SEGMENTATION DE LA CLIENTELE D'UN SITE DE E-COMMERCE

Jay Corentin – Avril 2023

PROBLEMATIQUE

À partir de la base de données fournie par Olist, identifier les comportements d'achats des clients et les segmenter afin d'aider l'équipe Marketing dans ses campagnes de communication.

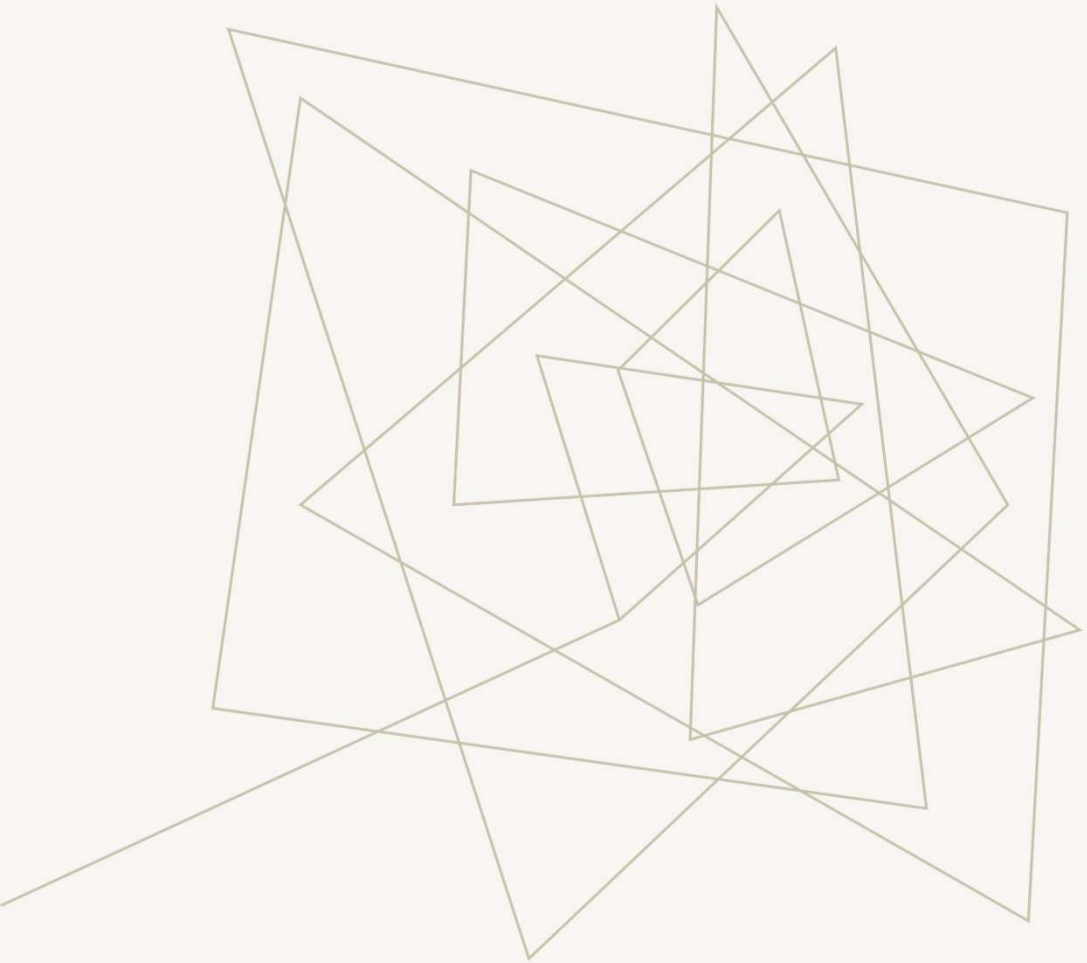
PLAN

I. Exploration des données

II. Essais de segmentation des clients

III. Simulation de maintenance

Jay Corentin – Avril 2023



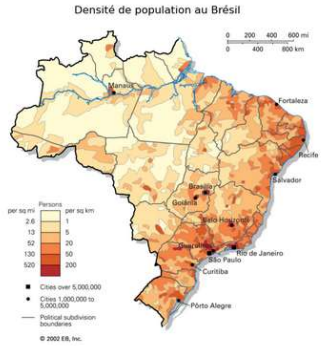
EXPLORATION DES DONNEES

Jay Corentin – Avril 2023

Dataset Olist

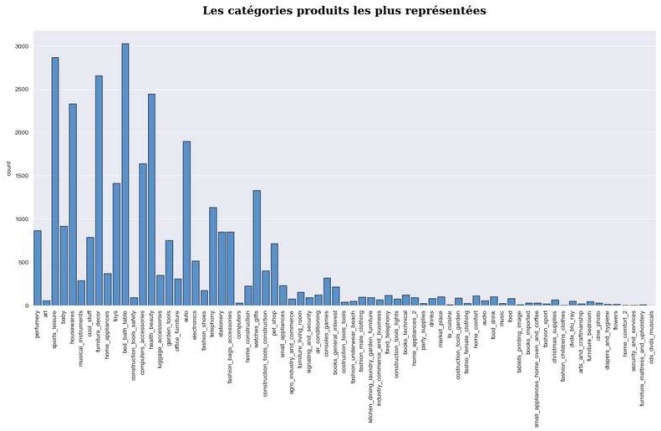
Composition du Dataset Olist : 9 fichiers

- geolocation_dataset
- customers_dataset
- order_items_dataset
- order_payment_dataset
- order_reviews_dataset
- orders_dataset
- products_dataset
- sellers_dataset
- product_category_name_translation

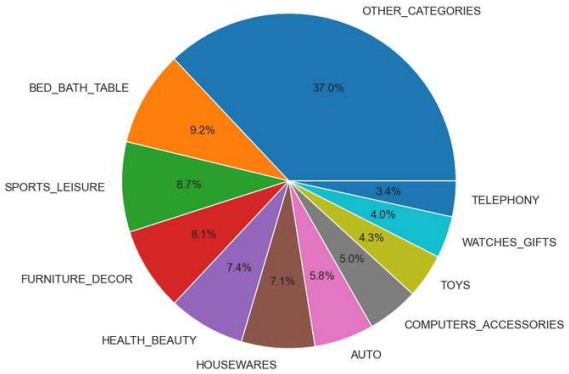


État avec les plus de commandes : Sao Paulo

Focus produits

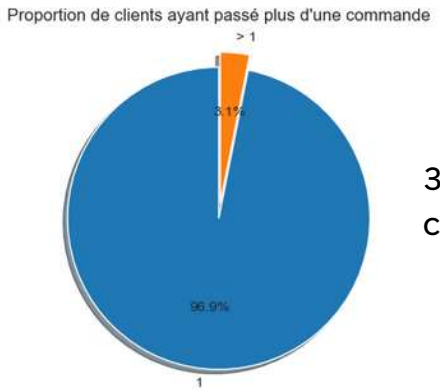


Initialement 71 catégories différentes de produits, dont certaines beaucoup plus représentées que d’autres.



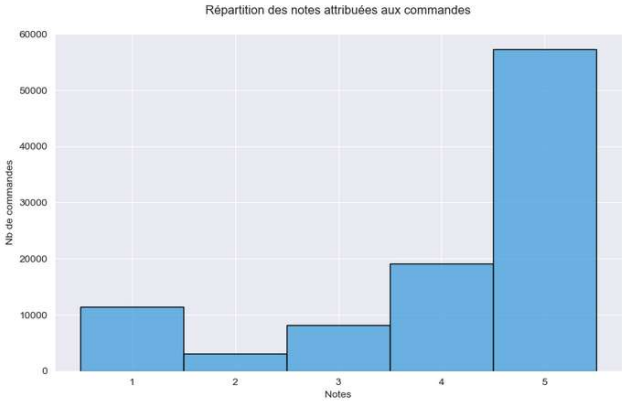
Réduction du nombre de catégories à 11, classées par volumes.
Catégories les moins représentées regroupées dans ‘Other categories’.

Focus clients



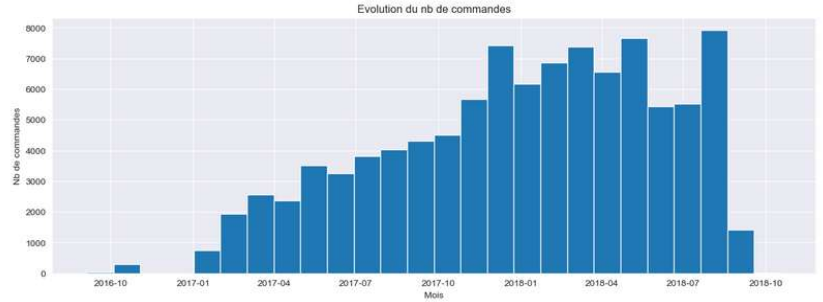
3,1% des clients ont passé au moins deux commandes (clients récurrents)

Le moyen de paiement le plus utilisé est la carte de crédit.

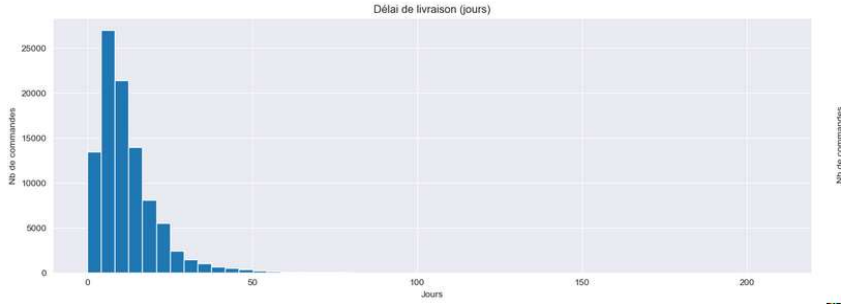


La note maximale (5) est attribuée dans une bonne proportion des commandes, mais présence de mauvais scores également.

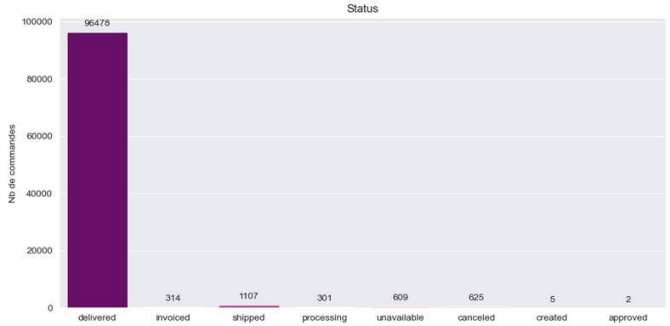
Focus commandes



Faible volume de commandes avant 2017-01 et après 2018-09. Période retenue de 20 mois.



Majorité des délais de livraison en quelques jours, mais parfois très longs.

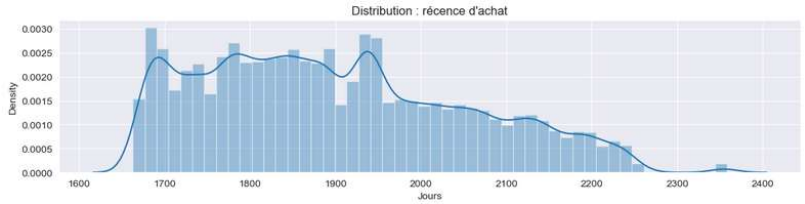


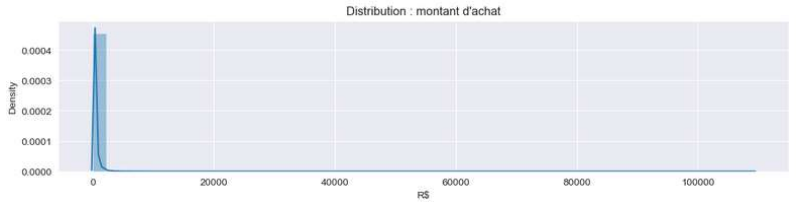
Nous ne retiendrons que les commandes avec le statut ‘Delivered’.

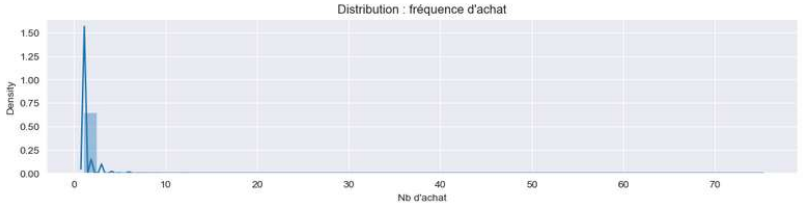
Feature engineering

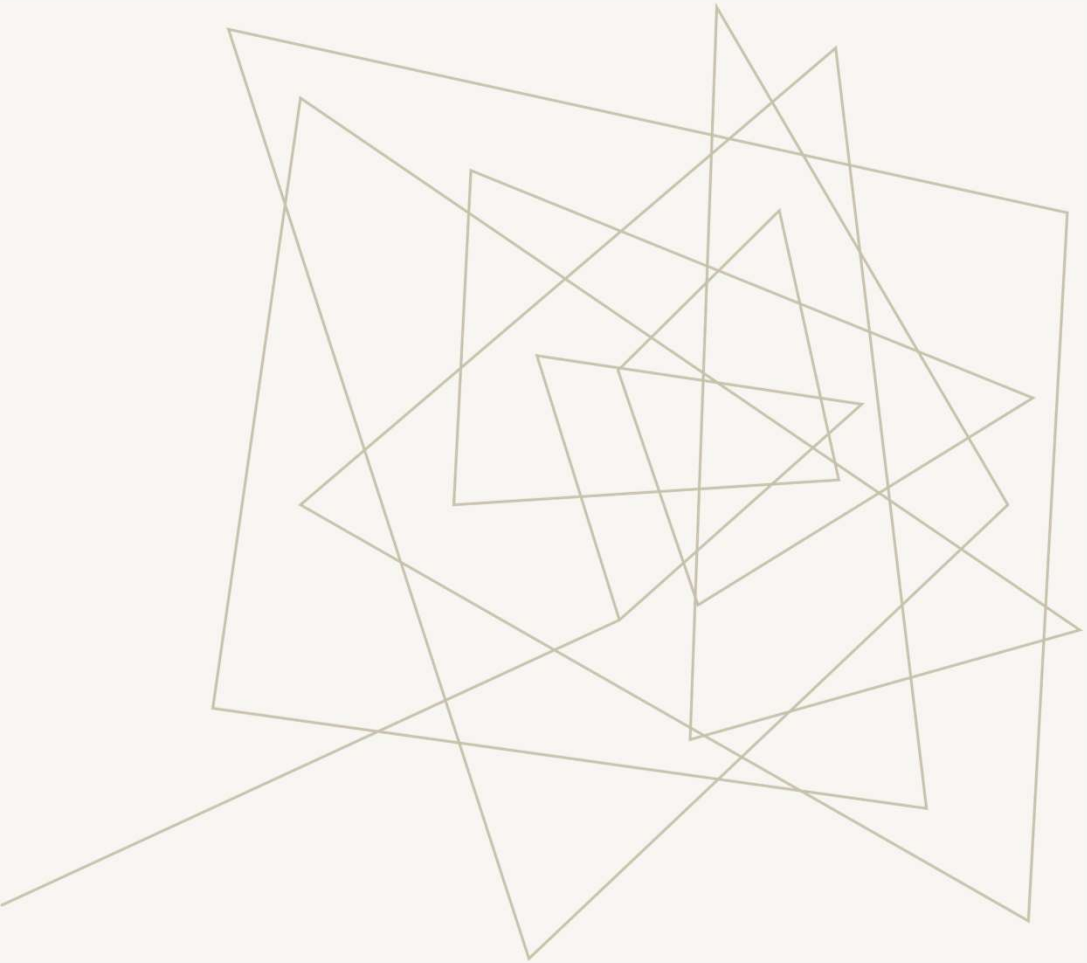
Création de variables :

- Freight ratio : part des frais de port sur le prix de la commande
- Variables RFM :
  - **R**écence (Nombre de jours depuis le dernier achat) ;
  - **F**réquence (Nombre de commandes) ;
  - **M**ontant d'achat (Montant total dépensé).









ESSAIS DE SEGMENTATION

Jay Corentin – Avril 2023

Approche de modélisation

Modèles utilisés (modélisation non supervisée) :

- **KMeans** : définition de centroïdes par moyennes arithmétiques, et attribution des points à chaque cluster en fonction de leur distance vis-à-vis des centroïdes.
- **DBSCAN** : définition de clusters par densités d'individus, les points trop éloignés sont considérés comme du bruit.
- **CAH (Classification Ascendante Hiérarchique)** : calcul de la dissimilarité entre les objets, puis regroupement selon un critère d'agrégation.

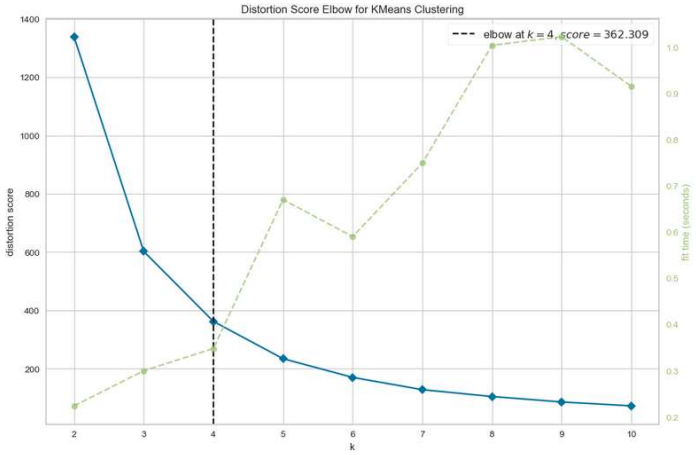
Datasets utilisés :

- **Variables RFM** (Récence, Fréquence, Montant d'achat) ;
- **Dataset global** : toutes les variables ;
- **Dataset hybride** : variables sélectionnées.

Critères de sélection du modèle :

- **Score d'évaluations** (Silhouette, Calinski-Harabasz, Davies-Bouldin) ;
- **Intérêt Marketing** : pluralité des clusters et explicabilité ;
- **Temps d'exécution**.

Variables RFM



Modélisation via Kmeans : méthode du coude (‘Elbow method’) avec un nombre de clusters allant de 2 à 10.

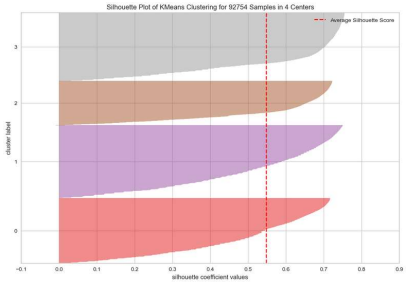
La modélisation suggère 4 clusters. Evaluons nos modèles sur un range de clusters allant de 3 à 6.

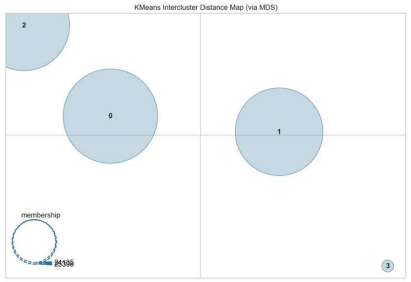
En combinant nos critères de sélection, la modélisation la plus pertinente est un nombre de clusters égal à 4.

Clusters	Silhouette	Calinski-Harabasz	Davies-Bouldin	Chrono (s)
3	0.571863	298946.309049	0.523522	69.88
4	0.548467	351913.536528	0.529494	68.07
5	0.547312	420610.205307	0.525351	67.57
6	0.542738	470379.396922	0.528654	66.08

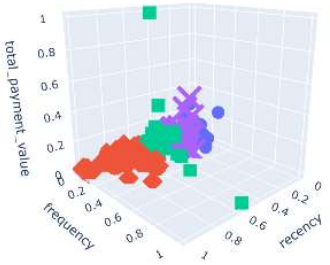
Variables RFM

Analyse de la segmentation RFM en 4 clusters

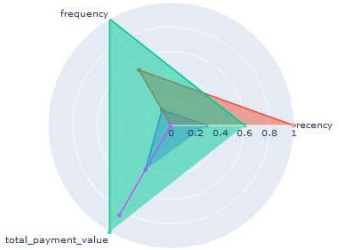




- Clusters bien distincts et homogènes.
- Scores intéressants



Comparaison des moyennes par variable des clusters



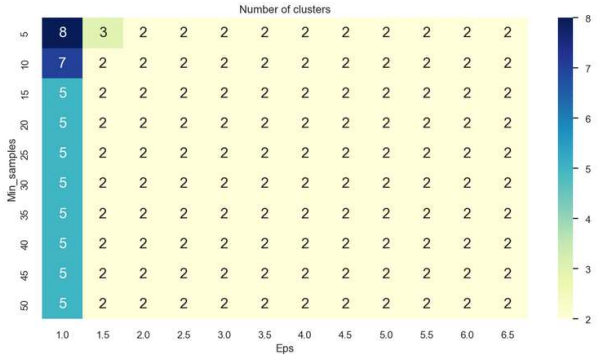
Trop peu de variables, faible intérêt Marketing.

Variables RFM

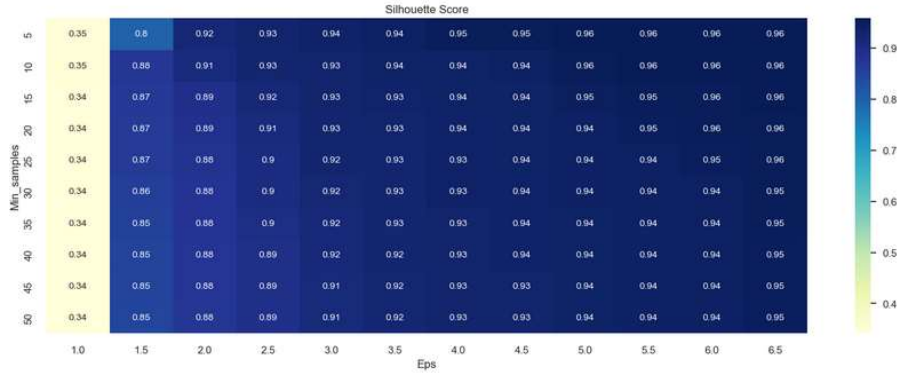
Modélisation via DBSCAN

Echantillonage du Dataset à 15%
Hyper paramètres testés :

- Eps values : entre 1 et 7 (gap de 0,5) ;
- Min samples : entre 5 et 55 (gap de 5).



Le nombre de clusters suggérés évolue entre 2 et 8, la valeur 2 étant la plus présente.

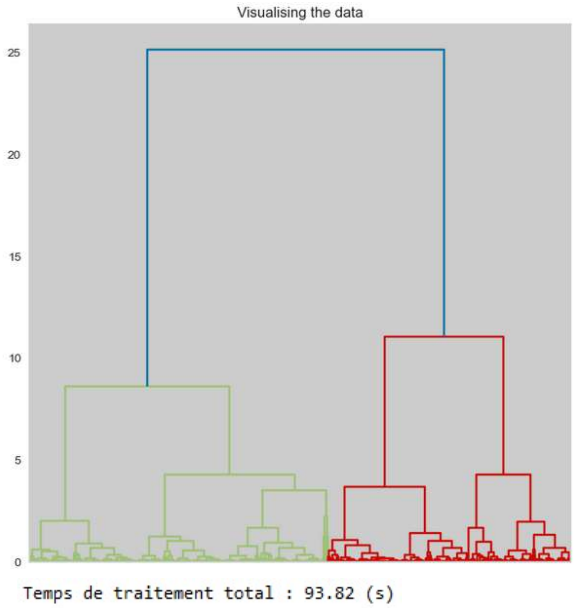


Les meilleurs scores de Silhouette (de 0,8 à 0,96) se trouvent sur 2 ou 3 clusters, ce qui est trop faible d'un point de vue Marketing.
Pour les nombres d clusters plus élevés, les scores chutent drastiquement (0,3).
Temps de traitement très long, même sur un échantillon.

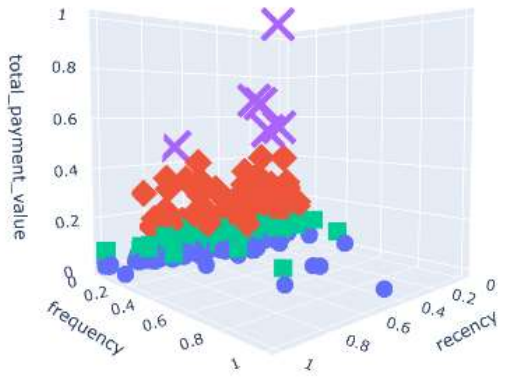
```
Temps de traitement total : 510.51 (s)
```

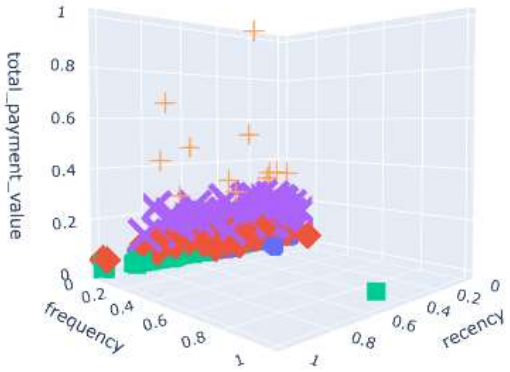
Variables RFM

Modélisation CAH (Classification Ascendante Hiérarchique) :
échantillonnage de la donnée à hauteur de 10%



Représentation 3D d'une segmentation en 4 et 5 clusters :



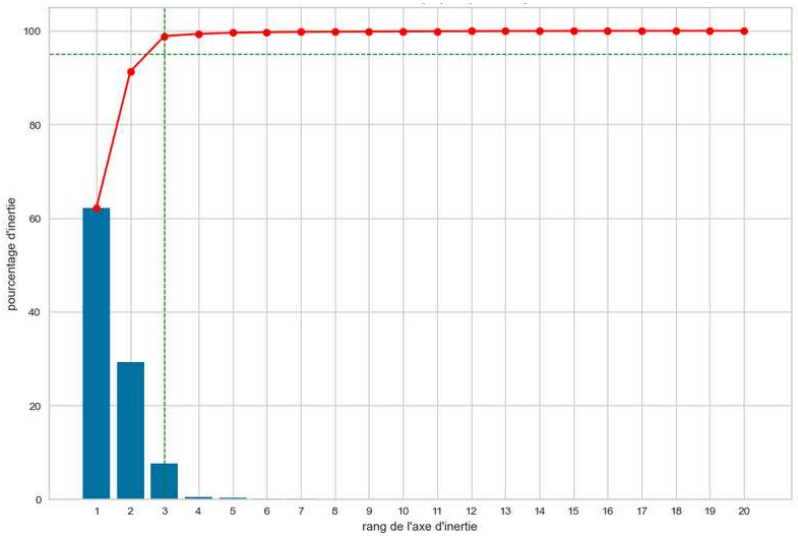


Ce modèle présente des temps de traitement très longs, même en ne prenant qu'un échantillon de la donnée, ce qui provoque une perte d'information, et biaise la fiabilité du modèle.
En outre, les clusters créés ne peuvent pas être identifiés assez distinctement, ne présentant donc pas d'intérêt Marketing.

Ce modèle sera écarté de notre approche.

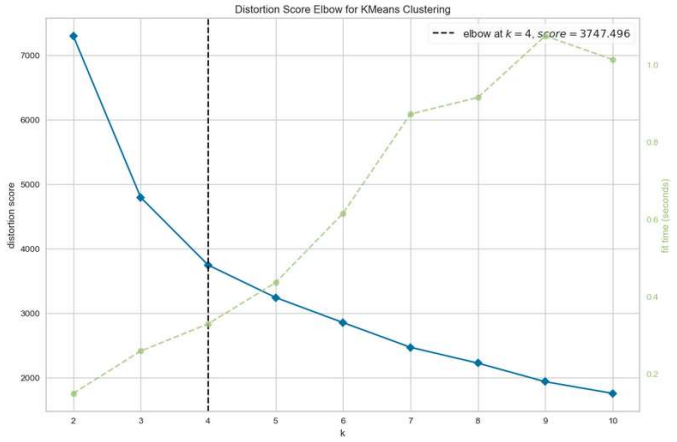
Dataset global

Le Dataset global comporte 20 variables, dont la moitié constituée des catégories produits réduites et agrégées. Par conséquent, nous utiliserons une Analyse en Composante Principale (ACP) pour réduire la dimension.



Selon l'ACP, il est possible de réduire la dimension à 3 variables, tout en conservant plus de 95% de la variance, donc de l'information.

Dataset global

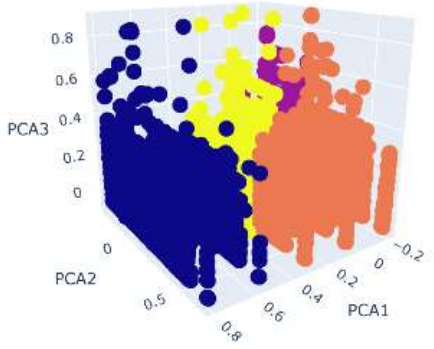


Clusters	Silhouette	Calinski-Harabasz	Davies-Bouldin	Chrono (s)
3	0.403692	95241.681767	0.888138	69.97
4	0.404325	89913.709063	0.924656	68.81
5	0.398288	82865.758435	0.972401	68.19
6	0.374904	78968.980460	0.924970	67.13

Modélisation via Kmeans (ACP en 3 dimensions) : méthode du coude ('Elbow method') avec un nombre de clusters allant de 2 à 10.

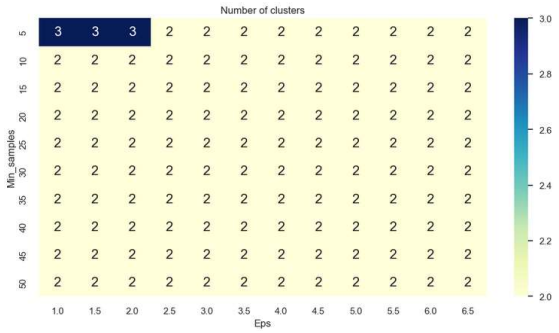
La modélisation suggère 4 clusters. Evaluons nos modèles sur un range de clusters allant de 3 à 6.

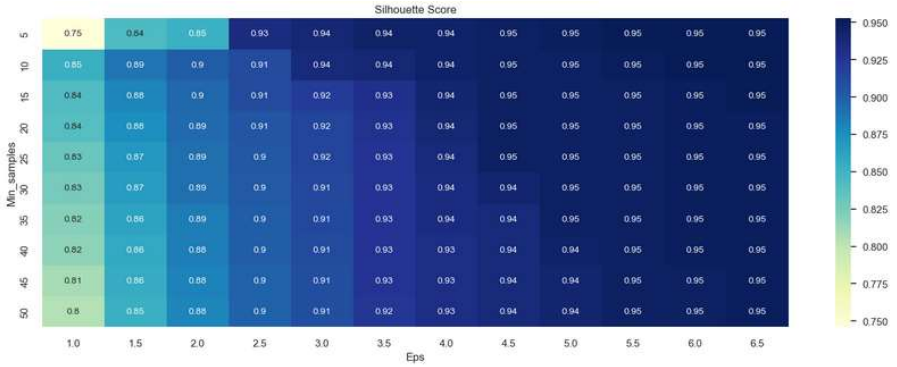
Les scores sont relativement plus faibles, et nous ne disposons que de trois variables, ce qui réduit significativement l'intérêt Marketing.



Dataset global

Modélisation via DBSCAN (ACP en 3 dimensions)





Temps de traitement total : 247.17 (s)

Echantillonage du Dataset à 10%
Hyper paramètres testés :

- Eps values : entre 1 et 7 (gap de 0,5) ;
- Min samples : entre 5 et 55 (gap de 5).

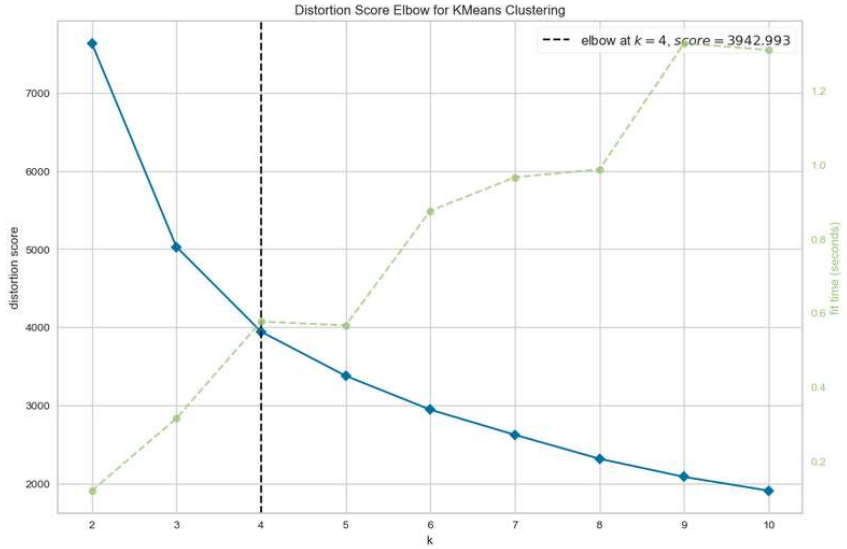
Le nombre de clusters suggérés évolue entre 2 et 3, la valeur 2 étant la plus présente.

Les scores de Silhouette sont relativement très bons, cependant le modèle ne nous suggère que peu de clusters, donc pas d'intérêt Marketing.
Temps de traitement très long, même sur un échantillon.

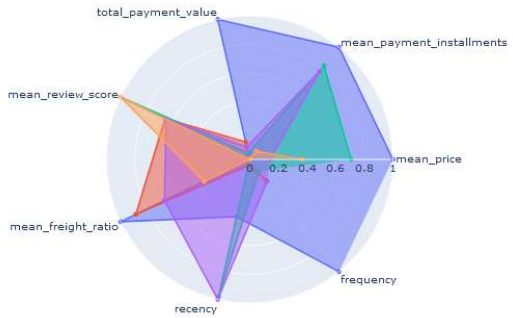
Dataset hybride

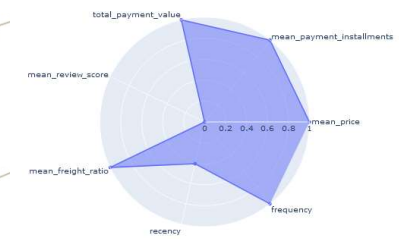
Variables retenues :

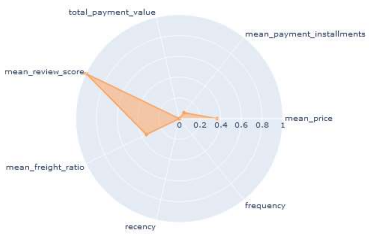
- RFM (Récence, Fréquence, Montant d'achat) ;
- Prix moyen d'achat ;
- Echelonnement moyen des paiements (Nb d'échéances) ;
- Score moyen ;
- Part moyenne des frais de port dans le prix.

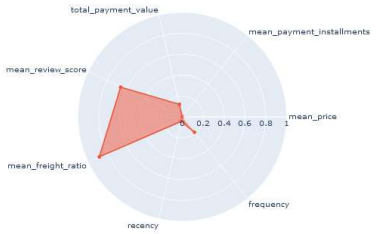


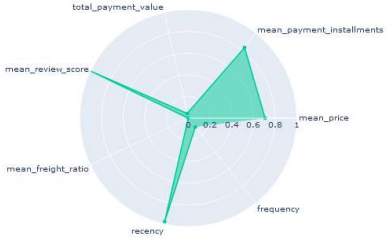
Clusters	Silhouette	Calinski-Harabasz	Davies-Bouldin	Chrono (s)
3	0.400208	94166.379695	0.892635	70.85
4	0.398992	88559.623043	0.930578	70.32
5	0.391943	81424.078206	0.982202	69.49
6	0.367001	77451.285652	0.934347	68.34
7	0.355070	74372.649914	0.997079	68.67
8	0.372350	73944.607633	0.950309	67.64

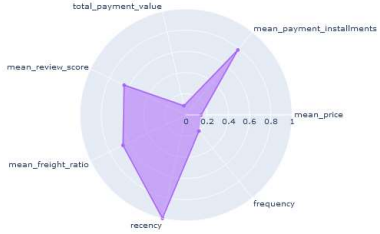












Dataset hybride

Clusters	**0**	**1**	**2**	**3**	**4**
Individus	11 688	15 086	23 387	11 199	31 394
Proportion	12,6%	16,3%	25,2%	12,1%	33,8%

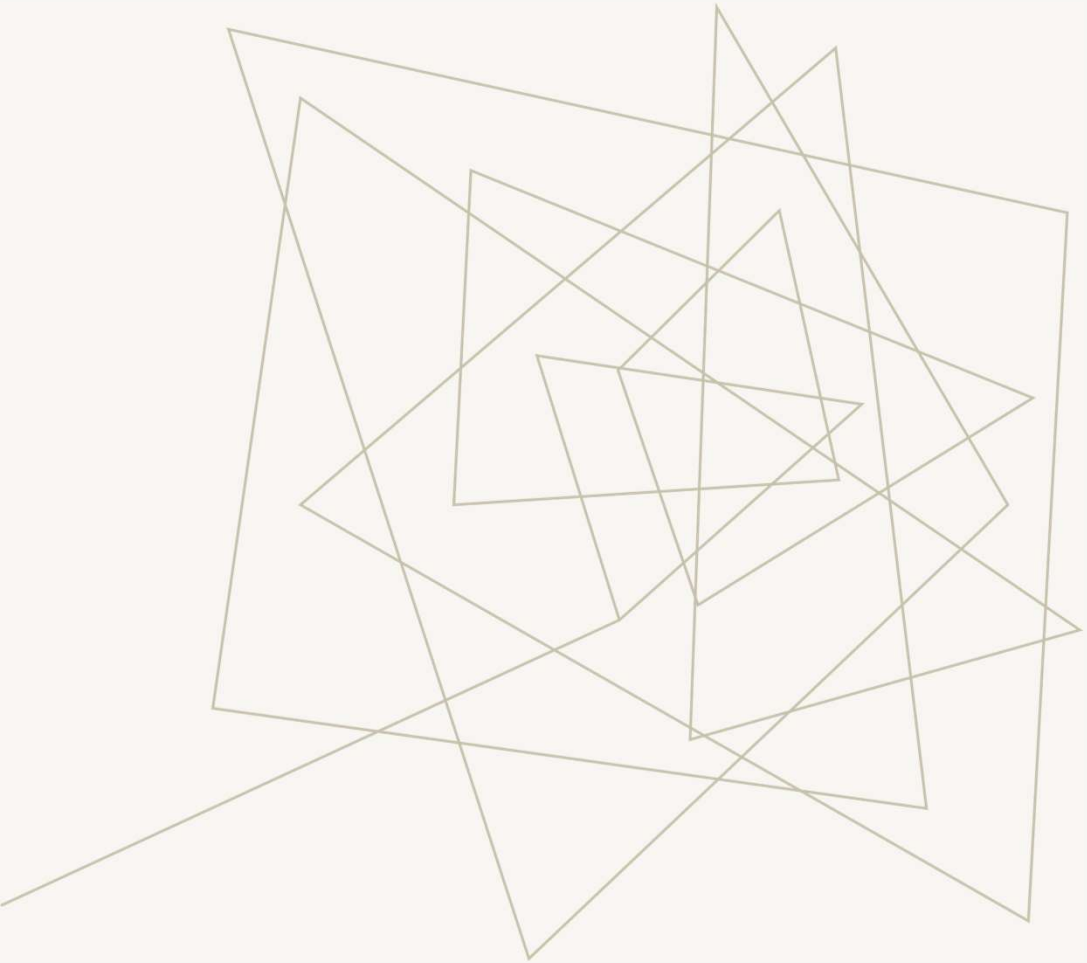
kmeans_label	mean_price	mean_payment_installments	total_payment_value	mean_review_score	mean_freight_ratio	recency	frequency
0	124.548508	2.719402	190.656514	4.996828	0.305722	1785.305028	1.192952
1	133.130872	3.152926	317.764840	1.237283	0.323110	1896.176747	1.412182
2	129.224368	3.075802	196.408483	4.997652	0.296190	2050.801011	1.216128
3	119.608892	2.685653	206.469739	3.703525	0.320001	1796.982432	1.234354
4	121.571291	3.051266	202.048188	3.675635	0.314006	2056.298929	1.235804

Modélisation retenue

Nous retiendrons le modèle du **Kmeans**, car ce dernier permet une segmentation avec un fort intérêt Marketing (clusters multiples), et présente des vitesses de calcul intéressantes.

Nous retiendrons le **Dataset hybride**, composé de sept variables dont les RFM.

Nous retiendrons une segmentation en **5 clusters**, celle-ci présentant la meilleure catégorisation des groupes de clients en fonction de leurs attitudes d'achat ('customer behavior').



SIMULATION DE MAINTENANCE

Jay Corentin – Avril 2023

Simulation de maintenance

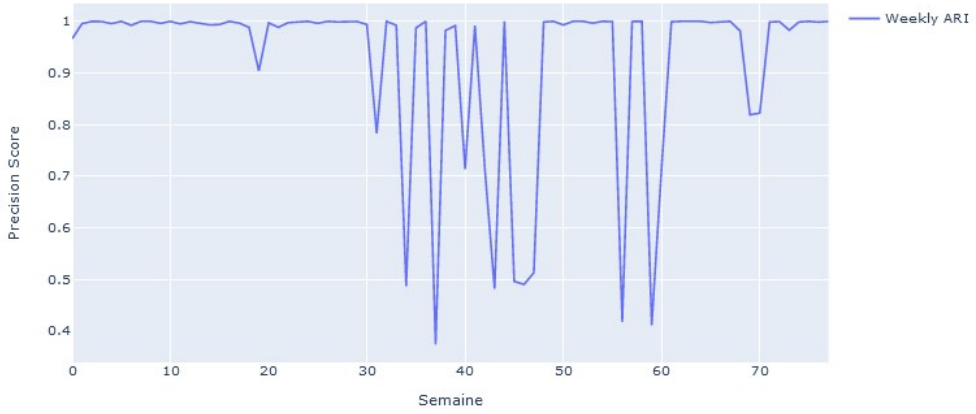
Suppression des périodes avant 2017-03 qui contiennent trop peu de commandes.

Classement temporel du Dataset par date de commande, puis split hebdomadaire.

=> Nous obtenons 79 périodes.

=> Entraînement du Kmeans (clusters = 5)

=> Calcul du score de précision entre chaque période



Le score ARI (score de précision) descend en dessous du seuil d'acceptabilité (0,8) à partir de la 31ème semaine (0,78), soit environ 7 mois.

Afin d'assurer une bonne stabilité de la segmentation, une **maintenance semestrielle** est recommandée.

QUESTIONS