# Rapport d'analyse de données

Youssef SHALABY Lyse LEBON

2026-01-14

## Table of contents

# 1 Introduction

Dans ce rapport, nous allons étudier un jeu de données fournissant un aperçu détaillé des routines d'exercice, des attributs physiques et des mesures de la condition physique (*les modalités*) de 973 membres d'une salle de sport (*les individus*). Ces individus sont répertoriés dans une base de données nommée "DataGym3MIC" que nous avons renommée pour la suite "gym".

**Lorsqu'il y aura la présence de ce signe : [Q], cela signifiera qu'il y a un complément sur le document quarto**.

Dans cet échantillon, nous retrouvons deux types de données : les données **qualitatives** ("gender" [nominale] et "level" [ordinale]) et les données **quantitatives continues** ("weight", "height", "duration", "calories", "fat", "water" et "bmi"). Nous avons affiché le sommaire des données sur la Table 1.

Table 1: Les premières lignes du jeu de données Gym

| gender | weight | height | duration | calories | fat | water | level | bmi |
|---|---|---|---|---|---|---|---|---|
| Length:973 | Min. : 40.00 | Min. :1.500 | Min. :0.500 | Min. : 303.0 | Min. :10.00 | Min. :1.500 | Min. :1.00 | Min. :12.32 |
| Class :character | 1st Qu.: 58.10 | 1st Qu.:1.620 | 1st Qu.:1.040 | 1st Qu.: 720.0 | 1st Qu.:21.30 | 1st Qu.:2.200 | 1st Qu.:1.00 | 1st Qu.:20.11 |
| Mode :character | Median : 70.00 | Median :1.710 | Median :1.260 | Median : 893.0 | Median :26.20 | Median :2.600 | Median :2.00 | Median :24.16 |
| NA | Mean : 73.85 | Mean :1.723 | Mean :1.256 | Mean : 905.4 | Mean :24.98 | Mean :2.627 | Mean :1.81 | Mean :24.91 |
| NA | 3rd Qu.: 86.00 | 3rd Qu.:1.800 | 3rd Qu.:1.460 | 3rd Qu.:1076.0 | 3rd Qu.:29.30 | 3rd Qu.:3.100 | 3rd Qu.:2.00 | 3rd Qu.:28.56 |
| NA | Max. :129.90 | Max. :2.000 | Max. :2.000 | Max. :1783.0 | Max. :35.00 | Max. :3.700 | Max. :3.00 | Max. :49.84 |

Nous avons spécifié à R que les deux variables gender et level étaient des variables qualitatives et nous avons également renommé les niveaux avec "débutant", "intermédiaire" et "avancé".

# 2 Statistique uni et bi-dimensionnelle

## 2.1 Étude statistique unidimensionnelle

### 2.1.1 Cas des variables qualitatives

Dans cette partie, nous nous focaliserons sur les deux variables qualitatives **"gender"** et **"level"**.

Tout d'abord, pour la variable **"gender"**, qualitative nominale. On constate sur la Figure 1, qu'un peu plus de 50% des inscrits à la salle sont des hommes. La répartition des sexes dans cette salle est équilibrée.

La variable **"level"** est une variable qualitative ordinale. On peut donc utiliser les fréquences afin de poursuivre l'étude. Nous avons choisi pour la Figure 1 de n'afficher que le camembert

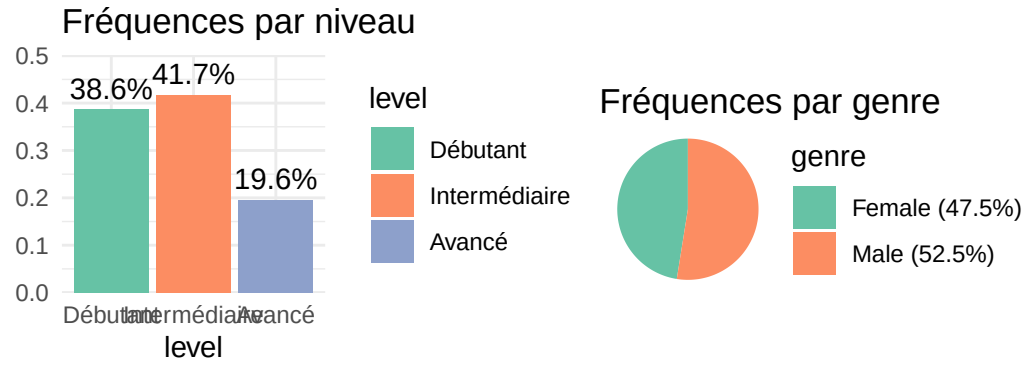



Figure 1: Diagramme en bâtons des différents niveaux (gauche) et graphe camembert de la répartition homme/femme (droite)

et le diagramme en bâtons car ils nous permettent de comprendre rapidement les valeurs des deux variables.

On constate avec la Figure 1 qu'il y a presque autant de débutants que d'intermédiaires et qu'ils sont plus nombreux que les avancés (plus de 75%). Ces résultats sont cohérents puisque maîtriser un sport demande beaucoup de temps et de persévérance.

### 2.1.2 Cas des variables quantitatives

Dans cette partie, nous allons analyser les Boxplot de toutes les variables quantitatives et en extraire quelques-uns qui peuvent être intéressants.

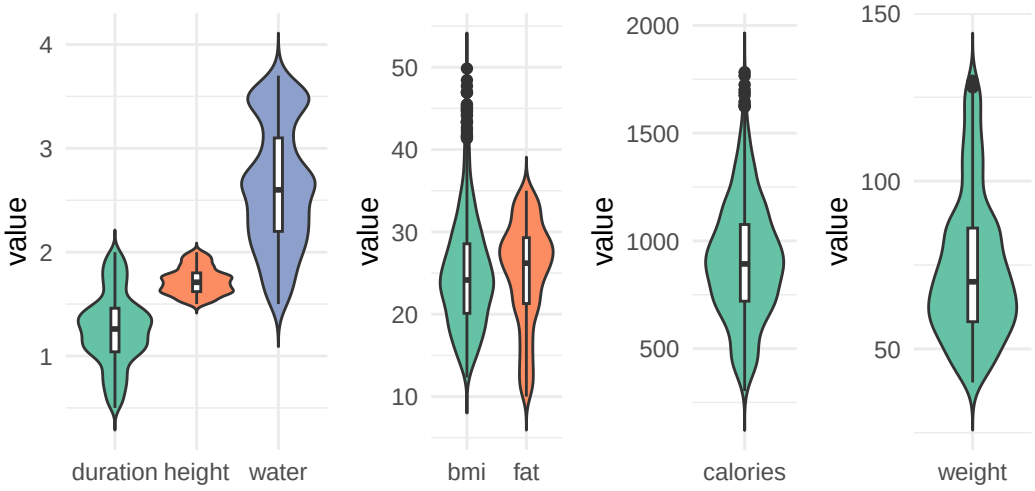



Figure 2: Visualisation de la répartition de toutes les variables quantitatives de gym par des violin plot

Grâce à la Figure 2, on constate que les différents quartiles des variables correspondent aux valeurs de la Table 1. On observe également que les variables “weight”, “calories” et “bmi” possèdent de nombreux outliers avec un étalement vers la droite. Dans le cadre de cette étude, nous allons conserver ces outliers qui restent pertinents dans le cadre d’une salle de sport. On constate également une forme particulière du violin plot de la variable “water”. Nous allons désormais étudier plus en profondeur les variables “bmi” et “water”.

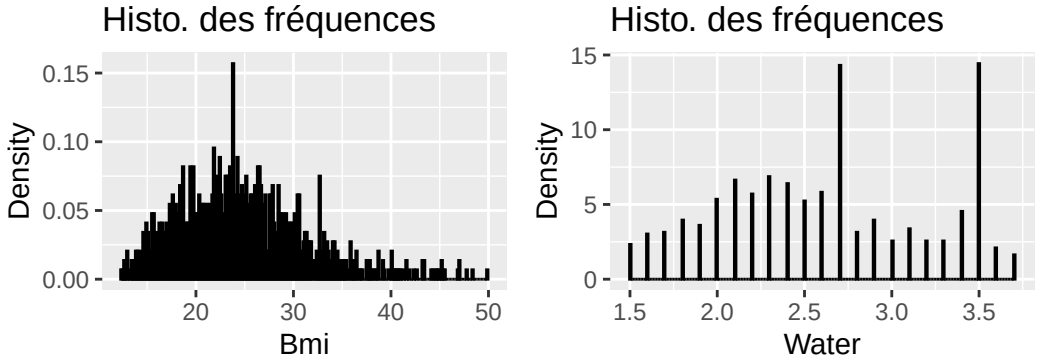



Figure 3: Histogramme des fréquences pour la variable water (à droite) et bmi (à gauche)

Analysons alors leurs histogrammes afin de comprendre ces valeurs “aberrantes”. D’après la Figure 3, on constate en effet un fort étalement vers la droite pour la variable “bmi”. Cet étalement peut s’expliquer par le fait que les inscrits à la salle développent du muscle (et perdent du gras), mais leur IMC reste élevé, voire augmente. Un autre facteur serait que des inscrits viennent à la salle afin de perdre du poids et ont donc un IMC élevé. Les deux pics sur l’histogramme de la variable “water” peuvent s’expliquer par le fait qu’il y ait des inscrits qui font des séances très longues (d’où la nécessité de boire beaucoup). On pourrait également soumettre l’hypothèse que les niveaux avancés prennent de la créatine, des protéines ou des électrolytes qui nécessitent un apport d’eau important. Le fait que cette courbe ne soit pas lisse pourrait s’expliquer par le fait qu’il s’agisse peut-être d’une variable discrétisée, les inscrits arrondissent leur consommation à des bouteilles de 0,5L ou 1,5L.

## 2.2 Étude statistique bi-dimensionnelle

### 2.2.1 Étude entre deux variables qualitatives

Dans cette partie, nous analysons les deux variables qualitatives “level” et “gender”. Grâce à la Table 2, nous pouvons observer la table de contingence de ces deux variables.

Table 2: Matrice de contingence entre level et sexe

| | Débutant | Intermédiaire | Avancé | Sum |
|---|---|---|---|---|
| Female | 179 | 193 | 90 | 462 |
| Male | 197 | 213 | 101 | 511 |
| Sum | 376 | 406 | 191 | 973 |

On constate toujours qu'il y a presque autant d'hommes que de femmes mais que, malgré tout, les hommes restent plus présents dans tous les niveaux confondus.

L'indice de Cramér = 0.0036 est quasiment nul, on en conclut que la liaison est très peu significative. Ainsi, on en déduit que dans ce club de gym, le niveau de compétence ("level") des individus est indépendant du genre ("gender").

### 2.2.2 Étude entre une variable qualitative et une quantitative

Dans cette partie nous avons décidé d'analyser la variable "level" avec la variable "fat" ainsi que la variable "level" avec la variable "duration". En effet, nous pensons que ces deux couples sont fortement liés. Vérifions ce prédicat à l'aide de la Figure 4.

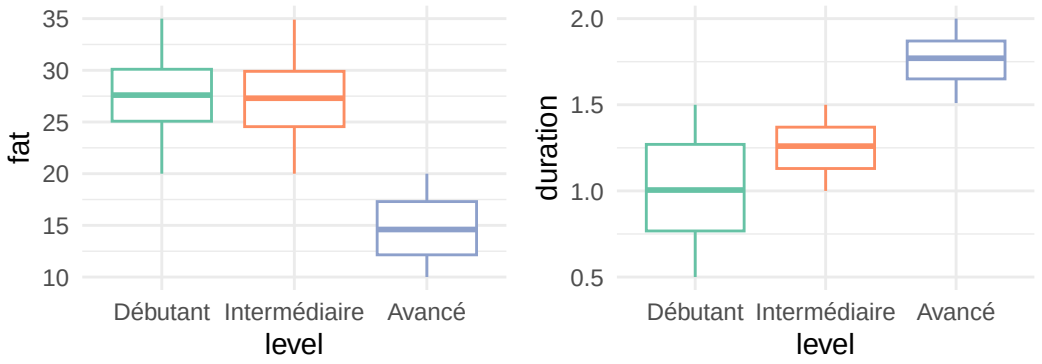



Figure 4: Boxplot entre level et fat (à gauche) et entre level et duration (à droite)

Après la création du boxplot de la Figure 4 pour ces deux couples, on constate bien qu'il y a une forte liaison entre la variable "level" et la variable "fat". En effet, bien que pour les débutants et intermédiaires le boxplot est plutôt similaire, on constate une nette différence avec le niveau avancé.
Idem pour la variable "level" et "duration", on peut constater la même chose que précédemment: il y a une nette différence entre les boxplots pour les deux premiers niveaux avec celui du niveau avancé.
[Q].

Table 3: Les rapports de correlation entre les variables quantitatives et qualitatives

| | height | weight | duration | calories | fat | bmi | water |
|---|---|---|---|---|---|---|---|
| gender | 0.3405 | 0.3356 | 1e-04 | 0.0227 | 0.1659 | 0.0973 | 0.4458 |
| level | 0.0009 | 0.0004 | 0.6218 | 0.5093 | 0.648 | 0.0020 | 0.1685 |

Grâce à la Table 3, qui affiche le rapport de corrélation entre toutes les variables quantitatives avec toutes les variables qualitatives, nous pouvons maintenant analyser si nos deux couples ont un rapport de corrélation faible ou important. Le rapport de corrélation entre "fat" et "level" (en vert dans la Table 3) est de : 0.648 et le rapport entre "duration" et "level" (en orange dans la Table 3) est de: 0.6218. On constate ainsi que les variables "fat" et "level" ainsi que duration et "level" sont fortement corrélées (le rapport de correlation est proche de 1). Ces valeurs paraissent logiques. En effet, plus on a un bon niveau dans un sport, plus on est apte à durer longtemps. Idem, plus on fait de sport, moins on a de graisse.
On peut également observer un rapport de corrélation élevé entre les variables "level" et "calories" ainsi que les variables "water" et "gender".

### 2.2.3 Étude entre deux variables quantitatives

Nous allons poursuivre cette partie en faisant l'étude des variables qui sont le plus corrélées, comme le montre la Figure 5.

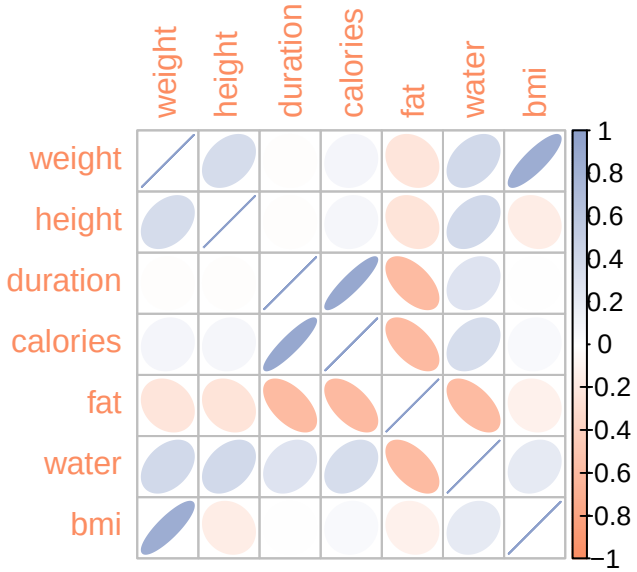



Figure 5: Matrice de corrélation des 7 variables quantitatives

À l'aide de la Figure 5, on constate que la variable "fat" est fortement corrélée négativement avec "duration", "calories" et "water". La variable "bmi" est corrélée positivement avec

“weight” ou encore la variable “calories” avec “duration”.

Grâce à la forte corrélation entre certaines variables quantitatives, il est pertinent de mener une Analyse en Composantes Principales (ACP). Cela permettra de résumer l'information en éliminant le bruit et la redondance, et de visualiser la structure des données sur un plan factoriel.

# 3 Analyse en Composantes Principales

Dans cette section, nous allons faire une Analyse en Composantes Principales sur les données centrées réduites.
Nos variables étant exprimées dans des unités de mesure hétérogènes (poids, temps, calories), l'ACP centrée réduite est indispensable. Elle permet de standardiser les données et ramène toutes les variables à une variance unitaire, ce qui revient à analyser la matrice des corrélations. Cela garantit que chaque variable contribue de manière équitable à la construction des axes factoriels.

L'inertie globale du nuage de point notée $\mathcal{J}$, est, dans notre cas d'étude avec des données centrées réduites, égale au nombre de variables quantitatives, ici $\mathcal{J} = 7$ (on veut calculer la trace de ΓM qui est la matrice des corrélations [qui vaut 1 sur la diagonale]).
Cette valeur représente la quantité d'information disponible que nous cherchons à résumer.
Par la suite, on nommera notre jeu de données centré et réduit GymCR.

## 3.1 Choix du nombre d'axes factoriels

Grâce à l'ACP centrée réduite, sur la Figure 6, on constate que l'étude des variables quantitatives est portée par au moins 3 dimensions (la somme des trois pourcentages vaut $85,9\%$). Cela signifie que l'essentiel de l'information contenue dans les données peut être visualisé dans un espace tridimensionnel.

Remarque : La décomposition de l'inertie globale sur les axes factoriels, observée dans la Figure 6, nous donne les inerties axiales, correspondant aux valeurs propres de la matrice de corrélation. L'axe 1 capture une inertie axiale de $\lambda_1 = 2.905$, l'axe 2, lui, capture une inertie axiale de $\lambda_2 = 1.876$ et l'axe 3 $\lambda_3 = 1.232$. Ainsi, notre réduction dimensionnelle préserve 85.9% de la variance du jeu de données original.

## 3.2 Analyse de la distribution des variables quantitatives selon les différents axes factoriels

Comme nous ne pouvons pas projeter en 3D, nous allons d'abord faire l'analyse du premier plan factoriel, puis le plan factoriel porté par l'axe factoriel 1 et l'axe factoriel 3.

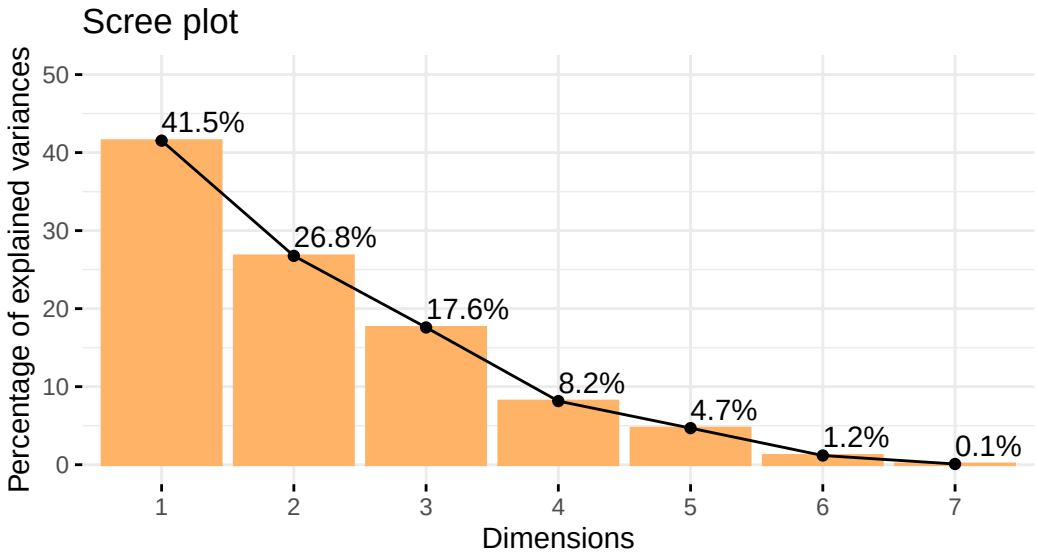



Figure 6: Pourcentage d'inertie expliquée par les axes factoriels

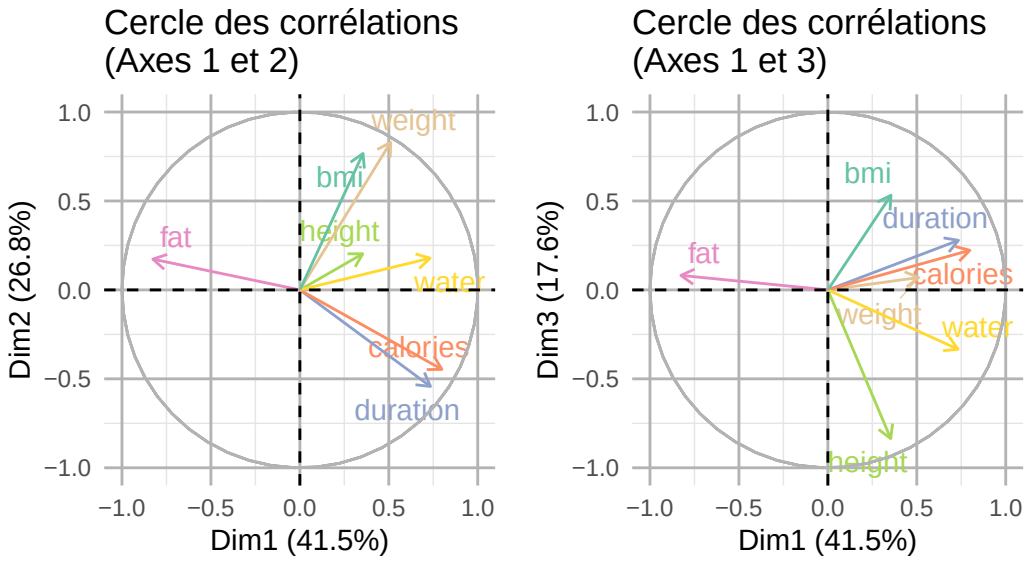



Figure 7: Cercles des corrélations des variables quantitatives projetées sur les plans factoriels (1,2) et (1,3)

Ainsi avec la Figure 7, on constate que la distribution des variables est équitablement répartie pour les 2 associations différentes, les dimensions ne sont pas définies par une seule variable.

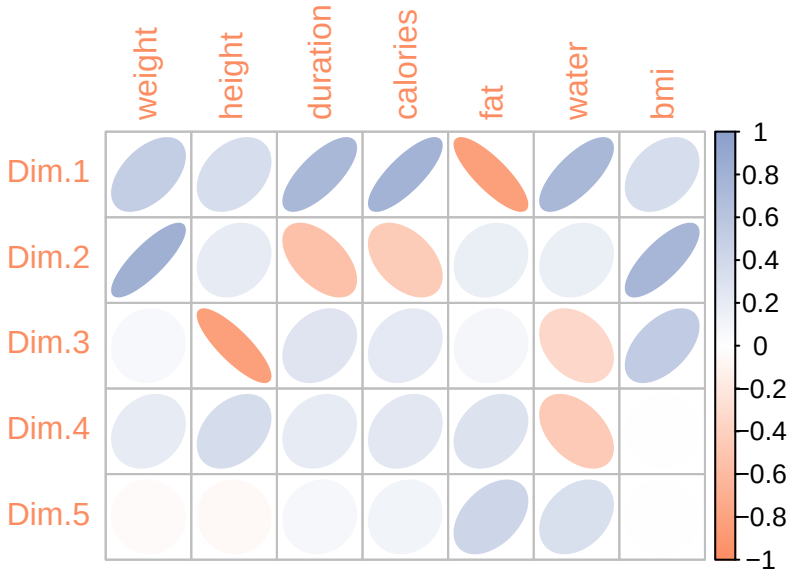



Figure 8: Matrice de corrélation entre les variables initiales et les méta-variables

Analyse du premier axe

Grâce à la Figure 8, on constate que le premier axe factoriel est fortement corrélé positivement avec la variable "calories" et corrélée positivement avec la variable "duration". Il est également corrélé négativement avec la variable "fat". On peut donc supposer que le premier axe structure l'échantillon selon le niveau d'entraînement.
Afin de faire le lien avec la Figure 7 (à l'aide du graphique de gauche), on retrouve à droite les individus qui s'entraînent longtemps, brûlent beaucoup de calories et boivent beaucoup d'eau. À gauche, on trouve les individus ayant un taux de graisse élevé. Cela traduit une logique physiologique : plus l'intensité sportive est élevée, plus le taux de graisse tend à être bas.

Analyse du deuxième axe

Grâce à la Figure 8, on constate que le deuxième axe factoriel est corrélé positivement avec les variables "weight" (très forte corrélation) et "bmi". Il peut représenter la corpulence brute des individus. Sur la Figure 7 (à l'aide du graphique de gauche), on retrouve en haut les individus lourds et à fort IMC et en bas, les individus légers.

Analyse du troisième axe

Grâce à la Figure 8, on constate que le troisième axe factoriel est corrélé négativement avec la variable "height". D'après la Figure 7, il est presque exclusivement défini par la taille mais de façon inversée (les grands en bas).

On note un phénomène intéressant avec le BMI : il est corrélé positivement au deuxième axe et au troisième axe. C'est mathématiquement logique car la formule de l'IMC est $Poids/Taille^2$.

L'ACP a permis la décomposition de l'IMC en ses deux composantes primaires: le poids sur l'axe 2 et la taille sur l'axe 3.

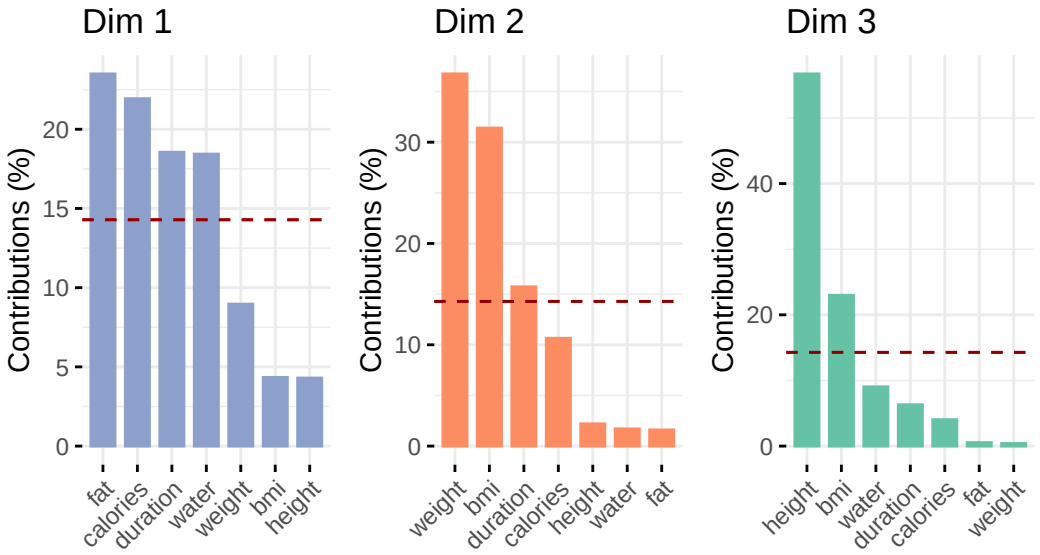



Figure 9: Graphique des contributions des variables quantitatives pour la dim 1 (gauche), dim 2 (centre) et la dim 3 (droite)

Ainsi, grâce à la Figure 9 et l'analyse faite avec la Figure 7, on constate que les variables citées contribuent fortement avec les différentes dimensions. On peut observer que le premier axe est aussi construit par la variable "water" et le deuxième axe est également definit par la variable "duration", qui ne se voyait pas dans les Figure 7 et Figure 8.

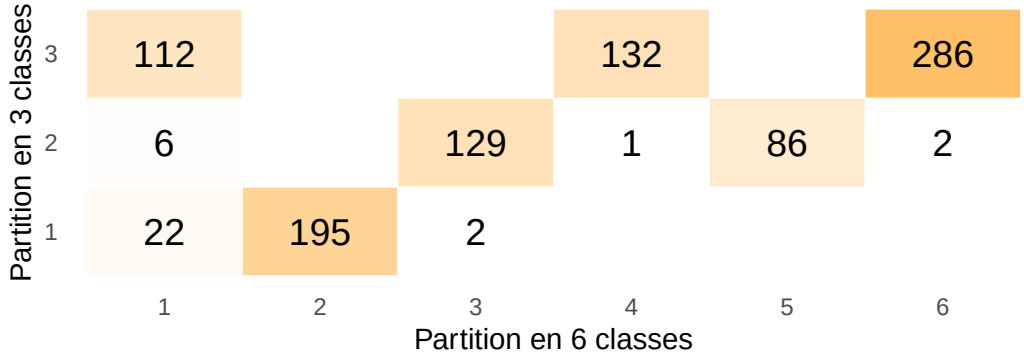
# 4 Clustering

Dans cette partie, nous allons essayer de partitionner un ensemble de données en sous-groupes homogènes. Nous allons tenter de minimiser la variance intra-classe et de maximiser la variance inter-classe.

## 4.1 Méthode de type K-means

Dans cette section nous allons aborder la méthode des K-means, un algorithme d'apprentissage non supervisé. Son but est de segmenter la population de la salle de sport afin d'identifier des profils types.

Table 4: Matrice de contingence croisant la partition en 3 classes et la partition en 6 classes




#### 4.1.1 Choix du nombre de classes

Le choix du nombre de classes K peut se déterminer grâce aux critères suivants: le critère du coefficient de Silhouette et le critère Inertie Intra. L'analyse de la Figure 10 montre que la valeur optimale est K=3, correspondant au maximum observé sur la figure de gauche. On retrouvera sur la figure de droite la valeur K=6, correspondant au coude sous la figure.

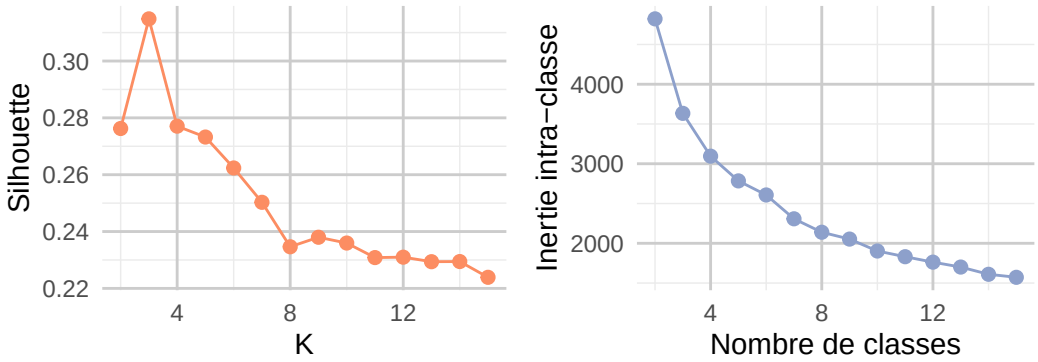



Figure 10: Méthode Silhouette (gauche) et Inertie intra-classe (droite)

Afin de déterminer si on conserve K = 3 ou 6, nous allons comparer les mesures d'agrégation pour ces deux valeurs. Pour K = 3 on obtient: 0.2028 pour gender et 0.2912 pour level. Pour K = 6 on obtient: 0.2042 pour gender et 0.256 pour level. Les mesures d'agrégation obtenues ne nous permettent pas de conclure sur le choix de K.

Grâce à la Table 4, on constate que le premier groupe (lié au clustering pour K=3) se retrouve majoritairement dans le deuxième groupe (lié au clustering pour K=6). Idem, le deuxième groupe (lié au clustering pour K=3) se retrouve majoritairement dans le troisième (lié au clustering pour K=6). Le dernier groupe (lié au clustering K=3) est répartit dans les groupes 1, 4 et 6 du clustering pour K=6. La répartition en 6 classes peut ainsi se retrouver facilement à partir de K=3. On en conclut qu'il y a peu de perte d'information.

On observe sur la Figure 11 avec K=3 qu'on obtient un coefficient de silhouette moyen de 0,31, ce score indique une structure de clustering relativement faible.
Ce résultat suggère qu'il n'y a pas de groupes naturellement disjoints dans la salle de sport. Toutefois, l'analyse des trois classes formées nous permet de comprendre la répartition des individus.

Toujours sur la Figure 11, la deuxième classe (en orange) est la plus large des trois et au dessus de la moyenne. Elle pourrait représenter les utilisateurs typiques qui sont compris entre le premier et le troisième quartile. La troisème classe (en violet) comporte des valeurs négatives qui représente des individus mal classés. Ceux-ci se rapprochent plus de la première ou la deuxième classe. Ces valeurs pourraient représenter les individus qui changent de niveau: de débutant vers intermédiaire ou de intermédiaire vers expert.
Enfin, la première classe (en vert) semble également former un groupe distinct: elle possède des valeurs au dessus de la moyenne et elles sont toutes positives. Toutefois, elle est de taille plus fine.
Ainsi, la première et deuxième classe sont bien identifiées contrairement à la troisième classe qui manque de stabilité, ce qui pourrait expliquer la valeur de la moyenne globale. Ces résultats mettent en évidence la complexité de notre jeux de données.

Ainsi avec K=6 la Figure 11 montre qu'on obtient un coefficient de silhouette moyen de 0,24 (bien plus faible que pour la valeur de K précédente). On constate que cinq classes comportent des valeurs négatives correspondant à des individus mal classés. De plus, les épaisseurs des classes 1, 3, 4 et 5 sont minimes, ce qui indique que peu d'individus se trouvent dans ces classes. La répartition des individus dans ces 6 classes semble plus légère et moins efficace que l'analyse faite précédemment.
Ainsi, on poursuivra cette étude avec K=3.

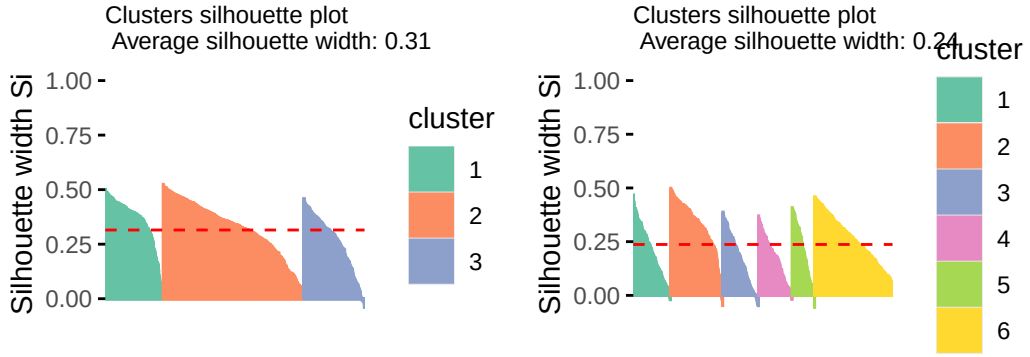



Figure 11: Graphique des silhouettes des individus pour la classification retenue pour K=3 (gauche) et K=6 (droite)

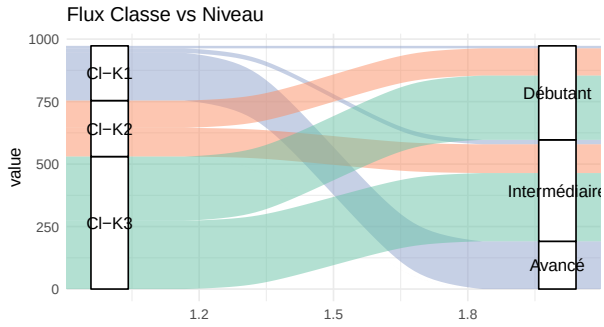
[Q].

### 4.1.2 Analyse du clustering avec le nombre de classes retenu (3)

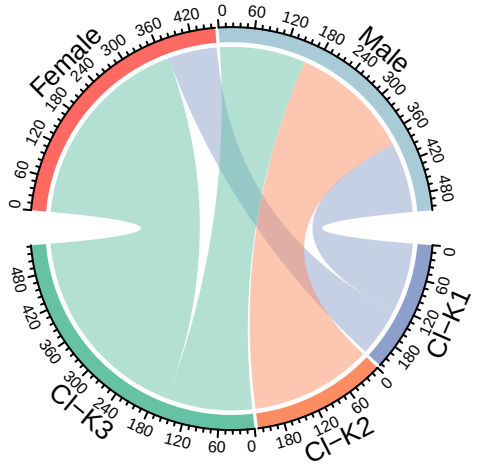
Maintenant que nous avons fixé le nombre de classe à K=3, nous pouvons analyser à quels critères (variables qualitatives) correspondent ces différentes classes.

On observe sur Figure 12a une distinction très nette pour la Classe 1, qui capte la quasi-totalité des profils "Avancé". Cela indique que l'algorithme a su isoler efficacement les performances élevées. En revanche, les Classes 2 et 3 se mélangent pour alimenter les catégories "Débutant" et "Intermédiaire". Ainsi on peut supposer qu'il est difficile de séparer ces deux niveaux en se basant uniquement sur les données physiologiques.

De même, la Figure 12b met en évidence la répartition des genres au sein de nos groupes. Le résultat le plus marquant concerne la Classe 2, qui est composée exclusivement d'hommes (elle regrouperait les hommes débutants et intermédiaires). À l'inverse, la Classe 3 est majoritairement féminine, bien qu'elle comporte une portion masculine. Enfin, la Classe 1 (les sportifs performants ou de niveau avancé d'apres la Figure 12a) est mixte. Cela suggère que si le genre joue un rôle déterminant pour la Classe 2, la haute performance athlétique de la Classe 1 est, elle, indépendante du genre.




(a) Flux de répartition des individus entre les classes du clustering et les niveaux




(b) Diagramme à cordes de la répartition du Genre par classe

Figure 12: Comparaison des classes selon le niveau (gauche) et le genre (droite)

Pour ce qui est de l'influence des variables quantitatives sur les différents cluster, on peut s'appuyer sur la Figure 13. Ici, nous le ferons pour K=3 mais l'étude peut se faire aussi sur K=6.
On constate que la première classe est constituée des individus qui restent longtemps à la salle, qui brûlent beaucoup de calories et qui ont peu de masse graisseuse (ce qui, d'après notre étude, s'apparente aux experts). Les deux autres classes sont presque similaires en terme d'influence des variables: elles comprennent les inscrits qui restent moins longtemps, avec un taux de

graisse plus élévé et brûlent moins de calories. La seule petite différence peut se faire sur la quantité d'eau bue au cours de la séance qui est plus élévée pour la deuxième classe que pour la dernière.


Figure 13: Distribution des variables quantitatives pertinentes de 'gym' en fonction de la classification en 3 classes

## 4.2 Méthode de classification ascendante hiérarchique (CAH)

La méthode CAH nous permet de construire un arbre complet des différentes classes et c'est à nous de "couper" l'arbre afin de choisir le nombre de groupe voulu.

En observant la Figure 14, nous constatons que la fusion des individus se fait d'abord à faible distance, puis la fusion nécéssite des distances de plus en plus grandes. La hauteur du dendrogramme représente la distance minimale entre les deux classes précédentes.

Grâce à la méthode de Calinski (à droite) affichée sur la Figure 15, on analyse un pic à K=3. Pour cette méthode, on retient alors 3 classes. Pour la méthode SPRSQ, on observe une décroissance rapide de l'inertie jusqu'à 4 classes. Au-delà de ce point, la courbe s'aplatit, indiquant que l'ajout de groupes supplémentaires n'améliore que marginalement l'homogénéité de la partition. Nous retenons donc une solution à 4 classes pour cette méthode.

[Q]

Afin de déterminer si on conserve K=3 ou 4, nous allons comparer les mesures d'agrégation pour ces deux valeurs. La valeur obtenue avec la mesure d'agrégation en 4 classes on obtient 0.1266 pour "gender" et 0.2664 pour "level". Si on choisit 3 classes on obtient 0.1294 pour "gender" et 0.3326 pour "level", on peut donc en conclure que la répartition en 4 classes n'est pas adaptée pour ces variables.

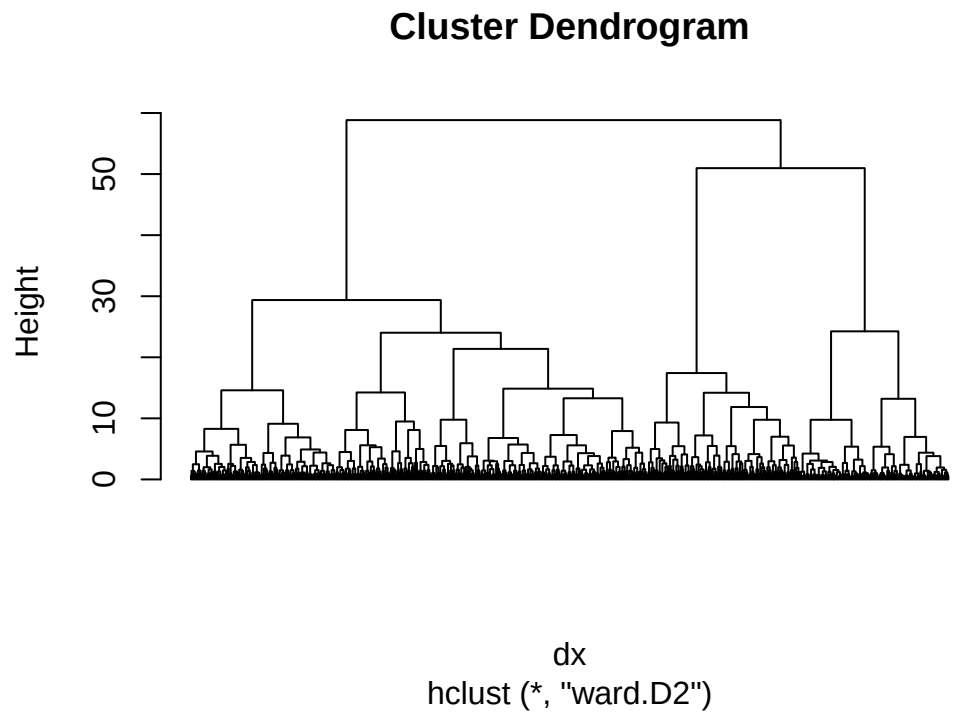



Figure 14: Dendrogrammes du jeu de données gym avec la méthode Ward

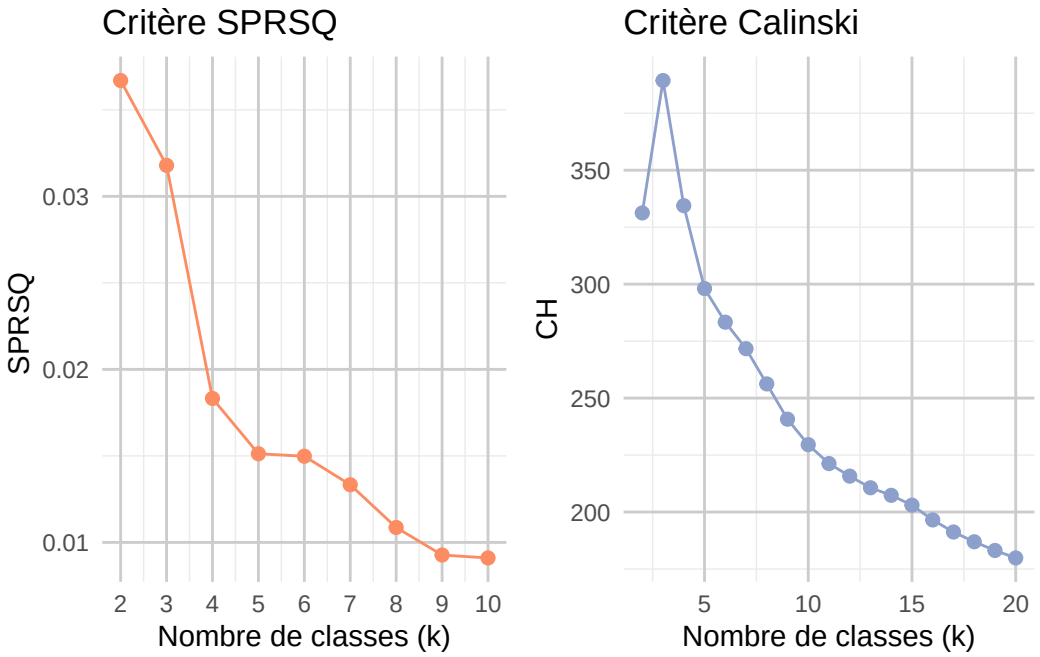



Figure 15: Détermination du nombre de classes grâce aux méthodes SPRSQ (à gauche) et Calinski-Harabasz (à droite)

Les valeurs de mesure d'agrégation sont plus importantes lorsqu'on utlise 3 classes, nous allons ainsi rester sur ce nombre de classes par la suite.

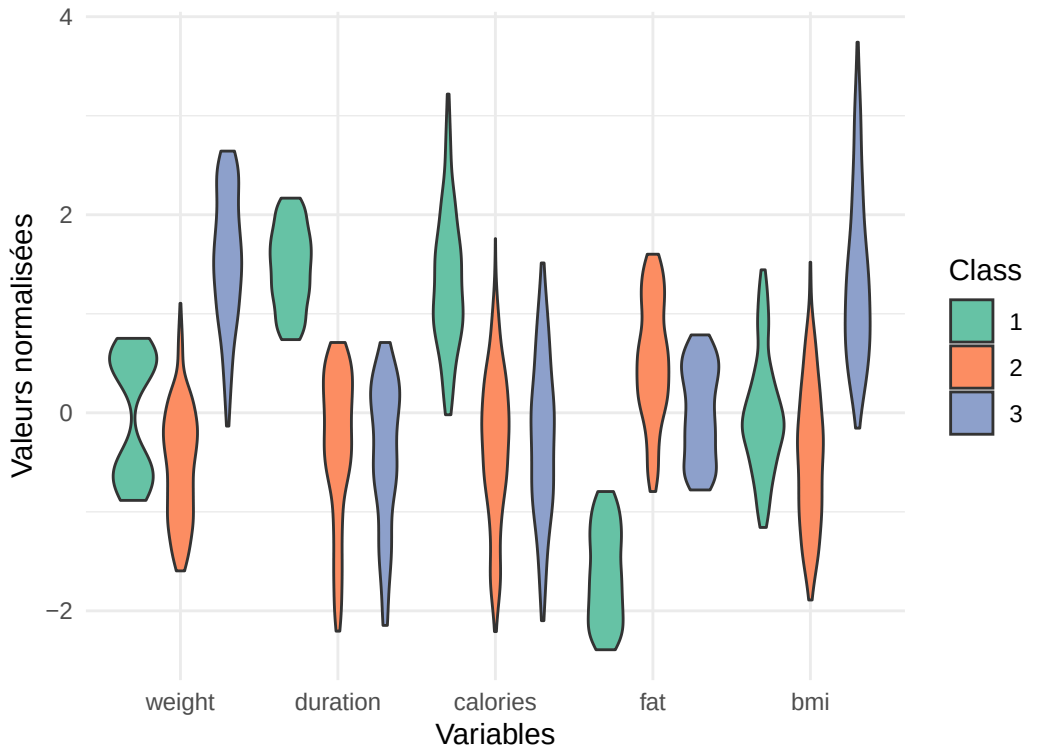


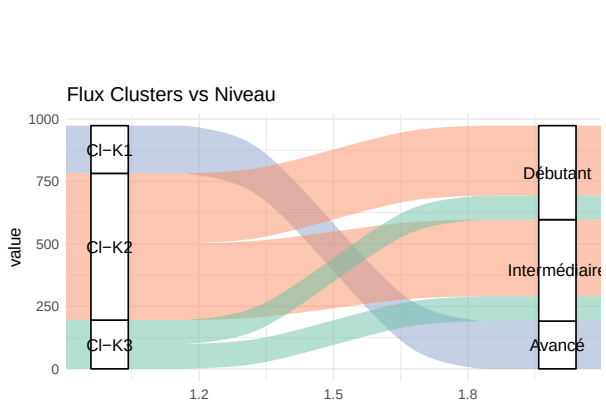

Figure 16: Distribution des variables quantitatives pertinentes de 'gym' en fonction de la classification en trois classes

Grâce à la Figure 16, on constate que la classe 1 comporte les individus qui restent longtemps à la salle, qui brûlent beaucoup de calories et qui ont peu de masse graisseuse (comme sur la Figure 13). La classe 3 comporte les individus qui pèsent lourds, avec une masse graisseuse plus importante (et donc un bmi élévé) et qui brûlent peu de calories. La classe 2 et 3 sont presque similaires, ce qui peu expliquer la difficulté à classer les individus de la salle de sport.

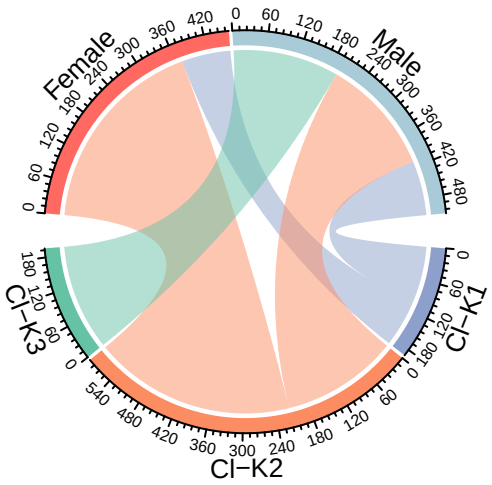
Grâce à la Figure 17a, on observe maintenant que la classe K1 est composée uniquement des individus de la salle qui sont avancés. On rentrouve ainsi dans la Figure 17b qu'il y a presque autant d'homme que de femme qui ont un niveau avancé. La classe K3 est uniquement composée d'hommes.

### 4.3 Comparaison de la méthode des K-means avec la classification hiérarchique

Si on compare cette étude avec la méthode de K-means étudiée précédemment, on constate que 52.62% des données sont classées de la même manière dans les deux méthodes.




(a) Flux entre Clusters et Niveau




(b) Répartition du Genre par Cluster

Figure 17: Comparaison des classes avec le Niveau (gauche) et le Genre (droite)

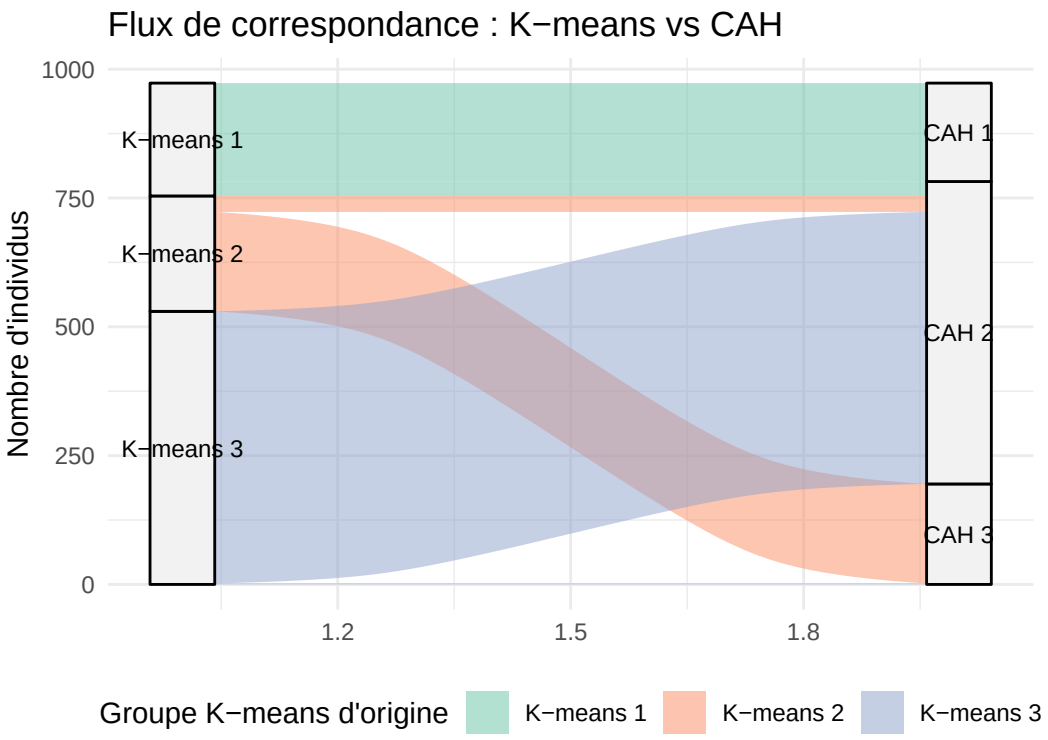



Figure 18: Diagramme alluvial des correspondances entre les classes des K-means et de la CAH

Ainsi, grâce à la Figure 18 on constate que la répartition des individus dans les 3 classes se ressemblent fortement entre les deux méthodes. On en conclut que la classification hiérarchique donne un résultat légèrement plus précis pour la classification des niveaux (grâce à la Figure 17a).

# 5 Conclusion

L'analyse du jeu de données "Gym", portant sur 973 individus (avec plus ou moins la même quantité d'hommes que de femmes), nous a permis de dégager les profils types au sein de la salle de sport. Nous avons pu mettre en lumière plusieurs résultats majeurs.

Premièrement, grâce à l'analyse bidimensionnelle on a démontré une indépendance entre le genre et le niveau. Être un homme ou une femme n'influe pas sur le fait d'être débutant ou avancé. En revanche, le niveau est fortement corrélé aux variables physiologiques : le niveau "avancé" se distingue nettement par une durée d'entraînement plus longue et un taux de graisse plus faible.

Deuxièmement, l'Analyse en Composantes Principales nous a permis de structurer l'information selon trois méta-variables expliquant près de 86% de l'inertie.
La première méta variable (41,5%) reflète l'intensité de l'entraînement (durée, calories brûlées contre le taux de graisse).
La deuxième méta variable (26,8%) représente la corpulence (poids et IMC).
La troisième méta variable (17, 6%) reflète la taille des individus de la salle.

Enfin, le clustering, mené comparativement via la méthode des K-means et la Classification Hiérarchique (CAH), a convergé vers une partition optimale en 3 classes. Bien que le coefficient de silhouette moyen (0.31) indique une séparation des groupes relativement faible, l'interprétation des classes est riche de sens : la classe "Performance" (Classe 1) regroupe quasi exclusivement les profils "Avancés". Fait notable : cette classe est mixte. Cela confirme que la haute performance athlétique lisse les différences de genre. Deux classes "Loisirs/Intermédiaires" (Classes 2 et 3) : Ces groupes rassemblent les débutants et intermédiaires, qui sont difficilement séparables sur le plan physiologique. Ici, la distinction s'opère principalement par le genre, avec une classe majoritairement masculine et l'autre féminine.

En conclusion, la classification hiérarchique s'est avérée légèrement plus précise que les K-means pour isoler les profils performants et les hommes des niveaux "débutants" et "intermédiaires".