Soutenance Projet 5

Segmentez des clients d'un site e-commerce

Étudiant: Hang ZHONG
Évaluateur: Denis Lecoeuche
Mentor : Arthur MELLO
Date: 06/04/2021

OPENCLASSROOMS

- 5 min

I. Problématique, interprétation et pistes de recherche envisagées

- 5 min

II. Nettoyage et Exploration

- 10 min

III. Modélisation

- 5 min

IV. Modèle final et stabilité

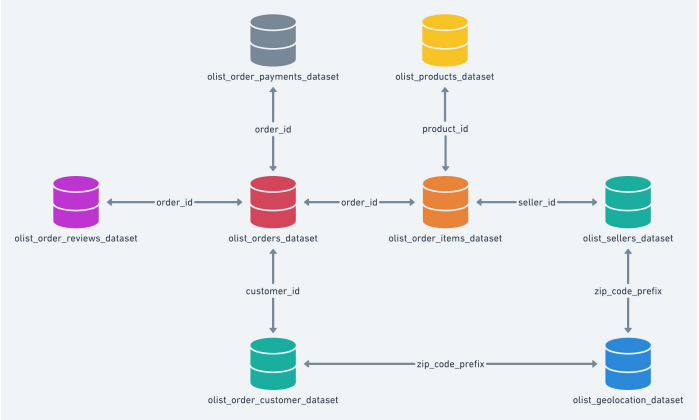
- 5-10 min

V. Questions-réponses

Partie I

Problématique, Interprétation et pistes de recherche envisagées

olist



Problématique

- Solution de vente sur les marketplaces en ligne
- Segmentation des clients -> Campagne

Interprétation

- Choix de features caractérisant le comportement d'un client
- Description actionnable de segmentation
- Analyse de stabilité au cours du temps -> Service de maintenance régulière

Pistes de recherche

- Analyse d'exploration : sélection de features
- Feature engineering
- Algorithmes Non-supervisés : Kmeans, DBSCAN, Hiérarchique
- Stabilité : ARI Score

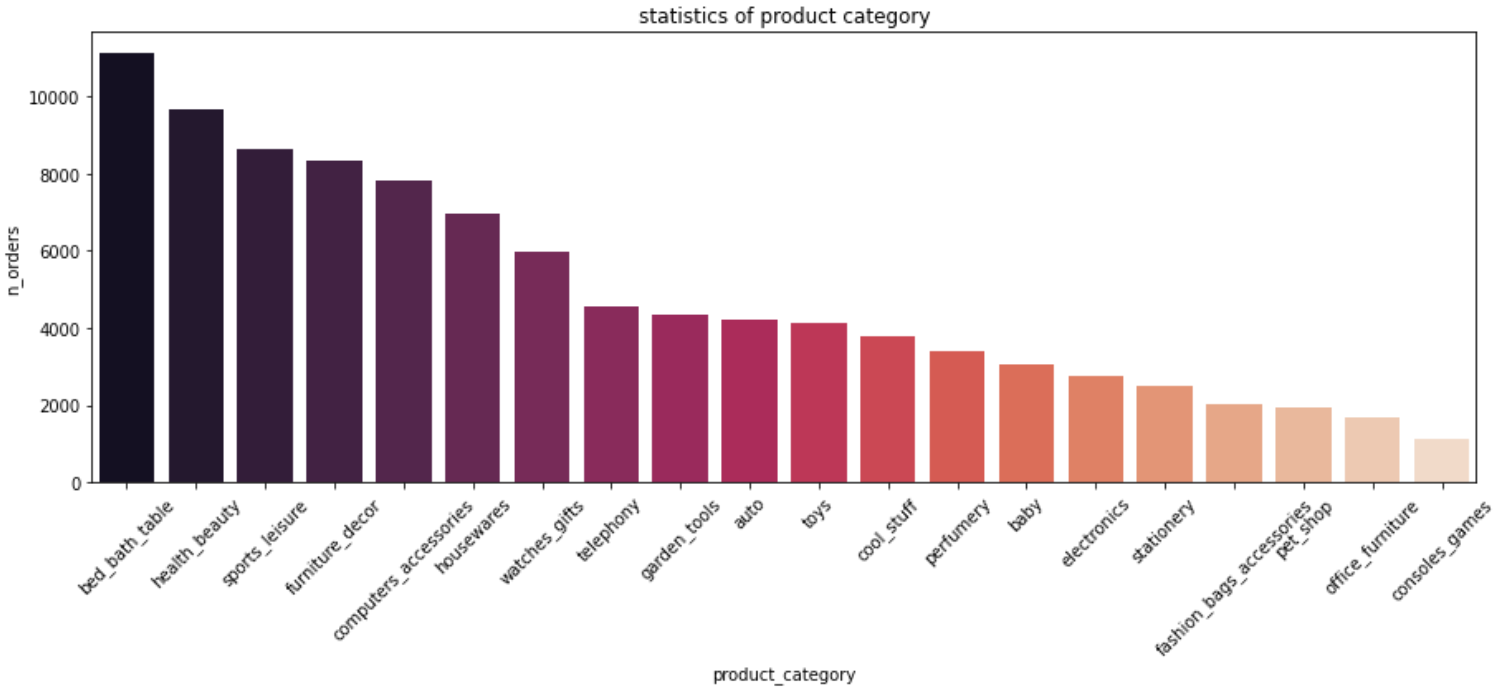
Partie II
Nettoyage et exploration

- 9 fichiers csv :
  - 120MB
  - 99441 enregistrements d'achats
  - location, valeur, date, vendeur, paiement, catégorie de produit, etc.
- Analyse d'exploration :
  - Variables pertinentes existantes :

    n_ items, reviews, vol/weight
  - Création de variables :

    RFM, credit_card_pct

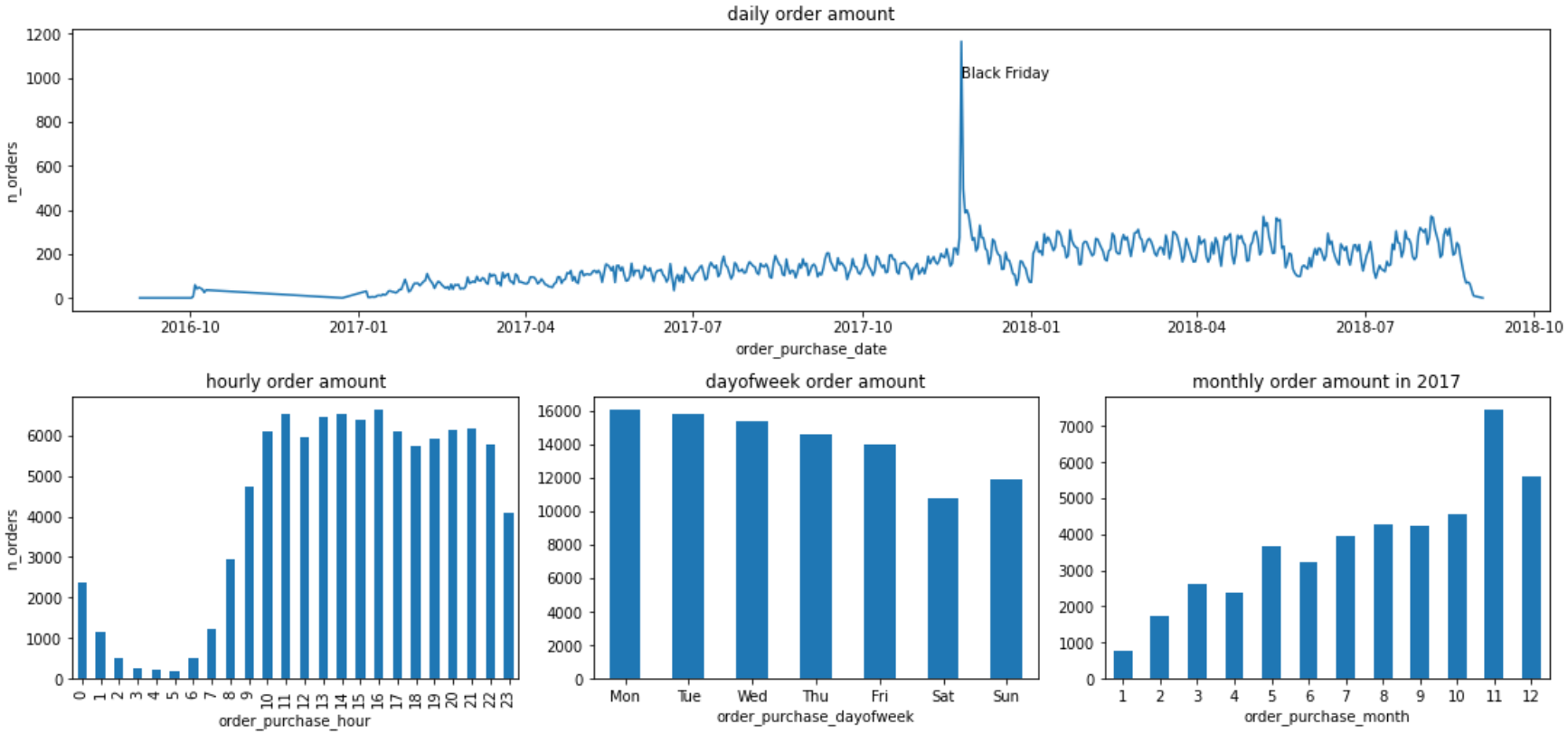
Partie II Nettoyage et exploration



Variables existantes :

- Volume et poids de produit

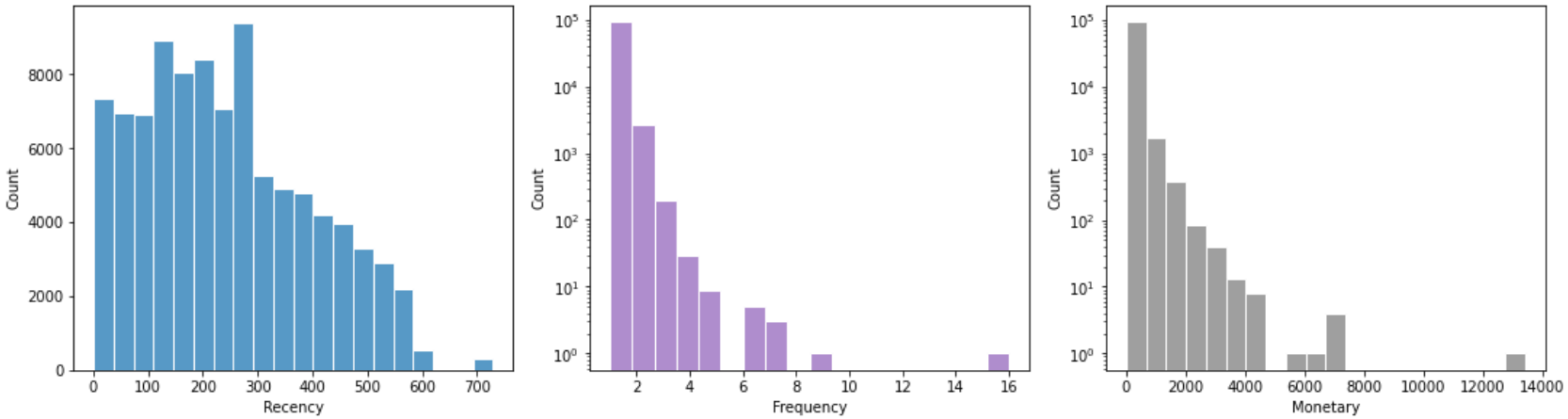
Partie II Nettoyage et exploration



Ce graphe nous présente le volume de vente au cours du temps. Cela peut nous aider à choisir le moment de campagnes publicitaires. Mais il n'arrive pas caractériser le profil de client.

Partie II
Nettoyage et exploration

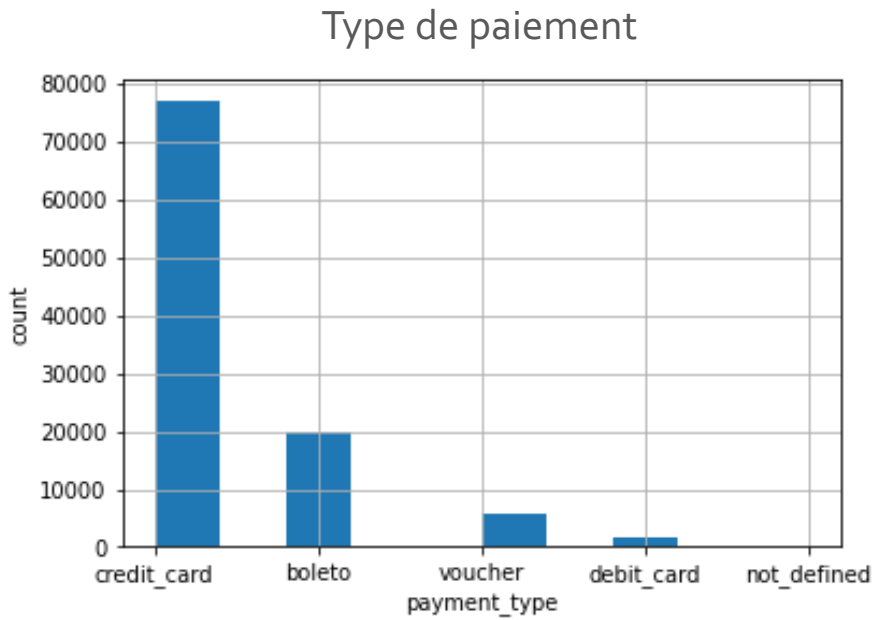
RFM de clients



Création de variable :
Récence, Fréquence de visite, Montant dépensé total

Remarque :
La récence des clients est relativement homogène.
Plus de 96 % des clients n'ont qu'une seule trace d'achat.
Plus de 95 % des clients dépensent moins de 500 euros au total.

Partie II Nettoyage et exploration



Création de variable :
Credit_card_pct: la proportion du paiement par la carte bleue.

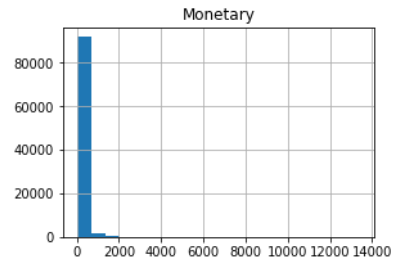
Remarque :
Le mode de paiement nous aide à déduire l'identité de consommateur. Par exemple, les personnes à faible revenu ou sans revenu sont moins susceptibles de détenir une carte de crédit.

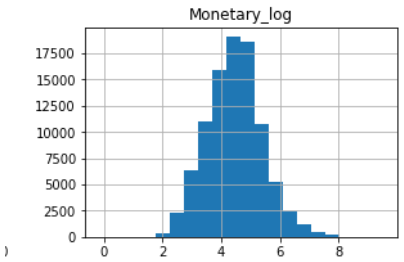
Partie II Nettoyage et exploration

- Concaténation des 8 features

  n_ items, note, volume/poid, récence, montant, credit_card_pct, payInstalment

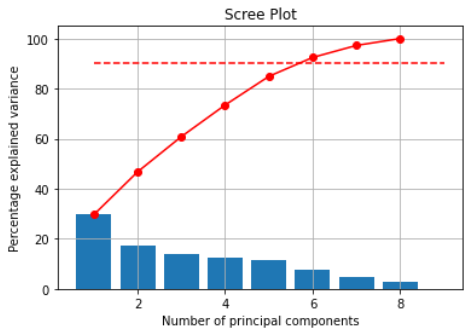
- Transformation logarithmique





- Normalisation (μ=0, δ=1)

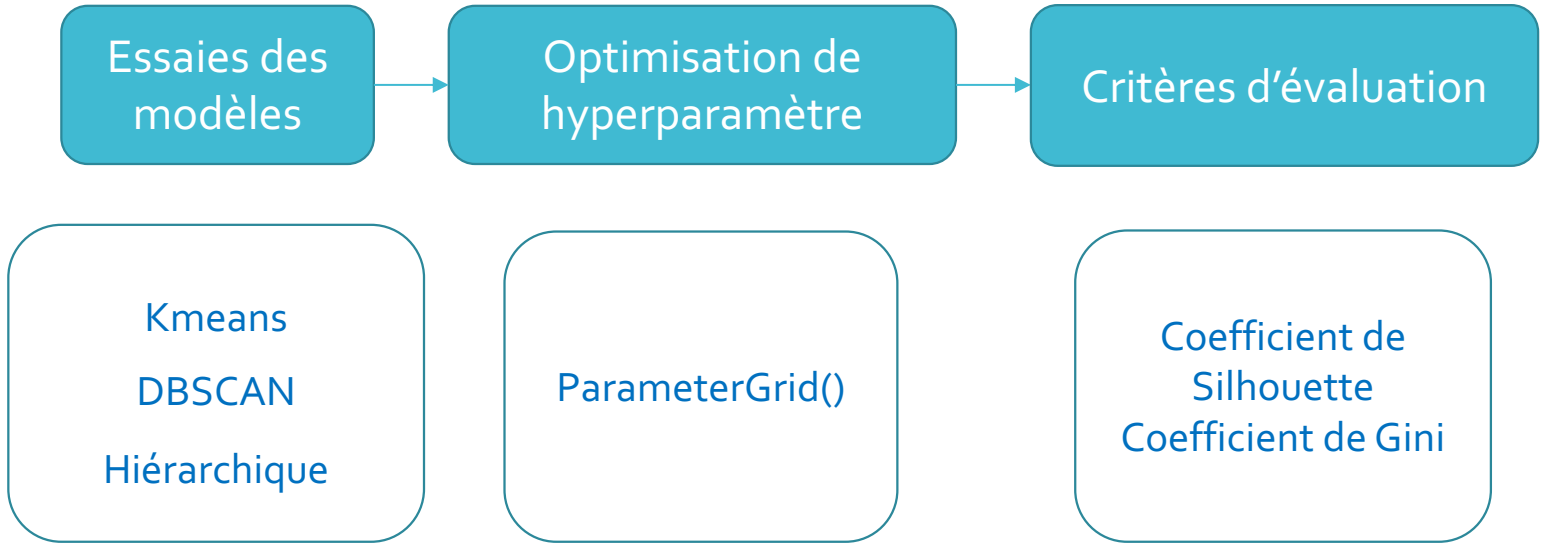
- PCA



Partie III
Modélisation

1. KMeans
2. DBSCAN
3. Hiérarchique

Processus de modélisation

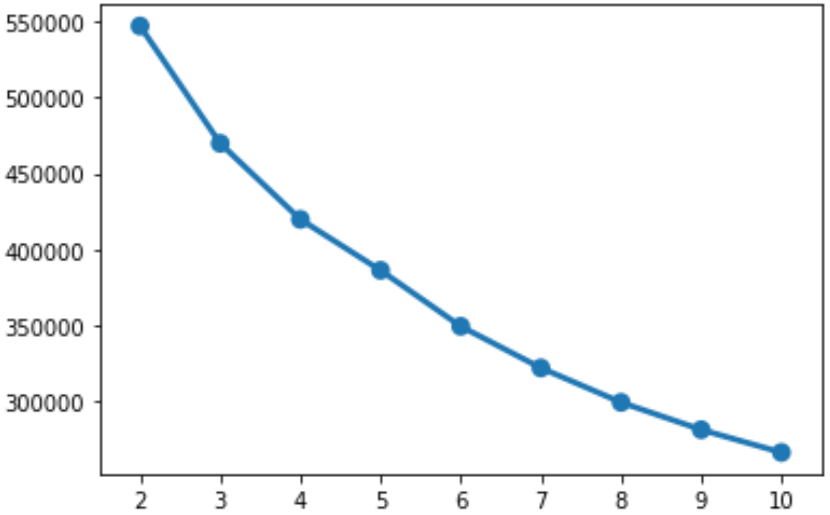


Partie III
Modélisation

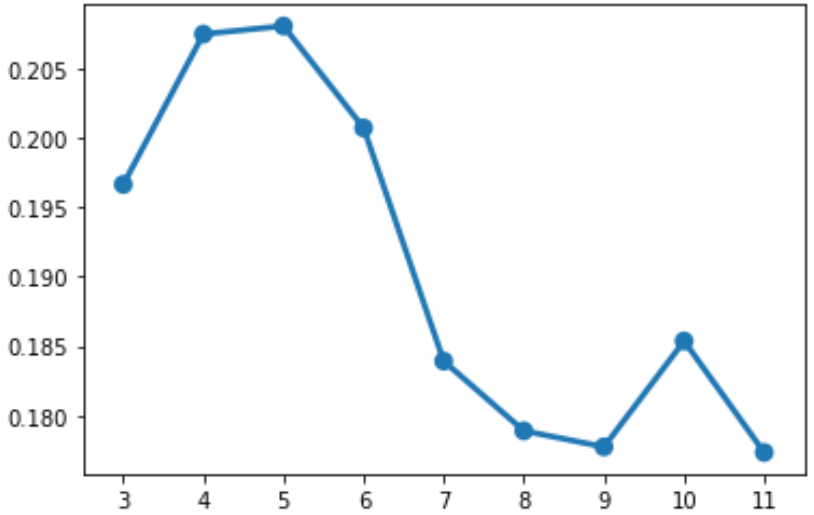
1. KMeans
2. DBSCAN
3. Hiérarchique

Optimisation de nombre de cluster

Méthode du coude(SSE)



Silhouette Score

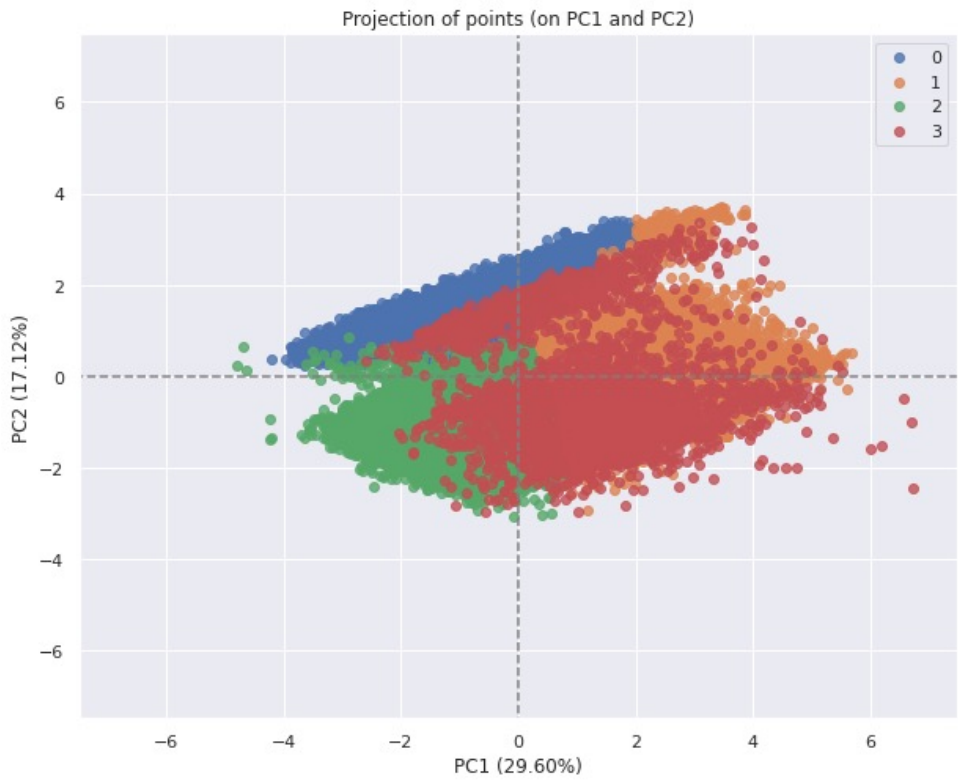


n_clusters = 4 pour KMeans

Partie III
Modélisation

1. KMeans
2. DBSCAN
3. Hiérarchique

Visualisation des clusters KMeans : PCA

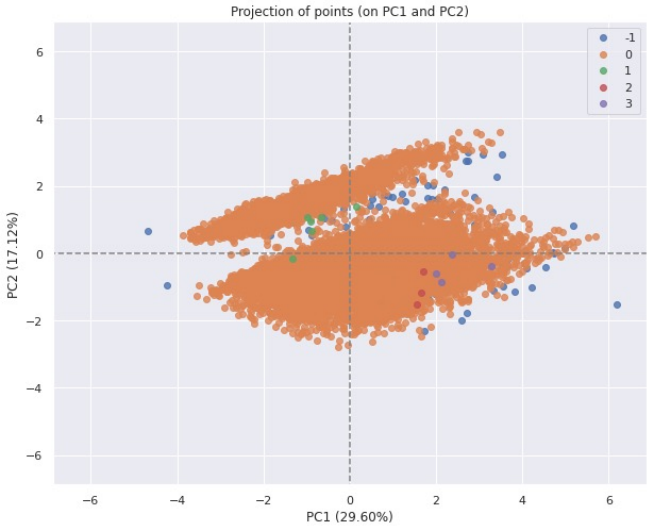


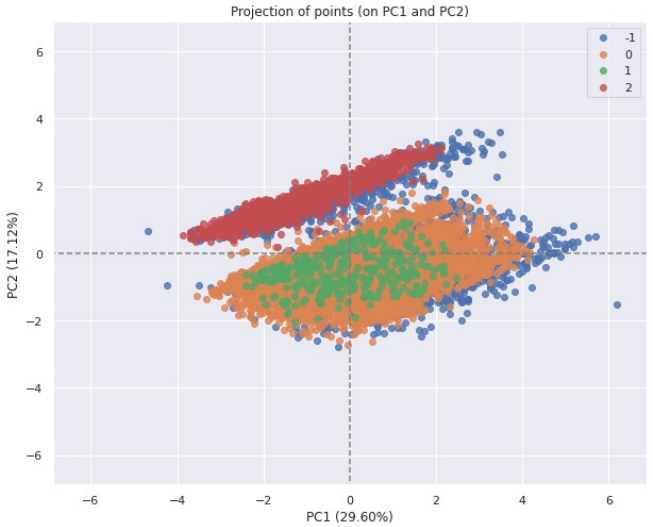
Coefficient de Gini : 0.30,
Silhouette score: 0.2216

Partie III
Modélisation

1. KMeans
2. DBSCAN
3. Hiérarchique

Optimisation de min_sample et eps





	score	gini	n_clus
31	0.351458	0.795140	5
25	0.283778	0.743232	4
29	0.228222	0.791176	5
21	0.189750	0.776894	5
22	0.188254	0.777193	5
20	0.182578	0.838720	7
23	0.178216	0.779921	5
24	0.175676	0.780475	5
13	0.146853	0.402084	4
14	0.144599	0.394677	4

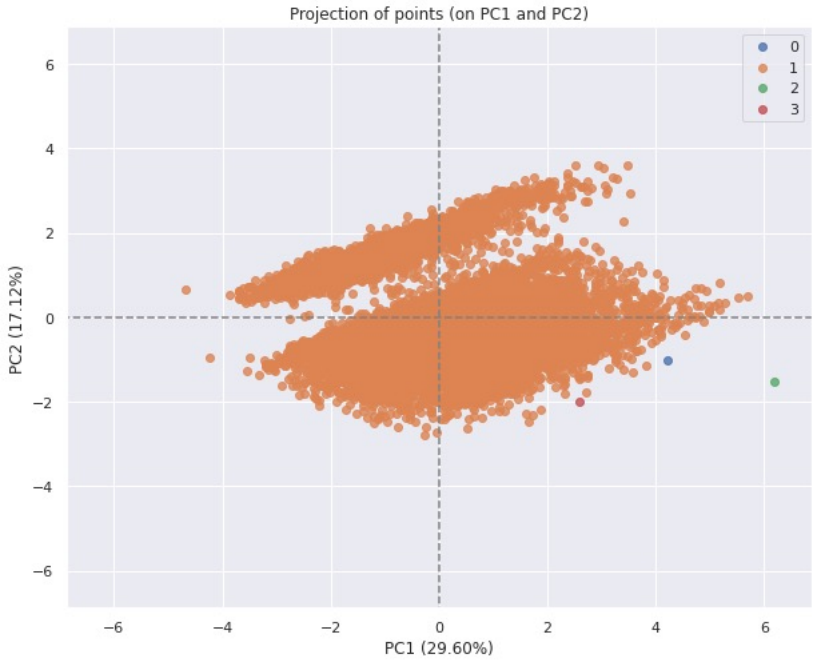
Silhouette : 0.3515,
Gini : 0.80
DBSCAN cluster labels :
[-1, 0, 1, 2, 3]
cluster size :
[78, 9292, 6, 3, 4]

Silhouette : 0.1469,
Gini : 0.40
DBSCAN cluster labels :
[-1, 0, 1, 2]
cluster size :
[2017, 5291, 362, 1713]

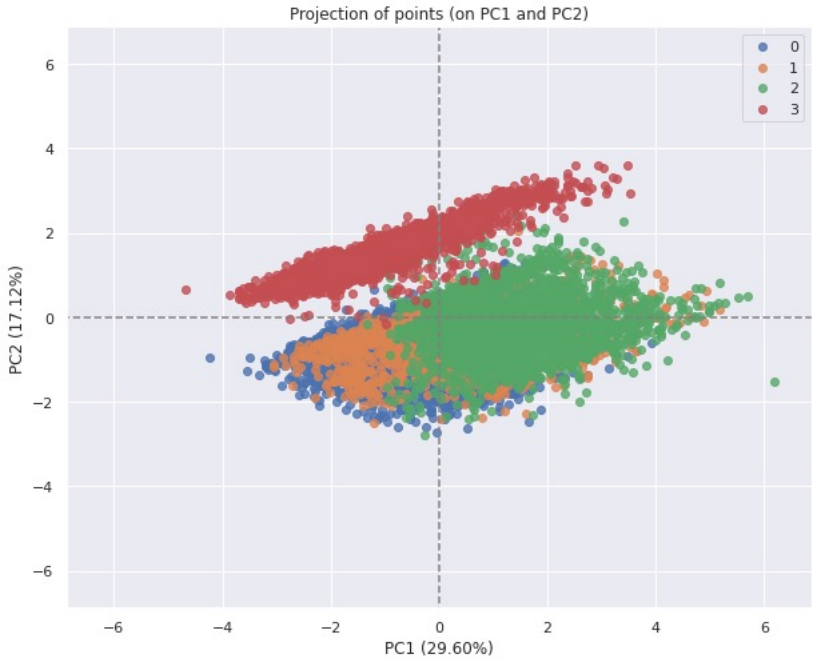
Partie III
Modélisation

1. KMeans
2. DBSCAN
3. Hiérarchique

Optimisation de affinity, linkage et n_clusters



Score: 0.7970,
Gini : 0.75
Hiérarchique cluster labels [0, 1, 2, 3]
cluster size:[2, 9379, 1, 1]

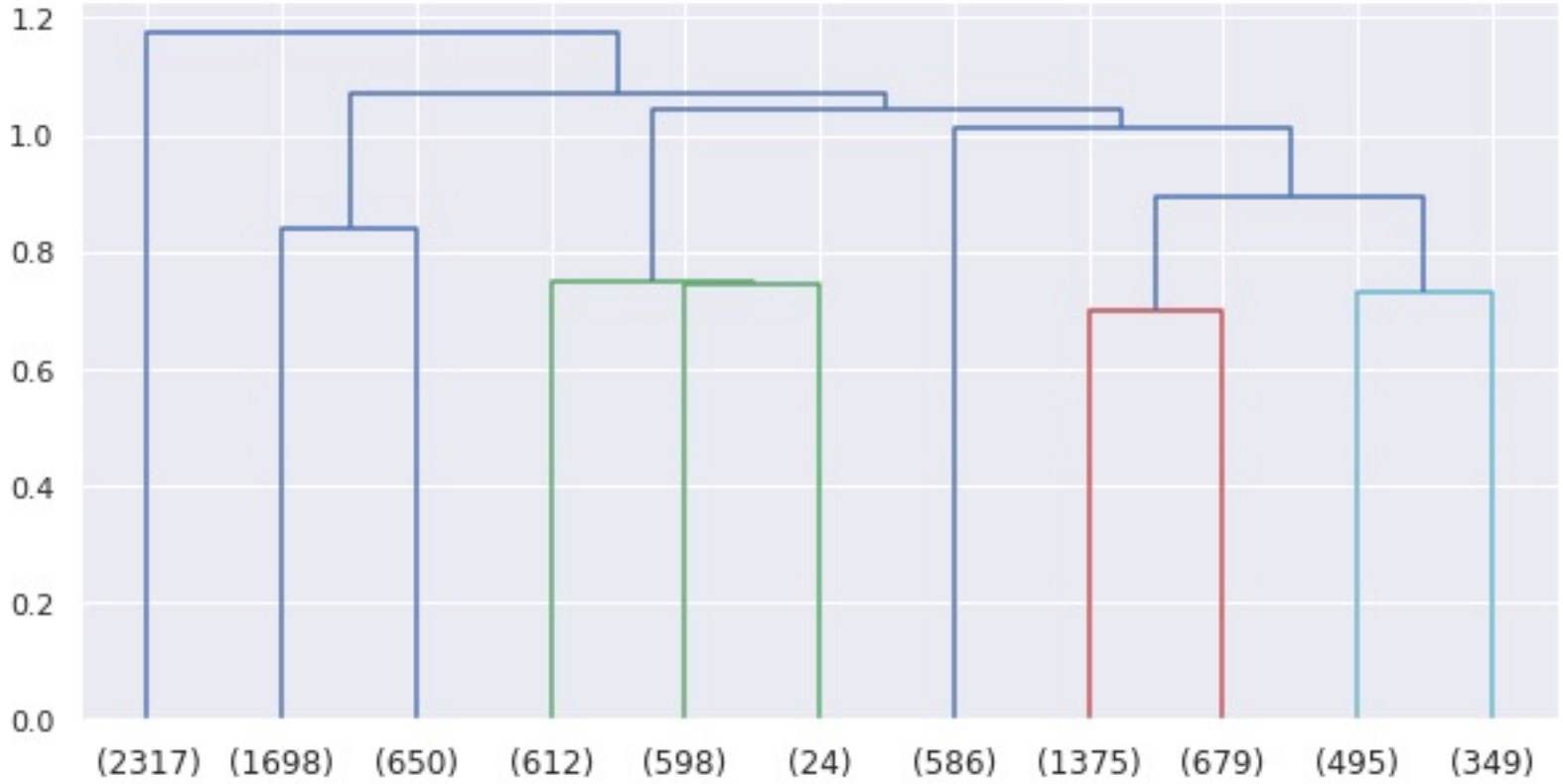


Score: 0.1785,
Gini : 0.16
DBSCAN cluster labels:[0, 1, 2, 3]
cluster size:[3149, 1236, 2684, 2314]

Partie III
Modélisation

1. KMeans
2. DBSCAN
3. Hiérarchique

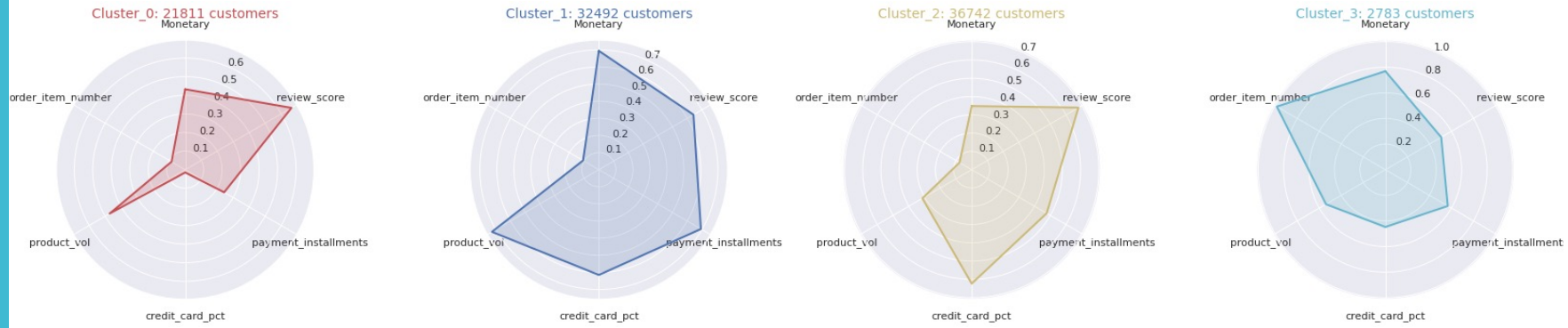
Visualisation des clusters avec un dendrogramme

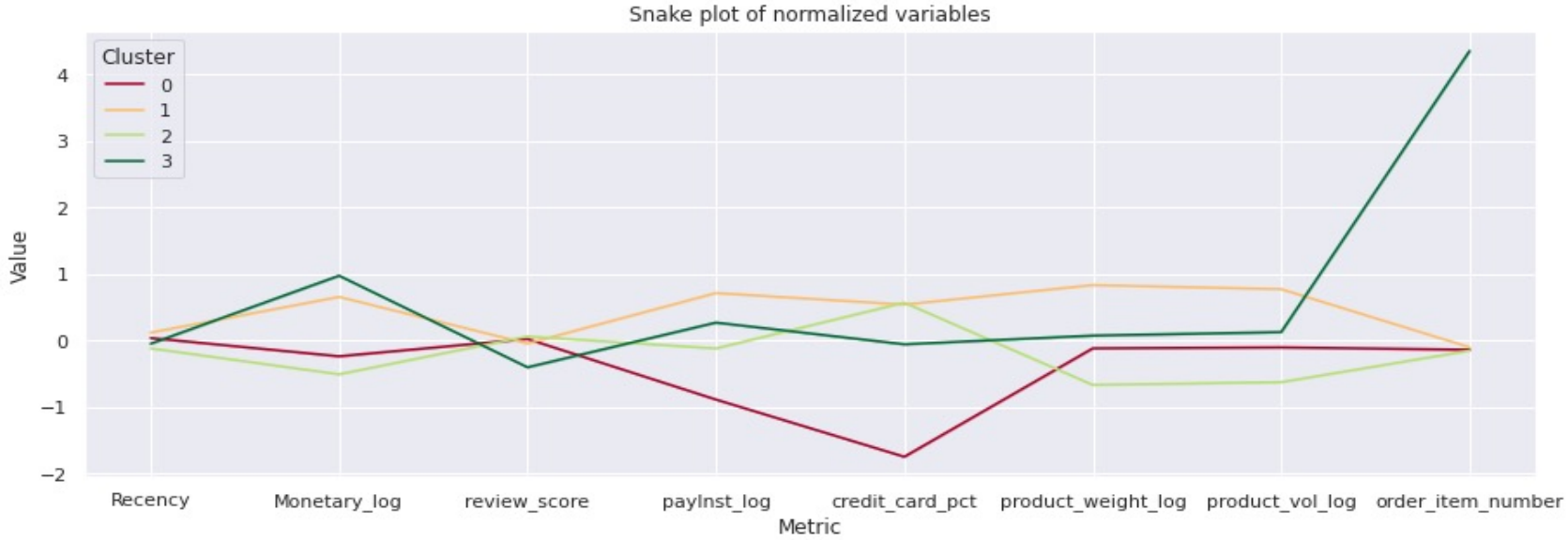


Parametre: { 'affinity': 'cosine', 'linkage': 'average', 'n_clusters' : 4}

Partie IV Modèle final

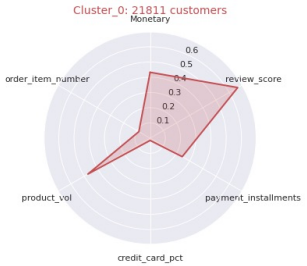
1. KMeans
2. Analyse de stabilité





Partie IV Modèle final

1. KMeans
2. Analyse de stabilité



Top 5 catégories intra-cluster

cluster 0
bed_bath_table
computers_accessories
health_beauty
sports_leisure
furniture_decor

audio	books_general_interest	books_imported	books_technical	party_supplies
0.005	0.006	0.001	0.003	0.001
0.003	0.003	0.000	0.001	0.000
0.004	0.008	0.001	0.004	0.000
0.003	0.010	0.000	0.001	0.000

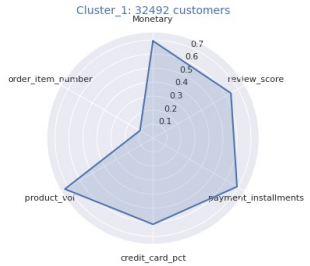
fashio_female_clothing	fashion_bags_accessories	fashion_childrens_clothes	fashion_male_clothing
0.001	0.024	0.0	0.001
0.000	0.005	0.0	0.001
0.000	0.032	0.0	0.001
0.003	0.066	0.0	0.006

Cluster 0:

1. Niveau de consommation **faible**
2. Peu sensibles à la qualité des produits, donne souvent des avis positifs.
3. Pas de paiements en plusieurs fois, pas de carte de crédit, principalement par **boleto**, susceptibles d'être stimulés par des **bons d'achat**.
4. Produits suggérés : audio, books, food, fashion_clothings, party_supplies

Partie IV Modèle final

1. KMeans
2. Analyse de stabilité



Top 5 catégories intra-cluster

cluster 1
bed_bath_table
furniture_decor
sports_leisure
housewares
health_beauty

air_conditioning	auto	baby	computers	cool_stuff	housewares	pet_shop	toys	luggage_accessories
0.003	0.043	0.028	0.001	0.037	0.061	0.019	0.038	0.009
0.004	0.050	0.038	0.005	0.064	0.085	0.023	0.053	0.020
0.002	0.038	0.029	0.000	0.021	0.052	0.017	0.037	0.005
0.008	0.095	0.055	0.007	0.052	0.343	0.061	0.063	0.014

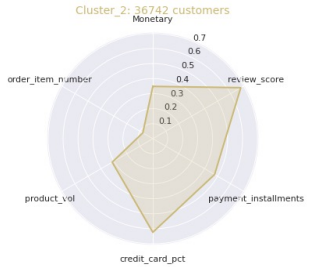
furniture_bedroom	furniture_decor	furniture_living_room	furniture_mattress_and_upholstery	garden_tools
0.001	0.076	0.004	0.000	0.046
0.002	0.097	0.009	0.001	0.062
0.000	0.047	0.001	0.000	0.018
0.005	0.641	0.029	0.000	0.238

Cluster 1:

1. Niveau de consommation **moyenne**
2. utilise beaucoup les **cartes de crédit** et les paiements échelonnés
3. ont l'habitude d'acheter des produits lourds et encombrants
4. Produits suggérés : climatisation, automobile, **bébé**, ordinateur, **meubles**, outils de jardin, appareils ménagers, articles ménagers, accessoires de bagage, **animalerie**, jouets.

Partie IV Modèle final

1. KMeans
2. Analyse de stabilité



Top 5 catégories intra-cluster

cluster 2
health_beauty
computers_accessories
watches_gifts
sports_leisure
telephony

books_general_interest	books_imported	books_technical	consoles_games	electronics
0.006	0.001	0.003	0.013	0.034
0.003	0.000	0.001	0.006	0.008
0.008	0.001	0.004	0.017	0.044
0.010	0.000	0.001	0.015	0.056

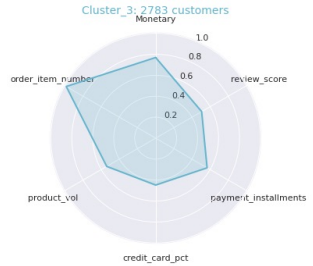
fashion_bags_accessories	food	food_drink	perfumery	telephony
0.024	0.006	0.003	0.033	0.053
0.005	0.001	0.001	0.024	0.007
0.032	0.008	0.004	0.047	0.080
0.066	0.021	0.017	0.056	0.070

Cluster 2:

1. Niveau de consommation **faible**
2. ont l'habitude d'acheter des produits légères et petits.
3. pas sensibles à la qualité, donnent souvent des avis positifs.
4. utilise beaucoup les **cartes de crédit** et parfois les paiements échelonnés
5. produits recommandés : livres, **consoles_jeux**, **électronique**, mode, alimentation, parfumerie, **téléphonie**.

Partie IV Modèle final

1. KMeans
2. Analyse de stabilité



Top 5 catégories intra-cluster

cluster 3
furniture_decor
bed_bath_table
housewares
computers_accessories
sports_leisure

auto	baby	construction_tools_construction	construction_tools_lights	construction_tools_safety
0.043	0.028	0.007	0.003	0.002
0.050	0.038	0.010	0.003	0.001
0.038	0.029	0.007	0.002	0.002
0.095	0.055	0.065	0.023	0.011

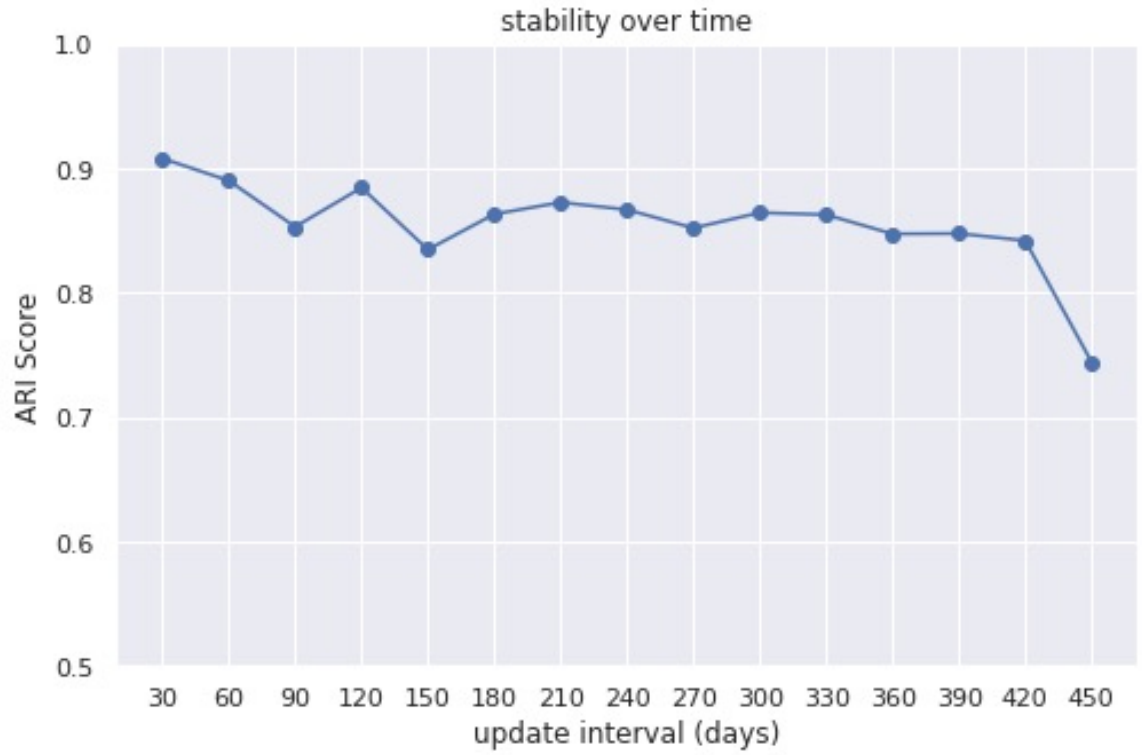
fashion_male_clothing	fashion_shoes	fashion_sport	fixed_telephony
0.001	0.002	0.000	0.003
0.001	0.003	0.000	0.001
0.001	0.002	0.000	0.003
0.006	0.005	0.002	0.014

Cluster 3:

1. Niveau de consommation **élevé**, fréquence élevée et grande quantité
2. Critique en termes d'avis sur les produits
3. utilise les **cartes de crédit**, **boleto** et souvent les paiements échelonnés, légèrement stimulés par les bons d'achat
4. produits recommandés : automobile, **bébé**, outils de construction, mode, **téléphonie fixe**, outils de jardinage, santé et beauté, appareils **ménagers**, mobilier de **bureau**, parfumerie, animalerie, signalisation et sécurité, montres et cadeaux.

Partie IV Modèle final

1. KMeans
2. Analyse de stabilité



- Clients de référence : clients des 6 premiers mois
- Meilleur stabilité : mettre à jour tous les 60 jours
- Stabilité de base: mettre à jour tous les 420 jours (14 mois)

Partie V Q&A

MERCI