Ce projet a pour objectif de mettre en œuvre un processus complet de raffinement des données à partir d'un jeu de données volontairement brut et imparfait. Le travail réalisé vise à appliquer les notions de data quality, de nettoyage et de transformation afin de produire un jeu de données cohérent et exploitable.

Le jeu de données utilisé représente des transactions de ventes réalisées par un café. Chaque ligne correspond à une transaction et contient des informations telles que le produit vendu, la quantité, le prix, la date, le mode de paiement et la localisation.
La phase d'exploration a permis d'identifier plusieurs problèmes de qualité, notamment des valeurs manquantes, des types de données inadaptés et des incohérences dans certaines colonnes.

Une phase de nettoyage a été réalisée afin de corriger les principaux problèmes de qualité identifiés. Les doublons ont été supprimés afin d'éviter toute redondance.
Les colonnes essentielles à l'interprétation d'une transaction (produit, quantité, prix et date) ont été traitées en priorité. Les lignes contenant des valeurs manquantes sur ces champs ont été supprimées, car elles ne permettaient pas une analyse fiable.
En revanche, les colonnes secondaires telles que la méthode de paiement et la localisation ont été conservées, et leurs valeurs manquantes ont été remplacées par "Unknown" afin de limiter la perte d'information.
La phase de transformation a permis de rendre les données exploitables. Les noms de colonnes ont été standardisés, les types de données corrigés et les dates converties dans un format approprié. La colonne représentant le montant total de la transaction a été recalculée à partir de la quantité et du prix unitaire, afin de garantir la cohérence des valeurs. Le jeu de données final a ensuite été exportéès dans un fichier CSV prêt à être utilisé.

Ce projet a permis de mettre en application un processus structuré de raffinement des données, depuis l'exploration jusqu'à la production d'un jeu de données propre et cohérent. Les choix effectués ont été guidés par des considérations de qualité et de logique métier, aboutissant à un dataset exploitable pour des analyses futures.