T. D. n° 3
Classification ascendante hiérarchique et la méthode des K-moyenne

Résumé

Ce document est le TD n° 3 du module Analyse exploratoire. Il reprend rapidement des éléments du cours et propose une mise en pratique de deux approches : la classification ascendante hiérarchique (CAH) avec *hclust()* et la méthode des K-moyenne (k-means) avec *kmeans()*.

1 World of Worcraft

FIGURE 1 – World of WorldCraft

source :`http://skalbunk.deviantart.com`

1.1 Chargement des données

Commencez par charger les données du fichier *wow_ royaumes.csv*. Associez ces données à un dataframe et appliquez la fonction *summary()*. Il est également possible de télécharger le fichier *wow_ royaumes.csv* depuis la plate-forme Kaggle : `https://www.kaggle.com/mylesoneill/warcraft-avatar-history`. Ce fichier provient

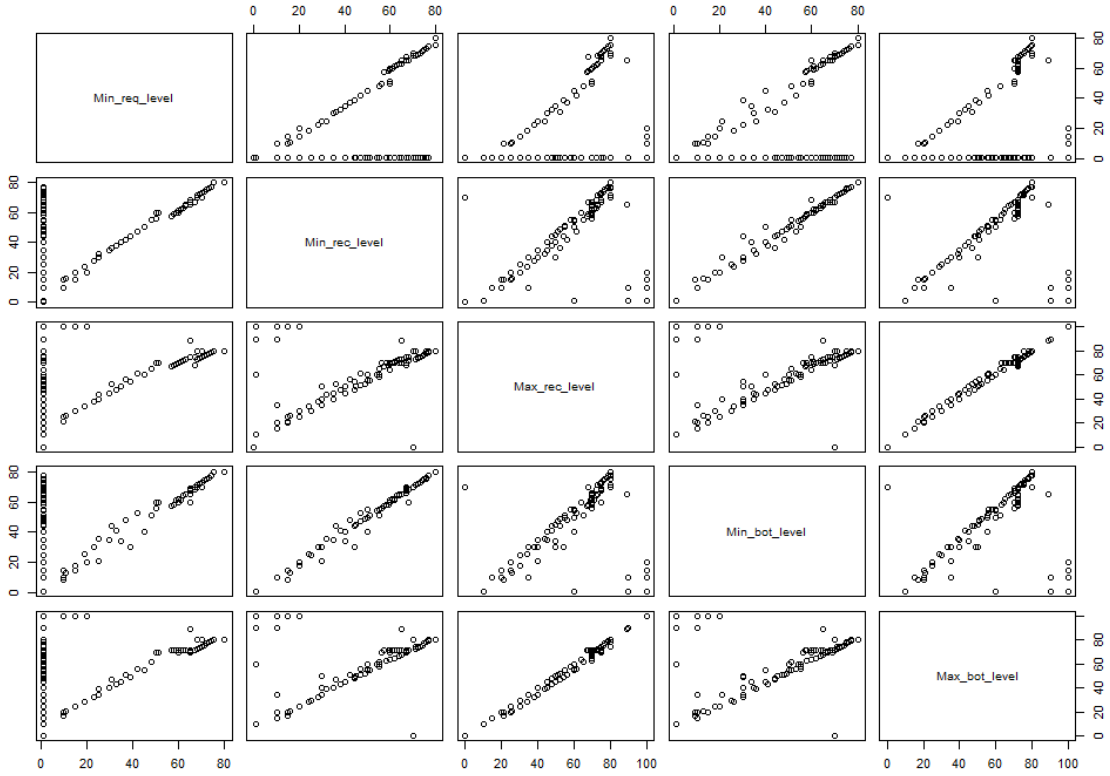
d'une collection d'informations à propos des joueurs du célèbre jeu World of Warcraft depuis 2008. Dans ce TP, nous observerons les royaumes dans lesquels évoluent les joueurs.
Affichez un croisement deux à deux pour chaque variable quantitative grâce à la fonction *pairs()*. Créez un sous ensemble de données composées uniquement des variables quantitatives.

```
> royau <- read.table('C:/Users/claey/Documents/cour/My TD/TD3/wow-
    royaumes.csv', sep =',',header=T,row.names=1,dec=".")

> print(head(royau))
                           Continent             Area
Durotar                     Kalimdor Central Kalimdor
The Barrens                 Kalimdor Central Kalimdor
Silverpine Forest    Eastern Kingdoms        Lordaeron
Stonetalon Mountains        Kalimdor Central Kalimdor
Thunder Bluff               Kalimdor Central Kalimdor
Dustwallow Marsh            Kalimdor Central Kalimdor
                                    Zone Subzone Type Size
                                       Controlled
Durotar                          Durotar         Zone   NA
    Horde
The Barrens                   The Barrens         Zone   NA
    Contested
Silverpine Forest        Silverpine Forest         Zone   NA
    Horde
Stonetalon Mountains Stonetalon Mountains         Zone   NA
    Contested
Thunder Bluff               Thunder Bluff         City   NA
    Horde
Dustwallow Marsh          Dustwallow Marsh         Zone   NA
    Contested
                     Min_req_level Min_rec_level Max_rec_level
Durotar                          1             1            10
The Barrens                      1            10            35
Silverpine Forest                1            10            20
Stonetalon Mountains             1            25            30
Thunder Bluff                    1             1           100
Dustwallow Marsh                 1            35            40
                     Min_bot_level Max_bot_level
Durotar                          1            10
The Barrens                     10            35
Silverpine Forest               10            20
Stonetalon Mountains            25            30
Thunder Bluff                    1           100
Dustwallow Marsh                35            40
> print(summary(royau))
            Continent                  Area                     Zone
 Eastern Kingdoms:54   Northrend        :30                       :
    7
 Kalimdor        :45   Outland          :29   Dragonblight        :
    7
 Northrend       :30   Lordaeron        :18   Netherstorm         :
```

```
   6
 Other            : 1   Central Kalimdor :15   Blackrock Mountain:
   5
 Ouland           : 1   Khaz Modan       :14   Hellfire Peninsula:
   5
 Outland          :28   Northern Kalimdor:14   Tanaris           :
   5
 The Great Sea    : 1   (Other)          :40   (Other)
   :125
               Subzone              Type          Size
                   :108   Zone         :76   Min.   : 5.00
 Auchindoun        :  4   Dungeon      :63   1st Qu.: 5.00
 Coilfang Reservoir:  4   Arena        : 5   Median : 5.00
 Caverns of Time   :  3   Battleground: 5   Mean   :11.54
 Coldarra          :  3   Sea          : 5   3rd Qu.:16.25
 Tempest Keep      :  3   City         : 4   Max.   :40.00
 (Other)           : 35   (Other)      : 2   NA's   :92
    Controlled  Min_req_level   Min_rec_level   Max_rec_level
 Alliance : 17   Min.   : 1.00   Min.   : 0.00   Min.   :  0.00
 Contested:114   1st Qu.: 1.00   1st Qu.:10.00   1st Qu.: 42.00
 Horde    : 17   Median : 1.00   Median :45.00   Median : 70.00
 PvP      : 10   Mean   :22.98   Mean   :39.98   Mean   : 61.92
 Sanctuary:  2   3rd Qu.:57.25   3rd Qu.:67.00   3rd Qu.: 80.00
                 Max.   :80.00   Max.   :80.00   Max.   :100.00
                                 NA's   :1       NA's   :1
 Min_bot_level   Max_bot_level
 Min.   : 1.00   Min.   :  0.00
 1st Qu.:10.00   1st Qu.: 40.75
 Median :46.00   Median : 70.00
 Mean   :40.52   Mean   : 61.97
 3rd Qu.:68.00   3rd Qu.: 80.00
 Max.   :80.00   Max.   :100.00
 NA's   :2       NA's   :2
> pairs(royau[,c(8:12)])
> royau.x <- royau[c(8:12)]
```

FIGURE 2 – Croisement deux à deux des variables quantitatives



À vous !

a) Omettez les variables avec NA au dataframe *royau.x*.

b) Centrez et réduisez les variables grâce à la fonction *scale()*, stockez ce résultat dans *royau.cr*.

c) Justifiez le centrage-réduction sur les variables.

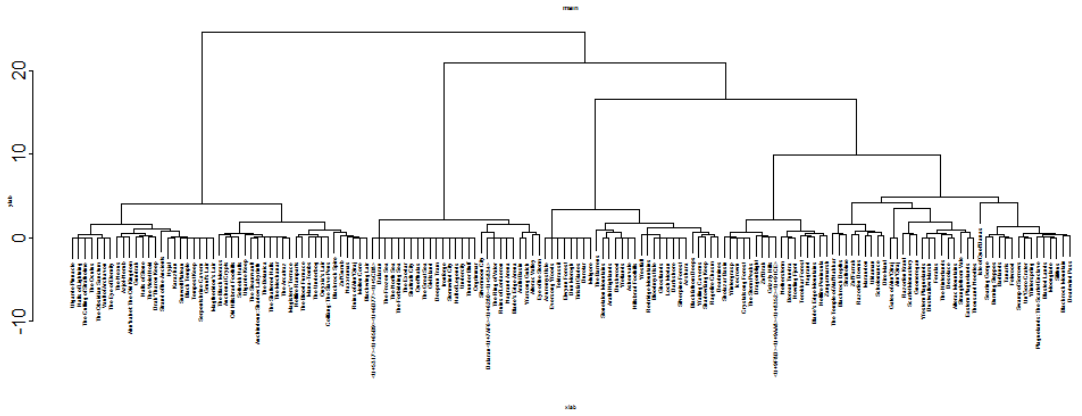
1.2 CAH

La fonction *dist()* calcule et renvoie la matrice des distances calculées en utilisant une mesure de distance spécifiée (ou par défaut la distance L2) pour calculer les distances entre les lignes d'une matrice de données. Appliquez cette fonction à votre dataframe. *royau.cr*.
La méthode suppose qu'on dispose d'une mesure de dissimilarité entre les individus, dans le cas de points situés dans un espace L2, on peut utiliser la distance comme mesure de dissimilarité. La classification ascendante hiérarchique est dite ascendante car elle part d'une situation où tous les individus sont seuls dans une classe, puis sont rassemblés en classes de plus en plus grandes. Le qualificatif hiérarchique vient du fait qu'elle produit une hiérarchie H à l'ensemble des classes à toutes les étapes de

l'algorithme. Dans ce TD nous utilisons la méthode de Ward pour séparer des individus en classes. La méthode de Ward propose qu'à chaque pas, on cherche à obtenir un minimum local de l'inertie intraclasse ou un maximum de l'inertie interclasse. En d'autres termes, elle consiste à réunir les deux clusters dont le regroupement fera le moins baisser l'inertie interclasse. On suppose tout de même l'existence de distances euclidiennes entre observations. Cette technique tend à regrouper les petites classes entre elles.

```
d.royau <- dist(royau.cr)
cah.ward <- hclust(d.royau,method="ward.D2")
par(cex=0.3, mar=c(5, 8, 4, 1))
plot(cah.ward, xlab="", ylab="", main="", sub="", axes=FALSE)
par(cex=1)
title(xlab="xlab", ylab="ylab", main="main")
axis(2)
```

FIGURE 3 – Classification ascendente hiérarchique



On décide constituer quatre regroupements de classes.

```
groupes.cah <- cutree(cah.ward,k=4)
```

À vous !

a) Proposez une autre visualisation de *cah.ward*.

b) Affichez les quatres groupes en utilisant la fonction *rect.hclust()*.

c) Affichez le contenue de chaque groupe.

1.3 K-mean

La méthode dite *K-means* propose de choisir aléatoirement un point comme le barycentre de chaque groupe puis de le recalculer à chaque nouvel individu introduit dans le groupe. Cette technique est intéressante car elle s'adapte lorsqu'on injecte de nouveaux individus. *K-means*, à la différence de la CAH, ne fournit pas d'outil d'aide à la détection du nombre de classes. Nous devons les programmer sous R ou utiliser

des procédures proposées par des packages dédiés. En faisant varier le nombre de groupes on observe l'évolution d'un indicateur de l'aptitude des individus à être plus proches entre eux dans un même groupe, que des individus des autres groupes. Cet indicateur traduit de la qualité d'une solution.

```
> groupes.kmeans <- kmeans(royau.cr,centers=4,nstart=5)
> print(groupes.kmeans)
K-means clustering with 4 clusters of sizes 41, 47, 29, 40

Cluster means:
  Min_req_level Min_rec_level Max_rec_level Min_bot_level
     Max_bot_level
1     -0.5503317     0.5517434   -0.01098919     0.5501013
    -0.03881017
2      1.4401329     0.9946827    0.47164460     0.9939750
    0.48303454
3     -0.7021036    -1.2931871    1.22025443    -1.2808127
    1.23421432
4     -0.6190411    -0.7967285   -1.42760295    -0.8031852
    -1.42259053

Clustering vector:
                          Durotar                       The Barrens
                                4                                 4
                Silverpine Forest              Stonetalon Mountains
                                4                                 4
                    Thunder Bluff                  Dustwallow Marsh
                                3                                 4
                        Orgrimmar                        Undercity
                                3                                 3
                        Ashenvale                Stranglethorn Vale
                                4                                 1
                   Wailing Caverns                          Tanaris
                                4                                 1
                         Maraudon                   The Hinterlands
                                4                                 4
                   Un'Goro Crater                           Felwood
                                1                                 1
       The Temple of Atal'Hakkar                  Blackrock Depths
                                1                                 1
                     Winterspring                Eastern Plaguelands
                                1                                 1
                          Azshara                       Scholomance
                                4                                 1
                     Arathi Basin                    Blackrock Spire
                                3                                 2
              Blackrock Mountain                          Moonglade
                                1                                 1
                          Mulgore                  Thousand Needles
                                4                                 1
             Hillsbrad Foothills                  Blackfathom Deeps
                                4                                 4
                         Desolace                         Darkshore
```

4	4
Alterac Mountains	Swamp of Sorrows
4	1
Tirisfal Glades	Searing Gorge
4	1
Burning Steppes	Zul'Farrak
1	1
Feralas	Arathi Highlands
4	4
Blasted Lands	Western Plaguelands
1	4
Stratholme	Scarlet Monastery
1	4
Shadowfang Keep	Badlands
4	1
Wetlands	Redridge Mountains
4	4
Hall of Legends	Gnomeregan
3	4
Razorfen Downs	Duskwood
1	4
Loch Modan	Silithus
4	1
Deadwind Pass	Razorfen Kraul
1	4
Ragefire Chasm	Dire Maul
4	1
Uldaman	Dun Morogh
4	4
Elwynn Forest	Zul'Gurub
4	2
Teldrassil	Westfall
4	4
Stormwind City	Molten Core
3	2
Onyxia's Lair	Warsong Gulch
2	3
Blackwing Lair	Deadmines
2	4
Gates of Ahn'Qiraj	Alterac Valley
3	3
Ironforge	Hyjal
3	2
Deeprun Tram	Ruins of Ahn'Qiraj
3	2
Ahn'Qiraj	Naxxramas
3	2
GM Island	The Great Sea
3	3
Nagrand Arena	Blade's Edge Arena
3	3
Quel'thalas	Hellfire Peninsula
3	1
Ghostlands	Silvermoon City

4	3
Zangarmarsh	Terokkar Forest
1	1
Shattrath City	Hellfire Ramparts
3	2
Eversong Woods	Nagrand
4	1
Blade's Edge Mountains	Netherstorm
1	1
The Blood Furnace	Coilfang: The Slave Pens
2	2
The Steamvault	Eye of the Storm
2	3
Azuremyst Isle	Bloodmyst Isle
4	4
Mana-Tombs	Old Hillsbrad Foothills
2	2
Auchenai Crypts	Sethekk Halls
2	2
The Black Morass	The Shattered Halls
2	2
Auchindoun: Shadow Labyrinth	The Botanica
2	2
The Arcatraz	Gruul's Lair
2	2
Karazhan	Magtheridon's Lair
2	2
The Exodar	The Forbidding Sea
3	3
Ruins of Lordaeron	Serpentshrine Cavern
3	2
Tempest Keep	The Underbog
2	2
The Mechanar	Black Temple
2	2
Magisters' Terrace	Isle of Quel'Danas
2	4
Sunwell Plateau	The North Sea
2	3
The Ring of Valor	Dalaran<U+7AF6><U+6280><U+5834>
3	3
Plaguelands: The Scarlet Enclave	Borean Tundra
1	1
Dragonblight	Howling Fjord
1	1
The Nexus	Grizzly Hills
2	1
Sholazar Basin	Crystalsong Forest
1	1
The Frozen Sea	Utgarde Keep
3	2
Azjol-Nerub	The Storm Peaks
2	1
Icecrown	Dalaran

```
                                1                                3
      Ahn'kahet: The Old Kingdom                      Zul'Drak
                                2                                1
                The Violet Hold                     Wintergrasp
                                2                                1
          Strand of the Ancients                Drak'Tharon Keep
                                2                                2
                         Gundrak        The Culling of Stratholme
                                2                                2
                      The Oculus               Halls of Lightning
                                2                                2
                   Halls of Stone                 Utgarde Pinnacle
                                2                                2
               Vault of Archavon              The Eye of Eternity
                                2                                2
            The Obsidian Sanctum <U+9F8D><U+9AA8><U+8352><U+91CE>
                                2                                1
<U+5317><U+65B9><U+6D77><U+5CB8>
                                3

Within cluster sum of squares by cluster:
[1] 42.51555 12.94720 11.49509 37.11403
 (between_SS / total_SS =  86.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
>
> #correspondance avec les groupes de la CAH
> print(table(groupes.cah,groupes.kmeans$cluster))

groupes.cah  1  2  3  4
          1  0  0  0 28
          2  0  0 27  0
          3  0 41  2 12
          4 47  0  0  0
```

Le groupe 3 de la CAH coïncide avec le groupe 2 des K-Means. Il y a certes des correspondances, mais elles ne sont pas toujours exactes, c'est pourquoi l'on peut avoir des résultats différents entre la méthode CAH et la méthode K-Means.

À vous !

a) Cherchez la signification du paramètre *centers* de la fonction *kmeans()*.

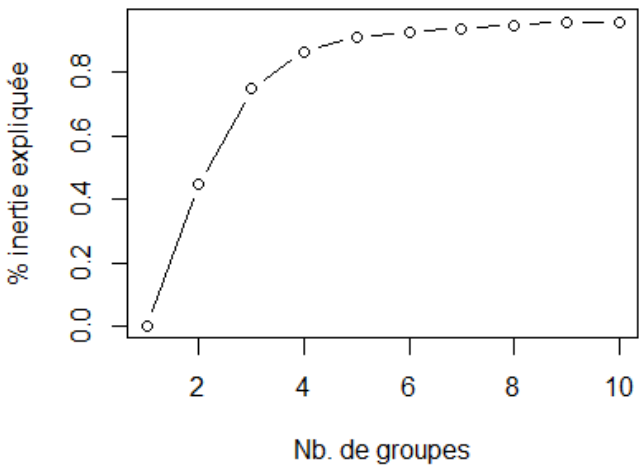
b) Cherchez la signification du paramètre *nstart* de la fonction *kmeans()*.

1.4 Détection des groupes pour le K-mean

On propose d'observer l'évolution de la proportion d'inertie pour différentes partitions, on cherche le point où la tangente devient horizontale dans le graphique (plus d'évolution).

```
> inertie.expl <- rep(0,times=10)
> for (k in 2:10){
  clus <- kmeans(royau.cr,centers=k,nstart=5)
  inertie.expl[k] <- clus$betweenss/clus$totss
}
> plot(1:10,inertie.expl,type="b",xlab="Nb. de groupes",ylab="%
   inertie expliquee")
```

FIGURE 4 – Pourcentage d'inertie selon les groupes



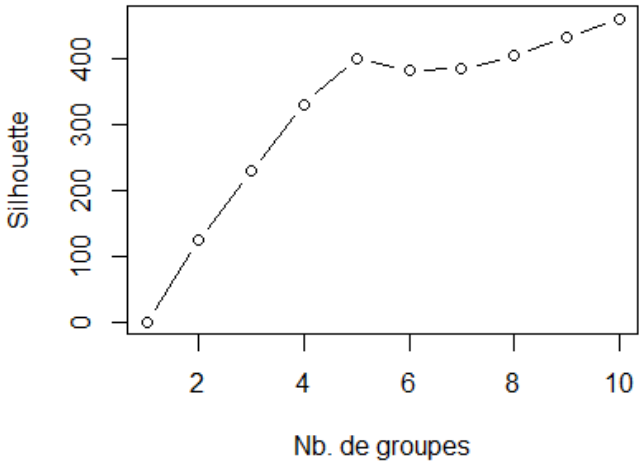
La largeur moyenne de la silhouette est la moyenne de s_i de tous les objets i dans les données, c'est à dire la largeur moyenne de la silhouette pour k classes. Ceci peut être utilisé pour sélectionner le "meilleur" nombre de classes, en choisissant le k qui donne la valeur la plus élevée de la moyenne de s_i.

À vous !

a) Utilisez la largeur moyenne de silhouette maximale avec la fonction *kmeansruns()* de la librairie *fpc*. Affichez ce second critère de façon graphique.

b) Interprétez ces résultats.

FIGURE 5 – Silhouette selon les groupes



1.5 Statistiques comparatives

Comparez les moyennes des variables actives conditionnellement aux groupes. Pour cela, quantifiez globalement l'amplitude des écarts avec la proportion de variance expliquée en créant la fonction *stat.com()* renvoyant :

- Le nombre de groupes.
- Le nombre d'observations.
- La moyenne globale.
- La variabilité totale.
- Les effectifs conditionnels.
- La moyennes conditionnelles.
- La variabilité expliquée.
- Les éléments du vecteur.

```
  > stat.comp <- function(x,y){
+   K <- length(unique(y))
+   n <- length(x)
+   m <- mean(x)
+   TSS <- sum((x-m)^2)
+   nk <- table(y)
+   mk <- tapply(x,y,mean)
+   BSS <- sum(nk * (mk - m)^2)
+   result <- c(mk,100.0*BSS/TSS)
+   names(result) <- c(paste("G",1:K),"% epl.")
+   return(result)
+ }
```

À vous !

a) Appliquez la fonction *stat.comp()* aux variables de la base originale royau.x et non pas aux variables centrées réduites.

b) Interprétez ces résultats.

1.6 ACP et CAH

Complétez l'analyse CAH avec l'ACP pour tenir compte des liaisons entre les variables. Affichez sur le graphique les groupes trouvés grâce à la classification ascendante hiérarchique précédente.

```
> acp <- princomp(royau.x,cor=T,scores=T)
> plot(1:5,acp$sdev^2,type="b",xlab="Nb. de facteurs",ylab="Val.
   Propres")
> biplot(acp,cex=0.65)
> plot(acp$scores[,1],acp$scores[,2],type="n",xlim=c(-5,5),ylim=c
   (-5,5))
> text(acp$scores[,1],acp$scores[,2],col=c("red","green","blue","
   black")[groupes.cah],cex=0.65,labels=rownames(royau.x),xlim=c
   (-5,5),ylim=c(-5,5))
```

FIGURE 6 – Valeurs propres selon le nombre de facteurs

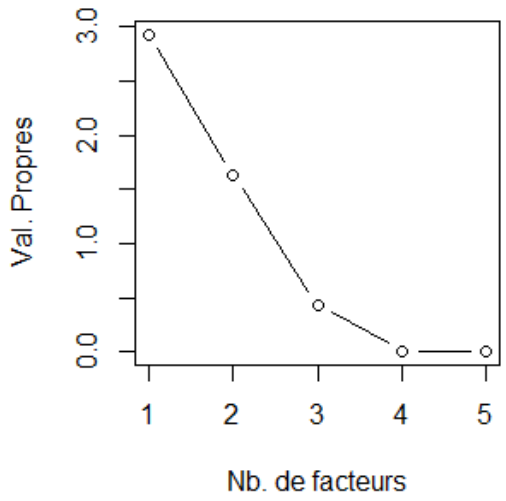


FIGURE 7 – Analyse en composantes principales

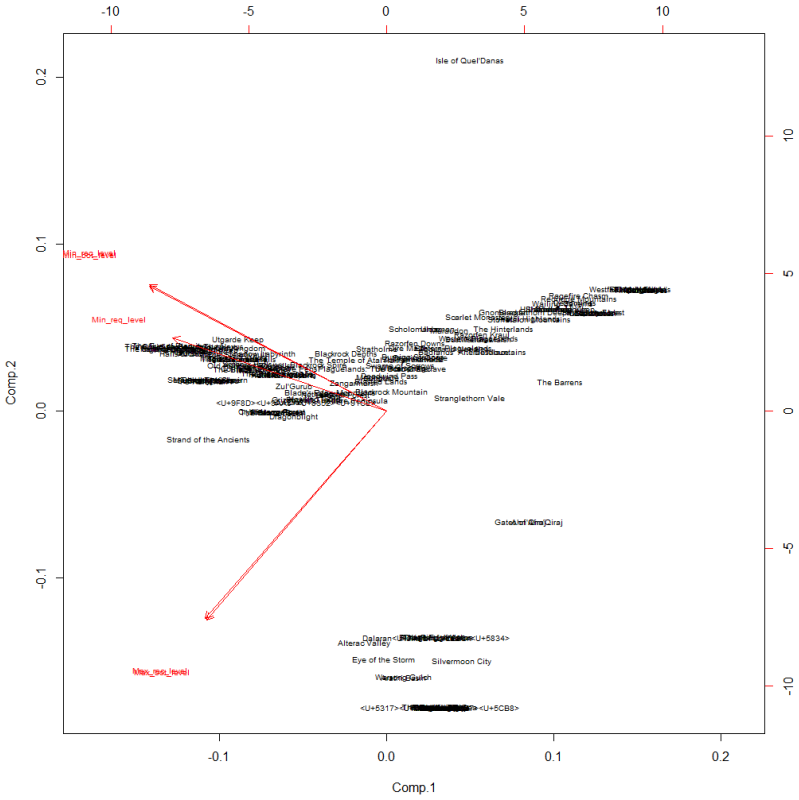
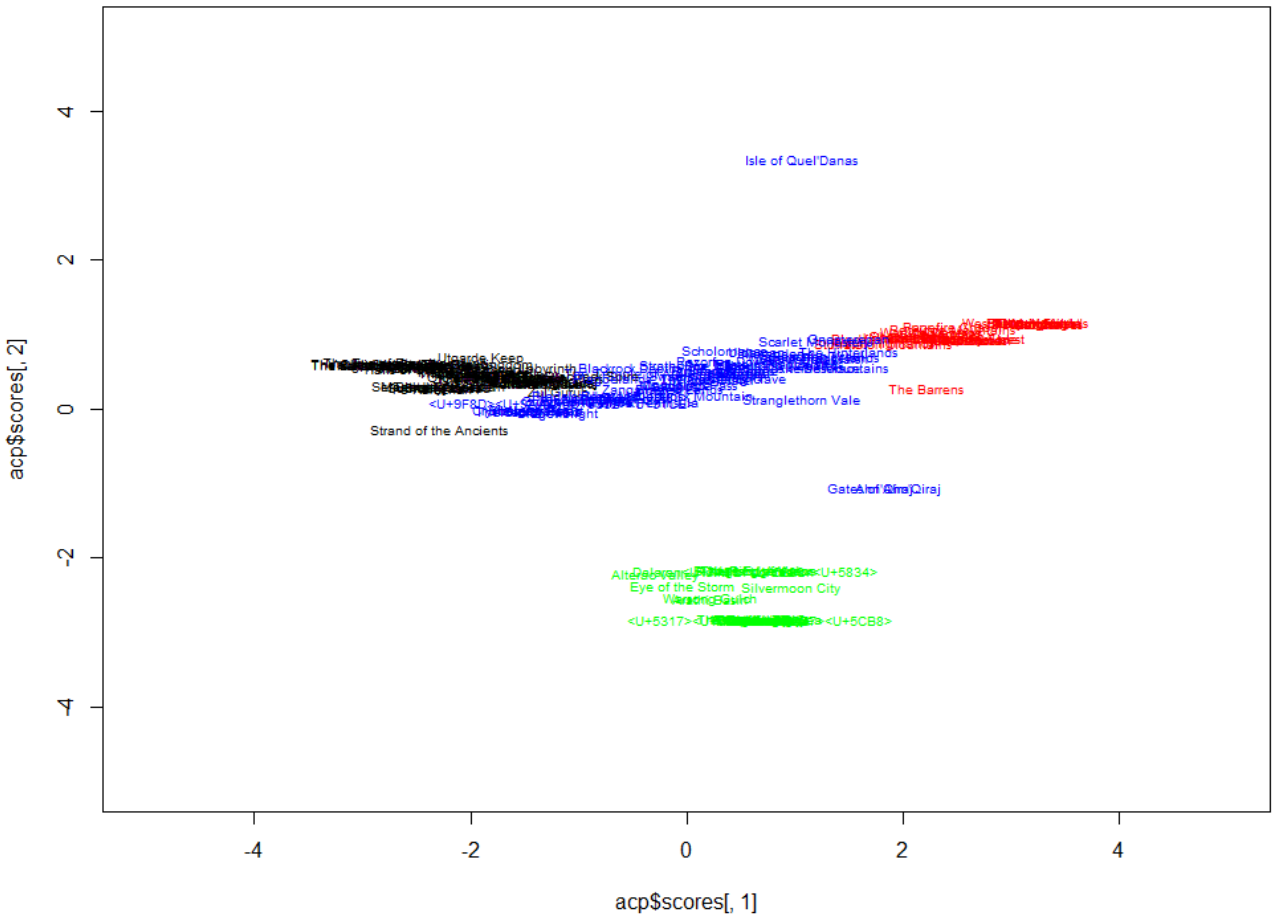


FIGURE 8 – Groupe de la CAH sur l'ACP



À vous !

a) Que pouvez-vous dire sur la ville *d'Isle of Quel'Danas* ?

b) Concluez sur l'interprétation des données.