Analyse exploratoire de données

Fabrice Rossi

Télécom ParisTech

Février 2010

Analyse d'un exemple

- Qualité de vins rouges du Portugal (Vinho Verde)
- 1599 vins
- 12 variables :
  - 11 mesures chimiques (pH, alcool, etc.) : variables numériques
  - une note de 0 à 10 : variable ordinale
- Plan d'analyse :
  - exploration univariée
  - exploration multivariée
  - classification (*clustering*)
- Données de issues de

  P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
  Modeling wine preferences by data mining from physicochemical properties.
  In Decision Support Systems, Elsevier, 47(4) :547-553. ISSN : 0167-9236.

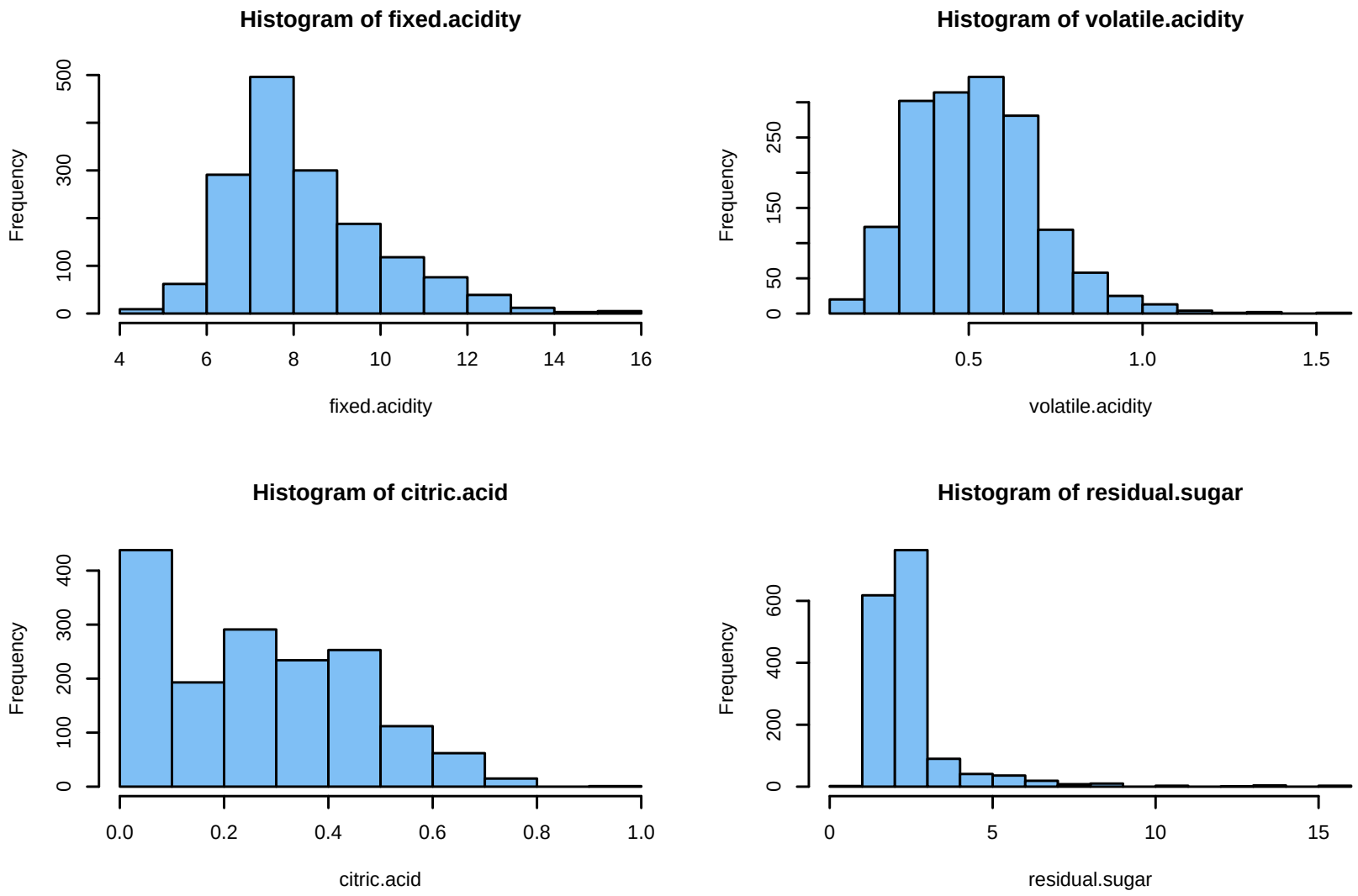
Plan

Analyses univariées

Analyses multivariées

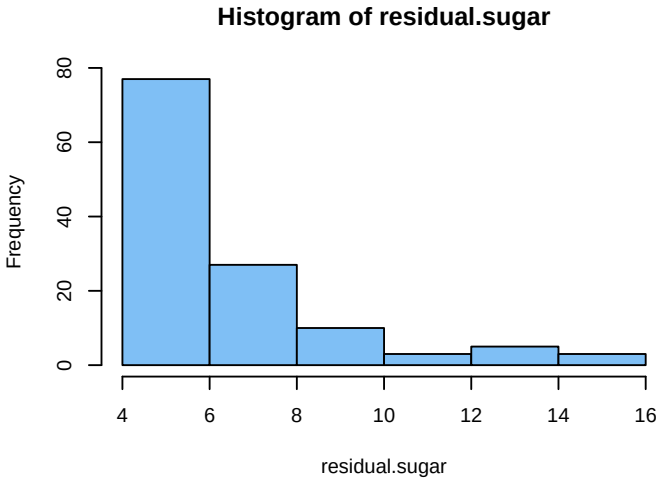
Classification

Histogrammes



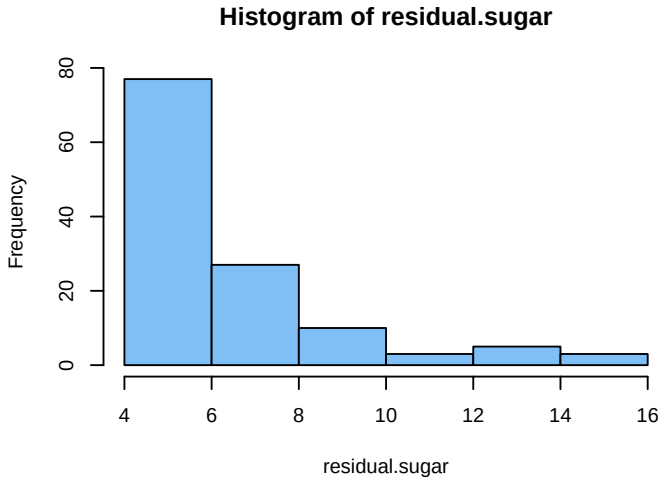
Analyses associées

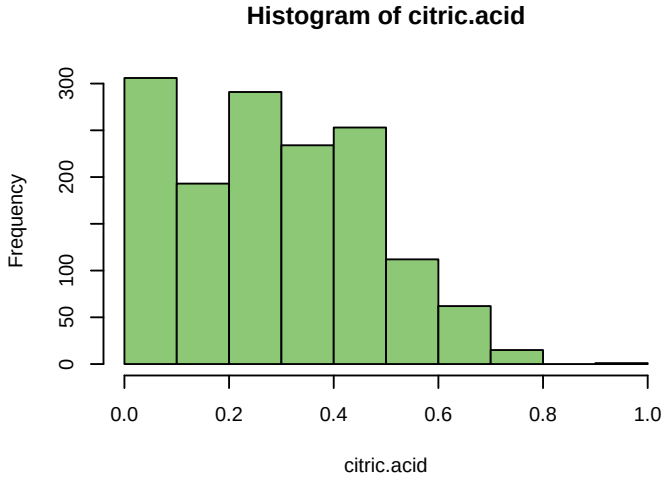
- Sucre résiduel :
  - 125 vins (7.8 %) avec plus de 4g/l (limite légale pour le vin sec)
  - 8 vins avec plus de 12g/l (passage de demi sec à moelleux)



Analyses associées

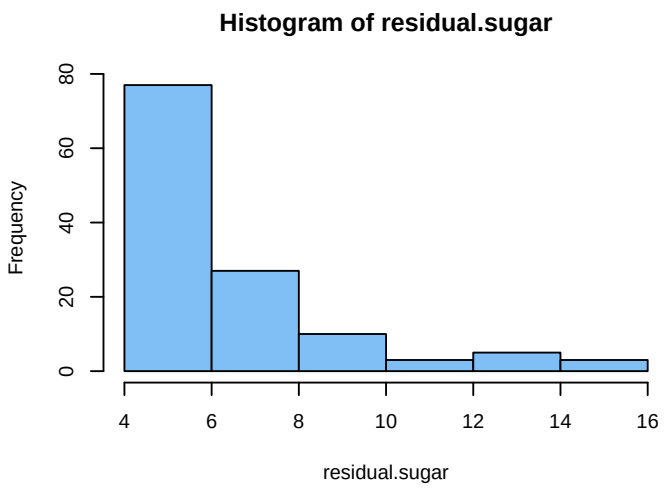
- Sucre résiduel :
  - 125 vins (7.8 %) avec plus de 4g/l (limite légale pour le vin sec)
  - 8 vins avec plus de 12g/l (passage de demi sec à moelleux)
- Acide citrique :
  - 132 vins sans acide citrique
  - un seul vin avec plus de 0.79 g/l

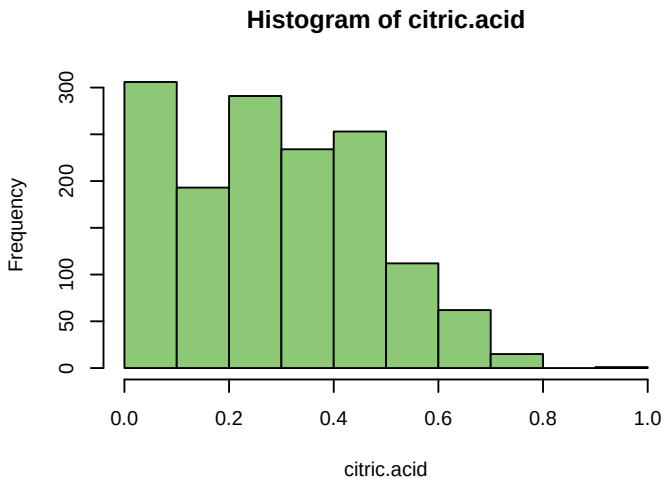




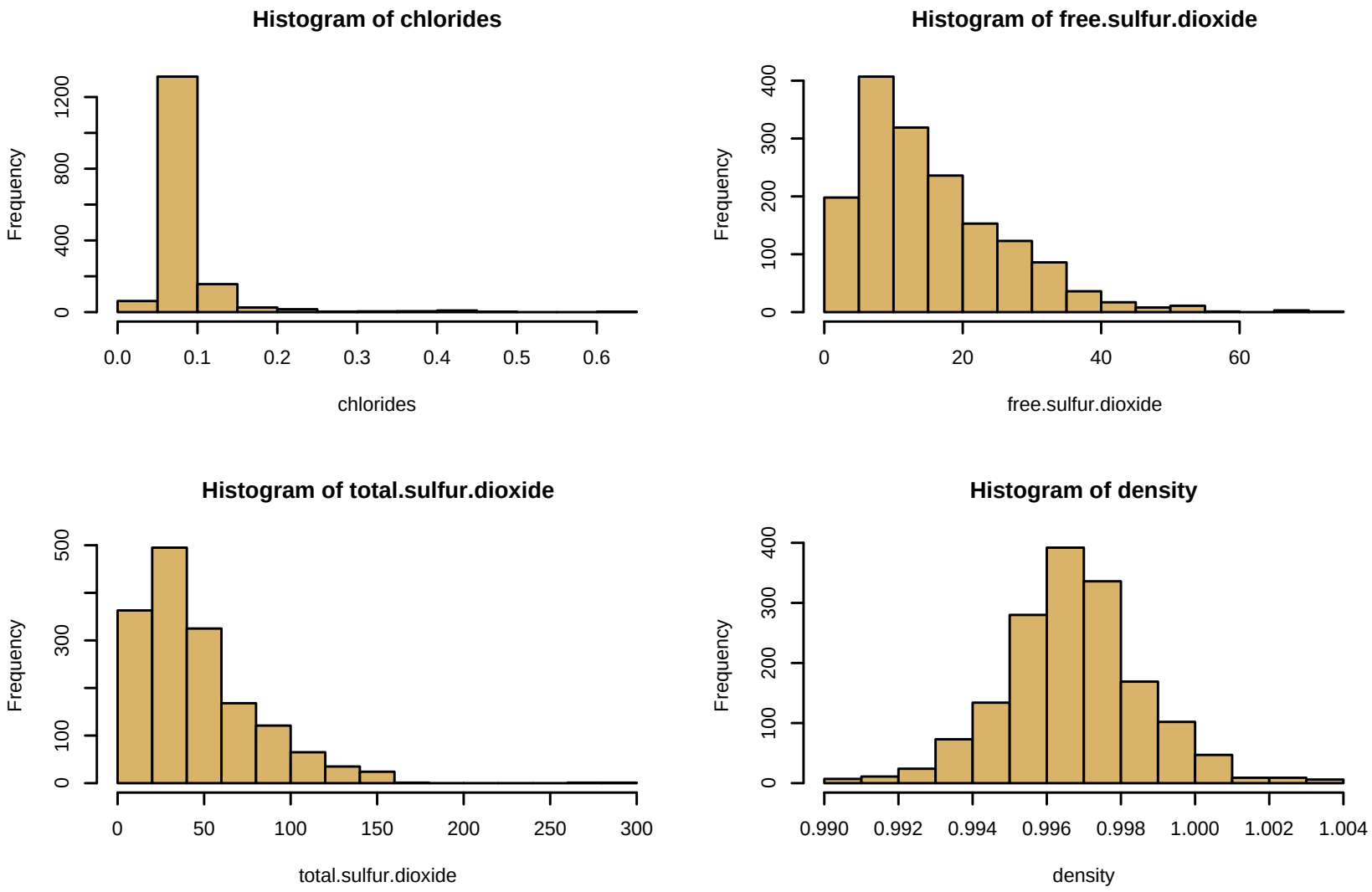
Analyses associées

- Sucre résiduel :
  - 125 vins (7.8 %) avec plus de 4g/l (limite légale pour le vin sec)
  - 8 vins avec plus de 12g/l (passage de demi sec à moelleux)
- Acide citrique :
  - 132 vins sans acide citrique
  - un seul vin avec plus de 0.79 g/l
- Acide acétique :
  - 4 vins au dessus de 1.2 g/l
  - 223 vins au dessus de 0.7 g/l (seuil de perception)



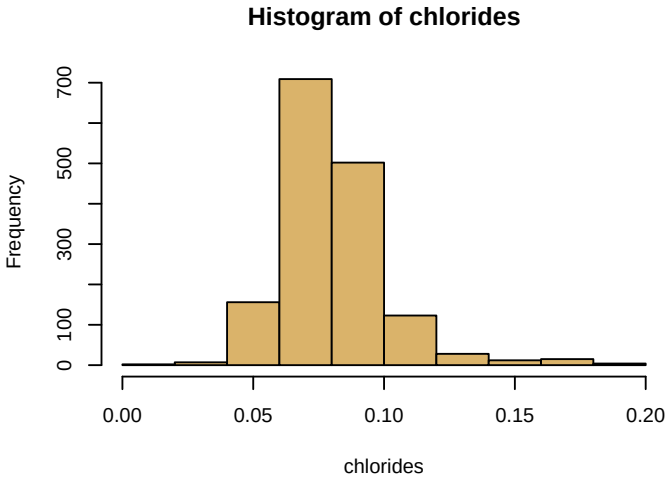


Histogrammes



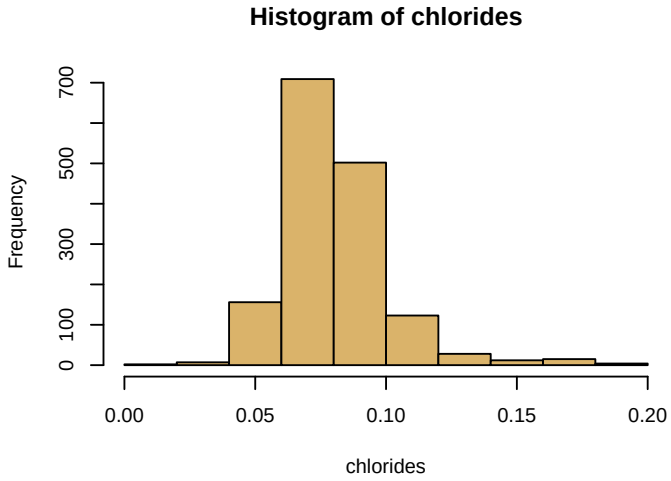
Analyses associées

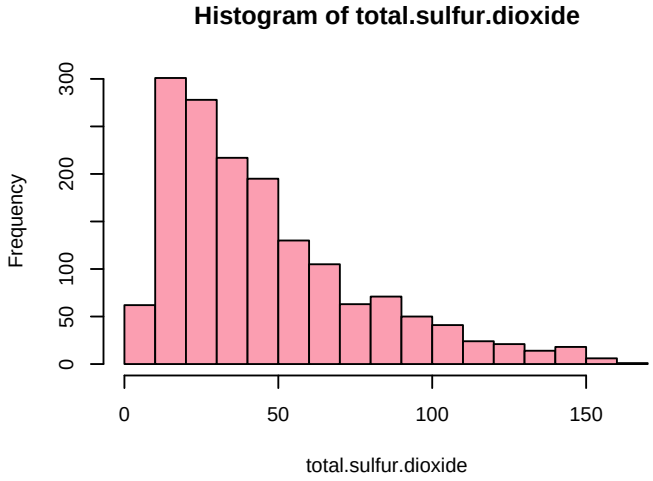
- Sel :
  - 41 vins avec plus de 0.2 g/l
  - $\mu = 0.0847$ $m = 0.079$
  - sans les *outliers* : $\mu = 0.081$ $m = 0.079$



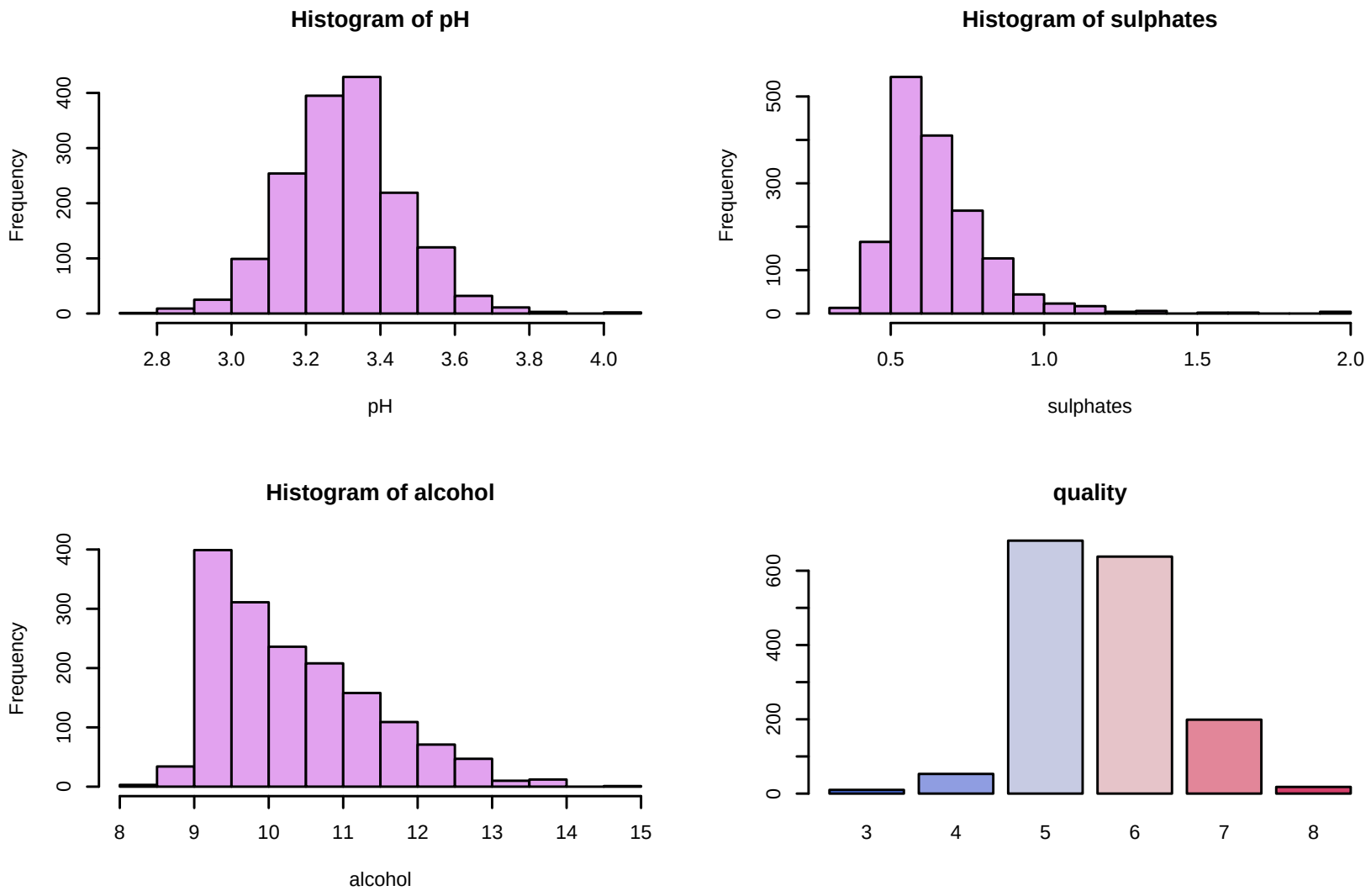
Analyses associées

- Sel :
  - 41 vins avec plus de 0.2 g/l
  - $\mu = 0.0847$ $m = 0.079$
  - sans les *outliers* : $\mu = 0.081$ $m = 0.079$
- SO_2 :
  - deux *outliers* : 289 et 278 mg/l
  - $\mu = 46.17$ $m = 38$





Histogrammes



Plan

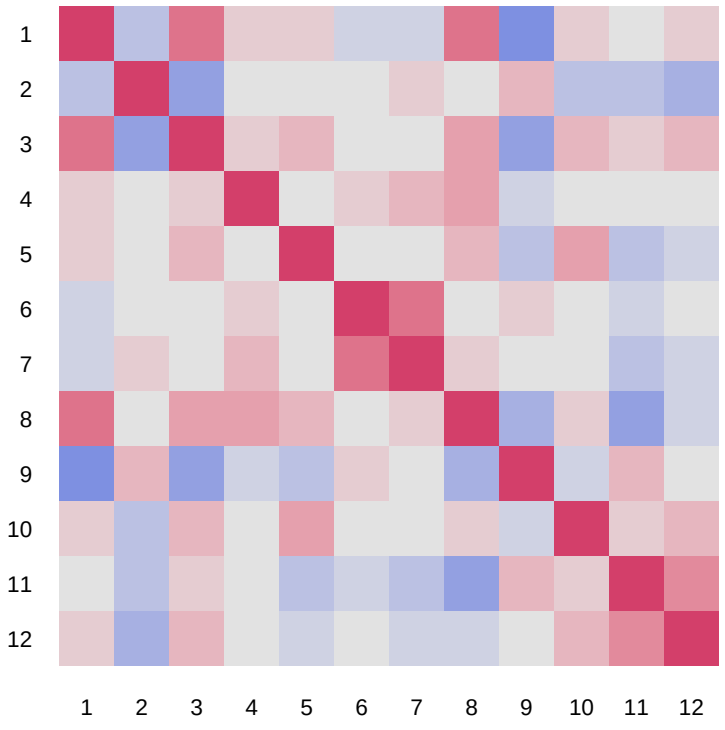
Analyses univariées

Analyses multivariées

Classification

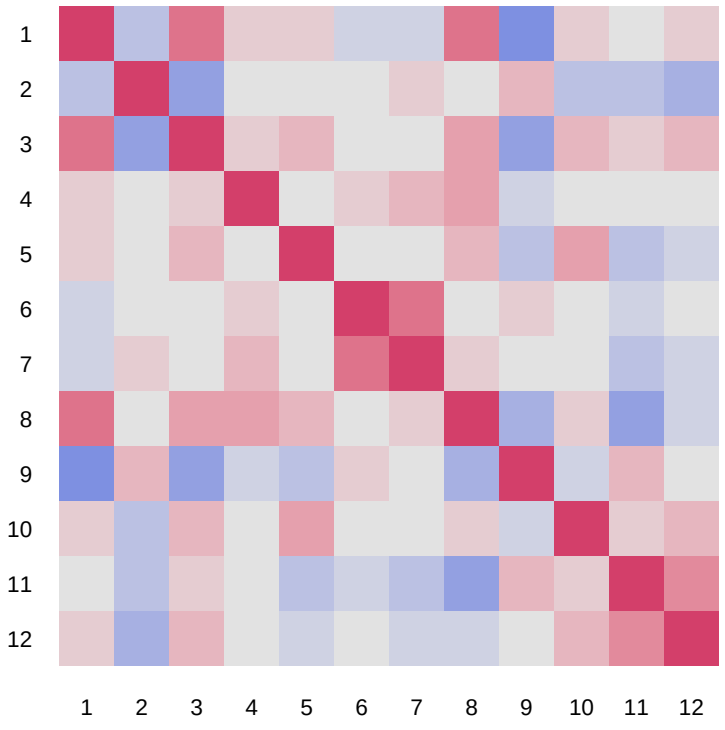
Corrélations

- Recherche de corrélations
- Représentation graphique de la matrice de corrélation :
  - rouge : forte corrélation positive
  - bleu : forte corrélation négative

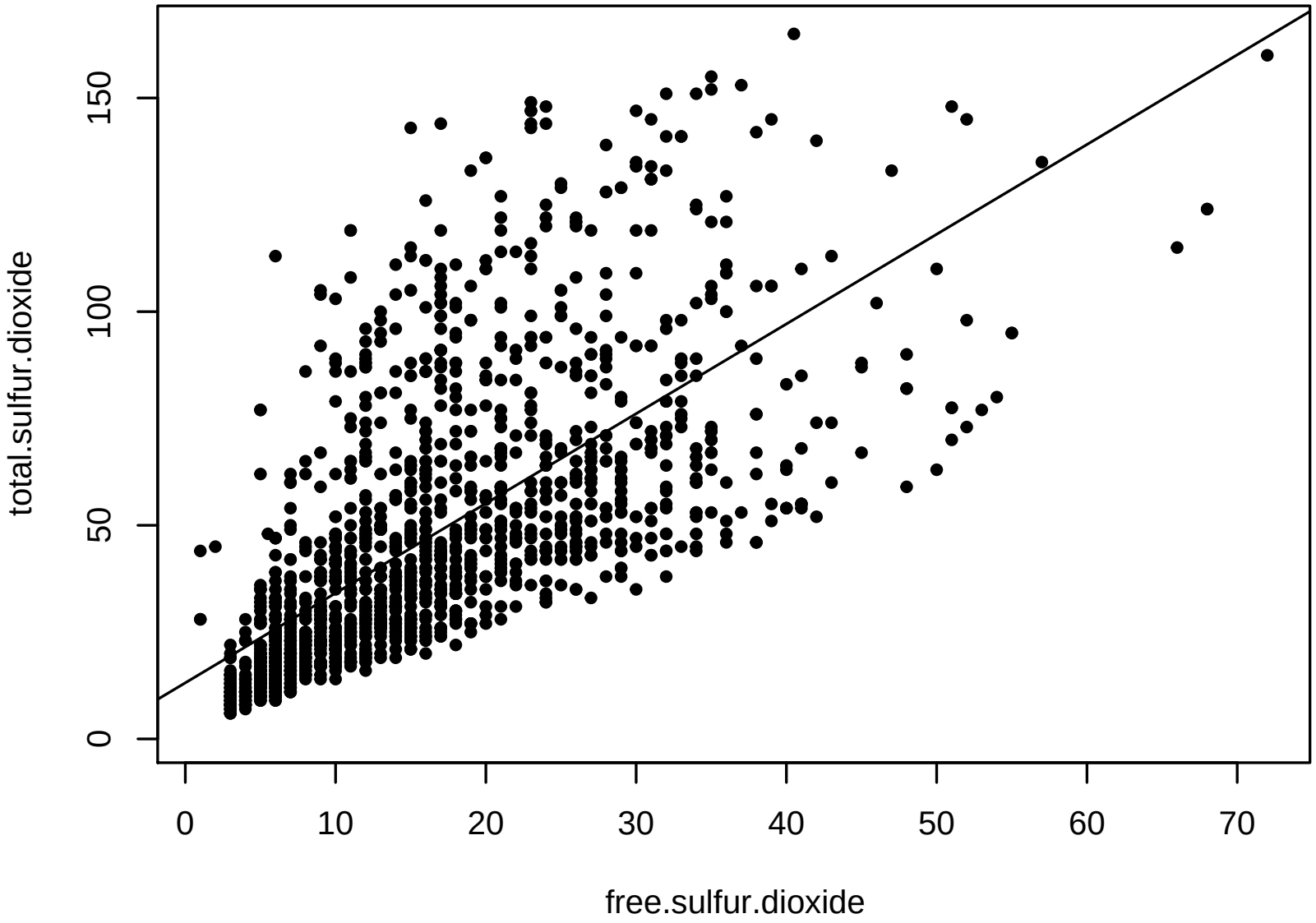


Corrélations

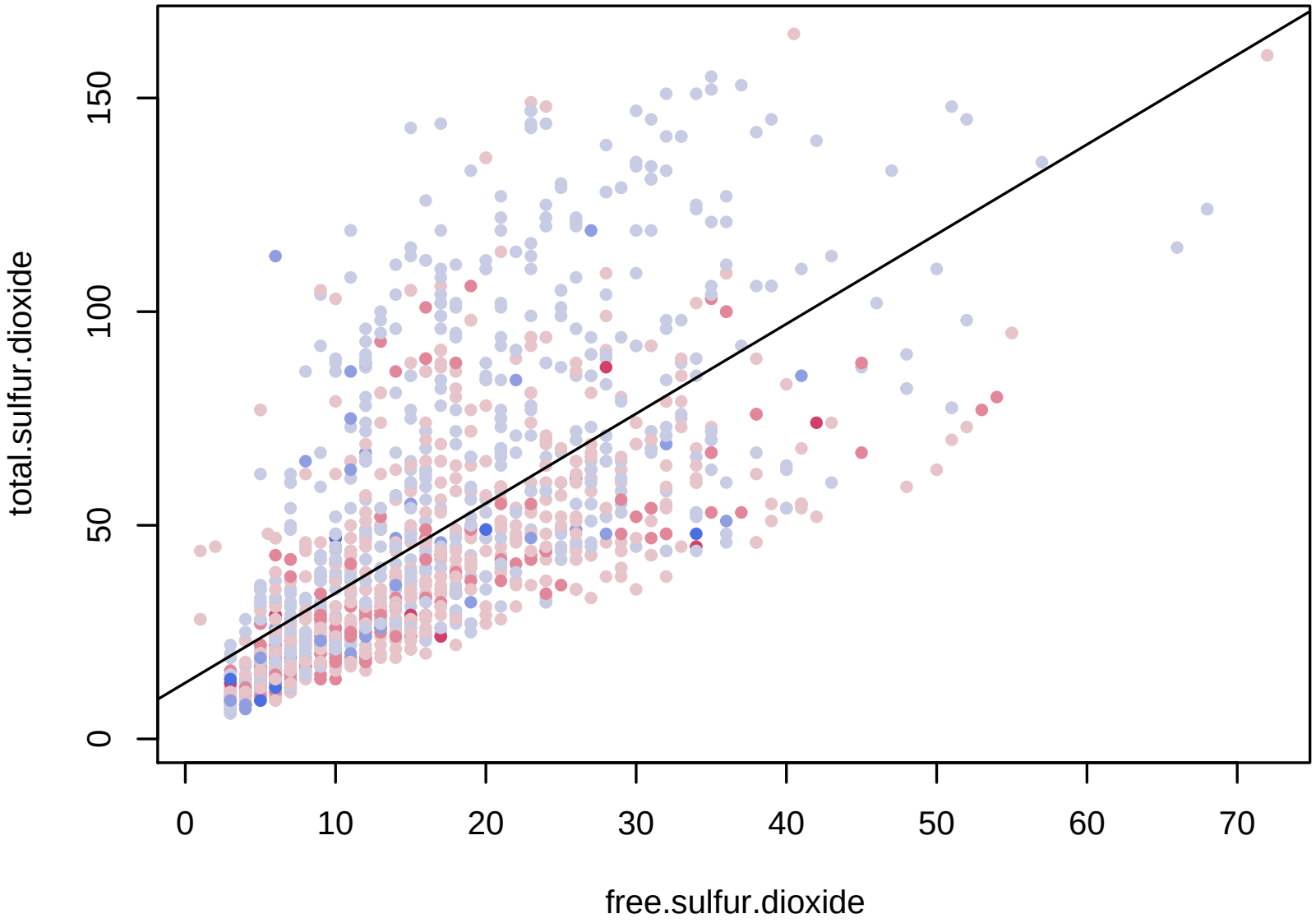
- Recherche de corrélations
- Représentation graphique de la matrice de corrélation :
  - rouge : forte corrélation positive
  - bleu : forte corrélation négative
- Ici :
  - alcool corrélé avec la qualité
  - acide acétique anti-corrélé avec la qualité
  - redondance sur le souffre et sur les acidités



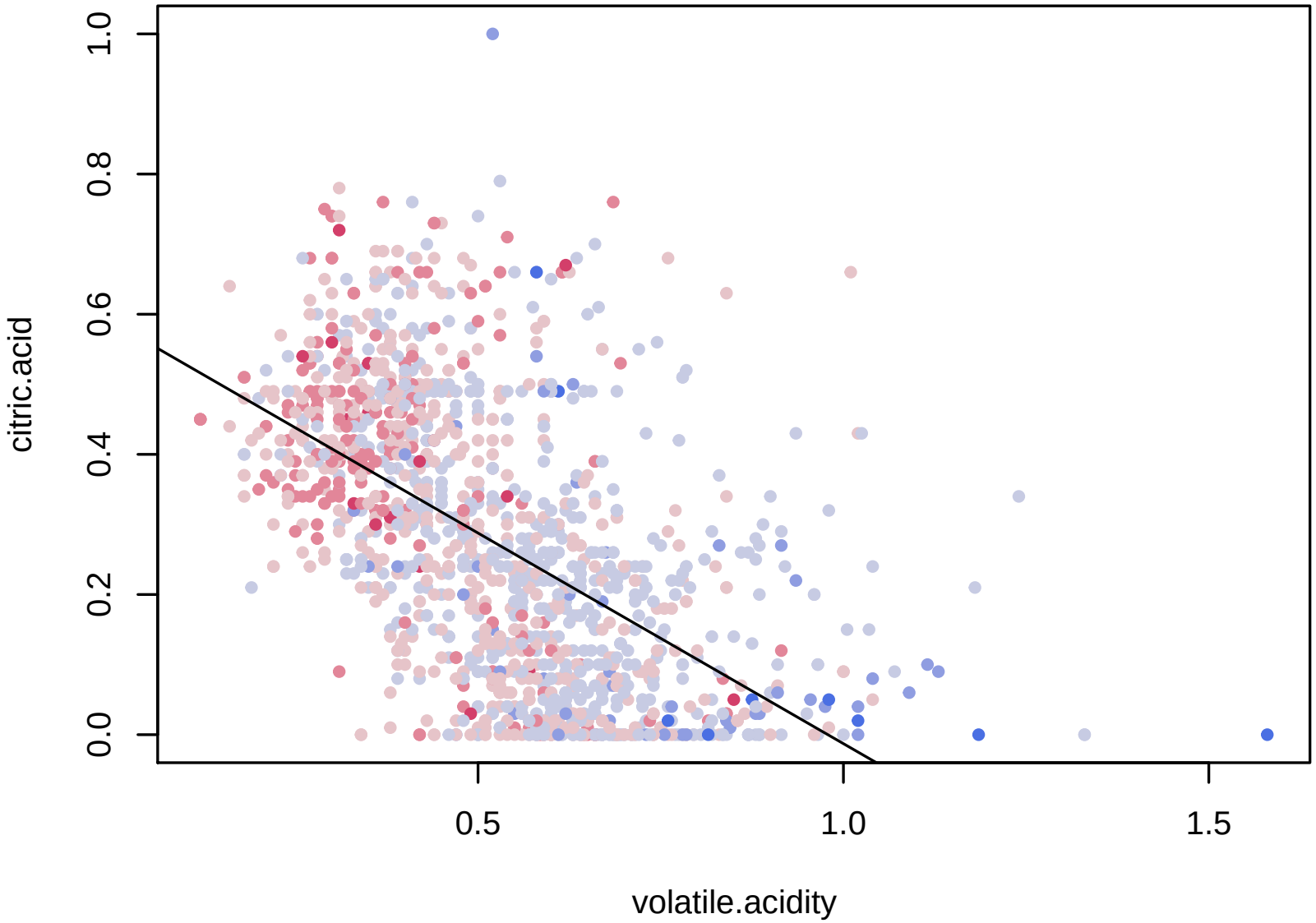
Soufre


total.sulfur.dioxide
0
50
100
150
0
10
20
30
40
50
60
70
free.sulfur.dioxide

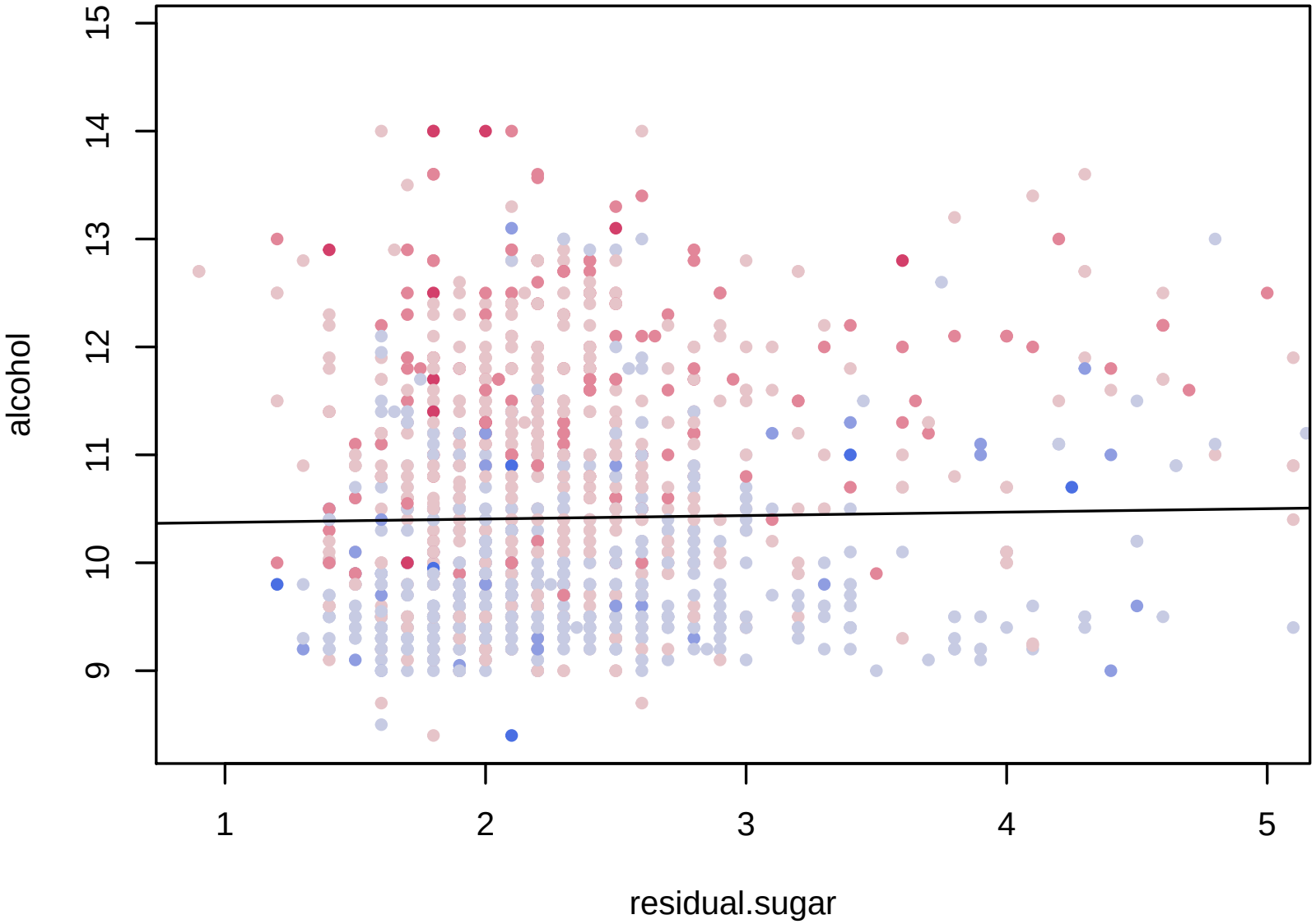
Soufre


total.sulfur.dioxide
0
50
100
150
0
10
20
30
40
50
60
70
free.sulfur.dioxide

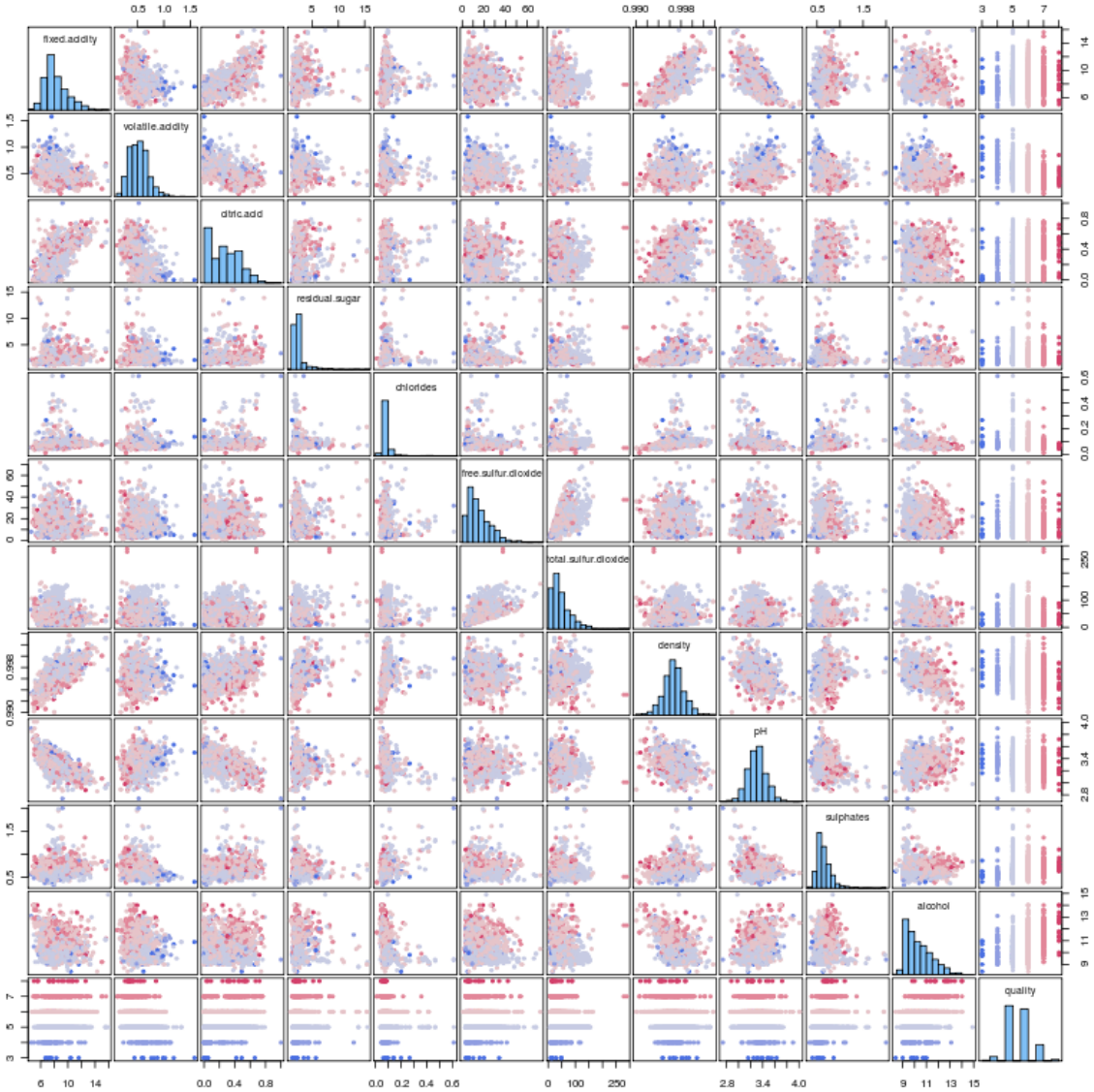
Acidité



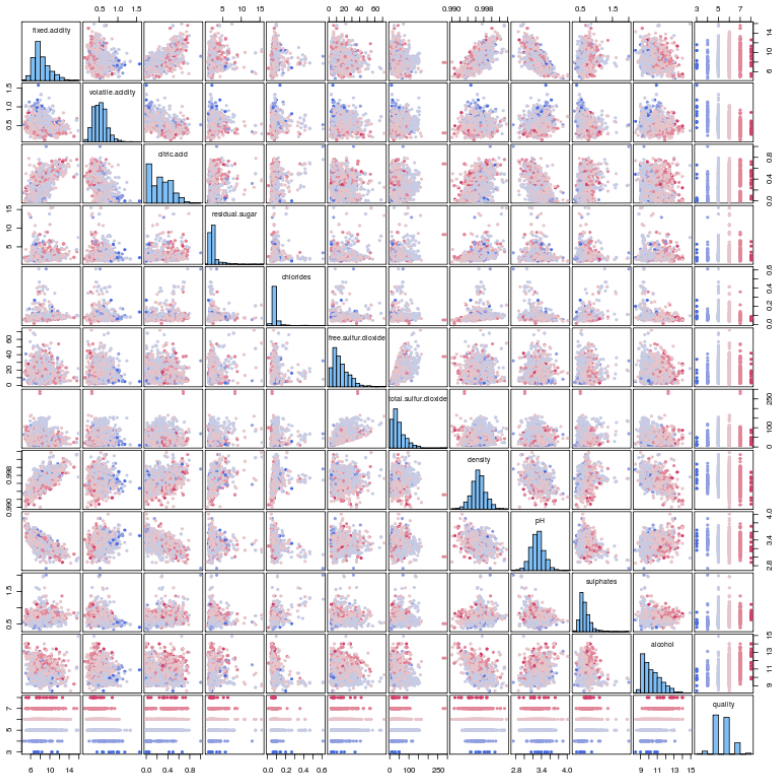
Alcool et sucre

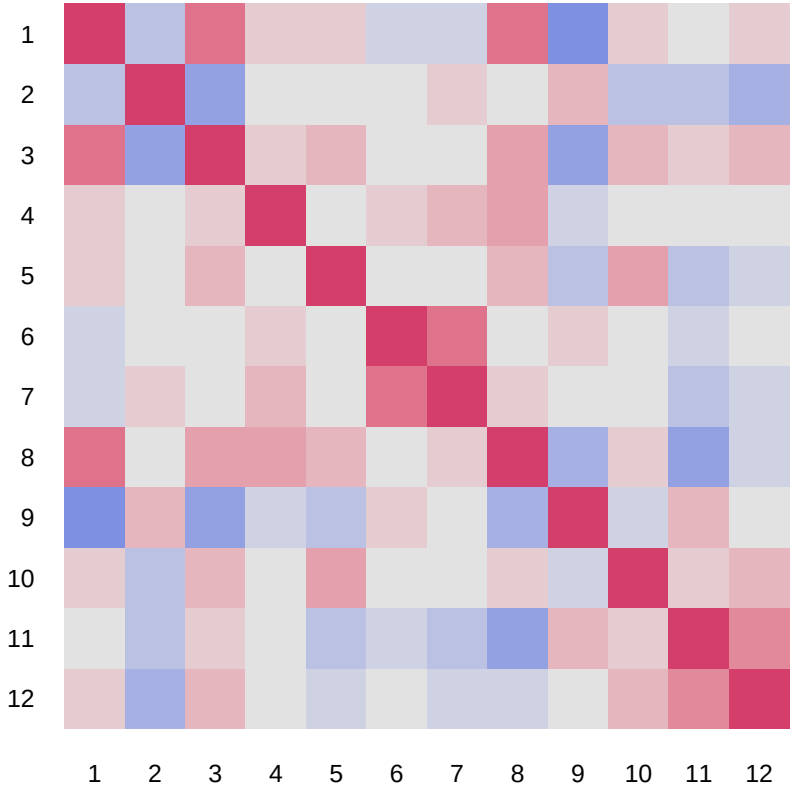


Vision globale

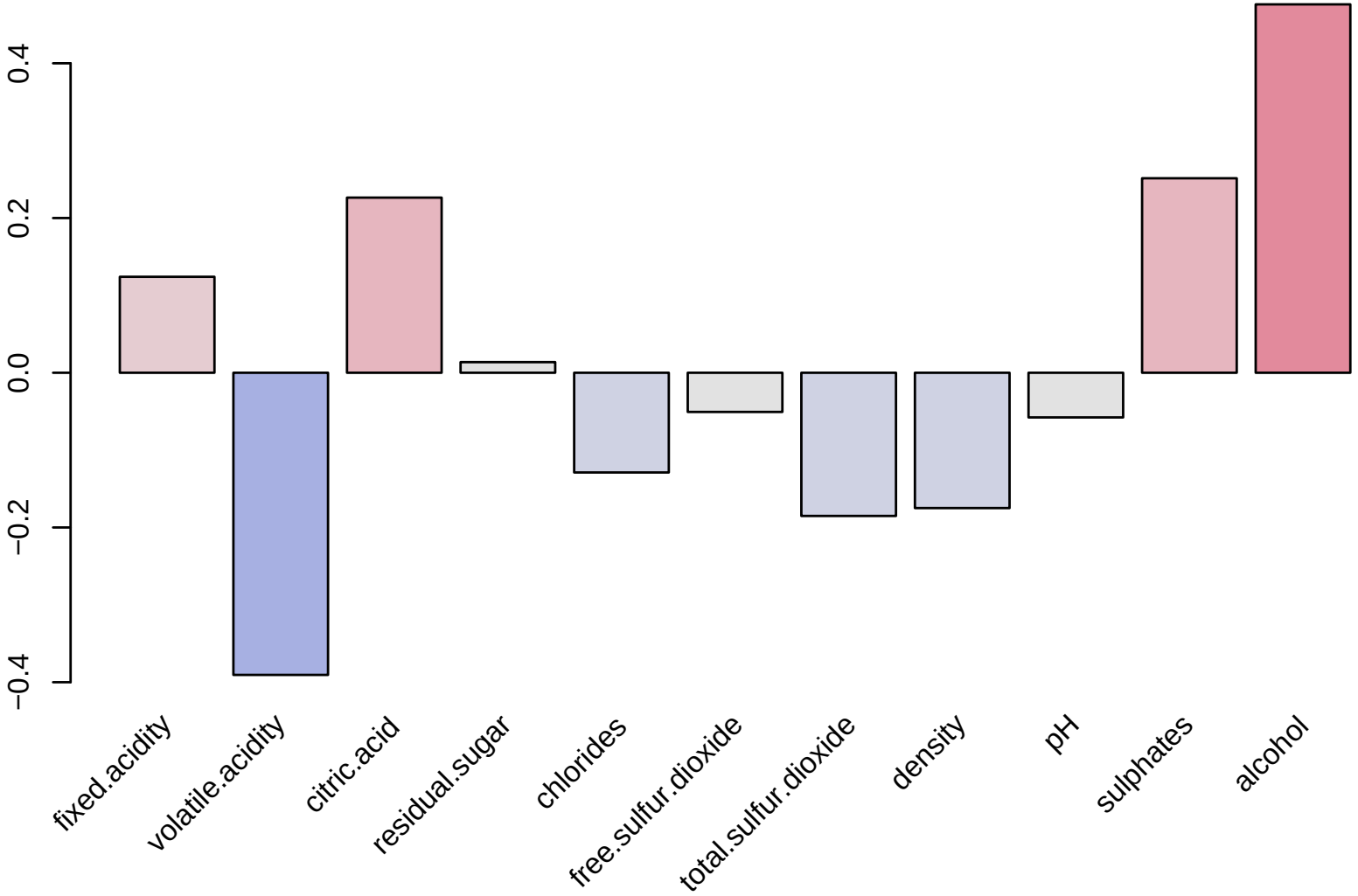


Vision globale

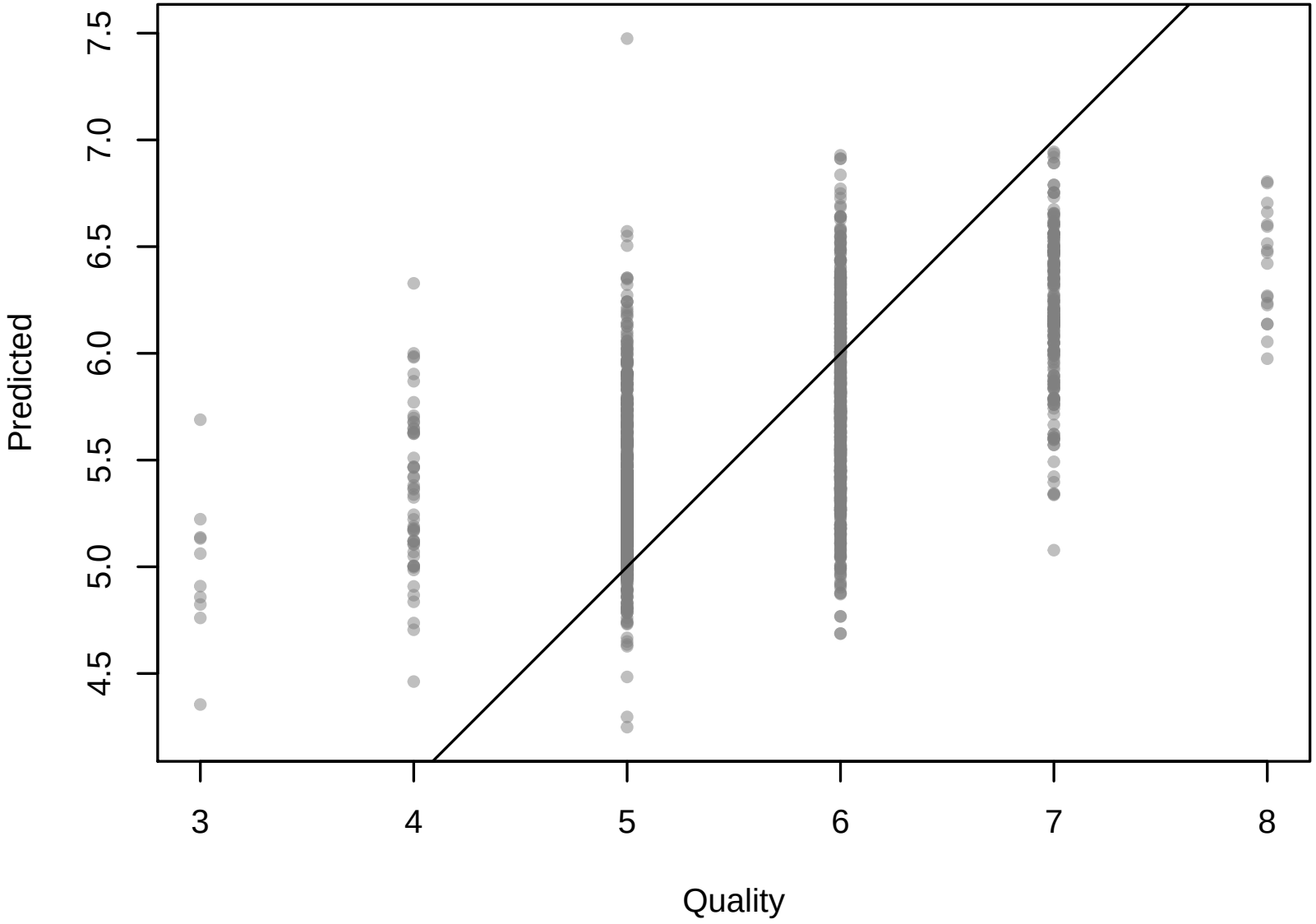




Qualité


0.4
0.2
0.0
−0.2
−0.4
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol

Modèle linéaire

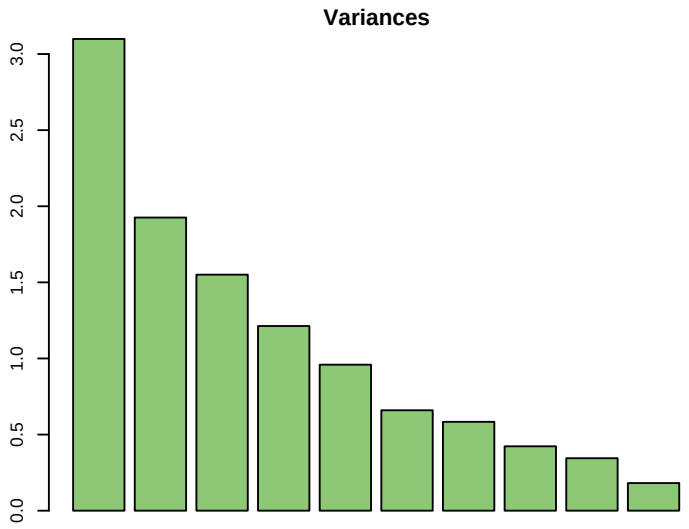


Analyse en composantes principales

Analyse sans la qualité

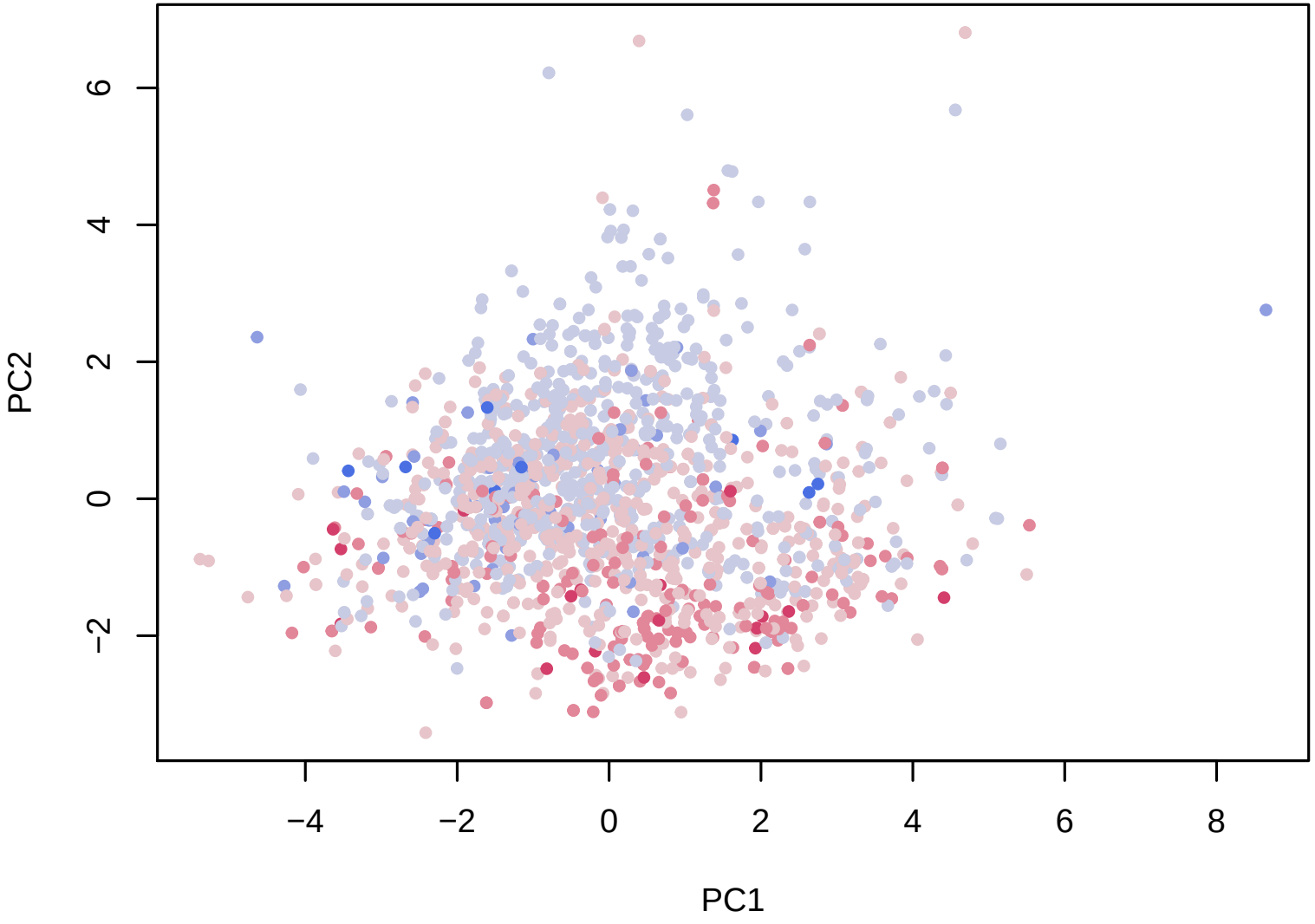
- pas de simplification forte
- seulement 46 % dans les deux premières composantes
- 80 % dans 5 composantes

Résultats similaires avec la qualité incluse



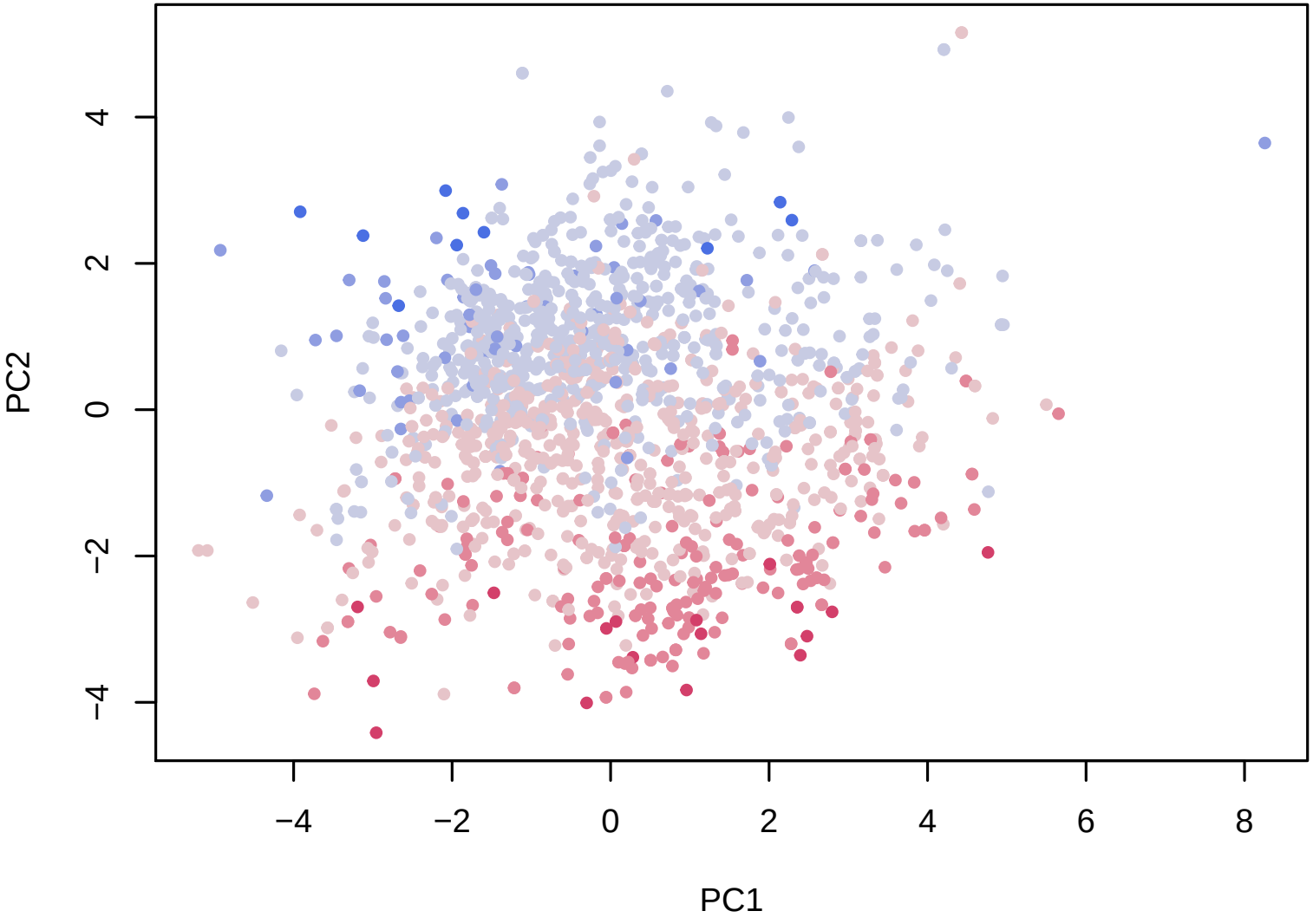
Analyse en composantes principales

Sans qualité (46 %)

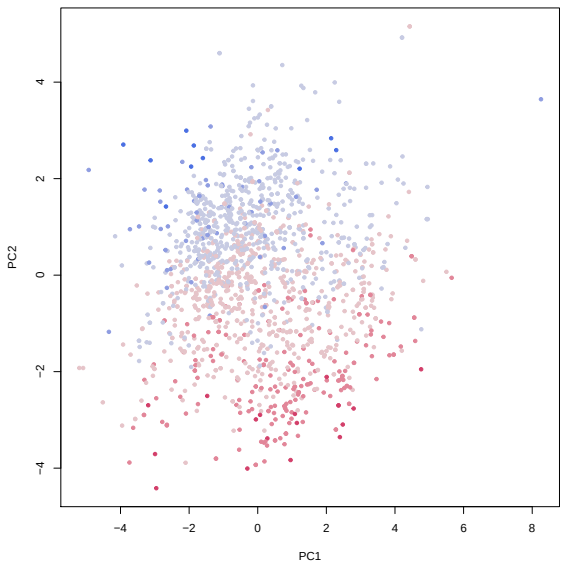


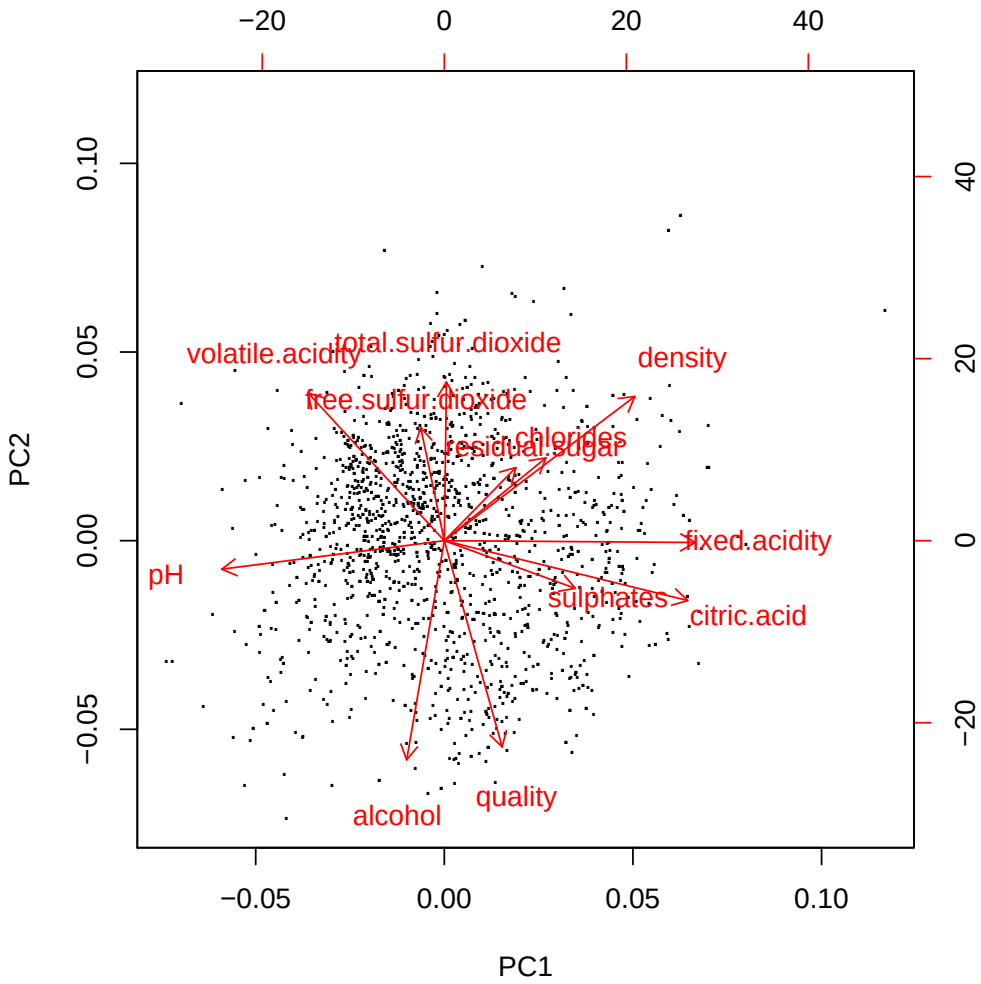
Analyse en composantes principales

Avec qualité (45 %)



Biplot


PC1
PC2


volatile.acidity
total.sulfur.dioxide
density
free.sulfur.dioxide
residual.sugar
chlorides
fixed.acidity
pH
sulphates
citric.acid
alcohol
quality
PC1
PC2

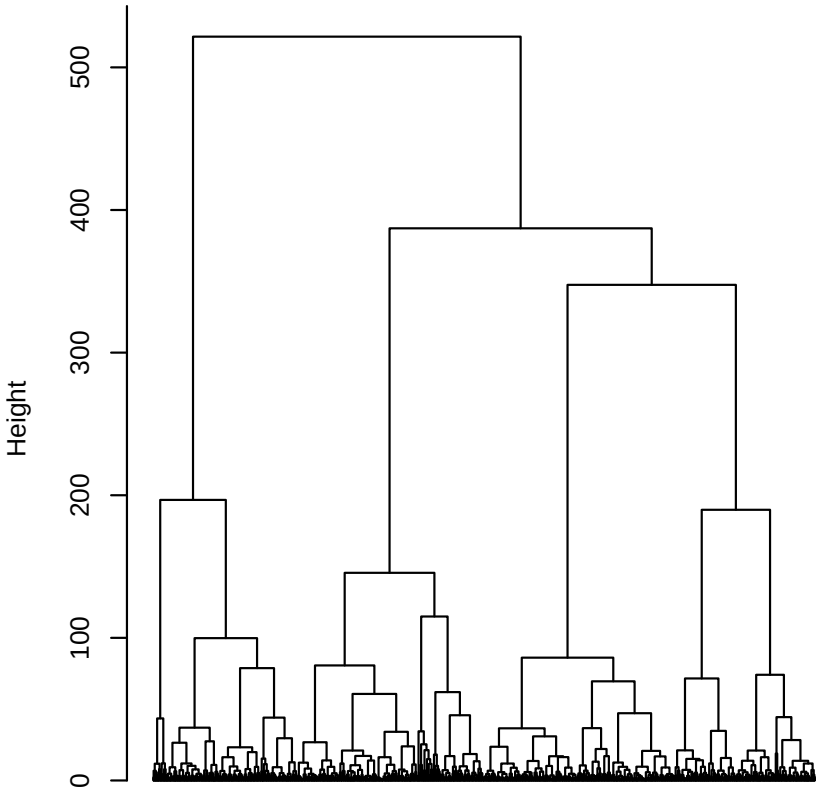
Plan

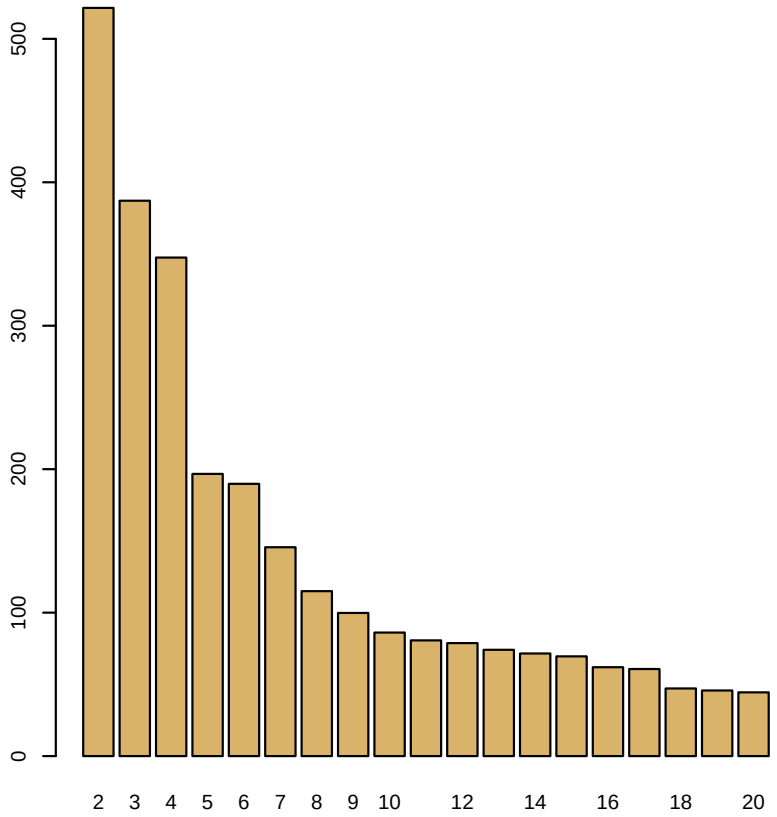
Analyses univariées

Analyses multivariées

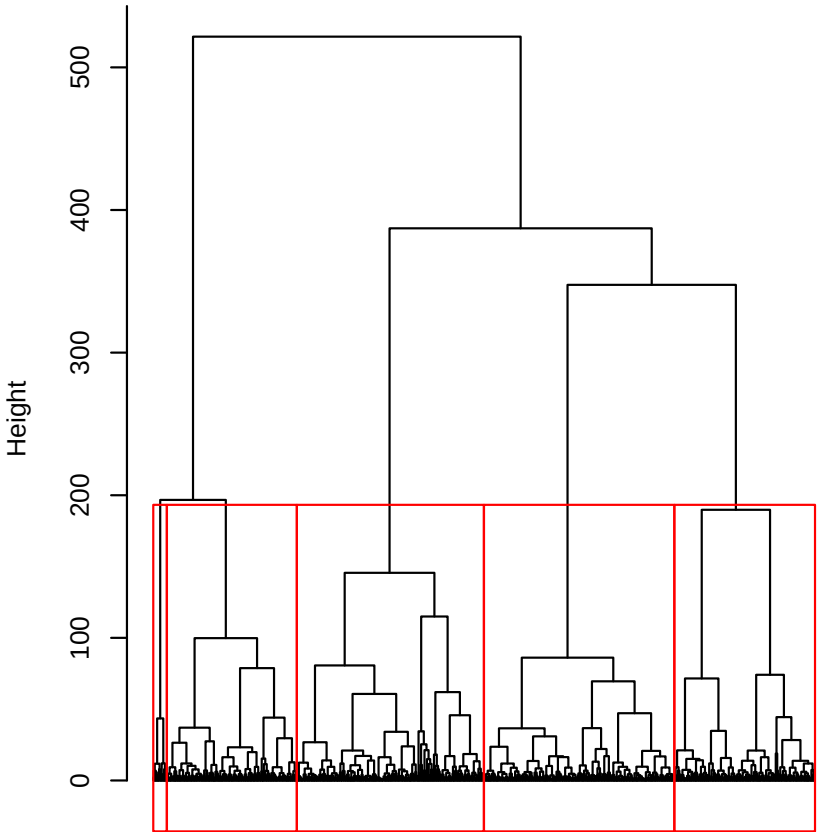
Classification

Classification hiérarchique (Ward)

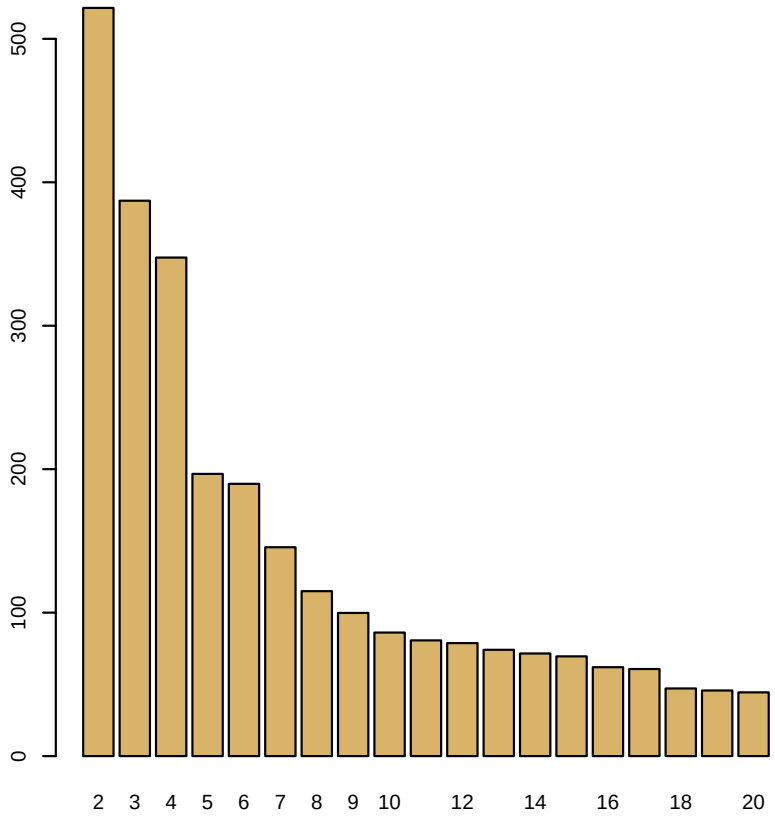




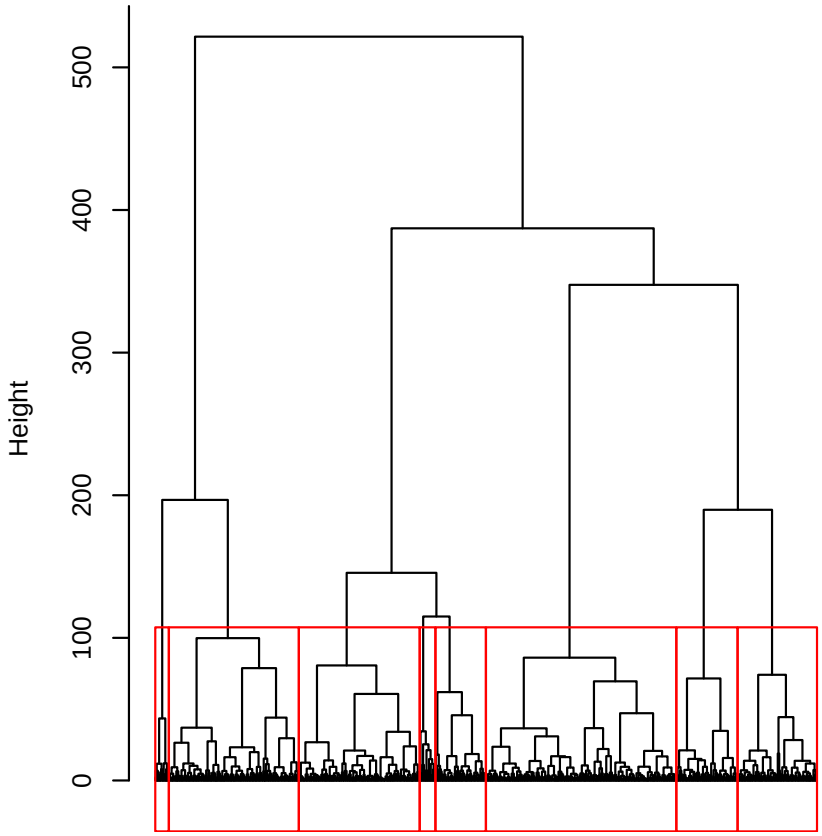
Classification hiérarchique (Ward)



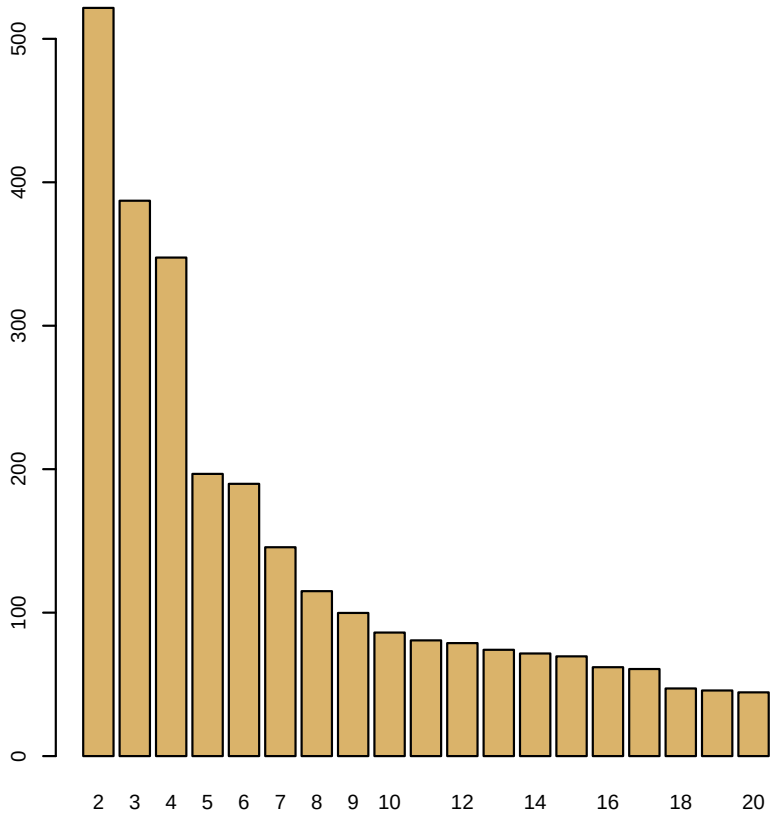
5 classes



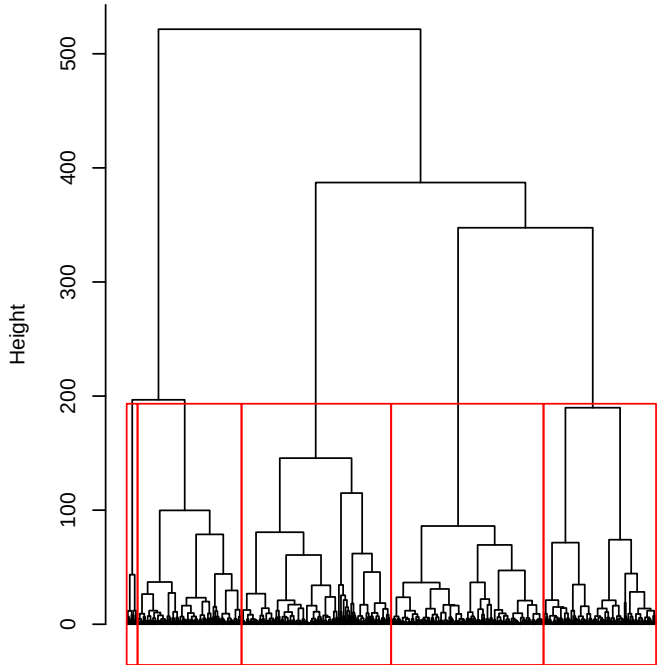
Classification hiérarchique (Ward)

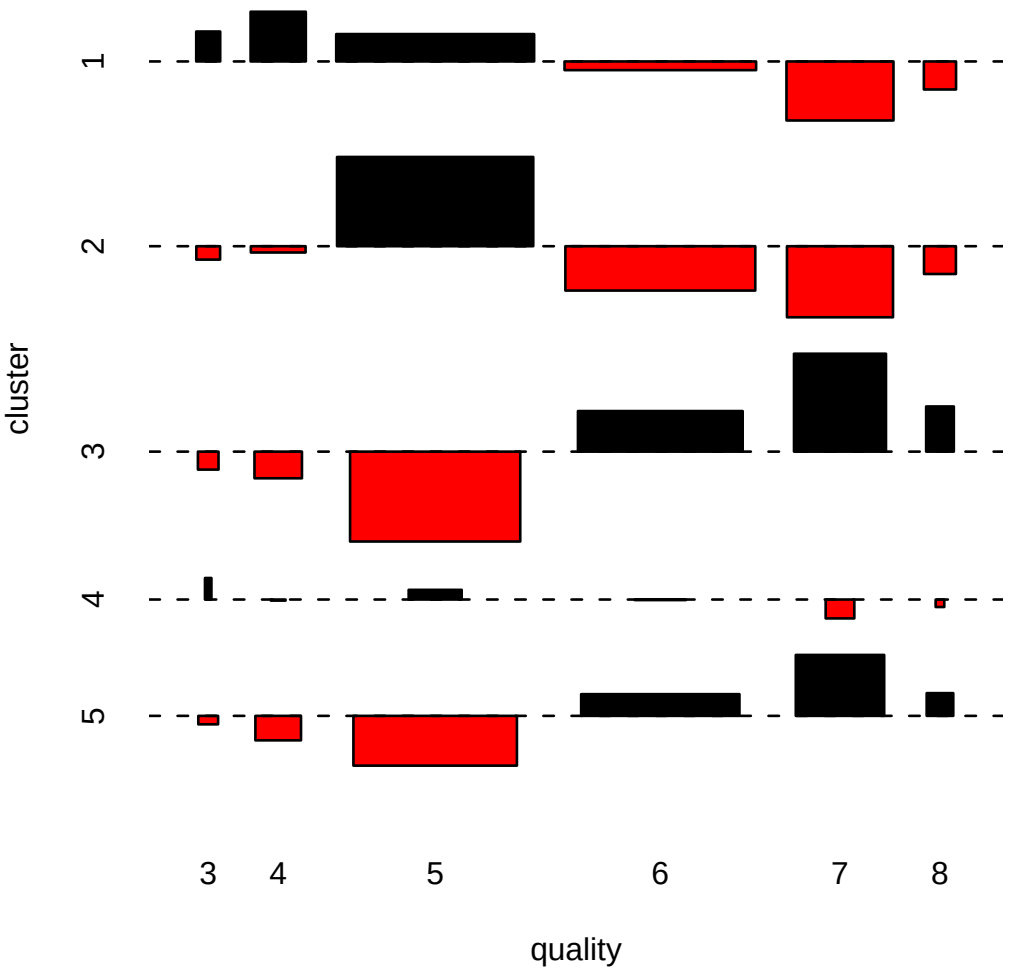


8 classes



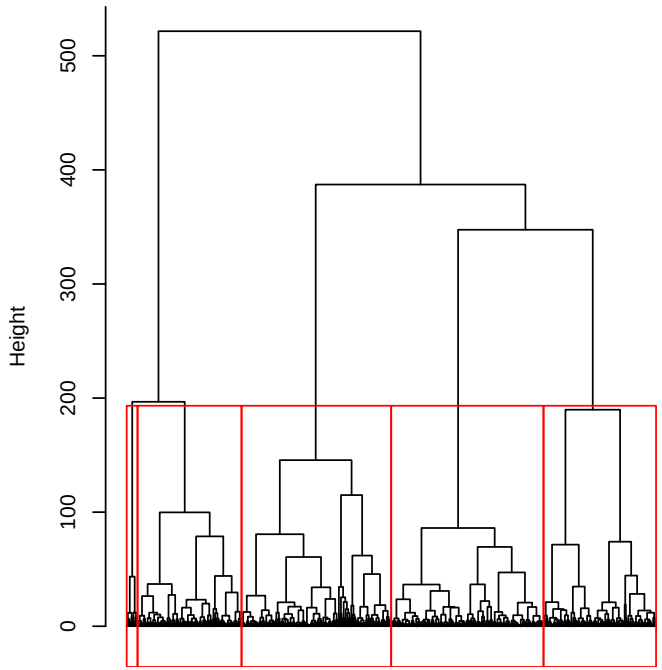
Analyse

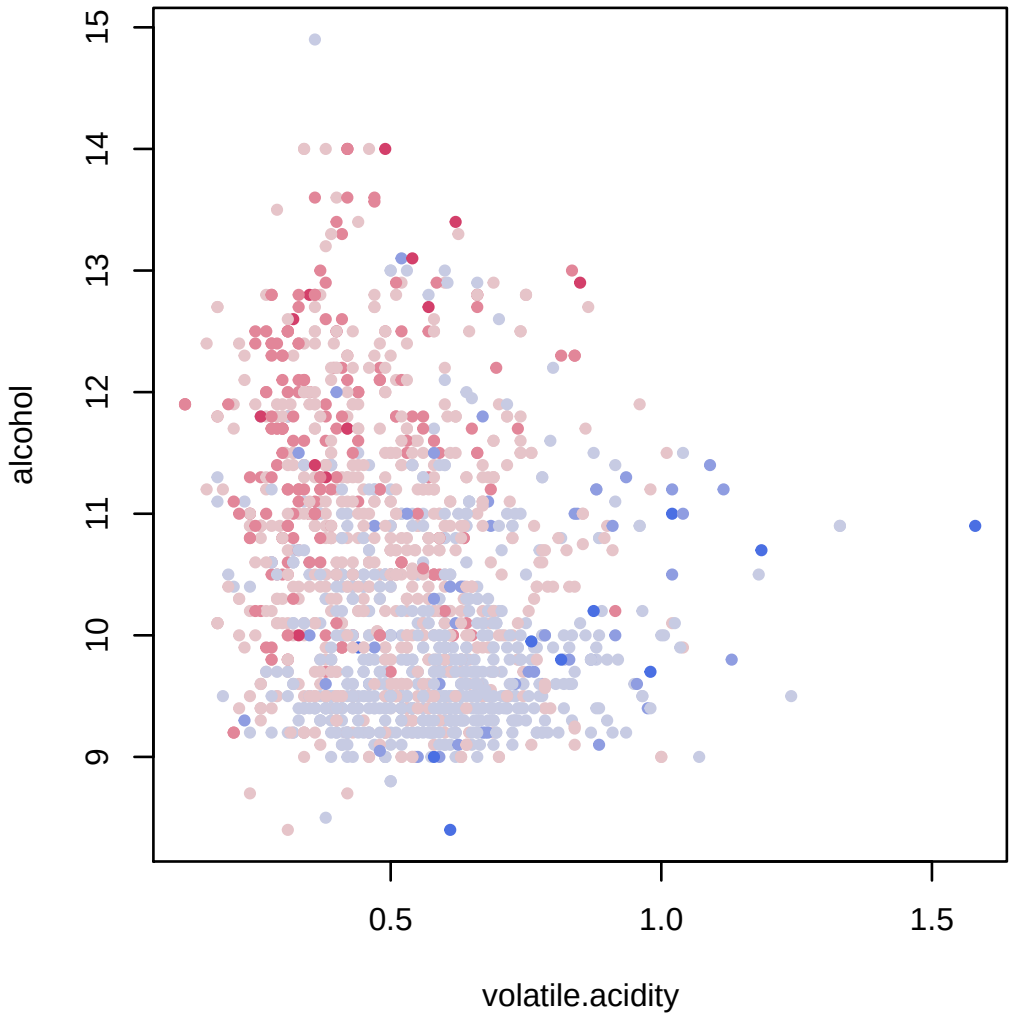




- 5 classes
- bonne séparation en qualité

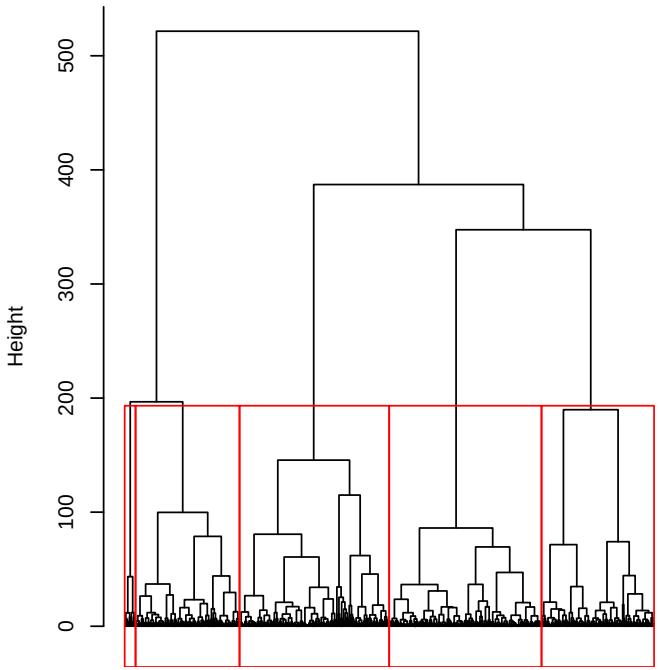
Analyse

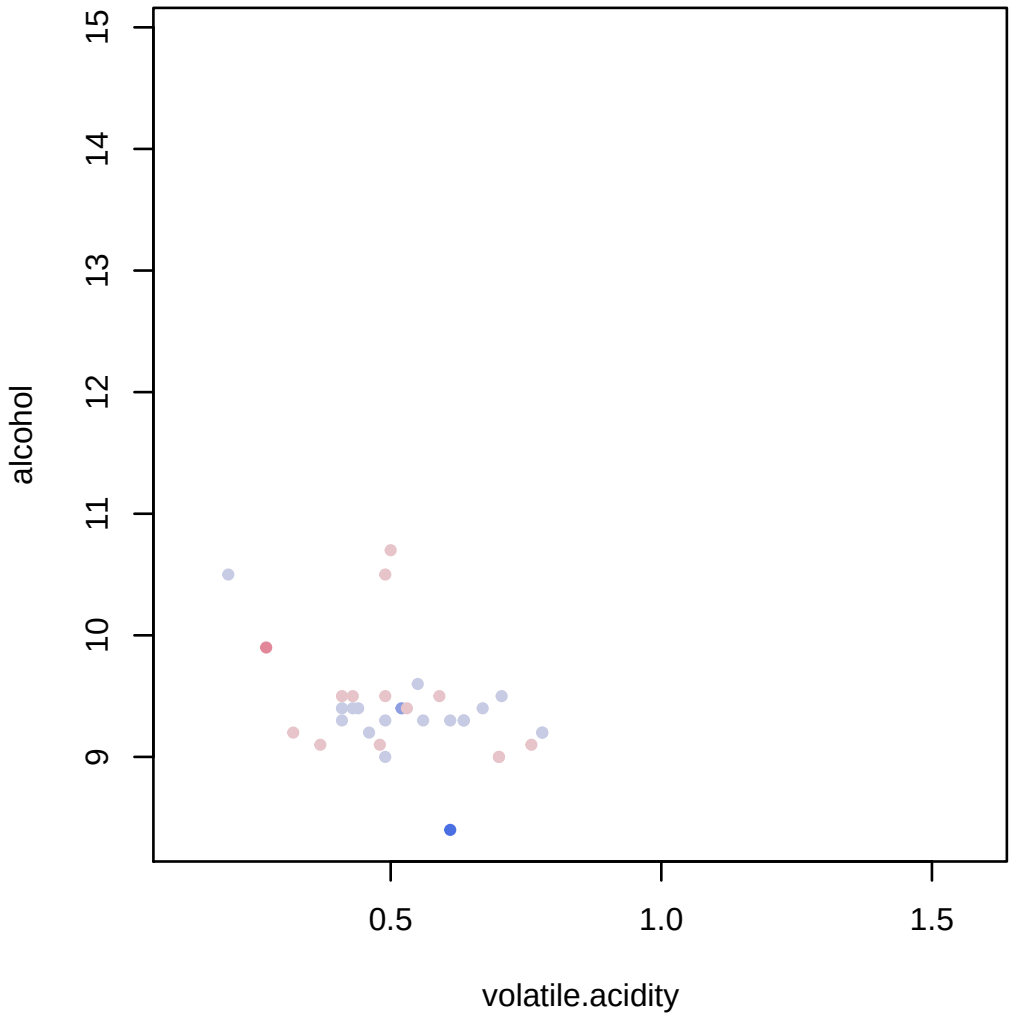




- 5 classes
- bonne séparation en qualité

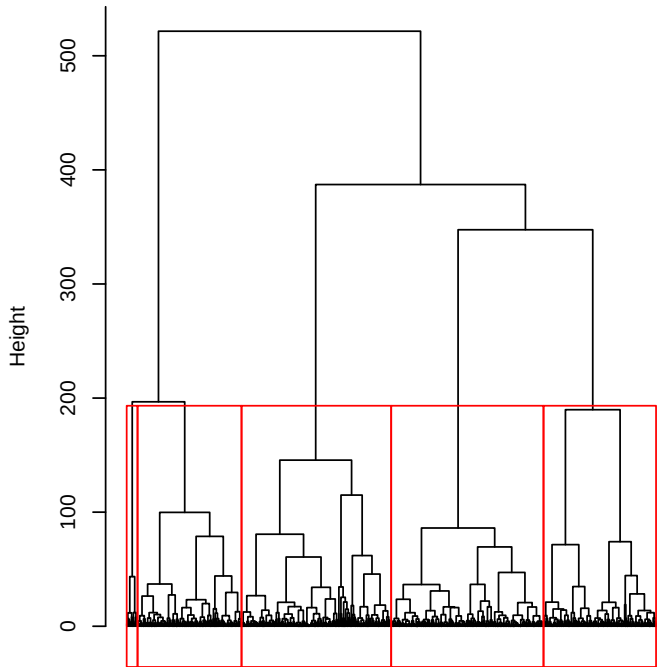
Analyse

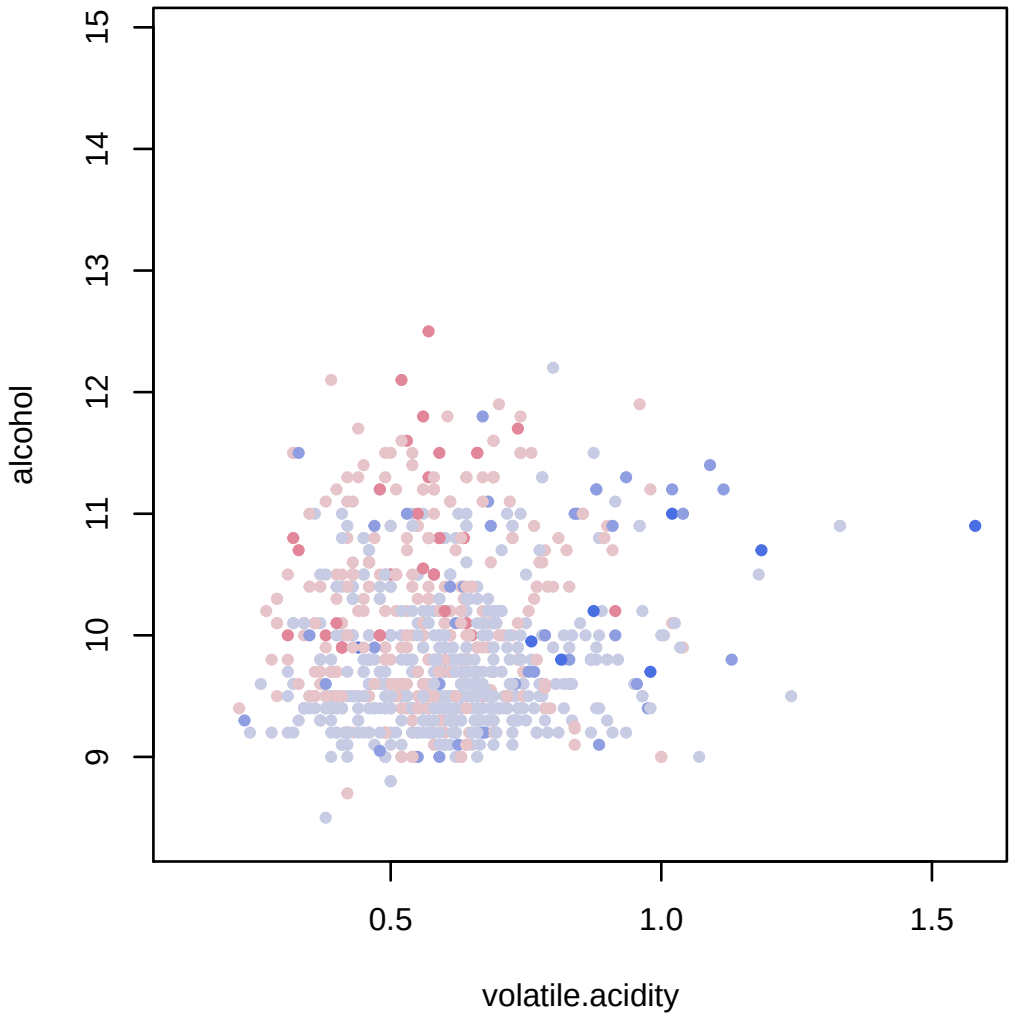




- 5 classes
- bonne séparation en qualité

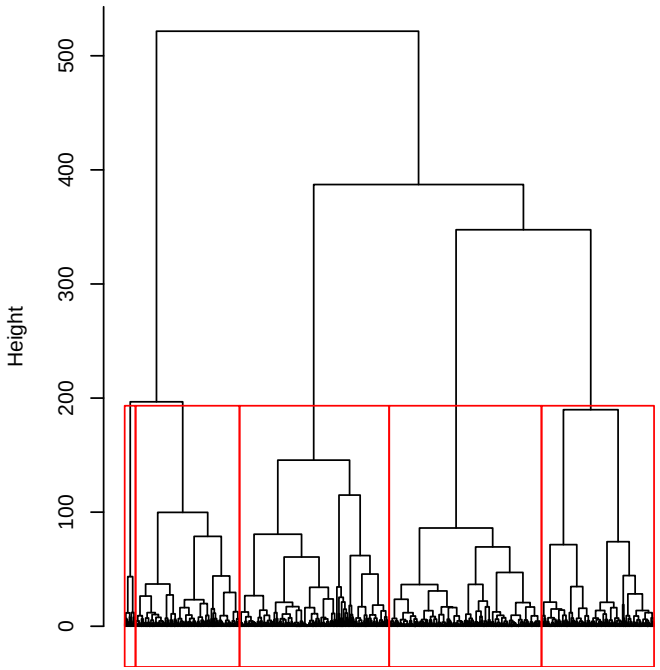
Analyse

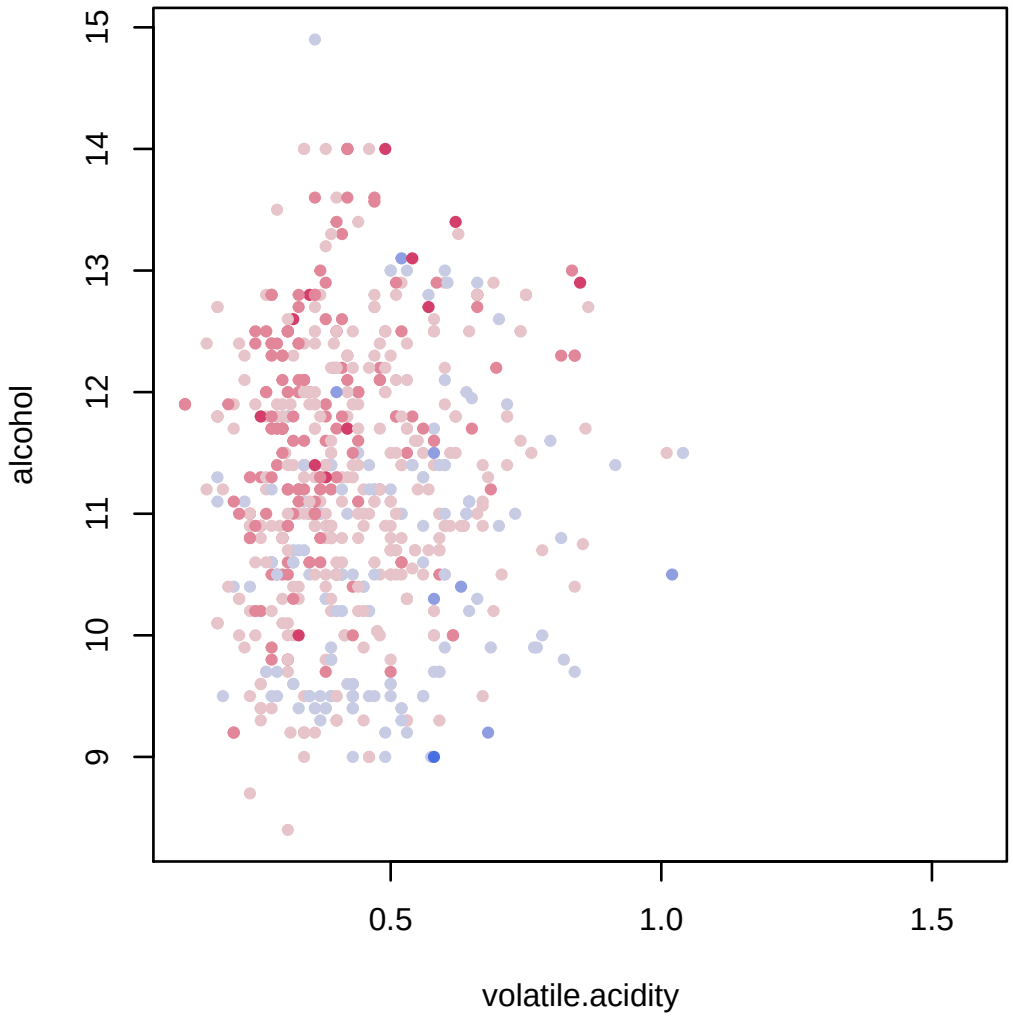




- 5 classes
- bonne séparation en qualité

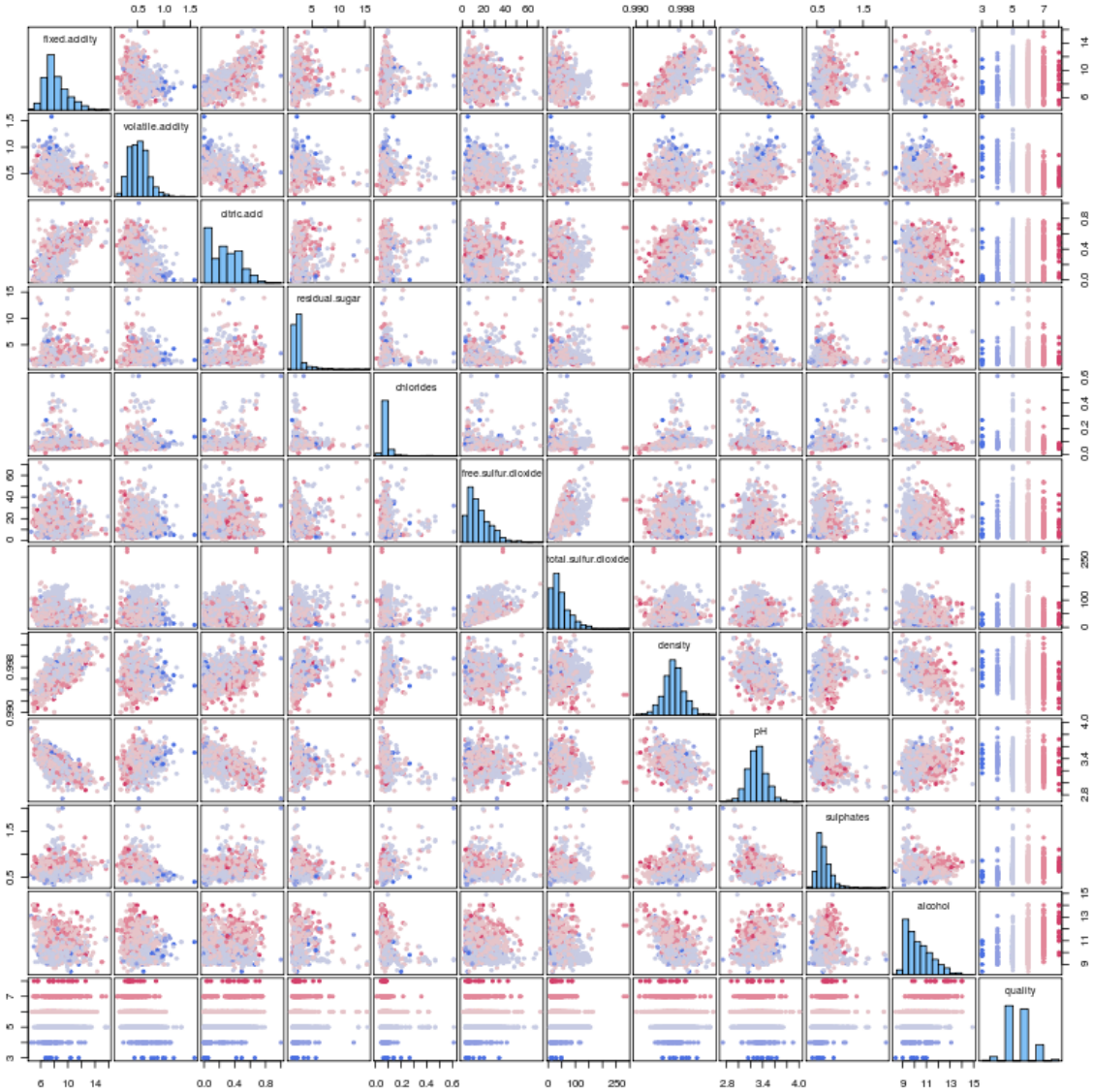
Analyse



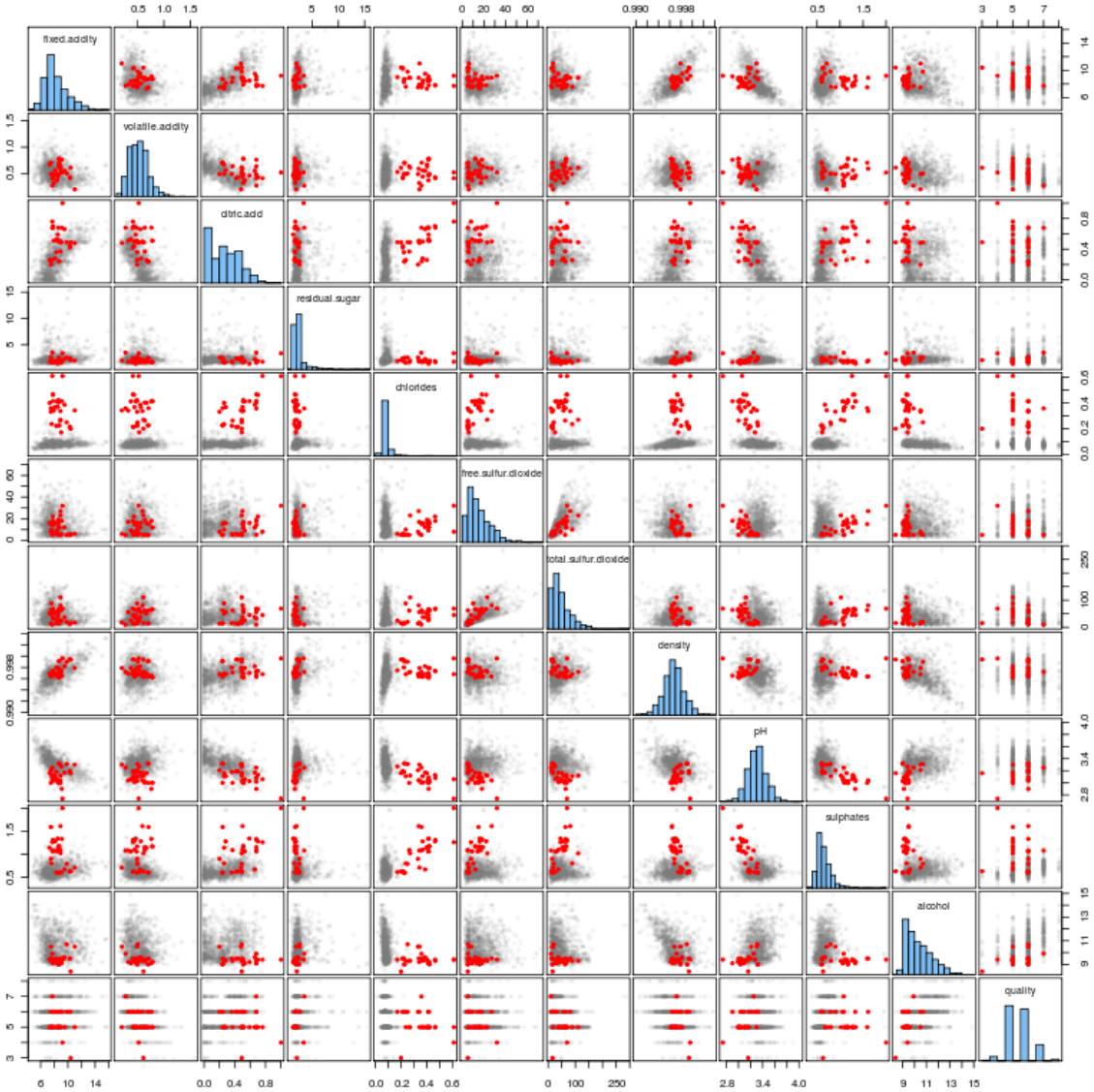


- 5 classes
- bonne séparation en qualité

Analyse


fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
quality

Analyse


fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
quality