

Projet d'analyse de données

Réalisé par :

Yahia SAADE
Mohamed CHEMLALI

13 Janvier 2026

INSA TOULOUSE

3^{ÈME} ANNÉE MIC-MA

Table of contents

1	Introduction	1
2	Analyse statistique du jeu de données	2
2.1	Exploration du jeu de données	2
2.2	Analyse uni-dimensionnelle	2
2.2.1	Variables quantitatives	2
2.2.2	Variables qualitatives	3
2.3	Analyse bi-dimensionnelle	4
2.3.1	Variables quantitatives - quantitatives	4
2.3.2	Variables quantitatives - qualitatives	5
2.3.3	Variables qualitatives - qualitatives	7
3	Analyse en Composantes Principales (ACP)	8
3.1	Choix du nombre d'axes	8
3.2	Analyse du Premier Plan Factoriel	8
3.3	Analyse de la troisième dimension	9
3.4	Synthèse de l'ACP	9
4	Classification (non-supervisée)	10
4.1	Classification par la méthode des K-means	10
4.1.1	Détermination du nombre de classes	10
4.1.2	Interprétation des classes	12
4.2	Classification avec la méthode CAH	12
4.2.1	Détermination du nombre de classes	13
4.2.2	Interprétation des classes	14
4.3	Comparaison des classifications obtenues	15
5	Conclusion	16

1 Introduction

Dans ce projet, nous effectuons une analyse statistique des habitudes d'entraînement et du profil physique d'un échantillon de 973 membres d'une salle de sport. Le jeu de données regroupe neuf variables réparties en trois catégories : les mesures physiques (poids, taille, graisse, IMC), les données liées à l'effort (durée, calories, eau consommée) et le profil des membres (genre, niveau).

Pour mener à bien ce projet, nous commencerons par une analyse statistique (unidimensionnelle et bidimensionnelle). Nous utiliserons ensuite des techniques de réduction de dimension afin de mieux visualiser les données. Cette étape sera suivie d'une classification (clustering), dont le but est de former des groupes homogènes d'individus. Cette approche permettra de révéler des liens entre les attributs physiologiques et le niveau des sportifs, en faisant apparaître des catégories naturelles au sein de l'échantillon. Ce travail vise à répondre aux interrogations suivantes : existe-t-il des liens ou des corrélations entre les caractéristiques physiques et le niveau de pratique des membres ? Par ailleurs, peut-on identifier des groupes d'individus aux profils similaires qui se forment naturellement au sein de ces données ?

Dans la suite du rapport, la présence du signe : “[Q]” signifiera qu'il y a un complément sur le document quarto.

2 Analyse statistique du jeu de données

2.1 Exploration du jeu de données

Le jeu de données contient deux variables qualitatives, nous distinguons la variable qualitative ordinale niveau (**level**) puisque les niveaux d'expérience sont classés suivant un ordre hiérarchisé et naturel, de la variable genre (**gender**) qui elle, est nominale n'ayant pas cette notion d'ordre, tandis que les sept autres indicateurs (**weight**, **height**, **duration**, **calories**, **fat**, **water** et **bmi**) sont quantitatives continues.

Table 1: Statistiques descriptives des variables quantitatives

	weight	height	duration	calories	fat	water	bmi
Min	40.00	1.50	0.50	303.00	10.00	1.50	12.32
Moyenne	73.85	1.72	1.26	905.42	24.98	2.63	24.91
Médiane	70.00	1.71	1.26	893.00	26.20	2.60	24.16
Max	129.90	2.00	2.00	1783.00	35.00	3.70	49.84
Ecart.type	21.21	0.13	0.34	272.64	6.26	0.60	6.66

L'analyse des statistiques descriptives met en évidence une disparité entre les variables, observable à deux niveaux :

- Les ordres de grandeur : Les variables taille (**height**), durée (**duration**) et eau (**water**) évoluent sur des plages restreintes (respectivement $[1.5, 2]$, $[0.5, 2]$ et $[1.5, 3.7]$). À l'inverse, graisse (**fat**) et IMC (**bmi**) présentent des valeurs intermédiaires ($[10, 35]$ et $[12, 49.84]$), tandis que le poids (**weight**) et surtout **calories** atteignent des niveaux bien supérieurs ($[40, 129]$ et $[303, 1783]$).
- La variance : taille (**height**), durée (**duration**) et eau (**water**) ont une variance très faible (< 1), celle-ci augmente pour la graisse (**fat**) et l'imc (**bmi**) (~ 6), s'élève à 21.21 pour le poids (**weight**), et explose à 272.64 pour **calories**.

Cette disparité de variances nous orientera potentiellement vers un choix d'une ACP particulière dans la suite de cette étude.

2.2 Analyse uni-dimensionnelle

2.2.1 Variables quantitatives

Pour visualiser les données, nous avons privilégié la combinaison du violin plot et du box plot. Cette approche permet d'analyser simultanément la densité (la forme de la distribution) et les statistiques descriptives (médiane, quartiles et valeurs extrêmes), offrant ainsi une vue d'ensemble plus riche de la structure des données.

L'examen des violin plots (Figure 1) montre des distributions différentes selon les variables. Nous observons une asymétrie positive (queue vers le haut) pour l'IMC (**bmi**) et le poids (**weight**), ce

qui révèle la présence d’individus ayant des valeurs particulièrement élevées. Cette corrélation est logique : les valeurs extrêmes d’IMC ($\geq 40 \text{ kg/m}^2$) sont liées aux poids les plus lourds ($> 125 \text{ kg}$). Bien que considérés comme des valeurs extrêmes, ces profils sont normaux et reflètent une réalité démographique : la prévalence de l’Obésité Sévère (IMC ≥ 40) atteint 5% à 10% de la population adulte. Ces observations confirment l’inclusion de sujets non-athlètes ou débutants dans l’échantillon. Parallèlement, les pics de dépenses caloriques (1500 – 1700 kcal) sont également des valeurs extrêmes, mais sont physiologiquement attendus et fréquents chez les Sportifs de Haut Niveau (SHN), ce qui valide la présence de profils “experts” dans le jeu de données. À l’inverse, la variable graisse (**fat**) présente une asymétrie négative (queue vers le bas), signe que les taux de masse grasse très faibles sont peu fréquents dans cet échantillon. Enfin, la densité de la variable durée (**duration**) forme une bosse autour de la médiane, ce qui montre que la durée des entraînements est très similaire pour la majorité des membres.

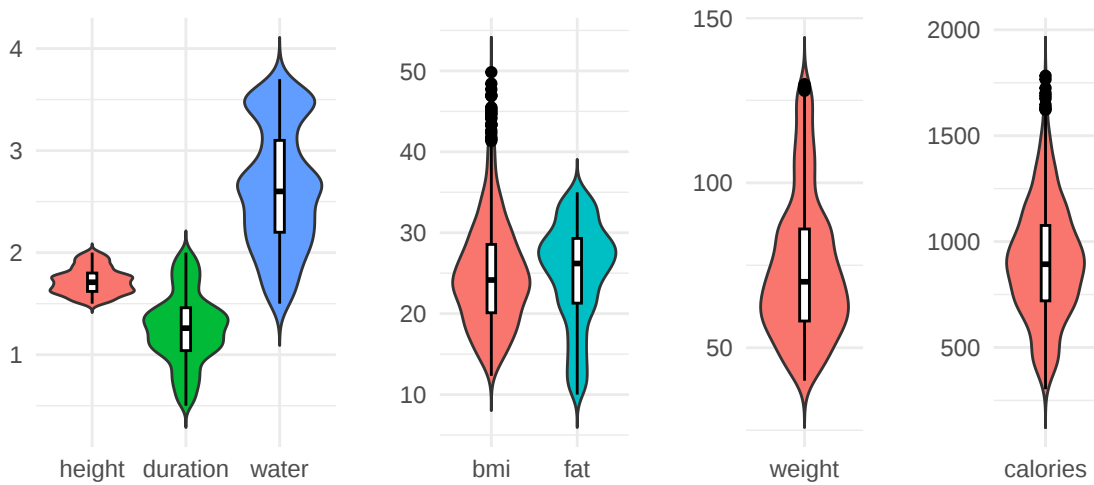


Figure 1: Violin plot des variables quantitatives

2.2.2 Variables qualitatives

Concernant le genre, on observe une parité quasi-parfaite (ratio $H/F \approx 1.1$), avec 53% d’hommes contre 47% de femmes (Figure 2).

Pour le niveau de compétence, l’échantillon se compose majoritairement de profils « Intermédiaire » (40 %) et « Débutant » (38,6 %). Le niveau « Expert » est minoritaire (21,4 %).

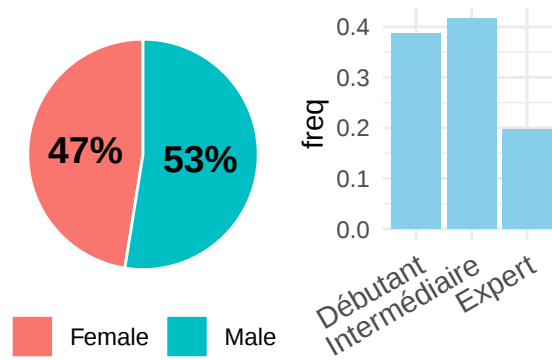


Figure 2: Répartition de l'échantillon selon le genre et le niveau

2.3 Analyse bi-dimensionnelle

2.3.1 Variables quantitatives - quantitatives

Nous analysons la matrice de corrélation (Figure 3) afin de synthétiser les relations linéaires entre les variables. Trois dynamiques distinctes émergent au sein du jeu de données:

- Tout d'abord, nous observons des relations morphologiques logiques : la variable poids (**weight**) est fortement corrélée positivement avec l'IMC (**bmi**) ($r = 0.85$). Cela s'explique mathématiquement, l'IMC étant calculé via la formule suivante : $IMC = \frac{Poids}{Taille^2}$.
- Ensuite, une dynamique liée à l'effort physique apparaît nettement : la variable durée (**duration**) présente une corrélation positive quasi-parfaite avec **calories** ($r = 0.91$). En effet, l'énergie dépensée est, évidemment, proportionnelle au temps d'effort. Nous notons que ces variables d'effort ne sont pas corrélées aux variables morphologiques (**height**, **weight**), ce qui suggère une absence de relation linéaire entre la durée de l'exercice et la corpulence de l'individu dans cet échantillon.
- Enfin, la masse grasse (**fat**) est corrélée négativement avec la variable calories ($r = -0.6$). Elle évolue en sens inverse de l'activité physique (**duration**, **calories**) et de la quantité d'eau consommée (**water**). Cela suggère que les profils les plus actifs et les plus hydratés tendent à avoir un taux de masse grasse plus faible, ce qui confirme le lien direct entre un entraînement intensif et la diminution de la masse grasse.

L'examen de la matrice de corrélation révélant de fortes corrélations (positives et négatives) entre les variables, nous opterons donc dans la partie suivante pour une réduction de dimension via une ACP. Cette méthode permettra de visualiser la structure des données sur des plans factoriels simplifiés sans perdre beaucoup d'information.

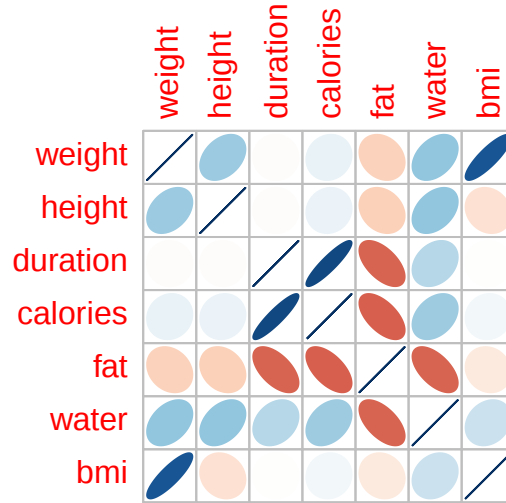


Figure 3: Matrice de corrélations entre les variables quantitatives

2.3.2 Variables quantitatives - qualitatives

Dans cette partie, nous nous intéressons au degré de liaison entre les variables qualitatives (**gender** et **level**) et les variables quantitatives. Plutôt que de visualiser la distribution de chaque variable quantitative en fonction des modalités des variables qualitatives afin d'évaluer l'existence d'une liaison, nous allons procéder au calcul du carré du rapport de corrélation empirique $\eta^2_{Y|X}$. Cet indicateur, défini par $\eta^2_{Y|X} = \frac{S^2_{Y,E}}{S^2_Y}$, mesure la proportion de la variance totale (S^2_Y) expliquée par la variance inter-classes ($S^2_{Y,E}$).

Table 2: Carré du rapport de corrélation empirique ($\eta^2_{Y|X}$)

	weight	height	duration	calories	fat	water	bmi
gender	0.336	0.340	0.000	0.023	0.166	0.446	0.097
level	0.000	0.001	0.622	0.509	0.648	0.169	0.002

En se référant au tableau du carré du rapport de corrélation empirique ci-dessus, nous pouvons tirer les conclusions suivantes sur le degré de liaison :

- Nous observons une liaison claire entre la variable genre (**gender**) et plusieurs variables physiologiques. Le genre explique une proportion significative de la variance de la taille ($\eta^2 \approx 0.34$) et du poids ($\eta^2 \approx 0.34$). L'effet le plus marqué est observé sur la variable eau (**water**), avec un η^2 d'environ 0.45, indiquant que près de la moitié de la variance de la consommation d'eau est expliquée par le genre.
- La variable niveau (**level**) est le facteur qui explique le plus fortement la variance de la masse grasseuse (**fat**, $\eta^2 \approx 0,65$), de la durée (**duration**, $\eta^2 \approx 0,62$) et des calories brûlées (**calories**, $\eta^2 \approx 0,51$). L'influence du niveau est donc majeure sur les variables directement liées à la performance physique et aux efforts.

En conclusion, si la variable genre est un excellent prédicteur des caractéristiques corporelles (**height**, **water**), la variable niveau est le meilleur prédicteur des variables d'effort et de dépense énergétique.

Pour aller un peu plus en détail, nous allons analyser la distribution des variables quantitatives en fonction de niveau (**level**) et / ou du genre (**gender**) là où le carré du rapport de corrélation est élevé pour observer la nature de la liaison entre les deux variables.

Distribution en fonction de level

L'analyse des boîtes à moustaches (Figure 4) révèle une séparation claire et progressive des distributions de valeurs pour les trois variables en fonction du niveau d'expertise (**level**):

- La durée (**duration**) : nous observons une augmentation monotone de la médiane de la durée de l'exercice, qui passe d'environ 60 minutes chez les Débutants à plus de 105 minutes pour les Experts. Il est frappant de constater une diminution de l'étendue interquartile entre les groupes Débutant, Intermédiaire et Expert. Alors que les débutants ont des durées hétérogènes, les experts suivent des séances calibrées et régulières. De plus, la distribution des Experts est nettement décalée vers des valeurs plus élevées, indiquant que le niveau est un puissant facteur discriminant de la durée d'activité physique.
- Calories : Similairement à la durée, la dépense calorique montre une tendance croissante marquée du groupe Débutant à l'Expert. Les distributions des groupes Débutant et Intermédiaire se chevauchent partiellement, tandis que celle des Experts est disjointe et décalée vers le haut (*mdiane* $\approx 1250kcal$), ce qui est cohérent avec une plus longue durée d'effort et une plus grande intensité d'effort.
- Fat : La variable fat présente un profil inverse aux deux précédentes. La distinction Débutant/Intermédiaire est floue (les boîtes se chevauchent largement), ce qui confirme que le passage au niveau intermédiaire est technique plutôt que physique. En revanche, le groupe Expert marque une rupture physiologique bien visible : leur médiane chute drastiquement (d'environ 28% à 15%), isolant ce groupe dans une catégorie athlétique distincte.

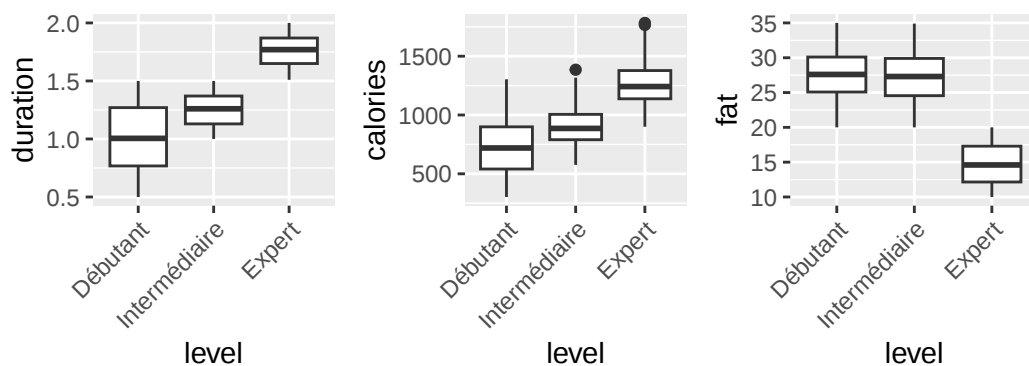


Figure 4: Distribution de la durée des séances, des calories et de la masse grasse selon le niveau

Distribution en fonction de gender

Le graphique (Figure 5) met en évidence des différences claires et systématiques entre les deux groupes de genre pour les trois variables mesurées (**height**, **weight**, **water**).

- Concernant la morphologie (taille et poids), les hommes présentent sans surprise des valeurs médianes plus élevées (environ 1,80 m contre 1,65 m pour la taille, et 85 kg contre 62 kg pour le poids). De plus, les écarts interquartiles du poids et de la taille sont plus grands chez les hommes, signe d'une plus grande hétérogénéité des gabarits masculins dans cet échantillon.
- La consommation d'eau (water) présente la différence la plus marquée : la boîte à moustaches des hommes est intégralement décalée au-dessus de celle des femmes, sans aucun chevauchement. Cela indique une consommation significativement plus élevée chez les hommes. Ce résultat valide l'analyse des corrélations : plus qu'un choix, c'est une nécessité métabolique. L'eau étant fortement liée au poids ($r = 0.39$), le gabarit masculin impose mécaniquement une hydratation supérieure.

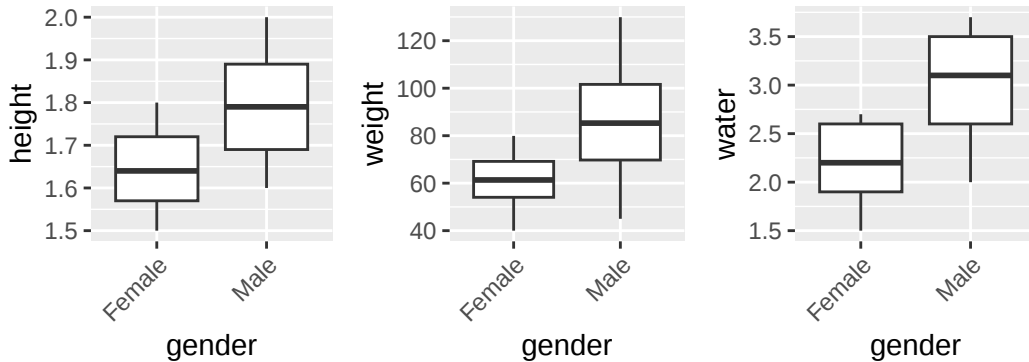


Figure 5: Distribution de la durée, des calories et de la masse grasse selon le genre

2.3.3 Variables qualitatives - qualitatives

Nous calculerons l'indice de Cramer pour étudier la liaison, celui-ci constituant un indice de liaison du χ^2 normalisé.

La formule est donnée par : $V = \sqrt{\frac{\chi^2}{n \times \min(J-1, K-1)}}$ où n est la taille de l'échantillon, χ^2 est l'indice de liaison khi-deux et ($J = 2, K = 3$) le nombre de modalités des deux variables.

Nous trouvons un indice de Cramer égal à 0.004. Ce résultat, très proche de 0, témoigne d'une lien très faible entre les deux variables, voire d'une indépendance. Cela confirme qu'il n'y a pas de biais de genre dans la progression sportive. La distribution des niveaux est indépendante entre les genres.

Table 3: Table de contingence entre Genre et Niveau

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

3 Analyse en Composantes Principales (ACP)

Comme observé précédemment, l'analyse descriptive unidimensionnelle a mis en évidence une disparité importante des variances : les écarts-types s'étendent de 0.13 pour la variable `height` à 272.64 pour la variable `calories`. Afin de ne pas donner un poids disproportionné aux variables à forte variance, nous réalisons une ACP centrée réduite.

L'analyse est réalisée sur les 7 variables quantitatives, les variables qualitatives `gender` et `level` serviront de support d'interprétation visuelle des résultats.

3.1 Choix du nombre d'axes

L'examen de la décroissance des valeurs propres (Figure 6) met en évidence une rupture nette d'inertie après la troisième dimension. De plus, les trois premiers axes capturent 85.9 % de l'inertie totale. Ce taux étant élevé (>80%), nous avons décidé de conserver ces trois dimensions pour l'interprétation.

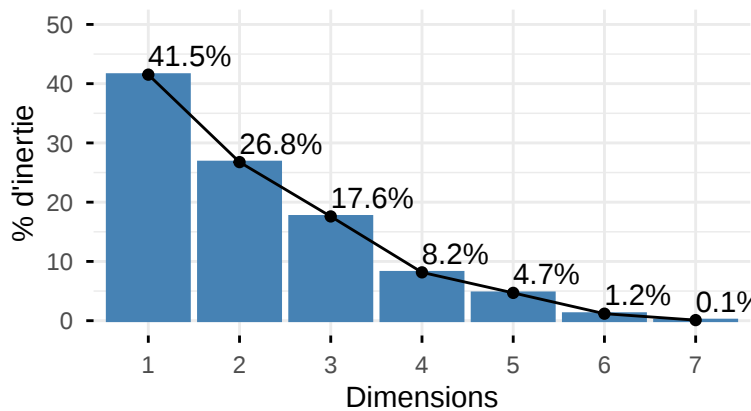


Figure 6: Pourcentage d'inertie expliqué par chaque dimension

3.2 Analyse du Premier Plan Factoriel

Dans cette partie, nous allons interpréter les résultats de l'ACP à l'aide d'une lecture croisée du cercle des corrélations (A) et des projections des individus selon le genre (B) et le niveau (C) (Figure 7).

Nous constatons que le premier axe factoriel est fortement corrélé positivement aux variables d'intensité (`calories` et `durée`) ainsi qu'aux variables de morphologie (taille, poids, IMC et eau), et est corrélé négativement avec la variable masse grasse (`fat`). La projection (Figure C) montre que les individus "Experts" se distinguent nettement en bas à droite du graphique, dans la direction des vecteurs `calories` et `durée`, en opposition à la masse grasse. Nous pouvons donc supposer que le premier axe synthétise la capacité athlétique globale, en liant directement la performance sportive à une constitution physique imposante.

Le deuxième axe factoriel est quant à lui fortement corrélé positivement avec les variables de morphologie (**weight**, **bmi**), et est corrélé négativement avec les variables d'intensité. Nous pouvons donc considérer que ce deuxième axe définit le gabarit physique, en opposant les gabarits lourds aux profils plus légers et sportifs.

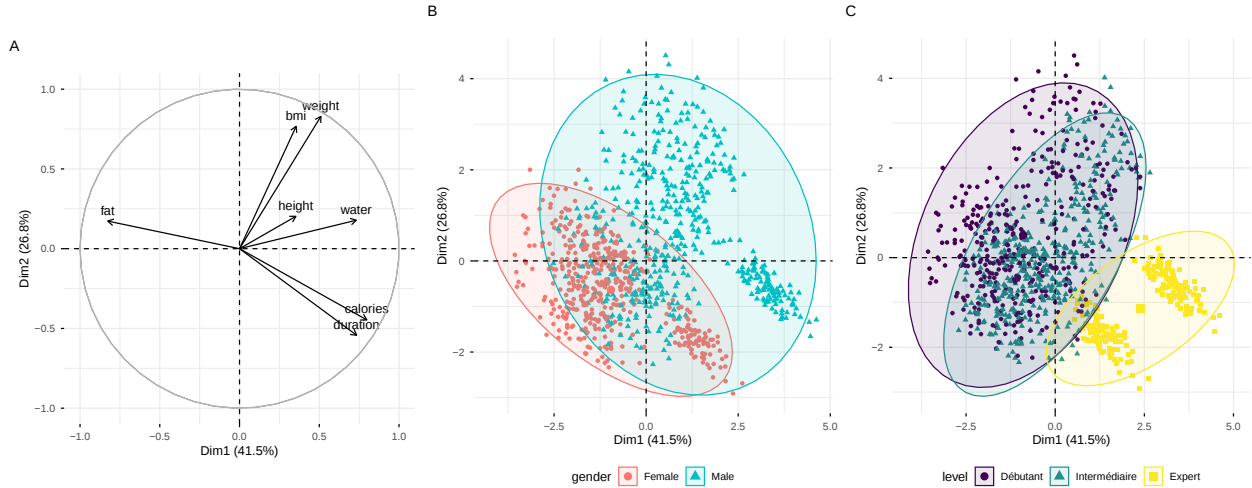


Figure 7: Cercle de corrélation entre les variables initiales et les deux premières composantes principales (A) et Projection des individus dans le 1er plan factoriel selon le Genre (B) et le Niveau (C)

3.3 Analyse de la troisième dimension

Nous complétons l'étude par l'analyse du troisième axe factoriel, qui capture 17.6 % de l'inertie totale.

L'observation de la Figure 8 montre que le troisième axe factoriel est fortement corrélé négativement à la taille, et corrélé positivement avec l'IMC. Cet axe est inversement proportionnel à la taille : plus un individu est grand, plus il se situe bas sur le graphique.

L'IMC présente un comportement intéressant puisqu'il est lié à la fois au deuxième et au troisième axe. Ceci est logiquement liée à la formule mathématique $IMC = \frac{Poids}{Taille^2}$. L'ACP a ainsi permis de décomposer l'IMC en ses deux éléments de base : le poids influence l'axe 2 tandis que la taille définit inversement l'axe 3.

3.4 Synthèse de l'ACP

En conclusion, l'ACP permet de caractériser la population selon deux logiques :

1. Une dimension de performance sportive portée par le premier axe factoriel, structurée par l'intensité de l'entraînement.
2. Une dimension morphologique portée par le deuxième et troisième axes factoriels : le deuxième axe est structuré par le poids (**weight**), tandis que le troisième est défini (inversement) par la taille (**height**).

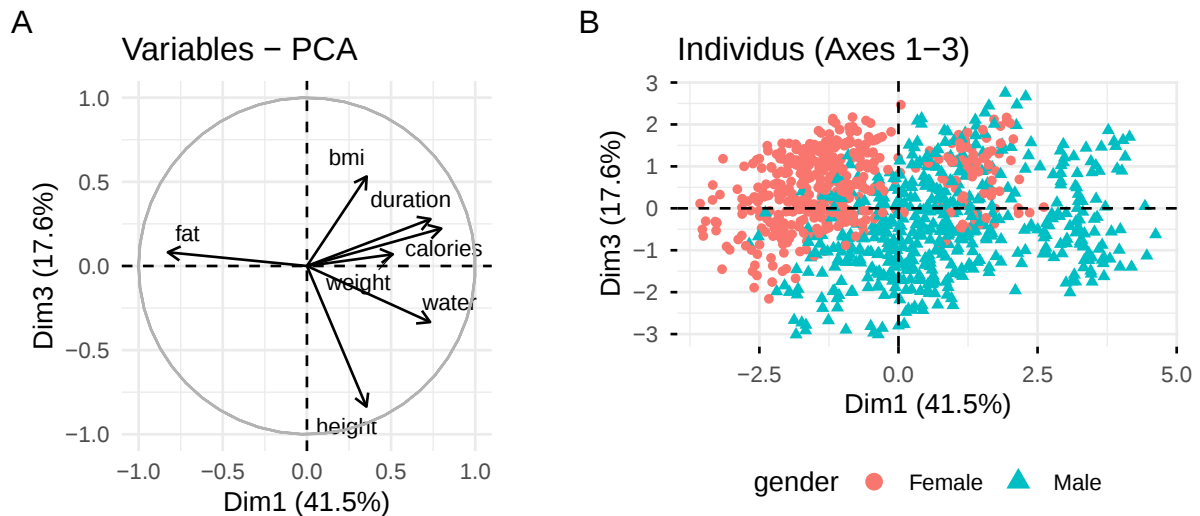


Figure 8: Cercle de corrélation sur les axes 1 et 3 (A) et Projection des individus sur les axes 1 et 3 (B)

4 Classification (non-supervisée)

4.1 Classification par la méthode des K-means

Dans cette partie, nous procédons à une classification non supervisée via la méthode des K-means. L'objectif est de segmenter l'échantillon en K classes homogènes en minimisant l'inertie intra-classe. L'analyse est effectuée sur les données centrées-réduites afin de ne pas biaiser le calcul des distances en accordant le même poids à toutes les variables.

4.1.1 Détermination du nombre de classes

Pour déterminer le nombre optimal de classes, nous avons analysé l'inertie intra-classe et le coefficient de silhouette pour des classifications allant de 2 à 15 classes (Figure 9). Les résultats sont partagés : on observe un coude vers $K = 6$, mais le score de silhouette, lui, atteint un pic pour $K = 3$.

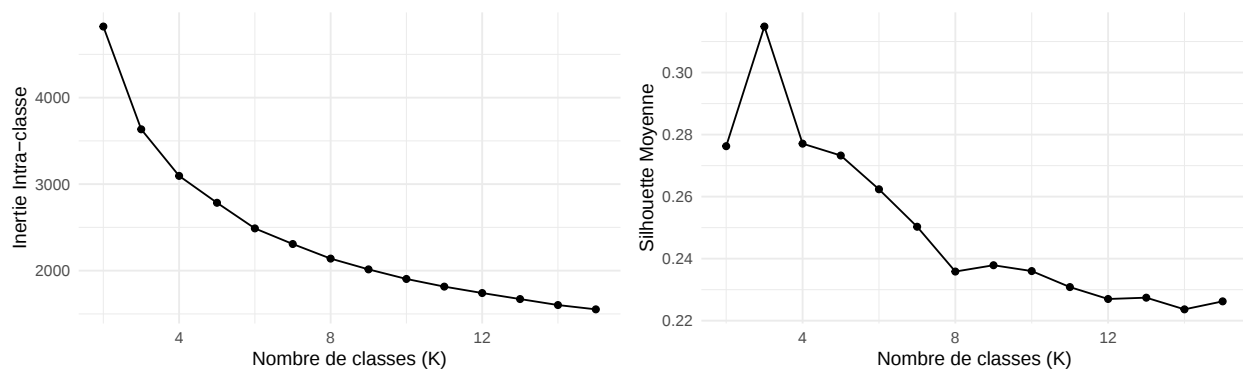


Figure 9: Inertie intra-classe (gauche) et méthode Silhouette (droite)

Pour trancher entre ces deux options ($K = 3$ ou $K = 6$), nous avons projeté les résultats sur le premier plan factoriel de l'ACP (Figure 10). Cela nous permet de juger la qualité visuelle et la cohérence des groupes.

Nous constatons que pour $K = 3$ (à gauche), les classes sont bien séparées dans l'espace factoriel. À l'inverse, pour $K = 6$ (à droite), bien que l'inertie soit plus faible, on observe une sur-segmentation. Les classes sont découpées en micro-groupes très proches géométriquement qui se chevauchent visuellement. Nous privilégions donc la solution $K = 3$ pour sa clarté.

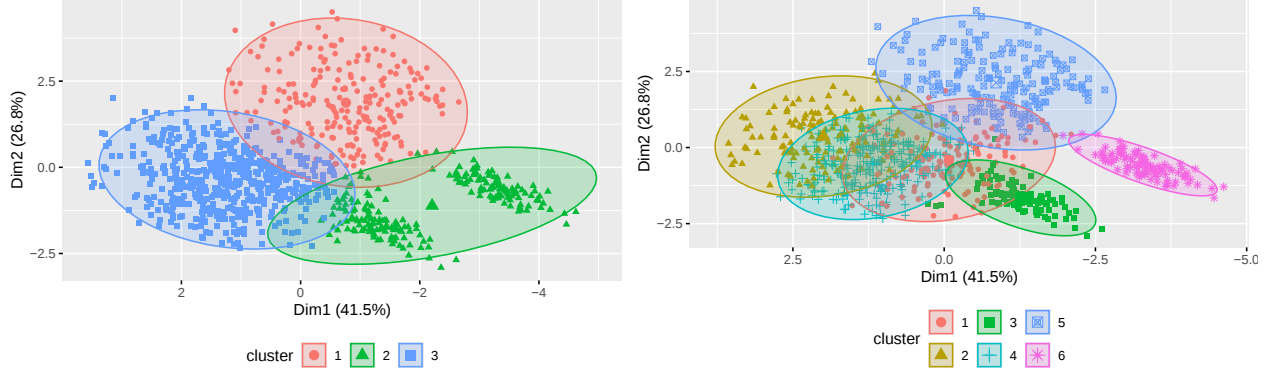
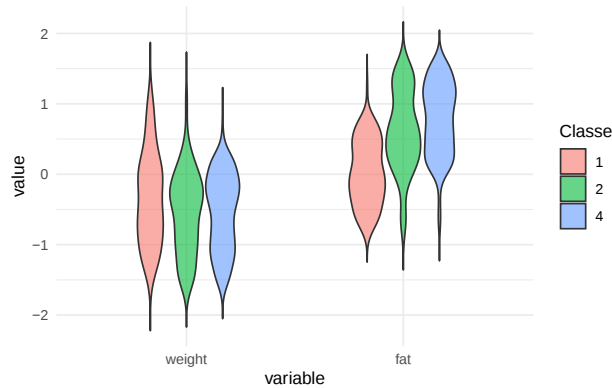


Figure 10: Comparaison des partitions : $K=3$ (gauche) vs $K=6$ (droite) sur le premier plan factoriel

Pour valider notre choix, nous avons croisé notre partition ($K = 3$) avec celle en 6 classes. Le tableau (Table 4) montre que passer à 6 classes n'apporte pas de gain structurel, mais génère une simple fragmentation des classes existantes. Le cas de la Classe 3 est très clair : elle se retrouve éclatée en trois sous-classes (1, 2 et 4). Nous avons comparé les distributions de ces trois sous-groupes sur la Figure 11. Le résultat est net : les courbes se superposent presque totalement. Nous constatons donc que l'algorithme a divisé artificiellement un groupe homogène sans apporter de vraie nuance. Cette redondance nous confirme qu'il est inutile de compliquer le modèle : nous conservons la solution à 3 classes.



	N°1	N°2	N°3	N°4	N°5	N°6
N°1	32	9	1	2	184	0
N°2	18	0	97	0	0	102
N°3	97	138	5	288	0	0

Table 4: Table de contingence entre la classification en 3 classes et 6 classes

Figure 11: Distributions des classes 1, 2 et 4

4.1.2 Interprétation des classes

Pour interpréter ces groupes, nous allons maintenant les croiser avec les variables qualitatives niveau (Table 5) et genre (Table 6) et nous analyserons les profils moyens par classe (Table 7), nous pouvons définir les profils suivants :

- La classe 1 est composée exclusivement d’hommes, tous de niveau Débutant ou Intermédiaire (aucun Expert). Sur l’ACP, cette classe occupe la partie en haut à droite correspondant aux fortes corpulences (avec un poids moyen très élevé de 103.7kg pour 1.79m), ce qui explique un IMC élevé. L’intensité d’entraînement y est modérée (moyenne de 850 calories pour 1.13h).
- La classe 2 est la plus homogène en termes de pratique sportive. Elle capture la totalité des individus “Experts”. Indépendante du genre, elle est caractérisée par une forte intensité d’entraînement (environ 1250 calories pour une durée moyenne de 1.72h).
- La classe 3 est majoritairement féminine (environ 70%). Comme la classe 1, elle ne contient aucun profil Expert et présente une intensité d’entraînement similaire (788 calories pour 1.12h). Cependant, elle s’y oppose regroupant des individus ayant un pourcentage de masse grasse plus élevé (28.66%) et une corpulence plus faible (poids moyen de 61.6kg pour une taille de 1.69m) caractéristiques d’un profil féminin.

Table 5: Répartition des Niveaux

Classe	Débutant	Intermédiaire	Expert
1	110	118	0
2	10	16	191
3	256	272	0

Table 6: Répartition des Genres

Classe	F	H
1	0	228
2	90	127
3	372	156

Table 7: Profils moyens des classes (K-means)

Classe	height	weight	bmi	fat	water	calories	duration
1	1.79	103.7	32.61	25.12	2.98	850.04	1.13
2	1.74	72.3	24.14	15.86	3.13	1248.36	1.72
3	1.69	61.6	21.90	28.66	2.27	788.39	1.12

4.2 Classification avec la méthode CAH

Suite à la classification par la méthode K-means, nous appliquons une Classification Ascendante Hiérarchique (CAH), basée sur la notion de ressemblance, afin de comparer ses résultats avec ceux de la classification précédente. Pour ce faire, nous utilisons la distance euclidienne, particulièrement adaptée à nos variables quantitatives (telles que le poids, la taille et la dépense calorique).

Nous observons des écarts significatifs lors du passage de 2 à 3 classes (Figure 12). Nous pouvons donc conjecturer que la solution à 3 classes est la plus pertinente, ce que nous vérifierons par la suite.

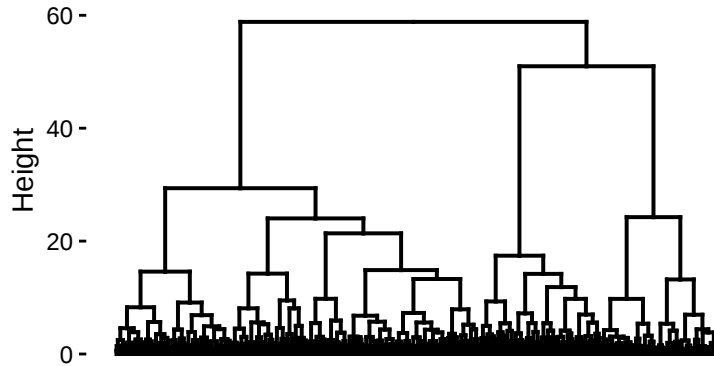


Figure 12: Dendrogramme avec la méthode de Ward (distance euclidienne)

4.2.1 Détermination du nombre de classes

Une fois le dendrogramme établi, nous allons déterminer le nombre optimal de classes. Pour ce faire, nous avons comparé l'évolution de deux indicateurs (Figure 13) : l'indice de Calinski-Harabasz et le coefficient de Silhouette. L'analyse de ces courbes révèle un pic pour $k = 3$. Ce résultat est cohérent avec le dendrogramme, qui présente un saut de dissimilarité significatif entre $k = 2$ et $k = 3$. Nous avons donc choisi de retenir une classification en trois classes.

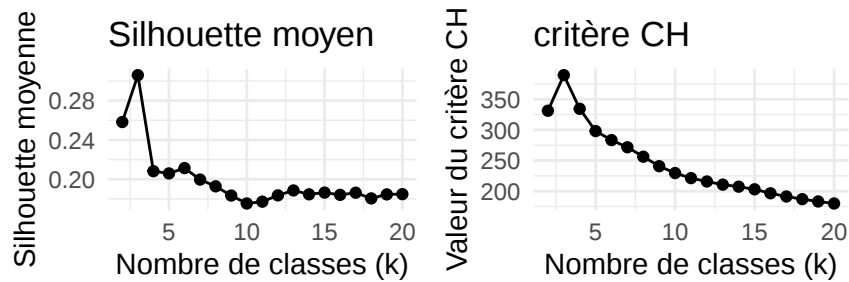


Figure 13: Détermination du nombre de classe

On constate que la projection des individus sur le premier plan factoriel (axes 1 et 2) (Figure 14) issue de la classification hiérarchique présente une allure très similaire à celle obtenue par la méthode des k-means. Cela signifie que les deux classifications obtenus ne sont pas significativement différents.

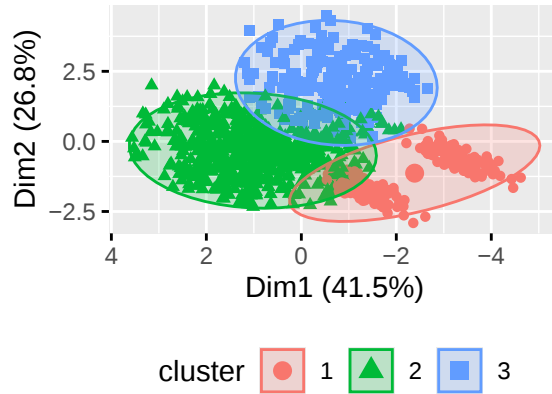


Figure 14: Projection des classes de CAH sur le premier plan factoriel

4.2.2 Interprétation des classes

En nous appuyant sur le cercle des corrélations (Figure 7) et les tables de contingence (Table 8) et la table des profils moyens (Table 11), nous distinguons un profil spécifique pour chaque classe :

- La Classe 1 se distingue par l'intensité d'activité la plus élevée, avec une dépense calorique (1265.34) et une durée d'entraînement (1.76) nettement supérieures aux autres groupes. Les individus présentent la masse grasse la plus faible (14.79 %) et un IMC sain (24.75). C'est le profil type des experts (hommes et femmes).
- La Classe 2 présente la taille moyenne la plus petite (1.70 m) et le poids le plus faible (63.05 kg). C'est cette classe qui possède la masse grasse la plus élevée (28.27 %), malgré l'IMC le plus bas (21.97). Cela correspond aux profils débutants et intermédiaires, majoritairement féminins, qui composent cette classe.
- La Classe 3 est composée uniquement d'hommes. Elle se caractérise par l'IMC moyen, la taille moyenne et le poids moyen les plus élevés (33.93, 1.78m et 106.79kg respectivement).

Table 8: Répartition des Niveaux par classe et répartition des Genres par classe

Cluster	Débutant	Intermédiaire	Expert	Cluster	F	H
1	0	0	191	1	90	101
2	280	307	0	2	372	215
3	96	99	0	3	0	195

Table 11: Profils moyens des classes (CAH)

Classe	height	weight	bmi	fat	water	calories	duration
1	1.72	73.43	24.75	14.79	3.12	1265.34	1.76
2	1.70	63.05	21.97	28.27	2.37	811.69	1.14

Table 11: Profils moyens des classes (CAH)

Classe	height	weight	bmi	fat	water	calories	duration
3	1.78	106.79	33.93	25.03	2.91	835.05	1.11

Le graphique ci-dessous (Figure 15) confirme notre analyse précédente des profils typiques des trois classes obtenues : les individus de la classe 1 se distinguent par une durée élevée associée au plus bas taux de masse grasse, tandis que ceux des classes 2 et 3 présentent une durée d'exercice plus courte et similaire. En revanche, la classe 3 se démarque de la classe 2 par un taux de masse grasse moins important.

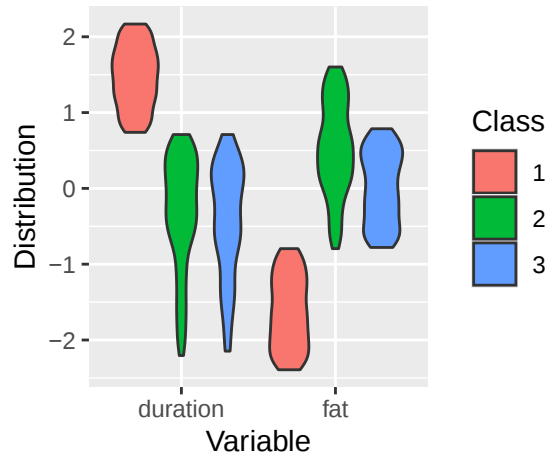


Figure 15: Distribution des variables 'durée' et 'Graisse' par classe

4.3 Comparaison des classifications obtenues

La projection sur le premier plan factoriel de l'ACP suggère une forte ressemblance entre les deux classifications. Ceci est confirmé par l'Adjusted Rand Index (ARI), dont la valeur de 0.8 témoigne d'une forte similarité entre les deux classifications obtenues.

On observe sur la Figure 16 une stabilité remarquable des classes formées par les classifications : la classe 3 (K-means) se retrouve quasi intégralement dans la classe 2 (CAH), tandis que la classe 2 (K-means) migre majoritairement vers la classe 1 (CAH). Quant à la classe 1 (K-means), sa partie majoritaire va dans la classe 3 (CAH) et l'autre partie minoritaire va à la classe 2 (CAH). Cette cohérence globale et le ARI confirment que les deux méthodes identifient des structures de données très similaires.

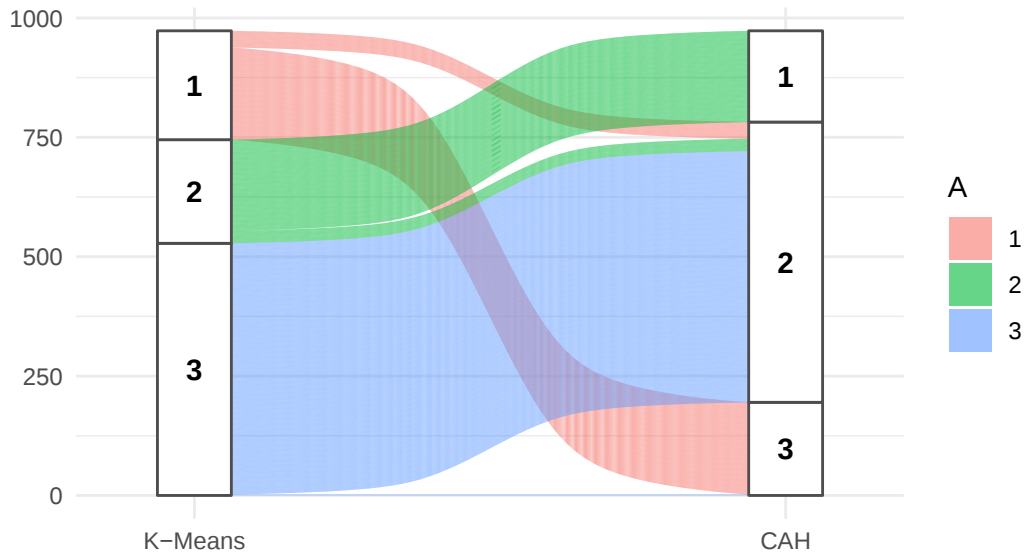


Figure 16: Diagramme d'alluvions des correspondances entre les clusters de K-means et de la classification hiérarchique

5 Conclusion

Au terme de cette étude portant sur un échantillon de 973 adhérents, nous pouvons conclure que dans la salle de sport étudiée, la progression athlétique s'opère indépendamment du genre. L'analyse statistique confirme via un indice de Cramer quasi-nul, une équité parfaite d'accès aux niveaux : le sexe biologique détermine certes les caractéristiques morphologiques (poids, taille), mais c'est uniquement l'investissement sportif qui dicte les performances (durée, calories). De plus, l'étude révèle que le statut Expert marque une véritable rupture physiologique. Contrairement à la transition fluide entre Débutants et Intermédiaires, nous observons une chute drastique de la masse grasse et une augmentation massive du volume d'entraînement chez les Experts. Enfin, les méthodes de classification ont permis d'identifier trois groupes de clientèle distincts : un groupe mixte d'experts focalisé sur la performance, un profil exclusivement masculin à forte corpulence et un profil majoritairement féminin avec une activité sportive modérée.