

# Projet - Analyse des données

Axel Struys - Alexis Buckens

December 19, 2016

## 1 Introduction

Nous avons choisi une base de données en provenance de l'UCI Machine Learning Repository. Celle-ci est composée d'échantillons de vins rouges portugais, ayant une appellation d'origine contrôlée "Vinho Verde". Ce vin provient du nord ouest du Portugal, réparti en 9 sous-régions ayant des sols et des climats différents. Cet éclatement de la production de ce vin à pour conséquence une grande variabilité des propriétés des différents échantillons. Dans ce dataset, chaque observation représente les différentes analyses physico-chimiques d'un échantillon de vin distinct, associées à une évaluation subjective, par un œnologue, de la qualité de ce vin.

### 1.1 Description des variables

Afin de mieux comprendre les analyses ultérieures, nous allons brièvement décrire les variables et ce qu'elles représentent.

**Fixed acidity (g/L)** Cette variable représente la concentration en acide tartarique présente dans le vin. C'est l'acide primaire présent dans les grappes de raisins, et son rôle est très important dans le goût du vin. De plus, il permet de contrôler la prolifération bactérienne en agissant comme conservateur.

**Volatile acidity (g/L)** C'est une mesure de l'acide acétique présente dans le vin. Cette acide est produite par l'activité métabolique des bactéries et levures. La concentration doit être idéalement de 0,3g/L; Une plus haute concentration peut altérer l'expérience gustative en donnant un goût sûr au vin.

**Citric acid (g/L)** C'est un composant intermédiaire du cycle de l'acide citrique, permettant aux bactéries et aux levures de produire de l'énergie. Normalement, presque toute l'acide citrique est consommée durant la fermentation du vin, mais les vignerons peuvent en ajouter afin de donner un goût frais au vin.

**Sulfur dioxide (mg/L)** Le dioxyde de soufre est un sous-produit de la fermentation, mais il est aussi utilisé comme additif par les vignerons. Il régule la croissance des bactéries et levures. Mais son intérêt majeur réside dans ses propriétés antioxydantes. En effet, le vieillissement du vin produit de l'acétaldéhyde, molécule qui à l'odeur d'une pomme brunie. Le dioxyde de soufre va se lier avec cette molécule, la rendant inodore. Le dioxyde de soufre permet donc de préserver le goût fruité du vin. Néanmoins, trop de dioxyde de soufre donne une odeur soufrée au vin, irritant les parois nasales. Il est présent à la fois sous forme dissoute et sous forme gazeuse (free sulfur dioxide)

**Density (g/mL)** C'est la masse du vin par unité de volume. L'éthanol étant peu dense (0.789 g/mL), plus un vin est alcoolisé, moins il est dense.

**Résiduel sugar (g/L)** Ce sont les sucres restant après la fermentation du sucre en alcool par les bactéries et levures. Ceci permet de distinguer un vin sec ( $\leq 4$ g/L) d'un vin moelleux (entre 12 g/L et 45 g/L). Le sucre a une grande importance dans les caractéristiques sensorielles du vin, permettant de balancer son amertume et donnant un goût fruité agréable.

**pH** Mesurant l'acidité, le pH ne corrèle pas totalement avec les autres variables représentant l'acides, car il est une mesure totale de l'acidité, incluant d'autres acides mineurs non représentés dans ce dataset.

**Chlorides (g/L)** C'est la concentration en ions chlorures dans le vin. Ils proviennent de la dissolution du chlorure de sodium (sel). Ils contribuent au goût salé du vin, mais trop de sel peut influencer négativement le goût.

**Sulphate (sulphate de potassium) (g/L)** C'est un engrais qui permet de compenser le manque en potassium du sol, favorisant la croissance des vignes. En outre, le sulfate est nécessaire pour la synthèse des protéines de la plante, et a un rôle bénéfique dans la formation de sucres et des composés organoleptiques (qui ont un rôle dans la perception du goût). Il a un effet sur les niveaux de sucres et donc d'alcool.

**Alcohol (vol.%)** Élément essentiel du vin, l'éthanol est le produit attendu de la fermentation des levures. Avec le sucre, l'alcool est responsable du goût sucré du vin, et est évidemment un facteur important dans la dégustation de ce vin. Néanmoins, un volume trop important d'alcool peut dégrader l'expérience sensorielle.

**Quality** C'est l'évaluation subjective, par des experts, de la qualité de l'échantillon de vin. C'est une variable discrète, dont le minimum est 0, et le maximum est 10. Dans l'échantillon sélectionné, celle-ci s'étend entre 3 et 8.

## 1.2 Intérêt des méthodes factorielles dans ce projet

La description des variables nous montrent que beaucoup peuvent influencer la qualité d'un vin, mais aussi, il apparaît qu'elles peuvent s'influencer entre elles. Aussi, il peut sembler plus adéquat de choisir un modèle d'entrée-sortie tel qu'une régression linéaire sur la qualité. Néanmoins, nous pensons que l'utilisation de méthodes factorielles peut présenter plusieurs avantages dans ce dataset.

Premièrement, les méthodes factorielles sont surtout descriptives. Ainsi, elles peuvent être considérées comme une des premières étapes dans la compréhension des données à analyser, nous permettant une analyse exploratoire des données. Deuxièmement, nous nous trouvons dans une situation où il y a beaucoup de variables, et il est possible que nous nous trouvons dans une situation où beaucoup de ces variables covarient ensemble. Ceci pose des problèmes évidents de multicolinéarités dans l'hypothèse où nous souhaitons utiliser un modèle linéaire ultérieurement. Enfin, ces méthodes nous permettent de représenter ces variables sur un plan de plus petite dimension, permettant une compréhension plus simple et intuitive des données.

## 2 Méthodologie

L'échantillon original était composé de 1689 observations. Chaque observation représente un échantillon d'un vin différent; Elles sont donc indépendantes. Une lacune dans cette base de donnée est que les auteurs ont choisi, pour des raisons de confidentialités, d'occulter l'origine et la marque de ces échantillons. Afin de simplifier les analyses, nous avons décidé en accord

avec l'assistant de réduire la taille du dataset. En procédant à un échantillonnage (voir code en annexe), nous avons sélectionné 200 observations.

Nous allons séparer notre analyse en 4 parties. Dans la première partie, nous allons présenter quelques statistiques descriptives. Ensuite, nous procéderons à une analyse en composantes principales sur toutes les variables, sauf Quality. Ensuite, nous réaliserons une analyse des correspondances multiples en discrétisant certaines variable, et en incluant la variable quality. Enfin, nous procéderons à une analyse de clustering, afin de voir si les variables permettent de séparer naturellement les échantillons en fonction de leur qualité

Avant de procéder à l'analyse proprement dite, il convient de rapidement préparer nos données. Tout d'abord, comme nous ne disposons que de variables continues, il est nécessaire de transformer certaines d'entre elles en variables discrètes afin de pouvoir ultérieurement procéder à la l'analyse en correspondance multiples. De plus, cela nous permettra d'analyser de manière plus adéquate les variables dont la relation n'est pas linéaire (voir scatterplot matrix figure 4). Cette transformation peut être faite en séparant les valeurs des 5 premières variables du dataset, à savoir fixed.acidity, volatile.acidity, citric.acid, residual.sugar and chlorides en 5 intervalles et en attribuant chaque intervalle a une catégorie.

### 3 Statistiques descriptives

	Mean	Std.dev	Median	Min	Max
Fixed acidity (g/L)	8.23	1.67	7.80	5.00	13.40
Volatile acidity (g/L)	0.53	0.17	0.54	0.16	1.00
Citric acid (g/L)	0.27	0.19	0.26	0.00	0.68
Residual sugar (g/L)	2.45	1.12	2.20	1.20	8.80
Chlorides (g/L)	0.08	0.04	0.08	0.04	0.47
Free Sulfur Dioxide (mg/L)	14.69	8.92	13.00	3.00	45.00
Total Sulfur Dioxide (mg/L)	43.48	32.73	35.00	7.00	278.00
Density (g/mL)	0.997	0.002	0.997	0.992	1.002
pH	3.31	0.14	3.33	2.94	3.72
Sulphates (g/L)	0.64	0.13	0.62	0.37	1.31
Alcohol (vol. %)	10.49	1.06	10.20	9.00	14.00
Quality	5.65	0.83	6.00	3.00	8.00

Table 1: Statistiques descriptives (Moyenne, Ecart-type, mediane, minimum et maximum) des 12 variables

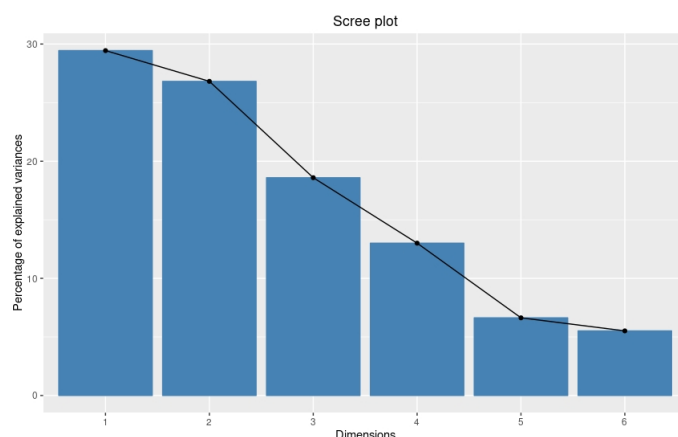
La table 1 présente les statistiques descriptives des variables du dataset. Dans l'annexe B se trouvent les boxplots, histogrammes, qq-plots, scatterplot matrix et tests de Shapiro-Wilk de la normalité des variables. Nous pouvons constater que seules les variables density et pH ont une distribution qui suit une loi normale. En outre, la relation entre certaines variables n'est pas linéaire. Ceci est à prendre en compte pour l'analyse en composante principale, car l'hypothèse sous-jacente de cette méthode suppose une relation linéaire entre les variables. Etant donné cette non-normalité, ainsi que la non-linéarité de la relation entre certaines variables, nous avons choisi d'utiliser la corrélation de Spearman afin de décrire les relations entre ces variables (voir table 2 de l'annexe A).

## 4 PCA

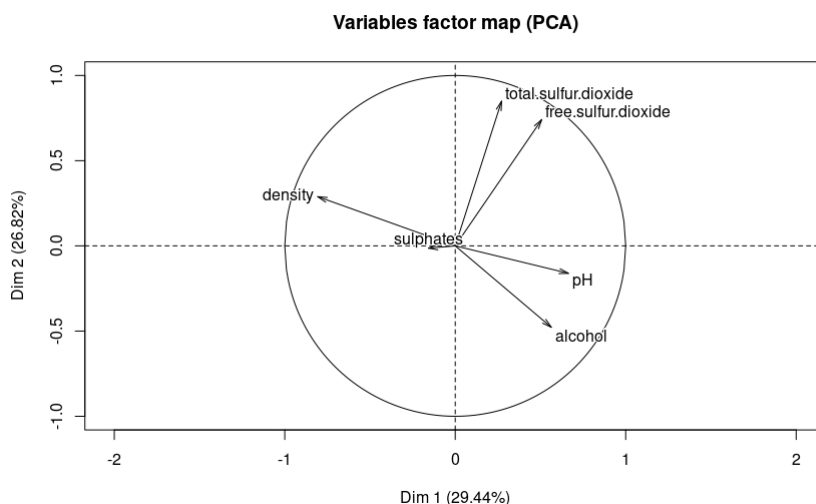
Nous allons commencer par nous intéresser aux variables continues en effectuant une ACP. Deux premières questions peuvent être posées : à partir de combien de dimensions suffisent à capturer l'essentiel de l'information, et comment les différentes variables contribuent-elles aux premières composantes.

Il nous est possible de répondre à la première question en observant les valeurs propres de l'analyse en composante principale. On peut observer que, comme prévu, la proportion de la variance expliquée par chacune des composantes est décroissante et que les 3 premières composantes permettent d'expliquer pratiquement 75% de la variance.

Le graphique suivant nous permet d'avoir une illustration visuelle et assez immédiate de la proportion de la variance expliquée par chacune des composantes principales :



La question de savoir comment les différentes variables contribuent aux deux premières composantes principales peut également être illustrée par un graphique :

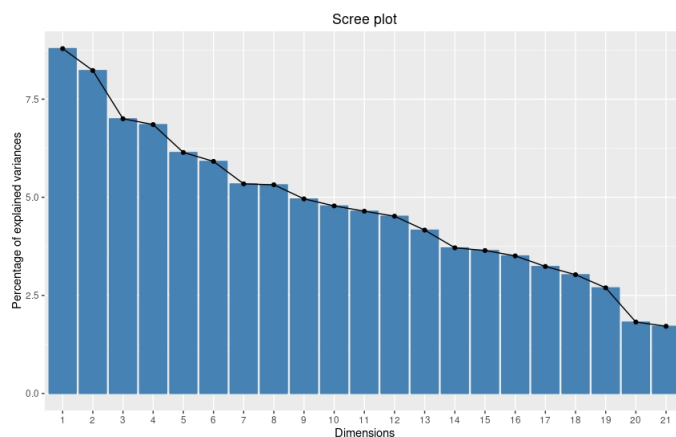


On peut voir sur celui-ci que la première dimension est plus liée aux variables "Ph" et "Density", et dans une moindre mesure "sulphates", tandis que la seconde dimension est elle plus liée aux variables "total.sulfur.dioxides" et "free.sulfur.dioxides", qui semblent elles-même proches l'une de l'autre (ce qui n'est naturellement pas surprenant).

## 5 MCA

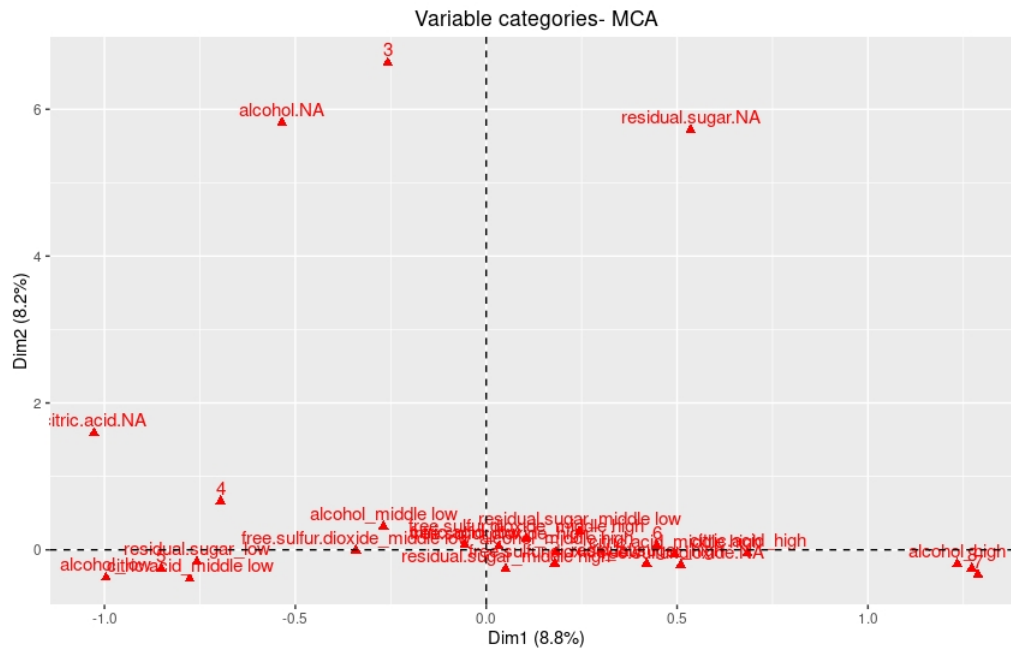
L'étape suivante consiste à effectuer une analyse en correspondance multiple. Pour cette analyse nous avons besoin de variables discrètes. La variable "quality" étant déjà une variable discrète, elle sera naturellement utilisée. Pour ce qui est des autres variables, nous les sélectionnerons sur base du nombre de levels qu'elles prennent si on les considère comme factors, afin de prendre celles qui naturellement semblent les plus proches d'une variable discrète et sur base de la non-linéarité de leur relations. Les variables sélectionnées seront alcohol, citric acid, residual.sugar et free.sulfur, et ces variables seront transformées en variables discrètes avec 5 catégories.

La première question est de savoir combien de dimensions conserver. Pour répondre à cette question, on peut de nouveau utiliser les valeurs propres et s'intéresser aux dimensions avec les valeurs propres les plus élevées. Si on trace un graphique reprenant les valeurs propres, de la plus élevée à la plus faible, on obtient le graphique suivant :



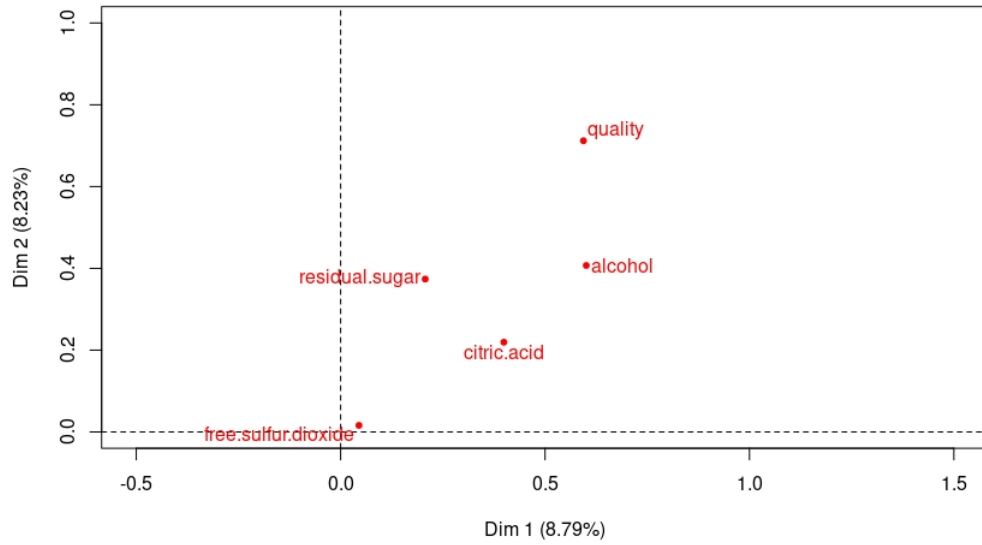
On peut donc constater que la proportion de variance expliquée par les différentes dimensions décroît lentement. Puisqu'il nous faut choisir un nombre de dimensions restreint pour expliquer nos données, il nous faut choisir également un critère de décision. Un premier critère est celui du "coup de coude" : Il s'agit de choisir le nombre de dimension à partir duquel la valeur des valeurs propres décroît brusquement. Un "coup de coude" peut être trouvé à plusieurs endroits, entre autre aux alentours de la 4eme dimension (mais également après deux dimensions), mais en ne choisissant que 4 dimensions, on ne peut expliquer que 30% de la variance. Un second critère consiste à choisir les dimensions pour lesquelles la valeur propres est supérieure à  $\frac{1}{nbvar}$ . Ce second critère nous pousse à choisir 10 dimensions, ce qui est fort élevé. On pourrait également choisir comme critère le % de variance expliquée, en fixant par exemple le seuil à 60%. Ce critère nous pousserait également à choisir 10 variables. On utilisera donc au final les 4 premières dimensions.

On peut également s'intéresser aux positions des individus, variables et categories dans les deux premières dimensions :



Sur ce graphique, on peut observer les catégories des variables (triangles rouges associés aux noms des variables), et les individus (ronds bleus). On peut ainsi voir quelles catégories/individus sont similaires. En l'occurrence, le graphe est un peu difficile à comprendre du fait que la plupart des valeurs sont proches de l'origine, à l'exception de deux outliers et des valeurs correspondant aux valeurs manquantes pour alcohol et residual.sugar qui s'en éloignent singulièrement. Néanmoins on peut déjà observer une certaine similarité entre les catégories "basses" de residual.sugar, citric.acid et alcohol, et on peut constater que alcohol "low" et alcohol "high" sont chacun à une des extrémités de l'axe correspondant à la première dimension. La première dimension semble donc correspondre essentiellement au taux d'alcool dans le vin. On pourra vérifier cela ultérieurement en observant les contributions des différentes variables aux premières dimensions. Finalement les vins de mauvaise qualité (4) se trouvent du côté gauche (alcohol low) tandis que les vins de bonne qualité (7) se trouvent du côté droit (alcohol high).

On peut aussi ne s'intéresser qu'aux variables analysées et tracer un graphique des coordonnées des variables dans le plan formé par les deux premières dimensions :



Aucune des variables ne semble être particulièrement associée avec l'un ou l'autre axe.

## 6 Conclusion

## 7 Annexes

### A Matrice de corrélation

	alcohol	chlorides	citric.acid	density	fixed.acidity	free.sulfur.dioxide	pH	quality	residual.sugar	sulphates	total.sulfur.dioxide	volatile.acidity
alcohol	1.00	-0.25	0.11	-0.41	-0.04	-0.03	0.13	0.48	0.19	0.21	-0.20	-0.18
chlorides	-0.25	1.00	0.06	0.31	0.12	0.09	-0.16	-0.11	0.23	-0.13	0.15	0.10
citric.acid	0.11	0.06	1.00	0.43	0.63	-0.17	-0.57	0.27	0.16	0.36	-0.00	-0.62
density	-0.41	0.31	0.43	1.00	0.65	-0.17	-0.31	-0.07	0.35	0.20	0.02	-0.02
fixed.acidity	-0.04	0.12	0.63	0.65	1.00	-0.23	-0.70	0.09	0.20	0.22	-0.06	-0.20
free.sulfur.dioxide	-0.03	0.09	-0.17	-0.17	-0.23	1.00	0.22	0.05	0.08	0.04	0.77	0.05
pH	0.13	-0.16	-0.57	-0.31	-0.70	0.22	1.00	-0.01	-0.03	-0.07	0.04	0.21
quality	0.48	-0.11	0.27	-0.07	0.09	0.05	-0.01	1.00	0.12	0.41	-0.07	-0.36
residual.sugar	0.19	0.23	0.16	0.35	0.20	0.08	-0.03	0.12	1.00	0.02	0.14	0.10
sulphates	0.21	-0.13	0.36	0.20	0.22	0.04	-0.07	0.41	0.02	1.00	-0.00	-0.43
total.sulfur.dioxide	-0.20	0.15	-0.00	0.02	-0.06	0.77	0.04	-0.07	0.14	-0.00	1.00	0.06
volatile.acidity	-0.18	0.10	-0.62	-0.02	-0.20	0.05	0.21	-0.36	0.10	-0.43	0.06	1.00

Table 2: Matrice de corrélation des 12 variables, en utilisant la corrélation de Spearman

### B Normalité

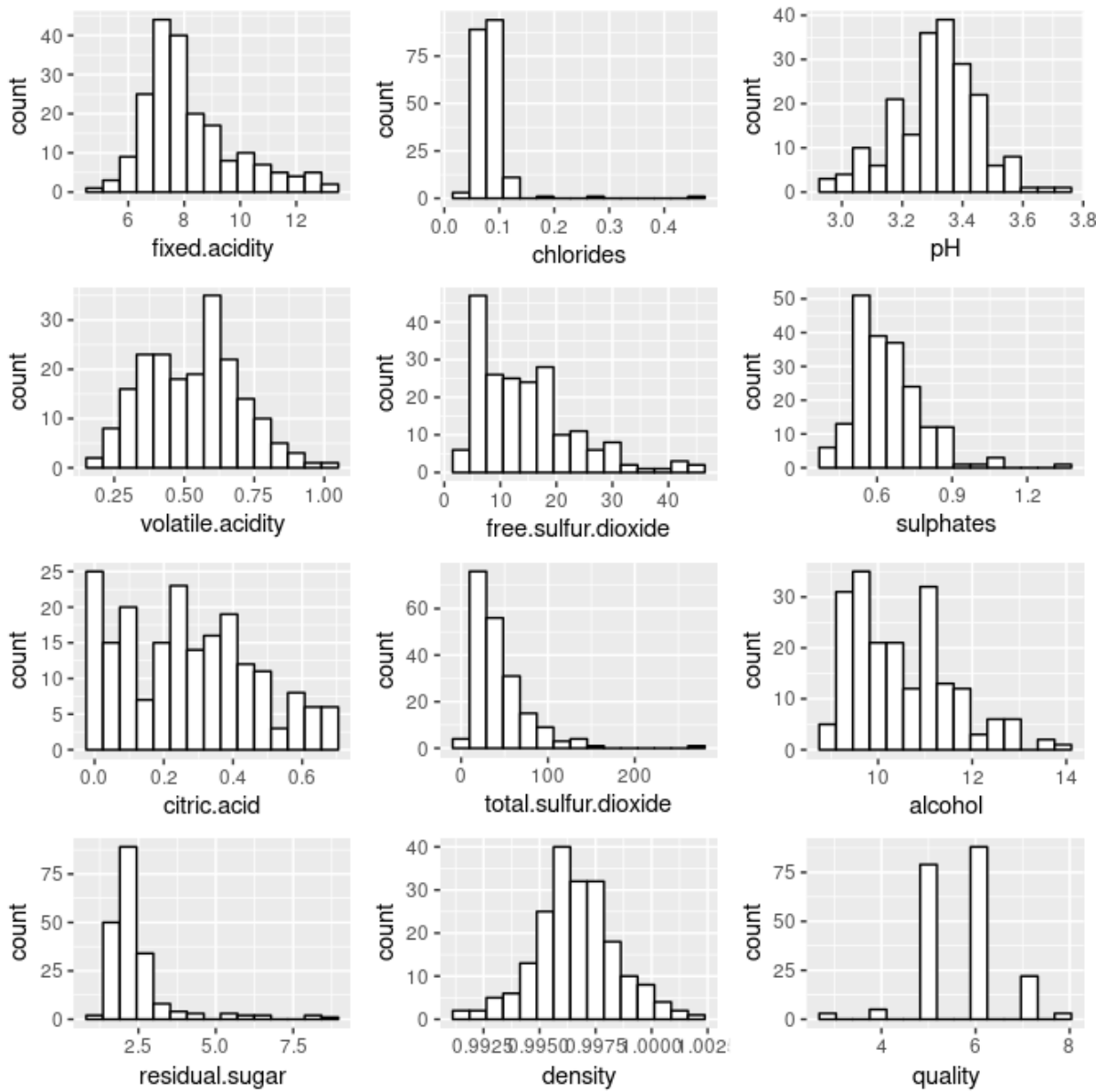


Figure 1: Histogrammes



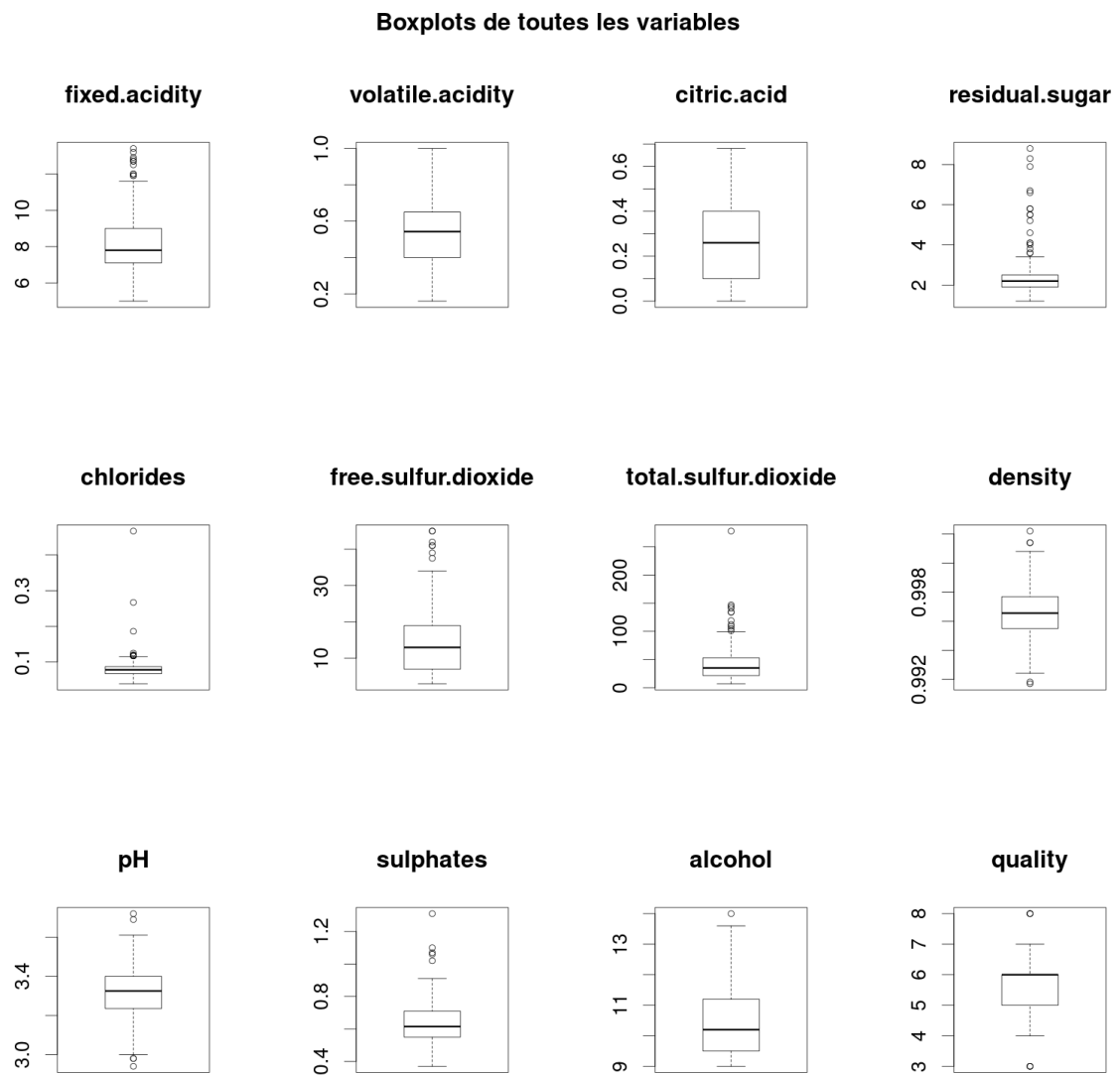


Figure 2: Boxplots

Quantile-quantile plot of normal distribution

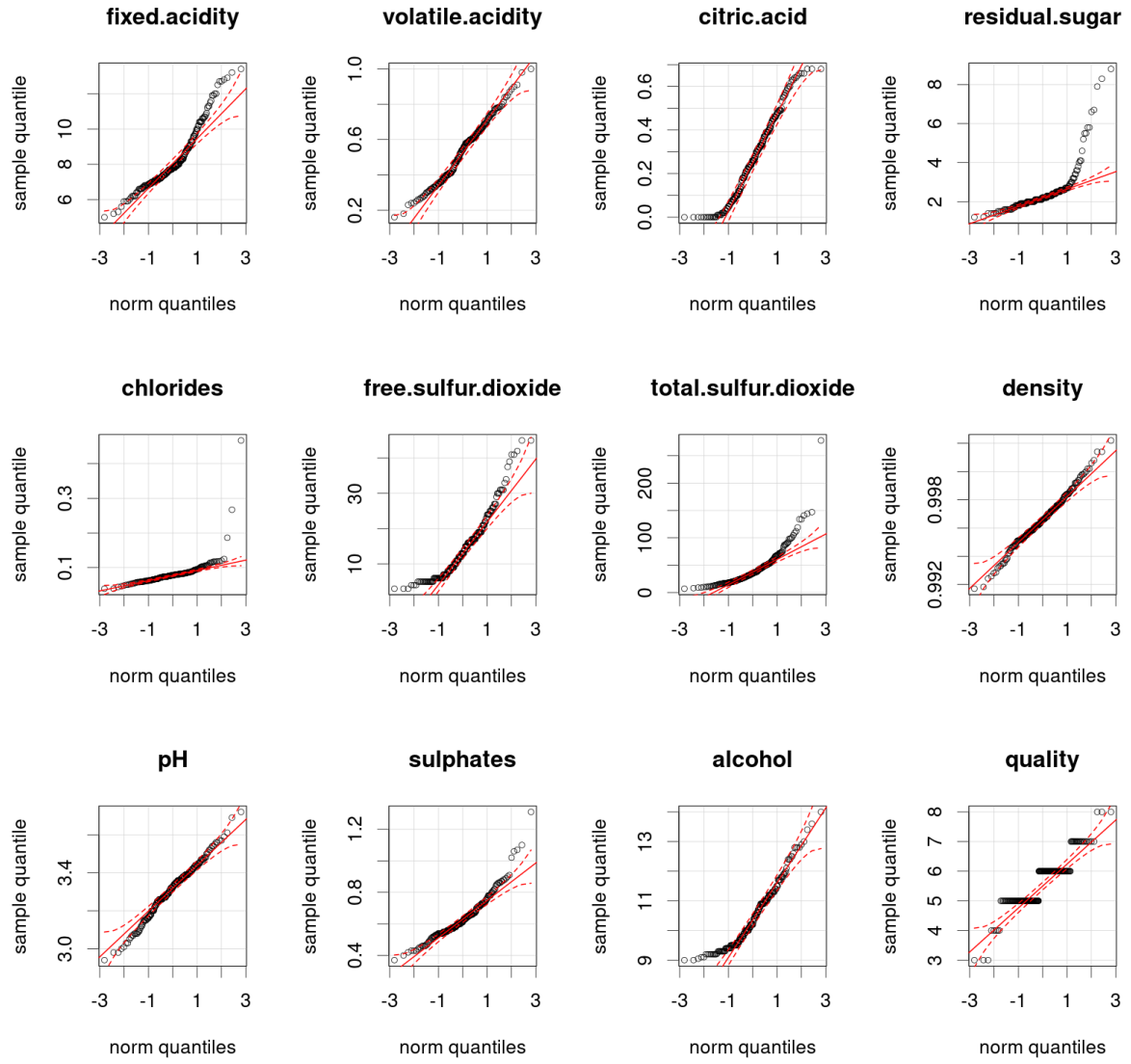


Figure 3: Q-Q Plots

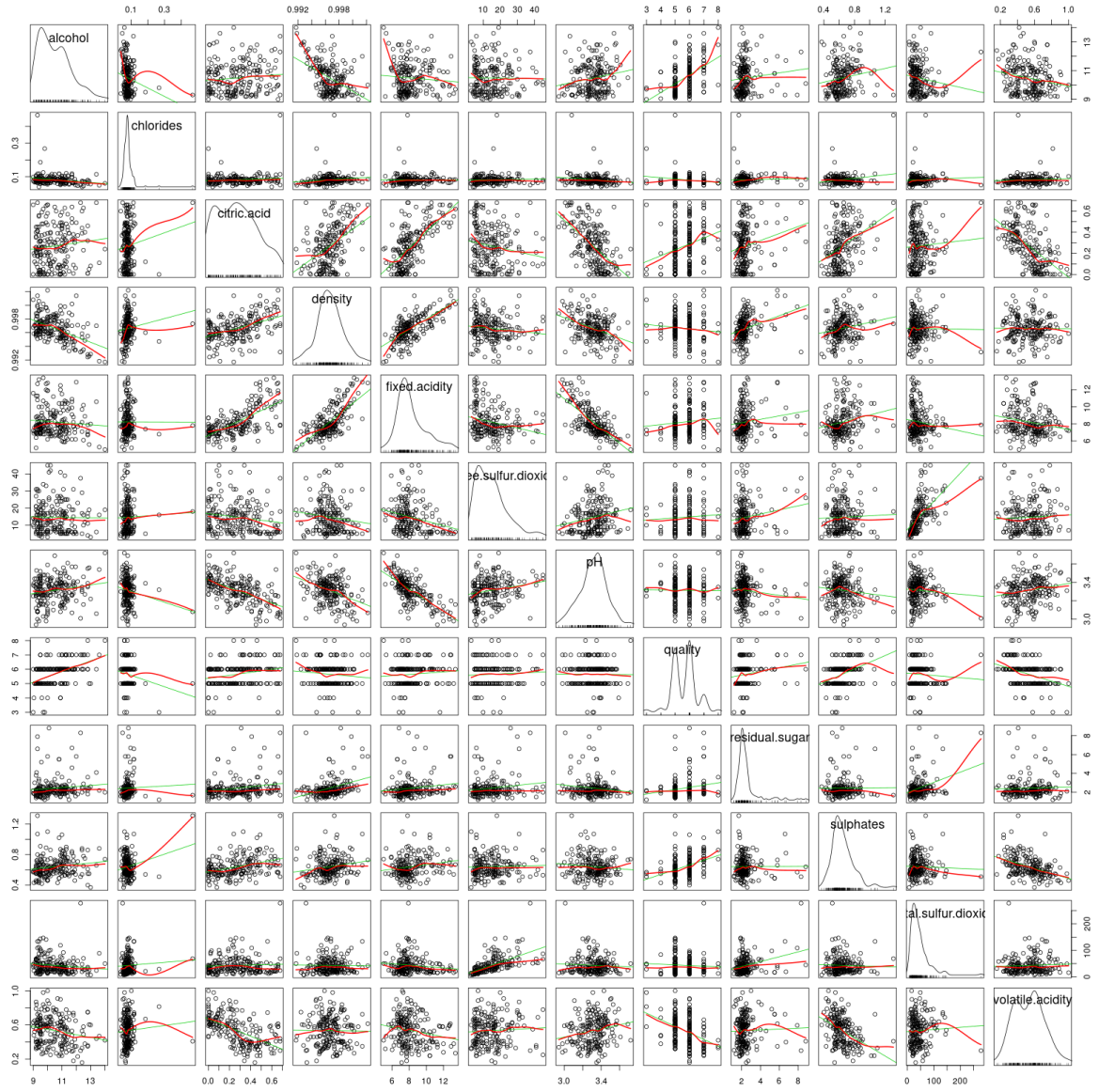


Figure 4: Scatterplot matrix

	Statistic	p.value
fixed.acidity	0.916	0.000
volatile.acidity	0.985	0.029
citric.acid	0.954	0.000
residual.sugar	0.627	0.000
chlorides	0.481	0.000
free.sulfur.dioxide	0.904	0.000
total.sulfur.dioxide	0.778	0.000
density	0.993	0.411
pH	0.991	0.286
sulphates	0.925	0.000
alcohol	0.933	0.000
quality	0.857	0.000

Table 3: Test de Shapiro-Wilk de normalité des variables

## C Code R

*##Code ici*