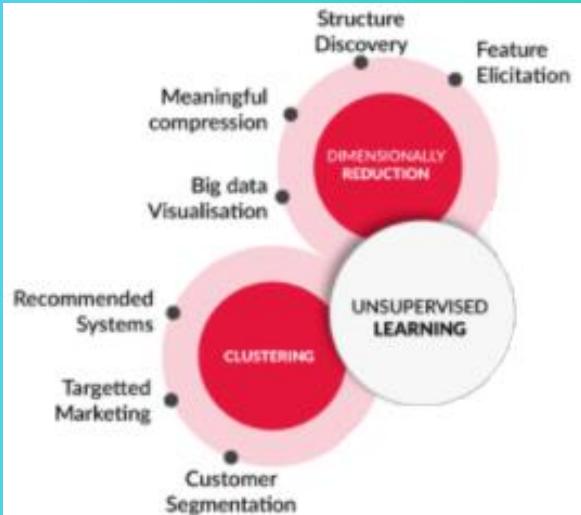


# Segmentez des clients d'un site e-commerce



## CLUSTERING



## SEGMENTATION



-  1. Problématique
-  2. Données
-  3. Modélisations
-  4. Conclusions



Problématique



Données



Modélisations



Conclusions

## Mission :

**Fournir** aux équipes Marketing de l'entreprise Olist (site de e-commerce) une **segmentation des clients** utilisables dans leurs campagnes de communication.



## Objectifs :



1. **Comprendre** les différents **types d'utilisateurs** (comportements, données personnelles)
2. **Fournir** une **description actionnable** de la segmentation avec une logique sous-jacente pour une optimisation optimale.
3. **Proposer** un **contrat de maintenance** basé sur une analyse de la stabilité des segments au cours du temps.

## Olist :

Plateforme vitrine de e-commerce au Brésil (2016).  
Met en lien acheteurs et vendeurs.  
Commande, paiement, suivi de livraison.  
Notation et avis sur la commande.



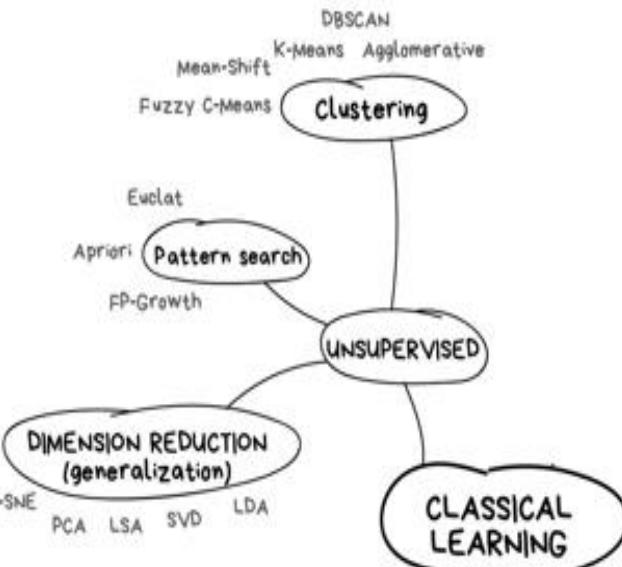
## Segmentation :

**Problème de marketing traditionnel segmentation RFM**

**Problème de classification non supervisée**



- Analyse centrée clients, 9 jeux de données :  
Fusion? Variables pertinentes?  
Feature engineering?
- Algorithmes ? Hyperparamètres? Métriques?



## Compréhension des clients :

- Interprétation des segments :  
Critères? Métriques adaptées?
- Description des actions à réaliser?

## Contrat de maintenance :

- Comment évaluer la stabilité?
- Maintenance sur x mois, ajout de clients?

## MÉTIER

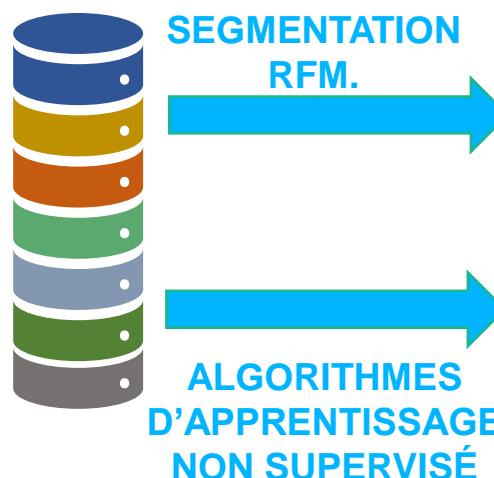
- 9 jeux de données séparés
- Plusieurs lignes par client
- Population hétérogène



## ANALYSE NETTOYAGE

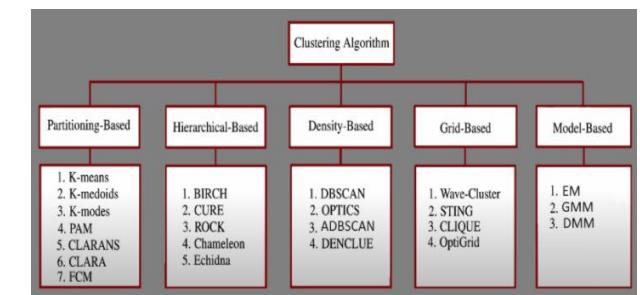
- 1 jeu de données assemblé
- Une seule ligne par client
- Indicateurs clients :
  - comportementaux
  - fréquence, valeur, récence

Fusion  
Nouvelles variables  
Standardisation  
Encodage  
Transfo. Log.



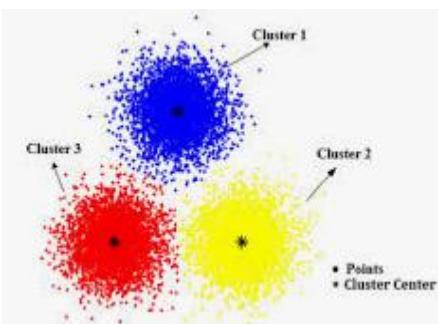
## SEGMENTATION

- Recherche de liens cachés
- sous-populations homogènes
- analyse multi-dimensionnelle
- Modèle applicable sur de nouveaux clients



## INTERPRÉTATION

- Connaissance client améliorée
- Segmentation intelligente
- Analyse descriptive des profils



## ANALYSE

## FEATURE ENGINEERING

## MODÉLISATION

## RAPPORT



Problématique



Données

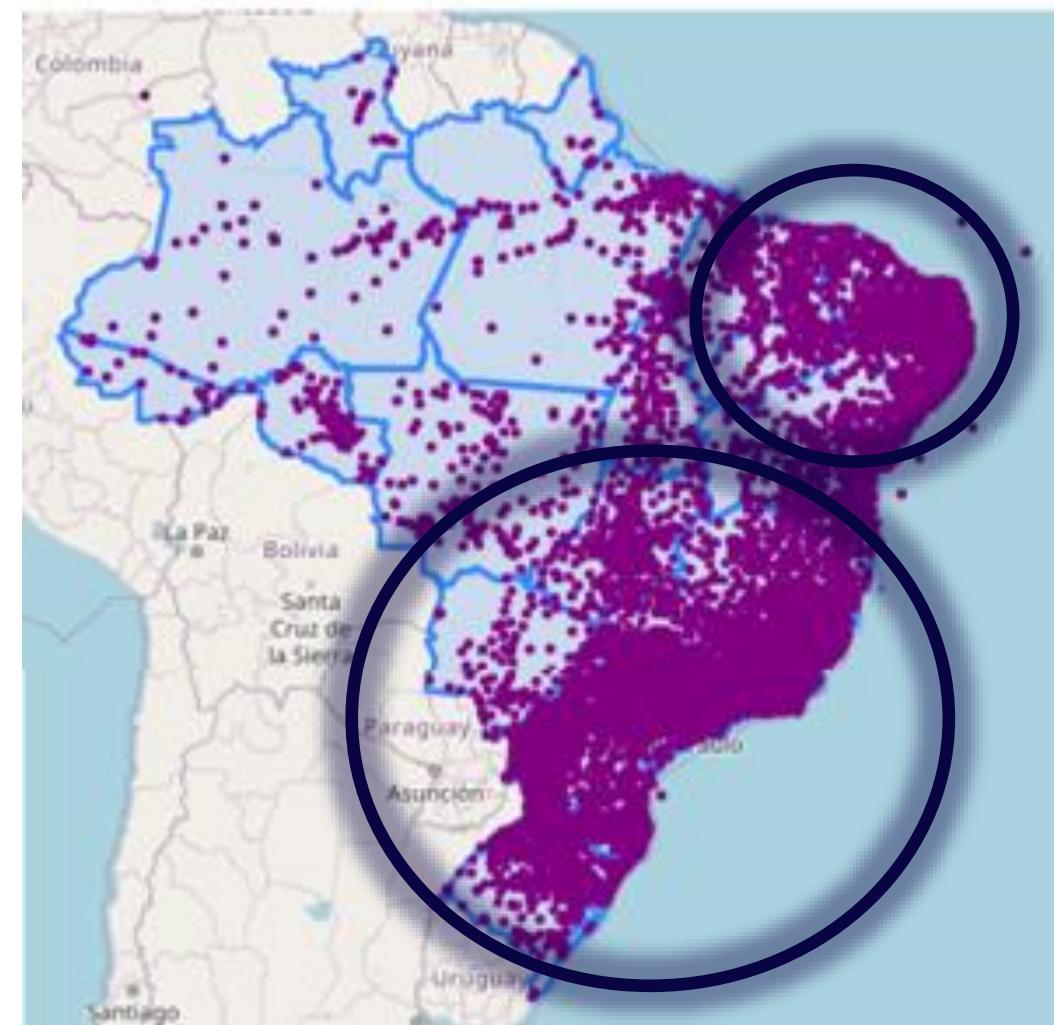
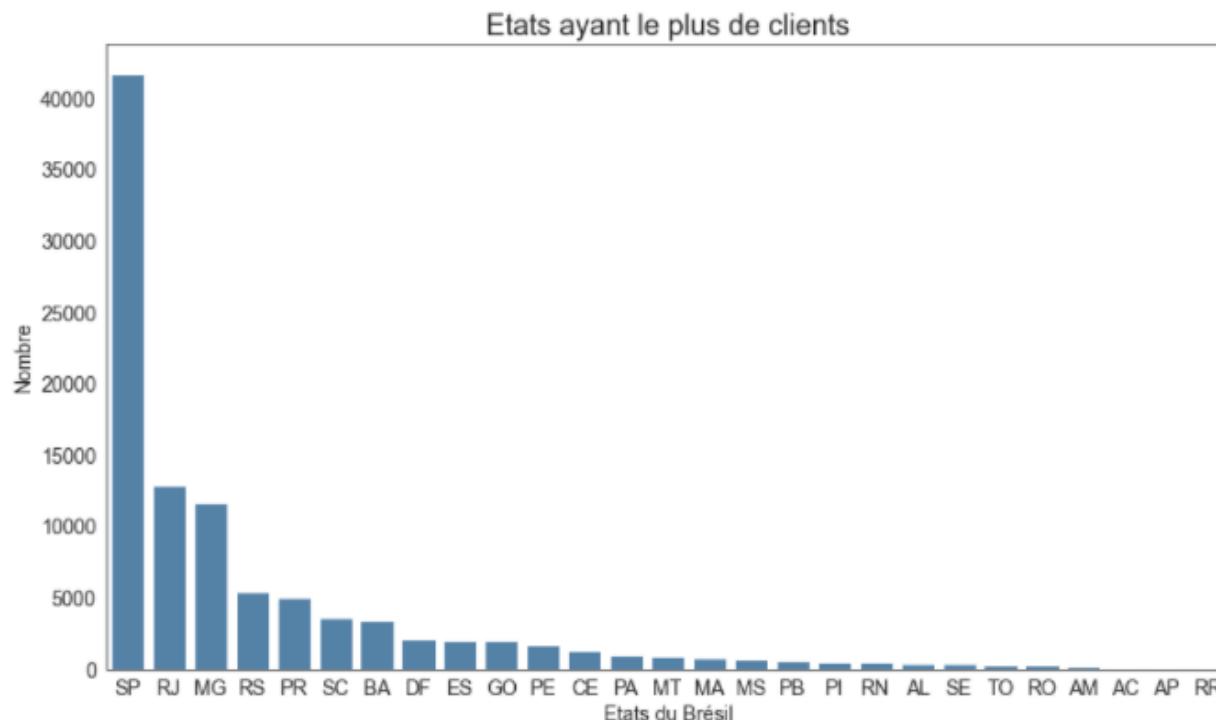


Modélisations



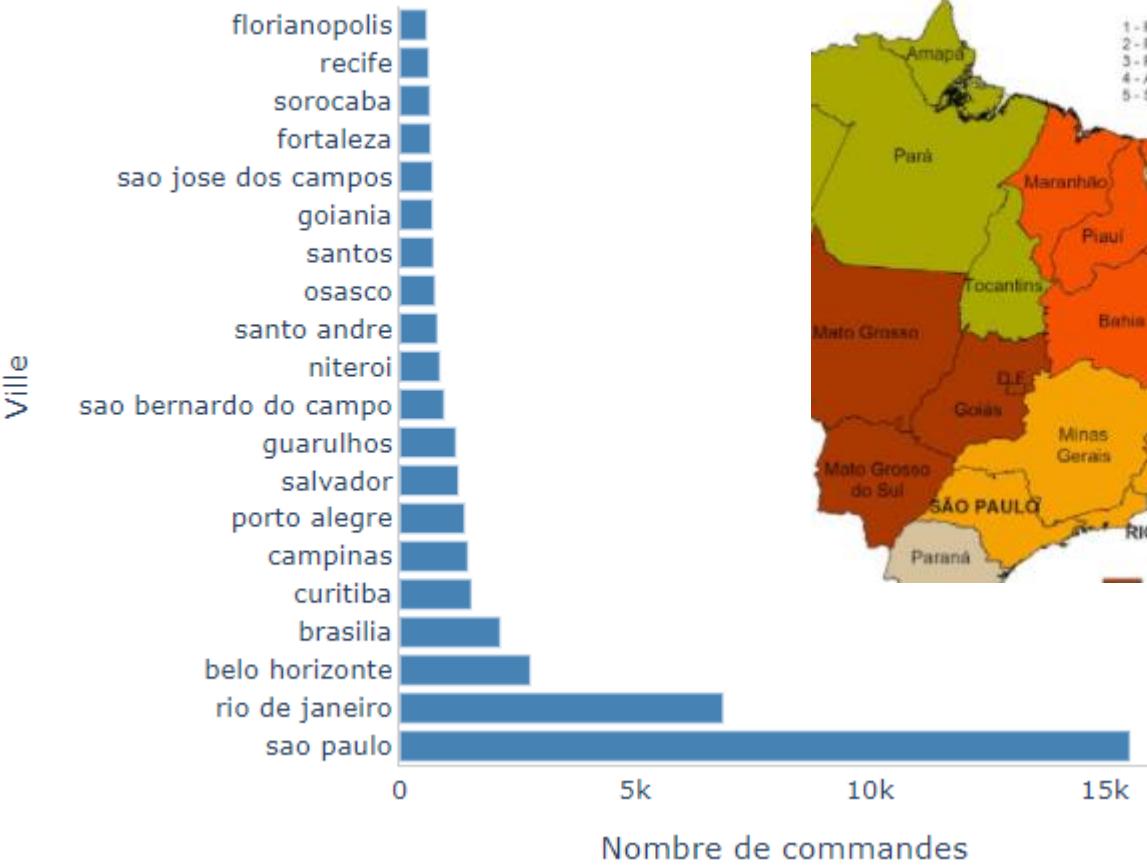
Conclusions

## Localisation des clients



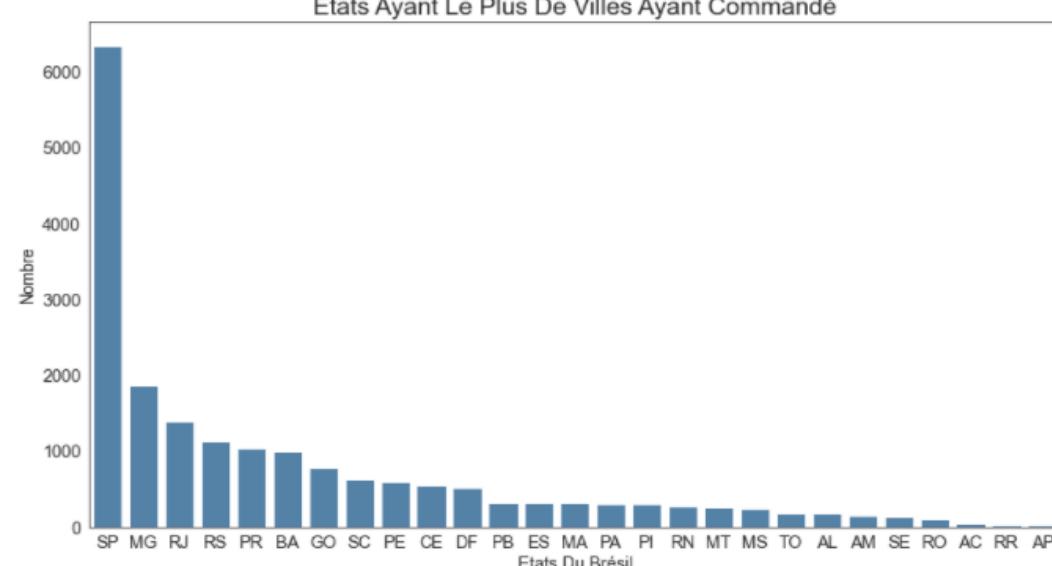
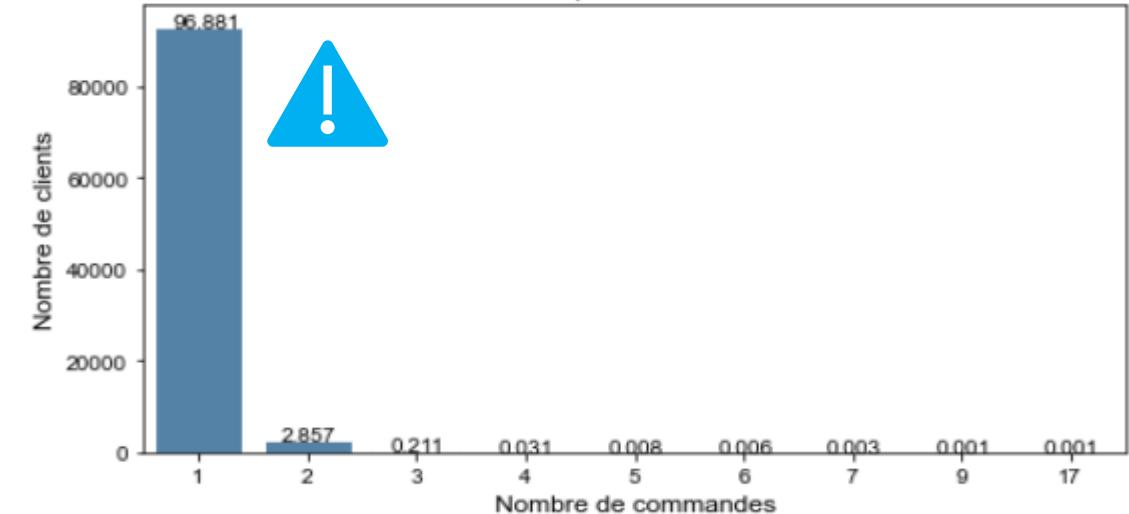
## Commandes

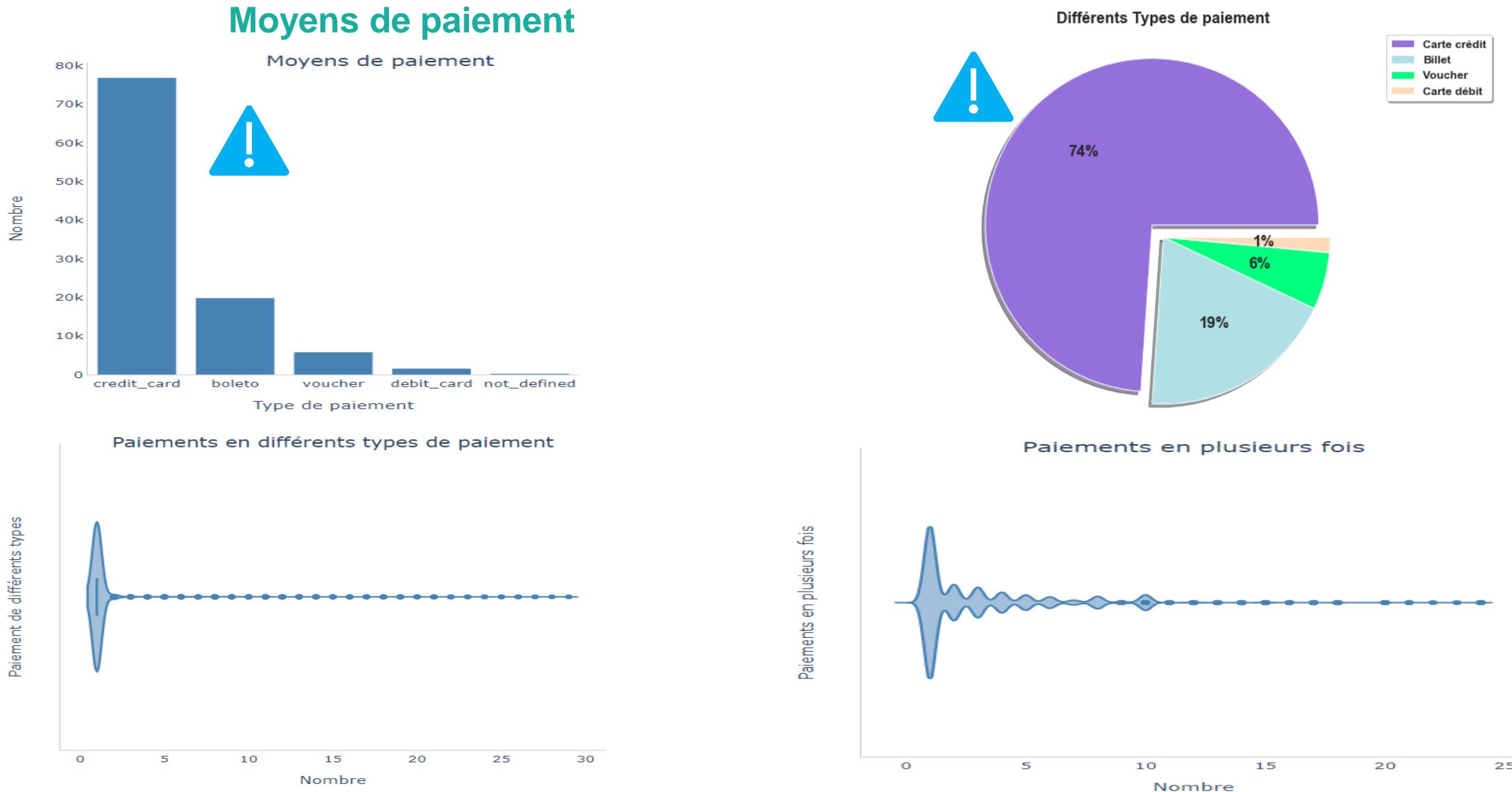
Top 20 des villes ayant le plus de client



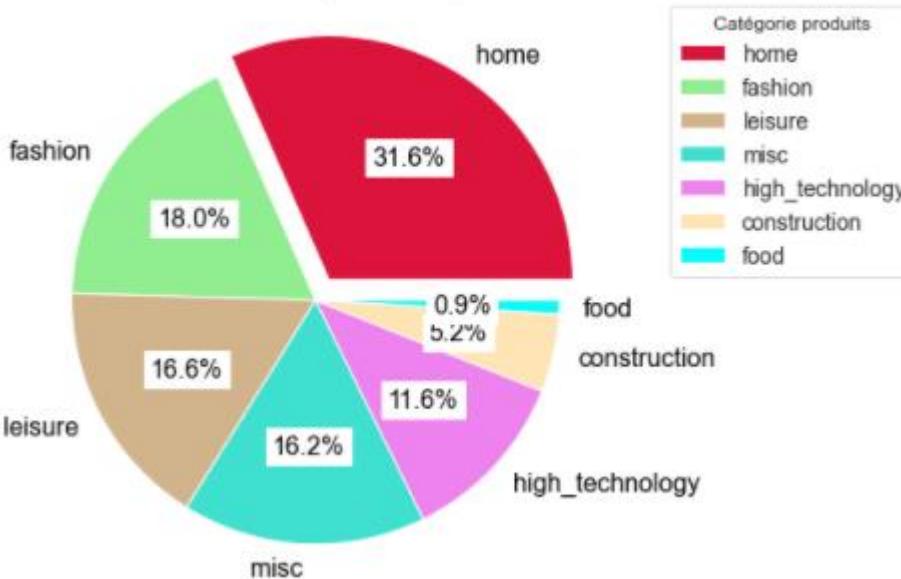
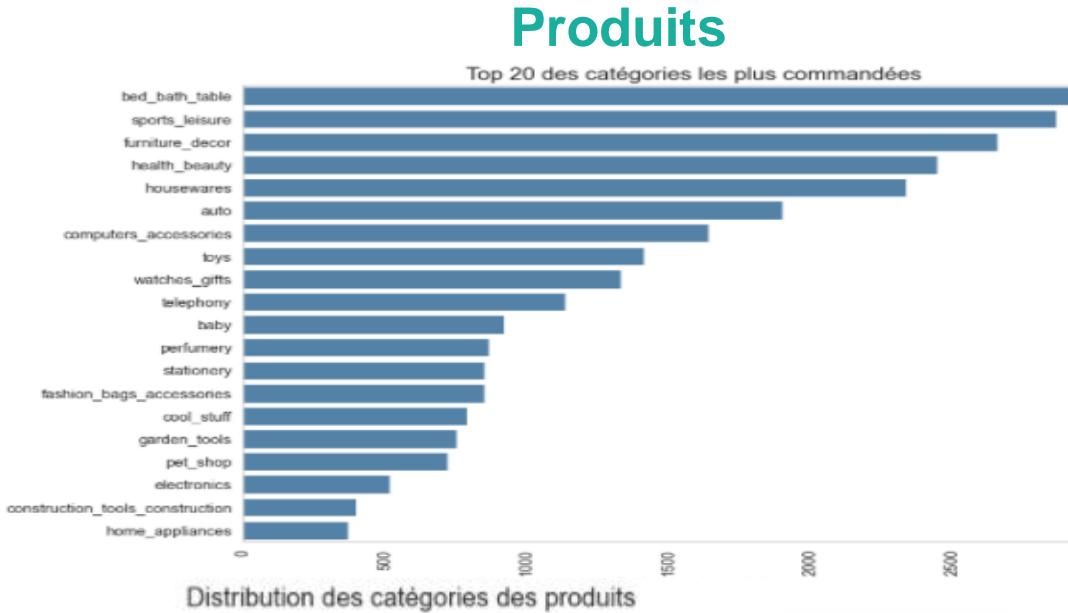
## Fréquence des commandes

Nombre de clients par nombre de commandes

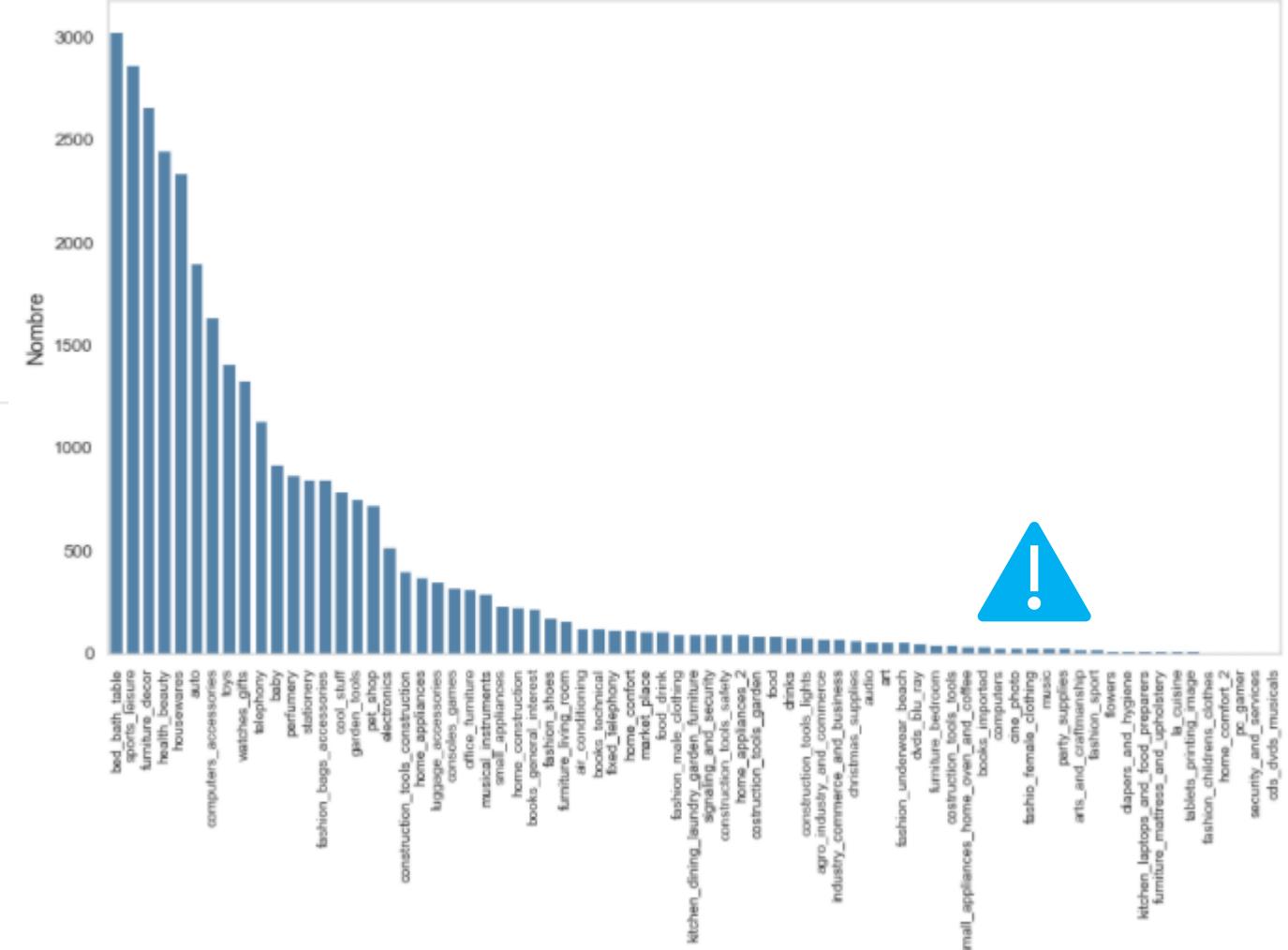




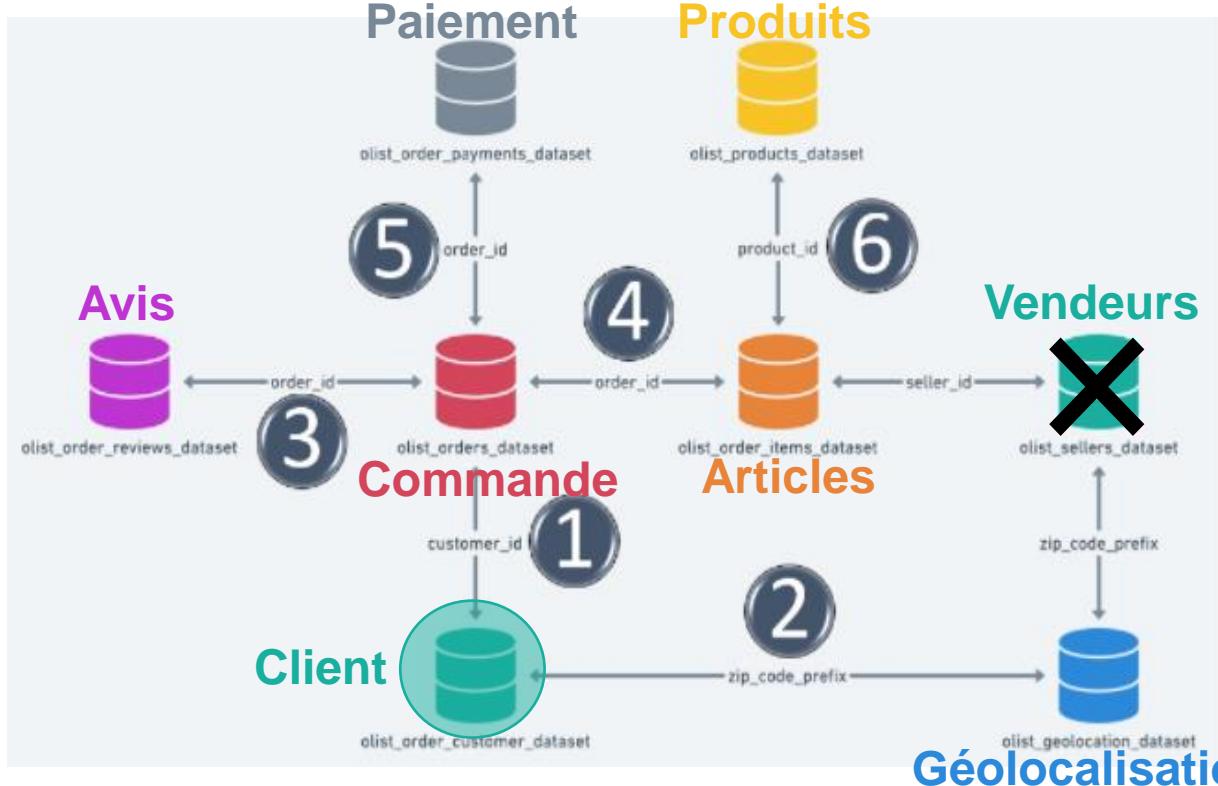
Nombre



Catégories des produits les plus commandés



## 1- FUSION - Clés et ordre d'assemblage



## 2- Nettoyage

Client/géolocalisation - **code postal** : int64 + suppr. '0'.  
 Géolocalisation : **moyenne latitudes/longitudes** par code postal.  
**Suppressions des variables inutiles** après fusion ou à l'analyse.  
 Transformer variable **date** de object en datetime.  
 Traduction de la **catégorie** de produits de portugais en **anglais** + imputation valeurs manquantes manuellement.  
**Valeurs manquantes** : peu → dropna().  
**Filtres** par commande livrée.  
**Valeurs aberrantes** : contrôle des différentes dates.

## 3- Feature Engineering

**Catégorie de produits** : de 73 à 7 catégories (cf annexe A).  
 Création de **nouvelles variables** (par aggrégation, cf annexe B).

customer_id	geolocation_zip	geo_relevance	geo_frequency	geo_neighborhood	date_purchased_mean	percentile_25th_mean	date_delivery_mean	delivery_mean	delivery_mean_25th	order_mean	order_mean_25th	order_mean_percentile_25th	order_mean_percentile_75th	order_mean_percentile_100th
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

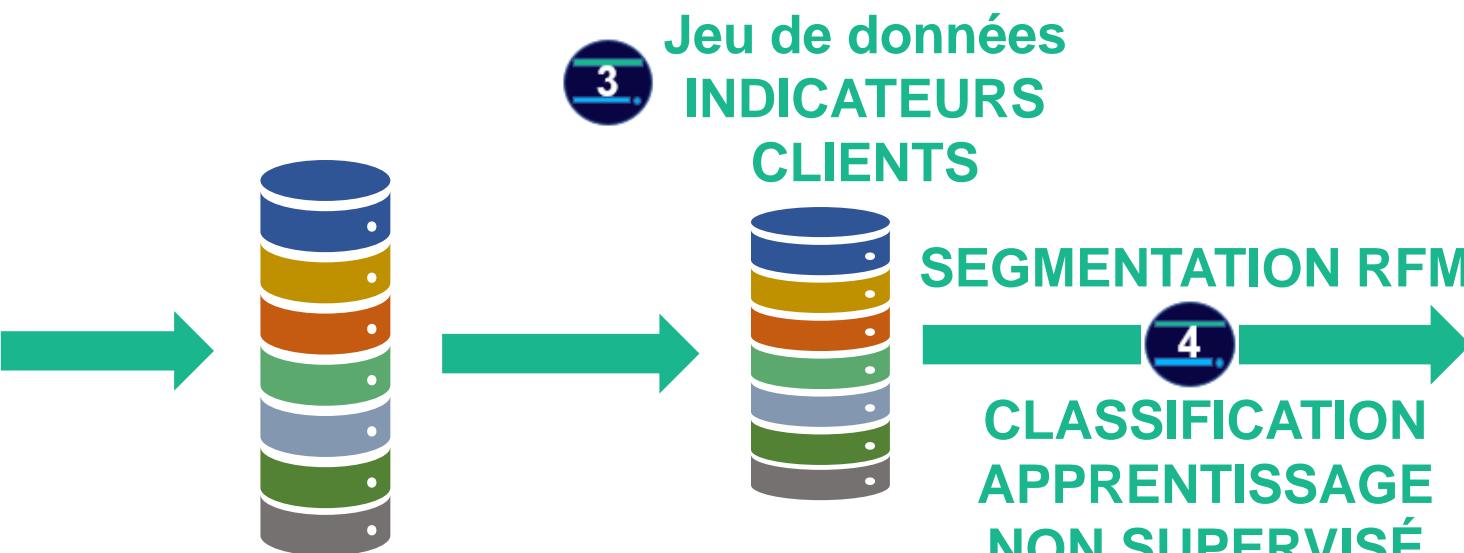
## 4- Jeu de données final

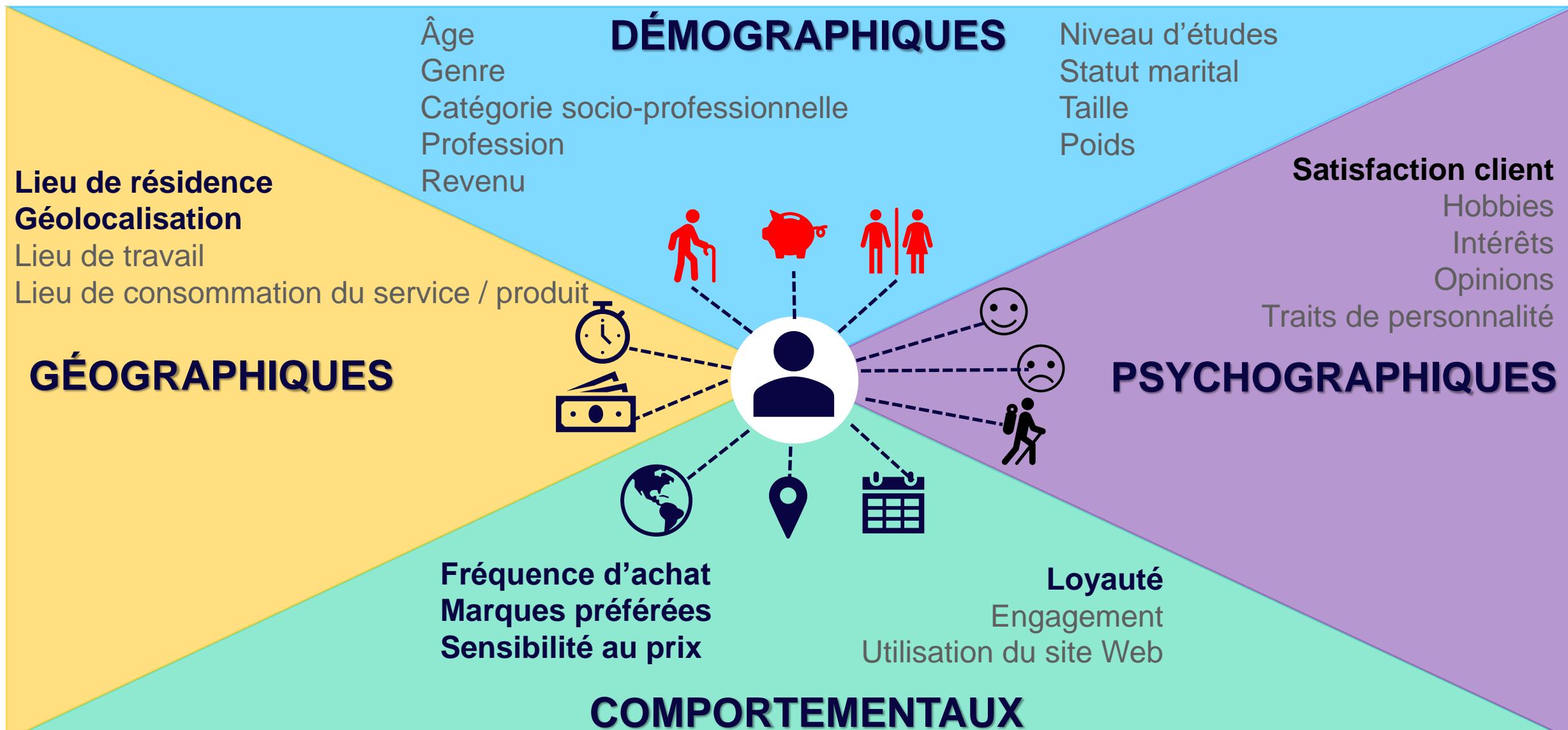
## Quels sont les indicateurs clients pertinents permettant d'effectuer une segmentation ?

### 1. Analyse métier + exploratoire



### 2. Jeu de données Assemblé





## GÉOGRAPHIQUES



Jour avec le plus de commandes



**Localisation :**  
ville de résidence  
état de résidence

**Géolocalisation :**

latitude  
longitude

Catégorie de produits

Catégorie la plus achetée

## DÉMOGRAPHIQUES



R

F

M

**Durée écoulée**  
depuis le dernier achat  
Heure, jour du dernier achat  
Mois, année du dernier achat

**Fréquence d'achat**

**Montant total des achats**  
Panier moyen

## PSYCHOGRAPHIQUES



**Note moyenne**  
de satisfaction  
Nombre d'avis



## COMMANDES

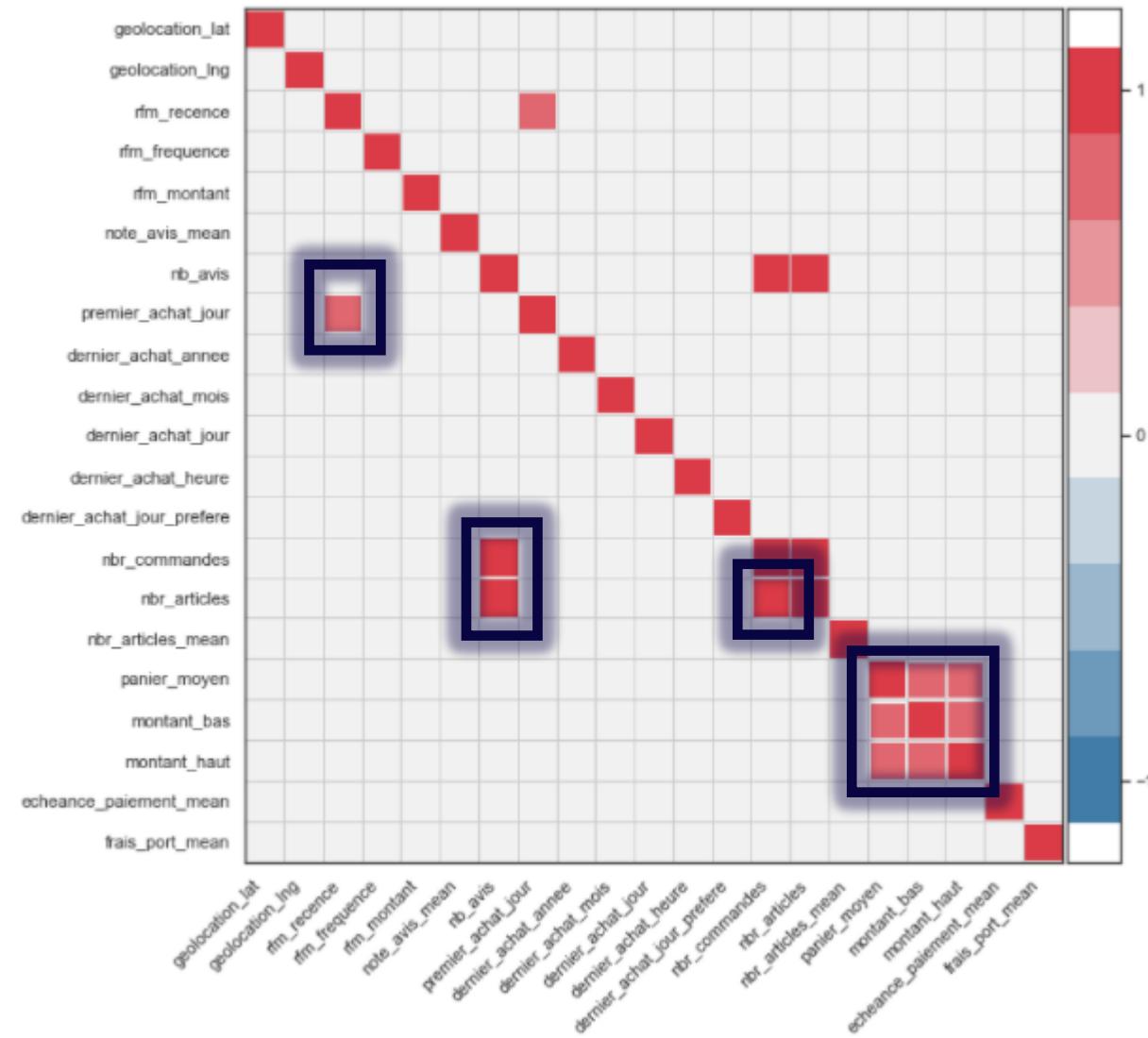
Nombre  
Nombre d'articles  
Nombre d'articles moyens

## PAIEMENTS

Moyen de paiement, haut, bas  
Facilités de paiement

## COMPORTEMENTAUX





2 Jeux de données finaux pour la segmentation

## Segmentation RFM



3 variables

## Algorithmes d'apprentissage non supervisé



20 variables



Problématique



Données



Modélisations



Conclusions

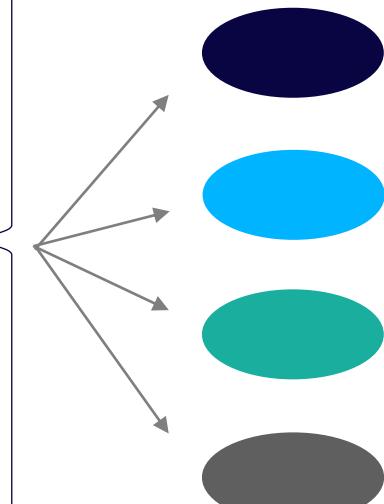
## 1. Segmentation RFM



3 variables



**Segmentation CLIENTS**

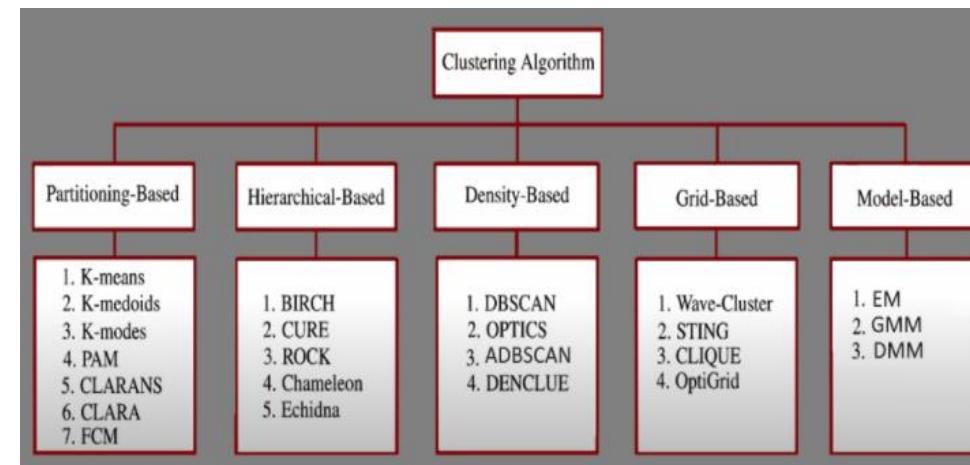


À interpréter

## 2. Algorithmes d'apprentissage non supervisé



20 variables





Problématique



Données



Modélisations – Segmentation RFM



Conclusions

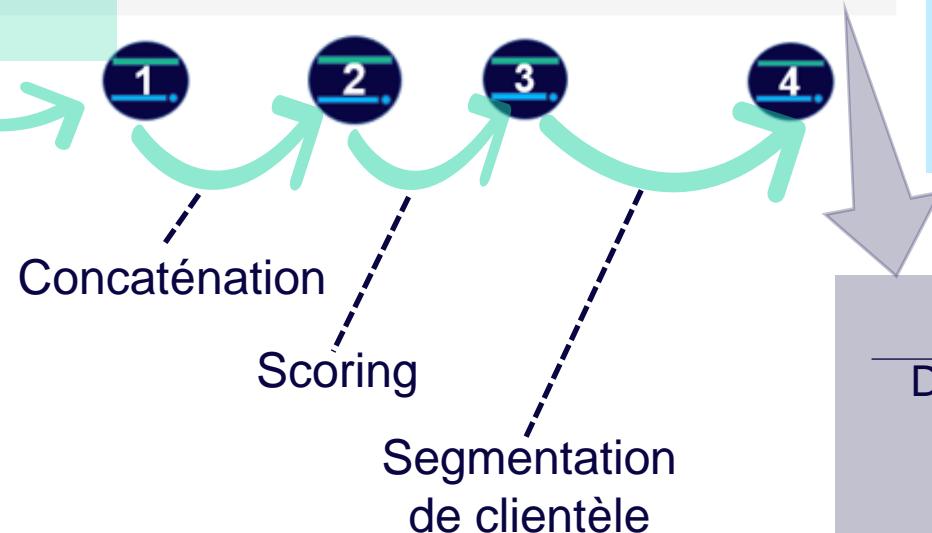
# 3. RFM - Scores et segments



Olist

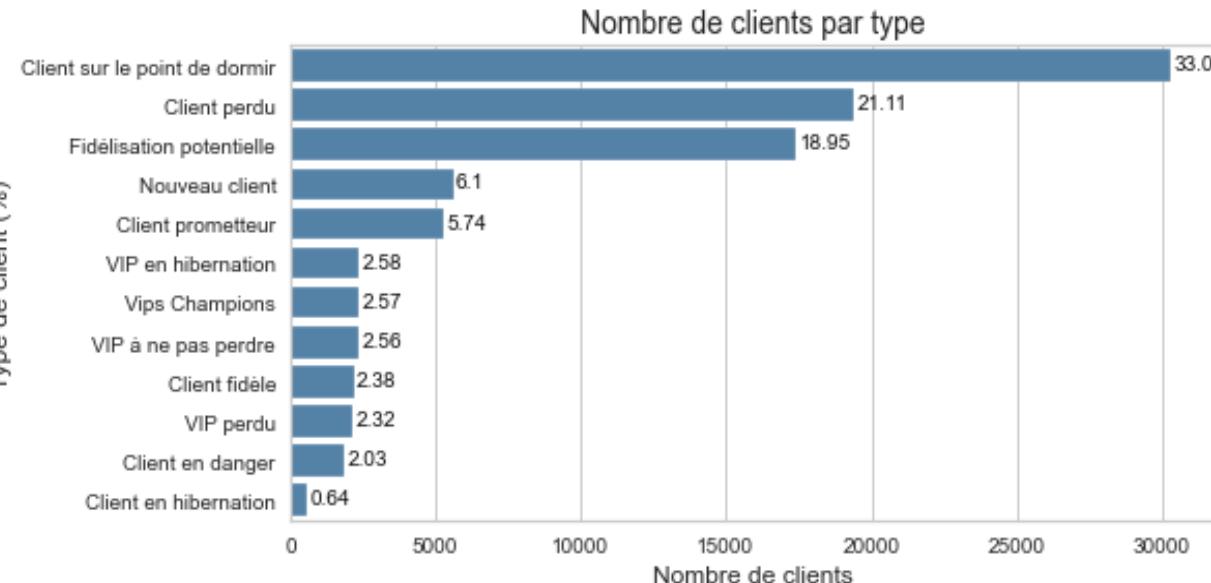
	R	F	M							
customer_unique_id	rfm_recence	rfm_frequence	rfm_montant	R	F	M	RFM_Segment	RFM_Score	RFM_Segm_Client	RFM_Niveau
0000366f3b9a7992bf8c76cfdf3221e2	112	1	141.90	4	1	3	413	8	Fidélisation potentielle	Argent
0000b849f77a49e4a4ce2b2a4ca5be3f	115	1	27.19	3	1	1	311	5	Client sur le point de dormir	Bronze
0000f46a3911fa3c0805444483337064	537	1	86.22	1	1	2	112	4	Client perdu	Bronze
0000f6ccb0745a6a4b88665a16c9f078	321	1	43.62	2	1	1	211	4	Client sur le point de dormir	Bronze
0004aac84e0df4da2b147fca70cf8255	288	1	196.89	2	1	3	213	6	Client sur le point de dormir	Bronze

Récence, Fréquence et Montant découpées en **1, 2, 3 ou 4** par quartiles ou Kmeans

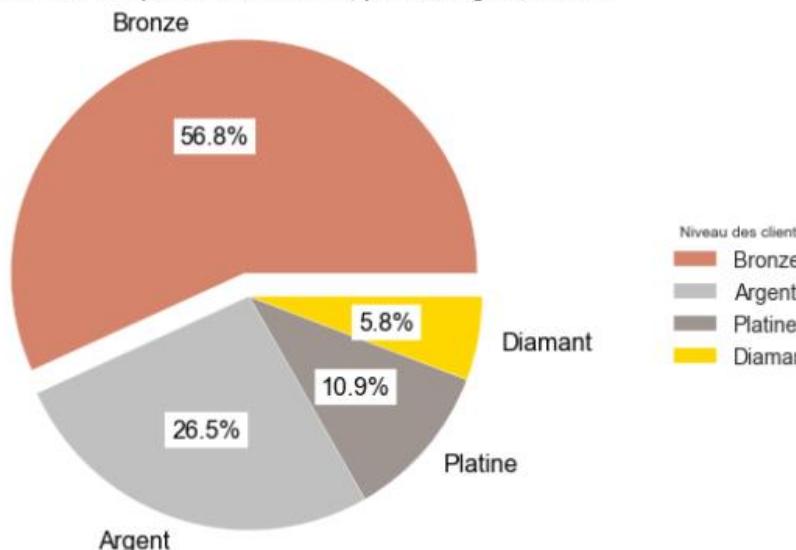


	R	F	M
VIPs Champions	4	4	4
Clients fidèles	3/4	1/2	3/4
Fidélisation potentielle	3/4	1/2	3/4
Nouveaux clients	4	1	1
Clients prometteurs	3/4	1/2	1/2
Clients ayant besoin d'attention	1/2	2/3	2/3
Sur le point de dormir	2/3	1/2	1-4
En danger	1/2	3/4	3/4
VIPs à ne pas perdre	3	4	4
VIPs en hibernation	2	4	4
Client en hibernation	1	4	1-4
VIPs perdus	1	4	4
Clients perdus	1	1/2	1

	RFM Score
Diamant	>= 11
Platine	entre 9 et 10
Argent	entre 7 et 8
Bronze	< 7



Distribution des clients par niveau diamant, platine, argent, bronze



## Segmentation RFM Olist



# RFM – Stabilité segments



O list

Periode	ARI
1-mois_Juillet_Août	0.73688
3-mois_Mai_Août	0.57709
6-mois_Février_Août	0.55590

Mars 2018

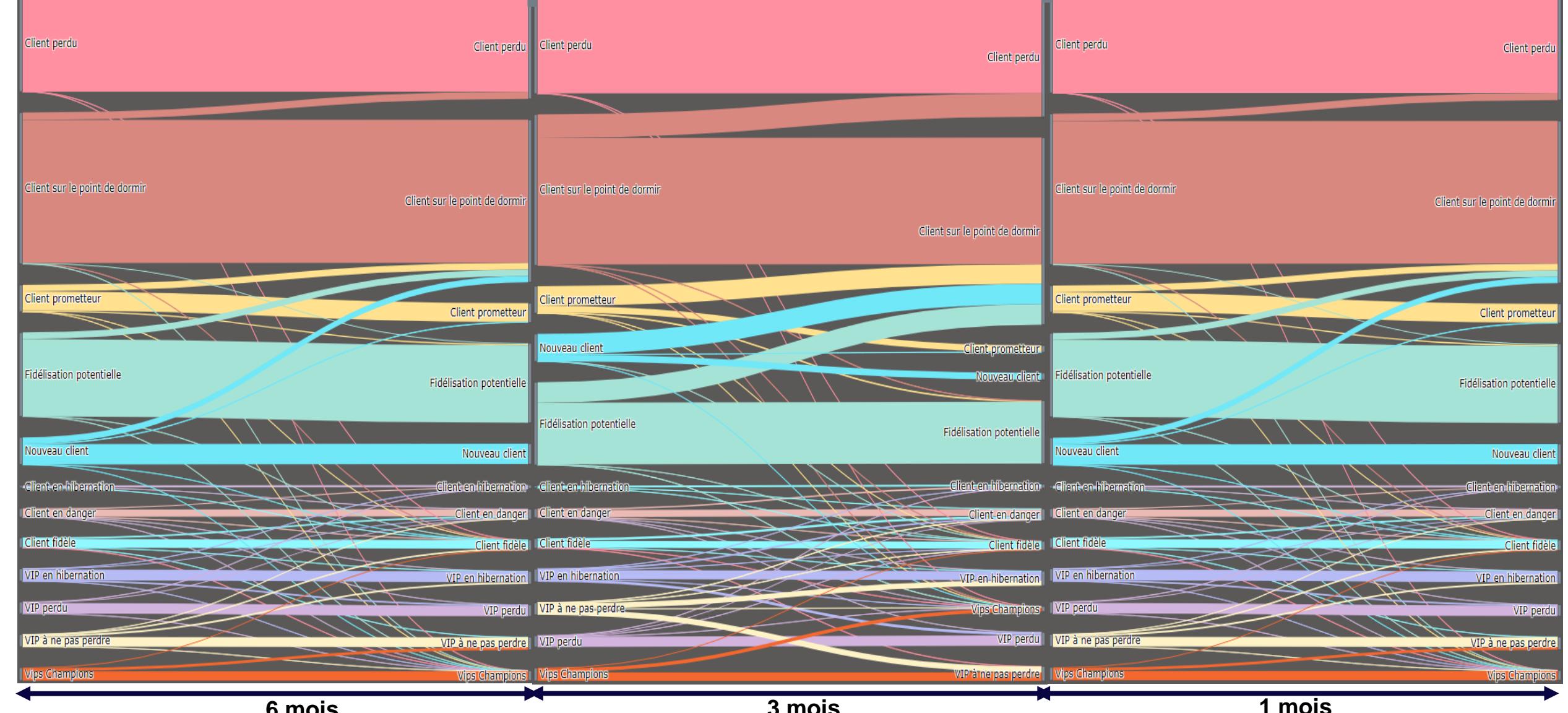
Août 2018

Mai 2018

Août 2018

Juillet 2018

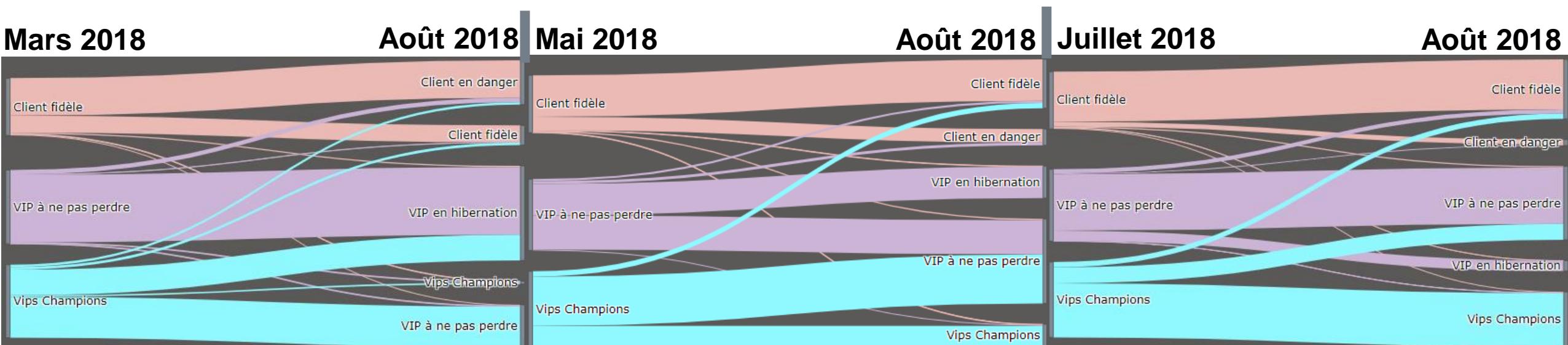
Août 2018



6 mois

3 mois

1 mois



# PROPOSITION DE CONTRAT DE MAINTENANCE

PRÉVISION DE CONTRAT DE MAINTENANCE	
Mise à jour trimestrielle	<p>Bonne stabilité des clients sur 1 mois.</p> <p>Bonne stabilité des clients fidèles sur 3 mois :</p> <ul style="list-style-type: none"><li>- mais 50% VIP à ne pas perdre deviennent des VIP en hibernation donc perdus,</li><li>- 2/3 des VIP champions deviennent des VIPS à ne pas perdre,</li><li>- peu de nouveaux clients ou clients avec potentiel de fidélisation ne deviennent des clients fidèles ou Vip.</li></ul> <p>➔ surveillance, actions à envisager.</p>

Type client	Attentions	Actions
<b>VIPs Champion</b>	Fortes	<b>Offrez des récompenses</b> (produits populaires). <b>Construisez votre crédibilité</b> (médias sociaux, commentaires).
<b>VIPs à ne pas perdre</b>	A stimuler	<b>Services sur mesure</b> (cibler les clients). <b>Passez un appel téléphonique</b> (être à l'écoute). <b>Connectez-vous sur les médias sociaux</b> (sur leurs réseaux).
<b>Fidélisation potentielles</b>	A privilégier	<b>Proposez un programme de fidélité</b> (adhésion à un club d'élite). <b>Organisez des concours.</b> <b>Faites en sorte qu'ils se sentent spéciaux</b> (souhaiter anniversaire, points cadeaux)
<b>Nouveaux clients</b>	A stimuler	<b>Fournissez une aide à l'accueil</b> (courrier bienvenue, kit d'accueil). <b>Offrez-leur des remises</b> (points de réduction, coupons). <b>Établissez une relation</b> (avis sur site, questions).
<b>Clients prometteurs</b>	A privilégier	<b>Proposez un essai gratuit</b> (produits hauts de gamme). <b>Faites connaître votre marque</b> (invitation à des évènements). <b>Proposez une annonce</b> (publicité en incluant les clients).
<b>Clients fidèles</b>	Fortes	<b>Prenez des retours d'information et des enquêtes</b> (avis). <b>Faites des ventes incitatives de vos produits.</b> <b>Offrez des bonus</b> (livraison gratuite, réductions).
<b>Clients ayant besoin d'attention</b>	A stimuler	<b>Proposez des produits combinés</b> (en fonction des anciens achats). <b>Soyez nostalgique</b> (offre sur les produits anciennement achetés). <b>Faites une farce</b> (avec les produits pour maintenir leurs intérêts).



Problématique



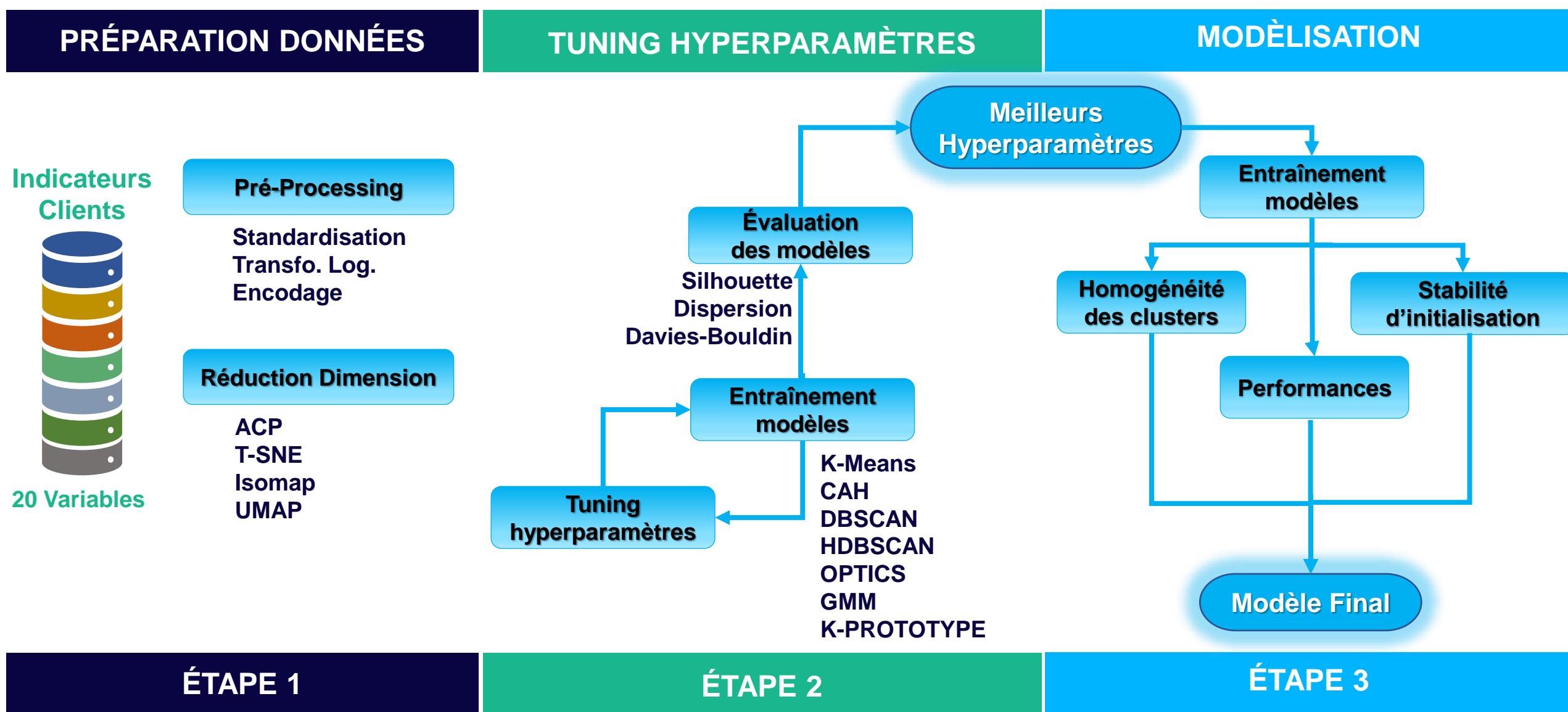
Données



Modélisations – Apprentissage  
Non supervisé



Conclusions



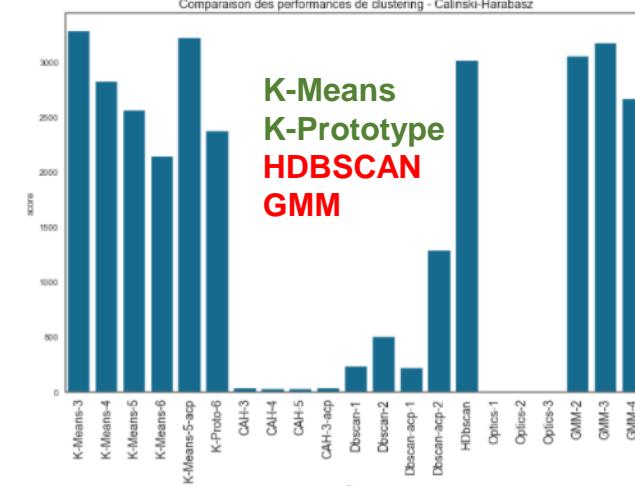
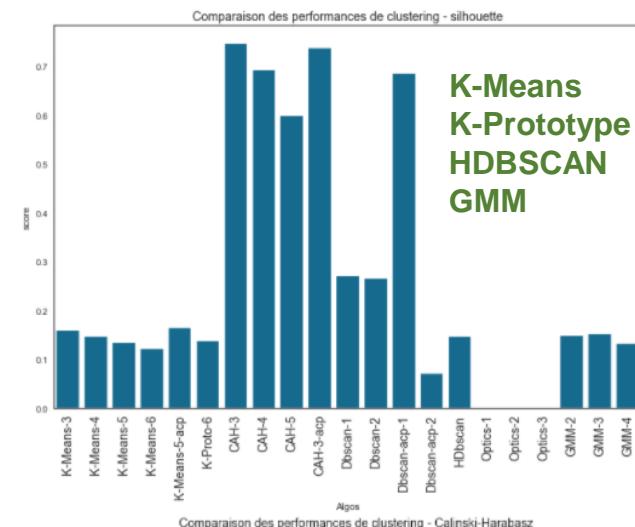
ÉTAPE 1

ÉTAPE 2

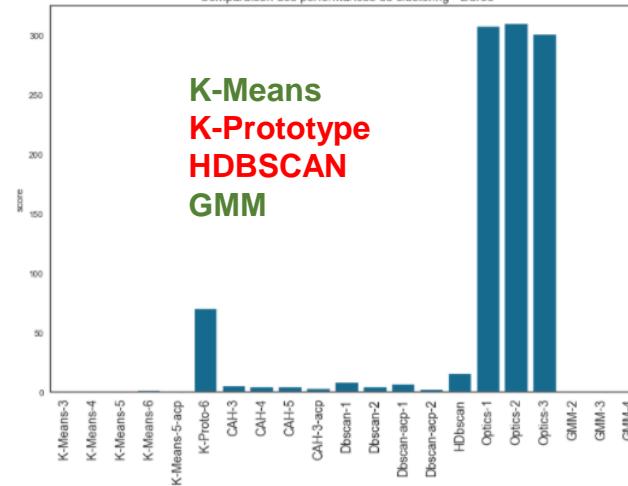
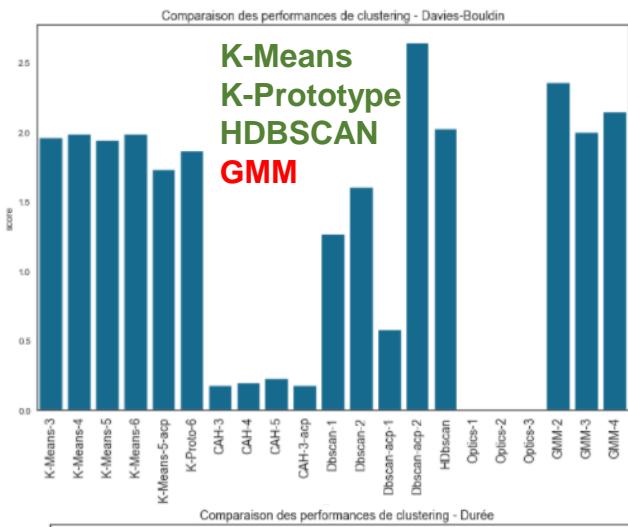
ÉTAPE 3

# Clustering – Performances (20000 clients)

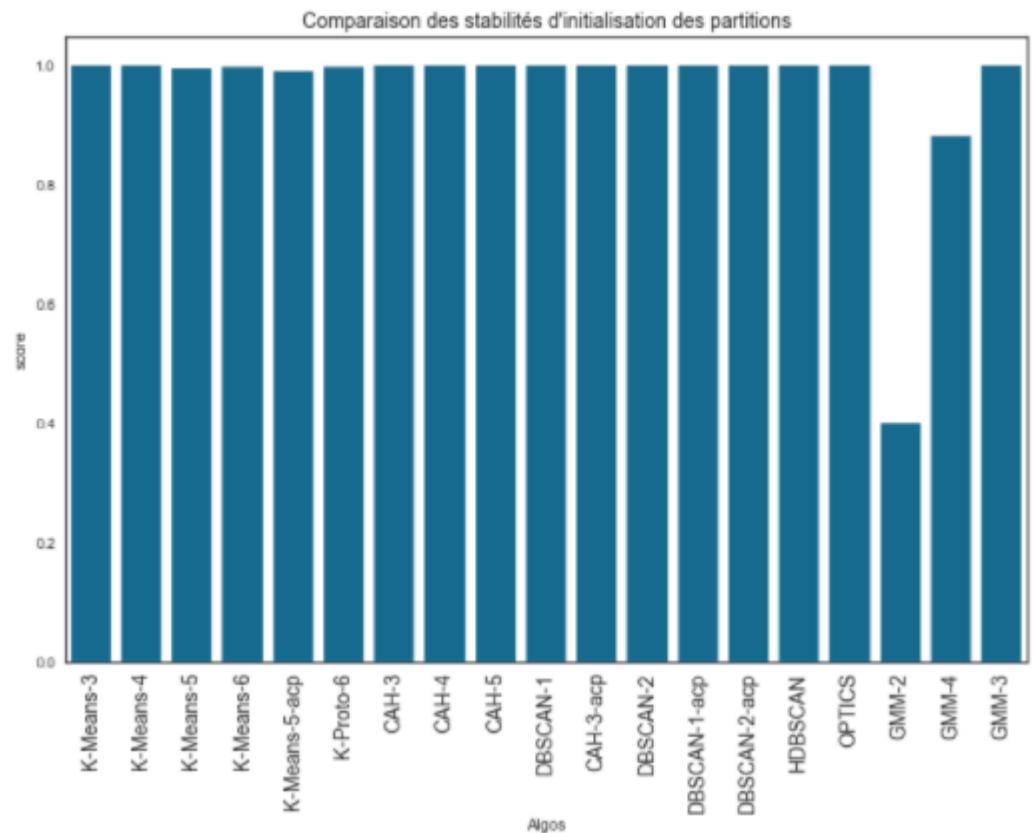
Algos	Nb_clusters	coef_silh	davies_bouldin	calinski_harabasz	Durée
K-Means-3	3	0.15883	1.95836	3281.44019	0.15855
K-Means-4	4	0.14617	1.98232	2818.52479	0.23335
K-Means-5	5	0.13466	1.93985	2564.19551	0.28454
K-Means-6	6	0.12082	1.98239	2140.17771	0.74199
K-Means-5-acp	5	0.16385	1.73178	3214.79648	0.20360
K-Proto-6	6	0.13665	1.86386	2369.06818	69.75964
CAH-3	3	0.74695	0.17329	30.51173	4.50845
CAH-4	4	0.69212	0.19346	26.76747	3.53690
CAH-5	5	0.59979	0.22597	22.82009	3.55105
CAH-3-acp	3	0.73880	0.17507	30.46068	2.65669
DbSCAN-1	3	0.27097	1.26156	227.54737	7.74423
DbSCAN-2	3	0.26518	1.59943	498.68343	3.49592
DbSCAN-acp-1	2	0.68634	0.57195	214.90855	6.34074
DbSCAN-acp-2	7	0.07081	2.64051	1281.09022	1.39324
HDBSCAN	3	0.14674	2.02266	3010.27275	14.94367
OPTICS-1	1	0.00000	0.00000	0.00000	307.50031
OPTICS-2	1	0.00000	0.00000	0.00000	309.63290
OPTICS-3	1	0.00000	0.00000	0.00000	300.88256
GMM-2	2	0.14902	2.35584	3053.91315	0.07876
GMM-3	3	0.15134	1.99644	3170.48551	0.14912
GMM-4	4	0.13233	2.14466	2660.01026	0.30985



Algorithme le plus performant et rapide :  
**K-Means**



Algos	ARI_mean	ARI_std
K-Means-3	0.99992	0.00005
K-Means-4	1.00000	0.00000
K-Means-5	0.99515	0.00498
K-Means-6	0.99790	0.00221
K-Means-5-acp	0.99102	0.00458
K-Proto-6	0.99847	0.00197
CAH-3	1.00000	0.00000
CAH-4	1.00000	0.00000
CAH-5	1.00000	0.00000
DBSCAN-1	1.00000	0.00000
CAH-3-acp	1.00000	0.00000
DBSCAN-2	1.00000	0.00000
DBSCAN-1-acp	1.00000	0.00000
DBSCAN-2-acp	1.00000	0.00000
HDBSCAN	1.00000	0.00000
OPTICS	1.00000	0.00000
GMM-2	0.40019	0.51623
GMM-4	0.88181	0.12334



**K-Means/K-Prototype stables.  
GMM moins stable.**

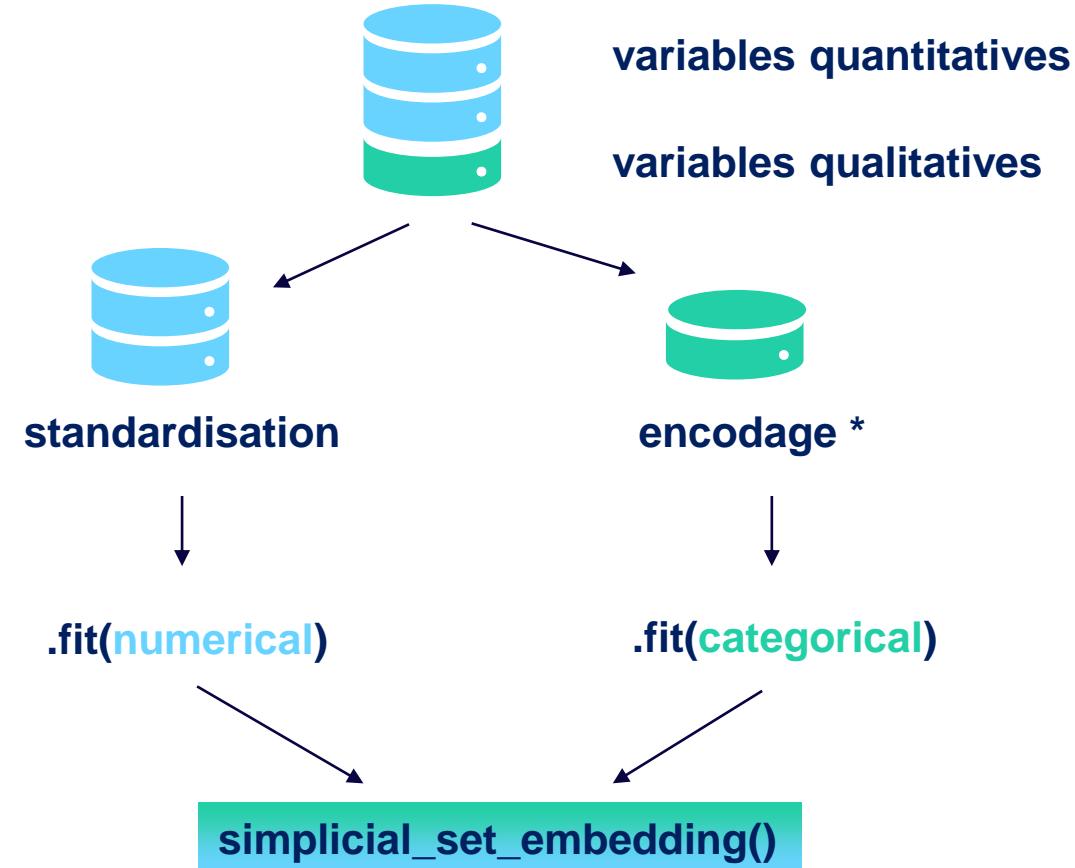
DONNÉE

PRE-PROCESSING

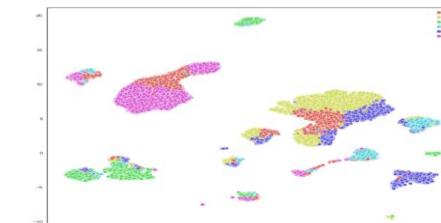
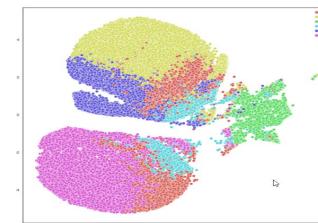
EMBEDDING

COMBINAISON

VISUALISATION

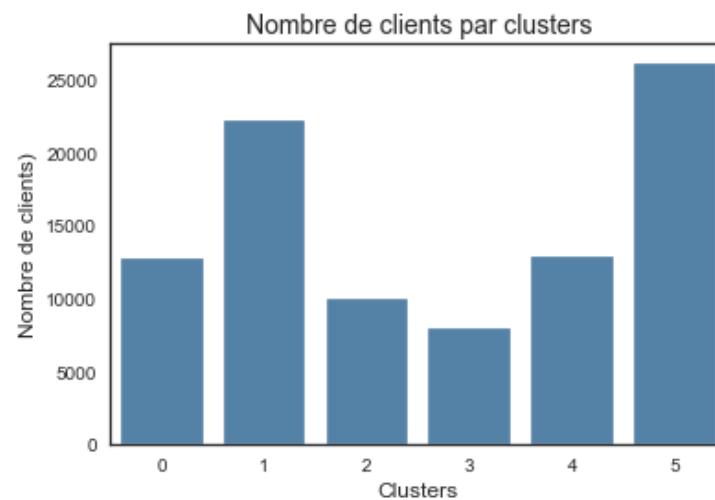


T-sne

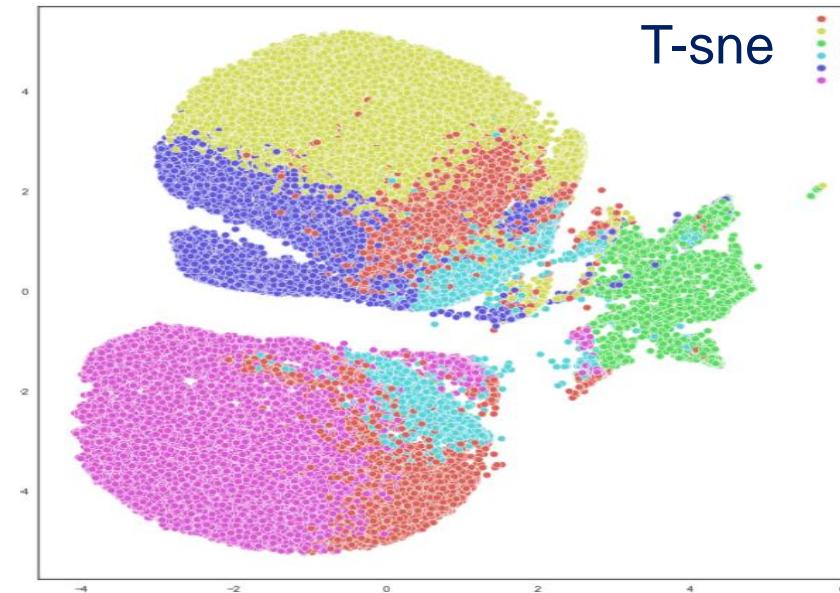


UMAP

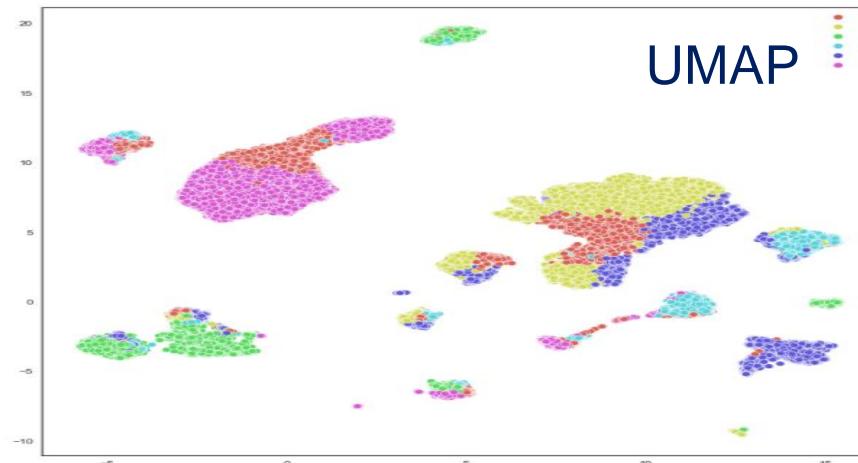
# 3. Clustering – K-Means k=6



 Clusters équilibrés



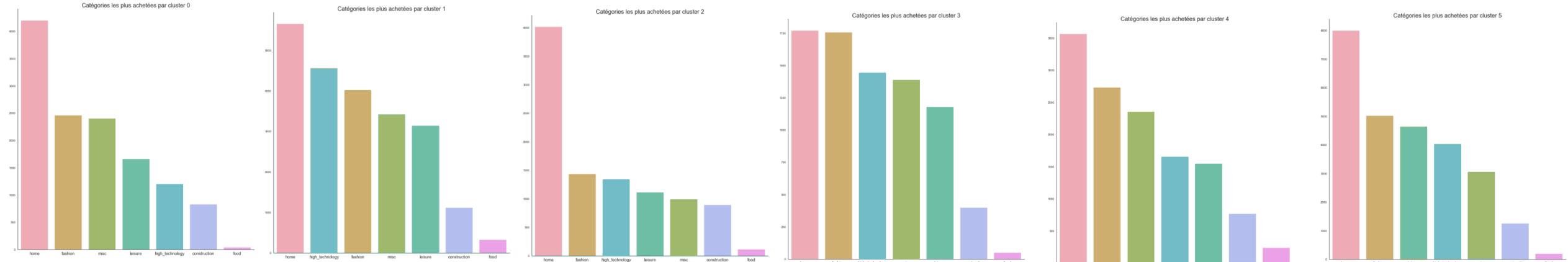
 Clusters distinguables



 Frontières nettes



## Catégories des produits par cluster



**Cluster 0**  
Meilleurs + échéance

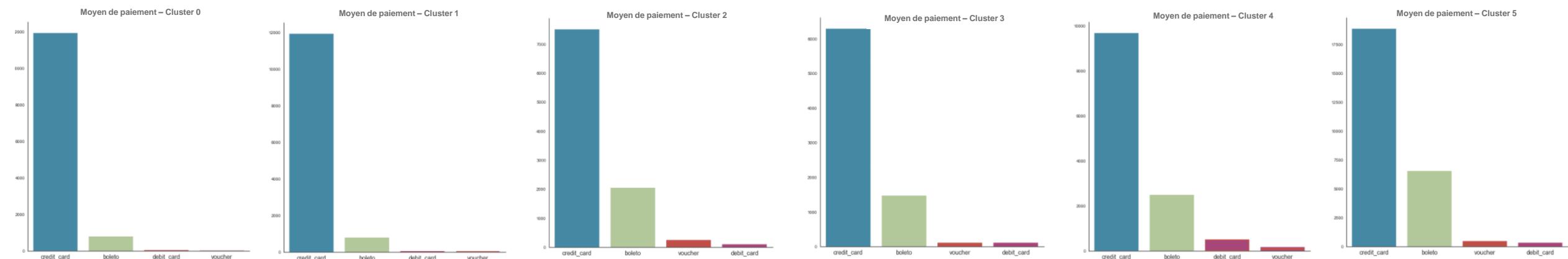
**Cluster 1**  
Nouveau

**Cluster 2**  
Meilleurs

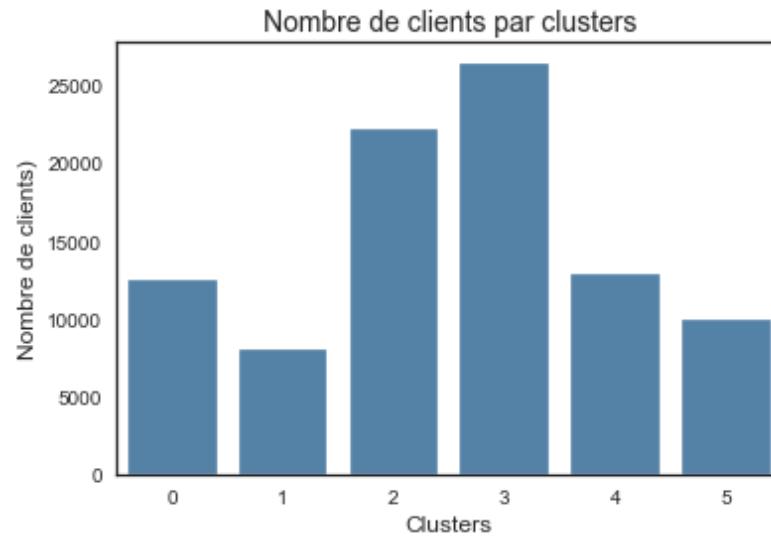
**Cluster 3**  
Fidèles Nord

**Cluster 4**  
Fidèles

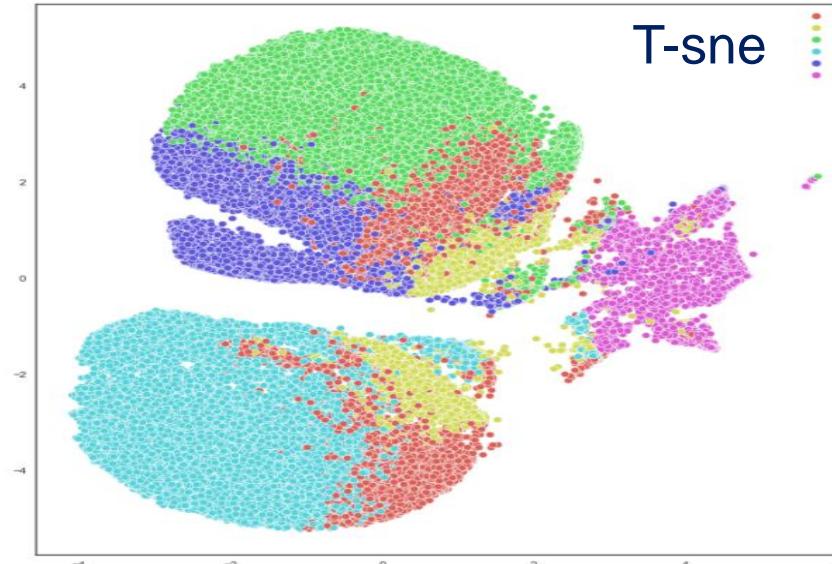
**Cluster 5**  
Perdus



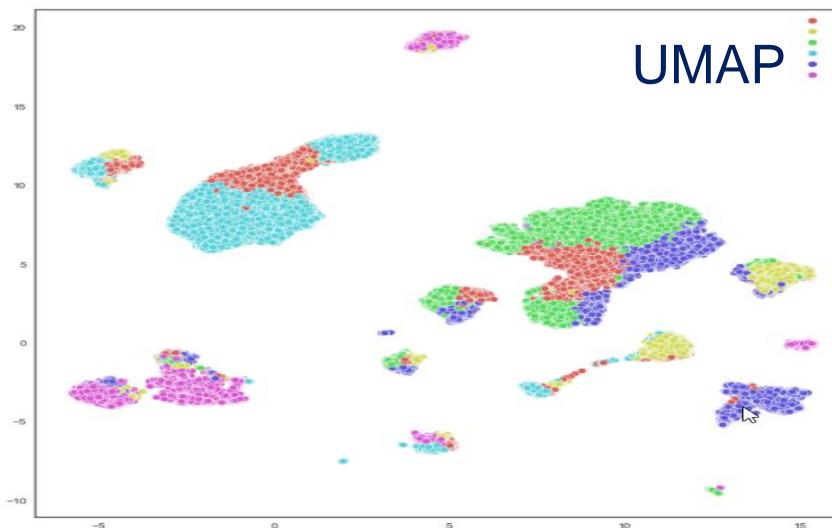
Catégories des moyens de paiement par cluster



Clusters équilibrés

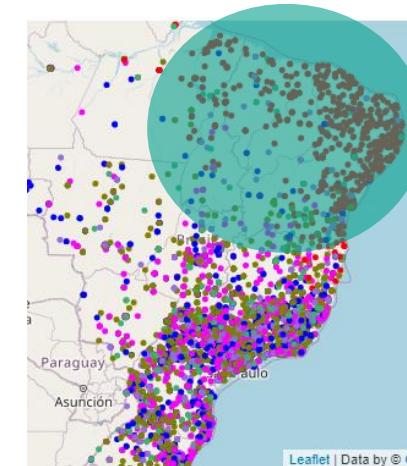
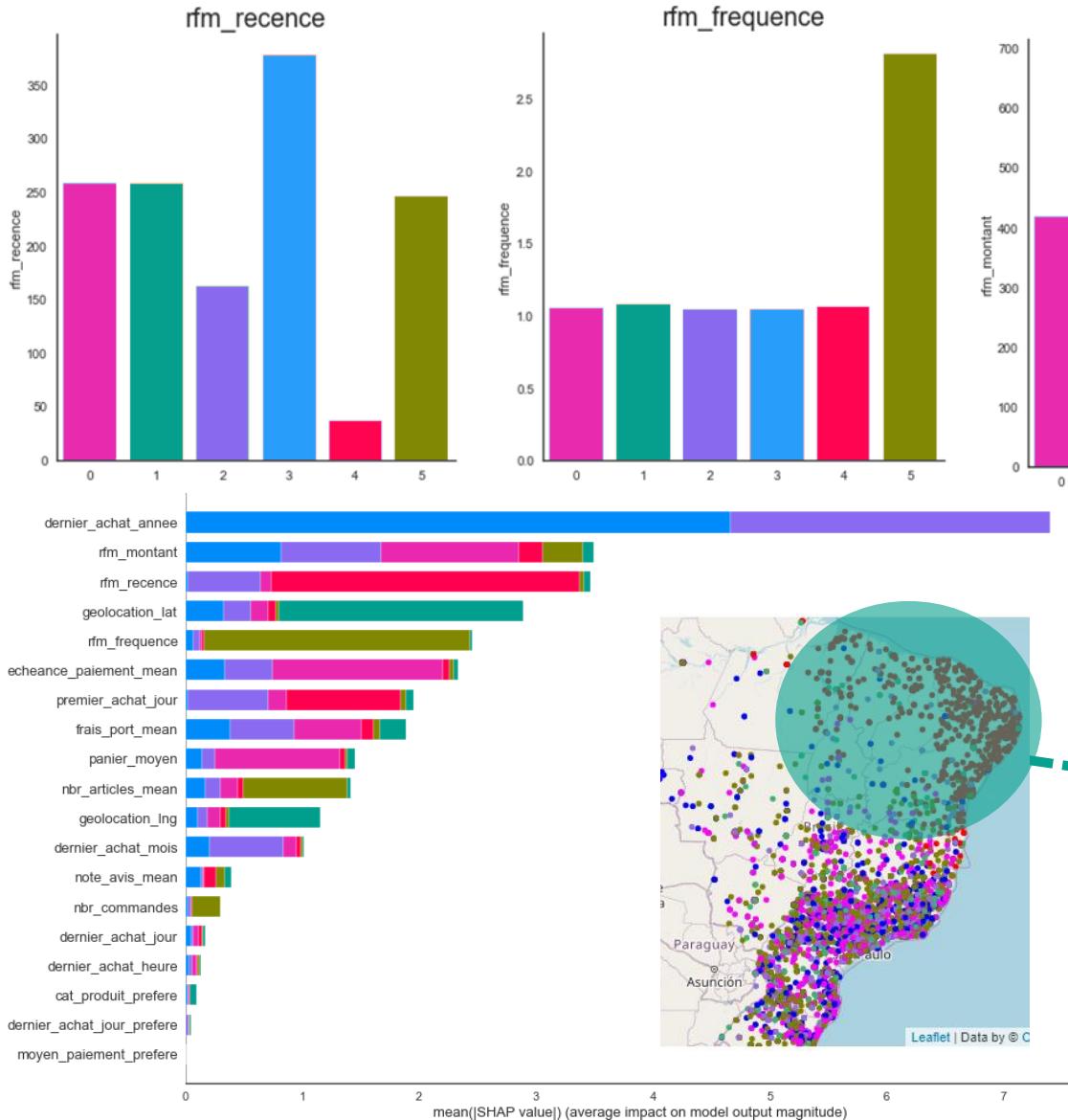


Clusters distinguables



Frontières nettes

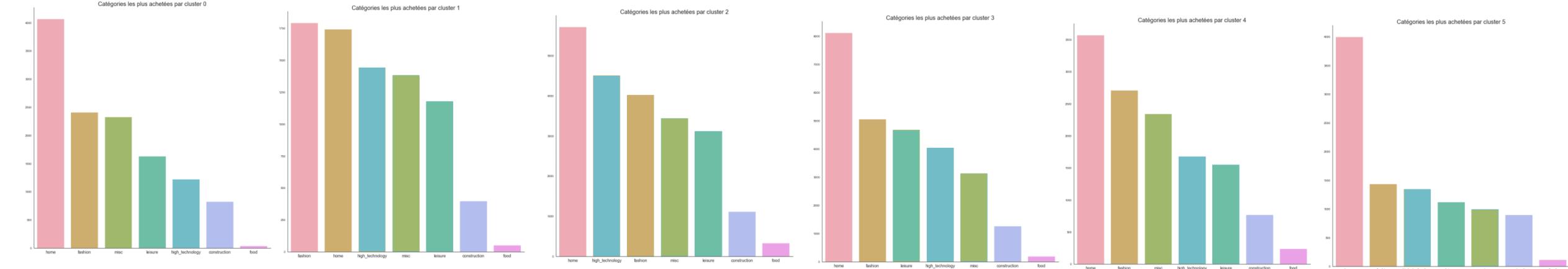
# Clustering – K-Prototype k=6



- Cluster 0 : Meilleurs clients utilisant les facilités de paiement ayant besoin d'attention.**
- Cluster 1 : Clients fidèles du Nord du Brésil achetant la catégorie 'fashion' ayant besoin d'attention.**
- Cluster 2 : Nouveaux clients.**
- Cluster 3 : Clients perdus.**
- Cluster 4 : Clients fidèles.**
- Cluster 5 : Meilleurs clients ayant besoin d'attention.**

Cluster_Kproto_6	cat_produit_prefere
4	home
2	home
5	home
0	home
1	<b>fashion</b>
3	home

## Catégories des produits par cluster



**Cluster 0**  
Meilleurs + échéance

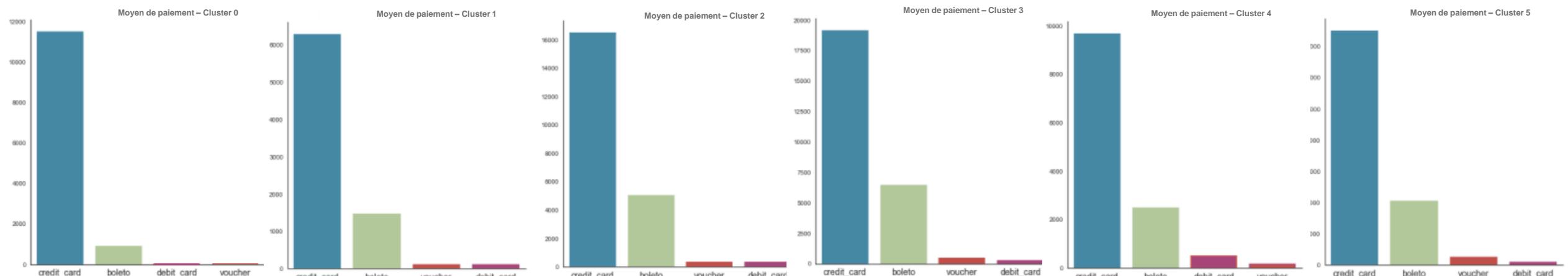
**Cluster 1**  
Fidèles Nord

**Cluster 2**  
Nouveaux

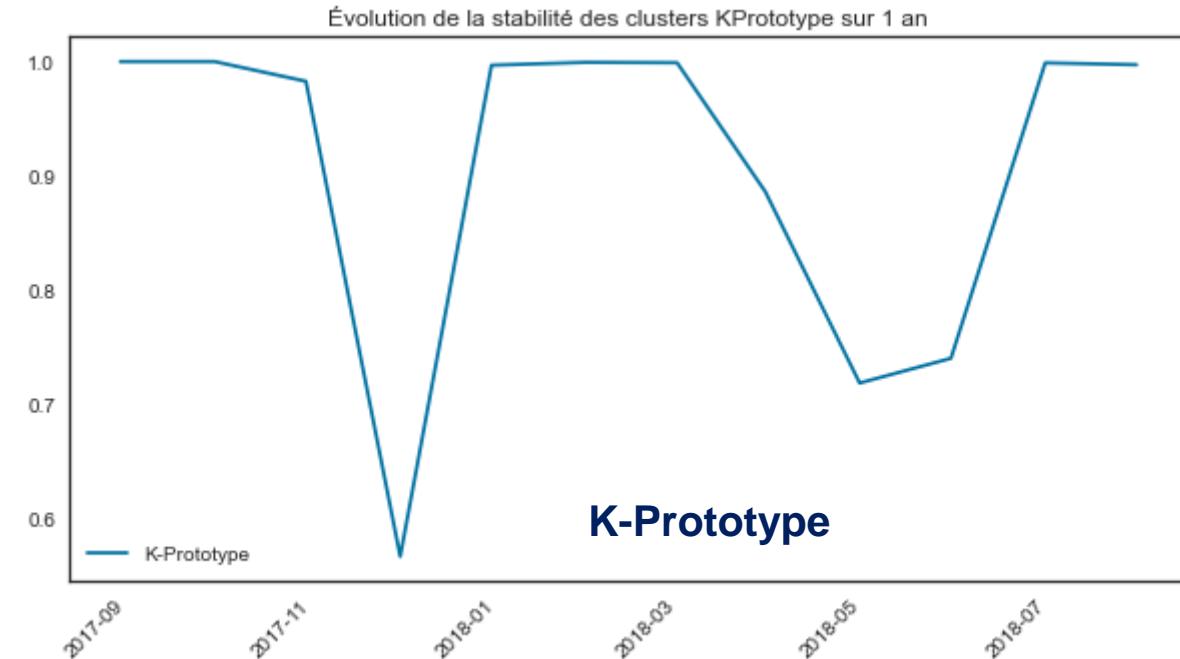
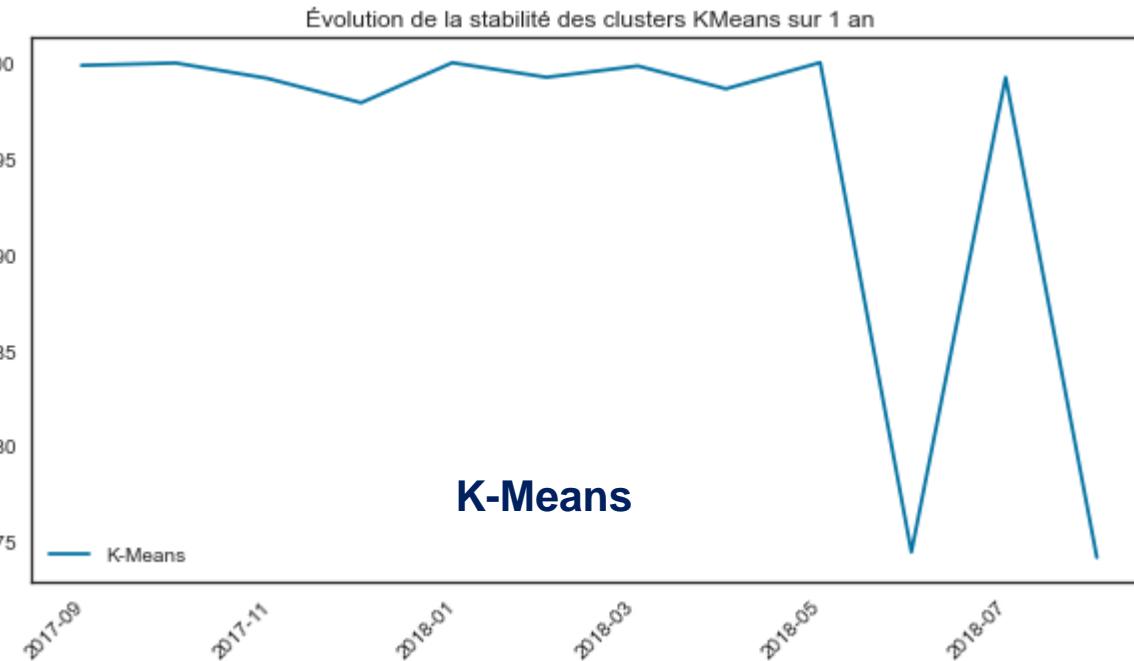
**Cluster 3**  
Perdus

**Cluster 4**  
Fidèles

**Cluster 5**  
Meilleurs



Catégories des moyens de paiement par cluster



Proposition d'un contrat de maintenance avec mise à jour trimestrielle



Problématique



Données



Modélisations



Conclusions

Modèle	Avantages	Inconvénients
Segmentation RFM	Simple à mettre en œuvre. Rapide. Marketing traditionnel.	3 variables seulement prise en compte. Tout refaire pour l'ajout de nouveaux clients.
K-Means	Tous les indicateurs clients numériques + qualitatifs encodés pris en compte.  Segments assez homogènes, interprétables, actionnables et assez stable dans le temps (3 mois), stable à l'initialisation.  Entraînement rapide.  Nouveaux clients pris en compte facilement (.predict).	Paramètre k à initialiser. Variables qualitatives mal prises en compte.
K-Prototype	Tous les indicateurs clients mixtes pris en compte.  Segments assez homogènes, interprétables, actionnables et assez stable dans le temps (3 mois) et stable à l'initialisation.  Nouveaux clients pris en compte facilement (.predict).	Paramètre k à initialiser. Lent. Paramétrage assez difficile.

### Modèle final :

**Kmeans avec 6 clusters** est une bonne piste de départ pour démarrer la segmentation de clientèle en partant d'un **contrat de maintenance avec mise à jour trimestrielle**.

Type client	Attentions	Actions
<b>Meilleurs clients ayant besoin d'attention</b>	Fortes	<b>Offrez des récompenses</b> (produits populaires). <b>Construisez votre crédibilité</b> (médias sociaux, commentaires).
<b>Clients fidèles (hors nord)</b>	A privilégier	<b>Prenez des retours d'information et des enquêtes</b> (avis). <b>Faites des ventes incitatives de vos produits.</b> <b>Offrez des bonus</b> (livraison gratuite, réductions).
<b>Clients fidèles du Nord du Brésil ayant besoin d'attention</b>	A privilégier	<b>Prenez des retours d'information et des enquêtes</b> (avis). <b>Faites des ventes incitatives sur les autres catégories de produits.</b> <b>Offrez des bonus</b> (livraison gratuite, réductions).
<b>Nouveaux clients</b>	A stimuler	<b>Fournissez une aide à l'accueil</b> (courrier bienvenue, kit d'accueil, 25% 2 <sup>ème</sup> commande). <b>Offrez-leur des remises</b> (points de réduction, coupons). <b>Établissez une relation</b> (avis sur site, questions).
<b>Meilleurs clients utilisant les facilités de paiement ayant besoin d'attention</b>	A privilégier	<b>Proposez un essai gratuit</b> (produits hauts de gamme). <b>Faites connaître votre marque</b> (invitation à des évènements). <b>Proposez une annonce</b> (publicité en incluant les clients). <b>Proposer plus de choix de facilité de paiements</b> (accessibles facilement).
<b>Clients perdus</b>	Choix	<b>Comprenez-les</b> (faites des recherches sur les clients perdus). <b>Faites une dernière promotion</b> (e-mail de promotion). <b>Ignorez-les complètement</b> (ne reviendront pas donc perte d'argent).

## Jeu de données

Nécessite plus de données :

- démographiques (âge, profession, sexe, nombre d'enfants..)
- psychographiques (avis sur le produit, centre d'intérêt...)

Biaisés :

96% des clients ne commandent qu'une seule fois.

Notes toutes très positives.

## Segmentation

Collaborer avec l'équipe Marketing métier :

Définir la finesse du nombre de segments souhaités par OLIST.

Valider le choix des variables ajoutées lors du feature engineering.

Valider les regroupements des catégories de produits.

Valider les premiers résultats (modifier le paramétrage/modèle si besoin).



# Annexes



## Catégorie de produits : de 73 à 7 catégories de produits

```

# Home
'furniture_bedroom': 'home',
'furniture_decor': 'home',
'furniture_living_room': 'home',
'furniture_mattress_and_upholstery': 'home',
'bed_bath_table': 'home',
'christmas_supplies': 'misc',
'kitchen_dining_laundry_garden_furniture': 'home',
'office_furniture': 'home',
'home_appliances': 'home',
'home_appliances_2': 'home',
'home_comfort_2': 'home',
'home_comfort': 'home',
'air_conditioning': 'home',
'housewares': 'home',
'art': 'home',
'arts_and_craftsmanship': 'home',
'party_supplies': 'misc',
'flowers': 'home',
'cool_stuff': 'home',

# Construction
'construction_tools_construction': 'construction',
'construction_tools_lights': 'construction',
'construction_tools_safety': 'construction',
'construction_tools_garden': 'construction',
'construction_tools_tools': 'construction',
'garden_tools': 'construction',
'home_construction': 'construction',

# High technology
'electronics': 'high_technology',
'audio': 'high_technology',
'tablets_printing_image': 'high_technology',
'telephony': 'high_technology',
'fixed_telephony': 'high_technology',
'small_appliances': 'high_technology',
'small_appliances_home_oven_and_coffee': 'high_technology',
'computers_accessories': 'high_technology',
'computers': 'high_technology',
'security_and_services': 'misc',
'signaling_and_security': 'misc',

```

```

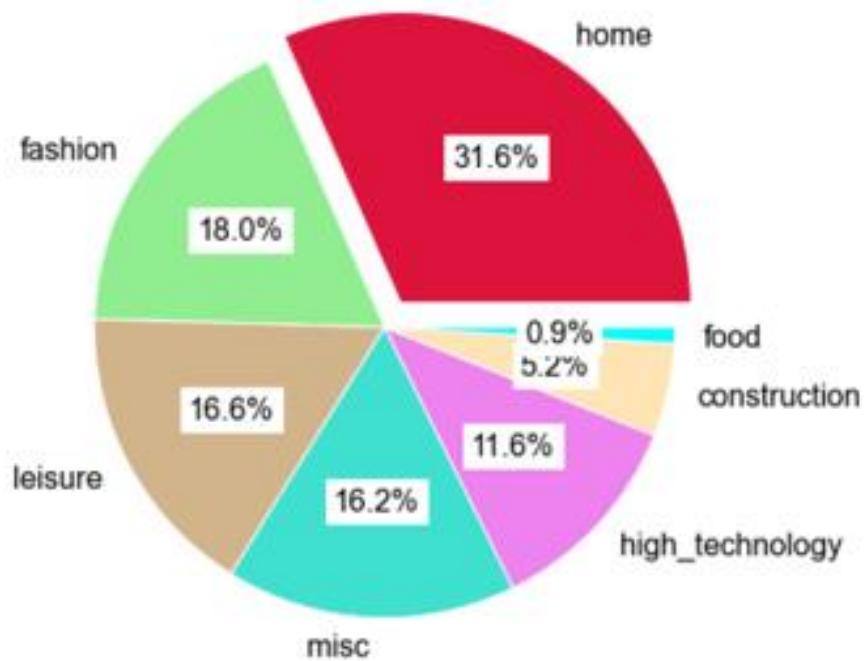
# Food
'drinks': 'food',
'food': 'food',
'food_drink': 'food',
'la_cuisine': 'food',
'kitchen_laptops_and_food_preparers': 'food',

# Fashion
'fashio_female_clothing': 'fashion',
'fashion_bags_accessories': 'fashion',
'fashion_childrens_clothes': 'fashion',
'fashion_male_clothing': 'fashion',
'fashion_shoes': 'fashion',
'fashion_sport': 'fashion',
'fashion_underwear_beach': 'fashion',
'health_beauty': 'fashion',
'perfumery': 'fashion',
'diapers_and_hygiene': 'fashion',
'baby': 'fashion',
'luggage_accessories': 'fashion',

# leisure
'sports_leisure': 'leisure',
'consoles_games': 'leisure',
'musical_instruments': 'leisure',
'toys': 'leisure',
'cine_photo': 'leisure',
'dvds_blu_ray': 'leisure',
'cds_dvds_musicals': 'leisure',
'music': 'leisure',
'books_general_interest': 'leisure',
'books_imported': 'leisure',
'books_technical': 'leisure',

```

Distribution des catégories des produits



## # Misc

```

'stationery': 'misc',
'auto': 'misc',
'watches_gifts': 'misc',
'agro_industry_and_commerce': 'misc',
'industry_commerce_and_business': 'misc',
'market_place': 'misc',
'pet_shop': 'misc',
'other': 'misc'

```

## Segmentation RFM

Description	Données	Variable	Action	Explication
Regroupement par client unique	customer	customer_unique_id	data.groupby('customer_unique_id').agg	1 ligne = 1 client
Récence 	orders	order_purchase_timestamp	'order_purchase_timestamp': lambda x: (date_ref - x.max()).days	Durée écoulée depuis le dernier achat
Fréquence 	orders	order_id	'order_id' : 'count'	Fréquence d'achat sur l'historique
Montant 	payments	payment_value	payment_value: 'sum'	Montant total des achats sur l'historique

## Clustering

Description	Données	Variable	Action	Explication
Segmentation géographique	customer	customer_city	Conserve les 200 villes les plus représentées et 'Autres' sinon 'customer_city': lambda x: x.mode()[0]	Ville de résidence du client
Segmentation géographique	Customer	customer_state	'customer_state': lambda x: x.mode()[0]	Etat de résidence du client
Segmentation géographique	Geolocation	geolocation_lat	'geolocation_lat': 'mean'	Latitude de la ville de résidence du client
Segmentation géographique	Geolocation	geolocation_lng	'geolocation_lng': 'mean'	Longitude de la ville de résidence du client
Segmentation psychographique	reviews	review_score	'note_avis_mean': 'mean',	Note de l'avis
Segmentation psychographique	reviews	review_score	'nb_avis': 'sum'	Nombre d'avis total

# Clustering

Description	Données	Variable	Action	Explication
Segmentation comportementale	orders	order_purchase_timestamp	'date_premier_achat': 'min'	Date du premier achat
Segmentation comportementale	orders	order_purchase_timestamp	'premier_achat_jour': lambda x: x.mode()[0]	Jour du premier achat
Segmentation comportementale	orders	order_purchase_timestamp	'date_dernier_achat': 'max'	Date du dernier achat
Segmentation comportementale	orders	order_purchase_timestamp	'dernier_achat_annee': lambda x: x.mode()[0]	Année du dernier achat
Segmentation comportementale	orders	order_purchase_timestamp	'dernier_achat_mois': lambda x: x.mode()[0]	Mois du dernier achat
Segmentation comportementale	orders	order_purchase_timestamp	'dernier_achat_jour': lambda x: x.mode()[0]	Jour du dernier achat

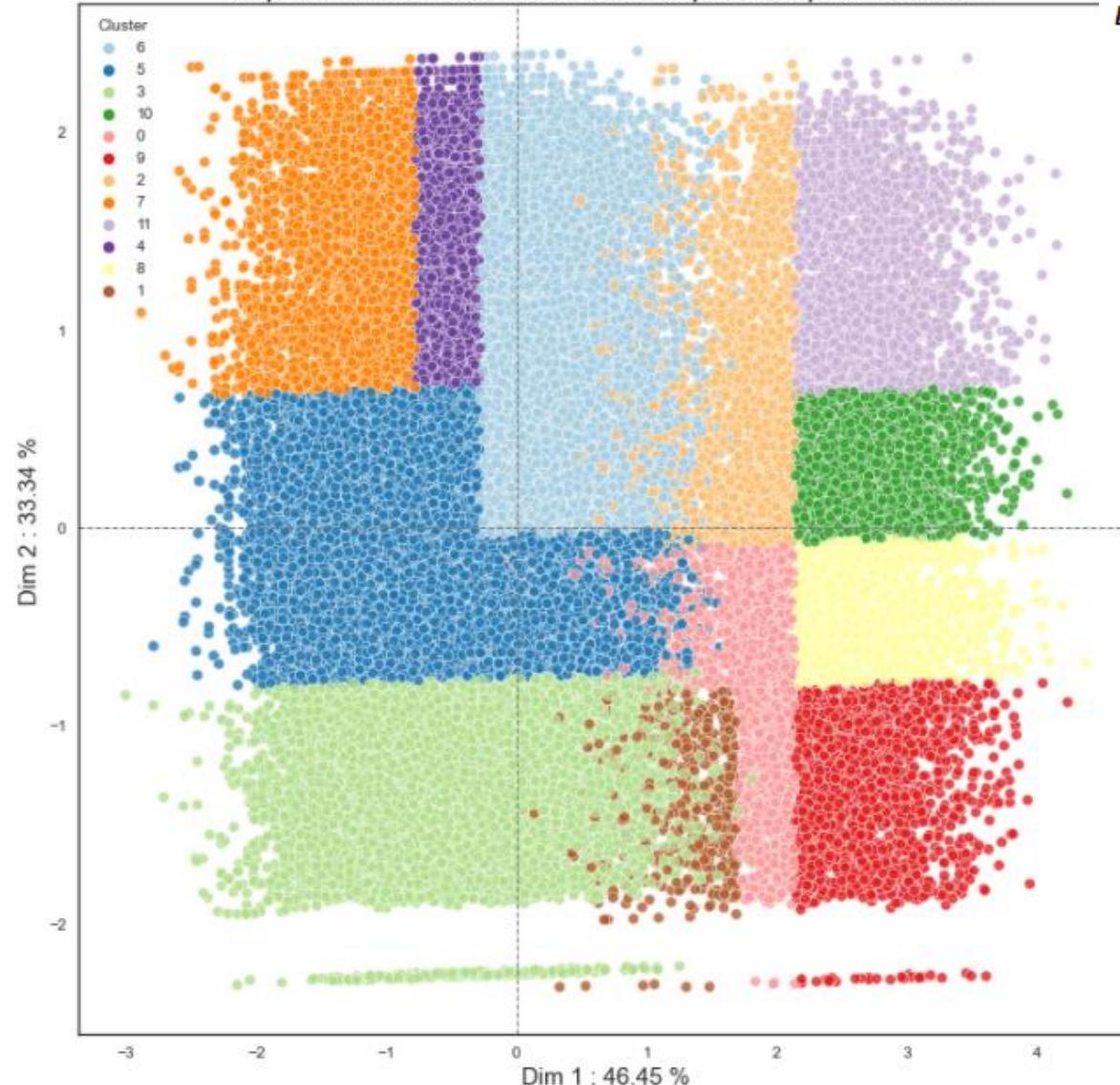
# Clustering

Description	Données	Variable	Action	Explication
Segmentation comportementale	orders	order_purchase_times_tamp	'dernier_achat_heure': lambda x: x.mode()[0]	Heure du dernier achat
Segmentation comportementale	orders	order_purchase_times_tamp	'dernier_achat_jour_prefere': lambda x: x.mode()[0]	Jour préféré du dernier achat
Segmentation comportementale	orders	orders_id	'nbr_commandes': 'sum'	Nombre de commandes sur l'historique
Segmentation comportementale	orders_items	order_item_id	'nbr_articles': 'sum'	Nombre d'articles sur l'historique
Segmentation comportementale	orders	order_item_id	'nbr_articles_mean': 'mean'	Nombre d'article moyen sur l'historique
Segmentation comportementale	payments	price	'panier_moyen': 'mean'	Prix du panier moyen sur l'historique

## Clustering

Description	Données	Variable	Action	Explication
Segmentation comportementale	payments	payment_value	'montant_bas': 'min'	Prix le plus haut sur l'historique
Segmentation comportementale	payments	payment_value	'montant_haut': 'max'	Prix le plus haut sur l'historique
Segmentation comportementale	payments	payment_type	'moyen_paiement_prefere': lambda x: x.mode()[0]	Moyen de paiement préféré
Segmentation comportementale	payments	payment_installments	'echeance_paiement_mean': 'mean'	Nombre d'échéance moyen sur l'historique
Segmentation comportementale	payments	freight_value	'frais_port_mean': 'mean'	Frais de port moyen sur l'historique
Segmentation comportementale	products	product_category_name_english	'cat_produit_prefere': lambda x: x.mode()[0]	Catégorie du produit préféré sur l'historique

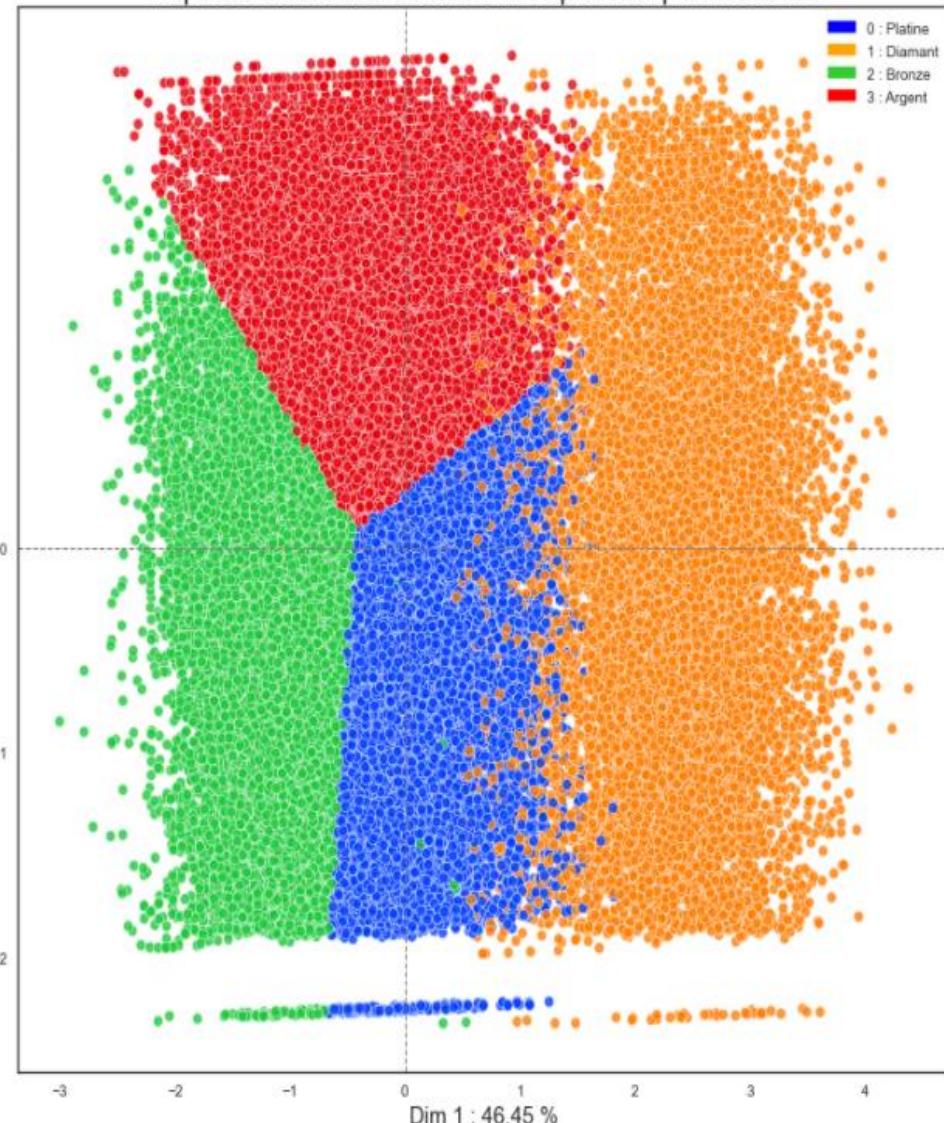
Représentation des clients sur le premier plan factoriel



Légende :

- 0 : Client en danger
- 1 : Client en hibernation
- 2 : Client fidèle
- 3 : Client perdu
- 4 : Client prometteur
- 5 : Client sur le point de dormir
- 6 : Fidélisation potentielle
- 7 : Nouveau client
- 8 : VIP en hibernation
- 9 : VIP perdu
- 10 : VIP à ne pas perdre
- 11 : Vips Champions

Représentation des clients sur le premier plan factoriel



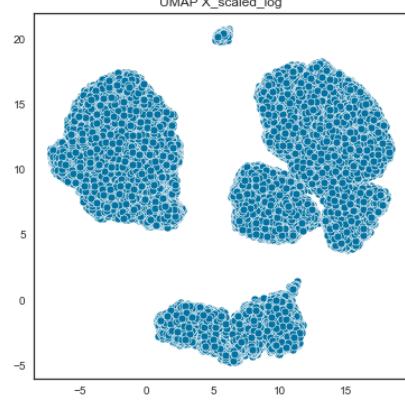
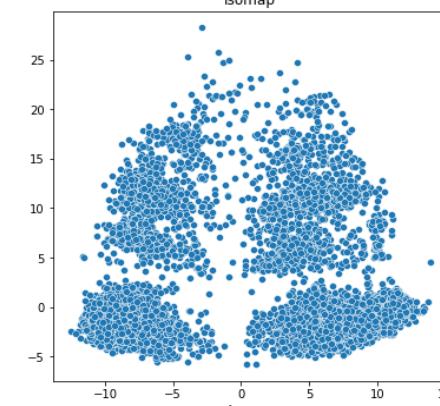
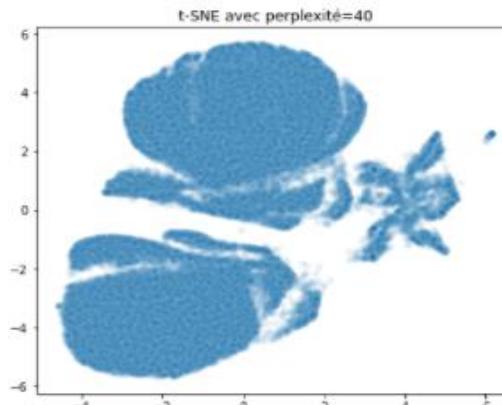
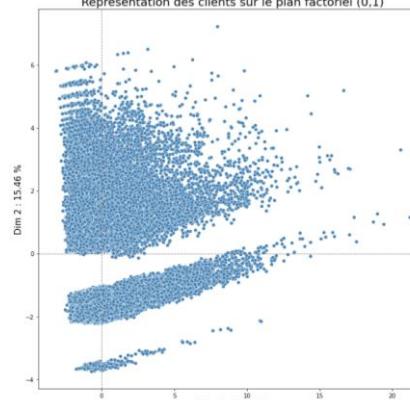
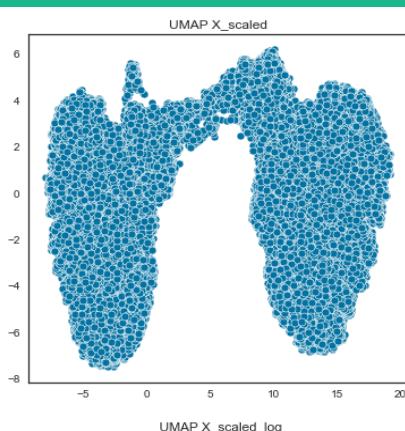
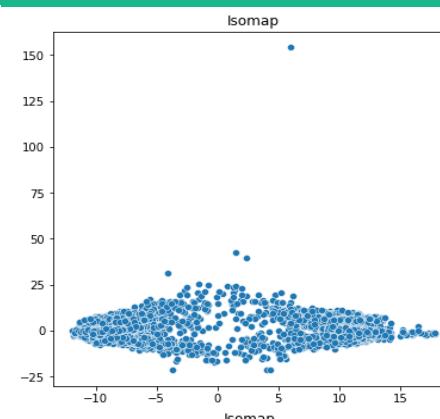
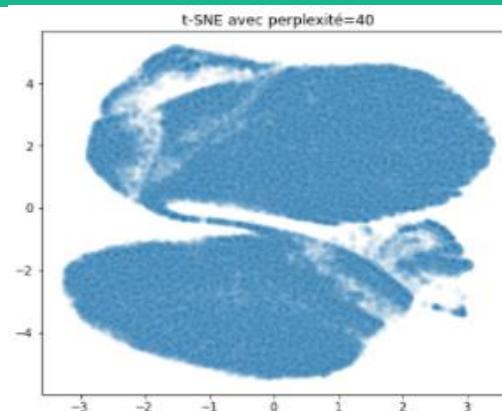
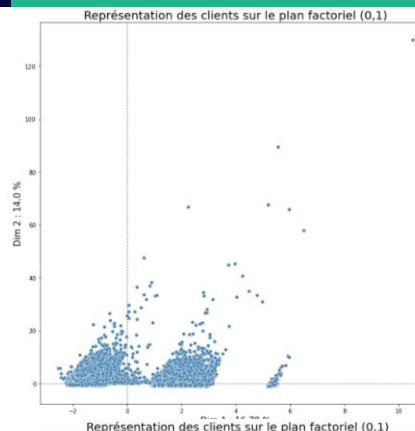
Type client	Attentions	Actions
Clients sur le point de dormir	A stimuler	<p><b>Partagez des ressources précieuses</b> (utilité des produits, liens utiles).</p> <p><b>Donnez votre analyse de la concurrence</b> (meilleurs que la concurrence?).</p> <p><b>Fournissez des mises à jour pertinentes</b> (contenu éducatif sur les produits).</p>
Clients en danger	A stimuler	<p><b>Offrez un crédit</b> (montant de crédit d'achat).</p> <p><b>Proposez une liste de souhaits</b> (organiser une vente en fonction des souhaits).</p> <p><b>Offres de mise à niveau</b> (gratuitement).</p>
VIPs à ne pas perdre	A stimuler	<p><b>Services sur mesure</b> (cibler les clients).</p> <p><b>Passez un appel téléphonique</b> (être à l'écoute).</p> <p><b>Connectez-vous sur les médias sociaux</b> (sur leurs réseaux).</p>
VIPs/Clients en hibernation	Choix	<p><b>Décidez si vous voulez qu'il revienne.</b></p> <p><b>Passez en revue votre produit</b> (consulter les produits).</p> <p><b>Envoyez une campagne personnalisée</b> (codes promos personnels).</p>
VIPs/Clients perdus	Choix	<p><b>Comprenez-les</b> (faites des recherches sur les clients perdus).</p> <p><b>Faites une dernière promotion</b> (e-mail de promotion).</p> <p><b>Ignorez-les complètement</b> (<i>ne reviendront pas donc perte d'argent</i>).</p>

## Pré Processing

## Réduction de dimension

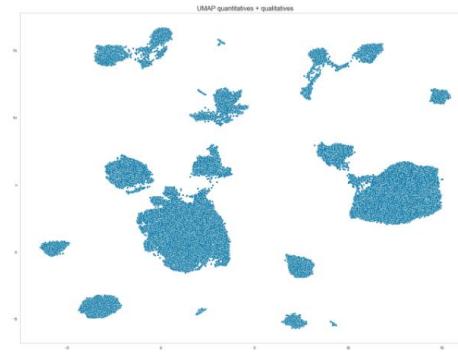
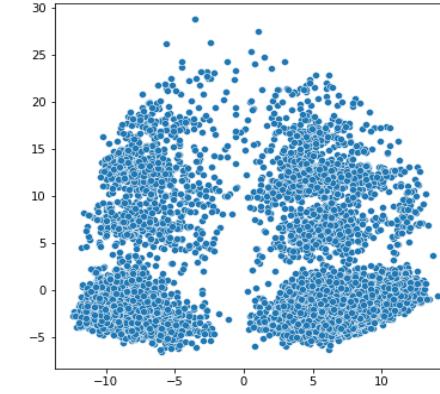
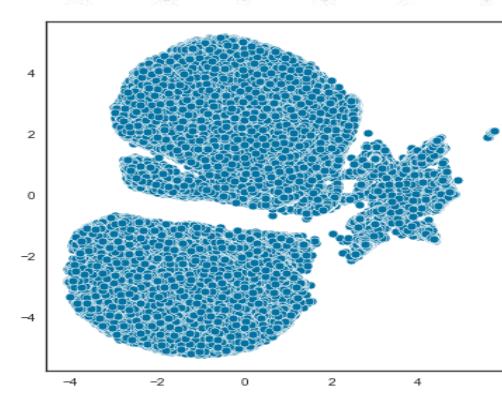
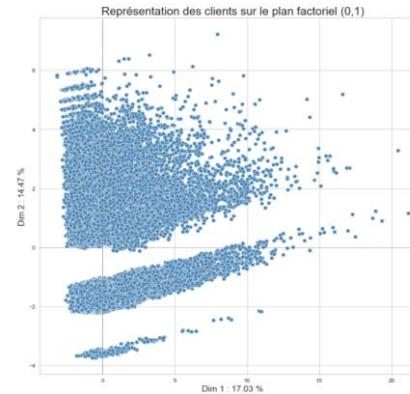
Quantitatives

StandardScaler()

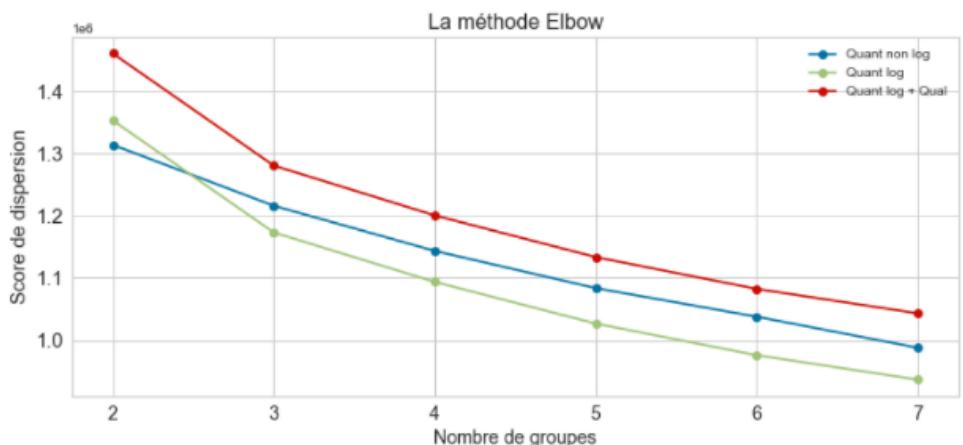


Quantitatives + Qualitatives

StandardScaler() pd.getDummies  
np.log()

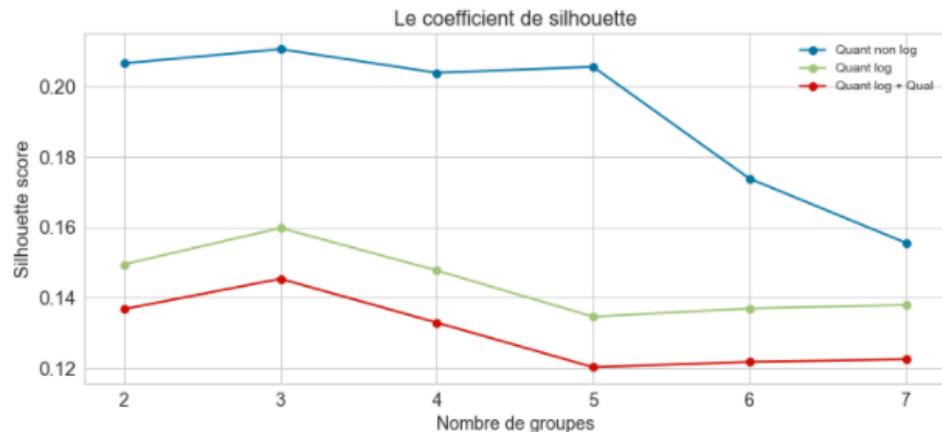


Modèle	Hyperparamètre	Grille de recherche	Meilleure performance
K-Means	n_clusters	[3, 4, 5, 6]	5, 6
	n_init	[10, 20, 30, 50, 70]	101
	init	['k-means++', 'random']	1
K-Prototype	n_clusters	6	6
	init	'Cao'	Cao
CAH	n_clusters	range(3, 10)	3
	linkage	['complete', 'average', 'single']	Single
	affinity	['euclidean', 'manhattan', 'cosine']	euclidean
DBSCAN	eps	[1, 1.5, 2, 2.5, 4]	3
	min_samples	[10, 20, 30, 40, 50]	10
HDBSCAN	min_samples	[1, 2, 3, 4, 5, 10]	1
	min_cluster_size	[300, 1000, 2000, 3000, 4000]	4000
OPTICS	min_samples	[1, 3, 5, 10]	1*
	xi	[0.05, 0.1, 0.2]	0,2*
	min_cluster_size	[0.01, 0.05]	0,01*
GMM	n_components	range(2,20)	2
	covariance_type	['full', 'spherical']	full



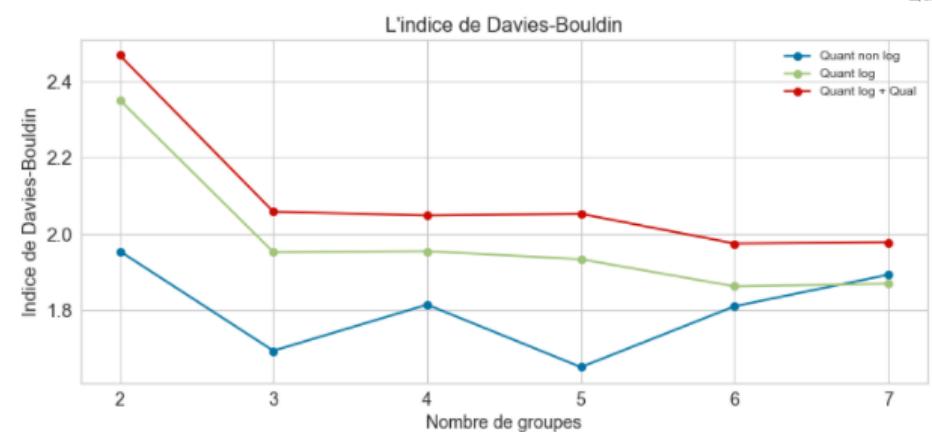
- Méthode Elbow :**

- Quantitatives non log : 3 ou 4 clusters.
- Quantitatives log : 3 ou 4 ou 5 clusters, difficile à interpréter.
- Quantitatives + qualitatives : 3 ou 6



- Coefficient de silhouette :**

- Quantitatives non log : 4 ou 6 clusters.
- Quantitatives log : 4 ou 5 ou 6.
- Quantitatives + qualitatives : 5 ou 6.



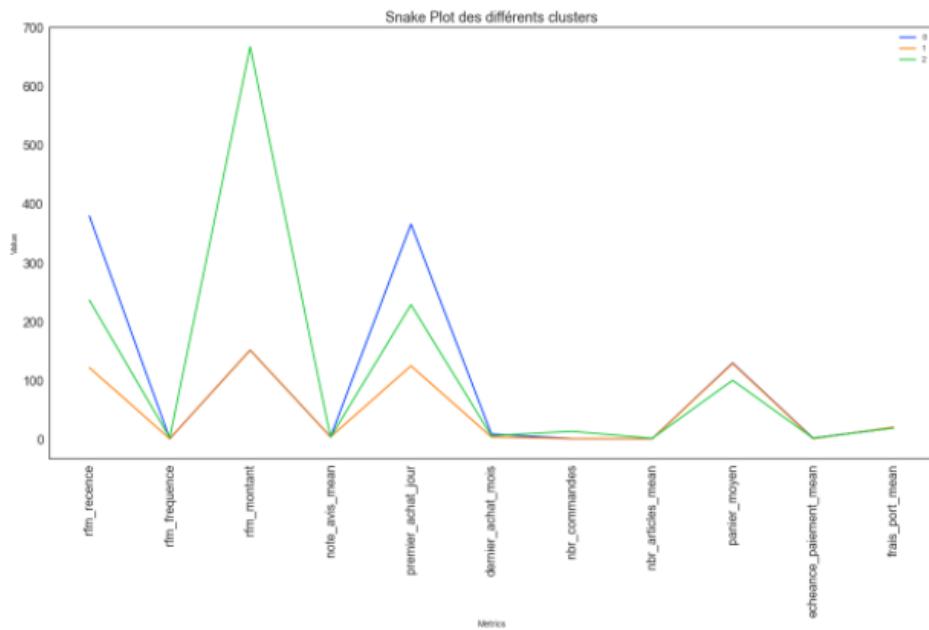
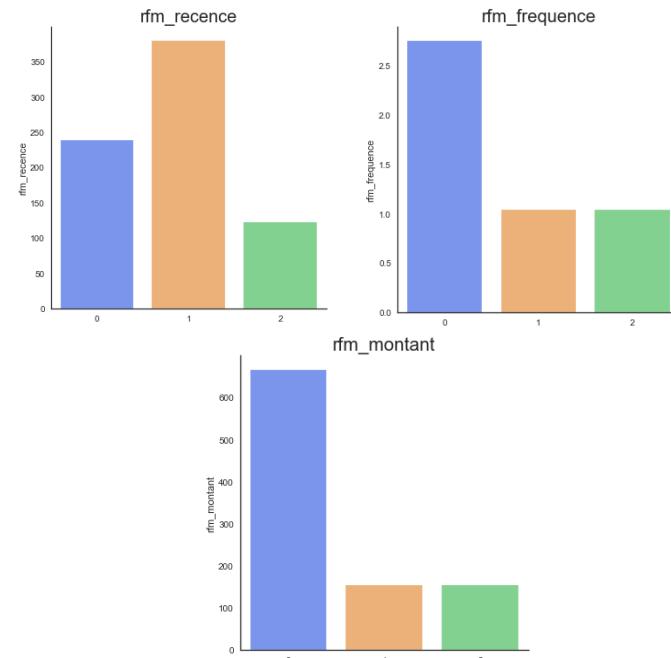
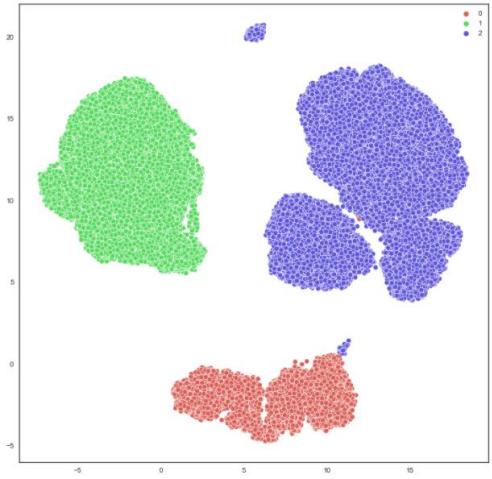
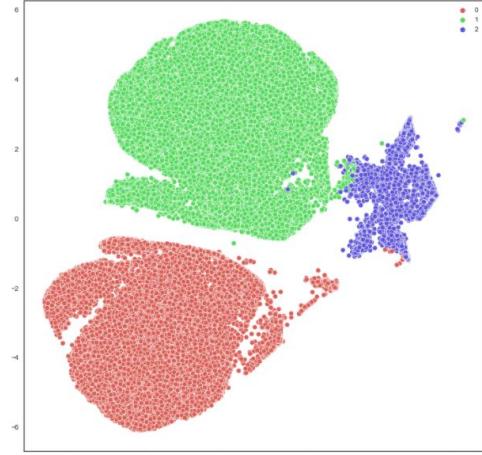
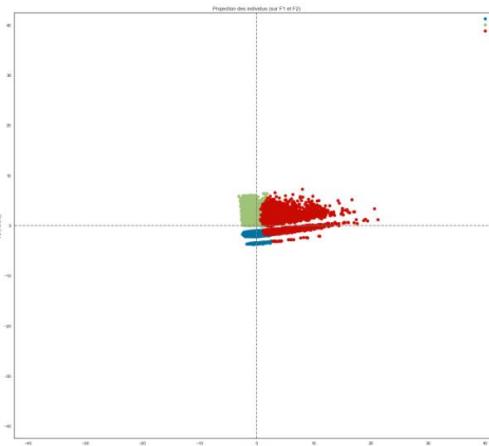
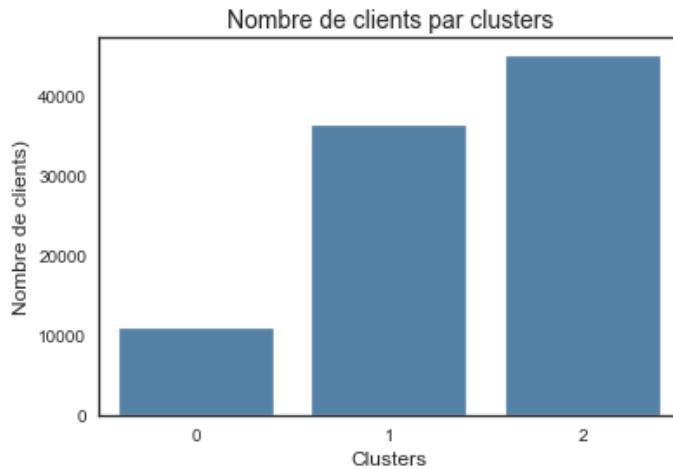
- Indice de Davies-Bouldin :**

- Quantitatives non log : 3 ou 4 ou 5 clusters.
- Quantitatives log : 3 ou 4 ou 5 clusters.
- Quantitatives + qualitatives : 5 ou 6.

# F Annexes – K-Means k=3



O list



**Cluster 0 :** commandes anciennes, peu fréquentes et d'un faible montant, dernier achat très ancien

**==> clients presque perdus.**

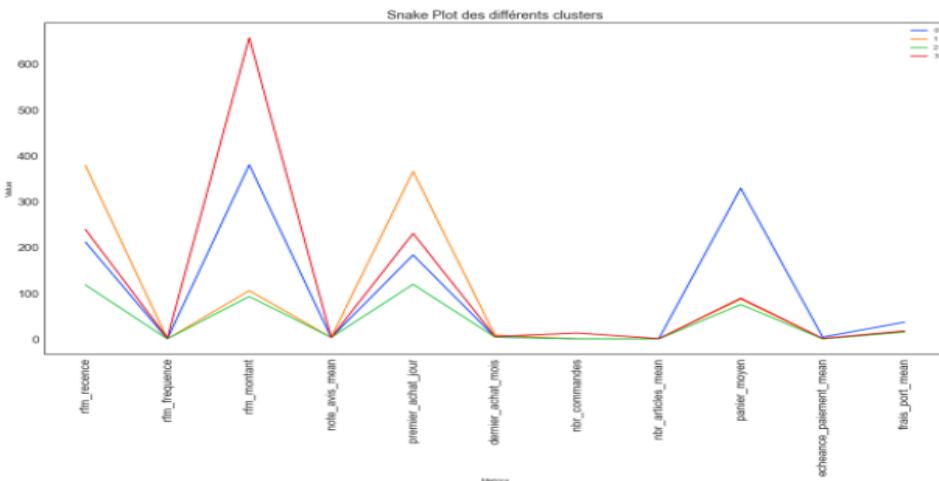
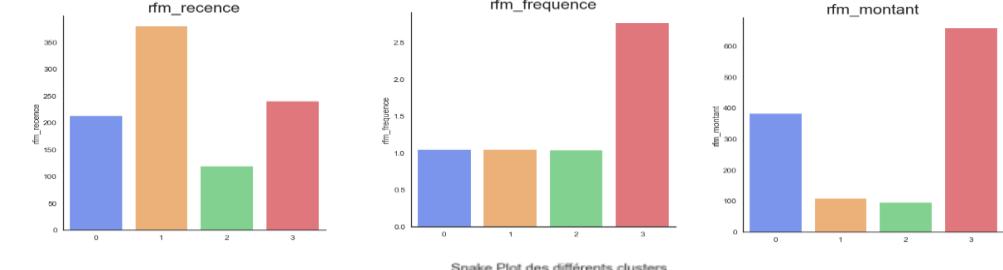
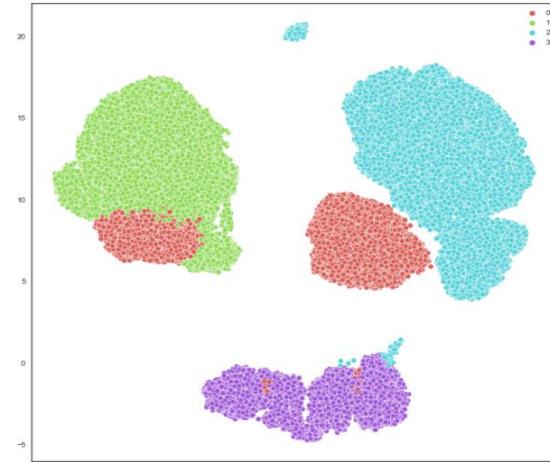
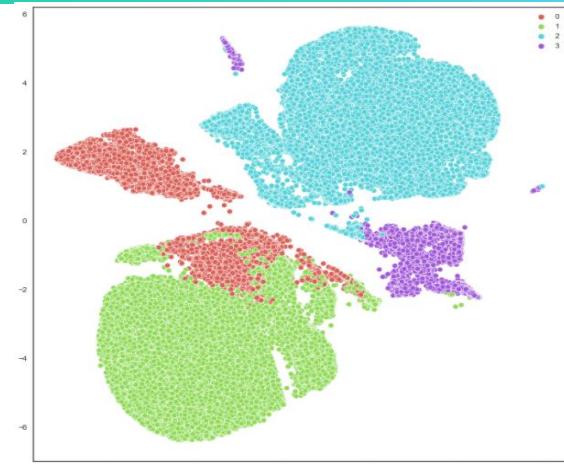
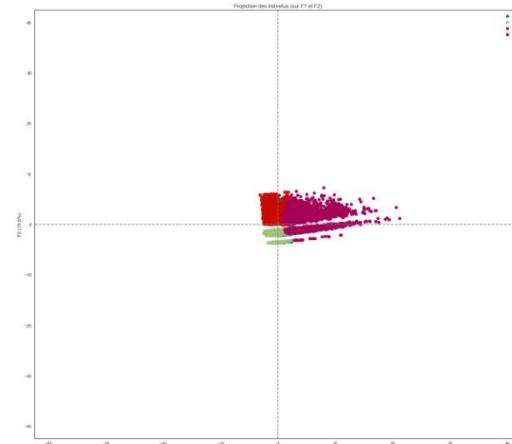
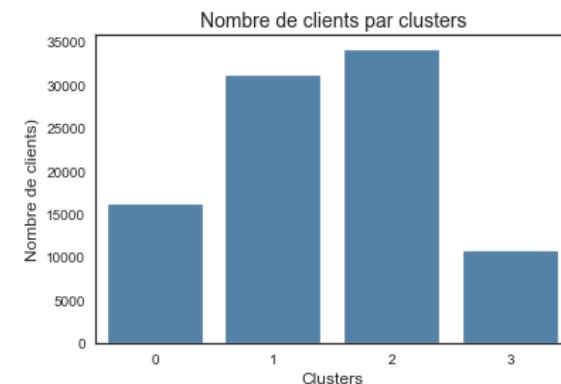
**Cluster 1 :** commandes très récentes, peu fréquentes et d'un faible montant

**==> nouveaux clients.**

**Cluster 2 :** commandes récentes, fréquentes et d'un bon montant

**==> meilleurs clients.**

# F Annexes – K-Means k=4



**Cluster 0** : Commandes moins récentes, moyennement fréquentes avec une bonne moyenne de montant

**==> Bon client sur le point de dormir.**

**Cluster 1** : commandes très anciennes, moyennement fréquentes et d'un montant correct, dernier achat très ancien

**==> Clients en hibernation ou perdus.**

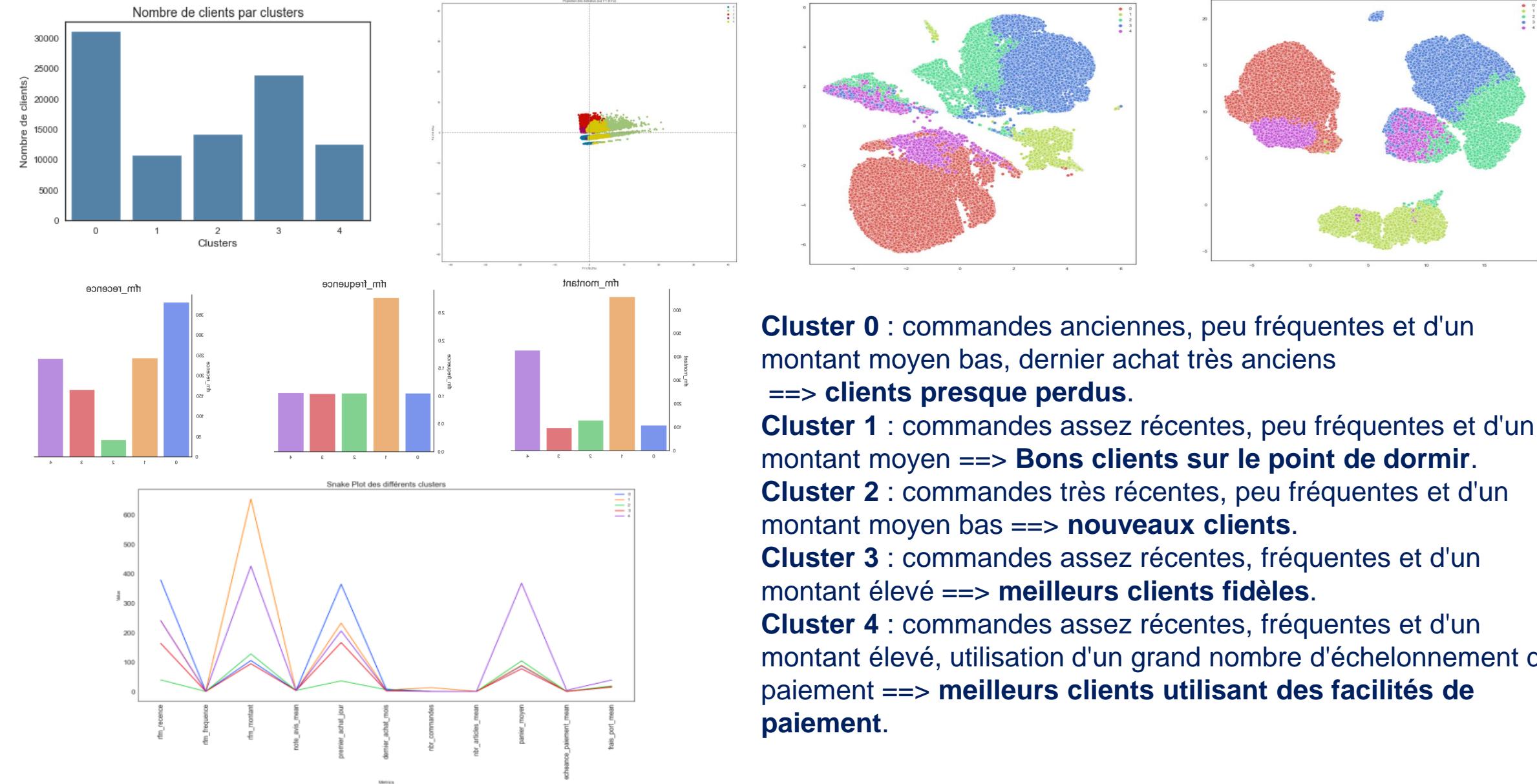
**Cluster 2** : commandes très récentes, peu fréquentes et d'un faible montant

**==> nouveaux clients.**

**Cluster 3** : commandes récentes, fréquentes et d'un bon montant

**==> meilleurs clients.**

# F Annexes – K-Means k=5



**Cluster 0 :** commandes anciennes, peu fréquentes et d'un montant moyen bas, dernier achat très ancien  
**==> clients presque perdus.**

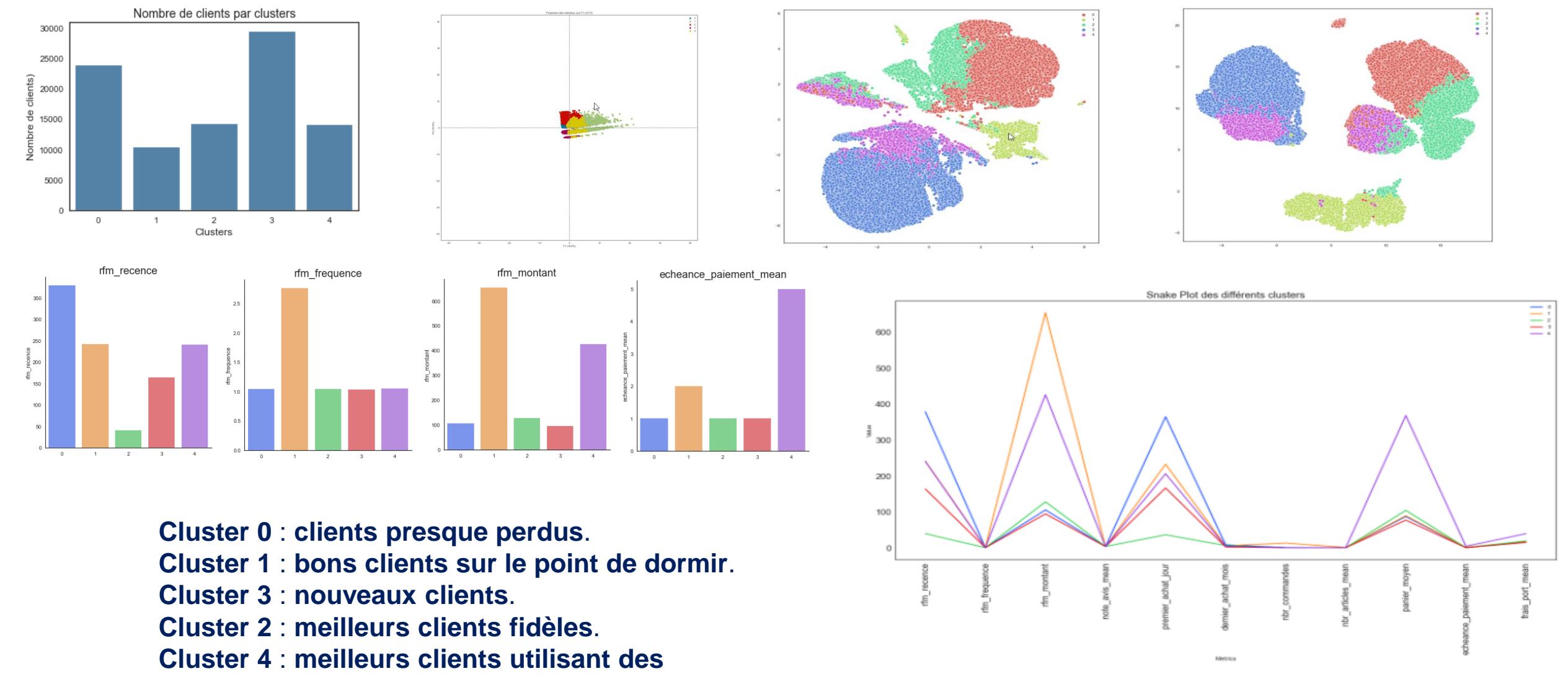
**Cluster 1 :** commandes assez récentes, peu fréquentes et d'un montant moyen ==> **Bons clients sur le point de dormir.**

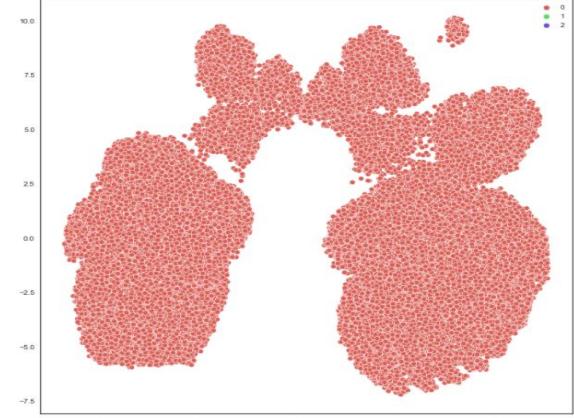
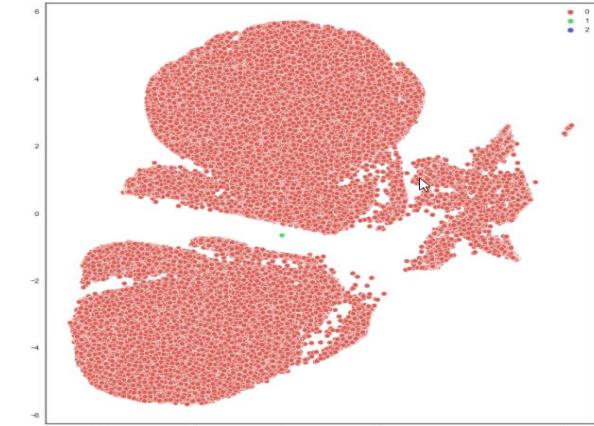
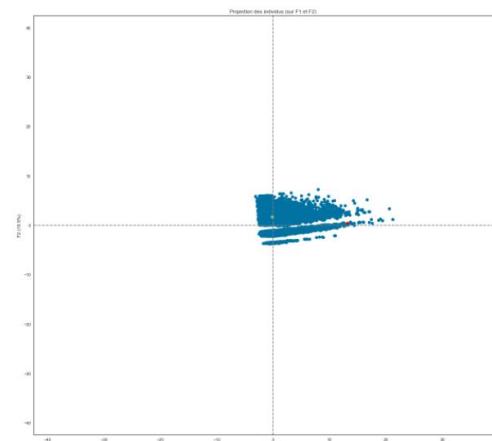
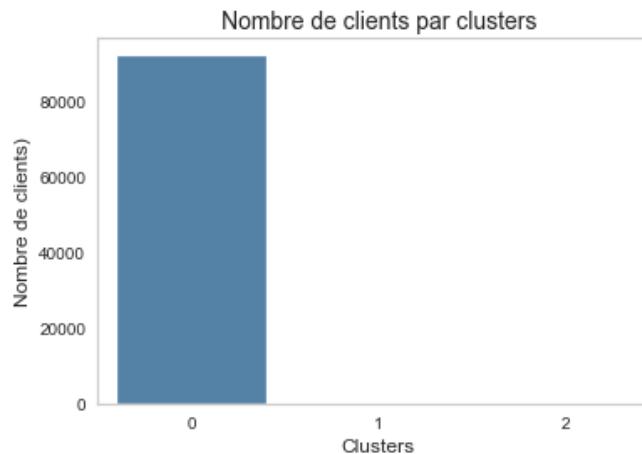
**Cluster 2 :** commandes très récentes, peu fréquentes et d'un montant moyen bas ==> **nouveaux clients.**

**Cluster 3 :** commandes assez récentes, fréquentes et d'un montant élevé ==> **meilleurs clients fidèles.**

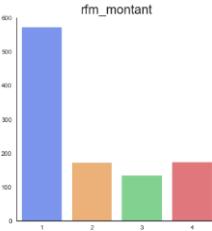
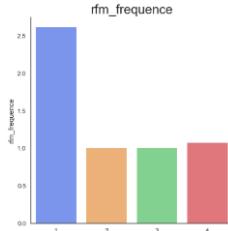
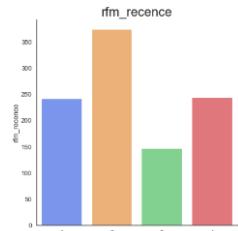
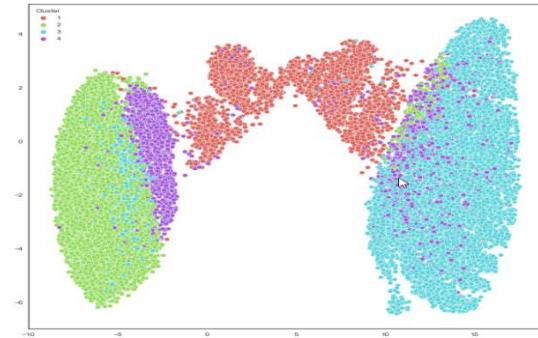
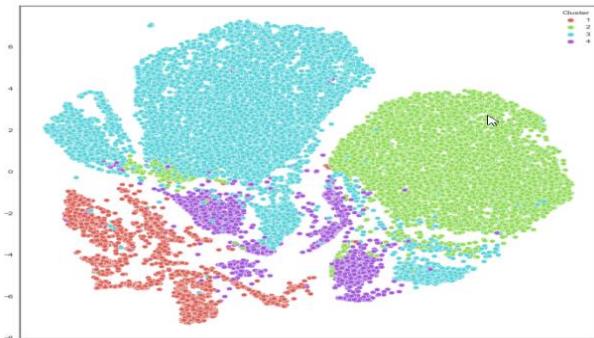
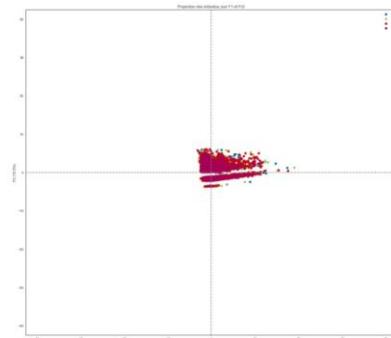
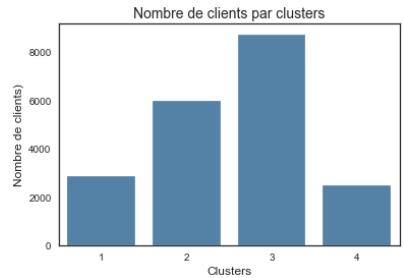
**Cluster 4 :** commandes assez récentes, fréquentes et d'un montant élevé, utilisation d'un grand nombre d'échelonnement de paiement ==> **meilleurs clients utilisant des facilités de paiement.**

# F Annexes – ACP + K-Means k=5

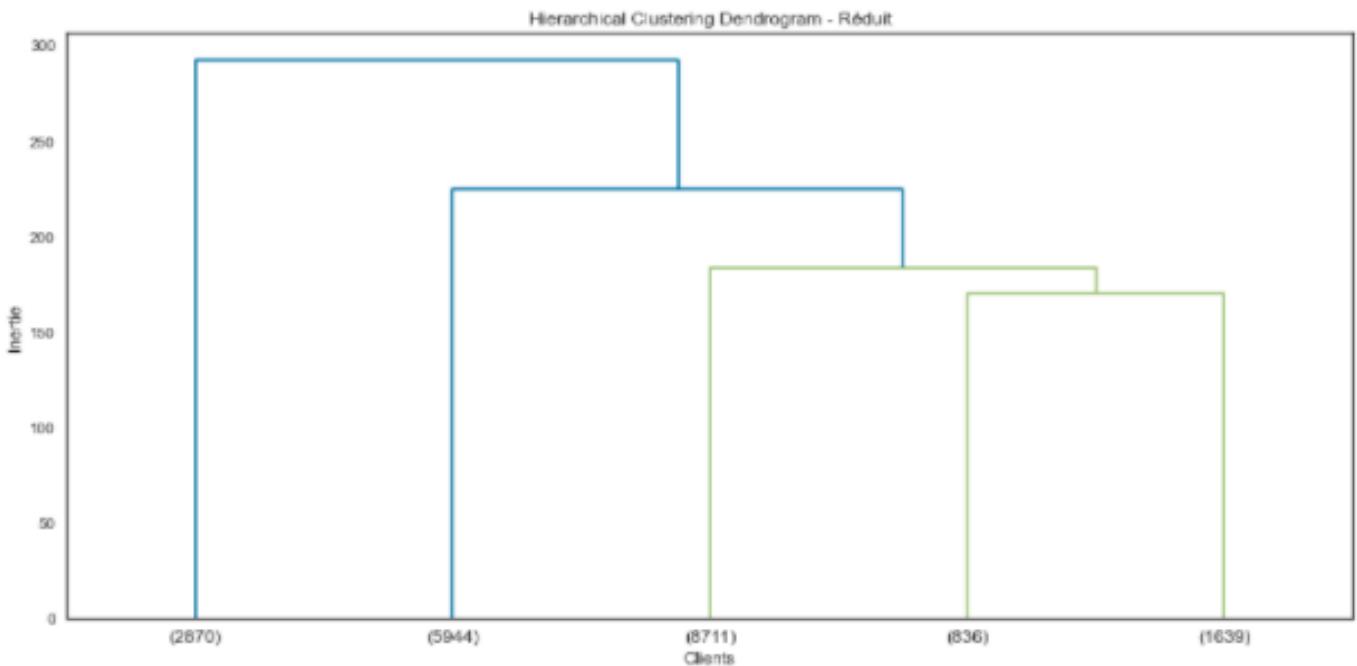


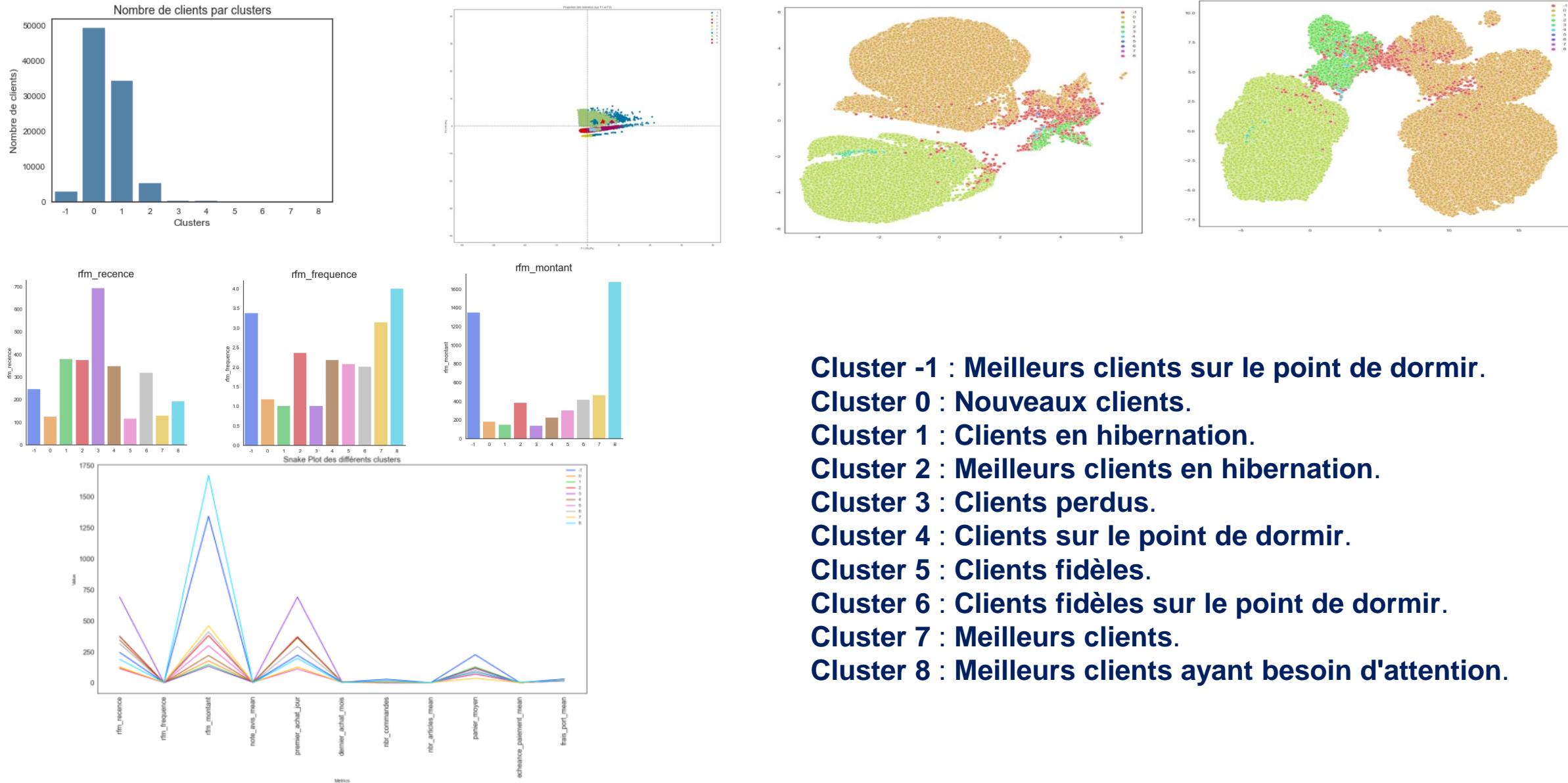


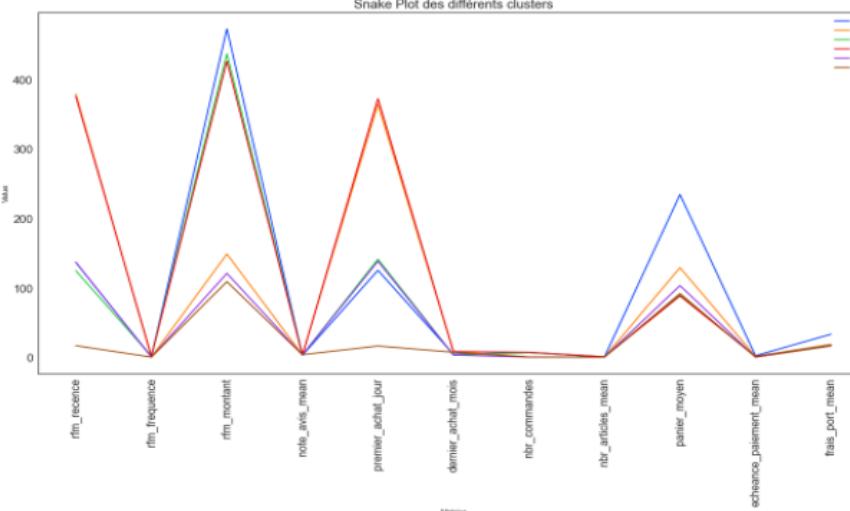
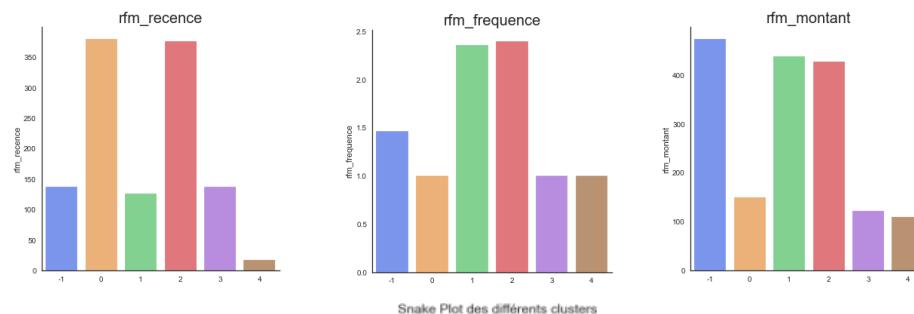
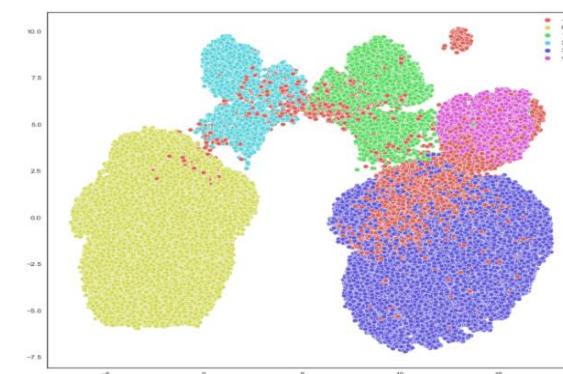
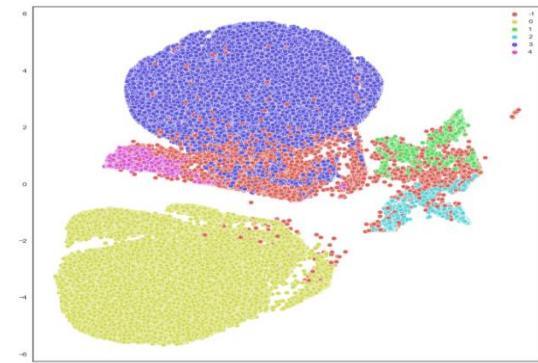
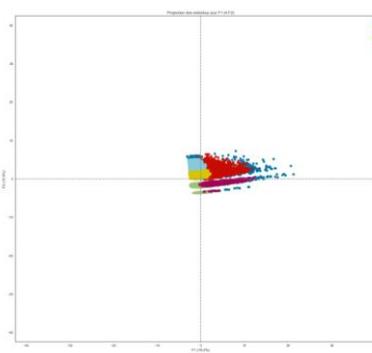
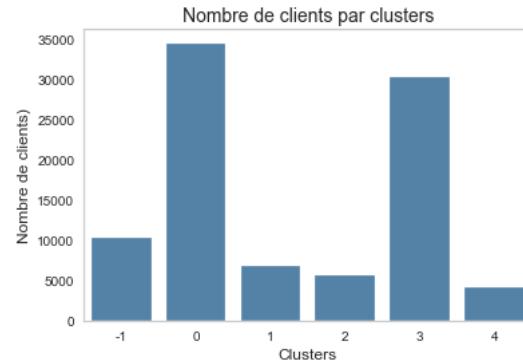
Tous les clients sont regroupés en 1 seul cluster  
→ segmentation impossible



**Cluster 1 : Meilleurs clients.**  
**Cluster 2 : Clients perdus.**  
**Cluster 3 : Nouveaux clients.**  
**Cluster 4 : Clients fidèles.**







**Cluster -1 : Meilleurs clients/Clients fidèles ayant un besoin d'attention.**

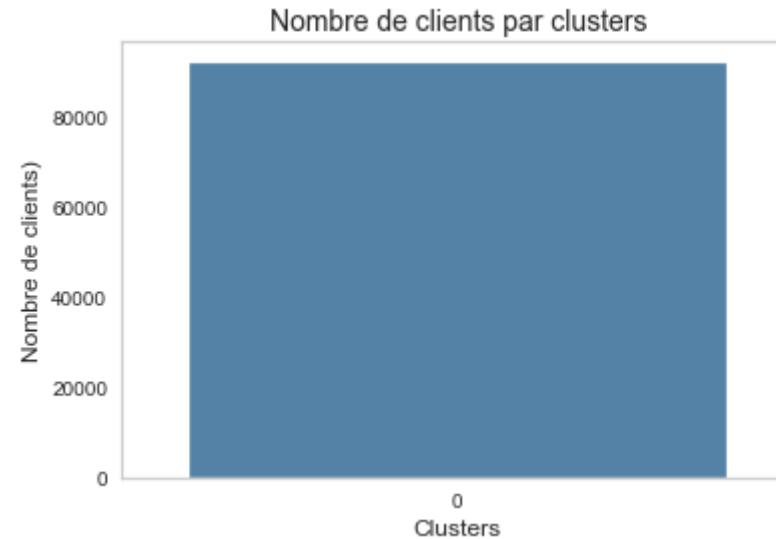
**Cluster 0 : Clients perdus.**

**Cluster 1 : Meilleurs clients fidèles.**

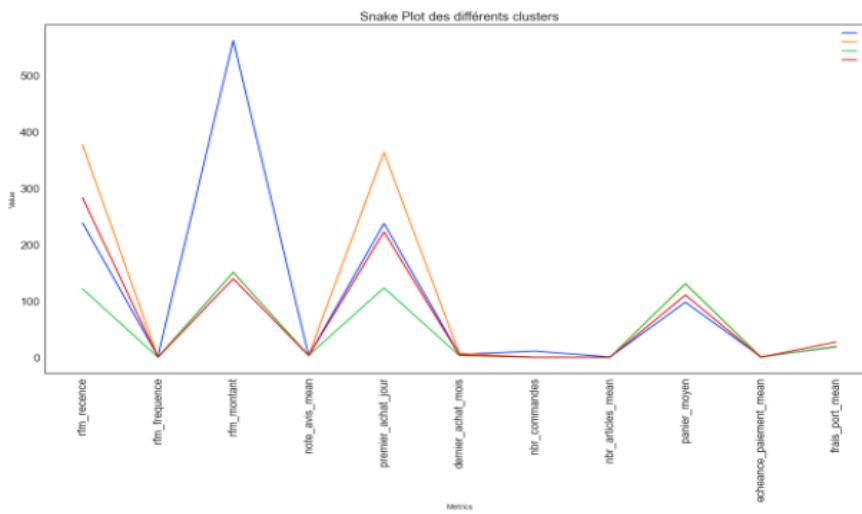
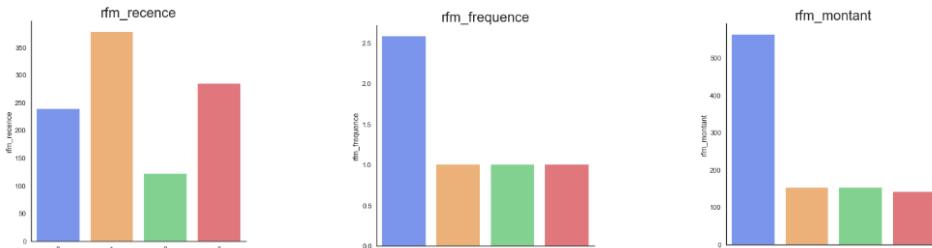
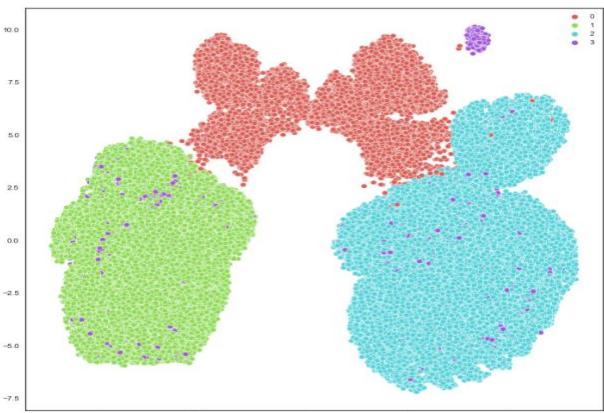
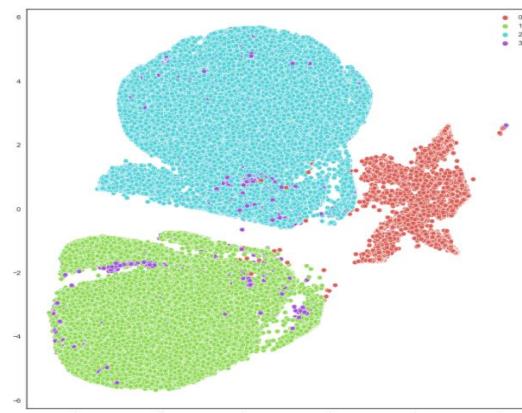
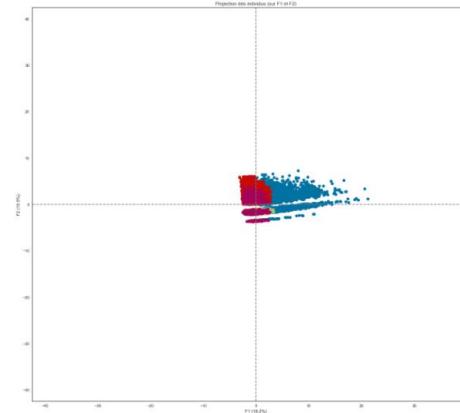
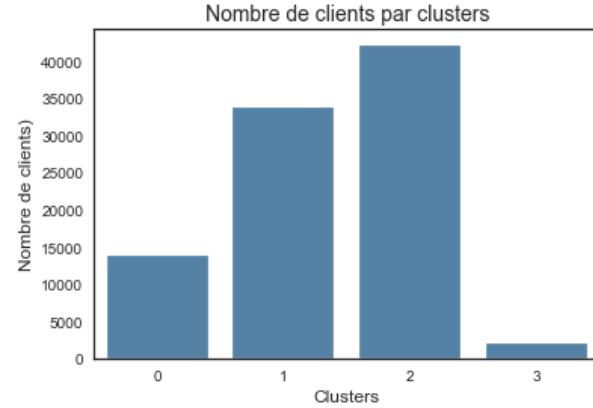
**Cluster 2 : Meilleurs clients perdus.**

**Cluster 3 : Clients ayant un besoin d'attention.**

**Cluster 4 : Nouveaux clients.**



Tous les clients sont regroupés en 1 seul cluster  
→ segmentation impossible



**Cluster 0 : meilleurs clients.**  
**Cluster 1 : clients perdus.**  
**Cluster 2 : nouveaux clients**  
**Cluster 3 : clients fidèles.**