Projet STA101

Analyse des données : méthodes descriptives

Étude des profils sportifs

LANQUETIN – 2024

Table des matières

1. Introduction

Dans un contexte où le sport et l'activité physique jouent un rôle crucial dans la promotion de la santé et du bien-être, comprendre les interactions entre les caractéristiques physiologiques, les comportements sportifs et les performances individuelles est essentiel.

Les salles de sport, fréquentées par des personnes aux profils variés, offrent un terrain d'observation privilégié pour étudier ces dynamiques.

En outre, l'émergence de technologies telles que les applications sportives, ainsi que les capteurs et montres connectés ont permis de collecter une quantité considérable de données, ouvrant la voie à des analyses plus précises et adaptées.

Ce projet se concentre sur l'exploitation d'un dataset simulé représentant une population fréquentant une salle de sport. Les objectifs principaux sont d'explorer les relations entre caractéristiques individuelles (comme l'âge, le genre, ou la masse musculaire) et comportements d'entraînement, et de segmenter cette population en groupes homogènes.

Cette segmentation peut non seulement guider les entraîneurs et les professionnels du sport dans l'élaboration de programmes sur mesure, mais aussi encourager une meilleure compréhension des besoins et des capacités des pratiquants.

L'étude vise à répondre aux questions suivantes :

- Quels sont les principaux profils d'individus selon leur condition physique ?
- Comment les variables physiologiques, comme la masse musculaire ou la fréquence cardiaque, évoluent-elles en fonction de l'expérience sportive ?
- Comment la durée et la fréquence des séances d'entraînement influencent-elles les performances physiques et la composition corporelle ?
- Peut-on identifier des segments pertinents pour personnaliser des programmes d'entraînement ?

2. Description des données

Chaque individu représente une personne distincte avec des informations détaillées sur ses caractéristiques physiques et cardiovasculaires, ses performances d'entraînement et ses habitudes sportives.

Le jeu de données comprend **322 individus**, ce qui offre une base suffisamment large pour obtenir des résultats statistiquement significatifs et représentatifs de la population habituelle d'une salle de sport.

Le jeu de données est structuré autour de **17 variables**, dont 15 quantitatives, qui offrent un aperçu détaillé des caractéristiques des individus.

Variable	Description	Type
Age	Âge de l'individu (années)	Quantitative (continue)
Gender	Genre de l'individu. Valeurs : Male, Female	Binaire, Qualitative supplémentaire
Weight	Poids de l'individu (en kg)	Quantitative supplémentaire (continue)
Height	Taille de l'individu (en m)	Quantitative supplémentaire (continue)
Max_BPM	Fréquence cardiaque maximale atteinte durant la séance (battements par minute)	Quantitative (continue)
Avg_BPM	Fréquence cardiaque moyenne durant la séance (battements par minute)	Quantitative (continue)
Resting_BPM	Fréquence cardiaque au repos (battements par minute)	Quantitative (continue)
Session_Duration	Durée des séances d'entraînement (minutes)	Quantitative (discrète)
Calories_Burned	Nombre de calories brûlées par séance	Quantitative (continue)
Fat_Percentage	Pourcentage de masse grasse	Quantitative (continue)
Water_Intake	Quantité d'eau ingérée dans la journée (en litres)	Quantitative (continue)
Workout_Frequency	Fréquence d'entraînement hebdomadaire (jours par semaine)	Quantitative (discrète)
Experience_Level	Années d'expérience. Valeurs allant de 0 à 5	Quantitative (ordinale)
BMI	Indice de masse corporelle (Poids/Taille²)	Quantitative (continue)
Muscle_Mass	Masse musculaire (en kg)	Quantitative (continue)
Recovery_Time	Temps de récupération après une séance d'exercice (minutes)	Quantitative (discrète)
Exercise_Type	Type d'exercice pratiqué par l'individu (cardio ou force). Valeurs : Cardio, Strength	Binaire, Qualitative supplémentaire

Tableau 1 – *Description des variables*

Pour mener à bien cette étude, nous envisagerons une analyse en composantes principales (ACP) pour identifier les principaux axes expliquant la variabilité des préférences sportives.

Pour cela, on définira les variable `BMI` (Indice de masse corporelle = Poids/Taille^2), `Fat_Percentage`, et `Muscle_Mass` étant plus pertinentes comme mesures de caractéristiques corporelles d'un individu, les variables `Height` (taille) et `Weight` (poids) seront définies comme illustratives.

Les variables quantitatives telles que le BMI, la fréquence cardiaque, et les durées d'entraînement seront standardisées et considérées comme actives, tandis que des variables qualitatives comme `Gender` (genre) et `Exercise_type` (type d'exercice) seront prises en compte à titre illustratif.

Une ACP centrée réduite sera effectuée du fait du grand nombre de variables quantitative et des unités différentes de ces variables.

Étant donné le grand nombre d'individus, une Classification Ascendante Hiérarchique sera utilisées pour déterminer le nombre de classes, et permettra d'identifier les corrélations entre les caractéristiques des individus et leurs préférences sportives.

3. Statistique exploratoire et descriptive

3.1. Analyse univariée

Une première analyse univariée des variables quantitatives sous forme d'indicateurs statistiques à l'aide de la fonction `summary()` nous donne les valeurs suivantes :

Variable	Min	X1st.Qu.	Median	Mean	X3rd.Qu.	Max
Age	18	23	29	32,26	35	69
Weight	50	69	76	76,75	87	100
Height	1,53	1,663	1,74	1,74	1,81	2,02
Max_BPM	160	181	190	190,2	200	220
Avg_BPM	120	136	142	142,7	150	165
Resting_BPM	60	66	70,5	70,92	75	85
Session_Duration	45	55,25	67	67,39	79	90
Calories_Burned	514	713	820,5	825,3	942,8	1116
Fat_Percentage	12	15	18	19,23	23	30
Water_Intake	2,3	2,3	2,5	2,784	3,1	5,4
Workout_Frequency	1	2	4	3,972	6	7
BMI	16,8	22,7	25,2	25,31	27,77	36,8
Muscle_Mass	15	20,7	29,6	28,25	34,8	40
Recovery_Time	30	43	57	59,07	75	90

Tableau 2 – *Valeurs statistiques*

La population étudiée présente une grande diversité en termes d'âge, de poids, de taille et de niveaux de forme physique. Bien que la majorité des participants soient jeunes, avec une médiane d'âge de 29 ans, il existe des individus allant jusqu'à 69 ans.

Les données montrent une variation importante des caractéristiques physiques, avec des poids allant de 50 kg à 100 kg, et des tailles de 1,53 m à 2,02 m. Les fréquences cardiaques maximales et au repos, ainsi que la durée des séances d'entraînement et les calories brûlées, indiquent des niveaux de forme physique variés, allant des débutants aux athlètes expérimentés.

Les habitudes d'entraînement, telles que la fréquence d'entraînement et la consommation d'eau, varient selon les besoins individuels, avec une médiane de 4 jours d'entraînement par semaine et une consommation d'eau de 2,3 L à 5,4 L par jour.

A l'aide de la fonction `stargazer()`, nous affichons l'écart type pour chaque variable :

Statistic	St. Dev.
Age	12.796
Weight	13.081
Height	0.099
Max_BPM	15.075
Avg_BPM	11.310
Resting_BPM	6.444
Session_Duration	13.565
Calories_Burned	142.813
Fat_Percentage	5.347
Water_Intake	0.645
Workout_Frequency	2.038
Experience_Level	1.710
BMI	3.560
Muscle_Mass	7.687
Recovery_Time	17.957

Les écarts types élevés dans des variables comme l'âge, le poids, la durée des séances d'entraînement et les calories brûlées suggèrent une population avec une diversité de profils typique d'une salle de sport. En revanche, la taille (`Height`) et la fréquence cardiaque au repos (`Resting_BPM`) présentent des écarts plus faibles, car moins influencées par des facteurs extérieurs.

Tableau 3 – *Écart type*

La fonction `ggplot()` sur la variable (`Exercise_Type`) nous permet d'obtenir le diagramme à barres suivant afin d'observer les préférences sportives globales :

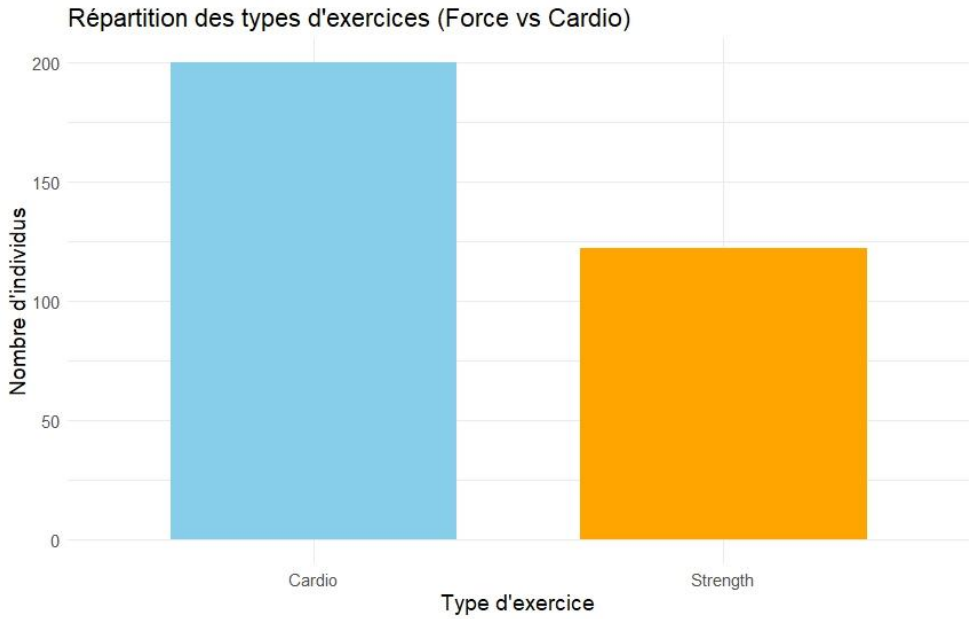

Figure 1 – *Répartition du type d'exercice*

La prédominance des exercices de cardio peut s'expliquer par leur accessibilité et leurs bénéfices généralisés, comme l'amélioration de l'endurance, de la santé cardiovasculaire et la perte de poids, qui sont des objectifs courants pour beaucoup de personnes.

En revanche, les exercices de force, bien qu'importants pour développer la musculature et la force physique, demandent souvent plus de matériel ou de technique, ce qui peut expliquer leur proportion moindre dans la population étudiée.

En complément de cela, la fréquence relative des niveaux d'expérience des usagers fournit une meilleure compréhension de la répartition des utilisateurs en fonction de leur degré d'expertise.

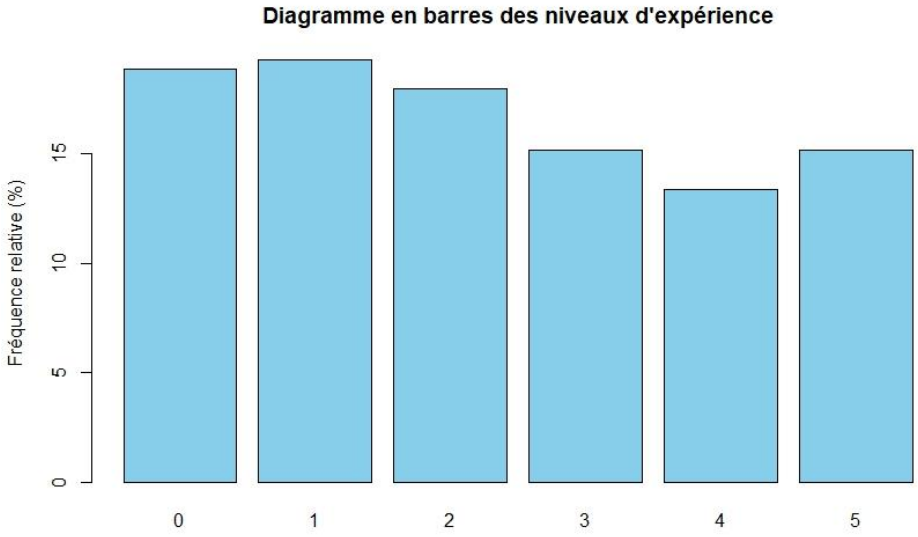

Figure 2 – *Pourcentage d'individus selon le nombre d'années d'expérience*

Le jeu de données représente une distribution homogène parmi les usagers, avec une concentration un peu plus élevée de débutants. On notera qu'au vu de la nature simulée du jeu de données, les taux de débutants devraient être bien plus élevés par rapport au reste en raison de la persévérance requise, des facteurs de motivation, et des défis croissants associés à la progression.

3.2. Analyse bivariée

Comparer la masse musculaire (`Muscle_Mass`) avec les catégories de type d'exercice (`Exercise_Type`) permet d'évaluer l'impact des différents types d'activités physiques sur le développement musculaire.

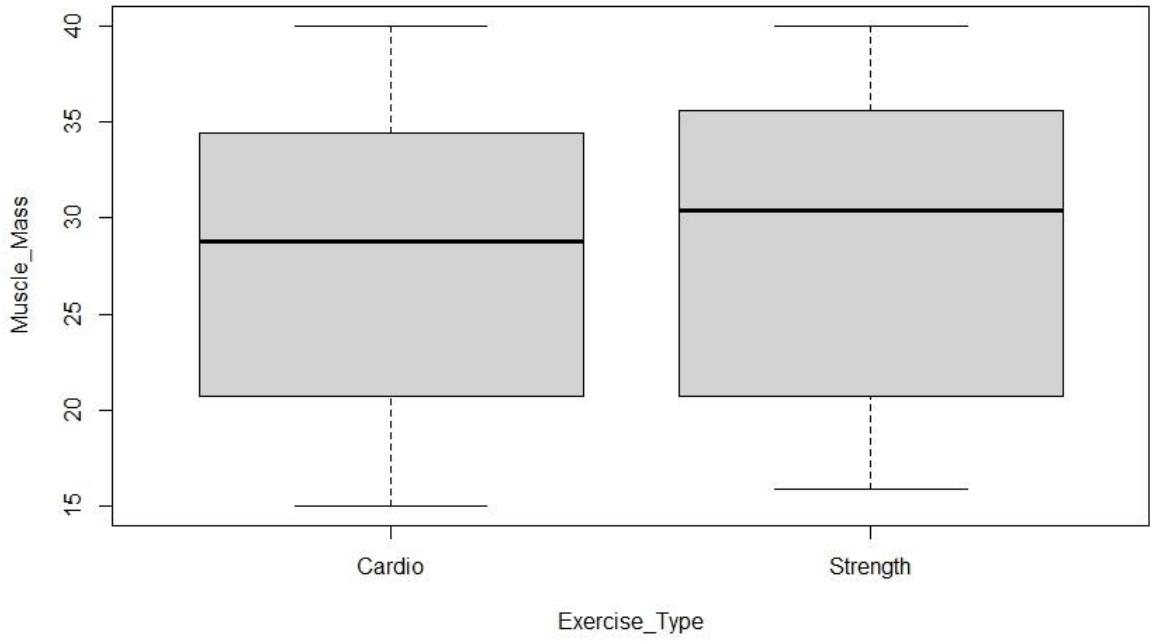

Figure 3 – *Répartition de la masse musculaire selon le type d'exercice*

Source	Df	Sum_Sq	Mean_Sq	F_value	Pr_F
dataset$Exercise_Type	1	110,4	110,387	1,8734	0,172
Residuals	320	18855,7	58,924		

Tableau 4 – *Table des variances*

Les résultats montrent que l'effet du type d'exercice (`Exercise_Type`) sur la masse musculaire (`Muscle_Mass`) n'est pas statistiquement significatif, avec une valeur p de 0.172. Il n'y a donc pas suffisamment de preuves pour conclure que le type d'exercice influence significativement la masse musculaire.

De plus, un coefficient η^2 de 0.0058 représente un effet très faible. Cela pourrait s'expliquer par le fait que les exercices de cardio et de force sollicitent des types de muscles différents et influencent la masse musculaire de manière distincte. On peut supposer également que la plupart des individus de l'échantillon combinent ces deux types d'exercices.

4. Analyse factorielle

4.1. Analyse en Composantes Principale

Une Analyse en Composantes Principales (ACP) est effectuée à l'aide du package `FactoMineR` afin de calculer les valeurs propres associées à chacune des 13 composantes principales.

	Valeurs propres	% Variance	% Variance cumulée
comp 1	2,63561	20,27	20,27
comp 2	2,12105	16,32	36,59
comp 3	1,91991	14,77	51,36
comp 4	1,12492	8,65	60,01
comp 5	1,06629	8,2	68,21
comp 6	0,93669	7,21	75,42
comp 7	0,88792	6,83	82,25
comp 8	0,84185	6,48	88,72
comp 9	0,71001	5,46	94,19
comp 10	0,56295	4,33	98,52
comp 11	0,1283	0,99	99,5
comp 12	0,06417	0,49	100
comp 13	0,00033	0,00	100

Tableau 5 – *Valeurs propres*

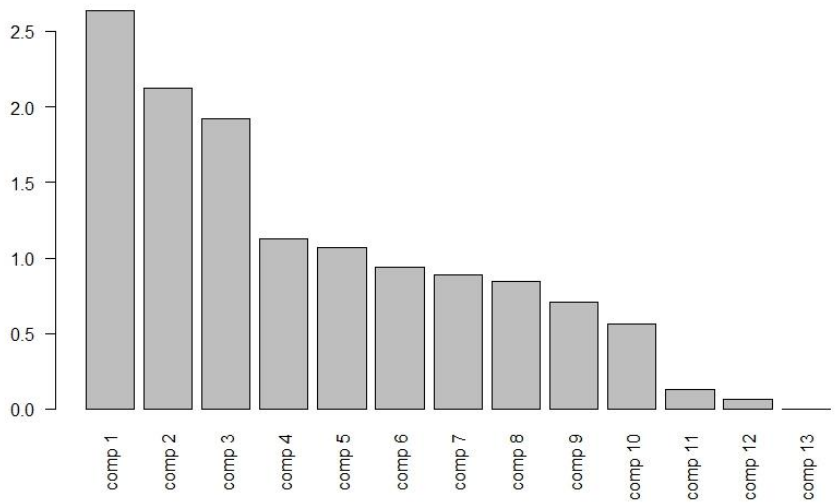

Figure 3 – *Éboulis des valeurs propres de l'ACP*

Selon la règle du coude, l'éboulis des valeurs propres indique qu'il est pertinent de conserver les trois premières composantes principales qui portent 51,36 % de l'information.

La règle de Kaiser (conserver les valeurs propres supérieures à 1 en ACP normée) pourrait nous inciter à retenir deux axes supplémentaire (pour un gain de 16,85 % d'informations), mais un travail beaucoup plus conséquent. Le choix est donc fait de retenir les quatre premiers axes, qui expliquent à eux seuls 60,1 % de l'inertie totale.

Nous tirons parti de la capacité de l'ACP à projeter les individus dans un même sous-espace défini par les axes principaux tout en gardant le genre et le type d'exercice sous forme de variable illustrative.

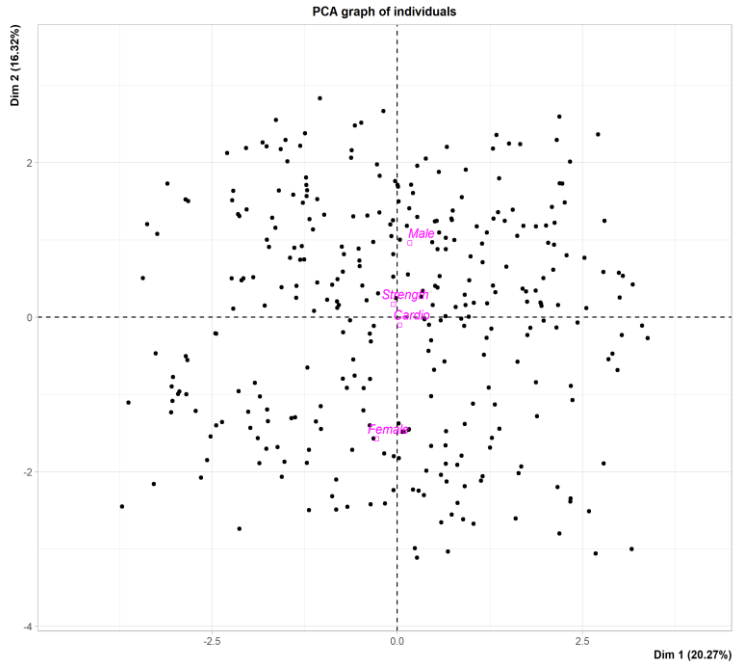

Figure 4 – *Premier plan de l'ACP*

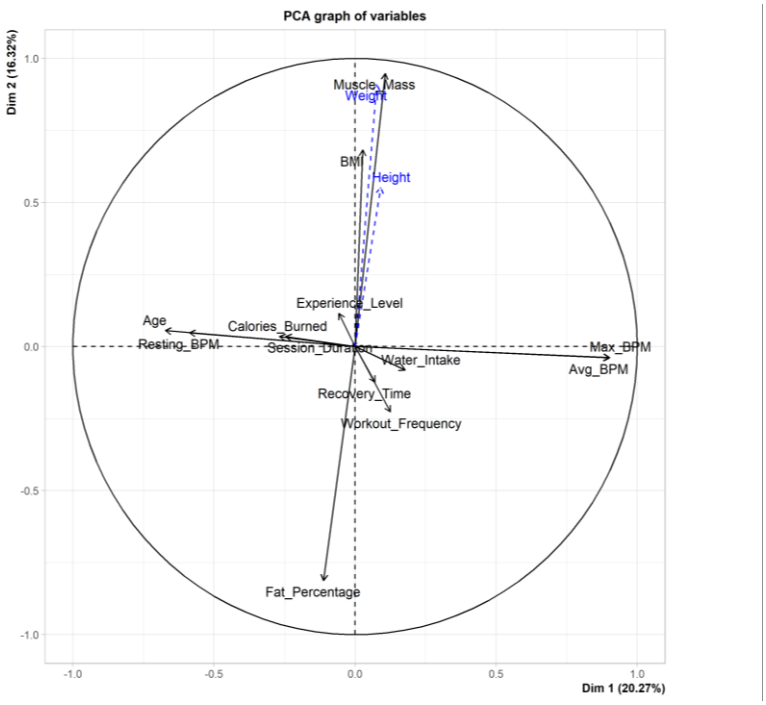

Figure 5 – *Premier cercle des corrélations*

En analysant la position des variables illustratives sur le premier axe factoriel, on constate que les individus placés dans la partie supérieure du nuage sont principalement des hommes (`Male`) favorisant les exercices de force (`Strength`). À l'inverse, ceux situés dans la partie inférieure sont majoritairement des femmes (`Female`) qui privilégient les exercices d'endurance (`Cardio`).

Bien que le nuage de points présente une disposition globalement homogène des individus, le cercle des corrélations des deux premiers axes offre une vision plus nette, permettant de mettre en évidence une relation marquée entre la fréquence cardiaque moyenne (Avg_BPM) et la fréquence cardiaque maximale (Max_BPM), ainsi que modérément entre l'âge (Age) et la fréquence cardiaque au repos (Resting_BPM). En outre, une opposition nette se dégage entre la masse musculaire (Muscle_Mass) et la masse grasse (Fat_Percentage), ce qui suppose un taux très faible d'individus forts mais corpulent dans le jeu de donnée.

Cela suggère que le premier axe reflète la dimension de la santé cardiovasculaire (bonne endurance opposée à faible endurance), tandis que le second axe représente des indicateurs physiologiques musculaires externes (force musculaire opposée à corpulence).

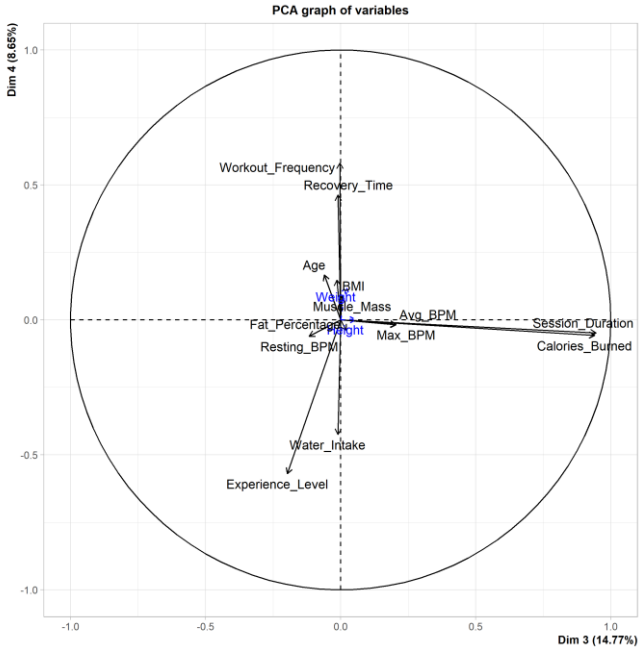

Figure 6 – *Cercle des corrélations du second plan*

D'autre part, un graphique circulaire basé sur le deuxième plan factoriel met en lumière la relation entre la durée des sessions (`Session_Duration`) et les calories brûlées (`Calories_Burned`).

Une matrice des corrélations générée à l'aide de la fonction `corrplot()` nous permet de confirmer les associations et oppositions précédentes de manière numérique.

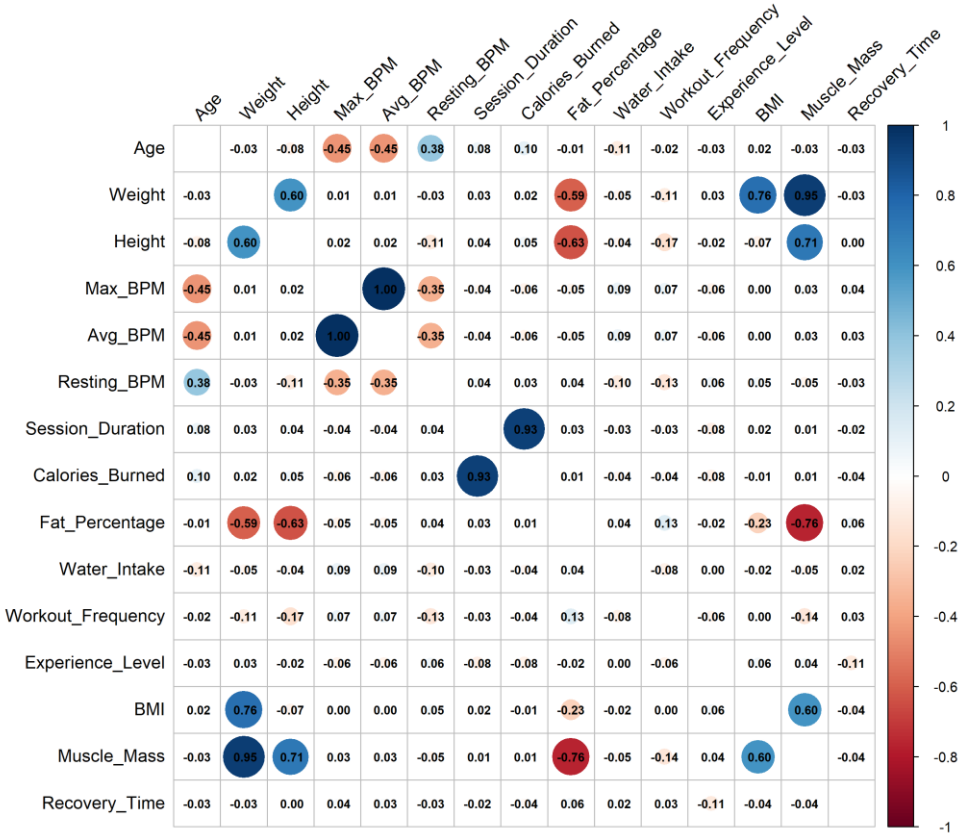

Figure 7 – *Corrélations des variables*

4.2. Classification non-supervisée

Nous effectuons une Classification Ascendante Hiérarchique (CAH) à l'aide de la fonction `cluster()`, afin de déterminer le nombre optimal de classes pour la partition, puis de consolider cette partition. La distance choisie est la distance euclidienne, qui est particulièrement adaptée aux données quantitatives continues. Quant à la méthode d'agrégation, nous utilisons la méthode de Ward, qui optimise le critère de minimisation de l'inertie intra-classe.

Cette classification est effectuée sur les données centrées et réduites obtenues grâce à l'analyse en composantes principales (ACP), ce qui permet de donner davantage de sens à la partition. Une première classification est réalisée à l'aide de cette approche, et le diagramme des gains d'inertie suggère que la partition optimale pourrait être constituée de 4 ou 6 classes. (En rouge sur le diagramme).

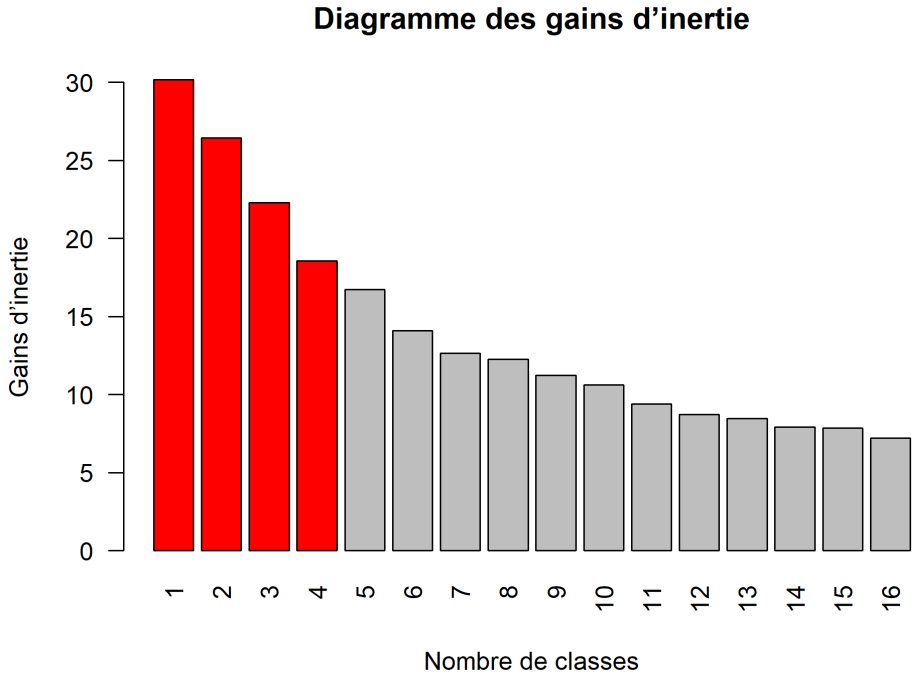

Figure 8 – *Variation de l'inertie intra-classe*

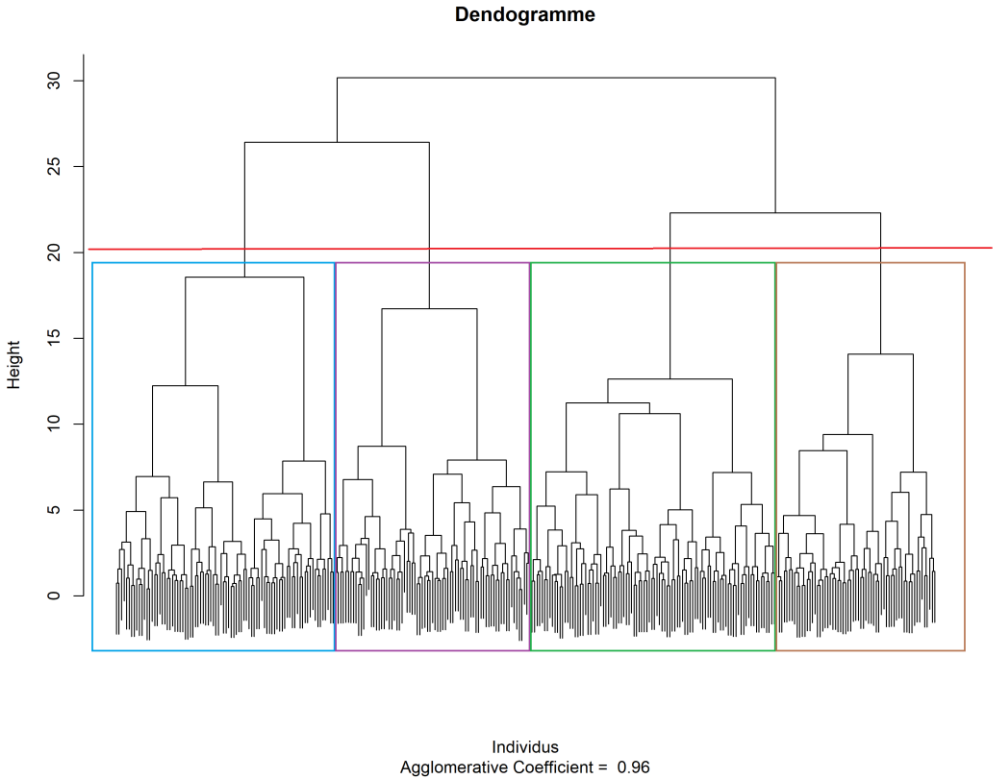

Figure 9 – *Dendogramme de la CAH*

Au vu du dendrogramme, un découpage en 4 classes semble assez naturel, car il marque une séparation nette entre les groupes, tout en maintenant une cohérence interne. Ce choix de 4 classes est renforcé par l'analyse du diagramme des gains d'inertie, qui suggère également que cette partition est optimale. Cela signifie que les individus d'une salle de sport se répartissent naturellement en 4 groupes distincts, chacun présentant des profils d'entraînement et des caractéristiques physiques similaires.

4.3. Interprétation des classes

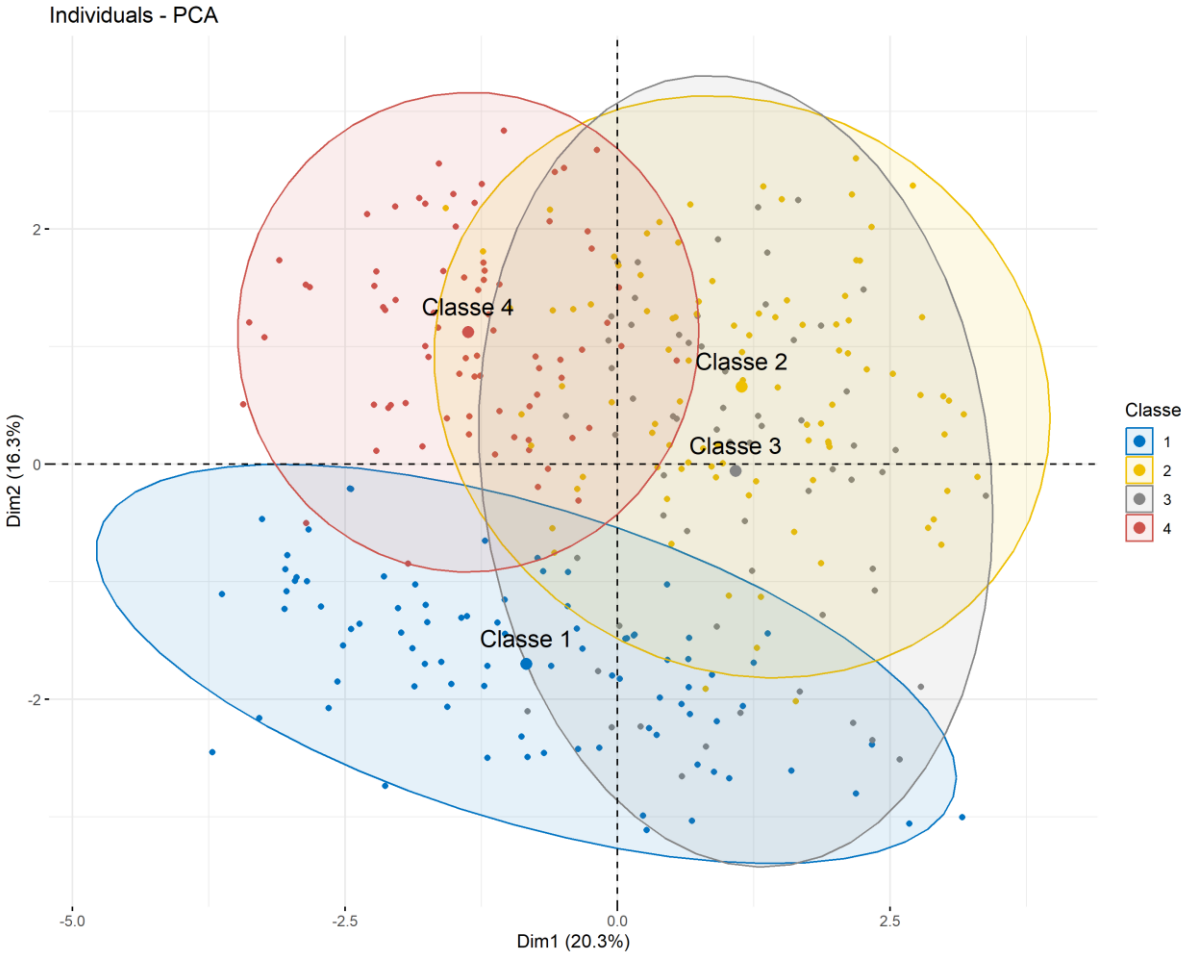

Figure 10 – *Projection des partitions sur le premier plan*

L'analyse préemptive de l'axe 1 permet de séparer distinctement les classes 1 et 4 des classes 2 et 3 : les premières rassemblent des individus globalement peu endurants, tandis que les secondes sont composées d'individus ayant une bonne endurance. La classe 1 semble regrouper des débutants, tandis que la classe 2, caractérisée à la fois par une grande endurance et une masse musculaire élevée, suggère des athlètes expérimentés. Enfin, la classe 3 se distingue par une pratique centrée exclusivement sur le cardio, tandis que la classe 4 met l'accent sur la force, au détriment de l'endurance, au vu de sa position négative sur le deuxième axe.

En outre, l'analyse des résultats présentés dans les tableaux 6, 7, 8 et 9 (en annexe), issus de la sortie de la fonction `catdes()`, permet d'examiner en détail les relations entre les variables quantitatives et les catégories, en mettant en évidence les différences significatives ainsi que les caractéristiques distinctives des groupes analysés.

Ainsi nous constatons que la Classe 1 est dominée par des individus peu endurants, avec une masse musculaire faible, un pourcentage de graisse élevé et une corpulence relativement faible, suggérant des débutants ou des pratiquants moins axés sur les performances intenses.

Quant à la Classe 2, elle regroupe des personnes expérimentées, combinant une grande endurance, une masse musculaire élevée et un gabarit plus important, indiquant une pratique importante de cardio et force.

La Classe 3 est caractérisée par des individus jeunes, axés sur des activités de cardio intenses, avec une grande consommation calorique et une endurance élevée, mais un faible engagement dans les exercices de force.

Enfin, la Classe 4 rassemble des individus plus âgés, très musclés, avec une masse corporelle élevée et une endurance moindre, reflétant une priorité accordée à des exercices de force et de musculation.

5. Conclusion

L'analyse des différentes classes d'individus a permis d'identifier plusieurs profils distincts en fonction de leur condition physique, confirmant ainsi la présence de groupes variés parmi la population étudiée. Les individus se regroupent principalement en fonction de leurs priorités en matière d'entraînement, ce qui se traduit par des différences notables dans leur endurance, leur masse musculaire, leur corpulence et leur composition corporelle.

Les personnes qui privilégient le cardio affichent une endurance supérieure, tandis que celles orientées vers la musculation présentent une masse musculaire plus développée. L'expérience sportive semble avoir un impact significatif sur ces caractéristiques : les individus plus expérimentés combinent souvent une bonne endurance et une forte masse musculaire, tandis que les débutants ou ceux moins axés sur la performance sont moins développés physiquement dans ces domaines. On notera également une présence non négligeable d'individus qui privilégient totalement une activité au détriment de l'autre.

Par ailleurs, la durée et la fréquence des séances d'entraînement influencent de manière marquée les performances physiques et la composition corporelle. Une pratique régulière et prolongée, notamment pour ceux qui privilégient le cardio, contribue à améliorer l'endurance, tandis que des séances orientées vers la musculation favorisent le développement musculaire. En résumé, ces résultats montrent qu'il existe une forte interaction entre l'expérience, les choix d'entraînement et les résultats physiques observés.

6. Annexe

Variable	v.test	Mean_in_category	Overall_mean	sd_in_category	Overall_sd	p.value
Fat_Percentage	11,85	25,08	19,23	3,53	5,34	2,03044E-32
Workout_Frequency	4,35	4,79	3,97	1,92	2,03	1,34853E-05
Age	3,2	36,05	32,26	13,39	12,78	0,001364579
Resting_BPM	3,18	72,81	70,92	6,51	6,43	0,001474911
Water_Intake	-3,38	2,58	2,78	0,42	0,64	0,00071977
Avg_BPM	-4,11	138,38	142,68	10,64	11,29	3,8753E-05
Max_BPM	-4,17	184,43	190,23	14,23	15,05	3,05248E-05
BMI	-6,84	23,06	25,31	3,05	3,55	7,79302E-12
Height	-9,29	1,66	1,74	0,07	0,1	1,49629E-20
Weight	-11,34	63,06	76,75	8,05	13,06	8,09108E-30
Muscle_Mass	-12,8	19,17	28,25	3	7,67	1,63608E-37

Tableau 6 – *Description de la Classe 1 selon les principaux axes*

Variable	v.test	Mean_in_category	Overall_mean	sd_in_category	Overall_sd	p.value
Muscle_Mass	5,79	32,06	28,25	5,34	7,67	7,16669E-09
Weight	5,03	82,39	76,75	9,86	13,06	4,79306E-07
Max_BPM	5,02	196,7	190,23	11,52	15,05	5,28997E-07
Avg_BPM	4,98	147,5	142,68	8,69	11,29	6,23247E-07
Height	4,01	1,77	1,74	0,09	0,1	6,18881E-05
BMI	3,26	26,3	25,31	3,34	3,55	0,001127571
Recovery_Time	-2,93	54,56	59,07	17,56	17,93	0,003360934
Resting_BPM	-5,11	68,1	70,92	5,08	6,43	3,14629E-07
Age	-5,32	26,45	32,26	8,72	12,78	1,06247E-07
Fat_Percentage	-6,06	16,46	19,23	3,91	5,34	1,34352E-09
Calories_Burned	-8,9	716,65	825,32	88,26	142,59	5,60862E-19
Session_Duration	-9,14	56,78	67,39	8,11	13,54	5,96521E-20

Tableau 7 – *Description de la Classe 2 selon les principaux axes*

Variable	v.test	Mean_in_category	Overall_mean	sd_in_category	Overall_sd	p.value
Session_Duration	7,42	78,76	67,39	8,19	13,54	1,16692E-13
Avg_BPM	7,08	151,73	142,68	8,37	11,29	1,42688E-12
Max_BPM	7,07	202,27	190,23	11,12	15,05	1,58989E-12
Calories_Burned	7,02	938,62	825,32	81,8	142,59	2,20088E-12
Water_Intake	4,95	3,14	2,78	0,79	0,64	7,43135E-07
Workout_Frequency	-3,33	3,21	3,97	1,68	2,03	0,000881905
Resting_BPM	-4,04	67,98	70,92	4,9	6,43	5,4578E-05
Age	-5,45	24,38	32,26	4,84	12,78	4,97791E-08

Tableau 8 – *Description de la Classe 3 selon les principaux axes*

Variable	v.test	Mean_in_category	Overall_mean	sd_in_category	Overall_sd	p.value
Age	7,45	41,74	32,26	13,02	12,78	9,34059E-14
Muscle_Mass	6,89	33,52	28,25	4,5	7,67	5,67287E-12
Weight	6,14	84,74	76,75	9,42	13,06	8,17722E-10
Resting_BPM	5,94	74,73	70,92	6,2	6,43	2,85768E-09
BMI	4,51	26,9	25,31	3,33	3,55	6,45799E-06
Height	3,86	1,78	1,74	0,07	0,1	0,000114483
Calories_Burned	2,27	857,55	825,32	150,44	142,59	0,0232086
Workout_Frequency	-3,26	3,31	3,97	1,93	2,03	0,001113157
Fat_Percentage	-5,91	16,09	19,23	3	5,34	3,51173E-09
Max_BPM	-7,63	178,81	190,23	10,76	15,05	2,42528E-14
Avg_BPM	-7,66	134,06	142,68	8,05	11,29	1,82193E-14

Tableau 9 – *Description de la Classe 4 selon les principaux axes*

```
# Initialisation des paramètres
set.seed(42)  # Pour la reproductibilité des résultats

# Nombre d'individus
n <- 322

# Distribution de l'âge, plus de jeunes que de personnes âgées
age <- c(sample(18:35, size = 250, replace = TRUE), sample(36:70, size = 72, replace = TRUE))

# Sexe : Répartition réaliste entre hommes et femmes
gender <- sample(c("Male", "Female"), n, replace = TRUE, prob = c(0.6, 0.4))

# Poids : Répartition réaliste pour les hommes et femmes
weight <- ifelse(gender == "Male",
                 sample(70:100, n, replace = TRUE),
                 sample(50:80, n, replace = TRUE))

# Taille : Répartition réaliste
height <- ifelse(gender == "Male",
                 rnorm(n, mean = 1.80, sd = 0.07),
                 rnorm(n, mean = 1.65, sd = 0.06))

# Fréquence cardiaque maximale (Max_BPM) : Une estimation réaliste
max_bpm <- ifelse(age < 30,
                  sample(180:220, n, replace = TRUE),
                  sample(160:200, n, replace = TRUE))

# Fréquence cardiaque moyenne (Avg_BPM) : Environ 75% du Max_BPM
avg_bpm <- max_bpm * 0.75

# Fréquence cardiaque au repos (Resting_BPM)
resting_bpm <- ifelse(age < 30,
                      sample(60:75, n, replace = TRUE),
                      sample(65:85, n, replace = TRUE))

# Durée des séances (Session_Duration) : en moyenne 45 à 90 minutes
```

```
   session_duration <- sample(45:90, n, replace = TRUE)

   # Calories brûlées par séance
   calories_burned <- session_duration * 10 + rnorm(n, mean = 150, sd = 50)

   # Quantité d'eau ingérée par jour (minimum 2.3 litres)
   water_intake <- pmax(rnorm(n, mean = 2.5, sd = 1), 2.3)

   # Fréquence d'entraînement hebdomadaire (Workout_Frequency) : entre 1 et 7 jours
   workout_frequency <- sample(1:7, n, replace = TRUE)

   # Années d'expérience (Experience_Level)
   experience_level <- sample(0:5, n, replace = TRUE)

   # Calcul de l'IMC (BMI)
   bmi <- weight / (height^2)

   # Type d'exercice (Exercise_Type)
   exercise_type <- sample(c("Cardio", "Strength"), n, replace = TRUE, prob = c(0.6, 0.4))

   # Ajustement des valeurs de masse musculaire et masse grasse en fonction du type d'exercice et de
l'expérience
   muscle_mass <- ifelse(exercise_type == "Strength",
                         weight * (0.35 + (experience_level / 10)),  # Plus d'expérience => plus de
muscle
                         weight * sample(0.30:0.35, n, replace = TRUE))  # Si Cardio, masse musculaire
un peu moins élevée

   fat_percentage <- ifelse(exercise_type == "Cardio",
                            sample(12:20, n, replace = TRUE),  # Plus faible masse grasse en cardio
                            sample(20:30, n, replace = TRUE))  # Plus élevée en force

   # Ajustement de la masse musculaire et de la masse grasse en fonction du sexe
   muscle_mass <- ifelse(gender == "Male",
                         weight * sample(0.40:0.45, n, replace = TRUE),
                         weight * sample(0.30:0.35, n, replace = TRUE))

   # Ajustement du pourcentage de masse grasse en fonction du sexe
   fat_percentage <- ifelse(gender == "Male",
                            sample(12:20, n, replace = TRUE),
                            sample(20:30, n, replace = TRUE))

   # Temps de récupération (Recovery_Time) : en fonction de l'intensité de l'exercice
   recovery_time <- sample(30:90, n, replace = TRUE)

   # Formater les variables pour respecter le nombre de chiffres après la virgule
   max_bpm <- round(max_bpm, 0)
   avg_bpm <- round(avg_bpm, 0)
   resting_bpm <- round(resting_bpm, 0)
   session_duration <- round(session_duration, 0)
   calories_burned <- round(calories_burned, 0)
   fat_percentage <- round(fat_percentage, 1)
   water_intake <- round(water_intake, 1)
   bmi <- round(bmi, 1)
   muscle_mass <- round(muscle_mass, 1)
   recovery_time <- round(recovery_time, 0)
   height <- round(height, 2)

   # Création du data frame
   dataset <- data.frame(
     Age = age,
     Gender = gender,
     Weight = weight,
     Height = height,
```

```
    Max_BPM = max_bpm,
    Avg_BPM = avg_bpm,
    Resting_BPM = resting_bpm,
    Session_Duration = session_duration,
    Calories_Burned = calories_burned,
    Fat_Percentage = fat_percentage,
    Water_Intake = water_intake,
    Workout_Frequency = workout_frequency,
    Experience_Level = experience_level,
    BMI = bmi,
    Muscle_Mass = muscle_mass,
    Recovery_Time = recovery_time,
    Exercise_Type = exercise_type
  )

  # Affichage de l'échantillon du jeu de données
  head(dataset, 10)  # Affiche les 10 premières lignes pour vérification
```

Extrait 1 – *Code R exécuté pour générer le jeu de donnée*