SEGMENTEZ LES CLIENTS D'UN SITE E-COMMERCE

Voahangy Joan ALEONARD – 25/02/2021

AGENDA DU JOUR

PRÉSENTATION DU PROJET

ANALYSE EXPLORATOIRE

ESSAIS DE MODÉLISATION

MODÈLE FINAL ET MAINTENANCE

PRÉSENTATION DU PROJET

APPEL À PROJET

olist

CONTEXTE

OLIST est un site brésilien proposant une solution de marché en ligne (marketplace) : elle met en relation les acheteurs et les vendeurs sur un espace / plateforme sécurisé.

BESOIN

OLIST mandate un Data Scientist pour fournir aux équipes Marketing e-commerce une **segmentation des clients**, qui pourra être utiliser quotidiennement pour des campagnes de communication.

MISSION

La **segmentation** doit répondre à 2 nécessités :

- Fournir une description **actionnable**, c'est-à-dire exploitable et facile d'utilisation pour l'équipe Marketing;
- Garantir une **stabilité dans le temps** sur la base d'un contrat de maintenance clair.

RÉFLEXION SUR LA MÉTHODOLOGIE

olist

PROBLEMATIQUE

Comme les **segments de clients** ne sont pas définies 'à priori', nous sommes donc face à une problématique de **clustering** : c'est-à-dire, que nous allons partitionner notre jeu de données en **sous-groupes** de clients **similaires.**

DEMARCHE

Compréhension globale des **relations** entre les différents datasets.

Sélection et/ou création de **variables pertinentes** permettant de caractériser au mieux le profil des clients (similarité ou différence) afin de fournir une description actionnable.

Utilisation d'un **algorithme de clustering (K-Means)** sur différentes combinaisons de variables.

Interprétation des clusters, c'est-à-dire, les **caractériser** pour obtenir des segments exploitables.

Recommandations de **maintenance temporelle** des segments.

ANALYSE EXPLORATOIRE DES DONNÉES

CONNEXIONS ENTRE LES DIFFÉRENTS DATASETS

NETTOYAGE DES DONNÉES

Principales étapes de nettoyage

- Agrégation des 73 produits en **11 nouvelles catégories principales** + une catégorie « unknown »
- Filtre sur le statut de commandes : uniquement celles qui ont été livrées (**delivered**)
- Filtre temporel dû à des incohérences sur 2016 et fin 2018: **janvier 2017 à aout 2018**
- Création d'une variable « **Régions** » pour simplifier la localisation client
- Assemblage des données en une table unique avec comme index l'identifiant unique du client
- Suppression des variables redondantes (matrice de corrélation)

MASTER DATAFRAME dimension : (93104, 12)

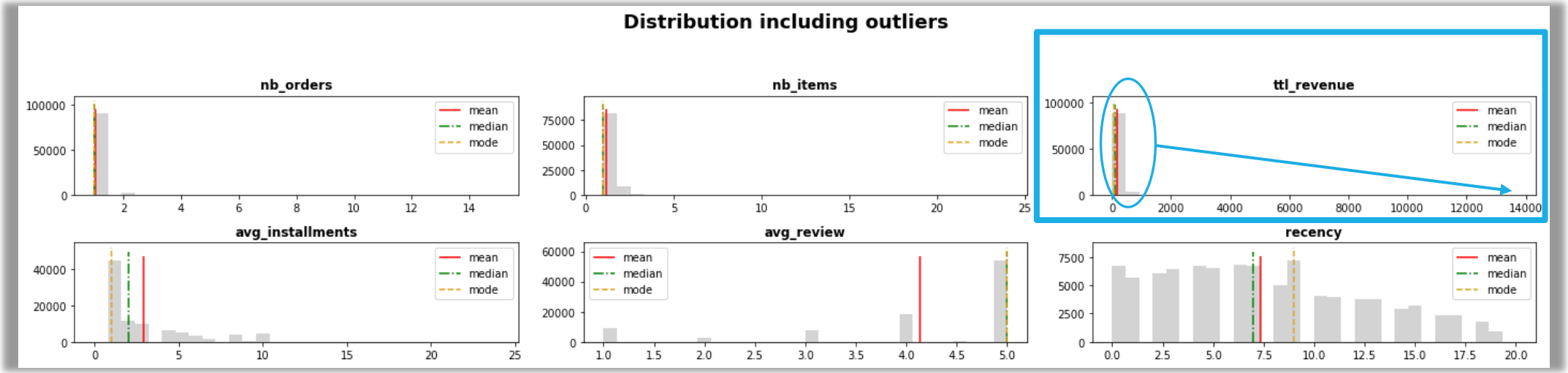
Missing values: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

main_region	main_category	min_purchase_date	max_purchase_date	nb_orders	nb_items	ttl_revenue	avg_installments	avg_review	purchase_time	recency
Southeast	furnitures	2018-05-10 10:56:27	2018-05-10 10:56:27	1	1	141.90	8.00	5.00	6-11h	3.00
Southeast	health_beauty	2018-05-07 11:11:27	2018-05-07 11:11:27	1	1	27.19	1.00	4.00	6-11h	3.00
South	supplies	2017-03-10 21:05:03	2017-03-10 21:05:03	1	1	86.22	8.00	3.00	18-21h	17.00
North	electronics	2017-10-12 20:29:41	2017-10-12 20:29:41	1	1	43.62	4.00	4.00	18-21h	10.00
Southeast	electronics	2017-11-14 19:45:42	2017-11-14 19:45:42	1	1	196.89	6.00	5.00	18-21h	9.00

SPÉCIFICITÉ DU JEU DE DONNÉES

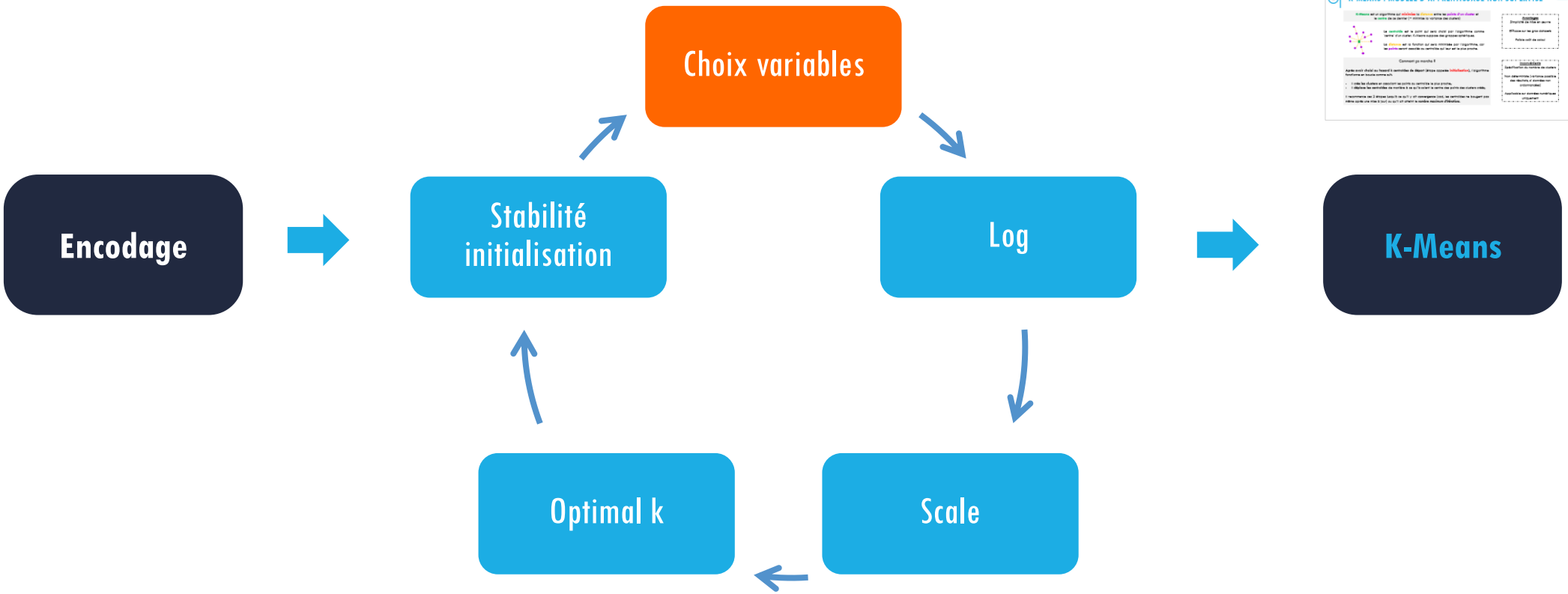
97% des clients ont fait un **achat unique** (3% de réachat)

Les outliers n'ont pas été exclus : si on prend l'exemple du Revenu, la moyenne du revenu par client est de 165, mais certains clients ont fait un achat allant jusqu'à 13.700 ; ils représentent de fait **un segment particulier de clients**
(*voir figure ci-dessous*)

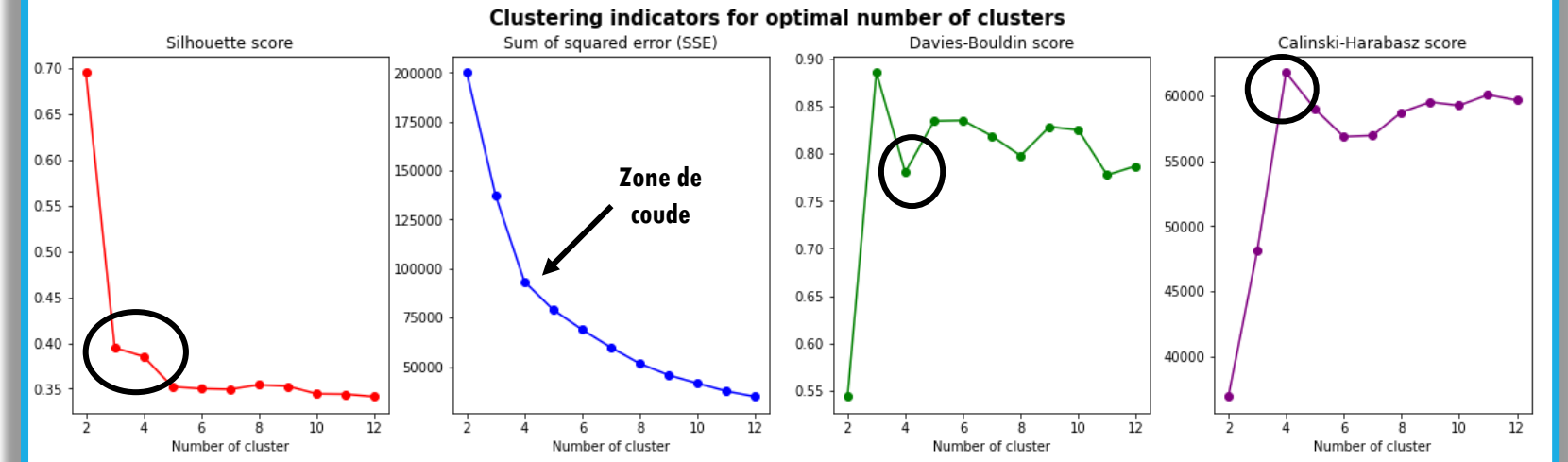


ESSAIS DE MODÉLISATION

PIPELINE DE MODÉLISATION



ÉVALUATION DU NOMBRE OPTIMAL DE CLUSTERS



Coefficient de silhouette
(à maximiser)

Différence entre les distances intra-cluster et les distances inter-cluster le plus proche → 0 (pire) et 1 (meilleur)

Méthode du coude
(Méthode visuelle)

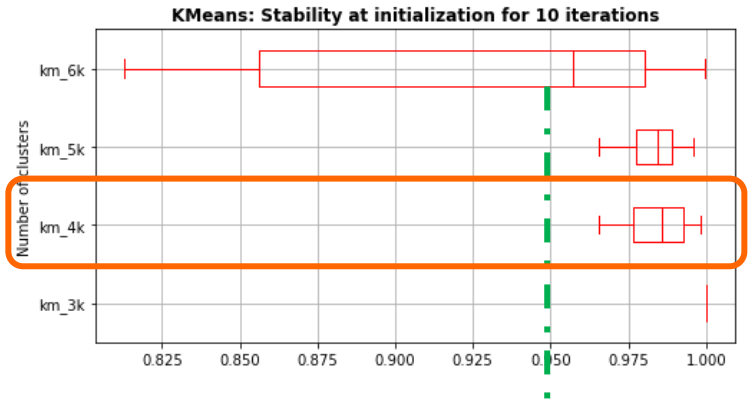
Evalue la fonction coût du K-Means – inertia, qui minimise la distance intra-cluster, tout en conservant un nombre de clusters optimal

Indice Davies-Bouldin
(à minimiser)

Rapport entre la distance des points à leur centroïde et la distance entre deux centroïdes → 0 (meilleur) et +infini (pire)

Indice Calinski-Harabasz
(à maximiser)

Rapport entre la variance inter-cluster et la variance intra-cluster



Stabilité à l'initialisation
Indice de Rand Ajusté (ARI)

Mesure la proportion de paires de points qui sont groupées **de la même façon** dans deux partitions.

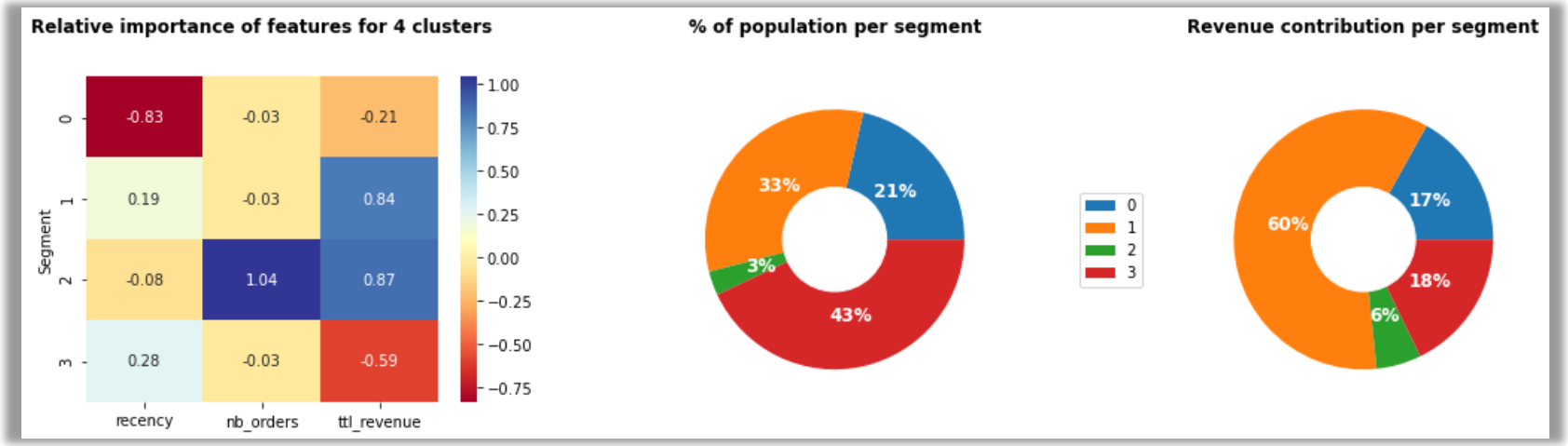
Sur **10 itérations d'initialisation aléatoire**, si l'index est proche de 1 (>95%), le clustering correspond quasiment à la partition initiale.

Chacun de ces indices fournit une mesure de la qualité du clustering (homogénéité, séparation, stabilité) en fonction du nombre de clusters.
Ici, le nombre optimal de clusters est 4.

SEGMENTATION RFM : 4 SEGMENTS

La **segmentation RFM** prend en compte:

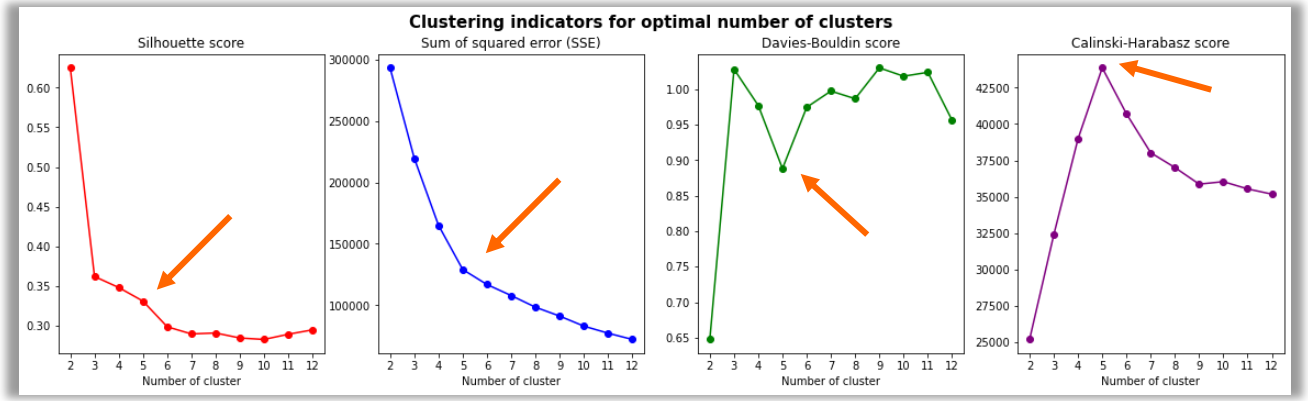
- La **Récence** : différence entre la dernière date de commande de l'intégralité de la base et la dernière date de commande du client;
- La **Fréquence**: ici, le nombre de commande;
- Le **Monétaire**: le montant de la commande

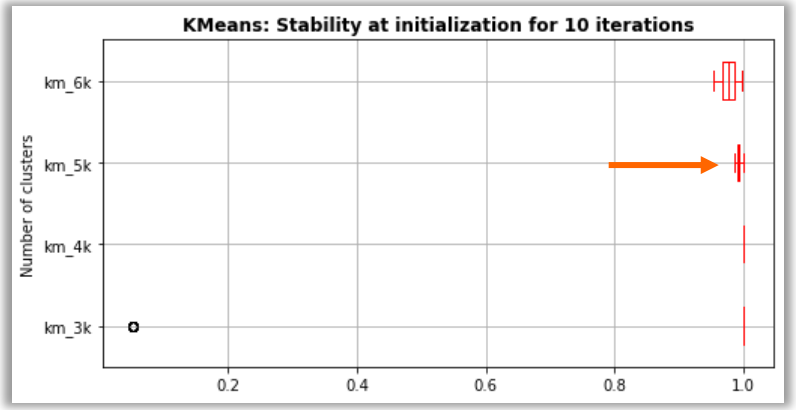


segment	recency	nb_orders	ttl_revenue
0	1.24	1.00	130.14
1	8.76	1.00	303.12
2	6.80	2.11	308.62
3	9.42	1.00	68.26
mean_pop	7.37	1.03	165.15

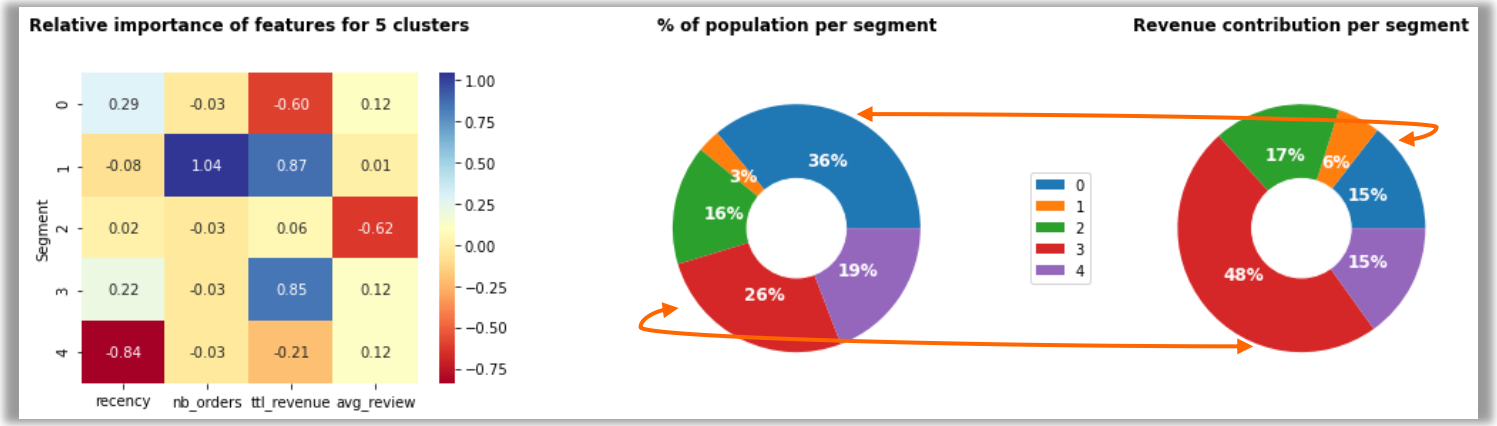
Attention à la récence et la fréquence
Avec 97% des clients en commande unique, les clients de 2017 sont pénalisés par rapport à ceux de 2018 en termes de récence.
De même, la fréquence avantage les 3% de clients ayant fait du réachat.

RFM + SATISFACTION : 5 SEGMENTS





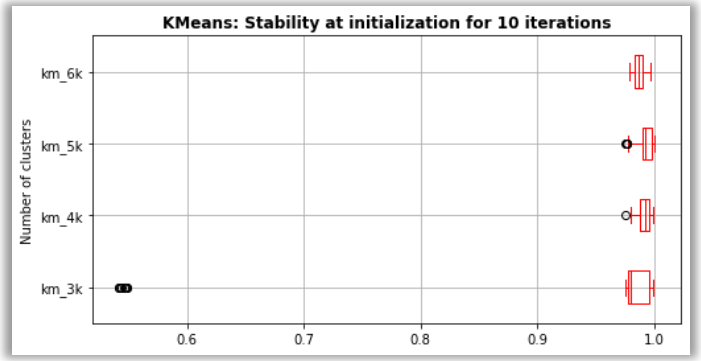
segment	recency	nb_orders	ttl_revenue	avg_review
0	9.49	1.00	66.38	4.62
1	6.80	2.11	308.62	4.19
2	7.50	1.00	174.95	1.57
3	8.96	1.00	305.09	4.65
4	1.19	1.00	129.86	4.63
mean_pop	7.37	1.03	165.15	4.14

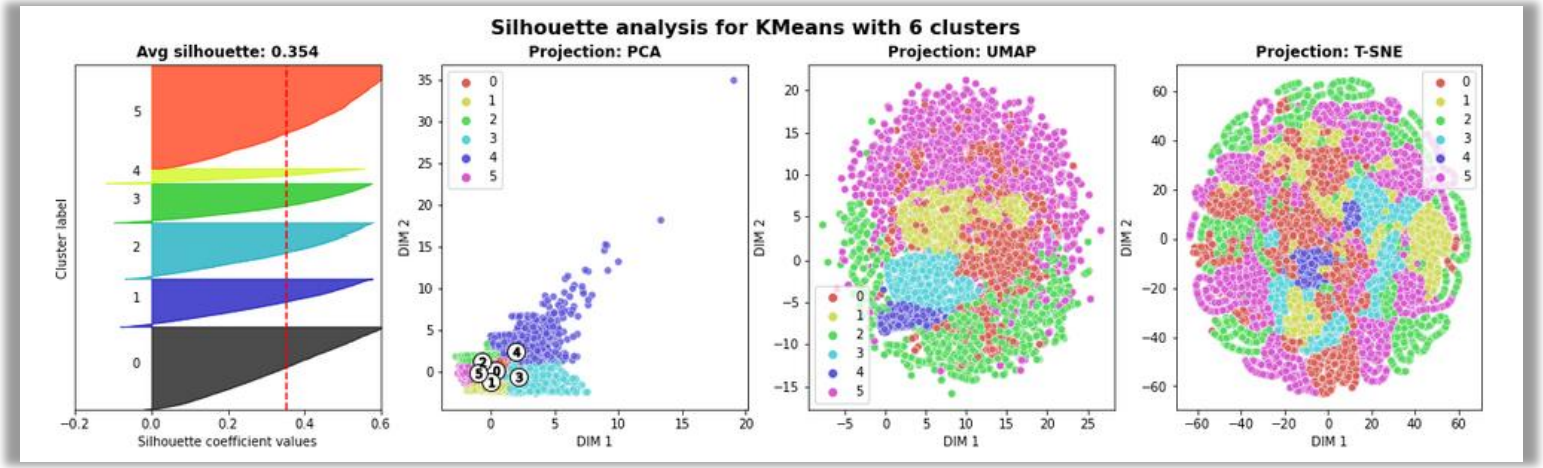


Un groupe de clients très mécontents

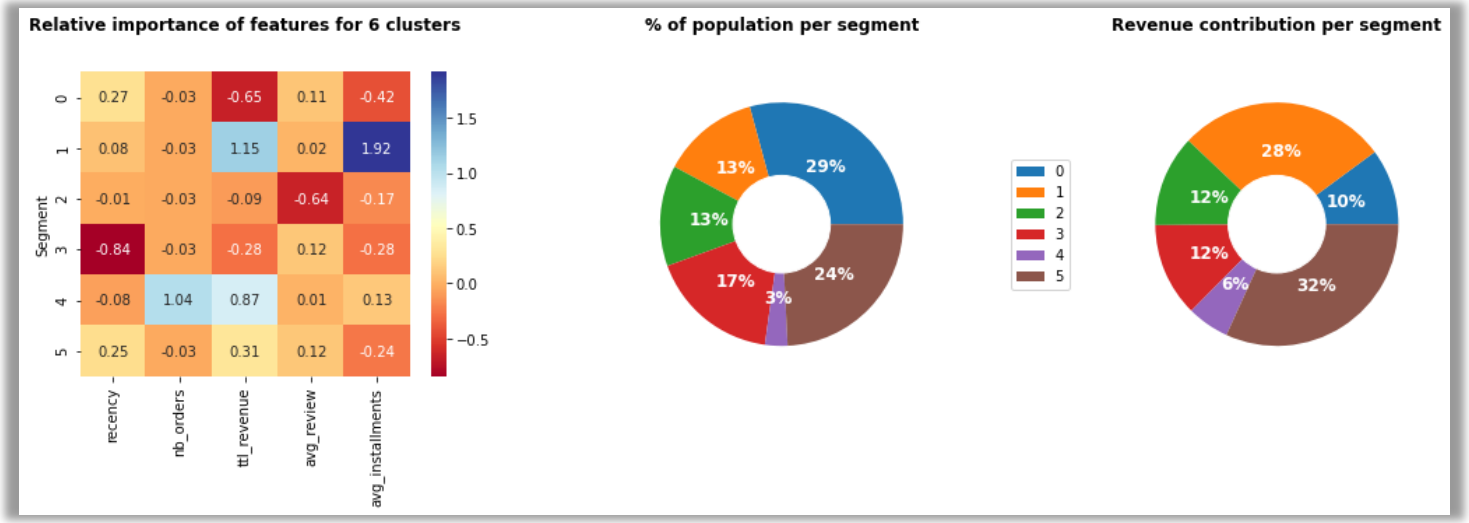
Proportion population/revenu déséquilibrée

RFM + SATISFACTION + FACILITÉS DE PAIEMENT : 6 SEGMENTS





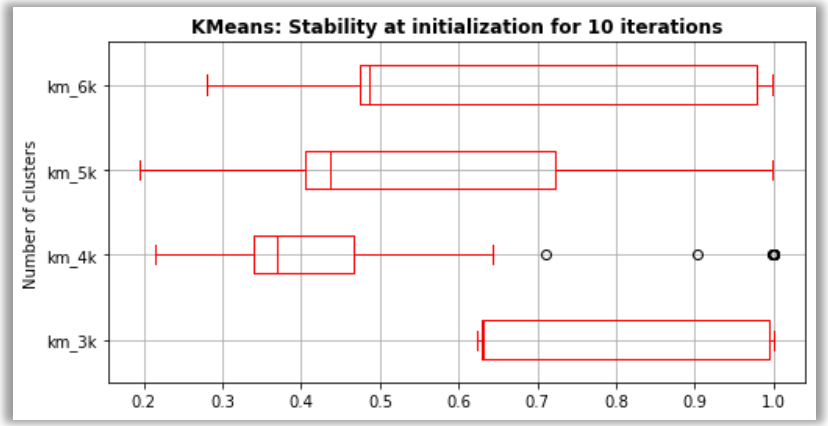
segment	recency	nb_orders	ttl_revenue	avg_review	avg_installments
0	9.35	1.00	57.08	4.60	1.69
1	7.99	1.00	355.20	4.24	8.50
2	7.27	1.00	149.90	1.49	2.42
3	1.15	1.00	119.48	4.63	2.09
4	6.80	2.11	308.62	4.19	3.30
5	9.19	1.00	217.05	4.65	2.21
mean_pop	7.37	1.03	165.15	4.14	2.91



On note ici une bonne stabilité à l'initialisation, l'homogénéité des clusters, avec des caractéristiques spécifiques par segment.

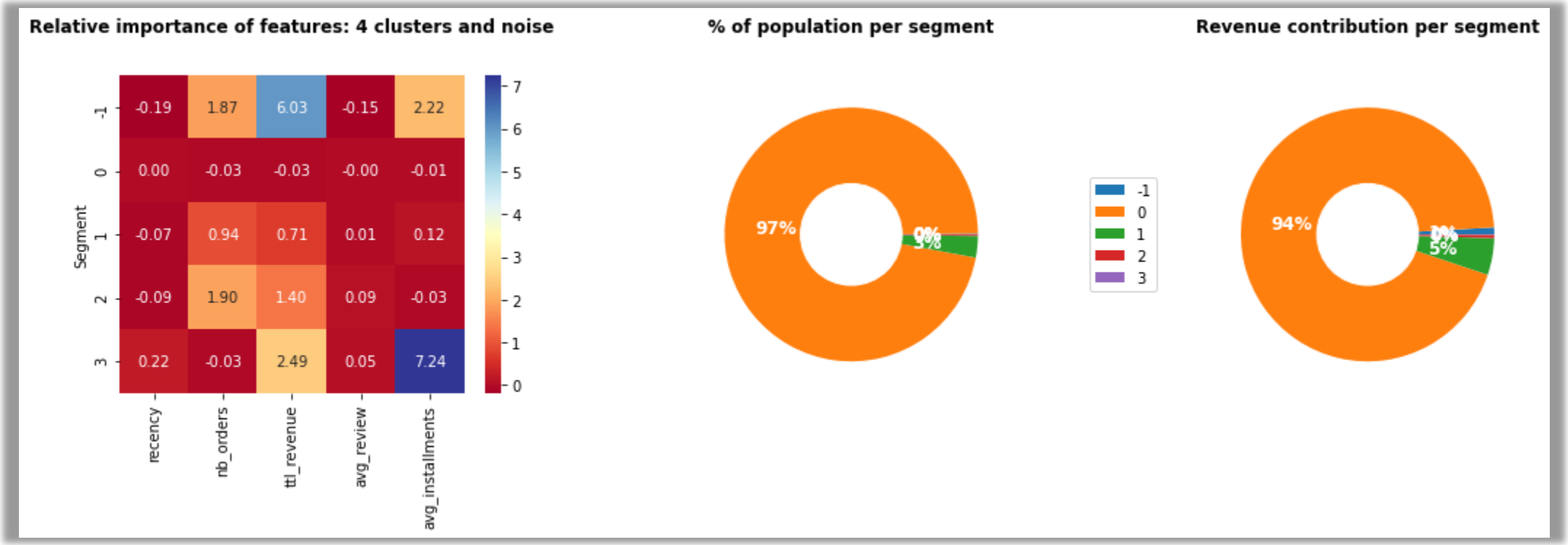
RAJOUT DES VARIABLES CATÉGORIELLES : NON CONCLUANT

	recency	nb_orders	ttl_revenue	avg_review	avg_installments	main_region	main_category	purchase_time
count	93,104.00	93,104.00	93,104.00	93,104.00	93,104.00	93,104.00	93,104.00	93,104.00
mean	-0.00	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
std	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
min	-2.36	-0.16	-2.93	-2.43	-1.08	-3.04	-1.71	-1.77
25%	-0.62	-0.16	-0.70	-0.11	-0.71	-0.34	-1.03	-0.48
50%	0.26	-0.16	-0.05	0.67	-0.34	0.56	-0.01	0.17
75%	0.77	-0.16	0.60	0.67	0.40	0.56	0.67	0.82
max	1.47	66.82	5.93	0.67	7.84	0.56	2.03	1.46



Pas de stabilité à l'initialisation

ESSAI AVEC DBSCAN : NON CONCLUANT



! Pas de tuning d'hyperparamètres

JEU DE DONNÉES : ANALYSE

SEGMENTATION MÉTIER ET CAMPAGNES DE COM'

Champions

13% des clients, 28% du revenu, satisfaits, utilisateurs de facilités de paiement

Personnalisation produits, facilités de paiement améliorés ou nouvelles solutions de financement, …

Unsatisfied

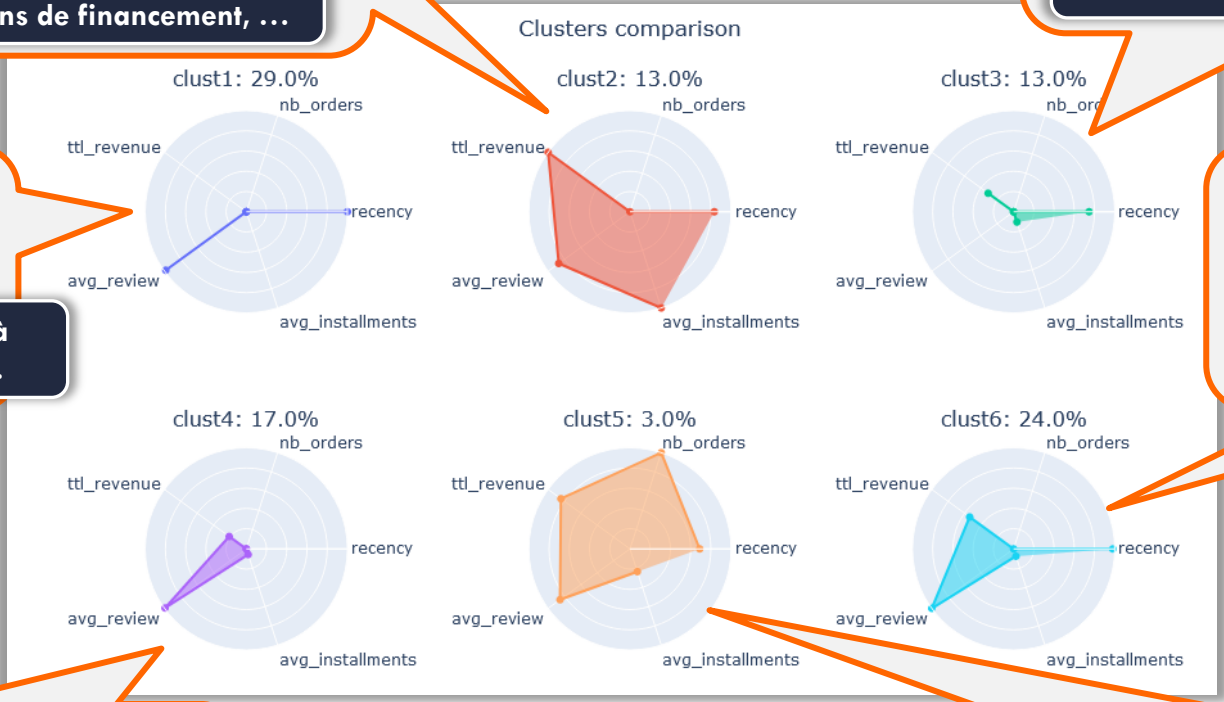
13% des clients, 12% du revenu, très mécontents

Adressage des points douloureux, com° sur la qualité de service, presque perdus

Small pocket

Presque 1/3 des clients, 10% du revenu, satisfaits

Incitations tarifaires agressives: offres à durée limitée, coupons de réductions,…



Inactive good value

24% des clients, 32% du revenu, satisfaits

Offres spéciales basées sur l'historique, recommandations cross-sell ou upsell, com° ('Vous nous manquez')…

Relatively recent

17% des clients, 12% du revenu, satisfaits, achats relativement récents

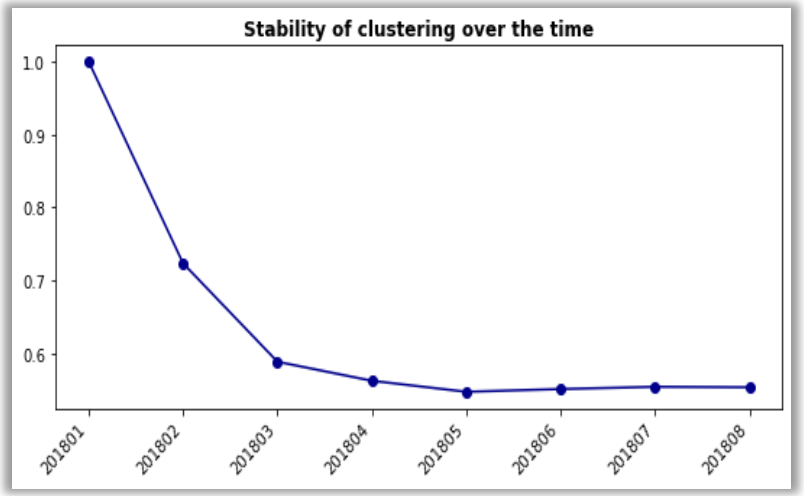
Offre de bienvenue pour bâtir la fidélité

Loyal

3% des clients, 6% du revenu, multiples achats

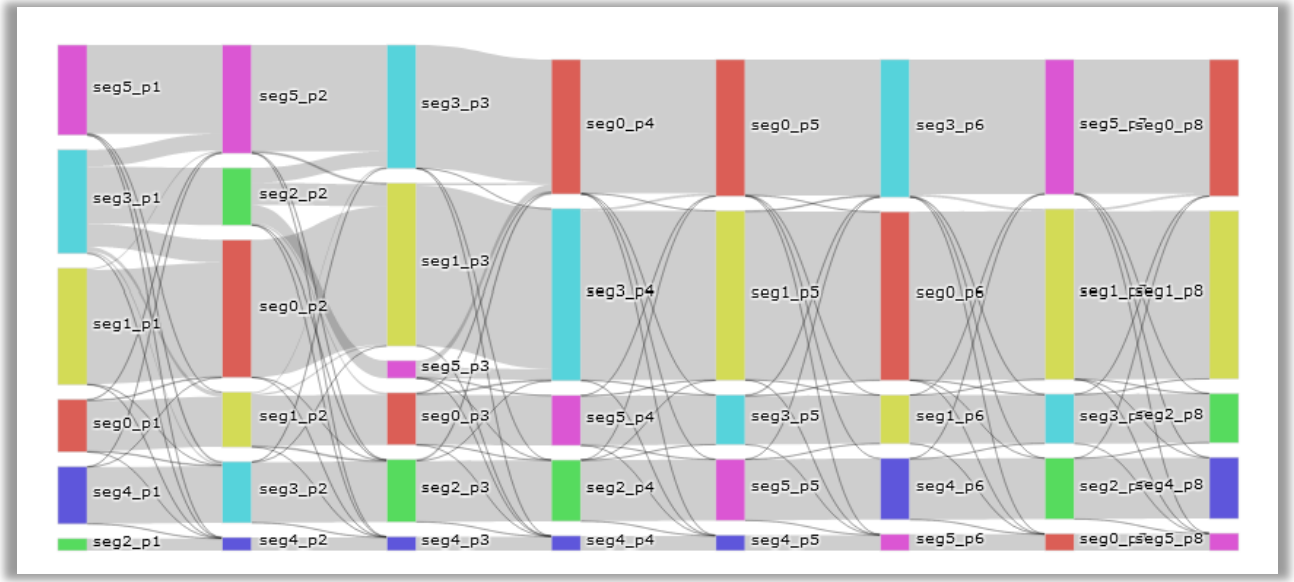
Programmes de fidélité, système de parrainage,…

STABILITÉ TEMPORELLE ET CONTRAT DE MAINTENANCE



Pour vérifier la stabilité temporelle de notre segmentation, nous avons :

- entraîné le modèle sur les clients de la 1ère année (janv-2017 + 365 jours)
- suivi les prédictions du numéro de segment des mêmes clients sur les périodes suivantes (janv-2018 à aout-2018)



Puis nous avons dessiné un diagramme de Sankey pour visualiser les flux

Recommandation quant à la fréquence de mise à jour : **tous les 2-3 mois**

SYNTHÈSE

QUE POUVONS-NOUS EN CONCLURE ?

Une structure de données qui complexifie la segmentation clients : 97% des clients à commande unique, récence biaisée, …

Des données clients incomplètes, d'un point de vue socio-démographique: âge, sexe, CSP, etc. pour parvenir à une segmentation plus pointue

Manque de visibilité sur les marges de manœuvre des équipes Marketing pour construire des variables pertinentes, réellement orientées métier

Proposition de segmentation sur 6 clusters, qui apporte une relative finesse d'analyse

5 Thématiques principales dégagées pour le clustering : récence, fréquence, valeur, satisfaction, facilités de paiement

Segmentation nécessitant une mise à jour tous les trimestres (maintenance régulière)

QUESTIONS / RÉPONSES

ANNEXES

BRAZIL REGIONS

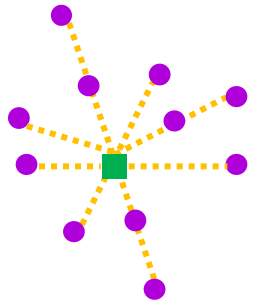
Based on Wikipedia details, a specific CSV file has been created to add CUSTOMERS dataframe with relevant information on **regions**, linked to states information :

Brazil regions

Regions	States
North	**AC** (Acre), **AP** (Amapa), **AM** (Amazonas), **PA** (Para), **RO** (Rondonia), **RR** (Roraima), **TO** (Tocantins)
NorthEast	**AL** (Alagoas), **BA** (Bahia), **CE** (Ceara), **MA** (Maranhao), **PB** (Paraiba), **PE** (Pernambuco), **PI** (Piaui), **RN** (Rio Grande Do Norte), **SE** (Sergipe)
CenterWest	**DF** (Distrito Federal), **GO** (Goias), **MT** (Mato Grosso), **MS** (Mato Grosso Do Sul)
SouthEast	**ES** (Espirito Santo), **MG** (Minas Gerais), **RJ** (Rio De Janiero), **SP** (Sao Paulo)
South	**PR** (Parana), **RS** (Rio Grande Del Sul), **SC** (Santa Catarina)

K-MEANS : MODÈLE D'APPRENTISSAGE NON SUPERVISÉ

K-Means est un algorithme qui **minimise** la **distance** entre les **points d'un cluster** et le **centre** de ce dernier (~ minimise la variance des clusters)



Le centroïde est le point qui sera choisi par l'algorithme comme 'centre' d'un cluster. K-Means suppose des grappes sphériques.

La distance est la fonction qui sera minimisée par l'algorithme, car les **points** seront associés au centroïde qui leur est le plus proche.

Comment ça marche ?

Après avoir choisi au hasard k centroïdes de départ (étape appelée **initialisation**), l'algorithme fonctionne en boucle comme suit:

- Il **crée les clusters** en associant les points au centroïde le plus proche;
- Il **déplace les centroïdes** de manière à ce qu'ils soient le centre des points des clusters créés;

Il recommence ces 2 étapes jusqu'à ce qu'il y ait ***convergence*** (cad, les centroïdes ne bougent pas même après une mise à jour) ou qu'il ait atteint le ***nombre maximum d'itérations.***

Avantages

Simplicité de mise en œuvre

Efficace sur les gros datasets

Faible coût de calcul

Inconvénients

Spécification du nombre de clusters

Non déterministe (variance possible des résultats, si données non ordonnancées)

Applicable sur données numériques uniquement

RÉFÉRENCES

- Jeu de données : Kaggle du dataset OLIST
- Régions du Brésil : wikipedia
- Vue d'ensemble des algorithmes de clustering avec scikit-learn - Clustering
- Méthodes/Indices de qualification du nombre de clusters : Elbow method, Davies-Bouldin
- Evaluation de la performance des clusters par visualisation graphique : scikit-learn - Silhouette Analysis, PCA et T-SNE , UMAP
- Comparaison de 2 clusters avec scikit-learn - Rand Index
- Sankey diagram : documentation, plotly.com, towardsdatascience.com, python-graph-gallery.com

Ce document a été produit dans le cadre de la soutenance du projet n°5 du parcours Ingénieur IA d'OpenClassrooms :
« Segmentez des clients d'un site e-commerce »

Mentor : Thierno DIOP
Evaluateur : Bertrand BEAUFILS

OPENCLASSROOMS
En partenariat avec
CentraleSupélec Microsoft