Gao Erwann DIA2

Reports Projet Final DATA Refinement

Choix du DATASET : dirty_cafe_sales

La motivation première de ce choix est que ce DATASET provient de Kaggle un site que j'utilise déjà pour des projets DATA-IA, je compte donc aller plus loin avec ce DATASET, potentiellement réaliser un projet DATA-IA avec ce DATASET nettoyé par la suite.

Objectif : Mettre évidence la demande et le chiffre d'affaires pour chaque produit.

01_EXPLORATION

J'ai dans ce notebook mis en évidence les principales caractéristiques du DATASET utilise pour pouvoir mieux comprendre les tenants et aboutissant pour pouvoir correctement nettoyer pour par la suite le transformer plus facilement.

Voici un résumé des informations que j'ai pu relever :

Le nombre de colonnes est : 8

Les colonnes sont : [« ID de transaction », « Article », « Quantité », « Prix par unité », « Total dépensé », « Méthode de paiement », « Emplacement », « Date de transaction »]

Le nombre de lignes est : 10 000

Il n'y a aucune ligne dupliquée

Il y a 5450 lignes avec des valeurs manquantes

Résumé des erreurs dans le DATASET d'origine :

Transaction ID: {}

Item: {'UNKNOWN': 344, nan: 333, 'ERROR': 292}

Quantity: {'UNKNOWN': 171, 'ERROR': 170, nan: 138}

Price Per Unit: {'ERROR': 190, nan: 179, 'UNKNOWN': 164}

Total Spent: {nan: 173, 'UNKNOWN': 165, 'ERROR': 164}

Payment Method: {nan: 2579}

Location: {nan: 3265, 'ERROR': 358, 'UNKNOWN': 338}

Transaction Date: {nan: 159, 'UNKNOWN': 159, 'ERROR': 142}

02_CLEANING

Dans cette partie l'objectif était de corriger les erreurs relevées plus tôt, on ajoute une vérification et correction des valeurs de quantity, price per unit et total spent car ce sont les valeurs les plus importantes pour la transformation.

D'abord on harmonise toutes les valeurs erronées en les changeant en valeurs manquante pour faciliter le traitement de celle-ci par la suite.

On supprime les lignes dont le nom de l'objet est manquant car elles ne peuvent servir notre objectif.

On change les types des colonnes numérique pour pouvoir faire des opérations dessus.

On de la cohérence des colonnes : ['Quantity', 'Price Per Unit', 'Total Spent']

On supprime les lignes ou ce n'est pas possible

On sauvegarde les données nettoyer dans le fichier cleaned_cafe_sales.csv

03_TRANSFORMATION

On peut d'abord améliorer le repère temporel en changeant le type de Transaction Date en datetime et en créant les colonnes Year Month Day weekday. Cela pourra être utile pour de plus ample recherche.

On peut enfin mettre en évidence le revenu généré par chaque produit, d'autres informations peuvent être relevé comme le nombre d'exemplaire moyen d'un produit acheté en une seul fois et la demande global de chaque produit.