MACHINE LEARNING ET MODELES LINEAIRES GENERALISES

Rapport des notes de mathématiques

ACTUARIAT S7 - ESILV

Mathilde Barbier, Edmée Hogenmuller, Léa Pausé, Louis Bolzinger

TABLE DES MATIERES

INTRODUCTION

Les données de notre base de données concernent les notes de mathématiques des étudiants de deux lycées au Portugal. Il est intéressant de chercher à trouver ce qui explique une note bonne ou une mauvaise, et ce que nous allons essayer de faire à l'aide de ce dataset, du moins dans la limite de notre échantillon.

Les données concernent les notes des élèves, les caractéristiques démographiques et sociales des élèves. Elles ont été recueillies à l'aide de rapports scolaires et de questionnaires. Dans notre étude, nous allons donc essayer de comprendre et d'expliquer les notes de mathématiques annuelles des élèves. Nous ne nous occupons que d'une matière, elle combine différents paramètres pour performer, un sens logique, des connaissances théoriques, qui sont autant de paramètres qui pourraient être modélisés par nos variables.

PRESENTATION DE NOTRE BASE ET DES VARIABLES

Notre base de données contenait initialement 33 variables et 395 individus. Après avoir passé au peigne fin nos données pour déceler d'éventuels doublons, des anomalies inter et intra variables, nous n'avons pas trouvé d'erreurs de ce type, et n'avons donc pas eu à remplacer ou supprimer de la donnée.

Nos variables sont majoritairement qualitatives, même si elles sont exprimées selon des facteurs chiffrés, nous avons quelques variables quantitatives intéressantes comme le nombre d'absence. Nous avons toutefois dû poser certaines limites à notre étude, nous nous sommes aperçus que nous avions peu de variables quantitatives continues, ce qui nous a donné des résultats peu interprétables pour certaines méthodes.

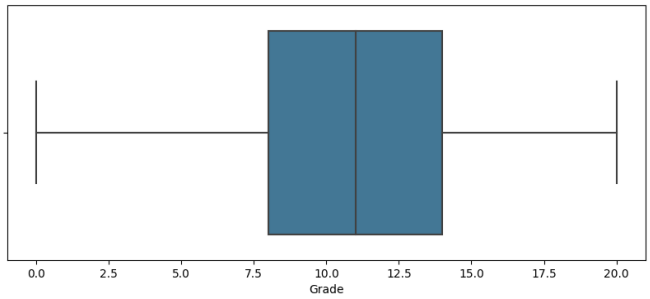
La page suivante présente les variables de notre base de données.

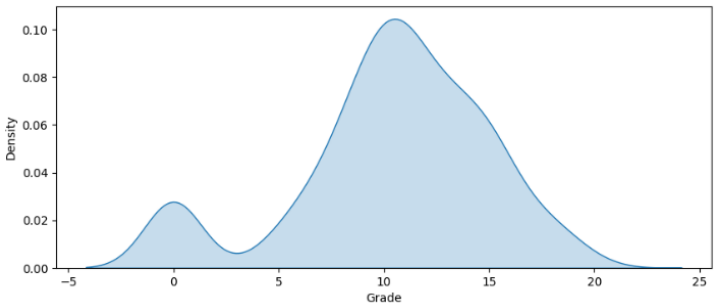
Présentation de nos variables :

Noms de variables	Description
school	École de l'étudiant, 2 possibilités : Gabriel Pereira ('GP') ou Mousinho da Silveira ('MS')
sex	Genre de l'élève : masculin ou féminin
age	Age de l'élève
famsize	Taille de la famille, 2 possibilités : moins de 3 ('LE3') et plus de 3 personnes ('GT3')
Pstatus	Situation des parents, 2 possibilités : habitent ensemble ('T'), n'habitent pas ensemble ('A')
Medu	Éducation scolaire de la mère de l'élève, note de 0 à 5
Fedu	Éducation scolaire du père de l'élève, note de 0 à 5
Mjob	Métier de la mère de l'élève : ('teacher', 'health', 'services', 'at_home', 'other')
Fjob	Métier du père de l'élève, mêmes possibilités que la mère
traveltime	Temps pour aller à l'école, note de 1 à 4 (entre <15min à plus 1h)
studytime	Temps accordé à étudier après les cours
failures	Nombre de fois où ils ont redoublé
schoolsup	Pratique sportive (oui ou non)
famsup	Support d'éducation familiale (oui ou non)
paid	Paiement cours supplémentaire pour étudier les cours (aide au devoir)
activities	Pratique d'activités extrascolaires
higher	Volonté de faire des études supérieure
internet	Accès à l'internet à la maison (oui ou non)
romantic	Élève est dans une relation amoureuse (oui ou non)
famrel	Qualité de la relation familiale, note de 1 à 5 (pas bien à très bien)
freetime	Temps libre après les cours, note de 1 à 5 (pas beaucoup à beaucoup)
goout	Temps passé à sortir avec des amis, note de 1 à 5 (pas beaucoup à beaucoup)
Dalc	Consommation d'alcool les jours de cours, note de 1 à 5 (pas beaucoup à beaucoup)
Walc	Consommation d'alcool le week-end, note de 1 à 5 (pas beaucoup à beaucoup)
health	Statut de santé de l'étudiant, note de 1 à 5 (pas bonne à très bonne)
absences	Nombre d'absences de l'étudiant (entre 0 à 93)
G3	Note finale en mathématiques = variable target
Grade_binomial	Classement bonne note à mauvaise note : (0 = en dessous de 10, 1 = au-dessus)

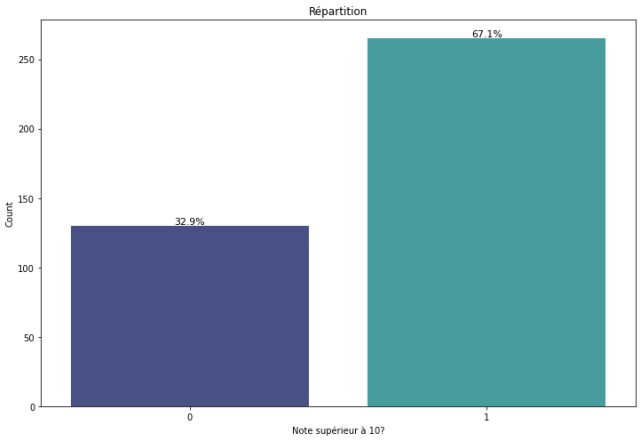
ANALYSE UNIVARIEE

1- PRESENTATION DE LA VARIABLE NOTES



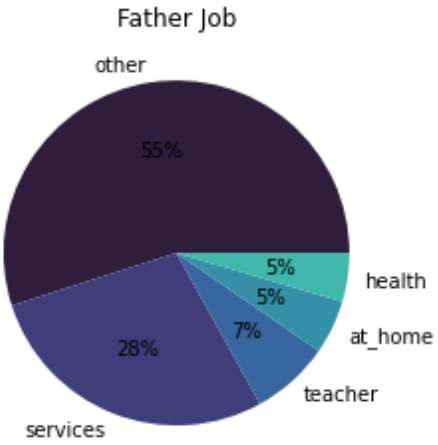


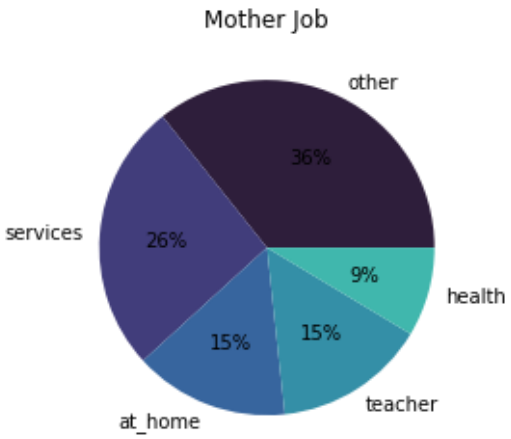
Les notes vont de 0 à 20, avec une majorité d'élève ayant une note entre 7,5 et 14 sur 20.
La moyenne des notes est de 10,45 et la médiane de 11.



La plupart des élèves (67,1%) ont une moyenne en maths au-dessus de 10.

2- LE TRAVAIL DES PARENTS

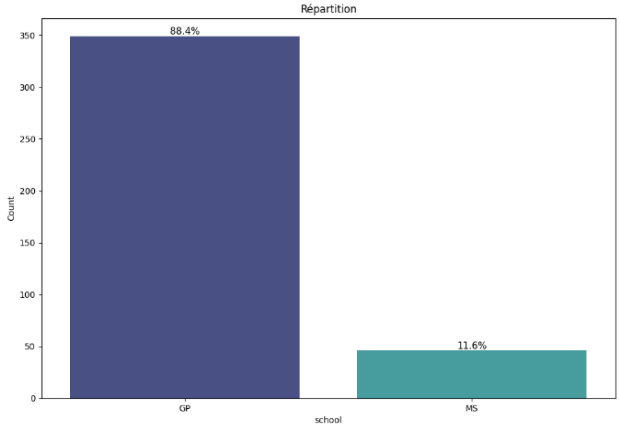


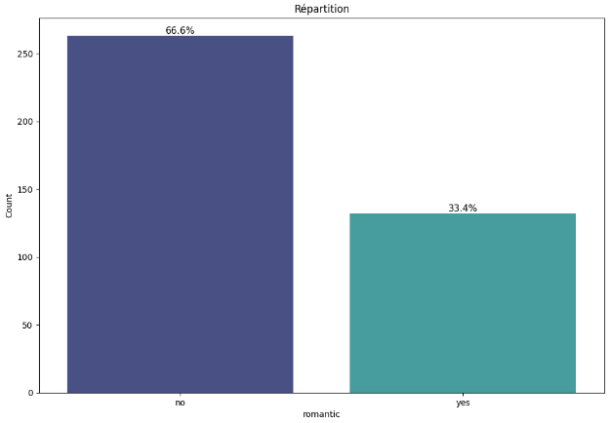


Le travail des parents des élèves est réparti en 5 groupes : Santé, au foyer, éducation, services publics ou autre. Le travail des parents est majoritairement caractérisé par other, ce qui ne nous apprend pas grand-chose sur leur travail. Les catégories choisies de travail ne permettent pas de décrire le travail de l'ensemble des parents des élèves.

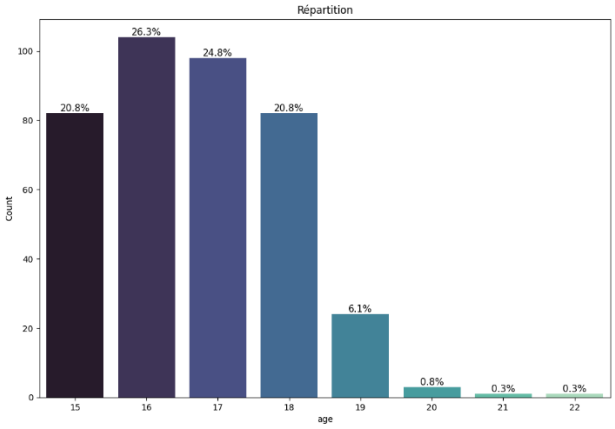
De plus les catégories choisies ne semblent pas pertinentes, les agriculteurs, cadres privés, artisans et les ouvriers ne sont pas représentés. Nous pouvons quand même remarquer que près d'un quart des pères et un quart des mères travaillent dans les services publics.

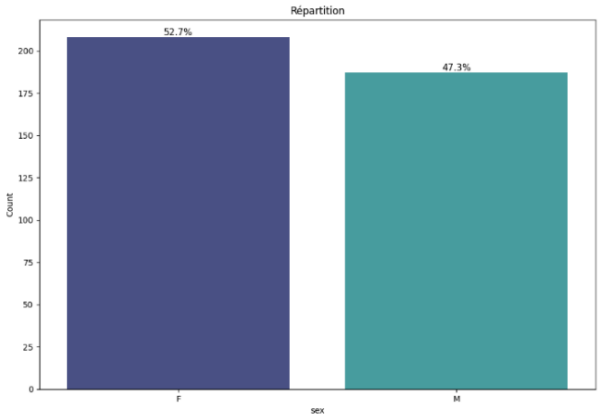
3- AUTRES REPARTITIONS





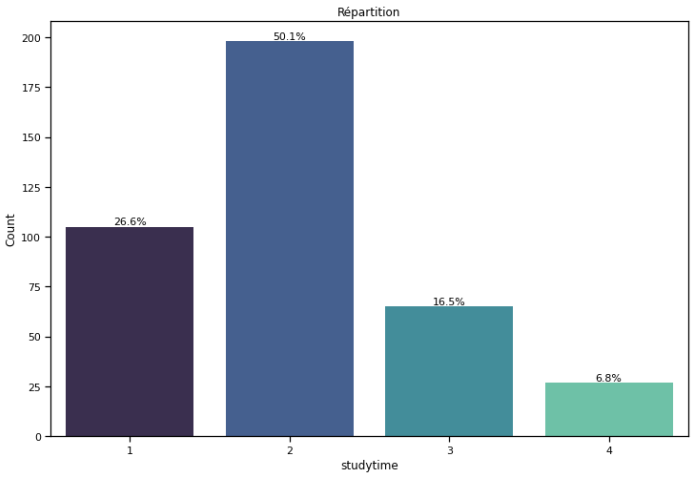
La plupart des élèves viennent du lycée GP (graphe de gauche) et sont célibataires (graphe de droite).

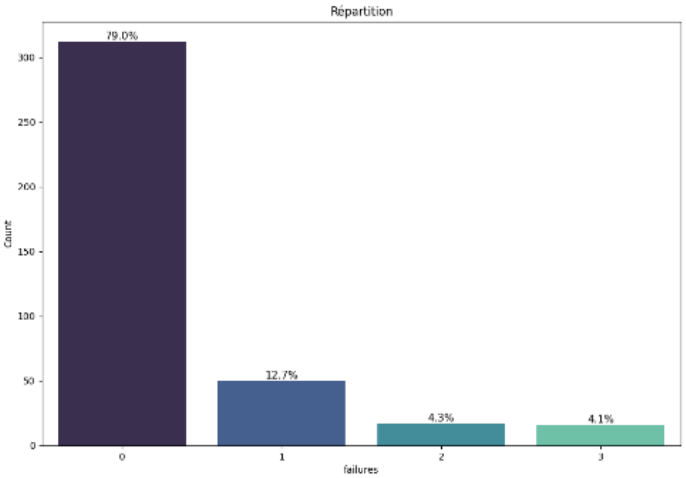




Les élèves ont entre 15 et 22 ans, et majoritairement entre 15 et 18 ans, (ce qui est logique car nous étudions des lycéens).

La répartition entre femmes et hommes est à peu près égalitaire, avec une légère majorité de femme. (Graphique de gauche)

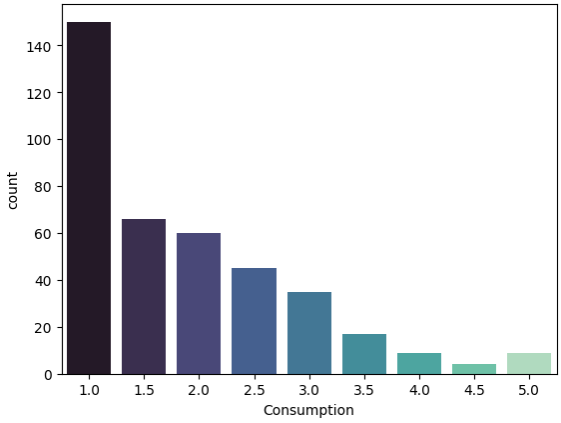


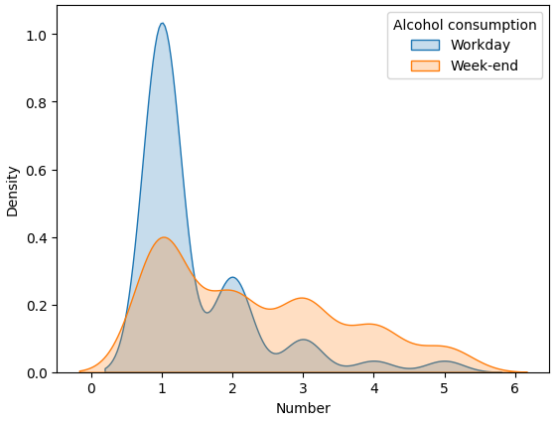


La plupart des élèves n'ont jamais redoublé. Le maximum de redoublement est de 4. (Graphique de droite)

75% des élèves sont dans les 2 premières catégories de temps de travail, avec 50% dans la 2e (Graphique de gauche). Cela signifie qu'une grosse majorité travaille moins de 5h par semaine.

4- DENSITES



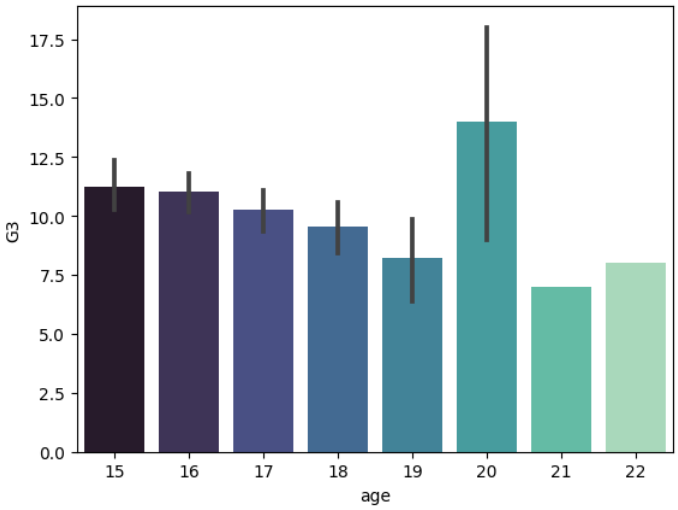


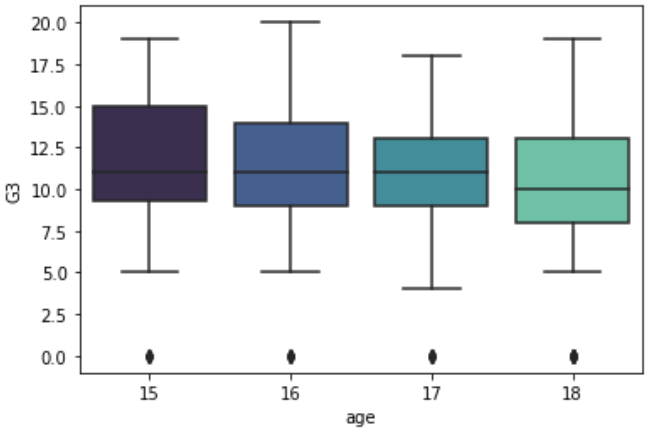
Concernant la consommation d'alcool, la plupart des élèves ont une faible consommation d'alcool. Peu d'élèves consomment beaucoup d'alcool en semaine, tandis que la répartition de la consommation le week-end est plus homogène. Nous observons cependant que plus nous montons de catégorie, moins il y a d'élèves.

Cependant, la consommation d'alcool étant noté sur une échelle de 1 à 5, nous ne savons pas à quoi correspond une faible consommation, ou une forte consommation.

ANALYSE BIVARIEE

1- ENTRE LA VARIABLE NOTE ET LES AUTRE VARIABLES

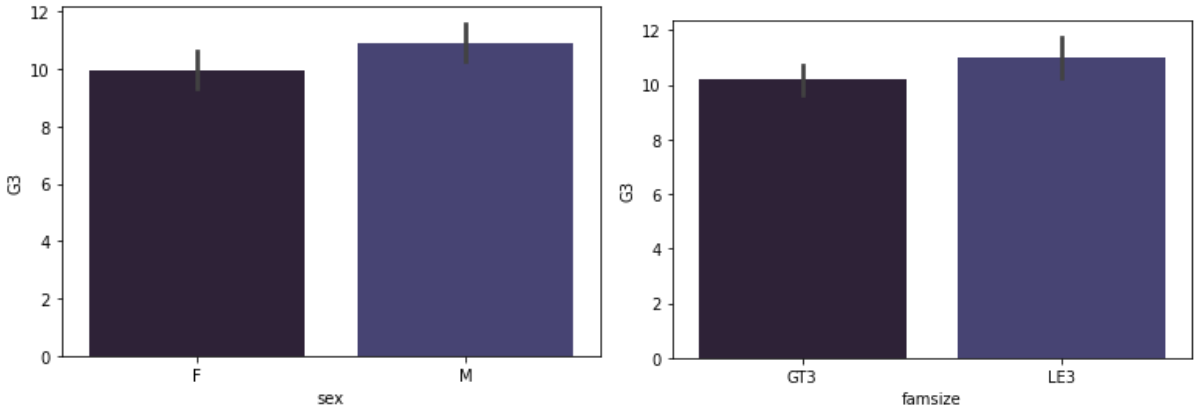




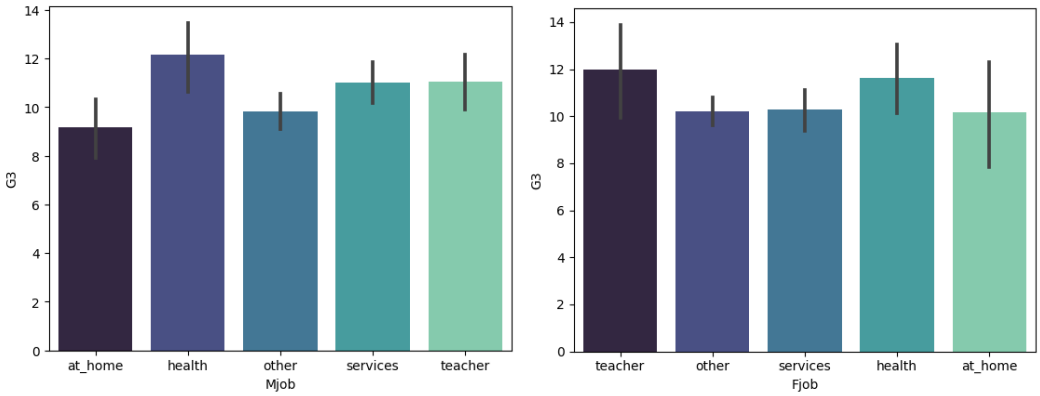
Nous pouvons observer sur le premier graphique que les étudiants de 20 ans sont ceux avec les meilleures notes en moyenne.

Entre 15 et 18 ans (l'âge majoritaire des élèves), les notes baissent en moyenne avec l'âge.

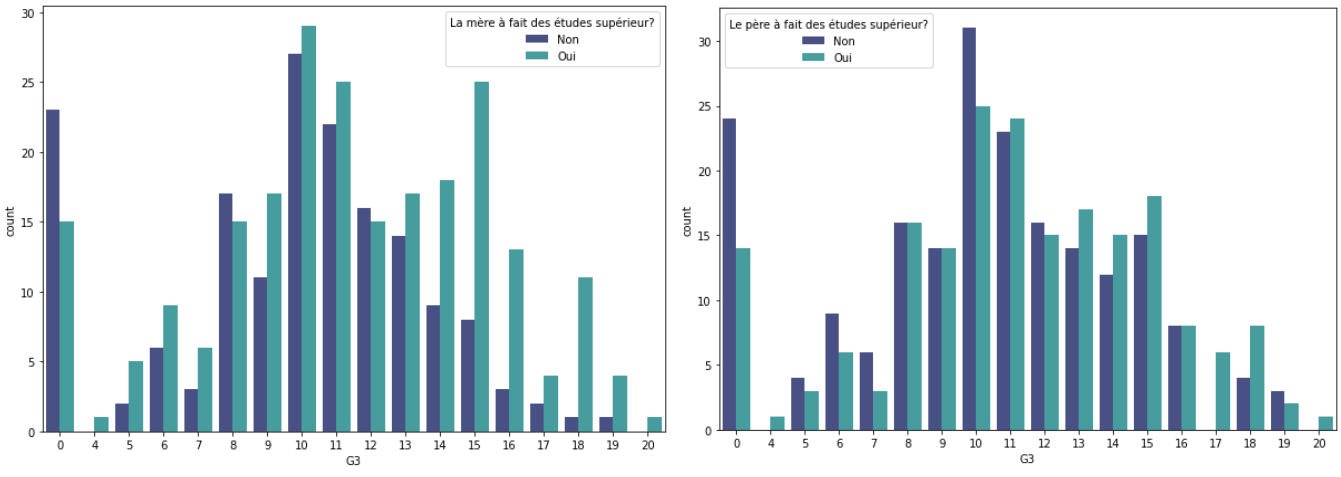
En observant les boxplots des notes en fonction de l'âge (pour les élèves de 15 à 18), nous observons que la médiane est similaire pour tous les âges. Cependant, les 25% premiers élèves des plus jeunes âges ont de meilleures notes que ceux plus vieux. Au contraire, les 25% derniers ont des notes similaires pour tous les âges.



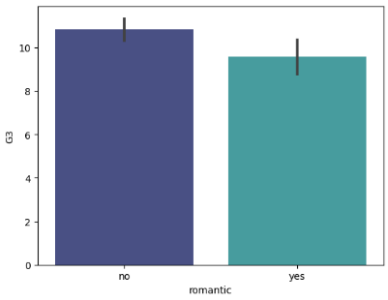
Les hommes ont en moyenne de meilleures notes que les femmes (graphique de gauche). Les élèves avec une famille de moins de 3 membres ont de meilleures notes, mais la différence n'est pas considérable. (Graphique de droite).

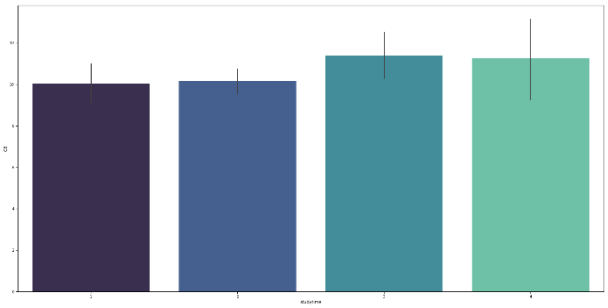


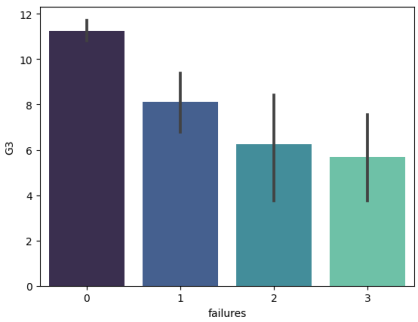
Les élèves dont la mère travaille dans la santé semblent réussir légèrement mieux, mais hormis cela le travail des parents ne semble pas avoir un gros impact sur les notes.



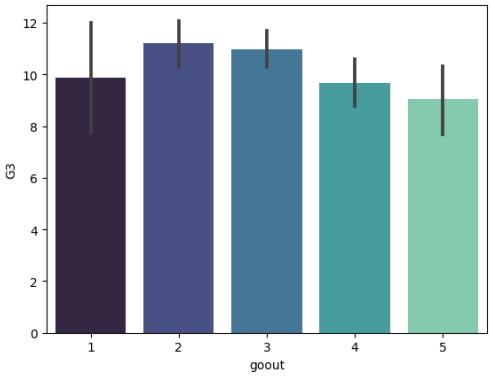
Les élèves dont la mère a fait des études supérieures ont tendance à avoir de meilleures notes que les autres.
Ce constat est moins clair pour les études du père, même s'il semble être vrai également.



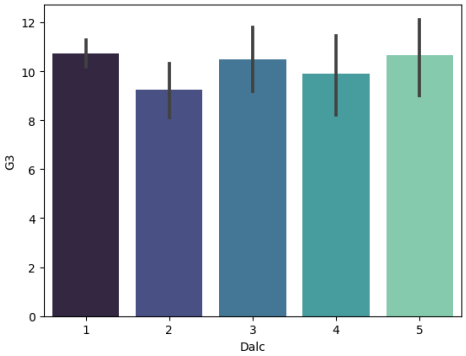


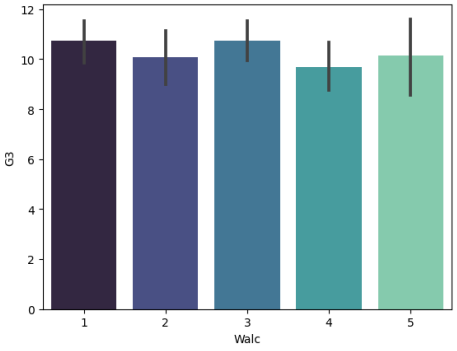


Les élèves en couple ont de moins bonnes notes que ceux célibataires (graphique de gauche). Les élèves travaillant plus de 5h par semaine (3e et 4e groupe) ont de meilleures notes que ceux travaillant moins de 5h. De plus, plus les élèves ont redoublé, moins ils ont de bonnes notes.

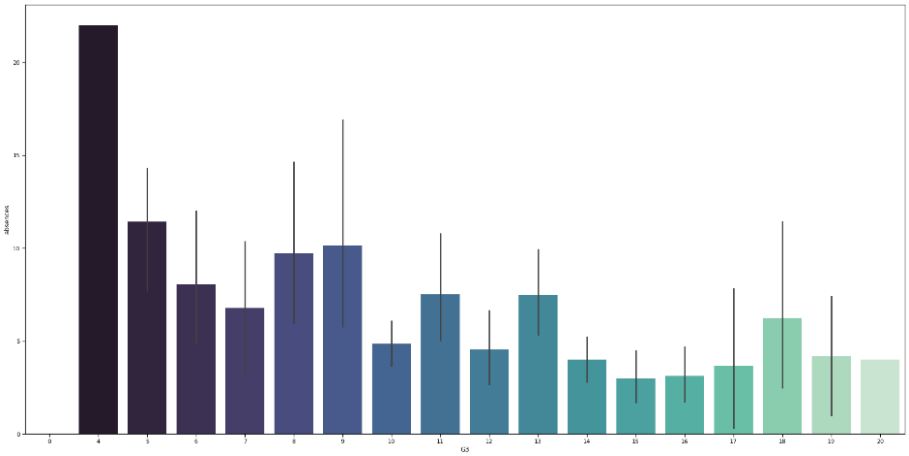


Les élèves sortant le moins n'ont pas de meilleures notes que les autres, mais à partir de la catégorie 2, plus les élèves sortent, moins ils ont de bonnes notes



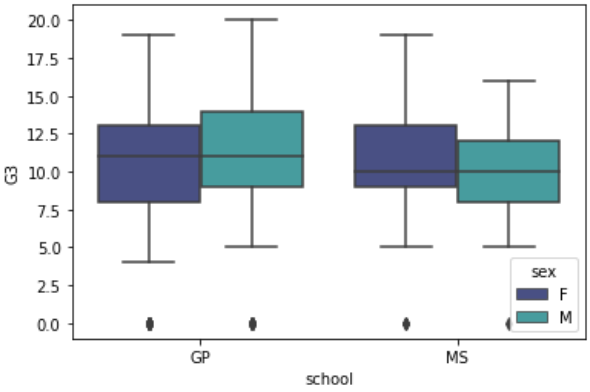


Les moyennes de notes des groupes de consommation d'alcool sont constantes. Sortir le week-end ou la semaine ne semble pas avoir d'effet sur les notes.

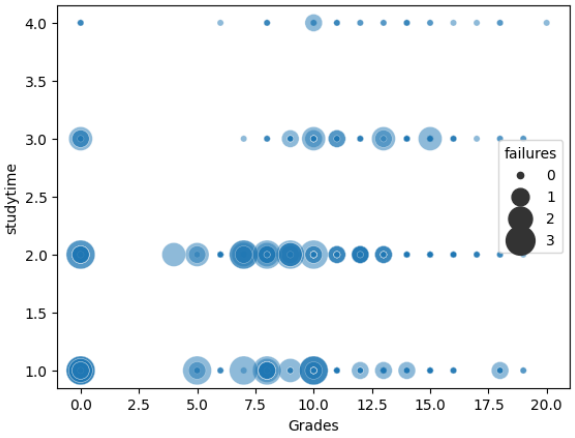


Nous observons que les élèves avec les pires notes ont le plus d'absences. Cependant, ceux qui ont plus le 13 de moyenne, leurs nombres d'absences ne varie pas énormément, mais reste bas.

2- CROISEMENT DE PLUSIEURS VARIABLES

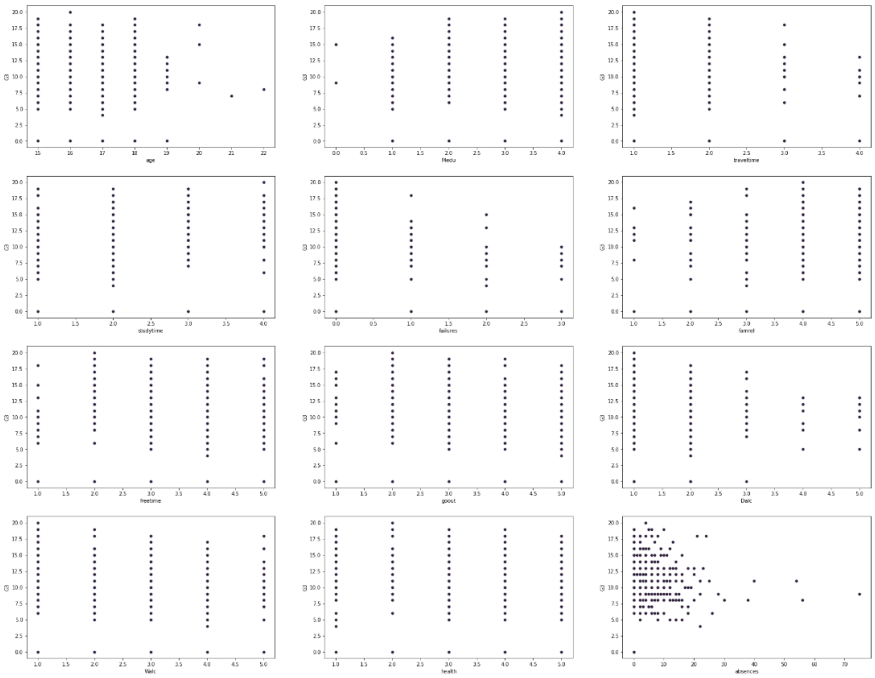


A l'école GP, les hommes réussissent mieux que les femmes, tandis que à MS, ce sont les femmes qui réussissent mieux.



Ce graphique nous permet de voir la répartition des redoublements en fonction des notes et du temps de travail. Ainsi, nous observons que les élèves travaillant le plus, redoublent très peu, et ont en général de bonnes notes, tandis que ceux qui travaillent le moins ont de moins bonnes notes, et ont plus redoublés.

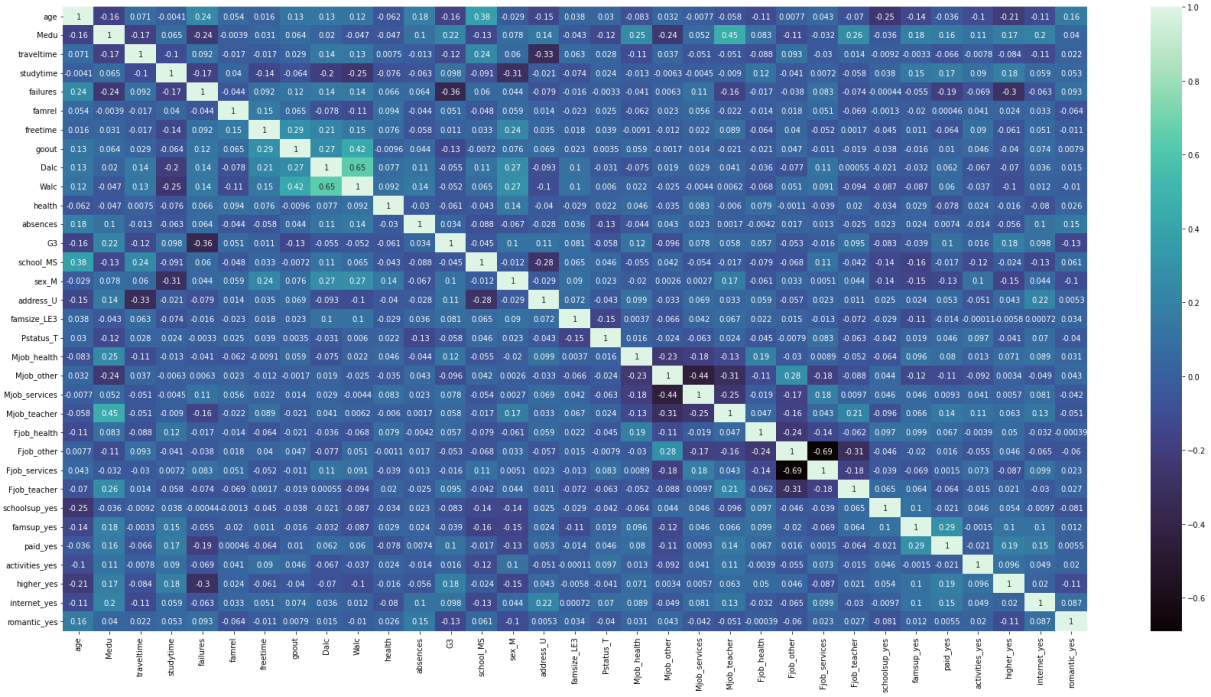
3- ETUDE DE LIEN ENTRE LES VARIABLES QUANTITATIVES ET LA NOTE



En représentant les notes en fonction des différentes variables quantitatives, nous n'observons aucun lien apparent entre la variable target et les différentes variables. De plus, nous voyons que nos variables ne sont pas continues, ce qui créé des lignes verticales.

4- CORRELATION DES VARIABLES :

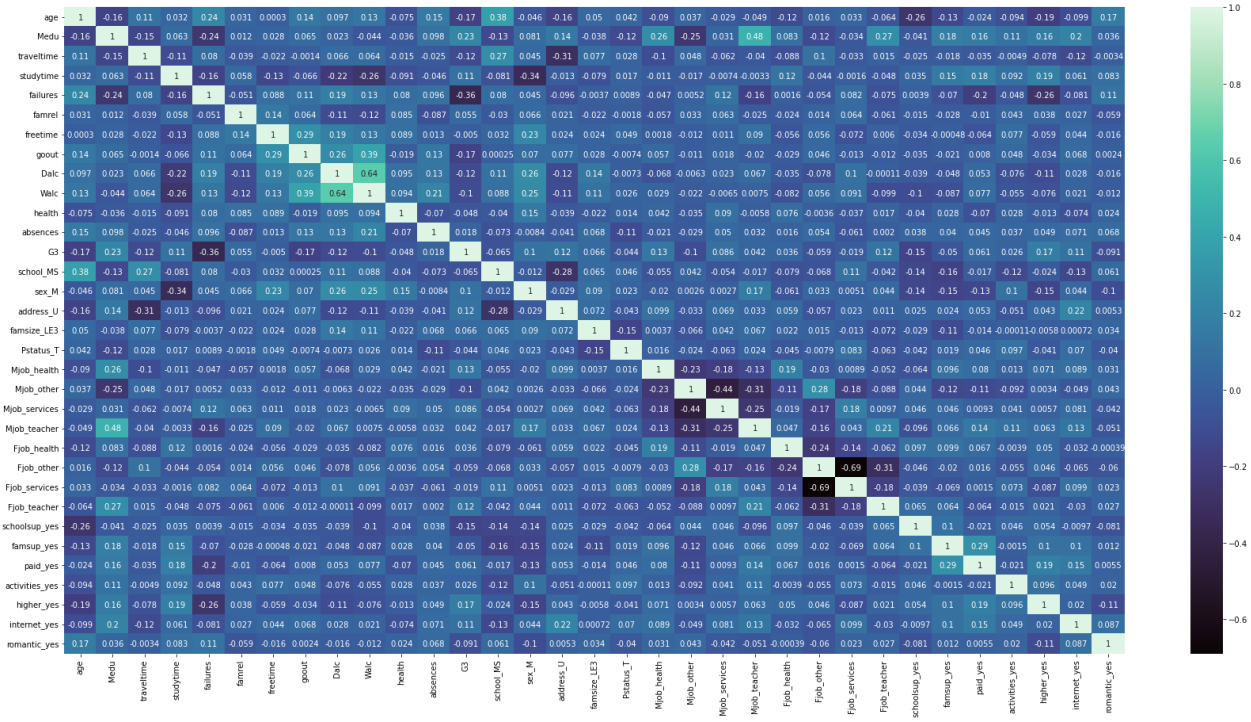
Nous avons d'abord étudié les coefficients de corrélation de Pearson :



Nous observons que notre variables target « G3 » est fortement corrélé avec aucune autre variable. Ainsi, il n'y a pas de relation linéaire simple entre notre variable target et les différentes variables.

Cependant, il y a une forte corrélation entre les variables Dalc, Walc et goout, qui décrivent la consommation d'alcool en semaine, en week-end, et le nombre de sorties par semaine avec des amis. Nous pouvons interpréter que le fait de sortir avec des amis est lié avec une forte consommation d'alcool et que les gens qui boivent en semaine boivent également le week-end.

Nous avons ensuite étudié les corrélations de Spearman :



Nous observons ici également que notre variable target est peut corrélée avec les autres. La variable la plus corrélée avec notre target est « failures » qui a un coefficient de corrélation de -0,36. Une augmentation de redoublement amènera à une baisse des notes. Ainsi, il ne semble pas exister de relation entre la variable target et les différentes variables

Nous allons cependant tenter de créer des modèles de régression linéaire, poisson et binomial, pour voir si nous pouvons tout de même expliquer une partie des notes.

5- TEST ANOVA

Nous avions expliqué précédemment que nous ne pensions pas que boire en semaine ou boire le week-end ait un impact sur les notes.
Nous allons effectuer un test ANOVA sur chacune de ces variables pour en être sûr, avec comme hypothèses :

H0 : La moyenne de la variable note par rapport à chaque catégorie de buveur est égale.
H1 : Au moins une des moyennes de la variable G3 par rapport à chaque catégorie de buveur est inégale.

	df	sum_sq	mean_sq	F	PR(>F)
C(Dalc)	4.0	132.173885	33.043471	1.583605	0.177864
Residual	390.0	8137.734976	20.865987	NaN	NaN

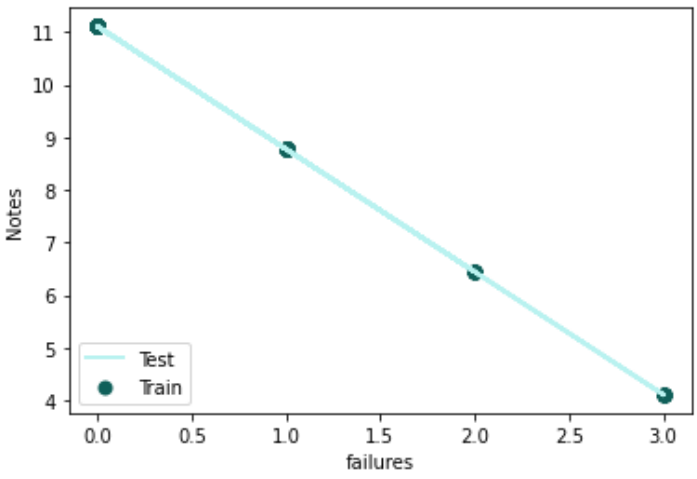
	df	sum_sq	mean_sq	F	PR(>F)
C(Walc)	4.0	61.722394	15.430599	0.733162	0.56975
Residual	390.0	8208.186467	21.046632	NaN	NaN

Nous observons que pour chacune variable, la p-value est supérieure à 0,05. Nous ne rejetons donc pas l'hypothèse nulle.
Ainsi, boire ne semble pas avoir un impact direct sur les notes.

REGRESSION LINEAIRE SIMPLE

Bien que nous ayons montré précédemment qu'il ne semblait pas y avoir de lien simple entre notre variable target et nos autres variables, nous avons tout de même implémenté des modèles de régressions linéaires simples avec les variables les plus corrélées avec la variable target. Les résultats étaient non concluants, il n'y a pas de lien linéaire simple entre notre target et une de nos variables.

Pour la variable failure, nous obtenons un modèle sous la forme : Y=11.108354994430004 - 2.33245451X

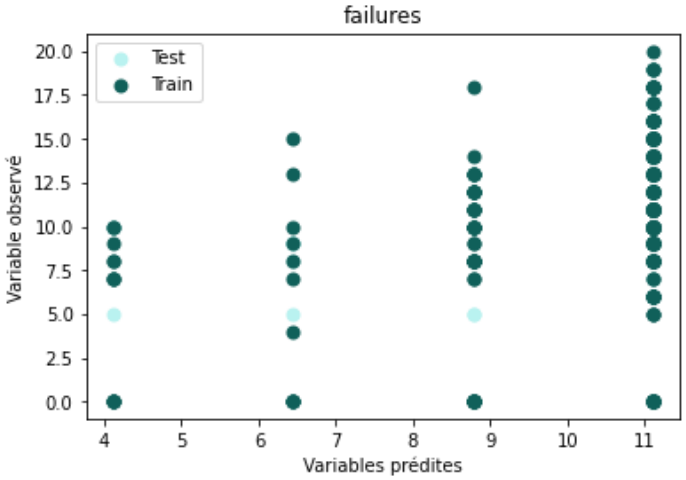


Ainsi, d'après notre modèle, plus le nombre d'absence serait grand, plus la note diminuerait. Cela semble logique.

Cependant, notre modèle n'est pas bon.

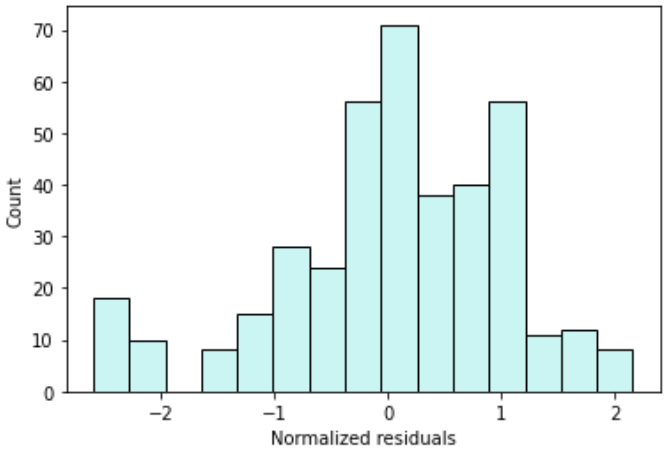
En effet, le R2 de la population d'apprentissage est de 0.14, et celui de la population test est encore plus bas 0.08.

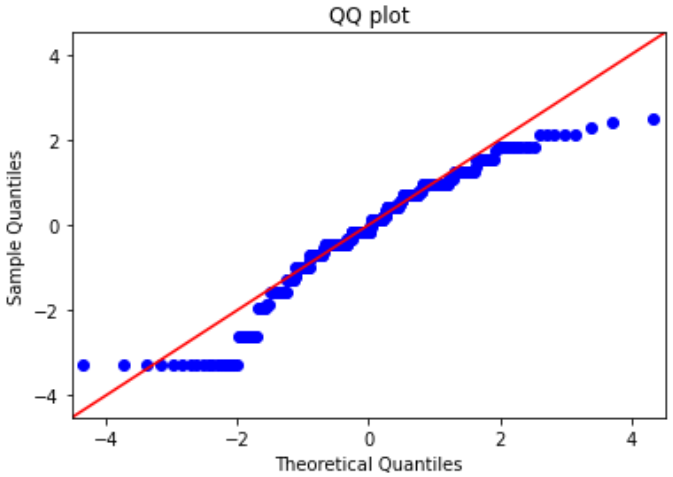
L'erreur quadratique de la population d'apprentissage est de 4.25 et celui de la population test est encore plus haut, 4.34



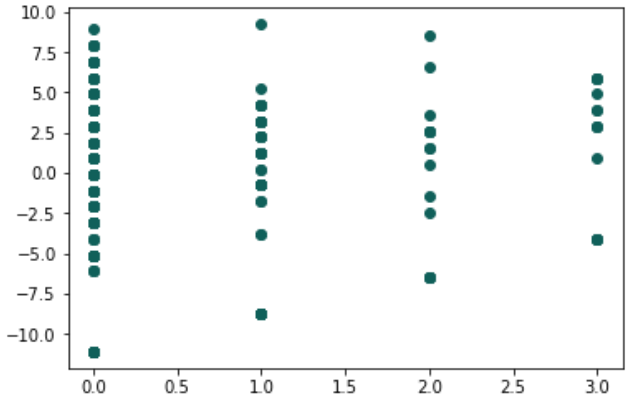
En représentant sur un graphique les notes observées en fonction des prédites de la population test, nous observons que les prédictions sont très mauvaises, car aucune droite x=y ne se dessine.

En continuant sur l'étude des résidus, le constat est le même : notre modèle n'est pas bon. En effet nos résidus ne semblent pas suivre une loi normale centrée-réduite, comme nous le montre l'histogramme ainsi que le qq-plot





La variance des résidus est cependant plutôt constante.



REGRESSION LINEAIRE MULTIPLE

Ayant un nombre important de variables, la première étape pour créer un modèle de régression linéaire multiple était de sélectionner un nombre limité de variable, les plus influentes, afin d'avoir un modèle d'explication des notes à la fois performant (résidus les plus petits possibles) et économique (le moins possible de variables explicatives).

Pour comparer nos variables, nous prenons comme critère le p-value et le R2 ajusté.
Nous avons ainsi effectué un Backward Elimination, et un Forward Elimination pour chacun de ces critères.
Nous avions auparavant modifié nos variables catégorielles en variables « dummies »

Nous avons sélectionné 3 combinaisons de variables :

1. ['Medu', 'failures', 'goout', 'sex_M', 'romantic_yes'], sélectionnées pour le critère de p-value en backward et en forward elimination
2. ['age', 'Medu', 'studytime', 'failures', 'freetime', 'health', 'absences', 'sex_M', 'address_U', 'famsize_LE3', 'schoolsup_yes', 'famsup_yes', 'higher_yes', 'romantic_yes'], sélectionnées pour le critère de R2 ajusté en backward elimination .
3. ['paid_yes', 'Pstatus_T', 'Dalc', 'Walc', 'internet_yes', 'school_MS', 'famrel', 'traveltime', 'activities_yes']], sélectionnées pour le critère de R2 ajusté en forward elimination .

Nous avons ainsi 3 modèles de régressions linéaires pour chaque sélection de variables. Nous avons divisé notre jeu de données en 2 parties (80%, pour l'apprentissage et les 20% restant pour le test), afin d'évaluer la performance du modèle sur des données non vues auparavant par le modèle.

Nous obtenons encore une fois, des modèles de régressions linéaire non concluants.

Pour le modèle avec les variables : ['failures', 'Medu', 'sex_M', 'goout', 'romantic_yes']. Nous obtenons ceci :

Generalized Linear Model Regression Results

Dep. Variable:	G3	No. Observations:	395
Model:	GLM	Df Residuals:	389
Model Family:	Gaussian	Df Model:	5
Link Function:	identity	Scale:	17.417
Method:	IRLS	Log-Likelihood:	-1121.8
Date:	Thu, 03 Nov 2022	Deviance:	6775.2
Time:	17:10:33	Pearson chi2:	6.78e+03
No. Iterations:	3		
Covariance Type:	nonrobust		

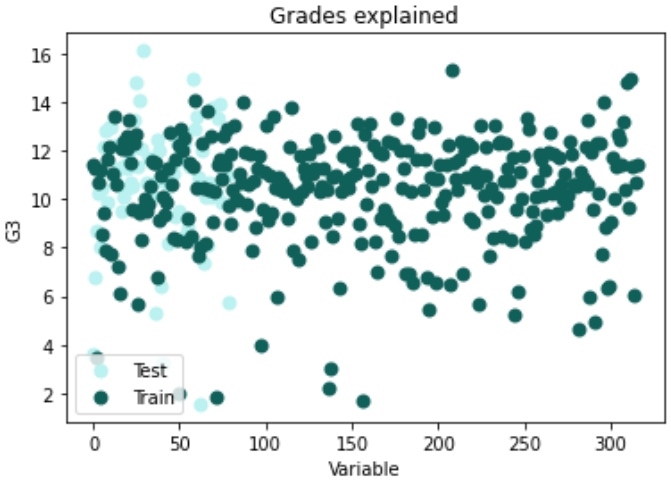
	coef	std err	z	P>|z|	[0.025	0.975]
const	10.6342	0.824	12.900	0.000	9.018	12.250
Medu	0.6157	0.200	3.080	0.002	0.224	1.008
failures	-1.8943	0.296	-6.391	0.000	-2.475	-1.313
goout	-0.4571	0.192	-2.386	0.017	-0.833	-0.082
sex_M	0.9557	0.426	2.242	0.025	0.120	1.791
romantic_yes	-0.9279	0.451	-2.058	0.040	-1.812	-0.044

Les variables Medu et sex_M ont un coefficient positif, contrairement à failure, goout et romantic_yes. Ainsi, d'après le modèle, être de sexe masculin augmente de 0.96 point la note, et plus la mère à une bonne éducation, plus la note est meilleure. Au contraire, être en couple fait baisser de 0.93 la note, sortir et avoir redoublé font également baisser la note.

Cela est cohérent avec nos observations précédentes, car nous avions vu par exemple que les hommes réussissaient mieux que les femmes, que les personnes avec une mère ayant la meilleure éducation possible réussissaient mieux que les autres et que ceux qui avaient le plus redoublé avaient un moins bon taux de réussite que les autres.

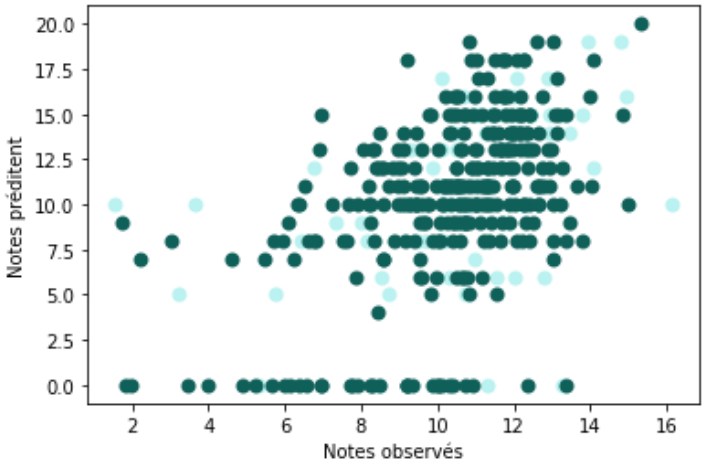
La p-value de toutes les variables est bien inférieure à 0,5, mais en observant le modèle, nous remarquons que nous n'avons pas un modèle linéaire du tout.

Voici la représentation du modèle que nous obtenons :



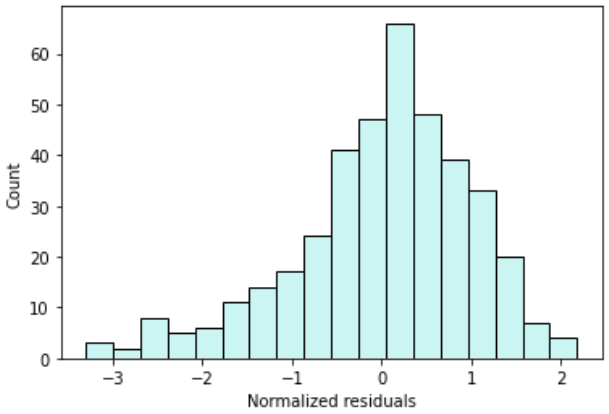
Sur notre modèle d'apprentissage, nous obtenons une erreur quadratique moyenne de 4.24 et un score R2 de 0.2, et sur notre modèle test, nous obtenons une erreur quadratique moyenne de 3.79 et un score R2 est 0.02. Ainsi, notre modèle explique assez mal nos notes.

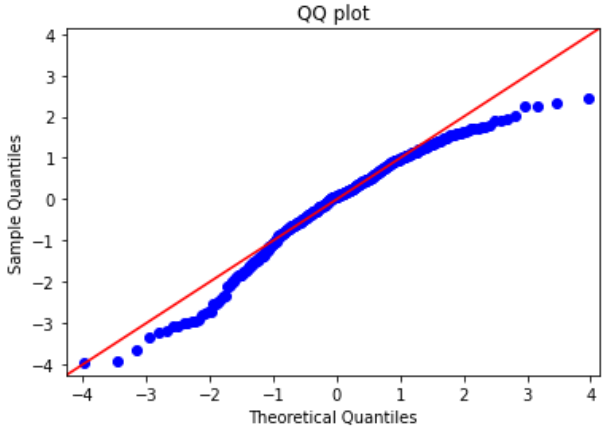
En étudiant le graphe des notes prédites en fonction des observées, nous observons que notre modèle ne prédit pas bien les notes, car aucune droite x=y ne se dessine.

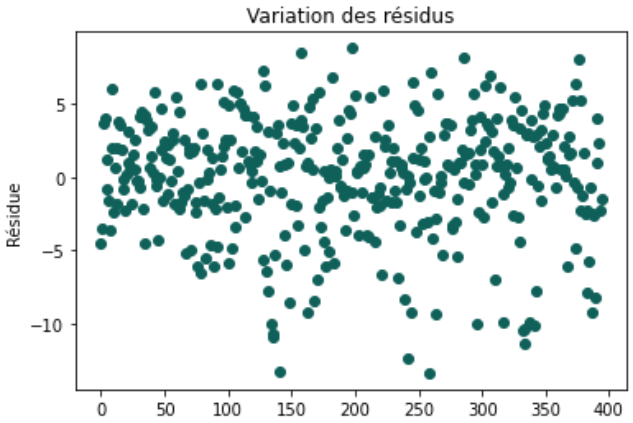


Cependant, nous observons que quand la note est réellement mauvaise, notre modèle ne prédit pas de bonnes notes. Ainsi nous ne pouvons pas prédire les bonnes notes, mais nous pouvons prédire les mauvaises.

En étudiant l'histogramme et le qq-plot, nous observons que nos résidus ne semblent pas suivre une loi normale centrée-réduite.







La variance des résidus est cependant constante.

MODELE GENERALISE BINOMIAL

Pour effectuer une régression binomiale, nous avons tout d'abord mis notre variable target sous forme d'une variable binomiale. Elle prend la valeur 0 quand la note est inférieure à 10, et 1 sinon. Nous testions ensuite avec un modèle d'apprentissage une régression binomiale pour chacun des trois sets de variables sélectionnées précédemment.

Les résultats sont ici plus concluants. Nous allons étudier le même set de variables que précédemment :

Generalized Linear Model Regression Results

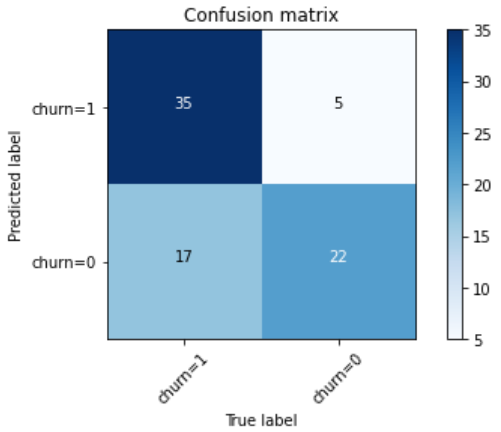
Dep. Variable:	Grade_binomial	No. Observations:	395
Model:	GLM	Df Residuals:	389
Model Family:	Binomial	Df Model:	5
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-220.46
Date:	Sun, 30 Oct 2022	Deviance:	440.92
Time:	16:39:10	Pearson chi2:	396.
No. Iterations:	4		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.7577	0.462	3.805	0.000	0.852	2.663
Medu	0.1200	0.111	1.082	0.279	-0.097	0.337
failures	-0.9242	0.177	-5.217	0.000	-1.271	-0.577
goout	-0.3556	0.108	-3.288	0.001	-0.568	-0.144
sex_M	0.4377	0.239	1.833	0.067	-0.030	0.906
romantic_yes	-0.3144	0.245	-1.283	0.199	-0.795	0.166

Comme pour la régression linéaire multiple étudiée précédemment, les variables Medu et sex_M ont un coefficient positif, contrairement à failure, goout et romantic_yes.

Mais elles n'ont pas le même impact. En effet, être du sexe masculin multipliera les chances d'avoir une note supérieure à 10 par $e^{0,43}$, tandis qu'être en couple les fera diviser par $e^{0,31}$
Encore une fois, la variable ayant le plus gros impact est la variable failures, où chaque échec supplémentaire réduit la probabilité d'avoir une note supérieure à la moyenne de $e^{0,92}$

En étudiant la matrice de confusion de la régression binomiale, nous observons que notre régression n'est pas mauvaise :



En effet, nous observons que les bonnes notes sont bien prédites à 81%, tandis que les mauvaises le sont à 67% seulement.

Ainsi nous en déduisons que notre modèle réussit à prédire les mauvaises notes, mais pas les bonnes.

MODELE GENERALISE POISSON

Pour la régression de poisson, nous avons dû choisir une autre variable, qui comptait quelque chose. Nous avons décidé de choisir la variable absences.

Pour choisir les variables explicatives, nous avons sélectionné les variables ayant la plus grosse corrélation de Spearman avec notre variable absence.

Nous avons ensuite testé plusieurs modèles, en supprimant certaines variables.

Pour comparer ces modèles, nous avons pris la déviance de chaque modèle, que nous avons divisé par le nombre de variables.

La meilleure combinaison de variables que nous ayons pu trouver est :

["age","Medu","Fedu","goout","internet_yes","failures","romantic_yes","higher_yes"]

Generalized Linear Model Regression Results

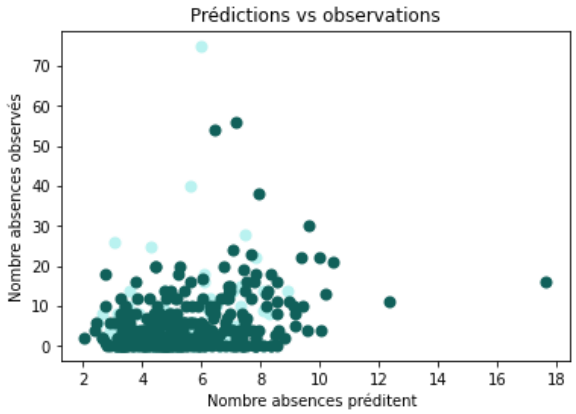
Dep. Variable:	absences	No. Observations:	395
Model:	GLM	Df Residuals:	386
Model Family:	Poisson	Df Model:	8
Link Function:	log	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1962.9
Date:	Tue, 29 Nov 2022	Deviance:	2916.4
Time:	20:24:34	Pearson chi2:	3.64e+03
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.0048	0.329	-6.085	0.000	-2.651	-1.359
age	0.1738	0.017	10.152	0.000	0.140	0.207
Medu	0.1693	0.026	6.523	0.000	0.118	0.220
Fedu	-0.0410	0.025	-1.623	0.105	-0.090	0.009
goout	0.0058	0.019	0.302	0.763	-0.032	0.044
internet_yes	0.3921	0.068	5.732	0.000	0.258	0.526
failures	0.0551	0.030	1.830	0.067	-0.004	0.114
romantic_yes	0.2890	0.044	6.554	0.000	0.203	0.375
higher_yes	-0.0452	0.093	-0.486	0.627	-0.227	0.137

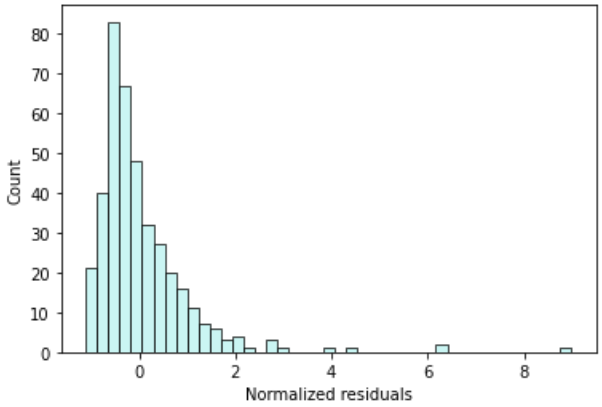
Nous pouvons déduire de ce tableau que le nombre d'absences augmente lorsque l'âge augmente, le niveau d'éducation de la mère augmente, l'étudiant est en couple et qu'il a accès à internet.
Au contraire, ce nombre d'absences diminue plus le père a une éducation haute, et si l'étudiant souhaite poursuivre ses études.

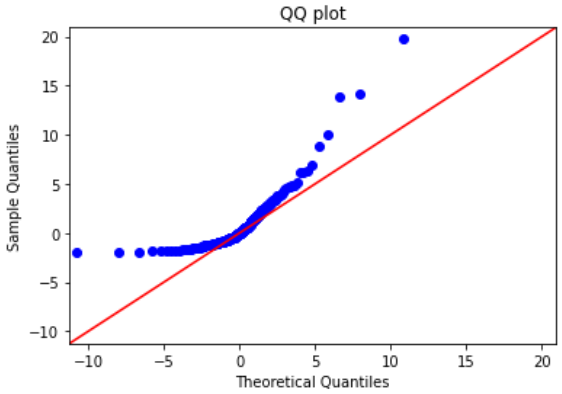
Cela peut sembler étrange car on aurait tendance à croire que plus l'étudiant sort dans la semaine, plus il est absent, alors que le fait de sortir une fois de plus par semaine divise le nombre d'absence par (e^-0.0058).

En étudiant le graphe du nombre d'absences prédites en fonction du nombre d'absences observées, nous observons que notre modèle ne prédit pas bien les notes, car aucune droite x=y ne se dessine.

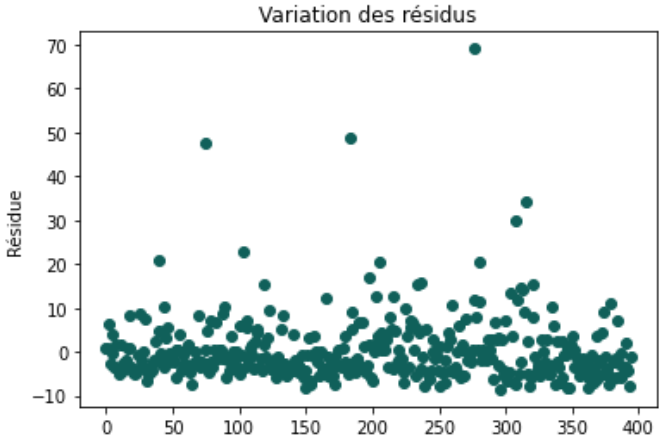


Nos résidus ne semblent pas suivre une loi normale centrée-réduite, comme nous le montre l'histogramme ainsi que le qq-plot





La variance des résidus est cependant constante, et les résidus semblent indépendants



ACP

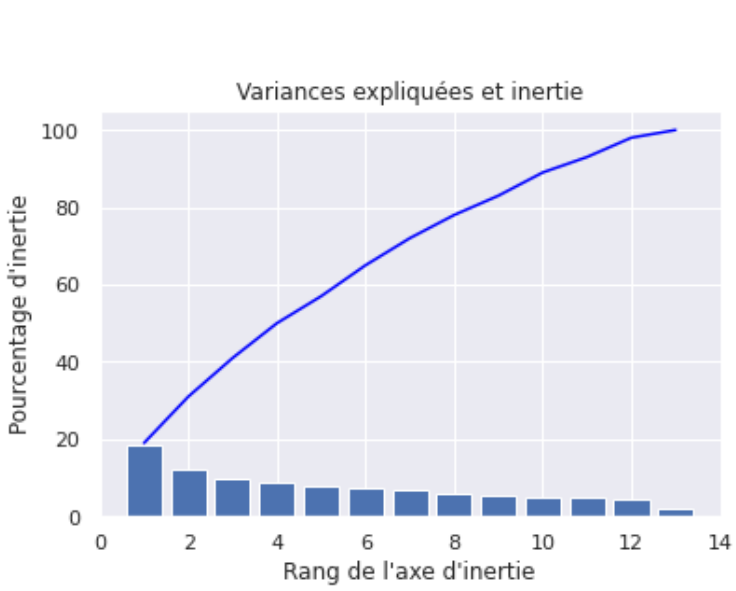
L'Analyse en Composantes Principales nous permet d'étudier des tableaux de données avec des individus en lignes (les étudiants) et des variables quantitatives en colonne.

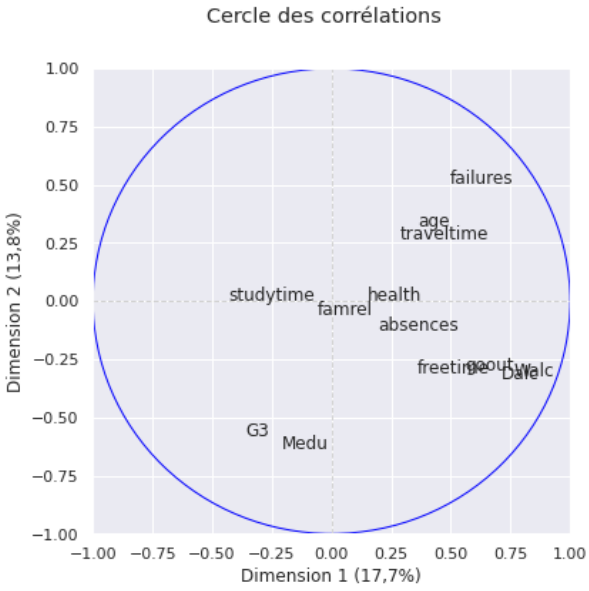
Notre objectif à ce stade est d'avoir une vision globale de notre base de données, de comprendre les éventuels liens qui existent entre les variables.

Résultats sur les variables :

Pour déterminer les variables influençant le plus nos axes d'ACP, nous avons réalisé un graphe du rang d'inertie en fonction du pourcentage d'inertie. Ce graphe nous indique que les 2 premiers axes contribueront à environ 17% et 13% de l'inertie totale.

Graphique sur les variables quantitatives et cercle de corrélation des variables quantitatives





Le premier facteur est corrélé positivement et assez fortement avec certaines variables, on note que « freetime », « goout », « Dalc » et « Walc » sont situés au même endroit. Ces variables sont corrélées entre elles et certaines d'entre elles sont bien représentées : notamment « goout » (sorties), « Dalc » et « Walc » (consommation d'alcool). Les évènements « sortir entre amis » et « consommer de l'alcool » évoluent ensemble à une vitesse constante. La variable « failures » est aussi bien représentée sur l'axe 1. En revanche, l'angle quasi droit entre la variable target (« G3 ») et les variables relatives à l'alcool montre que ces deux variables sont indépendantes entre elle.

En ce qui concerne l'axe 2, on note que « Medu » est bien représentée. Notre variable target se situe à proximité de celle-ci mais n'est pas forcément très bien représentée.

De ces observations, nous pouvons en conclure que l'axe 1 correspond au style de vie de de l'individu tandis que l'axe 2 pourrait correspondre à son éducation.

Cette interprétation pourra être précisée avec les tableaux relatifs aux individus.

Résultats sur les individus :

Dans cette partie de l'analyse, nous cherchons à déterminer les individus les mieux représentés pour déterminer les axes. Nous cherchons les contributions et les qualités des points individus.

Les individus les mieux représentés sont les individus 247, 61, 150, 349, 149 etc… L'objectif ici est de les trouver sur le graphe des individus et de relever leurs caractéristiques.

Caractéristiques des individus et interprétation des axes :

- Zoom sur le point 247 (extrémité gauche axe Dim 1) : individu âgé de 22 ans, ne prend pas beaucoup de temps à venir en cours (moins d'une demi-heure), travaille moins de 2h par semaine, a redoublé 3 fois, entretient une bonne relation avec sa famille et il consacre beaucoup de temps libre après les cours (score = 4/5). Nous nous concentrons maintenant sur les variables bien représentées, goout, dalc et walc : il sort souvent avec ses amis (score = 5, maximum), consomme beaucoup d'alcool en jour de semaine et le week-end (les 2 scores = 5, maximum). Son score final est de 8.

- Zoom sur le point 47 (extrémité droite axe Dim 1) : étudiant âgé de 16 ans, travaille beaucoup (4/5), n'a jamais redoublé, a une bonne relation avec sa famille. Concernant les variables bien représentées : il a une note de 2/5 concernant les sorties entre amis et 1/5 concernant sa consommation d'alcool. Sa note finale est de 20.

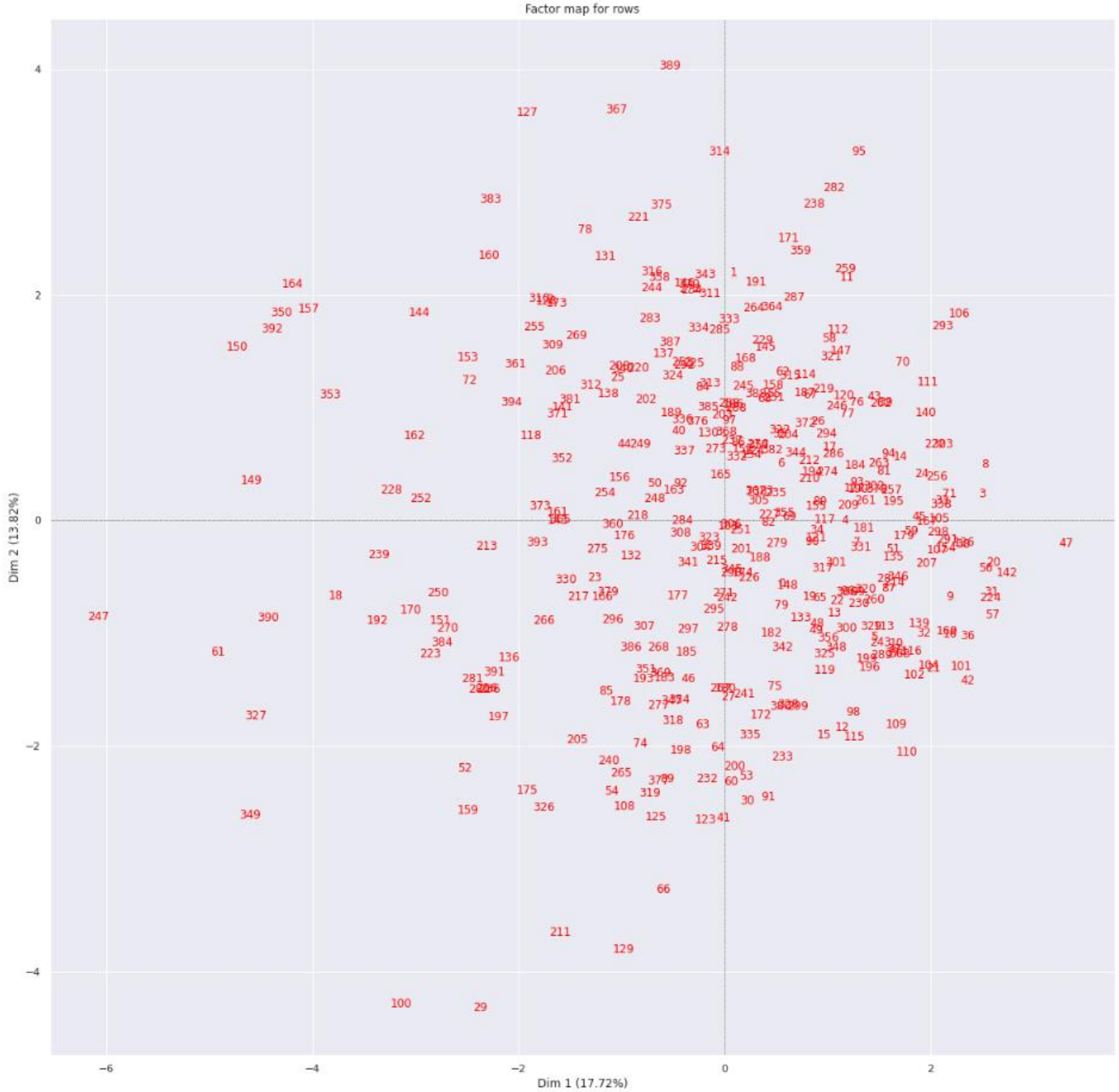
- Zoom sur le point 389 (extrémité haut axe Dim 2) : âgé de 18 ans, cet étudiant n'a pas une bonne relation avec ses parents. Il ne sort pas beaucoup et ne consomme pas beaucoup d'alcool. Ses parents n'ont pas fait beaucoup d'études. Il a une note finale de 0.

- Zoom sur le point 100 (extrémité bas axe Dim 2) : âgé de 16 ans, l'étudiant ne travaille pas beaucoup, sort souvent et consomme de l'alcool (score maximum). Son score à l'examen est de 5.

Résumé des individus dans un tableau

Élève	Age	Medu/Fedu	Travel time	Devoirs	Redoublement	Famille	Temps libre	Sorties	Alcool	Absences	Note finale
247	22	3-1	1	1	3	5	4	5	5	16	8
61	16	1-1	4	1	0	5	5	5	5	6	11
47	16	4-3	1	4	0	4	2	2	1	4	20
389	18	1-1	2	2	1	1	1	1	1	0	0
100	16	4-4	1	1	0	4	5	5	5	14	5

Nous notons la présence d'individus bien représentés vers la partie gauche de l'axe 1 : individus 247, 61. Cependant, les individus à droite ne sont pas bien représentés. Ce sont les individus ayant des bonnes notes : nous pouvons penser que cela est dû au fait qu'avoir des bonnes notes ne s'explique par forcément mais en avoir des mauvaises s'explique en fonction de notre qualité de vie. Cette analyse nous permet de conclure que le premier axe correspond au fait d'avoir obtenu une bonne note finale. A gauche, on observe les étudiants ayant obtenu une moins bonne note que les étudiants à droite. Nous pouvons penser que le deuxième l'axe quant à lui est celui qui représente l'éducation fournie par les parents.

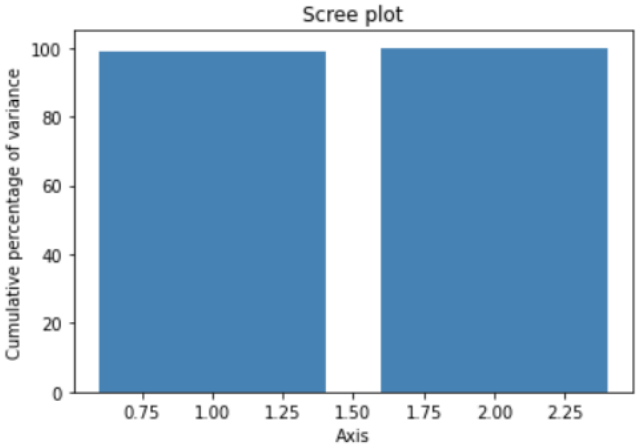


AFC

L'AFC est utile pour analyser le lien entre deux variables qualitatives. Nous avons choisi d'observer ce lien entre notre variable target des notes discrétisées, et la variable qualitative décrivant le temps de travail.

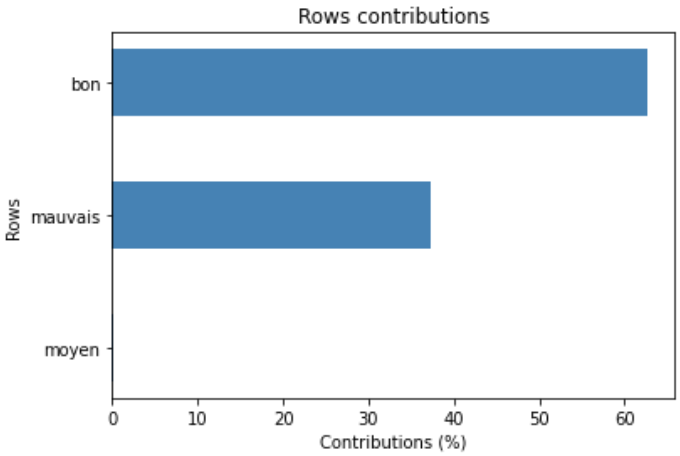
studytime	1	2	3	4
Note Qualitative				
bon	18	31	17	7
mauvais	37	70	16	7
moyen	50	97	32	13

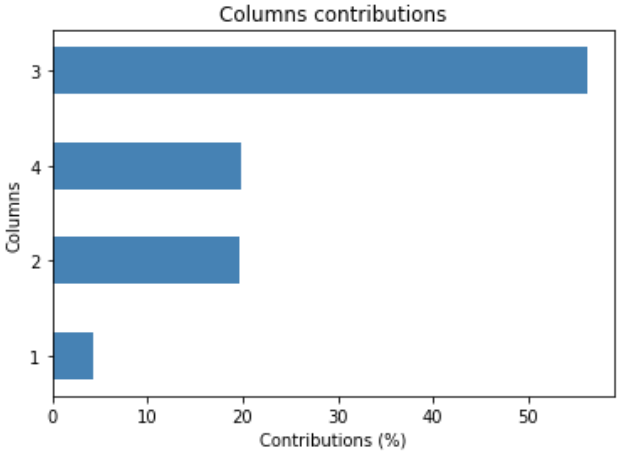
Nous avons ainsi obtenu le tableau de contingence ci-contre.



Dans note analyse, les 2 axes expliquent 100% de la variance totale, il est donc évident de les retenir pour dessiner par la suite la carte de nos facteurs.

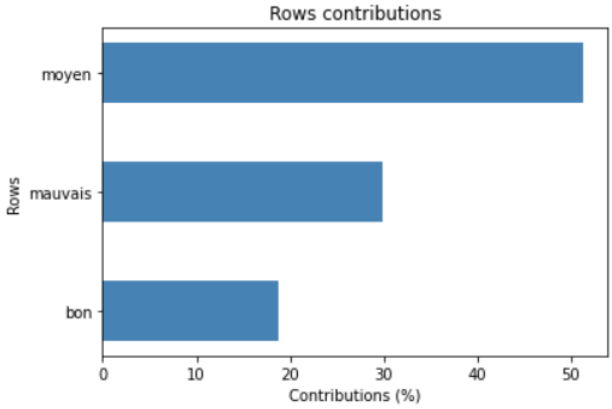
L'axe 1 explique d'ailleurs à lui tout seul 98,98% de la variance, il est donc judicieux de s'interroger sur sa composition, quels sont les paramètres de nos variables qui ont permis de le déterminer ?





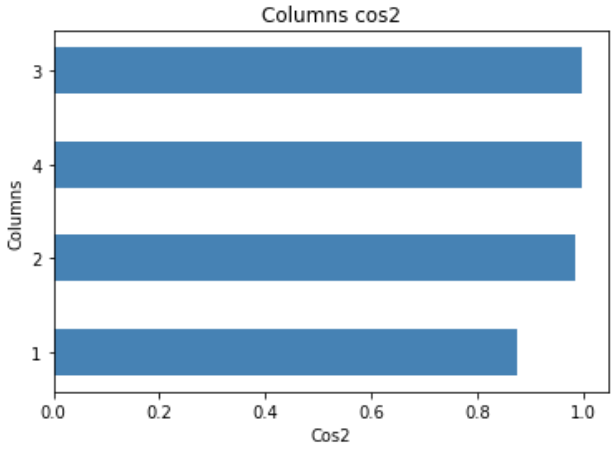
Une bonne ou mauvaise note permet ainsi de façonner efficacement notre axe1 : ce sont des facteurs déterminants, contrairement à une note moyenne, mais qui judicieusement se situe à l'origine de notre axe 1.

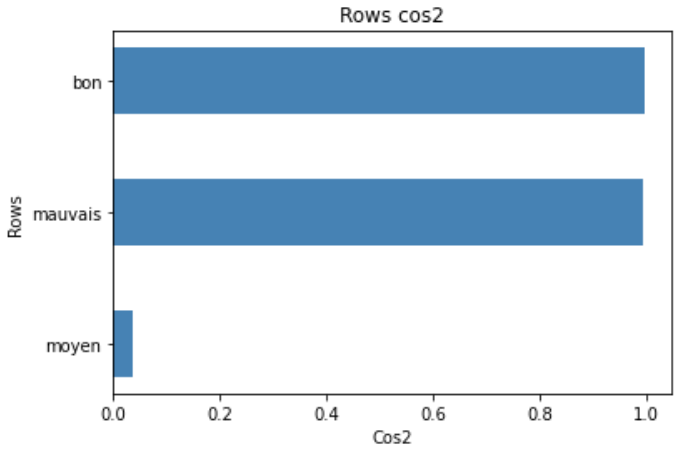
Également, nous pouvons tirer des conclusions similaires pour le temps de travail 3, 4 et 2 qui définissent bien l'axe 1.



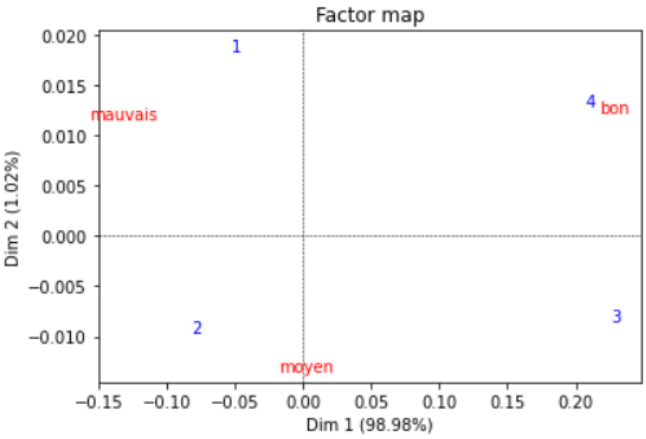
Il est de plus intéressant de relever la contribution d'une note moyenne à l'axe 2. Eteinte lors de l'axe1, il semblerait que l'axe2 ait donc tendance à séparer les valeurs extrêmes de la moyenne.

En observant nos différents cos² pour l'axe 1 nous retrouvons des conclusions semblables à précédemment, les colonnes 1 2 3 4 sont respectivement bien représentées sur le graphique ci-dessous de l'AFC, de même que les paramètres bon et mauvais.





Finalement, voici la Factor Map pour les variables Notes (qualitativement) et temps de travail.

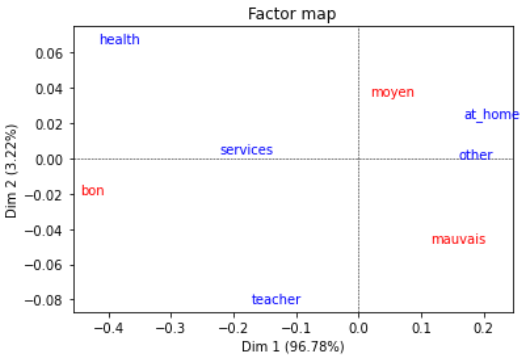


L'axe 1 sépare très clairement les bonnes des mauvaises notes, il retranscrit donc le niveau de l'individu.

L'axe 2 sépare les valeurs extrêmes, d'un côté les meilleurs et pires résultats avec les plus haut et plus bas temps de travail, et de l'autre les valeurs moyennes.

On remarque ainsi que les temps de travail les plus élevés sont très proches du résultat « bon », avec le 4 au plus près et le 3 du même côté de l'axe 1. On peut déduire un certain lien entre ces variables.

Si les temps de travail 1 et 2 sont du côté du niveau « mauvais », leur interprétation reste limitée étant donnée la disparité du modèle.



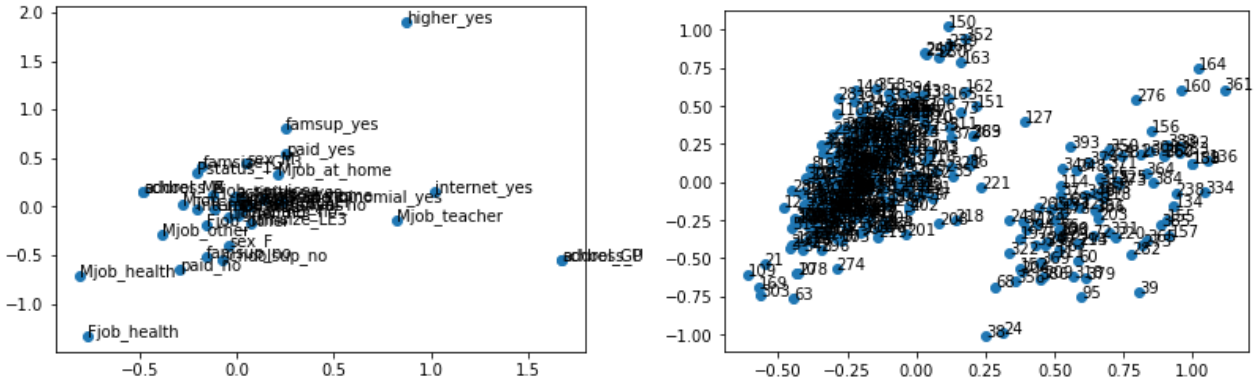
Nous avons réalisé d'autres AFC comme par exemple entre la note qualitative et le travail de la mère, qui ne nous ont pas permis de tirer d'avantages de conclusions, si ce n'est un certain raprochement entre les bonnes notes et un métier nécessitant des études relativement avancées: santé, fonctionnaire, professeur.

Ainsi une bonne note serait liée avec un haut temps de travail et dans une proportion moindre un certain type de travail de la mère.

ACM

L'ACM est l'analyse des correspondances multiples. Tandis que l'AFC analyse les liaisons entre 2 variables qualitatives, l'ACM permet d'analyser la liaison entre un nombre multiple de variables qualitatives. En sélectionnant l'ensemble de nos variables qualitatives et en les catégorisant, nous obtenons une inertie environ égale à 1,39 pour les 2 premiers axes. L'interprétation des résultats n'est pas forcément évidente. En effet, nous avons 16 variables qualitatives avec au total 36 modalités.

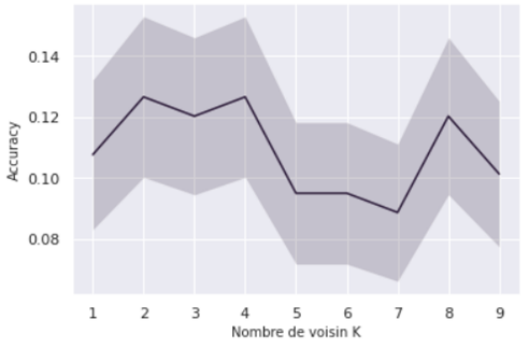
Graphique de l'ensemble des modalités et des individus en fonction des deux premières dimensions



KNN

Le KNN K-Nearest-Neighbors cherche à classer une nouvelle observation dans un dataset en le comparant avec ses K voisins.

Nous avons choisi ici de tenter d'expliquer la moyenne d'un individu (G3) en fonction de son âge, son temps de trajet, sa santé, ses absences, son temps de sortie, et son temps de travail. Après avoir normalisé les données, et séparé notre base de données en un jeu de test et un jeu d'entrainement, nous obtenons des prédictions exactes en fonction de K, détaillées dans le graphe suivant :



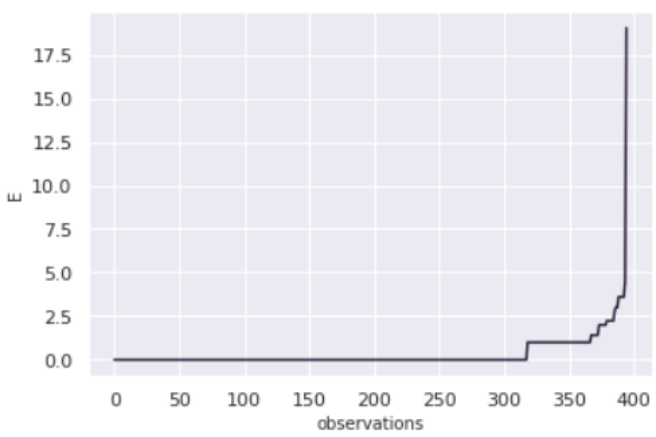
La meilleure précision est obtenue pour K = 2, avec plus ou moins une marge d'erreur. Toutefois elle n'est pas assez significative pour pouvoir en tirer des conclusions.

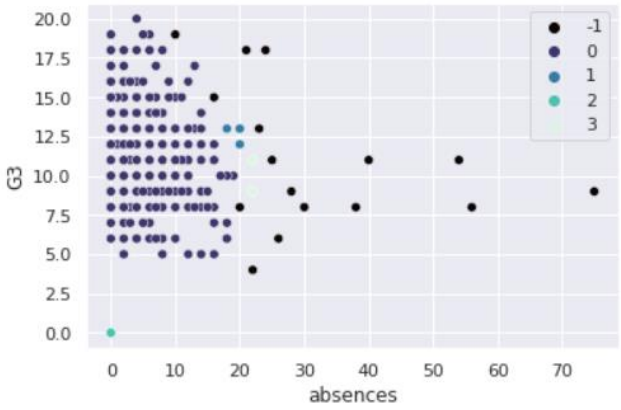
Il semblerait donc qu'il n'y ait pas de combinaisons judicieuses de ces paramètres qui puisse expliquer la note d'un individu.

DBSCAN

Le DBSCAN permet de séparer et regrouper des données en groupes homogènes ayant des caractéristiques communes, en se basant sur différentes variables. On choisit les paramètres du modèle pour qu'il soit le plus efficace possible : avoir un nombre limité de cluster pertinent, en se basant sur une courbe du nombre de cluster en fonction des paramètres.

Nous avons réalisé un DBSCAN en fonction de deux paramètres : le nombre d'absences et la note reçue. Nous avons obtenu ces graphes :





Du fait du manque de données continues, l'utilisations de DBSCAN est relativement limité à ces deux variables. La répartition des individus ayant déjà été interprétée auparavant, ici nous nous intéressons davantage à leur répartition entre eux.

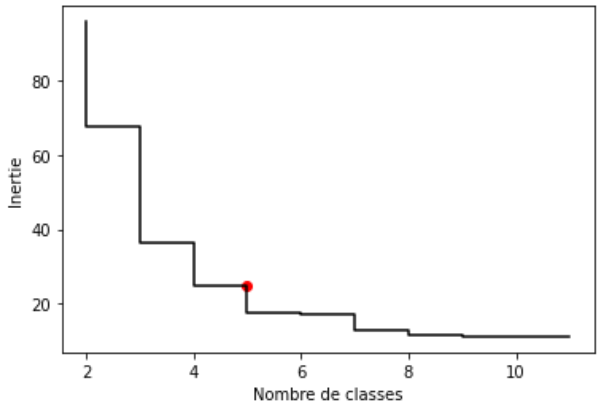
Il est clair que nous ne pouvons observer de clusters pertinents, et ce même en ayant choisi un nombre de clusters bien définit par une approche Elbow.

CAH ET KMEANS

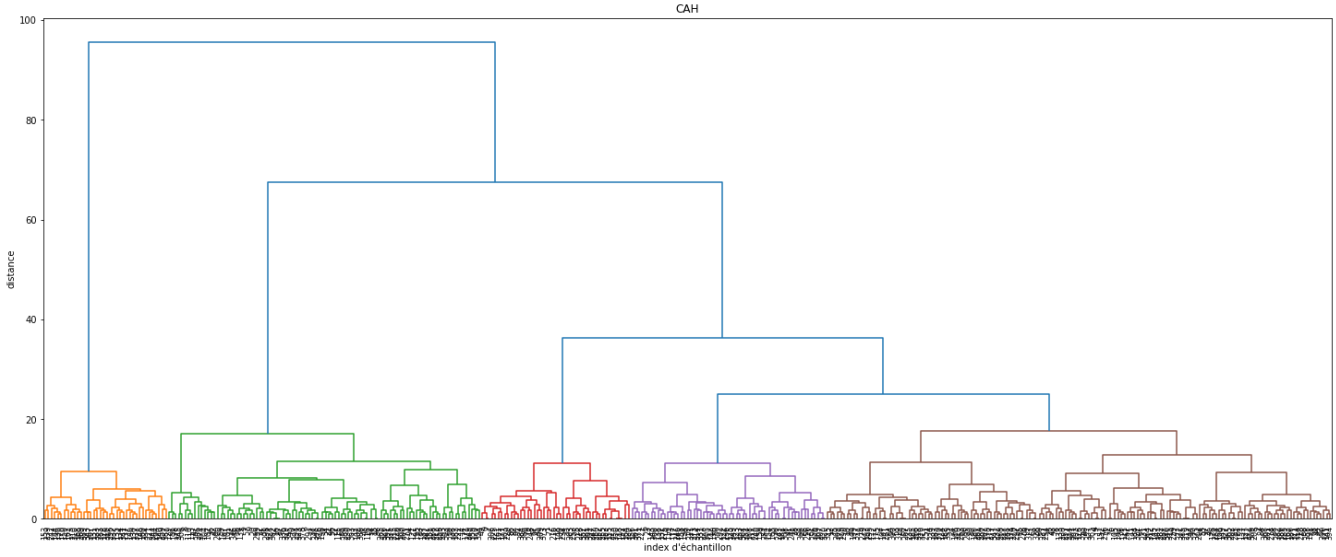
La Classification Ascendante Hiérarchique est une méthode de classification, qui permet de partitionner une population en différentes classes ou sous-groupes. Il faut au préalable définir le nombre optimal de classes avec un diagramme d'inertie.

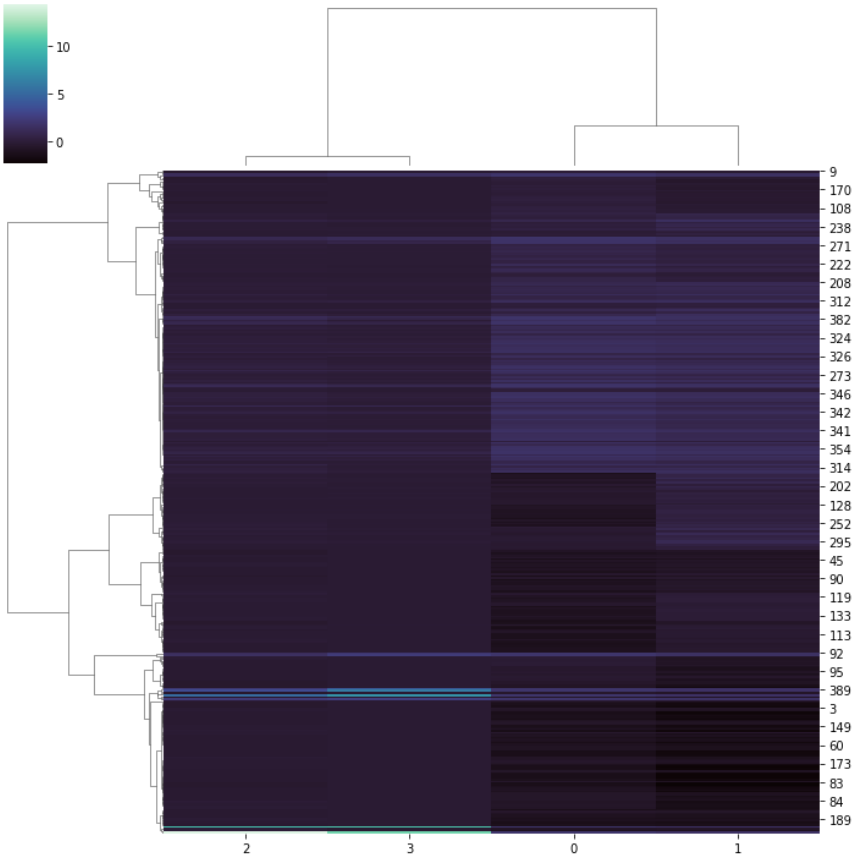
Pour la première classification ascendante hiérarchique non avons utilisé les résultats du test Backward Forward c'est à dire les variables suivantes : "Medu", "studytime", "goout", "sex_M", "famsize_LE3", "higher_yes", "romantic_yes","G3"

On a tracé le diagramme d'inertie de ces variables et on obtient le résultat suivant :



On constate que pour le CAH il est pertinent de déterminer 5 classes. Car après 5 classes l'inertie est peu modifiée.

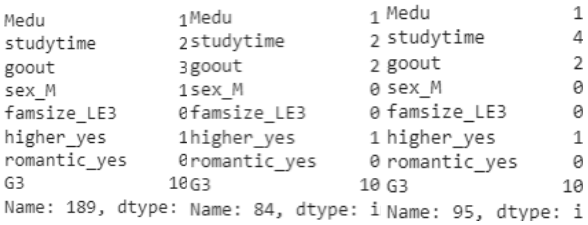




On observe à droite les individus et à gauche leurs classification.

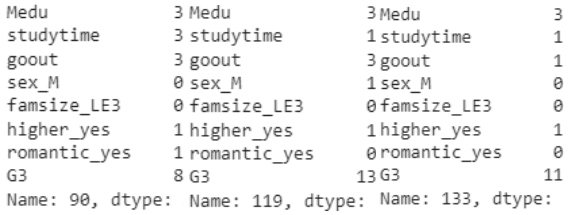
Nous allons donc comparer les individus entre eux afin de voir comment ils ont été classifiés.

Classe 1 (en bas à droite)



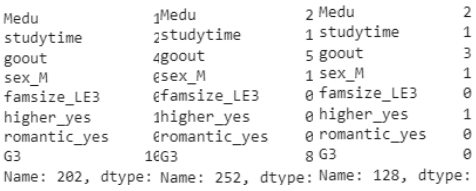
--> Profil de cette classe : étudiant ayant une famille de plus de 3 personnes, une mère qui a un niveau de scolarité faible, ils sont n'ont pas de relations amoureuses, ils ont envie de faire des études, passent un peu de temps à sortir avec leurs amis, ils ont des notes en mathématiques de 10.

Classe 2 :



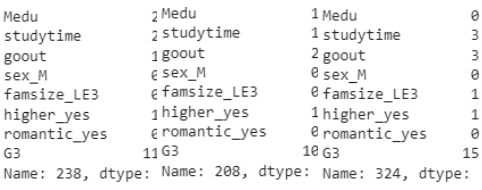
--> profil de cette classe : étudiants ayant une mère qui a un bon niveau scolaire, famille de plus de 3 personnes, ont envie de faire des études, sortent davantage avec leurs amis, passent moins de temps à étudier que la classe 1. Ils ont des notes en mathématiques un peu moins homogènes.

Classe 3 :



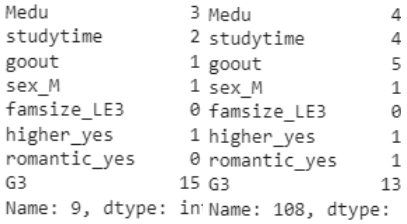
--> profil de cette classe : étudiants qui sortent beaucoup avec leurs amis, n'ont pas de relations amoureuses, étudient très peu, ont une famille de plus de 3 personnes, ont une mère qui n'a pas fait beaucoup d'études, n'ont pas de bonnes notes en mathématiques.

Classe 4 :



--> profil de cette classe : ce sont des filles, ayant toutes envie de faire des études, n'ont pas de relations amoureuses, leurs mères ont un niveau scolaire faible, elles ont plus de 10 en mathématiques.

Classe 5 (en haut à droite) :

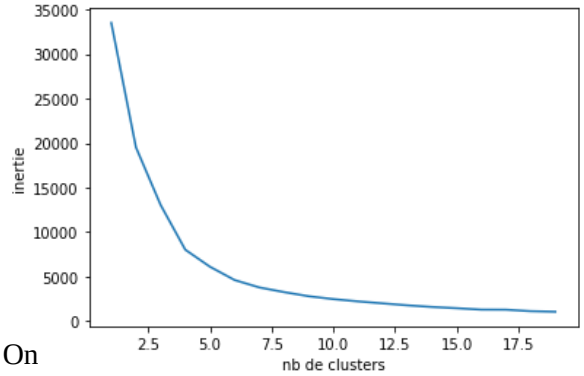


--> profil de cette classe : ce sont des garçons, leurs mères ont fait des études, ils passent du temps à étudier, font parties d'une famille de plus de 3 personnes, ont envie de faire des études, ont de bonnes notes en mathématiques.

Via ce CAH, on constate que les individus en haut du graphe ont leurs mères qui ont fait des études, ont des notes plus élevées.

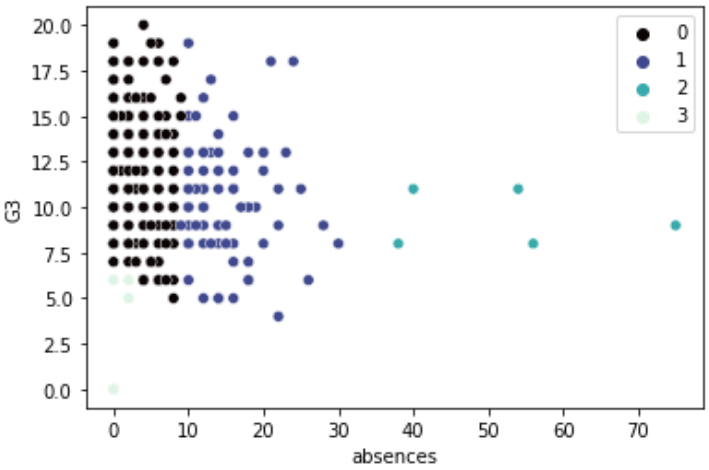
KMeans de la variable "absences" par rapport à G3, notre variable target :

Afin de savoir combien de clusters nous voulons obtenir on utilise la méthode du coude :



On constate que 4 clusters semblent être le plus appropriés.

On obtient le KMeans suivant :

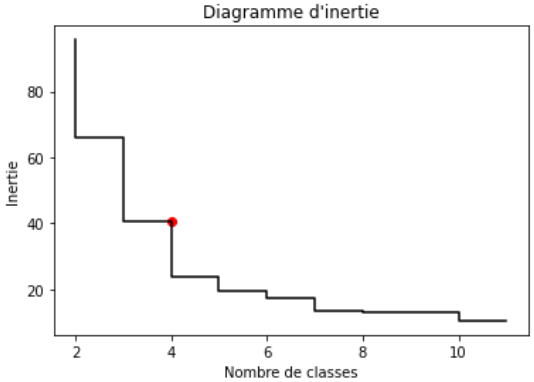


Nos centroïdes sont les suivants :

[3,11.98] , [14.85 ,10.31] ,[0.1, 0.41],[52.6, 9.4].

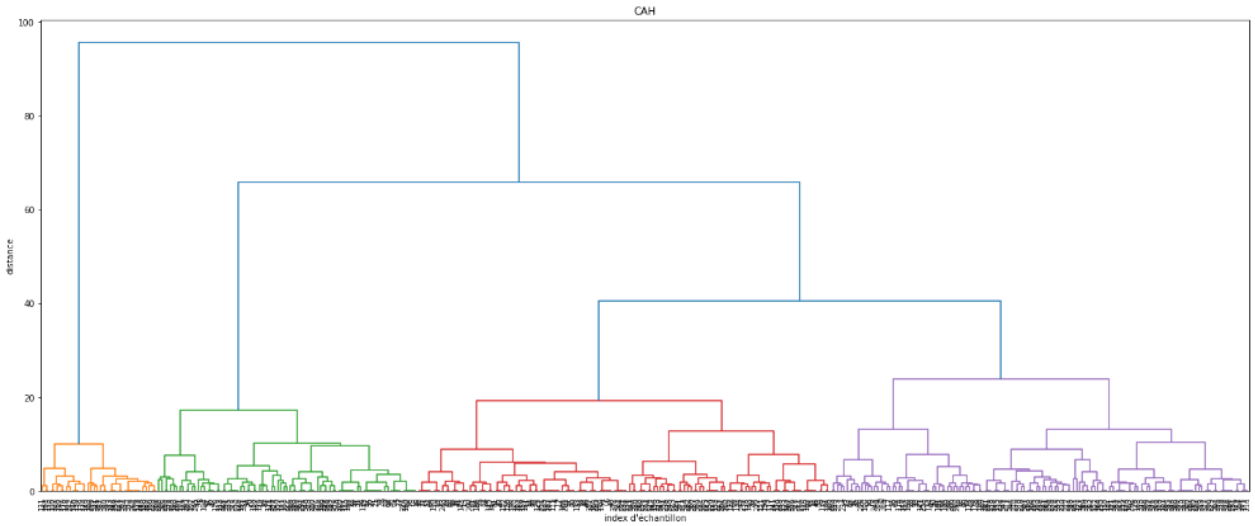
On observe 4 profils d'élèves. Ceux ayant entre 0 et absences et ayant des notes comprises entre 5 et 20, ceux ayant entre 10 et 30 absences et des notes comprises entre 4 et 19, ceux ayant plus de 40 absences avec des notes entre 8 et 12 et un petit groupe d'étudiants ayant moins de 5 absences et des notes inférieures à 8. On constate que les étudiants qui ont plus de 30 absences n'ont jamais une note supérieure à 12. Par ailleurs, les étudiants qui ont plus que 18 ont moins de 10 absences. On constate également qu'il y a un étudiant ayant 0 absences et 0 de moyenne.

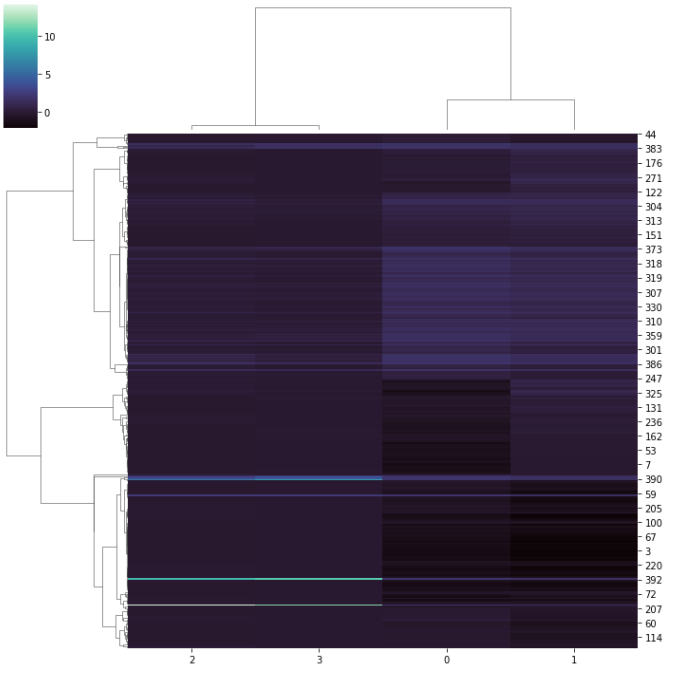
On effectue également un deuxième CAH avec les variables suivantes : "Medu","age","higher_yes","G3"



En effectuant le graphique d'inertie on constate que 4 classes différentes semblent être le mieux.

On obtient le CAH suivant :





Classe 1 :

Medu	2	Medu	4	Medu	1	Medu	4
age	15	age	16	age	15	age	16
higher_yes	1	higher_yes	1	higher_yes	1	higher_yes	1
G3	9	G3	11	G3	5	G3	16
Name: 114, dtype:		Name: 60, dtype:		Name: 72, dtype:		Name: 59, dtype:	

--> Education de la mère soit très faible soit élevée. Ils ont tous envie de faire des études supérieures, leurs âges sont de 15-16 ans avec un écart-type des notes important.

Classe 2:

```
Medu          2
age          17
higher_yes    1
G3           13
Name: 236, dtype

Medu          4
age          18
higher_yes    1
G3           11
Name: 325, dtype

Medu          3
age          22
higher_yes    0
G3            8
Name: 247, dtype
```

--> On constate un âge plus élevé pour cette classe, entre 17 et 22 ans, avec notes plus concentrées. L'éducation de la mère semble être plus élevée pour ces étudiants.

Classe 3

```
Medu          2
age          16
higher_yes    1
G3           11
Name: 176, dtype

Medu          2
age          18
higher_yes    1
G3           14
Name: 271, dtype

Medu          4
age          19
higher_yes    1
G3            8
Name: 307, dtype

Medu          4
age          17
higher_yes    1
G3           10
Name: 301, dtype

Medu          3
age          17
higher_yes    1
G3           10
Name: 318, dtype
```

--> On constate des notes très homogènes, en moyenne 10. Des étudiants d'environ 17 ans, ayant tous envie de faire des études supérieures.

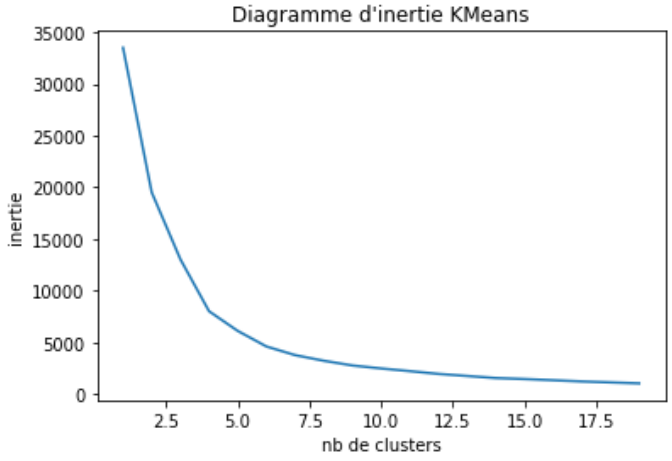
Classe 4 :

```
Medu          2
age          16
higher_yes    1
G3            9
Name: 44, dtype:

Medu          1
age          19
higher_yes    1
G3            0
Name: 383, dtype
```

--> On retrouve dans cette classe des étudiants ayant envie de faire des études, n'ayant pas forcément des bonnes notes et ayant une mère qui a un faible niveau de scolarité.

On réalise un Kmeans avec les variables suivantes : 'Medu','age','higher_yes'



4 clusters nous semblent être le plus pertinent.

Nous obtenons les centroïdes suivants :

[3.52252252, 17.7027027, 0.96396396],[3.61344538, 15.52941176, 0.98319328],[1.68656716, 15.6119403 , 0.98507463] ,[1.55102041, 17.71428571, 0.86734694]

On constate que pour les 4 centroïdes, les étudiants ont tous globalement envie de faire des études supérieures. Que ceux qui ont le moins envie sont ceux dont leurs mères ont le moins d'éducation.

Ceux qui en général veulent le plus faire des études sont ceux dont leur mère a un niveau de scolarité élevé.

Il a été difficile de faire des KMeans vu notre base de données car nous avons très peu de valeurs quantitatives continues.

CONCLUSION

Durant cette étude, nous nous sommes aperçus qu'il était dur de prédire les notes avec les données et variables de notre base de données. En effet, notre variable note n'était corrélé que très faiblement avec les autres.

Nous avons cependant vu dans nos régressions linéaires multiples, nos GLM binomial et notre ACP que nous pouvions expliquer les mauvaises notes, mais assez peu les bonnes. Cela est lié au fait que les variables de notre base de données sont des facteurs d'échecs. En effet, des variables comme sortir souvent, avoir des parents avec peu d'éducation, avoir déjà redoublé, être en couple auront des impacts négatifs sur les notes.

La seule variable qui aurait peu expliquer nos bonnes notes est le temps d'étude, comme nous avons pu l'observer lors de l'AFC, et dans une plus faible mesure, le travail des parents.

Cependant, une minorité d'élèves travaillent beaucoup. Pour ces élèves, travailler beaucoup est un facteur de réussite, mais pour les autres travailler peu n'est pas forcément un facteur d'échec.

Pour expliquer l'ensemble des notes, il nous faudrait ainsi d'autres variables, qui explique pourquoi des élèves qui n'ont pas de facteur d'échec ne réussissent pas tous pareil.