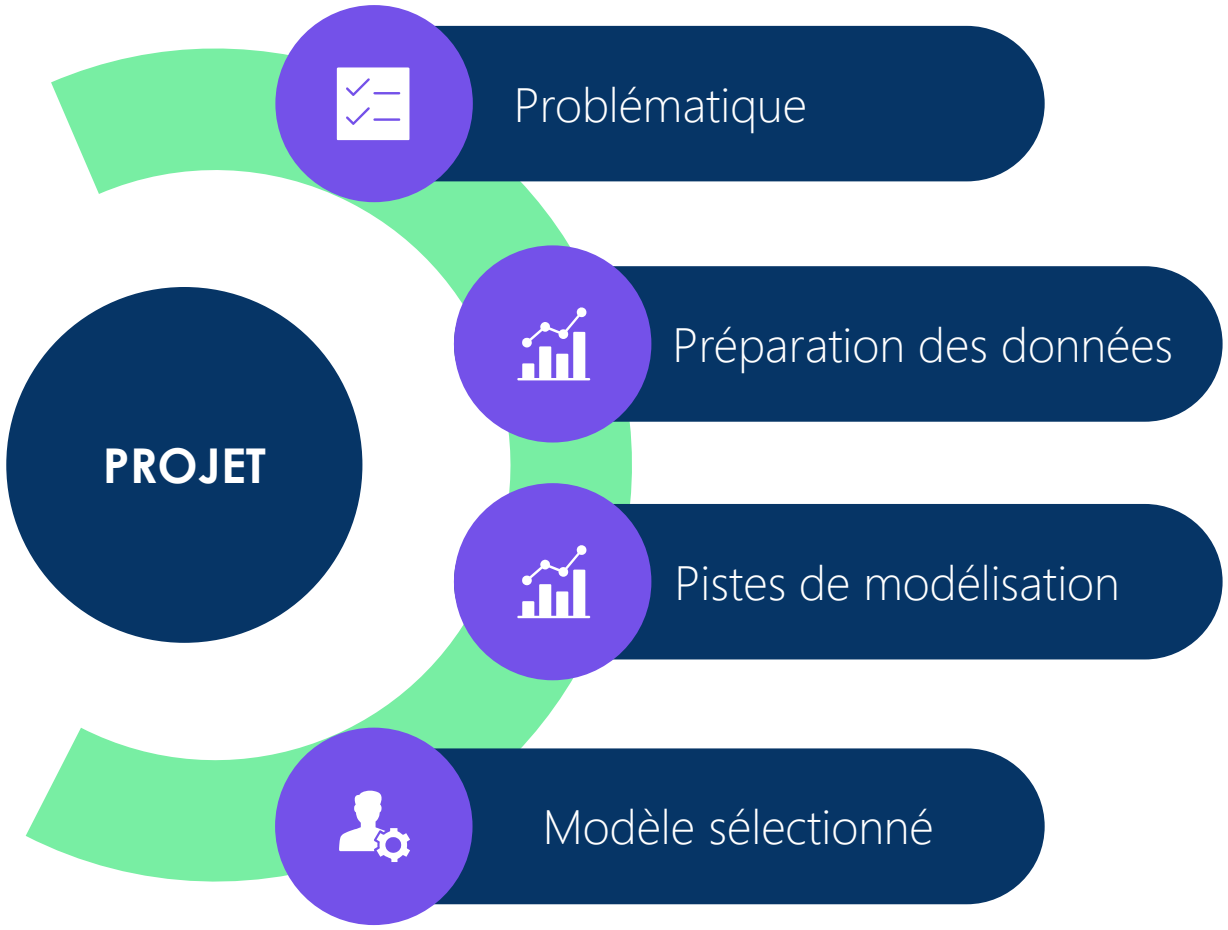
Projet N°5 :
Segmentez des clients d'un site e-commerce

Agustin Bunader (autofinancé)
Soutenance de Projet
Avril 2021

Programme


PROJET
Problématique
Préparation des données
Pistes de modélisation
Modèle sélectionné

Problématique – Présentation

Enterprise :

- Olist fournit un support opérationnel complet aux clients (commerçants) en gérant leurs catalogues de produits, l'inventaire, la tarification, l'exécution, le service client et les paiements à un seul endroit

Problématique :

- Olist souhaite une segmentation des clients à utiliser au quotidien pour leurs campagnes de communication

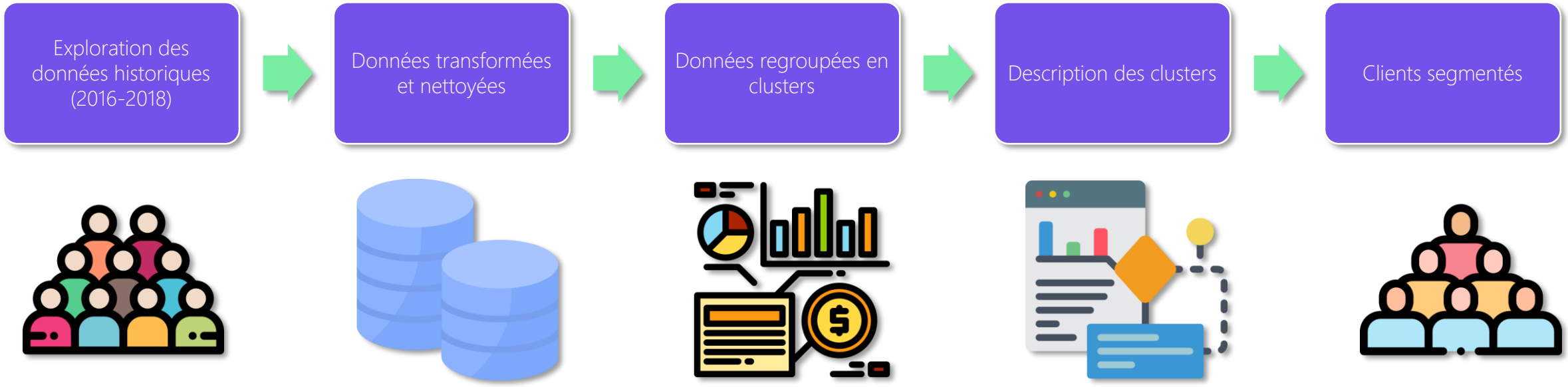
Mission :

- Comprendre les différents types d'utilisateurs d'Olist
- Utiliser des méthodes non-supervisées pour regrouper l'ensemble des clients de profils similaires
- Fournir une description actionnable de la segmentation et sa logique sous-jacente pour une utilisation optimale
- Fournir une proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps

Problématique – Interprétation

Interprétation :

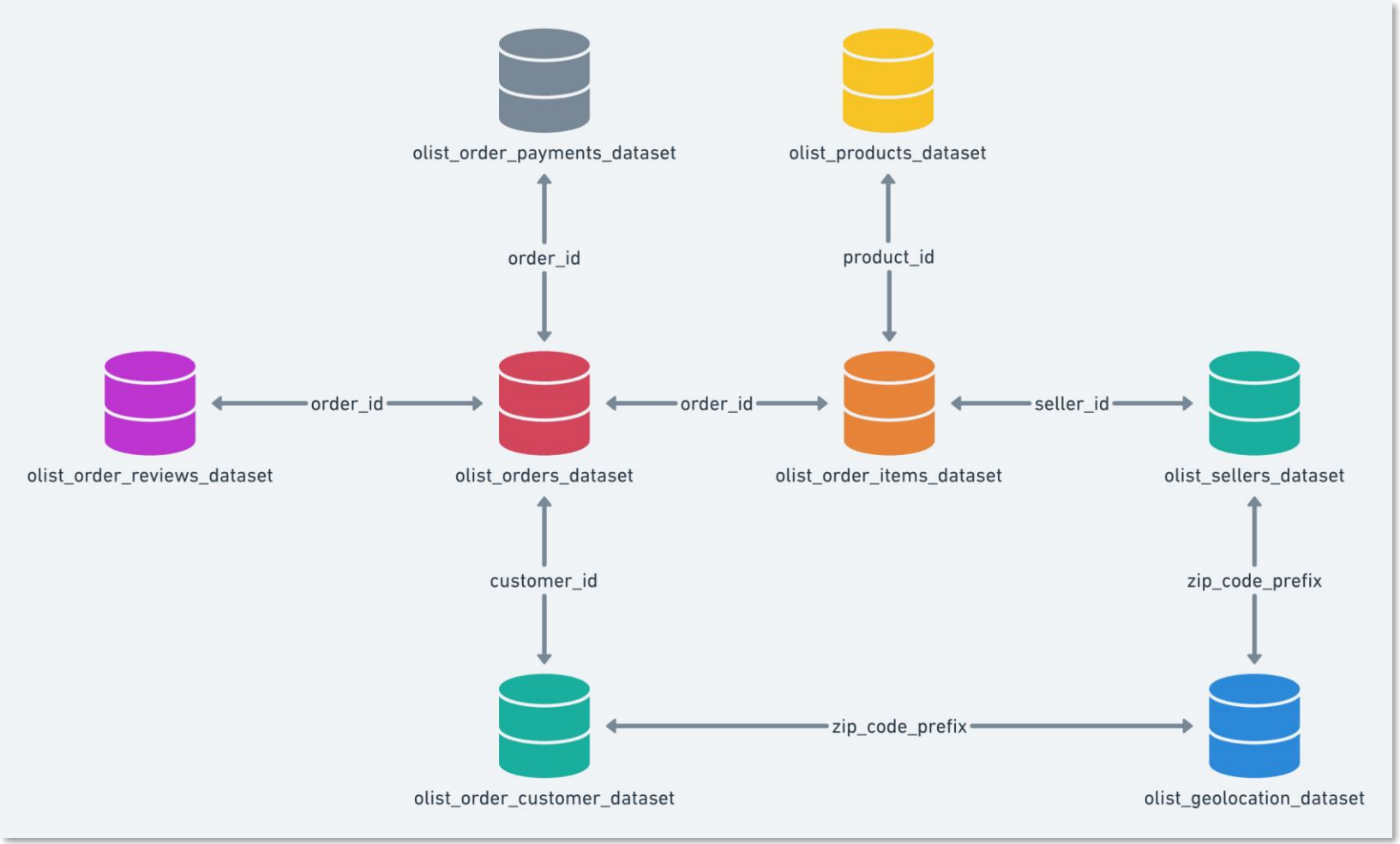
- Exploration des données et choix des features adaptés
- Classification non-supervisée avec des clusters qui devront être explicables et réutilisables par l'équipe de marketing



Problématique – Data Schema

Présentation des données :

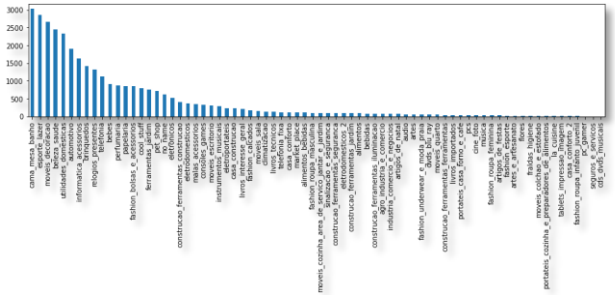
- Les données sont organisées dans une base de données relationnelle distribuée en 8 fichiers csv
- Un 9ème fichier contient les traductions de chaque catégorie de produits du portugais brésilien à l'anglais

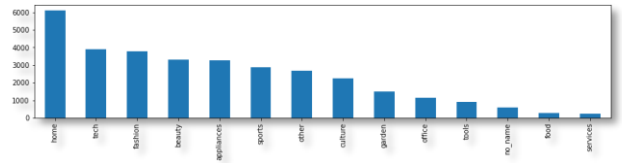


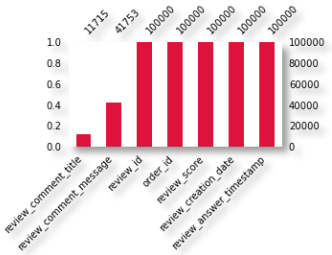
Préparation des données – Nettoyage

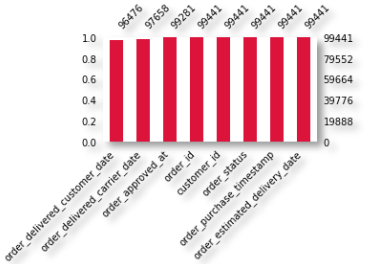
Nettoyage sur les csv sources

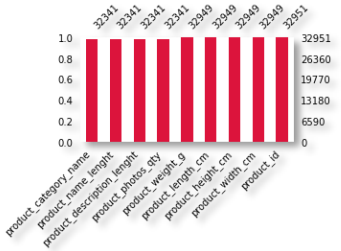
Quoi ?	Situation	Résultat
translation : product_category_name	72 des 74 catégories de produits ont des traductions du portugais à l'anglais	• Les traductions manquantes ont été ajoutées • Les catégories ont été réduites en 15 groupes
geolocation : geolocation_city **customers** : customer_city **sellers** : seller_city	Plusieurs champs de villes sont remplis d'erreurs, par exemple la ville de São Paulo est écrite au moins de 37 manières différentes	Correction des erreurs orthographiques dans les noms des villes
geolocation : geolocation_lat **geolocation** : geolocation_lng	Une vingtaine des coordonnées aberrantes selon les coordonnées maximales du Brésil (source : Wikipedia)	Suppression des coordonnées aberrantes
Doublons	261831 doublons sur **geolocation**	Suppression des doublons
NaN	**order_reviews** : 0.21% **orders** : 0.01% **products** : 0.01%	• **order_reviews** : les NaN sont remplacés par *no_title* et *no_message* selon le cas • **orders** : les NaN sont sur les dates et elles sont remplis par 1970-01-01 • **products** : les produits sans catégorie reçoivent la catégorie *no_name*. Les mesures NaN sont remplis par 0
orders	NaN remplacés par 1970-01-01	Les commandes avec la date 1970-01-01 sont effacés











Préparation des données – Assemblage

Assemblage :

- Union des données dans une table unique avec *customer_unique_id* comme index
- Pour éviter les biais lors de la modélisation, nous retirons le 1% (outliers) avec la méthode des quantiles
- On ajoute des colonnes RFM pour aider la clustérisassions des données
- Création des nouveaux features (en plus de RFM)

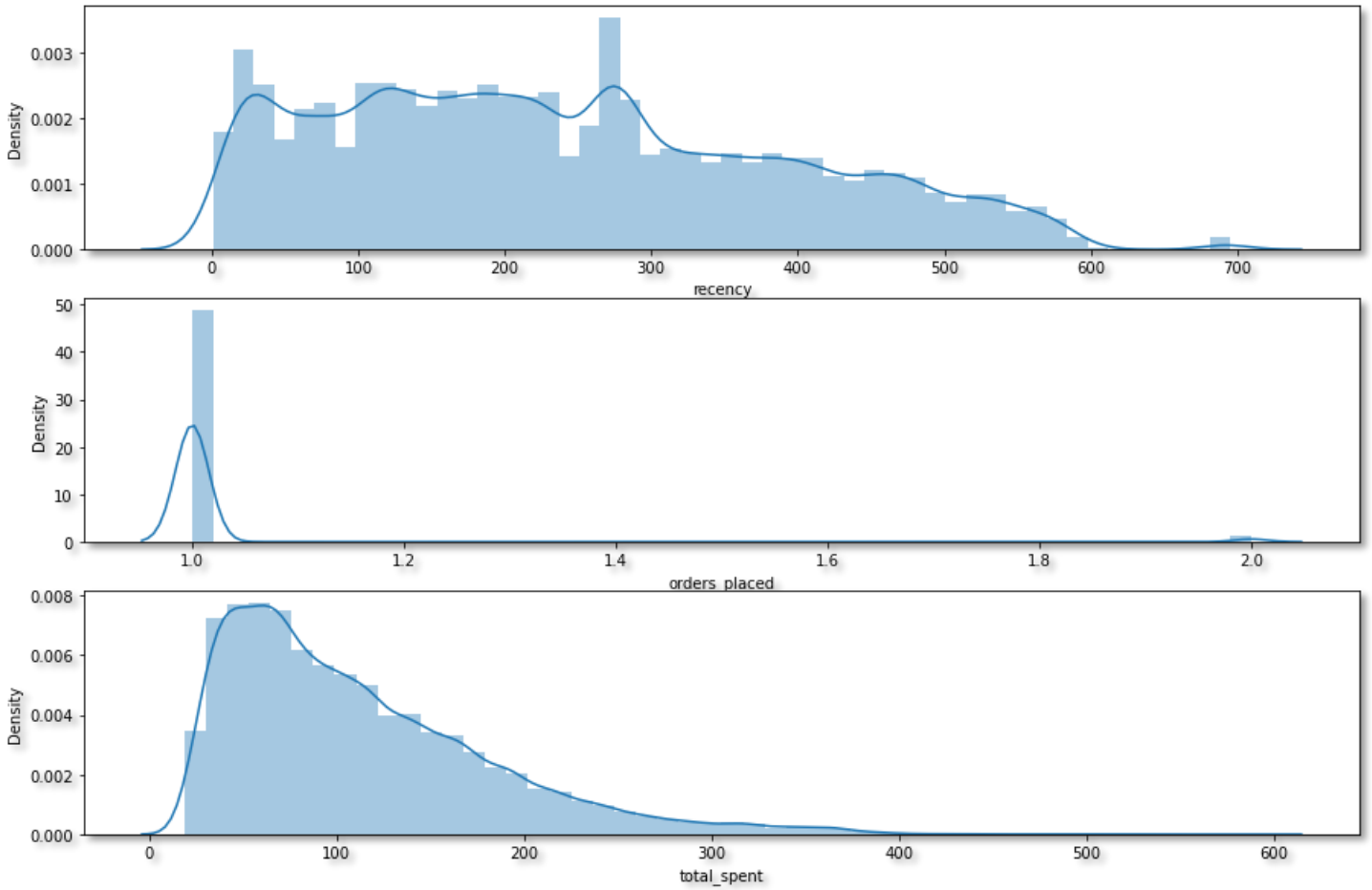
Jeu assemblé :
73866 lignes
39 colonnes

Feature	Description (par client unique)
total_spent	Total dépensé (*payment_value*)
order_placed	Quantité de commandes faites
total_products_bought	Quantité totale de produits achetés
prefered_product_category	Catégorie de produit préférée
prefered_payment_type	Méthode de paiement préférée
avg_review_score	Review score moyen
max_order_products / avg_order_products	Quantité maximale / moyenne des produits achetés par commande
max_order / avg_order	Montant maximal / moyenne dépensée par commande
total_weight_g	Poids total de toutes les commandes
avg_weight_g	Poids moyen de commande
max_delivery_delay / avg_delivery_delay	Délai maximal / moyen de traitement de commande
first_order_date	Date de la première commande
last_order_date	Date de la dernière commande
appliances, beauty, culture, fashion, food, garden, home, no_name, office, other, services, sports, tech & tools	Montant dépensé par catégorie (sans inclure les frais de livraison)

Préparation des données – Assemblage

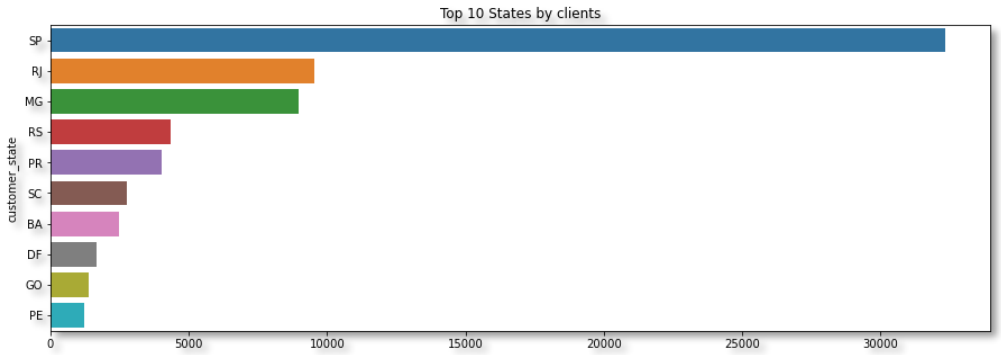
RFM :

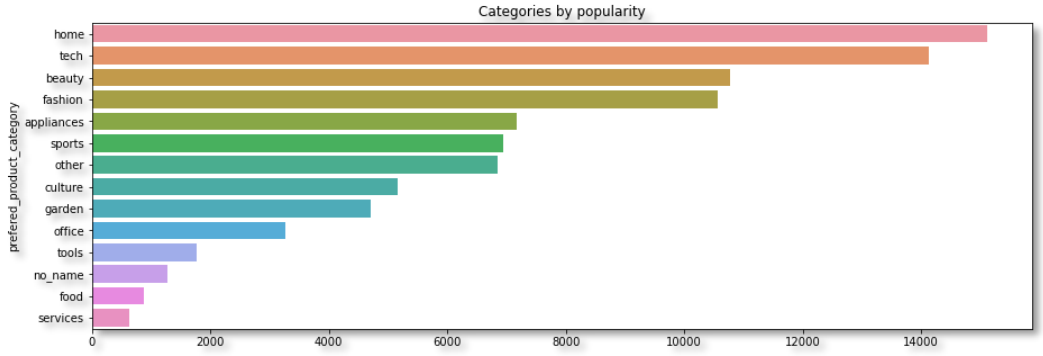
- Récence (R) : date du dernier achat ou dernier contact client. Valeurs entre 1 et 4.
- Fréquence (F) : fréquence des achats sur une période de référence donnée. Valeurs entre 1 et 4.
- Montant (M) : Somme des achats cumulés sur cette période. Valeurs entre 1 et 4.

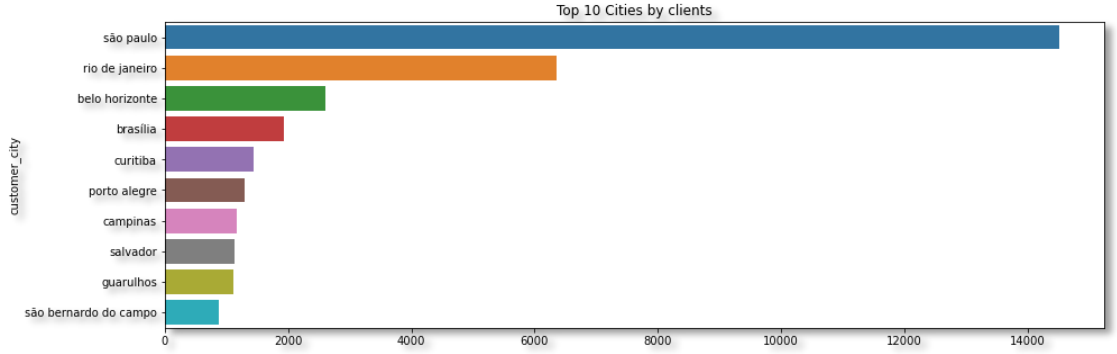


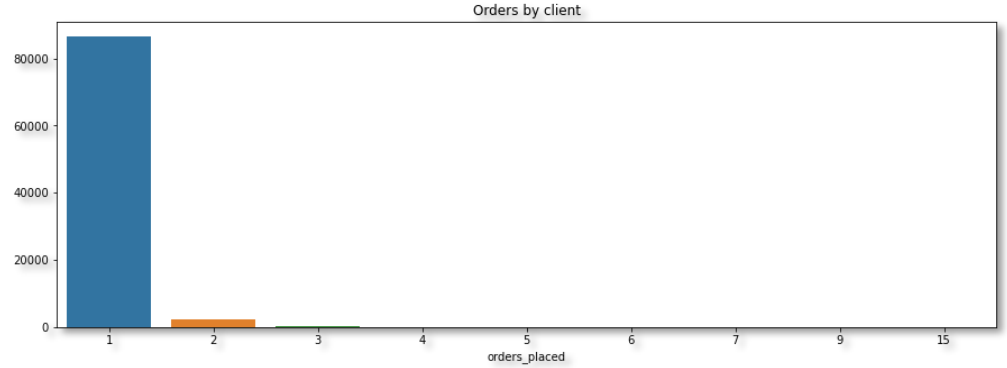
Préparation des données – Exploration

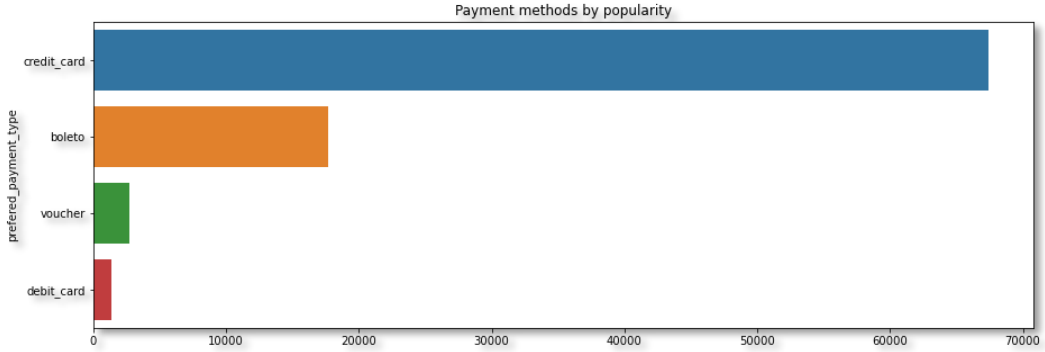
Exploration univariée





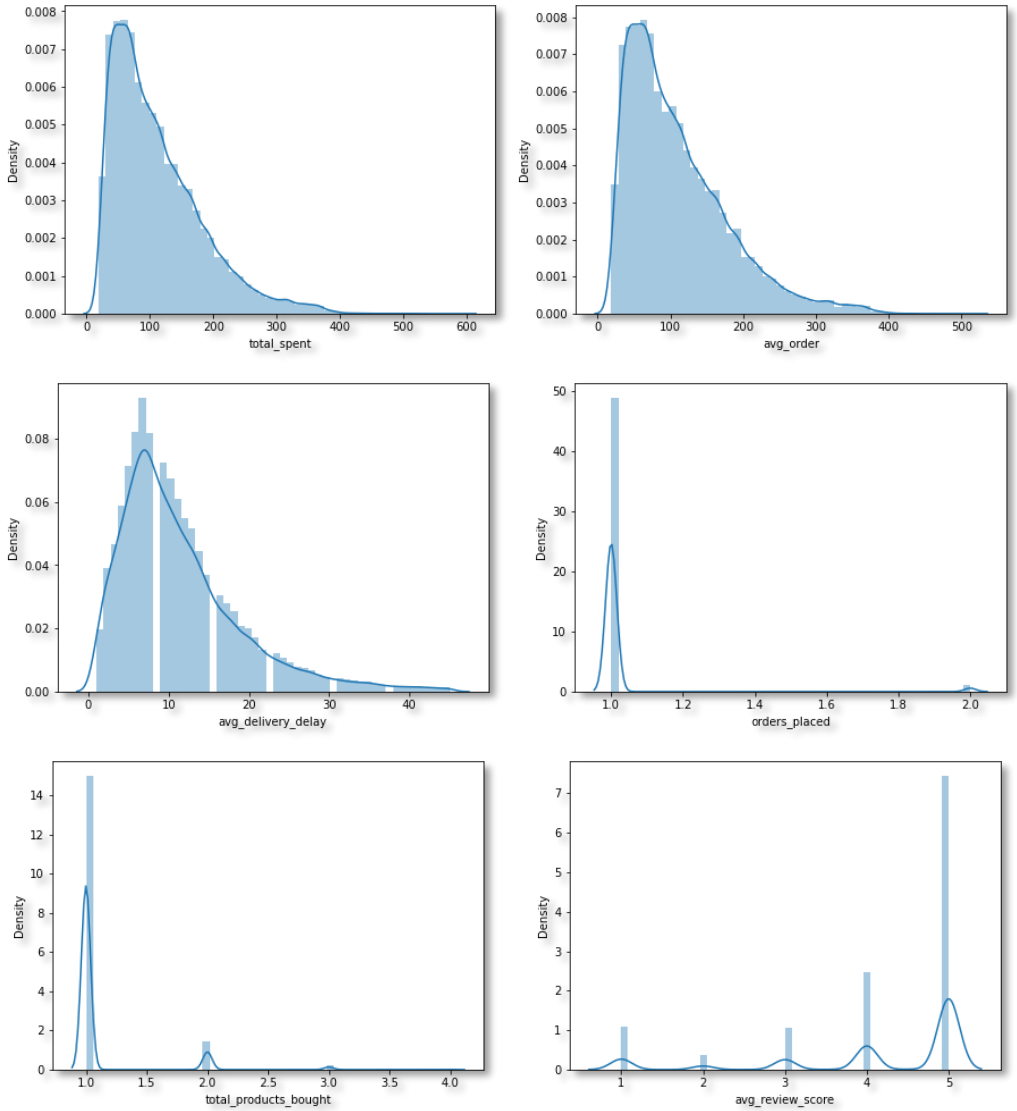




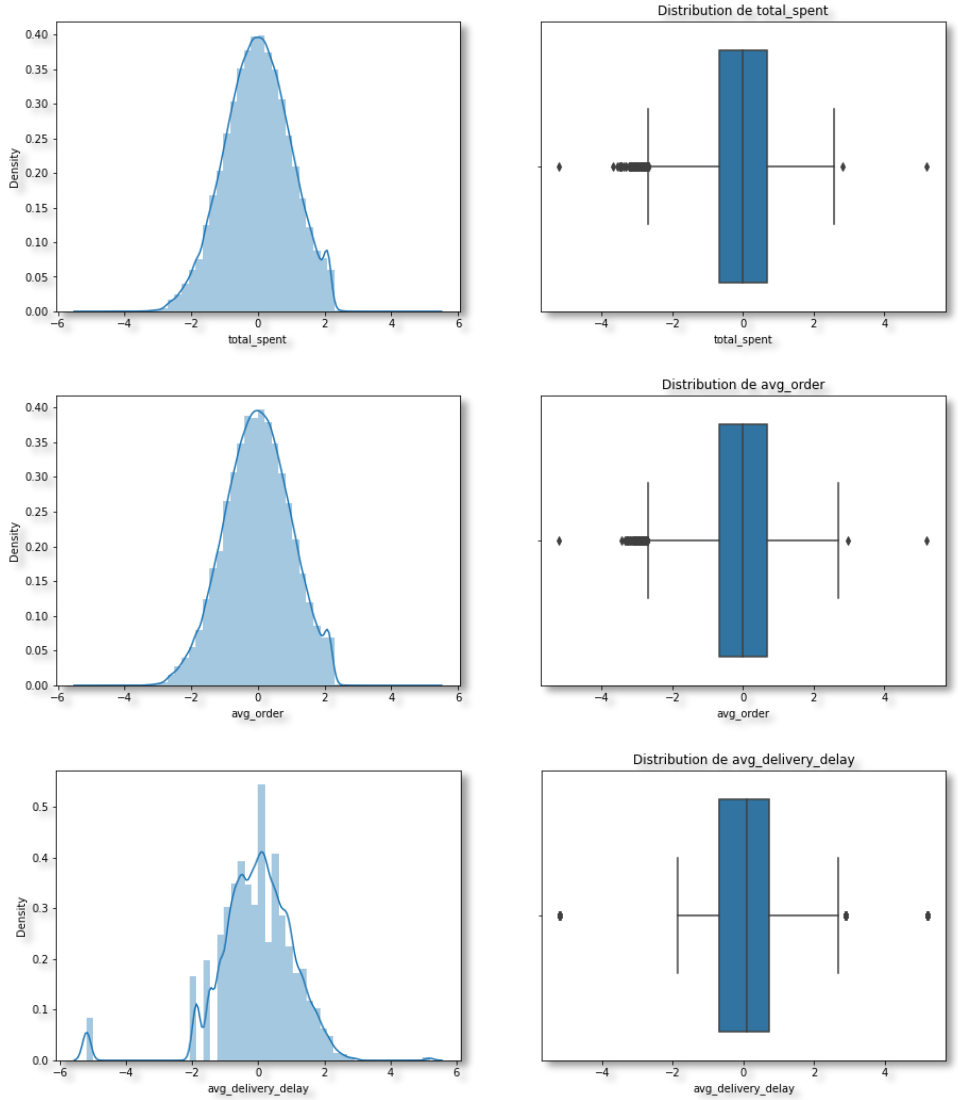


Préparation des données – Exploration

Exploration univariée

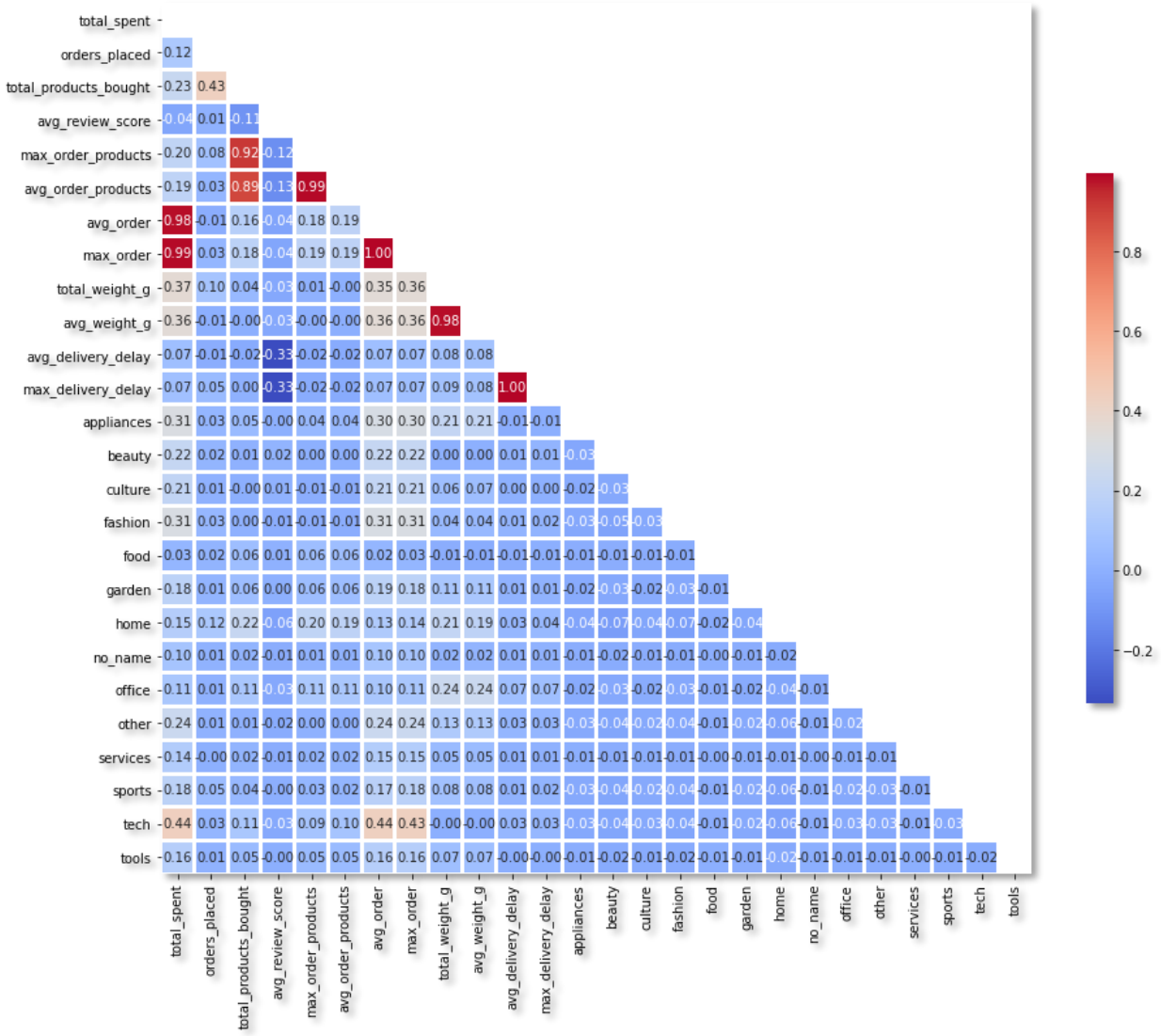


Exploration univariée – Distribution normale



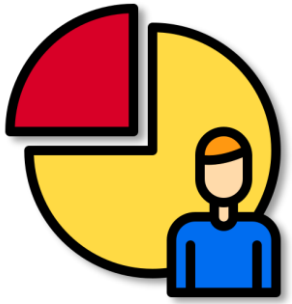
Préparation des données – Exploration

Exploration bivariée



Résultat :

Il est compliqué de déterminer la pertinence des features car la plupart du jeu de données est composé de clients avec un seul achat d'un seul article



Pistes de modélisation – Processus

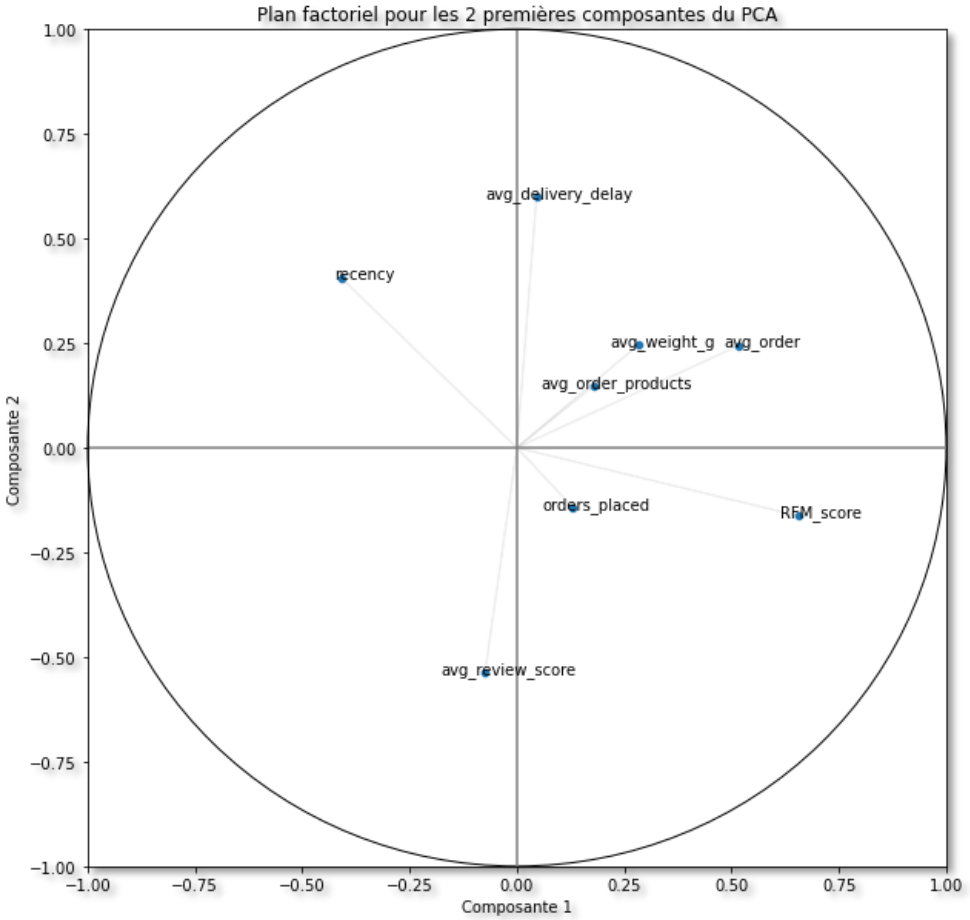
Modèles :

- K-Means, DBScan, AgglomerativeClustering
- Taille originale : 73866
- Taille réduite : 18467

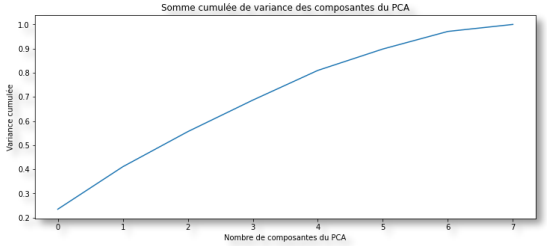
Choix de modèle et nombre de clusters sur échantillon réduit (25%) → Application sur la totalité des données → Description des clusters → Offre commerciale (mise à jour des clusters)

Pistes de modélisation – PCA et K-Means

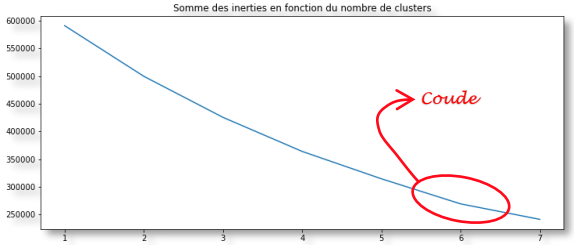
Visualisation PCA :



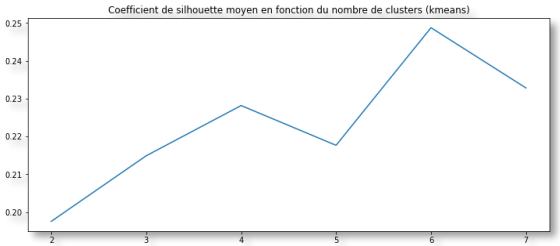
PCA & K-Means :



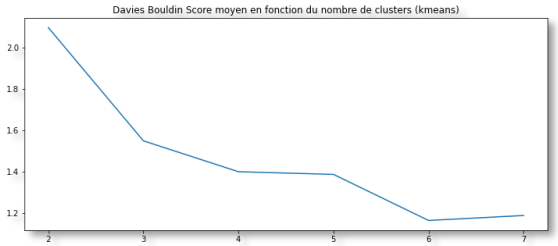
PCA
- 7 variables
- 1 taux de variance cumulé



K-Means
- Coude



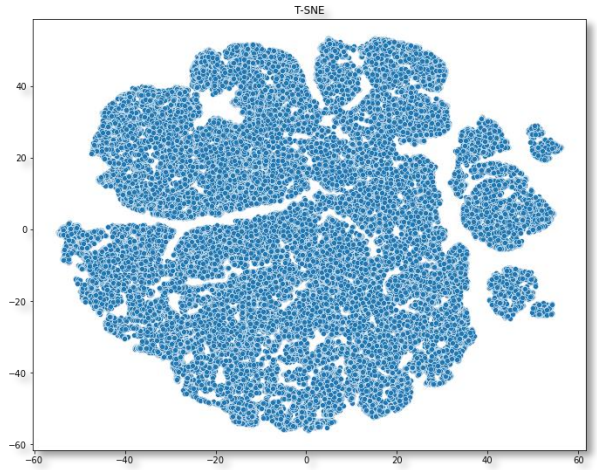
K-Means Silhouettes Score
- 6 clusters
- 0.2488 score

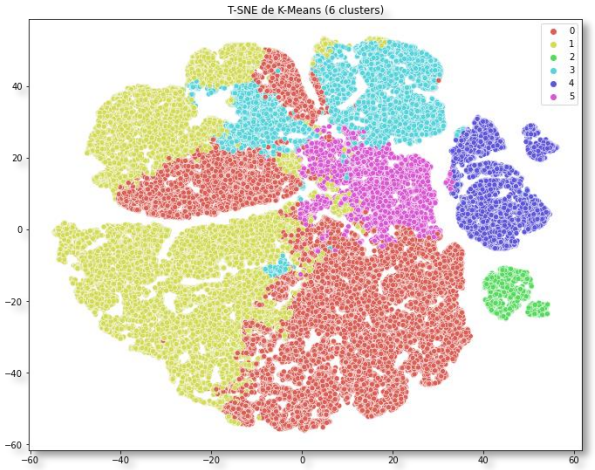


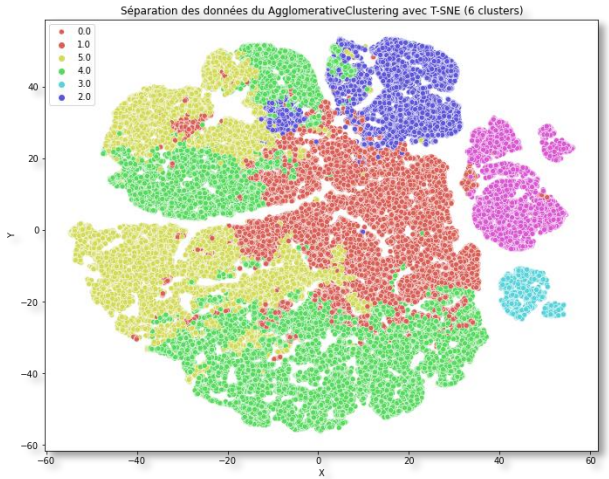
K-Means Davies Bouldin
- 6 clusters
- 1.1642 score

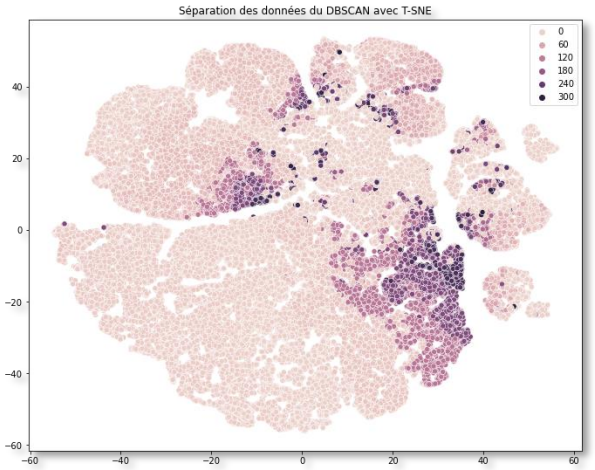
Pistes de modélisation – Clusters

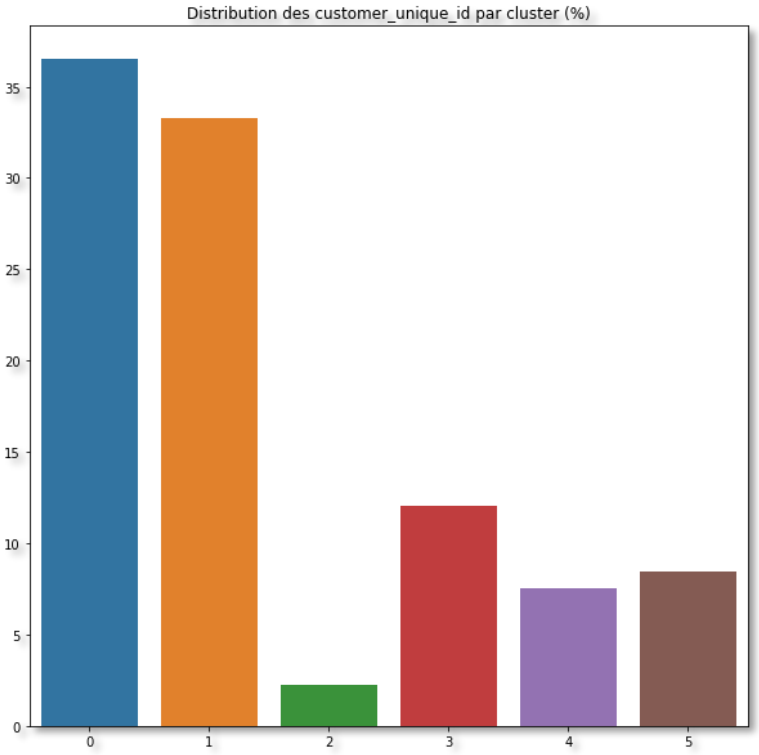
Visualisation t-SNE de K-Means / AgglomerativeClustering / DBSCAN :











Label	Taille
0	26969
1	24570
2	8879
3	6233
4	5538
5	1676

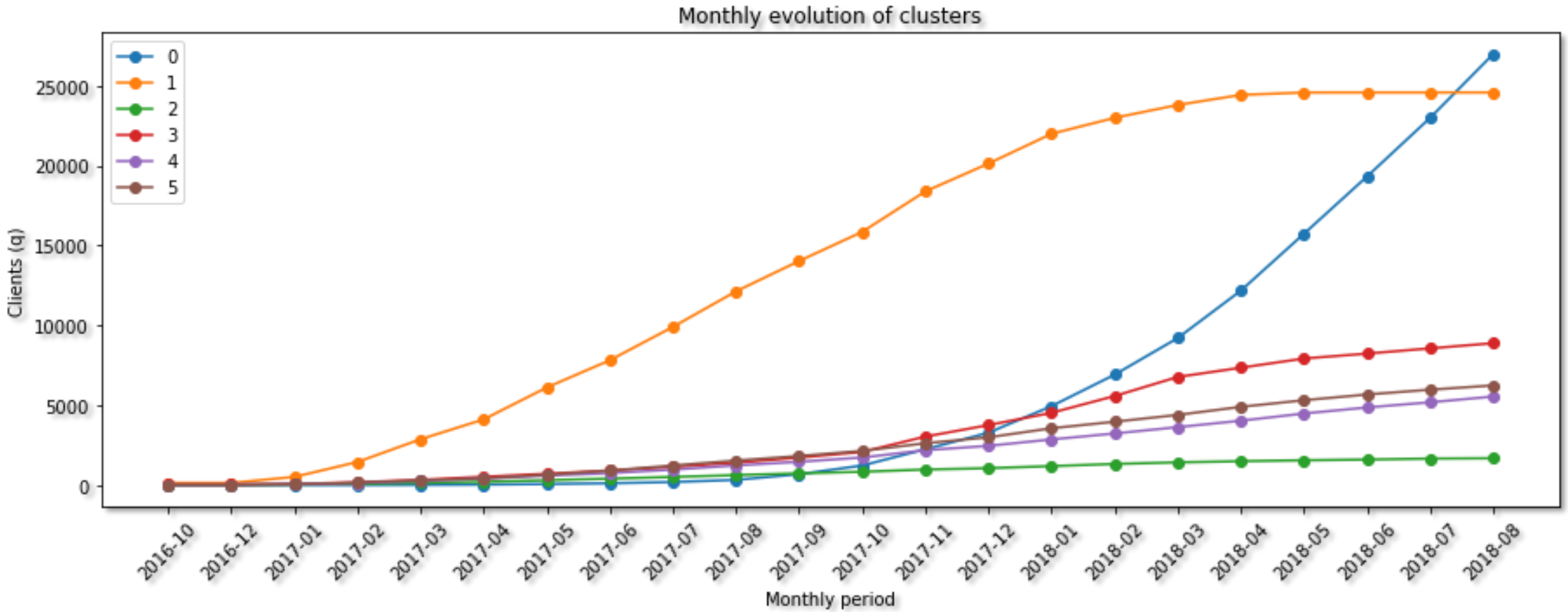
Modèle sélectionné – Clusters & Proposition

Description des clusters :

Cluster	Description	Nom	Action / Proposition	
Cluster 0	• Meilleure récence • Délai de livraison le plus bas (en moyenne) • Les meilleures notes • Deuxième RFM Score	Loyal	Action seulement si leur avis change.	
Cluster 1	• Pire récence • Moins dépensé et le moins de poids acheté • Très bonnes notes • Pire RFM Score	Nécessite une activation	Nous devons les reconquérir.	
Cluster 2	• Le plus de commandes passées • Très bonnes notes • Premier RFM Score	Ambassadeur	Action seulement si leur avis change.	
Cluster 3	• Les pires notes • Délai de livraison le plus élevé	Malheureux	Problème de livraison. Enquêter sur les délais de livraison et un éventuel changement de transporteur.	
Cluster 4	• Plus d'un produit acheté par commande • Notes moyennes	Nécessite de l'attention	Les clients ont fait un premier achat mais on ne les a pas revus depuis. Mécontents ? Il faut créer la notoriété de la marque ?	
Cluster 5	• Le plus dépensé et le plus de poids acheté • Très bonnes notes • Deuxième pire récence	Prometteur	Ils montrent des signes prometteurs avec la valeur de leur achat, mais cela fait un moment qu'ils n'ont pas passé une commande. Cibler les articles de leur liste de souhaits avec une réduction temporelle ?	

Modèle sélectionné – Evolution

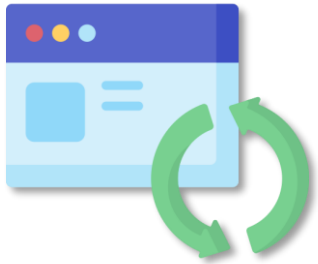
Evolution des clusters dans le temps :



Modèle sélectionné – Conclusion & Contrat

A remarquer :

- On a créé et transformé des variables malgré la faiblaisse du jeu de données (la plupart des clients ont un unique achat et leur comportement est assez similaire)
- Utilisation d'un modèle non-supervisé K-Means adapté à la problématique de segmentation de clientèle
- Optimisation de clusters (hyperparamètre du K-Means) avec PCA, Silhouettes et Davies Bouldin
- Utilisation du PCA et t-SNE pour représentation graphique des clusters
- Suivi de l'évolution des segments clients à travers la période disponible 10/2016 – 08/2018
- Rétrospectivement, une segmentation des données par périodes temporelles plus courtes avec une récence correspondant à chaque période aurait pu être testée
- Connaître l'âge et le sexe des clients peut aider à mieux créer les clusters



Maintenance :

Itération mensuelle de l'algorithme et contrôle de l'évolution des clusters. Selon la situation, de nouveaux clusters peuvent être ajoutés ou peuvent remplacer des clusters existants.

L'image Pickle créée lors de la première itération permet de suivre l'évolution des clients déjà existants. Cette image est la clé pour suivre l'évolution des clients.