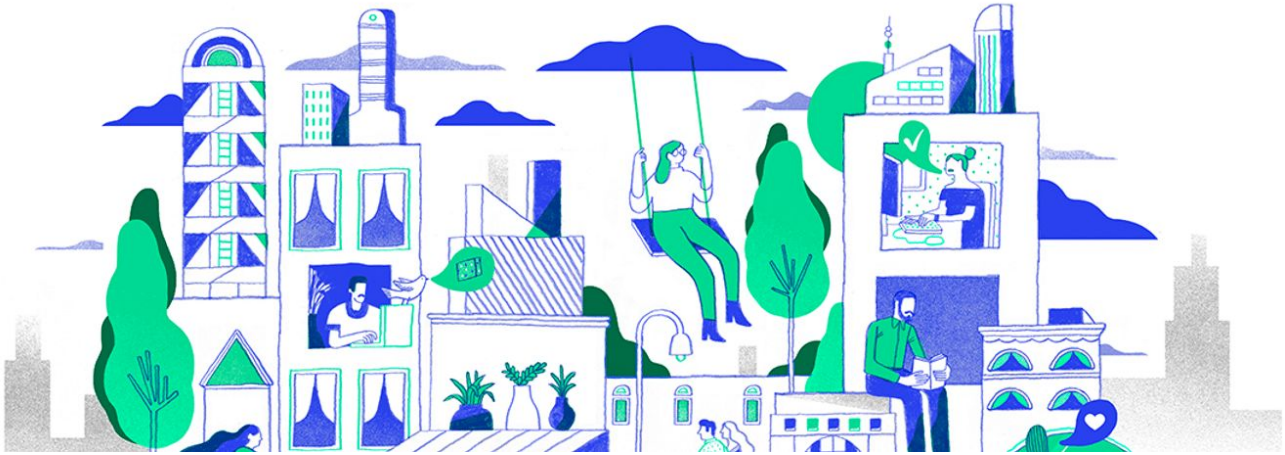
Formation DS-IML PROJET 4 : OLIST Customers Segmentation



olist

empowering commerce

olist
empowering commerce

MISSION

Conseil marketing

Analyse et Segmentation de la clientèle Olist.

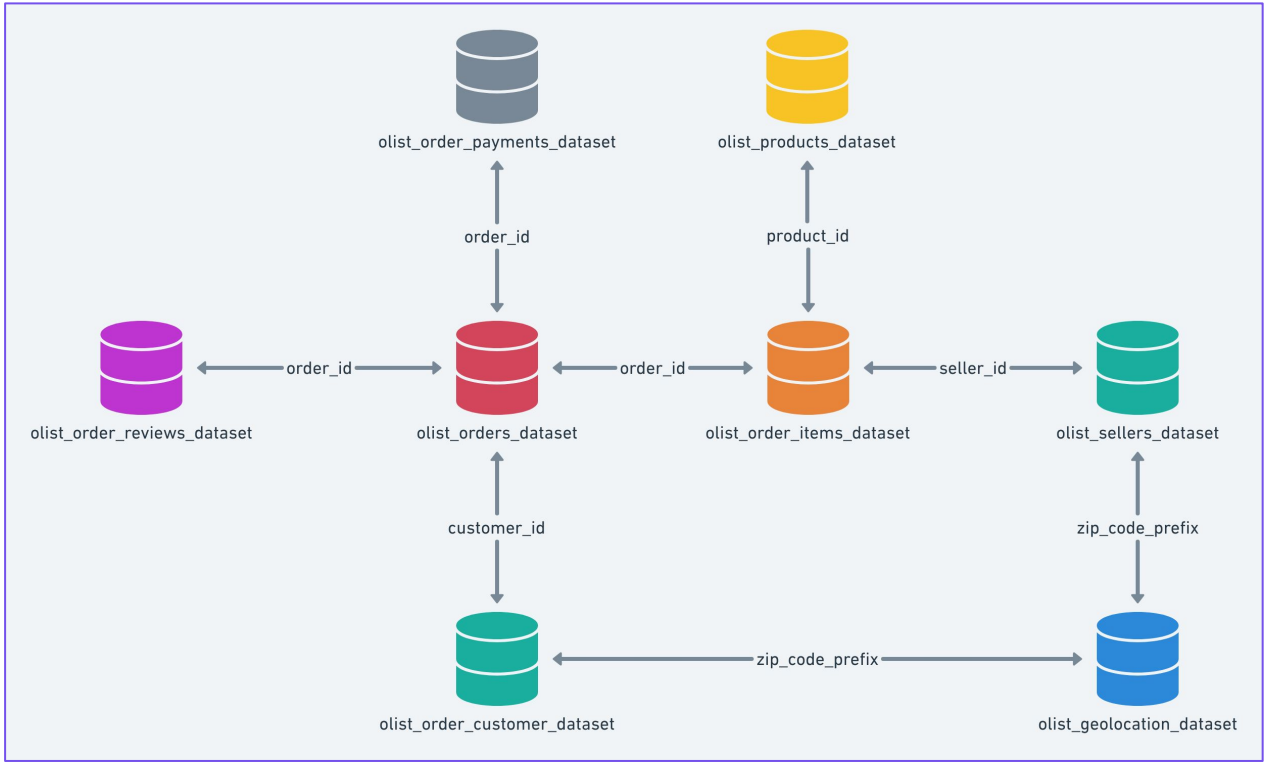
Résultats utilisables et opérationnels

SOMMAIRE

Données à disposition

9 fichiers *.csv

- Clients
- Statut Com.
- Compo. Com.
- Paiements
- Reviews
- Produits
- Revendeurs
- Traduction
- Géo localisation



Après réunion des données

116 581 individus

	customer_unique_id	customer_city	customer_state	order_status	order_purchase_timestamp
0	861eff4711a542e4b93843c6dd7febb0	franca	SP	delivered	2017-05-16 15:05:35
1	9eae34bbd3a474ec5d07949ca7de67c0	santarem	PA	delivered	2017-11-09 00:50:13
2	9eae34bbd3a474ec5d07949ca7de67c0	santarem	PA	delivered	2017-11-09 00:50:13

Client **Commande** **Produit Commandé**

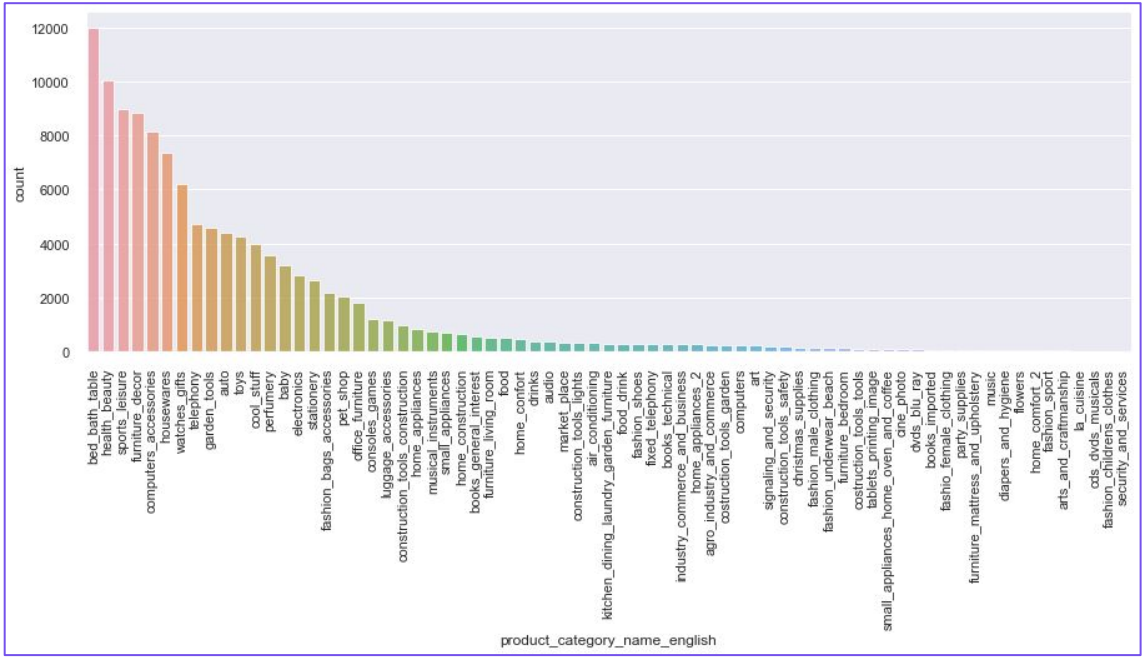
FOCUS CLIENT

40 à 9 colonnes

```
0   customer_unique_id               116581 object
1   customer_state                   116581 object
2   order_id                         116581 object
3   order_purchase_timestamp         116581 object
4   payment_type                     116581 object
5   payment_installments             116581 int64
6   payment_value                    116581 float64
7   review_score                     116581 int64
8   product_category_name_english    116581 object
```

Date d'achat

- Mois
- Semaine du mois
- Jour de la semaine
- Créneau Horaire



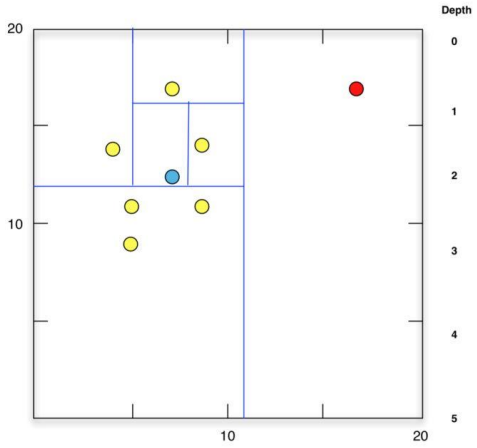
Réduction catégorielles

Catégories de Produits : 18 cat.

Provinces des clients : 6 cat.

Valeurs manquantes : Aucune !

Valeurs extrêmes : Recherche avec Isolation Forest



(a) (b)

Hyper-paramètre de CONTAMINATION

Proportion d'individus les plus "anormaux" devant être considérés comme "outliers"

SOMMAIRE

Pourquoi segmenter des clients ?

Segmentation RFM ?

Une clientèle est variée, identifier des groupes permet de cibler ses actions commerciales.
Procédé classique en marketing. 1er niveau de segmentation sur des critères fondamentaux.

On note les clients sur les critères suivants :

- **R**ECENCY - Date du dernier achat
- **F**REQUENCY - Nombre de commandes passées
- **M**ONETARY VALUE - Montant total dépensé

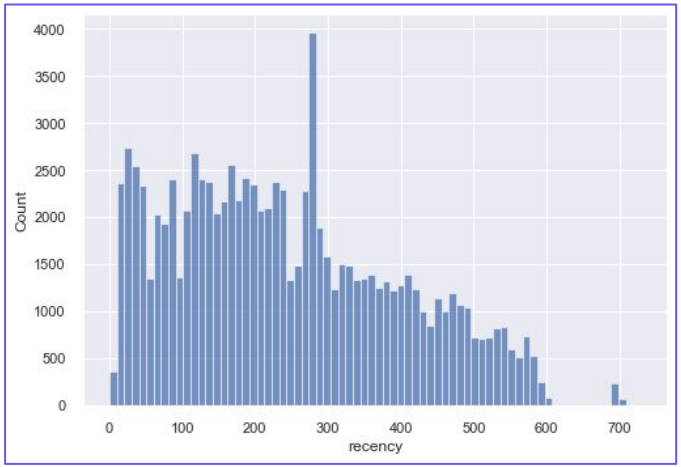
Classe RFM

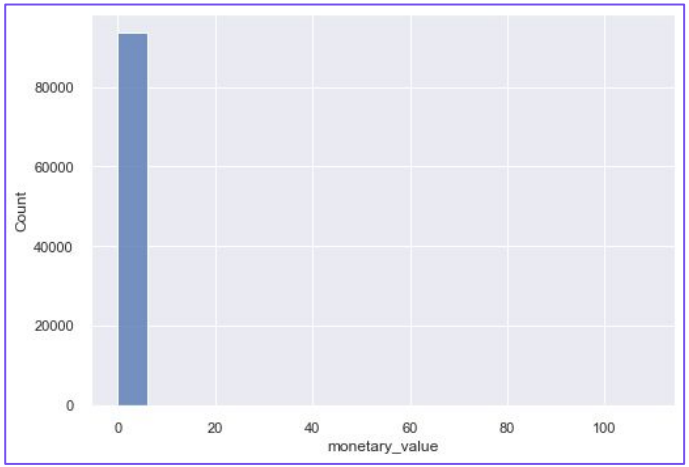
111 333 etc...

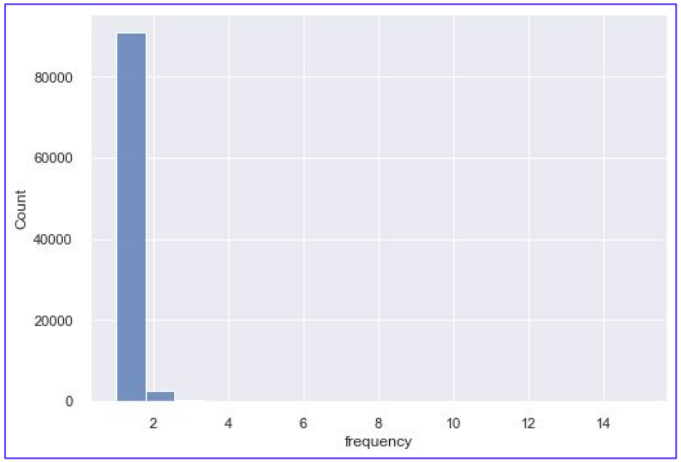
On part d'une Table RFM

Opération sur le dataset

customer_unique_id	recency	frequency	monetary_value
8d50f5eadf50201ccdcedfb9e2ac8455	14	15	820.15
3e43e6105506432c953e165fb2acf44c	188	9	1963.58
ca77025e7201e3b30c44b472ff346268	94	7	2126.44







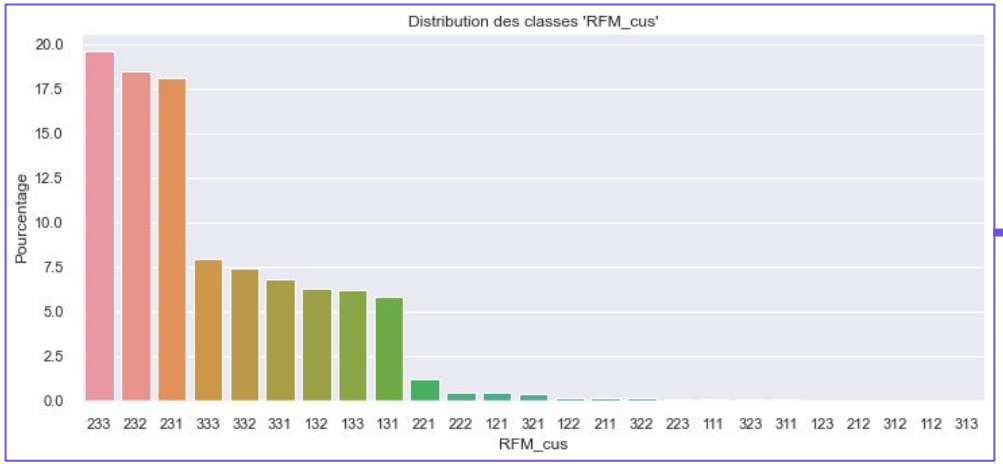
1	91238
2	2617
3	185
4	30
5	9
7	3
6	3
15	1
9	1

Chantier prioritaire : FIDÉLISATION CLIENT

Changement de plan

RFM “ad-hoc” sur 3 groupes

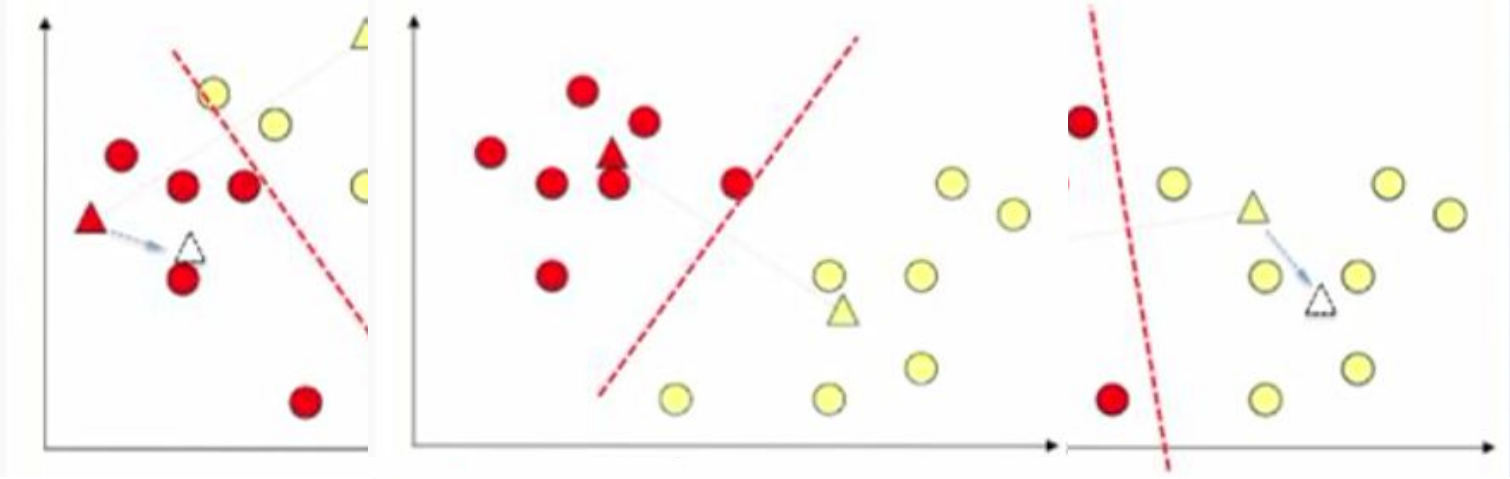
Montant dépensé :	**Quantiles 2/3 et 1/3.**		
Fréquence :	**3+ commandes**	**2 commandes**	**1 commande**
Récence :	**- de 90 jours**	**90 à 365 jours**	**+ de 365 jours**



27 classes en tout.
La majorité peu représentée

Machine Learning | Supervisé | Renforcé | Non Supervisé

K-Means Segmente des données numériques en K groupes

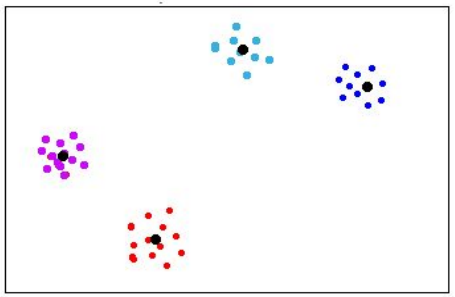


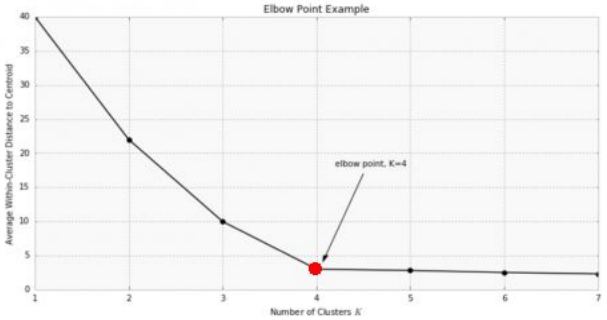
1- Calculs de plusieurs k-segmentations

2- Choix basé que plusieurs éléments...

Méthode du coude

cost = somme des variances des clusters





Elbow method

Silhouette Score

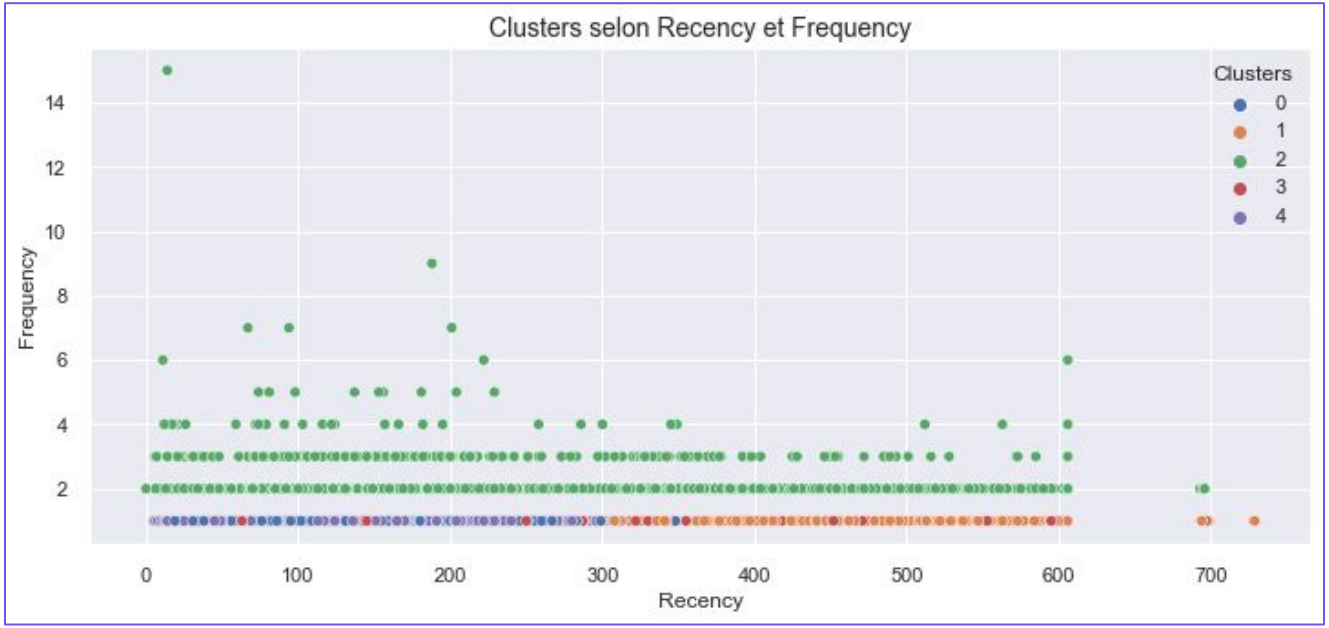
Note de “qualité” de la segmentation

Observation des clusters

Sens, interprétation...

F et M "passées au log" & standardisation

K = 5



Cela nous aide-t-il ?

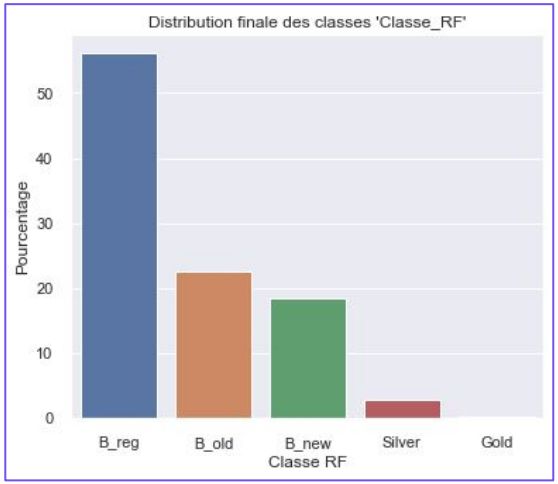
NON

1 - Un cluster regroupe les clients à au moins deux achats.

2 - Les autres sont partagés sur R et M

Objectif : augmentez fréquence/fidélité

Gold	F > 2
Silver	F = 2
B new	F = 1 & R <90
B reg	F = 1 & 90< R <365
B old	F = 1 & 365< R



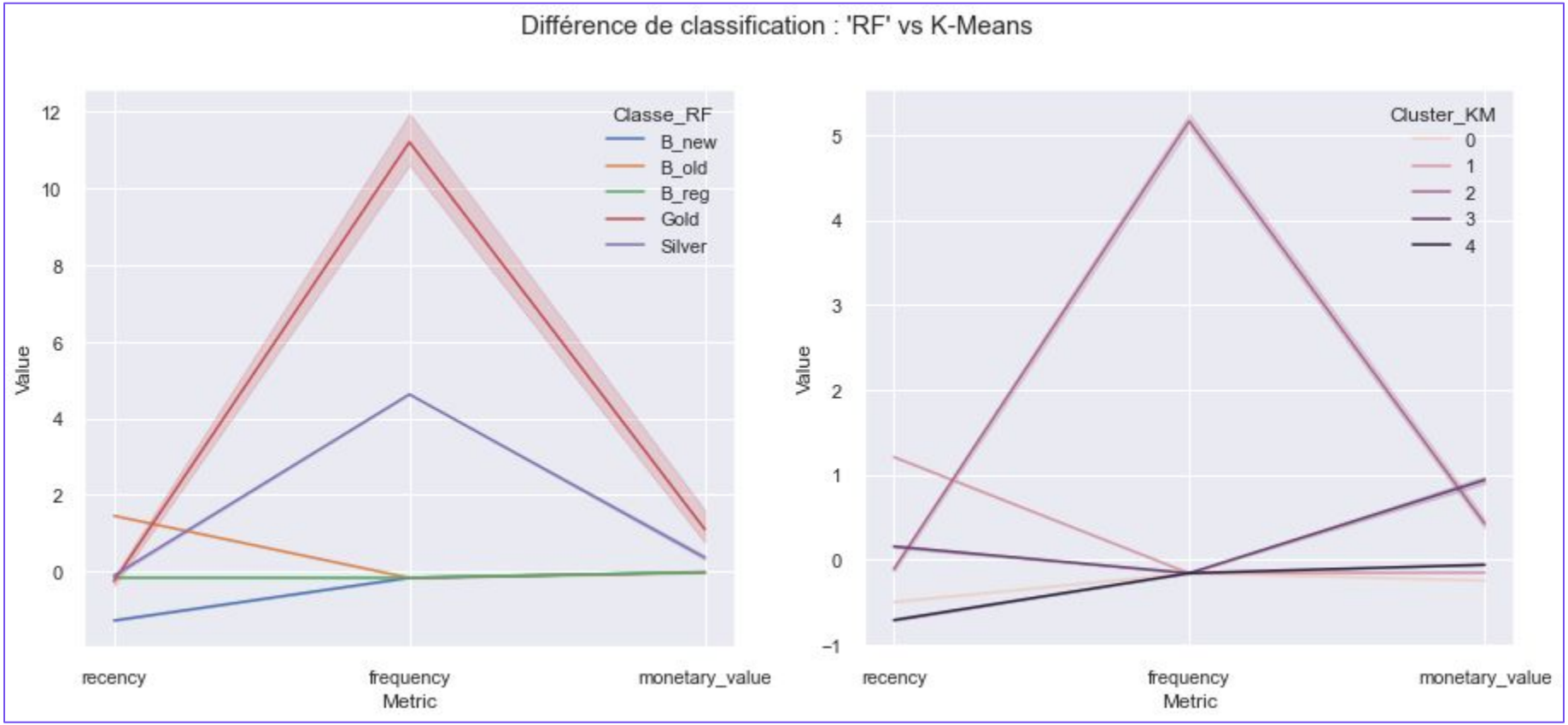
Globale :

Mécanisme de fidélisation

Ciblée :

Communiquer avec le client en fonction de leur classe “RF”.

Snake Plots


Différence de classification : 'RF' vs K-Means
Classe_RF
B_new
B_old
B_reg
Gold
Silver
Value
12
10
8
6
4
2
0
recency
frequency
monetary_value
Metric
Cluster_KM
0
1
2
3
4
5
4
3
2
1
0
−1
recency
frequency
monetary_value
Metric

Stacked Area Plot

Percentage Stacked Area Plot

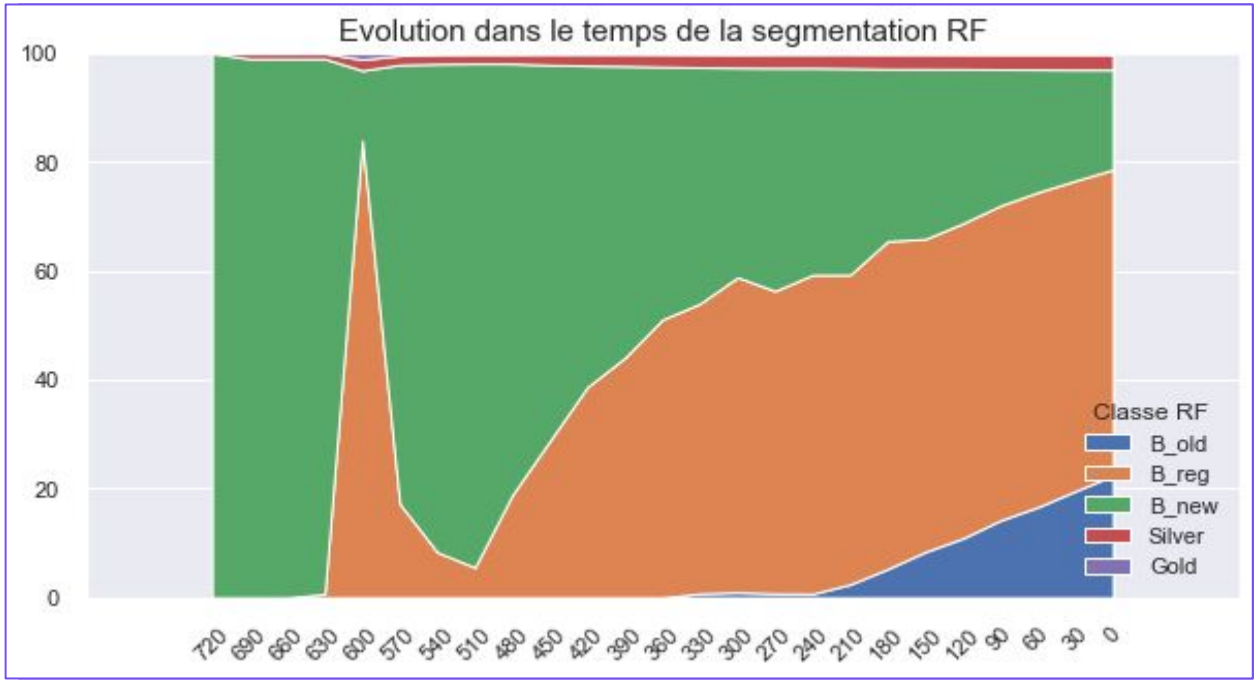
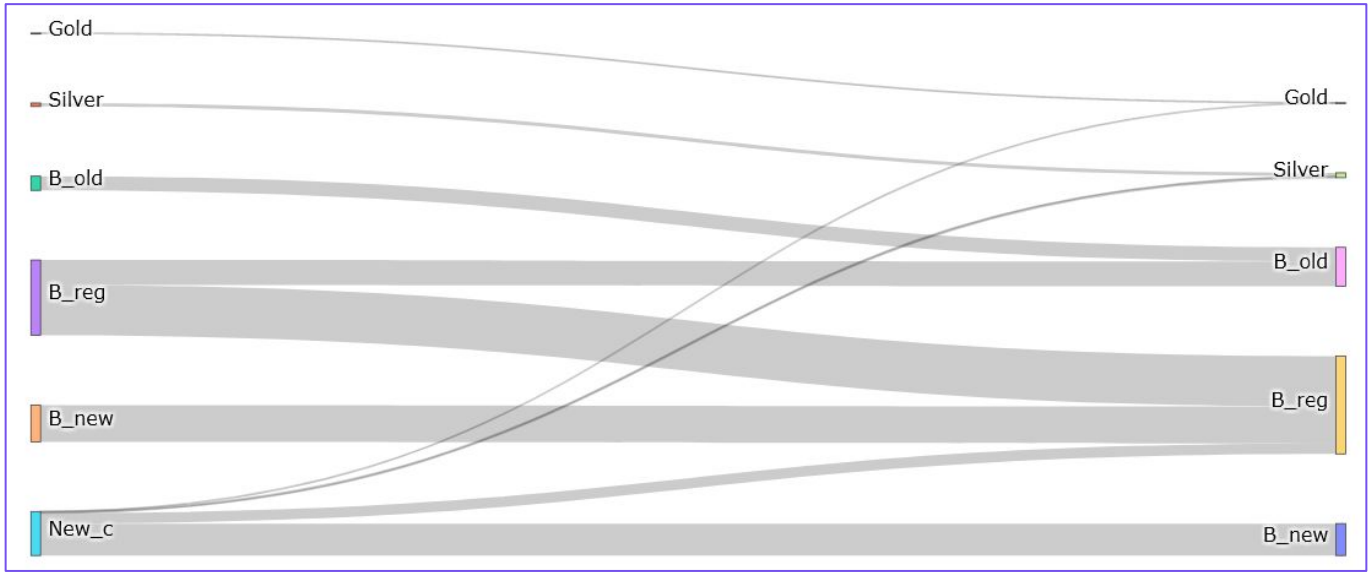


Diagramme Sankey | Flux entre classes sur une période données



Nous fournirons les listes des clients changeant de classe...

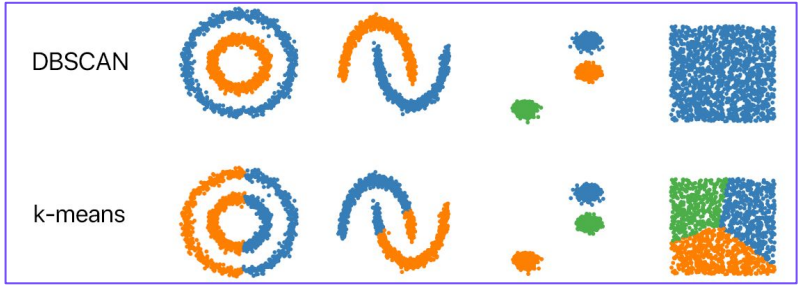
SOMMAIRE

ML non-supervisé

- Découverte de groupes pertinents
- Construction de profils types

Algorithmes

K-Means	Data numériques
K-Modes	Data catégorielles
K-Prototypes	Mixed data
DBSCAN	Densité de points



Principe Simple

Pas d’assurance de résultat

1 - Cadre de recherche

2 - Algo.

3 - Observation

Acte d'achat : M_class / paiement / product_cat

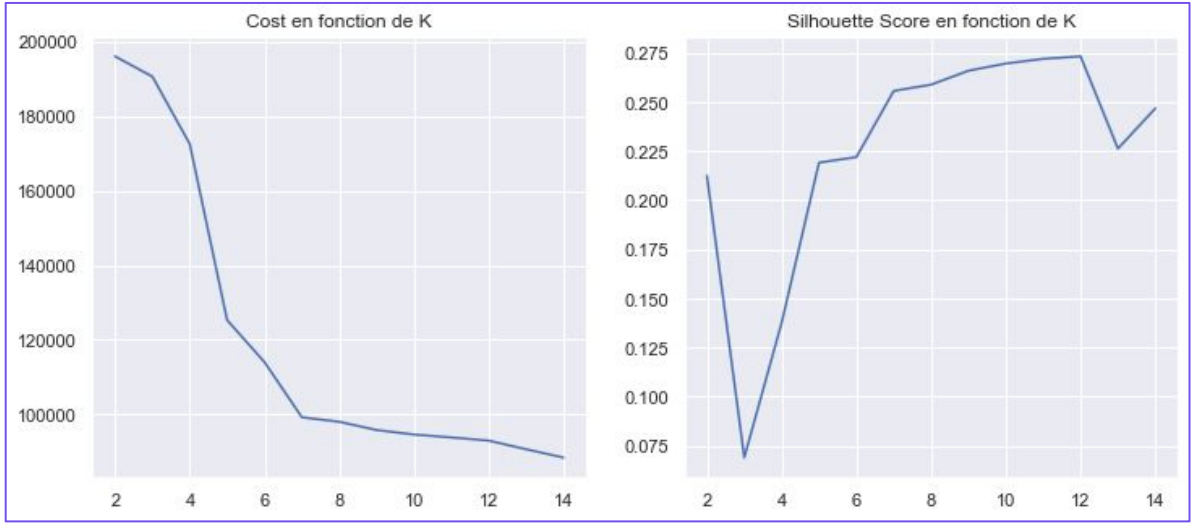
Que des variables catégorielles : K-Modes

Choix de K

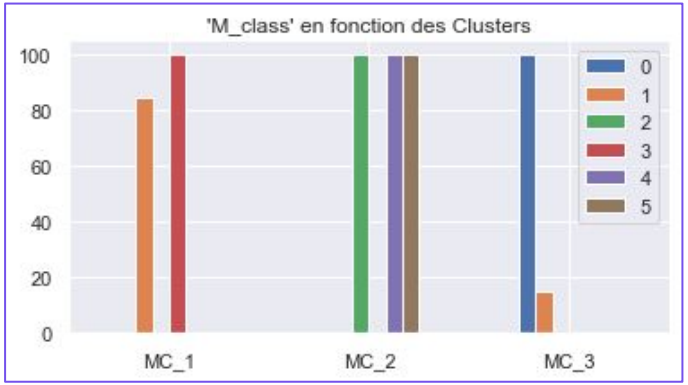
15 minutes

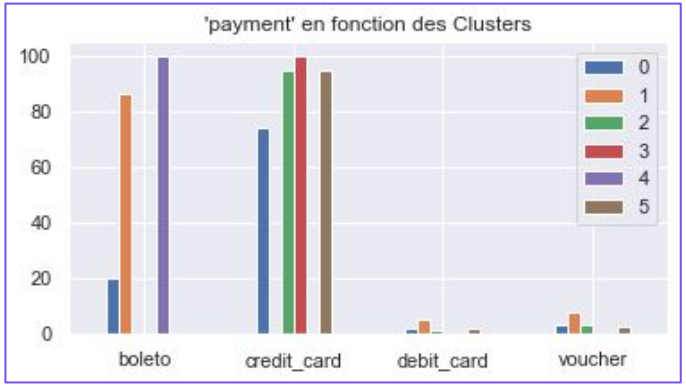
ualité
diocre

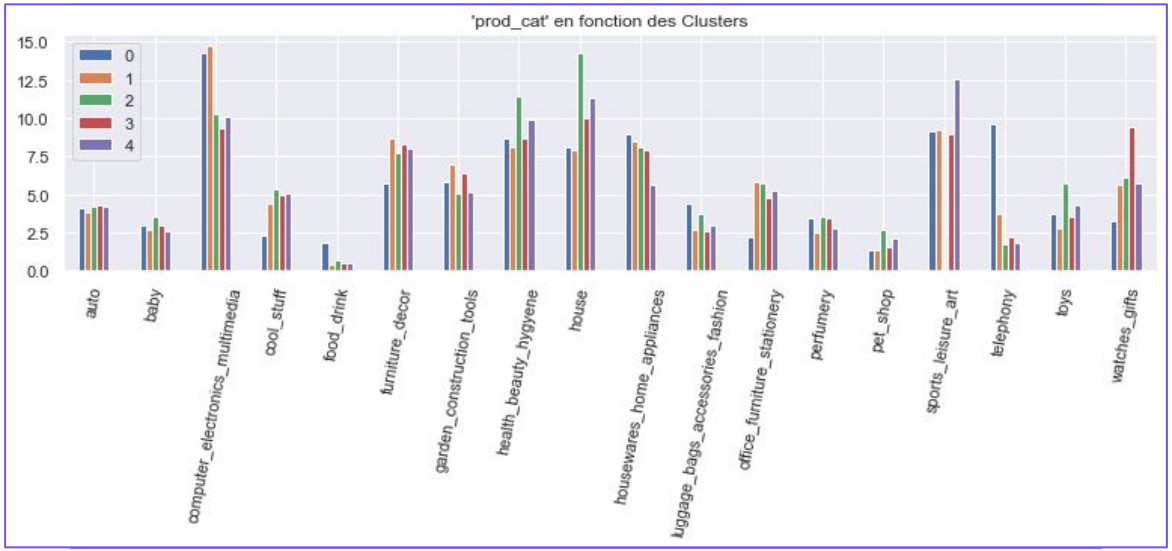
K = 6



0	30848
3	24729
2	22320
1	7468
4	6009
5	2713

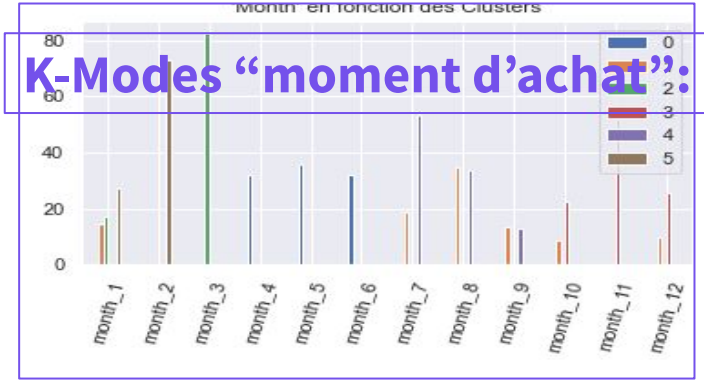


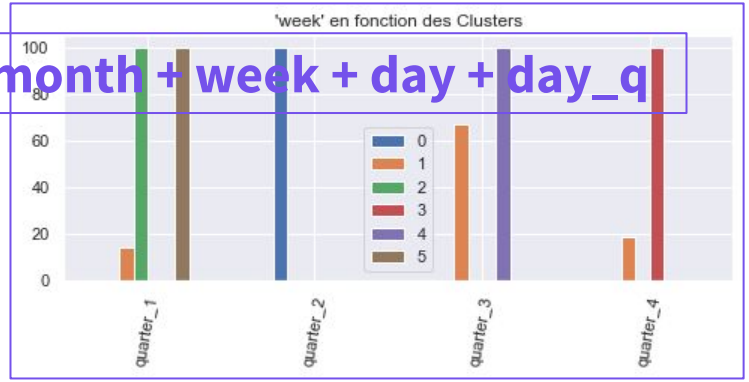


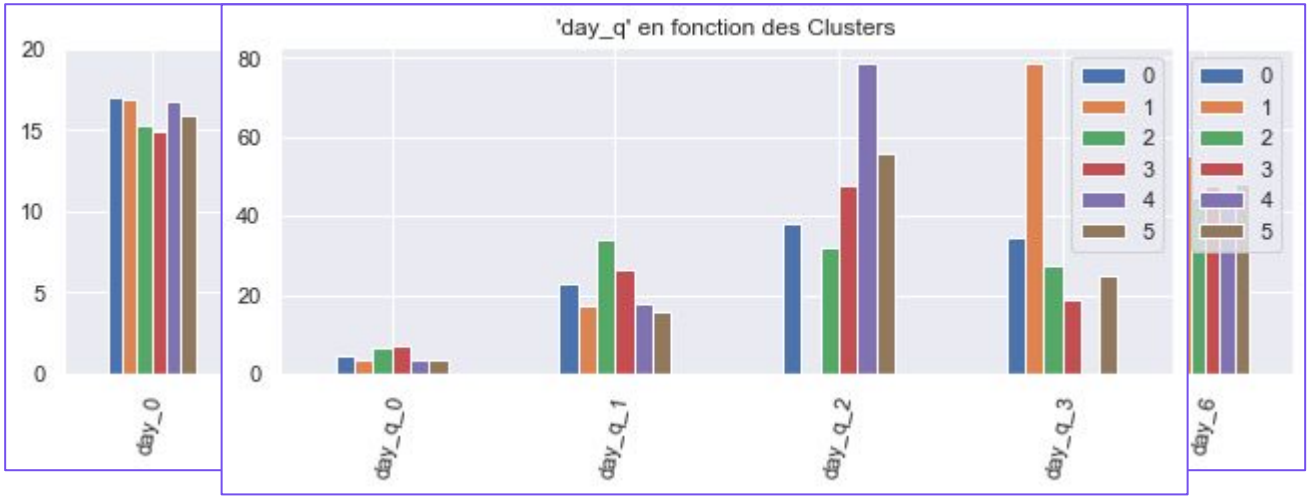


0	27993
1	18203
3	13670
4	11990
2	11321
5	10910

K-Modes "moment d'achat": month + week + day + day_q







Intégration facile et rapide

	customer_unique_id	cl_achat	cl_moment
0	861eff4711a542e4b93843c6dd7febb0	2	0
1	9eae34bbd3a474ec5d07949ca7de67c0	0	0
2	9eae34bbd3a474ec5d07949ca7de67c0	2	2

états civil

analy-tics

(...)

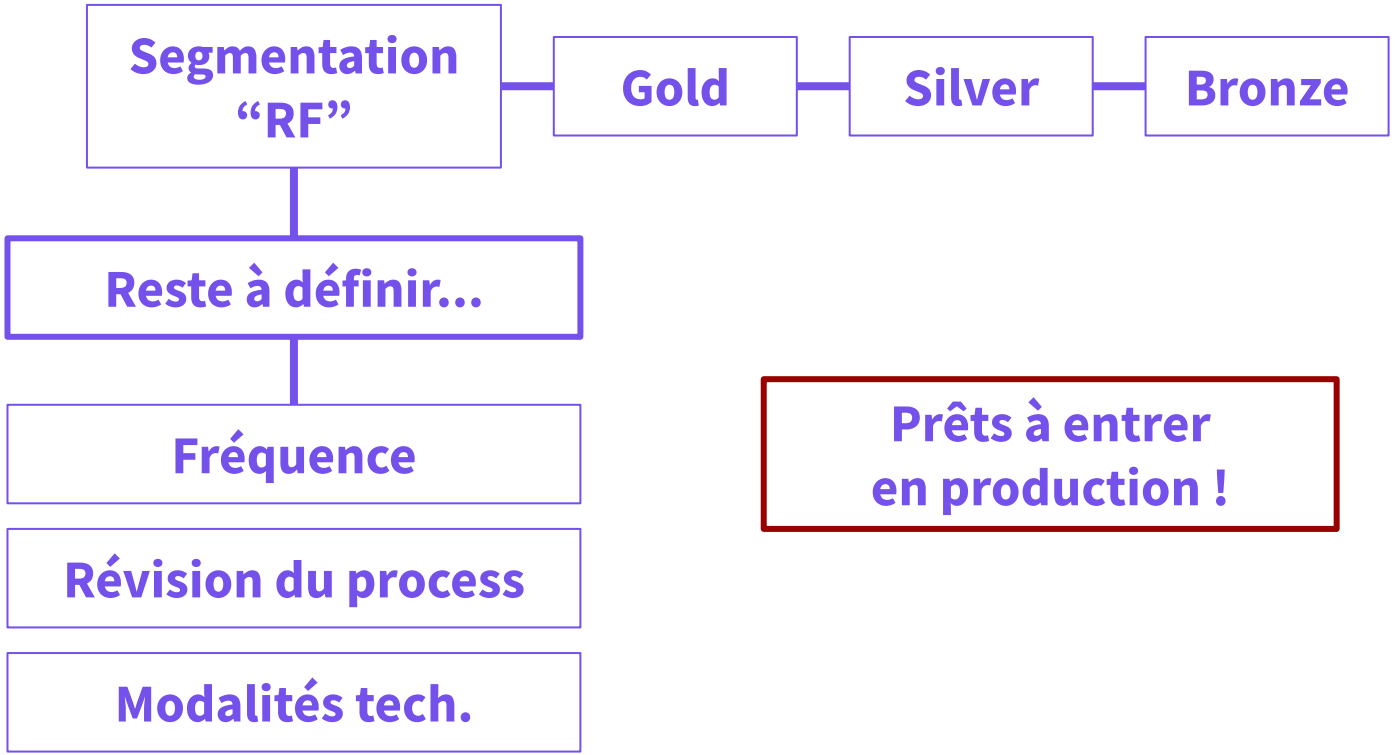
Plusieurs approches possible

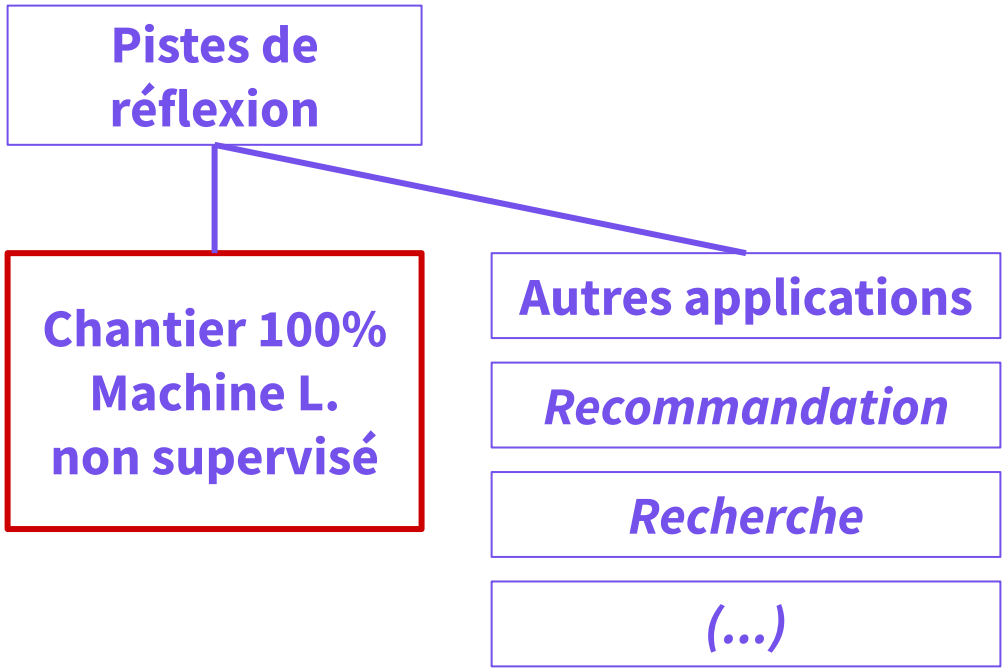
Utiliser toutes les données à disposition

STRATEGIE TEST

Clients GOLD

SOMMAIRE


Segmentation “RF”
Gold
Silver
Bronze
Reste à définir...
Fréquence
Révision du process
Modalités tech.
Prêts à entrer en production !


Pistes de réflexion
Chantier 100% Machine L. non supervisé
Autres applications
Recommandation
Recherche
(...)