Dossier Clustering

DORONZO Franck,
DUDOIT Romain,
DARANKOUM Davy,
COSTE Louis

M1 Informatique - Université Lumière Lyon 2

18 avril 2021

Table des matières

1 Introduction

Nous avons choisi de travailler sur un jeu de données disponible sur le site UCI Machine Learning Repository à l'adresse archive.ics.uci.edu/ml/datasets/Student+Performance. Il s'agit d'un jeu de données qui porte sur les caractéristiques (essentiellement sociales) et les notes obtenues sur trois périodes d'élèves de l'enseignement supérieur de deux écoles portugaises. Ces données ont été collectées à l'aide de rapports scolaires et de questionnaires. Deux ensembles de données sont fournis concernant les performances dans deux matières distinctes : les mathématiques et la langue portugaise. Notre étude s'est portée sur celui qui contient les notes en langue portugaise pour lequel il y a plus d'observations (649 contre 395).

Les variables du jeu de données sont les suivantes :

1 school - école de l'élève (binaire : 'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira)
2 sex - sexe de l'élève (binaire : 'F' ou 'M')
3 age - âge de l'élève (de 15 à 22)
4 address (binaire : 'U' - urbain ou 'R' - rural)
5 famsize - taille de la famille (binaire : 'LE3' - plus ou moins de 3 ou 'GT3' - plus que 3)
6 Pstatus - cohabitation des parents (binaire : 'T' - vivent ensemble ou 'A' - séparés)
7 Medu - niveau de scolarité de la mère (0 - aucun, 1 - enseignement primaire (4th grade), 2 - 5th to 9th grade, 3 - enseignement secondaire ou - niveau plus élevé)
8 Fedu - niveau de scolarité du père (0 - aucun, 1 - enseignement primaire (4th grade), 2 - 5th to 9th grade, 3 - enseignement secondaire ou -4 niveau plus élevé)
9 Mjob - emploi de la mère ('teacher', 'health', 'services' (administratif ou police), 'at_home' ou 'other')
10 Fjob - emploi du père ('teacher', 'health', 'services' (service administratif ou police), 'at_home' ou 'other')
11 reason - raison de choisir l'école (close to home, 'reputation', 'course' ou 'other')
12 guardian - tuteur de l'élève ('mother', 'father' ou 'other')
13 traveltime - temps de trajet entre la maison et l'école (1 - <15 min., 2 - 15 à 30 min., 3 - 30 min. à 1 heure, ou 4 - >1 heure)
14 studytime - temps d'étude dans la semaine (1 - <2 heures, 2 - 2 à 5 heures, 3 - 5 à 10 heures, or 4 - >10 heures)
15 failures - nombre d'échec passés (n si 1<=n<3, sinon 4)
16 schoolsup - soutien scolaire (binaire : yes or no)
17 famsup - soutien familial (binaire : yes or no)
18 paid - cours payants supplémentaires (binaire : yes or no)
19 activities - activités parascolaires (binaire : yes or no)
20 nursery - est allé à l'école maternelle (binaire : yes or no)
21 higher - veut aller dans l'enseignement supérieur (binaire : yes or no)
22 internet - accès à internet à la maison (binaire : yes or no)
23 romantic - relation amoureuse (binaire : yes or no)
24 famrel - qualité des relations familiales (de 1 - très mauvais à 5 - excellent)
25 freetime - temps libre après les cours (de 1 - très peu à 5 - très élevé)
26 goout - sort avec des amis (de 1 - très peu souvent à 5 - très souvent)

27 Dalc - consommation d'alcool en semaine (de 1 - très peu à 5 - très élevée)
28 Walc - consommation d'alcool le weekend (de 1 - très peu à 5 - très élevée)
29 health - état de santé actuel (de 1 - très mauvais à 5 - très bon)
30 absences - nombre total d'absence (de 0 à 93)

Notes en langue portugaise obtenues sur 3 périodes :
31 G1 (de 0 à 20)
32 G2 (de 0 à 20)
33 G3 (de 0 à 20, cible finale)

2 Formulation des problématiques d'étude

Étant en face d'un jeu de données qui a initialement été constitué pour pouvoir étudier la performance des élèves sur les cours Portugais, notre première problématique s'est articulée autour de cette dernière. Étudier la performance des élèves dans le cours de Portugais. Dans un second temps, nous nous sommes intéressé à la qualité de vie des élèves. Afin de mieux organiser notre analyse, nous avons exploré cette problématique sous 2 angles. Le premier a concerné l'étude de la qualité des relations familiales des élèves et le deuxième, l'étude des conditions de santé de ces mêmes élèves. Ainsi avec ces deux dernières études, le but sera de voir le rôle que joue la qualité de santé ou la qualité de relation familiale dans la performance des élèves. Enfin nous nous sommes intéressé à la constitution de classes d'élèves dans le but de mettre en évidence les variables qui caractérisent la plupart des groupes et d'établir par la même occasion un éventuel lien entre ces classes et les notes obtenues.

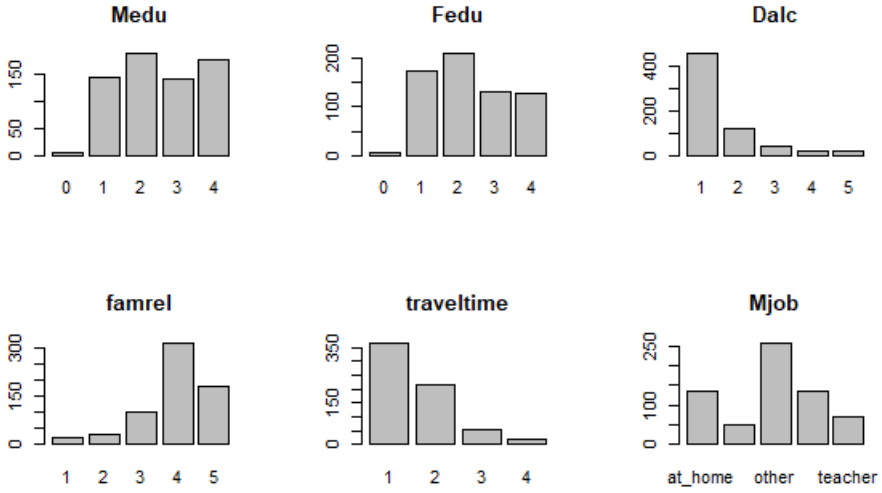
3 Étude de la performance des élèves

3.1 Choix de la méthode utilisée et pré-traitements

Notre première problématique est de déterminer si il y a un lien entre les caractéristiques d'un élève et le fait qu'il soit plus ou moins bon à l'école. Les variables du jeu de données étant majoritairement qualitatives. Nous avons alors envisagé de réaliser une analyse des correspondances multiples (ACM). Rappelons que dans une ACM, deux individus sont proches si ils ont choisi les mêmes modalités. Rappelons aussi que la proximité entre deux modalités de variables différentes sous entend quelles sont communes aux mêmes individus et que la proximité entre deux modalités de la même variable sous entend qu'elle sont communes à des individus qui se ressemblent.

Avant toute ACM, il est indispensable de réaliser une analyse préliminaire de chaque variable, afin de voir si toutes les classes sont aussi bien représentées ou s'il existe un déséquilibre. En effet, de part une distance élevée au centre de gravité du plan factoriel, une modalité rare risque de contribuer trop fortement à certains axes. Il convient ainsi d'éviter de travailler avec des modalités d'effectifs trop faibles qui risquent de perturber les résultats de l'analyse

en regroupant par exemple une modalité dont l'effectif est faible avec une autre modalité proche. Nous avons ainsi converti en facteur les variables du jeu de données qui devaient être considérées comme qualitatives et avons donc analysé les effectifs des modalités de chacune de ces variables. Nous avons trouvé 6 variables pour lesquelles les effectifs de certaines modalités sont inférieurs à 5% de la population d'élèves :



Nous avons ainsi regroupé les modalités 0 et 1 des variables 'Medu' et 'Fedu', les modalités 4 et 5 de la variable 'Dalc', les modalités 3 et 4 de la variable 'famrel', 3 et 4 de la variable 'traveltime'. Dans les graphiques qui suivront, le nom des variables dont des modalités ont été regroupées à été concaténé à la chaîne de caractère '.reg'. Concernant la variable 'Mjob', il y a peu (23) d'élèves dont la mère est professeur mais nous avons gardé les modalités tels quel de manière a rester cohérent par rapport à la variable 'Fjob'.

Afin de garder des variables actives homogènes, nous avons décider de décomposer le problème en réalisant 3 ACM liées à 3 thèmes dont les variables actives sont les suivantes :

- thème 1 (famille) : 'famsize', 'famrel', 'guardian', 'Pstatus', 'Fedu', 'Medu', 'Fjob', 'Mjob'
- thème 2 (travail, activités, soutien scolaire ...) : 'activities', 'paid', 'studytime', 'freetime', 'schoolsup', 'higher', 'internet', 'nursery', 'address'
- thème 3 (consommation d'alcool, sorties, santé) : 'Dalc','Walc','goout','romantic','health'

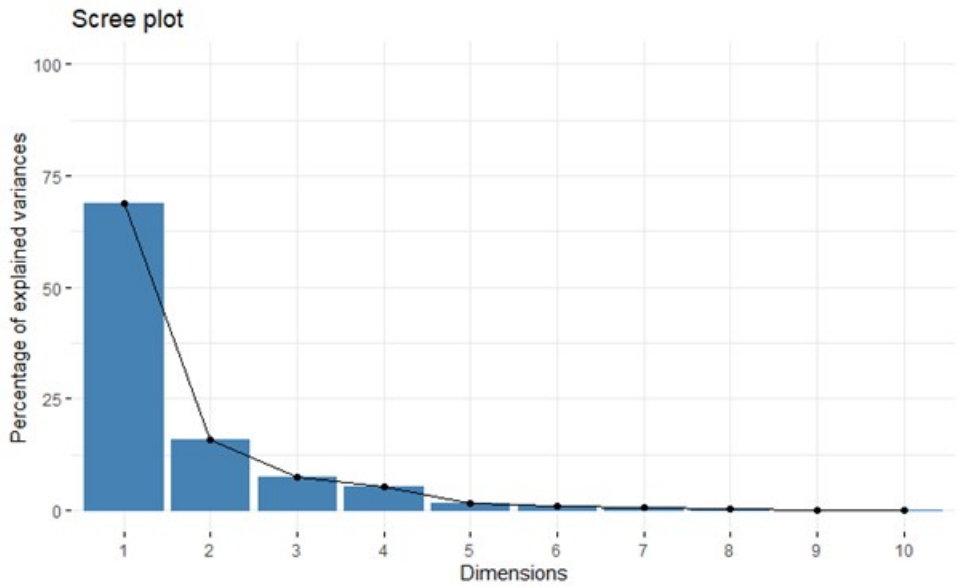
En variables supplémentaires qualitatives nous avons considéré les variables 'sex' et 'school' et en supplémentaires quantitatives les variables 'G1', 'G2', 'G3', 'failures' et 'absences'.

3.1.1 1ère ACM

Nous considérons les variables 'famsize', 'famrel', 'guardian', 'Pstatus', 'Fedu', 'Medu', 'Fjob','Mjob'.

Choix du nombre d'axe

De par un codage disjonctif sur les variables qualitatives, il est difficile de concentrer l'inertie sur les premiers facteurs. Afin de mieux rendre compte de l'intérêt de ces derniers, nous choisirons le nombre d'axes après avoir appliqué la correction de Benzécri.

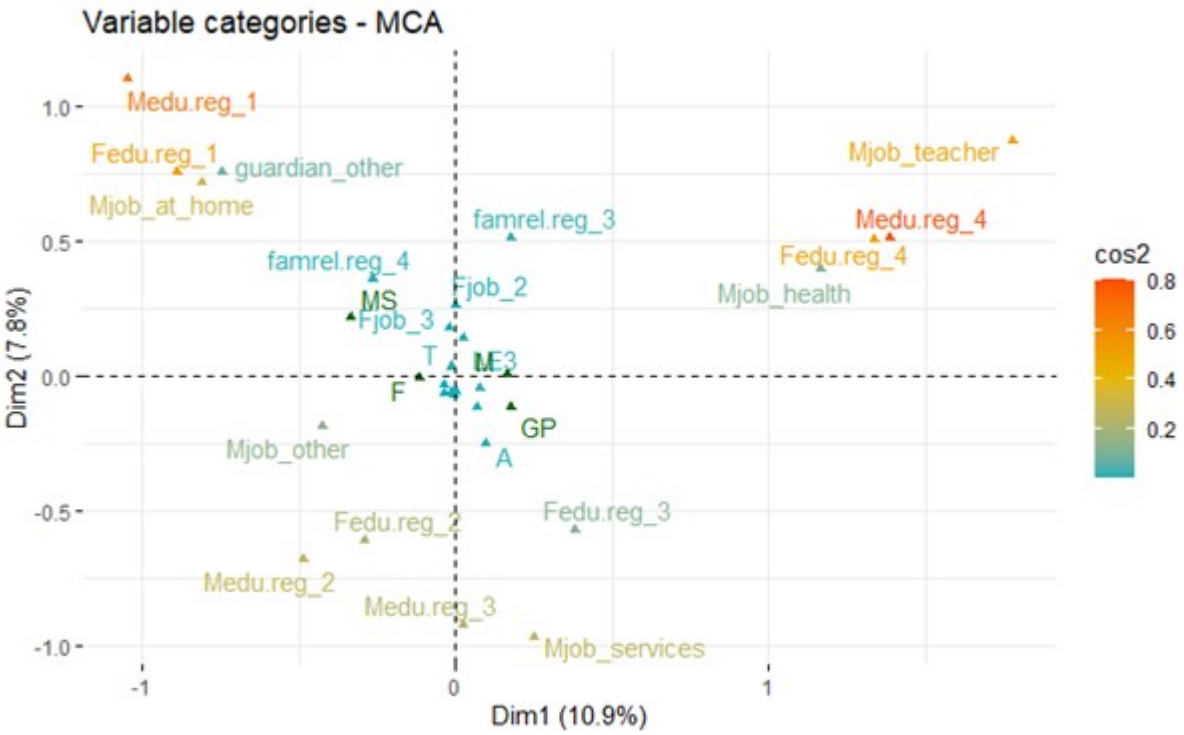


Le graphique ci-dessus suggère de retenir les deux premiers axes qui permettent d'expliquer 84% de l'inertie.

Analyse des résultats des modalités

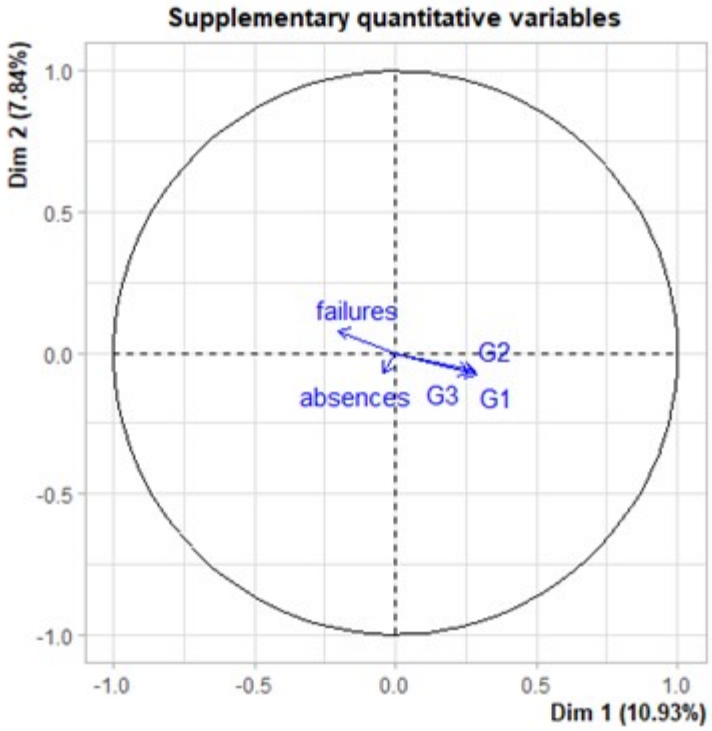
	coord	contrib	cos2
Medu.reg_4	1.385469959	2.366708e+01	7.086861e-01
Fedu.reg_4	1.338373574	1.615389e+01	4.400753e-01
Mjob_teacher	1.777687463	1.603082e+01	3.943370e-01
Medu.reg_1	-1.046094235	1.148789e+01	3.261053e-01
Fedu.reg_1	-0.888724703	1.007223e+01	3.054691e-01

Les modalités les plus contributives (contribution élevée) et les mieux représentées (cos2 élevé) du premier axe sont affichées sur le tableau ci-dessus. On remarque que le coté positif de l'axe 1 est caractérisé par les élèves dont les parents ont un niveau scolaire élevé ou dont la mère est professeur. Le coté négatif est caractérisé par les élèves dont les parents ont un faible niveau scolaire.



Le deuxième axe est plus délicat à interpréter. Son côté positif est plutôt lié à un faible niveau d'enseignement des parents et au fait que la mère soit sans travail.

Interprétation des variables supplémentaires quantitatives

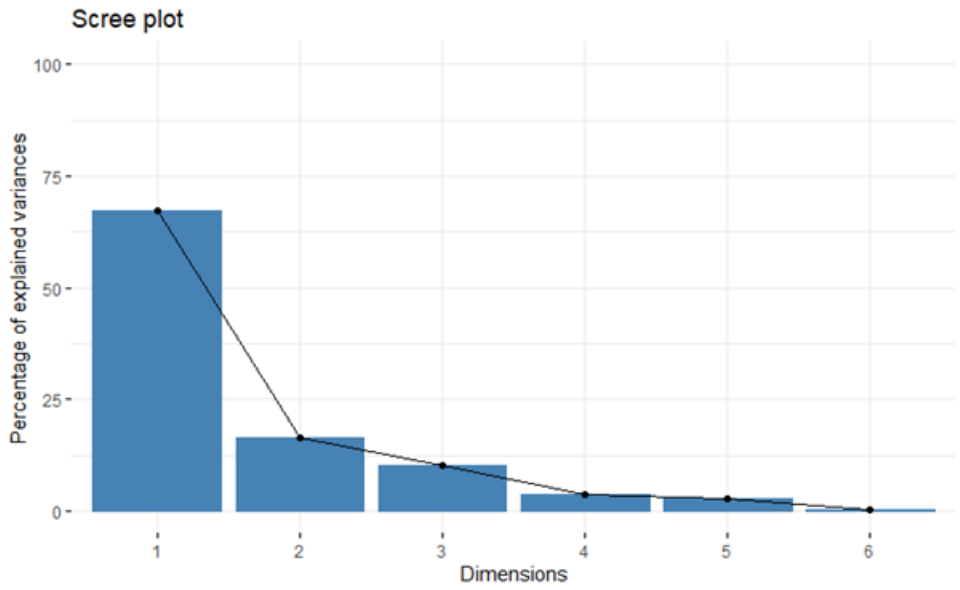


Le cercle des corrélations nous montre que les variables 'G1', 'G2', ou 'G3' sont projetées du côté positif de l'axe 1 mais sont faiblement liées à celui-ci. De ce fait, notre ACM ne nous permet pas de conclure quand à une liaison entre le fait d'avoir de bonnes notes et le fait d'avoir des parents avec un niveau scolaire élevé.

3.1.2 2ème ACM

Nous considérons les variables 'activities', 'paid', 'studytime', 'freetime', 'schoolsup', 'higher', 'internet', 'nursery', 'address'.

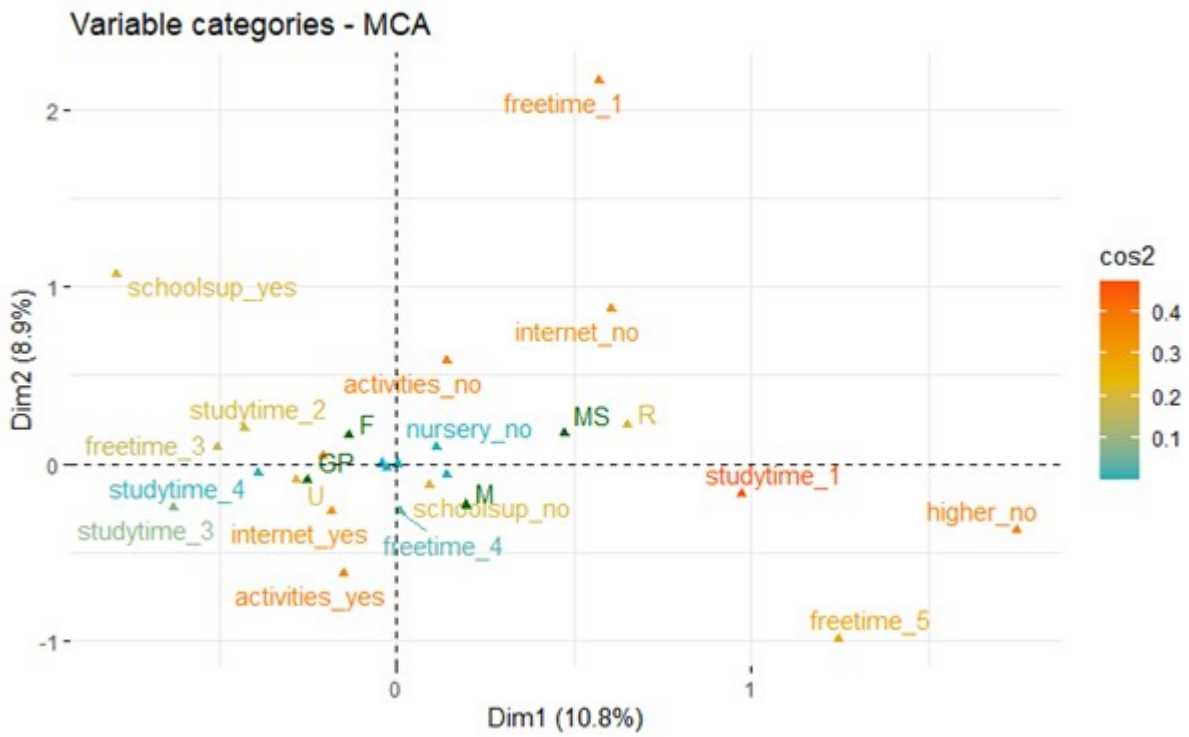
Choix du nombre d'axe



Comme précédemment, le graphique suggère de retenir les 2 premiers axes qui permettent d'expliquer 83% de l'inertie.

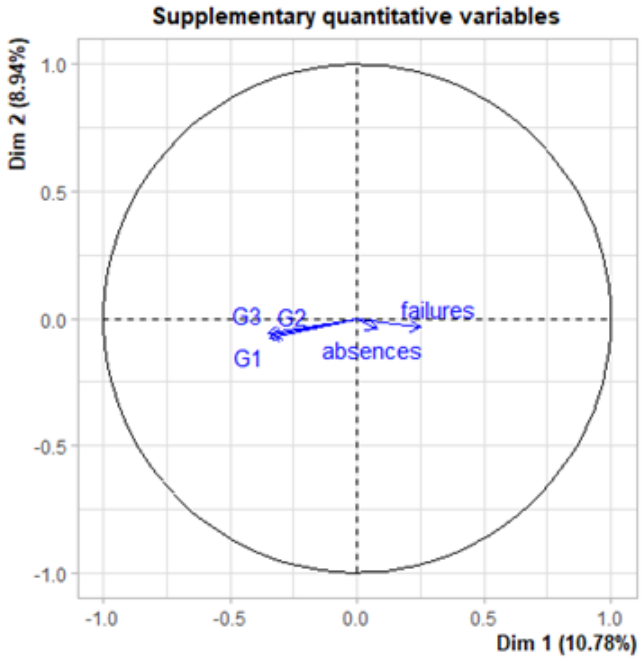
Analyse des résultats des modalités

Le coté positif de l'axe 1 est caractérisé par les élèves qui n'étudie pas beaucoup et qui ne veulent pas poursuivre dans l'enseignement supérieur. Il n'y a pas vraiment de variable qui caractérise son côté négatif comme on peut le voir ci-dessous :



Le côté positif de l'axe 2 est caractérisé par les élèves qui n'ont pas beaucoup de temps libre après les cours et qui ne pratiquent pas d'activités périscolaires. Son côté négatif est caractérisé par les élèves qui pratiquent une activité périscolaire.

Interprétation des variables supplémentaires quantitatives

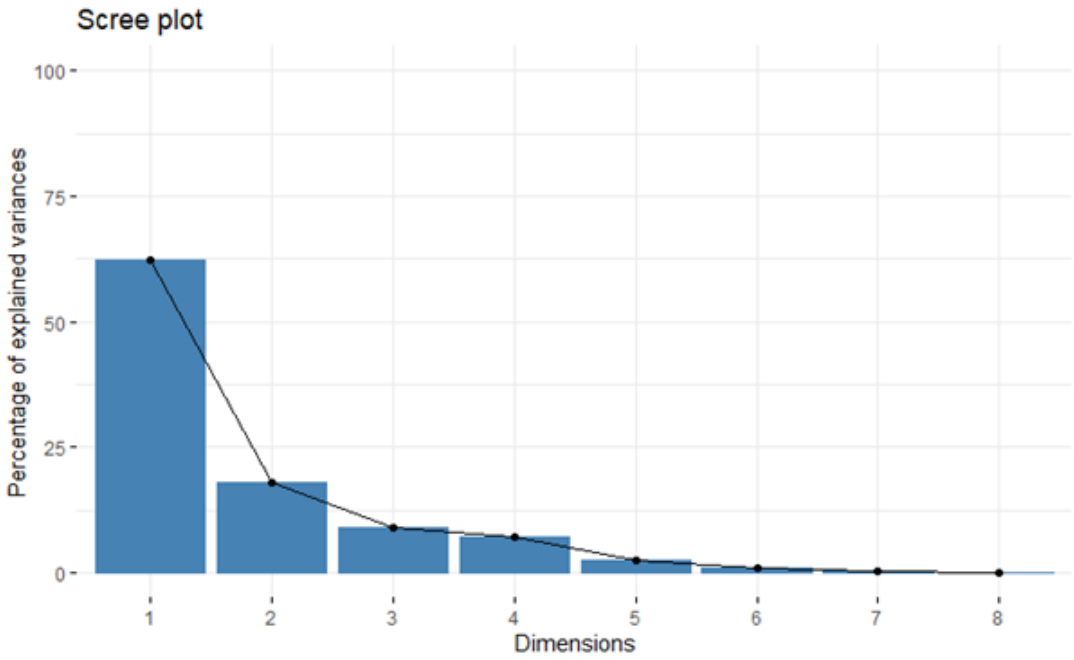


Le cercle des corrélations ne met pas suffisamment en évidence un lien entre l'obtention de bonnes notes en portugais et le fait de beaucoup travailler, d'avoir eu recours à des cours payants ou encore la pratique d'activité périscolaire.

3.1.3 3ème ACM

Nous considérons les variables 'Dalc','Walc','goout','romantic','health'.

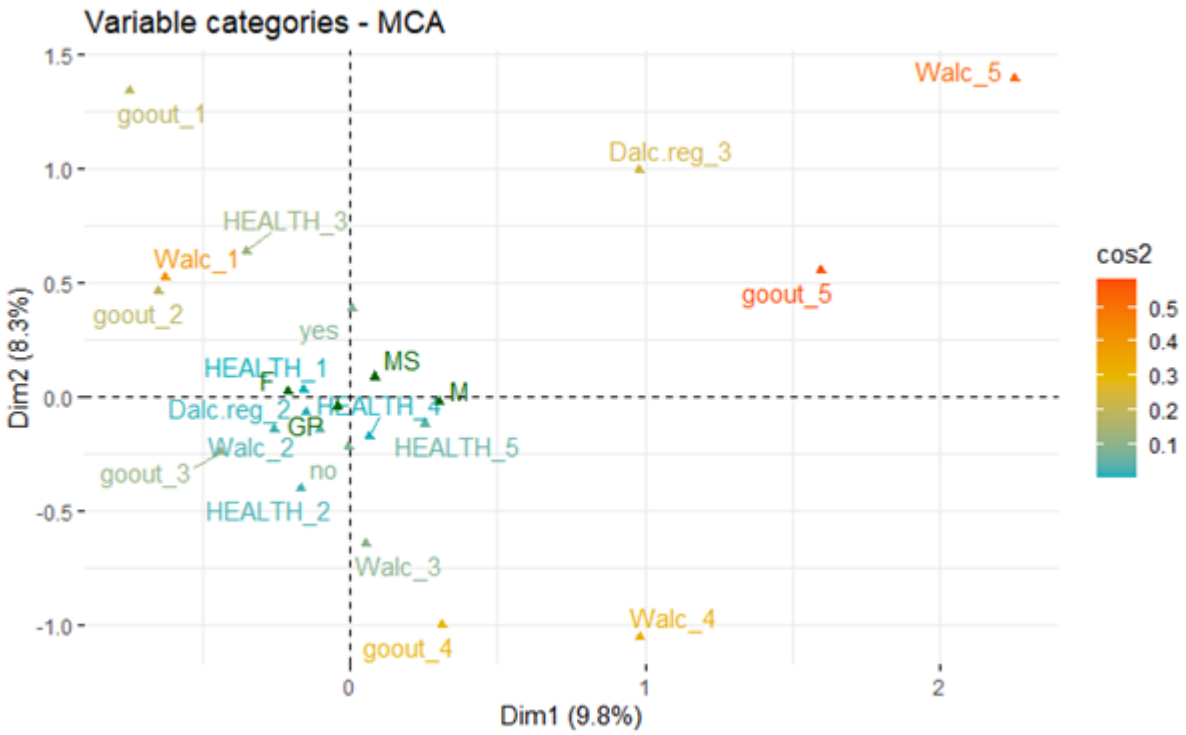
Choix du nombre d'axe



Nous retenons les 2 premiers axes pour lesquels 80% de l'inertie est expliquée.

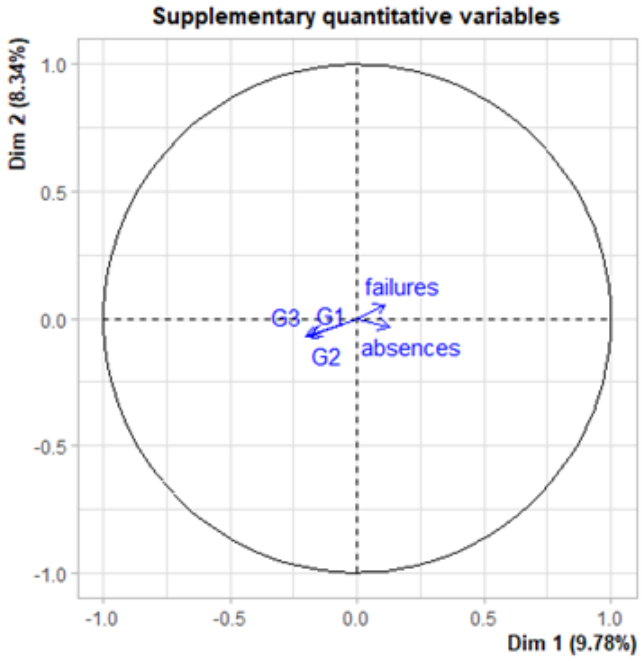
Analyse des résultats des modalités

Le côté positif de l'axe 1 est caractérisé par les élèves qui sortent beaucoup avec leur amis et qui consomment beaucoup d'alcool le week-end. Et le côté négatif par les élèves qui consomment peu d'alcool le week-end.



L'axe 2 n'apporte pas beaucoup d'information supplémentaire. Son côté négatif semble plutôt caractérisé par les élèves qui consomment assez souvent d'alcool le week-end et qui sortent souvent avec leur amis.

Interprétation des variables supplémentaires quantitatives

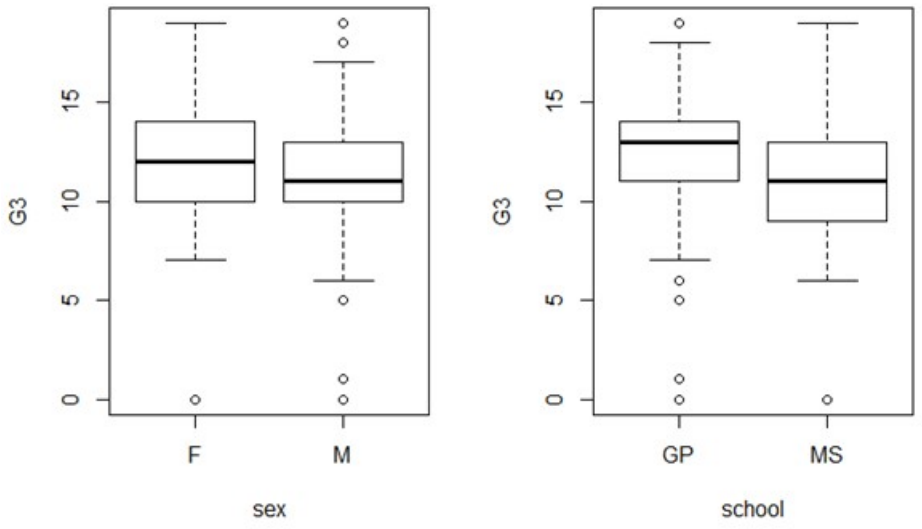


Là encore, le cercle des corrélations nous montre qu'il n'y a pas suffisant de lien entre le fait d'avoir de bonne notes et le fait de beaucoup sortir avec des amis ou de consommer beaucoup d'alcool.

3.2 Conclusion

Les ACM que nous avons réalisé ont mis en évidence différents profils d'élèves. La première ACM à mis en évidence une différence entre les élèves marquée par le niveau d'étude des parents. Il y a une association forte entre des modalités : les élèves dont la mère à un niveau d'étude élevé semblent aussi avoir un père dont le niveau d'étude est élevé. Il en est de même pour les parents avec un faible niveau d'étude. La deuxième ACM semble montrer une association entre le fait de travailler peu et le fait de ne pas vouloir poursuivre dans l'enseignement supérieur et la dernière montre une association entre le fait de beaucoup sortir entre amis et le fait de consommer beaucoup d'alcool le week-end. Après étude de la projection des variables supplémentaires quantitatives pour chacune de ces ACM, nous ne pouvons tirer aucune conclusion entre les notes obtenues et le profil des élèves suivant les différentes ACM réalisées.

Notons que pour chacune des ACM nous avons considéré les variables 'sex' et 'school' comme illustratives. Pour terminer sur cette partie sur la 'performance' des étudiants, nous avons voulu tester si le sexe et l'école étaient liées aux notes obtenues. Les variables 'G1', 'G2', 'G3' étant très corrélées, nous avons uniquement considéré 'G3'.



Les boxplots montre que en moyenne les filles ont eu de meilleures notes que les garçons et les élèves de l'école GP ont eu de meilleures notes que l'école MS. Nous avons réalisé deux tests de student et ces derniers nous ont confirmé que les différences de notes entre filles et garçons et entre les deux écoles étaient significatives.

4 Étude de la qualité des relations familiales des élèves

L'objectif de cette étude est de pouvoir répondre à certaines questions qui permettront d'avoir un aperçu sur la qualité des relations familiales des élèves de notre jeu de données. Il s'agira principalement d'observer les influences des autres variables par rapport à celle qui mesure la qualité des relations familiales à savoir la variable « FAMREL ». Dans la suite, nous tenterons de répondre aux questions suivantes :
i) Quelles sont les variables qui expliquent au mieux la qualité de relation familiale des élèves ?
ii) Y'a-t-il des modalités qui sont liées ou qui s'opposent ?
iii) Quelles sont les modalités qui sont associées à une bonne ou mauvaise qualité de relation familiale d'un élève ?

4.1 Analyse à Composantes Multiples

Dans une première approche, nous allons nous intéresser aux variables qualitatives de notre jeu de données qui seront considérées comme variables explicatives et à la variable cible «FAMREL ». C'est la raison qui explique notre choix pour la technique de réduction de dimensions ACM.

4.1.1 Pré-traitements des données

L'ACM étant très sensible aux disparités qui peuvent exister sur les effectifs d'une variable qualitative, nous avons commencé par faire un regroupement de modalités en vue de faire disparaître les problèmes liés aux modalités de faible effectif. Nous avons regroupé par exemple les modalités de :
• La variable cible «FAMREL » en :
o 1 ou 2 => FamrelFaible
o 3 => FamrelMoyen
o 4 ou 5 => FamrelBon
• La variable «FAILURES » en :
o 0 => Failures0
o 1, 2 ou 3 => FailuresAtLeast1
• La variable « Medu » :
o 0 ou 1 => MeduFaible
o 2 ou 3 => MeduMoyen
o 4 => MeduBon
• Etc.

Nous rappelons que ces regroupements de modalités se sont faits une fois après avoir observé les effectifs résultants du croisement entre la variable «FAMREL » et chacune des autres

variables concernées. Par exemple, en ce qui concerne le regroupement de la variable « Medu », vous pouvez observer sur les tableaux ci-dessous que nous avons fait un choix de regroupement qui permet d'aboutir à des effectifs tous supérieurs à 5.

	0	1	2	3	4
famrel_Bon	0	38	52	39	51
famrel_Faible	4	34	41	31	42
famrel_Moyen	2	71	93	69	82

FIGURE 1 – Avant le regroupement

	Medu_Bon	Medu_Faible	Medu_Moyen
famrel_Bon	51	38	91
famrel_Faible	42	38	72
famrel_Moyen	82	73	162

FIGURE 2 – Après le regroupement

4.1.2 Sélection de variables

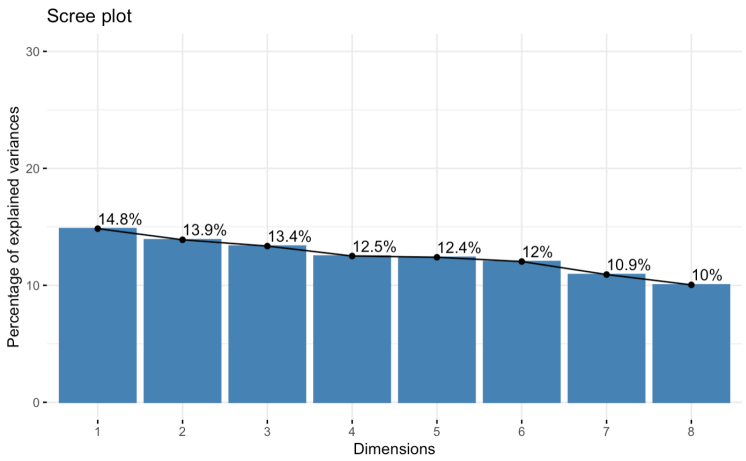
Étant en présence de 26 variables qualitatives si on omet la variable cible FAMREL, effectuer une ACM avec toutes ces dernières contribuera à énormément diluer l'inertie des axes et complexifier les interprétations. Pour ce faire, nous avons décidé de réaliser des tests de Khi2 entre FAMREL et chacune des autres 26 variables en vue de sélectionner que celles qui donnent une P-Value < 0.05 dans les résultats du test.

À l'issu de ces tests, seulement 4 variables sur 26 se retrouvent à avoir une influence sur la variable FAMREL. Il s'agit de : "Fjob", "failures","internet" et "freetime".

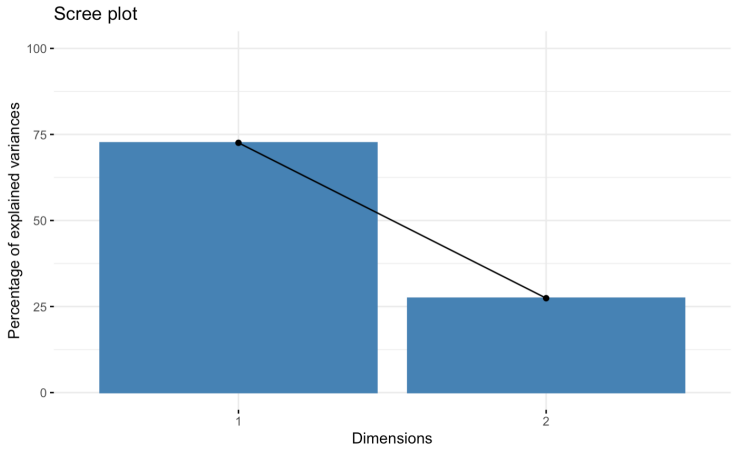
L'ACM a donc été implémentée avec ces 4 variables et avec FAMREL comme variable qualitative supplémentaire.

4.1.3 Choix du nombre d'axes

Malgré le fait qu'on ait réduit le nombre de variables, l'ACM implémentée présente quand même une inertie assez diluée sur les axes factoriels. Pour atteindre 50% de variance cumulée, il faut choisir les 4 premières dimensions, ce qui fait beaucoup et ne permettra pas d'expliquer une bonne partie de l'inertie du nuage des modalités.



Pour y pallier, comme vu en cours, nous avons appliqué la correction de Benzécri :



Avec la correction de Benzécri, plus de 73% de l'inertie du nuage est expliquée par le premier axe. Puis environ 27% est expliquée par le second axe.Nous allons alors baser notre analyse sur les 2 premiers axes.

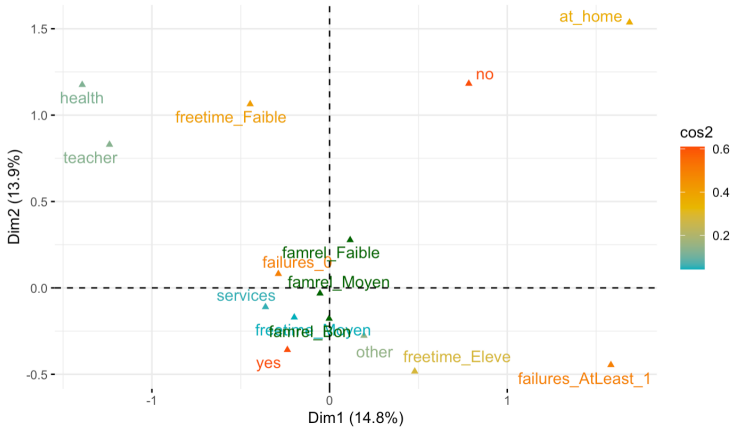
4.1.4 Interprétation des modalités sur le premier axe

En récupérant les coordonnées des modalités sur le premier axe factoriel, leurs contributions à l'axe et leurs Cos2 qui représentent la qualité de leur représentation sur l'axe dans le tableau ci-dessous, nous avons pu aboutir aux interprétations suivantes :

	coord	contrib	cos2
failures_0	-0.2883913	5.924086	0.45660069
failures_AtLeast_1	1.5832681	32.523232	0.45660069
at_home	1.6881809	15.530021	0.19719621
no	0.7835438	12.027859	0.18615478
yes	-0.2375806	3.647001	0.18615478
freetime_Eleve	0.4791567	7.327825	0.14014747
teacher	-1.2386691	7.166348	0.09010578
health	-1.3922812	5.784510	0.07122090
freetime_Faible	-0.4471288	3.942699	0.06114380
services	-0.3607594	3.056319	0.05033477
other	0.1934834	1.782536	0.04871968
freetime_Moyen	-0.1988406	1.287563	0.02493450

Le côté positif de l'axe 1 est matérialisé par les modalités failuresAtLeast1(cette modalité indique que l'élève a déjà redoublé une classe au moins 1 fois), athome(Le père ne travaille pas) et no(C'est-à-dire qu'il n'y a pas internet dans le domicile familial de l'élève). Le côté négatif de l'axe est matérialisé par les modalités failures0(l'élève n'a jamais redoublé) et yes(Présence d'internet à la maison). Ces modalités sont caractérisées par les Cos2 et contributions les plus élevés sur l'axe 1.

Ci-dessous, nous affichons la projection des modalités sur le plan factoriel avec un gradient de couleur informant sur la qualité de représentation des modalités (le Cosinus carré). Nous avons également projeté les modalités de la variable cible FAMREL :



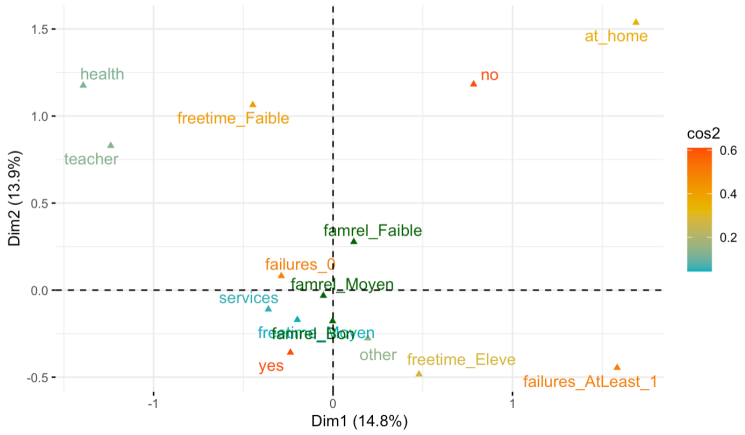
Sur ce nuage, on peut voir comme annoncé précédemment que les modalités failuresAtLeast1, athome et no s'opposent aux modalités failures0 et yes sur l'axe 1. Les modalités de la variable cible (FamrelFaible, FamrelMoyen et FamrelBon) sont assez proches du centre de gravité du nuage. On n'observe pas une tendance qui se dessine sur l'axe 1 avec ces modalités. On peut même voir que la modalité FamrelMoyen correspond au profil moyen. C'est-à-dire qu'une grande partie des élèves sont dans la catégorie de ceux qui ont une qualité de relation familiale moyenne. On n'observe pas particulièrement de modalités pouvant être associées.

4.1.5 Interprétation des modalités sur le deuxième axe

	coord	contrib	cos2
no	1.18234430	29.2615338	0.423872780
freetime_Faible	1.06379737	23.8447971	0.346102732
at_home	1.53686730	13.7516321	0.163430586
yes	-0.35850199	8.8724731	0.423872780
freetime_Eleve	-0.48338858	7.9681975	0.142633923
health	1.17555089	4.4059840	0.050773414
other	-0.27640652	3.8868237	0.099429102
teacher	0.82932673	3.4323064	0.040391814
failures_AtLeast_1	-0.44595197	2.7568199	0.036224619
freetime_Moyen	-0.17045263	1.0109120	0.018323063
failures_0	0.08122987	0.5021530	0.036224619
services	-0.11050092	0.3063674	0.004722419

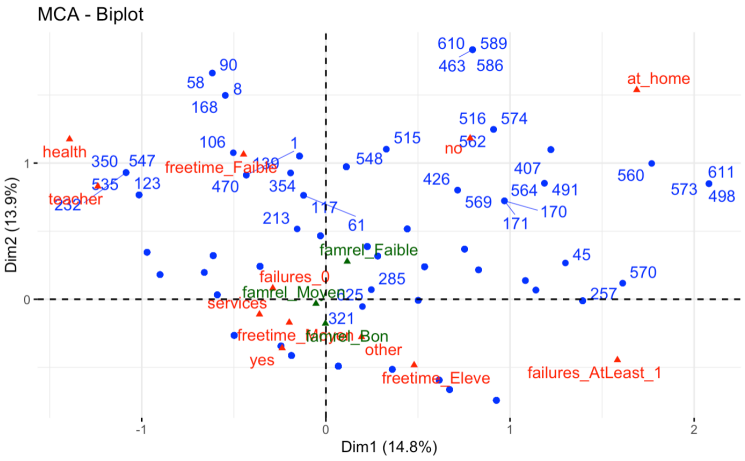
Le côté positif de l'axe est caractérisé par les modalités no(Pas de présence d'internet), freetimeFaible(Faible quantité de temps libres après les cours) et athome(Le père ne travaille pas). Le côté négatif de l'axe 2 lui est caractérisé par les modalités yes(Présence d'internet à la maison) et freetimeelevee(Grande quantité de temps libres après les cours).

Ci-dessous est affiché la projection des modalités sur le plan factoriel avec comme gradient de couleur, la qualité de représentation des modalités sur le deuxième axe factoriel :



Sur ce nuage, on peut confirmer l'opposition entre les modalités référencées plus haut selon le 2ème axe. Alors les élèves qui ont une faible quantité de temps libre après les cours, qui n'ont pas internet à la maison et dont le père ne travaille pas s'opposent à ceux qui ont une quantité de temps libre élevée après les cours et qui ont internet chez eux.

4.1.6 Interprétation du croisement entre les modalités et les individus sur le même plan factoriel



Sur ce nuage modalités/individus, on n'observe pas particulière un regroupement de certains individus autour d'une modalité donnée. On y voit des individus très dispersés sur le nuage. On ne peut donc pas aboutir à une conclusion concrète.

4.1.7 Conclusion

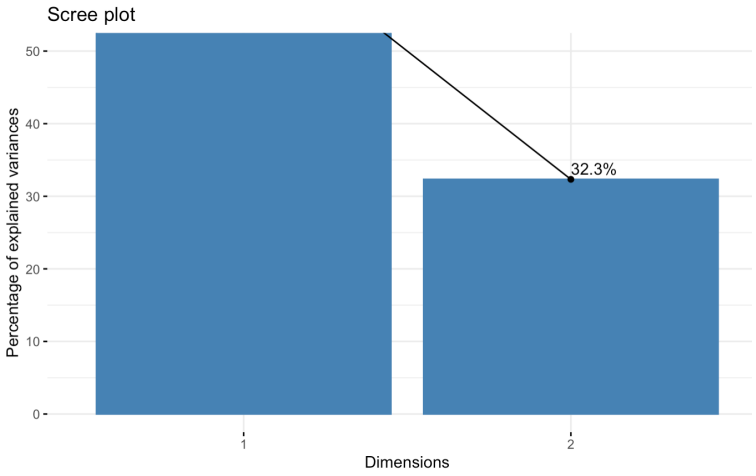
Face à aux modalités de la variable cible qui se retrouvent quasiment au centre de gravité du nuage, cette analyse n'a pas réussi à pouvoir dégager les modalités qui sont liées à la bonne

ou mauvaise qualité de relation familiale des élèves. Que ce soit selon l'axe 1 ou l'axe 2 du plan factoriel, on n'arrive pas à observer une tendance claire en ce qui concerne la qualité des relations familiales. Cependant grâce à l'analyse des modalités selon l'axe 1, nous avons réussi à identifier 2 groupes d'élèves qui s'opposent : ceux qui ont redoublé au moins 1 fois, dont le père ne travaille pas et qui n'ont pas internet chez eux d'un côté et ceux qui ont internet chez eux et qui n'ont jamais redoublé d'un autre. L'analyse sur le deuxième axe a permis d'opposer les élèves qui ont une faible quantité de temps libre après les cours, qui n'ont pas internet à la maison et dont le père ne travaille pas à ceux qui ont une quantité de temps libre élevée après les cours et qui ont internet chez eux. De part ces 2 analyses, on a pu mettre en évidence des modalités qui s'opposaient par contre aucune association de modalités significative n'a été remarquée.

4.2 AFC

Comme vu dans l'étude précédente qui a fait intervenir une ACM, nous n'avons pas pu répondre à certaines questions de notre problématique. De ce fait, nous avons décidé de faire une AFC qui permettra de croiser la variable cible FAMREL(mesurant la qualité de relation familiale des élèves) et la variable FAILURES(mesurant le nombre de classes redoublées). Pourquoi avoir choisi FAILURES ; parce que d'une part, le test de KHI2 avec la variable Famrel a donné une p-value$<$0.05 et d'autre part, parce qu'on a estimé que dans la vie courante, la qualité des relations qu'entretiennent les enfants et leurs parents influe beaucoup sur leurs résultats scolaires. Ceci dans le but d'observer les éventuelles associations de profils qui pourraient exister et de pourquoi pas pouvoir répondre aux questions du type :
Quels profils sont associés au profil des élèves ayant une bonne ou une mauvaise qualité de relation familiale ?

4.2.1 Choix du nombre d'axes



A l'aide de ces histogrammes, on peut voir qu'environ 68% de l'inertie du nuage est représenté par l'axe 1. Nous avons alors fait le choix de ne baser notre interprétation uniquement sur cet axe.

4.2.2 Interprétation sémantique de l'axe 1 avec les profils lignes

	coord	contrib	cos2
famrel_Bon	-0.06328064	22.84592	0.9881229
famrel_Moyen	0.23355986	63.24531	0.8619931
famrel_Faible	0.15413595	13.90877	0.2711863

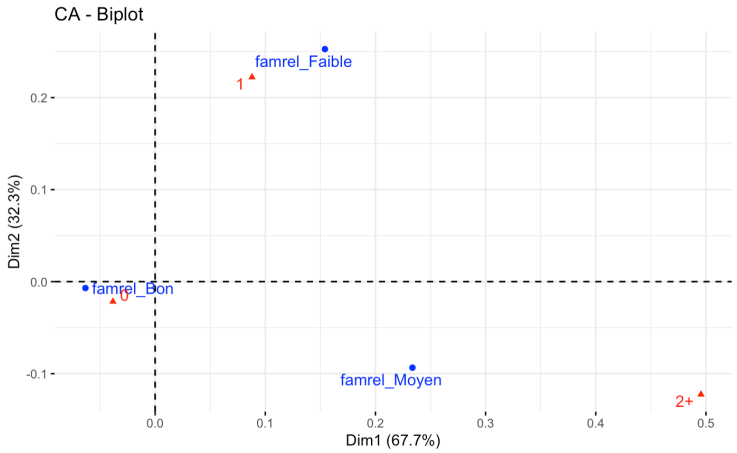
On constate à l'aide de ce tableau que le côté négatif de l'axe 1 est caractérisé par le profil des élèves qui ont une bonne qualité de relation familiale. Ce profil s'oppose à celui des élèves qui ont une qualité de relation familiale moyenne du côté positif de l'axe.
En d'autres termes, cela veut dire que la distribution du niveau d'échec scolaire n'est pas la même chez les élèves qui ont une bonne qualité de relation familiale que pour ceux qui ont une qualité de relation familiale moyenne.

4.2.3 Interprétation sémantique de l'axe 1 avec les profils colonnes

	coord	contrib	cos2
2+	0.49550599	84.553116	0.9423582
0	-0.03828539	9.237384	0.7580408
1	0.08790711	6.209500	0.1353739

En s'intéressant maintenant aux profils colonnes, on remarque que le côté positif de l'axe 1 est représenté par le profil des élèves qui ont redoublé au minimum 2 fois. Ce profil s'oppose à celui des élèves qui n'ont jamais redoublé. Ce qui montre que la distribution de la qualité de relation familiale en fonction des élèves qui ont redoublé plus d'une fois ou qui n'ont jamais redoublé n'est pas la même.

4.2.4 Représentation graphique simultanée des profils lignes et des profils colonnes



Par rapport à l'axe 1, on voit bien que le profil d'élèves ayant redoublés 2 fois ou plus s'oppose à celui des élèves qui n'ont jamais redoublé.
On peut également voir sur ce graphique que le profil des élèves qui n'ont jamais redoublé est très proche du centre de gravité du nuage. Ce qui amène alors à dire que le profil des élèves n'ayant jamais redoublé a une distribution de la qualité de relation familiale qui ressemble à la distribution de la qualité de relation familiale dans tout le jeu de données. Il est qualifié de profil moyen.

On remarque aussi que le profil des élèves qui ont une bonne qualité de relation familiale est assez proche du centre de gravité du nuage. Ce profil a une distribution du niveau d'échec scolaire qui ressemble à la distribution du niveau d'échec dans tout le jeu de données.

Les deux profils dernièrement décrit étant eux-mêmes proches l'un de l'autre, on peut conclure en disant qu'ils sont associés. De cette association, on peut déduire que chez les élèves qui n'ont jamais redoublé, on en a beaucoup plus qui ont une bonne qualité de relation familiale que dans l'ensemble du jeu de données.

4.2.5 Conclusion

En conclusion de cette étude, on se rend compte que le profil des élèves qui ont une bonne qualité de relation familiale s'oppose à celui des élèves qui ont une qualité de relation familiale moyenne. D'un autre côté, le profil des élèves qui ont redoublé au moins 2 fois s'oppose à celui des élèves qui n'ont jamais redoublé. Grâce à cette AFC entre FAMREL et FAILURES, l'observation des profils lignes et colonnes dans le même plan factoriel a permis de déduire une association entre le profil des élèves qui ont une bonne qualité de relation familiale et

celui des élèves qui n'ont jamais redoublé. C'est-à-dire que ceux qui n'ont jamais redoublé ont le plus souvent une bonne qualité de relation familiale.

5 Étude sur la santé des élèves

L'objectif de cette étude est de pouvoir déterminer les variables qui sont liées à la santé des élèves. Au terme de cette analyse, nous voulons être en mesure de pouvoir définir les modalités qui sont liées à la bonne ou mauvaise santé des élèves en fonction des données qui sont à notre disposition.

5.1 Pré-traitements des données

Notre jeu de données étant essentiellement composé de variables qualitatives, une ACM semblait être la technique idéale à utiliser en vue de pouvoir répondre aux questions que l'on se posait dans cette étude. Ainsi, tout comme dans les ACM réalisées précédemment, nous avons réalisé des regroupements de modalités pour garantir de grands effectifs et réduire le biais dans l'implémentation de l'ACM.

Il est important de souligner que dans cette étude, notre variable cible est la variable HEALTH qui est une variable mesurant le niveau de santé d'un élève donné. Cette variable qui était initialement composée de 5 modalités (1=> niveau de santé très bas à 5=>Très bonne santé) a été regroupée en 3 modalités :
o 1 ou 2 => santé faible
o 3 => santé moyenne
o 4 ou 5 => santé bonne

5.2 Sélection de variables

Face aux grands nombre de variables, nous avons effectué un test de KHI2 entre HEALTH et chacune des autres variables qualitatives. En ne retenant que les variables qui ont donnée une P-Value inférieure à 5%, 3 variables ont été retenues : SEX, MEDU(Le nniveau d'éducaton de la mère) et STUDYTIME(Le temps de travail de l'élève à la maison).

Dans la suite de cette étude, nous interpréterons les résultats d'une ACM réalisée avec ces 3 variables en incluant la variable HEALTH comme variable qualitative supplémentaire.

5.3 Choix du nombre d'axes

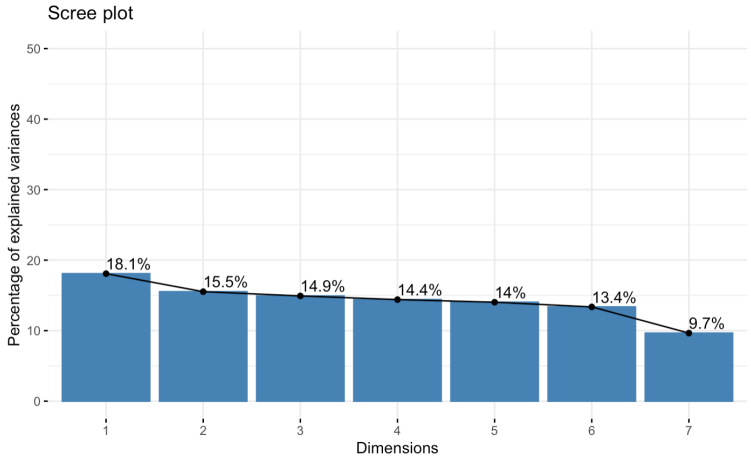


FIGURE 3 – Pourcentages de variances expliquées par les axes

Comme vu sur le graphique ci-dessus, l'ACM implémentée présente également ici une inertie assez diluée sur les axes factoriels. Avec l'application de la correction de Benzécri, On a :

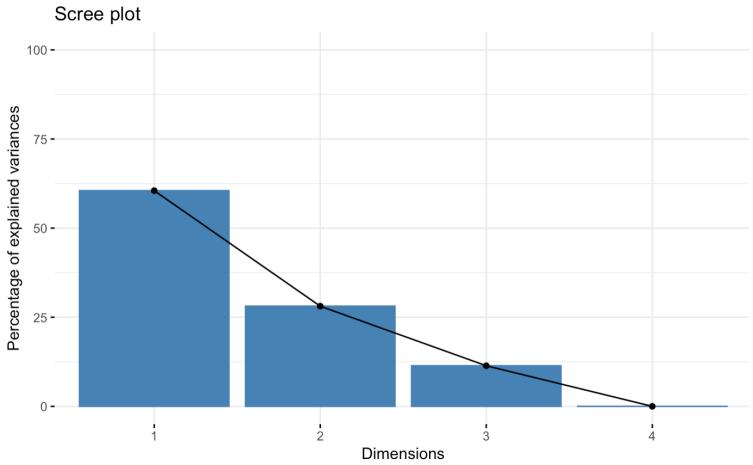


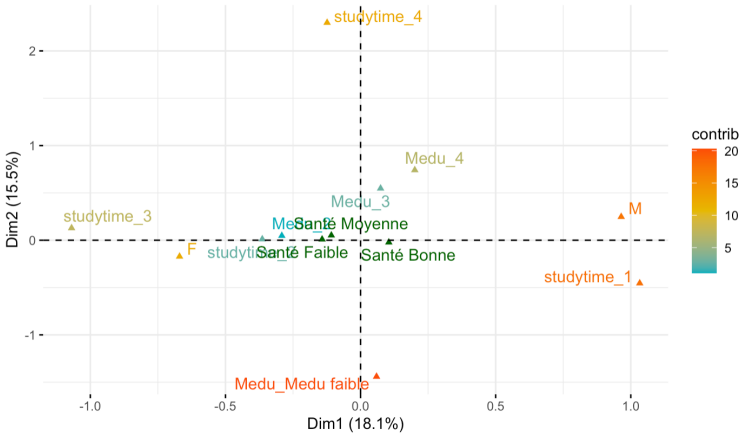
FIGURE 4 – Pourcentages de variances expliquées par les axes avec la correction de Benzécri

On arrive à obtenir environ 60% d'inertie expliquée par le premier axe et environ 27% d'inertie expliquée par le second axe. Nous baserons alors notre interprétation sur les deux premiers axes.

5.4 Analyse des résultats des modalités sur le premier axe

	coord	contrib	cos2
M	0.96453095	30.12274881	0.6461229948
studytime_1	1.03304841	27.53961299	0.5177209829
F	-0.66988311	20.92076027	0.6461229948
studytime_3	-1.06941002	13.50332931	0.2009653355
studytime_2	-0.36373633	4.91194869	0.1173045210
Medu_2	-0.29142026	1.92279534	0.0341170475
Medu_4	0.20062363	0.85739878	0.0148601739
Medu_3	0.07406840	0.09282422	0.0014952386
studytime_4	-0.12382601	0.06532395	0.0008740242
Medu_Medu faible	0.05905722	0.06325763	0.0010393510

Avec ce tableau, on voit que le côté positif de l'axe 1 est caractérisé par les modalités M et Studytime 1. Et le côté négatif quant à lui par les modalités F et studytime 3. Ce qui se dégage ici, c'est que les élèves de sexe masculin et de temps d'étude faible s'opposent à ceux de sexe féminin et de temps d'étude moyen.



Sur le nuage des modalités ci-dessus, on peut effectivement observer l'opposition mise en évidence dans l'analyse du tableau précédent. On peut aussi remarquer que par rapport aux autres modalités, les modalités "Masculin" et "Studytime faible" sont assez proches. On peut donc déduire avec beaucoup de réserve une association entre elles. Ce qui est également très perceptible sur ce nuage, c'est qu'on ne distingue pas de tendance en observant les modalités de la variable cible HEALTH. En effet, celles-ci sont toutes regroupées autour du centre de gravité du nuage. On ne peut donc pas conclure à un effet santé sur l'axe 1.

5.5 Analyse des résultats des modalités sur le deuxième axe

	coord	contrib	cos2
Medu_Medu faible	-1.44086327	43.846217577	0.6186739109
studytime_4	2.29838869	26.206880253	0.3011248699
Medu_4	0.74134787	13.632718512	0.2029101622
studytime_1	-0.45434028	6.202955574	0.1001421488
Medu_3	0.54738033	5.903281595	0.0816625611
M	0.24756754	2.310837364	0.0425667291
F	-0.17193986	1.604915767	0.0425667291
studytime_3	0.12845083	0.226853701	0.0028993888
Medu_2	0.04767141	0.059914144	0.0009129522
studytime_2	0.01120264	0.005425514	0.0001112710

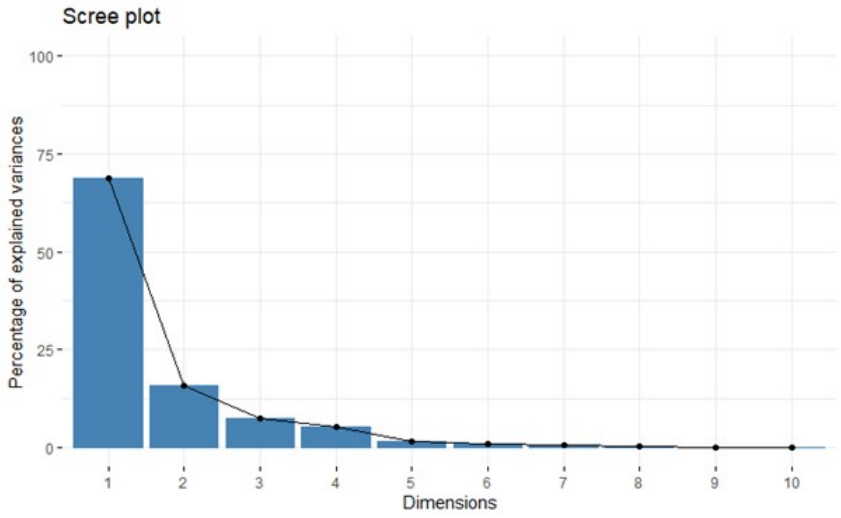
Avec ce tableau, on voit que le côté négatif de l'axe 2 est caractérisé par la modalité Medu Faible et le côté positif de l'axe par les modalités studytime 4 et Medu 4. Cela veut dire que les élèves dont la mère a un niveau d'éducation faible s'opposent aux élèves dont la mère a un niveau d'éducation élevé et qui étudient beaucoup à la maison.

5.6 Conclusion

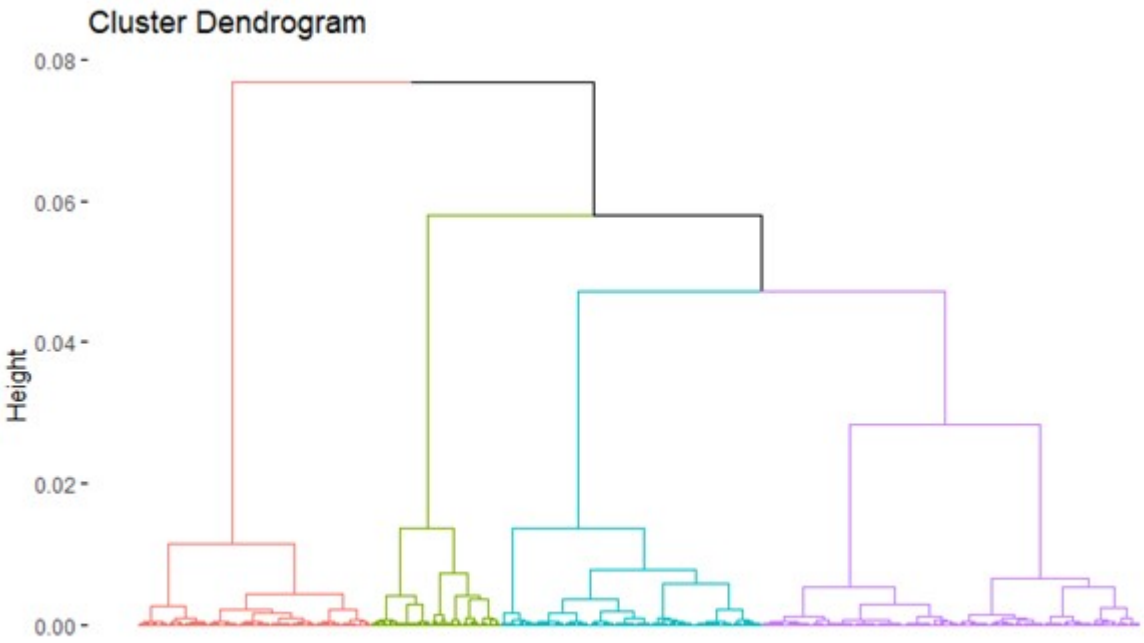
Au cours de cette étude, nous avons été à même de détecter des oppositions existantes entre les élèves en ce qui concerne certaines modalités. Nous savons en l'occurrence que les élèves qui sont de sexe masculin et qui étudient peu à la maison, s'opposent à ceux de sexe féminin et qui ont un temps d'étude moyen chez eux. D'un autre côté, les élèves dont la mère a un niveau d'éducation faible s'opposent à ceux dont la mère à un niveau d'éducation élevé et qui ont un temps d'étude élevé chez eux.
Cependant, nous n'avons pas été en mesurant de déterminer les modalités qui pourraient être associées à la bonne ou mauvaise qualité de santé d'un élève.

6 Étude générale et constitution de classes d'élèves

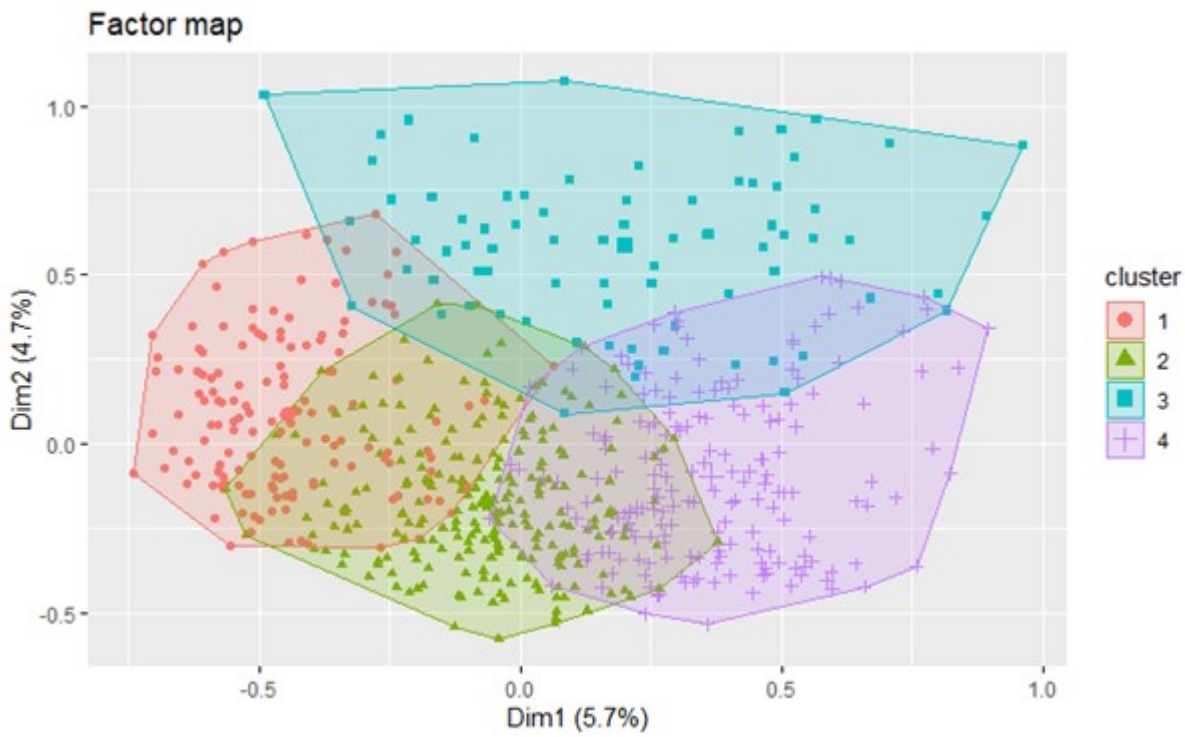
Nous souhaitons pour cette dernière partie regrouper les élèves dans des classes (clusters) et voir quelles sont leurs spécificités. Pour cela nous allons réaliser une classification ascendante hiérarchique (CAH) sur les composantes de l'ACM dont les variables d'études sont l'union des variables utilisées dans les trois ACM successives de la première partie sur l'étude des performances des élèves. Le graphique ci-dessous suggère de retenir les 4 premiers axes qui permettent d'expliquer 97% de l'inertie.



Nous avons donc réaliser une CAH sur les 4 composantes issues de cette ACM en utilisant la fonction HCPC du package 'factominer' avec une distance de Ward. Nous obtenons le dendrogramme suivant :



L'arbre suggère une coupure en 4 classes. Nous pouvons visualiser les individus et leurs classes d'appartenance sur le premier plan factoriel de l'ACM :



Les classes 1,2,3 et 4 sont respectivement composée de 149, 239, 76 et 185 élèves.

Nous souhaitons maintenant déterminer en quoi ces 4 classes diffèrent, voir quelles sont les caractéristiques des élèves représentées dans ces classes. Elles peuvent être décrites par les variables/modalités, les axes principaux et les individus. Le petit tableau ci-dessous nous montre quelles sont les variables qui caractérisent la plupart des groupes.

	p.value	df
Medu.reg	6.769650e-154	9
Mjob	1.022605e-96	12
Fedu.reg	1.341566e-70	9
Walc	2.225303e-65	12
famrel.reg	1.841396e-47	9
goout	2.257584e-32	12

Il s'agit des variables liées aux niveau d'étude des parents, au métier de la mère, à la consommation d'alcool le week-end, aux sorties entre amis et à la qualité des relations familiales. On peut regarder le détail de chaque classe, modalité par modalité. Ci-dessous, les colonnes correspondent à des proportions. La 1ère correspond à la proportion d'élèves possédant une modalité qui se trouvent dans une classe. La 2ème correspond à la proportion d'élève se trouvant dans la classe qui possède la modalité. La 3ème correspond à proportion d'individus possédant la modalité toute classe confondue.

	Cla/Mod	Mod/Cla	Global
Medu.reg=Medu.reg_4	82.285714	96.6442953	26.964561
Mjob=Mjob_teacher	91.666667	44.2953020	11.093991
Fedu.reg=Fedu.reg_4	67.187500	57.7181208	19.722650
Mjob=Mjob_health	75.000000	24.1610738	7.395994
internet=internet_yes	27.510040	91.9463087	76.733436
higher=higher_yes	25.517241	99.3288591	89.368259
Dalc.reg=Dalc.reg_1	29.781421	73.1543624	56.394453
school=GP	28.132388	79.8657718	65.177196
address=U	26.991150	81.8791946	69.645609
nursery=nursery_yes	25.911708	90.6040268	80.277350
activities=activities_yes	28.888889	61.0738255	48.536210
freetime=freetime_2	34.579439	24.8322148	16.486903
HEALTH=HEALTH_5	27.710843	46.3087248	38.366718
studytime=studytime_3	31.958763	20.8053691	14.946071
freetime=freetime_1	11.111111	3.3557047	6.933744

	Cla/Mod	Mod/Cla	Global
Medu.reg=Medu.reg_3	71.942446	41.8410042	21.417565
Medu.reg=Medu.reg_2	62.365591	48.5355649	28.659476
Mjob=Mjob_services	62.500000	35.5648536	20.955316
Fedu.reg=Fedu.reg_2	55.023923	48.1171548	32.203390
famrel.reg=famrel.reg_1	43.237251	81.5899582	69.491525
Dalc.reg=Dalc.reg_1	45.355191	69.4560669	56.394453
address=U	43.141593	81.5899582	69.645609
school=GP	43.498818	76.9874477	65.177196
Mjob=Mjob_other	47.674419	51.4644351	39.753467
freetime=freetime_3	47.808765	50.2092050	38.674884
Fedu.reg=Fedu.reg_3	54.198473	29.7071130	20.184900
Walc=Walc_1	46.963563	48.5355649	38.058552
sex=F	42.819843	68.6192469	59.013867
internet=internet_yes	40.562249	84.5188285	76.733436
schoolsup=schoolsup_yes	57.352941	16.3179916	10.477658

Classe 1 et classe 2

Commentaires sur la classe 1 : La quasi totalité des étudiants de la classe 1 ont une mère dont le niveau d'enseignement est élevé et la majeur partie des étudiants qui ont une mère avec un niveau d'enseignement élevé sont dans cette classe. Concernant les lignes suivantes, en regardant la 2ème colonne, nous pouvons dire par rapport aux élèves se trouvant dans cette classe que plus de la moitié ont un père possédant un niveau d'enseignement élévé (57%) , plus de 90% possèdent internet et la quasi totalité veulent poursuivre dans l'enseignement supérieur (modalité partagée par 89% des étudiants toute classe confondue). 73% consomment presque pas d'alcool en semaine et 80% sont issus de l'école GP. 81% habitent en milieu rural.

Commentaires sur la classe 2 : Par rapport à la deuxième colonne, on remarque que 81% des élèves de cette classe ont une mauvaise relation avec leur famille. 76% proviennent de l'école GP et 68% sont des filles. 70% ne consomment presque pas d'alcool en semaine. 89% ont une mère dont le niveau d'éducation est modéré (2 ou 3) et 77% ont un père dont le niveau d'éducation est aussi modéré.

	Cla/Mod	Mod/Cla	Global
Walc=Walc_5	91.1111111	53.947368	6.933744
goout=goout_5	48.1818182	69.736842	16.949153
famrel.reg=famrel.reg_4	79.4117647	35.526316	5.238829
freetime=freetime_5	42.6470588	38.157895	10.477658
sex=M	21.8045113	76.315789	40.986133
famrel.reg=famrel.reg_3	46.5116279	26.315789	6.625578
studytime=studytime_1	22.6415094	63.157895	32.665639
Dalc.reg=Dalc.reg_3	28.5714286	26.315789	10.785824
Walc=Walc_4	22.9885057	26.315789	13.405239
activities=activities_yes	15.2380952	63.157895	48.536210
Mjob=Mjob_services	18.3823529	32.894737	20.955316
higher=higher_no	21.7391304	19.736842	10.631741
Fjob=Fjob_2	23.5294118	15.789474	7.858243
HEALTH=HEALTH_5	15.6626506	51.315789	38.366718
paid=paid_yes	23.0769231	11.842105	6.009245

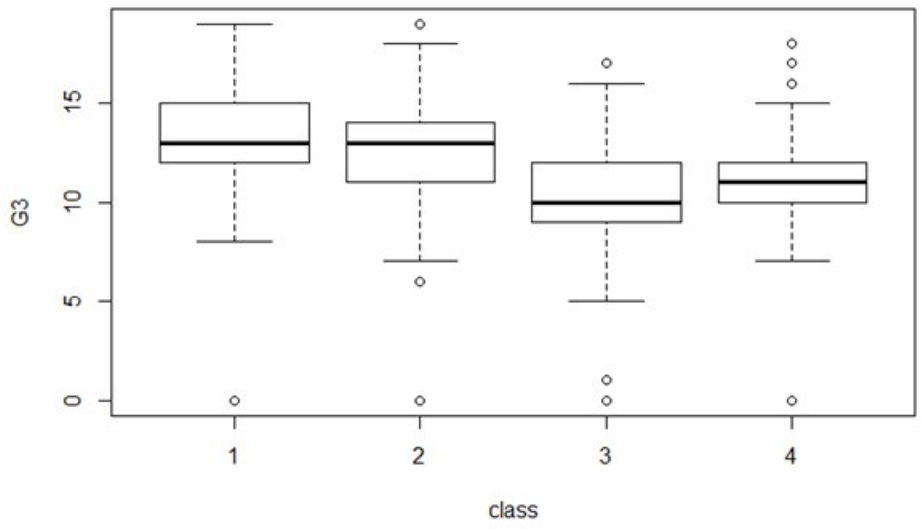
	Cla/Mod	Mod/Cla	Global
Medu.reg=Medu.reg_1	81.879195	65.9459459	22.958398
Fedu.reg=Fedu.reg_1	66.298343	64.8648649	27.889060
Mjob=Mjob_at_home	69.629630	50.8108108	20.801233
school=MS	49.557522	60.5405405	34.822804
internet=internet_no	56.291391	45.9459459	23.266564
address=R	49.746193	52.9729730	30.354391
higher=higher_no	57.971014	21.6216216	10.631741
Dalc.reg=Dalc.reg_2	41.314554	47.5675676	32.819723
Dalc.reg=Dalc.reg_3	54.285714	20.5405405	10.785824
activities=activities_no	35.628743	64.3243243	51.463790
guardian=guardian_other	56.097561	12.4324324	6.317411
studytime=studytime_1	37.264151	42.7027027	32.665639
sex=F	32.114883	66.4864865	59.013867
Walc=Walc_2	36.000000	29.1891892	23.112481
nursery=nursery_no	36.718750	25.4054054	19.722650

Classe 3 et classe 4

Commentaires sur la classe 3 : Par rapport à la 2ème colonne, plus de la moitié consomment beaucoup d'alcool le week-end, 70% sortent beaucoup avec leurs amis. Plus de 70% ont une relation avec leur famille qui est bonne ou moyenne. On peut noter que 76% sont des hommes. Une part non négligeable (63%) étudie très peu.

Commentaires sur la classe 4 : Par rapport à la 2ème colonne, une part importante des élèves ont une mère ou un père qui ont un faible niveau d'enseignement (plus de 64% pour les deux parents). La moitié ont leur mère qui n'a pas de travail (at_home). 60% proviennent de l'école MS. Une part non négligeable (45%) ne possèdent pas internet et plus de la moitié vivent en milieu rural.

Pour finir, ci-dessous les boxplots des notes obtenues en G3 pour chacune des classes (de 1 à 4) :



On constate que les élèves qui ont obtenu en moyenne les meilleures notes sont dans les classes 1 et 2. La classe 3 est celle des élèves qui ont obtenu en moyenne les moins bonnes notes. Pour conclure à un effet de la classe sur les notes obtenues, nous avons réalisé une anova à un facteur. Si la taille des effectifs de chaque classe est suffisante, l'homogénéité des variances n'est toutefois pas respectée. Nous nous sommes donc rabattus sur un test de Kruskal pour lequel la p-value à été bien inférieure au risque alpha de 0.05. Nous pouvons donc conclure à un effet significatif de la classe d'appartenance des individus sur les notes obtenues en G3. Compte tenu des observations précédentes et des boxplots, nous pouvons dire qu'une consommation d'alcool élevée le week-end et le fait de sortir beaucoup avec des amis (classe 3) ou le fait d'avoir notamment des parents de faibles niveau d'enseignement (classe 4) ont un impact négatif sur les notes.

7 CONCLUSION

Au terme de nos études, les premières ACM organisées par 'thème' ont permis de mettre en évidence l'association entre certaines modalités et des profils d'élèves. Néanmoins, au regard des projections des notes 'G1','G2' ou 'G3', ces ACM n'ont pas permis établir un lien suffisant avec les notes obtenues.
Pour ce qui est de la qualité des relations familiales, on observe que certaines modalités ont tendance à apparaître le plus souvent lorsque l'élève affirme avoir une bonne qualité de relation familiale chez lui. On sait par exemple que la plupart des élèves qui n'ont jamais redoublé ont affirmé avoir une bonne qualité de relation familiale chez eux.
En ce qui concerne la santé des élèves, aucune association de modalités n'a été détectée. Nous avons cependant réussi à observer lors de cette étude, des groupes d'élèves qui s'opposent ; À savoir les élèves dont la mère a un niveau de d'éducation scolaire faible et les élèves qui étudient beaucoup chez eux et dont la mère a fait de longues études.

Dans ces deux dernières études, nous n'avons pas fait mention de la performance des élèves mesurées par leurs notes, car les variables mesurant cette performance étaient très mal représentées et faiblement corrélées aux axes factoriels. Nous n'avons donc pas pu établir un lien entre d'une part, la qualité de vie d'un élève à travers ses rapports familiaux et sa santé et d'autre part ses performances dans le cours de Portugais.
Notre dernière étude a eu pour but de constituer des classes d'élèves en partant d'une ACM plus globale. Elle nous permis de mettre plus en lumière les différences entre les individus et de conclure à un effet de la classe d'appartenance des élèves sur les notes obtenues.