Travaux Pratiques 2 – Analyse exploratoire et statistiques

Cours Big Data – L3

Objectifs du TP

Ce deuxième travail pratique a pour objectif de :

- approfondir la manipulation de données avec `pandas` et `numpy` ;
- réaliser une analyse exploratoire complète ;
- produire des visualisations pertinentes ;
- interpréter les résultats obtenus de manière critique.

Ce TP n'est **pas noté**, mais son rendu est **obligatoire** sur Community. Il prépare directement le **Projet 1**. Le rendu permet de vérifier l'engagement dans la démarche d'apprentissage et d'apporter un retour individualisé.

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`

Vous pouvez réutiliser l'environnement en ligne Jupyter ou un environnement local.

Jeu de données

Vous utiliserez le **jeu de données nettoyé lors du TP1**. Si nécessaire, vous pouvez améliorer ou corriger le nettoyage effectué précédemment.

Travail demandé

1. Chargement des données

— Charger le jeu de données nettoyé.
— Vérifier la cohérence globale des données.

2. Statistiques descriptives

Calculer et commenter :

— statistiques de base (moyenne, médiane, minimum, maximum) ;
— mesures de dispersion (variance, écart-type) ;
— statistiques pertinentes selon le jeu de données.

3. Analyse exploratoire

Réaliser une analyse exploratoire du jeu de données afin de mettre en évidence des éléments d'intérêt. L'analyse s'appuie notamment sur les points suivants :

— choisir une ou plusieurs variables quantitatives et étudier leur évolution ;
— mettre en évidence des tendances globales ou des variations marquées ;
— comparer les données selon au moins un critère pertinent (par exemple : catégories, groupes, périodes ou zones géographiques).

Chaque étape de l'analyse est accompagnée d'un court commentaire expliquant les choix effectués et les résultats observés.

4. Visualisations

Produire des visualisations adaptées, par exemple :

— histogrammes ;
— courbes d'évolution ;
— diagrammes en barres ;
— nuages de points.

Chaque figure doit :

— comporter un titre ;
— avoir des axes correctement nommés ;
— être accompagnée d'un commentaire interprétatif.

5. Question d'analyse

Formuler une **question simple** à laquelle les données permettent de répondre, par exemple :

— Comment évolue une variable donnée au cours du temps ?
— Existe-t-il une différence significative entre deux groupes ?

À titre d'exemple, les questions suivantes peuvent être explorées :

— Quelle est la catégorie de résolution la plus (ou la moins) représentée ?
— Quelle catégorie de résolution a été la plus retweetée ?
— À quelle heure de la journée les résolutions sont-elles le plus souvent tweetées ?
— Quelle région ou quel État des États-Unis publie le plus de résolutions ?

Présenter une réponse argumentée à une question principale. Des analyses complémentaires peuvent être ajoutées si pertinentes.

Livrables attendus

Vous devez rendre :

— un notebook Jupyter clair et structuré ;
— des graphiques lisibles et commentés ;
— une courte conclusion synthétisant les résultats principaux.

Remarques importantes

— Ce TP n'est pas noté, mais le rendu est obligatoire.
— La clarté du code et des commentaires est essentielle.
— Les résultats doivent être interprétés et discutés, pas seulement présentés.
— Ce travail constitue une base directe pour les **Projets**.