Exercice – Préparation et nettoyage des données (en Python)

1. Découpage des données

Séparer le jeu de données en **Train / Validation / Test** (par exemple : 60% / 20% / 20%).

2. Typage

Convertir les colonnes numériques importées en texte en types numériques appropriés :
Montant_raw, RevenuMensuel_raw, Dette_raw, etc.

3. Unités / devises

- Nettoyer la colonne Montant_raw.
- Convertir tous les montants en **HTG** (taux : **1 USD = 130 HTG**).
- Gérer différents formats, par exemple :
    - '1,200.50'
    - '20k'
    - '500 USD'
    - '12.5OUSD'

4. Dates

- Parser la colonne DateTransaction_raw (formats multiples + valeurs invalides).
- Créer des variables temporelles :
    - jour_semaine
    - mois
    - heure (si l'heure peut être reconstruite)

5. Doublons

Détecter et supprimer les lignes dupliquées (certaines sont **exactement identiques**).

6. Valeurs manquantes

- Traiter les valeurs manquantes (NaN, 'N/A', 'Unknown').
- Ajouter des indicateurs binaires si pertinent (ex. : revenu manquant).

7. Variables catégorielles

- Normaliser Ville_raw (variantes PAP, P-au-P, etc.).
- Appliquer un **One-Hot Encoding** pour :

    - Canal
    - Device
    - Ville_norm
- Appliquer un **Ordinal Encoding** pour NiveauEtude.

8. Valeurs aberrantes (Outliers)

Identifier et traiter les valeurs aberrantes, par exemple :

- Age ∈ [-4, 150]
- NbTrans_24h ∈ [120, 250]
- Montants très élevés

Tester différentes approches :

- Clipping
- log1p
- RobustScaler

9. Mise à l'échelle

Standardiser ou normaliser les variables numériques pour les modèles sensibles aux distances ou au gradient :

- Régression logistique
- KNN
- SVM

10. Data leakage

Respecter la règle suivante :

- **Fit** des transformations uniquement sur X_train
- **Transform** ensuite sur X_validation et X_test