Segmentez des clients d'un site e-commerce

Data Science | Projet 5

Fırat Yasar

19/11/2021

Sommaire

Présentation

Présentation de la problématique

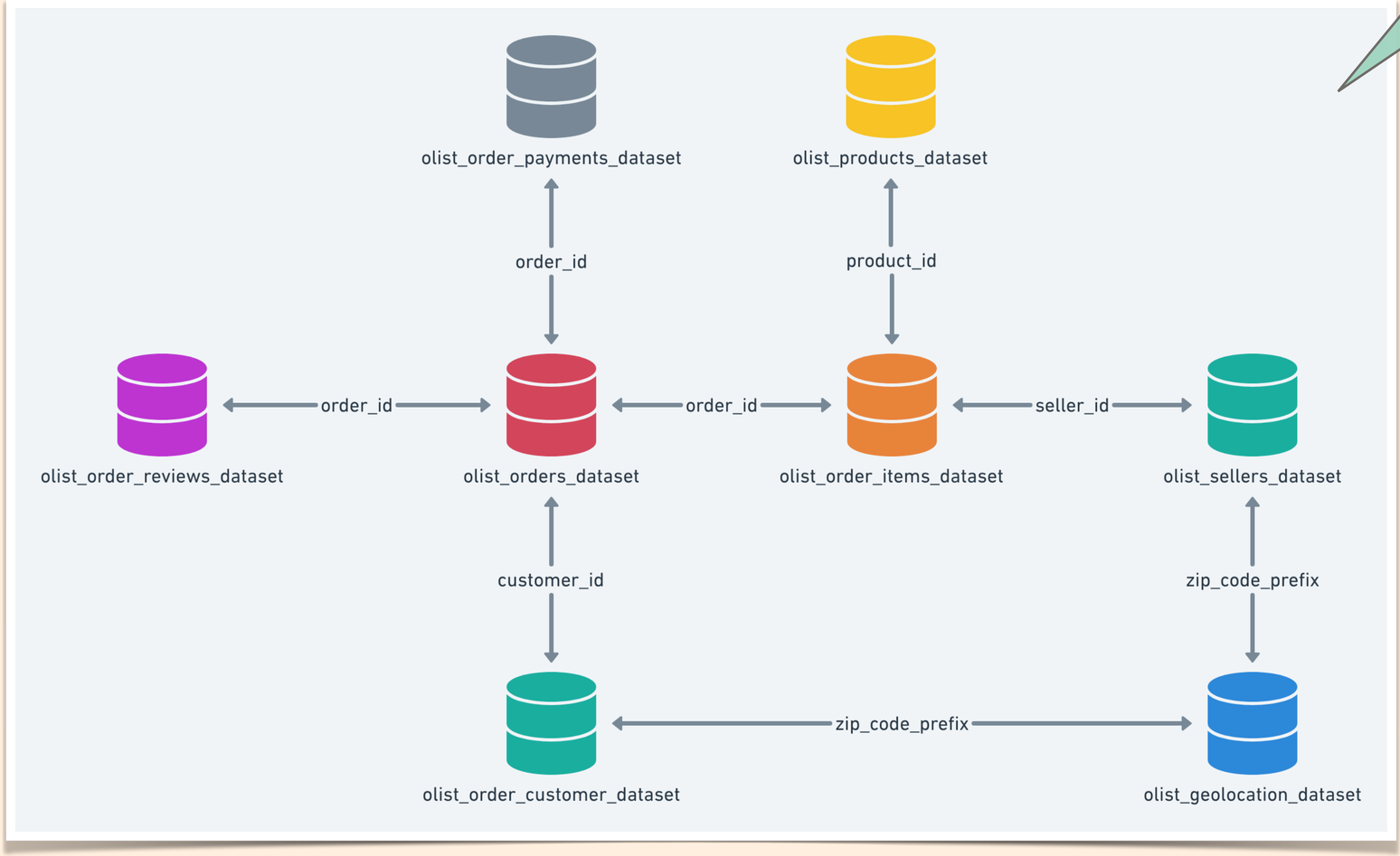
- L'objectif : Segmentation des clients
  - Notre objectif est de comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.
- Les missions
  - La segmentation fourni doit être **exploitable** et facile d'utilisation pour l'équipe marketing
  - Une **proposition de contrat de maintenance** basée sur une analyse de la stabilité des segments au cours du temps
  - Le code fourni doit respecter la convention **PEP8**, pour être utilisable par Olist.
- Pour cela, nous avons à notre disposition la base de données :

https://www.kaggle.com/olistbr/brazilian-ecommerce

Découverte du jeu de données

1 dataset supplémentaire pour la traduction des 71 catégories de produits :

Portugais == Anglais



DATA

119 151 lignes
30 colonnes

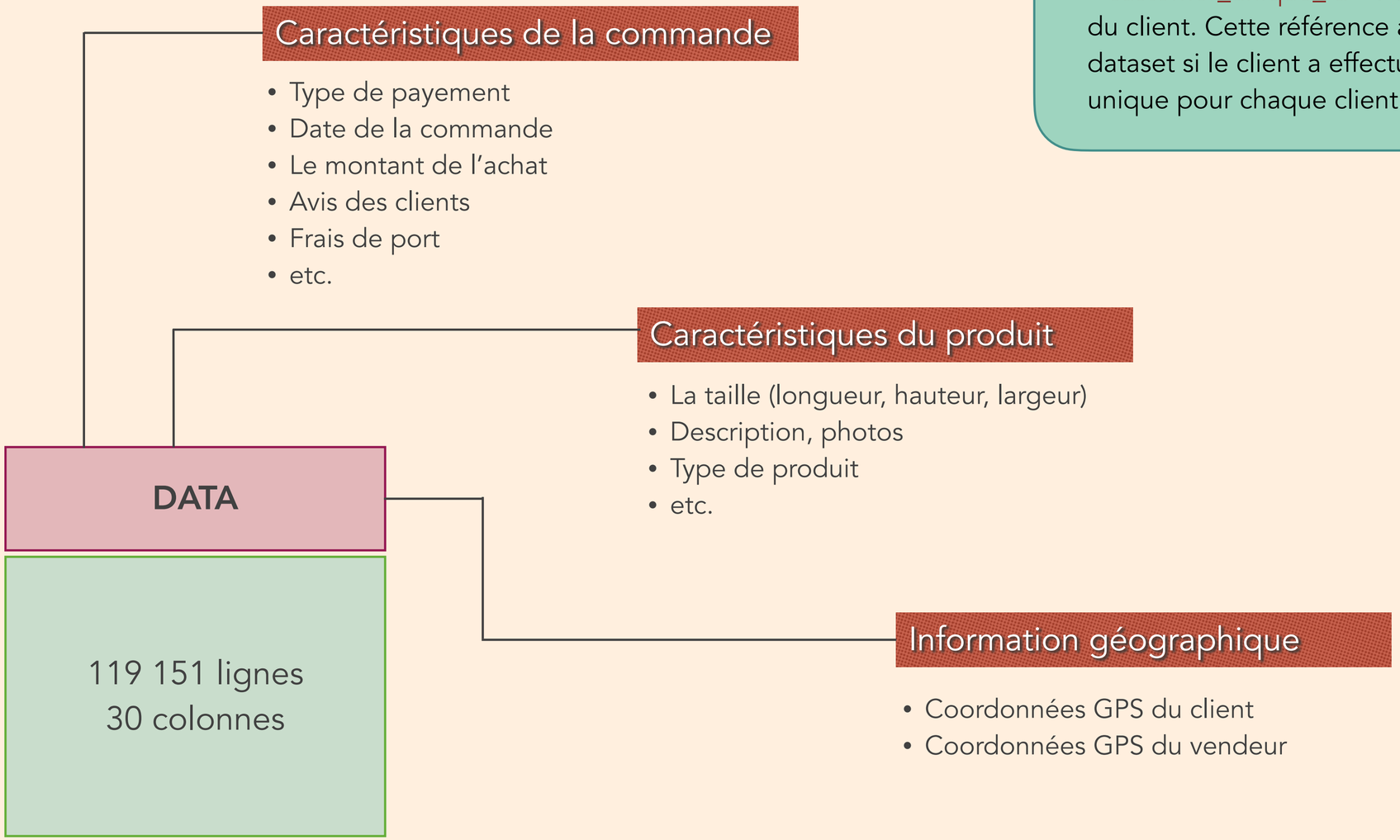
Concaténation des 8 datasets

Découverte du jeu de données

- « customer_id » : Cette clé représente la référence de la commande. Elle est toujours unique.

Une ligne = une commande

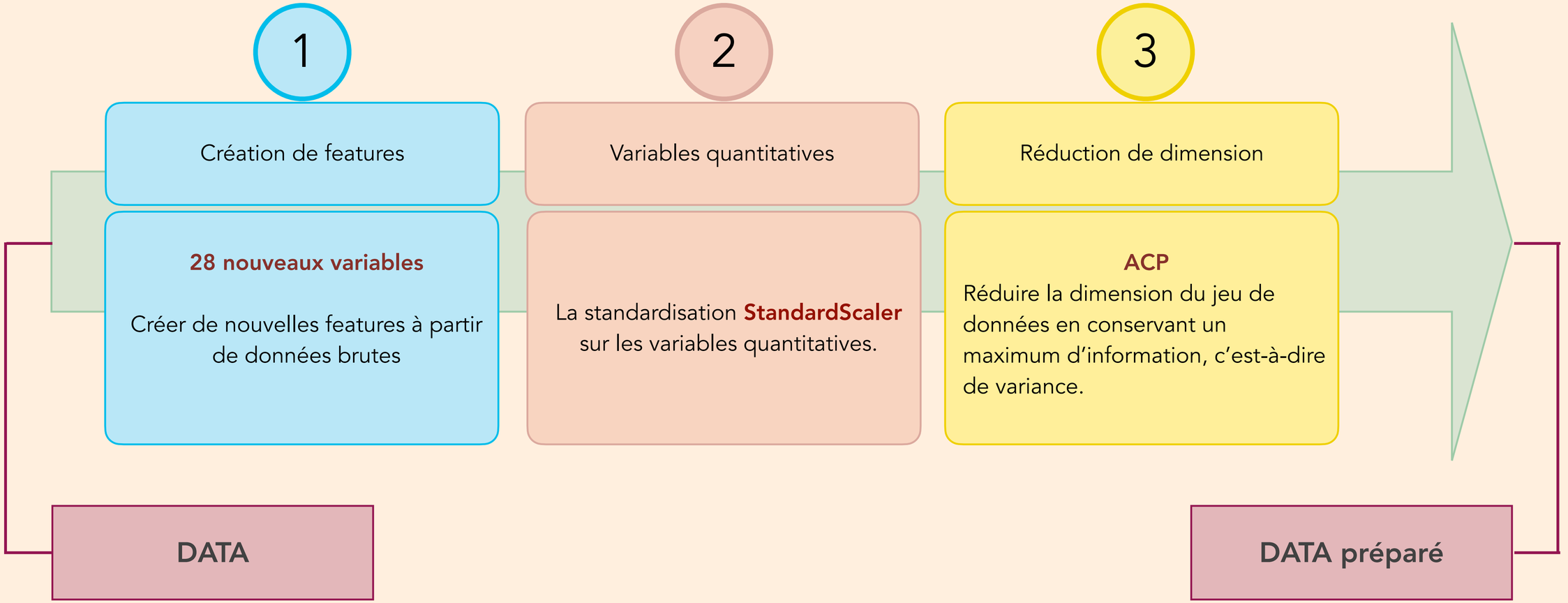
- « customer_unique_id » : Cette clé représente la référence du client. Cette référence apparaît plusieurs fois dans le dataset si le client a effectué plusieurs commandes. Elle est unique pour chaque client.



Feature engineering

Preprocessing

- Préparation des données
- Chaque ligne de DATA préparé représente un client.



Nouveaux features (attributs)

- Transformation des données brutes en la forme la plus appropriée pour les algorithmes de machine learning

- 3 catégories de nouveaux features

Categorie A.

Caractéristiques du client

Feature	Explication
ancienneté_du_client	premier achat (en jours)
date_derniere_commande	date du dernier achat
nb_commandes	le nombre de commandes effectués par le client
nb_payements	le nombre de payements effectués par le client
nb_produits	le nombre de produits achetés par le client
depense_total	somme des achats cumulés
review_score	le note donné par le client (min)

Categorie B.

Caractéristiques de commande

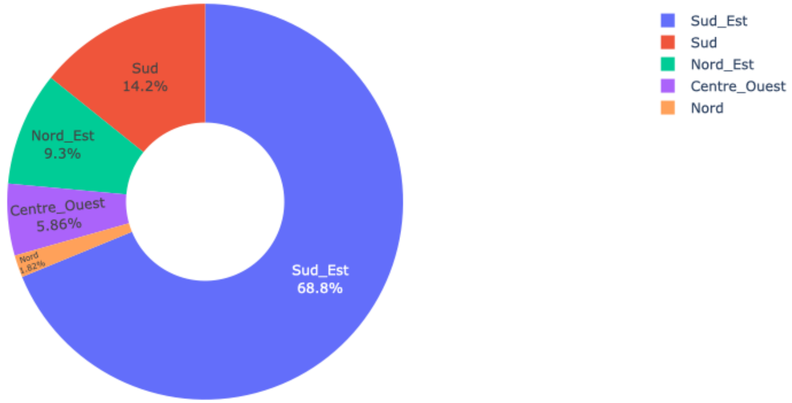
Feature	Explication
montant_commande_max	le montant maximum dépensé par le client
montant_produit_max	le prix maximum
frais_de_port_moy	le frais de port en moyenne
delai_livraison_max	(en jours)
volume_produit_moy	volume maximum (en cm3)
densite_produit_max	la dénsité du produit (en gr/cm3)
distance	la distance entre le client et le vendeur (en km)

Categorie C.

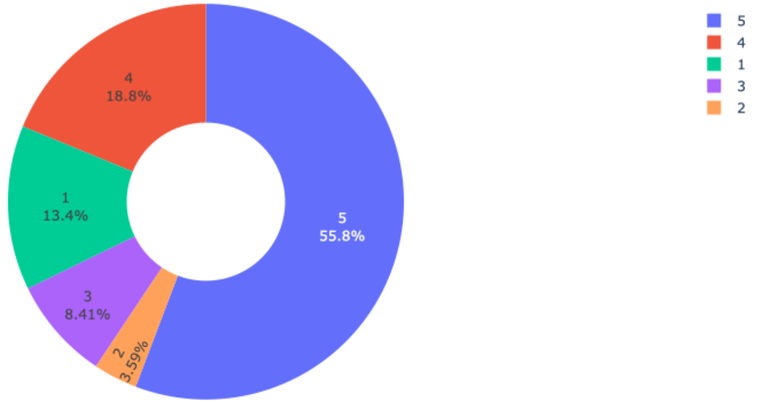
Traitement des variables categorielles

Feature	Explication
taux_type_payement	pourcentage de type de paiement
categorie_produit	8 catégories pricipales des produits
region_client	5 regions où les clients vivent

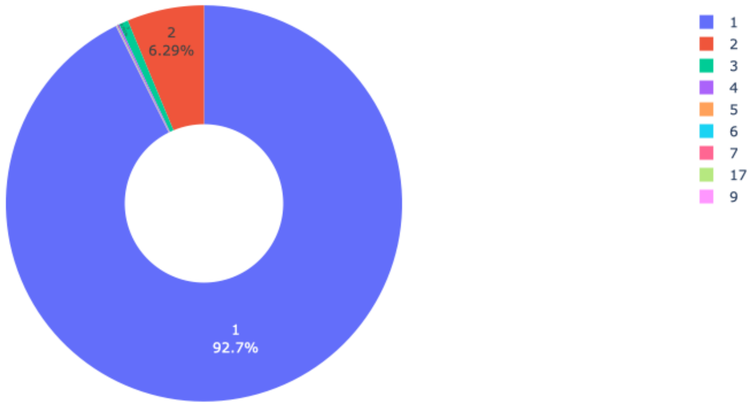
Population des clients par régions



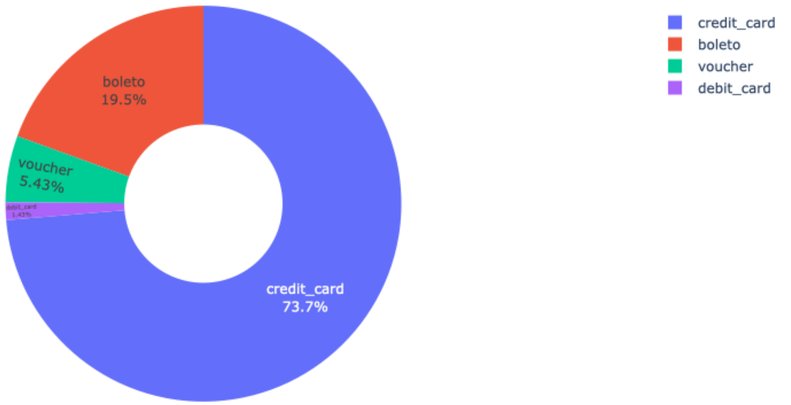
Review scores



Nombre de commandes



Type de payement



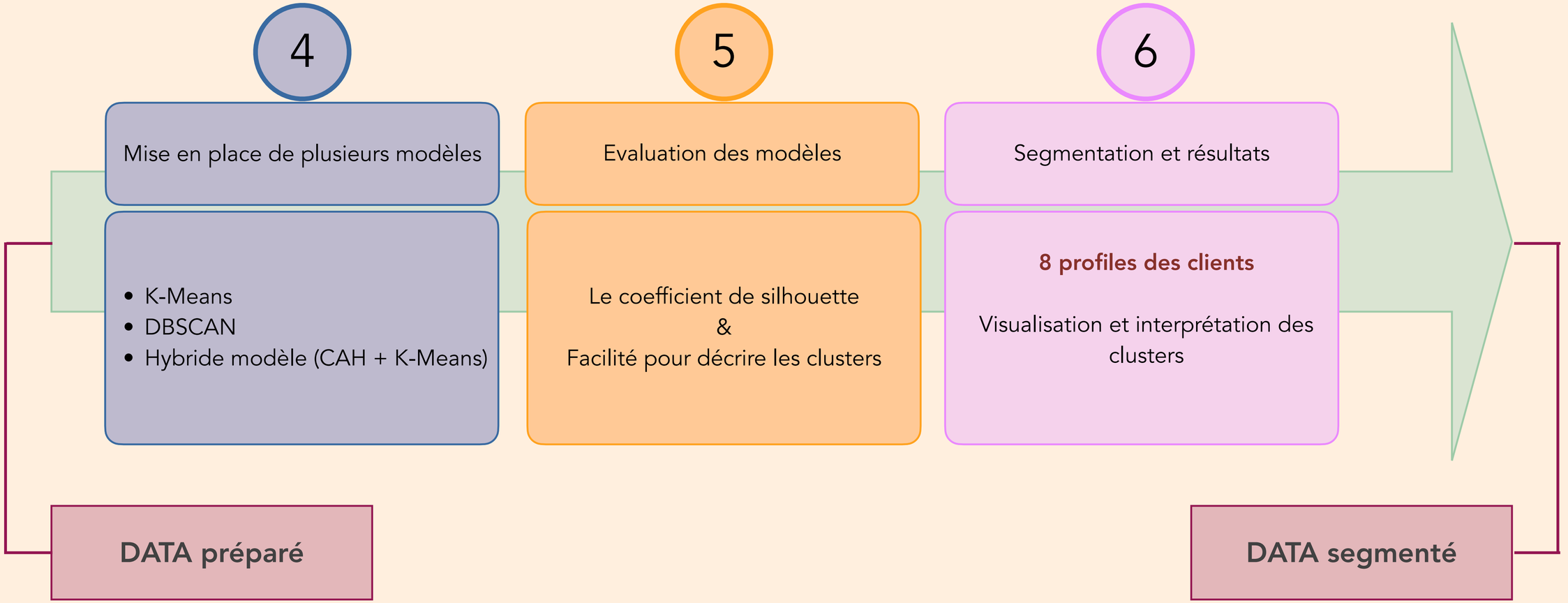
Modélisation

Non supervisée

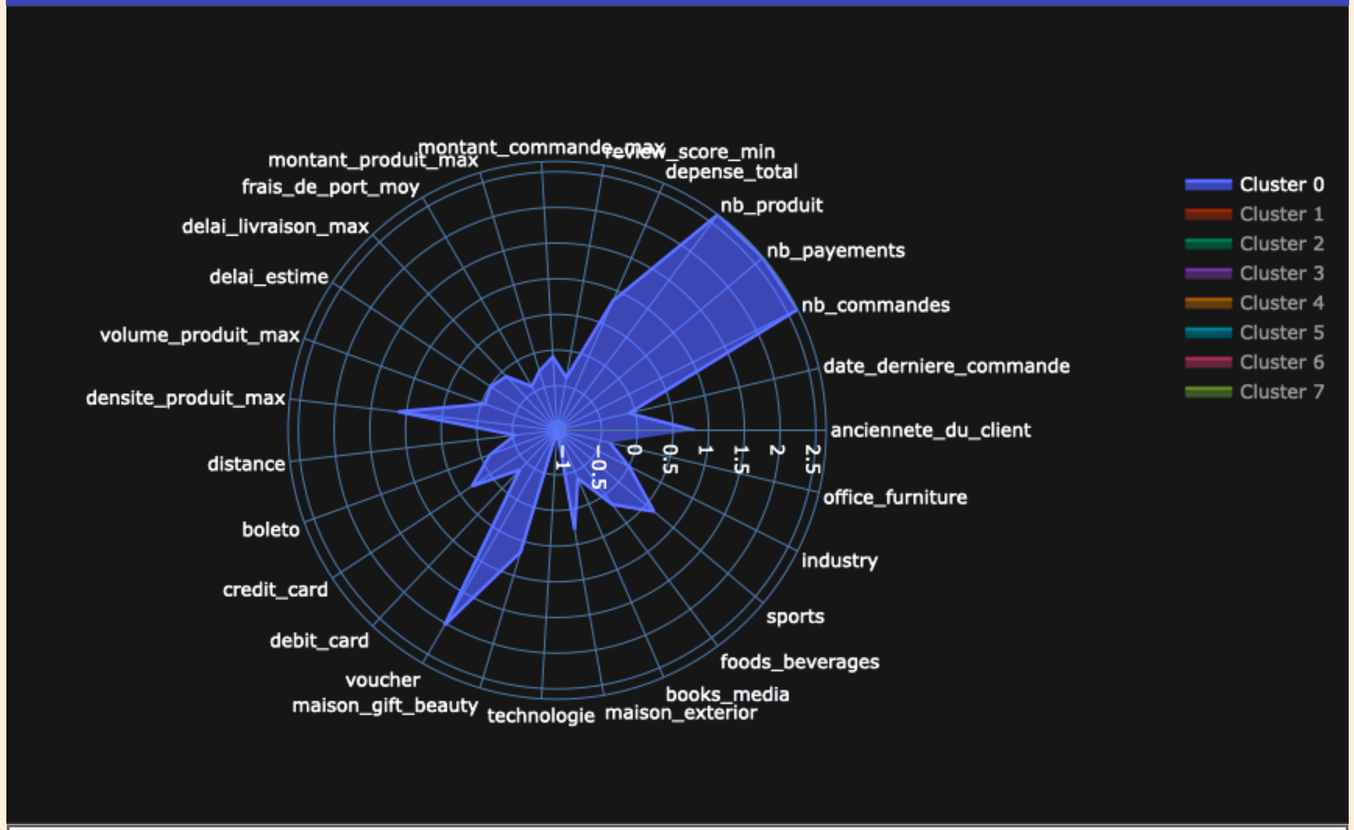
Clustering

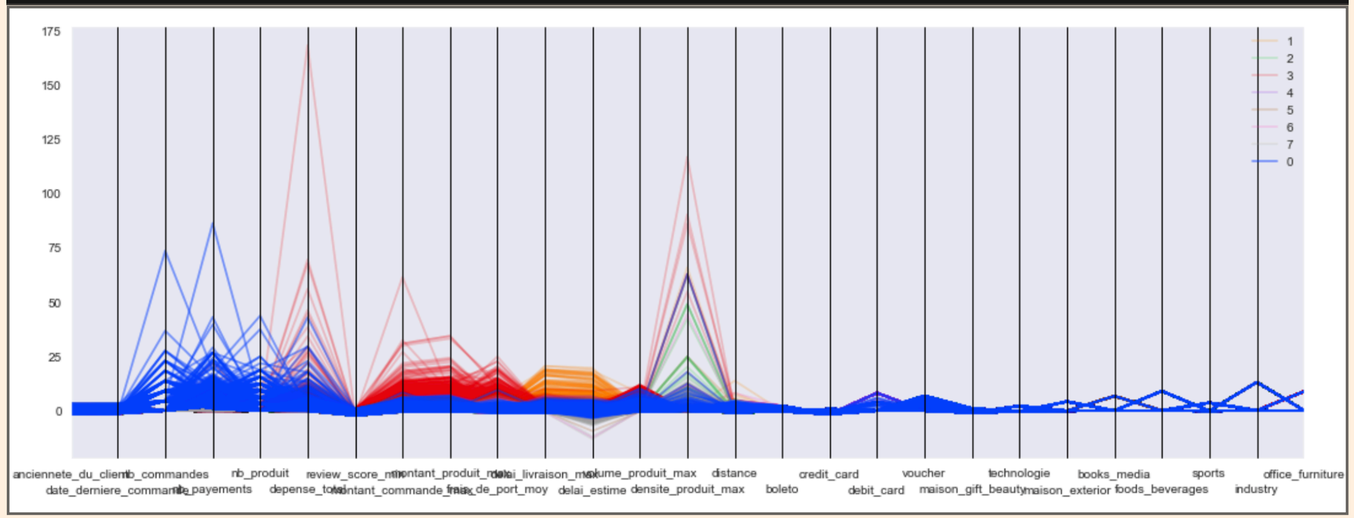
Modélisation

- Non Supervisée
- Clustering



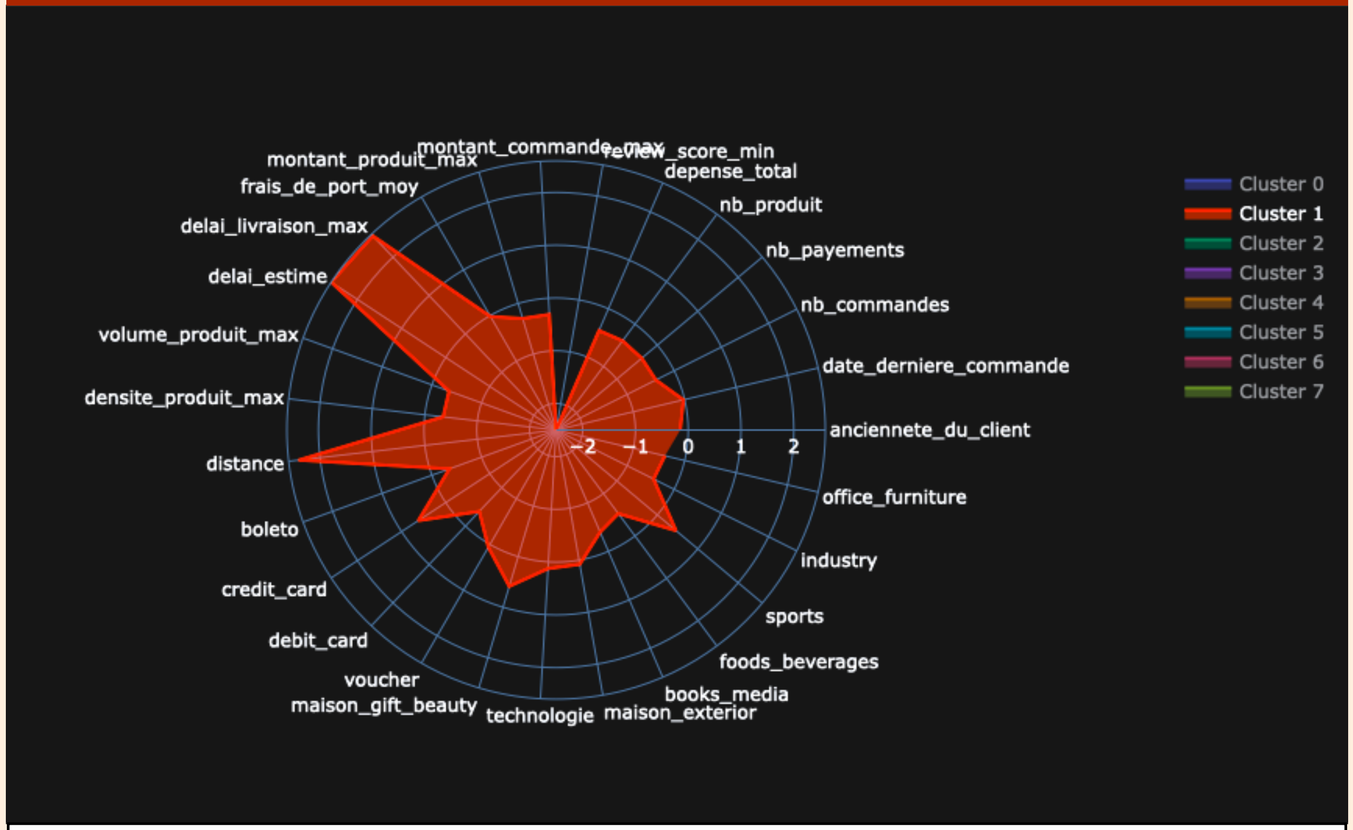
Cluster 0

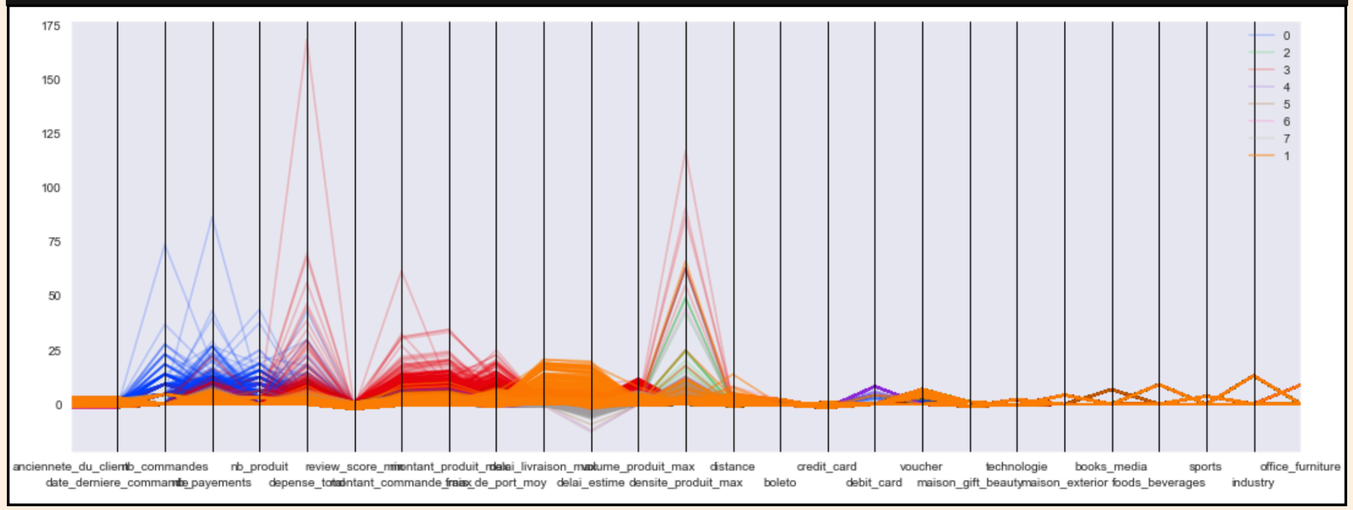




Paiements par voucher
Nombreuses commandes
Nombreux produits achetés
Paiements par tranche

Cluster 1

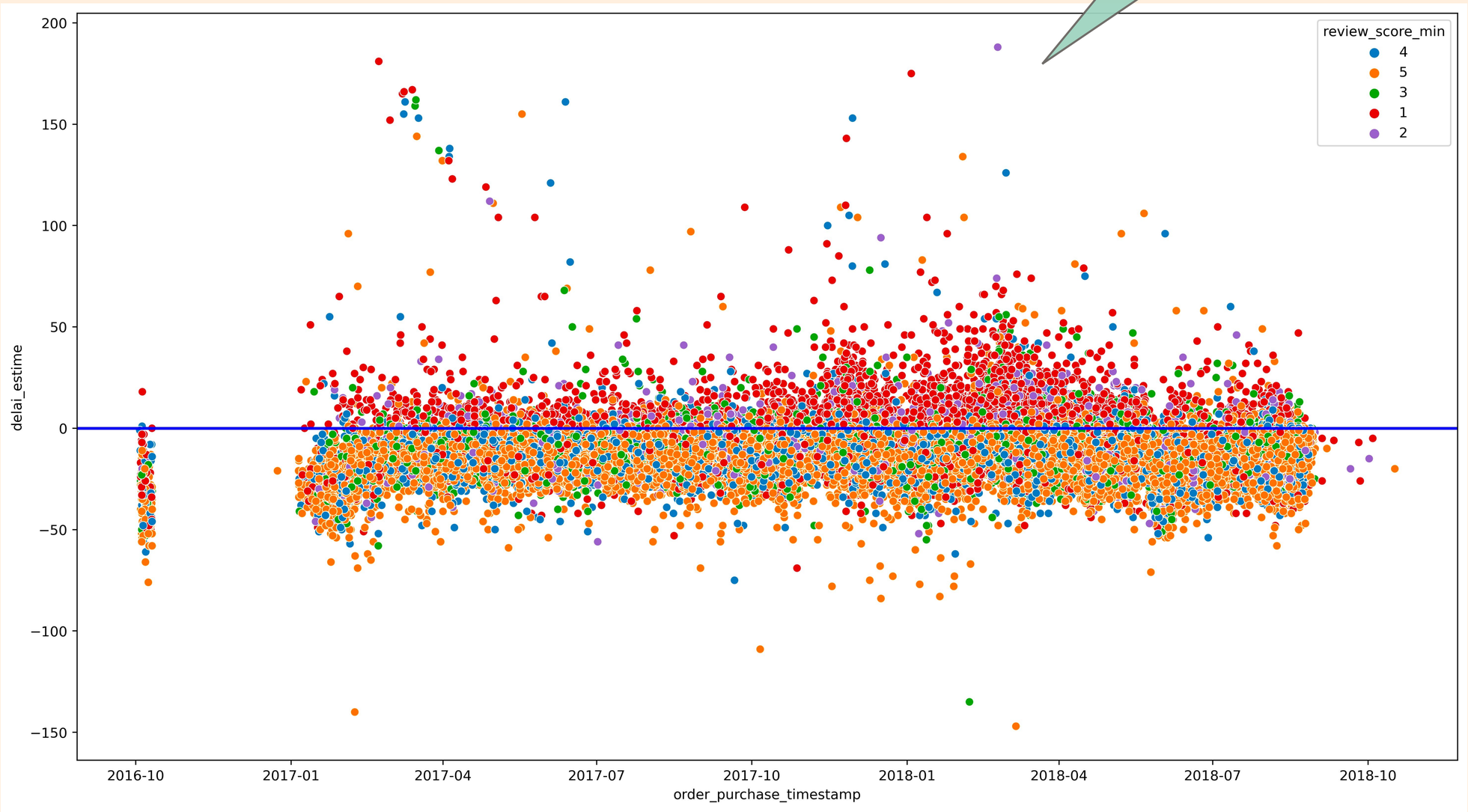




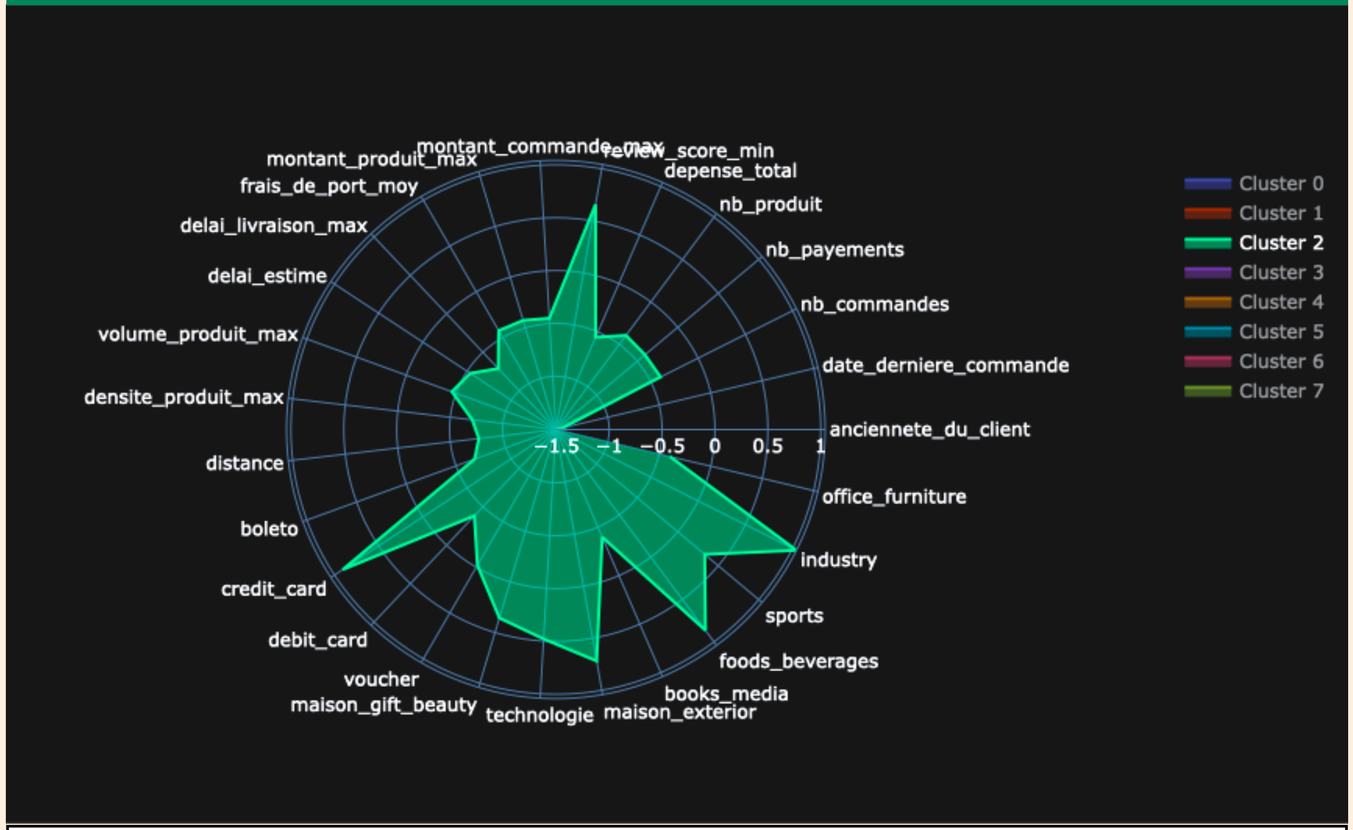
Les notes basses
Livraisons en retard
Délais de livraison très longs
Achats non locaux
Sports

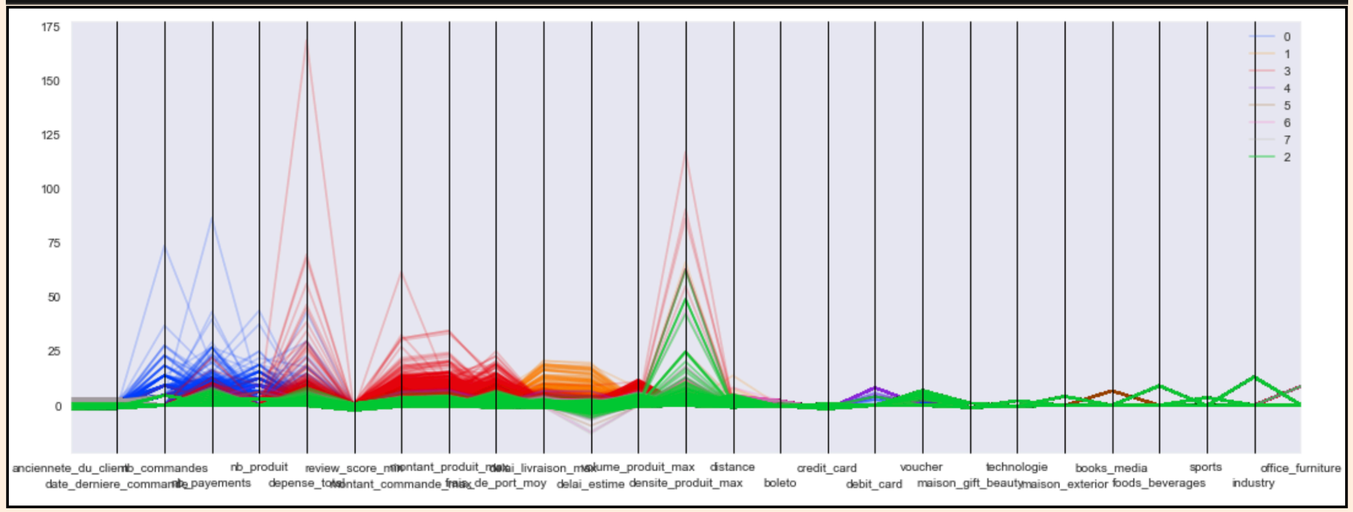
Cluster 1

colis arrivés dans les temps = les notes élevés
colis arrivés en retard = les notes faibles



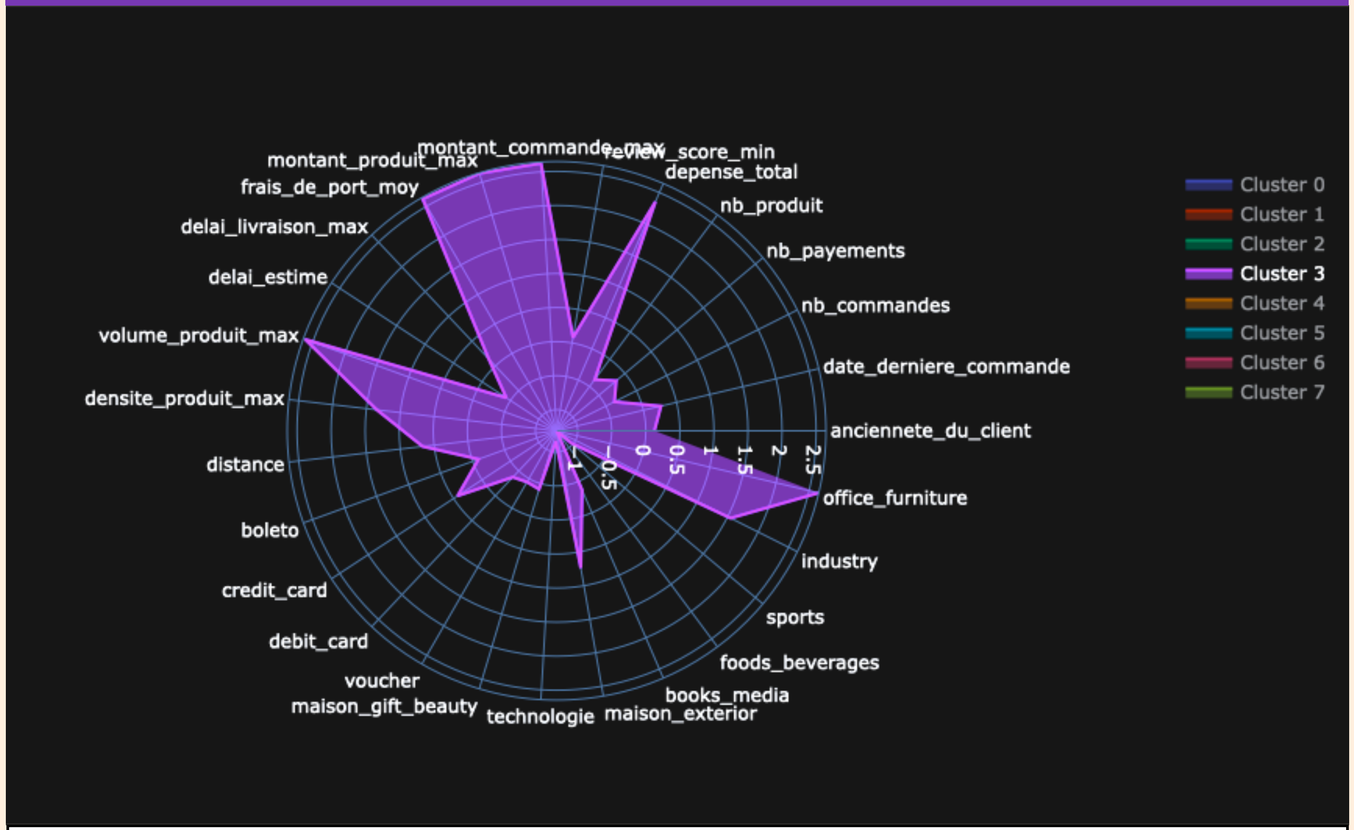
Cluster 2

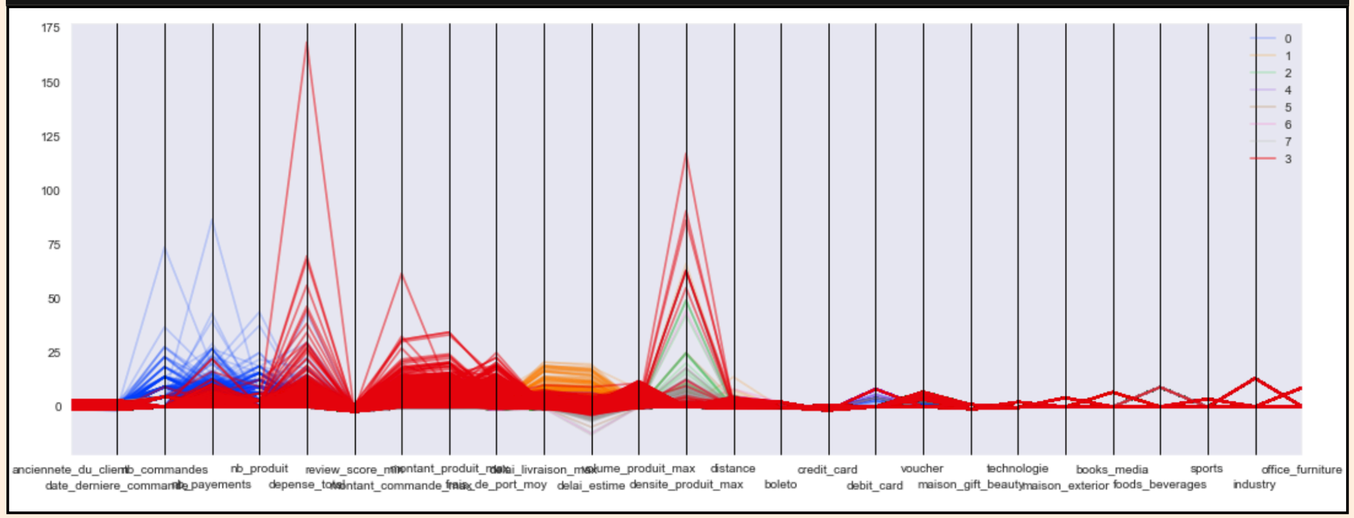




Clients anciens
Carte de credit
Achats divers
Les notes élevés

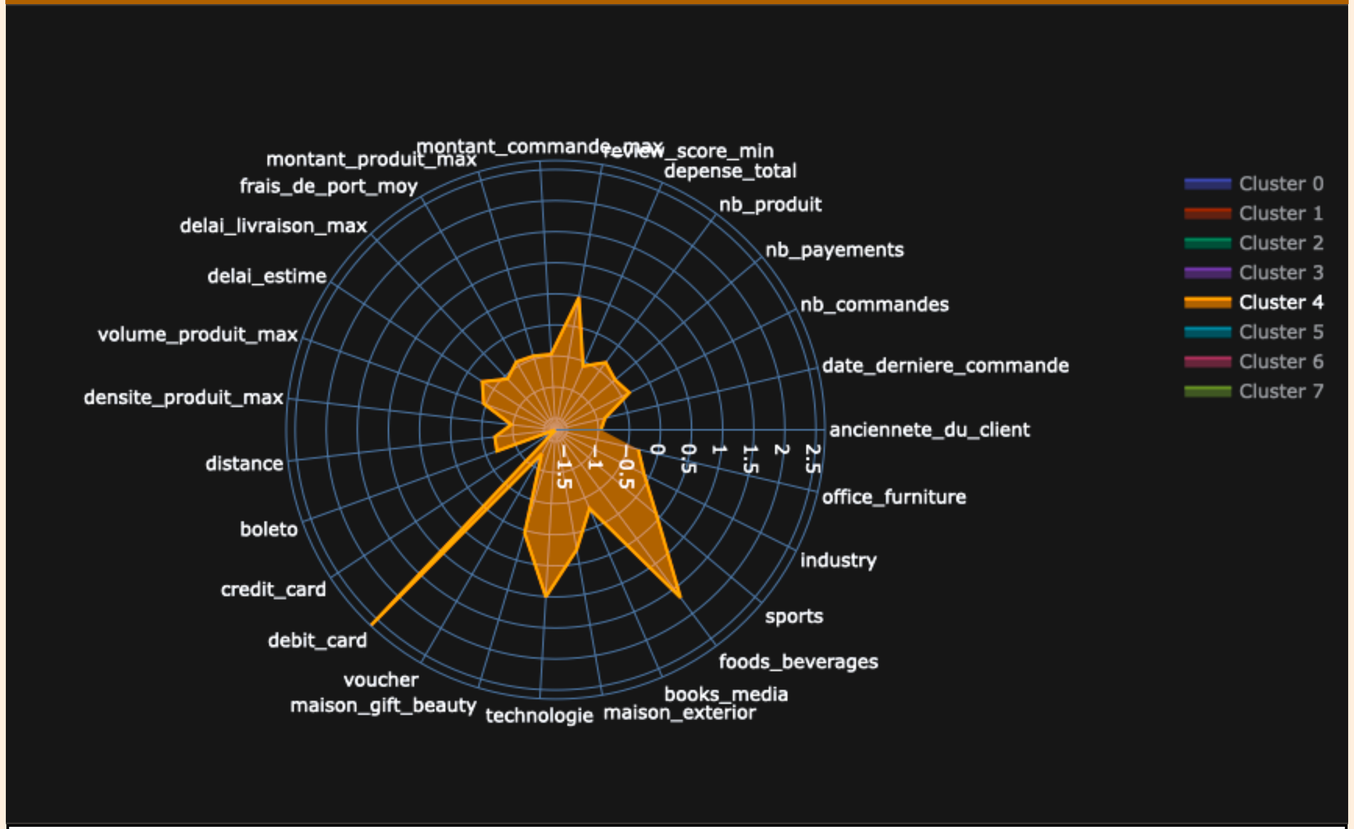
Cluster 3

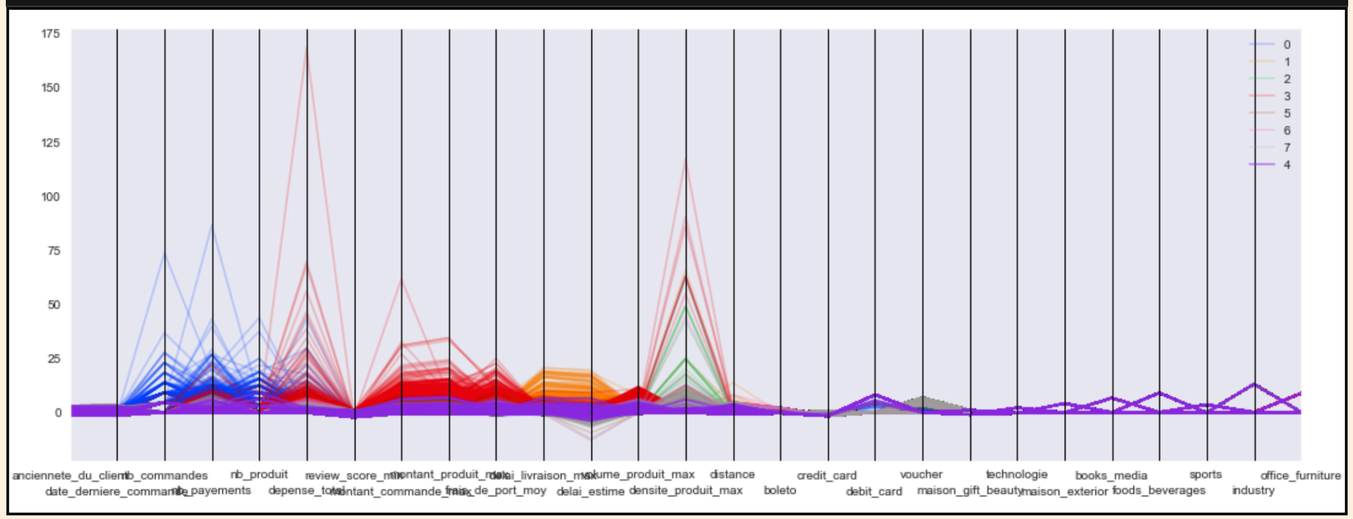




Produits volumineux
Office/meuble
Industrie
Frais de port élevés
Montant très élevé

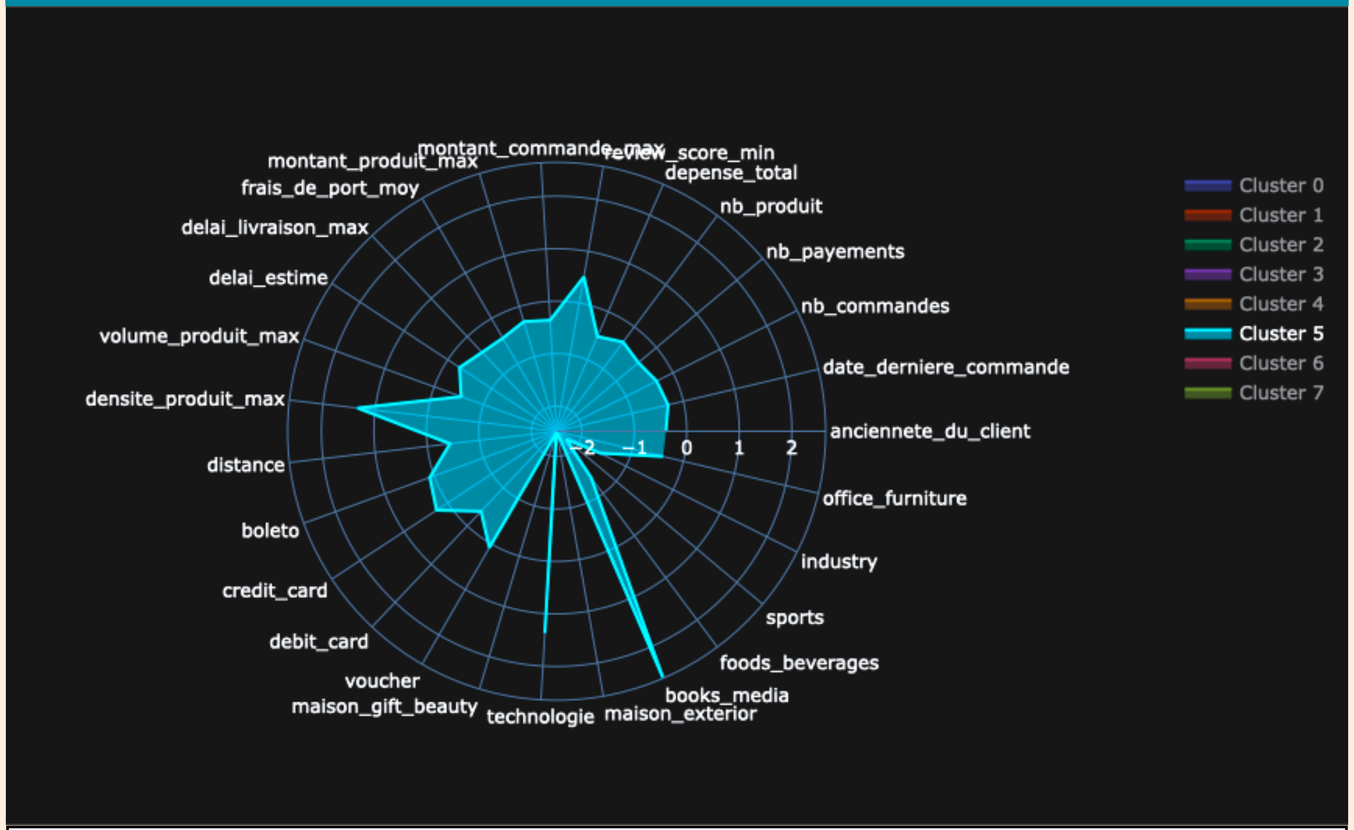
Cluster 4

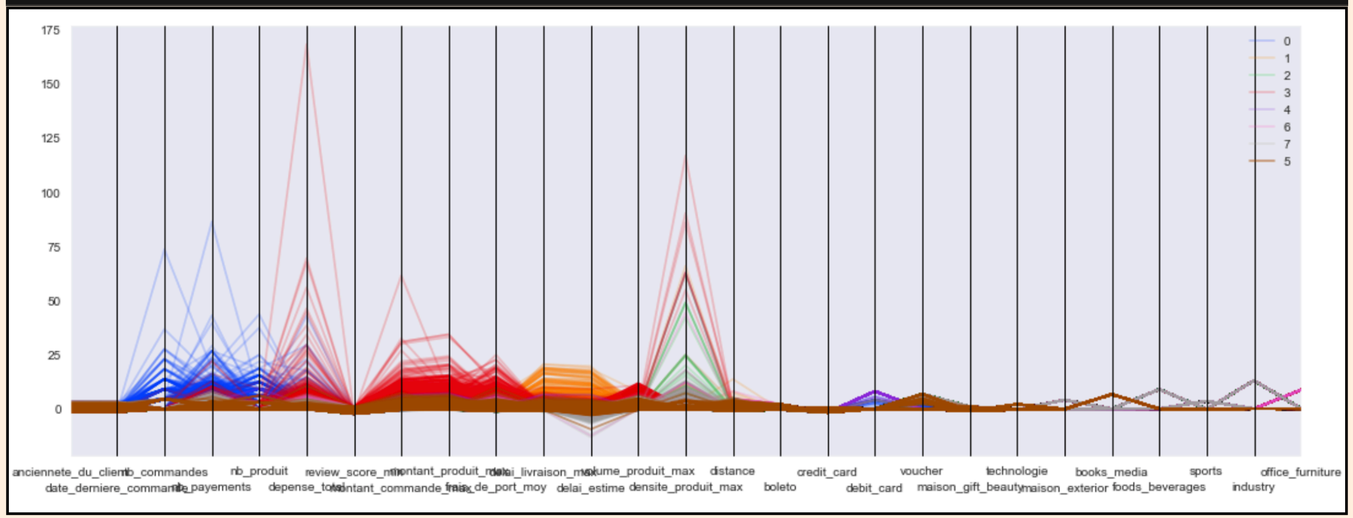




Carte de débit
Nourriture / Boissons
Les produit légers

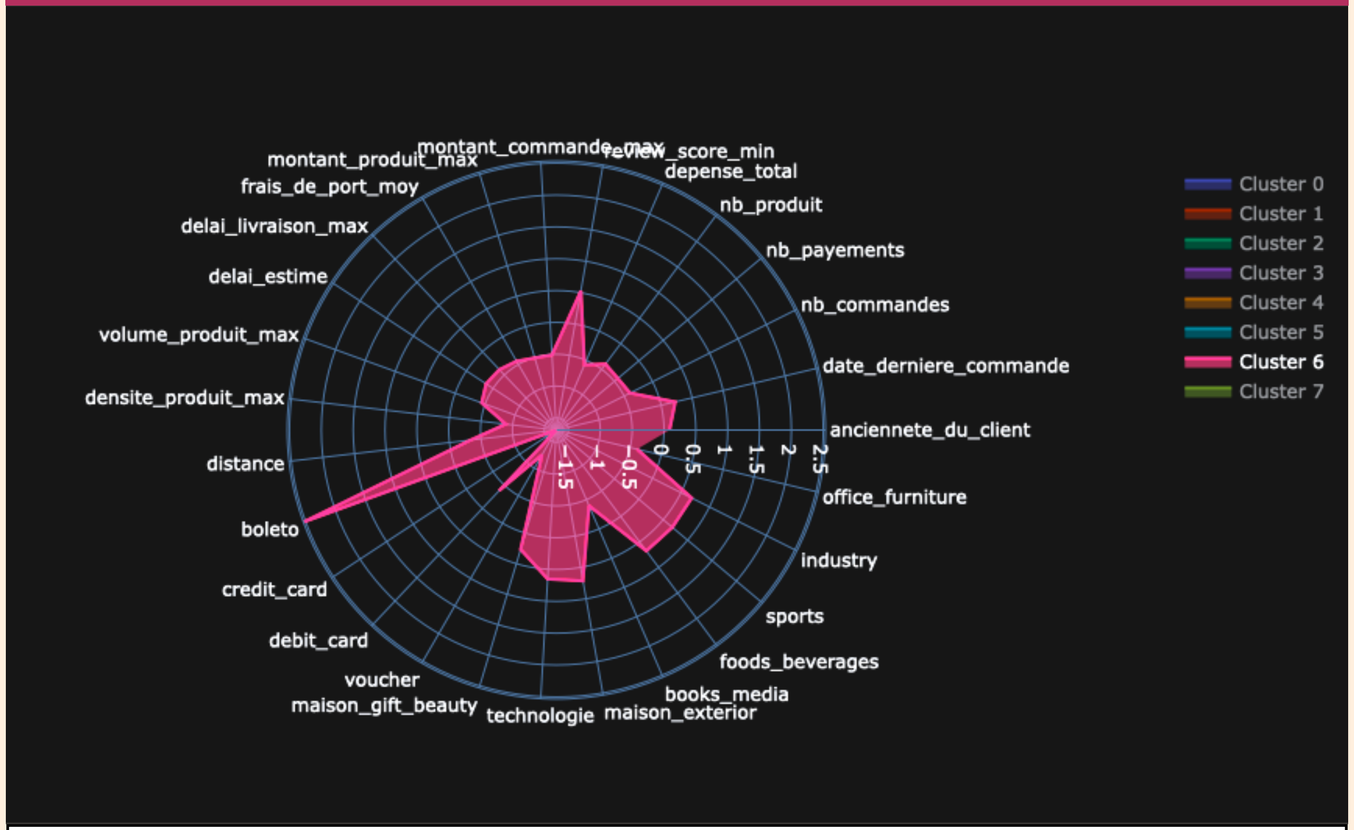
Cluster 5

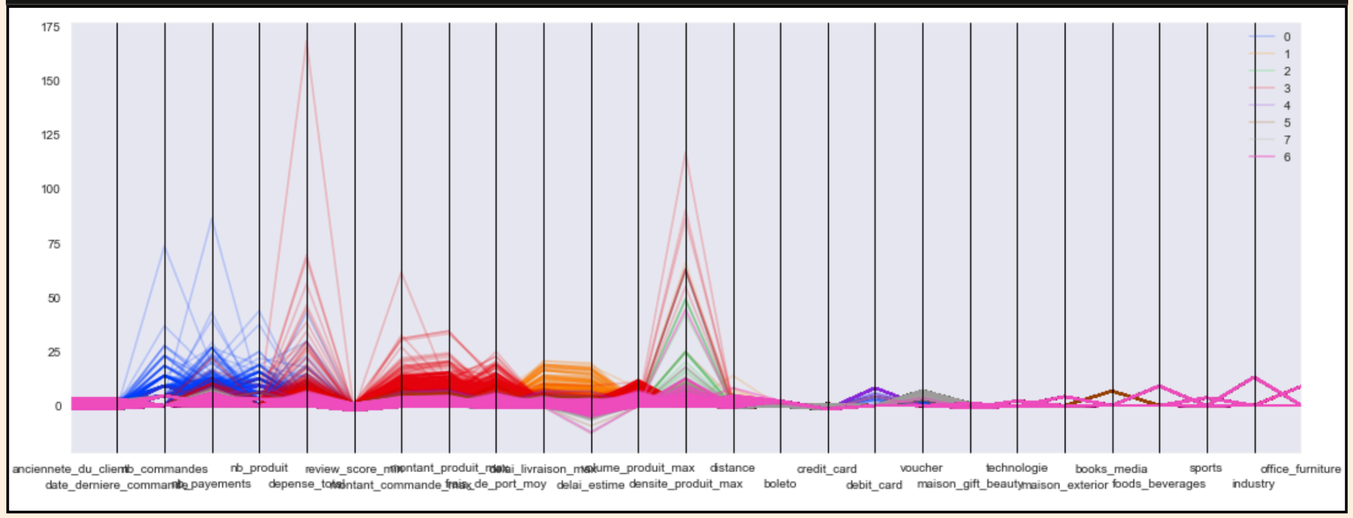




Technologie
Livres et media
Les notes élevé

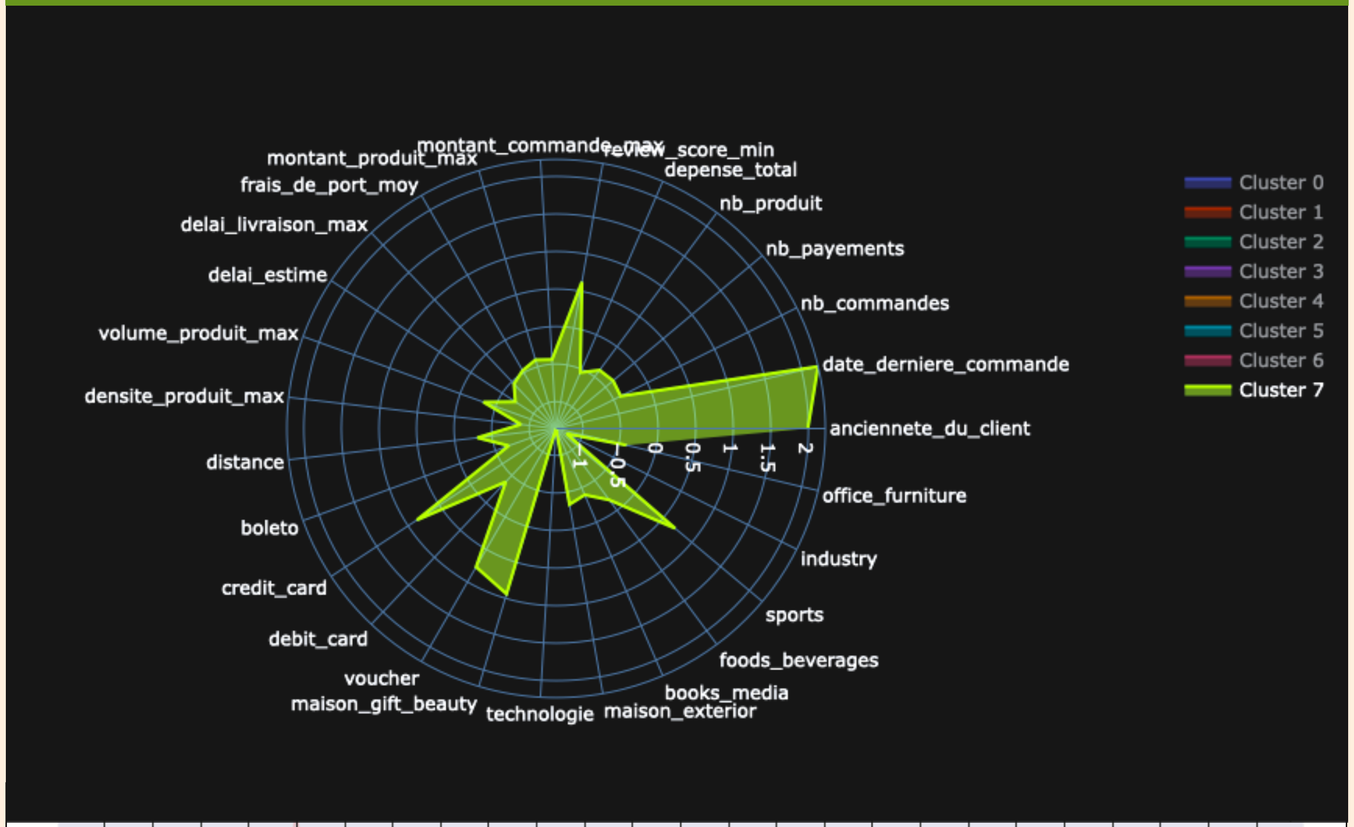
Cluster 6

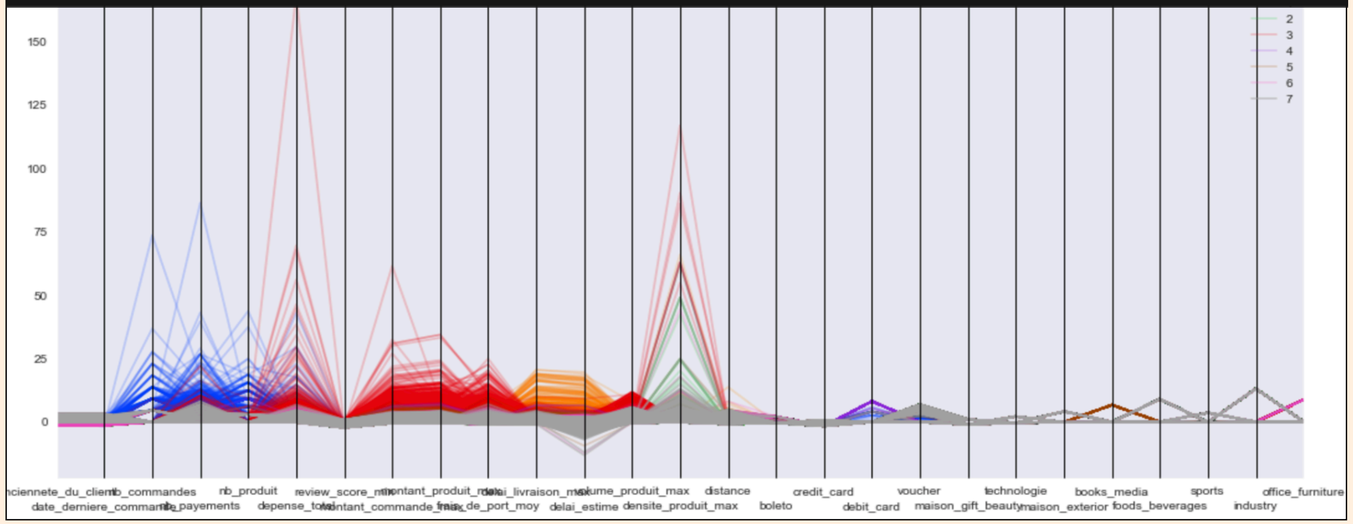




Boleto (ticket)
Achats divers

Cluster 7

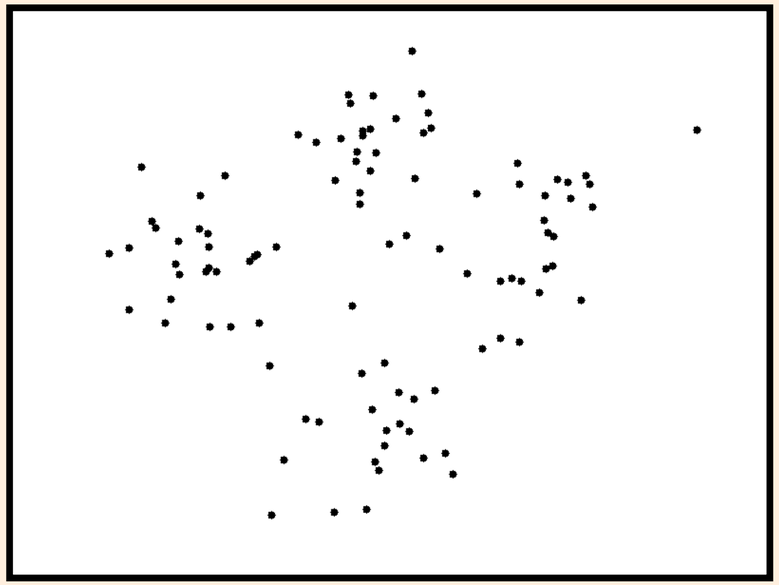




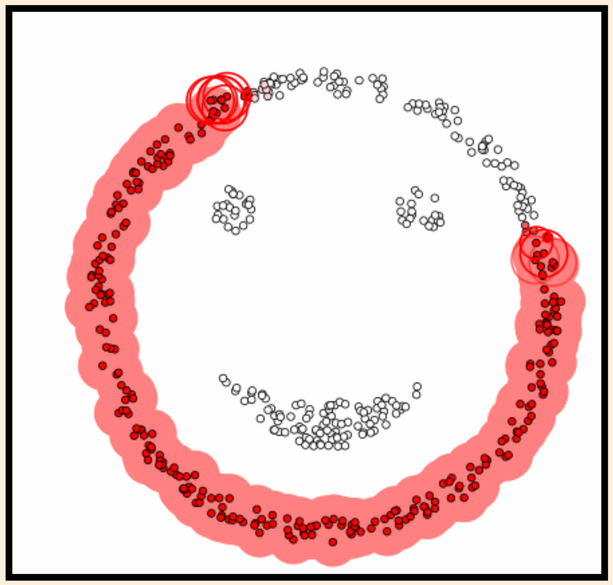
Achat récents
Cadeaux - Maison - Bébé - Beauté
Produits légers et petits
Carte de crédit

Etapes d’optimisation

K-Means



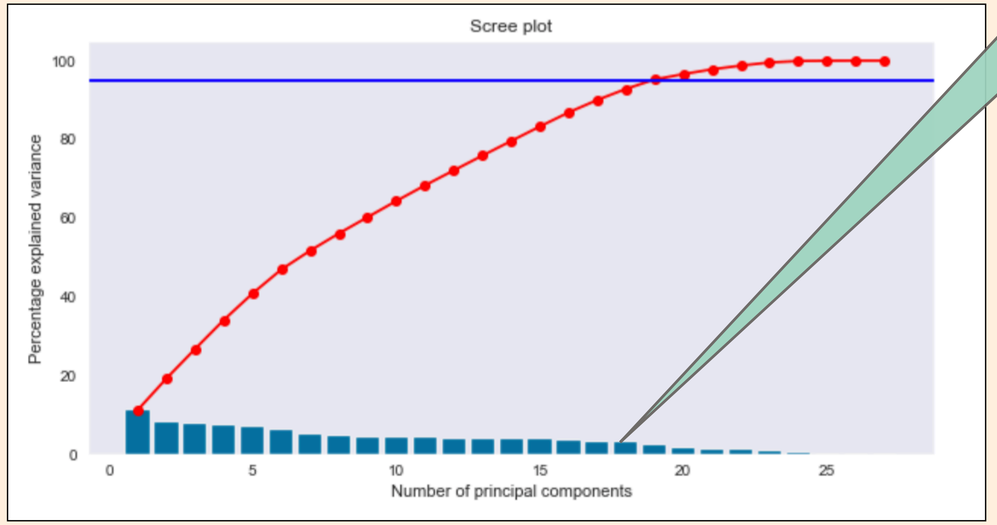
DBSCAN



Credits : https://github.com/learn-co-students/ds-k-means-clustering-readme-data-science

K-Means

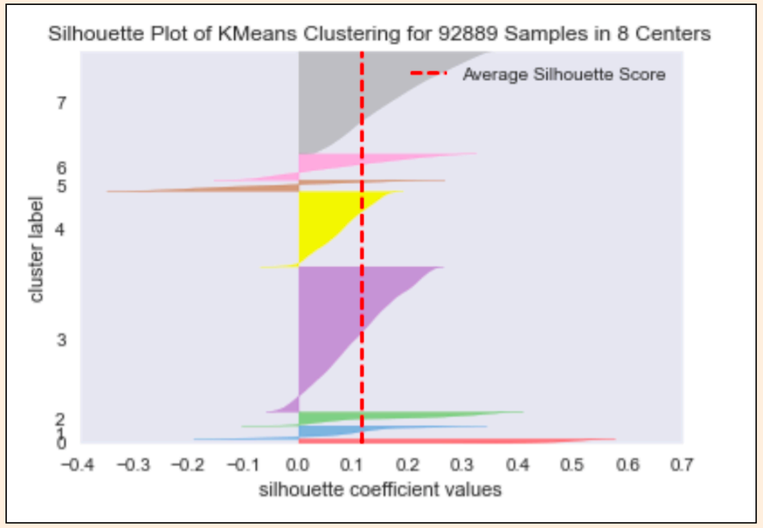
- L'analyse en composantes principales



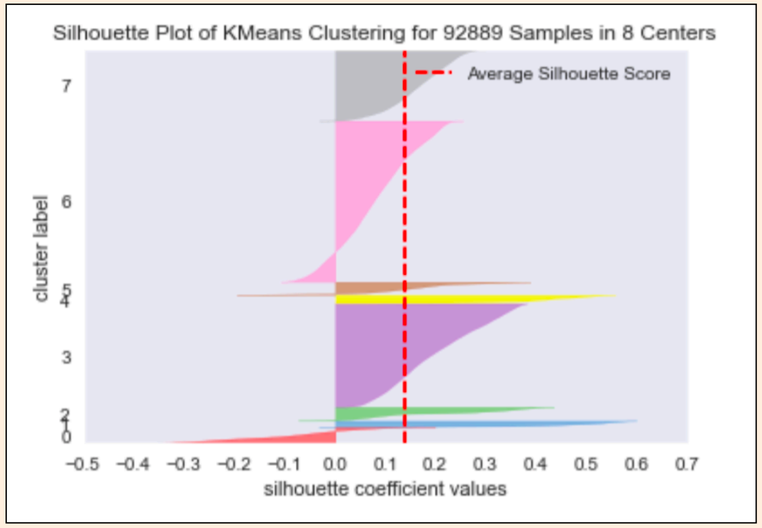
La plupart (95 %) des informations sont conservées dans 18 composantes
ACP

- **Hyperparamètres testé :**
  - n_cluster : 3 - 15
  - init='random', 'k-means++'
  - n_init = 10, 20, 30
- Avec ou sans les catégories de produits

- Le coefficient de silhouette



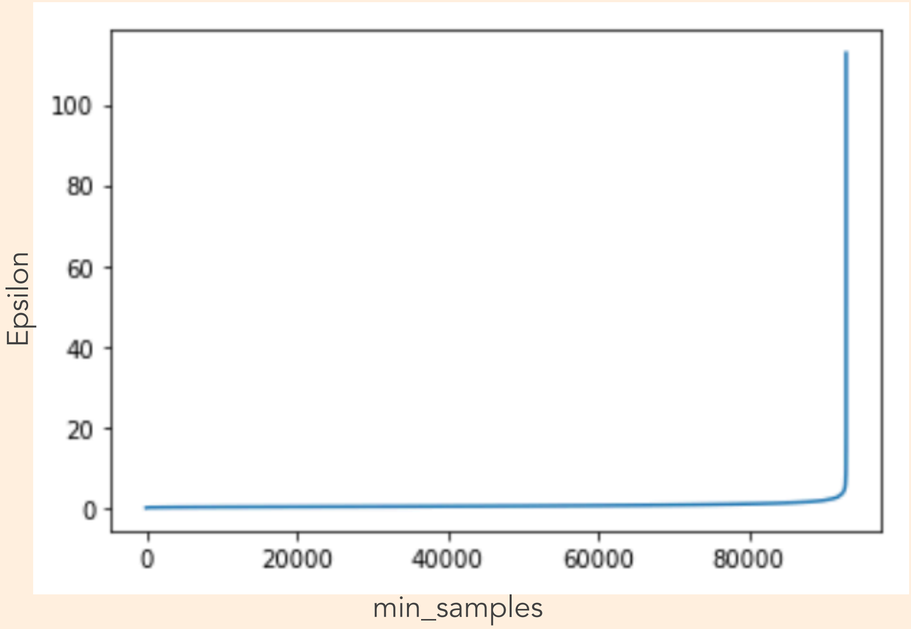
ACP

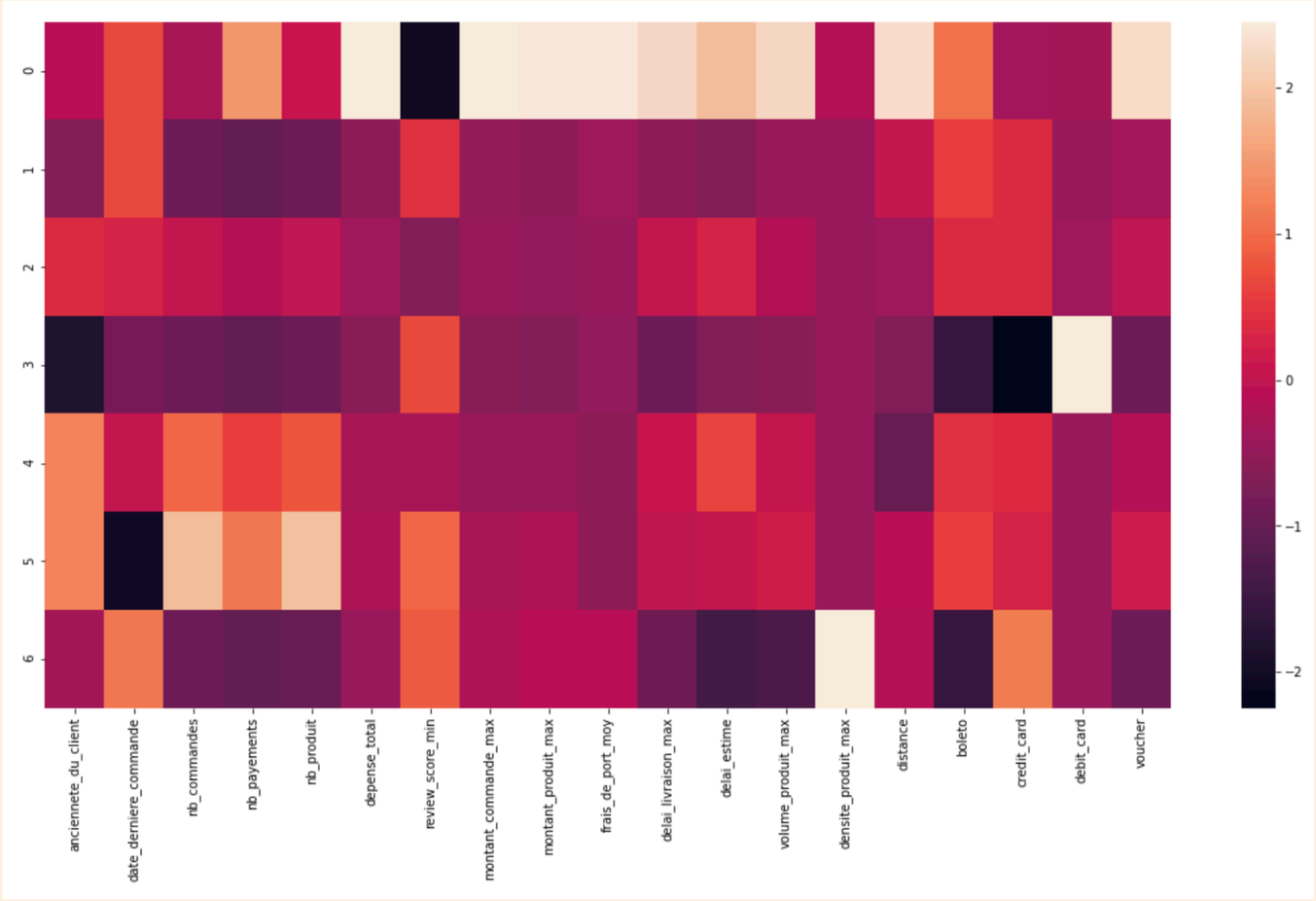


DBSCAN

- Hyperparamètres

- **Hyperparamètres testé :**
  - eps : 1.2 - 4.5
  - min_samples : 2 - 10

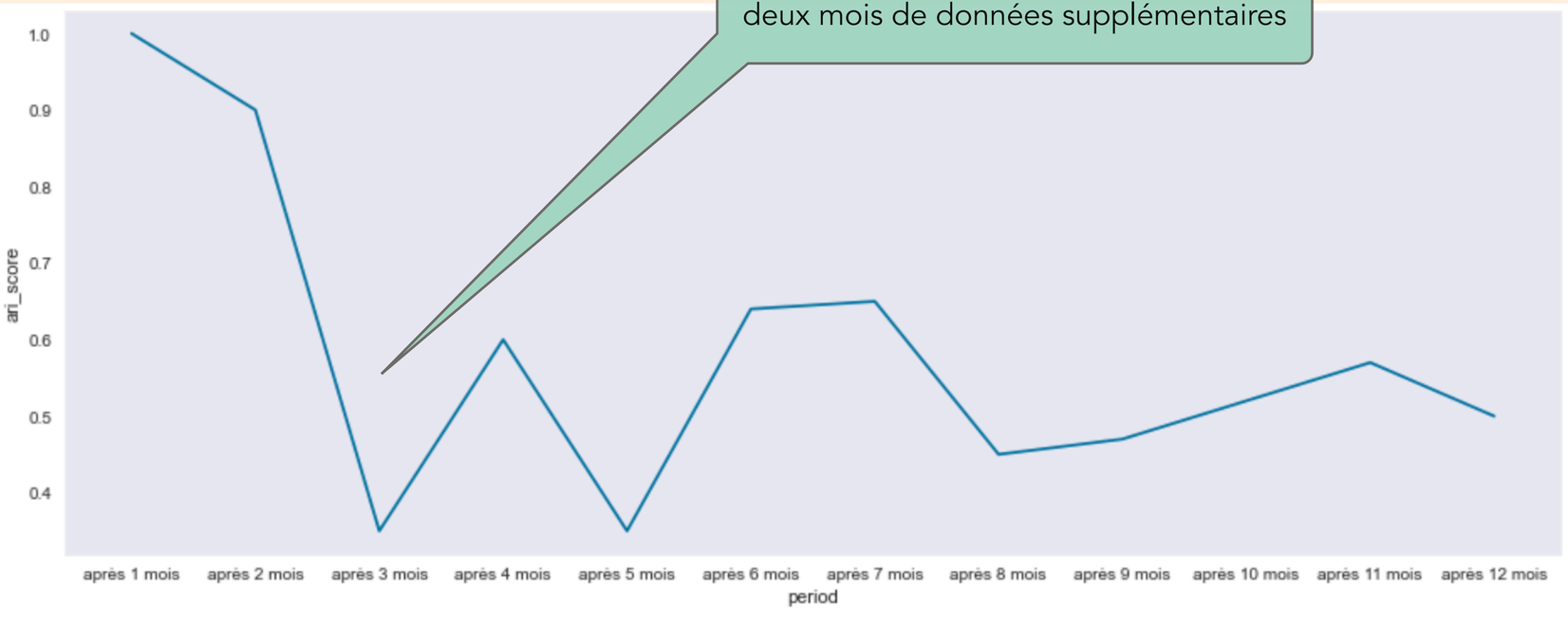




Maintenance

Stabilité de la segmentation

- Evolution de l'indice Rand ajusté **(ARI)**



Un intervalle de temps de 2 mois pour la maintenance de la segmentation

Conclusion

- 8 clusters exploitables facilement
- Entraînement de 2 modèles de classification non supervisée K-Means et DBSCAN
- Le modèle choisi : **K-means**
  - K-Means : Détection des groupes de clients est intéressante, facile à interpréter
  - DBSCAN : un modèle très compliqué à régler

- Il est nécessaire d'effectuer la segmentation deux mois après sa mise en production
- Les notebooks ont été réalisés dans le respect des règles de **PEP8**

https://www.python.org/dev/peps/pep-0008/

Merci de votre attention