Parcours Ingénieur Machine Learning

Session Mars 2021

OPENCLASSROOMS

Projet 4
Segmentez des clients d'un site e-commerce

25/07/2021

Etudiante : QITOUT Kenza

Mentor : Maïeul Lombard

Evaluateur : Zied Jemai

CONTEXTE DU PROJET

olist site de e-commerce brésilien

Connecte les entreprises au Brésil qui peuvent vendre leurs produits sur le site www.olist.com et les livrer aux clients

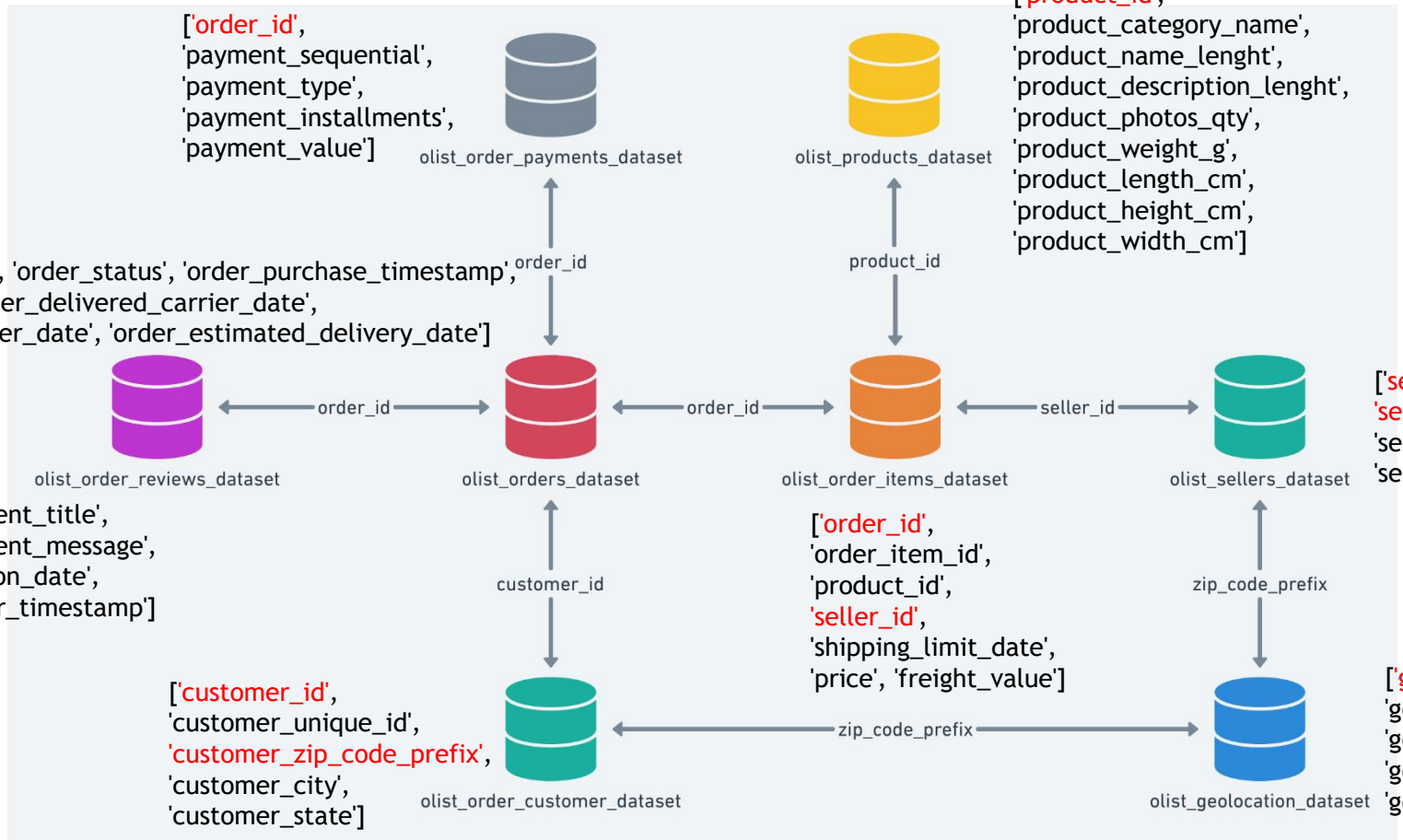
Problématique :

Se place en tant que consultant pour Olist => Aider l'équipe marketing en leur fournissant les profils des clients

Objectifs : Comprendre les différents types d'utilisateurs en réalisant une segmentation des clients facilement exploitable et compréhensible

BASES DE DONNEES

9 DataFrames disponibles sur https://www.kaggle.com/olistbr/brazilian-ecommerce, composés de données anonymisées :



PISTES DE RECHERCHE

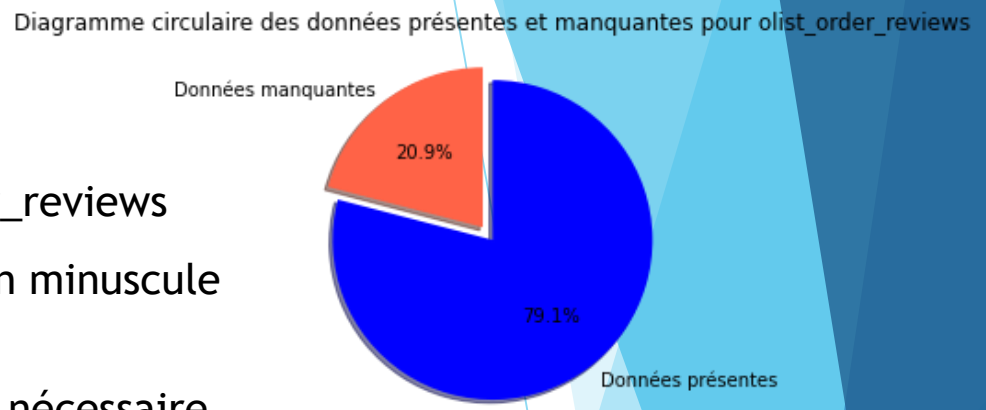
Missions :

- Réaliser une courte analyse exploratoire après avoir nettoyé les jeux de données
- Tester différentes approches de modélisation
- Evaluer la fréquence de mise à jour de la segmentation

Méthodologie :

- Modèle de base = segmentation RFM
- Construire d'autres modèles qui apportent plus d'informations à l'équipe marketing
- Clustering des clients via des méthodes non supervisées pour regrouper les clients de profils similaires

1. NETTOYAGE ET EXPLORATION



Nettoyage des différents DataFrames séparément :

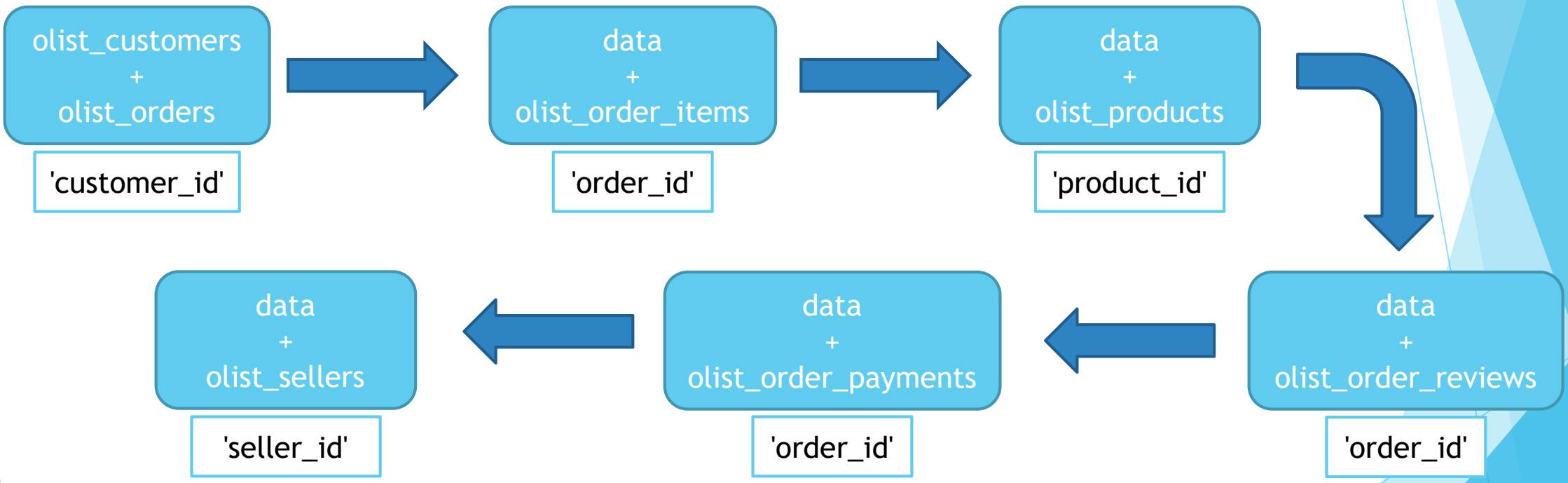
- 0% à 0.8% de NaN pour tous les DataFrames sauf pour olist_order_reviews
- Vérification de l'orthographe et transformation des catégories en minuscule pour les variables qualitatives
- Vérification des Outliers (avec les boxplot) => pas de traitement nécessaire

olist_customers	olist_geolocation	olist_order_items	olist_order_payments	olist_order_reviews	olist_orders	olist_products	olist_sellers
99441 * 5 Vérification des informations sur la ville et l'état des clients avec le DataFrame olist_geolocation avec une jointure via le ZipCode => 0 ligne supprimée	1000163 * 5 Agrégation des moyennes de 'geolocation_lat' et 'geolocation_lng'	112650 * 7 Création de la variable 'price_freight_sum' arrondie à 10^{-3}	103886 * 5 Création de la variable 'payment_type_credit_card'	100000 * 7 Création de la variable 'temps_reponse_enquete'	99441 * 8 Création des variables 'durée_livraison', et 'statut_livraison' Imputation des lignes sans duree_livraison	32951 * 9 Création de la variable 'volume_produit' Regroupement catégories en anglais avec le DataFrame product_category_name_translation	3095 * 4

Vérification de l'égalité au seuil de 10.0 entre la somme de 'freight_value' et 'price' et 'payment_value' arrondie à 10^{-3} avec une jointure via 'order_id' => Suppression de 122 lignes

1. NETTOYAGE ET EXPLORATION

Jointures internes des variables utilisées pour la segmentation :



Suppression des 7043 lignes contenant des données manquantes

DataFrame de 111 150 lignes et 27 colonnes des articles vendus pour chaque commande pour chaque client

1. NETTOYAGE ET EXPLORATION

Agrégations des informations pour chaque client :

- Groupby de 'price', 'freight_value', 'product_photos_qty', 'product_weight_g', 'volume_produit' par 'order_id'
- Puis Groupby de toutes les variables par **'customer_unique_id'**

Dans l'ordre croissant par 'order_purchase_timestamp' et en supprimant les doublons en fonction de la variable

Agrégation des variables quantitatives par la somme ou par la moyenne

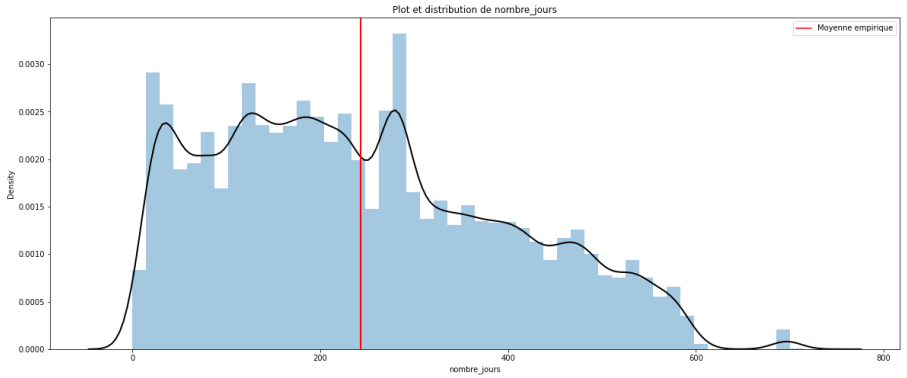
Agrégation des variables qualitatives par le mode (en cas d'ambiguïté, la catégorie la plus récente a été privilégiée)

→ DataFrame final de 91 701 lignes et de 30 colonnes

1. NETTOYAGE ET EXPLORATION

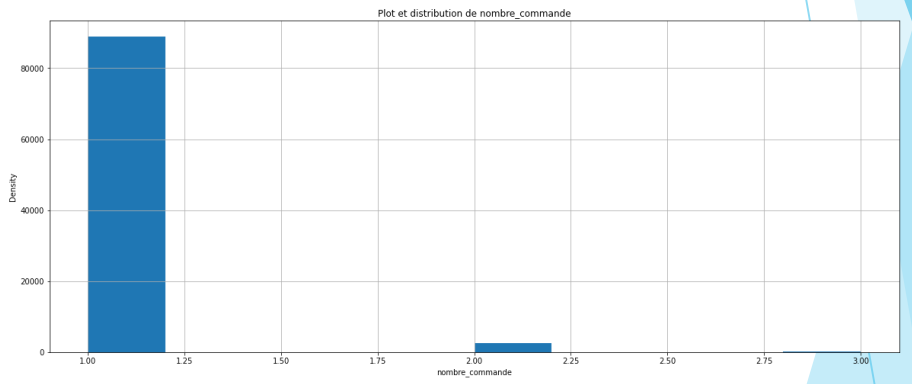
Indicateurs de la segmentation RFM :

- Récence = durée depuis le dernier achat



Moyenne = 243.05

- Fréquence = nombre de commandes



Moyenne = 1.03

- Montant = montant moyen des commandes

Moyenne = 141.81

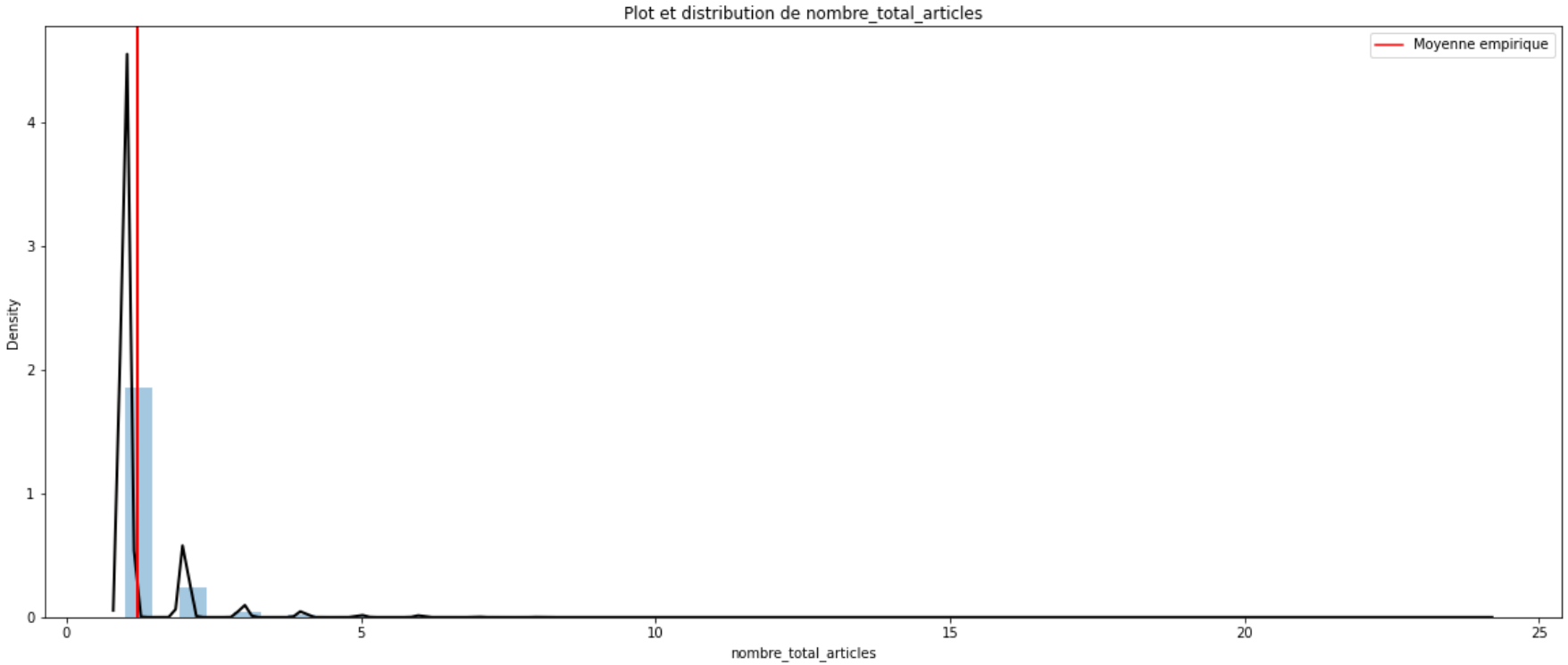
1. NETTOYAGE ET EXPLORATION

Autres indicateurs pour améliorer la segmentation :

- Nombre total d'articles
- Moyenne du nombre d’articles par commande
- Statut de la livraison de la dernière commande (‘Livré’ ou ‘Non livré’)
- Moyenne de la proportion des frais de transport
- Moyenne de la note du client et de la durée de réponse à l'enquête de satisfaction
- Moyenne de la durée de livraison
- Moyenne du poids des articles
- Moyenne du volume des articles
- Moyenne du nombre de photos par article
- Localisation la plus fréquente du vendeur par rapport au client (‘Local’ ou ‘Non local’)
- Moyenne de la distance entre le client et le vendeur
- Paiement par carte de crédit le plus fréquent ('credit_card', 'autre')
- Moyenne du nombre de versements par article
- Catégorie de l’article la plus fréquente

1. NETTOYAGE ET EXPLORATION

Analyses univariées : 'nombre_total_articles'

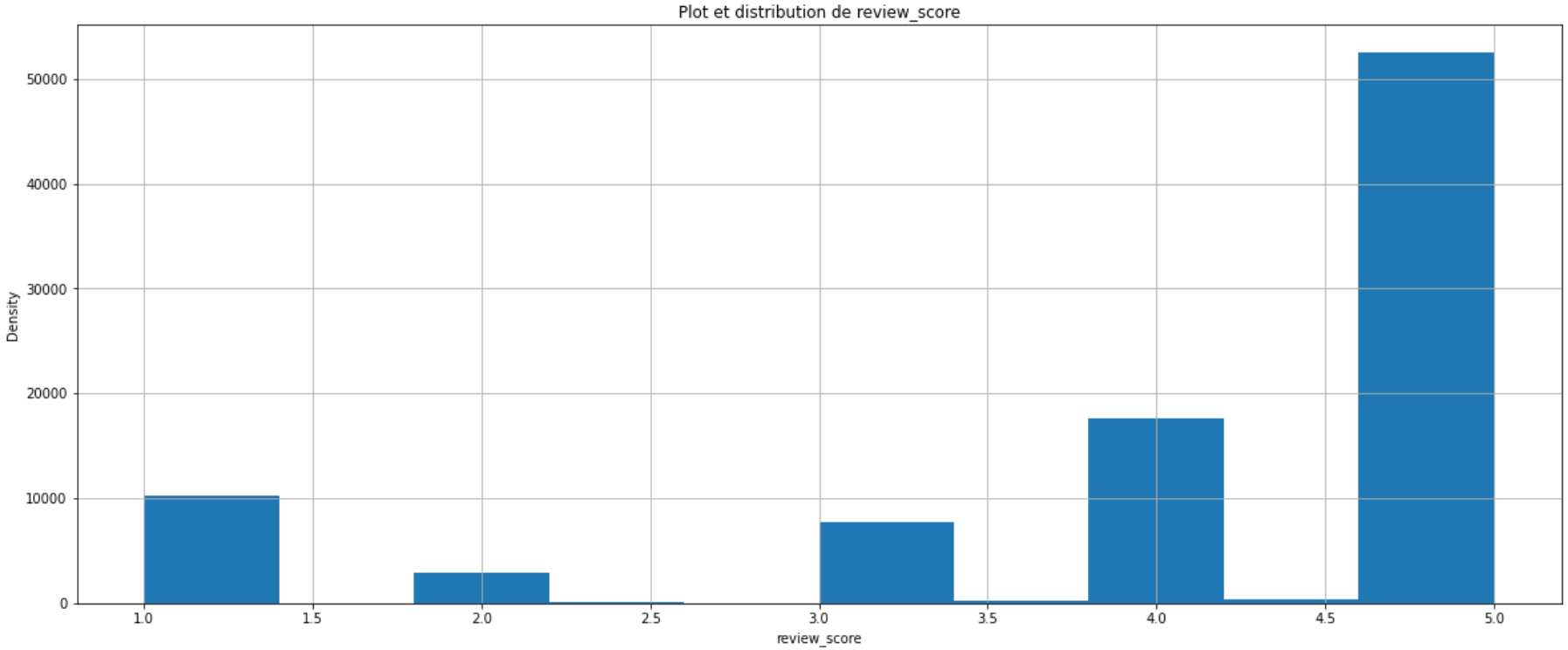


Moyenne : 1.21

Ecart-type : 0.46

1. NETTOYAGE ET EXPLORATION

Analyses univariées : 'review_score'

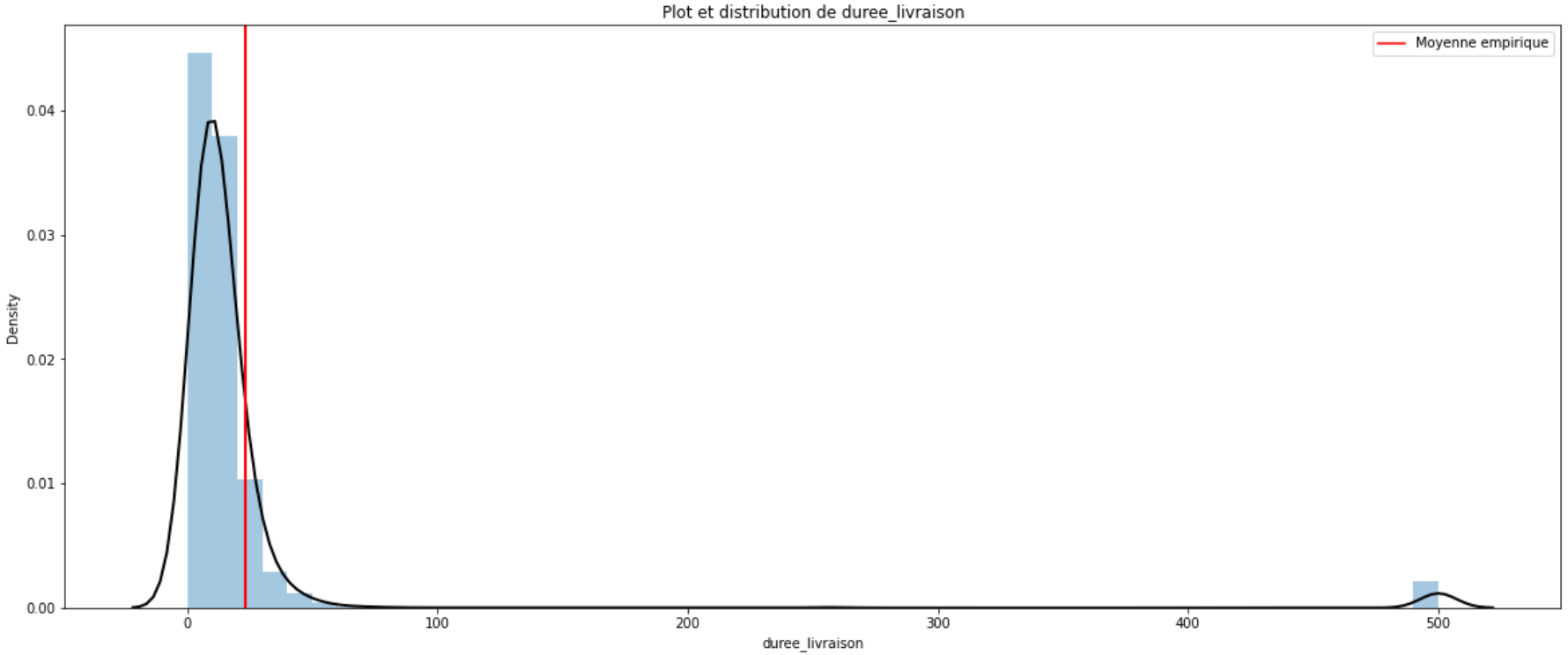


Moyenne : 4.08

Ecart-type : 1.78

1. NETTOYAGE ET EXPLORATION

Analyses univariées : 'duree_livraison'

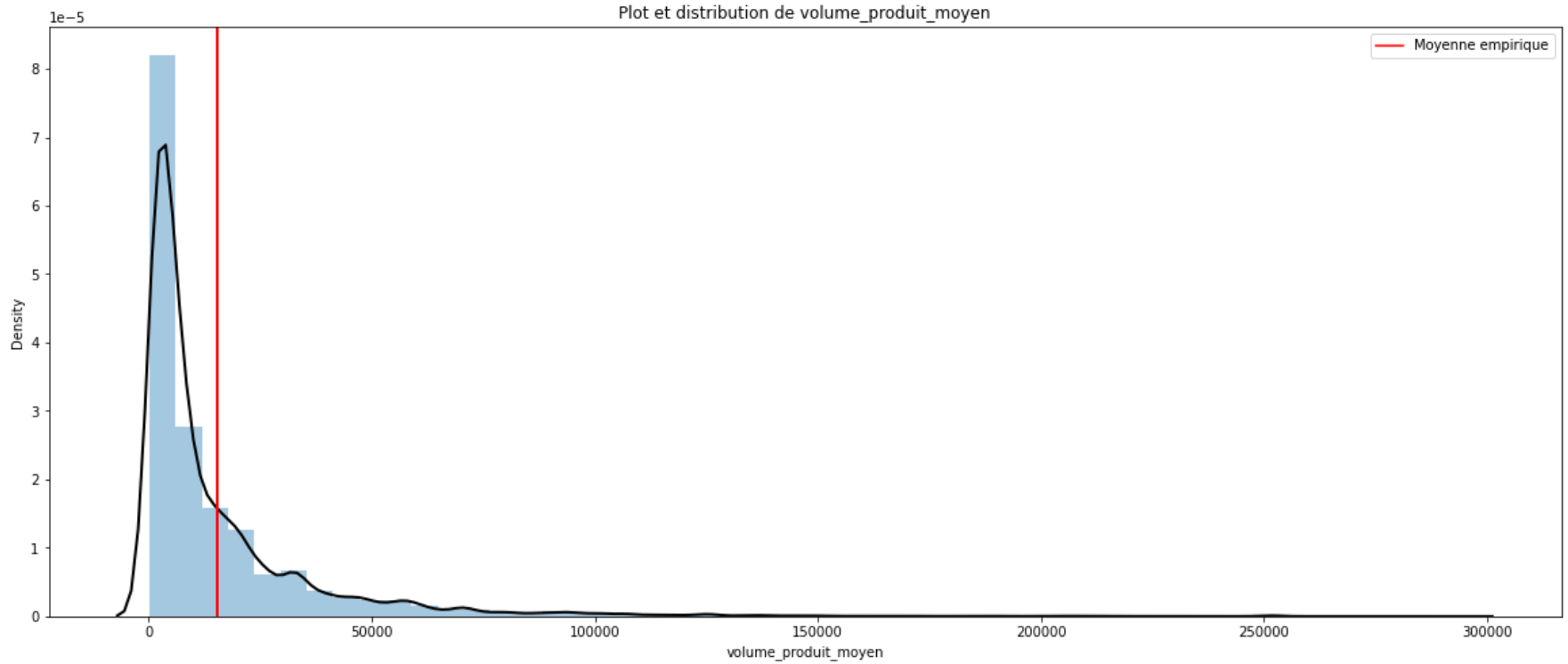


Moyenne : 23.08

Ecart-type : 5069.30

1. NETTOYAGE ET EXPLORATION

Analyses univariées : 'volume_produit_moyen'

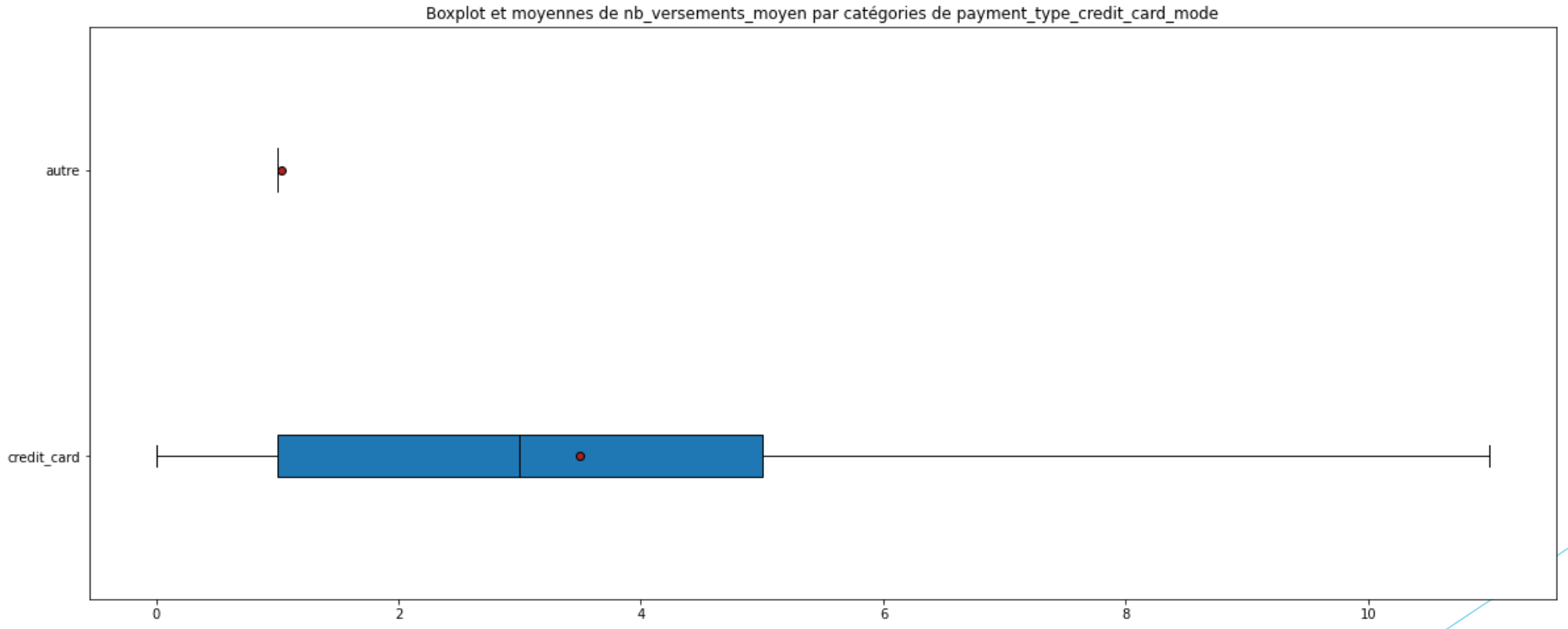


Moyenne : 15347.07

Ecart-type : 548609783.25

1. NETTOYAGE ET EXPLORATION

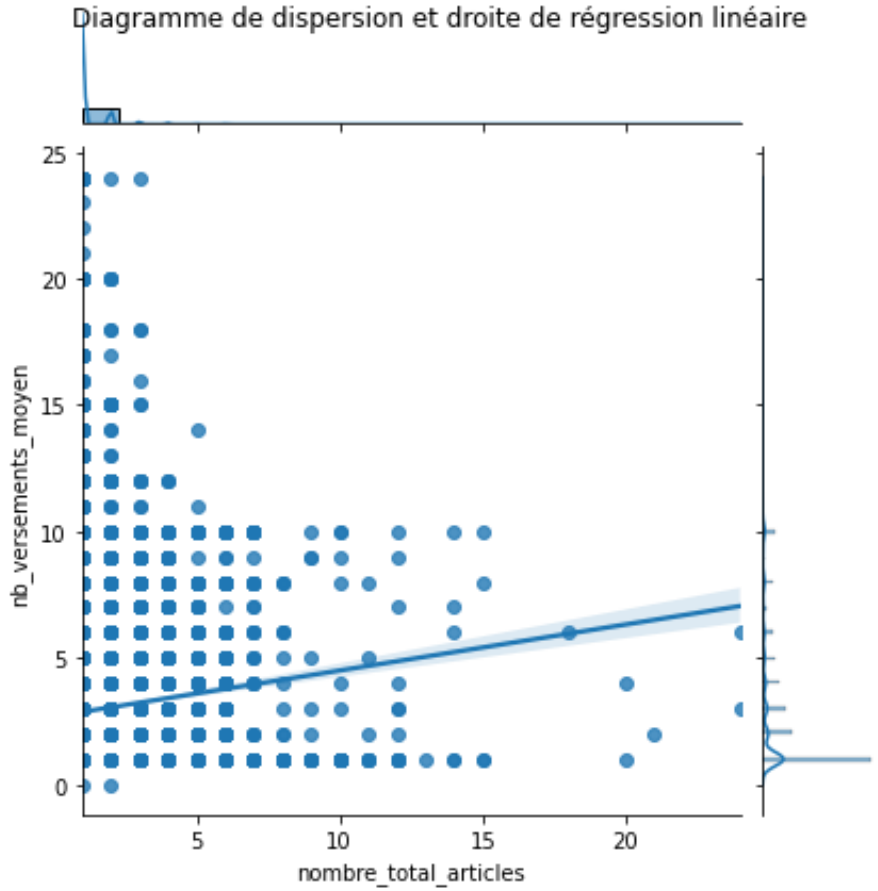
Analyses bivariées : Relation entre 'nb_versements_moyen' et 'payment_type_credit_card_mode'



ANOVA : $n^2 = 0.15$

1. NETTOYAGE ET EXPLORATION

Analyses bivariées : Relation entre 'moyenne_prix_commande' et 'volume_produit_moyen'



Corrélation de Pearson = 0.29

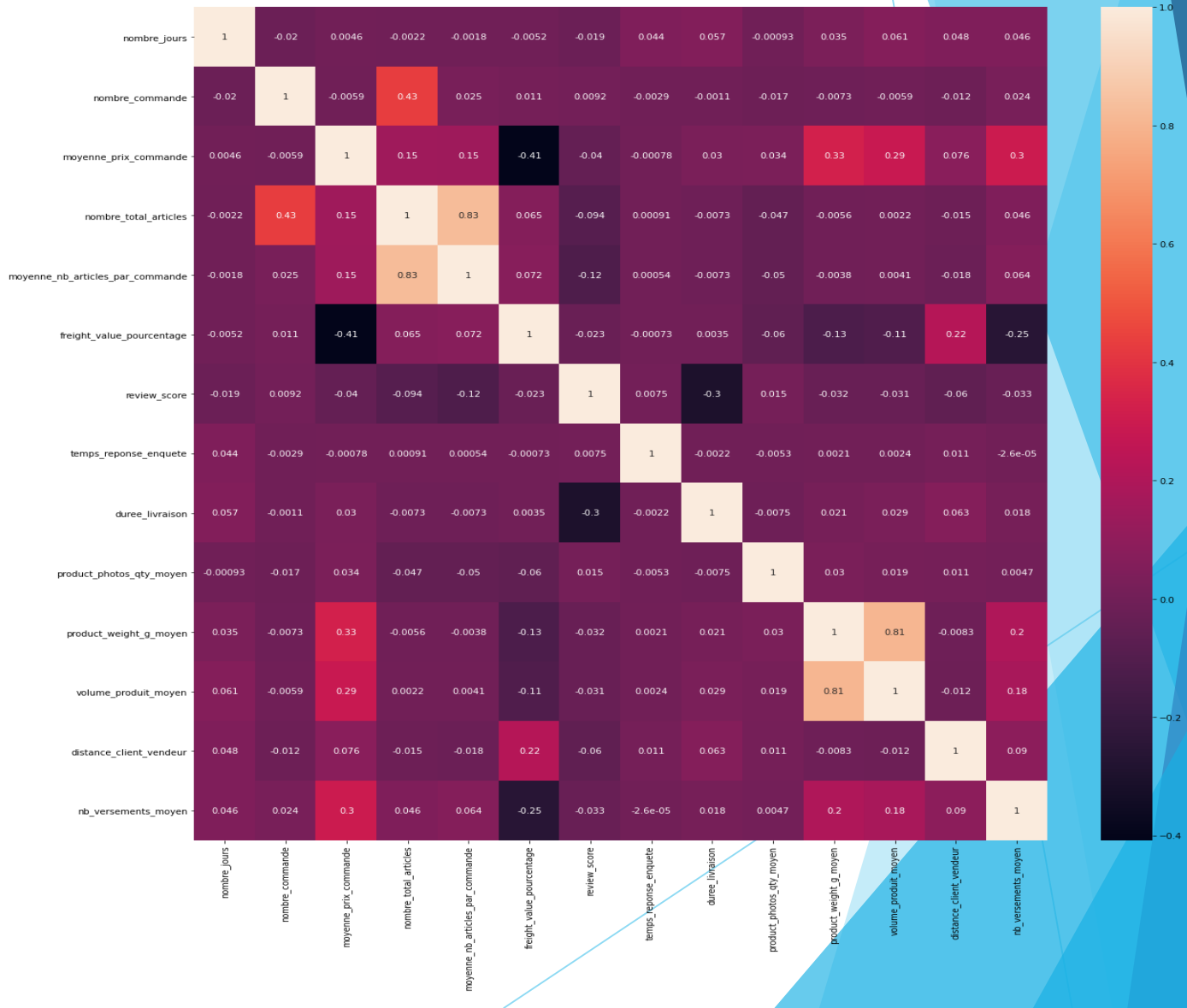
1. NETTOYAGE ET EXPLORATION

Vérification de la corrélation entre les variables :

Corrélation entre les variables utilisées dans la segmentation < |0.9|

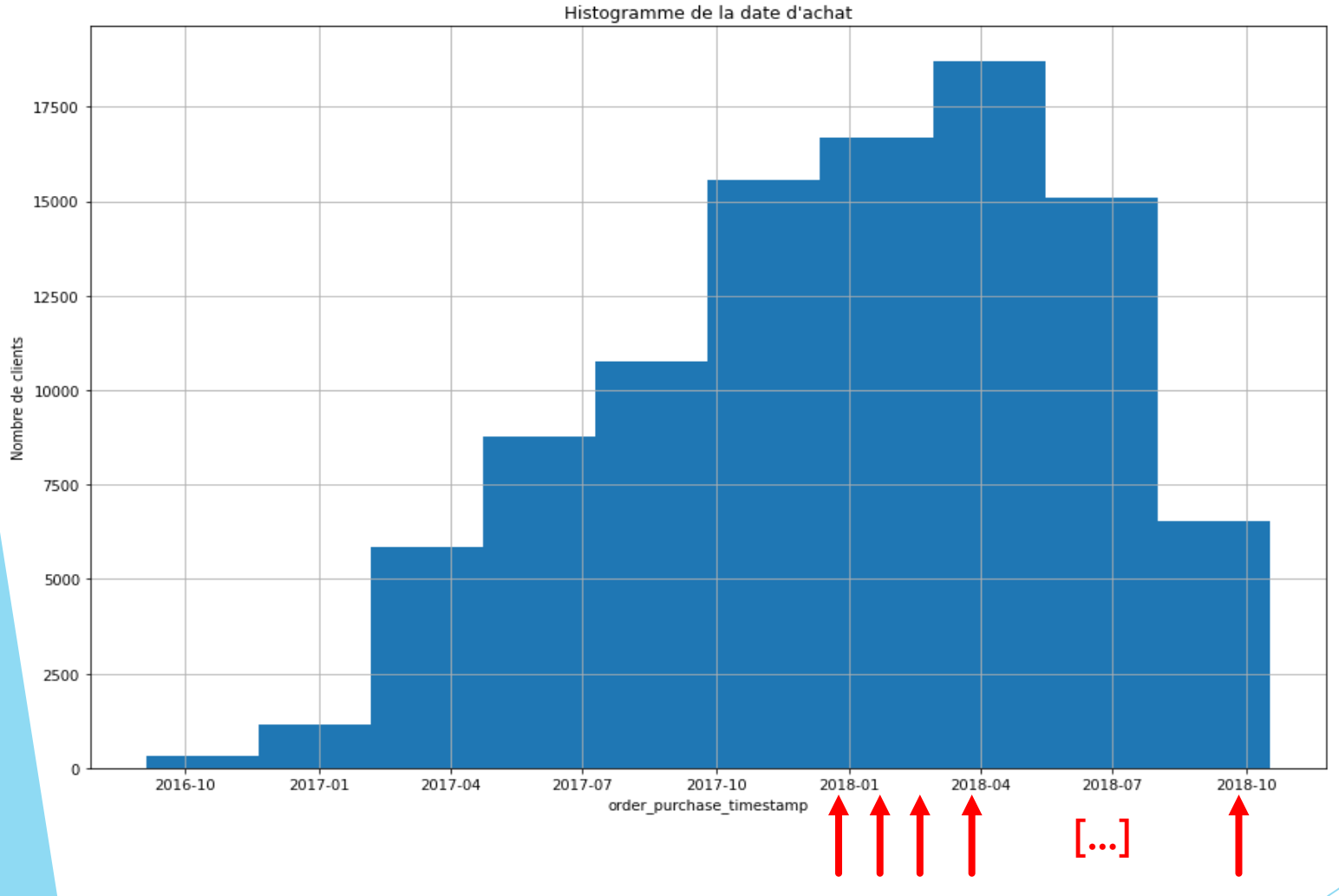
Suppression de 9 variables

DataFrame final de 91 701 lignes et 21 variables



2. MODELISATION

Méthode utilisée pour réaliser les analyses sur les différents modèles et la stabilité du modèle :



1e segmentation = utiliser les données de l'année 2017

2e segmentation = ajouter les nouveaux clients apparus X mois après 2017

Utiliser le même modèle que celui de l'année 2017

Comparaison du numéro du cluster attribué aux clients de 2017 avec la 2e segmentation

A partir de quelle fréquence, les nouveaux clients déstabilisent-ils la segmentation des clients ?

2. MODELISATION

Liste des modèles testés :

- KMeans (n_clusters entre 1 et 15) → Inertie et Coefficient de silhouette
- DBSCAN (x_eps entre 0.1 et 3.0, min_samples entre 2 et 10, pas = 2) → Coefficient de silhouette
- KModes (n_clusters entre 1 et 15) → Cost et Coefficient de silhouette
- KPrototypes (n_clusters = [2, 5, 8, 10, 13, 15]) → Cost

Recherche de l'hyperparamètre optimal

Comparaison des modèles entre eux :

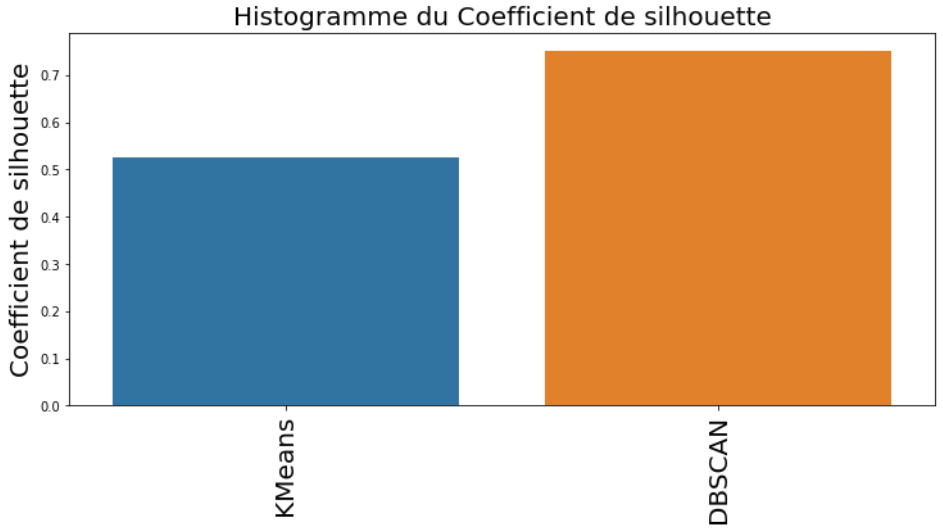
- Coefficient de silhouette (metrics.silhouette_score)
- Indice de Calinski-Harabasz (metrics.calinski_harabasz_score)
- Temps d'éxecution

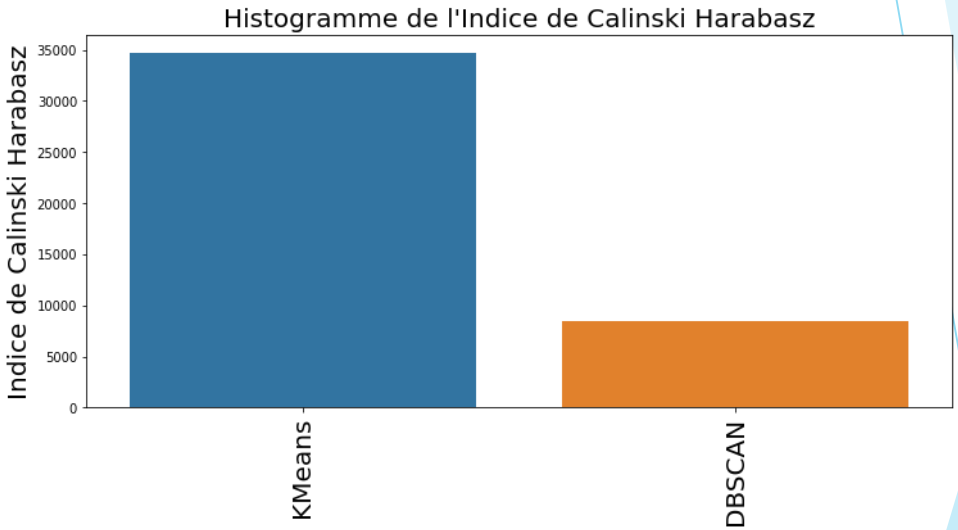
Etude de la stabilité des segments au cours du temps :

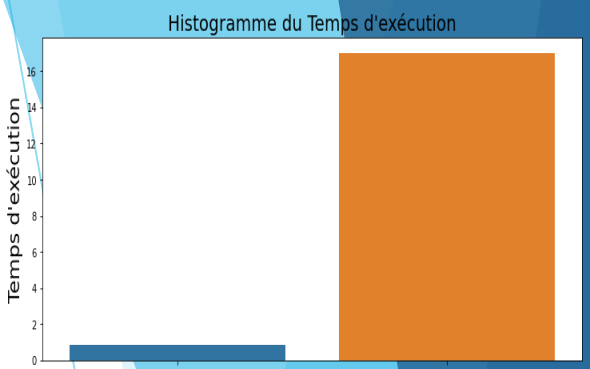
- Indice de Rand entre les clusters de 2017 et les clusters de X mois après (metrics.adjusted_rand_score)

3. Modèle optimal

Segmentation RFM : Comparaison des modèles



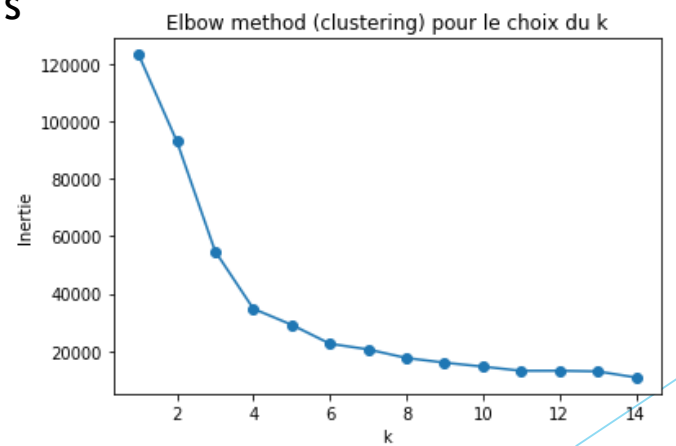


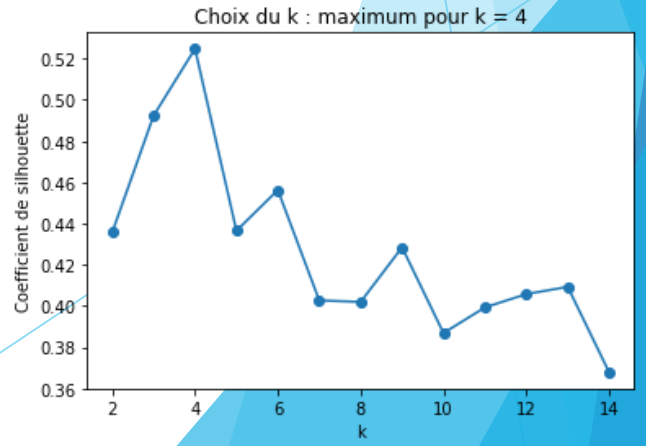


Sélectionne le KMeans avec un coefficient de silhouette à 0.52 et un indice de Calinski Harabasz à 34764, avec un temps d'exécution de 0.81 s

Méthode du coude :

➡ choix du n_clusters = 4

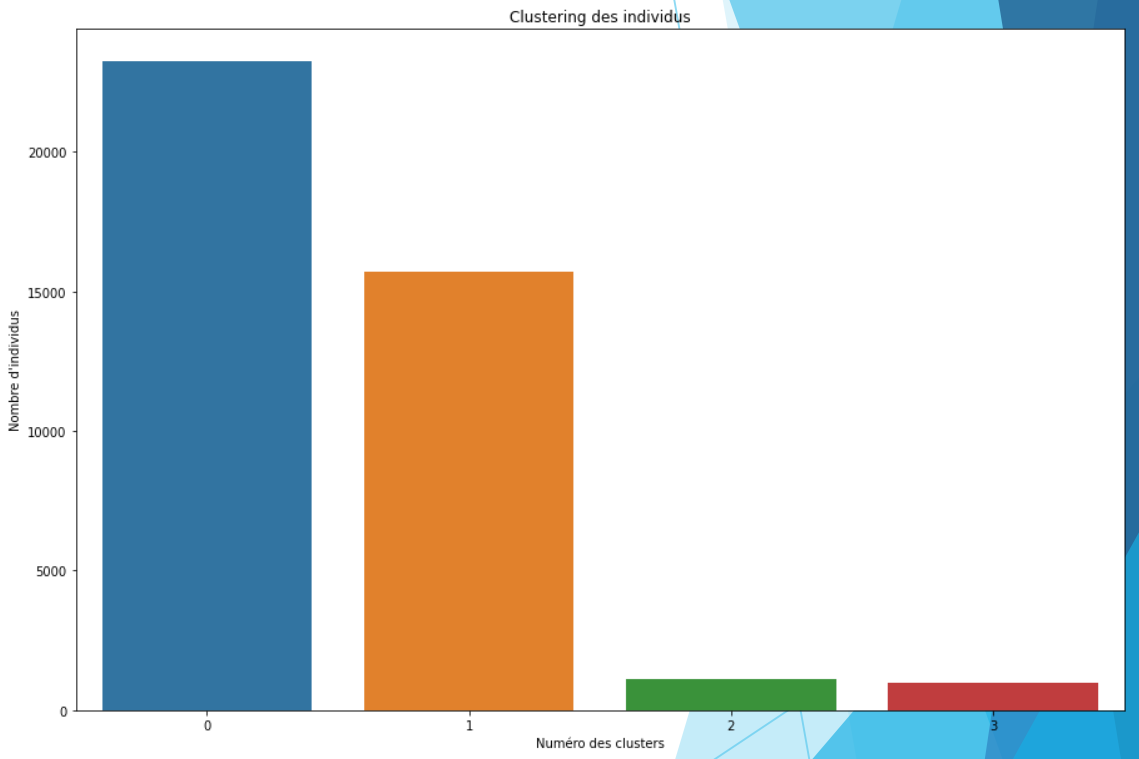




3. Modèle optimal

Segmentation RFM : Résultats du clustering avec KMeans

	nombre_jours		nombre_commande		moyenne_prix_commande		
	mean	std	mean	std	mean	std	size
clusters							
0	315.706161	44.146781	1.000000	0.000000	117.634100	108.390278	23244
1	489.085369	60.125911	1.000000	0.000000	116.238812	109.561442	15720
2	373.207840	90.788017	2.065634	0.247753	130.377698	130.293663	1097
3	393.161060	94.281864	1.004077	0.063757	1182.500836	767.877152	981



Profils des différents clients

Profil 0 = Clients récents

Profil 1 = Clients anciens

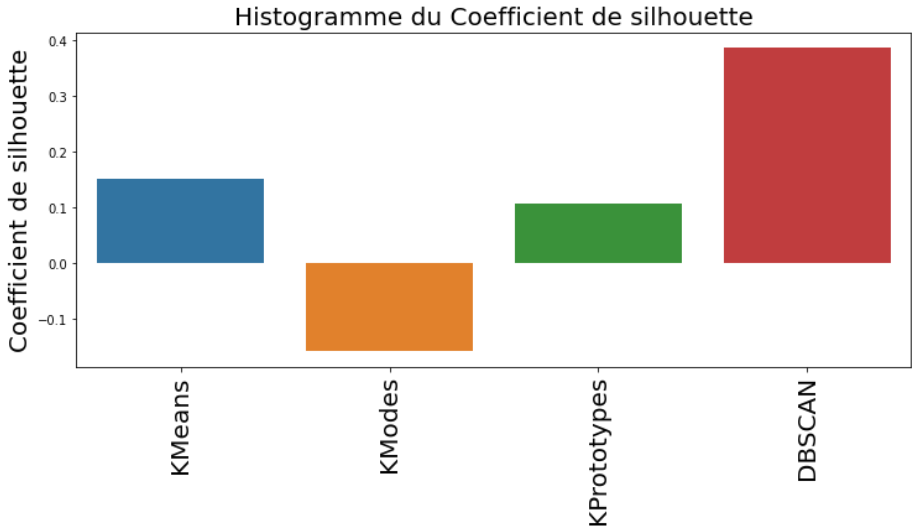
Profil 2 = Clients ayant passé plus de 2 commandes

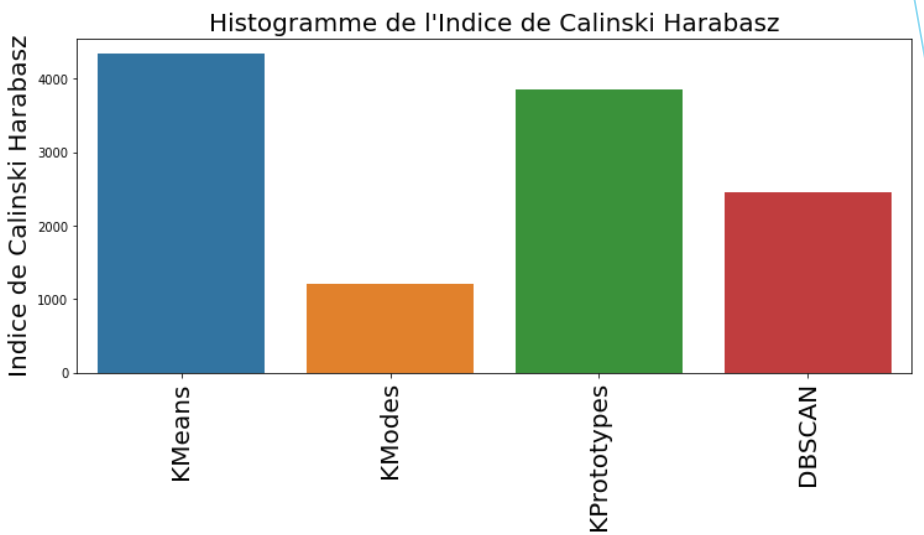
Profil 3 = Clients ayant beaucoup dépensé

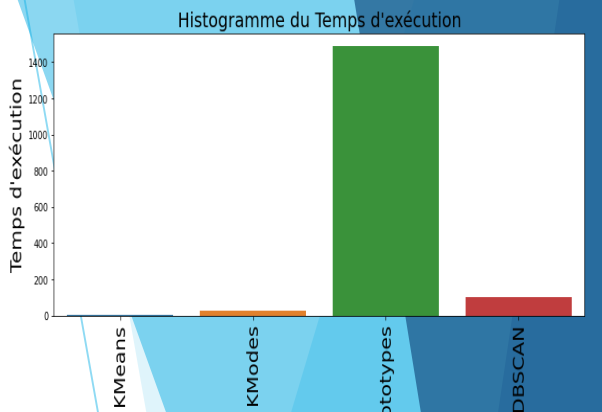
Groupes 2 et 3 peu représentés dans le DataFrame

3. Modèle optimal

Segmentation en ajoutant des indicateurs : Comparaison des modèles



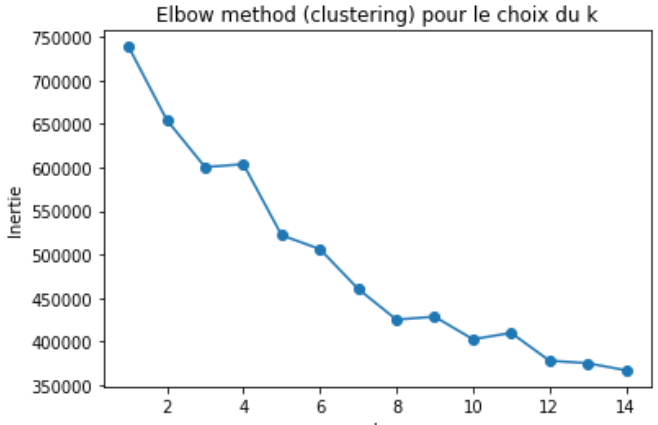


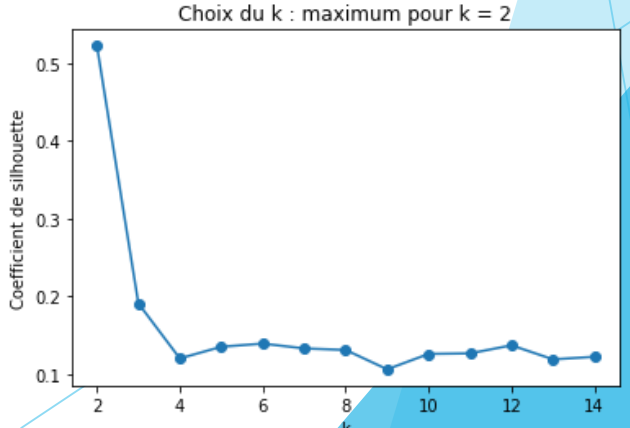


Sélectionne le KMeans avec un coefficient de silhouette à 0.15 et un indice de Calinski Harabasz à 4336, avec un temps d'exécution de 2.67 s

Méthode du coude :

➡ choix du n_clusters = 8

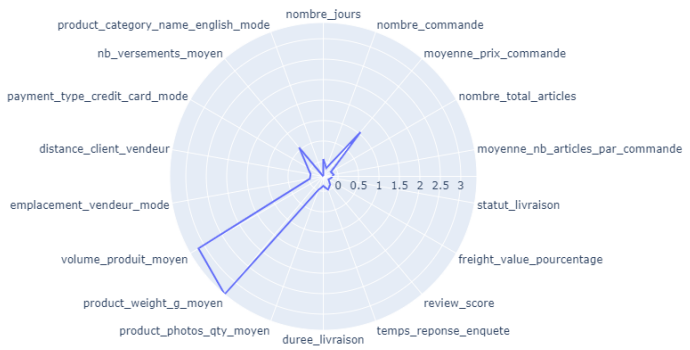




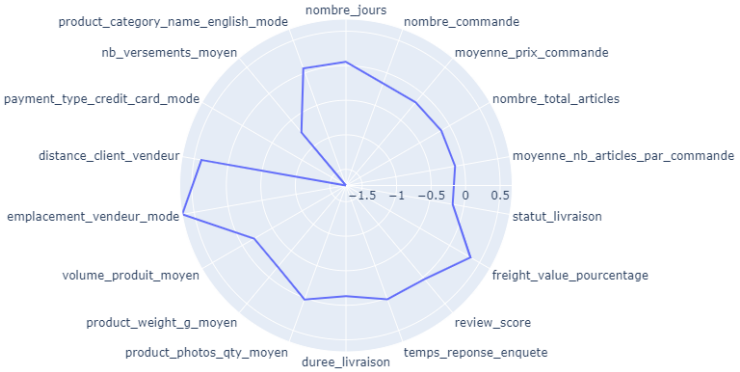
3. Modèle optimal

Segmentation en ajoutant des indicateurs : Résultats du clustering avec Kmeans

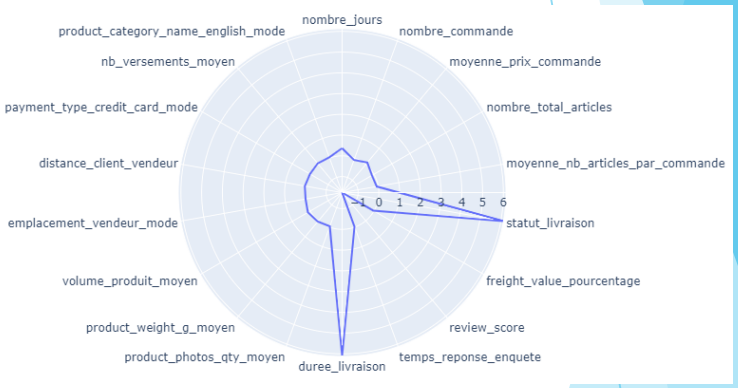
Profil 0



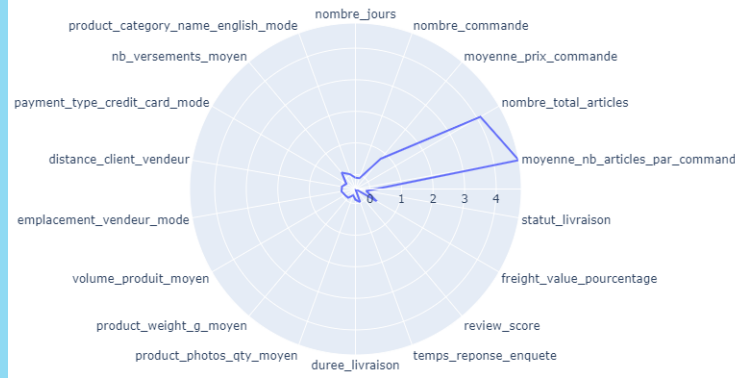
Profil 1



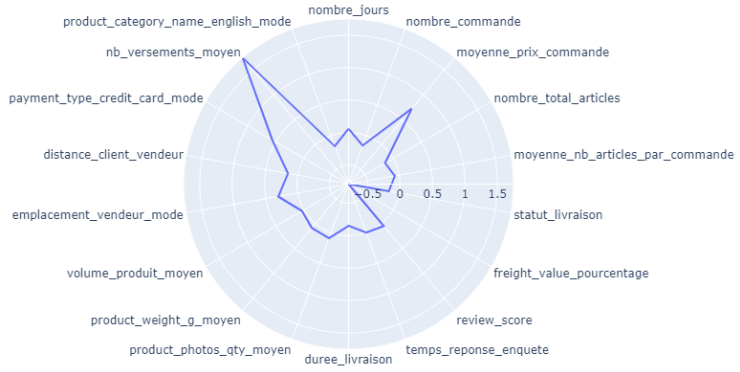
Profil 3



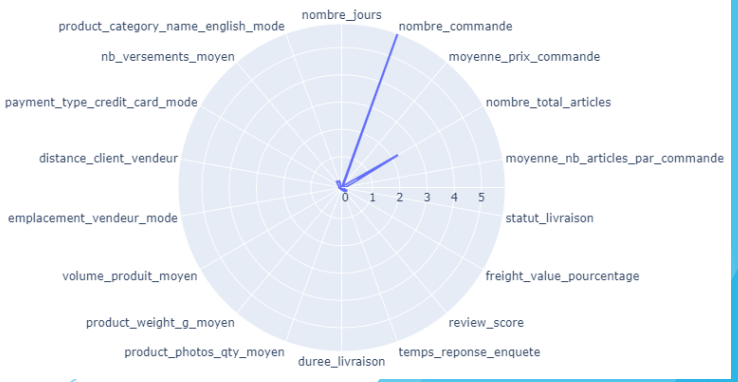
Profil 4



Profil 5



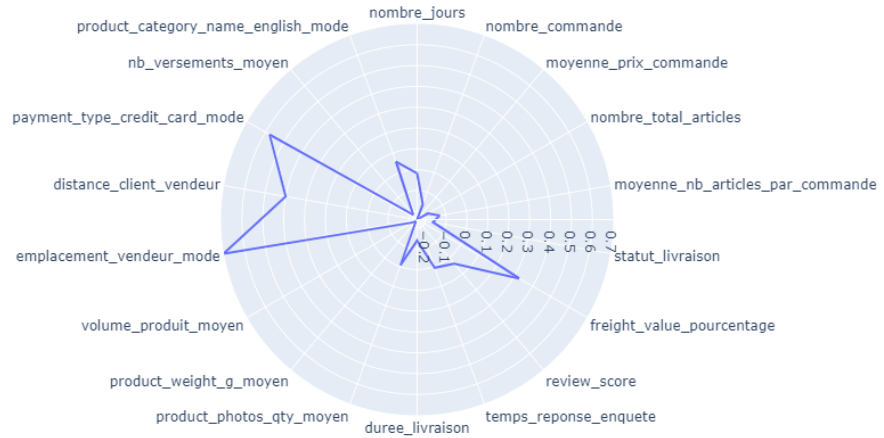
Profil 7



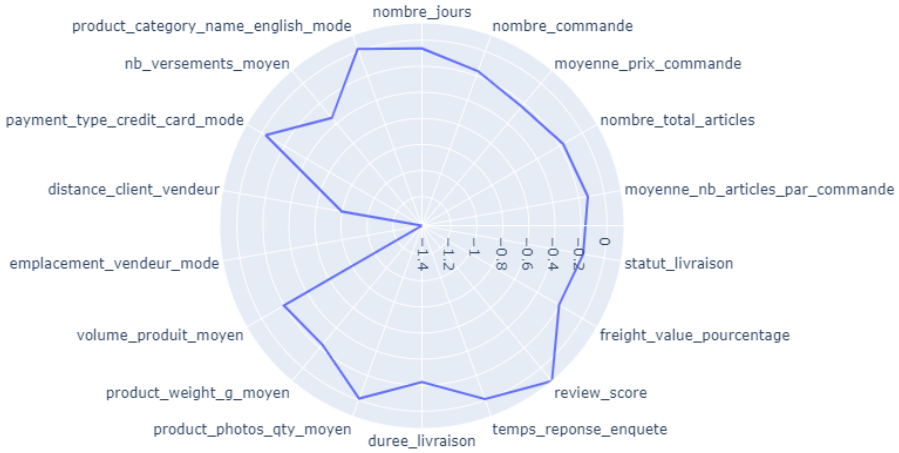
3. Modèle optimal

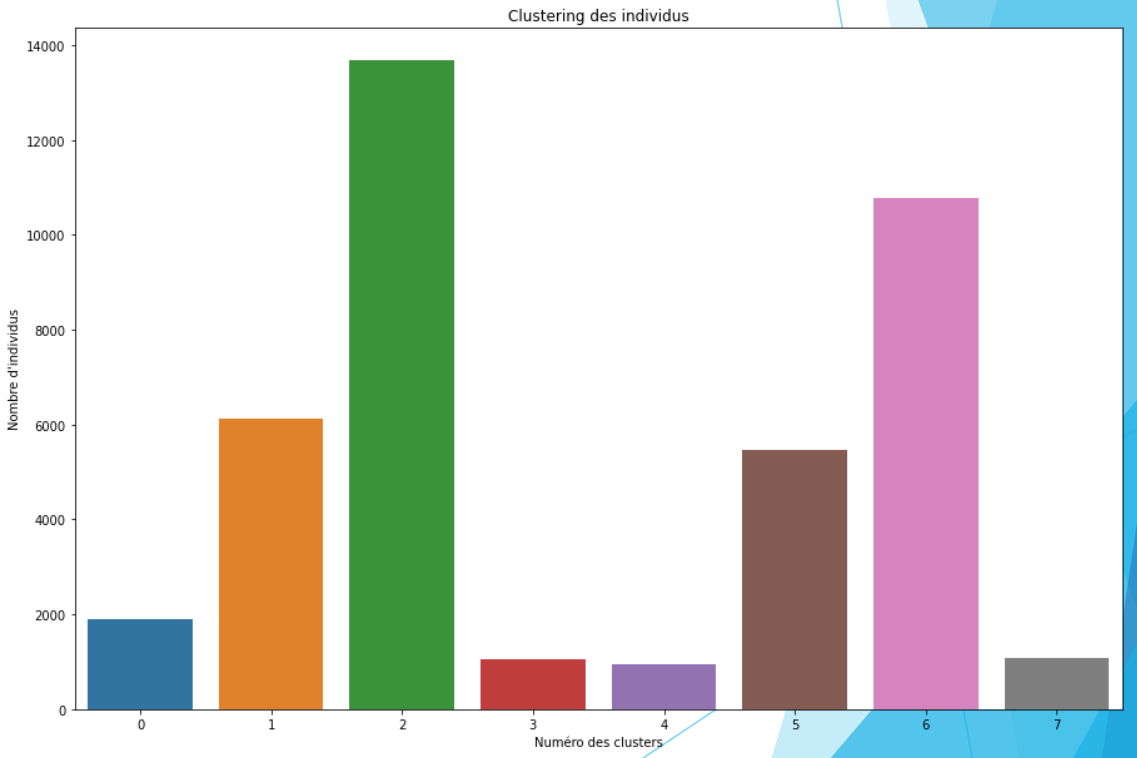
Segmentation en ajoutant des indicateurs : Résultats du clustering avec KMeans

Profil 2



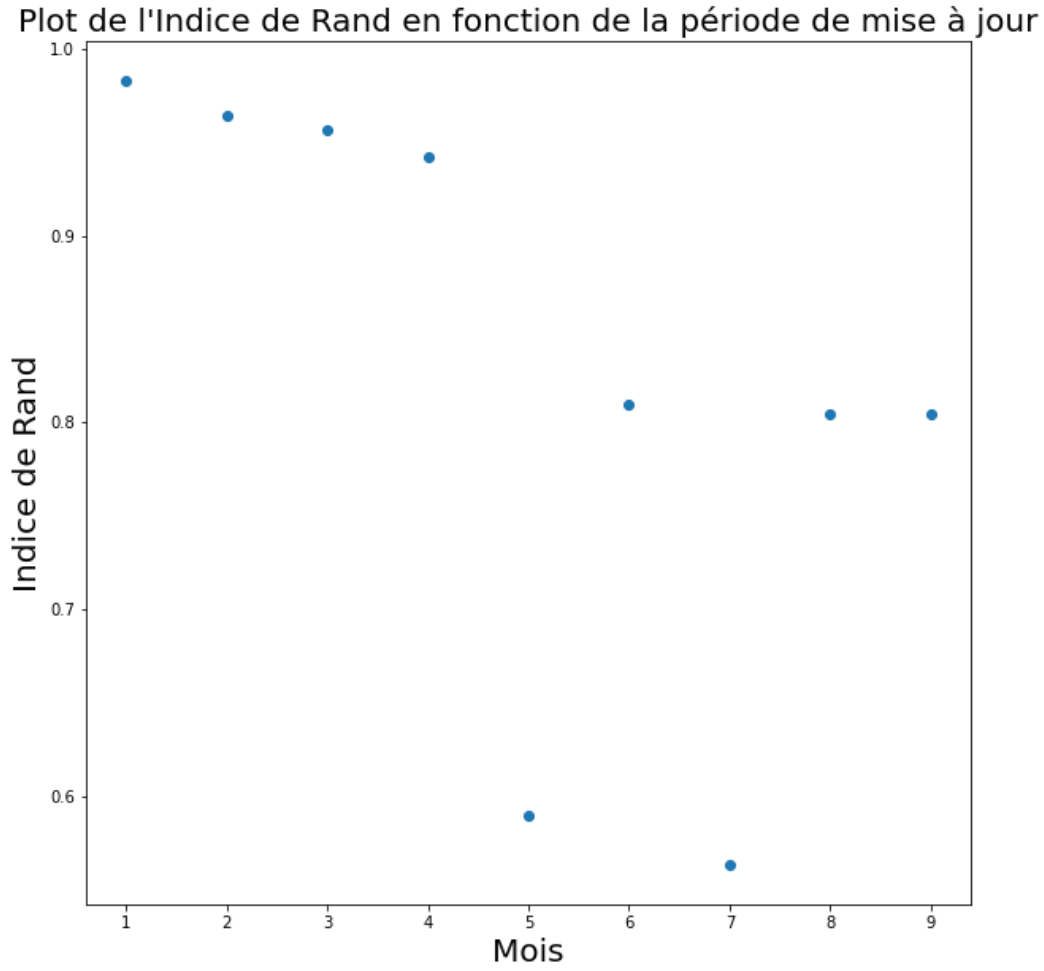
Profil 6





3. Modèle optimal

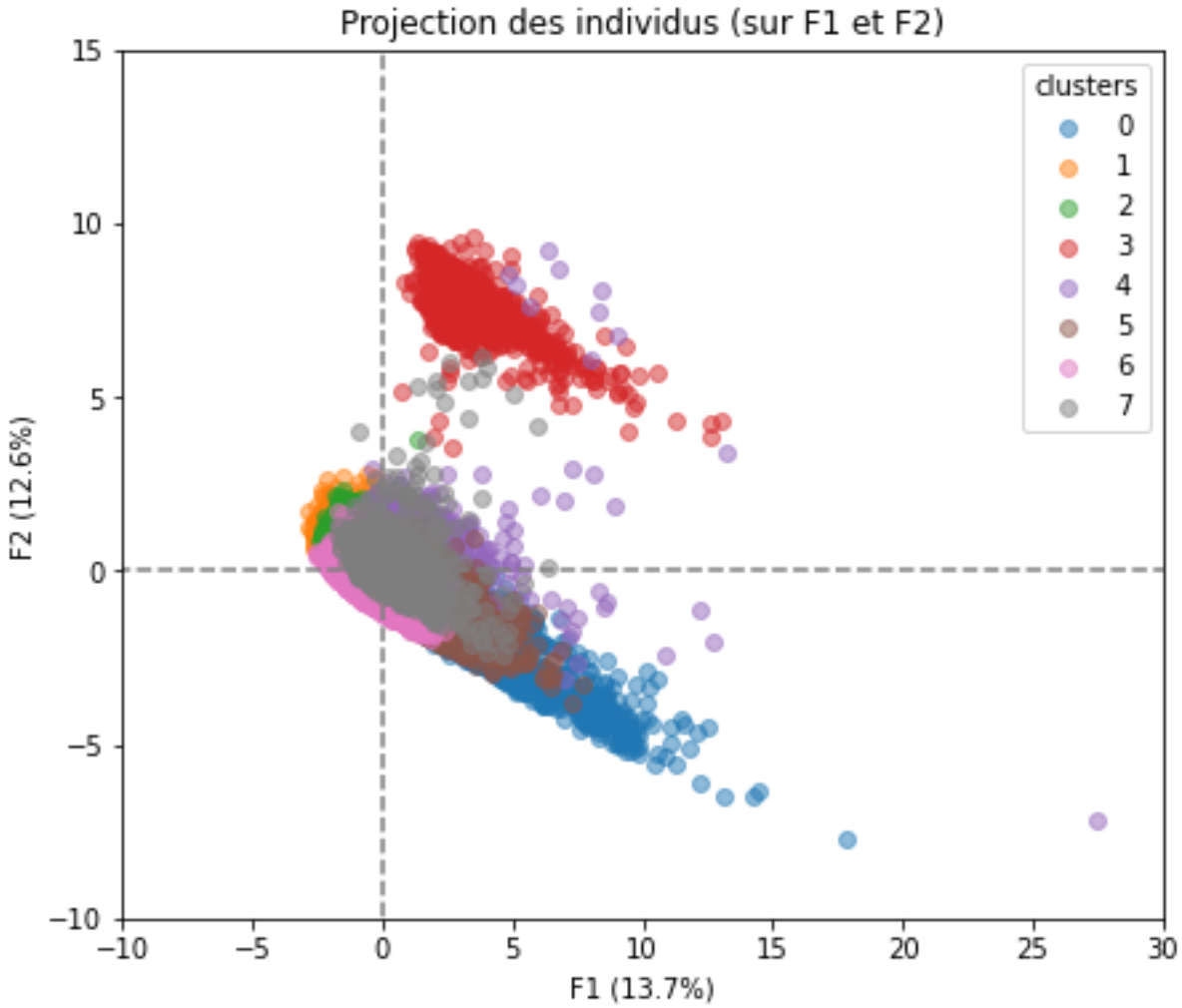
Modèle sélectionné : Kmeans avec l'ensemble des indicateurs

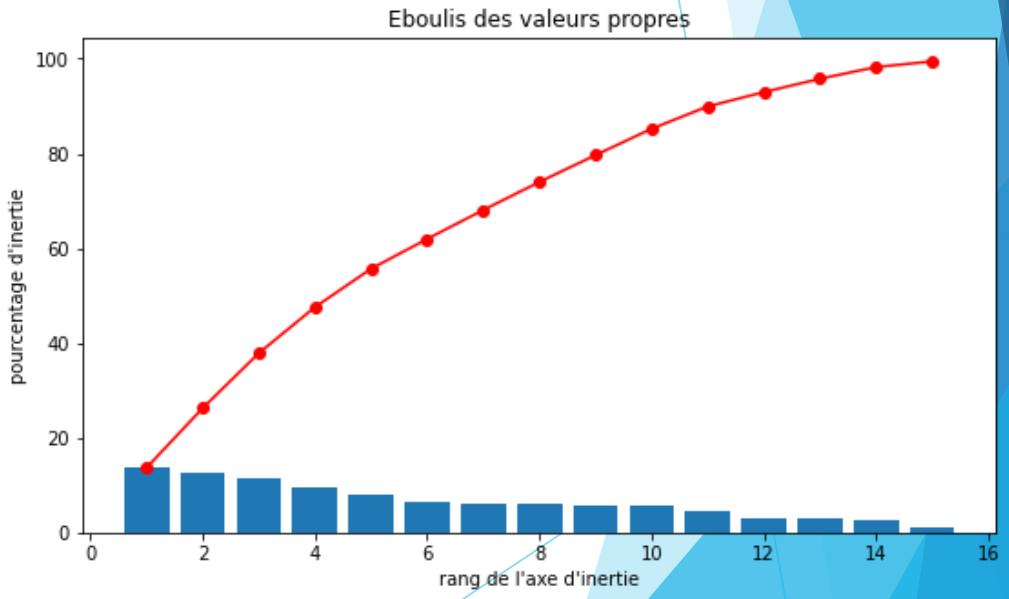


Mise à jour tous les 5 mois

3. Modèle optimal

Modèle sélectionné : Kmeans avec l'ensemble des indicateurs





Conclusion

Réponse à la problématique : Comprendre les différents types d'utilisateurs

Segmentation RFM => Apporte plus d'informations à l'équipe marketing en ajoutant des indicateurs (profils clients plus détaillés)

Modèle optimal = **Kmeans** VS KPrototypes, KModes et DBSCAN

Semble apporter le plus d'informations dans la segmentation des clients et plutôt rapide

- Segments plus représentatifs que d'autres
- Profils exploitables

Etude de la stabilité des segments au cours du temps = Mise à jour tous les 5 mois (après chute de l'Indice de Rand)

MERCI DE VOTRE ATTENTION

QUESTIONS – REPONSES