MINI-PROJET PANDAS

A remettre le 06/12/2026

Vous êtes Data Analyst pour EduMart, une boutique (web/app/boutique) qui vend : Cours, livres, logiciels, crédits cloud, laptops et accessoires. Vous avez à votre disposition trois Datasets au format csv contenus dans le fichier .zip nommé "pandas_mini_projet_dataset.zip".

Votre **objectif** est de : nettoyer et analyser les ventes + comportement client (retours, avis, ...) avec **Pandas**.

Vous allez pratiquer :

1. Import / inspection : `read_csv`, `info`, `describe`, `value_counts`
2. Qualité des données : `isna`, `duplicated`, `astype`, `to_datetime`, `to_numeric`
3. Nettoyage : traitement des NA, correction types, filtrage anomalies
4. Agrégation : `groupby`, `agg`, `pivot_table`
5. Jointures : `merge` (clients + produits + ventes)
6. Export : `to_csv` (données nettoyées / tables de résultats)

Description des fichiers et colonnes

`customers.csv`

- `customer_id` (clé)
- `age` (quelques valeurs `"unknown"` pour forcer le nettoyage)
- `gender` (valeurs manquantes)
- `city` (valeurs manquantes)
- `segment` (Étudiant/Professionnel/Entreprise/Indépendant)
- `signup_date`

`products.csv`

- `product_id` (clé)
- `category` (Cours/Livre/Laptop/Accessoire/Cloud/Logiciel)
- `brand`, `product_name`
- `unit_price`

`order_lines.csv`

- `order_id`, `customer_id`, `product_id`, `order_date`
- `quantity` (inclut quelques quantités négatives = anomalies)

- `discount_pct` (quelques valeurs stockées en texte)
- `gross_amount`, `net_amount`
- `payment_method`, `channel`, `marketing_source`
- `delivery_days` (quelques NaN + quelques valeurs extrêmes)
- `returned` (0/1), `review_score` (NaN possibles)
- `segment`, `city`, `category` (présents pour faciliter certaines analyses, mais vous pouvez aussi les recalculer via `merge`)

Consignes du mini-projet

Étape 1 — Charger et comprendre

1. Charger les 3 CSV.
2. Afficher `head()`, tailles, `info()` et `describe(include="all")`.
3. Lister les colonnes à problèmes (types incohérents, NA, valeurs suspectes).

Livrable : un mini "rapport d'audit" (5–10 lignes) sur la qualité des données.

Étape 2 — Nettoyage

1. **Types**
   - `signup_date`, `order_date` => datetime
   - `discount_pct` => numérique (attention aux valeurs texte)
   - `age` => numérique (gérer `"unknown"`)
2. **Valeurs manquantes**
   - Proposer une stratégie
3. **Doublons**
   - Détecter et décider : supprimer ou garder (justifier)
4. **Anomalies**
   - `quantity <= 0`
   - `delivery_days` très élevé (ex: > 30)
   - Vérifier la cohérence `net_amount ≈ gross_amount*(1-discount_pct)` (tolérance)

Livrable : `order_lines_clean.csv` + 5 règles de nettoyage documentées.

Étape 3 — ICP (Indicateur Clé de Performance)

Calculer (après nettoyage) :

- Chiffre d'affaires total, CA mensuel

- Panier moyen (AOV) = CA / nombre de commandes
- Taux de remise moyen (global et par catégorie)
- Taux de retour (global, par catégorie, par canal)
- Score d'avis moyen (global, par catégorie, par délai de livraison)

Fonctions attendues : `groupby`, `agg`, `pivot_table`, `sort_values`, `nlargest`.

Étape 4 — Jointures (Merge) pour enrichir les ventes

Objectif: Enrichir `order_lines.csv` avec les informations clients (`customers.csv`) et produits (`products.csv`) afin de produire une **table d'analyse unique** et de recalculer quelques ICP "propres".

1. **Charger les 3 fichiers**
   - `order_lines.csv`
   - `customers.csv`
   - `products.csv`
2. **Vérifier les clés**
   - `customer_id` doit exister dans `orders` et `customers`
   - `product_id` doit exister dans `orders` et `products`
3. **Faire les jointures**
   - Jointure ventes ↔ clients sur `customer_id`
   - Jointure résultat ↔ produits sur `product_id`

Consigne : utilisez **`how="left"`** à partir de `order_lines` (on ne veut pas perdre de ventes).

1. **Contrôler la qualité de la jointure**
   - Vérifier la taille avant/après merge (nombre de lignes)
   - Compter les lignes sans match :
     - `city` ou `segment` manquant après merge (client non trouvé)
     - `category` ou `unit_price` manquant après merge (produit non trouvé)
2. **Recalculer des colonnes business après enrichissement**
   À partir des colonnes produits et commandes :
   - `gross_amount_calc = unit_price * quantity`
   - `discount_pct` converti en numérique
   - `net_amount_calc = gross_amount_calc * (1 - discount_pct)`

Comparer avec `gross_amount` et `net_amount` existants :

- Créer une colonne `amount_diff = net_amount - net_amount_calc`
- Identifier les lignes “suspectes” (ex : `abs(amount_diff) > 0.01`)

1. **Mini-analyse demandée (après jointure)**
   Produire un tableau (DataFrame) : **CA net par segment client et par catégorie produit**, trié décroissant.

- Lignes = `segment`
- Colonnes = `category`
- Valeurs = somme de `net_amount` (ou `net_amount_calc` si vous avez choisi de recalculer)

Livrables

- `orders_enriched.csv` (table finale après merge + colonnes recalculées)
- Un tableau `pivot` : CA par **segment × catégorie**
- 5 lignes de commentaires : ce que vous avez observé (valeurs manquantes, incohérences, etc.)