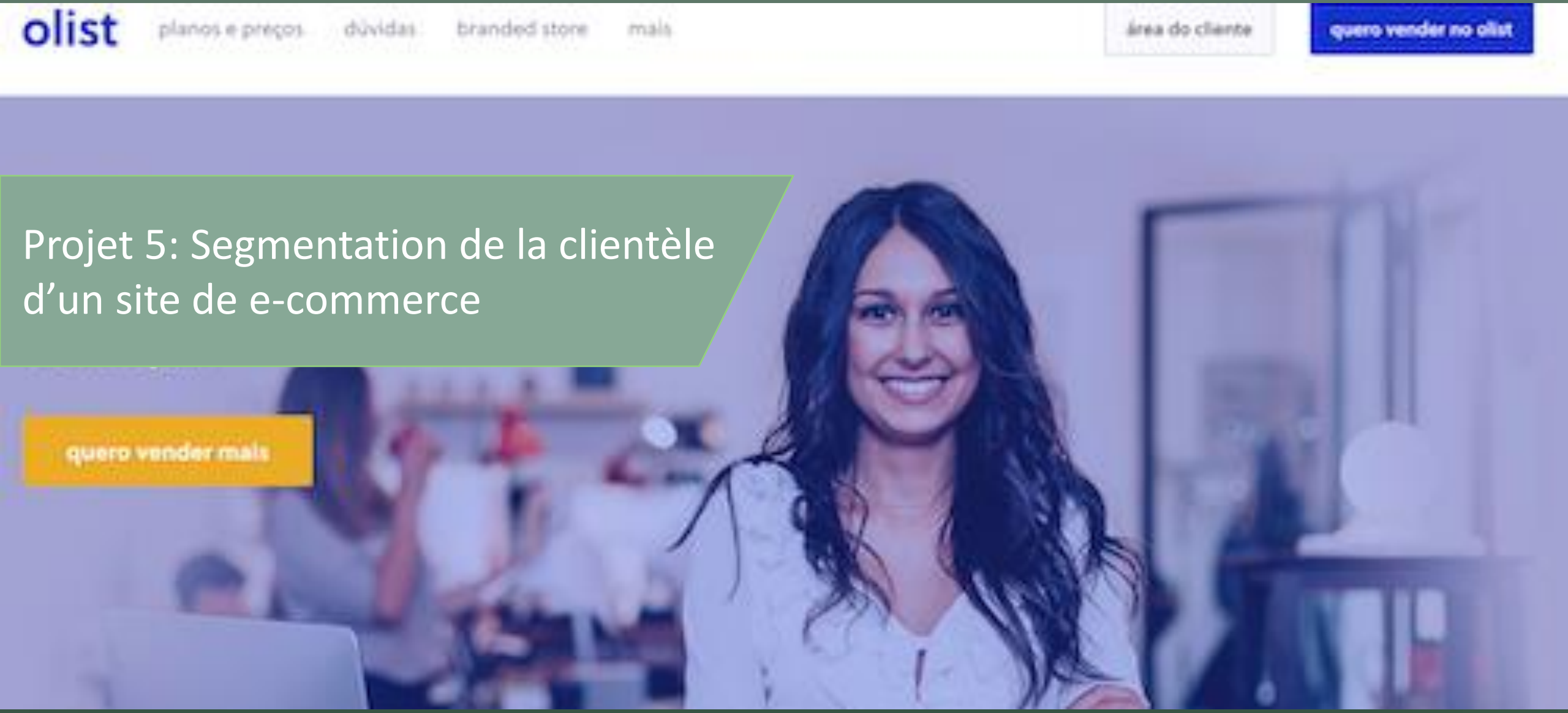
Projet 5: Segmentation de la clientèle d’un site de e-commerce



Missions

- Assembler un dataset avec des variables qui informent sur le comportement des clients
- Regrouper les clients par caractéristiques communes
- Simuler les prédictions du modèle pour évaluer sa pertinence dans le temps

Variables:

- RFM:
  - Récence
  - Fréquence
  - Montant

- Satisfaction: review_score
- Type de produits: product_weight_g

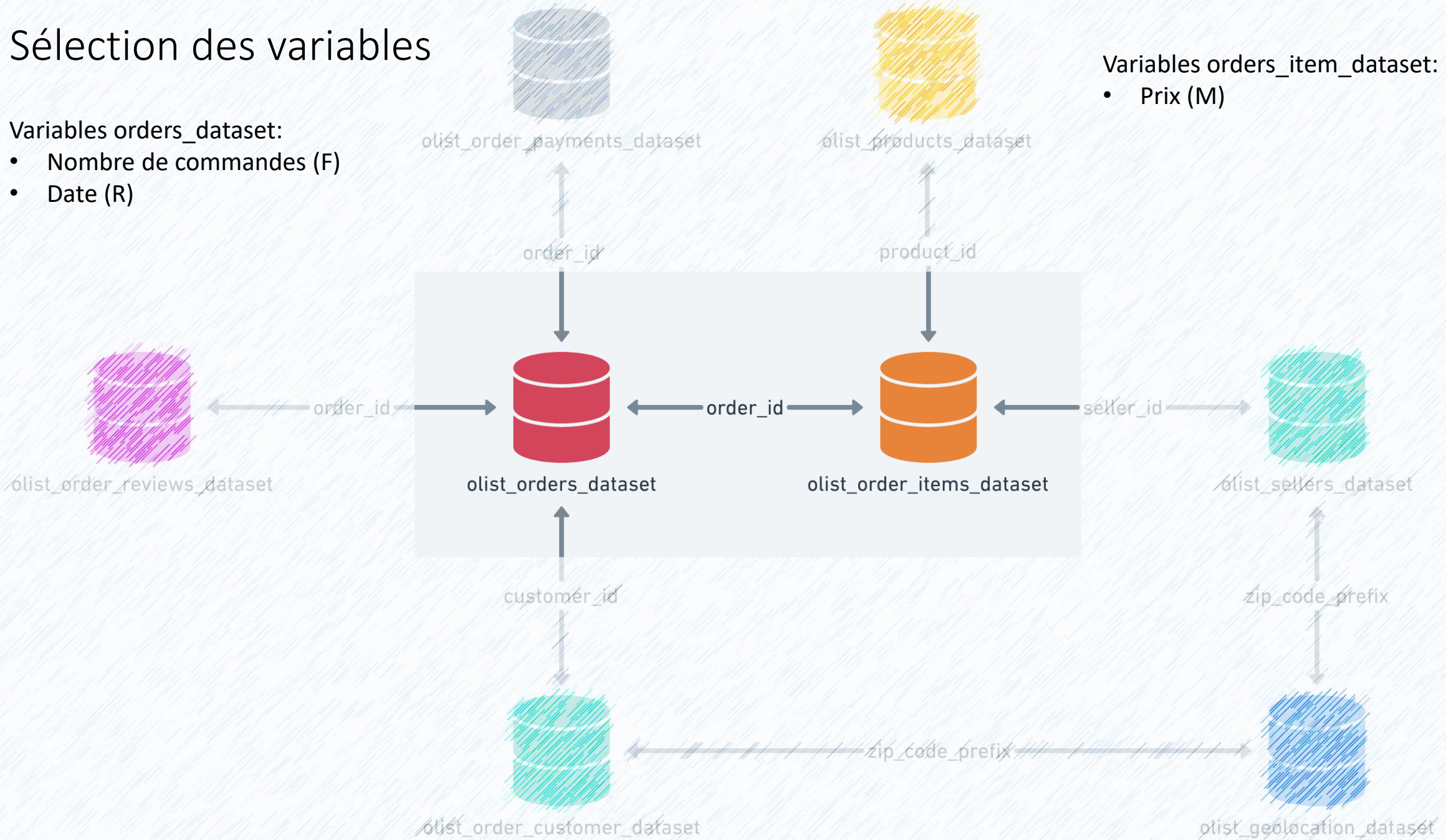
Sélection des variables

Variables orders_dataset:

- Nombre de commandes (F)
- Date (R)

Variables orders_item_dataset:

- Prix (M)



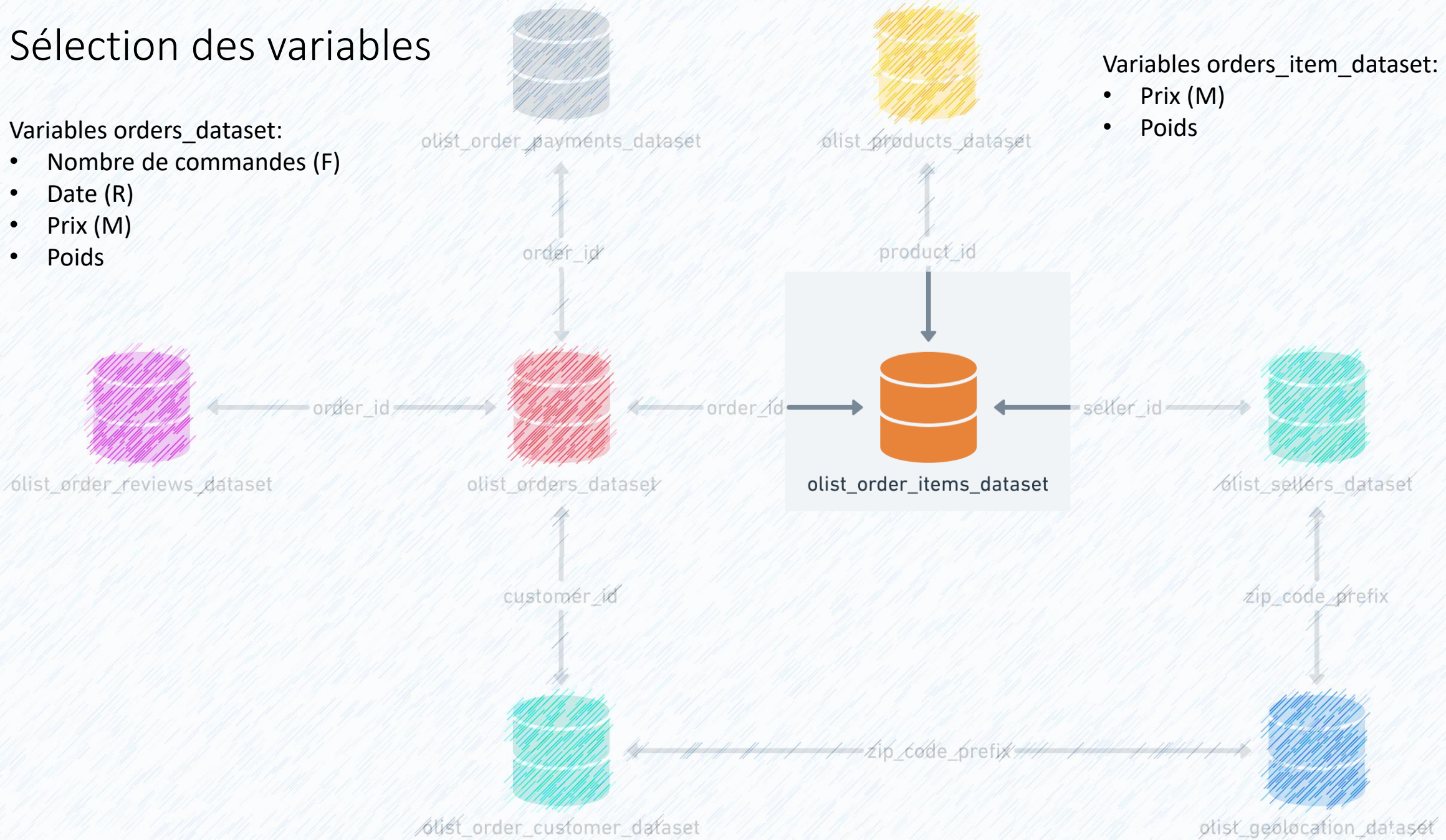
Sélection des variables

Variables orders_dataset:

- Nombre de commandes (F)
- Date (R)
- Prix (M)
- Poids

Variables orders_item_dataset:

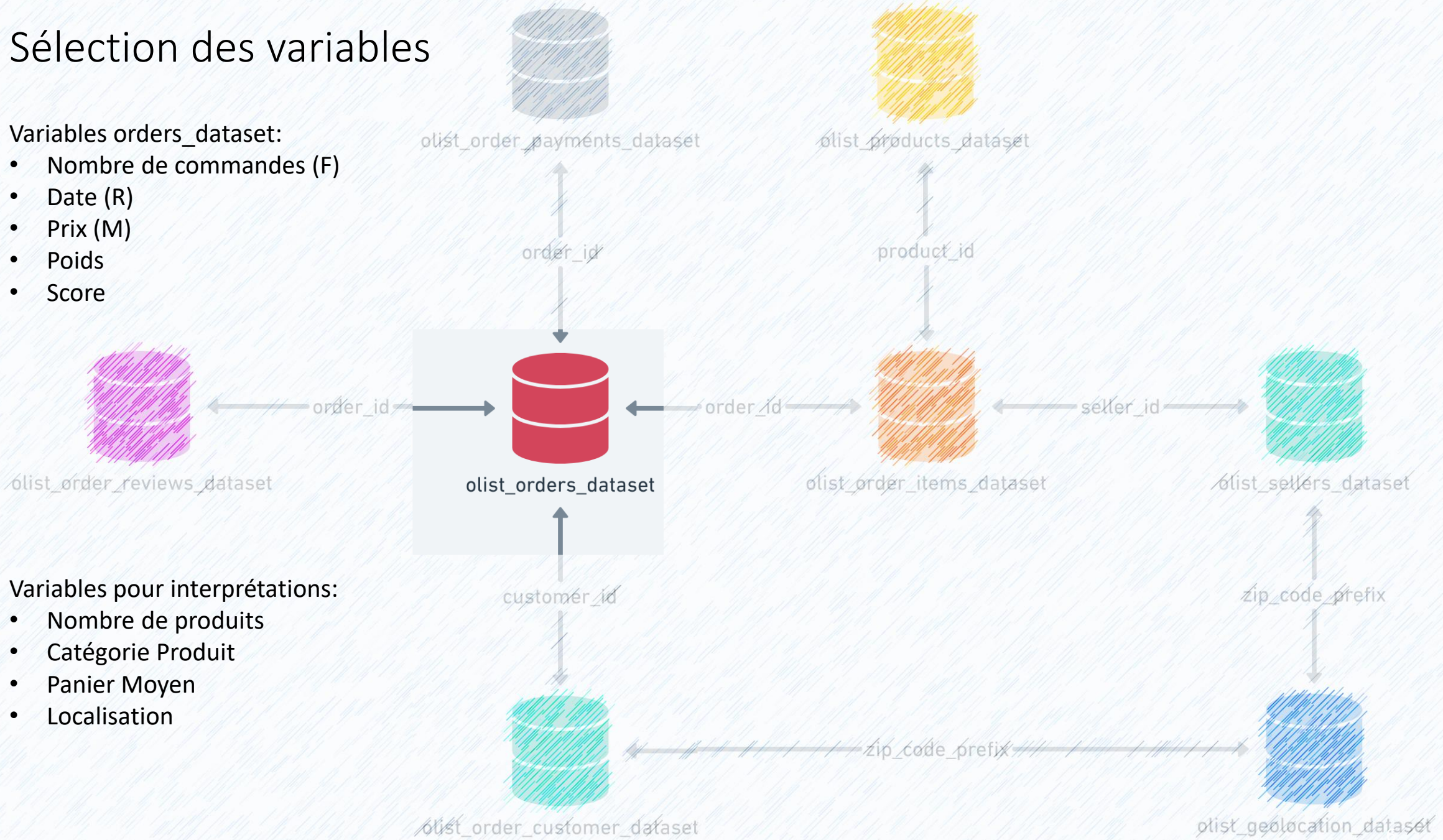
- Prix (M)
- Poids



Sélection des variables

Variables orders_dataset:

- Nombre de commandes (F)
- Date (R)
- Prix (M)
- Poids
- Score



Variables pour interprétations:

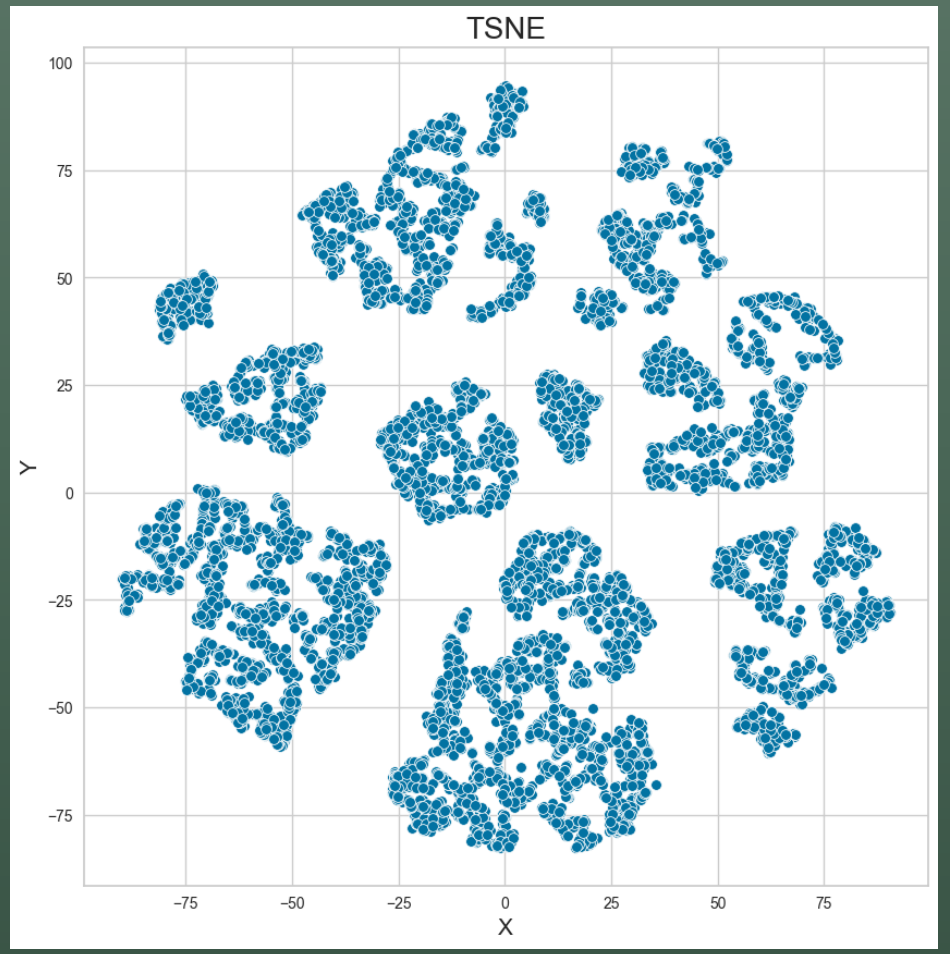
- Nombre de produits
- Catégorie Produit
- Panier Moyen
- Localisation

Nouveau dataset

- Sélection d’une période:
  - 1 janvier 2017 – 31 août 2018
- 95 762 Clients
- 3% des clients ont commandé plusieurs fois
- 12% des clients ont acheté plusieurs produits

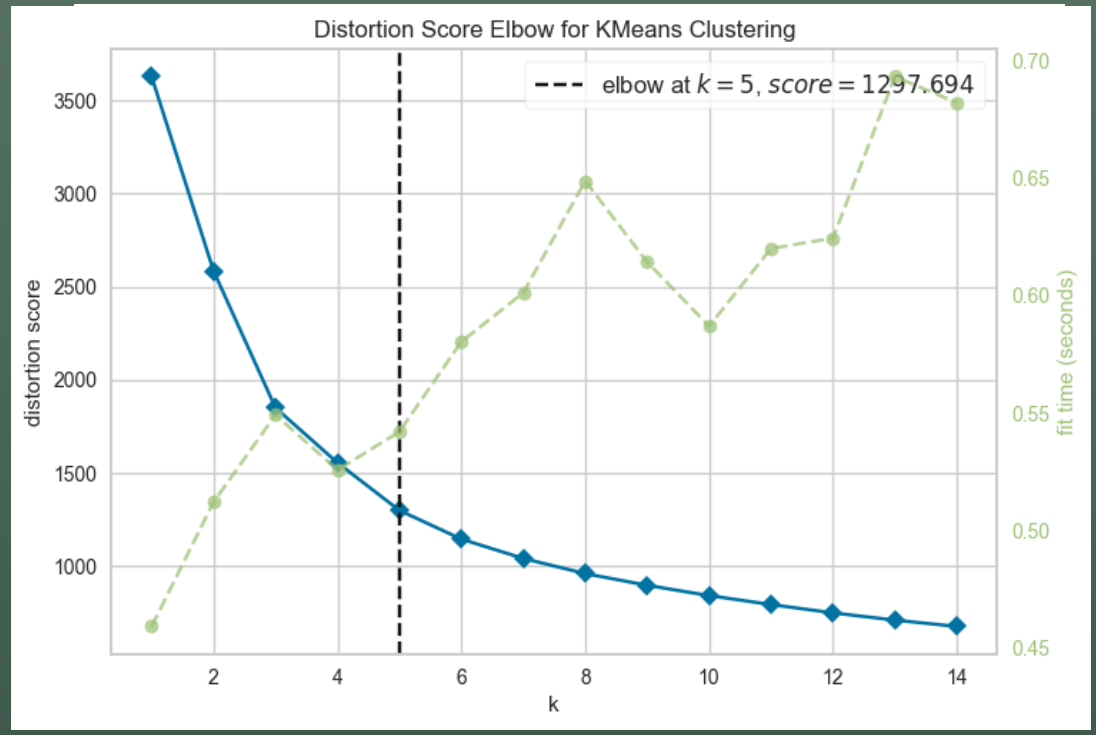
Algorithme non-supervisé

- Kmeans
  - Regroupement basé sur la proximité
- DBSCAN
  - Prend compte la densité
- Clustering Hiérarchique
  - Division ou Agglomération



Kmeans

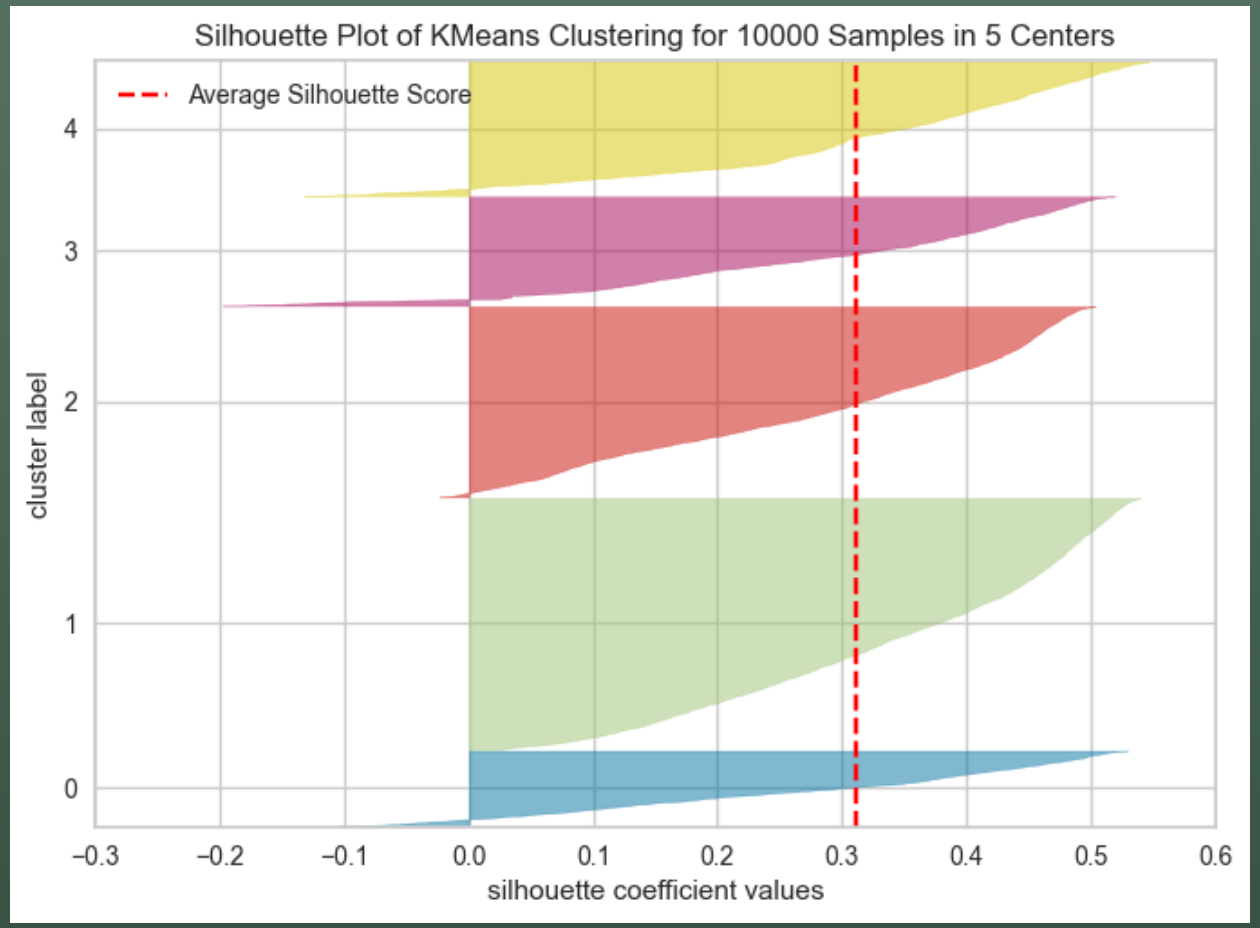
- Méthode du coude pour le choix de la valeur du paramètre k
- Sensible aux outliers



Le score silhouette pour 5 clusters

- Mesure:
  - Distinction entre cluster
  - Densité

- Score: 0,31



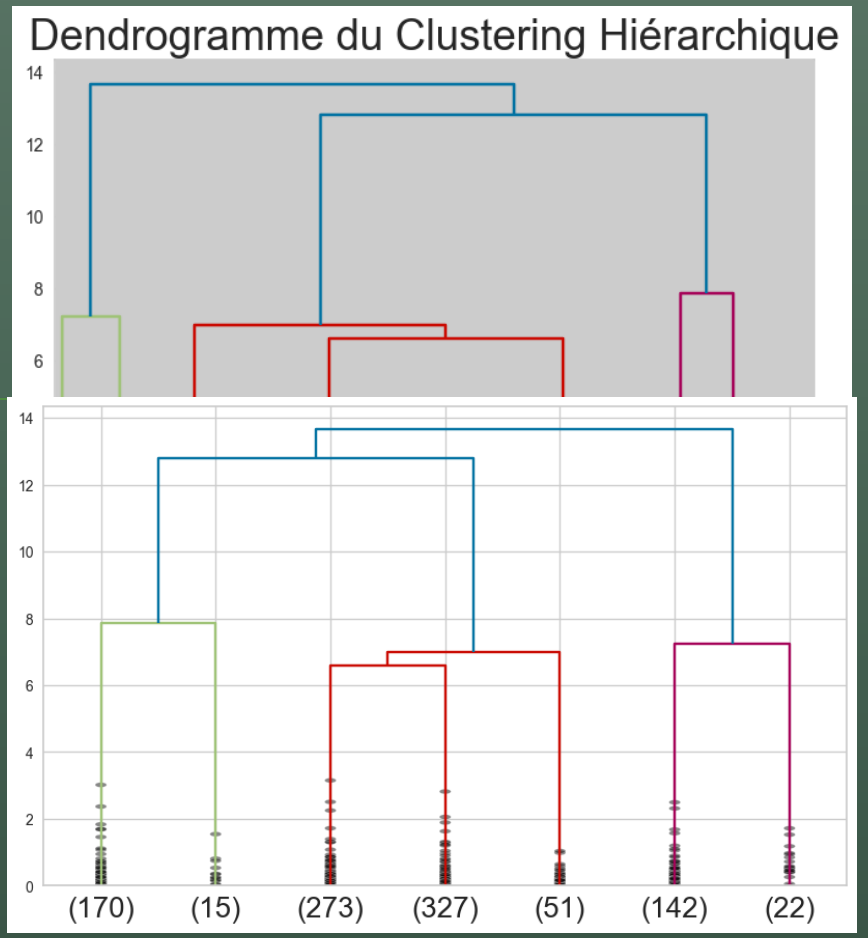
Clustering Hiérarchique

- Intuitif et flexible: Niveau de coupe
- Utile pour interpréter les relations

- Temps de calcul important

Pour un échantillon de 1000 :

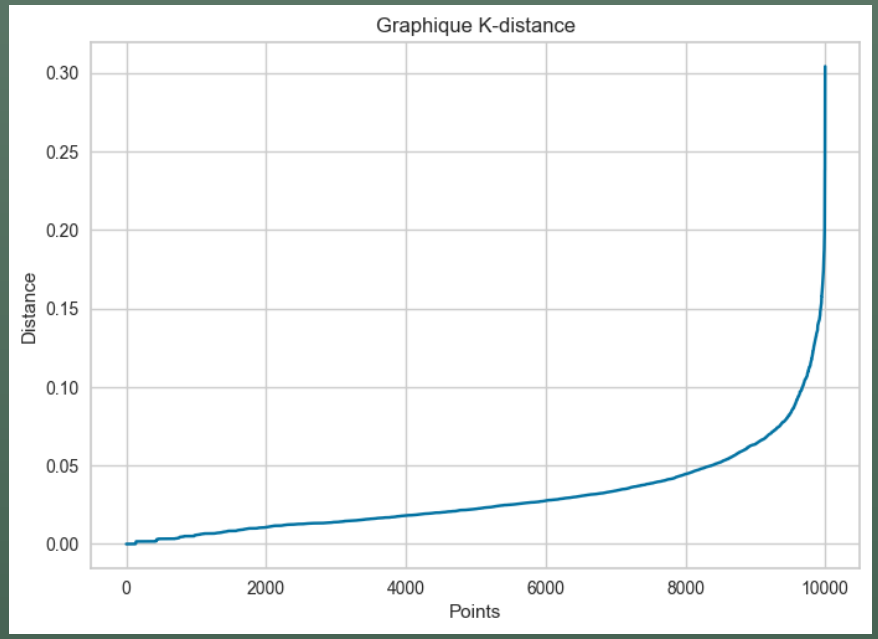
- Pour 7 clusters:
  - Score silhouette: 0,29



DBSCAN

- Choix des Paramétrages:
  - Distance esp
  - taille de l'échantillon mini
- Exclut le bruit

- Nécessite un paramétrage subtil



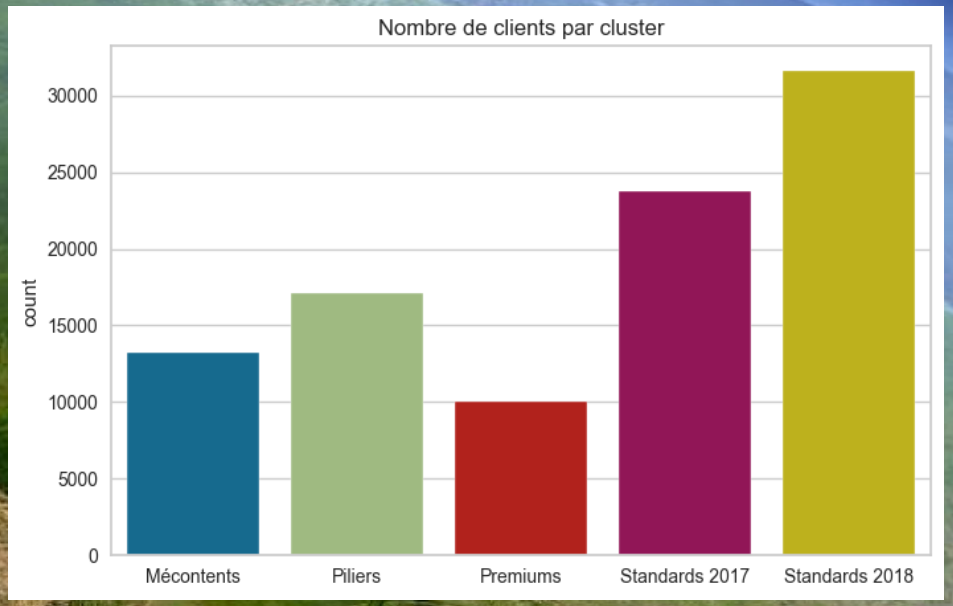
Exemple sur un échantillon de 10 000 clients:

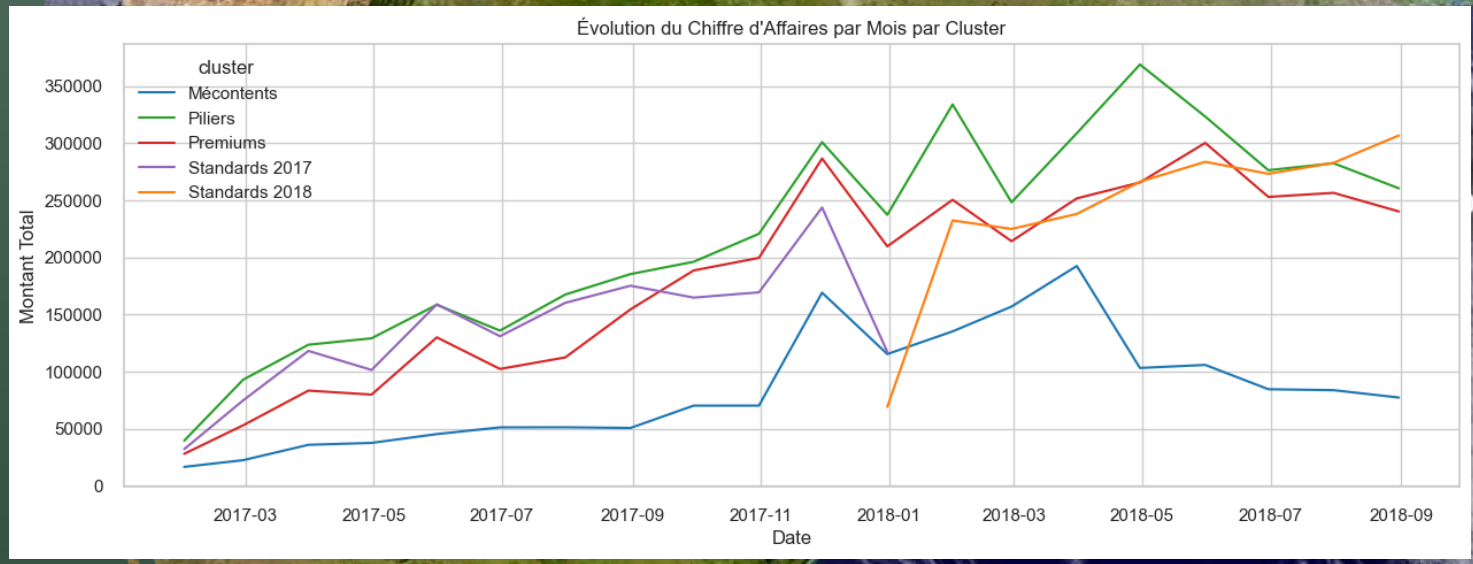
- esp=0.16 , min_samples=100 (1%)
  - 4 clusters
  - Bruit: 3 608 (36%)
  - Score silhouette:0,07

- L'augmentation du min_sample entraine une augmentation du bruit.
- 1 cluster identifier principalement.

5 Segments

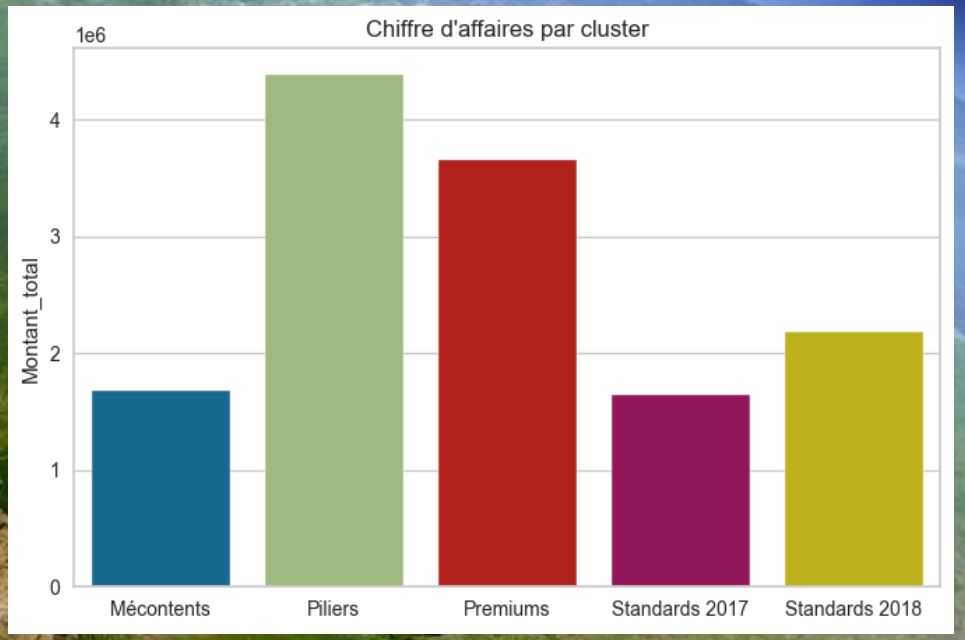
- Standards: 58%
  - 2017
  - 2018
- Mécontents: 13,8%
  - Score moyen: 1,2

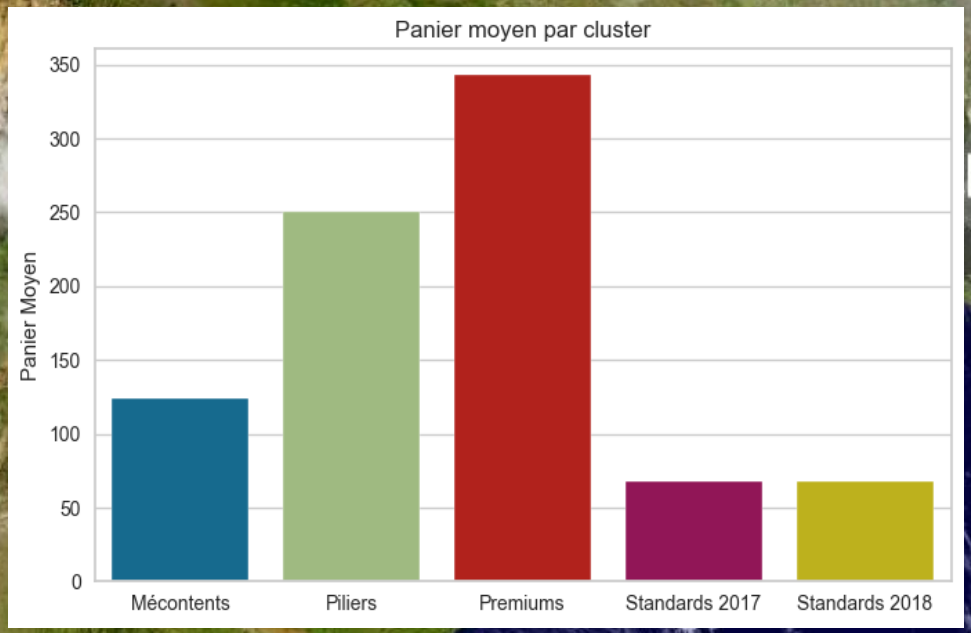




5 Segments

- Piliers:
  - 32,4% du CA
  - Panier moyen: 250 réales
  - Poids produits: élevé
- Premiums:
  - 27% du CA
  - Panier moyen: 342 réales
  - 10,5 % sont récidivistes

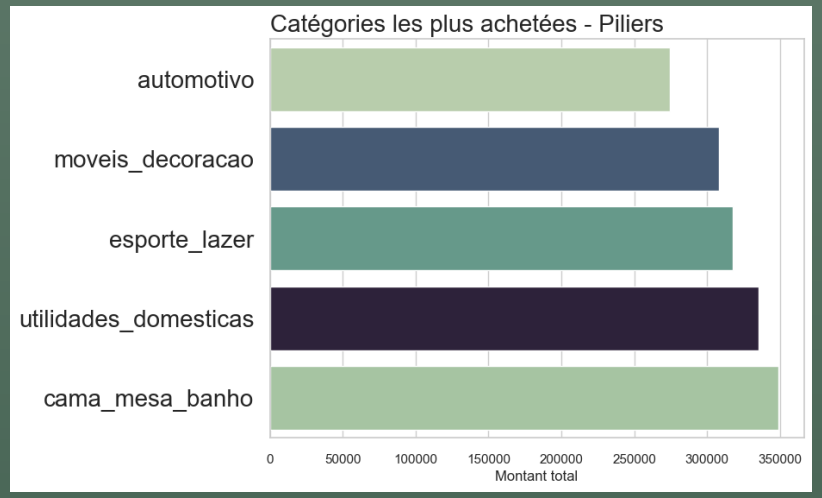


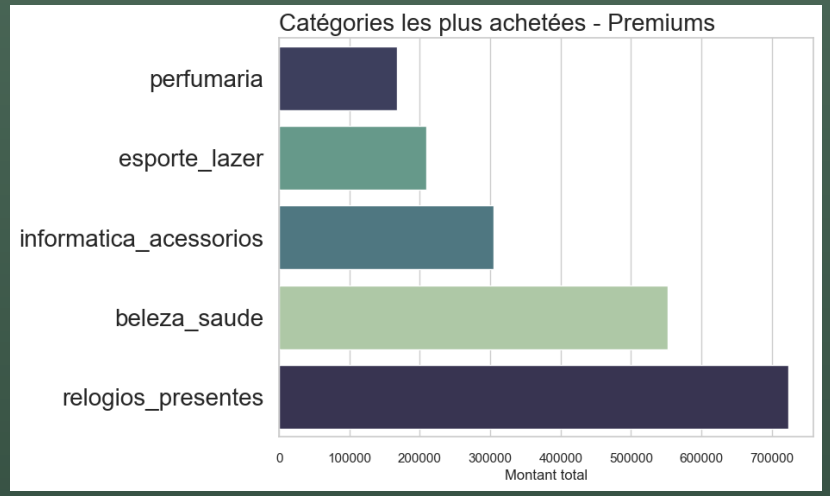


Quels produits?

- Piliers:
  - Meubles
  - Salle de bain

- Premium:
  - Montres
  - Produits de beauté





Simulation d’ajout de nouveaux clients

- Stabilité du modèle dans le temps
- Optimisation de la fréquence des segmentations
- Adaptation à l’évolution des comportements des clients

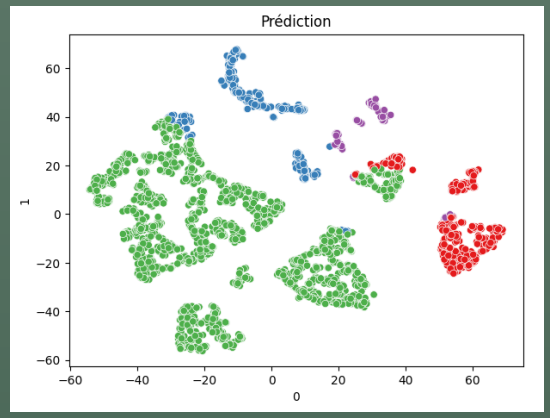
Prédiction vs Nouvelle segmentation

31 janvier 2018

Date variable

Segmentation (13 mois)

Prédiction
(1 à 7 mois)



Février:
ARI = 0,99

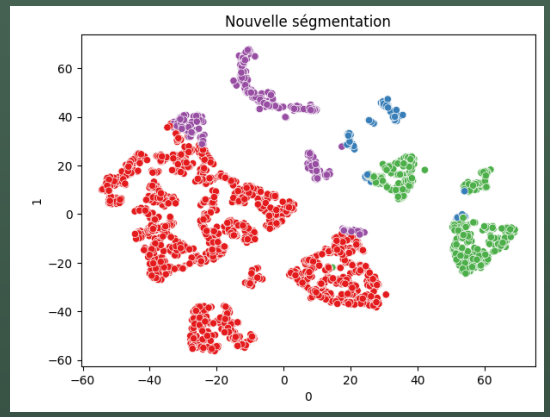
1 février au 10 août: 0,80
1 février au 11 août: 0,79

Février-août:
ARI = 0,78

0,8

Mesure de l'ARI

Segmentation (période totale)



Sensible aux choix de période, nombre de cluster, outliers…

Conclusion

- Sélection des variables pour comprendre le comportement des clients
- Choix de Kmeans:
  - Facile à paramétrer
  - Utilisable sur des grands ensembles de données
- 5 comportements distincts de la clientèle
- Simulation: modèle entrainé sur 13 mois et pour 5 clusters apporte des prédictions proches d’une nouvelle segmentation pendant plusieurs mois.