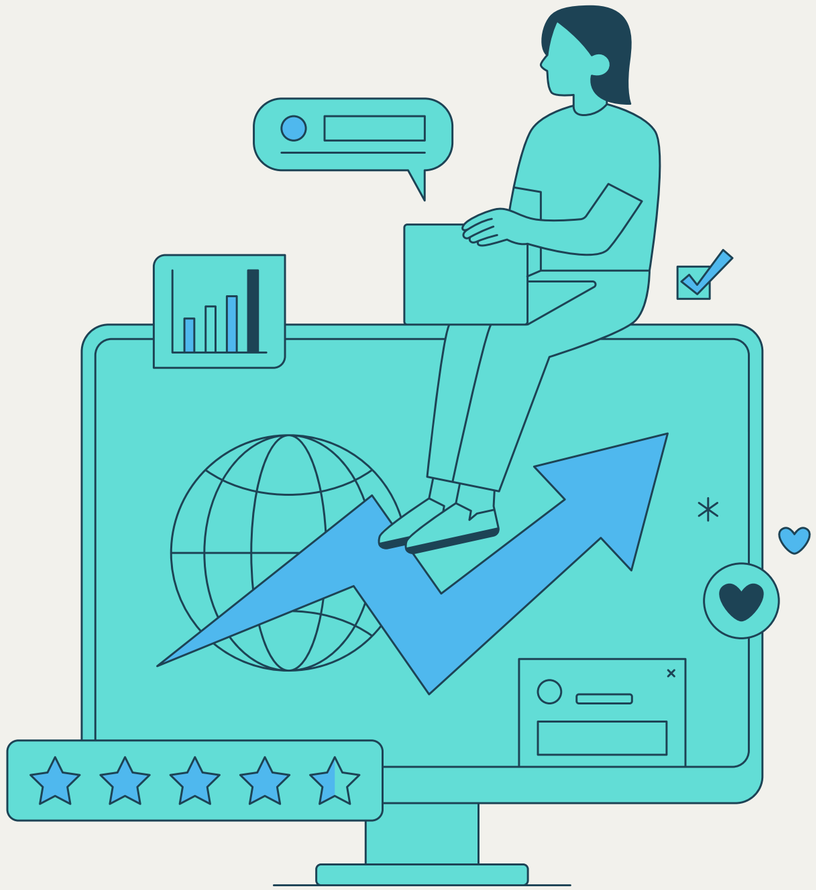
Segmentation Clients & Contrat de Maintenance

Mission Olist – Consultant Data Science

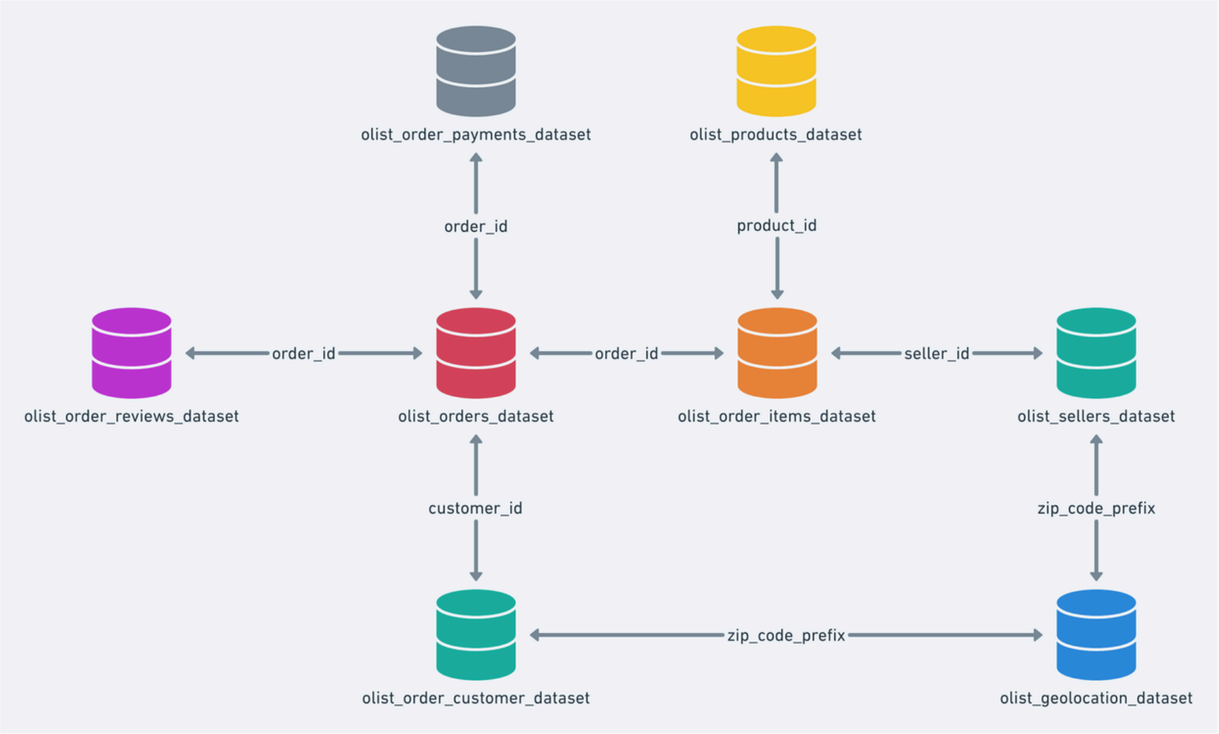
Mai 2025

Natascha Minnitt



Contexte & Problématique

- Commandes : 98 666
- Clients : 96 096
- Produits : 32 951
- Real brésilien : + 16 millions



1. Segmentation pertinente

2. Maintenance

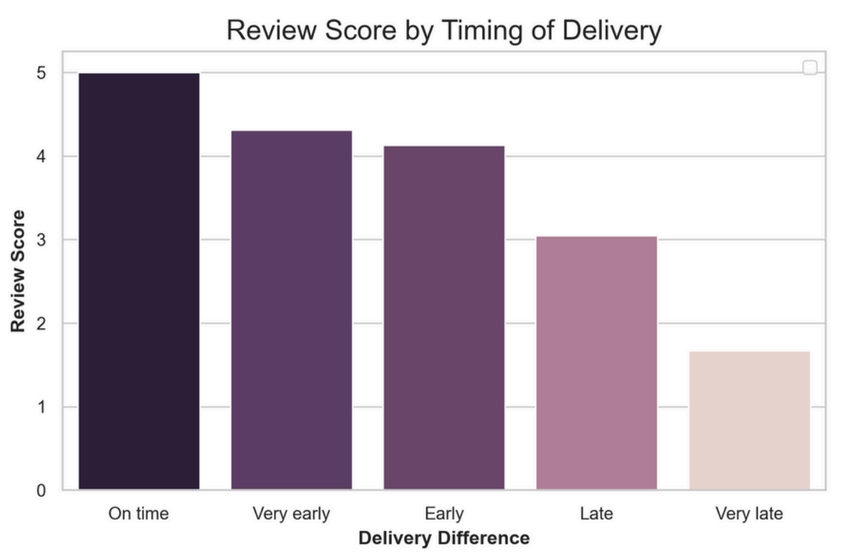
Traitement des données

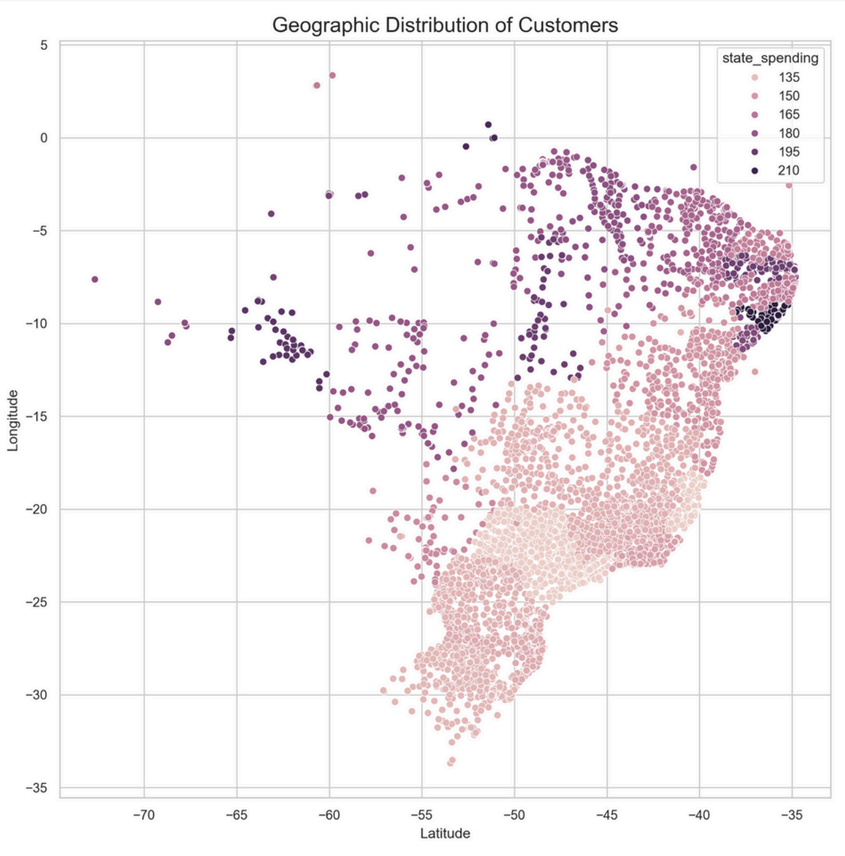
1. Nettoyage & création du fichier client

- 2017-05 → 2018-05
- Agrégation par client (3% returning)

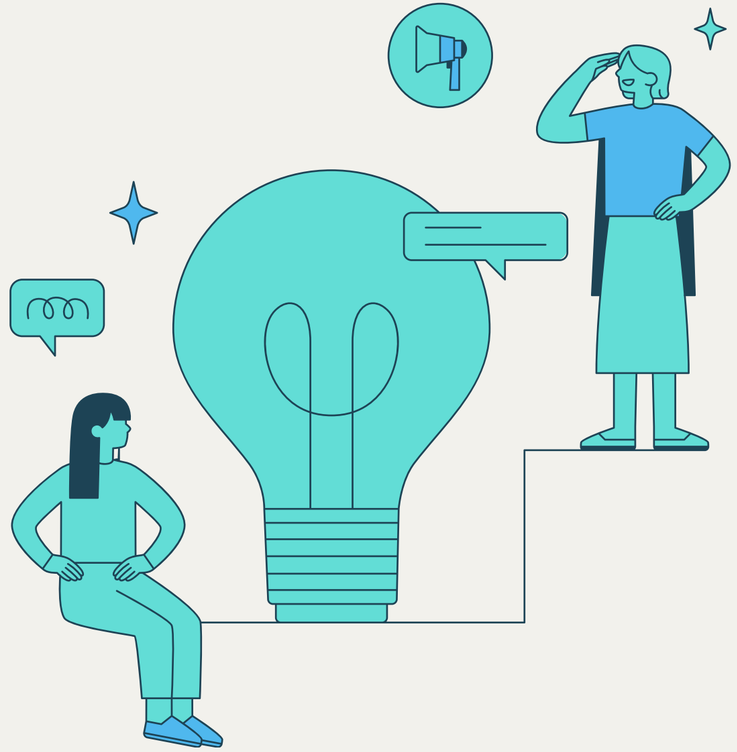
2. Feature Engineering

[1] Récence	Date de command max
[2] Fréquence(retour)	oui = 1 ‖ non = 0
[3] Montant	Somme et transformation log
[4] Note d'évaluation	Moyenne (1 - 5)
[5] Différence de livraison	attendu - réel
[6] Echéances	Moyenne
[7] Pouvoir d'achat	1 = SP / RJ ‖ 0 = autre
[8] Dépenses de la région	Moyenne
[9] Popularité du produit	Récence + retour par catégorie





Modélisation & sélection

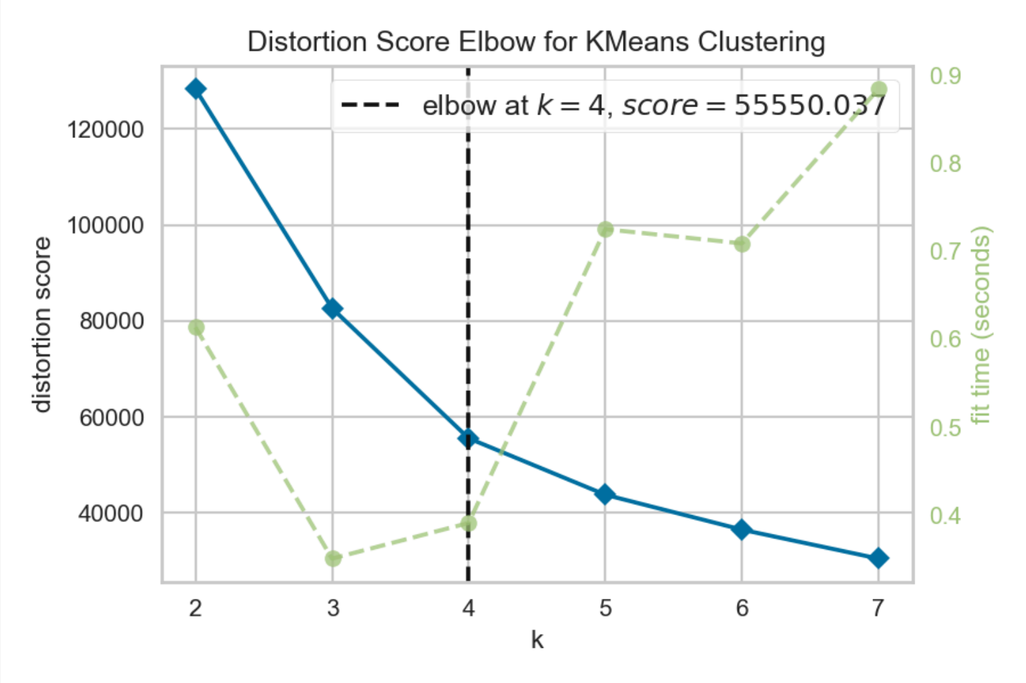


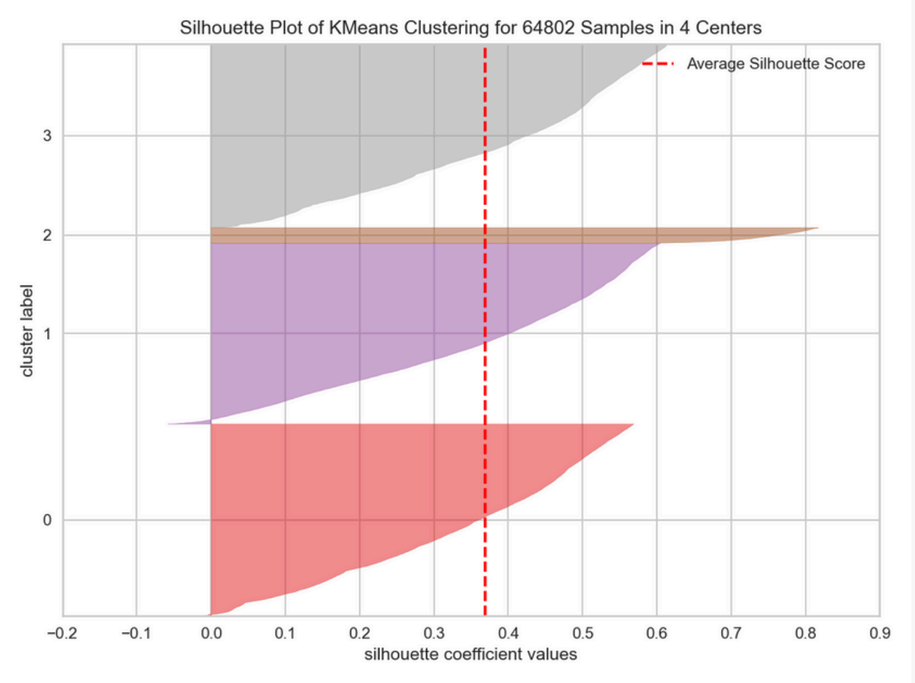
K-Means Clustering

	K	Silhouette score ↑	Calinski–Harabasz ↑	Davies–Bouldin ↓	Distortion ↓	Temps ↓
M0 : RFM	4	**0.37**	**53 990**	**0.76**	**55 550**	**0.4**
M1 : RFM + Note d'évaluation	4	0.33	32 779	0.97	102 957	0.4
M2 : RFM + Différence de livraison	4	0.26	25 580	1.09	118 667	0.5
M3 : RFM + Pouvoir d'achat	3	0.37	32 252	0.94	129 898	0.4
M4 : RFM + Dépenses de la région	4	0.31	29 993	1.00	108 517	0.5
M5 : RFM + échéances	4	0.33	35 549	1.06	97 967	0.5
M6 : RFM + Popularité du produit	4	0.29	27 851	1.06	113 217	0.6

K-Means Clustering

	K	Silhouette score ↑	Calinski–Harabasz ↑	Davies–Bouldin ↓	Distortion ↓	Temps ↓
M0 : RFM	4	**0.37**	**53 990**	**0.76**	**55 550**	**0.4**

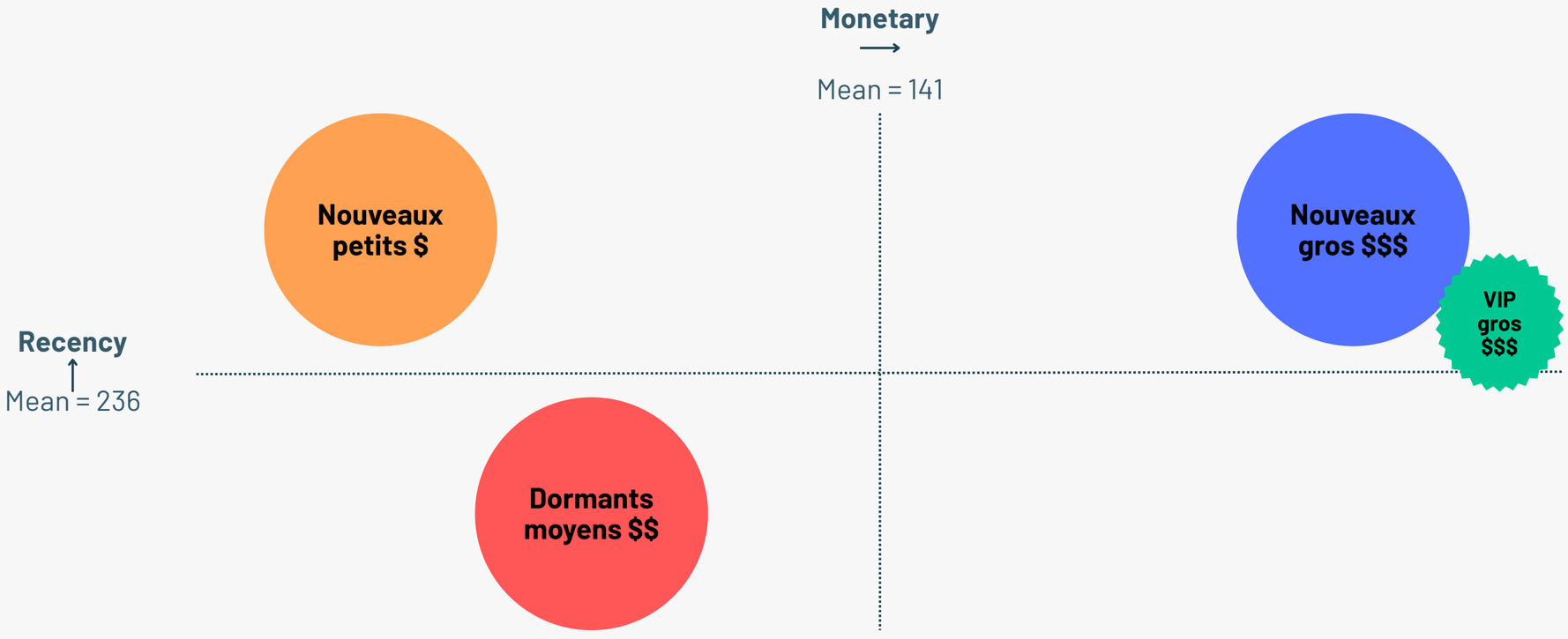




K-Means Clustering

M0 : RFM avec 4 clusters

Profile	N	Récence [m = 156]	Retour [m = 0.03]	Montant [m = 139]
[0] Nouveaux gros dépensiers	21 759	101	0	251
[1] Nouveaux petits dépensiers	20 799	100	0	43
[2] Clients récurrents, gros dépensiers [VIP]	1766	142	1	253
[3] Clients dormants, dépensiers moyens	20 479	271	0	109



K-Means Clustering

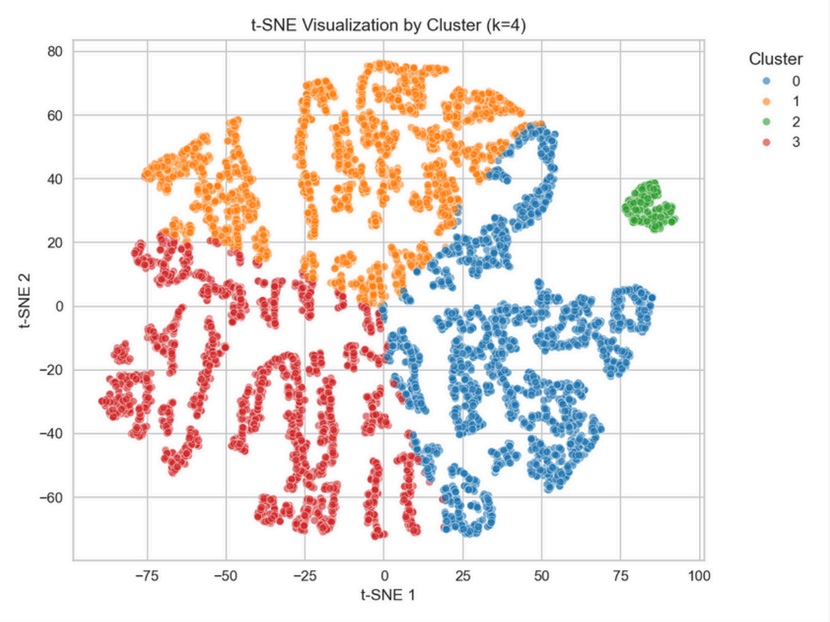
M0 : RFM avec 4 clusters

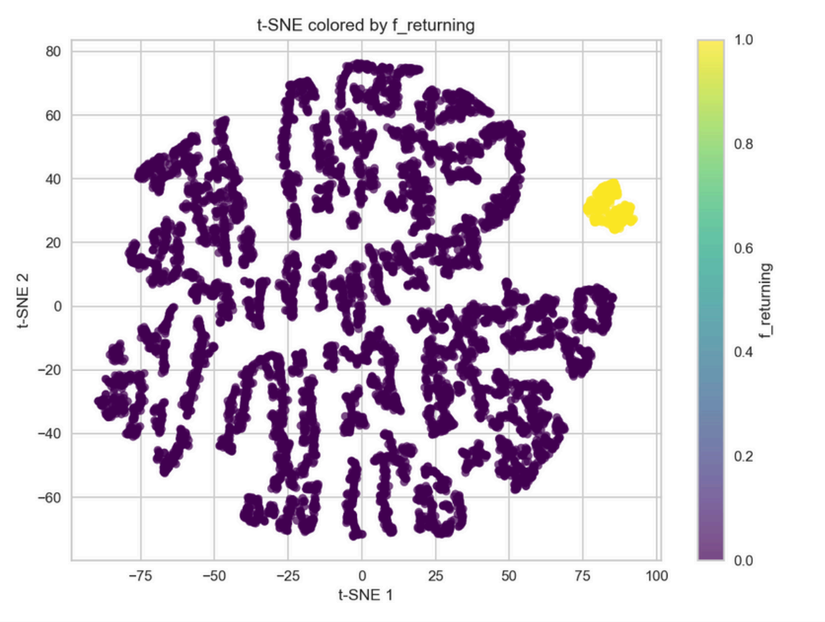
[0] Nouveaux gros dépensiers

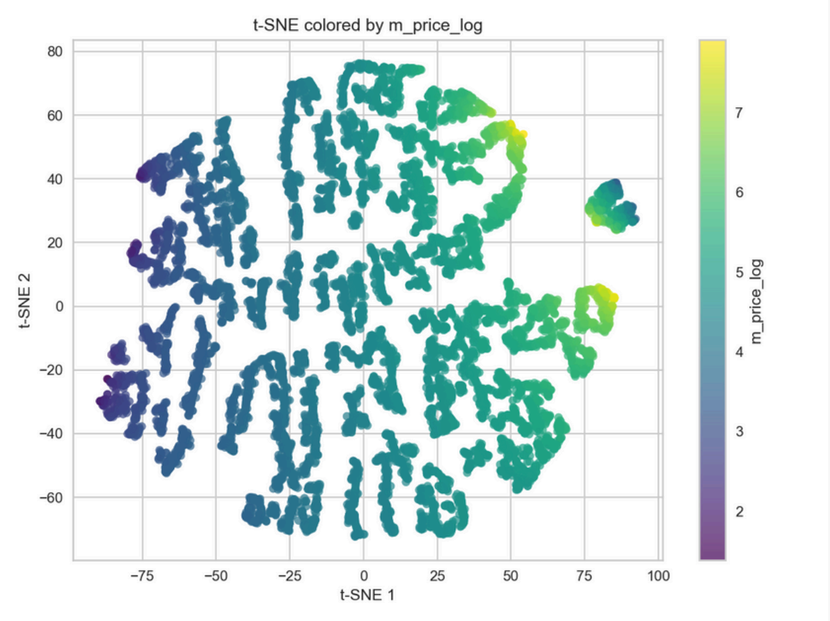
[1] Nouveaux petits dépensiers

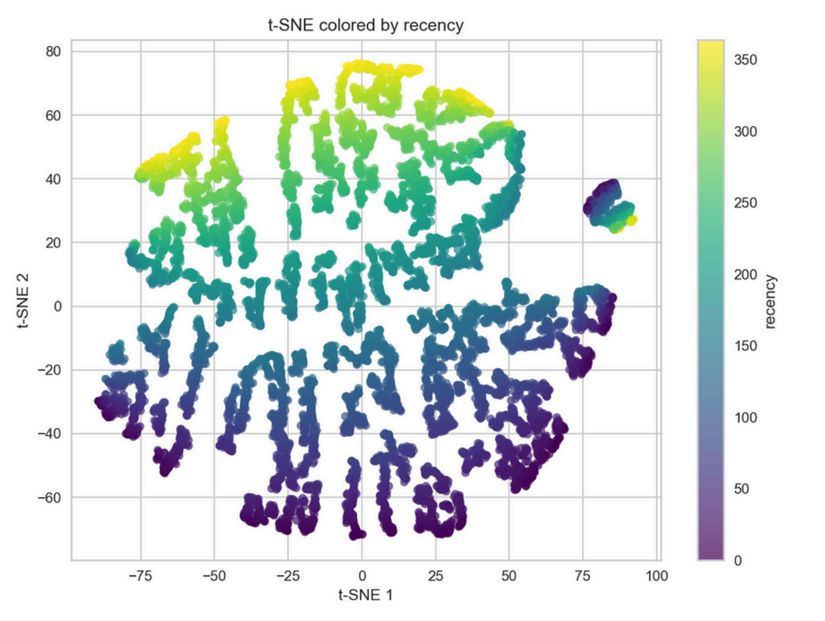
[2] Clients récurrents, gros dépensiers [VIP]

[3] Clients dormants, dépensiers moyens



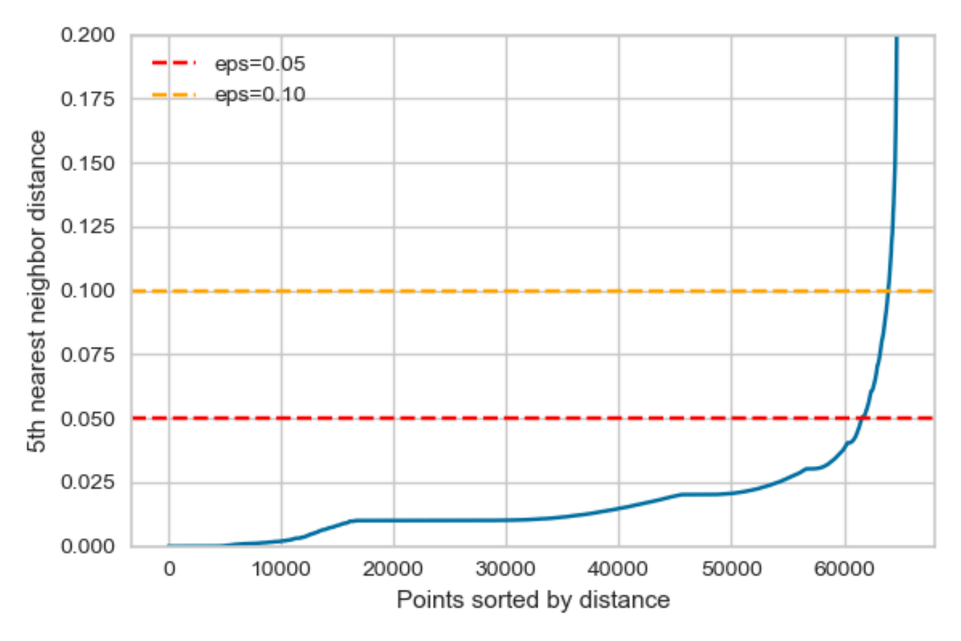


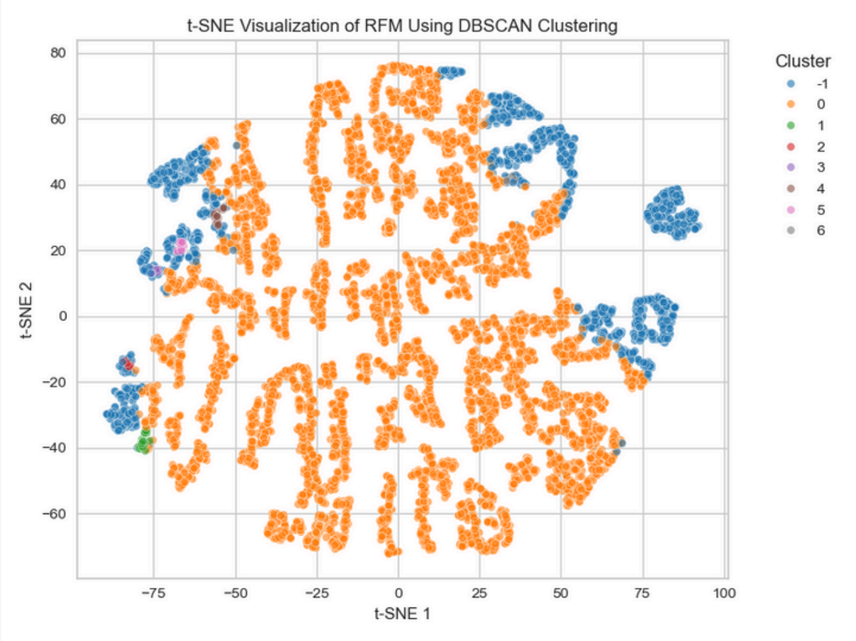




DBSCAN Clustering

M0 : RFM





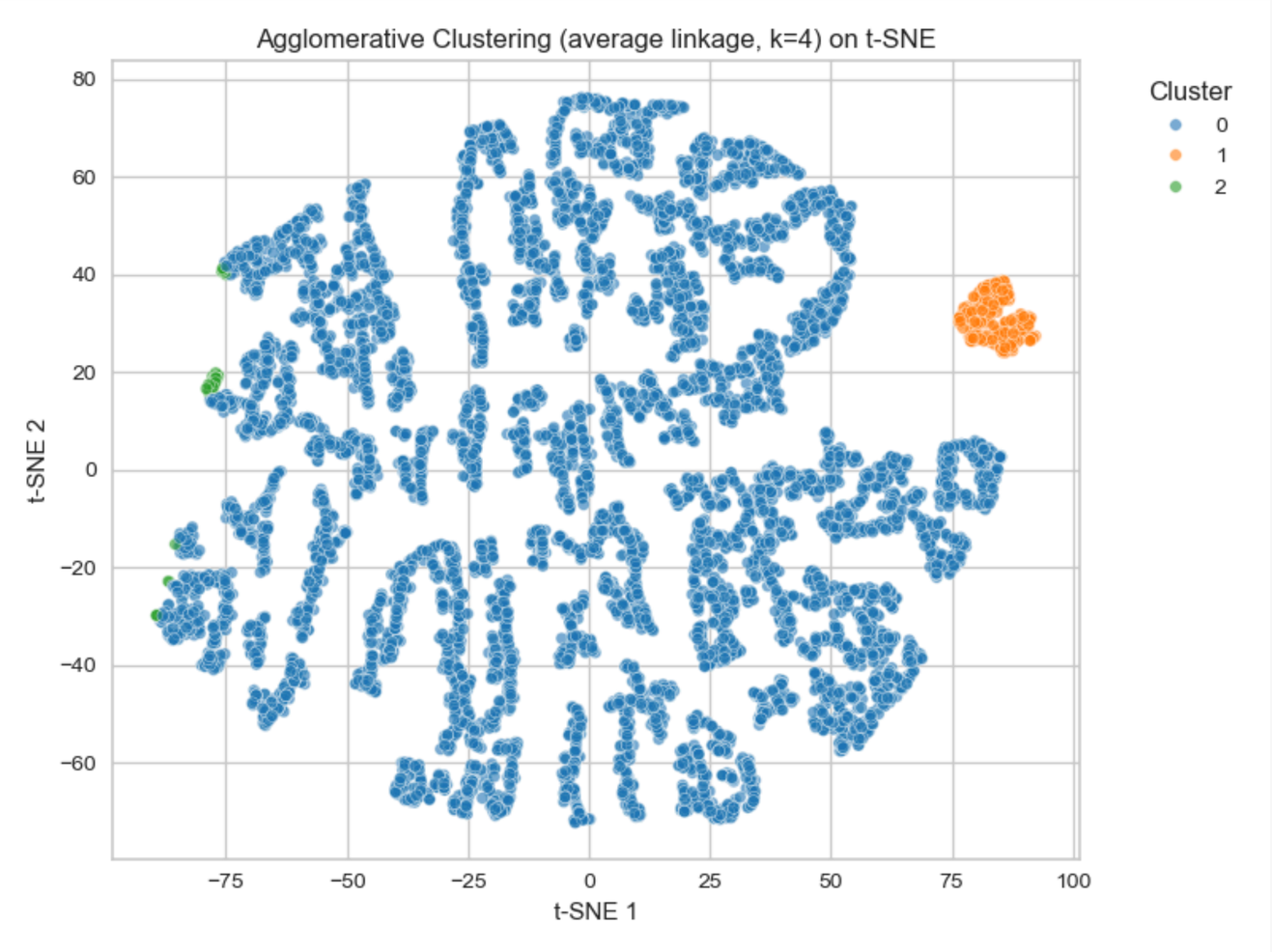
epsilon	min sample	clusters	bruit	silhouette
0.10	10	41	53 821	0.46
0.10	100	27	1 400	0.31
0.05	10	69	4 383	-0.25
0.05	100	8	11 065	-0.33

N Clusters

- -1: 11 065
- 0: 52 795
- 1: 262
- 2: 130
- 3: 119
- 4: 158
- 5: 163
- 6: 101
- 7: 9

Agglomératif Clustering

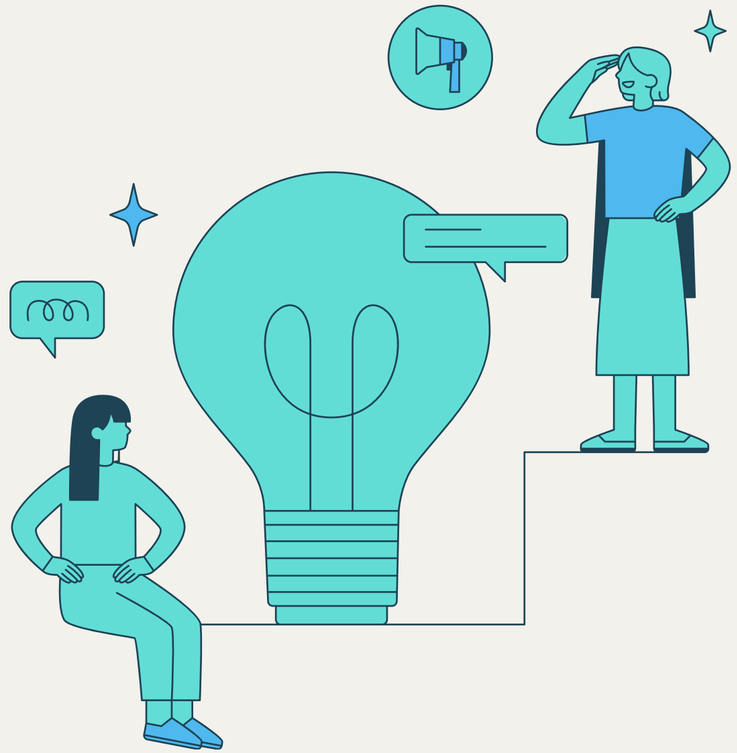
M0 : RFM



k = 3 → silhouette = 0.34

n = 10 000

Contrat de maintenance



Simulation de dérive – ARI dans le temps

Dataframe complet

09-2016 → 08-2018

Dataframe essaie : M0

05-2017 → 05-2018

M1

M0 + 2 semaines

M2

M0 + 4 semaines

M3

M0 + 6 semaines

M4

M0 + 8 semaines

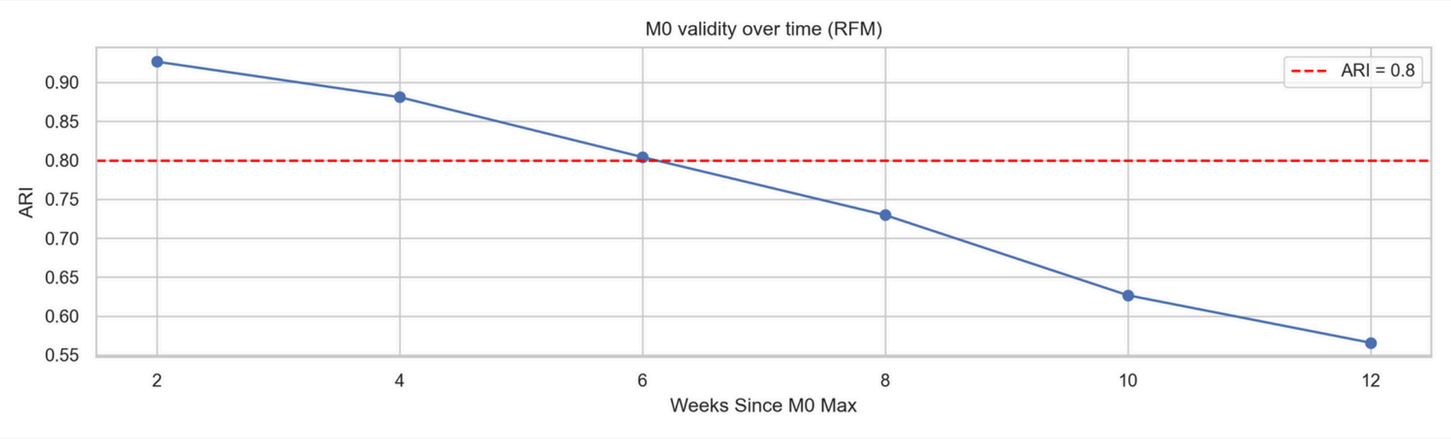
M5

M0 + 10 semaines

M6

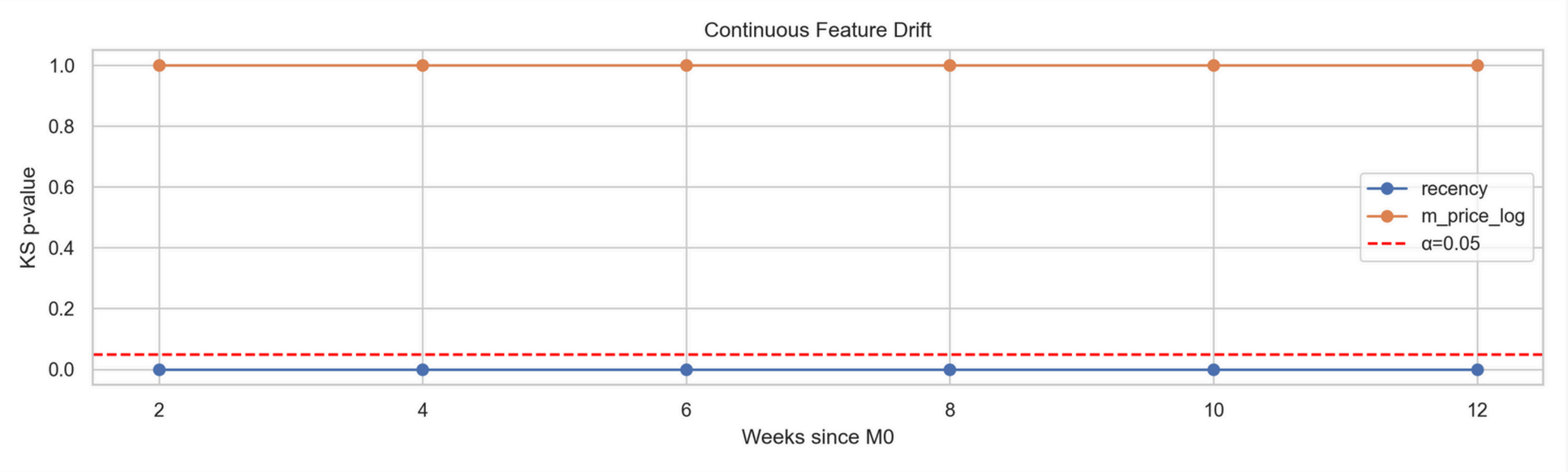
M0 + 12 semaines → 08-2018

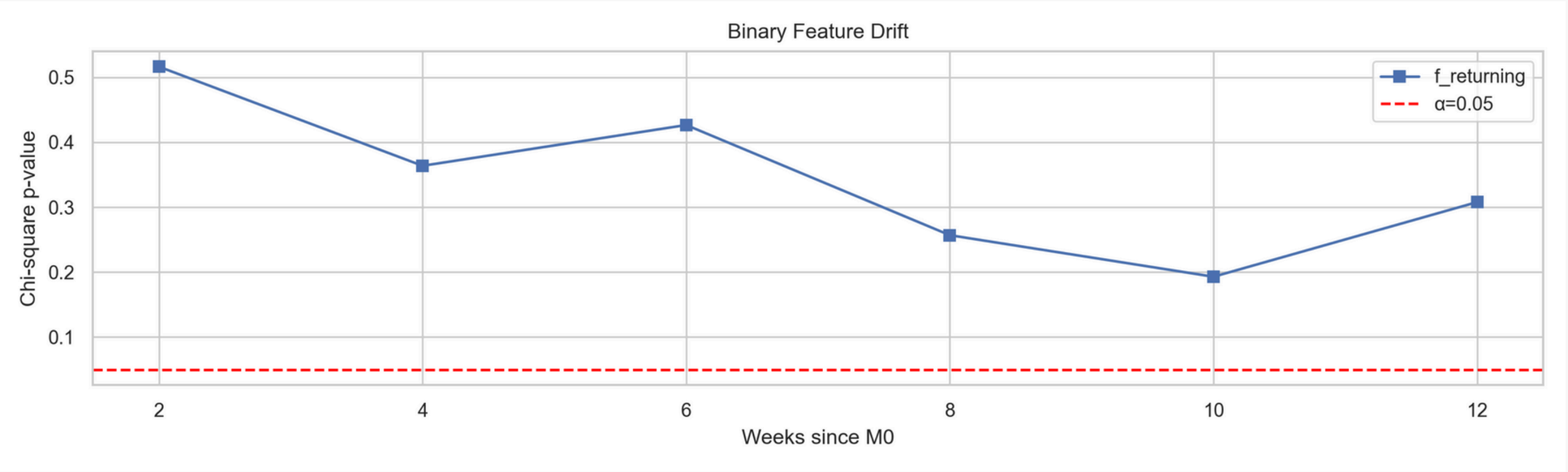
Réentraîner le modèle



Tests de dérive des features

M0 : RFM





Conclusions

- Le K-means sur RFM est le plus performant
- 4 segments exploitables
- Les segments restent valides environ 6 semaines
- La dérive des variables confirme cette cadence