

A photograph of a man with a beard and dark hair, wearing a black leather jacket over a light-colored shirt. He is seated at a wooden desk, looking down at a laptop screen. On the desk, there is a computer monitor displaying the "olist" logo in blue. To the left of the monitor, there's a keyboard and some papers. To the right, there's a mouse and a small stack of books or papers. The background is slightly blurred, showing what appears to be an office environment with other desks and windows.

Projet 5 – Parcours Data Scientist

Segmenter des clients d'un site e-commerce

olist

Fiacre KAKPO - août 2021

Plan de travail

1. **Contexte et objectifs**
2. **Exploratory Data Analysis** : exploration, description des données, analyses uni et multivariées (taille et type de données, correlations, Nan ?, corrélations ?, insights ?...)
3. **Segmentation RFM** basée sur la **Récence** du dernier achat, la **Fréquence** d'achat du client et le montant total (**Monetary**) dépensé par le client
4. **Sélection et Apprentissage** d'un modèle de clustering
5. **Évaluation** des métriques (choix du nombre de clusters)
6. **Caractérisation** des clusters
7. **Evaluation de la stabilité** des clusters et proposition d'un contrat de maintenance

Contexte

Généralités

- Olist est une solution de vente sur les marketplaces en ligne au Brésil

Objectifs généraux:

- Segmenter les clients pour les campagnes de communication futures :
 - ◆ campagnes d'emailing;
 - ◆ lancement produits;
 - ◆ augmentation des recettes par client.



olist

para todos os tipos de negócio, inclusive o seu

Com olist shops você constrói uma loja virtual para divulgar produtos ou serviços por WhatsApp e redes sociais



olist

planos e preços dúvidas parceiros d2c loja olist mais

área do cliente querer vender no olist

olist store

suas vendas em todos os lugares

quero saber mais

Introduction

Cette étude consistera à segmenter la base clientèle d'Olist afin de **comprendre le comportement d'achat des clients.**

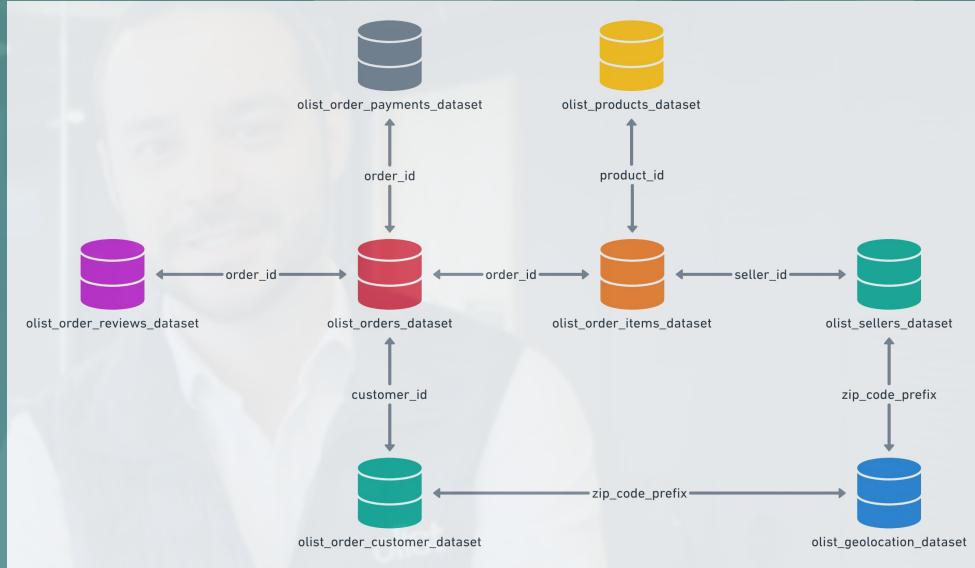
De façon spécifique, il s'agira de :

- Associer chaque client unique à un groupe caractéristique (cluster);
- **Caractériser les clusters**, les variables qui les définissent, leurs tailles;
- Évaluer la **stabilité temporelle** des clusters

Structure du jeu de données

9 datasets dont 8 retenus pour l'étude :

- Clients
- Produits
- Lignes de commande
- Entête de commande
- Règlement des commandes (paiement)
- Commentaires des clients
- Géolocalisation des clients
- Catégorie des produits

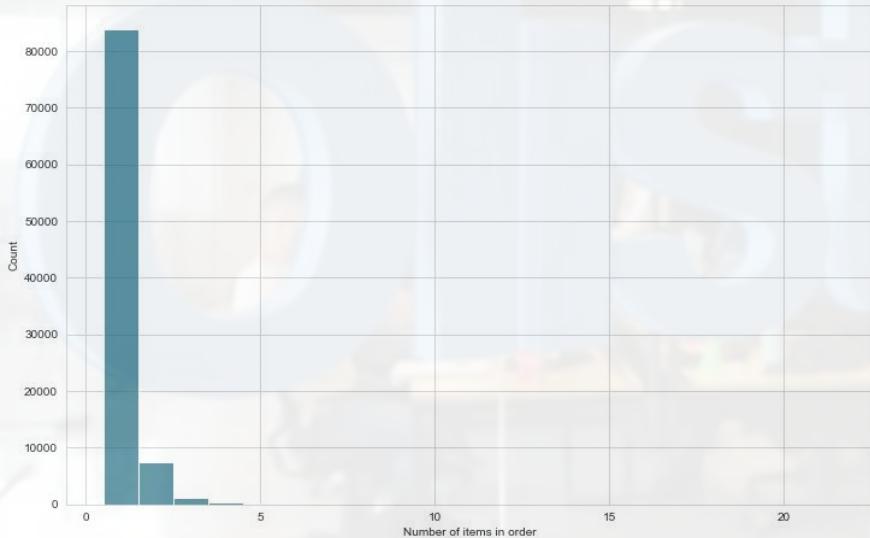


Cette base de données s'appuie sur un modèle relationnel. Ceci facilite la jointure (concaténation des tables) et se basant sur les différentes clés (primaires et secondaires)

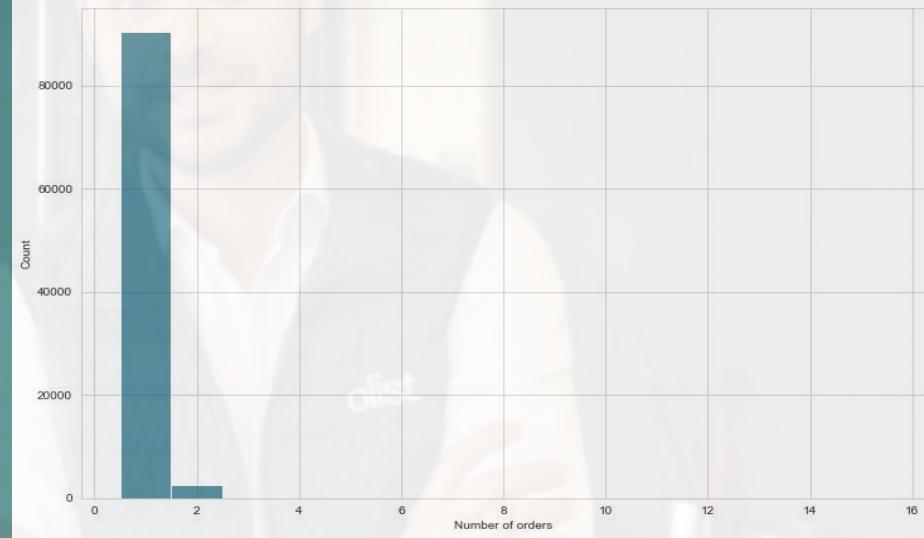
Les grandes étapes de l'EDA (1/7)

Nombre moyen d'articles par commande

Nombre moyen d'articles par commande



Nombre de commandes par client

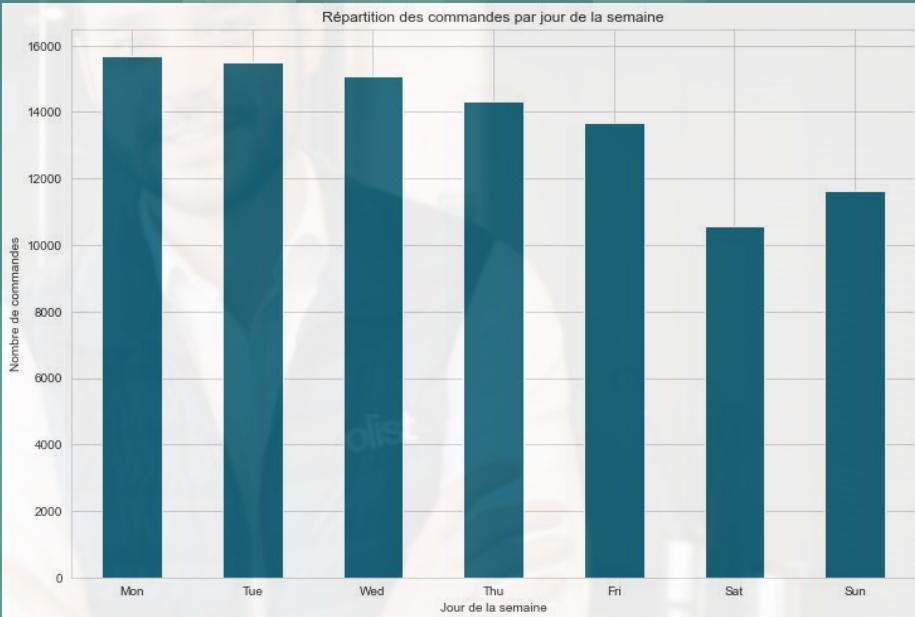
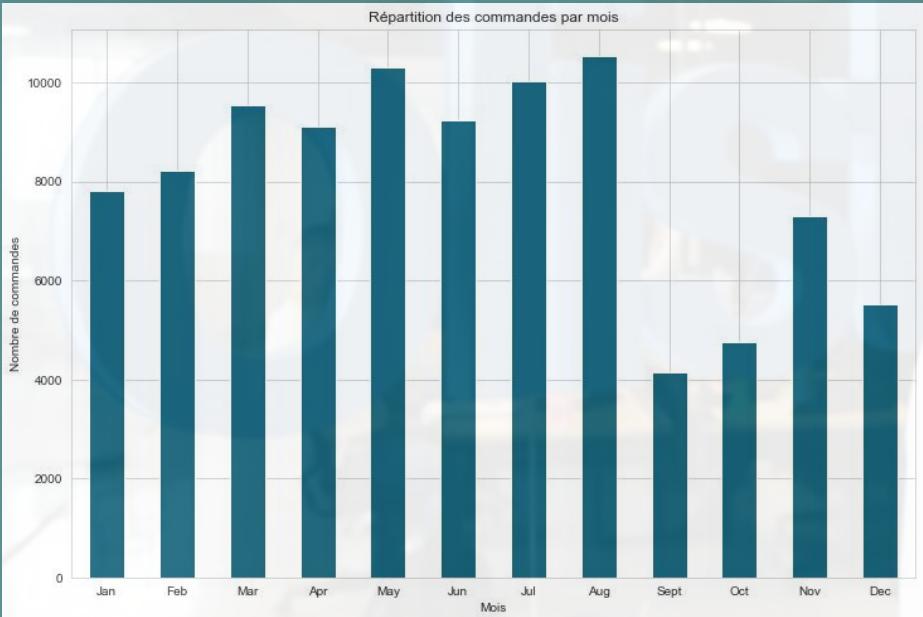


La plupart des clients a en moyenne une commande.

La plupart des commandes concerne en moyenne un article, donc une seule catégorie

Les grandes étapes de l'EDA (2/7)

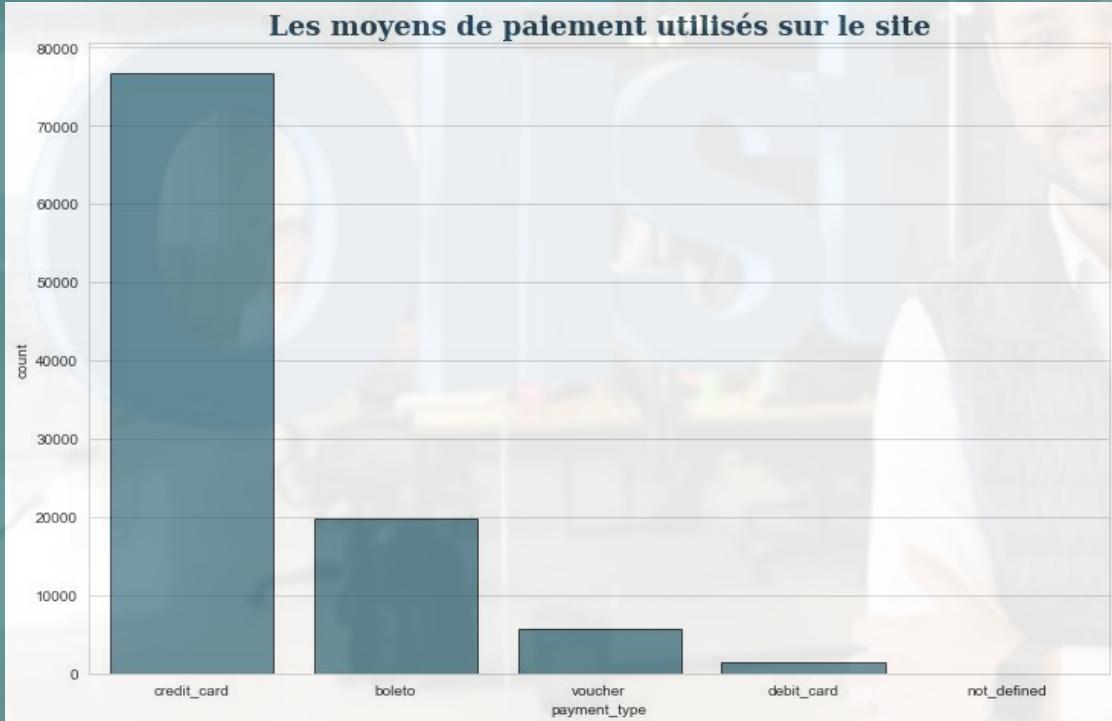
Répartitions des commandes par mois et par jour de semaine



On observe que les ventes évoluent en **fonction du mois ou du jour**. Ce constat a entraîné la création de la variable “**sale_month**”. **Sale_month** recueille le **mois d'une commande**.

Les grandes étapes de l'EDA (3/7)

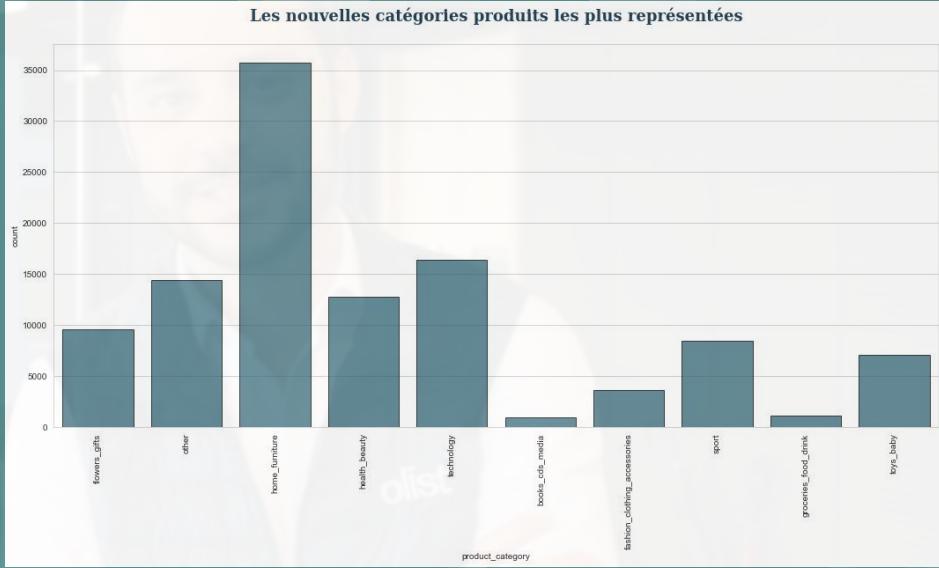
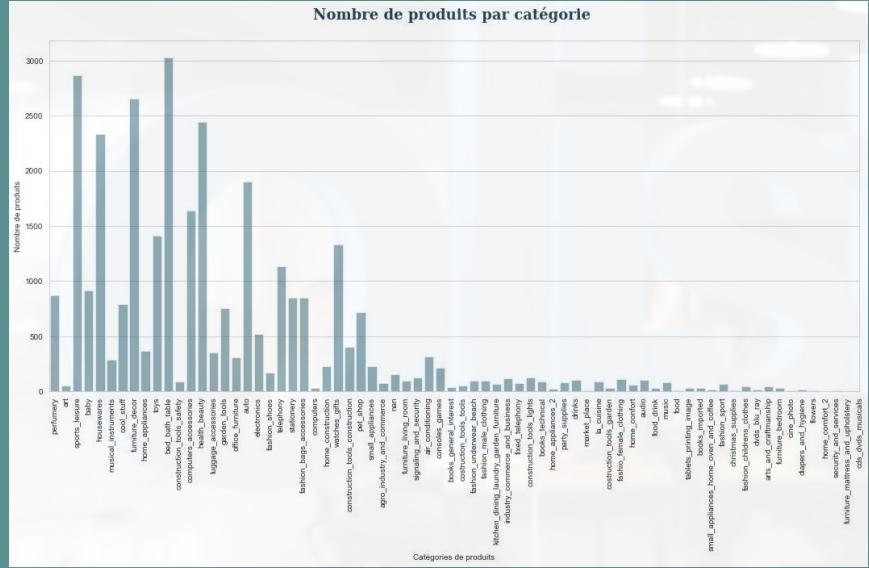
Les moyens de paiement utilisés sur Olist



Les **moyens de paiement** semblent très fortement **dominés** par les **cartes de crédits**, à plus de 90%. La variance étant faible, il semble plus judicieux de **supprimer la variable payment_type** (type de paiement)

Les grandes étapes de l'EDA (4/7)

Transformation des catégories de produits (1/2)

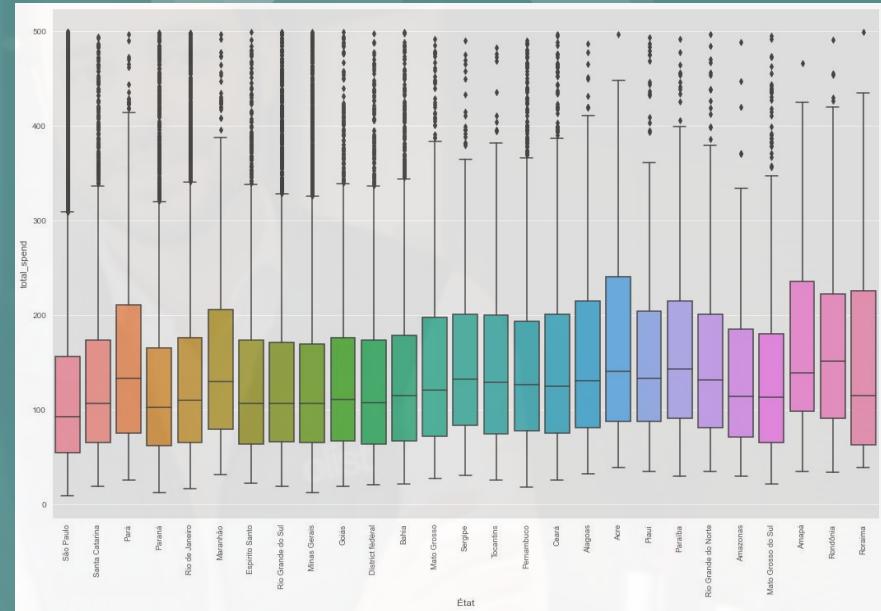
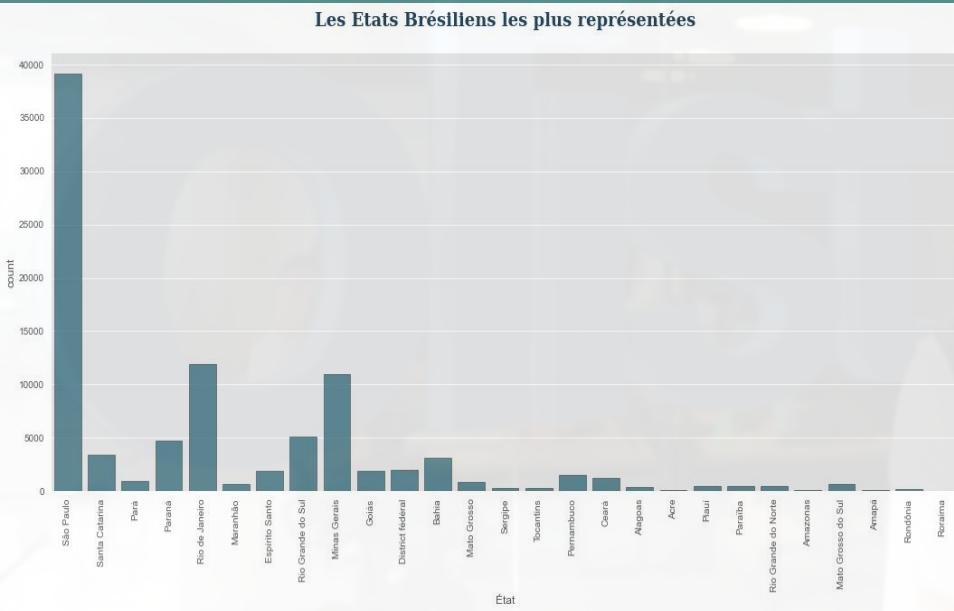


Passage de plus d'une soixantaine de catégories produits à 10, grâce au regroupement sur la base des catégories principales des sites de vente en ligne en 2017.

Les grandes étapes de l'EDA (5/7)

Les Etats les plus représentés et les dépensiers

Les Etats Brésiliens les plus représentées

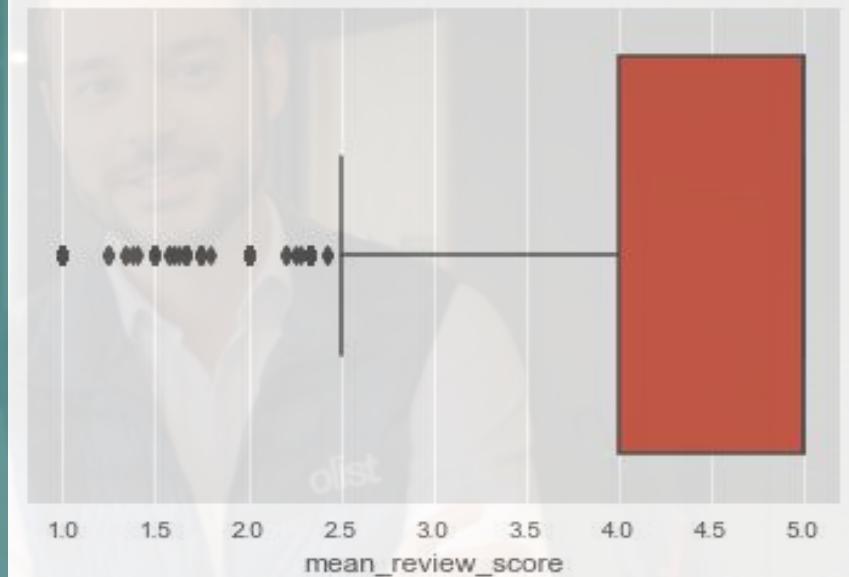
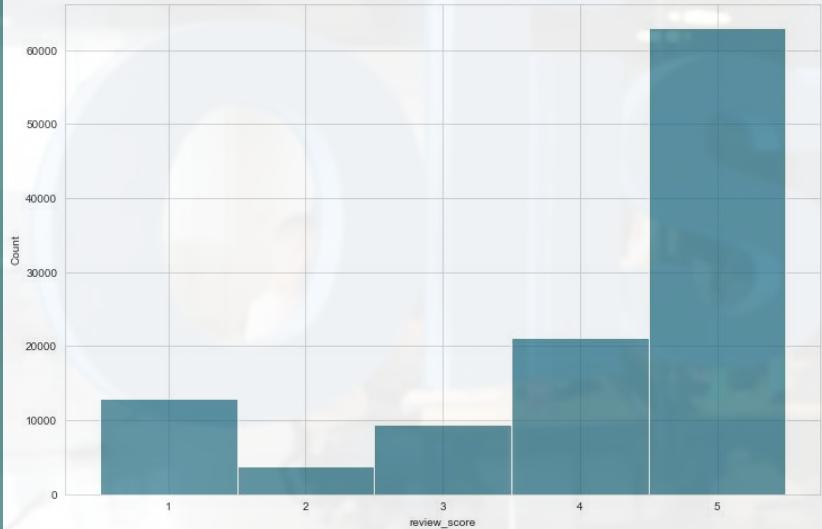


On remarque que la plupart des clients uniques sont surtout présents dans l'État de São Paulo. Par contre, sur le volet dépense, les disparités avec les autres Etats sont peu visibles ou quasi-inexistant. La corrélation Etat - dépense est peu évidente. Nous avons décidé donc de supprimer la variable liée à la géolocation du client.

Les grandes étapes de l'EDA (6/7)

Répartition des notes clients

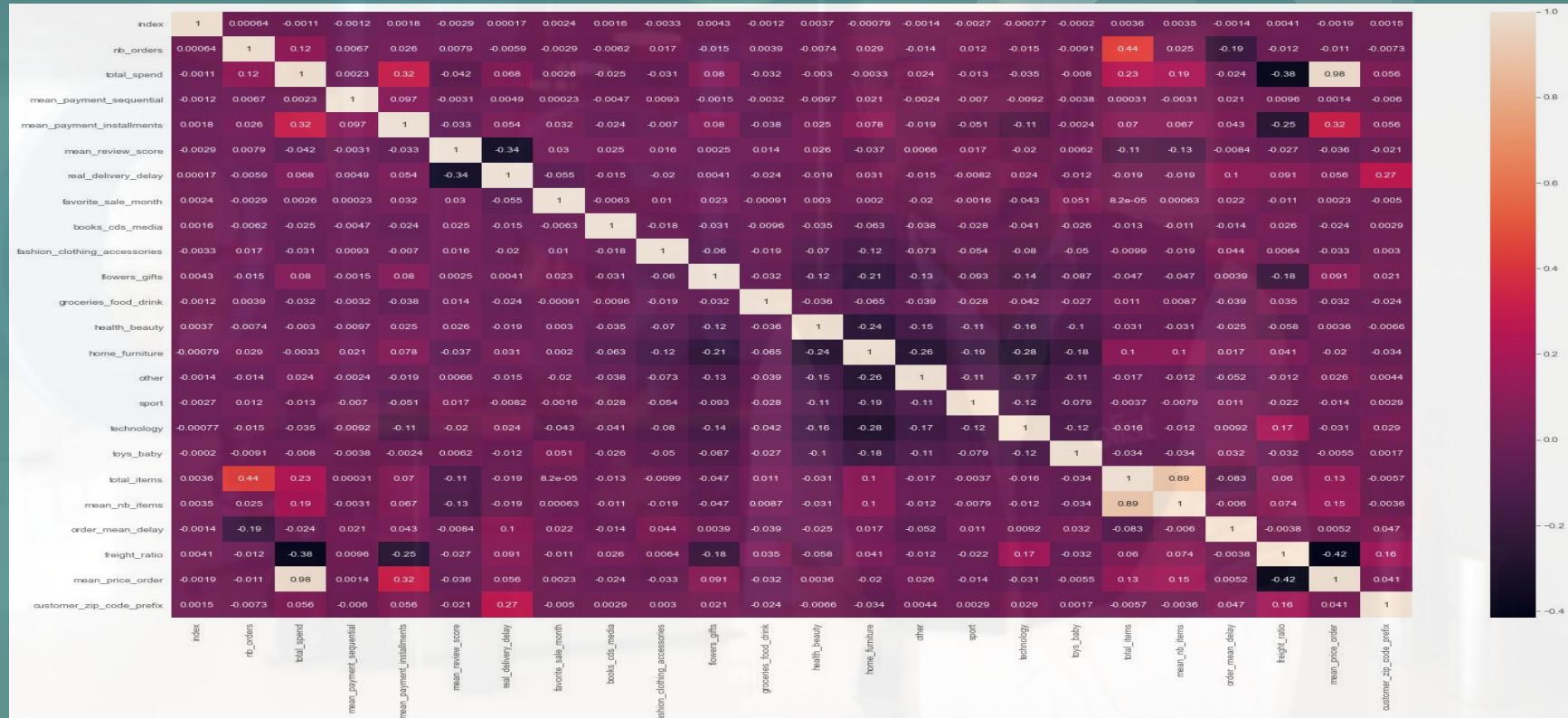
Répartition des notes attribuées aux commandes



Note client dominé par la note 5.

Les grandes étapes de l'EDA (7/7)

Heatmap des corrélations linéaires



Autres créations de variables

```
cust_data_master["order_total_delay"] = cust_data_master["order_total_delay"] / cust_data_master["nb_orders"]
cust_data_master = cust_data_master.rename(columns={"order_total_delay": "order_mean_delay"})
```

```
cust_data_master["freight_ratio"] = round(cust_data_master["total_freight"] / (
    cust_data_master["total_spend"] + cust_data_master["total_freight"]), 2)
cust_data_master["mean_price_order"] = round(
    cust_data_master["total_spend"] / cust_data_master["nb_orders"], 2)
cust_data_master["total_spend"] = (
    cust_data_master["total_spend"] + cust_data_master["total_freight"])
cust_data_master1 = cust_data_master.drop("total_freight", axis=1)
```

```
data["real_delivery_delay"] = (data.order_delivered_customer_date
                               - data.order_purchase_timestamp).dt.round('1d').dt.days
```

- La part des frais de livraison dans les dépenses totales du client;
- Les dépenses moyennes du client;
- Montant total dépensé par le client;
- Durée entre la date d'achat et la date de livraison.

Segmentation selon la méthode RFM

		
Récence La proximité du dernier achat Ex : durée depuis le dernier achat	Fréquence Récurrence des achats sur une période Ex : nombre d'achats sur la dernière année	Montant Valeur client sur une période Ex : Somme de tous les montants d'achat sur la dernière année

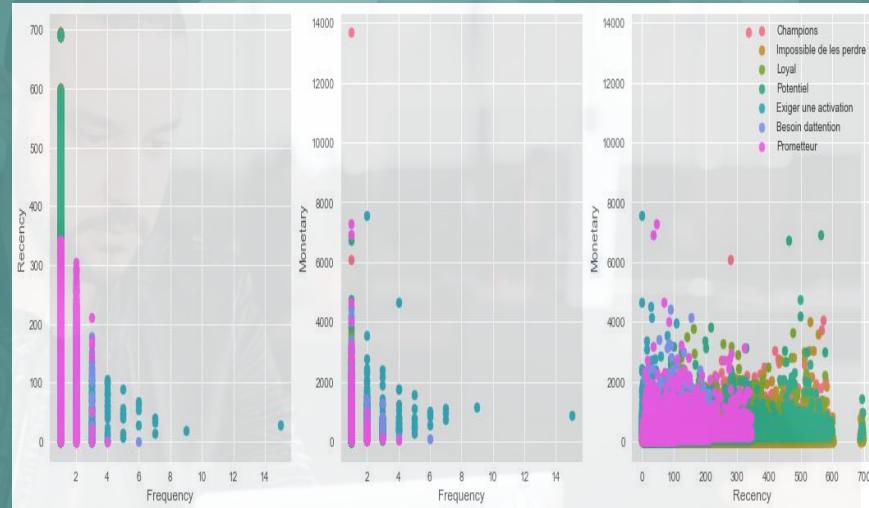
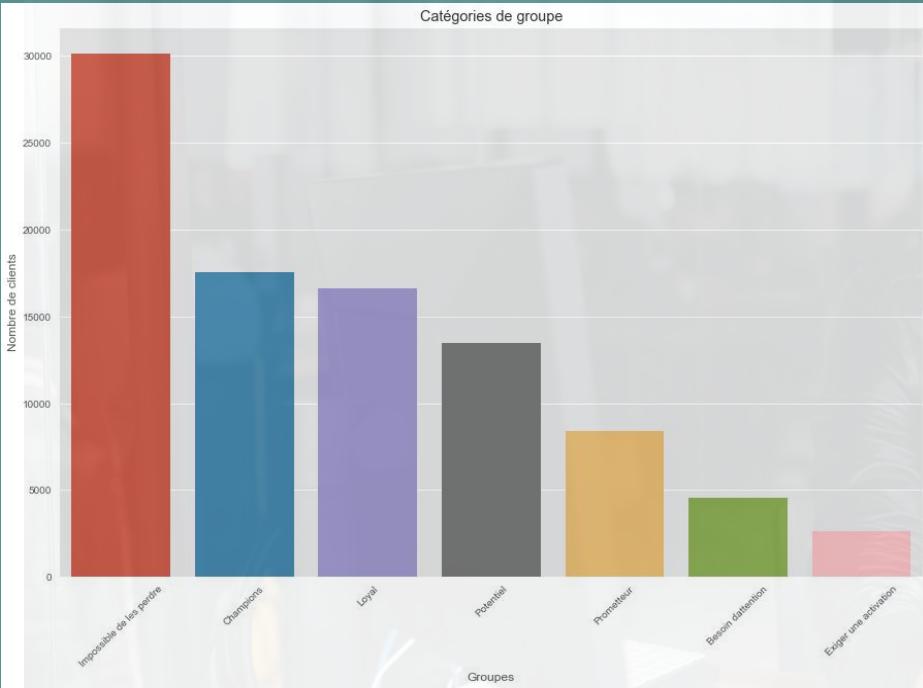
- Technique utilisée en Marketing;
- Basée sur l'historique d'achat des clients;
- 3 Variables calculées par client.

Recency : Durée entre aujourd'hui (date la plus récente enregistrée) et le dernier achat

Frequency : Fréquence d'achat

Monetary : Somme que le client a dépensé

Segmentation selon la méthode RFM



	Recency	Frequency	Monetary	r_quartile	f_quartile	m_quartile	RFM_Score	RFM_Cluster
0	111	1	141.90	2	4	2	8	Champions
1	114	1	27.19	2	4	4	10	Impossible de les perdre
2	537	1	86.22	4	4	3	11	Impossible de les perdre
3	321	1	43.62	3	4	4	11	Impossible de les perdre
4	288	1	196.89	3	4	1	8	Champions
5	146	1	166.98	2	4	2	8	Champions
6	131	1	35.38	2	4	4	10	Impossible de les perdre
7	182	1	419.18	2	4	1	7	Loyal
8	543	1	150.12	4	4	2	10	Impossible de les perdre
9	170	1	129.76	2	4	2	8	Champions

Modèle de clustering

Les modèles de clustering testés :

- K-Means
- K-Prototype

Métriques testés :

- Inertie (“Méthode de coude” - “Elbow Method”);
- Coefficient de Silhouette;
- Coefficient de Calinski Harabasz;
- Adjusted Rand Index (ARI).

La méthode K-means (méthodes des centroïdes)

Principe :

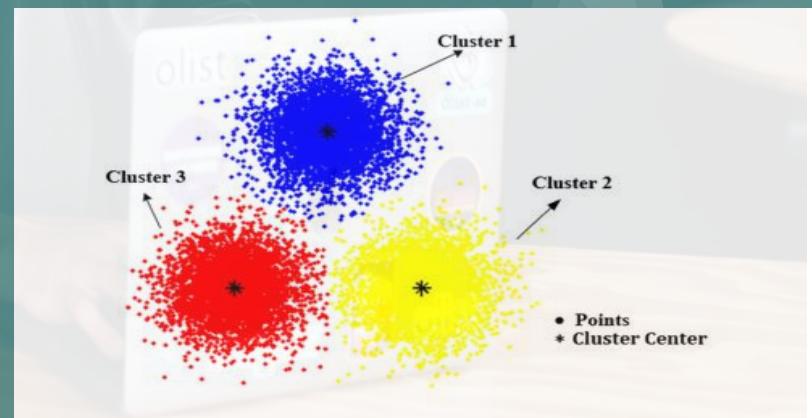
- Le K-means (k-moyennes) est un algorithme non supervisé de clustering;
- Il permet de regrouper en K clusters distincts les observations d'un dataset. Ainsi les données similaires se retrouveront dans un même cluster;
- L'algorithme du k-means travaille avec les centres de gravité des groupes;
- La méthode des k-means repose sur la minimisation de la somme des distances euclidiennes au carré entre chaque objet (ou sujet, ou point) et le centroïde (le point central) de son cluster.

The diagram shows the mathematical expression for the K-means objective function:

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2$$

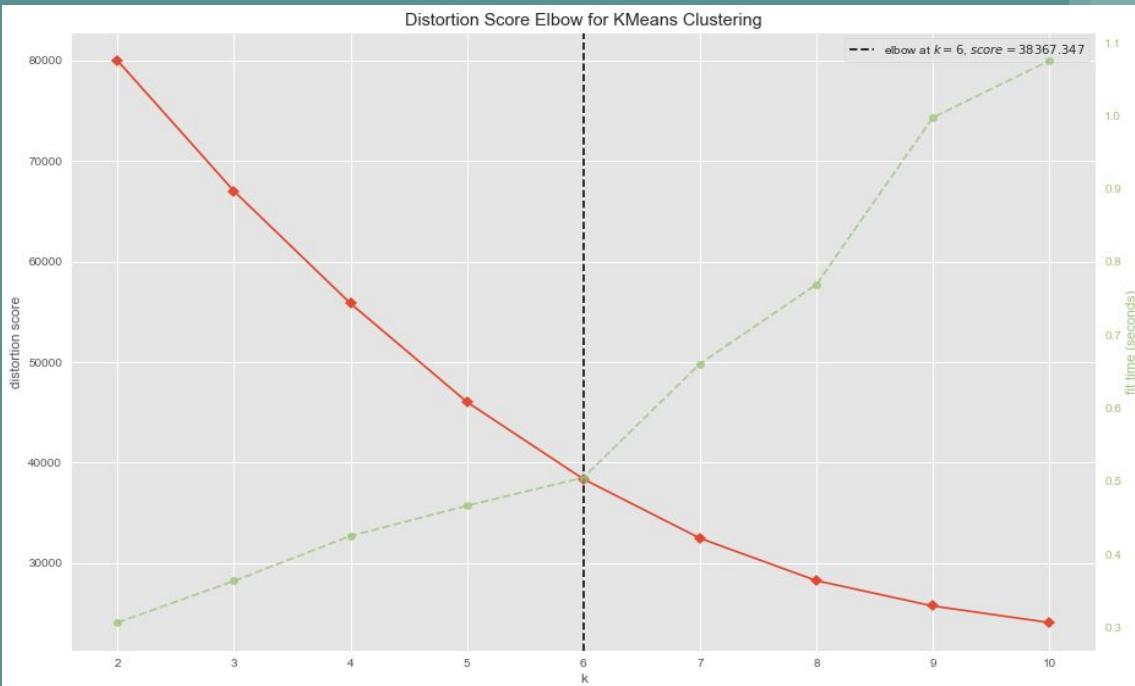
Annotations explain the components:

- number of clusters $\rightarrow k$
- number of cases $\rightarrow n$
- case i $\rightarrow x_i^{(j)}$
- centroid for cluster j $\rightarrow c_j$
- Distance function $\rightarrow \| \cdot - \cdot \|^2$



K-Means - la méthode du coude (Elbow)

La méthode du coude est une approche utilisée dans détermination du nombre de clusters dans un ensemble de données. Elle consiste à tracer la variation expliquée en fonction du nombre de clusters, et en choisissant le coude de la courbe comme le nombre de clusters à utiliser.



Grâce à la méthode du coude basée sur **le score de distorsion** (somme moyenne des carrés des distances des points par rapport aux centres des clusters), une segmentation en **K = 6** clusters serait la meilleure option.

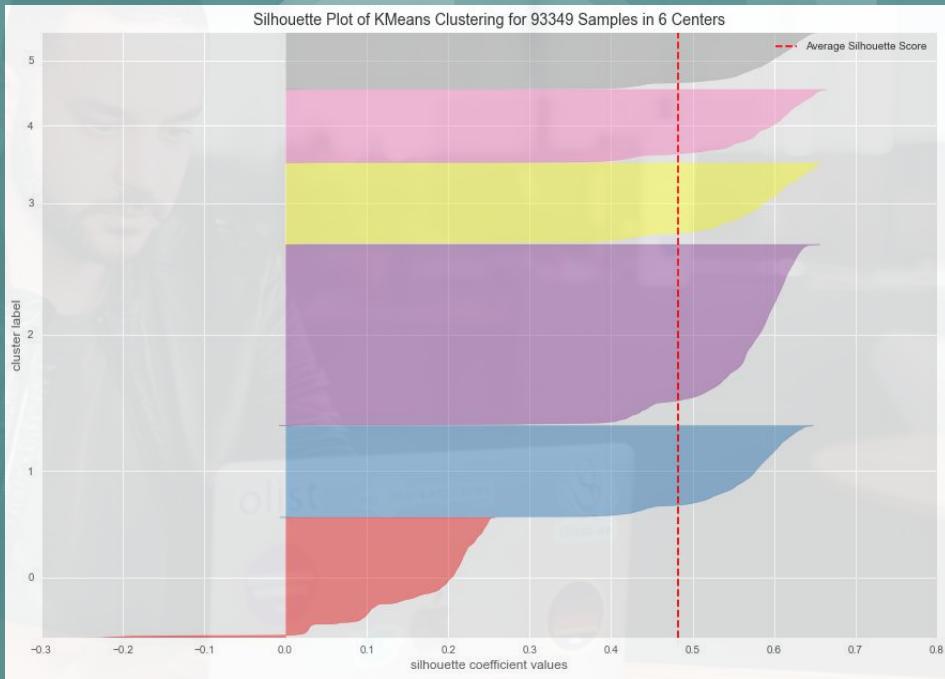
K-Means - Coefficient de silhouette

Le **coeffcient de silhouette** mesure la qualité (homogénéité et séparation) des clusters.

Pour chaque point, son coefficient de silhouette est la différence entre la distance moyenne avec les points du même groupe que lui (cohésion ou homogénéité) et la distance moyenne avec les points des autres groupes voisins (séparation).

- Si **cette différence est négative**, le point est en moyenne plus proche du groupe voisin que du sien : **il est donc mal classé**.
- A l'inverse, si **cette différence est positive**, le point est en moyenne plus proche de son groupe que du groupe voisin : il est donc **bien classé**.

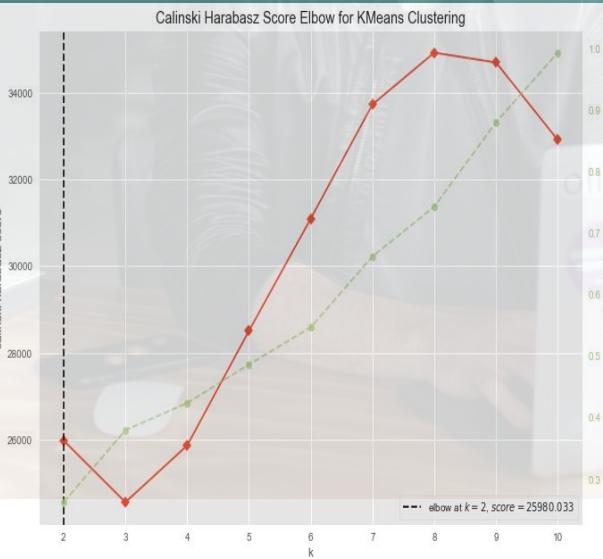
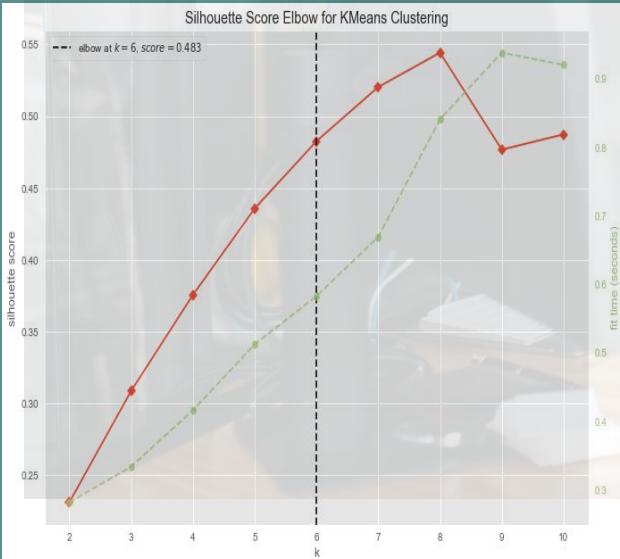
Le **coeffcient de silhouette proprement dit** est la moyenne du coefficient de silhouette pour tous les points.



Dans notre cas, les clusters semblent relativement bien répartis et les séparations sont relativement claires avec cependant quelques erreurs sur l'un des clusters (le 1er cluster)

K-Means - Test d'autres métriques

- Coefficient de Silouhette : rapport moyen entre la distance intra-cluster et la distance du cluster le plus proche
- Indice de Calinski-Harabasz : rapport entre la dispersion des grappes dans et entre les groupes.



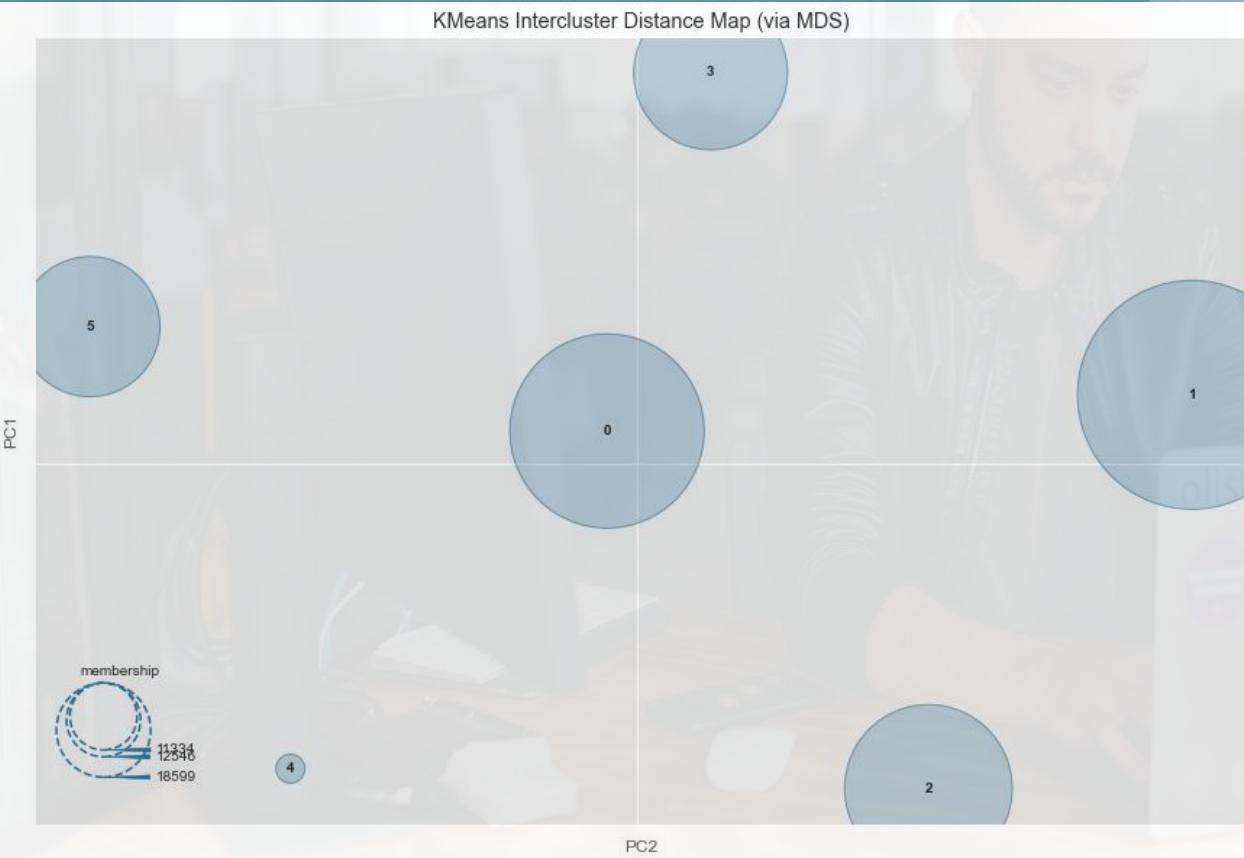
Observation :

Pour la métrique score silhouette, le nombre de K est également de 6.

Pour le score Calinski Harabasz, le meilleur K est plus incertain.

Les scores sur la répartition en 6 clusters semblent être meilleurs.

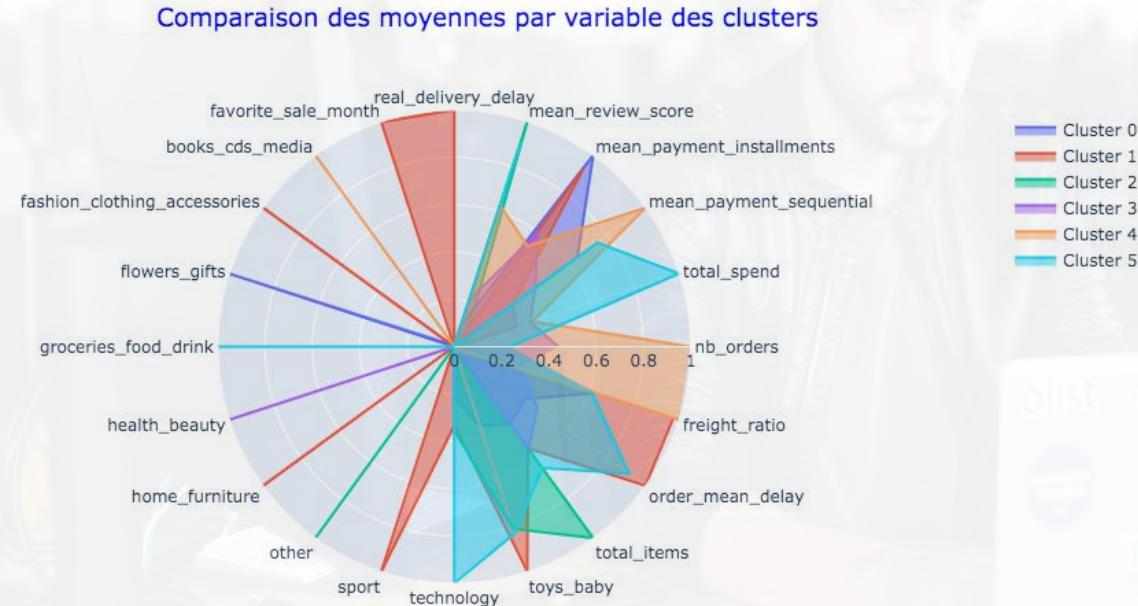
K-Means - Distances interclusters



Sur cette projection 2D, les différents clusters semblent bien séparés sur les 2 premières composantes principales. Le clustering semble donc performant. Il sera important d'**identifier les composantes métier de chaque cluster**.

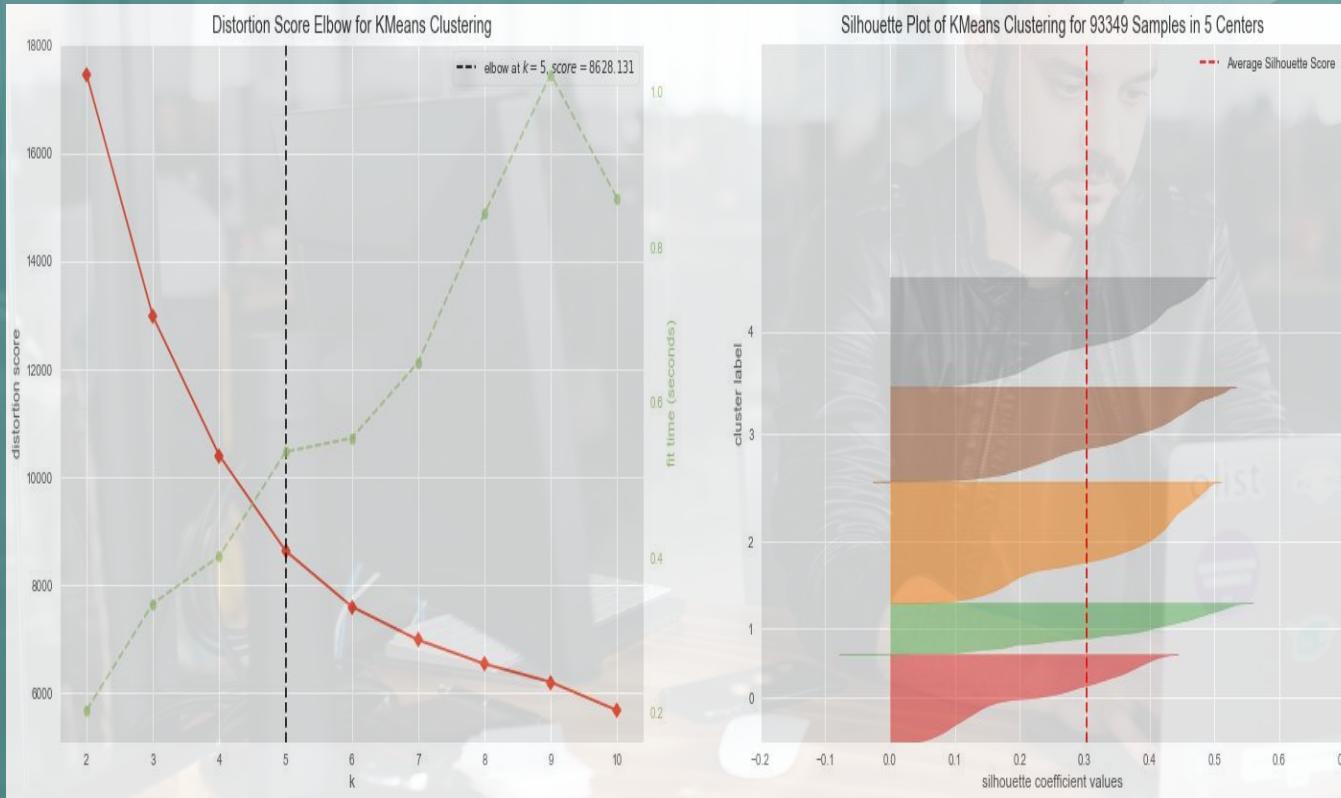
K-Means - Diagramme en Radars (Caractérisation des clusters)

Visualisation avec le poids des catégories produits



Les catégories produits semblent masquer les autres variables; I

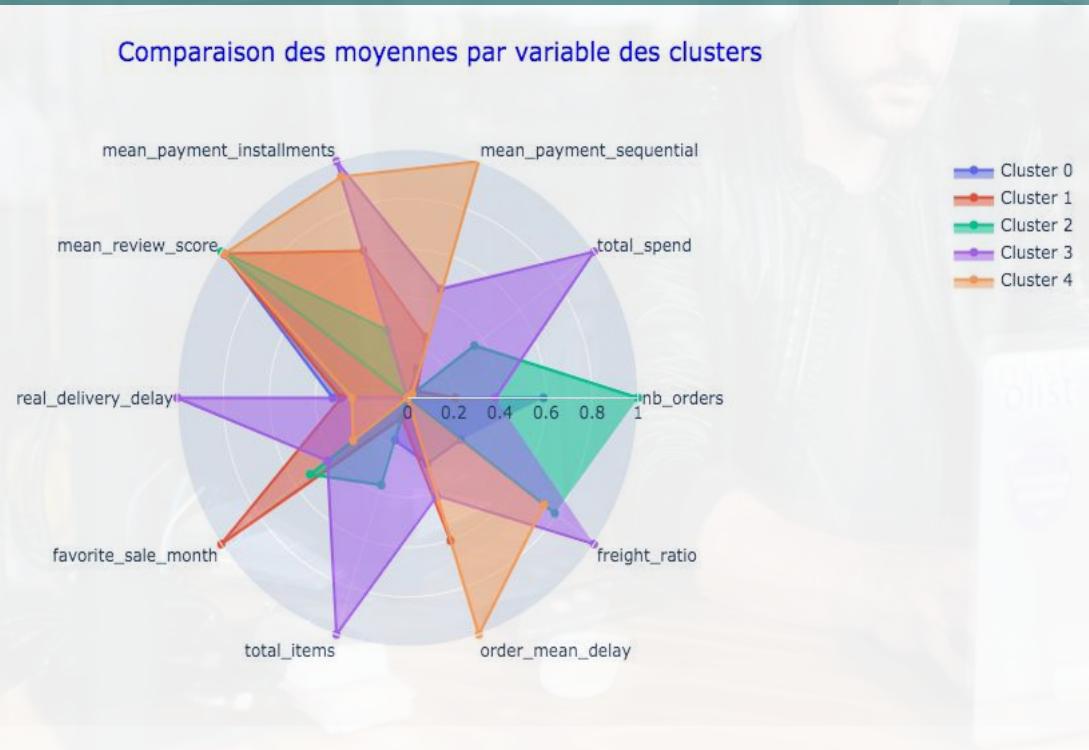
K-Means SANS les catégories produits - méthode de coude et coefficient de silhouette



Le nombre de clusters est passé à :
5. Les Clusters semblent denses et nettement séparés

Diagramme en radars (Sans les variables catégories Produits)

Le diagramme de Kiviat, diagramme en radar , en étoile ou encore en toile d'araignée sert à représenter sur un plan en deux dimensions au moins trois ensembles de données multivariées. Chaque axe, qui part d'un même point, représente une caractéristique quantifiée.



Nombre de cluster : 5

Caractérisation des clusters

Groupe 1 : Ce sont des clients très satisfaits, commandant généralement en fin d'année pour des montants faibles. Ils bénéficient d'un court délai de livraison, utilisent relativement plusieurs moyens de paiement. Ils commandent peu d'articles pour un nombre faible de commandes.

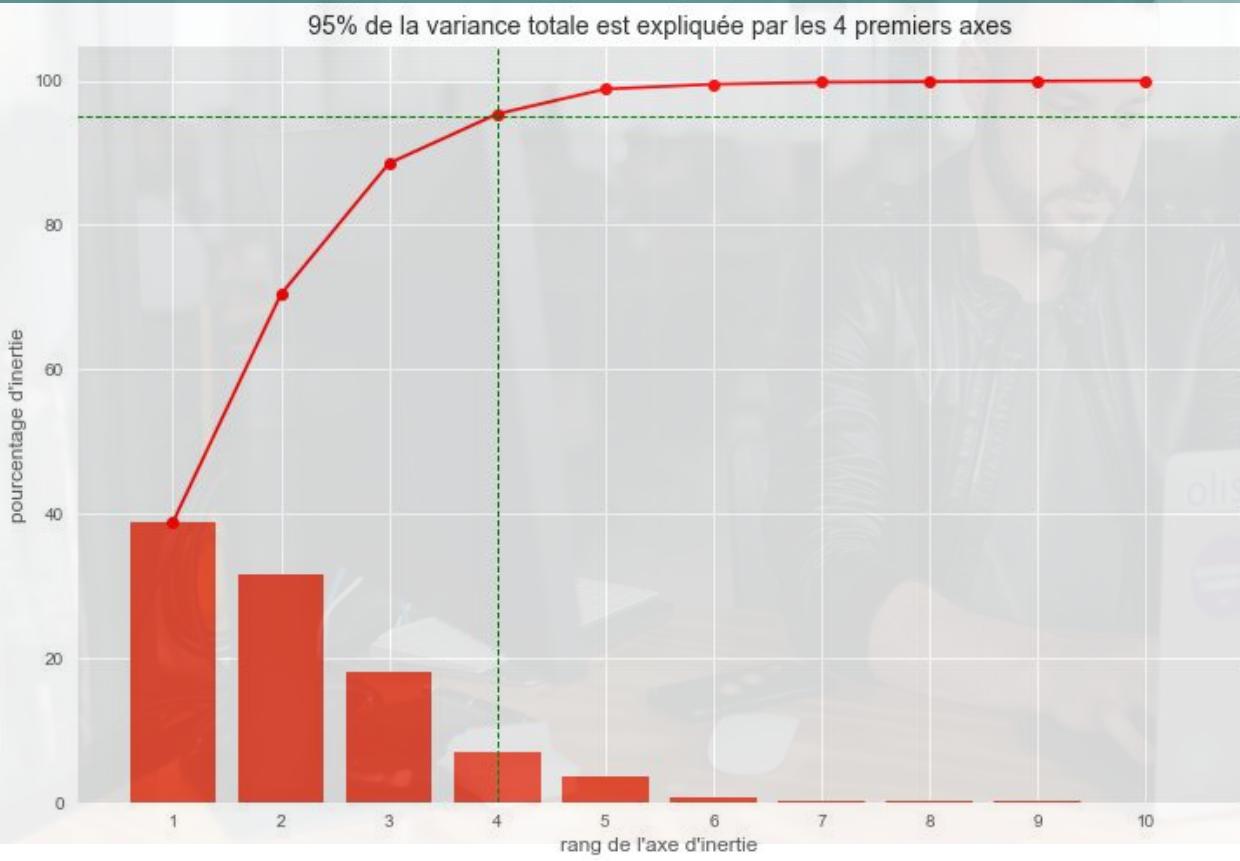
Groupe 2 : Clients satisfaits, de début d'année, avec plusieurs commandes mais un faible nombre d'articles. Ils dépensent peu, se font livrer dans un délai raisonnable. Ce sont des clients qui paient en une seule tranche.

Groupe 3 : Clients mécontents, effectuant leurs achats généralement en milieu d'année. Leurs commandes contiennent un nombre important de produits, ce qui nécessite des frais de livraison élevés, et des délais de livraison assez longs. Ils dépensent énormément et préfèrent échelonner le paiement. Les moyens de paiement utilisés varient peu.

Groupe 4 : Clients qui ont en général, une seule commande. Leur retour est positif d'autant que la livraison se fait assez rapidement. Ils achètent généralement au premier trimestre. Et même si le montant total dépensé est faible ainsi que le nombre d'articles achetés, les frais de livraison sont relativement élevés. Il n'est pas exclu que les articles en question aient de grandes dimensions ou qu'ils soient éloignés des boutiques vendeuses. Ce sont des acheteurs qui préfèrent faire les règlements avec plusieurs méthodes de paiement et en plusieurs tranches malgré un montant des achats assez faible.

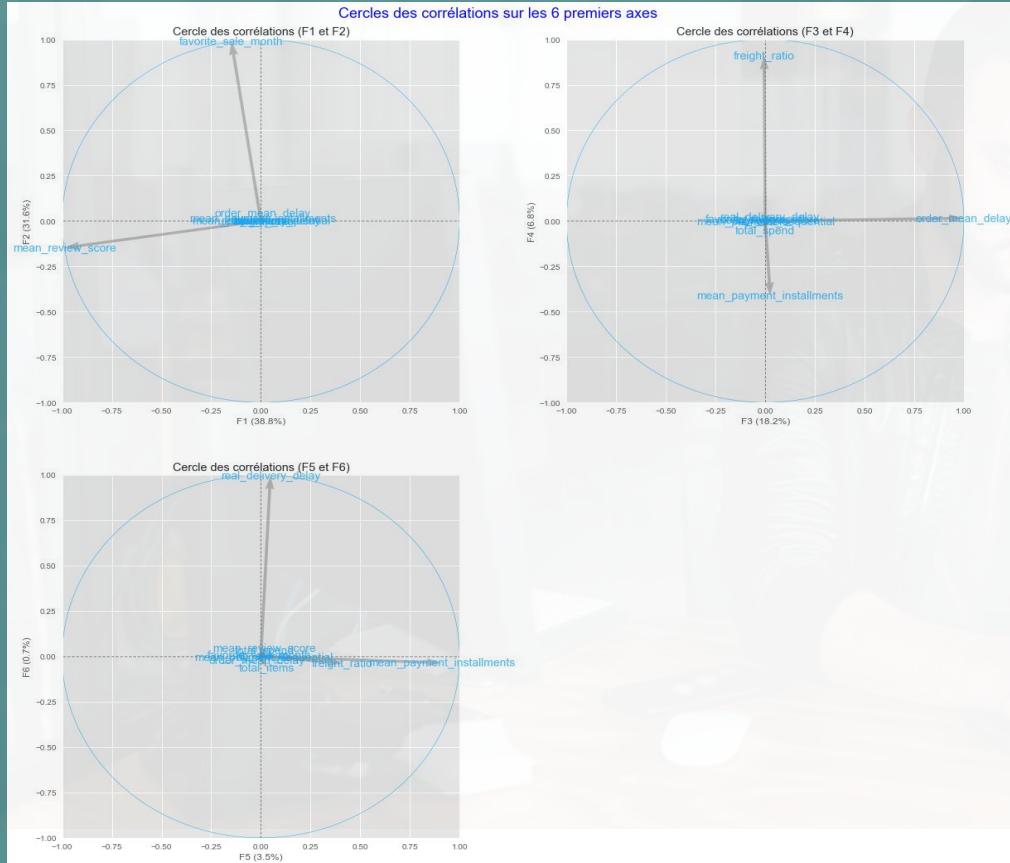
Groupe 5 : Clients fidèles satisfaits aux dépenses faibles. Ils achètent en été, et sont livrés très rapidement.

Réduction dimensionnelle (PCA)



95% de la variance totale est expliquée par les 4 premiers axes.

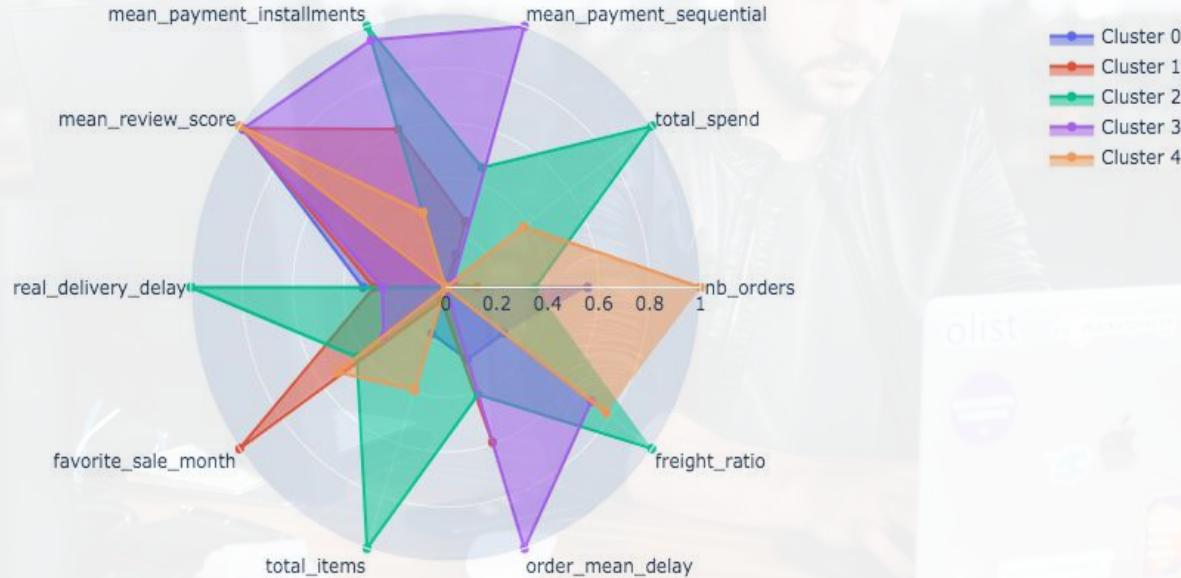
Réduction dimensionnelle (PCA)



Cercles des corrélations

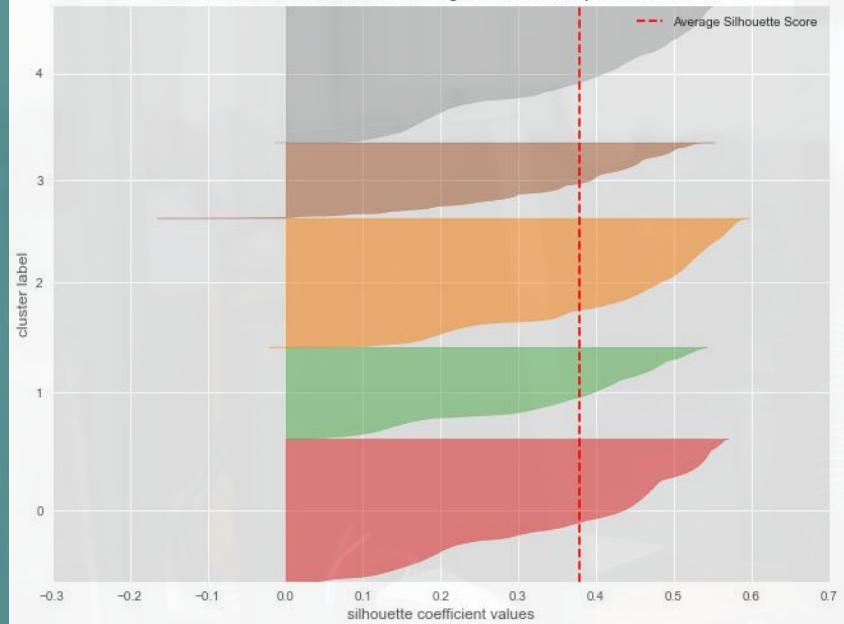
Diagramme en radars

Comparaison des moyennes par variable des clusters

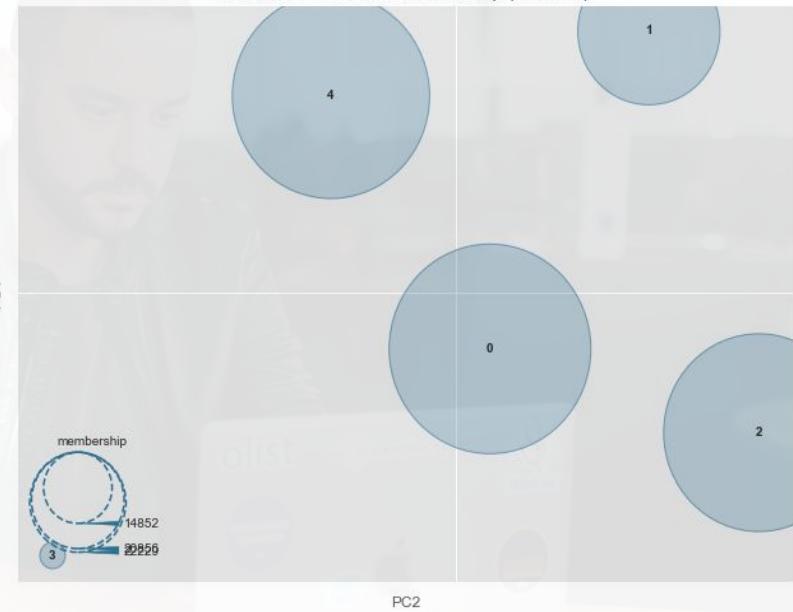


Coefficient de silhouette & Distances interclusters après PCA

Silhouette Plot of KMeans Clustering for 93349 Samples in 5 Centers



KMeans Intercluster Distance Map (via MDS)



Clusters relativement bien séparés et densément répartis

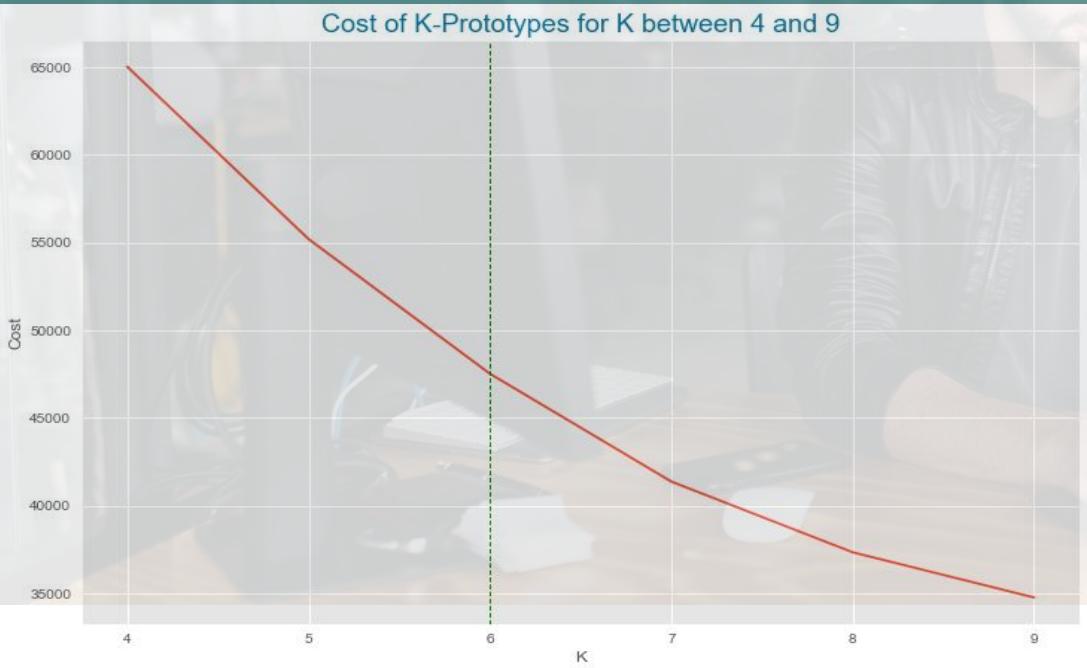
Réduction dimensionnelle (PCA)

Scores de stabilité à l'initialisation

Iteration	FitTime	Inertia	Homo	ARI	AMI
Iter 0	0.068s	6150	0.773	0.648	0.781
Iter 1	0.080s	6165	0.628	0.578	0.645
Iter 2	0.077s	6154	0.771	0.665	0.787
Iter 3	0.067s	5482	0.999	1.000	0.999
Iter 4	0.069s	5482	0.999	1.000	0.999
Iter 5	0.075s	5482	1.000	1.000	1.000
Iter 6	0.085s	6308	0.690	0.616	0.702
Iter 7	0.069s	5482	1.000	1.000	1.000
Iter 8	0.097s	6190	0.607	0.555	0.621
Iter 9	0.082s	6226	0.608	0.560	0.623

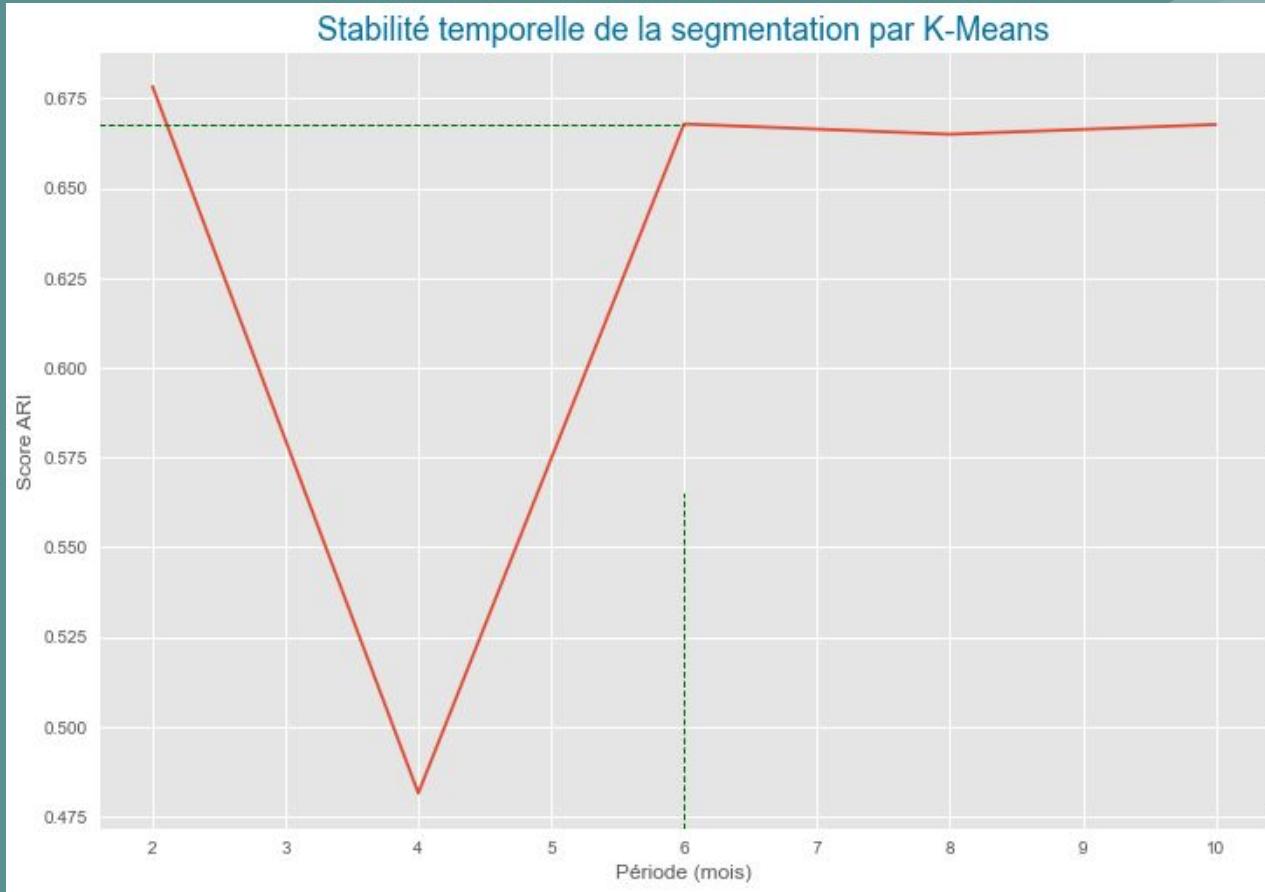
Clustering avec l'algorithme K-Prototype

L'algorithme k-prototypes combine les « moyennes » des variables numériques et les « modes » des variables catégorielles pour construire un nouveau « prototype » hybride du Cluster. Il est cependant gourmand.



K = 6

Stabilité des clusters



Conclusion

- Segmentation RFM limitée, nécessité d'une activité récurrente
- Améliorer la qualité du jeu de données
- Nécessité d'introduire des variables sur les goûts d'achat et type de produits, fréquence de consultation du site,...
- Nécessité d'améliorer le coefficient de similarité des clusters avant d'améliorer la stabilité
- DBSCAN

A man with a beard and dark hair, wearing a black leather jacket over a dark t-shirt, is seated at a wooden desk in an office environment. He is looking down at a silver laptop screen which displays various logos for e-commerce platforms like Olist, Mercado Livre, and AliExpress. On the desk in front of him are several items: a white keyboard, a black smartphone, a blue mouse, a small white notepad, and a pair of glasses. The background is slightly blurred, showing office cubicles and windows.

Merci de votre
attention !!!