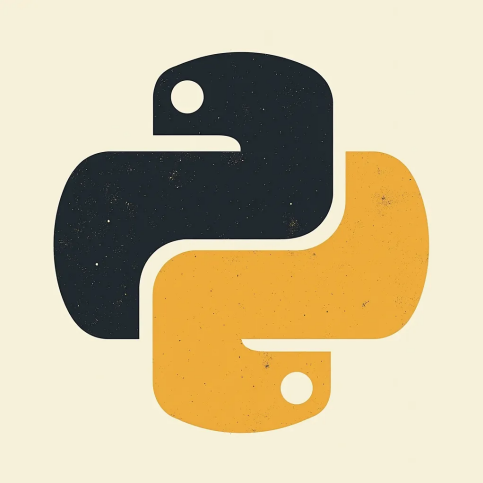
Travaux Pratiques 1 – Introduction à l'analyse de données

Cours Big Data – L3



Objectifs du TP

Ce premier travail pratique a pour objectif de vous permettre de :

- prendre en main l'environnement de travail (`Python`, `Jupyter Notebook`) ;
- manipuler un jeu de données réel ;
- identifier les problèmes classiques de qualité des données ;
- effectuer un premier nettoyage et une exploration simple des données.

Ce TP n'est **pas noté** mais doit être **rendu** sur Community, il constitue une étape essentielle pour la suite du cours et les projets. Le rendu permet de vérifier l'engagement dans la démarche d'apprentissage et d'apporter un retour individualisé.

Pour les séances de travaux pratiques, l'environnement *Jupyter Notebook* avec kernel *Python* sera utilisé. Pour les premiers TP (TP1 et TP2), il est possible d'utiliser l'environnement en ligne :

`https://jupyter.org/try-jupyter/lab/`

⚠ Attention ⚠

Cet environnement ne nécessite aucune installation et permet une prise en main rapide. Attention toutefois : les sessions sont temporaires et les fichiers doivent être sauvegardés localement avant la fermeture du navigateur. Il est fortement recommandé de sauvegarder régulièrement son travail et de conserver une copie locale des notebooks.

Une alternative recommandée consiste à installer un environnement Python local via *Miniconda*. Cette installation peut être réalisée directement dans le répertoire personnel de l'utilisateur, et

permet de disposer d'un environnement Jupyter persistant. Cette solution est fortement conseillée pour les TP suivants et pour la réalisation des projets.

Jeu de données

Le jeu de données utilisé pour les deux TPs est mis à disposition sur Community. Il s'agit d'un jeu de données réel, conçu pour refléter des situations couramment rencontrées en analyse de données.

Il se présente sous la forme d'une table contenant **499 tweets** relatifs aux résolutions du Nouvel An 2015. Chaque enregistrement correspond à un tweet unique et inclut notamment la date et l'heure de publication, des informations de localisation, le texte original du tweet ainsi que la catégorie de résolution associée.

Comme tout jeu de données réel, les données fournies peuvent présenter :

— des valeurs manquantes ;
— des incohérences de format ou de contenu ;
— des valeurs aberrantes.

Ces imperfections sont **volontaires** et font pleinement partie du travail attendu, l'objectif étant de confronter les étudiants à des données non idéales et de les amener à mettre en œuvre des méthodes de nettoyage et d'analyse adaptées.

Environnement de travail

Le travail devra être réalisé dans un **notebook Jupyter** en utilisant Python et les bibliothèques suivantes :

— `pandas`
— `numpy`
— `matplotlib` et/ou `seaborn`

Travail demandé

1. Chargement et inspection des données

— Charger le jeu de données dans un notebook Jupyter.
— Afficher les premières lignes du jeu de données.
— Indiquer le nombre de lignes et de colonnes.
— Identifier les types des différentes variables.

2. Description du jeu de données

Décrire brièvement le jeu de données :

— signification des principales variables ;
— type de données (numériques, catégorielles, temporelles, etc.) ;
— présence éventuelle de valeurs manquantes.

3. Nettoyage des données

Effectuer un premier nettoyage du jeu de données :

— gestion des valeurs manquantes (suppression ou remplacement justifié) ;

— correction ou suppression des valeurs aberrantes si nécessaire ;
— conversion des types de variables si besoin (dates, nombres, catégories).

Les choix effectués doivent être brièvement justifiés dans le notebook.

4. Analyse exploratoire simple

Réaliser une première analyse exploratoire :

— statistiques descriptives (moyenne, médiane, minimum, maximum, etc.) ;
— visualisations simples (histogrammes, courbes, diagrammes en barres).

Les graphiques doivent être lisibles et accompagnés d'un court commentaire.

5. Sauvegarde des données nettoyées

Exporter le jeu de données nettoyé dans un nouveau fichier (`CSV` par exemple), qui sera réutilisé dans les prochains TP.

Livrables attendus

À la fin du TP, vous devez fournir :

— un notebook Jupyter clairement commenté ;
— le fichier de données nettoyé ;
— une courte conclusion (quelques lignes) résumant les observations principales.

Remarques importantes

— Ce TP n'est pas noté, mais il prépare directement le **TP 2** et les projets.
— La clarté du code et des commentaires est essentielle.
— Les résultats obtenus doivent être interprétés, pas seulement affichés.