

Problème d'analyse de données

Fabrice Rossi

Il est **très vivement** conseillé de lire l'intégralité du sujet avant de répondre aux questions. Il est aussi **très vivement** conseillé de tenir compte des analyses précédentes lors d'une réponse à une question (et donc de traiter l'énoncé dans l'ordre).

Je ne m'attends pas à des réponses complètes à l'intégralité des questions. J'attends des réponses concises et précises.

1 Description des données

On étudie dans ce problème un jeu de données constitué de 178 échantillons décrits par 13 variables numériques. Chaque échantillon correspond à un vin italien. Les variables sont issues d'analyses physico-chimiques pratiquées sur le vin. On dispose aussi d'une variable qualitative à trois modalités qui décrit le type du vin (il ne s'agit pas d'une mesure de qualité du vin, les modalités ne sont pas ordonnées).

Voici quelques remarques sur les variables, aucune expertise sur le sujet n'est nécessaire pour répondre correctement aux questions. La numérotation des variables est la même dans tout l'énoncé (cf. par exemple les figures 1 et 2).

1. **Type** : trois variantes de vignes sont considérées dans cette étude ;
2. **Alcool** : c'est le taux d'alcool du vin ;
3. **Acide malique** : l'acide malique est présent naturellement dans le jus de raisin, il disparaît en partie lors de la fermentation malolactique ;
4. **Cendres** : le vin contient toujours des cendres dont la quantité en grammes par litre est indiquée par cette variable ;
5. **Alcalinité** : cette variable mesure le pouvoir réducteur des cendres du vin, c'est une indication de sa minéralisation ;
6. **Magnesium** : la quantité de magnesium est une autre indication de la minéralisation du vin ;
7. **Phénols** : cette variable indique la quantité totale des molécules de la famille des phénols contenue dans le vin (les phénols sont des composés aromatiques) ;
8. **Flavonoïdes** : les flavonoïdes sont une famille de molécules généralement colorées en rouge, dont certaines sont des phénols ;
9. **Phénols non flavonoïdes** : mesure la quantité de phénols qui ne tombent pas dans la classe des flavonoïdes ;
10. **Proanthocyanidines** : il s'agit de flavonoïdes particulier, les tanins du vin ;
11. **Intensité de couleur** : mesure l'inverse de la transparence du vin ;
12. **Teinte** : mesure la couleur (au sein usuel) du vin, sans tenir compte de l'intensité (la teinte est la *hue* en anglais) ;
13. **OD280/OD315** : il s'agit du ratio d'absorbances d'un échantillon de vin à deux longueurs d'onde bien choisies. Il est utilisé pour mesurer la quantité d'acide nucléique (ARN/ADN) dans le vin ;
14. **Proline** : cette variable mesure la quantité de l'acide aminé L-Proline contenu dans l'échantillon.

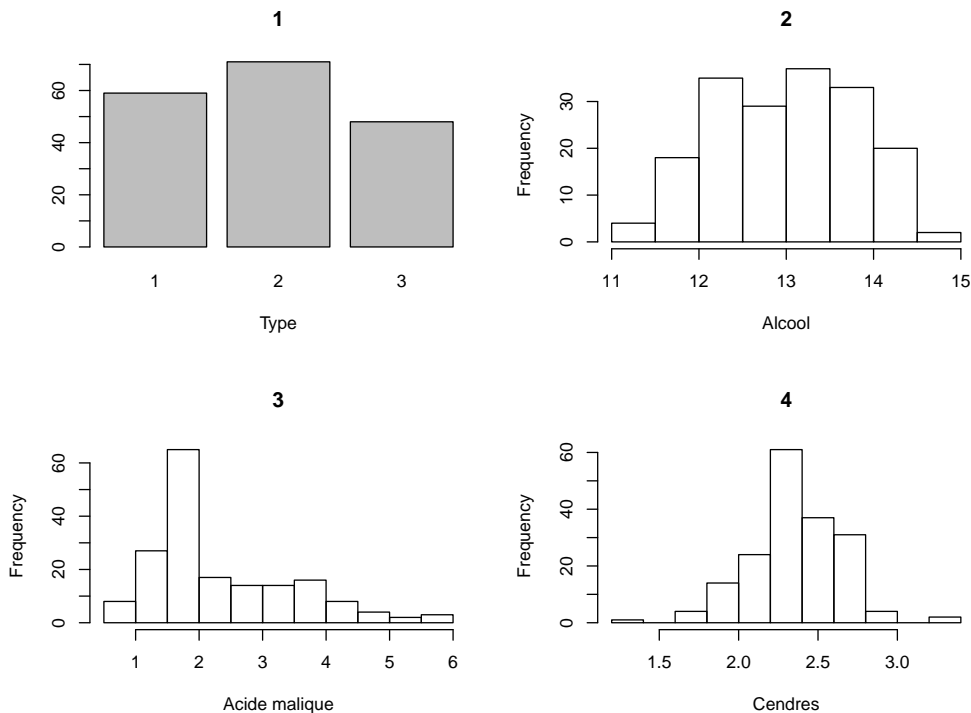


FIG. 1 – Histogrammes des données vin

2 Analyse univariée

On commence par étudier les variables de façon séparées. Les figures 1 et 2 représentent l'ensemble des variables sous forme d'histogrammes (ou de diagramme à bâtons). La séparation en deux groupes d'histogrammes n'a aucun sens particulier, elle répond uniquement à une contrainte de place sur le papier.

Question 1 Commenter les histogrammes des variables numériques en répondant aux questions suivantes :

1. identifiez les variables dont la distribution est symétrique.
2. indiquez les distributions qui font apparaître des points atypiques.
3. identifiez les variables unimodales et multimodales.
4. pour chaque statistique proposée dans la liste qui suit, indiquez si elle pourrait être utile pour résumer la distribution de la variable : moyenne, variance, médiane, intervalle interquartile.
5. pour chaque variable, proposez une statistique intéressante, par exemple la probabilité d'observer une valeur inférieure à un certain seuil (vous n'utiliserez pas les statistiques de la question précédente).

3 Analyses multivariées

3.1 Corrélation

La figure 3 représente les valeurs absolues des corrélations entre toutes les variables numériques (avec la numérotation utilisée dans les figures 1 et 2) sous forme de niveaux de gris : plus la case

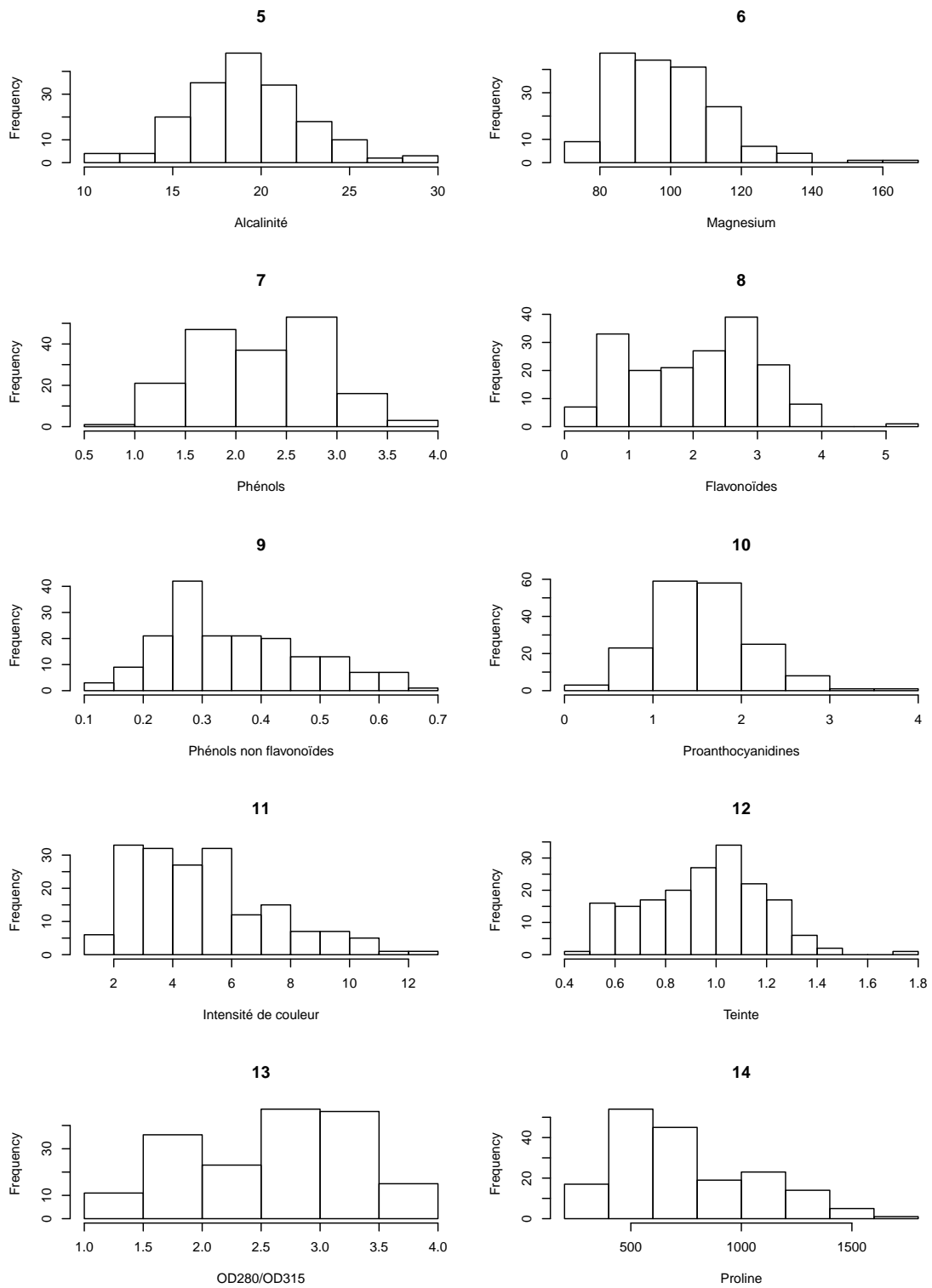


FIG. 2 – Histogrammes des données vin

est foncée, plus la valeur absolue de la corrélation est proche de 1. La diagonale a été laissée blanche pour ne pas perturber la lecture des autres cases. La figure 4 donne un exemple de la correspondance entre niveaux de gris et valeurs numériques.

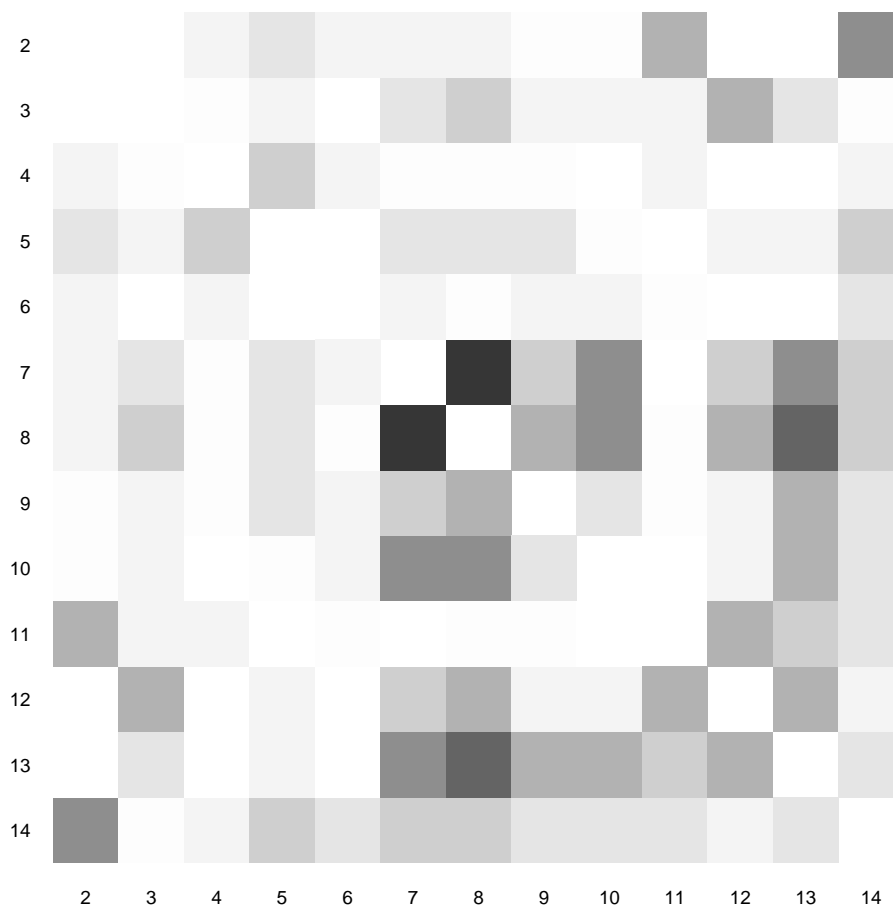


FIG. 3 – Valeurs absolues des corrélations

Question 2 *L'analyse de corrélation n'inclut pas la variable qualitative de type (numéro 1). Pourquoi ?*

Question 3 *Quelles variables pourraient être intéressantes dans le cadre d'une analyse bivariée ? Il est conseillé de tenir compte ici des résultats de l'analyse univariée ainsi que de la description de variables. En vous appuyant sur cette dernière indiquez en particulier quelles sont les corrélations attendues.*

3.2 Analyse bivariée

La figure 4 représente les corrélations de la variable Flavonoïdes avec toutes les autres variables, en utilisant les mêmes niveaux de gris que sur la matrice 3. La figure 5 représente les données selon deux variables. La droite sur la figure correspond au meilleur modèle de prédiction des Flavonoïdes par les Phénols.

Question 4 *Commenter la figure 4. Faites en particulier le lien avec vos réponses à la question 3.*

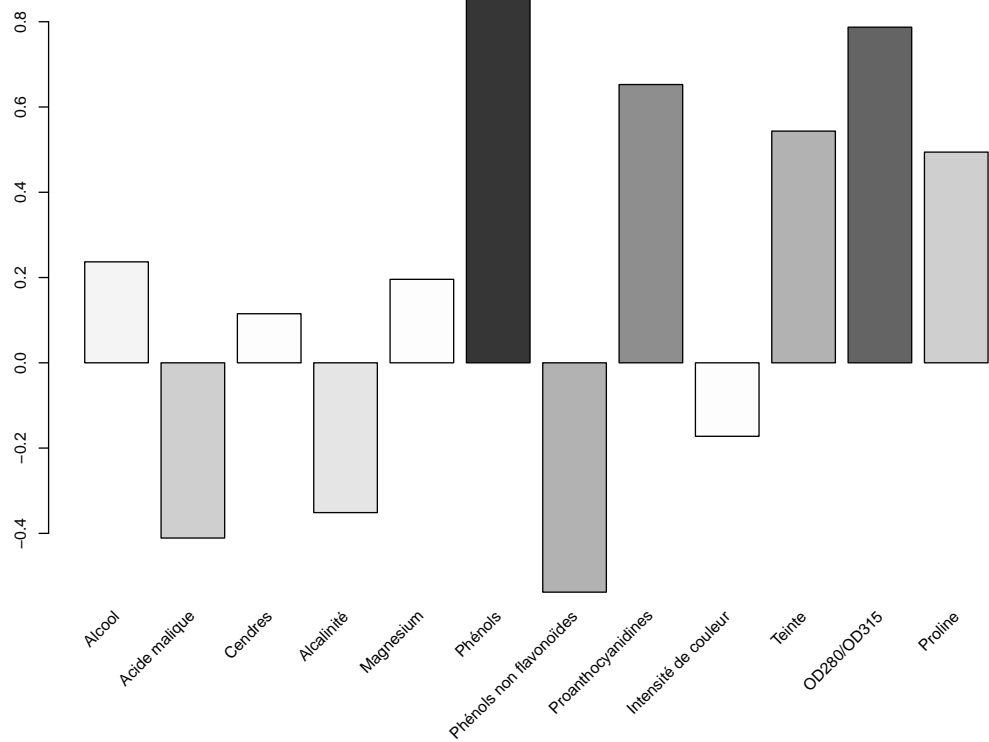


FIG. 4 – Corrélations de la variable Flavonoïdes avec toutes les autres variables numériques

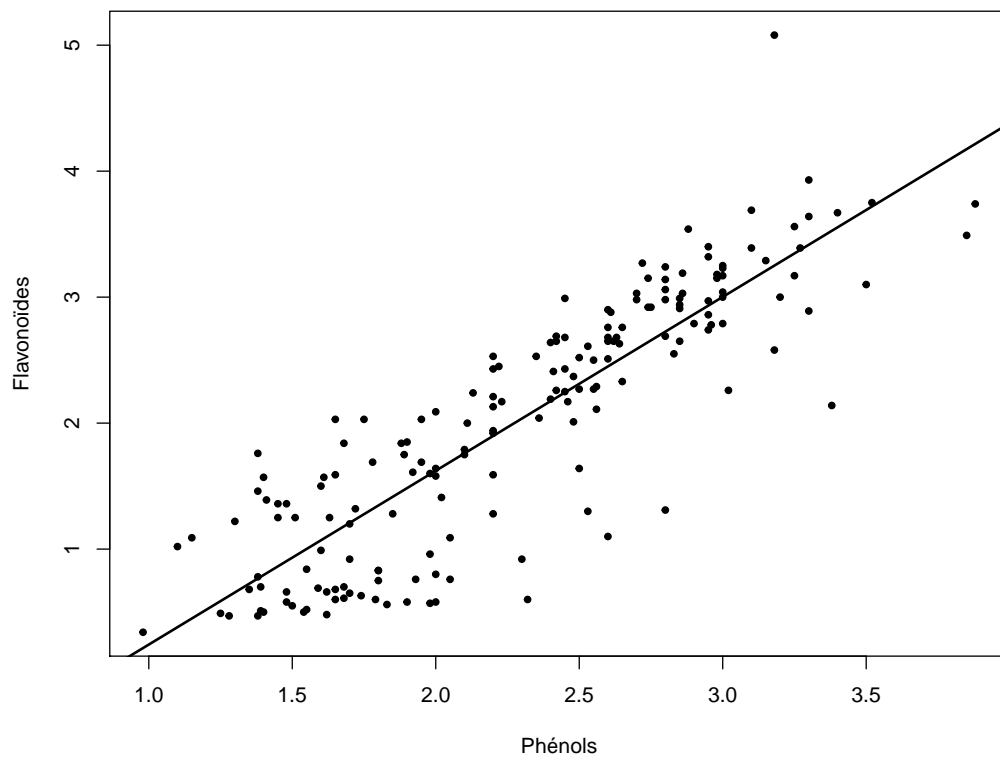


FIG. 5 – Diagramme de dispersion Flavonoïdes/Phénols

Question 5 Commenter la figure 5, en particulier en comparant à ce qui pouvait être attendu à l'observation des figures 3 et 4.

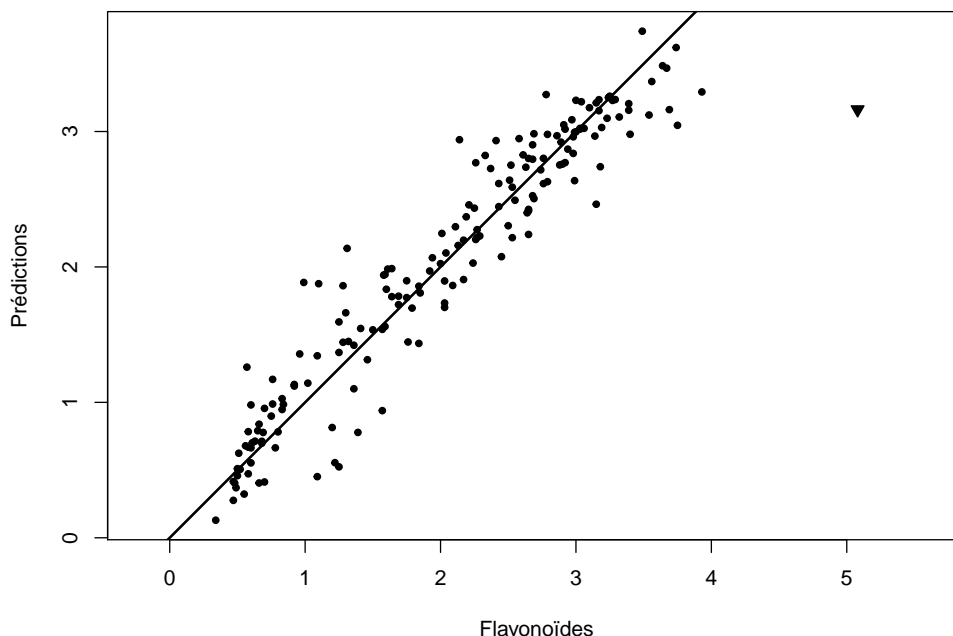


FIG. 6 – Prévisions d'un modèle linéaire des Flavonoïdes construit à partir de toutes les autres variables numériques. La droite est celle d'équation $y = x$.

Question 6 La figure 6 représente les prévisions d'un modèle linéaire qui tente d'expliquer la variable Flavonoïdes en fonction de toutes les autres variables numériques. Commenter les résultats obtenus. Comparer en particulier avec la figure 5.

Le triangle en haut à droite de la figure correspond à une observation choisie parmi tout le jeu de données, dont la description numérique est donnée ci-dessous.

Type	Alcool	Acide malique	Cendres	Alcalinité	Magnesium	Phénols	Flavonoïdes
2	11.56	2.05	3.23	28.5	119	3.18	5.08
Phénols non flavonoïdes		Proanthocyanidines	Intensité de couleur	Teinte			
		0.47	1.87	6	0.93		
OD280/OD315	Proline						
3.69	465						

Ce vin est-il typique (prendre en compte les analyses précédentes et justifier la réponse) ?

3.3 Analyse sur trois variables

Question 7 La figure 7 donne un autre diagramme de dispersion des données portant sur deux variables numériques, le ratio OD280/OD315 et la quantité de Proline. La variable de type est représentée grâce à des symboles. Commenter le diagramme. On pourra d'abord étudier les aspects bivariés (en liaison avec la figure 3) puis passer au trivarié.

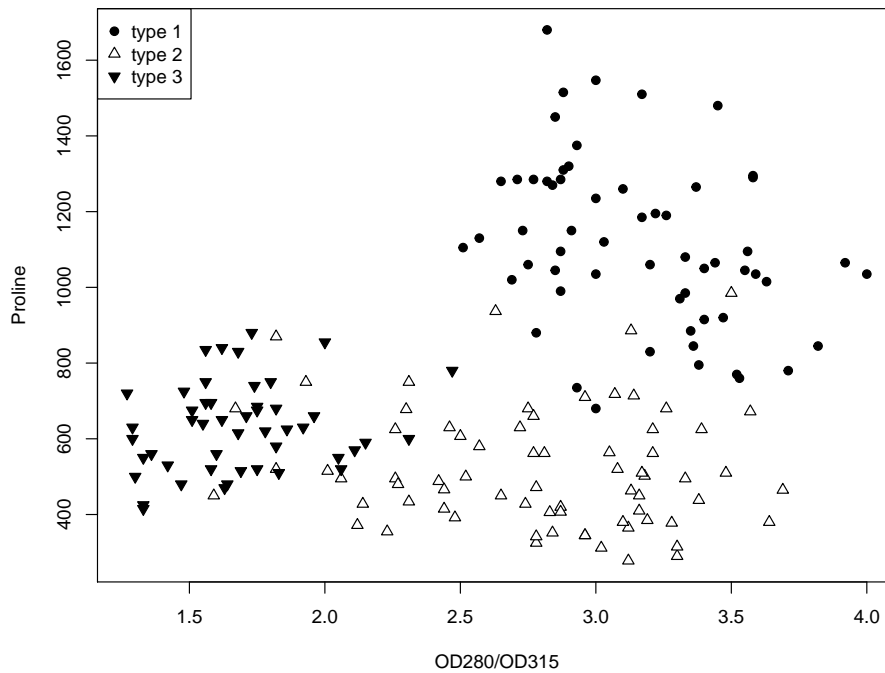


FIG. 7 – Diagramme de dispersion Proline/OD280/OD315. Le type de chaque vin est indiqué selon la légende.

3.4 Analyse en composantes principales

On réalise une analyse en composantes principales des données, en excluant la variable qualitative. Les variances expliquées sont les suivantes :

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	2.169	1.580	1.203	0.9586	0.9237	0.8010	0.7423	0.5903
Proportion of Variance	0.362	0.192	0.111	0.0707	0.0656	0.0494	0.0424	0.0268
Cumulative Proportion	0.362	0.554	0.665	0.7360	0.8016	0.8510	0.8934	0.9202
	PC9	PC10	PC11	PC12	PC13			
Standard deviation	0.5375	0.5009	0.4752	0.4108	0.32152			
Proportion of Variance	0.0222	0.0193	0.0174	0.0130	0.00795			
Cumulative Proportion	0.9424	0.9617	0.9791	0.9920	1.00000			

Question 8 *Combien de composantes faudrait-il conserver pour ne pas perdre trop d'information ? Que dire a priori d'une représentation sur les deux premiers axes ? Commenter la représentation obtenue (Figure 8) ainsi que le Biplot associé (Figure 9), notamment en comparant avec les figures 3 et 7.*

4 Classification

On conduit finalement une classification hiérarchique sur les données, en excluant toujours la variable de type de la construction du résultat.

Question 9 *La figure 10 représente deux dendrogrammes obtenus respectivement en utilisant le critère du lien simple (dendrogramme de gauche) et le critère de Ward (dendrogramme de droite). Commenter les résultats. Quel critère vous semble le plus pertinent ?*

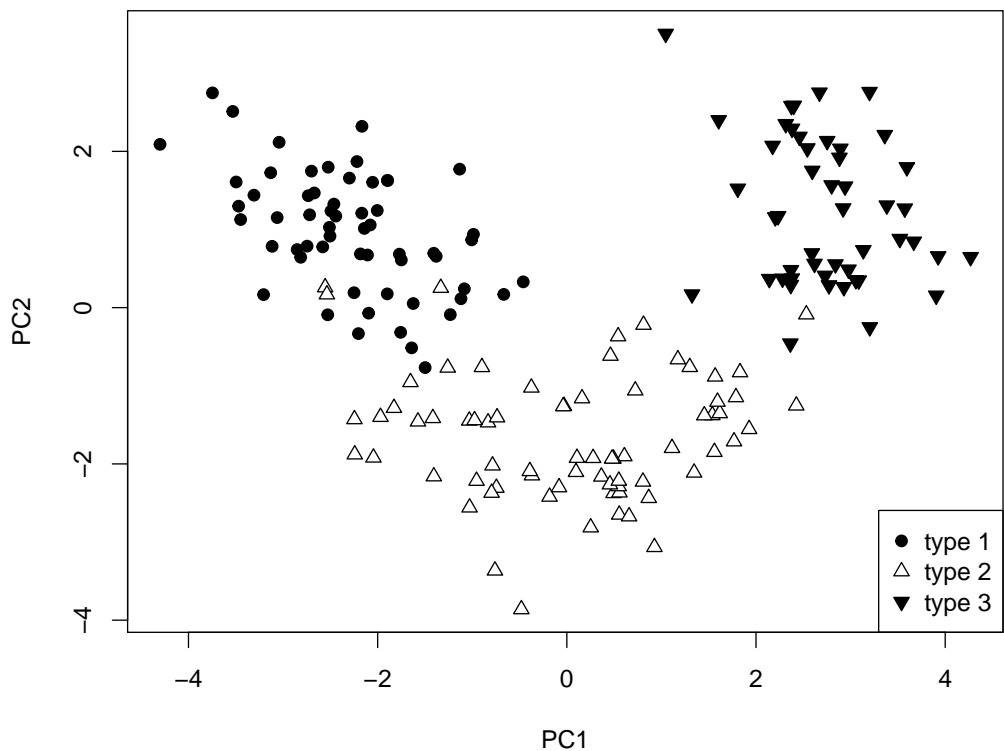


FIG. 8 – Représentation des données selon les deux premières composantes. Le type de chaque vin est indiqué selon la légende.

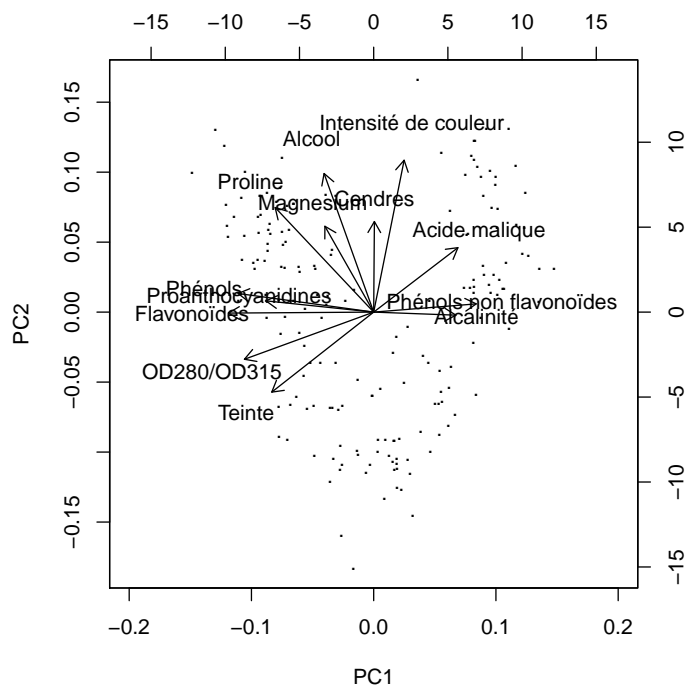


FIG. 9 – Biplot de l'Analyse en Composantes Principales

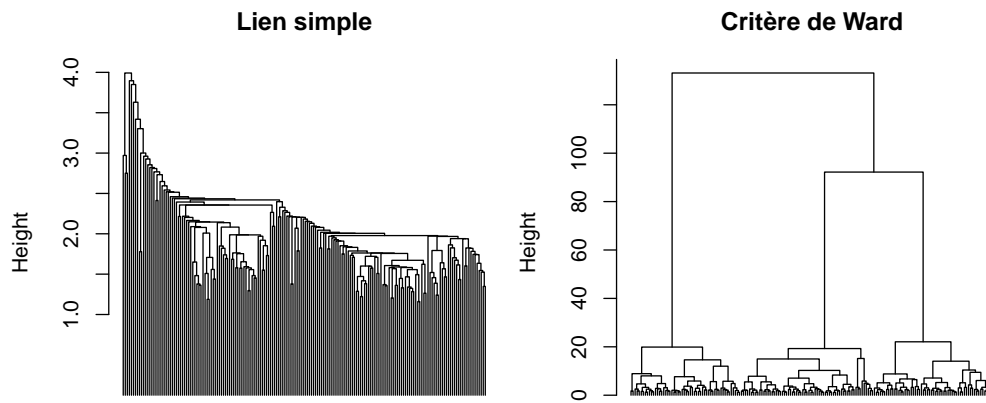


FIG. 10 – Dendrogrammes de deux classifications hiérarchiques des données

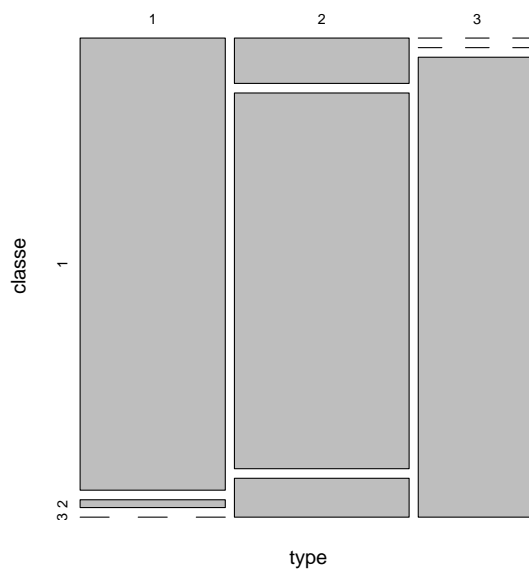


FIG. 11 – Mosaicplot pour le type de vin et la classe

Question 10 *En utilisant le dendrogramme associé au critère de Ward, on choisit de faire une partition des données en trois classes. Commenter ce choix.*

Question 11 *La figure 11 est un mosaïcplot de deux variables qualitatives : le type du vin (en abscisse) et la classe du vin telle qu'elle a été obtenue dans la question précédente (en ordonnée). Commenter le diagramme.*

Question 12 *On souhaite construire automatiquement un modèle déterminant le type du vin en fonction des 13 variables numériques. Estimer la difficulté d'une telle tâche en utilisant les analyses conduites dans les questions précédentes. Peut-on en particulier identifier des variables numériques qui expliquent bien le type du vin ? Que dire des composantes principales ?*