Projet de statistique

Classification des pays

Moussa DEME - Aïba FOFANA

tuteurs :
MR. PREDA - MR. GRIMONPREZ

Polytech Lille
Avril 2020

Contents

1 Jeu de données

1.1 But du projet

Le jeu de données correspond à la consommation de plusieurs types de vins (exprimée en KiloLitres/an) dans huit pays d'Europe et d'Amérique. Il est composé de 18 variables quantitatives (type de vins) pour 8 observations (les pays).
L'objectif de notre étude est de faire une analyse descriptive multivariée ainsi qu'une classification des pays. Pour ce faire, nous allons utiliser la méthode de l'Analyse en Composantes Principales (ACP) car nos données sont quantitatives. De plus, l'ACP nous permettra de réduire la dimension de notre jeu de données initial tout en gardant le maximum d'information, rendant ainsi facile la classification.

```
vins = read.table("vins.csv", sep=",", header = TRUE, row.names = 1)
str(vins)
```

```
'data.frame':   18 obs. of  8 variables:
 $ BELGIQUE : int  7069 2436 3066 2422 22986 17465 3784 7950 2587 17200 ...
 $ NEDERLAND: int  3786 586 290 1999 22183 19840 2339 10537 600 22806 ...
 $ RFA      : int  12578 2006 10439 17183 21023 72977 4828 7552 2101 15979 ...
 $ ITALIE   : int  8037 30 1413 57 56 2364 98 24 0 50 ...
 $ UK       : int  13556 1217 7214 1127 30025 39919 7885 8172 7582 20004 ...
 $ SUISSE   : int  9664 471 112 600 6544 17327 3191 11691 143 1279 ...
 $ USA      : int  10386 997 3788 408 13114 17487 11791 1369 872 4016 ...
 $ CANADA   : int  206 51 330 241 3447 2346 1188 1798 131 944 ...
```

1.2 Pré-traitement et visualisation des données

Le jeu de données dans sa version initiale présente les variables en ligne et les observations en colonne. Pour faciliter la réalisation de notre étude, nous avons donc décidé de le transposer et de le mettre dans un dataframe. En effet, l'opération de transposition met le jeu de données dans un objet de type "matrix" R, qui considère que toutes les variables sont du même type. Ici, cela n'a pas d'impact car nos variables sont toutes entières. Mais pour pouvoir référer les variables par leur nom par exemple (avec le "dollar"), un data.frame est plus adapté, ce qui explique notre choix.
Nous avons aussi remarqué que deux variables portent approximativement le même nom (les variables RHON et RHONE). N'ayant pas plus d'information, nous avons décidé de considérer ces variables comme étant distinct. Notre jeu de données final comporte 18 variables pour 8 observations. Un extrait de ce jeu de données est le suivant:

```
vins = t(vins)
vins = as.data.frame(vins)
knitr::kable(vins[,sample(c(1:18),5)])
```

	VDQS	TRES_FORT	AUTRE_FORT	RHONE	MOUSSEAU_AOC
BELGIQUE	1976	2415	24	785	2436
NEDERLAND	1029	74	1533	1648	586
RFA	1346	208	160	1009	2006
ITALIE	0	8	0	6	30
UK	2258	1705	480	775	1217
SUISSE	212	12	0	643	471
USA	1017	36	0	542	997
CANADA	487	47	0	35	51

1.2.1 Somme par ligne (individus)

```
sommeLigne = apply(vins, 1, sum)
knitr::kable(t(sommeLigne))
```

BELGIQUE	NEDERLAND	RFA	ITALIE	UK	SUISSE	USA	CANADA
134463	112705	364707	20143	257379	53363	92747	50643

Le pays le plus consommateur de vins de notre jeu de données est la République Fédérale D'Allemagne (RFA). En effet, 364707 KiloLitres de vins ont été consommés dans ce pays sur une année. Il est suivi du Royaume Uni (UK) et de la Belgique.

1.2.2 Somme par colonne (variables)

```
sommeCol = apply(vins, 2, sum)
sommeCol
```

```
       CHMPAGNE    MOUSSEAU_AOC MOUSSEAU_SIMPLE          ALSACE
          65282            7794           26652           24037
        GIRONDE       BOJOLAIS        BORDEAUX            RHON
         119378          189725           35104           49093
           ANJOU     AOC_AUTRES            VDQS      AUTRE_VDQS
          14016           82278            8325          423862
       PROVENCE         MUSCAT           RHONE        AOC_FORT
           5969           20731            5443            1759
     AUTRE_FORT      TRES_FORT
           2197            4505
```

Le vin le plus consommé, tout pays confondu est "AUTRE_VDQS" avec 432862 kilolitres/an. Il s'agit d'un vin de qualité supérieure. Il est suivi par le vin "BOJOLAIS" avec 189725 kilolitres/an .

2 Analyse univariée des données

2.1 Résumés numériques

```
stat_uni = function(x) {
return(c(length(x),min(x),max(x),mean(x),median(x),sd(x)))
}
res_stat_uni = apply(vins, 2, "stat_uni")
row.names(res_stat_uni) = c("Nobs", "Min", "Max", "Moyenne", "Mediane", "Ecart-type")
knitr::kable(t(res_stat_uni), format = "markdown", align = 'r')
```

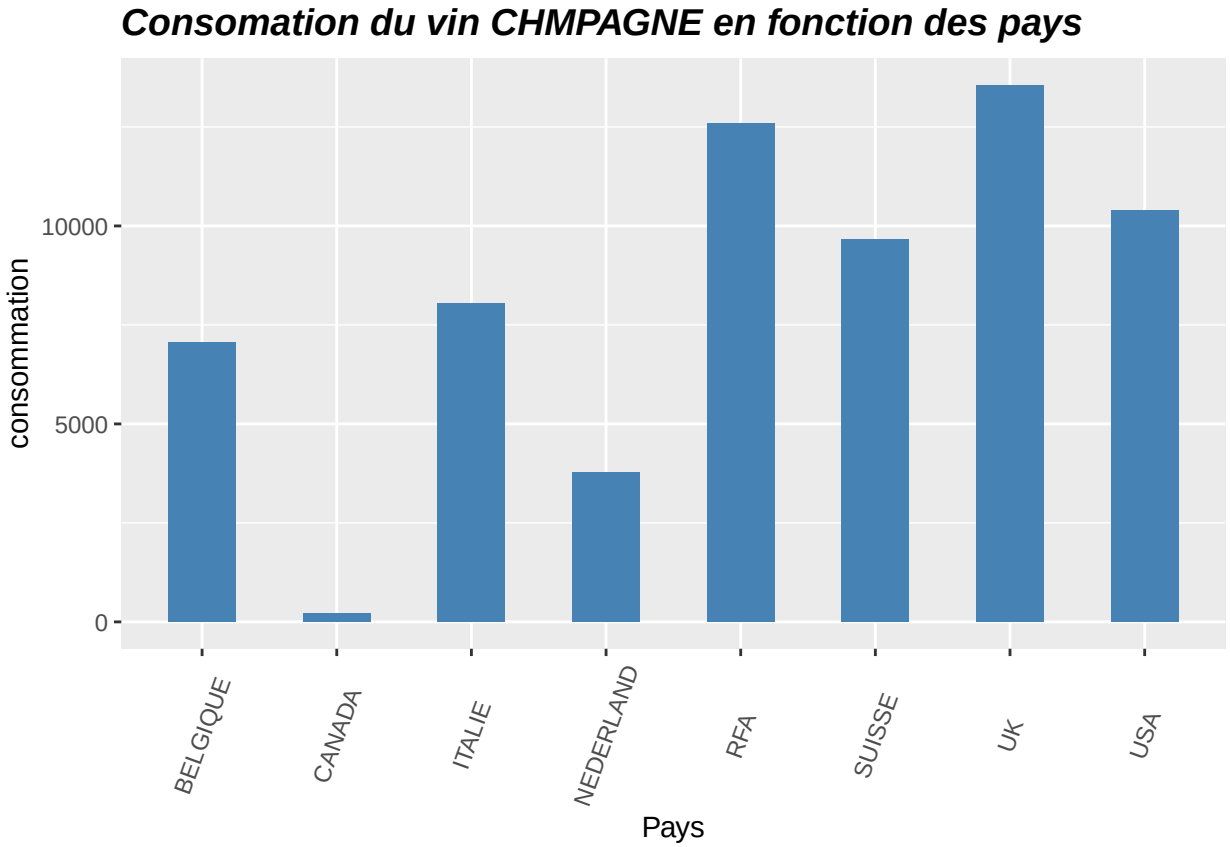
	Nobs	Min	Max	Moyenne	Mediane	Ecart-type
CHMPAGNE	8	206	13556	8160.250	8850.5	4463.4393
MOUSSEAU_AOC	8	30	2436	974.250	791.5	879.0143
MOUSSEAU_SIMPLE	8	112	10439	3331.500	2239.5	3744.9927
ALSACE	8	57	17183	3004.625	863.5	5791.0439
GIRONDE	8	56	30025	14922.250	17068.5	10751.7863
BOJOLAIS	8	2346	72977	23715.625	17476.0	23100.6687
BORDEAUX	8	98	11791	4388.000	3487.5	3814.7672
RHON	8	24	11691	6136.625	7751.0	4449.1869
ANJOU	8	0	7582	1752.000	736.0	2540.9631
AOC_AUTRES	8	50	22806	10284.750	9997.5	9590.8525
VDQS	8	0	2258	1040.625	1023.0	804.1762
AUTRE_VDQS	8	1029	191140	52982.750	32347.5	63697.3609
PROVENCE	8	0	2514	746.125	342.5	876.4323
MUSCAT	8	0	12891	2591.375	1122.5	4280.4211
RHONE	8	6	1648	680.375	709.0	528.6138
AOC_FORT	8	0	1177	219.875	80.5	397.0784
AUTRE_FORT	8	0	1533	274.625	12.0	534.9651
TRES_FORT	8	8	2415	563.125	60.5	945.2758

Ce tableau permet d'obtenir les moyennes pour chaque type de vins, les dispersions, les valeurs minimales, maximales et les médianes.
On peut voir que certains vins n'ont pas été consommés dans certains pays, leur minimum étant à zéro (vins VDQS, ANJOU, PROVENCE, MUSCAT, AOC_FORT, AUTRE_FORT). On constate aussi que le vin le plus consommé est "AUTRE_VDQS", qui correspond à un vin de qualité supérieur avec une moyenne de 52982.75 KiloLitres/an. Il est suivi du vin "BOJOLAIS" avec une moyenne de 23715.62 KiloLitres. La consommation des vins "AUTRE_VQS", "Bojolais", "Gironde" et "AOC_Autres" est très dispersée (variée) dans ces pays du fait des grandes valeurs des écart-types. Globalement, on peut donc dire que ces pays sont des consommateurs de vins de qualité, issus d'une zone géographique spécifique.

2.2 Représentation graphique

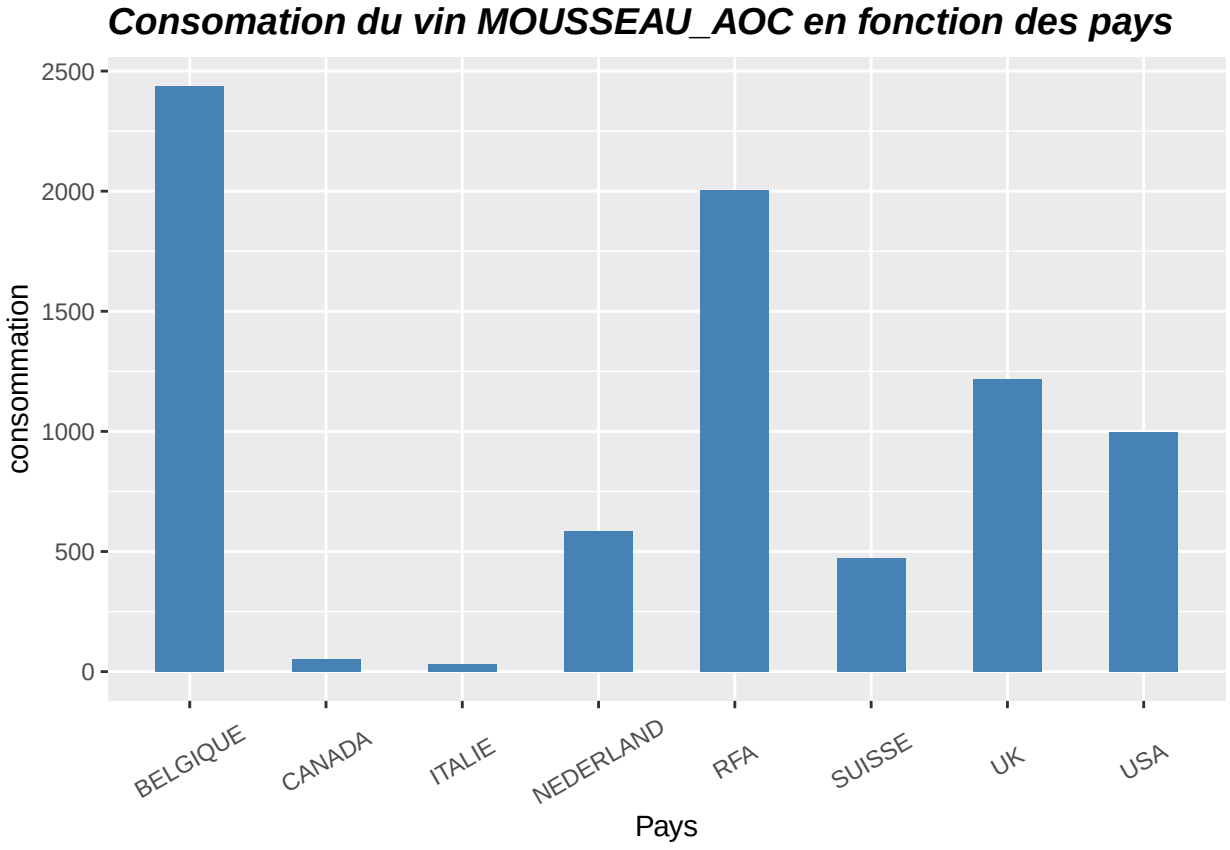
Les graphiques ci-dessous permettent de représenter la quantité de consommation des différents types de vins en fonction des pays.

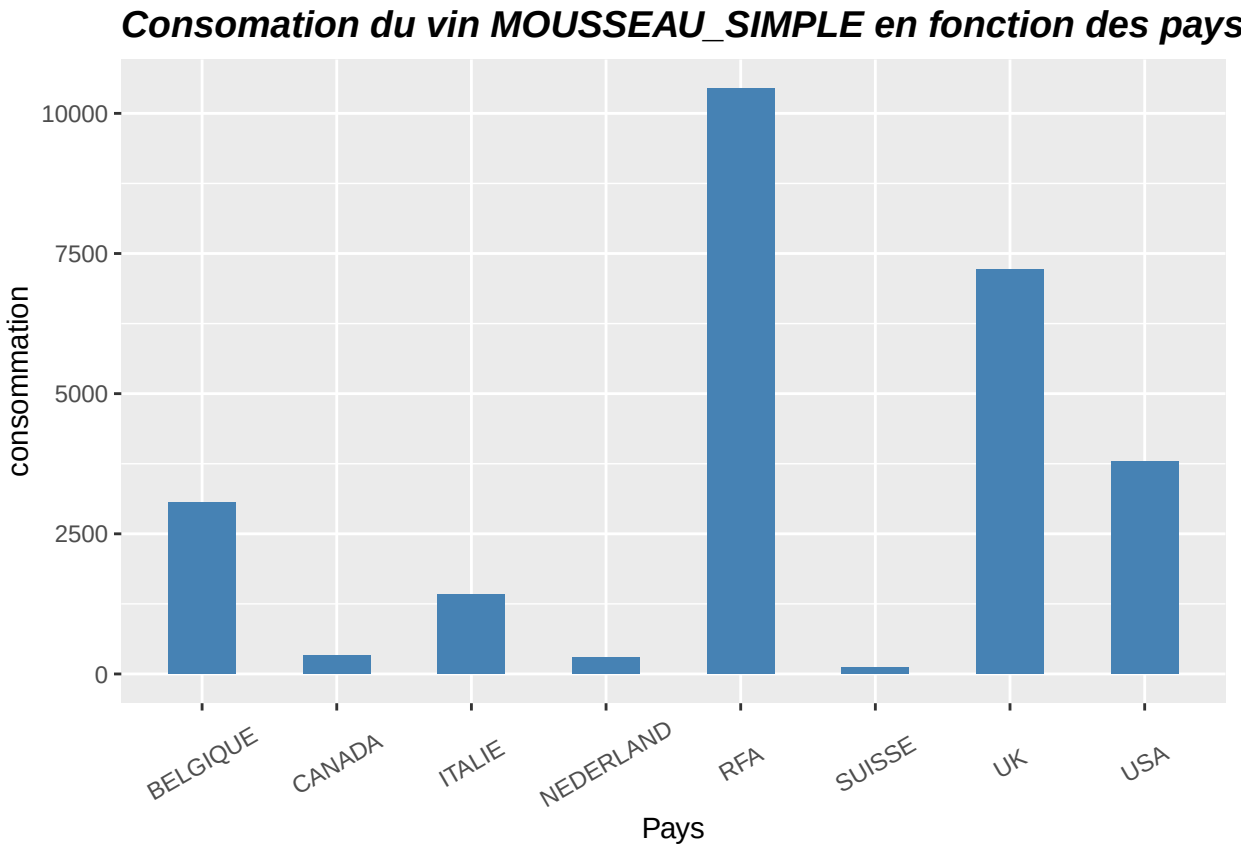
Le premier graphique ci-dessous nous donne la quantité de champagne consommée dans les huit pays de notre jeu de données. On constate que ce vin est consommé principalement au Royaume Uni et en Allemagne. Au Canada, ce vin n'est que très peu consommé.

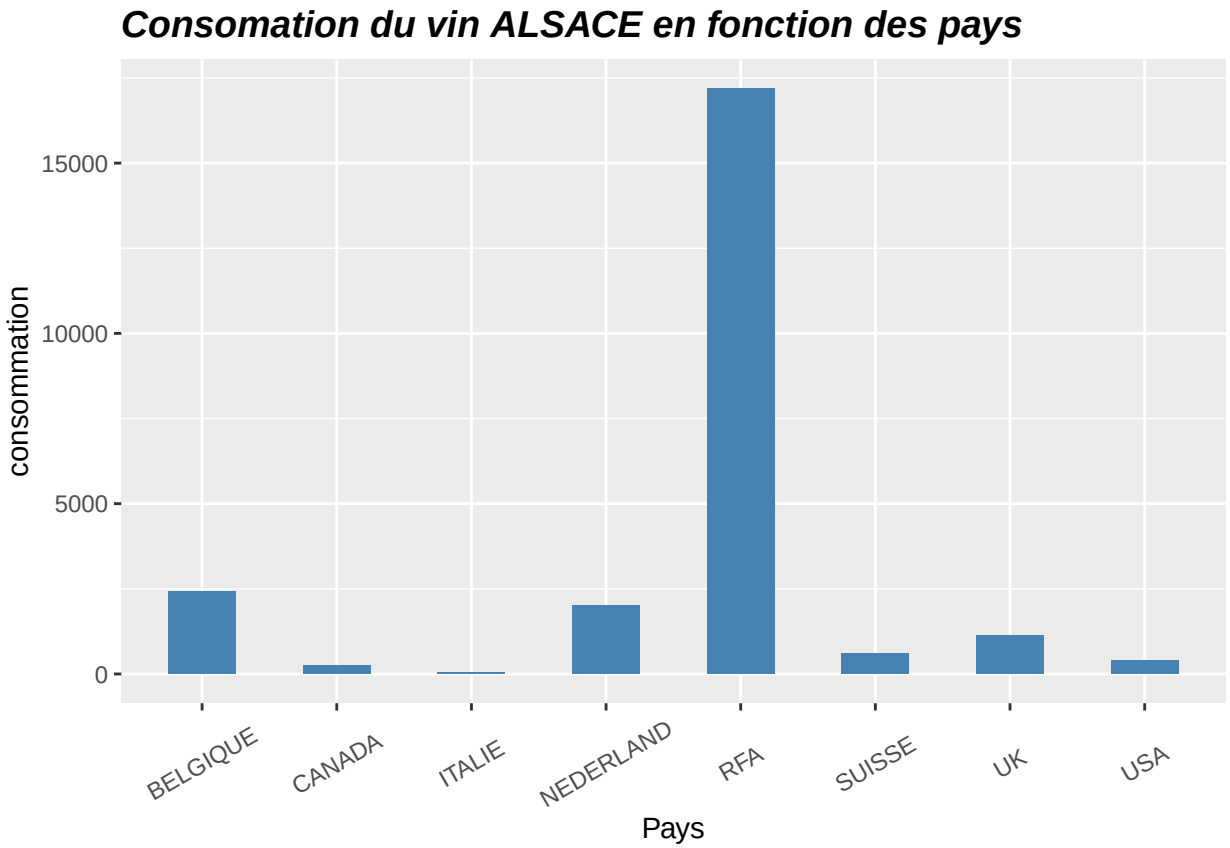


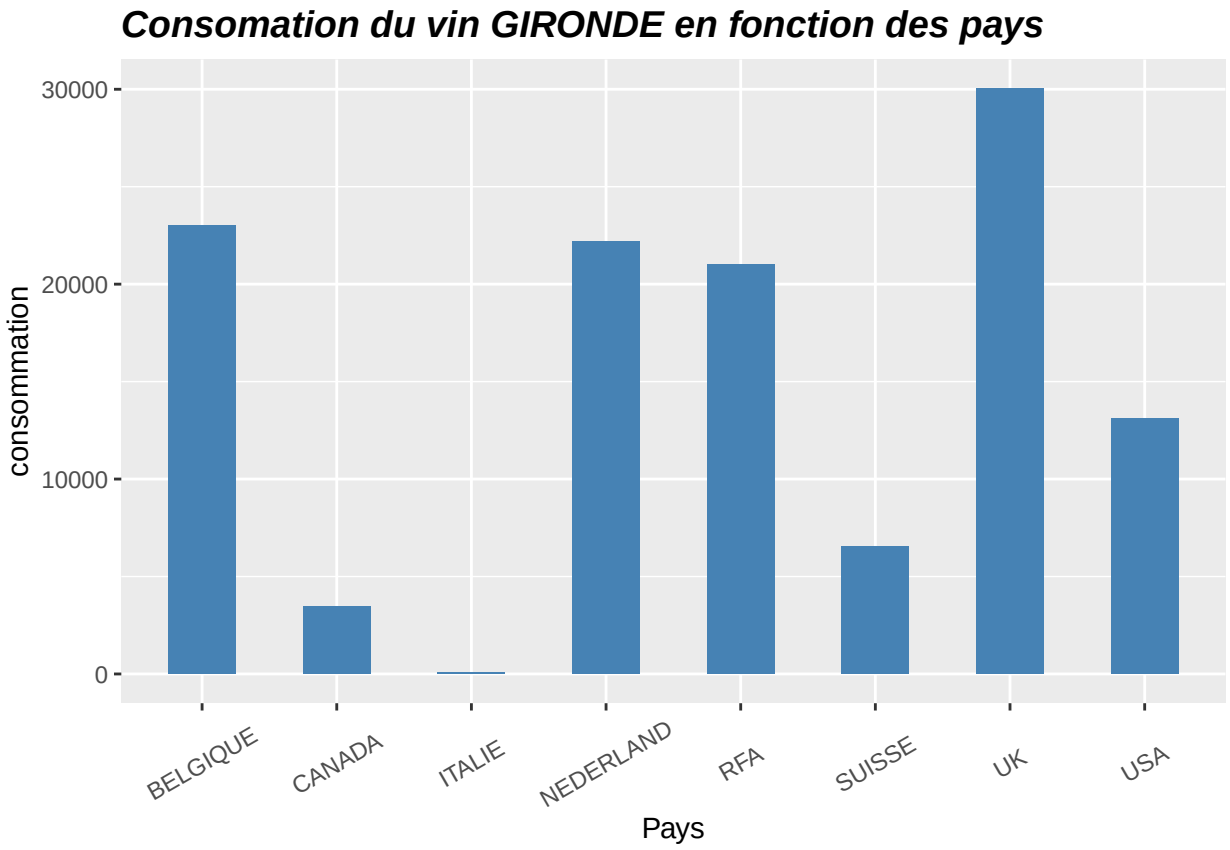
Les graphiques ci-dessous constituent la consommation des vins "MOUSSEAU_AOC", "MOUSSEAU_SIMPLE", "ALSACE", "GIRONDE", "BOJOLAIS", "BORDEAUX", "RHON" et "ANJOU" (dans cet ordre) en fonction des pays. On constate que les vins MOUSSEAU_AOC sont très consommés en Belgique et en Allemagne et peu consommés en Italie et au Canada (premier graphique).
Quand au MOUSSEAU_SIMPLE, il est largement consommé en Allemagne et très peu en Suisse, au Canada et au Nederland (deuxième graphique).
40% du vin ALSACE est consommé en l'Allemagne (troisième graphique).
Le vin GIRONDE est consommé en grande quantité dans tous les pays, sauf en Italie et au Canada où les quantités restent relativement faibles (quatrième graphique). Il en est de même pour le BOJOLAIS et le BORDEAUX (cinquième graphique).
Sur le dernier graphique, donnant la quantité du vin Anjou consommée, on constate que plus de la moitié du vin Anjou est consommée au Royaume Uni.

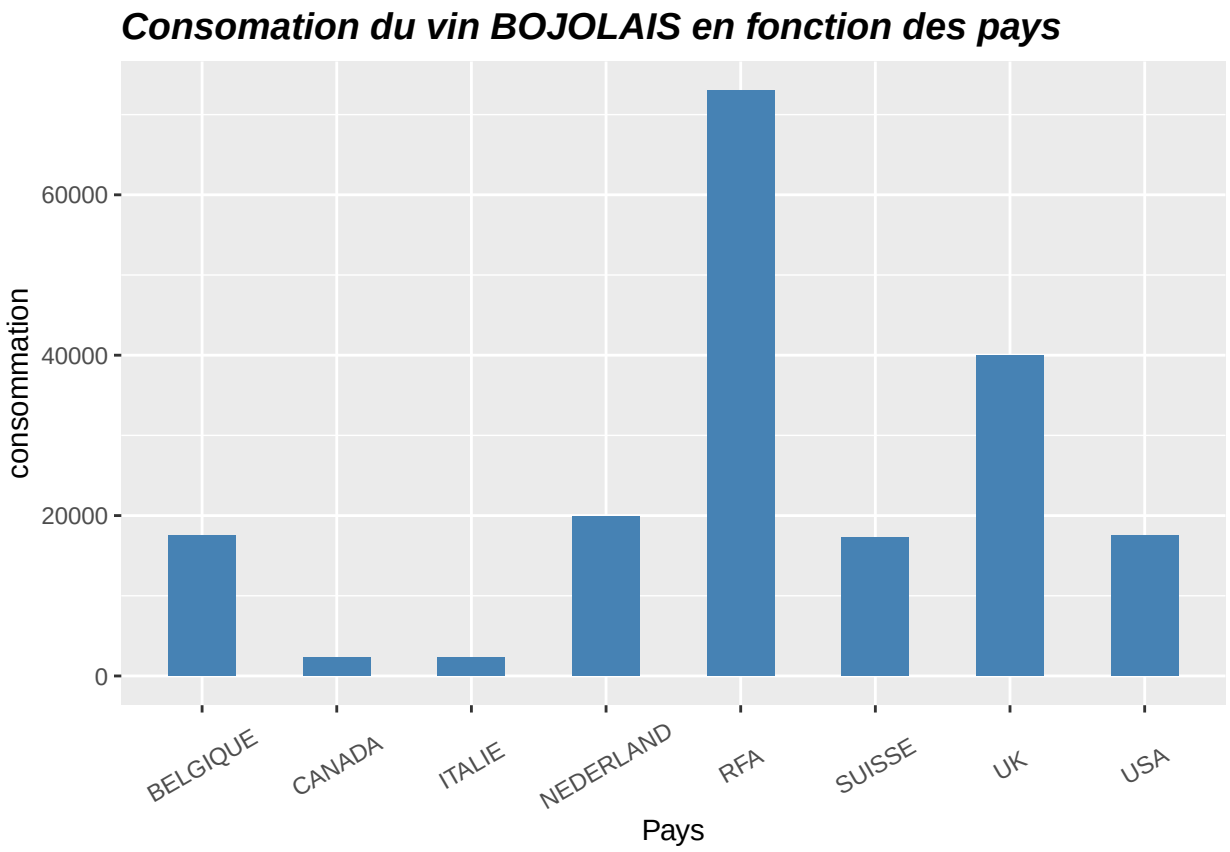
```
library(ggplot2)
library(ggthemes)
library(pander)
Pays = rownames(vins)
plot_list = list()
for(i in 2:9){
  df = data.frame(Pays,consommation = vins[,i])

  p<-ggplot(data=df, aes(x=Pays, y=consommation)) +
    geom_bar(stat="identity", fill="steelblue",width=0.5, size=0.1)+
    theme(axis.text.x = element_text(angle=30, vjust=0.5))+
    ggtitle(paste("Consomation du vin",vins.names[i],"en fonction des pays")) +
    theme(
    plot.title = element_text(color="black",
                          size=14, face="bold.italic"))
  plot_list[[i]] = p
  par(mfrow=c(4,4))
  print(plot_list[[i]])
}
```

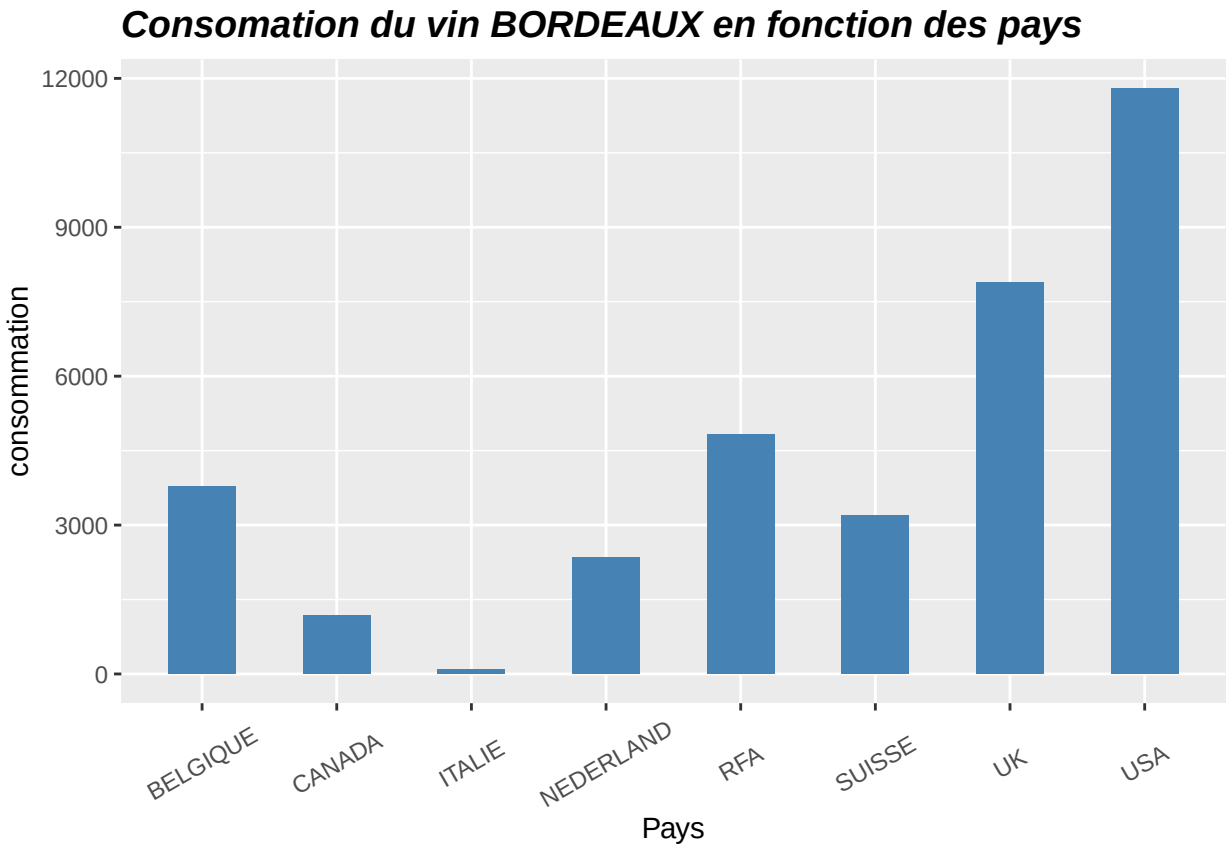


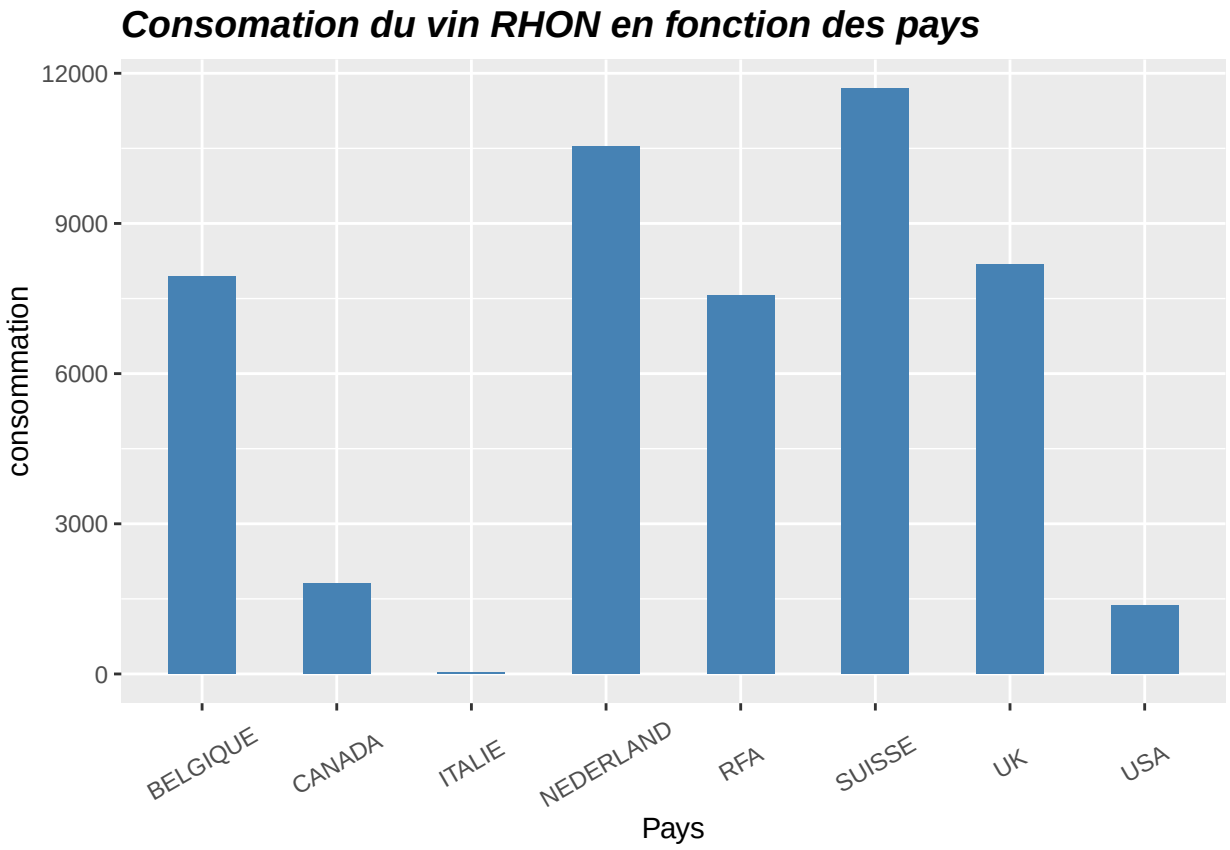

Consomation du vin MOUSSEAU_SIMPLE en fonction des pays
consommation
10000
7500
5000
2500
0
BELGIQUE
CANADA
ITALIE
NEDERLAND
RFA
SUISSE
UK
USA
Pays

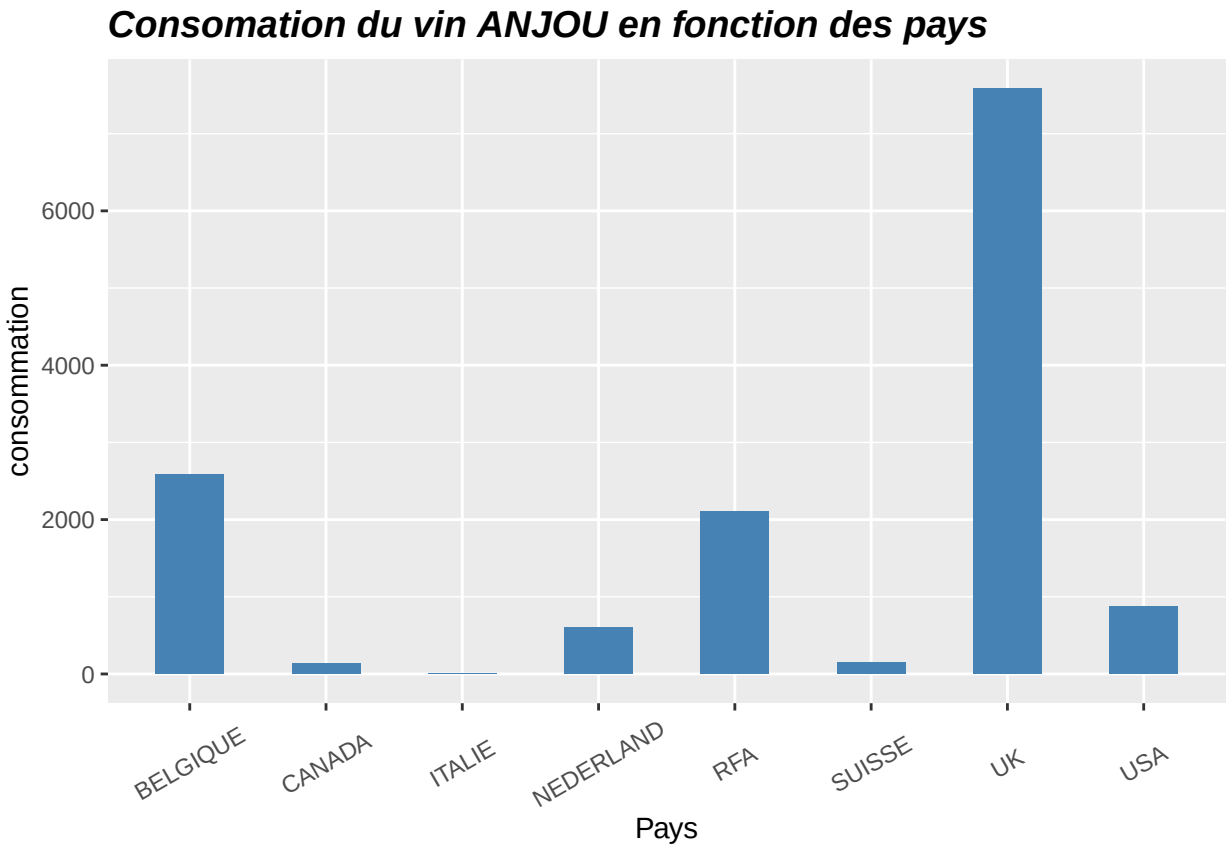

Consomation du vin ALSACE en fonction des pays
consommation
15000
10000
5000
0
BELGIQUE
CANADA
ITALIE
NEDERLAND
RFA
SUISSE
UK
USA
Pays


Consomation du vin GIRONDE en fonction des pays
30000
20000
10000
0
consommation
BELGIQUE
CANADA
ITALIE
NEDERLAND
RFA
SUISSE
UK
USA
Pays


Consomation du vin BOJOLAIS en fonction des pays
consommation
60000
40000
20000
0
BELGIQUE
CANADA
ITALIE
NEDERLAND
RFA
SUISSE
UK
USA
Pays


Consomation du vin BORDEAUX en fonction des pays
consommation
12000
9000
6000
3000
0
BELGIQUE
CANADA
ITALIE
NEDERLAND
RFA
SUISSE
UK
USA
Pays


Consomation du vin RHON en fonction des pays
12000
9000
6000
3000
0
consommation
BELGIQUE
CANADA
ITALIE
NEDERLAND
RFA
SUISSE
UK
USA
Pays


Consomation du vin ANJOU en fonction des pays
consommation
6000
4000
2000
0
BELGIQUE
CANADA
ITALIE
NEDERLAND
RFA
SUISSE
UK
USA
Pays

3 Analyse bivariée des données

3.1 Corrélation

Pour évaluer l'inter-dépendence de plusieurs variables simultanément, nous utilisons la fonction **rcorr de R**. Celle-ci permet en effet d'évaluer en une seule étape les coefficients de corrélation et les niveaux de significativité(p.value).

```
cor = rcorr(as.matrix(vins))
cor$r
```

```
                 CHMPAGNE MOUSSEAU_AOC MOUSSEAU_SIMPLE       ALSACE
CHMPAGNE         1.0000000   0.46333448      0.73693611  0.389588061
MOUSSEAU_AOC     0.4633345   1.00000000      0.66317213  0.566819069
MOUSSEAU_SIMPLE  0.7369361   0.66317213      1.00000000  0.772024929
ALSACE           0.3895881   0.56681907      0.77202493  1.000000000
GIRONDE          0.4297570   0.72155885      0.59209057  0.336165005
BOJOLAIS         0.6622271   0.63381413      0.90707806  0.885366296
BORDEAUX         0.5964842   0.38306440      0.48163282  0.046749067
RHON             0.2087084   0.35689399      0.07534305  0.220650004
ANJOU            0.5776359   0.47226452      0.62155756  0.101664537
AOC_AUTRES       0.2113951   0.61272310      0.44582687  0.359530757
VDQS             0.4052077   0.78667329      0.61661088  0.250372468
AUTRE_VDQS       0.5107641   0.58157262      0.93440358  0.882192931
PROVENCE         0.1772463   0.70588763      0.57072794  0.887007138
MUSCAT           0.4323210   0.22511476      0.44421366 -0.059543370
RHONE            0.1494274   0.42660037      0.21795405  0.357850733
AOC_FORT         0.4634902   0.22220046      0.44600753 -0.046020306
AUTRE_FORT      -0.2055172  -0.08460848     -0.11895191 -0.008684208
TRES_FORT        0.2330486   0.70038817      0.28462666 -0.052441685
                  GIRONDE  BOJOLAIS    BORDEAUX        RHON      ANJOU
CHMPAGNE        0.4297570 0.6622271  0.59648421  0.20870836 0.57763590
MOUSSEAU_AOC    0.7215589 0.6338141  0.38306440  0.35689399 0.47226452
MOUSSEAU_SIMPLE 0.5920906 0.9070781  0.48163282  0.07534305 0.62155756
ALSACE          0.3361650 0.8853663  0.04674907  0.22065000 0.10166454
GIRONDE         1.0000000 0.6293128  0.46285626  0.56349725 0.77529041
BOJOLAIS        0.6293128 1.0000000  0.34298013  0.40552195 0.48361886
BORDEAUX        0.4628563 0.3429801  1.00000000 -0.03541612 0.44967967
RHON            0.5634972 0.4055219 -0.03541612  1.00000000 0.27554767
ANJOU           0.7752904 0.4836189  0.44967967  0.27554767 1.00000000
AOC_AUTRES      0.9341941 0.5551382  0.17504577  0.59736395 0.60425378
VDQS            0.9337615 0.5264408  0.50773579  0.35248346 0.84404538
AUTRE_VDQS      0.5273186 0.9295528  0.24116457  0.14910778 0.49898944
PROVENCE        0.4999792 0.7601589 -0.09639982  0.47078910 0.08472363
MUSCAT          0.7215451 0.3493358  0.36540002  0.29035481 0.94968386
RHONE           0.7364102 0.4984682  0.17060762  0.75803275 0.21787990
AOC_FORT        0.7136148 0.3642399  0.33584292  0.32518562 0.94824830
AUTRE_FORT      0.4855522 0.1106056 -0.09846807  0.47972383 0.11647195
TRES_FORT       0.6494800 0.1374827  0.17871143  0.27553848 0.70199031
                AOC_AUTRES      VDQS  AUTRE_VDQS    PROVENCE      MUSCAT
CHMPAGNE         0.2113951 0.4052077  0.51076415  0.17724635  0.43232095
MOUSSEAU_AOC     0.6127231 0.7866733  0.58157262  0.70588763  0.22511476
MOUSSEAU_SIMPLE  0.4458269 0.6166109  0.93440358  0.57072794  0.44421366
ALSACE           0.3595308 0.2503725  0.88219293  0.88700714 -0.05954337
```

```
GIRONDE           0.9341941 0.9337615  0.52731860  0.49997922  0.72154513
BOJOLAIS          0.5551382 0.5264408  0.92955282  0.76015886  0.34933581
BORDEAUX          0.1750458 0.5077358  0.24116457 -0.09639982  0.36540002
RHON              0.5973640 0.3524835  0.14910778  0.47078910  0.29035481
ANJOU             0.6042538 0.8440454  0.49898944  0.08472363  0.94968386
AOC_AUTRES        1.0000000 0.8024936  0.45925050  0.59672500  0.60409970
VDQS              0.8024936 1.0000000  0.51976776  0.38978757  0.73120164
AUTRE_VDQS        0.4592505 0.5197678  1.00000000  0.71144609  0.35365415
PROVENCE          0.5967250 0.3897876  0.71144609  1.00000000 -0.06393624
MUSCAT            0.6040997 0.7312016  0.35365415 -0.06393624  1.00000000
RHONE             0.8387025 0.4689564  0.25907759  0.59854704  0.26680490
AOC_FORT          0.6015040 0.7117073  0.35467994 -0.05147062  0.99719789
AUTRE_FORT        0.6943362 0.2118014 -0.02934817  0.21184280  0.32724273
TRES_FORT         0.5440734 0.8178777  0.16911553  0.18412453  0.55655891
                     RHONE    AOC_FORT   AUTRE_FORT   TRES_FORT
CHMPAGNE          0.1494274  0.46349017 -0.205517193  0.23304855
MOUSSEAU_AOC      0.4266004  0.22220046 -0.084608482  0.70038817
MOUSSEAU_SIMPLE   0.2179540  0.44600753 -0.118951907  0.28462666
ALSACE            0.3578507 -0.04602031 -0.008684208 -0.05244168
GIRONDE           0.7364102  0.71361478  0.485552226  0.64948002
BOJOLAIS          0.4984682  0.36423986  0.110605638  0.13748267
BORDEAUX          0.1706076  0.33584292 -0.098468067  0.17871143
RHON              0.7580328  0.32518562  0.479723831  0.27553848
ANJOU             0.2178799  0.94824830  0.116471948  0.70199031
AOC_AUTRES        0.8387025  0.60150398  0.694336157  0.54407338
VDQS              0.4689564  0.71170726  0.211801359  0.81787772
AUTRE_VDQS        0.2590776  0.35467994 -0.029348171  0.16911553
PROVENCE          0.5985470 -0.05147062  0.211842801  0.18412453
MUSCAT            0.2668049  0.99719789  0.327242732  0.55655891
RHONE             1.0000000  0.27440019  0.800129513  0.14660800
AOC_FORT          0.2744002  1.00000000  0.325800326  0.55170047
AUTRE_FORT        0.8001295  0.32580033  1.000000000 -0.06048420
TRES_FORT         0.1466080  0.55170047 -0.060484200  1.00000000
```

```
#cor$r[which(cor$r <0, arr.ind = T)]
```

Le tableau ci-dessus est la matrice des corrélations. Elle donne les coefficients de corrélation linéaire des variables prises deux à deux. Il s'agit d'une succession d'analyses bivariées, constituant un premier pas vers l'analyse multivariée.
Une analyse du tableau ci dessus nous permet de constater que la plus part des corrélations sont positives et certaines sont très fortes (entre les variables BOJOLAIS et MOUSSEAU_SIMPLE, AUTRE_VDQS et MOUSSEAU_SIMPLE, GIRONDE et AOC_AUTRES, ANJOU et MUSCAT, ANJOU et AOC_FORT, AUTRE_FORT et RHONE etc) car la valeur du coefficient de corrélation dépasse les 0.8. D'autres sont moyennes (entre GIRONDE et PROVENCE etc) voire faibles (entre les variables PROVENCE et MUSCAT etc).
La formule ci-dessous permet de mieux visualiser :

```
symnum(cor$r, abbr.colnames=FALSE)
```

```
                CHMPAGNE MOUSSEAU_AOC MOUSSEAU_SIMPLE ALSACE GIRONDE
CHMPAGNE        1
MOUSSEAU_AOC    .        1
MOUSSEAU_SIMPLE ,        ,            1
ALSACE          .        .            ,               1
```

```
GIRONDE          .         ,             .                .       1
BOJOLAIS         ,         ,             *                +       ,
BORDEAUX         .         .             .                        .
RHON                       .                                      .
ANJOU            .         .             ,                        ,
AOC_AUTRES                 ,             .                .       *
VDQS             .         ,             ,                        *
AUTRE_VDQS       .         .             *                +       .
PROVENCE                   ,             .                +       .
MUSCAT           .                       .                        ,
RHONE                      .                              .       ,
AOC_FORT         .                       .                        ,
AUTRE_FORT                                                        .
TRES_FORT                  ,                                      ,
                 BOJOLAIS BORDEAUX RHON ANJOU AOC_AUTRES VDQS AUTRE_VDQS
CHMPAGNE
MOUSSEAU_AOC
MOUSSEAU_SIMPLE
ALSACE
GIRONDE
BOJOLAIS         1
BORDEAUX         .         1
RHON             .                   1
ANJOU            .         .               1
AOC_AUTRES       .                   .     ,      1
VDQS             .         .         .     +      +           1
AUTRE_VDQS       *                         .      .           .     1
PROVENCE         ,                   .           .           .     ,
MUSCAT           .         .               *      ,           ,     .
RHONE            .                   ,           +           .
AOC_FORT         .         .         .     *      ,           ,     .
AUTRE_FORT                           .           ,
TRES_FORT                                  ,     .           +
                 PROVENCE MUSCAT RHONE AOC_FORT AUTRE_FORT TRES_FORT
CHMPAGNE
MOUSSEAU_AOC
MOUSSEAU_SIMPLE
ALSACE
GIRONDE
BOJOLAIS
BORDEAUX
RHON
ANJOU
AOC_AUTRES
VDQS
AUTRE_VDQS
PROVENCE         1
MUSCAT                     1
RHONE            .                 1
AOC_FORT                   B               1
AUTRE_FORT                 .       +       .          1
TRES_FORT                  .               .                    1
attr(,"legend")
[1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

Comme indiqué dans la légende, les coefficients de corrélation entre 0 et 0.3 sont remplacés par un espace (" "); les coefficients de corrélation entre 0.3 et 0.6 sont remplacés par"."; etc.

Effectuons maintenant des tests statistiques pour evaluer le niveau de confiance aux valeurs des coefficients de corrélation présentés précedemment.

```
cor$P
```

```
                CHMPAGNE MOUSSEAU_AOC MOUSSEAU_SIMPLE      ALSACE
CHMPAGNE              NA   0.24757535    0.0370048239 0.340070847
MOUSSEAU_AOC   0.24757535           NA    0.0730269335 0.142910662
MOUSSEAU_SIMPLE 0.03700482  0.07302693              NA 0.024787420
ALSACE         0.34007085   0.14291066    0.0247874203          NA
GIRONDE        0.28792368   0.04332582    0.1220045722 0.415566946
BOJOLAIS       0.07358424   0.09151188    0.0018686419 0.003449614
BORDEAUX       0.11855793   0.34892397    0.2268752459 0.912473127
RHON           0.61988729   0.38547589    0.8592654795 0.599513433
ANJOU          0.13373669   0.23735806    0.0999517534 0.810688386
AOC_AUTRES     0.61528432   0.10630153    0.2682363481 0.381719333
VDQS           0.31930446   0.02055287    0.1034796270 0.549801353
AUTRE_VDQS     0.19584175   0.13048170    0.0006713756 0.003734825
PROVENCE       0.67455802   0.05039871    0.1395565052 0.003307829
MUSCAT         0.28473674   0.59195311    0.2701816861 0.888619783
RHONE          0.72396632   0.29187112    0.6040938211 0.384110957
AOC_FORT       0.24739535   0.59688440    0.2680189303 0.913833681
AUTRE_FORT     0.62536839   0.84211457    0.7790601391 0.983717929
TRES_FORT      0.57859773   0.05303465    0.4944473468 0.901851969
                   GIRONDE     BOJOLAIS  BORDEAUX       RHON
CHMPAGNE      0.2879236841 0.0735842358 0.1185579 0.61988729
MOUSSEAU_AOC  0.0433258239 0.0915118834 0.3489240 0.38547589
MOUSSEAU_SIMPLE 0.1220045722 0.0018686419 0.2268752 0.85926548
ALSACE        0.4155669462 0.0034496140 0.9124731 0.59951343
GIRONDE                NA 0.0945616886 0.2481287 0.14579573
BOJOLAIS      0.0945616886           NA 0.4055657 0.31889299
BORDEAUX      0.2481286581 0.4055656798        NA 0.93365028
RHON          0.1457957324 0.3188929924 0.9336503         NA
ANJOU         0.0238006326 0.2246841528 0.2636186 0.50890416
AOC_AUTRES    0.0006777193 0.1531965595 0.6784320 0.11787449
VDQS          0.0006909445 0.1801328873 0.1989567 0.39179577
AUTRE_VDQS    0.1792740012 0.0008285096 0.5650433 0.72453919
PROVENCE      0.2070531629 0.0285846823 0.8203670 0.23903108
MUSCAT        0.0433317377 0.3963336254 0.3734162 0.48540913
RHONE         0.0372110575 0.2086501304 0.6862638 0.02930059
AOC_FORT      0.0468306187 0.3750510148 0.4160422 0.43189708
AUTRE_FORT    0.2225616913 0.7942996037 0.8165623 0.22899165
TRES_FORT     0.0813458889 0.7454498665 0.6719823 0.50891887
                     ANJOU   AOC_AUTRES         VDQS   AUTRE_VDQS
CHMPAGNE      0.1337366914 0.6152843227 0.3193044564 0.1958417509
MOUSSEAU_AOC  0.2373580622 0.1063015268 0.0205528745 0.1304817011
MOUSSEAU_SIMPLE 0.0999517534 0.2682363481 0.1034796270 0.0006713756
ALSACE        0.8106883859 0.3817193332 0.5498013531 0.0037348251
GIRONDE       0.0238006326 0.0006777193 0.0006909445 0.1792740012
BOJOLAIS      0.2246841528 0.1531965595 0.1801328873 0.0008285096
```

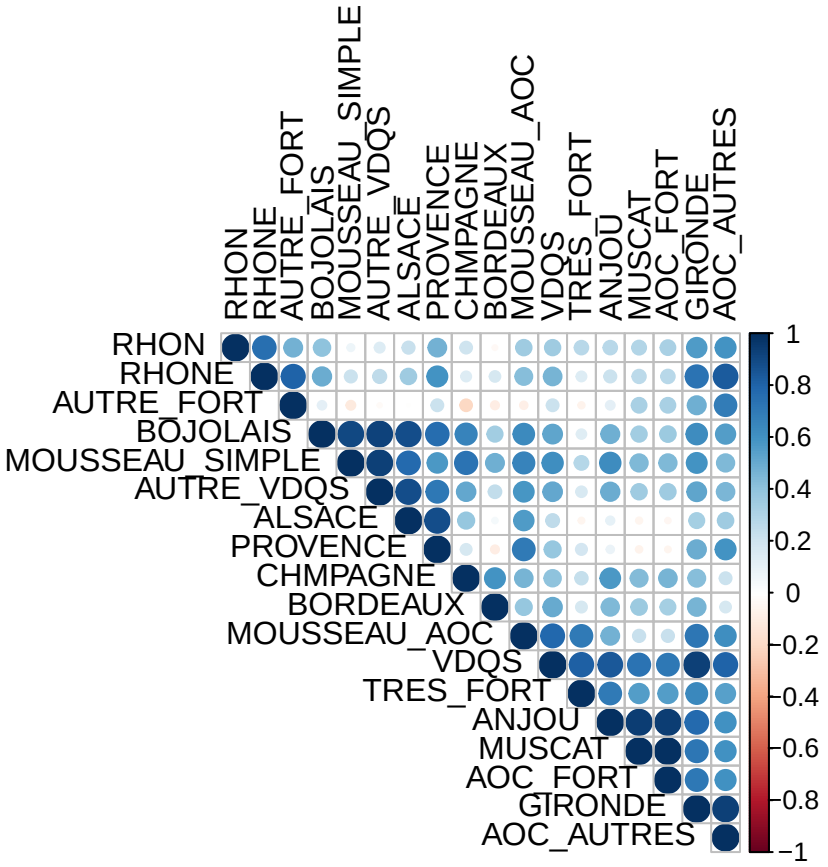
```
BORDEAUX        0.2636185857 0.6784320208 0.1989567056 0.5650432902
RHON            0.5089041563 0.1178744938 0.3917957705 0.7245391928
ANJOU                     NA 0.1125991464 0.0084081946 0.2080985156
AOC_AUTRES      0.1125991464           NA 0.0165207646 0.2523206543
VDQS            0.0084081946 0.0165207646           NA 0.1867342258
AUTRE_VDQS      0.2080985156 0.2523206543 0.1867342258           NA
PROVENCE        0.8419017585 0.1183706502 0.3398017512 0.0478161233
MUSCAT          0.0003065682 0.1127156263 0.0392913306 0.3901138785
RHONE           0.6042199527 0.0092629347 0.2411175575 0.5355288103
AOC_FORT        0.0003331983 0.1146881977 0.0476967891 0.3886427647
AUTRE_FORT      0.7835820906 0.0560289868 0.6145893275 0.9450037681
TRES_FORT       0.0522584013 0.1633023495 0.0131141705 0.6889024054
                   PROVENCE       MUSCAT       RHONE     AOC_FORT
CHMPAGNE        0.674558022 2.847367e-01 0.723966317 2.473954e-01
MOUSSEAU_AOC    0.050398713 5.919531e-01 0.291871116 5.968844e-01
MOUSSEAU_SIMPLE 0.139556505 2.701817e-01 0.604093821 2.680189e-01
ALSACE          0.003307829 8.886198e-01 0.384110957 9.138337e-01
GIRONDE         0.207053163 4.333174e-02 0.037211057 4.683062e-02
BOJOLAIS        0.028584682 3.963336e-01 0.208650130 3.750510e-01
BORDEAUX        0.820367020 3.734162e-01 0.686263844 4.160422e-01
RHON            0.239031083 4.854091e-01 0.029300589 4.318971e-01
ANJOU           0.841901759 3.065682e-04 0.604219953 3.331983e-04
AOC_AUTRES      0.118370650 1.127156e-01 0.009262935 1.146882e-01
VDQS            0.339801751 3.929133e-02 0.241117558 4.769679e-02
AUTRE_VDQS      0.047816123 3.901139e-01 0.535528810 3.886428e-01
PROVENCE                 NA 8.804458e-01 0.116958914 9.036629e-01
MUSCAT          0.880445846           NA 0.522974408 5.488876e-08
RHONE           0.116958914 5.229744e-01          NA 5.107426e-01
AOC_FORT        0.903662901 5.488876e-08 0.510742621           NA
AUTRE_FORT      0.614518440 4.288172e-01 0.017088546 4.309758e-01
TRES_FORT       0.662489853 1.519245e-01 0.729023565 1.562986e-01
                AUTRE_FORT  TRES_FORT
CHMPAGNE        0.62536839 0.57859773
MOUSSEAU_AOC    0.84211457 0.05303465
MOUSSEAU_SIMPLE 0.77906014 0.49444735
ALSACE          0.98371793 0.90185197
GIRONDE         0.22256169 0.08134589
BOJOLAIS        0.79429960 0.74544987
BORDEAUX        0.81656233 0.67198227
RHON            0.22899165 0.50891887
ANJOU           0.78358209 0.05225840
AOC_AUTRES      0.05602899 0.16330235
VDQS            0.61458933 0.01311417
AUTRE_VDQS      0.94500377 0.68890241
PROVENCE        0.61451844 0.66248985
MUSCAT          0.42881723 0.15192447
RHONE           0.01708855 0.72902357
AOC_FORT        0.43097579 0.15629858
AUTRE_FORT              NA 0.88686841
TRES_FORT       0.88686841         NA
```

Pour tester la significativité linéaire entre deux variables de notre échantillon (code ci-dessus), nous testons : h_0 : "il n'y a pas de corrélation linéaire entre les 2 variables" contre h_1 : "il existe un lien statistique et linéaire entre les 2 variables". Il s'agit d'un test paramétrique bilatéral. Au risque de prèmier espèce $\alpha =$

5%, on rejette h_0 si la p-value est inférieur à α et on favorise h_1 : il existe un lien linéaire entre les deux variables.
Par exemple, la p-value entre les variables "MOUSSEAU_SIMPLE" et "BOJOLAIS" est inférieure au seuil, on peut donc dire que les pays consommateurs du vin mousseau simple ont également tendance a consommé le vin bojolais(correlation positive). La corrélation entre ces deux vins est d'ailleurs très forte. A l'inverse, il n'existe aucun lien statistique entre les vins bordeaux et alsace d'après l'étude de la p-value (la corrélation est d'ailleurs très faibles entre les deux variables).

3.2 Graphique des corrélations

```
library(corrplot)
corrplot(cor$r, type="upper", order="hclust", tl.col="black")
```



Les corrélations positives sont affichées en bleu et les corrélations négatives en rouge. L'intensité de la couleur et la taille des cercles sont proportionnelles aux coefficients de corrélation. A droite du corrélogramme, la légende de couleurs montre les coefficients de corrélation et les couleurs correspondantes.

4 Analyse en Composante Principale (ACP)

Principe : L'idée à la base de l'analyse en composantes principales est de pouvoir expliquer ou rendre compte de la variance observée dans la masse de données initiales en se limitant à un nombre réduit de composantes, définies comme étant des transformations mathématiques des variables initiales. La composante (linéiare) s'écrit : $c = \mu_1 \times X_1 + \mu_2 \times X_2 + \cdots + \mu_p \times X_p$ avec $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$ des pods définissant la composante c. De plus il faut noter que ces poids sont normalisés, c'est à dire que : $\sum_{i=1}^{p} \mu_i^2 = 1$. Cette normalisation s'explique par le critère utilisé pour determiner c qui est celui de la variance maximale.

Elle permet donc de synthétiser l'information(variance) contenue dans un tableau de données (n colonnes × p lignes), d'identifier une éventuelle similarité entre les individus et de determiner la liaison entre les variables.
Si tous nos coefficients de corrélation étaient aussi faibles que ceux entre les variables "TRES_FORT" et "BOJOLAIS" ou entre "RHON" et "MOUSSEAU_SIMPLE" par exemple, il n'y aurait aucun intérêt à procéder à une analyse en composantes principales de ces données. En effet, pour pouvoir extraire une composante correspondant à une fonction linéaire des variables initiales, il faut nécessairement que ces variables soient intercorrélées plus ou moins fortement.
Nous faisons donc l'ACP normée (scale.unit=TRUE) sur les 18 variables de notre jeu de données avec la fonction **PCA**.

```
library(FactoMineR)
data.pca = PCA(vins, scale.unit=TRUE, ncp =18 , graph = FALSE)
```

L'objet créé avec la fonction **PCA** contient de nombreuses informations stockées dans de nombreuses listes et matrices différentes. Ces valeurs sont décrites dans les sections suivants.

4.1 Nombre de Composantes principales à retenir

Les valeurs propres mesurent la quantité de variance expliquée par chaque axe principal. Elles sont grandes pour les premiers axes et petites pour les axes suivants. Autrement dit, les premiers axes correspondent aux directions portant la quantité maximale de la variation du jeu de données.
Notre objectif ici est d'examiner les valeurs propres afin de déterminer le nombre de composantes principales à prendre en considération. Pour ce faire, nous allons utiliser deux méthodes.

Voici un aperçu de nos valeurs propres

```
knitr::kable(data.pca$eig)
```

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	8.8880543	49.378080	49.37808
comp 2	3.1963075	17.757264	67.13534
comp 3	2.7276039	15.153355	82.28870
comp 4	1.3266750	7.370416	89.65911
comp 5	0.9700696	5.389276	95.04839
comp 6	0.6845571	3.803095	98.85149
comp 7	0.2067326	1.148515	100.00000

4.1.1 Critère de la moyenne

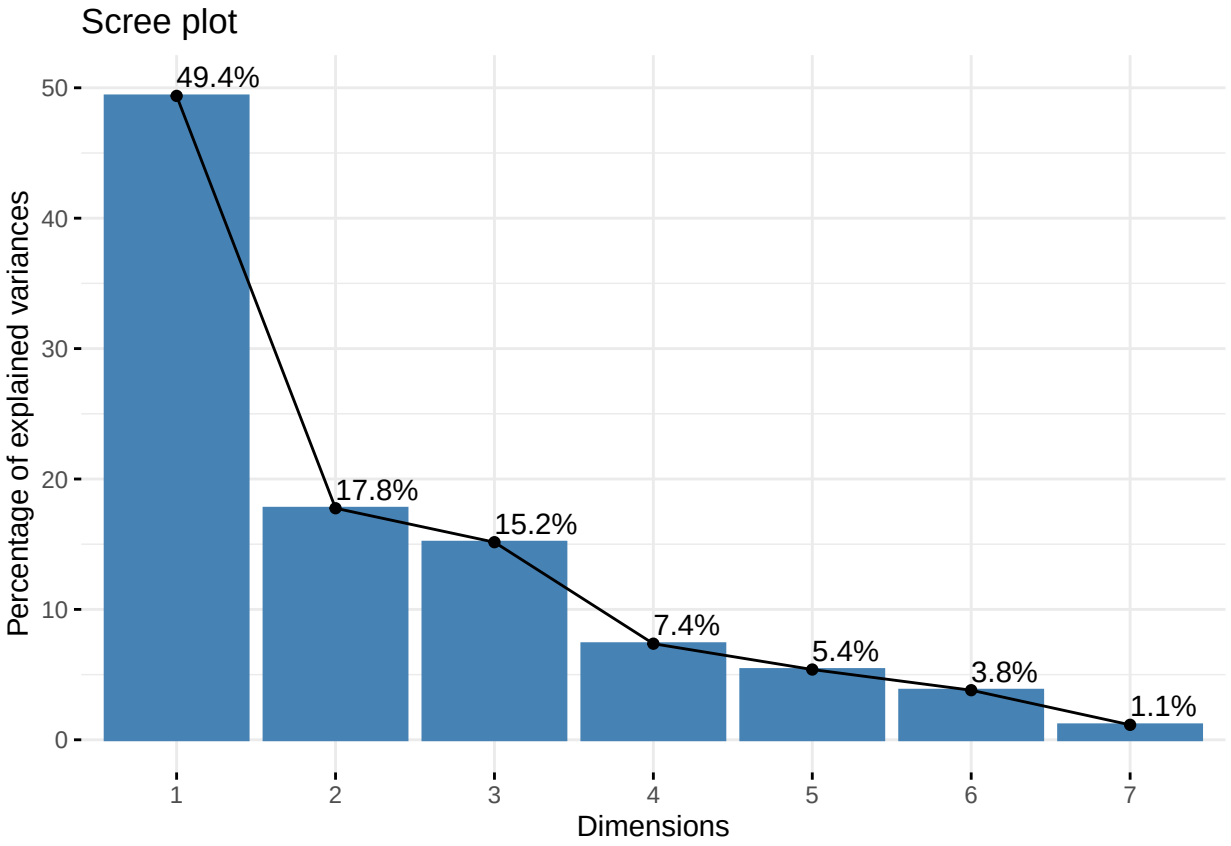
```
which(data.pca$eig[,1]>1)
```

```
comp 1 comp 2 comp 3 comp 4
     1      2      3      4
```

Si on considère le critère de l'inertie moyenne (valeur propre > 1), qui indique que la composante principale concernée représente plus de variance par rapport à une seule variable d'origine (lorsque les données sont standardisées), on constate qu'il faudrait garder les quatre premières composantes principales. Ces quatre composantes expliquent près de 90% de la variance.

4.1.2 Graphique des valeurs propres

```
library(factoextra)
fviz_eig(data.pca, addlabels = TRUE, ylim = c(0, 50))
```



Une autre méthode pour déterminer le nombre de composantes principales est de regarder le graphique des valeurs propres ci-dessus. Le nombre d'axes est déterminé par le point, au-delà duquel les valeurs propres restantes sont toutes relativement petites et de tailles comparables.
Au vu du graphique, on peut garder les trois premières composantes qui expliquent 82,3% de l'information contenue dans la variance.
Le but etant de reduire au maximum la dimension de notre jeu de données tout en gardant la variance maximale, **on peut donc raisonnablement retenir trois composantes principales**. Du reste, la dimension 3, c'est à dire l'espace, permet une très bonne visualisation de l'information contenue dans notre jeu de données.

4.2 Interprétation des axes par rapport aux variables

```
data.pca$var
```

```
$coord
                    Dim.1       Dim.2        Dim.3        Dim.4
CHMPAGNE        0.6072688  0.10245549 -0.504142159  0.265917797
MOUSSEAU_AOC    0.7775888  0.24707582 -0.081497120 -0.535805334
MOUSSEAU_SIMPLE 0.8021424  0.37333858 -0.423651049  0.142908304
ALSACE          0.5591280  0.81347935  0.007869155  0.077255318
GIRONDE         0.9448574 -0.21729730  0.182048256 -0.064771333
BOJOLAIS        0.8234525  0.49840344 -0.081965291  0.239634045
BORDEAUX        0.4479770 -0.15484690 -0.451148901  0.170421932
RHON            0.5193233 -0.05182179  0.580974609 -0.002183721
ANJOU           0.8057605 -0.45991898 -0.314633077  0.054335217
AOC_AUTRES      0.8500064 -0.15394690  0.459984631 -0.056935254
VDQS            0.8948636 -0.27998602 -0.077522665 -0.278259755
AUTRE_VDQS      0.7475147  0.50654587 -0.233948949  0.156806641
PROVENCE        0.6127968  0.66764947  0.349117520 -0.212437430
MUSCAT          0.6967702 -0.61237397 -0.156965370  0.244465430
RHONE           0.6271885  0.03771704  0.709931302  0.131611226
AOC_FORT        0.6995089 -0.59854682 -0.147958273  0.257145305
AUTRE_FORT      0.3001127 -0.26691177  0.802856582  0.358435345
TRES_FORT       0.5935437 -0.42362712 -0.134232143 -0.653292485
                      Dim.5         Dim.6        Dim.7
CHMPAGNE         0.29493350  0.3593299576  0.282404234
MOUSSEAU_AOC     0.19535444 -0.0109669605  0.047920796
MOUSSEAU_SIMPLE -0.07383411 -0.0938997632  0.054901486
ALSACE          -0.13811046 -0.0225619014 -0.003661895
GIRONDE          0.10349713 -0.1083538383 -0.015403403
BOJOLAIS        -0.02315895  0.0776020666 -0.053092387
BORDEAUX         0.67696263 -0.2460010138 -0.154806930
RHON             0.14292699  0.5742736100 -0.199657824
ANJOU           -0.18028938  0.0639093486 -0.026259300
AOC_AUTRES      -0.05141208 -0.1581695692  0.106306598
VDQS             0.03147649 -0.1764231465 -0.072613889
AUTRE_VDQS      -0.28315702 -0.1043729476 -0.119335857
PROVENCE        -0.10815453  0.0001140905  0.003763406
MUSCAT          -0.22959349  0.0218981618 -0.043763997
RHONE            0.28661893 -0.0201477877  0.036493212
AOC_FORT        -0.23862909  0.0856187857 -0.011800476
AUTRE_FORT      -0.01443898 -0.2252617778  0.121179752
TRES_FORT       -0.10203222  0.0885633370  0.071989670

$cor
                    Dim.1       Dim.2        Dim.3        Dim.4
CHMPAGNE        0.6072688  0.10245549 -0.504142159  0.265917797
MOUSSEAU_AOC    0.7775888  0.24707582 -0.081497120 -0.535805334
MOUSSEAU_SIMPLE 0.8021424  0.37333858 -0.423651049  0.142908304
ALSACE          0.5591280  0.81347935  0.007869155  0.077255318
GIRONDE         0.9448574 -0.21729730  0.182048256 -0.064771333
BOJOLAIS        0.8234525  0.49840344 -0.081965291  0.239634045
BORDEAUX        0.4479770 -0.15484690 -0.451148901  0.170421932
RHON            0.5193233 -0.05182179  0.580974609 -0.002183721
```

```
ANJOU           0.8057605 -0.45991898 -0.314633077  0.054335217
AOC_AUTRES      0.8500064 -0.15394690  0.459984631 -0.056935254
VDQS            0.8948636 -0.27998602 -0.077522665 -0.278259755
AUTRE_VDQS      0.7475147  0.50654587 -0.233948949  0.156806641
PROVENCE        0.6127968  0.66764947  0.349117520 -0.212437430
MUSCAT          0.6967702 -0.61237397 -0.156965370  0.244465430
RHONE           0.6271885  0.03771704  0.709931302  0.131611226
AOC_FORT        0.6995089 -0.59854682 -0.147958273  0.257145305
AUTRE_FORT      0.3001127 -0.26691177  0.802856582  0.358435345
TRES_FORT       0.5935437 -0.42362712 -0.134232143 -0.653292485
                      Dim.5         Dim.6        Dim.7
CHMPAGNE          0.29493350  0.3593299576  0.282404234
MOUSSEAU_AOC      0.19535444 -0.0109669605  0.047920796
MOUSSEAU_SIMPLE -0.07383411 -0.0938997632  0.054901486
ALSACE          -0.13811046 -0.0225619014 -0.003661895
GIRONDE           0.10349713 -0.1083538383 -0.015403403
BOJOLAIS        -0.02315895  0.0776020666 -0.053092387
BORDEAUX          0.67696263 -0.2460010138 -0.154806930
RHON              0.14292699  0.5742736100 -0.199657824
ANJOU           -0.18028938  0.0639093486 -0.026259300
AOC_AUTRES      -0.05141208 -0.1581695692  0.106306598
VDQS              0.03147649 -0.1764231465 -0.072613889
AUTRE_VDQS      -0.28315702 -0.1043729476 -0.119335857
PROVENCE        -0.10815453  0.0001140905  0.003763406
MUSCAT          -0.22959349  0.0218981618 -0.043763997
RHONE             0.28661893 -0.0201477877  0.036493212
AOC_FORT        -0.23862909  0.0856187857 -0.011800476
AUTRE_FORT      -0.01443898 -0.2252617778  0.121179752
TRES_FORT       -0.10203222  0.0885633370  0.071989670

$cos2
                     Dim.1       Dim.2        Dim.3        Dim.4
CHMPAGNE        0.36877534 0.010497128 0.2541593169 7.071227e-02
MOUSSEAU_AOC    0.60464437 0.061046459 0.0066417805 2.870874e-01
MOUSSEAU_SIMPLE 0.64343249 0.139381697 0.1794802115 2.042278e-02
ALSACE          0.31262409 0.661748658 0.0000619236 5.968384e-03
GIRONDE         0.89275551 0.047218118 0.0331415677 4.195326e-03
BOJOLAIS        0.67807401 0.248405988 0.0067183088 5.742448e-02
BORDEAUX        0.20068338 0.023977561 0.2035353310 2.904364e-02
RHON            0.26969669 0.002685498 0.3375314958 4.768637e-06
ANJOU           0.64925003 0.211525465 0.0989939731 2.952316e-03
AOC_AUTRES      0.72251096 0.023699649 0.2115858608 3.241623e-03
VDQS            0.80078090 0.078392169 0.0060097636 7.742849e-02
AUTRE_VDQS      0.55877819 0.256588719 0.0547321106 2.458832e-02
PROVENCE        0.37551990 0.445755819 0.1218830427 4.512966e-02
MUSCAT          0.48548866 0.375001881 0.0246381274 5.976335e-02
RHONE           0.39336536 0.001422575 0.5040024542 1.732151e-02
AOC_FORT        0.48931267 0.358258300 0.0218916506 6.612371e-02
AUTRE_FORT      0.09006763 0.071241892 0.6445786919 1.284759e-01
TRES_FORT       0.35229417 0.179459940 0.0180182683 4.267911e-01
                      Dim.5        Dim.6        Dim.7
CHMPAGNE        0.0869857695 1.291180e-01 7.975215e-02
MOUSSEAU_AOC    0.0381633580 1.202742e-04 2.296403e-03
MOUSSEAU_SIMPLE 0.0054514764 8.817166e-03 3.014173e-03
```

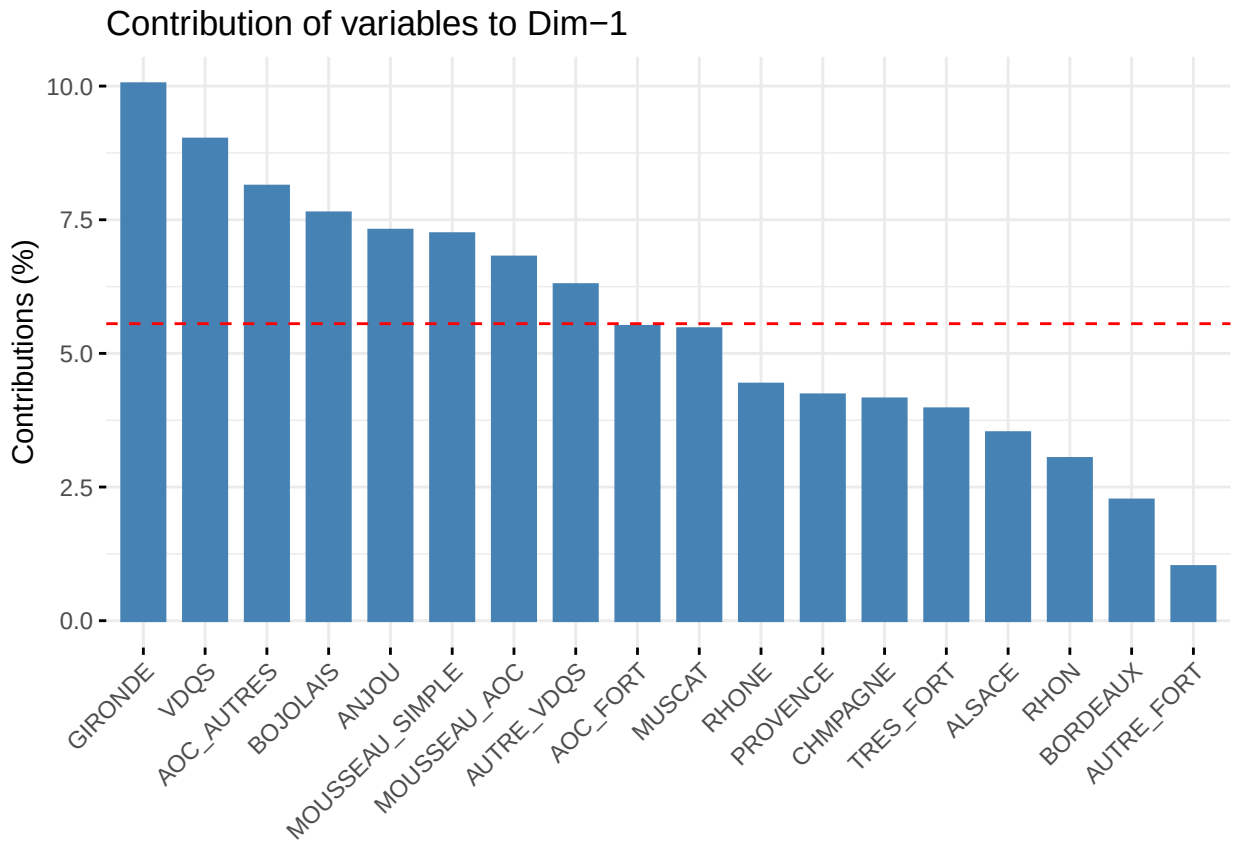
```
ALSACE          0.0190745000 5.090394e-04 1.340947e-05
GIRONDE         0.0107116567 1.174055e-02 2.372648e-04
BOJOLAIS        0.0005363370 6.022081e-03 2.818802e-03
BORDEAUX        0.4582784052 6.051650e-02 2.396519e-02
RHON            0.0204281240 3.297902e-01 3.986325e-02
ANJOU           0.0325042603 4.084405e-03 6.895508e-04
AOC_AUTRES      0.0026432016 2.501761e-02 1.130109e-02
VDQS            0.0009907692 3.112513e-02 5.272777e-03
AUTRE_VDQS      0.0801778976 1.089371e-02 1.424105e-02
PROVENCE        0.0116974023 1.301664e-08 1.416322e-05
MUSCAT          0.0527131726 4.795295e-04 1.915287e-03
RHONE           0.0821504115 4.059334e-04 1.331755e-03
AOC_FORT        0.0569438449 7.330576e-03 1.392512e-04
AUTRE_FORT      0.0002084842 5.074287e-02 1.468453e-02
TRES_FORT       0.0104105739 7.843465e-03 5.182513e-03

$contrib
                    Dim.1       Dim.2        Dim.3        Dim.4
CHMPAGNE         4.149112  0.32841421  9.318043534 5.330038e+00
MOUSSEAU_AOC     6.802888  1.90990568  0.243502387 2.163962e+01
MOUSSEAU_SIMPLE  7.239295  4.36070986  6.580142112 1.539396e+00
ALSACE           3.517351 20.70353539  0.002270256 4.498754e-01
GIRONDE         10.044443  1.47727081  1.215043281 3.162286e-01
BOJOLAIS         7.629049  7.77165484  0.246308084 4.328451e+00
BORDEAUX         2.257900  0.75016440  7.462056075 2.189205e+00
RHON             3.034373  0.08401877 12.374652287 3.594427e-04
ANJOU            7.304749  6.61780706  3.629338330 2.225350e-01
AOC_AUTRES       8.129011  0.74146963  7.757206332 2.443419e-01
VDQS             9.009631  2.45258531  0.220331246 5.836282e+00
AUTRE_VDQS       6.286845  8.02766060  2.006600410 1.853380e+00
PROVENCE         4.224996 13.94596161  4.468502322 3.401712e+00
MUSCAT           5.462260 11.73234675  0.903288326 4.504747e+00
RHONE            4.425776  0.04450683 18.477846366 1.305634e+00
AOC_FORT         5.505284 11.20850538  0.802596401 4.984168e+00
AUTRE_FORT       1.013356  2.22888104 23.631682623 9.684052e+00
TRES_FORT        3.963682  5.61460182  0.660589628 3.216998e+01
                     Dim.5        Dim.6        Dim.7
CHMPAGNE         8.96696128 1.886154e+01 38.577442463
MOUSSEAU_AOC     3.93408434 1.756964e-02  1.110808175
MOUSSEAU_SIMPLE  0.56196752 1.288010e+00  1.458005687
ALSACE           1.96630212 7.436040e-02  0.006486386
GIRONDE          1.10421522 1.715059e+00  0.114768942
BOJOLAIS         0.05528851 8.797047e-01  1.363501204
BORDEAUX        47.24180451 8.840242e+00 11.592359020
RHON             2.10584097 4.817570e+01 19.282515613
ANJOU            3.35071410 5.966493e-01  0.333547217
AOC_AUTRES       0.27247545 3.654569e+00  5.466526555
VDQS             0.10213382 4.546754e+00  2.550529892
AUTRE_VDQS       8.26516921 1.591352e+00  6.888631277
PROVENCE         1.20583119 1.901469e-06  0.006850986
MUSCAT           5.43395754 7.004960e-02  0.926456418
RHONE            8.46850657 5.929869e-02  0.644191824
AOC_FORT         5.87007801 1.070850e+00  0.067358133
AUTRE_FORT       0.02149167 7.412511e+00  7.103152608
```

```
TRES_FORT         1.07317798 1.145772e+00  2.506867599
```

Notons que les variables contribuant le plus à la formation des axes sont celles dont la coordonnée (et donc la corrélation) sur l'axe concerné sont proches de 1 en valeur absolue.
On regarde également les variables pour lesquelles la contribution est supérieure à la contribution moyenne, c'est à dire à la somme des contributions/n (ici 100/18 = 5.6).

4.2.1 Premier axe

```
library(factoextra)
fviz_contrib(data.pca, choice = "var", axes = 1, top = 18)
```



Obtenir la somme des contributions des variables contribuant au delà de la moyenne

```
sum(data.pca$var$contrib[which(data.pca$var$contrib[,1] >=  sum(data.pca$var$contrib[,1])/ncol(vins)),1
```

```
[1] 62.44591
```

Obtenir les variables ayant une coord > 0

```
which(data.pca$var$coord[,1]>=0)
```

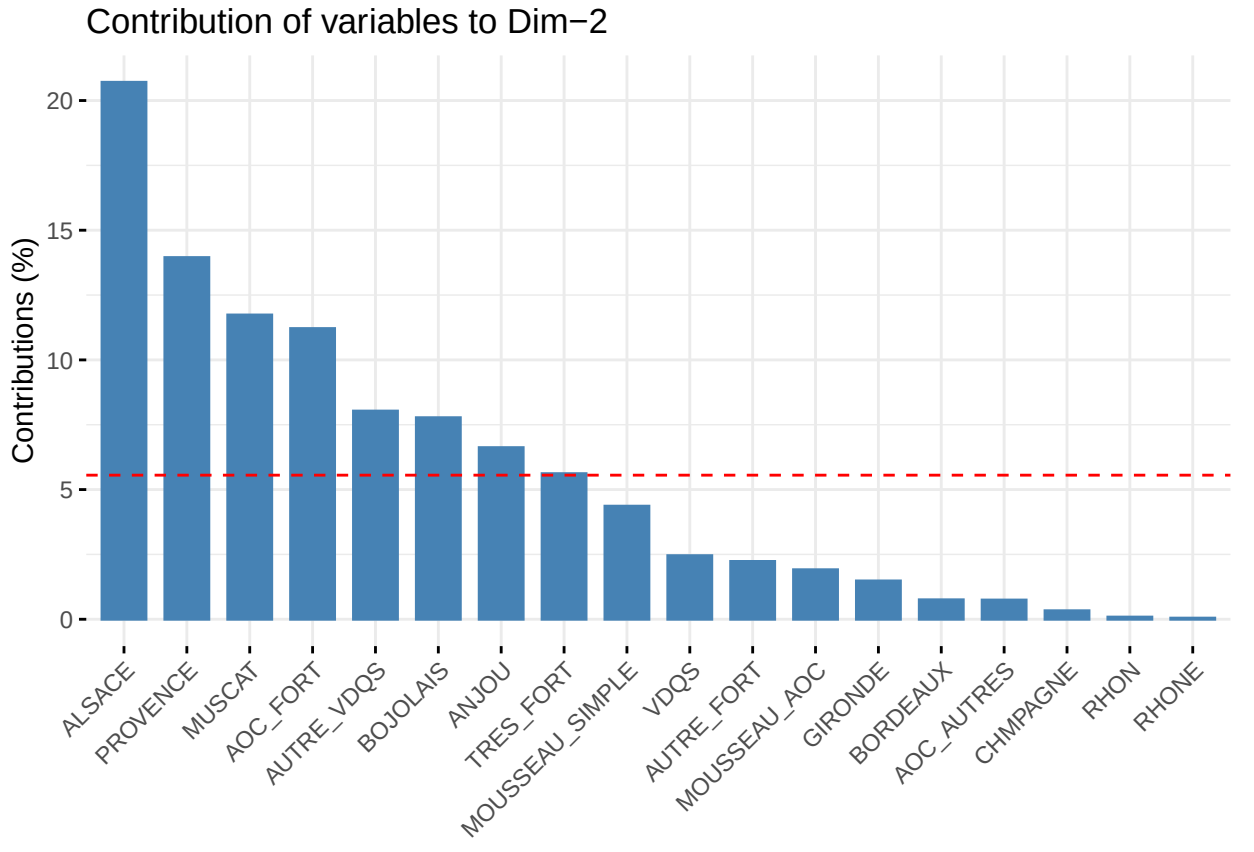
```
      CHMPAGNE    MOUSSEAU_AOC MOUSSEAU_SIMPLE         ALSACE
             1               2               3              4
      GIRONDE        BOJOLAIS        BORDEAUX           RHON
             5               6               7              8
        ANJOU      AOC_AUTRES            VDQS      AUTRE_VDQS
             9              10              11             12
```

```
     PROVENCE         MUSCAT          RHONE       AOC_FORT
           13             14             15             16
  AUTRE_FORT      TRES_FORT
           17             18
```

L'axe 1 est expliqué principalement par les variables GIRONDE, VDQS, AOC_AUTRES, ANJOU, AUTRE_VDQS, MOUSSEAU_SIMPLE, BOJOLAIS et MOUSSEAU_AOC dont les contributions (contrib) cumulées font 62,44 : Ces huit variables expliquent 62,44% de l'information contenue dans la première dimension. La contribution des deux derniers vins (AOC_FORT et MUSCAT) est plus importante sur l'axe 2, nous ne les retenons donc pas sur l'axe 1.
Cet axe regroupe du même côté ces vins (signe coord > 0), qui sont presque toutes très corrélées (cor) avec l'axe 1 (0.95 pour GIRONDE). Ils seront globalement bien représentés sur cet axe car leur qualité de représentation (cos^2) est pour la plupart bonne.
Cet axe regroupe ensemble les différents types de vins rouges (95% de la production des vins bojolais est rouge) et pétillants (mousseux simple et aoc) principalement, issus d'une zone géographique déterminée et ne pouvant être reproduit hors de cette zone (cela implique qu'il y aurait un lien entre le caractère du vin et sa provenance). Il s'agit des vins de qualité produits dans une région déterminée, comprenant les vins d'Appelation d'Origine Contrôlée (AOC) et des Vins Délimités de Qualités Superieurs (VDQS).
La variable GIRONDE a un rayon vecteur important sur cet axe 1, elle sera donc la mieux représentée. A l'inverse, la varible AUTRE_VDQS a un rayon vecteur petit.

4.2.2 Deuxième axe

```r
fviz_contrib(data.pca, choice = "var", axes = 2, top = 18)
```



Obtenir la somme des contributions des variables contribuant au delà de la moyenne à l'axe 2 exclusivement

```
sum(data.pca$var$contrib[which(data.pca$var$contrib[,2]
                               >= sum(data.pca$var$contrib[,2])/18),2])- # Enlève les v.a ret à axe 1
                              sum(data.pca$var$contrib[which
                              (data.pca$var$contrib[,1] >=
                               sum(data.pca$var$contrib[,1])/ncol(vins) &
                               data.pca$var$contrib[,2] >=
                               sum(data.pca$var$contrib[,2])/18),2])
```

```
[1] 63.20495
```

Obtenir les variables ayant une coord > 0

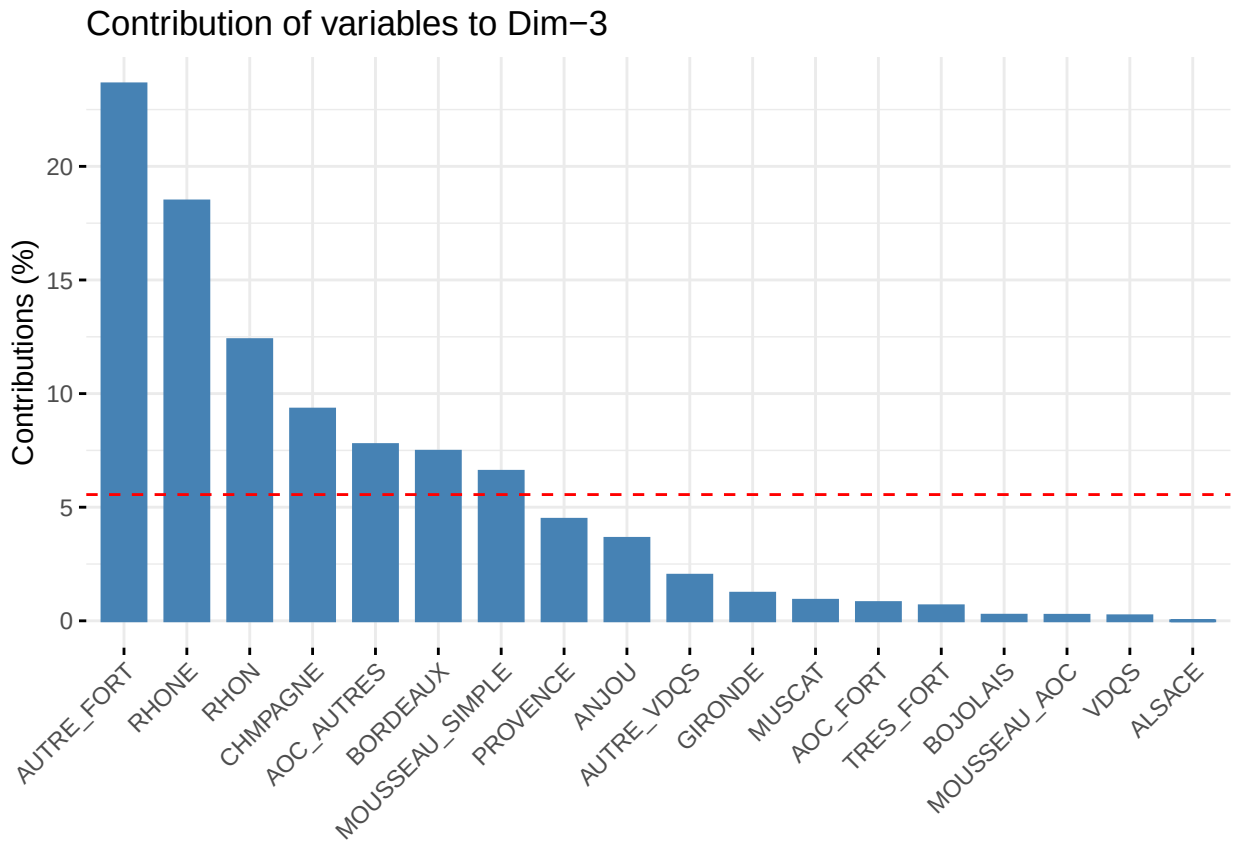
```
which(data.pca$var$coord[,2]>=0)
```

```
      CHMPAGNE    MOUSSEAU_AOC MOUSSEAU_SIMPLE          ALSACE
             1               2               3               4
      BOJOLAIS      AUTRE_VDQS        PROVENCE           RHONE
             6              12              13              15
```

L'axe 2 est dû en grande partie aux variables ALSACE, MUSCAT, AOC_FORT, PROVENCE et TRES_FORT (les autres vins ont été retenus sur l'axe 1), qui contribuent à 63,2% à la formation de cet axe (on enlève en effet les contributions des vins retenus sur l'axe 1). L'axe 2 sépare le vin alsace (côté positif) des quatre autres vins (côté négatif). En effet, cet axe differencie les différents types de vins "clairs" (vins blancs, vins rosés etc.). La corrélation entre cet axe et ces vins est négative, sauf pour la variable ALSACE pour laquelle elle est positive et forte. Globalement, les vins ne sont pas bien représentés sur cet axe car les cos^2 sont relativement faibles.

4.2.3 Troisième axe

```
fviz_contrib(data.pca, choice = "var", axes = 3, top = 18)
```



Obtenir la somme des contributions des variables contribuant au delà de la moyenne à l'axe 3 exclusivement

```
sum(data.pca$var$contrib[which(data.pca$var$contrib[,3]
                               >= sum(data.pca$var$contrib[,3])/18),3])- # Enlève les v.a retenues axe
                                sum(data.pca$var$contrib[which
                               (data.pca$var$contrib[,1] >=
                                sum(data.pca$var$contrib[,1])/ncol(vins) &
                                data.pca$var$contrib[,3] >=
                                sum(data.pca$var$contrib[,3])/18),3])
```

```
[1] 71.26428
```

```
                               -sum(data.pca$var$contrib[which  # Enlève les v.a ret axe 2
                               (data.pca$var$contrib[,2] >=
                                sum(data.pca$var$contrib[,2])/ncol(vins) &
                                data.pca$var$contrib[,3] >=
                                sum(data.pca$var$contrib[,3])/18),3])
```

```
[1] 0
```

Obtenir les variables ayant une coord > 0

```
which(data.pca$var$coord[,3]>=0)
```

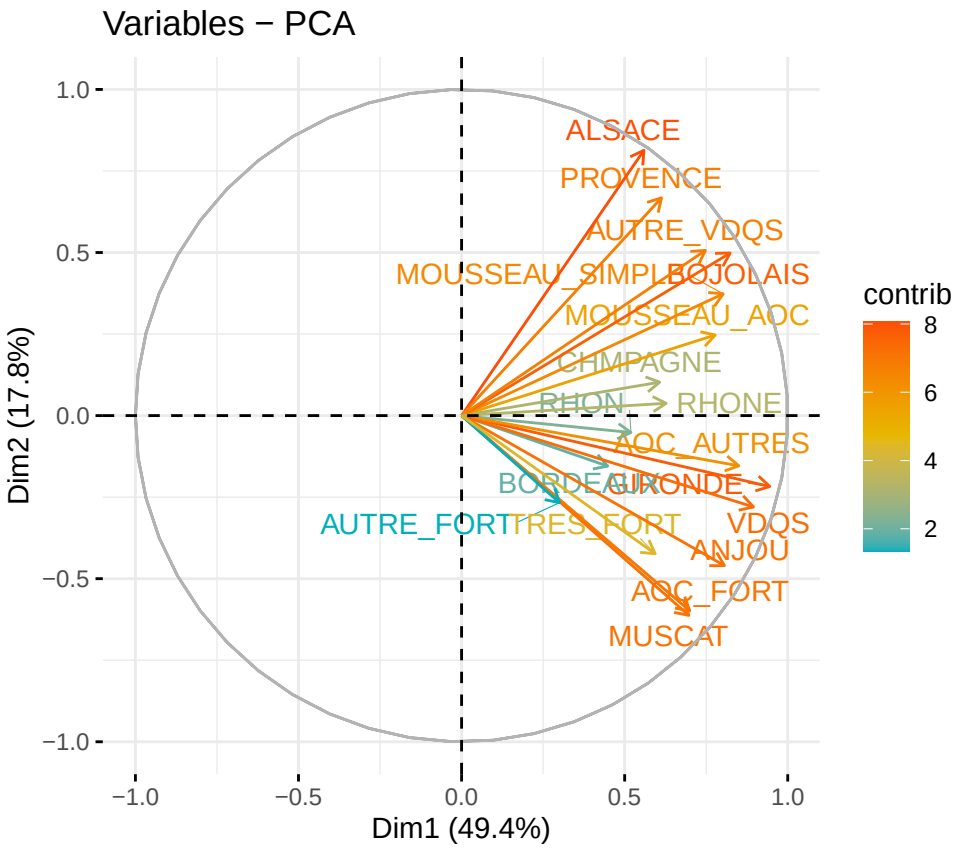
```
    ALSACE     GIRONDE        RHON AOC_AUTRES   PROVENCE      RHONE
         4           5           8          10          13         15
AUTRE_FORT
        17
```

L'axe 3 est formé grâce aux variables CHMPAGNE, RHON,RHONE, BORDEAUX et AUTRE_FORT (les autres ont été retenues sur les deux premiers axes) dont les contributions atteignent 71,26% ((on enlève en effet les contributions des vins retenus sur les axe 1 et 2). Cet axe separe les vins champagne et bordeaux (côté négatif de l'axe, vins de renommé) des vins rhône, rhon et autre_fort. Les vins du rhône sont à 61% blancs, les vins bordeaux sont connus pour être rouges (89% de la production) et les vins champagnes sont à 90% mousseux : on peut donc dire que cet axe sépare les vins rouges et mousseux des vins blancs.
Malheuresement, ce n'est pas sur cet axe que le vin bordeaux contribue le plus. En effet, sa contribution à l'axe 5 est de 47.24 (contre 7.46 sur l'axe 3). Les vins seront globalement mal représentés sur cet axe.

4.2.4 Cercle de corrélation

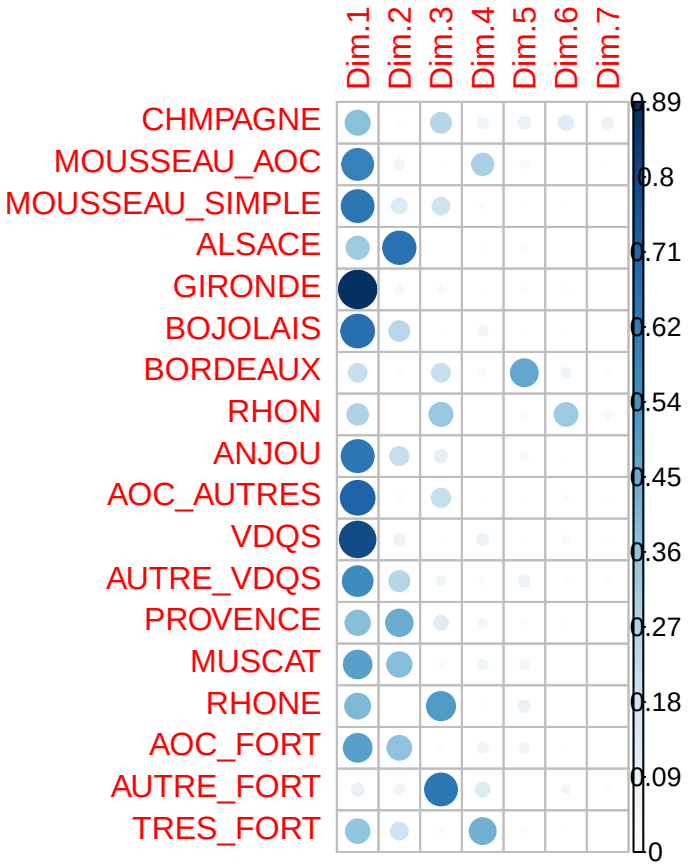
Le graphique ci-dessous correspond au cercle de corrélation des variables, avec une coloration en fonction de la contribution des variables. Il montre les relations entre les composantes et les vins qui contribuent le plus à la formation des axes.
Sur ce cercle de corrélation formé par les 2 premiers facteurs, on observe que les vins se projettent toutes du même côté de l'axe 1 : il s'agit d'un effet taille (elles contribuent toutes dans le même sens à la formation de l'axe 1). Elles sont donc toutes corrélées positivement avec la dimension horizontale. L'axe 1 est donc un axe d'échelle entre les fortes consommations de certains vins et les faibles consommations de ces mêmes vins.
La deuxième composante différencie alors les individus de "taille" semblable : on parle d'effet forme.
Du fait de « l'effet taille » sur l'axe 1, on en déduit que cet axe oppose essentiellement les pays ayant une forte consommation des vins GIRONDE, VDQS, AOC_AUTRES,ANJOU,AUTRE_VDQS etc (i.e. tous les vins de l'axe 1) à ceux ayant de faibles consommations de ces mêmes types de vins.
Sur ce graphique, on peut également voir que certaines variables sont très fortement corrélées(AOC_FORT et MUSCAT par exemple) et d'autres ont une corrélation faible (ALSACE et ANJOU par exemple)

```
fviz_pca_var(data.pca,
             col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,
             )
```



Nous pouvons aussi visualiser le cos^2 des variables sur toutes les dimensions en utilisant le package corrplot. Un cos^2 élevé indique une bonne représentation de la variable sur les axes principaux en considération. Dans ce cas, la variable est positionnée à proximité de la circonférence du cercle de corrélation. A l'inverse, un faible cos2 indique que la variable n'est pas parfaitement représentée par les axes principaux. Dans ce cas, la variable est proche du centre du cercle.
Le code R est le suivant :

```
library("corrplot")
corrplot(data.pca$var$cos2, is.corr=FALSE)
```



4.3 Interprétation des axes par rapport aux individus

```
data.pca$ind
```

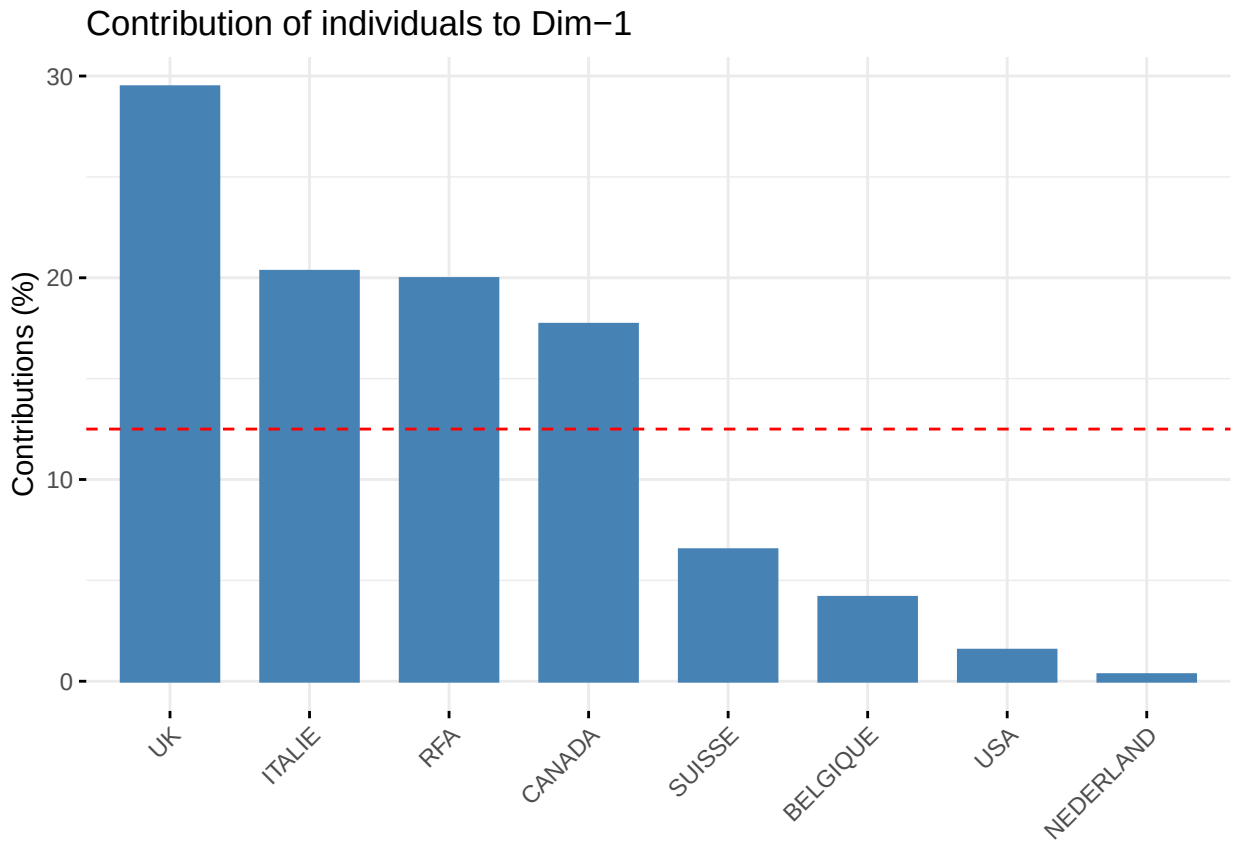
```
$coord
               Dim.1       Dim.2      Dim.3      Dim.4      Dim.5
BELGIQUE   1.7199251 -0.45666525  0.3039831 -2.9170688  0.1772091
NEDERLAND  0.4830365 -0.73341525  3.9433322  0.7888259  0.2015239
RFA        3.7675720  4.00695287 -0.3434222  0.4661844 -0.4040089
ITALIE    -3.8011925  0.09381538 -0.9857779  0.3110341 -0.8582476
UK         4.5776332 -2.95035530 -1.4359711  0.9144179 -0.6364164
SUISSE    -2.1532849  0.20311031  0.4326977  0.3211369  0.7650370
USA       -1.0464542 -0.09474951 -1.6667549  0.3907789  2.0377953
CANADA    -3.5472351 -0.06869325 -0.2480869 -0.2753094 -1.2828924
               Dim.6       Dim.7
BELGIQUE   0.07486158  0.15317875
NEDERLAND -0.56331337  0.16399938
RFA       -0.04121199 -0.02018547
ITALIE     0.10884192  0.92800823
UK         0.25464841 -0.06517461
SUISSE     1.88631572 -0.32721438
USA       -0.99701578 -0.03950606
CANADA    -0.72312649 -0.79310584

$cos2
               Dim.1        Dim.2       Dim.3       Dim.4       Dim.5
```

```
BELGIQUE  0.25007854 0.0176300403 0.007811892 0.719367360 0.002654785
NEDERLAND 0.01346503 0.0310418248 0.897375289 0.035909532 0.002343691
RFA       0.46159997 0.5221209555 0.003835301 0.007067375 0.005307921
ITALIE    0.84319933 0.0005136161 0.056708565 0.005645555 0.042984881
UK        0.63438394 0.2635231877 0.062425435 0.025313935 0.012261762
SUISSE    0.50295528 0.0044749717 0.020309329 0.011186823 0.063487939
USA       0.11924884 0.0009776147 0.302522213 0.016629376 0.452204506
CANADA    0.81061241 0.0003039915 0.003964980 0.004882871 0.106026340
                Dim.6        Dim.7
BELGIQUE  4.737785e-04 1.983601e-03
NEDERLAND 1.831249e-02 1.552143e-03
RFA       5.523189e-05 1.325014e-05
ITALIE    6.913260e-04 5.025672e-02
UK        1.963144e-03 1.285959e-04
SUISSE    3.859714e-01 1.161425e-02
USA       1.082475e-01 1.699578e-04
CANADA    3.368696e-02 4.052244e-02

$contrib
              Dim.1       Dim.2      Dim.3      Dim.4      Dim.5
BELGIQUE   4.160278  0.81556277  0.4234748 80.1749766  0.4046495
NEDERLAND  0.328143  2.10359113 71.2615805  5.8628374  0.5233114
RFA       19.963028 62.78991946  0.5404873  2.0476748  2.1032403
ITALIE    20.320905  0.03441990  4.4533503  0.9115104  9.4914445
UK        29.470350 34.04161022  9.4497450  7.8783442  5.2190299
SUISSE     6.520882  0.16133380  0.8580210  0.9716859  7.5417469
USA        1.540082  0.03510875 12.7312844  1.4388240 53.5091707
CANADA    17.696332  0.01845396  0.2820567  0.7141468 21.2074069
               Dim.6       Dim.7
BELGIQUE   0.10233363  1.41872452
NEDERLAND  5.79429342  1.62624312
RFA        0.03101327  0.02463648
ITALIE     0.21631804 52.07205174
UK         1.18408347  0.25683719
SUISSE    64.97243355  6.47389733
USA       18.15116188  0.09436881
CANADA     9.54836274 38.03324081

$dist
 BELGIQUE NEDERLAND       RFA    ITALIE        UK    SUISSE       USA
 3.439310  4.162712  5.545346  4.139565  5.747314  3.036245  3.030352
   CANADA
 3.939883
```

Ci-dessus les contributions des individus aux axes, leurs coordonées et qualités de représentation.

4.3.1 Premier axe

```
fviz_contrib(data.pca, choice = "ind", axes = 1, top = 8)
```



Obtenir la somme des contributions des variables contribuant au delà de la moyenne

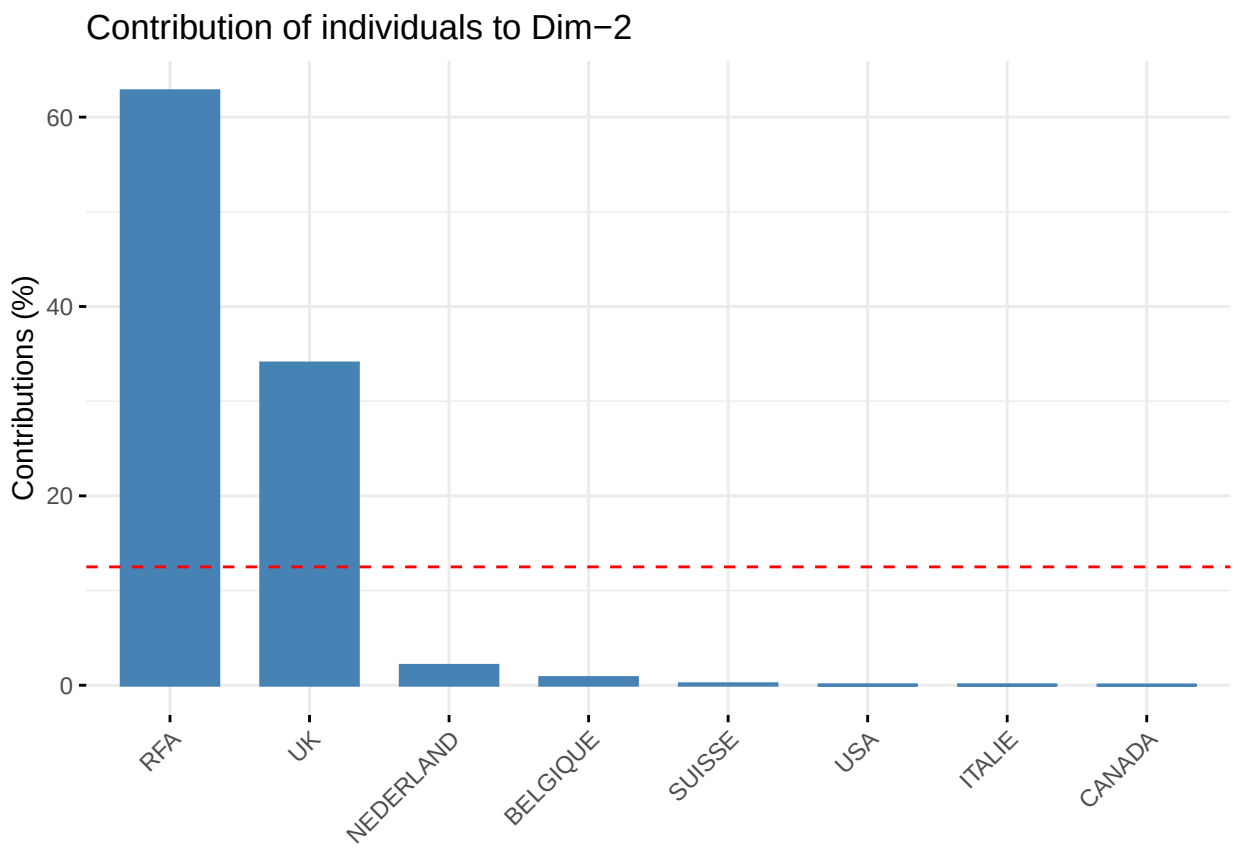
```
sum(data.pca$ind$contrib[which(data.pca$ind$contrib[,1] >= sum(data.pca$ind$contrib[,1])/8),1])
```

```
[1] 87.45062
```

L'axe 1 est formé du Canada, du United Kingdom (UK), de l'Italie et de l'Allemagne (RFA). Ces quatres individus contribuent à 87,45% à la formation de cet axe (UK est le pays qui contribue le plus). Il s'agit d'un axe de séparation des pays RFA et UK (du côté positif) des deux autres (côté négatif). En effet, les coordonées de ces derniers sur cet axe sont négatifs car la quantité de vins (constituant l'axe 1) consommés dans ces pays est faible. De plus, nous pouvons dire que l'Allemagne et le Royaume Uni sont des pays consommateurs de vins provenant d'une région géographique spécifique (en grande quantité), contrairement aux deux autres pays pour lesquels la consommation de ce type de vins est faible.
Néanmoins, l'individu RFA est le plus mal représenté sur cet axe.

4.3.2 Deuxième axe

```
fviz_contrib(data.pca, choice = "ind", axes = 2, top = 8)
```



Obtenir la somme des contributions des variables contribuant au delà de la moyenne

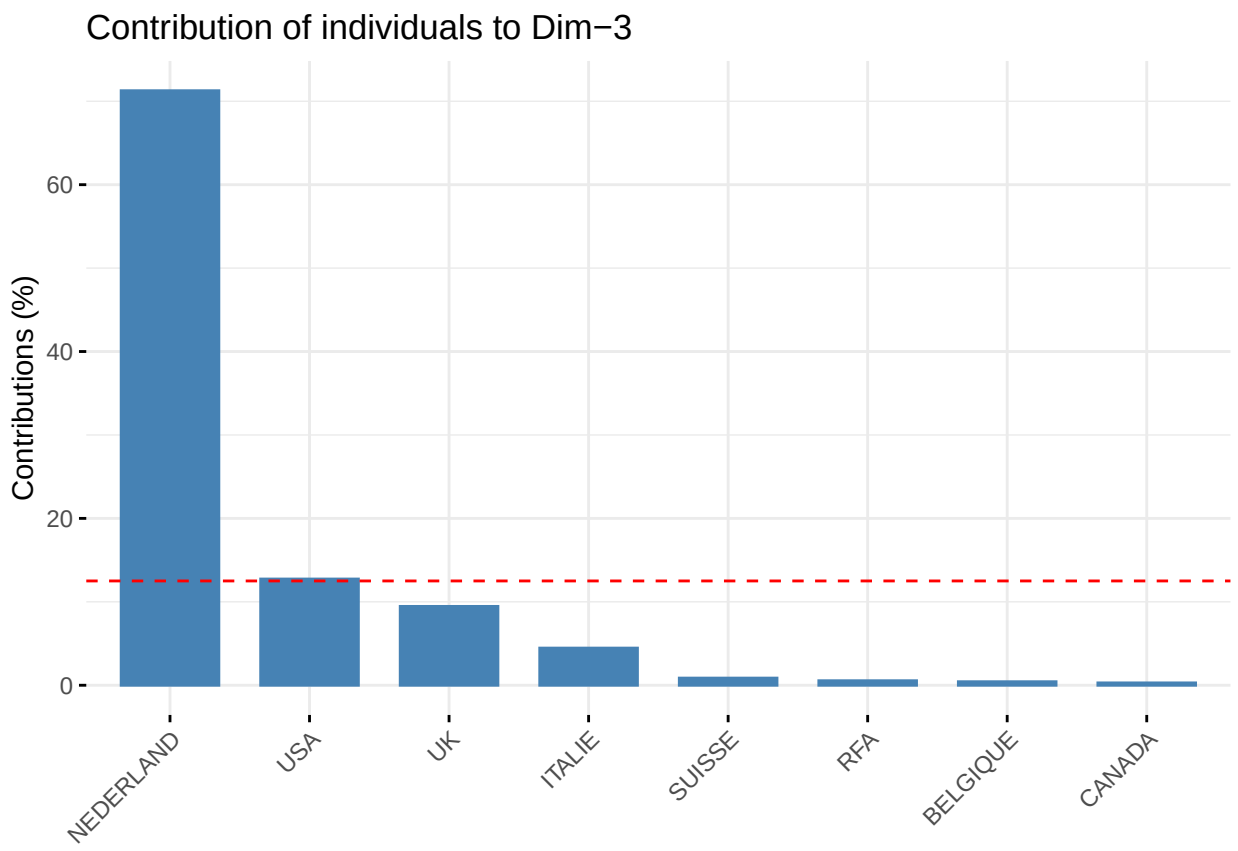
```
sum(data.pca$ind$contrib[which(data.pca$ind$contrib[,2] >= sum(data.pca$ind$contrib[,2])/8),2])
```

```
[1] 96.83153
```

L'axe 2 : ce sont les individus RFA et UK qui contribuent le plus à la formation de cet axe (à 96.83%). Il s'agit également d'un axe de séparation : RFA du côté positif qui est moyennement bien représenté et UK du côté négatif, mal représenté (cos^2 faible). Il y a donc une différence entre les types de vins "clairs" consommés dans ces deux pays. En effet, les allemands ont tendance à consommer du vin d'alsace alors que chez les anglais, il s'agit des vins MUSCAT, AOC_FORT, TRES_FORT et PROVENCE

4.3.3 Troisième axe

```
fviz_contrib(data.pca, choice = "ind", axes = 3, top = 8)
```



Obtenir la somme des contributions des variables contribuant au delà de la moyenne

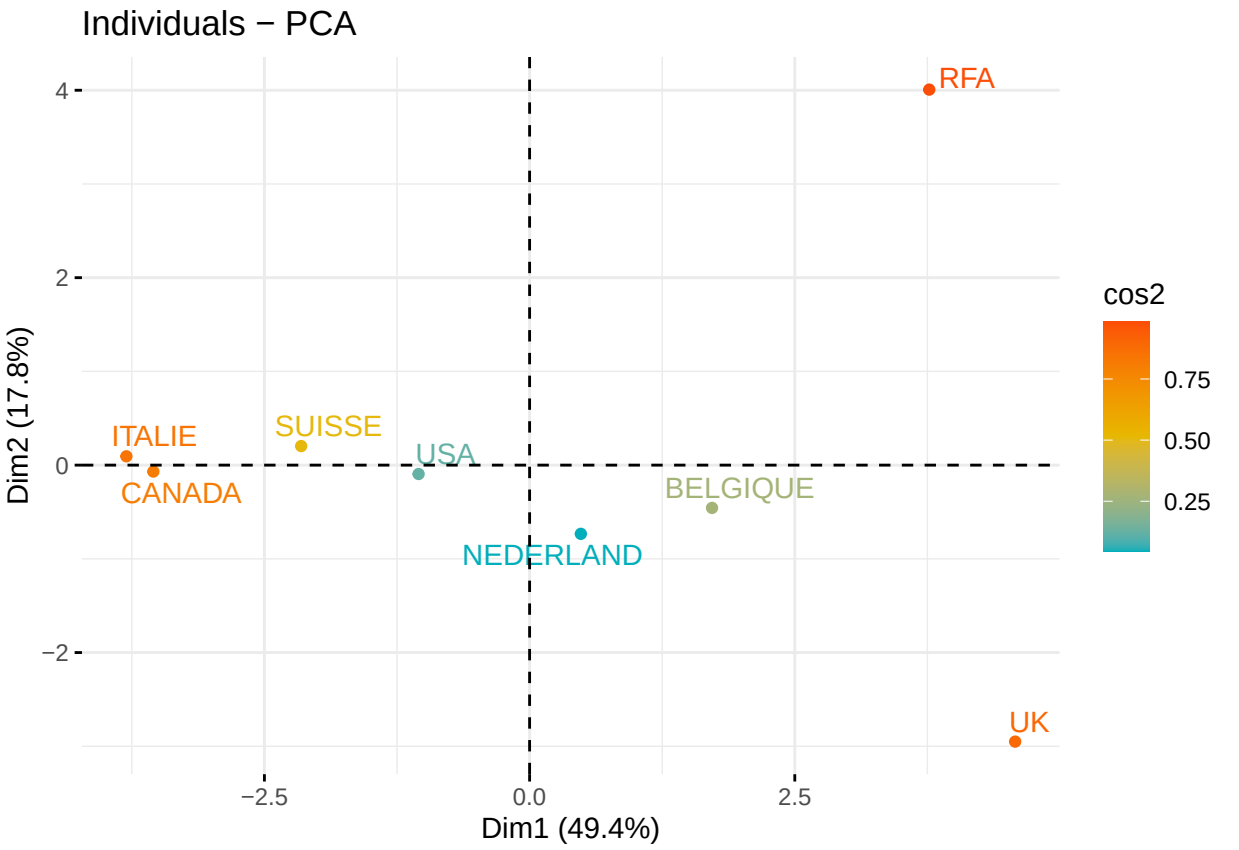
```
sum(data.pca$ind$contrib[which(data.pca$ind$contrib[,3] >= sum(data.pca$ind$contrib[,3])/8),3])
```

```
[1] 83.99286
```

Sur l'axe 3, ce sont les pays USA et NEDERLAND qui contribuent le plus (à 84%). Cet axe oppose les États-Unis (consommateurs de vins champagne et bordeaux) aux Pays-Bas (consommateurs de vins du rhône, rhon et des autres vins forts).
C'est à la formation de l'axe 4 que la Belgique contribue, mais nous n'avons pas gardé cet axe dans notre analyse.

4.3.4 Nuage des individus

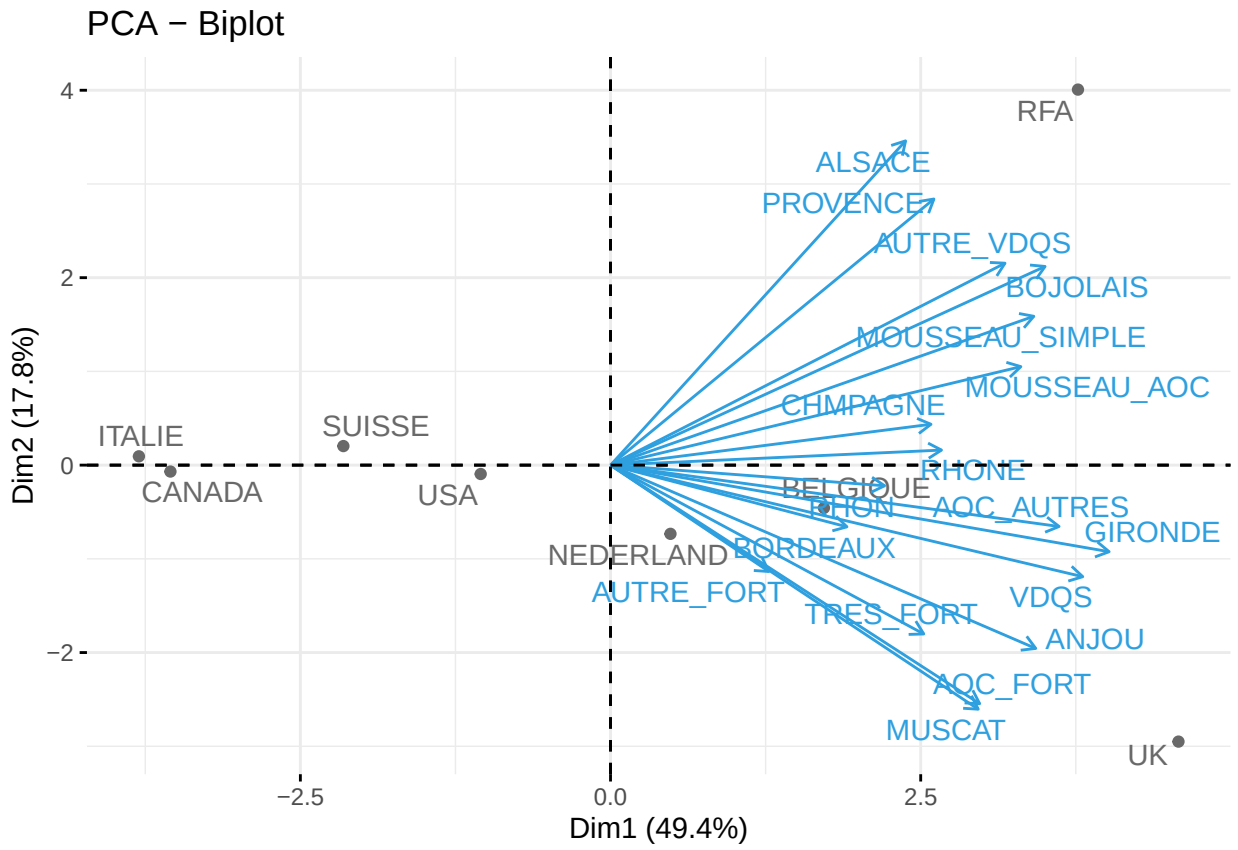
```
fviz_pca_ind(data.pca,
             col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE
             )
```



Au vu du graphique, nous pouvons dire que les types de vins consommés en Italie sont les mêmes qu'au Canada (individus similaires). Le Canada et le Royaume Uni ont des types de consommations inverses.
Le nuage des individus nous montre qu'il existe quatre classes au moins. La première classe serait constituée de l'Italie, du Canada, des USA et de la Suisse. La deuxième quant à elle contiendrait le Nerderland et la Belgique. Les troisième et quatrième seraient des singletons et contriendrait la RFA et UK. Nous effectuerons la classification plus tard pour approfondir notre réflexion.

4.4 Biplot des individus et des variables

```
fviz_pca_biplot(data.pca, repel = TRUE,
                col.var = "#2E9FDF",
                col.ind = "#696969"
                )
```



Globalement, le biplot peut être interprété comme suite:
- un individu qui se trouve du même côté d'une variable donnée a une valeur élevée pour cette variable;
- un individu qui se trouve sur le côté opposé d'une variable donnée a une faible valeur pour cette variable.
Ainsi, on voit que les pays comme l'Italie, la Suisse, les USA et le Canada ont des quantités faibles pour tous les types de vins. Il s'agit d'invidus simialires. RFA aura des quantités faibles pour les vins comme AOC_FORT, TRES_FORT, AUTRE_FORT etc… et des quantités fortes pour le vin BOJOLAIS par exemple. Au Royaume Uni (UK), ce sont les vins ANJOU et AOC_FORT qui sont principalement consommés et les vins de PROVENCE par exemple ne sont que très peu consommés.

5 Classification

Nous avons realiser l'ACP sur notre jeu de données et nous avons fait le choix des 3 premiers axes factoriels qui contiennent plus 82% de l'information des données brutes. Nous appliquerons dans la suite une **classification hierarchique ascandente** en utilisant la sortie de notre **ACP** realisée ci-dessus et la fonction **HCPC** de R. A l'issu de l'ACP, nous avons intuitivement détecté l'existence de quatre classes. Cependant nous allons s'en tenir au nombre de classes optimales proposé par le logiciel R.
Notre objectif est de régrouper les pays ayant des profils de consommation similiaire. Pour ce faire, nous allons construire nos classes de manière à miniser la variance intra-classe et à maximer la variance inter-classe (il faut que les individus de deux classes différentes aient des profils différents).
On met nb.clust à -1 car nous ne savons pas à priori en combien de classes nous allons séparer notre jeu de données, on laisse donc le logiciel décider.
D'abord, refaisons l'ACP avec les trois premiers axes.

```
library(FactoMineR)
data.pca = PCA(vins, scale.unit=TRUE, ncp =3 , graph = FALSE)
```

5.1 Nombre de classes proposé par R

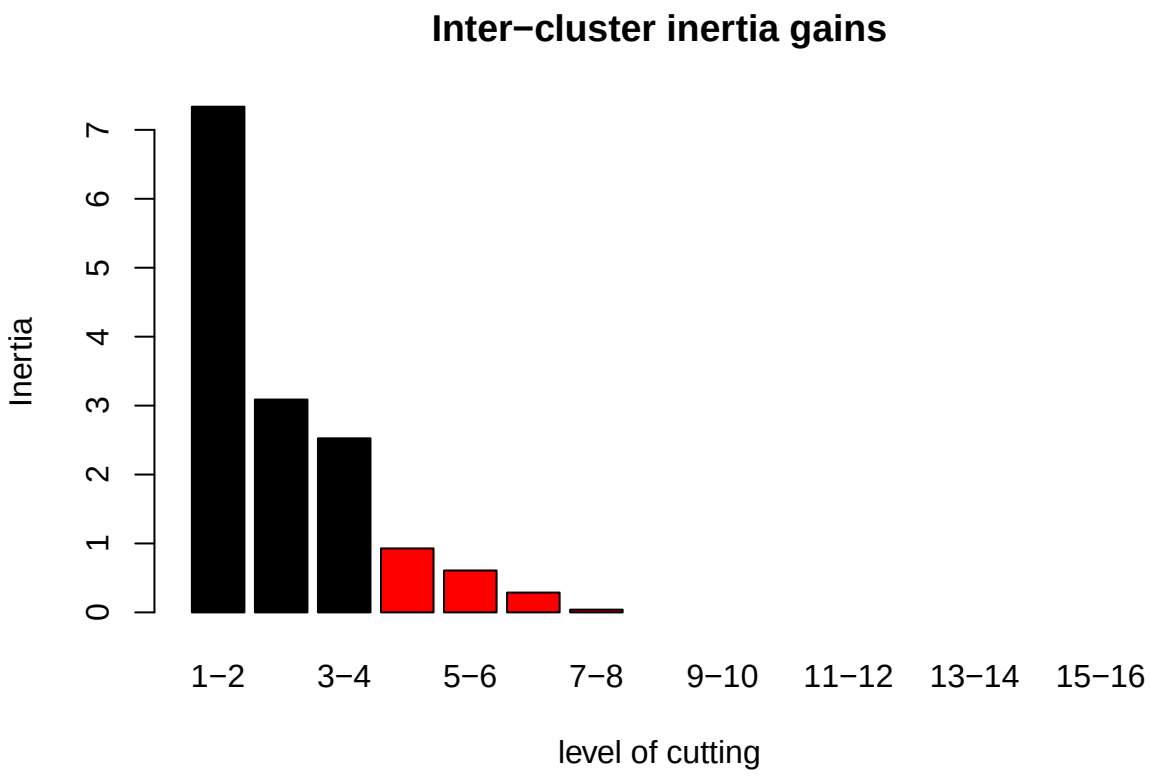
```
hcpc=HCPC(data.pca
          ,nb.clust=-1
          ,proba=1,method = "ward",graph = FALSE)
hcpc$call$max
```

```
[1] 4
```

Le logiciel R nous propose de faire une partition en quatre (4) classes. Dans cette partition, nous avons deux classes qui contiennent qu'un seul pays chacune. Nous avons donc décidé de garder ces quatre classes malgré l'existence de deux singletons. En effet, plus le nombre de classes est élévé, plus la classification est bonne, c'est à dire plus la variance inter-classe est grande.

5.2 Gain d'inertie

Le graphique suivant est aussi utilisé pour le choix du nombre de classes lors d'une classification ascendente hierarchique, il s'agit du graphique du **gain d'inertie inter-classe**.



En analysant ce graphique, on se rend compte que le passage d'une classe unique à deux classes permet de gagner enormement en inertie inter-classe. Il en est de même de 2 à 3 et 3 à 4. Par contre un nombre de classe au déla de quatre (4) a très peu d'intêret (très faible gain en inertie inter-classe). Et d'ailleurs on remarque qu'il est incensé de faire plus de 8 classe puisqu'on a que huit individus.

6 Composition des classes et paragons

Dans cette partie, nous allons découvrir pour chaque classe, le nombre de pays affecté ainsi que les noms de ces pays. De plus pour chaque classe, nous determinerons le **paragon** c'est à dire l'individu le plus proche du centre de gravité de cette classe ou tout simplement l'individu qui caracterise la classe.

6.1 Première classe

Pour connaitre les pays qui composent cette classe, voyons le tableau ci-dessous,

Pays	Classe
ITALIE	1
SUISSE	1
USA	1
CANADA	1

Nous pouvons donc observer que cette première classe est composée de quatre(4) pays à savoir l'**Italie** , la **Suisse**, les **USA** et le **Canada**. Mais quel est le paragon de cette classe ? Voyons le tableau suivant :

```
hcpc$desc.ind$para$`1`
```

```
   CANADA    SUISSE    ITALIE       USA
0.9873964 1.1681851 1.2226661 1.9100811
```

Ce tableau nous résume les ecarts entre chaque individu (pays) et le centre de gravité de la classe. Le pays ayant le plus petit écart correspond au paragon. Ainsi donc l'**Italie** est le paragon de cette première classe. En d'autres termes, l'Italie est le pays qui caracterise le mieux cette classe.

6.2 Deuxième classe

Observons le tableau ci-dessous

Pays	Classe
BELGIQUE	2
NEDERLAND	2

Nous constatons que cette deuxième classe est constituée de deux pays à savoir la **Belgique** et les **Pays-bas** (Nederland). Nous pouvons aussi constater dans le tableau ci-dessous que les deux pays sont equidistant du centre de gravité de la classe, ce qui est logique puisqu'on a que deux points (individus ayant le même poids) et donc le centre de gravité correspond au milieu du segment qui les separent : ils sont donc tous des paragons de cette classe, c'est à dire que l'un ou l'autre caracterise la classe.

```
hcpc$desc.ind$para$`2`
```

```
 BELGIQUE NEDERLAND
 1.926872  1.926872
```

6.3 Troisième classe

Cette classe est composé d'un seul pays qui est la **République Fédérale de l'Allemagne**(RFA). On peut vérifier cela dans le tableau ci-dessous. Etant composé d'un seul pays (individu), le paragon de cette classe est bien evidemment l'individu, lui seul qui caracterise sa classe.

```
knitr::kable(ind.per.class(hcpc,3))
```

Pays	Classe
RFA	3

6.4 Quatrième classe

Nous savons que sept pays sur huit ont été classés dans les trois premières classes, donc cette classe sera composé d'un seul pays qui est le **Royaume Uni(UK)**. On peut vérifier cela dans le tableau ci-dessous. Etant composé d'un seul pays (individu), le paragon de cette classe est bien evidemment l'individu, lui seul qui caracterise sa classe.

```
knitr::kable(ind.per.class(hcpc,4))
```

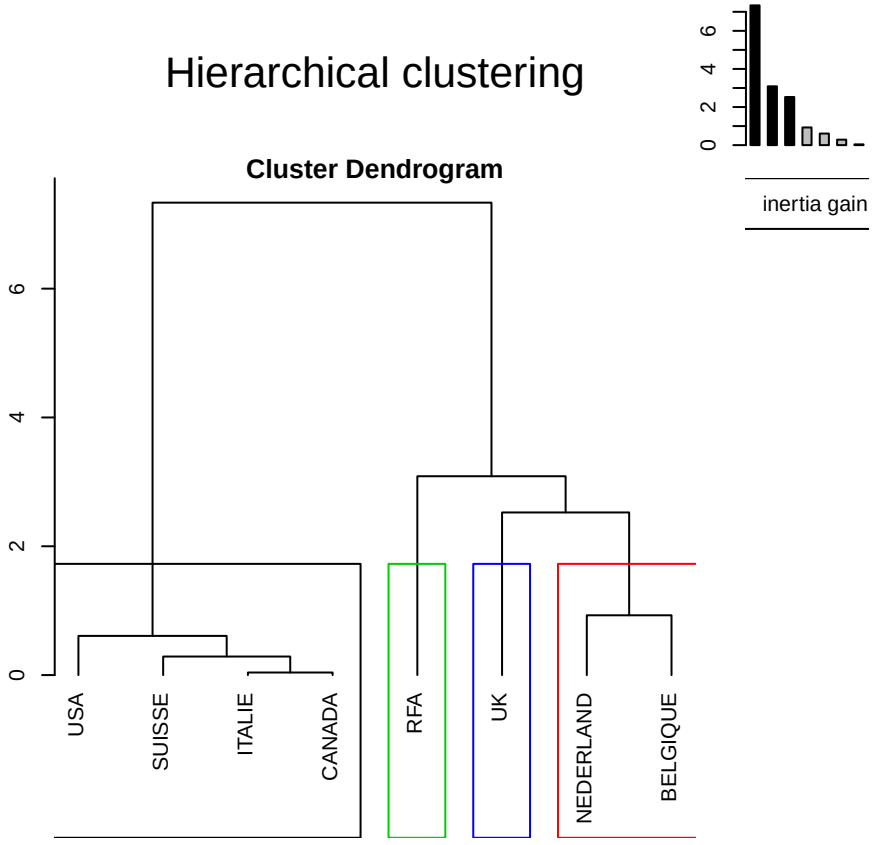
Pays	Classe
UK	4

6.5 Graphique récapitulatif (dendogramme)

La répartition des pays par classe peut être visualisée par un graphique : le **dendogramme** en dimension 2 ou 3.

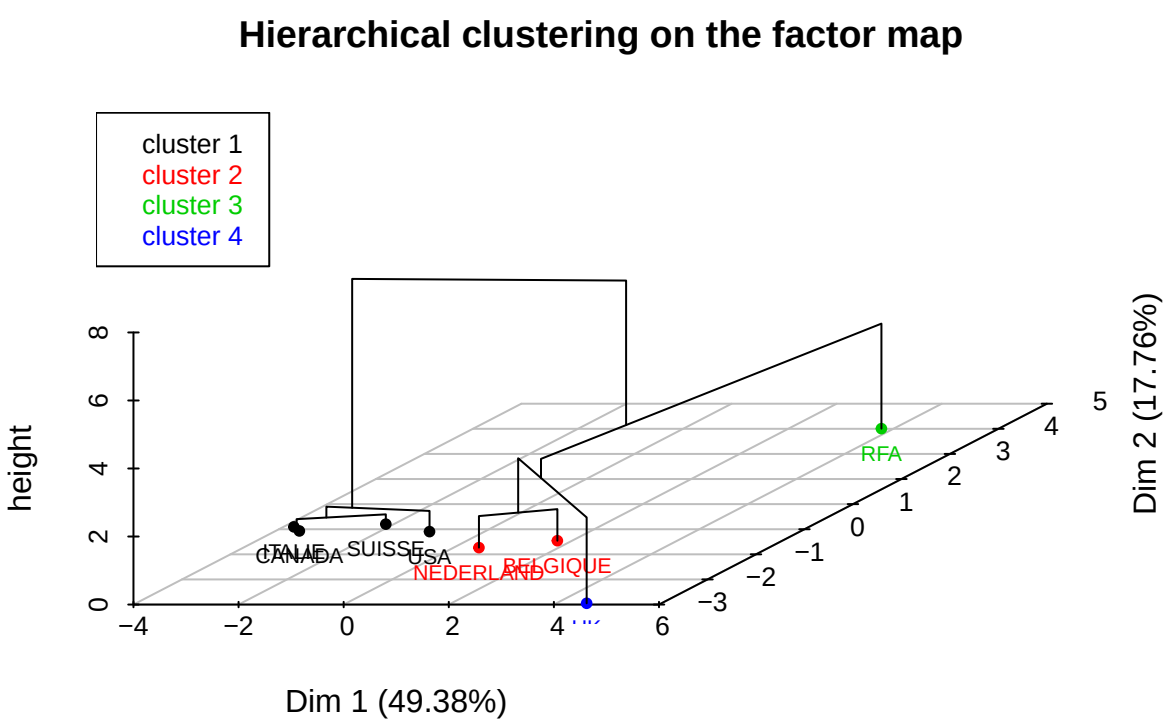
Dans le plan ou en dimension 2 on obtient le graphique suivant:

```
plot(hcpc,choice="tree")
```



Nous pouvons bien observer et en même temps les quatres classes avec leur composition.

Dans l'espace ou en dimension 3, nous avons :



Ce graphique,beaucoup plus riche en information permet en plus de la compostion des classes de connaitre le score d'un pays pour les deux premières composantes principales formées à l'issue de l'ACP. Ainsi on peut constater que la première classe est formée par les pays ayant un mauvais score pour les deux dimensions pendant que la deuxième classe est caractérisée par de bons scores pour la dimension 2 et la troisième classe est caractérisée par les pays ayant un score élévé pour la dimension 1 etc. Le dendogramme en dimension 3 nous a ainsi permis d'introduire la caracterisation des classes que nous allons développer dans la section suivante.

7 Caracterisation des classes

Comme annoncé précedemment, nous allons analyser de plus près la differenciation d'une classe à l' autre. Il s'agira de donner un sens à chaque classe. Pour cela nous allons dans un premier temps, identifier les facteurs qui caracterisent le mieux chaque classe et deuxièmement descendre plus en profondeur en essayant de trouver les types de vins qui differencient vraisemblablement chacune des quatre classes. Notons quand même que ces analyses se baseront sur un test statistique : la **valeur test** qui est décrit dans la sous section suivante.

7.1 Valeur test

La valeur test correspond au test d'egalité entre deux moyennes de deux echantillons differentes. Le premier echantillon correspond à la base de donnée initiale et le deuxième correspond aux individus affectés à une classe i donnée (i $\in$ {1,2,3,4}).
Soit **n** le nombre d'individus de l'echantillon 1(tous les individus) et n_g celui des individus de la classe i (sous ensemble de l'echantillon 1)(n_g < n). Considerons une variable quantitative **X** de notre jeu de données, μ sa moyenne dans l'echantillon global et μ_g dans la classe i.

La valeur test(statistique de test) est definie de la manière suivante :

$$t_c = \frac{\mu_g - \mu}{\sqrt{\frac{n-n_g}{n-1} \times \frac{\sigma^2}{n_g}}} \hookrightarrow \mathrm{N}(0,1)$$

L'hypothèse nulle de ce test est : $H_0 : \mu_g = \mu$ et la contre hypothèse est $H_1 : \mu_g \neq \mu$.
En prenant le niveau de risque $\alpha = 5\%$, on a :

- Si $|t_c| > 1.96$, alors X caracterise la classe i considéré.
- X caracterise d'autant mieux la classe i que $|t_c|$ est grande.

Nous analyserons ce test pour chacune des dimensions rétenues ou variables de notre jeu de donnée (c'est à dire chaque type de vin) et pour chacune des 4 classes.

7.2 Analyse des classes par rapport aux facteurs

7.2.1 Première classe

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Dim.2	0.0493845	0.0333707	0	0.1217548	1.787822	0.9606129
Dim.3	-0.9883933	-0.6169805	0	0.7867567	1.651546	0.3229600
Dim.1	-2.3402522	-2.6370417	0	1.1121700	2.981284	0.0192707

En analysant le tableau ci-dessus, nous conctatons que la valeur test est significative (p.value<5%) que pour la première dimension avec une valeur **v.test** de -2.34 < 0. Ce qui veut dire que les individus de la première classe sont caracterisés par un mauvais score pour le premier axe factoriel. On peut verifier cela avec le score de l'Italie qui est le paragon de la classe.

```
  BELGIQUE  NEDERLAND        RFA     ITALIE         UK     SUISSE
 1.7199251  0.4830365  3.7675720 -3.8011925  4.5776332 -2.1532849
       USA     CANADA
-1.0464542 -3.5472351
```

On remarque effectivement que l'ITALIE a le plus mauvais score pour la dimension 1 et tous les autres individus de cette classe à savoir CANADA, SUISSE et USA ont un mauvais score.

7.2.2 Deuxième classe

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Dim.3	1.9641844	2.1236577	0	1.8196746	1.651546	0.0495087
Dim.1	0.5643675	1.1014808	0	0.6184443	2.981284	0.5725041
Dim.2	-0.5084058	-0.5950403	0	0.1383750	1.787822	0.6111688

Nous remarquons qu'avec un risque de première espèce de 5%, la valeur test est significative que pour la dimension 3. En allant plus loin dans l'analyse, nous conctatons que la valeur **v.test** pour l' axe factoriel 3 vaut 1.96. Cela s'interprète par le fait que les individus de cette deuxième classe sont caracterisés par un bon score pour la dimension 3.

7.2.3 Troisième classe

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Dim.2	2.2412482	4.0069529	0	0	1.787822	0.0250100

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Dim.1	1.2637414	3.7675720	0	0	2.981284	0.2063229
Dim.3	-0.2079399	-0.3434222	0	0	1.651546	0.8352759

En analysant le tableau précedent, nous remarquons que la valeur test est significatif que pour la dimension 2 (p.value<5%). Par ailleurs la valeur **v.test** vaut 2.24. Cela veut dire que les individus de cette classe sont caracterisés par un très bon score pour la dimension 2.

7.2.4 Quatrième classe

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Dim.1	1.5354569	4.577633	0	0	2.981284	0.1246716
Dim.3	-0.8694709	-1.435971	0	0	1.651546	0.3845896
Dim.2	-1.6502511	-2.950355	0	0	1.787822	0.0988916

D'après le tableau précédent, la valeur test n'est significatif pour aucune des dimensions. En prenant plus de risque, 15% comme risque de première espèce, on constate que le test est significatif pour la dimension 1 et 2. En allant plus loin dans l'analyse, nous conctatons que les valeurs **v.test** pour les axes factoriels 1 et 2 valent respectivement 1.53 et -1.61. Cela s'interprète par le fait que l'individu de cette quatrième classe est caracterisé par un bon score pour la dimension 1 et un mauvais score pour la dimension 2.

7.3 Analyse des classes par rapport aux vins (variables)

7.3.1 Première classe

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
BORDEAUX	-0.2380028	4067.00	4388.000	4595.355100	3568.3880	0.8118790
CHMPAGNE	-0.6888187	7073.25	8160.250	4055.065806	4175.1652	0.4909374
ALSACE	-1.3080338	326.50	3004.625	200.863760	5417.0255	0.1908618
MOUSSEAU_SIMPLE	-1.4506574	1410.75	3331.500	1458.260501	3503.1199	0.1468753
AUTRE_FORT	-1.4519765	0.00	274.625	0.000000	500.4141	0.1465081
MUSCAT	-1.4832088	346.75	2591.375	338.543479	4003.9673	0.1380189
AOC_FORT	-1.4931787	10.25	219.875	9.601432	371.4328	0.1353904
AUTRE_VDQS	-1.5343299	18429.00	52982.750	14789.528508	59583.4252	0.1249485
RHON	-1.5359736	3720.50	6136.625	4648.078662	4161.8332	0.1245449
TRES_FORT	-1.6079180	25.75	563.125	16.284579	884.2246	0.1078531
ANJOU	-1.6312948	286.50	1752.000	342.660546	2376.8534	0.1028281
BOJOLAIS	-1.6939003	9881.00	23715.625	7526.215284	21608.6970	0.0902842
PROVENCE	-1.8867050	161.50	746.125	167.517910	819.8273	0.0592000
MOUSSEAU_AOC	-1.8888050	387.25	974.250	393.541850	822.2426	0.0589180
RHONE	-2.0004740	306.50	680.375	288.402930	494.4729	0.0454491
VDQS	-2.1511911	429.00	1040.625	380.866775	752.2380	0.0314611
GIRONDE	-2.4023168	5790.25	14922.250	4810.870419	10057.3751	0.0162916
AOC_AUTRES	-2.5693932	1572.25	10284.750	1480.696859	8971.4211	0.0101877

Apres une analyse du tableau ci dessus , nous constatons que le test est significatif (c'est à dire que p.value $< 5\%$) pour les vins **RHONE**, **VDQS**, **GIRONDE** et **AOC_AUTRES**. Pour les autres types de vins, le test est non significatif et cela veut dire que la consommation des pays de ce groupe ne diffèrent pas significativement de celle des autres pour les autres types de vins. La classe sera donc caracterisée

par le niveau de consommation de ces quatre type de vins. Les pays de cette classe sont-ils de grand consommateurs de vins? Voyons voir !

Pour mieux caracteriser cette classe, nous regarderons les valeurs **v.test** pour chaque type de vin. Nous observons donc que pour tous ces quatre types de vins la valeur **v.test** est negative , comprise entre -2,56 et -2. Une valeur négative pour v.test signifie que les individus de cette classe consomment moins en moyenne que les autres. Nous tirons donc la conclusion selon laquelle , le groupe 1 est caracterisé par les **pays à faible consommation de vins de types RHONE, VDGS, GIRONNE et AOC_AUTRES** .

Par ailleurs, on remarque que la valeur **v.test** est negative aussi pour toutes les autres types de vins et cela veut dire que les pays du groupe 1 sont de faible consommateurs de vins en moyenne tout type de vins confondus. En conclusion, le groupe 1 est constitué par les **pays les moins consommateurs de vins**.

7.3.2 Deuxième classe

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
RHONE	1.6561967	1216.5	680.375	431.5	494.4729	0.0976820
AOC_AUTRES	1.6546846	20003.0	10284.750	2803.0	8971.4211	0.0979885
AUTRE_FORT	1.5380898	778.5	274.625	754.5	500.4141	0.1240267
TRES_FORT	1.1770963	1244.5	563.125	1170.5	884.2246	0.2391571
GIRONDE	1.1637510	22584.5	14922.250	401.5	10057.3751	0.2445249
RHON	1.1403220	9243.5	6136.625	1293.5	4161.8332	0.2541522
MOUSSEAU_AOC	0.9971499	1511.0	974.250	925.0	822.2426	0.3186917
PROVENCE	0.9621243	1262.5	746.125	112.5	819.8273	0.3359871
VDQS	0.9379023	1502.5	1040.625	473.5	752.2380	0.3482946
MUSCAT	-0.0493569	2462.0	2591.375	446.0	4003.9673	0.9606348
AOC_FORT	-0.0693988	203.0	219.875	43.0	371.4328	0.9446722
ANJOU	-0.1018627	1593.5	1752.000	993.5	2376.8534	0.9188656
ALSACE	-0.2239321	2210.5	3004.625	211.5	5417.0255	0.8228101
BOJOLAIS	-0.3579138	18652.5	23715.625	1187.5	21608.6970	0.7204078
BORDEAUX	-0.5678369	3061.5	4388.000	722.5	3568.3880	0.5701458
AUTRE_VDQS	-0.6161472	28949.0	52982.750	9798.0	59583.4252	0.5377974
MOUSSEAU_SIMPLE	-0.7210039	1678.0	3331.500	1388.0	3503.1199	0.4709071
CHMPAGNE	-0.9998035	5427.5	8160.250	1641.5	4175.1652	0.3174056

En analysant ce tableau, nous constatons que la valeur test n'est significatif pour aucun type de vins. En prenant un seuil de 10% par exemple, le test devient significatif pour les vins **RHONE** et **AOC_AUTRES**

En outre, pour ces deux types de vins, la valeur **v.test** est strictement positive (1.656 et 1.654). Il s'en suit donc que le groupe 2 est caracterisé par les **pays fortement consommateur de AOC_AUTRES et de RHONE** .

Nous constatons de plus pour les autres types de vins que la valeur **v.test** est positive pour certains et négatif pour d'autres. Cela signifie qu'en moyenne les pays de ce groupe consomme plus certains type de vins (TRES_FORT, GIRONDE, RHON…) au detriment des autres (BOJOLAIS,CHAMPAGNE…) comparativement aux autres pays.

7.3.3 Troisième classe

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
ALSACE	2.6173727	17183	3004.625	0	5417.0255	0.0088610
AUTRE_VDQS	2.3187195	191140	52982.750	0	59583.4252	0.0204103

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
BOJOLAIS	2.2797013	72977	23715.625	0	21608.6970	0.0226254
PROVENCE	2.1563992	2514	746.125	0	819.8273	0.0310525
MOUSSEAU_SIMPLE	2.0289057	10439	3331.500	0	3503.1199	0.0424679
MOUSSEAU_AOC	1.2547999	2006	974.250	0	822.2426	0.2095514
CHMPAGNE	1.0581019	12578	8160.250	0	4175.1652	0.2900090
RHONE	0.6645965	1009	680.375	0	494.4729	0.5063086
AOC_AUTRES	0.6347099	15979	10284.750	0	8971.4211	0.5256176
GIRONDE	0.6065947	21023	14922.250	0	10057.3751	0.5441199
VDQS	0.4059553	1346	1040.625	0	752.2380	0.6847754
RHON	0.3400845	7552	6136.625	0	4161.8332	0.7337929
ANJOU	0.1468328	2101	1752.000	0	2376.8534	0.8832640
BORDEAUX	0.1233050	4828	4388.000	0	3568.3880	0.9018656
AOC_FORT	-0.2285070	135	219.875	0	371.4328	0.8192521
AUTRE_FORT	-0.2290603	160	274.625	0	500.4141	0.8188220
MUSCAT	-0.2653306	1529	2591.375	0	4003.9673	0.7907548
TRES_FORT	-0.4016231	208	563.125	0	884.2246	0.6879614

En analysant ce tableau, nous constatons que la valeur test est significatif pour les vins de type **ALSACE**, **AUTRES_VDQS**, **BOJOLAIS**, **PROVENCE** et **MOUSSEAU_SIMPLE**.

De plus nous observons que la valeur v.test est positive pour tous ses cinq types de vins. Donc cette classe est caractérisée par les **pays les plus grands consonmateurs de vins** particulièrement les cinq types de vins cités ci-dessus. AUTRES_VDQS et BOJOLAIS qui sont les deux types de vins les plus consonsommés font partis des cinq. On comprend pourquoi cette classe est composé uniquement de la **République Fédérale de l'Allemagne** qui est par ailleurs le pays le plus consonmmateurs de vins tous types de vins confondus !

7.3.4 Quatrième classe

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
AOC_FORT	2.5768453	1177	219.875	0	371.4328	0.0099707
MUSCAT	2.5723549	12891	2591.375	0	4003.9673	0.0101009
ANJOU	2.4528227	7582	1752.000	0	2376.8534	0.0141740
VDQS	1.6183376	2258	1040.625	0	752.2380	0.1055899
GIRONDE	1.5016592	30025	14922.250	0	10057.3751	0.1331851
CHMPAGNE	1.2923441	13556	8160.250	0	4175.1652	0.1962380
TRES_FORT	1.2913857	1705	563.125	0	884.2246	0.1965700
MOUSSEAU_SIMPLE	1.1082978	7214	3331.500	0	3503.1199	0.2677332
AOC_AUTRES	1.0833568	20004	10284.750	0	8971.4211	0.2786501
BORDEAUX	0.9799943	7885	4388.000	0	3568.3880	0.3270889
AUTRE_VDQS	0.8076953	101108	52982.750	0	59583.4252	0.4192660
BOJOLAIS	0.7498543	39919	23715.625	0	21608.6970	0.4533424
RHON	0.4890573	8172	6136.625	0	4161.8332	0.6248011
AUTRE_FORT	0.4104101	480	274.625	0	500.4141	0.6815051
MOUSSEAU_AOC	0.2952292	1217	974.250	0	822.2426	0.7678188
RHONE	0.1913654	775	680.375	0	494.4729	0.8482393
ALSACE	-0.3466155	1127	3004.625	0	5417.0255	0.7288802
PROVENCE	-0.5636858	284	746.125	0	819.8273	0.5729680

En analysant ce tableau, nous constatons que la valeur test est significatif pour les vins de type **AOC_FORT**, **MUSCAT** et **ANJOU**.

De plus nous observons que la valeur v.test est positive (mais rélativement faible comparativement aux valeurs pour la troisième classe) pour tous ses trois types de vins. Donc cette classe est caractérisée par les **pays consonmateurs de vins modérés** en général mais grand consommateurs des trois types de vins cités ci-dessus (**AOC_FORT**, **MUSCAT** et **ANJOU**).

8 Recapitulation de la classification (CAH)

Nous recapitulons notre classification ascendente hierarchique dans le tableau suivant :

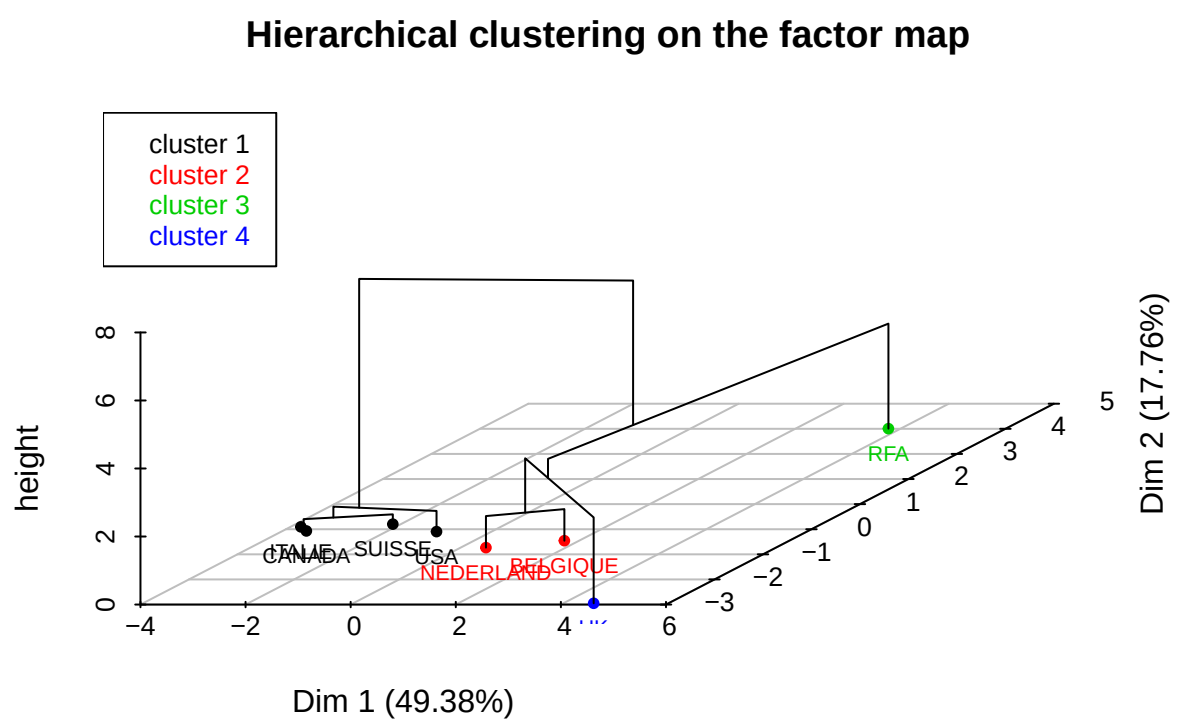
Classe	Caractérisation	Pays membres
1	Faible consommateur de vins	ITALIE, SUISSE, USA et CANADA
2	Gros consommateur de vins ciblés	NEDERLAND et BELGIQUE
3	Gros consommateur de vins	RFA
4	Consommateur de vins modéré	UK

9 Renforcement de la classification

9.1 Algorithme des centres mobiles (k means)

On peut consolider/améliorer les regroupements obtenus via l'algorithme de CAH en utilisant l'algorithme des centres mobiles. On prend alors pour centres initiaux les parangons obtenus lors de la CAH. Il est donc possible que des individus changent de groupes.

```
hcpc.consol = HCPC(data.pca,nb.clust = -1, consol = T,graph = F)
plot(hcpc.consol)
```



Nous obtenons exactement les mêmes classes que la CAH non consolidé. Donc celle-ci est meilleure.

9.2 Méthode des distances

Cette méthode est plus fine que les k-means car elle va regrouper sur le critère de la distance (la plus petite) séparant chaque point.
Nous commençons par calculer les distances euclidiennes:

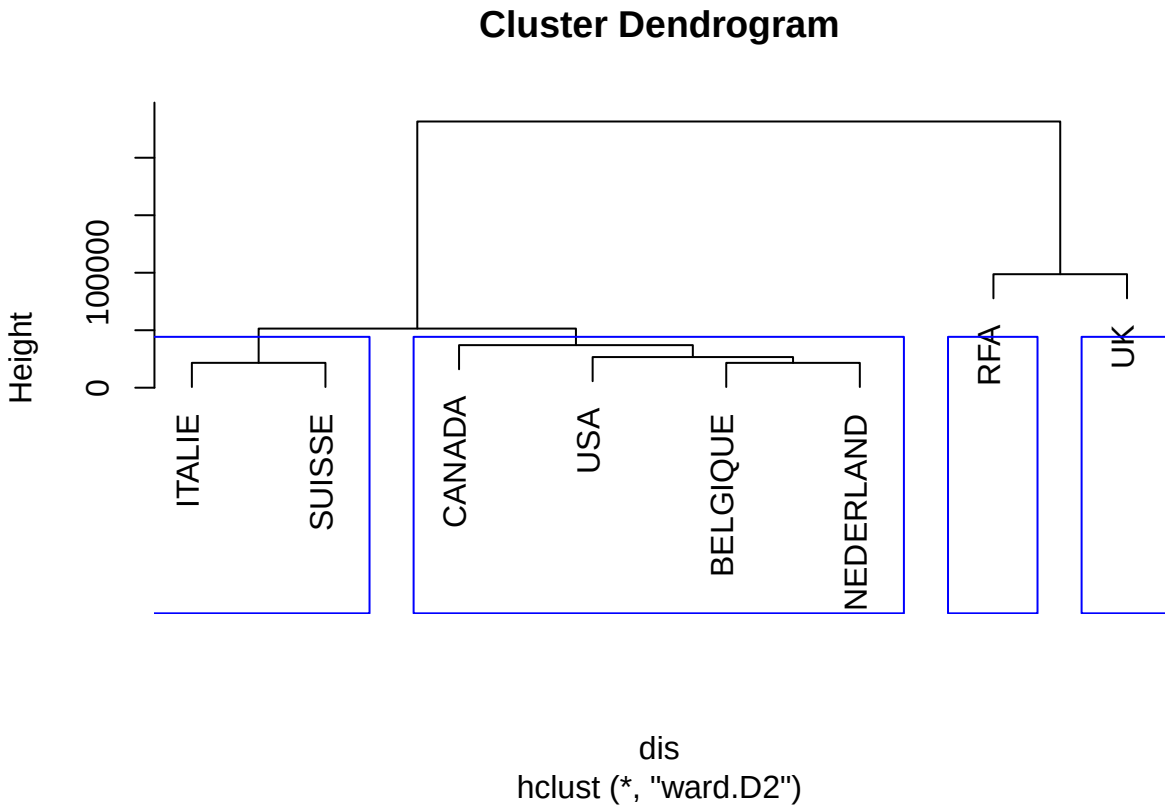
```
dis = dist(vins, method="euclidean")
dis
```

```
           BELGIQUE NEDERLAND       RFA    ITALIE        UK    SUISSE
NEDERLAND  21605.16
RFA       163161.55 181350.84
ITALIE     45915.72  39907.85 199291.35
UK         68381.55  86755.03  98710.48 108650.24
SUISSE     44758.36  33083.60 200222.94  21613.11 108437.62
USA        23814.59  27057.80 175839.88  29992.22  83309.06  29673.91
CANADA     31652.67  40185.34 171539.57  31793.48  83230.77  42790.21
               USA
NEDERLAND
RFA
ITALIE
UK
SUISSE
```

```
USA
CANADA     26720.07
```

Ensuite, nous mettons en place les clusters (en prenant 4 classes)

```
myclust = hclust(dis, method="ward.D2")
plot(myclust)
rect.hclust(myclust,4,border="blue")
```



ou encore avec la fonction eclust :

```
myclust = eclust(dis,k=4)
myclust$cluster
```

```
 BELGIQUE NEDERLAND       RFA    ITALIE        UK    SUISSE       USA
        1          1         3         4         2         4         1
   CANADA
        1
```

La méthode de la distance nous propose quatre classes :
- une classe avec la Belgique, le Nederland, les USA et le Canada,
- une classe avec UK,
- une classe avec RFA,
- et une classe avec l'Italie et la Suisse.

La classification proposée est donc différente de celle proposée par la CAH ou par la méthode des k-means.

10 Conclusion

Nous avons vu que le premier facteur est corrélé positivement, et assez fortement avec presque toutes les variables du jeu de données : plus un pays consomme du vins, plus il a un score élévé sur cet axe.
Le deuxième facteur quant à lui différencie les individus de "taille" semblable.
Réaliser d'abord l'ACP sur notre jeu de données a eu plusieurs avantages dans la réalisation de notre classification.
D'abord, elle a permis de réduire considérablement la dimension de 18 à 3. Sur un petit jeu de données comme celui à notre disposition, la différence est minime mais sur un plus grand avec des milliers de lignes et de colonnes, celà peut être un atout énorme.
Ensuite, nous avons eu une idée des vins ainsi que des pays qui étaient les mieux représentés sur chacune des dimensions. En faisant cela, nous avons su quel était le profil de chaque pays et donc avoir une idée du caractéristique des clusters. À l'issu de l'ACP, le nuage des individus permettait de distinguer intuitivement quatre (4) groupes de pays. Cela a été confirmé par la classification avec la CAH et les autres.
Les individus du groupe 1 (Italie, Canada, Suisse et USA) ne consomment que très peu de vins (comparativement aux autres pays réprésentés dans notre jeu de données). Il suffit de regarder les données du paragon de cette classe (Italie) pour s'en rendre compte.
Les individus du groupe 2 (Belgique, Nederland) sont des consommateurs des vins du rhône et des autres vins AOC.
Le groupe 3 ne contient que RFA, dont les habitants consomment principalement les vins ALSACE, AUTRES_VDQS, BOJOLAIS, PROVENCE et MOUSSEAU_SIMPLE.
Enfin, le groupe 4 est représenté par UK, consommateur des vins AOC_FORT, MUSCAT et ANJOU.
Comme perspective, il serait interesser de mener une analyse comparative des resulats de plusieurs méthodes de classification, la CAH et la methode des distances notomment pour comprendre les differences et surtout choisir la bonne classification selon bien entendu un critère.