Segmentez des clients d’un site e-commerce

P5 - OpenClassrooms

Sommaire

Introduction

Je suis consultant pour Olist, une entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne.

Mon rôle est d'accompagner Olist dans leur projet de monter une équipe Data et leur premier cas d'usage Data Science autour de la segmentation client.

Mon objectif est de comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.

Base de données

Customers : 6 colonnes, 99441 lignes

geoloc : 6 colonnes, 1000163 lignes

order_items : 8 colonnes, 112650 lignes

order_pymts : 6 colonnes, 103886 lignes

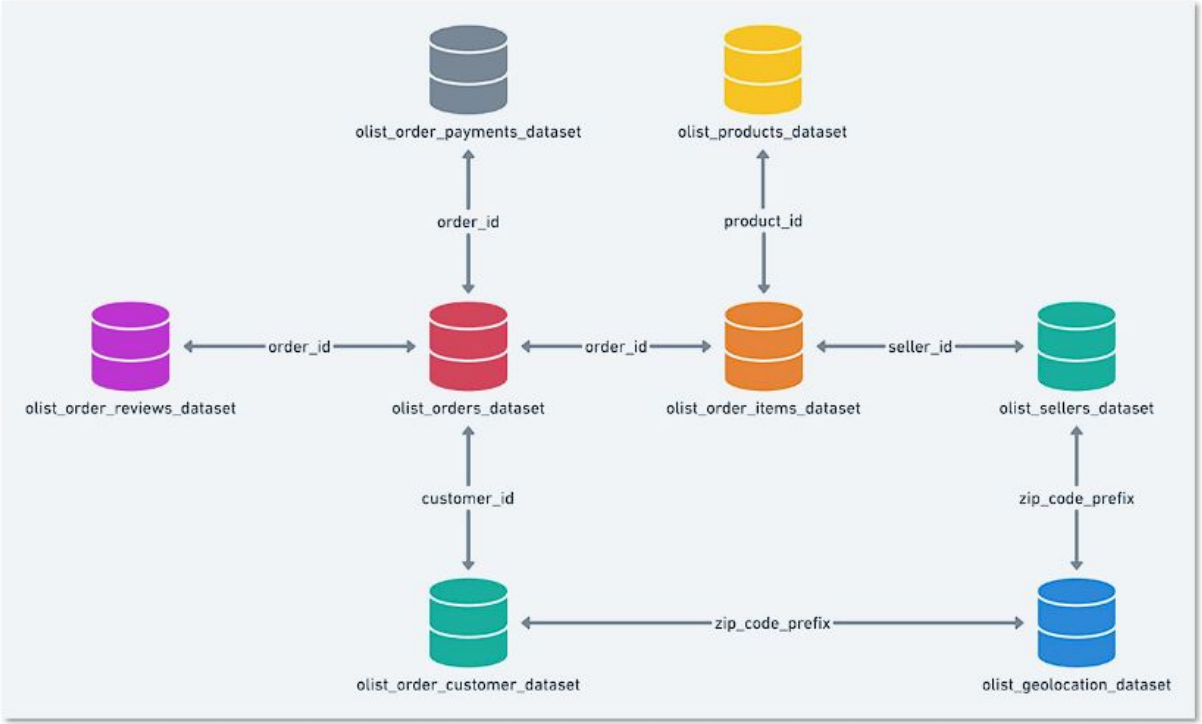
order_reviews : 8 colonnes, 99224 lignes

orders : 9 colonnes, 99441 lignes

products : 10 colonnes, 32951lignes

sellers : 5 colonnes, 3095 lignes

translation : 3 colonnes, 71 lignes



Dashboard

fonctionnalités (requêtes)

● En excluant les commandes annulées, quelles sont les commandes récentes de moins de 3 mois que les clients ont reçues avec au moins 3 jours de retard ?

● Qui sont les vendeurs ayant généré un chiffre d'affaires de plus de 100 000Real sur des commandeslivrées via Olist ?

● Qui sont les nouveaux vendeurs (moins de 3 mois d'ancienneté) qui sont déjà très engagés avec la plateforme (ayant déjà vendu plus de 30 produits) ?

● Question : Quels sont les 5 codes postaux, enregistrant plus de 30 reviews, avec le pire review score moyen sur les 12 derniers mois

Q1 — Commandes ≤ 3 mois livrées avec ≥ 3 jours de retard

order_id : identifiant unique de la commande

customer_id : identifiant du client qui a passé la commande

purchase_date : date d'achat

delivered_date : date où la commande a été livrée au client

eta_date : date de livraison estimée

	order id	customer id	purchase c	delivered c	eta date	delay days
1	b2997e1d7061605e9285496c581d1fbd	9e83d47684eb1a58b1c31830f5de10ac	2018-07-30	2018-10-02	2018-08-14	49
2	a2b4be96b53022618030c17ed437604d	ffa87b4246c4848711afb512bd51f161	2018-07-22	2018-09-27	2018-08-17	41
3	238652e39c5fdf89a8fd44776f532501	fc041ede47154c40f55455e20c1a1954	2018-07-25	2018-09-17	2018-08-10	38
4	4af2fb154881f350d8696f7f7a7f80d3	7c71fa0871e272a25eeccac52af90595	2018-07-23	2018-09-20	2018-08-13	38
5	7d09831e67caa193da82cfea3bee7aa5	1409b2945191b7aff1975ba2ce9918c5	2018-08-05	2018-09-25	2018-08-20	36
6	4505acb3759da6b9c7d79a80d29ab3bb	a35878bee339b45240b5a327d933509b	2018-08-06	2018-09-19	2018-08-17	33
7	84869ba3df14629b57ca40c491a842e6	8000d8c2201ad0577d5f459c6325ccdc	2018-07-27	2018-09-14	2018-08-13	32
8	84db939b1ab9686533e9a06a0354beb6	c0789eee49fe7d5d93d5e412c14181ce	2018-07-20	2018-09-17	2018-08-16	32
9	1e7d25f611e794f9614dd3e10a8596e7	8be45a1114ff0e79615f7b8189aec7df	2018-08-01	2018-09-21	2018-08-23	29
10	6325af88a0611fc357055cb87dcec11e	bfc24858928300e9b18e0d96637b8404	2018-07-18	2018-09-17	2018-08-21	27
11	c005c973843746a08a6ea826af4ce0c0	71ac72b29860fdac58666426bbe6b4ba	2018-07-29	2018-09-13	2018-08-21	23
12	826781a83e7fab9989e0882180d2adc2	8d8e4f7961914580fefe9e4ccb4df76f	2018-08-02	2018-08-28	2018-08-07	21
13	5f2aaf089bdf0d14f2357066f139c2a4	04fbe9cd7c2f918568c5b4e934a56316	2018-07-17	2018-08-28	2018-08-07	21
14	2591f6277be80b0c25627c745ec900c4	614e4a9149f6119fc5e8780ddfeaedfd	2018-08-04	2018-09-03	2018-08-14	20

delay_days : retard en **jours** = `delivered_date - eta_date`

`nombre de lignes: 100`

Q2 — Vendeurs avec CA > 100 000 BRL (sur commandes livrées)

seller_id : identifiant du vendeur (marchand) sur la marketplace

revenue_brl :**chiffre d'affaires** agrégé (**somme de** `price` des lignes `order_items` rattachées à des **commandes livrées**.)

items_sold : nombre **de lignes d'items** vendues par ce vendeur (chaque ligne = ~1 produit). Attention : `order_item_id` n'est **pas** une quantité, c'est juste un index de ligne.

delivered_orders : **nombre de commandes distinctes** (livrées) dans lesquelles ce vendeur apparaît (au moins un item).

	seller id	revenue brl	items sold	delivered orders
1	4869f7a5dfa277a7dca6462dcf3b52b2	226987.93	1148	1124
2	53243585a1d6dc2643021fd1853d8905	217940.44	400	348
3	4a3ca9315b744ce9f8e9374361493884	196882.12	1949	1772
4	fa1c13f2614d7b5c4749cbc52fecda94	190917.14	579	578
5	7c67e1448b00f6e969d365cea6b010ab	186570.05	1355	973
6	7e93a43ef30c4f03f38b393420bc753a	165981.49	322	319
7	da8622b14eb17ae2831f4ac5b9dab84a	159816.87	1548	1311
8	7a67c85e85bb2ce8582c35f2203ad736	139658.69	1155	1145
9	1025f0e2d44d7041d6cf58b6550e0bfa	138208.56	1420	910
10	955fee9216a65b617aa5c0531780ce60	131836.71	1472	1261
11	46dc3b2cc0980fb8ec44634e21d2718e	122811.38	523	503
12	6560211a19b47992c3666cc44a7e94c0	120702.83	1996	1819
13	620c87c171fb2a6dd6e8bb4dec959fc6	112461.5	778	722
14	7d13fca15225358621be4086e1eb0964	112436.18	571	558

nombre de lignes: 17

Q3 — Nouveaux vendeurs (< 3 mois d'ancienneté) déjà très engagés (> 30 produits)

seller_id : identifiant du vendeur.

first_sale_date : **date de 1re vente** observée pour ce vendeur (sur **commande livrée**).

days_since_first_sale : **ancienneté** en **jours** depuis cette première vente, calculée par rapport à la **date max** du dataset.

products_sold : volume d'items vendus = **COUNT(*)** des lignes dans `order_items` rattachées à des **commandes livrées** (chaque ligne ≈ 1 produit).

Le résultat est déjà filtré à `> 30`.

	seller id	first sale date	days since first sale	products sold
1	d13e50eaa47b4cbe9eb81465865d8cfc	2018-08-04	74	68
2	81f89e42267213cb94da7ddc301651da	2018-08-08	70	52
3	240b9776d844d37535668549a396af32	2018-07-17	92	35

Q4 — 5 codes postaux (≥ 30 reviews) avec le pire score moyen sur 12 mois

zip_prefix : préfixe du code postal client

(5 premiers chiffres du CEP brésilien), représentatif d'une zone.

avg_review_score : note moyenne des avis (1 à 5) sur les 12 derniers mois glissants (par rapport à la review la plus récente).

Plus c'est bas, plus la satisfaction est mauvaise

avg_review_score : note moyenne des avis (1 à 5) sur les 12 derniers mois glissants (par rapport à la review la plus récente).

Plus c'est bas, plus la satisfaction est mauvaise

	zip_prefix	avg_review_score	n_reviews
1	22753	2.868	53
2	22723	3	31
3	28893	3.125	32
4	22770	3.184	38
5	13056	3.273	33

focus sur les variables importantes

Le choix des variables

```
use_cols = ['recency_days','frequency','monetary','aov','avg_review_score','delay_rate_ge3d']
```

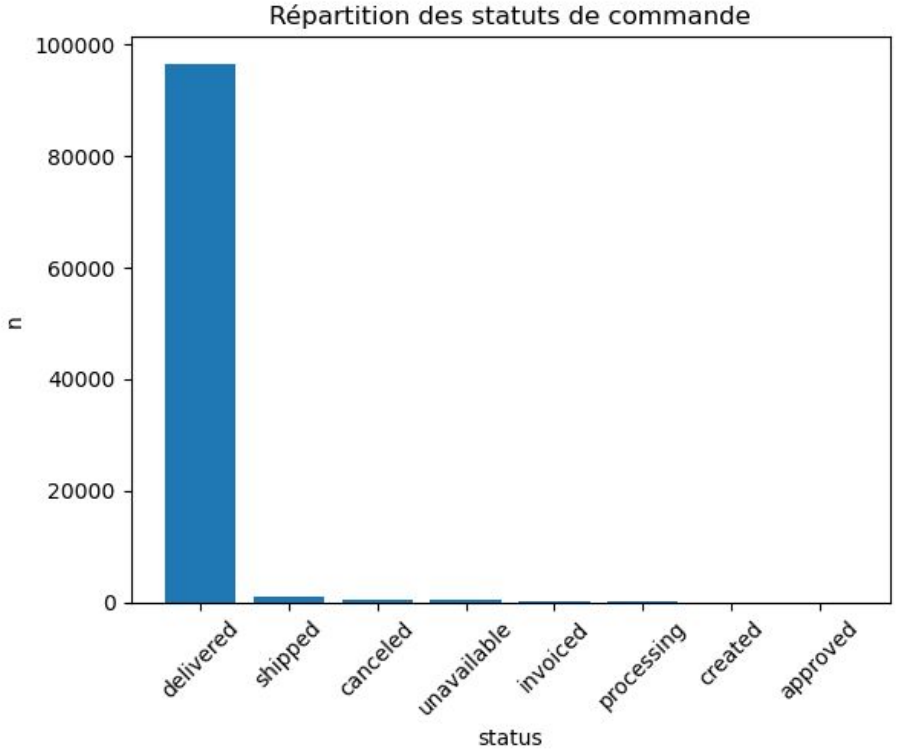
Le noyau : RFM (comportement d'achat)

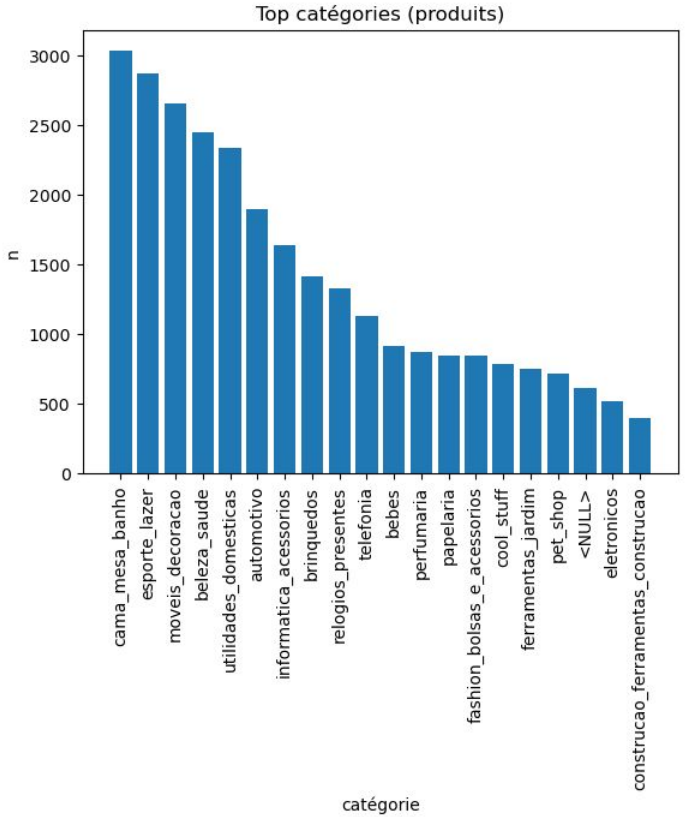
- **Recency (`recency_days`) Clients récents = Clients actif**
  Mesure la **fraîcheur de la relation**. Petit = client **récent** (forte propension à ré-acheter), grand = **dormant** → cible **win-back**.
- **Frequency (`frequency`) Loyauté des clients**
  Sépare **occasionnels** vs **récurrents**. Action : abonnements/bundles, cadence de relance, protection des bons clients.
- **Monetary (`monetary`) Chiffre d'affaire**
  Valeur cumulée à défendre/développer. On applique `log1p` pour compresser les gros montants (éviter qu'ils écrasent tout).

Enrichissements marketing

- **AOV (`aov = monetary / frequency`, puis `log1p`)**
  Capte le **ticket moyen par commande** (différent du Monetary total).
  → Distingue **petit panier** vs **premium**, utile pour **coupons**, **bundles**, seuils de livraison offerte.
- **Satisfaction (`avg_review_score`)**
  Synthèse de l'**expérience perçue** (1★–5★).
  → Oriente la **tonalité**: *satisfaits* → upsell/fidélité ; *insatisfaits* → réparation/geste avant relance.
- **Logistique (`delay_rate_ge3d`)**
  Taux de commandes livrées avec ≥3j de retard.
  → Indique une **friction opérationnelle** : avant de pousser une promo, **corriger SLA** / offrir un **express** / geste commercial.

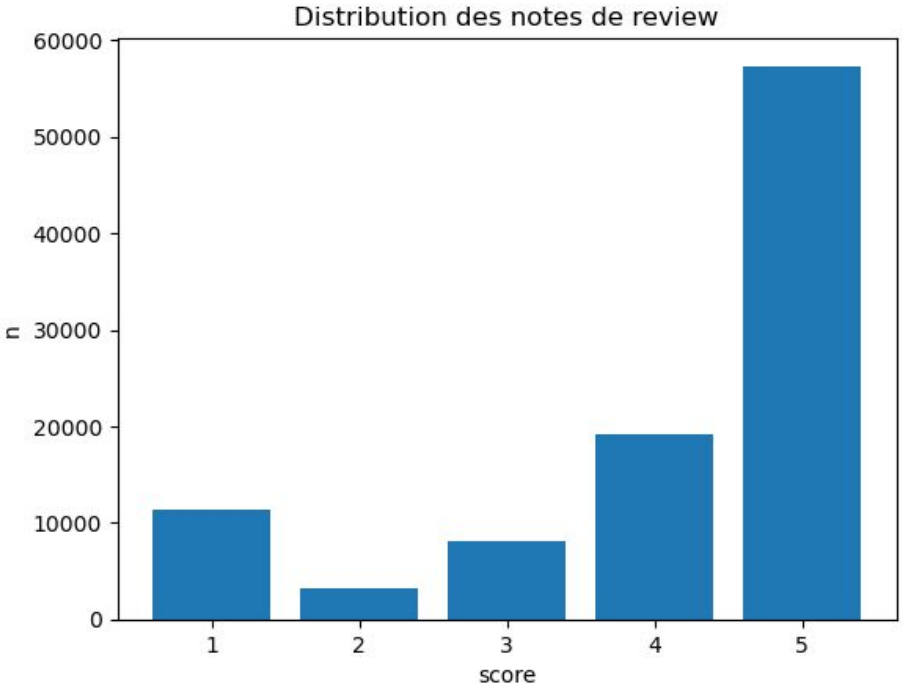
EDA (statuts commande et produits)

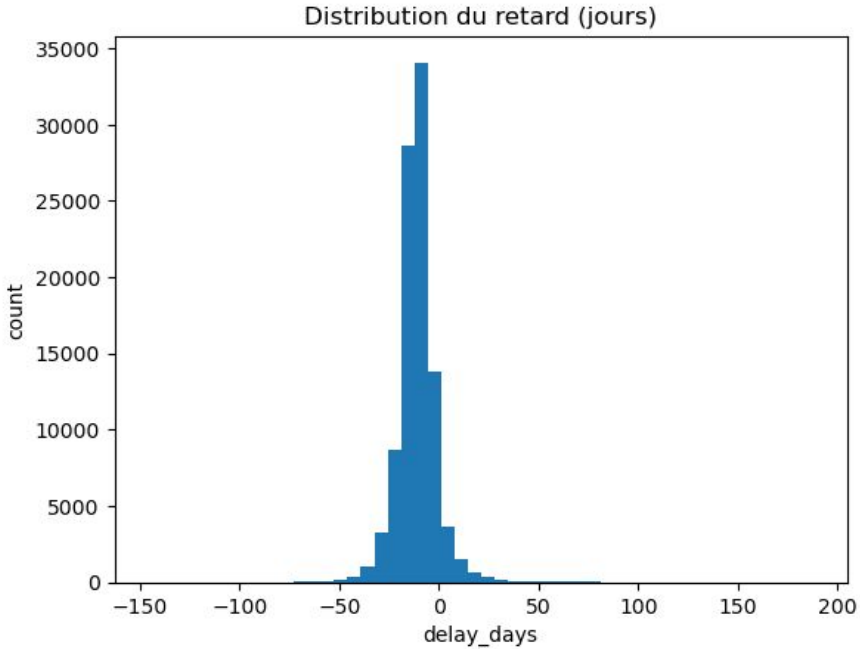




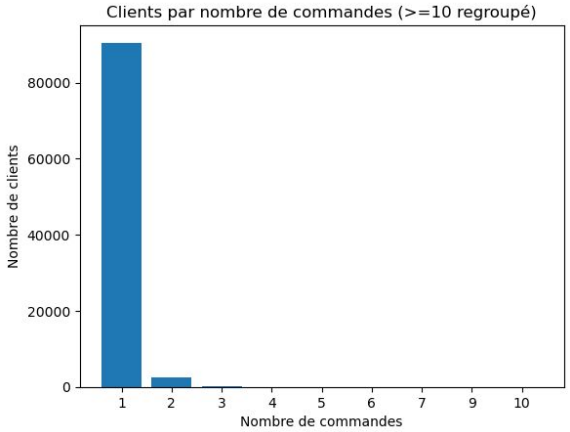
EDA (fréquence commandes)

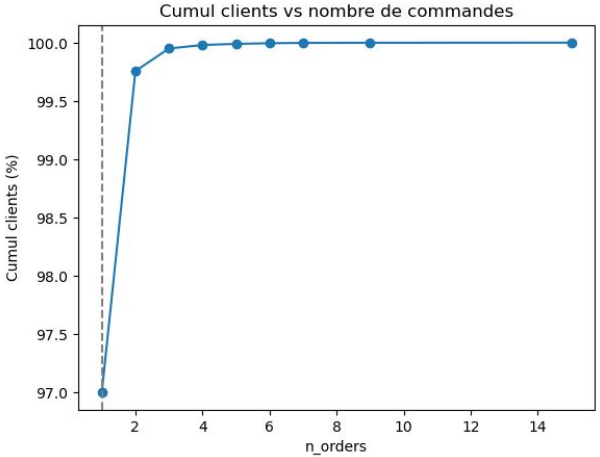
EDA (note client / retard)

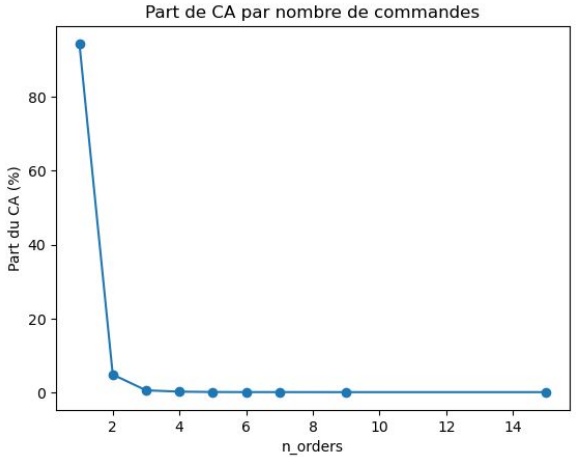




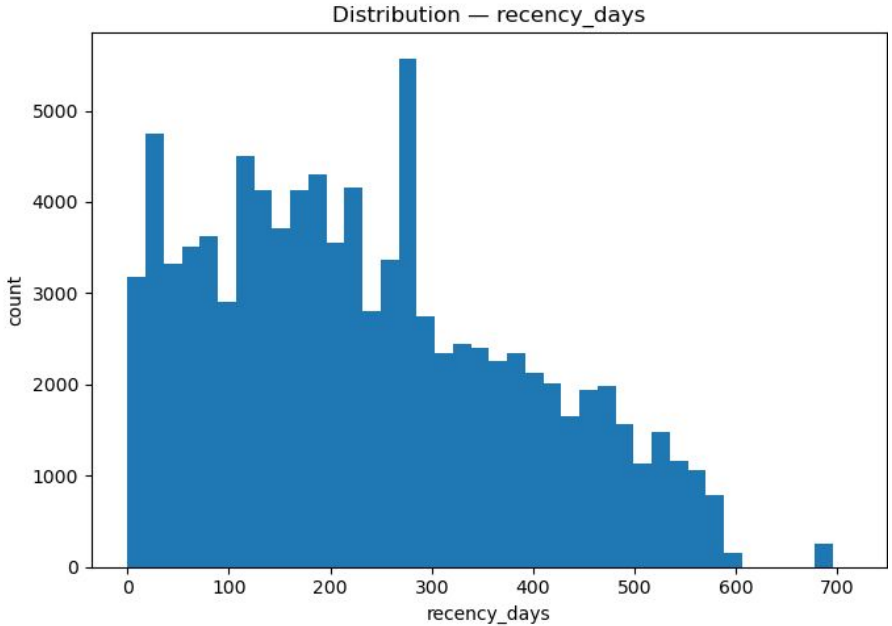
EDA (commande par client et CA)

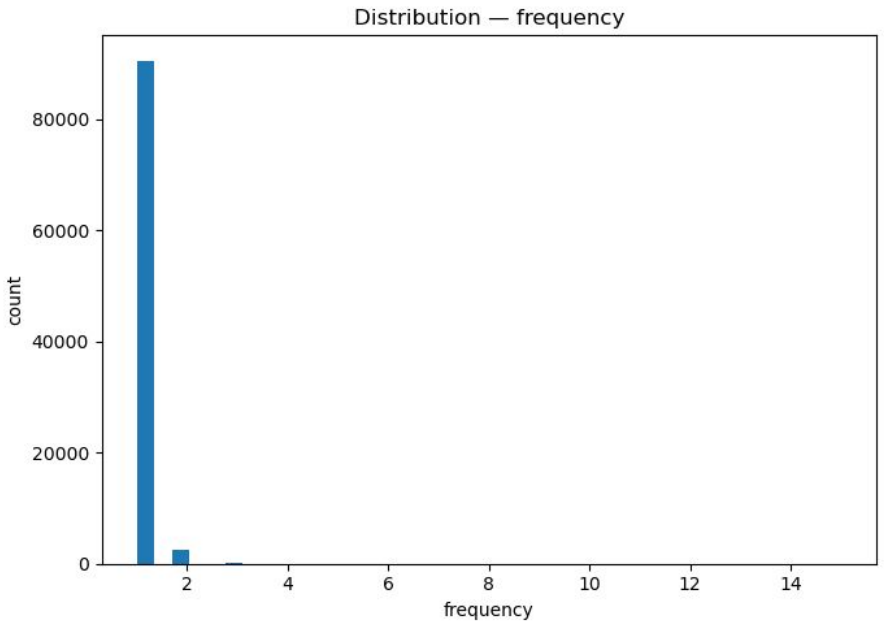




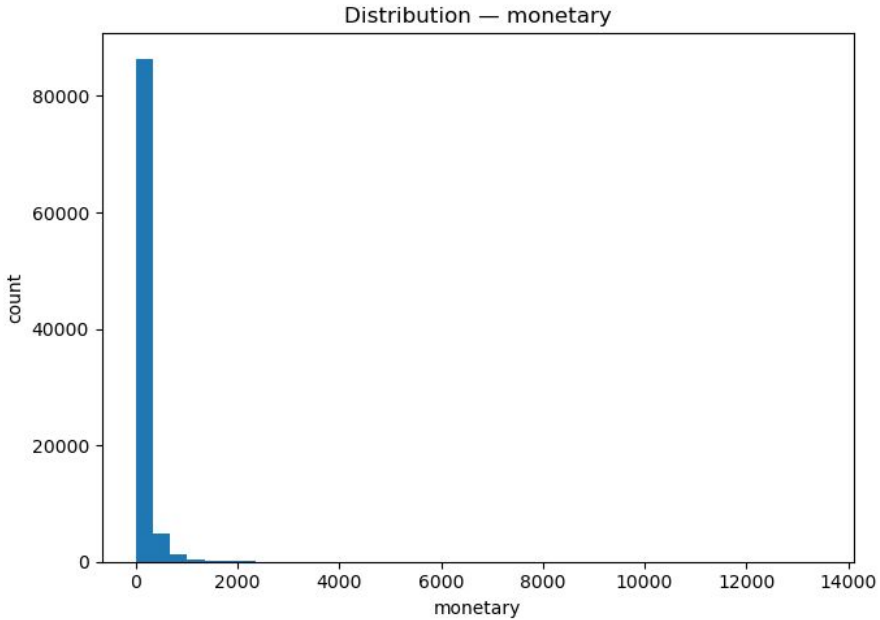


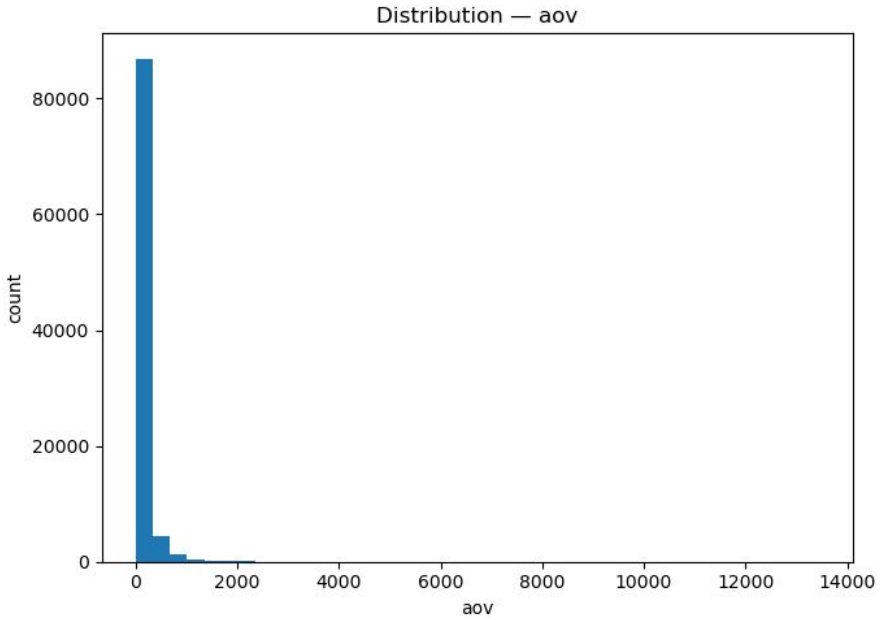
EDA (RFM) 1


Distribution — recency_days
count
recency_days
0
1000
2000
3000
4000
5000
0
100
200
300
400
500
600
700

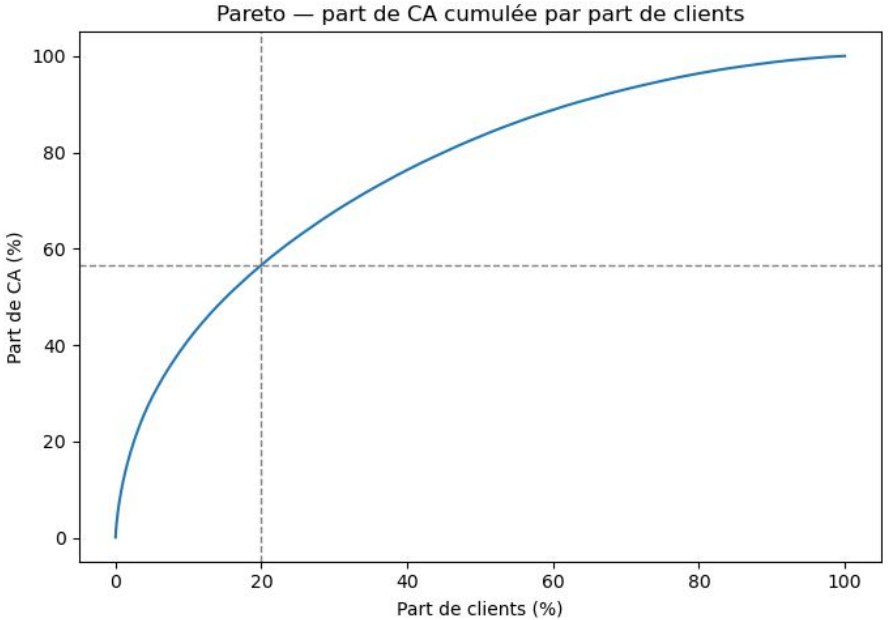

Distribution — frequency
count
frequency
0
20000
40000
60000
80000
2
4
6
8
10
12
14

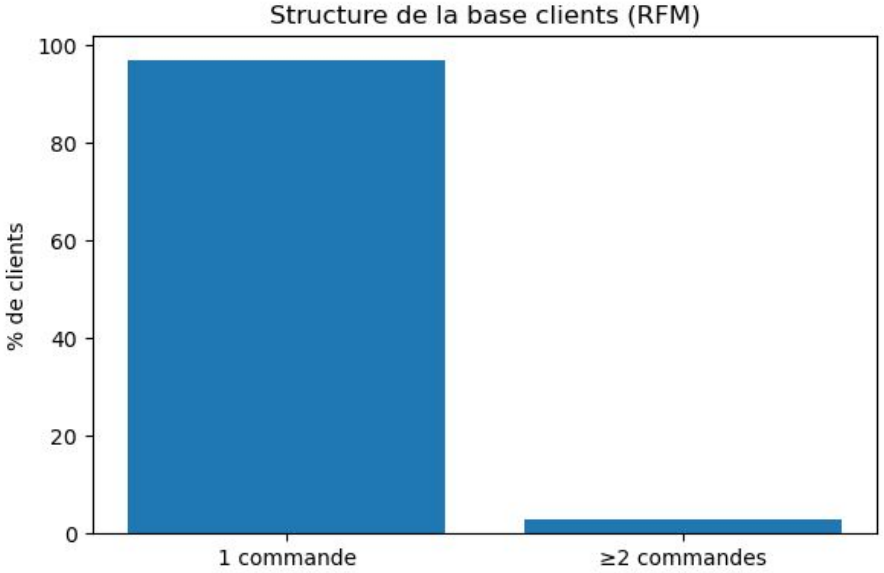
EDA (RFM) 2


Distribution — monetary
count
monetary
0
20000
40000
60000
80000
0
2000
4000
6000
8000
10000
12000
14000


Distribution — aov
count
aov
0
20000
40000
60000
80000
0
2000
4000
6000
8000
10000
12000
14000

EDA (part du CA vs part de clients)





EDA (conclusion)

KPIs RFM (base livrée)

Clients analysés : 93 358

One-timers (1 commande) : 90 557 (97 %)

Repeaters (≥ 2 commandes) : 2 801 (3 %)

Récence médiane : 218 jours

Fréquence médiane : 1

Monetary médian : 89,73 R$

Panier moyen (AOV) médian : 86,99 R$

97 % des clients n’achètent qu’une fois ; 3 % seulement reviennent.

Le top 10 % de clients réalise 41 % du CA (forte concentration).

Récence médiane 218 j → fort potentiel réactivation.

Rappel des variables

Le choix des variables

```
use_cols = ['recency_days','frequency','monetary','aov','avg_review_score','delay_rate_ge3d']
```

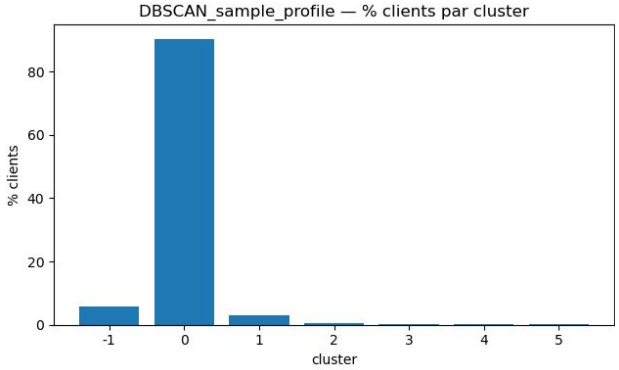
Le noyau : RFM (comportement d'achat)

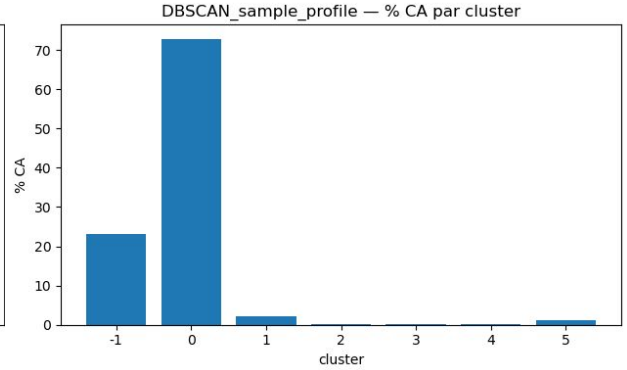
- **Recency (`recency_days`) Clients récents = Clients actif**
  Mesure la **fraîcheur de la relation**. Petit = client **récent** (forte propension à ré-acheter), grand = **dormant** → cible **win-back**.
- **Frequency (`frequency`) Loyauté des clients**
  Sépare **occasionnels** vs **récurrents**. Action : abonnements/bundles, cadence de relance, protection des bons clients.
- **Monetary (`monetary`) Chiffre d'affaire**
  Valeur cumulée à défendre/développer. On applique `log1p` pour compresser les gros montants (éviter qu'ils écrasent tout).

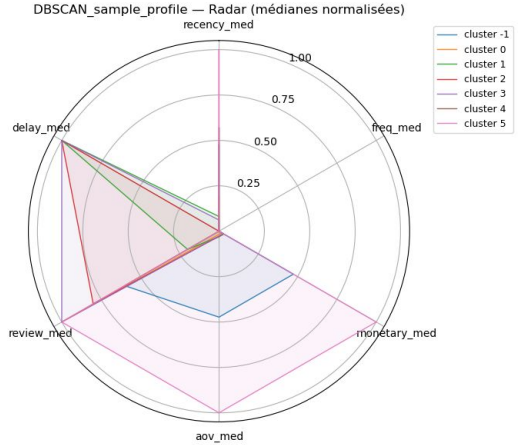
Enrichissements marketing

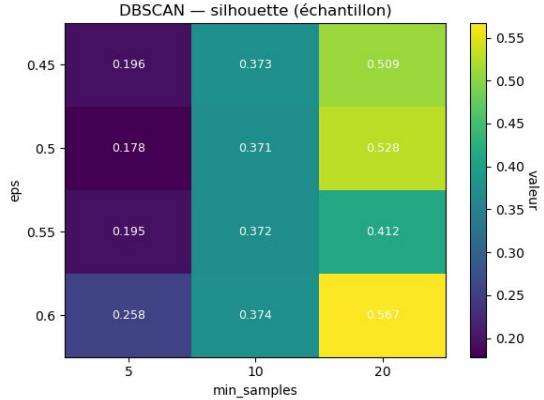
- **AOV (`aov = monetary / frequency`, puis `log1p`)**
  Capte le **ticket moyen par commande** (différent du Monetary total).
  → Distingue **petit panier** vs **premium**, utile pour **coupons**, **bundles**, seuils de livraison offerte.
- **Satisfaction (`avg_review_score`)**
  Synthèse de l'**expérience perçue** (1★–5★).
  → Oriente la **tonalité**: *satisfaits* → upsell/fidélité ; *insatisfaits* → réparation/geste avant relance.
- **Logistique (`delay_rate_ge3d`)**
  Taux de commandes livrées avec ≥3j de retard.
  → Indique une **friction opérationnelle** : avant de pousser une promo, **corriger SLA** / offrir un **express** / geste commercial.

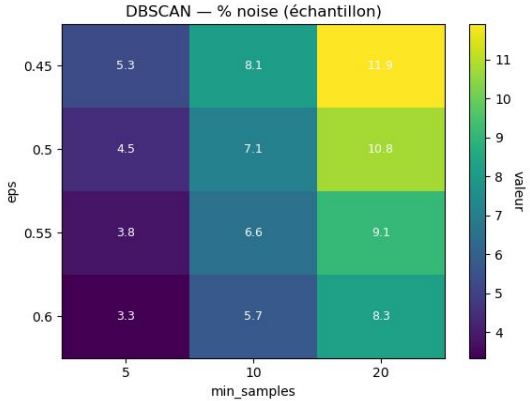
DBSCAN (Trop sensible, clusters peu actionnables)



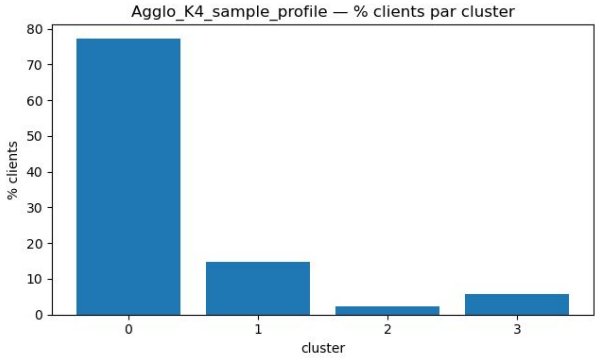


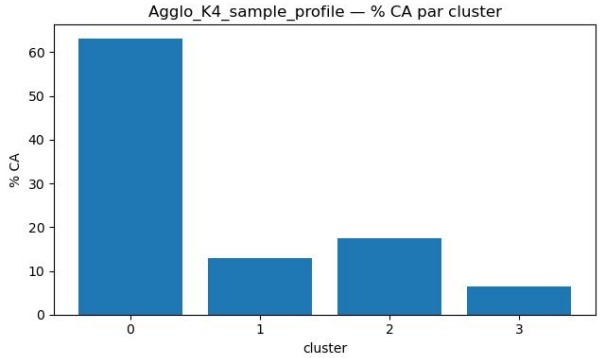


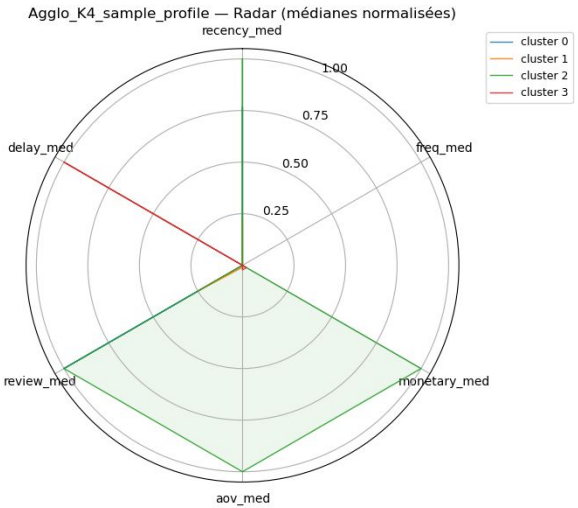


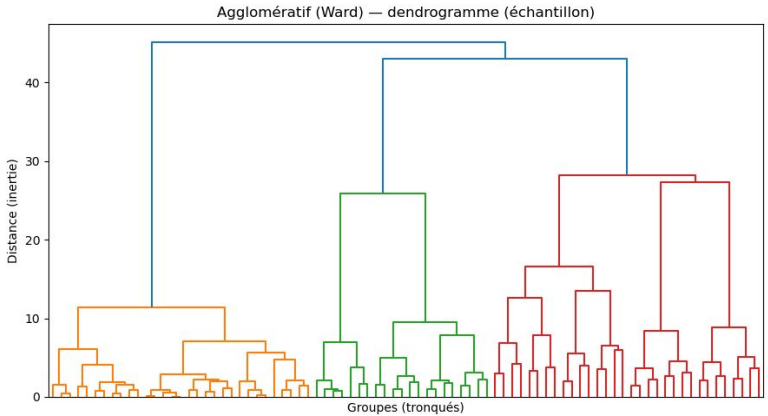


Hiérarchique (Gros cluster dominant, coûteux à maintenir)

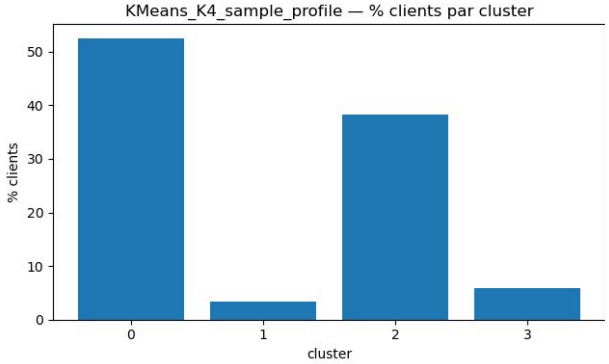


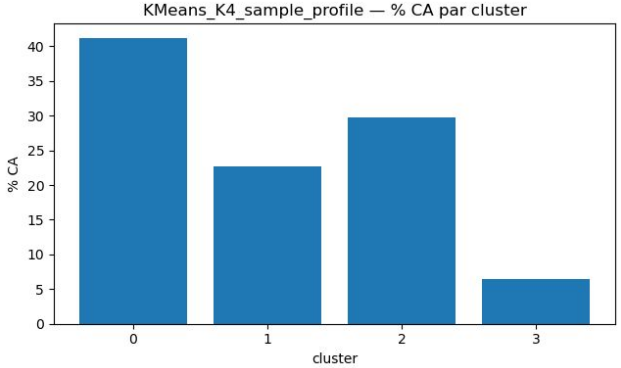


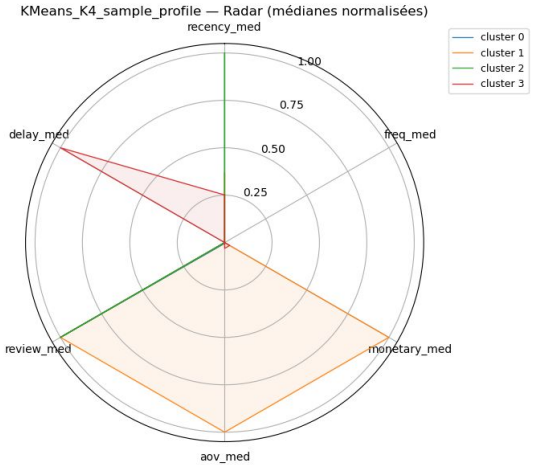




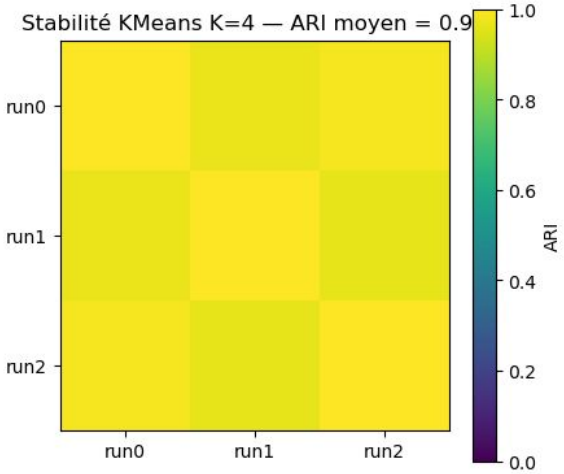
KMeans (Segments équilibrés et lisibles)

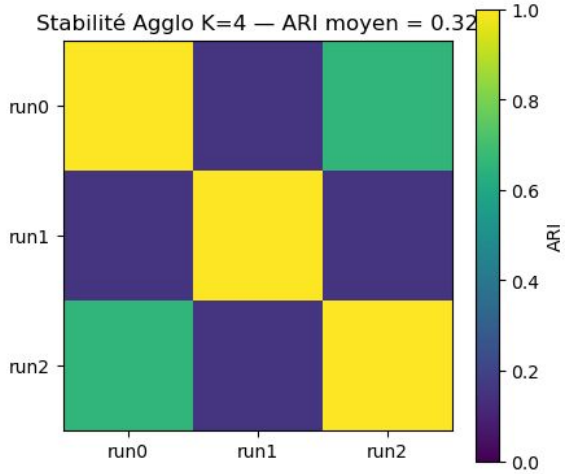


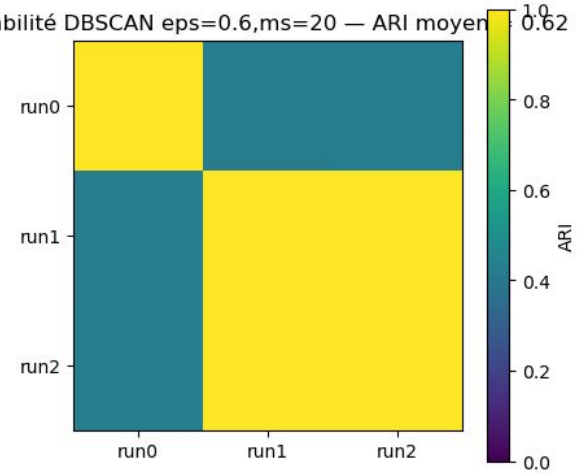




Stabilité des modèles

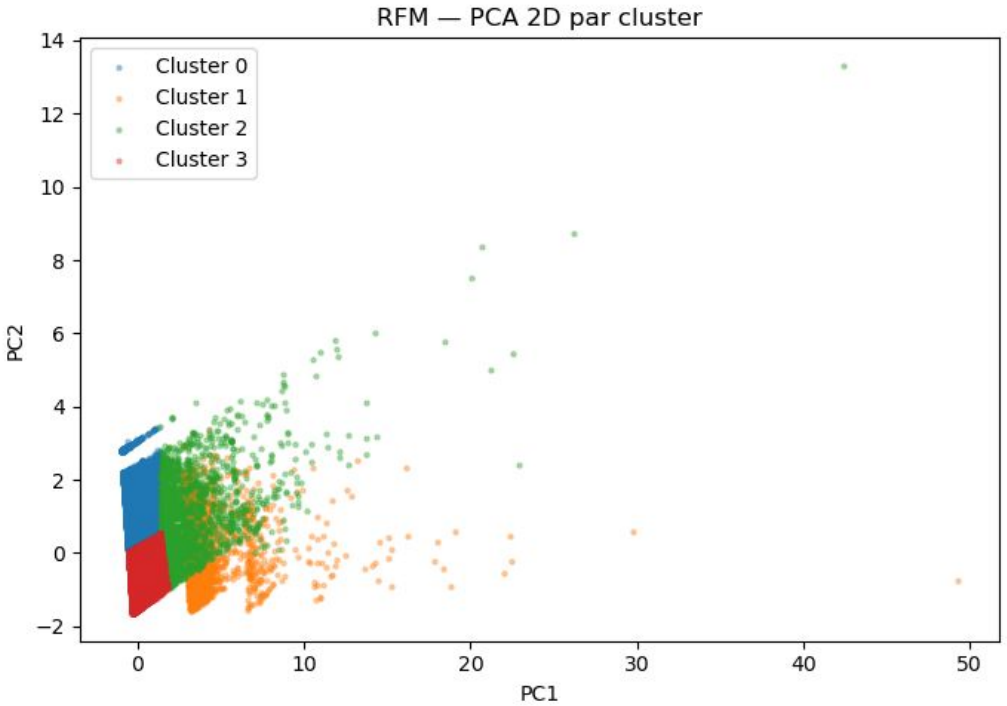


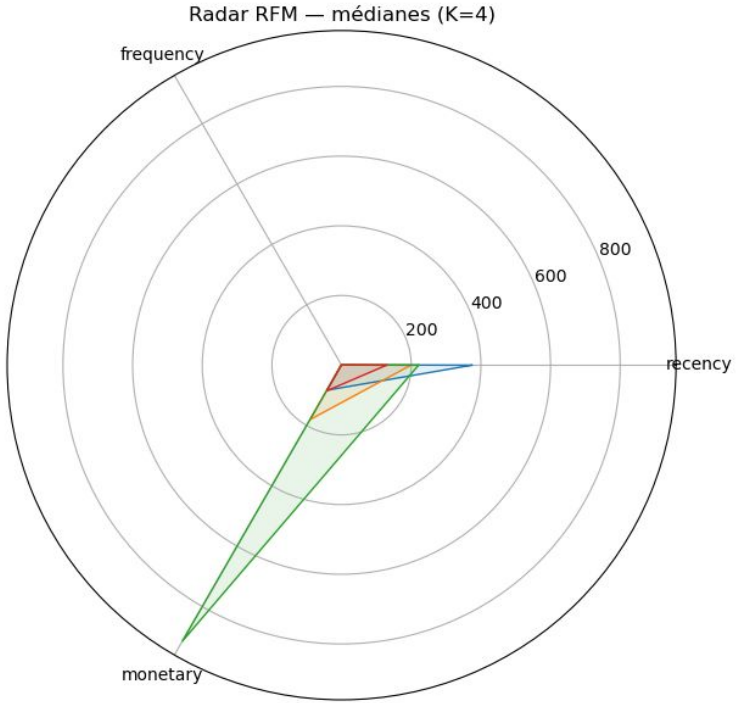




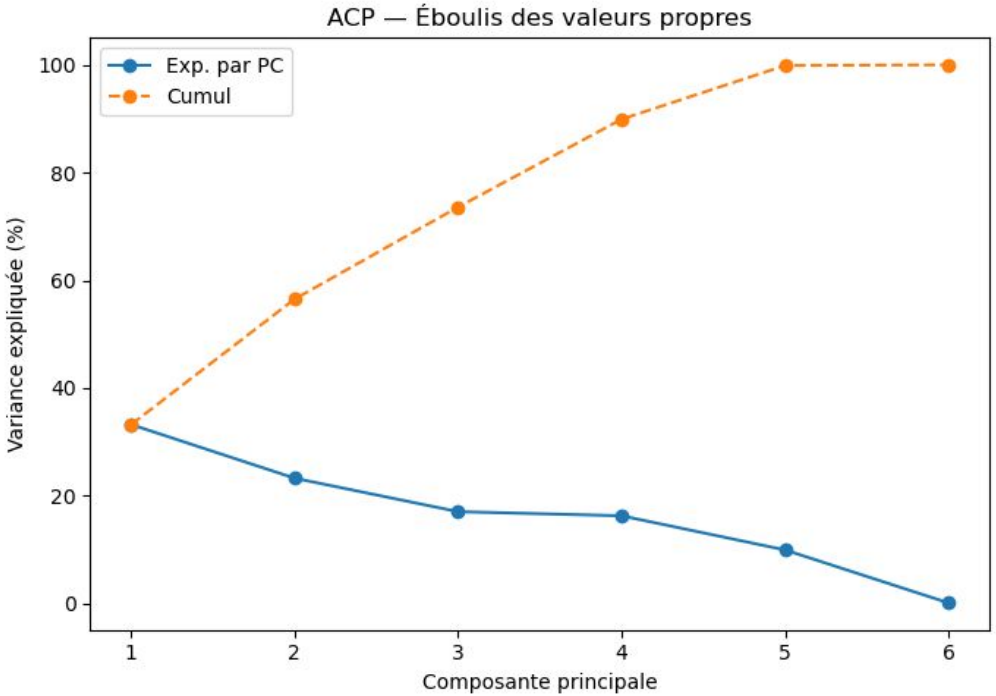
model	n_clusters	largest_share_%	silhouette	verdict
DBSCAN (best eps=0.6, ms=20)	2	97.04	0.665	⚠ Sensible aux params / bruit
Agglo K=4	4	77.26	0.395	⚠ Gros cluster dominant
KMeans K=4	4	52.5	0.347	Equilibré & segments nets

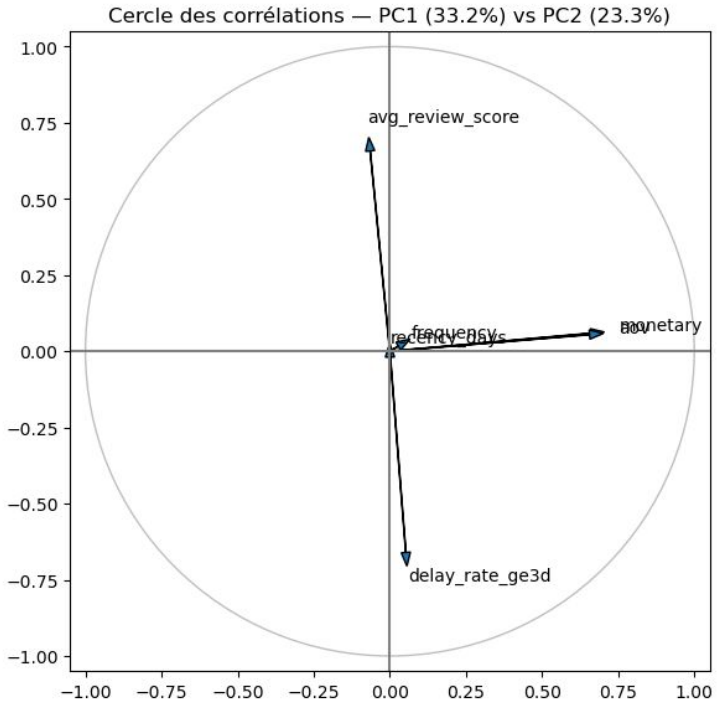
KMeans RFM (radar / PCA colorée)





Meilleurs modèle (éblouis + cercle de corrélation)



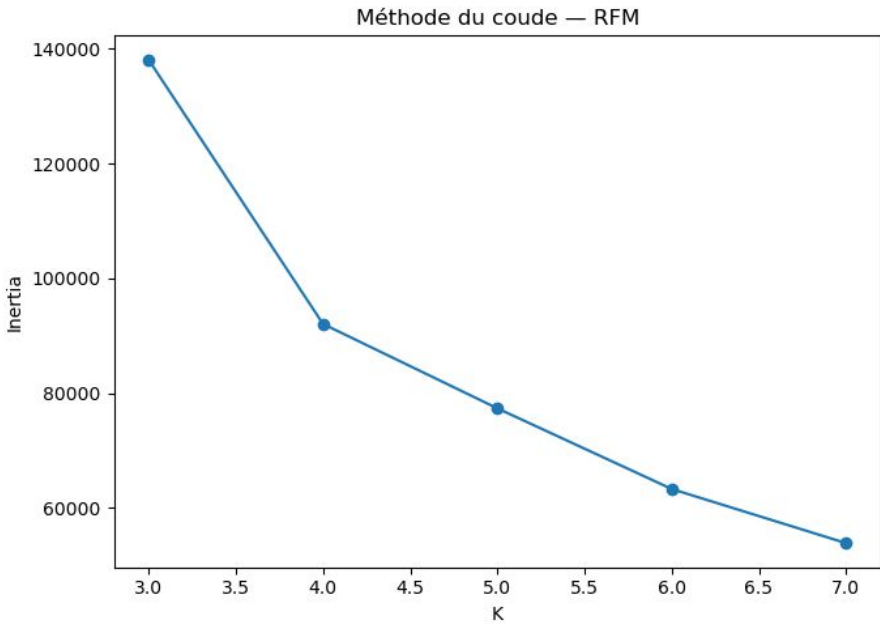


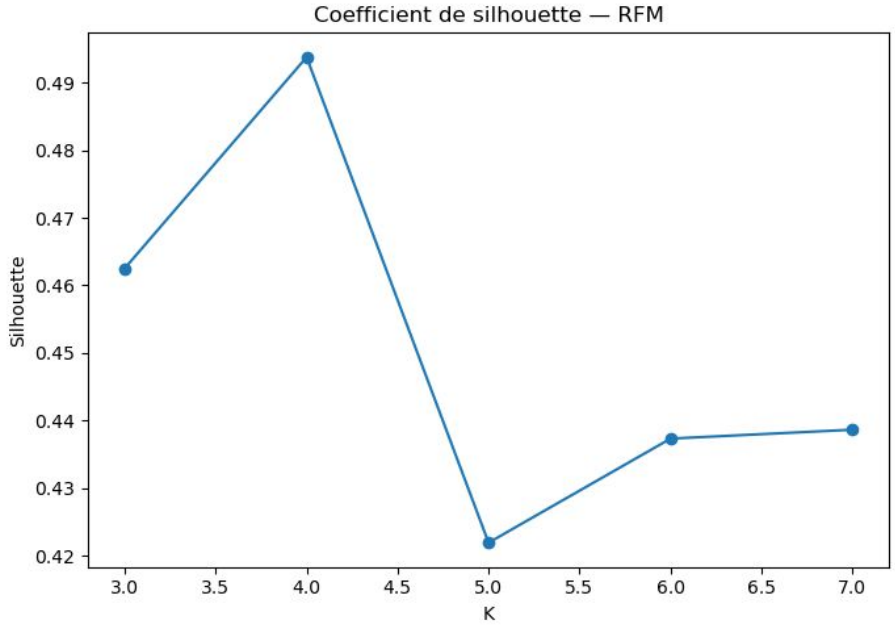
RFM (interprétation)

Cluster 3 — “Nouveaux/Actifs à convertir” (54% des clients, 44% du CA)	séries d’emails post-achat, bundles d’entrée, coupons 2e commande, recommandations produit
Cluster 0 — “One-shot anciens / froids” (40% des clients, 32% du CA)	campagnes win-back low-cost (email/SMS), offres flash
Cluster 1 — “Repeaters légers” (3% des clients, 5% du CA (2 achats)	programme fidélité simple, abonnements/auto-replenish, reco produits complémentaires
Cluster 2 — “High-Spenders / VIP dormants” (2% des clients, 19% du CA)	traitement VIP (support prioritaire, cadeaux), relance personnalisée haut de gamme, offres limitées

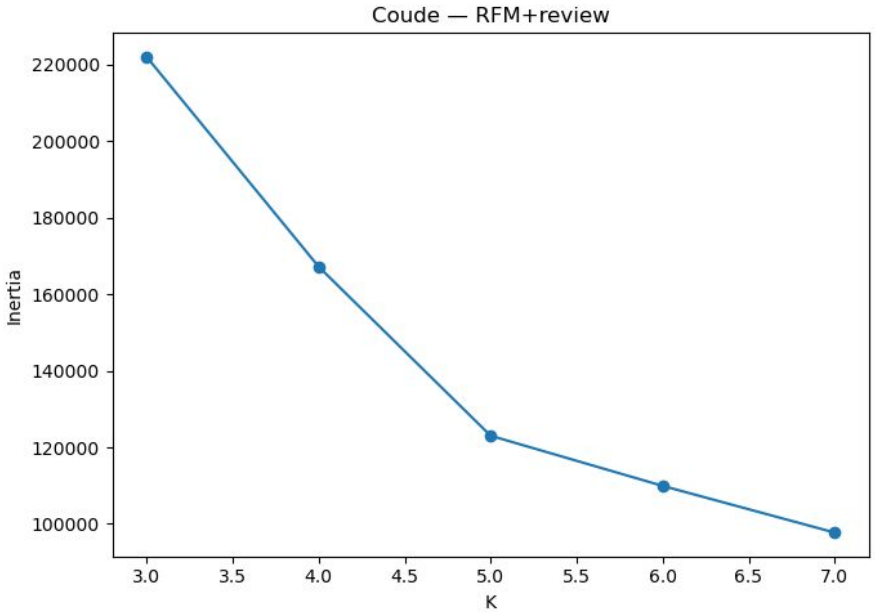
Cluster 3 à convertir / cluster 2 ne pas perdre

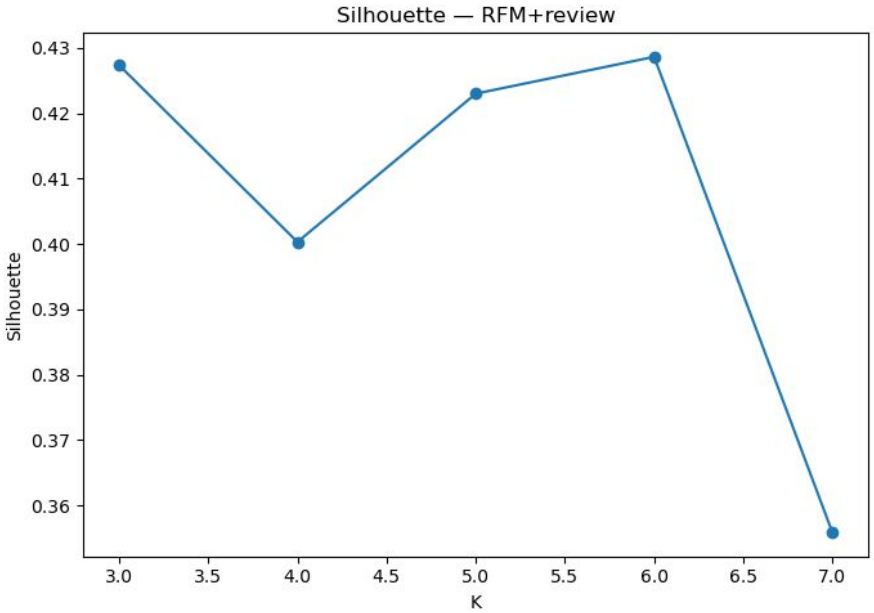
KMeans RFM (coude & silhouette)



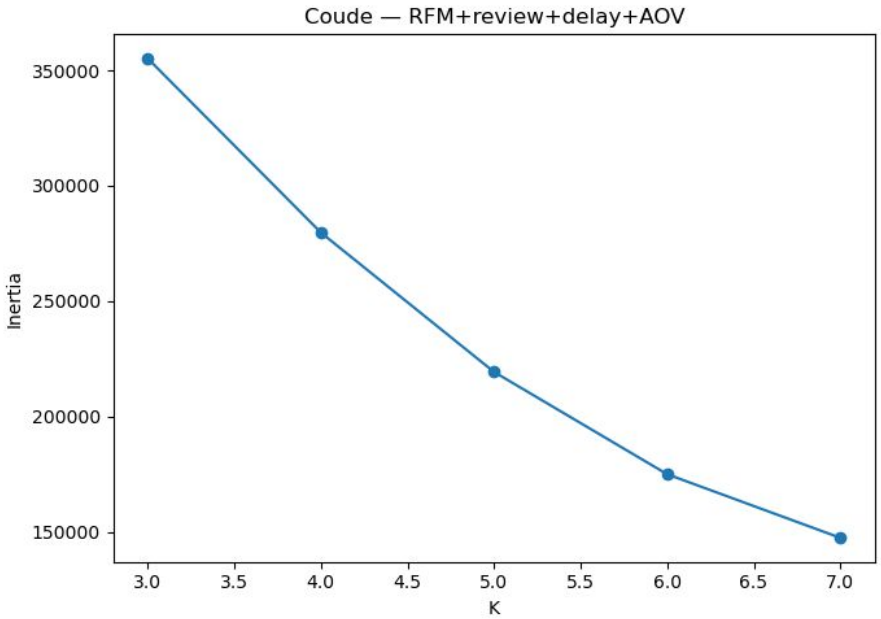


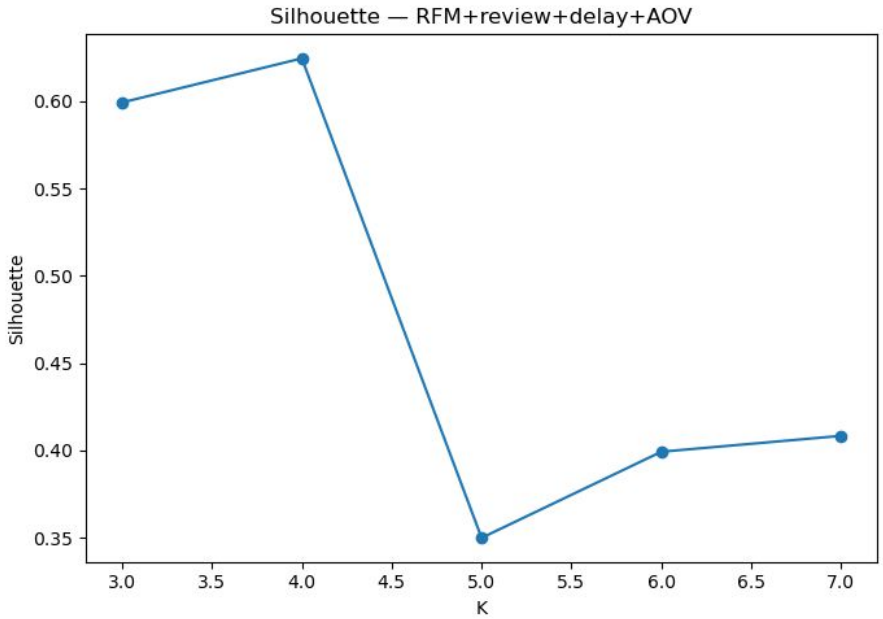
RFM + features supplémentaire



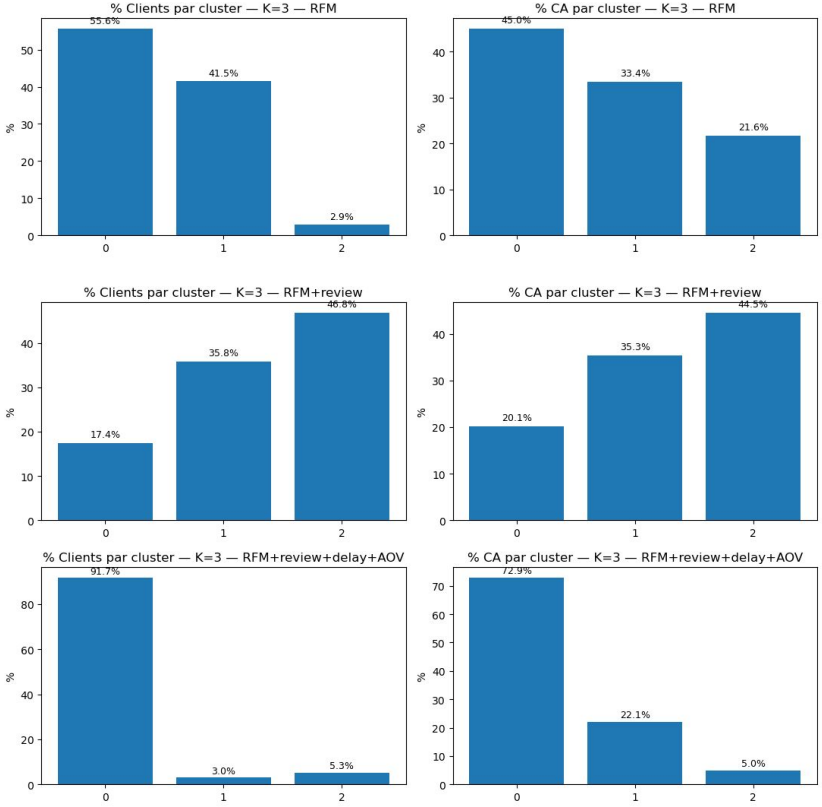


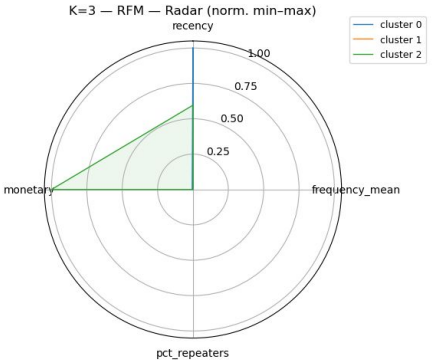
RFM + 3 features

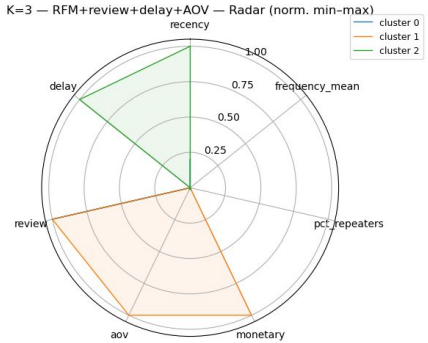


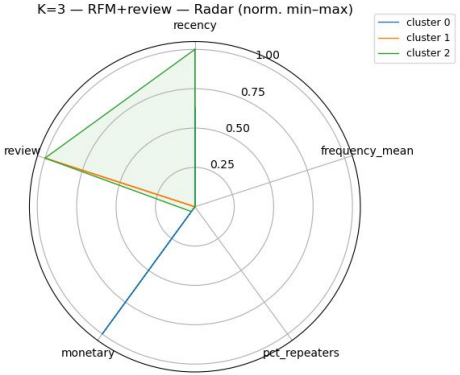


K3 RFM + variantes

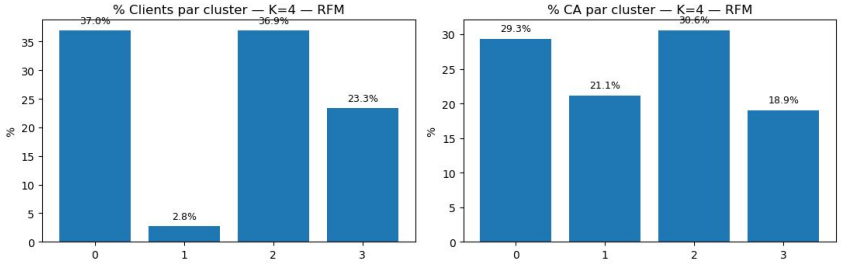


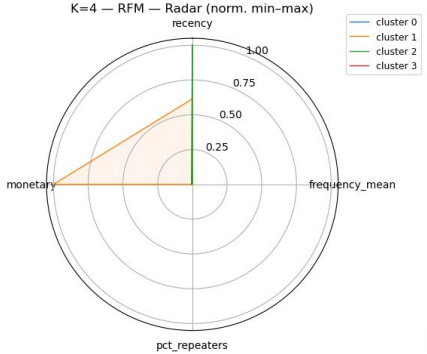


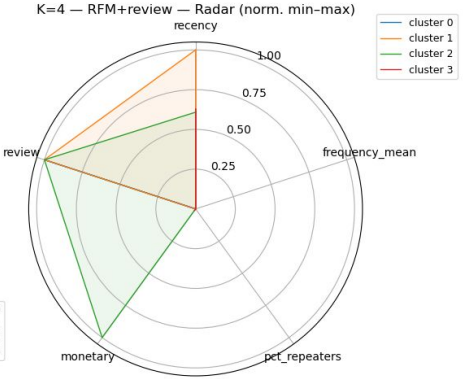


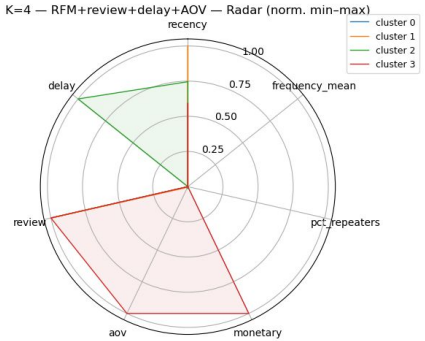


K4 RFM + variantes

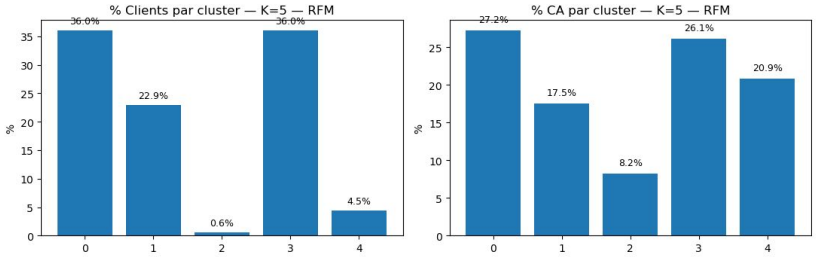


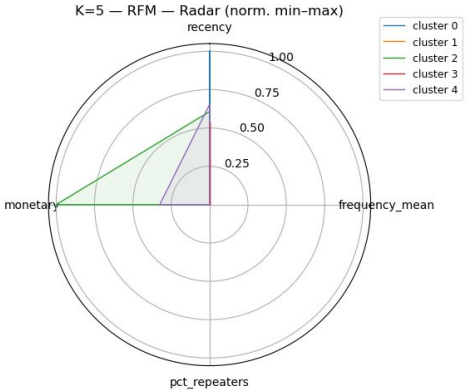


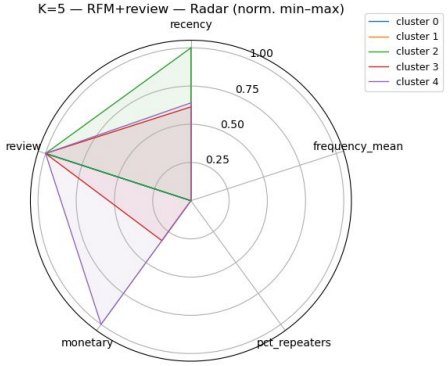


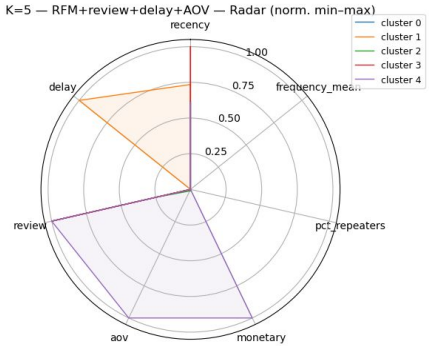


K5 RFM + variantes





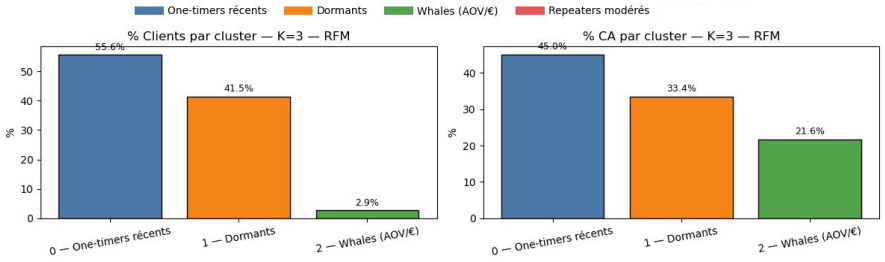


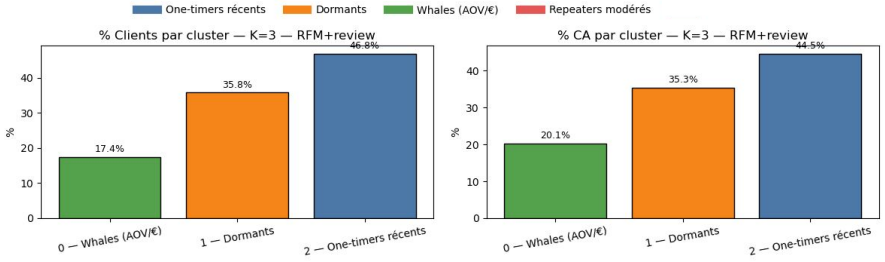


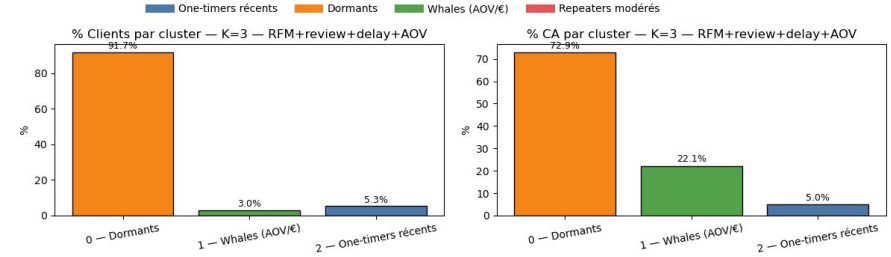
RFM (meilleur modèle)

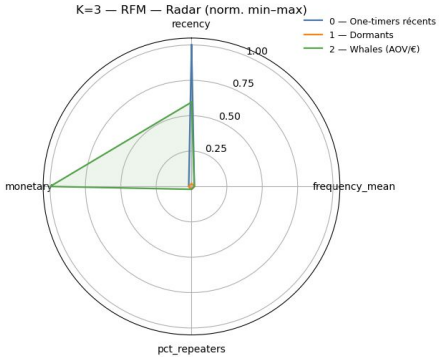
	k	3	4	5	6	7
variant	n_features					
RFM	3	0.463	0.494	0.422	0.437	0.439
RFM+review	4	0.427	0.400	0.423	0.429	0.356
RFM+review+delay+AOV	6	0.599	0.625	0.350	0.399	0.408

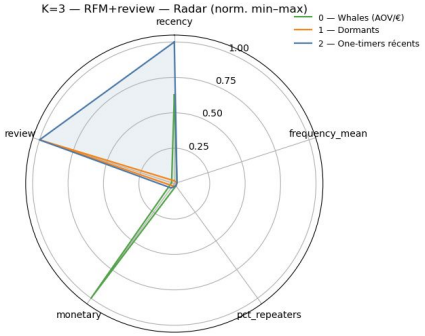
K3 marketing

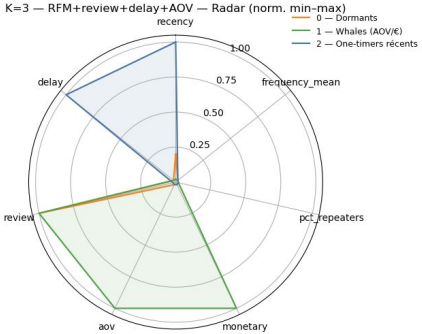




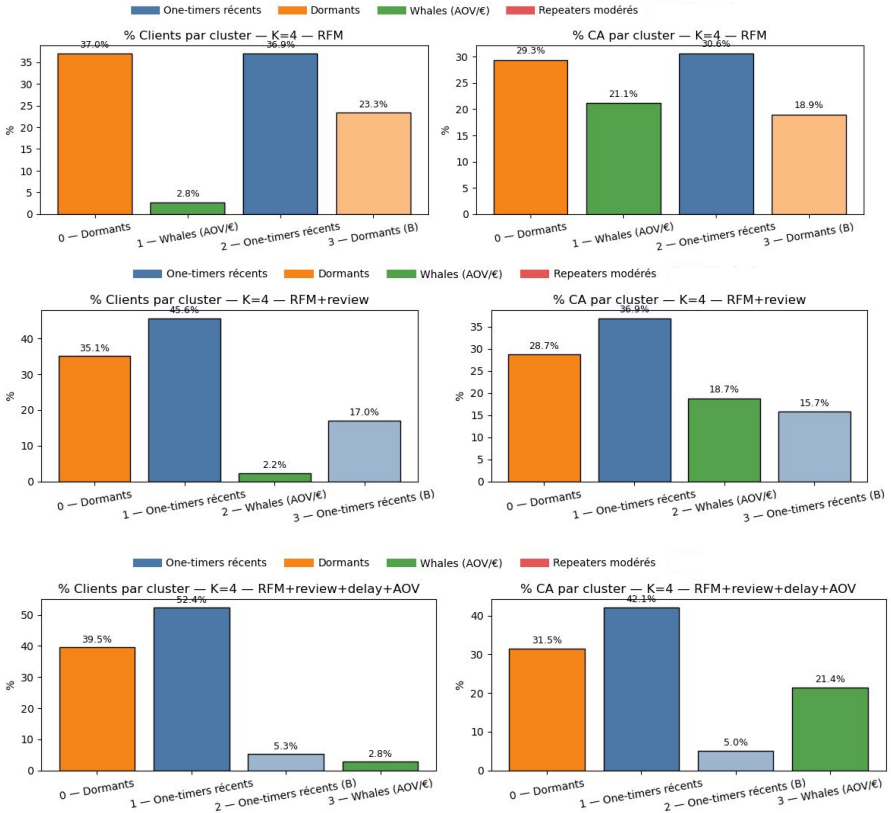


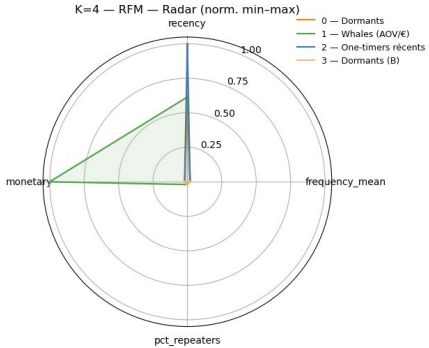


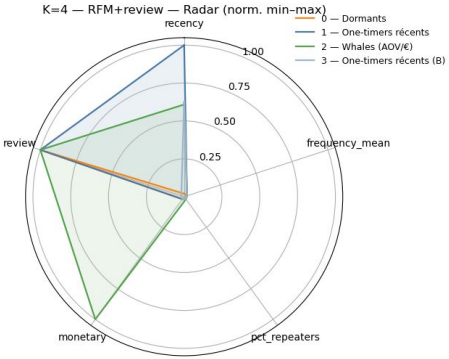


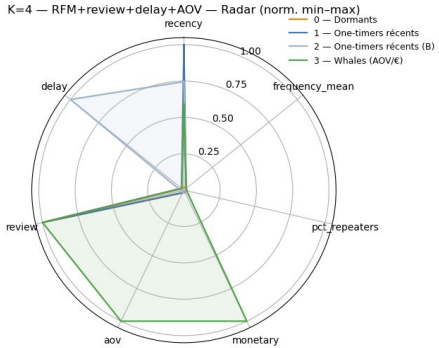


K4 marketing

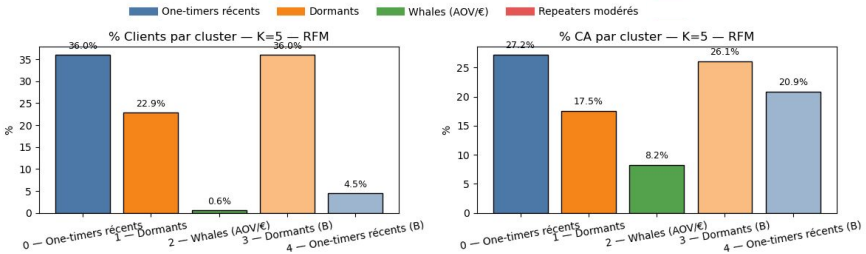


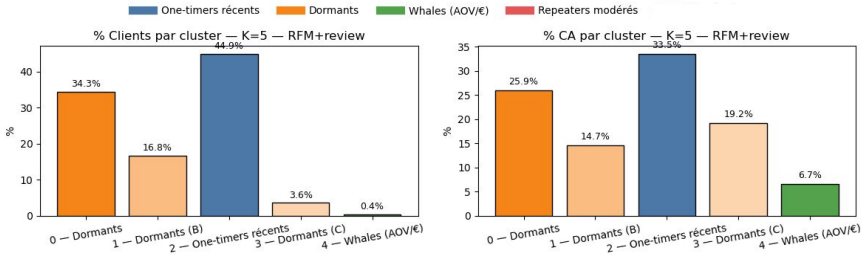


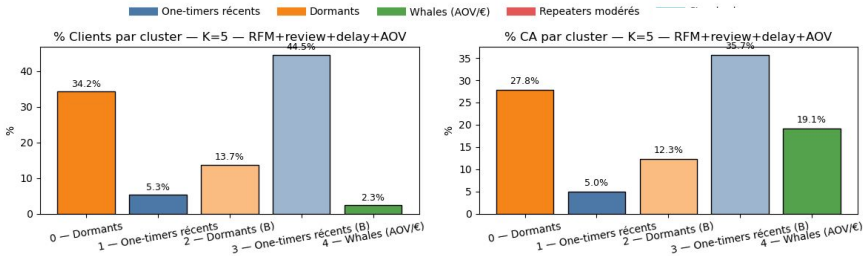


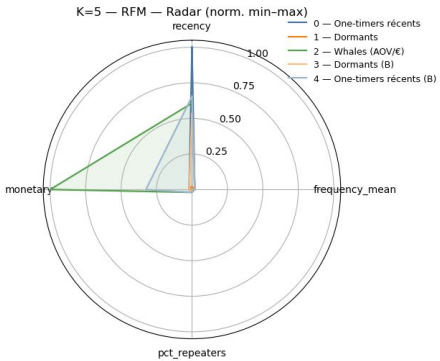


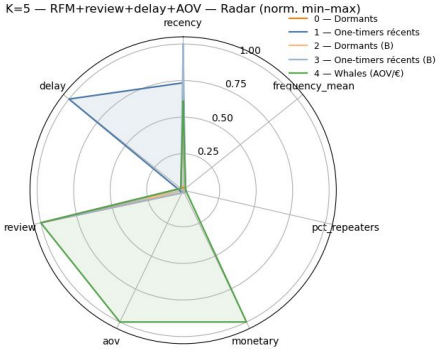
K5 marketing

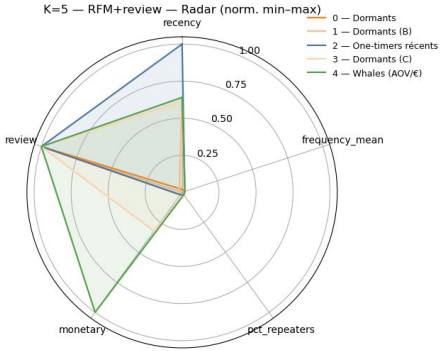


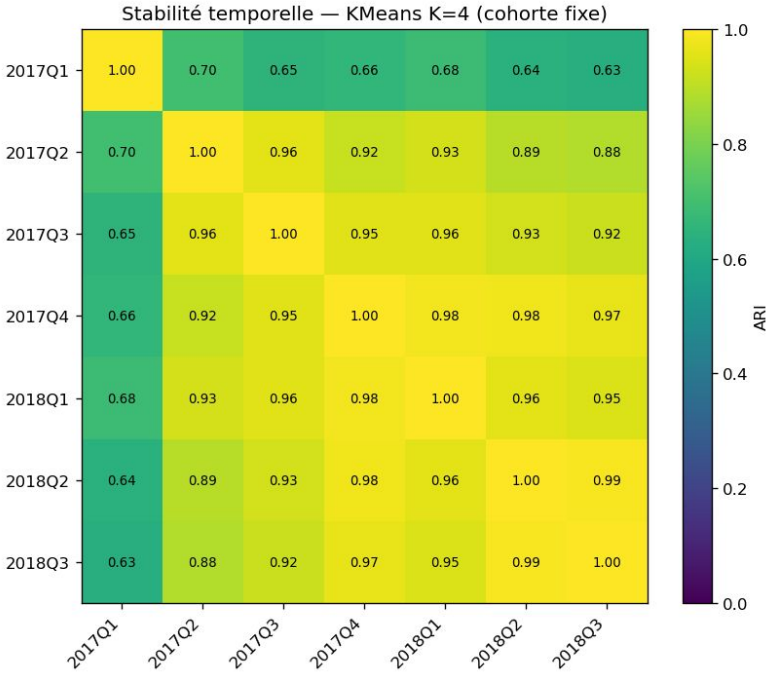


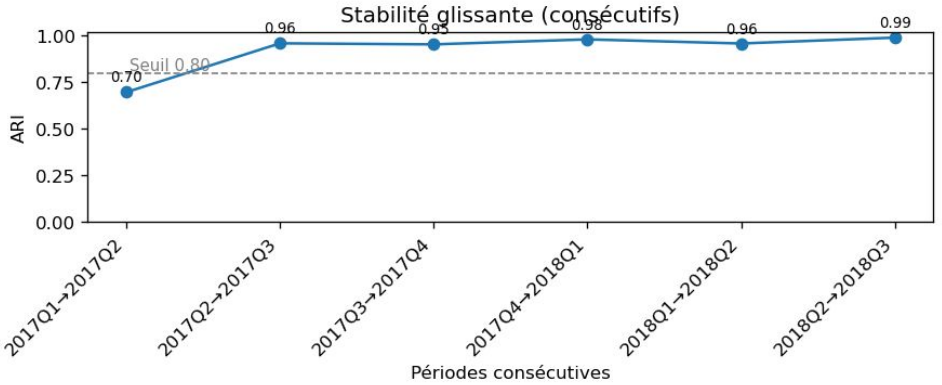


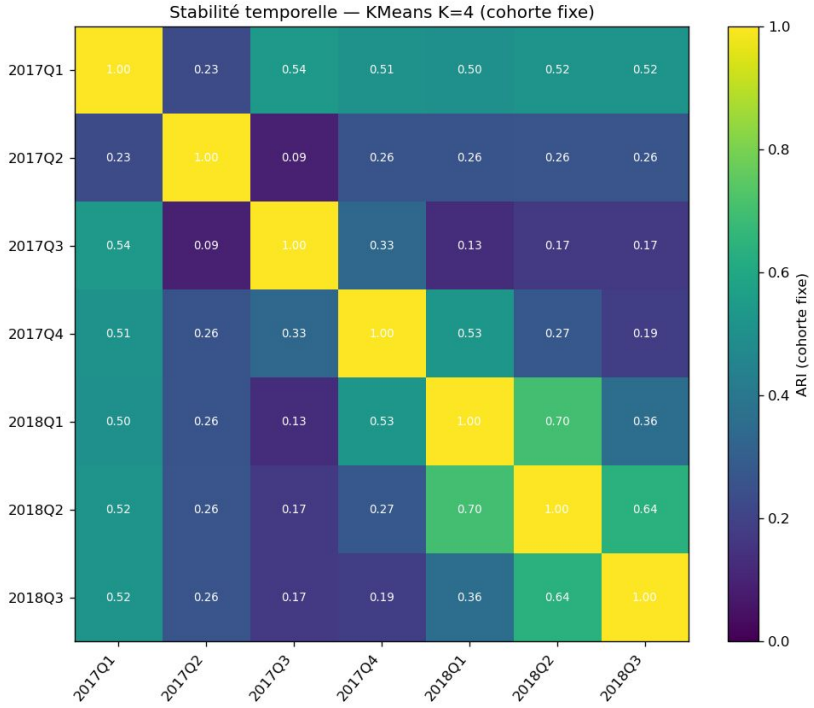


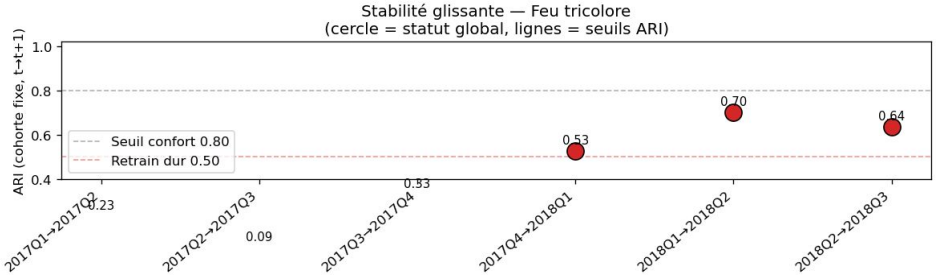


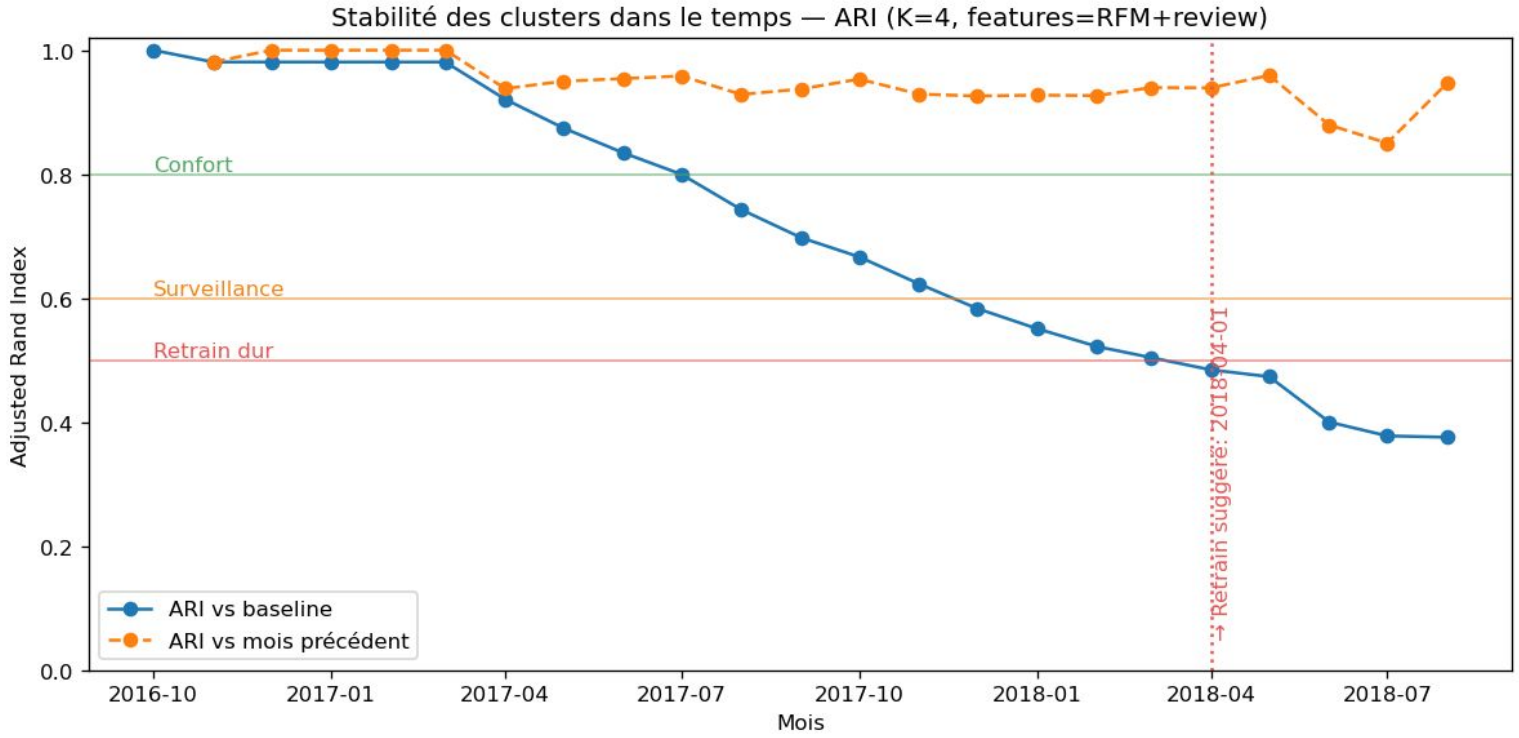



Stabilité temporelle — KMeans K=4 (cohorte fixe)
2017Q1
2017Q2
2017Q3
2017Q4
2018Q1
2018Q2
2018Q3
1.00 0.70 0.65 0.66 0.68 0.64 0.63
0.70 1.00 0.96 0.92 0.93 0.89 0.88
0.65 0.96 1.00 0.95 0.96 0.93 0.92
0.66 0.92 0.95 1.00 0.98 0.98 0.97
0.68 0.93 0.96 0.98 1.00 0.96 0.95
0.64 0.89 0.93 0.98 0.96 1.00 0.99
0.63 0.88 0.92 0.97 0.95 0.99 1.00
ARI
1.0
0.8
0.6
0.4
0.2
0.0


Stabilité glissante (consécutifs)
ARI
1.00
0.75
0.50
0.25
0.00
Seuil 0.80
0.70
0.96
0.95
0.98
0.96
0.99
2017Q1→2017Q2
2017Q2→2017Q3
2017Q3→2017Q4
2017Q4→2018Q1
2018Q1→2018Q2
2018Q2→2018Q3
Périodes consécutives


Stabilité temporelle — KMeans K=4 (cohorte fixe)
2017Q1
2017Q2
2017Q3
2017Q4
2018Q1
2018Q2
2018Q3
1.00 0.23 0.54 0.51 0.50 0.52 0.52
0.23 1.00 0.09 0.26 0.26 0.26 0.26
0.54 0.09 1.00 0.33 0.13 0.17 0.17
0.51 0.26 0.33 1.00 0.53 0.27 0.19
0.50 0.26 0.13 0.53 1.00 0.70 0.36
0.52 0.26 0.17 0.27 0.70 1.00 0.64
0.52 0.26 0.17 0.19 0.36 0.64 1.00
ARI (cohorte fixe)
1.0
0.8
0.6
0.4
0.2
0.0


Stabilité glissante — Feu tricolore
(cercle = statut global, lignes = seuils ARI)
ARI (cohorte fixe, t→t+1)
1.0
0.8
0.6
0.4
Seuil confort 0.80
Retrain dur 0.50
0.23
0.09
0.33
0.53
0.70
0.64
2017Q1→2017Q2
2017Q2→2017Q3
2017Q3→2017Q4
2017Q4→2018Q1
2018Q1→2018Q2
2018Q2→2018Q3


Stabilité des clusters dans le temps — ARI (K=4, features=RFM+review)
Adjusted Rand Index
1.0
0.8
0.6
0.4
0.2
0.0
Confort
Surveillance
Retrain dur
→ Retrain suggéré: 2018-04-01
ARI vs baseline
ARI vs mois précédent
2016-10
2017-01
2017-04
2017-07
2017-10
2018-01
2018-04
2018-07
Mois

Conclusion

One-timers récents (beaucoup de clients, récents, freq=1, € moyen)	**Objectif :** déclencher la 2e commande	séries d'emails avec top ventes, "acheté ensemble", et **offre douce** (ex : livraison offerte dès X€ plutôt que -10%).
Dormants (anciens, freq=1, peu engagés, € faible à moyen)	**Objectif :** réactiver / comprendre les freins.	nouveautés, best-sellers, livraison gratuite / -X€, Sondage
Whales (AOV/€ élevé ; reviews souvent bonnes ; fréquence pas forcément haute)	**Objectif :** préserver la valeur, encourager la récurrence	traitement VIP (support prioritaire, cadeaux), relance personnalisée haut de gamme, offres limitées
Repeaters modérés / "Prometteurs" (freq ≥2, % repeaters > moyenne, € moyen)	**Objectif :** faire passer de **2→3→4 achats** et stabiliser la cadence.	abonnement, remises