

FORMATION DATA SCIENTISTS
OPENCLASSROOMS
20 JUILLET 2021 - 18 MAI 2022

Romain Vaillant Janvier 2022



P5 : SEGMENTEZ DES CLIENTS D'UN SITE E- COMMERCE



Segmentez des clients d'un site e-commerce



Sommaire

1. Olist
2. Objectifs du projet
3. Nettoyage des données
4. Approches de modélisation
5. Simulation
6. Conclusion

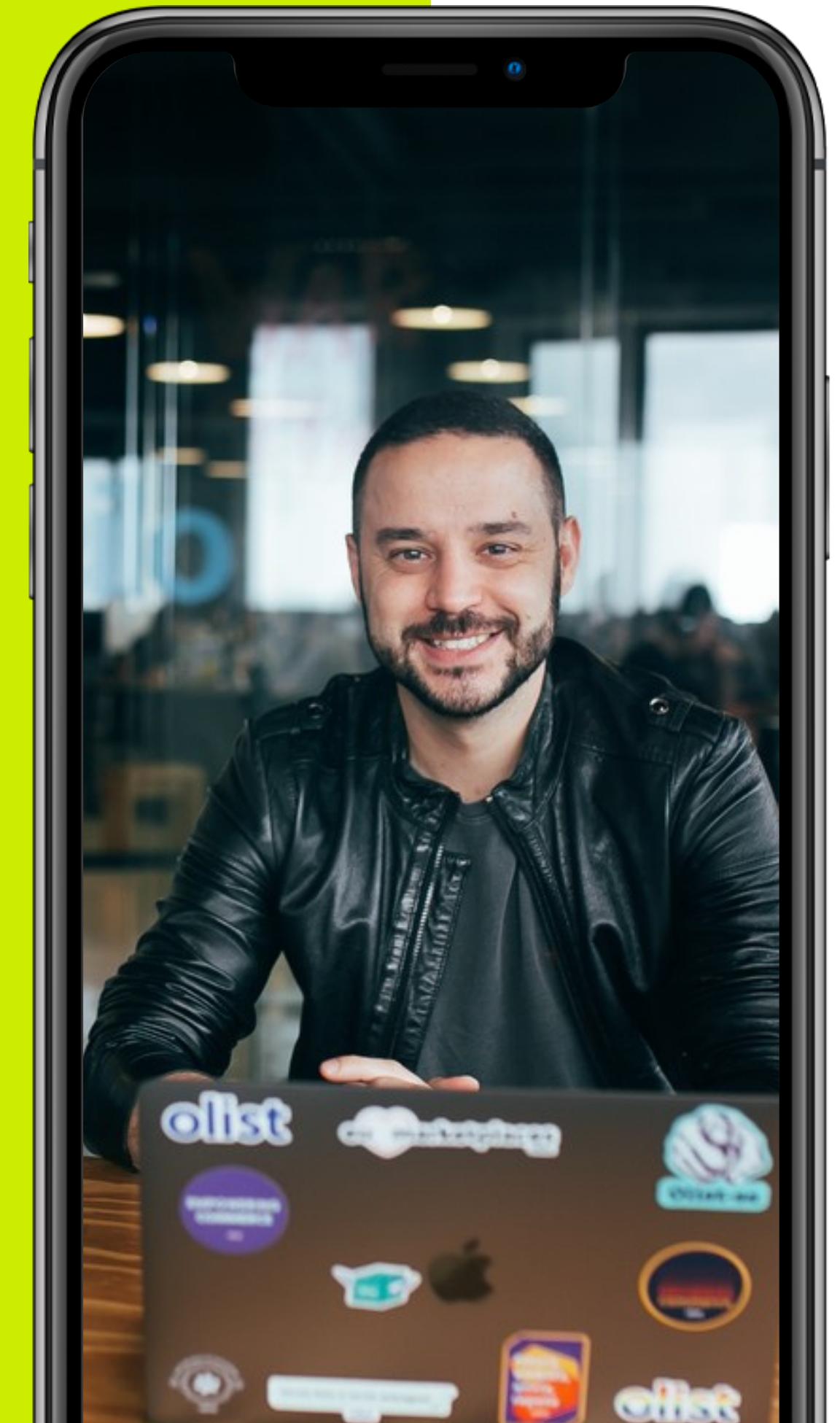
Olist

Olist

**Plate-forme de vente en ligne
Brésilienne. Ecosystème de mise en
relation clients-vendeurs pour
particuliers et entreprises**

3

- Précédemment marketplace pour particuliers depuis 2011
- Eclosion du projet en 2015 avec l'intégration des grandes enseignes de retail Brésiliennes "store in store" et l'arrivée d'investisseurs institutionnels
- Plus jeune licorne d'Amérique Latine
- Vocation mondiale



Objectifs

Objectifs :

1

Segmentation des clients

Comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles

3

Contrat de maintenance

Fournir une proposition de contrat de maintenance basée sur une analyse de la stabilité des clusters clients au cours du temps

2

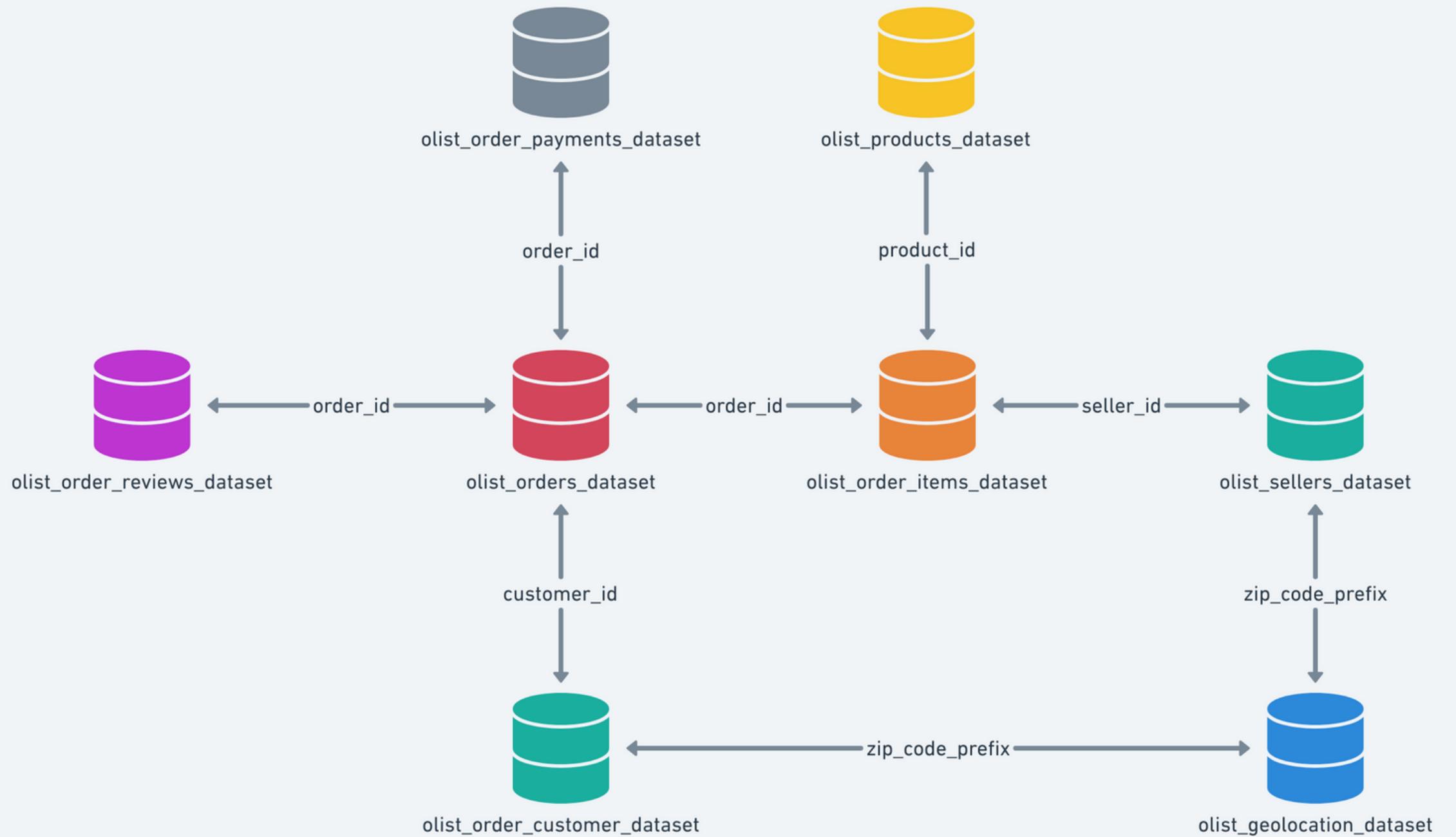
Description actionable

Fournir à l'équipe marketing une description actionable de la segmentation et de sa logique sous-jacente pour une utilisation optimale

5

Nettoyage des données

Jeux de données



Olist

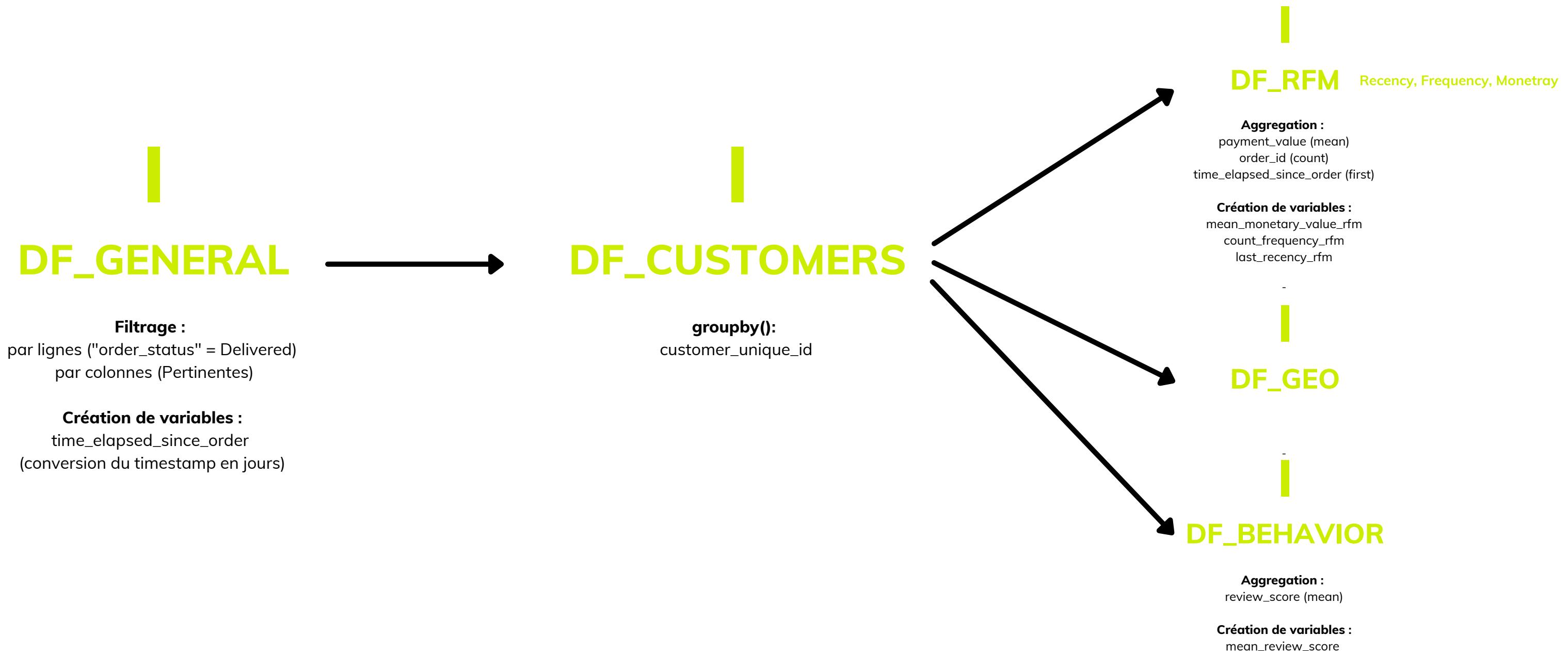
8 datasets

Merge

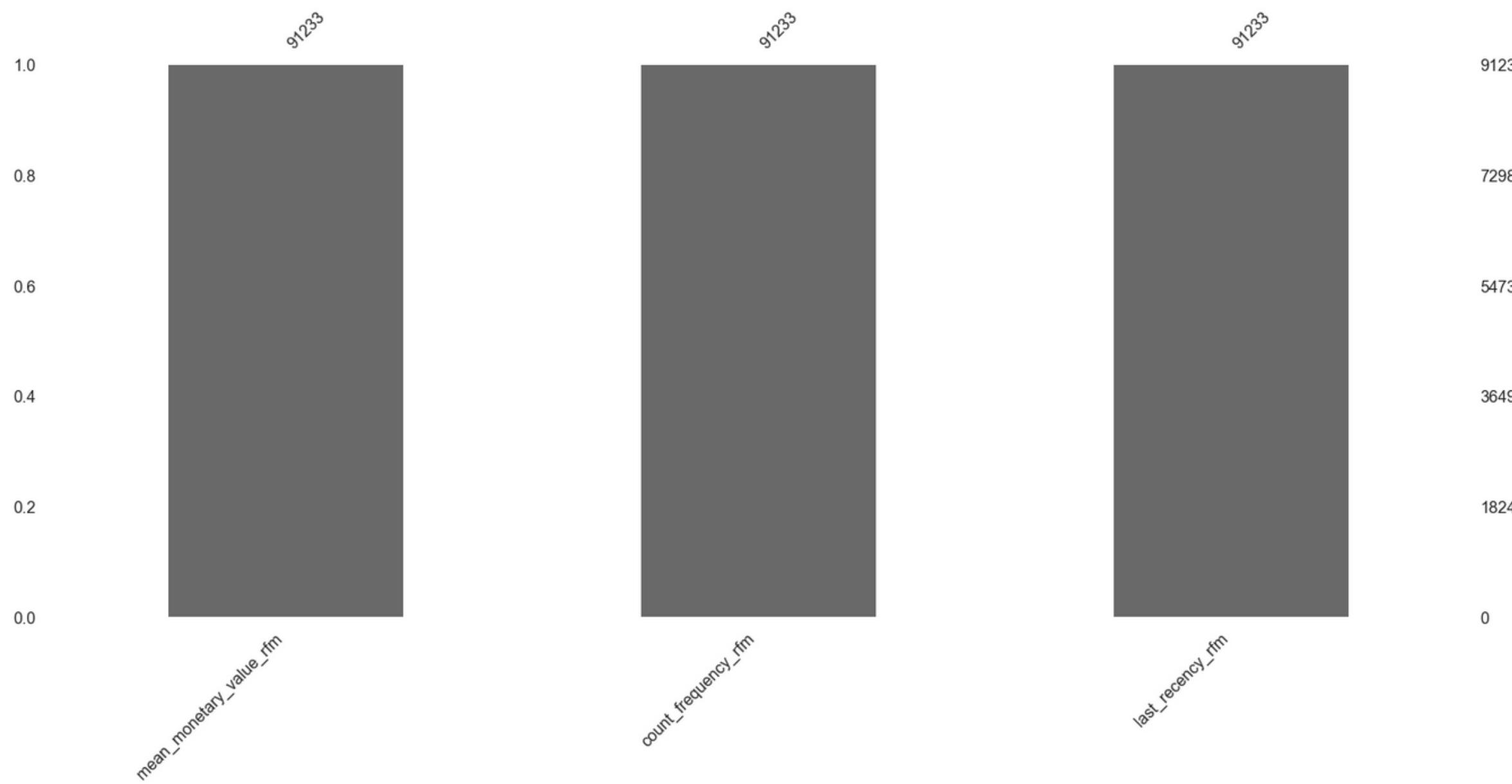
df_general

- 44 colonnes
- 115 299 lignes
- 3,4% de valeurs manquantes

NETTOAYGE DU JEU DE DONNEES



DATASET : DF_RFМ (BASELINE)



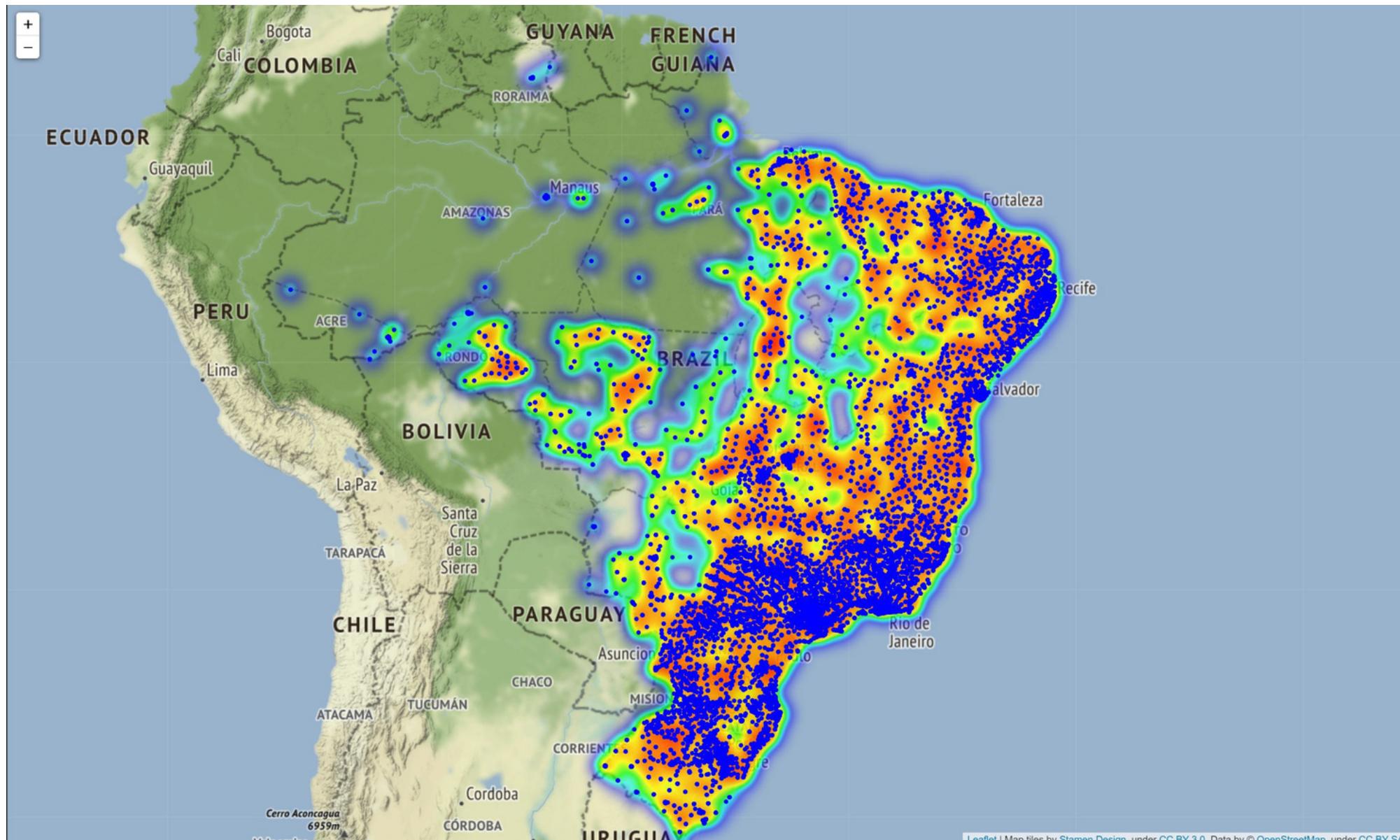
NETTOYAGE

115 299

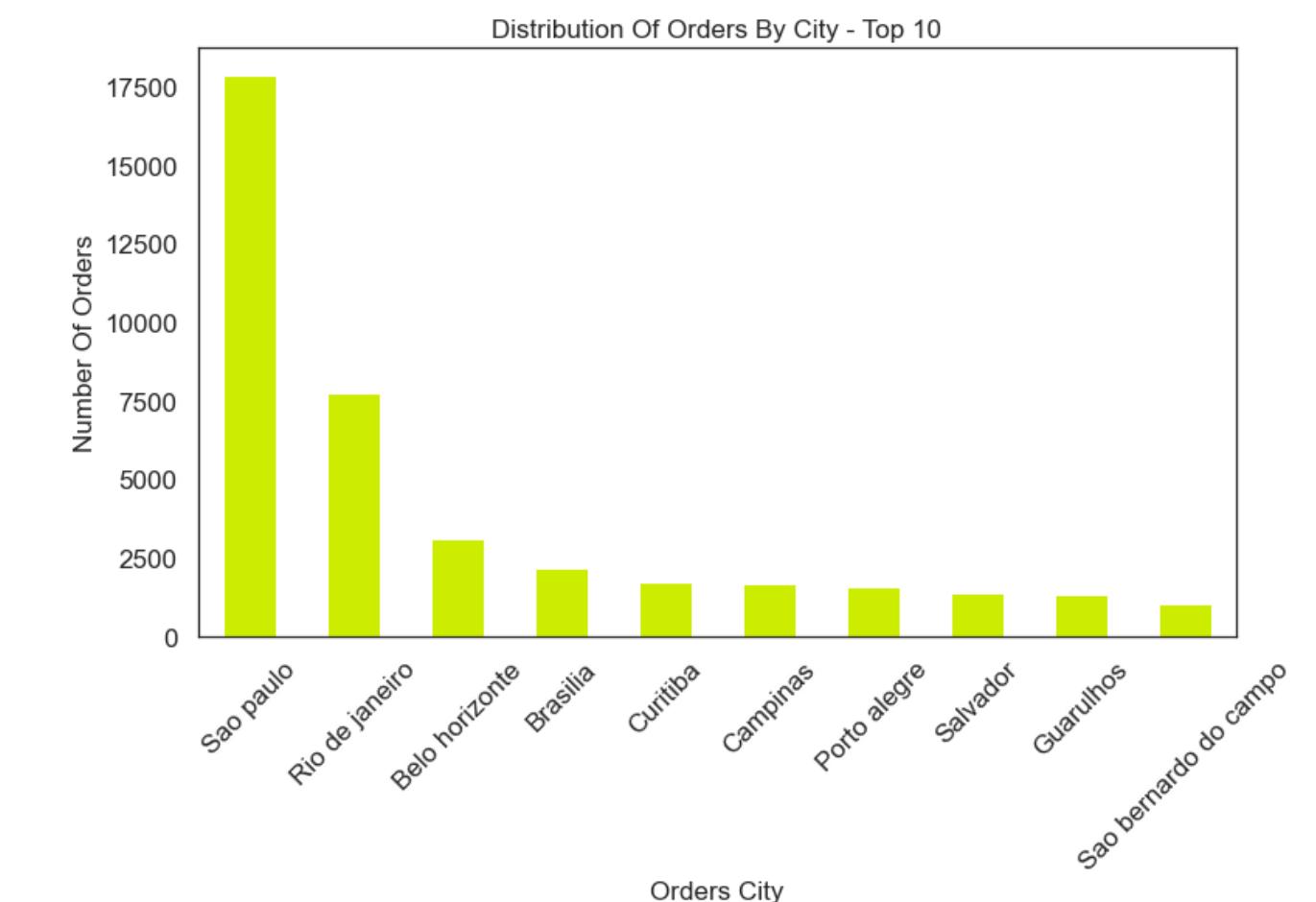


91 233 Lignes

APERCU DE LA LOCALISATION DES COMMANDES LIVREES

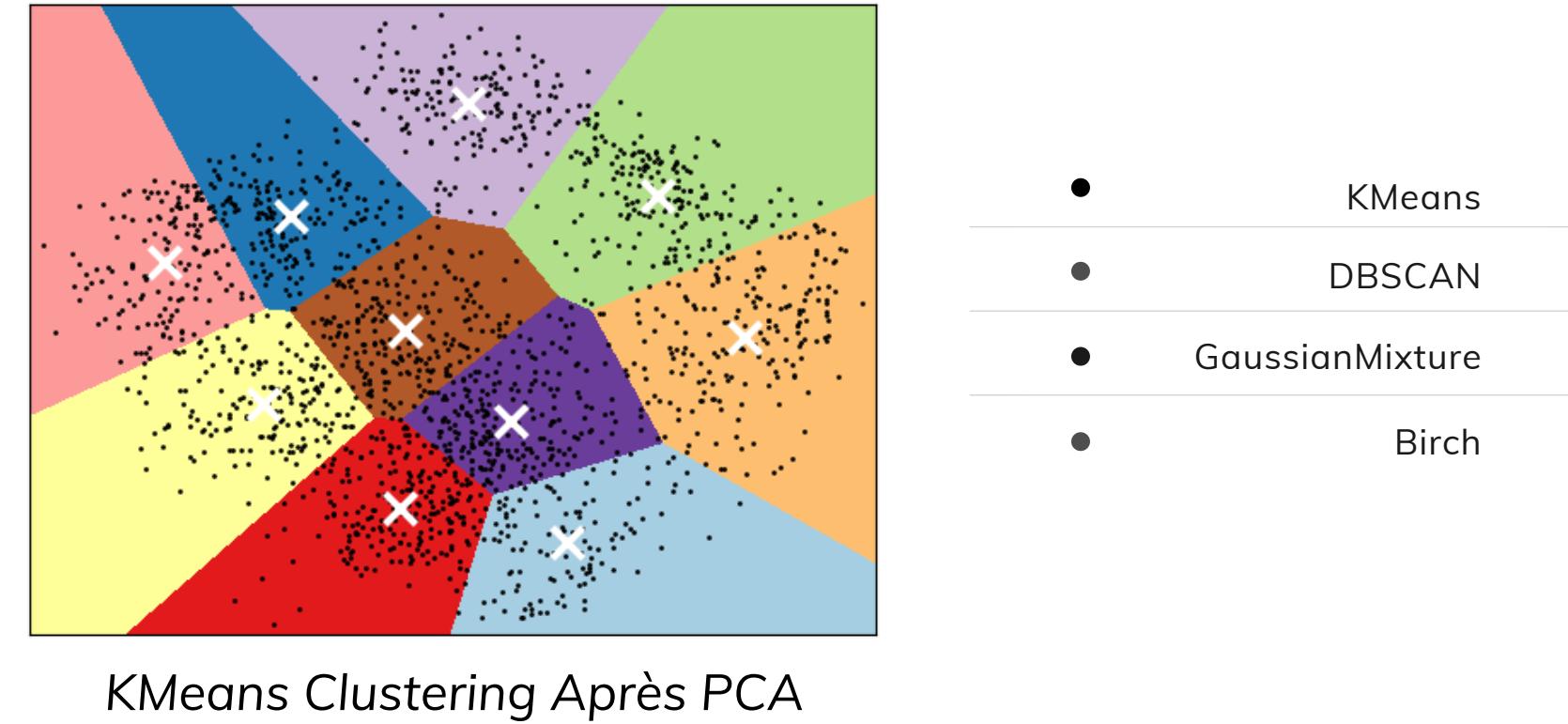


REPARTITION DES COMMANDES PAR VILLE TOP 10

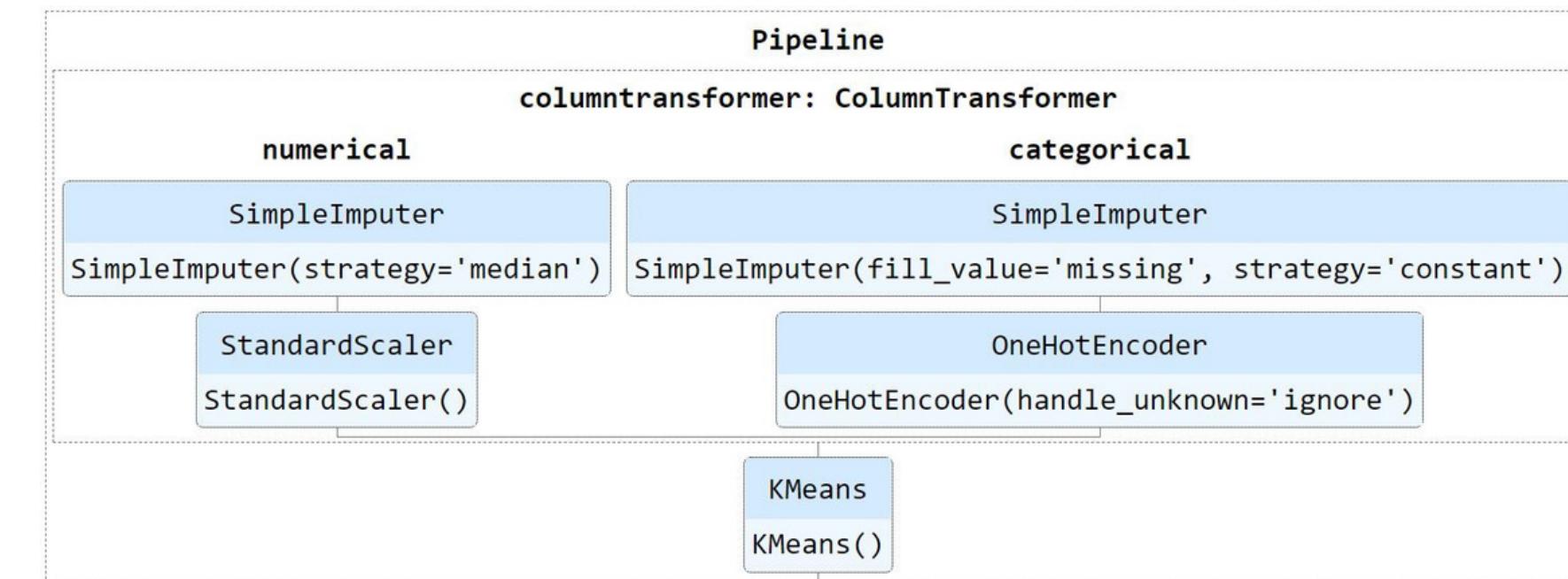


Approches de modélisation

CLUSTERINGS SELECTIONNES



FEATURES ENGINEERING



Quatre Use Cases

RFM (Baseline)

RFM
+ Geographie

RFM
+ Comportement

RFM
+ Geographie
+ Comportement

Variables

Recency
Frequency
Monetary Value (Total)

Variables

Recency
Frequency
Monetary Value (Total)

Customer state (C)
Geolocation city (C)
Zip code prefix (C)

Variables

Recency
Frequency
Monetary Value (Total)

Payment type (C)
Product Category (C)

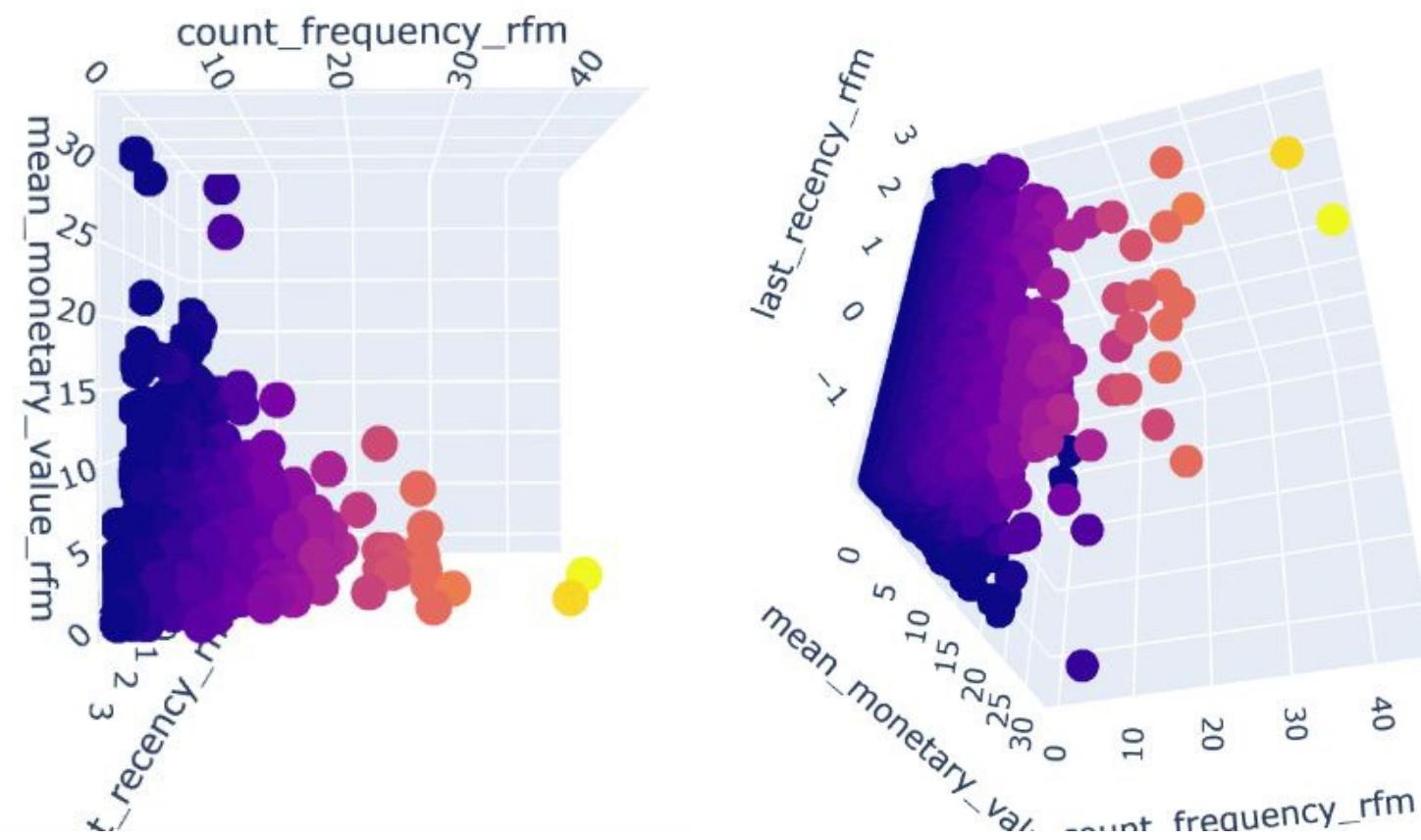
Variables

Recency
Frequency
Monetary Value (Total)

Customer state (C)
Geolocation city (C)
Zip code prefix (C)
Payment type (C)
Product Category (C)

PERFORMANCE DES USE CASES

BIRCH (TUNING)



PERFORMANCE DES ALGORITHMES DE CLUSTERINGS (TUNING)

USE CASE 3

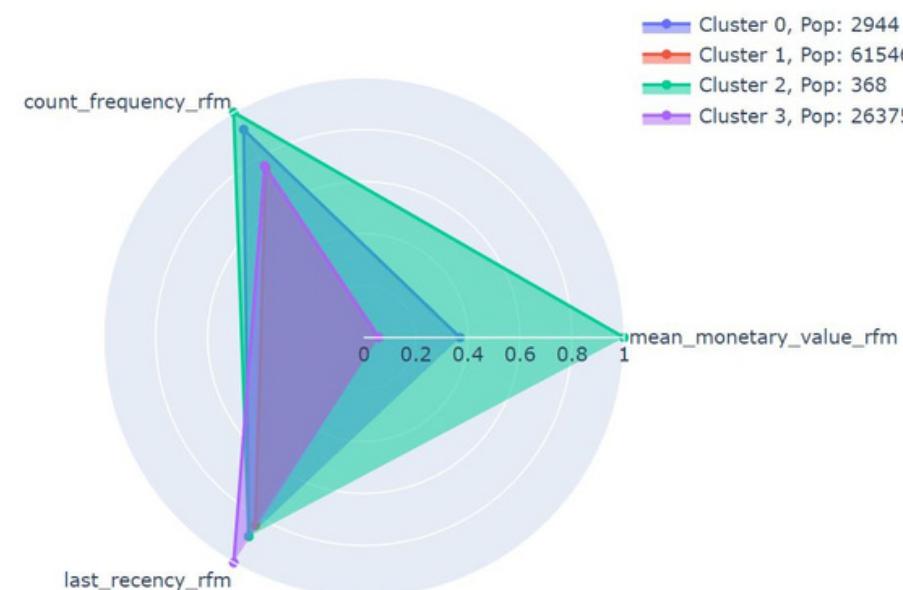
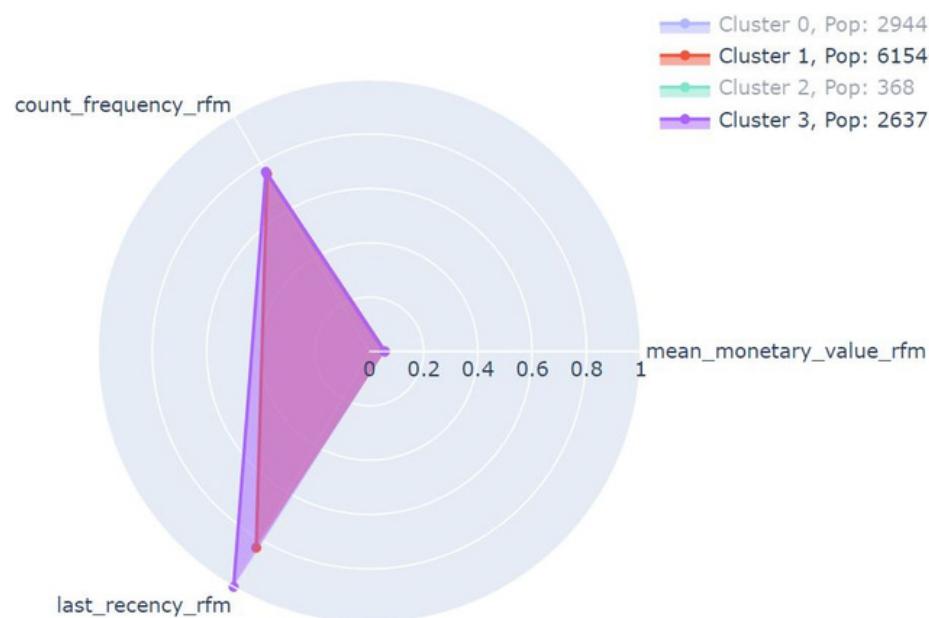
| | SILHOUETTE | CALINSKI-HARABASZ | DAVIES-BOULDIN | RUNNING TIME |
|-----------------|-------------|-------------------|----------------|--------------|
| KMeans | 0.38 | 23201 | 1.36 | 0.33 |
| DBSCAN | 0.09 | 1366 | 1.70 | 28.56 |
| GaussianMixture | 0.65 | 13702 | 0.82 | 1.23 |
| Birch | 0.83 | 6626 | 0.54 | 6.00 |

MODELE FINAL SELECTIONNE

ANALYSE DES CLUSTERS

Use case 1 GaussianMixture

Silhouette : 0.46



Clusters

Cluster 0:

Des clients sont à haute valeur ajoutée. Fréquents et dépensiers.

Cluster 1:

Des clients peu dépensiers et modérément réguliers. Récents,

Cluster 2:

Cluster à rapprocher du cluster 0 mais cluster qui risque d'être ignoré au vu de sa faible taille bien que les clients soient à très haute valeur ajoutée

Cluster 3:

Semblable au cluster 1 avec des clients qui ont effectué un achat encore plus récent.

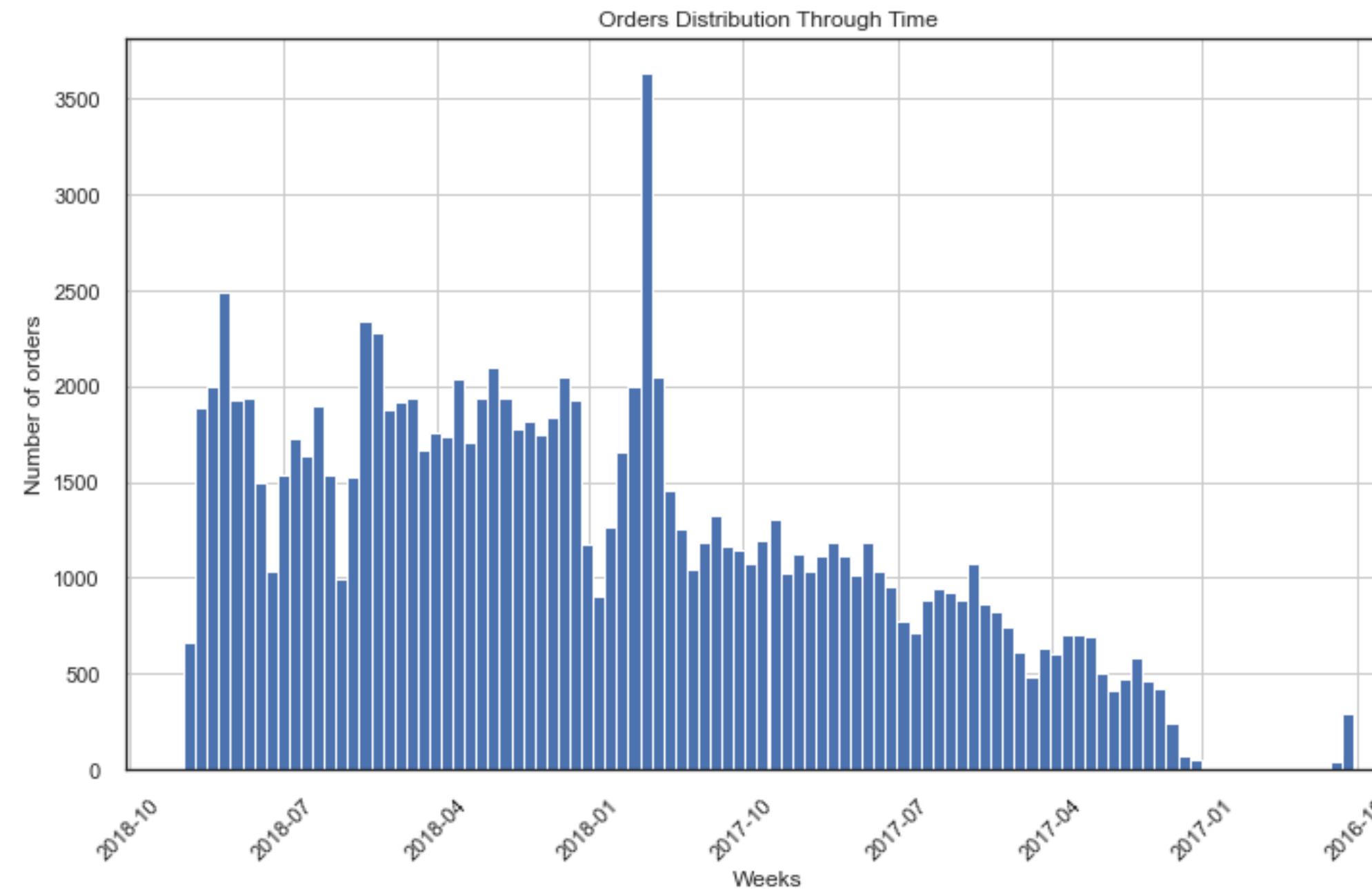
Agrégats calculés en fonction de la moyenne donc sensible aux outliers. La distribution est-elle identique?

- Standard
- Premium
- Gold

Ne pas perdre les nouveaux clients

Simulation

Répartition des commandes par semaines



Période
Oct. 2016 - Août 2018

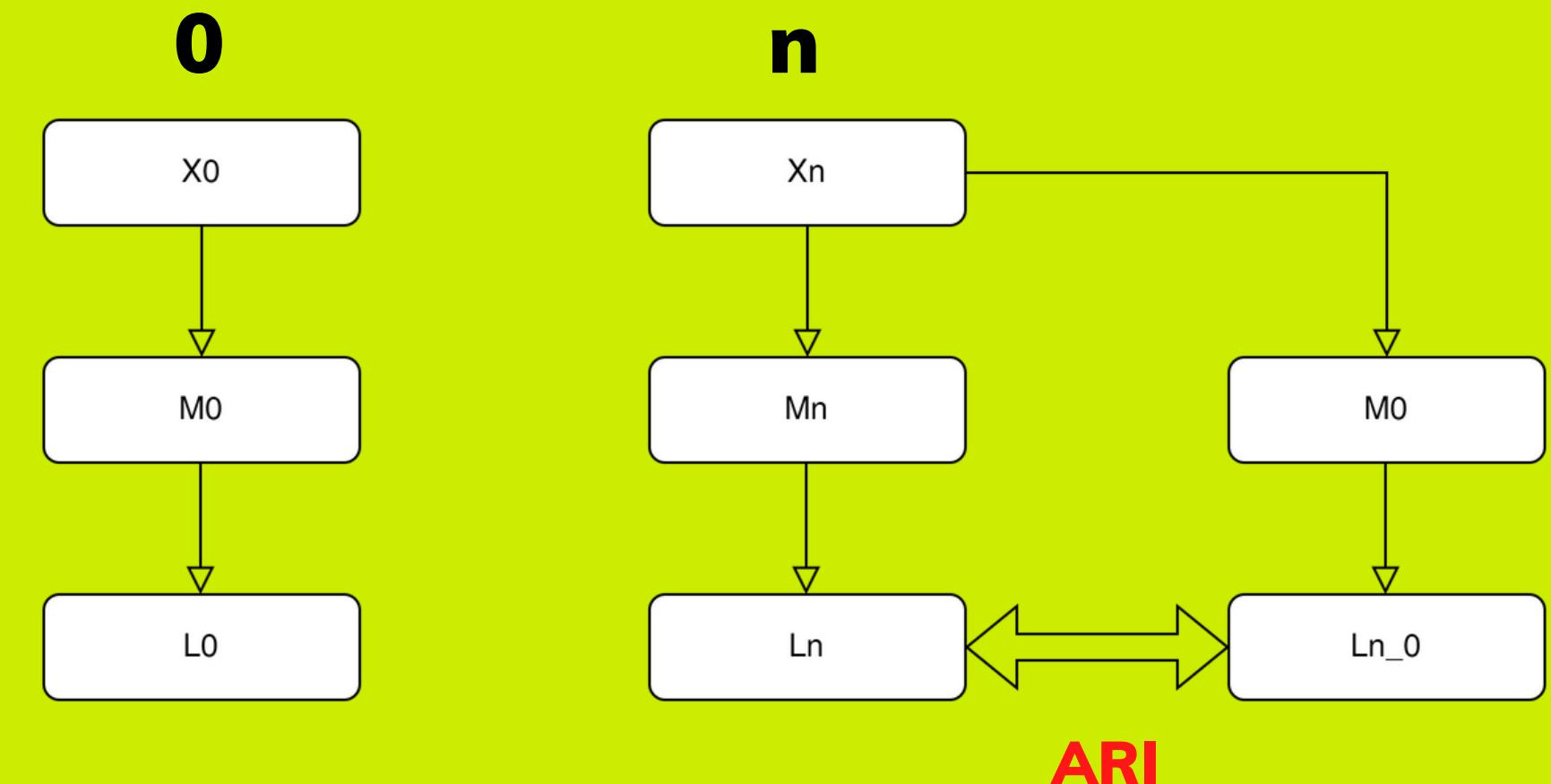
Durée
695 Jours/
100 Semaines/
1.9 Années



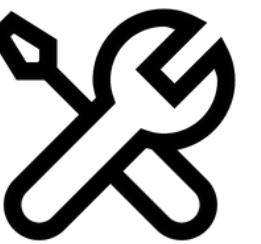
Démarche

- Calcul stabilité des clusters
Fonction rand_score de Sklearn
- A intervalles réguliers de 7 jours
- Position de départ 600/695 jours

Algorithme



Maintenance



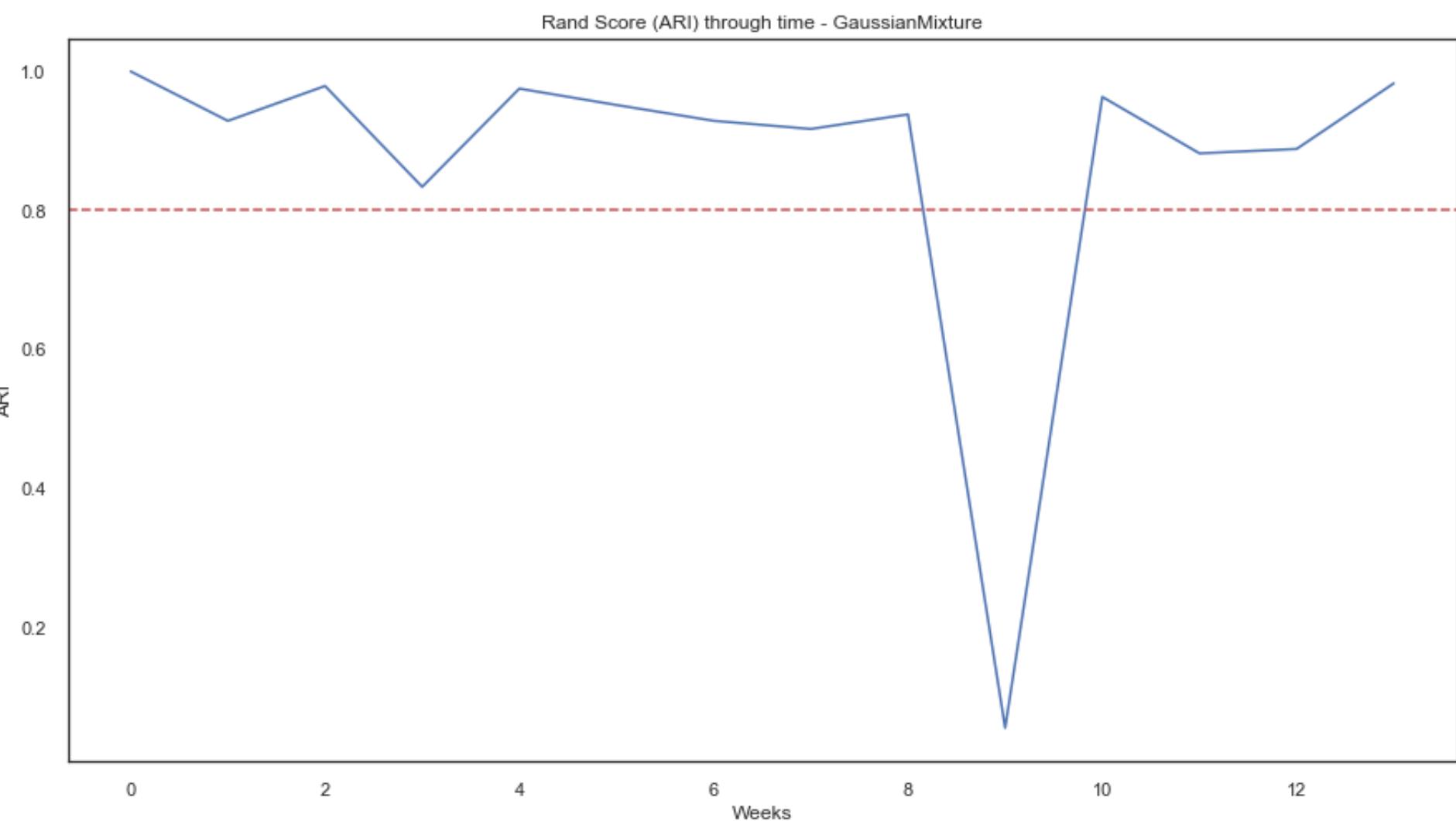
ARI Score GaussianMixture

Par semaine

Début :
600/695 jours

Cible :
 $ARI < 0.80$

Durée avant mise à jour :
8 semaines environ 2 mois





Conclusion

3 Points

Les clients réguliers ont panier moyen plus élevé

Ils représentent cependant des clusters de taille réduite qui risquent de ne pas être pris en compte

La Performance des algorithmes de clustering est fortement dépendante du calcul des agrégats

Ici deux clusters semblables peuvent avoir une répartition différente de leurs valeurs

En effet les agrégats sont calculés en fonction de la moyenne et donc sensible aux outliers

Enfin l'étude du score de ARI semble nous indiquer un besoin de mise à jour du modèle de clustering toutes les 8 semaines



Merci de votre écoute

Séance de questions

P5 : Segmentez des clients d'un site e-commerce

Romain Vaillant Janvier 2022