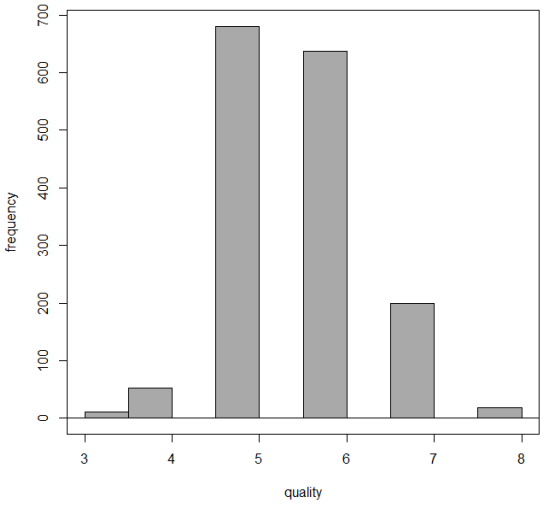
PRESENTATION DU JEU DE DONNEES

Notre jeu de données provient du site *https://archive.ics.uci.edu/ml/datasets/Wine+Quality*. Ses auteurs sont P. Cortez, A. Cerdeira, F. Almeida, T. Matos et J. Reis. Il traite d'un type de vin rouge portugais, le Vinho Verde. C'est un vin intermédiaire concernant le taux d'alcool, et il est apprécié pour sa fraîcheur. Il contient 12 colonnes, la dernière étant la variable à expliquer. Les colonnes 1 à 11 sont des variables actives continues, elles correspondent aux composantes des différents vins et la colonne 12 correspond à la note hédonique, une variable quantitative représentant la médiane de 3 notes venant de 3 experts différents, données pour un seul vin. Cette note est anonyme, donc aucun test sur l'effet juge dans l'effet global n'est réalisable, cependant le terme expert signifie que ceux-ci sont intransigeants et représentent aussi l'avis des consommateurs.

Les caractéristiques physico-chimiques (variables) sont les suivantes :

Acidité fixe : acidité naturelle du raisin (g/L)
Acidité volatile : acidité exogène due aux fermentations alcooliques microbiennes ou à l'extraction de l'acide acétique du bois du fût (g/L)
Acide citrique : présent dans la plupart des raisins, les vins rouges en sont souvent dépourvus (g/L)
Sucre résiduel : quantité de sucre restant après fermentations alcooliques (g/L)
Chlorure : quantité de sel dans le vin (g/L)
Dioxyde de soufre libre : La forme libre de SO2 libre existe en équilibre avec le SO2 sous forme de gaz dissous et les ions disulfures (mg/L)
Dioxyde de soufre total : Le taux de forme libres et liées de SO2 (mg/L)
Densité : dépend du pourcentage d'alcool et de la teneur en sucre (g/cm3)
Ph : potentiel hydrique (s.u.)
Sulfates : additif qui contribue au niveau de SO2 ce qui agit en antimicrobien et antioxydant (g/L)
Alcool (% volume)

Il y a 1599 lignes, autant d'individus, autant de vins. Une telle quantité de vins différents de même appellation (Vinho Verde) est explicable par le fait que c'est un vignoble morcelé à l'extrême où, d'une façon générale, chaque producteur ne possède guère plus d'un hectare. Ajoutons que ce vignoble s'étend sur 15% du territoire portugais. Il est donc normal d'avoir un échantillon de vin élevé, tous différents, mais avec une désignation commune. En 2009, l'année de la récolte de ces données on comptabilisait 27 662 producteurs. Au vu de l'histogramme (nombre de vins repartis en fonction de la note) la qualité des vins suit une courbe de Gauss. Ce qui est normal au vu de la quantité de vins représentatifs du vignoble, beaucoup de vins sont moyens, peu sont excellents ou médiocres. Avec une moyenne de 5.6 et un écart type de 0.8 celui-ci admet une faible dispersion autour de la moyenne. Les vins en grande majorité oscillent donc de 0.8 autour de 5.6.

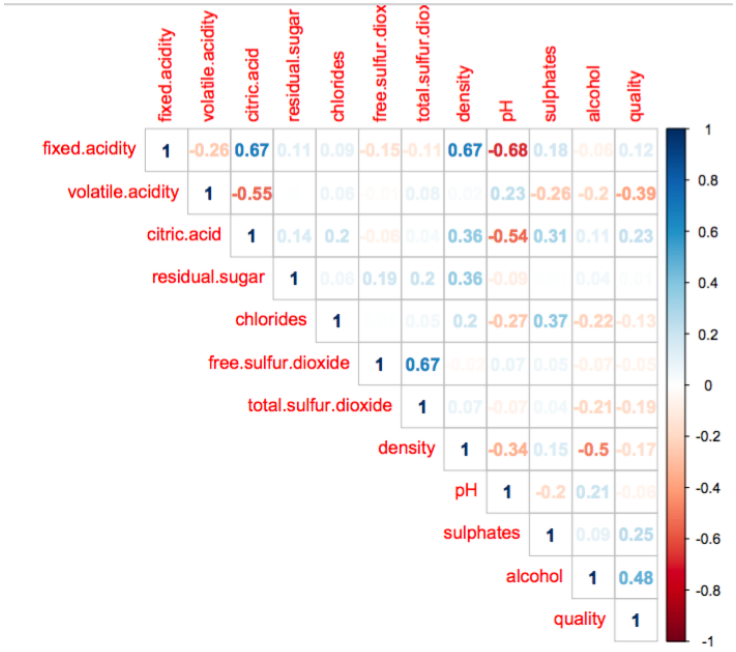


PROBLEMATIQUE

Comment expliquer l'appréciation d'un vin rouge à partir de ses propriétés physico-chimiques ?

ÉTUDE DES CORRELATIONS LINEAIRES ENTRE LES VARIABLES

Par cette étude nous cherchons les liens entre les variables caractérisant nos vins, en particulier quelles sont les variables prises seules à seules corrélées avec la qualité. Cette étude nous donne un a priori du résultat de la régression linéaire et va nous aider à l'analyser: les variables très corrélées entre elles vont voir le coefficient évoluer car une partie de l'une est expliquée par l'autre, ou bien disparaître.
Sur ce corrélogramme nous pouvons voir le taux de corrélation entre chacune des variables, les plus corrélées ayant leur taux en rouge et les moins corrélées en bleu.



On constate que le pH est fortement corrélé négativement avec l'acidité fixe (-0,68), ce qui signifie que le pH baisse quand l'acidité fixe augmente (logique car un pH bas est un pH acide). Le pH est aussi corrélé négativement avec l'acide citrique (-0,54, même remarque). L'acide citrique est corrélé positivement avec l'acidité fixe (0,67), l'acidité fixe étant l'acidité naturelle du raisin et l'acide citrique étant un acide présent dans le raisin lors de sa maturation. L'acide citrique est corrélé négativement avec l'acidité volatile (-0,55) qui est l'acidité exogène, on peut penser que cela est dû au fait que des bactéries transforment l'acide citrique en acide acétique (dosé pour déterminé l'acidité volatile). Les variables concernant l'acidité du vin sont donc corrélées entre elles. On peut s'attendre que le meilleur modèle pour déterminer la qualité d'un vin n'en sélectionne qu'un ou deux.

La densité est corrélée positivement avec l'acidité fixe (0,67) et modérément avec l'alcool (0,5).En effet on sait que la mesure du taux d'alcool se fait à partir de la densité du liquide (Loi définie par Louis Joseph Gay-Lussac en 1824 pour les besoins de la taxation des alcools)

Le taux de $S0_2$ libre et le taux de SO_2 total sont corrélés positivement(0,67), le premier étant compris dans le second.

Les sulfates n'ayant aucune corrélation forte avec les autres, on peut s'attendre à les retrouver dans le modèle car ses informations ne sont pas apportées par les autres.

Enfin, la variable la plus corrélée avec la qualité est l'alcool (0,48), plus le taux d'alcool est fort plus le vin est apprécié. On a aussi l'acidité volatile qui est corrélée négativement avec la qualité (-0,39).

On peut donc penser que l'appréciation d'un vin ne dépend pas que de quelques variables mais plus d'une combinaison de variables, ou bien que les variables recueillies pour l'étude ne sont pas celles qui influent sur la qualité.

NB : Nous parlons ici de corrélations linéaires, des taux proches de 0 n'implique pas une absence de corrélation autre que linéaire (exponentielles, en puissance...)

REGRESSION LINEAIRE MULTIPLE ET MEILLEUR MODELE

La régression linéaire multiple nous donne la relation de cause à effet entre nos variables explicatives, les paramètres physico-chimiques, et la variable à expliquer, la qualité. Ce qui nous intéresse ici est de déterminer le meilleur modèle expliquant la qualité, donc de savoir quels sont les paramètres les plus importants pour prédire l'appréciation d'un vin rouge. Le principe est de faire la régression multiple sur le modèle complet, puis d'enlever une à une les variables influant le moins tant que les autres sont dans le modèle (coefficient en valeur absolue le plus petit). Par soucis d'efficacité nous avons utilisé la fonction reg.best du package FactoMineR.

```
res = RegBest(y=VIN[,12],x=VIN[,-12])
res$best
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.4300987  0.4029168  10.995  < 2e-16 ***
volatile.acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
free.sulfur.dioxide   0.0050774  0.0021255   2.389    0.017 *
total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6477 on 1591 degrees of freedom
Multiple R-squared:  0.3595, Adjusted R-squared:  0.3567
F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

Les variables ayant été conservée dans le modèle sont l'acidité volatile, les chlorures, le SO_2 libre, le SO_2 total, le pH, les sulfates et l'alcool. Nous sommes passés de 11 variables explicatives à 7. Les coefficients des variables acidité volatile, chlorures, SO_2 total et pH sont négatifs, ils font donc baisser la qualité quand ils augmentent (voir corrélogramme en annexe pour leurs liens seuls avec la qualité). Au contraire, le taux de SO_2 libre, les sulfates et le taux d'alcool ont un effet positif sur l'appréciation des vins.

On constate que le taux de SO_2 libre pris seul est corrélé positivement avec la qualité (voir corrélogramme) tandis qu'avec les autres variables dans la régression il joue en défaveur de la qualité. Cela signifie qu'une autre variable corrélé avec lui faisait augmenter la qualité (brouillage), mais qu'à lui seul il a une influence négative.
Un pH généralement bon pour un vin rouge se situe entre 3,3 (acidulé) et 3,9 (douceâtre). Le Vinho Verde étant apprécié pour sa fraîcheur, son pH doit être relativement bas dans cette échelle ce que nous confirme la régression.
Le SO_2 (et les sulfates) peut engendrer un goût de soufre même à faible dose d'où la corrélation négative.
L'acide acétique est le principal acide volatile, en faible quantité son arôme est bénéfique mais pas à forte dose.
Selon cette fonction, ces sept variables sont déterminantes sur la qualité d'un vin rouge. Il doit être plutôt acide, riche en alcool, en sulfates et en SO_2 libre mais pauvre en acidité volatile (acidité exogène), en chlorures, et en SO_2 total.
Cependant, le modèle ne prend pas en compte les sucres résiduels connus pour donner de l'onctuosité aux vins et donc jouer sur la qualité.

ACP

OBJECTIFS

L'ACP permet de décrire un jeu de données, de le résumer, d'en réduire la dimensionnalité.
L'ACP réalisée sur les individus du tableau de données répond à différentes questions :

1. Etude des individus (des vins) : deux vins sont proches s'ils ont des caractéristiques similaires. On s'intéresse à la variabilité entre individus. Y a-t-il des similarités entre les individus pour toutes les variables ? Peut-on établir des profils de vins ? Peut-on opposer un groupe d'individus à un autre ?
2. Etude des variables (des propriétés physicochimiques) : on étudie les liaisons linéaires entre les variables. Les objectifs sont de résumer la matrice des corrélations et de chercher des variables synthétiques: peut-on résumer les propriétés des vins par un petit nombre de variables ?
3. Lien entre les deux études : peut-on caractériser des groupes d'individus par des variables ?

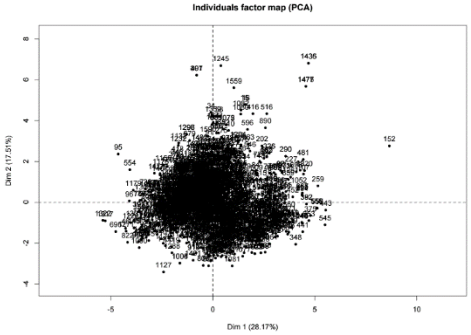
On étudie les profils de vins uniquement en fonction de leur propriété physicochimique. Les variables actives ne seront donc que celles qui concernent les propriétés physicochimiques. L'autre variable (*"Note hédonique"*) n'appartient pas aux profils de vins, elle n'est pas prise en compte dans la dimension, et utilisent une information déjà donnée par les autres variables mais il est intéressant de la confronter aux composantes principales. L'on n'attend pas à ce que ce soit corrélé, si ça l'est tant mieux. Nous l'utiliserons comme variable illustrative.

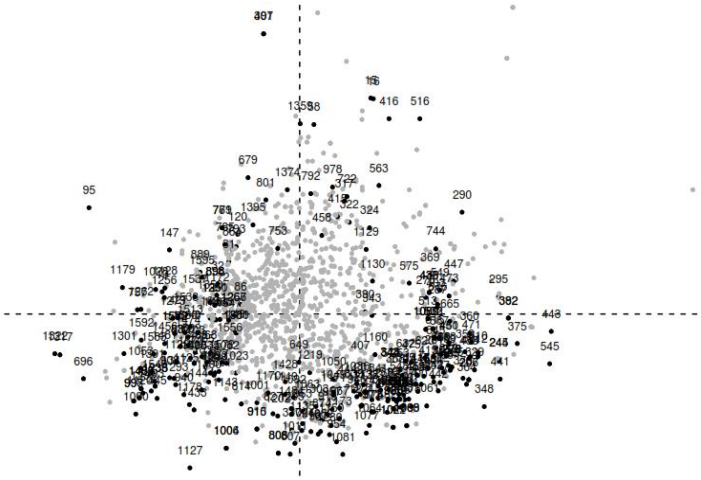
Dans le tableau de données, les variables ne sont pas mesurées dans les mêmes unités On va alors normer . On doit les réduire de façon à donner la même influence à chacune R le fera automatiquement si l'on utilise le booléen avec scale.unit.

```
res.pca = PCA(VIN[,1:11], scale.unit=TRUE, ncp=11, graph=T)
#VIN: le tableau de données utilisé
#scale.unit: pour choisir de réduire ou non les variables, ici oui
#ncp: le nombre de dimensions à garder dans les résultats
#graph: pour choisir de faire apparaître les graphiques ou non
```

Au vu du nuage de point, il est nécessaire de le clarifier

```
x11()
plot(res.pca)
plot(res.pca,cex=0.5,select="cos2 0.7",unselect"grey70")
#x11(): Ouvrir une nouvelle fenêtre graphique
#plot(res.pca): Y mettre le graphique
#res.pca: le résultat d'une PCA
#cex: Modifier la police d'écriture
#select: Ici sélectionner les individus dont la qualité de représentation évalué par cos² est supérieur à 0.7
#unselect: colorier les points qui ne remplissent pas la condition précédente en gris
```





Les deux premières dimensions contiennent 45,68 % de l'inertie totale (l'inertie est la variance totale du tableau de données, la somme des variances associés à chacune des variables). La première dimension vaut 28.17 % et la seconde vaut 17.51 %.Pour 1599 individus et 12 variables cette représentation reste relativement bonne.
La variable *"residual sugar"* est négativement corrélée à la variable *"alcohol"*. Quand un vin possède un taux de sucre résiduel élevé, le taux d'alcool est faible étant donné que les levures n'ont pas réalisé la fermentation alcoolique en utilisant la totalité du sucre.Il n'y a pas de corrélation entre la variable « total disulfure » et « sulphates ». Il faut faire attention ici une grande valeur pour la variable « *pH* » correspond à un liquide basique : plus un vin a un pH faible, plus il est acide, plus il se retrouve à droite dans le cercle des corrélations.
Le premier axe oppose les vins qui sont acides comme le *375* à ceux qui sont basiques comme le 95.

La fonction *dimdesc()* calcule le coefficient de corrélation entre une variable et une dimension et réalise un test de significativité.
Ces tableaux donnent le coefficient de corrélation et la probabilité critique des variables qui sont significativement corrélées aux dimensions principales. Les variables actives dont la probabilité critique est inférieure à 0.05 apparaissent.

```
dimdesc(res.pca, axes=c(1,2))
#res.pca: le résultat d'une PCA
#axes: les axes choisi
```

```
$Dim.1
$Dim.1$quanti
                     correlation       p.value
fixed.acidity         0.86140602  0.000000e+00
citric.acid           0.81619353  0.000000e+00
density               0.69599339 5.427610e-232
sulphates             0.42764729  4.238754e-72
chlorides             0.37364637  3.799188e-54
residual.sugar        0.25721219  1.410737e-25
quality               0.11034930  9.749578e-06
free.sulfur.dioxide  -0.06365298  1.089917e-02
alcohol              -0.19933772  8.578273e-16
volatile.acidity     -0.42001233  2.348725e-69
pH                   -0.77198543 1.383551e-316
```

```
$Dim.2
$Dim.2$quanti
                     correlation       p.value
total.sulfur.dioxide  0.79031776  0.000000e+00
free.sulfur.dioxide   0.71271338 3.229371e-248
volatile.acidity      0.38154068  1.443575e-56
residual.sugar        0.37758520  2.401559e-55
density               0.32414940  1.948327e-40
chlorides             0.20546173  1.057370e-16
sulphates            -0.05211625  3.717973e-02
fixed.acidity        -0.15335255  7.098289e-10
citric.acid          -0.21065171  1.701655e-17
quality              -0.38680247  3.226863e-58
alcohol              -0.53593092 1.427117e-119
```

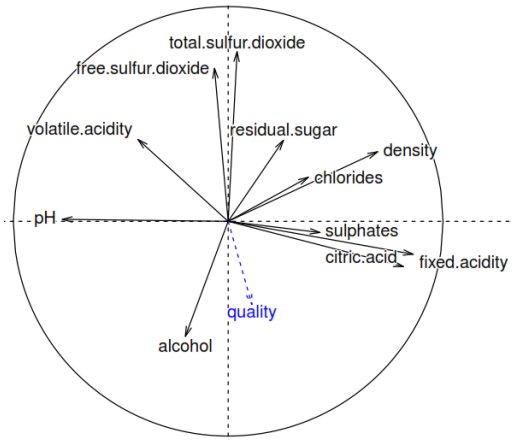
Les variables qui contribuent à la construction de cet axe sont le pH, l'acide citrique et l'acidité fixe. Cette dimension est donc à interpréter comme un gradient d'acidité.
Le deuxième axe oppose les vins avec un taux de dioxyde de soufre élevé et les autres. Les variables l'expliquant sont l'alcool, le dioxyde de soufre total et libre.
Les variables *"sulphates"* et *"chlorides"* ne sont pas très bien représentées dans ce plan 2D.
A l'issue de cette première approche, on peut diviser le premier plan factoriel en trois parties : les vins puissants en alcool, les vins acides, et les vins avec un pH plus élevé.
Les individus contribuant à la représentation du premier axe sont "*152*" et "*1322*", avec respectivement "*1.511*" et "*0.586*", pour le second "*1436*" et "*1127*", respectivement "*1.506*" et "*0.379*"

VARIABLES ILLUSTRATIVES

Les variables illustratives n'influencent pas la construction des composantes principales de l'analyse. Elles aident à l'interprétation des dimensions de variabilité. On peut ajouter deux types de variables : continues et qualitatives. On ajoute la variable *"Note"* comme une variable continue illustrative.

```
res.pca = PCA(VIN, scale.unit=TRUE, ncp=11, quanti.sup=12,
graph=T)

#VIN: le tableau de données utilisé
#scale.unit: pour choisir de réduire ou non les variables
#ncp: le nombre de dimensions à garder dans les résultats
#quanti.sup: vecteur des index des variables continues
illustratives
#graph: pour choisir de faire apparaître les graphiques ou non
```



Dans notre jeu de données la variable explicative est très mal représentée dans le plan. Elle servait à savoir si la structure induite se présentait comme une variable active. Comme la variable Qualité résulte d'une note attribué au vin par des juges, on peut émettre l'hypothèse que les vins extrêmes, c'est à dire immondes ou excellent font preuve d'un consensus dans la note mais que pour le reste, les vins intermédiaires, tous les goûts sont dans la Nature.

CONCLUSION

L'étude statistique des caractéristiques physico-chimiques des vins associées à une note hédonique nous a permis une meilleur compréhension de ce qui était apprécié dans ce type de vin et même d'avoir une équation reliant le profil du vin et sa note hédonique moyenne.

Les vins Vinho Verde étaient présentés comme étant appréciés pour leur fraîcheur, cela s'explique par un pH intermédiaire bas par rapport aux autres vins rouges. Les variables liées à l'acidité que sont le pH, l'acide citrique, l'acidité volatile et l'acidité fixe sont très liées, et vont donner une grande variabilité entre les différents vins rouges et donc des saveurs différentes. Une autre grande source de variabilité entre les vins rouges est le taux de dioxyde de soufre qui peut engendrer un gout désagréable en grande quantité. Enfin, la seule caractéristique physico-chimique très corrélée avec la qualité est taux d'alcool. Un bon Vinho Verde est donc un vin alcoolisé, acide, et pauvre en dioxyde de soufre. Les autres caractéristiques vont engendrer des profils variables, mais selon les personnes, différents profils pourront être appréciés. Il n'existe donc pas de recette miracle pour un grand vin, tous les goûts sont dans la nature, et heureusement. Certains paramètres dépendent intrinsèquement du raisin et donc ne sont pas à 100% contrôlables lors de la production, de plus cela apporte de la variété.

ETUDE D'UN VIN PORTUGAIS

AGRO CAMPUS OUEST

CORNANGUER LENAIG & SEROUART MARIO