RAPPORT DE RAFFINEMENT DE DONNÉES : CAFÉ SALES

Étudiant : Ethan

Projet : Nettoyage et fiabilisation d'un dataset de 10 000 transactions

Lien GitHub : data_refirment

1. Introduction et Objectif

Ce projet porte sur le traitement d'un jeu de données brut de 10 000 lignes représentant les ventes d'un café. L'objectif était de transformer des données sales (incohérentes, manquantes, mal typées) en une base de données propre, prête à être utilisée pour une analyse commerciale fiable.

2. Diagnostic Initial (Data Quality)

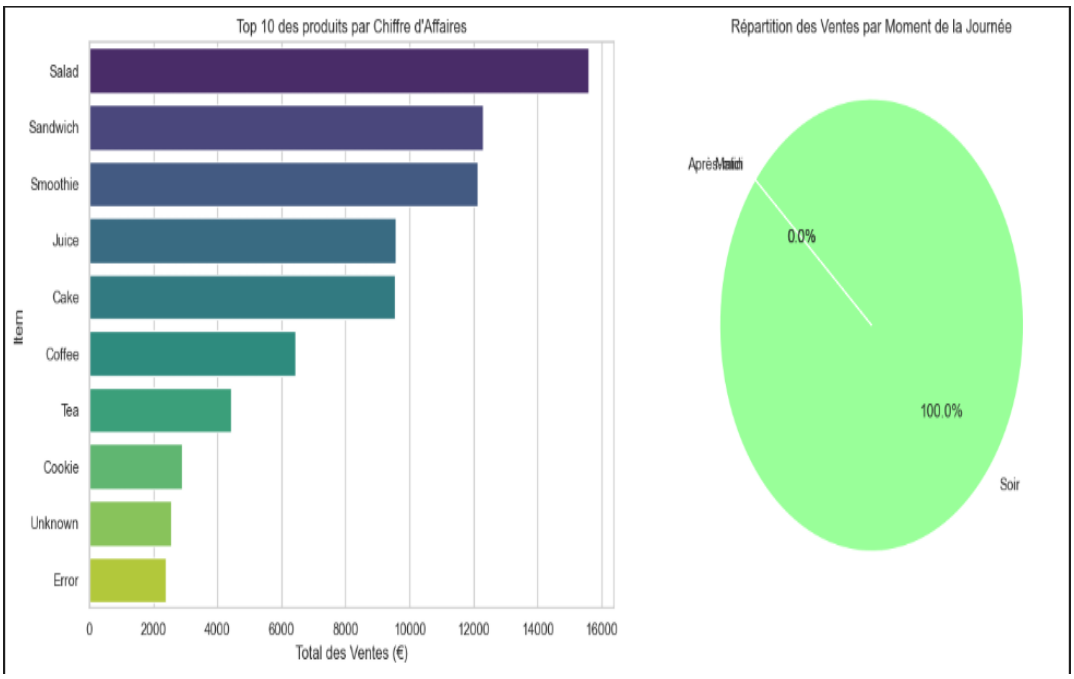
L'exploration du fichier dirty_cafe_sales.csv a révélé des problèmes critiques empêchant toute analyse immédiate : - Valeurs manquantes massives : 2 579 lignes sans mode de paiement et 3 265 sans localisation. - Types de données erronés : les prix et quantités étaient stockés comme texte. - Incohérences de saisie : produits identiques écrits différemment.

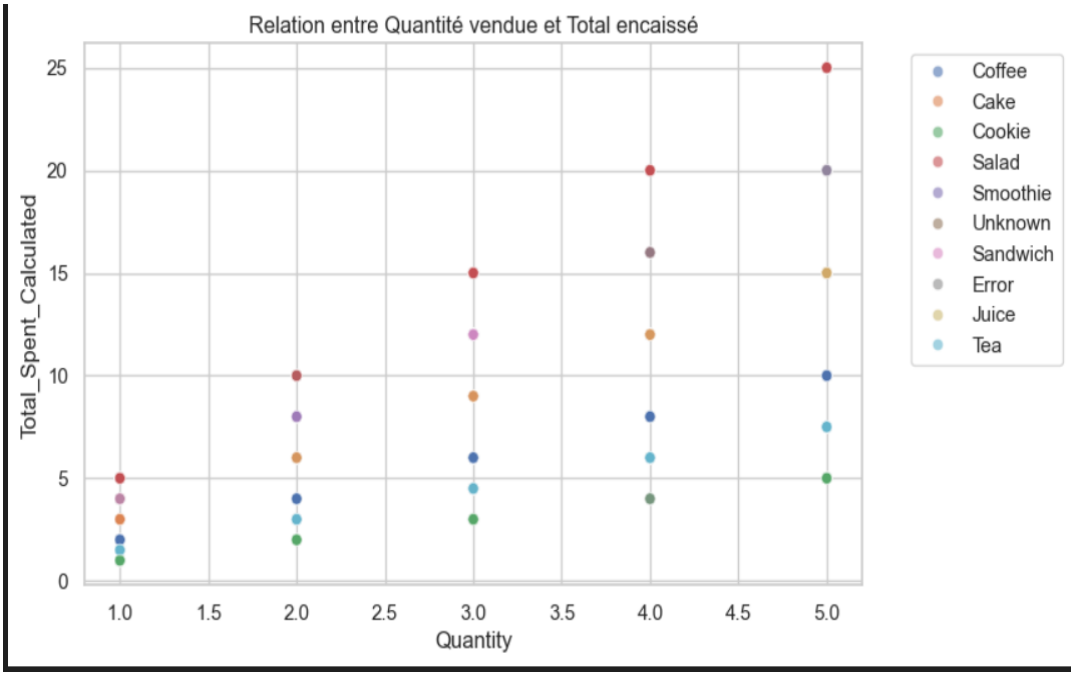
3. Justification des choix techniques (Méthodologie)

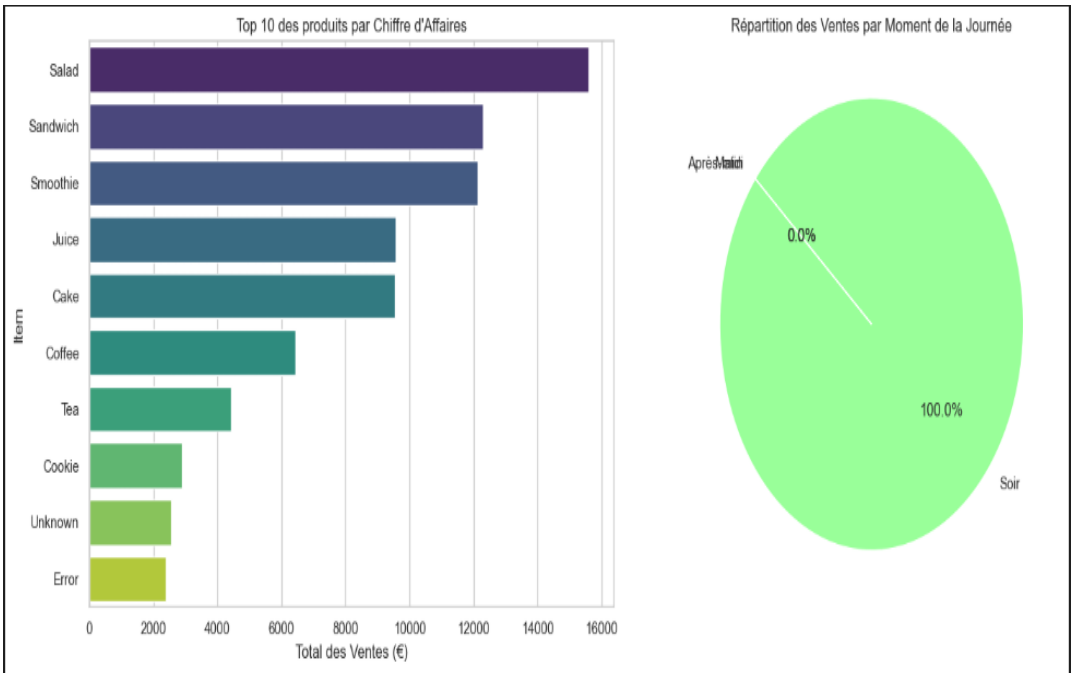
Suppression des lignes sans Item : suppression des transactions où le produit est manquant afin d'éviter toute fausse analyse. Imputation par 'Unknown' : remplacement des valeurs manquantes pour conserver le volume de données. Standardisation du texte : harmonisation des noms de produits pour garantir la cohérence analytique.

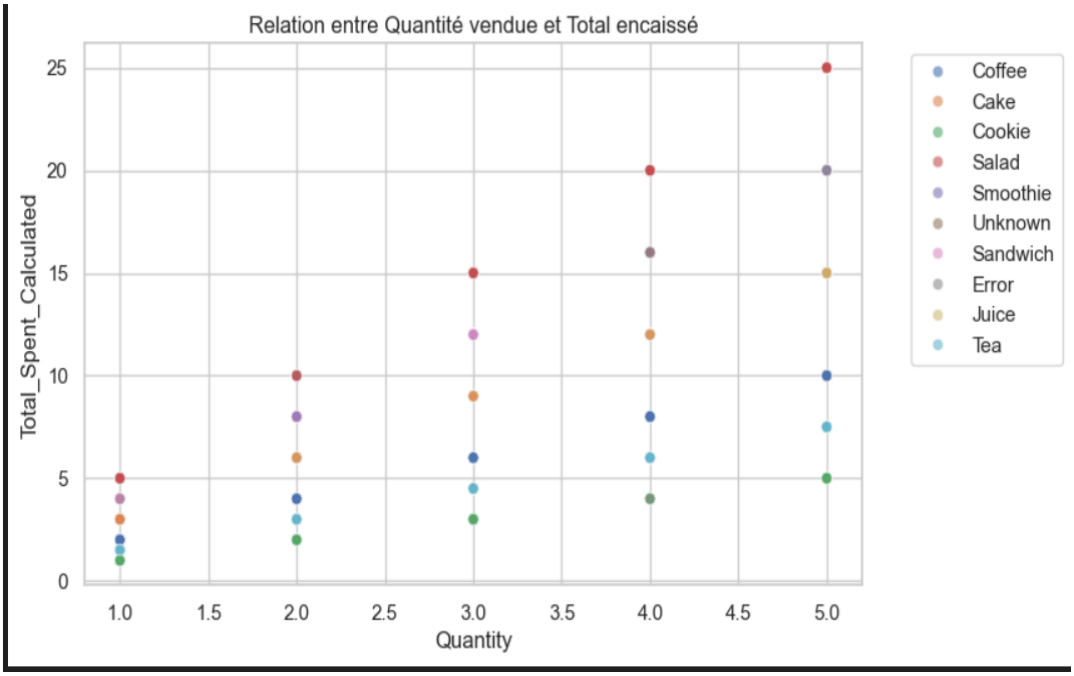
4. Visualisations et Analyse des Résultats

Les analyses montrent que la Salade, le Sandwich et le Smoothie sont les produits les plus rentables. La relation linéaire parfaite entre quantité et total encaissé valide la cohérence financière. La répartition temporelle révèle une concentration des ventes en soirée.




Relation entre Quantité vendue et Total encaissé
Total_Spent_Calculated
Quantity
0
5
10
15
20
25
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Coffee
Cake
Cookie
Salad
Smoothie
Unknown
Sandwich
Error
Juice
Tea


Top 10 des produits par Chiffre d'Affaires
Item
Salad
Sandwich
Smoothie
Juice
Cake
Coffee
Tea
Cookie
Unknown
Error
0
2000
4000
6000
8000
10000
12000
14000
16000
Total des Ventes (€)
Répartition des Ventes par Moment de la Journée
0.0%
100.0%
Soir



5. Conclusion

Le dataset final, situé dans data/PROCESSED/cleaned_cafe_sales.csv, est désormais fiable et exploitable pour des analyses Business Intelligence.