Projet de segmentation Olist

Support de présentation

Segmentation des clients
d’un site e-commerce

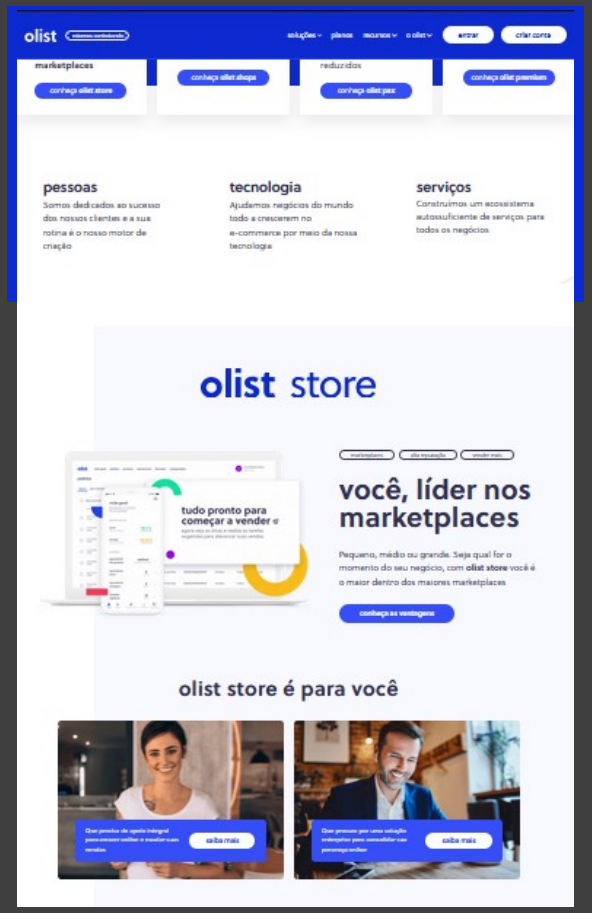
Juillet 2021

Sommaire

1. Rappel de la problématique
2. Analyse exploratoire et transformation des données
3. Méthodes employées : Clusterisations, NMF, RFM
4. Interprétation et exploitation possible des résultats
5. Proposition de maintenance
6. Bilan de projet

1. Rappel de la problématique

https://olist.com/



Mission :

Aider les équipes d’ Olist à comprendre les différents types de consommateurs.

➡ En utilisant des méthodes non supervisées dans le but de regrouper les utilisateurs ayant des comportements similaires.

1. Réaliser une segmentation clients sur la base des données fournies
2. Proposer une « notation » client exploitable pour l’équipe marketing
3. Proposer une maintenance pour la mise à jour de la segmentation

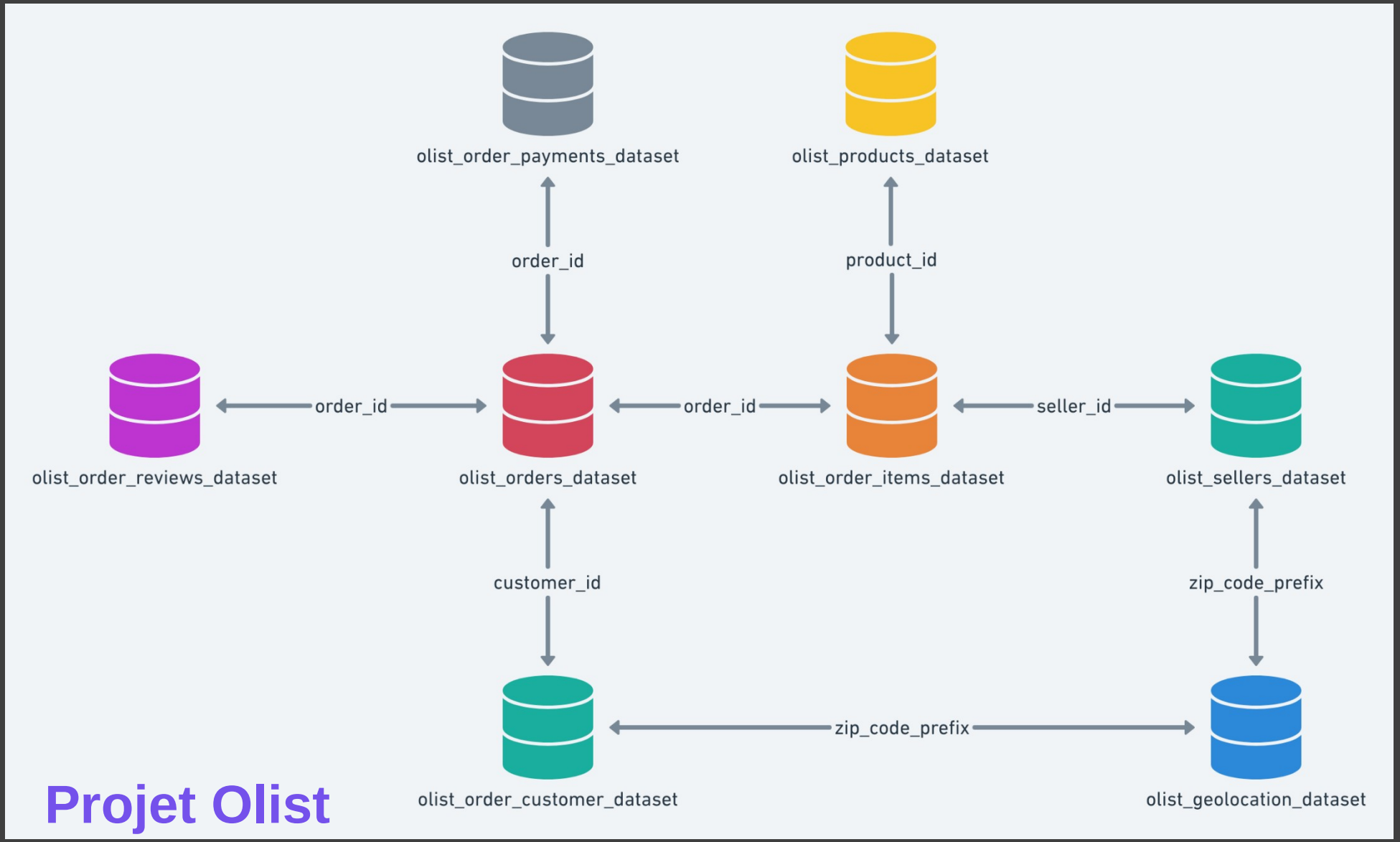
2. Analyse exploratoire des données

Source : données composées de 9 fichiers csv issus d'un SGBDR
(clients, commandes, paiements, produits, vendeurs, etc.)

But :
- fusionner ces données pour 1 entrée = 1 client avec ses caractéristiques
- explorer les données pour les caractériser pour une première compréhension
- nettoyer et identifier les éventuels problèmes à la fusion (outliers, imputation)

Opérations :
- fusion progressive des données avec exploration conjointe
- transformation de variables (extraction, sélection, transformation, création)

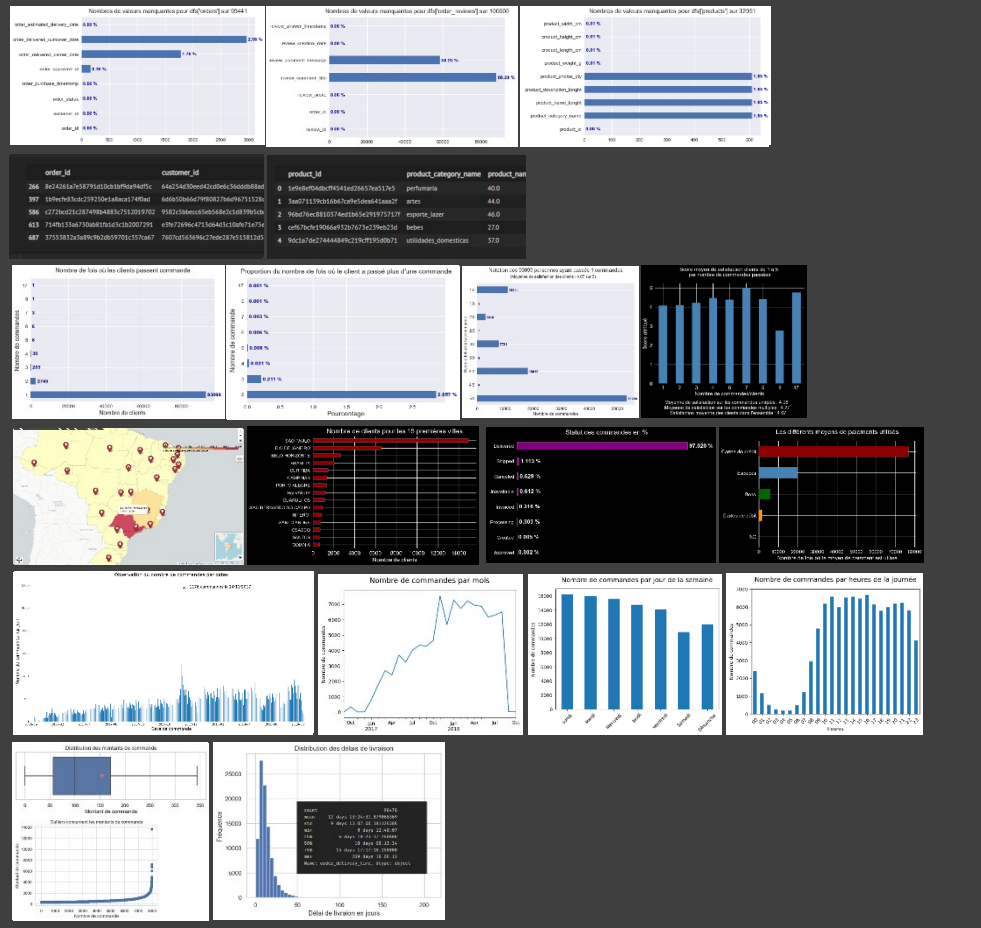
2. Analyse exploratoire des données



Projet Olist

2. Analyse exploratoire des données

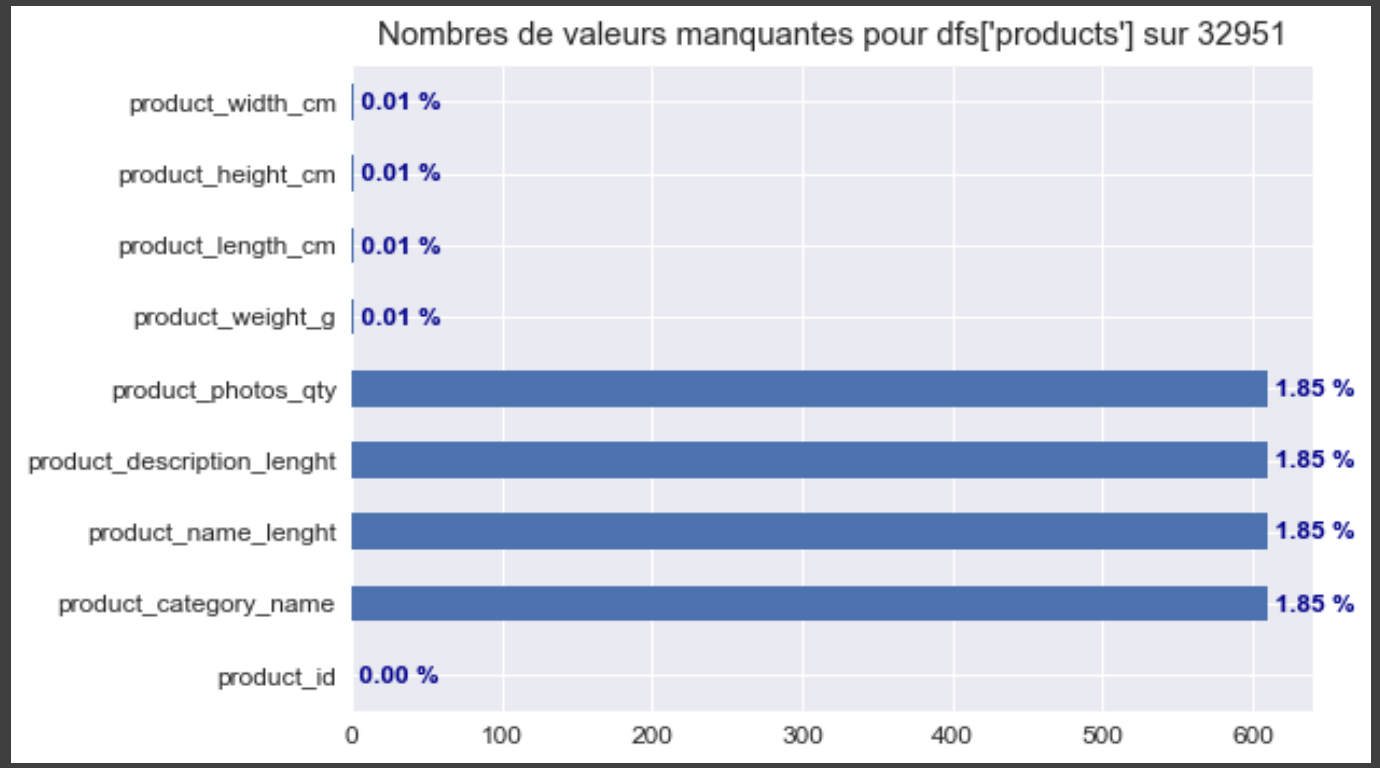
1. Taux de données manquantes (NaN) faible : 3 tables 9
2. Identification & compréhension des relations entre tables
3. Données clients et produits anonymisées (ID uniques)
4. Caractérisation du nombre de commandes passées
5. Niveau de satisfaction des clients / nombre de commandes
6. Répartition géographique des clients (Brésil, villes, états)
7. Moyens de paiement utilisés par les clients
8. Statut des commandes à l'instant t.
9. Vue globale de commandes par dates, mois, heures
10. Montant de commande moyen des clients + (outliers)
11. Délais de livraisons des commandes

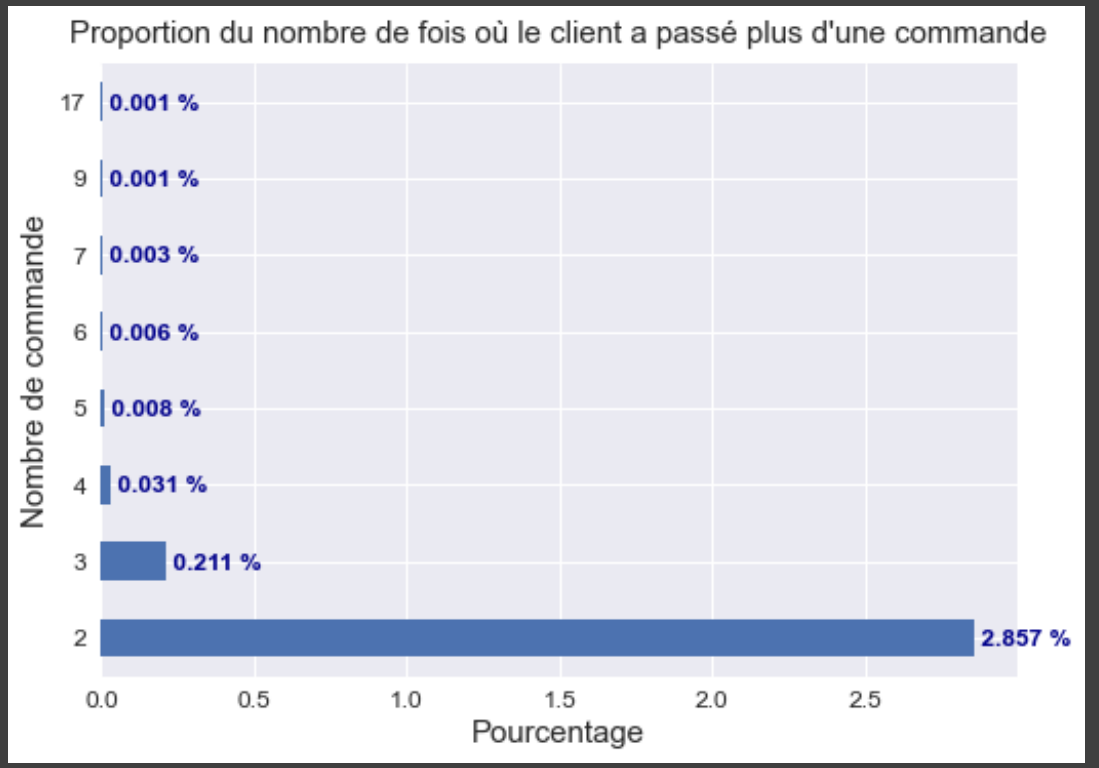


Projet Olist

Segmentation des clients d'un site e-commerce

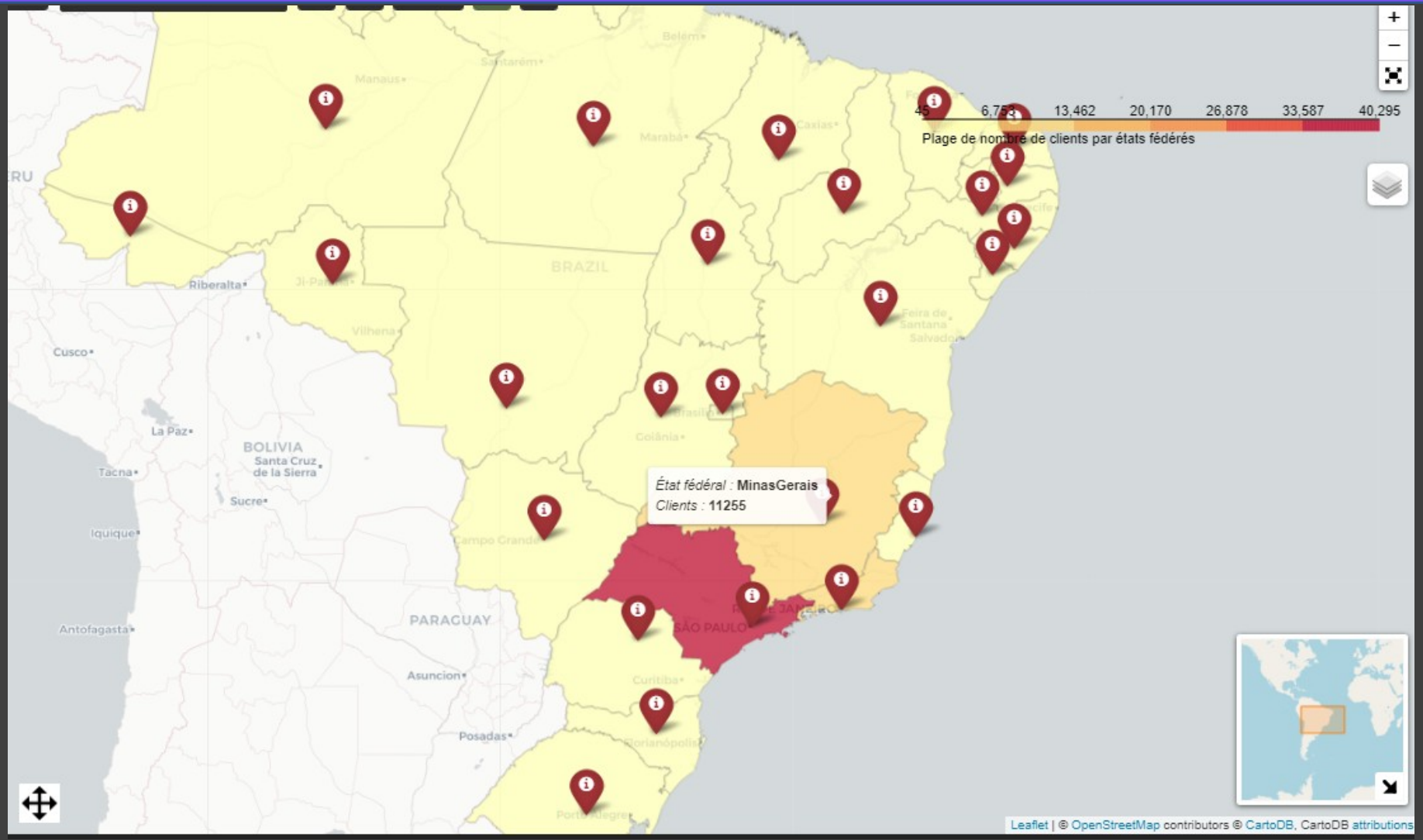
2. Analyse exploratoire des données



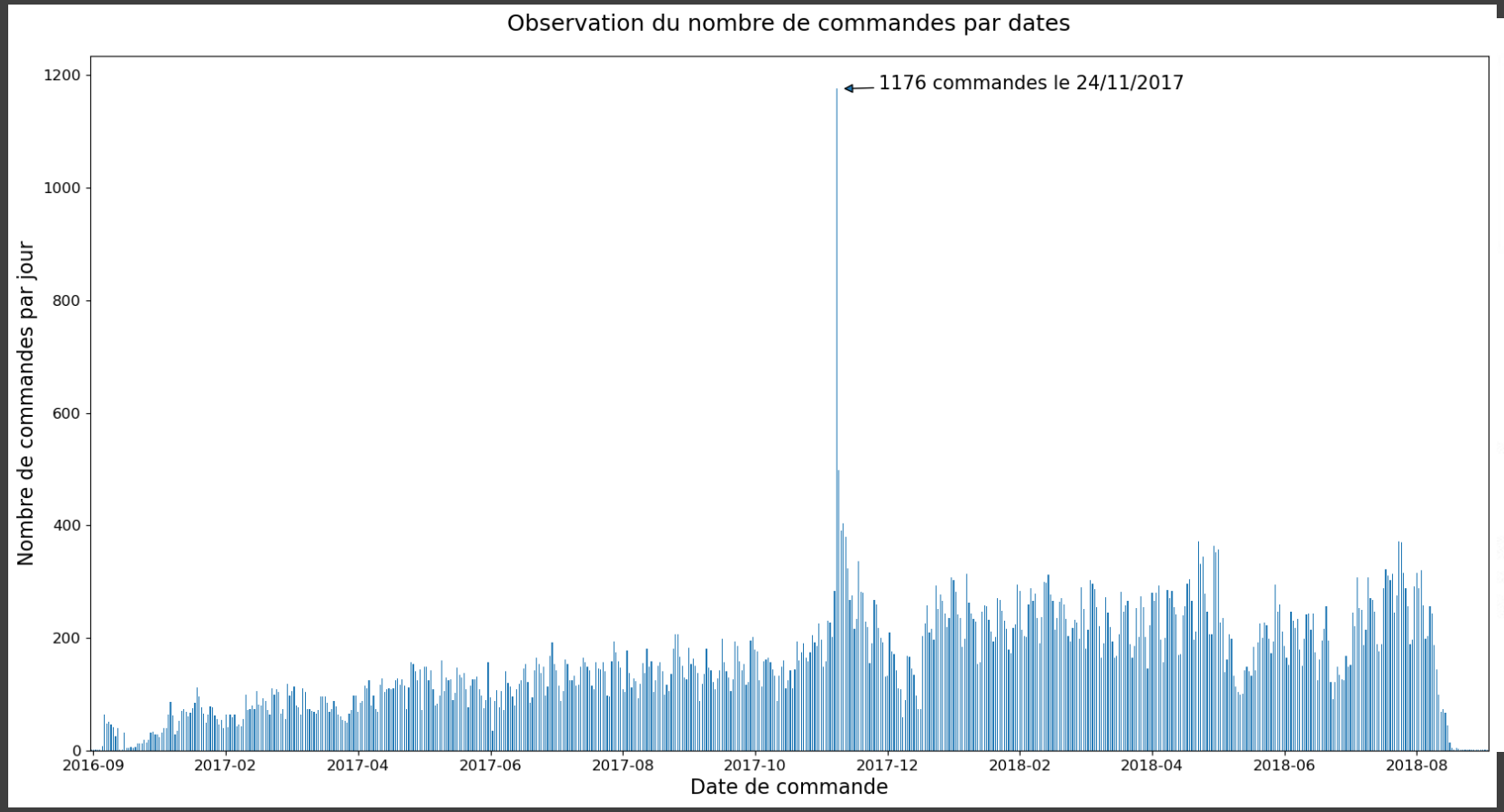


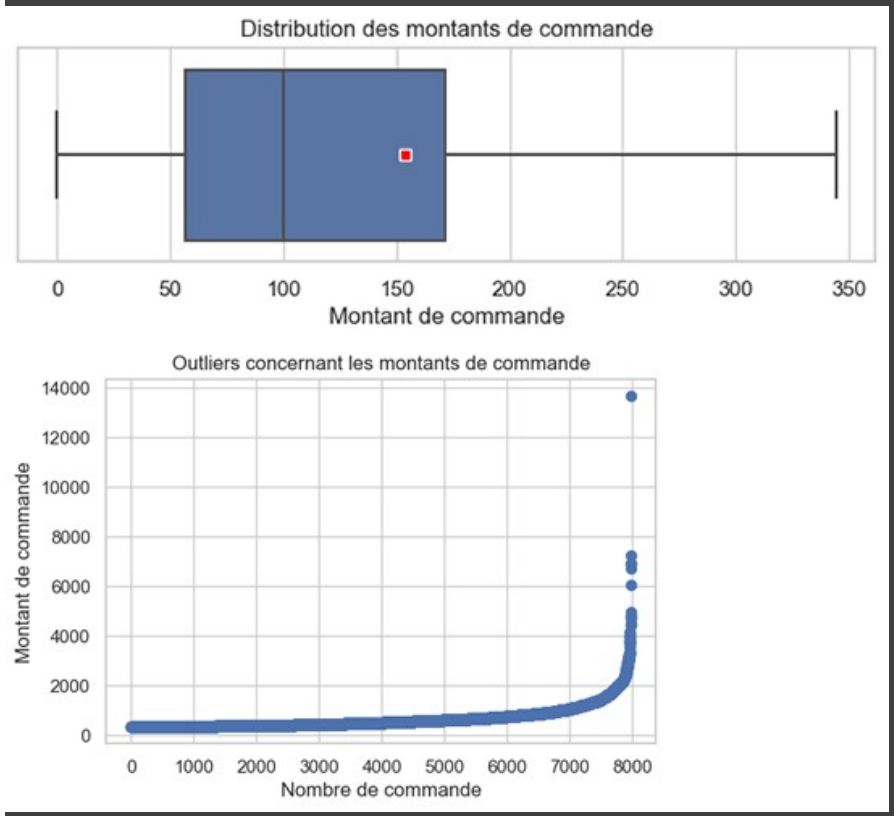
	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at
266	8e24261a7e58791d10cb1bf9da94df5c	64a254d30eed42cd0e6c36dddb88adf0	unavailable	2017-11-16 15:09:28	2017-11-16 15:26:5
397	1b9ecfe83cdc259250e1a8aca174f0ad	6d6b50b66d79f80827b6d96751528d30	canceled	2018-08-04 14:29:27	2018-08-07 04:10:2
586	c272bcd21c287498b4883c7512019702	9582c5bbecc65eb568e2c1d839b5cba1	unavailable	2018-01-31 11:31:37	2018-01-31 14:23:5
613	714fb133a6730ab81fa1d3c1b2007291	e3fe72696c4713d64d3c10afe71e75ed	canceled	2018-01-26 21:34:08	2018-01-26 21:58:3
687	37553832a3a89c9b2db59701c357ca67	7607cd563696c27ede287e515812d528	unavailable	2017-08-14 17:38:02	2017-08-17 00:15:1

2. Analyse exploratoire des données



2. Analyse exploratoire des données





Projet Olist

Segmentation des clients d'un site e-commerce

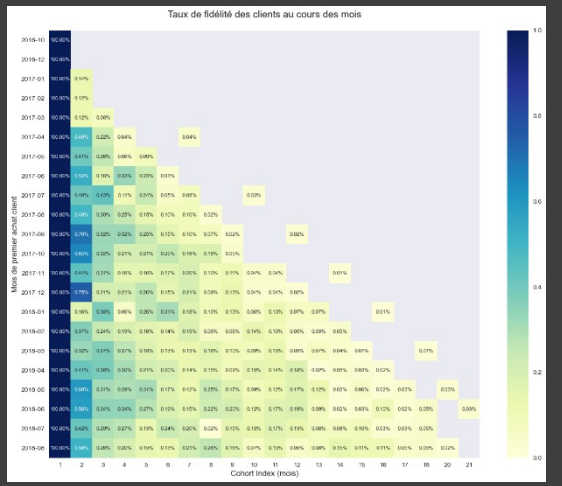
2. Analyse exploratoire complémentaire des données

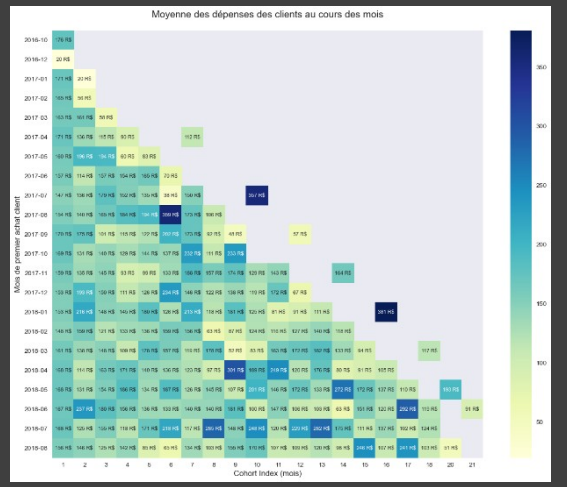
Analyse préliminaire et complémentaire à la segmentation pour observer le jeu de donnée dans sa globalité (approche commerciale) :

Cohortes clients (Fidélité & Moyenne de commande/Mois)

Utilisation de seulement quelques variables spécifiques:

- Client unique par entrée
- Nombre de commande(s)
- Dépense totale en achats
- Date d’achat(s)
- Statut de commande(s) -> doit être effectif (livré/validé)

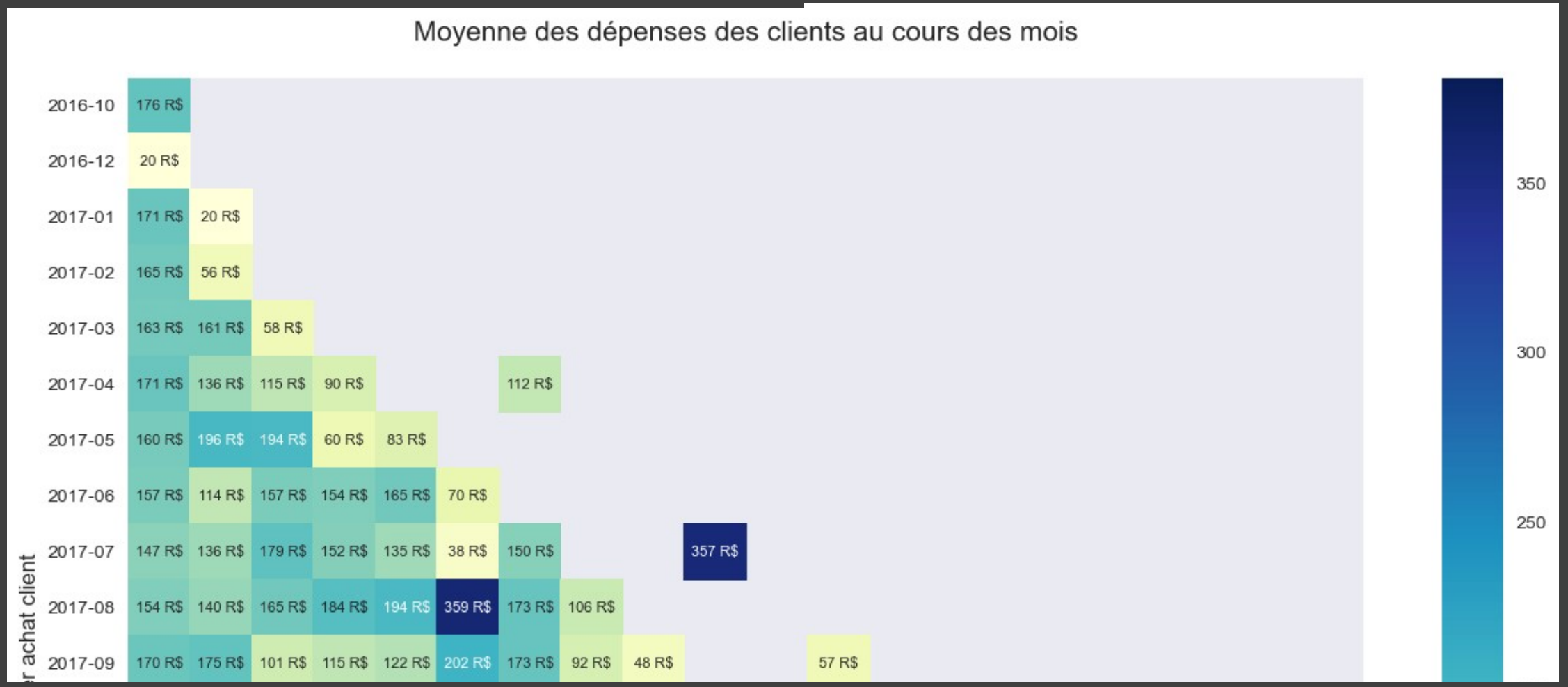




2. Analyse exploratoire complémentaire des données

2. Analyse exploratoire complémentaire des données

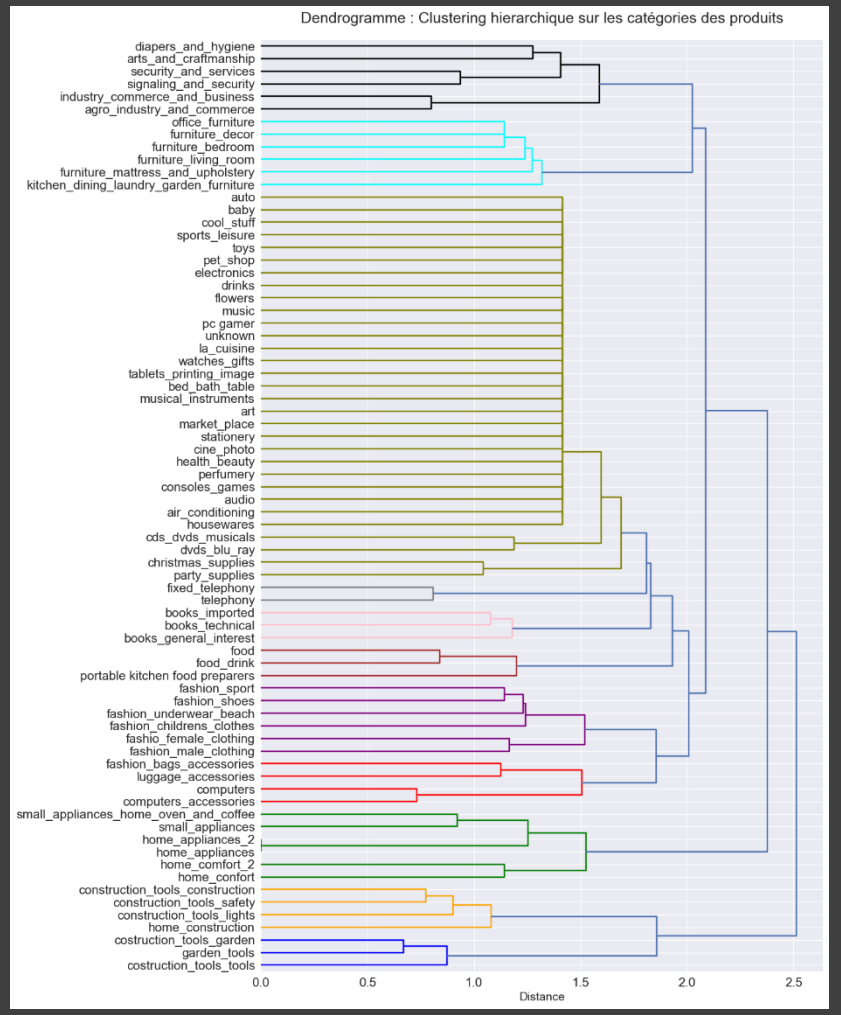
2. Analyse exploratoire complémentaire des données



2. Analyse exploratoire complémentaire des données

2. Transformation des données

1. Opération de fusions successives pour obtenir une entrée par client résumant son activité complète
2. Exemple : création de catégories produits spécifiques pour regrouper les domaines d'achats clients. Réduction de 74 à 16 catégories plus compréhensibles.
3. Utilisation de nombreux « groupby » avec des fonctions d'agrégation dans le but de résumé l'activité d'un client en ligne.
4. Création de variables (but réduire le nombre total de variables et éviter les fortes corrélations) : délai entre commandes, proportion des frais de port dans la commande, densité produit, dépense moyenne totale client
5. Imputation, correction de données, e.g. données de géolocalisation, données manquantes (moyenne ou suppression, etc.)



2. EDA : Corrélations linéaires ?

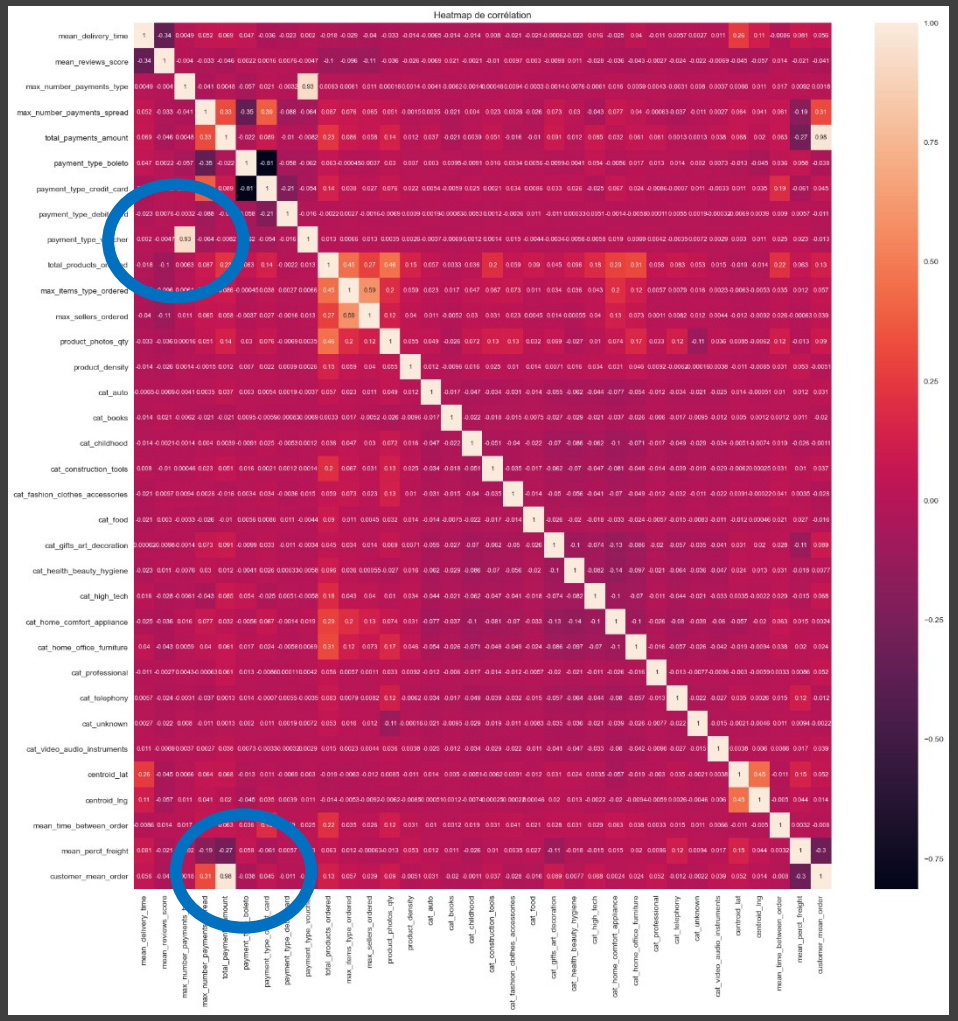
Corrélations linéaires particulières :

Rien de surprenant, ni utilisable pour de la transformation ou réduction de variables supplémentaires.

- payment_type_voucher & max_number_payment_type
- customer_mean_order & total_payments_amount

DataFrame final créé pour la segmentation :

- 95420 entrées clients
- 43 variables
- Pas de données manquantes sauf « reviews » clients



3. Méthodes employées : Clusterisation, NMF, RFM

- Utilisation d’un modèle de clustering (K-Means, DBSCAN)
- Compréhension de certains types de variables: RFM, NMF
- Recoupement des informations (Clustering/RFM): Score
- Comparaison: méthodes RFM / clustering totales/partielles
- Interprétations possibles (axes: financier, produits, clients)
- Fréquence du contrat de maintenance / Objectifs OLIST

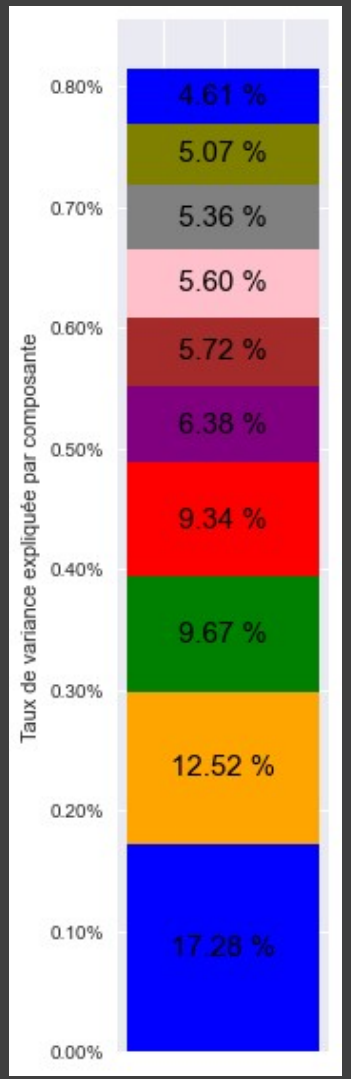
3. Méthodes employées : Clusterisation, NMF, RFM

ACP aide à la visualisation, réduction de dimensions ?

- 18 variables numériques de départ
- 14 variables expliquent 95% de variance totale…
- En conservant 80% de variance on peut réduire à 10 variables

Deux premières composantes ACP cumulent moins de 30 % variance expliquée

- Aide à la visualisation pour l'interprétation ? Non, pas significatif
- Réduction utile pour la clusterisation ? Non, pas utile, e.g. : perte de 20%



3. Méthodes employées : Clusterisation, NMF, RFM

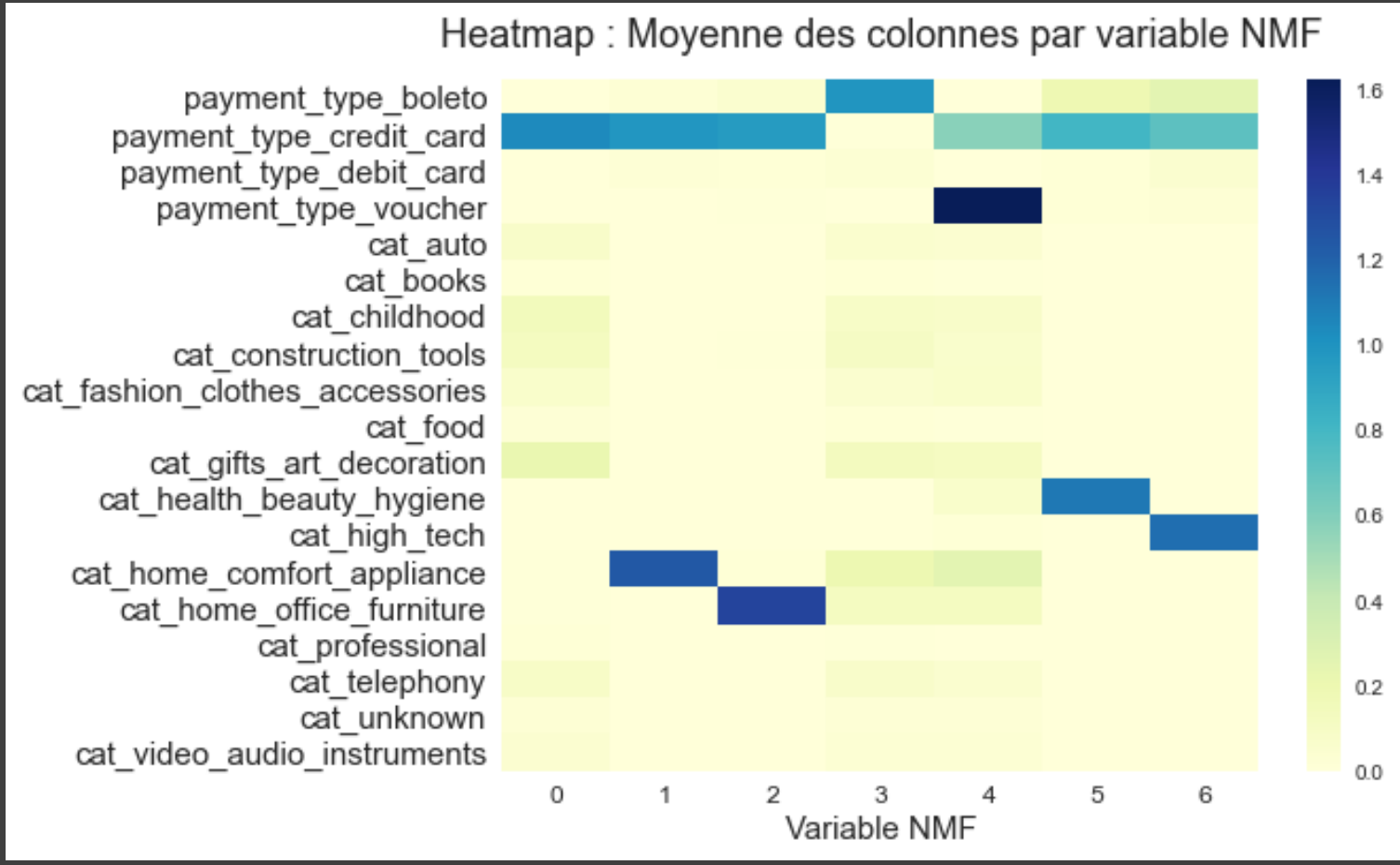
Caractérisation des habitudes de consommation clients en fonction des catégories des produits/paiements.

Intéressant : confirme l'EDA

Informations complémentaires:

- Catégories : home
- Catégories : health_beauty_hygiene
- Paiements : boleto & voucher

Majoritairement : « credit card » pour l'ensemble des features NMF



3. Méthodes employées : Clusterisation, NMF, RFM

4 types de clustering avec K-Means effectués en faisant varier les données:

- 1er type de clustering : variables numériques (7 clusters)
- 2ème type de clustering : variables numériques (12 clusters)
- 3ème type de clustering : avec variables RFM (sélection)
- 4ème type de clustering : variables RFM + variables catégorielles (meilleur compromis: résultats/dimensions)

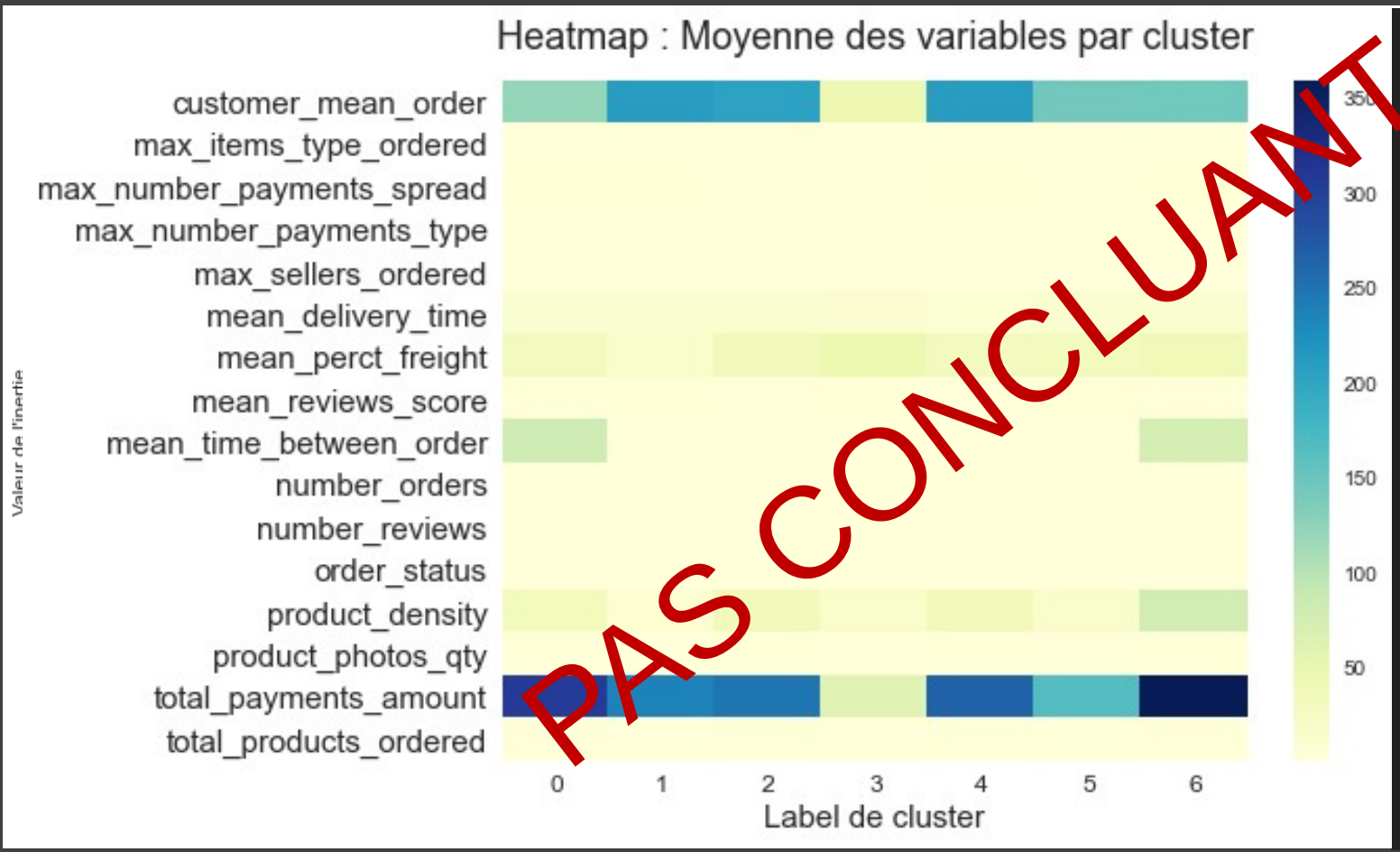
14 clusters qui sont interprétables
(meilleur que les résultats des 12 clusters obtenus)

- Meilleures métriques (SSE et coef. Silhouette et Index D.B.)
- Permet une caractérisation fine par groupe de clients
- Limitation : nombre important pour une lecture globale clients

Un nombre important de clusters :
intéressant pour la caractérisation par groupe produits.

3. Méthodes employées : Clusterisation, NMF, RFM

1ère approche via K-Means sur les données numériques exclusivement.




```
silhouette_avg : 0.21865898721018584
Davies-Bouldin index :  1.078879130607698
KMeans labels :  [1 3 2 5 4 0 6]
Taille du cluster numéro 0 : 2759 clients

Taille du cluster numéro 1 : 42149 clients

Taille du cluster numéro 2 : 7524 clients

Taille du cluster numéro 3 : 38950 clients

Taille du cluster numéro 4 : 1172 clients

Taille du cluster numéro 5 : 2747 clients

Taille du cluster numéro 6 : 119 clients
```

3. Méthodes employées : Clusterisation, NMF, RFM

Approche RFM pour caractériser plus facilement les clients :

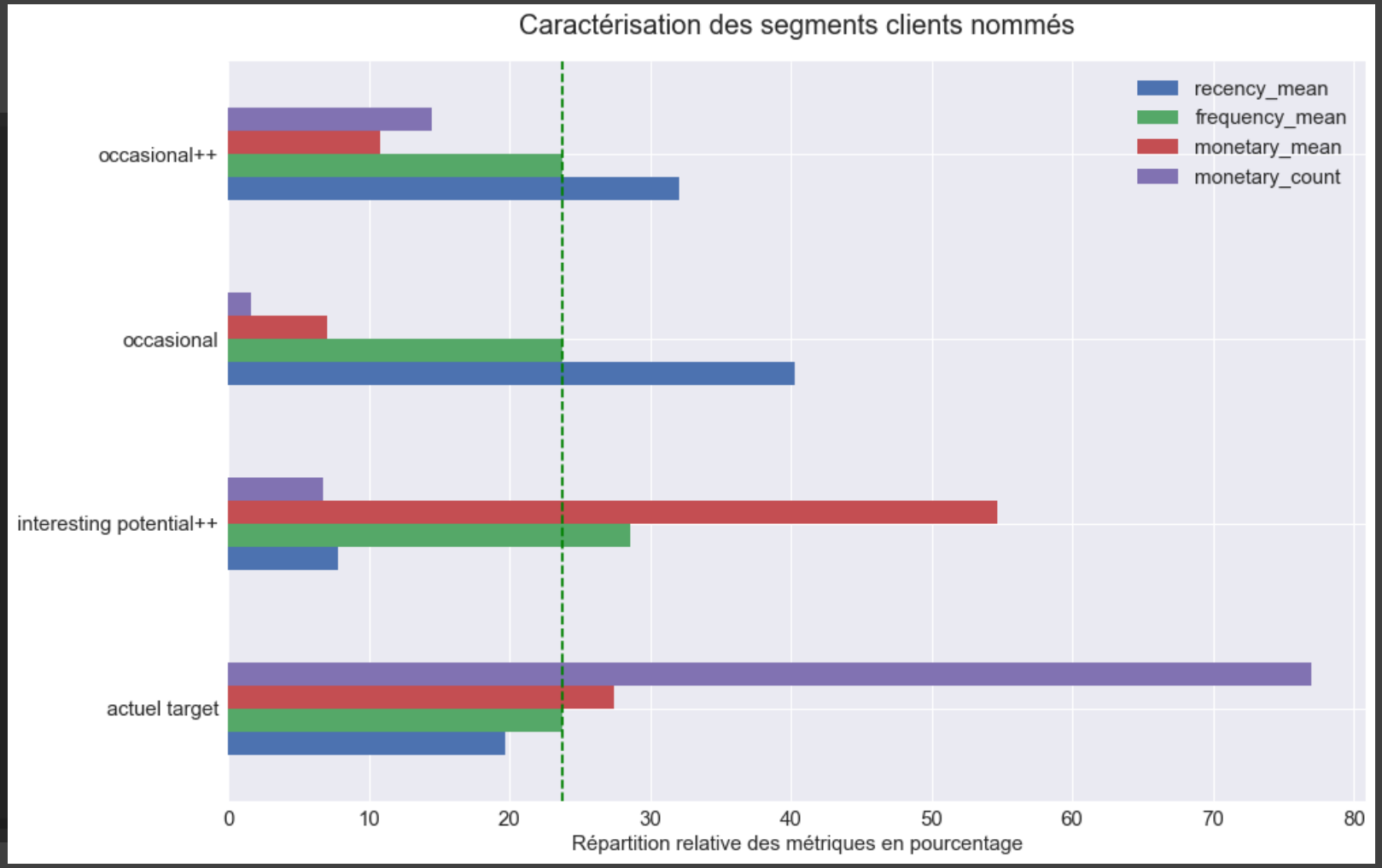
	recency	frequency		monetary	
	mean	mean	max	mean	count
rfm_score					
3	455.1	1.0	1	43.9	1519
4	395.2	1.0	1	56.9	4578
5	346.9	1.0	1	72.5	8970
6	300.9	1.0	3	109.3	14486
7	252.7	1.0	4	138.0	17457
8	212.0	1.0	6	170.4	16929
9	175.6	1.0	4	219.9	14173
10	134.1	1.1	5	257.0	8872
11	100.1	1.2	9	314.6	4592
12	58.6	1.4	15	408.2	1781

Système de notation clients exploitable :

	recency	frequency	monetary	
	mean	mean	mean	count
segment_label				
actuel target	223.0	1.0	170.7	71917
interesting potential++	88.5	1.2	340.8	6373
occasional	455.1	1.0	43.9	1519
occasional++	363.2	1.0	67.3	13548

3. Méthodes employées : Clusterisation, NMF, RFM

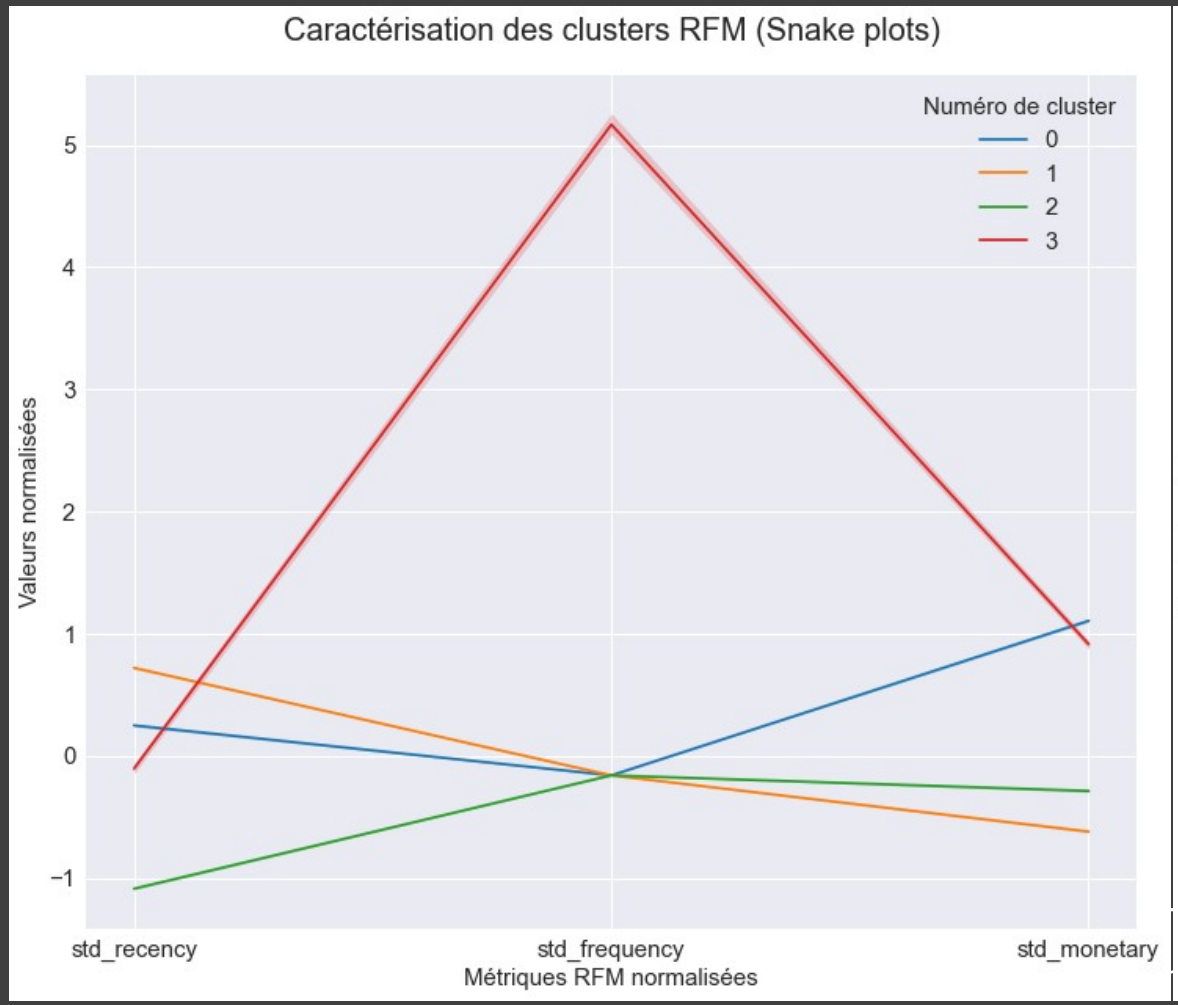
	recency	frequency		monetary	
	mean	mean	max	mean	count
rfm_score					
3	455.1	1.0	1	43.9	1519
4	395.2	1.0	1	56.9	4578
5	346.9	1.0	1	72.5	8970
6	300.9	1.0	3	109.3	14486
7	252.7	1.0	4	138.0	17457
8	212.0	1.0	6	170.4	16929
9	175.6	1.0	4	219.9	14173
10	134.1	1.1	5	257.0	8872
11	100.1	1.2	9	314.6	4592
12	58.6	1.4	15	408.2	1781

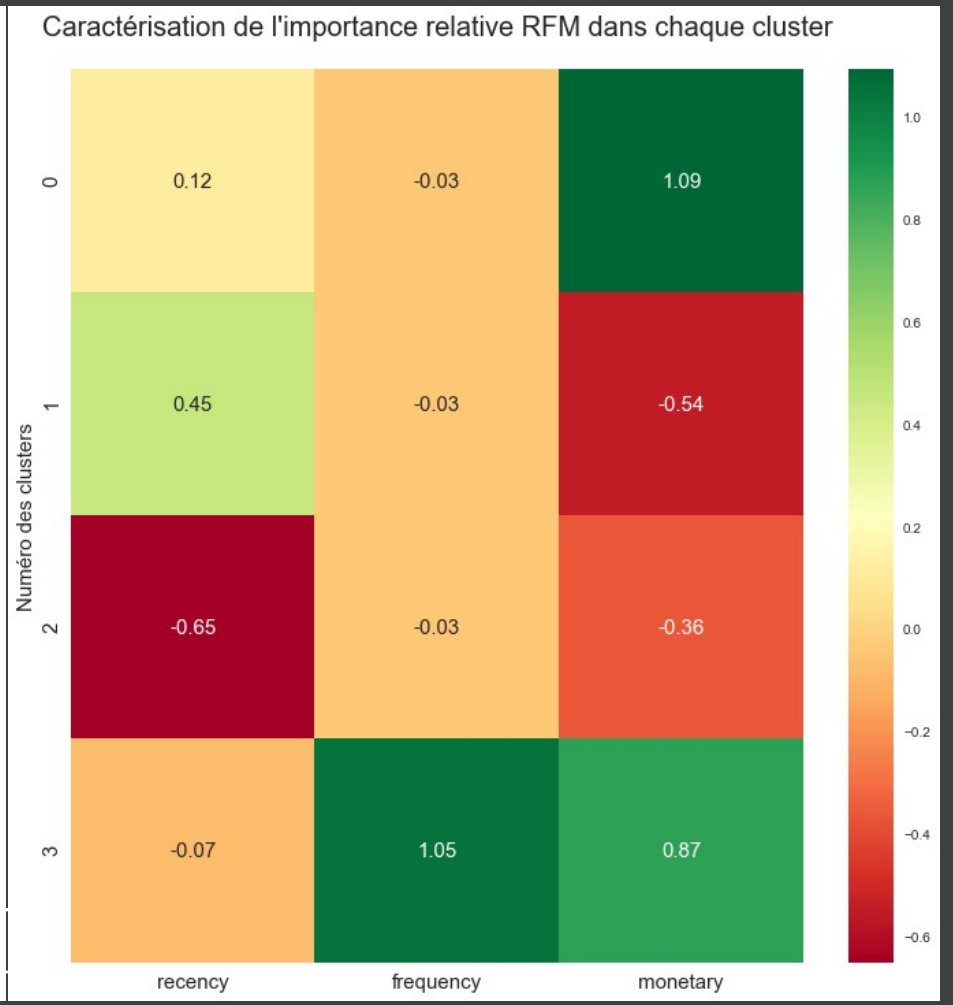


3. Méthodes employées : Clusterisation, NMF, RFM

Opération : Clustering sur les variables RFM normalisées

But : vérifier si le système de notation clients est cohérent du point de vue d'un algorithme non supervisé





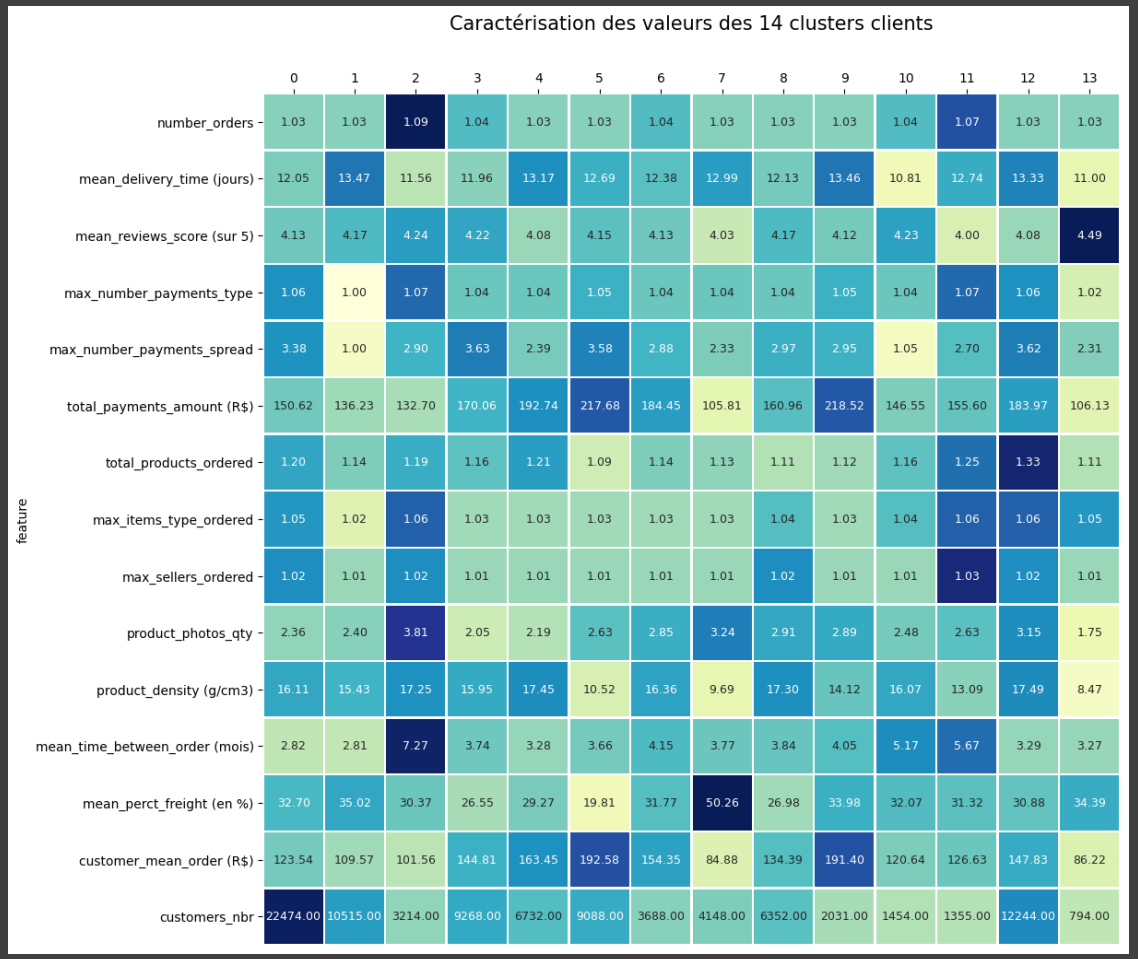
3. Méthodes employées : Clusterisation, NMF, RFM

14 clusters (utiles pour une caractérisation précise) :

- Majorité des clients des 14 clusters proviennent de : Sao Paulo (SP), Segment RFM « actual target »

- Catégories spécifiques quasiment à chaque clusters: intéressant commercialement -> segmentation par catégorie de produits

- La Majorité des clients payent avec la « credit card », seul le cluster 2 -> paiement en espèces

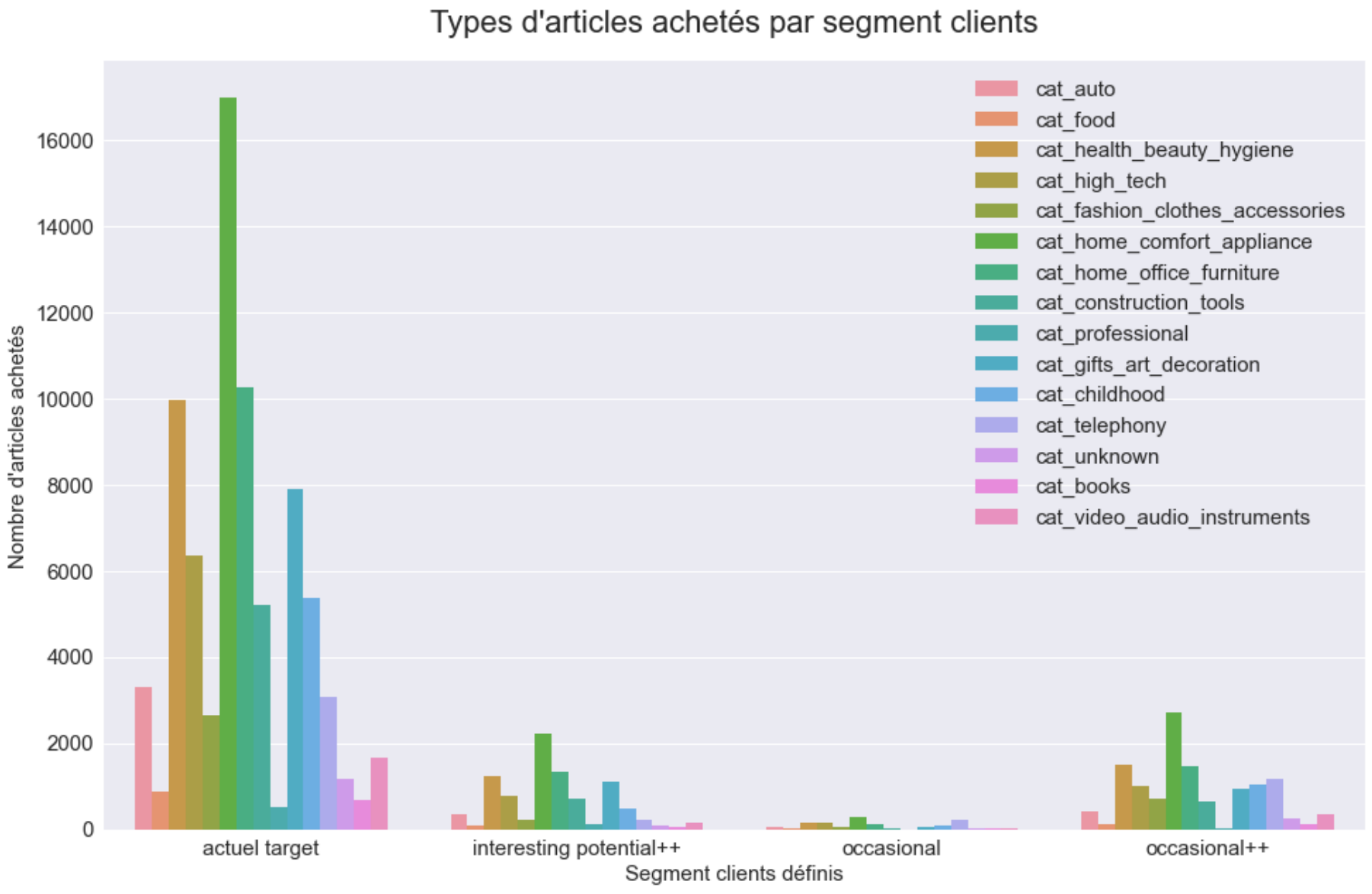


3. Méthodes employées : Clusterisation, NMF, RFM

Cluster	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Catégorie Produits :	home_ confort_ appliance	home_ confort_ appliance + home_ office_ furniture	fashion_ clothes_ access.	healt_ beaut_ hygiene	high_ tech	gifts_ art_ deco.	auto	telephony	chilhood	video_ audio_ intrum.	home_ confort_ appliance + health_ beauty_ hygiene	unknown	home_ office_ furniture + construc._ tools	books

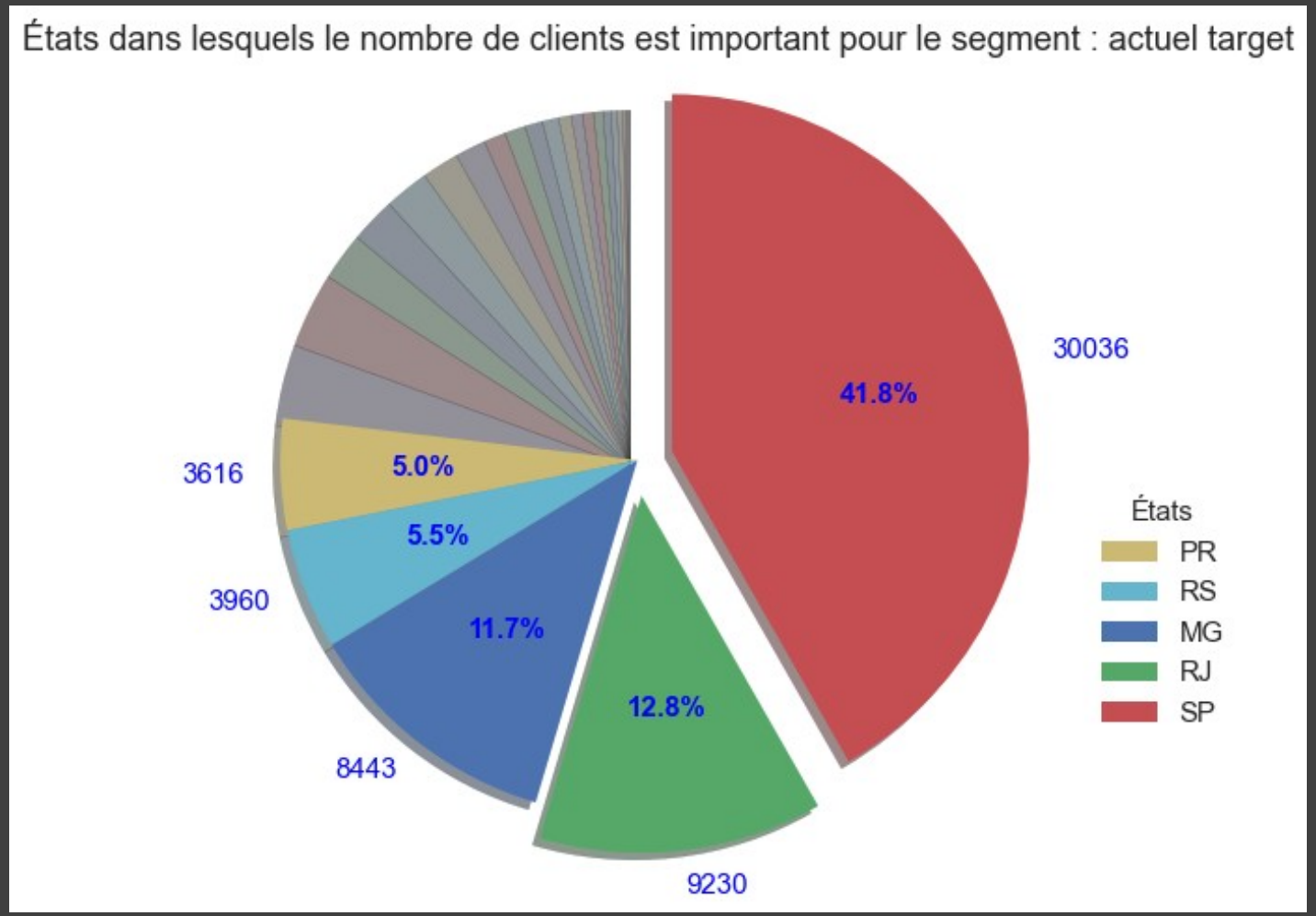
Cluster	0	1	2	3	4	5	6	7	8	9	10	11	12	13
max_number_payments_spread	3.38	1.00	2.90	3.63	2.39	3.58	2.88	2.33	2.97	2.95	1.05	2.70	3.62	2.31
total_payments_amount (R$)	150.62	136.23	132.70	170.06	192.74	217.68	184.45	105.81	160.96	218.52	146.55	155.60	183.97	106.13
mean_time_between_order (mois)	2.82	2.81	7.27	3.74	3.28	3.66	4.15	3.77	3.84	4.05	5.17	5.67	3.29	3.27
mean_perct_freight (en %)	32.70	35.02	30.37	26.55	29.27	19.81	31.77	50.26	26.98	33.98	32.07	31.32	30.88	34.39
mean_reviews_score (sur 5)	4.13	4.17	4.24	4.22	4.08	4.15	4.13	4.03	4.17	4.12	4.23	4.00	4.08	4.49
total_products_ordered	1.20	1.14	1.19	1.16	1.21	1.09	1.14	1.13	1.11	1.12	1.16	1.25	1.33	1.11
number_orders	1.03	1.03	1.09	1.04	1.03	1.03	1.04	1.03	1.03	1.03	1.04	1.07	1.03	1.03

4. Interprétation et exploitation possible des résultats



Note : confirme les tendances observées via la NMF sur les catégories seules

4. Interprétation et exploitation possible des résultats



Unités fédérales (états + 1 district fédéral : DF, Brasilia)

PR : Paraná

RS : Rio Grande do Sul

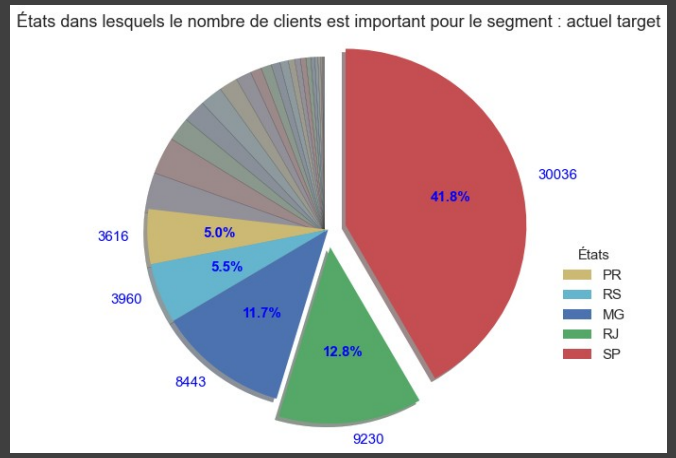
MG : Minas Gerais

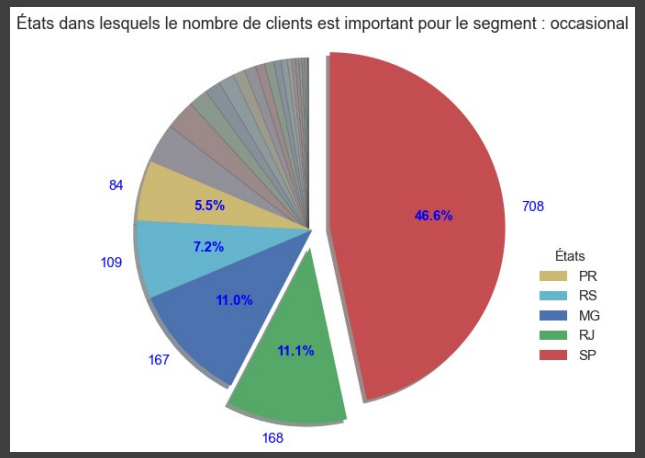
RJ : Rio de Janeiro

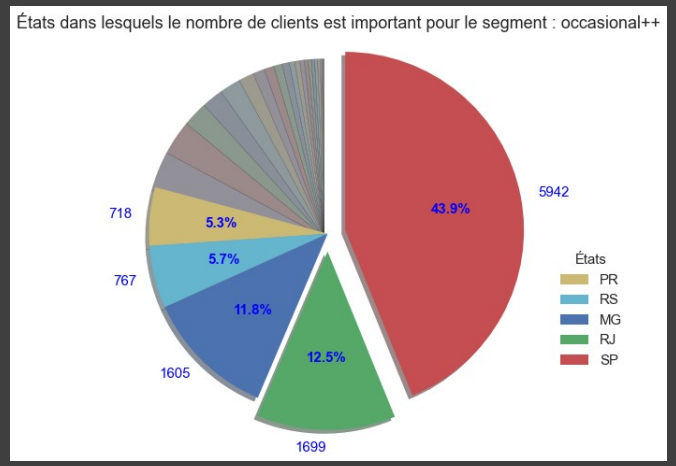
SP : São Paulo

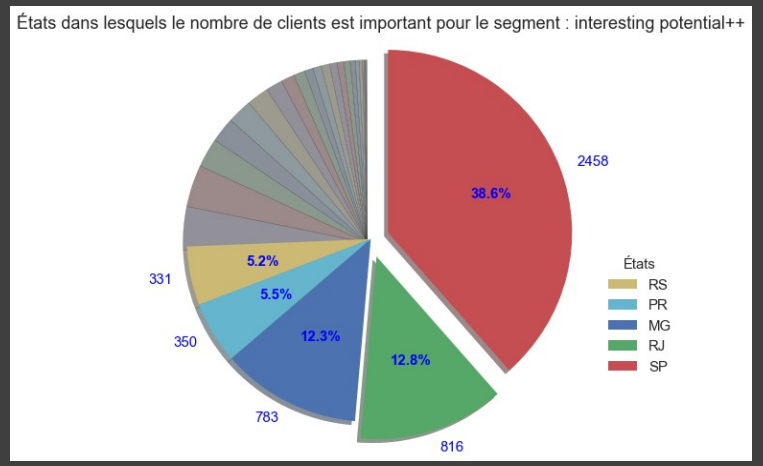
4. Interprétation et exploitation possible des résultats

Répartition géographique par états des clients par segment









Unités fédérales

(états + 1 district fédéral : DF, Brasilia)

PR : Paraná

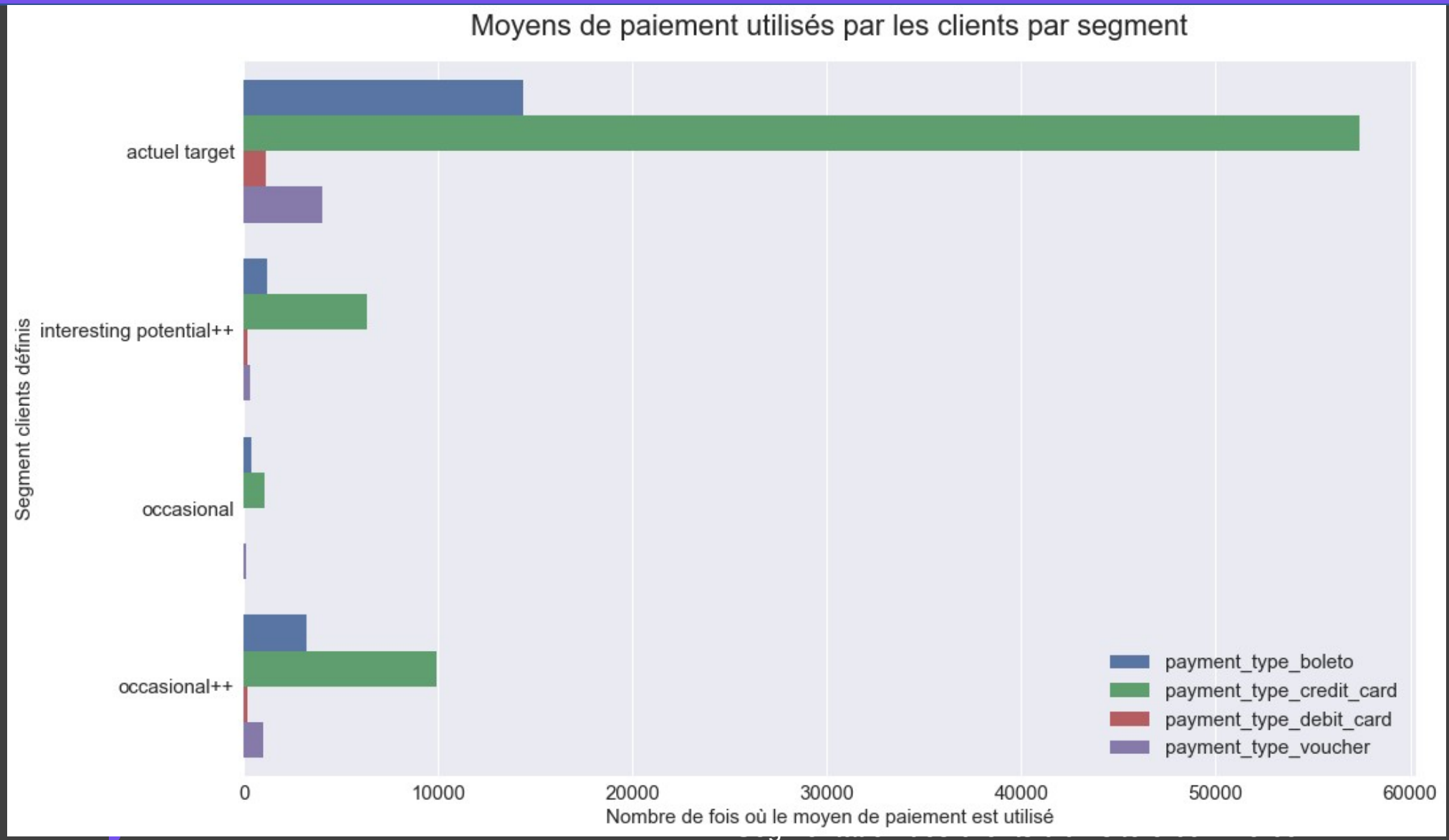
RS : Rio Grande do Sul

MG : Minas Gerais

RJ : Rio de Janeiro

SP : São Paulo

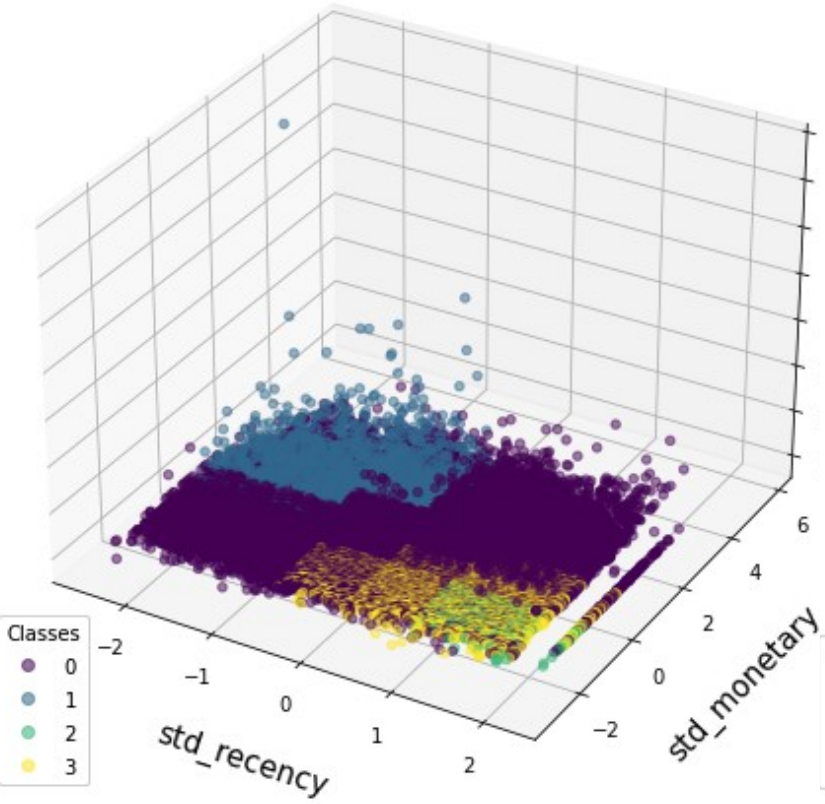
4. Interprétation et exploitation possible des résultats



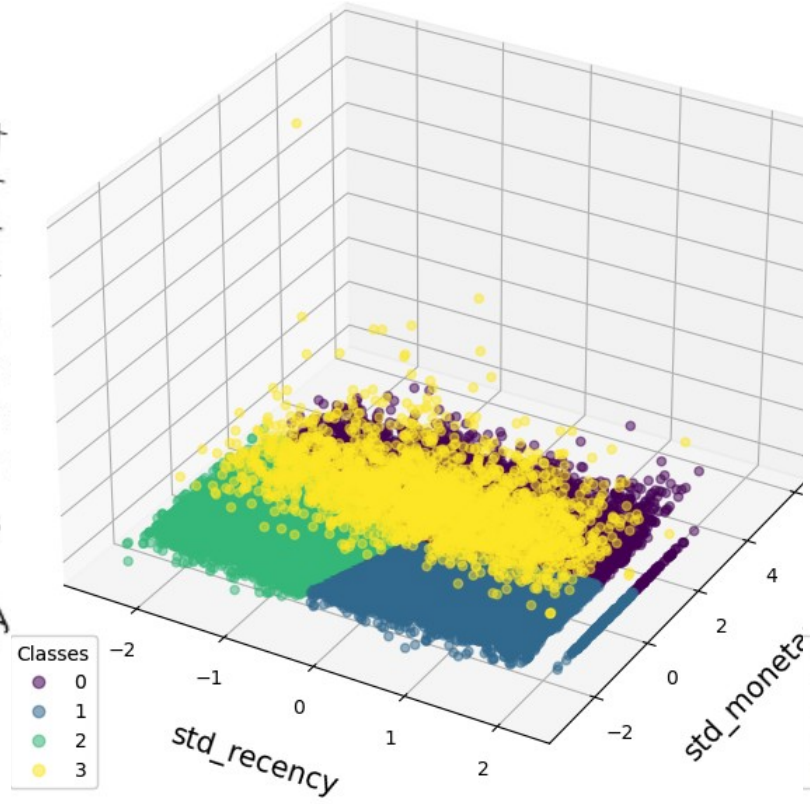
4. Interprétation et exploitation possible des résultats

Comparaison des résultats avec DBSCAN avec la même méthodologie que pour K-Means / RFM.

Visualisation des segments clients labellisés

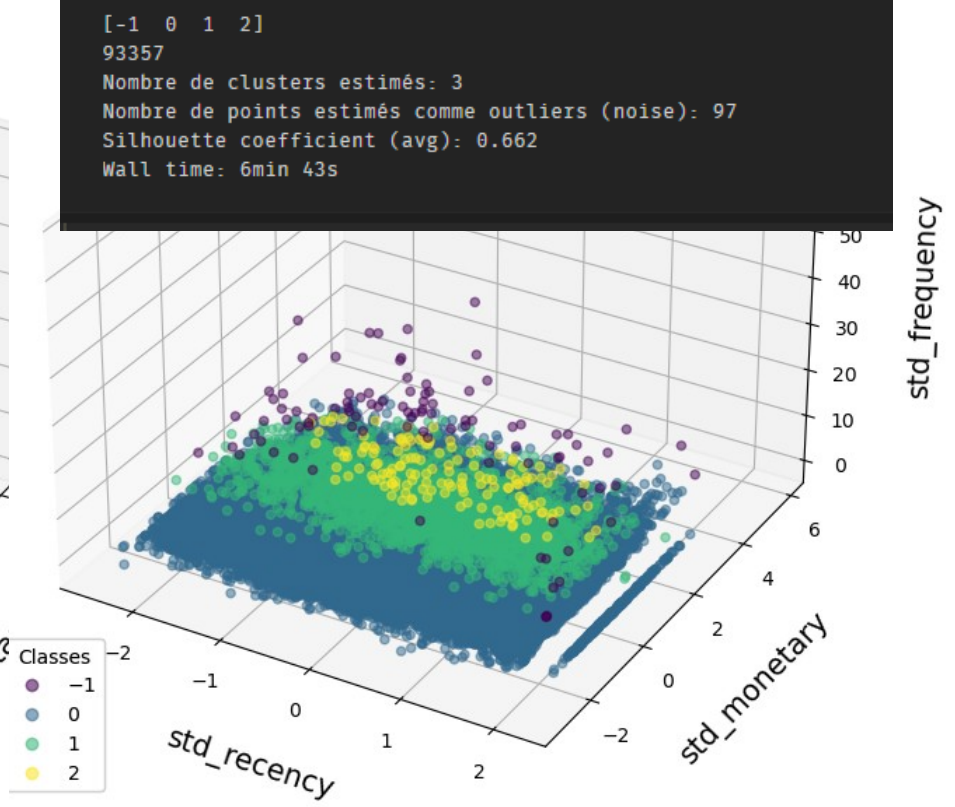


Visualisation des clusters KMeans/RFM



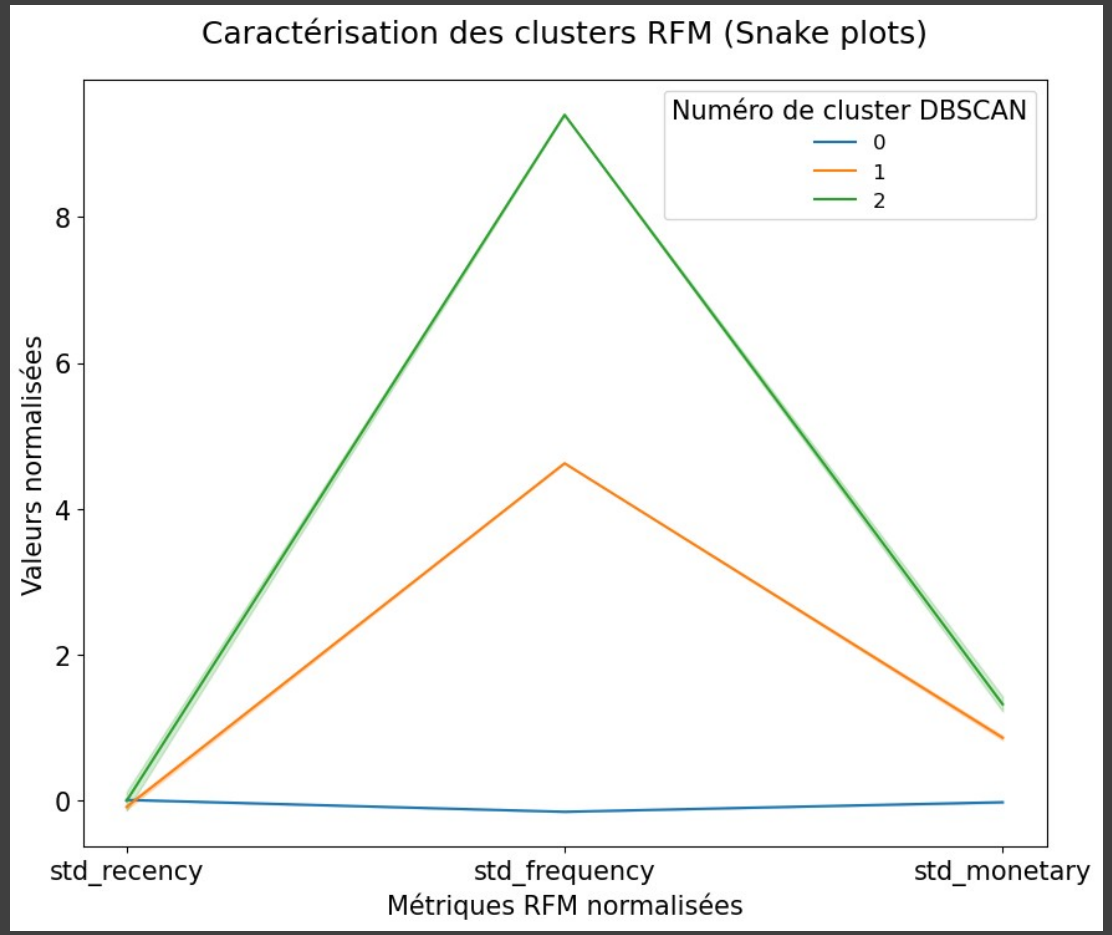
Visualisation des clusters DBSCAN/RFM

```
[-1  0  1  2]
93357
Nombre de clusters estimés: 3
Nombre de points estimés comme outliers (noise): 97
Silhouette coefficient (avg): 0.662
Wall time: 6min 43s
```



4. Interprétation et exploitation possible des résultats

Comparaison des résultats avec DBSCAN avec la même méthodologie que pour K-Means / RFM.



Regroupement en seulement 3 catégories de clients :

- La distinction s'effectue au niveau de la fréquence
- Et nettement au niveau monétaire

Ce qui pourrait laisser penser que les segments clients avec occasional et occasional++ ont été fusionnés. Mais en réalité les résultats sont différents du clustering K-Means

5. Proposition de maintenance

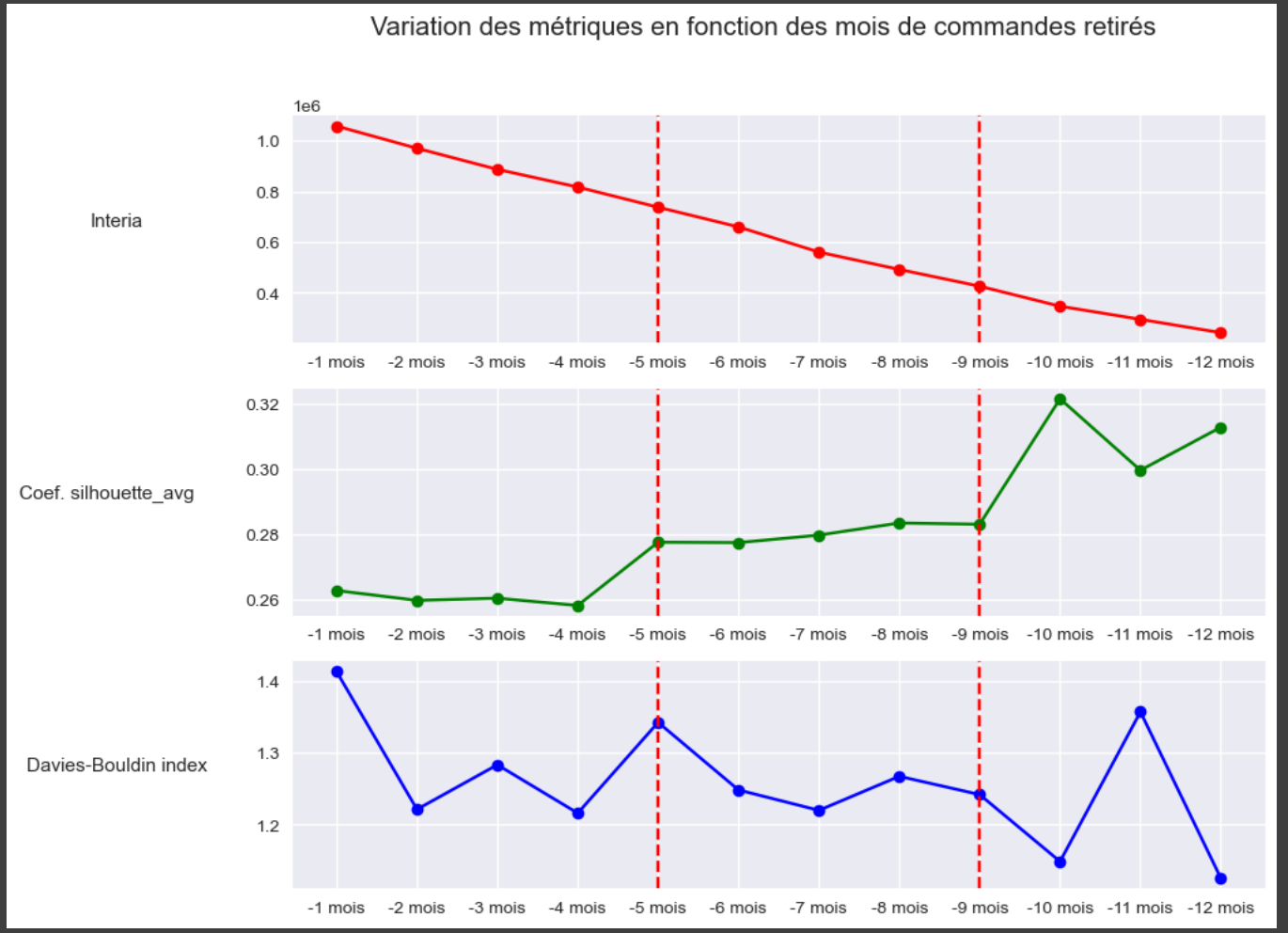
Reprise des résultats à 14 clusters

Détermination de la variation des métriques en fonction du nombre de mois de commandes impliqués

Observation des valeurs en partant de la dernière commande – 1 mois, -2 mois, etc.

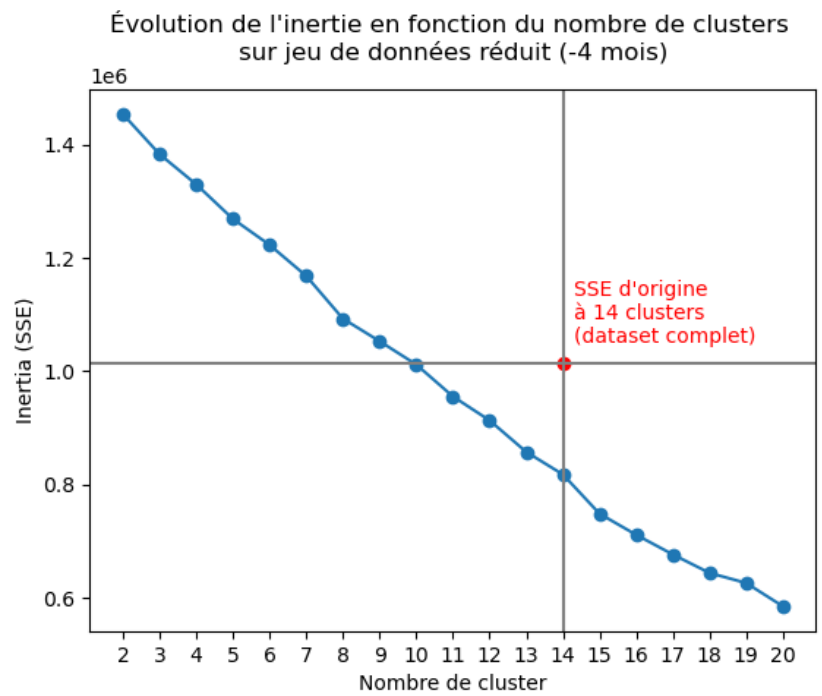
Coef. Silhouette moy. Variation significative. (plage de 4/5 mois max.)

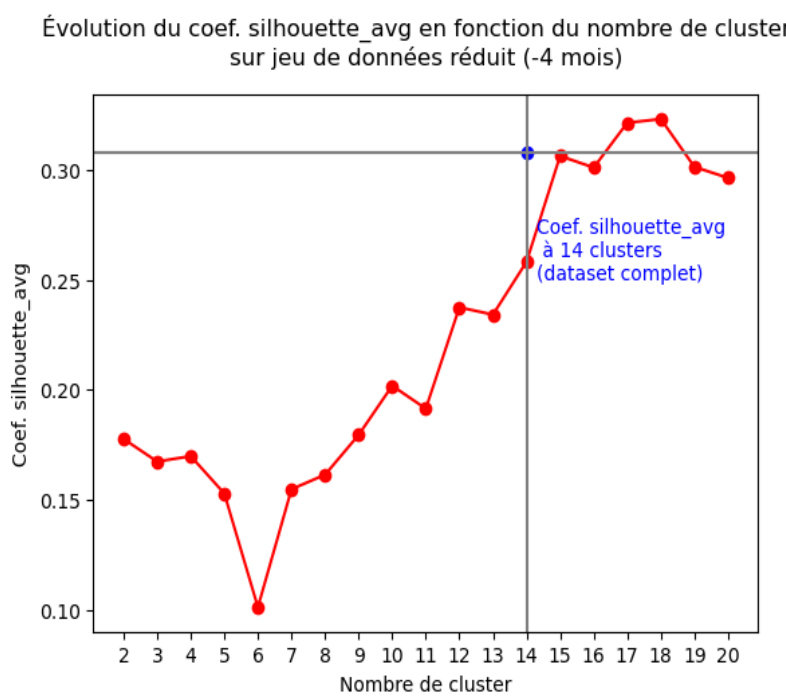
1ère Maintenance à envisager par exemple dans les 4 premiers mois et réévaluée avec une fréquence différente si le pool de clients varie rapidement.

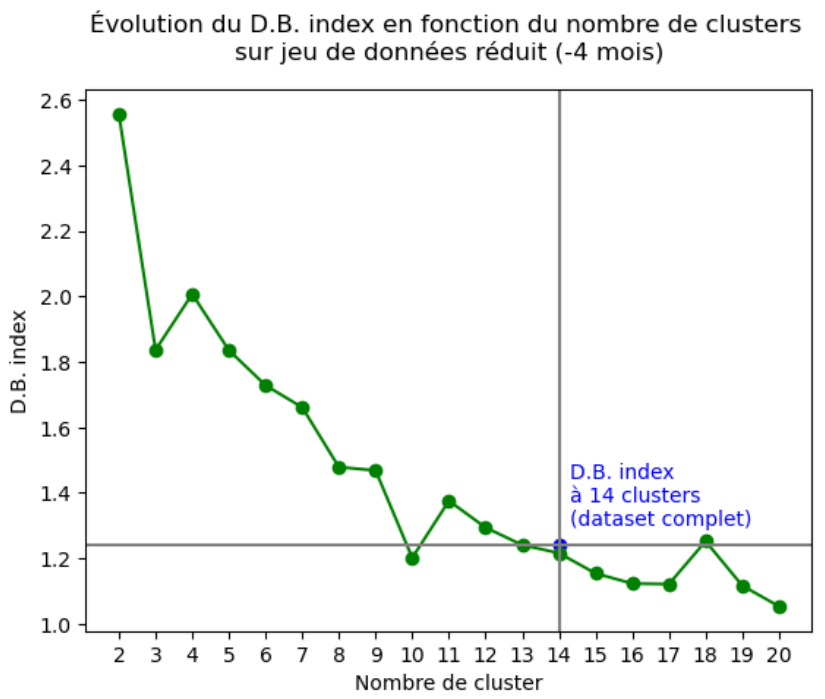


5. Proposition de maintenance

Comparaison avec les valeurs des métriques obtenues sur la clusterisation à 14.







5. Bilan de projet

1. Segmentation RFM : rapide & caractérisation intelligible
2. K-Means peut amener un complément de caractérisation précis.
3. Tout dépend des axes (commerciaux/marketing) -> Adaptable
4. Il serait intéressant de retester les modèles avec d'autres types de variables en fonction de l'évolution du « pool » clients.
5. Dataset très particulier (constaté via EDA : 1 commande = 1 client)
6. L'accès à d'autres informations clients (anonymes) pourrait être un apport très intéressant (âge, sexe, navigateur, terminal de commande, page(s), visité(s), Mesure de trafic, d'audience, etc.)
7. Caractérisation clients avec le nom des produits achetés -> ALS (recommandation -> en « vectorisant » les clients autrement)