



2025

RAPPORT BIG DATA

Réalisé par

Georgy GUEI
Verdiane KOCOUISSO PLOMEY
Guillaume LE FORMAL
Mohamed BENTAMA SERROUKH



SOMMAIRE

- **1 Introduction**
 - Présentation du sujet et objectifs
- **2 Présentation du Dataset**
 - Description des données et de leurs structures
- **3 Outils et Architecture**
 - Architecture en zones et outils utilisés
- **4 Traitement de la donnée**
 - Etapes d'ingestion, de transformation et de structuration
- **5 Analyse et Résultats**
 - Résultats corrélation et visualisation
- **6 Conclusion**
 - Synthèse des analyses (et ouverture)

Source de nos deux notebooks:

<https://shorturl.at/GWZp7>

<https://shorturl.at/agK7B>





Introduction

Dans un monde où les rythmes de la vie moderne influencent directement la santé et le bien-être des individus, il est essentiel d'étudier les interactions entre le sommeil, l'activité physique, l'occupation, le stress et d'autres facteurs liés au mode de vie.

Ce projet a pour but d'explorer un ensemble de données combinant des informations démographiques, de santé et de sommeil afin de mieux comprendre ces relations et de suggérer des moyens d'améliorer le bien-être général.

CONTEXTE DU PROJET

Le sommeil joue un rôle crucial dans la santé physique et mentale. Cependant, il n'agit pas seul : des facteurs comme l'activité physique, l'activité professionnelle, les niveaux de stress et les caractéristiques démographiques (âge, sexe) interagissent pour influencer la qualité et la durée du sommeil.


En analysant ces données, nous visons à répondre à des questions clés, telles que :

Comment les groupes démographiques dorment-ils ?

Existe-t-il une corrélation entre l'activité physique et la qualité du sommeil ?

Les niveaux de stress affectent-ils directement la durée ou la qualité du sommeil ?

Y a-t-il un lien entre la profession exercée, le niveau de stress et la qualité du sommeil ?





Introduction

NOS OBJECTIFS

Les objectifs de ce projet sont les suivants :


Nettoyage et préparation des données : Identifier et traiter les données manquantes, aberrantes ou en double, et structurer les données selon un modèle logique pour faciliter l'analyse.

Analyse descriptive et exploratoire : Comprendre la distribution des variables (âge, profession, activité physique, sommeil, etc.) et identifier les tendances clés.

Création de modèles en étoile : Mettre en place une architecture de données intuitive pour réaliser des analyses multidimensionnelles, notamment sur les relations entre le sommeil, la santé et les facteurs démographiques.

Génération d'idées : Formuler des observations et des recommandations concrètes sur la base des résultats analytiques.

Rapport final : Fournir un résumé des résultats et des implications potentielles pour l'amélioration du mode de vie.





Présentation du Dataset

DESCRIPTION DES DONNÉES ET DE LEURS STRUCTURES

Dans ce projet, nous avons utilisé un dataset riche et varié comprenant plusieurs tables pour explorer les interactions entre le sommeil, l'activité physique, le stress et d'autres facteurs liés au mode de vie. Voici une description détaillée des différentes tables et de leur structure:

Table de Faits (Fact_Heath) :

La table de faits contient les informations principales sur les comportements et les états des individus au quotidien. Elle inclut :

- **Person_ID** : Identifiant unique de chaque individu.
- **Sleep_Duration** : La durée de sommeil de chaque individu, mesurée en heures.
- **Physical_Activity_Level** : Le niveau d'activité physique des individus, généralement exprimé en nombre de minutes d'activité physique par jour ou en intensité.
- **Daily_Steps** : Le nombre de pas quotidiens effectués par les individus, mesuré par des dispositifs de suivi de l'activité physique comme les podomètres.
- **Stress_Level** : Le niveau de stress ressenti par les individus, souvent mesuré par des auto-évaluations ou des dispositifs de suivi de la santé.



Présentation du Dataset

(8) Spark Jobs

fact_health_df: pyspark.sql.dataframe.DataFrame = [Person_ID: string, Sleep_Duration: float ... 3 more fields]

gold_fact_health_df: pyspark.sql.dataframe.DataFrame

Schema Details History

Person_ID: string
Sleep_Duration: float
Physical_Activity_Level: integer
Daily_Steps: integer
Stress_Level: integer

Person_ID	Sleep_Duration	Physical_Activity_Level	Daily_Steps	Stress_Level
135	7.3	60	8000	5
150	8.0	80	7500	3
252	6.8	30	6000	6
368	8.0	75	7000	3
1	6.1	42	4200	6

only showing top 5 rows

Tables de Dimensions :

Les tables de dimensions apportent des informations contextuelles qui enrichissent les analyses effectuées à partir de la table de faits. Elles incluent :

- **Dimensions Démographiques (Dim_Person) :**
 - **Person_ID :** Identifiant unique de chaque individu.
 - **Gender :** Le genre des individus.
 - **Age :** L'âge des individus, permettant de segmenter les analyses par groupe d'âge.
 - **Occupation :** La profession des individus, permettant d'explorer les différences en fonction des types d'occupation.
- **Dimensions du Sommeil :**
 - **Person_ID :** Identifiant unique de chaque individu.
 - **Quality_of_Sleep :** Qualité du sommeil, mesurée par des indicateurs tels que la durée du sommeil profond et le nombre de réveils nocturnes.
 - **Sleep_Disorder :** Indication de troubles du sommeil, tels que l'insomnie ou l'apnée du sommeil.

Présentation du Dataset

▶ (7) Spark Jobs

▼ dim_person_df: pyspark.sql.dataframe.DataFrame

Person_ID: string
Gender: string
Age: integer
Occupation: string

▶ gold_dim_person_df: pyspark.sql.dataframe.DataFrame = [Person_ID: string, Gender: string ... 2 more fields]

Person_ID	Gender	Age	Occupation
30	Male	30	Doctor
178	Male	42	Salesperson
150	Female	39	Accountant
183	Male	42	Lawyer
223	Male	44	Salesperson

only showing top 5 rows

Dim_Person - Table de dimension démographique

▶ (8) Spark Jobs

▶ fact_health_df: pyspark.sql.dataframe.DataFrame = [Person_ID: string, Sleep_Duration: float ... 3 more fields]

▼ gold_fact_health_df: pyspark.sql.dataframe.DataFrame

Schema Details History

Person_ID: string
Sleep_Duration: float
Physical_Activity_Level: integer
Daily_Steps: integer
Stress_Level: integer

Person_ID	Sleep_Duration	Physical_Activity_Level	Daily_Steps	Stress_Level
135	7.3	60	8000	5
150	8.0	80	7500	3
252	6.8	30	6000	6
368	8.0	75	7000	3
1	6.1	42	4200	6

only showing top 5 rows

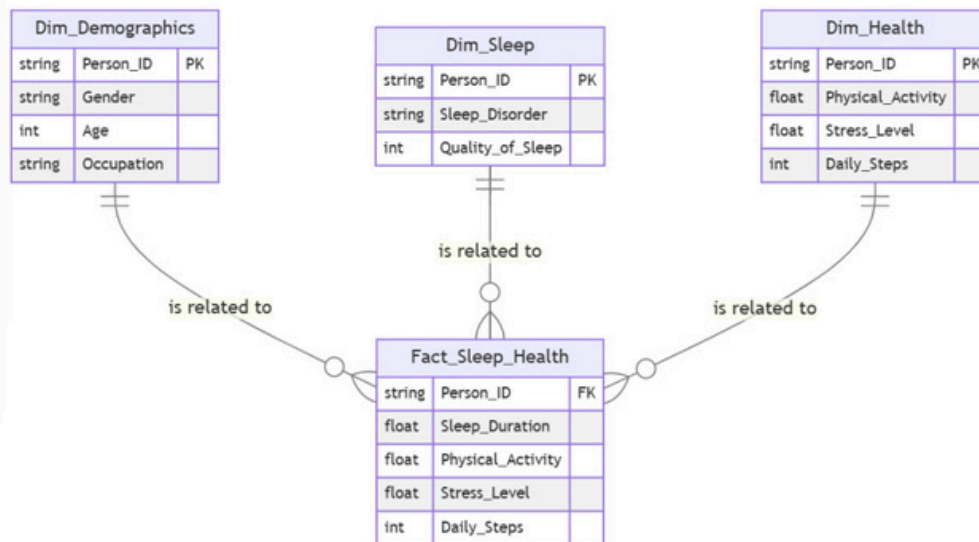
Dim_Sleep - Table de dimension du sommeil

Présentation du Dataset

STRUCTURE DES DONNÉES

Les données sont structurées sous forme de tables relationnelles, ce qui facilite les jointures et les analyses croisées. Voici un aperçu de la structure relationnelle :

- La **table de faits** est au centre de l'analyse, reliée aux différentes tables de dimensions par des clés étrangères (Foreign Keys), comme *Person_ID*.
- Les **tables de dimensions** fournissent des informations supplémentaires qui permettent de contextualiser les données de la table de faits et d'effectuer des analyses plus détaillées.

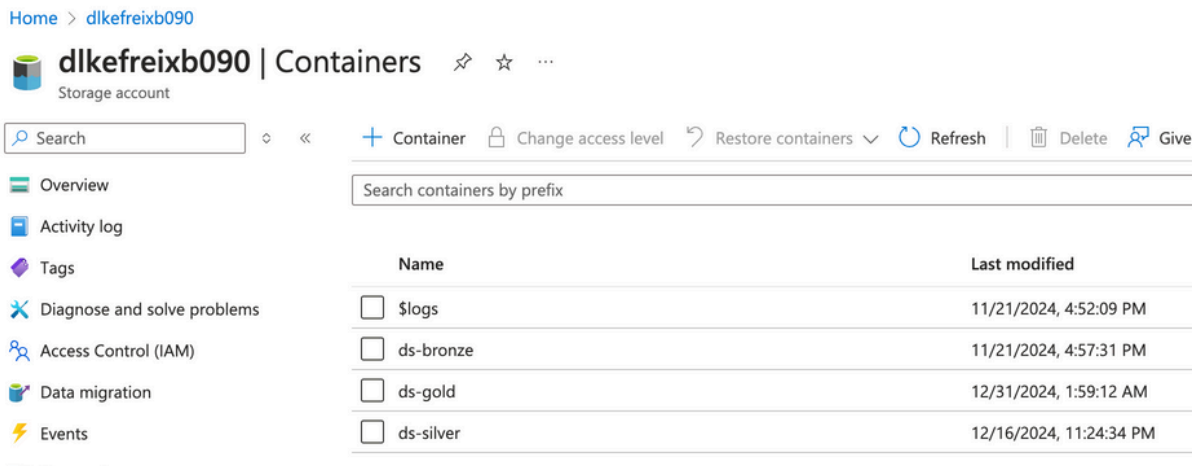


Modélisation : Une structure en étoile est conçue avec une table de faits (mesures) et des tables de dimensions (attributs descriptifs).

Outils et Architecture

ARCHITECTURE EN ZONES ET OUTILS UTILISÉS

L'architecture de notre projet est divisée en plusieurs zones fonctionnelles, chacune jouant un rôle clé dans le traitement et l'analyse des données. Voici une description détaillée des différentes zones et des outils utilisés :

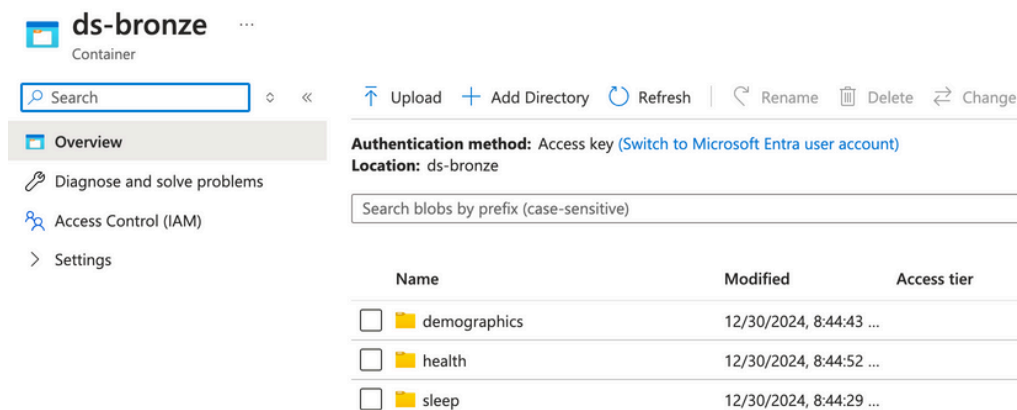


Aperçu des zone sur Azure

1. Zone d'Ingestion (ds-bronze) :

La zone d'ingestion est le point d'entrée des données brutes dans notre système. Les données sont collectées à partir de diverses sources et formats, y compris des fichiers CSV, des bases de données relationnelles, et des API de suivi de la santé. Les principales étapes et outils utilisés dans cette zone sont :

- **Sources de Données** : Fichiers CSV, bases de données, API.
- **Spark** : Pour lire et charger les données de manière distribuée et efficace.
- **DataFrames PySpark** : Utilisés pour structurer les données ingérées de manière tabulaire.



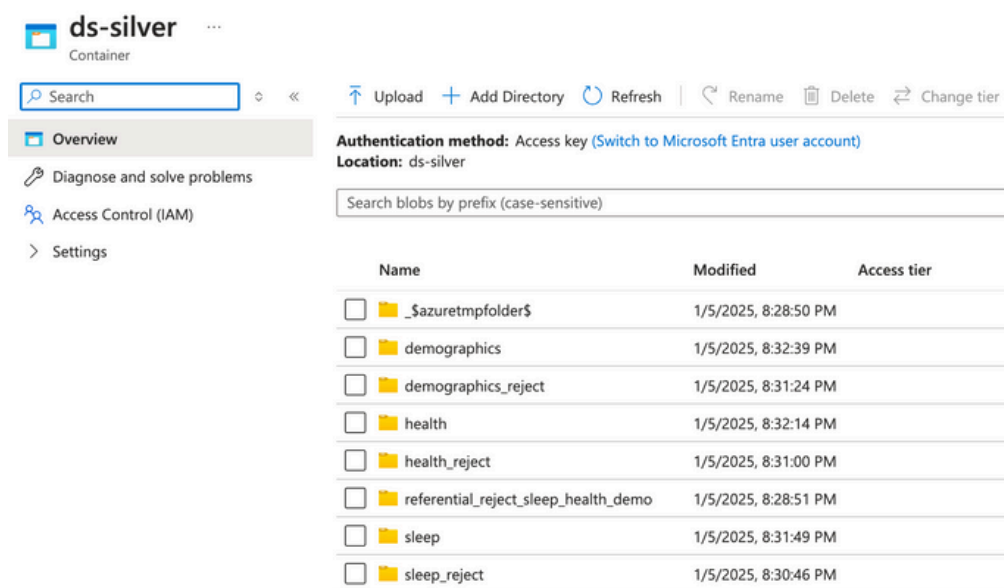
Aperçu de la zone de Transformation (ds-silver)

Outils et Architecture

2. Zone de Transformation (ds-silver) :

La zone de transformation est où les données brutes sont nettoyées, transformées et enrichies pour préparer les analyses ultérieures. Les principales étapes et outils utilisés dans cette zone sont :

- **Nettoyage des Données** : Traitement des valeurs manquantes, correction des anomalies et filtrage des données indésirables.
- **Transformation des Données** : Agrégation, normalisation et création de nouvelles variables.
- **PySpark** : Pour les opérations de transformation et de manipulation des données à grande échelle.



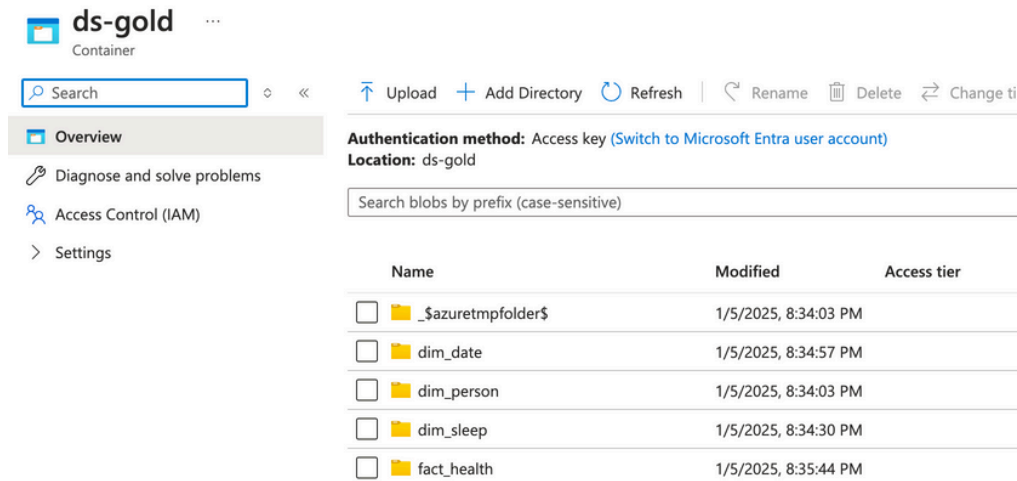
Aperçu de la zone de Transformation (ds-silver)

Outils et Architecture

3. Zone de Structuration (ds-gold) :

La zone de structuration est où les données transformées sont organisées et stockées de manière structurée, prêtes pour l'analyse. Les principales étapes et outils utilisés dans cette zone sont :

- **Organisation des Données** : Structurer les données sous forme de tables de dimensions et de faits, facilitant les jointures et les analyses croisées.
- **DataFrames PySpark** : Pour stocker et organiser les données structurées.
- **Jointures des Tables** : Pour relier les tables de faits et de dimensions et créer un DataFrame enrichi.

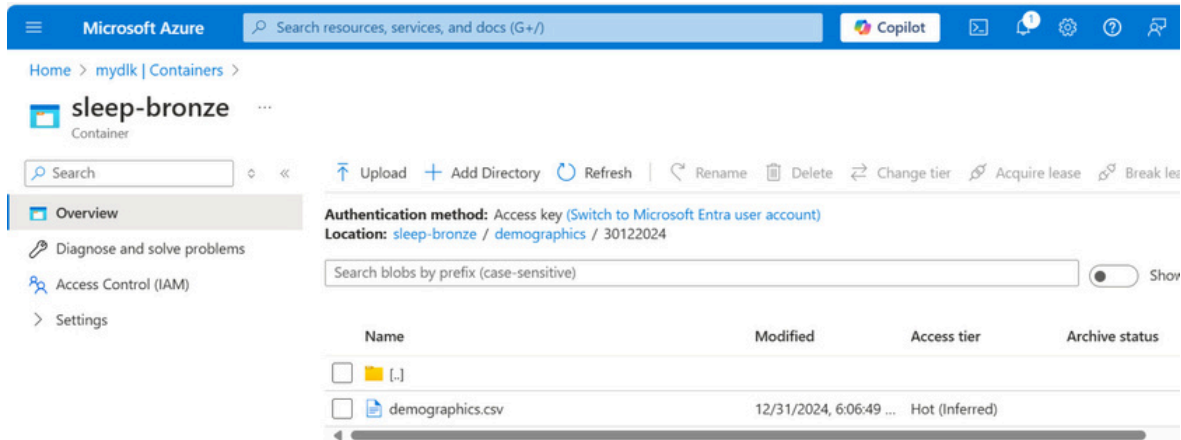


Aperçu de la zone de Structuration (ds-gold)

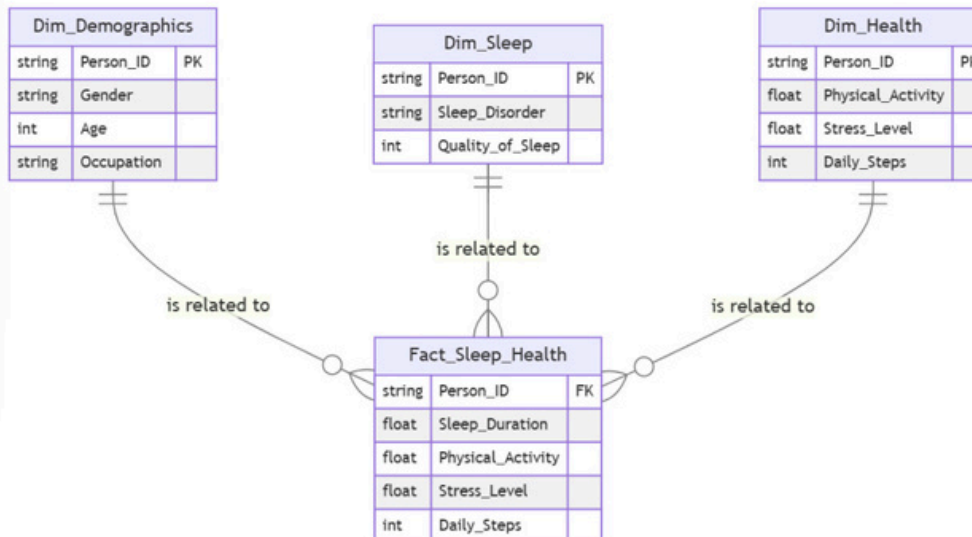
Modélisation : Une structure en étoile est conçue avec une table de faits (mesures) et des tables de dimensions (attributs descriptifs).

Outils et Architecture

ARCHITECTURE



Modélisation : Une structure en étoile est conçue avec une table de faits (mesures) et des tables de dimensions (attributs descriptifs).





Outils et Architecture

ARCHITECTURE


Analyse statistique et interprétation : Des mesures descriptives (moyenne, médiane, écarts types) et des visualisations sont générées pour répondre aux questions d'analyse.

LES OUTILS

Databricks/PySpark: Databricks permet le traitement de Dataframe. Par des requêtes PySpark on peut modifier la forme et le contenu de la donnée pour du séquençement, du nettoyage ou bien de l'analyse.

Azure: Nous avons utilisé un groupe de ressource et un compte de stockage pour initier le projet. Mais aussi Azure Blob Storage qui représente le container général ayant la donnée de nos 3 zones peut contenir une connexion directe avec Databricks.

Matplotlib: C'est avec matplotlib que nous faisons ressortir graphiquement nos résultats statistiques.



Traitement de la donnée

Zone Silver : Le premier traitement se fait lors de la mise en place des “normes” qu’on veut appliquer à la donnée, notamment avec l’intégration de la table de rejet.

```
from pyspark.sql.types import StructType, StructField, IntegerType, StringType

# Insertion des causes de rejet
reject_data = [
    (1000, 'Sleep', 'Durée de sommeil manquante ou invalide'),
    (1001, 'Sleep', 'Qualité de sommeil manquante ou invalide'),
    (1002, 'Sleep', 'Trouble de sommeil non spécifié'),
    (2000, 'Health', 'Niveau d\'activité physique manquant ou invalide'),
    (2001, 'Health', 'Nombre de pas quotidien non spécifié'),
    (2002, 'Health', 'Niveau de stress non valide'),
    (3000, 'Demographics', 'Genre non spécifié ou invalide'),
    (3001, 'Demographics', 'Âge manquant ou invalide'),
    (3002, 'Demographics', 'Profession non spécifiée'),
    (4000, 'General', 'Identifiant Person ID manquant ou non unique')
]

# Définir le schéma de la table des causes de rejet
reject_schema = StructType([
    StructField("IdReject", IntegerType(), nullable=False),
    StructField("targetTable", StringType(), nullable=False),
    StructField("rejectCause", StringType(), nullable=False)
])
```

```
# Charger les données de la table des causes de rejet
reject_df = spark.read.format("delta").load(reject_table_path)

# Afficher les données de la table des causes de rejet
reject_df.show()
```

```
▶ reject_df: pyspark.sql.dataframe.DataFrame = [IdReject: integer, targetTable: string ... 1 more field]
```

```
+-----+-----+-----+
|IdReject| targetTable|      rejectCause|
+-----+-----+-----+
| 2000|      Health|Niveau d'activité...|
| 2001|      Health|Nombre de pas quo...|
| 1001|       Sleep|Qualité de sommei...|
| 1000|       Sleep|Durée de sommeil ...|
| 3000|Demographics|Genre non spécifi...|
| 3002|Demographics|Profession non sp...|
| 4000|      General|Identifiant Perso...|
| 1002|       Sleep|Trouble de sommei...|
| 3001|Demographics|Âge manquant ou i...|
| 2002|      Health|Niveau de stress ...|
+-----+-----+-----+
```

Traitement de la donnée

On a donc la définition de conditions à respecter et nous filtrons la donnée.

Zone Gold : Mise en place des tables de dimensions, nous en avons 3, “*dim_person*”, contenant des informations comme le genre, l'âge ou l'occupation.

“*dim_sleep*”, on y trouve des détails sur la qualité et troubles du sommeil et finalement “*dim_date*” ajoutant une dimension temporelle avec des périodes analytiques.

Sans oublier notre table de fait “*fact_health*” permettant la corrélation entre nos tables de dimensions et apporte une combinaison entre les données du sommeil, de l'activité physique et du stress.

30

```
# Joindre les DataFrames health_df et sleep_df provenant de ds-silver
fact_health_df = health_df.join(sleep_df, "Person_ID") \
    .select("Person_ID", "Sleep_Duration", "Physical_Activity_Level", "Daily_Steps",
           "Stress_Level")

# Spécifier le chemin de stockage pour la table fact_health dans ds-gold
gold_fact_health_table_path = gold_path + "fact_health"

# Enregistrer le DataFrame fact_health au format Delta
fact_health_df.write.format("delta").mode("overwrite").save(gold_fact_health_table_path)

# Vérifier les données dans la table fact_health
gold_fact_health_df = spark.read.format("delta").load(gold_fact_health_table_path)
gold_fact_health_df.show(5)
```

```
► fact_health_df: pyspark.sql.dataframe.DataFrame = [Person_ID: string, Sleep_Duration: float ... 3 more fields]
► gold_fact_health_df: pyspark.sql.dataframe.DataFrame = [Person_ID: string, Sleep_Duration: float ... 3 more fields]
```

Person_ID	Sleep_Duration	Physical_Activity_Level	Daily_Steps	Stress_Level
135	7.3	60	8000	5
150	8.0	80	7500	3
252	6.8	30	6000	6
368	8.0	75	7000	3
1	6.1	42	4200	6

only showing top 5 rows

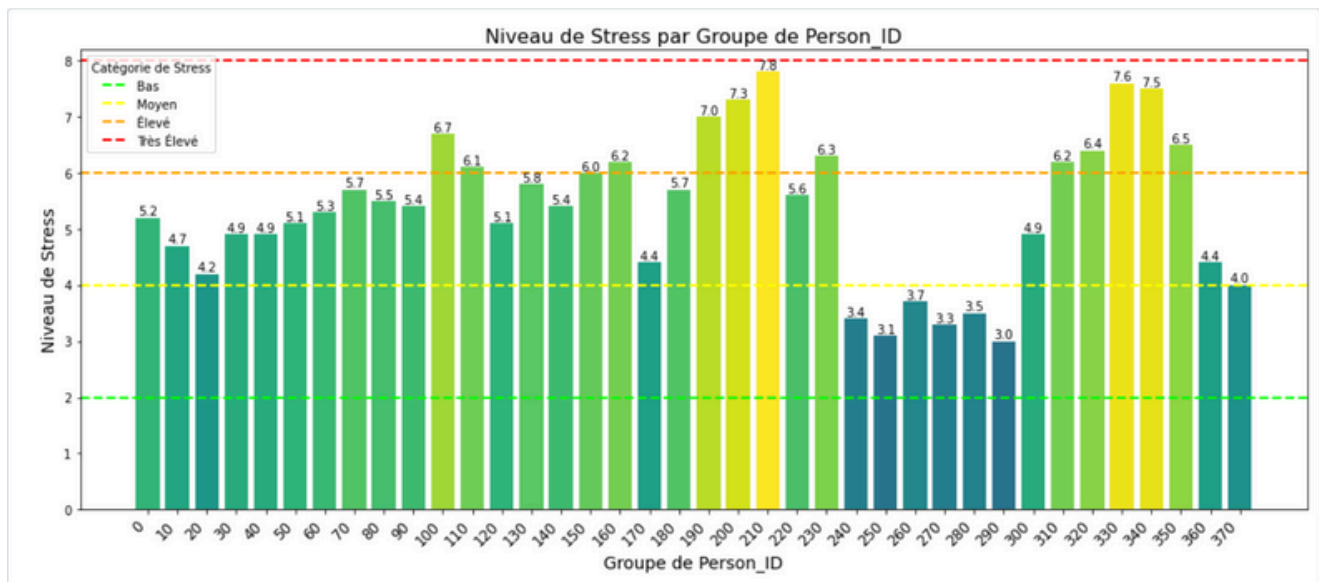
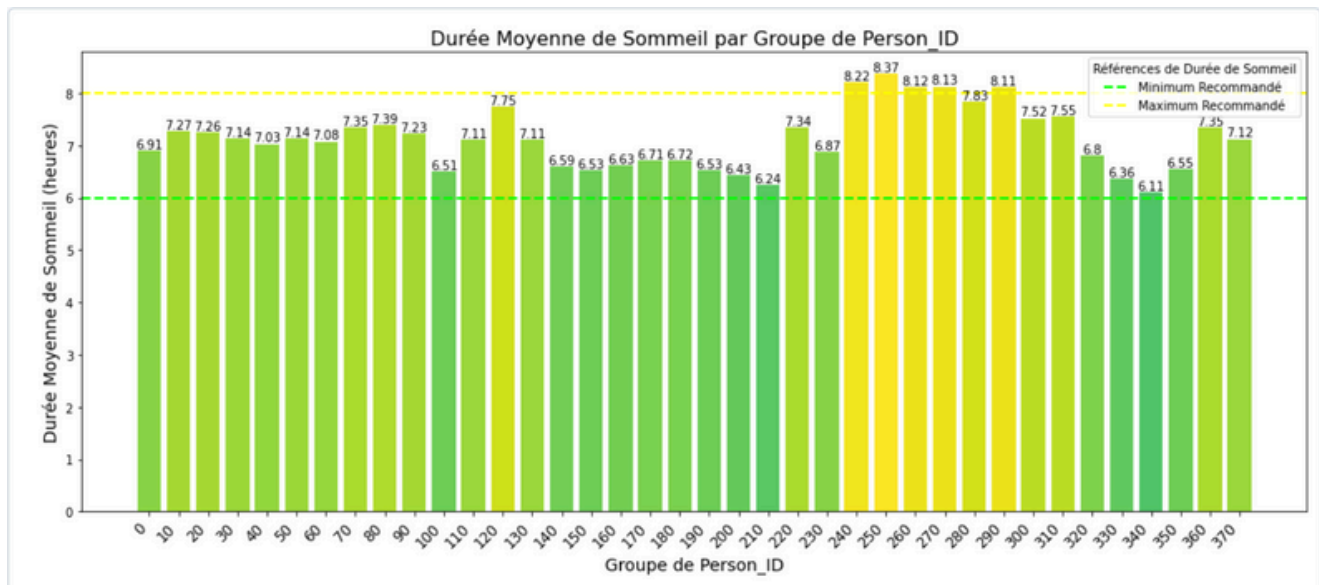
Analyse et Résultats

ANALYSE

Nous faisons différents calculs qui nous aideront pour la visualisation et l'établissement de résultats.

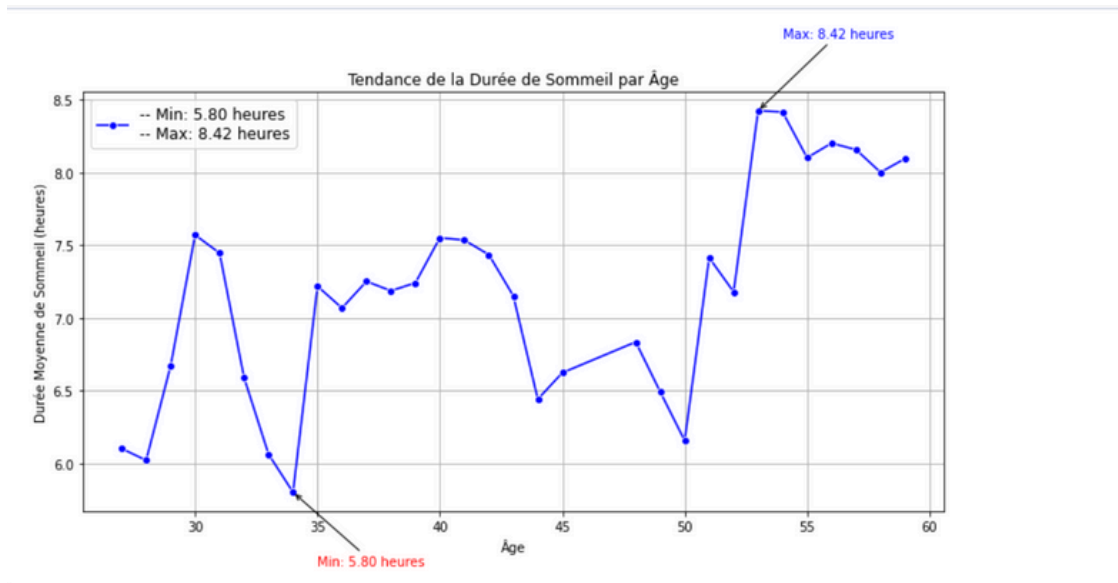
- *Sleep_Duration*: Durée moyenne du sommeil.
- *Physical_Activity_Level*: Niveau moyen d'activité.
- *Stress_Level*: La Distribution des niveaux de stress.

A partir de ces données on peut mettre en place des graphes sur ces 3 grands sujets, sommeil, activité physique et stress.



Analyse et Résultats

ANALYSE



RÉSULTATS

Avec des durées de sommeil variant entre **5,8 et 8,42 heures**, des niveaux d'activité allant de faibles (**<40**) à **très élevés (>75)**, et **des niveaux de stress dépassant parfois 8**, ces facteurs montrent des interactions complexes.

Une stratégie globale, combinant amélioration du sommeil, augmentation de l'activité physique, et réduction du stress, est essentielle pour améliorer la santé et le bien-être des individus.



Conclusion

CONCLUSION

Ces analyses mettent en évidence des interactions fortes entre le sommeil, l'activité physique, et le stress, qui sont des piliers essentiels du bien-être. Une approche holistique combinant des recommandations spécifiques pour améliorer le sommeil, encourager l'exercice, et réduire le stress peut transformer positivement la santé des individus.

L'intégration d'autres variables telles que l'alimentation, les habitudes numériques, ou les données biométriques avancées pourrait fournir une vue encore plus complète des facteurs influençant le bien-être.
A suivre :)

Source de nos deux notebooks:

<https://shorturl.at/GWZp7>

<https://shorturl.at/agK7B>

