Projet d'analyse de données

Qu'est ce qui fait un bon étudiant ?

FIDA CYRILLE Rudio et BAUDET Léo-Paul
MAIN4 - Polytech Sorbonne

2023-05-09

On a beaucoup d'idées reçues par rapport aux facteurs qui feraient d'un étudiant un bon étudiant, notamment sur l'alcool. Le jeu de données que nous avons étudié permet alors de confronter ces idées que nous nous faisons à des données concrètes.

1 Présentation du jeu de données.

Le jeu de données, nommé "Student alcohol consumption", est constitué d'informations sur la vie d'étudiants dans un lycée du Portugal. Ces informations vont de leur résultats universitaires ou de leur vie familiale à leur consommation d'alcool. Le jeu a été construit à partir d'une enquête menée auprès d'étudiants en mathématiques et en portugais.

L'objectif serait alors d'analyser le jeu de données afin de comprendre les facteurs qui impactent la réussite scolaire de ces étudiants. L'intérêt du jeu est la grande variété de facteurs proposée qui permet de couvrir un maximum d'hypothèses, notamment celles sur la consommation d'alcool proposée directement par le nom du jeu de données.

Voici les variables présentent dans ce jeu de données ;

- **school** - école (binaire: 'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira)
- **sex** - sexe (binaire: 'F' - female ou 'M' - male)
- **age** - age (numérique: de 15 à 22)
- **address** - adresse (binaire: 'U' - urbain or 'R' - rural)
- **famsize** - taille de la famille (binaire : 'LE3' - inférieur ou égal à3 or 'GT3' - supérieur à 3)
- **Pstatus** - parents qui habitent ensemble ? (binaire: 'T' - ensemble or 'A' - séparés)
- **Medu** - niveau d'études de la mère (numerique: 0 - vide, 1 - primaire (4th grade), 2 – collège, 3 – lycée or 4 – supérieur)
- **Fedu** - niveau d'études du père (numerique: 0 - vide, 1 - primaire (4th grade), 2 – collège, 3 – lycée or 4 – supérieur)
- **Mjob** - travail de la mère (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- **Fjob** - travail du père (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- **reason** - raison derrière le choix d'école (nominal:'home', school 'reputation', 'course' preference ou 'other')
- **guardian** - représentant légal (nominal: 'mother', 'father' ou 'other')
- **traveltime** - temps de trajet (numérique en min)
- **studytime** - temps d'étude hebdomadaire (numérique en h)
- **failures** - nombre d'échecs (numeric: n if 1<=n<3, else 4)
- **schoolsup** - aide scolaire supplémentaire (binaire : yes ou no)
- **famsup** - support famillial (binaire : yes ou no)

- **paid** - cours supplémentaires (binaire : yes ou no)
- **activities** - activités extra-scolaires (binaire : yes ou no)
- **nursery** - est allé à la crèche (binaire : yes ou no)
- **higher** - volonté de poursuite d'études (binaire : yes ou no)
- **internet** - accès à internet (binaire : yes ou no)
- **romantic** - en couple ? (binaire : yes ou no)
- **famrel** - états des relation famillale (numérique: de 1 - très mauvais à 5 - excellent)
- **freetime** - temps libre après les cours (numérique: de 1 - très mauvais à 5 - excellent)
- **goout** - sortie entre amis (numérique: de 1 - très bas à 5 - très élevé)
- **Dalc** - consommation d'alcool en semaine (numérique: de 1 - très basse à 5 - très élevée)
- **Walc** - consommation d'alcool le week-end (numérique: de 1 - très basse à 5 - très élevée)
- **health** - état de santé (numeric: from 1 - very bad to 5 - very good)
- **absences** - nombre d'absences (numérique: de 0 à 93)
- **G1** - note du 1^{er} (numérique: de 0 à 20)
- **G2** - note du $2^{ème}$ (numérique: de 0 à 20)
- **G3** - note du $3^{ème}$ (numérique: de 0 à 20)

Au cours de ce projet, nous nous concentrons sur la moyenne des étudiants qui est à calculer et représente la note des élèves sur l'année. En addition, nous nous intéressons aussi à la réussite scolaire des élèves sur l'année qui va directement découler de leur moyenne. Il s'agirait donc ici d'étudier un problème de classification supervisée sur la réussite et de régression sur la moyenne. Le but final serait alors d'avoir une meilleure compréhension des facteurs qui impacteraient la réussite scolaire et de les confronter à nos propres expériences en tant qu'étudiants.

Dans un premier temps, nous avons étudié chaque variable notamment leur corréléation avec la moyenne et la réussite. Ensuite, nous avons mis en place des modèles de régression linéaire poir prédire la moyenne annuelle des élèves. Pour finir, nous avons utilisé des méthodes et comparé des méthodes de Machine Learning dans le but de prédire la réussite des élèves.

Voici les bibliothèques à installer : ggplot2, FactoMineR, pROC, MASS, randomForest, gbm, gridExtra, dplyr, klaR, rpart, rpart.plot, corrplot, GGally, glmnet, e1071

2 Les données

2.1 Chargement des données

Le dataset est composé de 2 fichiers csv représentant les élèves de portugais et de maths. Il faut donc concaténer les deux jeux de données pour obtenir le jeu final. On peut noter qu'il y a 382 élèves qui suivent les deux cours. De base, il contient 33 variables dont 8 quantitatives et 25 qualitatives.

2.2 Nettoyage et vérification des données

Afin d'adapter le jeu de données à notre étude, nous l'avons modifié. Nous avons notamment modifié en amont les variables *traveltime* et *studytime* afin de les rendre numérique.

On transforme les variables qualitatives en factor et on vérifie que le jeu ne contient pas de NaN.

Pour préparer concrètement les données, nous avons calculé la moyenne pour chaque élève (variable Moy), et nous avons rajouté une variable pour la réussite scolaire (variable RS). On garde 3 modalités différentes pour RS :

1. "admission" pour des Moyennes supérieures à 10
2. "redoublement" pour des Moyennes entre 8.50 et 10
3. "exclusion" pour des Moyennes inférieures à 8.50

On a choisi cette séparation étant donné qu'elle est plus intéressante à étudier qu'une simple variable binaire (on a essayé). Cette répartition a été calcquée sur celle appliquée en France pour le lycée.

3 Exploration des données : analyse des variables

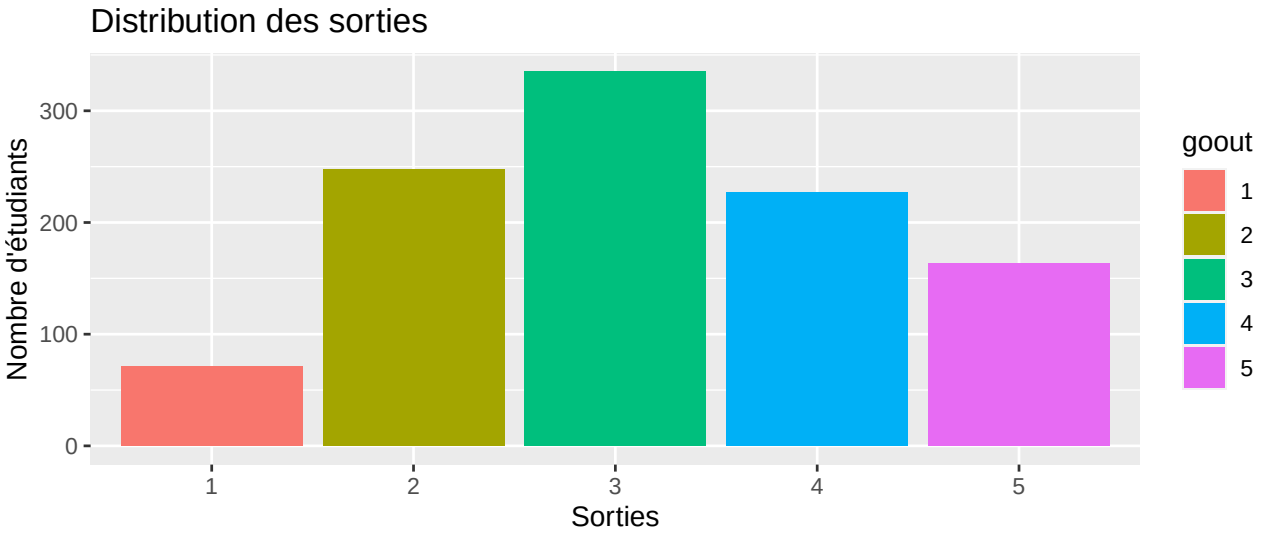
Cette partie consiste à appliquer d'abord des méthodes de statistiques descriptives afin de mieux comprendre le jeu de données et d'analyser les variables qui nous semblent intéressantes. La suite de l'analyse consiste alors à vérifier la corrélation des variables avec la moyenne et la réussite scolaire. On présente dans cette partie les résultats sur les variables les plus intéressantes, le reste des résultats est disponible en annexe mais nous n'avons pas traité toutes les variables étant donné leur nombre conséquent. Les sorties sont également présentées en annexe.

3.1 Les variables qualitatives

Pour chaque variable, nous étudions sa distribution avec soit un diagramme en bâton soit un diagramme circulaire. On effectue ensuite une ANOVA1 entre la variable et la moyenne pour en connaître l'impact à partir, notamment du test de Fisher. Un test de χ^2 est alors effectué pour vérifier la corrélation entre la variable et la réussite scolaire. S'il y a corrélation, on effectue alors une Analyse Factorielle des Correspondances (AFC).

3.1.1 Les sorties

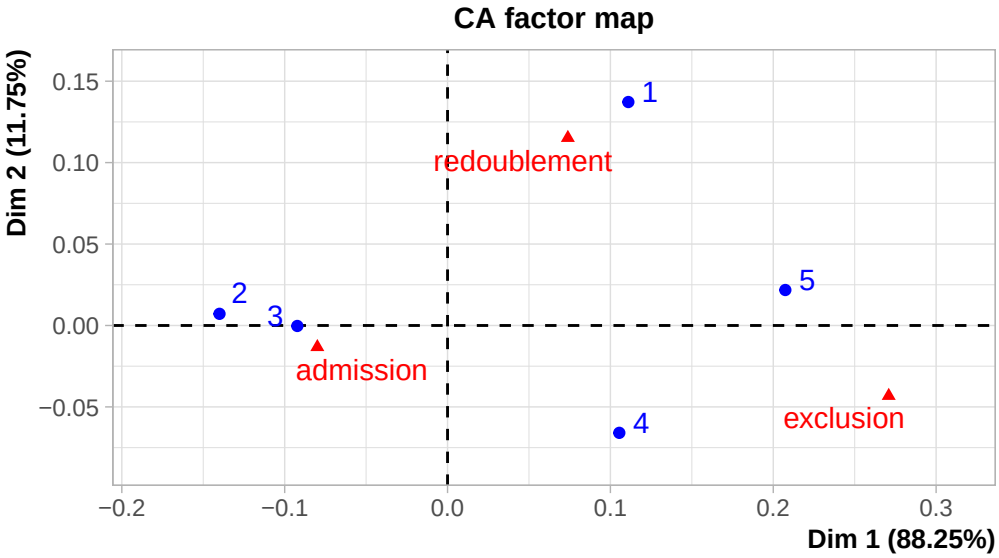
On remarque que les élèves maintiennent leur vie sociale. Le diagramme en bâton nous permet de voir que la majorité des étudiants sortent de manière modéré (modalité 3). Il y a quand même plus de personnes qui sortent vraiment beaucoup que de personnes qui ne sortent pas. Le test de Fisher indique les sorties sont très corrélées au notes et le test de Chi2 montre que la réussite scolaire est aussi corrélée aux sorties. Ainsi, on retrouve des résultats qui semblent cohérents et représentatifs de la vie étudiante.




```
##
## Call:
## lm(formula = Moy ~ goout, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5887  -1.8876  -0.0015   2.1124   7.6652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.5493     0.3770  27.980  < 2e-16 ***
## goout2        1.3727     0.4276   3.210  0.00137 **
## goout3        1.0049     0.4151   2.421  0.01564 *
## goout4        0.4522     0.4320   1.047  0.29548
```

```
## goout5       -0.1853     0.4517  -0.410  0.68178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.177 on 1039 degrees of freedom
## Multiple R-squared:  0.02957,    Adjusted R-squared:  0.02583
## F-statistic: 7.915 on 4 and 1039 DF,  p-value: 2.766e-06
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$goout and df$RS
## X-squared = 20.537, df = 8, p-value = 0.008485
```

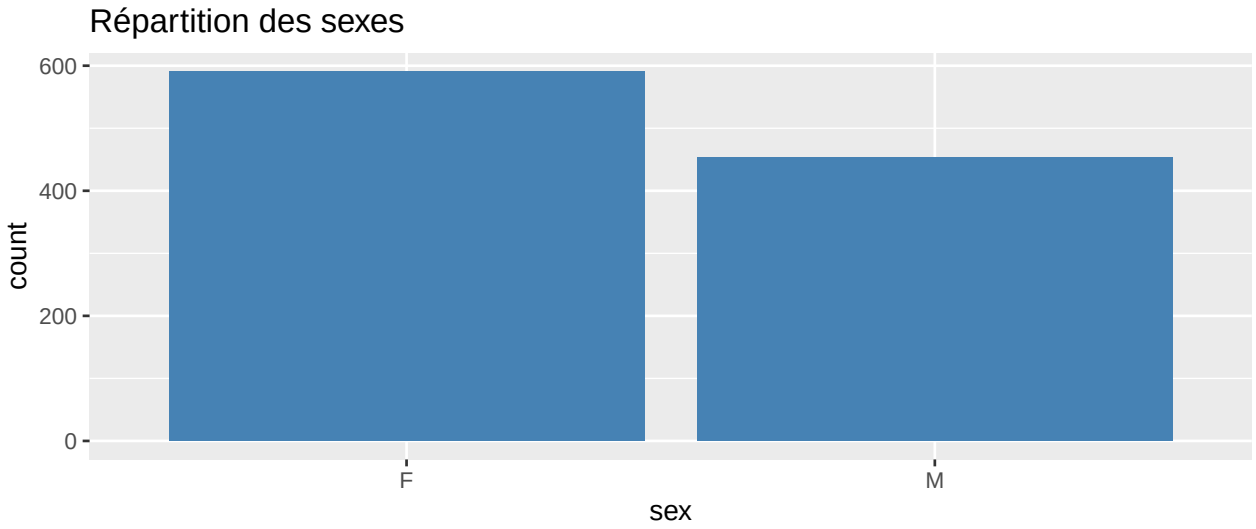


Avec un test du χ^2 on observe que les variables goout et RS sont corrélées (p-valeur petite devant 5%). Nous allons donc réaliser une AFC dessus. Egalement la p-valeur associée au test de fisher (sortie de anova) sur les variables Moy et goout montre que ces grandeurs sont aussi corrélées.

L'AFC nous montre ici que les étudiants qui sortent raisonnablement sont ceux qui réussissent le plus. En effet, ceux qui sortent le plus consacrent moins de temps à leur études ce qui peut expliquer ce résultat. Egalement les étudiants qui ne sortent quasiment pas échouent aussi beaucoup. Ce manque de sortie peut dénoter d'un défaut de socialisation ou des problèmes de santé qui impact gravement la réussite de l'élève. L'AFC nous montre que cela n'est pas un frein à leur réussite.

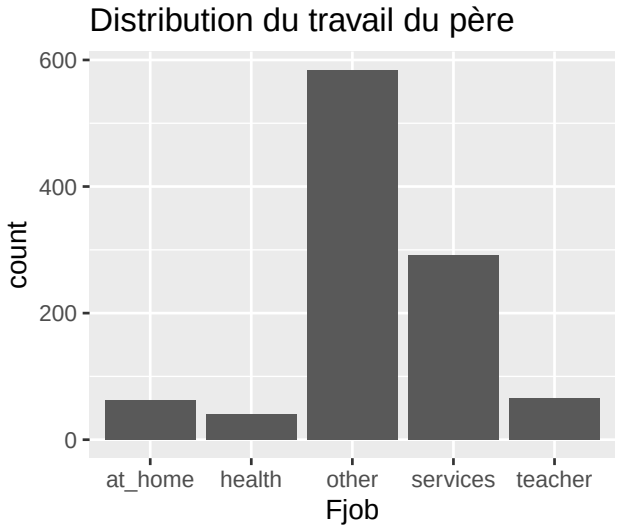
3.1.2 Le sexe des étudiants

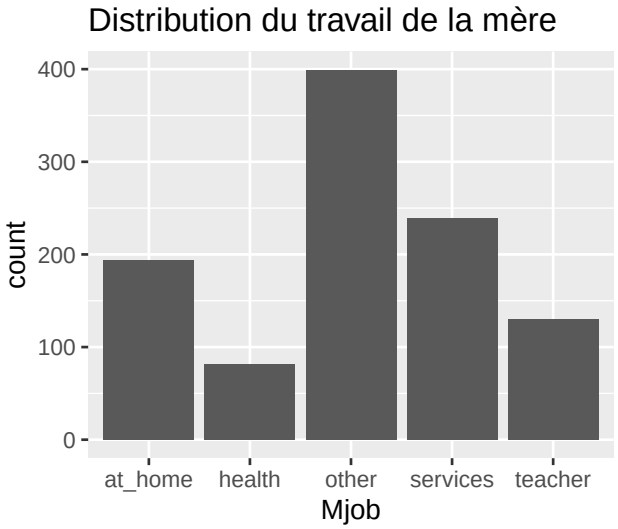
D'après le diagramme, le dataset est plutôt équilibré en terme d'hommes et de femmes,il y même plus de femmes que d'hommes dans ce lycées. On étudie ensuite le lien entre le sexe et les notes en effectant une ANOVA1. D'après le test de Fisher, p-value $> 5\%$ donc il n'y a pas d'effet du sexe sur les notes. D'après le test d'indépendances de Chi2 avec l'admission, le sexe des élèves n'a pas de lien avec leur réussite scolaire, ce qui est plutôt rassurant en terme d'équité.

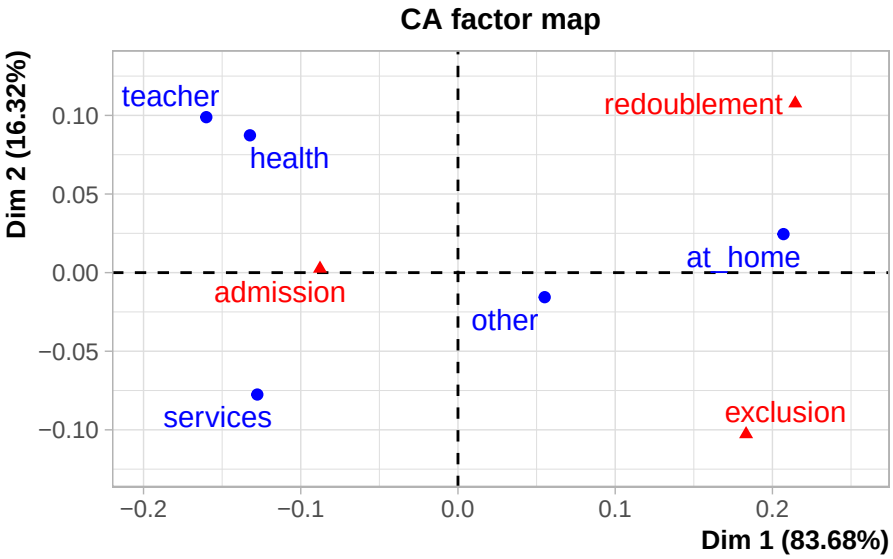


3.1.3 Travail des parents

Dans les deux cas, others et services sont les catégories qui dominent. Une différence notable est la que la proportion de femme au-foyer est bien plus élevée que celle des hommes. D'après le test de Fisher, le travail de la mère a un impact sur les notes, contrairement à celui du père. Les résultats des test de χš suivent les résultats des test de Fisher : le travail de la mère et la réussite scolaire sont bien corrélés mais celui du père n'a pas d'impact, ce qui est assez surprenant.



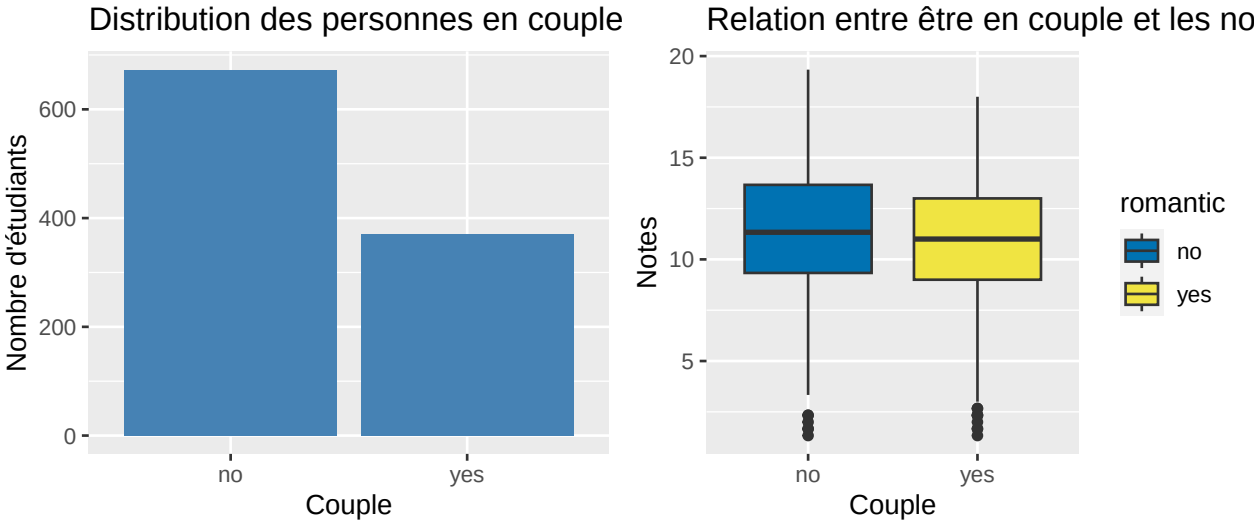




Etant donné les résultats du test de corrélation, on applique une AFC entre les deux variables. On observe alors que les élèves dont les mères travaillent dans la santé, l'enseignement ou les services auront tendances à être en réussite. A l'opposé, les élèves dont les mères sont femmes au foyer seront plutôt en situation d'échec.

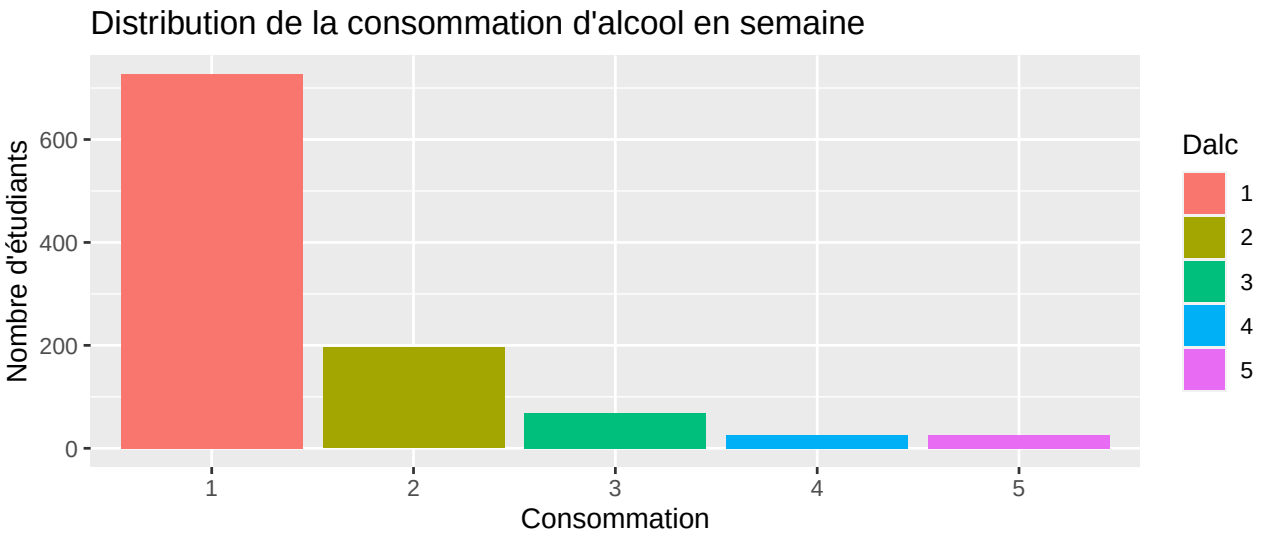
3.1.4 Les relations

Il y a environ deux fois plus de jeunes célibataires que de jeunes en couple. On peut penser qu'être en couple réduit le temps passé à étudier et rajoute des distractions, donc il devrait avoir un impact négatif sur les notes. D'après le test de Fisher, la p-value est fortement inférieure à 5%, donc on rejette H0: il y a bien un lien entre situation romantique et notes, ce qui rejoint bien l'idée de départ. Il serait donc intéressant d'étudier la distribution des notes selon la situation romantique. D'après les boxplots, les différences sont assez minimes, même si on peut apercevoir que les notes des célibataires sont légèrement meilleures. Cependant, d'après le test de χš la présence de relation amoureuse n'a pas d'impact sur la réussite scolaire. Ainsi, être en couple fait baisser la moyenne mais n'est pas un facteur d'échec.

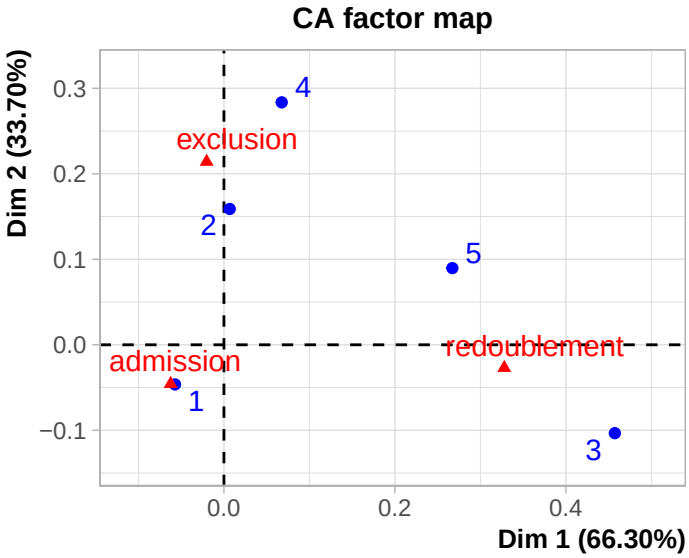


3.1.5 La consommation d'alcool

On s'intérésse enfin à la feature "principale" de ce jeu de données, la consommation d'alcool des étudiants.



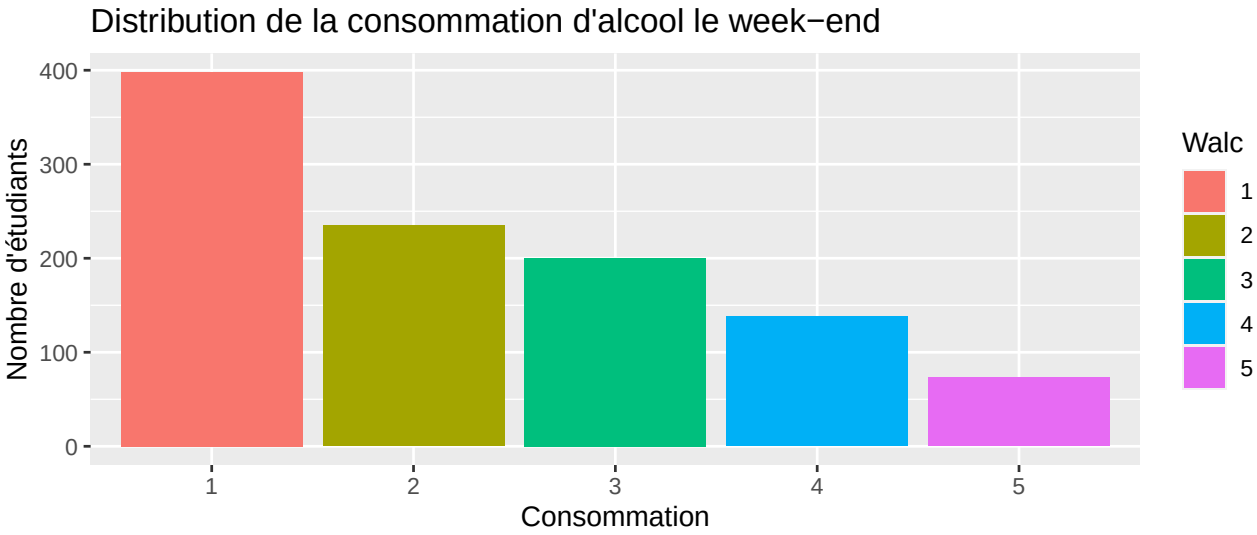

```
## Warning in chisq.test(df$Dalc, df$RS): Chi-squared approximation may be
## incorrect
```

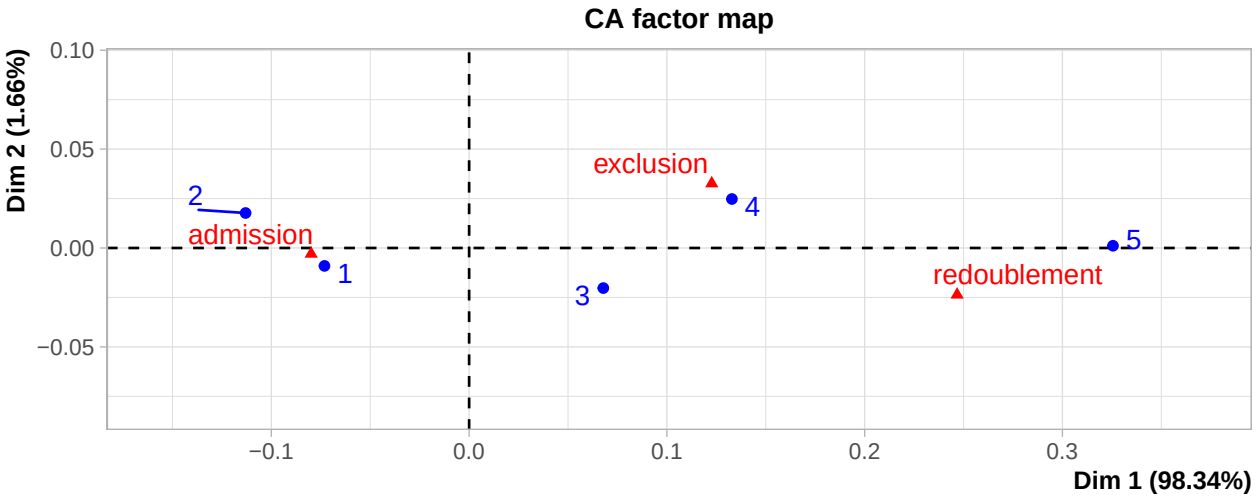


Avec le test du χ^2 on voit que les variables Dalc et RS sont corrélées. On va donc réaliser une AFC dessus. De même avec la p-valeur du test de Fisher sur les variables Moy et Dalc, on voit que ces variables sont aussi corrélées (et c'est logique au vu du test du χ^2).

On voit très clairement avec l'AFC que les étudiants qui consomment le plus d'alcool sont ceux qui réussissent le moins. En effet, une forte consomation d'alcool témoigne d'un grand nombre de sortie ou bien d'un grave problème de santé (alcoolisme). Ceux qui réussissent le plus sont ceux qui consomment le moins d'alcool.

Avec le diagramme en bâton, on voit que la majorité des étudiants ne consomme quasiment pas d'alcool en semaine. L'AFC montre que cela n'a pas du tout été un frein pour leur réussite.





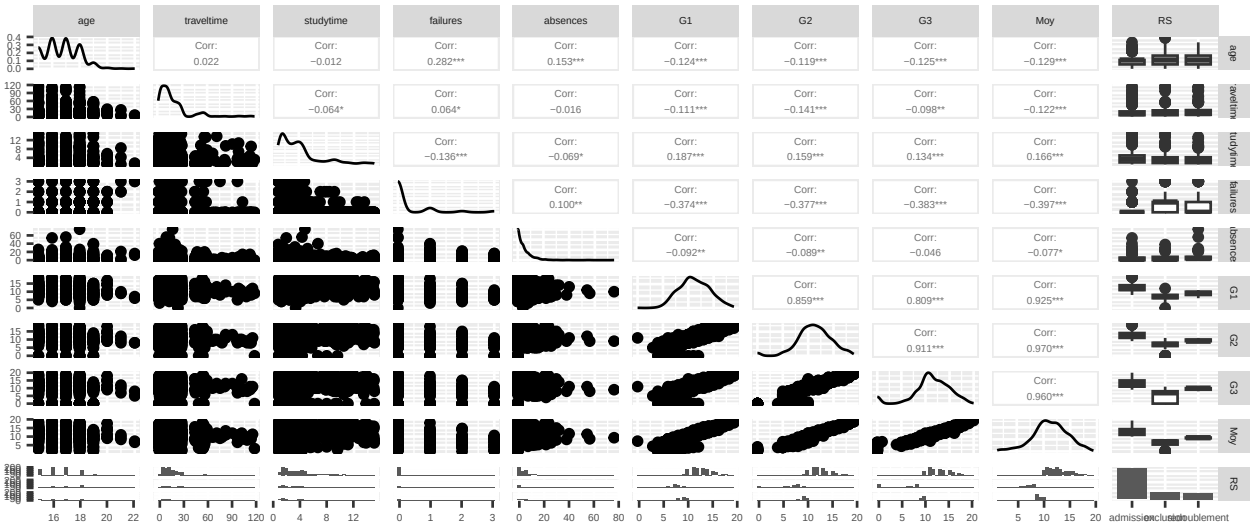
Tout d'abord on obtiens une p-valeur plus petite que 5% avec le test du χ^2 ce qui montre que les variables Walc (consommation alcool le week end) et RS (réussite scolaire) sont corrélées. Nous allons réaliser une AFC dessus afin de mieux les expliquées. De même avec le test de fisher (réalisé à l'aide de l'anova) réalisé sur les variables Walc et Moy montre qu'elles sont corrélées.

De même on obtiens le même résultat avec la consommation d'alcool le week end (ceux qui consomment le moins réussissent le plus), un peu plus nuancé cependant. En effet, on voit à travers les différents diagrammes en bâtons que globalement il y a plus d'étudiants qui consomment de l'alcool le week-end qu'en semaine. On voit donc grâce aux deux AFC que les étudiants qui consomment plutôt de l'alcool le week-end réussissent mieux que les étudiants qui consomment de l'alcool la semaine et le week end. Ainsi la variable avec la modalité 2 témoigne bien du fait que consommer de l'alcool en semaine est bien plus néfaste qu'en consommer en week-end (dans un contexte de soirée). En même temps, la consommation d'alcool en semaine de manière élevée montre des tendances alcooliques de la part des étudiant, notamment lycéens dans notre étude.

3.2 Les variables quantitatives

Dans cette partie, on s'intéresse aux variables quantitatives du jeu de données. De même que pour les variables qualitatives, on cherche à identifier les facteurs qui impactent la moyenne ainsi que la réussite scolaire.

3.2.1 Statistiques descriptives et corrélation



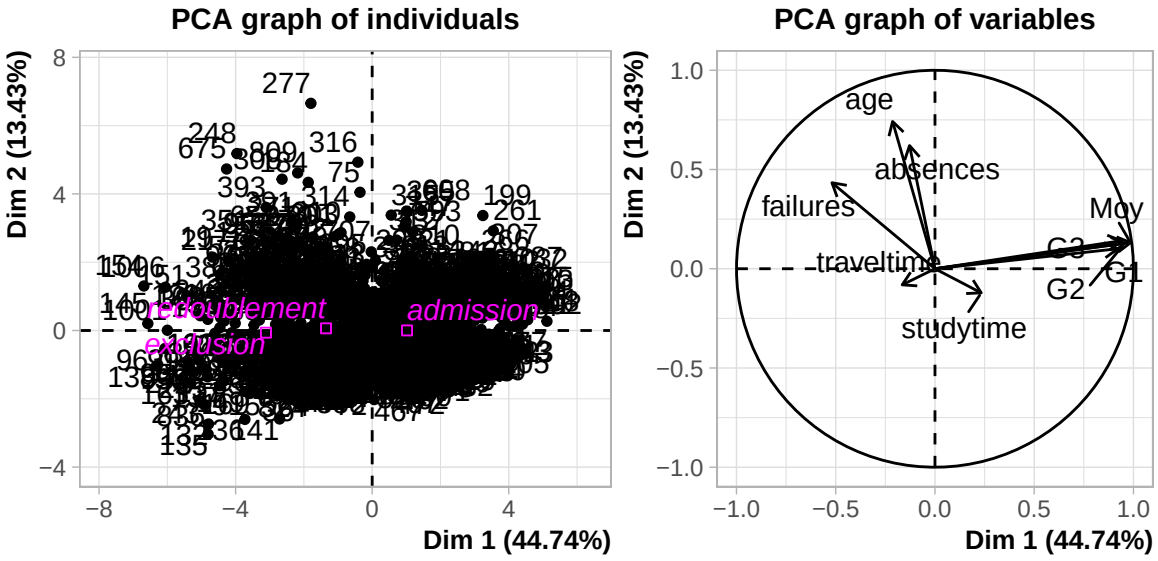
Le graphe nous montre la répartition bivariée des variables quantitatives ainsi que la corrélation entre les variables. On s'intéressera notamment à la corrélation par rapport à la moyenne.

On en déduit les informations suivantes:

- la majorité des étudiants ont entre 15 et 19 ans. L'âge est corrélé négativement avec la moyenne ce qui est plutôt surprenant.
- la plupart des étudiants ont moins de 30 minutes de temps de trajet. On a une corrélation négative pour le temps de trajet, ce qui paraît logique.
- les étudiants travaillent généralement moins de 4h. Corrélation positive prévisible.
- globalement peu d'absences et d'échecs. Pas de corrélation pour les abscences et forte corrélation pour les échecs.
- quasiment la même distribution de notes aux 3 semestres et donc de même pour la moyenne. Les élèves ont des moyennes qui tournent majoritairement entre 10-12. On remarque également que toutes les notes sont très fortement corrélées entre elles.

3.2.2 ACP

On effectue une Analyse en composantes principales (ACP) sur nos données quantitatives afin d'avoir plus d'informations sur nos données.



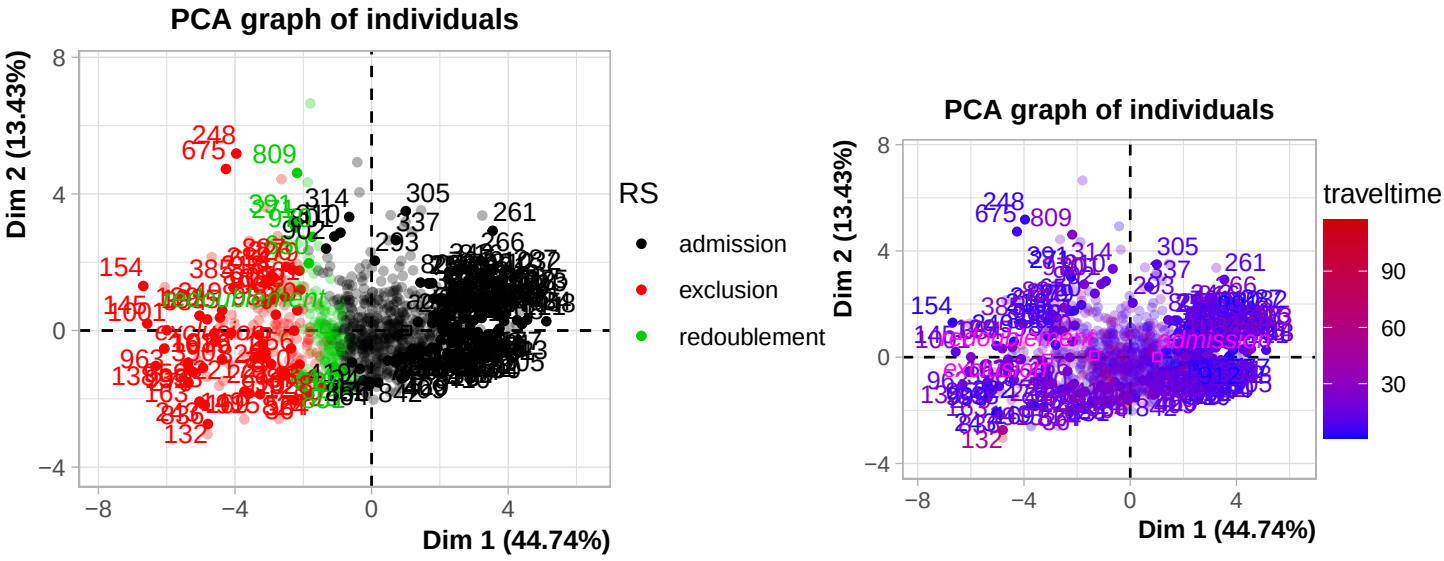
On voit que les variables studytime et traveltime sont très mal représentées, on ne peut donc pas les interpréter. Failure, age et absences ne sont pas particulièrement bien représentées non plus, mais elles sont interprétables. En interprétant le cercle de corrélation, on remarque que le premier axe sépare les bons des mauvais élèves : bons élèves à droite. Le second axe sépare les élèves plus âgés et absents (vers le haut) des élèves moins âgés et plus assidus.

On peut afficher les pourcentages d'explications d'inertie pour chacun des axes :

```
##          eigenvalue percentage of variance cumulative percentage of variance
## comp 1 4.026781e+00           4.474201e+01                          44.74201
## comp 2 1.208335e+00           1.342594e+01                          58.16796
## comp 3 1.011244e+00           1.123605e+01                          69.40401
## comp 4 9.600833e-01           1.066759e+01                          80.07160
## comp 5 8.730811e-01           9.700901e+00                          89.77250
## comp 6 6.445814e-01           7.162015e+00                          96.93451
## comp 7 1.961581e-01           2.179535e+00                          99.11405
## comp 8 7.973555e-02           8.859506e-01                         100.00000
## comp 9 1.041806e-30           1.157562e-29                         100.00000
```

On remarque que seul le premier axe explique une grande partie de l'inertie, les axes 2,3 et 4 expliquent chacun 10% d'inertie. Il faut donc 4 axes pour expliquer 80% d'inertie dans notre cas.

On affiche alors le graphe des individus pour avoir une meilleure idée.



On distingue assez clairement les 3 groupes d'individus sur le graphe. De manière logique on retrouve que les élèves ayant une bonne moyenne vers la droite. De la même manière on voit que malgré tout le temps de trajet semble quand même avoir une influence négative sur la réussite. On pourra noter que les personnes qui vont être exclues sont celles qui ont le plus grand nombre d'échecs, ce qui est cohérent. On remarque que les élèves les plus âgés se dirigent vers un redoublement ou une exclusion, ce qui est plutôt curieux mais qui pourrait s'expliquer par la présence d'élèves redoublants en difficulté.

Malheureusement, à notre niveau, nous ne disposons pas de réel moyen d'évaluer l'influence des variables quantitatives sur la réussite scolaire. Nous partirons donc du principe que les variables qui impactent le plus les notes auront un réel impact sur la réussite scolaire. On pensera notamment aux échecs.

4 Régression linéaire

Dans cette partie, nous mettons en place un modèle de régression linéaire gaussien afin de prédire la variable Moy à partir de toutes les autres. On s'intéressera ici uniquement à des modèles dont les variables auront été sélectionnées par step étant donné la grande quantité de variables. Le but sera donc de comparer les différents modèles : forward, stepwise et backward. Les modèles étant construits de manière plus ou moins similaires, la comparaison n'a pas forcément d'intérêt, mais cela permettra d'observer s'il existe de réelles différences entre chacune des méthodes.

Le modèle sera construits à partir de toutes les variables sauf G1, G2, G3 et RS, il est donc nécessaire de préparer les données avant. Pour évaluer le modèle, nous mettrons le modèle en place sur un jeu "d'entraînement" et nous l'évaluerons sur un jeu de test en utilisant le ratio $\frac{1}{5}$. La métrique que nous utiliserons sera la Mean Absolute Error (MAE), il s'agit de la moyenne des valeurs absolues des erreurs effectuées pour toutes les prédictions. Les modèles seront évalués $N = 10$ fois avec des répartitions aléatoires, la MAE obtenues à la fin sera donc une moyenne des MAE : MMAE. Cela nous permet d'avoir des résultats plus généraux sur ces modèles.

Voici les résutats obtenus:

```
##       forward stepwise backward
## MMAE 2.235727 2.241836 2.207396
```

Dans un premier temps, on se rend compte que les résultats sont tout de même assez élevés (de l'ordre de 2 ce qui est beaucoup pour des notes sur 20), indiquant donc que les modèles ne sont pas assez précis pour prédire la Moyenne. On peut imaginer que cela provient d'une insuffisance de données par exemple. On pourrait alors penser à rajouter de nouvelles variables.

Dans un second temps, on peut remarquer que les 3 modèles n'ont pas les mêmes résultats, ce qui montre qu'on n'obtient pas tout le temps le même modèle pour les différentes méthodes de sélection. De plus, on peut remarquer que les résultats changent beaucoup trop à chaque fois que l'on relance l'évaluation, il n'est donc pas possible d'élire de meilleur modèle selon cette procédure d'évaluation. Cependant, étant donné les résultats, on peut conclure que les 3 procédures de sélection se valent pour notre tâche et leurs résultats sont plutôt décevants.

5 Machine Learning : Classification de la réussite scolaire

Dans cette partie, nous nous concentrons sur la mise en place de méthodes de classification afin de prédire la variable RS (réussite scolaire).

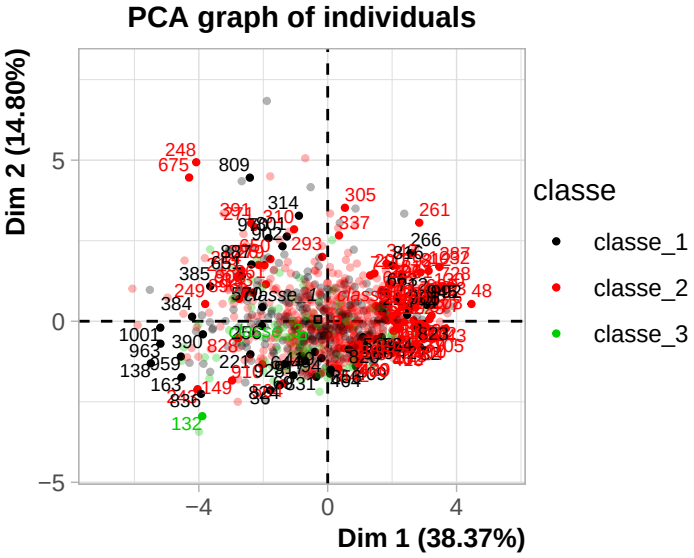
5.1 Classification non supervisée

L'intérêt d'effectuer de la classification non supervisé serait de voir comment seraient répartis les étudiants à partir des données quantitatives. On peut imaginer notamment que la classification pourrait s'effectuer sur les notes des élèves, mais selon quelles frontières ?

5.1.1 Kmeans

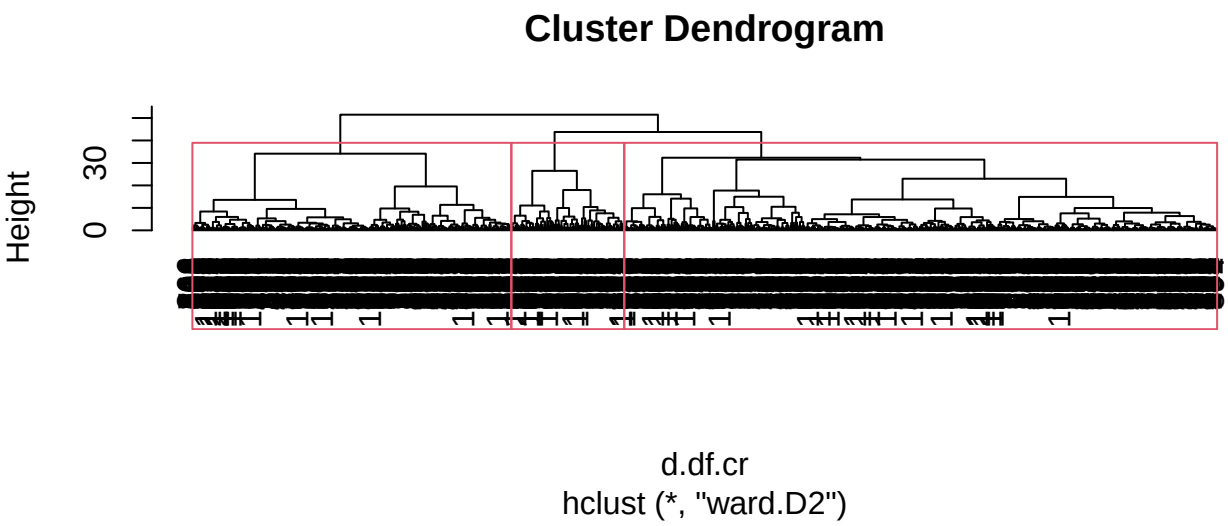
On applique l'algorithme de Kmeans avec 100 itérations et sans initialisation sur les données pour 3 groupes. Les résultats obtenus sont bien loin de la classification déjà présente . On en déduit donc que l'algorithme ne classe pas en fonction de la réussite au final. Le graphe d'ACP associé aux clusters montre qu'il n'y a pas de signification concrète à ces groupes et les 3 groupes formés ne sont pas particulièrement distinguables sur le graphe de l'ACP.

```
##
##     admission exclusion redoublement
##   1       217        66           45
##   2       448        88           79
##   3        58        19           24
```

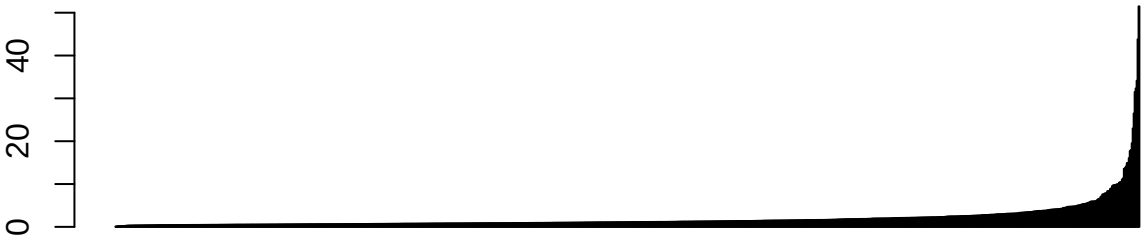


5.1.2 CAH

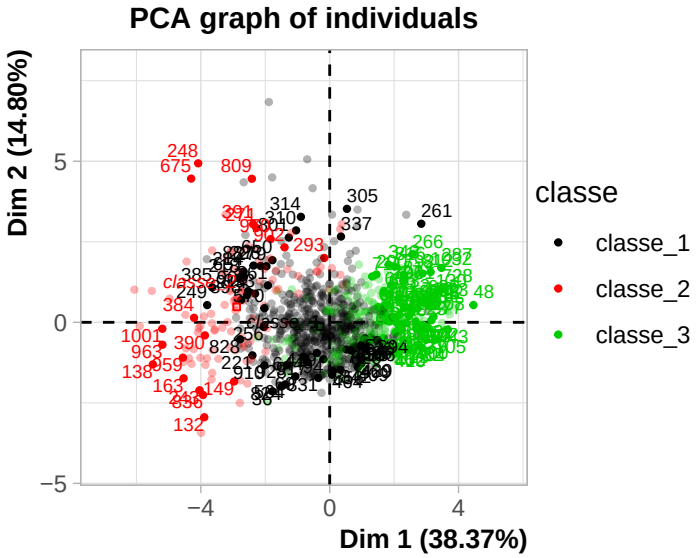
On lance un algorithme de cah sur nos données quantitatives. On obtient le dendogramme suivant duquel on extrait nos 3 classes qui nous intéressent.



On peut alors observer le graphe des hauteurs pour chaque branche. En se basant sur la perte d'inertie, il est clair que l'on aurait gardé plus de 3 classes. Cela laisse penser que l'algorithme classerait pourrait donc raffiner les classes plus que nous l'avons déja fait en distinguant 3 classes de réussite. Pour avoir une meilleure interprétation sur ces classes, on peut alors effectuer une ACP.



En observant l'ACP, on se rend compte que contrairement à la méthode de kmeans, les classes de la cah se distinguent bien sur le graphe des individus. On pourra également remarquer que la classification donnée est similaire à celle que nous avons mis en place pour RS (ACP section 3.2.2) ce qui est assez intéressant. On pourrait donc penser que l'algorithme classifie selon les notes, mais ce qui est le plus intéréssant est que les frontières pour chaque groupe sont vraiment proches de celles que nous avons mis en place.



5.2 Classification supervisée

Il s'agit ici de la partie la plus intéressante : prédire la réussite scolaire d'un élève à partir de données. Notre objectif est donc de trouver un modèle qui aurait des résultats fiables pour cette tâche. Nous nous sommes donc essentiellement intéressé à la comparaison des résultats de chacune des méthodes. Les méthodes utilisées seront évaluées avec leur accuracy et leur courbe ROC.

Pour évaluer les modèles de manière plus précise, on calcule la moyenne d'accuracy sur N configurations de jeu de données et de jeu de test. Cela nous permet d'avoir des résultats plus généraux sur les performances des modèles.

Nous avons mis en place une procédure pour évaluer nos modèles. Afin de mettre en place notre procédure d'évaluation, nous avons donc implémenté des fonctions qui prennent en entrée le jeu d'entraînement et le jeu de test. Ces fonctions entraînent les modèles correspondants, effectuent les prédictions sur le jeu de test et renvoient une liste contenant l'accuraccy, la table de confusion et la courbe ROC. Les modèles ont ensuite été évalués N=10 fois avec à chaque fois une séparation différente dont le ratio est $\frac{1}{5}$. Cela permet d'avoir des résultats plus généraux étant donné que l'on teste les modèles dans plusieurs conditions différentes.

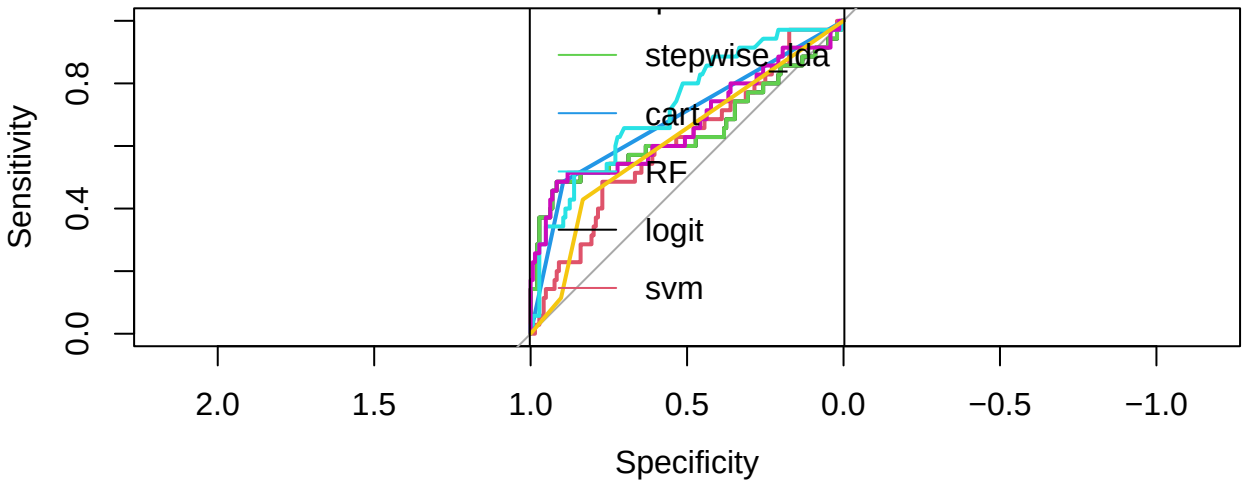
Voici les modèles que nous avons testés :

1. LDA
2. QDA
3. Stepwise lda
4. Random forest
5. Cart avec l'arbre optimal
6. Regression logistique
7. Support vector machine (bonus) avec un noyau radial avec c=10.

Notons également que nous avons retiré les variables de note du jeu de données puisque celles-ci réduiraient l'intérêt de la classification (dans notre cas la réussite scolaire sur l'année est calculée à partir des notes).

5.3 Comparaison

```
##                lda       qda stepwise_lda      cart        RF     logit
## accuracy 0.7200957 0.6626794    0.7177033 0.7153110 0.6985646 0.7224880
## AUC      0.6466270 0.6079365    0.6466270 0.6907738 0.7263889 0.6694444
##                svm
## accuracy 0.6339713
## AUC      0.6196429
```



Après calculs, on obtient des résultats plutôt décevants pour cette classification. Pour l'accuracy, les résultats tournent autour des 70% ce qui est assez faible, de même pour la courbe ROC. Le meilleur résultat en terme d'accuracy s'obtiennent avec une régression logistique et le pire avec une SVM. Pour l'aire sous la courbe ROC, les meilleurs résultats sont obtenus avec le modèle cart. Globalement, on en déduit que ces données ne se prêtent pas bien à la classification de la réussite scolaire dans leur état actuel.

6 Conclusion

Au final, ce jeu de données est vraiment intéressant étant donné qu'il renferme une grande quantité d'informations sur les étudiants. Il nous a permis d'appliquer une grande partie des méthodes statistiques apprises cette année de manière plus ou moins pertinente. On pourra noter également que la présence d'une grande quantité de variables qualitatives et de moins de variables quantitatives a rendu l'application de certaines méthodes plus compliquées. Cependant, globalement il s'agit d'un dataset intéressant pour apprendre à appliquer ces méthodes dans un cas moins trivial.

De plus, il nous a permis d'identifier les facteurs qui ont un impact sur les notes et la réussite scolaire des étudiants. On a ainsi pu observer des relations plutôt inattendues notamment par rapport à nos propres idées et expériences.

Cependant, les données du jeu ne sont pas forcément adaptées à la classification ni à la régression linéaire. On obtient des résultats assez bas pour toutes les méthodes utilisées dans les 2 tâches. On peut donc imaginer qu'une manière d'améliorer les résultats serait alors de rajouter de nouvelles variables qui pourrait avoir un impact sur la moyenne et la réussite car d'après notre étude, toutes les variables n'ont pas toutes un impact sur elles.

Pour finir, on pourra dire que ce jeu de données montre un cas pratique non trivial pour l'application des méthodes d'analyse de données et met en valeurs l'apport d'informations qu'elle donnent.

A Suite de l'étude des variables et sorties manquantes

A.1 Les sorties

```
# Distribution
ggplot(data = df, aes(x =goout, fill = goout)) +
  geom_bar() +
  labs(title = "Distribution des sorties",
       x = "Sorties", y = "Nombre d'étudiants")
```

```
# Lien avec la moyenne
summary(lm(Moy ~ goout,data=df))
```

```
##
## Call:
## lm(formula = Moy ~ goout, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5887  -1.8876  -0.0015   2.1124   7.6652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.5493     0.3770  27.980  < 2e-16 ***
## goout2        1.3727     0.4276   3.210  0.00137 **
## goout3        1.0049     0.4151   2.421  0.01564 *
## goout4        0.4522     0.4320   1.047  0.29548
## goout5       -0.1853     0.4517  -0.410  0.68178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.177 on 1039 degrees of freedom
## Multiple R-squared:  0.02957,    Adjusted R-squared:  0.02583
## F-statistic: 7.915 on 4 and 1039 DF,  p-value: 2.766e-06
```

```
# Lien avec la réussite
chisq.test(df$goout,df$RS)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$goout and df$RS
## X-squared = 20.537, df = 8, p-value = 0.008485
```

A.1.1 Le sexe des étudiants

```
# Distribution
ggplot(df, aes(x = sex)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Répartition des sexes")
```

```
# Lien avec la moyenne
summary(lm(Moy ~ sex,data=df))
```

```
##
## Call:
```

```
## lm(formula = Moy ~ sex, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0152  -2.0152  -0.0152   2.1722   8.1722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.3486     0.1324  85.706   <2e-16 ***
## sexM         -0.1874     0.2010  -0.932    0.351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.219 on 1042 degrees of freedom
## Multiple R-squared:  0.0008335,  Adjusted R-squared:  -0.0001254
## F-statistic: 0.8693 on 1 and 1042 DF,  p-value: 0.3514
```

```
# Lien avec la réussite
chisq.test(df$sex,df$RS)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$sex and df$RS
## X-squared = 1.1035, df = 2, p-value = 0.5759
```

A.2 Travail des parents

```
#Distributions
g2=ggplot(data = df, aes(x = Mjob)) +
  geom_bar() +
  labs(title = "Distribution du travail de la mère") +
  scale_fill_manual(values = c("#7570b3","#0072B2", "#E69F00", "#009E73", "#F0E442"))

g1=ggplot(data = df, aes(x = Fjob)) +
  geom_bar() +
  labs(title="Distribution du travail du père") +
  scale_fill_manual(values = c("#7570b3","#0072B2", "#E69F00", "#009E73", "#F0E442"))

grid.arrange(g1, g2, ncol = 2)
```

```
# Lien avec les notes
summary(lm(Moy ~ Medu+Fedu,data=df))
```

```
##
## Call:
## lm(formula = Moy ~ Medu + Fedu, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.265  -1.732   0.068   2.126   7.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.4233     0.2595  36.316  < 2e-16 ***
```

```
## Medu          0.5214     0.1125   4.635 4.02e-06 ***
## Fedu          0.2037     0.1150   1.771   0.0769 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.133 on 1041 degrees of freedom
## Multiple R-squared:  0.05434,    Adjusted R-squared:  0.05252
## F-statistic: 29.91 on 2 and 1041 DF,  p-value: 2.344e-13
```

```
# Lien avec la réussite
chisq.test(df$Mjob,df$RS)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$Mjob and df$RS
## X-squared = 21.736, df = 8, p-value = 0.005429
```

```
chisq.test(df$Fjob,df$RS)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$Fjob and df$RS
## X-squared = 7.4964, df = 8, p-value = 0.4841
```

A.2.1 Les relations

```
# Distribution
gr1=ggplot(df, aes(x = romantic)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Distribution des personnes en couple",
       x = "Couple", y = "Nombre d'étudiants")

# Lien avec les notes
summary(lm(Moy~ romantic,data=df))
```

```
##
## Call:
## lm(formula = Moy ~ romantic, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1486  -1.9455   0.1222   2.1847   7.8514
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.4819     0.1236  92.871  < 2e-16 ***
## romanticyes  -0.6041     0.2074  -2.913  0.00366 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.207 on 1042 degrees of freedom
## Multiple R-squared:  0.008077,   Adjusted R-squared:  0.007125
## F-statistic: 8.485 on 1 and 1042 DF,  p-value: 0.003658
```

```r
yes = df$Moy[df$romantic=='yes']
no = df$Moy[df$romantic=='no']

# Boxplot des notes
gr2=ggplot(data = df, aes(x = romantic, y = Moy, fill = romantic)) +
  geom_boxplot() +
  scale_fill_manual(values = c("#0072B2", "#F0E442")) +
  labs(title = "Relation entre être en couple et les notes",
       x = "Couple", y = "Notes")

grid.arrange(gr1, gr2, ncol = 2)
```

```r
# Lien avec la réussite
chisq.test(df$romantic,df$RS)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$romantic and df$RS
## X-squared = 5.5477, df = 2, p-value = 0.06242
```

A.3 La consommation d'alcool

```r
# Distribution
ggplot(data = df, aes(x =Dalc, fill = Dalc)) +
  geom_bar() +
  labs(title = "Distribution de la consommation d'alcool en semaine",
       x = "Consommation", y = "Nombre d'étudiants")
```

```r
# Lien avec la moyenne
summary(lm(Moy ~ Dalc,data=df))
```

```
##
## Call:
## lm(formula = Moy ~ Dalc, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.244  -1.911   0.089   2.089   7.756
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.5777     0.1181  98.001  < 2e-16 ***
## Dalc2        -0.8889     0.2564  -3.467 0.000547 ***
## Dalc3        -0.8917     0.4013  -2.222 0.026475 *
## Dalc4        -2.0008     0.6358  -3.147 0.001696 **
## Dalc5        -1.3982     0.6358  -2.199 0.028079 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.185 on 1039 degrees of freedom
## Multiple R-squared:  0.02443,    Adjusted R-squared:  0.02068
## F-statistic: 6.505 on 4 and 1039 DF,  p-value: 3.594e-05
```

```r
# Lien avec la réussite
chisq.test(df$Dalc,df$RS)
```

```
## Warning in chisq.test(df$Dalc, df$RS): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$Dalc and df$RS
## X-squared = 28.342, df = 8, p-value = 0.0004134
```

```r
# Distribution
ggplot(data = df, aes(x =Walc, fill = Walc)) +
  geom_bar() +
  labs(title = "Distribution de la consommation d'alcool le week-end",
       x = "Consommation", y = "Nombre d'étudiants")
```

```r
# Lien avec la moyenne
summary(lm(Moy ~ Walc,data=df))
```

```
##
## Call:
## lm(formula = Moy ~ Walc, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2831  -1.9050   0.0503   2.0614   7.7169
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.6164     0.1600  72.593  < 2e-16 ***
## Walc2        -0.1398     0.2626  -0.532 0.594577
## Walc3        -0.3781     0.2767  -1.366 0.172115
## Walc4        -1.1768     0.3154  -3.731 0.000201 ***
## Walc5        -1.2831     0.4065  -3.157 0.001642 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.192 on 1039 degrees of freedom
## Multiple R-squared:  0.0201, Adjusted R-squared:  0.01633
## F-statistic: 5.328 on 4 and 1039 DF,  p-value: 0.0003004
```

```r
# Lien avec la réussite
chisq.test(df$Walc,df$RS)
```

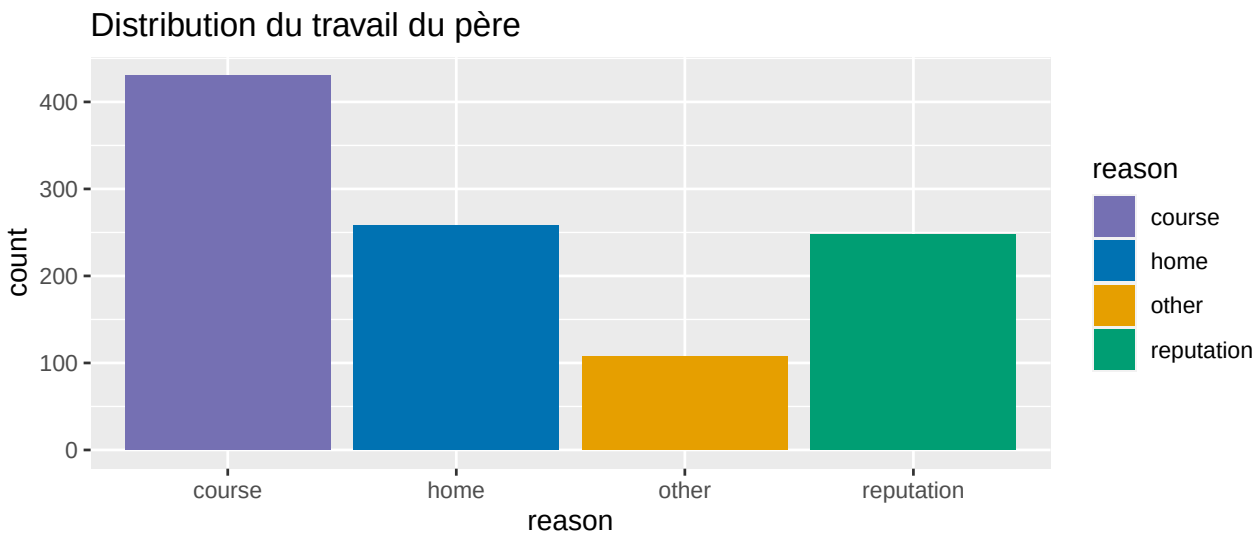
```
##
##  Pearson's Chi-squared test
##
## data:  df$Walc and df$RS
## X-squared = 16.5, df = 8, p-value = 0.03576
```

A.4 Les raisons du choix d'école

D'après le digramme circulaire, seule "other" possède un petit effectif alors que "course" domine. Ainsi, les élèves vont majoritairement en cours car ils les apprécient. D'après l'ANOVA1, il est clair que la raison

d'aller en cours impacte les notes des étudiants (p-value $< 5\%$). Cela paraît cohérent étant donné que cela détermine leur motivation à avoir de bonnes notes. De la même manière, la raison est bien corrélé avec la réussite scolaire, ce qui parâit bien cohérent.

```
# Distribution
ggplot(data = df, aes(x = reason, fill = reason)) +
  geom_bar() +
  labs(title="Distribution du travail du père") +
  scale_fill_manual(values = c("#7570b3","#0072B2", "#E69F00", "#009E73", "#F0E442"))
```




```
# Lien avec les notes
summary(lm(Moy~ reason,data=df))
```

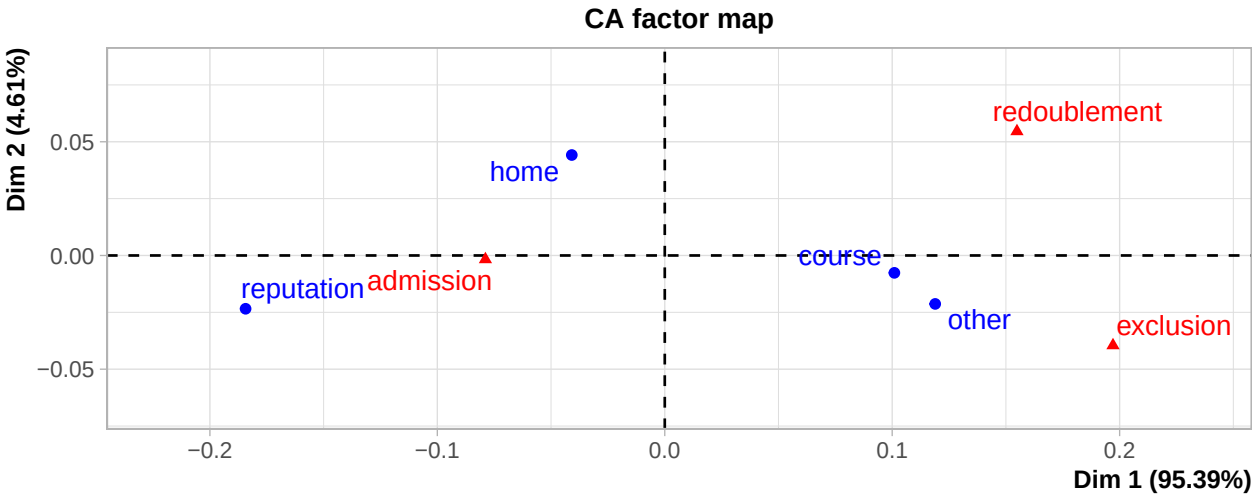
```
##
## Call:
## lm(formula = Moy ~ reason, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3858  -1.8791  -0.0052   2.1209   7.7876
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.87907    0.15372  70.771  < 2e-16 ***
## reasonhome        0.45943    0.25103   1.830   0.0675 .
## reasonother      -0.03956    0.34309  -0.115   0.9082
## reasonreputation  1.17335    0.25417   4.616 4.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.188 on 1040 degrees of freedom
## Multiple R-squared:  0.02209,    Adjusted R-squared:  0.01927
## F-statistic: 7.832 on 3 and 1040 DF,  p-value: 3.587e-05
```

```
# Lien avec la réussite
chisq.test(df$reason,df$RS)
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  df$reason and df$RS
## X-squared = 15.479, df = 6, p-value = 0.01684
```

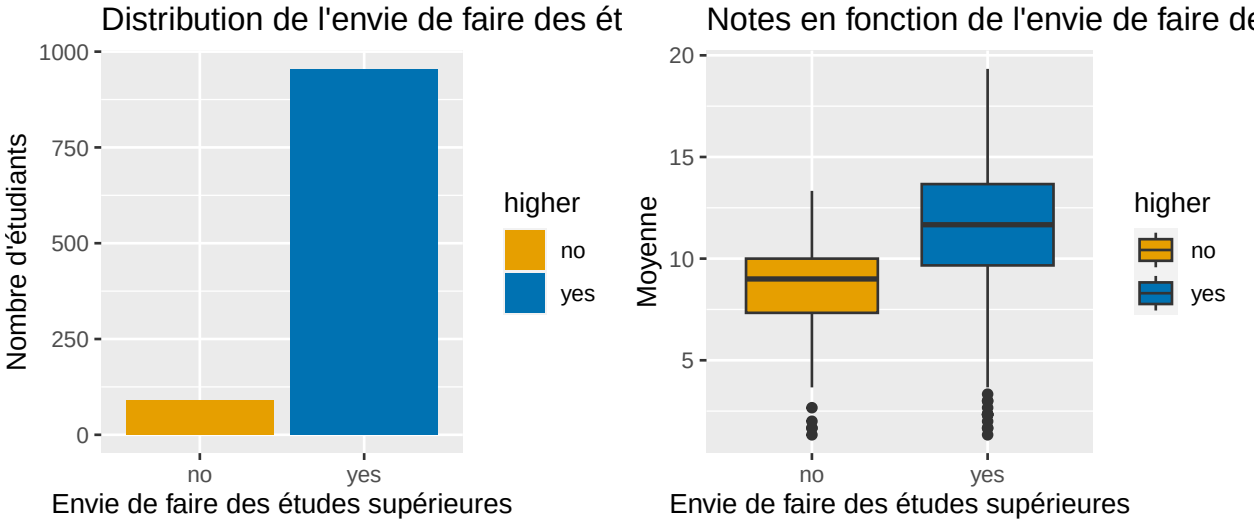
```
# AFC sur le travail de la mère
df.reason = data.frame(df$reason,df$RS)
table.reason = table(df.reason)
res = CA(table.reason)
```



On voit bien avec l'AFC que les personnes étant admises sont celles qui choisissent l'école pour sa réputation et sa proximité par rapport à leur domicile. A l'inverse on voit que les étudiants qui ont échoués sont ceux qui ont choisis l'école pour les cours ou d'autres raisons. On voit ici une des limite de cette méthode, en effet, on peut penser que les élèves qui réussissent le mieux sont ceux qui sont le plus motivés et donc qui ont chosies l'école pour les cours plus que pour sa réputation.

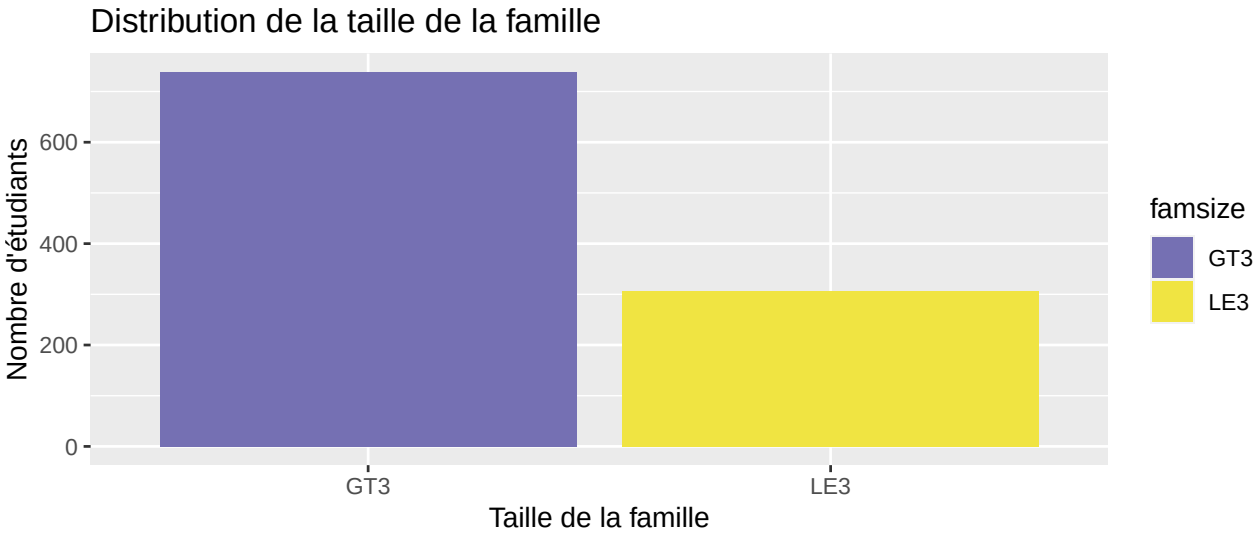
A.5 Volonté de faire des études supérieures

On observe qu'au moins 80% des élèves veulent continuer leur études après le lycée, ce qui est plutôt rassurant. De plus, d'après le test de Fisher, les deux variables sont corrélées. On peut également annoncer que ceux qui veulent faire des études supérieures tendent à avoir de meilleures notes grâce au test unilatéral. A priori, la volonté de faire des études supérieures est corrélée à la réussite scolaire. Donc, ceux qui veulent poursuivre leurs études auront de meileures notes et tendance à ne pas être en échec.



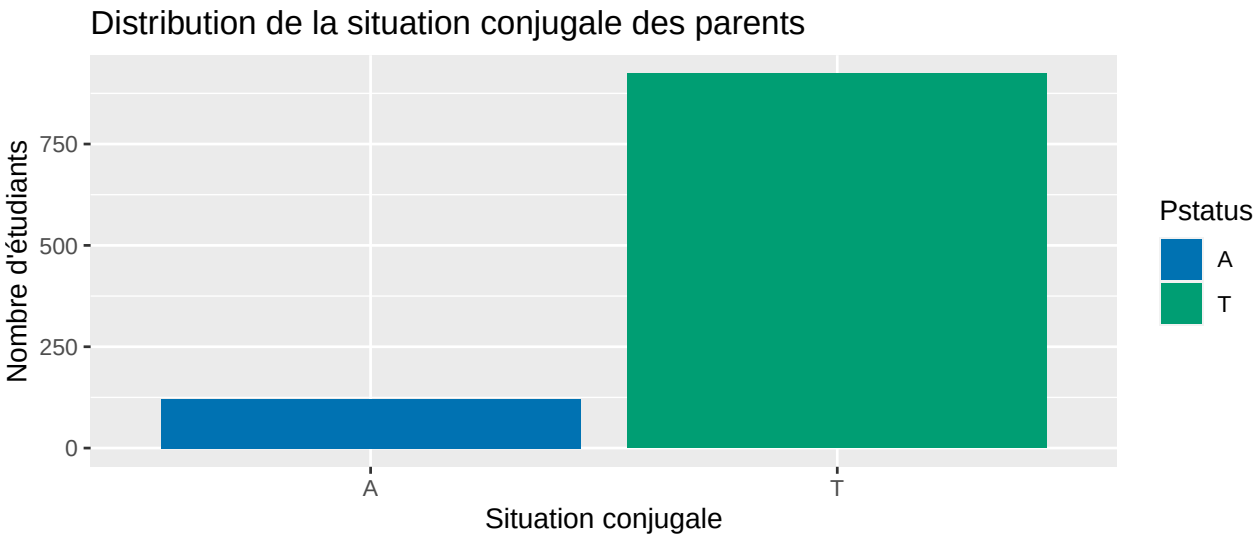
A.6 La taille de la famille

On a deux fois plus de grandes familles que de petites familles. D'après le test de Fisher, il y a bien un impactde taille de la famille sur les notes. Le test d'indépendance avec la réussite indique cependant que la taille de la famille n'est pas liée à la réussite scolaire.



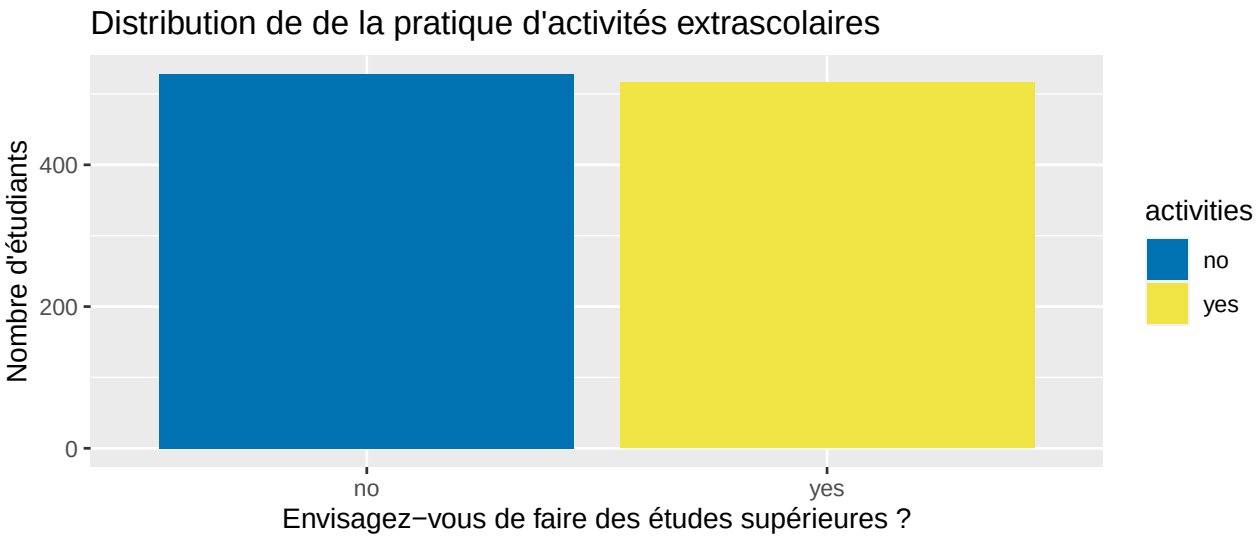
A.7 Situation familliale : séparation des parents

Le jeu est très déséquilibré au sujet de la situation famille : il y a 4 fois plus d'étudiants qui ont leurs parents qui vivent ensemble. De plus, le test de Fisher indique que la situation familliale n'a pas d'impact sur les notes. Le test de Chi2 soutient que le status des parents et la réussite scolaire sont indépendants.



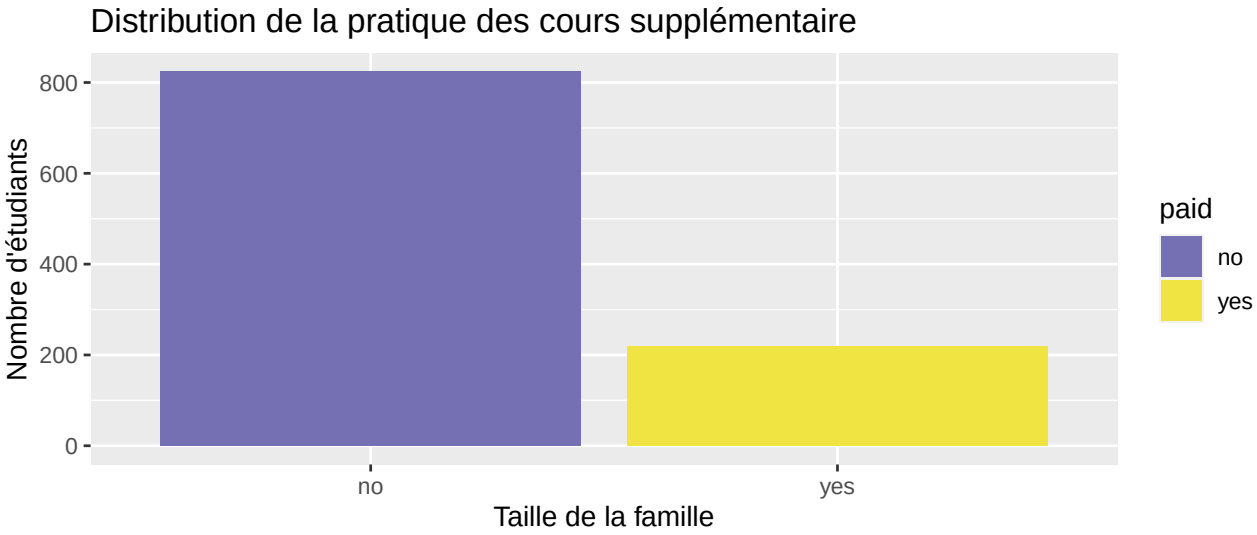
A.8 Activités extrascolaires

On a autant d'élèves qui pratiquent des activités extrascolaires que d'élèves qui n'en pratiquent pas, ce qui est plutôt intéréssant. De plus, le test de Fisher indique plutôt qu'il n'y a pas de liens entre les activités extrascolaires et les notes, ce qui est plutôt surprenant étant donné que l'on aurait tendance à penser que les étudiants ayant des activités, ont moins de temps pour étudier. Dans la même lignée, les activités sont plutôt indépendates de la réussite d'après le test de Chi2.



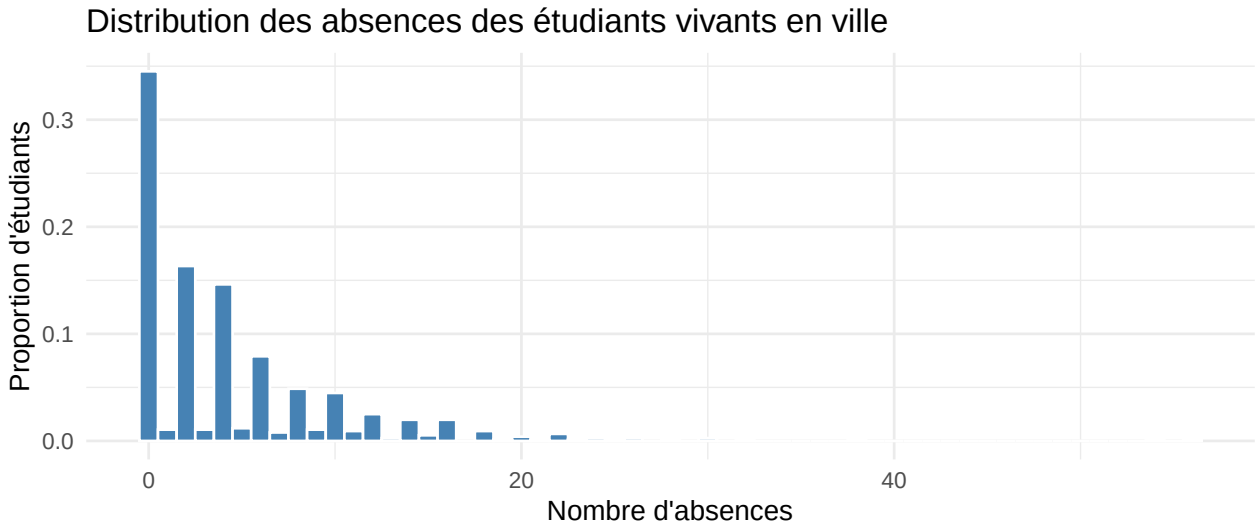
A.9 Cours supplémentaires

Il y a bien plus d'élèves qui ne suivent pas de cours supplémentaires que d'élèves qui en suivent. Cett distributution est cohérente avec l'idée qu'on oeut se faire. Le test de Fisher indique plutôt que les suivis de cours supplémentaires n'a pas d'impact sur la moyenne. De même, le suivi de cours supplémentaire n'est pas lié à la réussite.

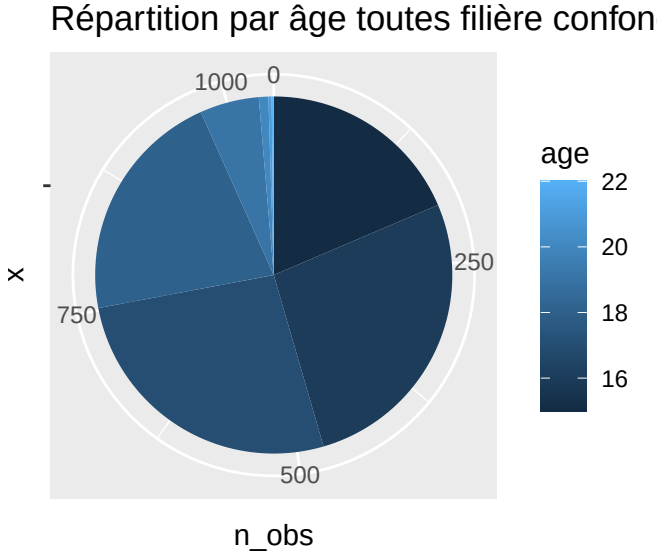


A.9.1 Absences des étudiants

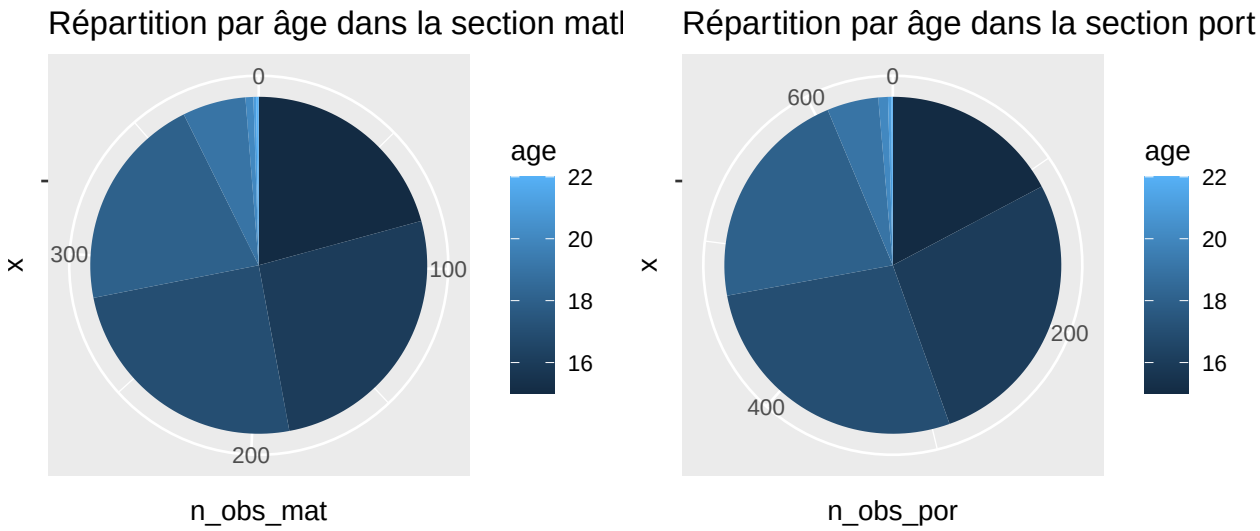
```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



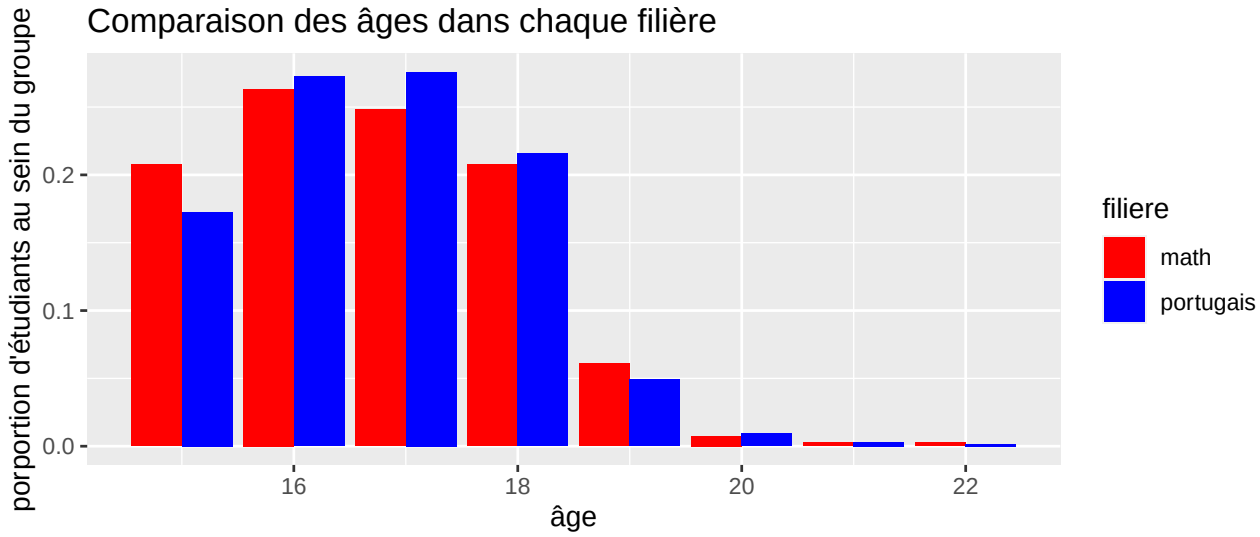
A.9.2 L'âge des élèves



La couleur la plus claire correspond à l'âge le plus grand (22 ans), dès que l'on passe à une couleur plus foncée, on diminue l'âge de 1. On voit clairement ici que la majorité des étudiants ont entre 15 et 19 ans.

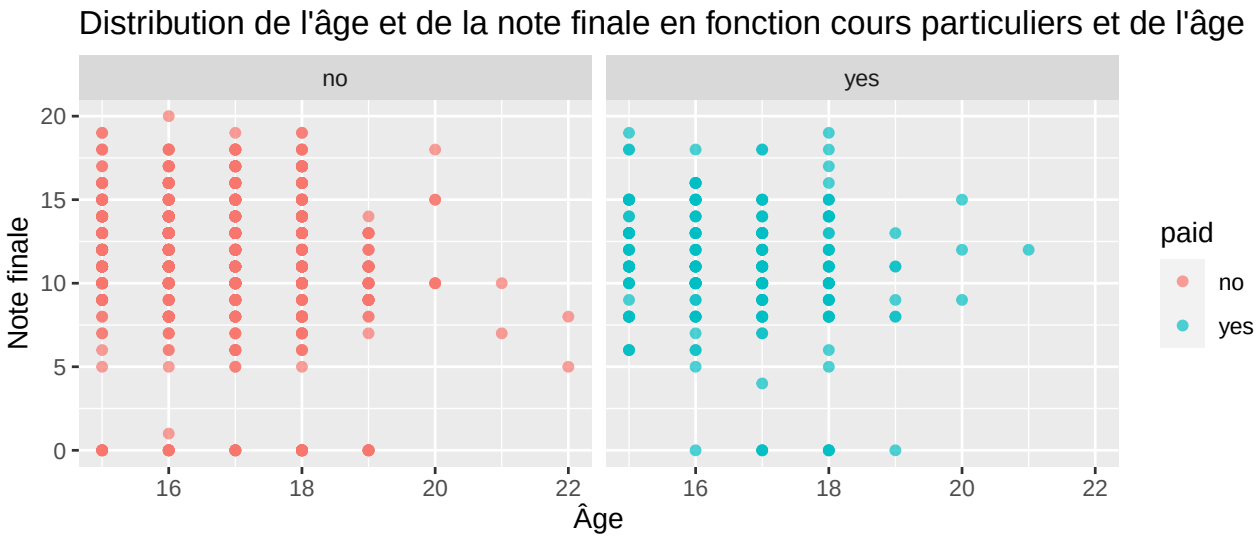


On voit que la répartiion semble être la grossièrement la même, en effet:



A.10 Cours particuliers

```
ggplot(df, aes(x = age, y = G3, color = paid)) +
  geom_point(alpha = 0.7) +
  facet_wrap(~paid) +
  labs(title = "Distribution de l'âge et de la note finale en fonction cours particuliers et de l'âge",
       x = "Âge", y = "Note finale")
```



Curieusement, les résultats semblent meilleur pour ceux qui n'ont pas pris de cours

B Codes pour la partie Régression

Code pour l'évaluation :

```
X = subset(df, select = -c(G1,G2,G3,RS) )

mae_forward = 0
mae_step = 0
mae_back = 0
N=10

for(i in 1:N)
{
  # Génération des jeux d'entraînement et de test
  set.seed(i)
  n <- nrow(X)
  p <- ncol(X)-1
  test.ratio <- .2 # ratio of test/train samples
  n.test <- round(n*test.ratio)
  tr <- sample(1:n,n.test)
  df.reg.test <- X[tr,]
  df.reg.train <- X[-tr,]

  # Modèles de bases
  res0=lm(df.reg.train$Moy ~ 1,df.reg.train)
  resT = lm(df.reg.train$Moy~.,df.reg.train)

  # Modèles par sélection
  linforward = step(res0,scope=formula(resT),direction="forward")
  linstep = step(res0,scope=formula(resT),direction="both")
  linback = step(resT,direction="backward")

  mae_forward = mae_forward + mean(abs(df.reg.test$Moy - predict(linforward,newdata = df.reg.test)))
  mae_step = mae_step + mean(abs(df.reg.test$Moy - predict(linstep,newdata = df.reg.test)))
  mae_back = mae_back + mean(abs(df.reg.test$Moy - predict(linback,newdata = df.reg.test)))
```

```
}
```

Code pour la comparaison :

```
mae_forward = mae_forward/N
mae_step = mae_step/N
mae_back = mae_back/N

result=matrix(NA, ncol=3, nrow=1)
rownames(result)=c('MMAE')
colnames(result)=c('forward','stepwise','backward')
result[1,]= c(mae_forward,mae_step,mae_back)
result
```

C Codes pour la partie Machine Learning

Voici un exemple de code pour l'architecture des fonctions d'évaluation:

```
LDA = function(data.train,data.test)
{
  res_lda=lda(data.train$RS ~., data=data.train)
  pred_lda <- predict(res_lda,newdata=data.test)$posterior[,2]

  # Table de confusion
  tab = table(data.test$RS,predict(res_lda,newdata=data.test)$class)

  # Courbe ROC
  ROC_lda <- roc(data.test$RS, pred_lda)

  # Accuracy
  accuracy_lda = mean(data.test$RS==predict(res_lda,newdata=data.test)$class)

  res = list(accuracy_lda,tab,ROC_lda)
  return(res)
}
```

Voici le code pour l'évaluation des modèles :

```
# Suppresion des colonnes
X = subset(df, select = -c(G1,G2,G3,Moy) )

for(i in 1:N)
{
  # Génération des jeux d'entraînement et de test
  set.seed(i)
  n <- nrow(X)
  p <- ncol(X)-1
  test.ratio <- .2 # ratio of test/train samples
  n.test <- round(n*test.ratio)
  tr <- sample(1:n,n.test)
  df.test <- X[tr,]
  df.train <- X[-tr,]

  # LDA
  res = LDA(df.train,df.test)
```

```
  a_lda[i] = res[[1]]
  ROC_lda = res[[3]]

  # QDA
  res = QDA(df.train,df.test)
  a_qda[i] = res[[1]]
  ROC_qda = res[[3]]

  # Stepwise
  res = stepwise(df.train,df.test)
  a_step = res[[1]]
  ROC_step = res[[3]]

  # cart
  res = Cart(df.train,df.test)
  a_cart[i] = res[[1]]
  ROC_cart = res[[3]]

  # Random forest
  res = RF(df.train,df.test)
  a_RF[i] = res[[1]]
  ROC_RF = res[[3]]

  # Regression logistique
  res = logit(df.train,df.test)
  a_logit[i] = res[[1]]
  ROC_logit = res[[3]]

  # Regression logistique
  res = SVM(df.train,df.test)
  a_svm[i] = res[[1]]
  ROC_svm = res[[3]]

}
```