Année universitaire	**2021-2022**		
Département	**Informatique**	**Année**	**L3**
Matière	**LIFO65**		
Enseignant	**Khalid Benabdeslem, Haytham Elghazel, Mehdi Hennequin**		
Intitulé TD/TP :	**TP1 Analyse de données (ACP) avec Python**		
Contenu	• **Réduction de dimensions (ACP)** • **Visualisation**		

Pour ce TP, Il faut rendre votre notebook python sur claroline. Sous la forme TP1_ADM_Nom_Prenom.ipynb.

Dans ce TP, vous allez expérimenter des algorithmes de traitement de données pour répondre à différents problèmes liés à l'analyse de données multidimensionnelles avec le langage Python.

Python est un langage de programmation très polyvalent et modulaire, qui est utilisé aussi bien pour écrire des applications comme YouTube, que pour traiter des données scientifiques. Par conséquent, il existe de multiples installations possibles de Python. L'utilisateur débutant peut donc se sentir dérouté par l'absence d'une référence unique pour Python scientifique. Le plus simple pour ce TP est d'installer, la suite scientifique Anaconda développée par l'entreprise Continuum (http://continuum.io/downloads.html). Anaconda rassemble tout le nécessaire pour l'enseignement de Python scientifique : le langage Python et ses modules scientifiques. Sur le plan des packages Python, vous allez utiliser **Scikit-learn**. Cette libraire montre dans cette situation tout son intérêt. La plupart des techniques récentes d'apprentissage sont en effet expérimentées avec Scikit-learn et le plus souvent mises à disposition de la communauté scientifique.

Pour plus de détails concernant :

- le langage Python vous pouvez aller sur le site suivant : http://www.python-course.eu/index.php
- la librairie Scikit-learn vous pouvez aller sur le site suivant : http://scikit-learn.org

Pour lancer le notebook Python, il faut taper la commande **jupyter notebook** dans votre dossier de travail. Une fenêtre va se lancer dans votre navigateur pour ouvrir l'application Jupyter. Créer un nouveau notebook Python et taper le code suivant dans une nouvelle cellule :

```
import numpy as np
np.set_printoptions(threshold=10000,suppress=True)
import pandas as pd
import warnings
import matplotlib.pyplot as plt
warnings.filterwarnings('ignore')
```

Le fichier **"villes.csv"** comporte 32 villes françaises décrites par les températures moyennes dans les 12 mois de l'année.

L'objectif dans cette partie est de représenter graphiquement le plus d'informations possibles contenues dans ce fichier de données et de déceler une éventuelle segmentation topologique des villes.

1. Importer ce jeu de données avec la librairie **pandas** (*c.f.* ***read_csv***)

```
data = pd.read_csv('./villes.csv', sep=';')
X = data.iloc[:, 1:13].values
labels = data.iloc[:, 0].values
```

2. Réaliser une Analyse en Composantes Principales (module **PCA** de Scikit-learn) sur ce jeu de données centrées réduites (**StandardScaler**)
    - Quel est le nombre d'axes à retenir pour conserver un minimum de 90% de l'information représentée dans le nuage initial.
    - Donner une interprétation des deux premiers axes principaux.

- En suivant le code ci-dessous, donner une visualisation graphique des villes projetées dans le plan principal.

X_pca étant la matrice des données transformées par l'ACP, **labels** étant le vecteur contenant le nom des instances (ici les villes).

```
import matplotlib
plt.scatter(X_pca[:, 0], X_pca[:, 1])
for l, x, y in zip(labels, X_pca[:, 0], X_pca[:, 1]):
   plt.annotate(l, xy=(x, y), xytext=(-0.2, 0.2), textcoords='offset points')
plt.show()
```

- Essayer d'analyser les positions et oppositions des villes sur le plan projeté. Avec les éléments que vous avez, identifiez visuellement une typologie des états.
- Définir une fonction permettant de regrouper toutes les procédures précédentes.

3. Appliquer la fonction précédente sur le jeu de données **"crimes.csv".** Il s'agit des statistiques de criminalité dans 50 états américains. Dans chaque état, sept types de crimes ou délits sont repérés par leurs nombres annuels de faits constatés rapportés sur 100 000 habitants : *meurtres* (**Meutre**), *enlèvements* (**Rapt**), *vols avec violence*(**Vol**), *agressions* (**Attaque**), *viol* (**Viol**), *vols peu importants* (**Larcin**), *vols de voitures* (**Auto_Theft**). Interpréter et comparer les résultats obtenus pour ce Jeu de données. Avec les éléments que vous avez, peut-on visuellement identifier une typologie des individus pour ce jeu de données.