

RAPPORT DE PROJET : DATA REFINEMENT

Analyse et Nettoyage du Dataset "Dirty Cafe Sales"

Date : 06 Janvier 2026 **Auteur :** GNABRO ANGE TRESOR DORIAN **Cours :** Data Refinement - HETIC

1. Introduction et Objectifs

1.1 Contexte

Ce projet vise à traiter un jeu de données brut ("Dirty Cafe Sales") contenant 10 000 transactions de ventes d'un café. Le fichier d'origine a été volontairement altéré avec des erreurs de saisie, des valeurs manquantes et des incohérences pour simuler un cas réel de "Data Engineering".

1.2 Objectifs

L'objectif n'est pas seulement de nettoyer les données, mais de les rendre **exploitables pour la prise de décision**. Les étapes suivies respectent le pipeline ETL :

- **Audit (Exploration)** : Diagnostiquer l'ampleur des dégâts.
 - **Cleaning (Nettoyage)** : Réparer les données sans perte d'information.
 - **Transformation** : Enrichir la donnée pour l'analyse business.
-

2. Étape 1 : Audit et Exploration (Notebook 01)

L'analyse exploratoire a révélé un dataset initialement inutilisable pour une analyse financière fiable.

2.1 Diagnostic des Données Manquantes

L'analyse des valeurs NaN et des chaînes de caractères ERROR / UNKNOWN a révélé :

- **Localisation** : ~40% de données inexploitables (3265 vides + 696 erreurs).

Méthode de Paiement : ~35% de données inexploitable.

- **Produits (Items)** : ~6.3% de produits non identifiés.

- **Données Financières** : Environ 2 à 3% des lignes avaient des prix ou des totaux corrompus.

2.2 Analyse de Cohérence

Une vérification cruciale a été effectuée sur les prix unitaires.

- **Constat** : Malgré les erreurs, une règle métier stricte a été identifiée : chaque article possède un prix unique et fixe (ex: *Coffee* est toujours à 2.0\$, *Salad* toujours à 5.0\$).

- **Impact** : Cette cohérence a validé notre stratégie de reconstruction des données manquantes.

3. Étape 2 : Stratégie de Nettoyage (Notebook 02)

Plutôt que de supprimer les lignes problématiques (ce qui aurait engendré une perte de ~40% du Chiffre d'Affaires), nous avons appliqué une stratégie de **Conservation Maximale**.

3.1 Standardisation

Conversion de toutes les valeurs textuelles parasites (ERROR, UNKNOWN, NaN) en format standard null pour permettre le traitement numérique.

3.2 Imputation des Catégories (Lieu et Paiement)

- **Problème** : Impossible de deviner mathématiquement le lieu ou le moyen de paiement.

- **Action** : Remplacement des valeurs manquantes par la catégorie "Unknown".

- **Justification** : Cela permet de conserver la ligne pour le calcul du CA global, même si l'axe d'analyse géographique est partiel.

3.3 Reconstruction Logique (Produits et Prix)

Nous avons utilisé la redondance de l'information pour réparer les lignes :

- **Logique** : Si le Prix est connu (ex: 5.0\$), alors l'Item est déduit (Salad). Inversement, si l'Item est connu, le Prix est corrigé.
- **Gestion des ambiguïtés** : Les articles ayant le même prix (ex: *Juice* et *Cake* à 3.0\$) n'ont pas été devinés pour éviter d'introduire de fausses données. Ils ont été étiquetés "Unknown Item".

3.4 Consolidation Mathématique et Temporelle

- **Formule appliquée** : Recalcul systématique de la colonne Total Spent via la formule $\text{Quantité} * \text{Prix Unitaire}$ pour éliminer les erreurs de calcul du fichier source.
- **Dates** : Utilisation de la méthode *Forward Fill* (propagation de la dernière date valide) pour combler les trous temporels.

Bilan du Nettoyage

- **Lignes initiales** : 10 000
 - **Lignes finales** : 9 926
 - **Taux de conservation** : **99.26%** (Performance excellente).
-

4. Étape 3 : Transformation et Analyse (Notebook 03)

Une fois les données propres, nous avons créé de nouvelles variables (Feature Engineering) pour répondre aux questions business.

4.1 Nouvelles Variables Créées

- **Temporel** : Month (Mois) et Day_of_Week (Jour) pour analyser la saisonnalité et l'affluence hebdomadaire.
- **Segmentation** : Price_Category (Low/Medium/High Cost) pour analyser le comportement d'achat selon le pouvoir d'achat.

1.

4.2 Indicateurs Clés de Performance (KPIs)

A. Analyse Produit (Stratégie Long Terme)

L'analyse a mis en évidence une distinction claire entre volume et valeur :

- **Le Champion du Volume** : Le **Café** (Coffee) et la **Salade** sont les produits les plus commandés (>3800 unités). Ils sont les moteurs du trafic client.
- **Le Champion de la Rentabilité** : La **Salade** (Salad) génère le plus gros Chiffre d'Affaires (19 075\$), loin devant le Café.
- **Recommandation** : La Salade est le produit "Vache à Lait". La gestion de son stock est critique.

B. Analyse Saisonnière (Tendance)

- **Pic d'activité** : Octobre (~7 690\$ CA).
- **Creux d'activité** : Février (~6 890\$ CA).
- **Constat** : L'activité est relativement stable, ce qui suggère une clientèle fidèle et régulière, peu impactée par les saisons.

C. Analyse Opérationnelle (Court Terme)

L'analyse des volumes de ventes par jour nous permet d'identifier les pics de consommation (environ 1400 commandes/jour). Bien que ce tableau ne soit pas un outil de gestion de stock en temps réel, il offre une vision précise de l'historique des ventes. Cela permet aux gérants d'**anticiper la demande future** et d'ajuster les approvisionnements en amont pour éviter les ruptures

5. Conclusion

Le projet de Data Refinement a permis de transformer un fichier brut comportant plus de 40% d'anomalies en une base de données **fiable à 100%**.

Les techniques de nettoyage avancées (déduction logique par le prix) ont permis de sauver la quasi-totalité du chiffre d'affaires, offrant ainsi une vision juste de la réalité économique du café. Les tableaux de bord finaux permettent désormais à la direction d'**anticiper les volumes de vente** (prévision de la demande) et de piloter la stratégie financière (suivi de la rentabilité).