

Rapport Projet Data Refinement – Café Sales Dataset

1. Introduction : L'objectif du projet est de nettoyer un dataset brut de ventes d'un café pour le rendre exploitable. Le dataset provient de Kaggle ("cafe-sales-dirty-data-for-cleaning-training"). Les étapes réalisées sont : exploration, nettoyage et transformation.
2. Exploration des données Le dataset contient 10 000 transactions et 8 colonnes : Transaction ID, Item, Quantity, Price Per Unit, Total Spent, Payment Method, Location, Transaction Date. Toutes les colonnes étaient au format texte. Il y avait des valeurs manquantes, surtout dans Payment Method (environ 25 %) et Location (environ 32 %). Des incohérences textuelles et des valeurs numériques mal formatées ont été repérées. La date des transactions était en texte.
3. Nettoyage des données Les colonnes numériques (Quantity, Price Per Unit, Total Spent) ont été converties en nombres. La colonne Transaction Date a été convertie en format date. Le texte a été standardisé : espaces supprimés, tout en minuscules, fautes corrigées (ex. : "credot card" -> "credit card", "nan" -> "unknown"). Les valeurs manquantes critiques ont été supprimées. Les autres (Payment Method, Location) ont été remplacées par "unknown". Le montant Total Spent a été recalculé pour chaque ligne : Quantity * Price Per Unit. Les doublons sur Transaction ID ont été supprimés. Les valeurs aberrantes ont été retirées (Quantity entre 1 et 50, Price entre 0,5 et 50). Résultat : 8 596 transactions propres.
4. Transformation et analyse des informations ont été extraites de la date : jour, mois, année, jour de la semaine. Les ventes ont été analysées : produits les plus vendus, périodes les plus fortes, montants totaux par produit, par paiement et par localisation.
5. Visualisation des résultats Des graphiques ont été créés :
 - Bar chart pour les produits les plus vendus.
 - Line chart pour les ventes par mois.
 - Bar chart ou heatmap pour les ventes par jour de la semaine. Ces graphiques montrent les tendances hebdomadaires et mensuelles.

6. Conclusion Le projet a transformé un dataset brut et incohérent en un dataset propre et cohérent. Les étapes d'exploration, de nettoyage et de transformation ont permis d'obtenir des données fiables pour des analyses et visualisations.