

**Compte rendu personnel - Transformation de données avec Pandas**

**🧩 Objectif**

Nettoyer, transformer et sauvegarder un jeu de données CSV (jeu\_donnees\_etl\_5000\_lignes.csv) à l’aide de pandas, tout en journalisant les étapes dans un fichier .log.

**0. Configuration du logger**

Un logger est initialisé en début de script pour stocker dans un fichier (dossier logs/) toutes les opérations réalisées, horodatées et structurées selon un format de journal.

✅ Log file généré : logs/log\_pipeline\_YYYYMMDD\_HHMMSS.log

Une image contenant texte, Police, logiciel, capture d’écran

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, Police, capture d’écran, ligne

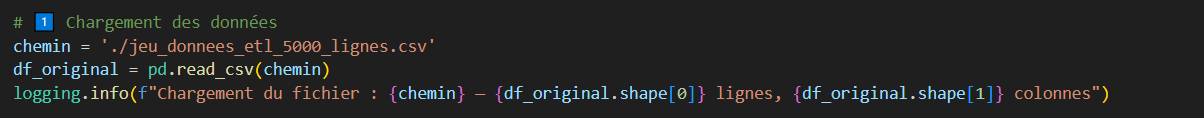
Le contenu généré par l’IA peut être incorrect.

**1️ Chargement des données :**

* Fichier lu : jeu\_donnees\_etl\_5000\_lignes.csv
* Dimensions initiales : 5266 lignes × 5 colonnes
* Colonnes :
  + ID\_produit, Nom\_produit, Quantite\_vendue, Prix\_unitaire, Date\_vente

**📌 Problèmes identifiés :**

* Valeurs nulles :
  + Nom\_produit → 11 lignes
  + Quantite\_vendue → 2672 valeurs manquantes
  + Prix\_unitaire → 2569 non-nulles
* Certaines dates invalides (invalid\_date apparaît en mode texte)
* Présence de doublons possibles



Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.

**2️ Suppression des doublons :**

Méthode : drop\_duplicates()

* Lignes supprimées : 301
* Nouvelles dimensions : 4965 lignes × 5 colonnes

✅ Action loggée correctement dans le fichier .log

Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, Police, capture d’écran, Graphique

Le contenu généré par l’IA peut être incorrect.

**3️ Winsorisation des valeurs extrêmes :**

Colonnes concernées :

* Quantite\_vendue
* Prix\_unitaire

Méthode utilisée :

* Calcul des bornes par IQR (Q1, Q3)
* Application de clip(lower, upper)

💡 Objectif : éviter les valeurs très extrêmes qui fausseraient les moyennes.

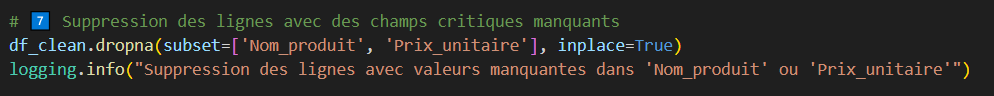
Une image contenant texte, capture d’écran, logiciel, Police

Le contenu généré par l’IA peut être incorrect.

**4️ Suppression des lignes critiques manquantes**

Colonnes critiques : Nom\_produit et Prix\_unitaire

* Les lignes où ces champs sont NaN sont supprimées (avec dropna).
* L’opération est bien loggée.
* La colonne Quantite\_vendue reste avec 1307 valeurs manquantes non traitées.



**5️ Validation des données nettoyées**

Contrôles effectués :

* 🔄 Doublons restants : 0 ✅
* ❗ Valeurs manquantes : Quantite\_vendue encore à 1307
* ❌ Valeurs négatives : aucune sur les quantités ou prix

📌 Ces résultats sont enregistrés dans un dictionnaire rapport et journalisés dans le log.

**6️ Création de la colonne Total\_HT**

Calcul : Quantite\_vendue \* Prix\_unitaire

✅ Effectué dans un bloc try/except avec journalisation.  
⚠️ Warning généré ici aussi : SettingWithCopyWarning.

➡️ Cela est dû à la modification d’un DataFrame dérivé (df\_clean). Pour éviter cela, il faut faire : Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.

**7️ Normalisation Min-Max de Total\_HT**

Nouvelle colonne : Total\_HT\_normalise  
Formule : (x−min)/(max−min)(x - min) / (max - min)

✅ Colonne bien ajoutée, mais les 1307 lignes avec NaN en Quantite\_vendue ont propagé des NaN dans Total\_HT et donc dans la normalisation aussi.

Une image contenant capture d’écran, texte, Police

Le contenu généré par l’IA peut être incorrect.

**8️ Agrégation par produit**

* Regroupement (groupby) sur Nom\_produit
* Somme du Total\_HT
* Résultat stocké dans df\_resume

✅ Renommage final en Chiffre\_affaires

Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.

**9️ Documentation finale & sauvegarde**

* Dictionnaire doc\_resume généré
* Résumé imprimé et loggé

✔️ Fichier final sauvegardé dans :  
archives/donnees\_filtrees\_20250523\_121210.csv

Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.

**✅ Conclusion**

Ce script est bien structuré, journalisé, et couvre les principales étapes du nettoyage de données. Il peut encore être renforcé avec :

* Un traitement complet des NaN (ex : imputation par médiane pour Quantite\_vendue)
* L’usage systématique de .loc pour éviter les SettingWithCopyWarning
* Une vérification des types pour Date\_vente (conversion en datetime + nettoyage de valeurs invalides)