

Projet d'Analyse de Données : Performance en Salle de Sport

Timeo Bossuet, Charles Ganne

Table des matières

1 Chargement et découverte du jeu de données	1
2 Analyse unidimensionnelle	2
2.1 Variables quantitatives	2
2.2 Variables qualitatives	9
3 Analyse bidimensionnelle	11
3.1 Corrélations entre variables quantitatives	11
3.2 Variables quantitatives en fonction de variables qualitatives	14
3.3 Association quantitatives vs qualitatives (η^2)	19
4 Analyse en composantes principales	20
4.1 Analyse du cercle des corrélations	23
5 Classification non supervisée (Clustering)	25
5.1 Classification Ascendante Hiérarchique (CAH)	26
5.2 Méthode des K-means	26
5.3 Comparaison des classifications	27
5.4 Croisement et Profilage	29
6 Bibliographie	30

1 Chargement et découverte du jeu de données

Ce projet a pour objectif de mener l'étude descriptive uni- et bi-dimensionnelle du jeu de données DataGym3MIC disponible sous Moodle. Voici les 6 premières lignes du jeu de données pour s'en faire une première idée.

	gender	weight	height	duration	calories	fat	water	level	bmi
1	Male	88.3	1.71	1.69	1313	12.6	3.5	3	30.20
2	Female	74.9	1.53	1.30	883	33.9	2.1	2	32.00
3	Female	68.1	1.66	1.11	677	33.4	2.3	2	24.71

4	Male	53.2	1.70	0.59	532	28.8	2.1	1	18.41
5	Male	46.1	1.79	0.64	556	29.2	2.8	1	14.39
6	Female	58.0	1.68	1.59	1116	15.5	2.7	3	20.55

Ce jeu de données comprend des mesures empiriques réalisées sur un échantillon de 973 usagers d'une salle de sport. Chaque personne est décrite par les variables suivantes :

- *gender* : Sexe du membre de la salle de sport (homme ou femme)
- *weight* : Poids du membre en kilogrammes
- *height* : Taille du membre en mètres
- *duration* : Durée de chaque séance d'entraînement en heures
- *calories* : Total des calories brûlées au cours de chaque séance
- *fat* : Pourcentage de graisse corporelle du membre (appelé "masse grasse")
- *water* : Consommation quotidienne d'eau pendant les séances d'entraînement
- *level* : Niveau d'expérience : débutant (1), intermédiaire (2) et expert (3)
- *bmi* : Indice de masse corporelle (IMC), calculé à partir de la taille et du poids

On observe que la variable *gender* est qualitative nominale et la variable *level* est qualitative ordinaire. De plus, la variable *calories* est quantitative discrète. Les autres variables sont quantitatives continues. Il faut préciser à R les variables qui doivent être considérées comme qualitatives. On utilise donc la fonction `as.factor()` sur les variables *gender* et *level* :

```
SalleDeSport$gender <- factor(SalleDeSport$gender, levels=c("Male", "Female"), labels = c("Homme", "Femme"))
SalleDeSport$level <- factor(SalleDeSport$level, levels=c("1", "2", "3"), labels=c("Débutant", "Intermédiaire", "Expert"))
```

2 Analyse unidimensionnelle

On commence par étudier chaque variable individuellement.

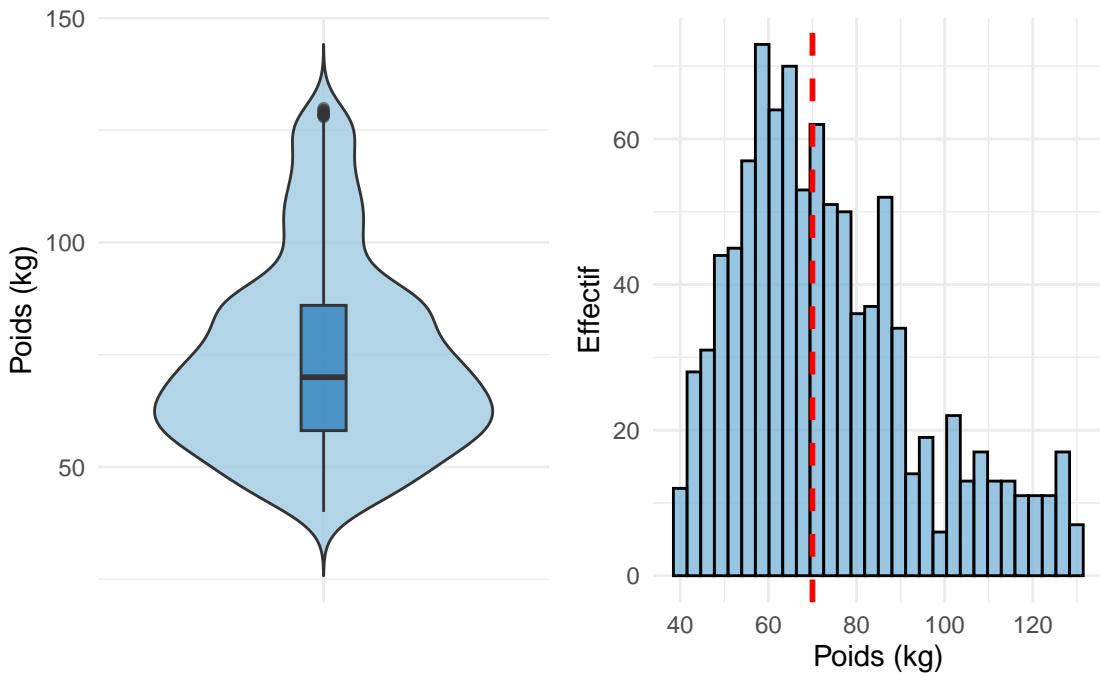
2.1 Variables quantitatives

Pour chaque variable quantitative, on examine des statistiques descriptives (moyenne, médiane, quartiles, etc.). On trace chaque boxplot, violinplot et histogramme.

2.1.1 Poids (weight)

Le boxplot montre une médiane de 70 kg, avec une dispersion notable (la moitié des individus se situe entre 58 et 86 kg). On observe des valeurs atypiques élevées (points au-dessus de ~125–130 kg). L'histogramme confirme une distribution asymétrique à droite : beaucoup d'individus entre ~50 et 85 kg avec un maximum de fréquence autour de 60–70 kg. On observe une queue à droite marquée, correspondant à des individus plus lourds (au-delà de 100 kg), déjà visibles comme valeurs atypiques/outliers sur le boxplot. La distribution n'est donc pas strictement gaussienne. Les 30 bins permettent de bien faire apparaître la queue à droite sans trop lisser la distribution. Un nombre plus faible masquerait les variations dans la zone centrale.

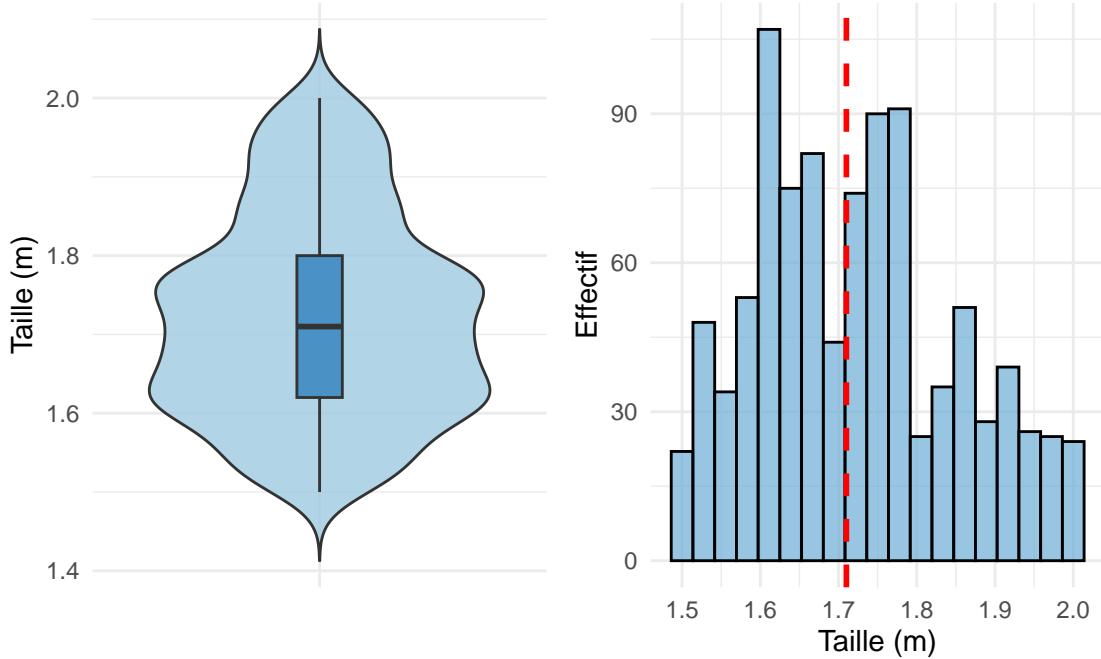
Distribution du poids



2.1.2 Taille (height)

La médiane est proche de 1,70–1,72 m, avec un intervalle interquartile entre 1,62 et 1,80 m. Le boxplot ne met pas en évidence d'outliers. L'histogramme suggère une distribution plutôt concentrée entre 1,60 et 1,80 m, avec deux petites “bosses” (distribution pas parfaitement en cloche, probablement liée au fait que hommes et femmes soient dans le même jeu de données), mais sans valeurs extrêmes dominantes. En traçant les graphiques avec différents nombres de bins, on a remarqué que certaines configurations masquaient le double pic tandis que d'autres le faisaient clairement apparaître. On a donc choisi 19 bins pour qu'ils apparaissent sans trop montrer les petites variations autour.

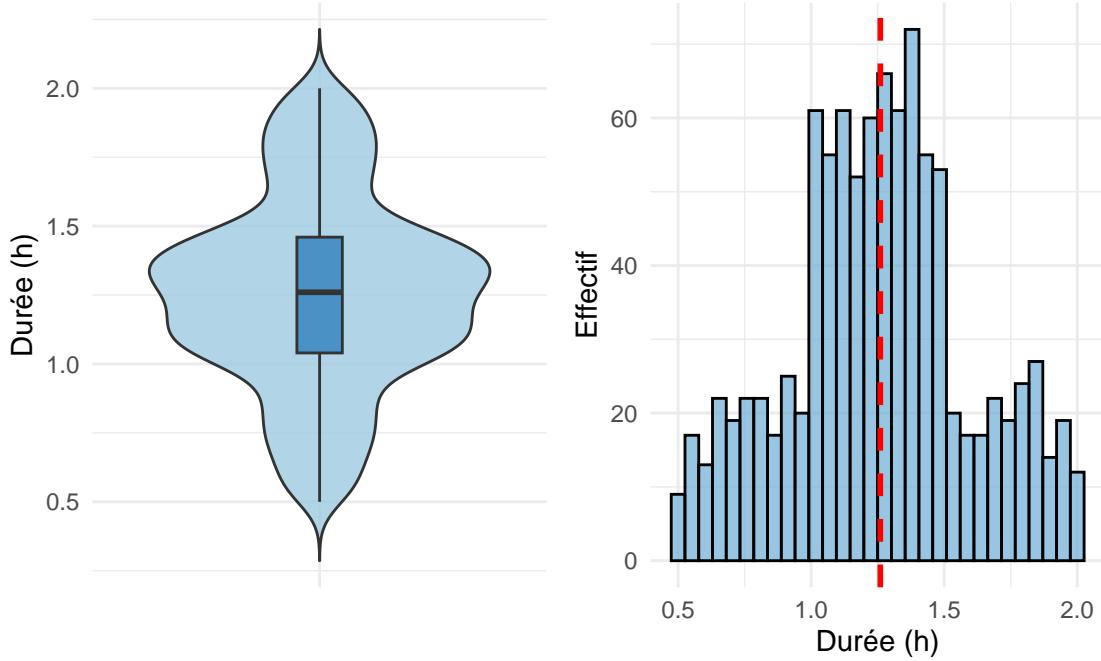
Distribution des tailles des adhérents



2.1.3 Durée (duration)

La médiane est autour de 1,25 h, et la moitié des séances se situe approximativement entre ~1,05 et ~1,45 h. Les durées extrêmes vont d'environ 0,5 h à 2 h. L'histogramme montre une concentration nette entre 1 h et 1,5 h, avec quelques séances plus courtes et plus longues. Ça pourrait par exemple être une recommandation de la salle de sport de faire une séance courte, deux séances moyennes et une séance longue par semaine. Les 30 bins sont pertinentes car elles permettent de bien visualiser le pic entre 1h et 1,5 h sans trop lisser.

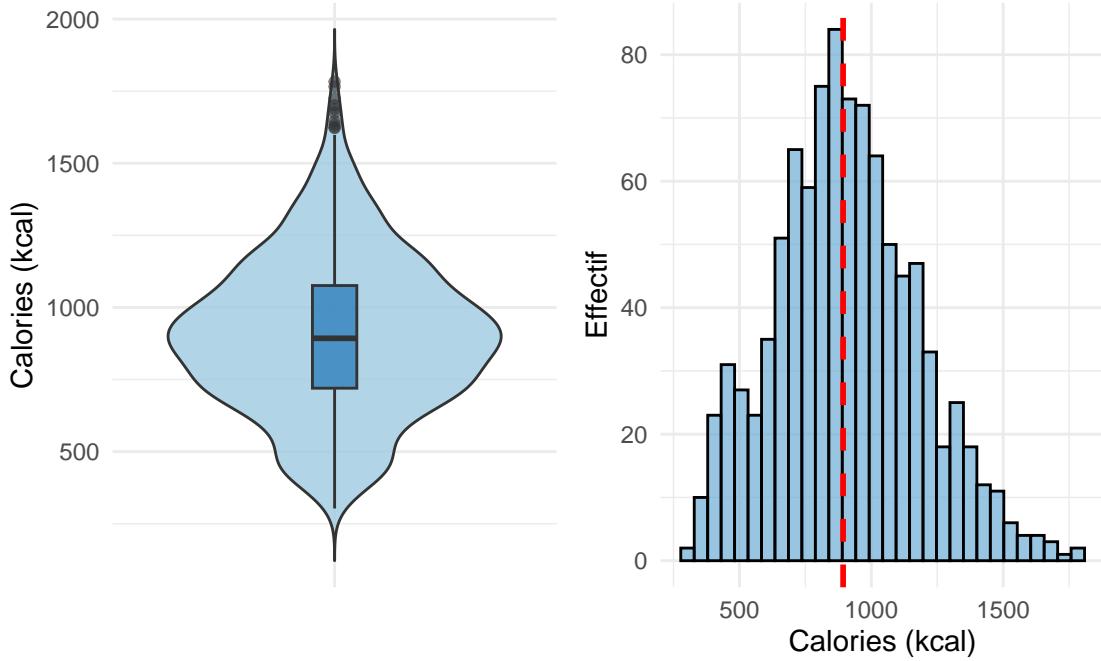
Distribution des durées des séances



2.1.4 Calories (calories)

Le boxplot indique une médiane proche de 900 kcal, avec un intervalle interquartile d'environ 750 à 1100 kcal. On voit plusieurs valeurs atypiques élevées (au-delà de 1600/1750 kcal), ce qui indique des séances exceptionnellement énergivores. Cette structure est cohérente avec une distribution asymétrique à droite : des séances “standard”, et quelques séances très intenses. Les 30 bins sont adaptés ici, ils révèlent la forme globale, tout en conservant la visibilité de la queue à droite. Un nombre plus faible masquerait les extrêmes.

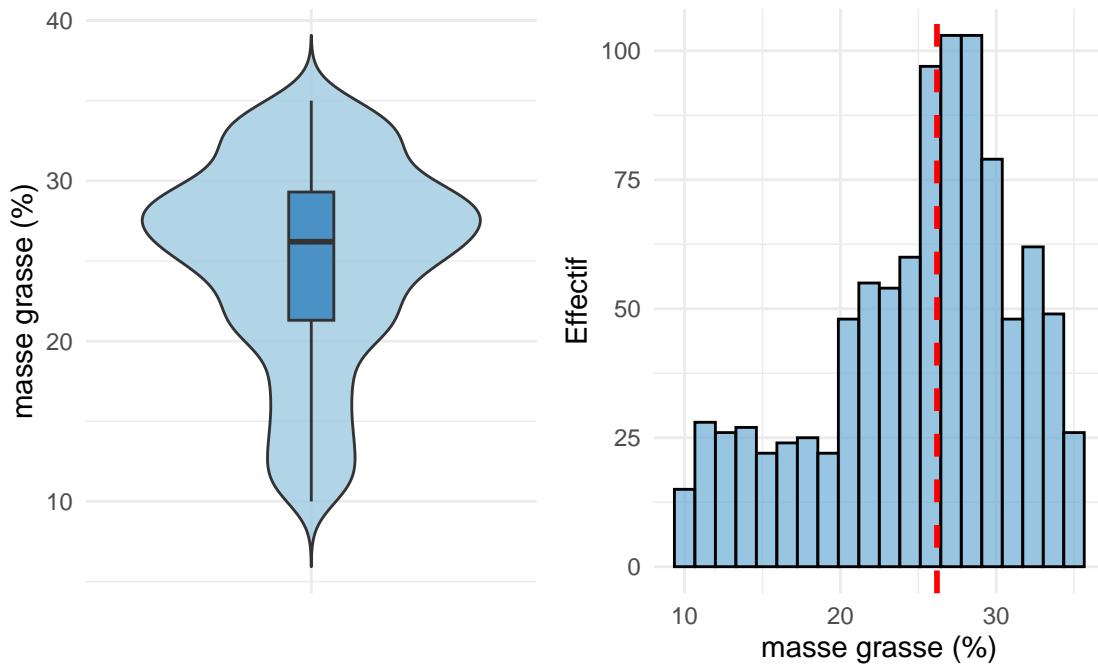
Calories brûlées pendant les séances



2.1.5 Masse grasse (fat)

La médiane est autour de 26%, avec la majorité des valeurs entre 21% et 29%. L'étendue va approximativement de 10% à 35%. Le boxplot ne montre pas de points atypiques extrêmes ; l'ensemble semble relativement homogène, avec une dispersion modérée. L'histogramme semble légèrement multimodal, ce qui peut refléter des différences physiologiques (genres ou niveaux). Les 20 bins sont acceptables, en avoir plus n'apporte pas plus d'informations.

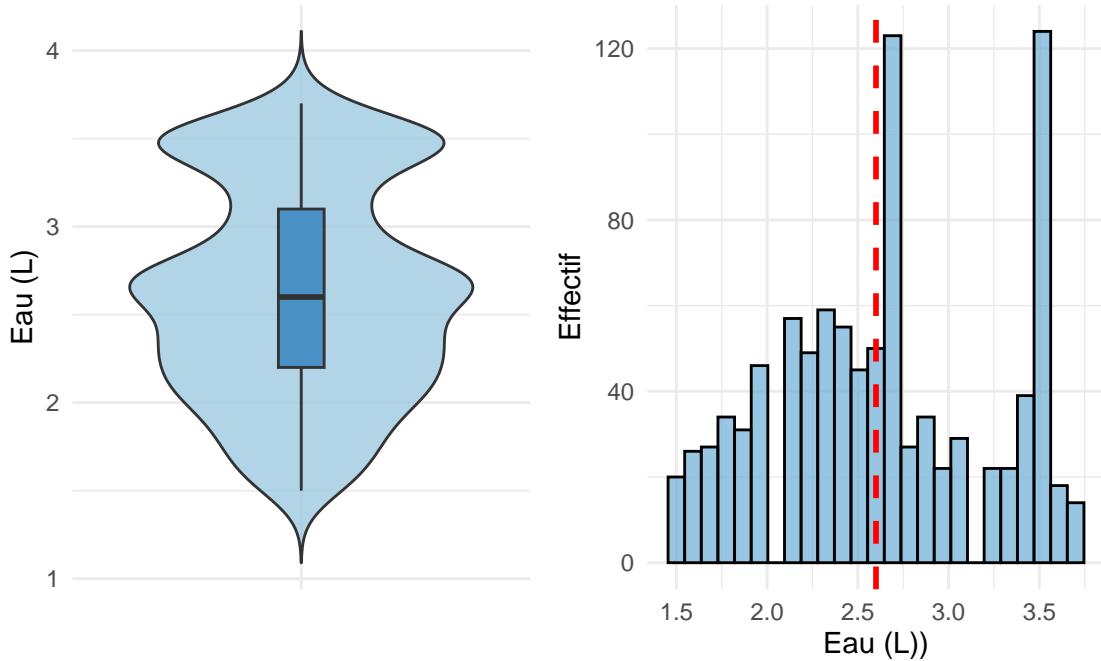
Distribution du masse grasse des adhérents



2.1.6 Eau (water)

La médiane est autour de 2,6 L, et l'intervalle interquartile va de 2,2 à 3,1 L. L'étendue est d'environ 1,5 à 3,7 L. La distribution paraît assez concentrée autour de 2–3 L, sans valeurs extrêmes marquantes. Toutefois on remarque des pics marqués autour de certaines valeurs (notamment 2,7 L et 3,5 L), suggérant des valeurs arrondies ou/et des comportements standardisés (bouteilles, recommandation de cette salle de sport).

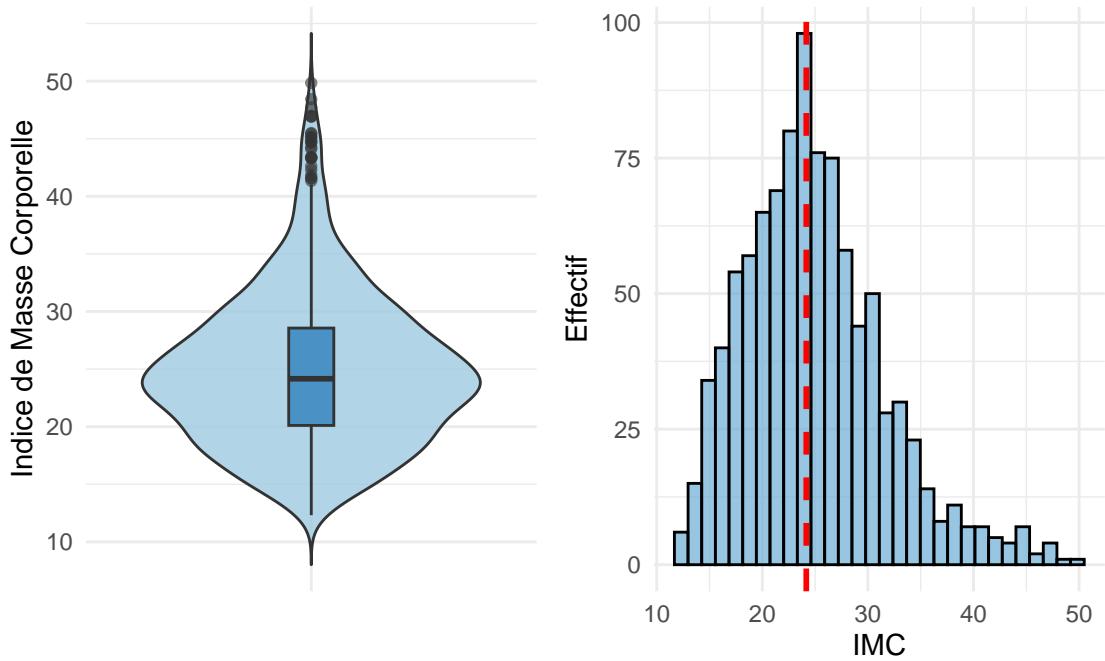
Distribution des consommations d'eau des adhérents



2.1.7 IMC (bmi)

On observe une médiane de 24, avec la moitié des individus entre 20 et 28. En revanche, on observe de nombreux outliers élevés (au-delà de 40 et jusqu'à 50), ce qui traduit une forte asymétrie à droite : la majorité des individus est dans une zone “classique”, mais une minorité présente des IMC très élevés. Cette queue droite importante confirme la présence d’individus en surpoids ou obésité, déjà visible sur le boxplot via de nombreux outliers. Cela rejoint les graphiques sur les poids où on avait remarqué déjà la présence d’outlier avec des individus bien au delà de 100kg.

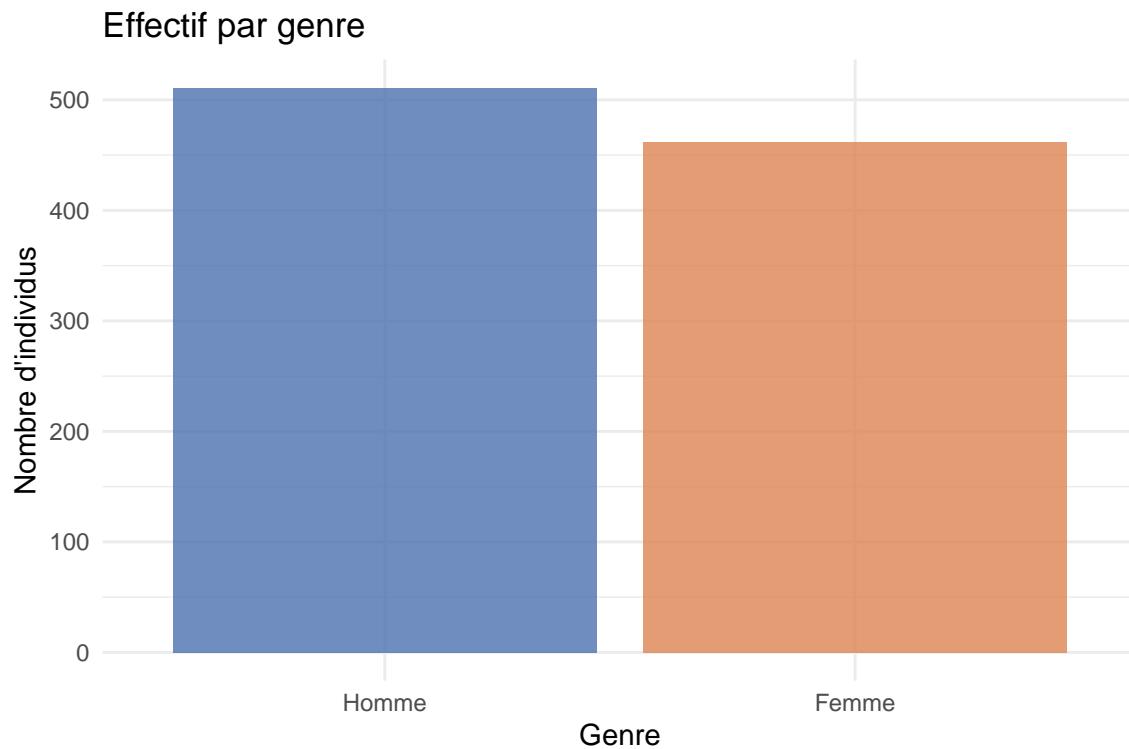
Distribution des IMC des adhérents



2.2 Variables qualitatives

Étudions à présent les distributions des variables qualitatives gender et level. On visualise ces répartitions avec des diagrammes en barres.

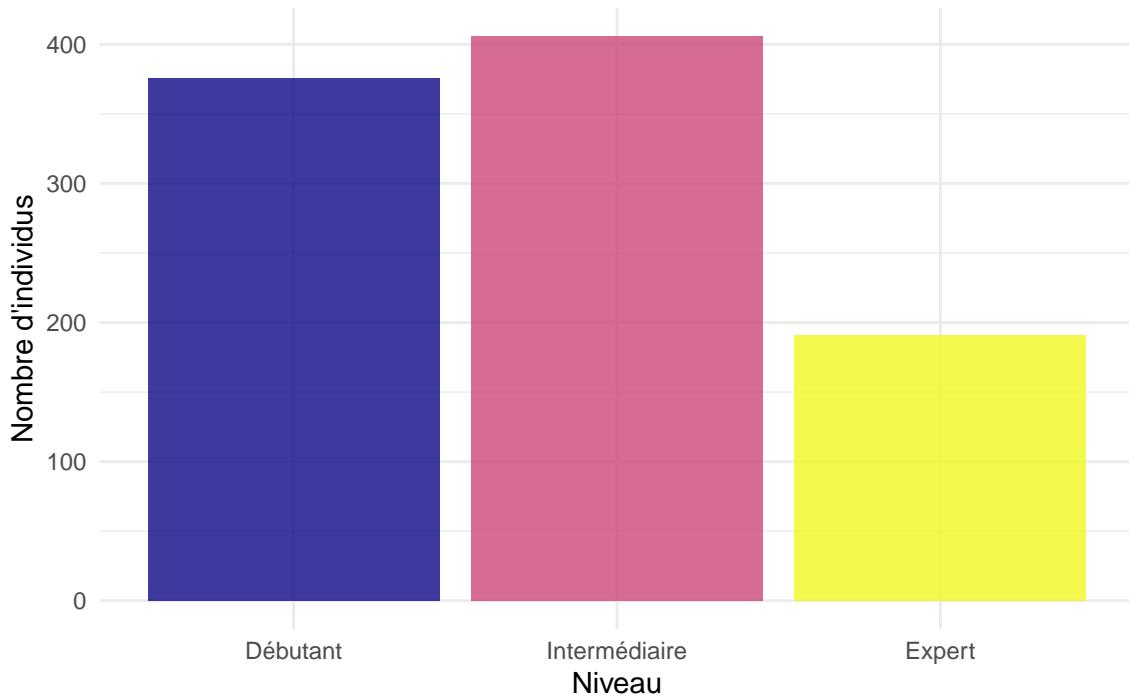
2.2.1 Le genre



Le diagramme en barres montre une répartition relativement équilibrée entre hommes et femmes, avec une légère majorité d'hommes dans l'échantillon. Cette distribution suggère une population globalement mixte, ce qui limite le risque de biais lié à une surreprésentation marquée d'un genre.

2.2.2 Le niveau

Effectif par niveau d'expérience



La répartition par niveau d'expérience met en évidence une majorité d'individus de niveau intermédiaire et de débutants. Les experts constituent le groupe le moins représenté. Cette structure est cohérente avec un contexte de pratique sportive, où les niveaux intermédiaires sont généralement les plus fréquents, tandis que les niveaux experts restent plus rares.

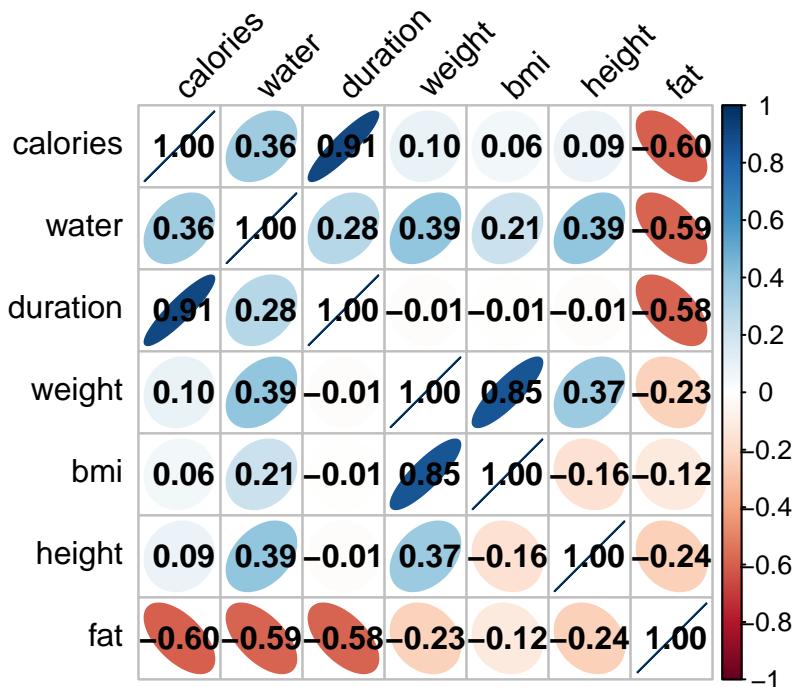
Ces distributions marginales fournissent un premier aperçu de la structure de l'échantillon. Elles serviront de référence pour l'analyse bidimensionnelle, notamment afin d'interpréter correctement les comparaisons entre groupes.

3 Analyse bidimensionnelle

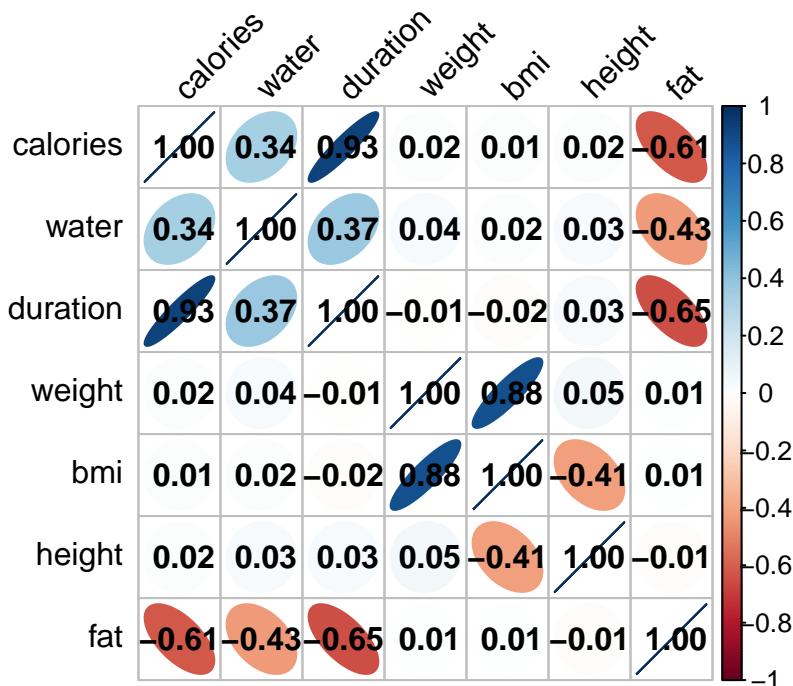
3.1 Corrélations entre variables quantitatives

On explore à présent les relations entre les variables quantitatives avec des matrices de corrélations.

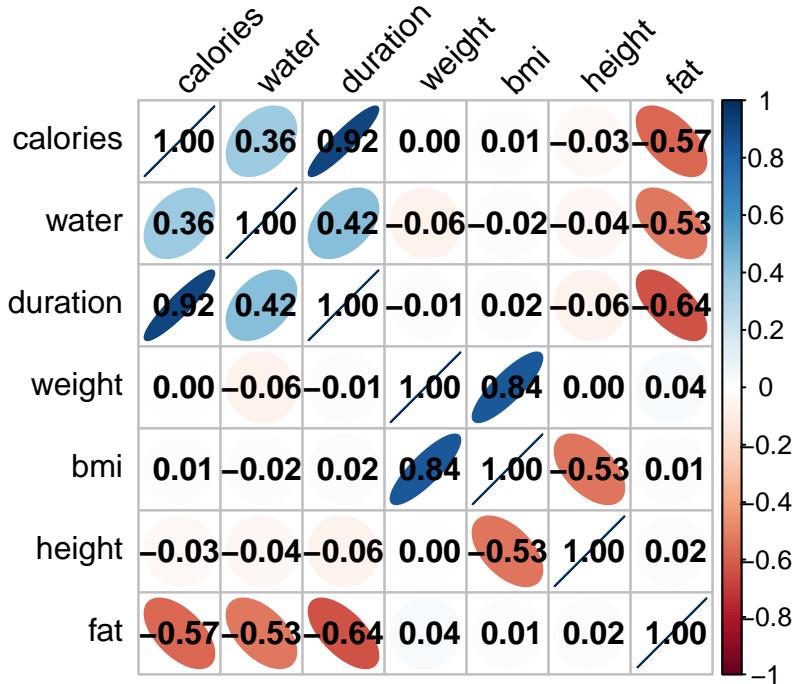
Matrice de corrélation



Matrice de corrélation – Homme



Matrice de corrélation – Femme



On observe d'abord une corrélation positive entre le poids et la taille, les individus plus grands auraient tendance à être plus lourds. L'IMC, défini à partir du poids et de la taille, est logiquement très fortement corrélé au poids. En revanche, sa corrélation avec la taille apparaît faible et légèrement négative dans l'échantillon global. Lorsque l'analyse est menée séparément chez les hommes et les femmes, cette corrélation négative devient plus marquée, ce qui est plus cohérent avec la formule de calcul de l'IMC et suggère que l'effet de la taille est en partie masqué dans l'analyse globale. De plus, la corrélation entre poids et taille disparaît complètement lorsqu'on sépare hommes et femmes.

La relation entre la durée des séances et les calories brûlées est particulièrement forte, ce qui confirme que la dépense énergétique dépend avant tout du temps consacré à l'effort. La consommation d'eau est également positivement liée à la durée et aux calories, bien que de manière plus modérée, traduisant des comportements d'hydratation variables d'un individu à l'autre. Ces relations restent globalement stables lorsque l'on distingue les hommes et les femmes, ce qui indique des mécanismes communs indépendants du genre.

Le masse grasse se distingue par des corrélations négatives marquées avec la durée des séances, les calories brûlées et la consommation d'eau. Cela suggère que les individus présentant un pourcentage de masse grasse plus élevé ont tendance à réaliser des séances plus courtes et moins intenses, accompagnées d'une hydratation plus faible. À l'inverse, les variables morphologiques telles que le poids et la taille ne sont pas liées aux indicateurs de performance sportive considérés ici. Cette dissociation indique que, dans ce jeu de données et avec ces variables, la morphologie seule n'explique pas directement les performances, qui semblent davantage dépendre de l'effort fourni et de la composition corporelle.

Enfin, il est étonnant de voir que le masse grasse n'est pas corrélé au poids ou à l'IMC. Cette variable n'augmente donc pas mécaniquement avec le poids, surtout dans une population sportive

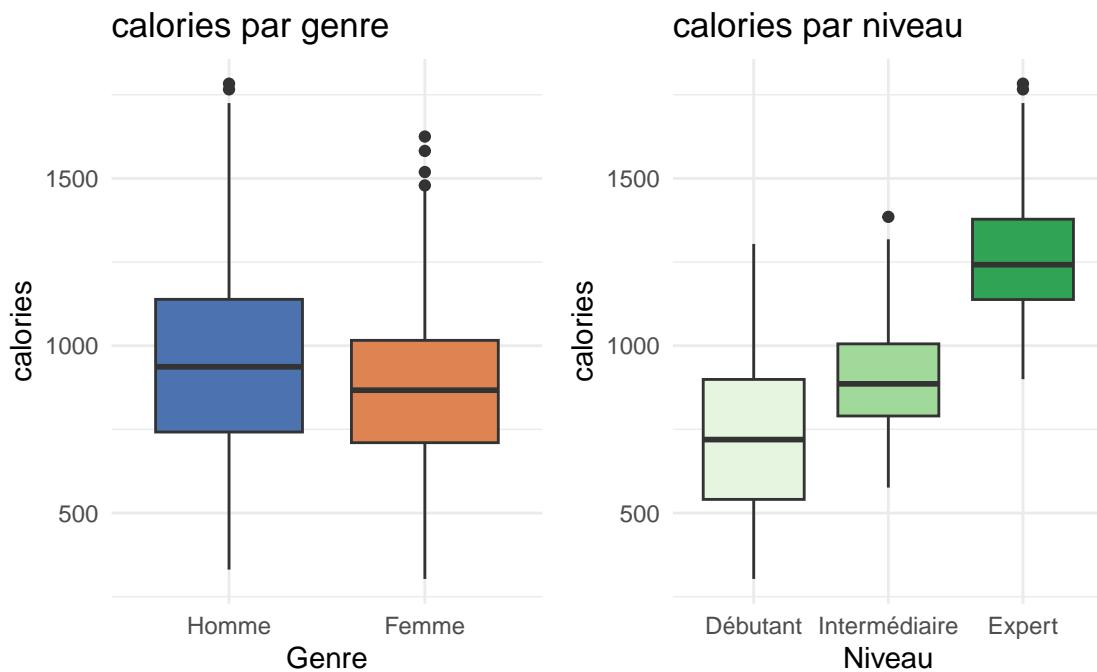
où la masse musculaire peut représenter une part importante du poids total. Cette observation avait déjà été faite lors de l'expérience super-size me. Le sujet n'avait pas pris beaucoup de poids mais était quand même devenu en très mauvaise santé, notamment à cause de son masse grasse (et beaucoup d'autres troubles liés à la junk-food).

Un nuage de points générique calories vs duration coloré par genre permet de visualiser ces effets :

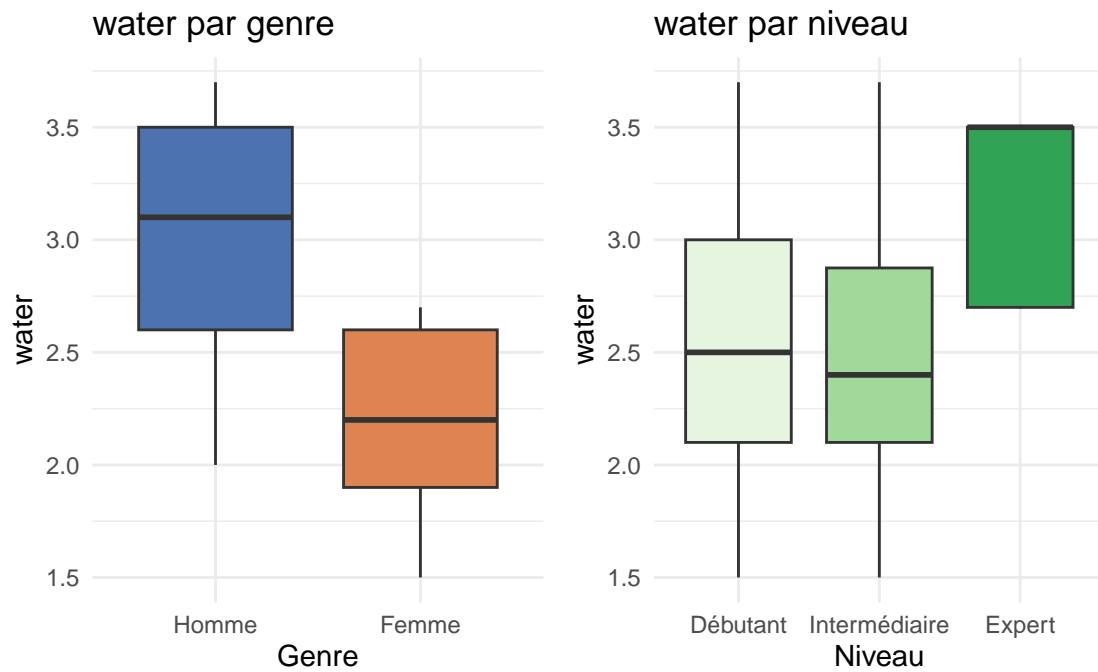
3.2 Variables quantitatives en fonction de variables qualitatives

Comparons les distributions des variables quantitatives selon les catégories de gender et de level à l'aide de boxplots.

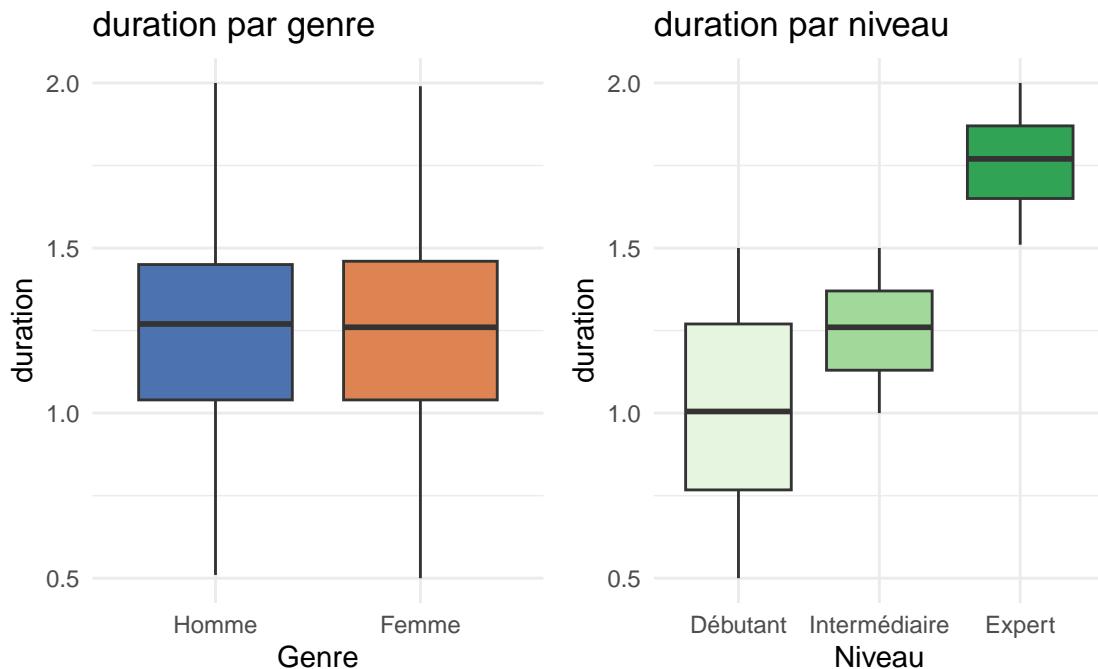
Boxplots de calories selon genre et niveau



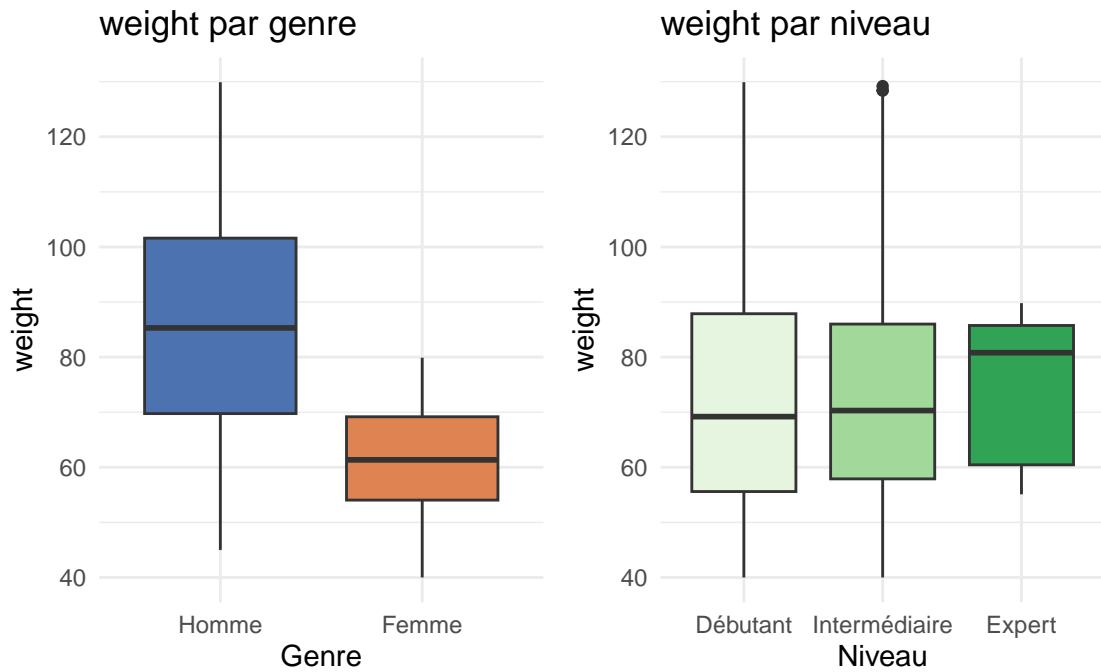
Boxplots de water selon genre et niveau



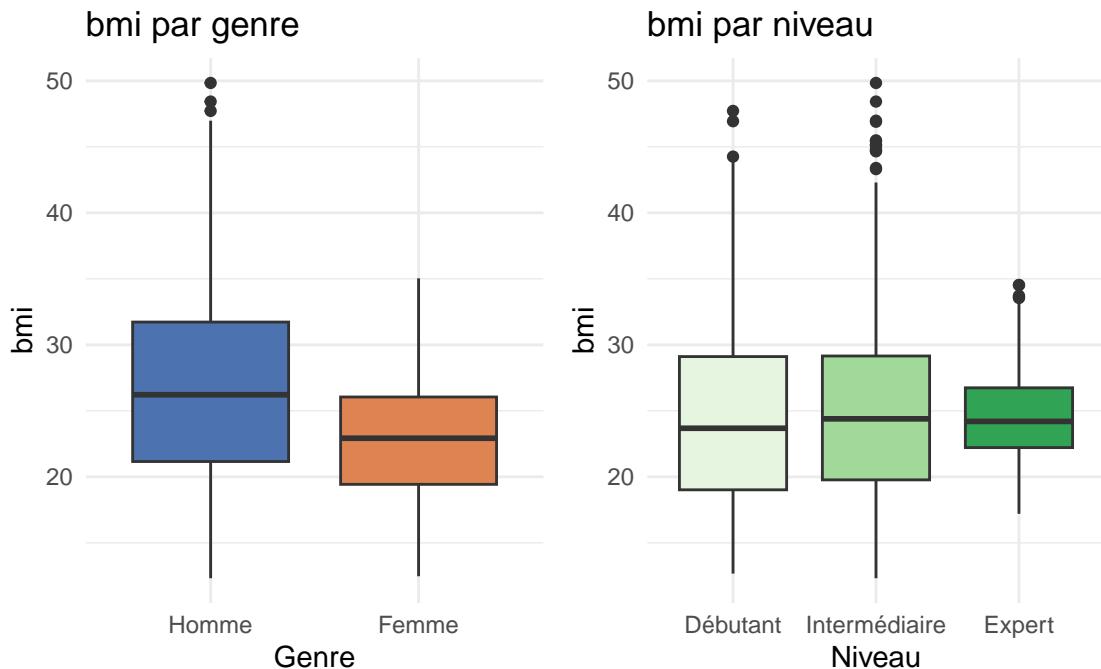
Boxplots de duration selon genre et niveau



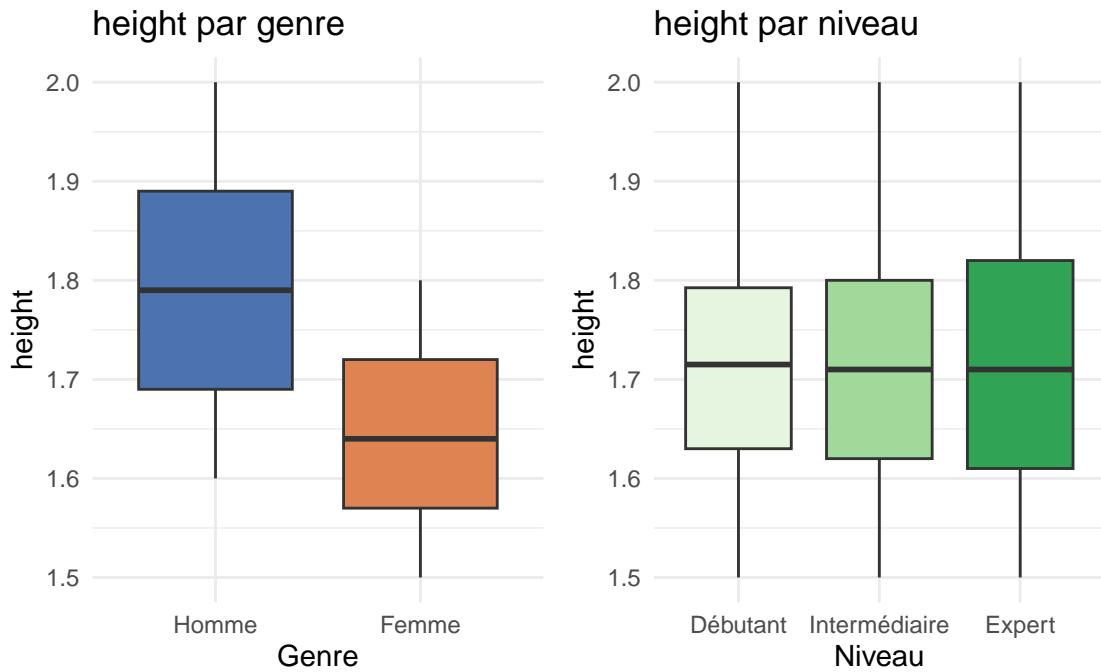
Boxplots de weight selon genre et niveau



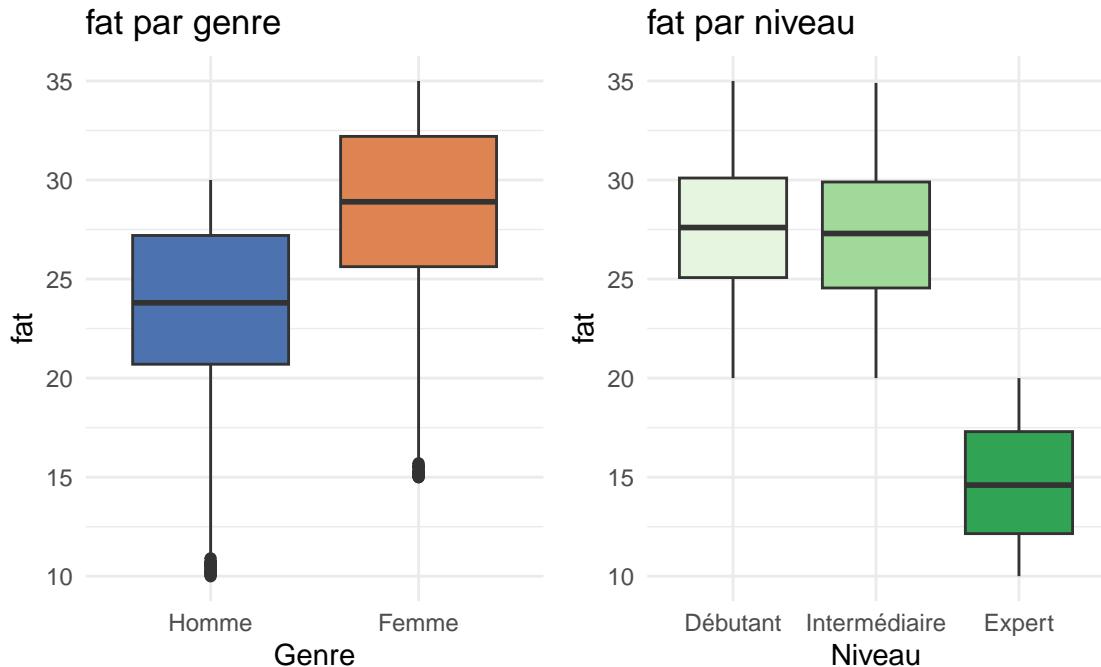
Boxplots de bmi selon genre et niveau



Boxplots de height selon genre et niveau



Boxplots de fat selon genre et niveau



Entre les hommes et les femmes, on observe une différence de poids notable. Celle-ci va de paire avec la taille. Bien que nous n'avions pas relevé de corrélation entre ces deux variables en séparant par sexe, le coefficient était tout de même de 0.37 sur la matrice mixte. Ces trois variables sont donc liées mais il n'y a pas d'observation particulière à faire à ce niveau là. La durée des séances

et le nombre de calories dépensées ne semblent pas liées au sexe. Nous avions vu que ces deux variables étaient fortement corrélées entre elles, d'où les observations similaires lorsqu'on les observe indépendamment. On relève toutefois la présence d'outliers chez les deux genres pour la variable calories et non pour la durée. Ainsi, certains individus consomment beaucoup plus de calories. Selon les corrélations établies précédemment, on s'attend à ce que ces mêmes individus aient des séances plus longues et un masse grasse plus bas. Et en effet, sur les boxplot du masse grasse, nous retrouvons ces outliers. Une autre observation plus notable est que les femmes ont un masse grasse nettement plus élevé. Il peut y avoir plusieurs explications à cela, et le jeu de données n'inclue pas certaines variables potentiellement clé à ce sujet telles que la maternité ou l'âge. Toutefois, une recherche sur google scholar tend à confirmer cette observation. Le graphique suivant, allant de paire avec la corrélation inverse entre la consommation d'eau et le masse grasse établie précédemment, montre que les femmes semblent consommer moins d'eau. On avait toutefois déjà observé que les personnes avec un masse grasse plus élevé consommaient moins d'eau. De plus, cette corrélation restait vraie quand on séparait les matrices par sexe, c'est donc le lien eau-fat plutôt que eau-genre que l'on gardera. Enfin, l'IMC par genre. On remarque la présence de certaines personnes en situation d'obésité sévère et morbide seulement chez les hommes, tandis qu'il n'y a pas de femmes avec un IMC supérieur à 35. La tendance au sur-poids est globalement plus prononcée chez les hommes que chez les femmes. Cette tendance est inverse à celle observée sur le masse grasse.

Sur le premier graphique, le poids ne semble pas lié au niveau, car les 3 niveaux ont leur intervalle inter-quartile entre 57 et 87. Toutefois, il est notable que les experts ont un poids clairement borné entre 55 et 90, ce qui n'est pas le cas des débutants et des intermédiaires. Ainsi, une tendance se dessine tout de même quand au poids qu'aurait un expert. La taille également semble non liée au niveau. On remarque toutefois que l'intervalle semble s'élargir quand le niveau augmente. Mais on avait observé sur le premier violin plot de la taille un double pic dans la distribution. Cela pourrait donc être un biais résiduel dû au genre. On remarque une tendance claire et prononcée au niveau de la durée des séances. Les experts ont tous des séances longues, tandis que les intermédiaires font des séances moins longues et les débutants encore moins. Étant donné la corrélation forte entre la durée des séances et les calories dépensées, on retrouve cette même tendance sur le graphique suivant, le niveau allant croissant avec les calories dépensées. Cette tendance semble moins prononcée pour la consommation d'eau. Elle l'est malgré tout, on la visualise simplement un peu moins parce que tout le monde consomme globalement des quantités d'eau comparables. Toutefois, on remarque encore que les experts (on le sait déjà, faisant des séances plus longues et plus énergivores) consomment plus d'eau que les débutants et intermédiaires. Cette tendance est toutefois liée aux deux précédentes, il n'y a pas de lien de causalité notable entre la consommation d'eau et l'expertise dans un sport. Le masse grasse quand à lui fortement corrélé à l'expertise. Il semble même se dessiner une condition nécessaire et suffisante car pas un seul adhérant avec un masse grasse entre 12 et 18 n'est débutant ou intermédiaire, et tous les gens ayant un masse grasse dans cet intervalle sont classés experts. À l'inverse, les gens ayant une masse grasse plus élevée (de 20 à 35%) seront forcément débutant ou intermédiaire. Cette observation est tout à fait cohérente avec ce qui se trouve dans la littérature (Sedukin D. V. et al., 2025), (Duncan M. J. et al., 2016). On observe également la non présence d'outliers, contrairement au graphique sur l'IMC. En effet, sur ce dernier on retrouve les outliers en situation d'obésité grave que nous avions identifié précédemment. Tandis que les experts ne sont que rarement en situation d'obésité. En effet, un IMC de 35 est déjà considéré comme extrême pour un expert. La boîte est également beaucoup plus fine, plaçant l'IMC d'un expert entre la fine fourchette de 23 à 27 avec une tolérance beaucoup plus réduite que pour les débutants et intermédiaires. Pour ces derniers, l'écart inter quartile est entre 20 et 30 avec une moustache s'étendant de 10 à 43, ce qui inclut les situations de maigreur et plusieurs nuances d'obésité et de sur-poids.

En conclusion, l'expertise se distingue par une masse grasse réduite (12-18 %). Bien que le genre influence la répartition des masses grasses et les extrêmes de l'IMC, la performance reste dictée par la durée des séances et l'intensité de la dépense calorique. Ces résultats valident la littérature scientifique en confirmant que l'assiduité à l'entraînement est le principal levier pour optimiser sa santé physique.

faire η^2 si le temps.

3.3 Association quantitatives vs qualitatives (η^2)

Pour chaque variable quantitative, on calcule le rapport de corrélation η^2 vis-à-vis de chaque variable qualitative (gender et level) à l'aide de la fonction eta2() du package BioStatR.

```
eta_gender <- sapply(variables_quantitatives, function(v) eta2(SalleDeSport[[v]], SalleDeSport$gender))
eta_level <- sapply(variables_quantitatives, function(v) eta2(SalleDeSport[[v]], SalleDeSport$level))
eta_tab <- data.frame(Variable = variables_quantitatives, Eta2_Gender = eta_gender, Eta2_Level = eta_level)
print(eta_tab)
```

Variable	Eta2_Gender	Eta2_Level
calories	0.0226943595	0.5093431711
water	0.4457674667	0.1685026462
duration	0.0001488394	0.6217788488
weight	0.3356351005	0.0004401156
bmi	0.0973253285	0.0019553712
height	0.3404763329	0.0008769585
fat	0.1659009315	0.6480331227

Le tableau obtenu donne, pour chaque variable quantitative, la part de variance expliquée par le genre ou par le niveau. Par exemple :

Une grande valeur de η^2 pour fat vs gender indiquerait que le pourcentage de masse grasse diffère significativement entre hommes et femmes (on s'y attend).

Une valeur élevée de η^2 pour duration vs level suggérerait que la durée moyenne de séance varie selon l'expérience (éventuellement les débutants s'entraînent moins longtemps que les experts, ou vice versa).

Des valeurs proches de 0 (pour water vs gender, par exemple) signifieraient une absence de lien fort.

[À compléter : interpréter les valeurs spécifiques d' η^2 et en tirer des conclusions sur quelles relations sont fortes]

Variables qualitatives entre elles

Enfin, on examine l'association entre les deux variables qualitatives gender et level.

Rédaction Table de contingence

```
tab_genre_niveau <- table(SalleDeSport$gender, SalleDeSport$level)
tab_genre_niveau
addmargins(tab_genre_niveau) # ajoute les totaux lignes/colonnes
```

Profils-lignes et profils-colonnes

```
prop.table(tab_genre_niveau, 1) # proportions par genre (lignes) prop.table(tab_genre_niveau,  
2) # proportions par niveau (colonnes)
```

Mosaic plot

```
mosaicplot(tab_genre_niveau, color=TRUE, main="Mosaic : Genre vs Niveau")
```

On obtient ainsi la table de contingence et les marges. Les profils-lignes (`prop.table(...,1)`) montrent la répartition proportionnelle des niveaux pour chaque genre : par exemple, chez les hommes, X% sont débutants, Y% intermédiaires, Z% experts. Les profils-colonnes (`prop.table(...,2)`) indiquent la proportion d'hommes et de femmes parmi les débutants, etc. Le mosaic plot met en évidence visuellement ces proportions.

[À compléter : commenter les résultats. Par exemple, si la proportion d'hommes est plus grande parmi les experts que parmi les débutants, cela se traduirait par un décalage visible. Sinon, on constate une répartition uniforme, etc.]

Conclusion partielle

En résumé, cette analyse descriptive montre les caractéristiques principales du jeu de données DataGym3MIC. La plupart des variables quantitatives ont une distribution unimodale sans valeurs aberrantes majeures. On observe des différences attendues entre les genres (par exemple poids/taille supérieurs chez les hommes, pourcentage de graisse supérieur chez les femmes). Le niveau d'expérience semble également associé à certaines variables d'entraînement (à détailler). Les représentations graphiques (histogrammes, boxplots, nuages de points, mosaic plots) sont adaptées pour visualiser ces distributions et relations.

La suite du rapport pourra compléter l'interprétation détaillée de chaque graphique produit, ainsi que l'exploration de relations plus fines entre les variables (test d'indépendance entre qualitatives, analyse de covariance, etc.). Chaque choix de représentation a été justifié par le type de variable et l'intérêt d'illustrer visuellement la distribution ou la relation étudiée.

Ces graphiques (à exécuter) montreront la répartition : par exemple, si les hommes représentent 55% et les femmes 45% de l'échantillon, ou si 50% sont débutants, 30% intermédiaires, 20% experts. [À compléter : indiquer les pourcentages exacts et observations principales sur ces distributions]

4 Analyse en composantes principales

Le principe de l'ACP repose sur la création de composantes principales par combinaison linéaire des variables originales. En éliminant la redondance d'information (corrélations), cette méthode permet de visualiser les individus sur des plans factoriels et d'identifier les axes qui expliquent la plus grande part de l'inertie du nuage de points.

Le jeu de données oppose ici la **morphologie** (poids, taille, IMC, graisse) à la **pratique sportive** (durée, calories, eau, niveau).

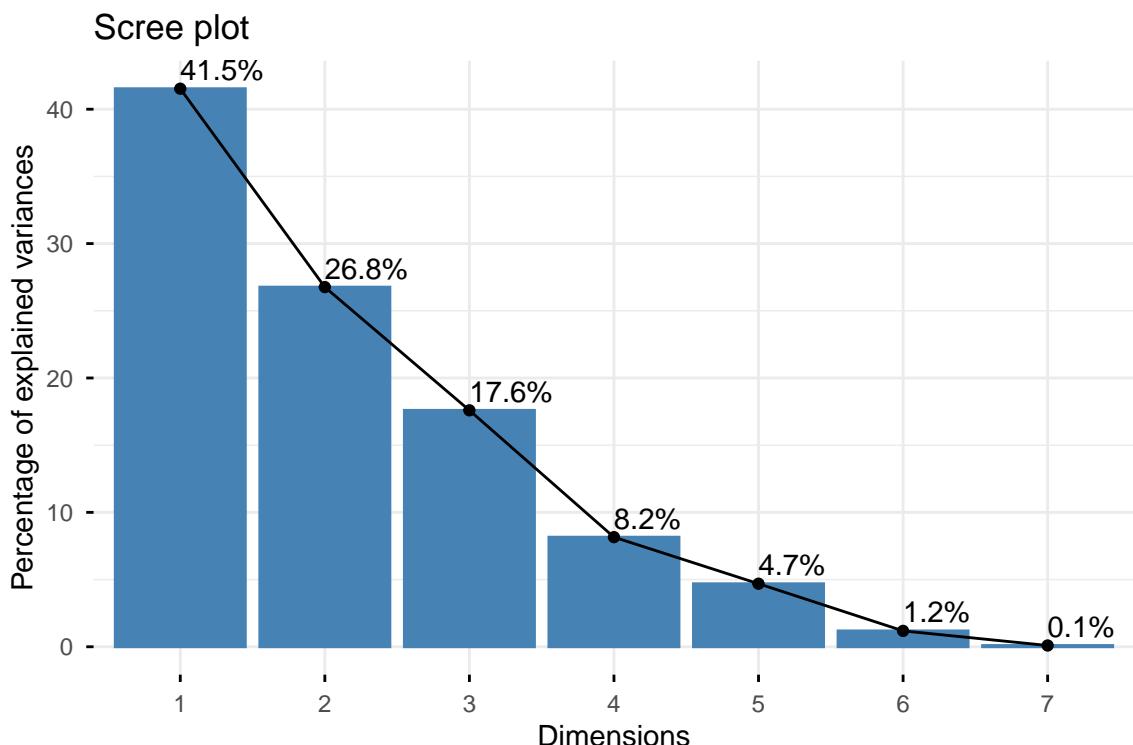
L'analyse préalable montre de fortes corrélations, notamment entre la durée et les calories, ou le poids et l'IMC. En raison de l'hétérogénéité des unités (kg, h, kcal), nous appliquons une **ACP centrée réduite**. Cette normalisation assure une contribution équitable de chaque variable à l'inertie, empêchant les variables à forte variance (comme les calories) de dominer l'analyse.

```

# ACP centrée-réduite
res_pca <- PCA(
  SalleDeSport,
  scale.unit = TRUE,
  quali.sup = c(which(names(SalleDeSport)=="gender"),
                which(names(SalleDeSport)=="level")),
  graph = FALSE
)

# Valeurs propres (choix du nombre d'axes)
#res_pca$eig
fviz_eig(res_pca, addlabels = TRUE)

```



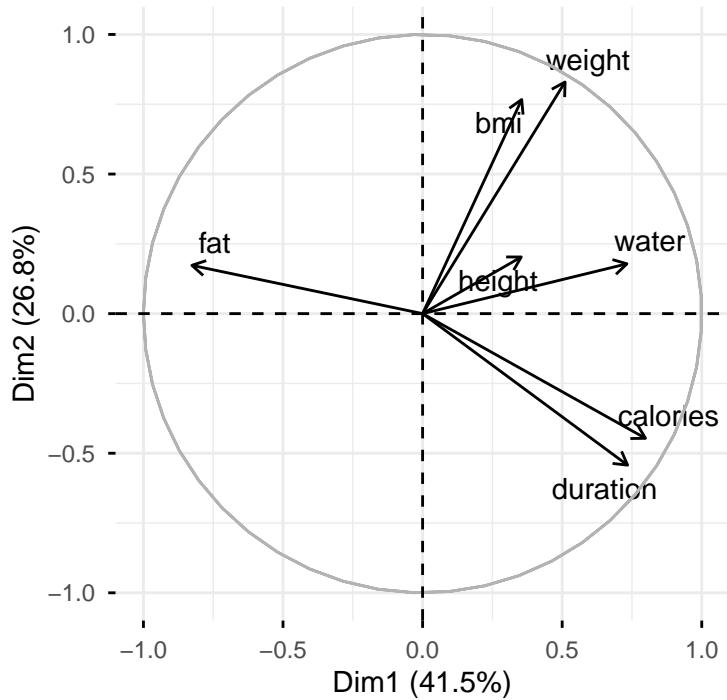
Les deux premiers axes factoriels expliquent 68,3% de l'inertie totale (41,5% pour l'axe 1 et 26,8% pour l'axe 2), assurant une représentation fidèle des données. Bien que l'essentiel de l'analyse repose sur ce premier plan factoriel, la contribution du troisième axe (17,6%) reste significative et ne sera pas totalement occultée lors de l'interprétation.

```

# cercle des corrélations
fviz_pca_var(res_pca,
             repel = TRUE)

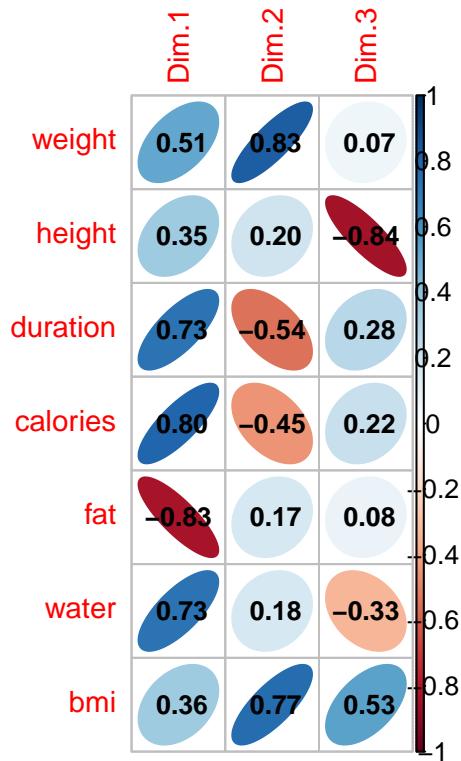
```

Variables – PCA



```
# matrice des corrélations
cor_3axes <- res_pca$var$cor[, 1:3, drop = FALSE]

corrplot(
  cor_3axes,
  method = "ellipse",
  addCoef.col = "black",
  number.cex = 0.8,
  tl.cex = 0.9
)
```



4.1 Analyse du cercle des corrélations

Les variables `weight`, `duration`, `calories`, `water` et `bmi` présentent des vecteurs longs, indiquant une excellente représentation sur le premier plan factoriel. À l'inverse, la variable `height` est plus proche du centre, signalant une moins bonne qualité de représentation sur ces deux axes.

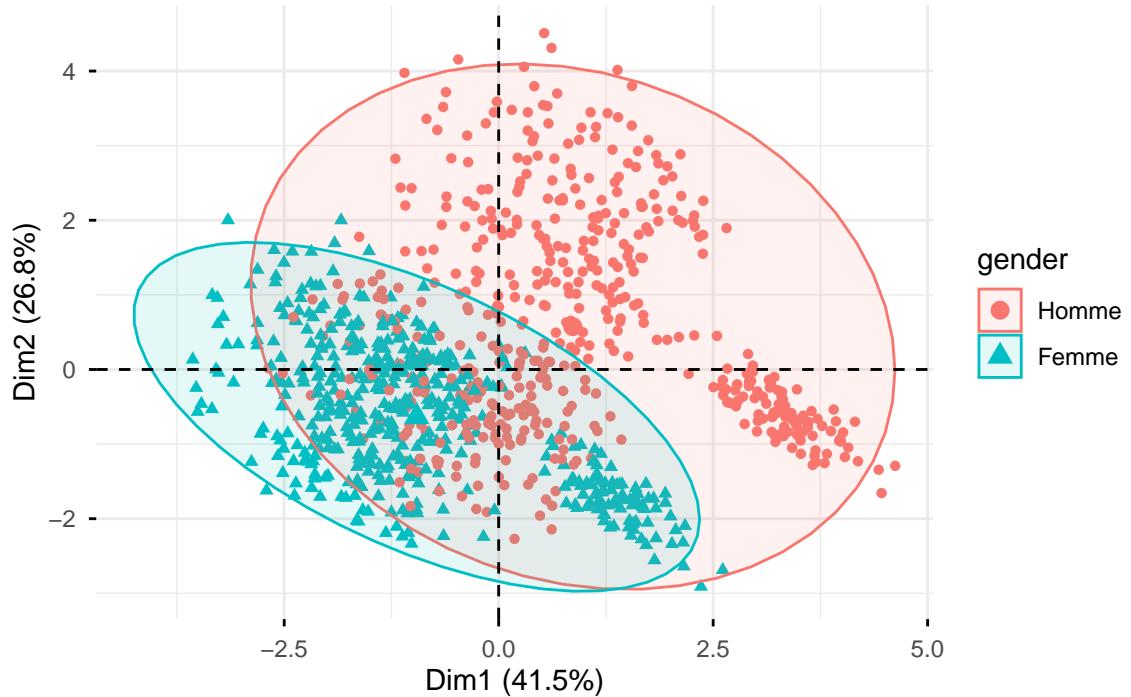
- **Axe 1 (41,5%)** : Intensité de l'entraînement. Cet axe oppose les variables de performance (`calories`, `duration`, `water`) à la masse grasse (`fat`). Il confirme que les profils ayant un taux de graisse élevé effectuent des séances moins intenses.
- **Axe 2 (26,8%)** : Morphologie. Il est fortement corrélé au poids et à l'IMC, tout en s'opposant légèrement à l'endurance. Cet axe permet de distinguer les individus selon leur corpulence.
- **Axe 3 (17,6%)** : Dimension verticale. Principalement lié à la taille (`height`), cet axe apporte une information complémentaire sur la stature, la taille étant peu corrélée aux deux premiers axes (0,35 et 0,20).

En conclusion, le premier axe oppose des profils à forte intensité d'entraînement à des profils à forte masse grasse. Le second axe traduit principalement des différences morphologiques liées au poids et à l'IMC. La troisième composante est surtout associée à la taille et apporte une information complémentaire.

```
# Graphique des individus
fviz_pca_ind(res_pca,
              geom = "point",
```

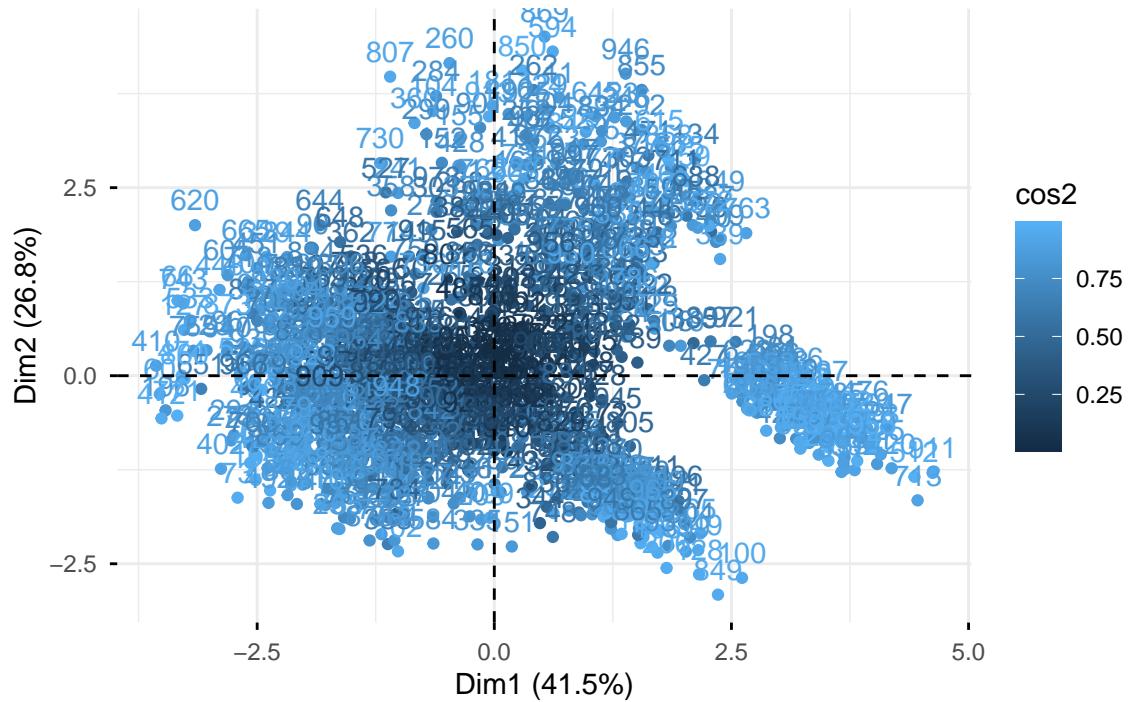
```
habillage = "gender",  
addEllipses = TRUE)
```

Individuals – PCA



```
# Autre lecture intéressante :  
fviz_pca_ind(res_pca, col.ind = "cos2")
```

Individuals – PCA



5 Classification non supervisée (Clustering)

Afin de regrouper les individus en classes homogènes, nous allons utiliser les résultats de notre ACP. Plutôt que de travailler sur les variables brutes, nous réalisons la classification sur les coordonnées des individus sur les **3 premiers axes factoriels**.

Cette approche permet de baser les groupes sur les dimensions structurantes identifiées précédemment (Intensité, Corpulence, Taille) en éliminant le bruit résiduel des derniers axes.

```
donnees_acp <- res_pca$ind$coord[, 1:3]

# Aperçu des données utilisées pour le clustering
head(donnees_acp)
```

	Dim.1	Dim.2	Dim.3
1	3.178026	-0.20613732	0.5326227
2	-1.128313	0.44504681	2.0336135
3	-1.652054	0.28775885	0.3246140
4	-2.683734	0.01647382	-0.8505631
5	-2.189516	-0.34673298	-1.9837715
6	1.138159	-1.68339118	0.1416179

5.1 Classification Ascendante Hiérarchique (CAH)

Nous effectuons une CAH sur ces coordonnées factorielles en utilisant la distance euclidienne et le critère de Ward (minimisation de l'inertie intra-classe).

Dendrogramme (CAH sur ACP)

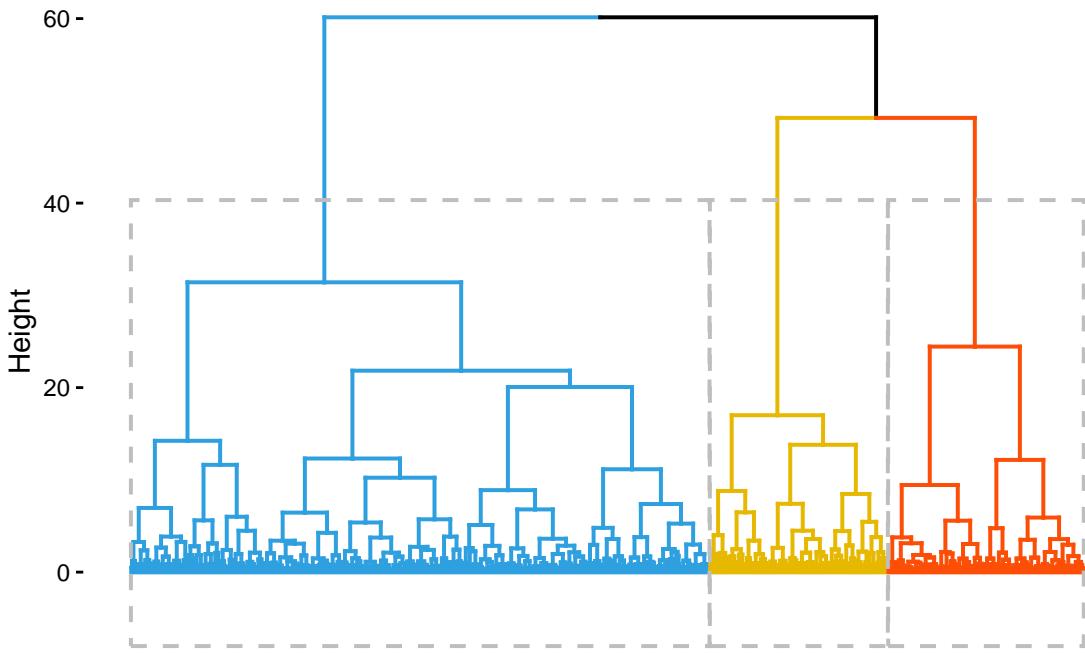


Figure 1: Dendrogramme de la classification hiérarchique sur composantes principales

5.1.1 Choix du nombre de classes

Le dendrogramme montre un saut d'inertie majeur lors du passage de 3 à 2 classes. Nous retenons donc une partition en **k=3 classes**. Ce choix est cohérent avec la structure de la variable `level` qui comporte 3 modalités (Débutant, Intermédiaire, Expert). Nous chercherons par la suite à vérifier si ces classes statistiques recouvrent effectivement ces niveaux d'expertise ou si elles révèlent une autre segmentation (ex: distinction Homme/Femme ou Morphologie).

5.2 Méthode des K-means

Nous appliquons maintenant l'algorithme des K-means, toujours sur les coordonnées des 3 premiers axes factoriels.

Projection des clusters K-means sur le plan ACP (1–2)

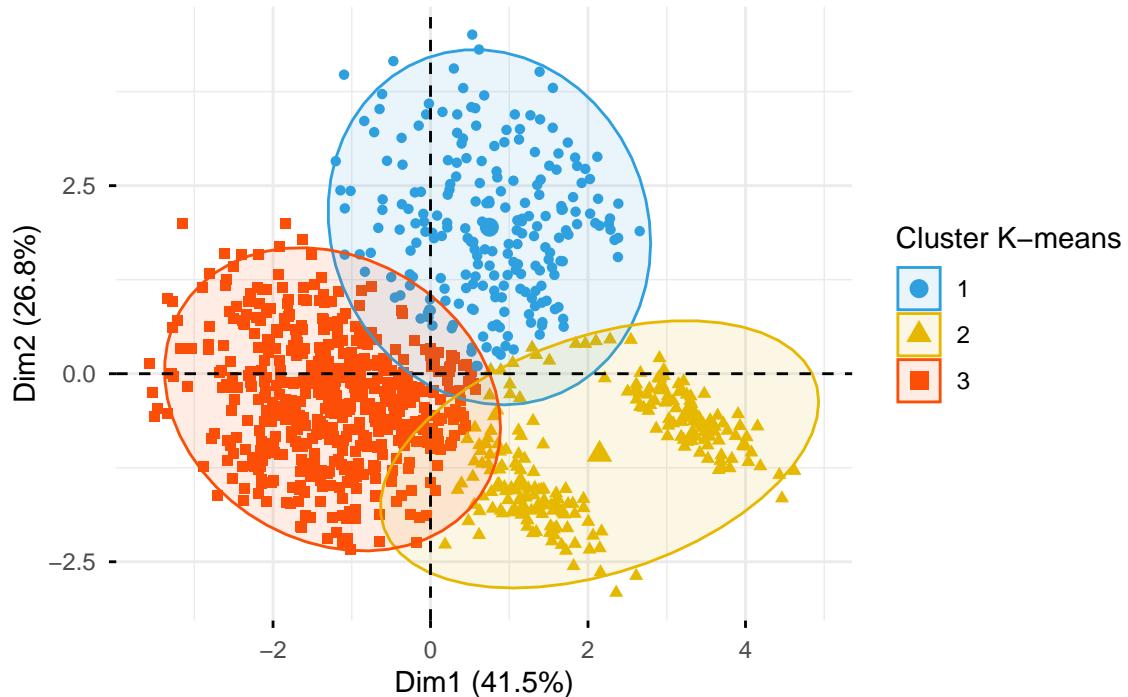


Figure 2: Partition K-means projetée sur le premier plan factoriel

L'avantage de cette représentation est immédiat : les classes sont, par construction, bien séparées sur les axes factoriels.

5.3 Comparaison des classifications

Comparons la partition obtenue par la CAH et celle des K-means.

Analyse de la stabilité des classes (Matrice de confusion) :

```
classes_cah <- cutree(res_cah, k = 3)

SalleDeSport$cluster_cah <- NA
SalleDeSport$cluster_kmeans <- NA

ids_a_classer <- rownames(donnees_acp)

SalleDeSport[ids_a_classer, "cluster_cah"] <- factor(classes_cah)
SalleDeSport[ids_a_classer, "cluster_kmeans"] <- factor(res_kmeans$cluster)

table_comp <- table(
  CAH = SalleDeSport$cluster_cah,
  Kmeans = SalleDeSport$cluster_kmeans
)
```

```
addmargins(table_comp)
```

Kmeans					
CAH	1	2	3	Sum	
1	0	200	0	200	
2	39	26	526	591	
3	179	3	0	182	
Sum	218	229	526	973	

En comparant la CAH et les K-means via la table de contingence, nous observons une stabilité remarquable pour deux groupes, mais une hésitation sur le troisième :

- **Le noyau dur (Diagonale forte) :**

- La classe 1 de la CAH correspond parfaitement à la classe 2 des K-means (200 individus, 0 erreur). C'est un groupe très homogène et distinct.
- La classe 3 de la CAH est quasi-identique à la classe 1 des K-means (179 individus sur 182).

- **La zone de flou (Classe centrale) :**

- La classe 2 de la CAH est la plus volumineuse (591 individus) et semble être une classe “fourre-tout”. L'algorithme des K-means, qui réalloue les individus pour optimiser les centres, a “nettoyé” cette classe : il a conservé 526 individus dans son Cluster 3, mais en a réassigné 39 vers le Cluster 1 et 26 vers le Cluster 2.

En conclusion, la méthode des K-means a permis d'affiner la partition en “récupérant” des individus frontières mal classés par la méthode hiérarchique (qui ne peut pas revenir en arrière une fois un regroupement effectué). Nous privilégierons donc la partition K-means pour l'interprétation finale.

5.4 Croisement et Profilage

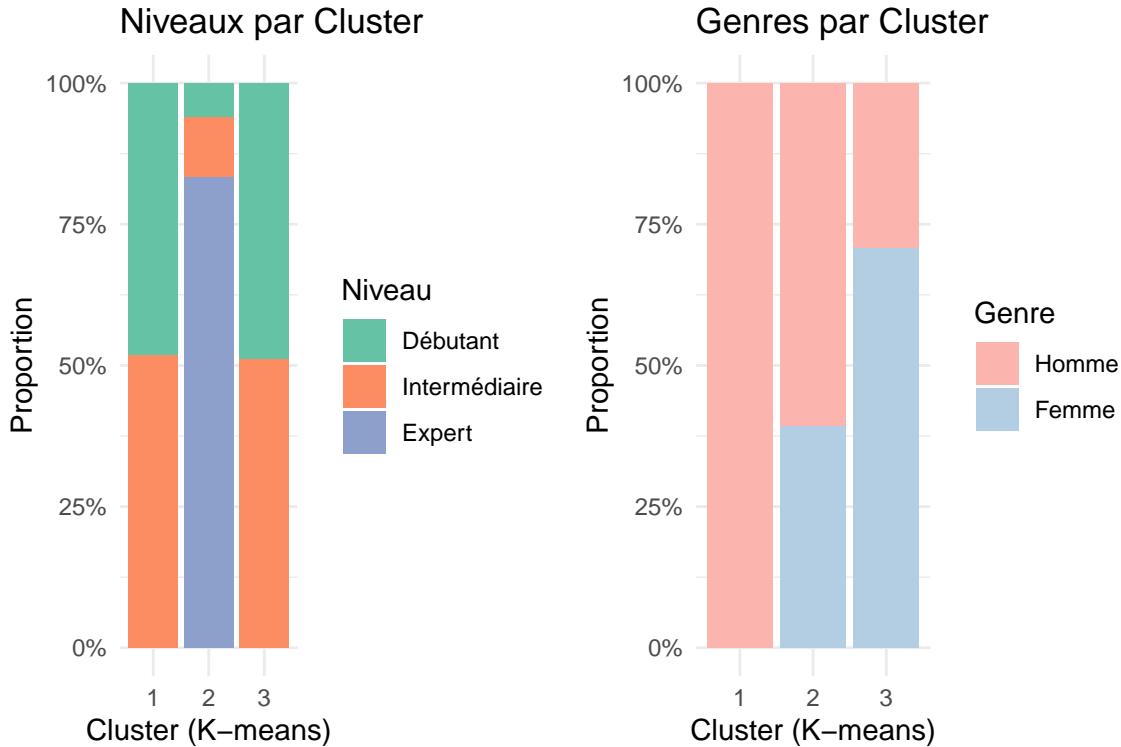


Figure 3: Analyse croisée des clusters avec les variables qualitatives

En projetant la partition K-means sur le premier plan factoriel et en la croisant avec les variables qualitatives, nous pouvons dresser les portraits-robots des 3 clusters :

- **Cluster 1 (correspondant probablement aux Experts)** : Ce groupe rassemble les individus ayant des séances longues, une forte dépense calorique et un faible taux de masse grasse.
- **Cluster 3 (correspondant probablement à une morphologie spécifique/Surpoids)** : Ce groupe se caractérise par un IMC et un poids élevés, potentiellement corrélés à une pratique moins intense.
- **Cluster 2 (Les “Intermédiaires”)** : Le groupe central et volumineux. Il rassemble la majorité des usagers, avec des performances et une morphologie moyennes. C'est ce manque de caractéristiques extrêmes qui rend ce groupe plus difficile à délimiter (comme vu dans la comparaison CAH/Kmeans).

```
# Une fois votre clustering terminé (res_kmeans)
SalleDeSport$cluster <- as.factor(res_kmeans$cluster)

# 1. Description automatique des classes (variables quantitatives)
# Calcule la moyenne des variables pour chaque cluster
aggregate(cbind(weight, height, duration, calories, fat, bmi) ~ cluster,
```

```
  data = SalleDeSport,  
  FUN = mean)
```

```
cluster      weight     height duration calories      fat      bmi  
1          1 104.51147 1.787752 1.119771  841.1651 25.10734 33.02060  
2          2  72.57991 1.742926 1.702314 1240.6943 16.33843 24.02301  
3          3  61.70399 1.686711 1.118935  786.0894 28.68346 21.93867
```

```
# 2. Croisement avec les variables qualitatives (Tableaux croisés)  
print("Répartition des Niveaux par Cluster :")
```

```
[1] "Répartition des Niveaux par Cluster :"
```

```
table(SalleDeSport$cluster, SalleDeSport$level)
```

	Débutant	Intermédiaire	Expert
1	105	113	0
2	14	24	191
3	257	269	0

```
print("Répartition des Genres par Cluster :")
```

```
[1] "Répartition des Genres par Cluster :"
```

```
table(SalleDeSport$cluster, SalleDeSport$gender)
```

	Homme	Femme
1	218	0
2	139	90
3	154	372

6 Bibliographie

- Nom .P et al., ANNÉE, [en ligne, consulté le XX YYYYYYYY] : lien
- Sedukin D. V. et al., 2025, [en ligne, consulté le 29 Décembre] : <https://vpbim.com.ua/wp-content/uploads/2025/10/54.pdf>
- Duncan M. J. et al., 2016, [en ligne, consulté le 29 Décembre] : <https://www.tandfonline.com/doi/abs/10.1080/02640414.2016.1258483>