# Atelier – Data Cleansing / Data Préparation

**Objectif principal** : Être capable de corriger les erreurs mineures et supprimer les données non pertinentes selon des critères définis, à travers une démarche structurée.

**Compétence à acquérir**

À l’issue de l’atelier, vous saurez :

* Identifier les erreurs courantes dans un jeu de données.
* Appliquer des techniques avancées de nettoyage.
* Utiliser un outil adapté pour nettoyer des données réelles.
* Automatiser et documenter un pipeline simple de data cleansing.
* Travailler en équipe autour d’un objectif de qualité de données (Digital Cleanup Day).

**LIEN POUR DEPOT DU PROJET :** [**https://forms.gle/efz2VJZcgYBiFwab9**](https://forms.gle/efz2VJZcgYBiFwab9)

**DATE LIMITE : 25/05/25**

## Introduction

Le **data cleansing**, ou **nettoyage de données**, est une étape critique dans tout projet IA ou analytique. Il vise à **détecter, corriger ou supprimer** des données :

* erronées (ex. : valeurs impossibles),
* redondantes (doublons),
* incomplètes,
* mal formatées,
* ou non pertinentes.

Un jeu de données propre garantit des **résultats plus fiables, plus rapides, et moins coûteux** à traiter. En 2025, le volume des données numériques continue d’exploser, ce qui rend leur nettoyage non seulement indispensable, mais aussi un acte éthique pour limiter l’empreinte numérique (d'où l’intérêt du volet “Digital Cleanup”).

## **Partie 1 : Qualité des données et Data Cleansing**

### Définitions et rôle

Le data cleansing n’est pas un simple “ménage”, c’est une **étape stratégique**. Il conditionne la qualité des prédictions, le respect des normes (ex : RGPD), la confiance des parties prenantes, et l’efficacité opérationnelle.

Il fait partie intégrante du **Data Quality Management System (DQMS)**, qui regroupe :

* l’intégrité,
* l’exactitude,
* la cohérence,
* la complétude,
* la fraîcheur des données.

Sans nettoyage, les modèles d’IA peuvent apprendre des biais, amplifier des erreurs ou simplement ne pas fonctionner.

### Problèmes fréquents à traiter

* **Doublons partiels ou exacts** : souvent liés à l’import de bases multiples.
* **Incohérences de format** : ex. nom = "dupont" vs "Dupont", dates inversées selon la norme.
* **Valeurs aberrantes** : ex. température = 350°C, salaire = 0€.
* **Champ vide ou NULL** : gérer leur sens (valeur manquante vs valeur valide absente).
* **Colonnes inutiles** : ex. ID techniques non pertinents, informations obsolètes.

### Techniques avancées de nettoyage

1. **Parsing (analyse syntaxique)** : extraire ou restructurer des infos avec regex (par ex. numéro de téléphone extrait d’un champ libre).
2. **Transformation** : standardiser les formats (ex. dates ISO, majuscules, unités).
3. **Renforcement des contraintes d’intégrité** : ajouter des règles métiers, contrôles automatiques, types forts.
4. **Analyse statistique** : identifier des anomalies via des histogrammes, boxplots, ou détection automatique d’outliers.
5. **Crowdsourcing et enrichissement externe** : combiner les données internes avec des bases fiables (ex. sirene.fr pour valider les entreprises).

### Étapes du processus de Data Cleansing

**1. Identifier les sources**

Tout nettoyage commence par comprendre d’où viennent les données, leur structure, leur format, leur fréquence d’actualisation. Un bon nettoyage commence **avant l’import**.

**2. Collecter**

Il s’agit de rassembler les données tout en respectant leur format d’origine. On veille aux :

* encodages (UTF-8 vs ISO),
* séparateurs (point-virgule vs virgule),
* types de fichiers (CSV, JSON, Excel).

**3. Supprimer ou fusionner les doublons**

On identifie les doublons **exactement identiques** (ligne répétée), mais aussi **approximate duplicates** (ex. : même prénom/nom avec variations typographiques).

Techniques : Fuzzy Matching, Levenshtein Distance, clustering.

**4. Gérer les valeurs manquantes**

Trois grandes stratégies :

* **Suppression** (si trop nombreuses ou inutilisables),
* **Imputation** : moyenne, médiane, prédiction,
* **Marquage** : ajout d’une colonne booléenne “is\_missing”.

**5. Nettoyer les données aberrantes**

On applique ici des méthodes de détection :

* IQR (interquartile range),
* Z-score,
* Visualisation (boxplots, scatter plots).

**6. Transformer / standardiser**

Harmonisation des noms, unités, formats :

* Dates ISO8601,
* Catégories réduites à des valeurs contrôlées,
* Standardisation des champs texte (capitalisation, accents).

**7. Automatiser**

Utiliser des scripts (Pandas, PySpark), des outils visuels (OpenRefine, Talend), ou des workflows ETL (Airbyte, Make, Alteryx). L’automatisation passe aussi par des **règles métier déclarées**.

**8. Évaluer et itérer**

Évaluer la qualité **avant/après nettoyage** :

* taux de complétude,
* taux de cohérence,
* taux d’erreurs résiduelles.

Le data cleansing est **itératif**. Plus on nettoie, plus on apprend à structurer les données en amont.

### Outils phares

* **OpenRefine** : manipulation visuelle, clustering, transformations.
* **Trifacta (Wrangler)** : nettoyage basé sur le ML, intuitif.
* **WinPure** : très utilisé en PME pour ses performances locales.
* **Drake** : pipeline extensible pour le traitement de dépendances.
* **KATARA & IntelliClean** : solutions avancées pour les gros volumes et projets collaboratifs.

## **Partie 2 : Mise en pratique guidée**

**Objectif :**

Appliquer une **méthode complète de nettoyage** sur un jeu de données réel. L’outil est au choix (Pandas, OpenRefine, Google Sheets, etc.).

**Étapes:**

1. **Inspection du dataset**  
   Comprendre les colonnes, détecter les incohérences.
2. **Nettoyage progressif**  
   Corriger les fautes, supprimer les doublons, compléter les champs, transformer les formats, etc.
3. **Documentation des choix**  
   Chaque étape doit être commentée : quelle règle ? pourquoi ? quelle alternative possible ?
4. **Export final + rapport**  
   Le fichier nettoyé est enregistré, et les étudiants produisent un **rapport de nettoyage** expliquant la démarche.

### Activité collective : Data Cleaning Day

Une animation inspirée du **Digital Cleanup Day** avec objectif écologique et pédagogique.

**Objectif : Nettoyer des fichiers numériques, comprendre leur empreinte, et améliorer la qualité globale des données.**

**Organisation type :**

* Équipe : par binôme ou trinôme
* Objectif chiffré (Go nettoyés, erreurs corrigées)
* Récompenses symboliques (badges, prix symboliques)
* Jeu de données fourni

**Objectif de l'exercice**

À partir d’un jeu de données brut contenant des erreurs, votre objectif est de :

* **Analyser** les anomalies présentes,
* **Proposer une stratégie de nettoyage** raisonnée,
* **Appliquer cette stratégie** à l’aide d’un outil adapté,
* **Documenter** vos choix méthodologiques,
* **Produire un fichier nettoyé + un mini rapport**.

**Jeu de données**

Fichier : dataset\_data\_cleaning\_5000\_rows.csv  
Format : CSV – Encodage UTF-8  
Nombre de lignes : > 5000  
Colonnes : id, first\_name, last\_name, email, birth\_year, country, gender, income

**Livrables attendus**

**1. Fichier nettoyé (au format .csv)**

Nom attendu : nom\_prenom\_cleaned.csv

**2. Rapport de nettoyage (format PDF ou Markdown)**

Nom attendu : nom\_prenom\_rapport.pdf

Contient les réponses aux questions ci-dessous

**Étapes**

Vous êtes libres dans la méthode, mais vous devez **justifier vos choix**. Réfléchissez **avant de coder**. Inspirez-vous des étapes suivantes :

**Étape 1 : Exploration initiale**

* Quelles sont les premières anomalies visibles ?
* Quels types d’erreurs semble-t-il y avoir (syntaxe, format, valeurs manquantes, incohérences…) ?

**Étape 2 : Stratégie de nettoyage**

* Quels critères allez-vous utiliser pour détecter :
  + les **doublons** ?
  + les **valeurs aberrantes** (par ex. revenus, années de naissance) ?
  + les **formats incohérents** (ex. pays ou genre) ?
* Allez-vous supprimer, corriger, ou ignorer ces anomalies ? Pourquoi ?

**Étape 3 : Implémentation**

Utilisez l’outil de votre choix :

* **Python (Pandas)** recommandé si vous êtes à l’aise avec les scripts,
* **OpenRefine**, **Google Sheets**, ou **Excel** si vous préférez une approche visuelle.

Important : **tracez vos opérations** pour pouvoir les expliquer.

**Étape 4 : Documentation**

Le rapport de nettoyage doit contenir :

1. **Votre analyse initiale** (résumé des problèmes identifiés),
2. **La méthodologie choisie** (étapes + outils),
3. **Des exemples concrets de nettoyage réalisé** (avant/après),
4. **Des réflexions personnelles** :
   * Quelles erreurs ont été les plus difficiles à gérer ?
   * Quelle partie auriez-vous pu automatiser davantage ?
   * Quel est, selon vous, l’impact de ce nettoyage sur une IA ?

**Barème indicatif d’évaluation**

| **Critère** | **Points** |
| --- | --- |
| Qualité de l’analyse initiale | 5 pts |
| Pertinence de la stratégie | 5 pts |
| Propreté et cohérence du fichier | 5 pts |
| Clarté et richesse du rapport | 5 pts |
| Capacité de justification des choix | 5 pts |

**Conseils méthodologiques**

* Ne partez **pas directement dans le nettoyage** : observez, notez, puis agissez.
* Un bon nettoyage est **documenté et reproductible**.
* Ce n’est **pas grave de supprimer des lignes** si vous justifiez pourquoi.
* Pensez à vérifier la **distribution statistique** des variables (ex. : histogrammes pour income ou birth\_year).

### Ressources complémentaires

* [Data Cleansing – DataManagement Wiki](https://datamanagement.wiki/data_quality_management_system/data_cleansing)
* [Blog DataBird : méthodes avancées de data cleaning](https://www.data-bird.co/blog/data-cleaning-methodes)
* [Unite AI – meilleurs outils 2025](https://www.unite.ai/fr/10-best-data-cleaning-tools/)
* [Guide Digital Cleanup – KABA Impact](https://www.kaba-impact.fr/blog/comment-organiser-un-digital-clean-up-dans-votre-entreprise)