Rapport du data refinement

Pour ce projet, j'ai travaillé sur un dataset brut des ventes d'un café appelé « cafe_sales_dirty.csv ». L'objectif était de nettoyer ce dataset pour qu'il soit exploitable et cohérent car le dataset brut avait des valeurs manquantes, des doublons, des erreurs de format et des incohérences.

Du coup, j'ai mis en pratique les notions de data quality et data refinement que nous avons vues en cours pour régler cela.

J'ai pu voir plusieurs problèmes au début du dataset comme certaines colonnes qui contenaient "ERROR" ou "UNKNOWN", ou des colonnes numériques qui étaient en format texte, ou encore des informations qui étaient manquantes et il y' avait aussi des doublons exacts. J'ai donc traiter ces problèmes pour ne pas perdre de données importantes et pour ne pas avoir de confusion sur le dataset nettoyé (final).

Les étapes que j'ai mis en place pour réalisé le dataset final.

Dans le notebook 1 (01_EXPLORATION) :

J'ai commencé par charger le dataset brut et observer les premières lignes et les statistiques générales.
Ensuite, j'ai remarqué les problèmes tels que les valeurs manquantes dans item, payment_method et location, des erreurs dans total_spent et quantity sous forme de "ERROR",des colonnes numériques en texte, et des doublons exacts dans quelques lignes.

Du coup, j'ai choisi de remplacer les valeurs incorrectes pour ne pas perdre d'informations que les supprimer parce que une fois supprimer je peux pas revenir en arrière.

Dans le notebook 2 (02_CLEANING) :

Pour commencer, j'ai standardisé les noms des colonnes pour faciliter la manipulation et éviter les erreurs de type KeyError. On peut prendre comme exemple dans Transaction ID qui est devenu transaction_id et Price Per Unit qui est devenu price_per_unit. Cette étape permet de travailler facilement sur les colonnes sans se tromper dans les noms.

Ensuite, j'ai remplacé toutes les valeurs "ERROR", "UNKNOWN" et vides par NaN et j'ai supprimé les doublons exacts. Cette étape m'a permis donc de nettoyer les erreurs tout en conservant les lignes importantes.

Puis,j'ai converti les colonnes numériques (quantity, price_per_unit, total_spent) en type numérique et transaction_date en type date. Du coup, cela me permet de pouvoir effectuer des calculs corrects sur ces colonnes et de détecter facilement les

incohérences, comme des totaux qui ne correspondent pas à la multiplication quantity * price_per_unit.

Puis, j'ai remplis les valeurs manquantes comme pour les colonnes numériques, j'ai aussi utilisé la médiane, et pour les colonnes catégorielles (item, payment_method, location), j'ai mis "unknown". Du coup sa permet de conserver toutes les lignes et de signaler clairement les informations manquantes sans supprimer de données, comme ça le dataset peut être complet et exploitable.

Dans le notebook 3 (03_TRANSFORMATION) :

J'ai commencé par mettre une colonne calculée total_spent_calculated, qui correspond à quantity * price_per_unit. Cela m'a permis de vérifier la cohérence avec la colonne total_spent et de repérer et corriger les erreurs éventuelles du dataset brut.

Puis j'ai vérifié les statistiques du dataset nettoyé, j'ai vue que les colonnes numériques étaient correctes, total_spent_calculated correspondait aux valeurs attendues, et les colonnes catégorielles contenaient "unknown" pour signaler les valeurs manquantes. Mais il y'avait encore quelques dates qui étaient manquantes (NaT).

Enfin, j'ai sauvegardé le dataset clean dans le dossier Data/processed sous le nom cafe_sales_clean.csv.

Ma Réflexion/ Ma Justification :

Pendant le projet, j'ai commencé par remplacer les valeurs incorrectes par NaN puis par "unknown" car cela m'a permis de ne pas perdre de données tout en signalant qu'il manquait des informations.

Puis, j'ai remplis les colonnes numériques manquantes par la médiane car cela m'a permis de garder des valeurs réalistes et de ne pas fausser les statistiques globales.

Ensuite, j'ai créer une colonne calculée (total_spent_calculated) qui m'a aider à vérifier que les calculs sont corrects et de pouvoir repérer les erreurs du dataset brut.

Et enfin, j'ai garder quelques valeurs comme "unknown" ou NaT d'après des recherches qui m'ont aider à savoir que cela est volontaire pour montrer où il manque des informations et pour ne pas supprimer inutilement des lignes.

Travailler sur le dataset brut m'a permis de comprendre l'importance de chaque étape du nettoyage et de réfléchir à des solutions simples. Mon dataset clean je pense qu'il est maintenant exploitable, cohérent.