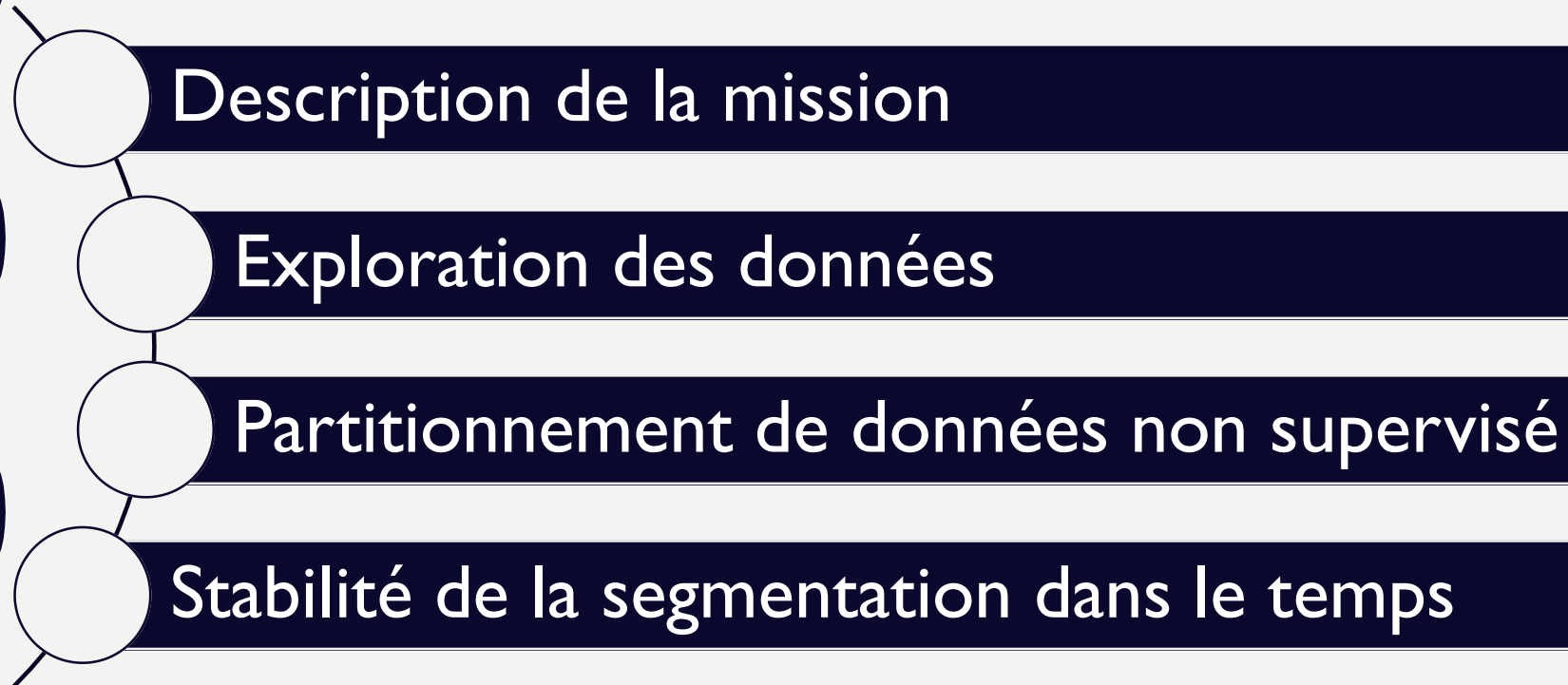


**PROJET 5 :  
SEGMENTEZ DES  
CLIENTS D'UN  
SITE E-  
COMMERCE**

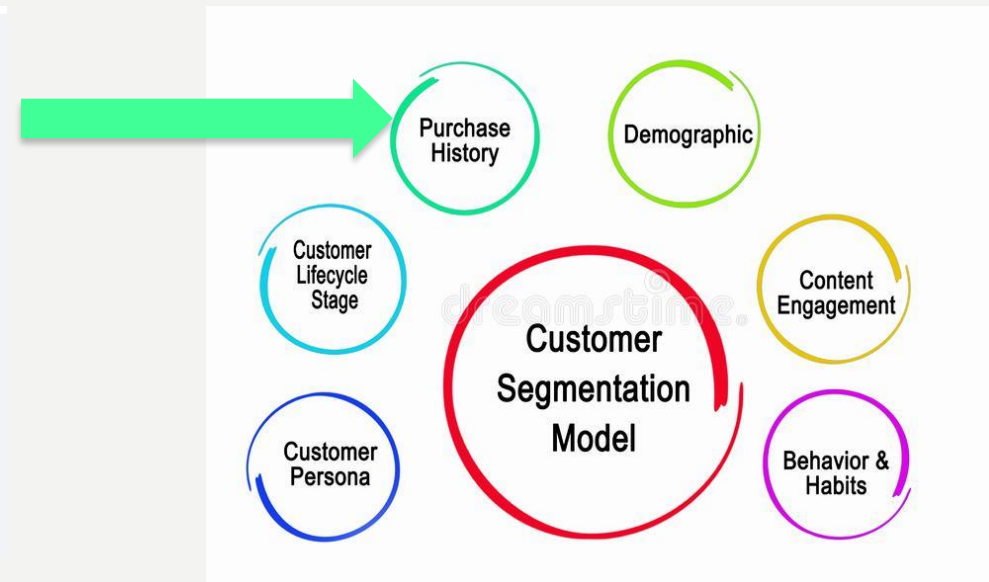
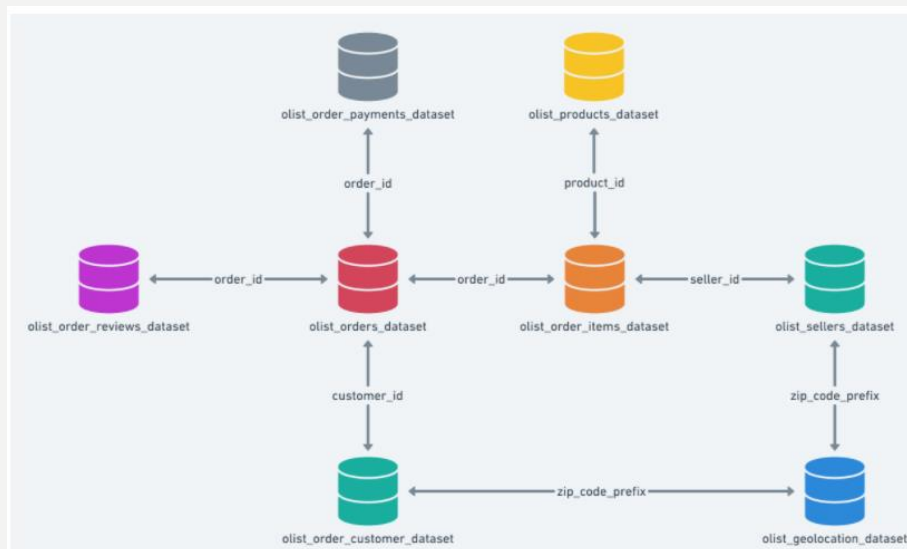
*XIAOFAN LEI*

# ORDRE DU JOUR

- 
- Description de la mission
  - Exploration des données
  - Partitionnement de données non supervisé
  - Stabilité de la segmentation dans le temps

# DESCRIPTION DE LA MISSION

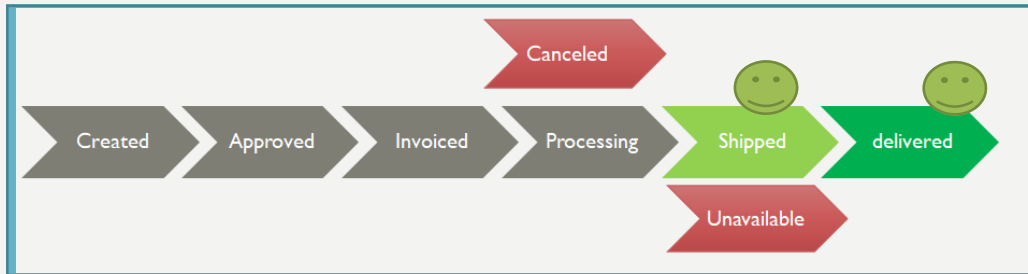
- Objectif
  - mieux connaître la clientèle d'Olist et d'être en mesure de fournir le support de travail en vu des campagnes de communication.
- Critères de segmentation
  - répartir la clientèle en catégories distinctes selon des critères comportementaux d'achat



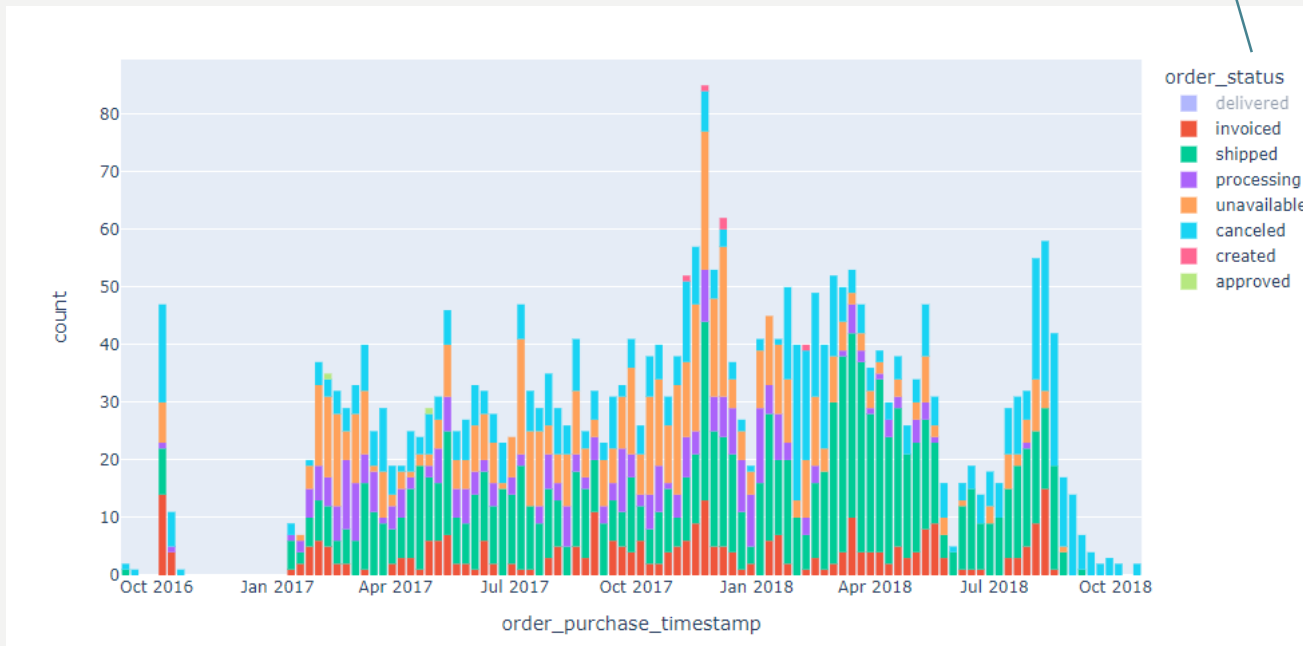


# **EXPLORATION DES DONNÉES**

# EXPLORATION DES DONNÉES : STATUT DE LA COMMANDE

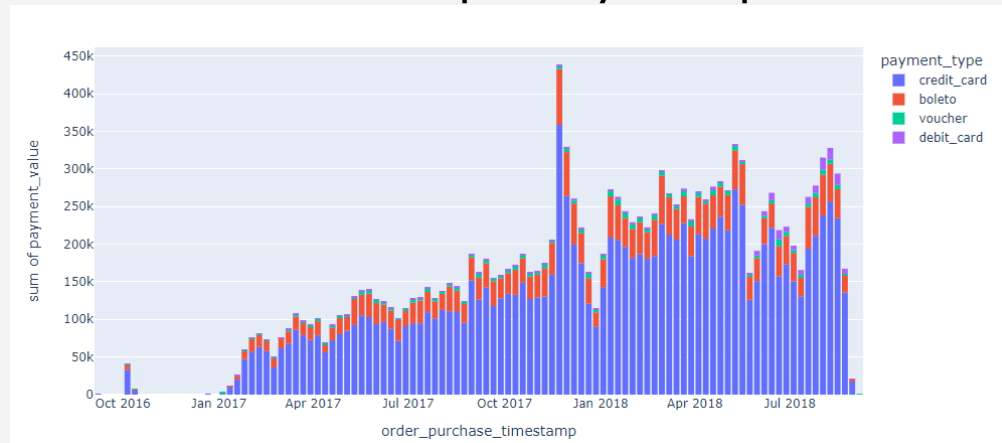


- Historique de deux ans d'achats
- Pas changement de statut automatique
  - Nous pouvons déduire que le statut reste en « shipped » car le client ne confirme pas la réception du colis, mais sans réclamation de la part du client, la commande peut être considéré comme terminée
  - La segmentation sera faite sur les clients ayant finalisé leur achat, c'est-à-dire en statut « shipped » et « delivered »

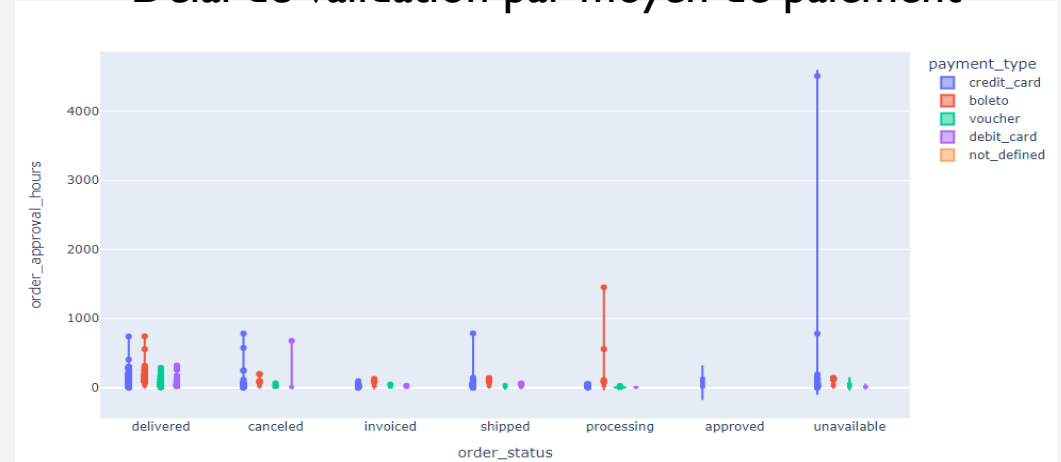


# EXPLORATION DES DONNÉES : MOYEN DE PAIEMENT

Chiffres d'affaires par moyen de paiement

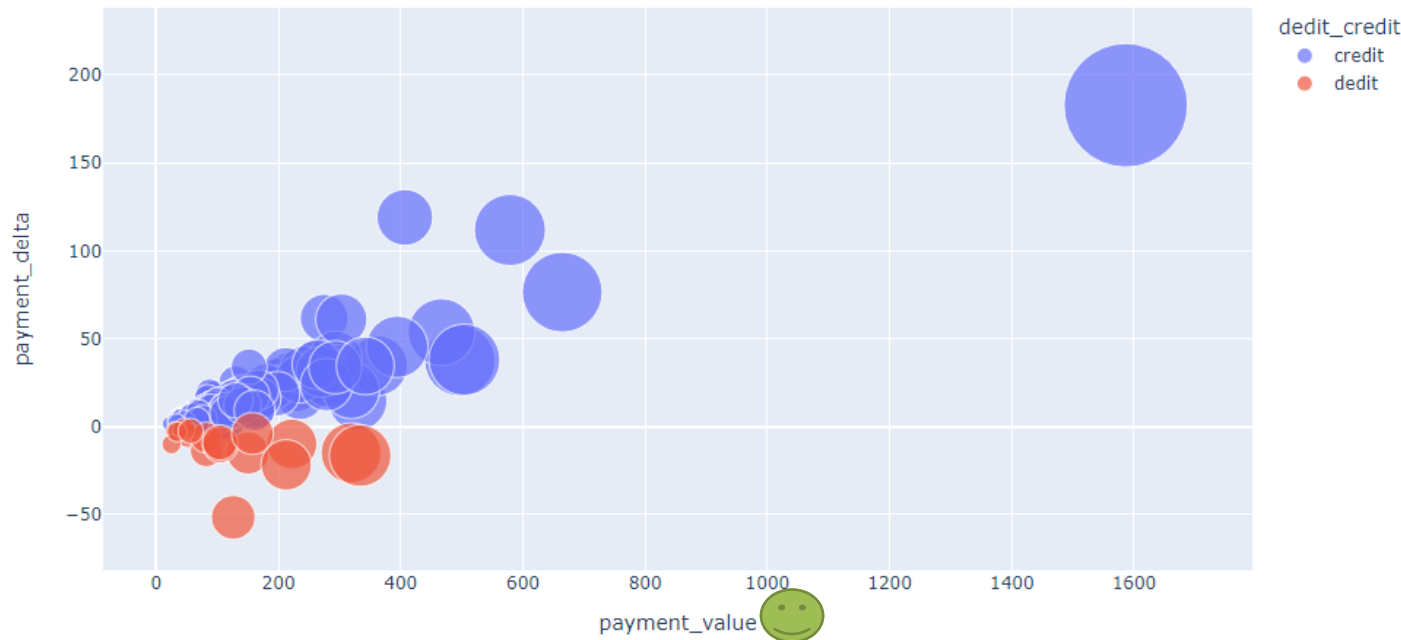


Délai de validation par moyen de paiement



- Une baisse de chiffres d'affaires considérable en janvier 2018
- La plupart des achats sont réalisés par carte de crédit
- La validation est relativement longue en cas d'indisponibilité des produits

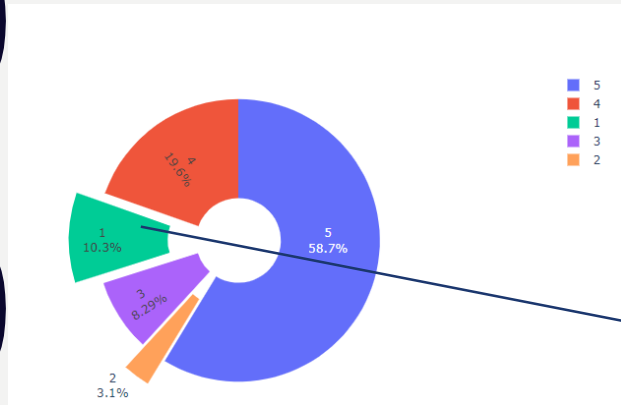
# EXPLORATION DES DONNÉES : CONTRÔLE DU PAIEMENT



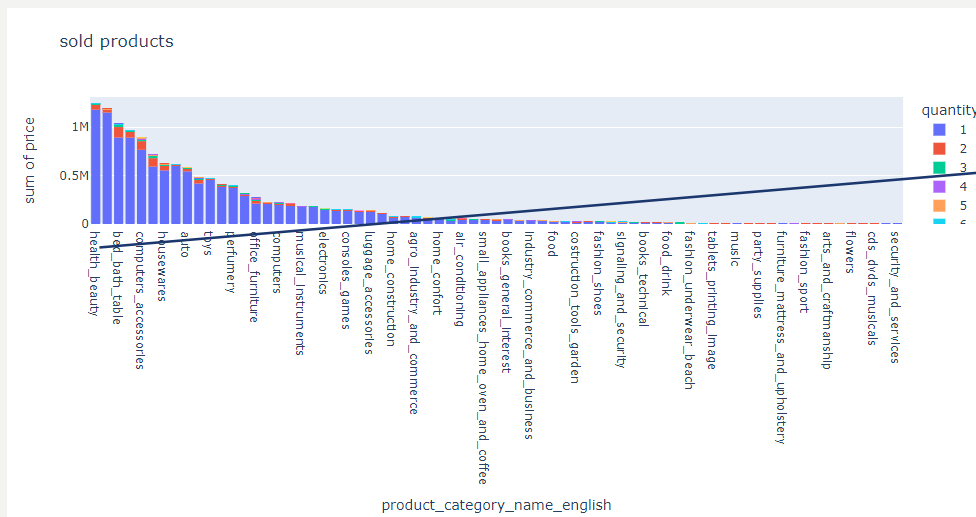
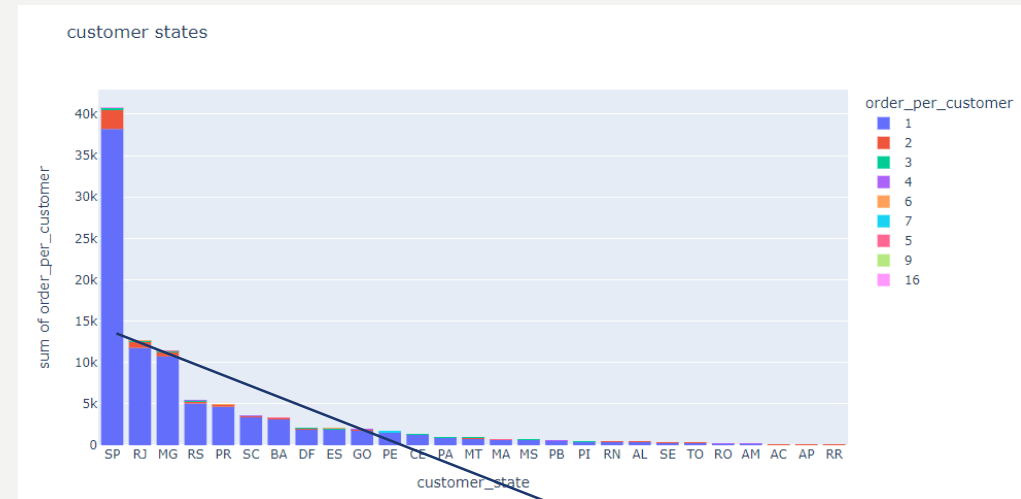
- Anomalies

- Parmi les commandes validées, 247 montants payés ne correspondent pas à la somme due (=le prix du produit + le frais de port)
- Le montant payé par le client sera prise en compte dans la segmentation des clients

# EXPLORATION DES DONNÉES : AUTRES CARACTÉRISTIQUES



- 13% de clients mécontents (notes 1 et 2)



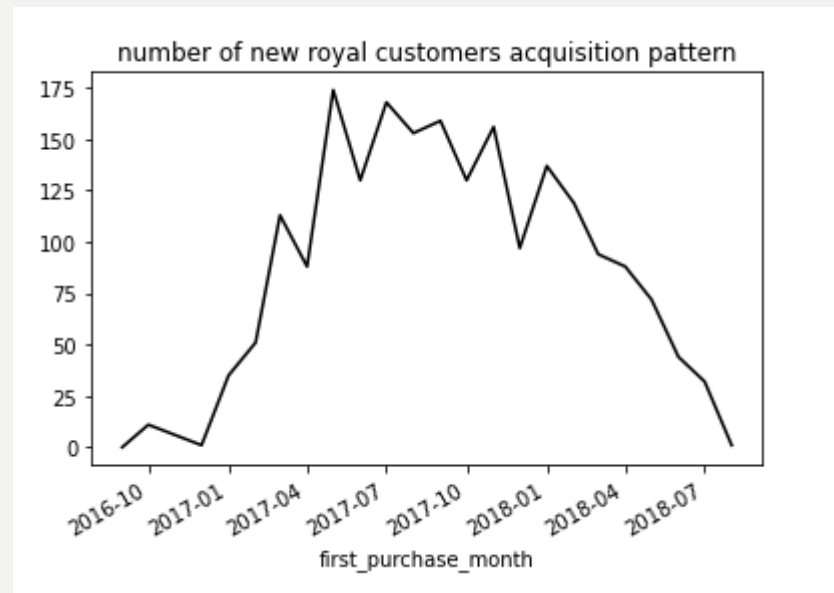
- Les produits de beauté et de santé sont les plus populaires

- La majorité des clients habitent dans l'état Sao Paolo



# ANALYSE DES COHORTES : RÉTENTION DES CLIENTS

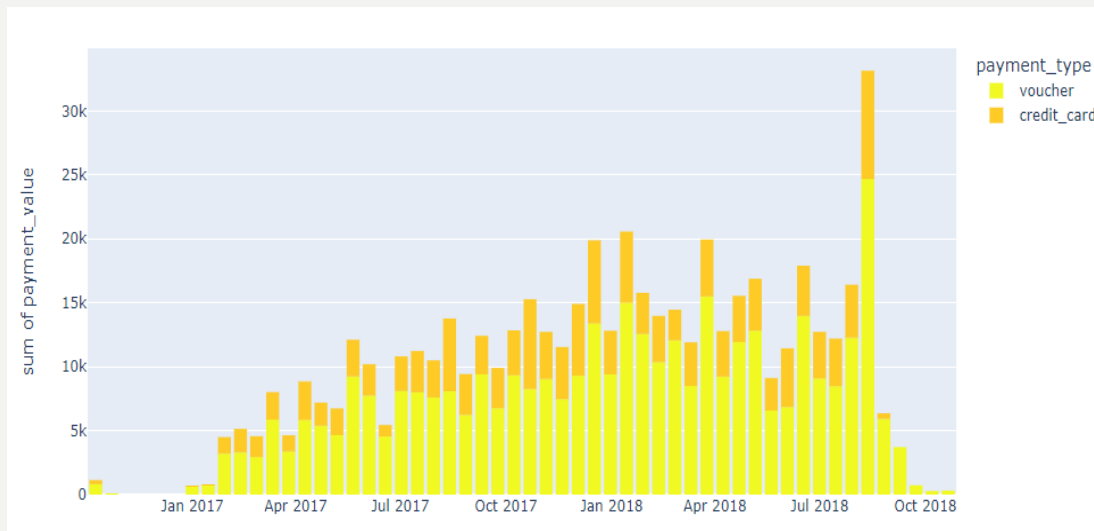
diff_month	customer_unique_id																			
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19	20
first_purchase_month																				
2016-09-01	2.0																			
2016-10-01	291.0						1			1		2		1		1		1	2	2
2016-12-01	1.0	1																		
2017-01-01	814.0	3	2	1	3	1	3	1	1		3	1	5	3	1	1	2	3	1	
2017-02-01	1770.0	4	5	2	7	2	4	3	2	3	2	5	2	3	2	1	1	3		
2017-03-01	2739.0	14	9	16	9	4	4	9	8	2	10	3	5	3	4	7	2	4		
2017-04-01	2495.0	15	6	5	8	6	11	8	7	4	6	2	1	1	2	2	4			
2017-05-01	3814.0	19	19	13	11	16	17	6	11	11	9	12	10	1	8	11				
2017-06-01	3321.0	16	12	12	9	11	13	7	5	8	10	11	5	4	7					
2017-07-01	4151.0	43	19	9	14	8	16	4	8	11	8	13	5	10						
2017-08-01	4350.0	30	16	11	14	24	12	11	6	6	10	8	5							
2017-09-01	4321.0	30	23	16	22	14	9	10	12	7	13	3								
2017-10-01	4630.0	36	11	4	11	9	9	17	15	9	9									
2017-11-01	7530.0	46	32	12	14	16	9	14	9	4										
2017-12-01	5663.0	12	16	21	15	12	9	1	11											
2018-01-01	7277.0	24	26	22	21	14	13	17												
2018-02-01	6710.0	24	28	20	19	14	14													
2018-03-01	7284.0	29	24	22	10	9														
2018-04-01	6992.0	39	22	16	11															
2018-05-01	6857.0	40	17	15																
2018-06-01	6188.0	28	16																	
2018-07-01	6270.0	32																		
2018-08-01	6400.0	1																		



- De manière générale, malgré le nombre important de nouveaux arrivants chaque mois, le taux de rétention des clients est très bas.

# ANALYSE DES COHORTES : BON D'ACHAT

## Chiffres d'affaire réalisés avec bon de réduction



- Le bon d'achat ne permet pas de faire revenir les clients.

[illegible]

# VUE GLOBALE DU DATASET

	count	mean	std	min	25%	50%	75%	max	na	na%
frequency	94399.0	1.0797	0.4576	1.0000	1.0000	1.0000	1.0000	33.0000	0	0.00%
recency	94399.0	-286.8575	152.7221	-772.8437	-395.7211	-267.7896	-163.1509	-44.3495	0	0.00%
order_approval_hours	94386.0	10.3077	20.6102	0.0000	0.2164	0.3472	14.6288	784.0456	13	0.01%
order_delivery_days	93349.0	9.3426	8.7651	-16.0962	4.1153	7.1060	12.0334	205.1910	1050	1.11%
order_delivery_delay	93350.0	-11.1509	10.1402	-146.0161	-16.2272	-11.7548	-6.3936	188.9751	1049	1.11%
diff_month	94399.0	0.0878	0.8294	0.0000	0.0000	0.0000	0.0000	20.0000	0	0.00%
monetary	94398.0	165.2543	226.1368	9.5900	63.1000	107.7800	182.6100	13664.0800	1	0.00%
payment_type_boleto	94398.0	0.2055	0.4181	0.0000	0.0000	0.0000	0.0000	6.0000	1	0.00%
payment_type_credit_card	94398.0	0.7991	0.4741	0.0000	1.0000	1.0000	1.0000	16.0000	1	0.00%
payment_type_debit_card	94398.0	0.0160	0.1267	0.0000	0.0000	0.0000	0.0000	2.0000	1	0.00%
payment_type_not_defined	94398.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1	0.00%
payment_type_voucher	94398.0	0.0591	0.4420	0.0000	0.0000	0.0000	0.0000	33.0000	1	0.00%
payment_sequential	94398.0	1.0805	0.4585	1.0000	1.0000	1.0000	1.0000	33.0000	1	0.00%
payment_installments	94398.0	3.0284	2.9323	0.0000	1.0000	2.0000	4.0000	64.0000	1	0.00%
prod-bed_bath_table	94398.0	0.1173	0.4067	0.0000	0.0000	0.0000	0.0000	13.0000	1	0.00%
prod-health_beauty	94398.0	0.1015	0.3525	0.0000	0.0000	0.0000	0.0000	21.0000	1	0.00%
prod-sports_leisure	94398.0	0.0904	0.3393	0.0000	0.0000	0.0000	0.0000	11.0000	1	0.00%
prod-furniture_decor	94398.0	0.0873	0.3887	0.0000	0.0000	0.0000	0.0000	18.0000	1	0.00%
prod-computers_accessories	94398.0	0.0818	0.3528	0.0000	0.0000	0.0000	0.0000	20.0000	1	0.00%
prod-others	94398.0	0.6847	0.7045	0.0000	0.0000	1.0000	1.0000	24.0000	1	0.00%
quantity	94398.0	1.4124	2.3752	1.0000	1.0000	1.0000	1.0000	231.0000	1	0.00%
review_score	93726.0	4.1309	1.3006	1.0000	4.0000	5.0000	5.0000	5.0000	673	0.71%
geolocation_lat	94138.0	-21.1741	5.6430	-36.6054	-23.5882	-22.9253	-20.0995	42.1840	261	0.28%
geolocation_lng	94138.0	-46.1712	4.0753	-72.6667	-48.1113	-46.6302	-43.5867	-8.5779	261	0.28%

Commande et satisfaction

Livraison et autres

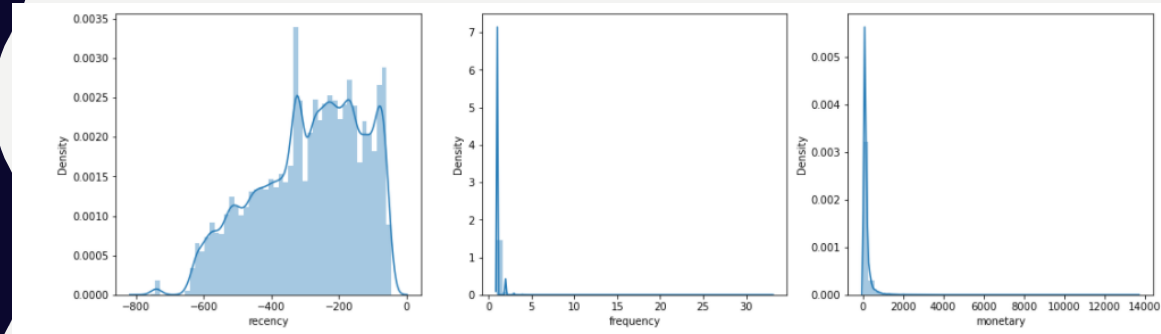
paiement

produit

géolocalisation

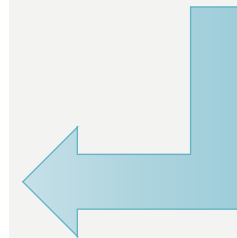
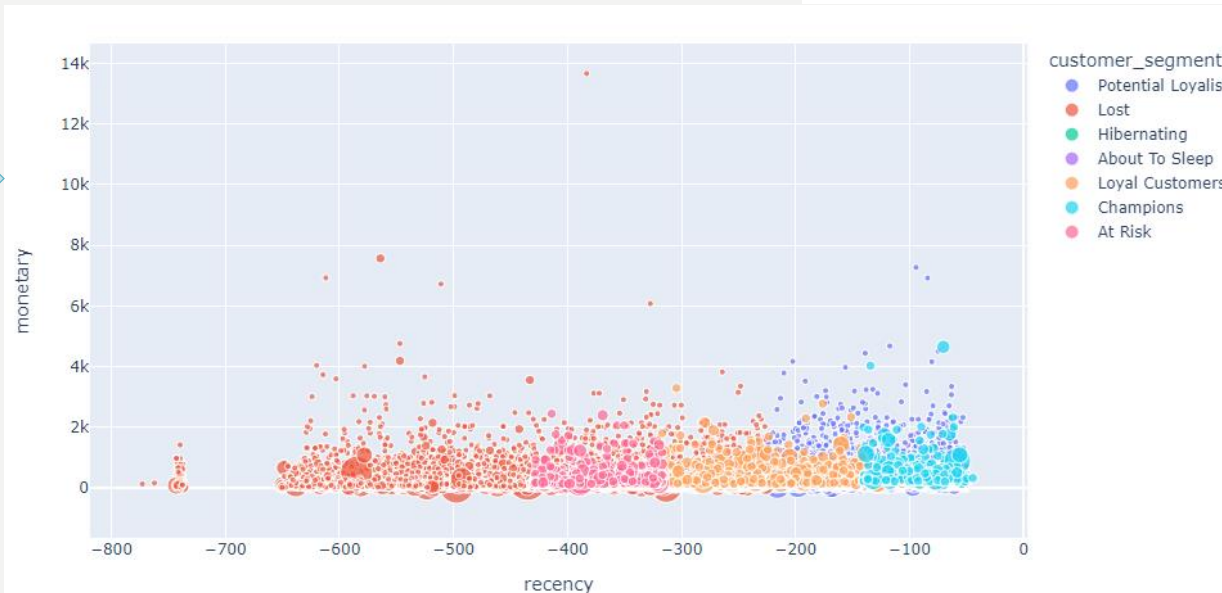
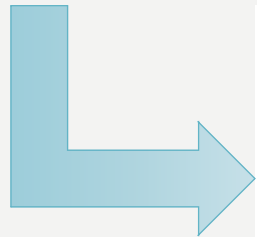
# EXEMPLE DE SEGMENTATION AVEC RFM À PARTIR DES VALEURS QUANTILES

Valeurs RFM de notre base de données



Définition manuelle de la segmentation

	customer_segment	activity	actionable_tip	r_range	fm_range
0	Champions	Bought recently, buy often and spend the most!	Reward them. Can be early adopters for new pro...	4-5	4-5
1	Loyal Customers	Spend good money with us often. Responsive to ...	Upsell higher value products. Ask for reviews....	2-5	3-5
2	Potential Loyalist	Recent customers, but spent a good amount and ...	Offer membership / loyalty program, recommend ...	3-5	0-5
3	Recent Customers	Bought most recently, but not often.	Provide on-boarding support, give them early s...	4-5	0-1
4	Promising	Recent shoppers, but haven't spent muc	Create brand awareness, offer free trials	3-4	0-1
5	Customers Needing Attention	Above average recency, frequency and monetary ...	Make limited time offers, Recommend based on p...	2-3	2-3
6	About To Sleep	Below average recency, frequency and monetary ...	Share valuable resources, recommend popular pr...	2-3	0-2
7	At Risk	Spent big money and purchased often. But long ...	Send personalized emails to reconnect, offer r...	1-2	3-5
8	Can't Lose Them	Made biggest purchases, and often. But haven't...	Win them back via renewals or newer products, ...	1-3	4-5
9	Hibernating	Last purchase was long back, low spenders and ...	Offer other relevant products and special disc...	2-3	0-4
10	Lost	Lowest recency, frequency and monetary scores.	Revive interest with reach out campaign, ignor...	0-2	0-5





# **PARTITIONNEMENT DE DONNÉES NON SUPERVISÉ**

# MÉTHODES DE PARTITIONNEMENT DE DONNÉES NON SUPERVISÉES

## Basée sur l'utilisation de noyaux

- utilise des noyaux afin de définir les cluster
- Algorithme à tester : Kmeans

## Basée sur des modèles

- utilise des distribution de données (Gaussienne, etc...) afin de créer les clusters dans lesquelles seront les données.
- Algorithme à test : Les mixtures de Gaussiennes (Gaussian-mixture).

## Basée sur la densité

- utilise le fait que dans l'espace, les données similaires sont toutes regroupées au même endroit formant ainsi dans l'espace des points de haute densité.
- Algorithme à tester : DBSCAN

## Basée sur la hiérarchie des données

- Deux approches : ascendante et descendante
- Algorithme à tester : ascendante

# EVALUATION DE LA PERFORMANCE : COEFFICIENT DE SILHOUETTE

Métrique



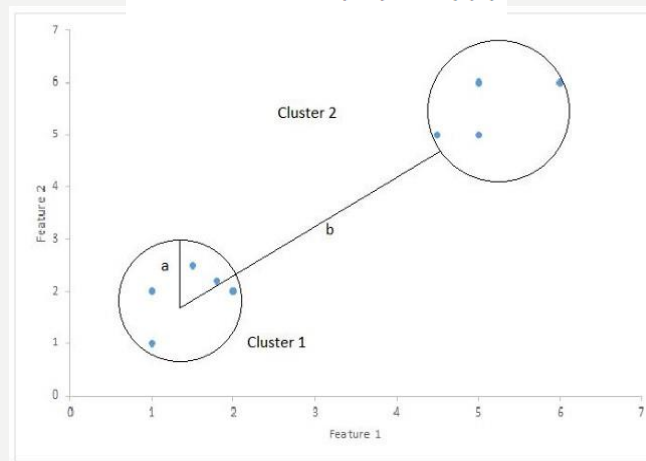
- ☐ Coefficient de silhouette : mesure la qualité d'une partition d'un ensemble de données en classification automatique

Expression



- ☐ Moyenne du coefficient de silhouette pour tous les points

$$\text{Silhouette Score} = (b-a)/\max(a,b)$$



Domaine de variation



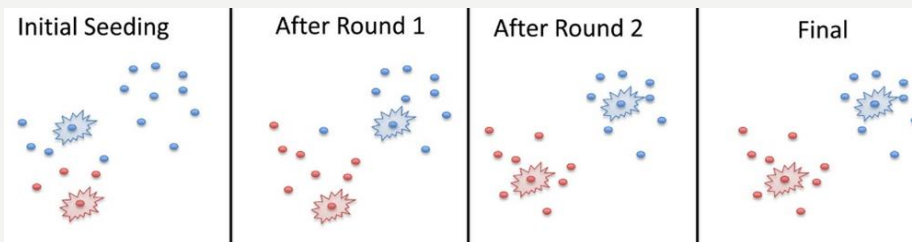
- ☐ Le coefficient de silhouette varie entre -1 (pire classification) et 1 (meilleure classification)

# KMEANS

## Méthode de calcul:



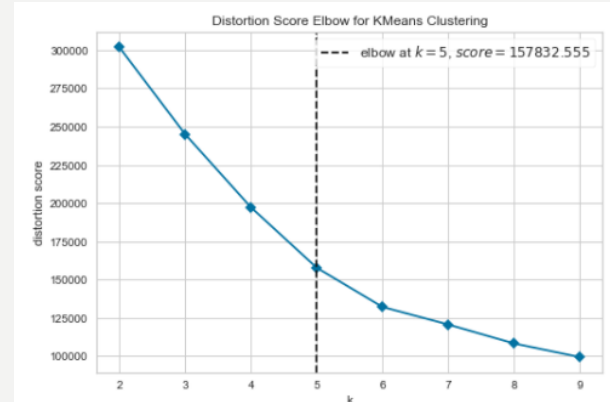
- ☐ Etape 0 : Initialisation des centrioles avec K-means++ : Choix des centres avec une probabilité liée à la distance au carré aux autres centres
- ☐ Etape 1 : Affectation de chaque individu au centre le plus proche
- ☐ Etape 2 : calcul des centres de gravité des groupes qui deviennent les nouveaux centroides
- ☐ Boucle itérative : itération des étapes 1, 2 jusqu'à ce que les centres ne bougent plus.



## Détermination du nombre optimal de clusters



- ☐ Méthode Elbow : distorsion (somme des distances au carré des centres)



Avantages

Faibles temps de calcul

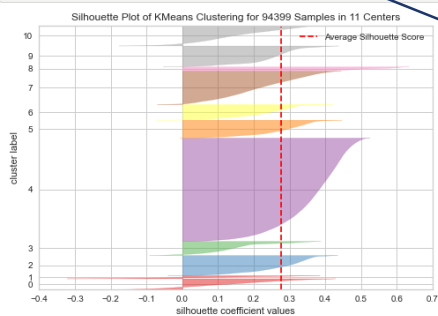
Inconvénients

Résultats non identiques entre 2 exécutions



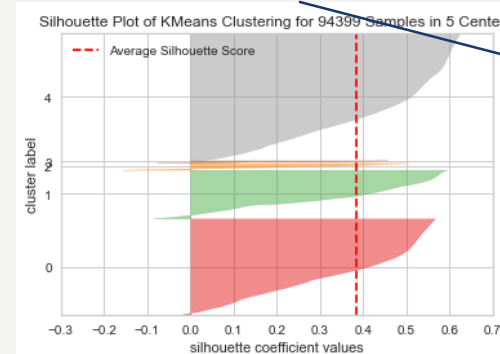
# KMEANS : DATASET COMPLET / RFM+SATISFACTION

## Tous les détails

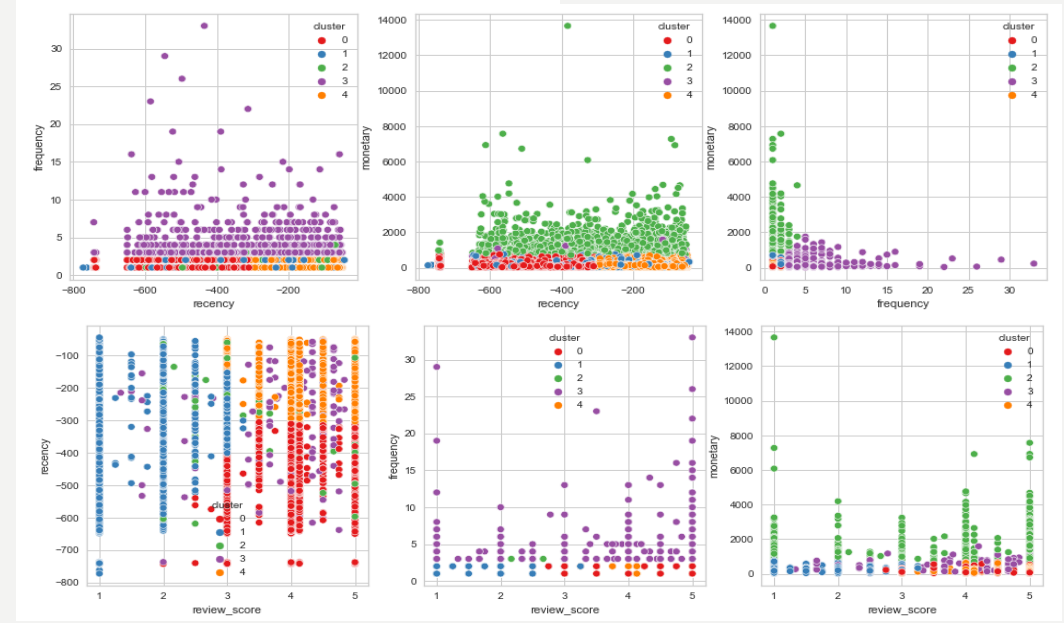
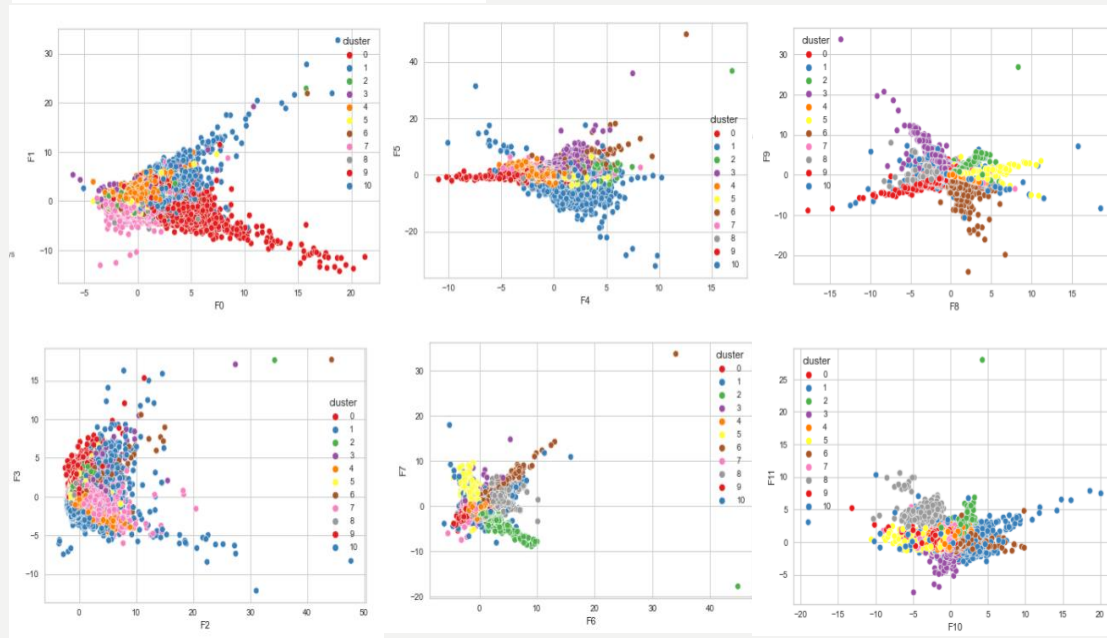


- 11 clusters
- Taux de silhouette 0,3
- clusters nombreux et confus

## RFM+Satisfaction

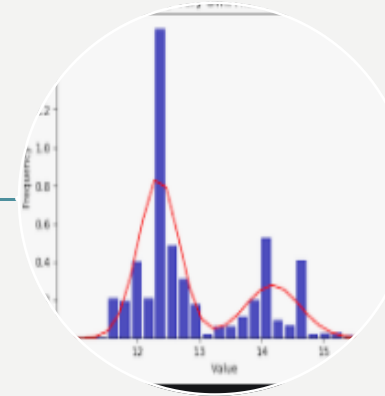
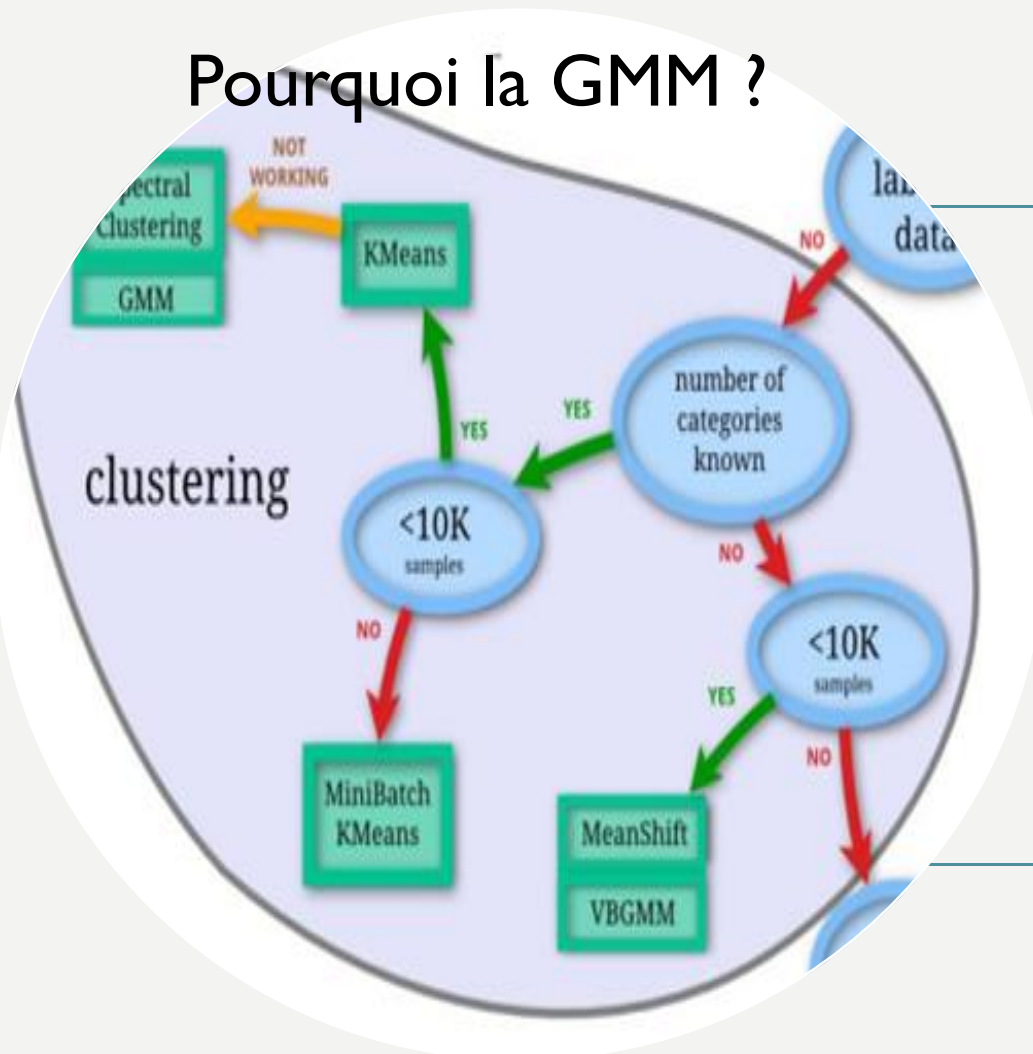


- 5 clusters
- Taux de silhouette 0,4
- clients relativement homogènes



# MODÈLE DE MÉLANGE DE GAUSSIENNES (GMM)

Pourquoi la GMM ?

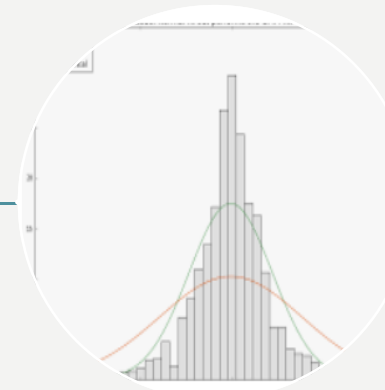


Comment la GMM fonctionne?

- Il suppose que les données suivent une distribution qui correspond à une moyenne pondérée de plusieurs distributions gaussiennes
- Les paramètres sont optimisés selon un critère de maximum de vraisemblance pour approcher le plus possible la distribution recherchée.

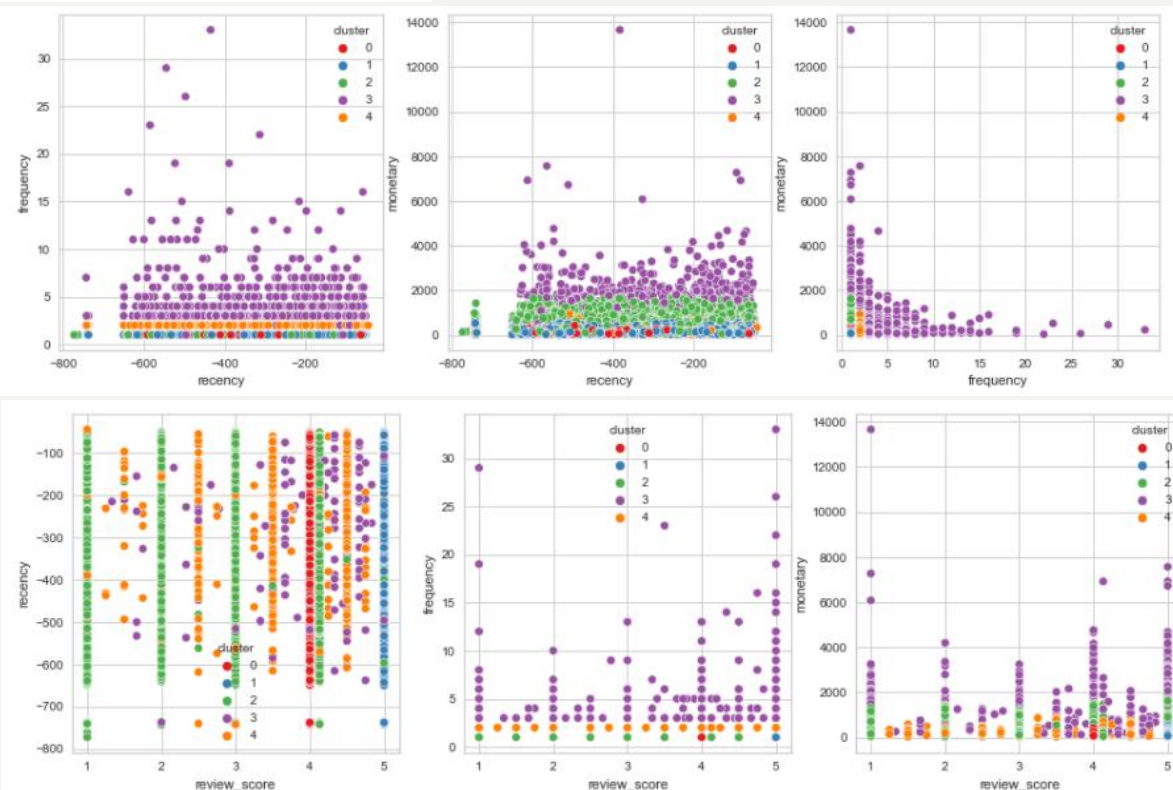
Différences avec Kmeans

- La GMM est paramétrique, le k-means non : il ne suppose pas que les clusters suivent une loi particulière.
- Réputé reconstruire de manière particulièrement efficace les données manquantes dans un jeu de données expérimentale



# GMM: RFM+SATISFACTION

Silhouette Coefficient: 0.136



Résultat similaire

Pas de clusters pour l'axe « Recency ».

Déclinaison en plan d'action de marketing moins évidente que Kmeans

# DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE (DBSCAN)

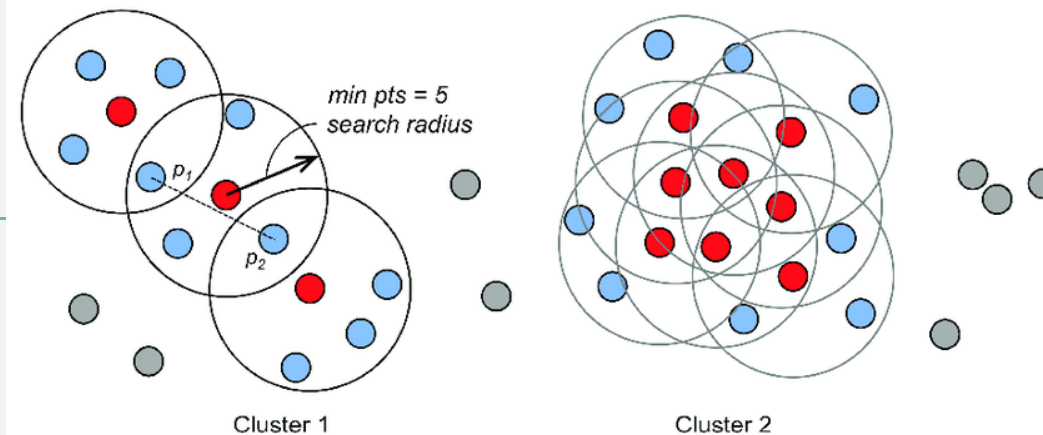
## Principe

Estimation de la densité locale

2 paramètres attendus :

Rayon de recherche :  $\epsilon$

Nombre minimum de points : MinPts



## VS Kmeans

Nombre de clusters non défini en amont

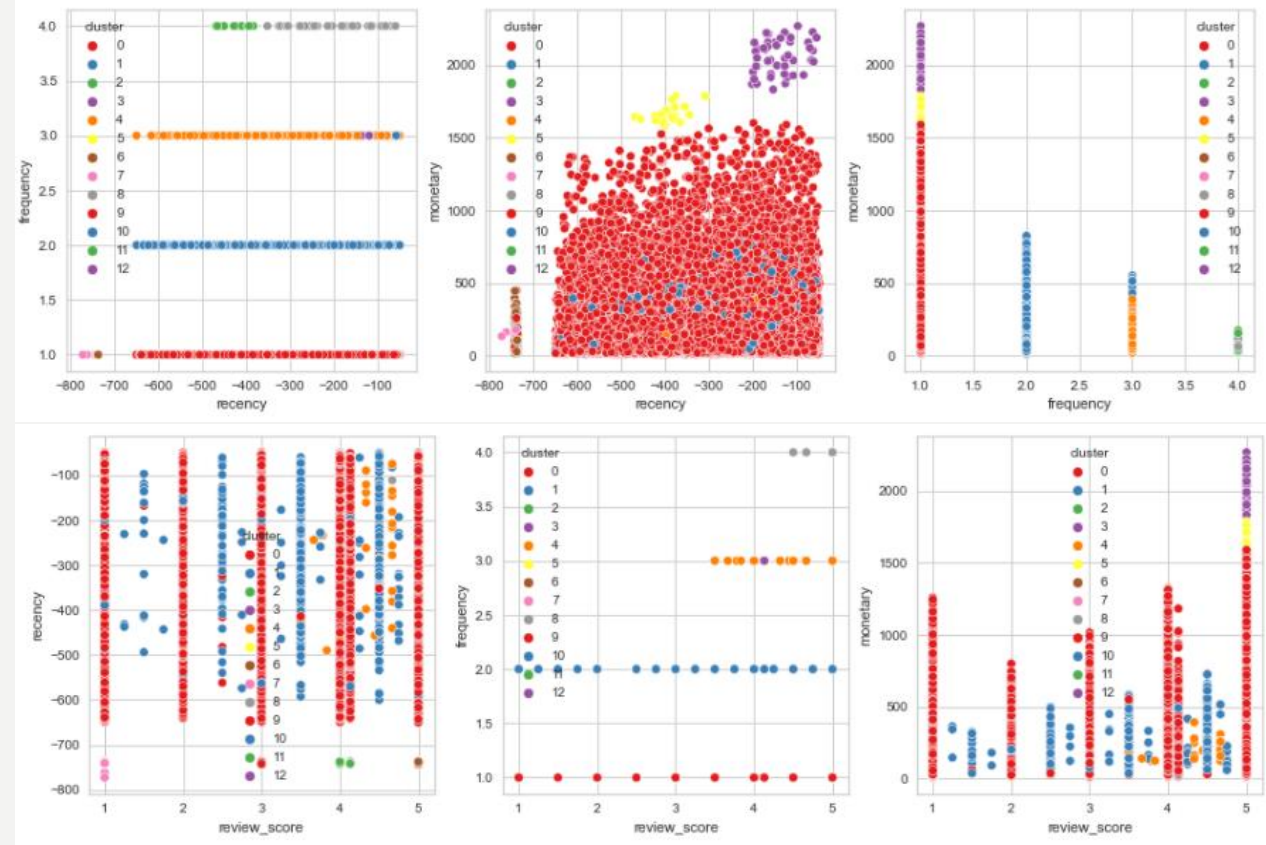
Capacité de gérer les valeurs aberrantes ou anomalies.

# DBSCAN : RFM+SATISFACTION

- Clusters nombreux et extrêmement déséquilibrés
- Des gros clients sont considérés comme des bruits

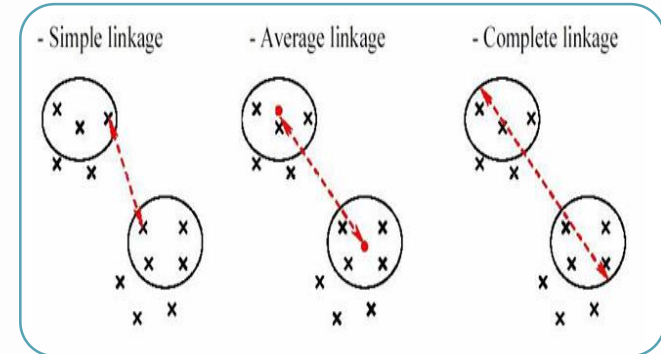
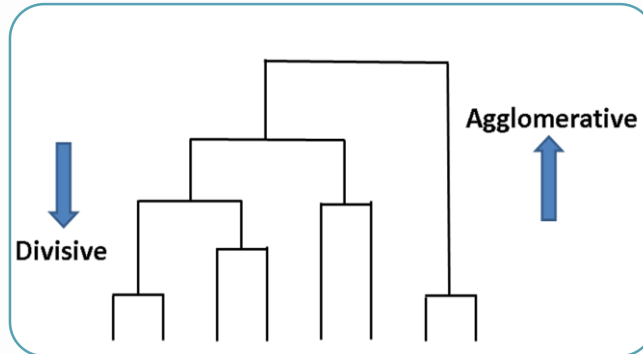
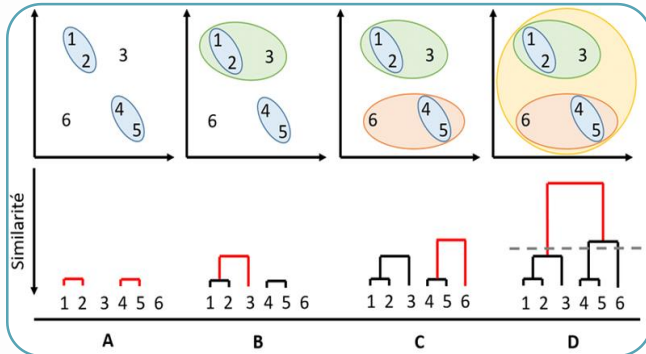
	recency	frequency	monetary	review_score	size	%total
cluster						
-1	-328.50	2.70	1,035.60	3.40	1402	0.01
0	-285.30	1.00	149.20	4.10	87999	0.93
1	-286.30	2.00	185.30	4.20	4349	0.05
2	-740.40	1.00	106.90	4.00	45	0.00
3	-140.60	1.00	2,059.10	5.00	46	0.00
4	-287.10	3.00	138.70	4.80	269	0.00
5	-389.20	1.00	1,663.60	5.00	23	0.00
6	-740.50	1.00	133.30	5.00	118	0.00
7	-742.90	1.00	108.00	1.00	34	0.00
8	-195.20	4.00	88.60	5.00	40	0.00
9	-740.50	1.00	114.70	3.00	17	0.00
10	-107.10	3.00	485.20	4.90	15	0.00
11	-430.20	4.00	83.80	5.00	15	0.00
12	-113.70	3.00	127.70	4.00	27	0.00

Silhouette Coefficient: 0.218





# PARTITIONNEMENT HIÉRARCHIQUE



## Principe :

- on cherche les 2 points les plus proches
- Et puis on cherche à nouveau les points les plus proches pour les regrouper en un cluster,
- on itère jusqu'à n'avoir plus qu'un seul cluster.

## Deux approches :

- Ascendante ou Agglomerative
- Descendante ou divisive

## Méthodes d'agrégation :

- simple linkage
- *average linkage*
- complet linkage
- ward (regrouper les classes de façon que l'augmentation de l'inertie interclasse soit maximale, ou l'inertie intraclasse soit minimum)
- ...

# MÉTHODE MIXTE : RFM+SATISFACTION

Méthode  
hiérarchique

Inconvénients

- Pas adapté aux grands volumes de données

Avantages :

- Aide au choix du nombre de groupes
- Hiérarchique
- Facile à utiliser
- Choix de la distance

Méthode  
mixte

Kmeans pour  
calculer un grand  
nombre de clusters  
temporaires

Créer des segments  
à partir des  
centroïdes de  
Kmeans

Ré-affecter chaque  
client à son cluster  
final



- Clusters trop déséquilibrés pour être exploitable

	recency	frequency	monetary	review_score	size	%total
cluster						
0	-286.80	1.10	161.50	4.10	94085	1.00
1	-300.30	6.80	291.90	4.20	206	0.00
2	-298.60	1.20	3,374.20	4.00	101	0.00
3	-469.70	24.40	229.20	3.60	7	0.00

# CONCLUSION

- Kmeans permet de réaliser la segmentation de manière rapide et efficace sur les données de commandes et de satisfactions.

	recency	frequency	monetary	review_score	size	%total
cluster						
0	-441.80	1.00	135.80	4.60	32333	0.34
1	-281.30	1.00	153.30	1.70	16383	0.17
2	-284.60	1.10	1,216.70	4.10	2242	0.02
3	-291.60	4.20	228.70	4.10	875	0.01
4	-171.40	1.00	135.60	4.70	42566	0.45

cluster 0 clients perdus :  
leur dernière commande date  
plus d'un an

cluster 1 clients mécontents  
: ont une opinion négative  
de leur achat

cluster 2 gros acheteurs:  
sont dépensiers

cluster 3 clients fidèles : ont  
acheté plusieurs fois sur le  
site

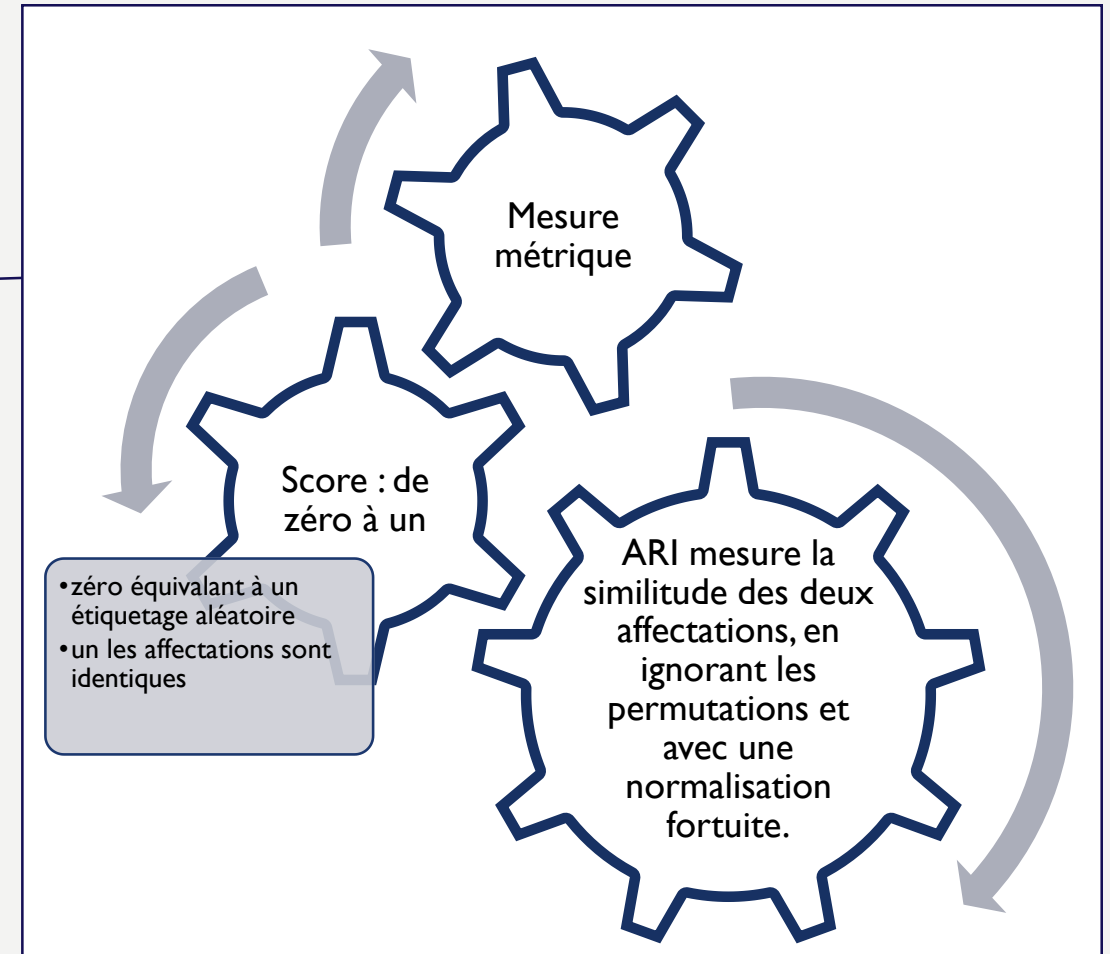
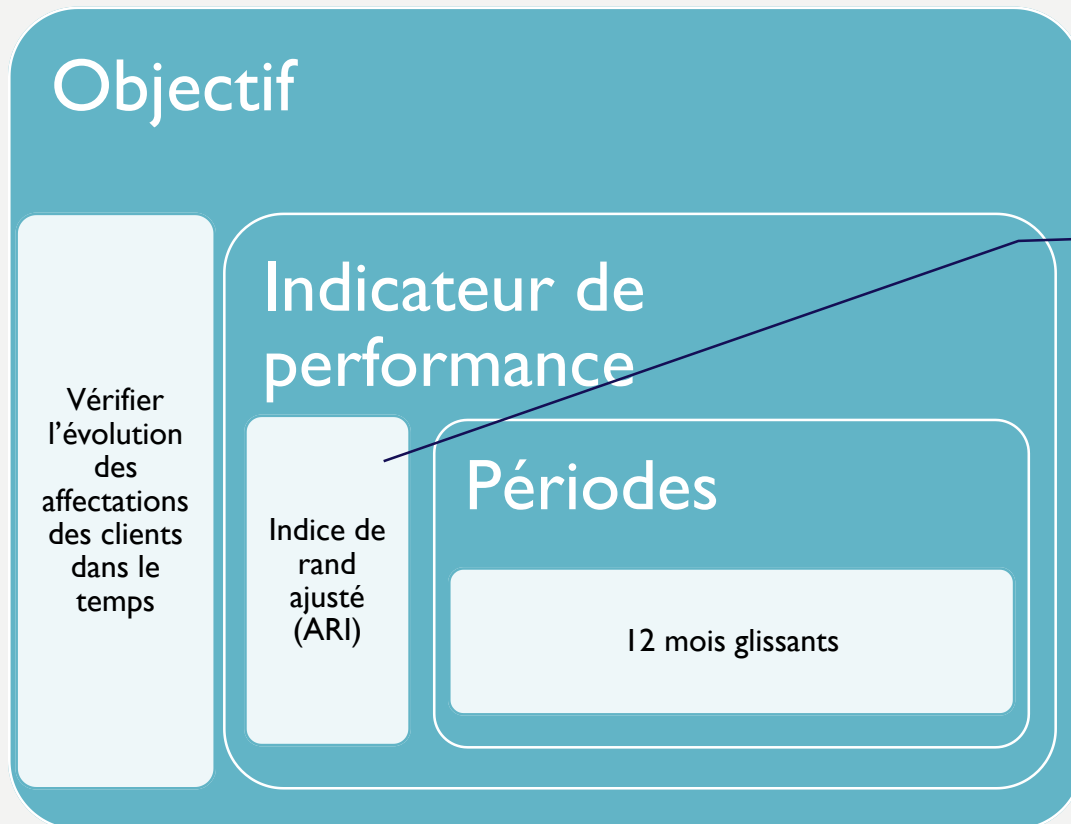
cluster 4 clients récents :  
ont effectué des commandes  
pendant la dernière année





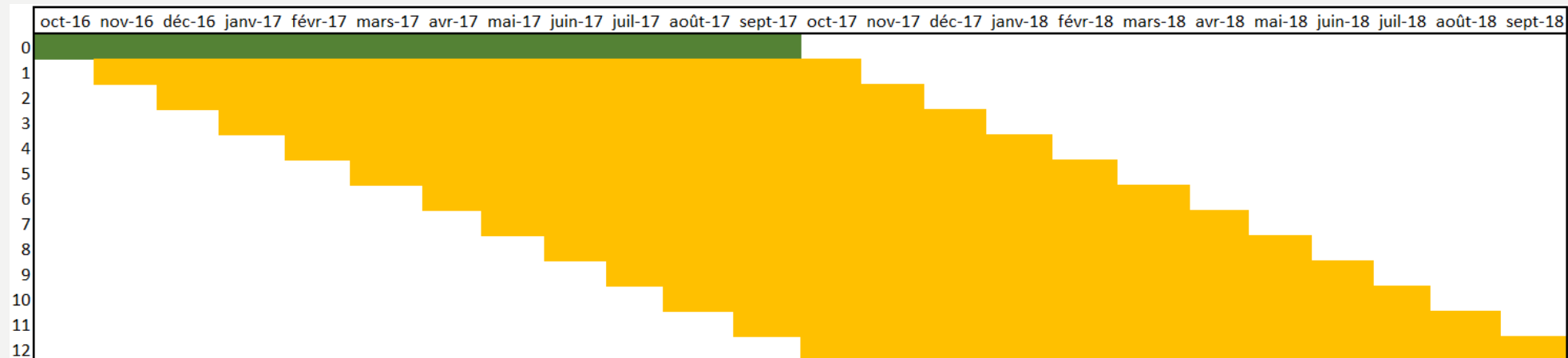
# **STABILITÉ DE LA SEGMENTATION DANS LE TEMPS**

# STABILITÉ DANS LE TEMPS : MÉTHODE DE LA SIMULATION



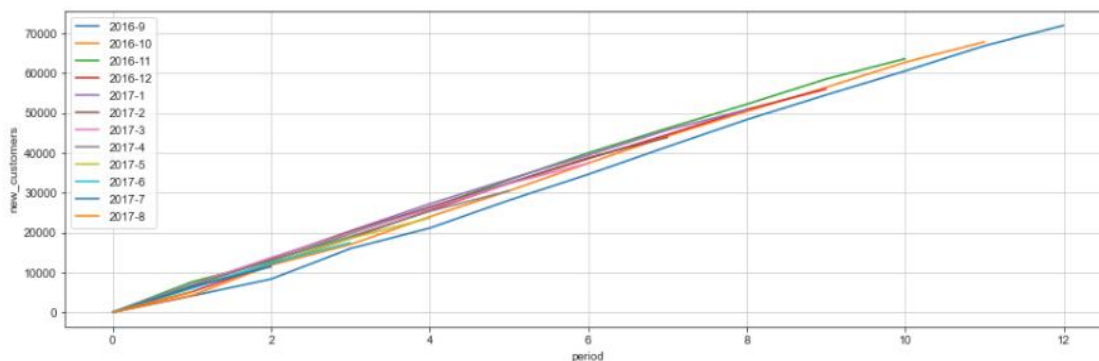
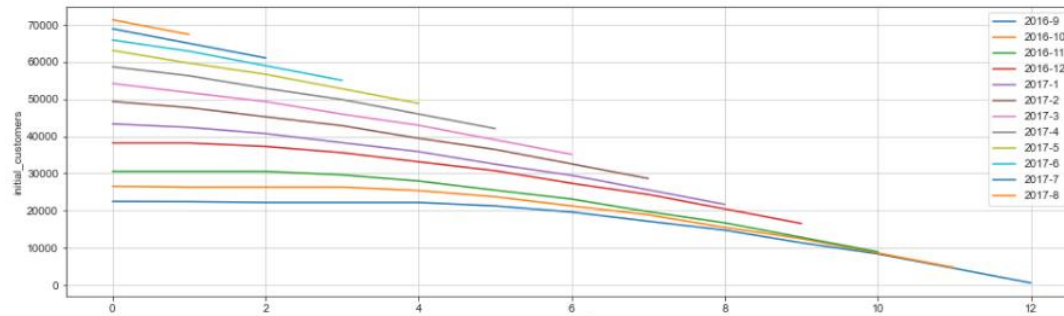
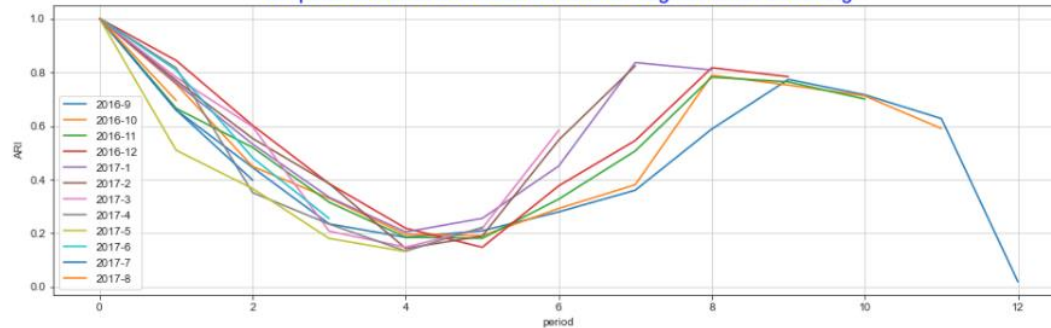
# STABILITÉ DANS LE TEMPS : UN EXEMPLE DE LA SIMULATION

- Population de référence (ligne 0) : la clientèle de la première année
- Période d'évolution ( ligne 1 ->12): du 10/2016 au 9/2018
- Nombre de simulation : 12 pour tracer leur évolution mensuelle
- Indicateurs calculés :
  - Indice Rand ajusté (ARI) : similitude des clusters de la population initiale
  - Population initiale restante : nombre de clients initiaux restant
  - Nouveaux entrants : nombre de nouveaux clients



# STABILITÉ DANS LE TEMPS: RÉSULTAT DE LA SIMULATION

temporal evolution of K-means customer segmentation clustering



- Quelque soit la période de référence, les indicateurs partagent la même tendance.
  - ARI : au bout de 4 mois, la plupart des clients changent de segment. La remontée du score après 4 mois est dû au départ important de clients initiaux.
  - clients existants : restent relativement stable dans les 2 mois qui suivent
  - Nouveaux clients : hausse constante et de manière considérable
- Conclusion : il convient de ré-effectuer la segmentation chaque **2 à 4 mois** afin de prendre en compte les nouveaux clients et mettre à jour le segment des clients existants.