Projet Databricks: Transformation des Données en Zones Bronze, Silver et Gold

Ce document récapitule toutes les étapes effectuées, de l'importation des données jusqu'à leur analyse dans la zone Gold.

Contexte : Nous avons récupéré une base de données regroupant un certain nombre d'indicateurs de santé et de facteurs de risques liés aux maladies cardiaques sur un panel de 9500 personnes sur le lien suivant :

https://www.kaggle.com/datasets/oktayrdeki/heart-disease?resource=download

Ex:



1. Importation des données dans la zone Bronze

Objectif: Créer une table brute contenant les données initiales sans modification.

Requête SQL:

```
%sql
CREATE OR REPLACE TABLE default.BronzeHeartDisease AS
SELECT *
FROM default.heart_disease_csv;
```

- Action effectuée : Les données issues du fichier CSV ont été importées dans la table BronzeHeartDisease.
- **Objectif**: Stocker les données telles quelles, sans transformation, dans une table intermédiaire.

2. Transformation et nettoyage des données dans la zone Silver

Objectif : Supprimer les lignes contenant des cellules vides, convertir les colonnes numériques et enrichir les données.

2.1 Suppression des lignes contenant des cellules NULL

Requête SQL:

```
CREATE OR REPLACE TABLE default.SilverHeartDisease AS
FROM default.BronzeHeartDisease
   Age IS NOT NULL
   AND Sexe IS NOT NULL
   AND Tension_arterielle IS NOT NULL
   AND Taux_de_cholesterol IS NOT NULL
   AND Habitudes_sportives IS NOT NULL
   AND Tabagisme IS NOT NULL
   AND Antecedents_Cardiaques IS NOT NULL
   AND Diabete IS NOT NULL
   AND IMC IS NOT NULL
   AND Hypertension IS NOT NULL
   AND Faible cholesterol HDL IS NOT NULL
    AND Cholesterol_LDL_eleve IS NOT NULL
   AND `Consommation_d'alcool` IS NOT NULL
   AND Niveau de stress IS NOT NULL
    AND Heures_de_sommeil IS NOT NULL
   AND Consommation de sucre IS NOT NULL
   AND Taux_de_triglycerides IS NOT NULL
    AND Glycemie_a_jeun IS NOT NULL
    AND Niveau_de_CRP IS NOT NULL
    AND Niveau_homocysteine IS NOT NULL
   AND `Maladie_cardiaque_?` IS NOT NULL;
```

• Action effectuée: Toutes les lignes ayant au moins une cellule NULL ont été supprimées.

2.2 Conversion des colonnes à plusieurs décimales

Requête SQL:

```
CREATE OR REPLACE TABLE default.SilverHeartDisease AS
   Age,
   Sexe,
   Tension arterielle,
   Taux_de_cholesterol,
   Habitudes_sportives,
   Tabagisme,
   Antecedents Cardiaques,
   Diabete,
   ROUND(CAST(REPLACE(IMC, ',', '.') AS DOUBLE), 2) AS IMC_Cleaned,
   Hypertension,
   Faible_cholesterol_HDL,
   Cholesterol_LDL_eleve,
    `Consommation_d'alcool`,
   Niveau de stress,
   ROUND(CAST(REPLACE(Heures_de_sommeil, ',', '.') AS DOUBLE), 2) AS Heures_de_sommeil_Cleaned,
   Consommation_de_sucre,
   ROUND(CAST(REPLACE(Taux_de_triglycerides, ',', '.') AS DOUBLE), 2) AS Taux_de_triglycerides_Cleaned,
   ROUND(CAST(REPLACE(Glycemie_a_jeun, ',', '.') AS DOUBLE), 2) AS Glycemie_a_jeun_Cleaned,
   ROUND(CAST(REPLACE(Niveau_de_CRP, ',', '.') AS DOUBLE), 2) AS Niveau_de_CRP_Cleaned,
   ROUND(CAST(REPLACE(Niveau_homocysteine, ',', '.') AS DOUBLE), 2) AS Niveau_homocysteine_Cleaned,
    `Maladie cardiaque ?`
FROM default.BronzeHeartDisease
   Age IS NOT NULL
   AND Sexe IS NOT NULL
   AND Tension_arterielle IS NOT NULL
   AND Taux_de_cholesterol IS NOT NULL
   AND Habitudes_sportives IS NOT NULL
   AND Tabagisme IS NOT NULL
   AND Antecedents_Cardiaques IS NOT NULL
   AND Diabete IS NOT NULL
   AND IMC IS NOT NULL
   AND Hypertension IS NOT NULL
   AND Faible_cholesterol_HDL IS NOT NULL
   AND Cholesterol_LDL_eleve IS NOT NULL
   AND `Consommation_d'alcool` IS NOT NULL
   AND Niveau de stress IS NOT NULL
   AND Heures_de_sommeil IS NOT NULL
   AND Consommation_de_sucre IS NOT NULL
   AND Taux_de_triglycerides IS NOT NULL
   AND Glycemie_a_jeun IS NOT NULL
   AND Niveau_de_CRP IS NOT NULL
   AND Niveau_homocysteine IS NOT NULL
   AND `Maladie_cardiaque_?` IS NOT NULL;
```

- Action effectuée: Toutes les colonnes contenant des valeurs avec plusieurs décimales ont été converties en format à deux décimales.
- Astuce technique: Les virgules ont été remplacées par des points avant la conversion.

2.3 Enrichissement des données : Ajout de colonnes calculées

Requête SQL:

```
%sql
CREATE OR REPLACE TABLE default.SilverHeartDisease AS
SELECT

*,
CASE

WHEN IMC_Cleaned < 18.5 THEN 'Maigreur'
WHEN IMC_Cleaned >= 18.5 AND IMC_Cleaned < 25 THEN 'Normal'
WHEN IMC_Cleaned >= 25 AND IMC_Cleaned < 30 THEN 'Surpoids'
ELSE 'Obésité'
END AS IMC_Category,
CASE

WHEN IMC_Cleaned < 18.5 THEN 'Risque élevé'
WHEN IMC_Cleaned >= 18.5 AND IMC_Cleaned < 25 THEN 'Risque faible'
WHEN IMC_Cleaned >= 25 AND IMC_Cleaned < 30 THEN 'Risque modéré'
ELSE 'Risque élevé'
END AS Risk_Score
FROM default.SilverHeartDisease;</pre>
```

Ajout des colonnes :

- o IMC_Category: Catégorisation de l'IMC (Maigreur, Normal, Surpoids, Obésité).
- o **Risk_Score :** Calcul du risque associé à l'IMC.

3. Transition vers la zone Gold

Objectif: Réutiliser les données Silver sans modifications supplémentaires.

Requête SQL:

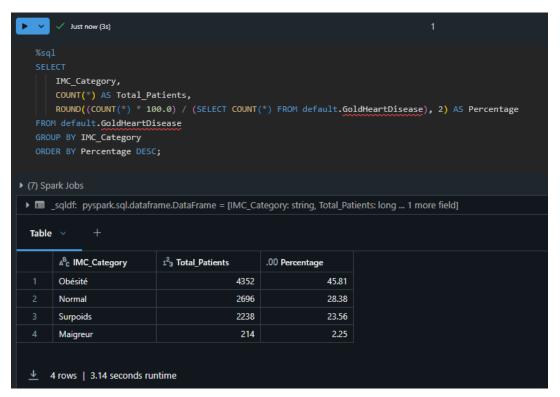
```
%sql
CREATE OR REPLACE TABLE default.GoldHeartDisease AS
SELECT *
FROM default.SilverHeartDisease;
```

- Action effectuée : Les données Silver ont été copiées dans la table Gold sans transformation.
- Objectif: La zone Gold représente les données finales, prêtes pour l'analyse.

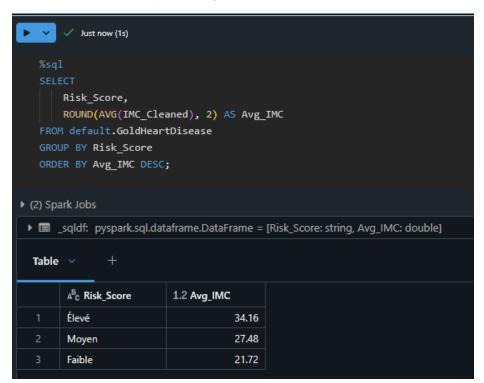
4. Analyses et requêtes sur la zone Gold

Objectif: Extraire des statistiques et tendances à partir des données Gold.

4.1 Pourcentage de patients par catégorie IMC



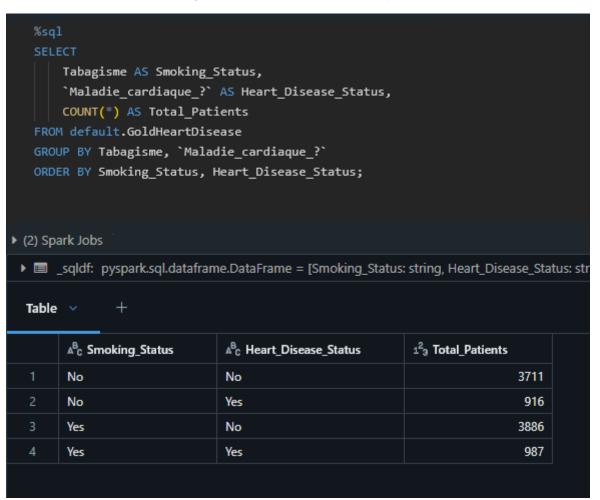
4.2 Moyenne de l'IMC par catégorie de risque



4.3 Pourcentage de patients atteints de maladies cardiaques

```
`Maladie_cardiaque_?` AS Heart_Disease_Status,
      COUNT(*) AS Total_Patients,
     ROUND((COUNT(*) * 100.0) / (SELECT COUNT(*) FROM default.GoldHeartDisease), 2) AS Percentage
  FROM default.GoldHeartDisease
  GROUP BY `Maladie_cardiaque_?`
  ORDER BY Percentage DESC;
(6) Spark Jobs
🕨 🔳 _sqldf: pyspark.sql.dataframe.DataFrame = [Heart_Disease_Status: string, Total_Patients: long ... 1 more field]
Table 🔻
       ABC Heart_Disease_Status
                                    123 Total_Patients
                                                          .00 Percentage
                                                    7597
                                                                        79.97
       Yes
                                                    1903
                                                                        20.03
```

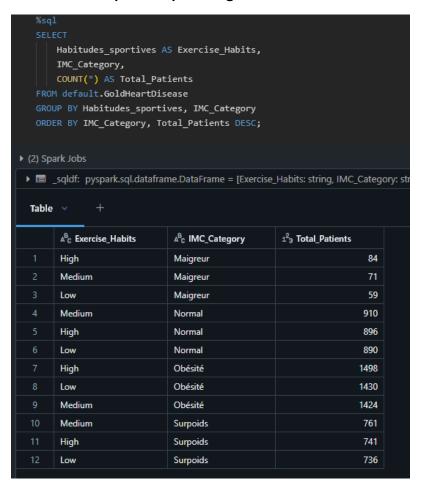
4.4 Corrélation entre le tabagisme et les maladies cardiaques



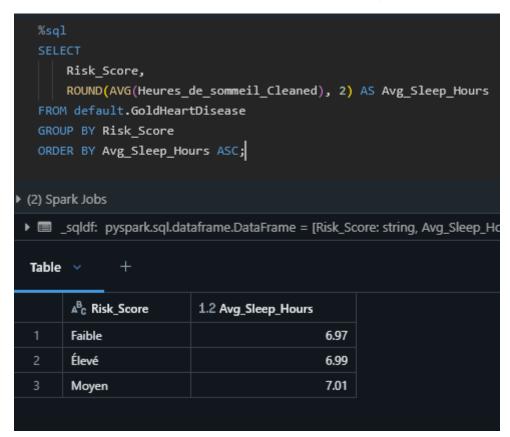
4.5 Pourcentage de patients ayant un stress élevé

```
Niveau_de_stress AS Stress_Level,
       COUNT(*) AS Total_Patients,
       ROUND((COUNT(*) * 100.0) / (SELECT COUNT(*) FROM default.GoldHeartDisease), 2) AS Percentage
   FROM default.GoldHeartDisease
   GROUP BY Niveau_de_stress
   ORDER BY Percentage DESC;
▶ (6) Spark Jobs
 ▶ ■ _sqldf: pyspark.sql.dataframe.DataFrame = [Stress_Level: string, Total_Patients: long ... 1 more field]
 Table v
        ABC Stress_Level
                             123 Total_Patients
                                                    .00 Percentage
        Medium
                                             3222
                                                                  33.92
        Low
                                             3158
                                                                  33.24
                                             3120
                                                                  32.84
        High
```

4.6 Habitudes sportives par catégorie IMC



4.7 Moyenne des heures de sommeil en fonction du risque



5. Conclusion

Ce projet illustre un pipeline complet de traitement des données depuis leur importation brute jusqu'à leur préparation pour l'analyse. Les zones Bronze, Silver et Gold ont permis de structurer le processus :

- Bronze: Conservation des données brutes pour assurer leur traçabilité.
- **Silver :** Nettoyage, normalisation et enrichissement des données pour les rendre exploitables.
- Gold: Présentation des données finales prêtes pour des analyses avancées.

Les analyses effectuées dans la zone Gold, telles que les statistiques sur l'IMC ou les risques de maladies cardiaques, offrent des perspectives précieuses pour comprendre les tendances dans les données. Ce pipeline peut être adapté à d'autres contextes pour gérer efficacement des données complexes.