Rapport de Data Refinement

Introduction

Ce rapport présente de manière détaillée le processus de data refinement appliqué à un dataset de transactions. L'objectif du projet est de transformer des données brutes en un jeu de données propre, cohérent et exploitable pour des analyses ultérieures.

Description du dataset

Le dataset contient des informations relatives à des transactions, incluant les produits achetés, les quantités, les montants dépensés, les méthodes de paiement, les localisations et les dates de transaction. Il comporte plusieurs milliers d'observations et présente divers problèmes de qualité des données.

Exploration des données

Une analyse exploratoire a été réalisée afin de comprendre la structure du dataset. Cette étape a permis d'identifier les types de variables, la présence de doublons et l'existence de valeurs manquantes ou incohérentes. Les colonnes relatives à la méthode de paiement et à la localisation présentaient notamment un taux élevé de valeurs manquantes.

Nettoyage des données

Le nettoyage des données a consisté à supprimer les doublons et à corriger les valeurs invalides. Les valeurs textuelles non informatives ont été assimilées à des valeurs manquantes. Les variables numériques ont été converties en types numériques et imputées par la médiane afin de limiter l'influence des valeurs extrêmes. Les variables catégorielles ont été traitées selon leur taux de valeurs manquantes.

Transformation des données

La variable de date a été convertie en format datetime afin de garantir la cohérence des informations temporelles. Des variables temporelles ont ensuite été créées à partir de cette date. Les variables catégorielles ont été encodées afin de rendre le dataset compatible avec des analyses statistiques ou des modèles prédictifs.

Organisation du projet et versionnement

Le projet a été structuré en plusieurs notebooks correspondant aux différentes étapes du data refinement : exploration, nettoyage et transformation. Conformément aux consignes pédagogiques, les datasets bruts et transformés ont été versionnés sur GitHub afin de garantir la traçabilité et la reproductibilité du projet.

Conclusion

Le processus de data refinement a permis d'améliorer significativement la qualité du dataset. Les données finales sont désormais propres, cohérentes et structurées, constituant une base fiable

pour des analyses avancées ou des projets de data science.