

Rapport de Projet : Nettoyage et Analyse des Ventes du Café

1. Introduction et Contexte

Ce projet porte sur le dataset `dirty_cafe_sales.csv`. L'objectif était de transformer des données brutes remplies d'erreurs en une base de données propre pour aider un gérant de café à comprendre ses ventes.

Note particulière : J'ai été absente les trois premiers jours du projet. J'ai donc dû rattraper les notions de Data refinement rapidement pour réaliser ce travail avec l'aide de mes camarades et l'a par la suite.

2. Identification des problèmes

Quand j'ai ouvert le fichier pour la première fois (Notebook 01), j'ai remarqué plusieurs problèmes :

- **Erreurs de calcul** : Beaucoup de lignes affichaient "ERROR" dans la colonne Total Spent.
- **Données manquantes** : Des informations comme le mode de paiement ou le lieu (In-store/Takeaway) étaient vides ou marquées "UNKNOWN".
- **Types de données** : Les prix et quantités étaient mélangés avec du texte, ce qui empêchait de faire des calculs mathématiques.

3. Ma logique de nettoyage (Notebook 02)

Ma priorité a été de ne pas supprimer de données, car chaque ligne représente une vente et donc de l'argent.

- **Le choix du recalcul** : Au lieu d'effacer les lignes avec "ERROR", j'ai forcé les colonnes en chiffres et j'ai demandé à Python de recalculer : $\text{Quantité} \times \text{Prix} = \text{Total}$. Cela m'a permis de "sauver" beaucoup de données.
- **La simplification** : J'ai remplacé tous les mots bizarres (ERROR, UNKNOWN) par "Non Spécifié" pour que les tableaux finaux soient lisibles.

4. Transformation et Résultats (Notebook 03)

Une fois le fichier propre (`cafe_sales_cleaned.csv`), je l'ai transformé pour qu'il soit utile:

1. **Analyse Temporelle** : J'ai ajouté les jours de la semaine. Cela nous permet de voir quel jour le café doit prévoir plus de serveurs.

2. Analyse Produit : J'ai créé un tableau qui classe les produits par rentabilité.

Résultat final : Le fichier final (cafe_sales_transformed.csv) est désormais totalement exploitable. On peut maintenant voir précisément quel produit est le plus vendu et quel est le jour le plus chargé de la semaine, ce qui était impossible au début à cause des erreurs.

5. Conclusion

Ce projet m'a appris que la qualité des données est la base de tout. Même si j'ai manqué le début, j'ai compris que la logique est plus importante que le code : il faut d'abord comprendre pourquoi la donnée est fausse avant de chercher à la réparer.