PROJECT 4

**To know**

Kmeans : random points as centroid, points are grouped based on their distances to the centroids. A new centroid is then calculated for each cluster, based on the average values of the data points within that cluster. This process of assigning data points and recalculating centroids is iteratively performed until the centroids' positions become stable, and the clusters remain unchanged.

SSE : somme des distances euclédiennes au carré de chaque point par rapport au centroide.

The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

Dbscan : given a set of points in some space, it groups together points that are closely packed (points with many [nearby neighbors](https://en.wikipedia.org/wiki/Fixed-radius_near_neighbors)), and marks as outliers points that lie alone in low-density regions (those whose nearest neighbors are too far away).

The main idea behind OPTICS is to extract the clustering structure of a dataset by identifying the density-connected points. The algorithm builds a density-based representation of the data by creating an ordered list of points called the reachability plot. Each point in the list is associated with a reachability distance, which is a measure of how easy it is to reach that point from other points in the dataset. Points with similar reachability distances are likely to be in the same cluster.

Hierarchical : We begin by treating each data point as its own cluster. Then, we join clusters together that have the shortest distance between them to create larger clusters. This step is repeated until one large cluster is formed containing all of the data points.

##### Définir la stratégie d’élaboration d’un modèle d’apprentissage non supervisé, sélectionner et entraîner des modèles adaptés à une problématique métier afin de réaliser une segmentation ou une réduction de données.

* CE1 Les variables catégorielles identifiées ont été transformées en fonction du besoin (par exemple via OneHotEncoder ou TargetEncoder). → THERE IS NONE IN THIS PROJECT.
* CE2 Vous avez créé de nouvelles variables à partir de variables existantes.
* CE3 Vous avez réalisé des transformations mathématiques lorsque c'est requis pour transformer les distributions de variables.
* CE4 Vous avez normalisé les variables lorsque c'est requis.
* CE5 Vous avez défini la stratégie d’élaboration d’un modèle pour répondre à un besoin métier.
* CE6 Vous avez proposé un nombre de segments et une répartition adaptés à la problématique métier.
* CE7 Vous avez explicité la stratégie d'ajout de nouveaux clients.
* CE8 Vous avez pris en compte la nature des variables d'entrée dans le choix des algorithmes et de la distance.
* CE9 Vous avez testé et comparé plusieurs algorithmes.

##### Évaluer les performances des modèles d’apprentissage non supervisé selon différents critères (scores, temps d'entraînement, etc.) en adaptant les paramètres afin de choisir le modèle le plus performant pour la problématique métier.

* CE1 Vous avez choisi des métriques adaptées pour évaluer les valeurs candidates pour le nombre de segments (selon algorithme, par exemple : coefficient silhouette / elbow pour le KMeans, eps/min samples pour le DBScan).
* CE2 Vous avez évalué la forme des clusters.
* CE3 Vous avez évalué la stabilité des clusters à l’initialisation.
* CE4 Vous avez optimisé les hyper-paramètres pertinents dans les différents algorithmes.
* CE5 Vous avez justifié le choix final d'algorithme et d'hyper-paramètres.
* CE6 Vous avez analysé la stabilité du modèle dans le temps afin d’en assurer sa maintenance
* CE7 La convention PEP8 est respectée. Le code est commenté (commentaires réguliers, docstrings dans les fonctions) .