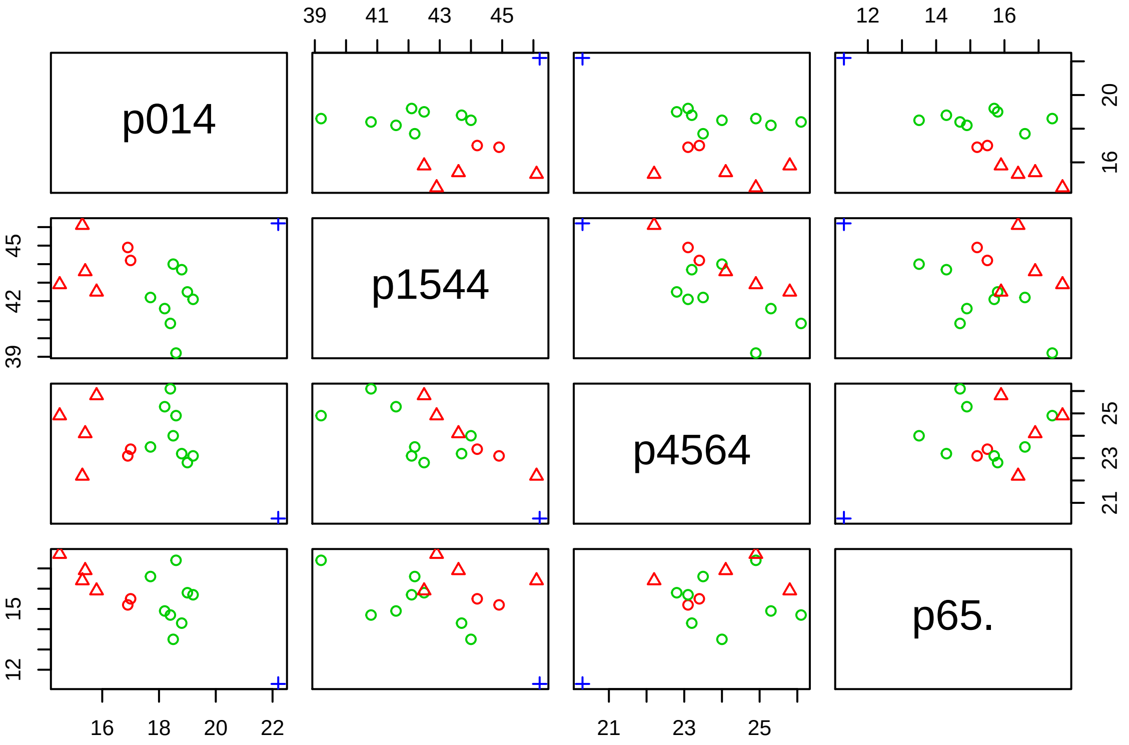
```
clusk <- km$cluster
o <- order(clusk)
stars(eupop[o,],nrow=3, col.stars=clusk[o]+1)
```



```
par(mar=c(3,6,3,2))
barplot(t(as.matrix(eupop[o,])), col=1:4, horiz=T, las=2)
```



```r
# another display, kmeans is colours, hclust is symbols
pairs(eupop, col = clusk+1, pch = cutree(h,3))
```

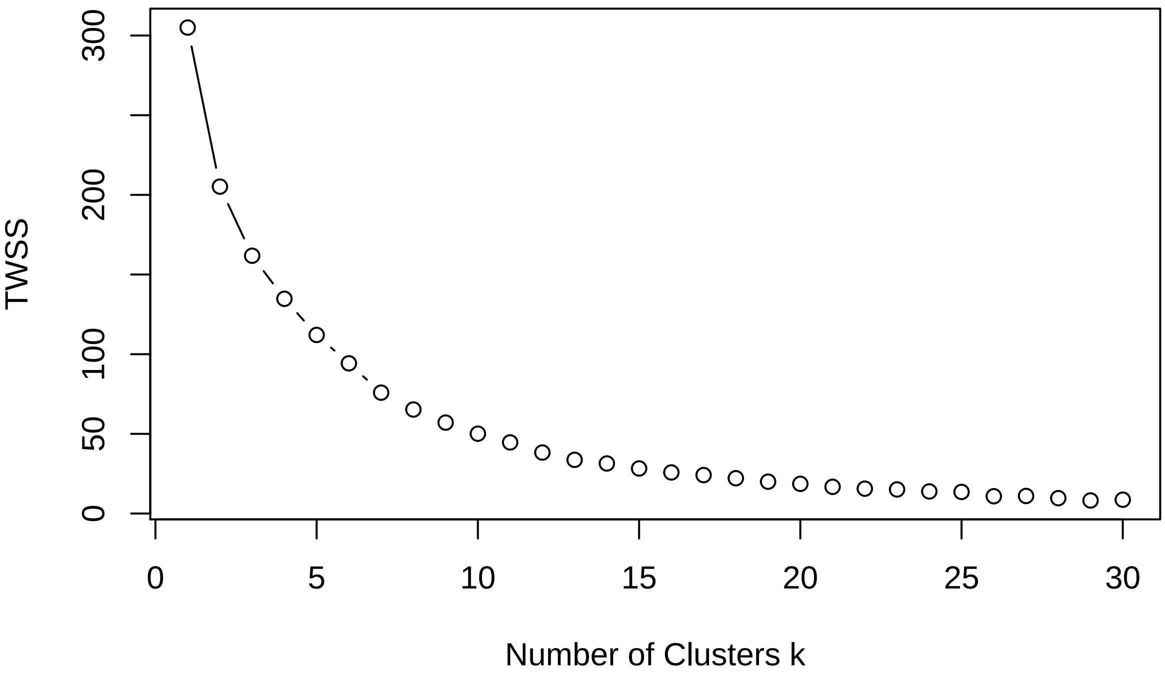


Q4

a)

```
music <- read.csv("data/music.csv")
music.feat <- music[, 4:8]
music.feat <- scale(music.feat)
wss <- vector(length=30)

for (k in 1:30) wss[k] <- kmeans(music.feat,centers=k, nstart = 25)$tot.withinss
plot(1:30, wss, type="b", xlab="Number of Clusters k",
     ylab="TWSS")
```



b)

TWSS declines slowly. Data does not partition into few, small, well-defined compact clusters.

```
clusk <- kmeans(music.feat,centers=5, nstart = 25)$cluster

table(music$Artist, clusk)
```

```
##               clusk
##                1 2 3 4 5
##     Abba       1 0 0 9 0
##     Beatles    0 0 8 2 0
##     Beethoven  1 5 0 2 0
##     Eels       0 0 7 3 0
##     Enya       2 0 0 1 0
##     Mozart     0 6 0 0 0
##     Vivaldi    3 5 0 1 1
```

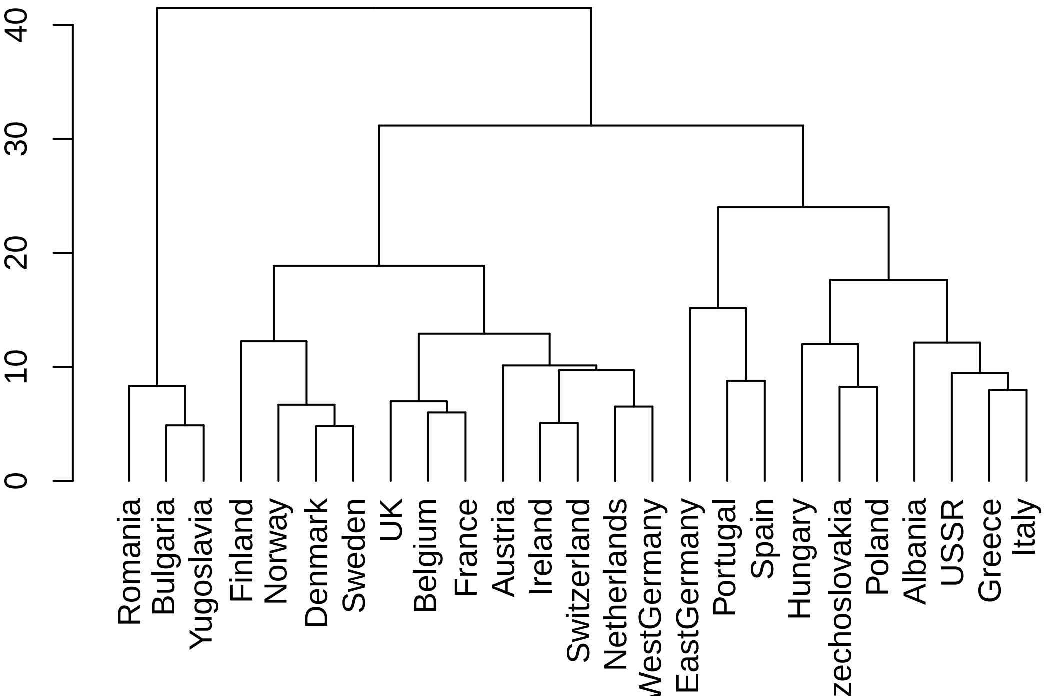
Students may not have set.seed, could have different output due to random starts.

All but one Abba tracks in a single cluster. Most of Beatles an the Eels tracks in a single cluster.

Q5

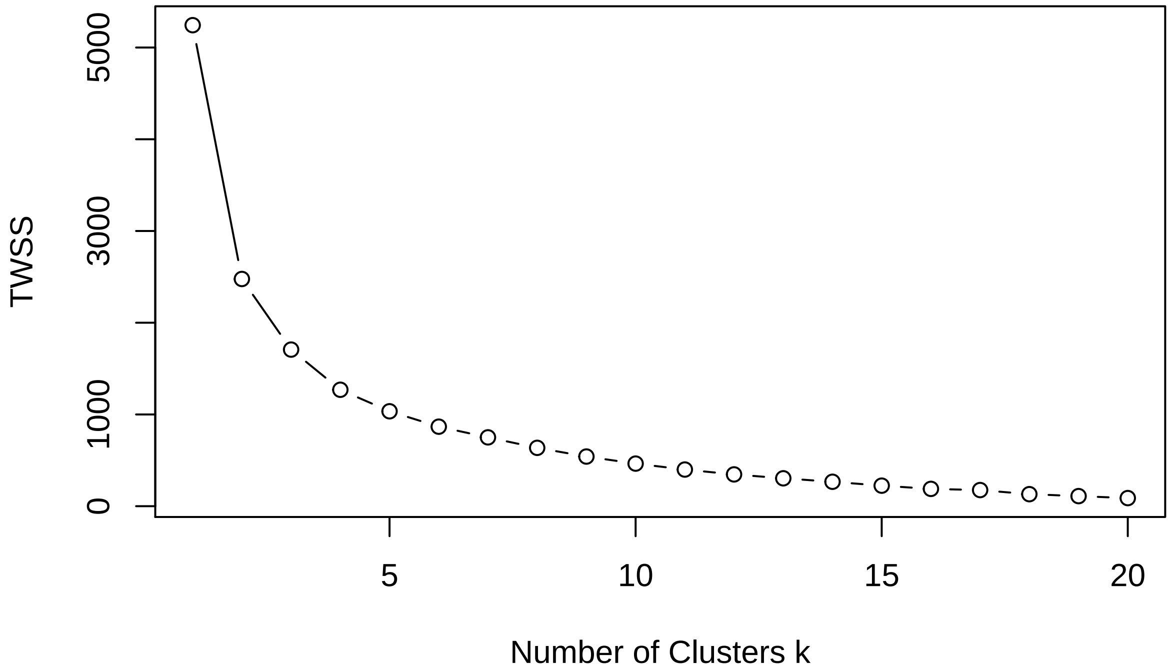
Anything sensible.

```r
protein <- read.csv("data/protein.csv")
protein.feat <- protein[, 2:10]
row.names(protein.feat) <- protein$Country
d <- dist(protein.feat)
h <- hclust(d, "complete")
plot(as.dendrogram(h))
```

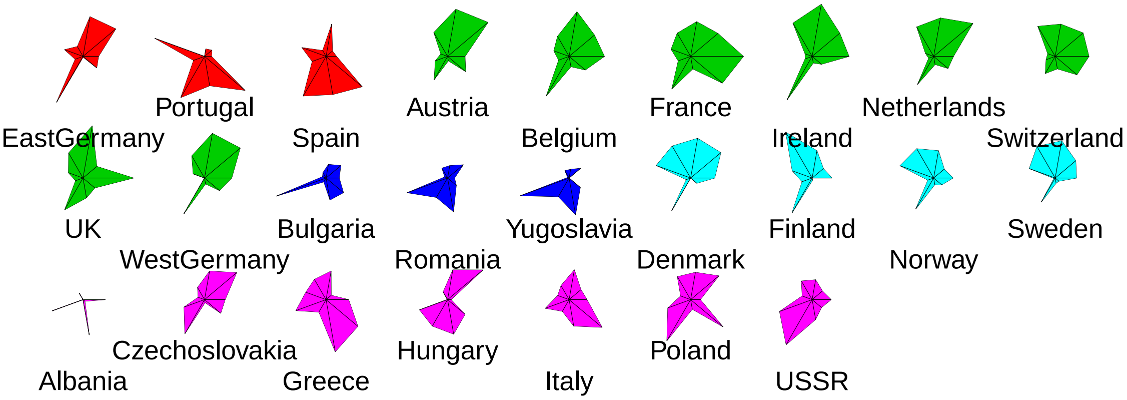


```r
hc <- cutree(h, 5)


wss <- vector(length=20)
for (k in 1:20) wss[k] <- kmeans(protein.feat,centers=k, nstart = 10)$tot.withinss
plot(1:20, wss, type="b", xlab="Number of Clusters k",
     ylab="TWSS")
```



```
clusk <-  kmeans(protein.feat,centers=5, nstart = 10)$cluster
o <- order(clusk)
stars(protein.feat[o,],nrow=3, col.stars=clusk[o]+1)
```



```
table(clusk, hc)
```

```
##      hc
## clusk 1 2 3 4 5
##     1 0 0 0 0 3
##     2 0 8 0 0 0
##     3 0 0 3 0 0
##     4 0 0 0 4 0
##     5 7 0 0 0 0
```