

RAPPORT DE PROJET – RAFFINEMENT DE DONNÉES (CAFÉ SALES)

Étudiant : Ethan

Projet : Raffinement complet d'un dataset brut

Dépôt GitHub : data_refirment

1. Introduction et Objectif

L'objectif de ce projet était d'appliquer les principes de Data Quality sur un dataset brut de 10 000 transactions issues des ventes d'un café. Le but était de nettoyer, transformer et produire une donnée fiable et exploitable pour une analyse commerciale réelle.

2. Diagnostic de la Qualité des Données

L'analyse initiale du fichier dirty_cafe_sales.csv a révélé plusieurs problèmes majeurs : - Valeurs manquantes importantes, notamment dans les colonnes Payment Method (2579 valeurs nulles) et Location (3265 valeurs nulles). - Problèmes de typage : toutes les colonnes étaient au format texte, empêchant les calculs numériques. - Absence de doublons stricts dans le dataset.

3. Justification des Choix de Nettoyage

Les décisions de nettoyage ont été prises selon une logique métier : - Suppression des lignes avec Item manquant afin d'éviter toute interprétation erronée des ventes. - Imputation des valeurs manquantes de Payment Method et Location par la catégorie "Unknown" afin de conserver le volume global des ventes. - Standardisation des champs texte pour corriger les incohérences de saisie (ex : coffee → Coffee).

4. Transformation et Visualisation

De nouvelles colonnes ont été créées pour renforcer la cohérence : - Création de Total_Spent_Calculated pour vérifier la cohérence financière. - Analyse temporelle des ventes afin d'identifier les périodes d'activité. Les visualisations montrent que les produits Salad et Sandwich génèrent le plus de chiffre d'affaires, et que l'activité est concentrée en soirée.

5. Analyse des Graphiques

Top 10 Produits : Les produits les plus rentables sont la Salad, le Sandwich et le Smoothie, suggérant une préférence pour des repas complets.

Répartition Temporelle : 100% des ventes sont concentrées le soir, indiquant soit une spécificité commerciale, soit un biais de collecte.

Relation Quantité / Total : La relation linéaire parfaite confirme la validité des calculs et l'absence d'anomalies financières.

6. Conclusion

Le dataset final nettoyé, stocké dans data/PROCESSED/cleaned_cafe_sales.csv, est désormais fiable, cohérent et prêt pour une utilisation en Business Intelligence. Ce projet démontre l'importance du raffinement des données avant toute prise de décision.