TP2: Gestion et visualisation des valeurs manquantes avec PySpark et Missingno

Objectif

Explorer, traiter et visualiser les valeurs manquantes d'un jeu de données de films (Netflix) à l'aide de PySpark et de la bibliothèque Missingno.

Étapes

1. Prise en main: Préparation des données

- Récupérer les données sur Kaggle : https://www.kaggle.com/datasets/ariyoomotade/netflix-data-cleaning-analysis-and-visualization
- Identifier et répertorier les valeurs considérées comme « vides » ou non significatives (ex: "Not Given", "N/A").
- Remplacer ces valeurs par `null` afin de normaliser les données manquantes.

2. Visualisation des données manquantes

- Convertir temporairement le DataFrame PySpark en DataFrame Pandas (`.toPandas()`).*Attention :* Cette opération peut être coûteuse en mémoire, utilisez-la sur un échantillon ou un DataFrame restreint.

- Utiliser Missingno (`msno.bar()`) pour visualiser la répartition des valeurs manquantes.
    - Interpréter le graphique pour identifier les colonnes ayant le plus de valeurs manquantes.

3. Nettoyage et réinspection

- Appliquer un filtre pour supprimer les lignes manquantes dans certaines colonnes clés (par exemple `director` ou `country`) à l'aide de `dropna()`.
- Générer un nouveau graphique Missingno sur le DataFrame ainsi nettoyé pour comparer et constater la réduction des données manquantes.
    - Analyser si le filtrage a amélioré la qualité globale du dataset.

4. Analyses complémentaires

- Réfléchir à d’autres approches de traitement des données manquantes (imputation, suppression ciblée, etc.).
- Éventuellement, intégrer ce nettoyage à la suite des transformations du TP1 (calculs de moyennes, filtrages par date ou note) afin de travailler sur un dataset plus propre.

Remarque :
Ce TP met l’accent sur la compréhension et la visualisation des données manquantes. Adaptez les opérations de nettoyage (remplacement, suppression, imputation) aux contraintes de votre projet et à la nature de vos données.