

# TP 2 Prétraitement des données

1. Importer les bibliothèques pandas, numpy, seaborn, matplotlib.pyplot, scipy, sklearn
2. Charger les données à partir du dataset diamonds.csv
3. Afficher les données.
4. Afficher le shape des données
5. Afficher les informations sur le dataset
6. Afficher les colonnes et supprimer les colonnes insignifiantes si elles existent.
7. Chercher et traiter les valeurs dupliquées.
8. Regarder la distribution des données à l'aide d'un histogramme.
9. Chercher et traiter les valeurs manquantes
10. Afficher les valeurs aberrantes à l'aide d'un boxplot.
11. Quantifier les valeurs aberrantes
12. Gérer les valeurs aberrantes pour l'attribut price
13. Afficher le shape des données
14. Calculer le taux des valeurs aberrantes sur le dataframe
15. Analyse Univariée :
  - a. Afficher les caractéristiques statistiques de l'attribut price
  - b. Calculer la dispersion des prix
  - c. Afficher la courbe de densité des prix
16. Analyse bivariée :
  - a. Faire un lmplot du carat par rapport au prix. Que remarquez-vous
  - b. Faire un lmplot du depth par rapport au prix. Que remarquez-vous
  - c. Afficher la matrice de corrélation- heat map
  - d. Est ce les coupes influent sur les prix ? faire un boxplot pour répondre à cette question
  - e. est ce que les couleurs aussi influent sur les prix ? faire un boxplot pour répondre à cette question
17. Diviser les données en Var Explicatives (X) et Var à Expliquer (Y)
18. Encoder les variables qualitatives
19. Diviser les données en training et test set