Assignment 1

```
library(tidyverse) # essential tools
library(ggdendro)
```

Question 1

For the dataset given, do these calculations by hand:

- Calculate a distance matrix using squared euclidean distance.
- Use hierarchical clustering with single linkage to cluster the data. Draw the dendro- gram and identify the two-cluster solution.
- Use hierarchical clustering with average linkage to cluster the data. Draw the den- drogram and identify the two-cluster solution.
- Cluster the data using kmeans with k = 2. Use starting clusters of (a,b,c) and (d,e).

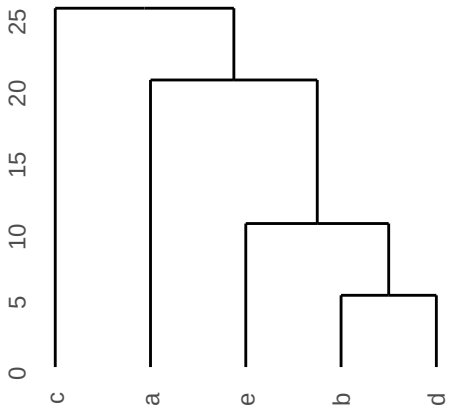
```
set.seed(123)
# Generating data
x <- matrix(sample(-5:5, 10), nrow = 5)
rownames(x) <- letters[1:5]
colnames(x) <- c("U", "V")
x
```

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

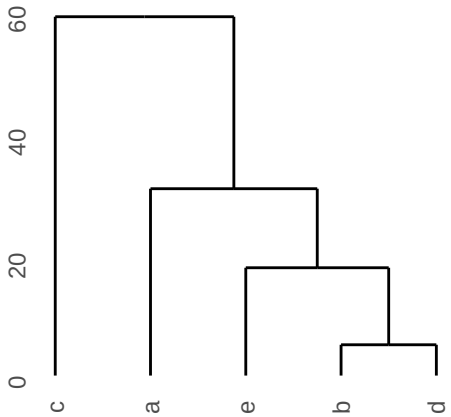
```
dx <- dist(x)^2 # finding distances
dx
```

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

```
# building  clusters
hs <- hclust(dx, "single")
ha <- hclust(dx, "average")


# ggplot version
ggdendrogram(as.dendrogram(hs))
```




```
ggdendrogram(as.dendrogram(ha))
```

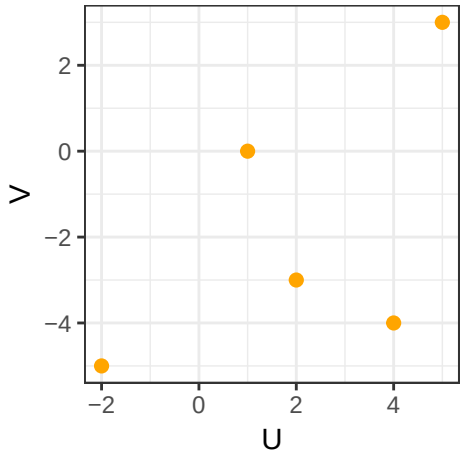



```
c1 <- apply(x[1:3,], 2, mean)
c2 <- apply(x[4:5,], 2, mean)


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"
```

```
x %>%
  as.data.frame() %>%
  ggplot(aes(U, V)) +
  geom_point(colour = "orange", size = 2) +
  theme_bw()
```



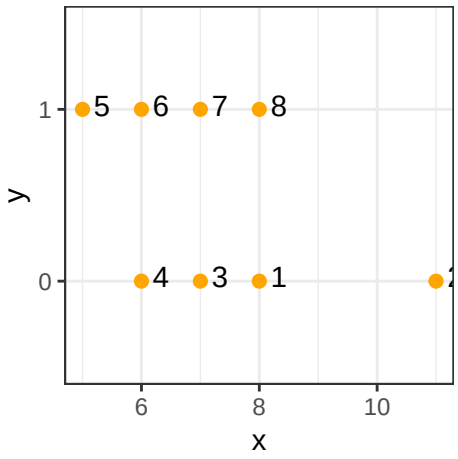
Question 2

Eight online shoppers buy 8, 11, 7 , 6, 5, 6 , 7, 8 pairs of socks. The same eight shoppers buy 0, 0, 0, 0, 1, 1, 1, 1 computers.

a) If you run kmeans on this data with $k = 2$,

with no scaling, what result would you expect?

```
my_df <- data.frame(x = c(8, 11, 7, 6, 5, 6, 7, 8),
                    y = c(0, 0, 0,0, 1, 1, 1, 1))

my_df %>%
ggplot(aes(x, factor(y))) +
  geom_point(colour = "orange", size = 2) +
  theme_bw() +
  geom_text(label = 1:8,hjust= -0.7, vjust = 0.3) +
  labs(y = 'y')
```




```
kmeans(my_df, 2)$cluster
```

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

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

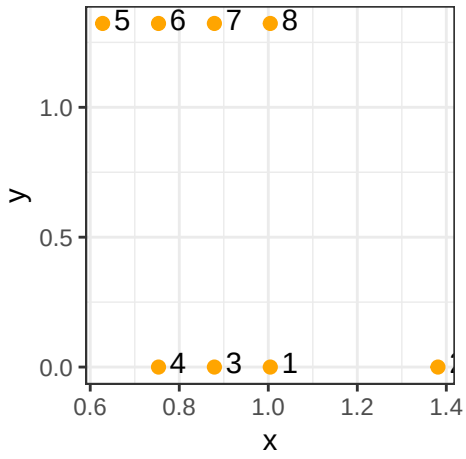
```
## [1] 2 1 2 2 2 2 2 2
```

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

b) If both variables are scaled to unit standard deviation, what will

kmeans with $k = 2$ give you?

```
d1 <- my_df %>%
  mutate_all(scale, center = FALSE)

d1 %>%
ggplot(aes(x, y)) +
  geom_point(colour = "orange", size = 2) +
  theme_bw() +
  geom_text(label = 1:8,hjust= -0.7, vjust = 0.3) +
  labs(y = 'y')
```




```
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"
```

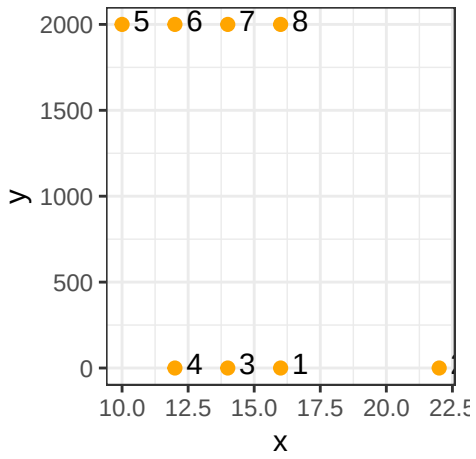
Points 1-4 in 1 cluster, 5-8 in the other.

c) Suppose socks cost 2 euro and the computer is 2000 euro. What is you

clustered the amount spent by each customer using kmeans with $k = 2$, with no scaling?

```
d <- my_df %>%
  mutate(x = 2*x, y = 2000*y)

d %>%
ggplot(aes(x, y)) +
  geom_point(colour = "orange", size = 2) +
  theme_bw() +
```

```
  geom_text(label = 1:8,hjust= -0.7, vjust = 0.3) +
  labs(y = 'y')
```




```
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.

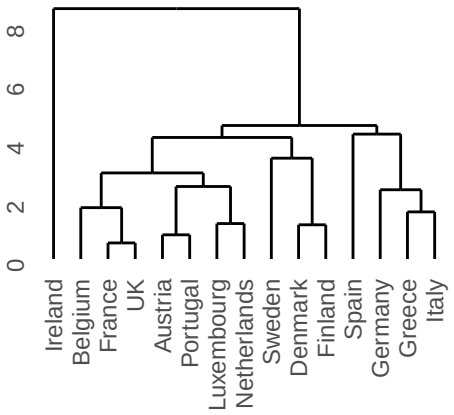
Question 3

The file `eupop.txt` contains the population and percentage distribution by age for EU countries in 1999. The age categories are 0-14 years, 15-44 years, 45-64 years and 65 years and over.

a) Construct the euclidean distance matrix of the percentage variables.

Use it to cluster the countries, using average linkage. Draw the dendrogram and interpret. Are there any outlier countries?

```
eupop <- read.table("data/eupop.txt") %>%
  select(-5)
d <- dist(eupop)
h <- hclust(d, "average")
ggdendrogram(as.dendrogram(h))
```



Ireland is an outlier.

b) Examine the 3-cluster solution. Which countries belong to each of

the three clusters? Summarise the partitions with sumPartition (in h1code.R) Interpret your findings.

```
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) Use the kmeans algorithm to find another 3-cluster grouping of

countries. Which countries belong to each of the three clusters?

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

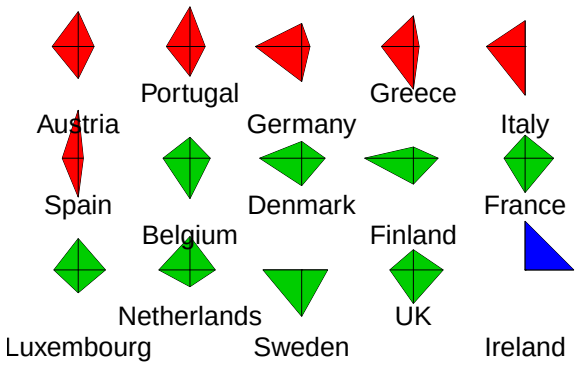
```
## K-means clustering with 3 clusters of sizes 6, 8, 1
##
## Cluster means:
##       p014    p1544    p4564     p65.
## 1 15.81667 44.03333 23.91667 16.26667
## 2 18.55000 42.01250 24.11250 15.36250
## 3 22.20000 46.20000 20.30000 11.30000
##
## Clustering vector:
##     Austria     Belgium     Denmark     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.

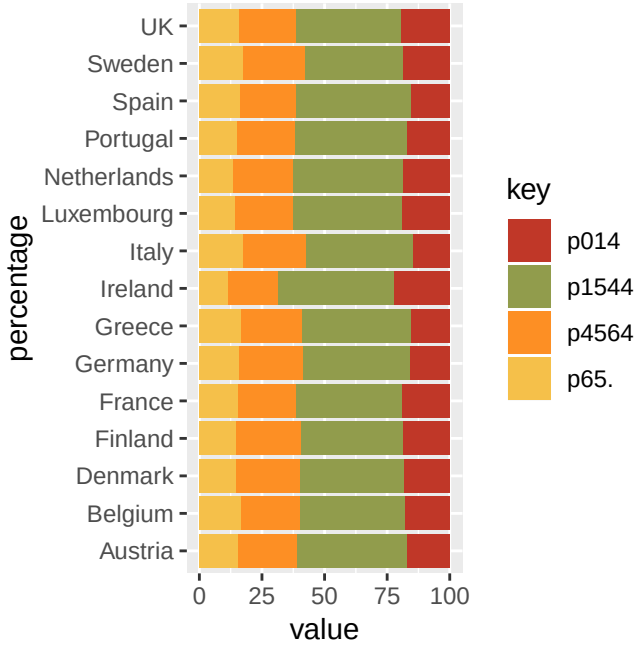
d) Construct a stars plot which shows the data and clustering obtained

from kmeans. Optional: can you think of a better way of showing the clusters? Can you think of a way to present the data and the clustering results of both methods on the same graphical display?

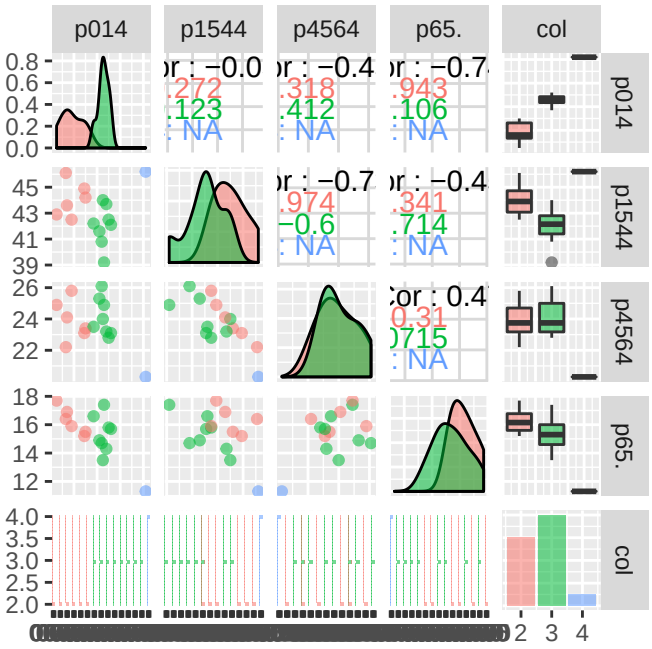
```
clusk <- km$cluster
o <- order(clusk)
stars(eupop[o, ], nrow = 3, col.stars = clusk[o] + 1)
```




```
eupop[o,] %>%
  mutate(country = rownames(.)) %>%
  gather(key, value, -country) %>%
  ggplot(aes(y = value, x = country, fill = key)) +
  geom_bar(stat = "identity") +
  labs(x = "percentage") +
  coord_flip() +
  ggpomological::scale_fill_pomological()
```



```
library(GGally)
eupop %>%
  mutate(col = clusk+1) %>%
  ggpairs(aes(colour = factor(col), alpha = 0.4))
```



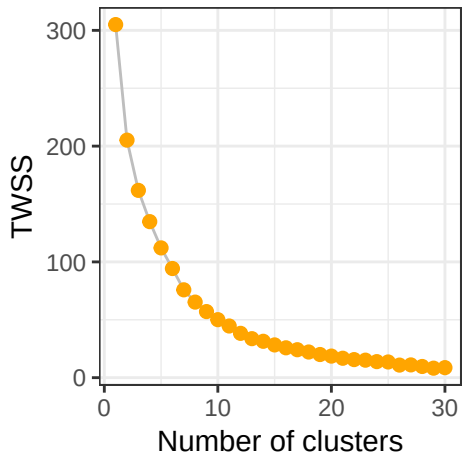
Question 4

Analyzing the music data.

a) Run the k-means algorithm over the range $k = 1, \ldots, 15$

clusters and record the total within cluster sum of squares (TWSS). Let nstart = 25. Plot k versus TWSS and choose the best fitting number of clusters. What do you observe? Note: remember to scale the data.

```
music <- read.table("data/music.txt")
music_feat <- music %>%
  select(3:7) %>% mutate_all(scale)

results <-
  data.frame(
    tot.withinss = 1:30 %>%
      purrr::map(kmeans, x = music_feat, nstart = 25) %>%
      purrr::map_dbl("tot.withinss"),
    ind = 1:30)


results %>%
ggplot(aes(ind, tot.withinss)) +
  geom_line(colour = 'grey') +
  geom_point(colour = "orange", size = 2) +
  theme_bw() +
  labs(y = 'TWSS', x = "Number of clusters")
```



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

b) Make a table of artist vs cluster solution from k = 5.

```r
clusk <- kmeans(music_feat, centers = 5, nstart = 25)$cluster


music %>%
  mutate(clusk = clusk) %>%
  group_by(Artist, clusk) %>%
  count()
```

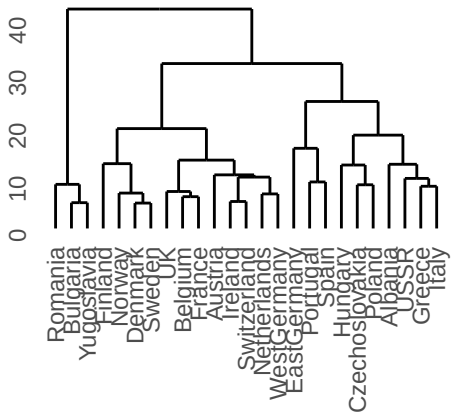
```
## # A tibble: 19 x 3
## # Groups:   Artist, clusk [19]
##    Artist    clusk     n
##    <fct>     <int> <int>
##  1 Abba          1     1
##  2 Abba          4     9
##  3 Beatles       3     8
##  4 Beatles       4     2
##  5 Beethoven     1     1
##  6 Beethoven     2     5
##  7 Beethoven     4     2
##  8 Eels          3     7
##  9 Eels          4     3
## 10 Enya          1     2
## 11 Enya          4     1
## 12 Mozart        2     6
## 13 Vivaldi       1     3
## 14 Vivaldi       2     5
## 15 Vivaldi       4     1
## 16 Vivaldi       5     1
## 17 <NA>          1     3
## 18 <NA>          2     1
## 19 <NA>          4     1
```

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

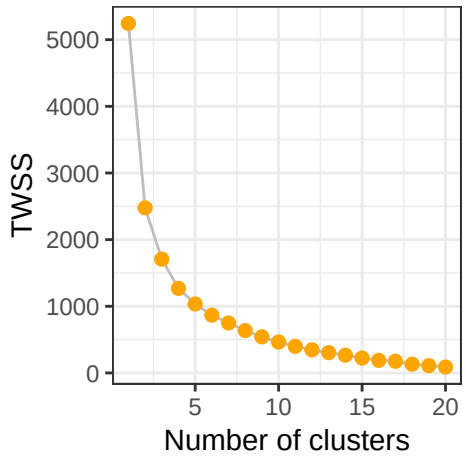
Question 5

Protein data. We want to study the similarities and differences in the protein composition of the diets of different countries. Using any methods that you choose from this course or otherwise, write a brief summary.

```
protein <- read.table("data/protein.txt")
protein_feat <- protein %>%  select(2:10)
row.names(protein_feat) <- protein$Country
d <- dist(protein_feat)
h <- hclust(d, "complete")
ggdendrogram(as.dendrogram(h))
```

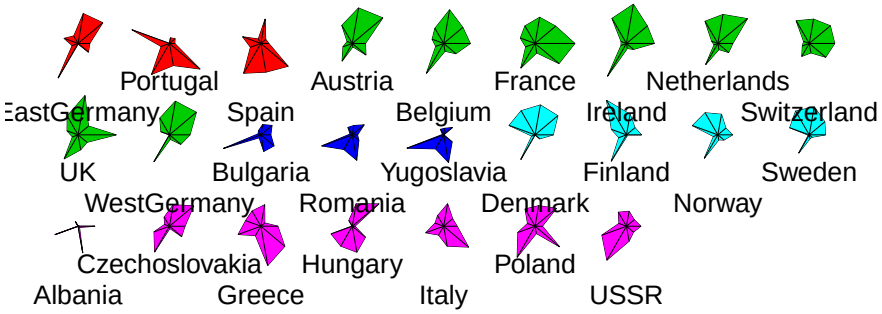



```
hc <- cutree(h, 5)


results <-
  data.frame(
    tot.withinss = 1:20 %>%
      purrr::map(kmeans, x = protein_feat, nstart = 10) %>%
      purrr::map_dbl("tot.withinss"),
    ind = 1:20)

results %>%
ggplot(aes(ind, tot.withinss)) +
  geom_line(colour = 'grey') +
  geom_point(colour = "orange", size = 2) +
  theme_bw() +
  labs(y = 'TWSS', x = "Number of clusters")
```




```
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
```