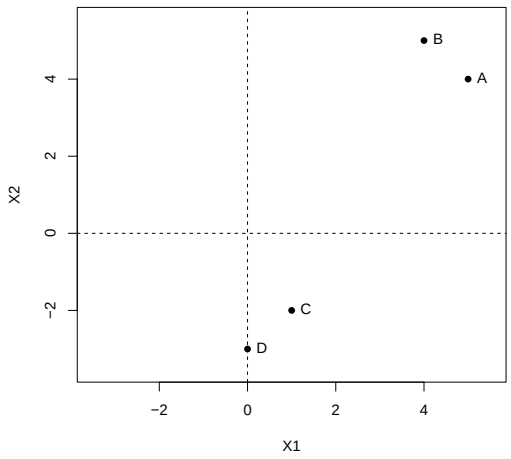
TP2 : partitionnement avec la méthode des kmeans

Exercice 1

On considère le tableau de données où 4 individus (ici des points) A, B, C et D sont décrit sur deux variables (X1 et X2):

	X1	X2
A	5	4
B	4	5
C	1	-2
D	0	-3



1. Rapeller la définition d'une partition.
2. Quel critère de qualité est optimisé par la méthode de **kmeans** ?
3. Ce critère est-il optimisé localement ou globalement ? Pourquoi ?
4. Proposer une mauvaise (au sens de ce critère) partition en deux classes de $\Omega = \{A, B, C, D\}$.
5. Appliquer à la main la méthode des **kmeans** en prenant comme centres initiaux les points A et B. Quelle partition en deux classes est obtenue ?
6. Appliquer avec R la méthode des **kmeans** en prenant comme centres initiaux les points A et B. Vérifiez que vous retrouver la même partition en deux classes.
7. A quoi correspondent les sorties `totss`, `tot.withinss` et `tot.bss` ?
8. En déduire le pourcentage d'inertie expliquée par cette patition.

Exercice 2.

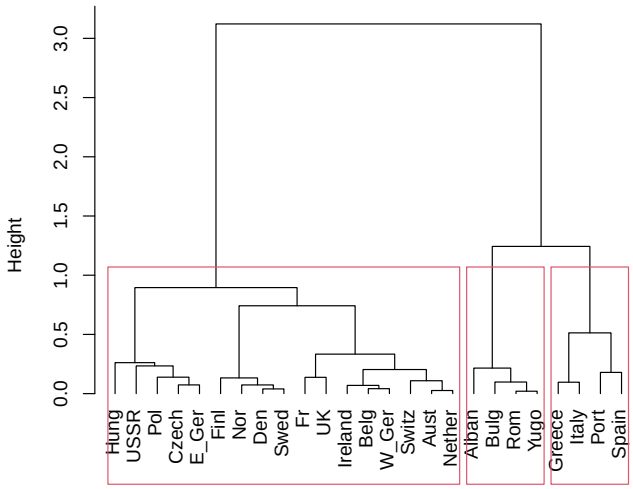
On considère le jeu de données `fromage` où $n = 29$ fromages sont décrit sur $p = 9$ variables quantitatives. On souhaite déterminer automatiquement des groupes de fromages avec la méthode des **kmeans** et les comparer à ceux obtenus par la CAH de Ward.

1. Importer les données depuis le fichier `fromage.txt`.
2. Standardiser les données (en utilisant la variance empirique non corrigée).
3. Constuire la CAH de Ward en pondérant tous les individus par $\frac{1}{n}$ et couper l'arbre pour obtenir une partition en 5 classes.
4. Quelle est le pourcentage de l'inertie des données expliquée par cette partition ?
5. Utiliser maintenant la fonction `kmeans` pour trouver une partition en 5 classes. Quel est le pourcentage d'inertie expliquée par cette partition ?
6. Recommencer plusieurs fois. Que constatez-vous ?
7. A quoi sert l'argument `nstart`. Augmentez la valeur de cet argument et relancer plusieurs fois la fonction **kmeans**. Que constatez-vous ?
8. Quelle méthode donne finalement la meilleure partition en 5 classes ? Pourquoi ?
9. Croiser les deux partitions pour les comparer.

Exercice 3 : consolider une partition

On reprend le jeu de données `protein`. La partition obtenue par CAH de Ward n'est pas optimale et peut être améliorée, consolidée par les **kmeans**.

1. Appliquer la CAH de Ward sur les données standardisées. Vérifiez que la somme des hauteurs est égale à l'inertie totale (ici le nombre de variables).
2. Quel est la proportion d'inertie expliquée par la partition en 3 classes obtenue en coupant cet arbre ?



3. Consolider cette partition en 3 classes en appliquant les **kmeans** à cette partition initiale.

4. Croiser les deux partitions (avant et après consolidation). Vous devez retrouver le tableau croisé ci-après.

```
##           part_km
## part_ward  1  2  3
##         1  4  0  0
##         2  2 15  0
##         3  0  0  4
```

5. Combien de pays ont changés de classe ? Quelle est l'amélioration du pourcentage d'inertie expliquée ?

Exercice 4 : clustering avec beaucoup d'individus

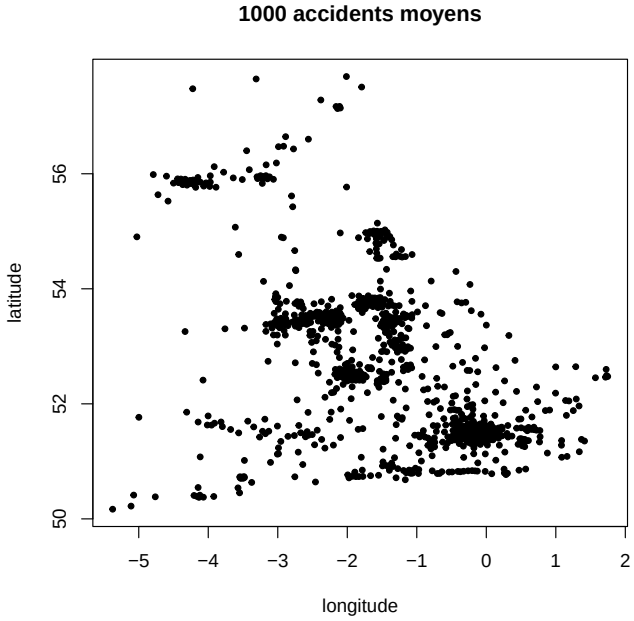
On regarde des données `urbanGB` récupérées sur le site de l'UCI Machine Learning Repository :

https://archive.ics.uci.edu/ml/datasets.php

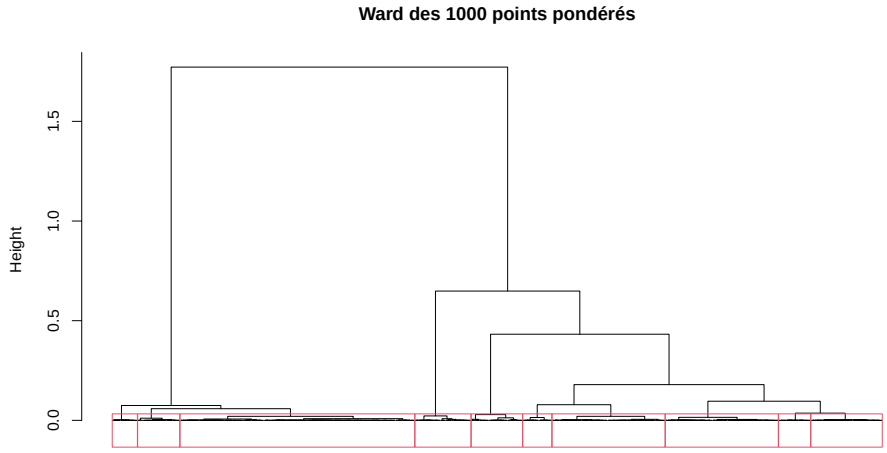
Ces données fournissent les coordonnées géographiques (longitude, latitude) de 360177 accidents de voitures ayant eu lieu dans des zones urbaines de Grande-Bretagne.

```
accidents <- read.table(file="../data/urbanGB/urbanGB.txt", sep = ",",
                  dec=".", header = FALSE)
colnames(accidents) <- c("longitude", "latitude")
```

1. Faire un plot des 360177 points.
2. Réduire le nombre de points à 1000 par clustering.



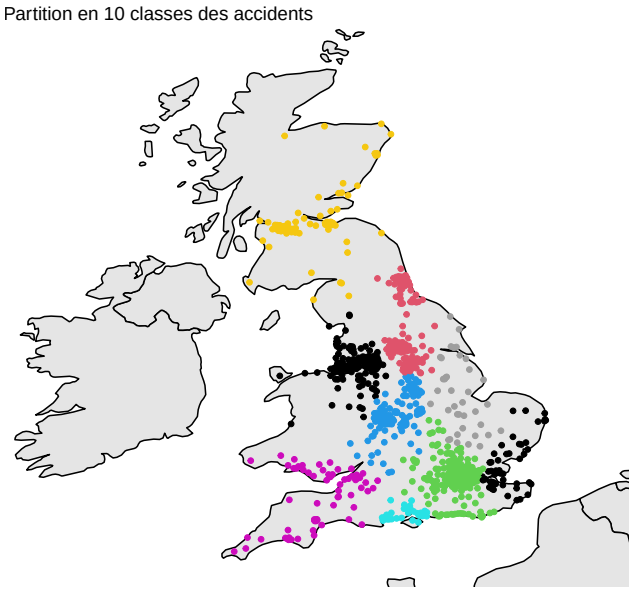
3. Appliquer la CAH de Ward aux 1000 points **pondérés par les poids des classes**. La partition en 10 classes de ces 1000 poins est finalement retenue.



4. Représenter la carte d'Angleterre avec le code ci-dessous.

```
library(RColorBrewer)
library(ggplot2)
library(maps)
worldmap = map_data('world')
carte <- ggplot() +
  geom_polygon(data = worldmap, aes(x = long, y = lat, group = group),
               fill = 'gray90',color = 'black') +
  coord_fixed(ratio = 1.3,  xlim = c(-10,3), ylim = c(50, 59)) +
  theme_void()
```

5. Ajouter à cette carte les 1000 points moyens d'accidents colorés en fonction de leur classe. Qu'en pensez-vous ?



Exercice 5 : clustering de données mixtes

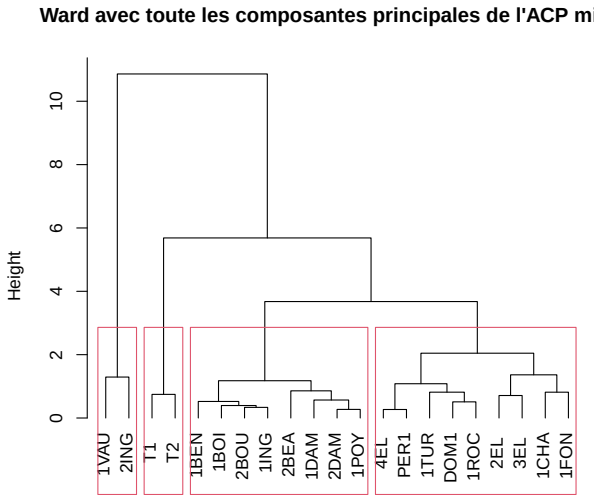
Le jeu de données `wine` décrit $n = 21$ vins sur un mélange de variables qualitatives (appellation d'origine et type de sol) et de quantitatives (descripteurs sensoriels).

```
library(FactoMineR)
data(wine)
print(wine[1:5,c(1:2,27,26,29)])
```

```
##           Label      Soil Bitterness Smooth Harmony
## 2EL      Saumur      Env1      1.926  2.731   3.143
## 1CHA     Saumur      Env1      1.926  2.500   2.964
## 1FON Bourgueuil      Env1      2.000  2.679   3.143
## 1VAU     Chinon      Env2      1.963  1.680   2.038
## 1DAM     Saumur Reference      2.071  3.036   3.643
```

L'idée est de trouver automatiquement des classes de vins ayant les mêmes profils sensoriels. Pour cela, il faut recoder les données en données quantitatives soit via une ACP sur données mixtes, soit via un clustering de variables.

1. Faire une `ACP mixte` de ces données avec la fonction `FAMD` (Factor Analysis of Mixed Data) du package `FactoMineR` et interpréter dans un premier temps les graphiques.
2. On sait qu'en `ACP mixte`, le nombre maximum de composantes principales est $min(n-1, p1+m-p2)$ p_1 est le nombre de variables quantitatives, p_2 est le nombre de variables qualitatives, et m est le nombre total de modalités. Quel est donc ici le nombre maximum de composantes principales ? Refaire l'`ACP mixte` en conservant cette fois toutes les composantes principales.
3. Constuire le dendrogramme de Ward avec toutes les composantes principales de l'`ACP mixte`. La partition en 4 classes est retenue.



3. Consolider cette patition avec la méthode des *kmeans*. Combien de vins ont changés de classe ?

```
##        part_ward
## part_km 1 2 3 4
##       1 7 0 0 0
##       2 0 2 0 0
##       3 2 0 8 0
##       4 0 0 0 2
```

3. Interpréter les 4 classes des vins via les graphiques ci-dessous.