Exercice : Analyse des ventes avec Pandas

Vous disposez d'un fichier ventes.csv (ou ventes.xlsx) contenant des données de ventes d'une entreprise. Le fichier contient les colonnes suivantes :

-**Date** : date de la vente

-**Produit** : nom du produit vendu

-**Catégorie** : catégorie du produit (Électronique, Mode, Alimentaire…)

-**Quantité** : nombre d'unités vendues

-**PrixUnitaire** : prix d'une unité

-**Magasin** : nom ou ville du magasin

Votre objectif est de réaliser une série d'analyses avec Pandas.

Partie A : Exploration des données

1. Charger le fichier CSV ou Excel avec pandas.
2. Afficher les 5 premières lignes du dataset.
3. Afficher les informations générales du dataset (.info()) et les statistiques descriptives (.describe()).
4. Identifier s'il existe des valeurs manquantes.

Partie B : Nettoyage et préparation

5. Convertir la colonne Date en type datetime.
6. Créer une nouvelle colonne **ChiffreAffaires = Quantité * PrixUnitaire.**
7. Vérifier s'il existe des doublons et les supprimer.

Partie C : Analyses simples

8. Trouver le produit le plus vendu en termes de quantité.
9. Calculer le chiffre d'affaires total par catégorie.
10. Calculer le chiffre d'affaires moyen par magasin.
11. Compter le nombre de ventes par mois.

Partie D : Analyses avancées

12. Identifier le top 5 des produits les plus rentables (en termes de CA).
13. Créer un tableau croisé (pivot table) : chiffre d'affaires par Catégorie et par Magasin.
14. Tracer un graphique de l'évolution du chiffre d'affaires par mois.

Partie E : Analyses supplémentaires

15. Identifier le magasin avec le plus grand chiffre d'affaires cumulé.
16. Trouver le mois où les ventes (quantité) ont été les plus fortes.
17. Calculer le prix unitaire moyen par produit.
18. Afficher le top 3 des catégories qui génèrent le plus de ventes en volume (quantité).
19. Analyser la répartition des ventes par jour de la semaine (CA total).
20. Créer un histogramme représentant la répartition des quantités vendues.

Exercice (Partie 2) : Traitement et modélisation avec Scikit-Learn

On repart du même dataset **`ventes.csv`** après l'avoir préparé (notamment avec la colonne **`ChiffreAffaires`**).

Partie F : Préparation des données

21. Sélectionner uniquement les colonnes utiles pour la modélisation :

- **`Quantité`**
- **`PrixUnitaire`**
- **`Catégorie`**
- **`Magasin`**
- **`ChiffreAffaires` (comme variable cible)**.

22. Transformer les variables catégorielles **(`Catégorie`, `Magasin`)** en variables numériques en utilisant Scikit-learn.

23. Séparer les données en **X (features)** et **y (target)** :

- X = toutes les colonnes sauf **`ChiffreAffaires`**
- y = **`ChiffreAffaires`**

24. Découper le dataset en **jeu d'entraînement** et **jeu de test** (**`train_test_split`**).

Partie G : Régression (prédiction du Chiffre d’Affaires)

25. Entraîner un modèle de **régression linéaire** pour prédire **`ChiffreAffaires`** à partir de **`Quantité`**, **`PrixUnitaire`**, **`Catégorie`**, **`Magasin`**.

26. Évaluer la performance du modèle. **(Score)**

27. Tester une autre méthode (ex. **`RandomForestRegressor`**) et comparer les résultats.