



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, l'étude des interactions entre le sommeil, l'activité physique, l'occupation, le stress et d'autres facteurs liés au mode de vie est essentielle. Ce projet a pour objectif 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 tels que l'activité physique, l'activité professionnelle, les niveaux de stress et les caractéristiques démographiques comme l'âge et le sexe interagissent pour influencer la qualité et la durée du sommeil. En analysant ces données, nous espérons pouvoir répondre à des questions clés telles que :

Comment les groupes démographiques se reposent-ils ?

L'activité physique influence-t-elle la qualité du sommeil ?

Le niveau de stress affecte-t-il directement la durée ou la qualité du sommeil ?

Existe-t-il un lien entre la profession exercée et le niveau de stress, et par extension, 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

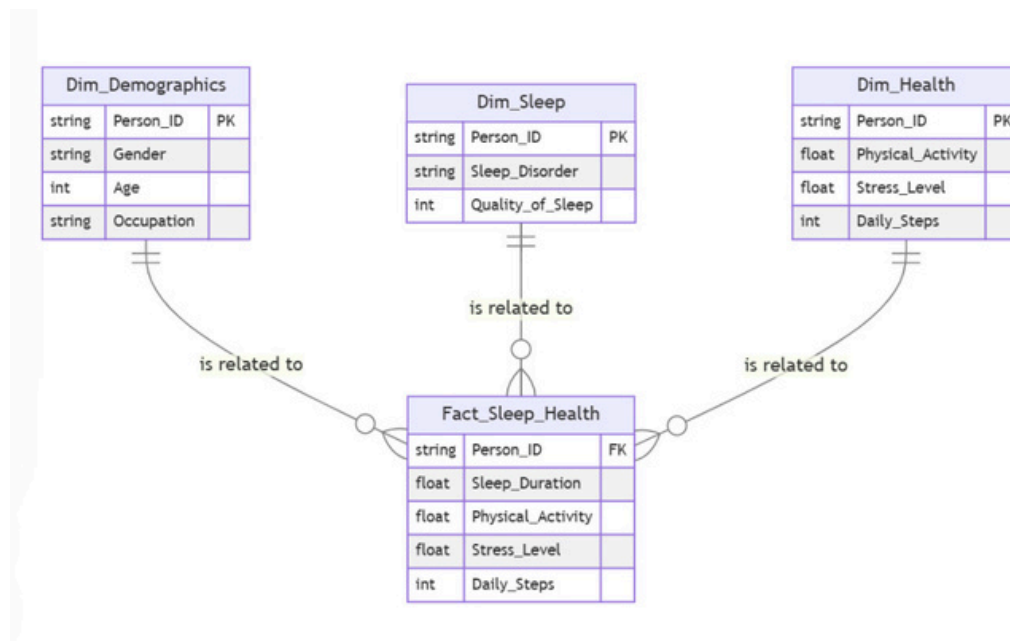
Dim_Sleep - Table de dimension du sommeil

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.

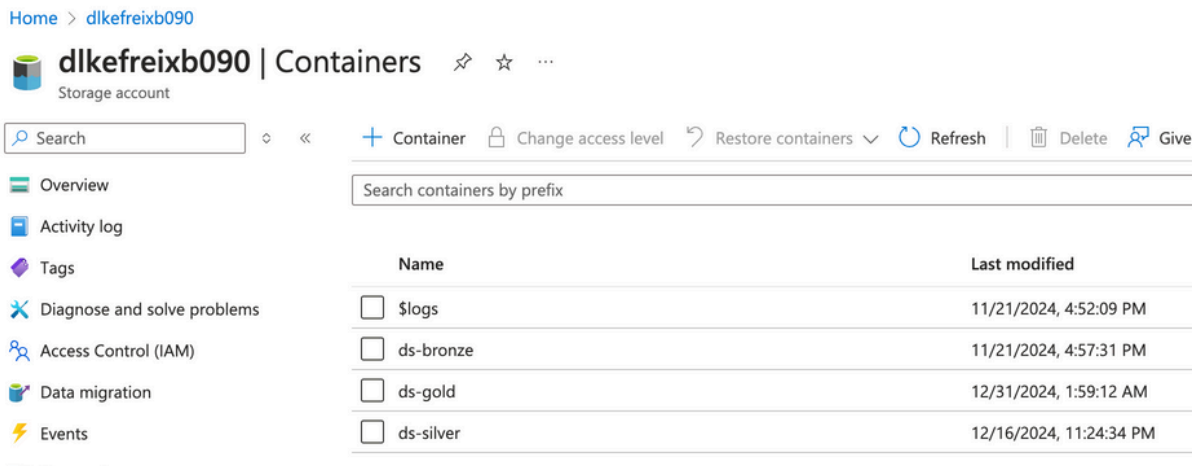


Modélisation : Une structure en étoile est conçue avec une table de faits (mesures) et des tables de dimensions (attributs descriptifs).

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 :

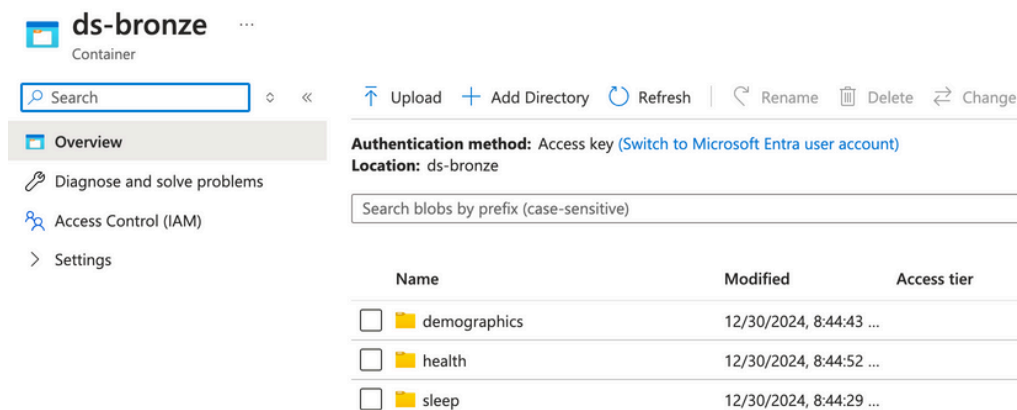


Aperçu des zone sur Azure

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.



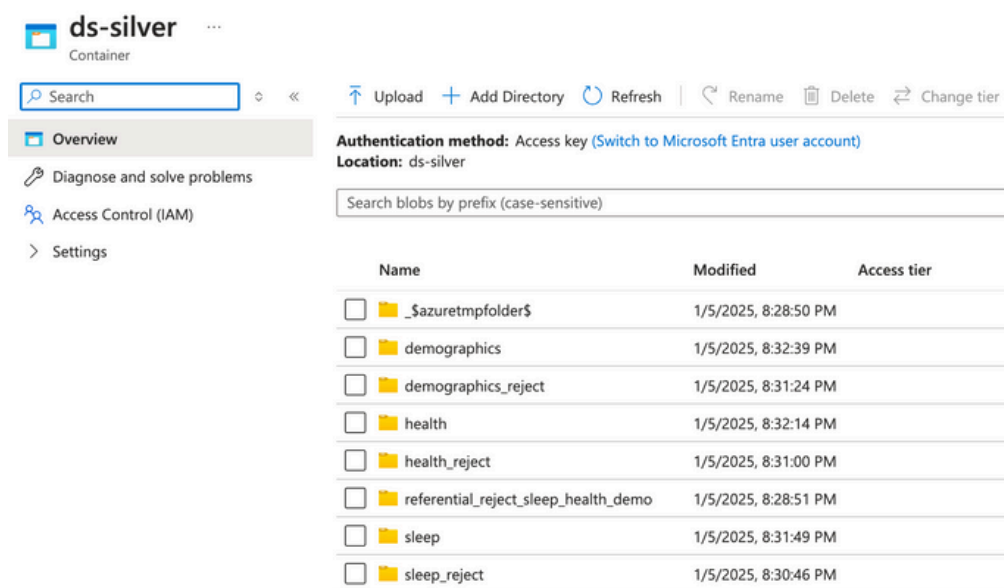
Aperçu de la zone de Transformation (ds-silver)

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.



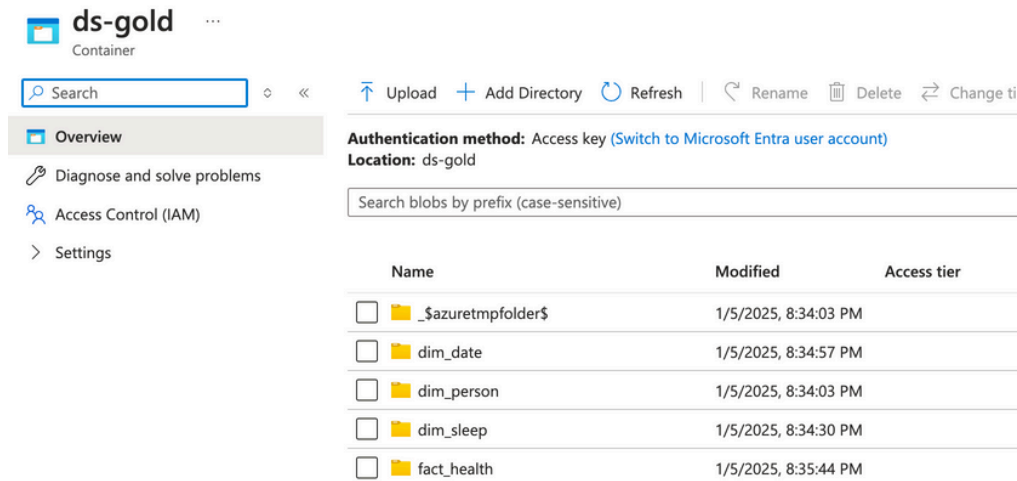
Aperçu de la zone de Transformation (ds-silver)

Outils et Architecture

3. Zone de Structuration (ds-gold) :

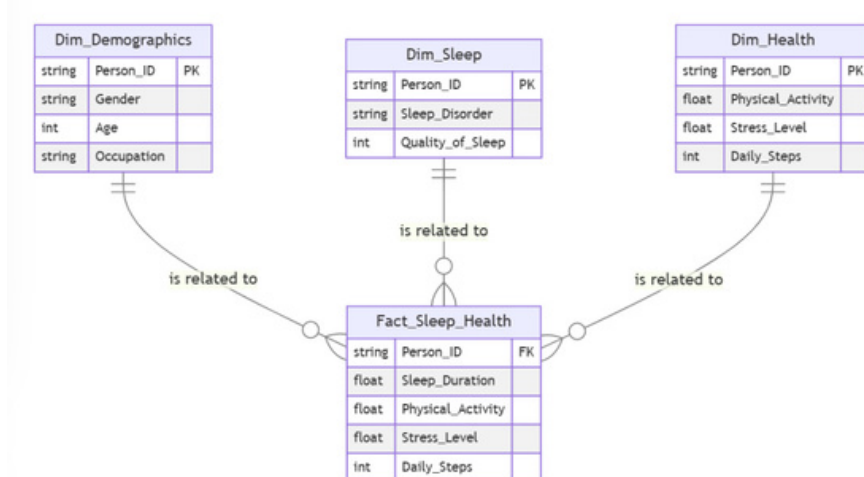
La zone de structuration est où les données transformées sont organisées et stockées de manière structurée, prêtes pour l'analyse. Les principales étapes et outils utilisés dans cette zone sont :

- **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).



Jointures pour Enrichir les Données

Outils et Architecture

LES OUTILS

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

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.



source: [Fichier:Microsoft-Azure.png — Wikipédia](#)



source: [Fichier:Logo Matplotlib.svg — Wikipédia](#)



databricks

source: [File:Databricks Logo.png - Wikipedia](#)

Traitement de la donnée

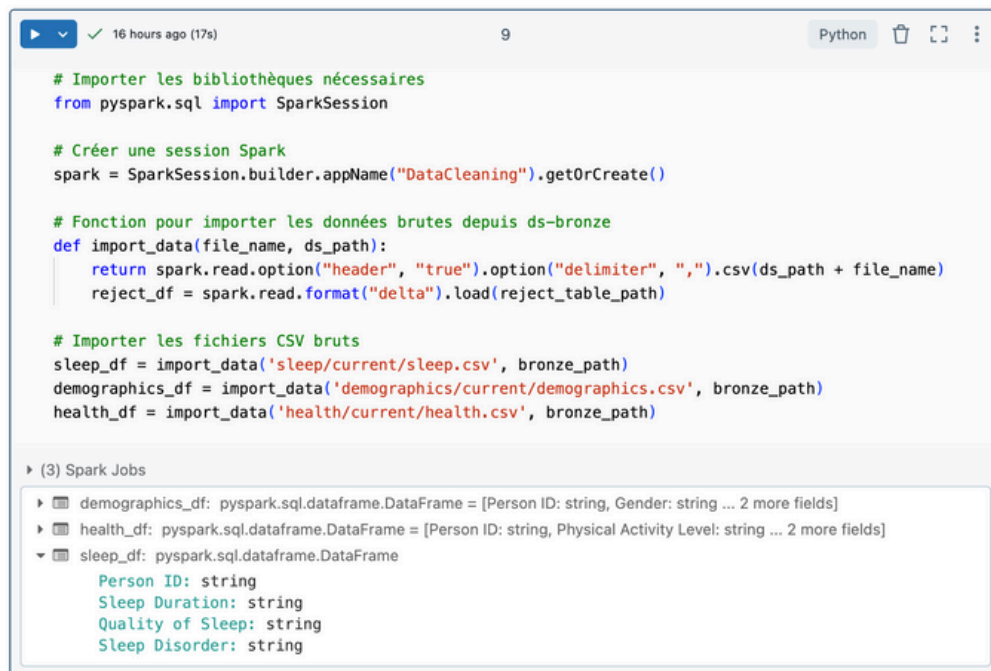
ÉTAPES D'INGESTION, DE TRANSFORMATION ET DE STRUCTURATION

Le traitement des données suit plusieurs étapes clés qui garantissent que les données brutes sont converties en un format structuré et prêt pour l'analyse. Voici une description détaillée des étapes suivies :

1. Ingestion des Données :

L'ingestion des données est la première étape du processus, où nous collectons les données brutes à partir de différentes sources. Dans notre projet, les sources principales sont des fichiers CSV contenant des informations sur la santé et le comportement des individus.

- **Chargement des Données** : Nous avons utilisé Apache Spark pour lire et charger les fichiers CSV dans des DataFrames PySpark. Spark est particulièrement adapté pour traiter de grandes quantités de données de manière distribuée et efficace.



```
# Importer les bibliothèques nécessaires
from pyspark.sql import SparkSession

# Créer une session Spark
spark = SparkSession.builder.appName("DataCleaning").getOrCreate()

# Fonction pour importer les données brutes depuis ds-bronze
def import_data(file_name, ds_path):
    return spark.read.option("header", "true").option("delimiter", ",").csv(ds_path + file_name)
    reject_df = spark.read.format("delta").load(reject_table_path)

# Importer les fichiers CSV bruts
sleep_df = import_data('sleep/current/sleep.csv', bronze_path)
demographics_df = import_data('demographics/current/demographics.csv', bronze_path)
health_df = import_data('health/current/health.csv', bronze_path)
```

▶ (3) Spark Jobs

- ▶ demographics_df: pyspark.sql.dataframe.DataFrame = [Person ID: string, Gender: string ... 2 more fields]
- ▶ health_df: pyspark.sql.dataframe.DataFrame = [Person ID: string, Physical Activity Level: string ... 2 more fields]
- ▼ sleep_df: pyspark.sql.dataframe.DataFrame
 - Person ID: string
 - Sleep Duration: string
 - Quality of Sleep: string
 - Sleep Disorder: string

Collecte des Données

Traitement de la donnée

2. Transformation des Données :

La transformation des données implique le nettoyage, l'enrichissement et la préparation des données pour l'analyse. Les principales étapes de cette phase comprennent :

- **Nettoyage des Données :**

- Traitement des valeurs manquantes : Suppression ou imputation des valeurs manquantes.
- Correction des anomalies : Identification et correction des valeurs aberrantes ou incohérentes.

```
✓ 16 hours ago (2m) 18

# Importer les bibliothèques nécessaires
from pyspark.sql.types import IntegerType, FloatType, StringType
from pyspark.sql.functions import col, lit, current_date, row_number, when

# Créer une vue temporaire pour les enregistrements à rejeter avec les colonnes supplémentaires
sleep_casted_df = sleep_df.withColumn("RowNumber", lit(1)) \
    .withColumn("Sleep_Duration", col("Sleep_Duration").cast(FloatType())) \
    .withColumn("Quality_of_Sleep", col("Quality_of_Sleep").cast(IntegerType())) \
    .withColumn("Sleep_Disorder", col("Sleep_Disorder").cast(StringType())) \
    .withColumn("IdCauseRejet", when(col("Person_ID").isNull(), 4000)
        .when(col("Sleep_Duration").isNull(), 1000)
        .when(col("Quality_of_Sleep").isNull(), 1001)
        .when(col("Sleep_Disorder").isNull(), 1002)
        .otherwise(None)) \
    .withColumn("InsertionDate", current_date())
```

Identification des Enregistrements à Rejeter

```
# Sélectionner les DataFrames de rejets
sleep_rejects_df = sleep_casted_df.filter(col("IdCauseRejet").isNotNull())
health_rejects_df = health_casted_df.filter(col("IdCauseRejet").isNotNull())
demographics_rejects_df = demographics_casted_df.filter(col("IdCauseRejet").isNotNull())

# Filtrer les enregistrements valides en utilisant une jointure anti
sleep_clean_df = sleep_casted_df.filter(col("IdCauseRejet").isNull()).select("Person_ID",
    "Sleep_Duration", "Quality_of_Sleep", "Sleep_Disorder")

health_clean_df = health_casted_df.filter(col("IdCauseRejet").isNull()).select("Person_ID",
    "Physical_Activity_Level", "Daily_Steps", "Stress_Level")

demographics_clean_df = demographics_casted_df.filter(col("IdCauseRejet").isNull()).select
    ("Person_ID", "Gender", "Age", "Occupation")

# Enregistrer les DataFrames rejetés au format Delta
sleep_rejects_df.write.format("delta").mode("overwrite").save(silver_path + "sleep_reject")
health_rejects_df.write.format("delta").mode("overwrite").save(silver_path + "health_reject")
demographics_rejects_df.write.format("delta").mode("overwrite").save(silver_path +
    "demographics_reject")

# Enregistrer les DataFrames nettoyés au format Delta
sleep_clean_df.write.format("delta").mode("overwrite").save(silver_path + "sleep")
health_clean_df.write.format("delta").mode("overwrite").save(silver_path + "health")
demographics_clean_df.write.format("delta").mode("overwrite").save(silver_path + "demographics")
```

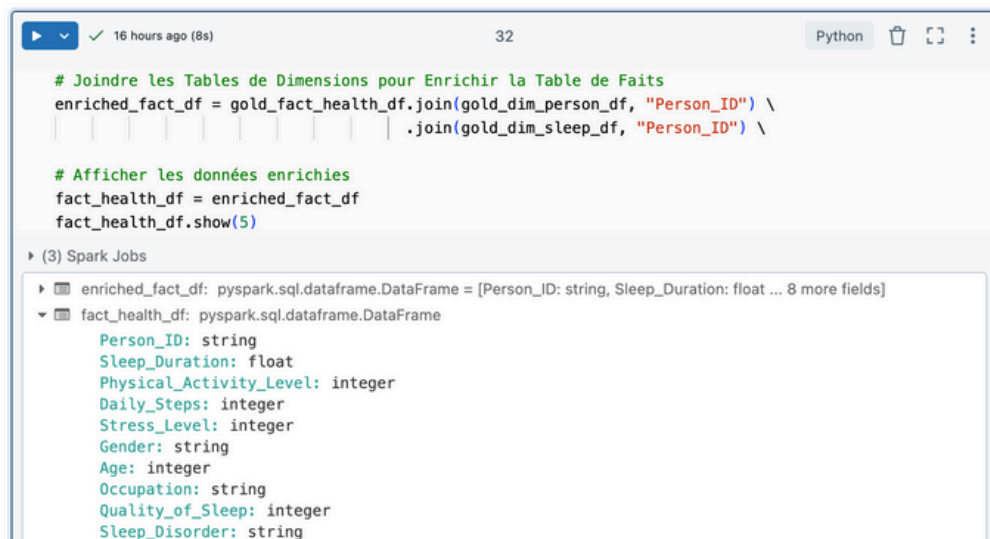
Gestion des Enregistrements à Rejeter

Traitement de la donnée

3. Transformation et Structuration des Données :

Cette étape consiste à enrichir et organiser les données transformées dans un format relationnel et à effectuer des jointures entre les différentes tables pour créer un DataFrame enrichi prêt pour l'analyse.

- **Création des Tables de dimensions et de faits** : Facilite les jointures et les analyses croisées.
- **Jointures des Tables** : Les tables de faits et de dimensions sont jointes sur la clé Person_ID pour créer un DataFrame enrichi.
- **Organisation des Données** : Les données sont organisées de manière à faciliter les analyses ultérieures, en structurant les variables d'intérêt et en garantissant leur intégrité et leur cohérence.



```
# Joindre les Tables de Dimensions pour Enrichir la Table de Faits
enriched_fact_df = gold_fact_health_df.join(gold_dim_person_df, "Person_ID") \
    .join(gold_dim_sleep_df, "Person_ID") \

# Afficher les données enrichies
fact_health_df = enriched_fact_df
fact_health_df.show(5)
```

▶ (3) Spark Jobs

- ▶ enriched_fact_df: pyspark.sql.dataframe.DataFrame = [Person_ID: string, Sleep_Duration: float ... 8 more fields]
- ▼ fact_health_df: pyspark.sql.dataframe.DataFrame
 - Person_ID: string
 - Sleep_Duration: float
 - Physical_Activity_Level: integer
 - Daily_Steps: integer
 - Stress_Level: integer
 - Gender: string
 - Age: integer
 - Occupation: string
 - Quality_of_Sleep: integer
 - Sleep_Disorder: string

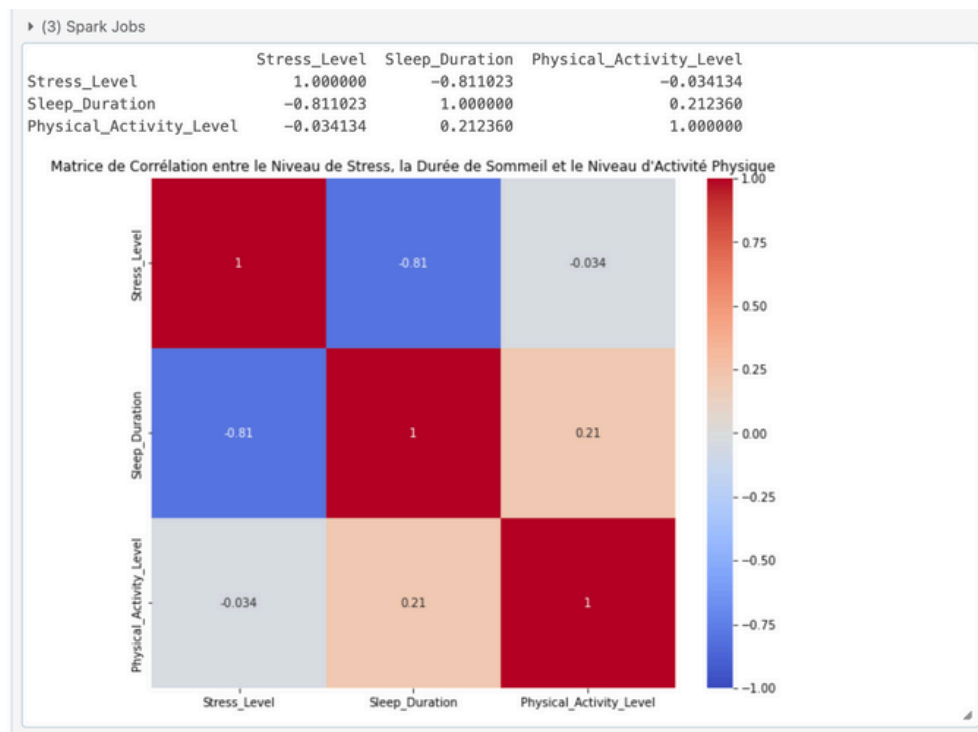
Jointures pour Enrichir les Données

Analyse et Résultats

RÉSULTATS CORRÉLATION ET VISUALISATION

L'analyse des données a révélé des insights précieux sur les relations entre le sommeil, l'activité physique et le stress des individus. Voici les principales conclusions de notre analyse de corrélation :

- **Corrélation entre le Niveau de Stress et la Durée de Sommeil :**
 - Une forte corrélation négative indique que les individus avec un niveau de stress plus élevé tendent à avoir une durée de sommeil plus courte.
- **Corrélation entre le Niveau de Stress et le Niveau d'Activité Physique :**
 - Une très faible corrélation négative suggère que l'activité physique n'a pratiquement aucune influence sur le niveau de stress.
- **Corrélation entre la Durée de Sommeil et le Niveau d'Activité Physique :**
 - Une faible corrélation positive montre que les individus avec un niveau d'activité physique plus élevé tendent légèrement à avoir une durée de sommeil plus longue.



Matrice de Corrélation

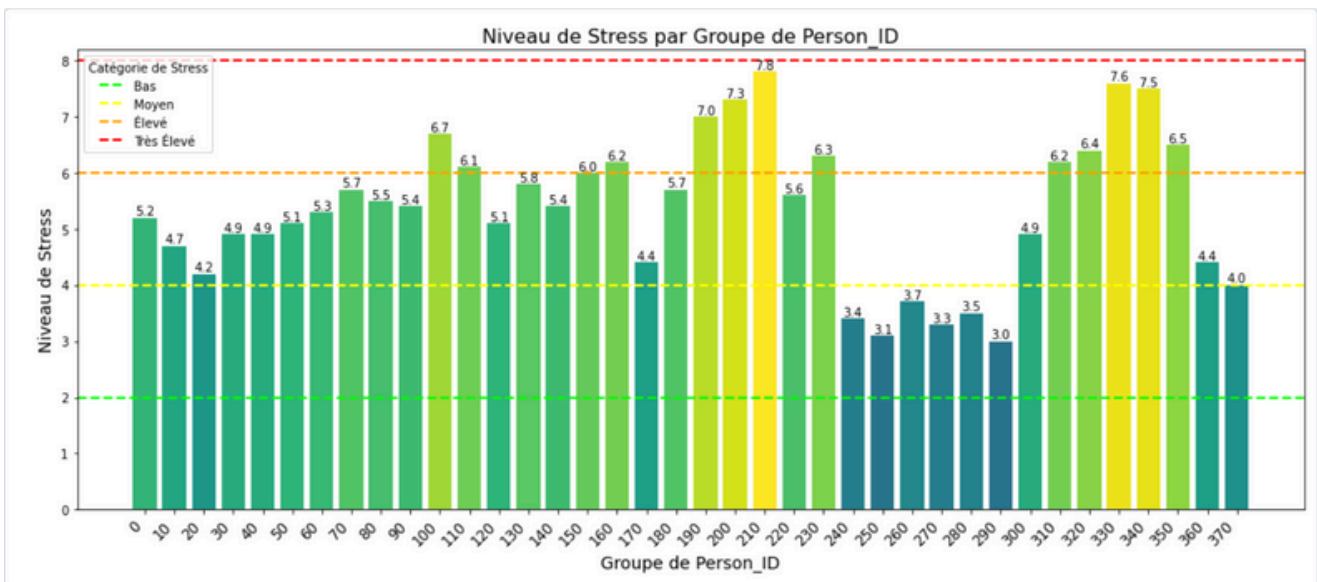
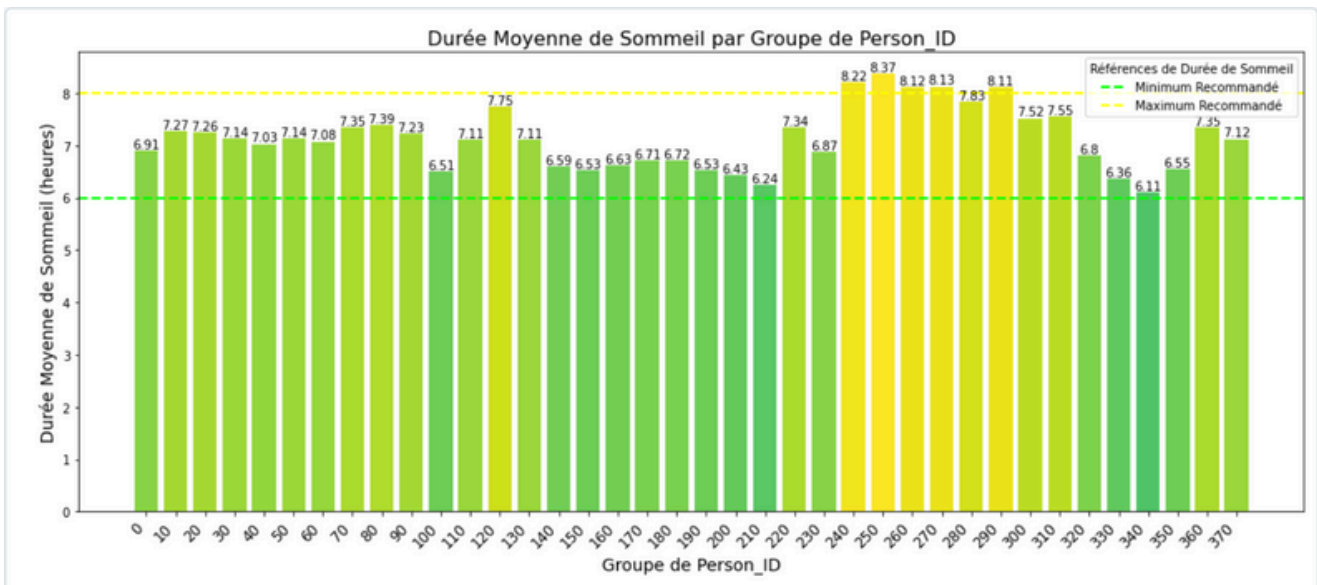
Analyse et Résultats

ANALYSE

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

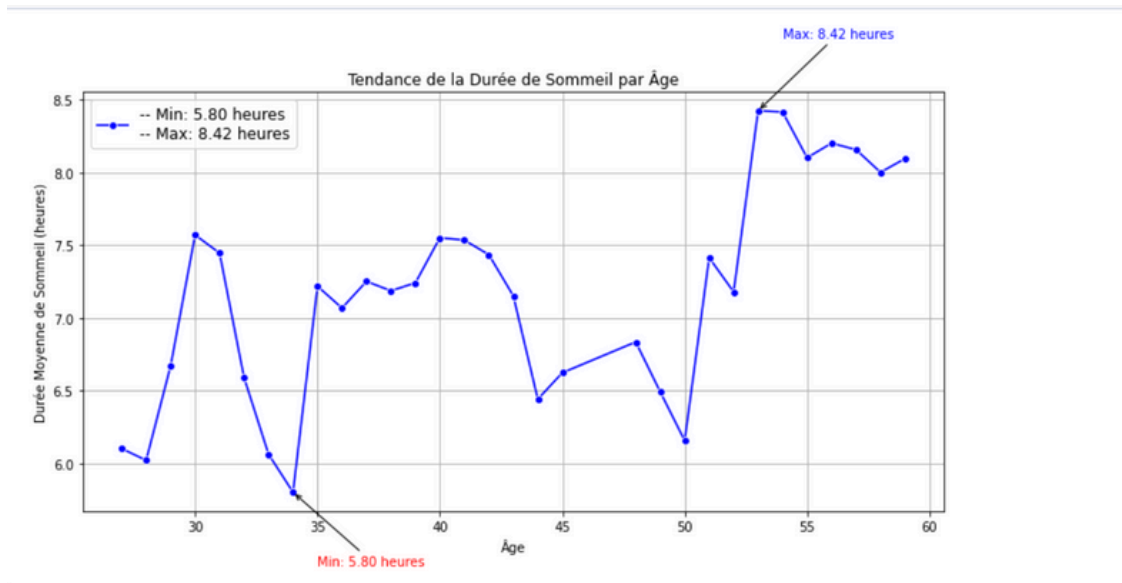
- *Sleep_Duration*: Durée moyenne du sommeil.
- *Physical_Activity_Level*: Niveau moyen d'activité.
- *Stress_Level*: La Distribution des niveaux de stress.

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.

Nous avons constaté que les individus avec des niveaux de stress plus élevés tendent à avoir une durée de sommeil plus courte, soulignant l'importance de la gestion du stress pour améliorer le sommeil. De plus, l'analyse a révélé que les niveaux d'activité physique ont une influence positive sur la durée de sommeil, bien que cette influence soit modérée. Ces insights permettent de formuler des recommandations spécifiques, comme encourager l'activité physique régulière et proposer des techniques de gestion du stress pour ceux qui en ont le plus besoin.



Conclusion

CONCLUSION

Nos analyses démontrent que nous avons atteint les objectifs fixés au début de ce projet. Nous avons une meilleure compréhension des relations entre les variables démographiques, de santé et de sommeil. Ces informations ont été exploitées pour formuler des recommandations spécifiques, telles que l'amélioration de la qualité du sommeil en fonction du groupe d'âge ou de la profession. Nous avons également établi une structure de données robuste permettant des analyses approfondies et des extensions futures, notamment l'intégration d'algorithmes d'apprentissage automatique pour prédire les troubles du sommeil. En comprenant mieux ces interactions, nous pouvons proposer des interventions et des recommandations personnalisées pour aider les individus à améliorer leur santé et leur qualité de vie.

Pour aller plus loin, il serait intéressant d'explorer comment d'autres facteurs, tels que l'alimentation et les habitudes de vie, peuvent également influencer le sommeil et le bien-être général.

Source de nos deux notebooks:

<https://shorturl.at/GWZp7>

<https://shorturl.at/agK7B>

