Segmentez des clients d'un site e-commerce

...

Septembre 2019

Antoine LEPAGE - yop1001@gmail.com

Démarche

Contexte :

- Consultant pour Olist est une solution de vente sur les marketplaces en ligne

Objectifs :

- Fournir aux équipes marketing une segmentation des clients
- Proposition de contrat de maintenance de la segmentation

Données:

- 9 datasets présentant les transactions du 4 septembre 2016 au 29 août 2018

Démarche :

- Construction de l'échantillon et nettoyage
- Nettoyage
- Exploration
- Pré-processing / Recherche de la meilleure modélisation
- Fréquence de mise à jour

Construction de l'échantillon

Etape 1 - Former un dataset de toutes les transactions à partir des 1 dataset composé de 9 tables

Etape 2 - Nettoyage du dataset

- Regroupement des 71 catégories de produit en 17 catégories
- Suppression de 49 transactions sans données de géolocalisation
- Suppression de 537 transactions non abouties
- Suppression de la seule transaction de septembre 2018

Etape 3 - Former un dataset compilant les données par clients à partir des transactions

- Nombre d'articles achetés
- Somme totale dépensée
- Calcul du nombre de paniers par clients
- Nb moyen d'articles par panier par client
- Fréquence des achats
- Ancienneté du client
- Dépense par catégorie pour chaque client
- Note de satisfaction moyenne
- Nombre de commentaires
- nombre de vendeurs différents par client
- indice de fidélité à un vendeur

111 099 transactions

20 variables

93 616 clients

39 variables

Exploration

...

Etendue de l'étude

111 099
transactions

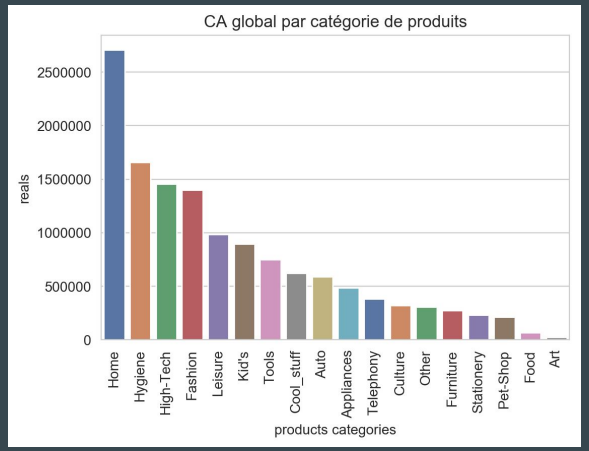
32 015 produits
différents vendus

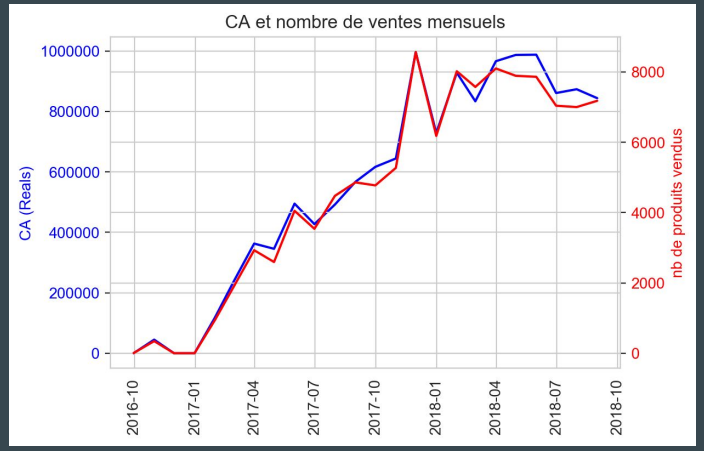
93 616
clients

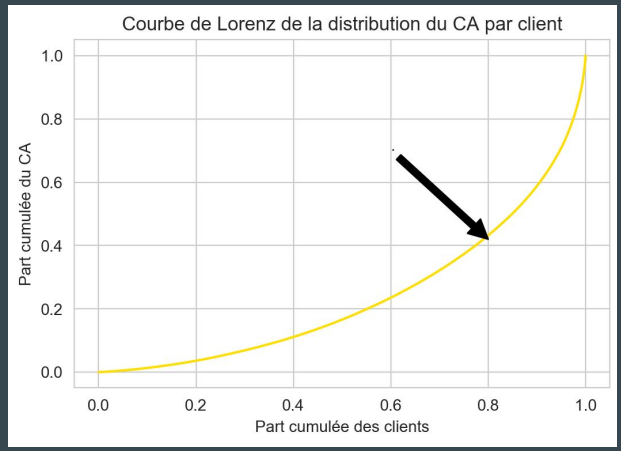
Sur 2 ans
09/2016 au 08/2018

Chiffre d'affaires

13 364 401 réals sur 2 ans

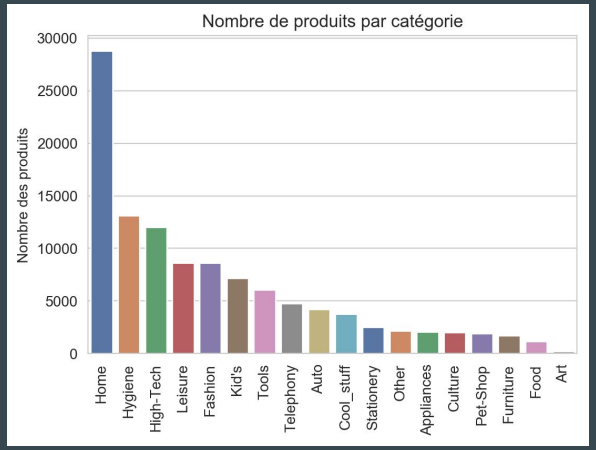


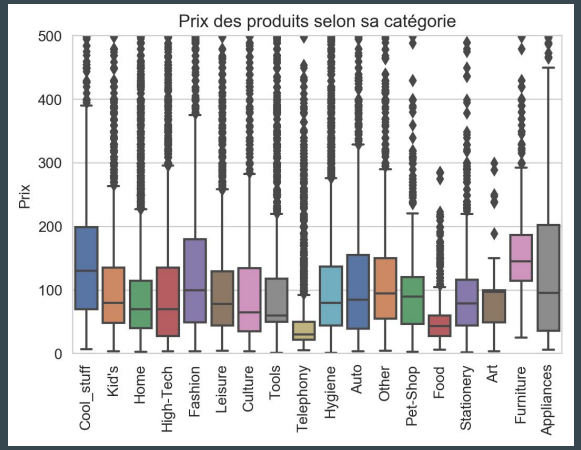


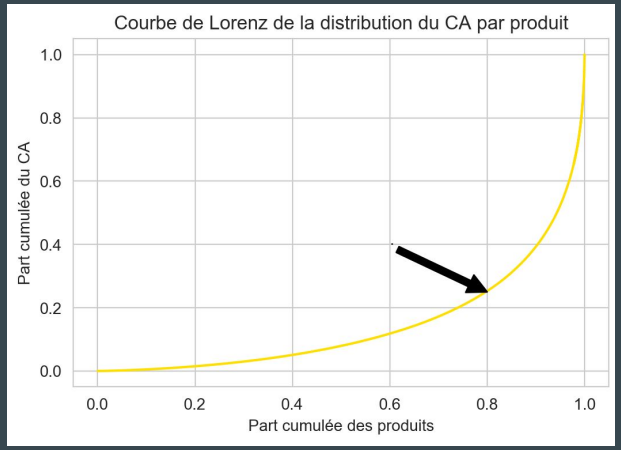


60 % du CA réalisé par 20 % des clients

Produits







20% des produits génèrent 75% du CA

Achats

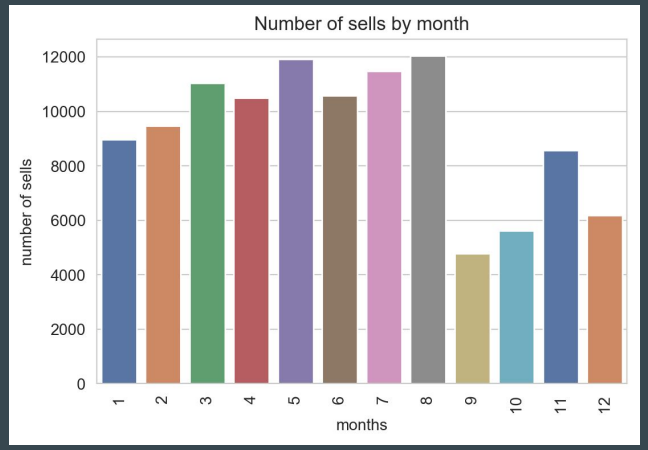
3 % des clients
ont commandé + d'une fois

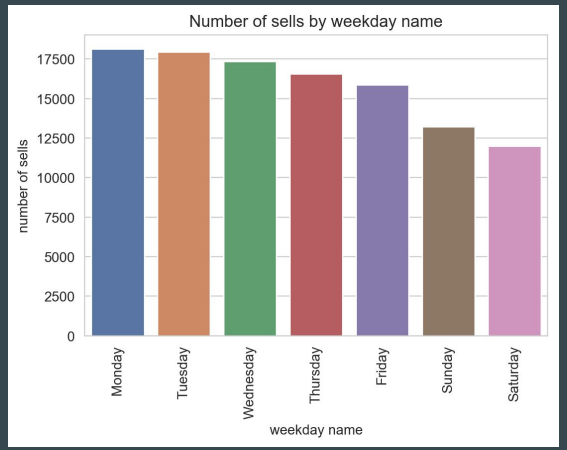
87% des clients
n'ont commandé qu'un seul article

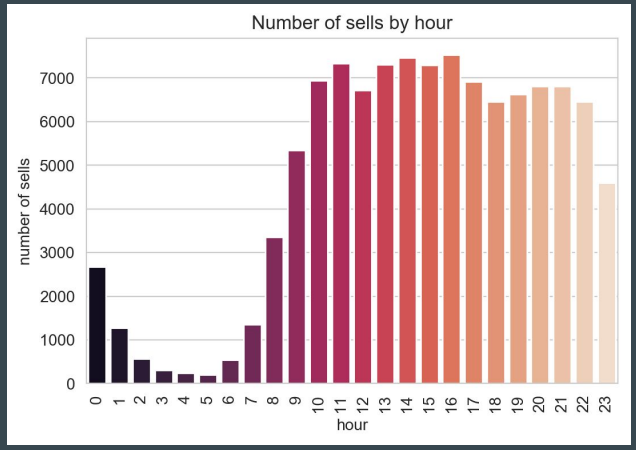
Panier moyen
138 réals

Satisfaction client
★★★★☆

Temporalité des achats





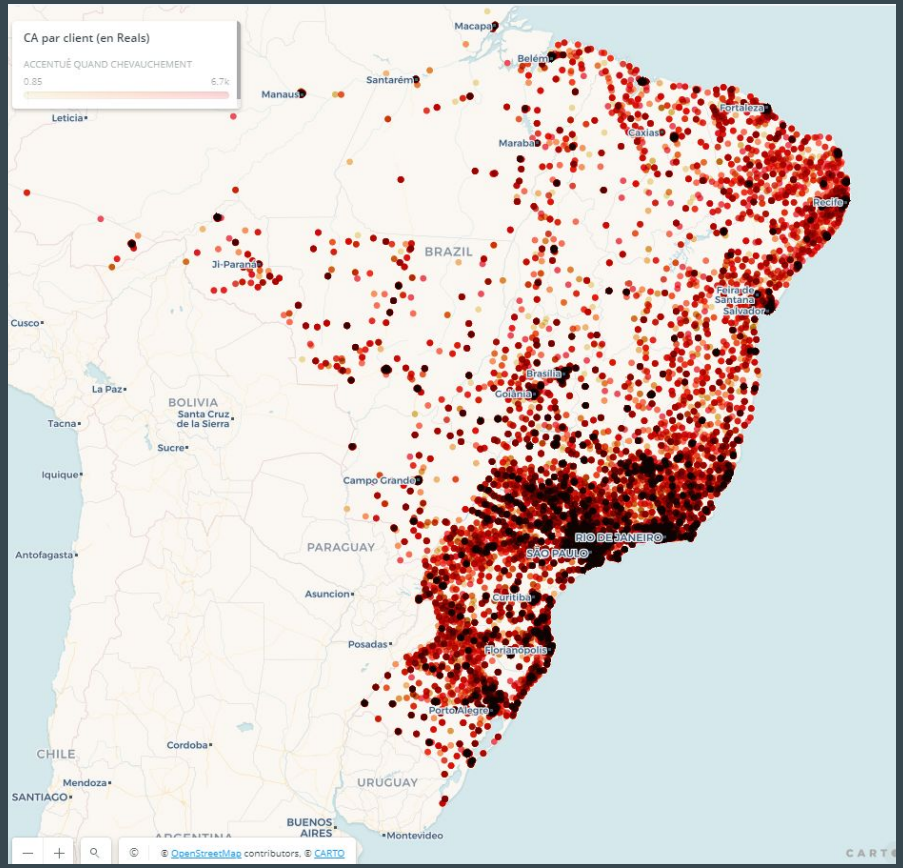


2 fois moins d'achats les 4 derniers mois

plutôt en semaine

entre 10h et 22h

Géolocalisation des clients



Modélisation

...

Stratégie

Objectif : Modéliser les données afin segmenter une base client

=> Apprentissage non supervisé

=> Clustering

Préprocessing

- Standardisation des données

Visualisation

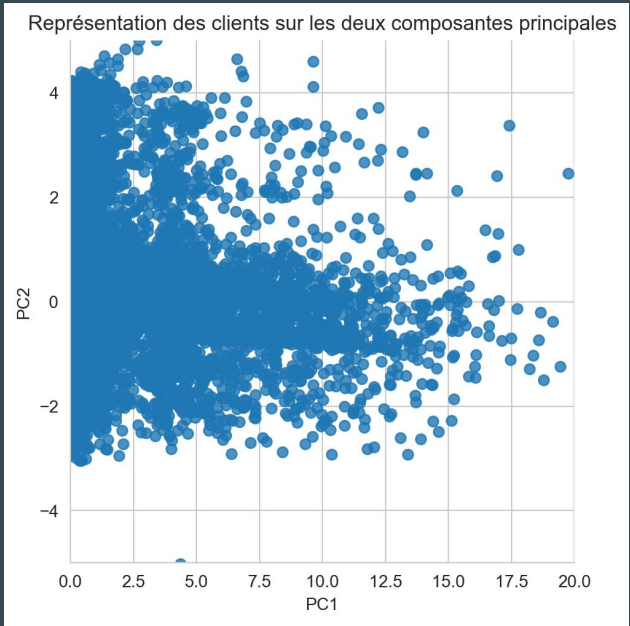
- ACP - 2 composantes principales : 42% de la variance

Modèles testés :

- Partitioning : KMeans
- Density : DBScan
- Hierarchical : Birch

Evaluation :

- Coefficient de silhouette
- Visualisation des groupes selon les 2 composantes principales (ACP)
- Interprétabilité des groupes



Choix des variables

Volume des dépenses :

- customer_unique_id
- nb_of_items_bought
- total_spent
- nb_of_baskets
- items_by_basket(mean)

Fréquence des achats

- time_length
- oldest_purchase
- purchase_frequency
- mean_purchase_frequency(days)
- td

Exprimer les dépenses dans les différentes catégories

- Appliances, Art, Auto, Cool_stuff, Culture, Fashion, Food, Furniture, High-Tech, Home, Hygiene, Kids, Leisure, Other, Pet-Shop, Stationery, Telephony, Tools

Ancienneté du client

- anciennete

Satisfaction client

- review_score
- nb_of_comments
- %_of_comments

Géolocalisation

- customer_zip_code_prefix
- customer_city
- customer_state,
- geolocation_lat
- geolocation_lng

Fidélité d'un client à un vendeur

- nb_of_sellers_used
- sellers_loyalty

Tests de 3 modèles

KMeans

Principes :

- non hiérarchique
- algorithme itératif qui minimise la somme des distances entre chaque individu et le centroïd.
- Le choix initial des centroïdes conditionne le résultat final.

Hyper-paramètre :

- n_clusters (elbow method)

DBScan

Principes :

- Clustering par densité
- Chemin pour passer de proche en proche en restant à l'intérieur du même cluster.
- dssdfs

Hyper-paramètres :

- eps=0.8
- min_samples=50

Birch

Principes :

- segmentation hiérarchique
- réduit la taille du jeu de données initial en le résumant sous la forme d'une structure hiérarchique à laquelle les points sont agrégés

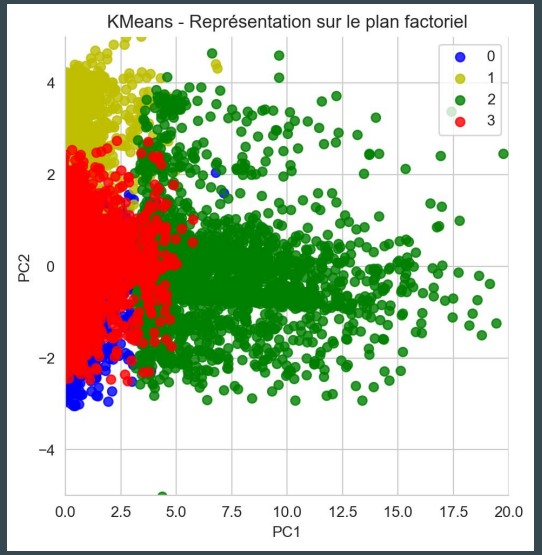
Hyper-paramètres :

- branching_factor=150
- n_clusters=5,
- threshold=1.3

Evaluations des modèles

KMeans

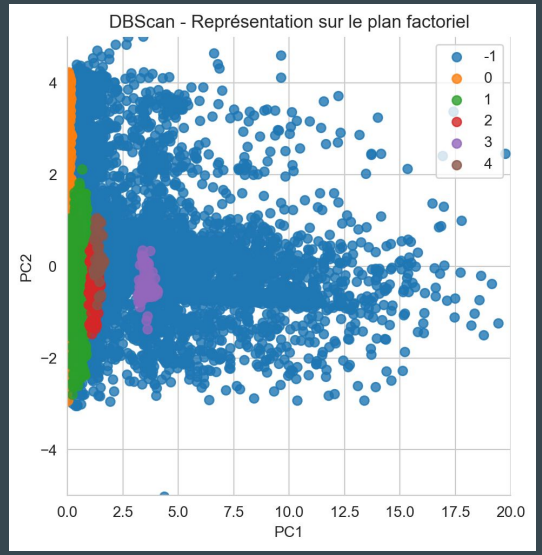
Coefficient de silhouette : 0.29



Interprétabilité possible

DBScan

Coefficient de silhouette : 0.38

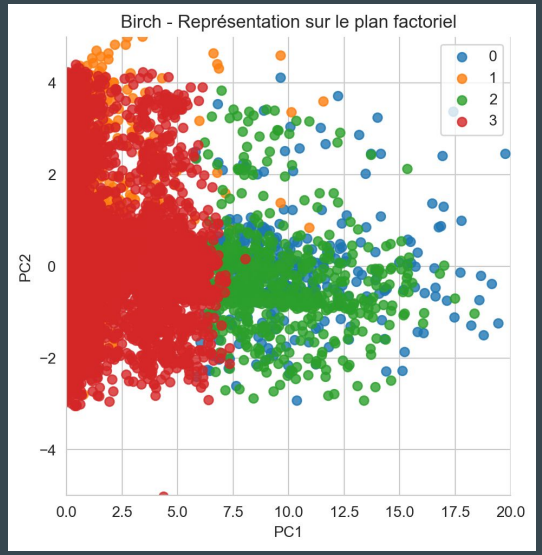


9% d'outliers

Interprétabilité difficile

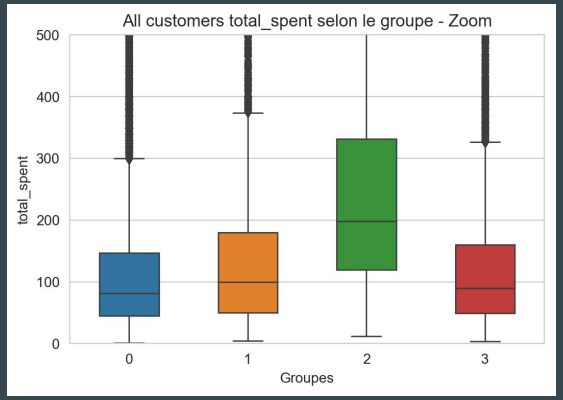
Birch

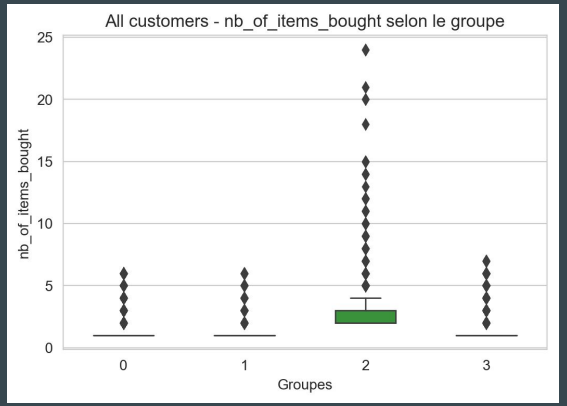
Coefficient de silhouette : 0.30

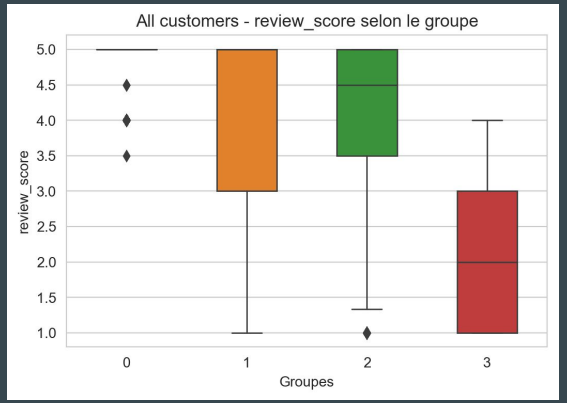


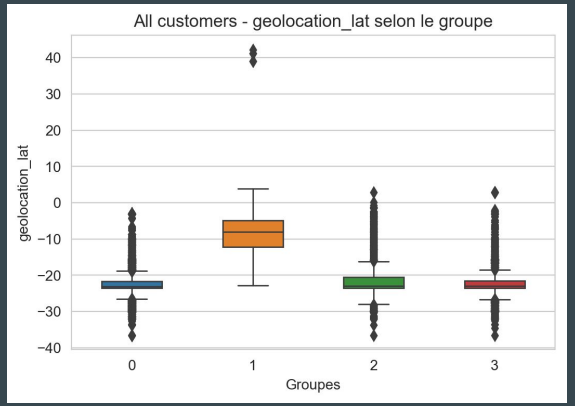
Interprétabilité difficile

Interprétation des groupes issus du KMeans









Non différenciant:

- anciennete
- nb_of_baskets
- mean_purchase_frequancy(days)

Description des groupes

	total_spent	*nb_of_items_bought*	*geolocalisation_lat*	*review_score*	description
Groupe 0	75	1	-22	5	Clients ayant fait qu'une seule commande Très satisfaits.
Groupe 1	100	1	-8	5	Client ayant fait qu'une seule commande De la région du Nordeste
Groupe 2	200	2-3	-22	4.5	Clients fidèles Clients satisfaits
Groupe 3	80	1	-22	2	Clients ayant fait qu'une seule commande Non satisfaits

Maintenance

Stratégie

- Prendre les clients de la première année et observer l'évolution de leur groupe tous les 3 mois
- = 23% des clients

Evaluation avec la *v_measure*

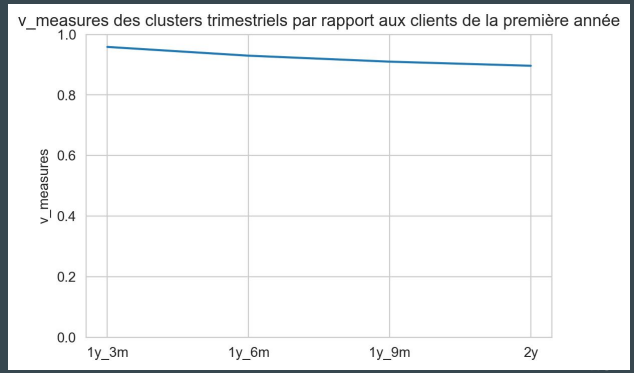
$$V_\beta = (1+\beta)\frac{h \cdot c}{\beta \cdot h + c}$$

Résultats :

- v_measure évolue peu sur N+1. Ce qui est normal car l'immense majorité des clients n'ont fait qu'une commande une fois

Recommendation :

- ne pas se contenter de ces résultats
- Améliorer la segmentation dans 6 mois quand la base clientèle aura évolué
- Permettra de mesurer les impacts des campagnes marketing



Conclusion

Difficultés rencontrées :

- première segmentation est exploitable mais n'est qu'un premier pas :
  - pas assez de données sur les clients (age, sexe, date d'inscription...)
  - business trop récent et en expansion (les groupes sont amenés à bouger)

- Beaucoup d'essais avant de trouver une segmentation interprétable
  - va et vients en ajoutant/retirant des variables en entrée
  - séparation en 2 datasets (les clients one-shot et les reguliers)

Améliorations :

- changer le calcul de la distance (Manhattan, Minkowski...)

- Poids décroissant des variables selon le temps

- PEP8 : commentaires en anglais, 80 caractères maxi