

SEGMENTATION CLIENTS

olist

Formation data scientist
Projet 5 | Alexis Marceau | Juin 2022

Sommaire

1. Contexte et problématique
2. Analyse exploratoire
3. Segmentation
4. Délai de maintenance du modèle



CONTEXTE ET PROBLEMATIQUE

Olist, spécialiste E-Commerce Brésilien souhaite qu'on fournit à ses équipes d'e-commerce une **segmentation des clients** qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.



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

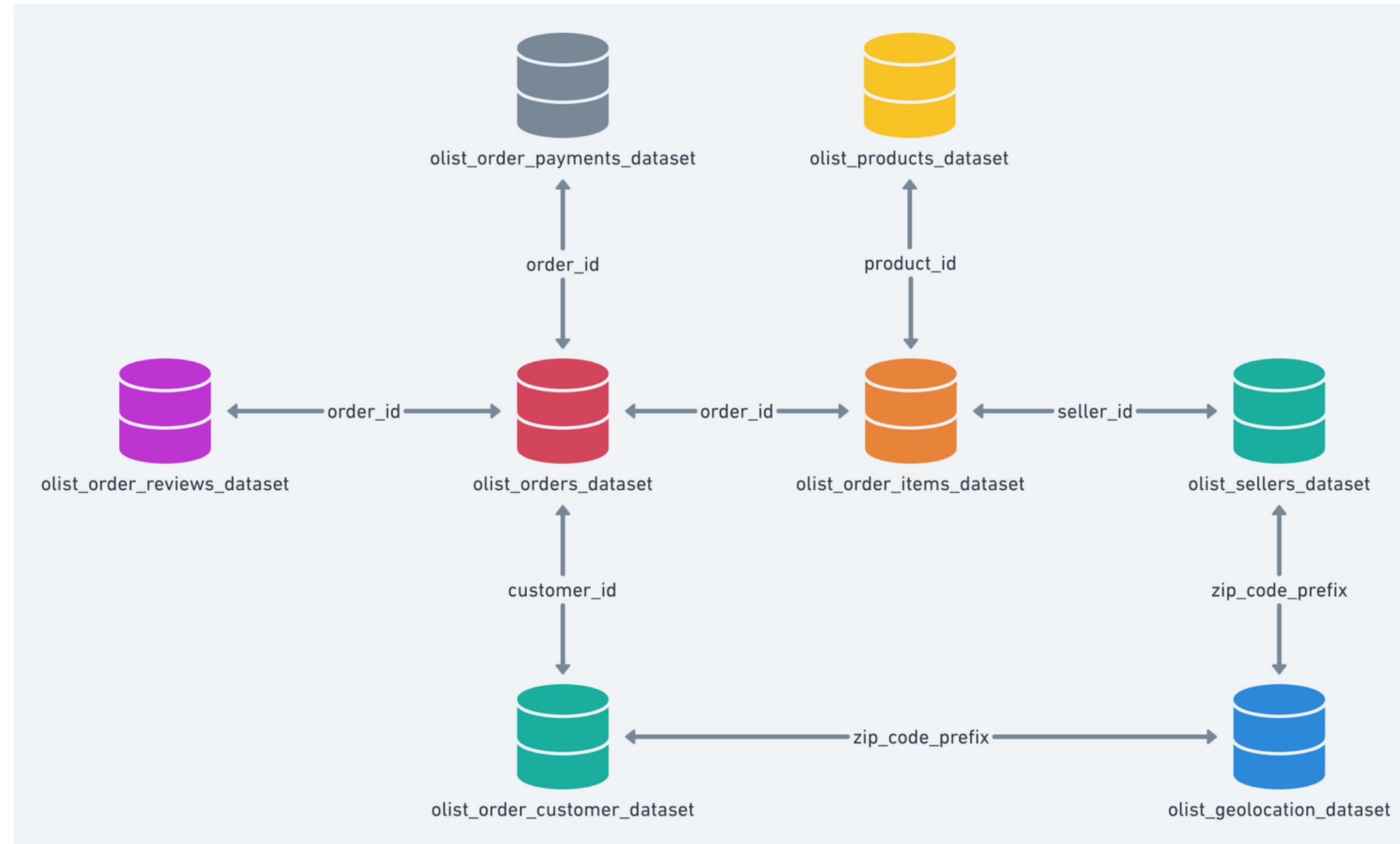
On doit fournir à l'équipe marketing une description actionnable de notre segmentation et de sa logique sous-jacente pour une utilisation optimale, ainsi qu'une proposition de **contrat de maintenance** basée sur une analyse de la stabilité des segments au cours du temps.



ANALYSE EXPLORATOIRE

Description, cleaning, feature engineering

DESCRIPTION DES DONNEES

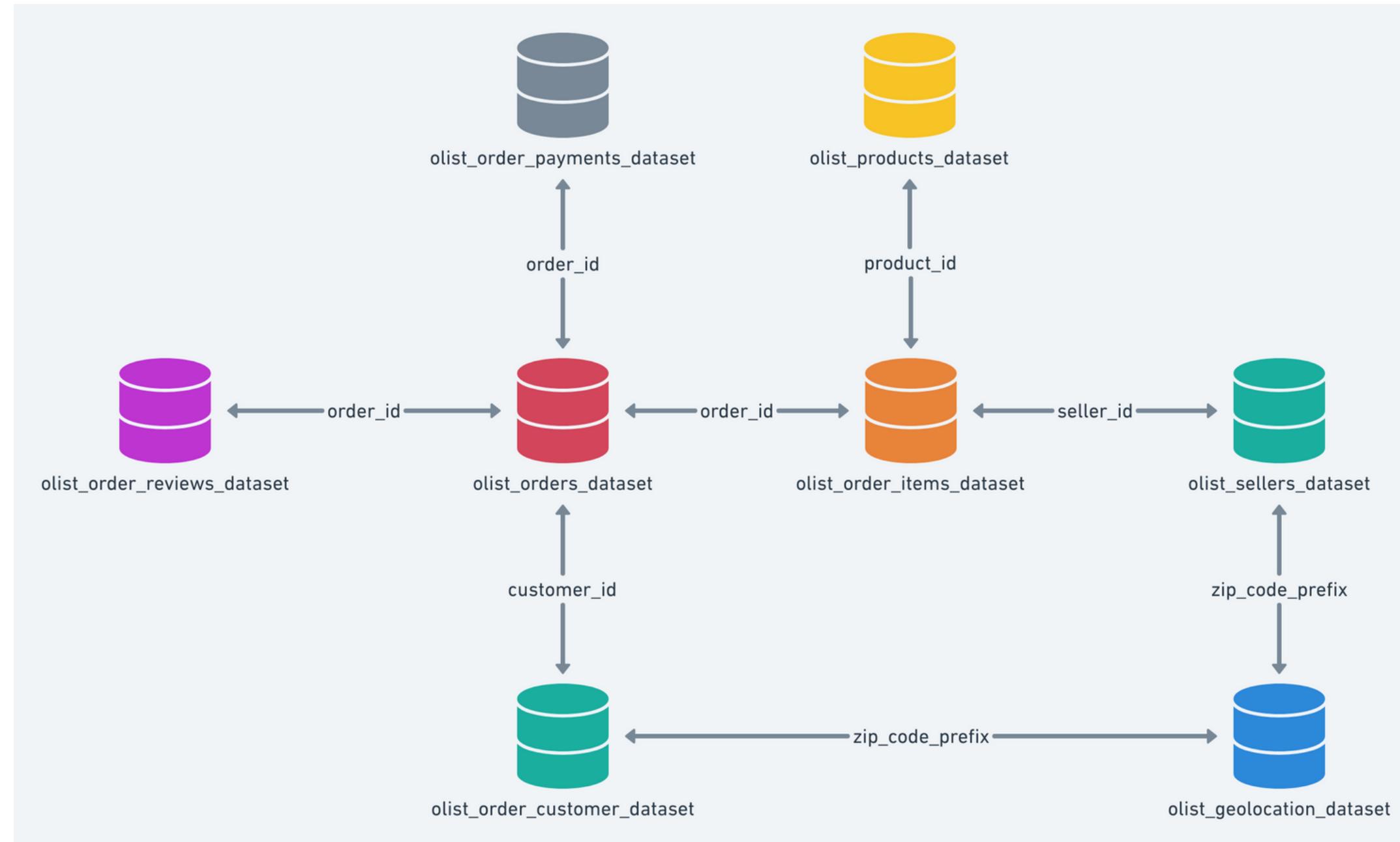


Nous disposons de différentes sources de données, chacune décrivant un sujet spécifique lié aux ventes en ligne.

En détail, nous disposons de 9 datasets (1 non représenté, *nr*) pour les données :

- **Les Clients :**
 - *olist_customers_dataset.csv*
- **Les Vendeurs:**
 - *olist_sellers_dataset.csv*
- **Les Géolocalisations :**
 - *olist_geolocation_dataset.csv*
- **Les Commandes clients:**
 - *olist_orders_dataset.csv*
 - *olist_order_items_dataset.csv*
 - *olist_order_payments_dataset.csv*
 - *olist_order_reviews_dataset.csv*
- **Les Produits :**
 - *olist_products_dataset.csv*
 - *product_category_name_translation.csv (nr)*

DESCRIPTION DES DONNEES



Les différentes tables sont globalement bien complétées. Peu de valeurs nulles, et clés primaires correctement établies.

L'analyse est réalisée uniquement pour les clients déjà livrés.

Je réalise des jointures durant l'analyse exploratoire, pour arriver à une seule table finale.

DISTRIBUTION DES COMMANDES

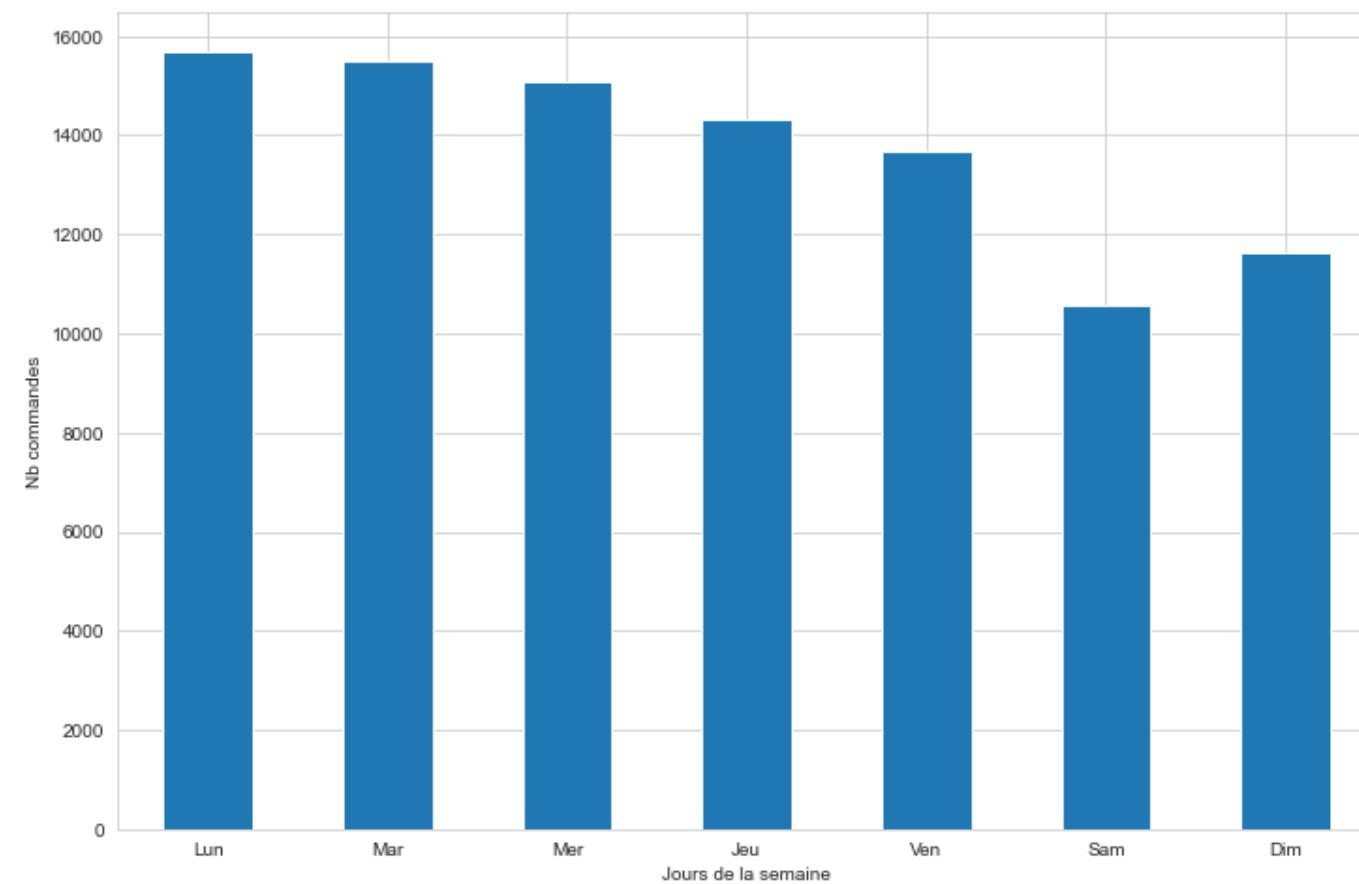


On observe :

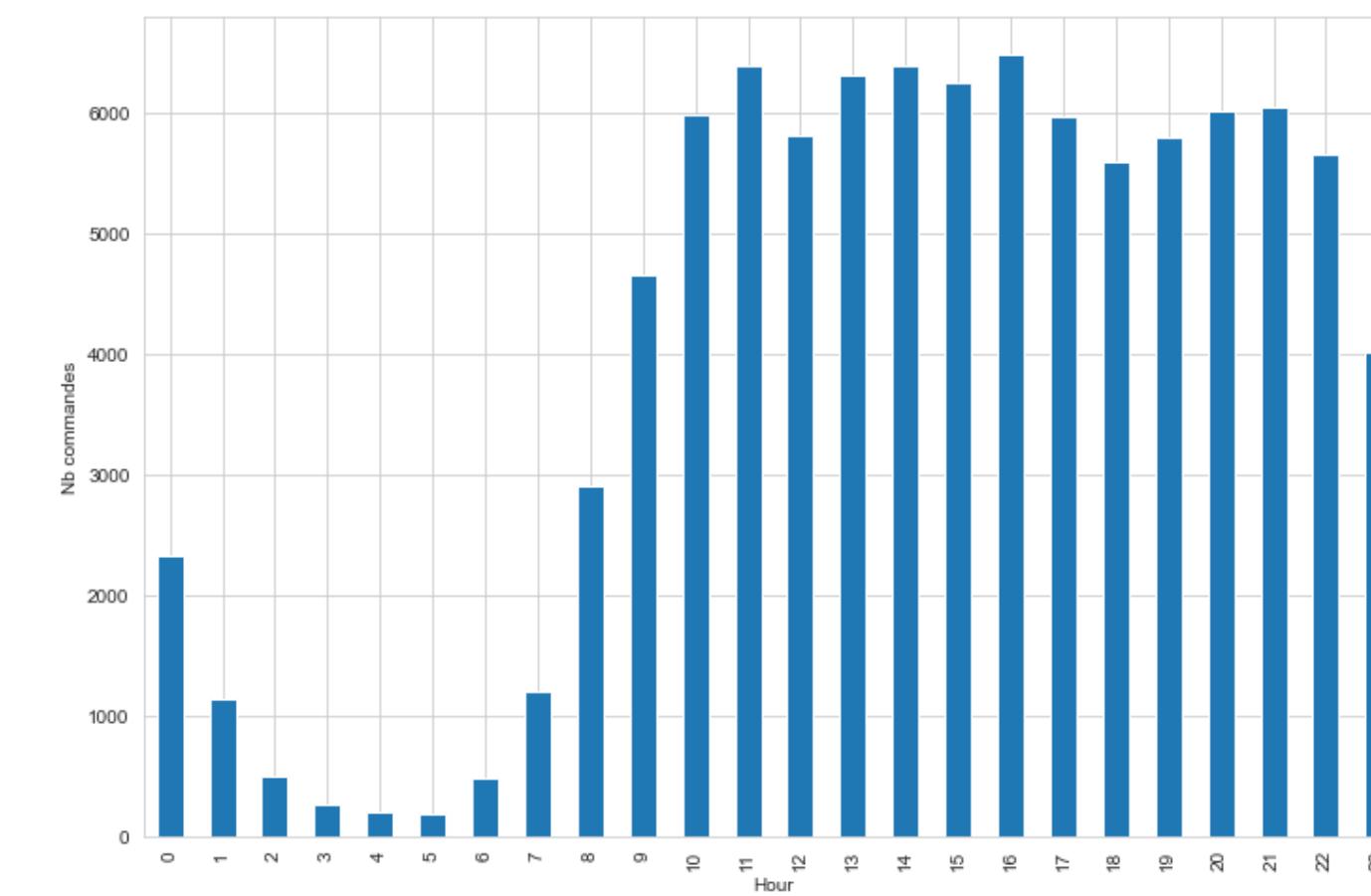
- Un plateau où les commandes sont presques nulles entre Octobre 2016 et Janvier 2017, ce qui peut correspondre au début des ventes de l'entreprise ou la mise en service de la base de données.
- Un pic de commande le 24 novembre 2017 ce qui peut correspondre au black friday de 2017 qui était le 24 novembre.

DISTRIBUTION DES COMMANDES

Nombre de commandes par jours de la semaine

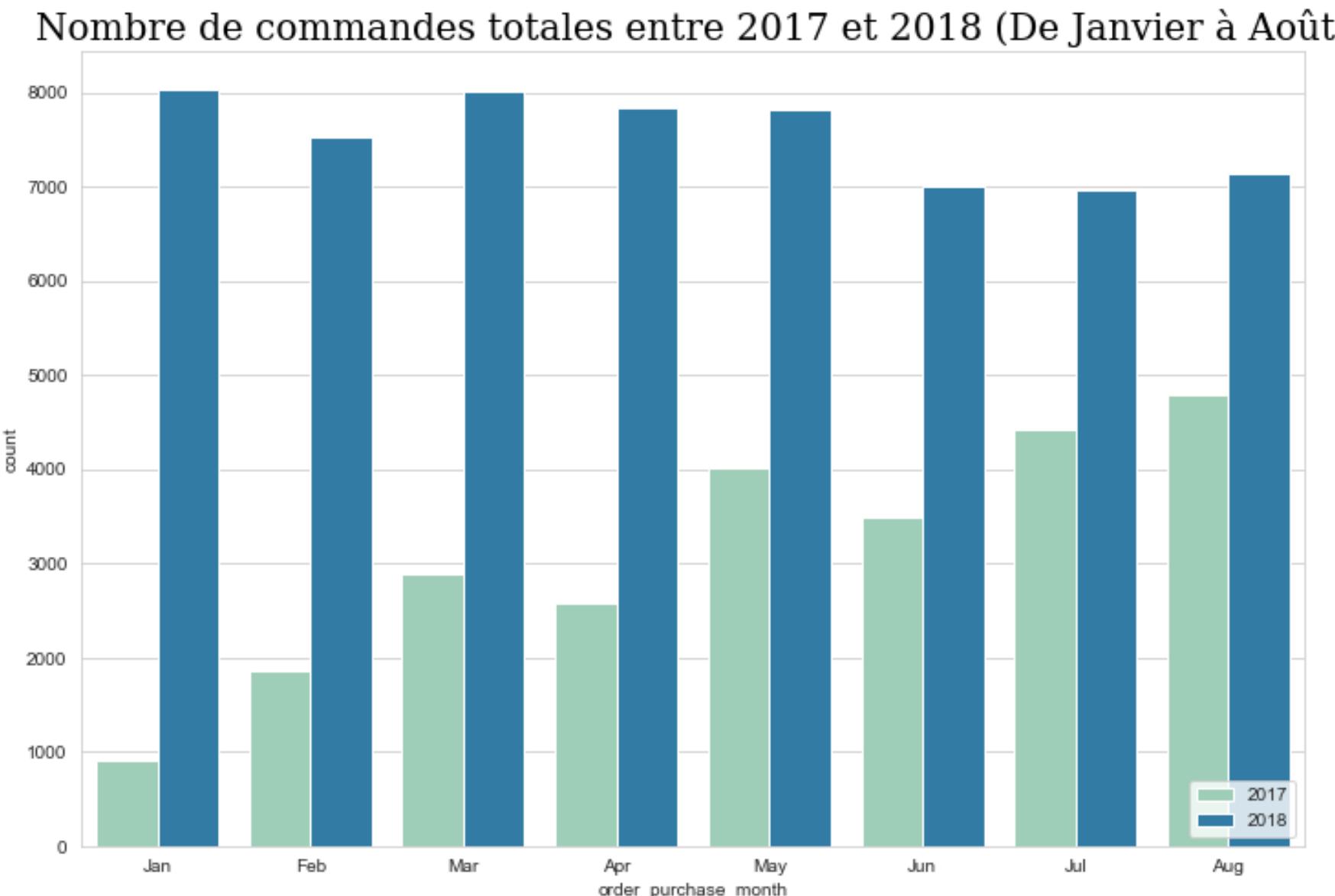


Nombre de commandes par heure de la journée



Les clients commandent davantage en début de semaine, et en fin de matinée ou d'après-midi.

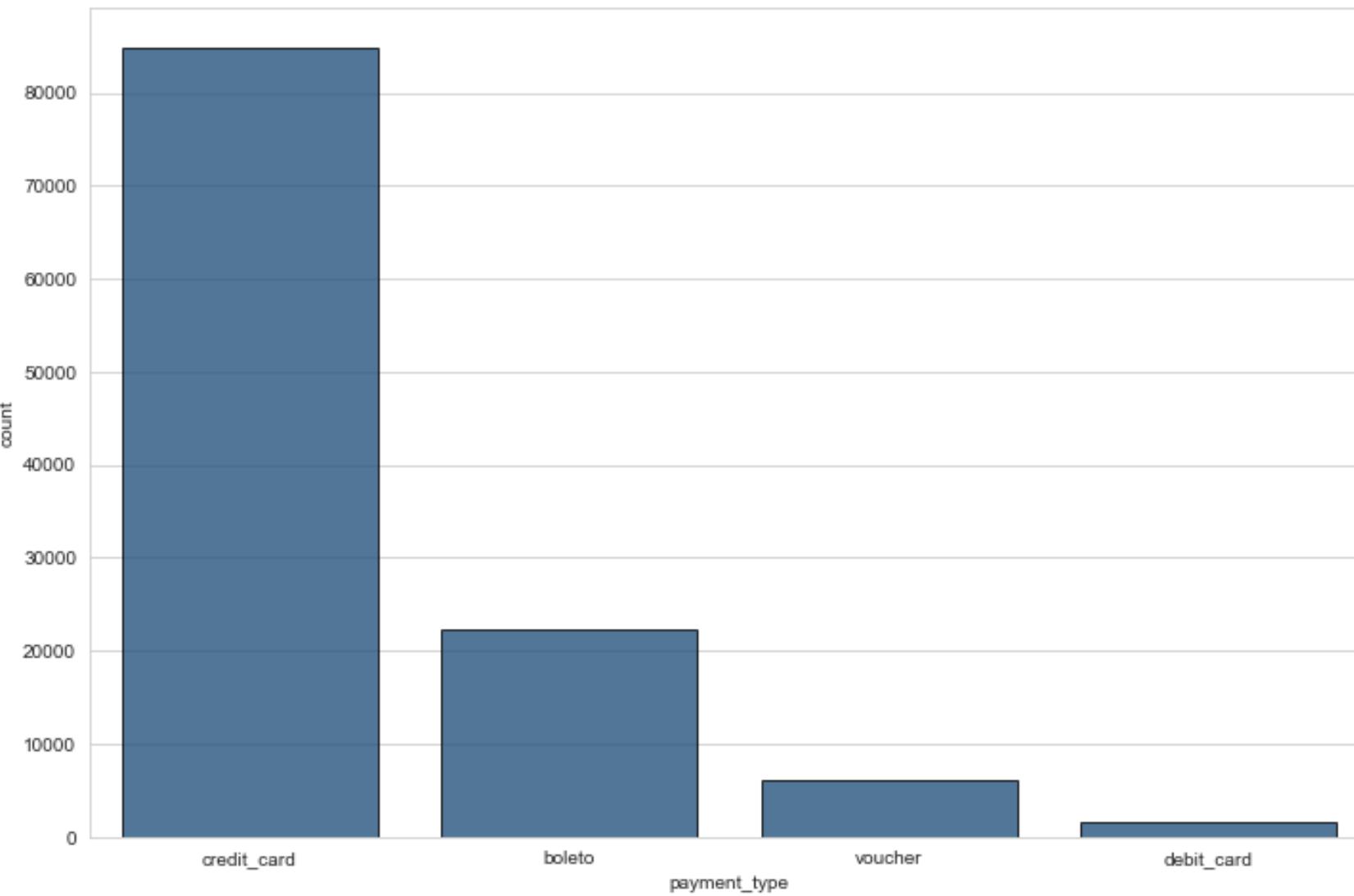
DISTRIBUTION DES COMMANDES



Augmentation significative du nombre de commandes entre 2017 et 2018 (pour les mois entre Janvier et Août, calculée à 141 %).

MOYENS DE PAIEMENT

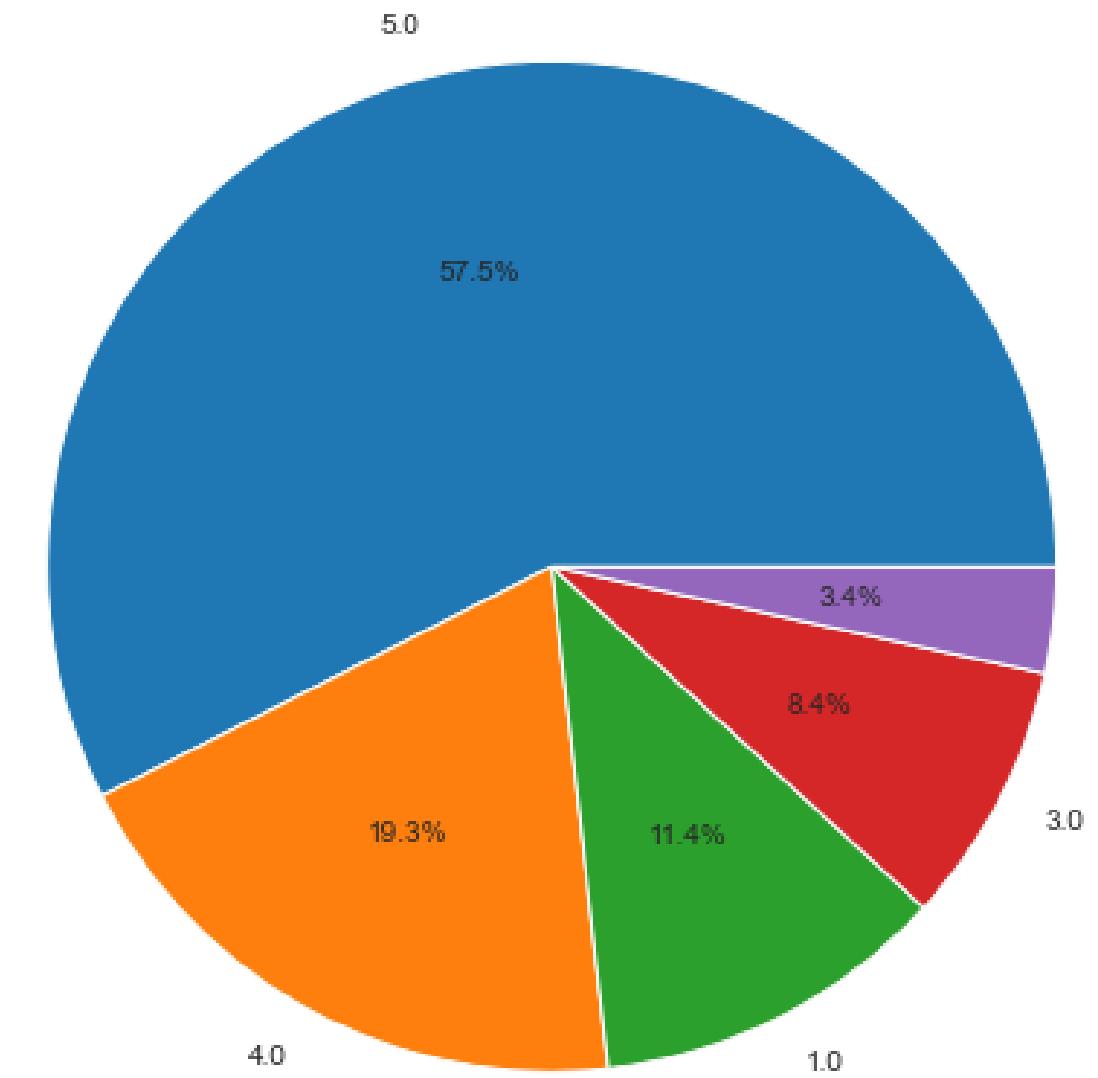
Les moyens de paiement utilisés sur le site



La majorité des paiements sont réalisés en carte bancaire.

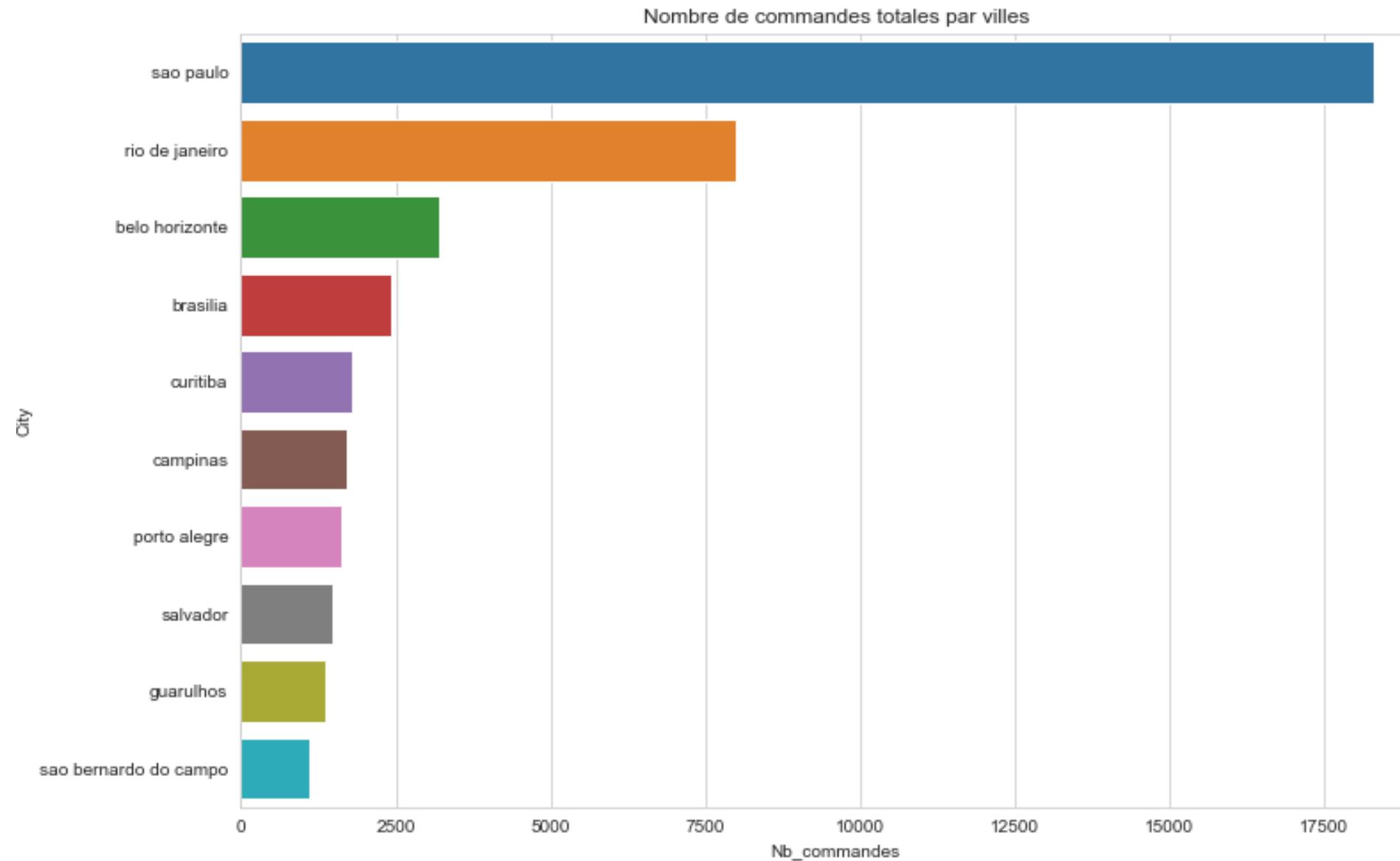
RÉPARTITION DES NOTES ATTRIBUÉES AUX COMMANDES

Répartition des notes attribuées aux commandes



- Plus des 3/4 des commandes du site ont reçues une note supérieur ou égale à 4.
- Moins d'1/4 des commandes du site ont reçues une note inférieur ou égale à 3.

RÉPARTITION DES COMMANDES PAR VILLES



La majorité des commandes ont été passées à Sao Paulo et Rio de Janeiro.

FEATURE ENGINEERING

Création d'un dataset client



On pourra caractériser un client selon les features suivantes :

- Son **identifiant** (`customer_unique_id`)

Les features relatives à une segmentation RFM :

- La **récence d'achat**, exprimée en nombre de jours depuis la date actuelle (`days_last_order`)
- Le **nombre de commandes effectuées** sur la période totale, correspondant à sa **fréquence d'achat** (`nb_orders`)
- La **dépense totale** (`total_spend`)

Ainsi que d'autres features caractérisant les clients, et pouvant compléter une segmentation :

- Le **montant moyen dépensé** (`mean_price_order`)
- La **note de satisfaction moyenne** (`mean_review_score`)
- Le **nombre total d'articles achetés** (`total_items`)
- Le **nombre moyen d'article par commande** (`mean_nb_items`)
- Son **zip code** (`customer_zip_code_prefix`)

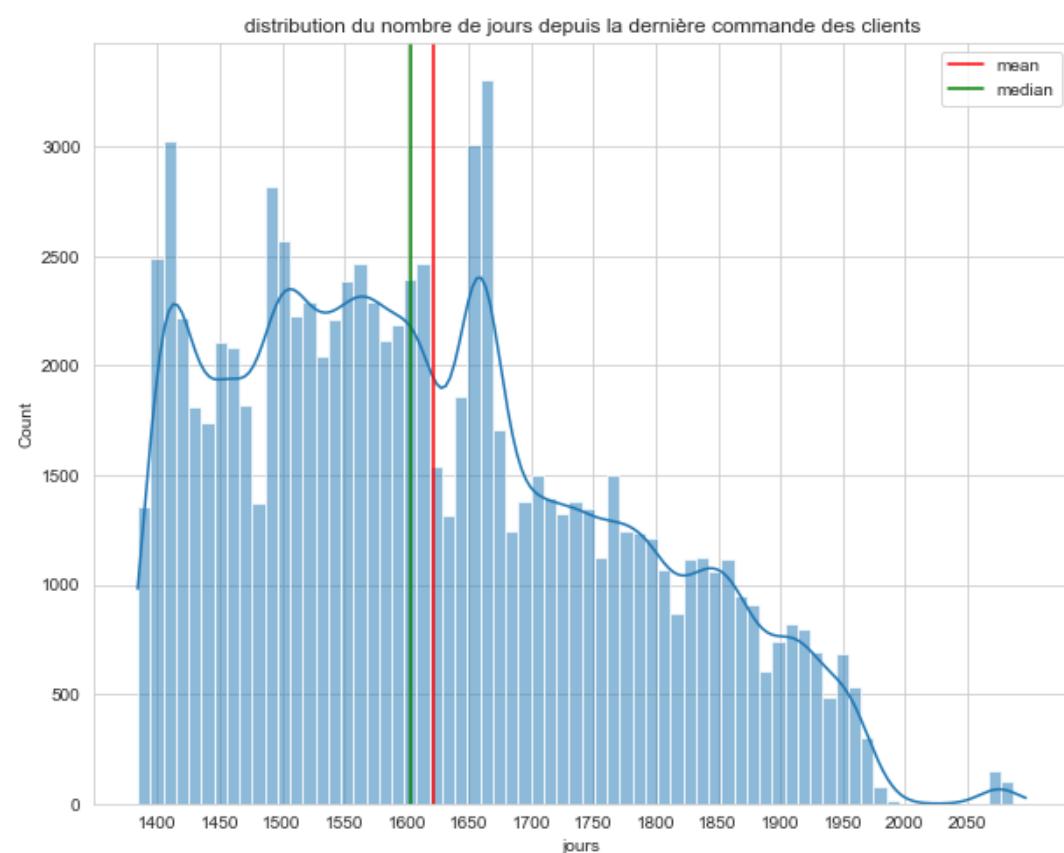
FEATURE ENGINEERING



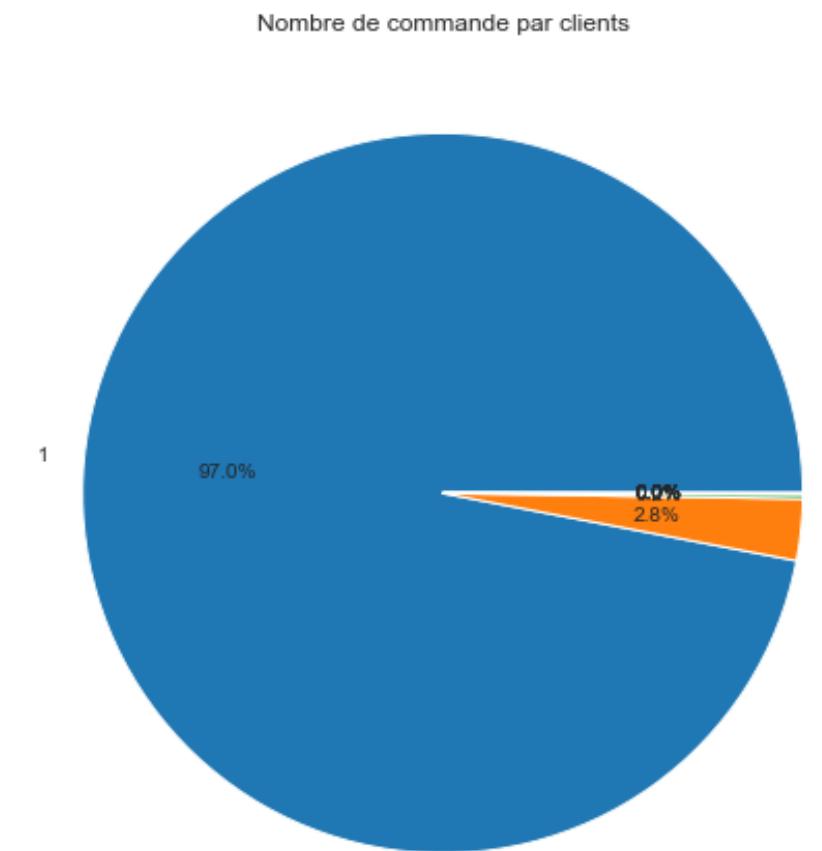
Dataset client



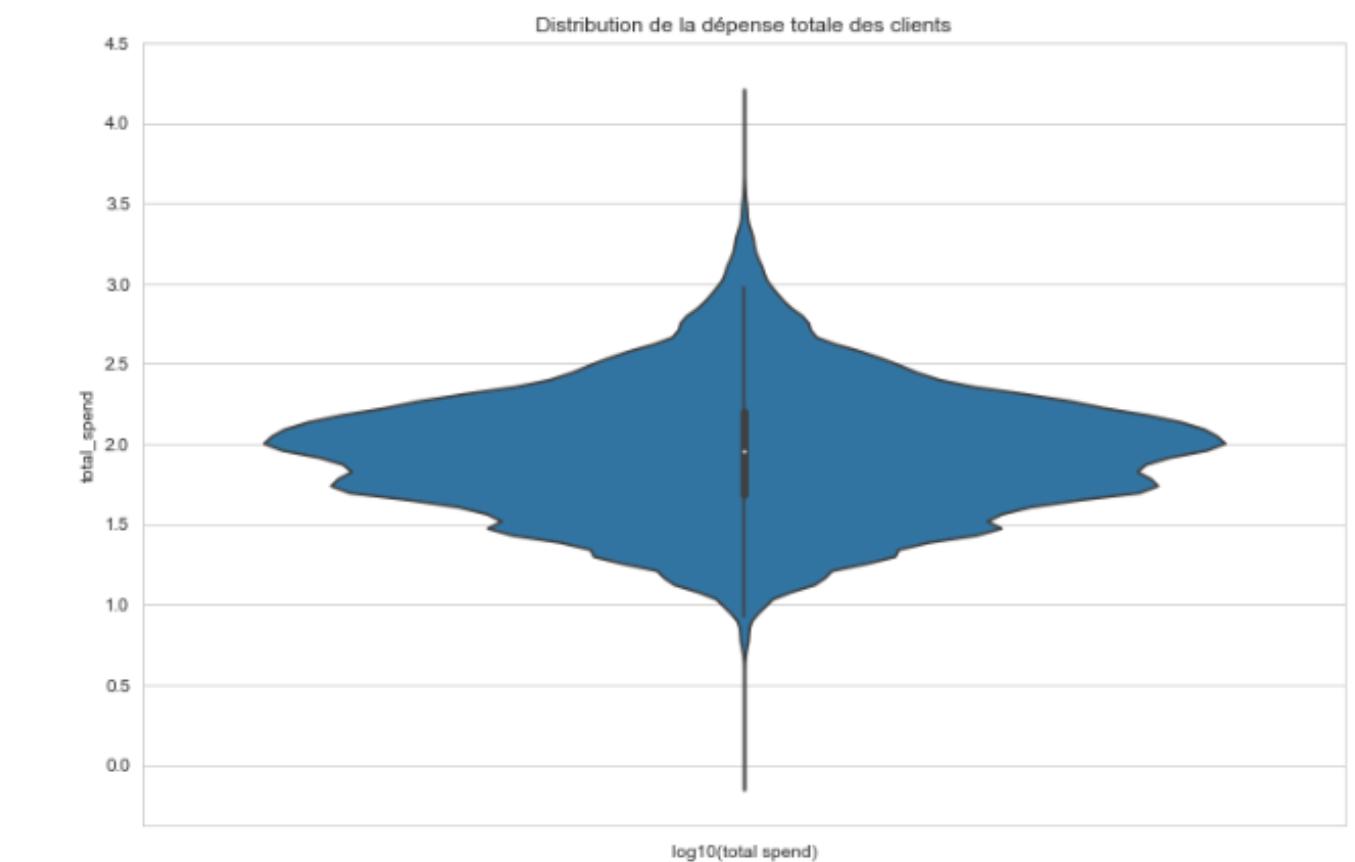
Création des variables RFM



Récence
Nombre de jours depuis de la dernière commande.



Fréquence d'achat client :
Comme indiqué dans la description du projet, seul 3% des clients ont passés plus d'une commande.



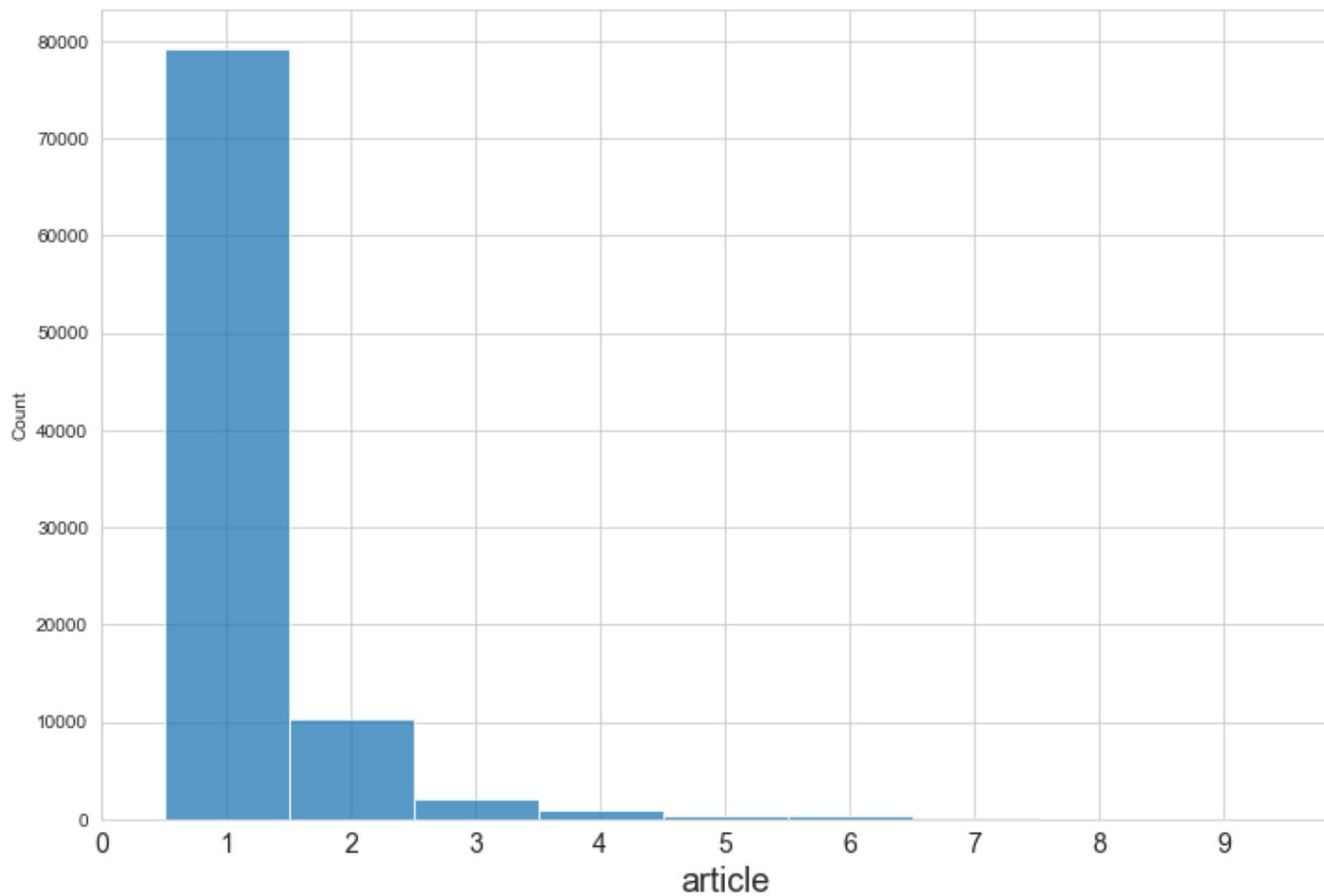
Dépense totale par client:
Médiane à environ 100 Réal brésilien (19 euros).

FEATURE ENGINEERING

Dataset client



Nombre d'articles par commande



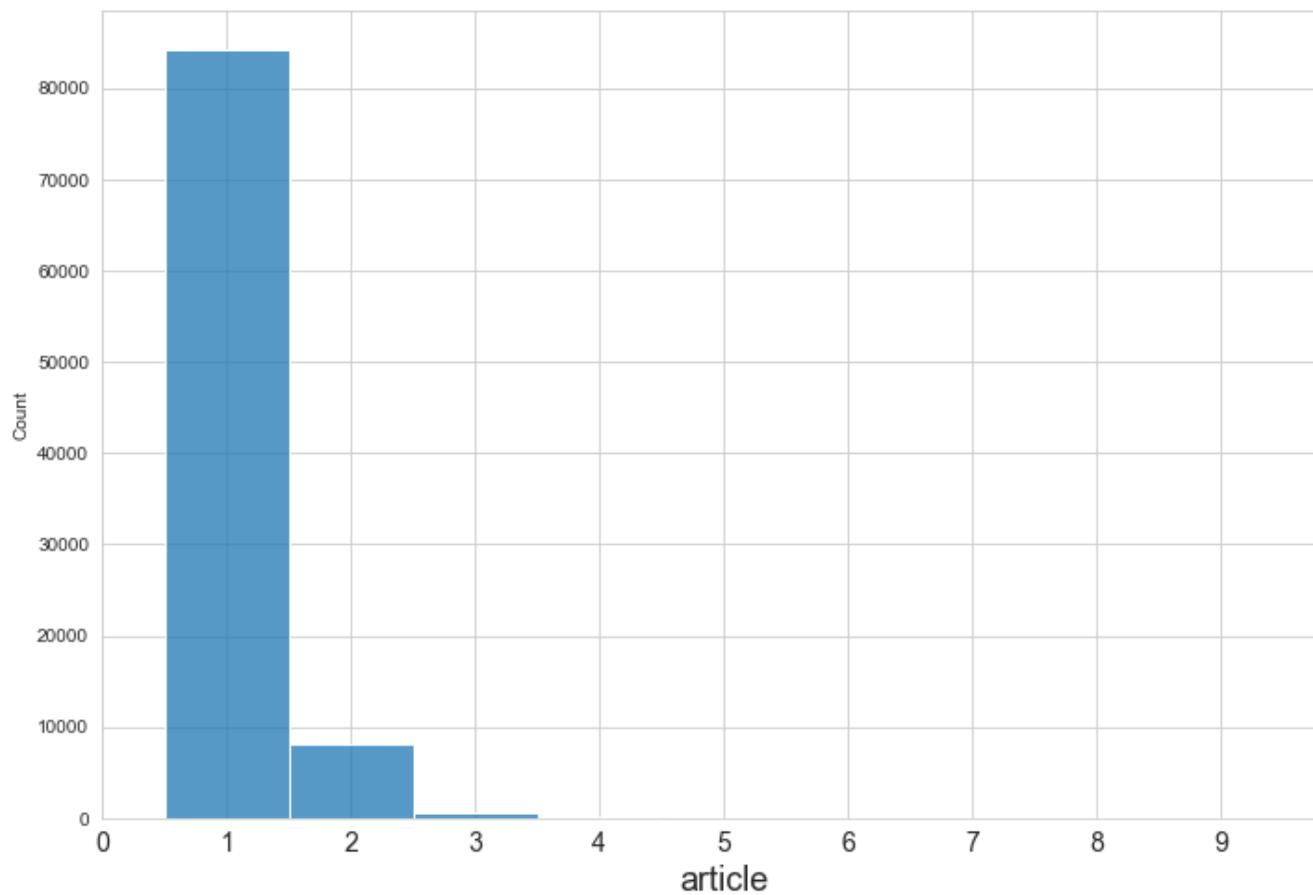
La majeur partie des commandes ne contiennent qu'un seul article.

FEATURE ENGINEERING

Dataset client



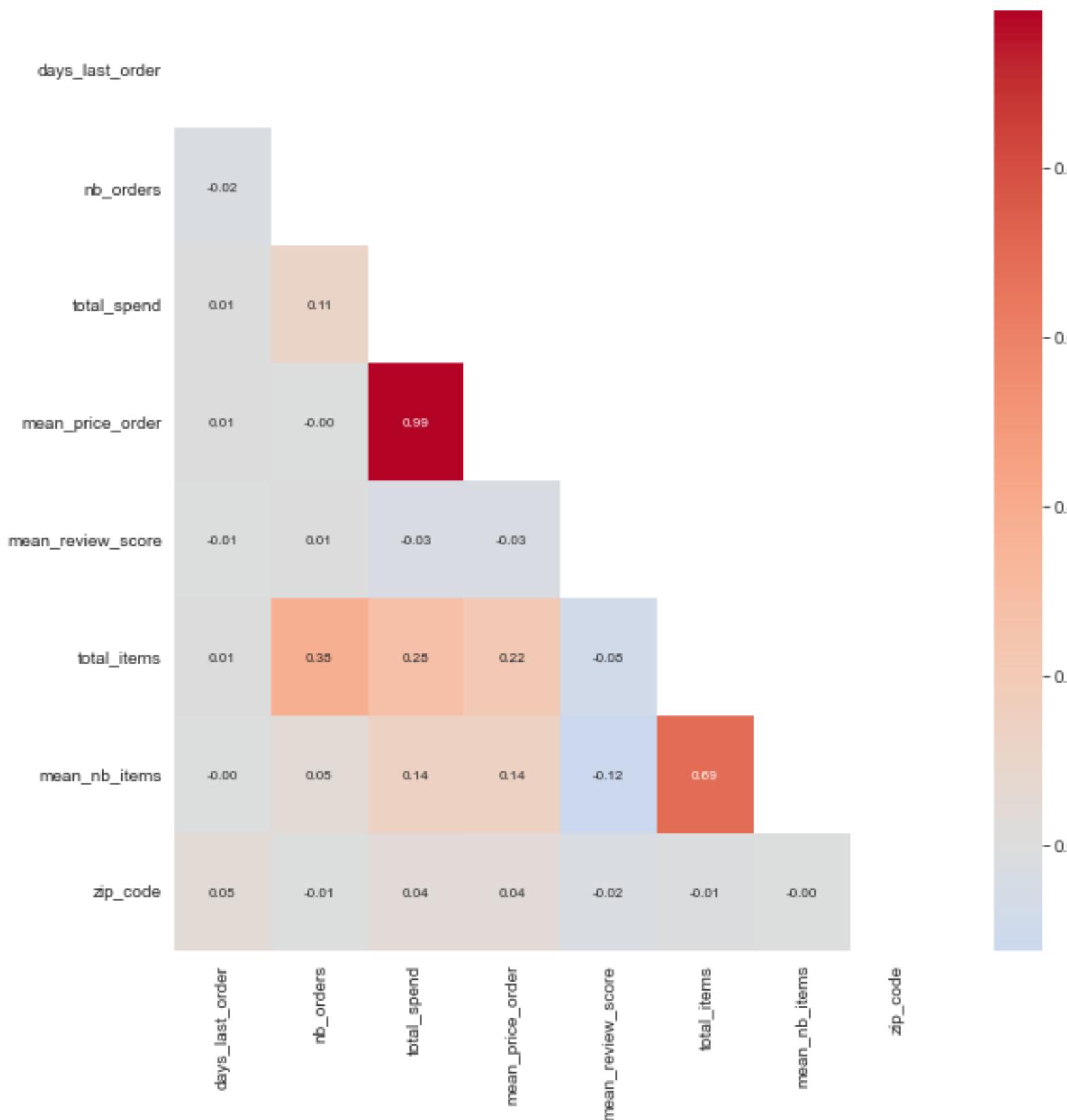
Nombre d'articles moyen par commande



Les distributions sont très similaires entre le nombre d'article total et le nombre d'article moyen. Cela peut s'expliquer par le fait que la majorité des clients n'ont commandé qu'une seule fois. Cette variable sera donc moins corrélée au fil du temps.

FEATURE ENGINEERING

Heatmap des corrélations linéaires df_clients



Dataset client

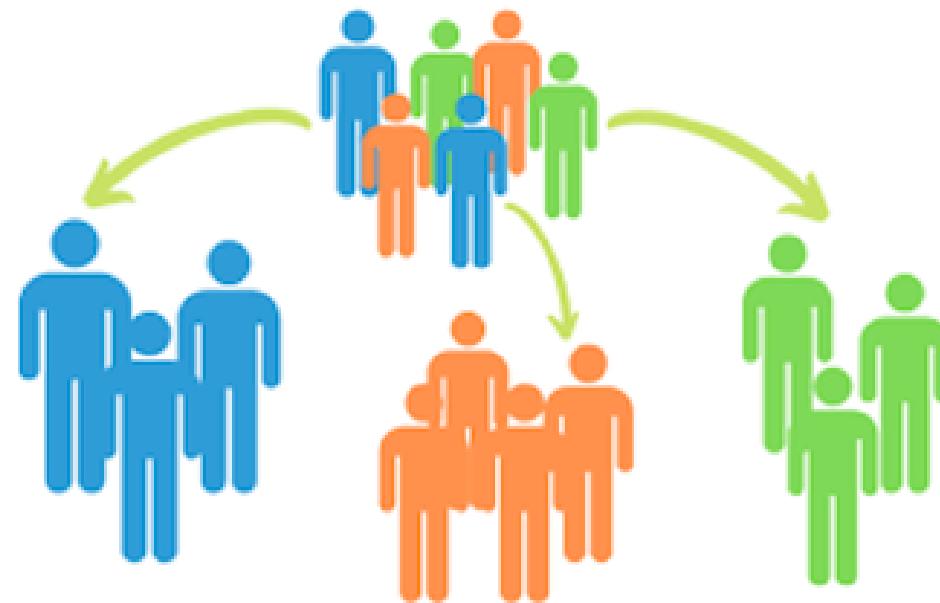


- Le prix total (`total_spend`) est très corrélé au prix moyen par commande (`mean_price_order`)
- Le nombre d'articles total (`total_items`) est très corrélé au nombre moyen d'article par commande (`mean_nb_items`).

Cela s'explique par le fait que la majorité des clients n'ont commandé qu'une seule fois.

SEGMENTATION

METHODE



1. Segmentation RFM

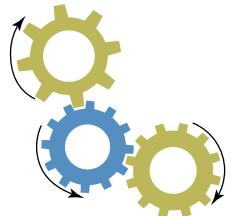
a. K-MEANS

b. DBSCAN

2. Segmentation sur l'ensemble des
variables avec le meilleur modèle

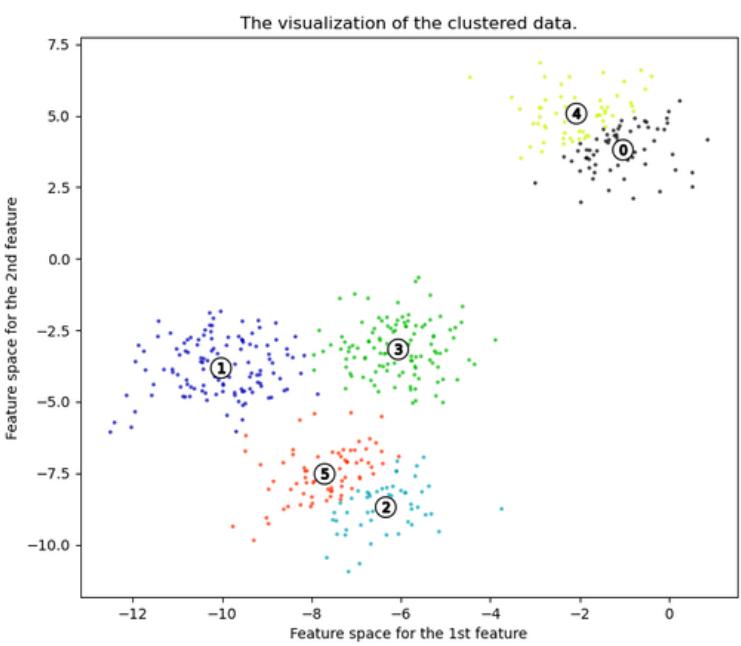
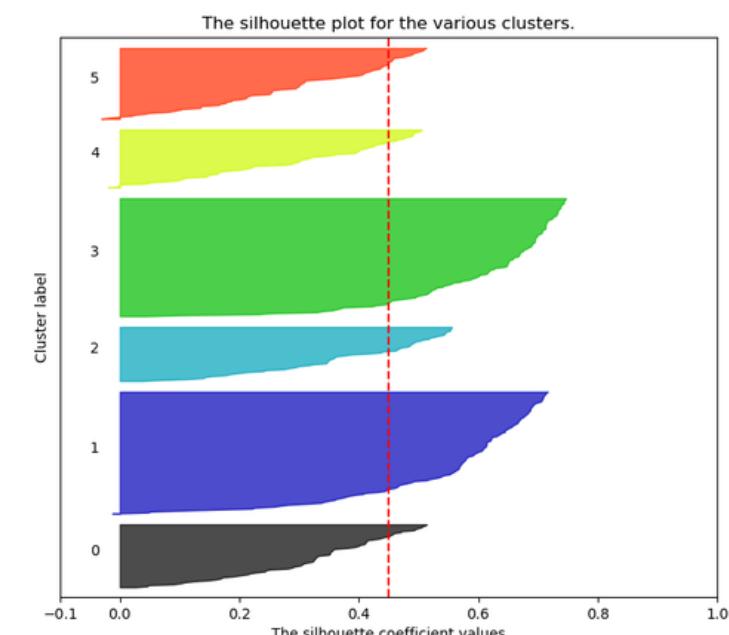
K-MEANS

Preprocessing : Normalisation MinMax

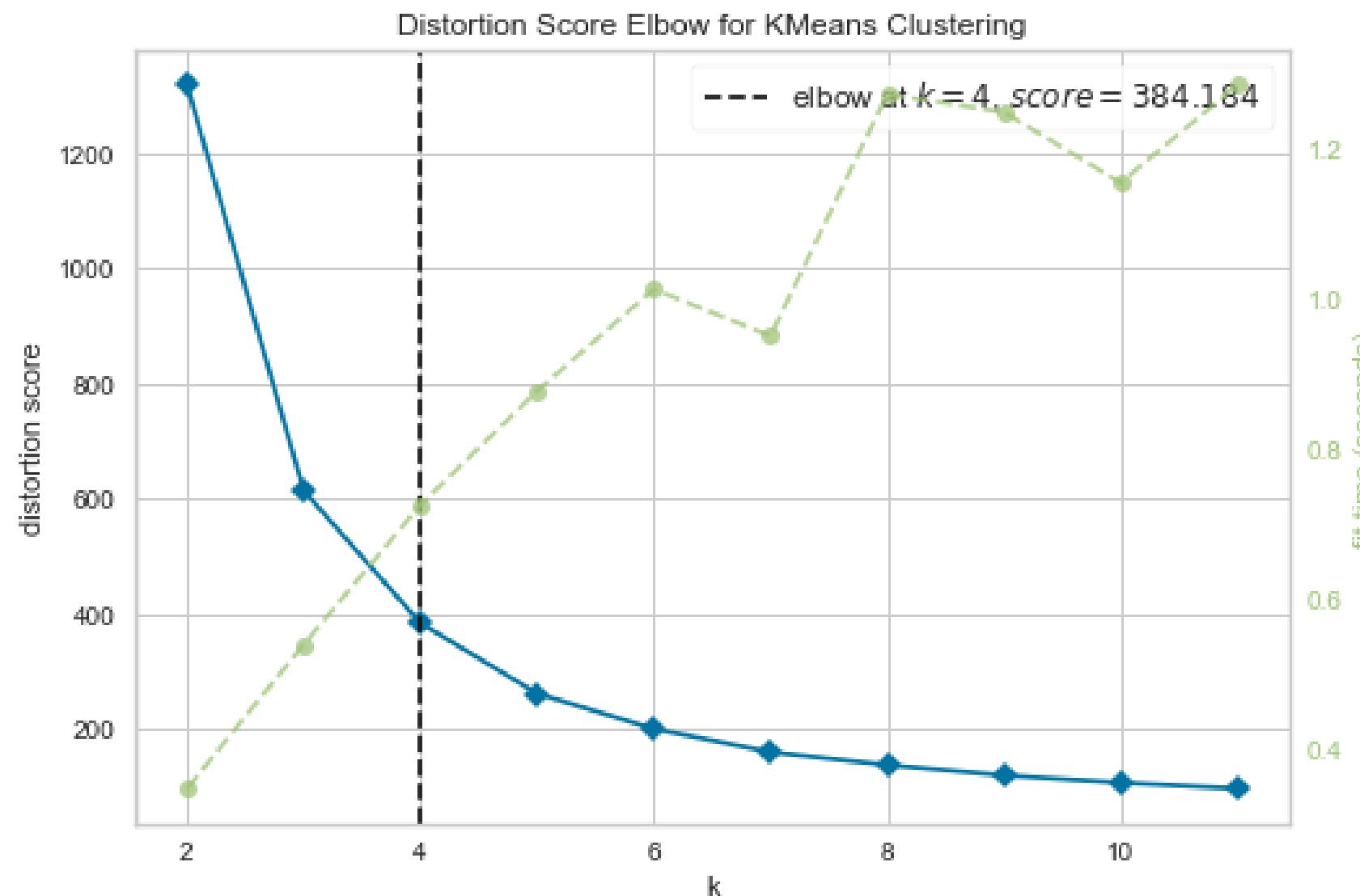


Critères d'évaluation :

- Clusters denses mais séparés entre eux
- Score moyen de silhouette (qui doit être maximal)
- Temps d'entraînement



Détermination du nombre de clusters



Méthode du coude (Elbow) :

On réalise une itération du K-Means sur un interval de nombre, les métriques (ici distorsion et temp d'entraînement) sont calculées et affichées sur une courbe.

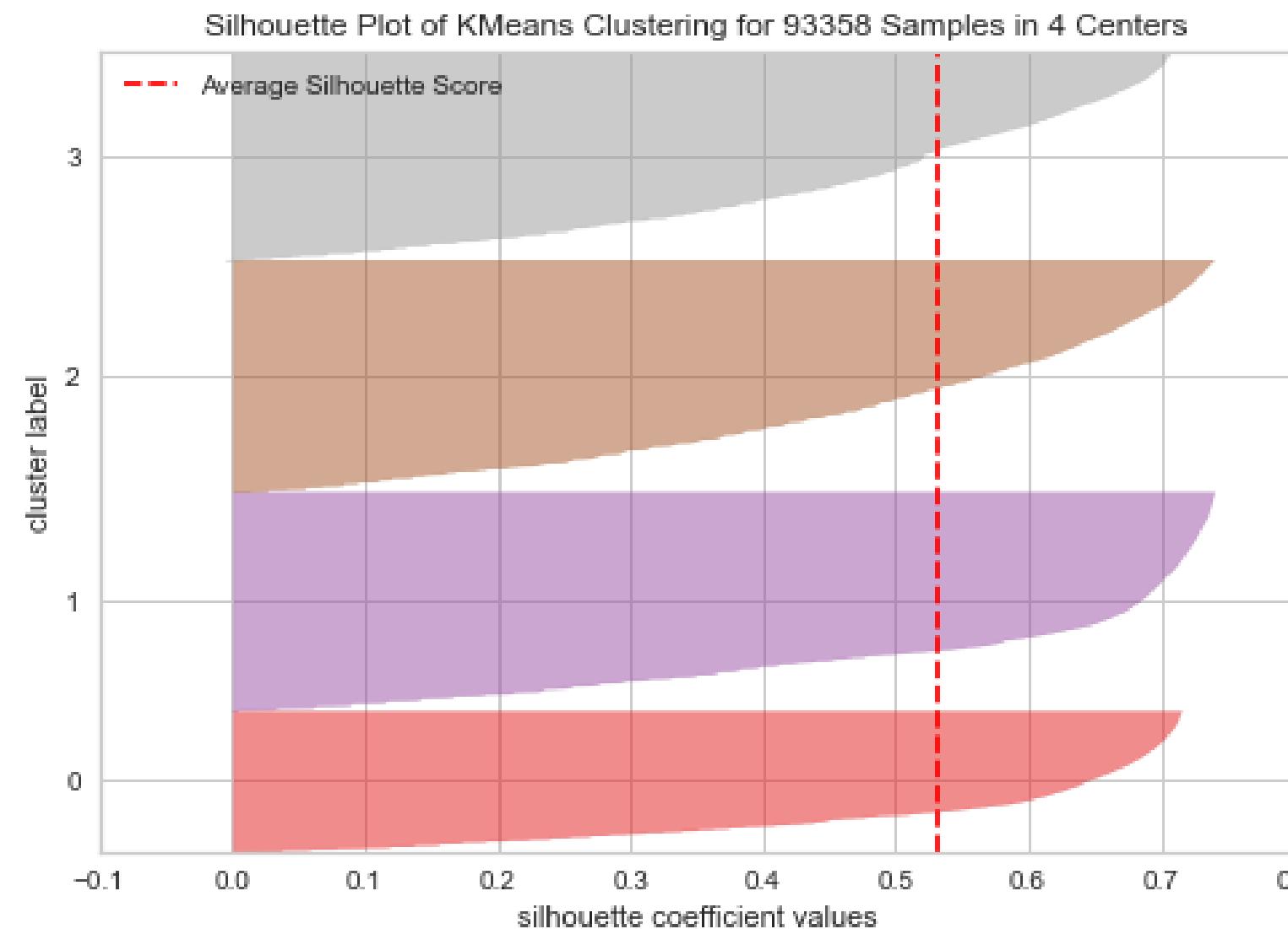
Le point d'inflexion représentant la distorsion (courbe bleu) indique le meilleur K. ici le meilleur K=4.

Le score de distorsion représente la somme des distances au carré de chaque point à son centre assigné.

Le calcul est réalisé grâce à la librairie Python Yellowbrick.

K-MEANS

Silhouette



Silhouette plot des 4 clusters

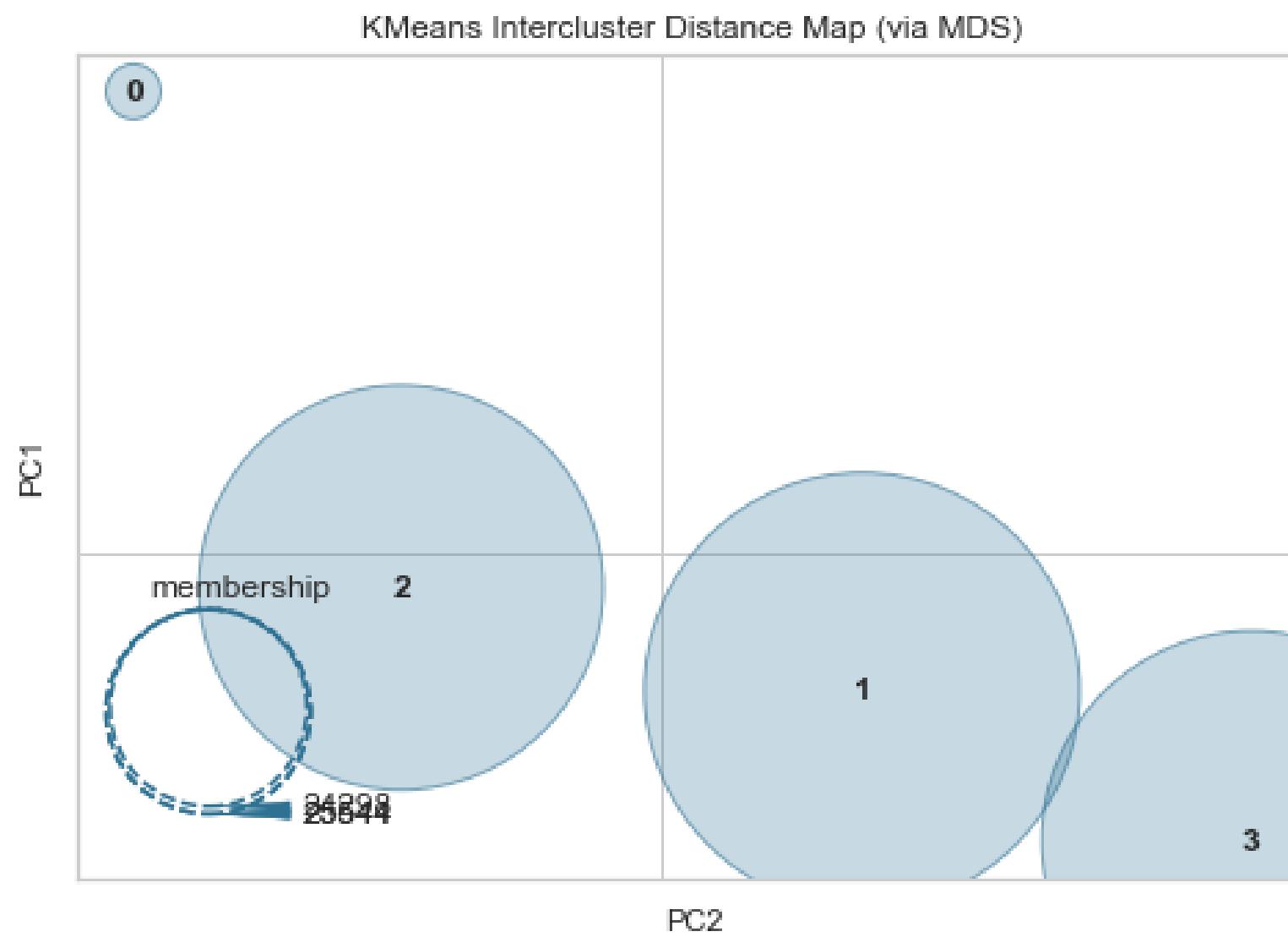
Le meilleur K déterminé étant de 4 groupes, une visualisation des coefficients de silhouette pour chaque cluster est affiché.

Ce graphique permet de visualiser la densité et la séparation des clusters.

Ici, les groupes semble bien répartis entre eux.

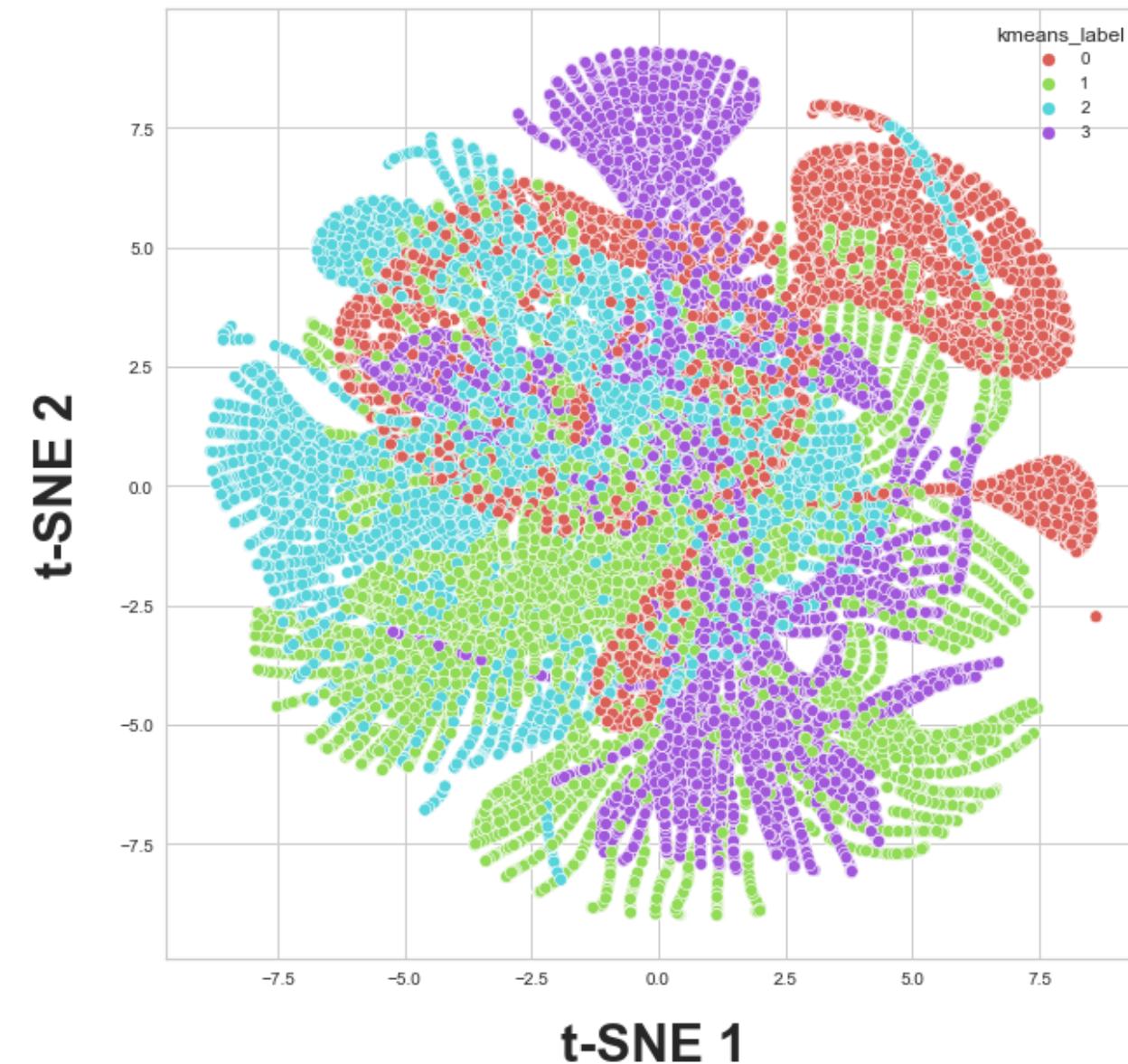
K-MEANS

Distance intercluster et visualisation t-SNE



En projetant les clusters sur les 2 premières composantes principales de la MDS (multidimensional scaling), on observe que les groupes ne sont pas tous bien séparés entre eux.

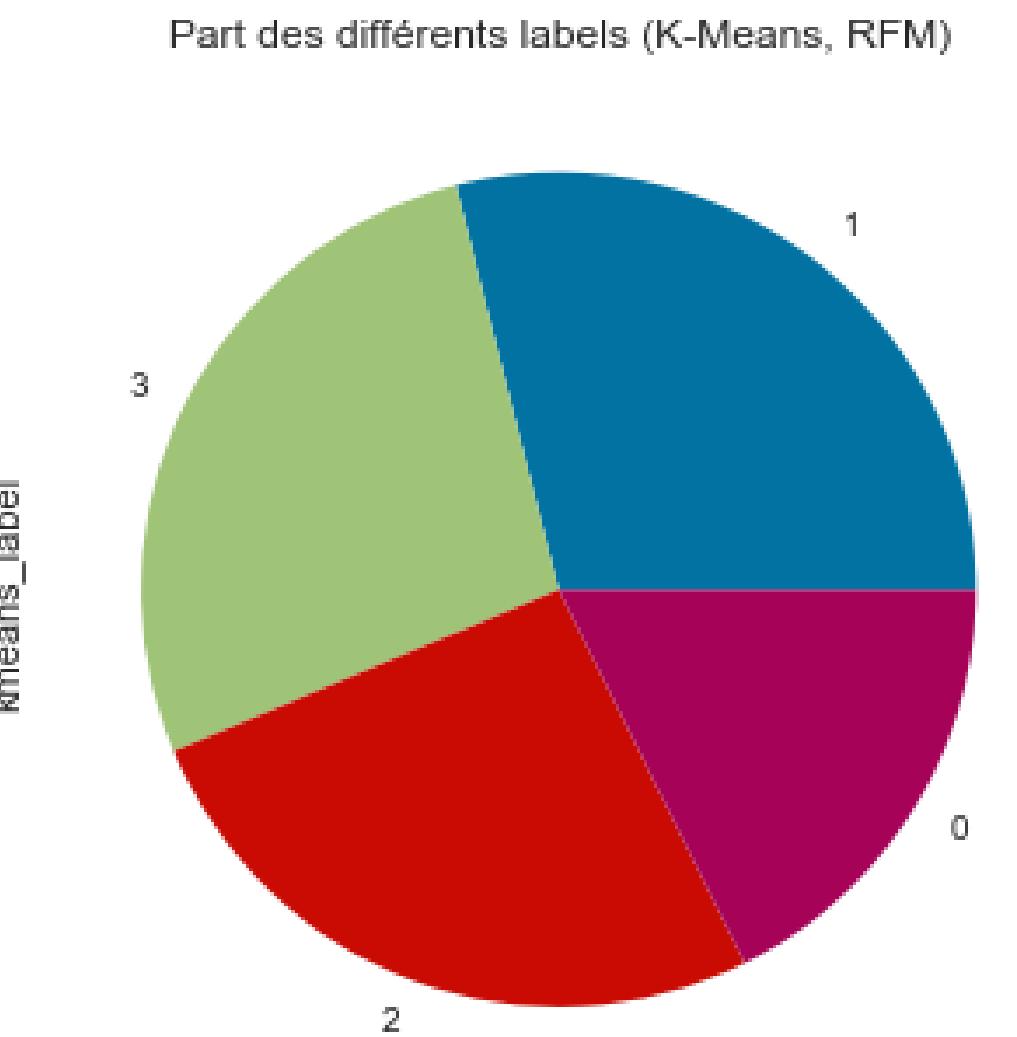
Mise en évidence des clusters t-SNE



La visualisation t-SNE semble indiquer également une mauvaise séparation des groupes (nombre d'itération égal à 300, et une complexité de 40).

K-MEANS

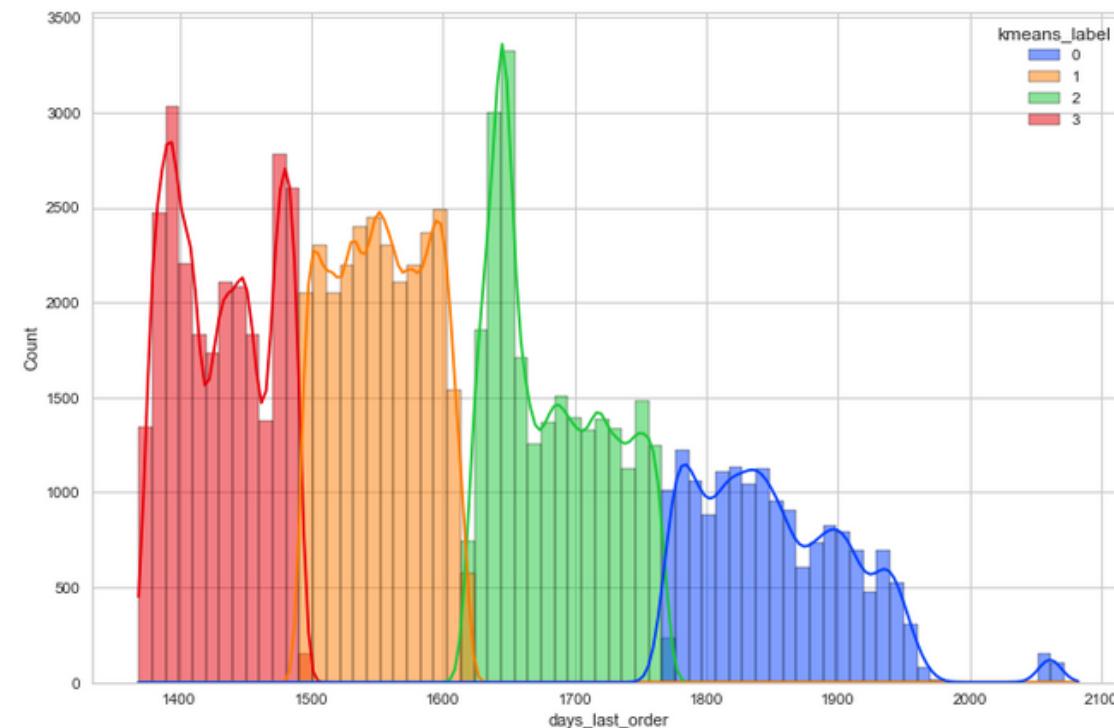
Interprétation métier des clusters



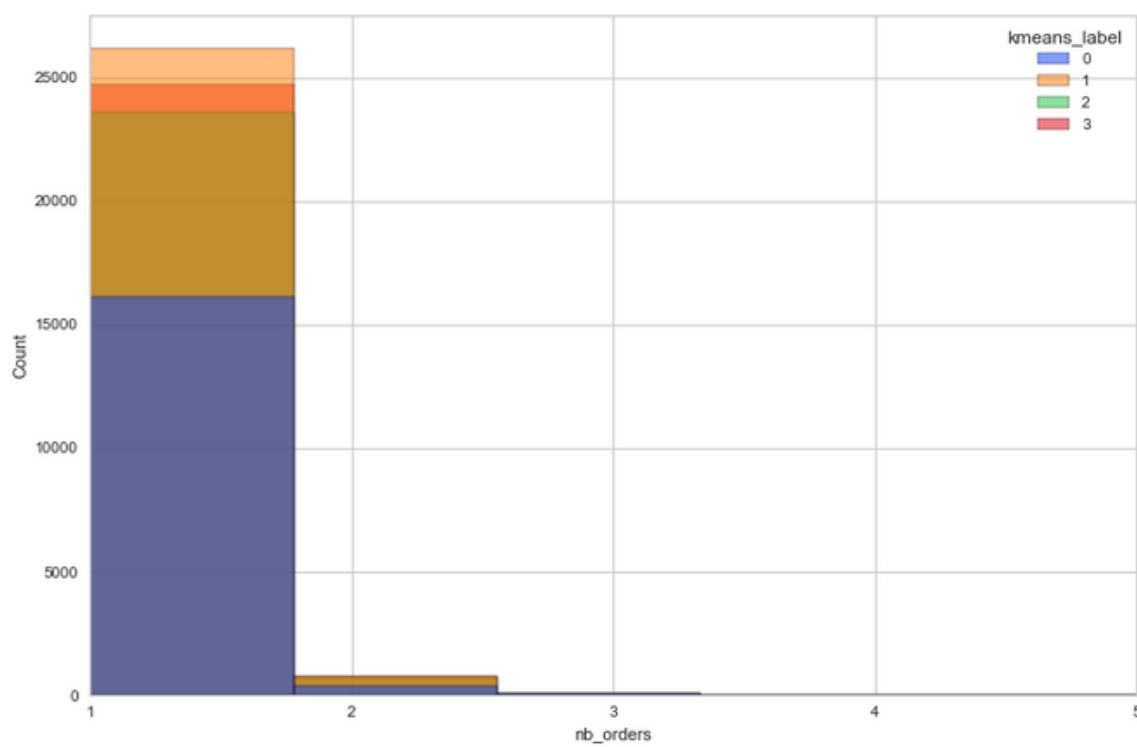
Comme indiqué par la visualisation silhouette précédemment et le pie plot ci-dessus, les groupes sont plutôt équilibré en nombre de clients.

K-MEANS

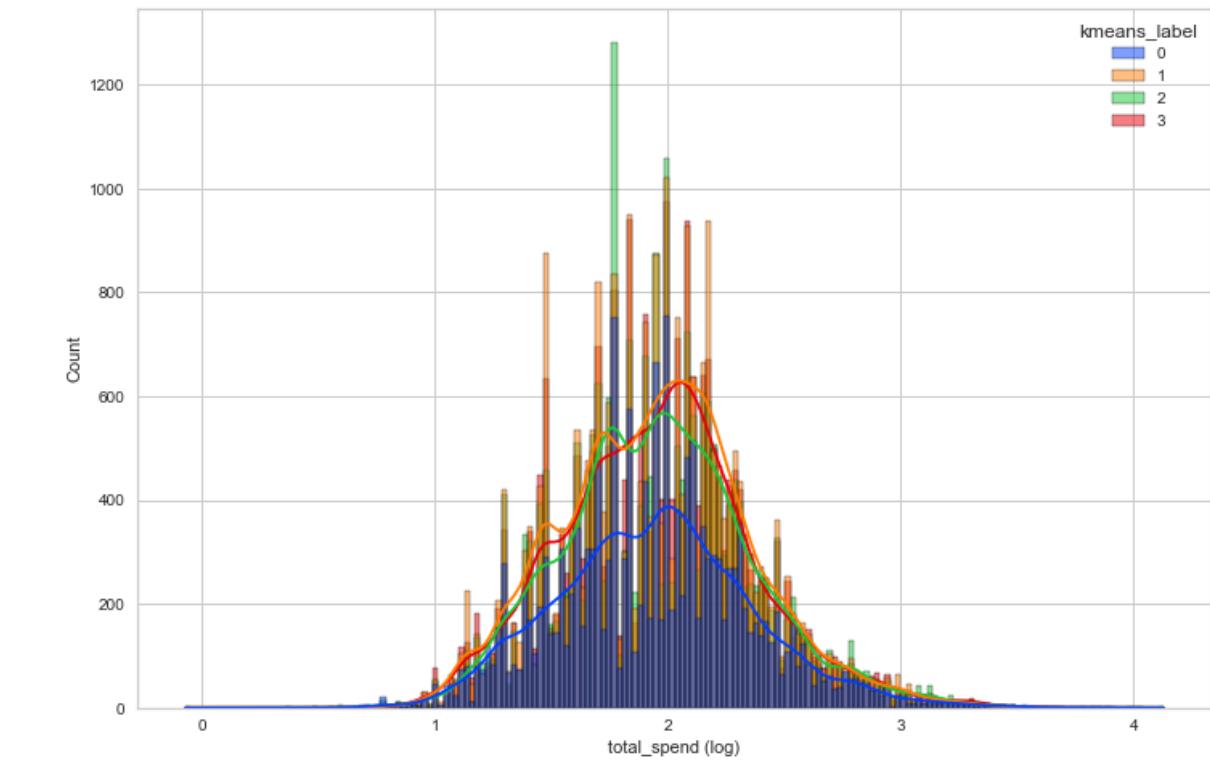
Interprétation métier des clusters



Récence



Fréquence



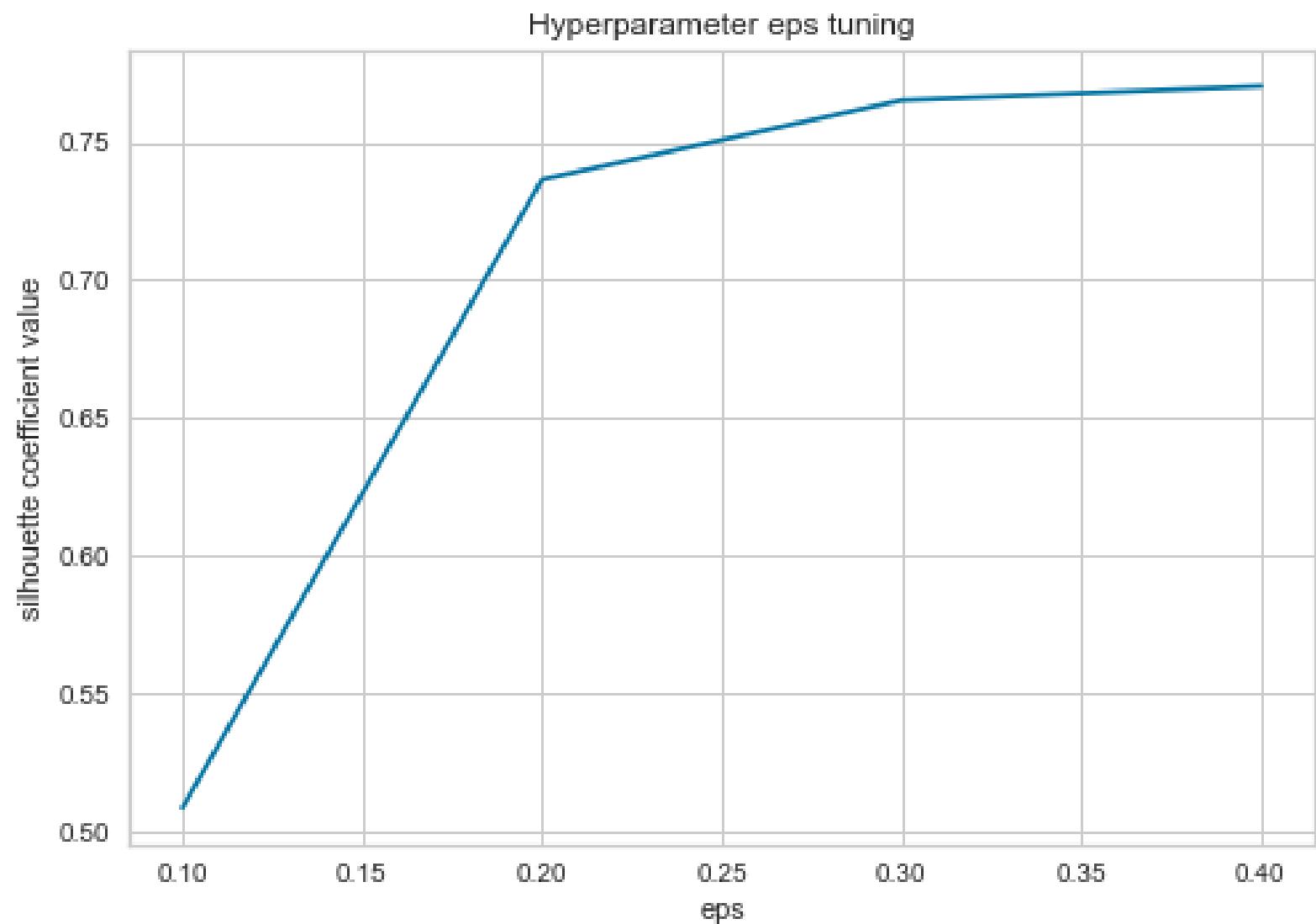
Montant

On observe une segmentation clair des différents clusters sur la variable lié à la récence.

- Label = 3 : clients qui ont achetés le plus récemment.
- Label = 0 : clients qui ont achetés le moins récemment. Ce qui explique le cluster très éloigné sur la visualisation MSD.

Il est cependant impossible de segmenter les clients à l'aide des variables F et M. La segmentation RFM ne semble pas efficace dans ce cas de figure. Ceci peut être expliqué par le fait que les clients n'ont commandés qu'une seule fois. Nous réaliserons plus tard une segmentation sur l'ensemble des variables, pour voir si cela permet d'améliorer le clustering.

Optimisation du modèle : recherche de l'hyperparamètre 'eps' optimal



On obtient un *eps* optimal (valeur où le score de silhouette est maximale) pour une valeur égale à
0.40

DBSCAN

Performance du modèle

	algorithme	silhouette	nb_clusters	time_sec
0	DBSCAN	0.770343	2	448.73

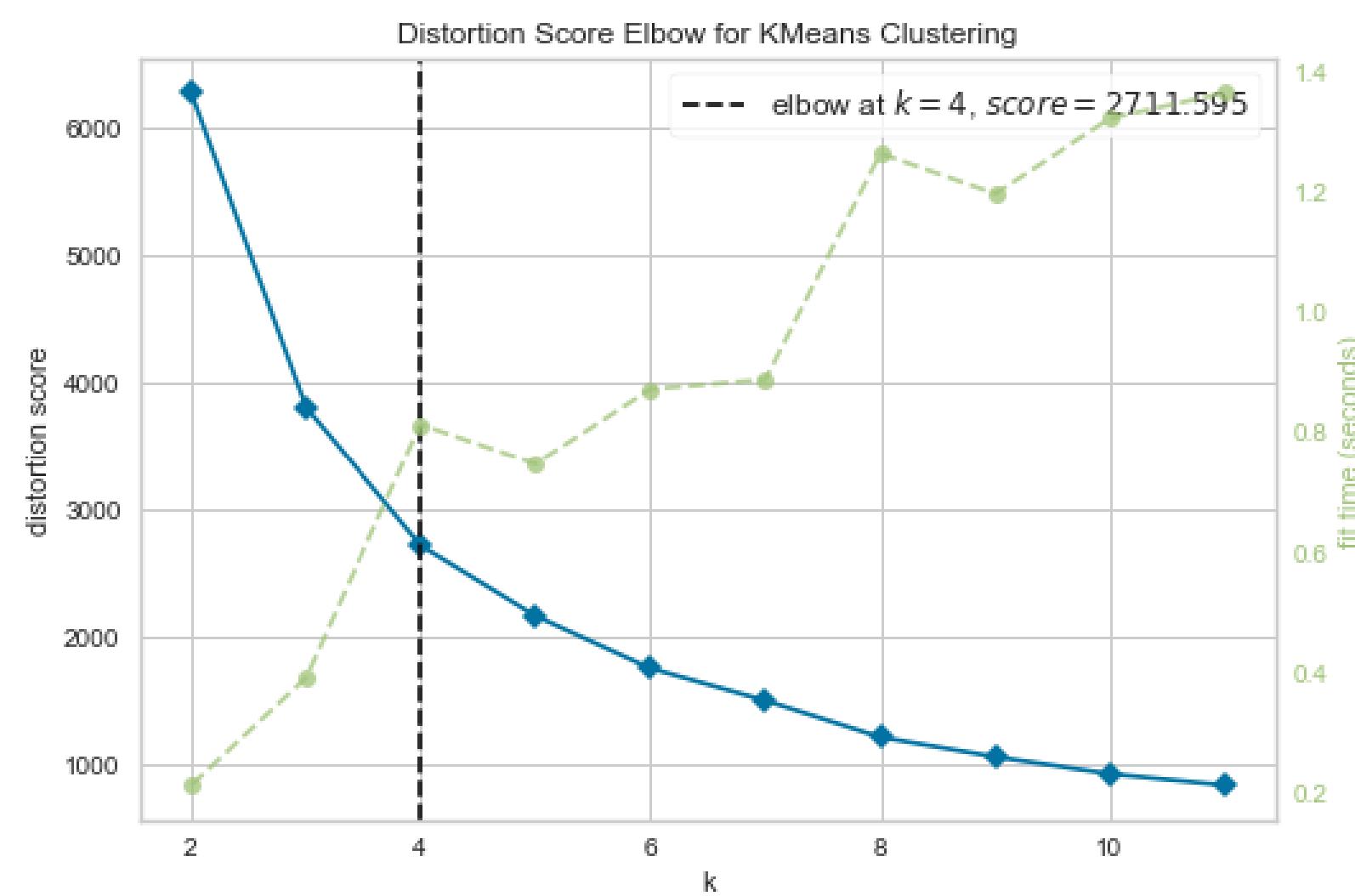
Distribution des groupes

0	93357
-1	1

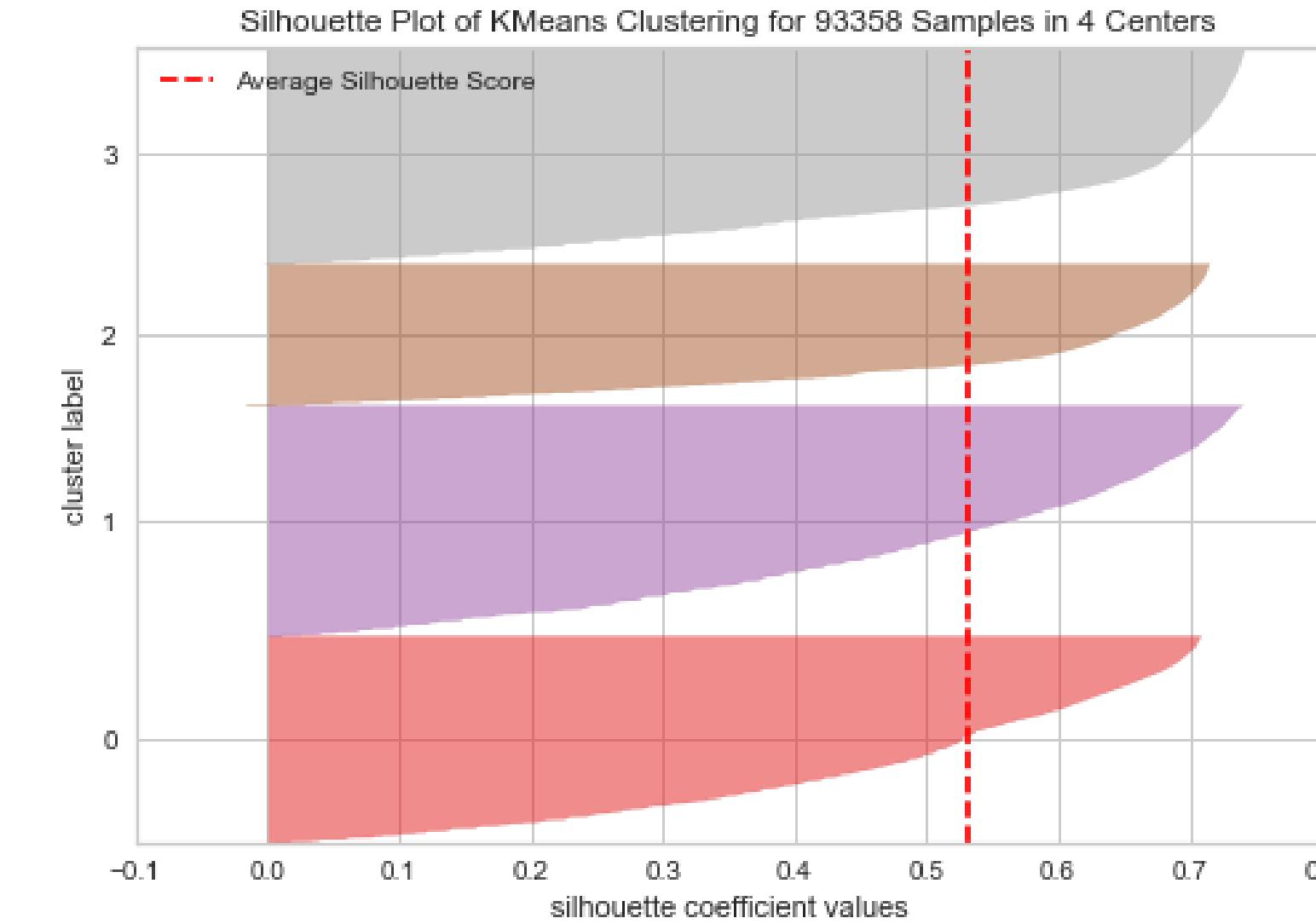
L'algorithme DBSCAN réalise un clustering de 2 clusters. En observant la proportion de clients dans chaque cluster on remarque une anomalie. Seulement 1 client est inclus dans un cluster, et tous les autres clients dans un autre cluster. De plus, l'algorithme est très coûteux en temps. (plus de 7min pour entraîner le modèle).

K-Means sera utilisé pour le reste de la modélisation.

SEGMENTATION SUR L'ENSEMBLE DES VARIABLES (K-MEANS)



Meilleur k pour k=4

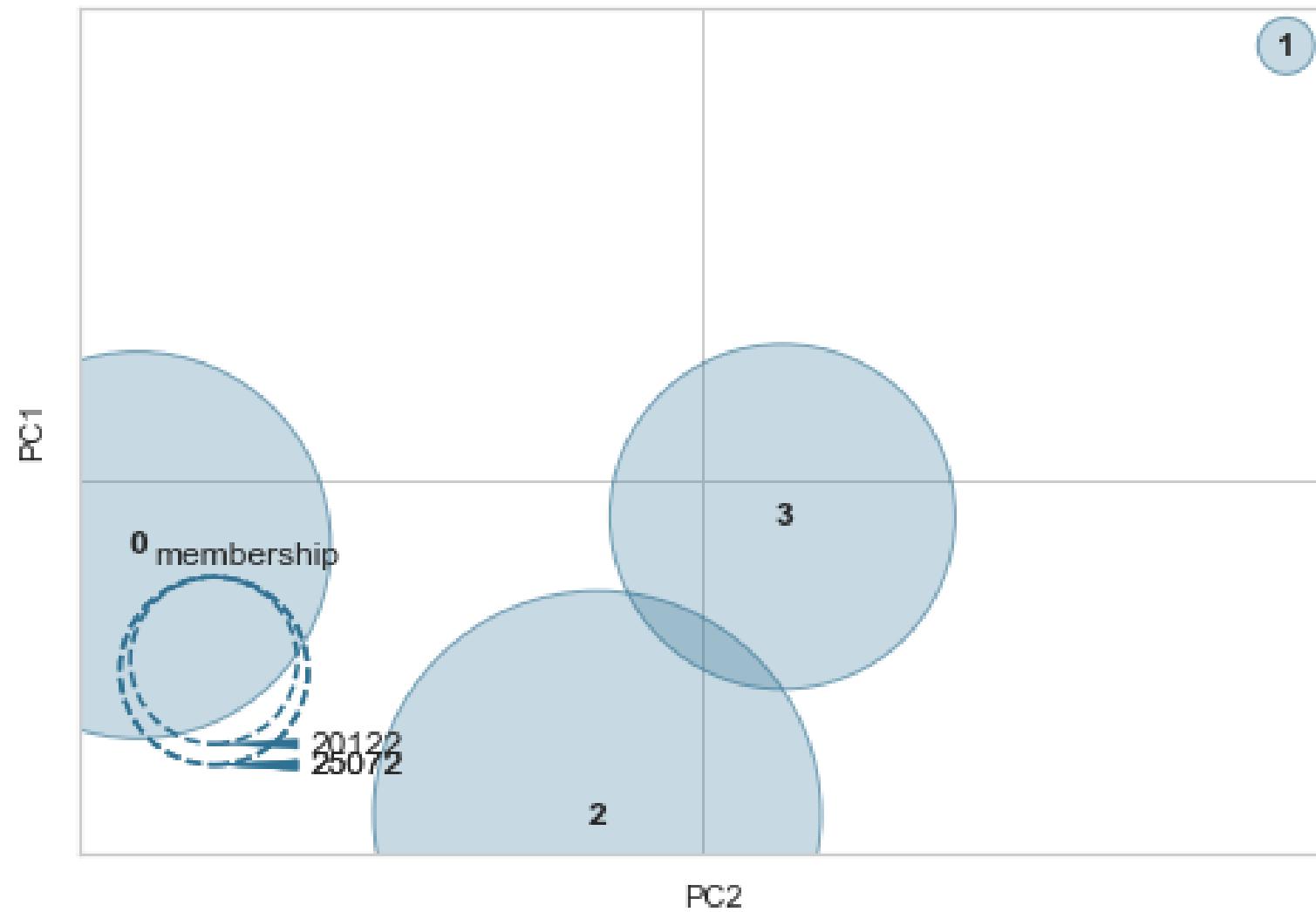


Les groupes semble bien répartis entre eux.

SEGMENTATION SUR L'ENSEMBLE DES VARIABLES (K-MEANS)

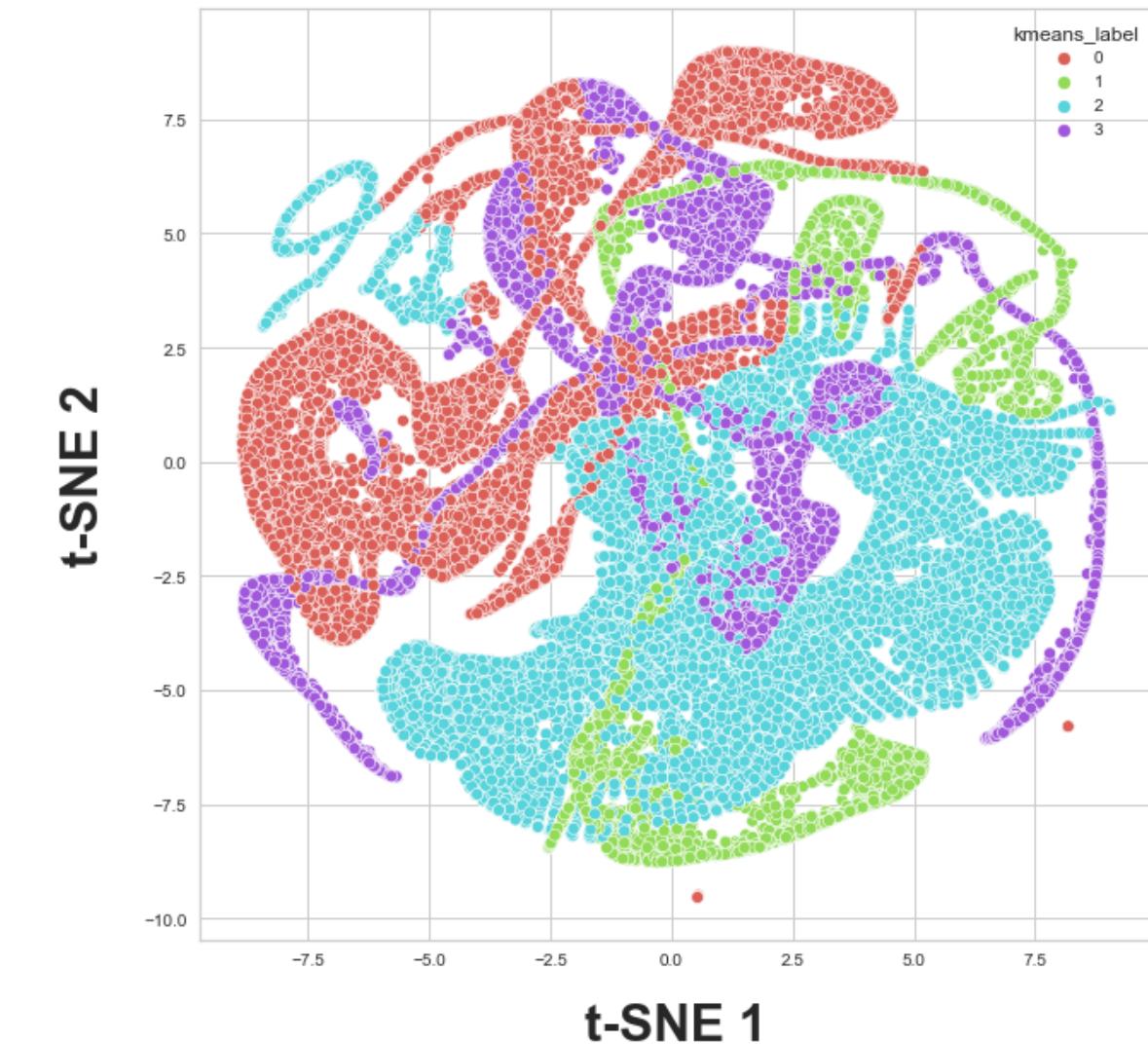
Segmentation sur l'ensemble des variables

KMeans Intercluster Distance Map (via MDS)



La séparation des clusters en utilisant l'ensemble des variables semble identique à une segmentation RFM. On a la même difficulté à séparer nettement les différents groupes.

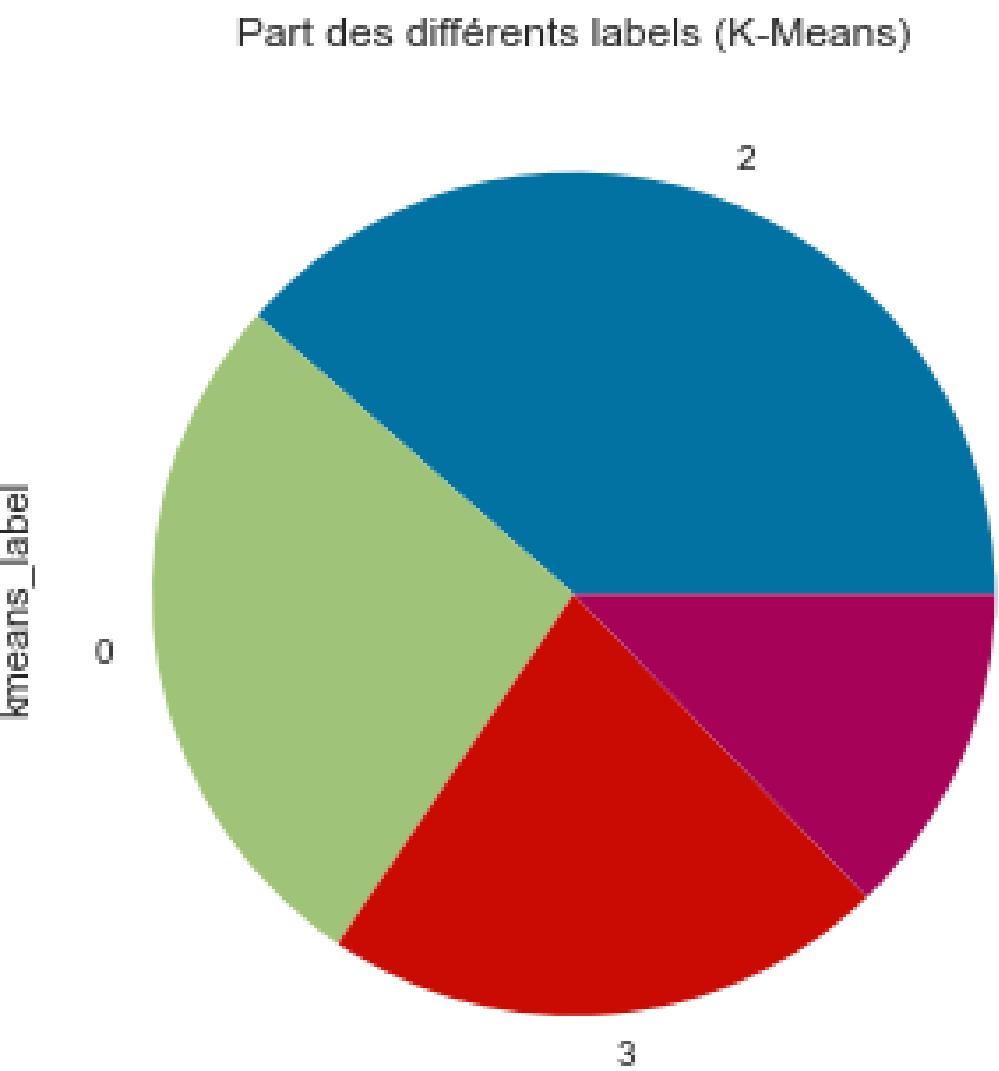
Mise en évidence des clusters t-SNE



La visualisation t-SNE semble indiquer également une mauvaise séparation des groupes (nombre d'itération égal à 300, et une complexité de 40).

K-MEANS (ALL FEATURES)

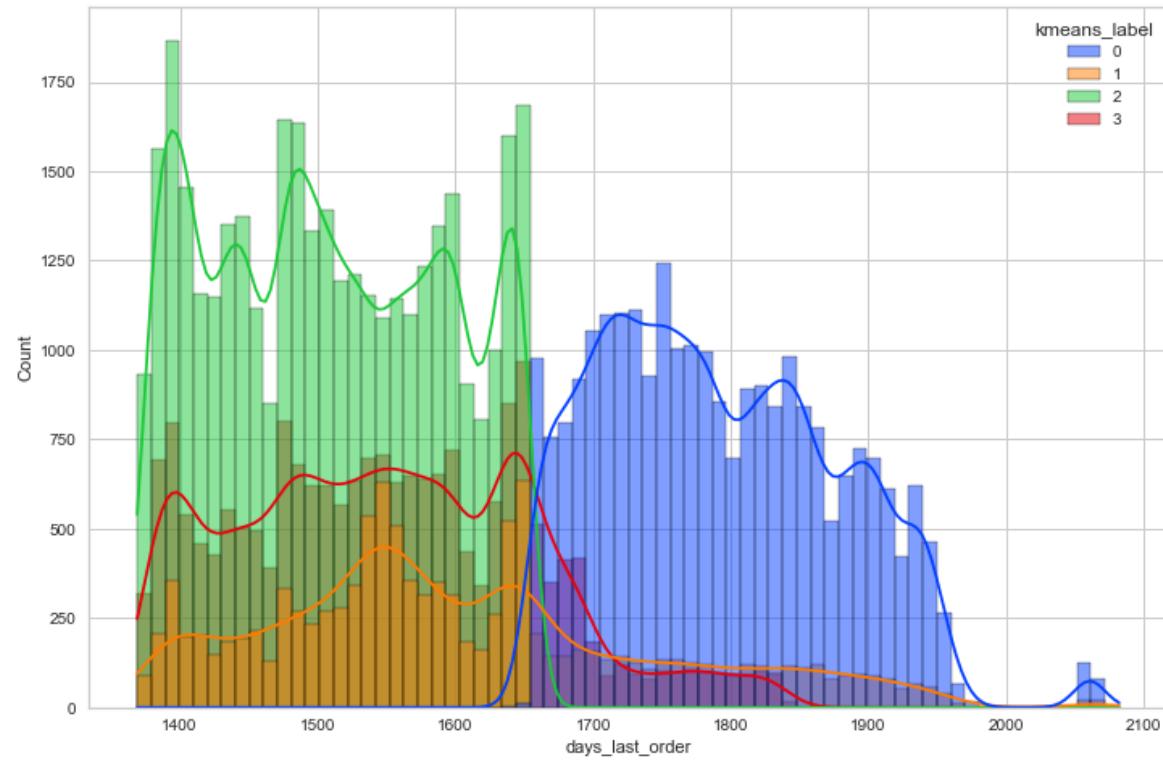
Interprétation métier des clusters



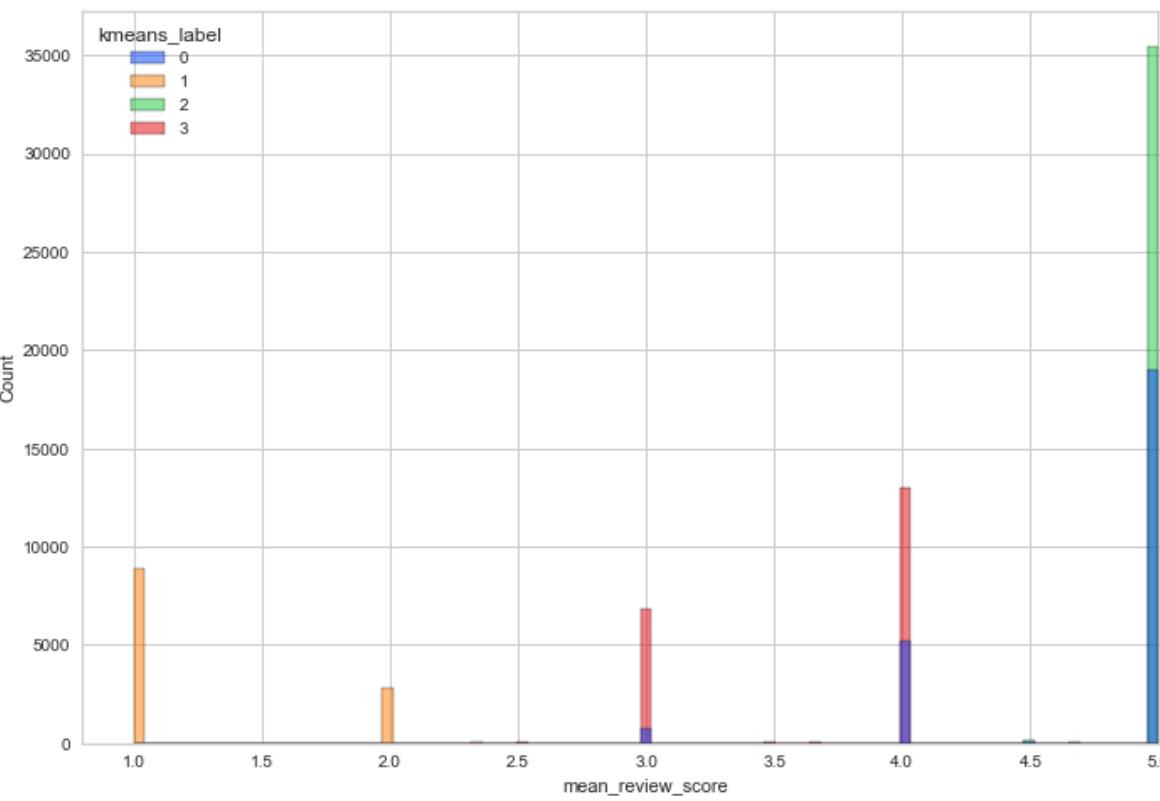
Comme indiqué par la visualisation silhouette précédemment et le pie plot ci-dessus, les groupes sont plutôt équilibré en nombre de clients.

K-MEANS (ALL FEATURES)

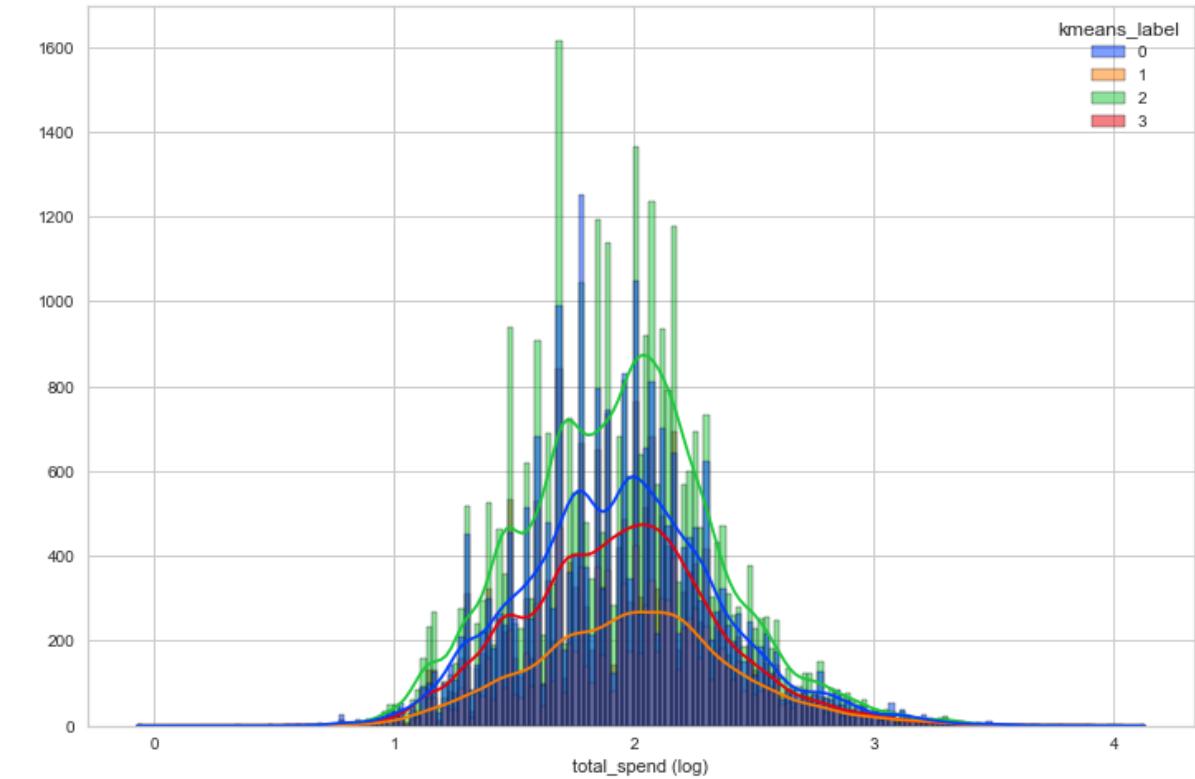
Interprétation métier des clusters



Récence



Score moyen



Montant

- La segmentation sur la récence semble moins évidente pour tous les clusters. (Cependant, label 0 bien séparé des autres labels).
- La segmentation sur le score moyen semble efficace pour certains groupes (Ex, label 1 bien séparés des autres labels).
- La segmentation sur le montant ne semble pas efficace (échelle log) sur la représentation.

K-MEANS (ALL FEATURES)

Interprétation métier des clusters

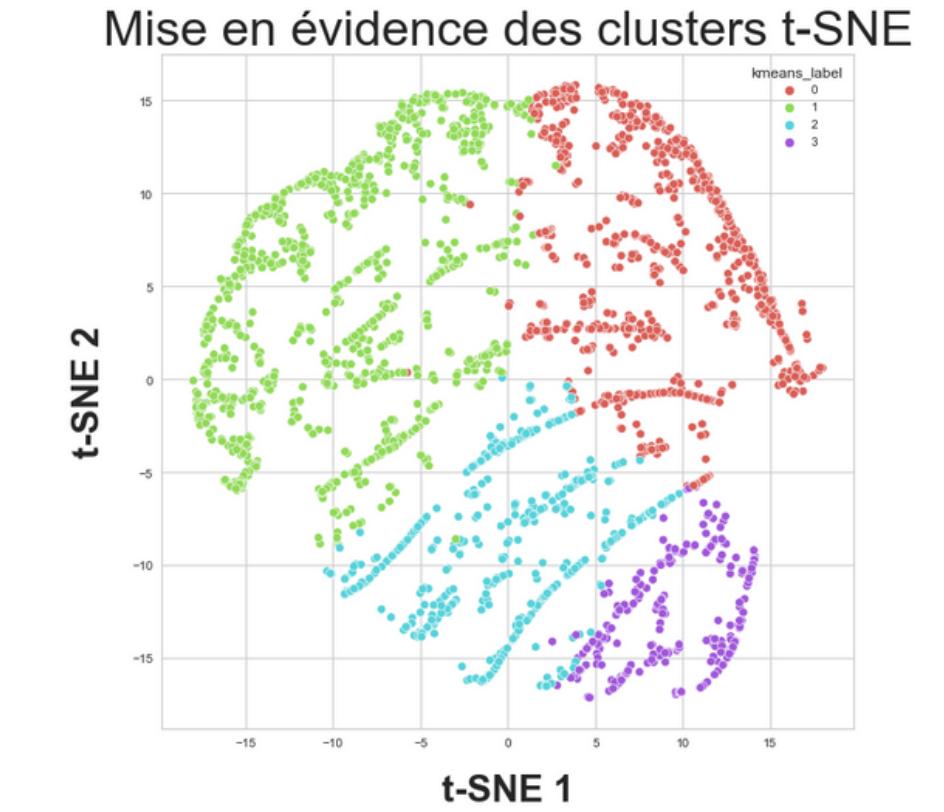
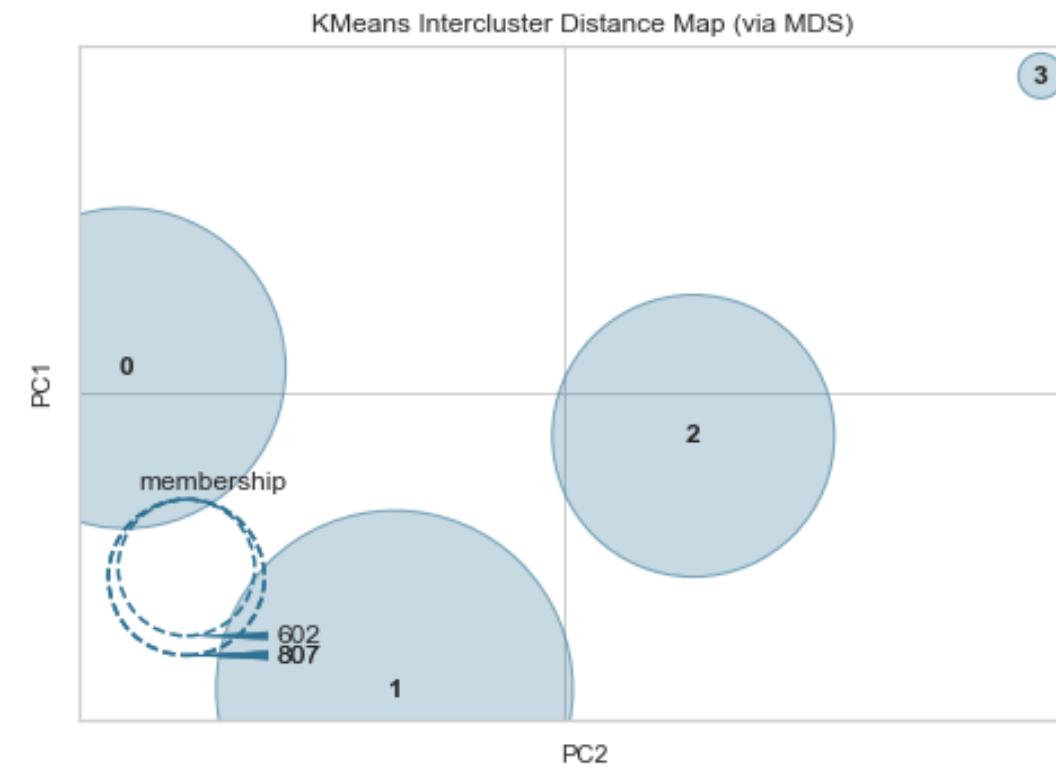
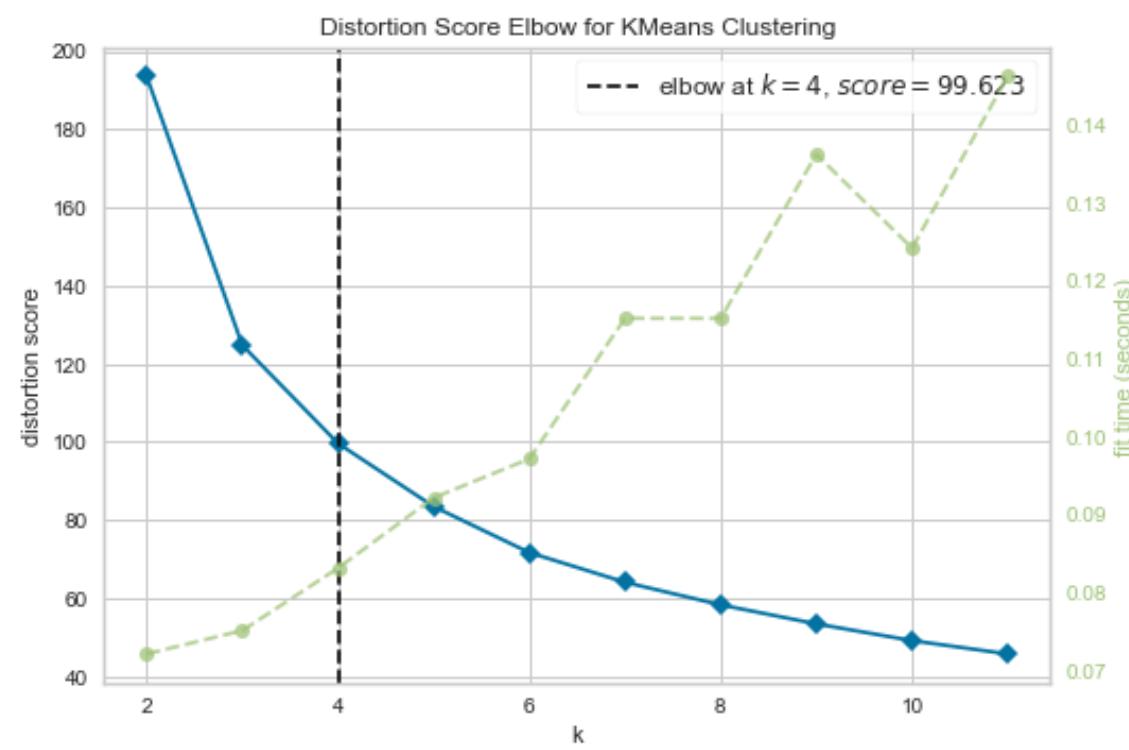
Valeur médiane par groupe de clients

kmeans_label	days_last_order	nb_orders	total_spend	mean_review_score	mean_nb_items
0	1782.0	1.0	89.9	5.0	1.0
1	1580.0	1.0	99.9	1.0	1.0
2	1504.0	1.0	89.9	5.0	1.0
3	1546.0	1.0	89.7	4.0	1.0

- Label 0 : Clients qui ont commandés il y a longtemps et qui sont très satisfait.
- Label 1: Clients insatisfaits.
- Label 2: Clients qui ont commandés récemment et qui sont très satisfait.
- Label 3: Clients qui ont commandés récemment et qui sont globalement satisfait.

SEGMENTATION SUR L'ENSEMBLE DES VARIABLES (K-MEANS)

Segmentation pour les clients qui ont passés plus d'une commande



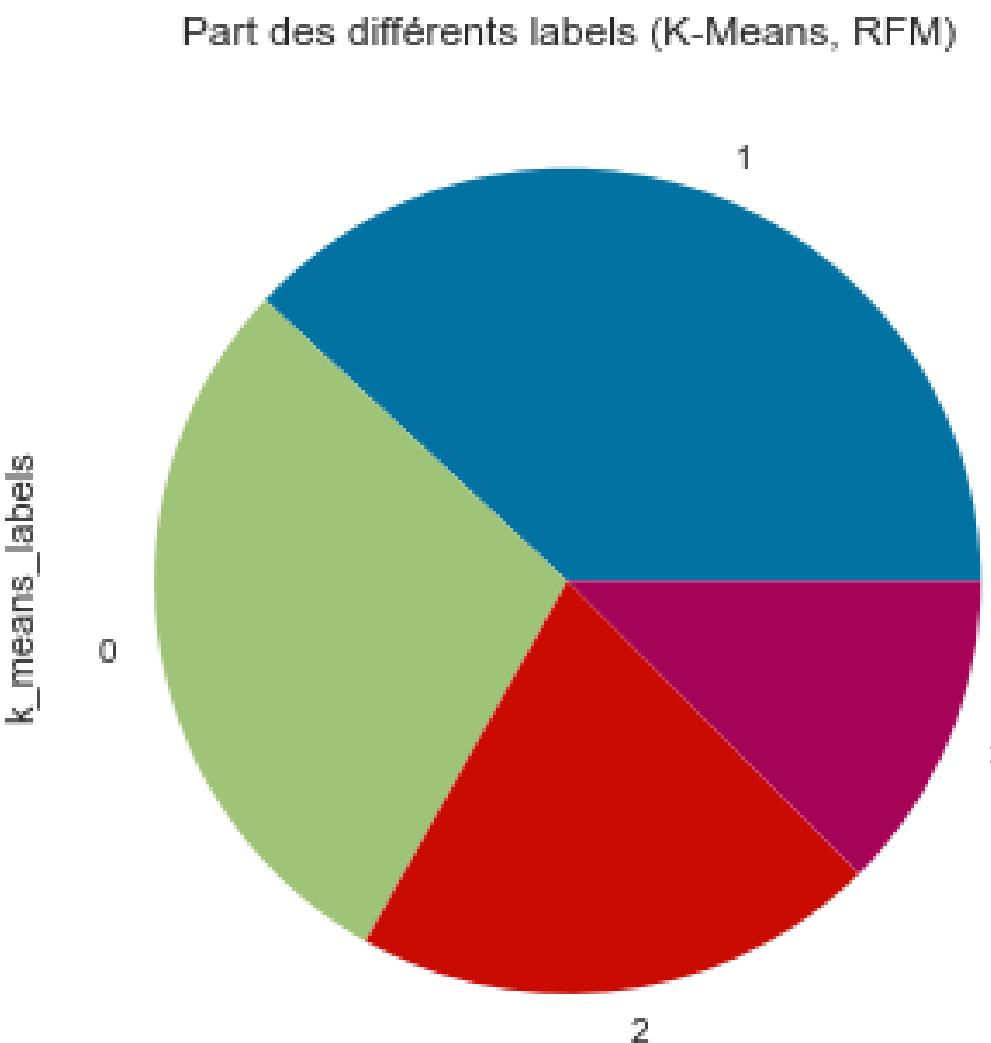
On obtient un bon clustering pour $k=4$

Les groupes semblent bien séparés.

La visualisation t-SNE semble indiquer également une bonne segmentation des clients.

K-MEANS (ALL FEATURES)

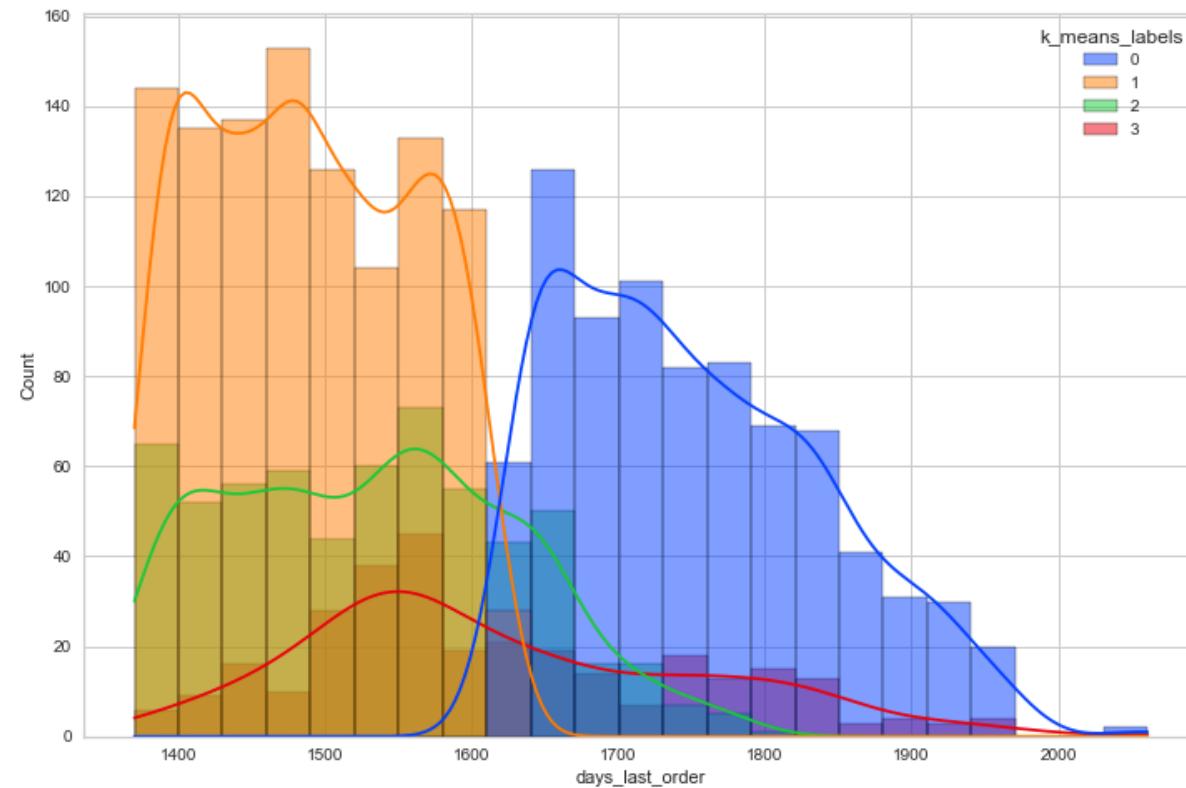
Interprétation métier des clusters : clients avec plus d'une commande



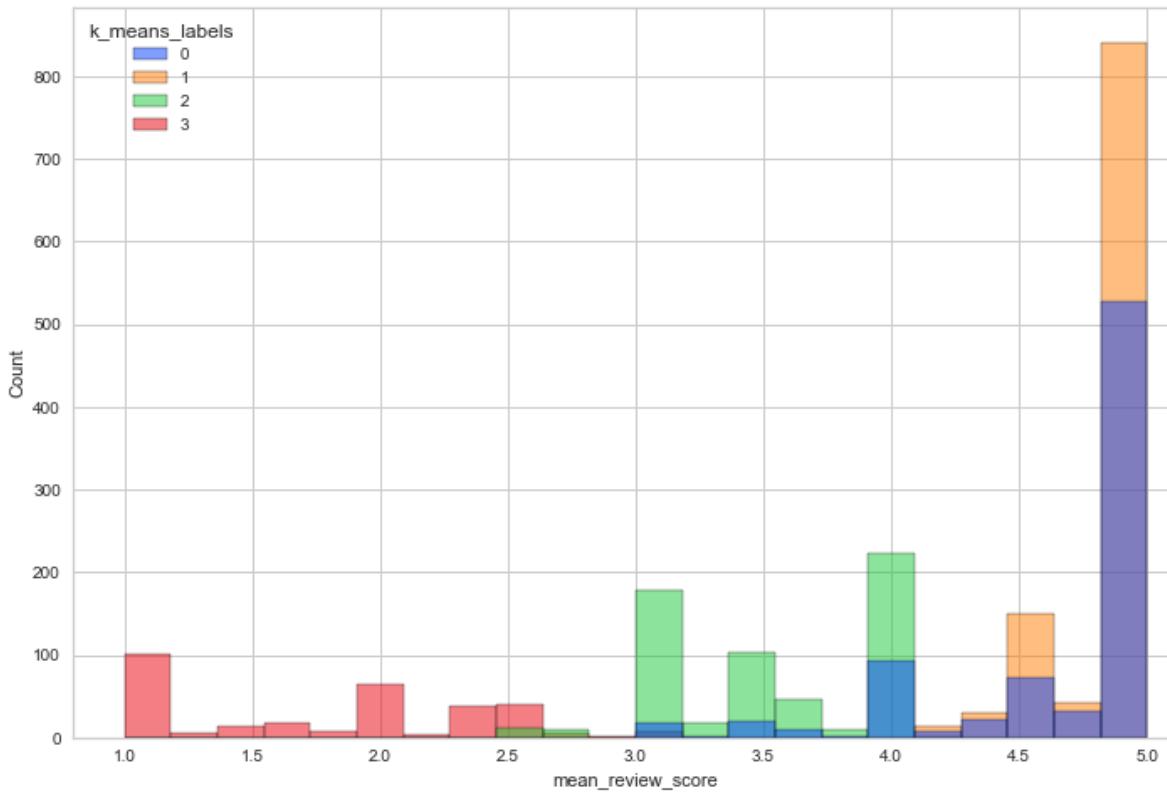
Les groupes sont plutôt équilibré en nombre de clients.

K-MEANS (ALL FEATURES)

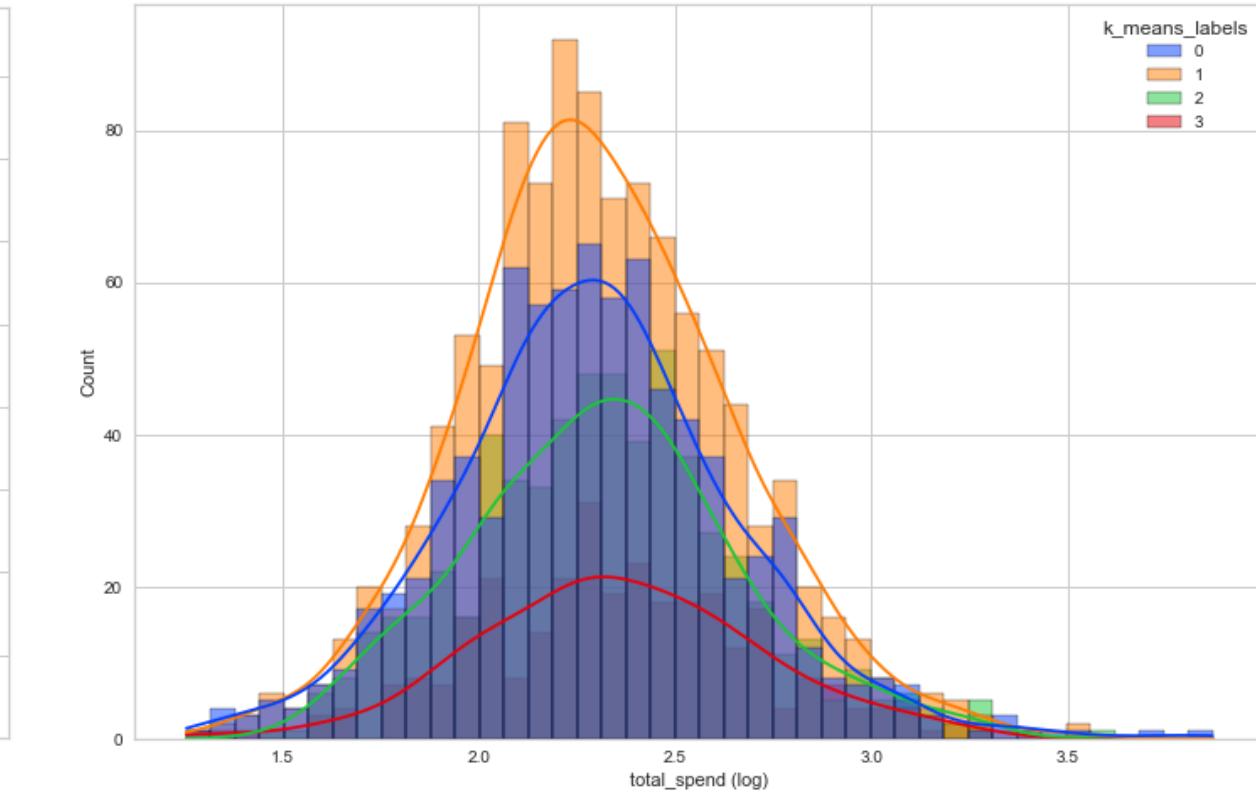
Interprétation métier des clusters : clients avec plus d'une commande



Récence



Score moyen



Montant

- La segmentation sur la récence semble moins évidente pour tous les clusters. (label 0 bien séparé des autres label)
- La segmentation sur le score moyen semble efficace pour certains groupes (label 3 bien isolé).
- La segmentation sur le montant semble être légèrement amélioré (échelle log) sur la représentation.

K-MEANS (ALL FEATURES)

Interprétation métier des clusters : clients avec plus d'une commande

Valeur médiane par groupe de clients

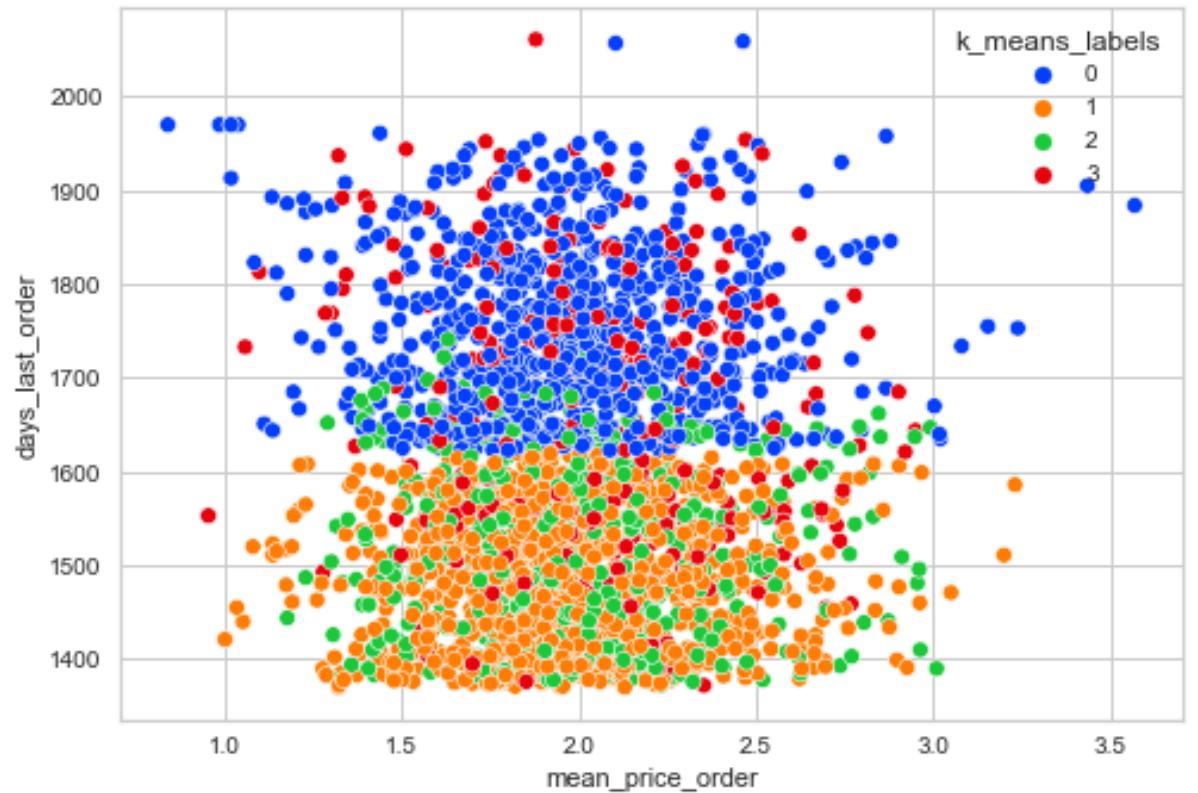
k_means_labels	days_last_order	nb_orders	total_spend	mean_review_score	mean_nb_items
0	1737.0	2.0	194.890	5.0	1.0
1	1485.0	2.0	191.990	5.0	1.0
2	1538.5	2.0	208.560	3.5	1.0
3	1584.5	2.0	223.345	2.0	1.0

- Label 0 : Clients qui ont commandés il y a longtemps mais qui sont très satisfait.
- Label 1: Clients qui ont commandés récemment et qui sont très satisfait mais qui dépensent moins.
- Label 2: Clients qui ont commandés assez récemment mais qui sont moyennement satisfait.
- Label 3: Clients qui ont commandés assez récemment et qui dépensent le plus mais qui sont peu satisfait.

K-MEANS (ALL FEATURES)

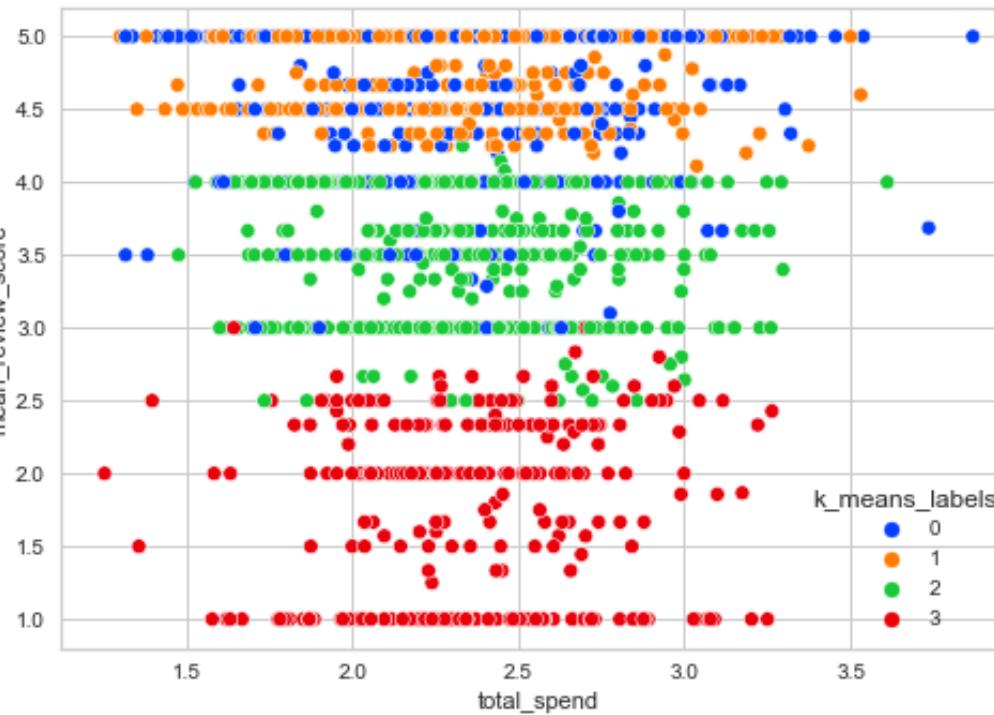
Interprétation métier des clusters : clients avec plus d'une commande

Scatter plots



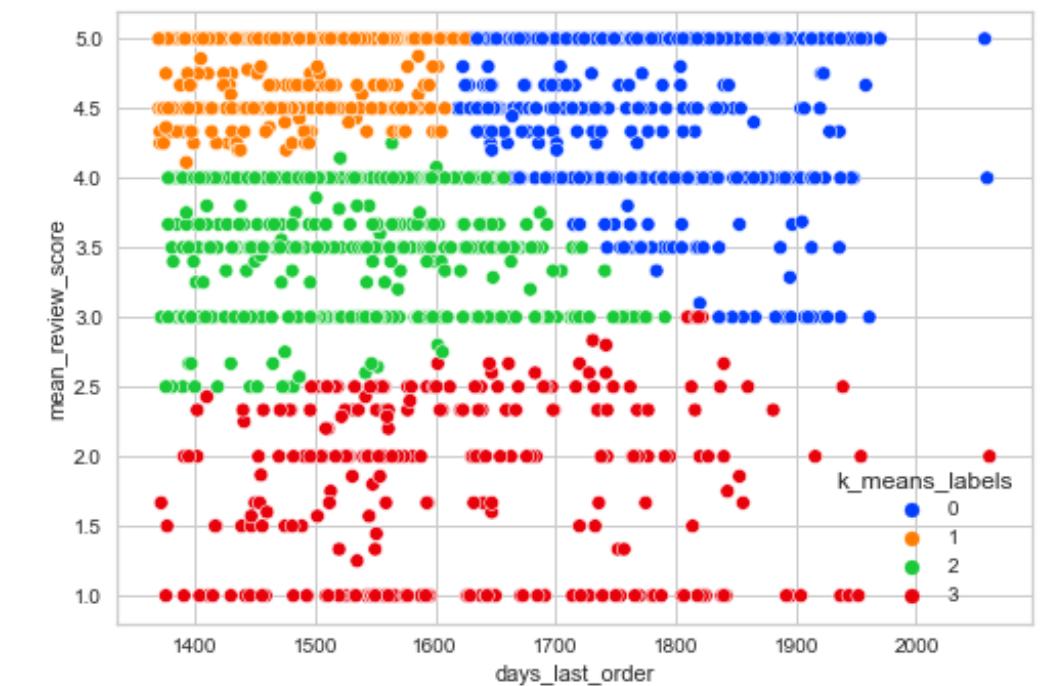
Récence vs Montant (log)

Comme vu précédemment la segmentation selon le montant n'est pas efficace. Cependant, la segmentation selon la récence pour les labels 0 et 1 donne de meilleurs résultats.



Review score vs Montant (log)

Comme vu précédemment la segmentation selon le review score est bon pour isoler les clients du label 3, qui sont plutôt insatisfaits.



Review score vs Récence

Distance intercluster proche mais la séparation est effective visuellement entre les clusters pour les deux variables.



Délai de maintenance du modèle

MISE A JOUR DU MODÈLE

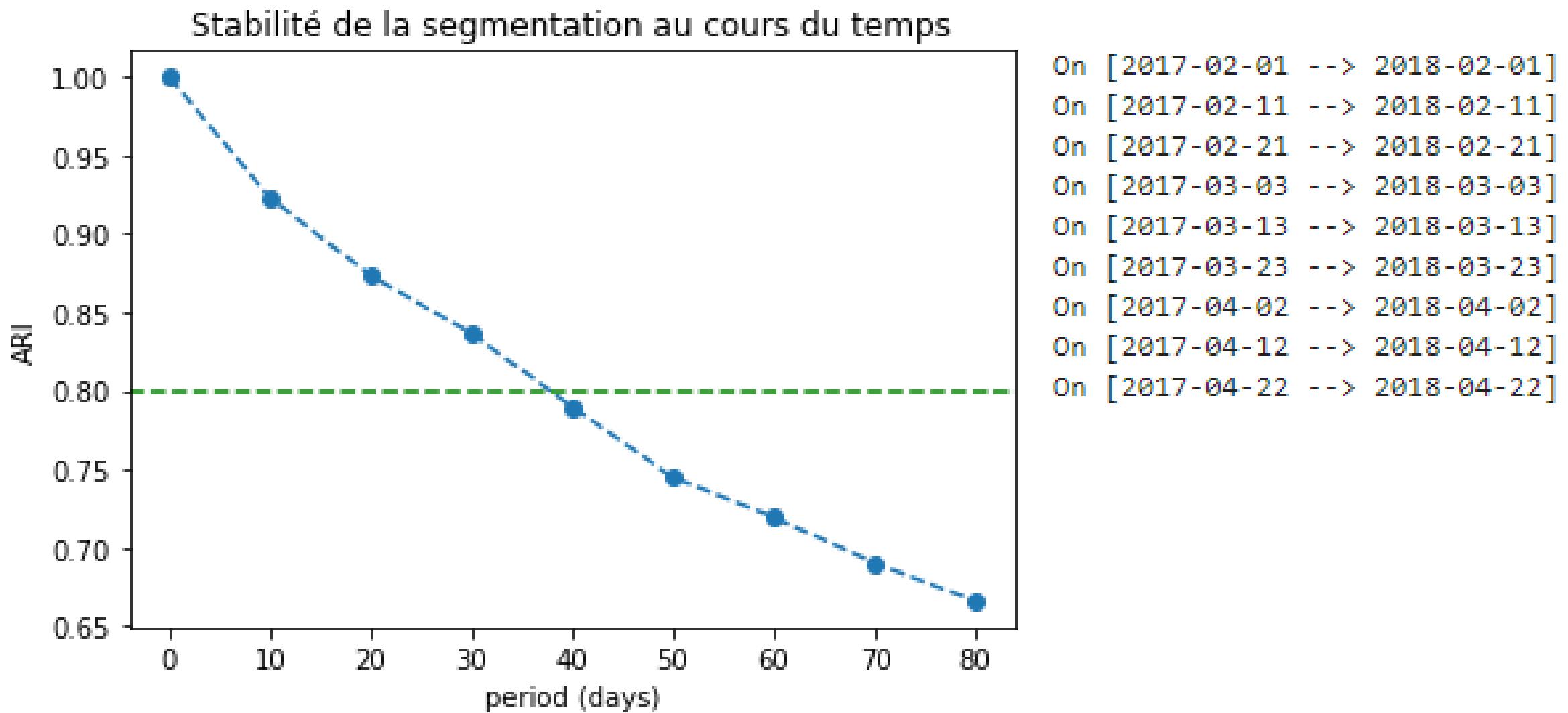
Nous devons vérifier à partir de quel moment les clients changent de segment. Dans le but d'établir un contrat de maintenant du programme de segmentation, nous devons tester sa stabilité dans le temps.

Méthode :

1. Segmentation de référence sur une période d'un an avec date de départ = t0 et avec un scaler initial.
2. On boucle en réalisant les opérations suivantes :
 - Segmentation sur une période d'un an avec date de départ décalé ($t0 + \text{delta_days}$) avec le scaler initial, réalisation des prédictions.
 - Création d'un nouveau scaler sur la nouvelle période
 - Segmentation avec le nouveau scaler et prédiction
 - Comparaison des prédictions grâce au calcul de l'indice de rang ajusté (ARI).

MISE A JOUR DU MODÈLE

Pour une période d'un an avec offset de 10 jours :



On constate qu'à partir d'environ 38 jours, le score ARI devient inférieur à 0.8. On pourra donc prévoir la maintenance du programme de segmentation tous les 30 jours pour garantir un score ARI > 0.8.

CONCLUSION

La segmentation RFM est efficace uniquement pour créer des groupes de clients par récence d'achat. Le clustering par dépense n'est pas concluant.

La majorité des clients ne commandent qu'une seule fois, ce qui rend difficile d'établir une segmentation par fréquence d'achat.

La segmentation sur l'ensemble des variables permet d'améliorer la segmentation et d'identifier des groupes d'individus selon la note de satisfaction et la récence d'achat.

Une mise à jour du modèle est nécessaire tous les 30 jours.

olist