Visualisation de données : rapport

AUTEF Lucas, BASSOUM Mohamed Emine, GUEMIMI Marouane,
JEANNE Arthur, WIATT Chloé

27/03/2025

---

1 Choix et présentation du jeu de données

Notre jeu de données est issu de Kaggle [1]. Il s'intitule *Sleep Health and Lifestyle Dataset* et traite donc du lien entre la qualité du sommeil et le mode de vie des individus. Il s'agit d'un jeu de données bien noté par Kaggle selon différents critères (complétude, crédibilité, etc.) ainsi que par les utilisateurs.

Il comporte 374 lignes et 13 colonnes. Il s'agit donc d'un échantillon assez modeste mais intéressant du point de vue du nombre et de la diversité des colonnes :

— Un identifiant unique par individu

— Le genre

— L'âge

— La profession

— La durée du sommeil (heures)

— La qualité du sommeil (1 à 10, évaluée subjectivement par la personne elle-même)

— Le niveau d'activité physique (temps quotidien d'activité sportive, en minutes/jour)

— Le niveau de stress (1 à 10, évalué subjectivement par la personne)

— La catégorie d'IMC (sous-poids, normal, surpoids)

— La tension artérielle (systolique/diastolique)

— La fréquence cardiaque (battements par minute)

— Le nombre de pas quotidiens

— Les troubles du sommeil (aucun, insomnie, apnée du sommeil)

Cette diversité a motivé notre choix de ce jeu de données, offrant un large éventail de questions possibles et de représentations possibles des données pour multiplier les points de vue. Le sujet a également semblé intéressant à l'ensemble du groupe, en ce que le sommeil est un sujet omniprésent dans notre quotidien.

1. `https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset/data`

2 Questions et objectifs posés

Nous avons divisé les colonnes en deux catégories :

— Les facteurs influençant potentiellement le sommeil (le genre, l'âge, la profession, le niveau d'activité physique, le niveau de stress, la catégorie d'IMC, la tension artérielle, la fréquence cardiaque et le nombre de pas quotidiens)

— Les variables permettant de mesurer la qualité générale du sommeil (durée du sommeil, qualité du sommeil, troubles du sommeil)

À la lumière de cette classification, les questions se divisent à leur tour en quatre grandes catégories, la deuxième étant notre problématique principale :

1. Comment sont répartis les différents caractères étudiés dans la population ? Cette question nous permet d'avoir un premier aperçu global du jeu de données. Pour y répondre, nous étudions la distribution de chaque variable une par une.

2. **Quelle est l'influence du mode de vie sur la qualité du sommeil, et notamment sur les troubles du sommeil ?** Pour répondre à cette problématique, nous étudions les corrélations entre une des trois variables étudiées et un ou plusieurs facteurs l'influençant. On peut s'attendre à quelques fortes corrélations, comme le niveau d'activité physique et le nombre de pas quotidiens, qui sont des données proches.

3. Existe-t-il une corrélation entre les différents facteurs étudiés ? Si une très forte corrélation est constatée, on pourra éventuellement regrouper des colonnes.

4. Existe-t-il une corrélation entre les variables étudiées ? On peut s'attendre à une forte corrélation ici puisque les trois variables mesurent la qualité du sommeil.

3 Pré-traitement

La base de données nécessitait peu de pré-traitements. En effet, comme on le voit sur le graphique ci-dessous, aucune colonne ne comportait de donnée manquante. Nous avons seulement rencontré un problème dans le code lié à la colonne Sleep disorder (troubles du sommeil) : la valeur "aucun trouble" était écrite comme "NaN" dans la base de données, ce qui était comptabilisé par Python comme une absence de donnée. Nous avons donc renommé les cellules correspondantes en "No disorder".

Un deuxième problème est lié à la saisie des données dans la colonne "BMI Category" (catégorie d'IMC) : d'après Kaggle, il existe normalement trois catégories : Underweight, Normal, Overweight (Sous-poids, Normal, Surpoids), mais certaines cases étaient labellisées "Normal" et d'autres "Normal weight". Nous avons donc réuni les deux catégories en une seule : "Normal".

4 Choix des représentations

Nous avons veillé à varier les formes de représentations, sans dénaturer le sens des données. Le choix des représentations est fortement lié à la nature des données elles-mêmes : s'il s'agit d'une distribution (une seule

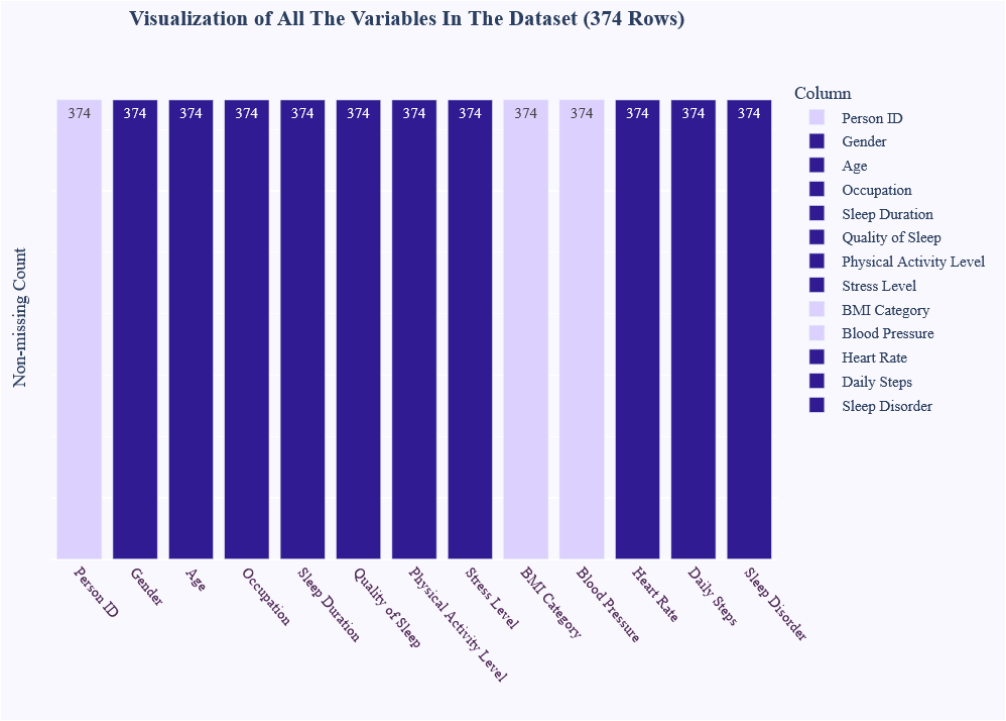

Figure 1 – Vérification des données manquantes

colonne), on privilégiera un diagramme circulaire, un diagramme en barres ou, de façon plus originale, un diagramme sunburst ou encore un treemap (carte proportionnelle).

Avec deux données, d'autres possibilités s'offrent à nous, suivant le type de variable : qualitative ou quantitative, continue ou discrète. On peut choisir une simple représentation sous forme de points ou de courbe, notamment pour les variables continues. Si l'une des variables est catégorielle ou discrète, on privilégiera les diagrammes en boîte, en violon ou en barres, ou encore une carte de densité.

Pour afficher davantage de variables, on peut utiliser un modèle 3D, ou garder un graphique en 2D et représenter une troisième variable à l'aide des couleurs ou de la taille des points (alors appelés bulles). Un diagramme sunburst interactif (Plotly express) permet de représenter de multiples variables sur un même graphique. On peut également garder deux données mais afficher davantage d'informations, comme la courbe moyenne, la distribution des données (violons, boîtes).

5 Notre dashboard

Voici le lien de notre dashboard avec tous nos graphiques, répartis suivant nos quatre grandes problématiques. Nous joignons également à ce rapport le code du Dashboard (Github) et nos codes pour les graphiques au format .ipynb.

6 Conclusions sur les questions posées

6.1 Comment sont répartis les différents caractères étudiés dans la population ?

Concernant les professions, les individus étudiés viennent notamment du domaine médical (infirmiers, médecins) ; on a également les domaines de l'éducation et de la justice, de la technologie et du business. L'échantillon représente donc seulement une fraction des métiers existants ; on peut penser que les auteurs de l'étude ont volontairement choisi des personnes avec des métiers stressants, pouvant donner lieu à des troubles du sommeil. À noter que dans la catégorie "Manager", on a un seul individu, ce qui n'est pas très représentatif.

Le nombre d'hommes et de femmes sont quasiment égaux dans la population étudiée (185 femmes contre 189 hommes).

Les âges vont de 26 à 59 ans, avec davantage de personnes au centre de la plage de valeurs (pic autour de 44 ans).

Concernant les IMC, on a une majorité de personnes avec un IMC normal (près de 60%), suivie de la catégorie surpoids (environ 40%). Les personnes obèses représentent un faible pourcentage (moins de 3%).

Concernant la qualité du sommeil, les valeurs ne descendent pas en dessous de 4/10, même pour les personnes atteintes de troubles du sommeil. Les valeurs sont concentrées entre 6 et 9, avec une absence de 10.

Les personnes atteintes de troubles du sommeil représentent environ 40% de l'échantillon, avec la moitié atteinte d'insomnie, et l'autre moitié d'apnée du sommeil.

Quant aux niveaux de stress, ils vont de 3 à 8 et sont presque équitablement répartis.

Le nombre de pas par jour va de 3000 à 11000, avec une majorité entre 5000 et 9000.

6.2 Quelle est l'influence du mode de vie sur la qualité du sommeil, et notamment sur les troubles du sommeil ?

On remarque que la durée du sommeil est dépendante de plusieurs facteurs biologiques :

— Les personnes obèses et en surpoids ont une durée de sommeil légèrement plus courte que les personnes avec un IMC standard

— Les femmes sont plus sujettes à l'apnée du sommeil que les hommes dans l'échantillon (ce qui n'est pas le cas dans la population générale d'après d'autres études ; c'est même l'inverse).

— Les personnes qui dorment moins longtemps ont une fréquence cardiaque plus élevée

— La pression artérielle est plus élevée chez les personnes ayant une faible qualité de sommeil

— L'insomnie est beaucoup moins présente chez les personnes pratiquant une activité physique régulière à contrario l'apnée du sommeil est plus prépondérante chez les personnes sportives

— La qualité et la durée du sommeil diminuent clairement lorsque le stress augmente

— L'insomnie apparaît plutôt à la quarantaine alors que l'apnée du sommeil se développe à la cinquantaine

— La répartition des temps de sommeil est très différente selon la catégorie socioprofessionnelle

6.3 Existe-t-il une corrélation entre les différents facteurs étudiés ?

— **Niveau de stress en fonction de l'âge :** le niveau de stress a été regroupé en trois catégories (3-4, 5-6 et 7-8, correspondant à stress faible, médium et élevé). On remarque un chevauchement important entre les catégories de stress, notamment entre 30 et 50 ans, ce qui indique que les niveaux de stress varient significativement dans cette tranche d'âge. Après 50 ans, la catégorie "Low Stress" domine, montrant une tendance à la diminution du stress avec l'âge.

— **Profession et niveau de stress :** on observe une corrélation entre les professions à haute responsabilité (médecins, managers) et un niveau de stress élevé. Les métiers nécessitant de la créativité et de la recherche (scientifiques, ingénieurs) montrent des niveaux de stress variés. Les métiers de la vente semblent être les plus stressants, probablement à cause de la pression financière et des objectifs de performance.

— **Fréquence cardiaque en fonction du nombre quotidien de pas et suivant la catégorie d'IMC :** les personnes actives (6 000 - 10 000 pas) ont une meilleure condition cardiovasculaire que celles marchant peu (3 000 - 4 000 pas). Les personnes obèses ont généralement une fréquence cardiaque plus élevée et marchent moins que celles ayant un IMC normal ou en surpoids. Ainsi, l'exercice régulier est bien associé à une meilleure santé cardiaque.

— **Fréquence cardiaque en fonction de l'âge :** le rythme cardiaque moyen fluctue avec l'âge sans tendance linéaire claire, avec plusieurs pics et creux. On constate des hausses notables autour de 28, 34, 48 et 55 ans, et des baisses vers 31, 35, 45 et 53 ans. Les barres d'erreur montrent une variabilité plus ou moins forte selon les âges.

6.4 Existe-t-il une corrélation entre les variables étudiées ?

Il semblerait que la durée de sommeil ait un impact sur la qualité du sommeil : plus la durée du sommeil est grande, plus les personnes interrogées pensent avoir un sommeil de qualité.

Les personnes insomniaques ont, sans surprise, un sommeil moins long, tandis que les personnes atteintes d'apnée du sommeil peuvent avoir un sommeil plus long que la normale, comme moins long (la répartition de la durée du sommeil pour les personnes atteintes d'apnée du sommeil est très étendue).