PROJECT 4

Questions :

* Quand je ne considère que les orders et pas products per customer j’en ai qu’1 par customer. 0, pas 3 % comme David disait. Je fais quoi du coup avec ma fréquence ?
  + nb\_products = len(set([order['order\_id'] for order in customer\_orders]))
  + Le nombre de produits est peut-être un métrique intéressant pour la fréquence  vu le manque d’order per customers ?
* Can I exclude 36 clients (nb\_products >= 8) to make my PCA graph more lisible ? Most likely yes.
* Min samples for optics and dbscan 100 ?
* Forme et stabilite des clusters ? CE 2-3. Stability = if you run your kmeans or other algo multiple times you should have very similar results. You can then using the same PCA use different colors for the different runs. The forme is so that the points with the same color are in the same region and that if you run the algos multiple times you get the same regions.
* Il serait intéressant également de regarder l’évolution de la la distribution de vos features numériques dans le temps, en utilisant des méthodes graphiques ou un test de Kolmogorov-Smirnov ?
  + Two time periods, A and B. Train a model A on period A, model B on A+B. Then calculate cluster C1 on model A on period B, C2 on model B on period B. Then get ARI on diff C1-C2 to determine if C2 makes sense as the model B. ARI acceptable > 70 %. Check a time period of 1-2-3-4-… weeks to get an acceptable ARI. Do a graph with ARI and number of weeks.

Number of orders is not just the len() but some orders are from the same order.

Your algos are expected to crash with all the rows, take a random sample of like 50-100k with a special Pandas method to have it shuffled.

* Un **notebook de l'analyse exploratoire** (non cleané, pour comprendre votre démarche) → Afficher les distributions suffisent.

##### 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) .