

## Exercice 1 : Analyse de Données Massives

A partir du fichier [diamonds.csv](#), analyser les données suivant le déroulement classique

1. Description de chaque variable
2. Recherche des liens entre le prix (price) et les autres variables

### Description des variables

Dans ce cadre, il faut décrire chaque variable séparément. Voici la répartition des variables et les étapes à réaliser pour chaque type de variable :

- *quantitative* : carat, depth, table, price, x, y, z
  - Moyenne, écart-type, médiane, minimum et maximum
  - Histogramme pour avoir un aperçu de la distribution, et éventuellement comparer à une gaussienne
    - Test de normalité à ajouter
  - Boîte à moustaches pour estimer la présence ou non d'outliers
- *qualitative* : cut, color, clarity
  - Table d'effectifs et de pourcentages
  - Diagramme en barres (diagramme circulaire à éviter ici, car trop de modalités)

Nous allons faire une fonction permettant la description d'une variable (en fonction de son type bien évidemment), et l'utiliser pour chacune des colonnes du jeu de données.

### Lien avec price

Ici, en fonction du type de la variable pour laquelle estimer le lien avec price, les opérations ne sont pas les mêmes à effectuer.

- *quantitative* : carat, depth, table, x, y, z
  - Corrélation, avec test de corrélation
  - Nuage de points
- *qualitative* : cut, color, clarity
  - ANOVA
  - Boîtes à moustaches pour chaque modalité

**Exercice 2** : Un jeu consiste à lancer une pièce de monnaie ; tant que Pile apparaît, on relance la pièce; si c'est Face qui apparaît le jeu se termine. Les instructions :

`from random import randint`

`a = randint(0,1)`

permettent d'affecter à la variable a la valeur 0 ou 1 de façon équiprobable et de simuler ainsi le jeu de Pile ou Face.

- Ecrire un programme qui compte le nombre de fois où Pile est apparu au cours d'un jeu.
- Compléter le programme afin de simuler cent fois ce jeu et d'afficher en sortie le nombre maximal de fois où Pile est apparu au cours d'un jeu.

**Exercice 2 :** On dispose des résultats d'une enquête concernant l'âge et les loisirs d'une population de 20 personnes :

|   |
|---|
| Age : 12 14 40 35 26 30 30 50 75 50 30 45 25 55 28 25 50 40 25 35 |
| Loisir : S S C C S T T L L L T C C C S L L C T T                  |

Codification : S : Sport, C : Cinéma, T : Théâtre, L : Lecture

- Faire l'étude du caractère « âge » : dresser le tableau statistique (effectifs, effectifs cumulés), calculer les valeurs de tendance centrale et ceux de la dispersion et tracez le diagramme en bâtons et la boîte à moustaches de cette distribution
- Faire l'étude du caractère « Loisir » dresser le tableau statistique, déterminer le mode et tracez le diagramme en bâtons, boîte à moustache et le diagramme à secteurs.

**Exercices 3 :** On observe 100 fois le nombre d'arrivées (variable X) de clients à un bureau de poste pendant un intervalle de temps (10 minutes) et on obtient les valeurs suivantes :

1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2  
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3  
 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4  
 4 4 4 4 4 5 5 5 5 5 5 5 6 6 6 6 6 6

- Dresser le tableau statistique de la distribution de la variable X (effectifs cumulés...).
- Calculer les valeurs de tendance centrale de la distribution : la moyenne, le mode et les trois quartiles Q1, Q2 et Q3.
- Calculer les valeurs de la dispersion de la distribution : variance, l'écart type et l'intervalle interquartile.
- Tracer le diagramme en bâtons et la boîte à moustaches de cette distribution.