TP – Prétraitement des données (Dataset bruité – Achats alimentaires)

Objectif : Nettoyer et préparer un jeu de données réaliste contenant des bruits (valeurs manquantes, doublons, fautes de frappe, catégories incohérentes, dates hétérogènes, espaces parasites, valeurs aberrantes, colonnes inutiles) afin de le rendre exploitable pour l'analyse et le data mining.

Jeu de données fourni

Fichiers : donnees_brutes_achats.xlsx

Schéma : TransactionID, Produit, Quantité, Prix, Catégorie, Date, Notes (Notes est à considérer comme colonne potentiellement inutile).

I. Travail demandé (étapes)

1. Importer les données dans Python (pandas).
2. Diagnostiquer la qualité des données : info, types, NA, doublons, valeurs uniques.
3. Gérer les doublons (exacts et par TransactionID).
4. Uniformiser les chaînes : trim espaces, casse cohérente, accents ; harmoniser les synonymes (pâtes/pates, tomate/tomato, café/cafe).
5. Harmoniser les catégories (Ex: 'épicerie', 'laitage', 'fruits & légumes'...).
6. Unifier le format des dates (YYYY-MM-DD) avec *pd.to_datetime* (dayfirst=True si besoin).
7. Gérer les valeurs manquantes (Quantité/Prix) : imputation justifiée ou suppression raisonnée.
8. Détecter/traiter les aberrants (quantités négatives/excessives, prix 999, lignes non pertinentes ").
9. Supprimer la colonne 'Notes' une fois l'analyse terminée si non utilisée.
10. Exporter une table finale propre (xlsx) et rédiger un court rapport des décisions.

Indications Python (pandas)

- *import pandas as pd*
- *df.info();*
- *df.describe(include='all')*
- *df.duplicated().sum();*
- *df.drop_duplicates(inplace=True)*
- *df['Produit'] = df['Produit'].str.strip()*
- *df['Produit_normalise'] = (df['Produit'].str.lower().str.replace('é','e').str.replace('è','e').str.replace('ê','e').str.replace('à','a').str.replace('â','a').str.replace('î','i').str.replace('ï','i').str.replace('ô','o').str.replace('ö','o').str.replace('û','u').str.replace('ü','u'))*

- *mapping_produits = {'pates':'pâtes','the':'thé','cafe':'café','tomato':'tomate','yaourts':'yaourt','riz ':'riz','frommage':'fromage','huile olive':'huile d\'olive','corn flakes':'corn-flakes'} # exemple*
- *df['Produit'] = df['Produit'].str.strip().str.lower().replace(mapping_produits)*
- *df['Catégorie'] = df['Catégorie'].str.strip().str.lower().replace({'epicerie':'épicerie','boisson':'boissons','fruits-legumes':'fruits & légumes','fruits/legumes':'fruits & légumes','laitage':'laitage','cremerie':'crèmerie','oeufs':'œufs & ovoproduits'})*
- *df['Date'] = pd.to_datetime(df['Date'], dayfirst=True, errors='coerce')*
- *df.loc[df['Quantité'] < 0, 'Quantité'] = pd.NA*
- *df.loc[df['Quantité'] > 100, 'Quantité'] = pd.NA # exemple de seuil*
- *df.loc[df['Prix'] > 100, 'Prix'] = pd.NA*
- *df = df[df['Produit'] != '—']*
- *df = df.drop(columns=['Notes'])*

Livrables attendus

1. Fichier propre : donnees_achats_propre.csv/ xlsx
2. Rapport court (1 page max) résumant : problèmes détectés, choix de nettoyage/harmonisation, méthode d'imputation, impact (nb de lignes/valeurs modifiées).

Barème indicatif (20 pts)

Diagnostic de qualité des données :	4 pts
Doublons & chaînes uniformisées :	4 pts
Dates & manquants traités :	4 pts
Détection/traitement des aberrants :	4 pts
Table finale propre + rapport :	4 pts