Projet 4

Segmentation de clients d’un site de e-commerce

Plan de présentation

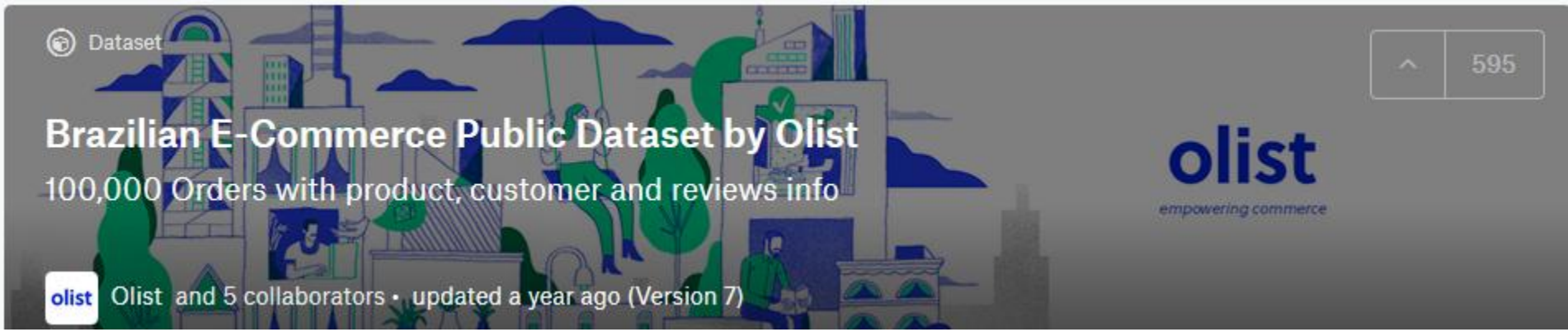
1. Présentation de problématique
2. Analyse exploratoire
    - Présentation de la base de données
    - Feature engineering
    - Analyses descriptives
3. Segmentation managériale
4. Réduction dimensionnelle
    - ACP
    - ACP à noyau
5. Clustering
    - Clustering hiérarchique
    - K-means
6. Evaluation de mise à jour
7. Conclusion

1. Présentation de problématique

- Site de e-commerce brésilien Olist a mis à disposition une base de données avec des commandes passées entre 2016 et 2018

- Notre mission est d'établir une segmentation de clients actionnable :
  - c'est-à-dire qu'elle devrait être interprétable
  - et utilisable par l'équipe commerciale, qui pourra utiliser l'information obtenue pour établir un plan d'actions afin de fidéliser les clients ou trouver des nouvelles pistes pour la publicité et autres démarches commerciales

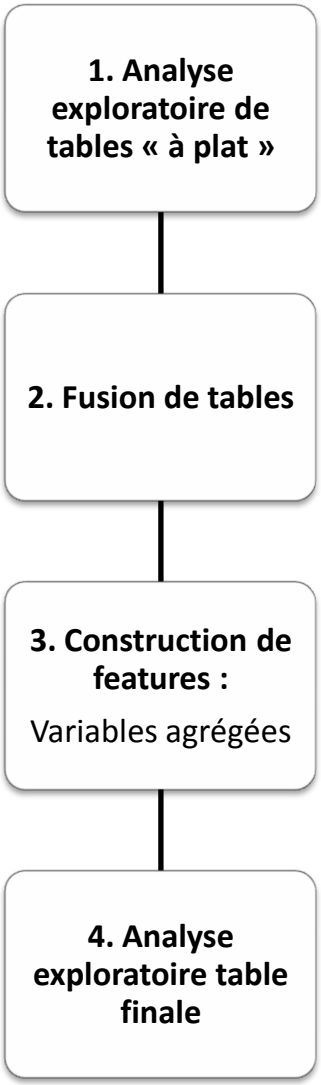
- Nous allons aussi évaluer la stabilité de clusters au fil de temps afin d 'établir un contrat de maintenance



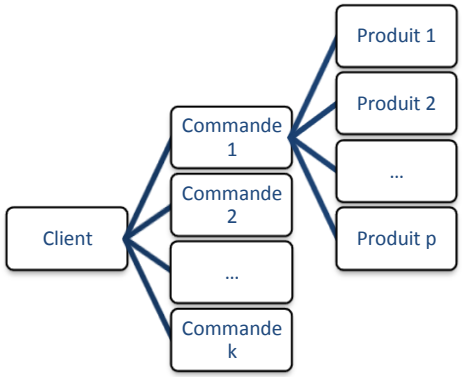
1. Présentation de problématique Outils

- Utilisation de bibliothèque de machine learning Scikit-learn
- Codage en PEP8 à l'aide de l'extension de Jupyter Notebook Autopep8
- Utilisation d'un module de fonctions créé pour le projet

2. Analyse exploratoire Workflow



2. Analyses exploratoires
Présentation de la base de données

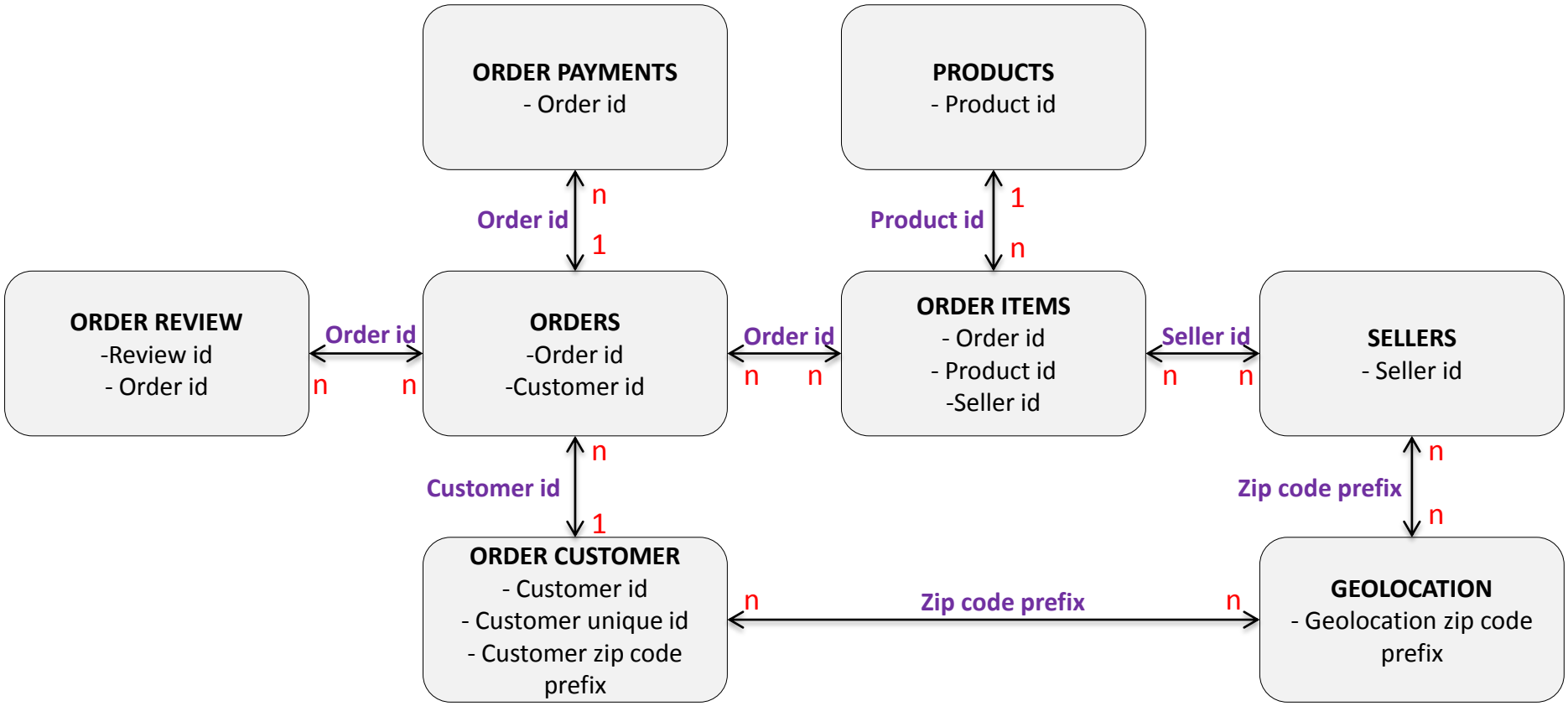


- Nombre de clients : 95k
- Nombre de commandes : 99k
- Nombre de produits achetés : 113k
- Nombre max de commandes par client : 16
- Nombre max d'articles par client : 24

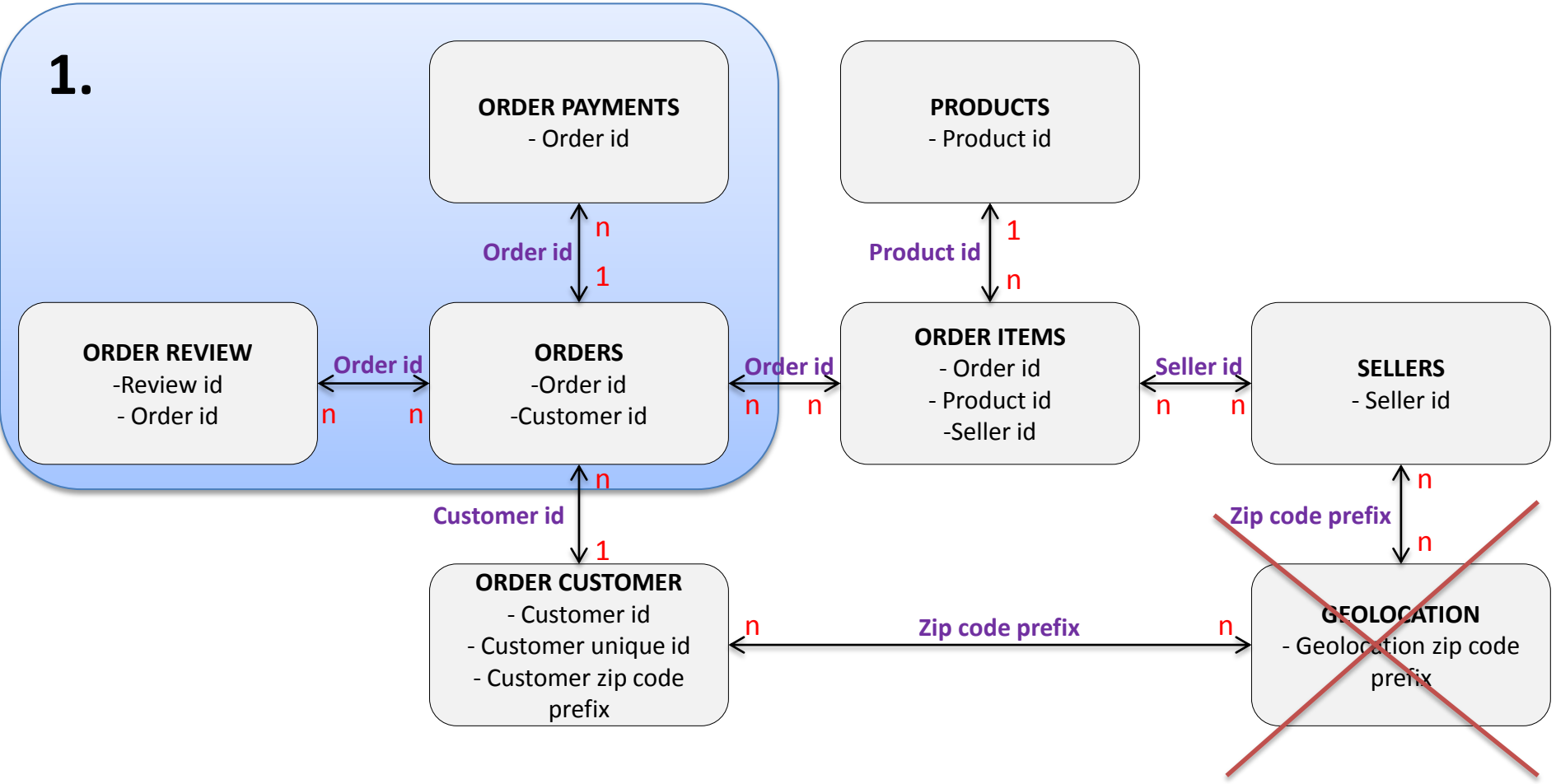
Nombre de commandes	Nombre de clients	Pourcentage
1	92507	96.95
2	2673	2.80
3	192	0.20
4	29	0.03
5	9	0.01
6	5	0.01
7	3	0.00
9	1	0.00
16	1	0.00

Nombre de produits	Nombre de clients	Pourcentage
1	83551	87.56
2	8996	9.43
3	1672	1.75
4	632	0.66
5	254	0.27
6	198	0.21
7	46	0.05
8	16	0.02
9	11	0.01
10	11	0.01
11	11	0.01
12	9	0.01
13	2	0.00
14	3	0.00
15	2	0.00
16	1	0.00
18	1	0.00
20	2	0.00
21	1	0.00
24	1	0.00

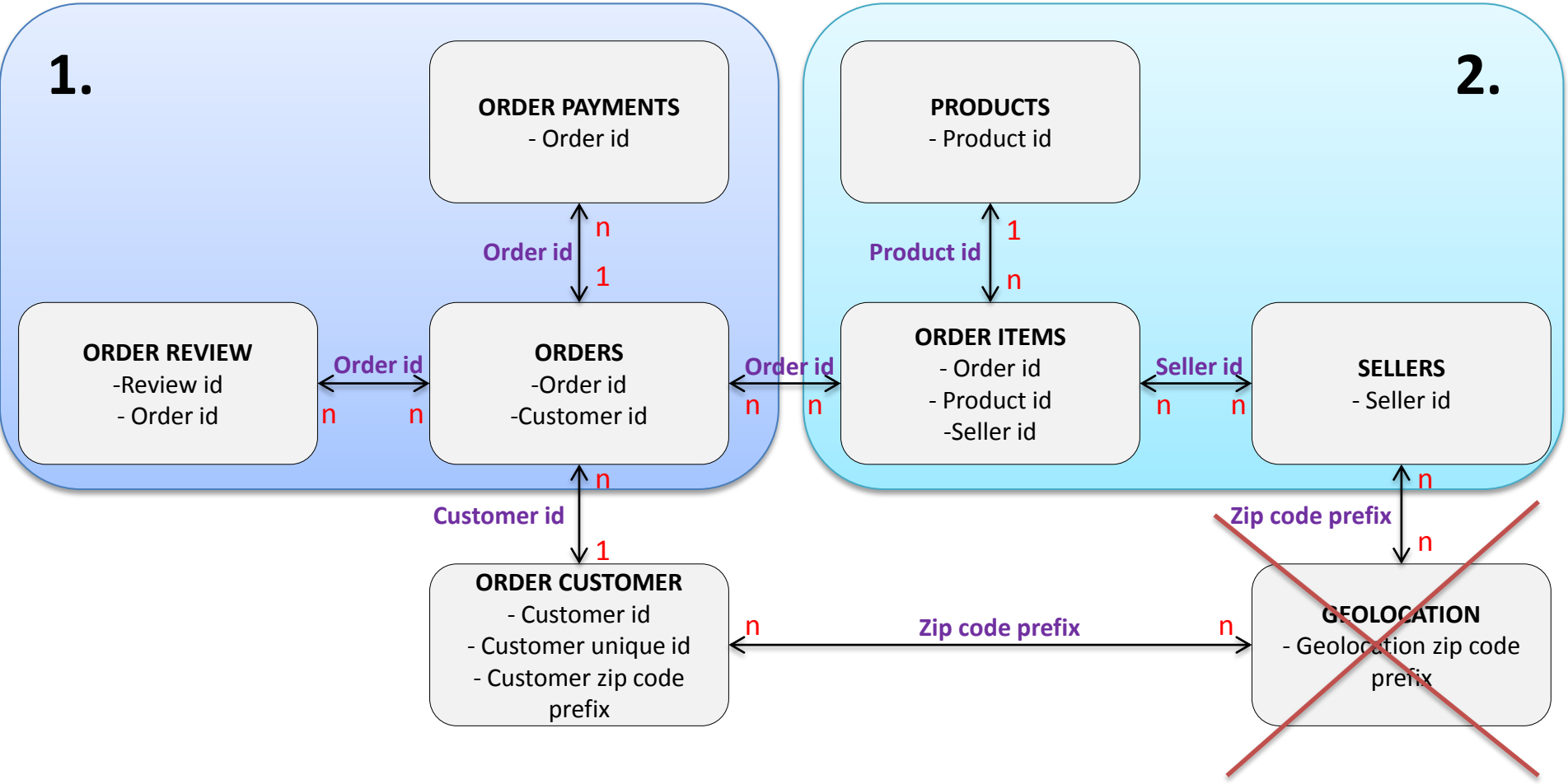
2. Analyse exploratoire
Présentation de la base de données



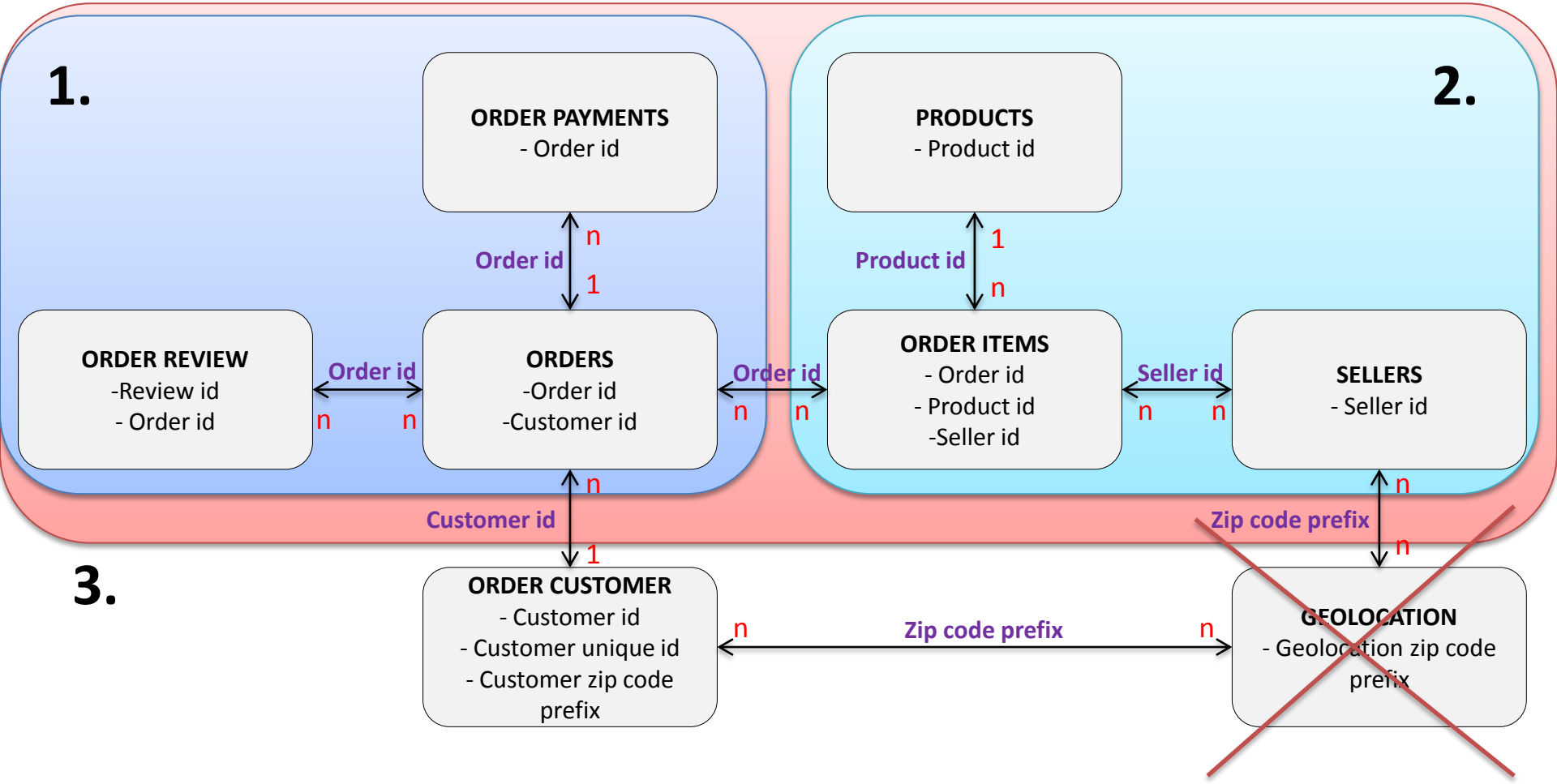
2. Analyse exploratoire
Présentation de la base de données



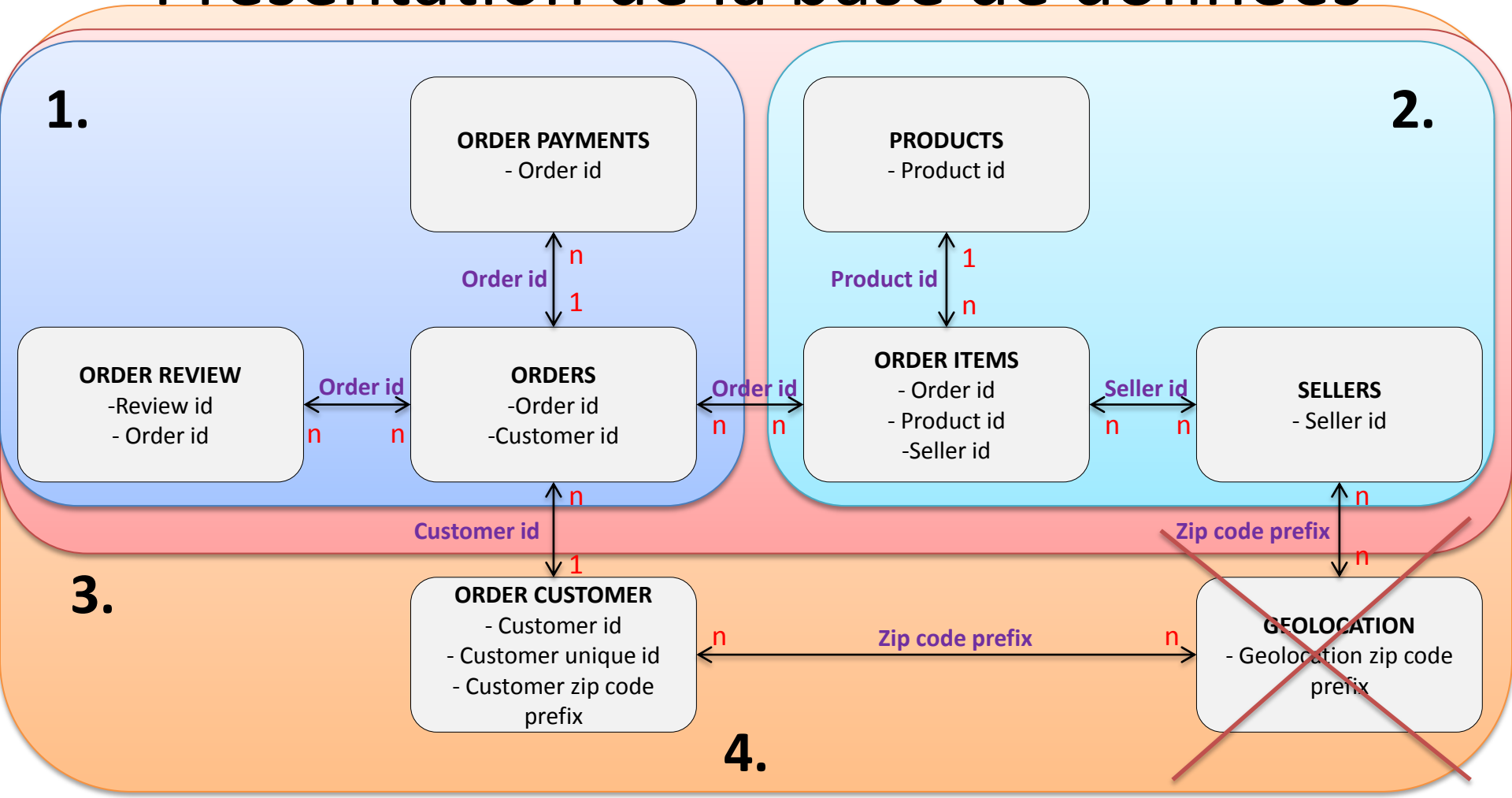
2. Analyse exploratoire
Présentation de la base de données



2. Analyse exploratoire
Présentation de la base de données



2. Analyse exploratoire
Présentation de la base de données



2. Analyse exploratoire
Features engineering

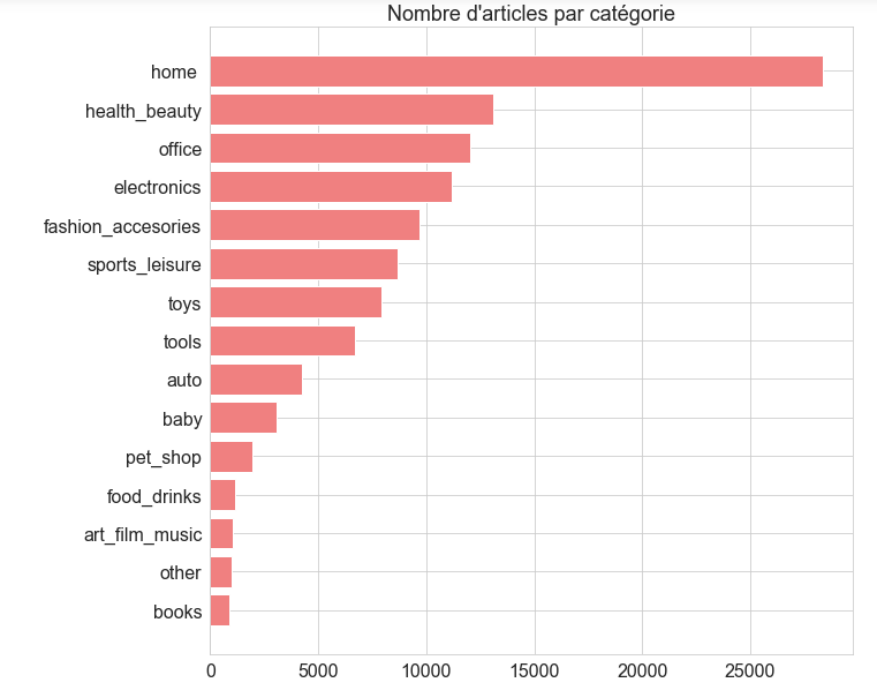
- **Création de features par agrégation**:
  - Min (Ex. : délais min : ancienneté de la dernière commande)
  - Max (Ex. nombre max d'échéances, prix max, …)
  - Moyen (Rating moyen, prix moyen, valeur moyenne de la commande, …)
  - Nombre (Nombre de commandes par client, nombre d'articles, …)

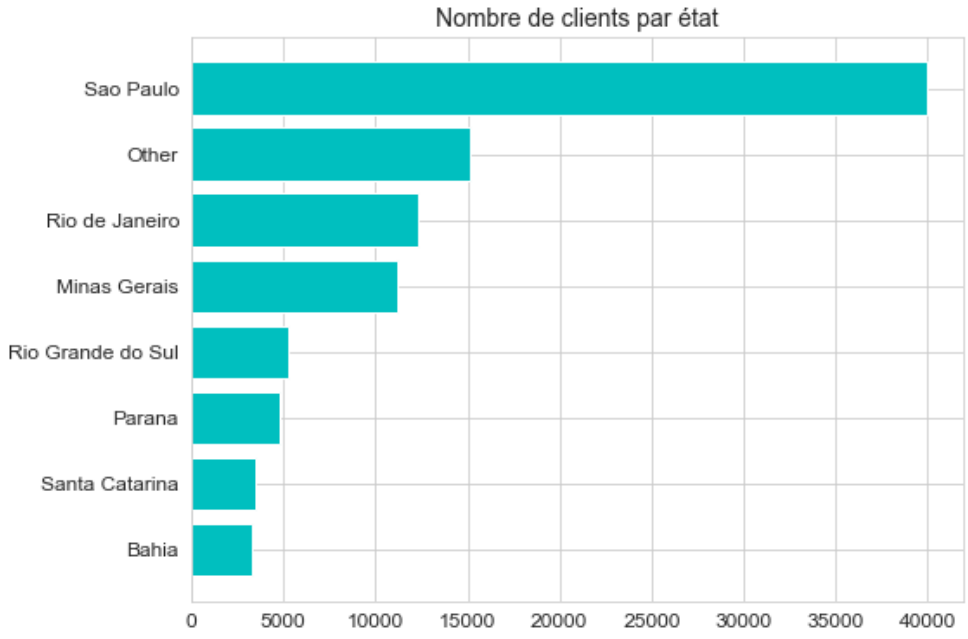
- **Recodage de variables catégorielles non ordonnées:**
  - Catégorie d'articles : 1 feature de nombre d'articles achetés par le client par classe
  - Etat d'origine de client : 1 feature binaire / classe

- **Création de nouvelles features:**
  - Ex. : Client vient d'une ville de + d'1milion d'habitants

2. Analyse exploratoire
Analyse descriptive

Features catégorielles recodées





3. Segmentation managériale
La méthode RFM

Méthode RFM :

- **R**écence: Nombre de jours depuis la dernière commande
- **F**réquence: nombre d'achats
- **M**ontant de l'achat

Le principe de segmentation est le suivant :

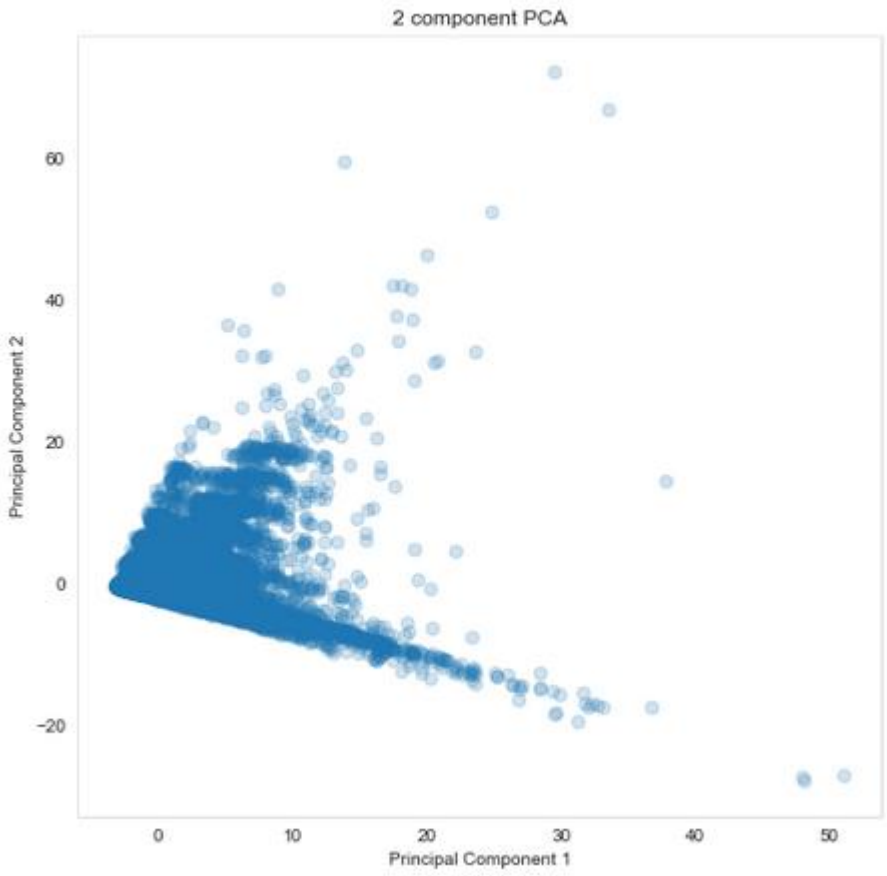
1. Calcul de scores
   - 4 classes de récence et montant (découpage par quartile)
   - 2 classes de fréquence (1 achat ou plus)

2. Segmentation dans les classes
   - Par récence :
     - Actif: R = 1
     - Warm: R = 2
     - Cold: R = 3
     - Presque perdu: R = 4
   - Autres centres d'intérêt :
     - Meilleur: R = 1, F = 1 et M = 1
     - Nouveau: R = 1 et F = 2
     - Actif - valeur élevée: R = 1 et M = 1 ou 2
     - Warm - valeur élevée: R = 2 et M = 1
     - A reconquérir: R = 3 ou 4 ET (F = 1 ou M = 1)

	Effectif	Pourcentage
RFM_meilleur	209	0.22
RFM_nouveau	23217	24.33
RFM_actif	24039	25.19
RFM_actif_valeur_elevee	12269	12.86
RFM_warm	23709	24.85
RFM_warm_valeur_elevee	5937	6.22
RFM_a_reconquerir	12790	13.40
RFM_cold	23886	25.03
RFM_presque_perdu	23786	24.93

4. Réduction dimensionnelle
Analyse en composantes principales

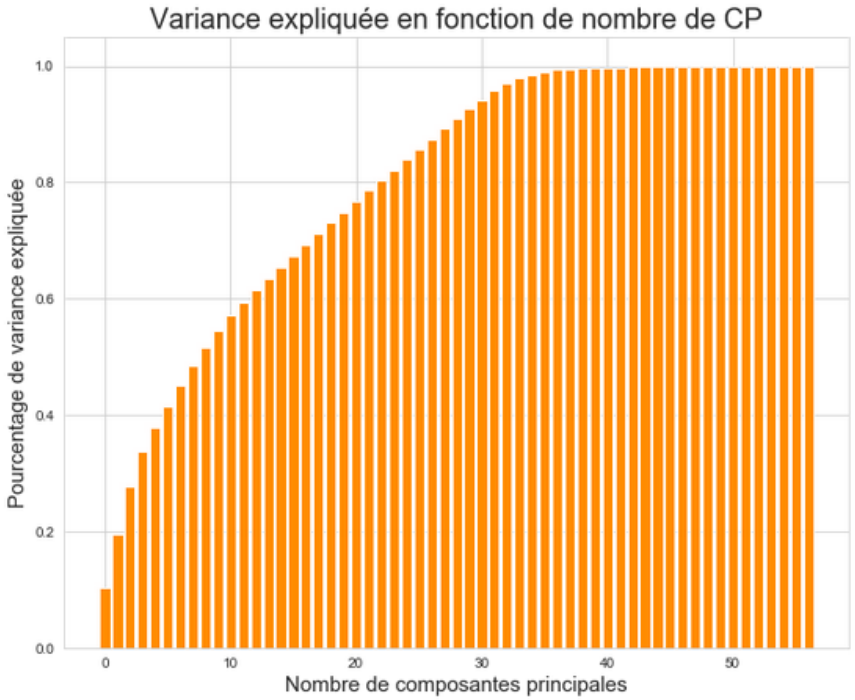
1. Visualisation de données



2. Réduction de dimensions

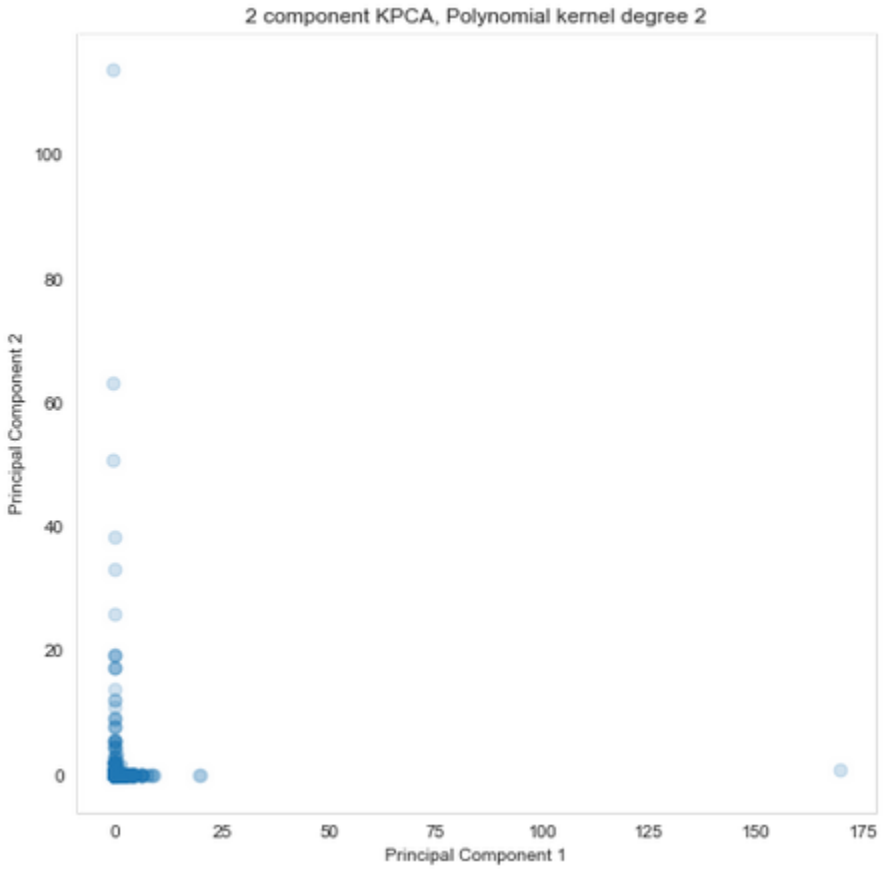
Variance expliquée:

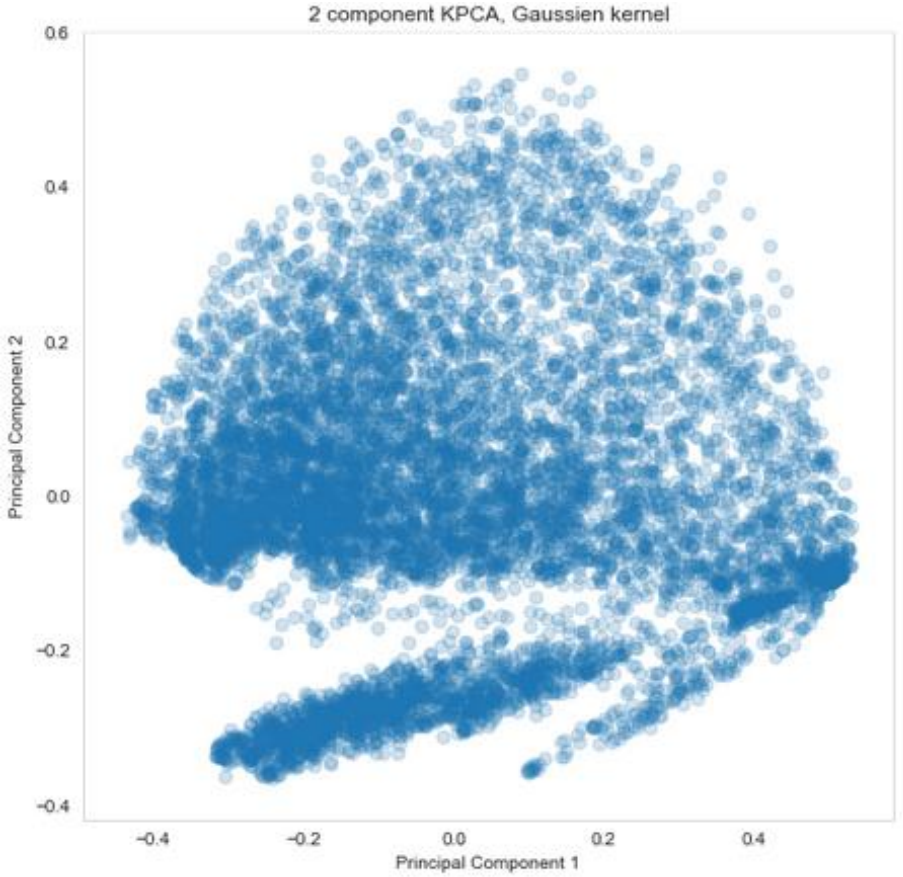
- 2 CP => 19 %
- 30 CP => 92 %



4. Réduction dimensionnelle
ACP à noyau

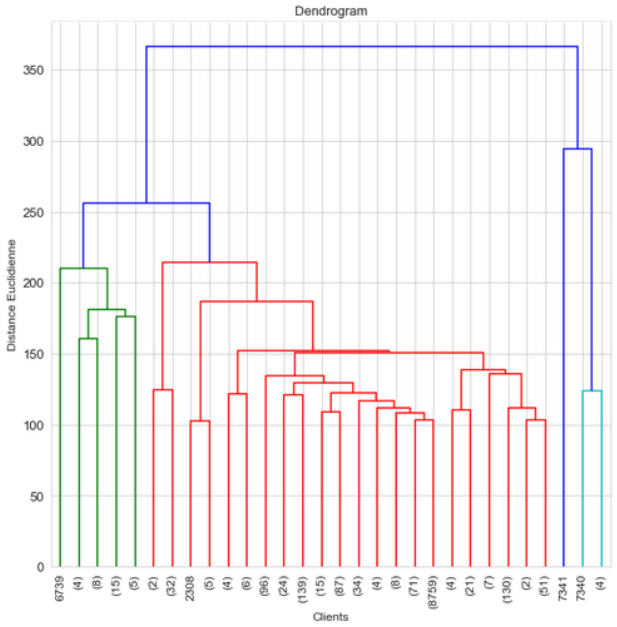
- L'analyse en composantes principales à noyau était effectué sur un échantillon de données qui contient 10 % (~9k) de clients.
- Le but : visualiser les données

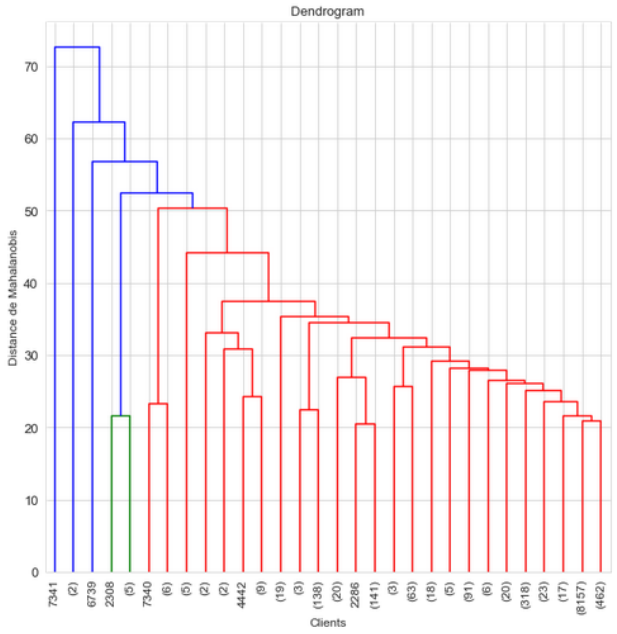


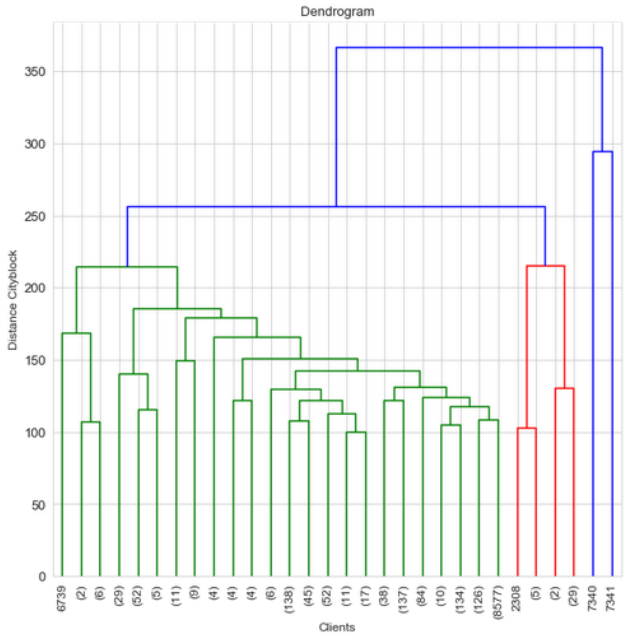


5. Clustering
Clustering hiérarchique

- Méthode appliquée sur l'échantillon + réduction dimensionnelle par ACP (30 CP)
- But : Visualiser le nombre de clusters optimal
- Distances utilisées :
  - Distance Euclidienne
  - Distance de Mahalanobis
  - Distance CityBlock



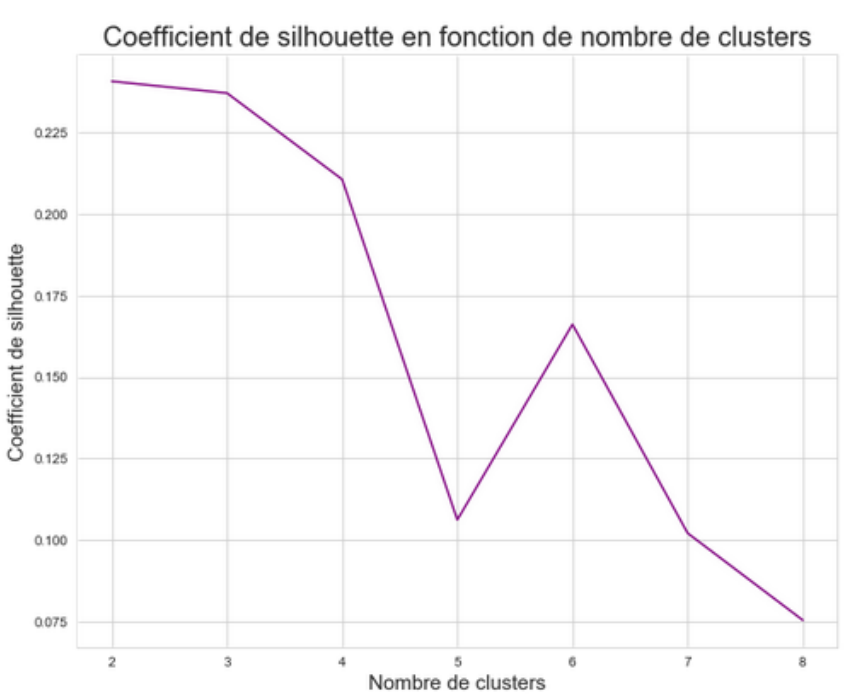




5. Clustering
K - means

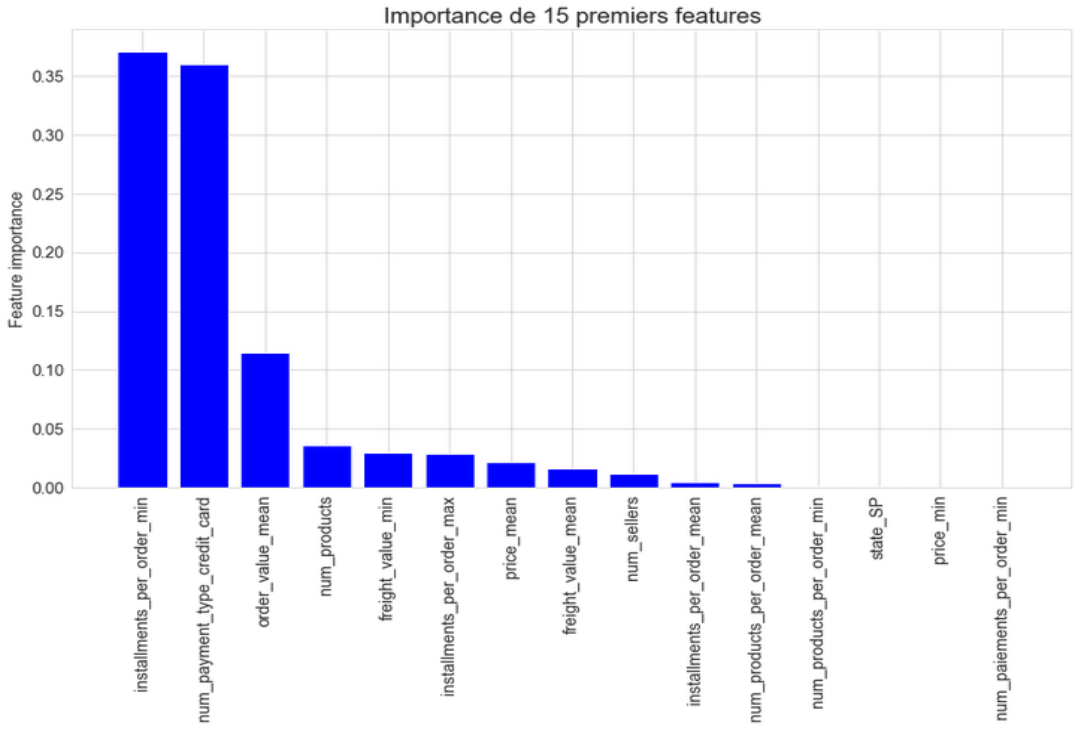
- Algorithme k-means était appliqué sur l'échantillon de données afin de trouver le nombre de clusters optimal
- Critère: coefficient de silhouette
- Coefficient optimal si k = 2
- Point de vue commercial, 3 clusters peuvent donner plus d'information sur les clients. La différence de coefficient de silhouette entre k = 2 et k = 3 n'est pas très importante

=> Choix de k = 3



5. Clustering
K-means – evaluation

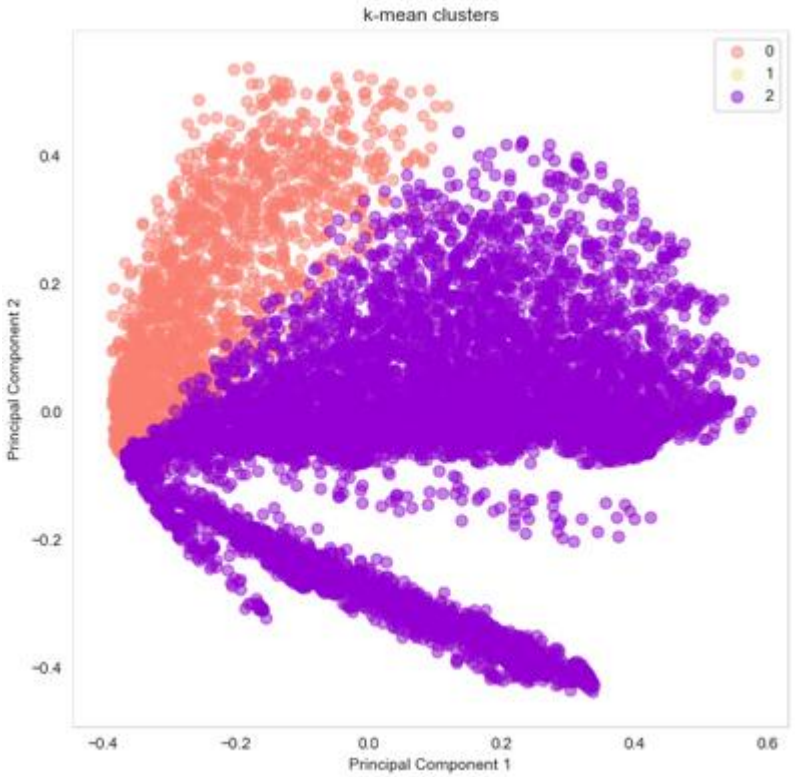
- Construction d'un modèle supervisé (arbre de décision) avec les clusters comme cible.
- But:
  - Définir l'importance de features
  - Calculer la précision de prédiction de modèle : Accuracy score = 96.8 %



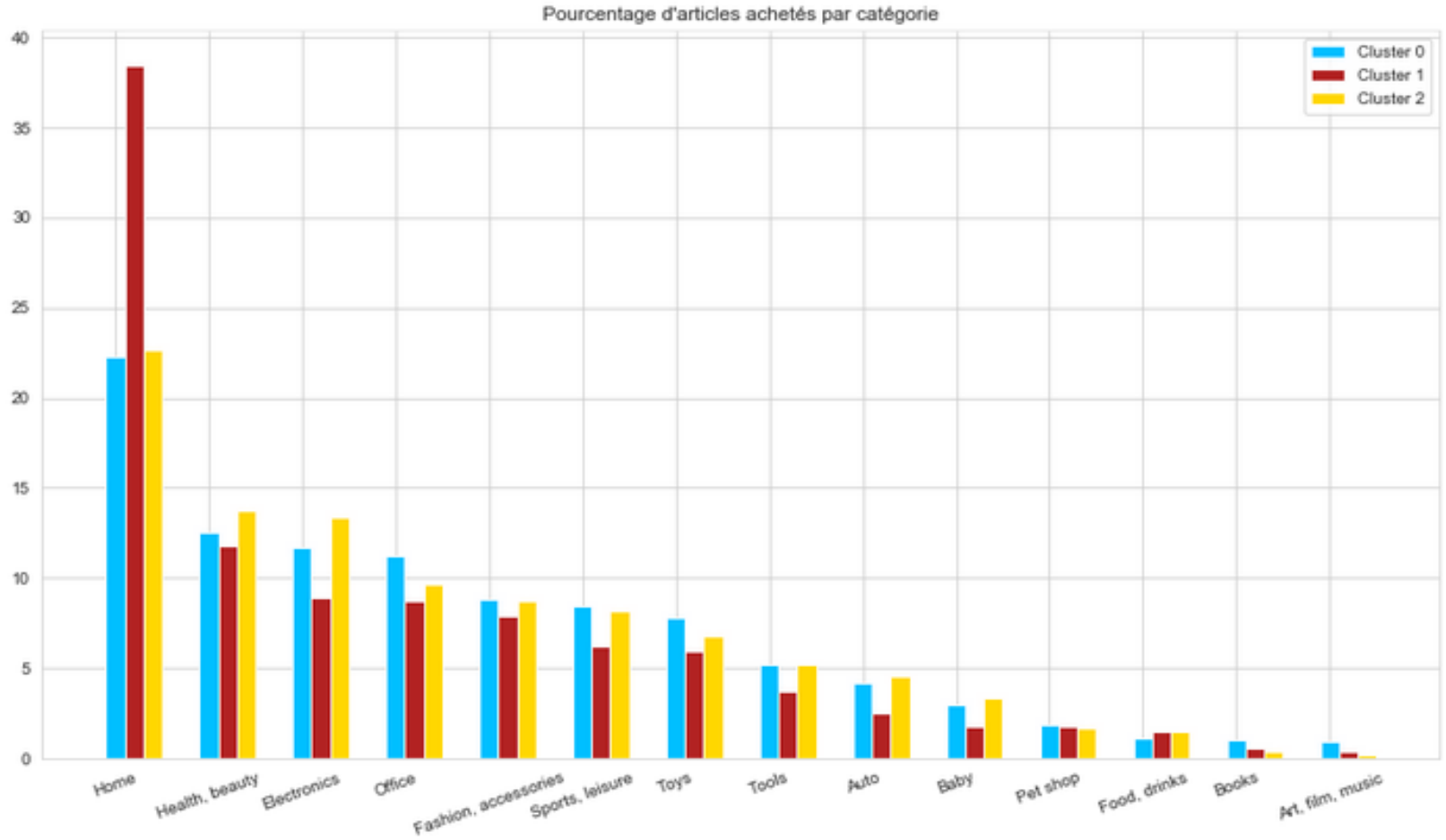
5. Clustering
K – means : Description de clusters

Workflow:

1. Reconstitution de table de départ, ajout de résultat de clustering
2. Effectif & pourcentage de clusters
3. Visualisation à l'aide de l'ACP à noyau sur l'échantillon de données
4. Test de Kruskal Wallis pour identifier les différences entre les clusters parmi les features numériques
5. Visualisation de catégories
6. Description de chaque cluster



5. Clustering
K – means : Description de clusters



5. Clustering
K – means : Description de clusters

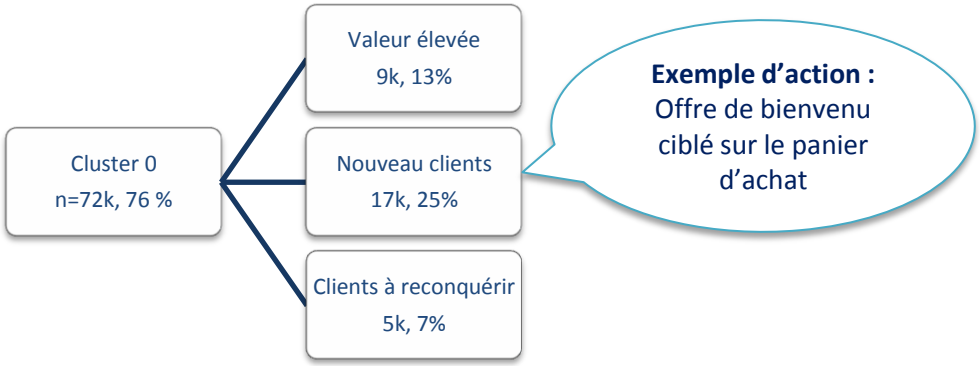
- **<u>Cluster 0:</u> Acheteur d'un article pas cher payé en une fois**
  - n = 72k, 76 % de population
  - Clients qui achètent un article pas très cher (med = 69 BRL) / Valeur commande (med = 85 BRL)
  - Paiement en une fois
  - Score de satisfaction maximal (= 5)

- **<u>Cluster 1:</u> Acheteur de plusieurs articles payé en plusieurs fois**
  - n = 9k, 10 % de population
  - Clients qui achètent plusieurs articles à la fois
  - Produits pas cher (med = 60 BRL) / Valeur de commande (med = 160 BRL)
  - Nombre moyen d'échéances = 3
  - Score de satisfaction en dessous de la moyenne (= 4)
  - Achètent plus de produits de catégorie « Home » que les autres segments

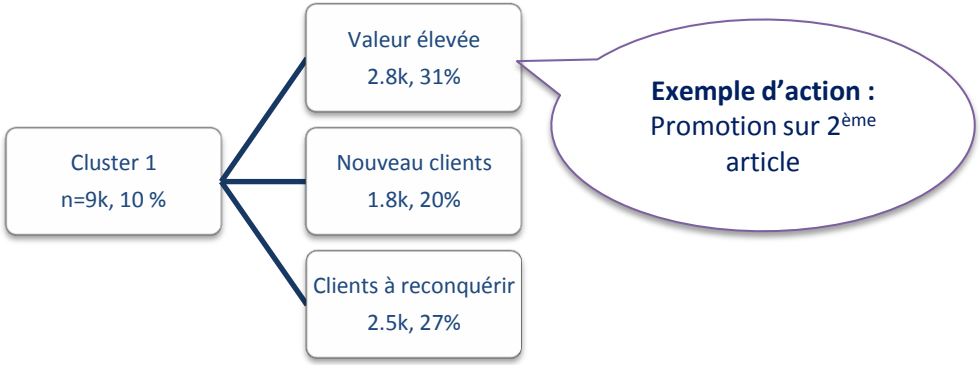
- **<u>Cluster 2:</u> Acheteur d'un article cher avec un paiement à beaucoup d'échéances**
  - n = 14k, 14 % de population
  - Clients qui achètent des produits cher (med = 230 BRL) / Valeur commande (med = 270 BRL)
  - Nombre moyen d'échéances = 8
  - Score de satisfaction maximal (= 5)
  - Clients sont moins représentés à Sao Paolo que dans les autres segments, sous-représentés dans une grande ville

5. Clustering
K – means + microsegmentation

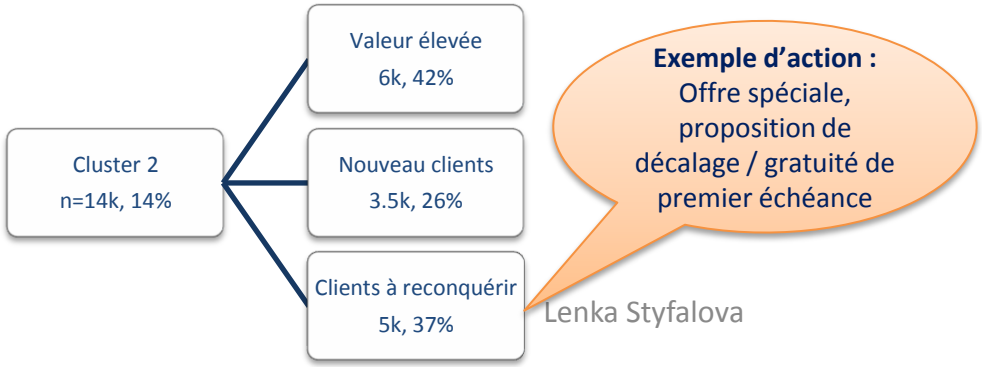
- **<u>Cluster 0:</u> Acheteur d'un article pas très cher payé en une fois**



- **<u>Cluster 1:</u> Acheteur de plusieurs articles pas cher payé en plusieurs fois**



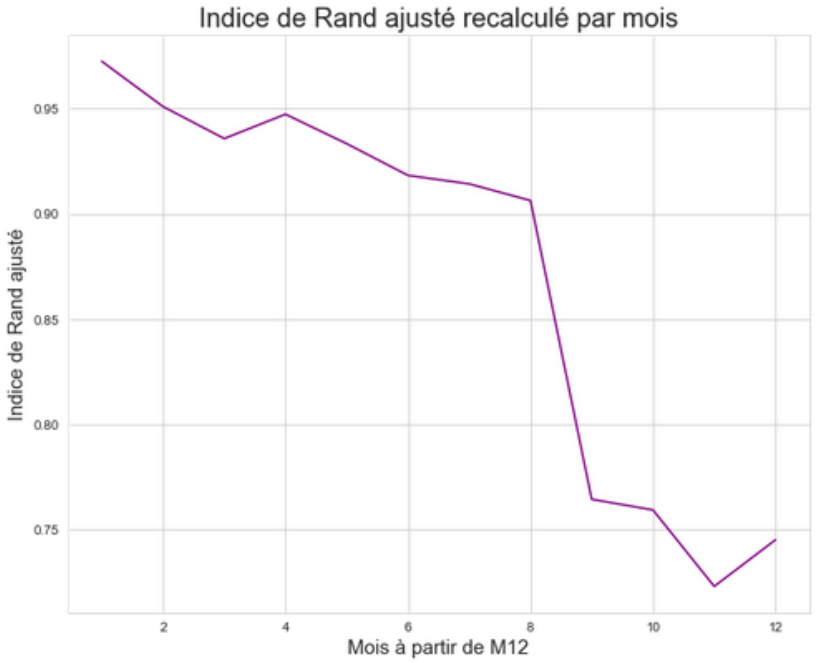
- **<u>Cluster 2:</u> Acheteur d'un article cher avec un paiement à beaucoup d'échéance**



6. Evaluation de mises à jour

- Séparation de données en « baseline » : 12 mois à partir de 1er achat
- Injection de données par mois à partir de M13, nouveau clustering
- Calcul d'Indice de rand ajusté entre les clusters de baseline et nouvelle segmentation
- Calcul approximatif : seulement la dernière commande de chaque client est prise en compte

Conclusion : Le score ARI commence à devenir instable avec une tendance à la baisse à partir de M7. Dans l'idéal, la maintenance de l'algorithme devrait être faire à fréquence biannuelle.



6. Conclusion

Nous avons réussi à créer des segments actionnable à l'aide d'algorithmes de machine learning en combinaison avec des techniques de marketing.

- La segmentation peut être améliorée par:
  - Création d'une macro-segmentation par type de produit (site vend des produits très hétérogènes)
  - Ajout d'information sur la réactivité de clients vis-à-vis des promotions
  - Analyse d'évolution de segments dans le temps

  => Selon les attentes de l'équipe marketing (quel type de clients cibler, pourquoi, quels sont les moyens, …)

- Eléments à préciser pour le contrat de maintenance:
  - Fonctionnement de l'algorithme : en temps réel / en batch, avec quel fréquence ?
  - Faut-il recalculer la segmentation après chaque achat d'un client ?
  - Serait-il souhaitable d'établir une macro segmentation statique (par exemple par catégorie de produit ou type de client – personne physique ou société etc.) ?