M2 IMSD

Projet DataMining

Etudiants : Laurie Courtant, Céline Le Bras, Issam Badache, Laurent Magon, Karona Oum

15/02/2016

Table des matières

Introduction

Le Datamining a pour objectifs de comprendre l'information comprise dans une base de données et ainsi aider à la prise de décision. Il est apparu dans un contexte où la donnée est passée de la rareté à l'abondance, rendant difficile l'extraction des informations. Cette « fouille de données » utilise des techniques statistiques d'analyse et de modélisation afin de soutirer des informations exploitables au sein de la base de données. Il est utilisé dans de nombreux secteurs comme le marketing, la banque, l'assurance, l'éducation, la détection de fraude etc. Dans le secteur du marketing, par exemple, le data mining permet d'analyser le comportement de la clientèle, pour en tirer des profils et trouver ceux qui seront appétants à acheter un nouveau produit.

Le Data Mining est une aide à la décision dans le sens où elle permet à l'entreprise d'analyser son environnement, ses clients. Son impact est à la fois stratégique mais aussi opérationnel grâce à la statistique décisionnelle. Par exemple, il est possible de créer des modèles pour optimiser la transformation d'un lead en un client. Le data mining permet également de segmenter le marché afin de développer des offres adaptées aux clients. Il ne sert pas simplement à des fins d'analyse mais permet désormais la personnalisation en termes de démarche commerciale mais aussi dans la définition des offres personnalisées.

La démarche du Datamining se base sur deux principes : la description de notre jeu de données et la prédiction. C'est la statistique qui permet l'analyse descriptive de nos données, mais surtout la modélisation de scoring et de prédictions. La puissance de la statistique est complémentaire et devient même principale face au requêtage.

En production le Datamining se base sur une dynamique de couple entre la compréhension métier et la compréhension des données. La compréhension métier permet de poser les conditions et les axes de recherche pour la fouille des données alors que la compréhension des données nécessite de suivre un processus spécifique. C'est-à-dire la préparation des données, la modélisation et l'évaluation statistique couplée avec une évaluation basée sur l'aspect métier. Il existe deux approches. Celle de prouver une intuition et celle de découvrir une information cachée. Par exemple, le site de rencontres Meetic a constaté une augmentation des connexions sur son application mobile lorsque le trafic routier est dense. Intuitivement, le lien entre ces deux phénomènes n'était pas évident.

Dans ce projet, notre objectif sera l'exploration des données de notre base, afin de pouvoir en présenter une analyse descriptive. Notre base provient d'une enquête en ligne sur la consommation de boissons en Europe. Nous poursuivrons notre exploration en créant une classification des individus interrogés en fonction de leur consommation, de leur comportement et de leur profil de consommateur, afin d'en dégager un profil général.

I°) Présentation de notre base

Nous avons choisi d'analyser une base de données sur la consommation de boissons. Cette dernière provient d'un sondage présent sur le logiciel Sphinx Campus. C'est une enquête en ligne, composée de 63 questions de différents types (ouverte, fermée). La base a été est alimentée par la diffusion du questionnaire via une campagne e-mailing dans différents pays européens.

1) Le questionnaire

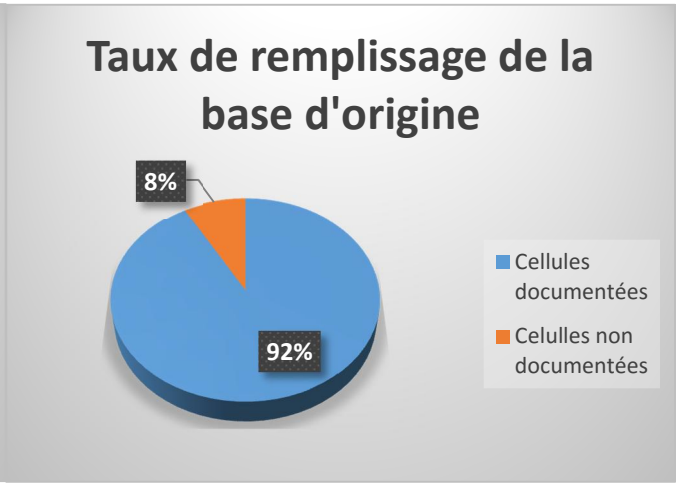
Cette enquête en ligne menée en 2010 en Europe par 30 universités pour « the Corberen Project » visait à étudier les corrélations entre la culture et la consommation. Nous n'effectuerons pas la même démarche : notre but est la détermination de profils généraux en fonction de 3 profils spécifiques qui sont :

1. Profil consommateur
2. Profil consommation
3. Profil comportement

La base de données a été alimentée par la diffusion du questionnaire via une campagne e-mailing effectuée par un insitut. Ce dernier a envoyé un mail à une liste de contact fournie par les universitaires qui a ensuite été diffusé par effet *boule neige*. Le fait que le travail ait été réalisé par un institut garantit une certaine qualité des données. Les pays concernés sont l'Allemagne, l'Espagne, la France, le Royaume-Uni, l'Italie. A noter que les poids démographiques ne sont pas respectés dans notre base, les populations de chaque pays dans la base ne sont en effet pas représentatives de la population réelle des pays.

Notre questionnaire contient 3082 observations et 63 variables. Le questionnaire en provenance de Sphinx comprend 15 questions de types :

Type de question	**Typologie de réponse**	
Fermée	Simple	9
	Echelle	2
	Numérique	1
	Multiple	1
	Image	1
Ouverte	Texte	1
Nb de questions		*15*



2) Nos variables

a) Présélection de variables

Après exportation de la base sous format Excel, nous avons pu constater que nous avions non pas 63 variables comme prévu mais 73 variables. Le surplus provient de variables ajoutées par le prestataire pour son traitement de données interne comme la date de création, son ID internet etc. Nous avons choisi de les éliminer afin de retrouver les variables dans le questionnaire d'origine (d'autant plus que certaines de ces variables étaient vides).

Puis nous avons fait un second nettoyage de notre base de données en enlevant :

- Les variables sur les images – le sondé devait choisir une combinaison de 3 images sur 36. Le but étant d'évaluer le sentiment associé à l'action de boire. Nous avons choisi de ne pas les traiter, car elles étaient très complexes à exploiter sur nos outils dans le temps imparti.
- Certaines variables sur les réponses libres – nous les avons aussi écartées car elles se réfèrent à la question portant sur les images. Seule la variable AlcMarq portant sur la marque de boisson alcoolisée préférée subsiste, car indépendante des autres.

b) Table avec les variables et les modalités

Après ce premier tri, il nous reste deux types de variables.

- Les variables quantitatives : La consommation et l'âge

Variables et modalités	Description	Type de variable	Groupe
ConsVin	Consommation moyenne de verres de vin/semaine (12cl)	Quantitative	Consommation
ConsBier	Consommation moyenne de verres de bière/semaine (25cl)	Quantitative	Consommation
ConsEau	Consommation moyenne de bouteille d'eau minérale/semaine (1L)	Quantitative	Consommation
ConsGaz	Consommation moyenne de boissons gazeuses/semaine (25cl)	Quantitative	Consommation
ConsSpir	Consommation moyenne de spiritueux/semaine (12cl)	Quantitative	Consommation
ConsCafe	Consommation moyenne de tasses de café/semaine	Quantitative	Consommation
ConsThe	Consommation moyenne de tasses de thé/semaine	Quantitative	Consommation
Age	Age de l'individu	Quantitative	Consommateur

Chaque variable de consommation est associée à un type de boisson. Elles permettent de connaître les quantités de boissons consommées de façon hebdomadaire pour chaque individu.

L'âge est quant à lui important car c'est une donnée du sondé qui permet une meilleure connaissance de celui-ci et aidera à la segmentation.

- Les variables qualitatives : le budget, les associations, l'alcool préféré, la marque préférée, la raison de boire (sa boisson préférée), le sexe, le niveau d'éducation, la situation professionnelle, le revenu net mensuel du ménage et le pays.

  Ces variables vont nous permettre de mieux connaitre nos sondés afin de faire ressortir leurs caractéristiques sociales. Par exemple, comprendre l'importance de la boisson dans son mode de vie, s'il y a un lien entre un type de boisson et une situation professionnelle particulière, ou le sexe etc.

  La question sur la marque de la boisson préférée est ouverte, sa variable contient donc du texte. Nous effectuerons ultérieurement une analyse de Textmining afin de visualiser les marques préférés de nos sondés.

  L'ensemble de nos variables se trouve en ANNEXE X. [Pour votre compréhension, vous trouverez un dictionnaire provisoire en PJ]

II°) Statistiques descriptives de la base initiale

Les statistiques descriptives sur les bases initiales permettent d'avoir un aperçu visuel de notre base, notamment de la distribution des variables. Elles peuvent mettre en évidence des défauts de la base, comme la présence d'outliers (aberrants ou extrêmes), des valeurs manquantes, ou des comportements illogiques (par exemple, un consommateur qui dépense 100 euros par semaine en café sans boire une seule tasse). Dans cette partie, nous verrons donc comment sont distribuées nos variables qualitatives et quantitatives.

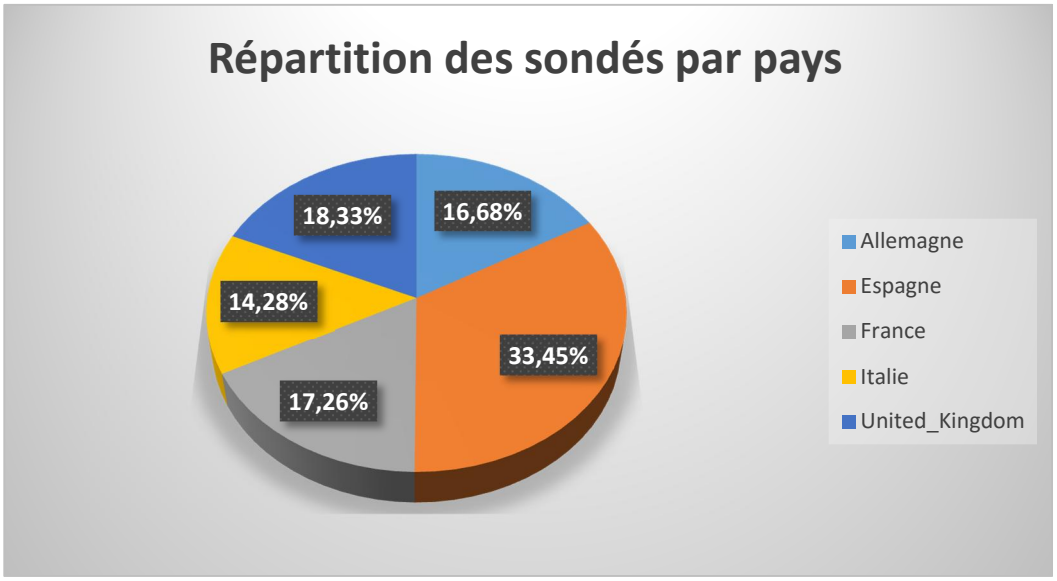
1) Variables qualitatives

Tri à plat :

Sexe				
Sexe	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
Femme	1555	50.45	1555	50.45
Homme	1527	49.55	3082	100.00

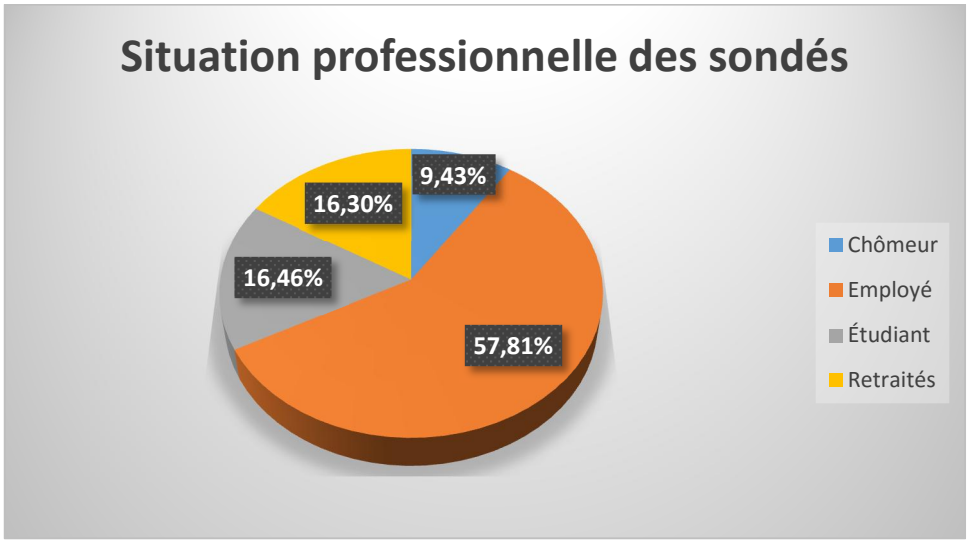
Notre variable sexe contient 50,45% de femmes contre 49,55% d'hommes. Les modalités sont donc réparties de façon homogène. Elle ne contient pas de valeurs manquantes.

Diagrammes circulaires :



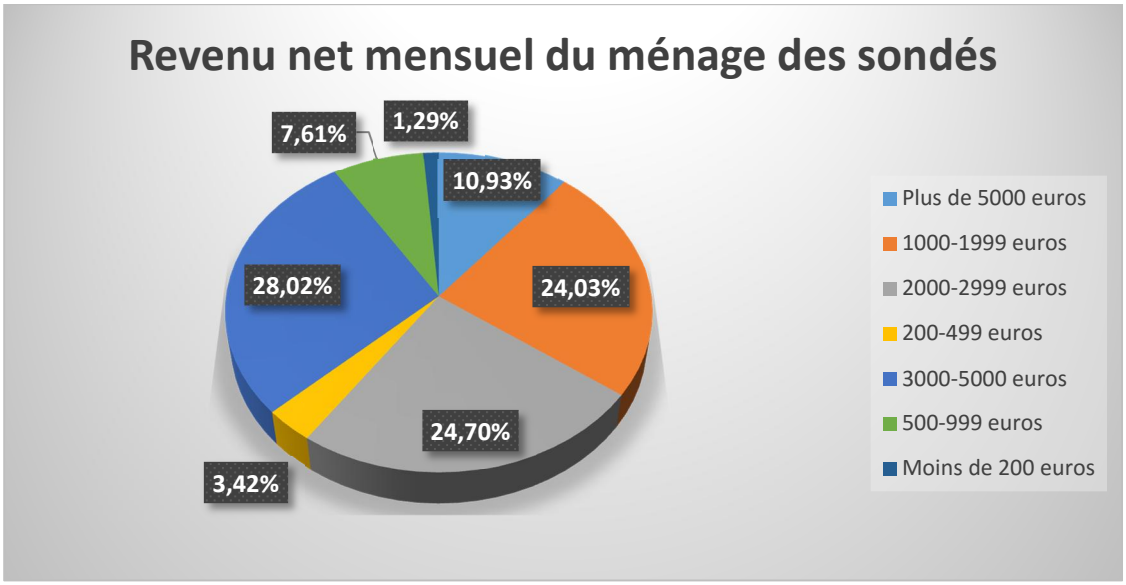
Comme explicité dans la partie I, la répartition des pays ne correspond pas aux poids démographiques réels de chaque pays mais à la répartition des sondés de notre base originale. Cinq pays sont

représentés, tous en Europe de l'Ouest. On remarque que l'Espagne est la plus représentée avec 33,45% de nos sondés, les autres pays ayant une part similaire tournant autour de 16%.

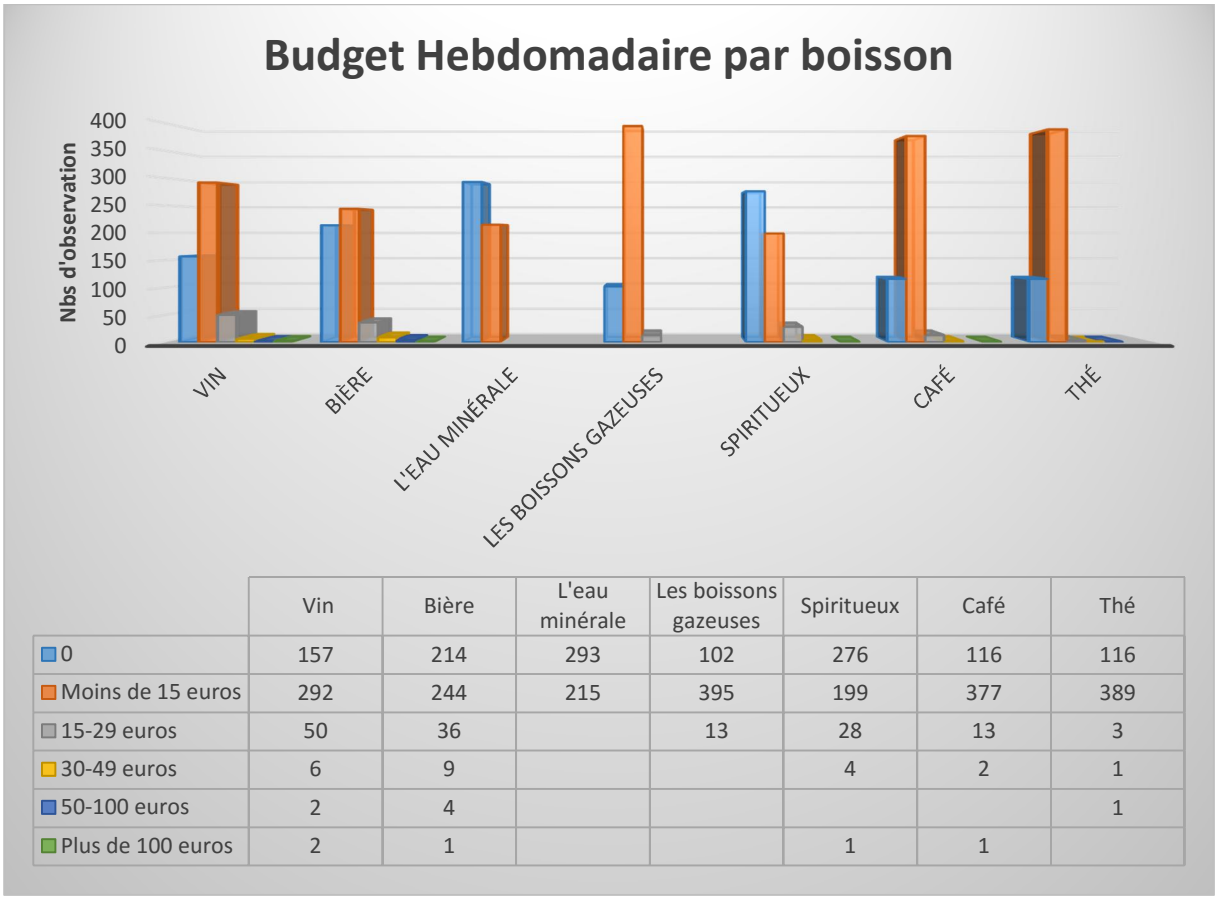


Ce diagramme permet de voir la répartition de la situation professionnelle de nos sondés présents dans la base. La majorité de nos sondés sont des employés. Nous avons à peu près autant d'étudiants que de retraités (environ 16%), et 9,43% de chômeurs.

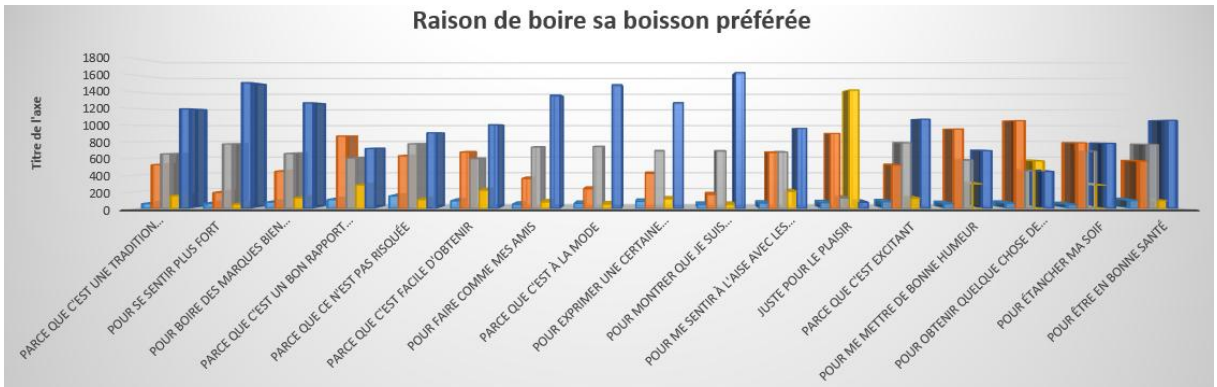
Sur ce point, la base est assez représentative de la réalité. En effet, les pourcentages du chômage d'Eurostat pour ces cinq pays est de 10,38% en 2009 et de 10,48% en 2011.



Diagrammes en bâtons :



	Vin	Bière	L'eau minérale	Les boissons gazeuses	Spiritueux	Café	Thé
0	157	214	293	102	276	116	116
Moins de 15 euros	292	244	215	395	199	377	389
15-29 euros	50	36		13	28	13	3
30-49 euros	6	9			4	2	1
50-100 euros	2	4					1
Plus de 100 euros	2	1			1	1	



	Parce que c'est une tradition familiale	Pour se sentir plus fort	Pour boire des marques bien connues	Parce que c'est un bon rapport qualité/prix	Parce que ce n'est pas risquée	Parce que c'est facile d'obtenir	Pour faire comme mes amis	Parce que c'est à la mode	Pour exprimer une certaine valeur personnelle	Pour montrer que je suis différent des autres gens	Pour me sentir à l'aise avec les autres	Juste pour le plaisir	Parce que c'est excitant	Pour me mettre de bonne humeur	Pour obtenir quelque chose de bien pour moi	Pour étancher ma soif	Pour être en bonne santé
Je ne sais pas	58	60	75	106	149	91	50	61	85	49	64	68	80	58	62	45	95
Je suis plutôt d'accord	525	192	446	870	632	677	359	237	421	167	672	902	520	954	1055	788	567
Je suis plutôt pas d'accord	654	775	660	605	776	596	737	743	691	689	677	125	789	581	452	686	769
Je suis totalement d'accord	153	50	126	281	107	219	78	53	116	49	204	1423	122	295	571	280	93
Je suis totalement en désaccord	1197	1511	1270	719	909	1006	1362	1488	1271	1632	961	74	1072	694	445	782	1059

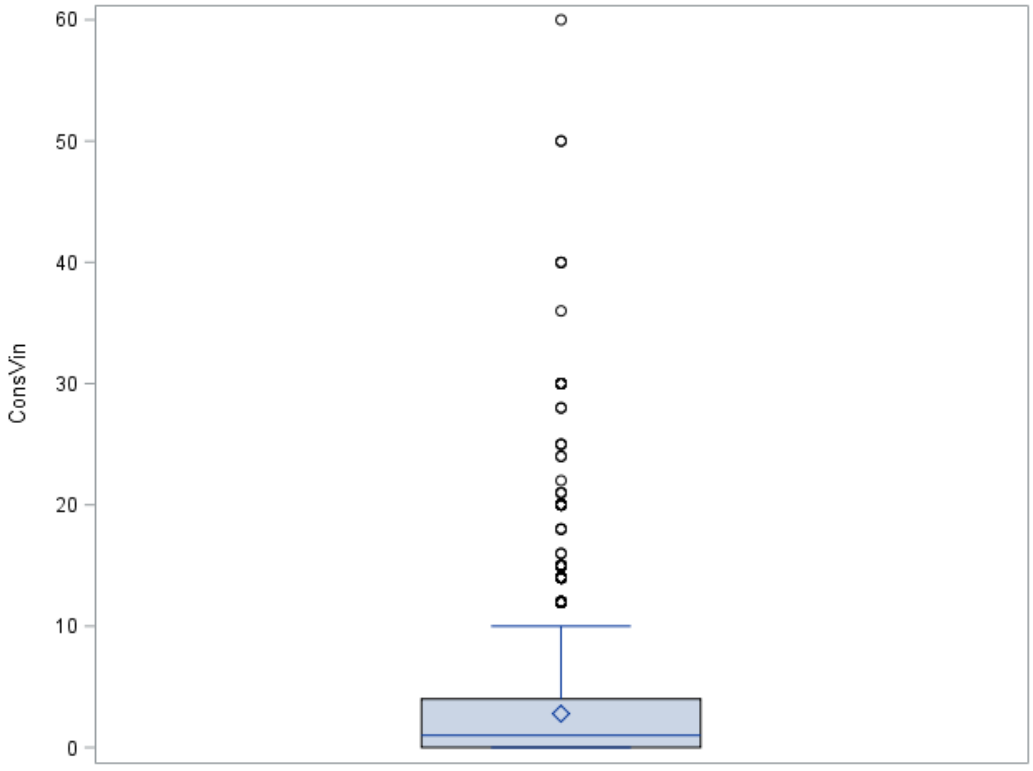
2) Variables quantitatives

BoxPlot :

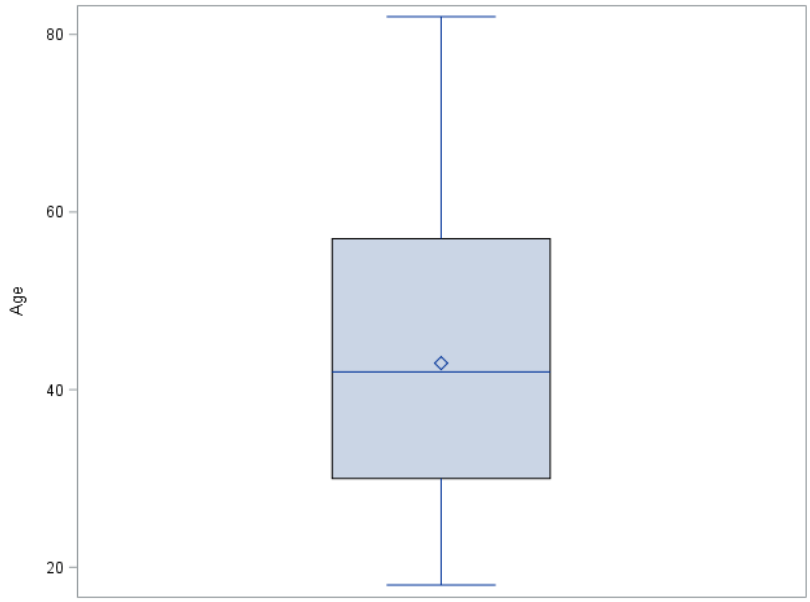
Les boxplot, ou boîtes à moustache, sont des représentations schématiques de la distribution des variables. Elles permettent de visualiser rapidement le profil d'une série statistique quantitative.

L'interprétation se fait comme suit :

- La longueur de la boîte détermine l'étendue de la partie centrale de la distribution.
- La ligne qui la « ceinture » représente la médiane.
- Les extrémités de la boîte représentent les premier et troisième quartiles.
- Les extrémités des moustaches représentent les valeurs minimum et maximum.
- Les outliers, s'il y en a, sont représentés par des ronds en dehors des limites des moustaches. Dans la mesure du possible, il vaut mieux les traiter pour ne pas fausser les statistiques faites sur les variables.
- Le losange indique la moyenne.

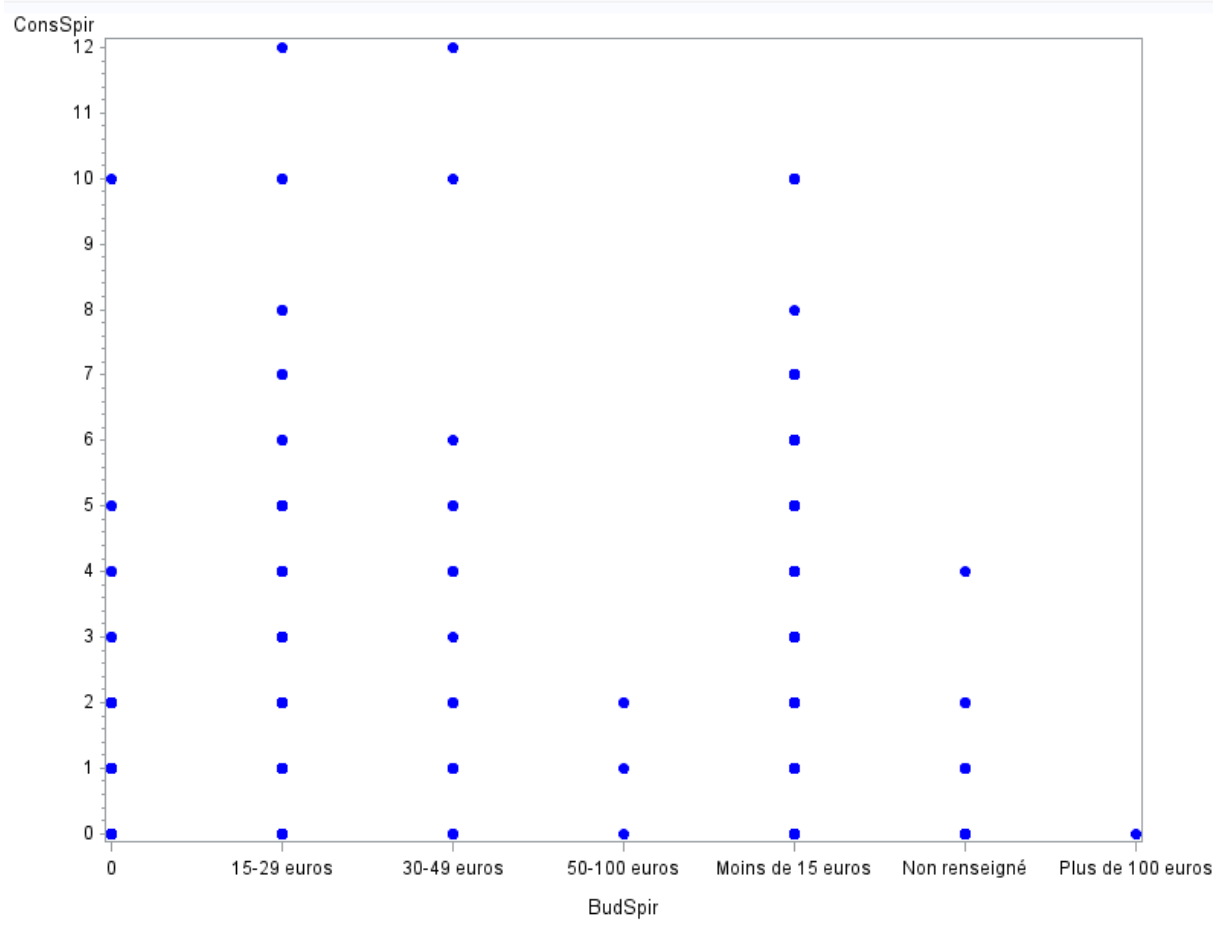


La variable ConsVin ci-dessus contient 16 valeurs aberrantes. La partie centrale de la distribution se situe entre 0 et 4 verres de vin, la médiane semble être à 1 et la moyenne à 3. Les autres variables de type ConsXX ont une boxplot similaire (**ANNEXE 1**).



La variable Age a une partie centrale plus étalée, l'âge de nos individus se situe globalement entre 30 et 58 ans. L'âge moyen est de 43, l'âge médian de 42. L'âge minimum est de 18 et le maximum de 82 ans.

Graphiques :



Ce graphique, représentant la consommation hebdomadaire de verres de spiritueux par rapport aux dépenses faites sur cette boisson, fait ressortir plusieurs points à première vue illogiques. Si les

individus ayant répondu qu'ils consommaient plusieurs verres de spiritueux sans utiliser leur budget peuvent avoir un sens (on peut en effet imaginer qu'ils se font inviter par des amis/de la famille), ceux qui dépensent de l'argent dans des boissons non consommées sont étranges. Peut-être que ces individus font des achats pour leur famille ? Peut-être font-ils des réserves en prévision d'une potentielle fin du monde ? Si ce comportement se retrouve sur plusieurs individus, il sera préférable de garder les observations concernées. En revanche, s'il est exceptionnel, nous devront supprimer ces individus.

Ce tour de la base a permis de mettre en évidence des données qu'il faudra traiter, ce qui fait l'objet de notre seconde partie.

III°) Traitement de la base

Avec cette nouvelle base que nous appelons Boisson propre sphinx, nous passons à une deuxième phase de préparation de nos données sur sas

- Les individus ont énormément de valeur manquante.
- Personne qui n'a pas répondu aux variables budgets ET variables consommations

1. Traitement des valeurs manquantes

Les valeurs manquantes sont les valeurs d'une observation n'ayant pas été renseignées. Il est indispensable de les traiter, notamment pour pouvoir faire une ACP, ou tout simplement parce que qu'un bon nombre d'outils d'analyse ne les prend pas en charge.

Selon le type de données manquantes et le contexte, le traitement à effectuer sera différent. D'où l'intérêt d'une analyse exploratoire des variables manquantes. Tout en gardant à l'esprit que tout choix aura un impact sur nos analyses, il est possible de :

- ✓ **Supprimer** les observations concernées si elles sont peu nombreuses, ou si leur présence laisse supposer un défaut de conception de la variable (ex : présence de valeurs manquantes sur une variable dont les observations sont supposées être relevées de façon automatique)
- ✓ **Ne pas utiliser** une variable contenant des valeurs manquantes.
- ✓ Si la donnée manquante est source d'information, elle peut être conservée, par exemple en la **remplaçant par** 0 (données quantitatives) ou NA/Inc… (données qualitatives). Elle est alors traitée comme une valeur à part entière.
- ✓ **Imputer les valeurs manquantes** en les remplaçant par une valeur par défaut, ou déduite des valeurs des autres variables (moyenne de la variable, médiane,…). Il ne faut pas négliger le fait que cette méthode est loin d'être neutre.

Dans notre base, il semblerait que certains sondés n'aient pas eu envie de répondre à des questions car elles demandaient trop de réflexion. D'autres fois, ce sont les questions qui ont pu dérouter les sondés, car ambigües ou difficiles à comprendre (ex : Raibien : Raison de boire sa boisson préférée : pour obtenir quelque chose de bien pour soi). Ou encore, ils se lassent du questionnaire et finissent par ne plus y répondre.

La présence de toutes ces variables manquantes diminuent la qualité de notre base. Cette dernière comportant initialement 3082 observations, nous pouvons nous permettre de retirer les individus n'ayant pas répondu à un/plusieurs questions, à condition qu'il en reste un nombre suffisant à la fin.

Ainsi, les individus n'ayant pas répondu aux variables suivantes ont été supprimés de la base : BudXX, RaiXX, AlcLieu, AlcAch. En effet, leur présence nous empêche de déterminer les profils que l'on souhaite.

De même pour ceux n'ayant pas répondu à une question de ConsX et de BudX simultanément. Par exemple, non-réponse sur les variables ConsVin (=Nombre de verres de vin par semaine) et BudVin

(Dépenses hebdomadaires en vin). Ainsi que ceux n'ayant répondu à aucune question de la série des AssoXX.

Nous avons réservé le même traitement aux individus ayant répondu « Je ne sais pas » aux questions sur les RaiXX (= Raison de boire sa boisson préférée), car cette réponse n'apporte pas d'information pouvant nous permettre de les classifier.

La variable AlcMarq est une variable qualitative à réponse libre. Elle représente la marque de la boisson alcoolisée préférée. Nous avons remplacé les valeurs manquantes par «Non renseigné »

Remplacement des valeurs manquantes de Cons*** par la moyenne des Cons*** ayant la même modalité que Bud

BudVin				
BudVin	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
0	482	26.20	482	26.20
15-29 euros	227	12.34	709	38.53
30-49 euros	34	1.85	743	40.38
50-100 euros	7	0.38	750	40.76
Moins de 15 euros	1085	58.97	1835	99.73
Plus de 100 euros	5	0.27	1840	100.00

Nous avons lancé des Proc Mean afin de visualiser la relation entre la consommation d'une boisson et sa part de budget dans un ménage.

Voici par exemple la relation entre la consommation de vin et son budget dans un ménage.

[Inclure un tri à plat avec avant remplacement par la moyenne de chaque modalité et après]

Par contre pour les valeurs manquantes, nous avons décidé de remplacer les valeurs manquantes par la moyenne d'une consommation ou de budget pour chaque boisson. Les individus ayant une valeur manquante à la fois sur les consommations et les budgets seront écartés de notre étude.

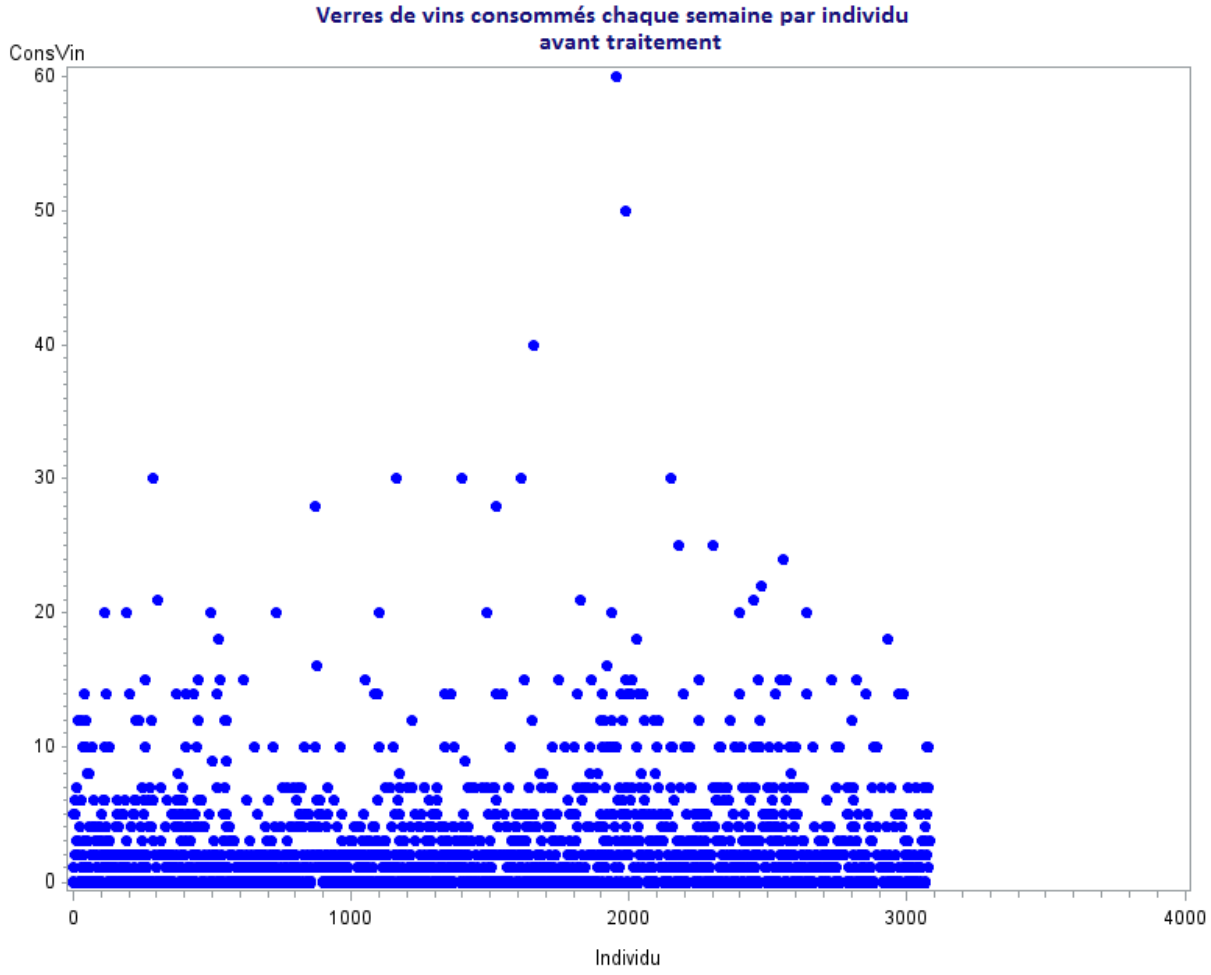
2. Traitement des outliers

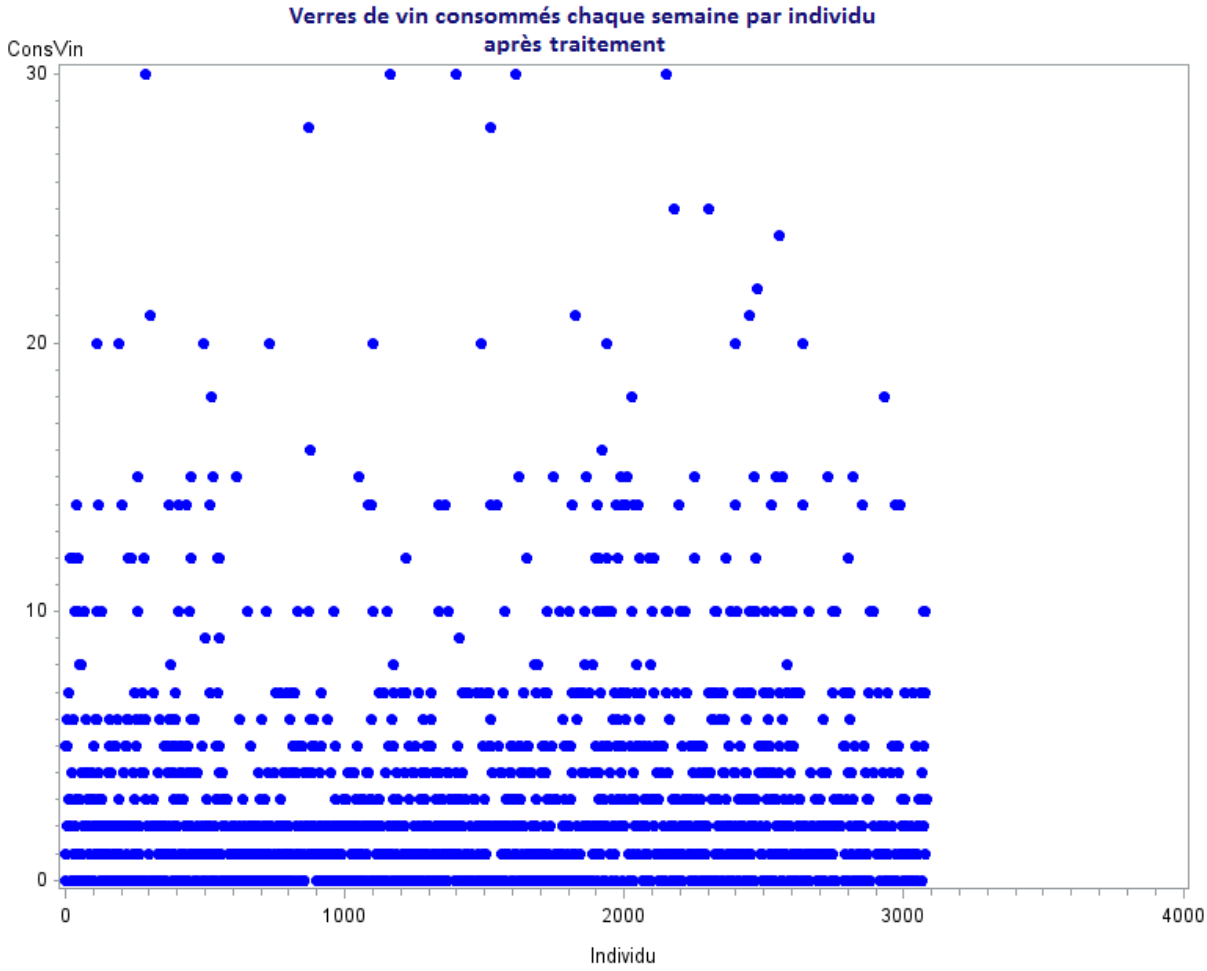
Un outlier est une valeur d'une observation anormalement éloignée de celles des autres. Il a un impact sur la formation des clusters. Parmi les outliers, on distingue les valeurs extrêmes et les valeurs aberrantes. Les outliers ne concernent que les variables quantitatives.

Les premiers désignent les valeurs sortant de l'ordinaire, mais qui sont exactes. Par exemple, un donateur régulier venant de gagner au loto pourrait donner une somme inhabituellement élevée en don aux associations. Les seconds sont des valeurs erronées, souvent dues à des erreurs humaines telles que l'ajout accidentel d'un 0 supplémentaire.

Il existe différents traitement pour les outliers, tels que la suppression, l'isolement ou l'imputation. Si les données sont assez nombreuses, on choisira souvent de supprimer les outliers. Pour les valeurs extrêmes (donc non aberrantes), on peut envisager de les isoler, c'est-à-dire les placer en illustratives, afin qu'elles restent visibles sans fausser les statistiques.

Dans notre base, nous avons 8 variables quantitatives : les 7 variables de consommation, ainsi que l'âge. Cette dernière ne contient pas de valeurs aberrantes ou extrêmes, il y a en effet plusieurs observations dans toutes les tranches d'âges. En revanche, à part pour la consommation d'eau, toutes les autres variables de consommation sont concernées par les outliers, que nous illustrons ci-dessous avec la variable ConsVin avant et après traitement :





Ainsi, nous avons retiré toutes les observations dont les consommations hebdomadaires dépassaient un des cas suivant :

- 30 verres de vin
- 40 verres de bière
- 36 verres de boissons gazeuses
- 19 verres de spiritueux
- 50 tasses de thé
- 50 tasses de café

On pourrait se dire que la consommation de 50 tasses de café soit faisable, et qu'il serait dommage de retirer des individus avec ce comportement car, même si ce dernier est atypique, il est réaliste. Le problème vient justement du fait de sa rareté, et la présence d'un simple individu avec un tel comportement peut modifier tout le clustering et fausser nos interprétations, ce qui explique notre choix de suppression.

3. Traitement des modalités de variables

a) Binarisation

Le questionnaire comprend des questions à réponses multiples. Afin de conserver toute l'information qu'elles contiennent, nous avons binarisé les modalités de chacune de ces questions. C'est le cas pour les 15 questions où les sondés choisissent les boissons qu'ils associent à une situation particulière. Dans notre cas, si le sondé choisi une boisson, alors nous aurons « 1 », sinon « 0 ». Au final, pour cette question nous avons créé 126 variables (14 associations * 9 boissons).

Il en est de même pour la question sur les lieux d'achat (AlcAchxx) des boissons alcoolisées, entraînant la création de 7 variables (1 question*7 lieux d'achats possibles) décrites ci-dessous :

Procédure CONTENTS

Liste alphabétique des variables et des attributs						
#	Variable	Type	Long.	Format	Informat	Libellé
33	AlcAch	Texte	107	$107.	$107.	AlcAch
195	AlcAchAut	Num.	8			
192	AlcAchBar	Num.	8			
194	AlcAchDis	Num.	8			
189	AlcAchMag	Num.	8			
190	AlcAchProx	Num.	8			
193	AlcAchRes	Num.	8			
191	AlcAchSpe	Num.	8			
32	AlcLieu	Texte	83	$83.	$83.	AlcLieu

Pareillement pour le lieu de consommation (AlcLieuXX) où nous avons créé 6 variables (1 question * 6 lieux de consommation possibles), visibles ci-dessous :

Liste alphabétique des variables et des attributs						
#	Variable	Type	Long.	Format	Informat	Libellé
32	AlcLieu	Texte	83	$83.	$83.	AlcLieu
188	AlcLieuAut	Num.	8			
187	AlcLieuBar	Num.	8			
183	AlcLieuMai	Num.	8			
186	AlcLieuRes	Num.	8			
184	AlcLieuTrav	Num.	8			
185	AlcLieuVoy	Num.	8			

b) Regroupement de modalités

Regroupement des modalités de revenu :

Afin de réduire le nombre de modalités, nous avons choisi de regrouper les modalités de la question sur le revenu. En effet nous avions 5 modalités de réponse (+ une modalité vide pour les non-réponses) :

- 'Moins de 200 euros' ➔'Moins de 1000 euros';
- '200-499 euros' ➔'Moins de 1000 euros';
- '500-999 euros' ➔ 'Moins de 1000 euros';
- '2000-2999 euros' ➔'2000-5000 euros';
- '3000-5000 euros'➔'2000-5000 euros';
- Valeur manquante ➔ 'inc';

Regroupement des modalités de RaiXX :

Nous avons également choisi de regrouper les modalités pour les 17 questions « Raison de boire votre boisson préférée », où le sondé doit donner un degré d'accord sur la raison proposée. Nous passons de 4 modalités à 2 modalités par type de raison.

- "Je suis totalement en désaccord"➔ "Non";
- "Je suis plutôt pas d'accord" ➔ "Non";
- "Je suis plutôt d'accord" ➔ "Oui";
- "Je suis totalement d'accord" ➔"Oui";

Changement de nom et regroupement des modalités :

Afin de faciliter la lecture sur les graphes et l'utilisation dans nos codes de la multitude de variables créées, nous avons renommé et groupé les modalités de variables suivantes :

Ci-dessous la sortie SAS correspondante avant renommage :

Procédure MEANS

Variable d'analyse : ConsBier ConsBier						
BudBier	N Obs	N	Moyenne	Ecart-type	Minimum	Maximum
0	565	512	0.1269531	0.7773054	0	16.0000000
15-29 euros	147	144	8.1944444	8.2481426	0	50.0000000
30-49 euros	32	31	13.2580645	11.5324694	2.0000000	50.0000000
50-100 euros	5	5	35.2000000	16.8878655	20.0000000	56.0000000
Moins de 15 euros	1089	1042	2.9222649	3.8230183	0	60.0000000
Plus de 100 euros	2	2	2.5000000	3.5355339	0	5.0000000

Après traitement, on obtient :

BudBier	*Nb Obs*	*Renommage*
0	565	BT1
Moins de 15€	1089	BT2
15-29€	147	BT3
30-49€	32	BT3
50-100€	5	BT3
Plus de 100€	2	BT3

Il en est de même pour les variables suivantes AlcPref, Educ, SituProf, RevNet et Pays.

c) Discrétisation des variables quantitatives

***Discrétisation des variables de consommation ConsXX* :**

La discrétisation consiste à transformer une variable quantitative en variable qualitative. Afin de discrétiser nos questions se rapportant à des boissons alcoolisées, nous nous sommes basés sur une échelle de consommation d'alcool éditée par le ministère de la santé, qui conseille de ne pas dépasser 2 verres d'alcool par jour pour les femmes et 3 verres pour les hommes, soit respectivement 14 et 21 verres par semaine. Dans notre questionnaire, la quantité d'alcool est identique pour toutes les boissons car chaque verre contient un volume différent de liquide (par exemple, 12 cl de vin et 25 cl de bière contiennent la même quantité d'alcool). Considérant que les hommes et les femmes n'ont pas la même tolérance à l'alcool, nous avons choisi de faire la moyenne des consommations entre les genres afin de pouvoir traiter notre base par individu et non pas par sexe.

Les variables de consommation de thé et de boissons gazeuses ont été discrétisées en fonction de la répartition des réponses sur ces questions.

Procédure MEANS

Variable d'analyse : ConsThe ConsThe						
constheq	N Obs	N	Moyenne	Ecart-type	Minimum	Maximum
T0	629	629	0	0	0	0
T1	788	788	4.4035533	2.4839178	1.0000000	9.0000000
T2	404	404	18.5742574	8.6939659	10.0000000	50.0000000

Ici la consommation de thé est discrétisée. Elle est répartie en 3 types de consommations ayant comme intervalle :

- Aucune consommation = T0, qui représente 629 individus
- Une consommation de 1 à 9 = T1, qui représente 788 individus
- Une consommation de 10 à 50 = T2, qui représente 404 individus

Enfin, la consommation de l'eau a été discrétisée en fonction des quantités recommandées par jour.

[Dans la conclusion de la partie du traitement de la base : Si on a le temps, refaire le diagramme ci dessous en comptant le nombre de cases non renseignées par rapport aux cases renseignées. But : Montrer que l'on a amélioré la qualité de la base sur le plan des valeurs manquantes.]

d) Traitement des variables

-Non complété-

Ont été supprimée :

- les variables RaiBien (= pour obtenir quelque chose de bien pour soi),… car peu claires et prêtant à confusion.

- Les variables AssoSport (car trop corrélé à AssoSoif)

IV°) Modélisation : Clustering et prédiction

1) Modélisation

a. Profil consommation

Ce profil a pour but de définir leurs habitudes de consommation. C'est-à-dire la quantité de consommation de telle ou telle boisson, leur boisson alcoolisée préférée et où ils consomment leur boisson préférée.

Pour cela nous avons donc utilisé pour ce type de profil les variables :

- Cons*** : Dans une semaine normale, quelle est votre consommation des boissons suivantes ?
- Alcpref : Quelle est la boisson alcoolisée que vous préférez ?
- Alclieu : où allez-vous souvent boire votre boisson préférée ?

Pour déterminer ce profil nous allons passer par plusieurs étapes :

- Une ACM pour nos variables qualitatives Alcpref et Alclieu afin de pouvoir en récupérer leurs coordonnées sur les axes
- Une Kmeans où l'on inclut les axes de l'ACM et nos variables quantitatives Cons***. Cette Kmeans nous permet de faire une première segmentation de notre profil et nous permet d'observer le R^2
- Une CAH où l'on utilisera les clusters donnés pas notre Kmeans. La classification ascendante étant assez longue, c'est pourquoi nous avons décidé de réduire ce temps en passant par une Kmeans.
- Caractérisation de nos clusters pour voir les différentes distributions de nos individus dans tel ou tel cluster

Voici notre résultat :

Description of "Cluster_HAC_2"

Cluster_HAC_2=c_hac_1			
Examples			[15,2 %] 264
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
ConsThe	33,09	21,28 (8,91)	5,94 (8,18)
ConsSpir	-0,56	0,71 (1,23)	0,75 (1,41)
ConsVin	-0,75	2,86 (3,39)	3,03 (4,05)
ConsGaz	-1,65	2,80 (3,71)	3,25 (4,77)
ConsBier	-2,22	2,04 (3,35)	2,55 (4,12)
ConsEau	-6,16	3,44 (3,87)	5,08 (4,71)
ConsCafe	-6,93	6,34 (7,61)	9,68 (8,52)
Discrete attributes : [Recall] Accuracy			
c2d_AlcLieuBar_1=_1_0,00	3,43	[17,4 %] 72,3 %	63,0 %
AlcPref=AlP3	2,85	[17,9 %] 52,7 %	44,6 %
c2d_AlcLieuMai_1=_2_1,00	2,58	[16,8 %] 72,0 %	65,0 %
AlcPref=AlP2	2,04	[22,6 %] 8,0 %	5,3 %
c2d_AlcLieuVoy_1=_2_1,00	1,42	[21,7 %] 4,9 %	3,5 %
AlcPref=AlP4	0,61	[17,3 %] 6,4 %	5,6 %
c2d_AlcLieuRes_1=_2_1,00	0,41	[15,7 %] 31,4 %	30,4 %
c2d_AlcLieuTrav_1=_1_0,00	0,09	[15,2 %] 99,2 %	99,2 %
c2d_AlcLieuTrav_1=_2_1,00	-0,09	[14,3 %] 0,8 %	0,8 %
c2d_AlcLieuRes_1=_1_0,00	-0,41	[14,9 %] 68,6 %	69,6 %
c2d_AlcLieuVoy_1=_1_0,00	-1,42	[14,9 %] 95,1 %	96,5 %
AlcPref=AlP1	-2,48	[12,2 %] 27,3 %	33,9 %
c2d_AlcLieuMai_1=_1_0,00	-2,58	[12,2 %] 28,0 %	35,0 %
AlcPref=AlP5	-2,76	[8,2 %] 5,7 %	10,5 %
c2d_AlcLieuBar_1=_2_1,00	-3,43	[11,3 %] 27,7 %	37,0 %

Cluster_HAC_2=c_hac_2			
Examples			[53,1 %] 923
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
ConsGaz	1,00	3,35 (5,34)	3,25 (4,77)
ConsEau	-2,88	4,78 (3,73)	5,08 (4,71)
ConsSpir	-4,23	0,62 (1,18)	0,75 (1,41)
ConsBier	-6,92	1,91 (2,51)	2,55 (4,12)
ConsVin	-11,26	2,00 (2,32)	3,03 (4,05)
ConsThe	-14,43	3,27 (3,72)	5,94 (8,18)
ConsCafe	-20,21	5,80 (4,35)	9,68 (8,52)
Discrete attributes : [Recall] Accuracy			
c2d_AlcLieuMai_1=_1_0,00	6,12	[63,1 %] 41,6 %	35,0 %
AlcPref=AlP5	5,87	[73,6 %] 14,5 %	10,5 %
c2d_AlcLieuBar_1=_2_1,00	5,29	[61,3 %] 42,8 %	37,0 %
c2d_AlcLieuTrav_1=_2_1,00	2,46	[85,7 %] 1,3 %	0,8 %
AlcPref=AlP1	1,81	[56,1 %] 35,9 %	33,9 %
AlcPref=AlP4	1,25	[59,2 %] 6,3 %	5,6 %
c2d_AlcLieuVoy_1=_1_0,00	0,75	[53,2 %] 96,9 %	96,5 %
AlcPref=AlP2	0,35	[54,8 %] 5,5 %	5,3 %
c2d_AlcLieuRes_1=_1_0,00	0,13	[53,2 %] 69,8 %	69,6 %
c2d_AlcLieuRes_1=_2_1,00	-0,13	[52,8 %] 30,2 %	30,4 %
c2d_AlcLieuVoy_1=_2_1,00	-0,75	[48,3 %] 3,1 %	3,5 %
c2d_AlcLieuTrav_1=_1_0,00	-2,46	[52,8 %] 98,7 %	99,2 %
c2d_AlcLieuBar_1=_1_0,00	-5,29	[48,2 %] 57,2 %	63,0 %
AlcPref=AlP3	-6,08	[45,0 %] 37,8 %	44,6 %
c2d_AlcLieuMai_1=_2_1,00	-6,12	[47,7 %] 58,4 %	65,0 %

Cluster_HAC_2=c_hac_3			
Examples			[31,7 %] 552
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
ConsCafe	27,01	17,77 (8,56)	9,68 (8,52)
ConsVin	12,65	4,83 (5,69)	3,03 (4,05)
ConsBier	9,13	3,88 (5,92)	2,55 (4,12)
ConsEau	7,83	6,38 (6,04)	5,08 (4,71)
ConsSpir	4,97	1,00 (1,77)	0,75 (1,41)
ConsGaz	0,20	3,28 (4,15)	3,25 (4,77)
ConsThe	-10,04	3,05 (4,12)	5,94 (8,18)
Discrete attributes : [Recall] Accuracy			
c2d_AlcLieuMai_1=_2_1,00	4,57	[35,5 %] 72,6 %	65,0 %
AlcPref=AlP3	4,32	[37,1 %] 52,2 %	44,6 %
c2d_AlcLieuBar_1=_1_0,00	3,03	[34,3 %] 68,1 %	63,0 %
c2d_AlcLieuTrav_1=_1_0,00	2,56	[32,0 %] 100,0 %	99,2 %
c2d_AlcLieuVoy_1=_1_0,00	0,30	[31,8 %] 96,7 %	96,5 %
c2d_AlcLieuRes_1=_1_0,00	0,18	[31,9 %] 69,9 %	69,6 %
AlcPref=AlP1	-0,03	[31,7 %] 33,9 %	33,9 %
c2d_AlcLieuRes_1=_2_1,00	-0,18	[31,4 %] 30,1 %	30,4 %
c2d_AlcLieuVoy_1=_2_1,00	-0,30	[30,0 %] 3,3 %	3,5 %
AlcPref=AlP4	-1,81	[23,5 %] 4,2 %	5,6 %
AlcPref=AlP2	-1,95	[22,6 %] 3,8 %	5,3 %
c2d_AlcLieuTrav_1=_2_1,00	-2,56	[0,0 %] 0,0 %	0,8 %
c2d_AlcLieuBar_1=_2_1,00	-3,03	[27,3 %] 31,9 %	37,0 %
AlcPref=AlP5	-4,17	[18,1 %] 6,0 %	10,5 %
c2d_AlcLieuMai_1=_1_0,00	-4,57	[24,8 %] 27,4 %	35,0 %

b. Profil comportement

Ce profil a pour but de définir le comportement de nos consommateurs. C'est-à-dire que l'on va observer leurs buget dans chaque boisson, à quelles situations ils associent de boire telle ou telle boisson, leur alcool préféré, leur lieu de consommation et d'achat et enfin pourquoi ils consomment telle ou telle boisson.

Pour cela nous avons donc utilisé pour ce type de profil les variables :

- Bud*** : A combien estimez-vous votre budget hebdomadaire ?
- Alcpref : Quelle est la boisson alcoolisée que vous préférez ?
- Alclieu : où allez-vous souvent boire votre boisson préférée ?
- Asso*** : Quelle boisson associez-vous le plus volontiers à chacune de ces circonstances ?
- Alcach : la plupart du temps où achetez-vous votre boisson préférée ?
- Rai*** : La liste suivante indique des raisons de boire votre boisson préférée ... Pour chacune d'elle pouvez-vous indiquer votre degré d'accord

Pour déterminer ce profil nous allons passer par plusieurs étapes :

- Pour les variables Asso***, nécessité de réduire leurs nombres:

  - ACM pour récupérer leurs coordonnées sur les axes
  - Une Kmeans incluant les principaux axes de l'ACM avec un R^2 acceptable
  - Une CAH pour classer chaque individu dans un cluster
  - Création d'une nouvelle variable contenant l'appartenance à tel ou tel cluster de nos individus

- Pour la segmentation du profil :

  - Une ACM pour nos variables qualitatives (la totalité ici) afin de pouvoir en récupérer leurs coordonnées sur les axes
  - Une Kmeans où l'on inclut les axes de l'ACM. Cette Kmeans nous permet de faire une première segmentation de notre profil et nous permet d'observer le R^2
  - Une CAH où l'on utilisera les clusters donnés pas notre Kmeans. La classification ascendante étant assez longue, c'est pourquoi nous avons décidé de réduire ce temps en passant par une Kmeans.
  - Caractérisation de nos clusters pour voir les différentes distributions de nos individus dans tel ou tel cluster

Résultats de notre segmentation sur les variables Asso*** :

Description of "Cluster_HAC_1"

Cluster_HAC_1=c_hac_1				Cluster_HAC_1=c_hac_2				Cluster_HAC_1=c_hac_3			
Examples		[14,9 %] 259		Examples		[62,9 %] 1094		Examples		[22,2 %] 386	
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			
c2d_AssoEfThe_1=_2_1,00	19,78	[49,4 %] 64,1 %	19,3 %	c2d_AssoCelPre_1=_1_0,00	15,43	[70,9 %] 93,9 %	83,3 %	c2d_AssoCelPre_1=_2_1,00	18,92	[64,3 %] 48,4 %	16,7 %
c2d_AssoAmiThe_1=_2_1,00	19,55	[61,7 %] 46,7 %	11,3 %	c2d_AssoAmiPre_1=_1_0,00	13,82	[69,3 %] 95,1 %	86,3 %	c2d_AssoAmiPre_1=_2_1,00	18,83	[69,3 %] 42,7 %	13,7 %
c2d_AssoMusThe_1=_2_1,00	17,88	[47,2 %] 57,9 %	18,3 %	c2d_AssoGalGaz_1=_1_0,00	11,97	[68,9 %] 92,4 %	84,4 %	c2d_AssoMusGaz_1=_2_1,00	15,11	[51,2 %] 49,0 %	21,2 %
c2d_AssoEnergThe_1=_2_1,00	17,05	[51,0 %] 47,9 %	14,0 %	c2d_AssoRepGaz_1=_1_0,00	11,58	[68,3 %] 93,4 %	86,0 %	c2d_AssoGalPre_1=_2_1,00	14,45	[71,9 %] 25,1 %	7,8 %
c2d_AssoTrisThe_1=_2_1,00	15,43	[41,5 %] 54,8 %	19,7 %	c2d_AssoAmiGaz_1=_1_0,00	11,03	[68,5 %] 91,2 %	83,7 %	c2d_AssoTrisPre_1=_2_1,00	14,28	[88,3 %] 17,6 %	4,4 %
c2d_AssoGalThe_1=_2_1,00	14,77	[68,9 %] 23,9 %	5,2 %	c2d_AssoMusPre_1=_1_0,00	11,03	[66,4 %] 98,2 %	93,0 %	c2d_AssoGalGaz_1=_2_1,00	13,81	[54,2 %] 38,1 %	15,6 %
c2d_AssoEntThe_1=_2_1,00	14,09	[45,5 %] 40,9 %	13,4 %	c2d_AssoEfThe_1=_1_0,00	10,98	[69,1 %] 88,7 %	80,7 %	c2d_AssoMusPre_1=_2_1,00	13,76	[72,1 %] 22,8 %	7,0 %
c2d_AssoSoifThe_1=_2_1,00	13,64	[49,7 %] 33,6 %	10,1 %	c2d_AssoGalPre_1=_1_0,00	10,93	[66,6 %] 97,6 %	92,2 %	c2d_AssoRepGaz_1=_2_1,00	13,32	[55,1 %] 34,7 %	14,0 %
c2d_AssoAmiEau_1=_2_1,00	12,25	[67,2 %] 17,4 %	3,9 %	c2d_AssoAmiThe_1=_1_0,00	10,56	[67,3 %] 94,9 %	88,7 %	c2d_AssoCelVin_1=_1_0,00	12,90	[38,5 %] 66,6 %	38,4 %
c2d_AssoChaThe_1=_2_1,00	11,82	[24,9 %] 84,6 %	50,7 %	c2d_AssoTrisPre_1=_1_0,00	10,00	[65,4 %] 99,4 %	95,6 %	c2d_AssoAmiGaz_1=_2_1,00	12,37	[50,2 %] 36,8 %	16,3 %
c2d_AssoRepThe_1=_2_1,00	11,57	[71,2 %] 14,3 %	3,0 %	c2d_AssoMusGaz_1=_1_0,00	9,61	[68,7 %] 86,0 %	78,8 %	c2d_AssoEfGaz_1=_2_1,00	11,30	[52,9 %] 28,2 %	11,8 %
c2d_AssoEfEau_1=_2_1,00	11,34	[33,2 %] 48,6 %	21,8 %	c2d_AssoAmiVin_1=_2_1,00	9,39	[76,0 %] 49,3 %	40,8 %	c2d_AssoTrisGaz_1=_2_1,00	10,44	[51,8 %] 25,6 %	11,0 %
c2d_AssoEnergEau_1=_2_1,00	11,33	[39,7 %] 35,1 %	13,2 %	c2d_AssoRepVin_1=_2_1,00	8,99	[69,3 %] 79,8 %	72,4 %	c2d_AssoEntGaz_1=_2_1,00	10,16	[47,2 %] 30,1 %	14,1 %
c2d_AssoRepEau_1=_2_1,00	11,00	[43,7 %] 28,2 %	9,6 %	c2d_AssoGalThe_1=_1_0,00	8,88	[65,3 %] 98,4 %	94,8 %	c2d_AssoAmiVin_1=_1_0,00	9,67	[30,2 %] 80,6 %	59,2 %
c2d_AssoEntEau_1=_2_1,00	10,03	[37,0 %] 32,4 %	13,1 %	c2d_AssoEnergThe_1=_1_0,00	8,86	[67,0 %] 91,7 %	86,0 %	c2d_AssoCelGaz_1=_2_1,00	9,57	[55,2 %] 19,2 %	7,7 %
c2d_AssoGalEau_1=_2_1,00	9,56	[55,1 %] 14,7 %	4,0 %	c2d_AssoCelGaz_1=_1_0,00	8,43	[65,7 %] 96,4 %	92,3 %	c2d_AssoRepVin_1=_1_0,00	9,48	[37,5 %] 46,6 %	27,6 %
c2d_AssoMusEau_1=_2_1,00	9,47	[41,5 %] 23,6 %	8,5 %	c2d_AssoMusThe_1=_1_0,00	8,35	[67,5 %] 87,7 %	81,7 %	c2d_AssoEnergGaz_1=_2_1,00	8,95	[37,1 %] 44,0 %	26,3 %
c2d_AssoTrisEau_1=_2_1,00	8,27	[35,8 %] 24,7 %	10,3 %	c2d_AssoCelVin_1=_2_1,00	7,88	[70,1 %] 68,6 %	61,6 %	c2d_AssoChaGaz_1=_2_1,00	8,32	[76,9 %] 7,8 %	2,2 %
c2d_AssoSoifEau_1=_2_1,00	7,88	[22,5 %] 66,4 %	44,0 %	c2d_AssoTrisThe_1=_1_0,00	7,51	[67,2 %] 85,8 %	80,3 %	c2d_AssoGalVin_1=_1_0,00	8,26	[30,2 %] 69,9 %	51,4 %
c2d_AssoChaEau_1=_2_1,00	7,75	[70,8 %] 6,6 %	1,4 %	c2d_AssoEntThe_1=_1_0,00	7,37	[66,3 %] 91,2 %	86,6 %	c2d_AssoChaPre_1=_2_1,00	7,47	[75,8 %] 6,5 %	1,9 %
c2d_AssoSpoThe_1=_2_1,00	7,66	[57,5 %] 8,9 %	2,3 %	c2d_AssoEfGaz_1=_1_0,00	7,16	[65,9 %] 92,4 %	88,2 %	c2d_AssoEntPre_1=_2_1,00	6,80	[85,0 %] 4,4 %	1,2 %
c2d_AssoCelEau_1=_2_1,00	6,70	[66,7 %] 5,4 %	1,2 %	c2d_AssoTrisGaz_1=_1_0,00	6,53	[65,6 %] 92,8 %	89,0 %	c2d_AssoSoifGaz_1=_2_1,00	6,42	[33,7 %] 35,8 %	23,5 %
c2d_AssoCelThe_1=_2_1,00	5,65	[66,7 %] 3,9 %	0,9 %	c2d_AssoEntGaz_1=_1_0,00	6,52	[66,0 %] 90,0 %	85,9 %	c2d_AssoEfEau_1=_1_0,00	6,03	[25,4 %] 89,4 %	78,2 %
c2d_AssoMusGaz_1=_1_0,00	4,60	[16,9 %] 89,6 %	78,8 %	c2d_AssoAmiEau_1=_1_0,00	6,48	[64,4 %] 98,4 %	96,1 %	c2d_AssoMusVin_1=_1_0,00	5,98	[25,2 %] 90,4 %	79,6 %
c2d_AssoRepMin_1=_2_1,00	4,49	[22,1 %] 32,8 %	22,1 %	c2d_AssoMusVin_1=_2_1,00	6,36	[77,5 %] 25,1 %	20,4 %	c2d_AssoChaThe_1=_1_0,00	5,83	[28,1 %] 62,4 %	49,3 %
c2d_AssoCelVin_1=_2_1,00	4,36	[17,8 %] 73,7 %	61,6 %	c2d_AssoRepThe_1=_1_0,00	6,33	[64,2 %] 99,0 %	97,0 %	c2d_AssoRepPre_1=_2_1,00	5,71	[75,0 %] 3,9 %	1,2 %
c2d_AssoTrisSpi_1=_1_0,00	3,64	[16,0 %] 95,4 %	88,8 %	c2d_AssoSoifThe_1=_1_0,00	6,12	[65,3 %] 93,3 %	89,9 %	c2d_AssoEfPre_1=_2_1,00	5,63	[100,0 %] 2,3 %	0,5 %
c2d_AssoEfGaz_1=_1_0,00	3,48	[16,0 %] 94,6 %	88,2 %	c2d_AssoRepMin_1=_1_0,00	5,88	[66,5 %] 82,4 %	77,9 %	c2d_AssoMusThe_1=_1_0,00	5,61	[24,8 %] 91,5 %	81,7 %
c2d_AssoSoifMin_1=_1_0,00	3,45	[18,4 %] 51,7 %	42,0 %	c2d_AssoChaGaz_1=_1_0,00	5,88	[63,9 %] 99,4 %	97,8 %	c2d_AssoSoifEau_1=_1_0,00	5,32	[26,9 %] 67,9 %	56,0 %
c2d_AssoEfCaf_1=_1_0,00	3,41	[17,5 %] 64,5 %	54,7 %	c2d_AssoGalVin_1=_2_1,00	5,80	[69,8 %] 53,9 %	48,6 %	c2d_AssoSpoGaz_1=_2_1,00	5,21	[34,8 %] 22,8 %	14,5 %
c2d_AssoChaCaf_1=_1_0,00	3,35	[17,5 %] 64,9 %	55,3 %	c2d_AssoCelEau_1=_1_0,00	5,55	[63,6 %] 99,9 %	98,8 %	c2d_AssoSpoSpi_1=_2_1,00	4,96	[100,0 %] 1,8 %	0,4 %
c2d_AssoTrisGaz_1=_1_0,00	3,33	[15,9 %] 95,0 %	89,0 %	c2d_AssoEnergGaz_1=_1_0,00	5,42	[66,7 %] 78,1 %	73,7 %	c2d_AssoMusCaf_1=_1_0,00	4,91	[24,2 %] 93,0 %	85,2 %

Libellés de nos clusters :

- Cluster 1 : C1_AssoNoAlc ➔les individus associant principalement des boissons non alcoolisées aux différentes situations
- Cluster 2 : C2_AssoNoCons ➔Les individus n'associant pas de boissons en particulier selon la situation
- Cluster 3 : C3_AssoAlc ➔ Les individus associant principalement des boissons alcoolisées aux différentes situations

Segmentation du profil comportement :

Description of "Cluster_HAC_1"

Cluster_HAC_1=c_hac_1				Cluster_HAC_1=c_hac_2				Cluster_HAC_1=c_hac_3			
Examples		[18,7 %] 326		Examples		[43,9 %] 764		Examples		[37,3 %] 649	
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			
RaiModet=Oui	22,50	[75,1 %] 49,1 %	12,2 %	AlcPref=AlP3	28,38	[81,6 %] 82,9 %	44,6 %	AlcPref=AlP1	19,56	[69,0 %] 62,7 %	33,9 %
RaiFortt=Oui	22,40	[81,3 %] 43,9 %	10,1 %	RaiSoift=Non	21,80	[66,1 %] 86,9 %	57,7 %	c2d_AlcLieuBar_1=_2_1,00	16,80	[62,7 %] 62,2 %	37,0 %
RaiFriendt=Oui	22,28	[63,3 %] 60,7 %	18,0 %	c2d_AlcLieuBar_1=_1_0,00	19,60	[61,8 %] 88,6 %	63,0 %	c2d_AlcAchBar_1=_2_1,00	14,40	[69,5 %] 39,6 %	21,3 %
RaiPersot=Oui	20,32	[53,9 %] 65,0 %	22,6 %	c2d_AlcAchBar_1=_1_0,00	16,00	[53,8 %] 96,5 %	78,7 %	RaiSoift=Oui	14,22	[56,6 %] 64,1 %	42,3 %
RaiDifft=Oui	20,17	[78,5 %] 38,0 %	9,1 %	c2d_AlcLieuRes_1=_2_1,00	12,61	[66,7 %] 46,1 %	30,4 %	BudVin=BV1	13,34	[63,7 %] 43,8 %	25,6 %
RaiAiset=Oui	19,50	[43,2 %] 82,5 %	35,8 %	c2d_AlcAchSpe_1=_2_1,00	12,55	[74,7 %] 32,5 %	19,1 %	c2d_AlcLieuMai_1=_1_0,00	12,44	[57,0 %] 53,5 %	35,0 %
RaiExcit=Oui	19,46	[49,4 %] 68,7 %	26,0 %	RaiAccesst=Non	12,38	[55,1 %] 79,6 %	63,4 %	RaiPersot=Non	11,22	[44,4 %] 92,0 %	77,4 %
RaiAccesst=Oui	18,08	[41,0 %] 80,1 %	36,6 %	RaiQPt=Non	11,95	[56,9 %] 70,8 %	54,7 %	AlcPref=AlP5	10,70	[73,6 %] 20,6 %	10,5 %
RaiMarquet=Oui	16,88	[48,4 %] 57,1 %	22,1 %	c2d_AlcLieuMai_1=_2_1,00	11,90	[54,3 %] 80,4 %	65,0 %	c2d_AlcAchSpe_1=_1_0,00	10,08	[43,0 %] 93,2 %	80,9 %
RaiMoodt=Oui	16,15	[33,9 %] 90,2 %	49,9 %	RaiModet=Non	11,58	[49,1 %] 98,0 %	87,8 %	RaiTradt=Non	10,06	[44,4 %] 86,9 %	73,0 %
RaiRiskt=Oui	16,14	[42,2 %] 66,0 %	29,3 %	BudBier=BB1	9,67	[61,4 %] 42,4 %	30,4 %	c2d_AlcLieuRes_1=_1_0,00	9,92	[44,9 %] 83,8 %	69,6 %
RaiQPt=Oui	13,85	[33,0 %] 79,8 %	45,3 %	RaiFriendt=Non	9,62	[49,3 %] 92,0 %	82,0 %	RaiSantt=Non	9,05	[43,5 %] 86,6 %	74,3 %
RaiSantt=Oui	13,80	[40,7 %] 55,8 %	25,7 %	RaiFortt=Non	9,34	[47,7 %] 97,5 %	89,9 %	RaiFortt=Non	8,49	[40,6 %] 97,8 %	89,9 %
RaiTradt=Oui	12,74	[38,4 %] 55,2 %	27,0 %	RaiRiskt=Non	9,09	[50,9 %] 81,9 %	70,7 %	RaiFriendt=Non	8,10	[41,7 %] 91,7 %	82,0 %
RaiSoift=Oui	10,10	[29,8 %] 67,2 %	42,3 %	RaiExcit=Non	8,48	[49,9 %] 84,0 %	74,0 %	RaiDifft=Non	8,10	[40,3 %] 98,2 %	90,9 %
BudBier=BB3	6,63	[37,4 %] 19,9 %	10,0 %	RaiAiset=Non	8,33	[51,3 %] 75,0 %	64,2 %	RaiAiset=Non	7,19	[43,5 %] 74,9 %	64,2 %
Clustering Asso=C3_AssoAlc	5,71	[28,8 %] 34,0 %	22,2 %	BudVin=BV2	7,98	[51,8 %] 70,0 %	59,4 %	RaiExcit=Non	7,01	[42,1 %] 83,5 %	74,0 %
AlcPref=AlP1	4,98	[25,3 %] 45,7 %	33,9 %	RaiDifft=Non	7,97	[46,9 %] 97,1 %	90,9 %	Clustering Asso=C3_AssoAlc	6,55	[51,6 %] 30,7 %	22,2 %
BudSpir=BS3	4,74	[36,6 %] 11,3 %	5,8 %	RaiMarquet=Non	6,95	[48,3 %] 85,7 %	77,9 %	RaiMarquet=Non	6,49	[41,3 %] 86,3 %	77,9 %
BudGaz=BG3	4,34	[35,0 %] 11,0 %	5,9 %	RaiMoodt=Non	6,46	[51,6 %] 58,9 %	50,1 %	RaiMoodt=Non	6,40	[44,7 %] 60,1 %	50,1 %
c2d_AlcLieuBar_1=_2_1,00	4,11	[23,8 %] 46,9 %	37,0 %	BudGaz=BG1	6,11	[56,6 %] 31,9 %	24,8 %	RaiModet=Non	6,27	[40,0 %] 94,1 %	87,8 %
c2d_AlcAchMag_1=_2_1,00	4,07	[21,4 %] 77,6 %	68,1 %	Clustering Asso=C2_AssoNoCons	5,54	[49,0 %] 70,2 %	62,9 %	BudThe=BT1	4,56	[44,6 %] 41,6 %	34,8 %
BudThe=BT3	3,97	[41,3 %] 5,8 %	2,6 %	Clustering Asso=C1_AssoNoAlc	5,18	[58,7 %] 19,9 %	14,9 %	BudBier=BB2	3,94	[41,1 %] 65,6 %	59,6 %
RaiPlaist=Oui	3,88	[19,8 %] 97,5 %	92,4 %	c2d_AlcAchRes_1=_2_1,00	5,17	[61,1 %] 15,8 %	11,4 %	RaiRiskt=Non	3,70	[40,1 %] 76,0 %	70,7 %
BudSpir=BS2	3,85	[23,2 %] 49,1 %	39,7 %	BudVin=BV3	5,12	[58,5 %] 19,9 %	15,0 %	RaiPlaist=Non	3,51	[51,5 %] 10,5 %	7,6 %
c2d_AlcLieuRes_1=_1_0,00	3,74	[21,1 %] 78,2 %	69,6 %	RaiPersot=Non	5,04	[47,2 %] 83,1 %	77,4 %	c2d_AlcAchRes_1=_1_0,00	3,26	[38,7 %] 91,8 %	88,6 %
c2d_AlcAchSpe_1=_1_0,00	3,48	[20,3 %] 87,7 %	80,9 %	BudThe=BT2	5,03	[48,6 %] 69,1 %	62,5 %	BudEau=BE3	3,19	[50,4 %] 10,0 %	7,4 %
AlcPref=AlP2	2,61	[29,0 %] 8,3 %	5,3 %	BudCafe=BC2	2,88	[46,0 %] 76,0 %	72,6 %	BudBier=BB3	3,15	[48,3 %] 12,9 %	10,0 %
c2d_AlcLieuVoy_1=_2_1,00	2,61	[31,7 %] 5,8 %	3,5 %	BudSpir=BS1	2,86	[47,0 %] 58,4 %	54,5 %	c2d_AlcAchMag_1=_1_0,00	2,47	[41,5 %] 35,4 %	31,9 %
c2d_AlcAchRes_1=_1_0,00	2,54	[19,6 %] 92,6 %	88,6 %	RaiSantt=Non	2,03	[45,4 %] 76,7 %	74,3 %	BudGaz=BG2	1,97	[38,8 %] 72,1 %	69,3 %
c2d_AlcAchBar_1=_2_1,00	2,50	[23,2 %] 26,4 %	21,3 %	c2d_AlcAchDis_1=_1_0,00	1,77	[44,0 %] 100,0 %	99,8 %	BudSpir=BS1	1,91	[39,3 %] 57,5 %	54,5 %
BudGaz=BG2	1,75	[19,8 %] 73,3 %	69,3 %	c2d_AlcLieuTrav_1=_1_0,00	1,70	[44,1 %] 99,6 %	99,2 %	RaiAccesst=Non	1,89	[39,0 %] 66,3 %	63,4 %

c. Profil consommateur

Ce profil a pour but de définir le profil de nos consommateurs. C'est-à-dire que l'on va observer les variables sociodémographiques de type âge, sexe, situation professionnelle, niveau d'étude et leur pays.

Pour cela nous avons donc utilisé pour ce type de profil les variables :

- Sexe : Votre genre
- Age : Votre âge
- Educ : Votre niveau d'éducation
- SituProf : Votre situation professionnelle
- RevNet : Quel est le revenu mensuel net de votre ménage ?

Pour déterminer ce profil nous allons passer par plusieurs étapes :

- Une ACM pour nos variables qualitatives (la totalité ici) afin de pouvoir en récupérer leurs coordonnées sur les axes
- Une Kmeans où l'on inclut les axes de l'ACM. Cette Kmeans nous permet de faire une première segmentation de notre profil et nous permet d'observer le R^2
- Une CAH où l'on utilisera les clusters donnés pas notre Kmeans. La classification ascendante étant assez longue, c'est pourquoi nous avons décidé de réduire ce temps en passant par une Kmeans.
- Caractérisation de nos clusters pour voir les différentes distributions de nos individus dans tel ou tel cluster

Segmentation du profil consommateur :

Description of "Cluster_HAC_3"

Cluster_HAC_3=c_hac_1			
Examples			[33,8 %] 587
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Age	-33,10	26,57 (4,93)	42,67 (14,47)
Discrete attributes : [Recall] Accuracy			
SituProf=SPr3	24,23	[98,1 %] 44,8 %	15,4 %
Pays=Ita	8,64	[62,2 %] 19,6 %	10,6 %
Pays=Alle	8,53	[55,0 %] 27,9 %	17,1 %
Sexe=Femme	6,53	[41,7 %] 57,6 %	46,6 %
Educ=Ed5	5,99	[43,4 %] 42,6 %	33,1 %
RevNet=RvN1	5,63	[51,2 %] 17,9 %	11,8 %
Educ=Ed3	2,75	[38,5 %] 34,1 %	29,8 %
RevNet=RvN2	0,81	[35,4 %] 23,9 %	22,7 %
SituProf=SPr1	-1,17	[29,6 %] 8,0 %	9,1 %
RevNet=RvN3	-2,00	[31,7 %] 50,8 %	54,1 %
Pays=Esp	-3,08	[28,9 %] 29,5 %	34,4 %
Educ=Ed4	-3,46	[23,4 %] 8,7 %	12,5 %
RevNet=RvN4	-3,64	[22,2 %] 7,5 %	11,4 %
Pays=Fr	-4,11	[23,9 %] 12,9 %	18,3 %
Educ=Ed2	-4,57	[22,7 %] 12,1 %	18,0 %
Educ=Ed1	-4,76	[13,3 %] 2,6 %	6,5 %
Sexe=Homme	-6,53	[26,8 %] 42,4 %	53,4 %
Pays=RU	-7,13	[17,4 %] 10,1 %	19,6 %
SituProf=SPr2	-8,49	[26,0 %] 46,8 %	60,8 %
SituProf=SPr4	-12,05	[0,8 %] 0,3 %	14,7 %

Cluster_HAC_3=c_hac_2			
Examples			[25,2 %] 439
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Age	32,65	62,17 (4,73)	42,67 (14,47)
Discrete attributes : [Recall] Accuracy			
SituProf=SPr4	27,24	[93,7 %] 54,4 %	14,7 %
Educ=Ed1	6,37	[50,4 %] 13,0 %	6,5 %
Educ=Ed4	5,33	[39,9 %] 19,8 %	12,5 %
Sexe=Homme	4,17	[29,3 %] 62,0 %	53,4 %
RevNet=RvN2	3,59	[32,2 %] 28,9 %	22,7 %
Educ=Ed2	3,16	[32,3 %] 23,0 %	18,0 %
Pays=RU	2,53	[30,6 %] 23,7 %	19,6 %
Pays=Esp	0,70	[26,3 %] 35,8 %	34,4 %
Pays=Alle	-0,03	[25,2 %] 17,1 %	17,1 %
RevNet=RvN1	-0,13	[24,9 %] 11,6 %	11,8 %
Pays=Fr	-0,61	[23,9 %] 17,3 %	18,3 %
SituProf=SPr1	-0,98	[22,0 %] 8,0 %	9,1 %
RevNet=RvN4	-1,56	[20,7 %] 9,3 %	11,4 %
RevNet=RvN3	-1,94	[23,4 %] 50,1 %	54,1 %
Educ=Ed5	-2,39	[21,7 %] 28,5 %	33,1 %
Pays=Ita	-3,53	[14,6 %] 6,2 %	10,6 %
Sexe=Femme	-4,17	[20,6 %] 38,0 %	46,6 %
Educ=Ed3	-7,48	[13,3 %] 15,7 %	29,8 %
SituProf=SPr3	-10,34	[0,0 %] 0,0 %	15,4 %
SituProf=SPr2	-11,51	[15,6 %] 37,6 %	60,8 %

Cluster_HAC_3=c_hac_3			
Examples			[41,0 %] 713
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Age	2,98	43,91 (5,27)	42,67 (14,47)
Discrete attributes : [Recall] Accuracy			
SituProf=SPr2	18,33	[58,4 %] 86,5 %	60,8 %
RevNet=RvN4	4,88	[57,1 %] 15,8 %	11,4 %
Pays=RU	4,62	[52,1 %] 24,8 %	19,6 %
Pays=Fr	4,49	[52,2 %] 23,3 %	18,3 %
Educ=Ed3	3,96	[48,2 %] 35,1 %	29,8 %
RevNet=RvN3	3,64	[45,0 %] 59,3 %	54,1 %
Sexe=Homme	2,59	[43,9 %] 57,1 %	53,4 %
Pays=Esp	2,34	[44,8 %] 37,6 %	34,4 %
SituProf=SPr1	2,00	[48,4 %] 10,8 %	9,1 %
Educ=Ed2	1,61	[45,0 %] 19,8 %	18,0 %
Educ=Ed1	-1,05	[36,3 %] 5,8 %	6,5 %
Educ=Ed4	-1,38	[36,7 %] 11,2 %	12,5 %
Sexe=Femme	-2,59	[37,7 %] 42,9 %	46,6 %
Educ=Ed5	-3,64	[34,9 %] 28,2 %	33,1 %
RevNet=RvN2	-3,95	[32,4 %] 18,0 %	22,7 %
Pays=Ita	-5,19	[23,2 %] 6,0 %	10,6 %
RevNet=RvN1	-5,30	[23,9 %] 6,9 %	11,8 %
Pays=Alle	-8,17	[19,8 %] 8,3 %	17,1 %
SituProf=SPr4	-12,48	[5,5 %] 2,0 %	14,7 %
SituProf=SPr3	-14,16	[1,9 %] 0,7 %	15,4 %

d. Fusion de nos profils

Finalement, une fois nos 3 profils déterminés, nous allons observer leurs interactions entre eux.

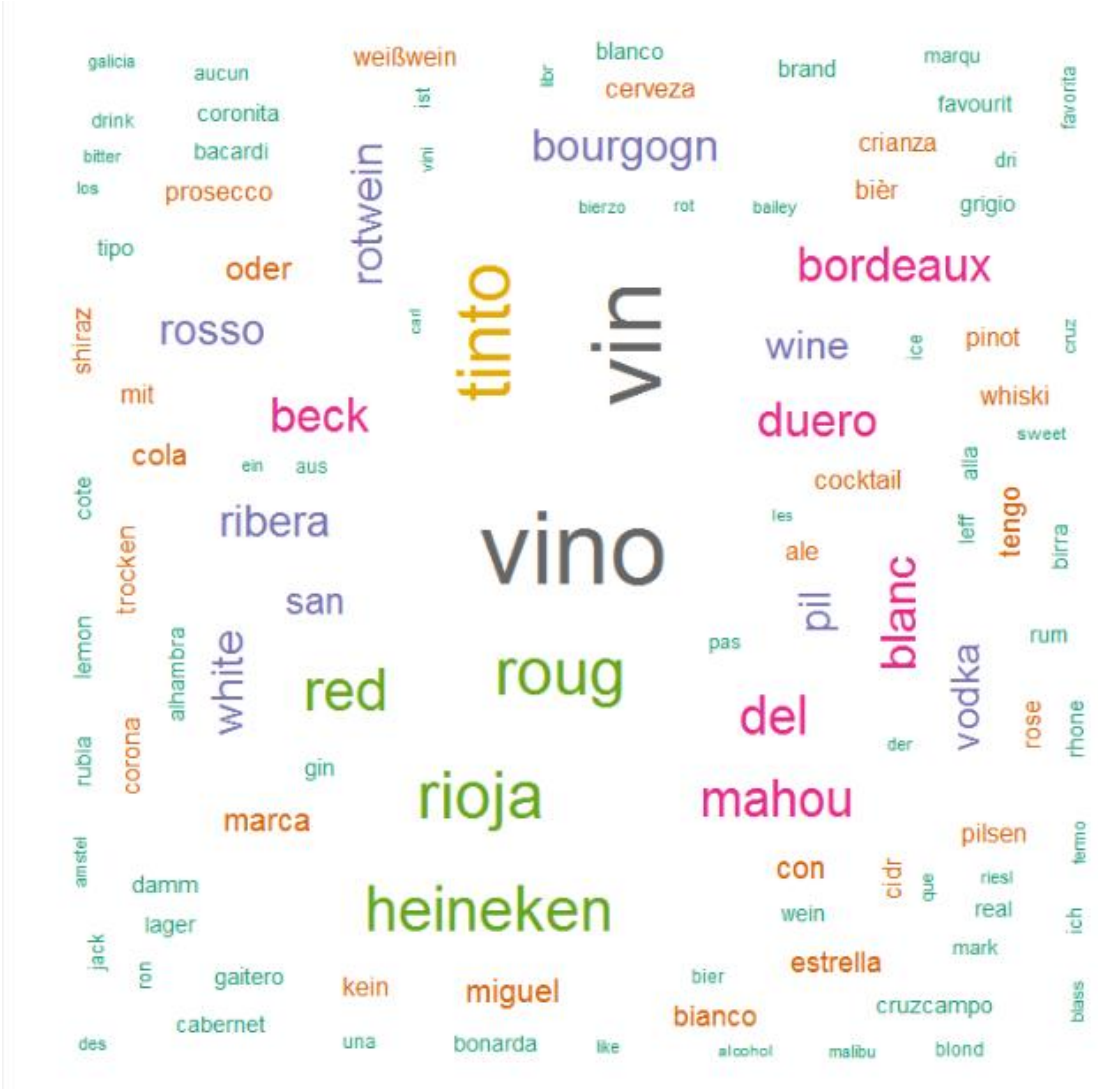
En effet, notre but est d'établir une segmentation globale comprenant à la fois les comportements et habitudes de consommation ainsi que le profil sociodémographique de nos consommateurs.

Pour cela nous allons effectuer trois méthodes différentes afin de comparer nos résultats :

- un test d'indépendance du Khi-2 entre nos différentes segmentations
- Une Analyse en composantes principales
- Une Kmeans + CAH afin de reconstruire une segmentation globale de nos 3 profils

Tout cela nous permettra d'étudier le lien potentiel entre chaque cluster de chacune de nos segmentations et établir s'il existe un lien potentiel entre eux.

V°) Analyse sémantique



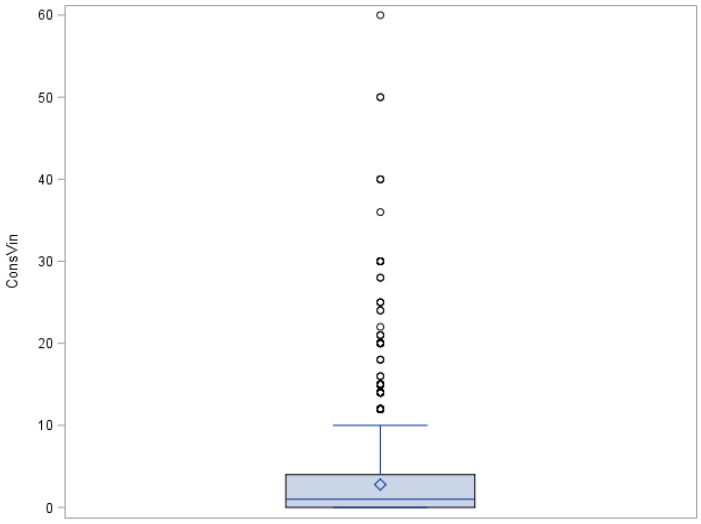
Nos résultats ne sont que provisoires pour le moment, ils sont susceptibles de changer au fur et à mesure de notre avancement.

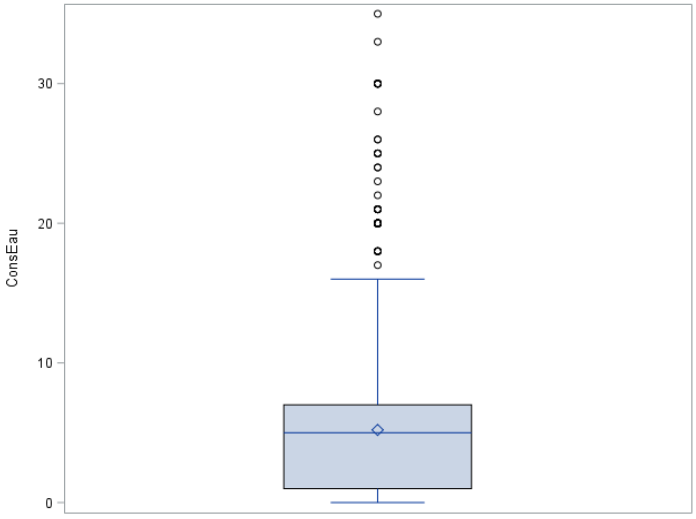
De plus, beaucoup plus de détails seront apportés à nos sorties, ainsi que des graphiques

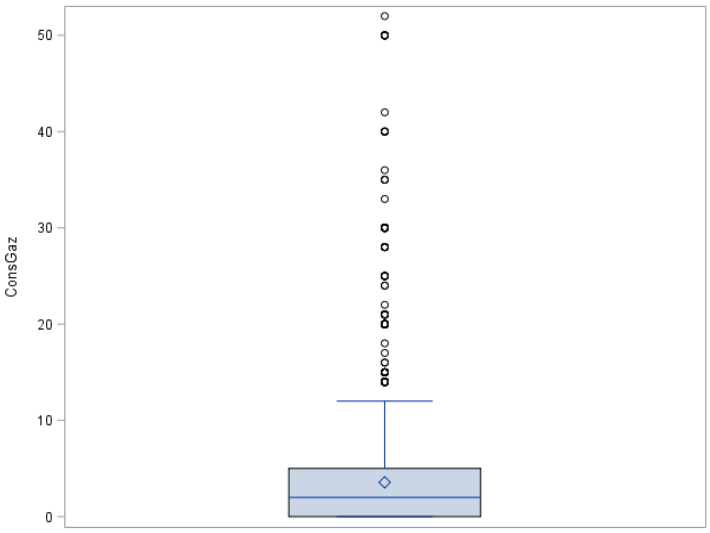
Les résultats ici présents correspondent à une démarche qui restera, quoi qu'il en soit, similaire.

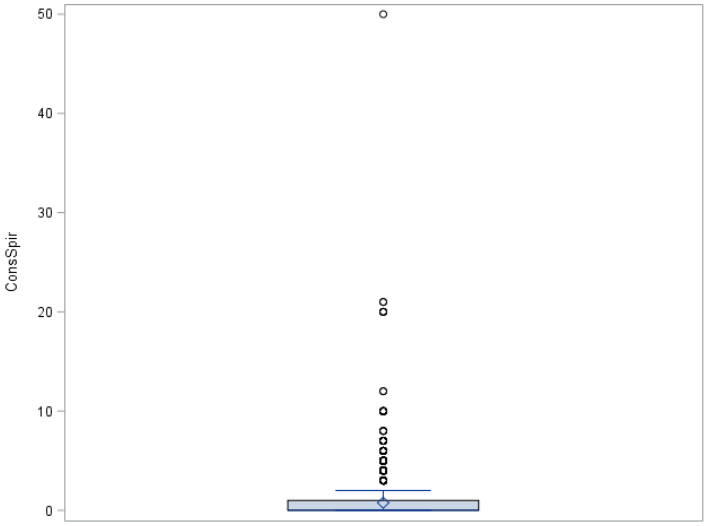
ANNEXES

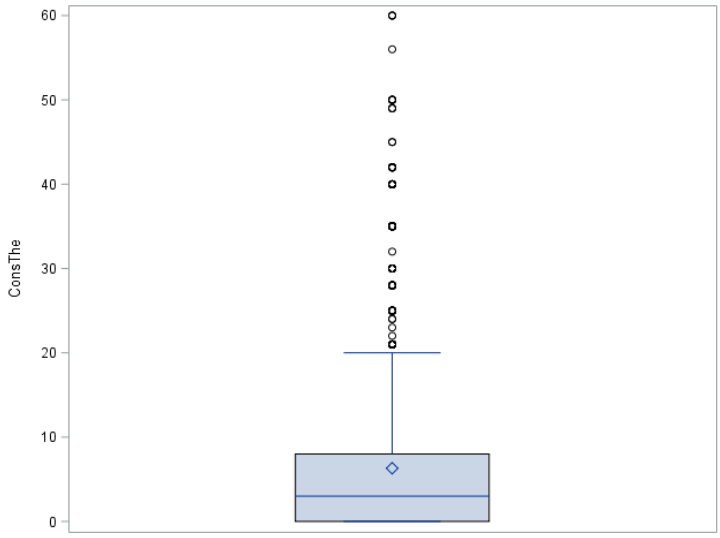
ANNEXE 1 : BOXPLOT des variables ConsXX

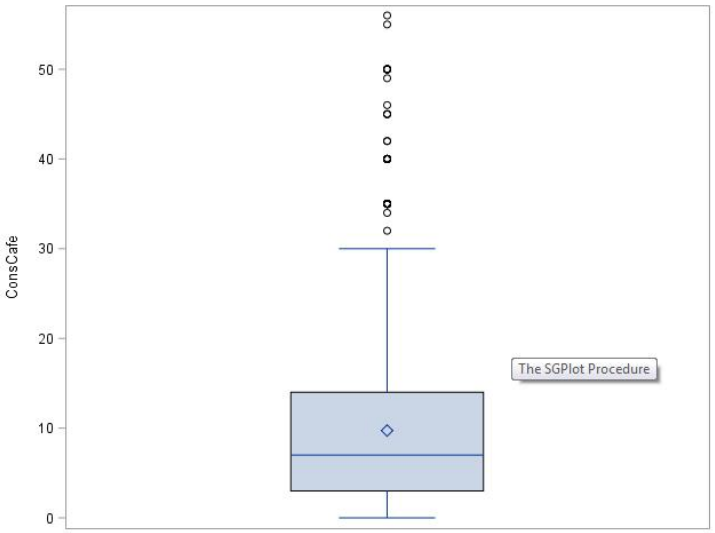


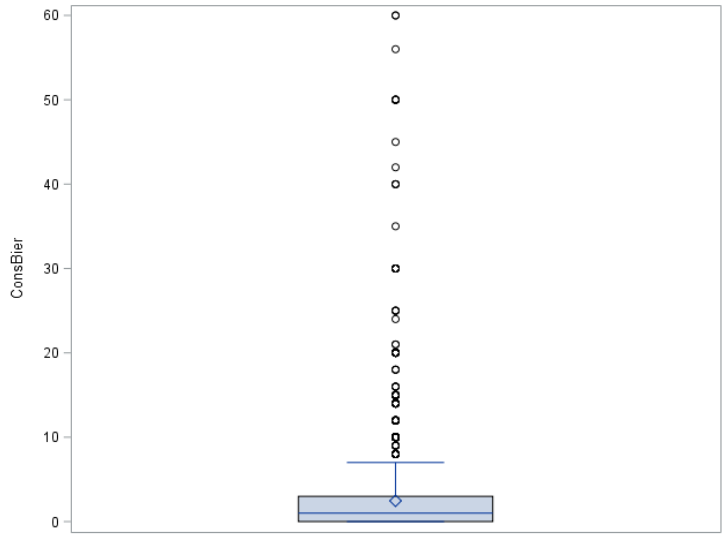



ConsGaz
0
10
20
30
40
50


ConsSpir
0
10
20
30
40
50


ConsThe
0
10
20
30
40
50
60


ConsCafe
0
10
20
30
40
50
The SGPlot Procedure


ConsBier
0
10
20
30
40
50
60