SEGMENTATION DES CLIENTS DU SITE E-COMMERCE OLIST

PROJET N°5
PARCOURS « DATA SCIENTIST »

ETUDIANT : SAHEL TAHERIAN
MENTOR : YANNICK SERGE
EVALUATEUR : ISAAC YIMGAING

SOUTENANCE DE PROJET

14 OCTOBRE 2021

Plan de la présentation

I. Présentation de la problématique

II. Préparation des données et exploration

III. Pistes de modélisations

IV. Présentation du modèle final

Rappel de la problématique

- Mission de consultant pour Olist, site de e-commerce brésilien
- solution de vente sur les marketplaces en ligne

OBJECTIFS:

- Fournir aux équipes d'e-commerce une segmentation des clients pour les campagnes de communication
- Comprendre les différents types d'utilisateurs
- Fournir une description actionnable de la segmentation
- Analyser la stabilité au cours du temps des segments (dans le but d'établir un contrat de maintenance)

II – PRÉPARATION DU JEU DE DONNÉES

Cleaning
Feature engineering
Exploration

Customers

- customer_id
- customer_unique_id
- customer_zip_code_prefix
- customer_city
- customer_state

Geolocation

- geolocation_zip_code_prefix
- geolocation_lat
- geolocation_lng
- geolocation_city
- geolocation_state

Orders

- order_id
- customer_id
- order_status
- order_purchase_timestamp
- order_approved_at
- order_delivered_carrier_date
- order_delivered_customer_date
- order_estimated_delivery_date

Order Items

- order_id
- order_item_id
- product_id
- seller_id
- shipping_limit_date
- price
- freight_value

Order Payments

- order_id
- payment_sequential
- payment_type
- payment_installments
- payment_value

Order Reviews

- review_id
- order_id
- review_score
- review_comment_title
- review_comment_message
- review_creation_date
- review_answer_timestamp

Products

- product_id
- product_category_name
- product_name_lenght
- product_description_lenght
- product_photos_qty
- product_weight_g
- product_lenght_cm
- product_height_cm
- product_width_cm

Sellers

- seller_id
- seller_zip_code_prefix
- seller_city
- seller_state

Product Category Name Translation

- product_category_name
- product_category_name_english

Inspection de l’intégrité des données et mesures curatives

1- Renommer "customer_zip_code_prefix" et "geolocation_zip_code_prefix" pour pouvoir merger les deux dataset.

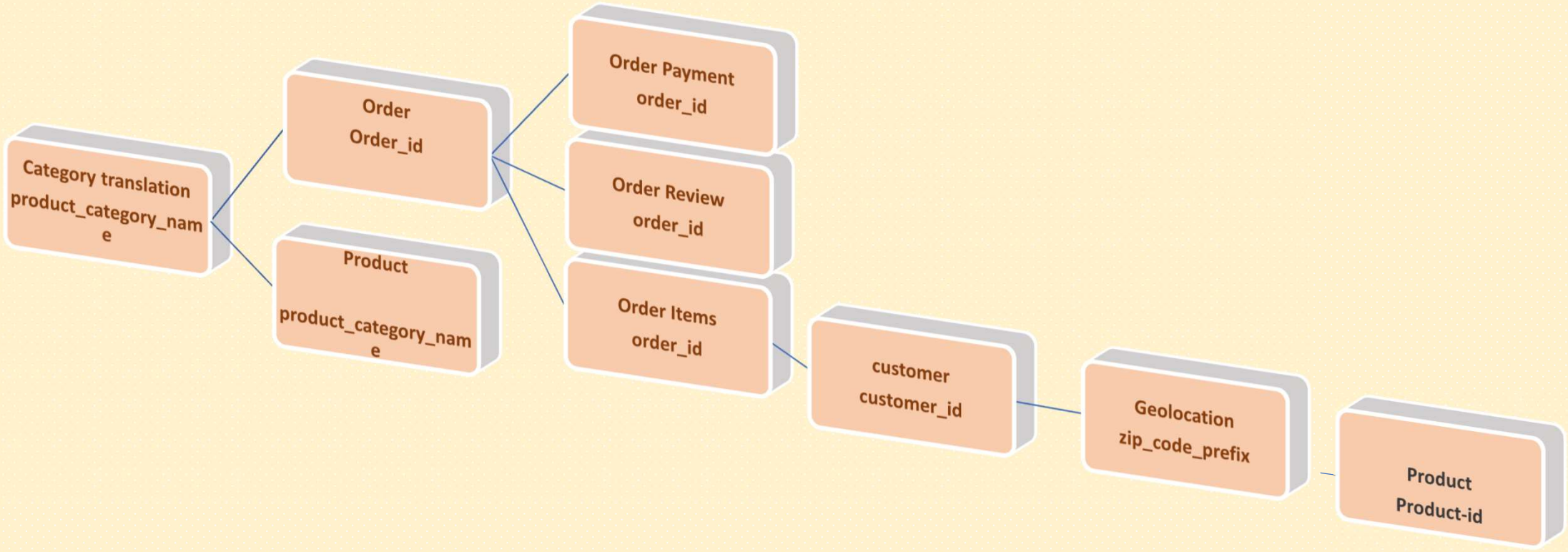
2- Suppression des lignes dupliquées :

- La présence de lignes dupliquées dans la Table de géolocalisation

3- La présence de valeurs manquantes ?

- Table des commandes
- Table des évaluations
- table des produits

Jointures des tables

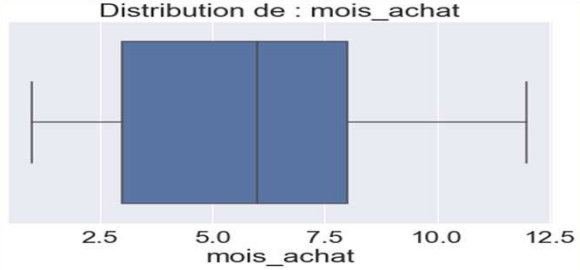


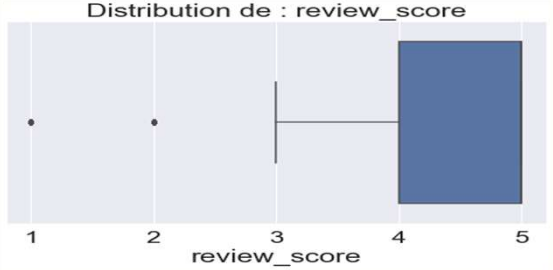
Quelques Modifications pendant les jointures des tables

- Suppression des variables inutiles comme "product_photos_qty,
- Création de variable "order_delivered" à la place de " order_status‘’
- Modification des types de données en datetime pour les colonnes temporelles
- Création des variables comme :
  - l’heure d’achat,
  - jour de la semaine d’achat,
  - mois_achat

Principales étapes du nettoyage après la jointure des tables

- **Suppression des lignes contenant des valeurs manquantes**
- **Suppression des lignes dupliquées**
- **Traitement des valeurs aberrantes**



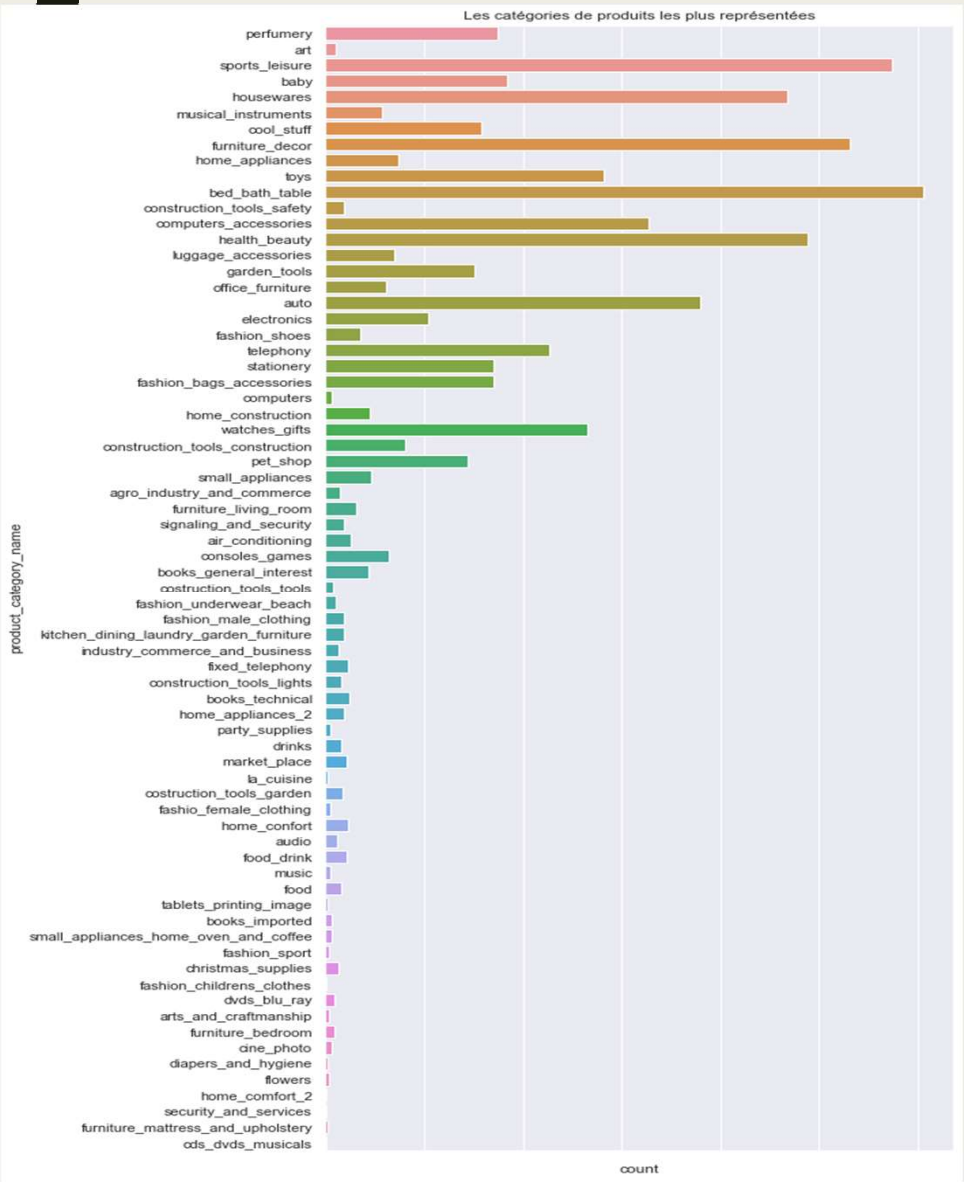


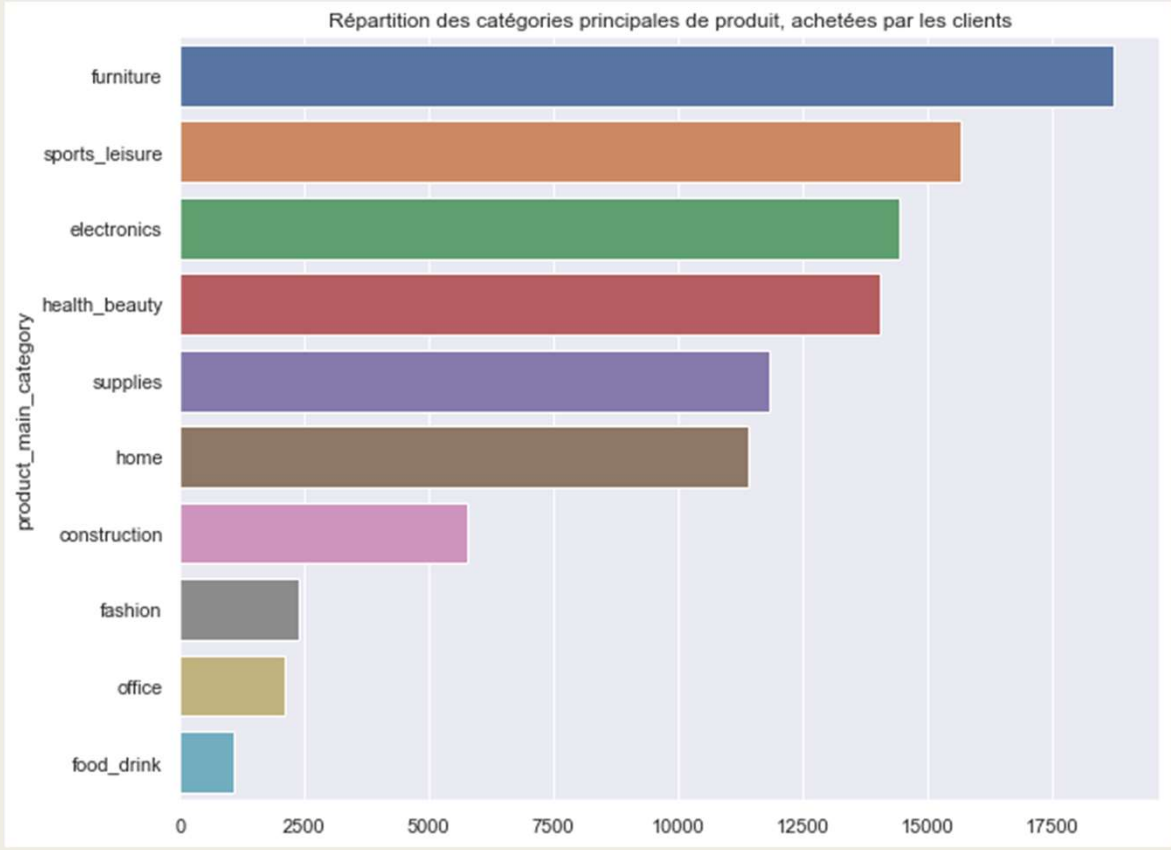
  - ***Isolation forest***
  - ***suppression de ligne : payment_installments ==0***

Feature engineering

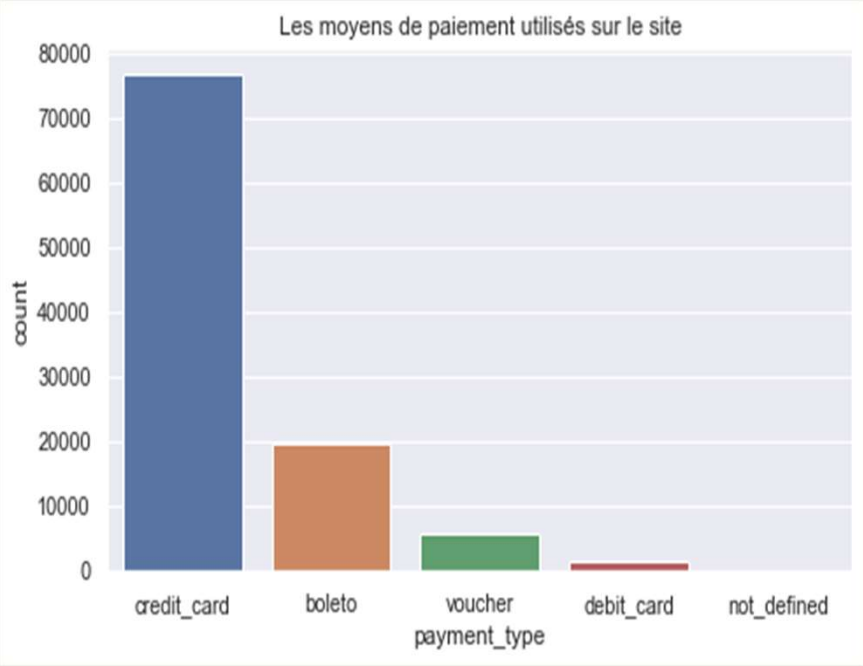
- **Réduction du nombre de catégories de produits (de 71 à 10)**
- **Suppression de 'geolocation_state' et 'geolocation_city'**
- **Création de nouvelles features :**
  - ***paid_credit_card***
  - ***first_order***
  - ***last_order***
  - ***Récence (date de la dernière commande) (R)***
  - ***Fréquence des commandes (F)***
  - ***Montant de la commande sur une période donnée(M)***
  - ***Distance Haversine entre l'état du client (moyenne des latitudes et longitudes de l'état) et le siège de Olist :***
  - ***Etc***
- ***Assemblage dans une table unique avec l'index l'id client***

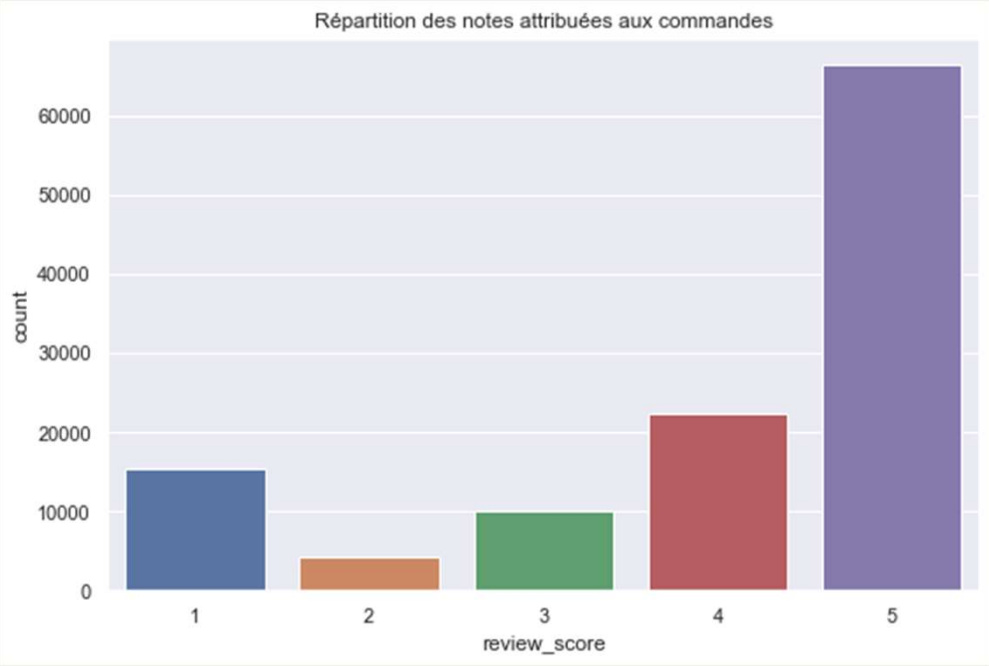
Exploration



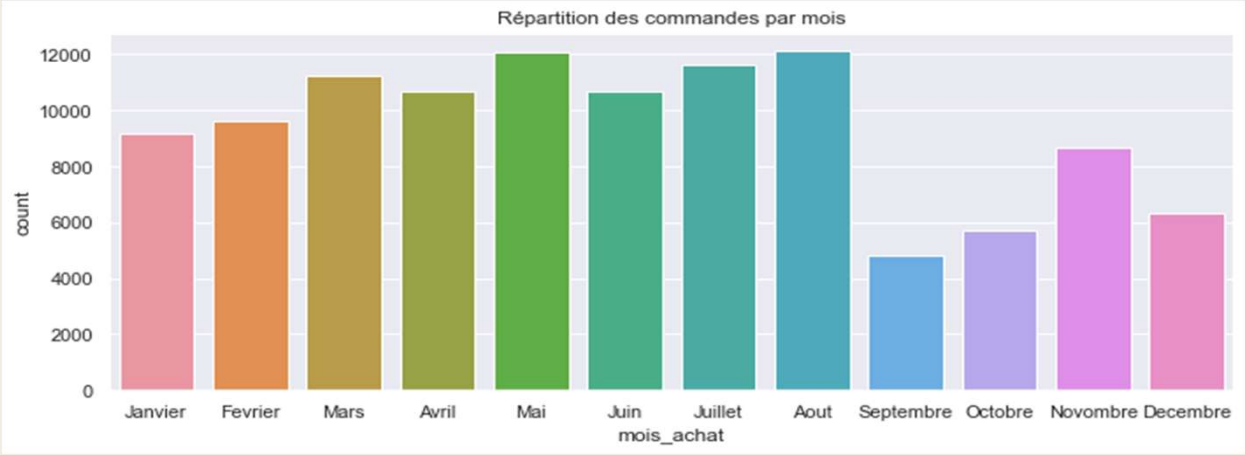


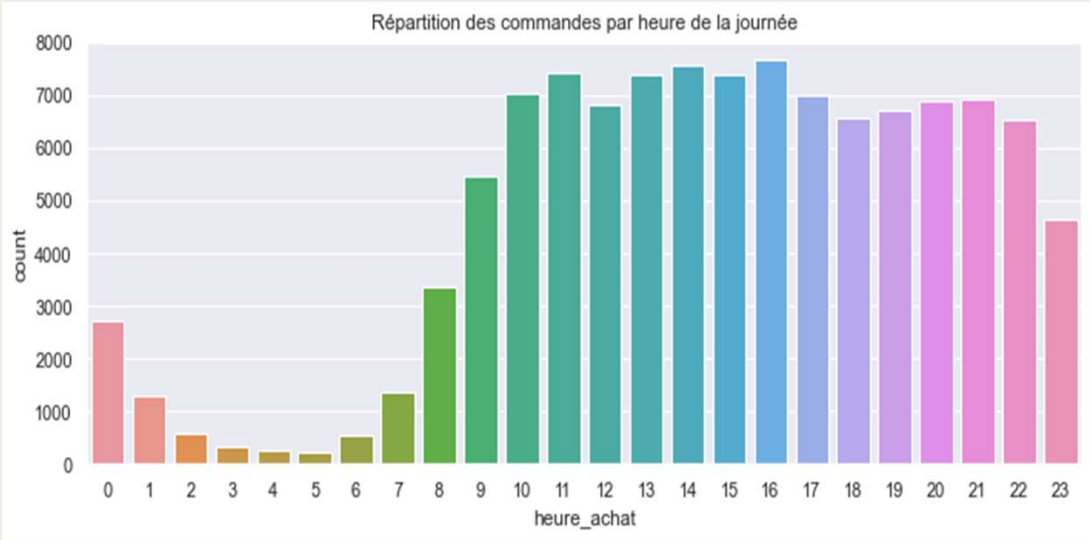
Exploration

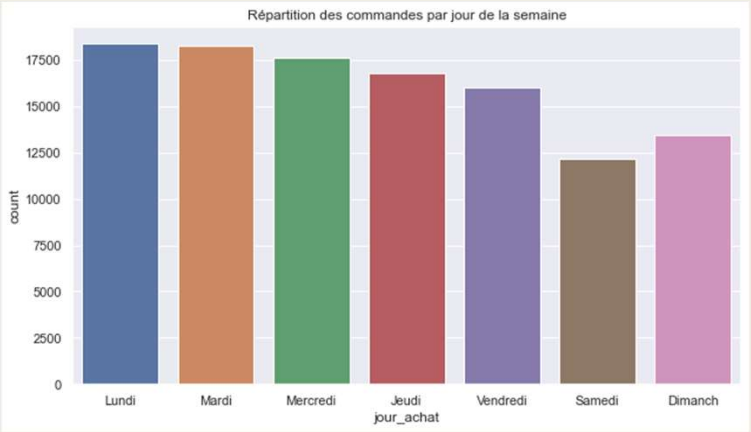




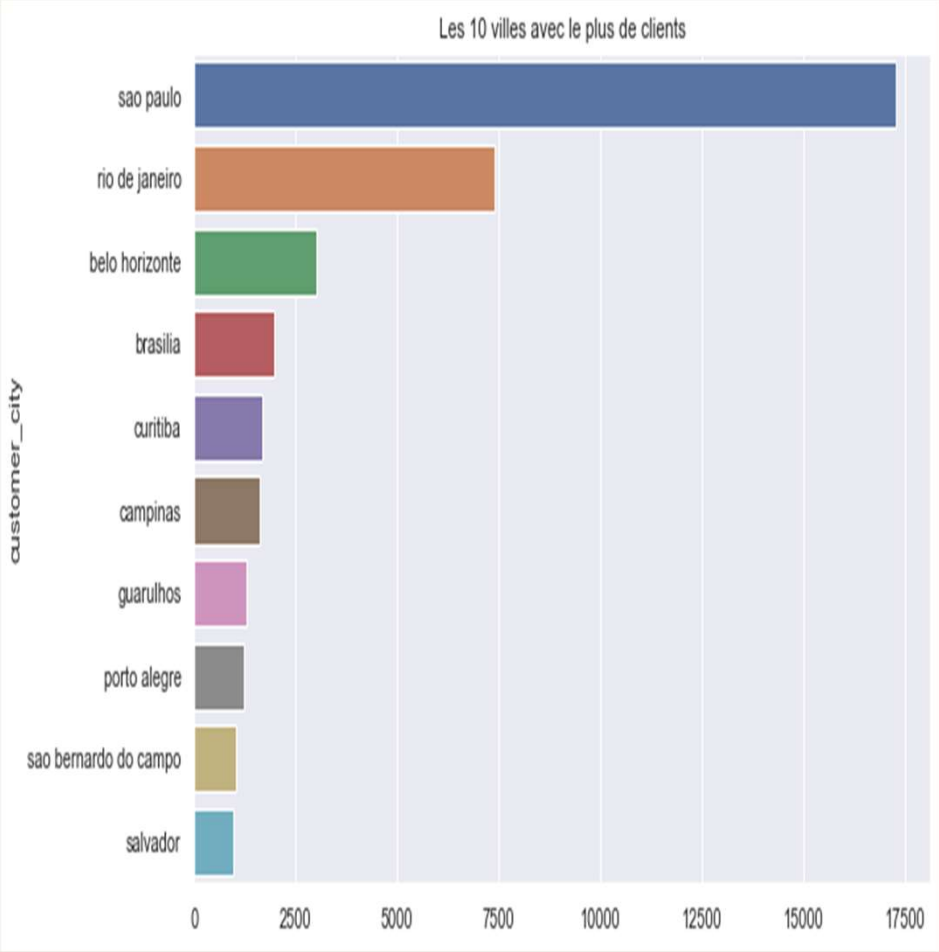
Exploration

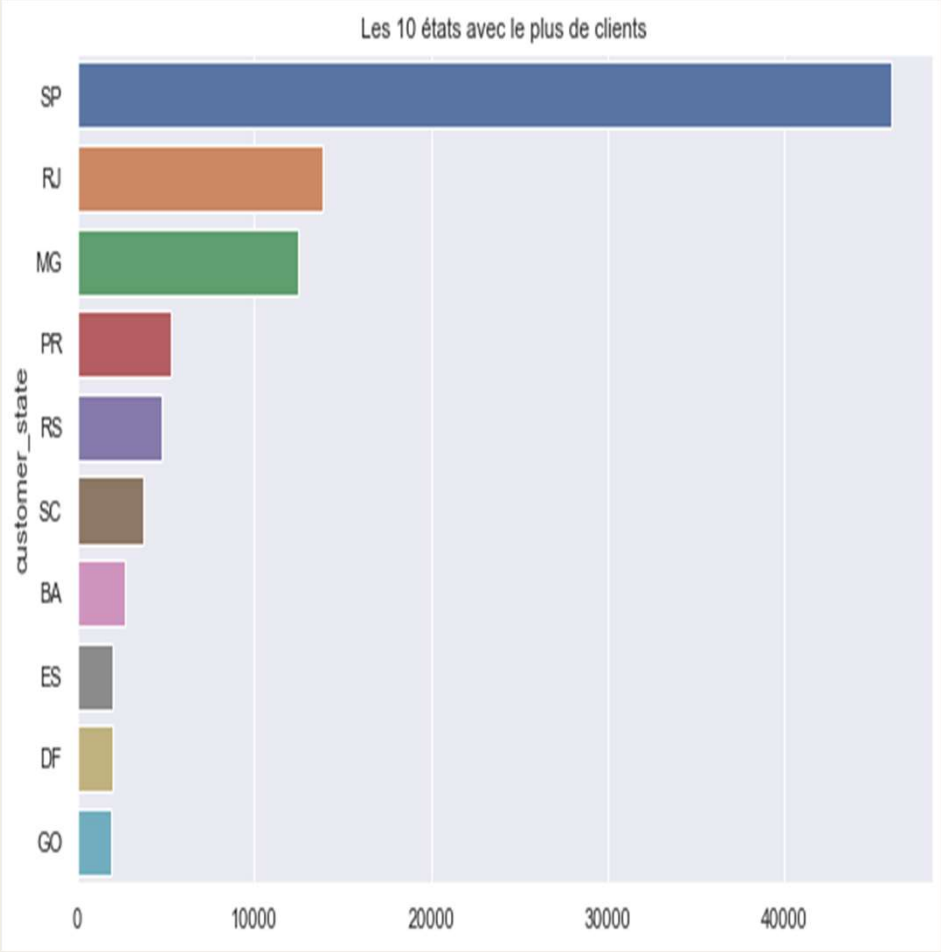






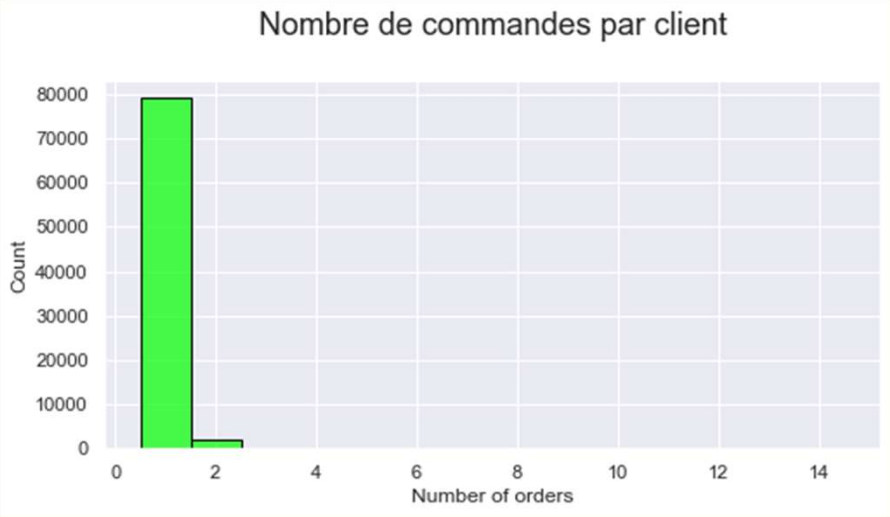
Exploration

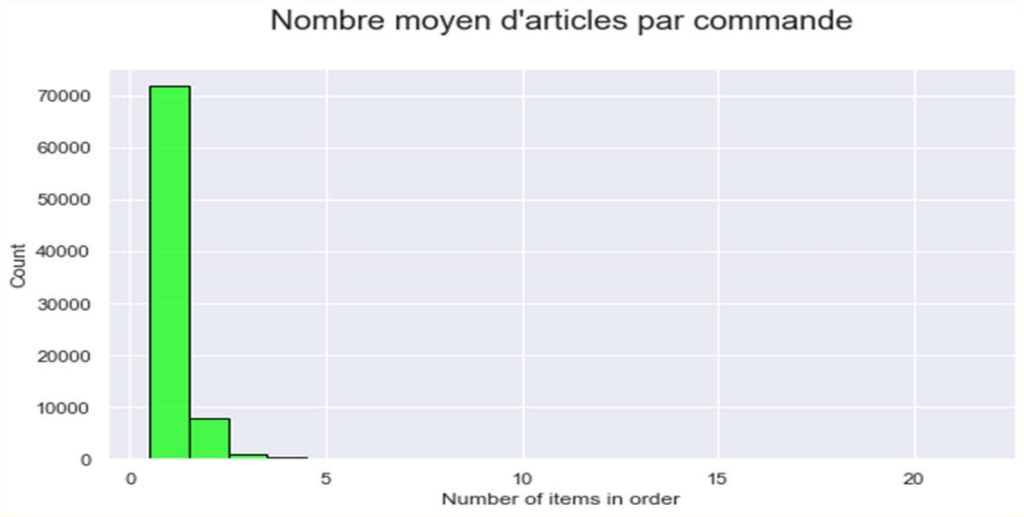




Exploration

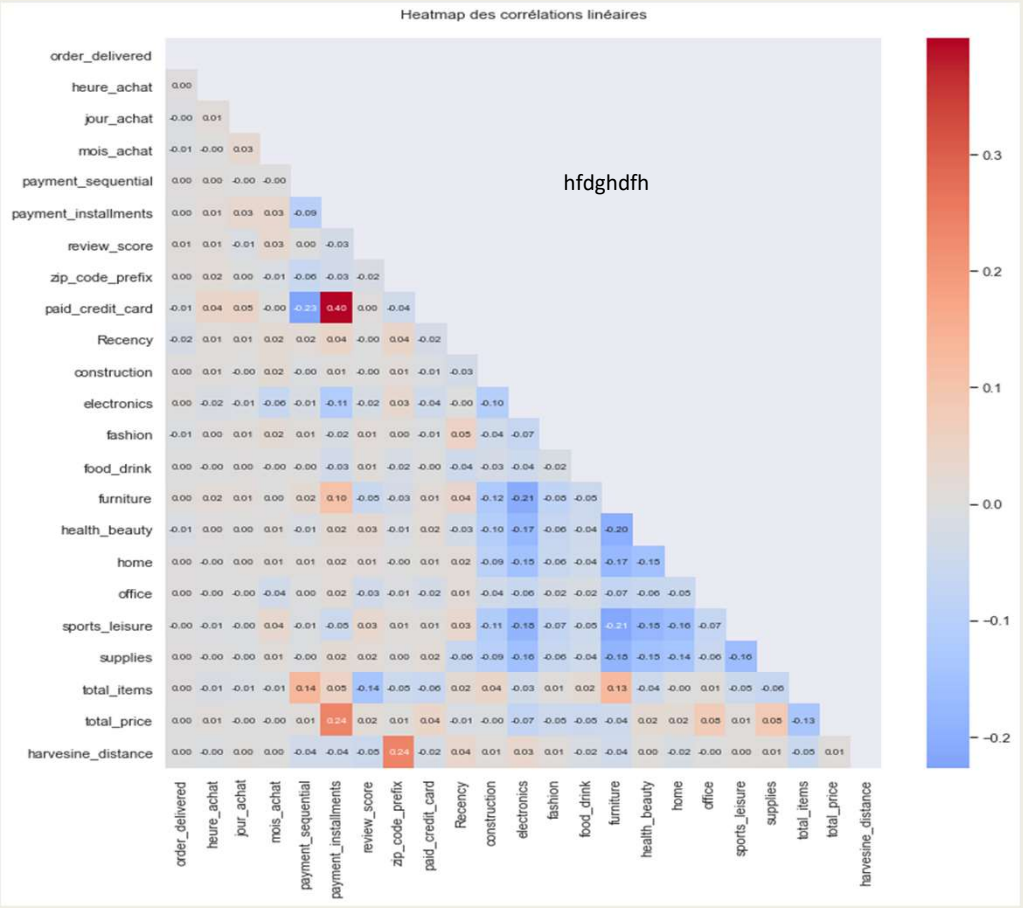
- Récence (R)- nombre de jours écoulés depuis le dernier achat
- Fréquence (F)

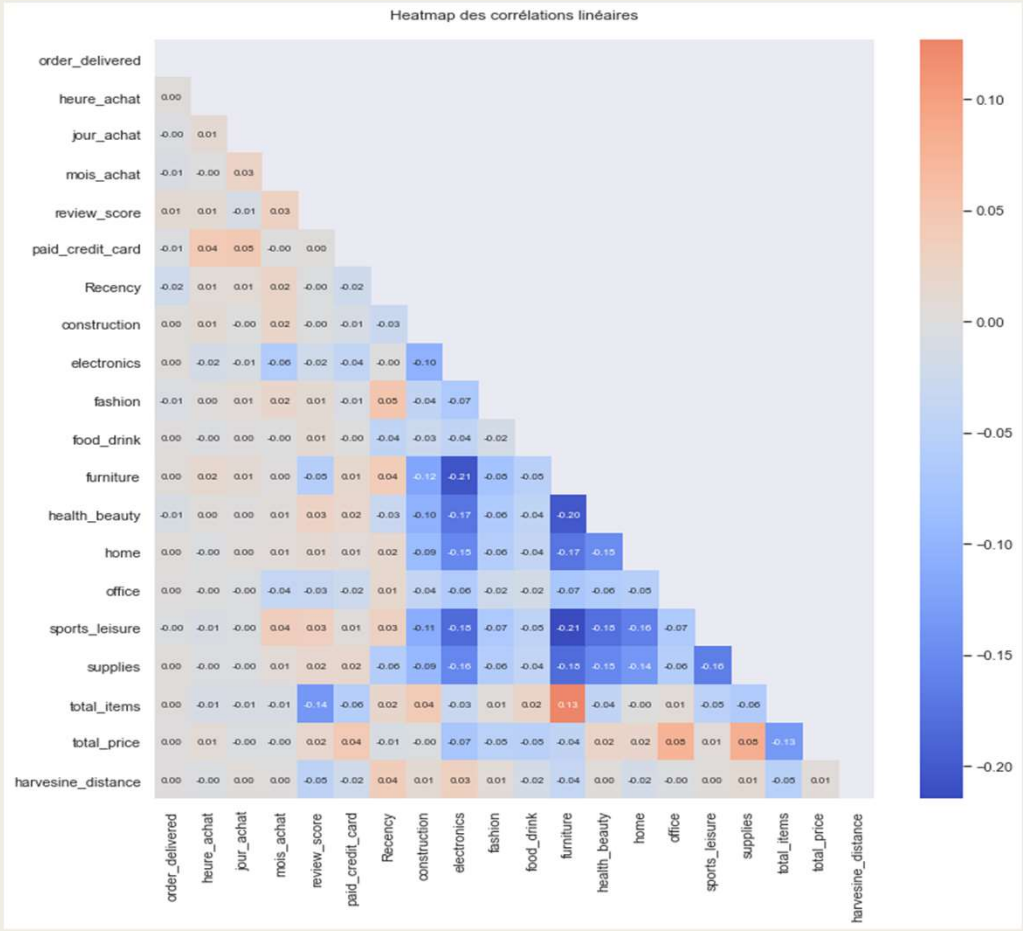




- Montant :
  - total_price = price + freight_value

Exploration





- **Suppression des variables corrélées: « payment_installments », «payment_sequential », « zip_code_prefix »**

Jeu final

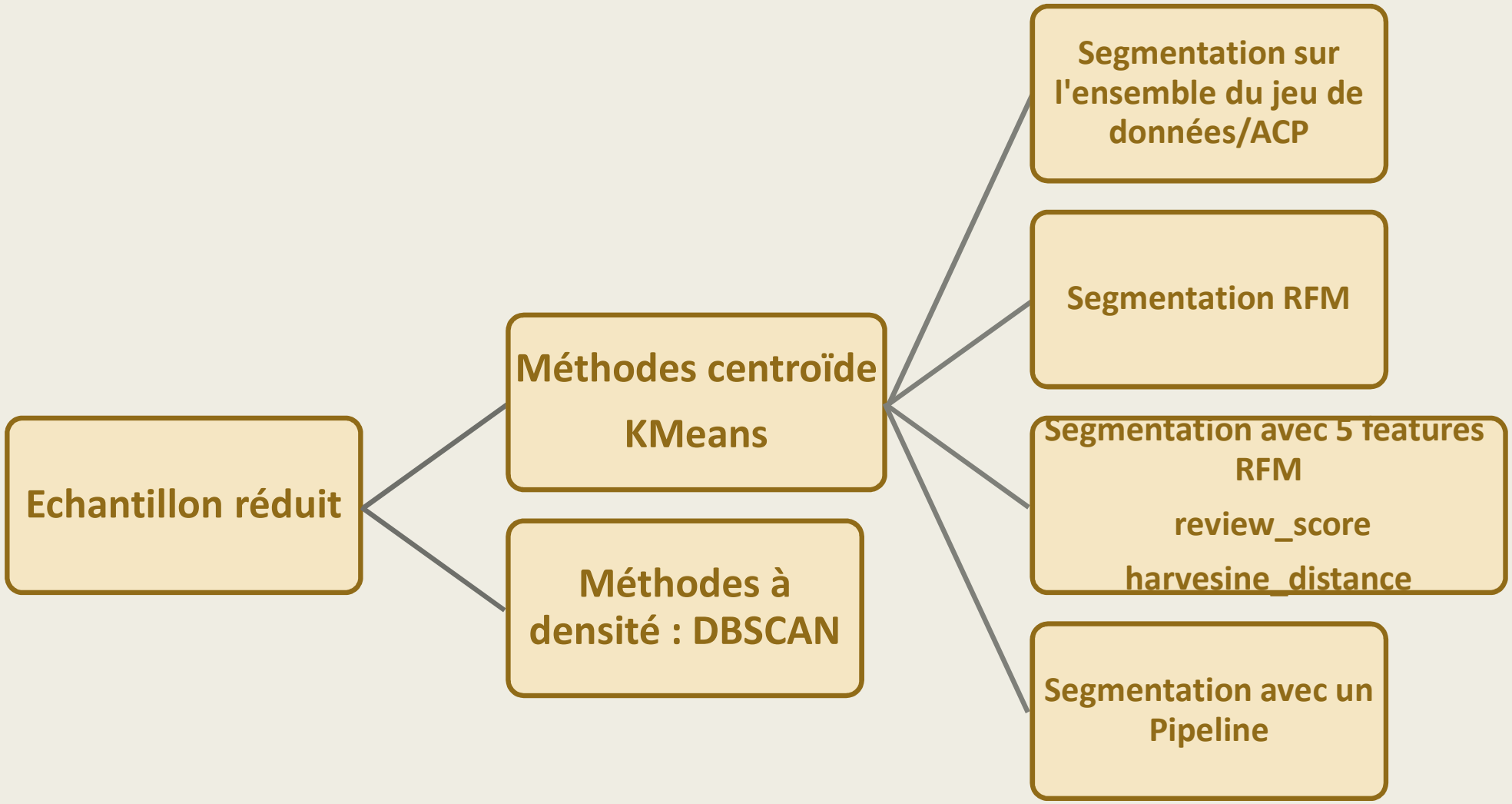
- **97445 lignes**
- **31 colonnes**

Essais des différentes approches de modélisation

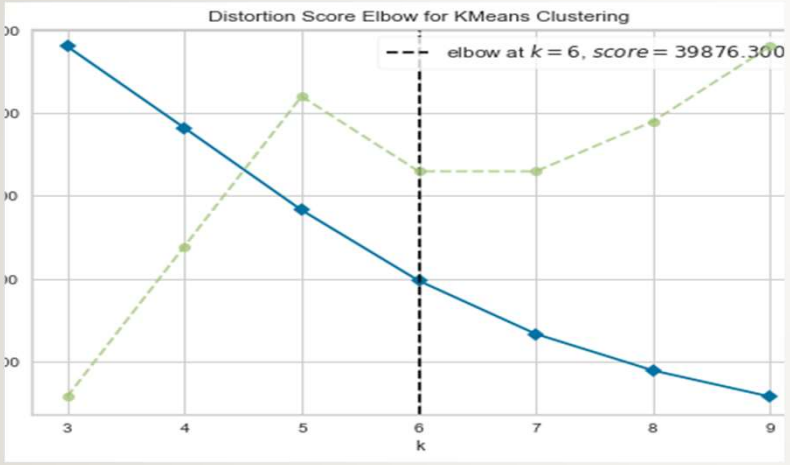
Préparation des données : Finalisation

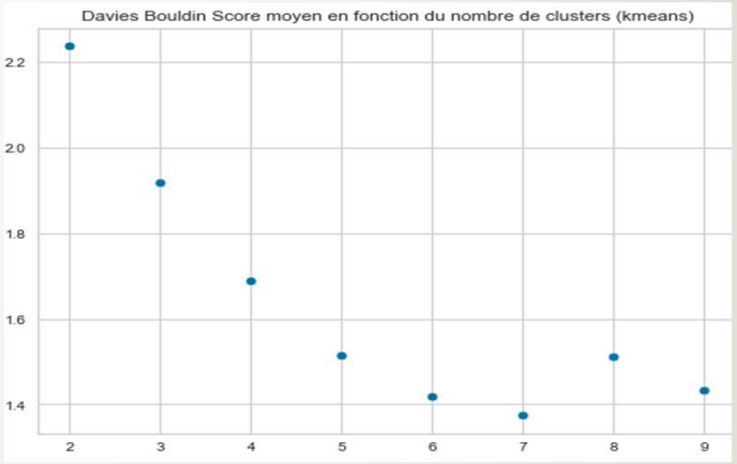
- ❑ Suppression des features non adaptées comme « product_id »
- ❑ Normalisation
- ❑ Sélection des features pour la segmentation

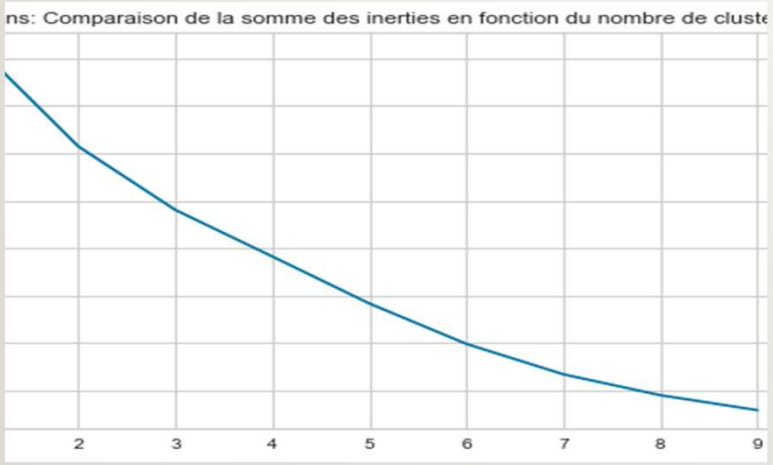
Processus de modélisation

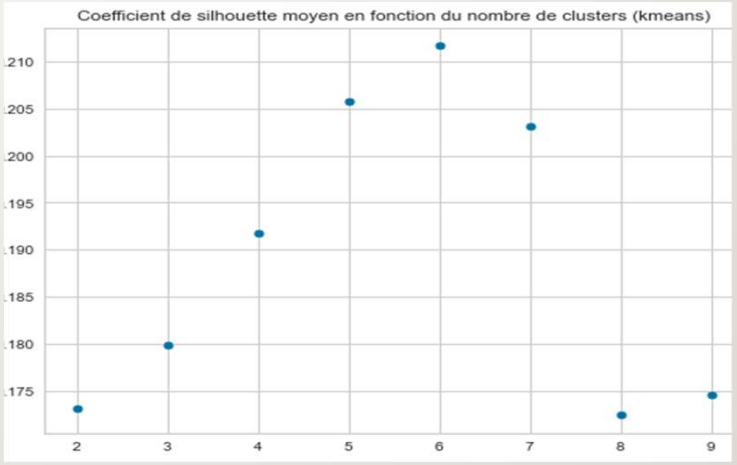


KMeans





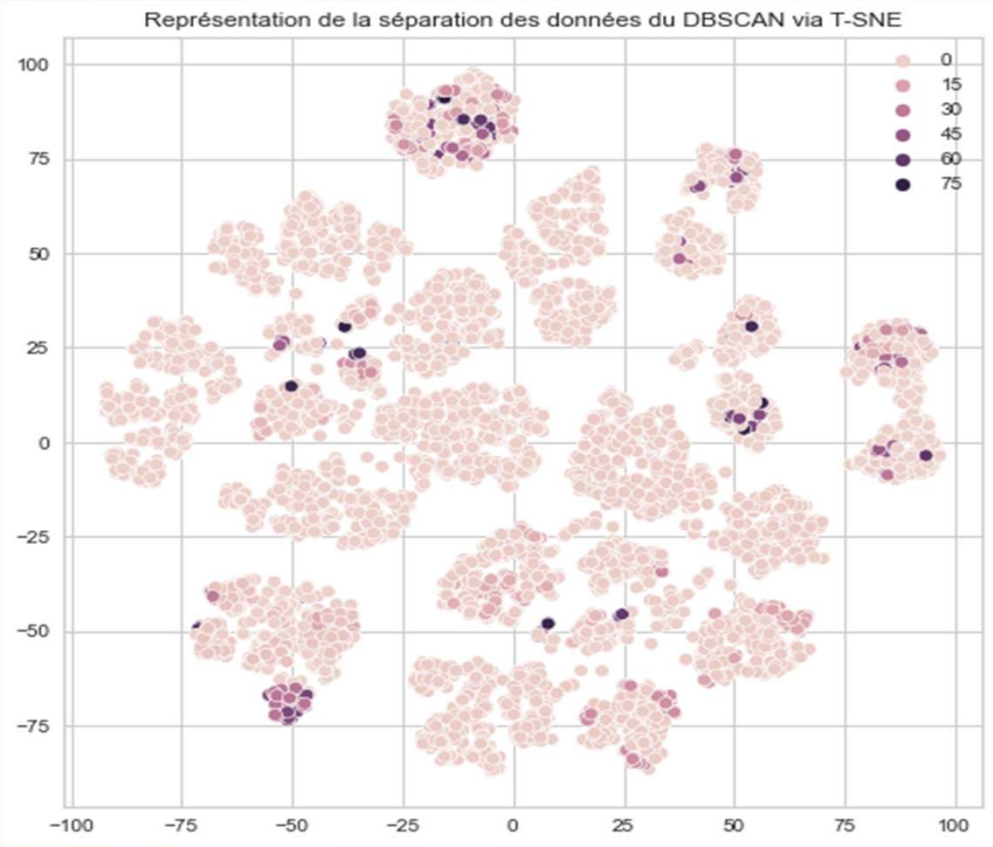




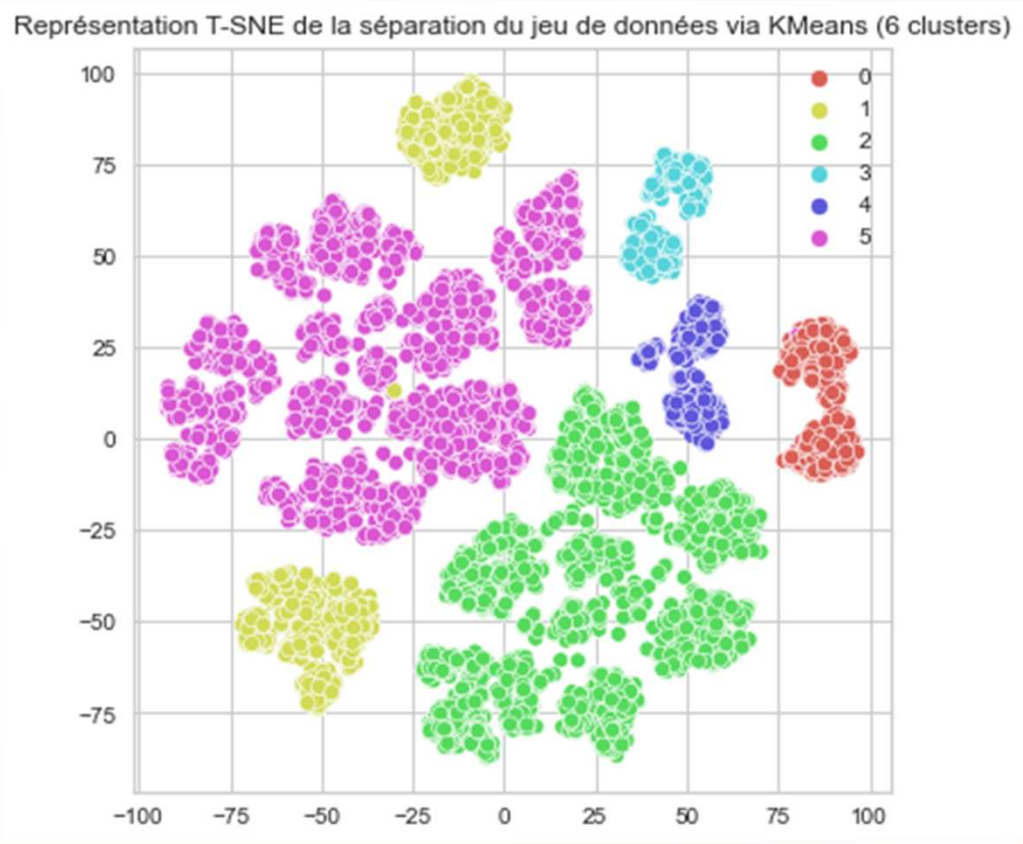
Le Best K

Visualisation via T-SNE

DBSCAN

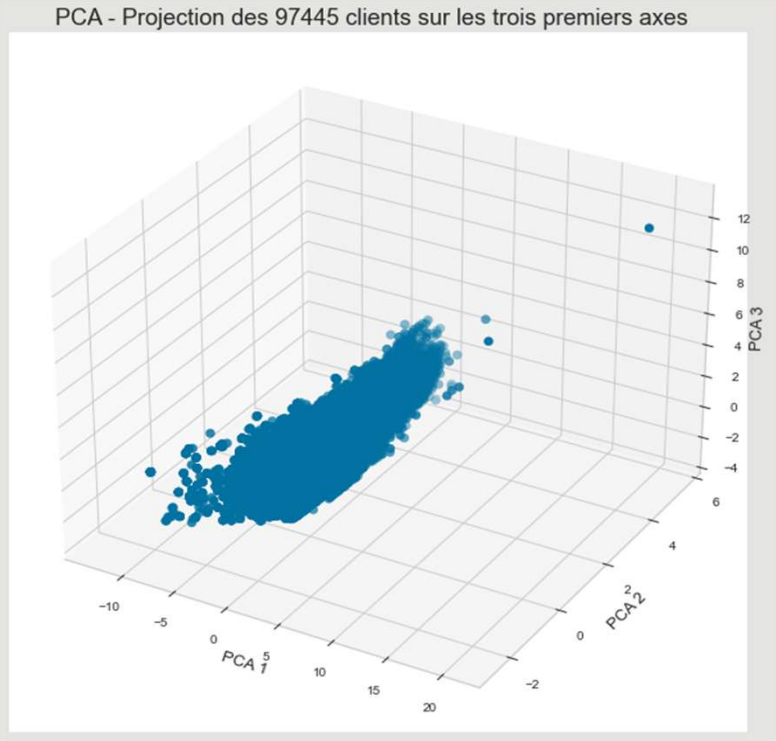


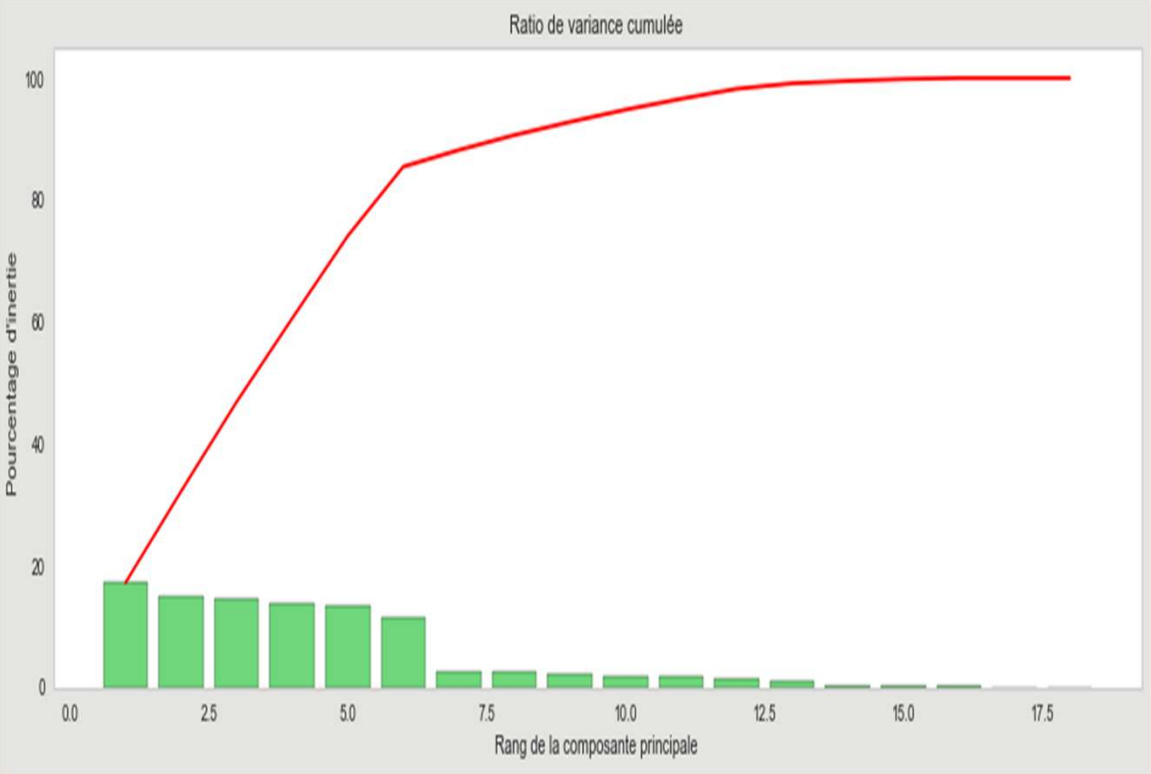
KMeans



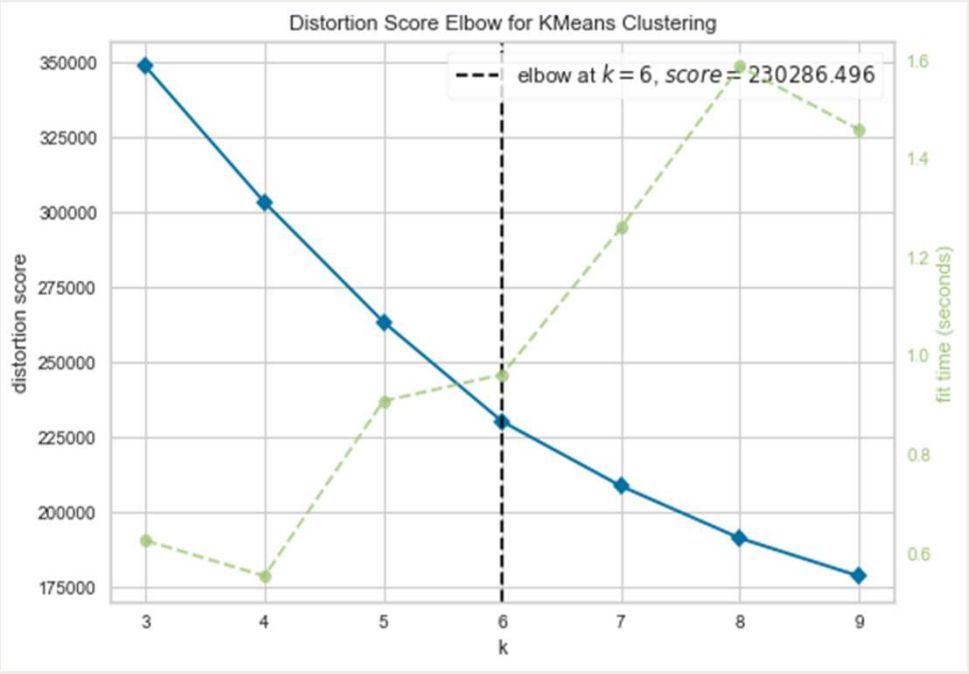
Segmentation sur l'ensemble du jeu de données

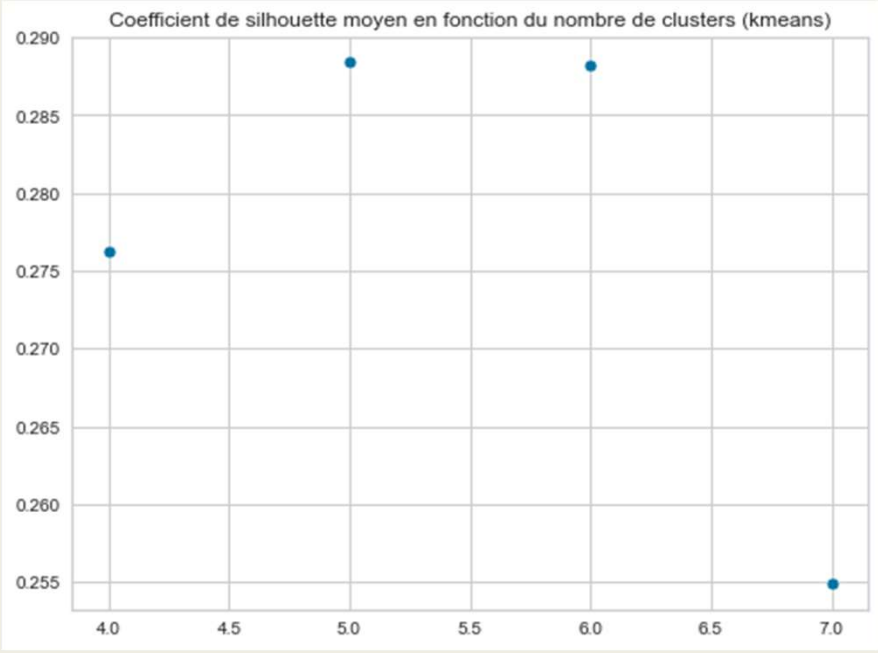
Réduction de dimension par ACP





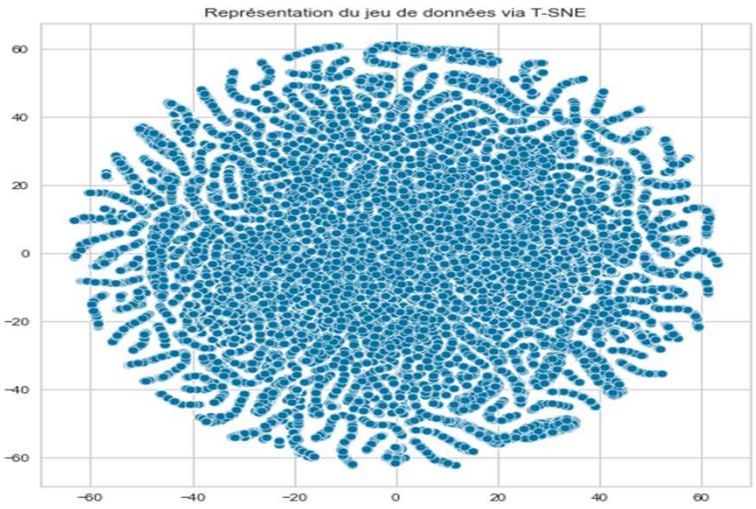
Segmentation sur l'ensemble du jeu de données

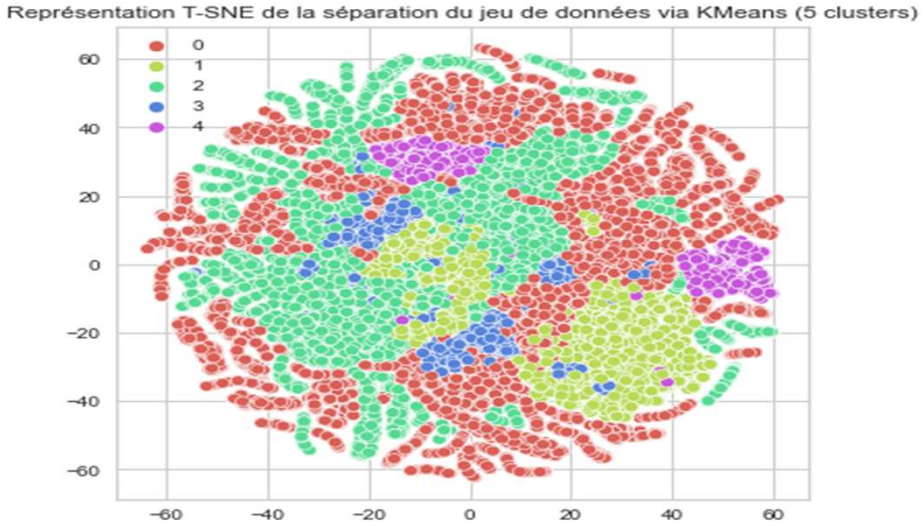


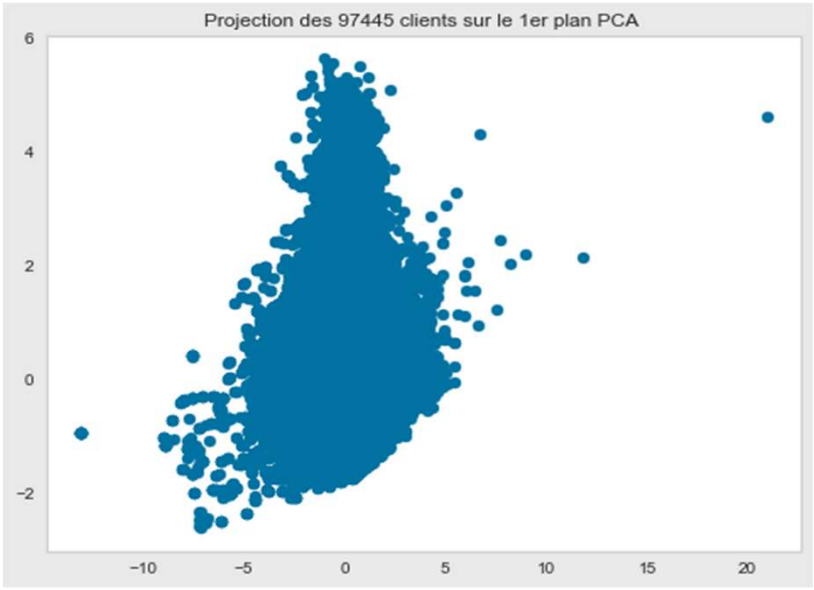


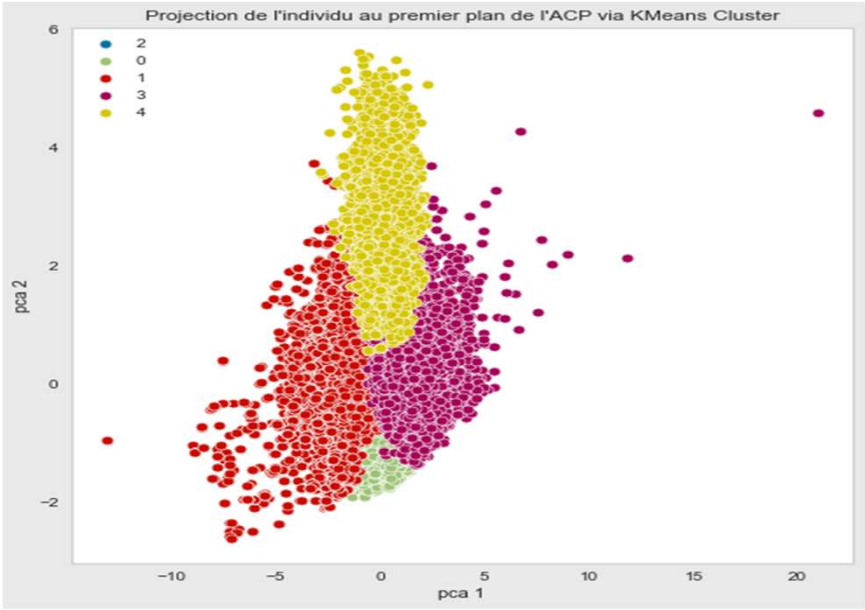
Le Best K

Visualisation

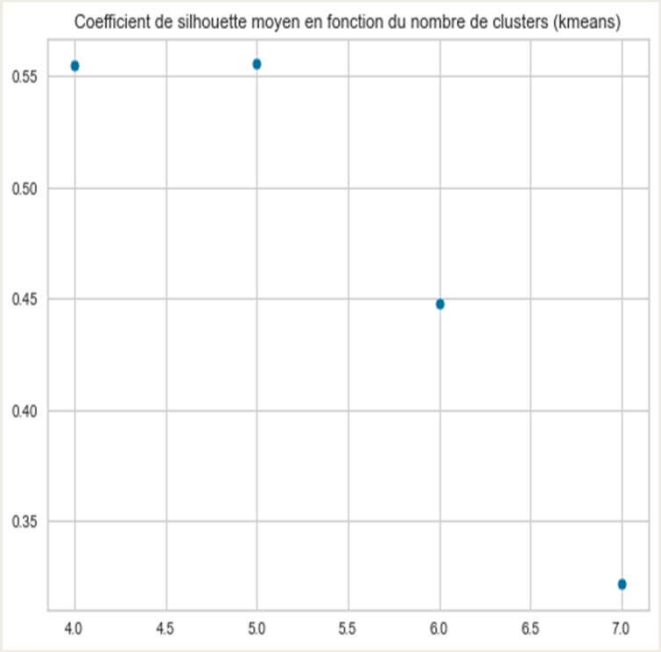


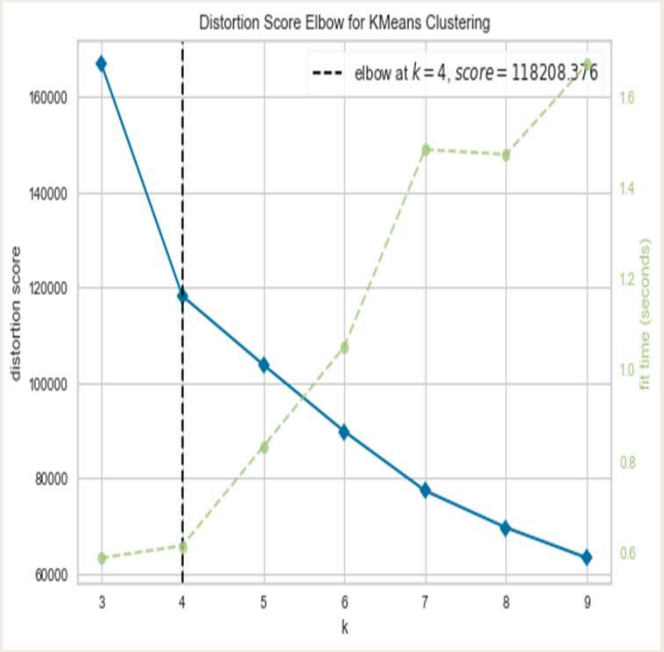


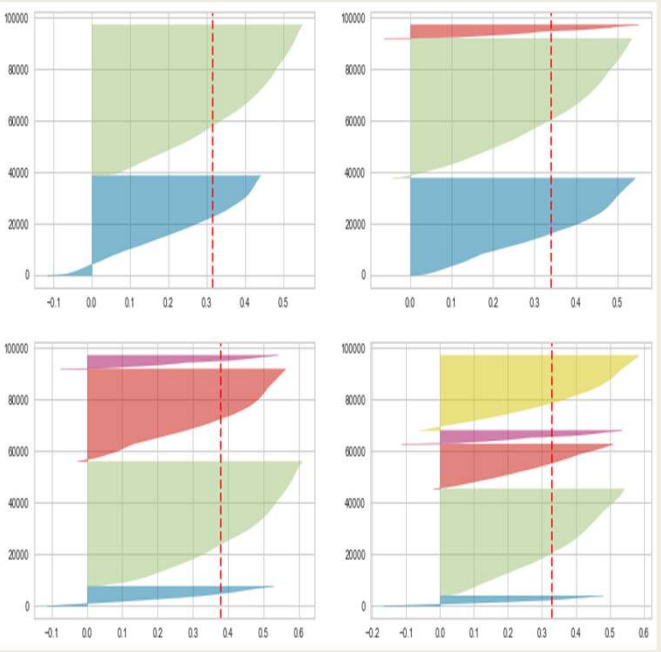




Segmentation RFM

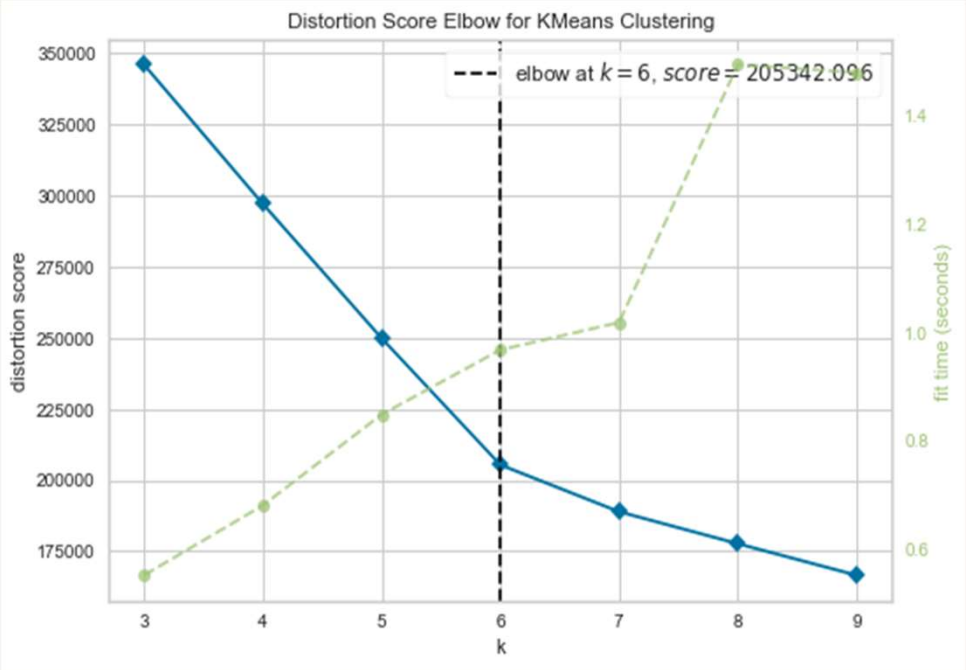


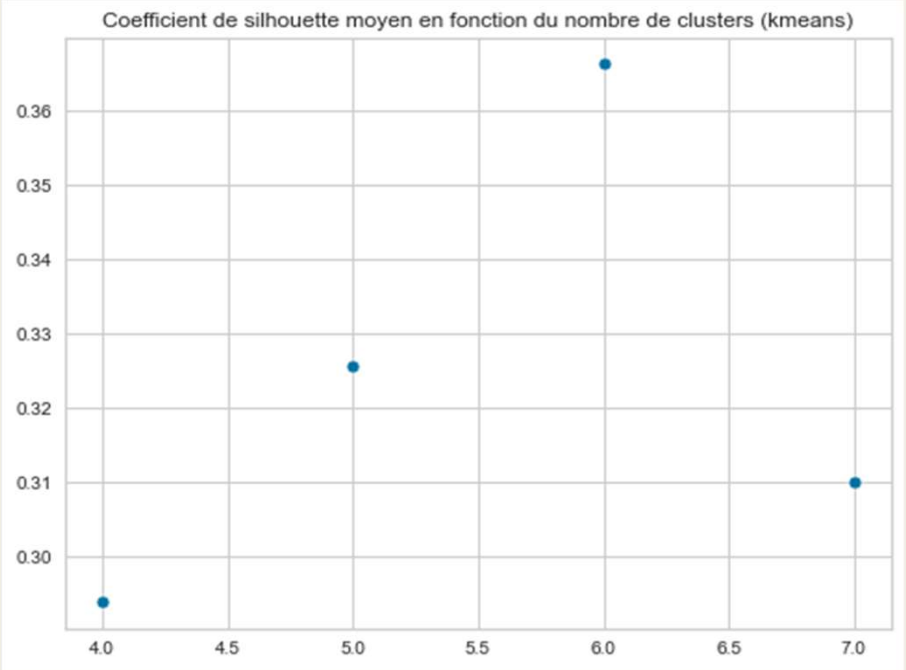




Le Best K

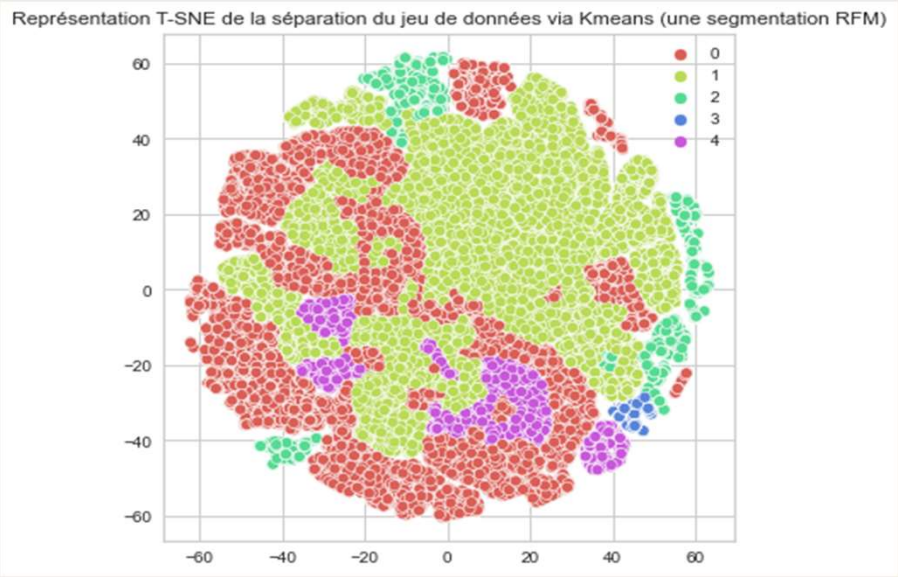
Segmentation avec 5 features
RFM + review_score+harvesine_distance

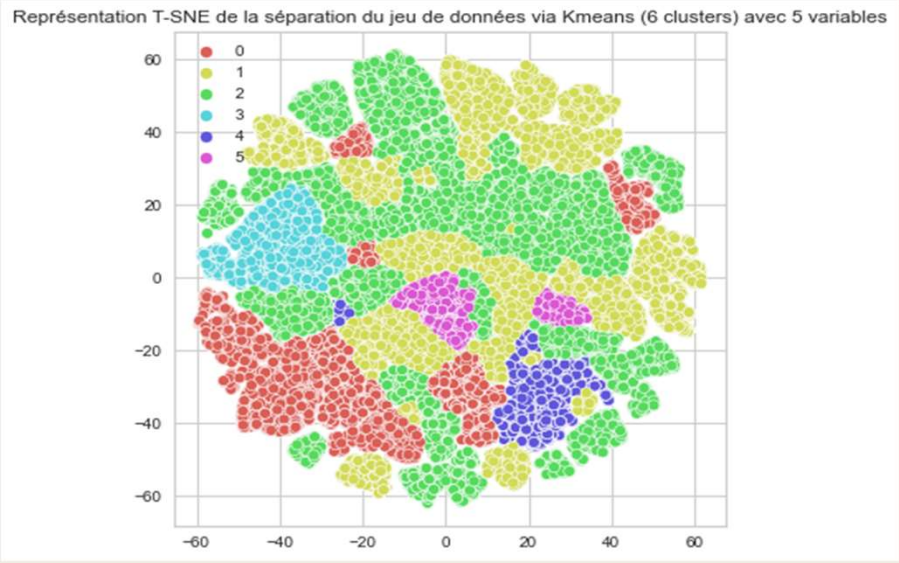




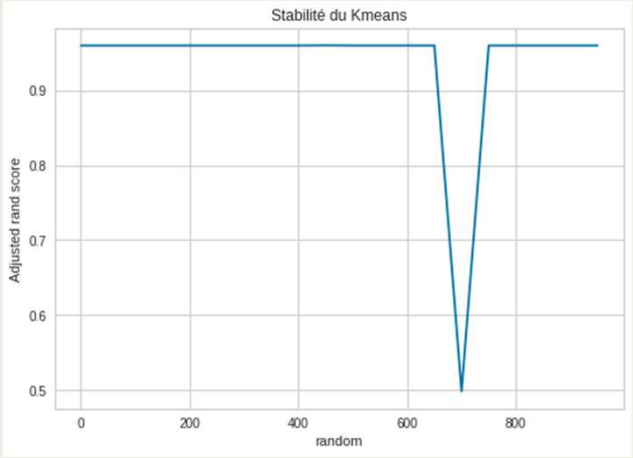
Le Best K

Visualization



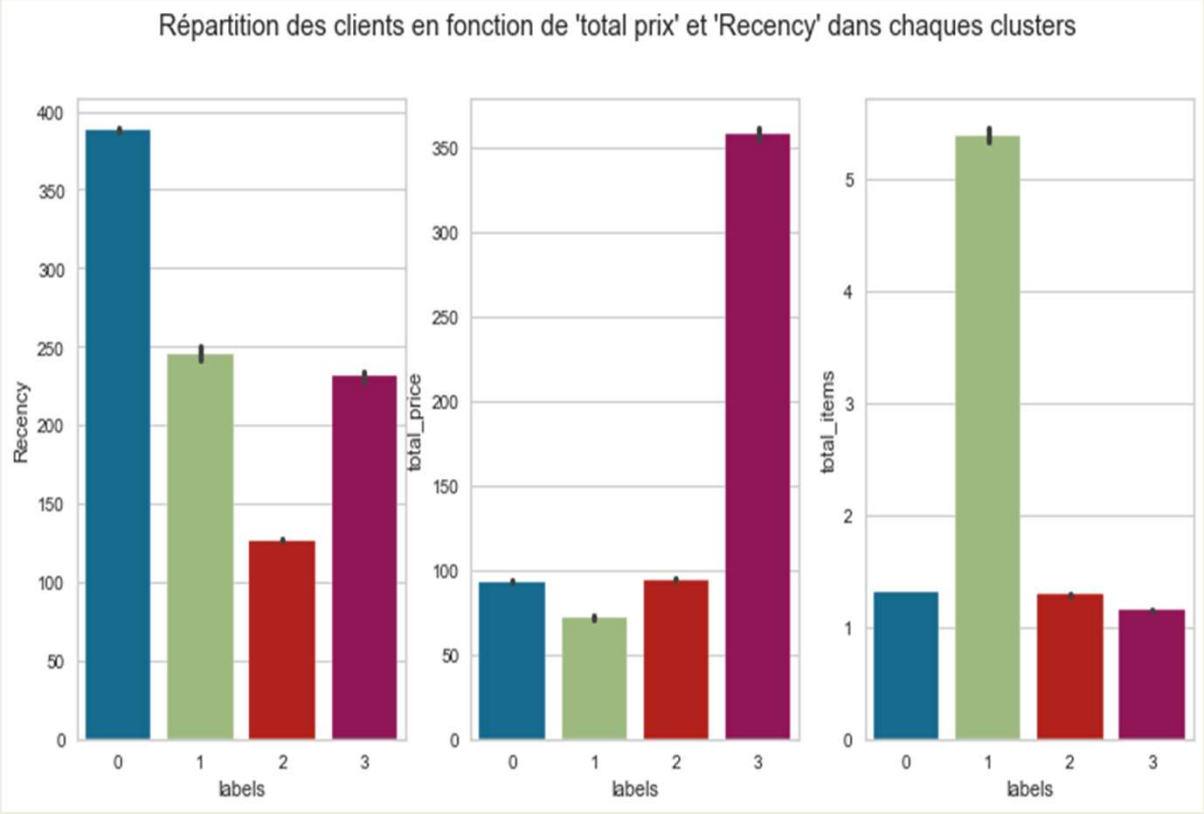


Stabilité du kMeans avec Adjusted_rand



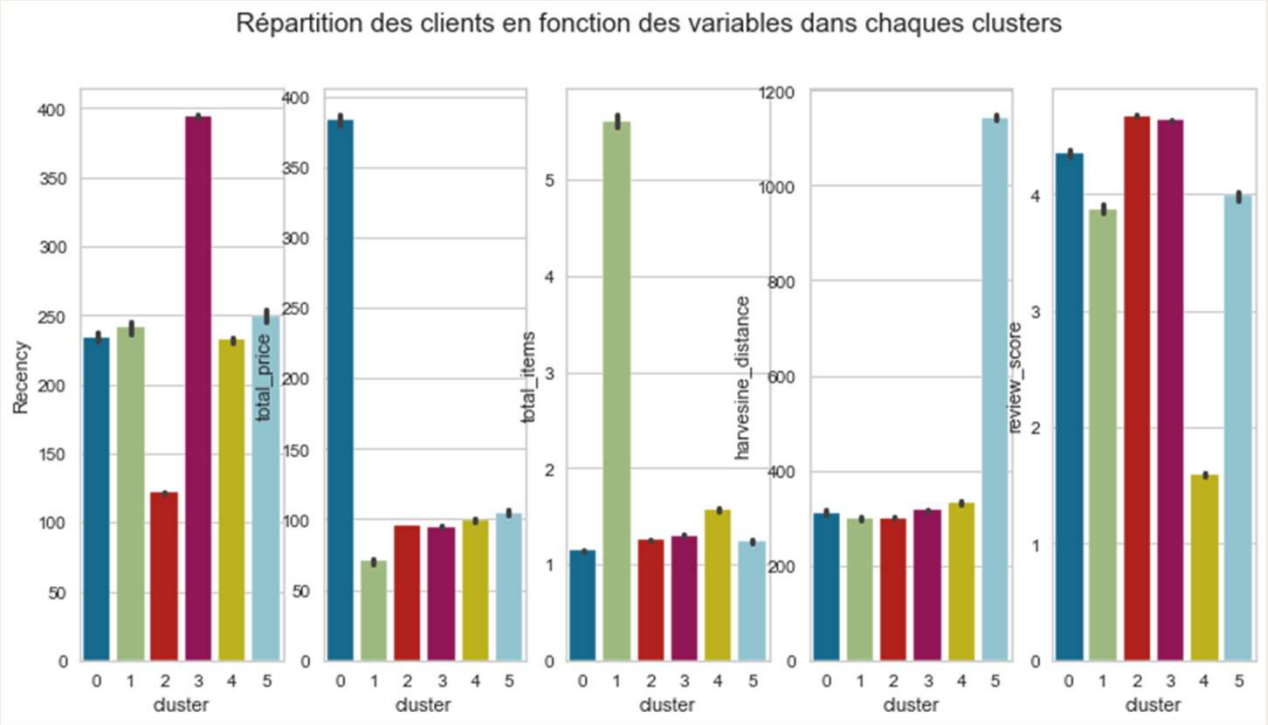
Types des clients

- Les nouveaux clients
- Les clients ayant acheté le plus de produits
- Les clients à reconquérir
- Les clients les plus dépensiers

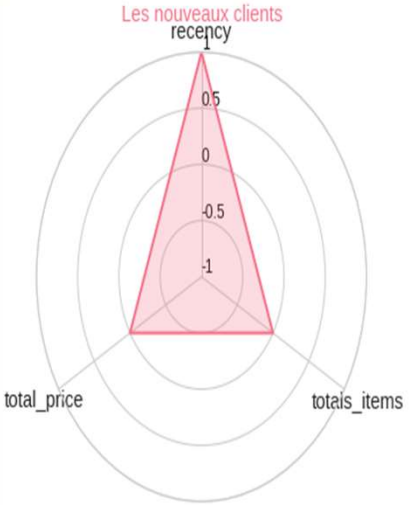


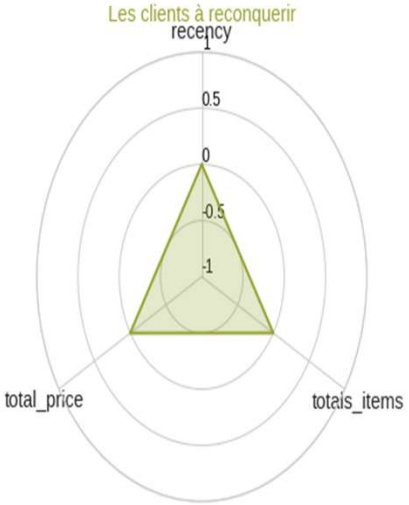
Types des clients

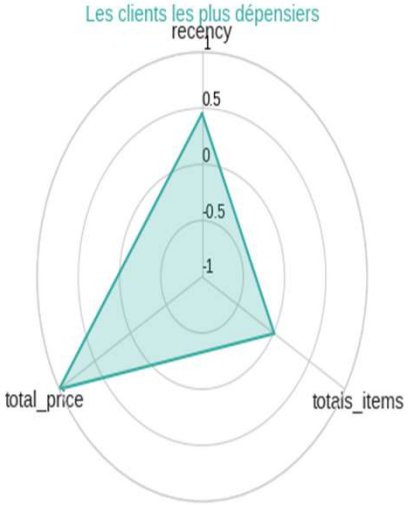
- Les nouveaux clients
- Les clients ayant acheté le plus de produits
- Les clients à potentiel
- Les clients les plus dépensiers
- Les clients insatisfaits
- Les clients les plus loin

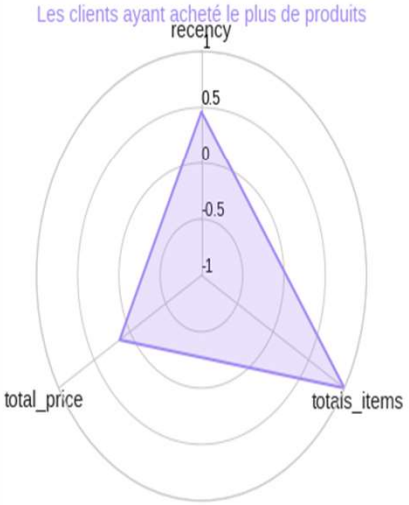


Visualization

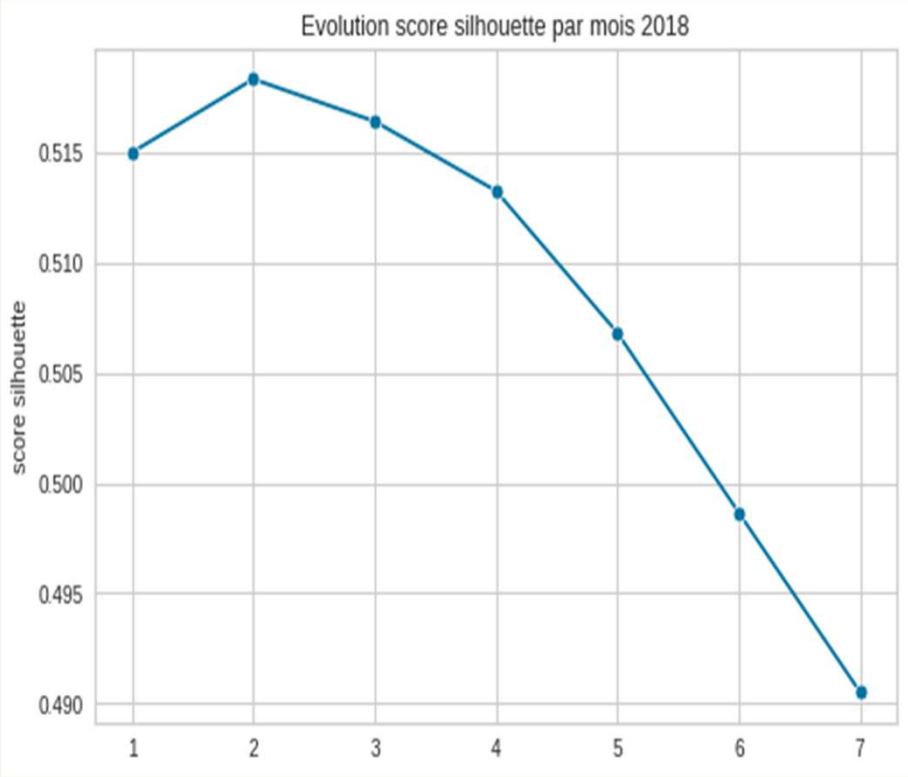


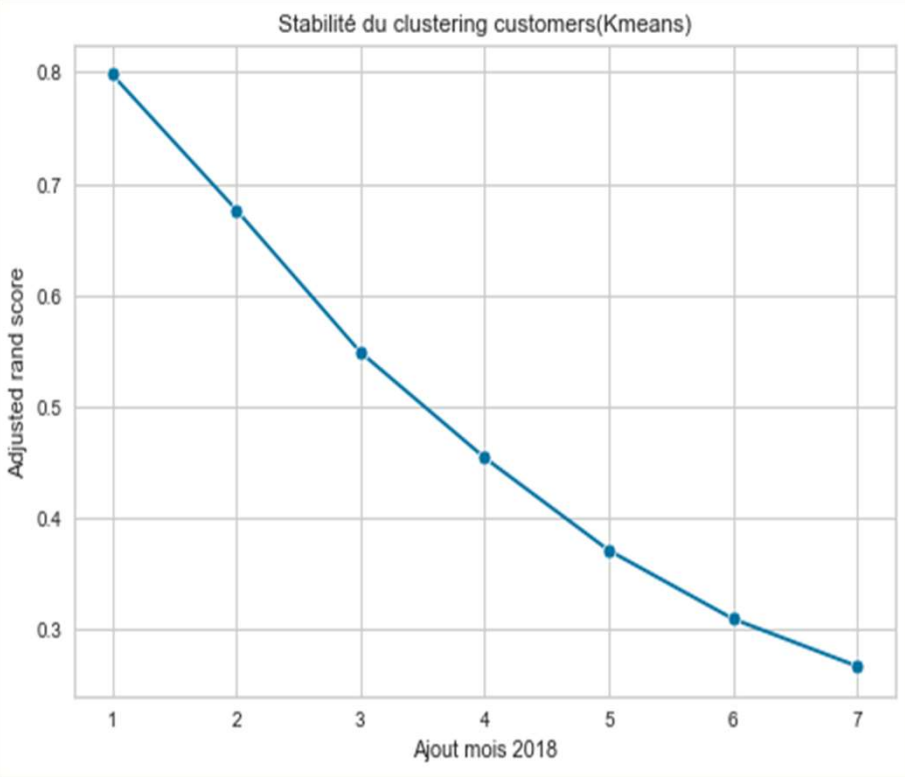






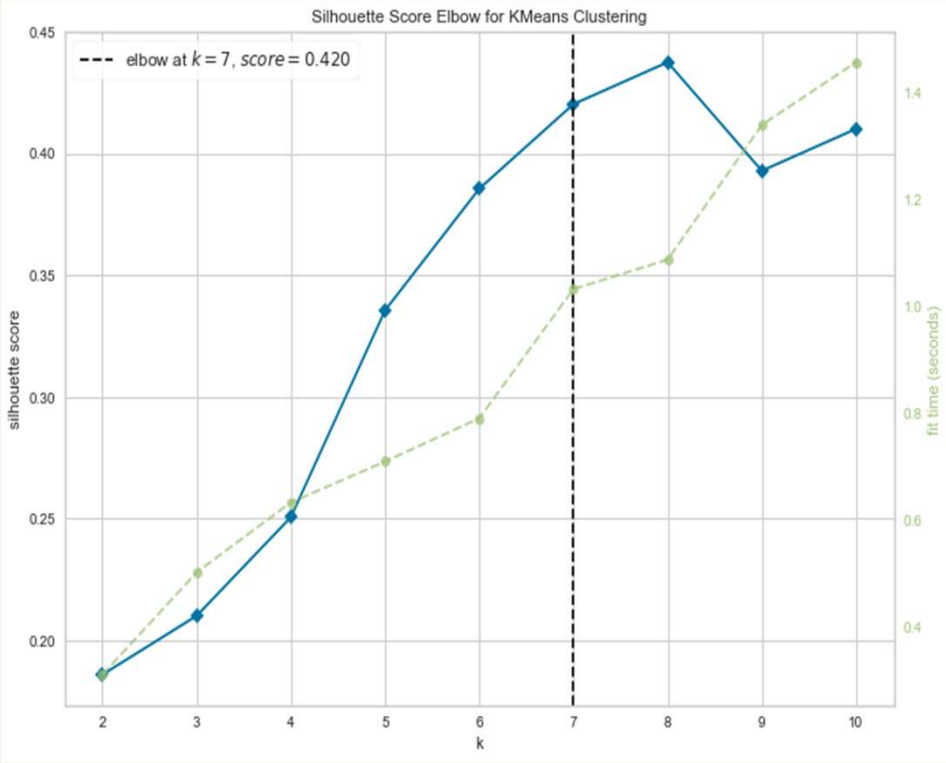
Stabilité temporelle de la segmentation

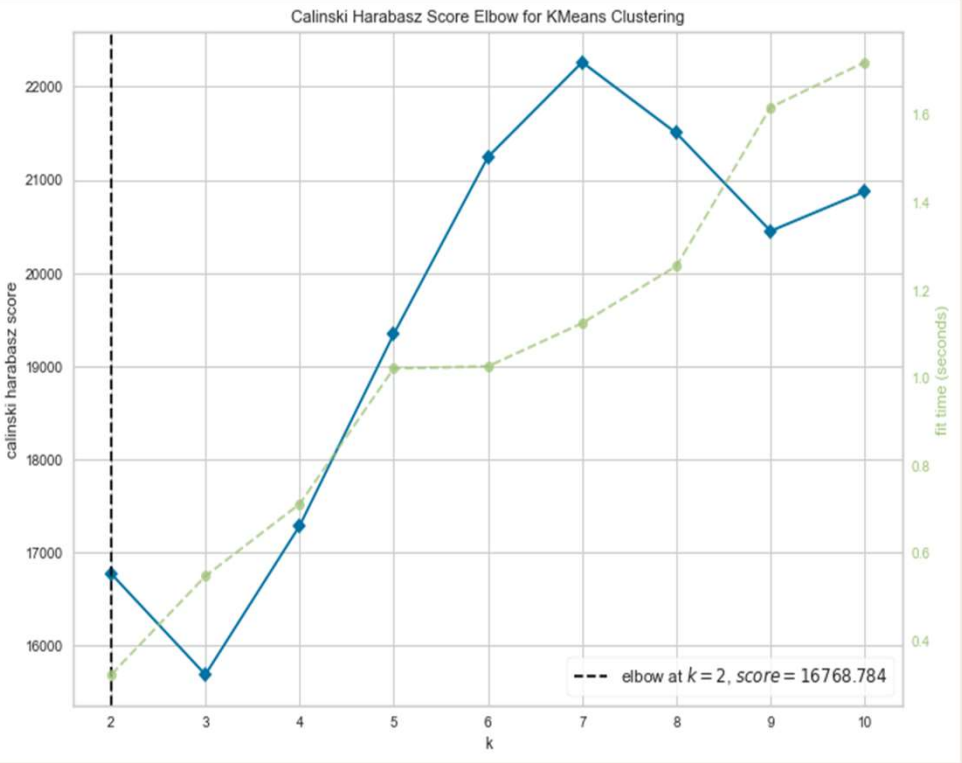




Segmentation avec un Pipeline

Elbow method with different metric





Scores de stabilité à l'initialisation

Iteration	FitTime	Inertia	Homo	ARI	AMI
Iter 0	0.086s	11471	0.918	0.912	0.905
Iter 1	0.065s	12634	0.583	0.481	0.632
Iter 2	0.075s	11645	0.644	0.530	0.638
Iter 3	0.072s	11458	1.000	1.000	1.000
Iter 4	0.056s	11458	1.000	1.000	1.000
Iter 5	0.057s	12634	0.583	0.481	0.632
Iter 6	0.064s	12634	0.583	0.481	0.632
Iter 7	0.051s	13189	0.583	0.459	0.623
Iter 8	0.062s	11472	0.918	0.912	0.905
Iter 9	0.098s	11472	0.920	0.913	0.906

Scores de stabilité à l'initialisation

Iteration	FitTime	Inertia	Homo	ARI	AMI
Iter 0	0.098s	1099	0.512	0.544	0.515
Iter 1	0.083s	1106	0.376	0.414	0.384
Iter 2	0.094s	1099	0.514	0.549	0.517
Iter 3	0.091s	1099	0.512	0.544	0.515
Iter 4	0.087s	1099	0.512	0.543	0.515
Iter 5	0.130s	1099	0.513	0.549	0.517
Iter 6	0.099s	1106	0.375	0.414	0.383
Iter 7	0.118s	1106	0.375	0.413	0.383
Iter 8	0.099s	1106	0.375	0.414	0.383
Iter 9	0.118s	1099	0.513	0.549	0.517

Scores de stabilité à l'initialisation

MERCI DE VOTRE ATTENTION