Projet n° 4

Parcours Ingénieur
Machine Learning

Segmentez des clients d'un site e-commerce

Présenté par
Alfred Bazin

Mentor
Amine Hadj-Youcef

25/09/2022

OPENCLASSROOMS

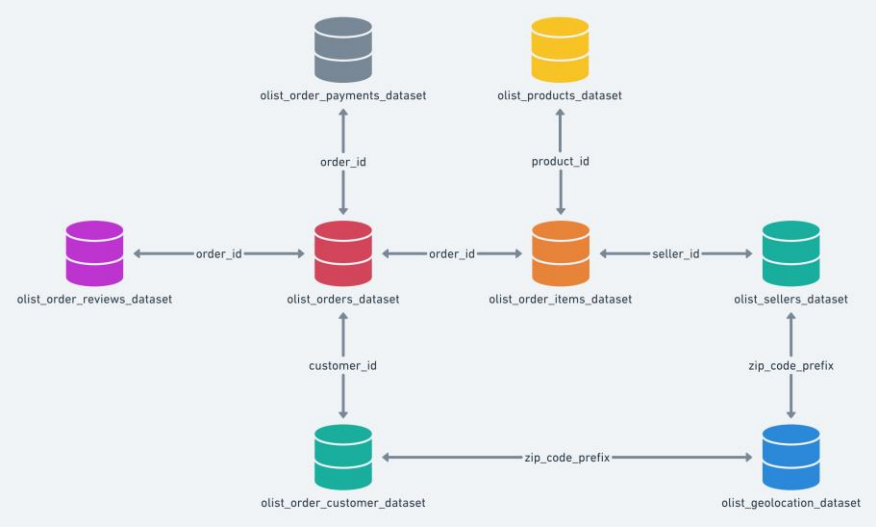
Sommaire

Présentation du projet

« **Segmentation** des clients utilisable au quotidien pour les campagnes de communication »

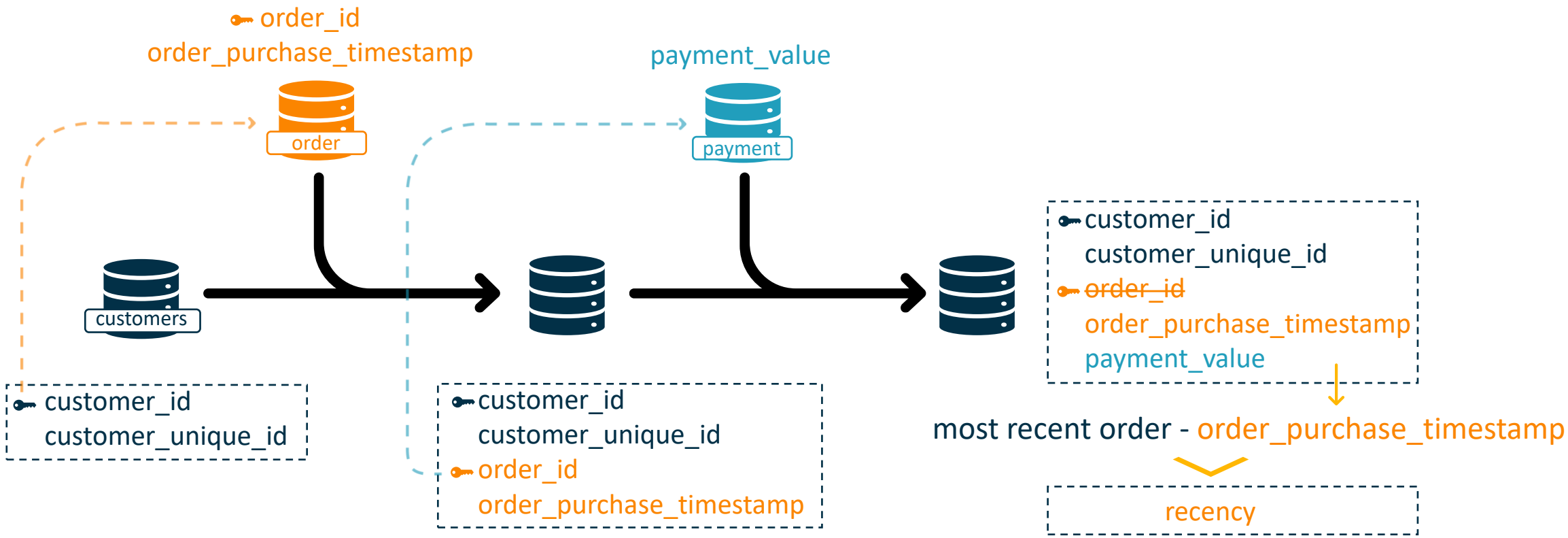
Données sur Kaggle :
https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

- **8** bases de données
- **99441** commandes
- **96096** utilisateurs



Nettoyage & Feature Engineering

Agrégation des bases de données (exemple RFM) :



Nettoyage & Feature Engineering

Agrégation des bases de données (exemple RFM) :

Groupés par :

customer_unique_id

customer_id	*count*	Frequency
recency	*min*	Recency
payment_value	*sum*	Monetary Value

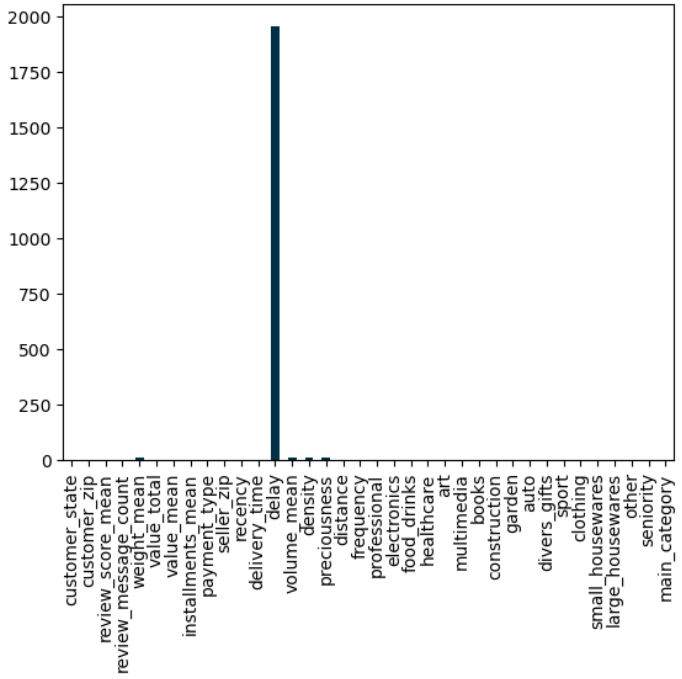
Autres features :

review_score (*mean*)
review_message (*count*)
Weight (*mean*)
Volume (*mean*)
Density (*mean*)
Preciousness (*mean*)
delivery_time (*mean*)
delay (*mean*)
Distance (*mean*)
Value (*mean*)
Installments (*mean*)
payment_type (*mode*)
customer_state (*mode*)
main_category (*Encoding…*)

Nettoyage & Feature Engineering

Valeurs manquantes et aberrantes

Valeurs manquantes

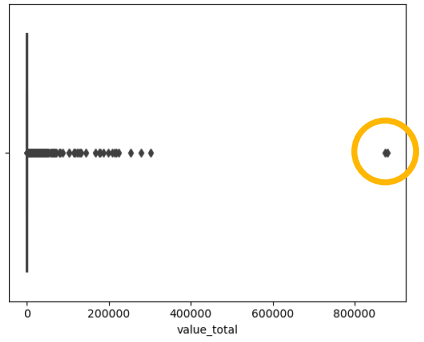


Weight
Volume
Density
preciousness
} Supprimées

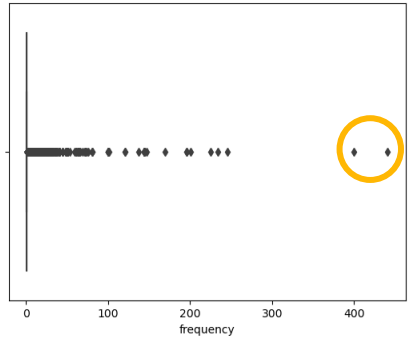
Delay → Imputés avec delivery time

Valeurs aberrantes et atypiques

Valeur monétaire totale



Fréquence

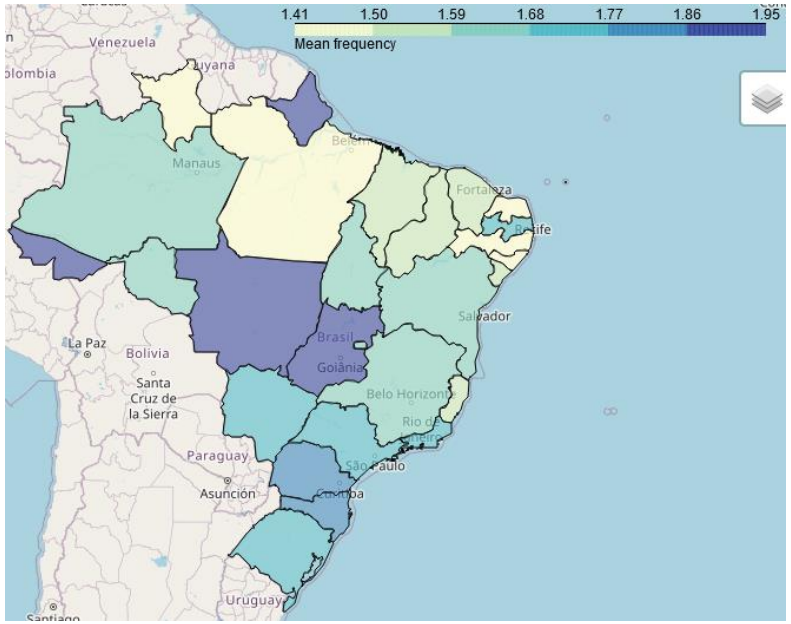


Nettoyage & Feature Engineering

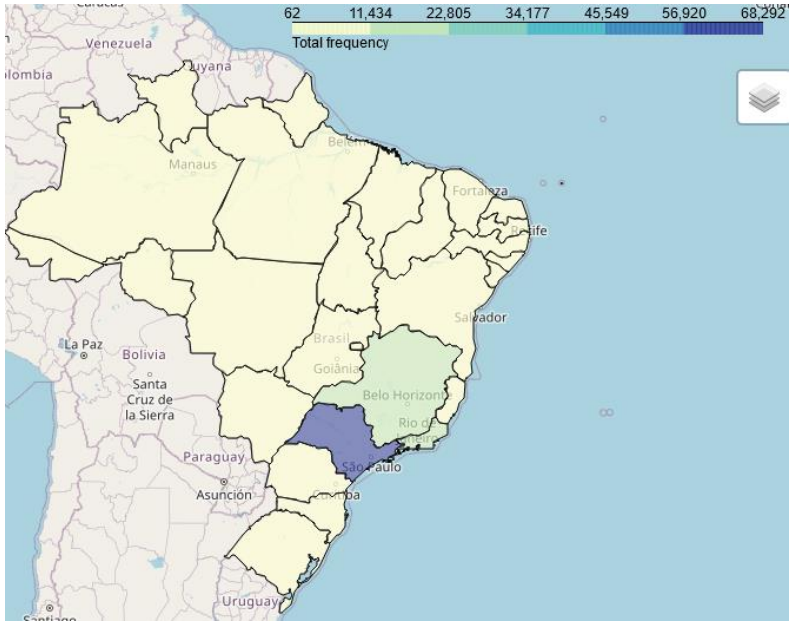
Analyse exploratoire

- La **fréquence moyenne** est **équilibrée** entre les états
- Les **états** plus **peuplés** représentent une **large majorité des commandes**

Fréquence **moyenne** par état

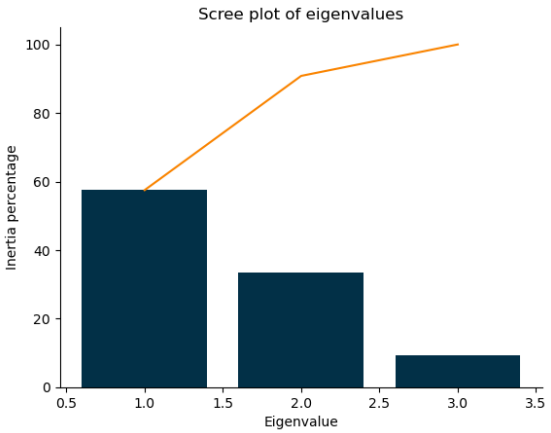


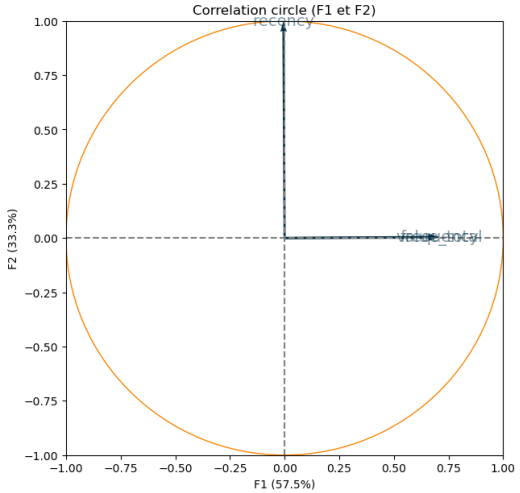
Fréquence **totale** par état

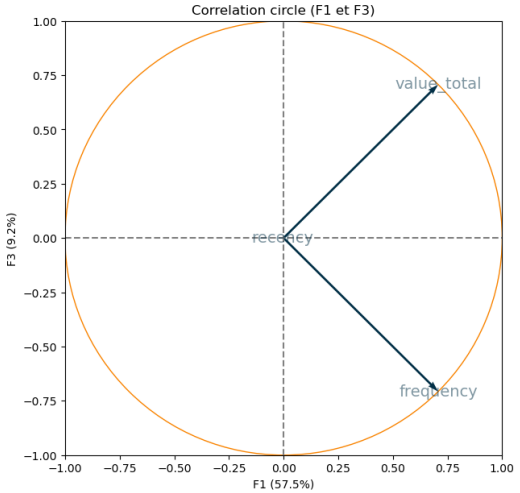


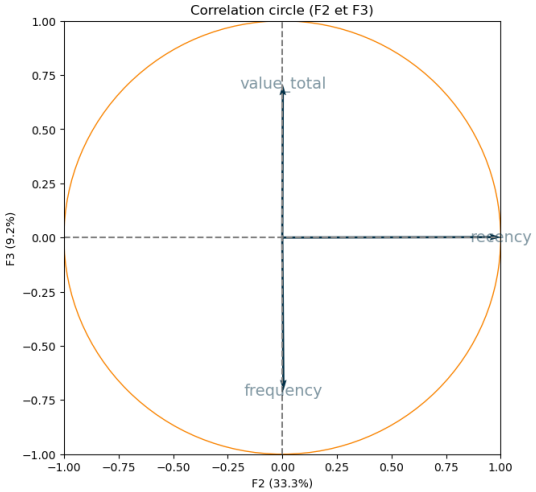
Nettoyage & Feature Engineering

Analyse exploratoire







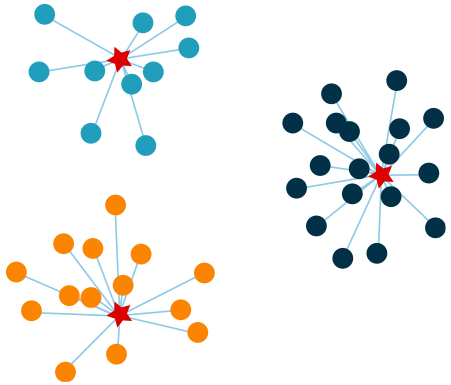


- F1 : Quantitée d'achats (*value_total* + *frequency*)
- F2 : La récence des achats (*recency*)
- F3 : Le prix des achats (*frequency* vs *value_total*)

Modélisation

RFM - KMeans

(Sur un échantillon de ~10 000 individus)

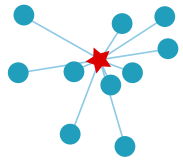


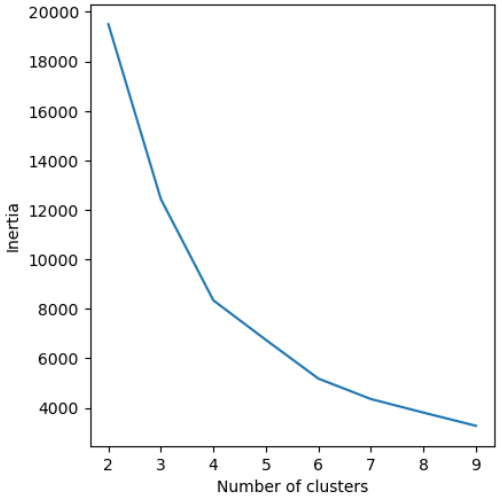
Paramètres :

- n_clusters
- init
- n_init

Inertie

Somme de distance des points au centroïde :





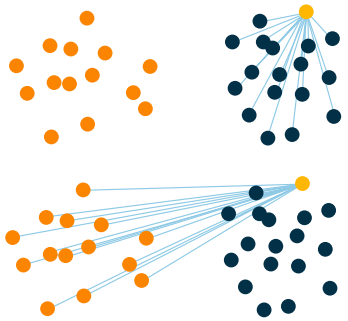
$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

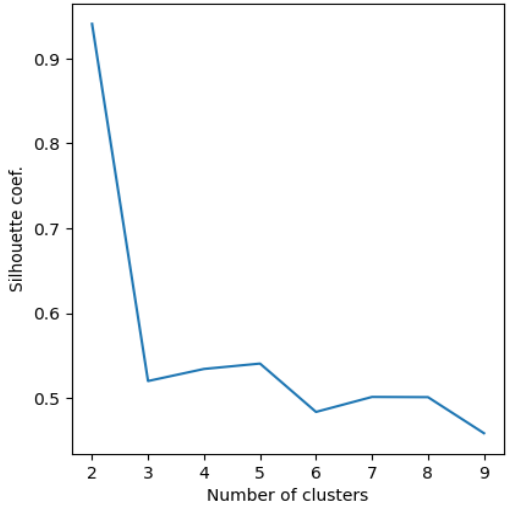
Métriques :

Coefficient de silhouette

a : Distance moyenne entre un **point** et ses **voisins** du **même cluster**

b : Distance moyenne entre un **point** et ses **voisins** du **cluster le plus proche**





$$s = \frac{b - a}{max(a, b)}$$

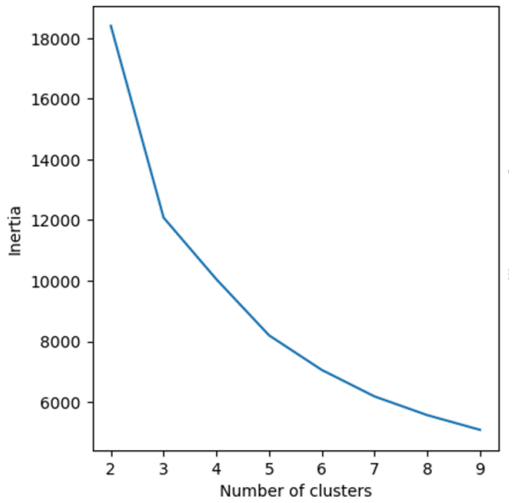
1 : Parfait

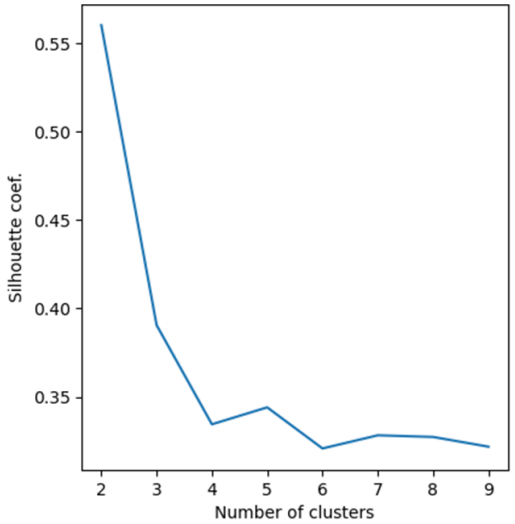
0 : Les clusters se recouvrent

-1 : Mauvais clustering

Modélisation

RFM (log) - KMeans



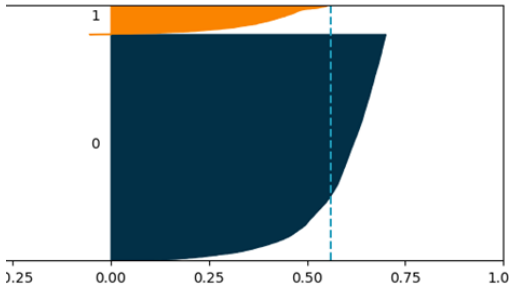


Inertie
- Coude entre **3** et **5** clusters

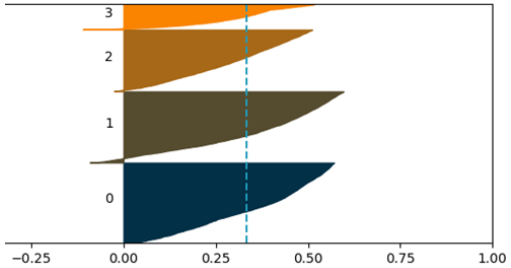
Coef. de silhouette
- Max à **2** clusters
- Chute avec un max à **5** clusters

Profil de silhouette

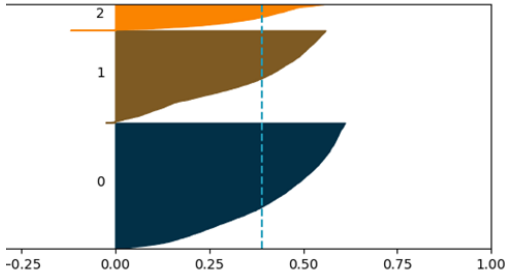
n clusters = **2**



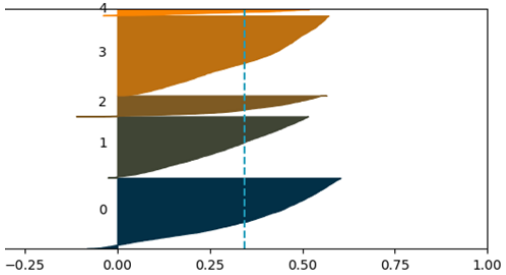
n clusters = **4**



n clusters = **3**



n clusters = **5**



Profil de silhouette
- Clusters plus **équilibrés**
- Toujours **chevauchement**

➡ Résultats **similaires** avec **ACP**

Modélisation

RFM (log) - KMeans

Ajout de features

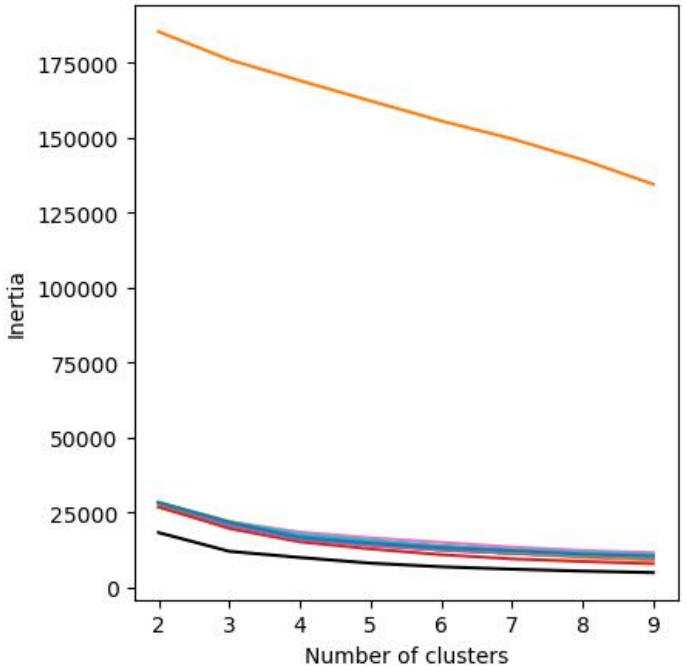
- Ajout de features à RFM :

Inertie ↗ Silhouette ↘

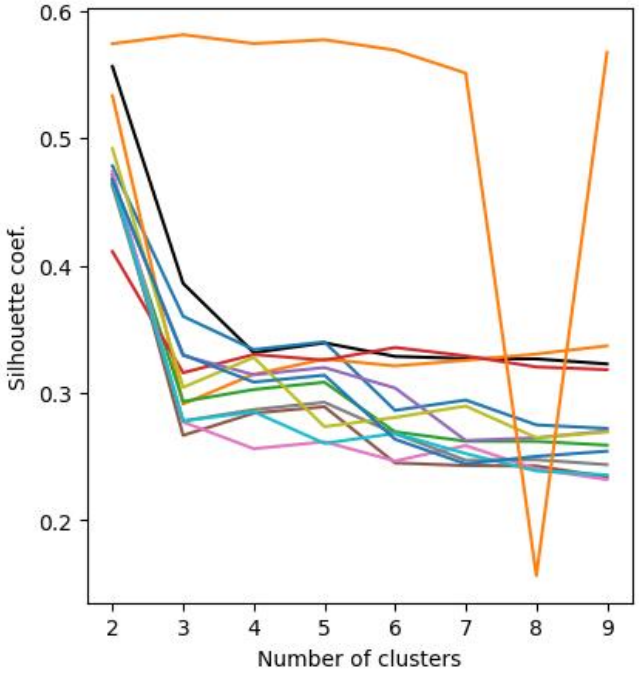
- Exceptions :
  - review_score_mean
  - review_message_count
  - value_mean
  - Cat. Features

↳ Explosion de l'inertie

Inertie



Coef. de silhouette



Features

Modélisation

RFM (log) - KMeans

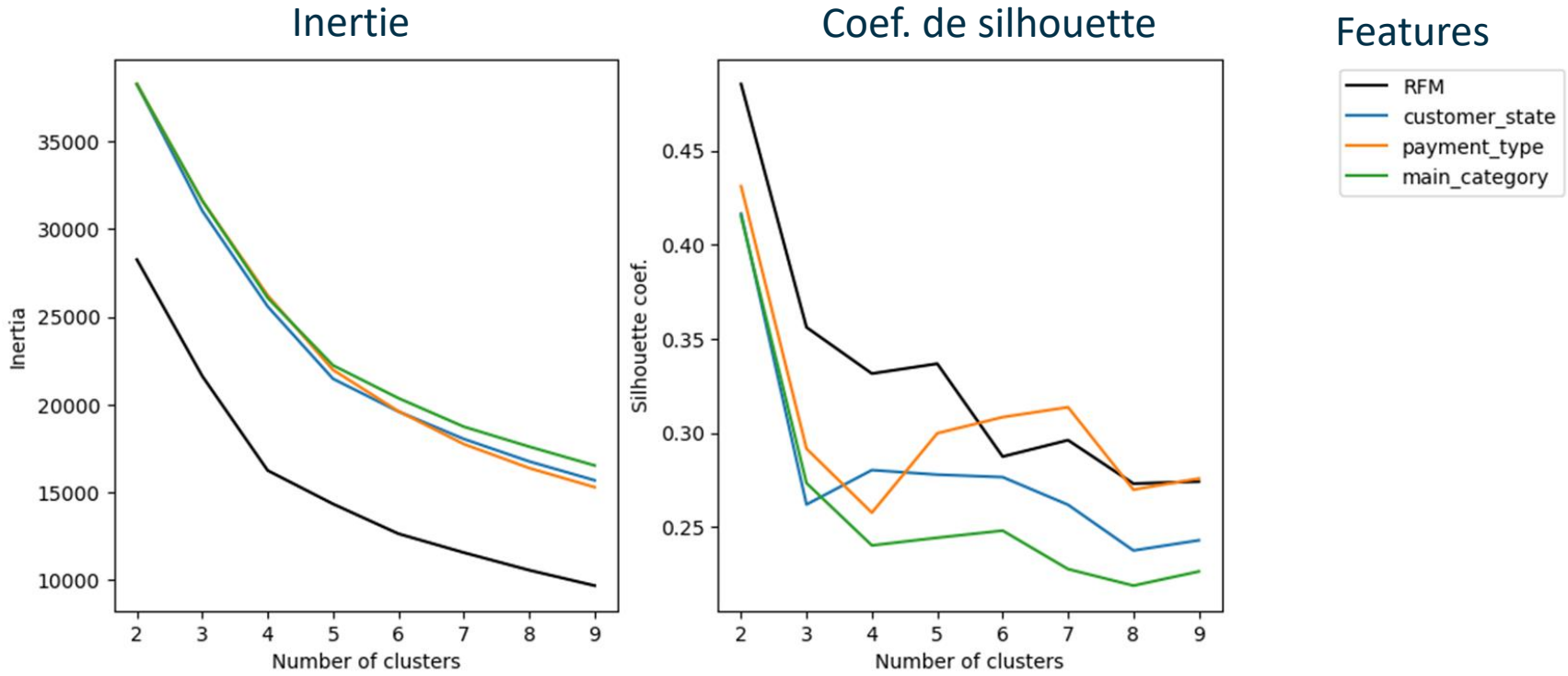
Ajout de features catégoriques

- Ajout de features à RFM :

Inertie ↗ Silhouette ↘

- Exceptions :
  - Payment_type

↳ Haut nombre de clusters avec recouvrement



Modélisation

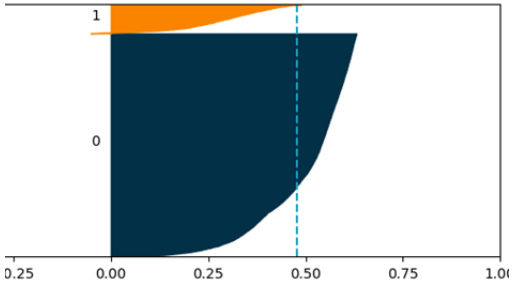
RFM+ (log) - KMeans

Features :

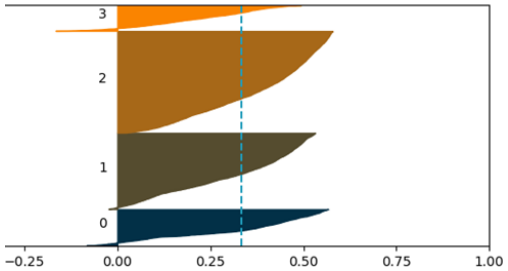
- recency
- frequency
- monetary value
- review_score_mean

Profil de silhouette

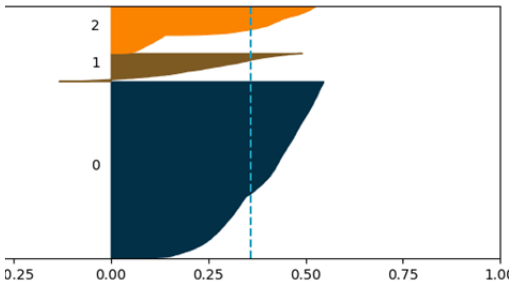
n clusters = **2**



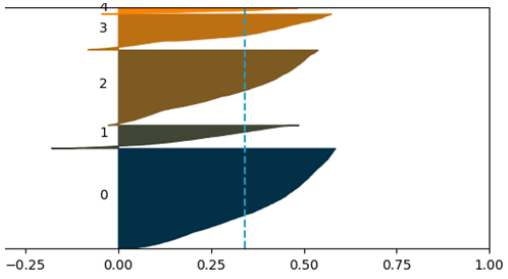
n clusters = **4**



n clusters = **3**



n clusters = **5**



Profil de silhouette

- Toujours **chevauchement**

➡ Résultats **similaires** avec **ACP**

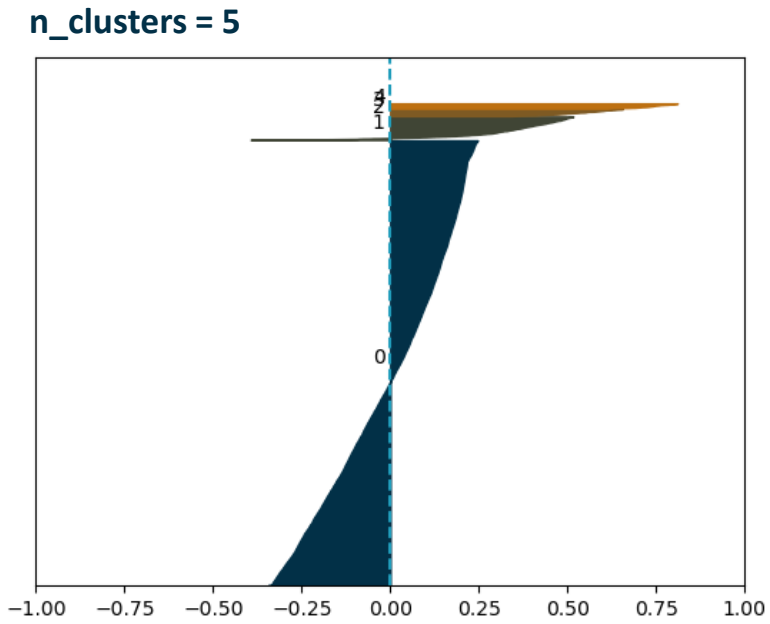
Modélisation

RFM+ (log) - DBSCAN

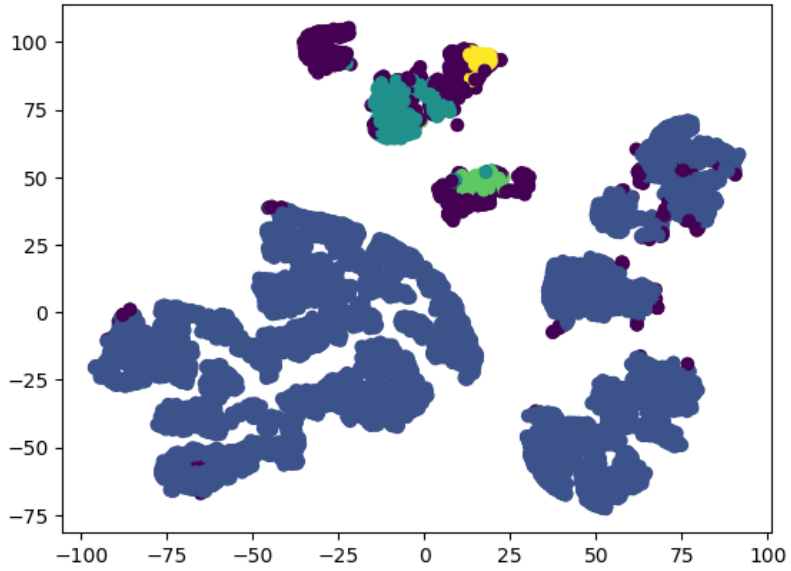
Paramètres :

- eps
- min_sample

Profil de silhouette



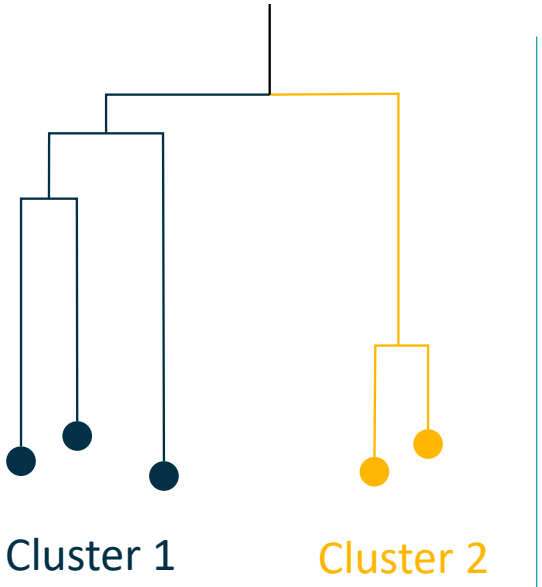
t-SNE



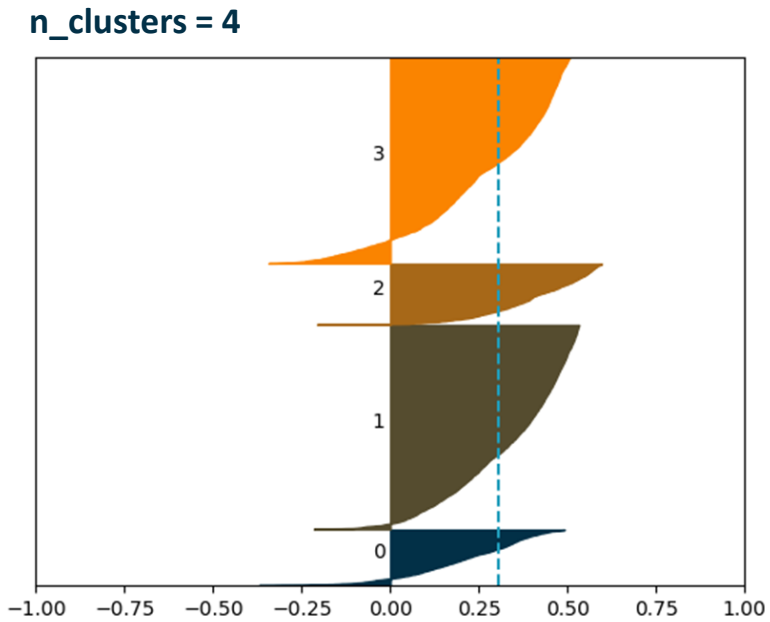
- Obtention de clusters **déséquilibrés** et se **chevauchant** fortement

Modélisation

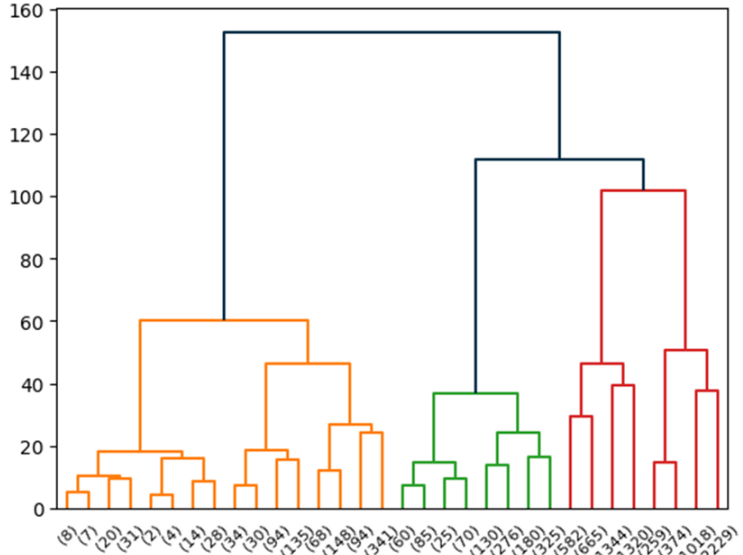
RFM+ (log) – Clustering hiérarchique



Profil de silhouette



Dendrogramme



- Résultats très proches du **KMeans** mais avec plus de **chevauchement**

Modélisation

RFM+ (log) – KMeans

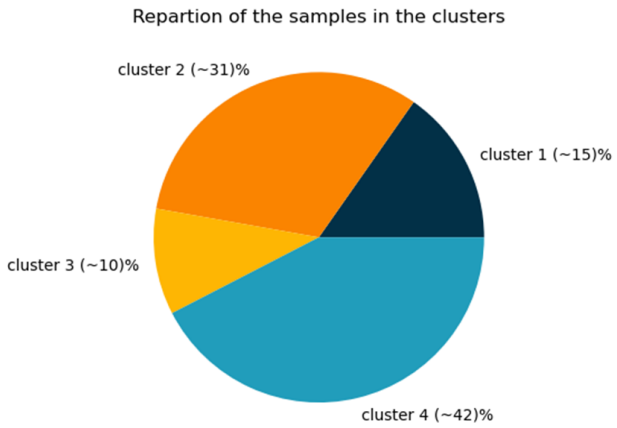
Interprétation des clusters (n_clusters = 4)

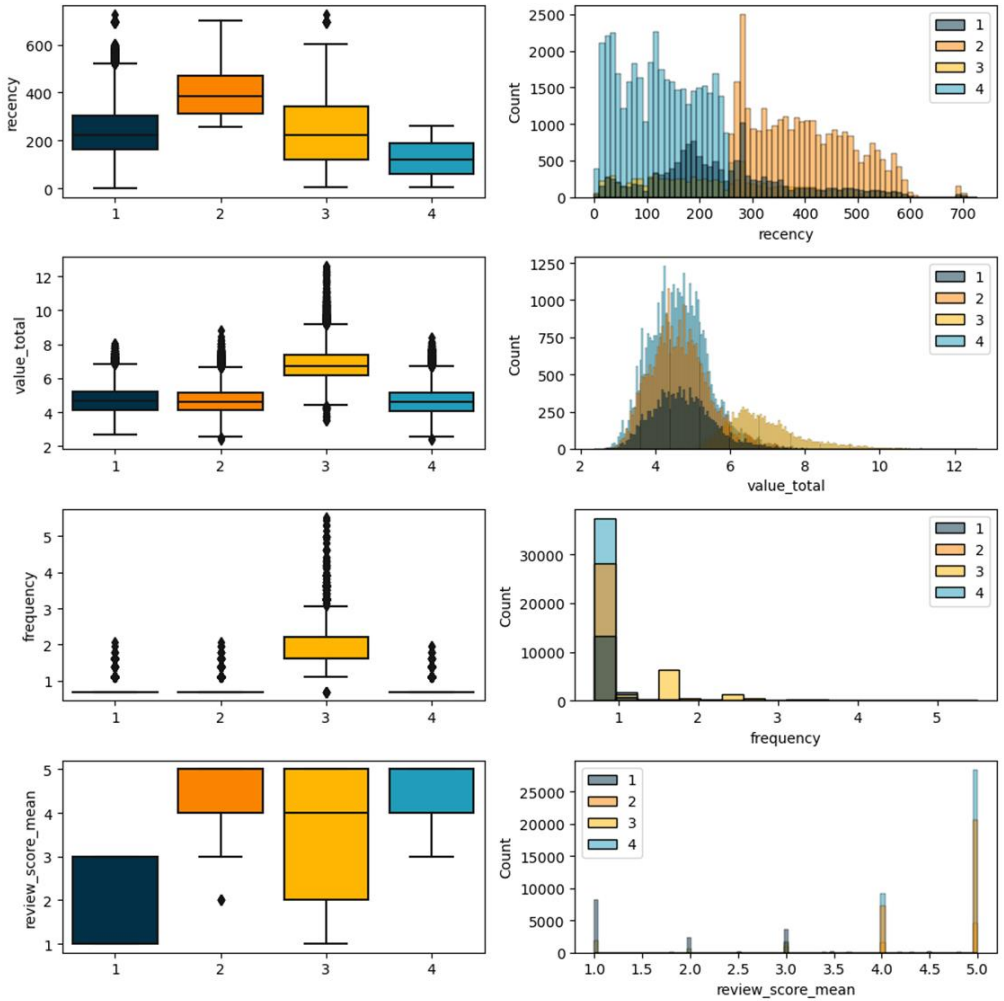
Cluster 1:
Clients qui ne sont **pas satisfaits** de leur commande.

Cluster 2 :
Clients qui ont commandé peu de fois **il y a longtemps.**

Cluster 3 :
Clients qui ont commandés **plusieurs fois** sur la plateforme.

Cluster 4 :
Clients **récents** qui n'ont pas commandé beaucoup et sont content.

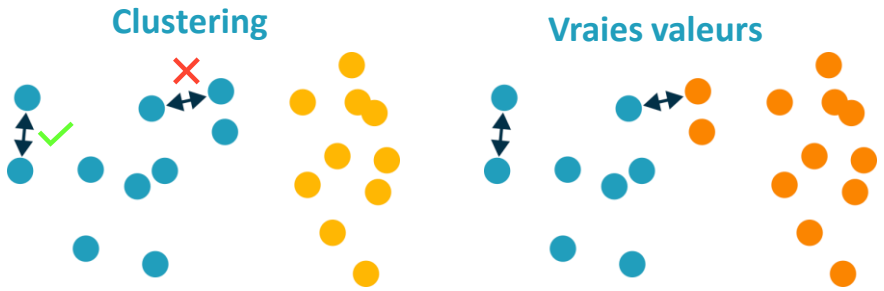




Simulation

ARI

Index de rand (RI)



$$RI = \frac{\sum_{k=0}^{n} \text{◀▶✓}}{n \text{◀▶}}$$ [1]

Index de rand ajusté (ARI)

$$ARI = \frac{RI - \mathbb{E}(RI)}{1 - \mathbb{E}(RI)}$$ [1]

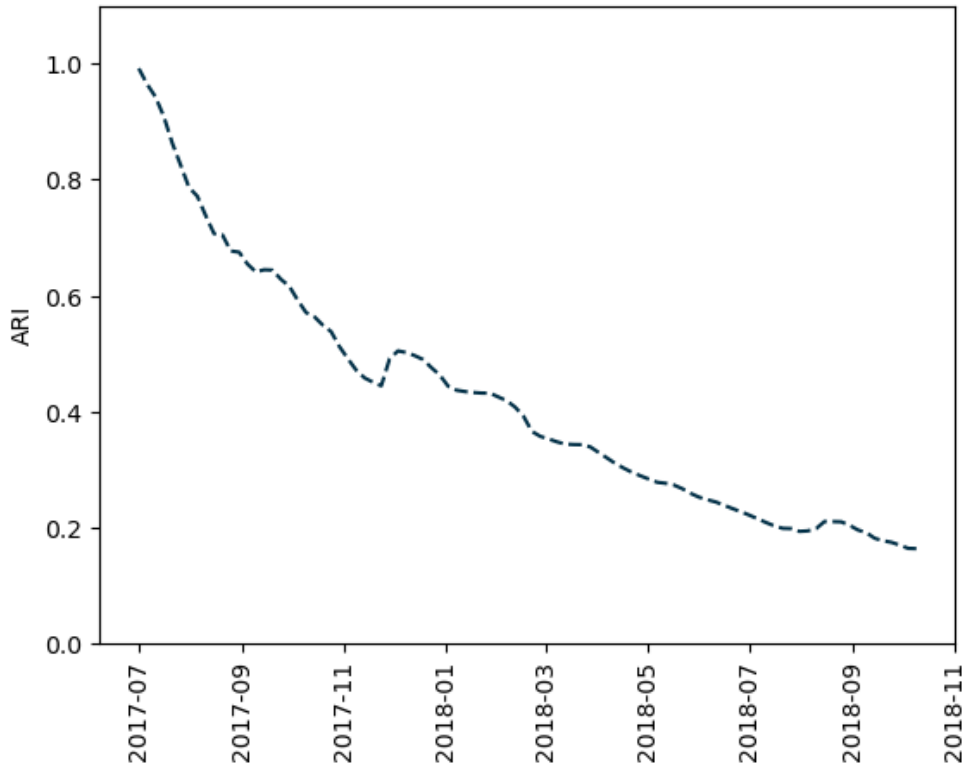
1. A. J. Gates, Y. Ahn, *The Impact of Random Models on Clustering Similarity*, Journal of Machine Learning Research 18 (2017) 1-28

Simulation

(début après un semestre de données)

- Sans maintenance
  - Diminution jusqu'à un score de ~0,2

ARI sans maintenance

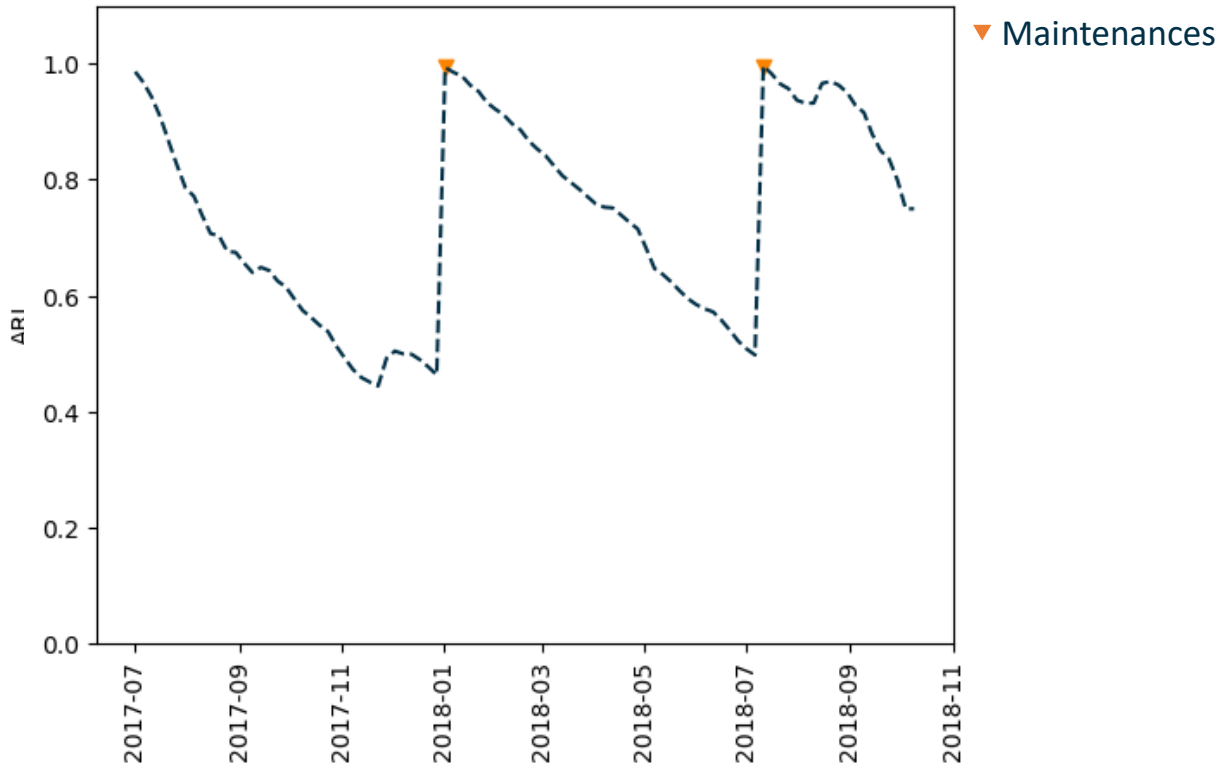


Simulation

(début après un semestre de données)

- Sans maintenance
  - Diminution jusqu'à un score de ~0,2
- 6 mois
  - Diminution jusqu'à un score de ~0,4

Maintenance tous les 6 mois

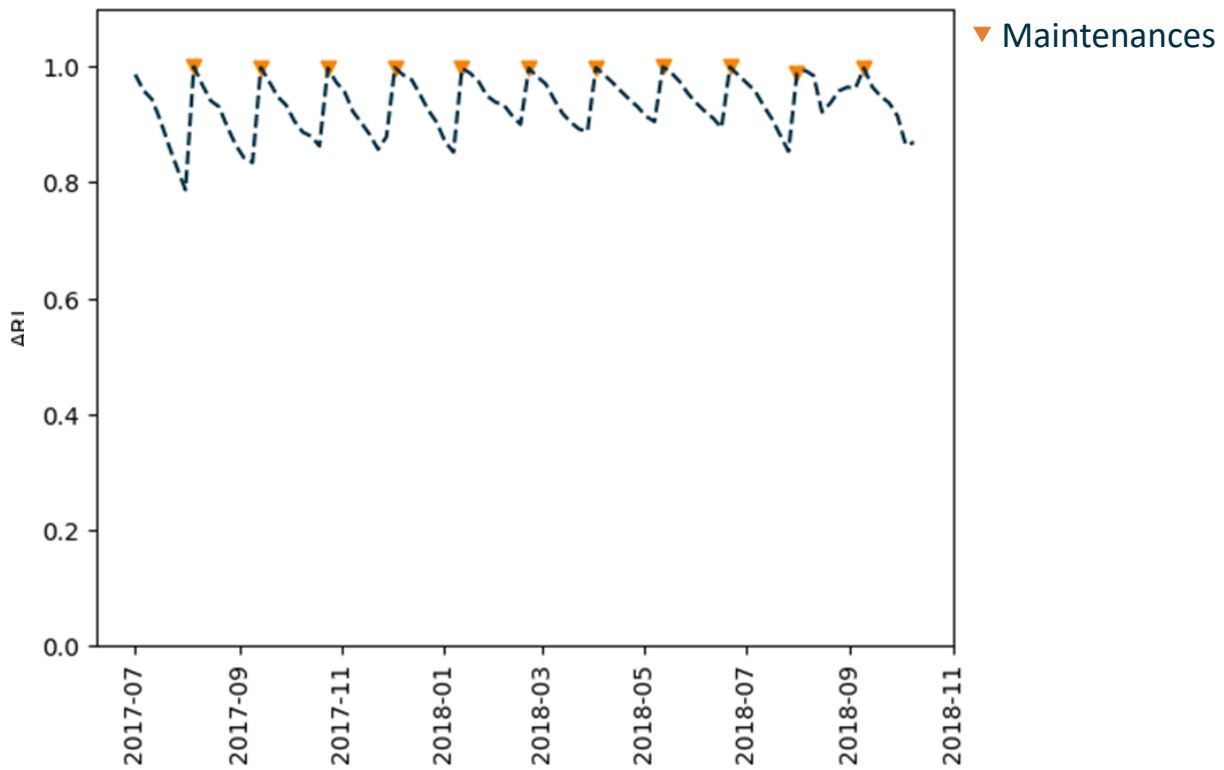


Simulation

(début après un semestre de données)

- Sans maintenance
  - Diminution jusqu'à un score de ~0,2
- 6 mois
  - Diminution jusqu'à un score de ~0,4
- 1 mois
  - Diminution jusqu'à un score de ~0,8

Maintenance tous les mois

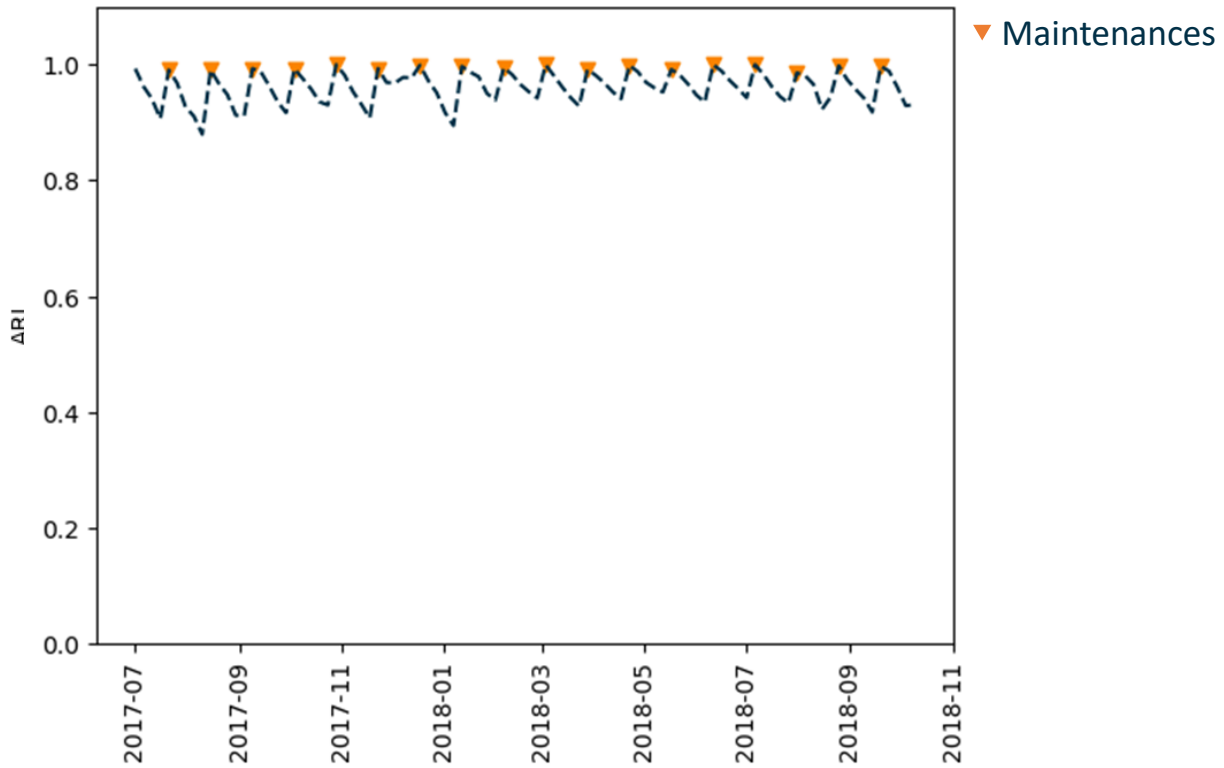


Simulation

(début après un semestre de données)

- Sans maintenance
  - Diminution jusqu'à un score de ~0,2
- 6 mois
  - Diminution jusqu'à un score de ~0,4
- 1 mois
  - Diminution jusqu'à un score de ~0,8
- 15 jours
  - Diminution jusqu'à un score de ~0,85

Maintenance tous les mois



Conclusion

- Clustering avec features **RFM** et **score moyen** :

Rency Frequency Monetary value Mean review score

- Parmi plusieurs modèles (**DBSCAN**, **Clustering hiérarchique**…), **KMeans** est le **plus performant**.

- **4 Clusters** formés :
  - Clients **récents**
  - Clients qui ne **reviennent pas**
  - Clients **fréquents**
  - Clients **mécontents**

- Maintenance de **15 jours** recommandée