TD 1

Exercice 1 : Tâches de FDD

Déterminer si chacune des activités suivantes est une tâche de fouille de données :

1. Diviser les clients d'une compagnie selon leur genre.
2. Calculer les ventes totales d'une compagnie.
3. Prédire les prix futurs d'un produit en utilisant les données passées.
4. Surveillance d'activités sismiques.
5. Extraire les fréquences d'un signal sonore.

Exercice 2 : Statistiques descriptives

Dans une étude portant sur 18 personnes adultes, nous avons obtenu les résultats suivants

age	23	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61
% mg	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

1. Calculer la moyenne, la médiane, l'écart-type de l'âge et du pourcentage de masse graisseuse
2. Calculer la distance Euclidienne entre l'individu 1 et l'individu 3. Cette distance est-elle adaptée ?

Exercice 3 : Mesures de similarité

Il est important de définir des mesures de similarité en analyse de données. Selon la métrique utilisée, les rangs entre objets varient. On suppose que l'on a les données 2D suivantes

	p_1	p_2
$\mathbf{x}_1$	1.5	1.7
$\mathbf{x}_2$	2	1.9
$\mathbf{x}_3$	1.6	1.8
$\mathbf{x}_4$	1.2	1.5
$\mathbf{x}_5$	1.5	1.0

1. Si l'on considère un nouveau point $\mathbf{x} = (1.4, 1.6)$ requête, trier les données selon leur ressemblance avec $\mathbf{x}$ en utilisant les distances Euclidiennes, et similarité cosinus (définie par $x'y/||x||||y||$)

2. Normaliser les données de telle sorte que chaque point ait une norme égale à 1, et procéder au tri des points selon la distance Euclidienne.

Exercice 4 : Réduction par binning

On considère les âges suivants pour des individus : 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

1. Réaliser un filtrage par moyenne, puis par borne des ces données en utilisant une profondeur de bin égale à 3. Quels sont les effets de ce filtrage ?
2. Comment pourrait-on déterminer des anomalies dans ces données ?

Exercice 5 : Normalisation

Avant de traiter les données, il est souvent recommandé de les normaliser, afin que les valeurs puissent être comparables.

1. Quelles sont les intervalles de valeurs des méthodes *min/max*, *z-score* et *mise à l'échelle décimale* ?
2. Si on considère les données 200, 300, 400, 600, 1000, quelles seront leurs valeurs après normalisation ?

Exercice 6 : Discrétisation

On dispose de la somme dépensée par les douze derniers clients d'un commerce : 5, 10, 11, 13, 15, 35, 50, 55, 72, 91, 204, 215.

1. Partitionner ces données en 3 bins par
   – fréquences égales
   – largeurs égales
   – regroupement