# Exercice : Web Scraping, Analyse de Données et Visualisation

## Objectifs pédagogiques

* Comprendre et appliquer les principes du web scraping avec BeautifulSoup et Requests
* Manipuler des données efficacement avec NumPy et Pandas
* Optimiser les calculs numériques à l’aide des tableaux NumPy
* Gérer et analyser des jeux de données en utilisant Pandas
* Créer des visualisations pertinentes et dynamiques avec Matplotlib et Seaborn

## Contexte

Vous disposez de trois fichiers HTML mensuels (janvier, février, mars 2025) contenant chacun un tableau de données de ventes (40 enregistrements par mois). Chaque enregistrement comporte les champs : Date, Product ID, Name, Category, Qty, Unit Price (€), Discount (%), Salesperson et Region.

Fichiers fournis :

* • index.html
* • sales\_january.html
* • sales\_february.html
* • sales\_march.html

## Partie 1 – Extraction & Consolidation

1. Écrire une fonction fetch\_sales(month\_url: str) -> list[dict] qui :  
 • Envoie une requête GET à month\_url avec Requests  
 • Parse le HTML avec BeautifulSoup  
 • Extrait chaque ligne du tableau et renvoie une liste de dictionnaires contenant les champs appropriés.  
2. Appliquer fetch\_sales aux trois fichiers et concaténer les résultats.  
3. Construire un DataFrame Pandas à partir de la liste et sauvegarder au format CSV.

## Partie 2 – Nettoyage & Analyse

1. Charger le CSV et parser la colonne date.  
2. Ajouter une colonne revenue = qty \* unit\_price \* (1 - discount/100) et extraire le mois.  
3. Calculer des statistiques descriptives : moyenne, médiane, écart-type, min, max sur revenue et discount.  
4. Grouper par produit pour obtenir mean, median et std sur qty et revenue.  
5. Créer un DataFrame prod\_info (product\_id → category) et fusionner avec df pour vérifier la cohérence.  
6. Analyser les ventes par région : CA total et nombre de ventes, puis identifier les 3 meilleures régions.

## Partie 3 – Visualisation

1. Histogramme de la distribution de revenue (Matplotlib).  
2. Boxplot de discount par catégorie (Seaborn).  
3. Barplot du CA moyen par produit (Seaborn).  
4. Lineplot de l’évolution mensuelle du CA global.  
5. Heatmap CA total par région × mois (Seaborn).  
6. Scatterplot qty vs revenue coloré par région (optionnel : pairplot).