Parcours Data Scientist

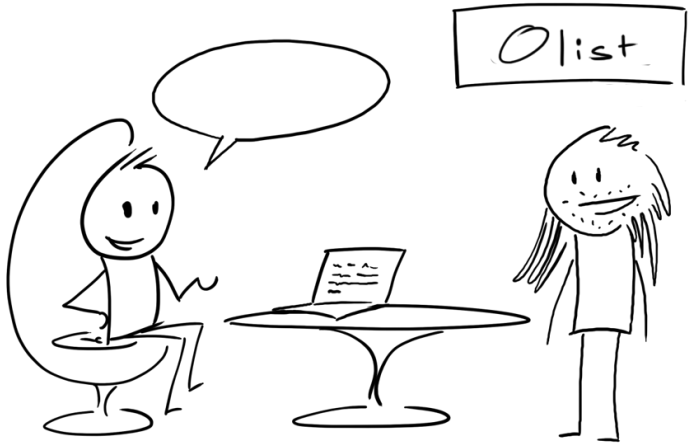
Projet N°5 :
Segmentez des clients d'un site e-commerce

Daniel CHASTANET

Sommaire

- Rappel de la problématique
- Analyse exploratoire
- Segmentation non supervisée
  - Modèles simples
  - Modèles avancés
- Maintenance du modèle
- Conclusion

Rappel de la problématique



Consultant pour la société Olist (solution de vente pour les marketplace)

Olist souhaite que l'on fournisse à ses équipes d'e-commerce une segmentation des clients qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

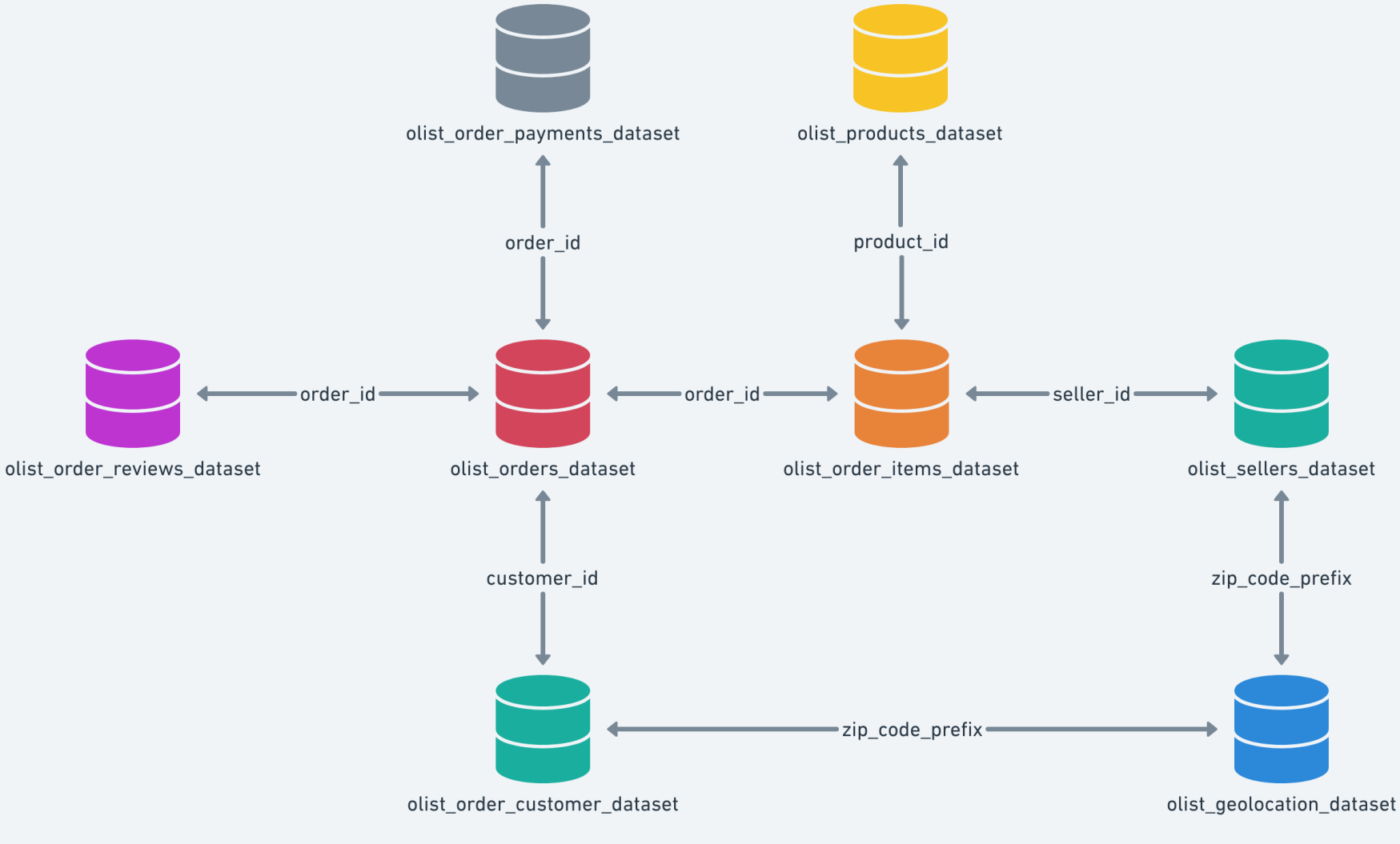
Les objectifs :
- Segmentation clients
- Description actionnable des groupes
- Proposition de contrat de maintenance
- Code en convention PEP8

Analyse exploratoire

Source : edx.org

Les données

- 9 bases de données :
- Merge des BBD selon le graphique (119143, 44)
- - de 1,5% des données manquantes (~1800)



Description des données

L'annonce :

nom, catégorie, description, photos, poids, longueur, largeur

Logistique :

géolocalisation (anonyme) du vendeur, de l'acheteur (pays, état etc)

Commandes :

numéro de commandes, frais de port, payement (type, facilités), dates (livraison etc), état

Client :

numéro d'identification, commandes, avis, dépenses

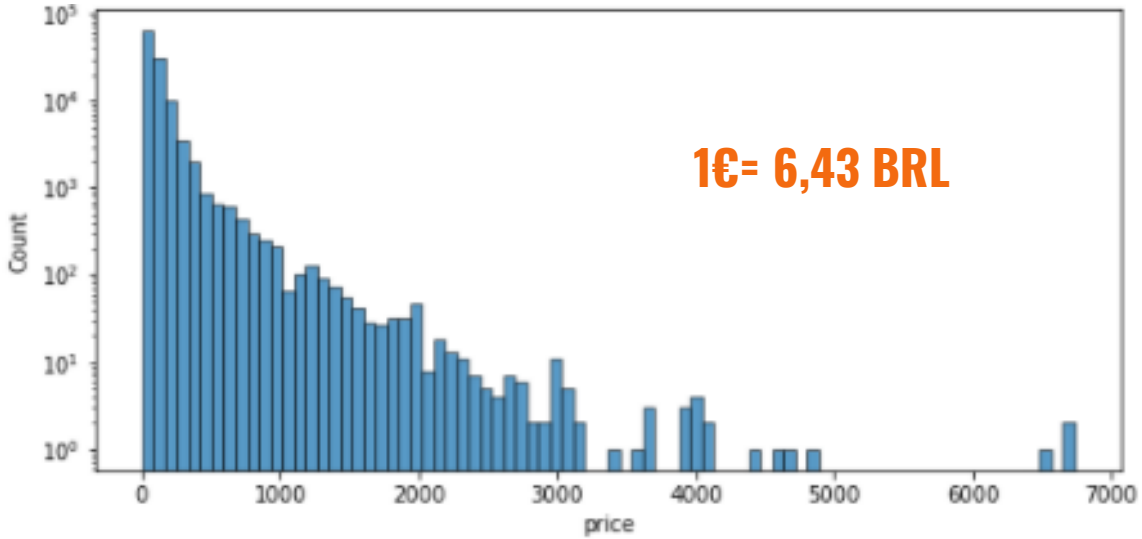
Préparations des données

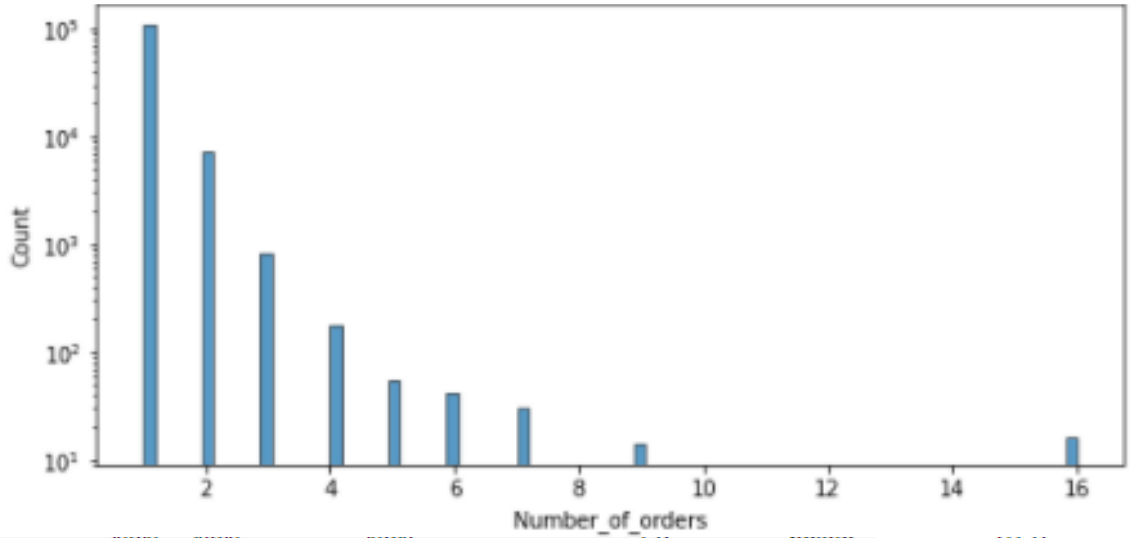
Nettoyage des données

Sélection des commandes en fonction de leur état :
delivered, shipped, invoiced, processing, canceled,
→ 2164

Valeurs manquantes ou abérhantes :
1 lignes retirés pour, quelques remplacement de
368 lignes retirés pour valeurs aberrantes

→ Base de données très « propre »
→ Base de données très peu équilibrée pour certain





		order_id	order_item_id	product_id	seller
489	442				
1488	e57d9	116978 a68ce1686d536ca72bd2dadc4b8671e5	NaN	NaN	NaN
1722	debc1				

Feature enginering

Les critères

Le but :

→ segmentation client

→ Créer des variables spécifiques pour des clients « uniques »

→ Garder le plus d'informations possible (concernant chaque client)

Les critères discriminants sélectionnés :

- critères « commandes »
  - nombre de commandes
  - nombre d'articles
  - montant total dépensé
  - montant moyen de transaction
  - fréquence
  - Date de dernier achat
- critère « catégoriel »
  - moyen de paiement utilisé
  - catégorie de produit
- critère de satisfaction
  - note moyenne des avis
- critère d'accessibilité
  - distance acheteur / vendeur
  - délais de livraison

Segmentation

Source : marketing-etudiant.fr

Segmentations basiques

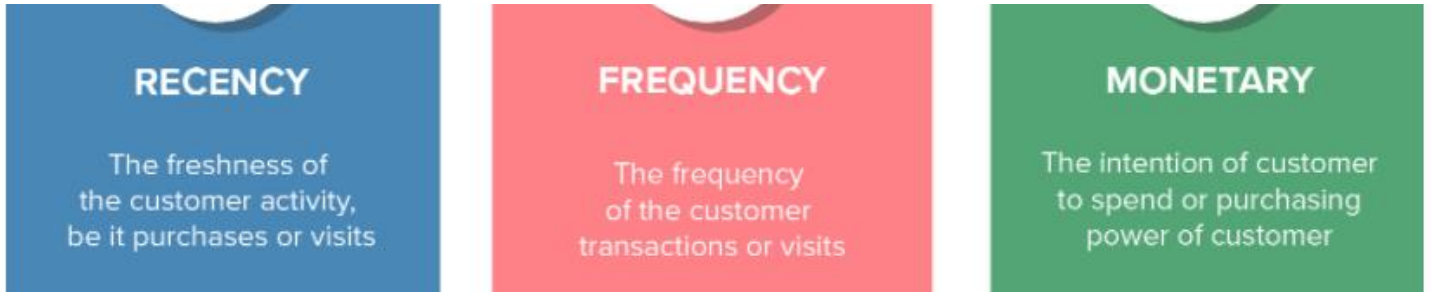
Segmentations sans persona

Pareto law (20/80%)

```
For the category : cool_stuff , the 80% total amount correspond to  50.39141780226152
For the category : pet_shop , the 80% total amount correspond to  55.54194733619106
For the category : nan , the 80% total amount correspond to  50.664697193500736
For the category : home_confort , the 80% total amount correspond to  52.18147448015123
For the category : health_beauty , the 80% total amount correspond to  50.76281529698942
For the category : electronics , the 80% total amount correspond to  41.50244964616222
For the category : industry_commerce_construction , the 80% total amount correspond to  50.33027227323989
For the category : arts_video_audio , the 80% total amount correspond to  38.38634600465477
For the category : clothing_and_bags , the 80% total amount correspond to  51.121605667060216
For the category : auto , the 80% total amount correspond to  45.42046063202999
```

ABC method (big / average / small)

RFM method



Segmentations basiques

Segmentations sans persona

Pareto law (20/80%)

ABC method
(big / average / small)

RFM method

Kmeans
7 clusters

Modèle basique et pas très adapté à notre base de donnée !

4	35732
5	30318
3	14358
6	11004
7	2185
8	685
9	115

à la main
3*3
→ 7 groupes

6	29035
0	20545
4	18860
1	14589
2	8809
3	2167
5	392

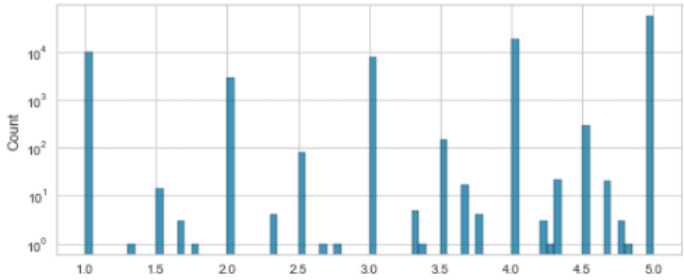
Modèles plus « avancés »

Source : stress.app

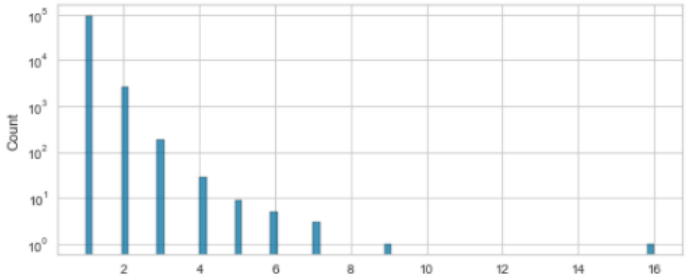
K Means – 5 attributs

Les attributs

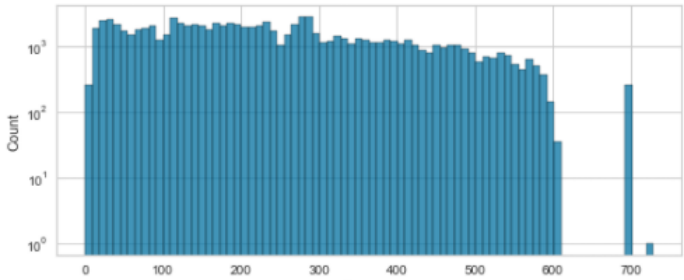
- nombre de commandes
- nombre d'articles
- date de dernier achat
- montant total dépensé
- note moyenne des avis



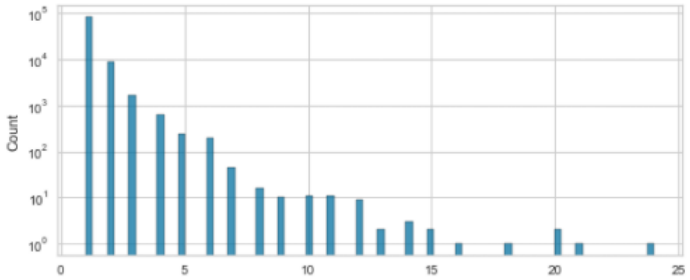
Avis clients



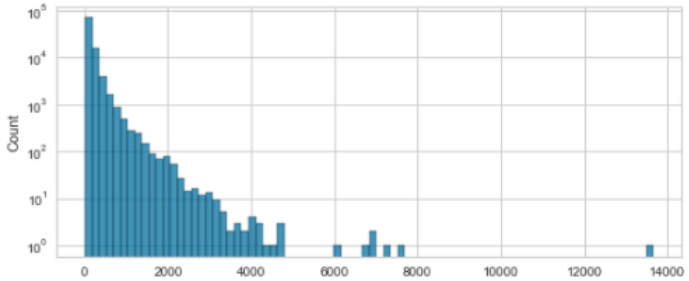
Nbr de commandes



Dernier achat



Nombre d'articles

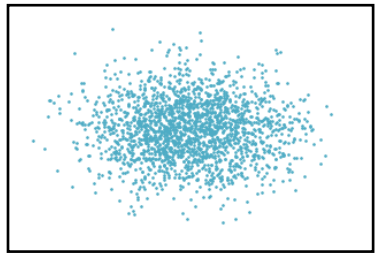


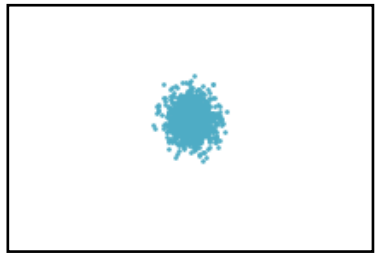
Dépense client

Choix du nombre de clusters

Forme/Distance/Stabilité/Compatibilité

Forme

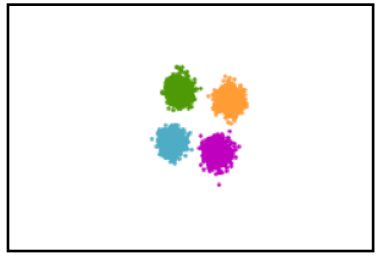


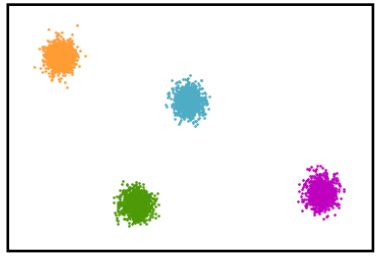


$$T_k = \frac{1}{|C_k|} \sum_{x \in C_k} d(x, \mu_k)$$

Avec
- C_k le cluster k
- μ_k le centroid

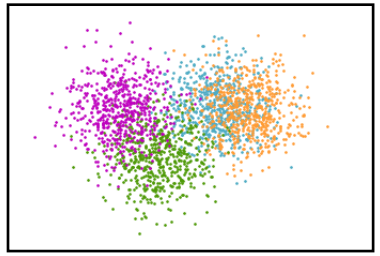
Distance

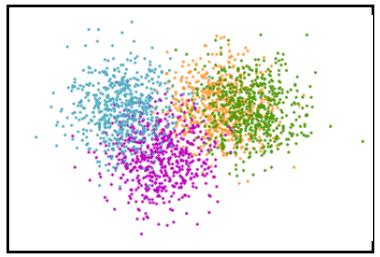




$$S_{k,l} = d(\mu_k, \mu_l)$$

Stabilité





Evaluation des clusters

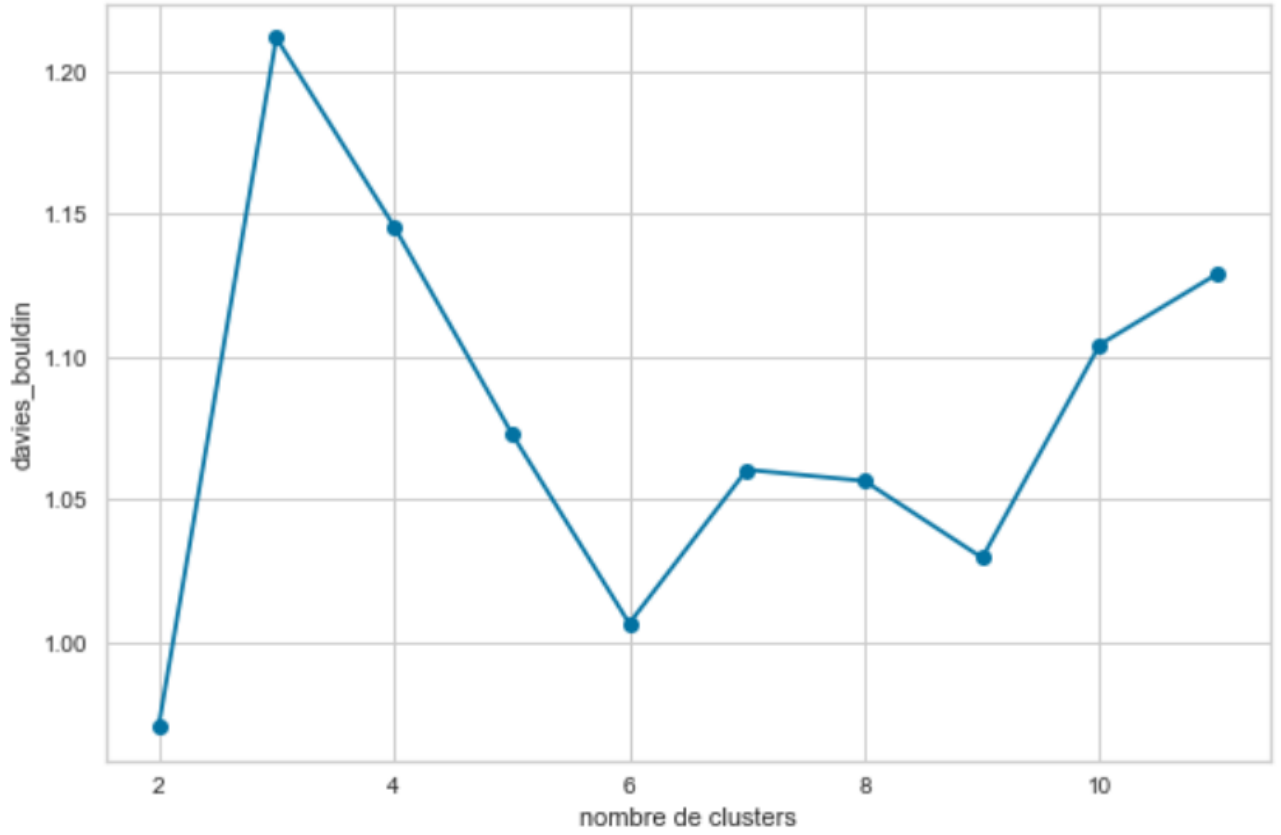
Forme/Distance/Stabilité/Compatibilité

Inertie / Methode du coude

Score de silhouette

Davies Bouldin

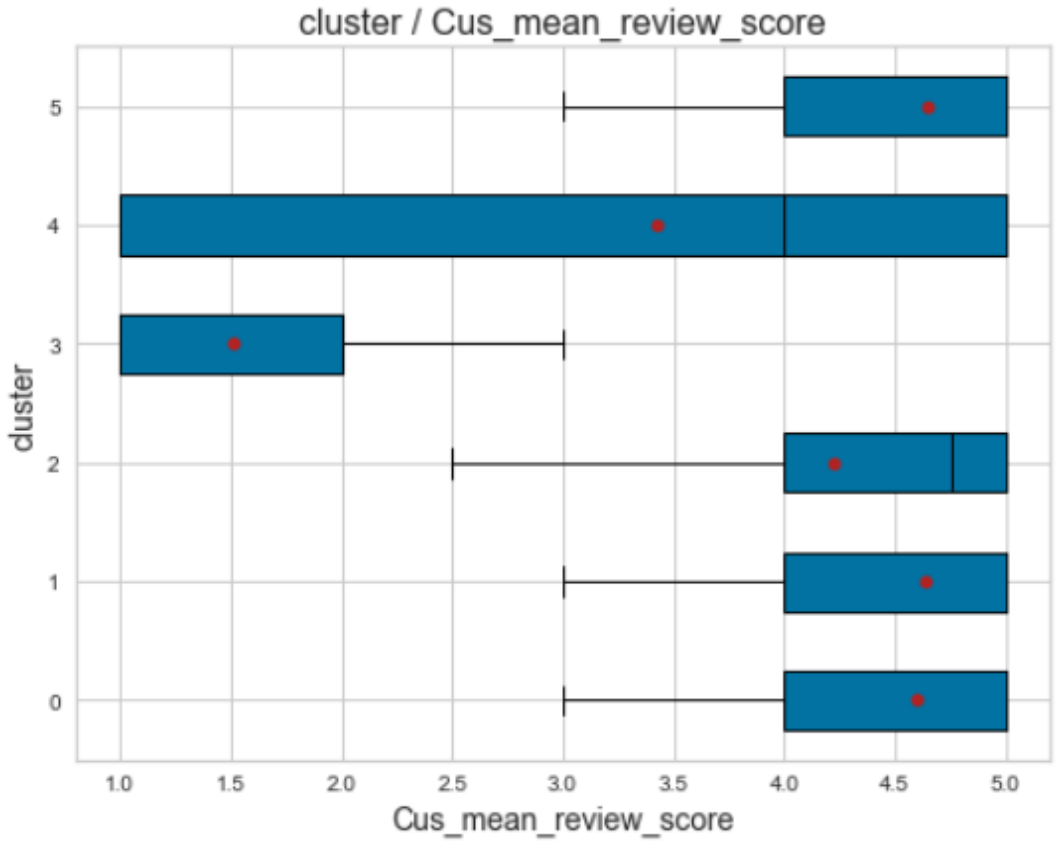
$$D = \frac{1}{K}\sum_{k=1}^{K} \max_{l \neq k} \frac{T_l + T_k}{S_{k,l}}$$



Evaluation des clusters

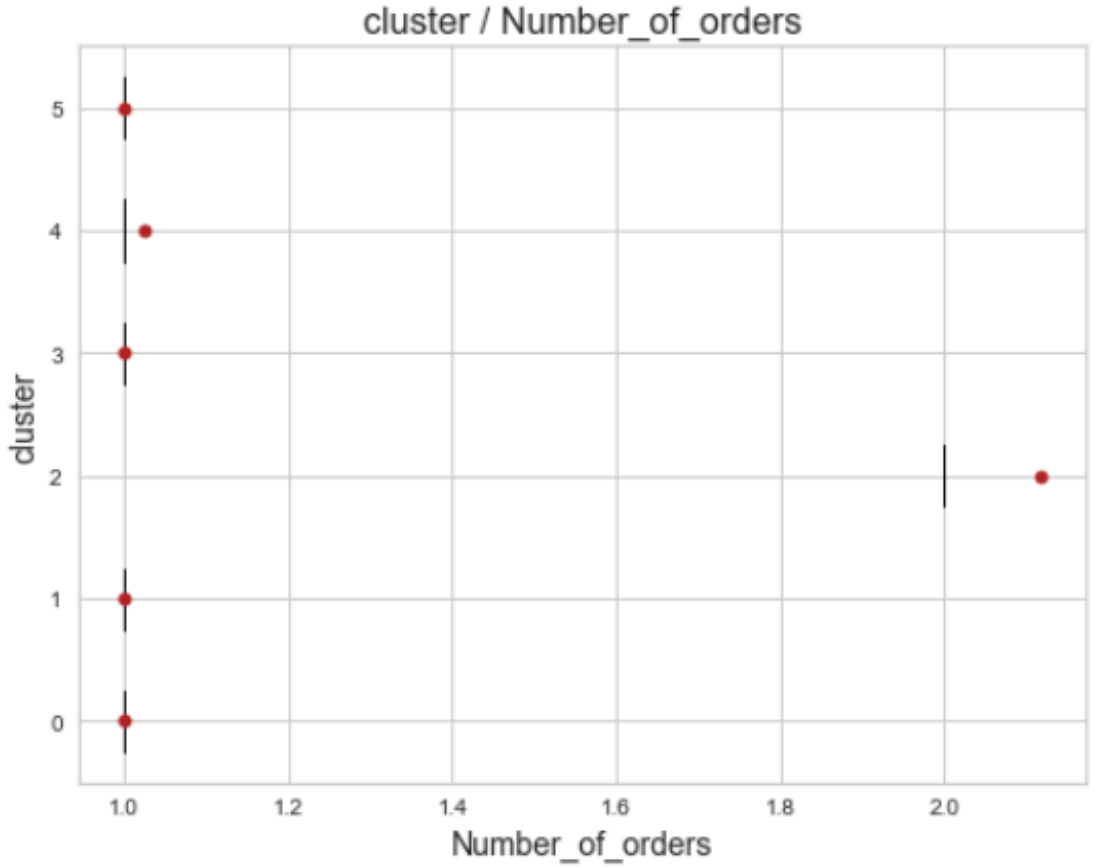
Compatibilité

Critères métiers : le sens !



Avis clients

Nombre de commandes



La segementation

Critère \ Cluster	0	1	2	3	4	5
Nombre de commandes	1	1	2+	1	1	1
Date du dernier achat	120-350	80-230	110-320	350-490	90-240	150-305
Nombre d'articles	3 - 4a	1	2 / 3a	1	1	1
Montant dépensé	m +++	m ---	m +	m --	m +	m -
Note Moyenne des avis	1 - 5*	4+	3,5+	4+	4+	1 - 2*
Nombre de clients	1973	29270	2707	23118	21262	12597
Global	**Les dépensiers** Montant élevés mais notes variables	**Les faible budget** Avis ok mais très faible dépense (ils sont nombreux)	**Les bons** Plusieurs achats, montant et note de satisfaction corrects	**Les fantômes** anciens, petites dé dépenses, avis corrects	**Les corrects** budget ok, avis ok et relativement récent	**Les insatisfaits** Mauvaise note, montant faible, un seul achat

Satisfaits **Insatisfaits**

K Means – all features

PCA (Principal Component Analysis)

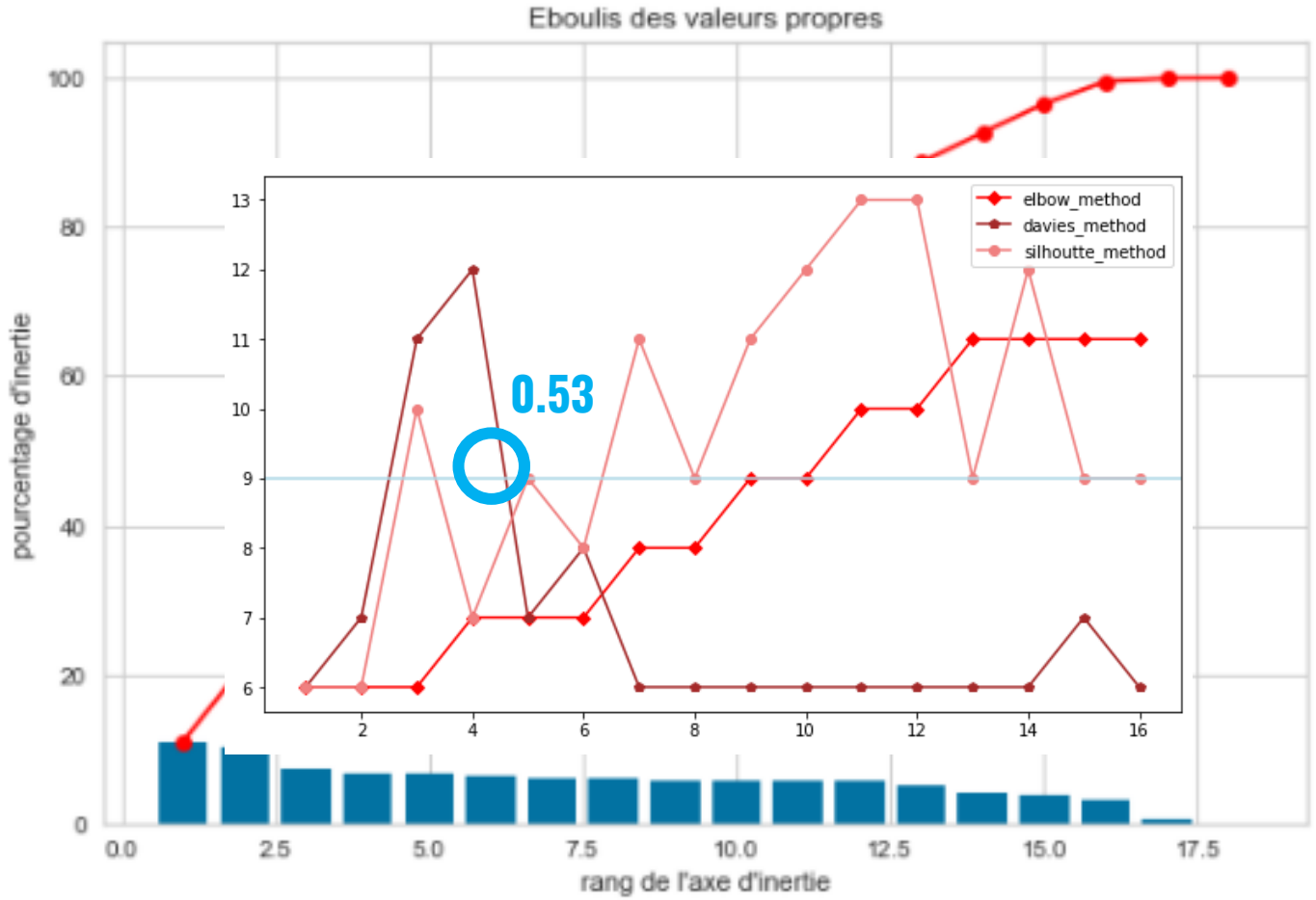
Les avantages de la PCA

Faciliter la visualisation

Réduction des coûts (calcul, stockage)

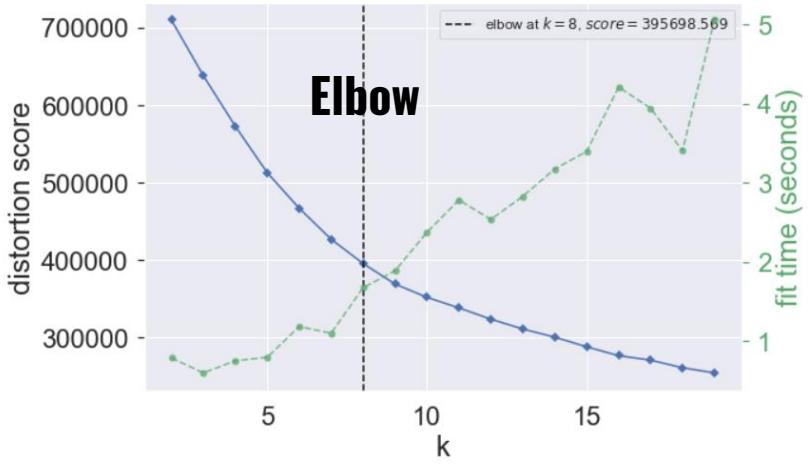
Améliorer de l'apprentissage

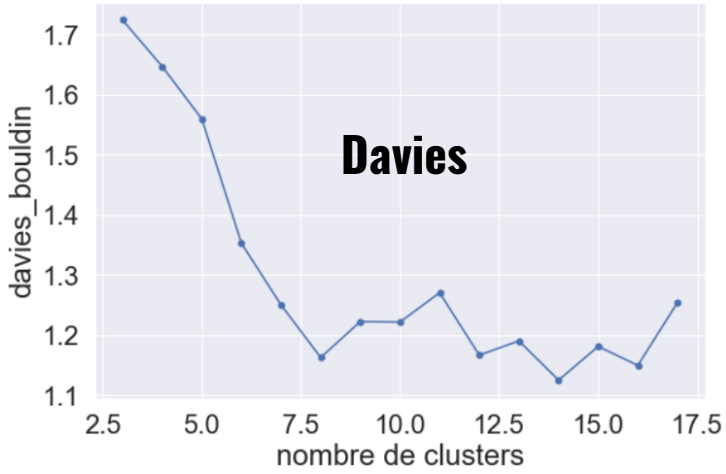
- modèle moins complexe
- élimination de variables non pertinentes
- combattre le fléau de la dimensionalité.

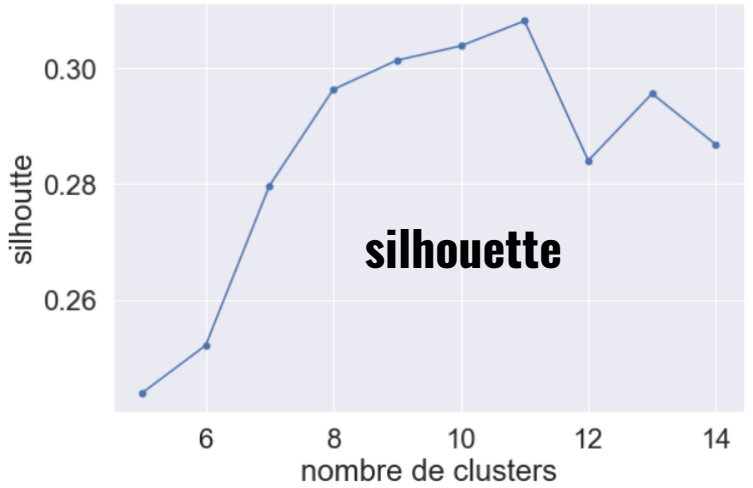


K Means PCA 6 (50%)

Lectures des clusters







Best number of clusters : 11 (sil, elbow, david)

Maximum number of clusters for a easy marketing use : 9

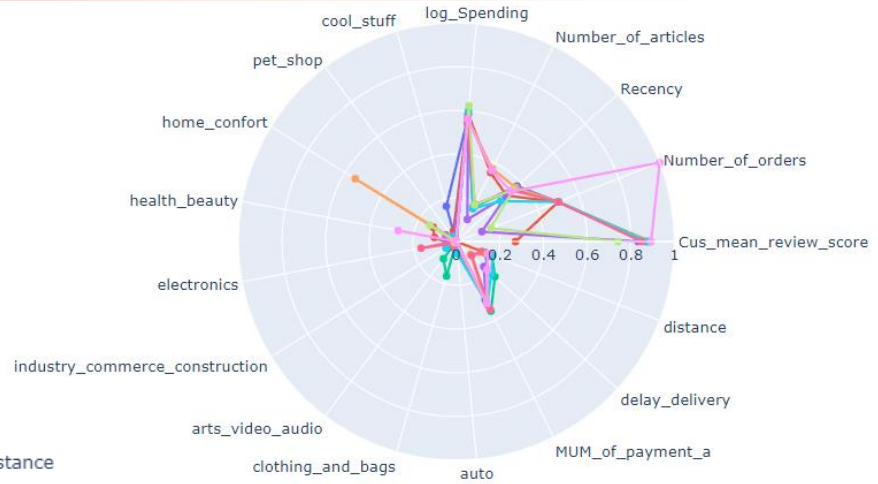
K = 9	
8	21822
4	17129
5	12456
1	10178
0	9542
2	9434
7	6560
6	2091
3	1715

K = 10	
4	21762
5	16963
1	12441
2	10070
9	9529
7	9456
3	6538
6	1709
0	1586
8	873

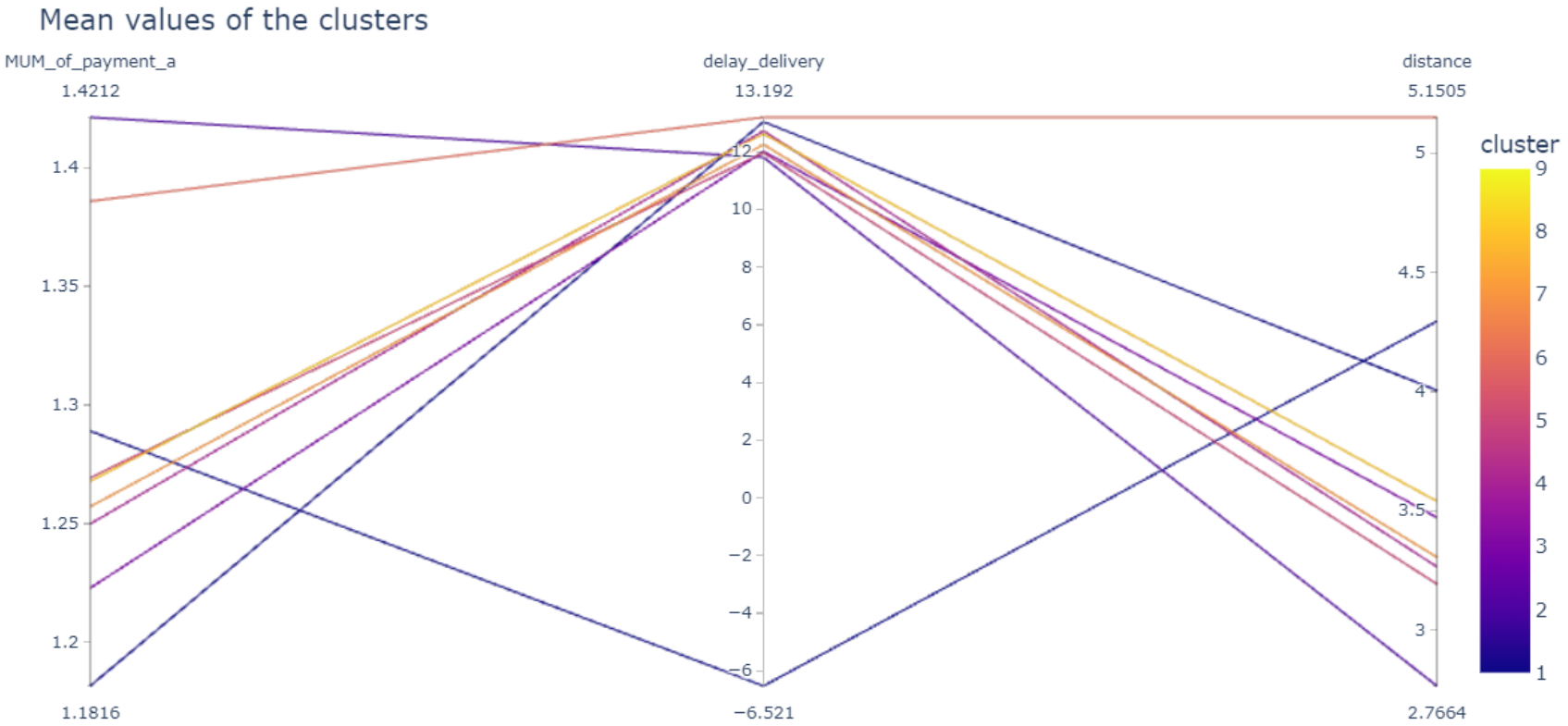
K = 11	
3	21682
6	16107
4	12422
1	9864
2	9504
5	9386
8	6549
0	1783
7	1743
9	1342
10	545

K Means PCA 6 (50%)

Description des données



Best number of clusters : 11 (sil, elbow, david)



K = 9

8	21822
4	17129
5	12456
1	10178
0	9542
2	9434
7	6560
6	2091
3	1715

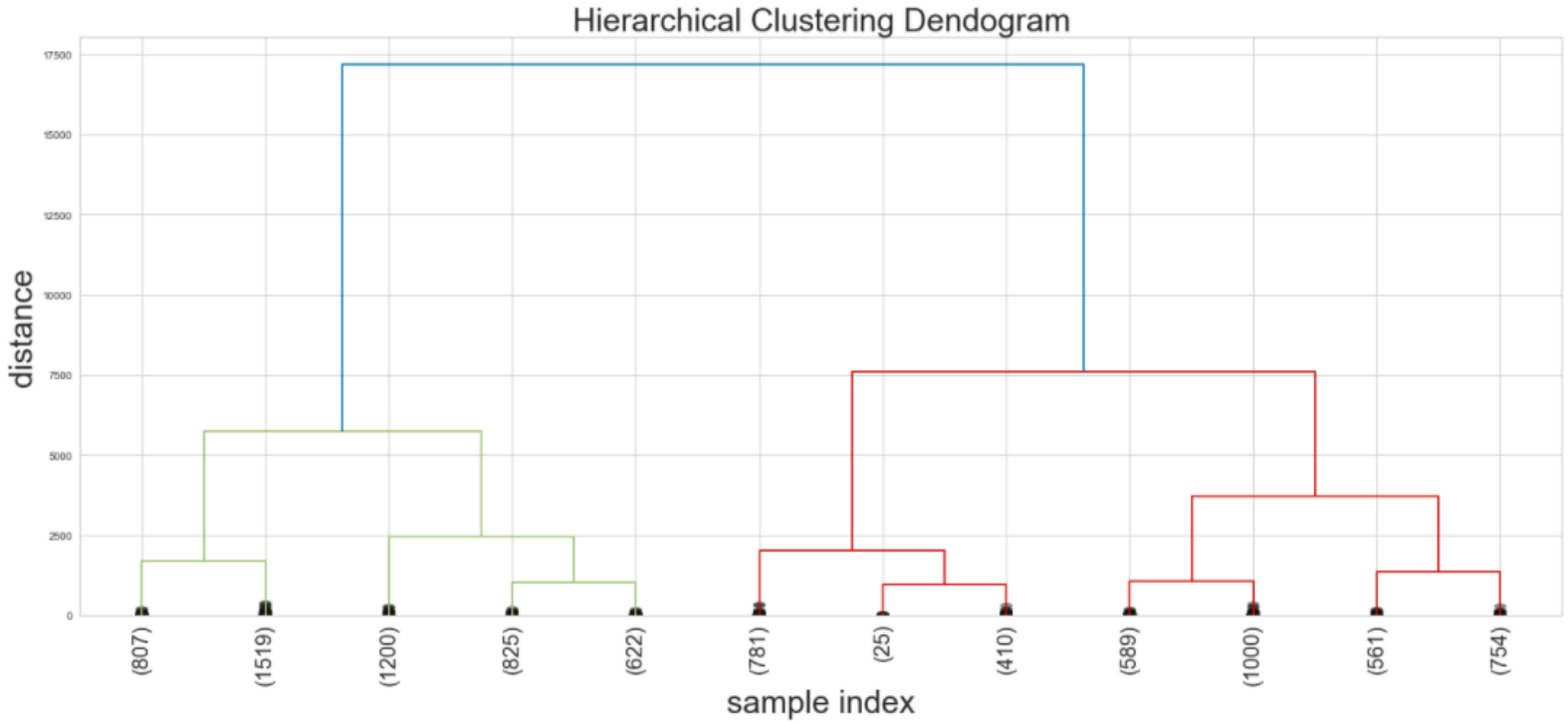
K Means - Segmentation

Critère \ Cluster	0	1	2	3	4	5	6	7	8
Note Moyenne des avis	4+	4+	4+	3,5+	4+	3+	4+	4+	1,4
Nombre de commandes				2		1-2			
Date du dernier achat	130-350	120-380	130-370	110-300	120-350	120-350	80-250	130-370	150-270
Nombre d'articles				2-3		2-4			
Montant dépensé	m+	m++	m+++	m++++	m++	m++++	m++	m+	m++
Cool_stuff		0,5	1						0+
pet				0+			0,5		
home		1		0-1		3-4			0-1
health				0-2	1				0+
electronics	1			0+					
commerce				0+			0-1		
art_vid_audio				0+				0-1	
clothing				0+				0-1	
auto							0-1		
mum	CC / boleto	CC	CC	CC	CC	CC	CC	CC / boleto	CC
delay									10 jours+
distance	+								++
Nombre de clients	9434	21825	12456	2091	17129	1715	9543	10179	6555
Global	**Les moins_a** Petites dépenses / boleto electronics	**Les moyens_a** Bonnes notes Dépenses moyennes home	**Les bons** Bonnes notes Dépenses élevées sur un achat / Cool stuff	**Dépensiers_a** Montant élevés notes correctes Catégories mixtes	**Les moyens** Dépenses moyennes santé	**Dépensiers_b** Montant élevés notes variables home	**Les clients récents** Dépenses moyennes	**Les moins_b** Petites dépenses	**Les insatisfaits** Mauvaise note, Long de livraison Longue distance

Hierarchical Clustering

Avantages : Pas besoin de connaître le nombre de cluster à l'avance

Désavantages : Peut être très long et très gourmant pour des BDD importantes

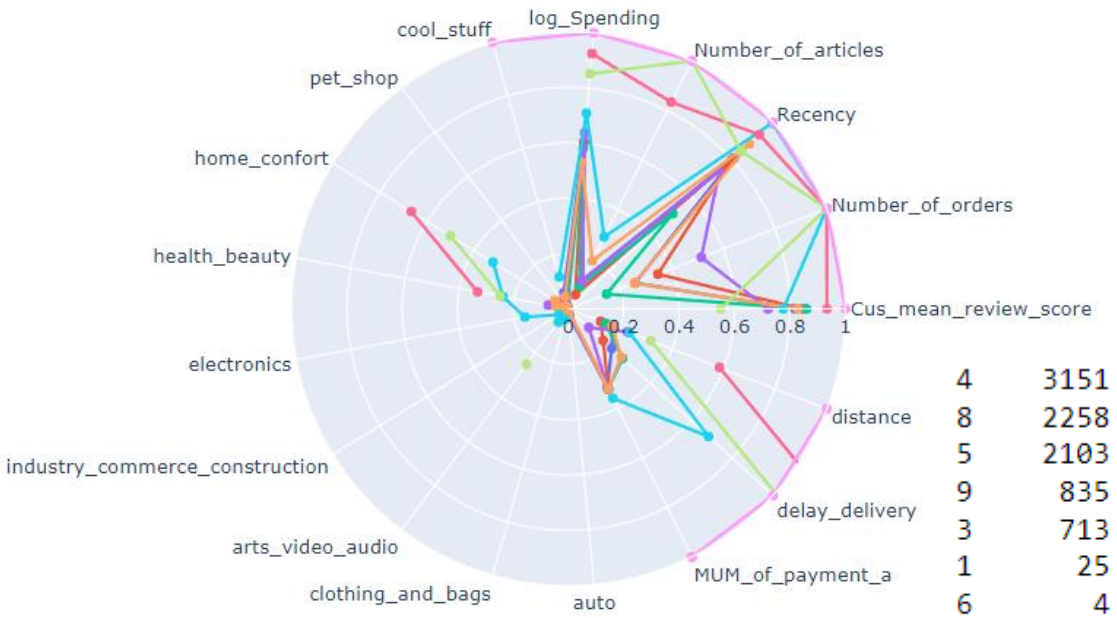


Autres méthodes de classification

Hierarchical Clustering

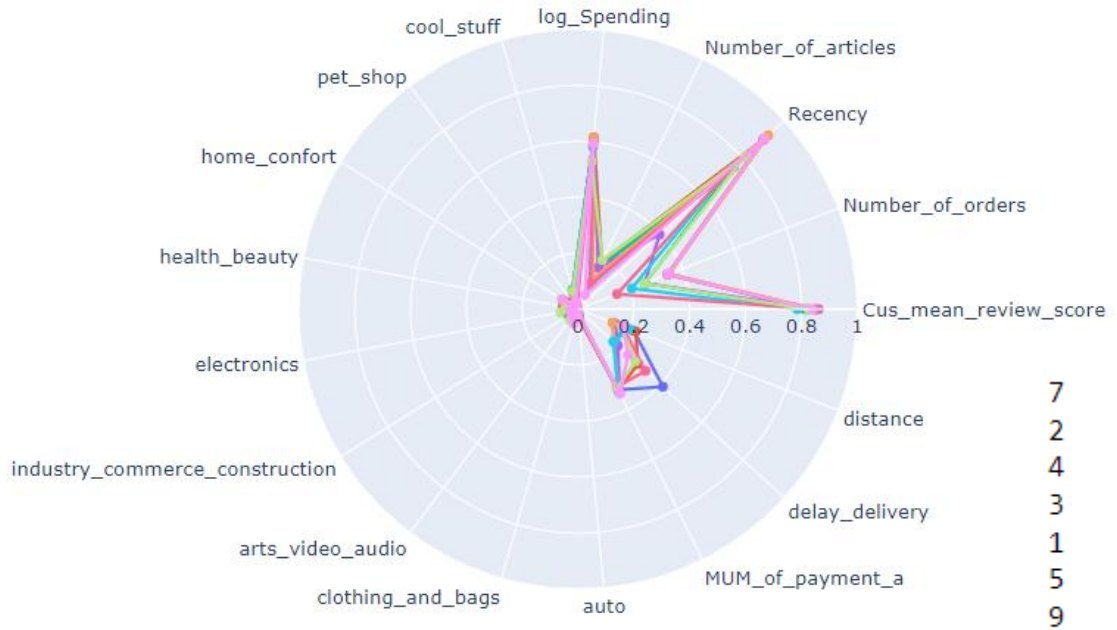
9 clusters

Meilleur score copenhet : 0,72 → method « centroid »



4	3151
8	2258
5	2103
9	835
3	713
1	25
6	4
2	3
7	1

Le plus proche du Kmeans : c = 0,69 → method « ward »



7	1589
2	1519
4	1447
3	1200
1	807
5	781
9	754
8	561
6	435

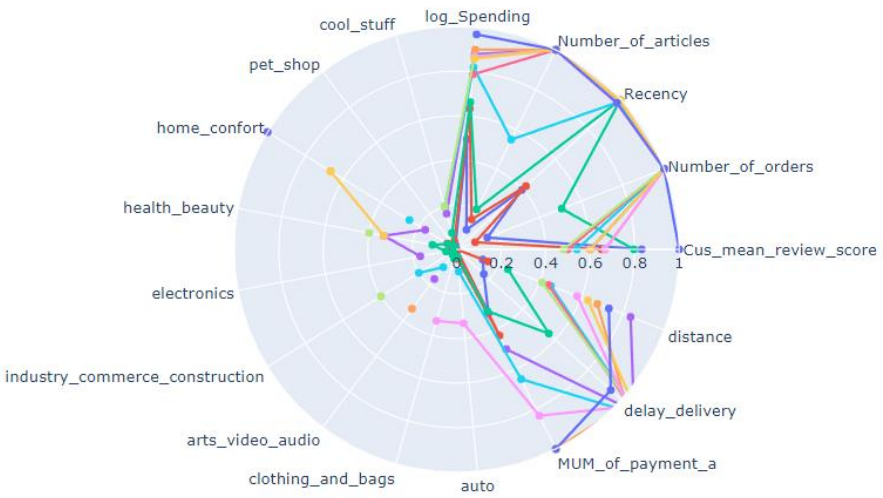
DBScan

Description des données

Avantages :
- Pas besoin de connaître le nombre de cluster à l'avance
- Identifie les valeurs aberrantes comme des bruits
- Capable de trouver assez bien des clusters de taille et de forme arbitraire

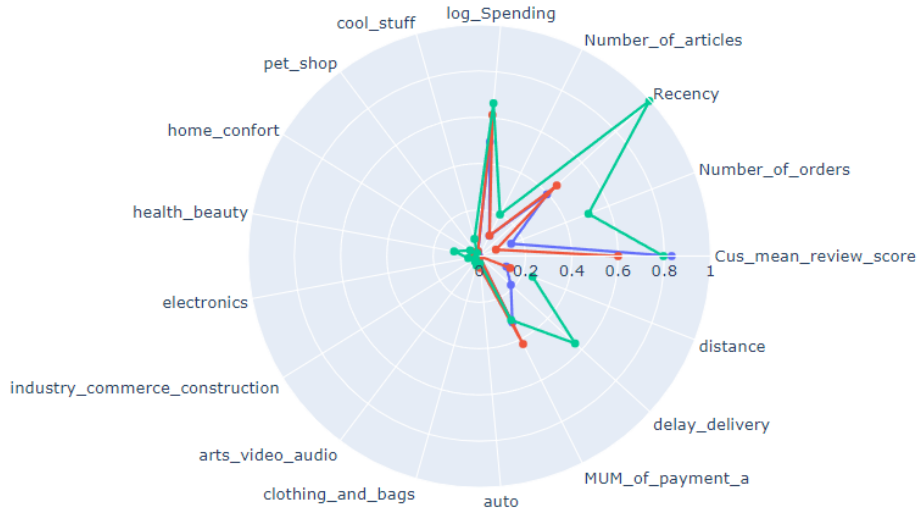
Désavantages :
- Faible face à des densités variables
- Difficile d'estimer ε en haute dimensions

```
eps / min_samples: 12 / 3
Estimated number of clusters: 10
Estimated number of noise points: 115
Silhouette Coefficient: -0.381
```



0	90530
1	246
-1	115
4	10
2	6
6	5
3	3
5	3
7	3
8	3
9	3

```
eps / min_samples: 12 / 10
Estimated number of clusters: 2
Estimated number of noise points: 230
Silhouette Coefficient: 0.188
```



0	90451
1	246
-1	230

Maintenance

Source : zdnet.fr

Au bout de combien de temps le modèle devient obsolète ?

A t0 → modèle M0
→ M0.fit(t0)

A t* → modèle M*
→ M*.fit(t*)

On les compare : M0.predict(t*) / M*.fit(t*)
(via le ARI Score)
ARI score : "*Rand index adjusted for chance*"

Similarity measurement between two clusterings.

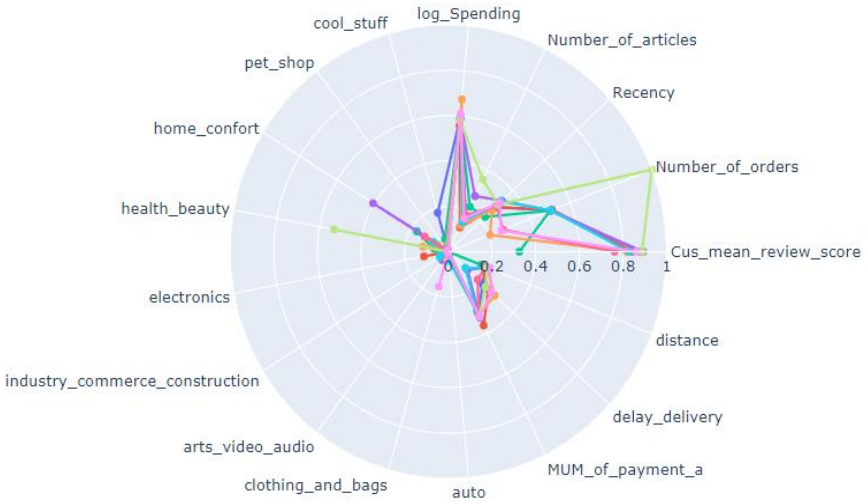
t0 = 1 an + 3 mois et 10 jours

t1 = t0 + 1 mois

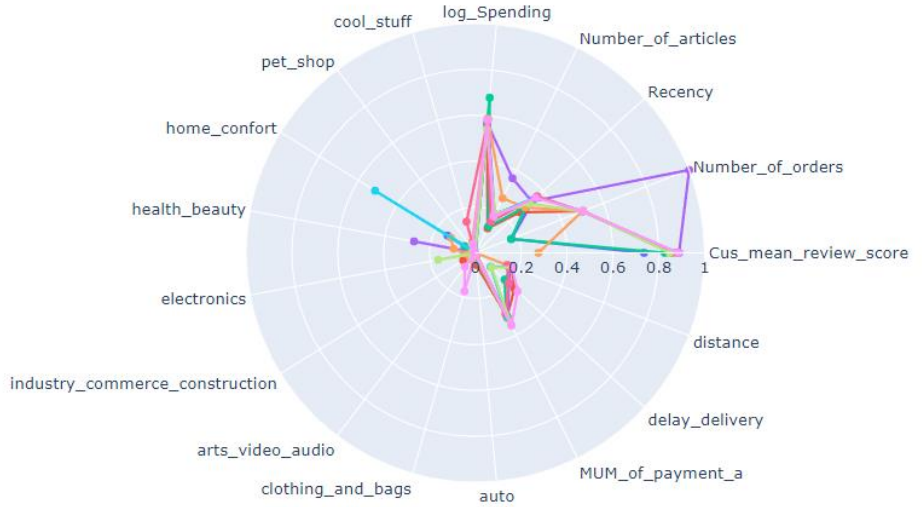
t2 = t0 + 3 mois

t3 = t0 + 6 mois

Au bout de combien de temps le modèle devient obsolète ?

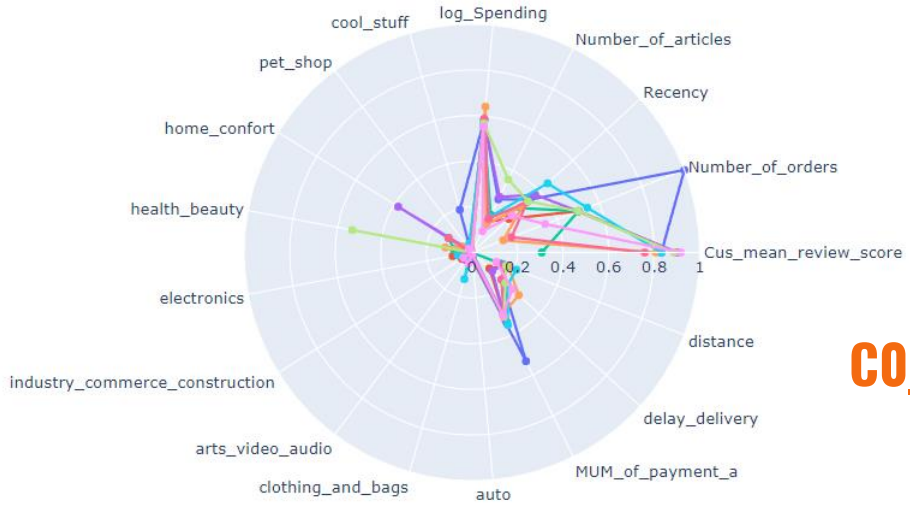


C0 = M0.fit(t0)



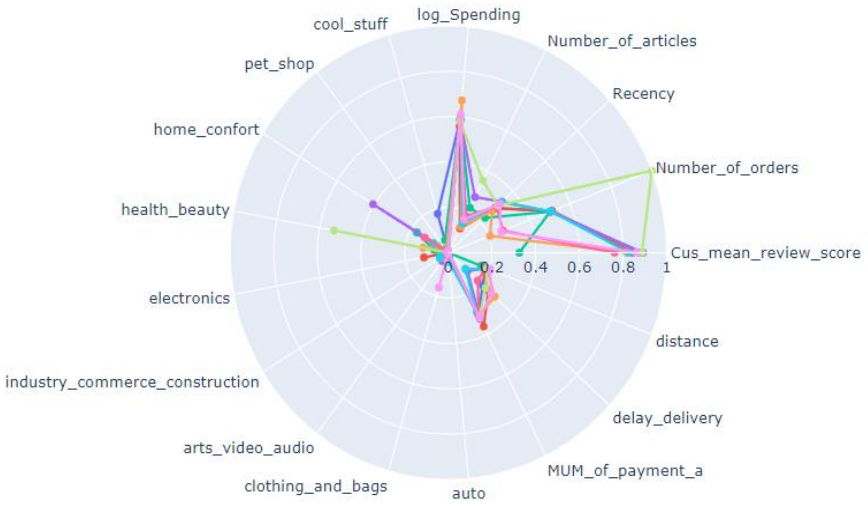
M3.fit(t3)

Ari =0.63

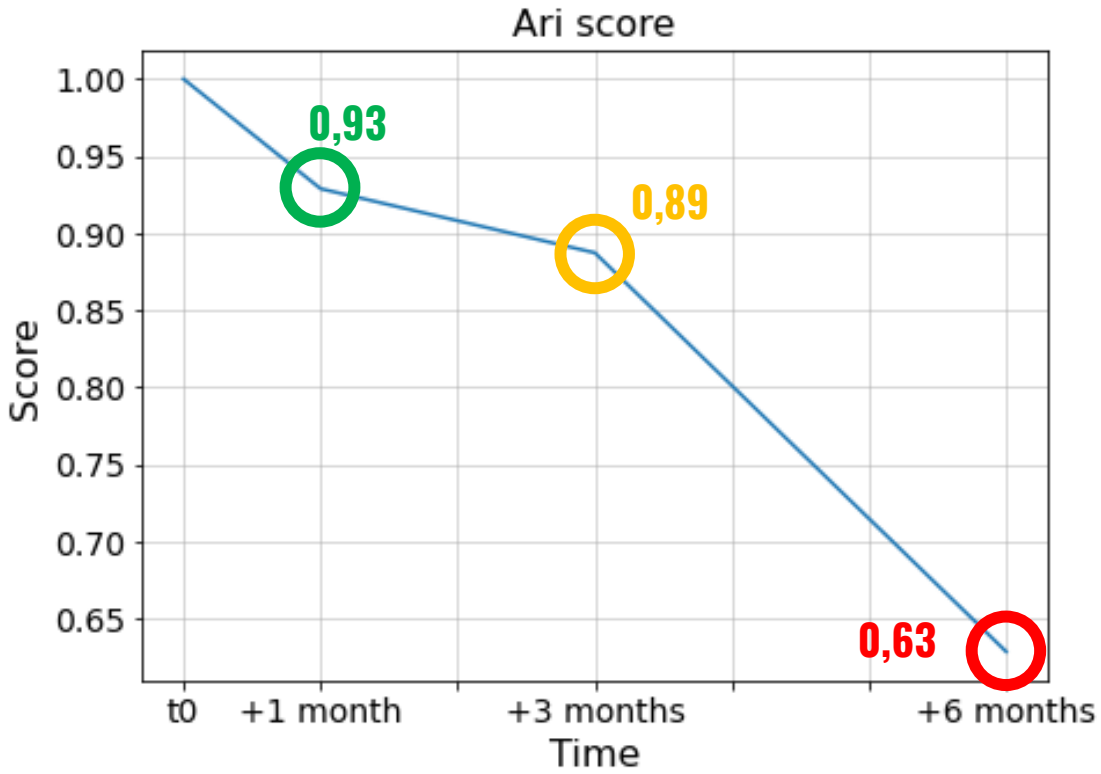


C0_3 = M0.predict(t3)

Au bout de combien de temps le modèle devient obsolète ?



C0 = M0.fit(t0)



ARI >= 0.90 excellent recovery
0.80 =< ARI < 0.90 good recovery
0.65 =< ARI < 0.80 moderate recovery
ARI < 0.65 poor recovery

Conclusions

Conclusions by Nick Youngson CC BY-SA 3.0 Alpha Stock Images

Conclusion

- BDD assez limitée en terme de nombre d'achat

  BDD assez limitée en terme de critères de persona précis
- Modèle du KMeans sélectionné

  Résultats obtenus
- Maintenance du modèle
- Pistes d'améliorations
- PEP8 ?

Fin de la présentation

Merci de m’avoir écouté !