

13/04/2025

# Projet STA101 :

Etude des facteurs impactant le niveau d'obésité

Sourou Alain NOUNAWON  
CNAME PARIS

## Table des matières

1	Présentation du projet.....	3
1.1	Contexte.....	3
1.2	Problématiques.....	3
1.3	Les données.....	4
1.4	Méthodologie.....	5
2	Description des données .....	6
2.1	Analyse univariée .....	6
2.1.1	Variables quantitatives.....	6
2.1.2	Variables qualitatives .....	7
2.2	Analyse bivariée.....	9
2.2.1	Lien entre variables quantitatives.....	9
2.2.2	Lien entre variables quantitatives et qualitatives .....	11
2.2.3	Lien entre variables qualitatives .....	12
3	Analyse factorielle : Analyse en Correspondances Multiples (ACM).....	13
3.1	Choix des variables actives .....	13
3.2	Choix du nombre de dimensions.....	14
3.3	Interprétation de l'ACM.....	14
3.3.1	Interprétation des graphes des variables.....	14
3.3.2	Interprétation des graphes des individus .....	16
3.4	Conclusion sur l'ACM.....	17
4	Classification non supervisée : Classification Ascendante Hiérarchique (CAH) .....	18
4.1	Mise en œuvre CAH .....	18
4.2	Description des groupes à partir des individus.....	19
4.3	Conclusion sur Classification non supervisée .....	20
5	Conclusion.....	20
	Annexes .....	21

## Les tableaux :

Tableau 1 : Variables renommées et description .....	5
Tableau 2 : Résumé numérique des variables quantitatives .....	6

## Les figures :

Figure 1 : Histogrammes de age, taille, poids et imc .....	7
Figure 2 : Répartition en effectif (à gauche) en proportion (à droite) de la variable grp_imc .....	8
Figure 3 : Nuage de points entre taille et poids (à gauche) et entre poids et imc (à droite).....	9
Figure 4 : Heatmap de corrélation .....	10
Figure 5 : Barplots bivariés de quelques paires de variables qualitatives .....	12
Figure 6 : Scree plot.....	14
Figure 7 : Graphe des variables sur les axes 1 et 2 .....	15
Figure 8 : Contribution des modalités sur les axes 1 (à gauche) et 2 (à droite).....	16
Figure 9 : Graphe des individus sur les axes 1 et 2, coloré en fonction du groupe IMC (à gauche) et sexe (à droite).....	17
Figure 10 : Dendrogramme issu de la distance de Ward (à gauche), saut d'inertie (à droite).....	18
Figure 11 : Dendrogramme représentant le nombre de classe basé sur le saut d'inertie .....	18
Figure 12 : Classes des individus .....	19

## Annexes :

Annexe 1 : Tableau des variables et leurs types/modalités à l'origine.....	21
Annexe 2 : Courbe de densité (à gauche), Boxplot (au milieu) et QQ Plot (à droite) de age, taille, poids et imc .....	21
Annexe 3 : Résumé numérique des variables qualitatives .....	22
Annexe 4 : Répartition en effectif (à gauche) en proportion (à droite) de quelques modalités.....	23
Annexe 5 : Résumé numérique bivarié entre variables qualitatives et IMC .....	25
Annexe 6 : Boxplots entre l'IMC et les différentes variables qualitatives .....	27
Annexe 7 : Tests d'hypothèses sur les relations bivariées entre les variables qualitatives et l'IMC .....	27
Annexe 8 : Barplots bivariés .....	30
Annexe 9 : Tests d'hypothèses des relations entre les variables qualitatives .....	31
Annexe 10 : Tableau des valeurs propres.....	33
Annexe 11 : Graphe des variables sur les axes 3 et 4 .....	34
Annexe 12 : Contribution des variables sur les axes 3 et 4 : .....	35

# 1 Présentation du projet

## 1.1 Contexte

Pour valider l'UE [STA101](#) du CNAM (Conservatoire national des arts et métiers), une étude de cas donnant lieu à la rédaction d'un rapport, doit être réalisée en mettant œuvre toutes les techniques vues au cours permettant d'explorer, décrire et interpréter des données dans leur aspect multidimensionnel.

L'étude menée ici, s'intéresse aux facteurs déterminant les risques d'obésité, laquelle constitue un problème majeur de santé publique. Les données étudiées sont issues de [Kaggle](#) : "*ObesityDataSet\_raw\_and\_data\_sinthetic.csv*". Elles comprennent des données permettant d'estimer les niveaux d'obésité chez les individus des pays du Mexique, du Pérou et de Colombie, en fonction de leurs habitudes alimentaires et de leur condition physique. « *77% des données ont été générées de manière synthétique à l'aide de l'outil Weka et du filtre SMOTE, 23% des données ont été collectées directement auprès des utilisateurs via une plateforme web.* » selon la [source<sup>1</sup>](#).

Le jeu de données comporte initialement 17 variables (dont 8 quantitatives et 9 qualitatives) et 2111 individus âgés de 14 à 61 ans. Il est important de préciser que le présent rapport fait l'objet d'une étude d'analyse descriptive (et non prédictive), l'objectif étant d'analyser les données au moyen des méthodes descriptives pour répondre à un certain nombre de questions posées sur les facteurs influençant l'obésité.

## 1.2 Problématiques

Grâce aux diverses informations de ce jeu de données, l'étude permettra de comprendre les facteurs influençant les niveaux d'obésité au Mexique, au Pérou et en Colombie en répondant aux questions suivantes :

- y a-t-il un lien entre le niveau d'obésité et la taille ? et le poids ?
- le fait qu'un membre de la famille ait souffert ou souffre de surpoids peut-il déterminer le niveau d'obésité de l'individu mesuré par l'IMC (Indice de Masse Corporelle) ?
- est-ce que la consommation fréquente d'aliments riches en calories favorise un niveau élevé d'obésité ?
- quel est l'effet de la fréquence de consommation de légumes dans les repas sur l'obésité ?
- le nombre de repas par jour caractérise-t-il la corpulence ? et manger entre les repas ? et le tabagisme ?
- la consommation d'eau quotidienne décrit-elle l'obésité ?
- quelle action exercent la surveillance des calories consommées, la pratique d'activité physique et le moyen de transport principal sur la corpulence ?
- existe-t-il un lien entre le niveau d'obésité et la fréquence de consommation d'alcool ?
- le temps passé à utiliser des appareils électroniques agit-il sur niveau d'obésité ?

---

<sup>1</sup> <https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub#undtbl1>

Avec l'analyse factorielle de correspondances multiples, le tableau disjonctif conjoint conduira à plus de variables, dû au nombre de modalités par variable ; certaines variables quantitatives seront d'ailleurs transformées en qualitatives. Cette analyse factorielle permettra :

- de réduire la dimensionnalité et ainsi construire de nouvelles variables expliquant les principaux axes ;
- d'identifier les ressemblances et oppositions entre les individus, puis visualiser l'ensemble des associations entre les modalités.

Les résultats de l'ACM favoriseront la classification au sens non supervisée des différents groupes pouvant ressortir, à l'aide de l'algorithme CAH (Classification Ascendante Hiérarchique).

### 1.3 Les données

Le jeu de données "*ObesityDataSet\_raw\_and\_data\_sinthetic.csv*" provient de [Kaggle](#) et se présente à l'origine à l'*Annexe 1*. On retient :

- 17 variables : 8 quantitatives et 9 qualitatives ;
- 2111 individus : âgés de 14 à 61 ans.

Il faut préciser que les variables ont été renommées (et leurs modalités traduites) en français afin d'être plus explicite :

Nº	Variables (renommées)	Description	Type Variables/modalités
1	sex	Genre	• femme • homme
2	age	Age	Numérique value
3	taille	Taille	Numérique (mètre)
4	poids	Poids	Numérique (en Kg)
5	ant_fam_obesite	Antécédents familiaux d'obésité	• oui • non
6	conso_freq_alim_cal	Consommation fréquente d'aliments riches en calories	• oui • non
7	freq_conso_legumes	Fréquence de consommation de légumes	Echelle 1 à 3
8	nbre_repas_jr	Nombre de repas par jour	Echelle 1 à 4
9	mange_entre_repas	Manger entre les repas	• non • souvent • frequemment • toujours
10	fume	Tabagisme	• oui • non
11	conso_eau_jr	Consommation d'eau quotidienne	Echelle 1 à 3
12	surveille_cal_conso	Surveillance des calories consommées	• oui • non
13	act_physique	Activité physique	Echelle 0 à 3
14	tps_use_tech	Temps passé à utiliser des appareils électroniques	Echelle 0 à 2
15	freq_conso_alcool	Fréquence de consommation d'alcool	• non • souvent • frequemment • toujours
16	moy_trans	Moyen de transport principal	• automobile • moto • velo • transports_en_commun • marche

17	niv_obesite	Niveau d'obésité	<ul style="list-style-type: none"> <li>• <i>sous_poids</i> • <i>normal</i></li> <li>• <i>surpoids_niv_1</i> •</li> <li><i>surpoids_niv_2</i></li> <li>• <i>obesite_type_1</i> •</li> <li><i>obesite_type_2</i> •</li> <li><i>obesite_type_3</i></li> </ul>
----	-------------	------------------	--

Tableau 1 : Variables renommées et description

En particulier pour l'analyse factorielle, les variables qualitatives correspondantes aux variables quantitatives de type échelle ont été ajoutées pour mieux représenter la catégorie, et ainsi que les groupes d'âge.

De nouvelles variables ont été ajoutées pour mieux représenter la catégorie. Il s'agit de :

- *grp\_freq\_conso\_legumes* à partir de « *freq\_conso\_legumes* » ;
- *grp\_nbre\_repas\_jr* à partir de « *nbre\_repas\_jr* » ;
- *grp\_conso\_eau\_jr* à partir de « *conso\_eau\_jr* » ;
- *grp\_act\_physique* à partir de « *act\_physique* » ;
- *grp\_tps\_use\_tech* à partir de « *tps\_use\_tech* » ;
- *grp\_age* à partir de « *age* ».

Ainsi que l'IMC (variable *imc*) calculée avec la formule (poids/taille<sup>2</sup>) et la variable qualitative correspondante *grp\_imc* obtenue selon les informations publiées par l'Assurance Maladie sur son site [ameli](#)<sup>2</sup> :

« Si l'IMC est :

- < 18,5 kg/m<sup>2</sup>, il s'agit d'une insuffisance pondérale ;
- = ou > 18,5 et < 25 kg/m<sup>2</sup>, la corpulence est normale ;
- = ou > 25 et < 30 kg/m<sup>2</sup>, il existe un surpoids ;
- = ou > 30 kg/m<sup>2</sup>, il s'agit d'obésité. »

La variable *grp\_imc* créée sera prise en compte dans la présente étude au détriment de la variable d'origine *niv\_obesite* constituant en effet une variable cible pour le cas d'une modélisation prédictive.

## 1.4 Méthodologie

A l'aide de l'outil R et après le chargement des librairies nécessaires, ont été réalisés l'import du jeu de données, la vérification de données manquantes et le formatage des variables. S'en est suivie l'analyse des données suivant les méthodes descriptives pour répondre aux différentes problématiques :

- analyse univariée pour décrire de manière numérique et graphique les individus suivant les différentes variables ;
- analyse bivariée pour identifier les relations entre les différentes variables ;

---

<sup>2</sup> <https://www.ameli.fr/yvelines/assure/sante/themes/surpoids-obesite-adulte/definition-causes-risques>

- analyse factorielle des correspondances multiples en raison de l'aspect qualitatif du jeu de données et les réponses à apporter aux problématiques.

Enfin, les groupes d'individus ont été mis en évidence grâce à la classification non-supervisée : CAH (Classification Ascendante Hiérarchique).

## 2 Description des données

Gender	Age	Height	Weight	Family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	EAF	TUE	CALC	MTRANS	NOObesidad
1 Female	21	1.62	64.0	yes	no	2	3	Sometimes	no	2	no	0	1	no	Public_Transportation	Normal_Weight
2 Female	21	1.52	59.0	yes	no	3	3	Sometimes	yes	3	yes	3	0	Sometimes	Public_Transportation	Normal_Weight
3 Male	73	1.80	77.0	yes	no	2	3	never	no	2	no	2	1	Frequently	Public_Transportation	Normal_Weight
4 Male	27	1.80	87.0	no	no	3	3	Sometimes	no	2	no	2	0	Frequently	Walking	Overweight_Level_I
5 Male	22	1.78	89.8	no	no	2	1	Sometimes	no	2	no	0	0	Sometimes	Public_Transportation	Overweight_Level_II
6 Male	29	1.62	53.0	no	yes	2	3	Sometimes	no	2	no	0	0	Sometimes	Automobile	Normal_Weight

### 2.1 Analyse univariée

#### 2.1.1 Variables quantitatives

Pour les variables numériques, on examine principalement :

- les statistiques descriptives : moyenne, médiane, écart-type, etc. ;
- la distribution via les boîtes à moustaches (boxplots), diagrammes Quantile-Quantile.

##### 2.1.1.1 Résumé numérique des variables quantitatives

Sur 2111 individus observés, le Tableau 2 présente les principaux indicateurs statistiques des variables quantitatives : *age*, *taille* et *poids*.

Variables	Moyenne	Médiane	Max	Min	Ecart-type	Q1	Q3	IQR	Asymétrie	Aplatissement
age	24.31	22.78	61.00	14.00	6.35	19.95	26.00	6.05	1.53	5.82
taille	1.70	1.70	1.98	1.45	0.09	1.63	1.77	0.14	-0.01	2.44
poids	86.59	83.00	173.00	39.00	26.19	65.47	107.43	41.96	0.26	2.30
imc	29.70	28.72	50.81	13.00	8.01	24.33	36.02	11.69	0.15	2.19

Tableau 2 : Résumé numérique des variables quantitatives

L'âge des personnes interrogées varie de quatorze (14) ans à soixante un (61) ans, leurs tailles de cent quarante cinq (145) à cent quatre vingt dix huit (198) centimètres. La première moitié a moins de vingt deux (22) ans, mesure moins de cent soixante dix (170) centimètres, pèse moins de quatre vingt trois (83) kilogrammes tandis que la seconde moitié en font plus.

Parmi les individus, on constate que :

- 25% ont moins de 20 ans, mesure moins de 1.63 m et pèsent moins de 65 Kg ;
- 50% ont entre 20 et 26 ans, mesure entre de 1.63 m et 1.77 m, et pèsent entre 65 et 107 Kg ;
- 25% sont âgés de plus de 26 ans, mesure 1.77 m et pèsent plus de 107 Kg.

L'individu moyen issu de l'enquête est âgé de 24 ans, mesure 1.70 m et pèse près de 87 Kg, ce qui correspond à une IMC proche de 30 qui le classe parmi les individus en surpoids.

Selon les indicateurs de forme, asymétrie et aplatissement, il semble que les distributions de l'âge, la taille, le poids, l'IMC ne suivent pas une loi normale.

### 2.1.1.2 Graphes des variables quantitatives

La Figure 1 illustre les histogrammes des variables *age*, *taille* (en haut), et *poids*, *imc* en bas respectivement de gauche à droite.

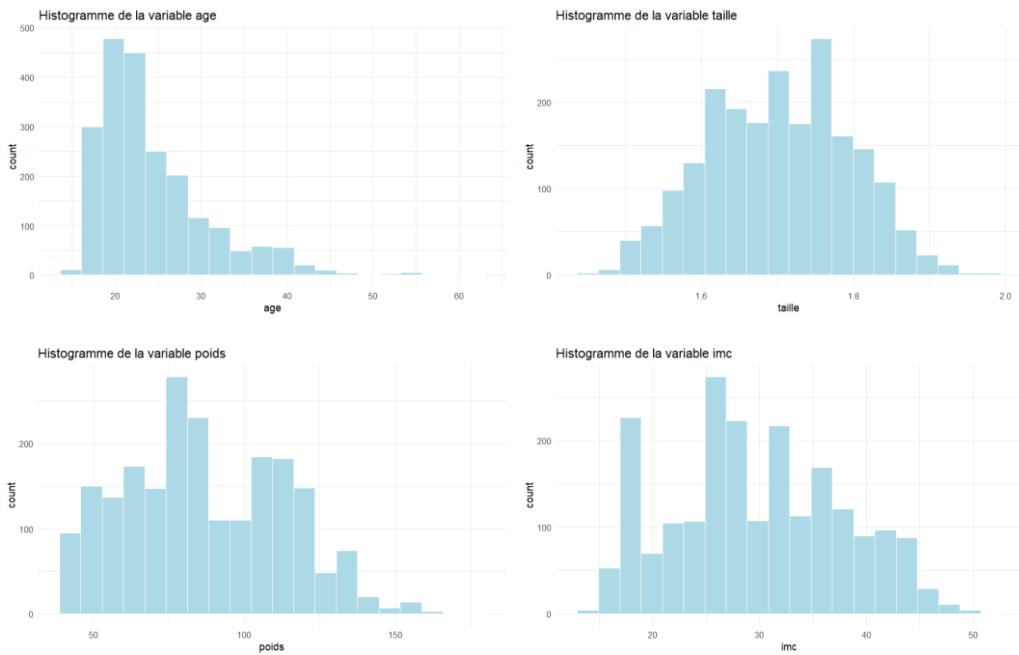


Figure 1 : Histogrammes de *age*, *taille*, *poids* et *imc*

La même interprétation que le résumé numérique se précise à travers ces histogrammes. A l'aide de l'*Annexe 2* qui illustre les courbes de densité, les boîtes à moustaches (Boxplot) et les diagrammes Quantile-Quantile (QQ Plot) des variables *age*, *taille*, *poids* et *imc*, on peut ajouter que la variable *age* comporte plusieurs valeurs extrêmes correspondant aux individus âgés de trente cinq ans et plus. Dans le même sens, on note quelques cas atypiques caractérisant les individus mesurant plus de cent quatre vingt dix centimètres (1.98 m) et pesant plus de cent soixante dix kilogrammes (170 Kg).

Les diagrammes Quantile-Quantile permettant de comparer la distribution d'une variable avec une distribution théorique (généralement la distribution normale), traduirait que les quatre variables (*age*, *taille*, *poids* et *imc*) ne suivent pas une loi normale car tous les points n'appartiennent pas à la droite en rouge, autrement les déviations par rapport à la ligne droite indiquent des écarts par rapport à la distribution normale. Cela se confirme d'ailleurs à l'aide du test de normalité de Jarque Bera qui rejette l'hypothèse nulle (distribution normale en termes d'asymétrie et d'aplatissement), la p-valeur étant < 0.05 :

- *age* : X-squared = 1519.4, df = 2, p-value < 2.2e-16 ;
- *taille* : X-squared = 28.083, df = 2, p-value = 7.979e-07 ;
- *poids* : X-squared = 66.152, df = 2, p-value = 4.33e-15 ;
- *imc* : X-squared = 66.061, df = 2, p-value = 4.552e-15.

### 2.1.2 Variables qualitatives

Le jeu de données présente en réalité beaucoup plus de variables catégorielles pour lesquelles, on analyse :

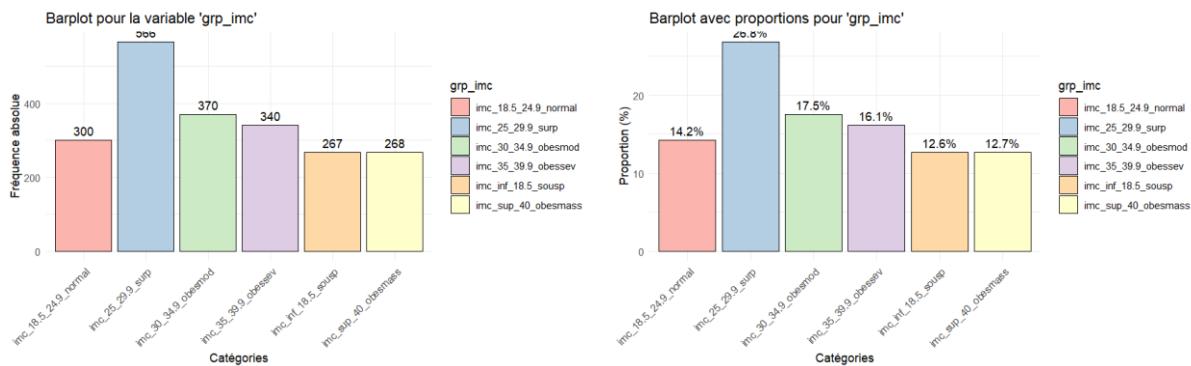
- la répartition des catégories qui permet de comprendre la diversité au sein d'une variable ;
- la fréquence des catégories à travers les diagrammes en barres (Barplots).

### 2.1.2.1 Résumé numérique des variables qualitatives

Le résumé numérique exhaustif des variables qualitatives peut être consulté à l'*Annexe 3*. On note par exemple 49% de femmes (1043) contre 51% d'hommes (1068).

### 2.1.2.2 Graphes des variables qualitatives

La *Figure 2* met en évidence la répartition des modalités de la classe d'obésité (basée sur le calcul de IMC et les groupes définis par l'OMS, variable *grp\_imc*) des individus ayant répondu à l'enquête.



*Figure 2 : Répartition en effectif (à gauche) en proportion (à droite) de la variable grp\_imc*

Elle permet de visualiser la répartition des modalités du niveau d'obésité et d'identifier la catégorie dominante ou les catégories les plus fréquentes. Il faut noter que les niveaux d'obésité renseignés dans le jeu de données sont basés sur l'indice de masse corporelle (IMC, poids divisé par taille au carré) et des informations de l'Organisation Mondiale de la Santé (OMS).

Ainsi, il en ressort qu'un peu plus 46% des individus souffrent d'obésité de différents types et ceux qui sont en surpoids représentent plus de 27%.

Que remarque-t-on de plus dans le jeu de données en analysant d'autres modalités ?

A l'aide du résumé numérique exhaustif des variables qualitatives à l'*Annexe 3* et en visualisant par exemple la répartition des modalités de quelques autres variables indiquée à l'*Annexe 4*, on observe chez les individus questionnés, que :

- 88% consomment des aliments très caloriques telles que les « fast food » ;
- plus de 95% ne surveillent pas les calories consommées ;
- environ 84% mangent souvent entre les repas, et 11.5% fréquemment ;
- 80% pratiquent très peu ou peu une activité physique ;
- 82% font partie d'une famille dont un membre a souffert ou souffre de l'obésité ;
- 75% empruntent les transports en commun, 22% en automobile, contre 2.7% qui marchent et 0.3% sont à vélo.

Ces observations impactent-elles réellement le niveau d'obésité. L'analyse bivariée nous permettra d'y répondre.

## 2.2 Analyse bivariée

### 2.2.1 Lien entre variables quantitatives

#### 2.2.1.1 Nuage de points

La *Figure 3* illustre le nuage de points entre taille et poids (à gauche) et entre poids et imc (à droite).

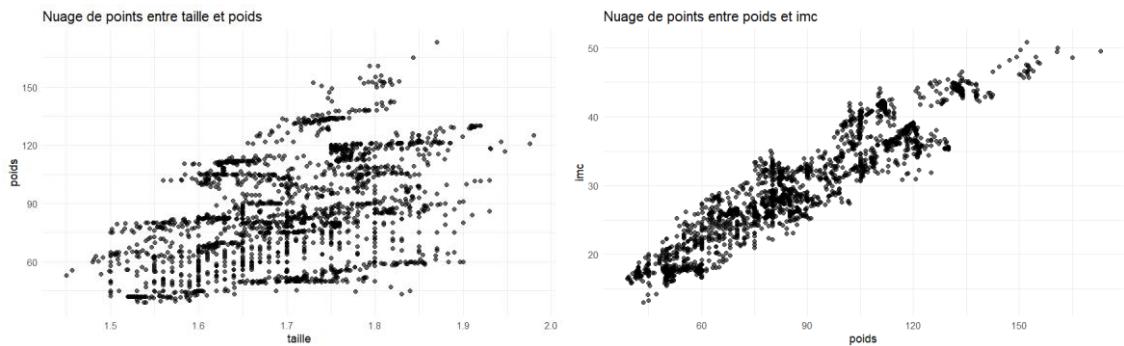


Figure 3 : Nuage de points entre taille et poids (à gauche) et entre poids et imc (à droite)

Le nuage de points à gauche illustre la relation entre la taille et le poids. Qu'observe-t-on ?

- Il y a une tendance générale où, à mesure que la taille augmente, le poids tend également à augmenter. C'est d'ailleurs en général l'intuition que l'on a.
- On note une dispersion significative des points, particulièrement pour les tailles moyennes entre 1.6 m et 1.8 m. Cela reflète une diversité de poids pour une même taille, ce qui est cohérent avec la variabilité des morphologies humaines.
- Pour des tailles plus petites (< 1.5 m), les poids sont concentrés à des niveaux plus bas.
- Quelques individus semblent avoir des poids très élevés (>150 kg) pour une taille donnée. Ils représenteraient des valeurs extrêmes.
- La majorité des points se situe dans une zone où les tailles sont entre 1.6 m et 1.8 m, avec des poids entre 50 kg et 100 kg. Cela pourrait refléter une répartition typique dans une population générale.

En conclusion, la corrélation entre la taille et le poids semble être positive, mais elle n'est pas parfaite ; le calcul de la corrélation aidera à quantifier la force de cette relation.

Quant au nuage de points à droite, il montre la relation entre le poids et l'indice de masse corporelle (IMC). Que peut-on en dire ?

- On observe que la disposition des points montre une tendance quasi-linéaire, claire où l'IMC augmente avec le poids. Cela indique une relation positive qui semble très forte entre ces deux variables. Plus une personne pèse lourd, plus son IMC est élevé, ce qui est attendu puisque l'IMC dépend directement du poids.
- Malgré cette tendance générale, il y a une dispersion autour de la droite suggérant que d'autres facteurs (comme la taille) influencent aussi l'IMC.
- Dans les poids très élevés (> 120 kg), figurent quelques points isolés qui pourraient indiquer des cas atypiques comme une personne avec un IMC plus bas qu'attendu pour son poids.

### 2.2.1.2 Corrélation entre les variables

La Figure 4, appelée « heatmap de corrélation » montre la corrélation entre les variables quantitatives.

Remarque : les variables autres que *age*, *taille*, *poids* et *imc* comprennent des valeurs de type échelle et il est pertinent de les traiter comme des variables catégorielles ; **il ne faut donc prêter attention qu'aux variables *age*, *taille*, *poids* et *imc* sur cette figure.**

On se concentre ici sur la relation qu'existe d'une part entre la taille et le poids, et d'autre part entre le poids et l'IMC.

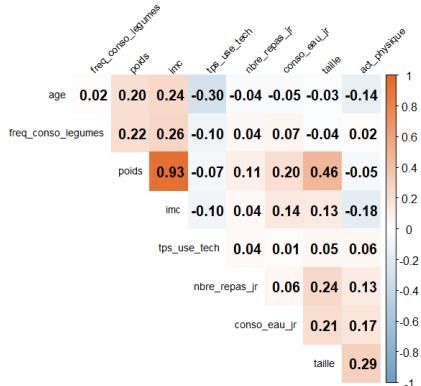


Figure 4 : Heatmap de corrélation

Les tests d'hypothèses peuvent-ils confirmer ces premières intuitions ?

### 2.2.1.3 Tests d'hypothèses sur les relations bivariées entre les variables quantitatives

On a constaté précédemment que les variables *age*, *taille*, *poids* et *imc* ne suivent pas une distribution Gaussienne mais compte tenu du grand nombre d'observations (2111), on peut supposer la condition de normalité remplie pour réaliser un test de corrélation de Pearson (coefficient de corrélation  $\rho$ , et p-valeur  $p$ ).

Le test permet de rejeter l'hypothèse nulle  $H_0$  (absence de corrélation linéaire) et d'accepter l'hypothèse alternative  $H_1$  (présence de corrélation linéaire) pour les relations :

- entre *poids* et *imc* :  $\rho = 0.93$ , avec  $p = 0$  ;
- entre *taille* et *poids* :  $\rho = 0.46$ , avec  $p = 1.02e-112$  ;
- entre *age* et *imc* :  $\rho = 0.24$ , avec  $p = 5.01e-30$  ;
- entre *age* et *poids* :  $\rho = 0.20$ , avec  $p = 5.52e-21$  ;
- entre *taille* et *imc* :  $\rho = 0.13$ , avec  $p = 1.22e-09$ .

En conclusion, il existe bien une relation linéaire très forte entre le poids et l'IMC (corrélation positive quasi-parfaite), faible entre la taille et le poids et entre l'âge et l'IMC. On ajoute que le coefficient de corrélation de Pearson pour les combinaisons de variables (*age* et *imc*, *age* et *poids*, *taille* et *imc*) est très faible et malgré leurs p-valeur bien inférieures à 0.05, une relation autre que linéaire peut être envisagée.

Cette figure montre que :

- la corrélation entre taille et poids est positive mais faible ( $\rho = 0.46$ , donc  $<0.5$ ) et,
- la corrélation entre poids et imc est positive et très forte ( $\rho = 0.93$ ) ce qui accentue l'intuition obtenue avec le nuage de points : relation linéaire entre poids et imc.

## 2.2.2 Lien entre variables quantitatives et qualitatives

### 2.2.2.1 Résumé numérique par modalité des variables qualitatives

Au vu nombre de pages limité, le présent rapport s'intéresse à la liaison entre les variables qualitatives et l'IMC ; le résumé numérique bivarié entre l'IMC et les variables qualitatives est indiqué à l'*Annexe 5*. Pour toutes les paires de variables qualitative quantitative, se reporter au fichier Quarto.

### 2.2.2.2 Boîtes à moustaches bivariées

Les boîtes à moustaches bivariées entre l'IMC et les variables qualitatives peuvent être observées à l'*Annexe 6*. Se reporter au fichier Quarto pour visualiser l'ensemble des boîtes à moustaches bivariées du jeu de données

En se basant sur les médianes, la hauteur et les outliers des boîtes représentant les différentes modalités, le jeu de données traduit les intuitions suivantes.

- Forte tendance :

Les antécédents familiaux (membre de famille ayant souffert ou souffrant d'obésité), la consommation fréquente des aliments très caloriques, la quantité quotidienne d'eau consommée, la surveillance des calories consommées, la pratique d'activité physique et les moyens de transport empruntés, le tabagisme auraient une influence remarquable sur l'IMC et donc sur le niveau d'obésité. On note quand même quelques cas atypiques : des individus qui se retrouvent dans l'obésité sévère alors qu'ils ne consomment pas fréquemment les aliments très caloriques, surveillent leurs calories consommées, marchent ou vont à vélo.

- Quelques particularités :

L'influence du genre sur l'IMC serait négligeable ; les hommes sont légèrement plus corpulents que les femmes et on observe une grande variabilité de l'IMC (entre 22 et 40) chez les femmes tandis que celle des hommes se situe entre 25 et 35.

Les individus qui consomment souvent (avec une forte variabilité) l'alcool auraient une IMC médiane plus élevée que ceux qui en consomment fréquemment, toujours ou pas du tout ; ce qui permettrait de dire que l'alcool ne représenterait pas le facteur principal d'obésité. On pourrait faire le parallèle avec la fréquence de consommation de légumes (l'IMC médiane des individus qui consomment toujours les légumes est la plus élevée), le nombre de repas par jour (l'IMC médiane des individus mangeant plus de 3 fois par jour est la plus faible), le fait de manger entre les repas (l'IMC médiane des individus qui ne mangent pas entre les repas est supérieure à celle de ceux qui mangent fréquemment ou toujours entre les repas mais inférieure à celle de ceux qui ont souvent ce comportement et varie fortement d'ailleurs).

Contrairement à ce que l'on pense, le temps passé devant les écrans ou les appareils technologiques n'aurait pas d'impact sur l'IMC car on remarque plutôt que la médiane des individus qui passent beaucoup plus de temps est inférieure aux autres et on observe une grande variabilité chez ceux qui passent moins de temps. Cela pourrait s'expliquer par d'autres facteurs comme l'âge (*et la catégorie socio professionnelle non présente dans le jeu de données*)

Que peut-on conclure avec les tests d'hypothèses de relations bivariées entre les variables quantitatives et qualitatives ?

### 2.2.2.3 Tests d'hypothèses sur les relations bivariées entre les variables quantitatives et qualitatives

A l'Annexe 7, les tests statistiques (*test Student pour 2 modalités et test ANOVA pour plus de 2 modalités, respectivement Wilcoxon et Kruskal-Wallis pour les tests non paramétriques*) montrent et confirment que les distributions de l'IMC sont différentes en fonction des modalités des variables qualitatives en raison des p-value < 0,05 sauf pour les variables « sexe » et « fume ». Il existe donc une relation entre l'IMC et les variables qualitatives « ant\_fam\_obesite », « conso\_freq\_alim\_cal », « mange\_entre\_repas », « surveille\_cal\_conso », « freq\_conso\_alcool », « moy\_trans », « grp\_freq\_conso\_legumes », « grp\_age », « grp\_nbre\_repas\_jr », « grp\_conso\_eau\_jr », « grp\_act\_physique », « grp\_tps\_use\_tech ».

## 2.2.3 Lien entre variables qualitatives

### 2.2.3.1 Diagrammes en barre

Les diagrammes en barre figurant à l'Annexe 8 illustrent les liaisons entre les variables qualitatives. Se reporter au fichier Quarto pour visualiser l'ensemble des diagrammes en barre des paires de variables qualitatives.

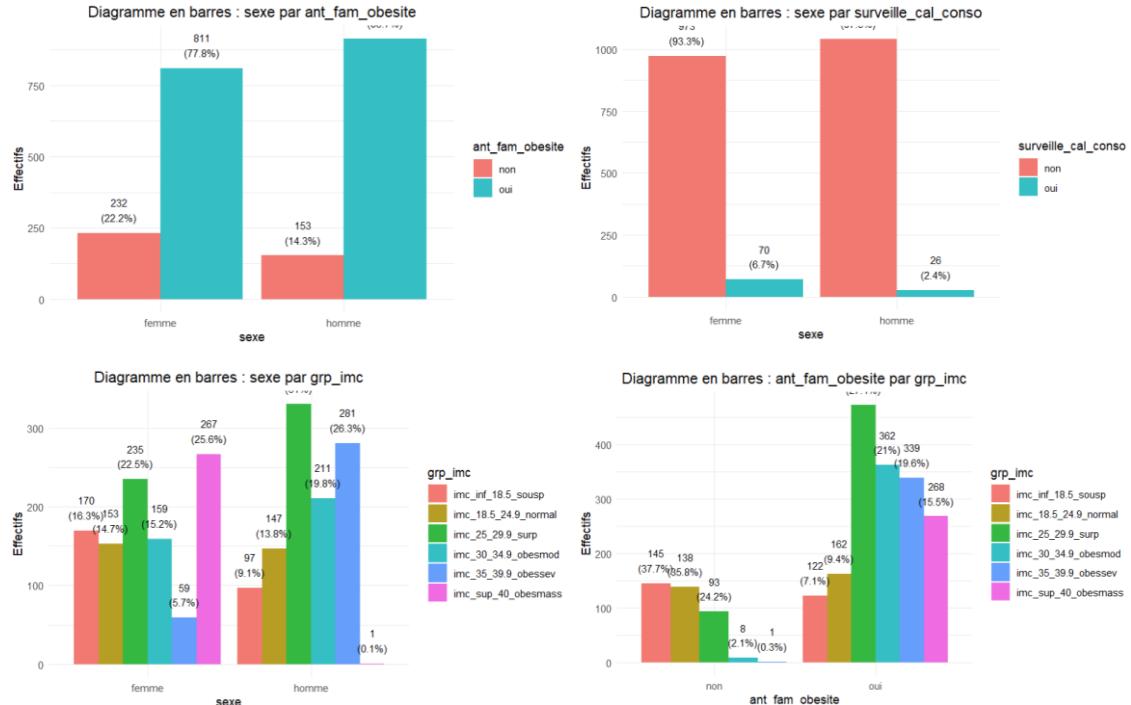


Figure 5 : Barplots bivariés de quelques paires de variables qualitatives

On note sur les graphes ci-dessus que la proportion d'hommes (85.7%) appartenant à une famille dont un membre souffrant ou ayant souffert d'obésité est supérieure à celle des femmes

(77.8%) et que ces dernières (femmes 6.7%) surveillent plus leur consommation de calories que les hommes (2.4%).

Il est également observé que les types d'obésité « modéré » et « sévère » touchent plus les hommes (respectivement 19.8% et 26.3%) que les femmes (respectivement 15.2% et 5.7%) ; à contrario, l'obésité « massive » atteint très largement les femmes (25.6%) que les hommes (0.1%).

Enfin, le graphe en bas à droite indique nettement que les individus touchés par les trois groupes d'obésité ont un antécédent familial d'obésité.

#### 2.2.3.2 Tests d'hypothèses sur les relations bivariées entre les variables qualitatives

L'ensemble des tests d'hypothèses entre toutes les paires de variables qualitatives peut-être lu dans le fichier Quarto.

L'*Annexe 9* relève particulièrement les tests d'hypothèses entre la variable « grp\_imc » et les autres variables qualitatives et on y observe l'existence de liaison à travers le test du Khi-2, la p-value < 0.05, et le V de Cramer qui mesure une intensité plus ou moins faible sur ces relations. Trois exemples ci-après :

```

o Analyse entre sexe et grp_imc
[1] "Table de contingence :"

    imc_inf_18.5_sousp imc_18.5_24.9_normal imc_25_29.9_surp imc_30_34.9_obesmod imc_35_39.9_obessev imc_sup_40_obesmass
femme      170             153            235           159            59             267
homme      97              147            331           211            281             1

Test du Khi-2 : X² = 452.4048, p-value = 0, V de Cramer = 0.4629

o Analyse entre ant_fam_obesite et grp_imc
[1] "Table de contingence :"

    imc_inf_18.5_sousp imc_18.5_24.9_normal imc_25_29.9_surp imc_30_34.9_obesmod imc_35_39.9_obessev imc_sup_40_obesmass
non        145             138            93             8              1              0
oui         122             162            473            362            339            268

Test du Khi-2 : X² = 586.5665, p-value = 0, V de Cramer = 0.5271
-----
o Analyse entre sexe et conso_freq_alim_cal
[1] "Table de contingence :"

    non oui
femme 143 900
homme 102 966

Test du Khi-2 : X² = 8.4999, p-value = 0.0036, V de Cramer = 0.0649

```

## 3 Analyse factorielle : Analyse en Correspondances Multiples (ACM)

### 3.1 Choix des variables actives

La présente étude se base sur un jeu de données qui relève plus de variables catégorielles. Le choix des variables actives porte naturellement sur toutes les variables qualitatives y compris celles qui ont été transformées. Ainsi, on a les variables actives : *sexe, ant\_fam\_obesite, conso\_freq\_alim\_cal, mange\_entre\_repas, fume, surveille\_cal\_conso, freq\_conso\_alcool, moy\_trans, grp\_imc, grp\_freq\_conso\_legumes, grp\_age, grp\_nbre\_repas\_jr, grp\_conso\_eau\_jr, grp\_act\_physique, grp\_tps\_use\_tech*.

Par ailleurs, les variables supplémentaires peuvent être : *imc*, *poids*, et/ou *age*.

### 3.2 Choix du nombre de dimensions

Le « screeplot » aide à déterminer le nombre de dimensions. On observe un premier coude (cassure) sur la 3ème dimension et un second à la 8<sup>ème</sup> dimension qui permet de choisir 8 dimensions malgré le cumul des valeurs propres dont le pourcentage d'inertie (variabilité exprimée) équivaut à 41% (cf. le tableau des valeurs propres à l'Annexe 10), ce qui paraît faible, mais plutôt considérable par rapport au nombre total d'axes (34 axes).

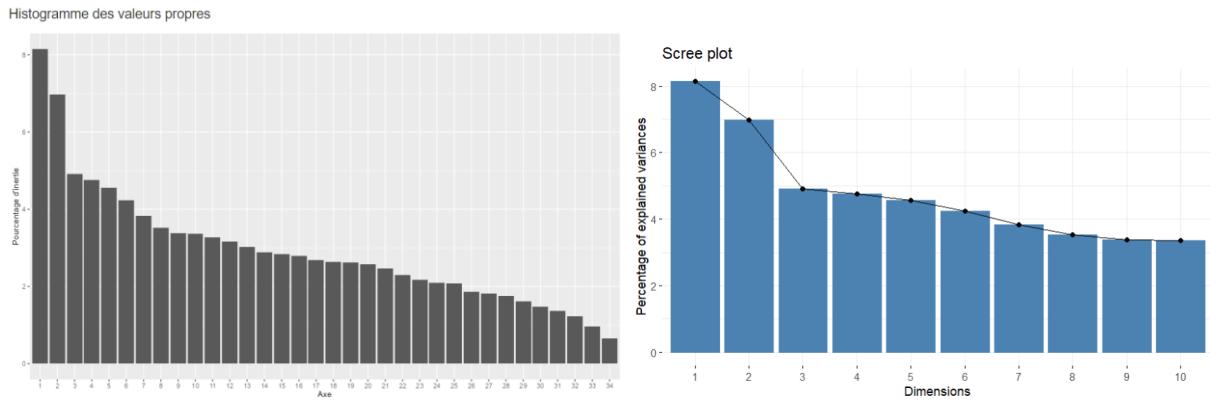


Figure 6 : Scree plot

Pour le présent rapport, on s'intéressera à l'interprétation des deux premières dimensions, éventuellement les axes 3 et 4.

### 3.3 Interprétation de l'ACM

L'exploration de l'ACM peut s'effectuer dans R, avec la commande : `explor::explor(acm)`

#### 3.3.1 Interprétation des graphes des variables

La Figure 7 montre le graphe des variables sur les axes 1 et 2 (à gauche), axes 3 et 4 (à droite), chaque modalité des variables qualitatives représentée comme un point. Les deux premiers axes sont les axes factoriels (à gauche), qui expliquent une part maximale de la variabilité des données, 8.15% sur l'axe 1 et 6.98% sur l'axe 2. Les distances et positions relatives des points permettent d'interpréter des proximités, oppositions, et groupements de modalités.

On verra que ce graphe permettrait de

- identifier des profils types, par exemple un profil « jeune, actif, mange équilibré » vs. « adulte, sédentaire » ;
- interpréter les liens entre modalités : ce sont celles proches les unes des autres ;
- repérer les modalités les plus discriminantes : les plus éloignées du centre et de grande taille.

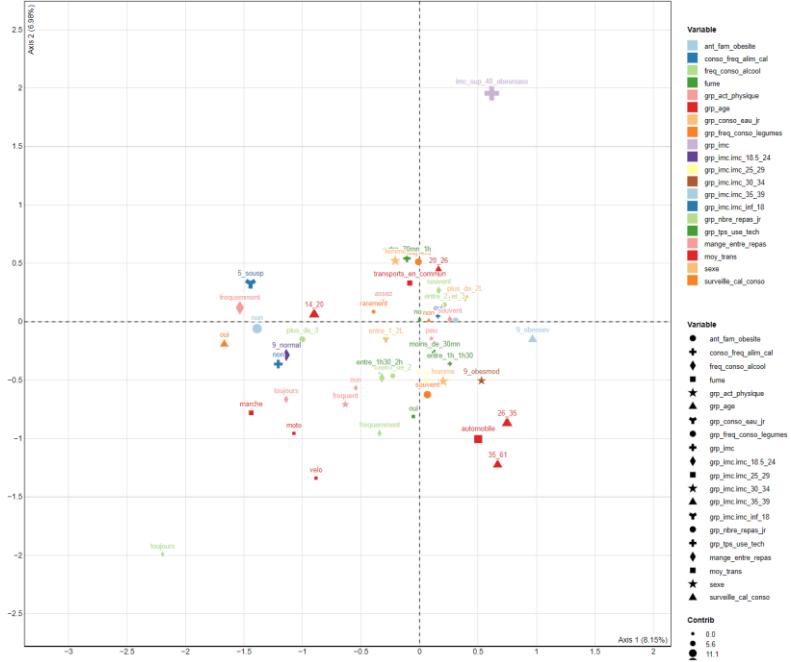


Figure 7 : Graphe des variables sur les axes 1 et 2

L'axe 1 semble opposer :

- à gauche, des comportements plutôt « positifs » : l'absence d'antécédent familial d'obésité (ant\_fam\_obesite.non), la non consommation d'aliments très caloriques (conso\_freq\_alim\_cal.non), la surveillance de calories (surveille\_cal\_conso.oui), la pratique fréquente d'activité physique (grp\_act\_physique.frequent), marche, vélo, corpulence normale (grp\_imc.imc\_18.5\_24.9\_normal), public jeune de 14 à 20 ans (grp\_age.14\_20), etc.
- à droite, des comportements plus sédentaires, à risque : corpulence d'obésités modérée et sévère (grp\_imc.imc\_30\_34\_obesmod , grp\_imc.imc\_35\_39\_obesev), la consommation d'aliments très caloriques (conso\_freq\_alim\_cal.oui), la non surveillance de calories (surveille\_cal\_conso.non), la faible pratique d'activité physique (grp\_act\_physique.peu), public plutôt plus âgé 26 à 35 ans et 35 à 61 ans, ayant comme moyen de transport l'automobile, etc.

En somme, l'axe 1 opposerait des profils plutôt actifs ou équilibrés (gauche) à des profils plus sédentaires ou avec des comportements liés à l'obésité (droite).

Quant à l'axe 2, il semble opposer :

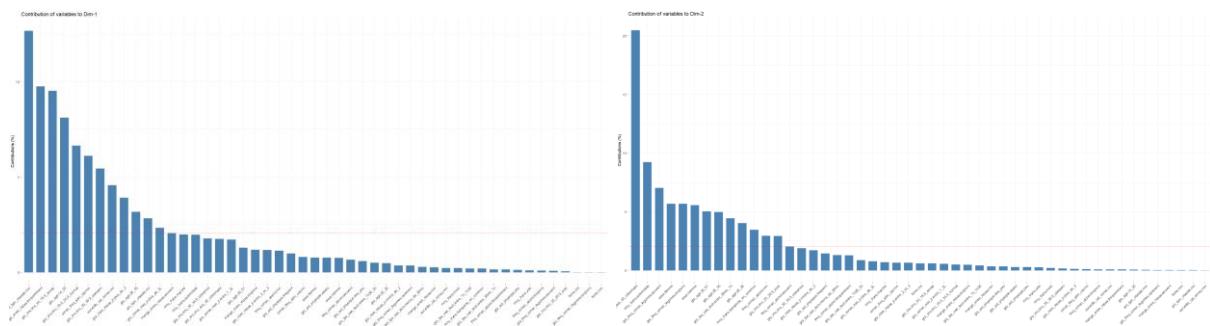
- En haut, la modalité (moy\_trans.transports\_en\_commun) proche du centre (l'origine) où se concentre la moyenne et le groupe d'obésité massive (grp\_imc.imc\_sup\_40\_obesmass) isolé et très éloigné du centre, ce qui s'apparente à une modalité très discriminante, probablement rare mais fortement associée à un profil spécifique.
- En bas, des modalités marche, moto, vélo situées à gauche rappelant des profils plutôt très actifs et sains et à droite la modalité automobile exprimant un profil sédentaire, d'ailleurs plus proche des modalités caractérisant les groupes d'obésité modérée et sévère (grp\_imc.imc\_30\_34\_obesmod , grp\_imc.imc\_35\_39\_obesev).

A noter que, le groupe d'obésité massive (`grp_imc.imc_sup_40_obesmass`) isolé et très éloigné du centre s'oppose à la modalité `freq_conso_alcool.toujours` située en bas à gauche et également très éloigné du centre ; cela pourrait traduire que la consommation fréquente d'alcool n'induit pas toujours à l'obésité et représenterait des cas atypiques.

Au final, l'axe 2 pourrait représenter un gradient de poids ou de sédentarité extrême : du plus sain (bas) au plus problématique (haut, obésité sévère).

Le graphe des variables sur les axes 3 et 4 se trouve à l'Annexe 11.

La *Figure 8* ci-dessous s'intéresse plutôt à la contribution des variables (modalités) sur les deux premiers axes. On constate par exemple que les modalités (par ordre décroissant) suivantes ont contribué à la construction du premier axe : `ant_fam_obesite.non`, `mange_entre_repas.frequemment`, `grp_imc.imc_inf_18.5_sousp`, `grp_age.14_20`, `grp_imc.imc_18.5_24.9_normal`, `conso_freq_alim_cal.non`, `grp_imc.imc_35_39.9_obesev`, `surveille_cal_conso.oui`, `grp_nbre_repas_jr.plus_de_3`, `grp_age.26_35`, `ant_fam_obesite.oui`, `grp_conso_eau_jr.plus_de_2L`, `mange_entre_repas.souvent`.



*Figure 8 : Contribution des modalités sur les axes 1 (à gauche) et 2 (à droite)*

La contribution des variables (modalités) sur les axes 3 et 4 peut être observée à l'Annexe 12.

### 3.3.2 Interprétation des graphes des individus

La Figure 9 présente le graphe des individus sur les axes 1 et 2, colorés en fonction du groupe d'IMC (à gauche) et sexe (à droite). Le fichier Quarto, notamment la commande « `explor::explor(acm)` » permet d'observer plus de cas à travers plusieurs visualisations.

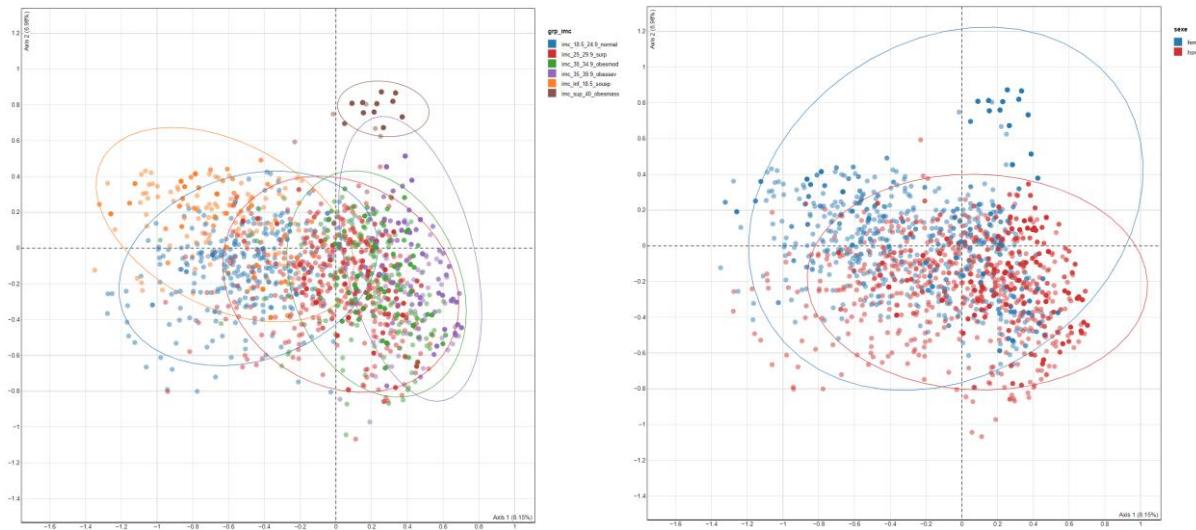


Figure 9 : Graphe des individus sur les axes 1 et 2, coloré en fonction du groupe IMC (à gauche) et sexe (à droite)

Sur le premier axe, les individus qui surveillent leur consommation calorique sont situés plus à gauche et ceux qui ne la surveillent pas sont plus répartis sur cet axe. On fait le parallèle avec les individus qui consomment des aliments très caloriques sachant que ceux consommant moins d'aliments très caloriques se trouvent plus à gauche, et les individus ayant un antécédent familial d'obésité sachant que ceux qui n'en présentent pas se retrouvent aussi à gauche. Également, les individus pratiquant fréquemment une activité physique figurent à gauche et les autres sont dispersés sur cet axe (plus de variabilité). Du point de vu du mode de transport, les individus qui marchent ou qui vont à vélo sont situés à gauche de l'axe et ceux qui sont en automobile sont plus concentrés à droite.

Le lien avec le graphe des variables se ferait en distinguant les individus sains (corpulence normale) des individus souffrant d'obésité ou en surpoids.

Sur le second axe, on distinguerait mieux par exemple le niveau d'obésité entre les hommes et les femmes chez qui on note plus l'obésité massive.

### 3.4 Conclusion sur l'ACM

L'ACM a permis de visualiser les modalités proches souvent associées et les modalités éloignées du centre qui relèvent des traits atypiques, spécifiques et discriminant.

Sur l'axe 1, on observe à gauche un comportement actif et plus ou moins sain contrairement à un comportement sédentaire ou à risque.

Et sur l'axe 2, on observe en haut l'IMC très élevé (obésité massive) et en bas les types d'obésité modérée et sévère d'ailleurs proche des profiles adultes se déplaçant en automobile.

## 4 Classification non supervisée : Classification Ascendante Hiérarchique (CAH)

### 4.1 Mise en œuvre CAH

La *Figure 10* montre le dendrogramme issu de la distance de Ward (à gauche) et le saut d'inertie (à droite) permettant et aidant au choix du nombre de classes.

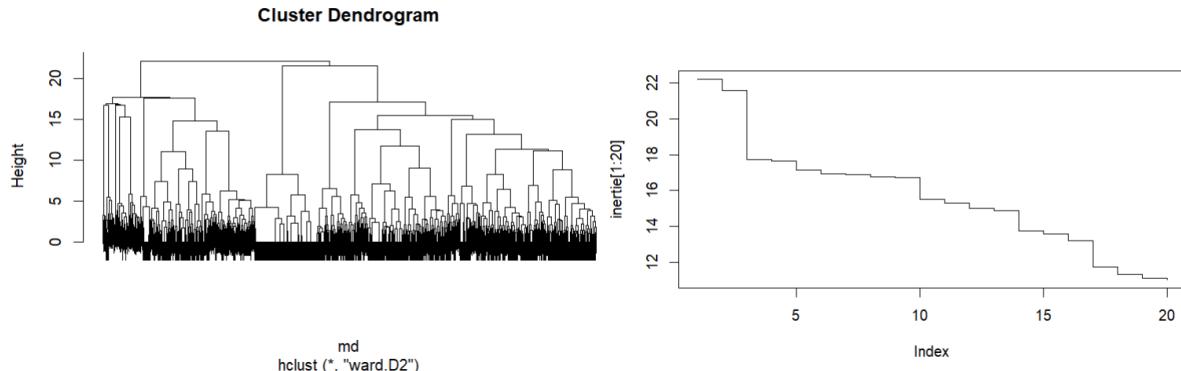


Figure 10 : Dendrogramme issu de la distance de Ward (à gauche), saut d'inertie (à droite)

Ici, on observe que le troisième saut d'inertie est plus haut que les autres, ce qui nous permettrait de choisir trois classes.

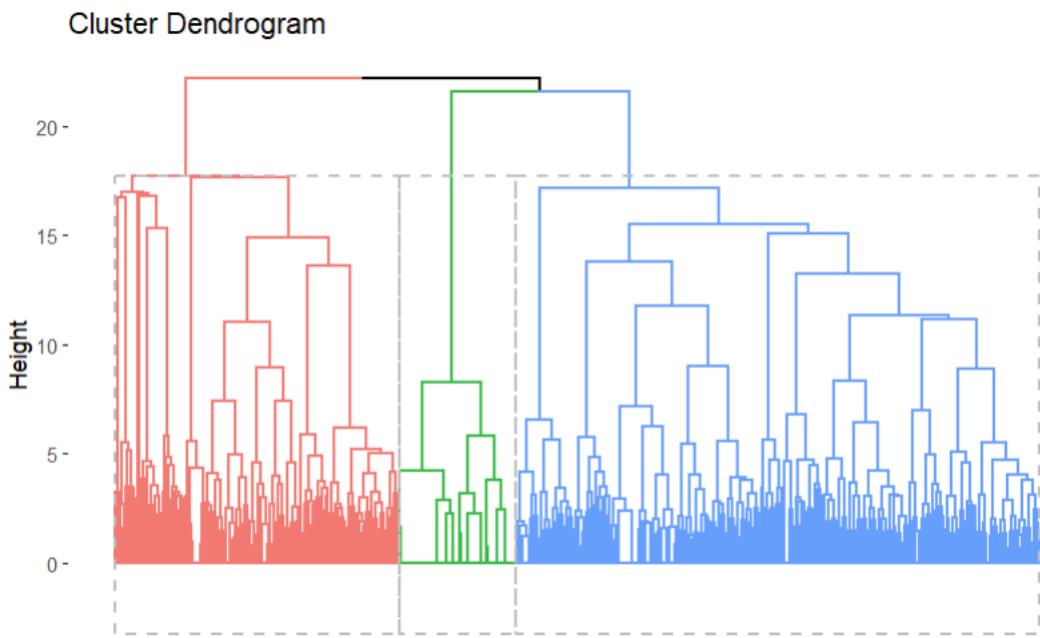


Figure 11 : Dendrogramme représentant le nombre de classe basé sur le saut d'inertie

A partir des plans de l'ACM, sont décrits ci-après les individus suite à la classification.

## 4.2 Description des groupes à partir des individus

La *Figure 12* montre trois classes d'individus.



*Figure 12 : Classes des individus*

Ce graphe représente les individus projetés sur les deux premiers axes factoriels de l'ACM (Dim1 et Dim2, expliquant respectivement 8.1% et 7% de l'inertie totale). Chaque point est un individu, coloré selon son appartenance à l'un des 3 groupes identifiés par la classification.

Groupe 1 (rouge)

- Réparti principalement à gauche de l'axe Dim1, avec une dispersion relativement étendue sur Dim2 ;
- Représente un groupe assez hétérogène, car les points sont épars ;
- Il pourrait représenter plus le profil "sain" de la population, public de corpulence normale.

Groupe 2 (vert)

- centré autour de l'origine, couvrant une grande surface autour du centre du plan ;
- c'est le groupe le plus nombreux et le plus dispersé, ce qui suggère une population atteinte d'obésité types modéré et sévère (par rapport aux dimensions retenues).

Groupe 3 (bleu)

- Très concentré et distinct en haut à droite du plan ( $\text{Dim1} > 0$  et  $\text{Dim2} > 0$ ).
- Ce groupe est clairement séparé des deux autres, ce qui indique qu'il possède des caractéristiques très spécifiques, propres à l'obésité massive qu'on a constaté plus tôt chez les femmes.

La compacité de ce groupe montre une forte homogénéité interne, ce qui peut en faire un groupe particulièrement intéressant à étudier.

### 4.3 Conclusion sur Classification non supervisée

Suite à l'ACM, la CAH nous a conduit à l'identification de trois classes issues du graphe de saut d'inertie. Ces classes mettent en évidence les individus de corpulence plus ou moins normale, ceux souffrant d'obésité de niveaux modéré et sévère et le groupe atypique de femmes atteintes d'obésité massive.

## 5 Conclusion

La présente étude de cas m'a permis de mettre en pratique toutes les méthodes descriptives vues au cours STA101 pour répondre aux questions posées au chapitre 1.2 Problématiques, les réponses étant décrites au fur et à mesure des analyses descriptives effectuées.

Il existe bel et bien un lien entre le niveau d'obésité et le poids, le fait qu'un membre de la famille ait souffert ou souffre d'obésité peut déterminer le risque d'obésité chez l'individu étudié, la consommation fréquente d'aliments très caloriques comme les « fast food » favorise l'obésité. La surveillance des calories consommées, la pratique d'activité physique et le moyen de transport principal représentent des facteurs pouvant influencer sur la corpulence. La fréquence de consommation d'alcool seule influe peu sur l'obésité, de même que le temps passé à utiliser des appareils électroniques.

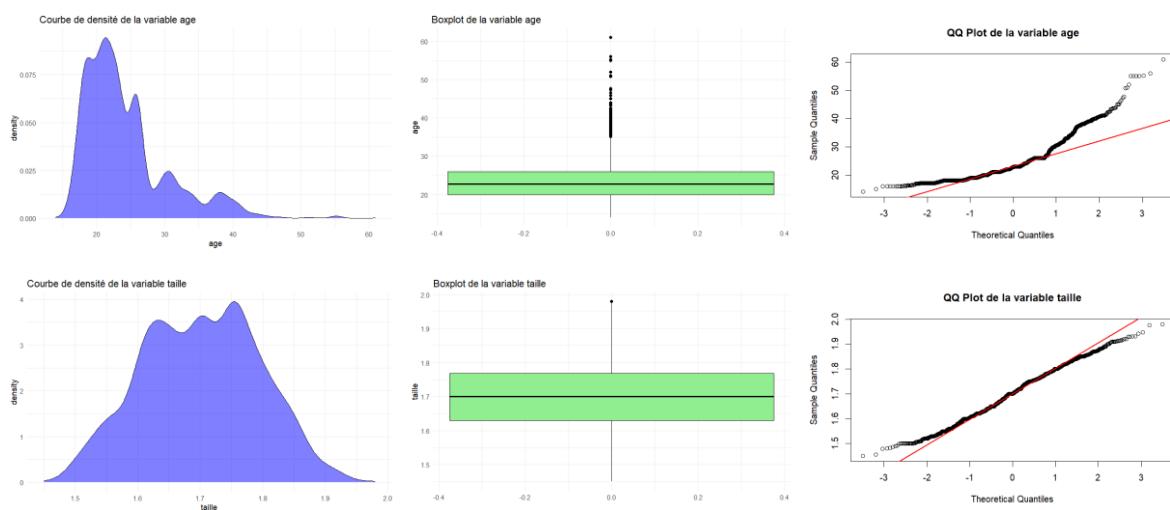
L'analyse factorielle, confirme les liaisons remarquées au cours de l'analyse bivariée et ressort la proximité entre les variables et les ressemblances entre individus. Ce qui a permis de résumer les individus en trois classes au moyen de la classification ascendante hiérarchique.

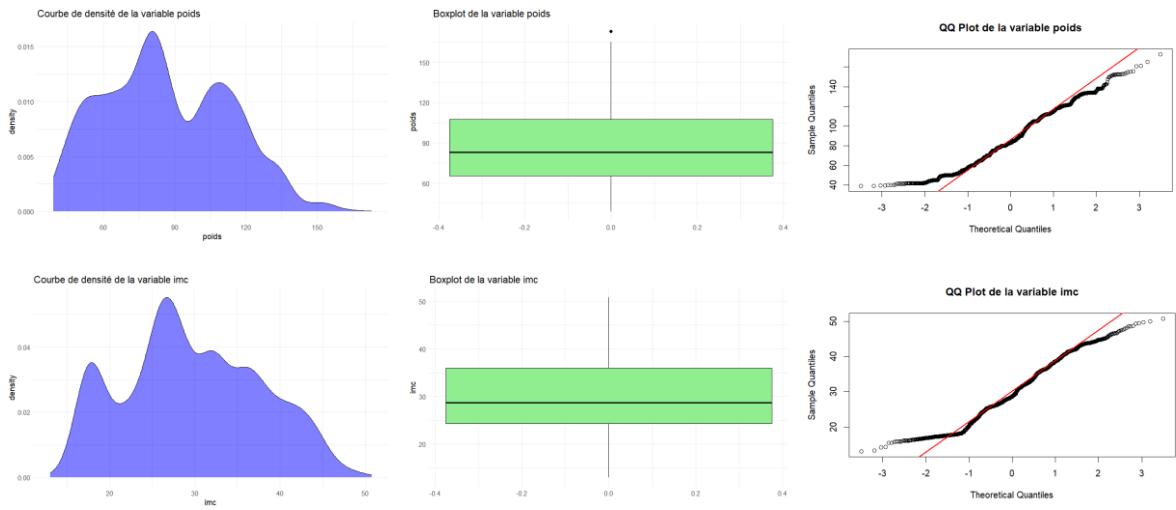
## Annexes

Annexe 1 : Tableau des variables et leurs types/modalités à l'origine

N°	Variables (Kaggle)	Types (Kaggle)	Type Variables/modalités
1	Gender	Categorical	• Female • Male
2	Age	Continuous	Numeric value
3	Height	Continuous	Numeric value in meters
4	Weight	Continuous	Numeric value in kilograms
5	family_history_with_overweight	Binary	• yes • no
6	FAVC	Binary	• yes • no
7	FCVC	Integer	Echelle 1 à 3
8	NCP	Continuous	Echelle 1 à 4
9	CAEC	Categorical	• no • Sometimes • Frequently • Always
10	SMOKE	Binary	• yes • no
11	CH2O	Continuous	Echelle 1 à 3
12	SCC	Binary	• yes • no
13	FAF	Continuous	Echelle 0 à 3
14	TUE	Integer	Echelle 0 à 2
15	CALC	Categorical	• no • Sometimes • Frequently • Always
16	MTRANS	Categorical	• Automobile • Motorbike • Bike • Public_Transportation • Walking
17	NOObeyesdad	Categorical	• Insufficient_Weight • Normal_Weight • Overweight_Level_I • Overweight_Level_II • Obesity_Type_I • Obesity_Type_II • Obesity_Type_III

Annexe 2 : Courbe de densité (à gauche), Boxplot (au milieu) et QQ Plot (à droite) de age, taille, poids et imc



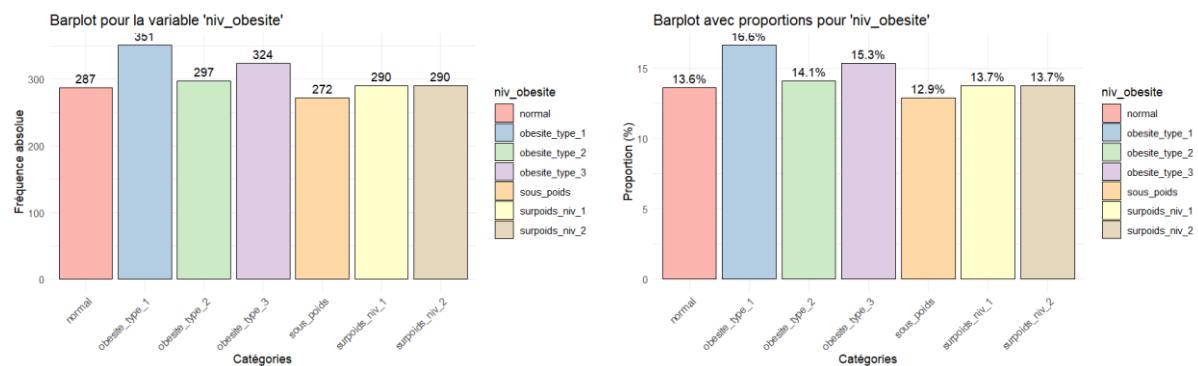


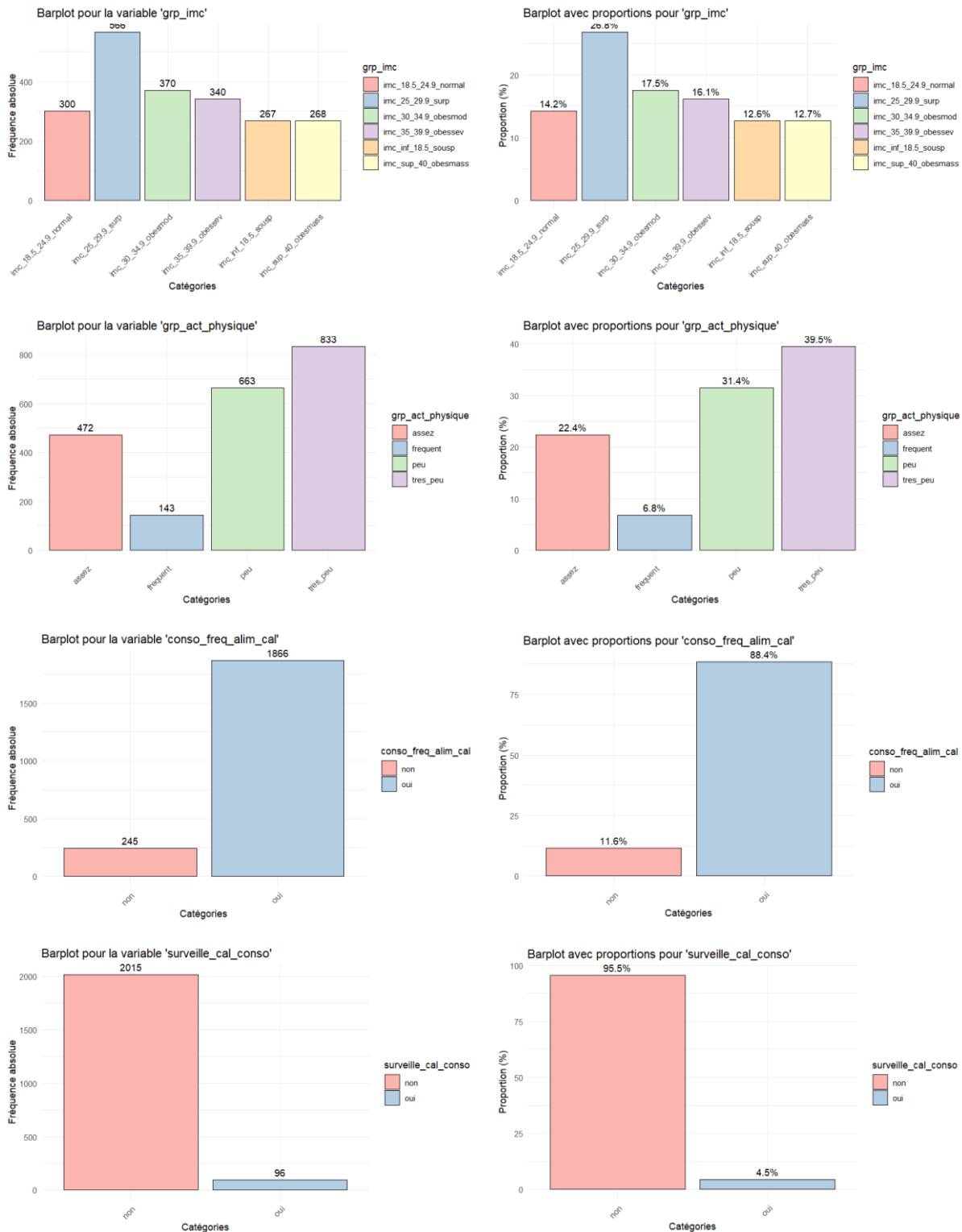
### Annexe 3 : Résumé numérique des variables qualitatives

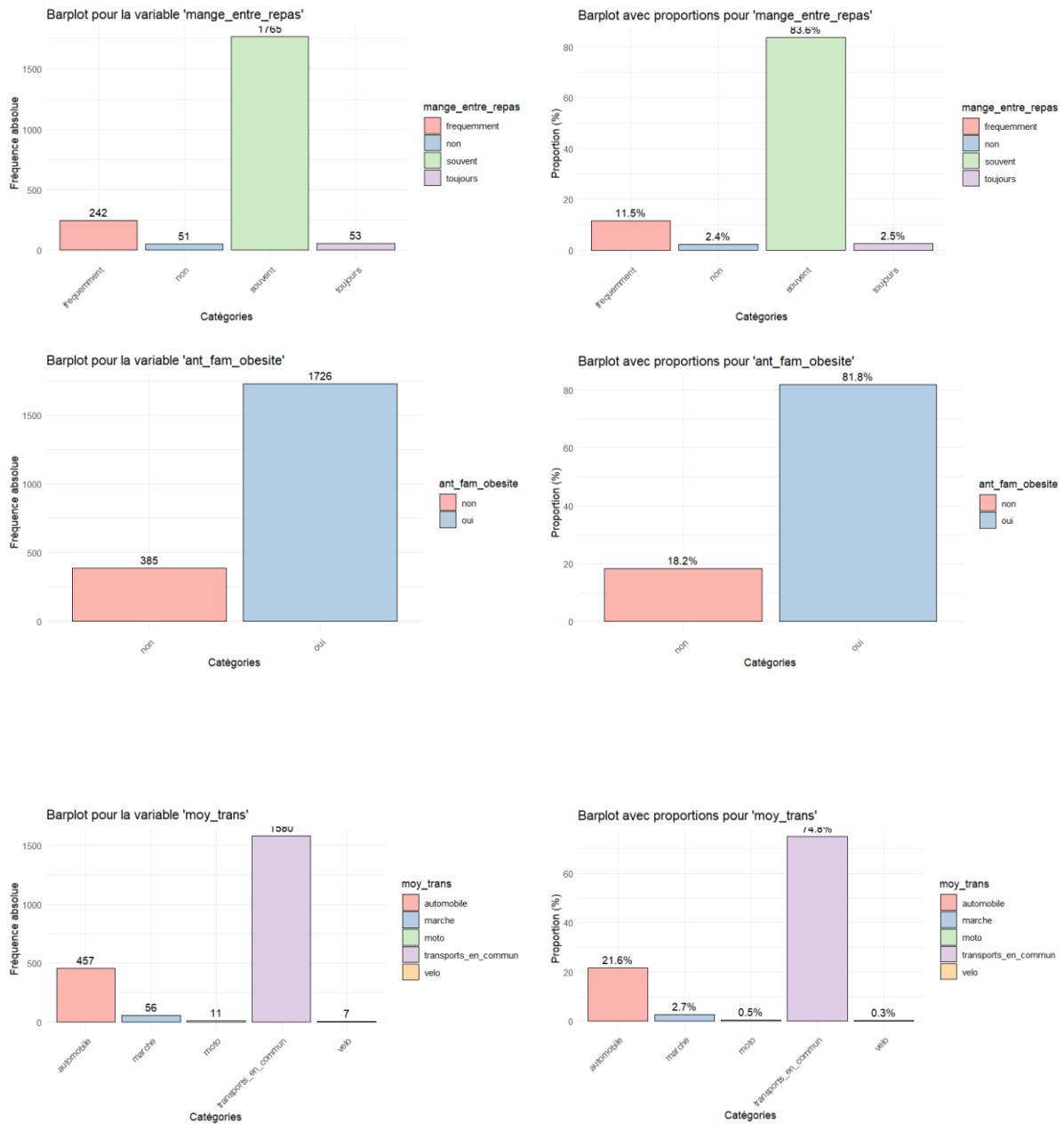
Variables	Catégorie	Fréquence.absolue	Proportion	Mode
sexe	femme	1043	49.41	homme
	homme	1068	50.59	<NA>
ant_fam_obesite	non	385	18.24	oui
	oui	1726	81.76	<NA>
conso_freq_alim_cal	non	245	11.61	oui
	oui	1866	88.39	<NA>
mange_entre_repas	toujours	53	2.51	souvent
	frequemment	242	11.46	<NA>
	non	51	2.42	<NA>
	souvent	1765	83.61	<NA>
fume	non	2067	97.92	non
	oui	44	2.08	<NA>
surveille_cal_conso	non	2015	95.45	non
	oui	96	4.55	<NA>
freq_conso_alcool	toujours	1	0.05	souvent
	frequemment	70	3.32	<NA>
	non	639	30.27	<NA>
	souvent	1401	66.37	<NA>
moy_trans	automobile	457	21.65	transports_en_commun
	velo	7	0.33	<NA>
	moto	11	0.52	<NA>
	transports_en_commun	1580	74.85	<NA>
	Marche	56	2.65	<NA>
niv_obesite	sous_poids	272	12.88	obesite_type_1
	normal	287	13.6	<NA>
	obesite_type_1	351	16.63	<NA>
	obesite_type_2	297	14.07	<NA>
	obesite_type_3	324	15.35	<NA>

	surpoids_niv_1	290	13.74	<NA>
	surpoids_niv_2	290	13.74	<NA>
grp_imc	imc_inf_18.5_sousp	267	12.65	imc_25_29.9_surp
	imc_18.5_24.9_normal	300	14.21	<NA>
	imc_25_29.9_surp	566	26.81	<NA>
	imc_30_34.9_obesmod	370	17.53	<NA>
	imc_35_39.9_obessev	340	16.11	<NA>
	imc_sup_40_obesmass	268	12.70	<NA>
grp_freq_conso_legumes	rarement	131	6.21	toujours
	souvent	903	42.78	<NA>
	toujours	1077	51.02	<NA>
grp_age	14_20	585	27.71	20_26
	20_26	1028	48.7	<NA>
	26_35	330	15.63	<NA>
	35_61	168	7.96	<NA>
grp_nbre_repas_jr	moins_de_2	395	18.71	entre_2_et_3
	entre_2_et_3	1488	70.49	<NA>
	plus_de_3	228	10.8	<NA>
grp_conso_eau_jr	entre_1_2L	1217	57.65	entre_1_2L
	plus_de_2L	894	42.35	<NA>
grp_act_physique	tres_peu	833	39.46	tres_peu
	peu	663	31.41	<NA>
	assez	472	22.36	<NA>
	frequent	143	6.77	<NA>
grp_tps_use_tech	moins_de_30mn	952	45.1	moins_de_30mn
	entre_30mn_1h	755	35.77	<NA>
	entre_1h_1h30	160	7.58	<NA>
	entre_1h30_2h	244	11.56	<NA>

Annexe 4 : Répartition en effectif (à gauche) en proportion (à droite) de quelques modalités







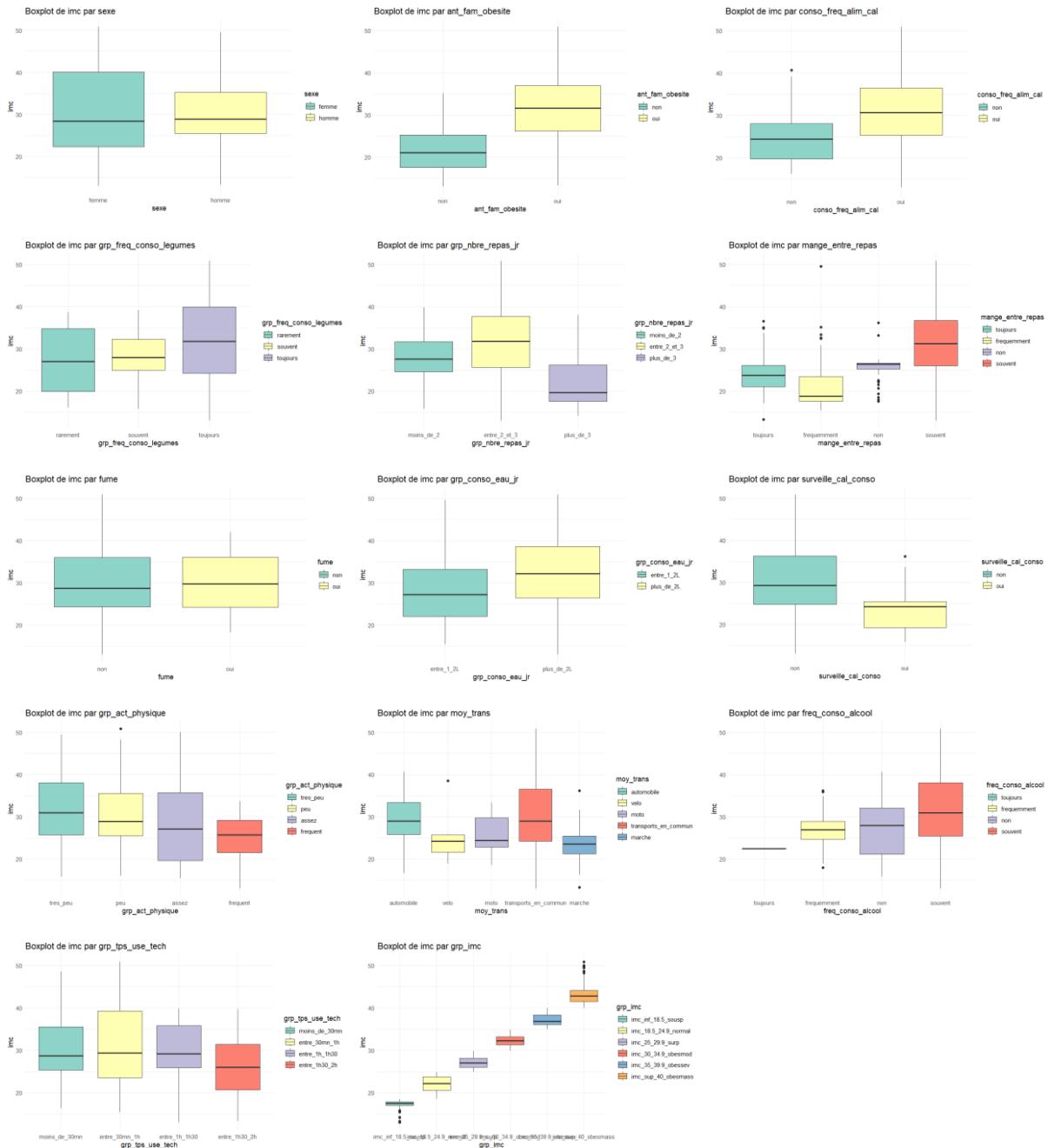
Annexe 5 : Résumé numérique bivarié entre variables qualitatives et IMC

Variables	Catégorie	Effectif	Moyenne	Mediane	Ecart_type
sexe	femme	1043	30.1	28.5	9.40
	homme	1068	29.3	28.9	6.35
ant_fam_obesite	non	385	21.5	21.0	4.21
	oui	1726	31.5	31.6	7.50
conso_freq_alim_cal	non	245	24.3	24.4	5.08
	oui	1866	30.4	30.7	8.05
mange_entre_repas	toujours	53	24.3	23.8	4.96
	fréquemment	242	20.9	18.9	4.46
	non	51	25.4	26.4	3.33

	souvent	1765	31.2	31.2	7.67
ant_fam_obesite	non	2067	29.7	28.7	8.04
	oui	44	29.7	29.7	6.60
	non	2015	30.0	29.3	8.01
surveille_cal_conso	oui	96	22.9	24.2	4.04
	toujours	1	22.5	22.5	NA
	frequemment	70	27.0	26.9	3.72
	non	639	27.1	28.0	6.39
freq_conso_alcool	souvent	1401	31.0	30.9	8.49
	automobile	457	29.2	29.0	6.09
	velo	7	25.2	24.2	6.56
	moto	11	25.8	24.4	4.79
moy_trans	transports_en_commun	1580	30.1	29.0	8.51
	marche	56	23.7	23.6	4.09
	sous_poids	272	17.4	17.5	0.786
	normal	287	22.0	22.1	1.84
	obesite_type_1	351	32.3	32.2	1.13
niv_obesite	obesite_type_2	297	36.7	36.4	1.29
	obesite_type_3	324	42.3	41.9	2.58
	surpoids_niv_1	290	26.0	26.0	0.661
	surpoids_niv_2	290	28.2	28.2	0.828
	imc_inf_18.5_sousp	267	17.4	17.5	0.771
	imc_18.5_24.9_normal	300	22.0	22.2	1.89
	imc_25_29.9_surp	566	27.1	27.0	1.27
grp_imc	imc_30_34.9_obesmod	370	32.3	32.2	1.20
	imc_35_39.9_obesnev	340	37.1	36.9	1.38
	imc_sup_40_obesmass	268	43.0	42.8	2.15
	rarement	131	27.2	27.0	7.41
	souvent	903	28.0	28.0	5.48
	toujours	1077	31.4	31.8	9.37
grp_age	14_20	585	24.7	23.7	7.64
	20_26	1028	31.8	31.8	8.24
	26_35	330	31.5	32.1	5.06
	35_61	168	31.0	31.0	4.14
grp_nbrepas_jr	moins_de_2	395	27.4	27.5	5.65
	entre_2_et_3	1488	31.5	31.8	8.03
	plus_de_3	228	22.1	19.6	5.53
grp_conso_eau_jr	entre_1_2L	1217	27.9	27.3	7.48
	plus_de_2L	894	32.2	32.2	8.04
grp_act_physique	tres_peu	833	30.9	30.9	7.76
	peu	663	29.9	28.9	7.45
	assez	472	28.7	27.1	9.26
	frequent	143	25.2	25.7	5.12
grp_tps_use_tech	moins_de_30mn	952	29.7	28.7	7.11

entre_30mn_1h	755	30.9	29.3	9.29
entre_1h_1h30	160	29.1	29.1	7.17
entre_1h30_2h	244	26.4	26.0	6.51

### Annexe 6 : Boxplots entre l'IMC et les différentes variables qualitatives



### Annexe 7 : Tests d'hypothèses sur les relations bivariées entre les variables qualitatives et l'IMC

◆ Analyse de imc en fonction de sexe

femme homme

1043 1068

Test t de Student entre imc et sexe : Statistique = 2.4282 , p-value = 0.0153

Test de Wilcoxon entre imc et sexe : Statistique = 568467 , p-value = 0.4113

---

◆ Analyse de imc en fonction de ant\_fam\_obesite

non oui

385 1726

Test t de Student entre imc et ant\_fam\_obesite : Statistique = 35.7676 , p-value = 0

Test de Wilcoxon entre imc et ant\_fam\_obesite : Statistique = 579810 , p-value = 0

---

◆ Analyse de imc en fonction de conso\_freq\_alim\_cal

non oui

245 1866

Test t de Student entre imc et conso\_freq\_alim\_cal : Statistique = -16.4352 , p-value = 0

Test de Wilcoxon entre imc et conso\_freq\_alim\_cal : Statistique = 125896 , p-value = 0

---

◆ Analyse de imc en fonction de mange\_entre\_repas

toujours frequemment non souvent

53 242 51 1765

Test ANOVA entre imc et mange\_entre\_repas : Statistique = 159.9893 , p-value = 0

Test Kruskal-Wallis entre imc et mange\_entre\_repas : Statistique = 409.6589 , p-value = 0

---

◆ Analyse de imc en fonction de fume

non oui

2067 44

Test t de Student entre imc et fume : Statistique = 0.0455 , p-value = 0.9639

Test de Wilcoxon entre imc et fume : Statistique = 45056.5 , p-value = 0.917

---

◆ Analyse de imc en fonction de surveille\_cal\_conso

non oui

2015 96

Test t de Student entre imc et surveille\_cal\_conso : Statistique = 15.7653 , p-value = 0

Test de Wilcoxon entre imc et surveille\_cal\_conso : Statistique = 150645.5 , p-value = 0

---

◆ Analyse de imc en fonction de freq\_conso\_alcool

toujours frequemment non souvent

1	70	639	1401
---	----	-----	------

Test ANOVA entre imc et freq\_conso\_alcool : Statistique = 41.3961 , p-value = 0

Test Kruskal-Wallis entre imc et freq\_conso\_alcool : Statistique = 98.2987 , p-value = 0

---

◆ Analyse de imc en fonction de moy\_trans

automobile	velo	moto transports_en_commun	marche
457	7	11	1580
			56

Test ANOVA entre imc et moy\_trans : Statistique = 10.8752 , p-value = 0

Test Kruskal-Wallis entre imc et moy\_trans : Statistique = 46.4791 , p-value = 0

---

◆ Analyse de imc en fonction de niv\_obesite

sous_poids	normal obesite_type_1	obesite_type_2	obesite_type_3	surpoids_niv_1	surpoids_niv_2
272	287	351	297	324	290

Test ANOVA entre imc et niv\_obesite : Statistique = 10085.48 , p-value = 0

Test Kruskal-Wallis entre imc et niv\_obesite : Statistique = 2059.783 , p-value = 0

---

◆ Analyse de imc en fonction de grp\_imc

imc_inf_18.5_sousp	imc_18.5_24.9_normal	imc_25_29.9_surp	imc_30_34.9_obesmod	imc_35_39.9_obesev
imc_sup_40_obesmass				

267	300	566	370	340	268
-----	-----	-----	-----	-----	-----

Test ANOVA entre imc et grp\_imc : Statistique = 12126.21 , p-value = 0

Test Kruskal-Wallis entre imc et grp\_imc : Statistique = 2034.513 , p-value = 0

---

◆ Analyse de imc en fonction de grp\_freq\_conso\_legumes

rarement souvent toujours

131	903	1077
-----	-----	------

Test ANOVA entre imc et grp\_freq\_conso\_legumes : Statistique = 54.2582 , p-value = 0

Test Kruskal-Wallis entre imc et grp\_freq\_conso\_legumes : Statistique = 84.2553 , p-value = 0

---

◆ Analyse de imc en fonction de grp\_age

14_20	20_26	26_35	35_61
-------	-------	-------	-------

585	1028	330	168
-----	------	-----	-----

Test ANOVA entre imc et grp\_age : Statistique = 123.9937 , p-value = 0

Test Kruskal-Wallis entre imc et grp\_age : Statistique = 340.0039 , p-value = 0

- ◆ Analyse de imc en fonction de grp\_nbre\_repas\_jr

moins\_de\_2 entre\_2\_et\_3 plus\_de\_3

395 1488 228

Test ANOVA entre imc et grp\_nbre\_repas\_jr : Statistique = 182.5111 , p-value = 0

Test Kruskal-Wallis entre imc et grp\_nbre\_repas\_jr : Statistique = 304.5377 , p-value = 0

- ◆ Analyse de imc en fonction de grp\_conso\_eau\_jr

entre\_1\_2L plus\_de\_2L

1217 894

Test t de Student entre imc et grp\_conso\_eau\_jr : Statistique = -12.4874 , p-value = 0

Test de Wilcoxon entre imc et grp\_conso\_eau\_jr : Statistique = 375315.5 , p-value = 0

- ◆ Analyse de imc en fonction de grp\_act\_physique

tres\_peu peu assez frequent

833 663 472 143

Test ANOVA entre imc et grp\_act\_physique : Statistique = 24.1846 , p-value = 0

Test Kruskal-Wallis entre imc et grp\_act\_physique : Statistique = 78.5013 , p-value = 0

- ◆ Analyse de imc en fonction de grp\_tps\_use\_tech

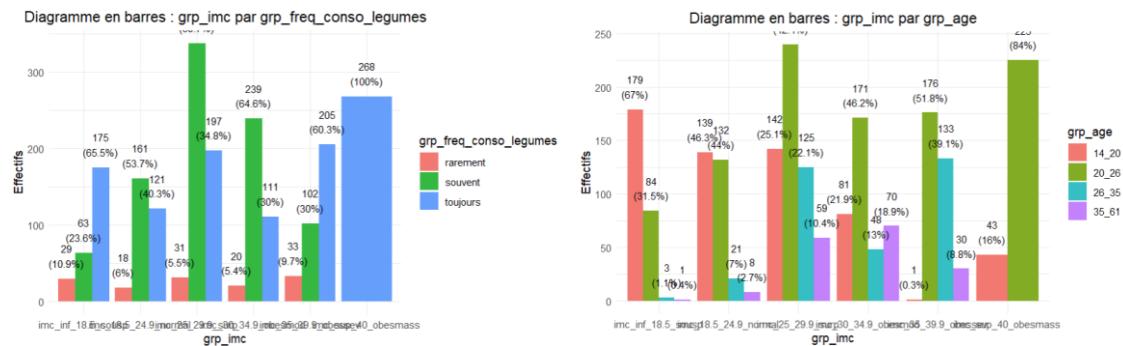
moins\_de\_30mn entre\_30mn\_1h entre\_1h\_1h30 entre\_1h30\_2h

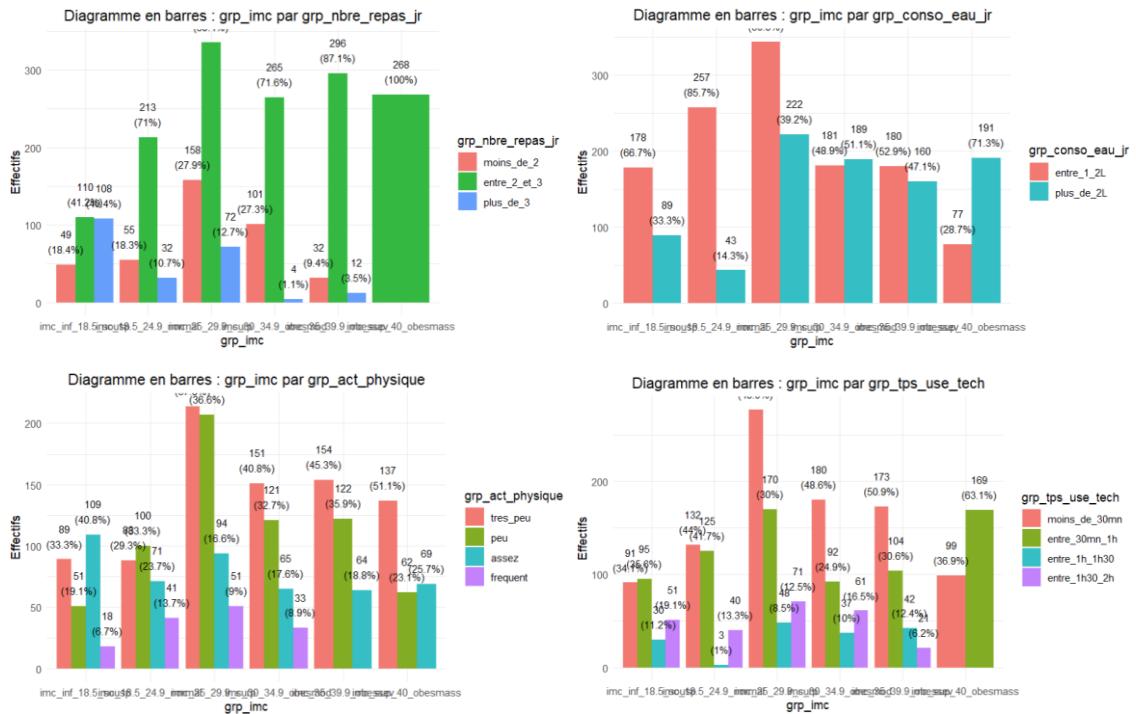
952 755 160 244

Test ANOVA entre imc et grp\_tps\_use\_tech : Statistique = 20.2365 , p-value = 0

Test Kruskal-Wallis entre imc et grp\_tps\_use\_tech : Statistique = 49.8675 , p-value = 0

### Annexe 8 : Barplots bivariés





### Annexe 9 : Tests d'hypothèses des relations entre les variables qualitatives

- ◆ Analyse entre grp\_imc et grp\_freq\_conso\_legumes

[1] "Table de contingence :"

	rarement souvent toujours		
imc_inf_18.5_sousp	29	63	175
imc_18.5_24.9_normal	18	161	121
imc_25_29.9_surp	31	338	197
imc_30_34.9_obesmod	20	239	111
imc_35_39.9_obesev	33	102	205
imc_sup_40_obesmass	0	0	268

Test du Khi-2 :  $X^2 = 482.3225$ , p-value = 0, V de Cramer = 0.338

- ◆ Analyse entre grp\_imc et grp\_age

[1] "Table de contingence :"

	14_20	20_26	26_35	35_61
imc_inf_18.5_sousp	179	84	3	1
imc_18.5_24.9_normal	139	132	21	8
imc_25_29.9_surp	142	240	125	59

imc_30_34.9_obesmod	81	171	48	70
imc_35_39.9_obesev	1	176	133	30
imc_sup_40_obesmass	43	225	0	0

Test du Khi-2 :  $X^2 = 730.2218$ , p-value = 0, V de Cramer = 0.3396

---

- ◆ Analyse entre grp\_imc et grp\_nbre\_repas\_jr

[1] "Table de contingence :"

	moins_de_2_entre_2_et_3_plus_de_3		
imc_inf_18.5_sousp	49	110	108
imc_18.5_24.9_normal	55	213	32
imc_25_29.9_surp	158	336	72
imc_30_34.9_obesmod	101	265	4
imc_35_39.9_obesev	32	296	12
imc_sup_40_obesmass	0	268	0

Test du Khi-2 :  $X^2 = 492.1717$ , p-value = 0, V de Cramer = 0.3414

---

- ◆ Analyse entre grp\_imc et grp\_conso\_eau\_jr

[1] "Table de contingence :"

	entre_1_2L_plus_de_2L		
imc_inf_18.5_sousp	178	89	
imc_18.5_24.9_normal	257	43	
imc_25_29.9_surp	344	222	
imc_30_34.9_obesmod	181	189	
imc_35_39.9_obesev	180	160	
imc_sup_40_obesmass	77	191	

Test du Khi-2 :  $X^2 = 214.0484$ , p-value = 0, V de Cramer = 0.3184

---

- ◆ Analyse entre grp\_imc et grp\_act\_physique

[1] "Table de contingence :"

	tres_peu	peu	assez	frequent
imc_inf_18.5_sousp	89	51	109	18
imc_18.5_24.9_normal	88	100	71	41
imc_25_29.9_surp	214	207	94	51
imc_30_34.9_obesmod	151	121	65	33
imc_35_39.9_obesev	154	122	64	0
imc_sup_40_obesmass	137	62	69	0

Test du Khi-2 :  $X^2 = 174.6178$ , p-value = 0, V de Cramer = 0.1661

---

- ◆ Analyse entre grp\_imc et grp\_tps\_use\_tech

[1] "Table de contingence :"

	moins_de_30mn	entre_30mn_1h	entre_1h_1h30	entre_1h30_2h
imc_inf_18.5_sousp	91	95	30	51
imc_18.5_24.9_normal	132	125	3	40
imc_25_29.9_surp	277	170	48	71
imc_30_34.9_obesmod	180	92	37	61
imc_35_39.9_obesev	173	104	42	21
imc_sup_40_obesmass	99	169	0	0

Test du Khi-2 :  $X^2 = 212.9649$ , p-value = 0, V de Cramer = 0.1834

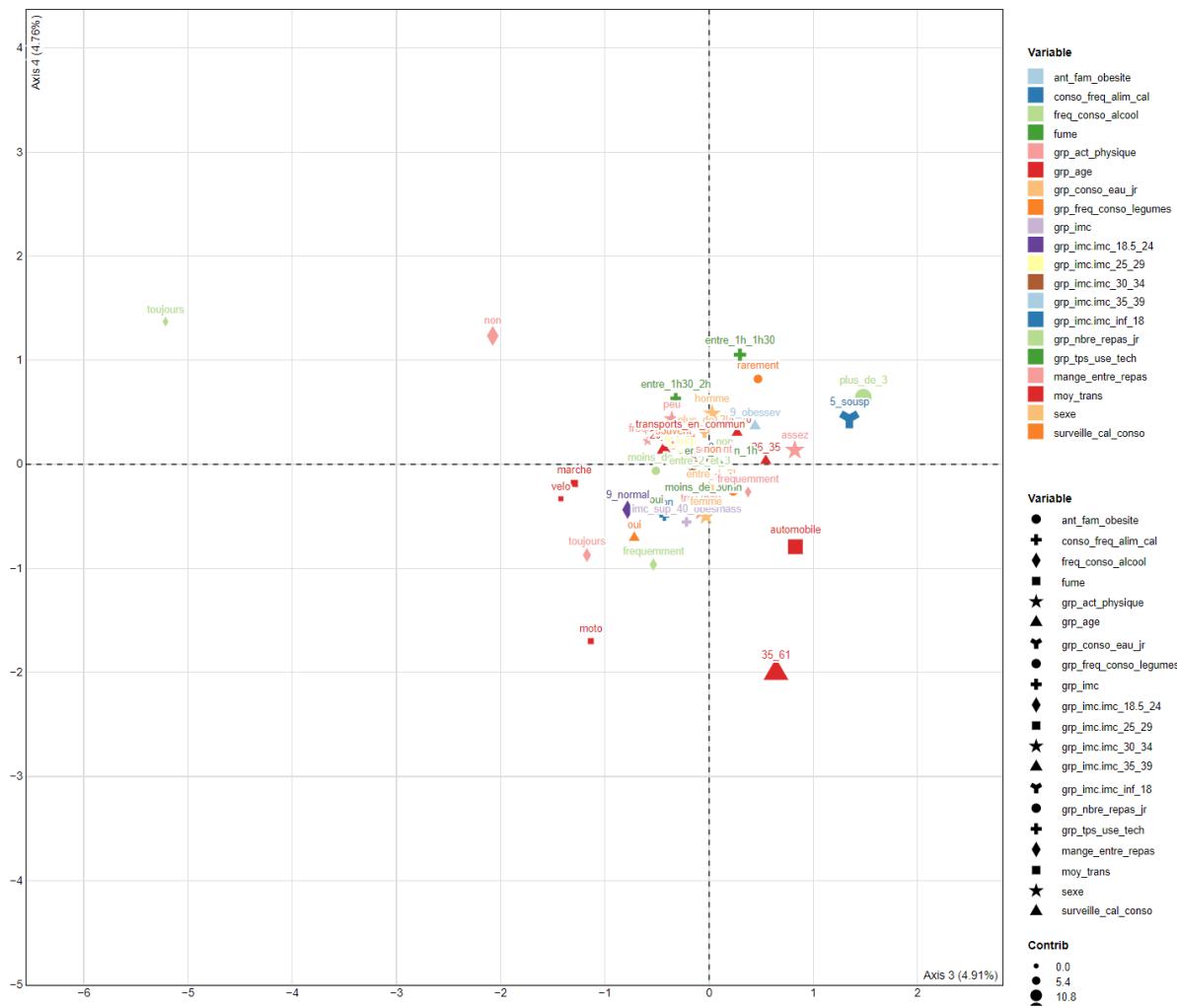
---

*Annexe 10 : Tableau des valeurs propres*

Tableau des valeurs propres

Axe	%	Cum. %
1	8.1	8.1
2	7.0	15.1
3	4.9	20.0
4	4.8	24.8
5	4.6	29.4
6	4.2	33.6
7	3.8	37.4
8	3.5	40.9
9	3.4	44.3
10	3.4	47.7
11	3.3	50.9
12	3.2	54.1
13	3.0	57.1
14	2.9	60.0
15	2.8	62.8
16	2.8	65.6
17	2.7	68.3
18	2.6	70.9
19	2.6	73.6
20	2.6	76.1
21	2.5	78.6
22	2.3	80.9
23	2.2	83.1
24	2.1	85.2
25	2.1	87.3
26	1.9	89.1
27	1.8	90.9
28	1.8	92.7
29	1.6	94.3
30	1.5	95.8
31	1.4	97.1
32	1.2	98.4
33	1.0	99.3
34	0.7	100.0

Annexe 11 : Graphe des variables sur les axes 3 et 4



### *Annexe 12 : Contribution des variables sur les axes 3 et 4 :*

