

**Module : Fouille de données**

**TP N° 1 : Découverte du Dataset, Analyse Statistique Descriptive et EDA (Exploratory Data Analysis)**

**Prérequis :**

- Connaissances de base en Python.
- Installation des bibliothèques Python : pandas, numpy.

**Dataset :**

Nous utiliserons le dataset "House Prices - Advanced Regression Techniques". Ce dataset est très utilisé dans le domaine de l'intelligence artificielle pour des prédictions de prix immobiliers.

Lien du dataset: <https://www.kaggle.com/datasets/yasserh/housing-prices-dataset>

**Chargement du Dataset**

1. Installer la Bibliothèque : Pandas.
2. Importer la bibliothèque.
3. Charger les Données.
4. Faire un aperçu des premières lignes du dataset.

**Exploration de la Structure des Données**

5. Combien de lignes et de colonnes contient le dataset ?
6. Quelles sont les colonnes présentes dans le dataset ?
7. Quels sont les types de données présents dans chaque colonne ?
8. Combien de colonnes sont numériques ? Combien sont catégoriques ?

**Informations et Statistique Descriptive du Dataset**

9. Quel est l'état général du dataset ? Y a-t-il des valeurs manquantes ?
10. Quels sont les indicateurs statistiques globaux des colonnes numériques ?
11. Quels sont le premier et le troisième quartile de chaque colonne numérique ?
12. Quelle est la moyenne et la médiane des prix des maisons ?
13. Quelle est la répartition des catégories pour chaque variable catégorielle ?
14. Combien de valeurs manquantes y a-t-il dans chaque colonne ?
15. Quels sont les 5 quartiers avec les prix de vente moyens les plus élevés ?
16. Quel est le prix moyen par statut d'ameublement
17. Quelle est la moyenne des prix pour les maisons ayant :
  - La climatisation (airconditioning = yes) ?
  - Un sous-sol (basement = yes) ?
  - Un accès à la route principale (mainroad = yes) ?
18. Quelle est la moyenne des prix en fonction du nombre d'étages