HESTIM
ENGINEERING & BUSINESS SCHOOL

18/05/2025

# TP: Data Mining avec NumPy, Pandas et Matplotlib

**Objectifs**

Dans ce TP, vous allez :

- Manipuler des données avec **Pandas**
- Effectuer des calculs avec **NumPy**
- Visualiser les données avec **Matplotlib**

## 1. Chargement et exploration des données

### 1.1 Importation des bibliothèques

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

### 1.2 Chargement d'un jeu de données

Utilisons un dataset fictif sur les ventes d'une entreprise:

```
data = {
   'Produit': ['A', 'B', 'C', 'D', 'E'],
   'Prix': [10, 20, 15, 30, 25],
   'Quantité': [100, 50, 80, 40, 60]
}
df = pd.DataFrame(data)
```

### 1.3 Aperçu des données

```
print(df.head())  # Affiche les premières lignes du dataset
print(df.info())  # Informations sur les colonnes et types de données
```

## 2. Manipulation des données avec Pandas

### 2.1 Ajout d'une nouvelle colonne : chiffre d'affaires

```
df['Chiffre_Affaires'] = df['Prix'] * df['Quantité']
print(df)
```

**2.2 Filtrage des données**

Affichons uniquement les produits avec un chiffre d'affaires > 1000 :

```
produits_rentables = df[df['Chiffre_Affaires'] > 1000]
print(produits_rentables)
```

**3. Analyse statistique avec NumPy**

**3.1 Calcul de statistiques basiques**

```
moyenne_ca = np.mean(df['Chiffre_Affaires'])
mediane_ca = np.median(df['Chiffre_Affaires'])
max_ca = np.max(df['Chiffre_Affaires'])
min_ca = np.min(df['Chiffre_Affaires'])
print(f"Moyenne CA: {moyenne_ca}, Médiane CA: {mediane_ca}, Max CA: {max_ca}, Min CA: {min_ca}")
```

**3.2 Détection des valeurs aberrantes**

Utilisons l'écart-type pour identifier des valeurs atypiques :

```
std_ca = np.std(df['Chiffre_Affaires'])
aberrantes = df[df['Chiffre_Affaires'] > (moyenne_ca + 2 * std_ca)]
print("Valeurs aberrantes:")
print(aberrantes)
```

**4. Visualisation des données avec Matplotlib**

**4.1 Histogramme des chiffres d'affaires**

```
plt.hist(df['Chiffre_Affaires'], bins=5, color='blue', edgecolor='black')
plt.xlabel('Chiffre d'affaires')
plt.ylabel('Fréquence')
plt.title('Distribution du chiffre d'affaires')
plt.show()
```

**4.2 Diagramme en barres des ventes par produit**

```
plt.bar(df['Produit'], df['Chiffre_Affaires'], color='green')
plt.xlabel('Produit')
plt.ylabel('Chiffre d'Affaires')
plt.title('Chiffre d'Affaires par Produit')
plt.show()
```

**4.3 Nuage de points : Prix vs Quantité**

```
plt.scatter(df['Prix'], df['Quantité'], color='red')
plt.xlabel('Prix')
plt.ylabel('Quantité Vendue')
plt.title('Relation entre Prix et Quantité')
plt.show()
```

**5. Questions:**

1. Quels sont les produits les plus rentables selon les données ?
2. Y a-t-il une corrélation entre le prix et la quantité vendue ? Justifiez avec les visualisations.
3. Quelle est la valeur maximale du chiffre d’affaires et quel produit l’a générée ?
4. Quelle est la distribution des chiffres d’affaires ? Que pouvez-vous en conclure ?