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.

tsne, viu map, pas du PCA vu qu-on perd de linfo -> pas trop important.

Au moins ne pas dire qu-on se base sur la visualisation pour faire des decisions.

Utilise le SSE/silhouette pour evaluer les clusters.

Tu peux egalement utiliser la distance nominale entre tous les points d-un cluster/intra-cluster/extra-cluster comme metriques additionels. A voir lesquels prennent pas trop de temps.

IMPORTANT

Prends une première période A avec un nombre fixe de clients sans commande dans le futur. Puis itère, prend les clients période A + 1 week avec la meme méthode SQL et entraine a chaque fois.

Puis compare labels période A avec chacune pour ARI.

tituscodes@gmail.com