Institut Sous-régional De Statistique et d'Économie Appliquée

PROJET DE BIG DATA

ANALYSE DE DONNEES DU SECTEUR SANTE AVEC PYSPARK

2025

Rédigé par :

ADOU Moussa

AYONTA NDJOUTSE Vanelle

BANZOUZI MIAMPASSI Hermann

BETCHE Ephraim Hamadjam

IYANOU PEMBA Julia Emraude

MOUBOKOUNOU Cherubin-Westeroph

OWONA BELINGA Axel Legrand

SADAI JAPHET

YABOYA Anicet Marius

Sous la supervision de :

Serge NDOUMIN

Table des matières

INTRODUCTION

Avec l'évolution rapide des technologies de collecte de données, l'analyse des données du secteur de la santé devient une priorité stratégique pour améliorer la qualité des soins et favoriser le bien-être des populations. Le projet *Analyse de données du secteur santé avec PySpark* s'inscrit dans cette dynamique en mettant à profit l'écosystème Big Data pour explorer, traiter et interpréter de grands volumes de données liés à la santé et au mode de vie. Grâce au framework **PySpark**, nous pouvons tirer parti du traitement en parallèle pour gérer efficacement des ensembles de données volumineux, tels que **FitLife360**, qui simule les données de 3 000 participants sur une période d'un an.

Ce projet vise à **identifier et analyser 10 indicateurs clés de santé et de bien-être**, allant des informations démographiques aux paramètres d'activité physique, en passant par les indicateurs de santé vitale et les habitudes de vie. Ces analyses fourniront des informations précieuses sur les comportements et les tendances en matière de santé, ainsi que sur l'impact du mode de vie sur la condition physique des individus. L'utilisation de **PySpark** permettra de réaliser des transformations et des actions sur ces données, en exploitant la puissance du calcul distribué pour obtenir des résultats en temps réel et détecter des modèles complexes.

L'objectif final est de démontrer la maîtrise des concepts fondamentaux de **PySpark**, tels que les **transformations** et les **actions**, tout en fournissant des analyses pertinentes qui pourraient, dans un contexte réel, soutenir la prise de décision dans le domaine de la santé publique.

Ce document s'articule autour de 4 principales sections. La première présente les données utilisées et les traitements nécessaires pour leur apurement avant analyses. La seconde section se charge de présenter les principaux résultats. La troisième section présente le tableau de bord construit et la dernière section présente les limites et recommandations de l'étude.

PRESENTATION DES DONNEES ET TRAITEMENT

1. Présentation des données

Le projet s'appuie sur **FitLife360**, un ensemble de données synthétiques représentant les données de suivi de la santé et de la condition physique de **3 000 participants** sur une période d'un an. Ces données regroupent des informations variées relatives aux **activités quotidiennes**, aux **mesures vitales** et aux **facteurs liés au mode de vie**, ce qui en fait une source précieuse pour l'analyse de la santé et la modélisation prédictive.

Les caractéristiques des données peuvent être regroupées en plusieurs catégories :

1. **Informations démographiques :**
    - **participant_id** : Identifiant unique pour chaque participant
    - **age** : Âge du participant (18 à 65 ans)
    - **gender** : Sexe (M/F/Autre)
    - **height_cm** : Taille en centimètres
    - **weight_kg** : Poids en kilogrammes
    - **bmi** : Indice de masse corporelle (calculé à partir de la taille et du poids)
2. **Paramètres d'activité physique :**
    - **activity_type** : Type d'exercice (course à pied, natation, vélo, etc.)
    - **duration_minutes** : Durée de la séance d'activité en minutes
    - **intensity** : Intensité de l'exercice (faible, moyenne, élevée)
    - **calories_burned** : Estimation des calories brûlées lors de l'activité

- **daily_steps** : Nombre de pas quotidiens

3. **Indicateurs de santé :**
    - **avg_heart_rate** : Fréquence cardiaque moyenne pendant l'activité
    - **resting_heart_rate** : Fréquence cardiaque au repos
    - **blood_pressure_systolic** : Tension artérielle systolique
    - **blood_pressure_diastolic** : Tension artérielle diastolique
    - **health_condition** : Présence de problèmes de santé
    - **smoking_status** : Statut tabagique (jamais, ancien, actuel)
4. **Paramètres du mode de vie :**
    - **hours_sleep** : Nombre d'heures de sommeil par nuit
    - **stress_level** : Niveau de stress quotidien (sur une échelle de 1 à 10)
    - **hydration_level** : Consommation quotidienne d'eau en litres
    - **fitness_level** : Score de forme physique basé sur l'activité cumulée

2. Traitement des données

Dans cette section, il est question de présenter les différentes vérifications qui ont été faites sur les données avant analyses. Il s'agissait principalement de vérifier l'existence de valeurs manquantes et de valeurs en doubles (doulants).

Après un examen rapide des données mises à notre disposition, il s'est avéré que la base de données ne présentait ni valeurs manquantes, ni doublons. Par conséquent, aucun traitement n'a été fait à ce niveau. Toutefois, pour des besoins d'analyses, la variable âge a été recodée en modalités.

Les participants de l'ensemble de données ont été répartis en cinq groupes d'âge distincts, reflétant une diversité de profils démographiques. Le groupe **18-25 ans** compte **492 participants**, représentant les plus jeunes individus suivis. La tranche d'âge **26-35 ans** regroupe **600 participants**, tandis que les **36-45 ans** sont légèrement plus nombreux avec **646 participants**. Le groupe **46-55 ans** rassemble **641 participants**, constituant une part

importante de la population étudiée. Enfin, le groupe des **56-65 ans** inclut **621 participants**, représentant les individus les plus âgés du panel. Cette classification par âge permet d'analyser les indicateurs de santé et d'activité en fonction des différentes étapes de la vie.

Après que toutes les vérifications et les traitements eurent été faits, la prochaine section se charge dès lors de présenter les principaux résultats des analyses effectuées avec PySpark.

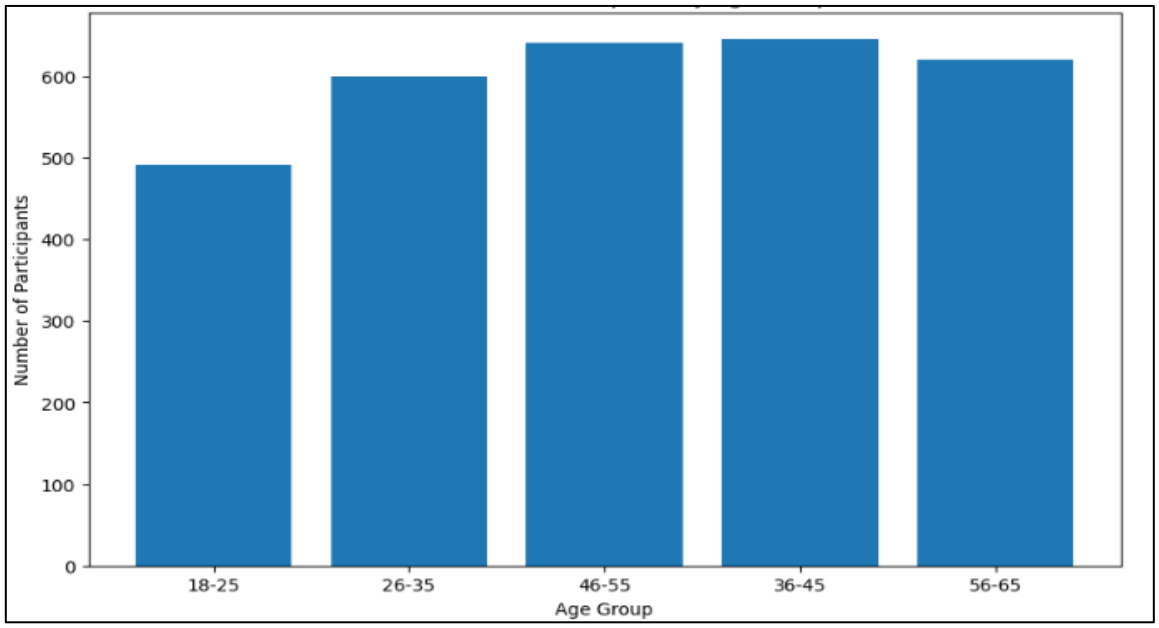
PRESENTATION DES PRINCIPAUX RESULTATS

Dans cette section, il est question de présenter les principaux résultats de nos analyses. En effet, nous avons choisi essentiellement **10 indicateurs** que nous avons jugés pertinents pour notre étude. Nous allons donc tour à tour les présenter et les analyser.

1. Analyse du nombre des participants par tranche d'âge

Le Graphique 1 ci-dessous présente la répartition des participants selon les groupes d'âge. Le groupe d'âge le plus représenté est celui des **36-45 ans** avec **646 participants**, suivi de près par le groupe **46-55 ans** (641 participants) et le groupe **56-65 ans** (621 participants). Le groupe **26-35 ans** compte **600 participants**, tandis que le groupe des **18-25 ans** est le moins représenté, avec **492 participants**. Cette répartition montre une prédominance des participants âgés de 36 ans et plus, représentant une majorité significative de l'échantillon étudié.

Graphique 1 : Répartition des participants par tranche d'âge

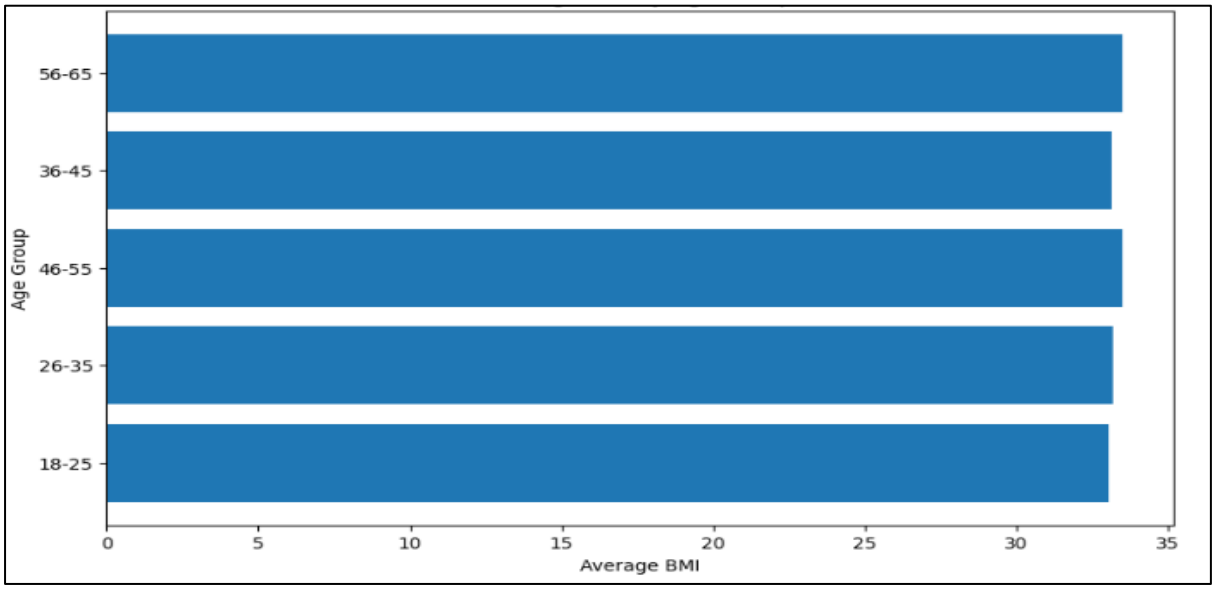


Source : Nos travaux, à partir des données fournies

2. Analyse du BMI moyen pour chaque tranche d'âge

Il ressort de l'analyse de ce graphique 2 que l'Indice de Masse Corporelle (IMC) moyen ne varient pas significativement dans les tranches d'âge. Dans toutes les tranches d'âges, l'IMC moyen dépasse 30.

Graphique 2 : Répartition de l'IMC moyen des participants par tranche d'âge

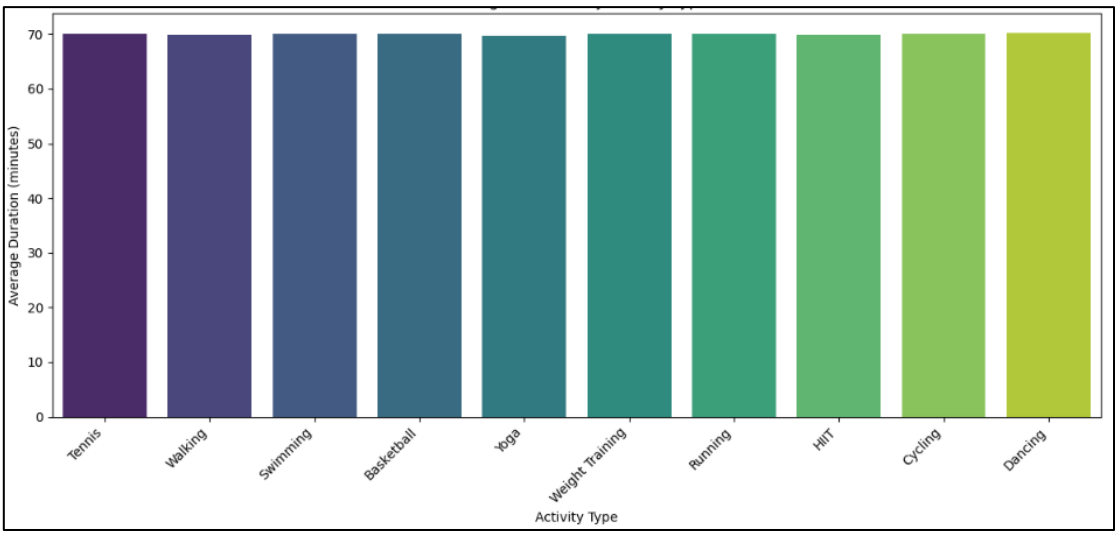


Source : Nos travaux, à partir des données fournies

3. Analyse de la durée moyenne d'activité pour chaque type d'activité

Il ressort de l'analyse du graphique 3 que les participants indépendamment du type d'activité ont pratiquement les mêmes durées moyenne d'activité. Généralement, cette durée se situait autour de 70 minutes.

Graphique 3 : Répartition de la durée moyenne par type d'activité

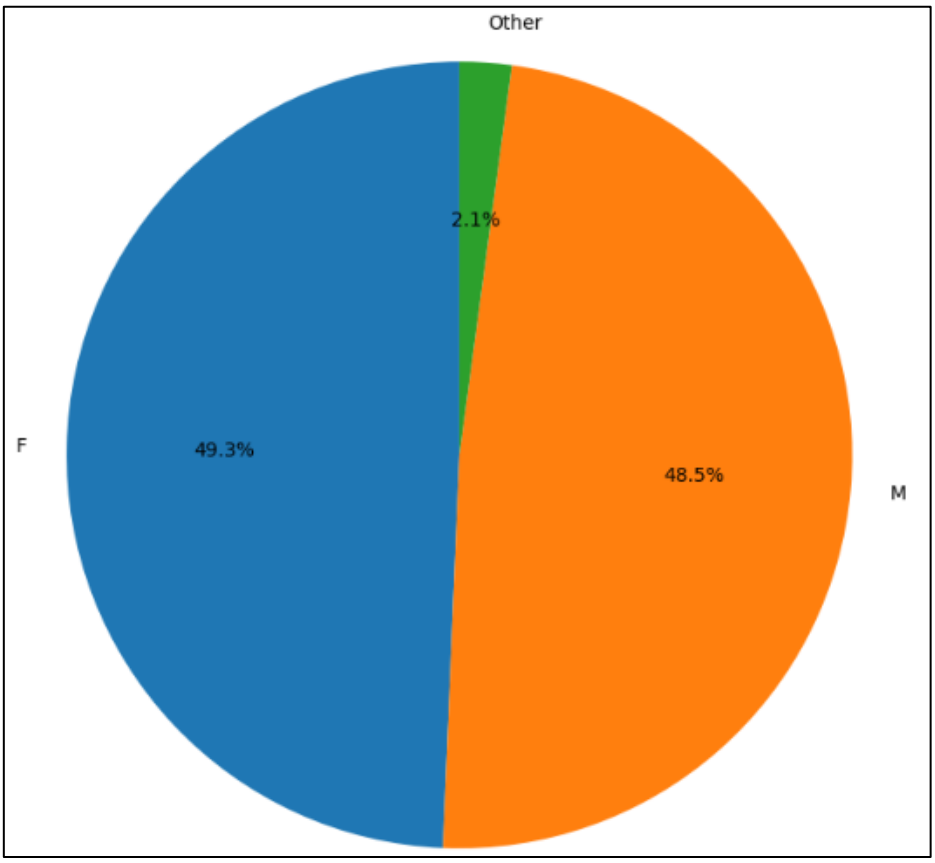


Source : Nos travaux, à partir des données fournies

4. Analyse du nombre de participants par sexe (gender)

Il ressort de l'analyse de ce graphique 4 que la proportion d'hommes et de femmes qui ont participé à l'étude sont sensiblement pareille, respectivement (49,3% et 48,5%). On note aussi une faible proportion de participants, soit 2,1%, qui ne s'identifient ni comme des hommes, ni comme des femmes mais comme des individus qui sont à la fois hommes et femmes et qui ne sont ni homme ni femmes.

Graphique 4 : Répartition des participants selon le sexe

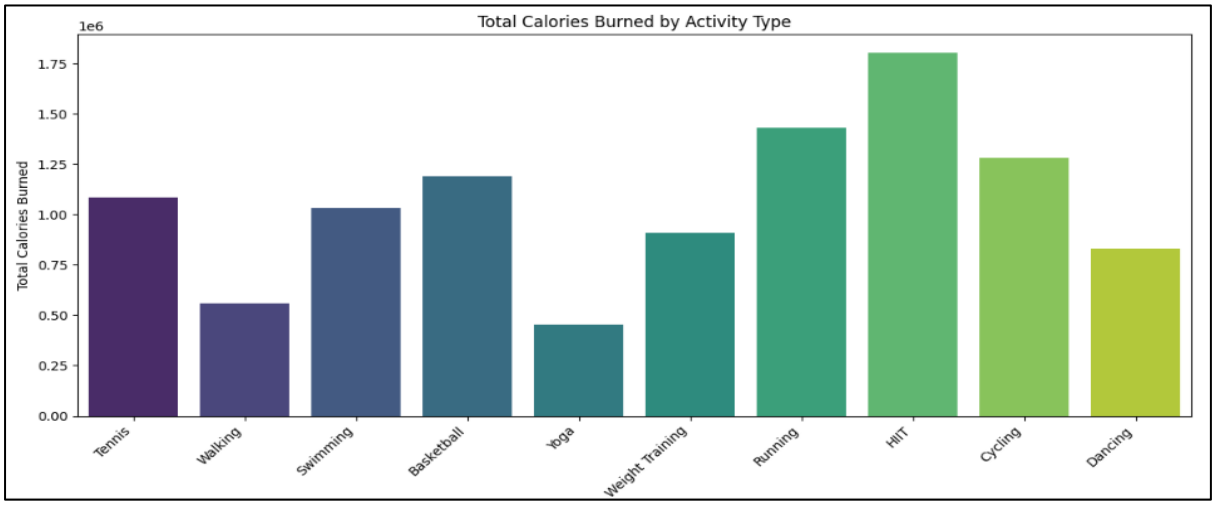


Source : Nos travaux, à partir des données fournies

5. Somme des calories brûlées pour chaque activité

Il ressort de l'analyse de ce graphique 5 que le **HIT** est l'activité qui a plus brulées de calories alors que le **yoga** est celle qui a le moins brules de calories.

Graphique 5 : Répartition du total des calories brulées selon l'activité

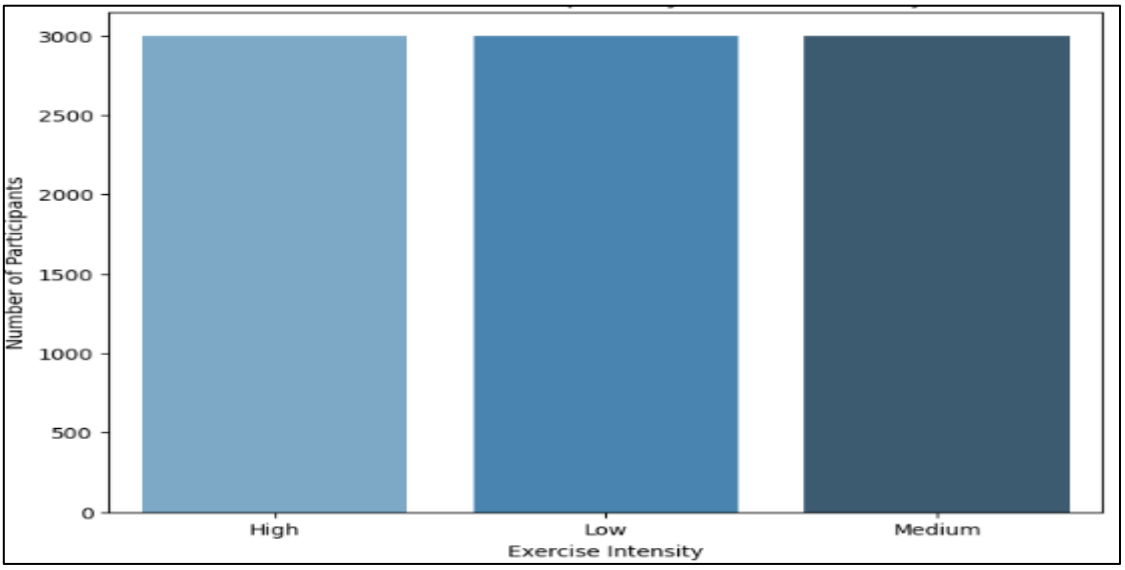


Source : Nos travaux, à partir des données fournies

6. Répartition des participants selon l'intensité de leurs exercices

Le graphique 6 ci-dessous montre la répartition du nombre de participants en fonction de l'intensité de leurs activités physiques. Chaque catégorie d'intensité – **Low (faible)**, **Medium (moyenne)** et **High (élevée)** – est associée à **3 000 participants**. Cela indique que chaque participant a effectué des activités à ces trois niveaux d'intensité au cours de la période étudiée. Autrement dit, l'ensemble des participants a une expérience variée, alternant entre des exercices de faible, moyenne et forte intensité, ce qui reflète une diversité dans les habitudes d'activité physique.

Graphique 6 : Répartition des participants selon l'intensité de leurs exercices



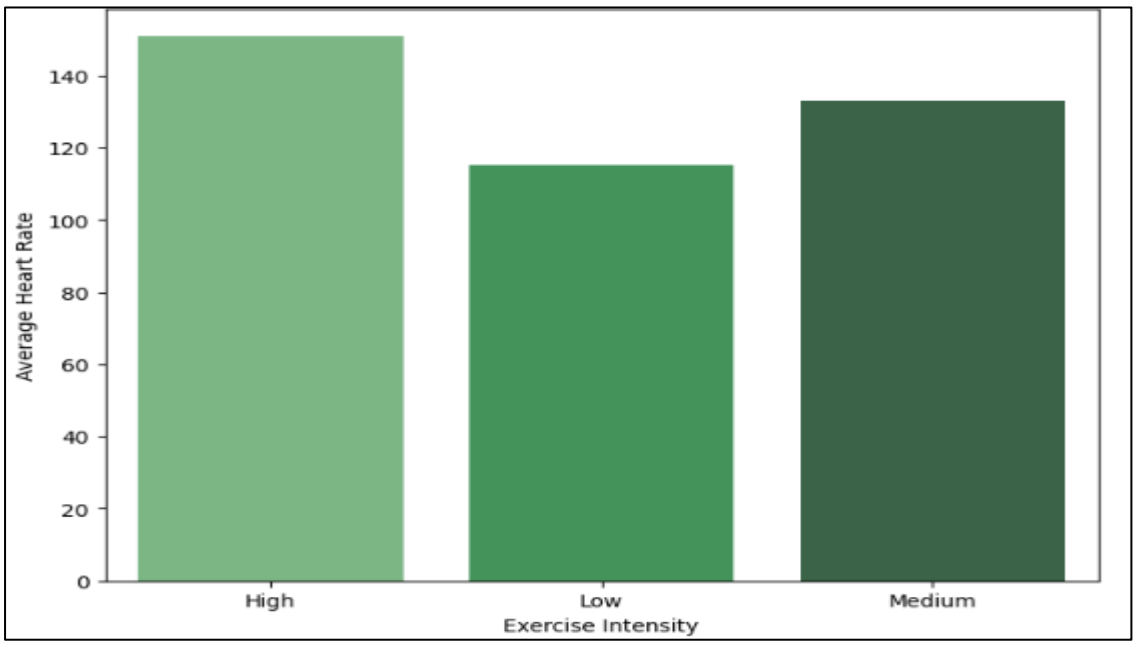
Source : Nos travaux, à partir des données fournies

7. Analyse de la fréquence cardiaque moyenne pour chaque niveau d'intensité

Ce graphique 7 montre une liaison directe entre l'intensité de l'exercice et la fréquence cardiaque moyenne, ce qui est attendu, car plus l'intensité augmente, plus le rythme cardiaque s'élève pour répondre aux besoins accrus en oxygène et en énergie du corps. Pour :

- **Intensité élevée (High)** : La fréquence cardiaque moyenne est de **151,08 bpm**, ce qui est cohérent avec l'augmentation de l'effort physique, impliquant une sollicitation cardiaque plus importante.
- **Intensité moyenne (Medium)** : La fréquence cardiaque moyenne est de **133,26 bpm**, indiquant une sollicitation modérée du système cardiovasculaire.
- **Intensité faible (Low)** : La fréquence cardiaque moyenne est de **115,43 bpm**, ce qui correspond à une activité physique légère, avec une sollicitation cardiaque réduite.

Graphique 7 : Fréquence cardiaque moyenne pour chaque niveau d'intensité



Source : Nos travaux, à partir des données fournies

8. Comparaison des fréquences cardiaques, des niveaux de stress et d'autres métriques entre les groupes "jamais", "ancien" et "actuel" fumeurs

Le tableau 1 présente les moyennes de la fréquence cardiaque, du niveau de stress, de l'IMC (indice de masse corporelle) et des calories brûlées selon le **statut tabagique** :

- **Ancien fumeur (Former)** : Fréquence cardiaque légèrement plus élevée (131,59 bpm) et IMC le plus élevé (33,35), ce qui pourrait indiquer un impact résiduel du tabagisme sur la santé.
- **Non-fumeur (Never)** : Niveau de stress légèrement plus élevé (5,25), mais IMC et calories brûlées similaires aux autres catégories.
- **Fumeur actuel (Current)** : Fréquence cardiaque moyenne la plus basse (131,31 bpm) et l'IMC le plus faible (33,25), avec des calories brûlées légèrement supérieures (15,42).

Tableau 1 : fréquences cardiaques, les niveaux de stress et d'autres métriques entre les groupes

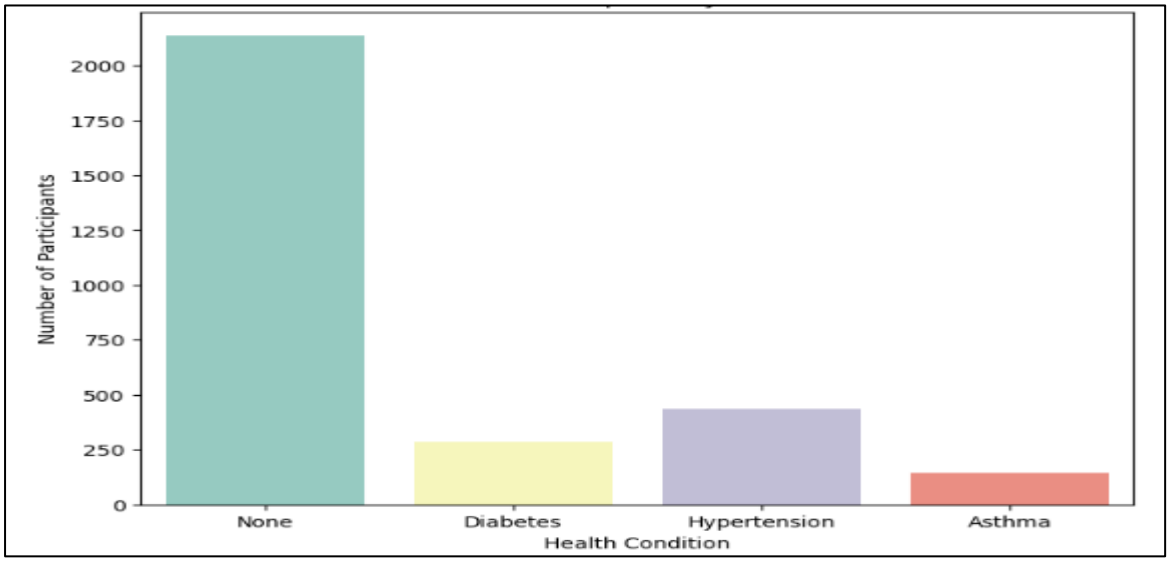
smoking_status	average_heart_rate	average_stress_level	average_bmi	average_calories_burned
Former	131.58530108768306	5.256778270644711	33.3529458589054	15.331317372546781
Never	131.43804702495203	5.253985124760077	33.27558248560398	15.390106285988736
Current	131.31438620910177	5.239337540322201	33.246729834196906	15.424153821557184

Source : Nos travaux, à partir des données fournies

9. Répartition des participants ayant ou non des problèmes de santé

Le graphique 8 montre que parmi les participants, **plus de 2 000** n'ont déclaré aucune condition de santé, ce qui indique une population majoritairement en bonne santé. Cependant, **environ 400** sont atteints d'hypertension, suivis de près de **283** présentant un diabète et **145** souffrant d'asthme. Ces chiffres soulignent la prévalence notable des maladies chroniques, notamment l'hypertension et le diabète, qui sont des indicateurs clés pour l'analyse de l'impact des habitudes de vie et de l'activité physique sur la santé.

Graphique 8 : Répartition des participants ayant ou non des problèmes de santé

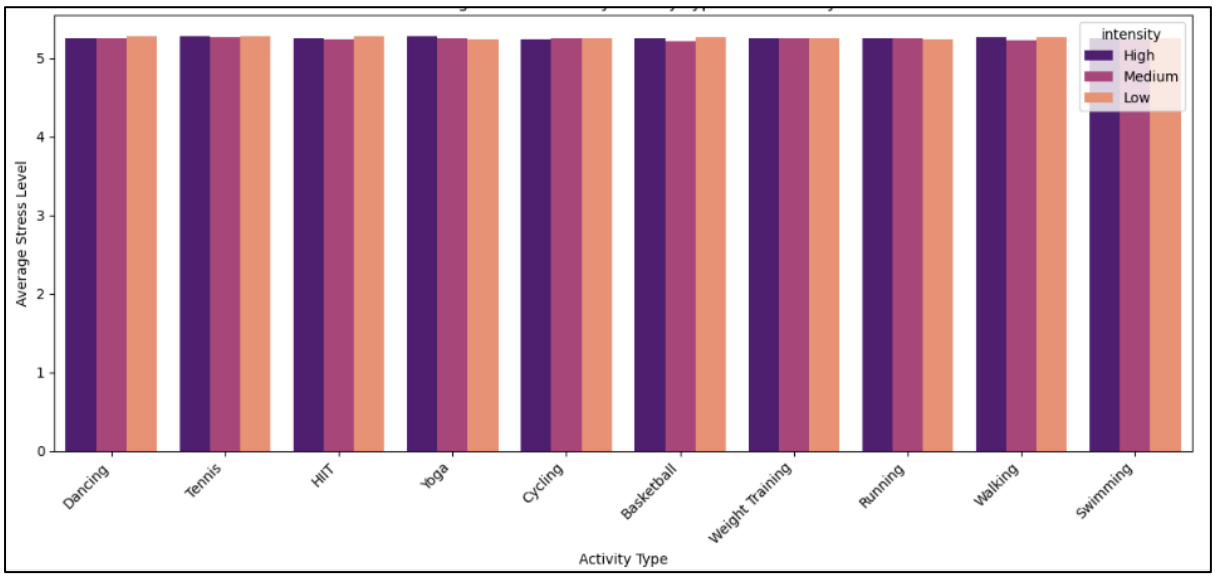


Source : Nos travaux, à partir des données fournies

10.Moyenne du niveau de stress pour chaque type et intensité d'activité

Le tableau montre que le **niveau de stress moyen** reste relativement stable, autour de **5.24 à 5.28**, quelle que soit l'activité ou l'intensité pratiquée. Cette homogénéité suggère que l'intensité de l'exercice ou le type d'activité a peu d'impact direct sur la variation du stress perçu, ou que d'autres facteurs externes pourraient jouer un rôle plus déterminant dans le niveau de stress quotidien des participants.

Graphique 9 : Moyenne du niveau de stress pour chaque type et intensité d'activité



Source : Nos travaux, à partir des données fournies

PRESENTATION DU TABLEAU DE BORD

Après analyse de ces indicateurs, nous avons mis en place un Dashboard avec le package Shiny Dashboard du logiciel R. Le Dashbord présente 7 onglets, à savoir l'onglet Démographie, Activités, Santé, Mode de vie, Rapport d'analyse, A propos du projet et Equipe. L'onglet Démographie fait présente quelques indicateurs en lien avec la pression artérielle systolique et diastolique moyenne, la répartition des participant par genre, par tranche d'âge et l'IMC moyen par tranche d'âge. Le second Onglet présente la durée moyenne par activité, la somme des calories brûlées par activité, la répartition des participants par intensité d'exercice et la moyenne des pas quotidiens. L'onglet Santé présente la fréquence cardiaque par intensité, l'état de santé et le statut de tabagisme des participants. L'onglet mode de vie présente les heures de sommeil par tranche d'âge, le niveau de stress par activité et la corrélation sommeil/fitness. L'application permet à l'utilisateur de choisir la période sur laquelle il veut visualiser les indicateurs et de télécharger ce rapport d'analyse dans l'onglet Rapport. Il a été déployé et est accessible à partir du lien suivant : FitLife360 Dashboard

Voici un aperçu du tableau de bord déployé :

LIMITES ET RECOMMANDATIONS DE L'ÉTUDE

1. Limites de l'étude

1. Limites méthodologiques :

- L'utilisation de données synthétiques (FitLife360) peut ne pas refléter parfaitement la réalité des comportements en matière de santé
- Absence de données longitudinales sur plusieurs années pour observer l'évolution des tendances
- Manque d'informations sur les antécédents médicaux des participants
- Absence de données qualitatives sur les motivations et les obstacles à l'activité physique

2. Limites liées à l'échantillon :

- Échantillon limité à 3000 participants, ce qui peut réduire la généralisation des résultats
- Possible biais de sélection dans la composition de l'échantillon
- Représentativité géographique non spécifiée
- Manque de diversité dans certaines catégories d'âge (sous-représentation des 18-25 ans)

3. Limites des variables :

- Absence de variables socio-économiques pouvant influencer les comportements de santé
- Mesure subjective du stress qui pourrait bénéficier d'indicateurs plus objectifs
- Manque de détails sur la nature exacte des problèmes de santé déclarés
- Absence d'information sur le régime alimentaire des participants

2. Recommandations

1. Pour améliorer la collecte de données :

- Élargir l'échantillon pour une meilleure représentativité
- Inclure des données socio-économiques et environnementales
- Mettre en place un suivi longitudinal sur plusieurs années
- Ajouter des mesures objectives du stress et de la qualité du sommeil
- Intégrer des données sur les habitudes alimentaires

2. Pour les futures recherches :

- Conduire des études qualitatives complémentaires pour comprendre les motivations
- Analyser l'impact des facteurs environnementaux sur l'activité physique
- Étudier les corrélations entre le mode de vie et les problèmes de santé spécifiques
- Développer des modèles prédictifs pour identifier les risques de santé

3. Pour l'amélioration des interventions :

- Développer des programmes personnalisés selon les tranches d'âge
- Mettre en place des interventions ciblées pour les groupes à risque (IMC élevé)
- Renforcer le suivi des anciens fumeurs qui présentent des indicateurs de santé préoccupants
- Créer des stratégies spécifiques pour la gestion du stress

4. Pour le développement technologique :

- Améliorer la précision des mesures des dispositifs de suivi
- Développer des outils plus sophistiqués pour l'analyse en temps réel
- Intégrer l'intelligence artificielle pour des recommandations personnalisées
- Renforcer la sécurité et la confidentialité des données collectées

Ces recommandations visent à améliorer la qualité et la portée des futures études dans ce domaine, tout en permettant une meilleure compréhension des facteurs influençant la santé et le bien-être des participants.

CONCLUSION

L'analyse des données met en évidence plusieurs tendances importantes. La **répartition par tranche d'âge** montre une prédominance des participants âgés de **36 ans et plus**, qui représentent la majorité de l'échantillon. Concernant l'**Indice de Masse Corporelle (IMC)**, il reste relativement élevé dans toutes les tranches d'âge, avec un IMC moyen supérieur à **30**, suggérant une prévalence de l'obésité.

La **durée moyenne des activités** est stable, autour de **70 minutes**, quel que soit le type d'activité. La **répartition par sexe** révèle une quasi-parité entre les hommes et les femmes, tandis qu'une petite proportion des participants s'identifie hors du binarisme traditionnel. L'**analyse des calories brûlées** montre que le **HIIT** est l'activité la plus intense, tandis que le **yoga** est celle qui consomme le moins de calories.

Chaque participant a pratiqué des exercices à **faible, moyenne et forte intensité**, et la **fréquence cardiaque moyenne** est proportionnelle à l'intensité de l'exercice : elle s'élève à **151 bpm** pour les exercices intenses, contre **115 bpm** pour les activités légères. En comparant les groupes de **fumeurs** et **non-fumeurs**, les anciens fumeurs présentent un IMC plus élevé et une fréquence cardiaque légèrement supérieure.

Enfin, l'analyse des **problèmes de santé** révèle que la majorité des participants ne présentent aucune pathologie. Cependant, **l'hypertension** et le **diabète** restent fréquents, touchant respectivement **434** et **283 participants**. Le **niveau de stress** reste globalement constant quelle que soit l'intensité ou le type d'activité, suggérant que d'autres facteurs externes influencent ce paramètre.