# Projet Data Clustering

*Costa Laura – M2 ATDM STAPS*

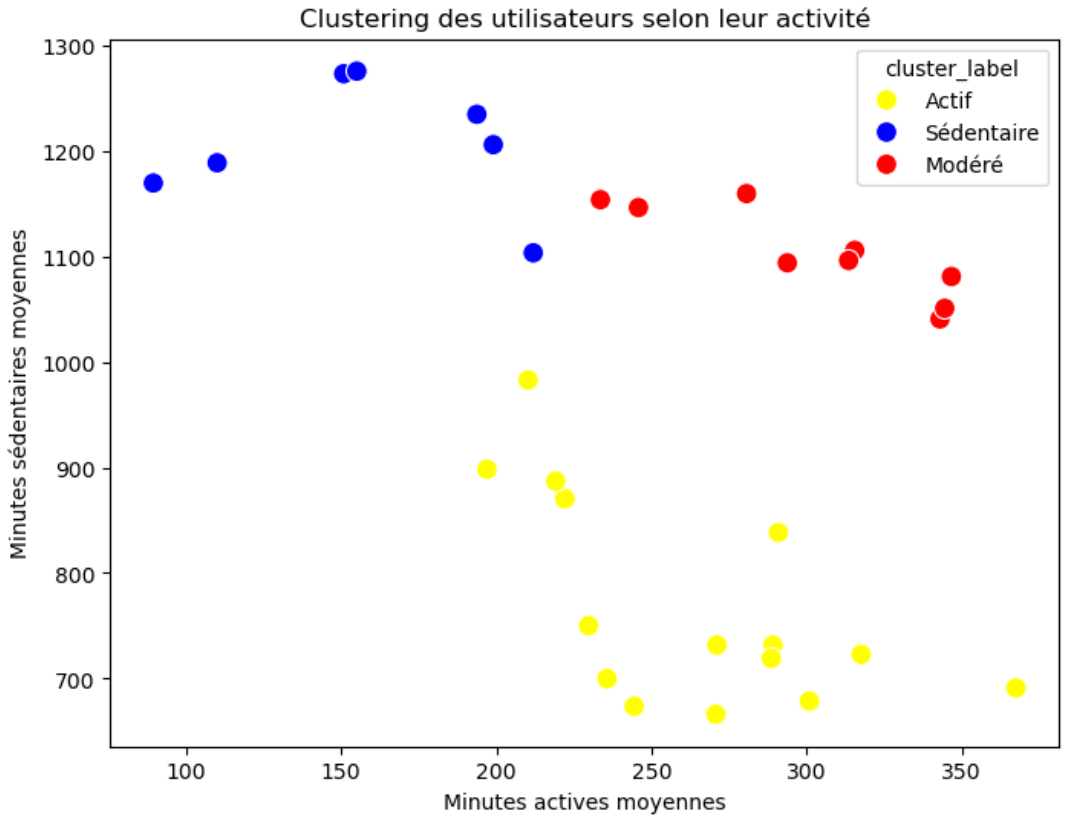
# Dataset : fitbit

## Clustering – Activité physique

- Objectifs : Regrouper les utilisateurs selon leur profil d'activité physique & identifier des profils types (actifs, modérément actifs, sédentaires)
- **Dataset : daily_activity**
  - 1 ligne = 1 journée d'un utilisateur
  - Suppression des doublons et des lignes avec valeurs manquantes
  - Transformation en format date pour le champ ActivityDate
  - Filtre sur SedentaryMinutes < 1440 minutes (24h), Calories > 0, TotalSteps > 100 pas
  - Création d'une nouvelle colonne TotalMinutes = VeryActiveMinutes + ModeratlyActiveMinutes + LightlyActiveMinutes + SedentaryMinutes puis filtre sur TotalMinutes < 1440 minutes (24h) et > 480 minutes (8h)
  - Création d'une nouvelle colonne TotalActiveMinutes = VeryActiveMinutes + ModeratlyActiveMinutes+LightlyActiveMinutes
  - Suppression des utilisateurs avec moins de 10 enregistrements
  - **906 lignes, 15 colonnes, 31 utilisateurs**

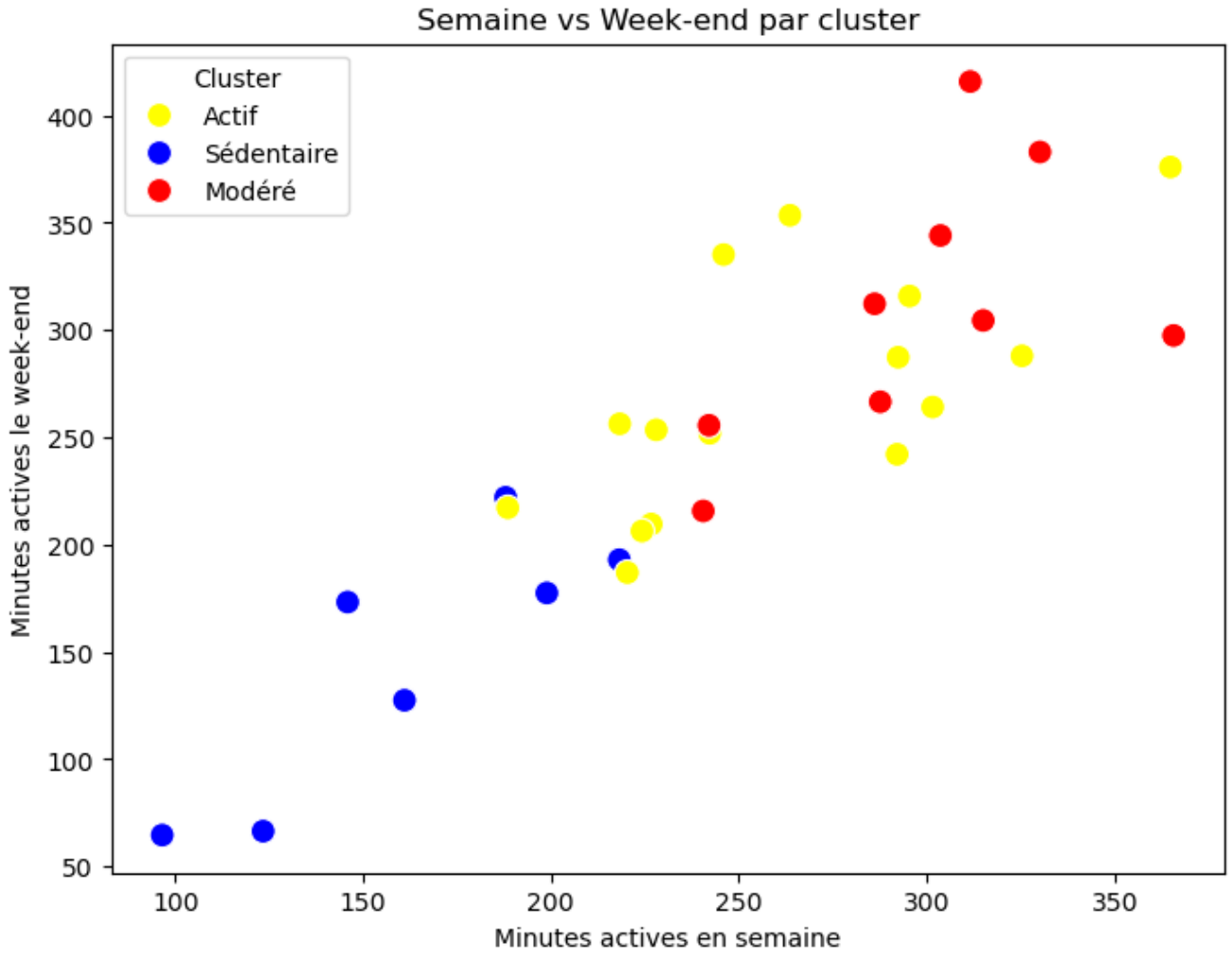
# Clustering avec Kmeans

- Calcul de la moyenne par utilisateur pour chaque variable
- Kmeans avec n=3 et mean_sedentary_minutes/mean_total_active_minutes




- **Sédentaire (peu d'activité, beaucoup de sédentarité)** :
  - Nombre de minutes sédentaires le plus élevé
  - Nombre de minutes actives le plus faible
- **Modéré (beaucoup d'activité, et de sédentarité)** :
  - Nombre de minutes sédentaires le plus élevé
  - Nombre de minutes actives le plus élevé
- **Actif (beaucoup d'activité et peu de sédentarité)** :
  - Nombre de minutes sédentaires le plus faible
  - Nombre de minutes actives le plus élevé

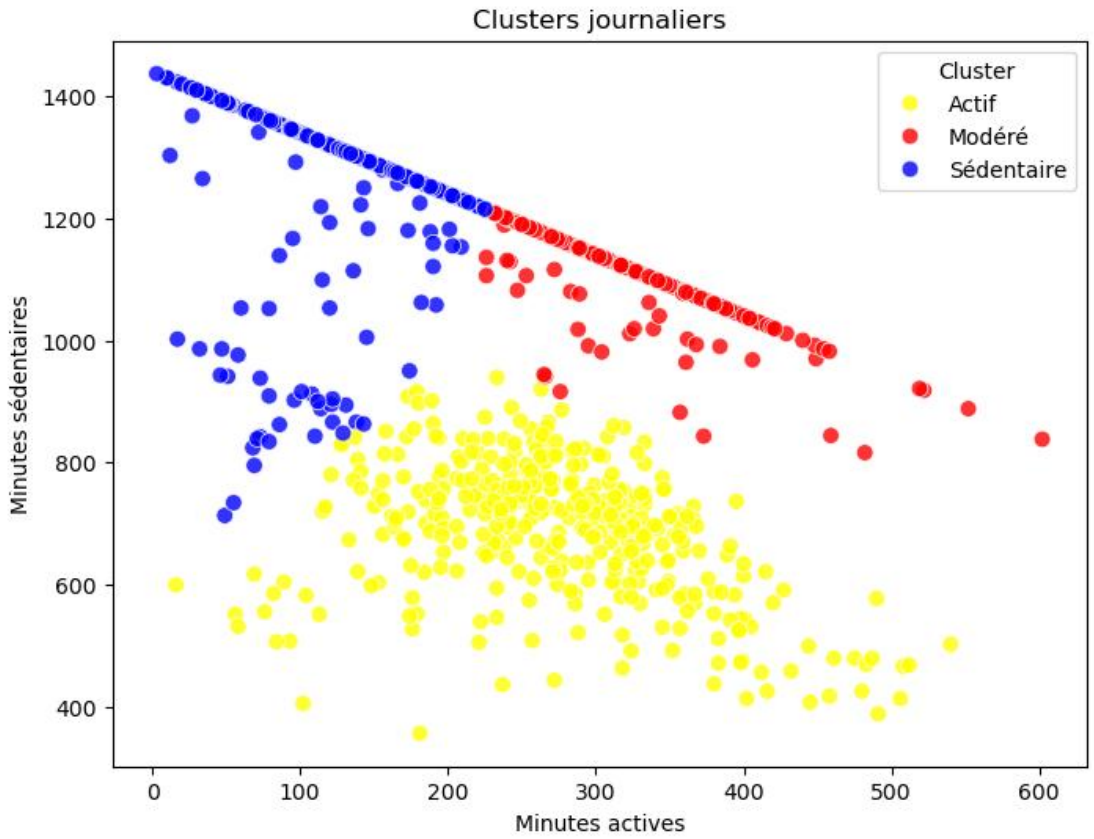
# Selon les clusters, l'activité physique est-elle pratiquée plutôt la semaine ou le week-end ?



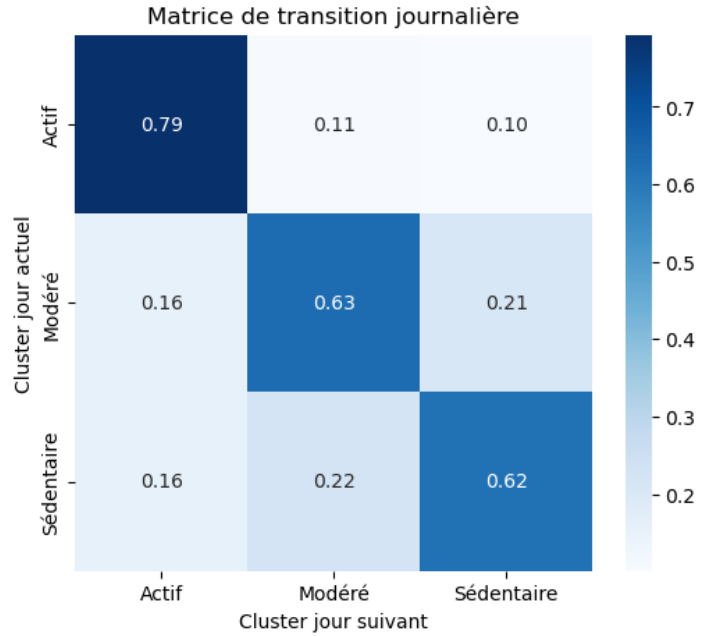

- **Sédentaire** :
  - Peu actifs en semaine ou le week-end
- **Modéré** :
  - Plus actif le week-end que la semaine
- **Actif** :
  - Actif la semaine et le week-end

# Matrice de transition (par jour)

- 1 point = 1 jour d'un utilisateur

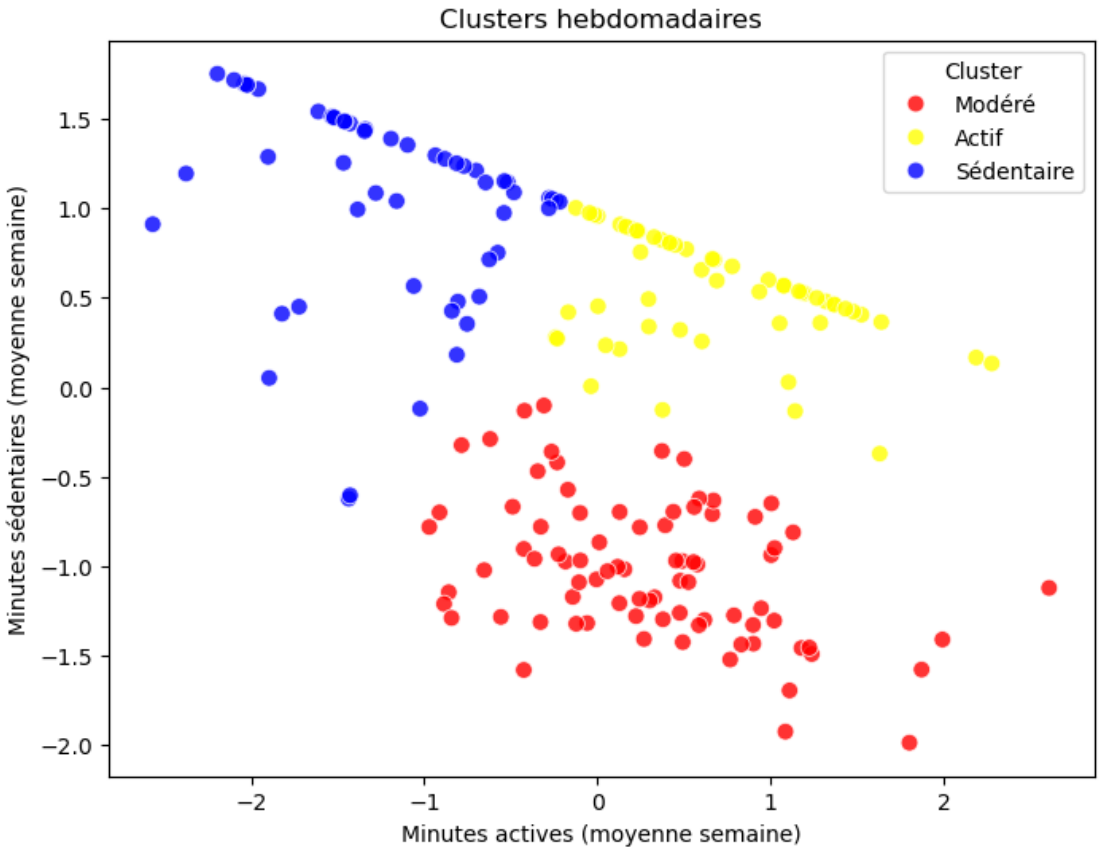



- Si je suis dans le cluster « Actif », j'ai 79% de chances d'y rester le lendemain (cluster stable)
- Si je suis dans le cluster « Modéré », j'ai 63% de chances d'y rester le lendemain (cluster un peu moins stable).
- Si je suis dans le cluster « Sédentaire », j'ai 62% de chances d'y rester le lendemain (cluster un peu moins stable).

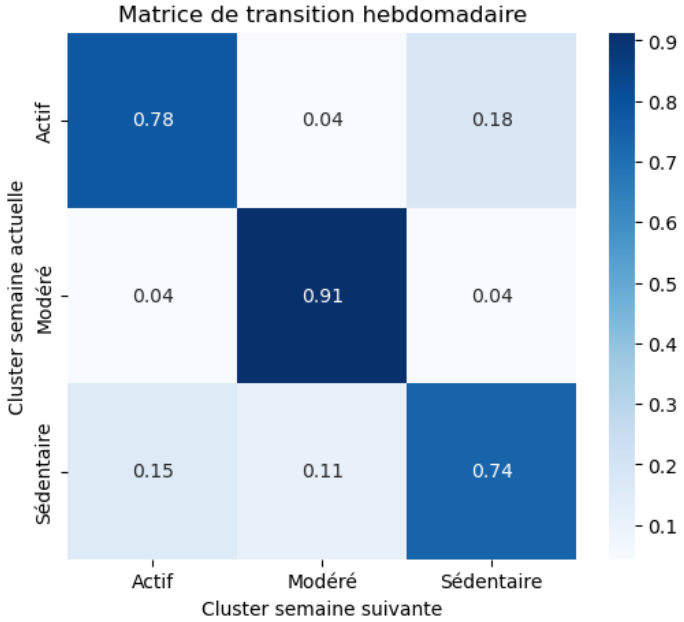


# Matrice de transition (par semaine)

- 1 point = 1 semaine d'un utilisateur

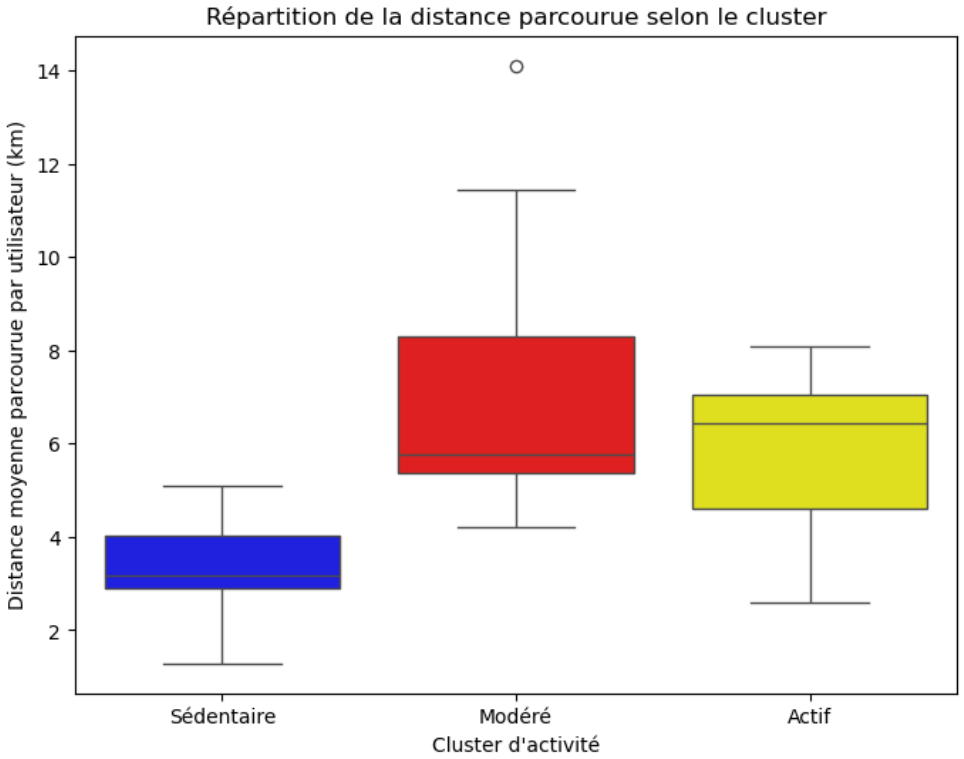


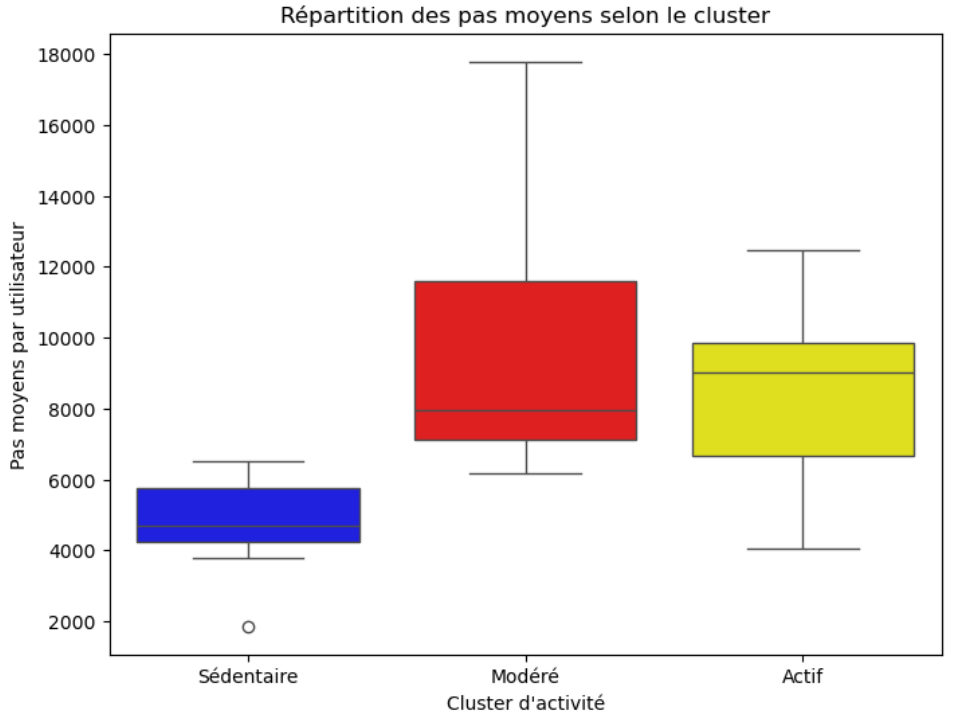

- Si je suis dans le cluster « Actif », j'ai 78% de chances d'y rester la semaine prochaine(cluster stable)
- Si je suis dans le cluster « Modéré », j'ai 91% de chances d'y rester la semaine prochaine (cluster stable).
- Si je suis dans le cluster « Sédentaire », j'ai 74 % de chances d'y rester la semaine prochaine(cluster un peu moins stable).



# Est-ce que le nombre de pas ou la distance par jour en moyenne sont différents selon les clusters ?

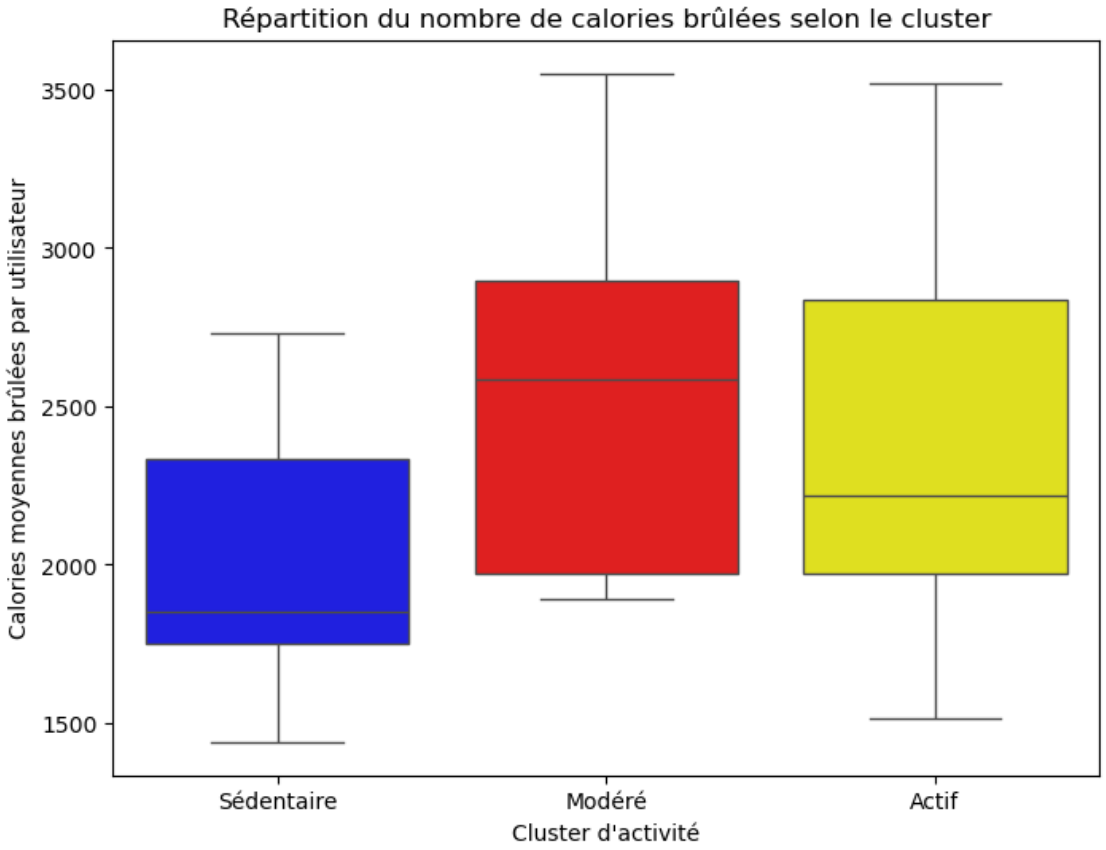
ANOVA 1 facteur : on pourrait penser que les sédentaires font moins de pas et de distance par jour par rapport aux 2 autres groupes



# Est-ce que le nombre de calories brûlées en moyenne est différente selon les clusters ?

ANOVA 1 facteur : pas de différence significative entre les groupes



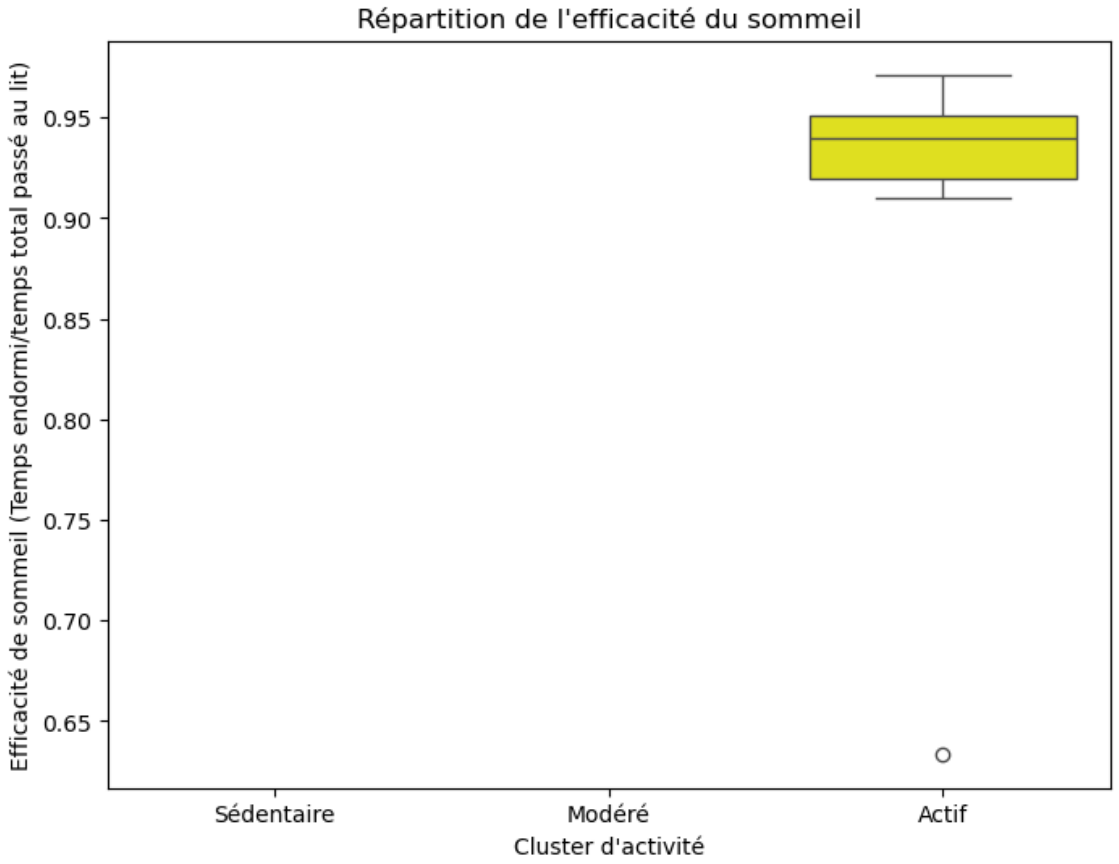
# Dataset : fitbit

**Clustering – Activité physique**

- Comparer avec le sommeil ou l'IMC
- **Dataset : sleep_data**
  - 1 ligne = 1 journée d'un utilisateur
  - Suppression des doublons et des lignes avec valeurs manquantes
  - Transformation en format date pour le champ Date
  - Création d'une nouvelle colonne « Efficacité du Sommeil » = TotalMinutesAsleep/TotalMinutesInBed
  - Suppression des utilisateurs avec moins de 10 enregistrements
  - **376 lignes, 6 colonnes, 15 utilisateurs**
- **Dataset : weight_data**
  - 1 ligne = 1 journée d'un utilisateur
  - Suppression des doublons et des lignes avec valeurs manquantes
  - Transformation en format date pour le champ Date
  - **98 lignes, 4 colonnes, 13 utilisateurs**

# Est-ce qu'il y a des différences en moyenne de « l'efficacité du sommeil » selon les clusters ?

Problème lors de la fusion des datasets : tous les clusters ne sont pas représentés



# Est-ce que l'IMC en moyenne est différent selon les clusters ?

- Pas assez d'individus par cluster pour faire une ANOVA, visuellement on voit cependant une différence entre sédentaire et les 2 autres groupes