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

OLIST

Jérémy VANGANSBERG - 2021

Introduction

- Qu’est ce qu’Olist ?
  - Plateforme de ventes en ligne, basée au **Brésil**
- Ma mission :
  - Aider les équipes d’Olist à comprendre les différents types d’utilisateurs en créant une **segmentation** des clients utilisable pour leurs campagnes de communication

Exemple d’un produit sur la market place

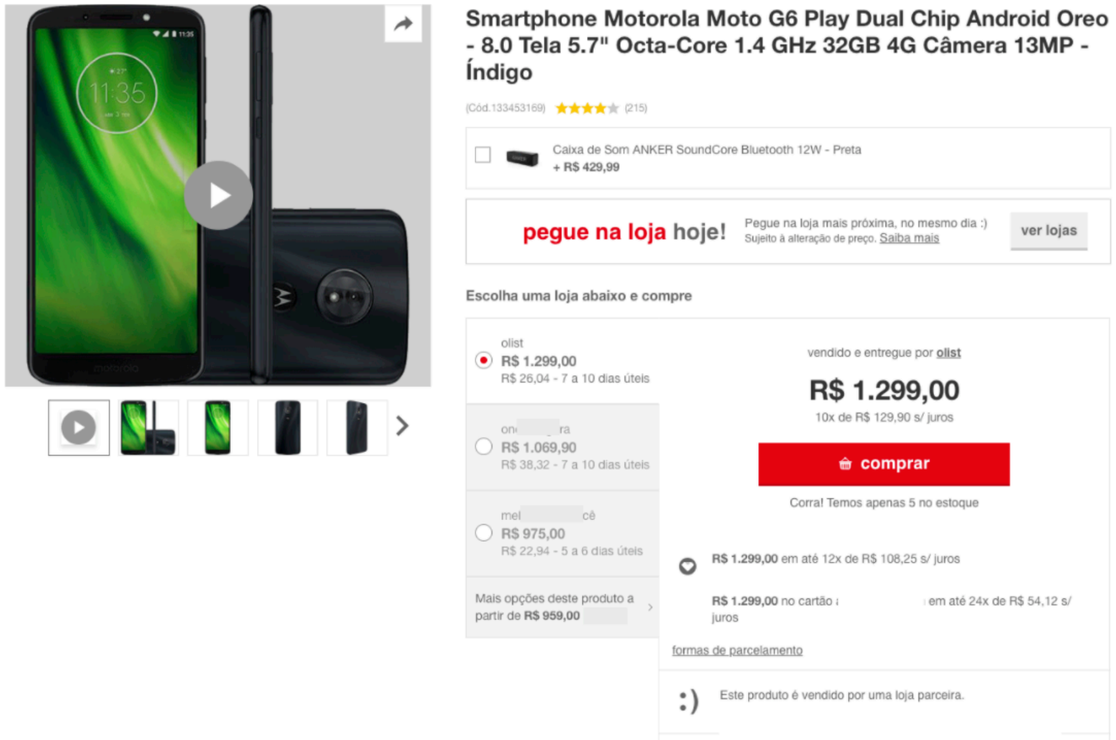


Table des matières

A. Problématique et pistes envisagées

B. Présentation des données

C. Nettoyage et *features engineering*

D. Exploration

E. Les modélisations

F. *Clusters* : analyse et visualisation

G. Maintenance

A. Problématique et pistes envisagées

Problématique

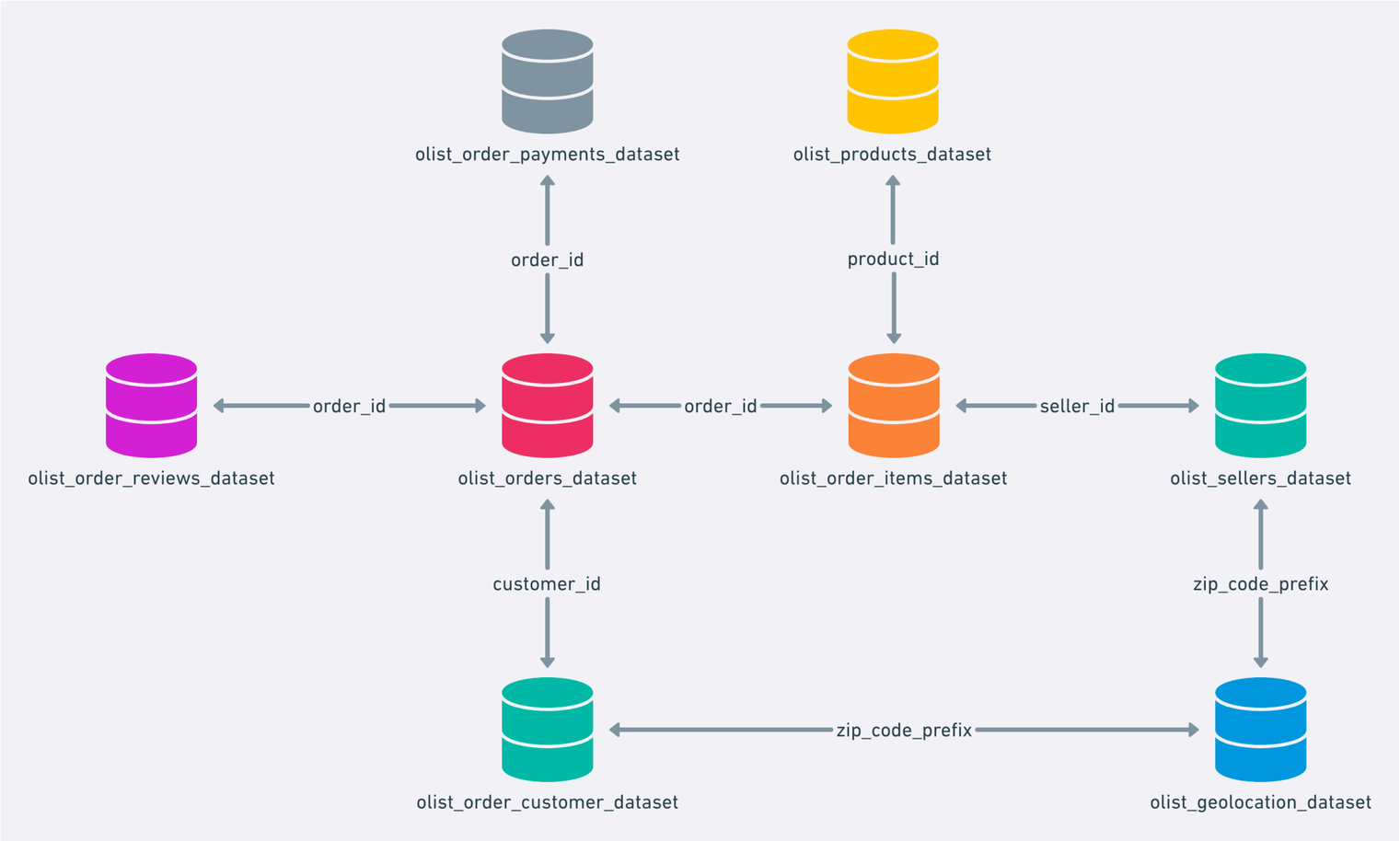
olist

- Objectifs :
  - Créer une segmentation des clients en fonction de leur **comportement d'achat** et de leurs **données personnelles**
  - Fournir une description actionnable des groupes de clients
  - Proposer un contrat de maintenance
- Moyens :
  - Utilisation de méthodes non supervisées pour regrouper les clients de profil similaire

B. Présentation des données

Les données (1)

- Base de données relationnelle composée de plusieurs jeux de données



Les jeux de données :

- Clients
- Vendeurs
- Géolocatisaion
- Produits
- Categories
- Commandes
- Commandes (produits)
- Commandes (paiement)
- Commandes (avis)

Les données (2)

olist

- Les données ont été regroupées en un **data set unique** en utilisant les clés primaires et étrangères qui lient les tables
- Caractéristiques de ce jeu de données :
  - 17 799 100 lignes
  - 45 variables

C. Nettoyage et *features engineering*

Actions diverses avant sélection

olist

- Données manquantes ?
  - Les méthodes `.info()` et `.describe()` nous indiquent que le jeu de données ne présente pas de valeurs manquantes
- Ajout de variables :
  - ‘same_state’: variable booléenne qui renvoie 1 si l’acheteur et le vendeur habitent dans le même état et 0 dans l’autre cas
  - ‘volume’ : longueur * hauteur * largeur
- Types de variables :
  - Les variables relatives aux dates sont encodées en tant que *object*, elles ont été transformées en *datetime*

Pré-sélection de variables

- Liste des 19 variables sélectionnées parmi les 45
  - **Client** : ’customer_unique_id', 'customer_city', ‘customer_state’, ’geolocation_lat', 'geolocation_lng'
  - **Commande** : ’order_id’, 'order_purchase_timestamp', ‘order_status', 'seller_id', ‘price', ‘review_score’, ‘same_state'
  - **Produits** : ‘product_id', ‘product_category_name_english’
  - **Livraison** : ’freight_value’, ’volume', 'product_weight_g', 'delivery_time'
  - **Paiement** : 'payment_type', ‘payment_installments'

Modifications

- Simplification et homogénéisation des **catégories** (De 71 à 22)
- Regroupement sur un plan sémantique, par exemple :
  - Les catégories 'books_general_interest', 'books_technical', ‘books_imported' ont été regroupé sous la catégorie ‘books’
- Objectif : créer des groupes de tailles homogènes et réduire le nombre de catégories pour faciliter l’analyse

Création des variables : (1)

De nouvelles variables on été créées en effectuant des *groupby* sur l'ID client

RFM (Récence, Fréquence, Montant) :

- Panier moyen
- Dernier achat en nombre de jours
- Total par client
- Fréquence d'achat
- *Note : le panier moyen et le total par client sont identiques lorsqu'un client n'a effectué qu'un seul achat*

Création des variables (2)

olist

- Note moyenne attribuée
- Volume moyen des colis
- Poids moyen des colis
- Nombre de paiements moyen
- Type de paiement le plus commun
- Temps de livraison moyen
- Frais de port moyen
- Distance vendeur-acheteur la plus commune(‘same_state’)

D. Exploration

L’index

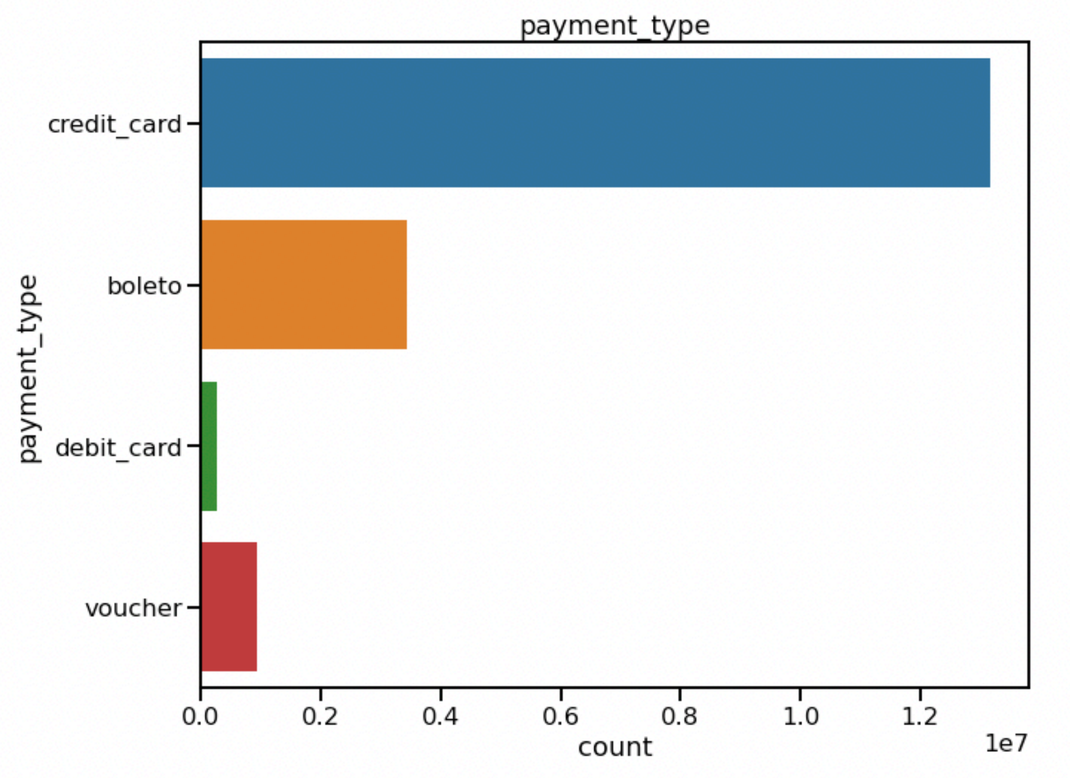
olist

- Deux candidats :
  - ‘customer_id’ : 96987 valeurs uniques. L’identifiant client est variable en fonction de la commande
  - ‘customer_unique_id’ : 93829 valeurs uniques. L’identifiant est unique.
- ‘customer_unique_id’ est donc retenu en tant qu’index

Les paiements

olist

Répartition des types de paiement



Proportion par nombre de paiements

Nombre de paiements	Proportion
1	50,3 %
2	11,3 %
3	10 %
4	6,6 %
10	5,9 %

Fréquence d’achat

olist

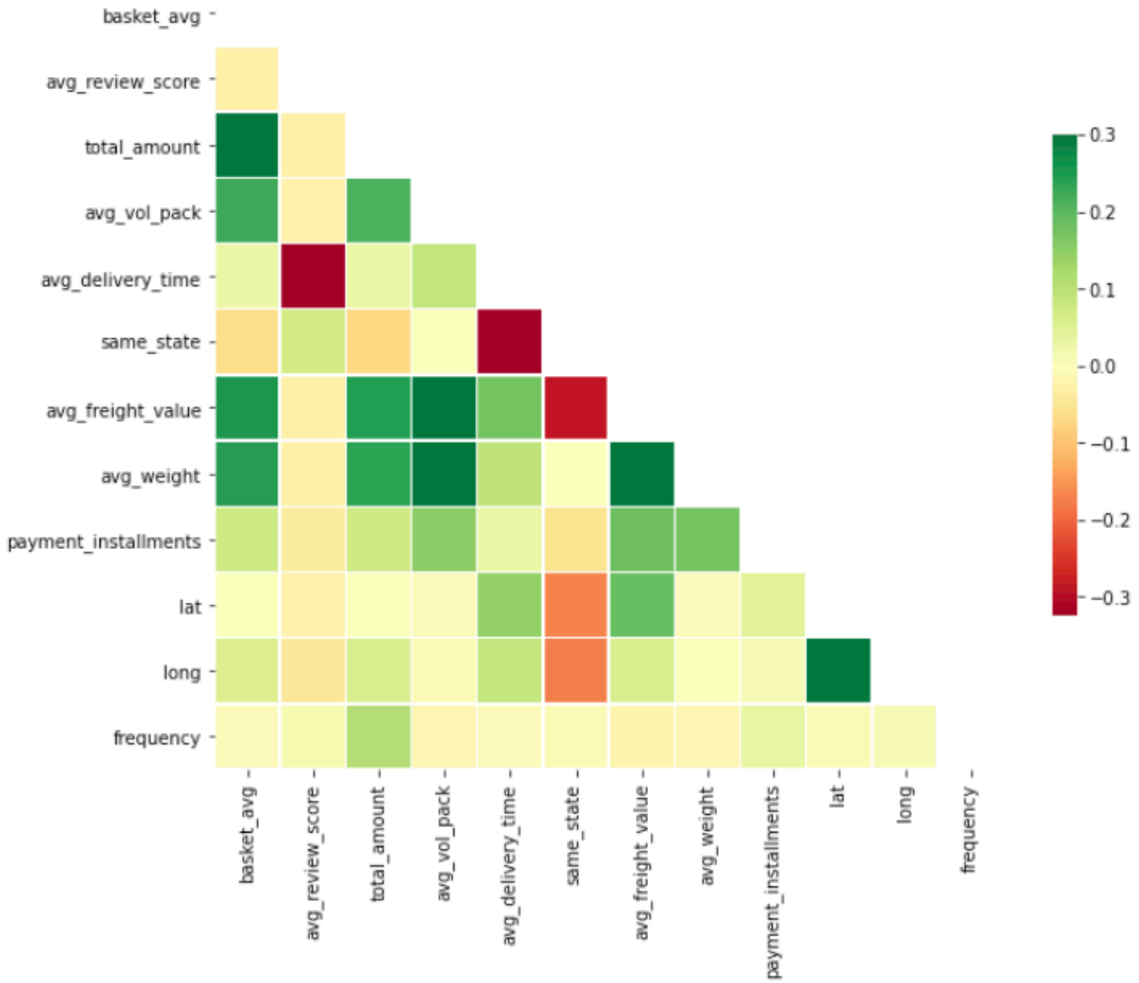
La proportion des clients par fréquence d’achat

Fréquence	Proportion
1	96,97 %
2	2,78 %
3	0,19 %
4	0,03 %

Corrélations entre les variables

olist

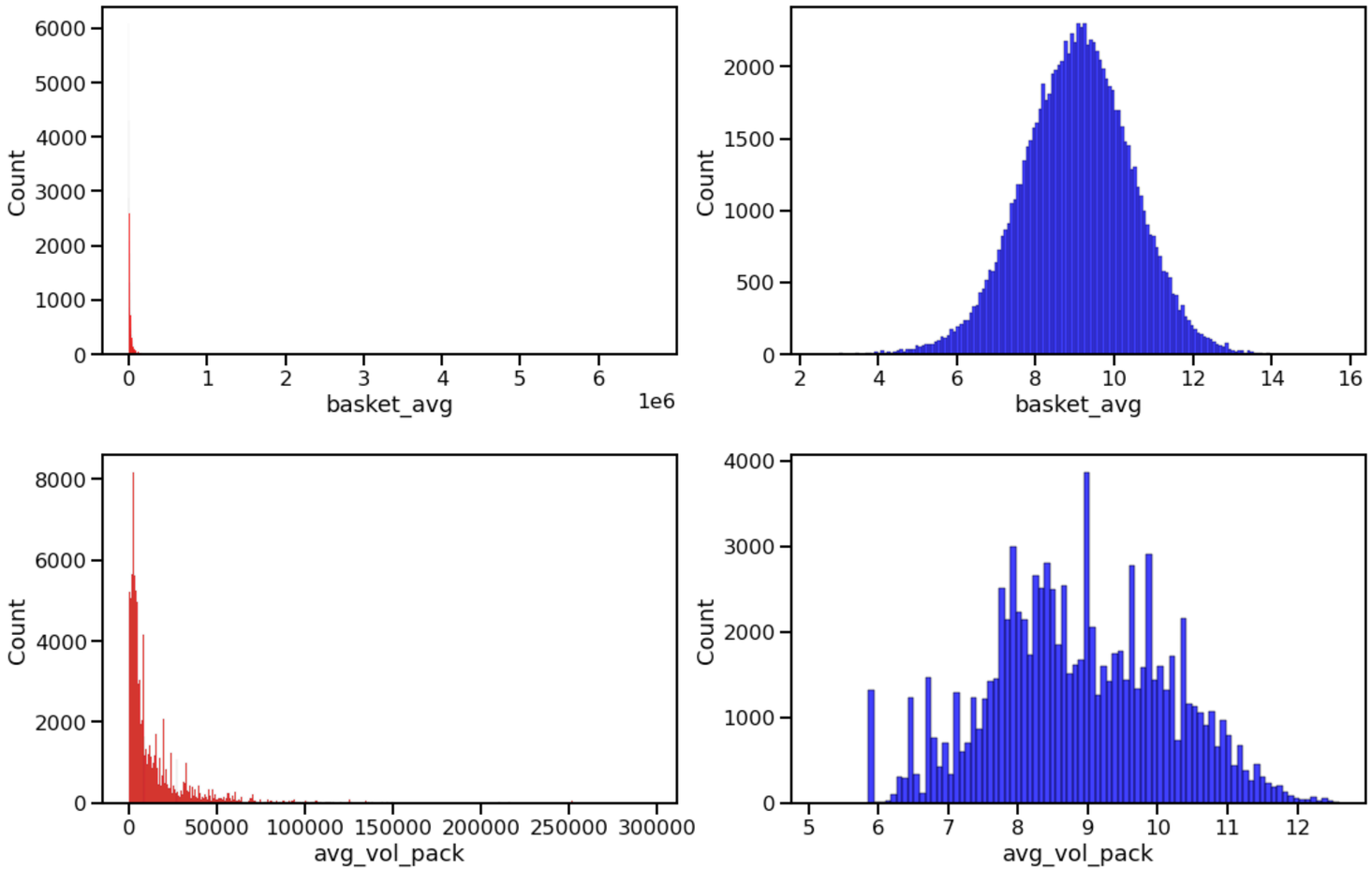
Heatmap des corrélations



- Corrélations positives :
  - Panier moyen et montant des commandes
  - Poids/Dimensions du colis et montant des frais de port
- Corrélations négatives :
  - Temps de livraison et note
  - Temps de livraison et la variable booléenne ‘same_state’
  - Montant des frais de port et la variable booléenne ‘same_state’

Distribution des données numériques

Distribution des variables rouge = données originales et bleu = données avec transformation de log x +1



E. Les modélisations

Avant-propos

olist

- Etant donné que le jeu de données est assez **volumineux**, je décide d'**éliminer le regroupement hiérarchique** qui n'est pas performant pour ce type de dataset
- Deux algorithmes de clustering seront utilisés :
  - **K-means**
  - DBScan (ou plus exactement **OPTICS** un algorithme proche de DBscan mais plus efficace pour les jeux de données important)
- Afin de visualiser les clusters, j'utiliserai deux algorithmes de réduction de dimensions :
  - L'Analyse en composantes principales (**ACP**)
  - **T-SNE**

Preprocessing

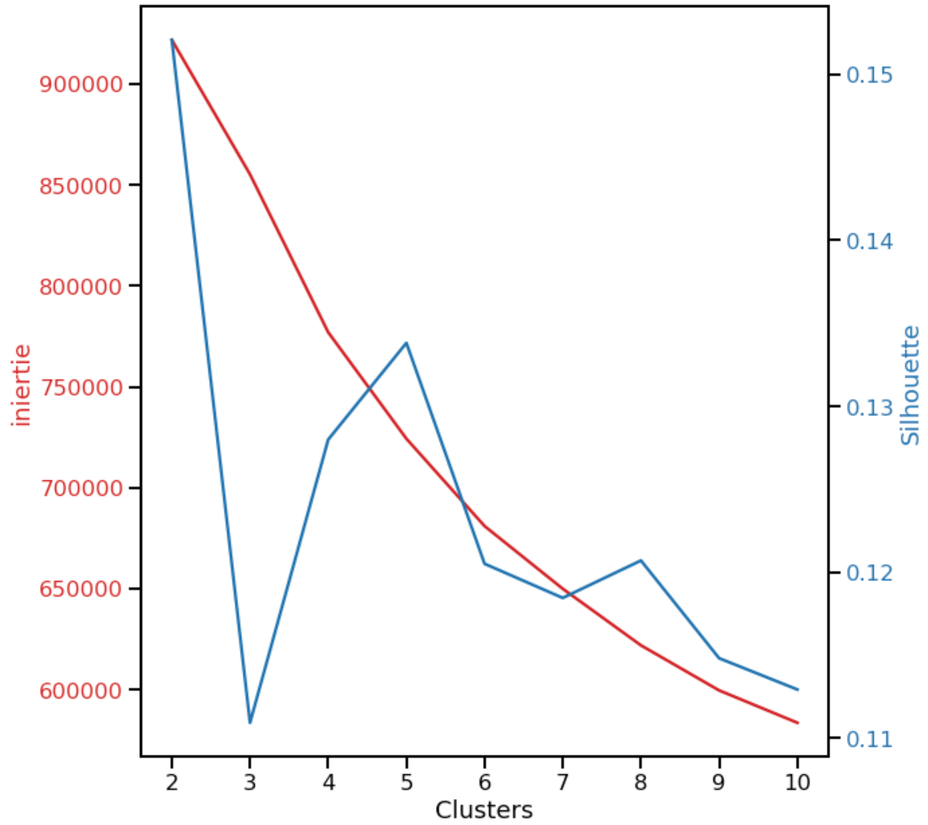
olist

- Variables numériques :
  - Pipeline :
    1. Valeur Absolue
    2. Log(x+1)
    3. Standard Scaler
- Variables catégorielles nominales
  - One Hot Encoding

K-means

olist

Silhouette et inertie



- La courbe de l'inertie n'a pas de coude évident
- On distingue **deux pics** sur la courbe des scores silhouette : 5 et 8 clusters
- Dans le but d'identifier des profils d'utilisateurs plus fins pour une segmentation **marketing**, je décide de choisir un nombre de clusters de 8

K-means

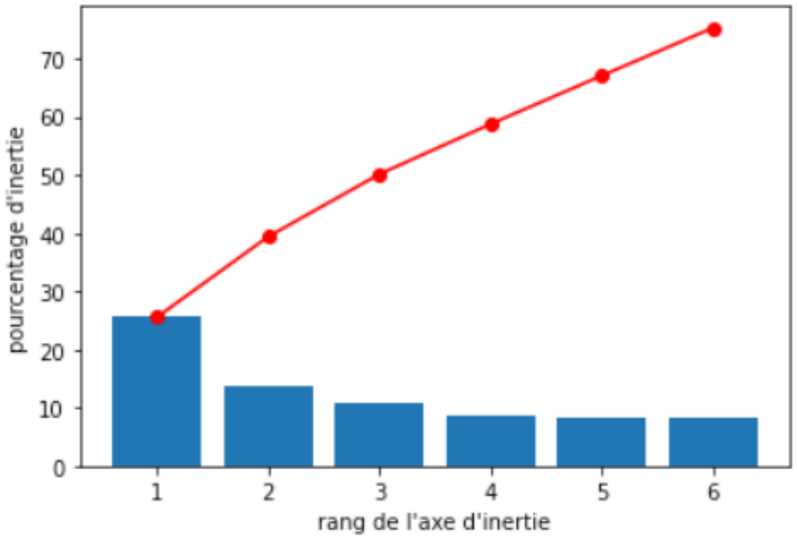
olist

Taille et score silhouette des clusters

- Avec le paramètre n_clusters = 8 :
  - Les groupes ont des tailles similaires
  - Chacun des groupes à un score supérieur à la moyenne

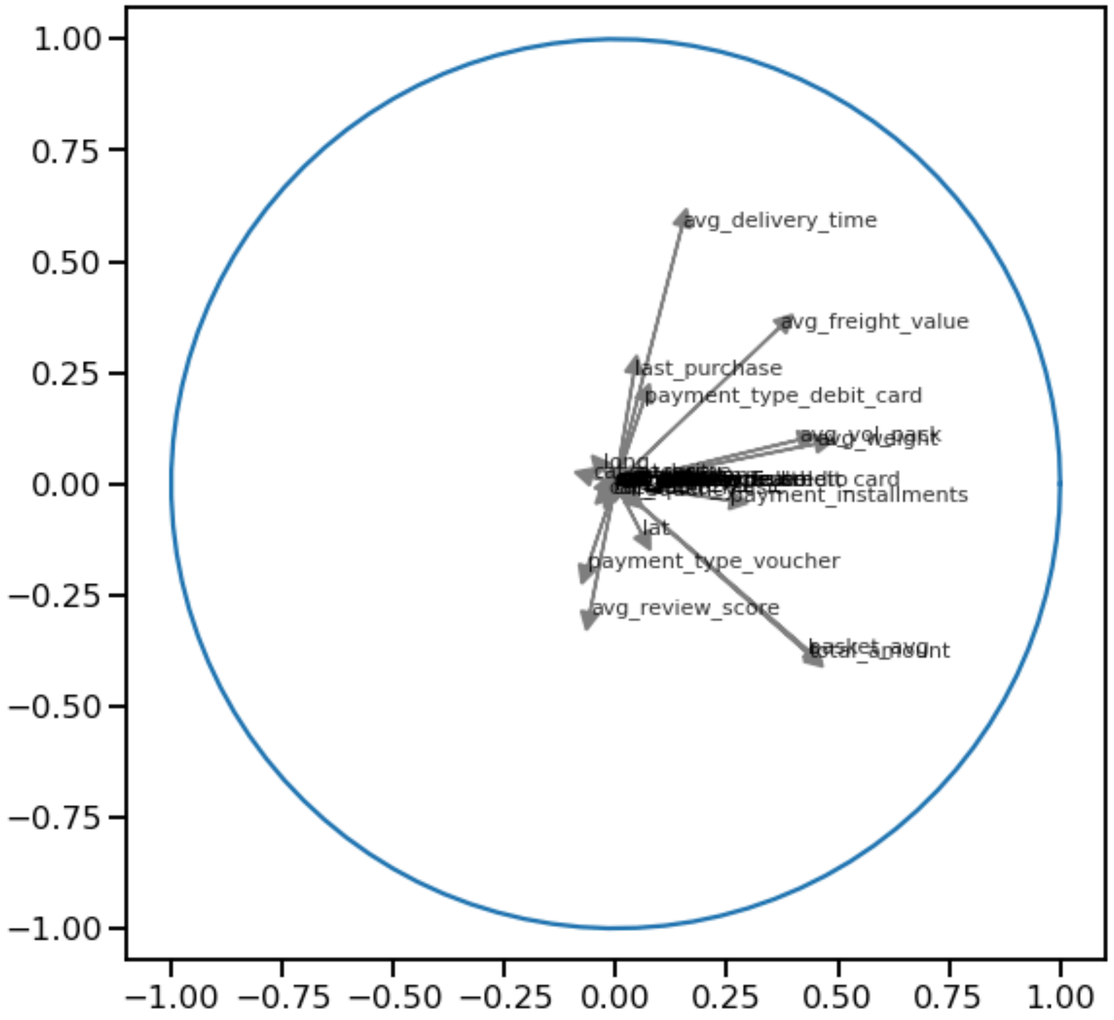
Analyse en Composantes Principales

Eboulis des valeurs propres



L’éboulis des valeurs propres indique que le premier plan factoriel explique environ **40%** de la variance

Analyse en Composantes Principales

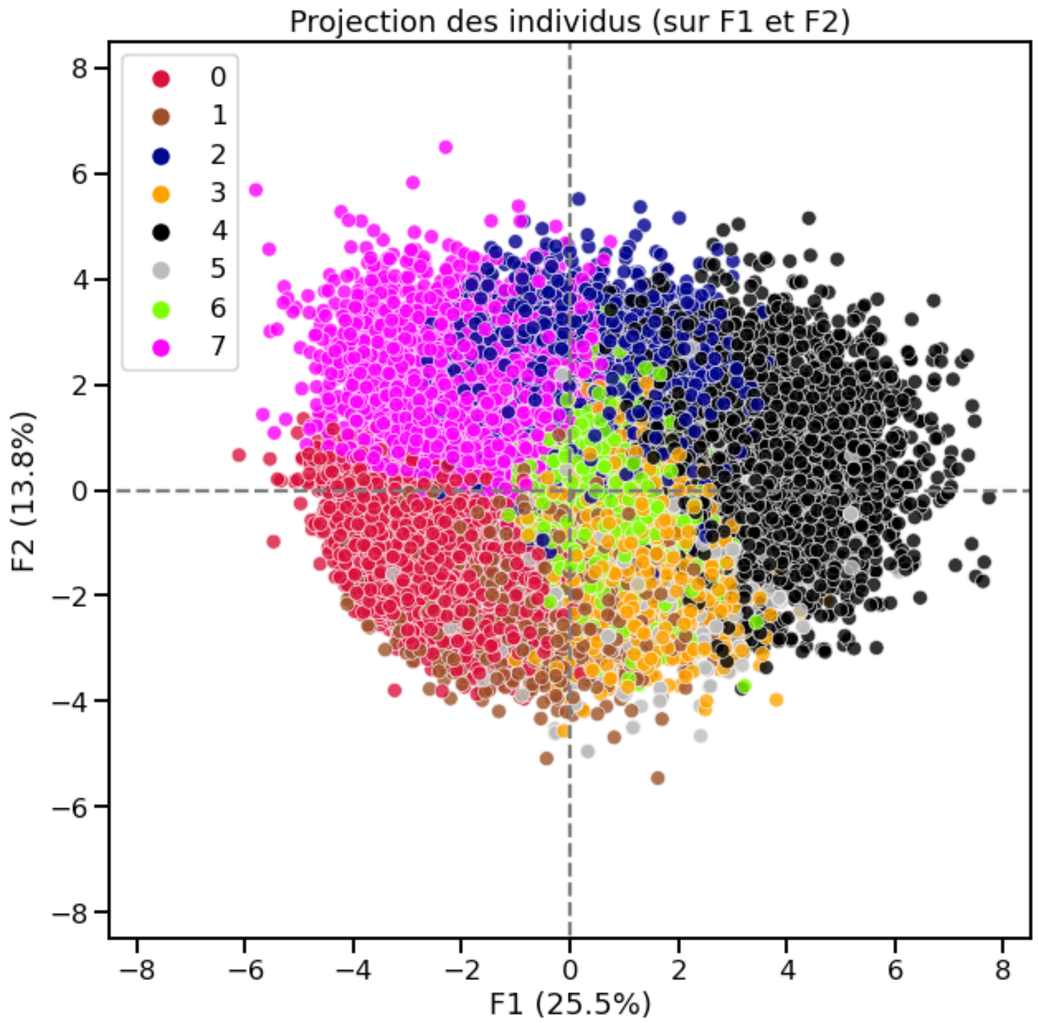


- Les variables les mieux représentées sur le cercles des corrélations sont :
  - Le panier moyen
  - le montant total des achats
  - Le temps de le livraison
  - Les frais de ports

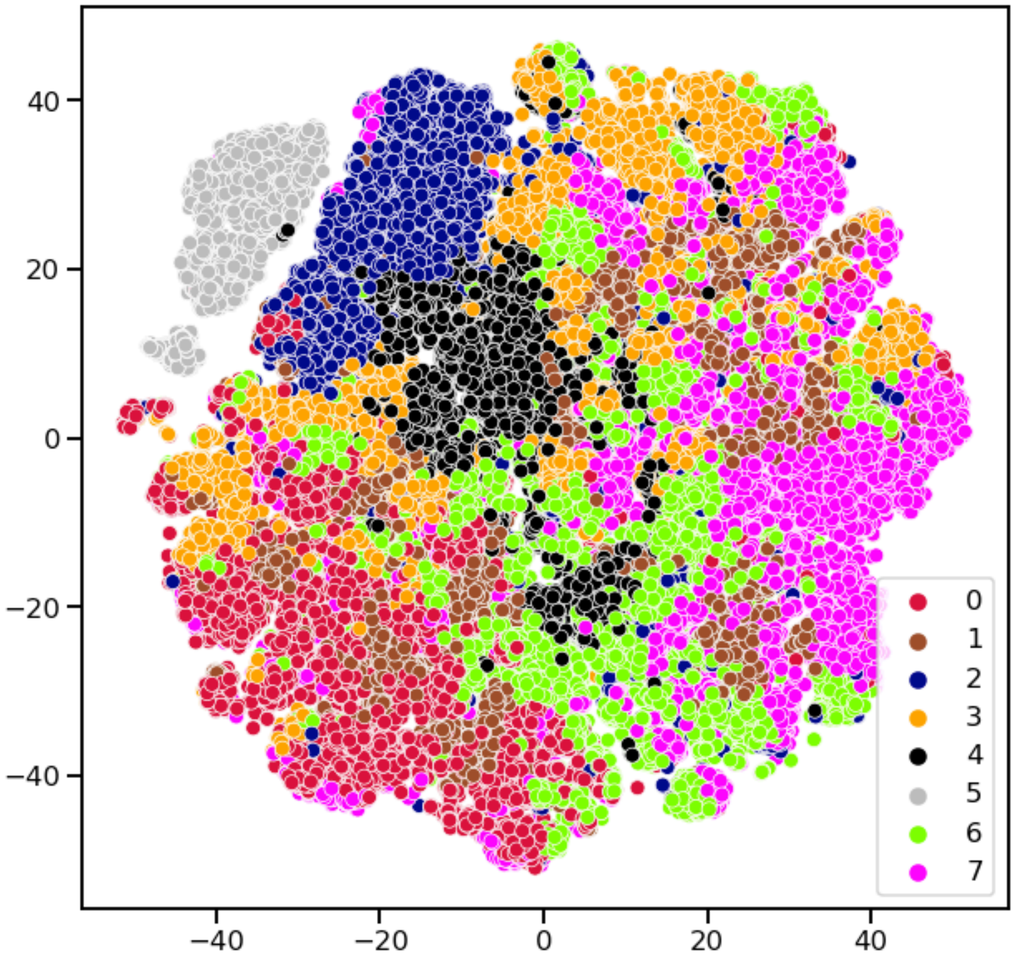
K-means - visualisation des clusters

olist

Premier plan factoriel de l'ACP



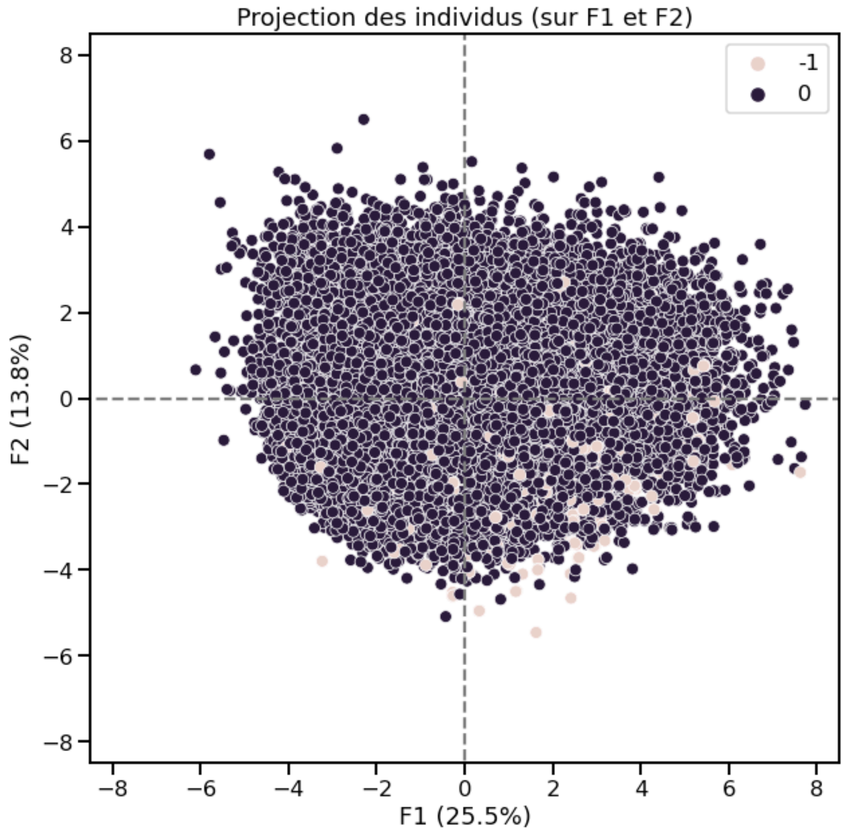
TSNE à deux dimensions



DBScan

olist

Visualisation des clusters du DBScan sur le premier plan factoriel de l'ACP



- J'ai utilisé l'implémentation **OPTICS** sous sklearn, qui est un algorithme proche de DBScan mais qui plus adapté au dataset volumineux
- En utilisant des hyperparamètres standards, l'algorithme n'a pas réussi à séparer les clients en des clusters exploitables

Choix du modèle

olist

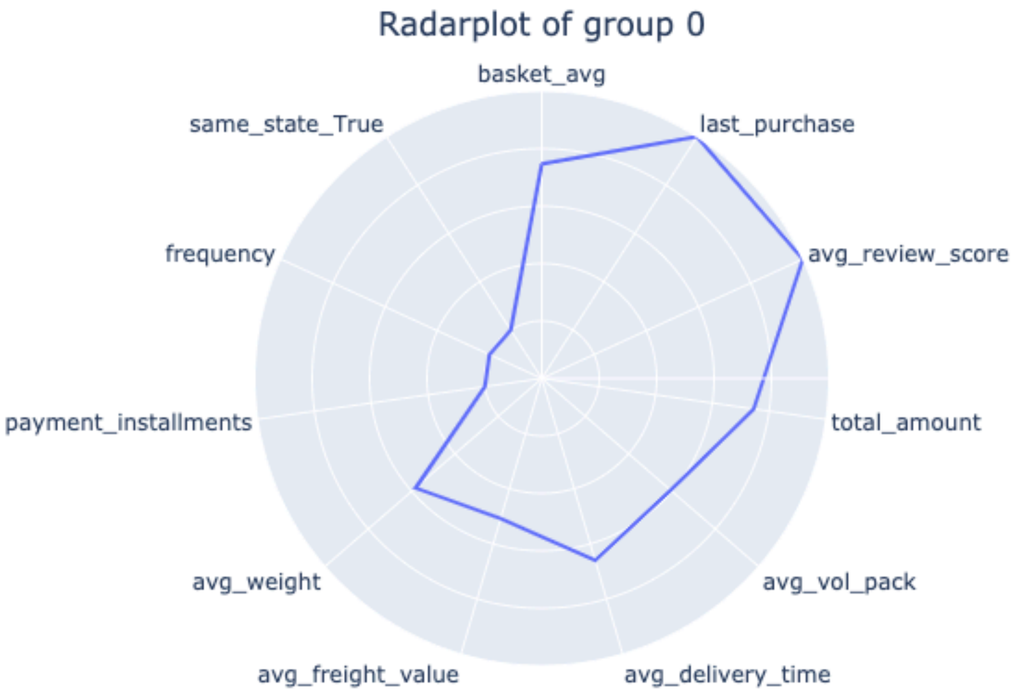
- Le **K-means** permet d'identifier plusieurs groupes tandis que l'implémentation du DBscan n'a pas été concluante
- C'est donc le K-means que je sélectionne
- Caractéristiques de l'algorithme :
  - **Distance et similarités** : la distance euclidienne a été retenu
  - **Forme des clusters** : grâce au score à l'utilisation du score silhouette, j'ai sélectionné un nombre de clusters qui tendent à être resserrés sur eux mêmes et loins les uns des autres
  - **Stabilité des clusters** : les paramètres *init='k-means++'* et *n_init=10* permettent de contrôler la stabilité des clusters

F. Analyse des groupes

Analyse du groupe 0

olist

Radar Plot du groupe 0

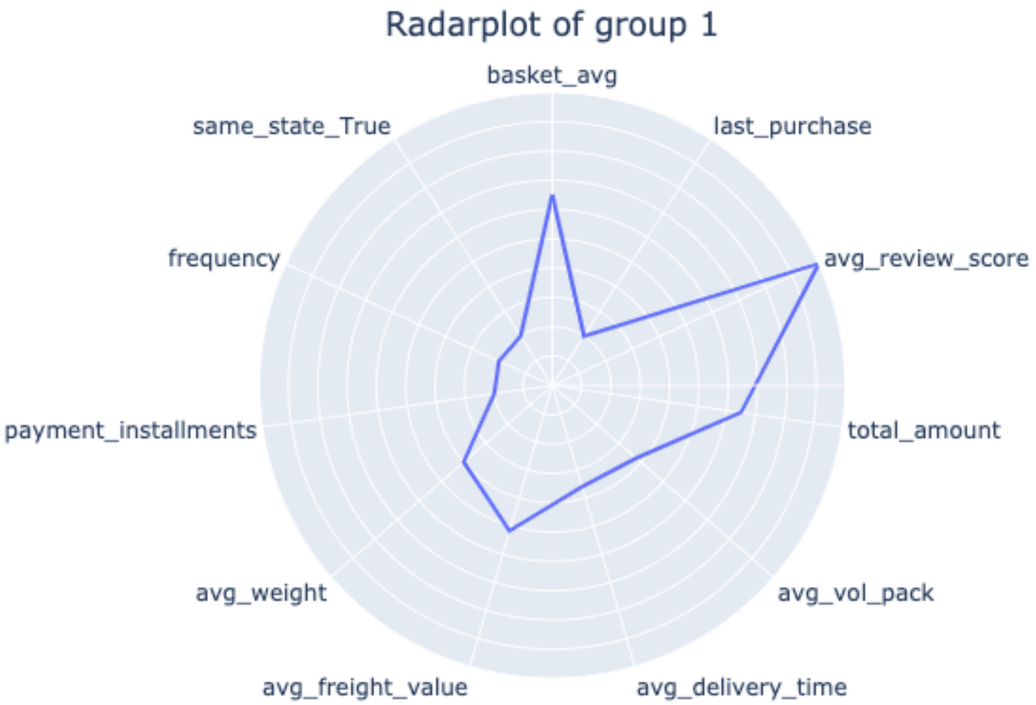


- Variable(s) de distinction :
  - last_purchase
- Description :
  - Ancien client, ils ont effectué leur dernier achat il y a longtemps sans effectuer une deuxième commande
- Proposition d’action :
  - Leur présenter les nouveautés

Analyse du groupe 1

olist

Radar Plot du groupe 1

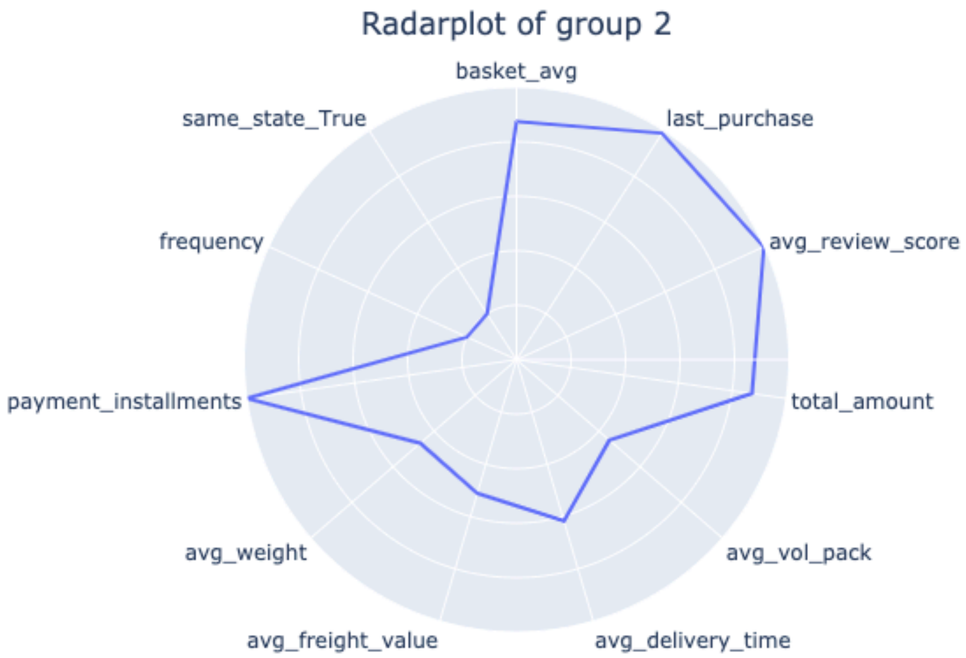


- Variable(s) de distinction :
  - last_purchase
- Description :
  - Nouveau client, leur dernier achat est très récent
- Proposition d’action :
  - Effectuer une enquête de satisfaction

Analyse du groupe 2

olist

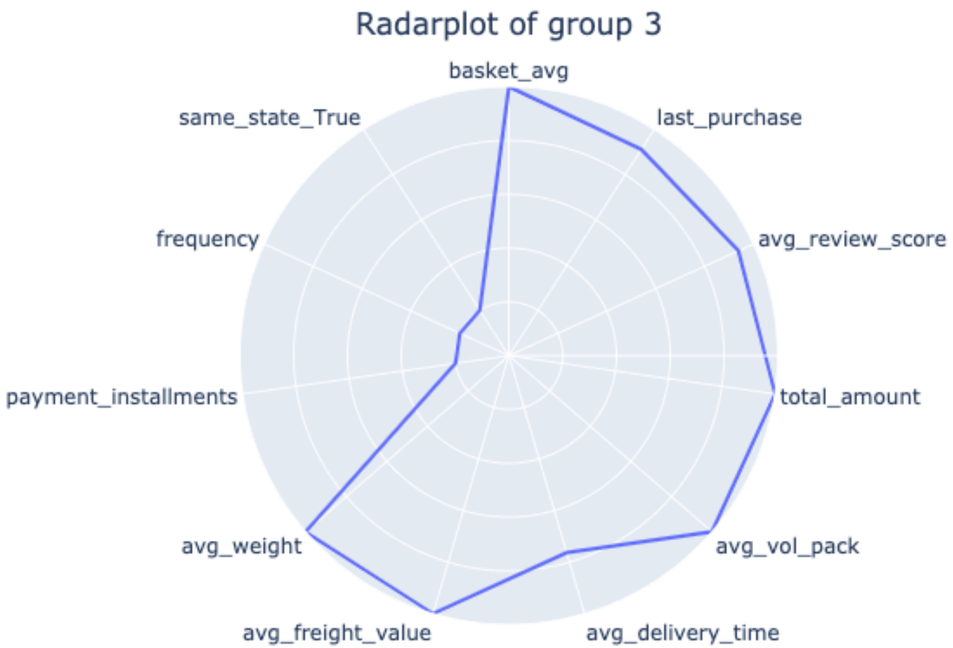
Radar Plot du groupe 2



- Variable(s) de distinction :
  - payment_installments
- Description :
  - Les clients ayant payés en plusieurs fois
- Proposition d’action :
  - Leur proposer des facilités de paiement

Analyse du groupe 3

Radar Plot du groupe 3

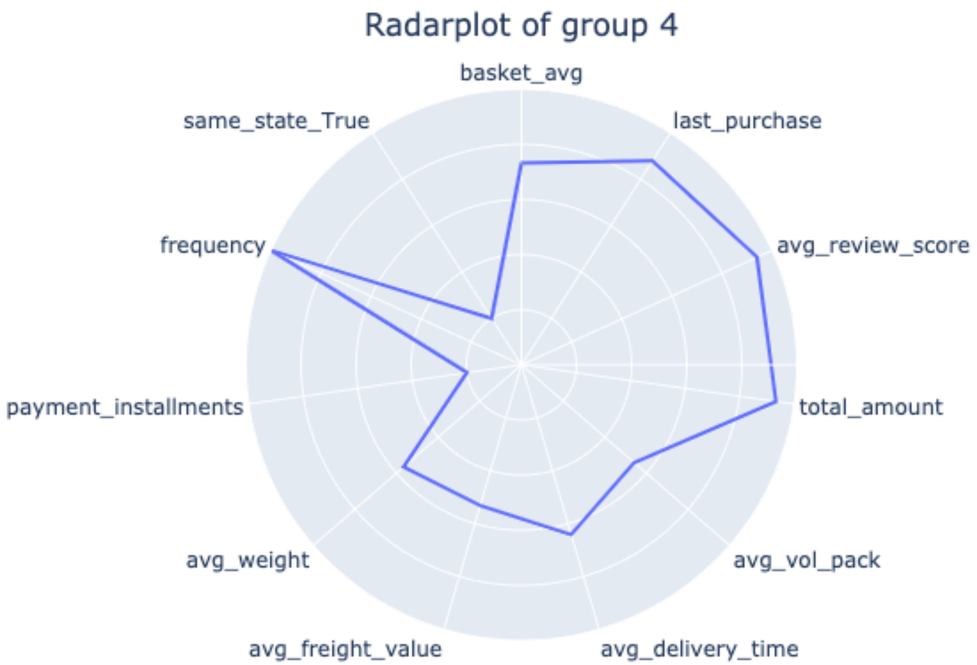


- Variable(s) de distinction :
  - basket_avg
  - total_amount
  - avg_weight
  - avg_freight_value
- Description :
  - Clients ayant effectués des achats onéreux. Leur colis est volumineux et les frais de port élevés
- Proposition d'action :
  - Invitation à des événements premiums (vente privée exclusive)

Analyse du groupe 4

olist

Radar Plot du groupe 4

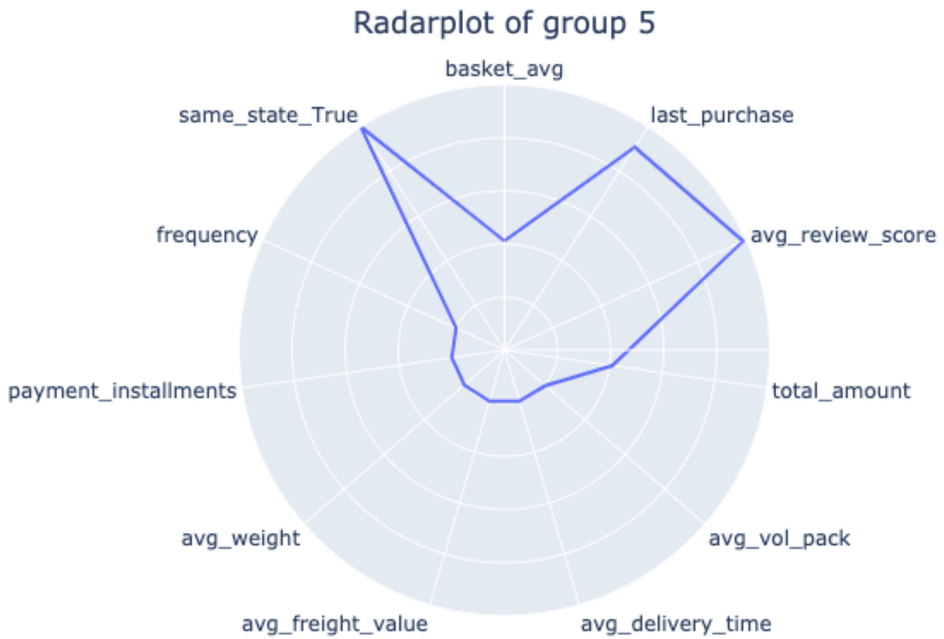


- Variable(s) de distinction :
  - Frequency
- Description :
  - Clients ayant effectués plusieurs achats
- Proposition d’action :
  - Leur proposer un abonnement annuel à la plate-forme

Analyse du groupe 5

olist

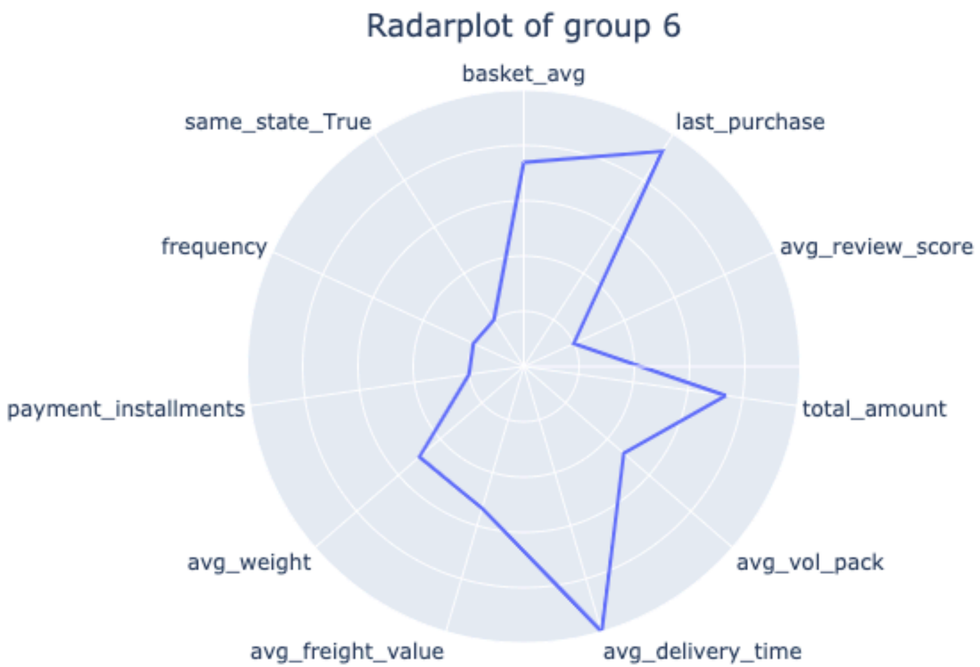
Radar Plot du groupe 5



- Variable(s) de distinction :
  - same_state
- Description :
  - Clients ayant effectués des petites commandes près de chez eux
- Proposition d'action :
  - Proposer un nouveau service de vente/ location d'objet pour répondre à des urgences

Analyse du groupe 6

Radar Plot du groupe 6

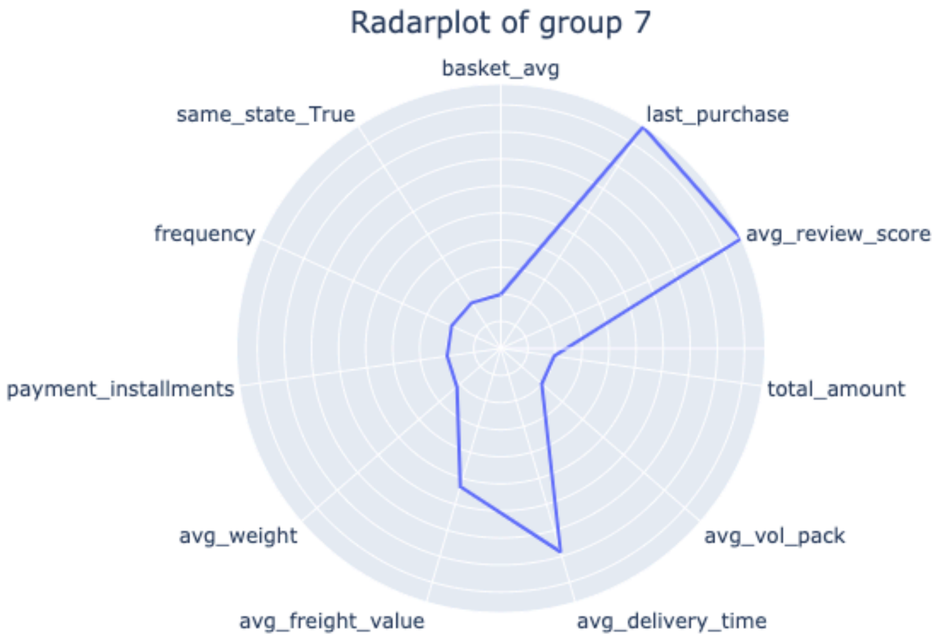


- Variable(s) de distinction :
  - avg_vol_pack
- Description :
  - Clients ayant attribués des mauvaises notes, leur commande a en général des délais de livraisons qui sont longs
- Proposition d’action :
  - Effectuer une enquête de satisfaction afin de comprendre l’origine de leur insatisfaction (Qu’est-ce-qui à causer ce long délai de livraison)

Analyse du groupe 7

olist

Radar Plot du groupe 7

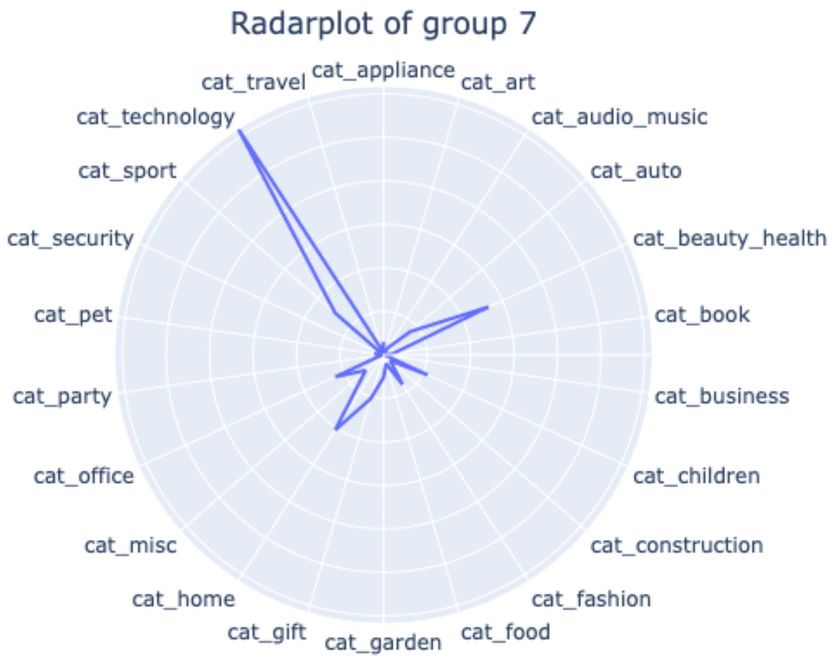


- Variable(s) de distinction :
  - basket_avg
  - total_amount
- Description :
  - Clients ayant effectués des petits achats
- Proposition d’action :
  - Communiquer sur des catégories de produits adaptés

Analyse du groupe 7 : catégories

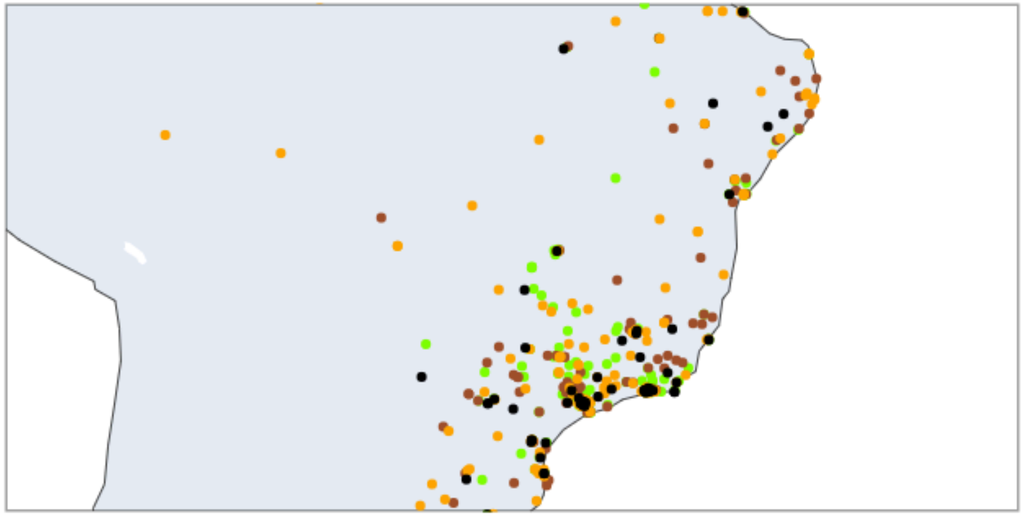
olist

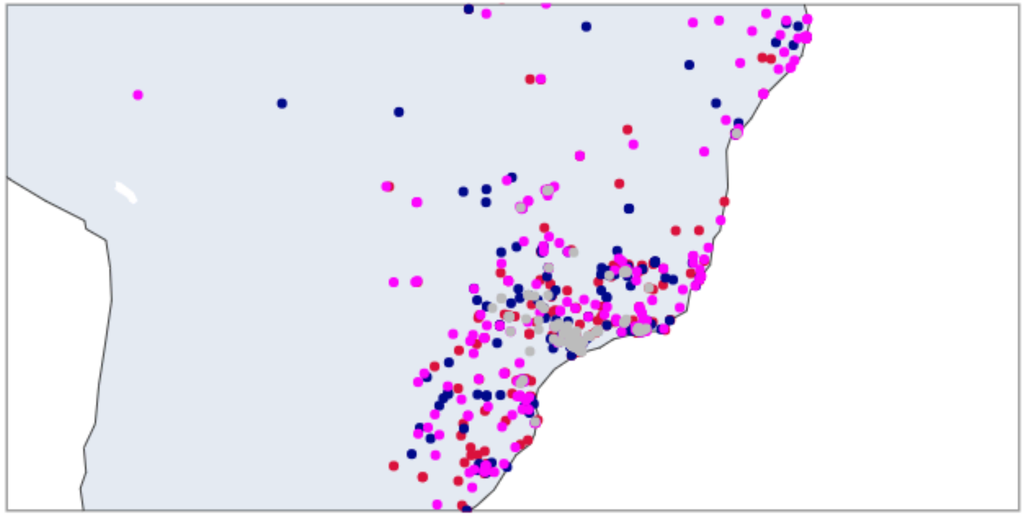
Radar Plot du groupe 7



- Catégorie préférentielle :
  - Technology

Origine géographique





group
- 0
- 6
- 2
- 1
- 7
- 3
- 5
- 4

Quelques tendances :

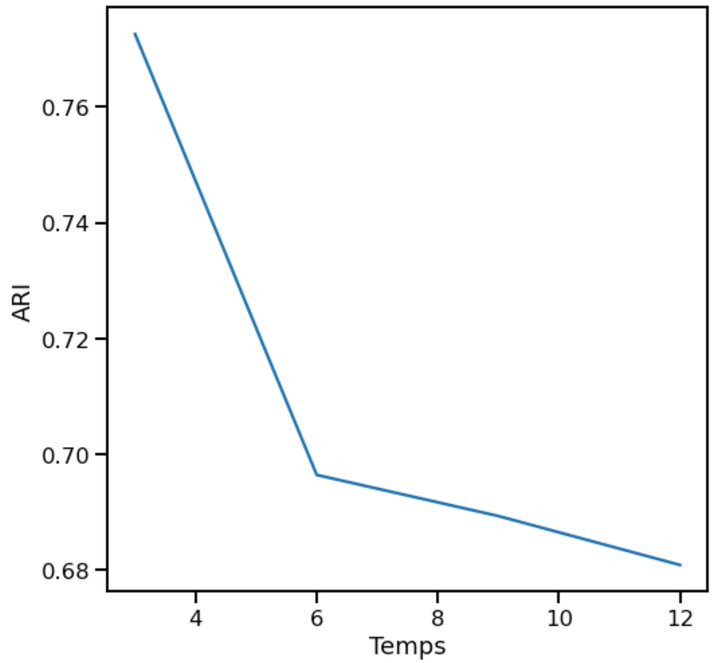
- Le groupe n°5 *(achats près de chez soi)* est concentré autour d'une ville
- Les clients du groupe n°3 *(achats onéreux)* ont des points isolés

G. Maintenance et stratégie pour les nouveaux clients

Contrat de maintenance

olist

Evolution du score ARI dans le temps



Qu’est-ce que le score ARI ?

L’**Adjusted Rand Index** est un score compris entre 0 et 1 qui permet de comparer la similitude de deux clusterings. Dans ce cas, on peut l’utiliser pour comparer notre clustering à t et t +1.

Graphiquement, on constate que le score ARI se dégrade dans le temps

Afin de préserver la qualité du modèle, on peut par exemple le mettre à jour tous les 3, 6 ou 9 mois en fonction du niveau de précision souhaité.

Stratégie pour les nouveaux clients

olist

- Les nouveaux clients seront attachés au groupe dont le **centroïde** est le plus proche grâce au **KNN**

- Cependant, comme il existe un groupe qui caractérise les nouveaux clients, je vais donc utiliser le KNN sans la dimension ‘last_purchase’

Conclusion

olist

- Le clustering des données effectué à l’aide du K-means, nous a permis de d'identifier 8 groupes d’utilisateurs ayant des comportements distinctifs
- Ces clients sont potentiellement actionnables par les équipes marketing de l’Olist
- En modifiant le nombre de clusters, il est possible d’ajuster la finesse de la segmentation des clients

Pistes d’amélioration

olist

- Effectuer une analyse fine des variables de distinction des groupes
- Pousser l’analyse géographique des clusters (relief, infrastructure, etc.)
- Optimiser la recherche d’hyperparamètres pour l’OPTICS