TP - K-MEANS

Objectif

- Réaliser une segmentation client à l'aide de l'algorithme **K-Means**.
- Le but est d'identifier des groupes de clients ayant des comportements similaires à partir de leurs caractéristiques socio-économiques et comportementales.

Compétences visées

- Comprendre et appliquer un algorithme de **clustering non supervisé**.
- Savoir **préparer, normaliser et visualiser** des données.
- Déterminer le **nombre optimal de clusters**.
- **Interpréter** les résultats de segmentation.

Données

Téléchargez le Dataset **Mall Customers** sur Kaggle. Mall_Customers.csv

Chaque observation correspond à un client avec les variables suivantes :

- `CustomerID`
- `Gender`
- `Age`
- `Annual Income (k$)`
- `Spending Score (1-100)`

Travail demandé

Partie 1 – Préparation des données

1. Charger et explorer le dataset.
2. Vérifier la présence de valeurs manquantes et traiter si nécessaire.
3. Analyser la distribution des variables (histogrammes, statistiques descriptives).
4. Sélectionner les variables pertinentes pour la segmentation (justifier votre choix).
5. Normaliser ou standardiser les variables retenues.

Partie 2 – Analyse exploratoire

1. Visualiser les relations entre les variables (scatter plots, pairplots, heatmap de corrélation).
2. Identifier des tendances ou regroupements visuels éventuels.
3. Émettre des hypothèses sur le nombre de segments possibles.

Partie 3 – Application de K-Means

1. Appliquer K-Means pour différents nombres de clusters (k = 2 à 10).
2. Utiliser la **méthode du coude** pour choisir le nombre optimal de clusters.
3. Calculer le **score de silhouette** pour valider le choix de k.
4. Entraîner le modèle final avec le k choisi.
5. Ajouter une colonne “Cluster” au dataset avec les labels obtenus.

Partie 4 – Interprétation

1. Calculer la moyenne des variables par cluster.
2. Décrire le profil typique de chaque segment (âge, revenu, score de dépense…).
3. Donner un **nom** ou une **étiquette descriptive** à chaque segment.
   (Exemple : “Jeunes dépensiers”, “Clients fidèles à haut revenu”, etc.)
4. Représenter les clusters par des visualisations 2D et 3D.

Partie 5 – Extension (bonus)

1. Comparer K-Means avec un autre algorithme de clustering (DBSCAN, CAH ou GMM).
2. Évaluer la stabilité du modèle (répéter avec d’autres random seeds).
3. Réduire la dimension avec **PCA** avant clustering et analyser l’impact.
4. Proposer des recommandations business basées sur la segmentation.

Rendu attendu

- Le **code** commenté
- Un **rapport PDF** contenant :
    - Description claire des étapes suivies.
    - Justifications des choix (k, variables, interprétations).
    - Visualisations propres et lisibles.
    - Analyse critique des résultats.
- Le tout dans un repo **GIT**