Projet pour segmenter des clients d'un site e-commerce

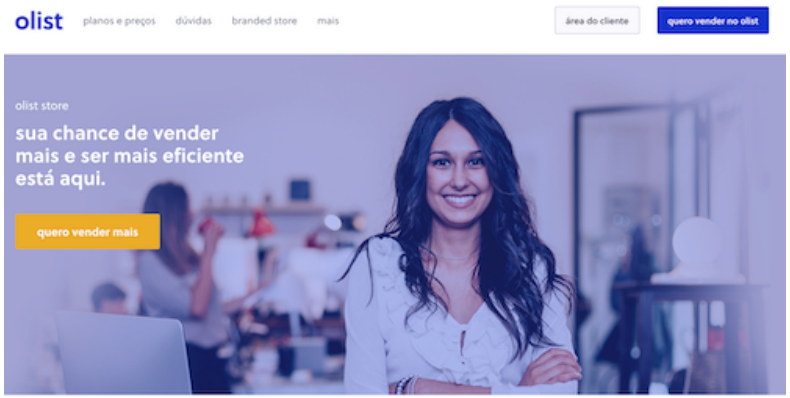
Cyril REGAN

Base de données : kaggle.com/olistbr/brazilian-ecommerce/

28 mai 2020

Table of Contents

Problématique



Besoin Olist : Segmentation des clients pour leur campagne de communication

- Comprendre les types d'utilisateurs
- *"Description actionable"* = Proposition de communications ciblées
- *"Proposition de contrat de maintenance"* = identifier la fréquence (ex : tous les mois) pour refaire l'analyse client

Table of Contents

Préparation des données

Nettoyage :

- Compilation des 9 fichers (products_category_translate - customers - orders - geolocation - products - items - sellers - payments - reviews) en un **unique** tableau
- Suppression commandes/produits mal définis (-4.3% des commandes)
- Imputation des données de géolocalisation.

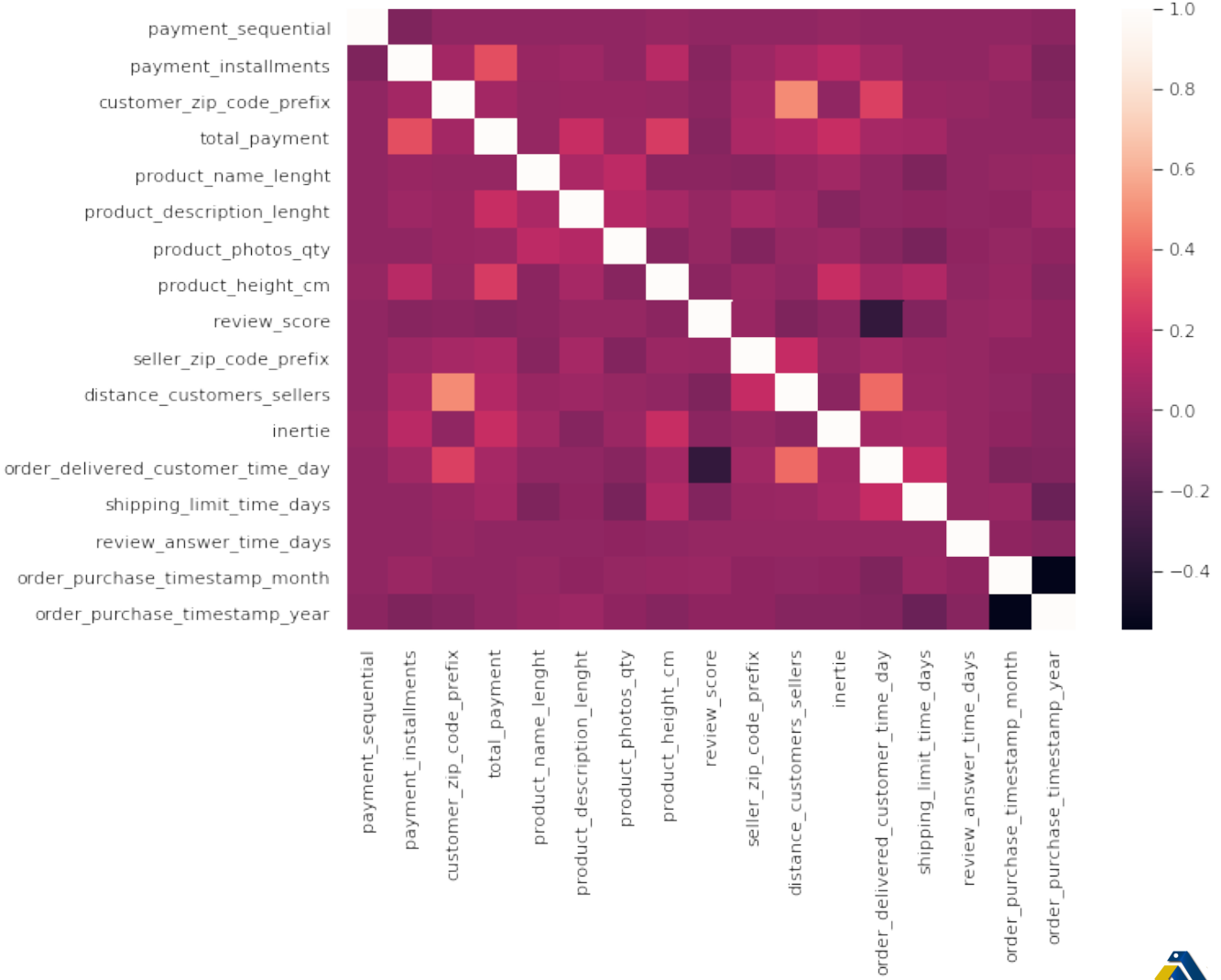
Transformation :

- des dates en **durée** ("order_purchase_timestamp", "order_delivered_customer_time","shipping_limit_time", "review_answer_time")
- des données géolocalisées en **distance** ("distance_customers_sellers"). Suppression des autres données géolocalisées autre que les pays.

Table of Contents

Corrélations

17 variables quantitatives

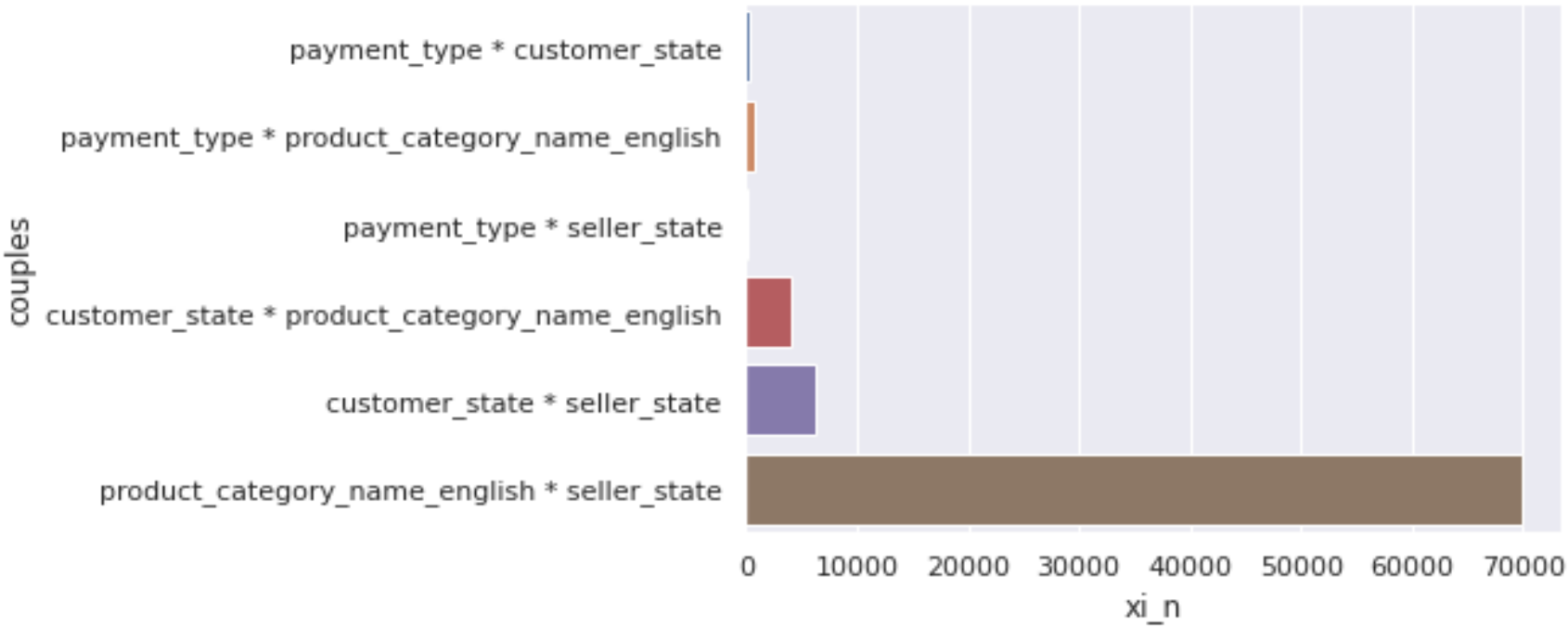


=> **Tranformation** largeur/longueur/poids des produits : **inertie** et **hauteurs**

Table of Contents

Analyse variables qualitatives

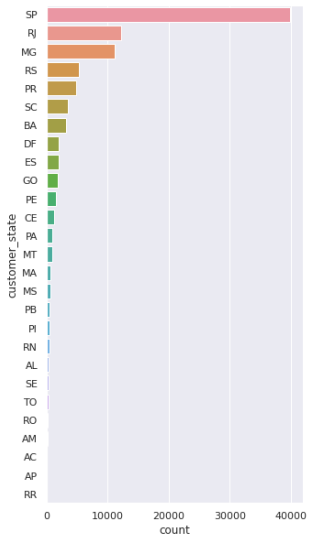
4 variables : type de paiement , catégorie de produit, états des clients, états des vendeurs. Contingence multiples (ξ_n) :



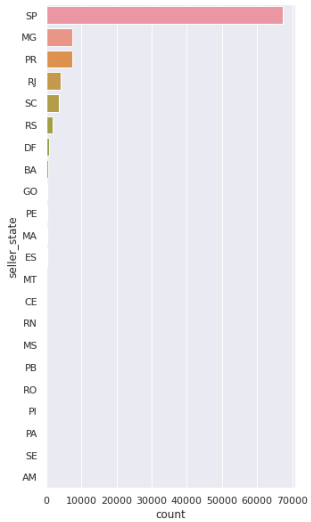
Fait marquant :
Contingence de états vendeurs * catégories produit :
$\xi_{computers,BA}/\xi_n = 0.26$, soit 144 ordinateurs sur les 549 vendus proviennent de l'état de Bahia

- **“Etats des vendeurs” fortement dépendant** des “catégories de produit” et des “Etats des clients”

états clients



états vendeurs



- **Etats des clients** mieux répartis
- **Etats des vendeurs** $\subset$ **distance client-vendeur** indirectement

$\Rightarrow$ On supprime la variable “états vendeurs”

Conclusion de l'analyse exploratoire

Tri par corrélations et contingences : 22 variables dont

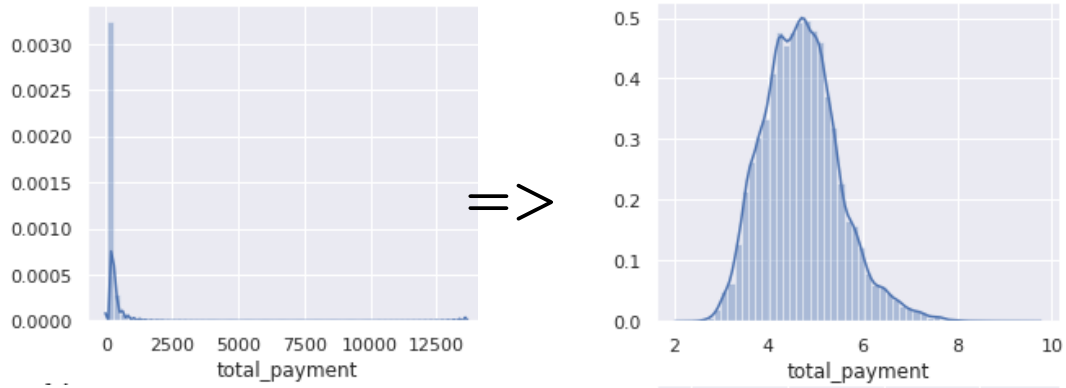
- 17 variables quantitatives
- 3 variables qualitatives

Fait marquant : **1/4** des ordinateurs sur les 549 vendus proviennent de l'état de Bahia.

Table of Contents

Transformations variables quantitatives

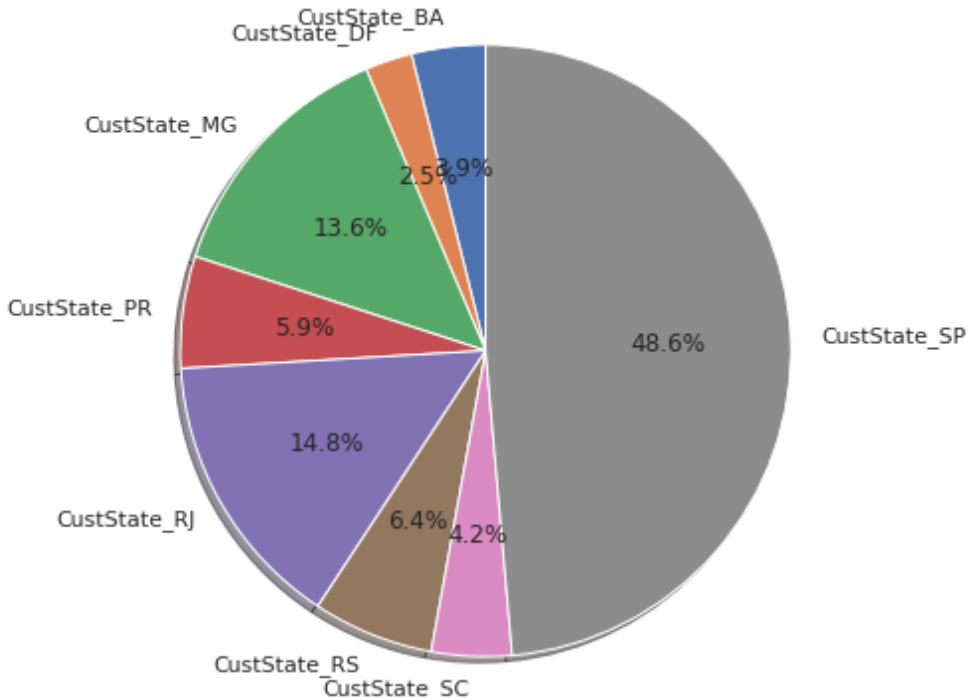
- $\ln(total\_payment)$



- $\sqrt{distance\_customers\_sellers}$
- $\sqrt{product\_photos\_qty}$
- $\sqrt{order\_delivered\_customer\_time\_day}$
- $\sqrt[4]{review\_answer\_time\_days}$
- $\sqrt[4]{shipping\_limit\_time\_day}$

Transformation variables qualitatives :

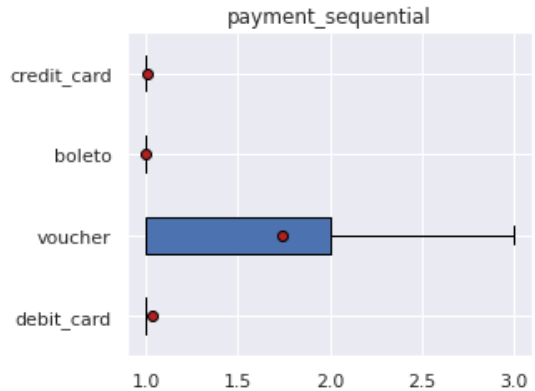
- One Hot Encoding de **états clients** avec filtre à 5% : selection de **8 états**



- LabelEncoding de **catégories produits** par la somme des achats par catégories $\sum$(**Total_payment**)

=> Permet de classer les catégories / volume de ventes

- Fusion en LabelEncoding de **type de paiement** par la moyenne de **paiement séquentiel**



=> Permet de supprimer l'analyse d'une variable quantitative

- L'encodage des variables qualitatives est **justifiée** par l'analyse graphique des structures =>

Table of Contents

Réduction des observations

$\sim$ 95000 observations : Temps de calcul **trop grand** pour les calculs de **réduction dimensionnelle** ou de **clustering** :

1. Segmenter en déciles chronologiques le set ($\sim$ 10000/décile)
2. **Construire** et **choisir** les meilleurs modèles de réduction dimensionnelle et clustering sur le **1er décile**.
3. Comparer avec ce modèles l'évolution temporelle sur les autres déciles (stabilité temporelle =>)

=> Suppression variables de type date [order_purchase _timestamp_month', 'order_purchase_timestamp_year] (différente période temporelle / décile).

Réduction dimensionnelle ?

- Visualiser les données
- Réduire les coûts
- Améliorer la qualité des modèles d'apprentissage

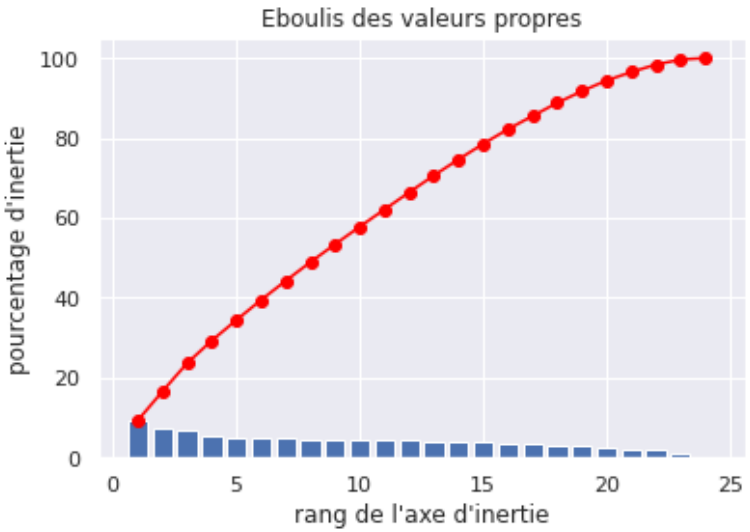
=> Analyse en composante principale (ACP) : maximise la variance

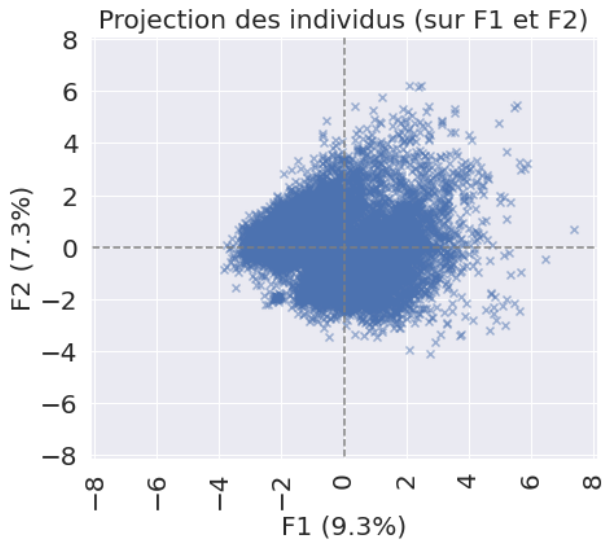
=> Isomap (non linéaire) : conserve la structure globale

=> t-Stochastic Neighbour Embedding (t-SNE) : favorise la structure locale

Table of Contents

Analyse en composante principale (ACP) : Eboulis des valeurs propres assez "plat"





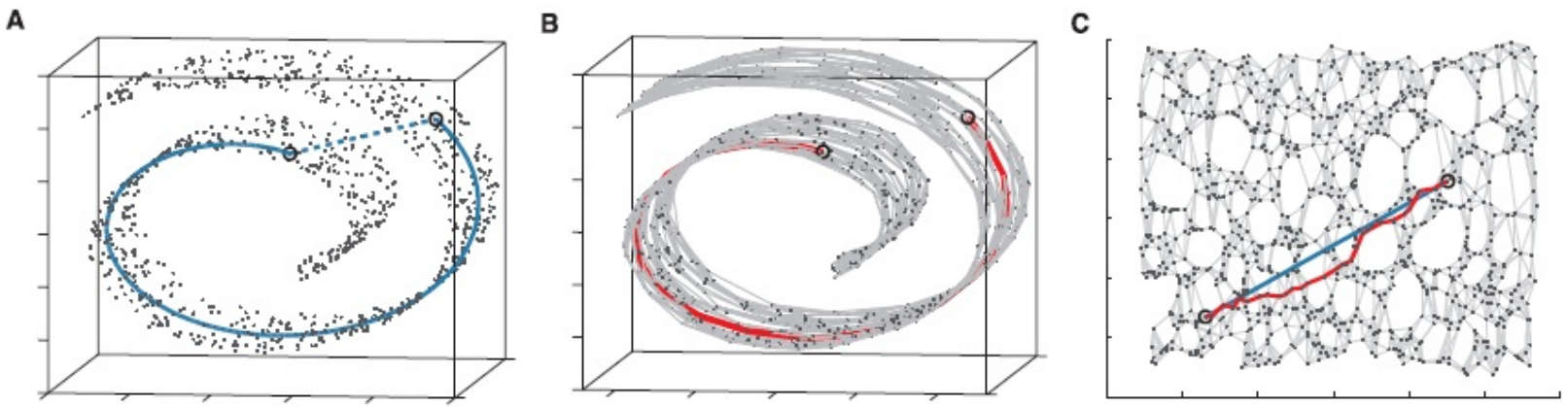
=> mauvaise visualisation des données du 1er plan factoriel (% d'inertie trop faible)

Table of Contents

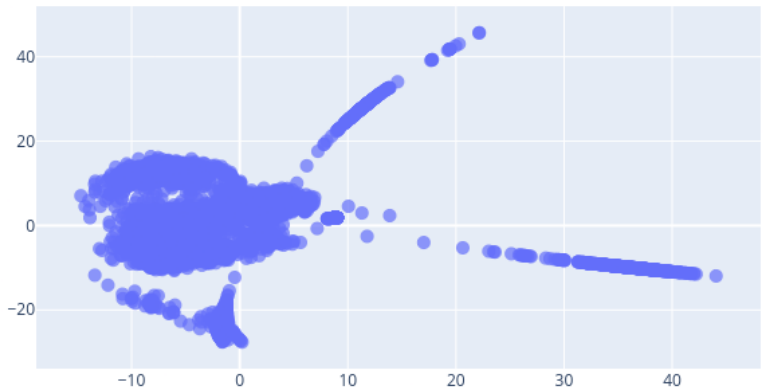
Multidimensional Scaling (MDS) : Minimiser l'erreur entre la distance d_{ij} des points x_i, x_j dans l'espace initiale avec la distance $||y_i - y_j||^2$ de la projection de x_i, x_j dans l'espace de dimension réduite

$$S(y_1, ..., y_N) = \sum_{i \neq j} \left(d_{ij} - ||y_i - y_j||\right)^2$$

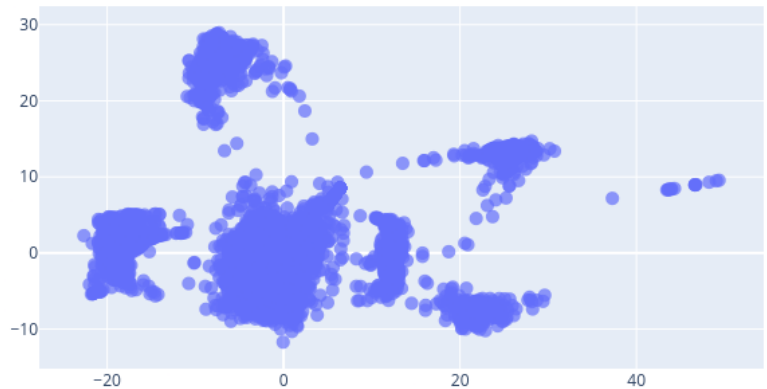
- **Isomap** : Appliquer l'algorithme du MDS à cette matrice de distances géodésique.



Isomap : nb voisin = 5



Isomap : nb voisin = 10



L'Isomap permet d'identifier quelques structures

Table of Contents

- **t-Stochastic Neighbour Embedding** (t-SNE) :
  - Calculer la similarité P_{ij} : à chaque point est associé une probabilité conditionnelle **gaussienne** en fonction de sa distance aux autres points dans l'espace initial :

$$p_{j/i} = \frac{\exp^{-||x_i - x_j||^2/2\sigma_i^2}}{\sum_{k \neq i} \exp^{-||x_i - x_k||^2/2\sigma_i^2}}$$

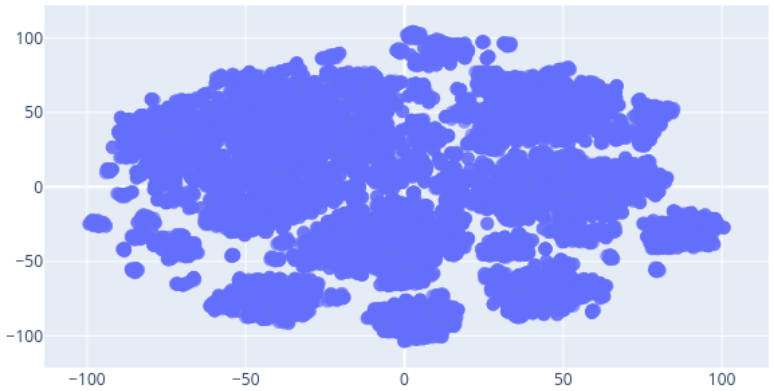
  - Calculer la similarité Q_{ij}, probabilité conditionnelle **t-Student**, dans l'espace de dimension inférieure

$$q_{j/i} = \frac{\exp^{-||y_i - y_j||^2}}{\sum_{k \neq i} \exp^{-||y_i - y_k||^2}}$$

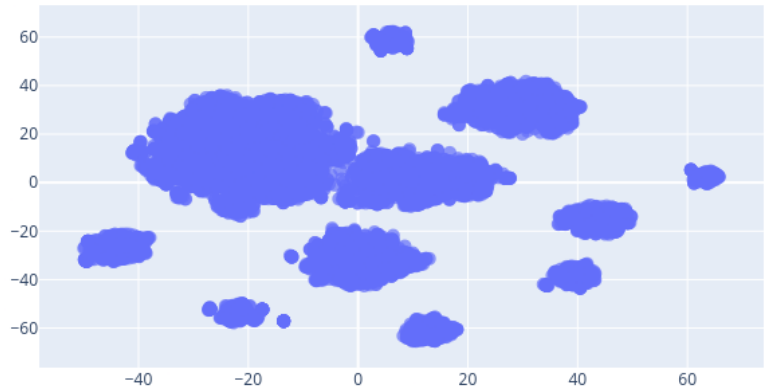
    telle que :
  - La KL divergence des similarités soit **minimisée** (par descente de Gradient)

$$C = \sum_i KL(P_i||Q_i||) = \sum_i \sum_j p_{j/i} \ln \frac{p_{j/i}}{q_{j/i}}$$

TSNE : perplexité = 10



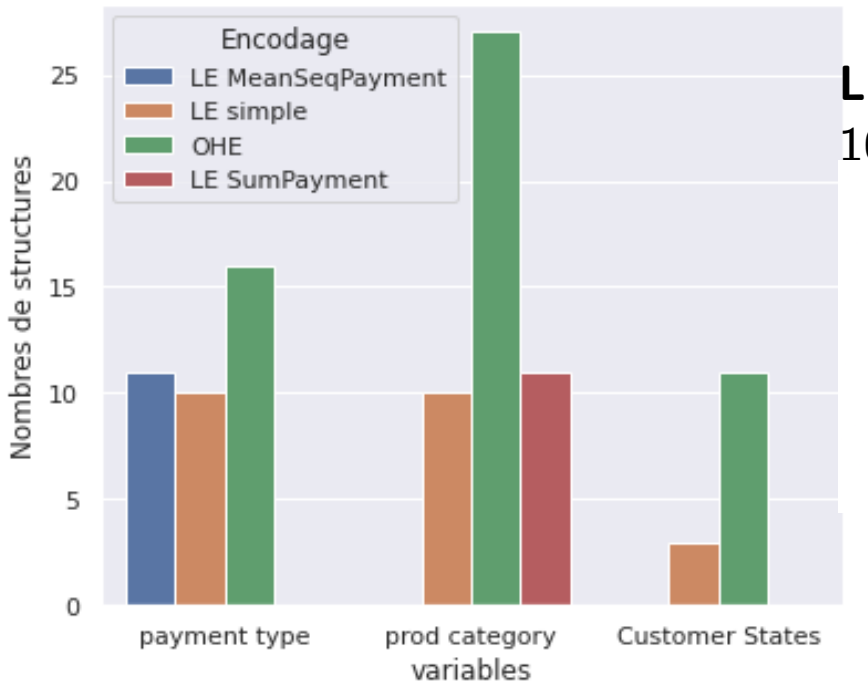
TSNE : perplexité = 100



Perplexité : Variance de P_{ij}. Estimation de la densité autour de chaque points.

=> TSNE avec une perplexité = 100 est le modèle de réduction dimensionnelle retenu.

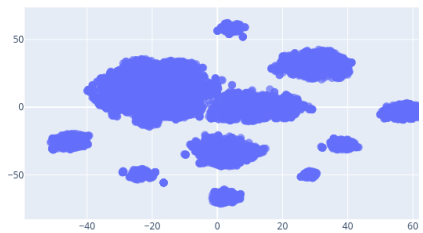
Evaluation de l'**encodage** des variables qualitatives

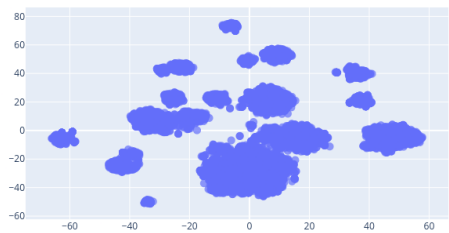


Encodage sur type_payment

LE simple :
10 structures

OHE :
15 structures





- l'OHE **renforce** l'influence de la variables pour la formation de clusters
- A l'inverse, le Label Encoding la **diminue**

- Objectif de nb de cluster <15 pour l'analyse

=> Choix de conserver l'encodage qui conserve 11 cluster <=

On s'attend à ce que la variable de **état des clients** soit la variable principale sur laquelle la formation de clusters s'appuie.

Table of Contents

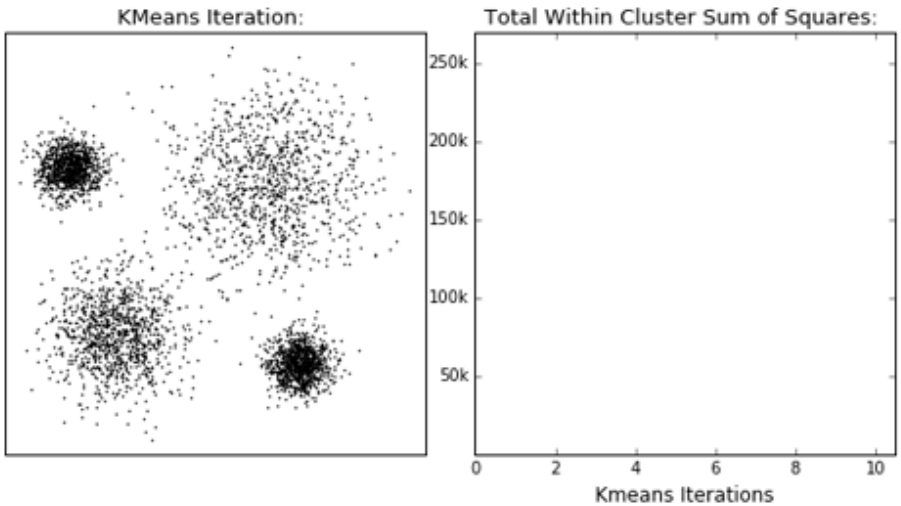
Présentation des modèles de clustering

- K-means
- DB-Scan
- Hierarchique

Table of Contents

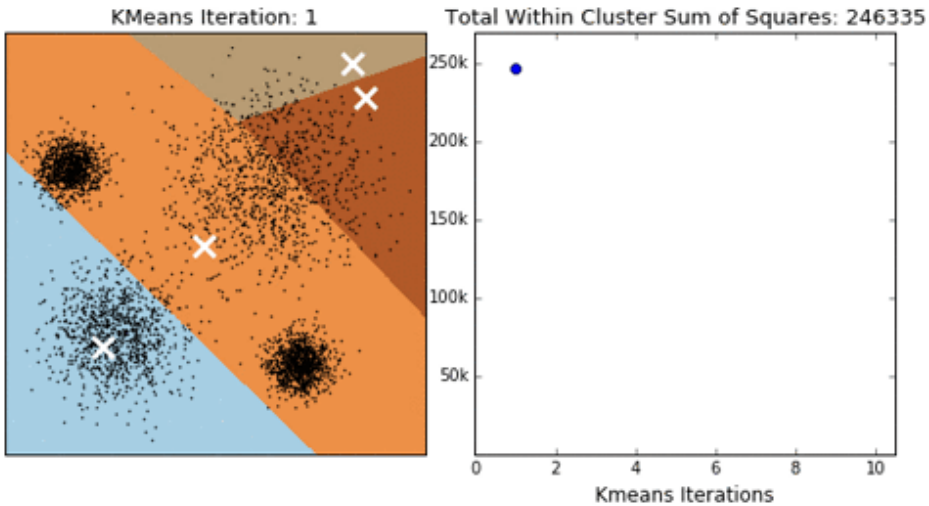
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



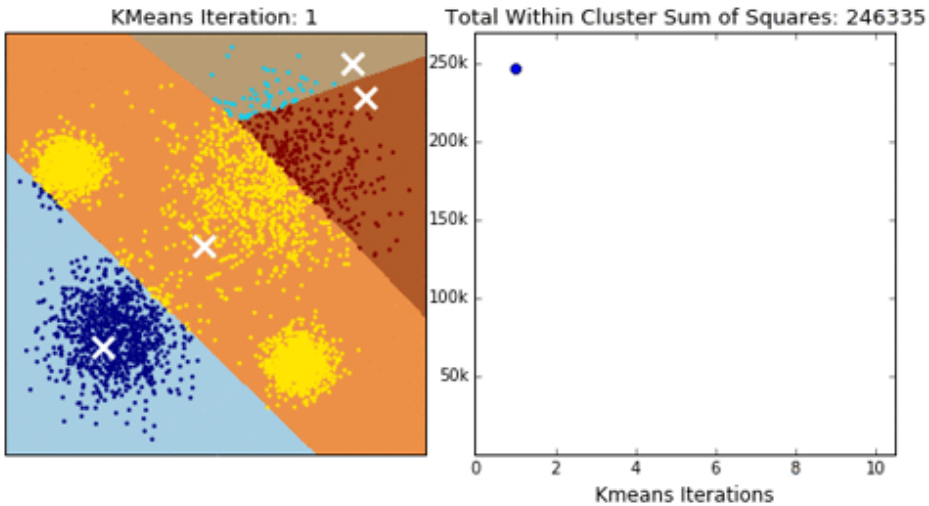
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



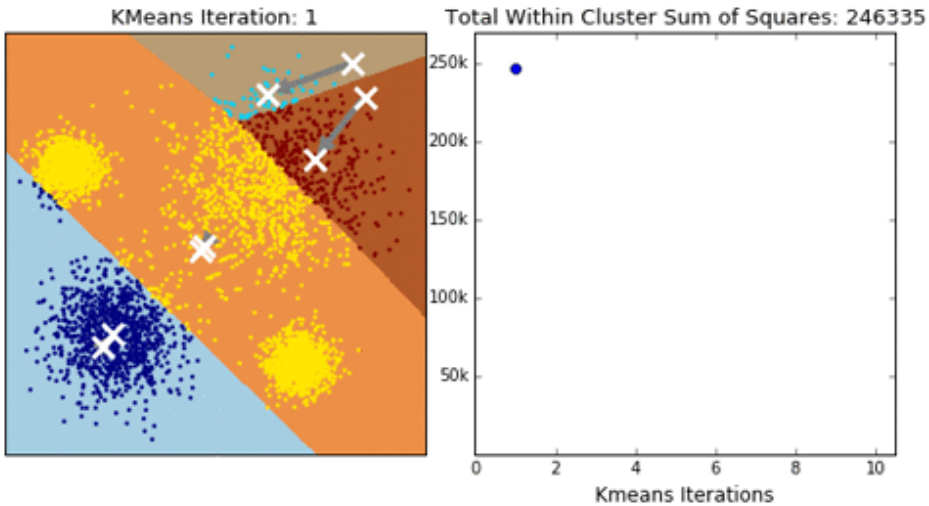
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



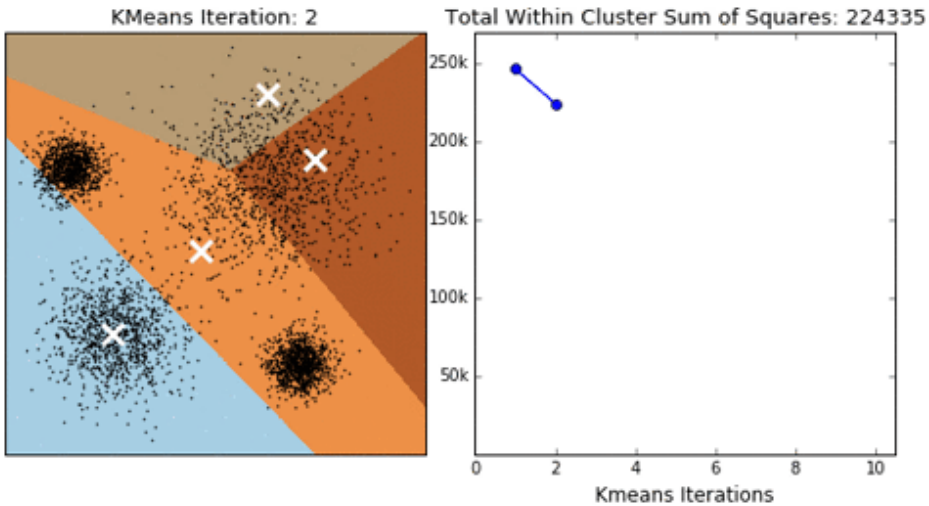
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



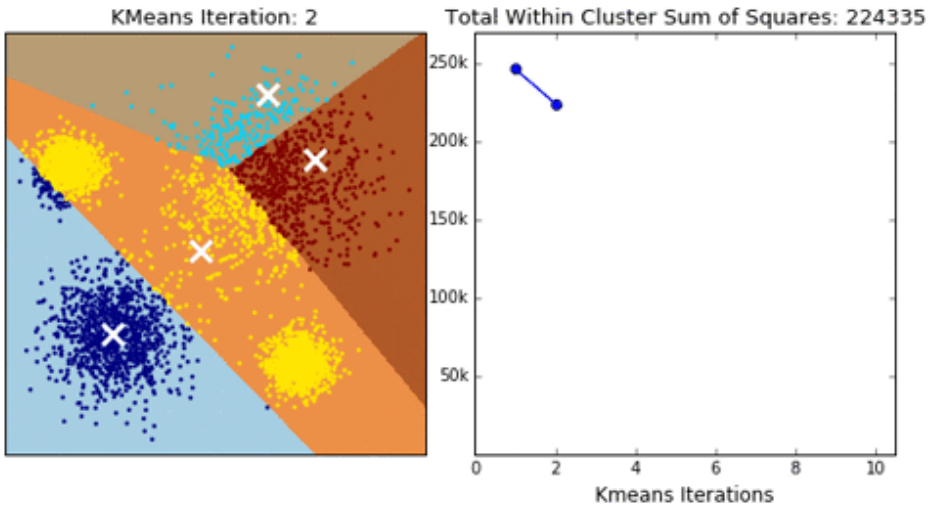
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



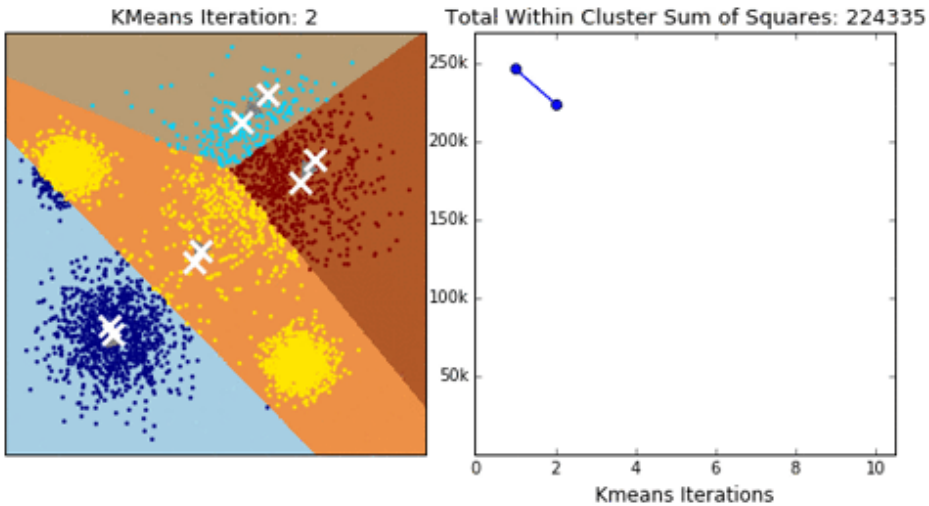
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



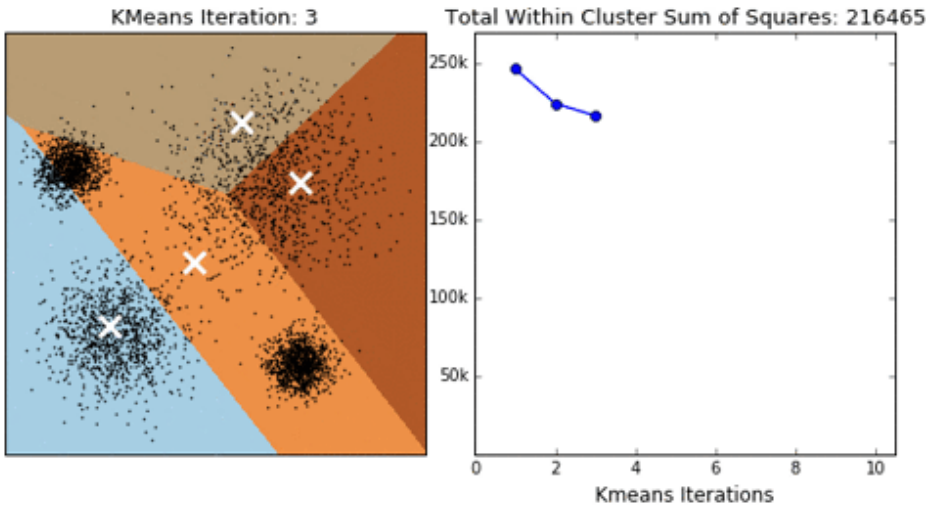
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



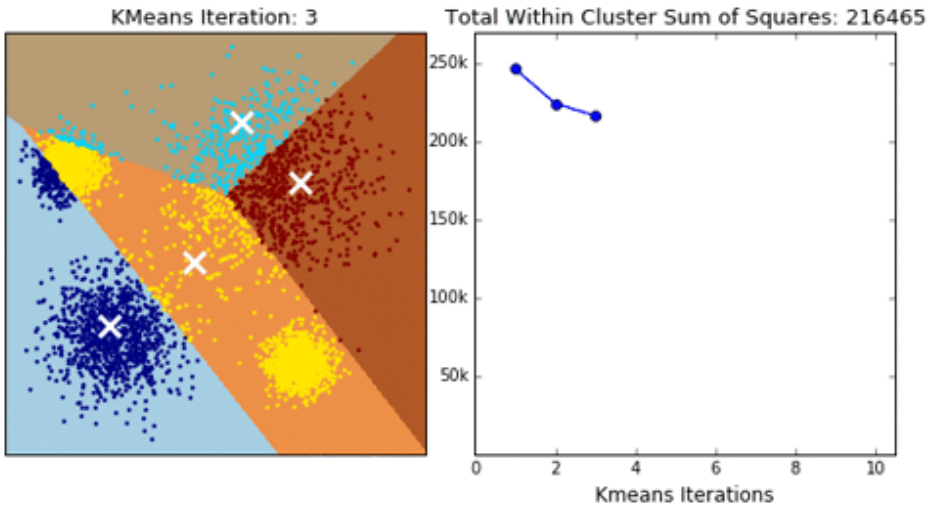
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



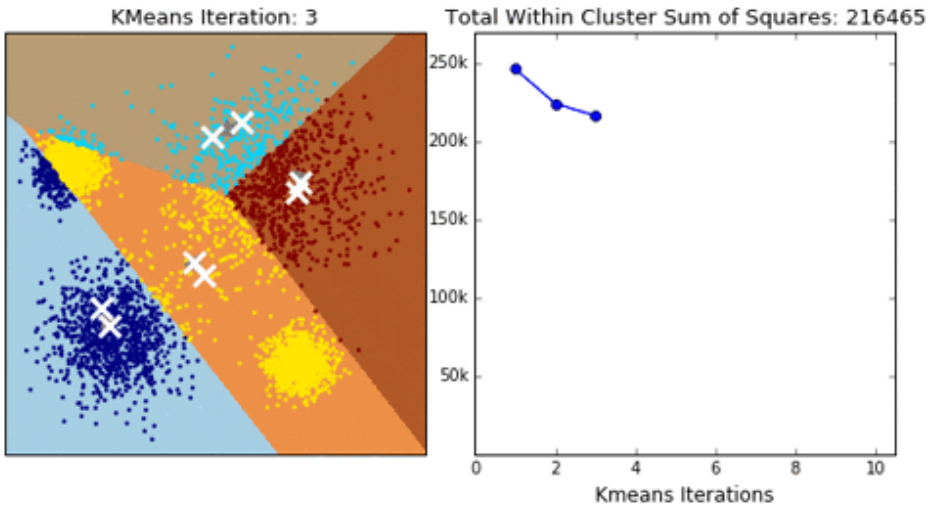
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



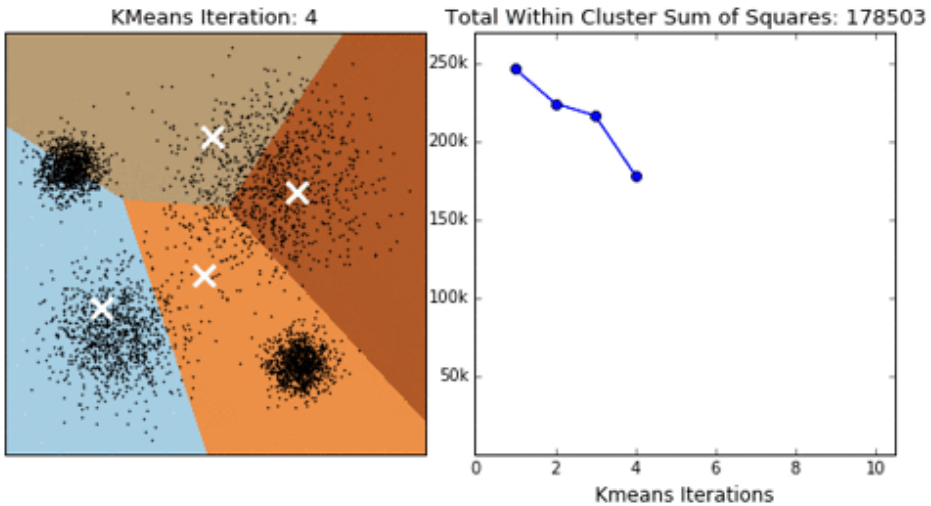
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



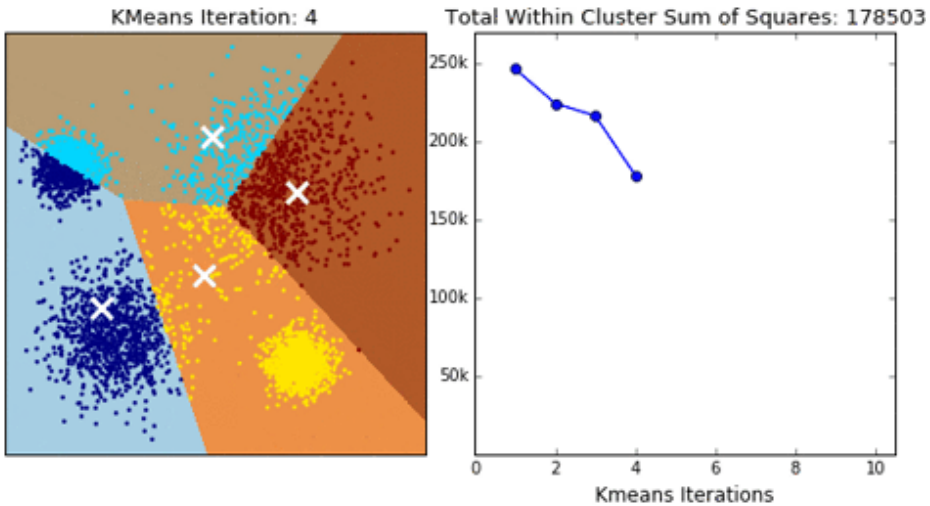
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



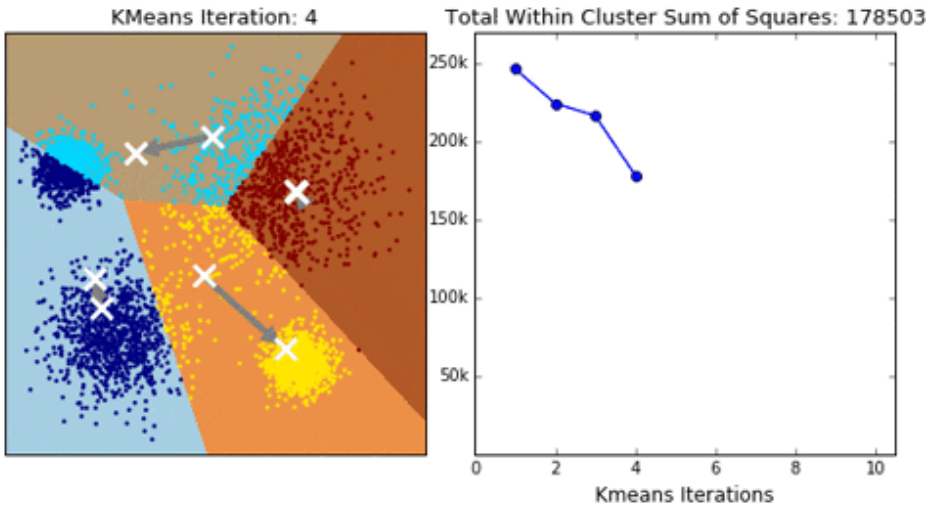
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



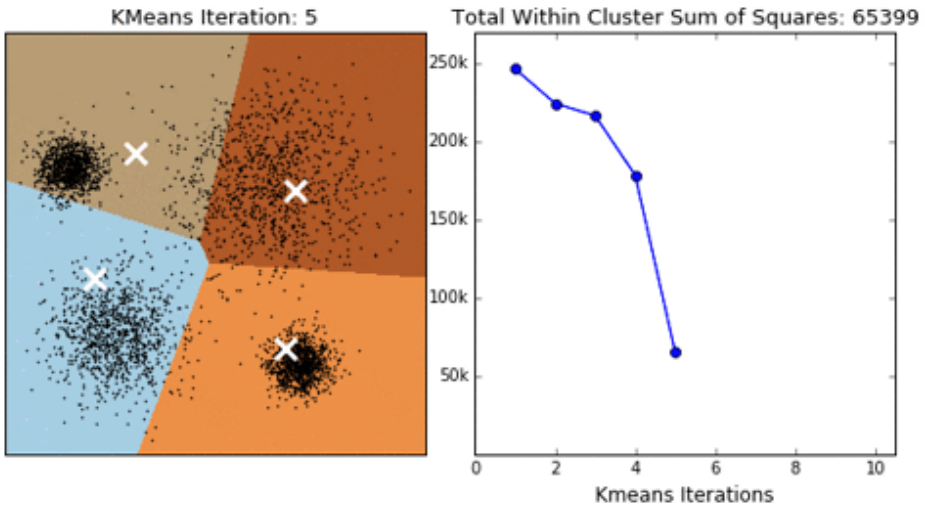
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



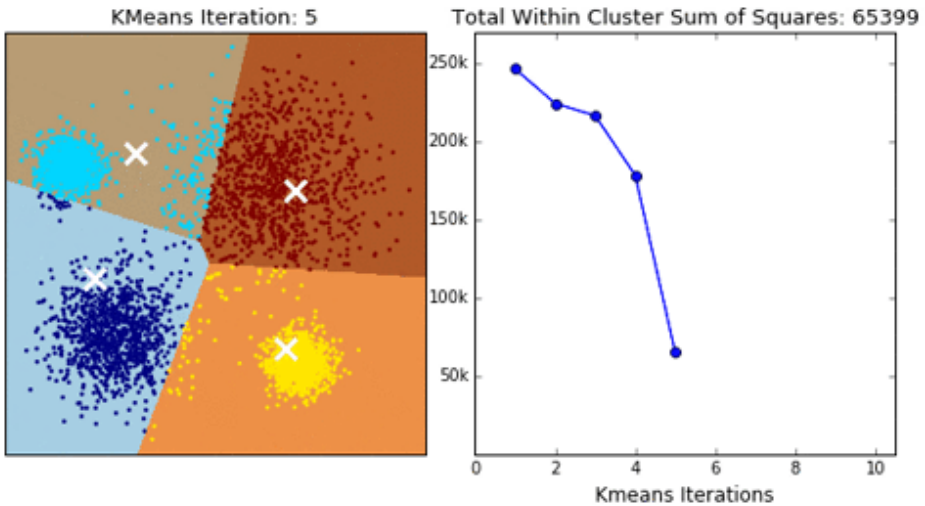
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



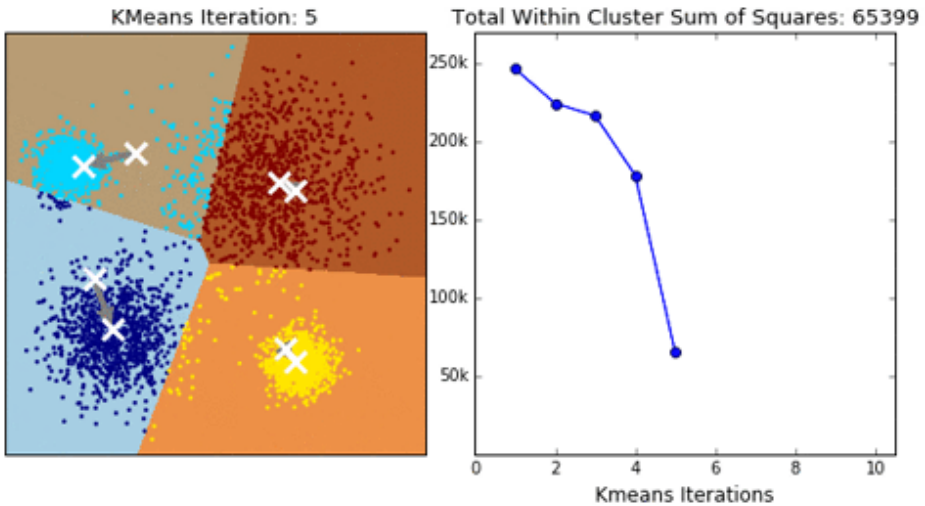
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



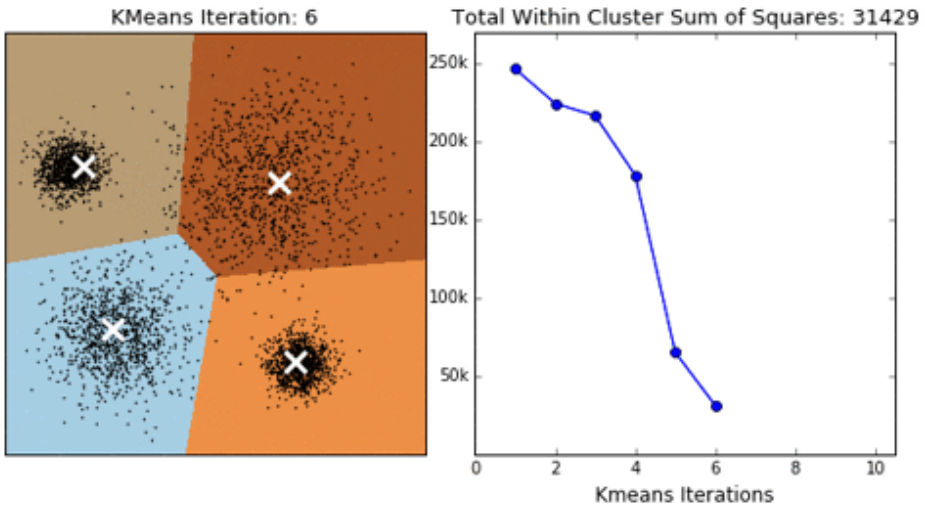
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



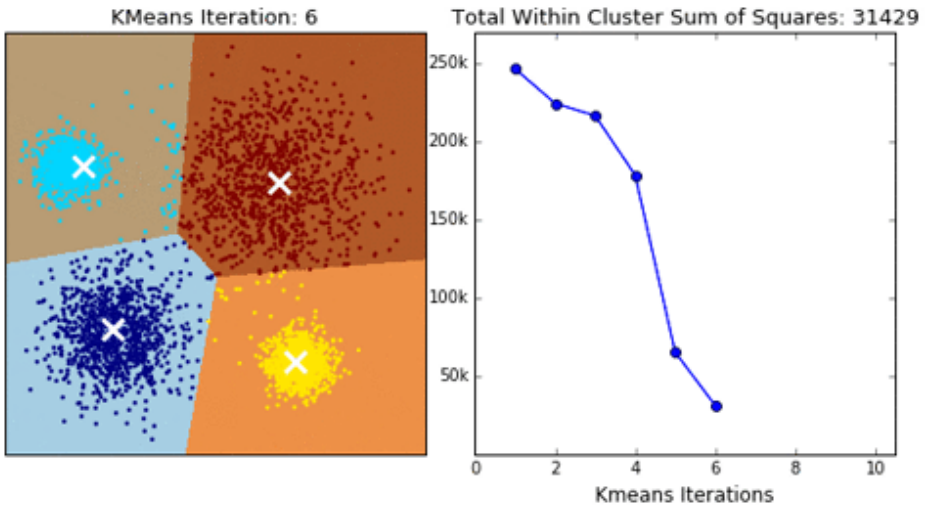
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



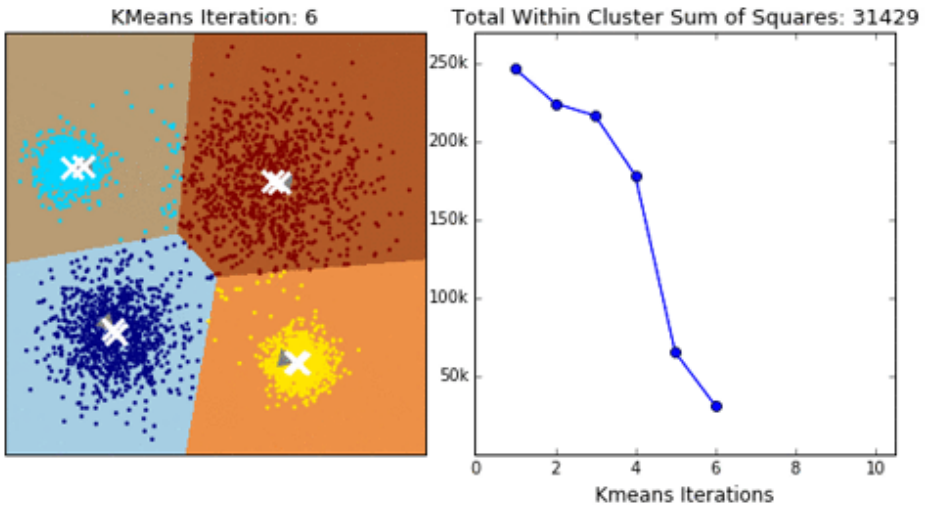
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



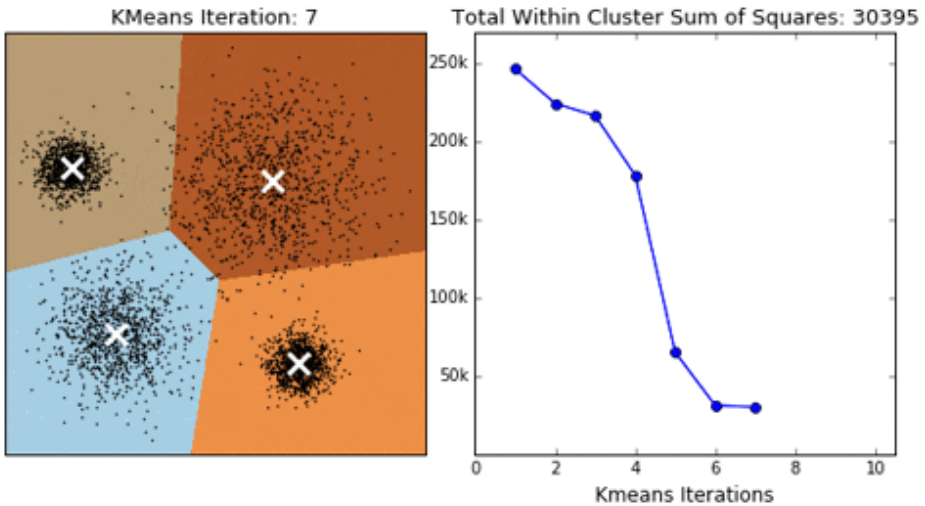
K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence



K-means

- Choix aléatoire de K centroïdes
- Association des points les plus proche à chaque centroïde
- Calcul les K centroïdes de chaque cluster
- Boucler jusqu'à convergence

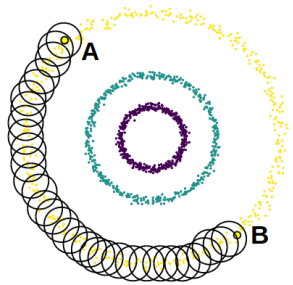


- Initialisation **kmean ++** : centroïdes initiaux sont choisis "éparpillés" le plus possible
- **random_state** paramétrée une initialisation **déterministe**.
- Avantages : Recherche **efficace** d'une partition de **variance intra-cluster minimale**.
- Inconvénients : limitation à des formes **cluster convexe** (corrigées avec Kernel KMeans). Trouve minimum **locaux** plutôt que globaux (corrigées en partie avec kmeans++ et les répétitions)

Table of Contents

DB-Scan

- Construction par densité : l'epsilon-voisinage ϵ et le nombre minimal de voisins n_{min} à l'intérieur.

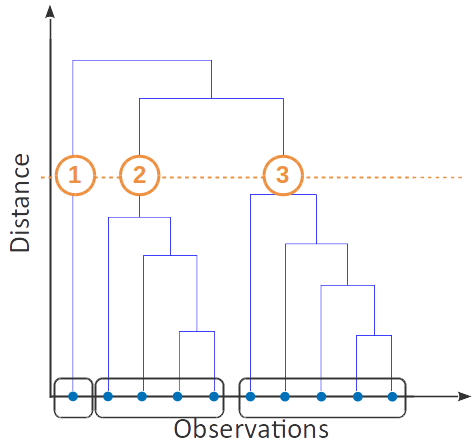


- Avantages : **Faible temps de calcul** sans définition à l'avance du nombre de clusters. Les clusters trouvés sont de **forme arbitraire**.
- Inconvénient : **fléau de la dimensionalité** : Les boules de rayon ϵ et de grande dimension ont tendance à ne contenir aucun autre point. Le choix de ϵ et n_{min} peut être délicat.

Table of Contents

Hierarchique : clustering de Ward

- Clustering séparatif ou aggrégatif
- Clustering de **Ward** : Clustering **aggrégatif** qui minimise l'augmentation de variance inter-cluster.
- Avantages : **pas de définition** à l'avance **du nombre de clusters**. Visualisation possible par dendogramme.



- Inconvénient : **complexité algorithmique lourde** : adapté aux *faibles nombre d'individus*.

Evaluation des modèles

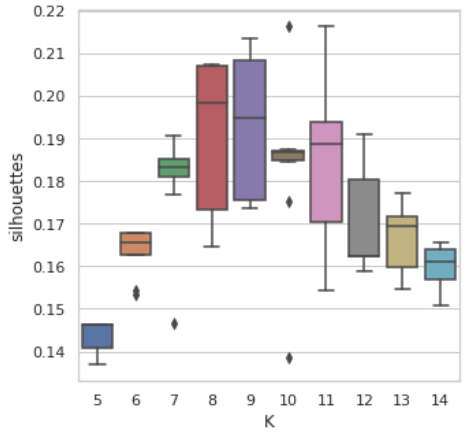
- Stabilité des clusters : étude de la variation des résultats par rapport à une initialisation différente.
- Forme des clusters : par le coefficient de silhouette :

$$s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}$$

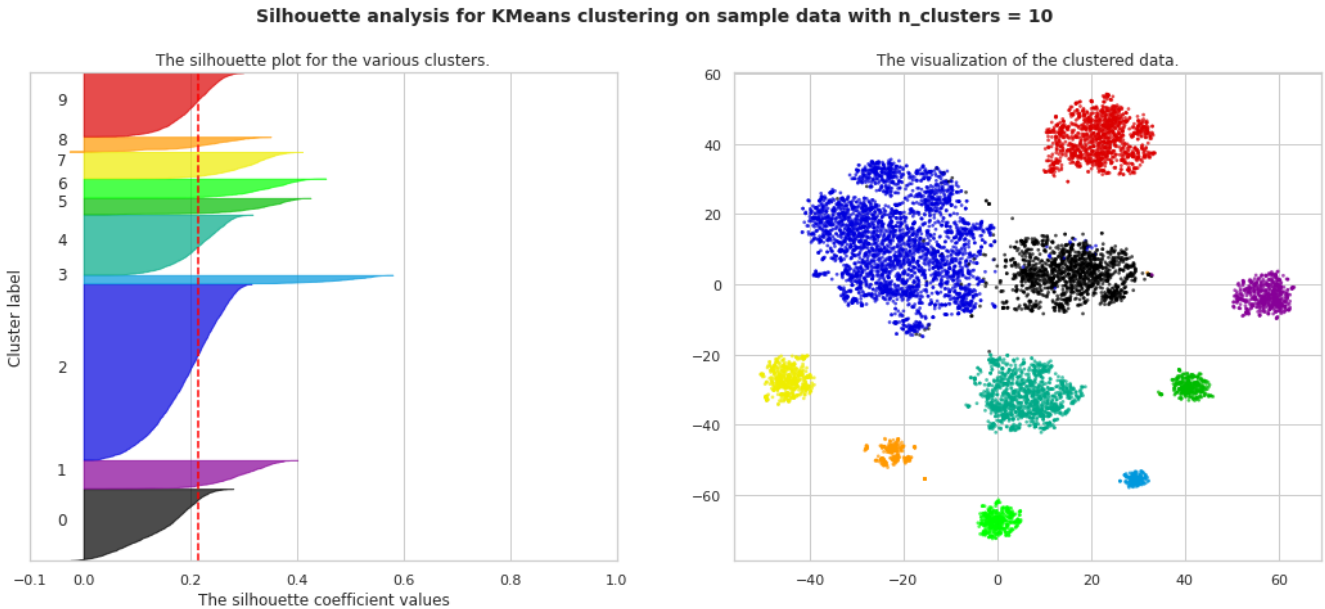
  - $a(x)$ distance moyenne de x aux autres points du cluster
  - $b(x)$ plus petite valeur de $a(x)$ si x était assigné à un autre cluster

K-Means

Stabilité : Initialisation avec différents **random_state**



Silhouette moyenne : 0.216

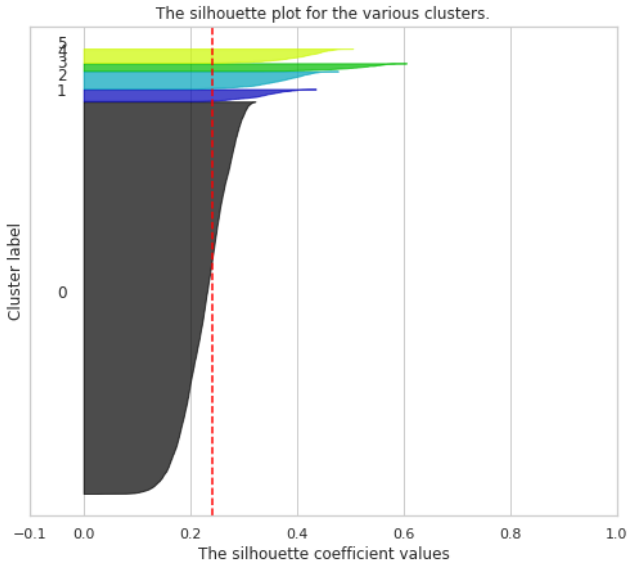


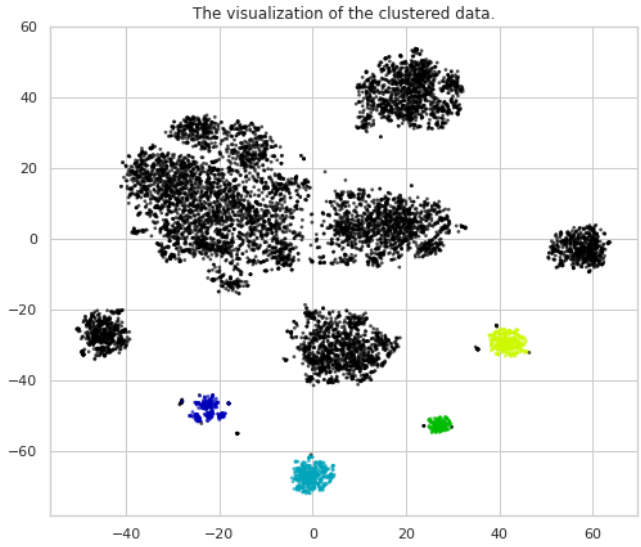
DBSCAN

$$\epsilon \in [0.5, 2, 5] \ , \ n_{min} \in [2, 10, 50]$$

Meilleurs paramètres : 6 clusters ($[\epsilon = 5, n_{min} = 50]$)

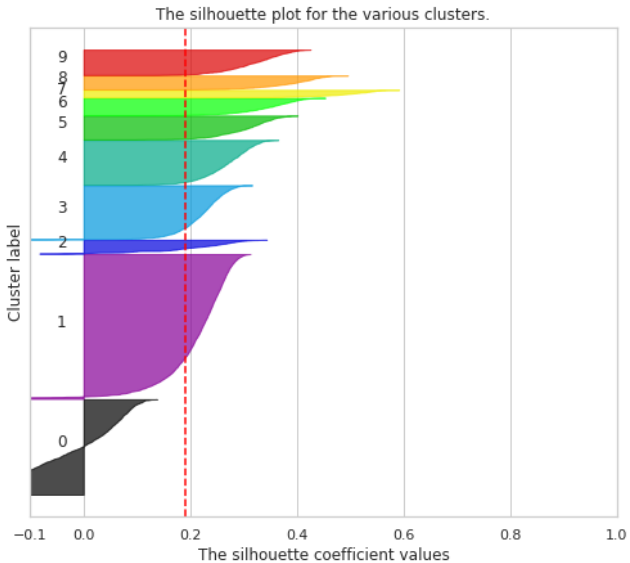
Silhouette moyenne : 0.241

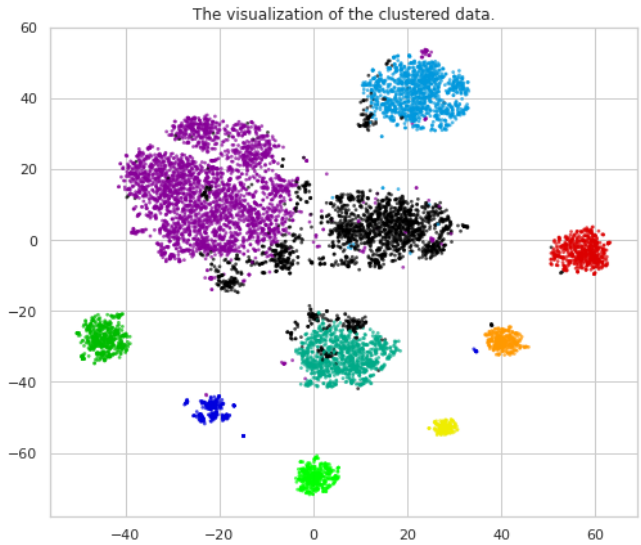




Hierarchique Ward

Silhouette moyenne : 0.192





Selection du meilleur modèle

- K-Means : coef silhouette = **0.22** et groupes homogènes
- DB-Scan : coef silhouette = **0.24** et groupes très inhomogènes
- Ward : coef silhouette = **0.20** et groupes homogènes

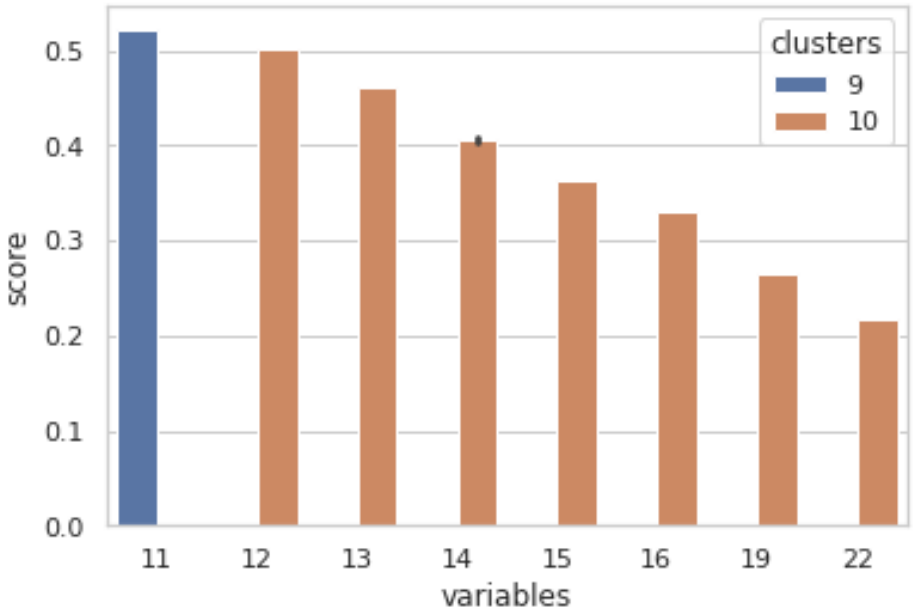
=> Selection de **K-Means**

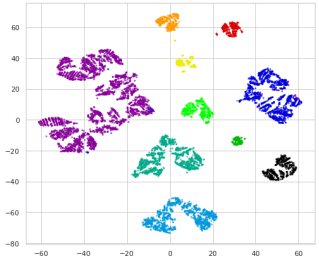
Réduction des variables avec le K-means

- Suppressions successives des variables les moins interpretables
- Etude d'impact sur le nombre et la forme des clusters.

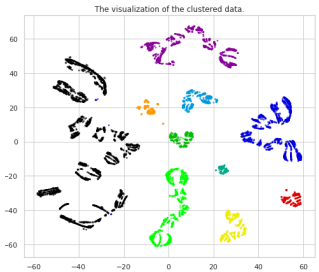
variables restantes	variables supprimées		
22			
19	product_name_lenght	product_photos_qty	prod_descript_lenght
16	inertie	product_height_cm	ship_lim_time_days
15	payment_installments		
14	total_payment OU review_ans_time_days		
13	total_payment	review_ans_time_days	
12	order_delivered_time_day		
11	distance_customers_sellers		

Score silhouette





13 variables



12 variables

11 variables

=> Choix de garder **13** variables : **10** clusters de **formes compactes**.

Table of Contents

Etude de la stabilité des clusters dans le temps <=

décile	années	mois
1	[2016, 2017]	[10, 12, 1, 2, 3, 4, 5]
2	[2017]	[5, 6, 7, 8]
3	[2017]	[8, 9, 10]
4	[2017]	[10, 11, 12]
5	[2017, 2018]	[12, 1, 2]
6	[2018]	[2, 3]
7	[2018]	[3, 4, 5]
8	[2018]	[5, 6]
9	[2018]	[6, 7, 8]
10	[2018]	[7, 8]

L'indice de Rand : proportion (x_1, x_2) groupés de la même façon dans les deux partitions (= 0 pour clustering 2 aléatoire, = 1 pour clustering 2 = clustering 1)

Clustering ($_i$ = décile n°i, / = comparaison Rand) :

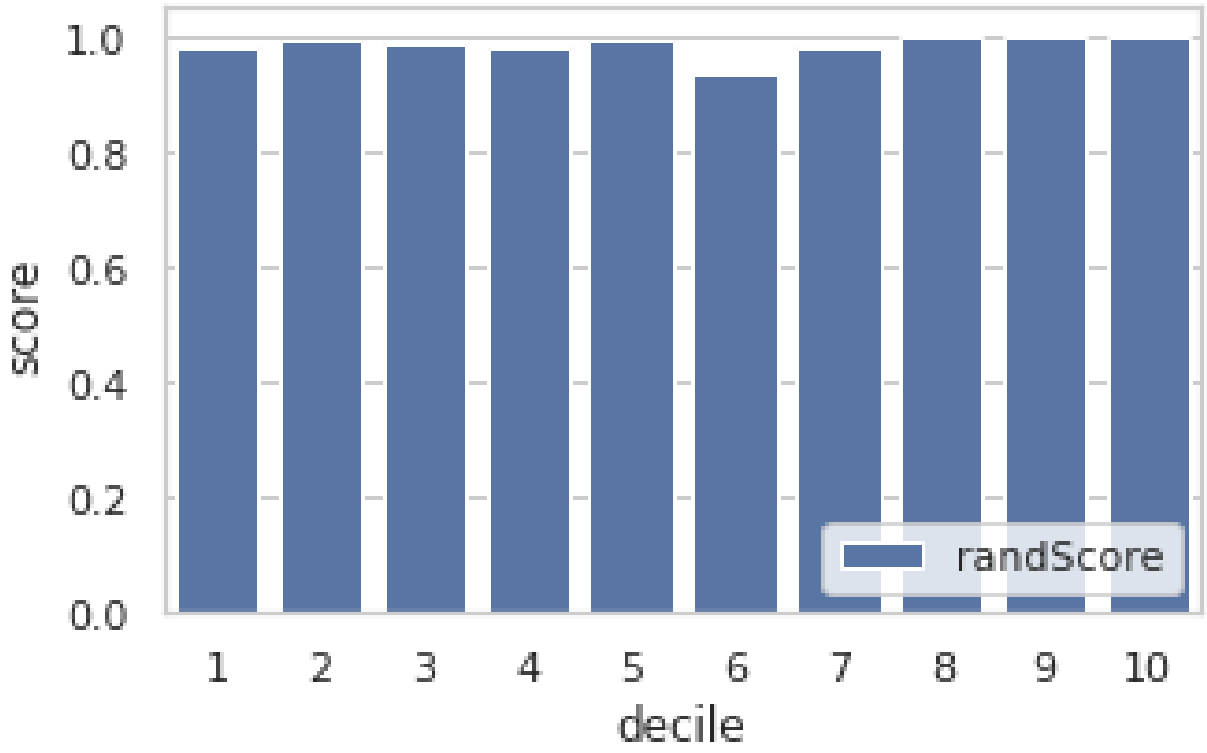
1 : $\text{fit}_1.\text{predict}_1$

2 : $\text{fit}_1.\text{predict}_2/\text{fit}_2.\text{predict}_2$

...

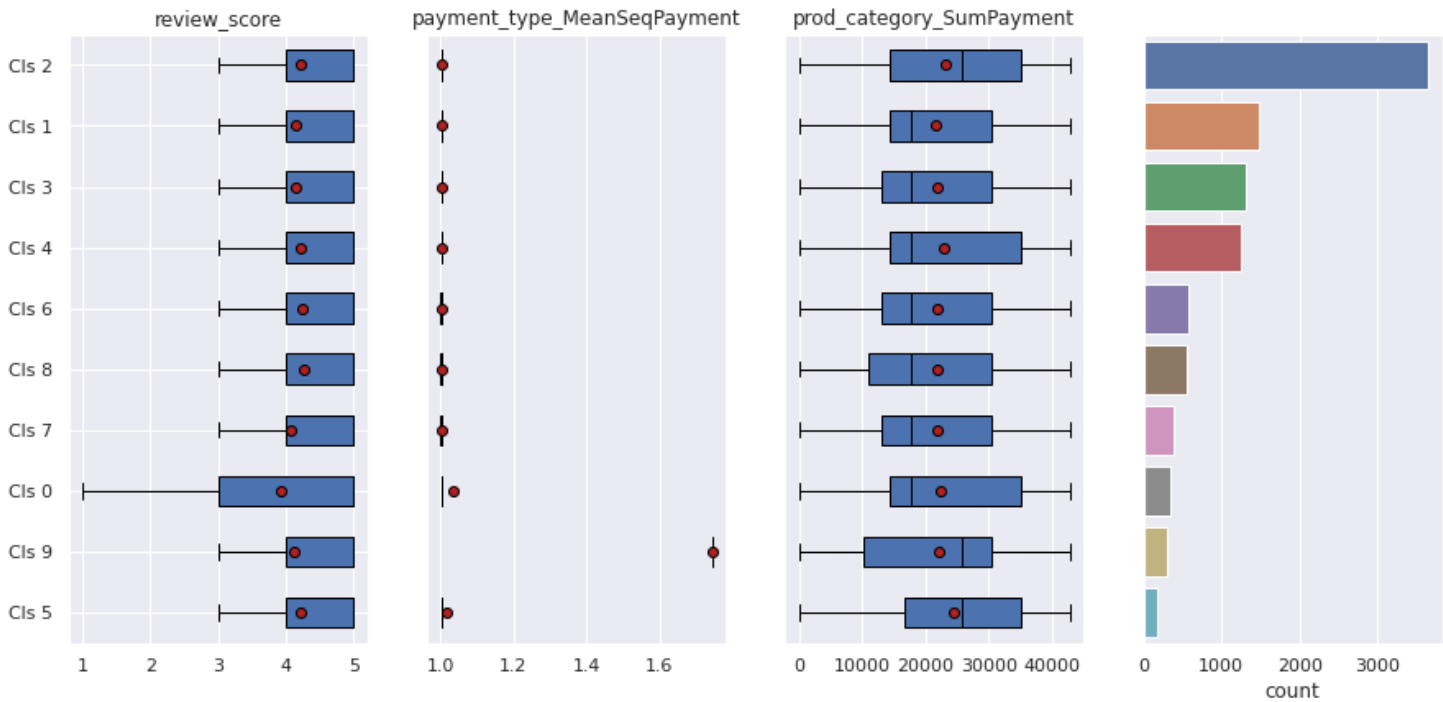
10 : $\text{fit}_1.\text{predict}_{10}/\text{fit}_{10}.\text{predict}_{10}$

=> Evaluation du Rand Scorce sur tous les déciles



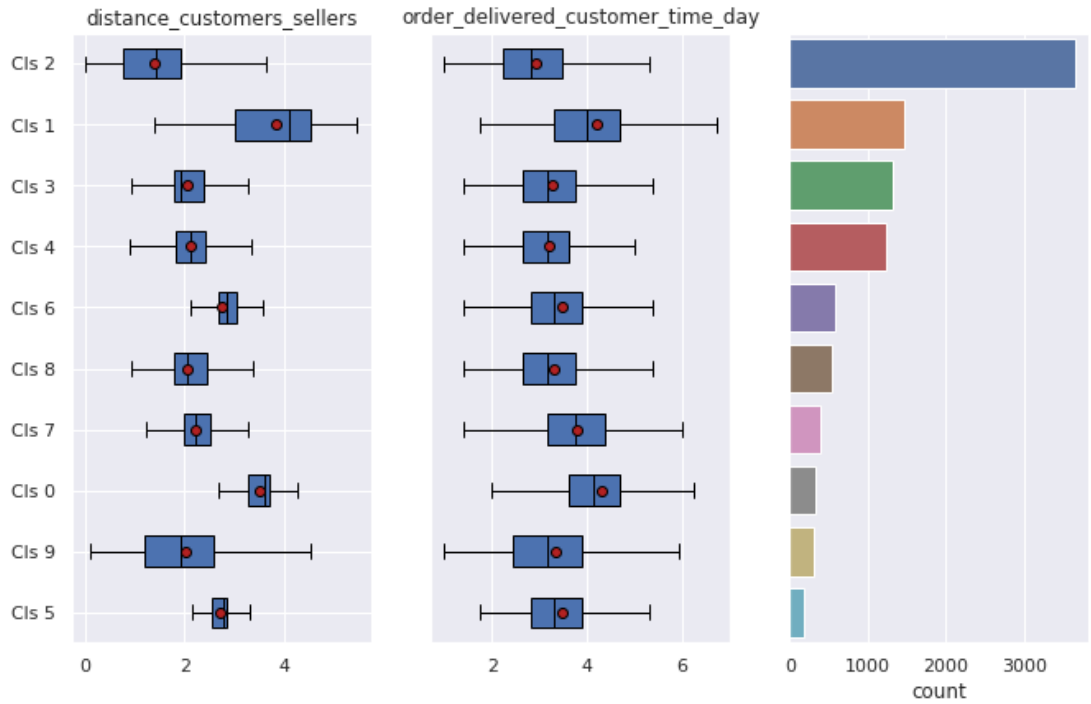
=> Clustering plutôt stable suivant les périodes étudiées

Analyse des clusters



=> Cls 9 se distingue par le type de paiement **séquencé**.

=> Cls 0 se distingue par un **mécontentement** des clients

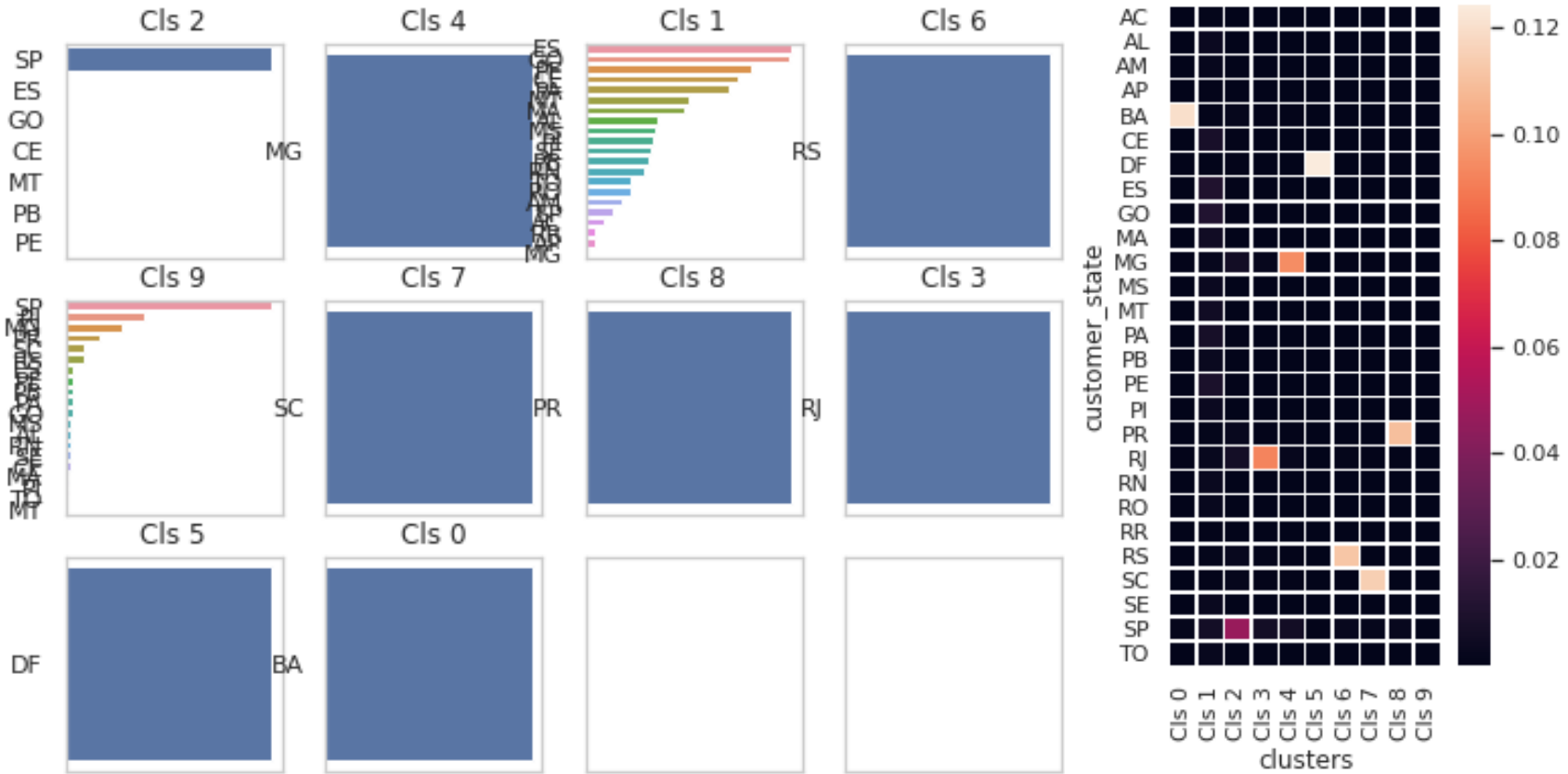


=> Cls 9 : type de paiement séquencé.

=> Cls 0 : mécontentement des clients, distance client-vendeur **grande**, temps de livraison **long**

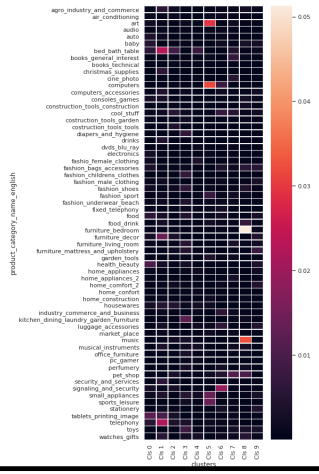
=> Cls 1 : distance client-vendeur **grande**, temps de livraison **long**

=> Cls 2 : distance client-vendeur **courte**, temps de libraison **court**



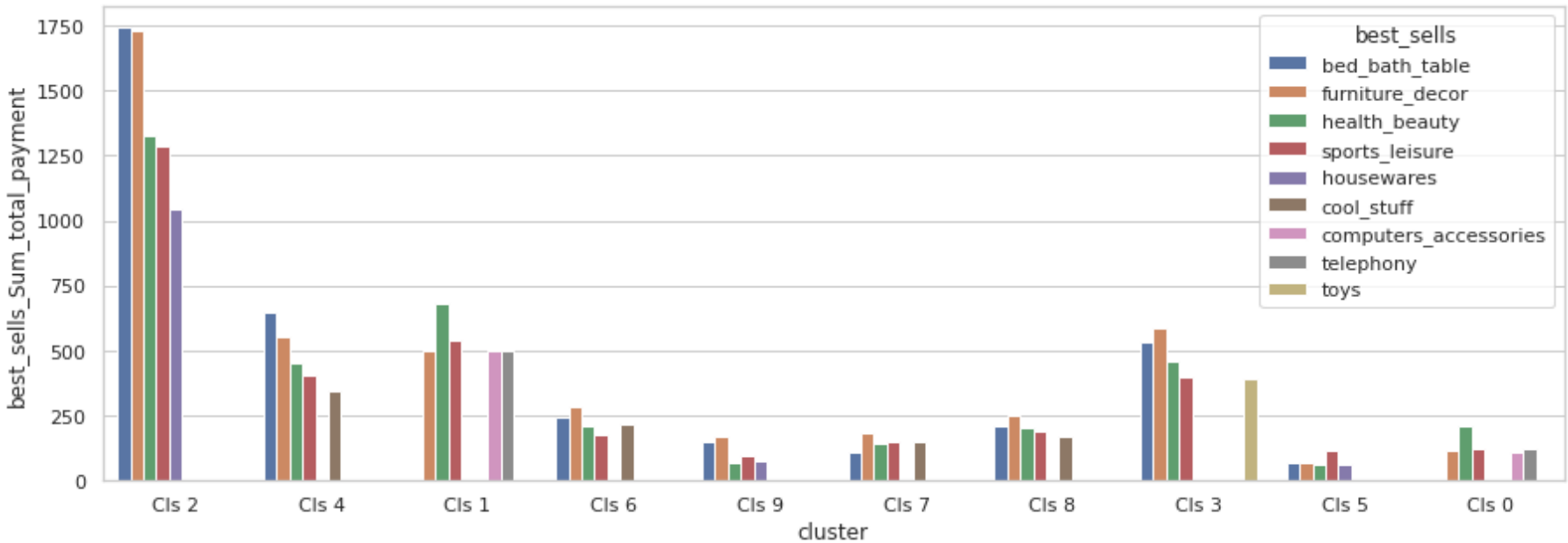
=> Dépendance forte de 8 clusters / 10 des Etats d'où proviennent les clients.

cluster	catégories spécifiques de produits vendus
Cls 2	
Cls 4	
Cls 1	bed_bath_table,furniture_decor,telephony
Cls 6	signaling_and_security
Cls 9	
Cls 7	pet_shop
Cls 8	furniture_bedroom,music
Cls 3	kitchen_dining_laundry_garden_furniture
Cls 5	art,computers,small_appliances,sports_leisure
Cls 0	health_beauty,tablets_printing_image



=> Identification de catégories de produits vendus spécifiquement dans certains cluster ($\xi_{\text{état,produit}}/\xi_n > 0.01$).

Clusters	Etat client	Distance client-vendeur	temps de livraison	Spécificité	Produit spécifique
Cls 0	BA	grande	long	Mécontentement des clients	=>
Cls 1		grande	long		=>
Cls 2	SP	courte	court		
Cls 3	RJ				=>
CLs 4	MG				
CLs 5	DF				=>
CLs 6	RS				=>
CLs 7	SC				=>
CLs 8	PR				=>
CLs 9				paiement séquencé	



=> Meilleurs ventes (en volume) des catégories de produit par clusters.

Table of Contents

Conclusion

Réponses aux besoins clients :

- Comprendre les types d'utilisateurs

=> Identification des clusters suivant :

- pays des clients
- mécontentement des clients
- distance client-vendeur
- temps de livraison
- type de paiement séquencé
- produits vendus spécifique

- *"Description actionable"* = Proposition de communications ciblées

=> Identifier en détail le mécontentement des clients (distance client-vendeur grande / temps de livraison grande / état client : Bahia)

=> Pub ciblé de ventes spécifiques ou importantes par cluster

- *"Proposition de contrat de maintenance"* = identifier la fréquence (ex : tous les mois) pour refaire l'analyse client

=> Pas de nécessité de refaire l'analyse client sur la période [2016-2018] : fréquence tous les 2 ans

Apprentissage majeur du projet :

- **Importance capitale de l'encodage** des variables qualitatives pour le clustering :
  => l'OHE renforce l'influence de la variable pour la formation de clusters
  => A l'inverse, le Label Encoding la diminue
- **Importance capitale du nombre de variables** pour le clustering :
  => **Réduire le nombre de variables** sans réduire le nombre de cluster améliore significativement les résultats : coéfficient de silhouette a **doublé** passant de **0.21** à **0.46** pour une réduction de **22** à **13** variables.

Perspectives

Problématique client :

- Etudier plus en détail les cluster

Perspectives personnelles :

- Creuser plus “systématiquement” l’influence de la transformation et la réduction de variable sur les modèles d’apprentisage artificiel

Merci de votre attention