Rapport de Data Refinement

Projet : Cafe Sales Dirty Data

Dataset source : Cafe Sales Dirty Data

Ce projet porte sur le nettoyage et la transformation du dataset "Cafe Sales Dirty Data" provenant de Kaggle. Ce dataset contient des données de ventes d'un café contenant des erreurs de qualité (valeurs null, doublons, formats incohérents, etc...).

On va donc faire une exploration, puis un nettoyage du dataset afin de pouvoir manipuler ces données sur un dataset clean !

État initial du dataset

- **Nombre de lignes :** 10 000
- **Nombre de colonnes :** 8
- **Format :** CSV
- **Colonnes :** Transaction ID, Item, Quantity, Price Per Unit, Total Spent, Payment Method, Location, Transaction Date

1. Exploration et Problème identifiés

- 1000 lignes et 8 colonnes
- Les colonnes sont tous en string, hors:
    - Les dates nécessitant une conversion en datetime
    - Les prix doivent être en float

- Grand nombre de valeurs null
- Pas de doublon

2. Nettoyage et transformation du dataset

✦ Gestion des valeurs calculables

Quantity et Total Spent (502 NaN chacun - 5.0%)

- **Décision :** Calcul intelligent avec formules mathématiques
- **Méthodologie appliquée :**
    1. Si Quantity manque et que Total Spent et Price Per Unit sont présents → Quantity = Total Spent / Price Per Unit
    2. Si Total Spent manque et que Quantity et Price Per Unit sont présents → Total Spent = Quantity × Price Per Unit
    3. Arrondi des quantités calculées à l'entier le plus proche
    4. On garde tout de même les lignes ou il y a des valeurs manquantes pour pas perdre de la donnée qui peuvent nous être utile

=> Cette approche permet de maximiser la conservation des données tout en garantissant leur cohérence mathématique. Selon moi les valeurs calculées sont fiables car basées sur la relation prix unitaire × quantité = total.

✦ Gestion des autres valeurs

Gestion des valeurs ‘ERROR’ et ‘UNKNOWN’

- Remplacer les valeurs ‘ERROR’ et ‘UNKNOWN’ par des valeurs null. Avec les valeurs nulle on pourra plus facilement les traiter et les manipuler.

Gestion des prix unitaire null

- Certains 'Item' étaient vides, on peut donc essayer de les remplir approximativement avec leur prix c'est-à-dire que les prix ont un seul prix dans le dataset et on va chercher ce prix là grâce à leur nom item

Autre transformations effectués:

1. **Conversion des dates :** Transformation de la colonne Transaction Date en format datetime pour permettre les analyses temporelles
2. **Création de nouvelles :** Colonnes pour Year, Month et Day de datime
3. **Extraction temporelle :** Création d'une colonne 'Month' pour faciliter l'agrégation mensuelle des ventes
4. **Normalisation des types :** Standardisation des types de données (int pour Quantity, float pour les montants)
5. **Arrondi des valeurs calculées :** Les quantités calculées ont été arrondies pour refléter la réalité des transactions (pas de demi-produit)

3. Résultats et qualité finale

✦ Métriques du dataset nettoyé

- **Dataset initial :** 10 000 lignes
- **Dataset nettoyé :** 8 569 lignes (85.7% conservé)
- **Lignes supprimées :** 1 431 (14.3%)
- **Valeurs manquantes restantes :** 0
- **Colonnes conservées :** 7 (suppression de Location)

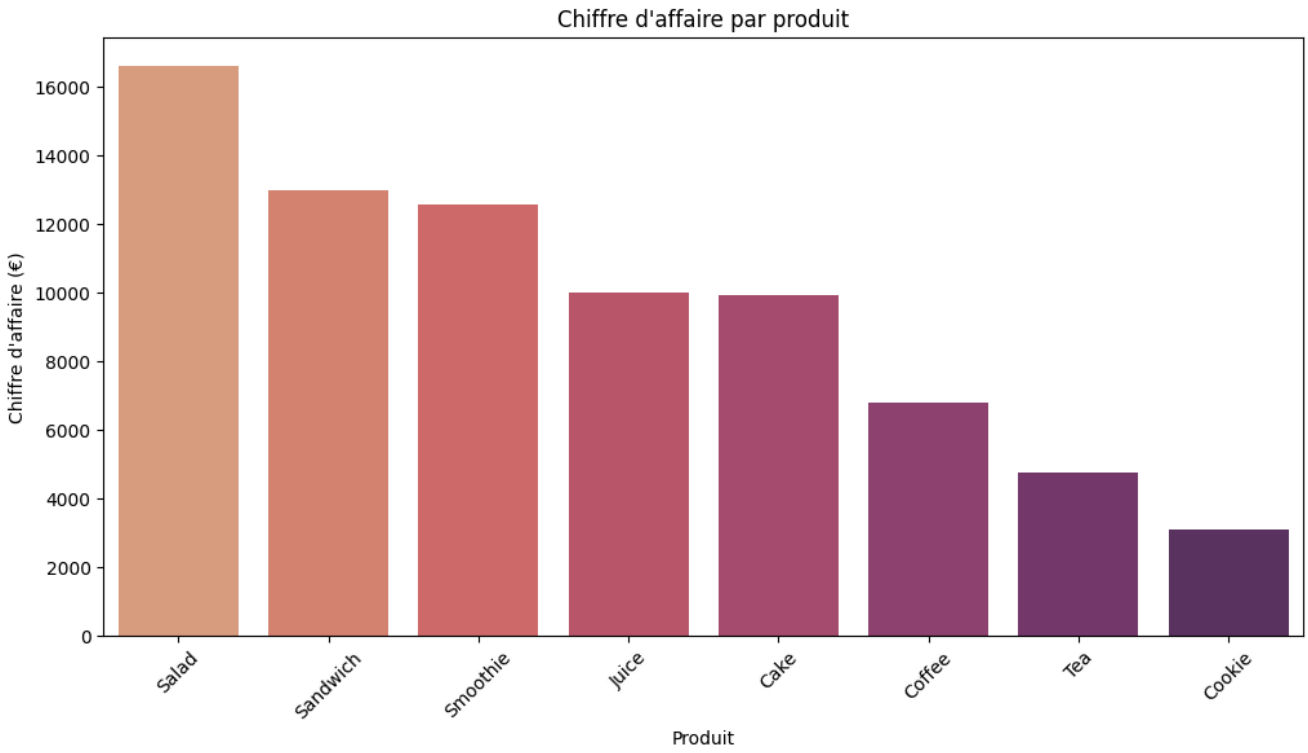
✦ Validation de la qualité

- Remplissage des valeurs null (dans la possibilité)
- Types de données corrects et standardisés
- Dates valides et exploitables
- Dataset prêt pour l'analyse et la visualisation

4. Insights et visualisations

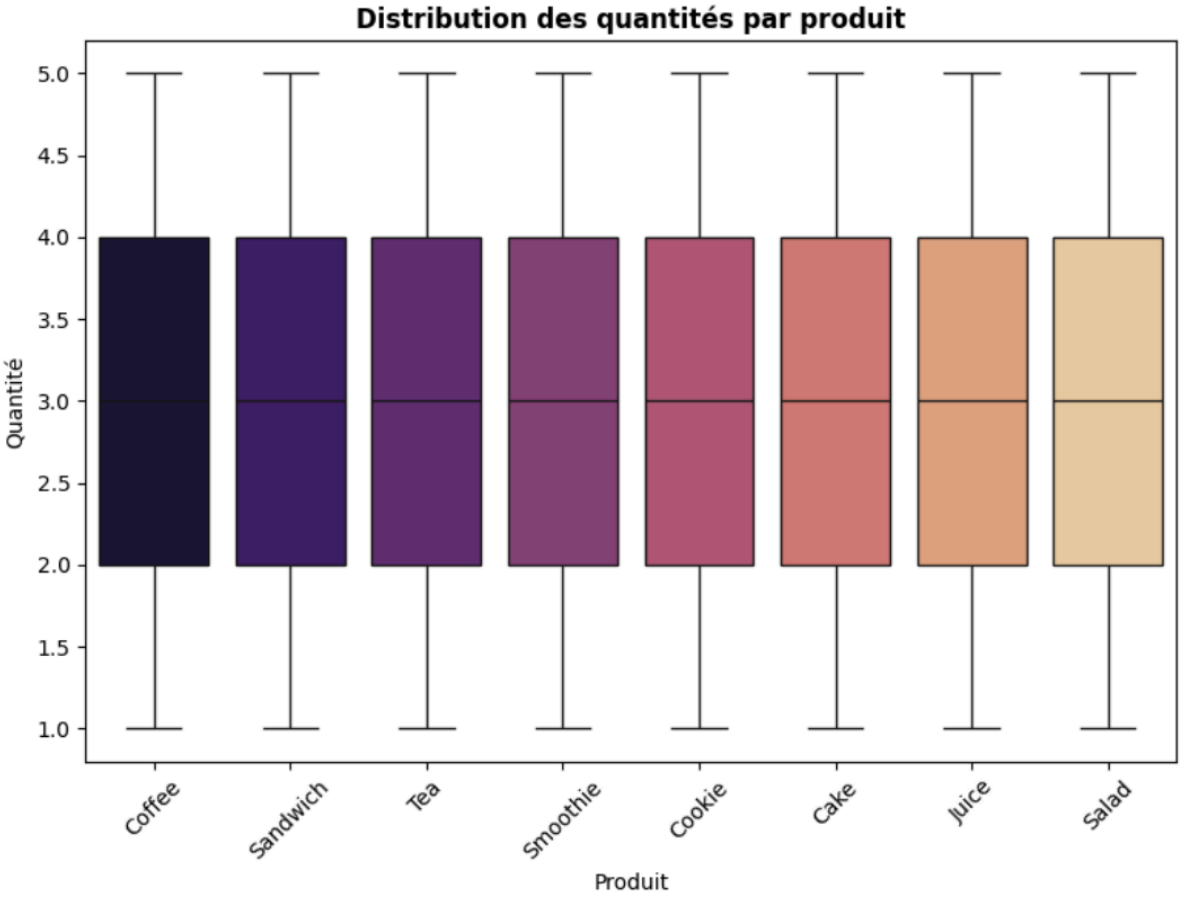
✦ Analyse des produits

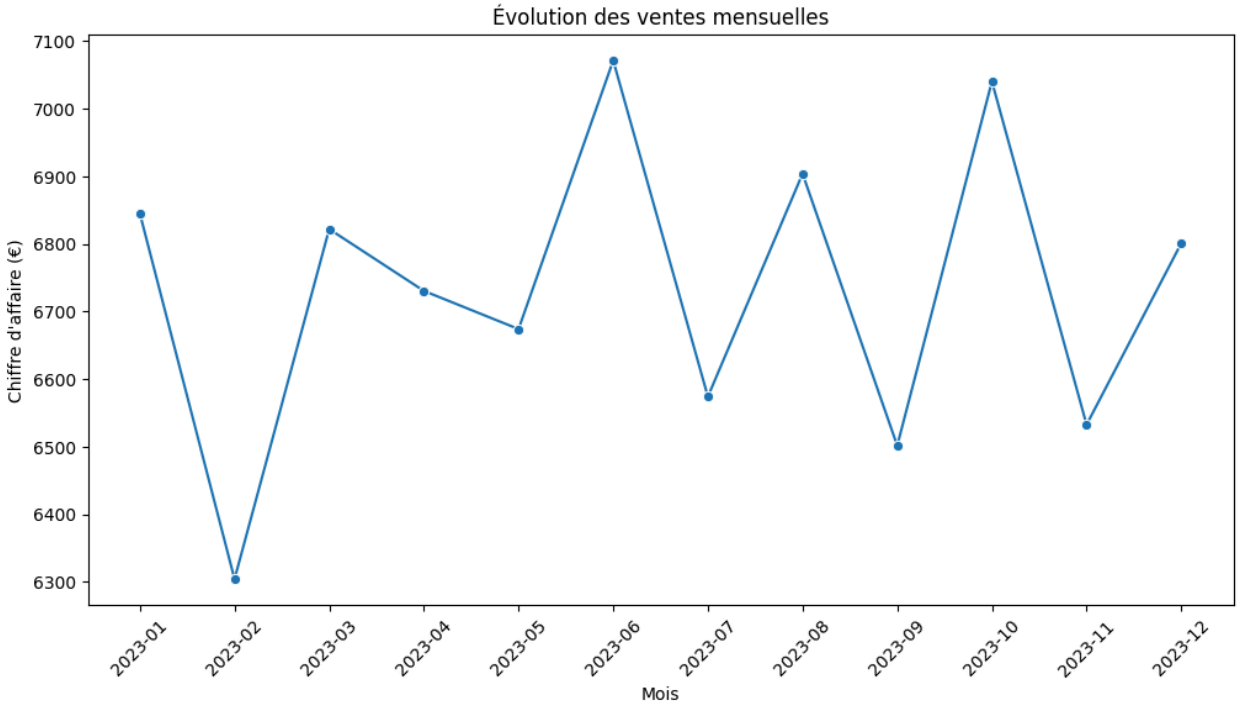
- **Distribution équilibrée :** Tous les produits ont un nombre de ventes similaire (~1100-1150 ventes)
- **Salad = produit star :** Génère le plus de revenus avec un panier moyen de ~15€
- **Prix fixes :** Le système de prix est stable



✦ Comportement d'achat

- **Quantité typique :** Les clients achètent en moyenne 2-4 unités par transaction
- **Stabilité temporelle :** Les ventes restent relativement stables sur l'année 2025


Distribution des quantités par produit
Quantité
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Coffee
Sandwich
Tea
Smoothie
Cookie
Cake
Juice
Salad
Produit


Évolution des ventes mensuelles
Chiffre d'affaire (€)
7100
7000
6900
6800
6700
6600
6500
6400
6300
2023-01
2023-02
2023-03
2023-04
2023-05
2023-06
2023-07
2023-08
2023-09
2023-10
2023-11
2023-12
Mois

5. Conclusion

Le processus de data refinement a permis de transformer un dataset brut de 10 000 lignes contenant de nombreuses erreurs en un dataset propre et exploitable de **10 000 lignes (100% des données conservées)**.

Les choix effectués ont privilégié **la conservation maximale des données** tout en garantissant leur qualité :

✦ Les valeurs de Quantity et Total Spent ont été calculées mathématiquement lorsque possible (formule : Quantity × Price = Total), permettant de récupérer des données au lieu de les supprimer

✦ Les colonnes contenant des 'ERROR' et 'UNKNOWN' on été remplacé par des valeurs null afin de les exploiter plus facilement (ex: Remplissage des colonnes par les calculs mathématiques)

✦ Transformation des dates en datetime et normalisation des formats numériques pour assurer la cohérence + création de nouvelles colonnes Year, Month, Day

✦ Vérification systématique que les relations entre quantités, prix et totaux sont respectées

Cette approche **orientée conservation** permet de maximiser le volume de données disponibles pour l'analyse, tout en maintenant un niveau de qualité élevé.

=> Les visualisations réalisées confirment la qualité des données nettoyées et permettent d'identifier des insights business actionnables, notamment la forte performance des salades et la stabilité du comportement d'achat client.

Ma citation: *"Ne jamais supprimer une donnée si elle peut être sauvée intelligemment."* ;)

✦ By Caroline CHEN 👩🏻‍💻 DIA2 ✦