Rapport Analyse de Données

Gwendoline Delestre

2022-11-30

Université de Grenoble

Table of Contents

Chapitre 0 4
Ex 13 4
Ex 14 6
Ex 15 6
Ex 16 9
Ex 17 11
Ex 18 12
Chapitre 1 14
Ex 19 14
Ex 20 15
Ex 21 17
Ex 22 26
Chapitre 2 52
Ex 23 52
Ex 24 55
Chapitre 3 65
Ex 31 65
Ex 32 65
Ex 33 82
Ex 34 86
Chapitre 4 91
Ex 27 91
Ex 28 104
Chapitre 5 105
CAH 105
K-means 107
Chapitre 6 : données personnelles 111
Les données : 111
Graphique du tableau de contingence 114
Calcul de l'AFC. 115
Significativité statistique 116

Valeurs propres/Variances 116
Biplot 118
Graphique des points lignes 119
Graphes des colonnes 129
Biplot symétrique. 132
Description des dimensions 135
Classification 137
CAH 137

Chapitre 0

Ex 13

Voici un aperçu des données utilisées pour cet exercice :

```
##   ID AGE LWT RACE SMOKE PTL HT UI FVT  BWT LOW
## 1 85  19 182    2     0   0  0  1   0 2523   0
## 2 86  33 155    3     0   0  0  0   3 2551   0
## 3 87  20 105    1     1   0  0  0   1 2557   0
## 4 88  21 108    1     1   0  0  1   2 2594   0
```

```
## 5 89  18 107    1     1   0  0  1   0 2600   0
## 6 91  21 124    3     0   0  0  0   0 2622   0
```

On convertie le poids de la mère en kilogrammes ce qui donne le tableau de données suivant :

```
##   ID AGE      LWT RACE SMOKE PTL HT UI FVT  BWT LOW
## 1 85  19 82.55381    2     0   0  0  1   0 2523   0
## 2 86  33 70.30682    3     0   0  0  0   3 2551   0
## 3 87  20 47.62720    1     1   0  0  0   1 2557   0
## 4 88  21 48.98798    1     1   0  0  1   2 2594   0
## 5 89  18 48.53438    1     1   0  0  1   0 2600   0
## 6 91  21 56.24545    3     0   0  0  0   0 2622   0
```

Voici le tri à plat de chacune des variables :

La distribution des âges :

```
## AGE
## 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 45
##  3  3  7 12 10 16 18 12 13 13 13 15  8  3  9  7  7  5  6  3  1  2  2  1
```

La distribution des poids de la mère en kilogrammes (on ne présente que les 6 premiers poids ici pour gagner de la place) :

```
## LWT
## 36.29 38.56 40.37 40.82 41.28 41.73
##     1     2     1     3     1     1
```

La distribution des races, 1 corresponds à Blanche, 2 corresponds à Noire, 3 correspond à Autre:

```
## RACE
##  1  2  3
## 96 26 67
```

La distribution du tabagisme, 1 correspond à fumeuse, 0 à non fumeuse :

```
## SMOKE
##   0   1
## 115  74
```

La distribution du nombres d'antécédents de prématurité©:

```
## PTL
##   0   1   2   3
## 159  24   5   1
```

La distribution des antécédents d'hypertension, 1 pour oui, 0 pour non:

```
## HT
##   0   1
## 177  12
```

La distribution de la présence d'irritabilité utérine, 1 pour oui, 0 pour non:

```
## UI
##   0   1
## 161  28
```

La distribution du nombre de visites chez le médecin durant le premier trimestre:

```
## FVT
##   0   1   2   3   4   6
## 100  47  30   7   4   1
```

La distribution du poids de naissance en grammes (on ne présente que les 6 premières lignes ici pour gagner de la place):

```
## BWT
##  709 1021 1135 1330 1474 1588
##    1    1    1    1    1    2
```

La distribution du poids de naissance inférieur ou égal a 2500g, 1 pour oui, 0 pour non:

```
## LOW
##   0   1
## 130  59
```

Ex 14

Voici les données de cet exercice :

```
##   Mort.à Année.de.carrière Nombre.de.film  Prénom       Nom Date.du.décès
## 1     93                66            221  Michel   Galabru    04-01-2016
## 2     53                25             58  André  Raimbourg    23-09-1970
## 3     72                48             98    Jean     Gabin    15-10-1976
## 4     68                37            140   Louis De Funès     27-01-1983
## 5     68                31             74    Lino   Ventura    22-10-1987
## 6     53                32             81 Jacques  Villeret    28-01-2005
```

Après avoir changé le nom de la première colonne, voici la colonne des prénoms et la table des données complètes ordonnées par âge de décès.

```
## [1] "Michel"  "André"  "Jean"    "Louis"   "Lino"    "Jacques"

##   Age.du.décès Année.de.carrière Nombre.de.film  Prénom       Nom Date.du.décès
## 2           53                25             58  André  Raimbourg    23-09-1970
## 6           53                32             81 Jacques  Villeret    28-01-2005
## 4           68                37            140   Louis De Funès     27-01-1983
## 5           68                31             74    Lino   Ventura    22-10-1987
## 3           72                48             98    Jean     Gabin    15-10-1976
## 1           93                66            221  Michel   Galabru    04-01-2016
```

Ex 15

Voici les données de cet exercice.

```
##      Y    X1    X2   X3
## 1 12.3 4.543 3.135 0.86
## 2 20.9 5.159 5.043 1.53
## 3 39.0 5.366 5.438 1.57
## 4 47.9 5.759 7.496 1.81
## 5  5.6 4.663 3.807 0.99
## 6 25.9 5.697 7.601 1.09
```

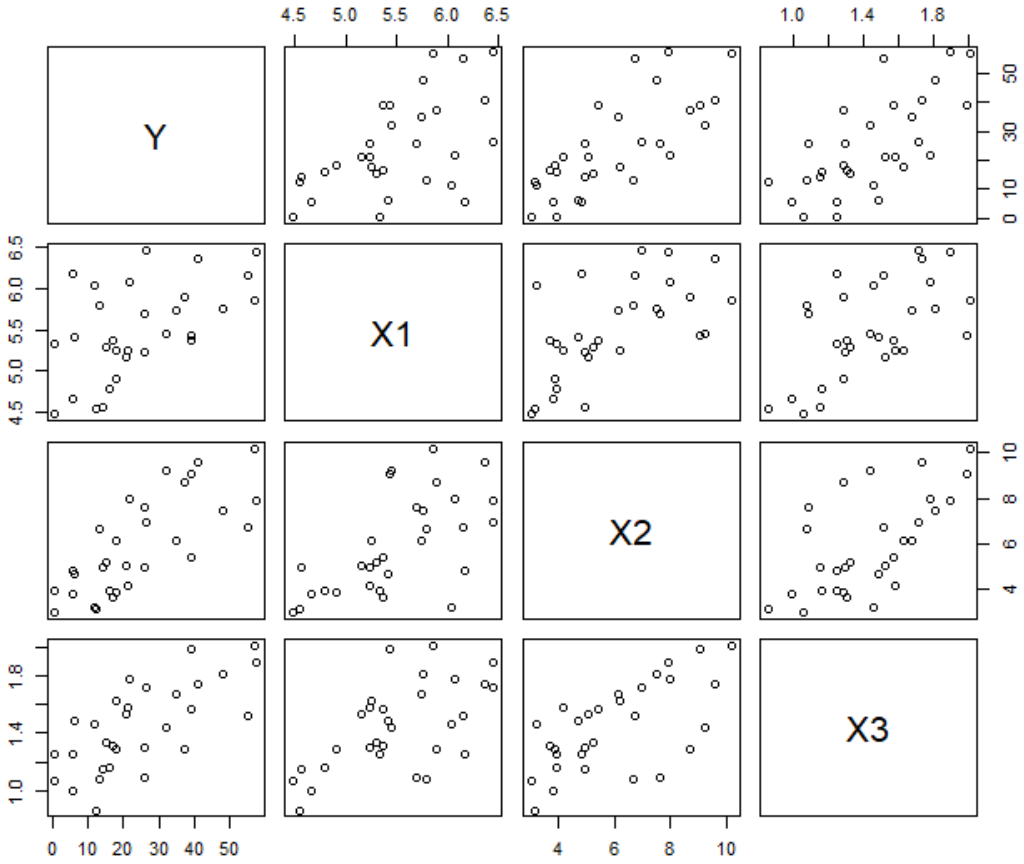
Voici la concentration d'acide acétique :

```
##  [1] 4.543 5.159 5.366 5.759 4.663 5.697 5.892 6.078 4.898 5.242 5.740
6.446
## [13] 4.477 5.236 6.151 6.365 4.787 5.412 5.247 5.438 4.564 5.298 5.455
5.855
## [25] 5.366 6.043 6.458 5.328 5.802 6.176
```

Les caractéristiques des données sont les suivantes ainsi que les paramètres élémentaires :

```
## 'data.frame':    30 obs. of  4 variables:
##  $ Y : num  12.3 20.9 39 47.9 5.6 25.9 37.3 21.9 18.1 21 ...
##  $ X1: num  4.54 5.16 5.37 5.76 4.66 ...
##  $ X2: num  3.13 5.04 5.44 7.5 3.81 ...
##  $ X3: num  0.86 1.53 1.57 1.81 0.99 1.09 1.29 1.78 1.29 1.58 ...

##        Y               X1              X2               X3
##  Min.   : 0.70   Min.   :4.477   Min.   : 2.996   Min.   :0.860
##  1st Qu.:13.55   1st Qu.:5.237   1st Qu.: 3.978   1st Qu.:1.250
##  Median :20.95   Median :5.425   Median : 5.329   Median :1.450
##  Mean   :24.53   Mean   :5.498   Mean   : 5.942   Mean   :1.442
##  3rd Qu.:36.70   3rd Qu.:5.883   3rd Qu.: 7.575   3rd Qu.:1.667
##  Max.   :57.20   Max.   :6.458   Max.   :10.199   Max.   :2.010
```

Sur les graphiques suivants, on peut voir la disposition des individus en fonction de deux variables et ce pour chacune des variables.



Ici, on reproduit les mêmes calculs mais on ne garde que les fromages ayant une concentration d'acide acétique supérieur à 5.1 et une concentration d'acide lactique inférieur à 1.77:

```
##       Y    X1    X2   X3
## 2  20.9 5.159 5.043 1.53
## 3  39.0 5.366 5.438 1.57
## 6  25.9 5.697 7.601 1.09
## 7  37.3 5.892 8.726 1.29
## 10 21.0 5.242 4.174 1.58
## 11 34.9 5.740 6.142 1.68

## 'data.frame':    19 obs. of  4 variables:
##  $ Y : num  20.9 39 25.9 37.3 21 34.9 25.9 54.9 40.9 6.4 ...
##  $ X1: num  5.16 5.37 5.7 5.89 5.24 ...
##  $ X2: num  5.04 5.44 7.6 8.73 4.17 ...
##  $ X3: num  1.53 1.57 1.09 1.29 1.58 1.68 1.3 1.52 1.74 1.49 ...

##        Y               X1              X2              X3
##  Min.   : 0.70   Min.   :5.159   Min.   :3.219   Min.   :1.080
```

```
##  1st Qu.:14.30   1st Qu.:5.313   1st Qu.:4.744   1st Qu.:1.295
##  Median :21.00   Median :5.455   Median :5.438   Median :1.460
##  Mean   :23.52   Mean   :5.654   Mean   :5.946   Mean   :1.435
##  3rd Qu.:33.45   3rd Qu.:5.968   3rd Qu.:6.857   3rd Qu.:1.575
##  Max.   :54.90   Max.   :6.458   Max.   :9.588   Max.   :1.740
```

Ex 16

Les données de cet exercice représentent la mesure de la qualité de l'air sur 5 mois. Il y a 6 colonnes et 153 lignes.

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

Les noms des variables considérées sont les suivantes, nous avons, 153 lignes (individus) et 6 colonnes (variables):

```
## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"

## [1] 153

## [1] 6
```

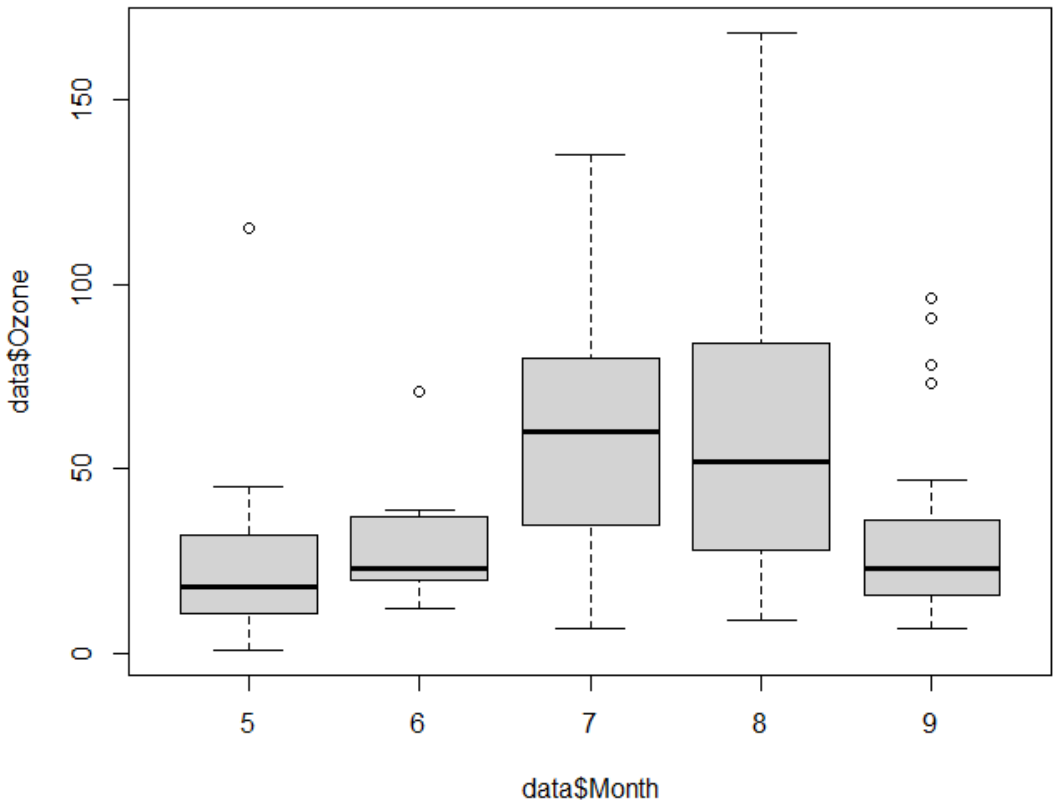
Voici les paramètres statistiques de bases de ces données:

La médiane de la totalité des données est égale à 31.5 pour l'Ozone et la moyenne est de 42.13.

```
##      Ozone           Solar.R          Wind             Temp
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
##  NA's   :37       NA's   :7
##      Month            Day
##  Min.   :5.000   Min.   : 1.0
##  1st Qu.:6.000   1st Qu.: 8.0
##  Median :7.000   Median :16.0
##  Mean   :6.993   Mean   :15.8
##  3rd Qu.:8.000   3rd Qu.:23.0
##  Max.   :9.000   Max.   :31.0
##
```

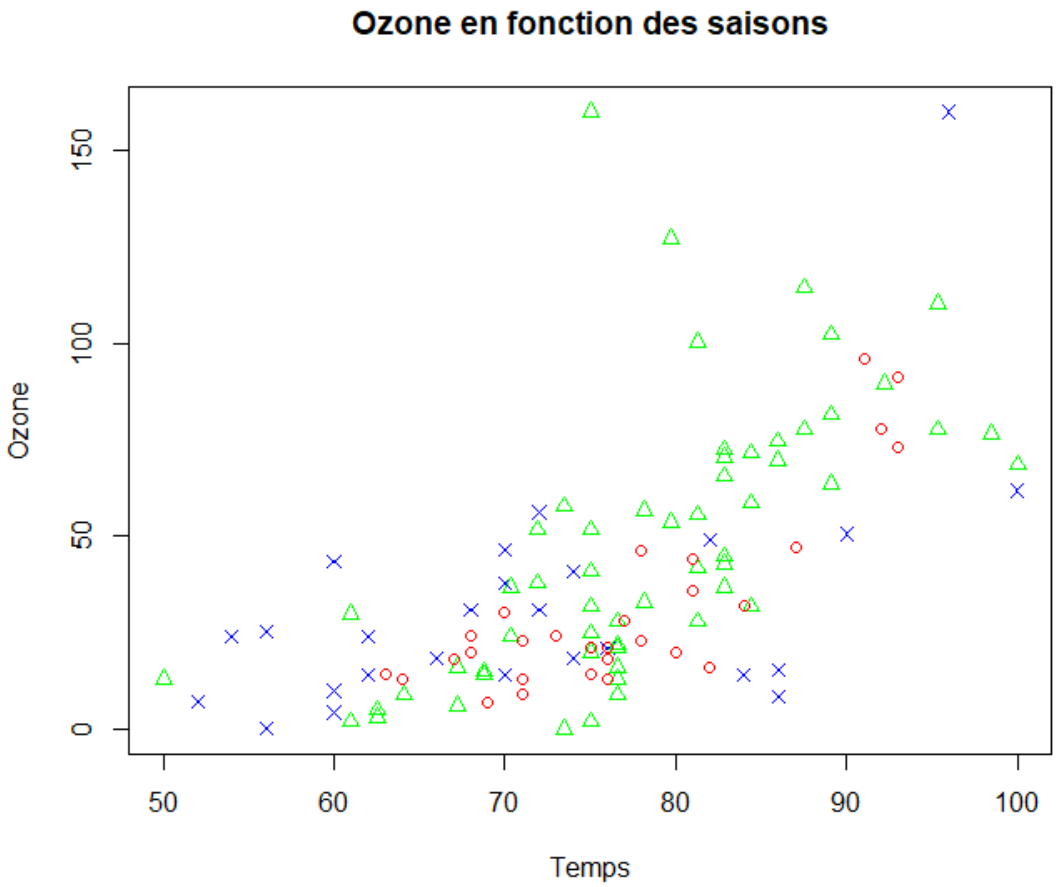
D'après ce graphique en boîte à moustache, on voit que les mois de juillet et août ont une plus forte concentration d'Ozone et aussi une plus grande amplitude de valeurs. La médiane

de ces deux mois se situe au dessus de 50. La médiane la plus faible se trouve au mois de mai. Il existe des valeurs aberrantes pendant les mois de mai, juin et septembre.



Ce graphique représente l'Ozone en fonction des saison, triangle pour été, rond pour l'automne et les croix pour le printemps. Pour ce graphique nous avons ajouté une colonne "Saison".

```
##   Ozone Solar.R Wind Temp Month Day    saison
## 1    41     190  7.4   67     5   1 printemps
## 2    36     118  8.0   72     5   2 printemps
## 3    12     149 12.6   74     5   3 printemps
## 4    18     313 11.5   62     5   4 printemps
## 5    NA      NA 14.3   56     5   5 printemps
## 6    28      NA 14.9   66     5   6 printemps
```



Ex 17

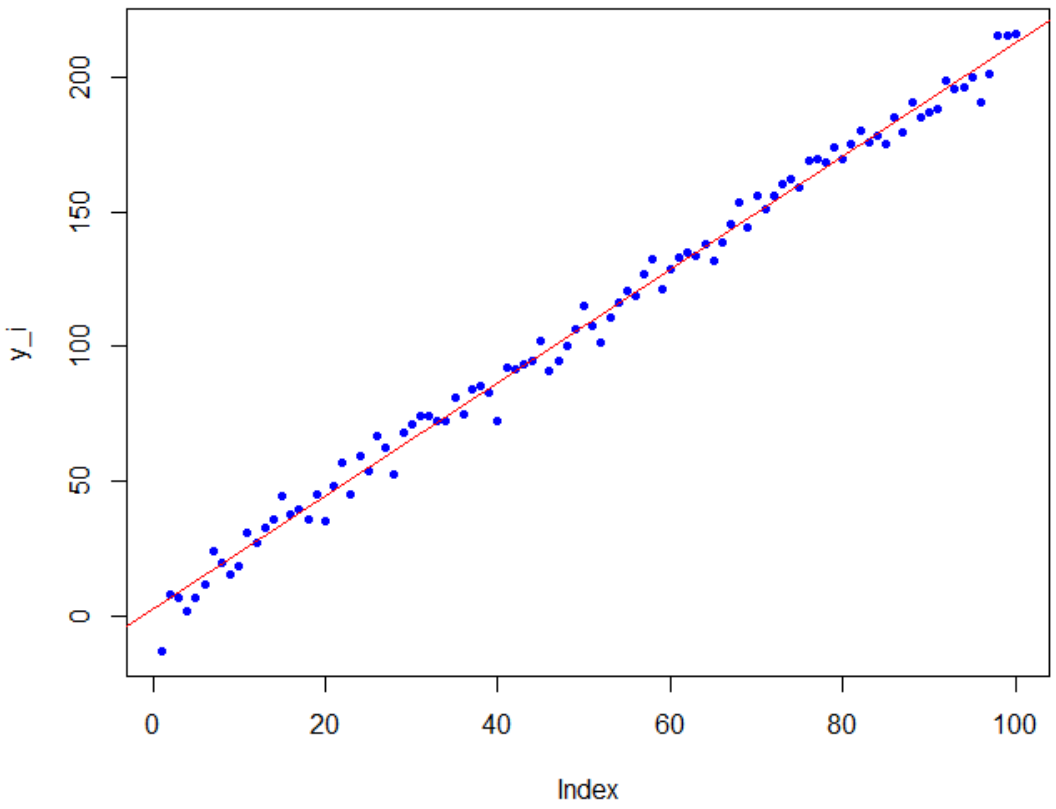
Voici un extrait des valeurs suivants la loi normale de moyenne 0 et de variances 25.

```
## [1] -17.012429   2.041144  -1.632296  -8.618130  -5.960283  -2.614278
```

y_i est donc le tableau qui suit:

```
## [1] -13.212429   7.941144   6.367704   1.481870   6.239717  11.685722
```

Voici le nuage de point représentant i en fonction de y_i :



Ex 18

Voici les données de cet exercice :

```
##          brun chatin roux blond
## marron     68    119   26     7
## noisette   15     54   14    10
## vert        5     29   14    16
## bleu       20     84   17    94
```

Nous pouvons ainsi obtenir la matrice des fréquences si dessous :

```
##          brun chatin roux blond
## marron   0.11   0.20 0.04  0.01
## noisette 0.03   0.09 0.02  0.02
## vert     0.01   0.05 0.02  0.03
## bleu     0.03   0.14 0.03  0.16
```

Nous obtenons les valeurs marginales pour la couleur des cheveux puis la couleur des yeux.

```
##   brun chatin   roux  blond
##   0.18   0.48   0.11   0.22

##   marron noisette     vert     bleu
##     0.36     0.16     0.11     0.36
```

Ici les profiles lignes :

```
##          brun chatin roux blond
## marron     68    119   26     7
## noisette   15     54   14    10
## vert        5     29   14    16
## bleu       20     84   17    94
```

Et les profils colonnes:

```
##                brun    chatin      roux      blond
## marron   0.6296296 0.4160839 0.3661972 0.05511811
## noisette 0.1388889 0.1888112 0.1971831 0.07874016
## vert     0.0462963 0.1013986 0.1971831 0.12598425
## bleu     0.1851852 0.2937063 0.2394366 0.74015748
```

La distance du khi deux entre les profiles lignes : la p-value est de 0.9996 qui est supérieur à 0.05. On peut donc supposer l'indépendance des données.

```
## Warning in chisq.test(profL): Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  profL
## X-squared = 0.78907, df = 9, p-value = 0.9998
```

Chapitre 1

Ex 19

Voici le tableau des données :

```
##                     BEPC BAC Licence Total
## Plus de 50 ans        15  12       3    30
## Entre 30 et 50 ans    10  18       4    32
## Moins de 30 ans       15   5       8    28
## Total                 40  35      15    90
```

Voici le tableau des fréquences correspondant aux données présentées au dessus:

```
##                     BEPC  BAC Licence Total
## Plus de 50 ans      0.17 0.13    0.03  0.33
## Entre 30 et 50 ans 0.11 0.20    0.04  0.36
## Moins de 30 ans    0.17 0.06    0.09  0.31
## Total               0.44 0.39    0.17  1.00
```

Voici la matrice des profils Lignes et Colonnes :

```
##                     BEPC BAC Licence Total
## Plus de 50 ans        15  12       3    30
## Entre 30 et 50 ans    10  18       4    32
## Moins de 30 ans       15   5       8    28
## Total                 40  35      15    90

##                         BEPC       BAC   Licence Total
## Plus de 50 ans     0.5000000 0.4000000 0.1000000     1
## Entre 30 et 50 ans 0.3125000 0.5625000 0.1250000     1
## Moins de 30 ans    0.5357143 0.1785714 0.2857143     1
## Total               0.4444444 0.3888889 0.1666667     1

##                      BEPC       BAC   Licence Total
## Plus de 50 ans     0.375 0.3428571 0.2000000    30
## Entre 30 et 50 ans 0.250 0.5142857 0.2666667    32
## Moins de 30 ans    0.375 0.1428571 0.5333333    28
## Total               1.000 1.0000000 1.0000000    90
```

On ne rejette pas H0 car la p-value est supérieur au seuil de 0.05. Il y a donc il y a indépendance des données.

```
## Warning in chisq.test(sweepTab): Chi-squared approximation may be
incorrect

##
##  Pearson's Chi-squared test
##
## data:  sweepTab
## X-squared = 11.175, df = 9, p-value = 0.2639
```

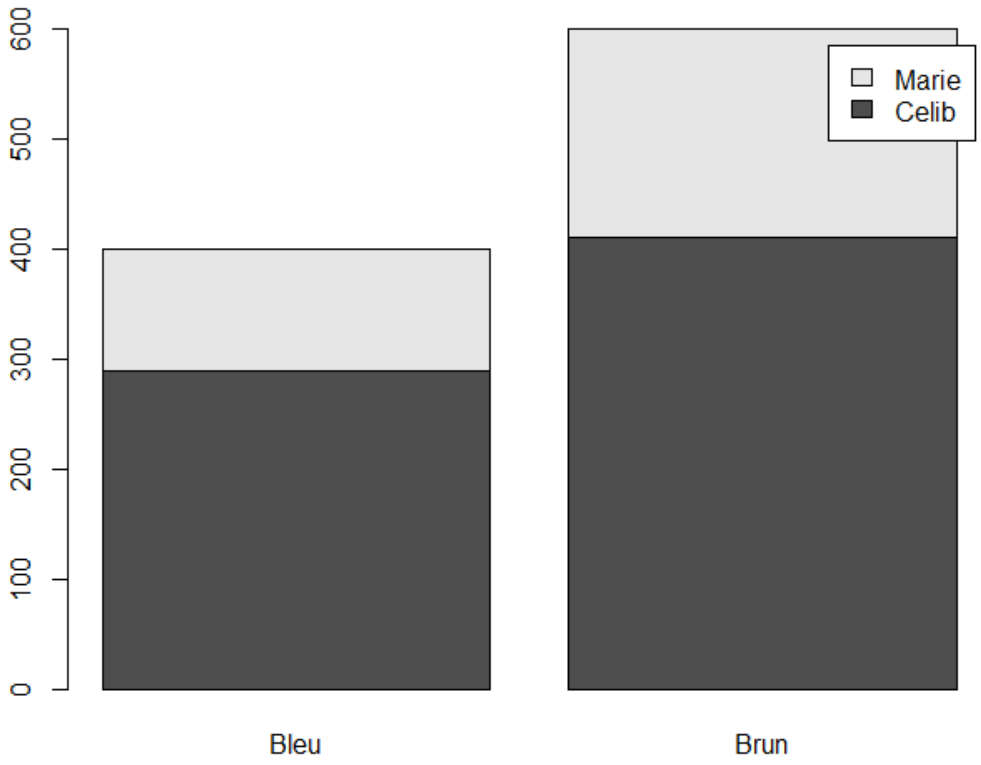
```
## Warning in chisq.test(profColonnes): Chi-squared approximation may be
incorrect

##
##  Pearson's Chi-squared test
##
## data:  profColonnes
## X-squared = 0.44457, df = 9, p-value = 1
```

Ex 20

Voici le tableau de données :

```
##       Bleu Brun
## Celib  290  410
## Marie  110  190
```

Voici un barplot montrant le nombre de personne marié ou célibataire ayant les yeux bleus et de la même façon ceux qui ont les yeux marrons.



Voici l'effectif total :

```
## [1] 1000
```

Voici l'effectif de chaque modalité pour la variables "statue Matrimonial"

```
## Celib Marie
##   700   300
```

Voici l'effectif de chaque modalité pour la variable "couleur yeux"

```
## Bleu Brun
##  400  600
```

Tableau des fréquences:

```
##       Bleu Brun
## Celib 0.29 0.41
## Marie 0.11 0.19
```

Voici les valeurs théoriques obtenues à partir des valeurs observées :

```
##       Bleu Brun
## Celib  280  420
## Marie  120  180
```

Voici le test du khi-deux, on remarque que la p-value est supérieur à 0.05 ce qui signifie que les deux variables sont indépendantes.

```
## Number of cases in table: 1000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1.9841, df = 1, p-value = 0.159

## Number of cases in table: 1000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1.154e-29, df = 1, p-value = 1
```

La valeur de la p-value est bien inférieur à 0.05 donc on peut dire que les deux variables sont non indépendantes lorsque l'on choisit de dire que les personnes aux yeux bleus sont tous mariés.

```
## Number of cases in table: 1000
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1000, df = 1, p-value = 1.796e-219
```

Dans le cas du jeu de données "HairEyeColor", la p-value du test du khi deux nous dit que les variables sont non indépendantes.

```
## Number of cases in table: 592
## Number of factors: 3
```

```
## Test for independence of all factors:
##  Chisq = 164.92, df = 24, p-value = 5.321e-23
##  Chi-squared approximation may be incorrect
```

Dans le cas du jeu de données “Titanic”, la p-value du test du khi deux nous dit que les variables sont non indépendantes.

```
## Number of cases in table: 2201
## Number of factors: 4
## Test for independence of all factors:
##  Chisq = 1637.4, df = 25, p-value = 0
##  Chi-squared approximation may be incorrect
```

Dans le cas du jeu de données “Titanic”, la p-value du test du khi deux nous dit que les variables sont non indépendantes.

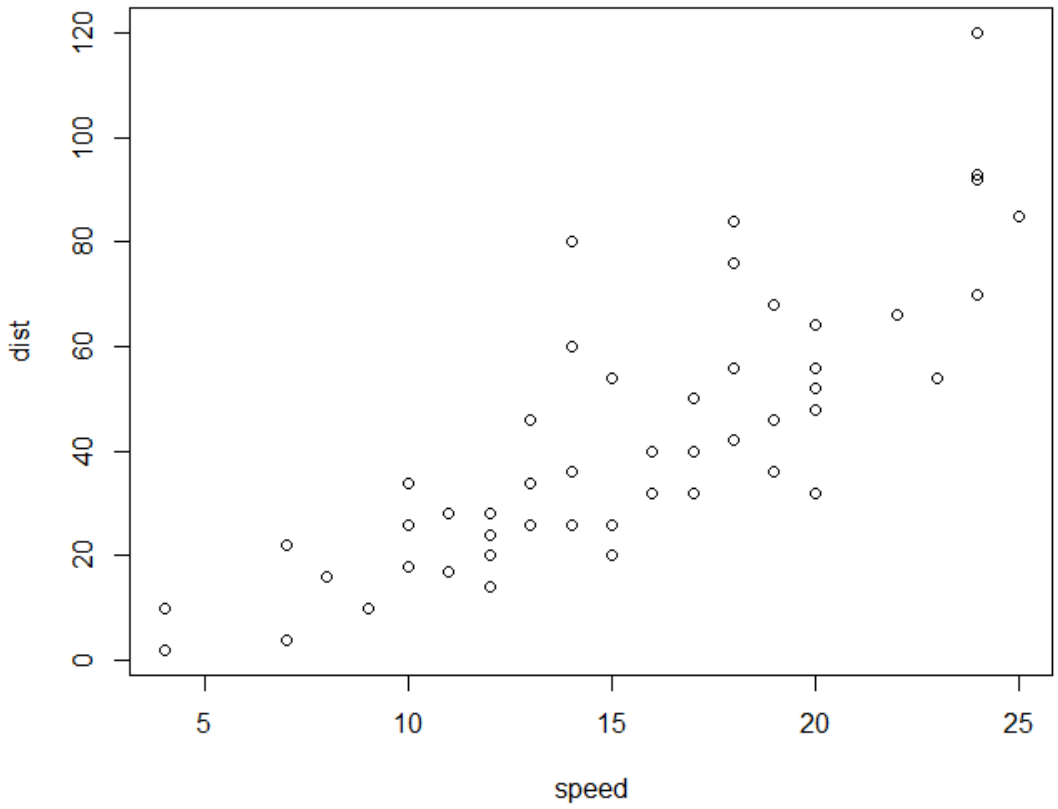
```
## Number of cases in table: 4526
## Number of factors: 3
## Test for independence of all factors:
##  Chisq = 2000.3, df = 16, p-value = 0
```

Ex 21

Voici les donnes de cet exercice : la variable “speed” représente la vitesse en miles par heure (mph) et la variable “dist” représente la distance d’arrêt en pied (ft). La matrice de données fait 50 lignes et 2 colonnes

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

Cette représentation graphique semble adaptée, on peut remarquer que plus la vitesse augmente plus la distance augmente.



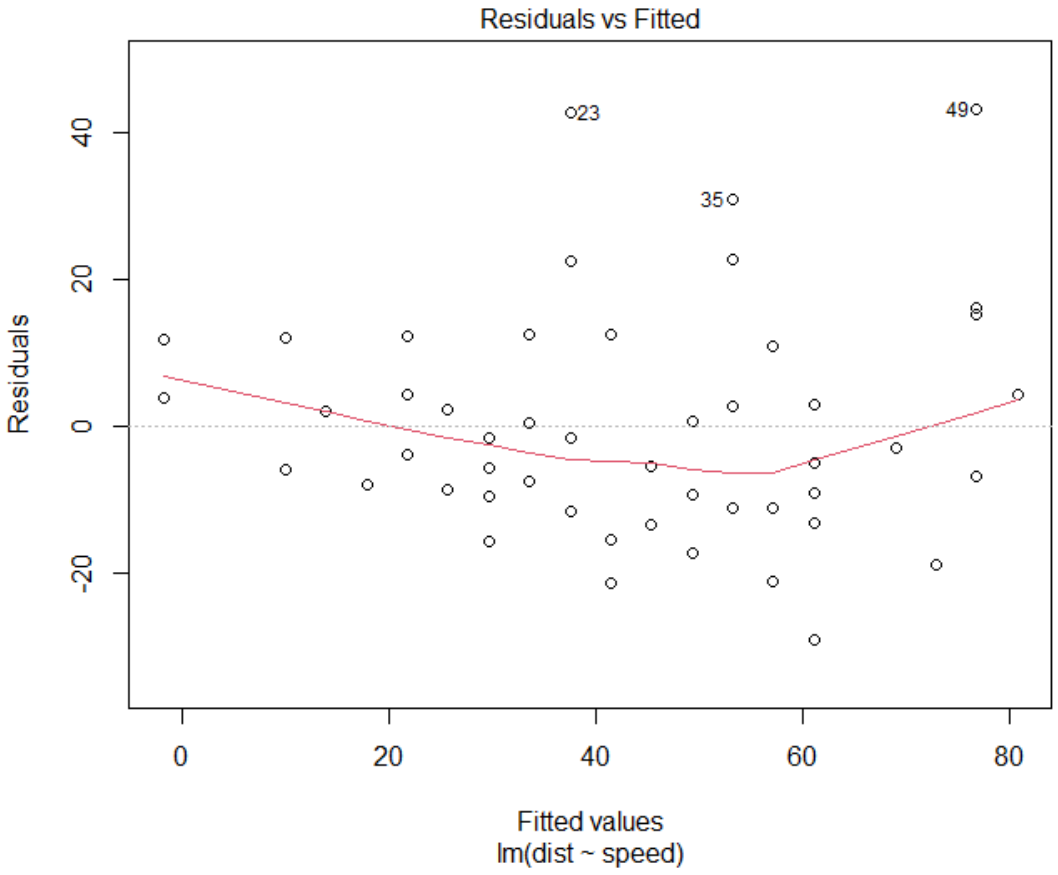
Les deux tableaux suivants représentent le résumé des caractéristiques de la variable que nous avons appelé reg qui suit un model linéaire et l'anova de reg. On voit que le résumé est plus détaillé que l'anova mais ils ne comportent pas les même informations à part la F value.

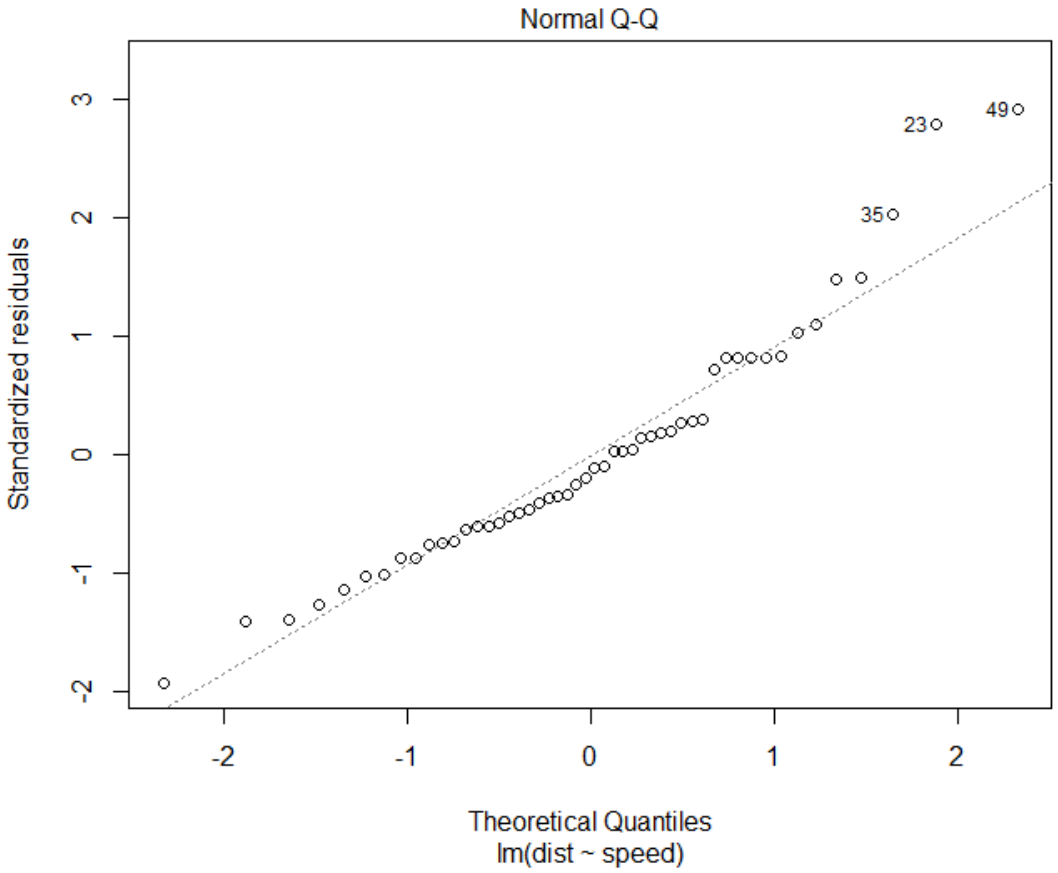
```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

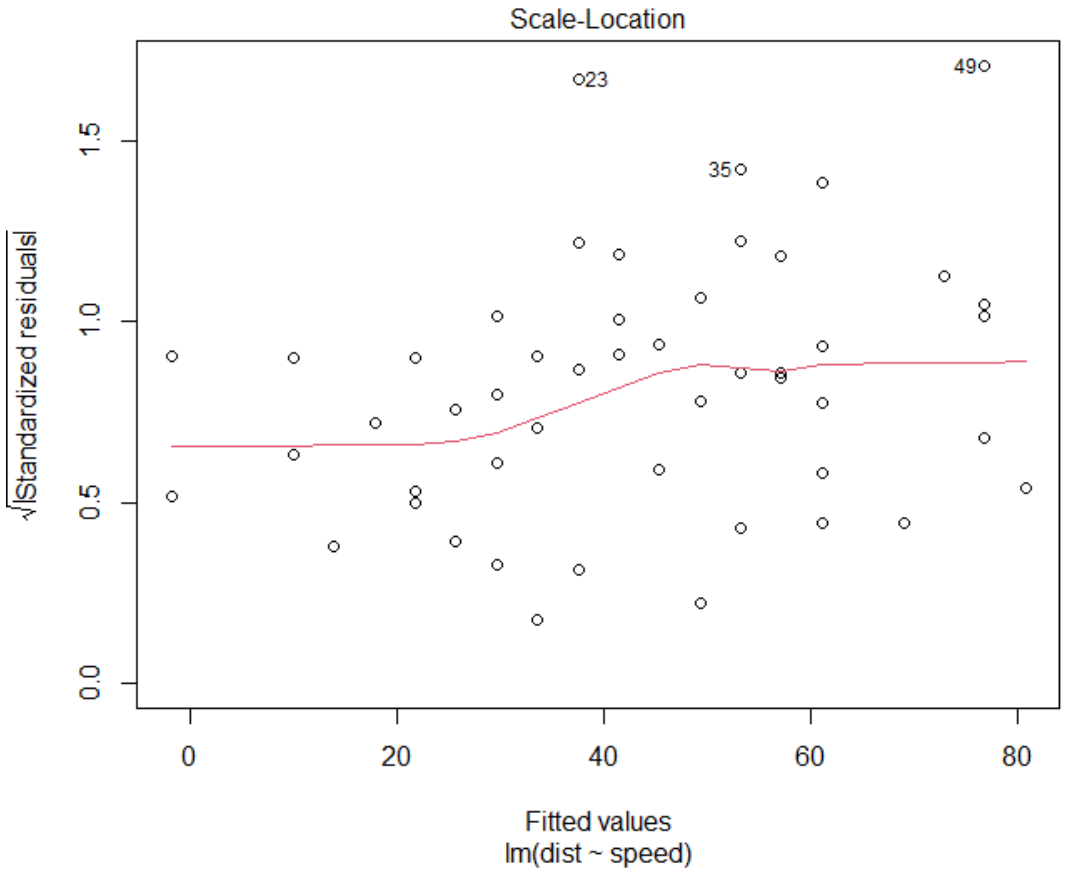
```
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12

## Analysis of Variance Table
##
## Response: dist
##           Df Sum Sq Mean Sq F value   Pr(>F)
## speed      1  21186 21185.5  89.567 1.49e-12 ***
## Residuals 48  11354   236.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

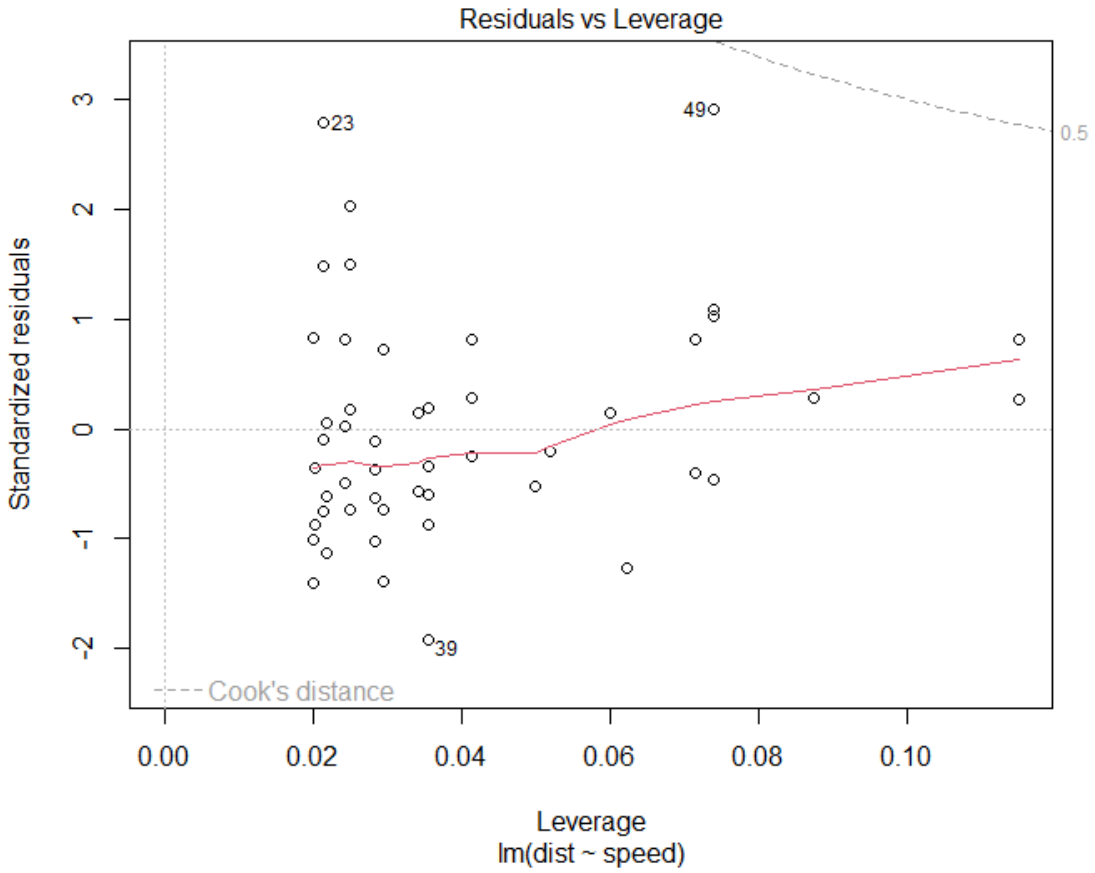
Voici le sparamètres utilisables de reg :

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

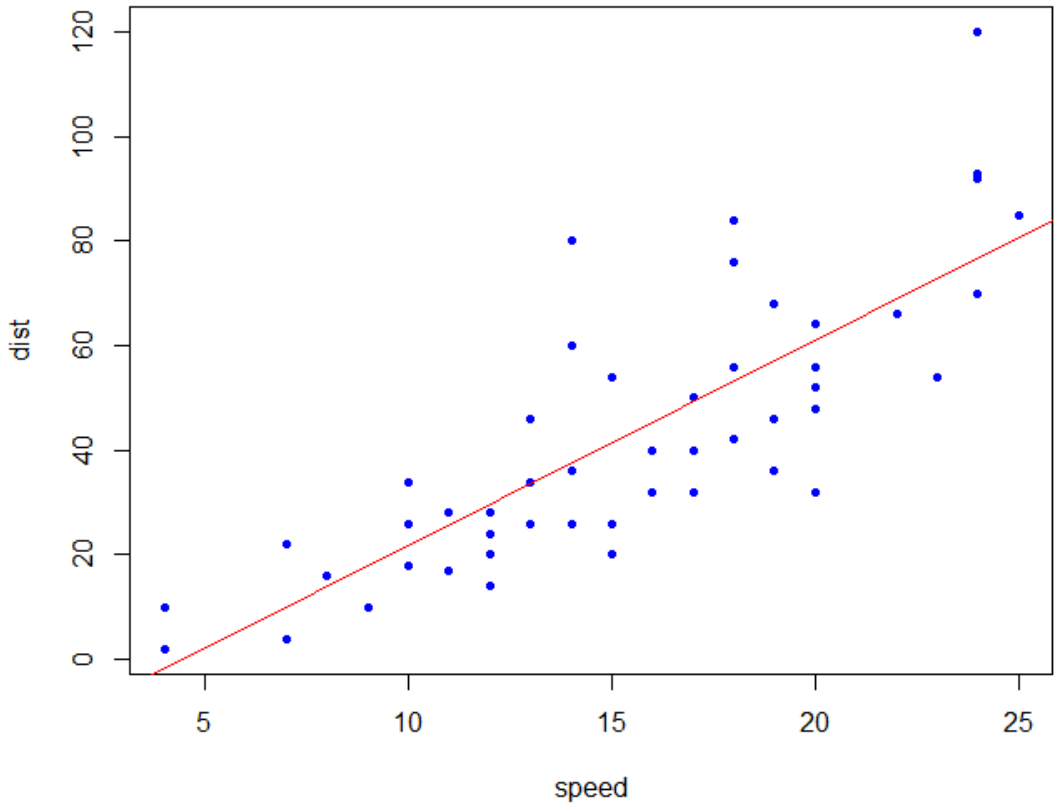

Residuals vs Fitted
23
49
35
Residuals
40
20
0
-20
0
20
40
60
80
Fitted values
lm(dist ~ speed)

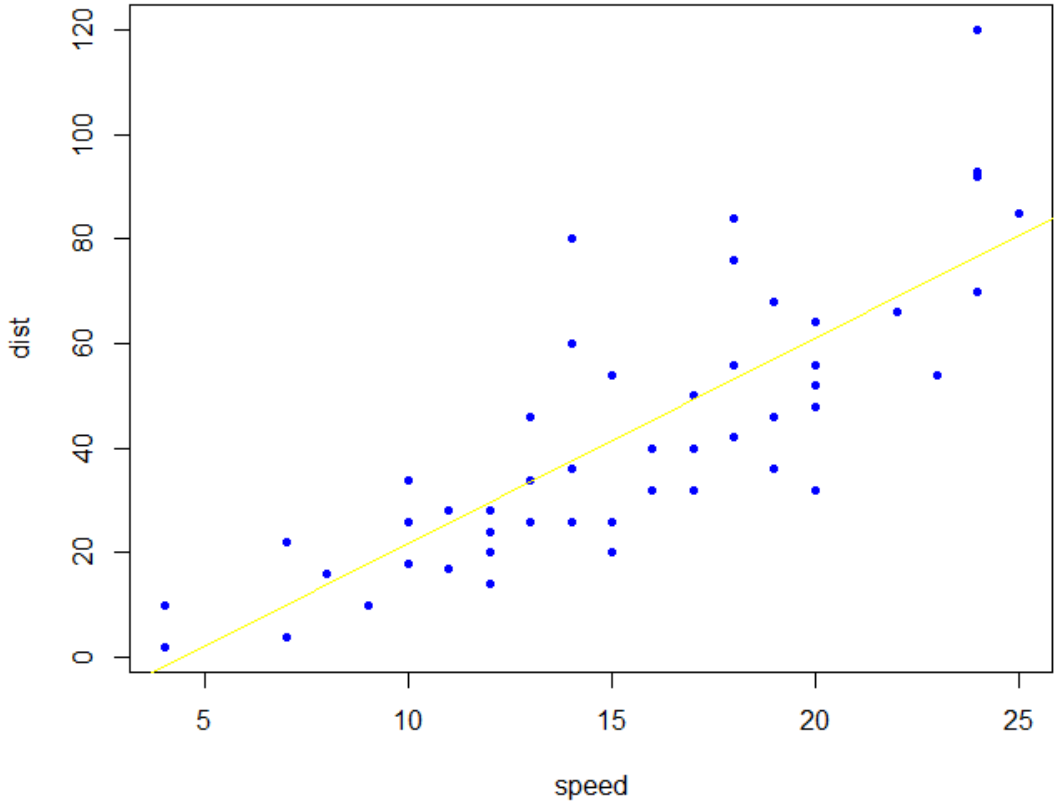

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(dist ~ speed)
49
23
35
-2
-1
0
1
2
3


Scale-Location
23
49
35
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
20
40
60
80
Fitted values
lm(dist ~ speed)


Residuals vs Leverage
Standardized residuals
Leverage
lm(dist ~ speed)
Cook's distance
0.5
23
49
39
-2
-1
0
1
2
3
0.00
0.02
0.04
0.06
0.08
0.10

Voici le graphique de nos données avec la droite représentant la modélisation linéaire. On peut remarquer qu'on obtient le même graphique que l'on fasse le plot de reg ou de reg$coeff





En voulant prédire la distance d'arrêt pour une vitesse de 20mph on peut prédire que l'on parcourra 61.07 ft qui est dans l'intervalle de confiance entre 55.25 pieds et 66.89 pieds, pour l'intervalle de prédiction est plus large.

```
##        1
## 61.06908

##        fit      lwr      upr
## 1 61.06908 55.24729 66.89088

##        fit      lwr      upr
## 1 61.06908 29.60309 92.53507
```

L'exemple de cars est adapté à la sélection de modèle et particulièrement de modèle linéaire. On le sait grâce à la corrélation qui est de 80%.

```
##           speed      dist
## speed 1.0000000 0.8068949
## dist  0.8068949 1.0000000

##             name syct mmin  mmax cach chmin chmax perf estperf
## 1  ADVISOR 32/60  125  256  6000  256    16   128  198     199
```

```
## 2  AMDAHL 470V/7   29 8000 32000   32     8    32  269     253
## 3  AMDAHL 470/7A   29 8000 32000   32     8    32  220     253
## 4 AMDAHL 470V/7B   29 8000 32000   32     8    32  172     253
## 5 AMDAHL 470V/7C   29 8000 16000   32     8    16  132     132
## 6  AMDAHL 470V/8   26 8000 32000   64     8    32  318     290
```

Ex 22

Voici les données de cet exercice : on a 168 individus et 11 variables.

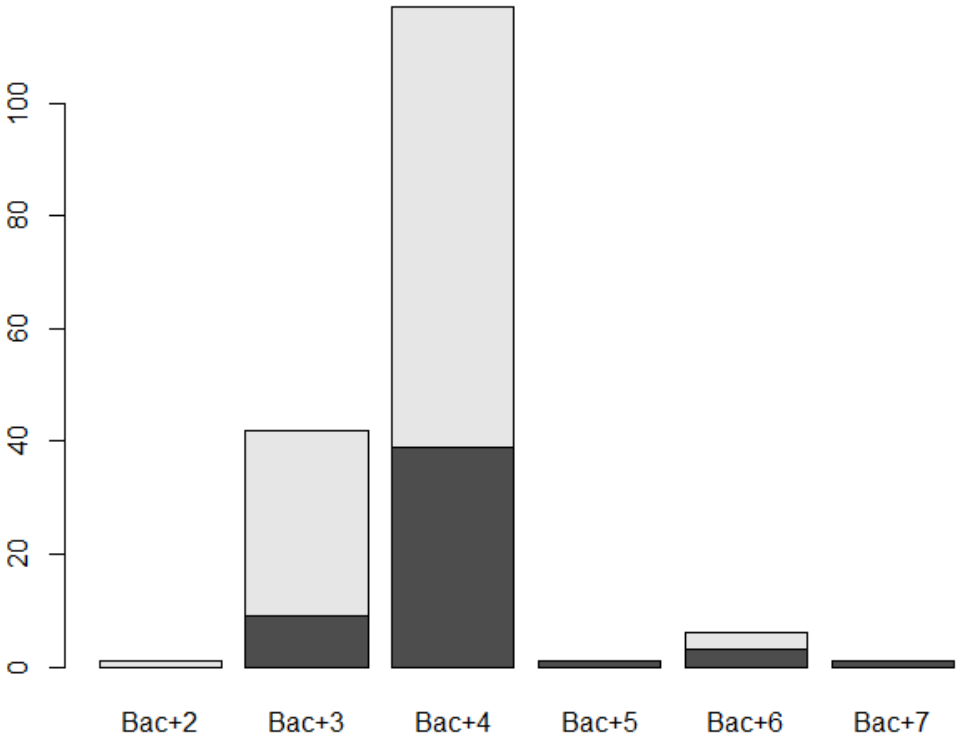
```
## # A tibble: 6 × 11
##   Sexe    Age EtatCivil   Nbenf…¹ Diplome Ancie…² Salaire Satis…³ Stress Estim…⁴
##   <chr> <dbl> <chr>         <dbl> <chr>     <dbl>   <dbl>   <dbl>  <dbl>   <dbl>
## 1 Homme    37 Célibataire       0 Bac+3        11    1600   14.4    15.7   16.2
## 2 Homme    38 Célibataire       2 Bac+3        14    1670   17.6    18.9   17.6
## 3 Femme    29 Célibataire       0 Bac+3         1    1600    4.05   21.4    4.31
## 4 Homme    53 Marié(e)          2 Bac+3        28    1896   32.6    13.9   34.6
## 5 Homme    30 Marié(e)          1 Bac+3         7    1996   10.5    17.9   10.0
## 6 Homme    44 Marié(e)          2 Bac+3        18    1960   22.2    18.8   22.6
## # … with 1 more variable: AvisReforme <chr>, and abbreviated variable names
## #   ¹Nbenfant, ²Anciennete, ³Satisfaction, ⁴EstimeSoi

## [1] 168  11
```

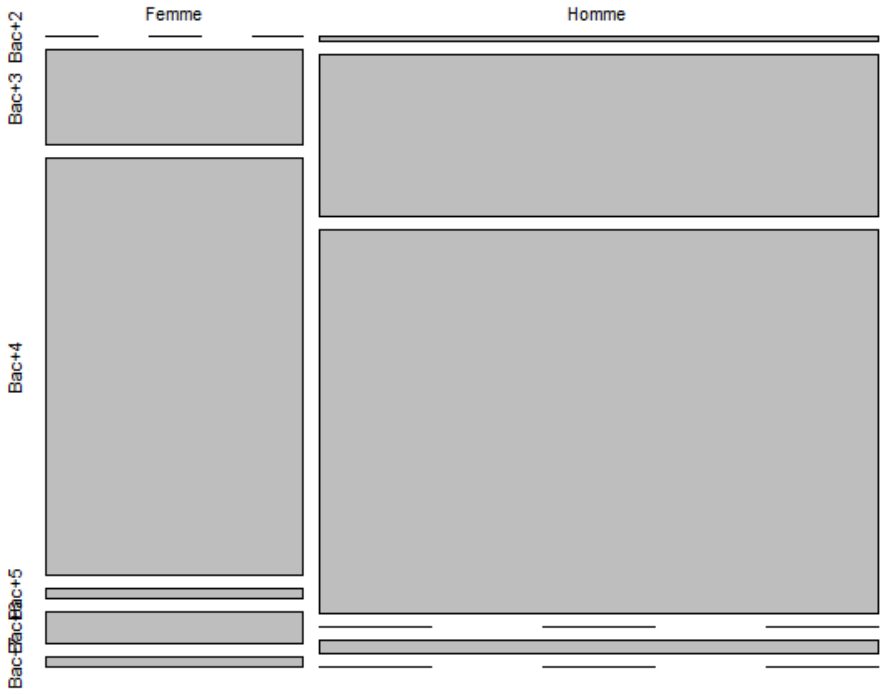
L'âge des enseignants est entre 25 et 57 ans, les salaires vont de 1200€ à 2200. 50 % des enseignants touche moins ou 1720€ et 50% touche 1720 ou plus. Le salaire moyen est de 1778 €.

```
##      Sexe                Age          EtatCivil            Nbenfant
##  Length:168         Min.   :25.00   Length:168         Min.   :0.00
##  Class :character   1st Qu.:37.00   Class :character   1st Qu.:1.00
##  Mode  :character   Median :41.00   Mode  :character   Median :2.00
##                     Mean   :41.99                      Mean   :1.72
##                     3rd Qu.:49.25                      3rd Qu.:2.00
##                     Max.   :57.00                      Max.   :5.00
##    Diplome            Anciennete       Salaire      Satisfaction
##  Length:168         Min.   : 1.00   Min.   :1200   Min.   : 3.85
##  Class :character   1st Qu.:10.00   1st Qu.:1650   1st Qu.:13.84
##  Mode  :character   Median :15.00   Median :1720   Median :19.17
##                     Mean   :16.55   Mean   :1778   Mean   :20.43
##                     3rd Qu.:24.25   3rd Qu.:1908   3rd Qu.:28.31
##                     Max.   :34.00   Max.   :2200   Max.   :38.45
##      Stress        EstimeSoi     AvisReforme
##  Min.   : 3.70   Min.   : 3.54   Length:168
##  1st Qu.:15.19   1st Qu.:14.03   Class :character
##  Median :18.19   Median :19.68   Mode  :character
##  Mean   :18.20   Mean   :21.08
##  3rd Qu.:21.11   3rd Qu.:29.84
##  Max.   :31.84   Max.   :42.15

##  [1] "Sexe"         "Age"          "EtatCivil"    "Nbenfant"     "Diplome"
##  [6] "Anciennete"   "Salaire"      "Satisfaction" "Stress"       "EstimeSoi"
## [11] "AvisReforme"
```

Voici les tableaux de contingence, de fréquence et de fréquence en pourcentage des variables “Sexe” et “Diplome” qui sont des données qualitatives suivie du graphique correspondant.

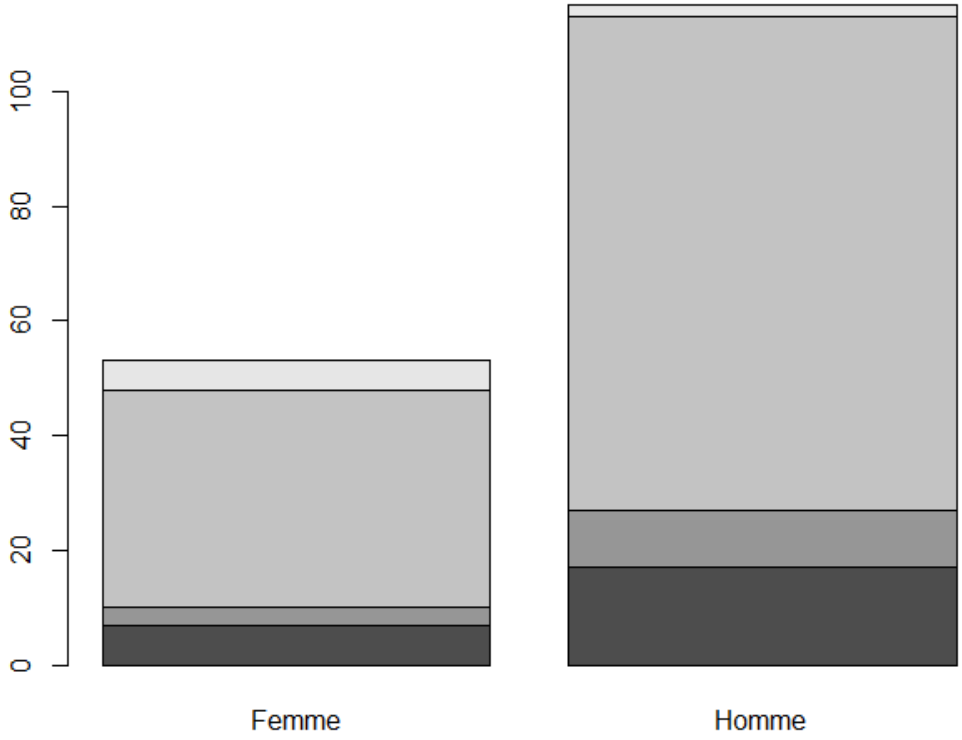
```
##
##         Bac+2 Bac+3 Bac+4 Bac+5 Bac+6 Bac+7
##   Femme     0     9    39     1     3     1
##   Homme     1    33    78     0     3     0

##
##               Bac+2       Bac+3       Bac+4       Bac+5       Bac+6       Bac+7
##   Femme 0.000000000 0.053571429 0.232142857 0.005952381 0.017857143 0.005952381
##   Homme 0.005952381 0.196428571 0.464285714 0.000000000 0.017857143 0.000000000

##
##               Bac+2      Bac+3      Bac+4      Bac+5      Bac+6      Bac+7
##   Femme  0.0000000  5.3571429 23.2142857  0.5952381  1.7857143  0.5952381
##   Homme  0.5952381 19.6428571 46.4285714  0.0000000  1.7857143  0.0000000
```


0
20
40
60
80
100
Bac+2
Bac+3
Bac+4
Bac+5
Bac+6
Bac+7



Voici les tableaux de contingence, de fréquence et de fréquence en pourcentage des variables “Sexe” et “EtatCivil” qui sont des données qualitatives.

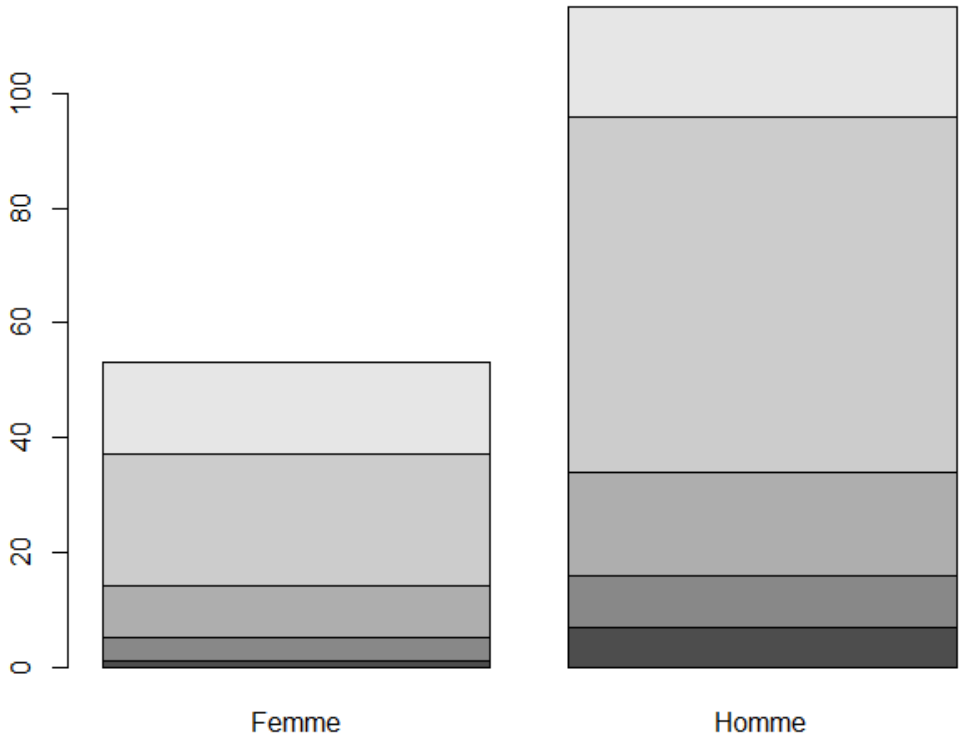
```
##
##                Femme Homme
##   Célibataire      7    17
##   Divorcé(e)       3    10
##   Marié(e)        38    86
##   Veuf(ve)         5     2

##
##               Bac+2       Bac+3       Bac+4       Bac+5       Bac+6       Bac+7
##   Femme 0.000000000 0.053571429 0.232142857 0.005952381 0.017857143 0.005952381
##   Homme 0.005952381 0.196428571 0.464285714 0.000000000 0.017857143 0.000000000

##
##               Bac+2      Bac+3      Bac+4      Bac+5      Bac+6      Bac+7
##   Femme   0.0000000  5.3571429 23.2142857  0.5952381  1.7857143  0.5952381
##   Homme   0.5952381 19.6428571 46.4285714  0.0000000  1.7857143  0.0000000
```


0
20
40
60
80
100
Femme
Homme

Voici les tableaux de contingence, de fréquence et de fréquence en pourcentage des variables “Sexe” et “Reforme” qui sont des données qualitatives.

```
##
##                     Femme Homme
##   Défavorable           1     7
##   Favorable             4     9
##   Neutre                9    18
##   Très défavorable     23    62
##   Très favorable       16    19

##
##                          Femme       Homme
##   Défavorable      0.005952381 0.041666667
##   Favorable        0.023809524 0.053571429
##   Neutre           0.053571429 0.107142857
##   Très défavorable 0.136904762 0.369047619
##   Très favorable   0.095238095 0.113095238

##
##                        Femme      Homme
##   Défavorable       0.5952381  4.1666667
```

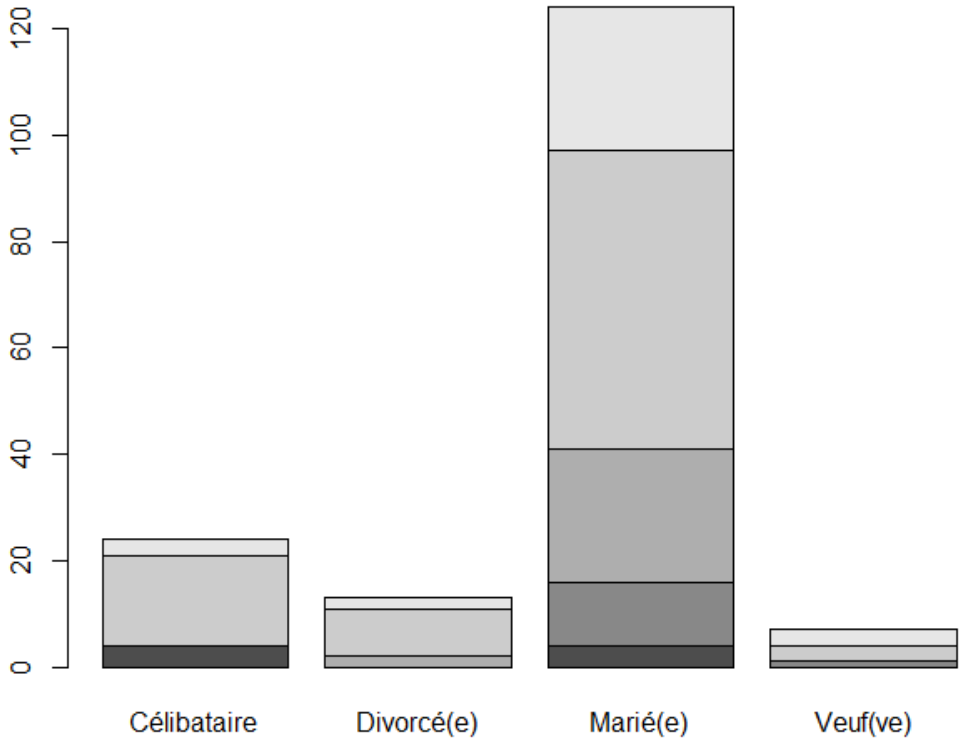
```
##   Favorable         2.3809524  5.3571429
##   Neutre            5.3571429 10.7142857
##   Très défavorable 13.6904762 36.9047619
##   Très favorable    9.5238095 11.3095238
```


0
20
40
60
80
100
Femme
Homme

Voici les tableaux de contingence, de fréquence et de fréquence en pourcentage des variables “EtatCivil” et “Reforme” qui sont des données qualitatives.

```
##
##                    Célibataire Divorcé(e) Marié(e) Veuf(ve)
##   Défavorable                4          0        4        0
##   Favorable                  0          0       12        1
##   Neutre                     0          2       25        0
##   Très défavorable          17          9       56        3
##   Très favorable             3          2       27        3

##
##                    Célibataire  Divorcé(e)    Marié(e)    Veuf(ve)
##   Défavorable      0.023809524 0.000000000 0.023809524 0.000000000
##   Favorable        0.000000000 0.000000000 0.071428571 0.005952381
##   Neutre           0.000000000 0.011904762 0.148809524 0.000000000
##   Très défavorable 0.101190476 0.053571429 0.333333333 0.017857143
##   Très favorable   0.017857143 0.011904762 0.160714286 0.017857143

##
##                    Célibataire Divorcé(e)   Marié(e)   Veuf(ve)
##   Défavorable        2.3809524  0.0000000  2.3809524  0.0000000
```

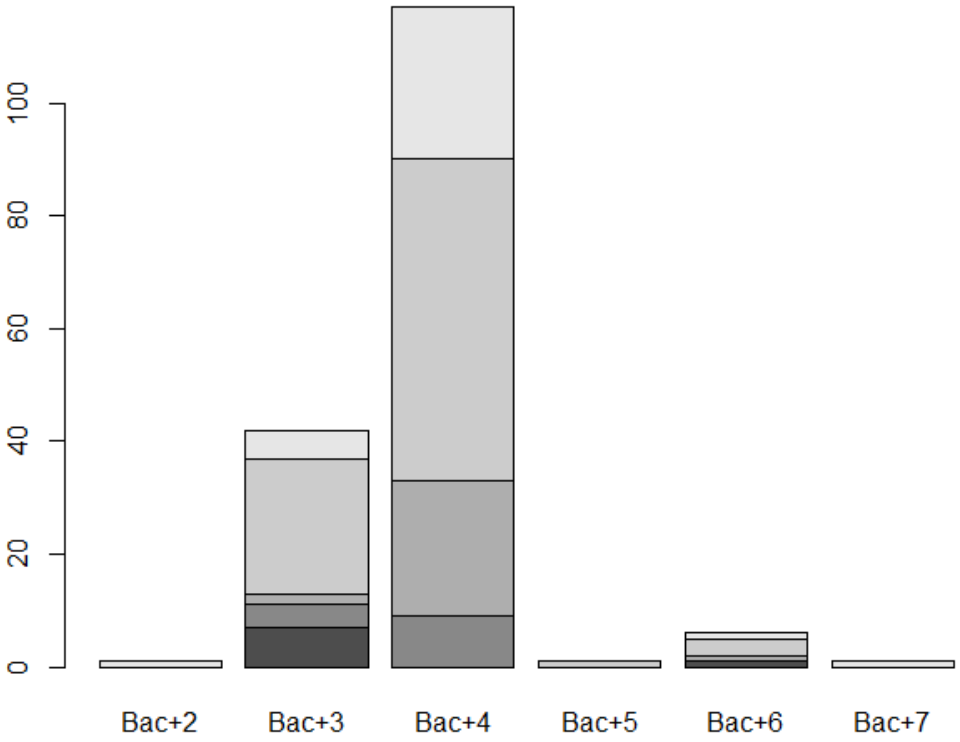
```
##   Favorable          0.0000000  0.0000000  7.1428571  0.5952381
##   Neutre             0.0000000  1.1904762 14.8809524  0.0000000
##   Très défavorable  10.1190476  5.3571429 33.3333333  1.7857143
##   Très favorable     1.7857143  1.1904762 16.0714286  1.7857143
```

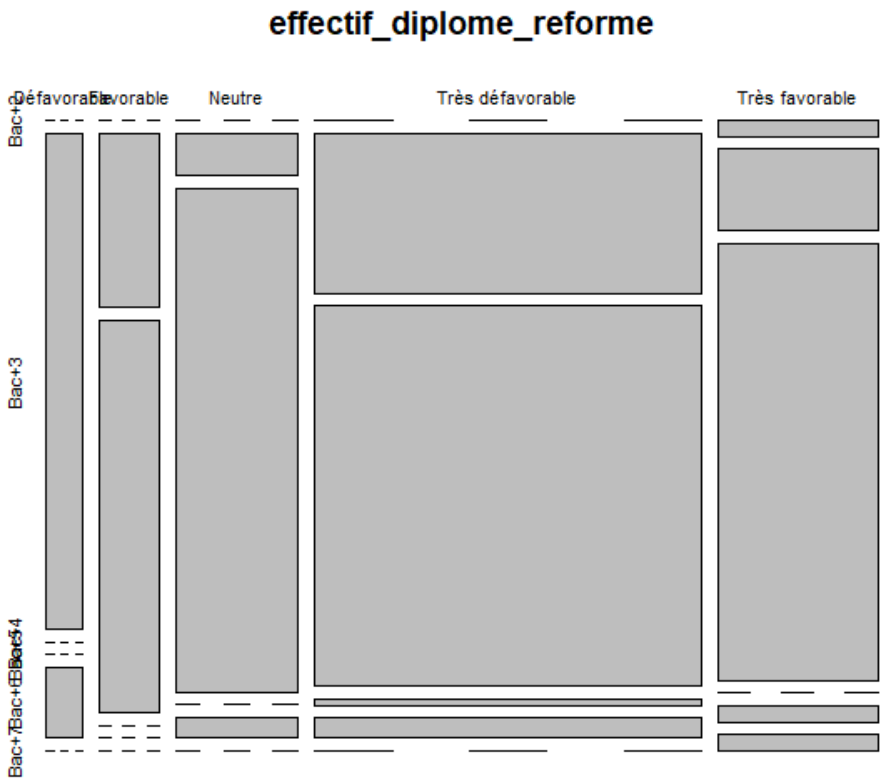

0
20
40
60
80
100
120
Célibataire
Divorcé(e)
Marié(e)
Veuf(ve)

Voici les tableaux de contingence, de fréquence et de fréquence en pourcentage des variables “Diplome” et “Reforme” qui sont des données qualitatives.

```
##
##                     Bac+2 Bac+3 Bac+4 Bac+5 Bac+6 Bac+7
##   Défavorable           0     7     0     0     1     0
##   Favorable             0     4     9     0     0     0
##   Neutre                0     2    24     0     1     0
##   Très défavorable      0    24    57     1     3     0
##   Très favorable        1     5    27     0     1     1

##
##                          Bac+2       Bac+3       Bac+4       Bac+5       Bac+6
##   Défavorable      0.000000000 0.041666667 0.000000000 0.000000000 0.005952381
##   Favorable        0.000000000 0.023809524 0.053571429 0.000000000 0.000000000
##   Neutre           0.000000000 0.011904762 0.142857143 0.000000000 0.005952381
##   Très défavorable 0.000000000 0.142857143 0.339285714 0.005952381 0.017857143
##   Très favorable   0.005952381 0.029761905 0.160714286 0.000000000 0.005952381
##
##                          Bac+7
##   Défavorable      0.000000000
##   Favorable        0.000000000
##   Neutre           0.000000000
```

```
##   Très défavorable 0.000000000
##   Très favorable   0.005952381

##
##                         Bac+2      Bac+3      Bac+4      Bac+5      Bac+6
##   Défavorable       0.0000000  4.1666667  0.0000000  0.0000000  0.5952381
##   Favorable         0.0000000  2.3809524  5.3571429  0.0000000  0.0000000
##   Neutre            0.0000000  1.1904762 14.2857143  0.0000000  0.5952381
##   Très défavorable  0.0000000 14.2857143 33.9285714  0.5952381  1.7857143
##   Très favorable    0.5952381  2.9761905 16.0714286  0.0000000  0.5952381
##
##                         Bac+7
##   Défavorable       0.0000000
##   Favorable         0.0000000
##   Neutre            0.0000000
##   Très défavorable  0.0000000
##   Très favorable    0.5952381
```

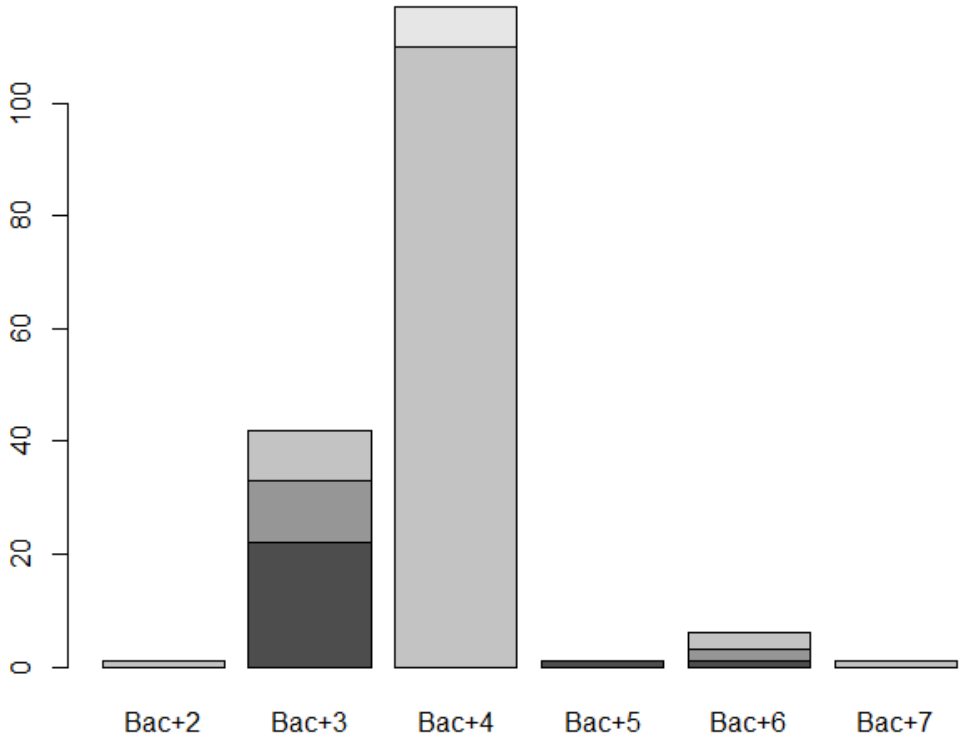

0
20
40
60
80
100
Bac+2
Bac+3
Bac+4
Bac+5
Bac+6
Bac+7

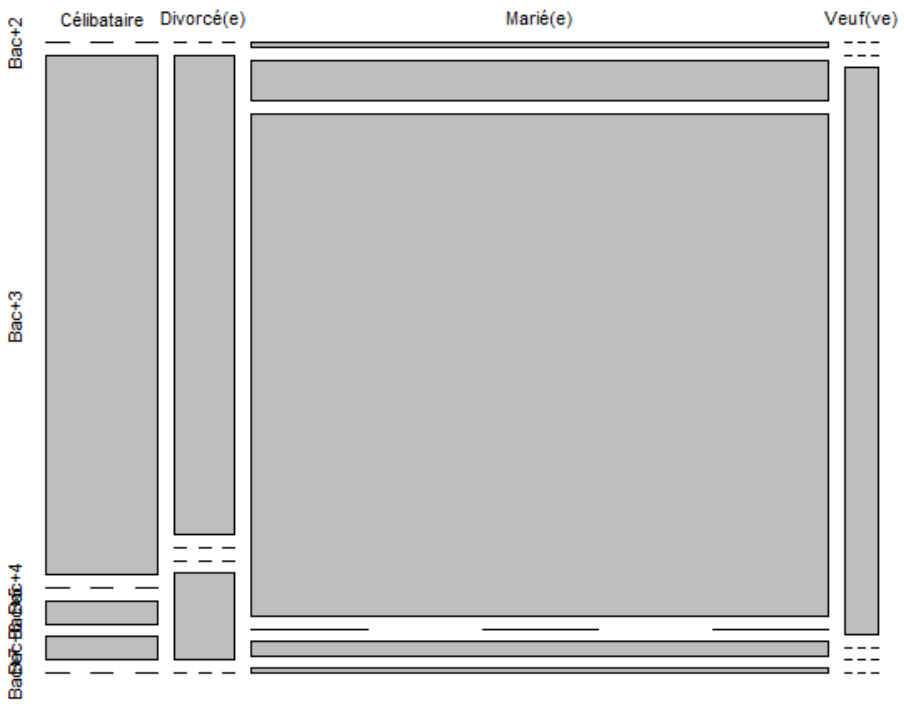


Voici les tableaux de contingence, de fréquence et de fréquence en pourcentage des variables “Diplome” et “EtatCivil” qui sont des données qualitatives.

```
##
##               Bac+2 Bac+3 Bac+4 Bac+5 Bac+6 Bac+7
##   Célibataire     0    22     0     1     1     0
##   Divorcé(e)      0    11     0     0     2     0
##   Marié(e)        1     9   110     0     3     1
##   Veuf(ve)        0     0     7     0     0     0

##
##                     Bac+2       Bac+3       Bac+4       Bac+5       Bac+6
##   Célibataire 0.000000000 0.130952381 0.000000000 0.005952381 0.005952381
##   Divorcé(e)  0.000000000 0.065476190 0.000000000 0.000000000 0.011904762
##   Marié(e)    0.005952381 0.053571429 0.654761905 0.000000000 0.017857143
##   Veuf(ve)    0.000000000 0.000000000 0.041666667 0.000000000 0.000000000
##
##                     Bac+7
##   Célibataire 0.000000000
##   Divorcé(e)  0.000000000
##   Marié(e)    0.005952381
##   Veuf(ve)    0.000000000
```

```
##
##                    Bac+2      Bac+3      Bac+4      Bac+5      Bac+6      Bac+7
##   Célibataire  0.0000000 13.0952381  0.0000000  0.5952381  0.5952381  0.0000000
##   Divorcé(e)   0.0000000  6.5476190  0.0000000  0.0000000  1.1904762  0.0000000
##   Marié(e)     0.5952381  5.3571429 65.4761905  0.0000000  1.7857143  0.5952381
##   Veuf(ve)     0.0000000  0.0000000  4.1666667  0.0000000  0.0000000  0.0000000
```


0
20
40
60
80
100
Bac+2
Bac+3
Bac+4
Bac+5
Bac+6
Bac+7



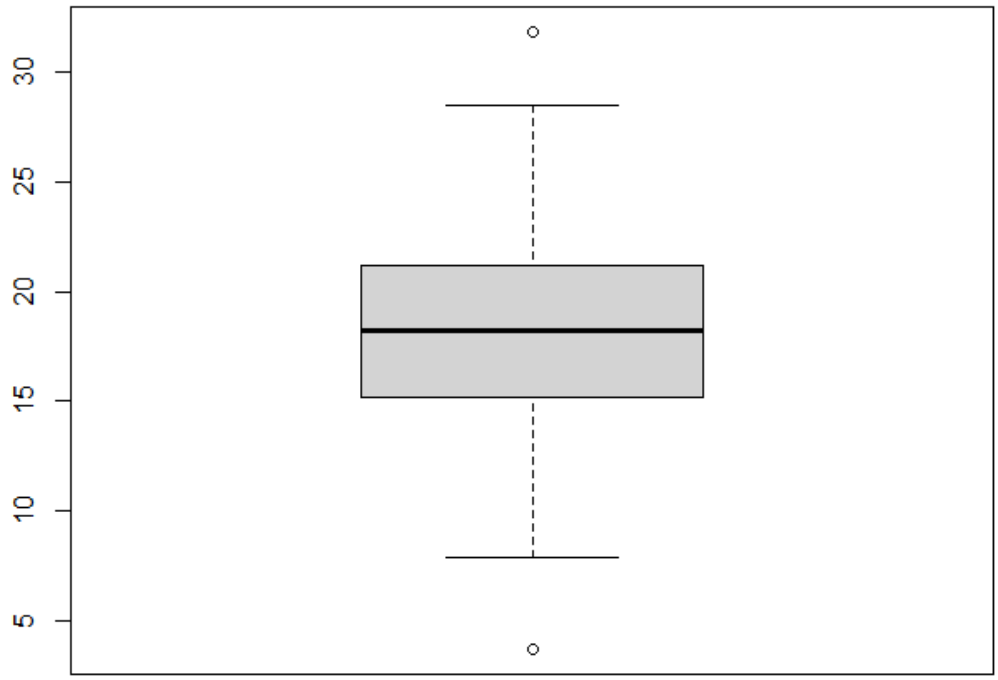
Voici un test de khi-deux des variables “Sexe” et “EtatCivil”. On observe une p-value de 0.1273 qui est supérieur au seuil de 0.05. On peut donc dire qu’il y a non indépendances des variables. L’état civil semble être lié au sexe des individus.

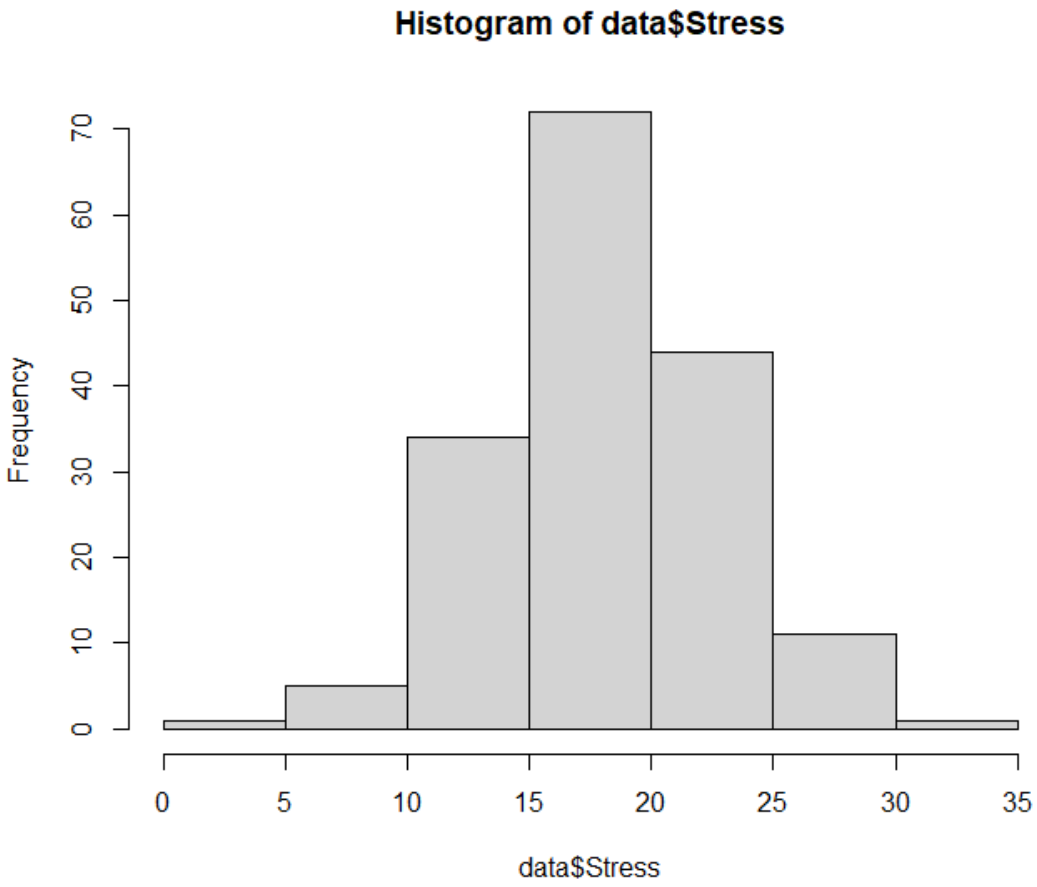
```
## Warning in chisq.test(data$Sexe, data$EtatCivil): Chi-squared
approximation may
## be incorrect

##
##  Pearson's Chi-squared test
##
## data:  data$Sexe and data$EtatCivil
## X-squared = 5.6972, df = 3, p-value = 0.1273
```

On se concentre sur les variables Stress et EtatCivil. Sur la boite à moustache de la variable “Stress”, on voit deux valeurs aberrantes, la médiane est d’environ 17 et on voit une grande amplitude de valeur.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.70   15.19   18.19   18.20   21.11   31.84
```

```
## [1] (15,20.6]   (15,20.6]   (20.6,26.2] (9.33,15]   (15,20.6]   (15,20.6]
## Levels: (3.67,9.33] (9.33,15] (15,20.6] (20.6,26.2] (26.2,31.9]
```

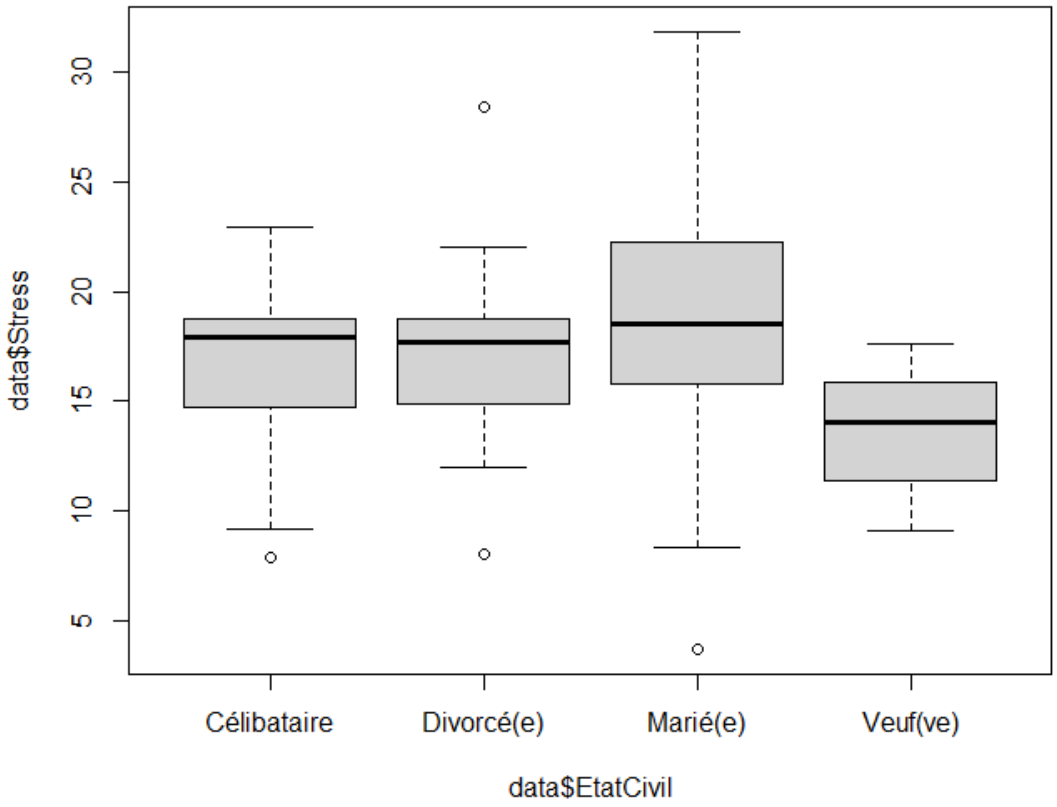




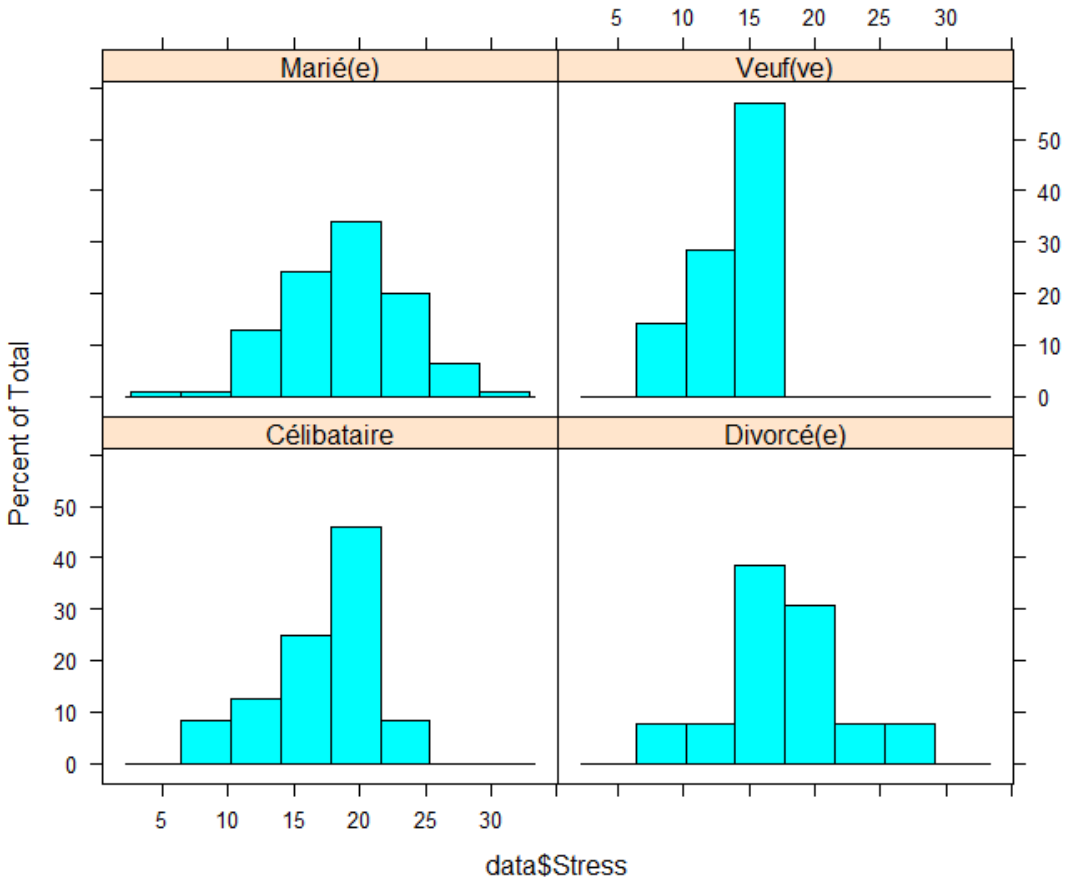
Voici le tableau de contingence de “Stress” et “EtatCivil”

Les enseignants mariés ont plus d’amplitude dans leur niveau de stress que les veufs. Médiane est un peu près égale pour les célibataires et les divorcés, un peu près pareil pour

les mariés mais plus basse pour les veufs.



On observe beaucoup moins d'amplitude de données chez les veufs que chez les enseignants mariés qui ont la plus grande, on observe un pic vers 15/20 chez chacun des états civils.

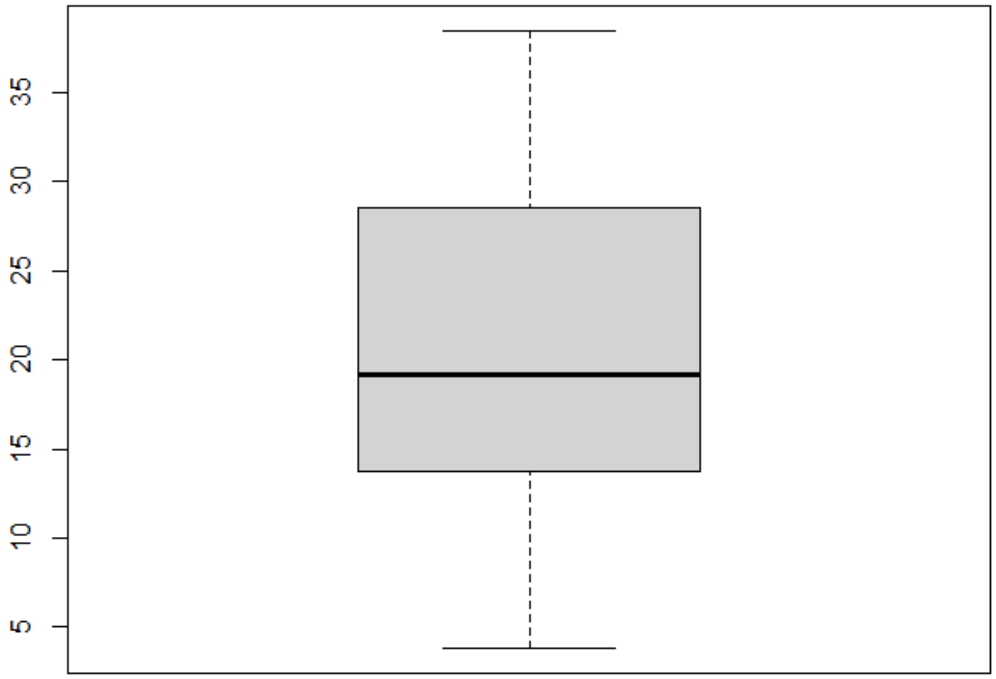



```
## data$EtatCivil: Célibataire
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.86   14.91   17.95   16.76   18.66   22.94
## ------------------------------------------------------------
## data$EtatCivil: Divorcé(e)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   14.86   17.72   17.17   18.74   28.40
## ------------------------------------------------------------
## data$EtatCivil: Marié(e)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.70   15.82   18.53   18.85   22.25   31.84
## ------------------------------------------------------------
## data$EtatCivil: Veuf(ve)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.12   11.37   14.04   13.60   15.86   17.58

## [1] 0.07492283
```

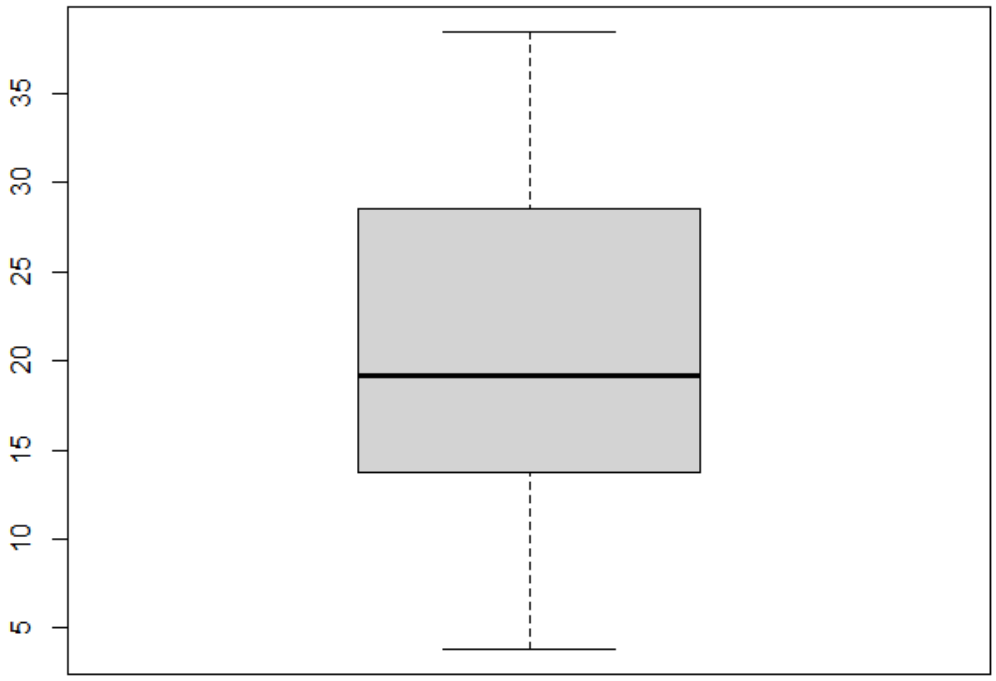
On se concentre maintenant sur les données quantitatives “Age” vs “Satisfaction”.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.85   13.84   19.17   20.43   28.31   38.45
```

La médiane est d'environ 18 sur ce graphique en boîte à moustache, on voit une amplitude de 30 points.



On ne détecte pas de valeurs aberrantes sur ce graphique.

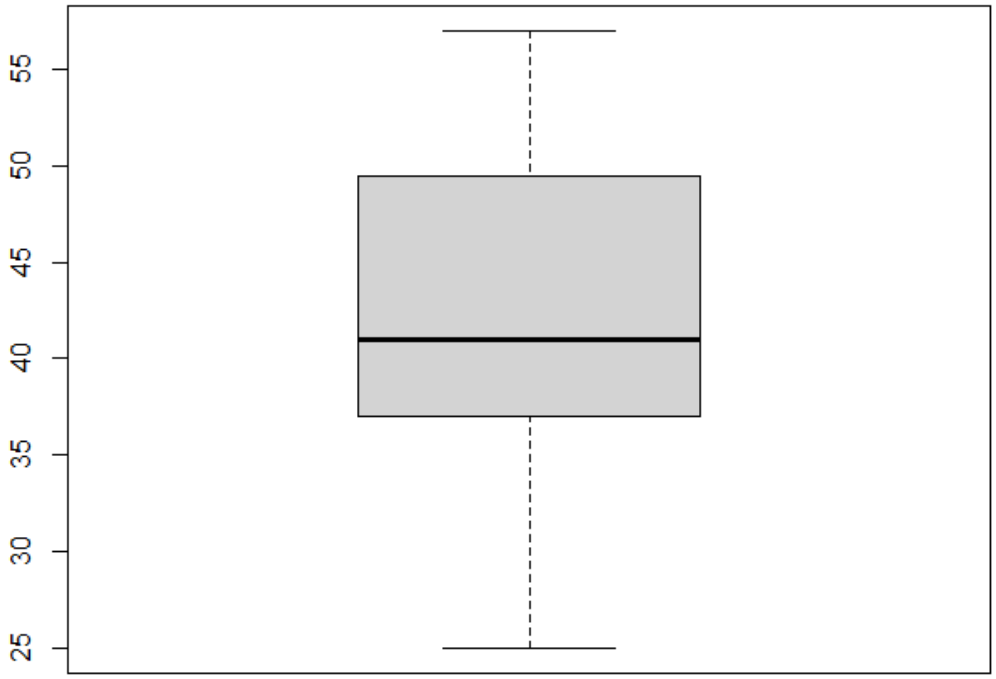



```
## integer(0)
```

Voici le résumé de la variable “Age”

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.00   37.00   41.00   41.99   49.25   57.00
```

Les âge sont entre 25 et 57, la médiane est de 41, on voit une grande amplitude de valeurs. On ne détecte pas de valeurs aberrantes sur ce graphique.



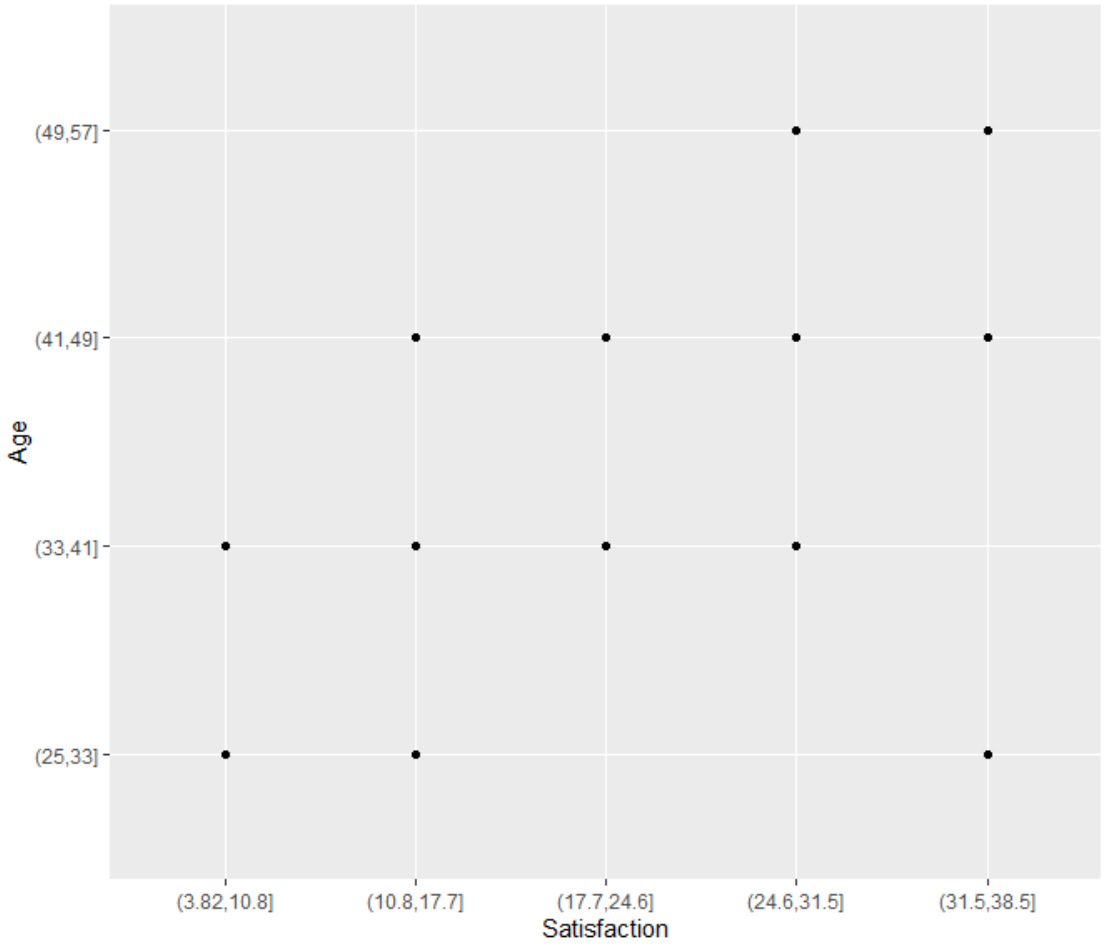
On constitue 5 classes pour la variable “Satisfaction” et 4 pour la variable “Age” :

Voici les tableaux de contingence, de fréquence et de fréquence en pourcentage des variables quantitatives “Satisfaction” et “Age”

```
##
##                (25,33] (33,41] (41,49] (49,57]
##   (3.82,10.8]      26       1       0       0
##   (10.8,17.7]       2      38       3       0
##   (17.7,24.6]       0      20      22       0
##   (24.6,31.5]       0       1      11      18
##   (31.5,38.5]       1       0       1      24

##
##                (25,33] (33,41] (41,49] (49,57]
##   (3.82,10.8]    0.15    0.01    0.00    0.00
##   (10.8,17.7]    0.01    0.23    0.02    0.00
##   (17.7,24.6]    0.00    0.12    0.13    0.00
##   (24.6,31.5]    0.00    0.01    0.07    0.11
##   (31.5,38.5]    0.01    0.00    0.01    0.14
```

```
##
##               (25,33] (33,41] (41,49] (49,57]
##   (3.82,10.8]   15.48    0.60    0.00    0.00
##   (10.8,17.7]    1.19   22.62    1.79    0.00
##   (17.7,24.6]    0.00   11.90   13.10    0.00
##   (24.6,31.5]    0.00    0.60    6.55   10.71
##   (31.5,38.5]    0.60    0.00    0.60   14.29

##
```

On peut remarquer un 'intru' qui se trouve à la fois dans la classe (32.5,38.5] de Satisfaction et dans la classe (25,33] en âge.



Chapitre 2

Ex 23

```
##    1  2 3 4  5
## Z1 1  2 3 4  9
## Z2 5 10 8 8 12
```

On calcul la moyenne et l'écart type de Z1:

```
## [1] 3.8

## [1] 3.114482
```

On calcul la moyenne et l'écart type de Z2

```
## [1] 8.6

## [1] 2.607681
```

Valeurs centrées réduites :

```
##             1          2          3           4        5
## Z1 -0.8990258 -0.5779452 -0.2568645  0.06421613 1.669619
## Z2 -1.3805370  0.5368755 -0.2300895 -0.23008950 1.303840
```

La matrice de corrélation indique qu'il existe une redondance dans les données.

```
##    1  2  3  4  5
## 1  1  1  1 -1 -1
## 2  1  1  1 -1 -1
## 3  1  1  1 -1 -1
## 4 -1 -1 -1  1  1
## 5 -1 -1 -1  1  1
```

On obtient les mêmes résultats que avec sd().

```
## Call:
## princomp(x = donnees[1, ])
##
## Standard deviations:
##    Comp.1
## 0.8944272
##
##  1  variables and  5 observations.

## Call:
## princomp(x = donnees[2, ])
##
## Standard deviations:
##    Comp.1
## 0.8944272
```

```
##
##  1  variables and  5 observations.

## Standard deviations (1, .., p=1):
## [1] 1
##
## Rotation (n x k) = (1 x 1):
##      PC1
## [1,]   1

## Standard deviations (1, .., p=1):
## [1] 1
##
## Rotation (n x k) = (1 x 1):
##      PC1
## [1,]   1
```

(d) Décrire et utiliser les fonctions PCA du package FactoMineR, et comparer avec les résultats trouvés

```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

##    1  2 3 4  5
## Z1 1  2 3 4  9
## Z2 5 10 8 8 12
```

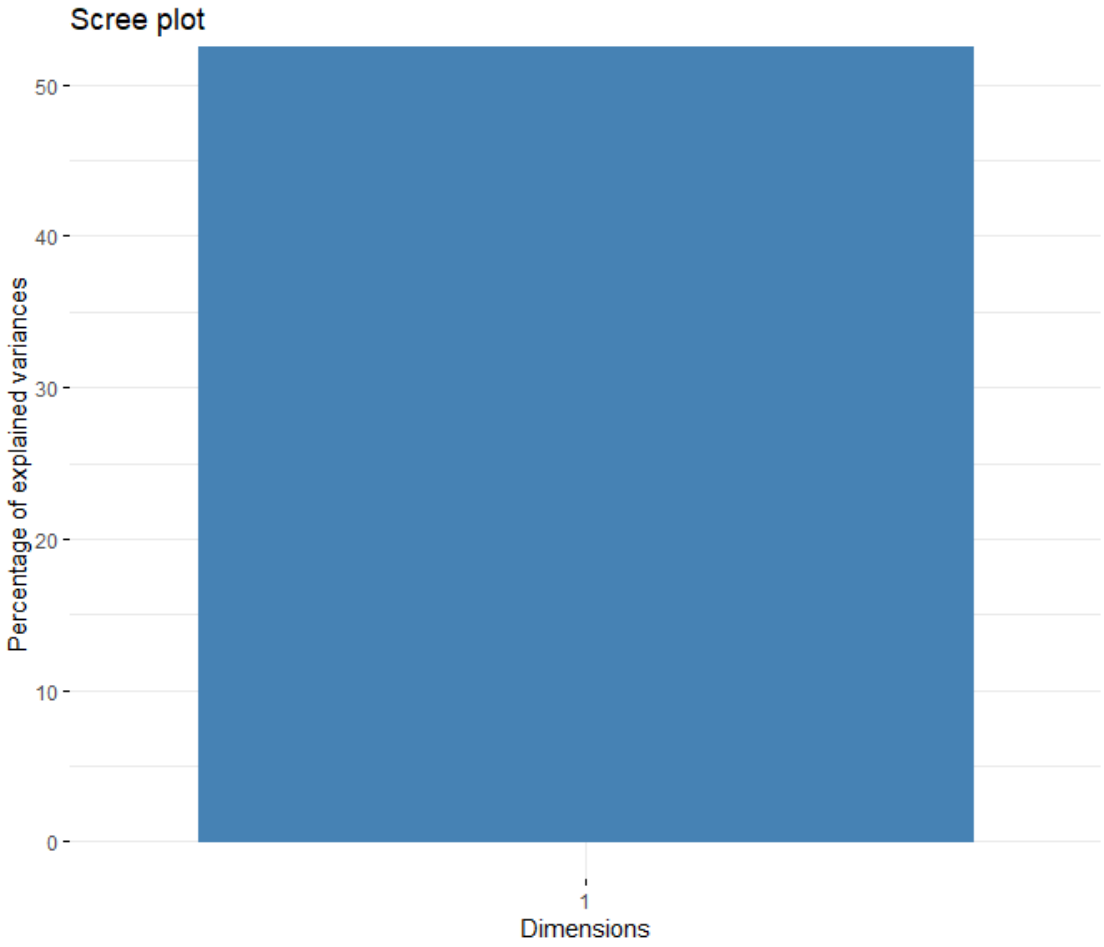
On lance la fonction PCA:

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 2 individuals, described by 5 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"
```

Extractions des valeurs propres : elles peuvent être utilisées pour déterminer le nombre d'axes principaux à conserver après l'ACP, ici 1 axe suffit à représenter la variance totale.

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1          5                    100                                100
```

Visualisation des valeurs propres : une seule variable explique toute la dimension, ce diagramme ne nous donne pas vraiment d'information dans ce cas précis.



Extraction des résultats pour les individus et les variables respectivement.

Dans l'immédiat on a les données suivantes :

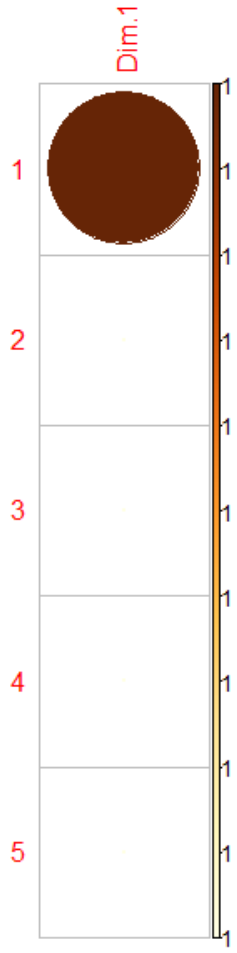
- var$cos2 : cosinus carré des variables, il représente la qualité de représentation des variables sur le graphique de l'ACP.
- var$contrib : contient les contributions (en pourcentage), des variables, aux composantes principales.

```
## Principal Component Analysis Results for variables
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the variables"
## 2 "$cor"     "Correlations between variables and dimensions"
```

```
## 3 "$cos2"    "Cos2 for the variables"
## 4 "$contrib" "contributions of the variables"
```

Le cercle de corrélation :

- la corrélation entre une variable et une composante principale (PC) est utilisée comme coordonnées de la variable sur la composante principale. La représentation des variables diffère de celle des observations : les observations sont représentées par leurs projections, mais les variables sont représentées par leurs corrélations



Ici toute la dimension une est expliquer par la première ligne.

Ex 24

```
##          Nom prixforf altmin altmax pistes kmfond remontee
## 1 LesAillons       76    900   2000     45     50       22
## 2    LesArcs      160    800   3226    117     30       69
## 3    Arèches       85    750   2300     30     47       15
## 4    Aussois       71    500   2750     21     10       11
## 5    Bessans       54   1710   2200      4     80        4
## 6   Bonneval       79   1850   3000     16      0       10
```

Pour la suite des opérations on enlève la première variable des données, c'est à dire le nom des stations car on analyse les données quantitative. On fait ensuite une analyse ACP.

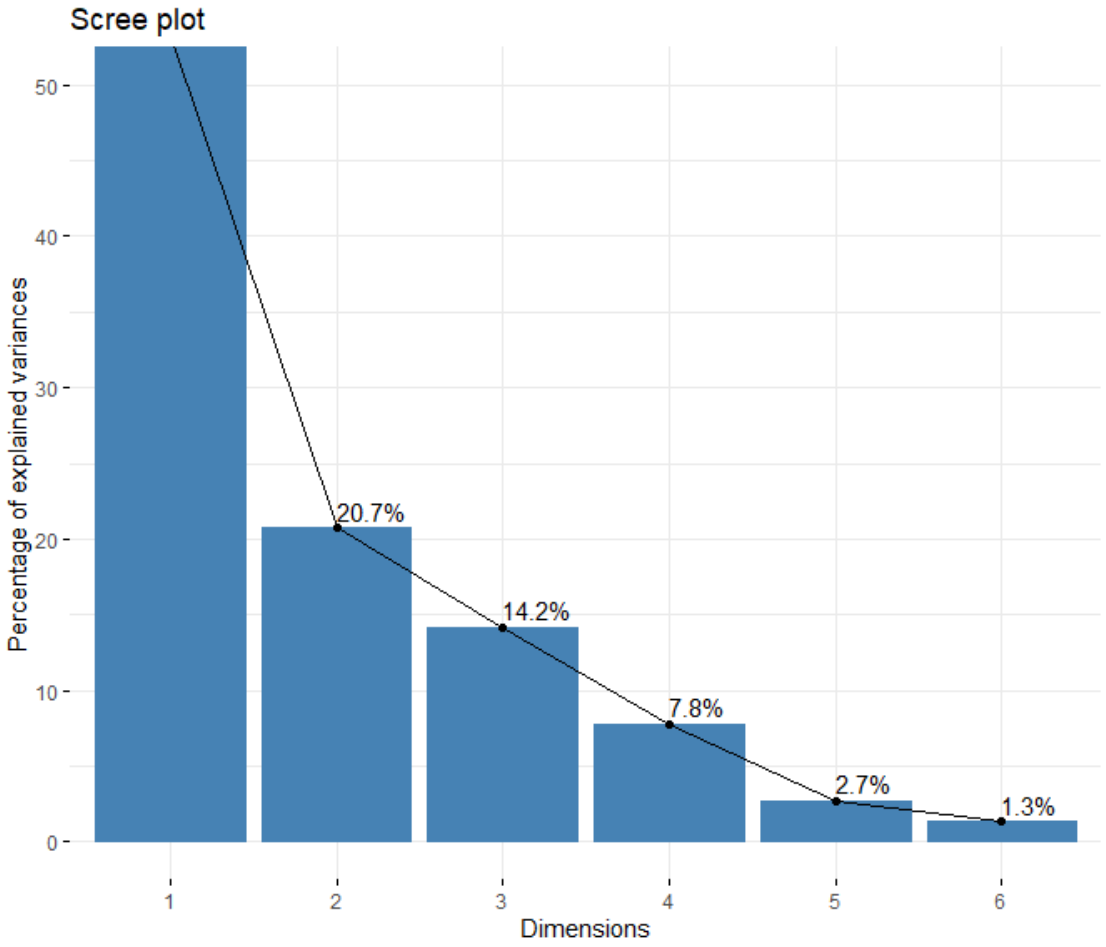
```
##   prixforf altmin altmax pistes kmfond remontee
## 1       76    900   2000     45     50       22
## 2      160    800   3226    117     30       69
## 3       85    750   2300     30     47       15
## 4       71    500   2750     21     10       11
## 5       54   1710   2200      4     80        4
## 6       79   1850   3000     16      0       10
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 32 individuals, described by 6 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"
```

Nous examinons les valeurs propres pour déterminer le nombre de composantes principales à prendre en considération. On garde les valeurs propres supérieur à 1, ici on a deux dimensions et on obtient plus de 70% de représentation des données.

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1 3.19906023              53.317670                           53.31767
## comp 2 1.24354998              20.725833                           74.04350
## comp 3 0.85138961              14.189827                           88.23333
## comp 4 0.46765198               7.794200                           96.02753
## comp 5 0.15977748               2.662958                           98.69049
## comp 6 0.07857073               1.309512                          100.00000
```

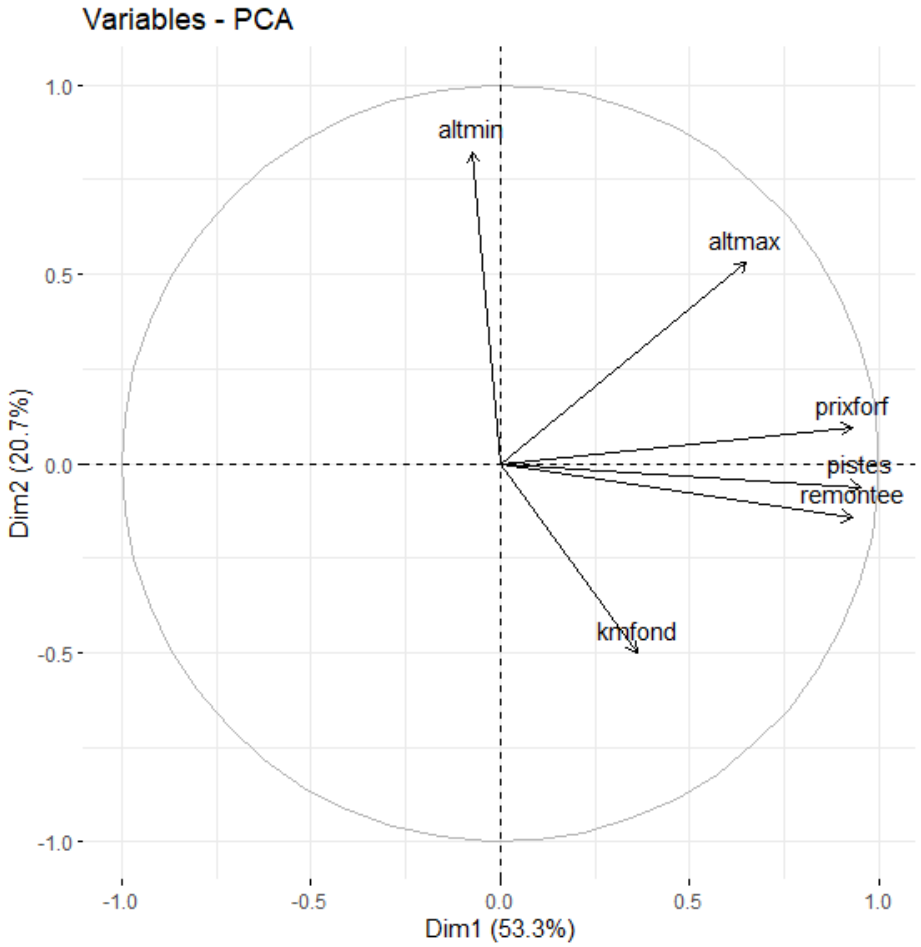
On peut conclure la même chose avec ce graphique, de plus on voit nettement un « coude » se dessiner ce qui nous indique le nombre de dimension à garder.



Le graphique suivant montre les relations entre toutes les variables. Il peut être interprété comme suit :

- Les variables positivement corrélées sont regroupées.
- Les variables négativement corrélées sont positionnées sur les côtés opposés de l'origine du graphique (quadrants opposés).
- La distance entre les variables et l'origine mesure la qualité de représentation des variables. Les variables qui sont loin de l'origine sont bien représentées par l'ACP.

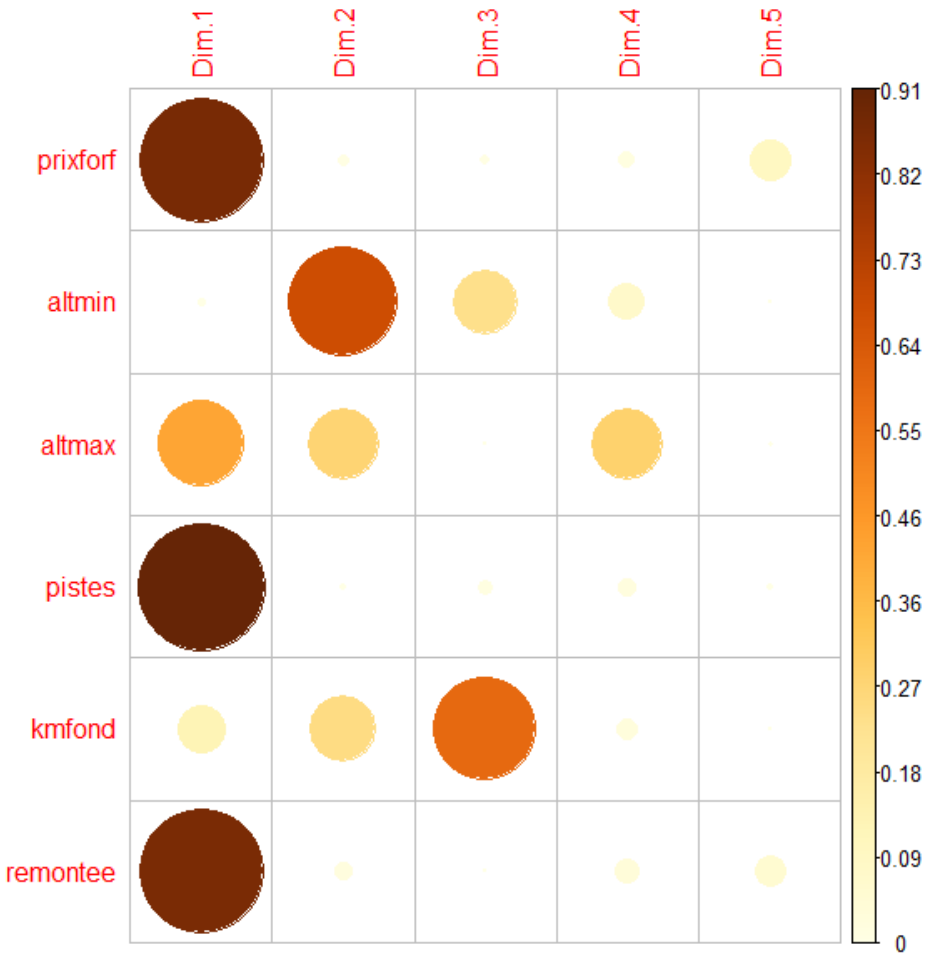
```
## Principal Component Analysis Results for variables
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the variables"
## 2 "$cor"     "Correlations between variables and dimensions"
## 3 "$cos2"    "Cos2 for the variables"
## 4 "$contrib" "contributions of the variables"
```



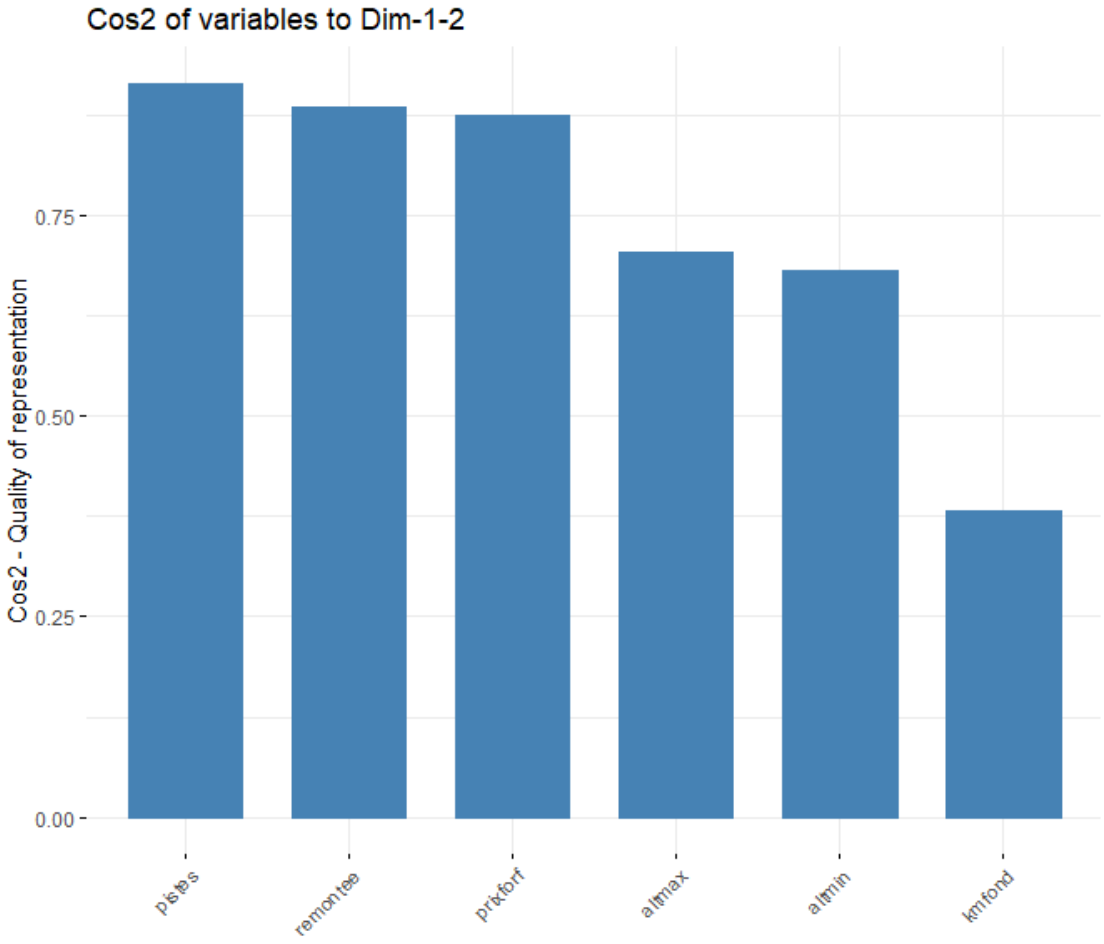
On voit ici que les variables prixfort, pistes et remontee sont positivement corrélées. La variable krnfond est la variable la moins bien représentée par l'analyse.

Ici on voit quelles variables est le mieux représenté suivant les dimensions. On peut déterminé de la même façon que les variables prixfort, pistes et remontee sont bien représentées sur la dimension 1 et altmin est la dimension la mieux représentée sur la

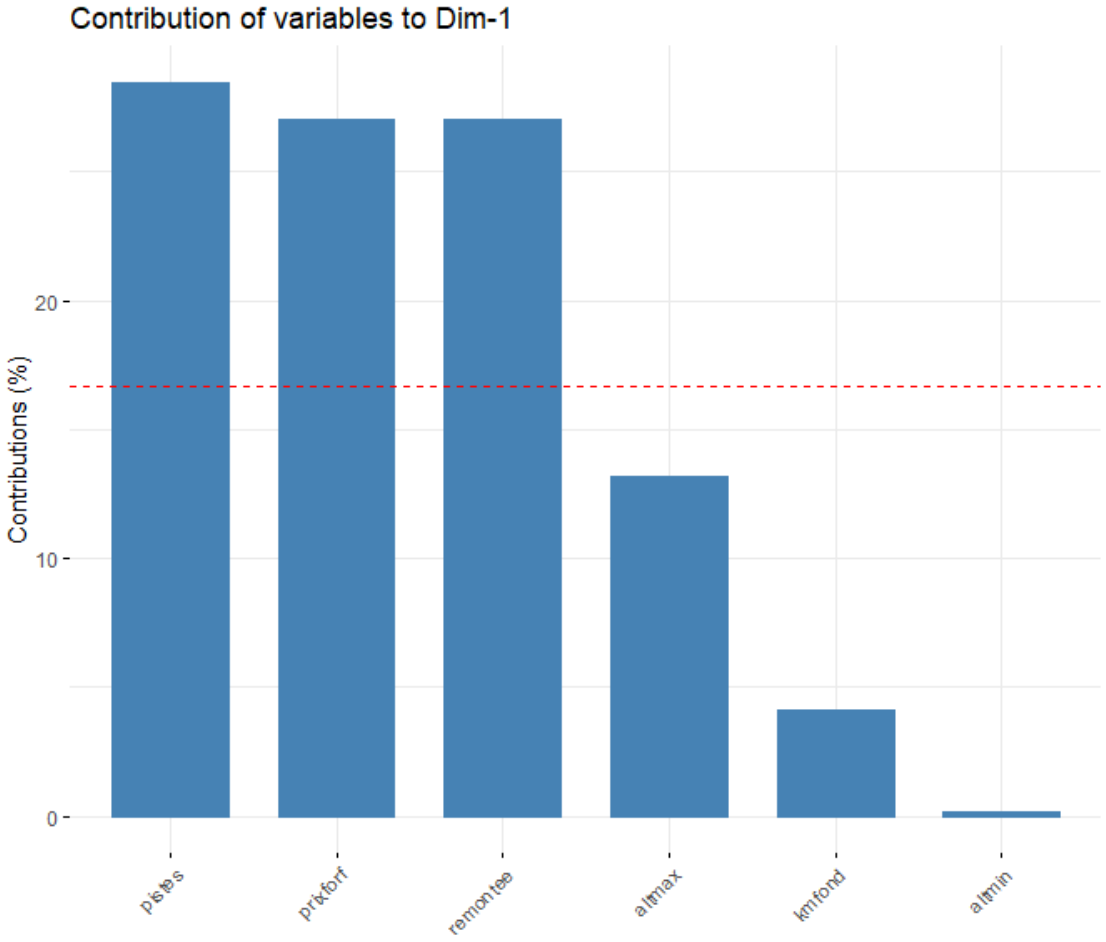
dimension 2.

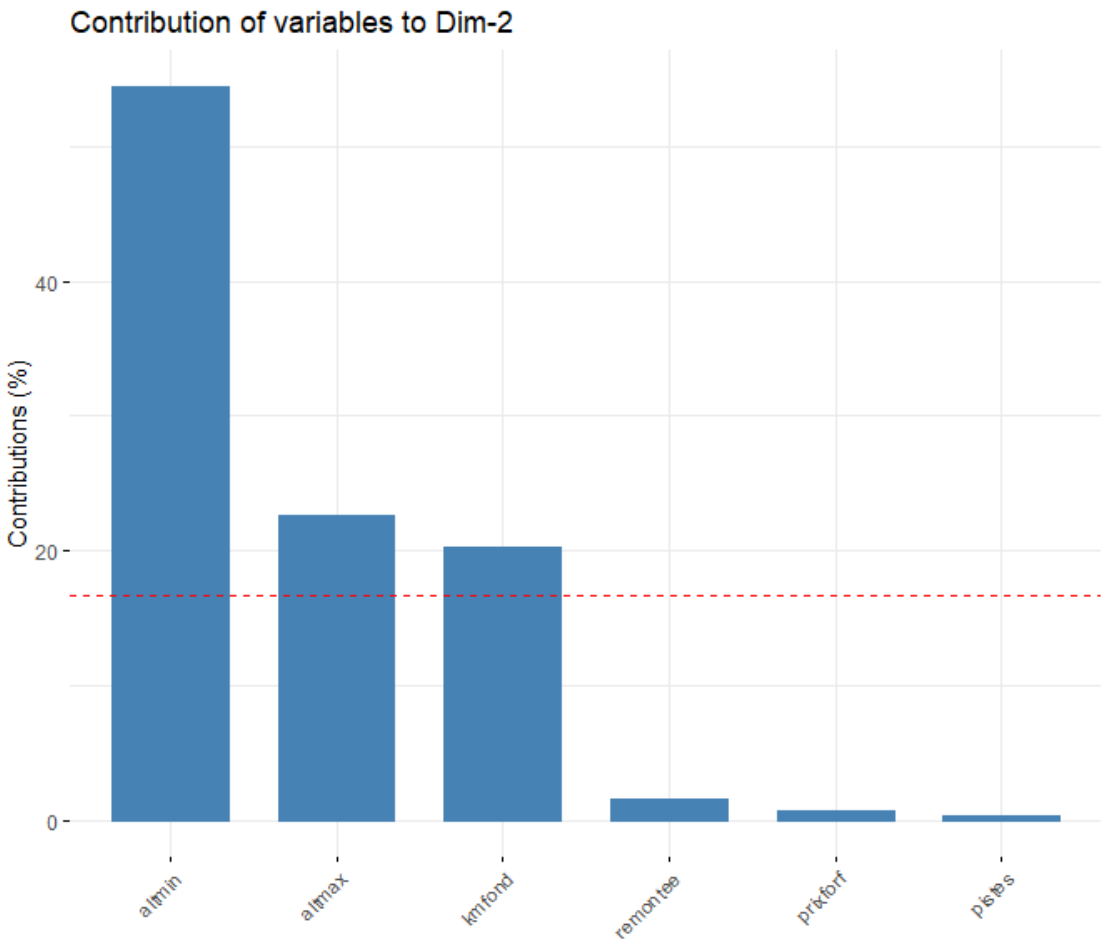

Dim.1
Dim.2
Dim.3
Dim.4
Dim.5
prixforf
altmin
altmax
pistes
kmfond
remontee
0.91
0.82
0.73
0.64
0.55
0.46
0.36
0.27
0.18
0.09
0

Ce graphique représente les variables les mieux représentés sur les dimensions 1 et 2. Ici les variables “pistes”, “remontee” et “prixfort” sont représentées à plus de 75%.



Ces graphiques représentent la contribution des variables au dimension 1 puis 2 :





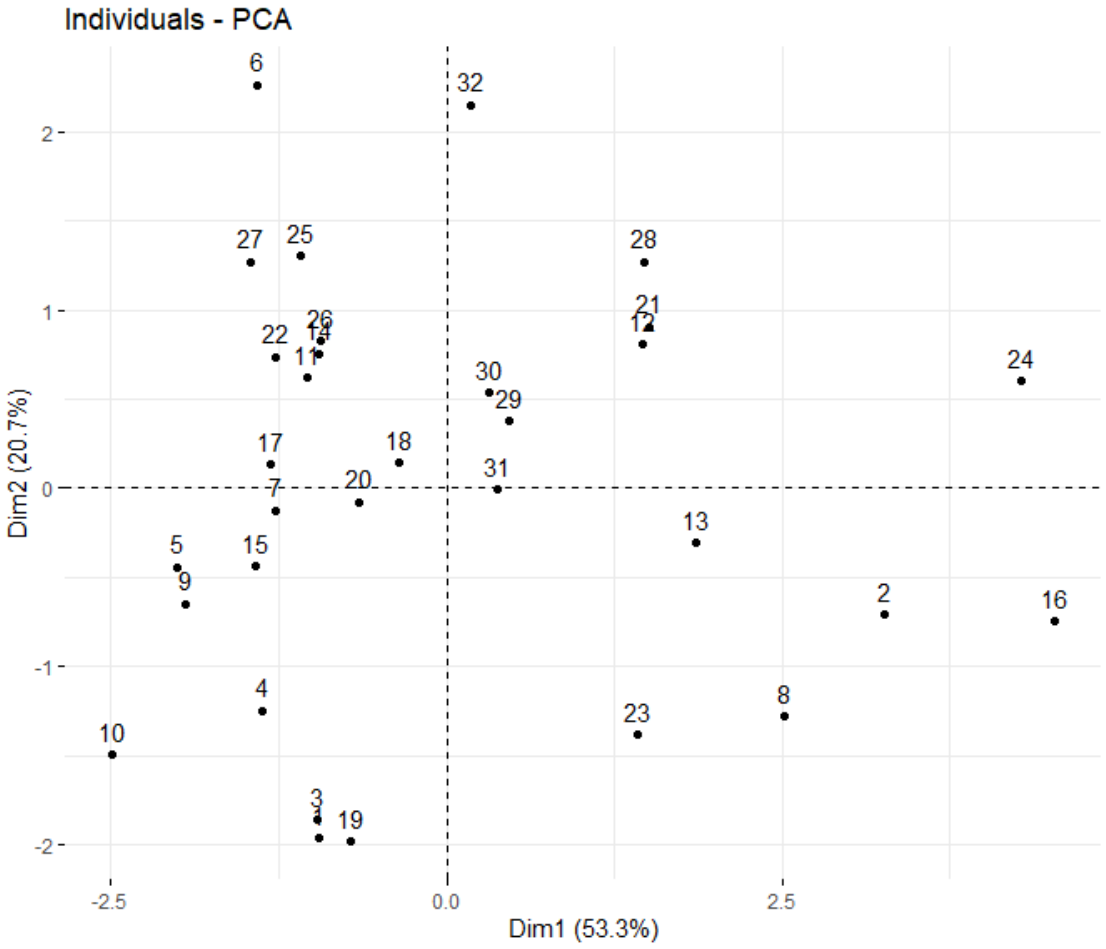
Identification des variables les plus significativement associées à une composante principale donnée.

```
## $Dim.1
##
## Link between the variable and the continuous variables (R-square)
##
================================================================================
====
##          correlation      p.value
## pistes     0.9540444 3.005797e-17
## prixforf   0.9303171 1.319383e-14
## remontee   0.9297367 1.488304e-14
## altmax     0.6500623 5.651902e-05
## kmfond     0.3619333 4.179647e-02
##
## $Dim.2
##
## Link between the variable and the continuous variables (R-square)
##
================================================================================
```

```
====
##        correlation      p.value
## altmin   0.8227049 7.576073e-09
## altmax   0.5309923 1.767126e-03
## kmfond  -0.5015475 3.450735e-03
```

On reprend le même procéder avec les individus.

```
## Principal Component Analysis Results for individuals
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the individuals"
## 2 "$cos2"    "Cos2 for the individuals"
## 3 "$contrib" "contributions of the individuals"
```

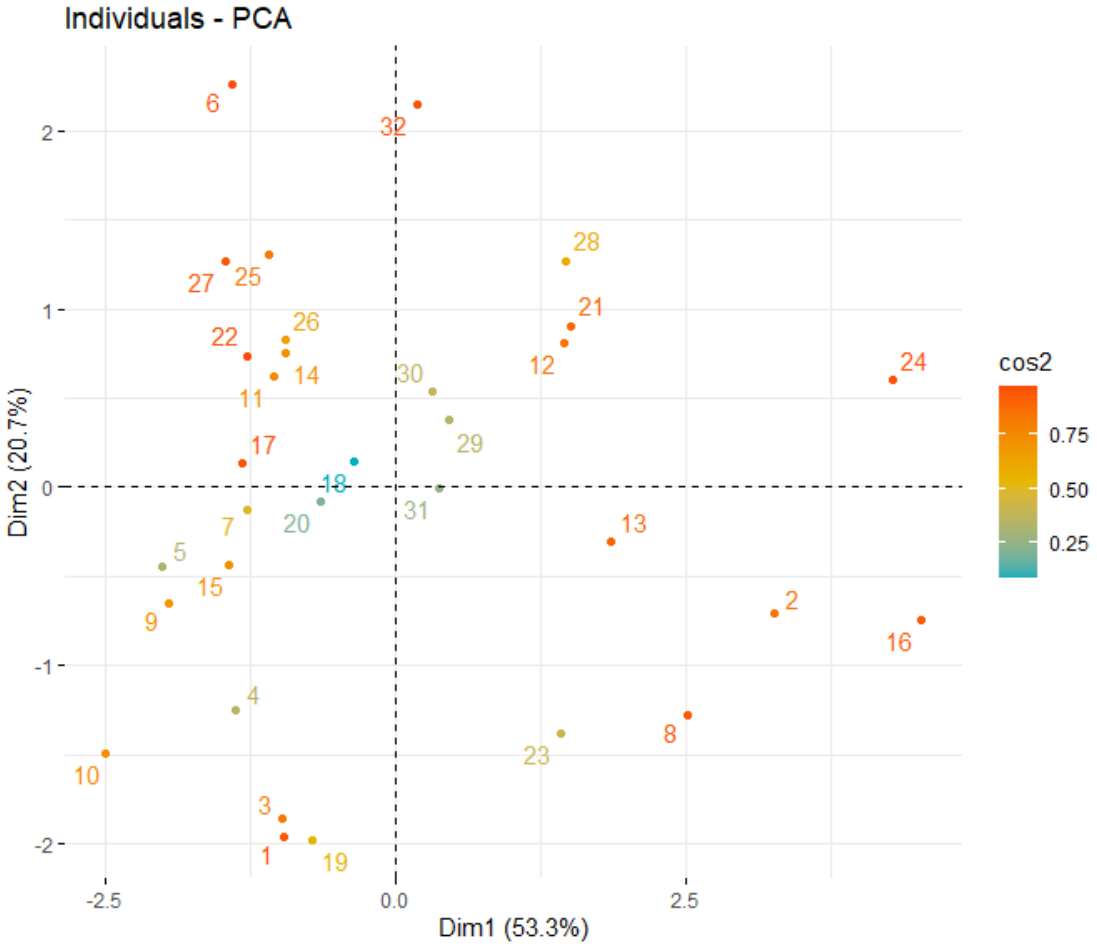


Un cos2 élevé indique une bonne représentation de la variable sur les axes principaux en considération.

- Un faible cos2 indique que la variable ou l'individu n'est pas parfaitement représenté par les axes principaux.

- Pour une variable ou un individu donné, la somme des cos2 sur toutes les composantes principales est égale à 1

On voit sur ce graphique que l'individu 18 est très mal représenté sur ces dimensions.



Chapitre 3

Ex 31

La classe de ces données USArrests est un data.frame.

Résultat avec les commandes princomp puis prcomp

```
##                Comp.1     Comp.2      Comp.3     Comp.4
## Alabama     0.9855659  1.1333924  0.44426879  0.1562671
## Alaska      1.9501378  1.0732133 -2.04000333 -0.4385834
## Arizona     1.7631635 -0.7459568 -0.05478082 -0.8346529
## Arkansas   -0.1414203  1.1197968 -0.11457369 -0.1828109
## California  2.5239801 -1.5429340 -0.59855680 -0.3419965

##                  PC1        PC2         PC3        PC4
## Alabama    -0.9756604  1.1220012 -0.43980366  0.1546966
## Alaska     -1.9305379  1.0624269  2.01950027 -0.4341755
## Arizona    -1.7454429 -0.7384595  0.05423025 -0.8262642
## Arkansas    0.1399989  1.1085423  0.11342217 -0.1809736
## California -2.4986128 -1.5274267  0.59254100 -0.3385592
```

3. La fonction gsvd réalise la décomposition en valeur singulières généralisée d'une matrice réelle Z de dimension n × p avec les métriques diagonales N = diag(r) sur R n et M = diag(c) sur R p . Le code de cette fonction est le suivant :

```
##               [,1]    [,2]    [,3]    [,4]
## Alabama    -0.9757  1.1220 -0.4398  0.1547
## Alaska     -1.9305  1.0624  2.0195 -0.4342
## Arizona    -1.7454 -0.7385  0.0542 -0.8263
## Arkansas    0.1400  1.1085  0.1134 -0.1810
## California -2.4986 -1.5274  0.5925 -0.3386
```

On obtient les mêmes résultats qu'avec prcomp.

```
##                Dim.1      Dim.2       Dim.3        Dim.4
## Alabama     0.9855659 -1.1333924  0.44426879  0.156267145
## Alaska      1.9501378 -1.0732133 -2.04000333 -0.438583440
## Arizona     1.7631635  0.7459568 -0.05478082 -0.834652924
## Arkansas   -0.1414203 -1.1197968 -0.11457369 -0.182810896
## California  2.5239801  1.5429340 -0.59855680 -0.341996478
## Colorado    1.5145629  0.9875551 -1.09500699  0.001464887
```

La fonction PCA donne les mêmes résultats que le princomp.

Ex 32

```
##      CAMP HOTEL LOCA RESI
## AGRI  239   155  129    0
## CADR 1003  1556 1821 1521
## INAC  682  1944  967 1333
## OUVR 2594  1124 2176 1038
```

Test du chi 2, on remarque que la p-value est inférieur à 0.05, la valeur seuil donc on peut supposer la non indépendance.

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 2067.9, df = 9, p-value < 2.2e-16
```

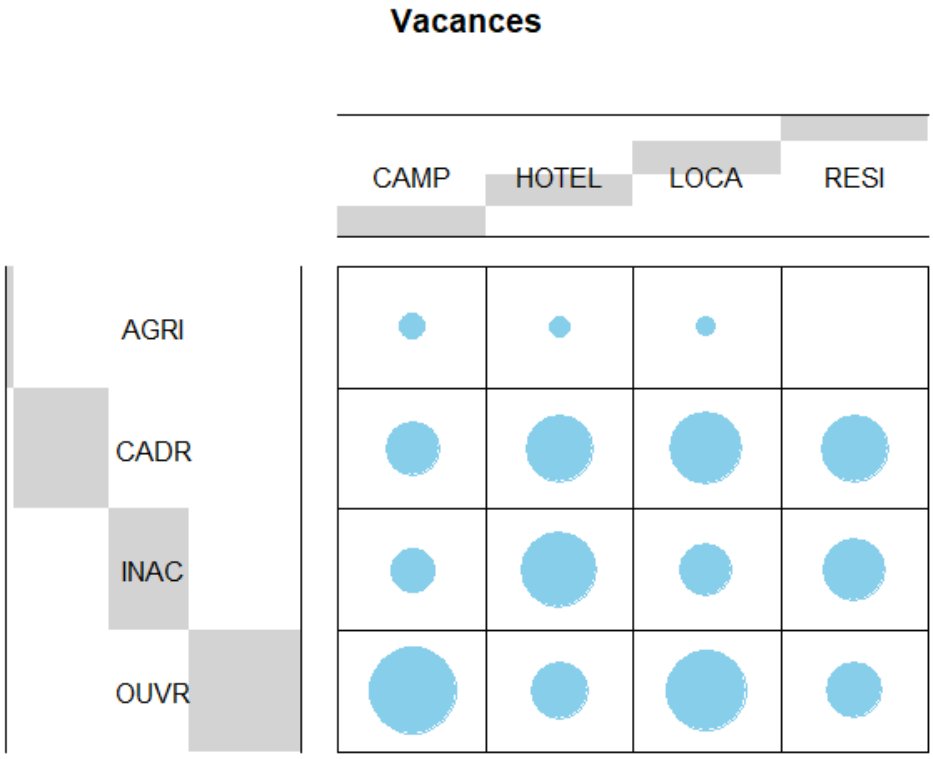
Voici les profils lignes puis colonnes

```
##           CAMP     HOTEL      LOCA      RESI
## AGRI 0.4569790 0.2963671 0.2466539 0.0000000
## CADR 0.1699712 0.2636841 0.3085918 0.2577529
## INAC 0.1384490 0.3946407 0.1963053 0.2706050
## OUVR 0.3742066 0.1621466 0.3139065 0.1497403
```

```
##            CAMP      HOTEL       LOCA      RESI
## AGRI 0.05289951 0.03243356 0.02532888 0.0000000
## CADR 0.22200089 0.32559113 0.35754958 0.3908016
## INAC 0.15095175 0.40677966 0.18986845 0.3424974
## OUVR 0.57414785 0.23519565 0.42725309 0.2667009
```

4 . Faire une AFC.

On voit sur ce graphique la répartition des individus en fonction des deux varibles.



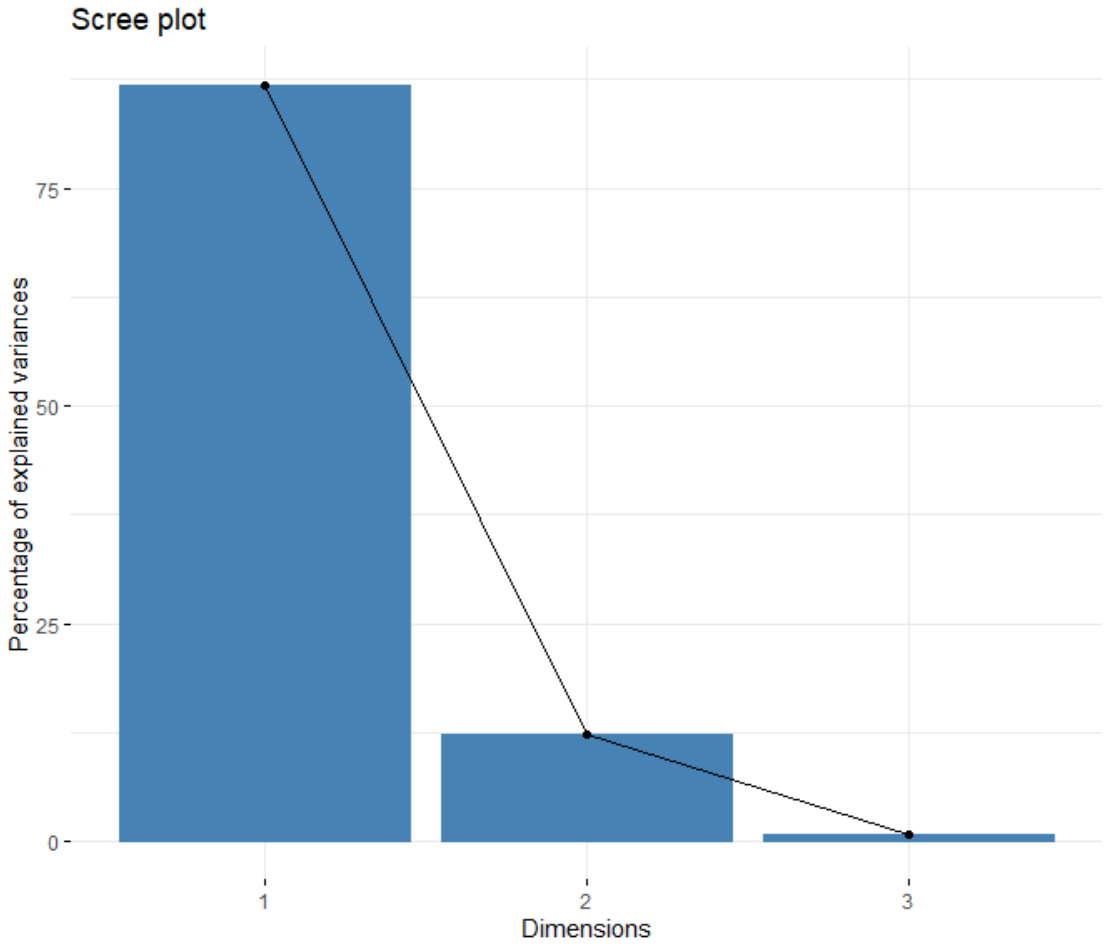
On fait une AFC sur les données :

```
## **Results of the Correspondence Analysis (CA)**
## The row variable has  4  categories; the column variable has 4 categories
## The chi square of independence between the two variables is equal to
2067.911 (p-value =  0 ).
## *The results are available in the following objects:
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$col"            "results for the columns"
## 3  "$col$coord"      "coord. for the columns"
## 4  "$col$cos2"       "cos2 for the columns"
## 5  "$col$contrib"    "contributions of the columns"
## 6  "$row"            "results for the rows"
## 7  "$row$coord"      "coord. for the rows"
## 8  "$row$cos2"       "cos2 for the rows"
## 9  "$row$contrib"    "contributions of the rows"
## 10 "$call"           "summary called parameters"
## 11 "$call$marge.col" "weights of the columns"
## 12 "$call$marge.row" "weights of the rows"
```

On extrait les valeurs propres expliquées par chaque axe principales

```
##        eigenvalue percentage of variance cumulative percentage of variance
## dim 1 0.098243388             86.8550689                           86.85507
## dim 2 0.013863055             12.2560576                           99.11113
## dim 3 0.001005421              0.8888735                          100.00000
```

Visualisation des valeurs propres, on voit sur ce graphique qu'une seule dimension suffit pour représenter plus de 75 pourcent des informations, avec deux dimensions on obtient presque 100%.



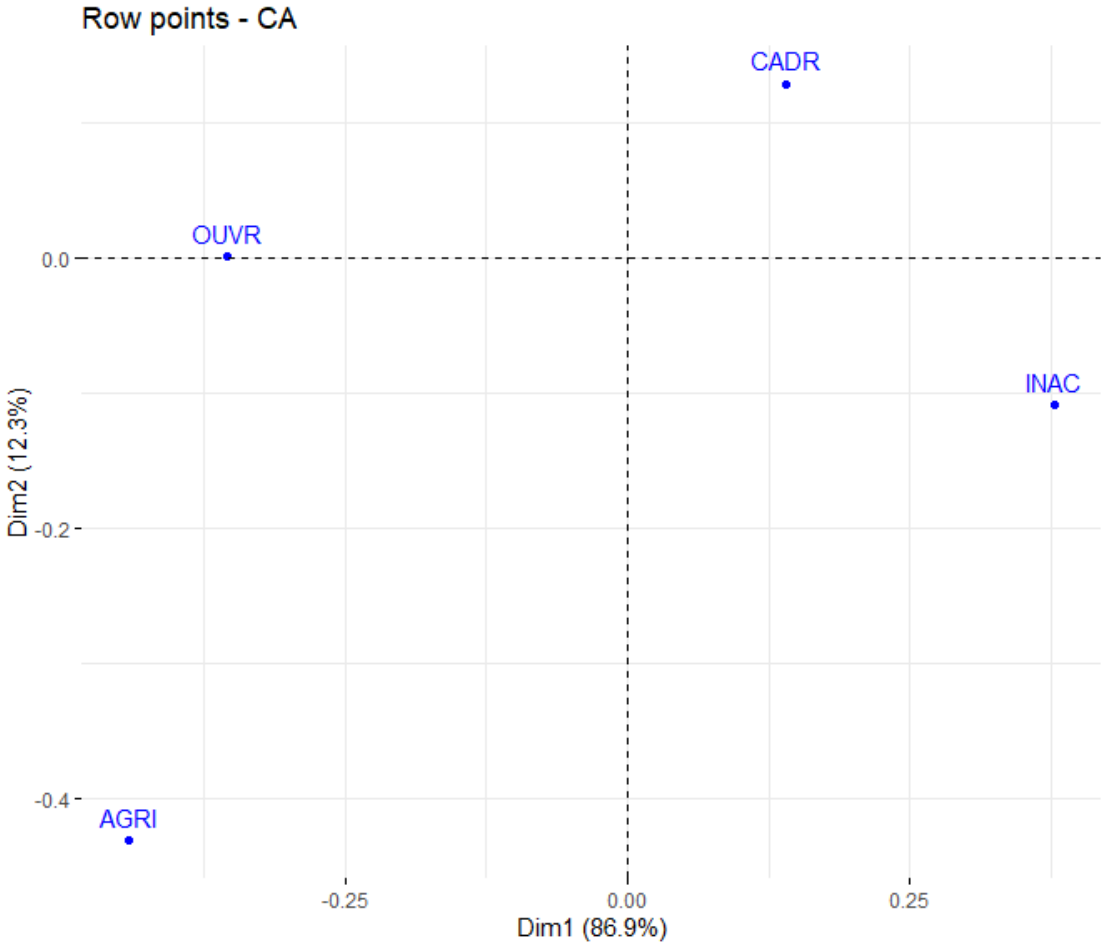
On extrait les résultats pour les lignes.

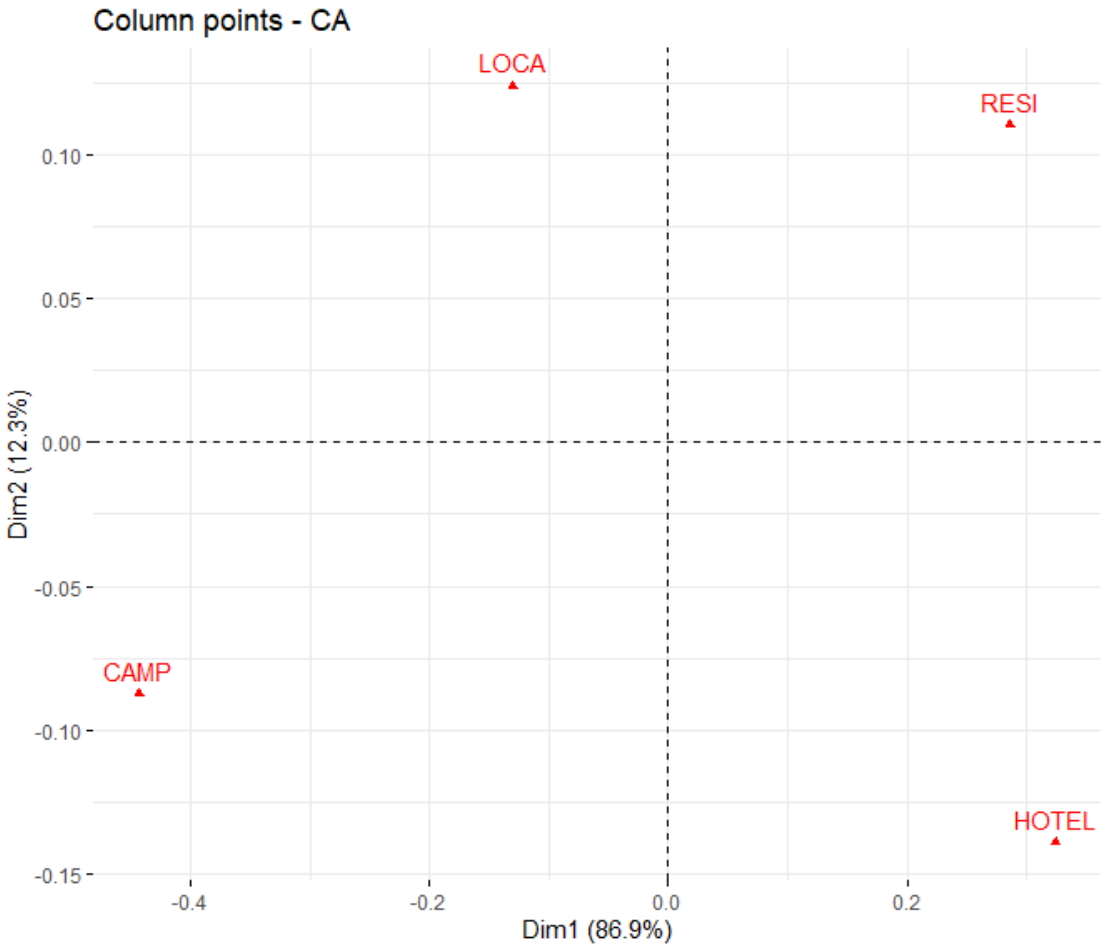
```
##          Dim 1        Dim 2       Dim 3
## AGRI 0.4880138 4.652482e-01 0.046738075
## CADR 0.5318771 4.487886e-01 0.019334324
## INAC 0.9207832 7.668811e-02 0.002528682
## OUVR 0.9971609 4.645805e-06 0.002834459

##          Dim 1        Dim 2    Dim 3
## AGRI  5.675896 38.346999815 53.11637
## CADR  6.430372 38.451297804 22.84068
```

```
## INAC 39.307445 23.200098198 10.54792
## OUVR 48.586287  0.001604183 13.49503
```

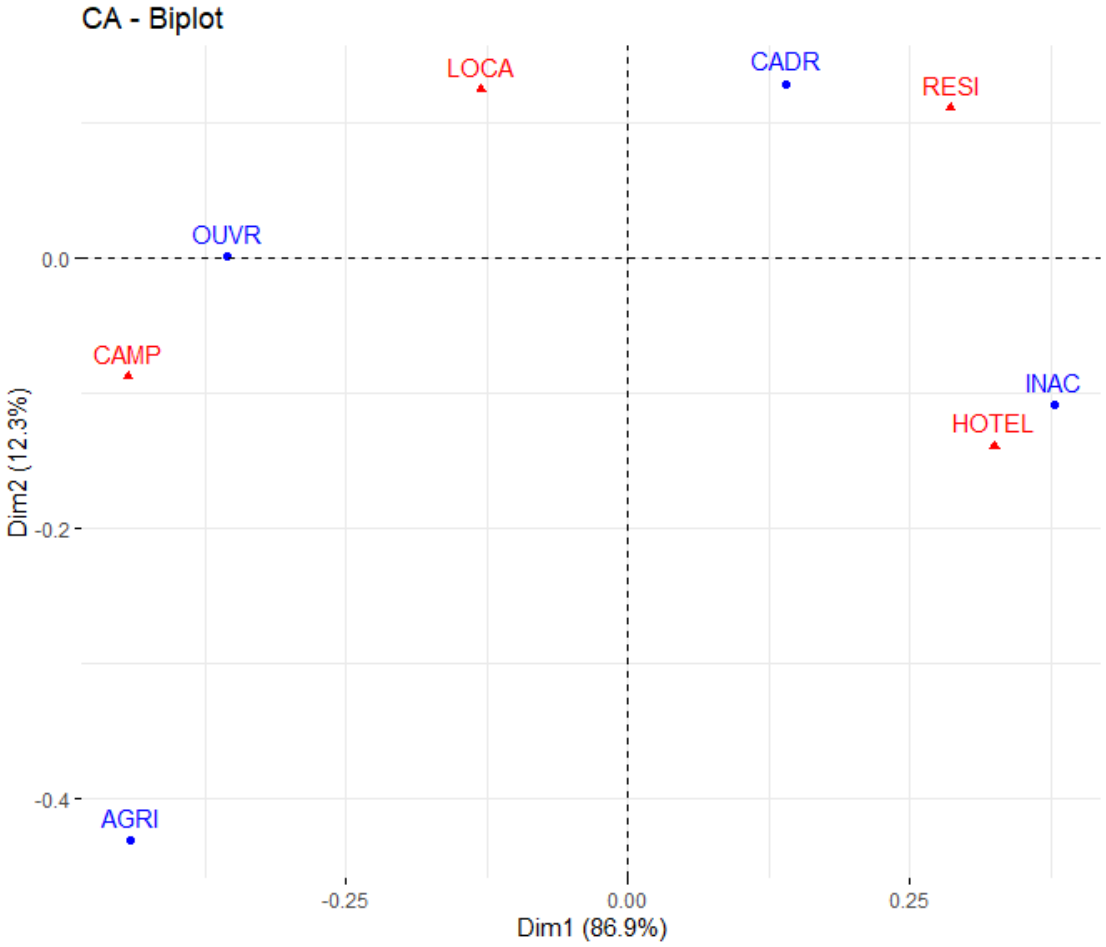
Les valeurs de cos2 sont comprises entre 0 et 1. La somme des cos2 pour les lignes sur toutes les dimensions de l'AFC est égale à 1. La qualité de représentation d'une ligne ou d'une colonne dans n dimensions est simplement la somme des cosinus carré de cette ligne ou colonne sur les n dimensions.

Visualisation des résultats pour les lignes et les colonnes, respectivement.

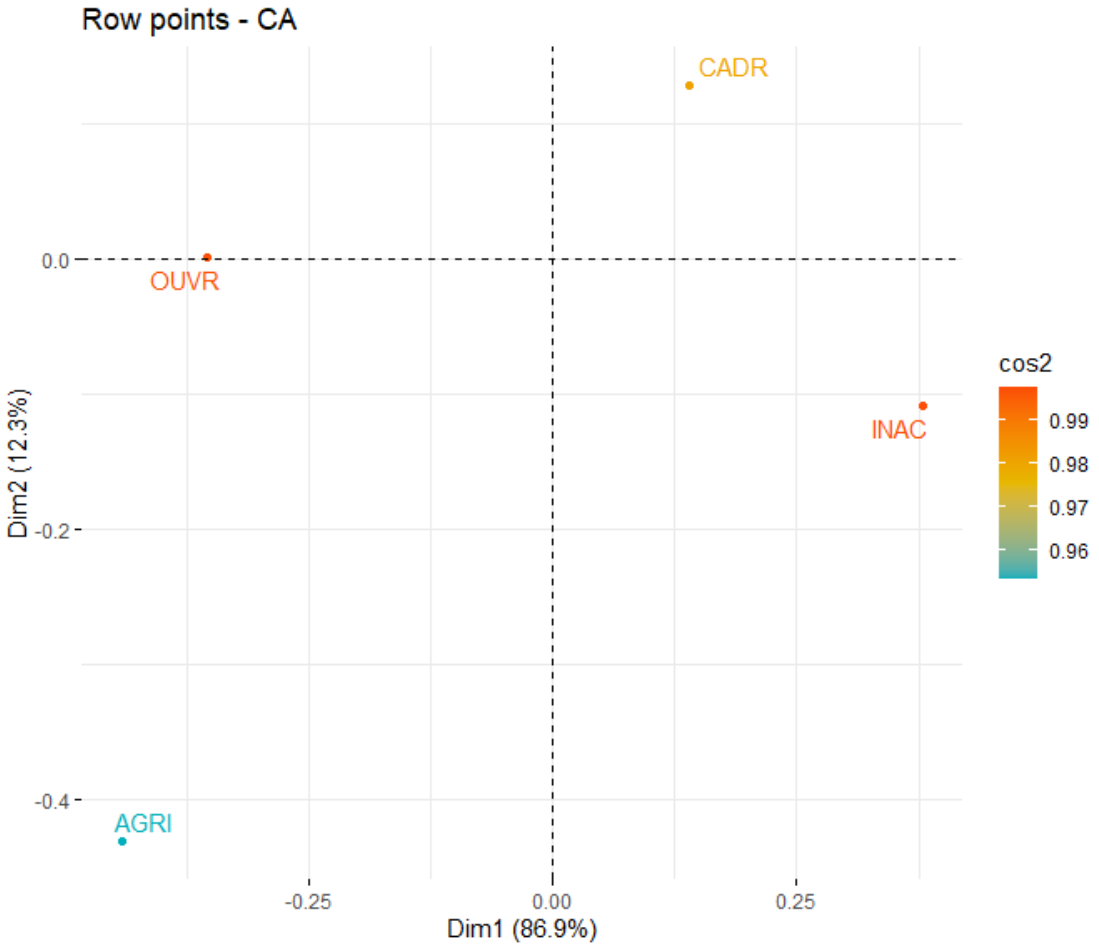



Column points - CA
LOCA
RESI
CAMP
HOTEL
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.4
-0.2
0.0
0.2
Dim2 (12.3%)
Dim1 (86.9%)

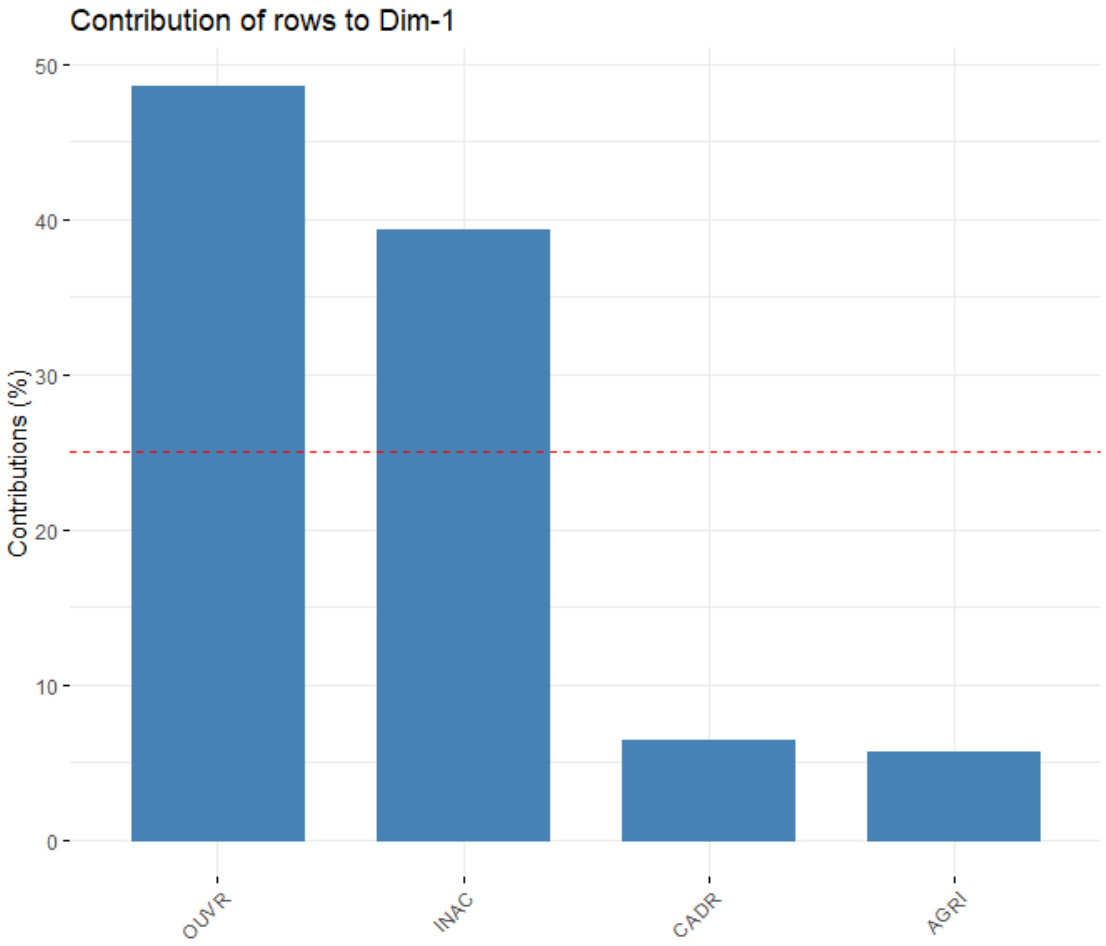
Superposition des deux graphiques précédents :

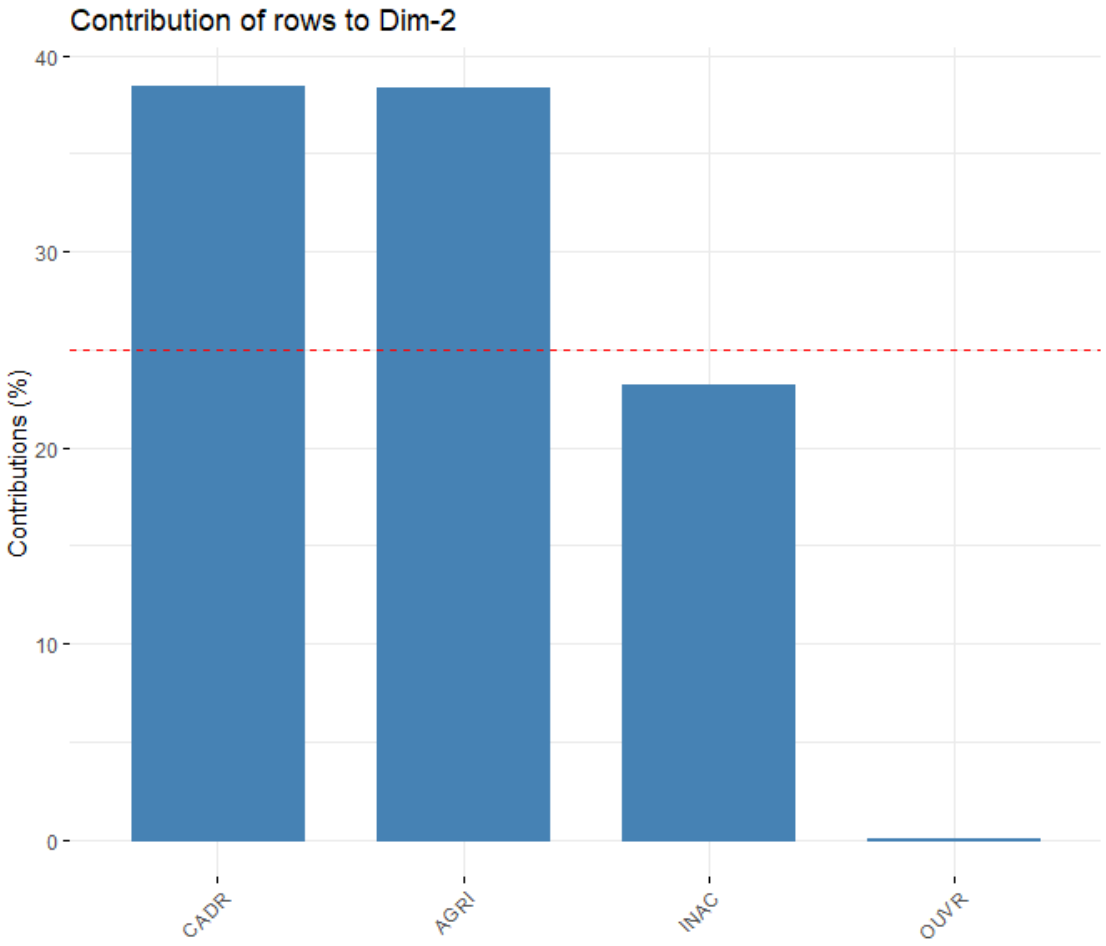


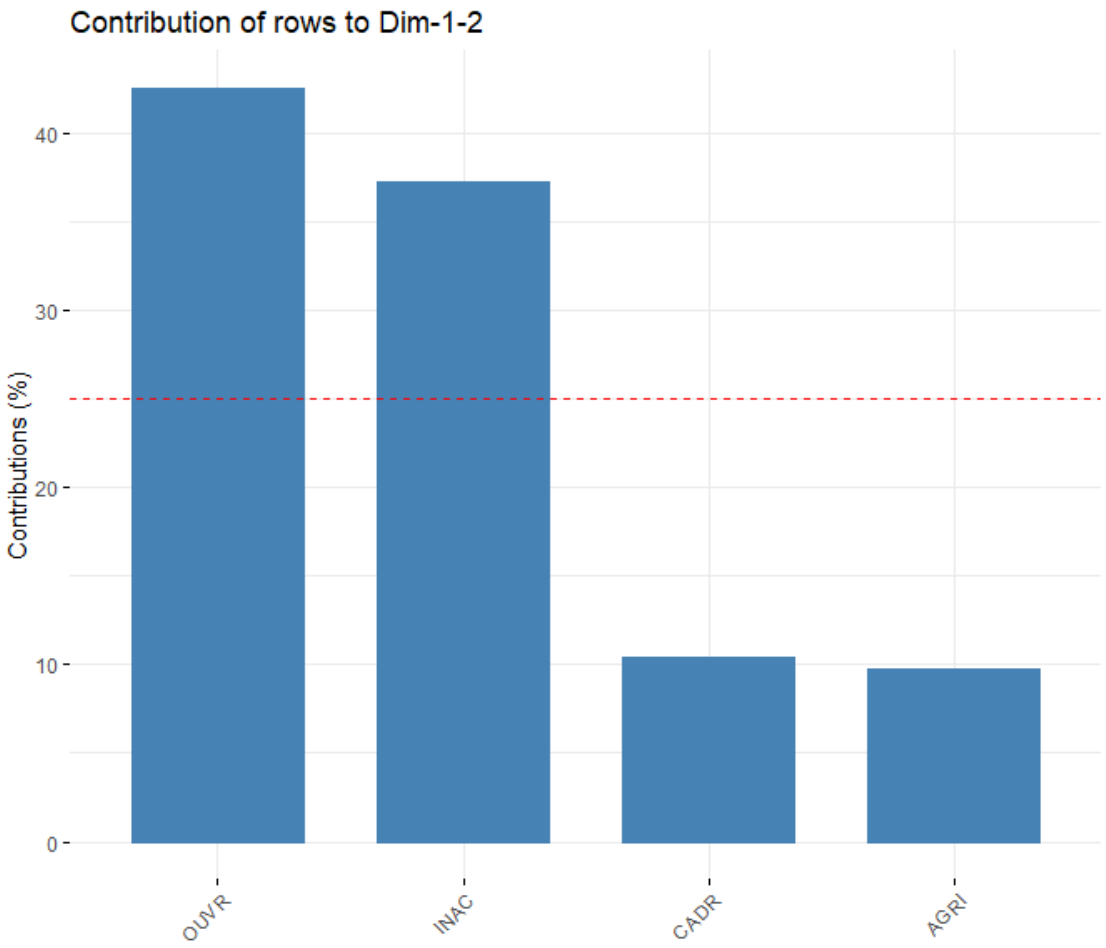
Ce graphique représente la contribution des individus aux dimensions grâce à un code couleur. On voit que les agriculteurs contribuent le moins à celles-ci.



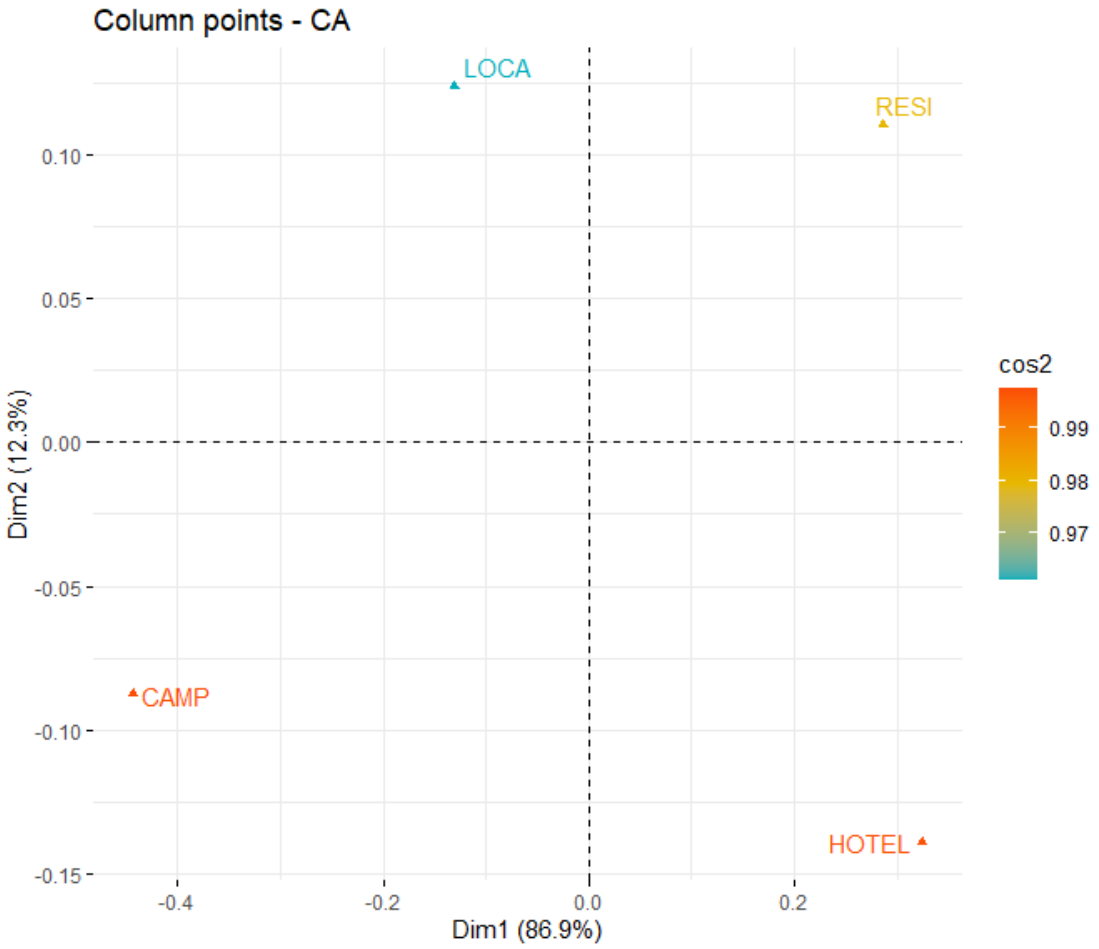
De façon différente on peut lire les mêmes informations en quelque peut plus détailler. Les ouvriers et inactifs contribuent le plus à la dimension 1 et les cadres et agriculteurs contribuent le plus à la dimension 2. Les ouvriers et inactifs contribuent le plus aux deux dimensions.

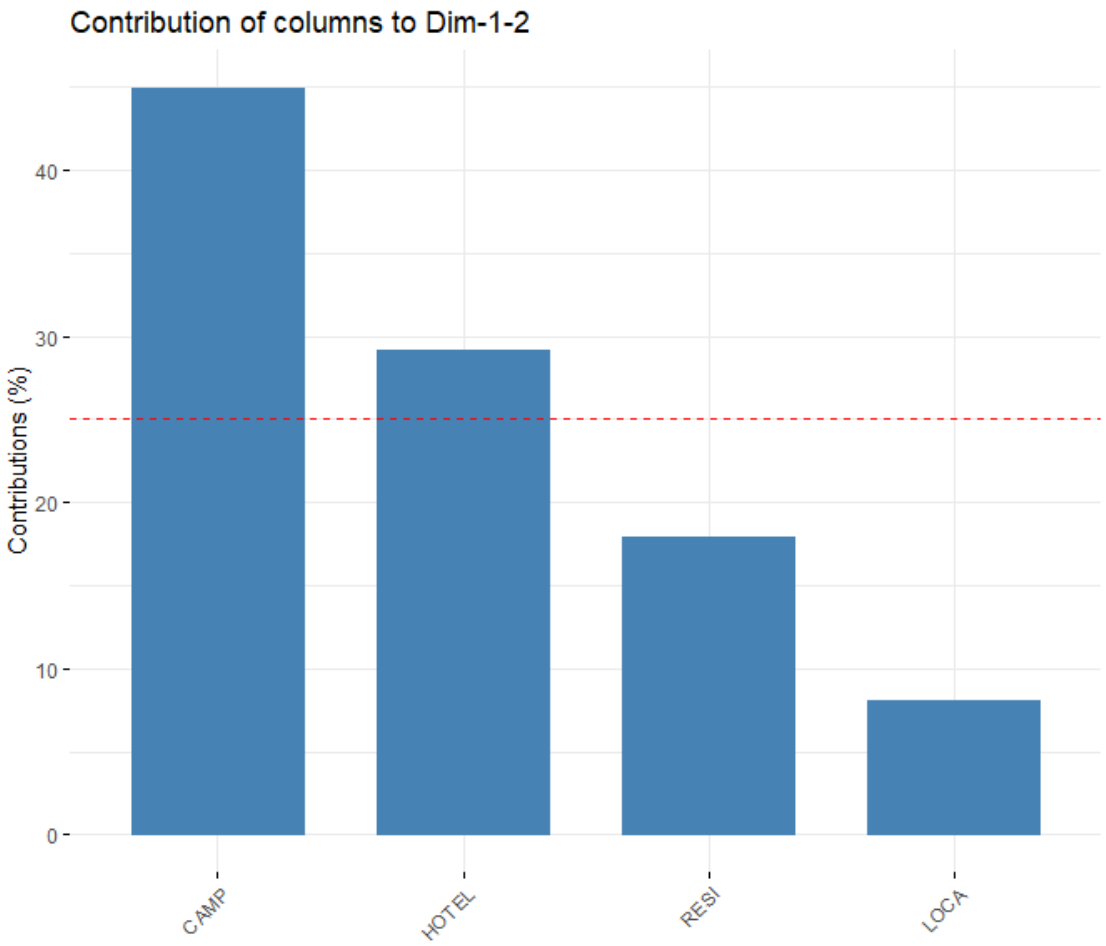

Contribution of rows to Dim-1
Contributions (%)
50
40
30
20
10
0
OUVR
INAC
CADR
AGRI


Contribution of rows to Dim-2
Contributions (%)
40
30
20
10
0
CADR
AGRI
INAC
OUVR

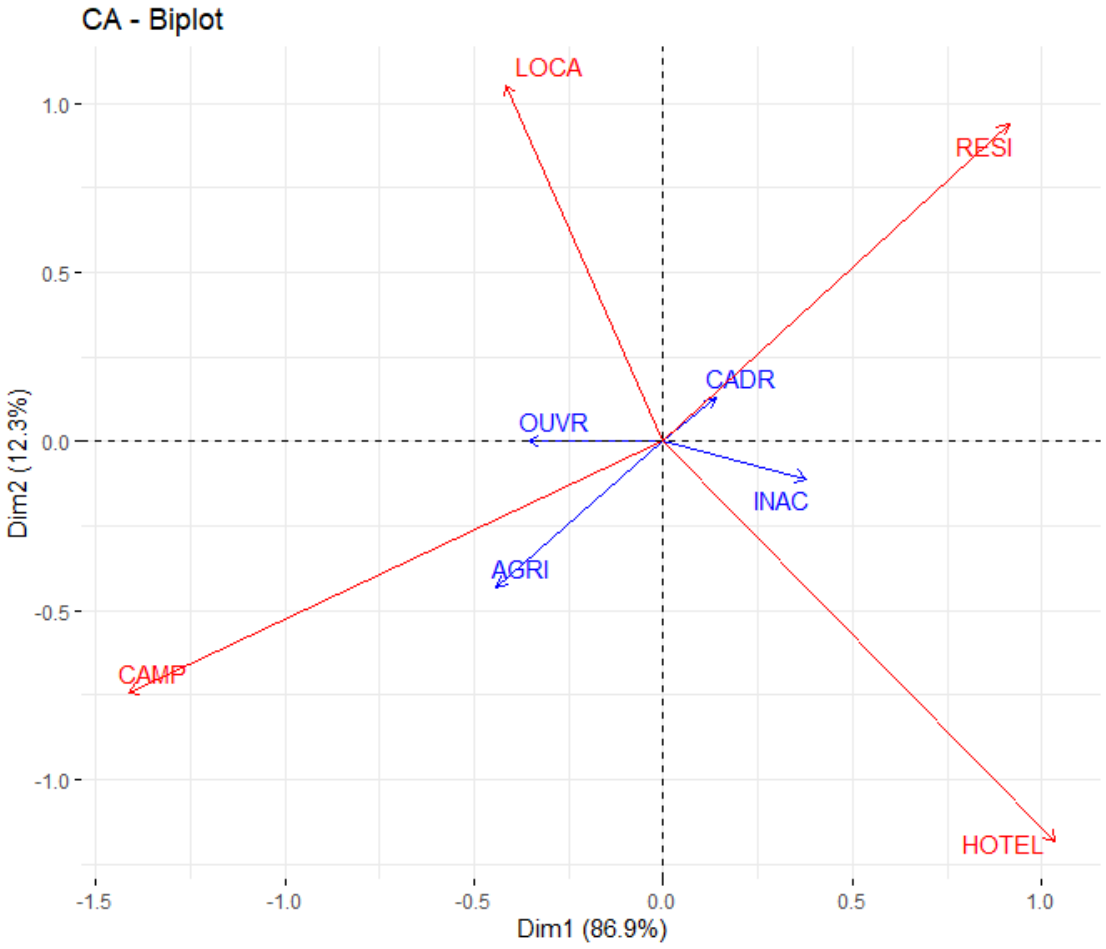


Du côté des colonnes, ce sont les locations qui contribuent le moins aux dimensions. Le camping et les hôtels sont bien représentés dans la dimension une et deux.


Column points - CA
LOCA
RESI
CAMP
HOTEL
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.4
-0.2
0.0
0.2
Dim2 (12.3%)
Dim1 (86.9%)
cos2
0.99
0.98
0.97

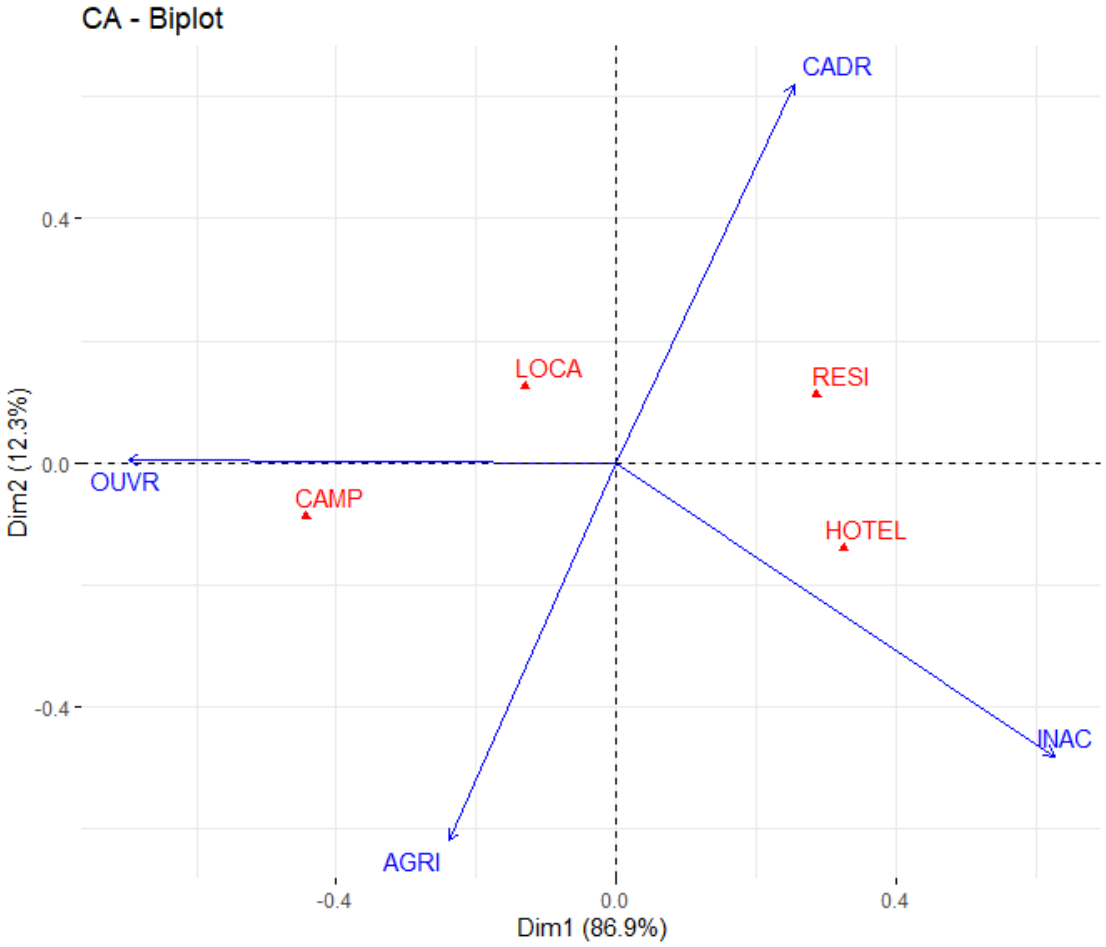

Contribution of columns to Dim-1-2
Contributions (%)
40
30
20
10
0
CAMP
HOTEL
RESI
LOCA

Si l'angle entre deux flèches est aigu, alors il y a une forte association entre les lignes et les colonnes correspondantes.



Cependant, les distances entre les points lignes et l'origine du graphique sont liées à leurs contributions aux axes principaux en considération. Plus une flèche est proche (en termes de distance angulaire) d'un axe, plus la contribution de la ligne sur cet axe par rapport à l'autre axe est importante. Si la flèche est à mi-chemin entre les deux axes, la ligne contribue aux deux axes de manière identique (c'est pratiquement le cas pour les individus inactifs et

c’est le cas pour les cadres et agriculteurs même s’ils contribuent peu).



Description des dimensions. Les lignes/colonnes sont triées en fonction de leurs coordonnées. Pour la dimension 1 :

```
##           coord
## AGRI -0.4414992
## OUVR -0.3548061
## CADR  0.1399003
## INAC  0.3785766

##            coord
## CAMP  -0.4430269
## LOCA  -0.1304042
## RESI   0.2862054
## HOTEL  0.3247191
```

Pour la dimension 2 :

```
##               coord
## AGRI -0.4310783200
## INAC -0.1092544256
```

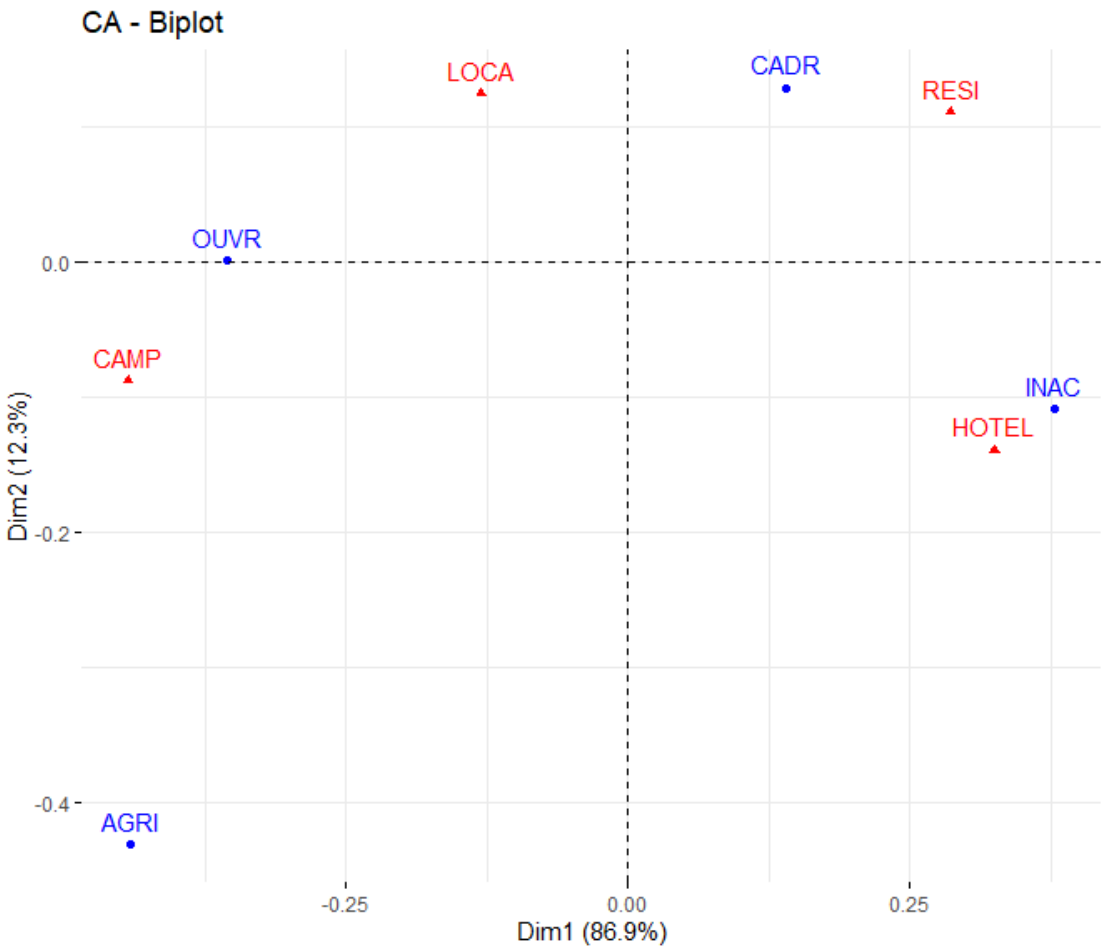
```
## OUVR  0.0007658417
## CADR  0.1285091421

##             coord
## HOTEL -0.13930796
## CAMP  -0.08771021
## RESI   0.11046638
## LOCA   0.12411002
```

4. Vérifier que la statistique du khi-deux égale la somme des valeurs propres multipliée par n. On n'observe pas ce phénomène ici. En revanche, la p-value est inférieur à 0.05 on peut donc penser que les données sont non indépendantes. Ceci parait logique puisque le type de logement de vacances dépend en parti des revenus des individus.

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1 3.19906023              53.317670                           53.31767
## comp 2 1.24354998              20.725833                           74.04350
## comp 3 0.85138961              14.189827                           88.23333
## comp 4 0.46765198               7.794200                           96.02753
## comp 5 0.15977748               2.662958                           98.69049
## comp 6 0.07857073               1.309512                          100.00000

## [1] 2465.25

##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 2067.9, df = 9, p-value < 2.2e-16
```

7. Y'a-t-il un effet Guttman ? Commenter.

Reprenons le graphique des modalités. Il permet de voir que les modalités sont disposées en arc de cercle. Il apparaît quand un ordre sous-tend les modalités.



On observe donc un effet Guttman.

Ici, l'ordre est le suivant : hotel, inac,resi, cadr, loca, ouvr,camp, agri.

Ex 33

Voici les données de cet exercice :

```
##    none light medium heavy
## SM    4     2      3     2
## JM    4     3      7     4
## SE   25    10     12     4
## JE   18    24     33    13
## SC   10     6      7     2
```

2. AFC et SVD généralisée.

(a) Construire la matrice F des fréquences, les vecteurs r et c des distributions marginales et la matrice Z des écarts à l'indépendance.

```
##     SM     JM     SE     JE     SC
## 0.0570 0.0933 0.2642 0.4560 0.1295
```

```
##      none     light    medium     heavy
## 0.3160622 0.2331606 0.3212435 0.1295337
```

(b) Calculer avec la fonction gsvd les matrices X et Y et d des coordonnées factorielles des profil-lignes et colonnes de l'AFC.

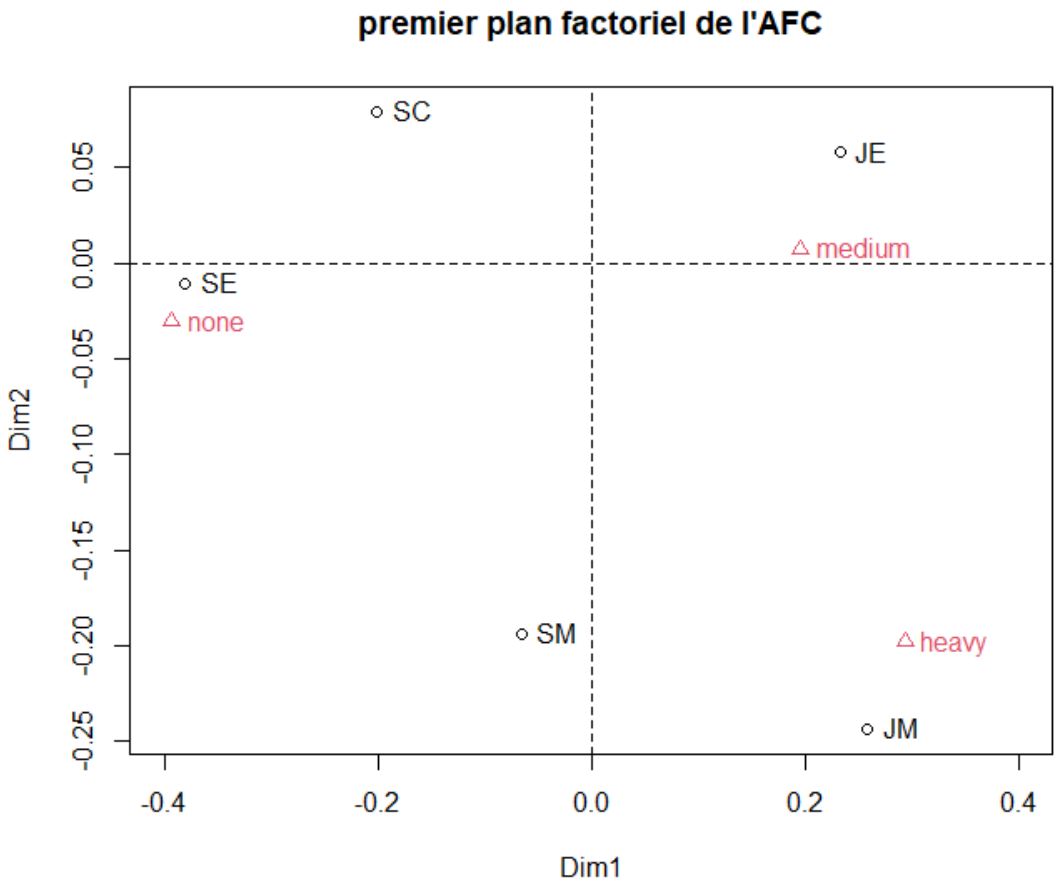
Voici la matrice X :

```
##           dim1        dim2         dim3
## SM -0.06576838 -0.19373700  0.070981028
## JM  0.25895842 -0.24330457 -0.033705190
## SE -0.38059489 -0.01065991 -0.005155757
## JE  0.23295191  0.05774391  0.003305371
## SC -0.20108912  0.07891123 -0.008081076
```

Voici la matrice Y :

```
##               dim1         dim2          dim3
## none   -0.39330845 -0.030492071 -0.0008904827
## light   0.09945592  0.141064289  0.0219980349
## medium  0.19632096  0.007359109 -0.0256590867
## heavy   0.29377599 -0.197765656  0.0262108499
```

(c) Représenter avec la fonction plot les profil-lignes et les profil-colonnes sur le premier plan factoriel de l'AFC.



(d) Quel est le pourcentage d'inertie expliquée par le premier plan factoriel de l'AFC pourcentages d'inertie des axes : Les pourcentages sont les suivant pour Dim 1 et Dim 2 respectivement :

```
## [1] 87.75587 11.75865
```

Voici le pourcentage d'inertie du plan :

```
## [1] 99.51453
```

3. Retrouver ces résultats avec le package FactoMineR et la fonction CA.

On obtient les valeurs propres. 2 dimensions suffisent pour obtenir 99.5% de représentation des données.

```
##        eigenvalue percentage of variance cumulative percentage of variance
## dim 1 0.0747591059             87.7558731                           87.75587
## dim 2 0.0100171805             11.7586535                           99.51453
## dim 3 0.0004135741              0.4854734                          100.00000
```

La matrice de res.carowcoord :

```
##           Dim 1       Dim 2        Dim 3
## SM -0.06576838  0.19373700  0.070981028
```

```
## JM  0.25895842  0.24330457 -0.033705190
## SE -0.38059489  0.01065991 -0.005155757
## JE  0.23295191 -0.05774391  0.003305371
## SC -0.20108912 -0.07891123 -0.008081076
```

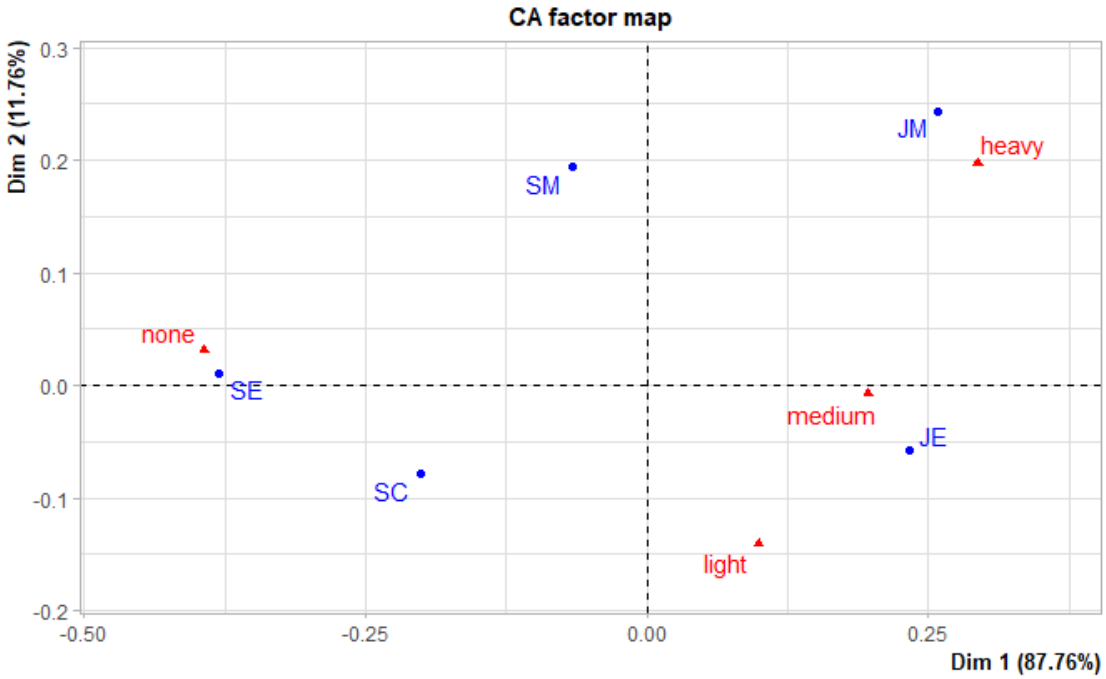
```
##           dim1        dim2         dim3
## SM -0.06576838 -0.19373700  0.070981028
## JM  0.25895842 -0.24330457 -0.033705190
## SE -0.38059489 -0.01065991 -0.005155757
## JE  0.23295191  0.05774391  0.003305371
## SC -0.20108912  0.07891123 -0.008081076
```

La matrice de res.cacolcoord :

```
##              Dim 1        Dim 2         Dim 3
## none   -0.39330845  0.030492071 -0.0008904827
## light   0.09945592 -0.141064289  0.0219980349
## medium  0.19632096 -0.007359109 -0.0256590867
## heavy   0.29377599  0.197765656  0.0262108499
```

La matrice Y : On peut noter que c'est la même que la précédente matrice

```
##               dim1         dim2          dim3
## none   -0.39330845 -0.030492071 -0.0008904827
## light   0.09945592  0.141064289  0.0219980349
## medium  0.19632096  0.007359109 -0.0256590867
## heavy   0.29377599 -0.197765656  0.0262108499
```



Ex 34

Voici les données de cet exercice :

```
##       B  C  D  F  G  H  I  L  M  N  P  R  S  U  W  Y
## CD1  34 37 44 27 19 39 74 44 27 61 12 65 69 22 14 21
## CD2  18 33 47 24 14 38 66 41 36 72 15 62 63 31 12 18
## CD3  32 43 36 12 21 51 75 33 23 60 24 68 85 18 13 14
## RD1  13 31 55 29 15 62 74 43 28 73  8 59 54 32 19 20
## RD2   8 28 34 24 17 68 75 34 25 70 16 56 72 31 14 11
## RD3   9 34 43 25 18 68 84 25 32 76 14 69 64 27 11 18
```

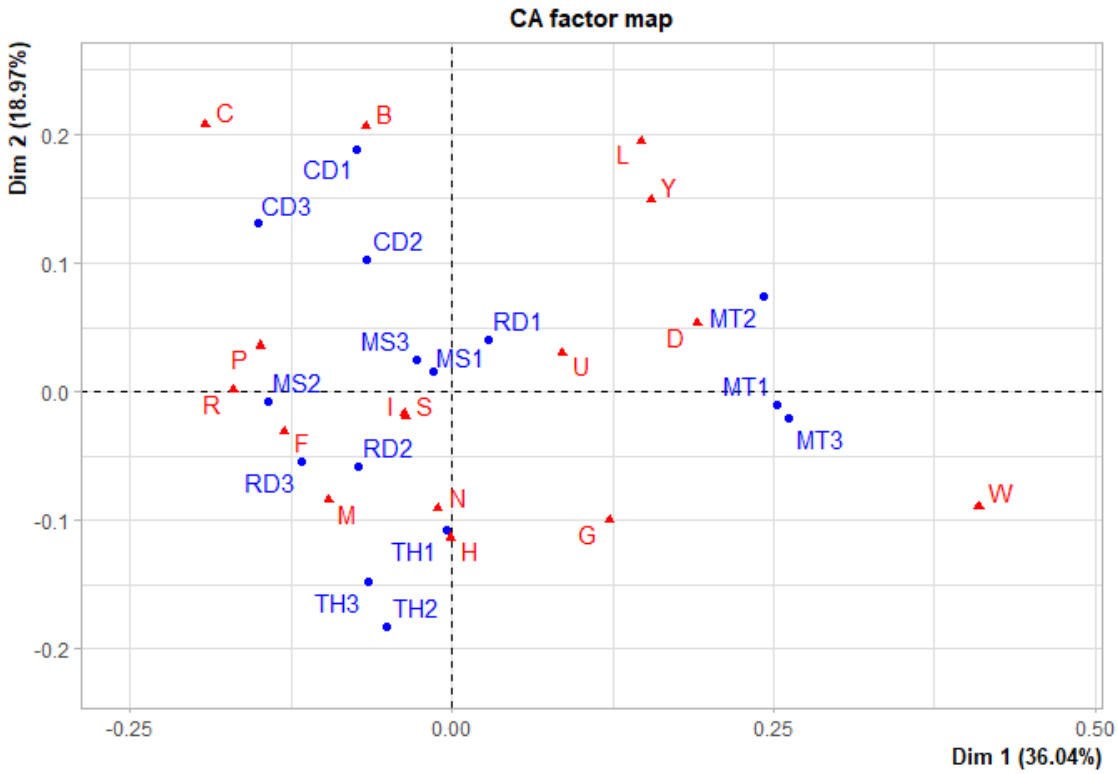
2. On considère dans un premier temps le tableau de contingence des 15 échantillons dont on connaît les auteurs. Effecter un test du χ 2 d'indépendence.

On transforme les données pour n'utiliser que les données quantitatives. La p-value < 0.05 donc il y a indépendance.

```
##       B  C  D  F  G  H  I  L  M  N  P  R  S  U  W  Y
## CD1  34 37 44 27 19 39 74 44 27 61 12 65 69 22 14 21
## CD2  18 33 47 24 14 38 66 41 36 72 15 62 63 31 12 18
## CD3  32 43 36 12 21 51 75 33 23 60 24 68 85 18 13 14
```

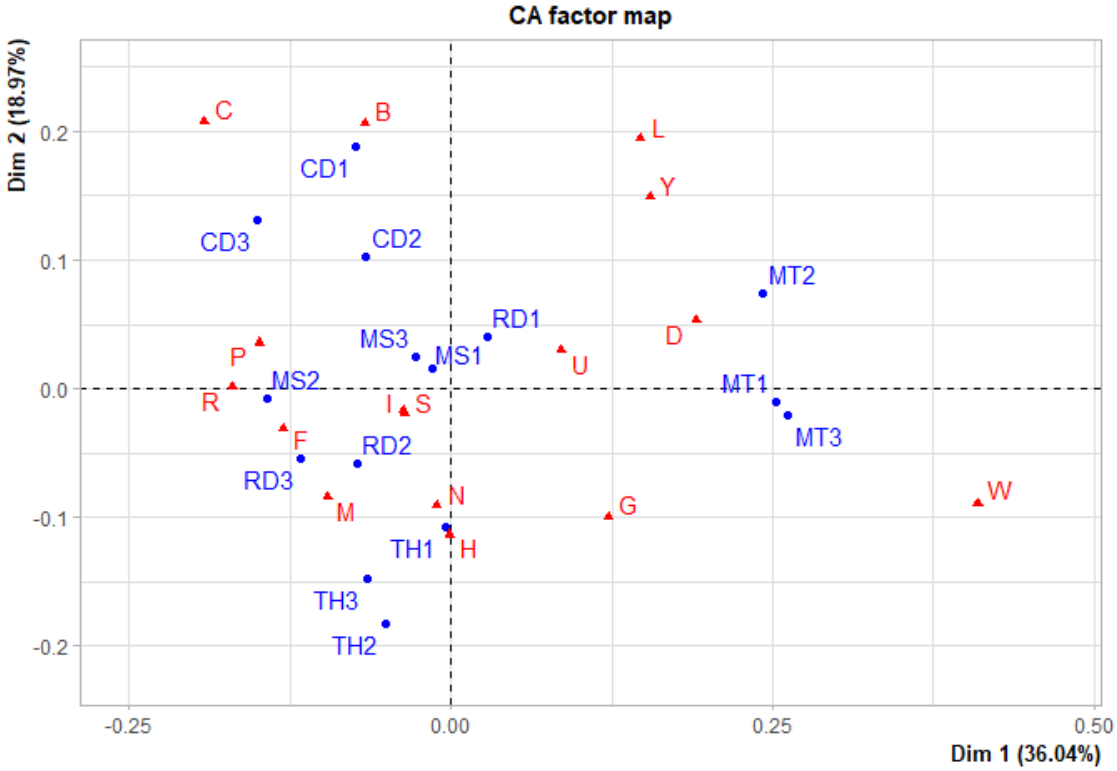
```
## RD1 13 31 55 29 15 62 74 43 28 73  8 59 54 32 19 20
## RD2  8 28 34 24 17 68 75 34 25 70 16 56 72 31 14 11
## RD3  9 34 43 25 18 68 84 25 32 76 14 69 64 27 11 18
## TH1 15 20 28 18 19 65 82 34 29 89 11 47 74 18 22 17
## TH2 18 14 40 25 21 60 70 15 37 80 15 65 68 21 25  9
## TH3 19 18 41 26 29 58 64 18 38 78 15 65 72 20 20 11
## MS1 13 29 49 31 16 61 73 36 29 69 13 63 58 18 20 25
## MS2 17 34 43 29 14 62 64 26 26 71 26 78 64 21 18 12
## MS3 13 22 43 16 11 70 68 46 35 57 30 71 57 19 22 20
## MT1 16 18 56 13 27 67 61 43 20 63 14 43 67 34 41 23
## MT2 15 21 66 21 19 50 62 50 24 68 14 40 58 31 36 26
## MT3 19 17 70 12 28 53 72 39 22 71 11 40 67 20 41 17

##
##  Pearson's Chi-squared test
##
## data:  data.active
## X-squared = 455.18, df = 210, p-value < 2.2e-16
```

3. Effectuer une AFC avec la fonction la fonction CA de FactoMineR et interpréter les résultats.

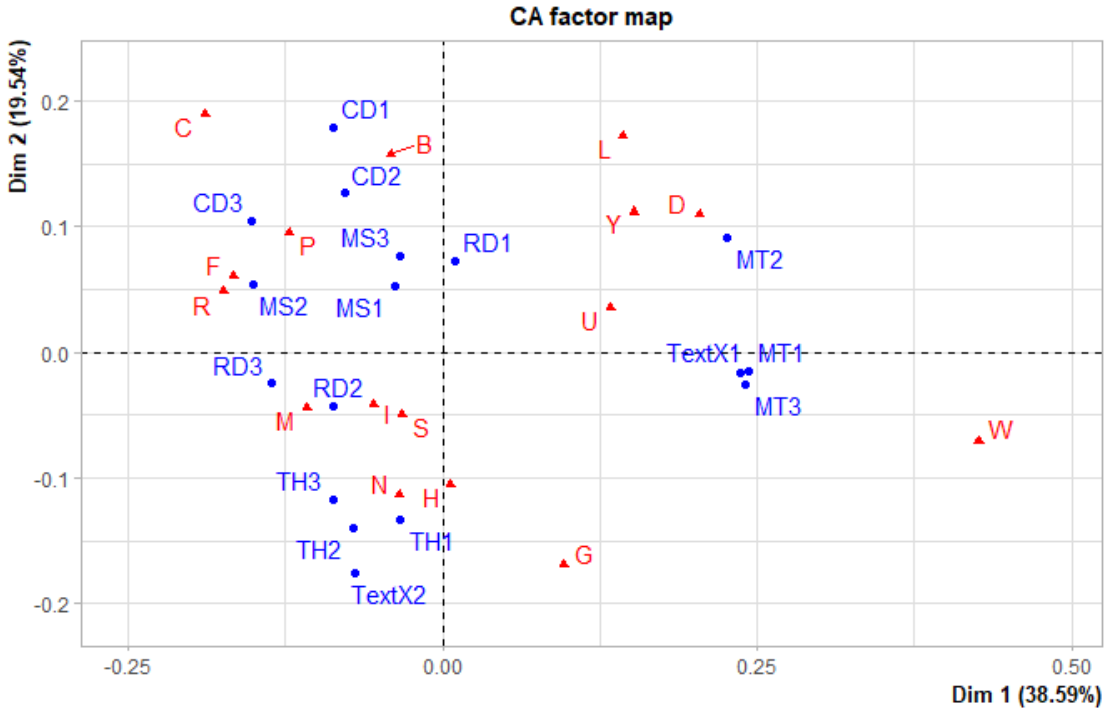


On extrait les valeurs propres, on remarque que deux dimensions nous donnent 55% de représentation des données.

```
##          eigenvalue percentage of variance cumulative percentage of variance
## dim 1  1.822781e-02             36.03691768                           36.03692
## dim 2  9.593661e-03             18.96694646                           55.00386
## dim 3  7.585251e-03             14.99626122                           70.00013
## dim 4  5.363102e-03             10.60300781                           80.60313
## dim 5  3.577226e-03              7.07227892                           87.67541
## dim 6  2.110781e-03              4.17307615                           91.84849
## dim 7  1.592456e-03              3.14833215                           94.99682
## dim 8  9.174775e-04              1.81387975                           96.81070
## dim 9  7.317600e-04              1.44671077                           98.25741
## dim 10 4.200503e-04              0.83045166                           99.08786
## dim 11 2.956366e-04              0.58448220                           99.67234
## dim 12 1.189202e-04              0.23510874                           99.90745
## dim 13 2.585120e-05              0.05110858                           99.95856
## dim 14 2.095969e-05              0.04143792                          100.00000
```

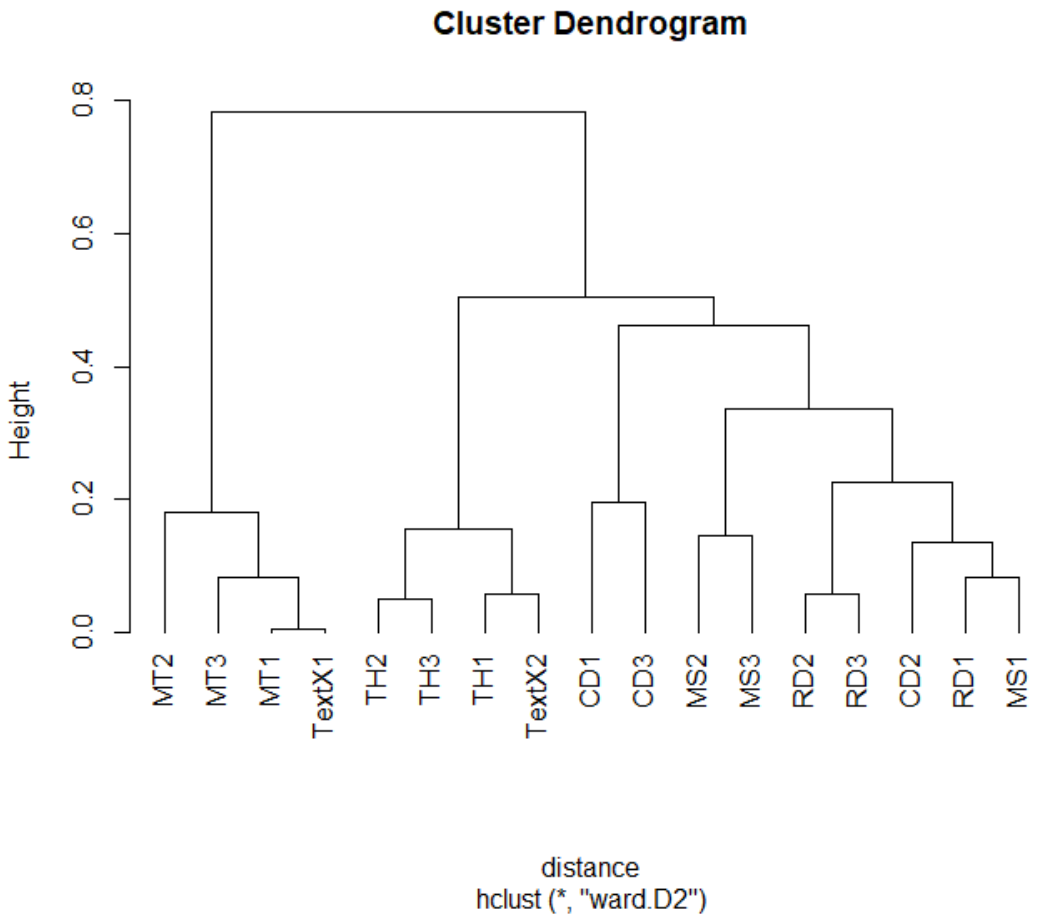


4. Effectuer une AFC avec la fonction CA de FactoMineR en ajoutant les deux textes inconnus en lignes supplémentaires.



5. Faire avec la fonction hclust une classification ascendante hiérarchique de Ward des 17échantillons décrits par leurs coordonnées factorielles sur les 4 premières dimensions de l'AFC. Quelle est la partition en 4 classes ?

Classification et dendrogramme :



Partition en 4 classes :

```
##     CD1     CD2     CD3     RD1     RD2     RD3     TH1     TH2     TH3     MS1
MS2
##       1       2       1       2       2       2       3       3       3       2
2
##     MS3     MT1     MT2     MT3 TextX1 TextX2
##       2       4       4       4       4       3
```

Chapitre 4

Ex 27

Voici les données de l'exercice : on a un data.frame.

```
##           taille poids velocite intellig affect agress  fonction
## beauceron    T++    P+      V++       I+    Af+    Ag+   Utilite
## basset        T-    P-       V-       I-    Af-    Ag+    Chasse
## ber_allem    T++    P+      V++      I++    Af+    Ag+   Utilite
## boxer         T+    P+       V+       I+    Af+    Ag+ Compagnie
## bull-dog      T-    P-       V-       I+    Af+    Ag- Compagnie
## bull-mass    T++   P++       V-      I++    Af-    Ag+   Utilite

## [1] "data.frame"

##           taille poids velocite intellig affect agress
## beauceron    T++    P+      V++       I+    Af+    Ag+
## basset        T-    P-       V-       I-    Af-    Ag+
## ber_allem    T++    P+      V++      I++    Af+    Ag+
## boxer         T+    P+       V+       I+    Af+    Ag+
## bull-dog      T-    P-       V-       I+    Af+    Ag-
```

3. On veut effectuer l'ACM de cette matrice H.

(a) Quelle décomposition en valeurs singulières généralisée (GSVD) faut-il faire ? Réaliser cette DSVG avec R. On calcul la matrice des fréquences, le poids des lignes et le poids des colonnes :

```
##                  T-         T+        T++         P-         P+        P++
## beauceron 0.00000000 0.00000000 0.00617284 0.00000000 0.00617284 0.00000000
## basset    0.00617284 0.00000000 0.00000000 0.00617284 0.00000000 0.00000000
## ber_allem 0.00000000 0.00000000 0.00617284 0.00000000 0.00617284 0.00000000
## boxer     0.00000000 0.00617284 0.00000000 0.00000000 0.00617284 0.00000000
## bull-dog  0.00617284 0.00000000 0.00000000 0.00617284 0.00000000 0.00000000
## bull-mass 0.00000000 0.00000000 0.00617284 0.00000000 0.00000000 0.00617284
##                  V-         V+        V++         I-         I+        I++
## beauceron 0.00000000 0.00000000 0.00617284 0.00000000 0.00617284 0.00000000
## basset    0.00617284 0.00000000 0.00000000 0.00617284 0.00000000 0.00000000
## ber_allem 0.00000000 0.00000000 0.00617284 0.00000000 0.00000000 0.00617284
## boxer     0.00000000 0.00617284 0.00000000 0.00000000 0.00617284 0.00000000
## bull-dog  0.00617284 0.00000000 0.00000000 0.00000000 0.00617284 0.00000000
## bull-mass 0.00617284 0.00000000 0.00000000 0.00000000 0.00000000 0.00617284
##                 Af-        Af+        Ag-        Ag+
## beauceron 0.00000000 0.00617284 0.00000000 0.00617284
## basset    0.00617284 0.00000000 0.00000000 0.00617284
## ber_allem 0.00000000 0.00617284 0.00000000 0.00617284
## boxer     0.00000000 0.00617284 0.00000000 0.00617284
## bull-dog  0.00000000 0.00617284 0.00617284 0.00000000
## bull-mass 0.00617284 0.00000000 0.00000000 0.00617284

##  beauceron     basset  ber_allem      boxer   bull-dog  bull-mass
## 0.03703704 0.03703704 0.03703704 0.03703704 0.03703704 0.03703704
```

```
##         T-         T+        T++         P-         P+        P++
## 0.04320988 0.03086420 0.09259259 0.04938272 0.08641975 0.03086420
```

(b) Montrer qu'en ACM, l'inertie totale des données vaut toujours m 1 ou m est le nombre total p de modalités et p le nombre de variables qualitatives. Vérifiez ensuite avec R que la somme des valeurs singulières trouvées à la question précédente vaut bien m / p - 1.

Sommes des valeurs singulières :

```
## [1] 1.666667
```

On obtient le même résultat

```
## [1] 1.666667
```

(c) Vérifiez également que le nombre maximum de dimension de cette ACM vaut bien min(n − 1, m − p). On a dix dimension maximum :

```
##  [1] 0.482 0.385 0.211 0.158 0.150 0.123 0.081 0.046 0.024 0.008
```

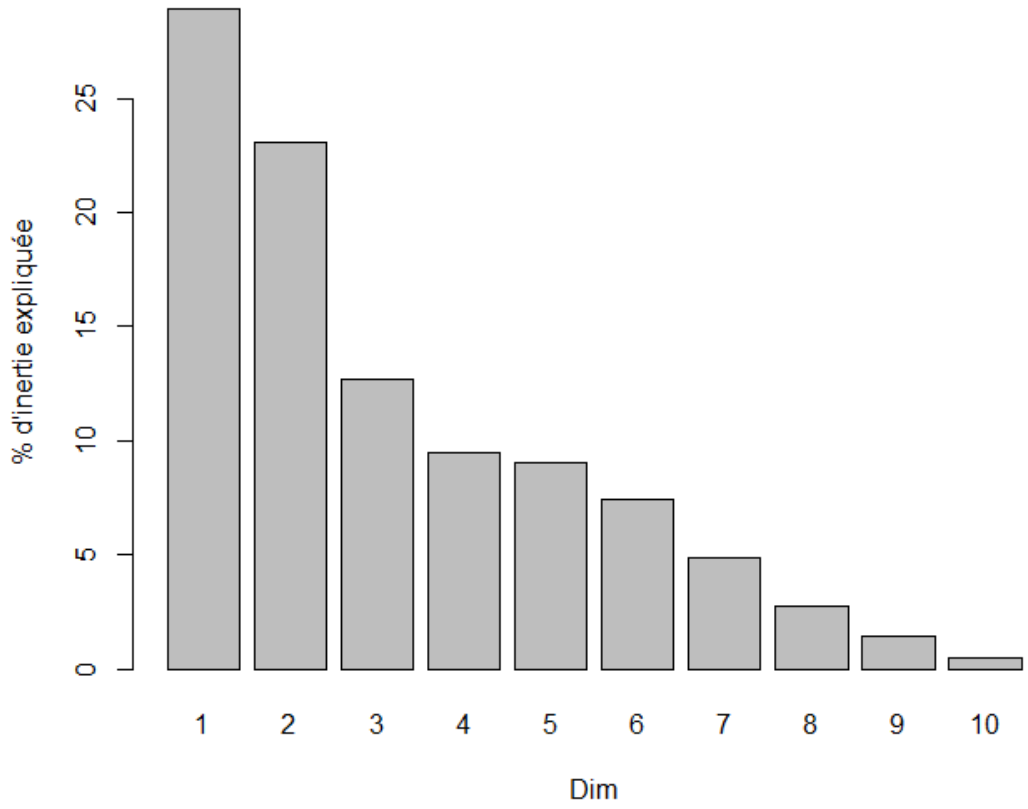
min (n-1,m-p) = 10 ont a donc une égalité.

```
## [1] 10
```

```
## [1] 10
```

(d) Représenter dans un diagramme en barre les pourcentages d'inertie expliquée par les dimensions de l'ACM. La dimension 1 et 2 contribuent beaucoup au pourcentage

d'inertie expliquée.



(e) Déterminer les matrices X et Y des coordonnées factorielles des races de chiens et des modalités des variables qualitatives sur les k = 3 premières dimensions. Modifier les noms des lignes et des colonnes dans X et Y afin qu'ils soient parlants.

Coordonnées factorielles des Races :

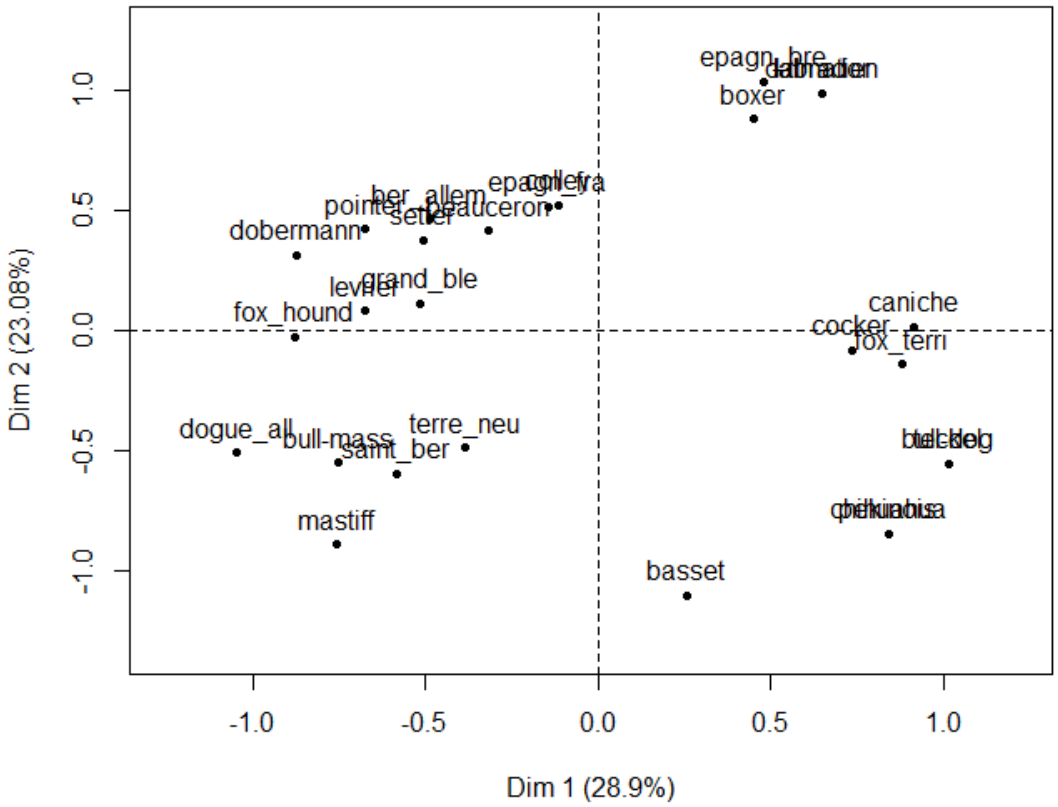
```
##                Dim 1       Dim 2       Dim 3
## beauceron -0.3172001  0.41770130  0.10146771
## basset     0.2541098 -1.10122699  0.19070097
## ber_allem -0.4863955  0.46444958  0.49813388
## boxer      0.4473649  0.88177794 -0.69201580
## bull-dog   1.0133522 -0.54987949  0.16342320
## bull-mass -0.7525745 -0.54691183 -0.49757307
## caniche    0.9123015  0.01618767  0.57656972
## chihuahua  0.8407994 -0.84385216  0.46994714
## cocker     0.7332953 -0.07907317 -0.66223042
## colley    -0.1173252  0.52610765  0.33489373
## dalmatien  0.6472398  0.99018429 -0.45858978
## dobermann -0.8732102  0.31548110  0.45231373
## dogue_all -1.0470168 -0.50695768 -0.16503476
```

```
## epagn_bre  0.4780443  1.03693257 -0.06192362
## epagn_fra -0.1449101  0.51578295 -0.11712661
## fox_hound -0.8765675 -0.02523985  0.36217150
## fox_terri  0.8816221 -0.13896696 -0.05352247
## grand_ble -0.5173377  0.11340393 -0.04402869
## labrador   0.6472398  0.99018429 -0.45858978
## levrier   -0.6766927  0.08316651  0.59559752
## mastiff   -0.7559318 -0.88763278 -0.58771530
## pekinois   0.8407994 -0.84385216  0.46994714
## pointer   -0.6733354  0.42388745  0.68573975
## saint_ber -0.5833790 -0.59366011 -0.89423924
## setter    -0.5041399  0.37713917  0.28907358
## teckel     1.0133522 -0.54987949  0.16342320
## terre_neu -0.3835042 -0.48525376 -0.66081322
```

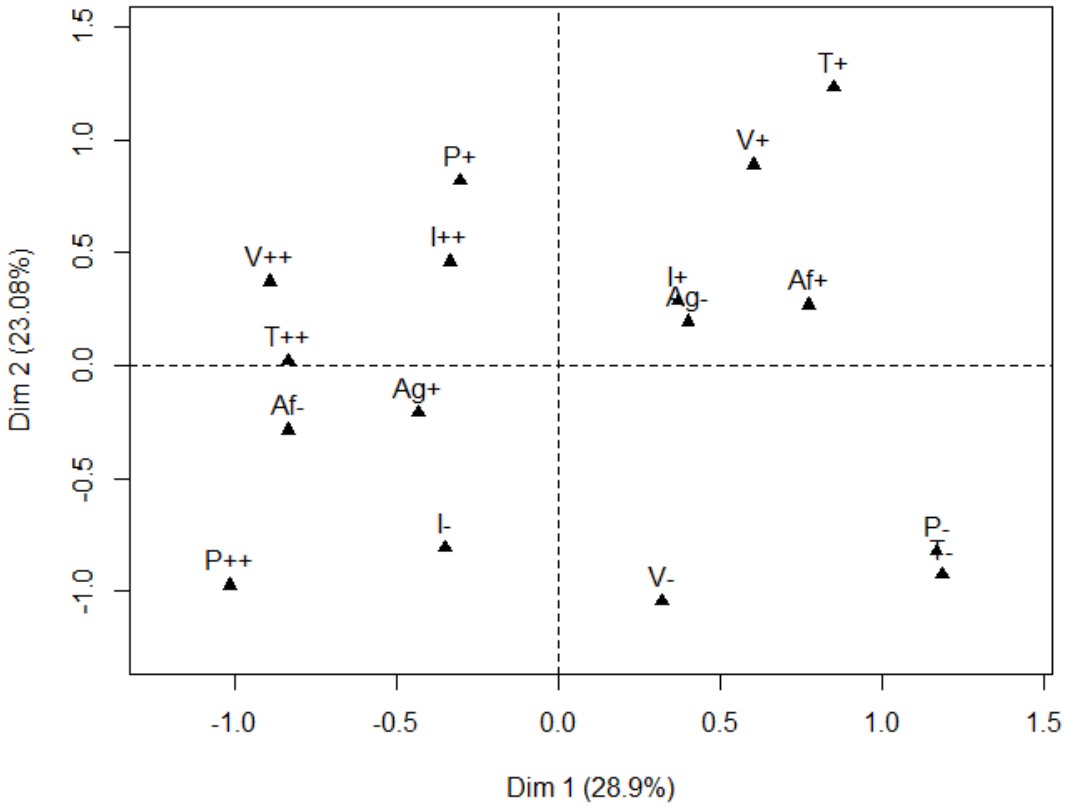
Coordonnées factorielles des modalités :

```
##          Dim 1       Dim 2       Dim 3
## T-   1.1849557 -0.92389650  0.61599962
## T+   0.8510880  1.23171972 -1.01605178
## T++ -0.8366753  0.02057846  0.05121744
## P-   1.1689180 -0.82434462  0.35877044
## P+  -0.3054053  0.81887572  0.23127208
## P++ -1.0151341 -0.97390062 -1.22159452
## V-   0.3199406 -1.04490006 -0.40172878
## V+   0.6036867  0.88781355 -0.35631249
## V++ -0.8920999  0.37183247  0.76308752
## I-  -0.3490450 -0.80855486  0.35151126
## I+   0.3694426  0.28550314 -0.49320252
## I++ -0.3350656  0.45948302  0.59992378
## Af- -0.8351500 -0.28746968 -0.06547357
## Af+  0.7754964  0.26693613  0.06079688
## Ag-  0.4007145  0.19425299  0.30972341
## Ag+ -0.4315386 -0.20919553 -0.33354829
```

(f) Faire un plot des individus et des modalités dans le premier plan factoriel. Individus:



Modalités :



(g) Utiliser la relation quasi-barycentrique pour retouver les coordonnées factorielles de la modalité T++ à partir des coordonnées factorielles des races de chiens. Indice des lignes des chiens T++

Moyenne des coordonnées factorielles des chiens T++

```
##      Dim 1      Dim 2      Dim 3
##  1.1849557 -0.9238965  0.6159996
```

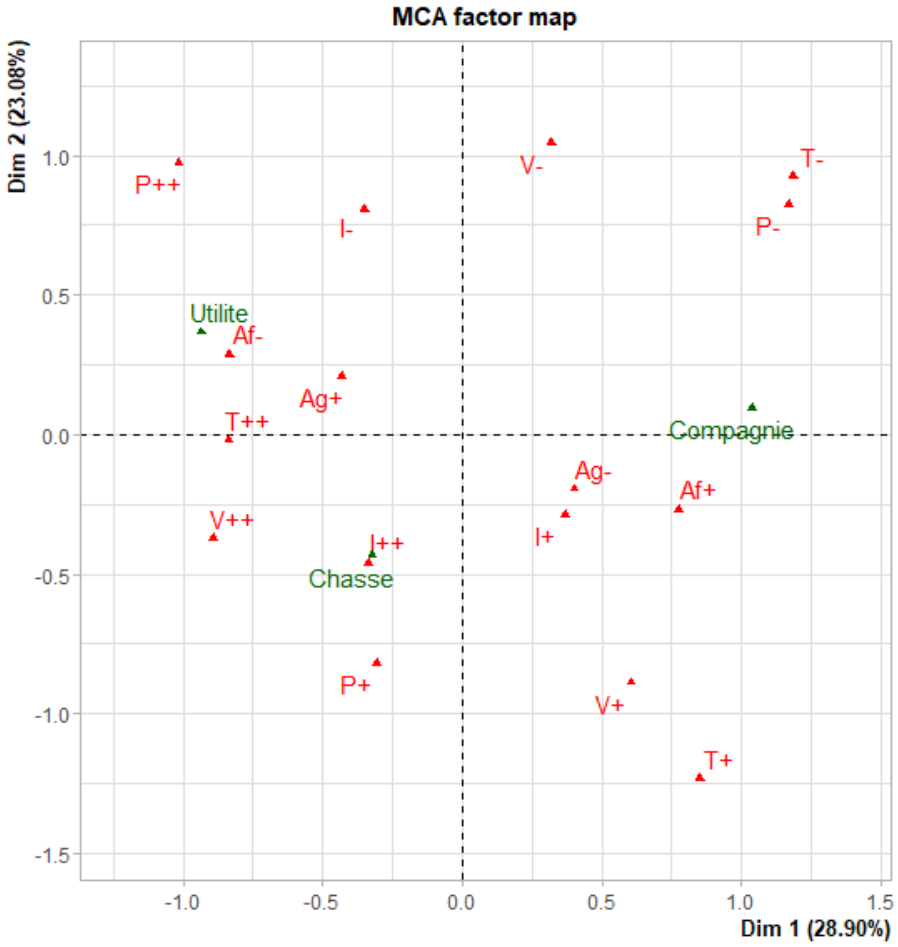
Coordonnées factorielles de T++

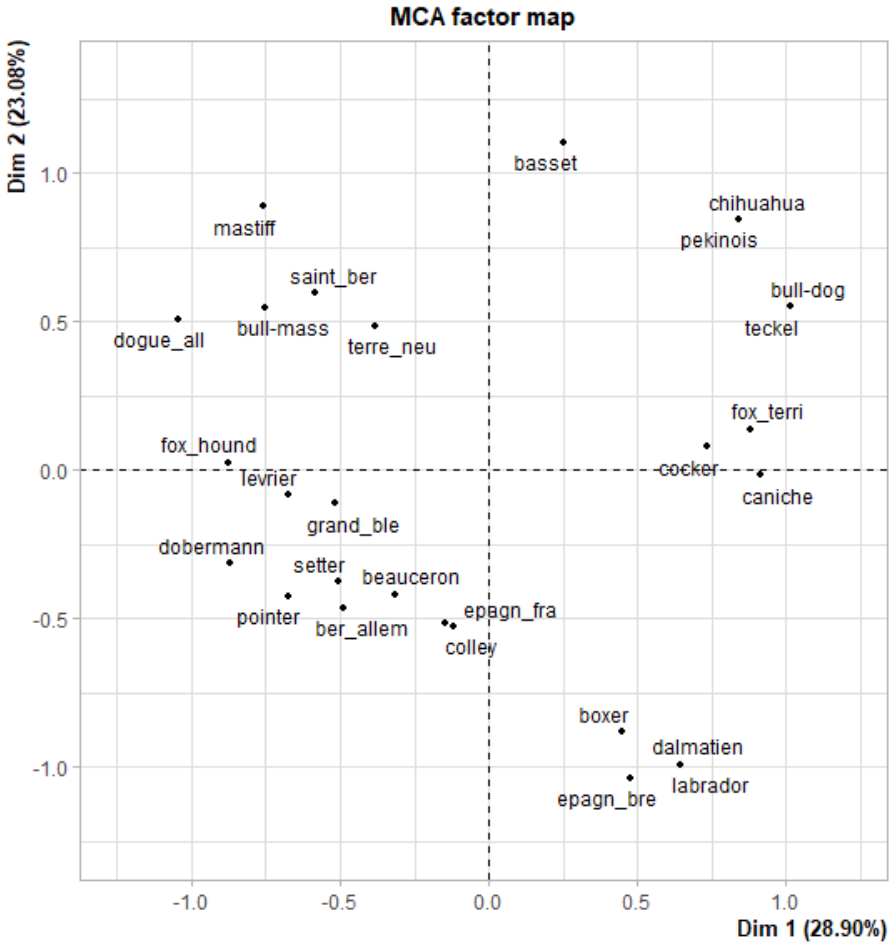
```
##       Dim 1      Dim 2     Dim 3
## T- 1.184956 -0.9238965 0.6159996
```

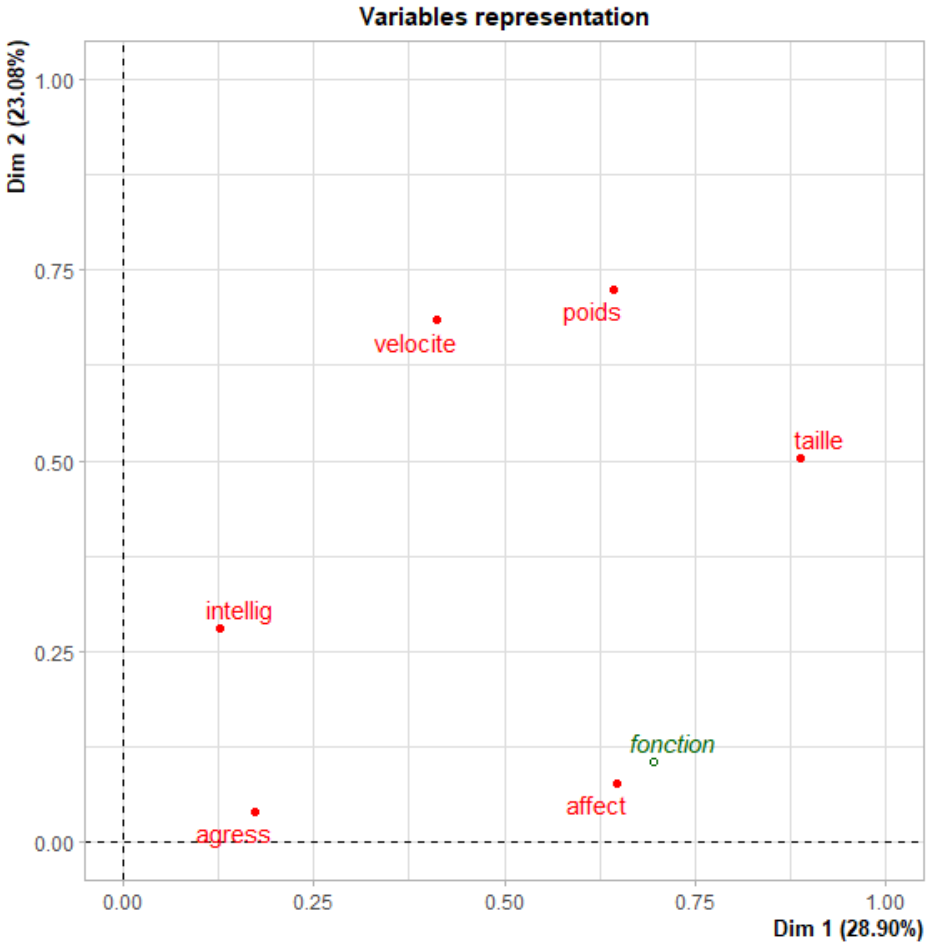
(h) Quels est le rapport de corrélation entre la variable taille avec la première composante principale ? Entre la variable taille et la seconde composante principale ?

4. On veut maintenant utiliser la fonction MCA du package FactoMineR.

(a) Faire l'ACM des données sur les races canines en mettant la variable fonction en illustratif.


MCA factor map
Dim 2 (23.08%)
Dim 1 (28.90%)
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
P++
I-
V-
T-
P-
Utilite
Af-
Ag+
T++
Compagnie
Ag-
Af+
V++
I++
I+
Chasse
P+
V+
T+

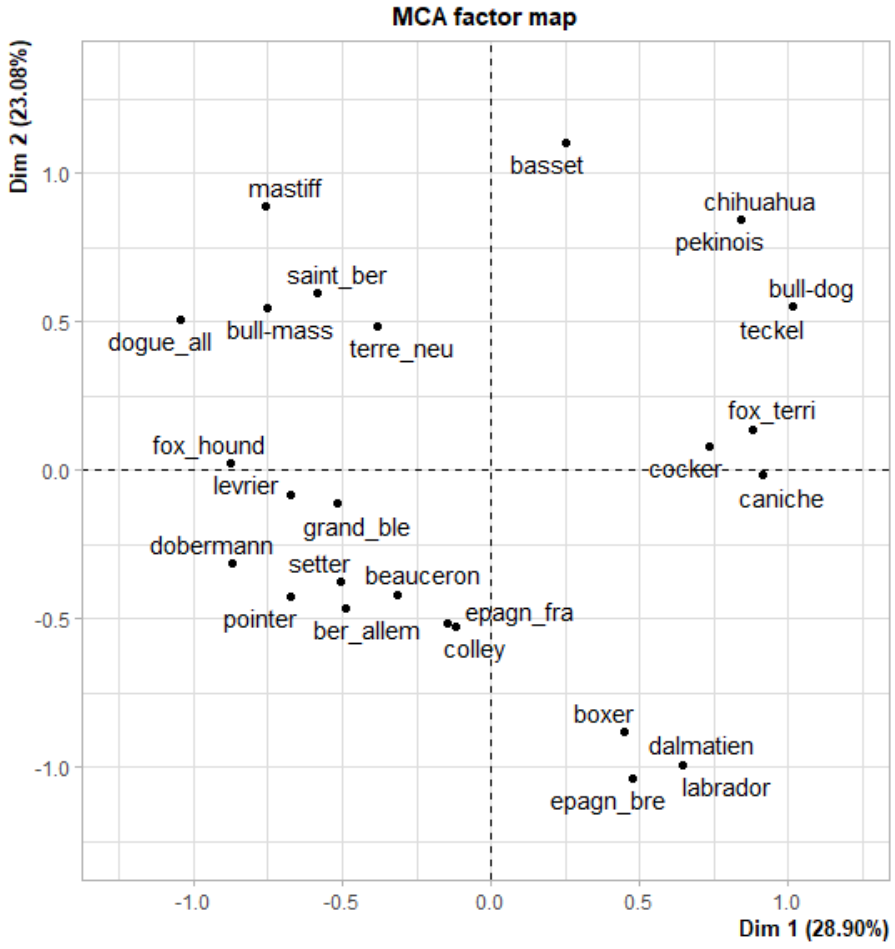

MCA factor map
Dim 2 (23.08%)
Dim 1 (28.90%)
1.0
0.5
0.0
-0.5
-1.0
basset
chihuahua
pekinois
mastiff
saint_ber
bull-dog
teckel
dogue_all
bull-mass
terre_neu
fox_terri
fox_hound
cocker
levrier
caniche
grand_ble
dobermann
setter
beauceron
pointer
epagn_fra
ber_allem
colley
boxer
dalmatien
labrador
epagn_bre

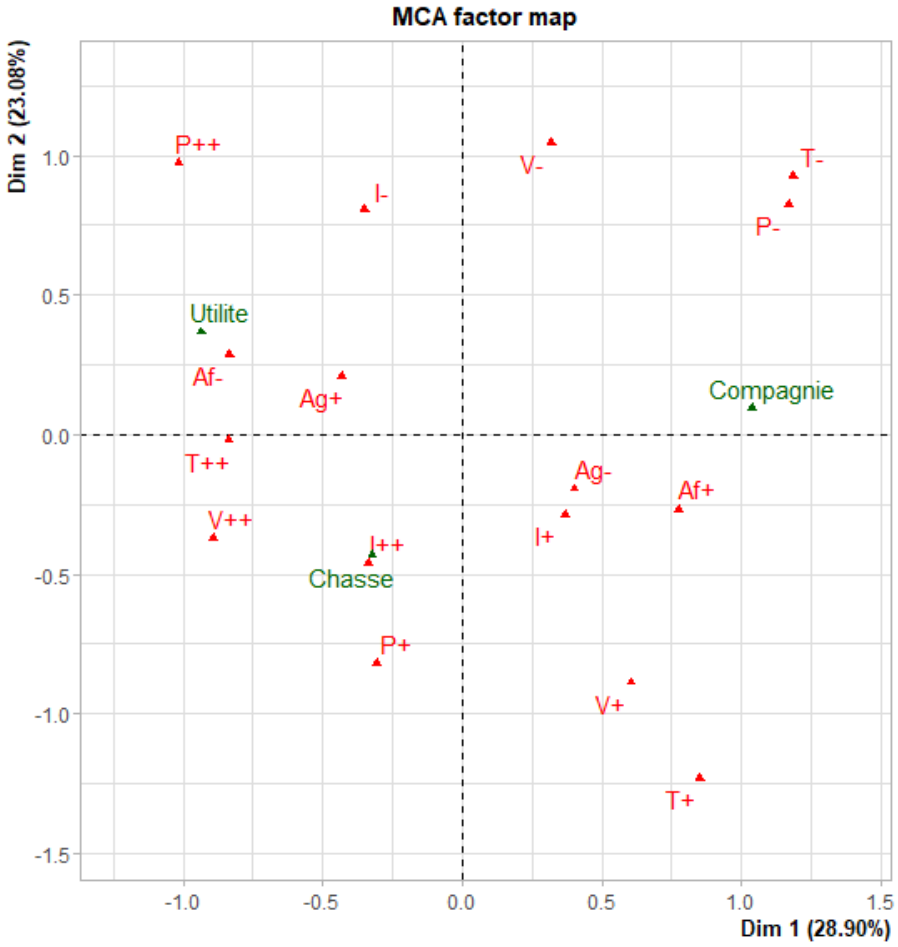


(b) Retrouvez les résulats numériques et les graphiques de la question 2.

```
##                Dim 1      Dim 2      Dim 3
## beauceron -0.3172001  0.4177013  0.1014677
## basset     0.2541098 -1.1012270  0.1907010
## ber_allem -0.4863955  0.4644496  0.4981339
## boxer      0.4473649  0.8817779 -0.6920158
## bull-dog   1.0133522 -0.5498795  0.1634232
## bull-mass -0.7525745 -0.5469118 -0.4975731

##                Dim 1      Dim 2      Dim 3      Dim 4      Dim 5
## beauceron -0.3172001 -0.4177013 -0.1014677 -0.2114363 -0.1185095
## basset     0.2541098  1.1012270 -0.1907010  0.2926373 -0.5240085
## ber_allem -0.4863955 -0.4644496 -0.4981339  0.5774253  0.2759021
## boxer      0.4473649 -0.8817779  0.6920158  0.2600018 -0.4555898
## bull-dog   1.0133522  0.5498795 -0.1634232 -0.3499193  0.3307865
## bull-mass -0.7525745  0.5469118  0.4975731  0.6551527  0.7219464

##          Dim 1       Dim 2       Dim 3
## T-   1.1849557 -0.92389650  0.61599962
## T+   0.8510880  1.23171972 -1.01605178
## T++ -0.8366753  0.02057846  0.05121744
```

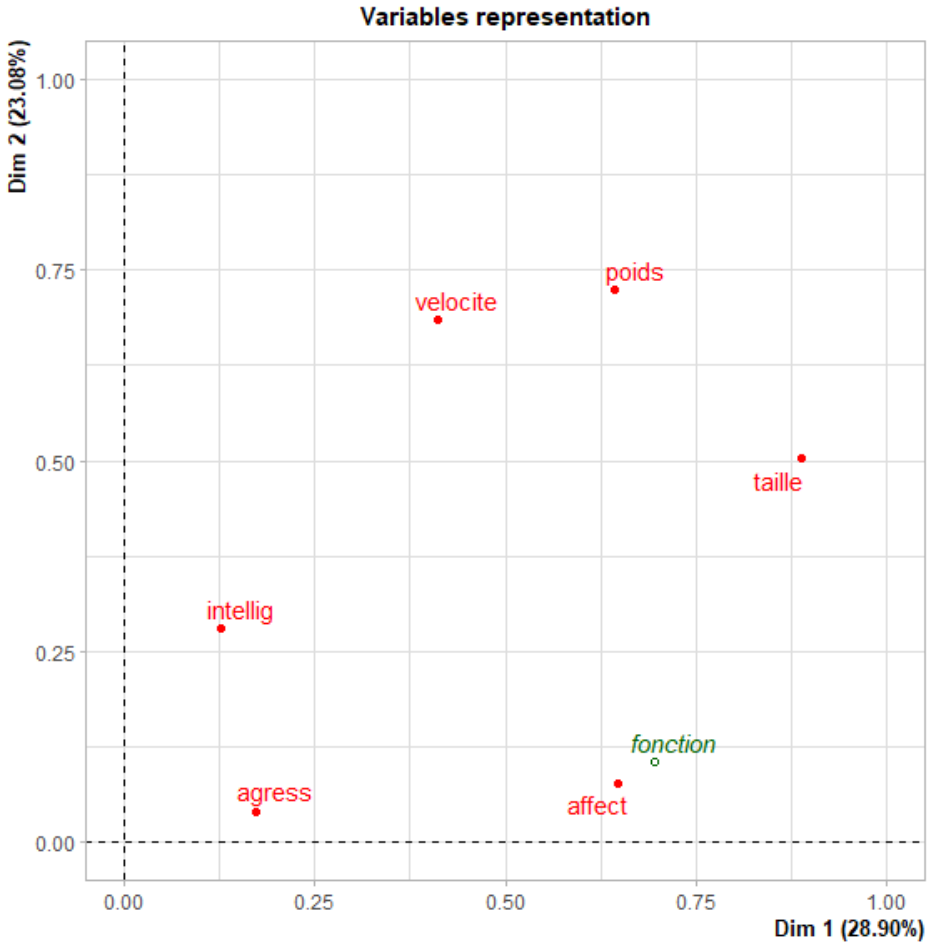
```
## P-   1.1689180 -0.82434462  0.35877044
## P+  -0.3054053  0.81887572  0.23127208
## P++ -1.0151341 -0.97390062 -1.22159452

##          Dim 1       Dim 2       Dim 3       Dim 4       Dim 5
## T-   1.1849557  0.92389650 -0.61599962  0.12014924 -0.01996350
## T+   0.8510880 -1.23171972  1.01605178  0.34245635 -0.31004022
## T++ -0.8366753 -0.02057846 -0.05121744 -0.17022176  0.11266304
## P-   1.1689180  0.82434462 -0.35877044  0.16488382 -0.05122143
## P+  -0.3054053 -0.81887572 -0.23127208 -0.11836395 -0.19020146
## P++ -1.0151341  0.97390062  1.22159452  0.06760494  0.61451838
```





(c) Retrouver les rapports de corrélations entre les variables qualitatives et les deux premières composantes principales. Faire le plot des variables en fonction de ces rapports de corrélation en utilisant la fonction plot.MCA.

```
##          Dim 1 Dim 2
## taille    0.89  0.50
## poids     0.64  0.72
## velocite  0.41  0.68
## intellig  0.13  0.28
## affect    0.65  0.08
## agress    0.17  0.04
```



(d) Mettre des données manquantes dans les données avec le code suivant : ?? Quel code ??

(e) Faire l'ACM de chiensNA. Comment les données manquantes sont-elles prises en compte dans la fonction MCA du package FactoMineR ? chiensNA ???

5. On veut maintenant comparer l'ACM et l'AFC dans le cas particulier de deux variables qualitatives.

(a) Avec la fonction CA de FactoMineR, effectuer l'AFC du tableau de contingence croisant les variables taille et poids.

```
##        poids
## taille P- P+ P++
##    T-   7  0   0
##    T+   1  4   0
##    T++  0 10   5

## **Results of the Correspondence Analysis (CA)**
## The row variable has  3  categories; the column variable has 3 categories
## The chi square of independence between the two variables is equal to
25.32857 (p-value =  4.320734e-05 ).
## *The results are available in the following objects:
```

```
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$col"            "results for the columns"
## 3  "$col$coord"      "coord. for the columns"
## 4  "$col$cos2"       "cos2 for the columns"
## 5  "$col$contrib"    "contributions of the columns"
## 6  "$row"            "results for the rows"
## 7  "$row$coord"      "coord. for the rows"
## 8  "$row$cos2"       "cos2 for the rows"
## 9  "$row$contrib"    "contributions of the rows"
## 10 "$call"           "summary called parameters"
## 11 "$call$marge.col" "weights of the columns"
## 12 "$call$marge.row" "weights of the rows"
```

(b) Avec la fonction MCA, effectuer l'ACM des deux premières colonnes des données chiens.

```
## **Results of the Multiple Correspondence Analysis (MCA)**
## The analysis was performed on 27 individuals, described by 2 variables
## *The results are available in the following objects:
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$var"            "results for the variables"
## 3  "$var$coord"      "coord. of the categories"
## 4  "$var$cos2"       "cos2 for the categories"
## 5  "$var$contrib"    "contributions of the categories"
## 6  "$var$v.test"     "v-test for the categories"
## 7  "$ind"            "results for the individuals"
## 8  "$ind$coord"      "coord. for the individuals"
## 9  "$ind$cos2"       "cos2 for the individuals"
## 10 "$ind$contrib"    "contributions of the individuals"
## 11 "$call"           "intermediate results"
## 12 "$call$marge.col" "weights of columns"
## 13 "$call$marge.li"  "weights of rows"
```

(c) Comparez les valeurs propres des deux analyses et vérifiez que vous retrouvez les relations du cours

```
##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1  0.9638515              48.192575                           48.19258
## dim 2  0.6391603              31.958016                           80.15059
## dim 3  0.3608397              18.041984                           98.19258
## dim 4  0.0361485               1.807425                          100.00000
```

Relation entre les valeurs propres des deux analyses, on retrouve d'abord les deux premières valeurs propres de MCA :

```
##     dim 1     dim 2     dim 3     dim 4
## 0.9908797 0.8997375 0.8003497 0.5950638
```

Pui les deux dernières valeurs de MCA :

```
##       dim 1       dim 2       dim 3       dim 4
## 0.009120305 0.100262487 0.199650336 0.404936208
```

Ex 28

1. Regarder les vidéos concernant ce package : https://www.youtube.com/user/HussonFrancois

1 ère vidéo : https://www.youtube.com/watch?v=bdD9P3fGb70

2 ème vidéo : https://www.youtube.com/watch?v=4F2C11hcvMM

2. Préparer un document avec Rmardown qui décrit les principales fonctionnalités de ce package, avec à chaque fois une explication de la méthode, des exemples et du code

Le package R missMDA permet de faire une ACP ou une ACM lorsqu'il y a des données manquantes. Dans le cas de l'ACP, on peut lancer directement une AFC mais dans ce cas-là les données sont remplacées par les moyennes de chaque variable avant que l'ACP soit vraiment calculé. Avec ce package, on peut imputer des données pour obtenir un tableau complet. On estime d'abord le nombre de dimension nécessaire pour compléter le tableau grâce à la fonction estim_ncpPCA(x, ncp.min, ncp.max), pour tester le nombre de dimension entre un nombre min et un nombre max. Ce procédé est long. Pour réduire ce temps lorsqu'on utilise de gros jeu de données on utilise la fonction method.cv="Kfold". Le processus est répété moins souvent donc il est plus rapide. Pour réaliser l'imputation on utilise la fonction imputePCA(x,np) ou np est le nombre de composantes de l'ACP. Ce procédé donne une meilleure estimation pour les valeurs manquantes que la moyenne. Les données déjà complété ne sont pas modifié ce qui permet de prendre en compte la liaison entre les variables et les ressemblances entre les individus. Après ça on peut lancer une ACP classique avec les données complété.

Pour l'ACM avec des données manquantes, on peut, comme avec l'ACP, commencer par utiliser MCA() pour voir ce que ça donne. On voit une nouvelle modalité 'NA' apparaître qui a été créée par le système. Ce n'est pas forcément intéressant pour visualiser les données de l'enquête. Encore une fois, on doit estimer le nombre de composantes nécessaires pour imputer les données, on utilise estim_ncpMCA() du package. Toutefois, obtenir la bonne réponse est plus dur. Une fois qu'on a le nombre de composante, on utilise imputeMCA(). On peut ensuite utiliser l'ACM qui permet de mieux visualiser les liaisons entres les modalités prises par chaque individu. Les valeurs imputées le sont de façon qu'elles ne contribuent pas dans la construction des axes. Le pourcentage d'inertie en revanche est affecté par ces données imputées et peut être surestimés. La structure du premier plan sera renforcée. Dans le cas où beaucoup de données sont imputées, le pourcentage d'inertie peut être fortement surestimé.

Chapitre 5

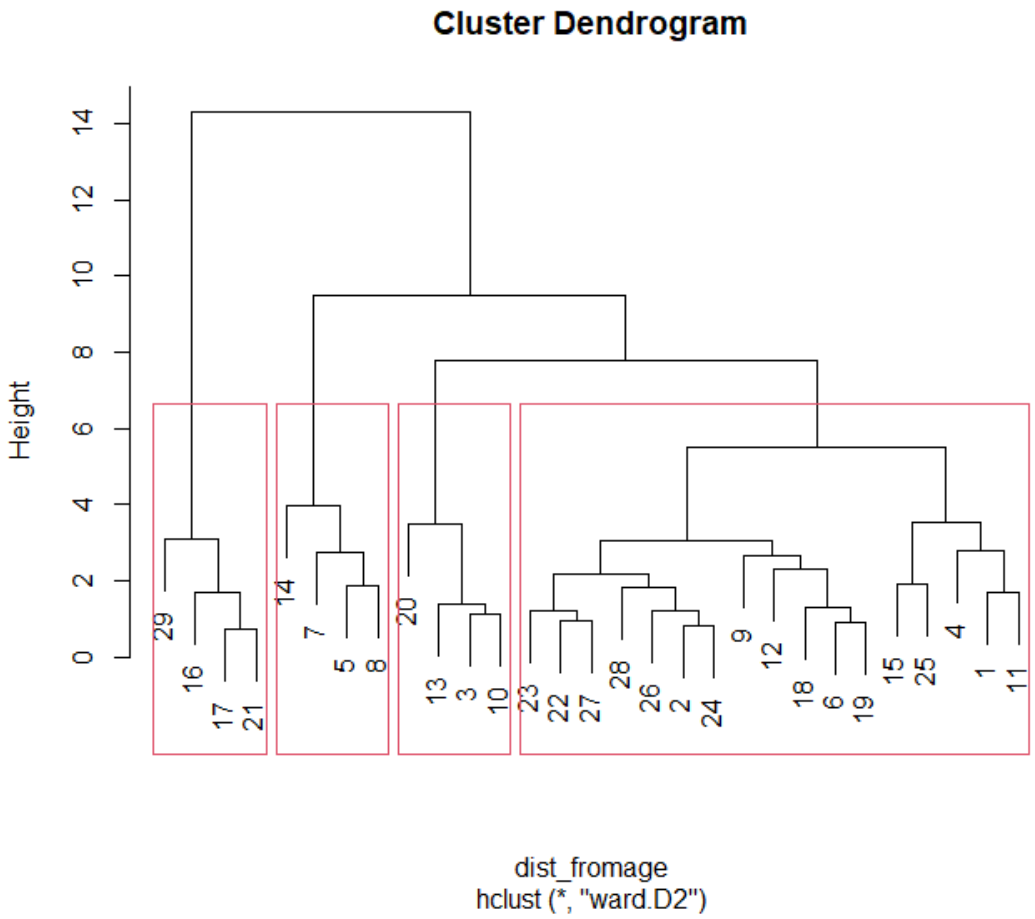
Voici les données de cet exercice :

```
##      Fromages calories sodium calcium lipides retinol folates proteines
## 1 CarredelEst      314  353.5    72.6    26.3    51.6    30.3      21.0
## 2     Babybel      314  238.0   209.8    25.1    63.7     6.4      22.6
## 3    Beaufort      401  112.0   259.4    33.3    54.9     1.2      26.6
## 4        Bleu      342  336.0   211.1    28.9    37.1    27.5      20.2
## 5   Camembert      264  314.0   215.9    19.5   103.0    36.4      23.4
## 6      Cantal      367  256.0   264.0    28.8    48.8     5.7      23.0
##   cholesterol magnesium
## 1          70        20
## 2          70        27
## 3         120        41
## 4          90        27
## 5          60        20
## 6          90        30

##    Fromages            calories       sodium         calcium
##  Length:29          Min.   : 70   Min.   : 22.0   Min.   : 72.6
##  Class :character   1st Qu.:292   1st Qu.:140.0   1st Qu.:132.9
##  Mode  :character   Median :321   Median :223.0   Median :202.3
##                     Mean   :300   Mean   :210.1   Mean   :185.7
##                     3rd Qu.:355   3rd Qu.:276.0   3rd Qu.:220.5
##                     Max.   :406   Max.   :432.0   Max.   :334.6
##     lipides         retinol          folates        proteines
##  Min.   : 3.40   Min.   : 37.10   Min.   : 1.20   Min.   : 4.10
##  1st Qu.:23.40   1st Qu.: 51.60   1st Qu.: 4.90   1st Qu.:17.80
##  Median :26.30   Median : 62.30   Median : 6.40   Median :21.00
##  Mean   :24.16   Mean   : 67.56   Mean   :13.01   Mean   :20.17
##  3rd Qu.:29.10   3rd Qu.: 76.40   3rd Qu.:21.10   3rd Qu.:23.40
##  Max.   :33.30   Max.   :150.50   Max.   :36.40   Max.   :35.70
##   cholesterol       magnesium
##  Min.   : 10.00   Min.   :10.00
##  1st Qu.: 70.00   1st Qu.:20.00
##  Median : 80.00   Median :26.00
##  Mean   : 74.59   Mean   :26.97
##  3rd Qu.: 90.00   3rd Qu.:30.00
##  Max.   :120.00   Max.   :51.00
```

CAH

On doit centrer et réduire les données pour éviter que les variables à forte variance n'influencent trop les résultats. On calcule la Matrice des distances entre individus.

Dendrogramme avec 4 classes:



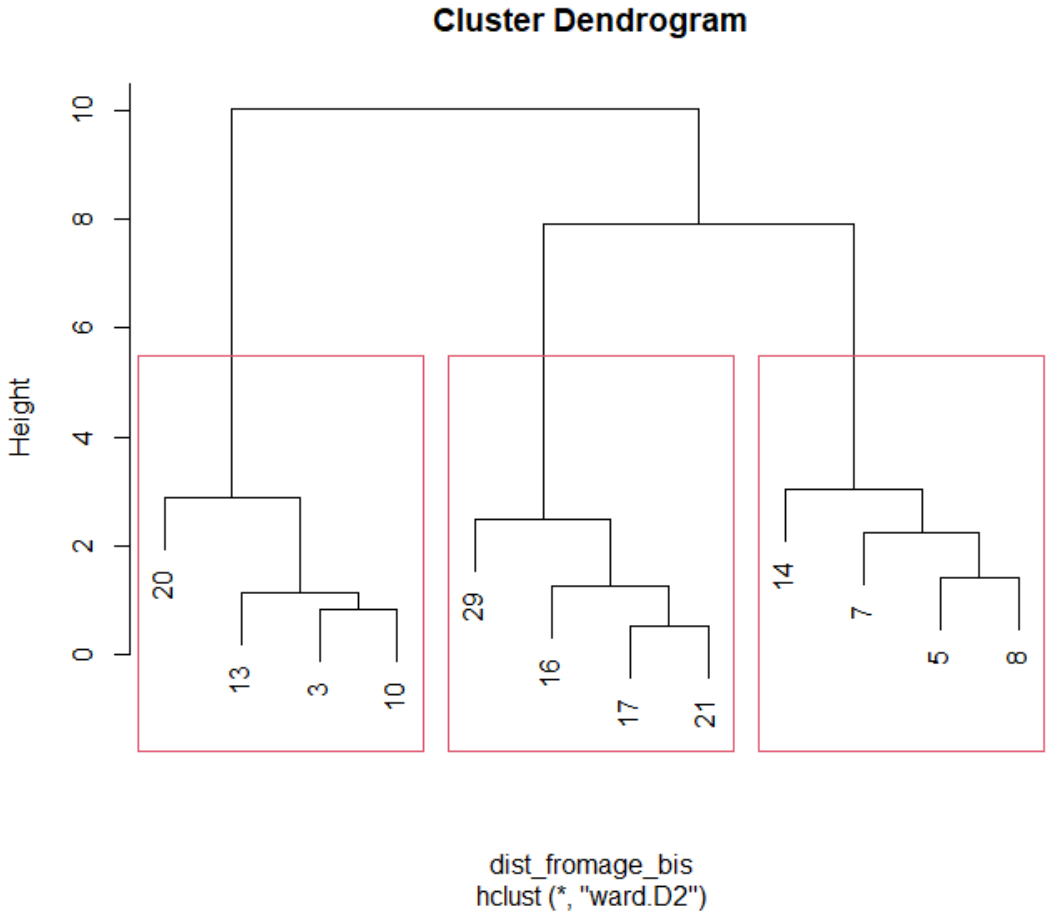
On enlève les fromages frais de notre data_set, c'est à dire la 4 ème classe.

```
##              Fromages calories sodium calcium lipides retinol folates proteines
## 29 Yaourtlaitent.nat.       70     91   215.7     3.4    42.9     2.9       4.1
## 16     Fr.frais20nat.       80     41   146.3     3.5    50.0    20.0       8.3
## 17     Fr.frais40nat.      115     25    94.8     7.8    64.3    22.6       7.0
## 21      Petitsuisse40      142     22    78.2    10.4    63.4    20.4       9.4
## 14 Fr.chevrepatemolle      206    160    72.8    18.5   150.5    31.0      11.1
## 7            Chabichou      344    192    87.2    27.9    90.1    36.3      19.5
## 5            Camembert      264    314   215.9    19.5   103.0    36.4      23.4
## 8             Chaource      292    276   132.9    25.4   116.4    32.5      17.8
## 20            Parmesan      381    240   334.6    27.5    90.0     5.2      35.7
## 13            Emmental      378     60   308.2    29.4    56.3     2.4      29.4
## 3             Beaufort      401    112   259.4    33.3    54.9     1.2      26.6
## 10               Comte      399     92   220.5    32.4    55.9     1.3      29.2
##    cholesterol magnesium
## 29          13        14
## 16          10        11
## 17          30        10
## 21          20        10
## 14          50        16
## 7           80        36
## 5           60        20
## 8           70        25
## 20          80        46
## 13         110        45
```

```
## 3          120        41
## 10         120        51
```

On applique les mêmes traitements que précédemment.

Dendrogramme sans les fromages frais avec 3 classes. Les classes sont les mêmes qu'avant mais évidemment on n'a plus que 3 classes, les fromages frais constituaient une classe entière dans le dendrogramme précédent.




```
## 29 16 17 21 14  7  5  8 20 13  3 10
##  1  1  1  1  2  2  2  2  3  3  3  3
```

K-means

On va appliquer la procédure Kmeans sur les données sans fromage frais, on lui demande 4 groupes en 5 essais.

```
## K-means clustering with 3 clusters of sizes 4, 4, 4
##
## Cluster means:
##     calories      sodium    calcium    lipides    retinol     folates
## 1  0.1593572  0.99873404 -0.5835202  0.2617469  1.1371899  1.13712342
## 2 -1.1990654 -0.90476489 -0.5118678 -1.2277355 -0.7093617 -0.08395259
```

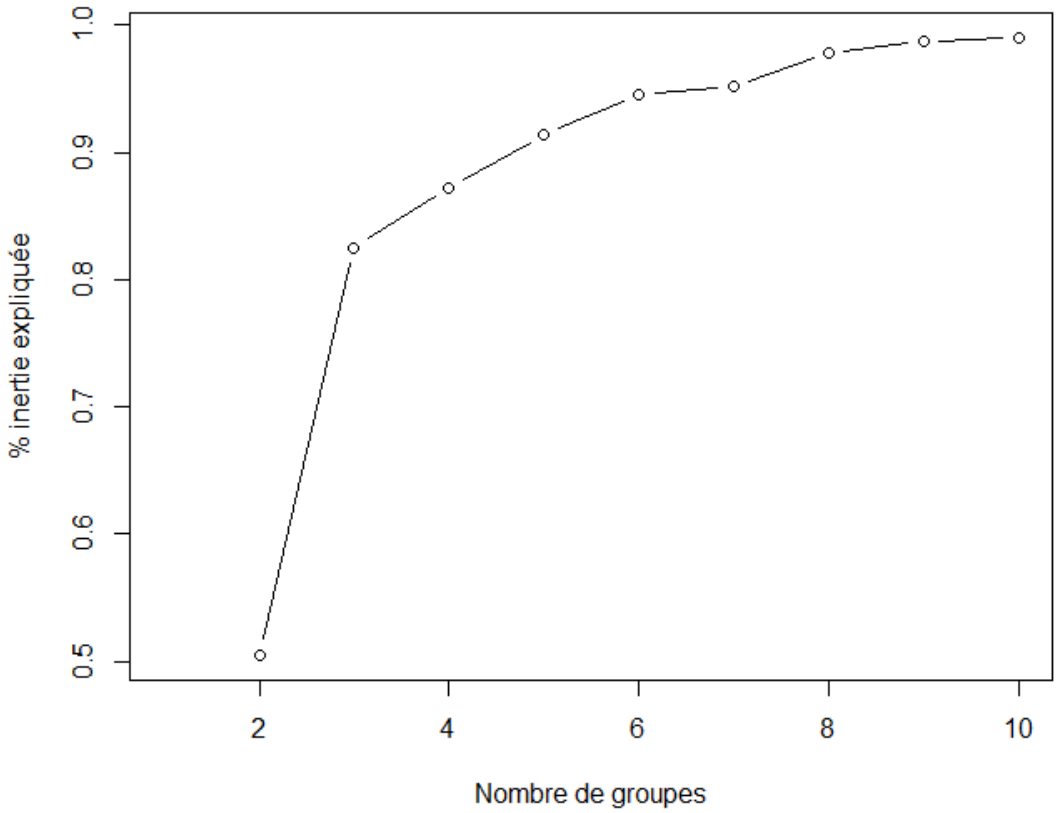
```
## 3  1.0397083 -0.09396915  1.0953881  0.9659886 -0.4278282 -1.05317082
##     proteines cholesterol  magnesium
## 1 -0.04863985  0.03526182 -0.1803498
## 2 -1.07725308 -1.12837817 -1.0078373
## 3  1.12589293  1.09311636  1.1881872
##
## Clustering vector:
## 29 16 17 21 14  7  5  8 20 13  3 10
##  2  2  2  2  1  1  1  1  3  3  3  3
##
## Within cluster sum of squares by cluster:
## [1] 8.107578 4.013674 5.162696
##  (between_SS / total_SS =  82.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

On regarde les correspondances des groupes entre CAH et KMeans :

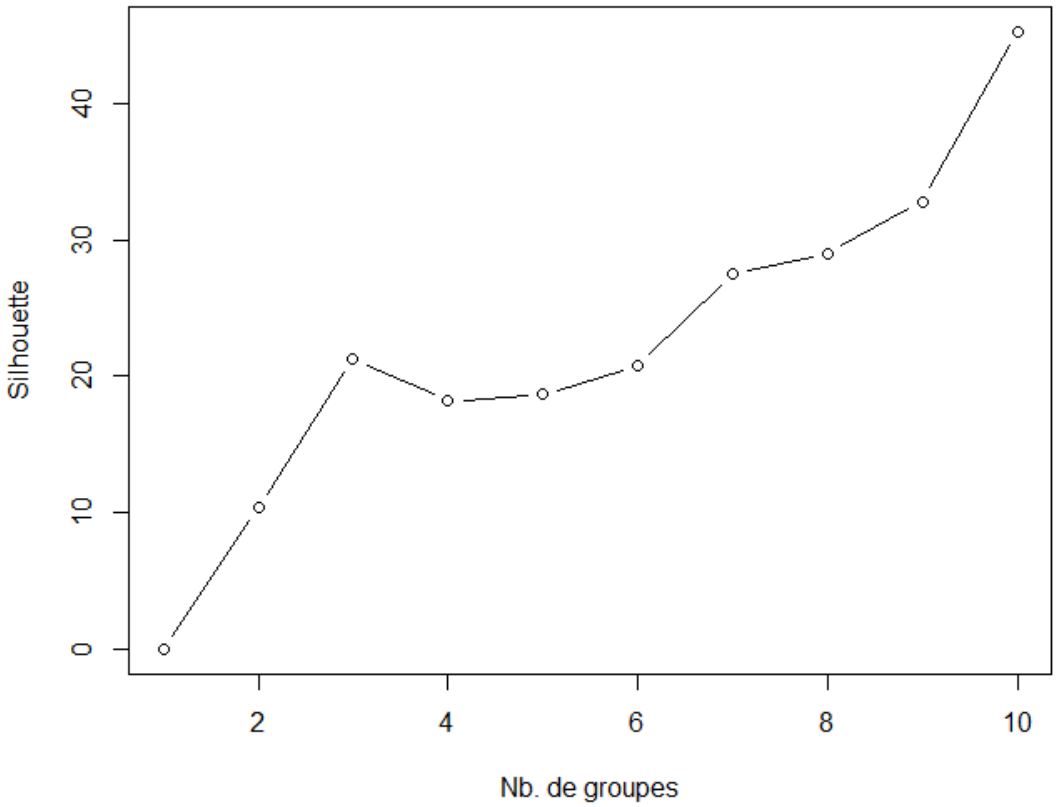
```
##
## groupes_cah_bis 1 2 3
##               1 0 4 0
##               2 4 0 0
##               3 0 0 4
```

On affiche le pourcentage d'inertie expliquée cumulative. On cherche à savoir le nombre de groupe nécessaire pour expliquer les données. Avec 4 groupes on expliquerait plus de 85% de l'inertie.

Evaluation de la proportion d'inertie expliquée : avec 5 groupes on expliquerai plus de 90 % des données.



Indice de Calinski Harabasz : on voit ici que seulement 3 classes serait un bon nombre de classe.


Silhouette
0
10
20
30
40
2
4
6
8
10
Nb. de groupes

Chapitre 6 : données personnelles.

Les données :

On nous est demandé de traiter les données sur le cancer du sein en fonction de l'âge, du centre et l'histologie. Voici un aperçu des données comme elles sont affichées dans le fichier de données :

```
#   Effectif Centre Age     `Centre-Age`  Survie Taille Inflamat…¹ Type …² Histo…³
##      <dbl> <chr>  <chr>   <chr>         <chr>  <chr>             <chr>   <chr>
## 1        9 Tokyo  Moins50 Tokyo-Moins50 Non    Minime            Maligne Minime…
## 2        7 Tokyo  Moins50 Tokyo-Moins50 Non    Minime            Benigne Minime…
## 3        4 Tokyo  Moins50 Tokyo-Moins50 Non    Grande            Maligne Grande…
## 4        3 Tokyo  Moins50 Tokyo-Moins50 Non    Grande            Benigne Grande…
## 5       26 Tokyo  Moins50 Tokyo-Moins50 Oui    Minime            Maligne Minime…
## 6       68 Tokyo  Moins50 Tokyo-Moins50 Oui    Minime            Benigne Minime…
## # … with abbreviated variable names ¹`Taille Inflamation`,
## #   ²`Type Inflammation`, ³Histologie
```

Nous avons 764 individus et 8 variables. Nous avons des données qualitatives. Ces données nous donnent le nombre de patients qui ont été soigné dans une des trois Centres (Tokyo, Boston, Glamorgan), la classe d'âge à laquelle ils appartiennent parmi moins de 50 ans, entre 50 et 69 et plus de 70 ans, nous avons le nombre de personne ayant survécut suivant le centre et leur âge puis nous avons des informations sur la taille de inflammation (grande ou minime) et sur le Type d' inflammation (Maligne ou Bénigne). Nous avons aussi les variables Centre-Age et Histologie qui résume les variables Centre et Age et Taille et Type d'inflammation respectivement.

Même si nous avons 78 lignes dans ce fichier nous avons en fait 764 individus car la première colonne correspond à l'effectif de chaque combinaisons de modalités.

```
## [1] 764
```

Dans un premier temps nous allons reconstruire les données de façon à avoir un tableau nous donnant seulement l'effectif de chaque combinaison :

```
##                                      Effectif
## Tokyo-Moins50 Non Minime-Maligne            9
## Tokyo-Moins50 Non Minime-Benigne            7
## Tokyo-Moins50 Non Grande-Maligne            4
## Tokyo-Moins50 Non Grande-Benigne            3
## Tokyo-Moins50 Oui Minime-Maligne           26
## Tokyo-Moins50 Oui Minime-Benigne           68
## Tokyo-Moins50 Oui Grande-Maligne           25
## Tokyo-Moins50 Oui Grande-Benigne            9
## Tokyo-50-69 Non Minime-Maligne              9
## Tokyo-50-69 Non Minime-Benigne              9
## Tokyo-50-69 Non Grande-Maligne             11
## Tokyo-50-69 Non Grande-Benigne              2
## Tokyo-50-69 Oui Minime-Maligne             20
## Tokyo-50-69 Oui Minime-Benigne             46
```

```
## Tokyo-50-69 Oui Grande-Maligne             18
## Tokyo-50-69 Oui Grande-Benigne              5
## Tokyo-Plus70 Non Minime-Maligne             2
## Tokyo-Plus70 Non Minime-Benigne             3
## Tokyo-Plus70 Non Grande-Maligne             1
## Tokyo-Plus70 Non Grande-Benigne             0
## Tokyo-Plus70 Oui Minime-Maligne             1
## Tokyo-Plus70 Oui Minime-Benigne             6
## Tokyo-Plus70 Oui Grande-Maligne             5
## Tokyo-Plus70 Oui Grande-Benigne             1
## Boston-Moins50 Non Minime-Maligne           6
## Boston-Moins50 Non Minime-Benigne           7
## Boston-Moins50 Non Grande-Maligne           6
## Boston-Moins50 Non Grande-Benigne           0
## Boston-Moins50 Oui Minime-Maligne          11
## Boston-Moins50 Oui Minime-Benigne          24
## Boston-Moins50 Oui Grande-Maligne           4
## Boston-Moins50 Oui Grande-Benigne           0
## Boston-50-69 Non Minime-Maligne             8
## Boston-50-69 Non Minime-Benigne            20
## Boston-50-69 Non Grande-Maligne             3
## Boston-50-69 Non Grande-Benigne             2
## Boston-50-69 Oui Minime-Maligne            18
## Boston-50-69 Oui Minime-Benigne            58
## Boston-50-69 Oui Grande-Maligne            10
## Boston-50-69 Oui Grande-Benigne             3
## Boston-Plus70 Non Minime-Maligne            9
## Boston-Plus70 Non Minime-Benigne           18
## Boston-Plus70 Non Grande-Maligne            3
## Boston-Plus70 Non Grande-Benigne            0
## Boston-Plus70 Oui Minime-Maligne           15
## Boston-Plus70 Oui Minime-Benigne           26
## Boston-Plus70 Oui Grande-Maligne            1
## Boston-Plus70 Oui Grande-Benigne            1
## Glamorgan-Moins50 Non Minime-Maligne       16
## Glamorgan-Moins50 Non Minime-Benigne        7
## Glamorgan-Moins50 Non Grande-Maligne        3
## Glamorgan-Moins50 Non Grande-Benigne        0
## Glamorgan-Moins50 Oui Minime-Maligne       16
## Glamorgan-Moins50 Oui Minime-Benigne       20
## Glamorgan-Moins50 Oui Grande-Maligne        8
## Glamorgan-Moins50 Oui Grande-Benigne        1
## Glamorgan-50-69 Non Minime-Maligne         14
## Glamorgan-50-69 Non Minime-Benigne         12
## Glamorgan-50-69 Non Grande-Maligne          3
## Glamorgan-50-69 Non Grande-Benigne          0
## Glamorgan-50-69 Oui Minime-Maligne         27
## Glamorgan-50-69 Oui Minime-Benigne         39
## Glamorgan-50-69 Oui Grande-Maligne         10
## Glamorgan-50-69 Oui Grande-Benigne          4
```

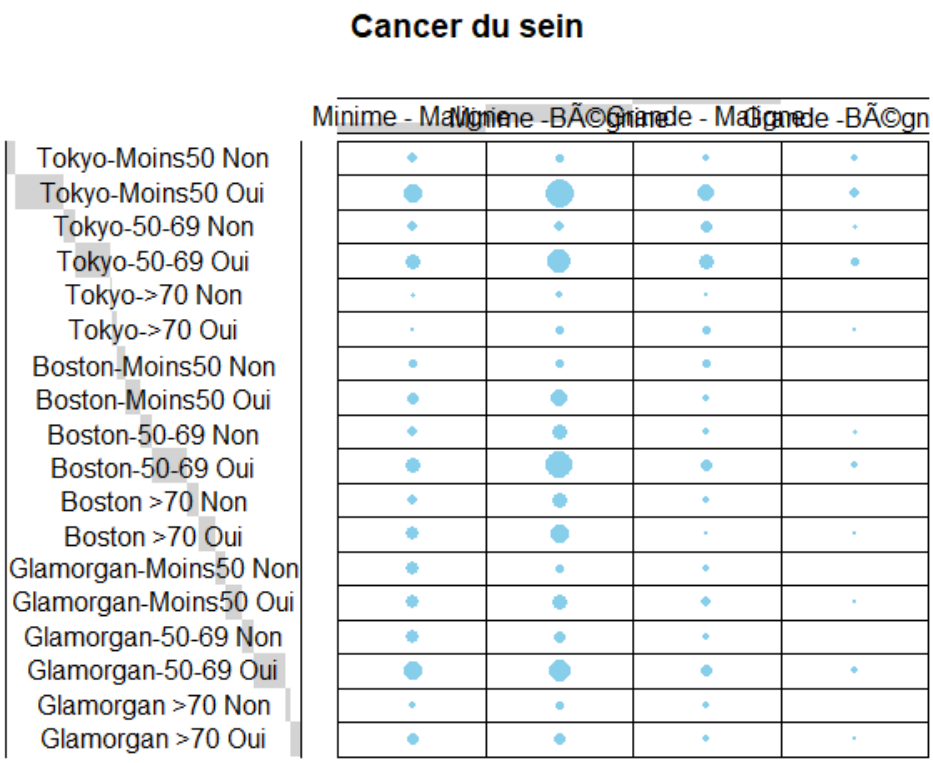
```
## Glamorgan-Plus70 Non Minime-Maligne         3
## Glamorgan-Plus70 Non Minime-Benigne         7
## Glamorgan-Plus70 Non Grande-Maligne         3
## Glamorgan-Plus70 Non Grande-Benigne         0
## Glamorgan-Plus70 Oui Minime-Maligne        12
## Glamorgan-Plus70 Oui Minime-Benigne        11
## Glamorgan-Plus70 Oui Grande-Maligne         4
## Glamorgan-Plus70 Oui Grande-Benigne         1

## [1] "data.frame"
```

Nous cherchons à savoir quel est la combinaison centre-age, survie et histologie qui admets le plus de patient: les gens qui ont survécu à des cancers minime et bénins de moins de 50 ans et soigné au centre de Tokio sont les plus nombreux dans cette étude.

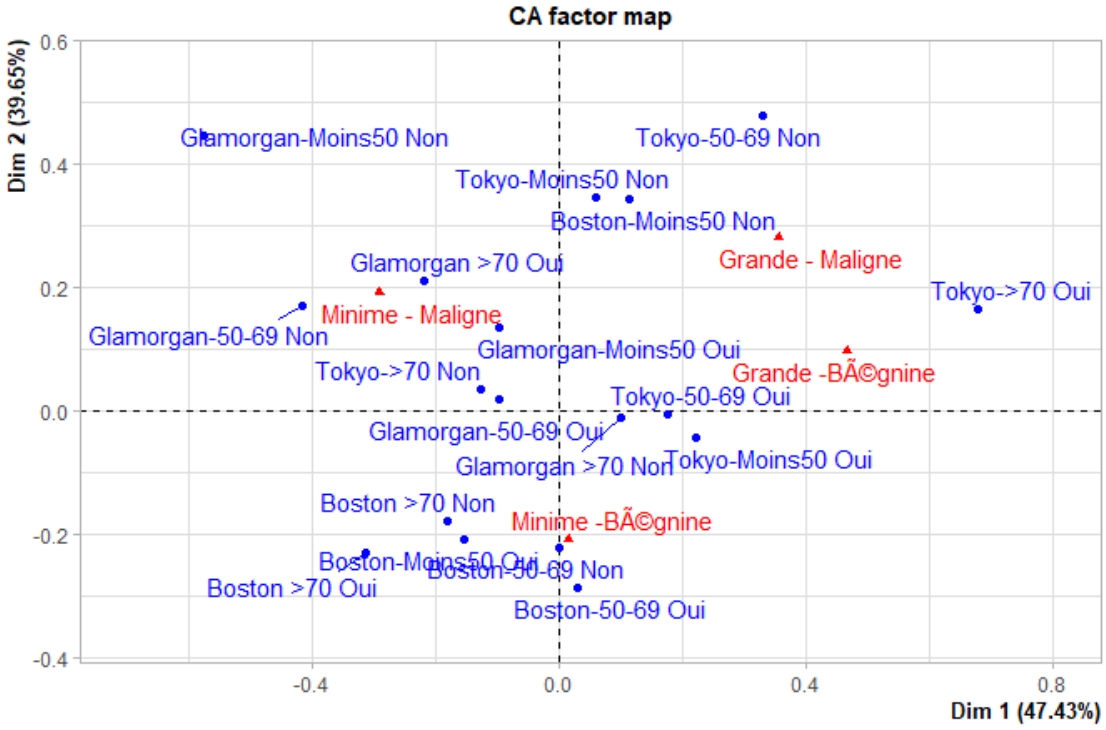
```
## [1] "Tokyo-Moins50 Oui Minime-Benigne"

## [1] "table"
```

Nous avons réduit le problèmes à deux variables qualitatives, nous aurions pu découpez le tableau différemment mais comme nous disposons que du tableau de contingence nous nous aurions dû effectué plus de traitement des données encore. Nous allons dons réaliser une AFC (Analyse factorielle des composantes) dans le tableau de contingence ci-dessus.

Graphique du tableau de contingence



Calcul de l'AFC.



La p-value du test d'indépendance du khi deux est inférieur à 0.05 ce qui signifie qu'il y a non-indépendance. C'est logique en premier lieu puisque le type d'inflammation peut avoir un effet sur le fait que le patient est survécu ou non par exemple.

```
## **Results of the Correspondence Analysis (CA)**
## The row variable has  18  categories; the column variable has 4 categories
## The chi square of independence between the two variables is equal to
87.22626 (p-value =  0.001188127 ).
## *The results are available in the following objects:
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$col"            "results for the columns"
## 3  "$col$coord"      "coord. for the columns"
## 4  "$col$cos2"       "cos2 for the columns"
## 5  "$col$contrib"    "contributions of the columns"
## 6  "$row"            "results for the rows"
## 7  "$row$coord"      "coord. for the rows"
## 8  "$row$cos2"       "cos2 for the rows"
```

```
## 9  "$row$contrib"    "contributions of the rows"
## 10 "$call"           "summary called parameters"
## 11 "$call$marge.col" "weights of the columns"
## 12 "$call$marge.row" "weights of the rows"
```

Significativité statistique

Ici, l'association est très significative car la valeur du chi deux est très petite.

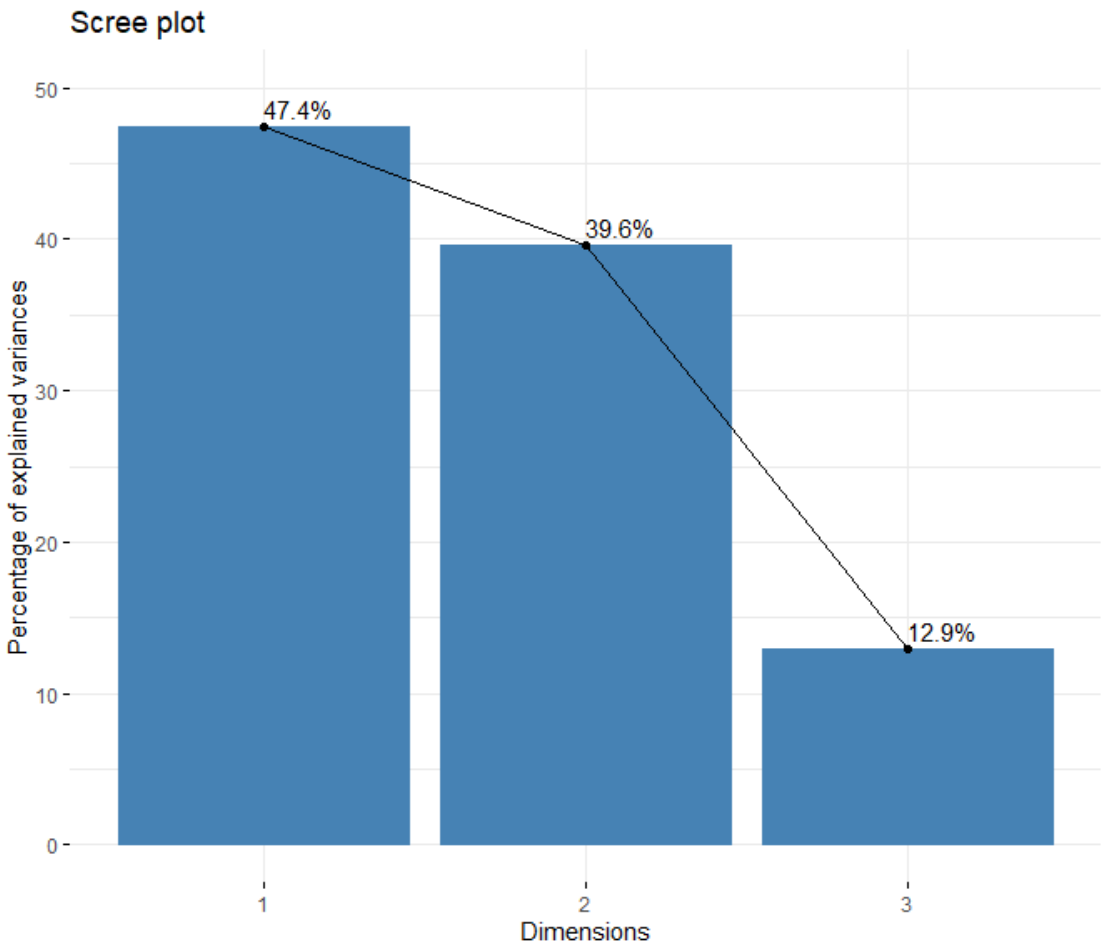
```
##
##  Pearson's Chi-squared test
##
## data:  housetasks
## X-squared = 1944.5, df = 36, p-value < 2.2e-16
```

Valeurs propres/Variances

L'examination des valeurs propres permet de déterminer le nombre d'axes principaux à considérer. Les valeurs propres correspondent à la quantité d'informations retenue par chaque axe. Elles sont grandes pour le premier axe et petites pour l'axe suivant.

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1 0.05414959         47.42870                    47.42870
## Dim.2 0.04526478         39.64666                    87.07536
## Dim.3 0.01475612         12.92464                   100.00000
```

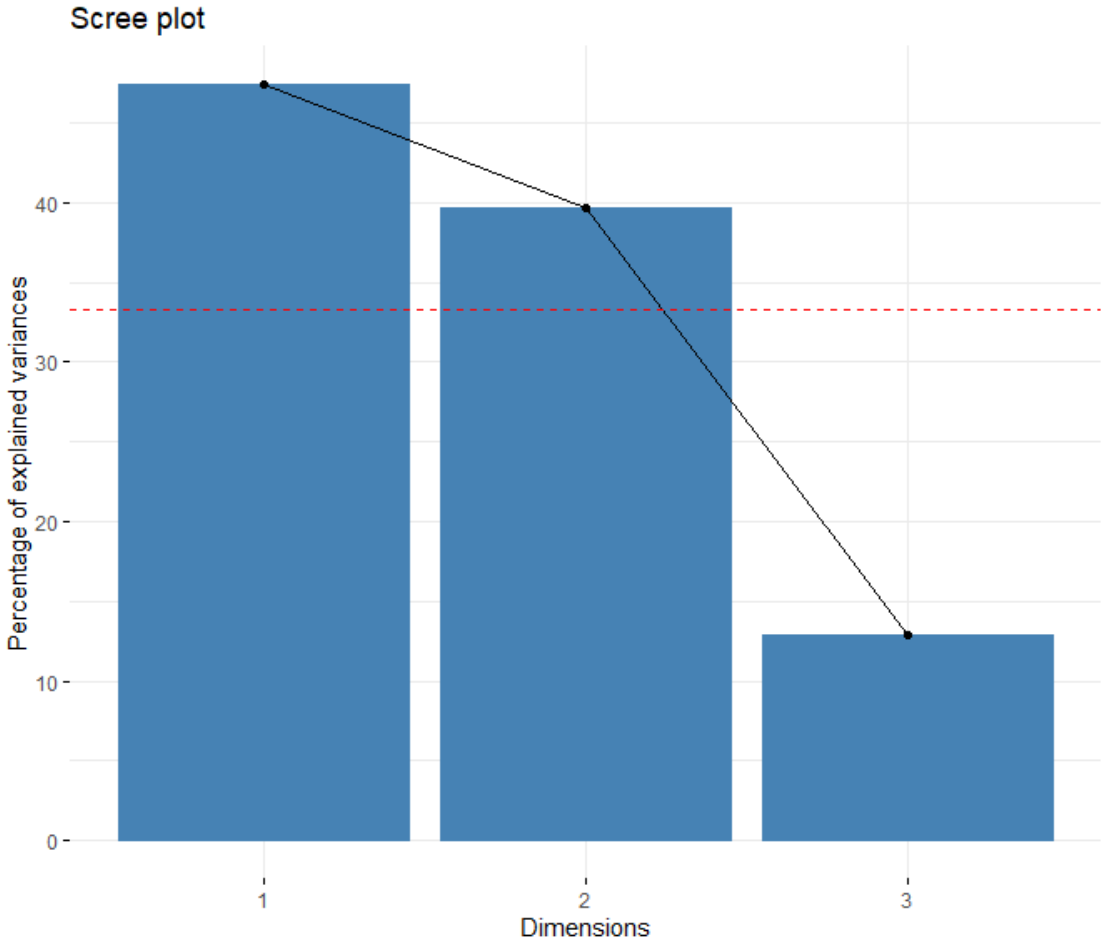
Dans notre analyse, les deux premiers axes expliquent 87% de la variance totale. C'est un pourcentage acceptable. Une autre méthode pour déterminer le nombre de dimensions est de regarder le graphique des valeurs propres, ordonnées de la plus grande à la plus petite valeur. Le nombre d'axes est déterminé par le au-delà duquel les valeurs propres restantes sont toutes relativement petites et de tailles comparables.



Le point auquel le graphique des valeurs propres montre un virage peut être considéré comme indiquant le nombre optimal d'axes principaux à retenir.

Il est également possible de calculer une valeur propre moyenne au-dessus de laquelle l'axe doit être conservé dans le résultat. On choisit dans notre cas de garder les deux premiers

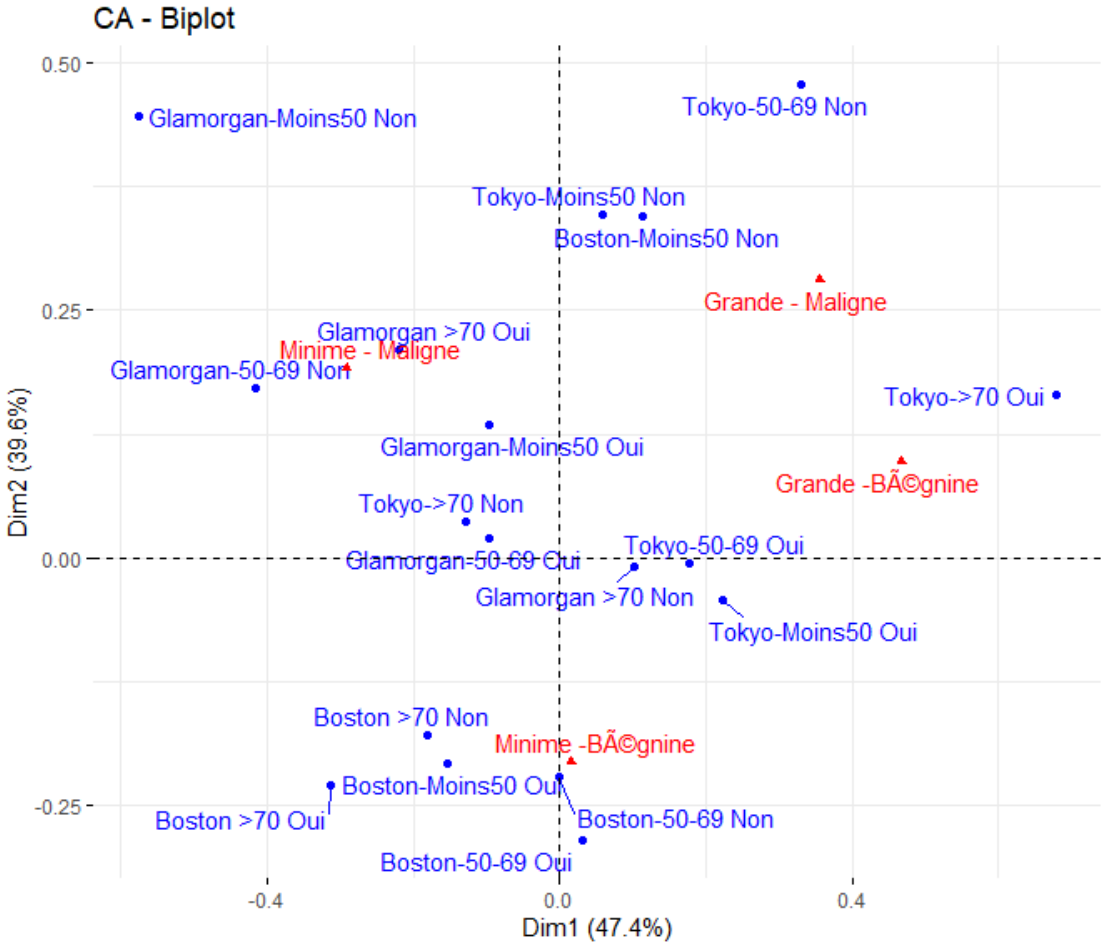
axes car ils expliquent 87 % de l'inertie total comme le montre le graphique suivant :



Biplot

La distance entre les points lignes ou entre les points colonnes donne une mesure de leur similitude (ou dissemblance). Les points lignes avec un profil similaire sont proches sur le graphique. Il en va de même pour les points colonnes. Ce graphique représente une analyse symétrique montrant les profils lignes et colonnes simultanément dans un espace commun. Dans ce cas, seule la distance entre les points lignes ou la distance entre les points colonnes peut être vraiment interprétée. La distance entre les points lignes et les points colonnes n'a pas de sens dans l'absolue. On peut remarquer un cluster en bas du graphique près de Minime – Bénine. Il y en a un autre au origine des axes puis lorsqu'on remontes, les profils

lignes sont plus éparses.



Graphique des points lignes

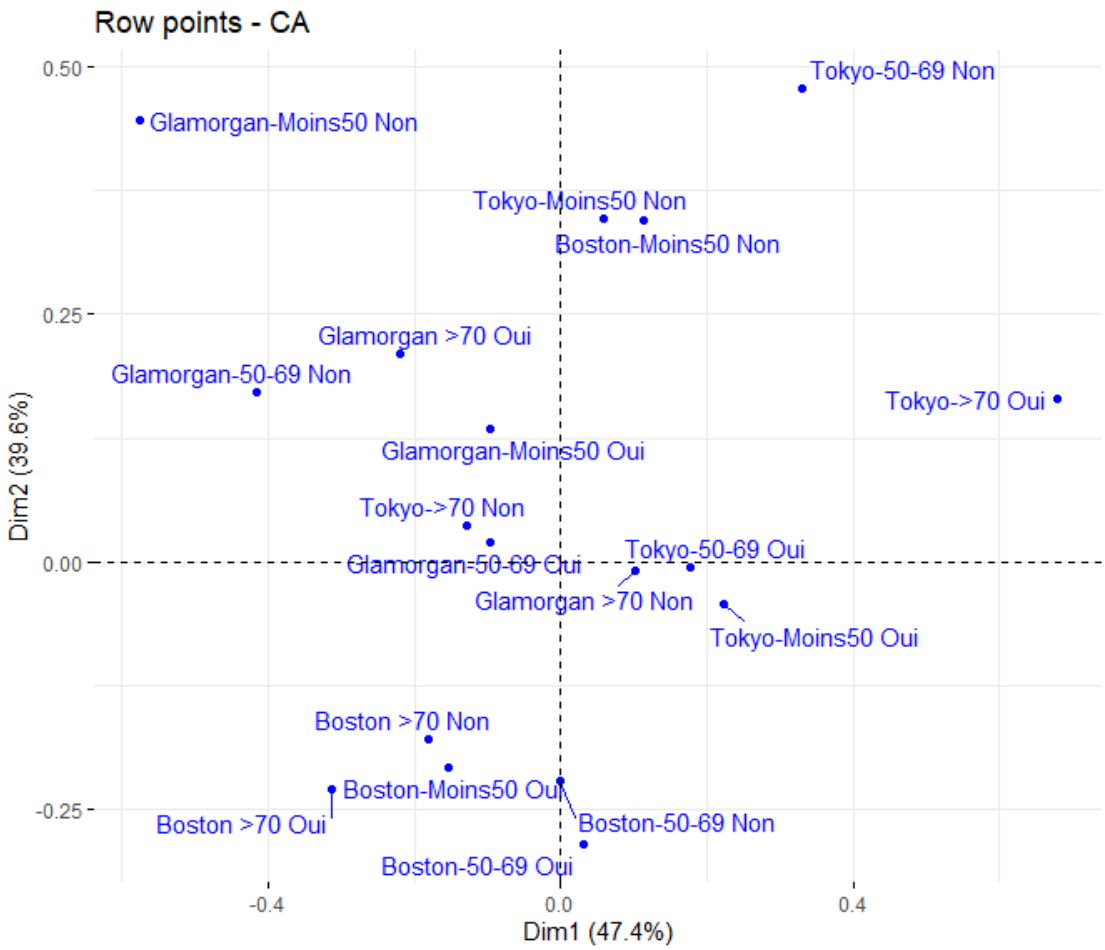
Cette fonction renvoie une liste contenant les coordonnées, les cos2 et les contributions des lignes :

```
## Correspondence Analysis - Results for rows
##  ================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the rows"
## 2 "$cos2"    "Cos2 for the rows"
## 3 "$contrib" "contributions of the rows"
## 4 "$inertia" "Inertia of the rows"

##                         Dim 1        Dim 2        Dim 3
## Tokyo-Moins50 Non  0.05984397  0.347242822  0.425183087
## Tokyo-Moins50 Oui  0.22261158 -0.041467169  0.058000495
## Tokyo-50-69 Non    0.32963070  0.478545525 -0.078988897
## Tokyo-50-69 Oui    0.17683947 -0.005357785 -0.002692771
## Tokyo->70 Non     -0.12748578  0.036221714 -0.176177265
## Tokyo->70 Oui      0.67869696  0.165272828 -0.139690409
```

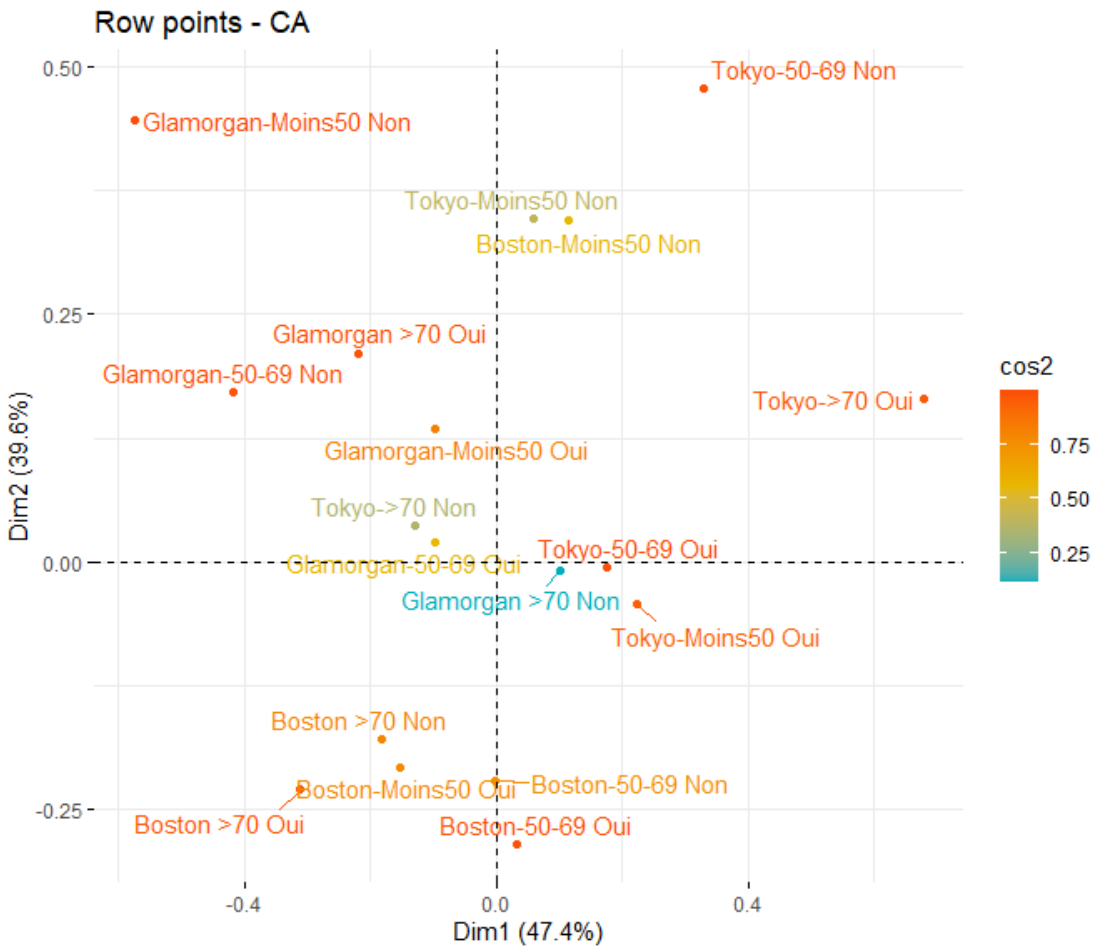
```
##                        Dim 1        Dim 2        Dim 3
## Tokyo-Moins50 Non 0.0117443 0.3954147085 0.5928409939
## Tokyo-Moins50 Oui 0.9069614 0.0314703869 0.0615682339
## Tokyo-50-69 Non   0.3159521 0.6659053345 0.0181425392
## Tokyo-50-69 Oui   0.9988515 0.0009168814 0.0002316015
## Tokyo->70 Non     0.3343950 0.0269944416 0.6386105227
## Tokyo->70 Oui     0.9077194 0.0538273181 0.0384532454

##                         Dim 1        Dim 2        Dim 3
## Tokyo-Moins50 Non  0.1991041  8.019375916 36.881972897
## Tokyo-Moins50 Oui 15.3326409  0.636450939  3.819510533
## Tokyo-50-69 Non    8.1419514 20.528366582  1.715647794
## Tokyo-50-69 Oui    6.7275959  0.007387658  0.005724312
## Tokyo->70 Non      0.2357145  0.022763311  1.651905974
## Tokyo->70 Oui     14.4746032  1.026815588  2.250146923
```

Le graphique suivant montre les relations entre les points lignes :

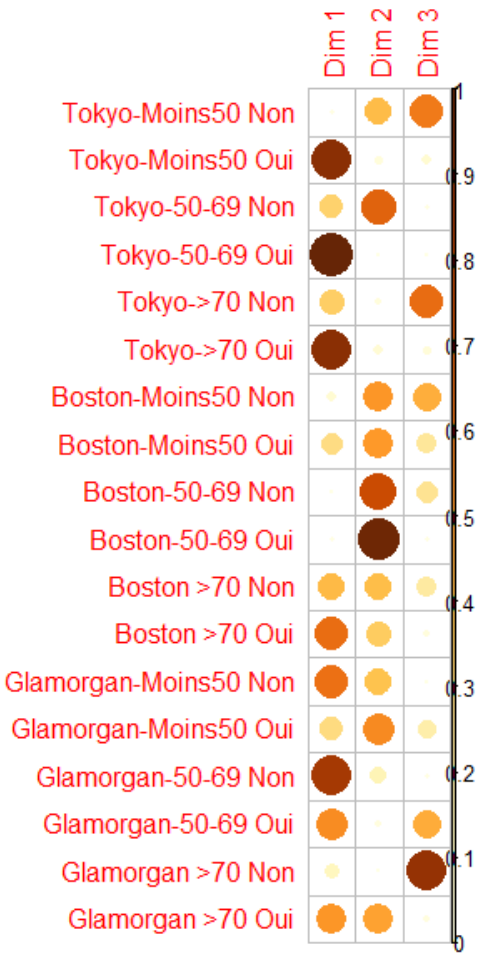
- Les lignes avec un profil similaire sont regroupées.
- Les lignes corrélées négativement sont positionnées sur des côtés opposés de l'origine du graphique
- La distance entre les points lignes et l'origine mesure la qualité des points lignes sur le graphique (on y voit un « cluster »). Les points lignes qui sont loin de l'origine sont bien représentés sur le graphique.
- Même si deux dimensions suffisent pour expliquer 87% des données, cela ne veut pas dire que tous les points sont aussi bien représentés sur ces dimensions.

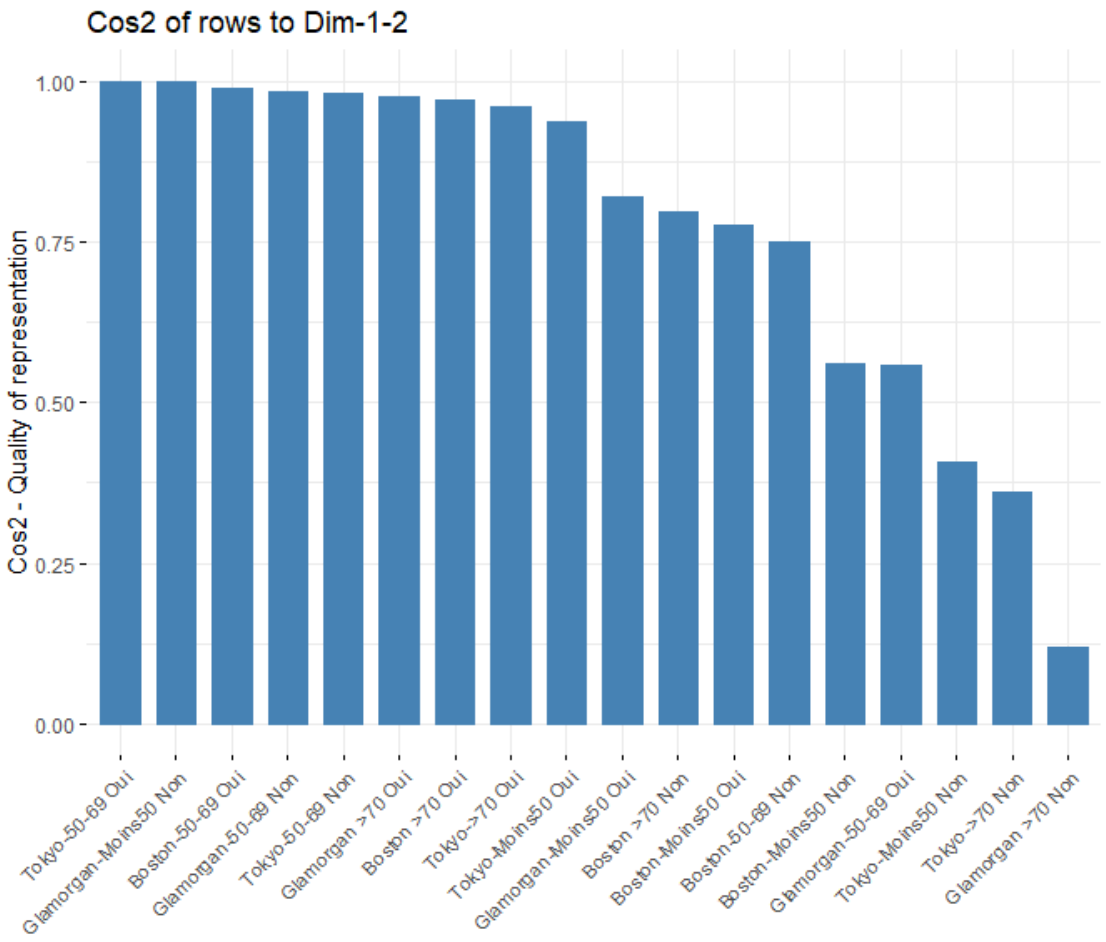


Sur ce graphique, on voit que les patients soignés à Glamorgan de plus de 70 ans qui n'ont pas survécu sont mal représentés sur ces dimensions. Les patients soignés à Tokyo de plus de 70 ans et de moins de 50 ans qui n'ont pas survécu ne sont pas bien représentés ici. Il en est de même pour les personnes soignées à Boston de moins de 50 ans qui sont morts et les patients de Glamorgan de 50-69 ans qui sont encore en vie ne sont pas très bien représentés dans ce graphique. Les autres combinaisons de modalités sont bien représentées sur le graphique.



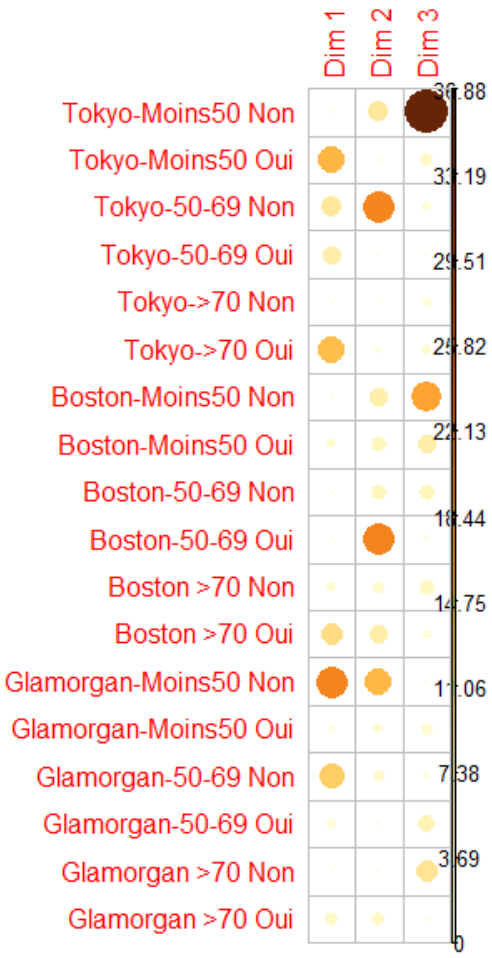
On peut visualiser différemment les mêmes informations.


Dim 1
Dim 2
Dim 3
Tokyo-Moins50 Non
Tokyo-Moins50 Oui
Tokyo-50-69 Non
Tokyo-50-69 Oui
Tokyo->70 Non
Tokyo->70 Oui
Boston-Moins50 Non
Boston-Moins50 Oui
Boston-50-69 Non
Boston-50-69 Oui
Boston >70 Non
Boston >70 Oui
Glamorgan-Moins50 Non
Glamorgan-Moins50 Oui
Glamorgan-50-69 Non
Glamorgan-50-69 Oui
Glamorgan >70 Non
Glamorgan >70 Oui
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0



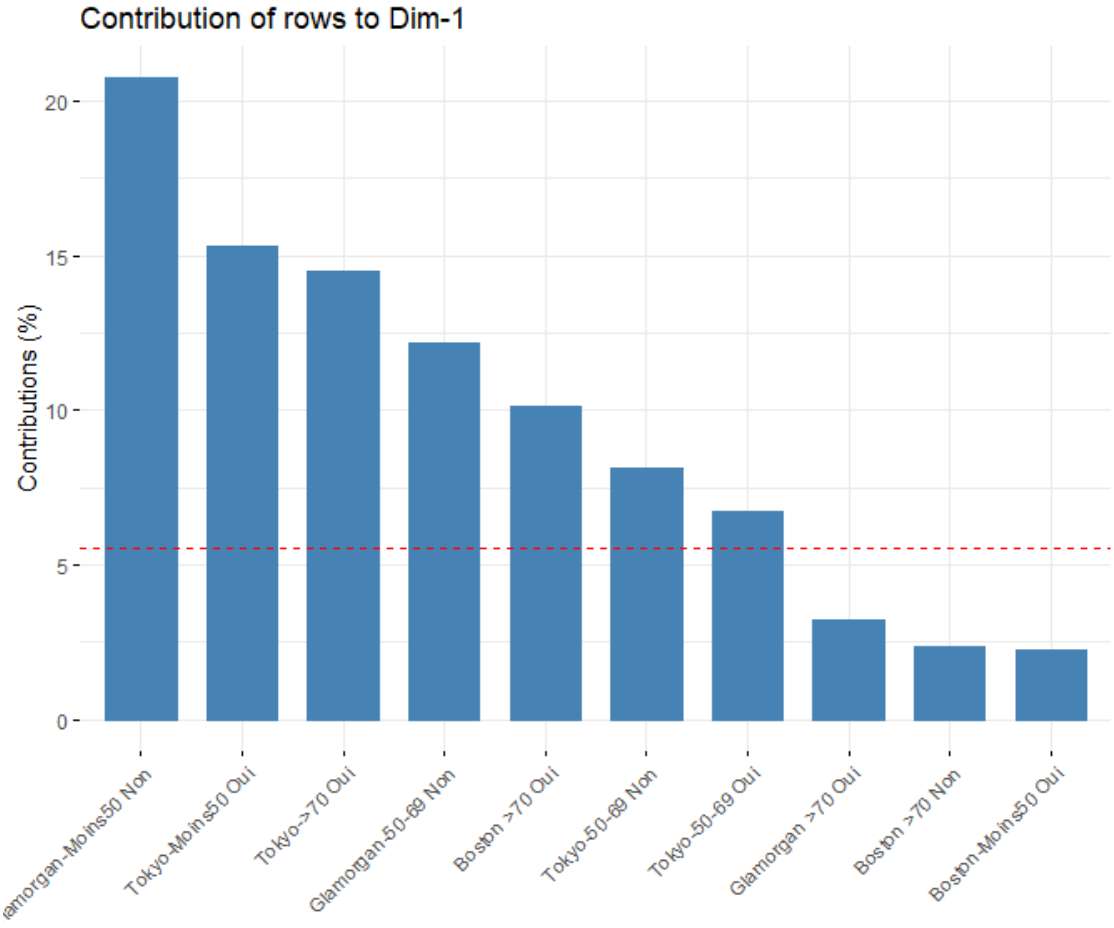
Les lignes avec des valeurs élevées, contribuent le mieux à la définition des dimensions.

- Les lignes qui contribuent le plus à Dim.1 et Dim.2 sont les plus importantes pour expliquer la variabilité dans le jeu de données.
- Les lignes qui ne contribuent pas beaucoup à aucune dimension ou qui contribuent aux dernières dimensions sont moins importantes. Ici, les lignes qui contribuent le plus au Dim 1 et 2 sont les personnes soignées à Tokyo de moins de 50 ans qui sont encore en vie, de 50 à 69 qui sont morts et de plus de 70 ans qui sont encore en vie. Les patients soignés à Boston qui ont entre 50 et 69 ans et qui sont encore en vie y contribuent aussi et finalement, les personnes soignées à Glamorgan qui ont moins de 50 ans décédés, et ceux de 50 à 69 ans qui sont aussi morts.

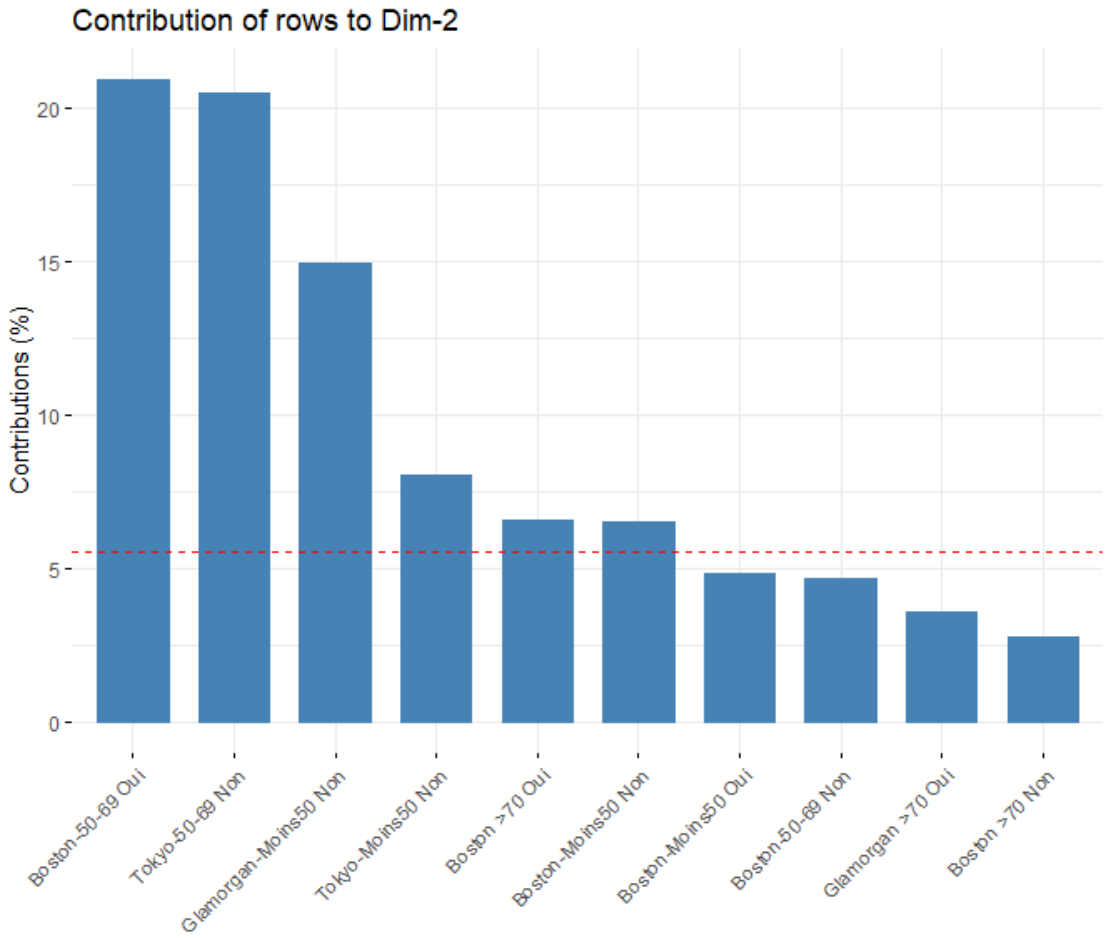


Sur ce graphique on peut observer le top 10 de la contribution de chaque ligne à la dimension 1. La droite en pointillée rouge, sur les graphiques suivant indique la

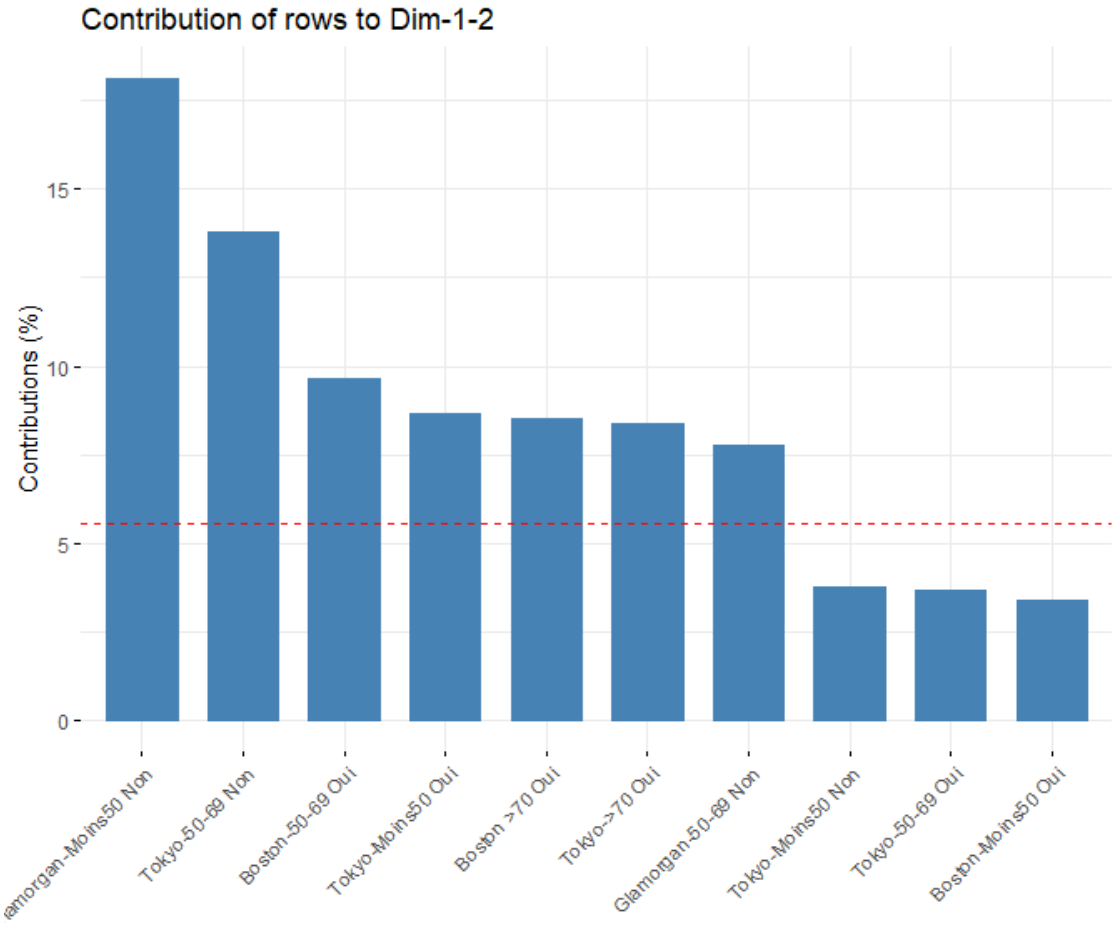
valeur moyenne attendue, si les contributions étaient uniformes



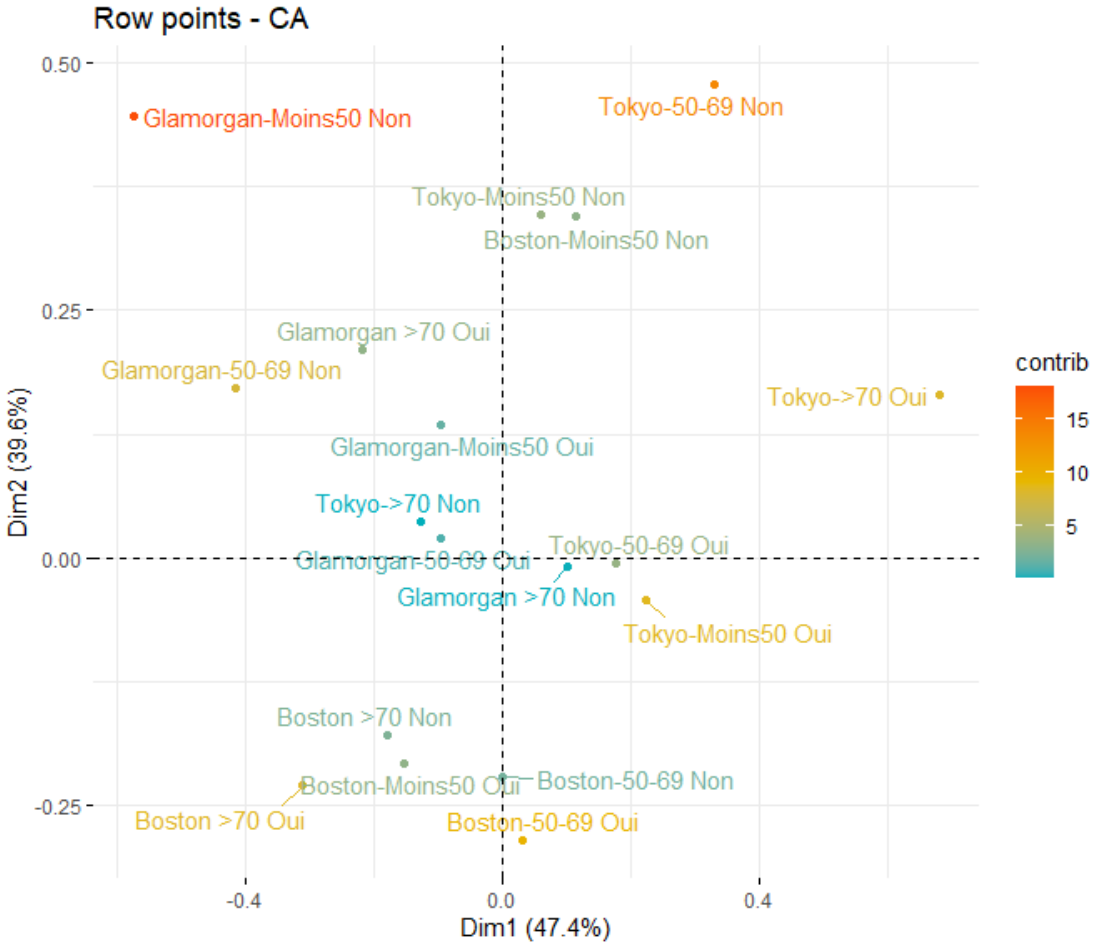
Sur ce graphique on peut observer le top 10 la contribution de chaque ligne à la dimension 2 :


Contribution of rows to Dim-2
Contributions (%)
20
15
10
5
0
Boston-50-69 Oui
Tokyo-50-69 Non
Glamorgan-Moins50 Non
Tokyo-Moins50 Non
Boston >70 Oui
Boston-Moins50 Non
Boston-Moins50 Oui
Boston-50-69 Non
Glamorgan >70 Oui
Boston >70 Non

Sur ce graphique on peut observer le top 10 la contribution de chaque ligne à la dimension 1 et 2 :



Le graphique donne une idée de la contribution des lignes aux différents dimensions.



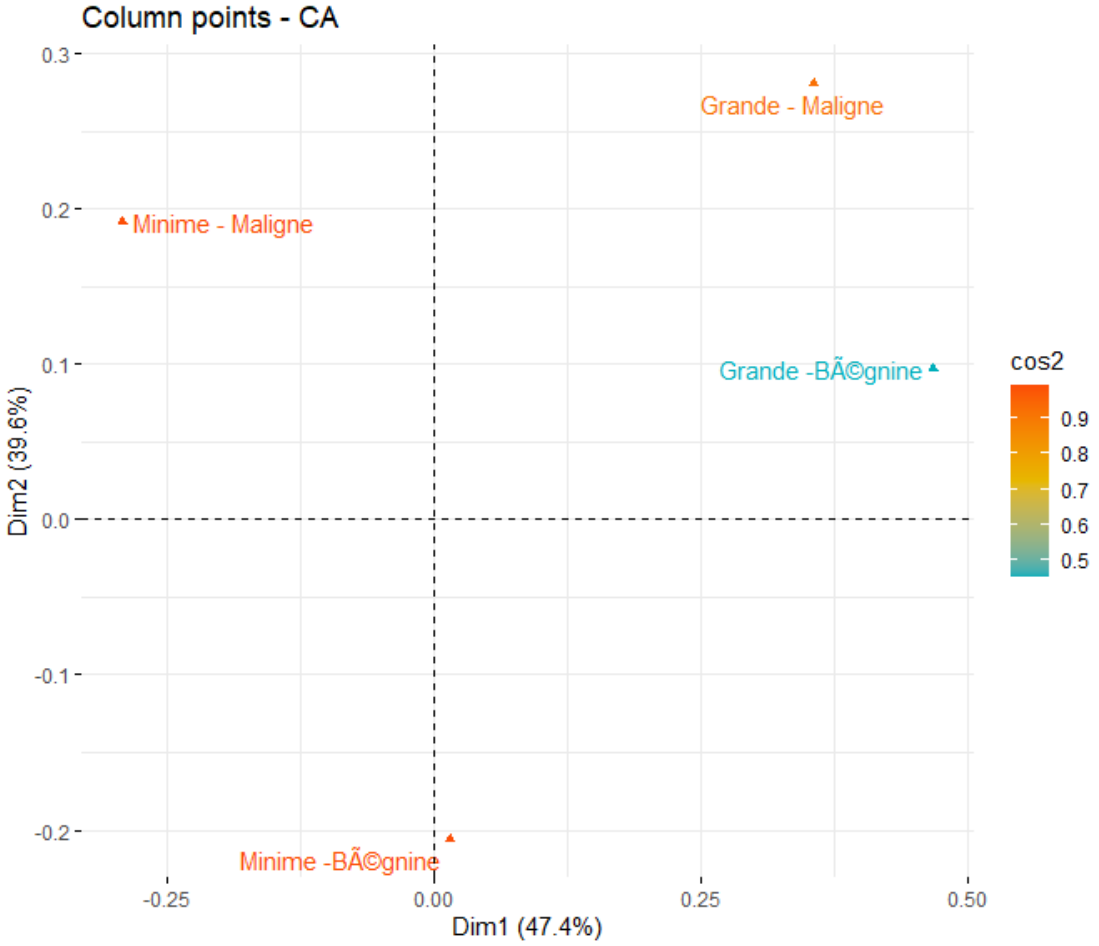
Graphes des colonnes

On reprend les mêmes étapes pour les colonnes

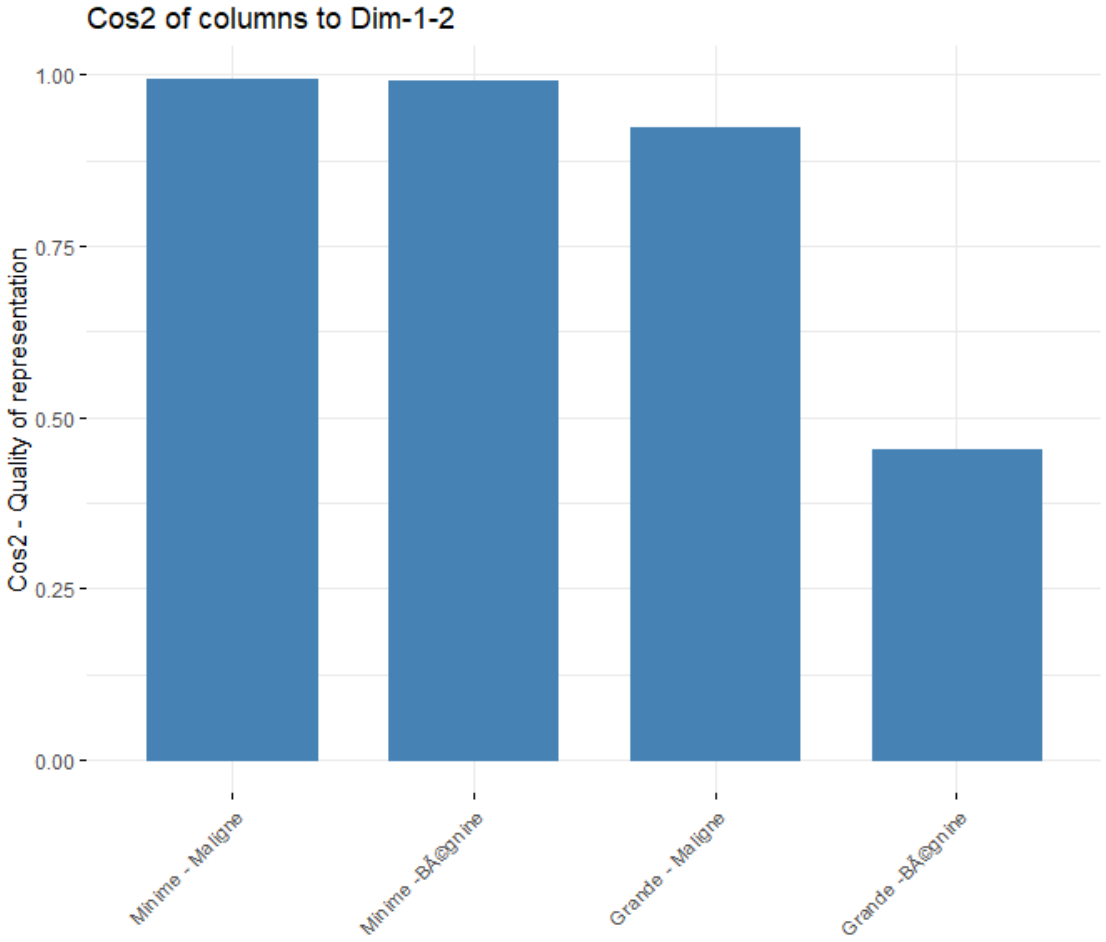
```
## Correspondence Analysis - Results for columns
##  ==================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the columns"
## 2 "$cos2"    "Cos2 for the columns"
## 3 "$contrib" "contributions of the columns"
## 4 "$inertia" "Inertia of the columns"

##                         Dim 1       Dim 2       Dim 3
## Minime - Maligne -0.29097573  0.19181658  0.03097367
## Minime -BÉgnine  0.01592089 -0.20610912 -0.01954645
## Grande - Maligne  0.35619268  0.28093234 -0.13171457
## Grande -BÉgnine  0.46761882  0.09729096  0.52428265

##                         Dim 1      Dim 2       Dim 3
## Minime - Maligne 0.691610711 0.30055258 0.007836706
## Minime -BÉgnine 0.005878829 0.98525998 0.008861194
```

```
## Grande - Maligne 0.568568701 0.35368485 0.077746454
## Grande -BÉgnine 0.434721879 0.01881796 0.546460163

##                        Dim 1      Dim 2     Dim 3
## Minime - Maligne 45.4336905 23.6195407  1.889177
## Minime -BÉgnine   0.2377265 47.6620054  1.314928
## Grande - Maligne 37.4146174 27.8425810 18.774215
## Grande -BÉgnine 16.9139656  0.8758729 78.021680
```

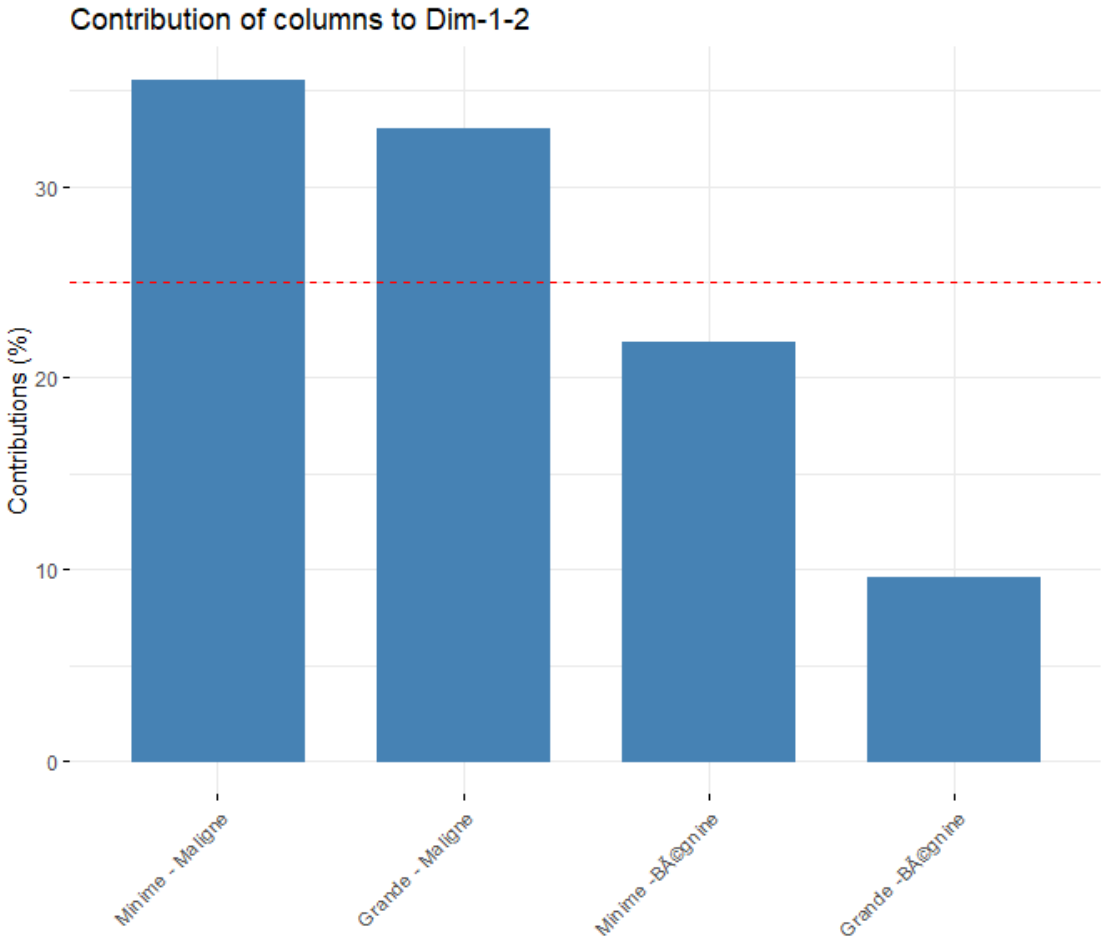
On voit ici que l'histologie "Grande inflammation bégnine" est mal représentée sur ces dimensions, "Grande inflammation Maligne" l'est un peu plus et les inflammations minime sont bien représentées.



On peut remarquer la même chose sur ce graphique :

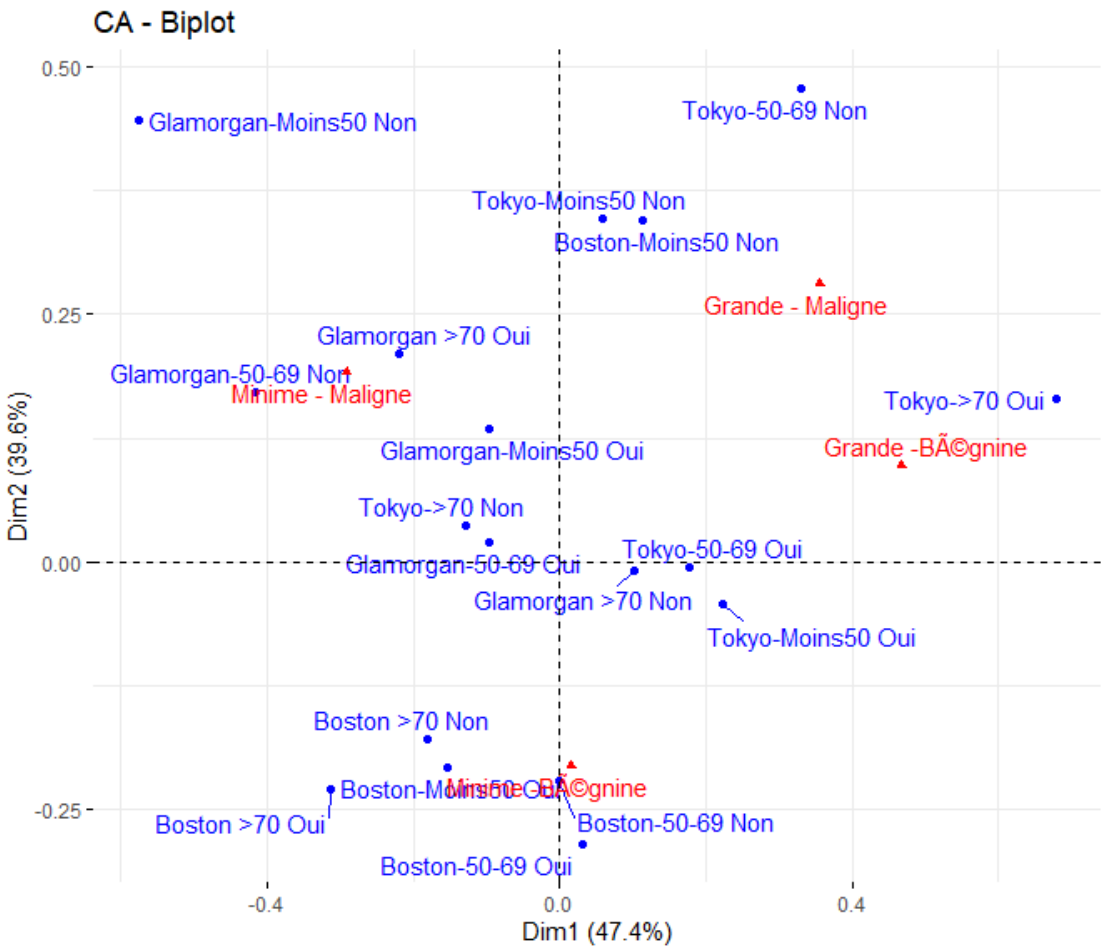


Sur ce graphique on remarque que les inflammations Maligne contribuent le plus aux dimensions 1 et 2.

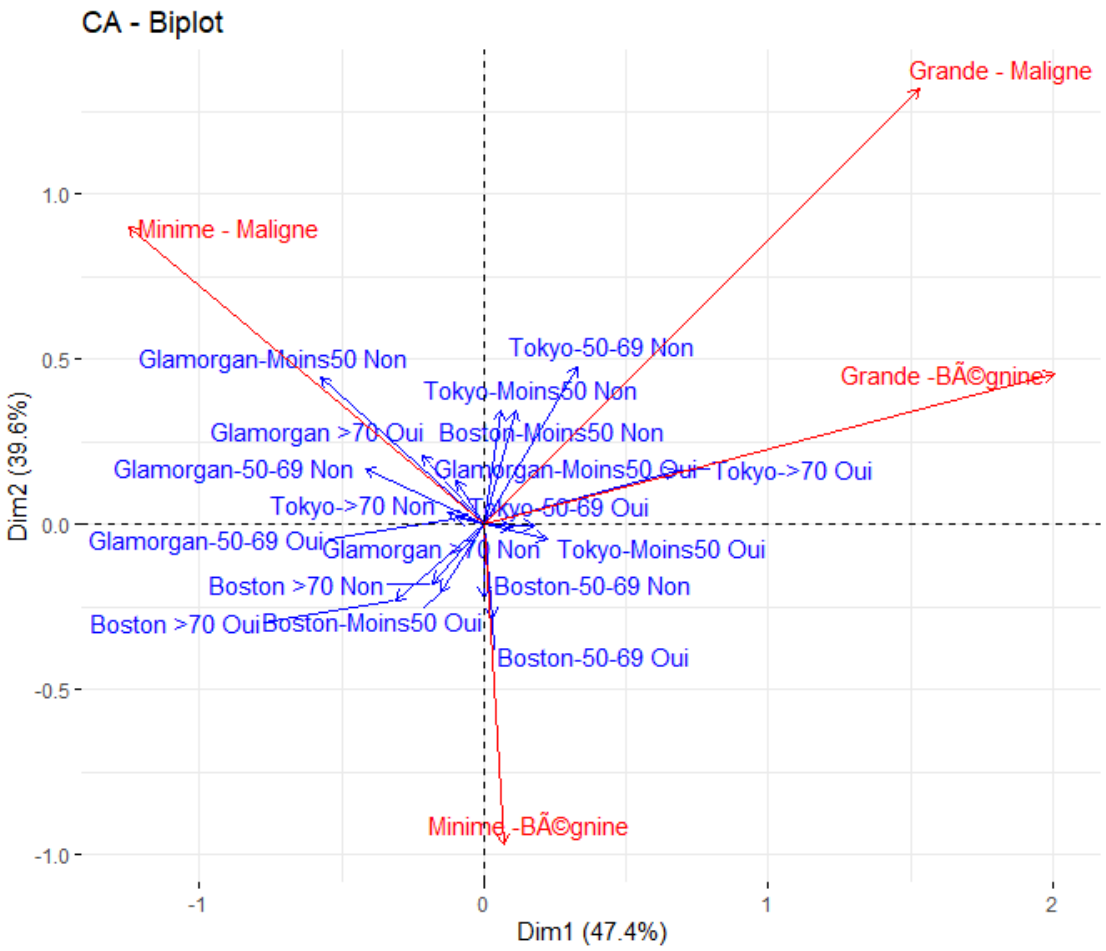


Biplot symétrique.

Comme mentionné ci-dessus, le graphique standard de l'AFC est un biplot symétrique dans lequel les lignes (points bleus) et les colonnes (triangles rouges) sont représentées dans le même espace à l'aide des coordonnées principales. Ces coordonnées représentent les profils des lignes et des colonnes. Dans ce cas, seule la distance entre les points lignes ou la distance entre les points colonnes peut être vraiment interprétée.



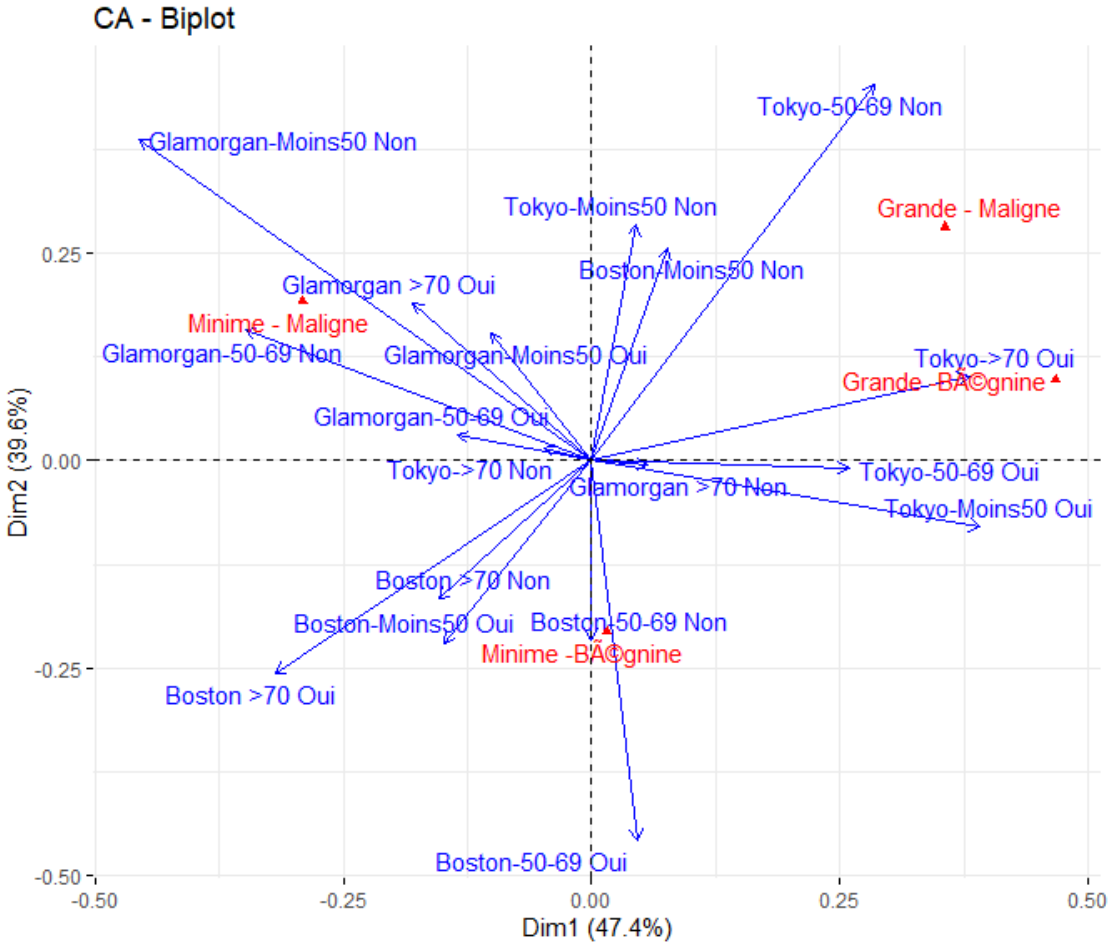
Pour interpréter la distance entre les points colonnes et les points lignes, le moyen le plus simple est de créer un biplot asymétrique. Cela signifie que les profils des colonnes doivent être représentés dans l'espace des lignes ou vice versa. Pour interpréter le résultat, il est important de regarder la position des observations et des variables sur le graphique, ainsi que la longueur et la direction des vecteurs. Les observations et les variables qui sont proches les uns des autres dans l'espace des composantes principales sont similaires, tandis que celles qui sont éloignées l'une de l'autre sont différentes. Les vecteurs des observations et des variables qui ont une longueur plus importante ont une plus grande influence sur les composantes principales, tandis que ceux qui ont une longueur plus petite exercent une influence moins importante. Les vecteurs qui pointent dans une direction similaire indiquent une relation positive entre les observations ou les variables, tandis que ceux qui pointent dans des directions opposées indiquent une relation négative.



Dans le graphique ci-dessus, la position des points colonnes est inchangée par rapport à celle du biplot conventionnel. Cependant, les distances entre les points lignes et l'origine du graphique sont liées à leurs contributions aux axes principaux en considération. Plus une flèche est proche (en termes de distance angulaire) d'un axe, plus la contribution de la ligne sur cet axe par rapport à l'autre axe est importante. Si la flèche est à mi-chemin entre les deux axes, la ligne contribue aux deux axes de manière identique.

Ici on voit que les inflammation grande et maligne contribuent énormément aux dimensions et de façon égale et positive. Les inflammations Minime et maligne contribuent bien aux dimensions et également mais de façons négatives pour la dimension 1. Les inflammations grande et bégnine contribuent plus à la dimensions 2. Les inflammations minime et bégnine semblent contribuer le moins mais elles contribuent plus à la première

dimension de façon positive.



De la même façon, les personnes soignées à Glamorgan de moins de 50 ans et qui sont décédés contribuent de façon négative pour la Dim 1. Les patients de Boston de 0 à 69 ans qui ont survécu contribuent eux beaucoup pour la dimension 1 et peu négativement pour la dimension 1.

Description des dimensions

```
## $`Dim 1`
## $`Dim 1`$row
##                              coord
## Glamorgan-Moins50 Non -0.5744567093
## Glamorgan-50-69 Non   -0.4169974512
## Boston >70 Oui        -0.3124966691
## Glamorgan >70 Oui     -0.2185808860
## Boston >70 Non        -0.1810090480
## Boston-Moins50 Oui    -0.1535880116
## Tokyo->70 Non         -0.1274857780
## Glamorgan-Moins50 Oui -0.0974102059
## Glamorgan-50-69 Oui   -0.0968534189
## Boston-50-69 Non      -0.0007256658
## Boston-50-69 Oui       0.0314157515
```

```
## Tokyo-Moins50 Non      0.0598439691
## Glamorgan >70 Non      0.1015160483
## Boston-Moins50 Non     0.1137101085
## Tokyo-50-69 Oui        0.1768394658
## Tokyo-Moins50 Oui      0.2226115771
## Tokyo-50-69 Non        0.3296307007
## Tokyo->70 Oui          0.6786969628
##
## $`Dim 1`$col
##                         coord
## Minime - Maligne -0.29097573
## Minime -BÉgnine  0.01592089
## Grande - Maligne  0.35619268
## Grande -BÉgnine  0.46761882

##                            coord
## Glamorgan-Moins50 Non -0.5744567
## Glamorgan-50-69 Non   -0.4169975
## Boston >70 Oui        -0.3124967
## Glamorgan >70 Oui     -0.2185809

##                         coord
## Minime - Maligne -0.29097573
## Minime -BÉgnine  0.01592089
## Grande - Maligne  0.35619268
## Grande -BÉgnine  0.46761882

##                              coord
## Boston-50-69 Oui      -0.285205849
## Boston >70 Oui        -0.229914288
## Boston-50-69 Non      -0.220807277
## Boston-Moins50 Oui    -0.206437675
## Boston >70 Non        -0.178737147
## Tokyo-Moins50 Oui     -0.041467169
## Glamorgan >70 Non     -0.008864303
## Tokyo-50-69 Oui       -0.005357785
## Glamorgan-50-69 Oui    0.019933669
## Tokyo->70 Non          0.036221714
## Glamorgan-Moins50 Oui  0.134910661
## Tokyo->70 Oui          0.165272828
## Glamorgan-50-69 Non    0.170978380
## Glamorgan >70 Oui      0.210775254
## Boston-Moins50 Non     0.344782115
## Tokyo-Moins50 Non      0.347242822
## Glamorgan-Moins50 Non  0.446359558
## Tokyo-50-69 Non        0.478545525

##                         coord
## Minime -BÉgnine -0.20610912
## Grande -BÉgnine  0.09729096
```

```
## Minime - Maligne  0.19181658
## Grande - Maligne  0.28093234
```

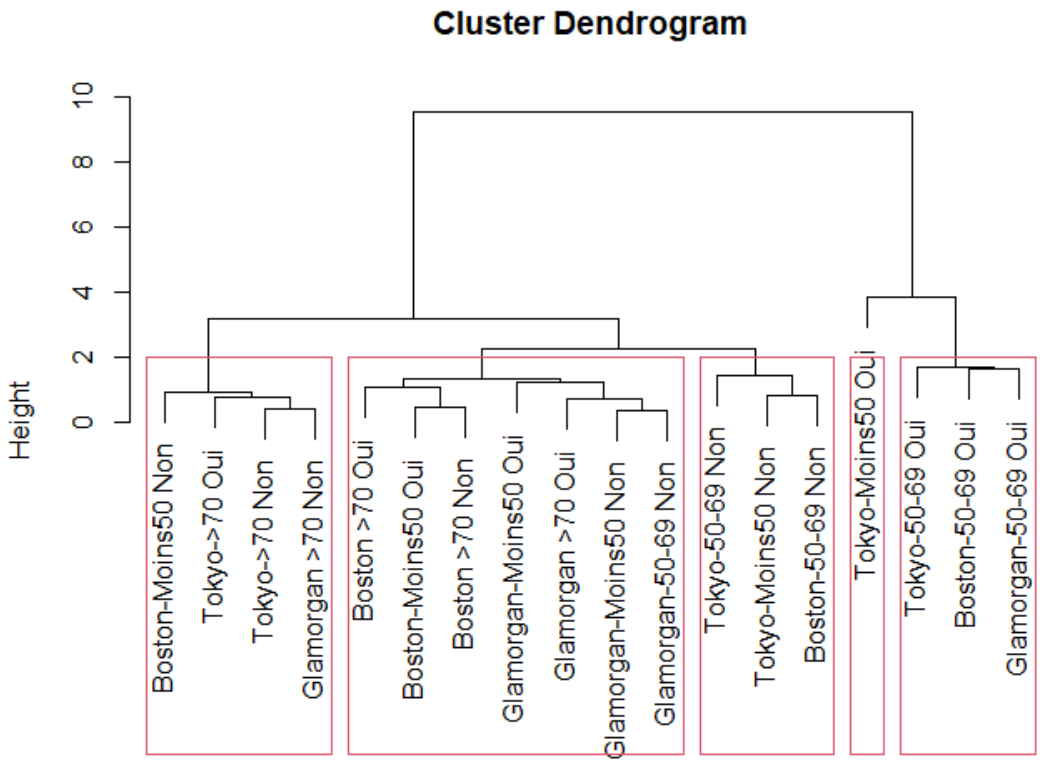
Classification

Rappel du tableau de données :

```
##                   Minime - Maligne Minime -BÉgnine Grande - Maligne
## Tokyo-Moins50 Non                9                7                4
## Tokyo-Moins50 Oui               26               68               25
## Tokyo-50-69 Non                  9                9               11
## Tokyo-50-69 Oui                 20               46               18
## Tokyo->70 Non                    2                3                1
## Tokyo->70 Oui                    1                6                5
##                   Grande -BÉgnine
## Tokyo-Moins50 Non                3
## Tokyo-Moins50 Oui                9
## Tokyo-50-69 Non                  2
## Tokyo-50-69 Oui                  5
## Tokyo->70 Non                    0
## Tokyo->70 Oui                    1

## Number of cases in table: 764
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 87.23, df = 51, p-value = 0.001188
##  Chi-squared approximation may be incorrect
```

CAH

On centre et réduit les données pour éviter que les variables à forte variance pèsent sur les résultats. On calcule la matrice des distances entre individus.



.

```
##                          Minime - Maligne Minime -BÉgnine Grande - Maligne
## Boston-Moins50 Non                     6                7                6
## Tokyo->70 Oui                          1                6                5
## Tokyo->70 Non                          2                3                1
## Glamorgan >70 Non                      3                7                3
## Boston >70 Oui                        15               26                1
## Boston-Moins50 Oui                    11               24                4
## Boston >70 Non                         9               18                3
## Glamorgan-Moins50 Oui                 16               20                8
## Glamorgan >70 Oui                     12               11                4
## Glamorgan-Moins50 Non                 16                7                3
## Glamorgan-50-69 Non                   14               12                3
## Tokyo-50-69 Non                        9                9               11
## Tokyo-Moins50 Non                      9                7                4
## Boston-50-69 Non                       8               20                3
## Tokyo-Moins50 Oui                     26               68               25
## Tokyo-50-69 Oui                       20               46               18
## Boston-50-69 Oui                      18               58               10
## Glamorgan-50-69 Oui                   27               39               10
```

```
##                         Grande -BÉgnine
## Boston-Moins50 Non                    0
## Tokyo->70 Oui                         1
## Tokyo->70 Non                         0
## Glamorgan >70 Non                     0
## Boston >70 Oui                        1
## Boston-Moins50 Oui                    0
## Boston >70 Non                        0
## Glamorgan-Moins50 Oui                 1
## Glamorgan >70 Oui                     1
## Glamorgan-Moins50 Non                 0
## Glamorgan-50-69 Non                   0
## Tokyo-50-69 Non                       2
## Tokyo-Moins50 Non                     3
## Boston-50-69 Non                      2
## Tokyo-Moins50 Oui                     9
## Tokyo-50-69 Oui                       5
## Boston-50-69 Oui                      3
## Glamorgan-50-69 Oui                   4
```