TP - Analyse de données avec Pandas, Matplotlib et Seaborn

ali.zainoul.az@gmail.com

4 juin 2025

Contexte

Une entreprise de e-commerce souhaite mieux comprendre le comportement d'achat de ses clients en Europe. Vous disposez pour cela d'un jeu de données réel disponible publiquement, contenant l'historique des ventes d'un magasin en ligne basé au Royaume-Uni.

Lien du jeu de données :
`https://archive.ics.uci.edu/ml/datasets/Online+Retail`

Le fichier contient notamment les colonnes suivantes :
— `InvoiceNo` : numéro de facture
— `StockCode` : code produit
— `Description` : nom du produit
— `Quantity` : quantité commandée
— `InvoiceDate` : date et heure de la commande
— `UnitPrice` : prix unitaire (en livres sterling)
— `CustomerID` : identifiant du client
— `Country` : pays du client

Objectif du TP

Explorer, visualiser et comprendre les données à l'aide de la bibliothèque `pandas` pour le traitement, et `matplotlib/seaborn` pour les visualisations.

Instructions de départ

```
import pandas as pd

df = pd.read_excel("data/Online_Retail.xlsx")

df.head()
```

Tâches demandées

1. **Nettoyage des données**
   — Identifier et supprimer les lignes contenant des valeurs manquantes.
   — Supprimer les commandes avec des quantités négatives ou nulles.

— Vérifier le type des colonnes, notamment les dates.

2. **Analyse exploratoire**
   — Calculer le nombre total de transactions par pays.
   — Identifier les 10 produits les plus vendus en termes de quantité.
   — Créer une colonne calculant le prix total de chaque ligne (quantité × prix unitaire).
   — Visualiser la répartition des ventes par pays à l'aide d'un diagramme à barres.
3. **Analyse temporelle**
   — Convertir la colonne `InvoiceDate` au bon format temporel.
   — Extraire le mois ou la semaine pour observer l'évolution des ventes dans le temps.
   — Tracer l'évolution mensuelle ou hebdomadaire du chiffre d'affaires.
4. **Analyse client**
   — Calculer le panier moyen par client (total des ventes / nombre de commandes).
   — Identifier les clients les plus rentables.
   — Visualiser la distribution des montants dépensés par client à l'aide d'un boxplot.
5. **Visualisations complémentaires (Seaborn)**
   — Corrélation entre la quantité achetée et le prix unitaire.
   — Utiliser un heatmap pour afficher les ventes par jour de la semaine et heure de la journée.

Recommandations

— Utilisez `groupby()` et `agg()` pour effectuer des agrégations.
— Pensez à créer des colonnes auxiliaires comme le mois, le jour ou l'heure à partir des dates.
— Soignez vos graphiques : ajoutez des titres, légendes et axes clairs.
— Utilisez les options de style de Seaborn pour améliorer la lisibilité.

Bibliothèques nécessaires

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```