La consommation d’alcool chez les étudiants

PERINET Anaëlle

Table des matières

1 Introduction

Dans le cadre de ce mini projet, j'ai choisi de travailler sur le jeu de données « student-por » disponible sur kaggle.com

J'ai choisi de traiter ce sujet car d'après une étude de la MAAF l'isolement provoqué par la crise sanitaire actuelle aurait conduit 15% des étudiants à consommer bien plus d'alcool qu'avant.

Le jeu de données contient 649 observations de 33 variables. Les données ont été obtenues dans le cadre d'une enquête menée auprès d'étudiants en portugais à l'école secondaire. Il contient beaucoup d'informations sociales, de genre et d'étude intéressantes sur les étudiants. Afin de répondre à notre question j'ai sélectionné uniquement les variables qui me semblaient pertinentes, et j'ai, pour la plupart changé leur type pour qu'elles soient toutes de type numérique.

L'objectif est de comprendre quels sont les critères (genre, âge, milieu social..) qui peuvent influer sur la consommation d'alcool des étudiants.

2 Dictionnaire des données

Le jeu de données que nous utilisons est disponible au lien suivant :
https://www.kaggle.com/uciml/student-alcohol-consumption?select=student-por.csv

Pour ce travail j'ai sélectionné les variables qui me semblaient les plus intéressantes et représentatives :

- Dalc
- Walc
- Age
- Medu
- Fedu
- Studytime
- Failures
- Famrel
- Freetime
- Goout
- Health
- Absences
- G1
- G2
- Sex
- Address
- Famsize
- Pstatus
- Famsup
- Paid
- Activities
- Higher
- Romantic

Vous trouverez joint à ce rapport un fichier excel contenant le dictionnaire des données détaillé.

3 Consommation d'alcool en Semaine

1. Statistiques Descriptives

Avant de commencer à travailler j'ai vérifié que le jeu de données n'avait pas de données manquantes.

Variable	Nbre manquant	Moyenne	Ec-type	Minimum	Maximum
age	0	16.7442219	1.2181376	15.0000000	22.0000000
Medu	0	2.5146379	1.1345520	0	4.0000000
Fedu	0	2.3066256	1.0999309	0	4.0000000
studytime	0	1.9306626	0.8295096	1.0000000	4.0000000
failures	0	0.2218798	0.5932351	0	3.0000000
famrel	0	3.9306626	0.9557169	1.0000000	5.0000000
freetime	0	3.1802773	1.0510926	1.0000000	5.0000000
goout	0	3.1848998	1.1757661	1.0000000	5.0000000
Dalc	0	1.5023112	0.9248344	1.0000000	5.0000000
Walc	0	2.2804314	1.2843800	1.0000000	5.0000000
health	0	3.5362096	1.4462591	1.0000000	5.0000000
absences	0	3.6594761	4.6407588	0	32.0000000
G1	0	11.3990755	2.7452651	0	19.0000000
G2	0	11.5701079	2.9136387	0	19.0000000
Sex	0	0.4098613	0.4921872	0	1.0000000
Address	0	0.3035439	0.4601426	0	1.0000000
Famsize	0	0.7041602	0.4567714	0	1.0000000
Pstatus	0	0.1232666	0.3289965	0	1.0000000
Famsup	0	0.6132512	0.4873809	0	1.0000000
Paid	0	0.0600924	0.2378414	0	1.0000000
Activities	0	0.4853621	0.5001712	0	1.0000000
Higher	0	0.8936826	0.3084812	0	1.0000000
Romantic	0	0.3682589	0.4827041	0	1.0000000

Nous allons dans un premier temps nous intéresser uniquement à la consommation d'alcool en semaine, voici la liste des variables qui vont nous intéresser et leurs attributs.

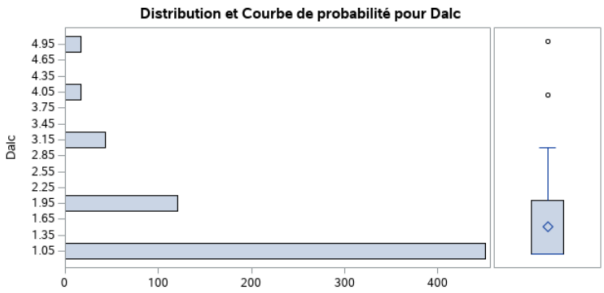
Liste alphabétique des variables et des attributs

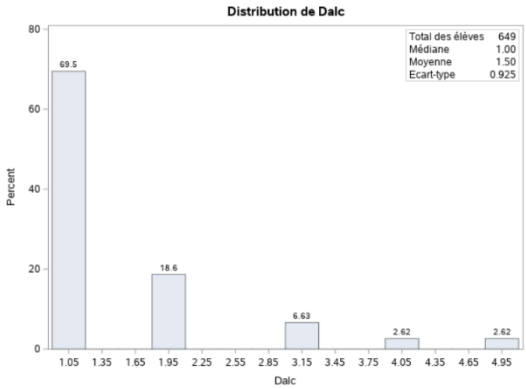
#	Variable	Type	Long.	Format	Informat
20	Activities	Num.	8		
15	Address	Num.	8		
9	Dalc	Num.	8	BEST12.	BEST32.
16	Famsize	Num.	8		
18	Famsup	Num.	8		
3	Fedu	Num.	8	BEST12.	BEST32.
12	G1	Num.	8	BEST12.	BEST32.
13	G2	Num.	8	BEST12.	BEST32.
21	Higher	Num.	8		
2	Medu	Num.	8	BEST12.	BEST32.
19	Paid	Num.	8		
17	Pstatus	Num.	8		
22	Romantic	Num.	8		
14	Sex	Num.	8		
11	absences	Num.	8	BEST12.	BEST32.
1	age	Num.	8	BEST12.	BEST32.
5	failures	Num.	8	BEST12.	BEST32.
6	famrel	Num.	8	BEST12.	BEST32.
7	freetime	Num.	8	BEST12.	BEST32.
8	goout	Num.	8	BEST12.	BEST32.
10	health	Num.	8	BEST12.	BEST32.
4	studytime	Num.	8	BEST12.	BEST32.

Les élèves évaluent leur consommation d'alcool sur un échelle de 1 à 5, le 1 étant une non-consommation/consommation très faible et le 5 étant une consommation très importante.

Moments			
N	649	Somme des poids	649
Moyenne	1.50231125	Somme des observations	975
Ecart-type	0.92483443	Variance	0.85531872
Skewness	2.14191336	Kurtosis	4.34929747
Somme des carrés non corrigée	2019	Somme des carrés corrigée	554.246533
Coeff Variation	61.5607739	Std Error Mean	0.03630293

La moyenne obtenue se situe donc entre le 1 et le 2 sur notre échelle, ce qui représente une consommation assez faible.





Ces deux graphiques nous confirment que l'immense majorité des élèves (69,5%) estiment leur consommation d'alcool en semaine au seuil 0, contre même pas 5% des élèves qui s'estiment aux seuils 4 et 5.

Table de Dalc par age

Dalc	age								
	17	16	18	15	19	20	21	22	Total
1	119	130	96	81	20	4	1	0	451
2	36	33	25	22	3	2	0	0	121
3	15	8	8	7	5	0	0	0	43
4	3	3	6	1	4	0	0	0	17
5	6	3	5	1	0	0	1	1	17
Total	179	177	140	112	32	6	2	1	649

Cette table des fréquences par âge nous permet de une fois de plus que tout âge confondu c'est la catégorie 1 qui est la plus présente.

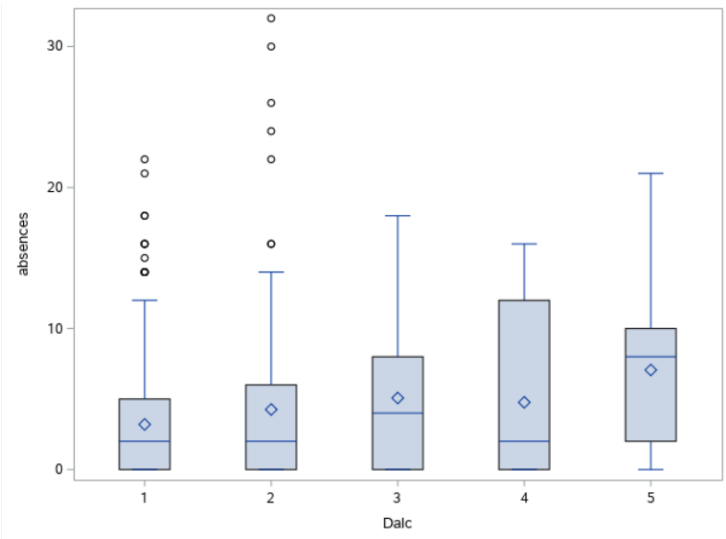
On peut constater que la catégorie d'âge pour laquelle le plus d'individu se situent dans la classe 1 sont les 16 ans, ils sont 73%.

A l'inverse, sans compter les 20, 21 et 22 ans car ils sont trop peu nombreux, la catégorie d'âge pour laquelle le plus d'individu se situent dans la classe 5 sont les 18 ans, ils sont 3,6%.

Table de Dalc par Sex			
	Sex		
Dalc	0	1	Total
1	305	146	451
2	58	63	121
3	11	32	43
4	7	10	17
5	2	15	17
Total	383	266	649

Cette table des fréquences par sexe, sachant que le 0 représente les filles et donc que le 1 représente les garçons, nous permet de voir seul au seuil 1, celui de la plus basse consommation d'alcool les filles sont devant les garçons :

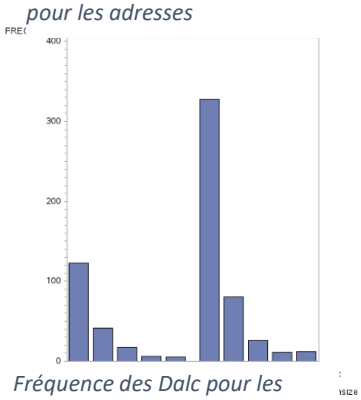
- 79,6% des filles se situent au niveau 1 contre 54,9% des garçons.
- 15,1% des filles se situent au niveau 2 contre 23,7% des garçons.
- 2,9% des filles se situent au niveau 3 contre 12% des garçons.
- 1,8% des filles se situent au niveau 4 contre 3,8% des garçons.
- 0,5% des filles se situent au niveau 5 contre 5,6% des garçons.



Boxplot des Absences en fonction de Dalc

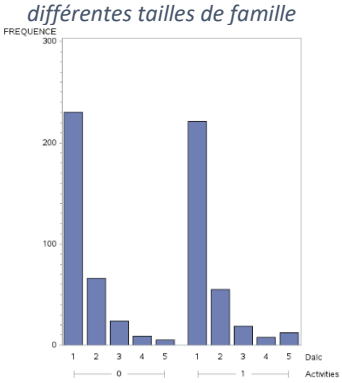
On constate ici que la consommation d'alcool en semaine à un léger impact sur le nombres d'absences en classe. La moyenne du nombre d'absences d'un élève qui classe sa consommation au niveau 5 est plus élevée que les autres en revanche la moyenne des absences des élèvent de niveau 4 est plus faible que les autres.

Fréquence des niveaux de Dalc pour les adresses



Ce graphique nous permet de voir que peut importe la taille de la famille la consommation d'alcool en semaine est inchangée, qu'on soit dans une petite ou dans une grande famille c'est le consommation de niveau 1 qui est la plus importante.

Fréquence des Dalc pour les différentes tailles de famille



Il en est de même avec ce graphique, qu'on pratique des activités extra-scolaire ou non la catégorie la plus représentée est la première.

En revanche on peut constater que les élèves qui pratiquent une activité extra-scolaire ont plus tendance à avoir une consommation de niveau 5.

La procédure CORR

22 Variables : age Medu Fedu studytime failures famrel freetime goout Dalc health absences G1 G2 Sex Address Famsize Pstatus Famsup Paid Activities Higher Romantic

Coefficients de corrélation de Pearson, N = 649

	age	Medu	Fedu	studytime	failures	famrel	freetime	goout	Dalc	health	absences	G1	G2	Sex	Address	Famsize	Pstatus	Famsup	Paid	Activities	Higher	Romantic
age	1.00000	-0.10783	-0.12105	-0.00842	0.31997	-0.02056	-0.00491	0.11280	0.13477	-0.00875	0.15000	-0.17432	-0.10712	-0.04366	0.02585	0.00247	0.00563	-0.10189	-0.00546	-0.05428	-0.26550	0.17881
Medu	-0.10783	1.00000	0.64748	0.09701	-0.17221	0.02442	-0.01969	0.00954	-0.00702	0.00461	-0.00858	0.26047	0.26404	0.11913	-0.19032	0.01432	0.05717	0.12049	0.11397	0.11935	0.21390	-0.03099
Fedu	-0.12105	0.64748	1.00000	0.05040	-0.16592	0.02026	0.00684	0.02769	0.00006	0.04491	0.02986	0.21750	0.22514	0.08391	-0.14149	0.03954	0.03186	0.13519	0.09463	0.07970	0.19174	-0.06757
studytime	-0.00842	0.09701	0.05040	1.00000	-0.14744	-0.00413	-0.06883	-0.07544	-0.13758	-0.05643	-0.11839	0.26088	0.24050	-0.20621	-0.06202	0.01094	0.00875	0.14351	-0.00231	0.07008	0.18826	0.03304
failures	0.31997	-0.17221	-0.16592	-0.14744	1.00000	-0.06265	0.10899	0.04508	0.10595	0.03559	0.12278	-0.38421	-0.38578	0.07389	0.06382	0.06607	0.00988	-0.00698	0.06942	0.00056	-0.30940	0.06990
famrel	-0.02056	0.02442	0.02026	-0.00413	-0.06265	1.00000	0.12922	0.08971	-0.07577	0.10956	-0.08953	0.04879	0.08959	0.08347	0.03390	-0.00464	-0.05130	0.01523	0.03194	0.05760	0.04824	-0.04492
freetime	-0.00491	-0.01969	0.00684	-0.06883	0.10899	0.12922	1.00000	0.34635	0.10990	0.08453	-0.01872	-0.09450	-0.10668	0.14631	0.03665	0.02126	-0.03759	0.00376	-0.04957	0.15033	-0.10262	0.02711
goout	0.11280	0.00954	0.02769	-0.07544	0.04508	0.08971	0.34635	1.00000	0.24513	-0.01574	0.08537	-0.07405	-0.07947	0.05818	-0.01548	0.00431	-0.03109	0.01726	-0.00668	0.08858	-0.06911	-0.00052
Dalc	0.13477	-0.00702	0.00006	-0.13758	0.10595	-0.07577	0.10990	0.24513	1.00000	0.05907	0.17295	-0.19517	-0.18948	0.28270	0.04730	-0.06048	-0.04151	-0.01684	0.05199	0.02259	-0.13166	0.06204
health	-0.00875	0.00461	0.04491	-0.05643	0.03559	0.10956	0.08453	-0.01574	0.05907	1.00000	-0.03023	-0.05165	-0.08218	0.13955	-0.00379	-0.00245	-0.01264	0.01880	0.06320	0.01300	0.01729	-0.01802
absences	0.15000	-0.00858	0.02986	-0.11839	0.12278	-0.08953	-0.01872	0.08537	0.17295	-0.03023	1.00000	-0.14715	-0.12474	0.02134	-0.07365	-0.00465	0.11749	0.04198	-0.03596	-0.01512	-0.12989	0.07949
G1	-0.17432	0.26047	0.21750	0.26088	-0.38421	0.04879	-0.09450	-0.07405	-0.19517	-0.05165	-0.14715	1.00000	0.86498	-0.10411	-0.15713	-0.04723	-0.01525	0.03826	-0.06278	0.08012	0.34903	-0.07497
G2	-0.10712	0.26404	0.22514	0.24050	-0.38578	0.08959	-0.10668	-0.07947	-0.18948	-0.08218	-0.12474	0.86498	1.00000	-0.10401	-0.15460	-0.03889	-0.01869	0.03814	-0.03393	0.06715	0.33195	-0.09794
Sex	-0.04366	0.11913	0.08391	-0.20621	0.07389	0.08347	0.14631	0.05818	0.28270	0.13955	0.02134	-0.10411	-0.10401	1.00000	-0.02550	-0.09820	-0.06470	-0.12947	0.07930	0.12471	-0.05813	-0.11014
Address	0.02585	-0.19032	-0.14149	-0.06202	0.06382	0.03390	0.03665	-0.01548	0.04730	-0.00379	-0.07365	-0.15713	-0.15460	-0.02550	1.00000	0.04611	-0.09464	-0.00556	0.03048	0.00928	-0.07671	0.03094
Famsize	0.00247	0.01432	0.03954	0.01094	0.06607	-0.00464	0.02126	0.00431	-0.06048	-0.00245	-0.00465	-0.04723	-0.03889	-0.09820	0.04611	1.00000	-0.23961	0.03982	0.05025	0.01479	-0.00452	0.03294
Pstatus	0.00563	0.05717	0.03186	0.00875	0.00988	-0.05130	-0.03759	-0.03109	-0.04151	-0.01264	0.11749	-0.01525	-0.01869	-0.06470	-0.09464	-0.23961	1.00000	-0.01020	-0.01592	-0.10156	-0.02273	0.05383
Famsup	-0.10189	0.12049	0.13519	0.14351	-0.00698	0.01523	0.00376	0.01726	-0.01684	0.01880	0.04198	0.03826	0.03814	-0.12947	-0.00556	0.03982	-0.01020	1.00000	0.09430	-0.00743	0.08534	-0.02340
Paid	-0.00546	0.11397	0.09463	-0.00231	0.06942	0.03194	-0.04957	-0.00668	0.05199	0.06320	-0.03596	-0.06278	-0.03393	0.07930	0.03048	0.05025	-0.01592	0.09430	1.00000	0.06578	0.02411	-0.01831
Activities	-0.05428	0.11935	0.07970	0.07008	0.00056	0.05760	0.15033	0.08858	0.02259	0.01300	-0.01512	0.08012	0.06715	0.12471	0.00928	0.01479	-0.10156	-0.00743	0.06578	1.00000	0.04491	0.05752
Higher	-0.26550	0.21390	0.19174	0.18826	-0.30940	0.04824	-0.10262	-0.06911	-0.13166	0.01729	-0.12989	0.34903	0.33195	-0.05813	-0.07671	-0.00452	-0.02273	0.08534	0.02411	0.04491	1.00000	-0.09939
Romantic	0.17881	-0.03099	-0.06767	0.03304	0.06990	-0.04492	0.02711	-0.00052	0.06204	-0.01802	0.07949	-0.07497	-0.09794	-0.11014	0.03094	0.03294	0.05383	-0.02340	-0.01831	0.05752	-0.09939	1.00000

La matrice des corrélations possède un grand nombre de d'informations.

On peut comprendre de ce tableau de la consommation d'alcool dépend beaucoup du sexe de l'élève et de si oui ou non l'étudiant sort avec ses amis.

On comprend également que les absences, le temps passer à étudier, les résultats du premier semestre et les résultats du second semestre sont influencés en grande part par la consommation d'alcool de l'élève en semaine.

2. Analyse en Composante Principale

Pour trouver le modèle le plus adapté à nos données nous en avons essayé trois :

- L'élimination descendante
- L'élimination ascendante
- La sélection Stepwise

L'élimination descendante nous donne un R^2 de 0,1881.

L'élimination ascendante nous donne un R^2 de 0,2003.

La sélection stepwise nous donne un R^2 de 0,1912.

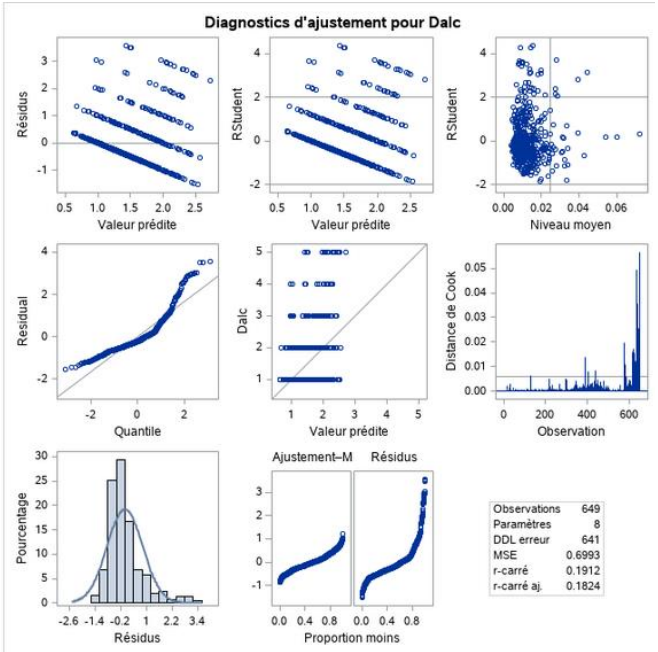
Même si l'élimination ascendante est celle qui a le R^2 le plus élevé, ce n'est pas celle qui que nous allons sélectionner car cette méthode conserve 12 variables, ce qui est trop.

Nous avons donc choisi de conserver la sélection stepwise car son R^2 est correcte et qu'elle ne conserve que 7 variables.

Cependant, le R^2 n'est pas très élevé donc le modèle n'explique pas beaucoup de variance.

Synthèse de Sélection Stepwise

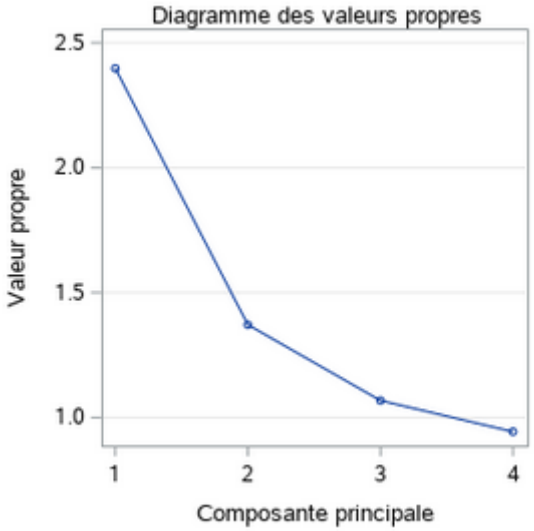
Etape	Variable entrée	Variable supprimée	Nombre var. dans	R carré partiel	R carré du modèle	C(p)	Valeur F	Pr > F
1	Sex		1	0.0799	0.0799	78.9307	56.20	<.0001
2	goout		2	0.0525	0.1324	39.6453	39.07	<.0001
3	G1		3	0.0229	0.1553	23.6307	17.48	<.0001
4	absences		4	0.0164	0.1717	12.7099	12.77	0.0004
5	famrel		5	0.0098	0.1815	7.0185	7.68	0.0057
6	age		6	0.0066	0.1881	3.8176	5.23	0.0226
7	Romantic		7	0.0031	0.1912	3.3405	2.50	0.1147



On obtient la nouvelle matrice des corrélations :

Matrice de corrélation									
	Dalc	age	famrel	goout	absences	G1	Sex	Romantic	numero_eleve
Dalc	1.0000	0.1348	-.0758	0.2451	0.1730	-.1952	0.2827	0.0620	0.7342
age	0.1348	1.0000	-.0206	0.1128	0.1500	-.1743	-.0437	0.1788	0.2948
famrel	-.0758	-.0206	1.0000	0.0897	-.0895	0.0488	0.0835	-.0449	-.0455
goout	0.2451	0.1128	0.0897	1.0000	0.0854	-.0741	0.0582	-.0005	0.1995
absences	0.1730	0.1500	-.0895	0.0854	1.0000	-.1471	0.0213	0.0795	0.1302
G1	-.1952	-.1743	0.0488	-.0741	-.1471	1.0000	-.1041	-.0750	-.1970
Sex	0.2827	-.0437	0.0835	0.0582	0.0213	-.1041	1.0000	-.1101	0.6352
Romantic	0.0620	0.1788	-.0449	-.0005	0.0795	-.0750	-.1101	1.0000	0.0312
numero_eleve	0.7342	0.2948	-.0455	0.1995	0.1302	-.1970	0.6352	0.0312	1.0000

En regardant la matrice des corrélations on peut considérer que les variables goout et Sex sont les plus corrélées avec la consommation d'alcool.



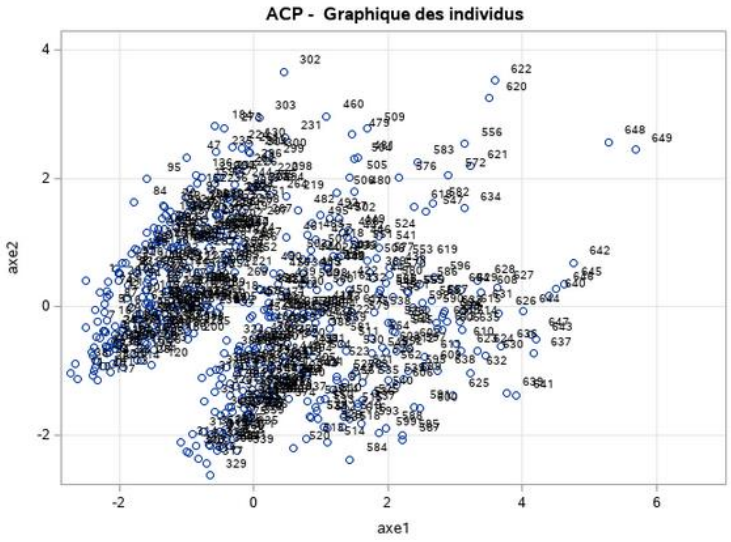
Sur ce graphique on peut voir que le coude est formé au niveau du deuxième axe.

Mais pour le choix du nombre d'axes de l'ACP définitif on va prendre les axes dont les valeurs propres sont supérieures à 1.

Valeurs propres de la matrice de corrélation				
	Valeur propre	Différence	Proportion	Cumulé
1	2.39771519	1.02625933	0.2664	0.2664
2	1.37145586	0.30332149	0.1524	0.4188
3	1.06813437	0.12490303	0.1187	0.5375
4	0.94323134		0.1048	0.6423

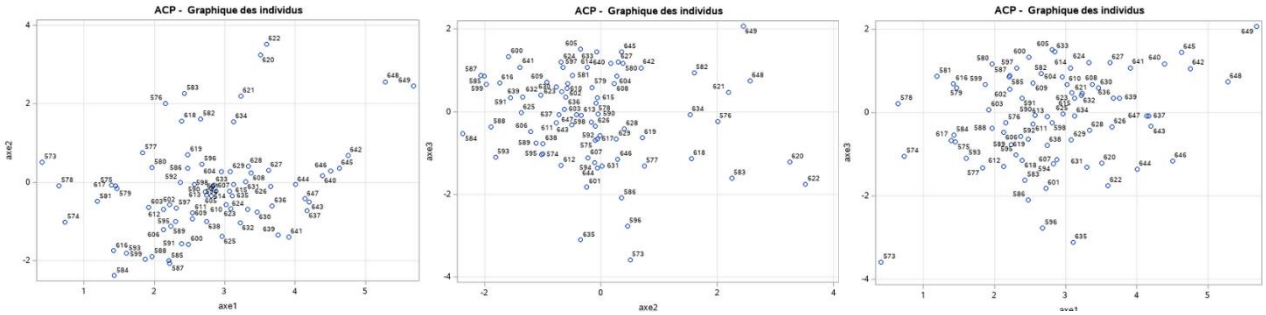
On va donc garder les 3 premiers axes pour un valeur cumulée de 53,75%.

De-là on obtient un graphique des individus :

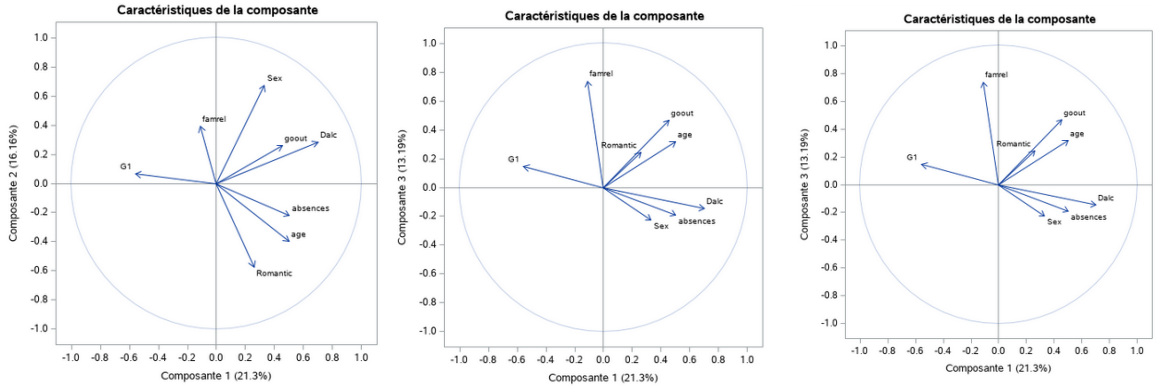


On voit tout de suite que ce graphique est illisible, il contient trop d'informations pour être étudié.

On va alors garder uniquement les consommations d'alcool de niveau 3, 4 et 5. On obtient alors :



On obtient également les graphiques des caractéristiques de la composante :



A l'observation de ses graphiques nous pouvons dire que les variables Dalc, Absences, Sex et goout sont reliées.

4 Consommation d'alcool le Week-end

3. Statistiques Descriptives

Nous allons maintenant nous intéresser uniquement à la consommation d'alcool le week-end, voici la liste des variables qui vont nous intéresser et leurs attributs.

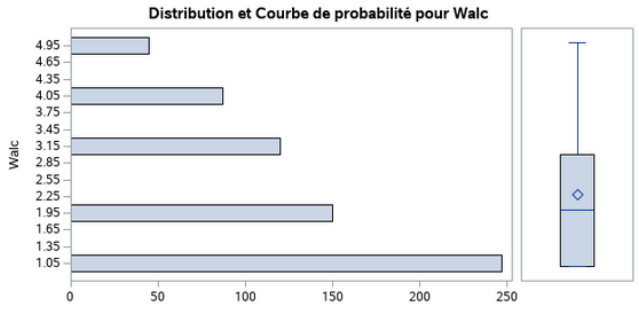
Liste alphabétique des variables et des attributs

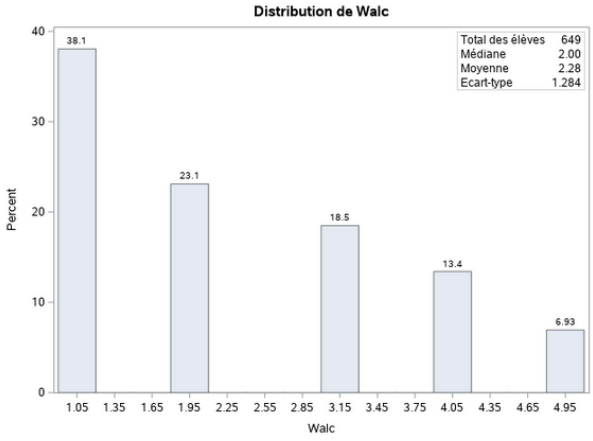
#	Variable	Type	Long.	Format	Informat
20	Activities	Num.	8		
15	Address	Num.	8		
16	Famsize	Num.	8		
18	Famsup	Num.	8		
3	Fedu	Num.	8	BEST12.	BEST32.
12	G1	Num.	8	BEST12.	BEST32.
13	G2	Num.	8	BEST12.	BEST32.
21	Higher	Num.	8		
2	Medu	Num.	8	BEST12.	BEST32.
19	Paid	Num.	8		
17	Pstatus	Num.	8		
22	Romantic	Num.	8		
14	Sex	Num.	8		
9	Walc	Num.	8	BEST12.	BEST32.
11	absences	Num.	8	BEST12.	BEST32.
1	age	Num.	8	BEST12.	BEST32.
5	failures	Num.	8	BEST12.	BEST32.
6	famrel	Num.	8	BEST12.	BEST32.
7	freetime	Num.	8	BEST12.	BEST32.
8	goout	Num.	8	BEST12.	BEST32.
10	health	Num.	8	BEST12.	BEST32.
4	studytime	Num.	8	BEST12.	BEST32.

Les élèves évaluent leur consommation d'alcool sur un échelle de 1 à 5, le 1 étant une non-consommation/consommation très faible et le 5 étant une consommation très importante.

Moments			
N	649	Somme des poids	649
Moyenne	2.28043143	Somme des observations	1480
Ecart-type	1.28437997	Variance	1.64963191
Skewness	0.63590427	Kurtosis	-0.7706892
Somme des carrés non corrigée	4444	Somme des carrés corrigée	1068.96148
Coeff Variation	56.3217974	Std Error Mean	0.05041632

La moyenne obtenue se situe donc entre le 2 et le 3 sur notre échelle, ce qui représente une consommation d'alcool moyenne.





Contrairement à ce que nous avions observé pour la consommation en semaine, les résultats sont moins tranchés, la consommation de niveau 1 arrive toujours en tête mais elle est suivie de plus près des autres niveaux de consommation.

Table de Walc par age									
	age								
Walc	17	16	18	15	19	20	21	22	Total
1	54	71	49	56	14	2	1	0	247
2	48	43	33	20	4	1	1	0	150
3	40	26	24	20	9	1	0	0	120
4	21	25	23	11	5	2	0	0	87
5	16	12	11	5	0	0	0	1	45
Total	179	177	140	112	32	6	2	1	649

On peut constater que :

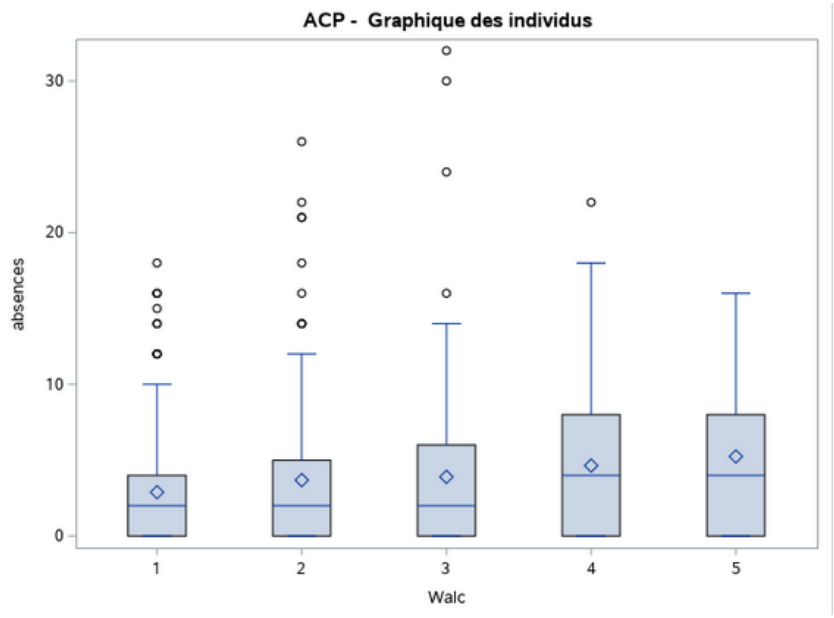
- 44% des 19 ans se situent dans la classe 1.
- 27% des 17 ans se situent dans la classe 2.
- 28% des 19 ans se situent dans la classe 3.
- 16% des 18 ans se situent dans la classe 4.
- 9% des 17 ans se situent dans la classe 5.

Table de Walc par Sex			
	Sex		
Walc	0	1	Total
1	176	71	247
2	99	51	150
3	71	49	120
4	30	57	87
5	7	38	45
Total	383	266	649

Cette table des fréquences par sexe, sachant que le 0 représente les filles et donc que le 1 représente les garçons, nous permet de voir seul que les filles sont plus nombreuses aux seuils 1, 2 et 3 tandis que les garçons sont en tête aux seuils 4 et 5.

On peut constater que :

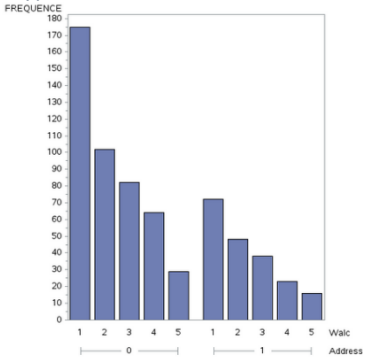
- 46,0% des filles se situent au niveau 1 contre 26,7% des garçons.
- 25,8% des filles se situent au niveau 2 contre 19,2% des garçons.
- 18,5% des filles se situent au niveau 3 contre 18,4% des garçons.
- 7,8% des filles se situent au niveau 4 contre 21,4% des garçons.
- 1,8% des filles se situent au niveau 5 contre 14,3% des garçons.



Boxplot des Absences en fonction de Walc

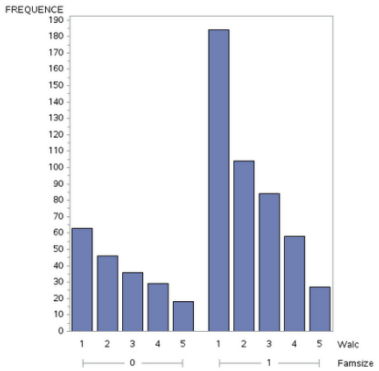
On constate ici que la consommation d'alcool en semaine à un très léger impact sur le nombres d'absences en classe. La consommation d'alcool le week-end n'a pas de réel impact sur les absences à l'école.

Fréquence des Walc pour les deux types d'adresse



Ce graphique nous permet de voir que peu importe la taille de la famille la consommation d'alcool en semaine est inchangée, qu'on soit dans une petite ou dans une grande famille c'est le consommation de niveau 1 qui est la plus importante.

Fréquence des Walc pour les deux tailles de familles



Il en est de même avec ce graphique, qu'on pratique des activités extra-scolaire ou non la catégorie la plus représentée est la première.

Coefficients de corrélation de Pearson, N = 649																						
	age	Medu	Fedu	studytime	failures	famrel	freetime	goout	Walc	health	absences	G1	G2	Sex	Address	Famsize	Pstatus	Famsup	Paid	Activities	Higher	Romantic
age	1.00000	-0.10783	-0.12105	-0.00842	0.31997	-0.02056	-0.00491	0.11280	0.08636	-0.00875	0.15000	-0.17432	-0.10712	-0.04366	0.02585	0.00247	0.00563	-0.10189	-0.00546	-0.05428	-0.26550	0.17881
Medu	-0.10783	1.00000	0.64748	0.09701	-0.17221	0.02442	-0.01969	0.00954	-0.01977	0.00461	-0.00858	0.26047	0.26404	0.11913	-0.19032	0.01432	0.05717	0.12049	0.11397	0.11935	0.21390	-0.03099
Fedu	-0.12105	0.64748	1.00000	0.05040	-0.16592	0.02026	0.00684	0.02769	0.03844	0.04491	0.02986	0.21750	0.22514	0.08391	-0.14149	0.03954	0.03186	0.13519	0.09463	0.07970	0.19174	-0.06767
studytime	-0.00842	0.09701	0.05040	1.00000	-0.14744	-0.00413	-0.06883	-0.07544	-0.21493	-0.05643	-0.11839	0.26088	0.24050	-0.20621	-0.06202	0.01094	0.00875	0.14351	-0.00231	0.07008	0.18826	0.03304
failures	0.31997	-0.17221	-0.16592	-0.14744	1.00000	-0.06265	0.10899	0.04508	0.08227	0.03559	0.12278	-0.38421	-0.38578	0.07389	0.06382	0.06607	0.00988	-0.00698	0.06942	0.00056	-0.30940	0.06990
famrel	-0.02056	0.02442	0.02026	-0.00413	-0.06265	1.00000	0.12922	0.08971	-0.09351	0.10956	-0.08953	0.04879	0.08959	0.08347	0.03390	-0.00464	-0.05130	0.01523	0.03194	0.05760	0.04824	-0.04492
freetime	-0.00491	-0.01969	0.00684	-0.06883	0.10899	0.12922	1.00000	0.34635	0.12024	0.08453	-0.01872	-0.09450	-0.10668	0.14631	0.03665	0.02126	-0.03759	0.00376	-0.04957	0.15033	-0.10262	0.02711
goout	0.11280	0.00954	0.02769	-0.07544	0.04508	0.08971	0.34635	1.00000	0.38868	-0.01574	0.08537	-0.07405	-0.07947	0.05818	-0.01548	0.00431	-0.03109	0.01726	-0.00668	0.08858	-0.06911	-0.00052
Walc	0.08636	-0.01977	0.03844	-0.21493	0.08227	-0.09351	0.12024	0.38868	1.00000	0.11499	0.15637	-0.15565	-0.16485	0.32078	0.01242	-0.08196	-0.07098	-0.06560	0.03568	0.03282	-0.08433	-0.01997
health	-0.00875	0.00461	0.04491	-0.05643	0.03559	0.10956	0.08453	-0.01574	0.11499	1.00000	-0.03023	-0.05165	-0.08218	0.13955	-0.00379	-0.00245	-0.01284	0.01880	0.06320	0.01300	0.01729	-0.01802
absences	0.15000	-0.00858	0.02986	-0.11839	0.12278	-0.08953	-0.01872	0.08537	0.15637	-0.03023	1.00000	-0.14715	-0.12474	0.02134	-0.07365	-0.00465	0.11749	0.04198	-0.03596	-0.01512	-0.12989	0.07949
G1	-0.17432	0.26047	0.21750	0.26088	-0.38421	0.04879	-0.09450	-0.07405	-0.15565	-0.05165	-0.14715	1.00000	0.86498	-0.10411	-0.15713	-0.04723	-0.01525	0.03826	-0.06278	0.08012	0.34903	-0.07497
G2	-0.10712	0.26404	0.22514	0.24050	-0.38578	0.08959	-0.10668	-0.07947	-0.16485	-0.08218	-0.12474	0.86498	1.00000	-0.10401	-0.15460	-0.03889	-0.01869	0.03814	-0.03393	0.06715	0.33195	-0.09794
Sex	-0.04366	0.11913	0.08391	-0.20621	0.07389	0.08347	0.14631	0.05818	0.32078	0.13955	0.02134	-0.10411	-0.10401	1.00000	-0.02550	-0.09820	-0.06470	-0.12947	0.07930	0.12471	-0.05813	-0.11014
Address	0.02585	-0.19032	-0.14149	-0.06202	0.06382	0.03390	0.03665	-0.01548	0.01242	-0.00379	-0.07365	-0.15713	-0.15460	-0.02550	1.00000	0.04611	-0.09464	-0.00558	0.03048	0.00928	-0.07671	0.03094
Famsize	0.00247	0.01432	0.03954	0.01094	0.06607	-0.00464	0.02126	0.00431	-0.08196	-0.00245	-0.00465	-0.04723	-0.03889	-0.09820	0.04611	1.00000	-0.23961	0.03982	0.05025	0.01479	-0.00452	0.03294
Pstatus	0.00563	0.05717	0.03186	0.00875	0.00988	-0.05130	-0.03759	-0.03109	-0.07098	-0.01284	0.11749	-0.01525	-0.01869	-0.06470	-0.09464	-0.23961	1.00000	-0.01020	-0.01592	-0.10156	-0.02273	0.05383
Famsup	-0.10189	0.12049	0.13519	0.14351	-0.00698	0.01523	0.00376	0.01726	-0.06560	0.01880	0.04198	0.03826	0.03814	-0.12947	-0.00558	0.03982	-0.01020	1.00000	0.09430	-0.00743	0.08534	-0.02340
Paid	-0.00546	0.11397	0.09463	-0.00231	0.06942	0.03194	-0.04957	-0.00668	0.03568	0.06320	-0.03596	-0.06278	-0.03393	0.07930	0.03048	0.05025	-0.01592	0.09430	1.00000	0.06578	0.02411	-0.01831
Activities	-0.05428	0.11935	0.07970	0.07008	0.00056	0.05760	0.15033	0.08858	0.03282	0.01300	-0.01512	0.08012	0.06715	0.12471	0.00928	0.01479	-0.10156	-0.00743	0.06578	1.00000	0.04491	0.05752
Higher	-0.26550	0.21390	0.19174	0.18826	-0.30940	0.04824	-0.10262	-0.06911	-0.08433	0.01729	-0.12989	0.34903	0.33195	-0.05813	-0.07671	-0.00452	-0.02273	0.08534	0.02411	0.04491	1.00000	-0.09939
Romantic	0.17881	-0.03099	-0.06767	0.03304	0.06990	-0.04492	0.02711	-0.00052	-0.01997	-0.01802	0.07949	-0.07497	-0.09794	-0.11014	0.03094	0.03294	0.05383	-0.02340	-0.01831	0.05752	-0.09939	1.00000

La matrice des corrélations possède un grand nombre de d'informations.

On peut comprendre de ce tableau de la consommation d'alcool dépend beaucoup du sexe de l'élève et de si oui ou non l'étudiant sort avec ses amis.

On comprend également que la consommation d'alcool le week-end et le temps passé à étudier sont corrélés.

4. Analyse en Composante Principale

Pour trouver le modèle le plus adapté à nos données j'en ai essayé trois :

- L'élimination descendante
- L'élimination ascendante
- La sélection Stepwise

L'élimination descendante nous donne un R^2 de 0,3053.

L'élimination ascendante nous donne un R^2 de 0,3148.

La sélection stepwise nous donne un R^2 de 0,3105.

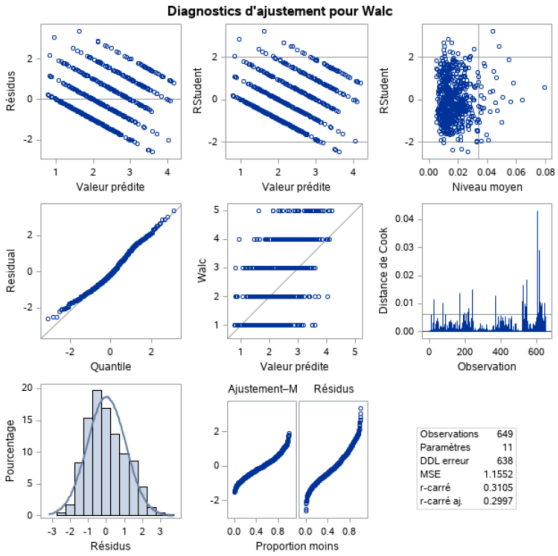
Même si l'élimination ascendante est celle qui a le R^2 le plus élevé, ce n'est pas celle qui que nous allons sélectionner car cette méthode conserve 13 variables, ce qui est trop.

J'ai donc choisi de conserver la sélection stepwise car son R^2 est correcte et qu'elle ne conserve que 10 variables.

Cependant, le R^2 n'est pas très élevé donc le modèle n'explique pas beaucoup de variance.

Synthèse de Sélection Stepwise

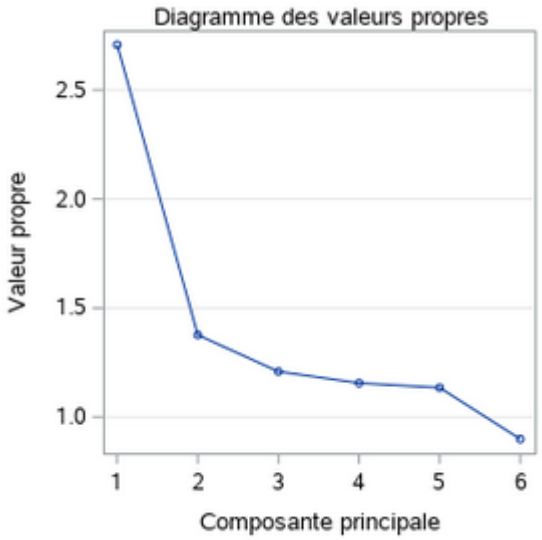
Etape	Variable entrée	Variable supprimée	Nombre var. dans	R carré partiel	R carré du modèle	C(p)	Valeur F	Pr > F
1	goout		1	0.1511	0.1511	133.736	115.14	<.0001
2	Sex		2	0.0892	0.2403	53.9034	75.86	<.0001
3	famrel		3	0.0234	0.2637	34.4638	20.47	<.0001
4	studytime		4	0.0157	0.2794	22.0508	14.04	0.0002
5	health		5	0.0085	0.2878	16.2987	7.63	0.0059
6	absences		6	0.0087	0.2965	10.3423	7.92	0.0051
7	Pstatus		7	0.0035	0.3000	9.1382	3.20	0.0742
8	Famsize		8	0.0053	0.3053	6.2727	4.89	0.0274
9	G2		9	0.0028	0.3081	5.7018	2.59	0.1081
10	freetime		10	0.0024	0.3105	5.4887	2.23	0.1356



On obtient alors cette matrice des corrélations :

Matrice de corrélation												
	Walc	goout	Sex	famrel	studytime	health	absences	Pstatus	Famsize	G2	freetime	numero_eleve
Walc	1.0000	0.3887	0.3208	-.0935	-.2149	0.1150	0.1564	-.0710	-.0820	-.1649	0.1202	0.9445
goout	0.3887	1.0000	0.0582	0.0897	-.0754	-.0157	0.0854	-.0311	0.0043	-.0795	0.3464	0.3468
Sex	0.3208	0.0582	1.0000	0.0835	-.2062	0.1395	0.0213	-.0647	-.0982	-.1040	0.1463	0.4770
famrel	-.0935	0.0897	0.0835	1.0000	-.0041	0.1096	-.0895	-.0513	-.0046	0.0896	0.1292	-.0727
studytime	-.2149	-.0754	-.2062	-.0041	1.0000	-.0564	-.1184	0.0087	0.0109	0.2405	-.0688	-.2332
health	0.1150	-.0157	0.1395	0.1096	-.0564	1.0000	-.0302	-.0126	-.0024	-.0822	0.0845	0.1246
absences	0.1564	0.0854	0.0213	-.0895	-.1184	-.0302	1.0000	0.1175	-.0046	-.1247	-.0187	0.1460
Pstatus	-.0710	-.0311	-.0647	-.0513	0.0087	-.0126	0.1175	1.0000	-.2396	-.0187	-.0376	-.0801
Famsize	-.0820	0.0043	-.0982	-.0046	0.0109	-.0024	-.0046	-.2396	1.0000	-.0389	0.0213	-.0909
G2	-.1649	-.0795	-.1040	0.0896	0.2405	-.0822	-.1247	-.0187	-.0389	1.0000	-.1067	-.1668
freetime	0.1202	0.3464	0.1463	0.1292	-.0688	0.0845	-.0187	-.0376	0.0213	-.1067	1.0000	0.1424
numero_eleve	0.9445	0.3468	0.4770	-.0727	-.2332	0.1246	0.1460	-.0801	-.0909	-.1668	0.1424	1.0000

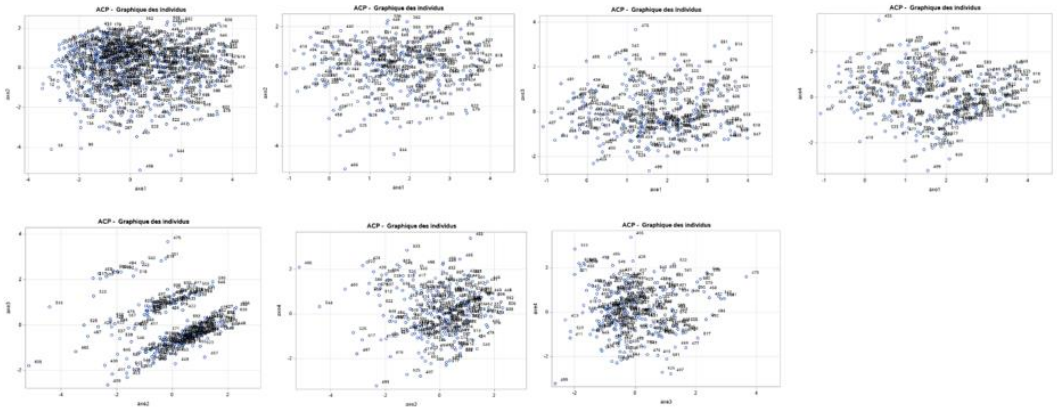
On voit que les variables les plus corrélées avec Walc sont gout, Sex et Studytime.



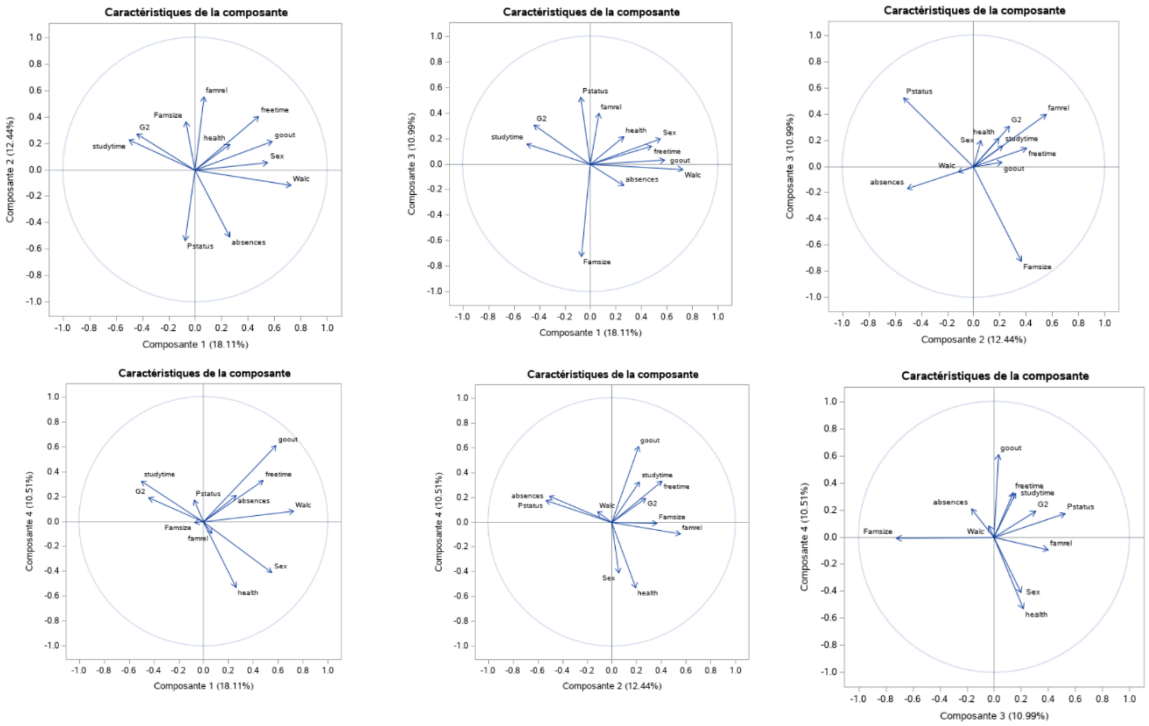
Avec le graphique « diagramme des valeurs propres » on peut voir que le coude est formé au deuxième axe mais pour le choix du nombre d'axes de l'ACP définitif, on prend ceux dont les valeurs propres sont supérieures à 1.

Valeurs propres de la matrice de corrélation				
	Valeur propre	Différence	Proportion	Cumulé
1	2.70802177	1.33153061	0.2257	0.2257
2	1.37649115	0.16791352	0.1147	0.3404
3	1.20857764	0.05257429	0.1007	0.4411
4	1.15600335	0.02138968	0.0963	0.5374
5	1.13461367	0.23584296	0.0946	0.6320
6	0.89877071		0.0749	0.7069

On choisit alors de garder les 4 premiers axes car le 5ème ne nous semble pas importer.
On a alors une valeur cumulée de 0,54%.



On voit que le premier graphique contient trop de données, il est illisible, on ne va alors conserver que les consommations de niveaux 3, 4 et 5.



A l'observation de ces graphiques nous pouvons dire que les variables Walc, Sex, goout, Freetime et absences ont un lien.

5 Conclusion

A travers ce travail j'ai pu constater que la consommation d'alcool chez les jeunes (15-22 ans) est influencée par différents facteurs tels que le sexe, les sorties entre amis, le temps libre, etc..

Mais cette consommation d'alcool a également des effets sur les études, en effet plus les étudiants consomment de l'alcool plus ils ont des absences, moins ils passent de temps à travailler sur leurs devoirs.

En revanche, on ne semble pas constater que la famille (Taille de la famille, niveau d'étude des parents, etc…) ait un lien avec une plus ou moins forte consommation.