epsi®
l'école d'ingénierie informatique

TD 2 : qualité de la donnée

Exercice 1 : vérification et correction de données

Contexte de l'exercice

Vous avez reçu une extraction brute du catalogue de produits d'un site e-commerce. Le fichier products.csv est connu pour contenir de nombreuses erreurs dues à des saisies manuelles et à des fusions de systèmes antérieures. Votre mission est de nettoyer ce catalogue pour le rendre utilisable par le système d'analyse des ventes.

Dataset : products.csv

En utilisant **Python** et/ou **SQL**, effectuez les tâches suivantes pour nettoyer le dataset :

1. Chargement et exploration: Chargez le fichier *products.csv* et affichez les premières lignes, les types de données et un résumé statistique pour identifier les potentiels problèmes.
2. Unicité : Identifiez et supprimez les doublons exacts.
3. Complétude : Gérez les valeurs manquantes :
    - Pour *stock_quantity*, remplacez les valeurs manquantes par 0 (en supposant qu'un stock non renseigné signifie "pas de stock").
    - Pour *price*, les produits sans prix ne peuvent être vendus. Supprimez les lignes où le prix est manquant.
4. Format et Type : La colonne *price* est au format chaîne de caractères (string) avec des symboles "€" et des virgules. Nettoyez-la et convertissez-la en un type numérique.
5. Validité : La colonne *stock_quantity* contient des valeurs invalides (négatives). Corrigez-les en les remplaçant par 0 (un stock ne peut être négatif). Convertissez ensuite la colonne en entier.
6. Cohérence : La colonne *category* a des problèmes de casse et d'accents (ex: "Électronique" vs "electronique"). Standardisez cette colonne en mettant tout en minuscules et en retirant les accents.

Fraicheur : Identifiez les produits "périmés" (ceux dont la date_added est antérieure à 2024) et affichez un avertissement ou un rapport sur ces produits.

Final : Affichez les informations (.info()) et les 5 premières lignes du DataFrame nettoyé.

Exercice 2 : REGEX

Contexte de l'exercice

Vous êtes Data Scientist dans une startup qui vient d'agréger les données de plusieurs CRM obsolètes. La base de données clients est chaotique. Votre mission est de nettoyer et de standardiser cette base pour la rendre exploitable par les équipes marketing.

1. Identifier les emails KO avec une expression régulière que vous allez définir
2. Créer une colonne email_validity_flag qui indique si l'email est correct ou non
3. Transformer la colonne 'Nom_Complet' pour qu'elle respecte le format standard "Prénom Nom" (ex: "Alice Dupont"), tout en gérant les noms composés (ex: "François-Pierre").
    a. Traiter les séparateurs : ',', '.', ...
    b. Traiter la casse : majuscule, minuscule, afin de mettre au format nom propre les noms
    c. Supprimer les espaces en trop
    d. Créer une colonne pour vérifier la validité du format du champ
4. Standardiser les numéros de téléphone
    a. Supprimer tous les caractères non numériques
    b. Traiter les différents cas : +33 ou 0
    c. Formatter le numéro pour qu'il s'affiche comme suit : 0X XX XX XX XX
    d. Afficher 'Numéro invalide' si le numéro ne contient pas 10 caractères