Projet LSTAT2110 – Analyse de données

KANA TSIGUIA GABIN, 05581900, DATS2M

Introduction

Le vin, boisson très prisée et jamais (ou quasi) absent à nos tables pendant des appéritifs ou diners, est généralement différent (gout, parfum…) en fonction de la marque, la région où il est fabriqué, les ingrédients etc… L'objet de ce projet, sera de trouver des liens, ou caractéristiques, permettant de regrouper et/ou différentier des composantes des vins, fabriqués dans une région donnée.

pour ce fait, nous disposons d'une base de données '**Wine**'.

Ces données sont le résultat d'une analyse chimique de vins cultivés dans la même région en Italie mais issus de trois cultivateurs différents. L'analyse a déterminé les quantités de 13 constituants retrouvés dans chacun des trois types de vins.

Présentation des données, analyse descriptive

Notre base de données, contient **14** variables…

```
    Alcohol        Malic_Acid         Ash         Ash_Alcanity
 Min.   :11.03   Min.   :0.740   Min.   :1.360   Min.   :10.60
 1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210   1st Qu.:17.20
 Median :13.05   Median :1.865   Median :2.360   Median :19.50
 Mean   :13.00   Mean   :2.336   Mean   :2.367   Mean   :19.49
 3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558   3rd Qu.:21.50
 Max.   :14.83   Max.   :5.800   Max.   :3.230   Max.   :30.00
   Magnesium      Total_Phenols     Flavanoids    Nonflavanoid_Phenols
 Min.   : 70.00   Min.   :0.980   Min.   :0.340   Min.   :0.1300
 1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205   1st Qu.:0.2700
 Median : 98.00   Median :2.355   Median :2.135   Median :0.3400
 Mean   : 99.74   Mean   :2.295   Mean   :2.029   Mean   :0.3619
 3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875   3rd Qu.:0.4375
 Max.   :162.00   Max.   :3.880   Max.   :5.080   Max.   :0.6600
 Proanthocyanins Color_Intensity       Hue             OD280
 Min.   :0.410   Min.   :  1.00   Min.   :0.4800   Min.   :1.270
 1st Qu.:1.250   1st Qu.: 32.25   1st Qu.:0.7825   1st Qu.:1.938
 Median :1.555   Median : 67.00   Median :0.9650   Median :2.780
 Mean   :1.591   Mean   : 64.59   Mean   :0.9574   Mean   :2.612
 3rd Qu.:1.950   3rd Qu.: 93.75   3rd Qu.:1.1200   3rd Qu.:3.170
 Max.   :3.580   Max.   :132.00   Max.   :1.7100   Max.   :4.000
    Proline       Customer_Segment
 Min.   : 278.0   Min.   :1.000
 1st Qu.: 500.5   1st Qu.:1.000
 Median : 673.5   Median :2.000
 Mean   : 746.9   Mean   :1.938
 3rd Qu.: 985.0   3rd Qu.:3.000
 Max.   :1680.0   Max.   :3.000
```

…Dont 13 continues(Numeriques) qui sont les constituants chimiques dans les vins, et une catégorielle(factor) qui représente les 3 régions d'Italie concernées par l'étude, dont on s'en privera pour pour l'analyse à composante principale. On l'utilisera plus tard pour l'analyse factorielle.

	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium
Alcohol	1.00	0.09	0.21	-0.31	0.27
Malic_Acid	0.09	1.00	0.16	0.29	-0.05
Ash	0.21	0.16	1.00	0.44	0.29
Ash_Alcanity	-0.31	0.29	0.44	1.00	-0.08
Magnesium	0.27	-0.05	0.29	-0.08	1.00
Total_Phenols	0.29	-0.34	0.13	-0.32	0.21
Flavanoids	0.24	-0.41	0.12	-0.35	0.20
Nonflavanoid_Phenols	-0.16	0.29	0.19	0.36	-0.26
Proanthocyanins	0.14	-0.22	0.01	-0.20	0.24
Color_Intensity	0.57	0.17	0.24	-0.14	0.22
Hue	-0.07	-0.56	-0.07	-0.27	0.06
OD280	0.07	-0.37	0.00	-0.28	0.07
Proline	0.64	-0.19	0.22	-0.44	0.39

	Total_Phenols	Flavanoids	Nonflavanoid_Phenols
Alcohol	0.29	0.24	-0.16
Malic_Acid	-0.34	-0.41	0.29
Ash	0.13	0.12	0.19
Ash_Alcanity	-0.32	-0.35	0.36
Magnesium	0.21	0.20	-0.26
Total_Phenols	1.00	0.86	-0.45
Flavanoids	0.86	1.00	-0.54
Nonflavanoid_Phenols	-0.45	-0.54	1.00
Proanthocyanins	0.61	0.65	-0.37
Color_Intensity	0.09	0.05	-0.01
Hue	0.43	0.54	-0.26
OD280	0.70	0.79	-0.50
Proline	0.50	0.49	-0.31

	Proanthocyanins	Color_Intensity	Hue	OD280	Proline
Alcohol	0.14	0.57	-0.07	0.07	0.64
Malic_Acid	-0.22	0.17	-0.56	-0.37	-0.19
Ash	0.01	0.24	-0.07	0.00	0.22
Ash_Alcanity	-0.20	-0.14	-0.27	-0.28	-0.44
Magnesium	0.24	0.22	0.06	0.07	0.39
Total_Phenols	0.61	0.09	0.43	0.70	0.50
Flavanoids	0.65	0.05	0.54	0.79	0.49
Nonflavanoid_Phenols	-0.37	-0.01	-0.26	-0.50	-0.31
Proanthocyanins	1.00	-0.06	0.30	0.52	0.33
Color_Intensity	-0.06	1.00	-0.28	-0.20	0.46
Hue	0.30	-0.28	1.00	0.57	0.24
OD280	0.52		-0.20	0.57	1.00	0.31
Proline	0.33		0.46	0.24	0.31	1.00

fig1. matrice des corrélations

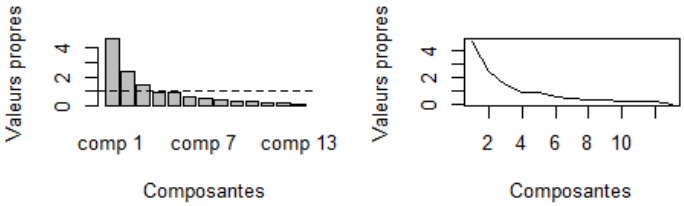
Analyse en composantes principales

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.68582787	36.0448298	36.04483
comp 2	2.42321480	18.6401139	54.68494
comp 3	1.46364026	11.2587712	65.94371
comp 4	0.91644116	7.0495474	72.99326
comp 5	0.86889966	6.6838435	79.67711
comp 6	0.58694378	4.5149522	84.19206
comp 7	0.51710036	3.9776950	88.16975
comp 8	0.37600483	2.8923448	91.06210
comp 9	0.32913117	2.5317782	93.59388
comp 10	0.28798926	2.2153020	95.80918
comp 11	0.22972201	1.7670923	97.57627
comp 12	0.21695698	1.6688998	99.24517
comp 13	0.09812788	0.7548298	100.00000

tab1. Valeurs propres

les valeurs propres superieures à 1 seront prises comme composantes principales.

le graphique suivant nous permet de fixer le nombre de composantes principales à 3; avec **54.68%** de l'information contenue dans le plan formé par les axes 1 et 2, et **65.9%** cumulée contenue dans les 3 axes (plan 1&2, 1&3 et 2&3).



Analyse et interpretation des variables:

coordonées factorielle des variables (coord) et qualité de représentation (cos2)

	Dim.1	Dim.2	Dim.3	Dim.4
Alcohol	0.37052838	0.742505883	-0.2192888	0.009765177
Malic_Acid	-0.50948384	0.412366524	0.1159302	0.565809867
Ash	0.02443169	0.494504785	0.7578815	-0.213037090
Ash_Alcanity	-0.52511228	-0.007889379	0.7308699	0.030906546
Magnesium	0.33553523	0.436259267	0.1900287	-0.217259945
Total_Phenols	0.86130947	0.018394825	0.1893507	0.159037856
Flavanoids	0.91459227	-0.069445436	0.1820531	0.116395866
Nonflavanoid_Phenols	-0.64360265	0.091619247	0.1937000	-0.252712437
Proanthocyanins	0.68098694	-0.072264065	0.2230336	0.373406841
Color_Intensity	0.09552089	0.809040927	-0.2054255	-0.036452563
Hue	0.60720792	-0.449243133	0.0721881	-0.443410009
OD280	0.79080151	-0.303170796	0.1893064	0.154316906
Proline	0.66373622	0.547611112	-0.1339346	-0.196903219

	Dim.5
Alcohol	0.18978982
Malic_Acid	0.03226788
Ash	0.11233350
Ash_Alcanity	-0.01770792
Magnesium	-0.72993746
Total_Phenols	0.16181613
Flavanoids	0.13459747
Nonflavanoid_Phenols	0.40911385
Proanthocyanins	-0.10415431
Color_Intensity	0.15243320
Hue	0.12118552
OD280	0.13933771
Proline	0.07897583

tab2. coordonnées factorielles

Variables les mal représentées sur les axes 1 et 2: **Ash, Ash_Alcanity, Magnesium**.

```
                Ash         Ash_Alcanity            Magnesium
              0.245                0.276                0.303
Nonflavanoid_Phenols          Malic_Acid      Proanthocyanins
              0.423                0.430                0.469
                Hue      Color_Intensity              Alcohol
              0.571                0.664                0.689
              OD280              Proline        Total_Phenols
              0.717                0.740                0.742
         Flavanoids
              0.841
```

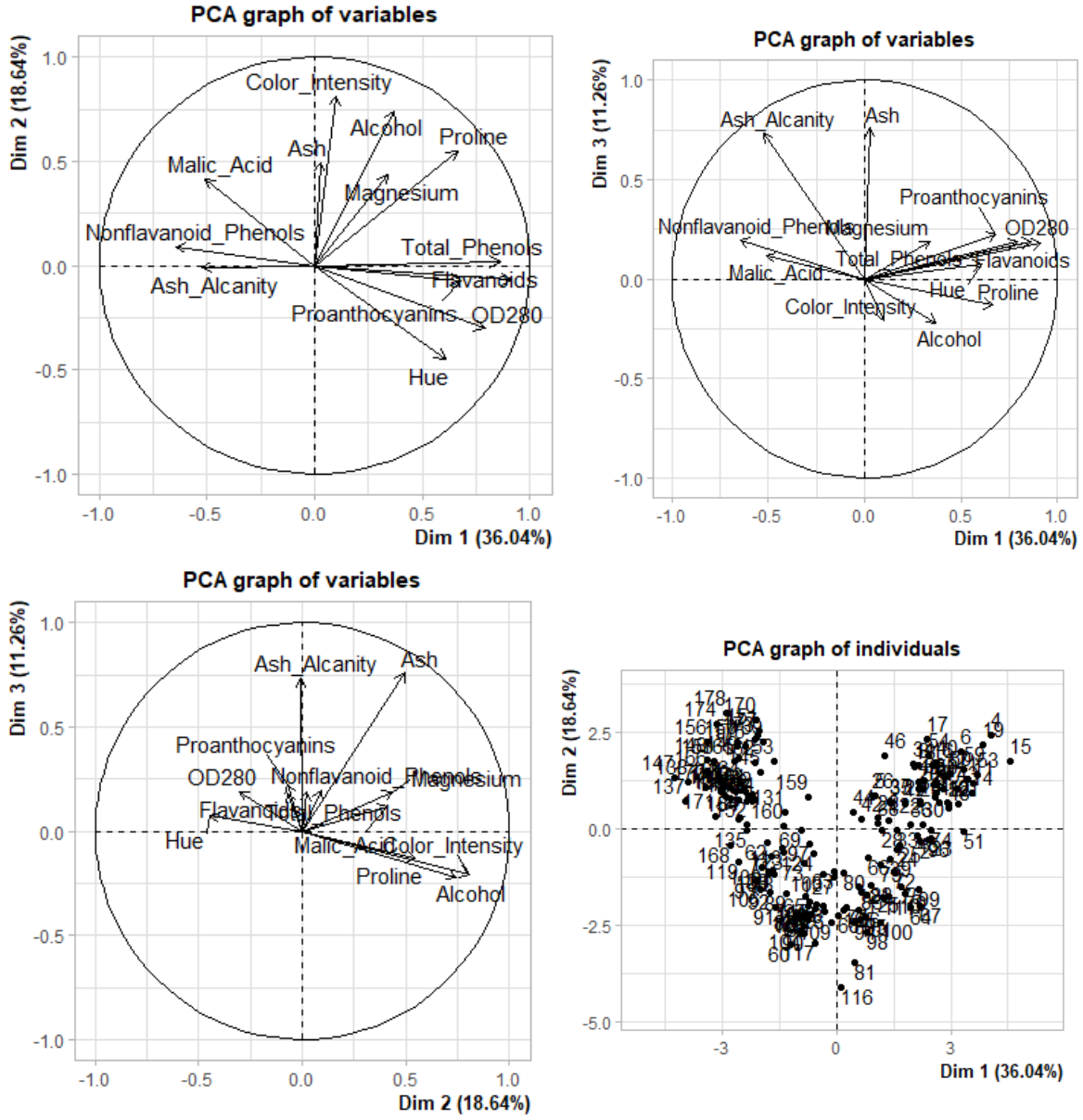
tab3. qualité de représentation 1-2

Variables les mal représentées sur les axes 2 et 3: **Total_Phenols , Flavanoids, Nonflavanoid_Phenols**.

```
       Total_Phenols           Flavanoids Nonflavanoid_Phenols
               0.036                0.038                0.046
     Proanthocyanins                OD280           Malic_Acid
               0.055                0.128                0.183
                 Hue            Magnesium              Proline
               0.207                0.226                0.318
        Ash_Alcanity              Alcohol      Color_Intensity
               0.534                0.599                0.697
                 Ash
               0.819
```

tab4. qualité de représentation 2-3

Cercle des correlations sur les axes 1 et 2 et individus



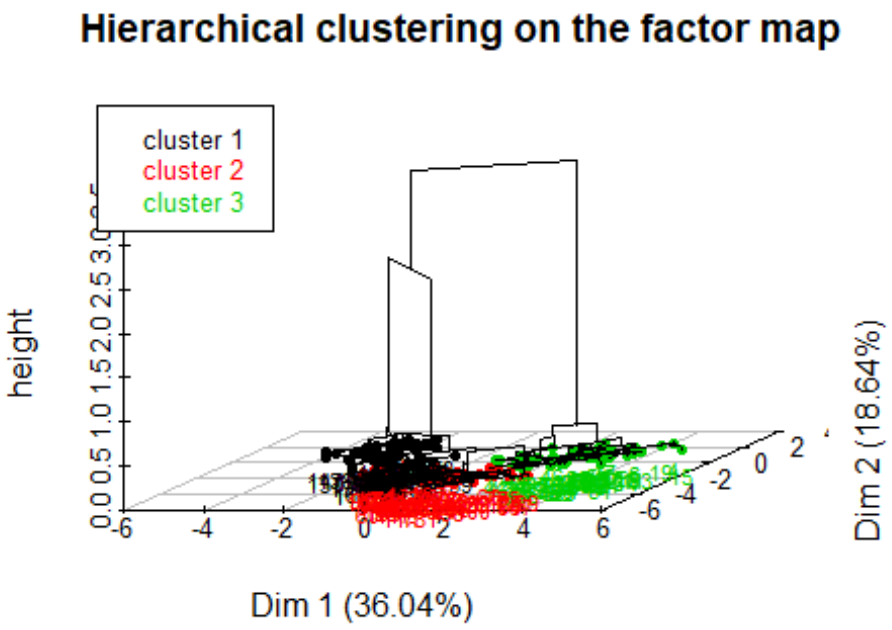
D'après le cercle des corrélations, on observe que les variables mieux représentées dans le premier plan sont: **Flavanoids, Total_Phenols, Proline, OD280, Alcohol,Color_Intensity**.

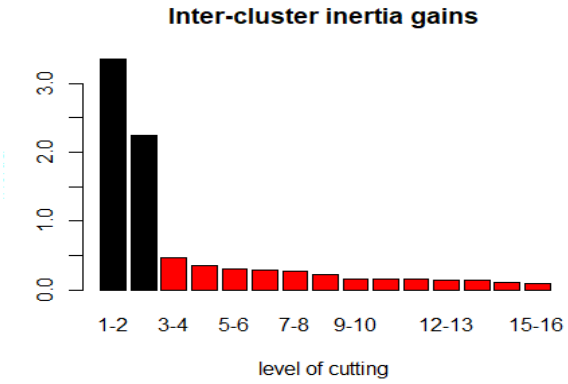
Sur le graphe des individus, on remarque que les axez 1 et 2 ont tendance à séparer les individus en 4 groupes:

- un premier groupe avec un fort taux de Malic_Acid, Nonflavanoid_Phenols
- un second groupe avec un fort taux de Alcohol et de Proline, qui, avec le premier groupe, ont une forte intensité de couleur

- un troisième et quatrième groupe avec moins de Malic_Acid,Nonflavanoid_Phenols,Alcohol et de Proline, qui pourraient etre regroupé en un meme groupe avec faible intensité de couleur. Une analyse discriminante ou un clustering pourrait confirmer cette division.

Clustering





le Deundogramme en 3D nous montre un une repartition des individus en 3 groupes, comme vu lors de l'ACP.

Par exemple les individus **19, 116** et **137**: L'individu 19 possède un taux de **14% d'alcool** et une **intensité de couleur de 127** qui correspondrait au groupe 2 de notre ACP, et à la région 1 de nos données.

L'individu 116 possède **11% d'alcool** pour une intensité de couleur de **3**, correspondant au groupe 3 de l'ACP, et à la région 2 de nos données.

L'individu **137** possède **12.27% d'alcool** pour un intensité de couleur de **50**.

Analyse des correspondances

L'analyse factorielle (AFC), étant une partie de l'ACP, permet d'analyser l'impact d'une variable categorielle sur notre jeu de donnée. Notre jeu de données conntient la variable **Customer_Segment**, qui représente les 3 regions concernées de l'étude.

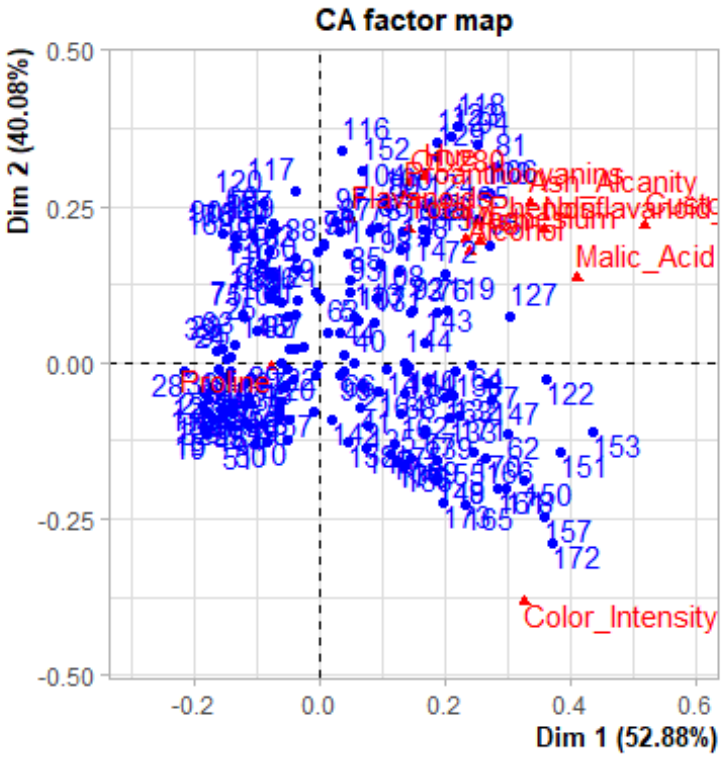
```
resCA <- CA(MyData, ncp = 5, graph = FALSE)
round(sort(rowSums(resCA$col$cos2[,1:2])), digits = 3)
```

```
          Flavanoids      Proanthocyanins           Malic_Acid
               0.216                0.432                0.437
      Total_Phenols Nonflavanoid_Phenols              OD280
               0.495                0.547                0.586
                 Hue     Customer_Segment         Ash_Alcanity
               0.626                0.773                0.879
                 Ash              Alcohol            Magnesium
               0.886                0.900                0.964
             Proline      Color_Intensity
               0.999                0.999
```

```
round(resCA$col$contrib[,c(1,2)], digits = 3)
```

```
                     Dim 1  Dim 2
Alcohol               3.558  2.588
Malic_Acid            1.876  0.270
Ash                   0.739  0.560
Ash_Alcanity         10.497  7.996
Magnesium            25.705 24.355
Total_Phenols         0.224  0.648
Flavanoids            0.024  0.666
Nonflavanoid_Phenols  0.223  0.102
Proanthocyanins       0.136  0.717
Color_Intensity      32.812 59.330
Hue                   0.127  0.525
OD280                 0.260  1.359
Proline              21.339  0.298
Customer_Segment      2.482  0.586
```

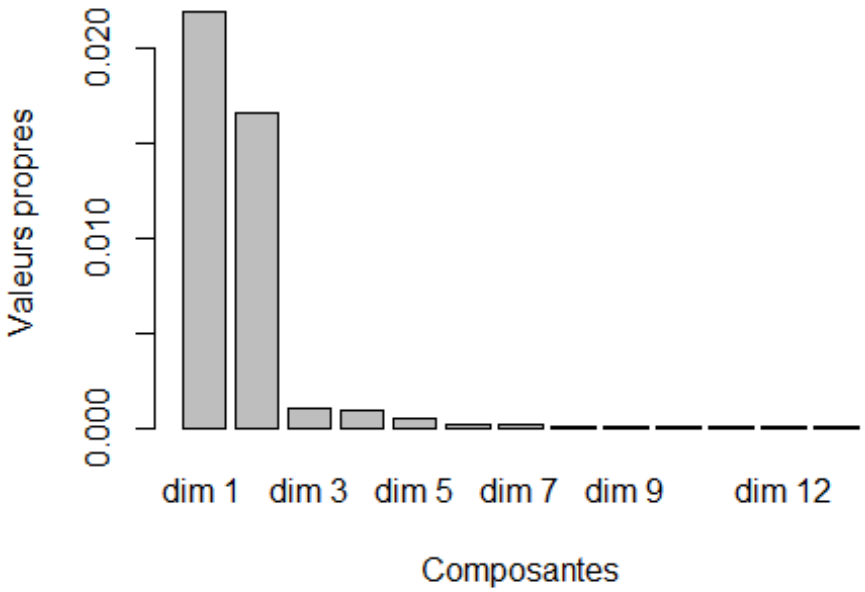
```
plot(resCA)
```




```
head(resCA$eig)

        eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.0219513471             52.8773846                           52.87738
dim 2 0.0166387832             40.0802434                           92.95763
dim 3 0.0010675398              2.5715374                           95.52917
dim 4 0.0008853465              2.1326621                           97.66183
dim 5 0.0004518049              1.0883277                           98.75016
dim 6 0.0001682946              0.4053956                           99.15555

barplot(resCA$eig[,"eigenvalue"], xlab = "Composantes", ylab="Valeurs propres")
abline(h = 1, lty = "dashed")
```



Notre analyse factorielle, assez similaire à l'ACP, pourrait se limiter à une analyse sur les dimensions 1 et 2, qui contiennent **93%** de l'information. Car, malgré l'inclusion de la variable discrète **Customer_Segment**, le tableau des variables mieux représentées est quasi le meme que celui de l'ACP, avec les variables mieux representées telles que: **Color_Intensity, Proline, Magnesium, Alcohol,Ash**

Conclusions

En conclusion, l'AFC tout comme l'ACP, nous permet de conclure que les regions concernées peuvent etre distinguées en fonction de:

- **leur intesité de couleur:** qui en moyenne est plus faible pour la region 2, mais plus forte en taux de proline, mais plus forte pour les region 1 et 3
- **Intensité de couleur:** Comme dit au point precédent, est plus forte en reion 1 et 3, et moins forte en region 2
- **Magnesium, Alcool,Ash:** Pour ne citer que celles-là parmis les plus importantes, permet de distinguer les region 1 et 3, qui sont moins riche en proline que la region 2.

Annexes