Segmentation des clients d'une plateforme de vente en ligne

présentation du 13 septembre 2021

Plan de la présentation

1) Problématique d'Olist

2) Description des données Olist

3) Feature engineering, exploration, nettoyage des données

4) Segmentation des clients par des modèles d'apprentissage non supervisé :
   a) modèles testés
   b) k-means

5) Stabilité dans le temps du modèle retenu

6) Description des segments identifiés

7) Conclusion

Problématique d'Olist

Contexte : société brésilienne SaaS (software as a service) qui facilite l'accès aux « market places » pour des sociétés de vente en ligne.

Objectif d'Olist : disposer d'une segmentation des clients qui achètent en ligne.

Missions confiées :

- aider les équipes d'Olist à comprendre les différents types d'utilisateurs.

- proposer une segmentation exploitable et facile d'utilisation pour l'équipe marketing.

- évaluer la fréquence à laquelle la segmentation doit être mise à jour.

Présentation des données

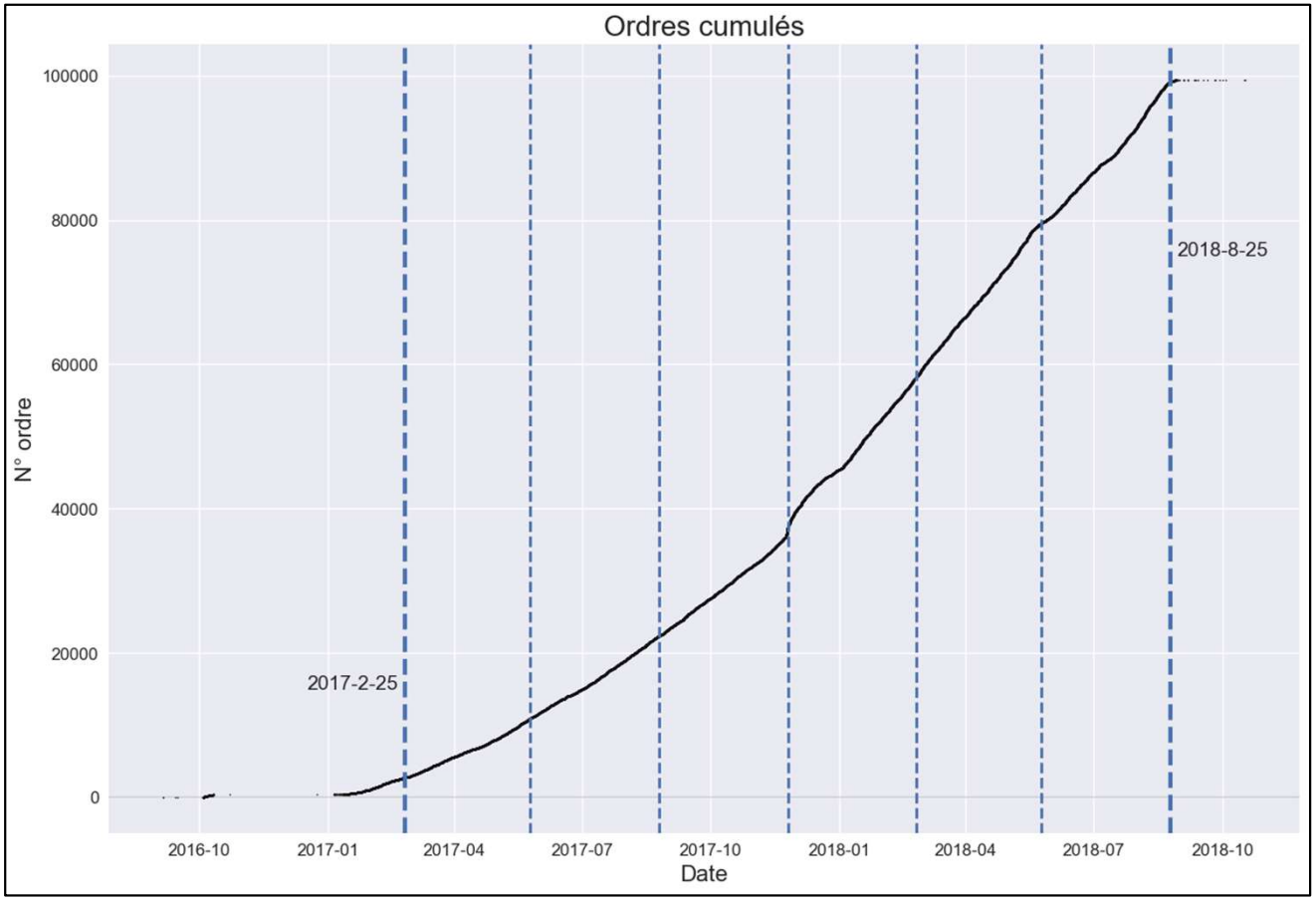
Source : https://www.kaggle.com/olistbr/brazilian-ecommerce

Fichiers .csv	Variables réutilisées	Nb de lignes
Orders	order_id, customer_id	99441
Customers	customer_id, customer_unique_id	99441
Order reviews	order_id, review_score	99441
Order payments	order_id, payment_sequential, payment_installments, payment_value	99440
Items	/	112650
Geolocation	/	19015
Translation	/	71
Products	/	32951
Sellers	/	3095

Note : nb de clients uniques = 96096

Séparation des données en 6 trimestres

Séparation en 6 trimestres (pour tester la stabilité des partitionnements obtenus)



Trimestre	Nombre d'ordres
1	8210
2	11408
3	15154
4	20786
5	21302
6	19567

Feature engineering

Inspiré par la segmentation RFM (récence, fréquence, montant).

Création des variables suivantes (calculées sur un trimestre et sur un client unique) :

- **récence** : durée (j) entre le dernier achat du trimestre et la fin du trimestre (pour un client, pour un trimestre).

- fréquence : **nombre de commandes** (pour un client, pour un trimestre).

- **montant total** : somme des commandes (pour un client, pour un trimestre) en reals.

- **montant moyen** : montant total / fréquence (pour un client, pour un trimestre).

- **installments** : moyenne du nombre d'installments par commande (pour un client, pour un trimestre).

Variables retenues pour l'apprentissage non supervisé

Dans une phase préliminaire (incluant l'exploration de données) :
- récence,
- nb de commandes moyen,
- montant total,
- montant moyen,
- nb d'installments moyen,
- review score moyen.

Finalement (pour l'apprentissage supervisé) :
- ~~récence,~~
- nb de commandes moyen,
- montant total,
- montant moyen,
- nb d'installments moyen,
- review score moyen.

Résultats dans la présentation : sans récence (sauf mentions particulières)

Données utilisées

On travaille avec un dataset par trimestre (6 datasets en tout).

Exemple :

extrait du tableau des données pour le trimestre 6 :

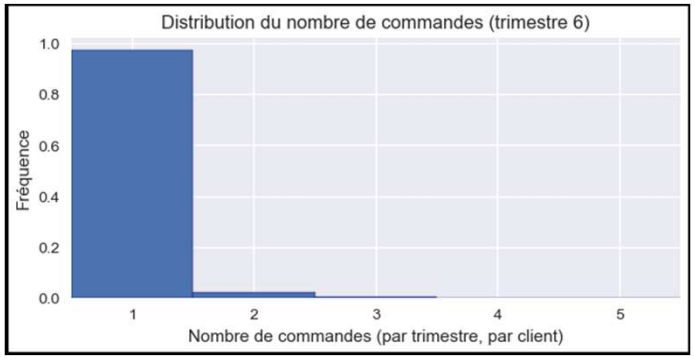
récence	nb_commandes	montant_total	montant_moyen	installments	review_score
80.53	1.0	40.36	40.36	4.0	4.0
33.51	1.0	201.34	201.34	3.0	5.0
60.25	1.0	137.70	137.70	10.0	4.0
34.17	1.0	83.79	83.79	8.0	5.0
11.62	2.0	244.14	122.07	1.5	4.0
1.11	1.0	22.29	22.29	1.0	5.0
79.08	1.0	61.96	61.96	4.0	4.0
10.05	1.0	118.98	118.98	6.0	5.0
5.29	1.0	60.05	60.05	1.0	5.0
48.55	1.0	83.26	83.26	1.0	5.0

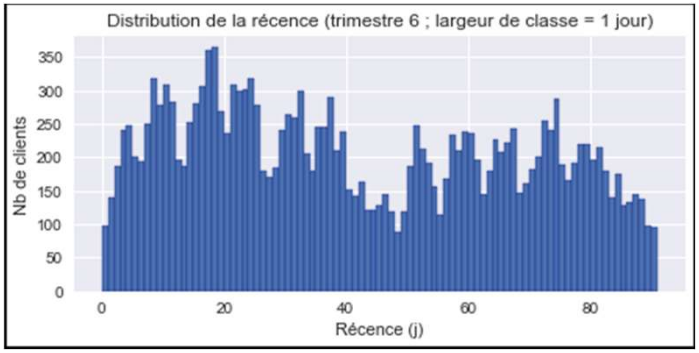
19058 lignes (1 ligne par client unique)

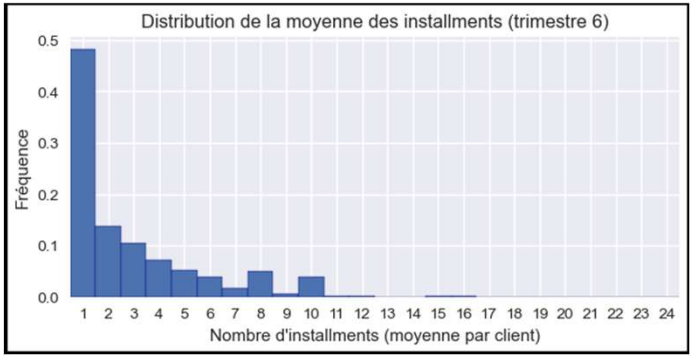
Exploration – analyse monovariée

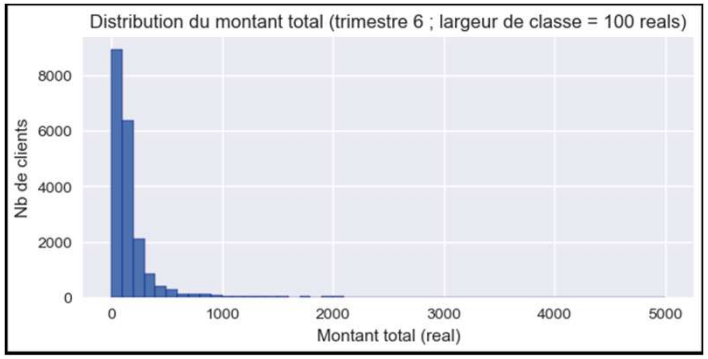
Données complètes et absence de données aberrantes pour les 6 datasets (trimestres)

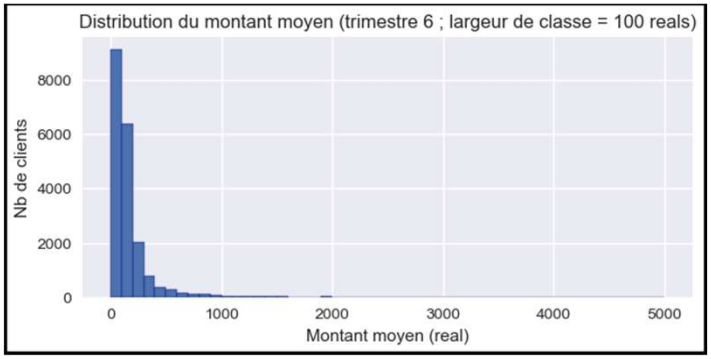
Distribution des variables : exemple représentatif (trimestre 6) :

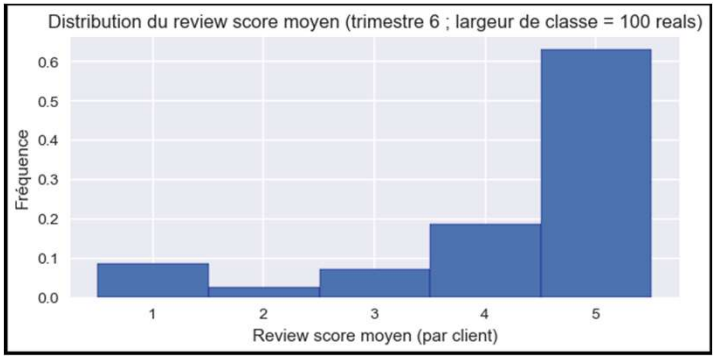








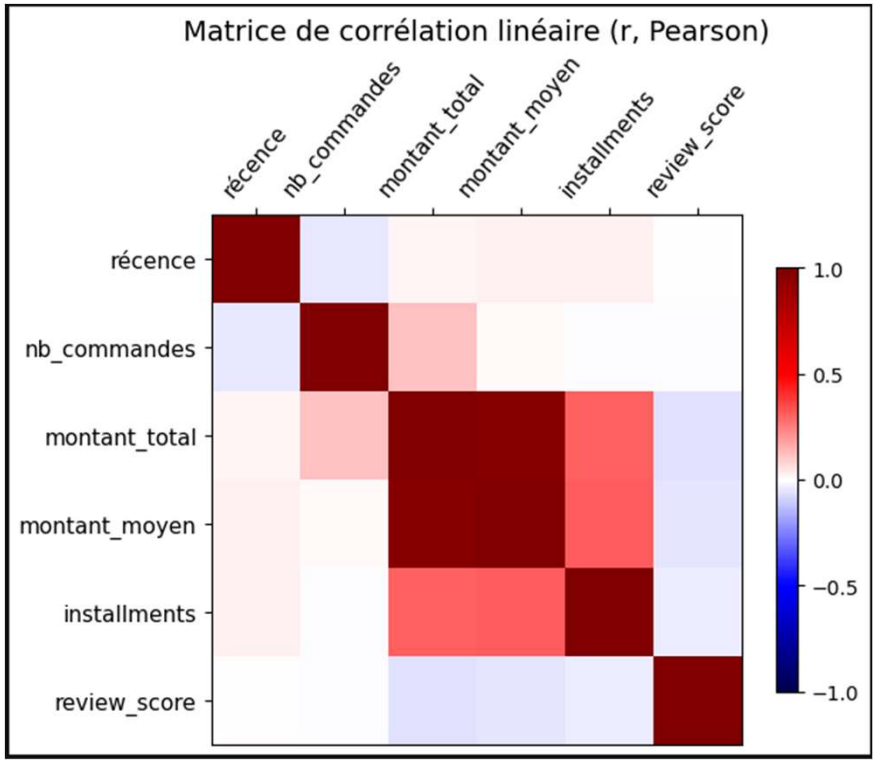




A chaque trimestre : 98 à 99% des clients qui passent une commande n'en passent qu'une seule.

Exploration – Matrice de corrélation linéaire (r, Pearson)

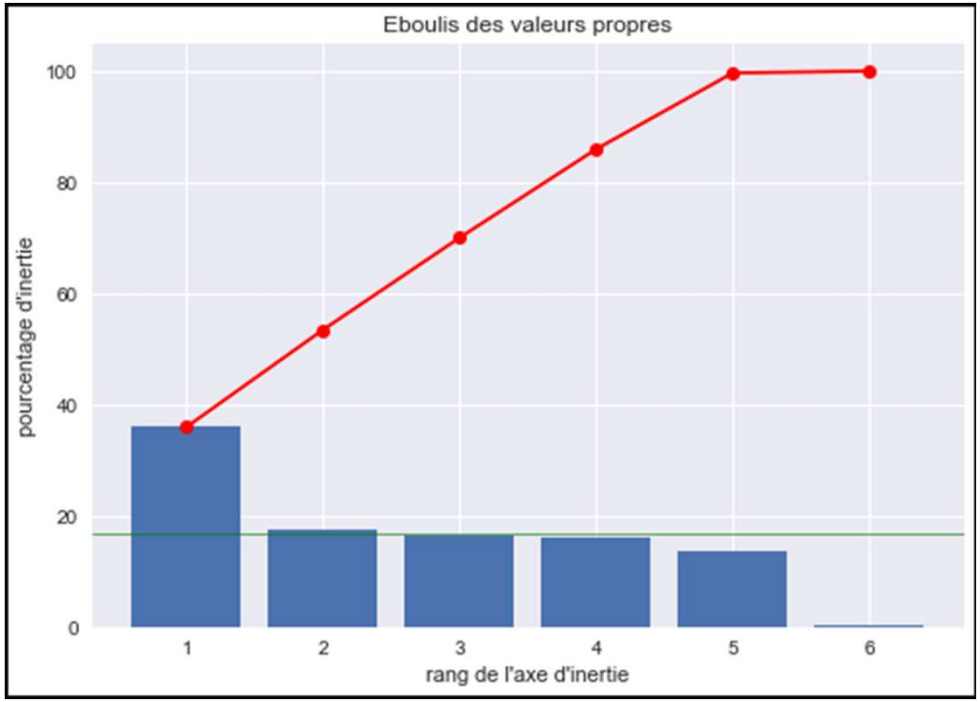
Exemple représentatif : trimestre 6



Très forte corrélation linéaire (r = 0.97 à 99 selon trimestre) entre montant total et montant moyen.
Corrélation linéaire faible (r = 0.30 à 0.37 à selon trimestre) entre entre le nb moyen d'installments et les montants.
Très faible corrélation linéaire (|r| < 0.1) entre les autres paires de variables

Exploration – analyse en composantes principales

Etude des corrélations entre les 6 variables utilisées - éboulis des valeurs propres :

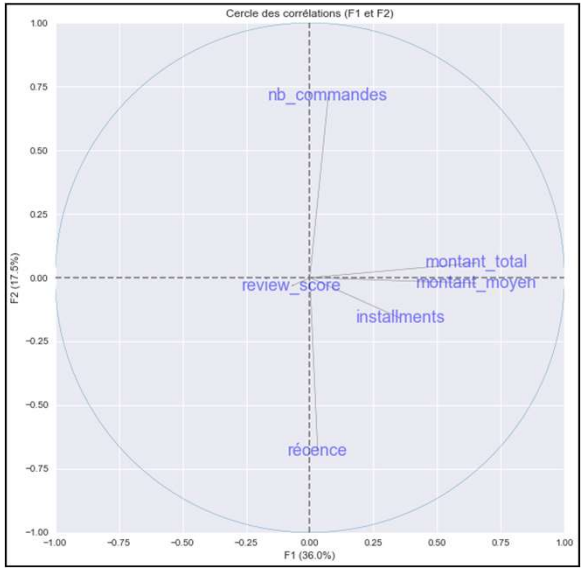


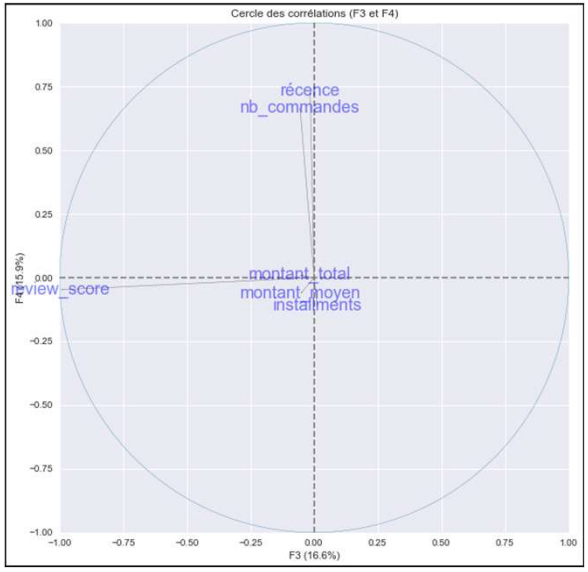
→ répartition des pourcentages d'inertie avec un ratio 2 / 1 / 1 / 1 / 1 / 0 entre les axes.

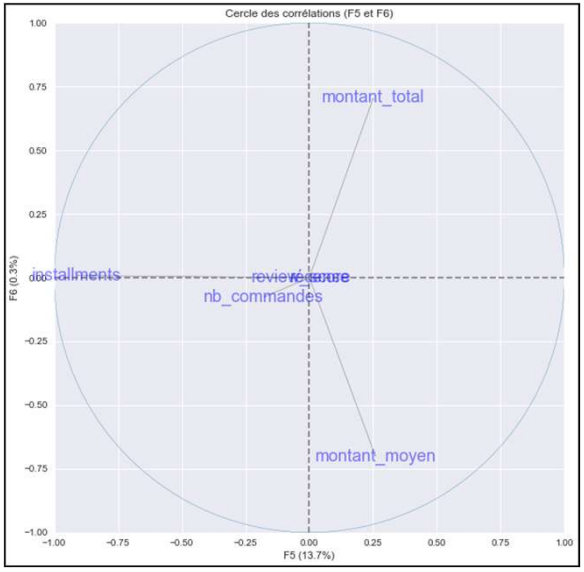
(data utilisée : trimestre 6)

Exploration – analyse en composantes principales

Cercles des corrélations (plans factoriels) :







Les variables suivantes sont principalement projetées selon les axes :

- montant total → F1.
- montant moyen → F1.
- nb de commandes → F2 et F4.
- récence → -F2 et F4.
- review score → -F3.
- installments → -F5.

Nettoyage des données

Dans les données analysées :

- Absence de donnée aberrante ou atypique.

- Absence de valeur manquante.

- Doublons dans les review scores :
  ⇒ nettoyage par conservation des reviews les plus récentes pour chaque commande.

Apprentissage non supervisé - généralités

Modules utilisés :
- scikit-learn.
- scipy (clustering hiérarchique agglomératif).

Analyse de la stabilité des segments au cours du temps :
- comparaison des 6 datasets / trimestres créés.
- métrique privilégiée pour tester la stabilité des clusters : index de Rand ajusté (ARI).

Variables utilisées (rappel) :
- nb de commandes,
- montant total,
- montant moyen,
- nb d'installments moyen,
- review score moyen,
- ~~récence.~~

Toutes les variables ont été **centrées et réduites** (StandardScaler) pour la phase d'apprentissage.

Apprentissage non supervisé – modèles testés

Modèle	Vitesse de calcul pour nos datasets	Stabilité de la convergence du modèle	Outil « predict »	Répartition équilibrée des clusters (choix métier)	Conclusion : modèle retenu
k-means	Elevée	Non	Oui	Oui	**Oui**
Classification ascendante hiérarchique	Suffisante (! RAM)	Oui	Non (sklearn et scipy)	Oui	**Non**
Affinity propagation	Trop lente (CA > $O(n^3)$)	Non	Oui	Oui	**Non**
Mean Shift	Suffisante	n.d.	Oui	Non	**Non**
BIRCH	Très élevée	n.d.	Oui	Non	**Non**

Outil « predict » du modèle :
- pour calculer à quel cluster appartient un nouveau client.
- indispensable pour tester la stabilité des clusters.

Apprentissage non supervisé – k-means

Plan de la présentation k-means :

1) comparaison de scores de silhouette

2) comparaison de silhouettes

3) stabilité inter-trimestres des partitions obtenues

4) partitionnement retenu

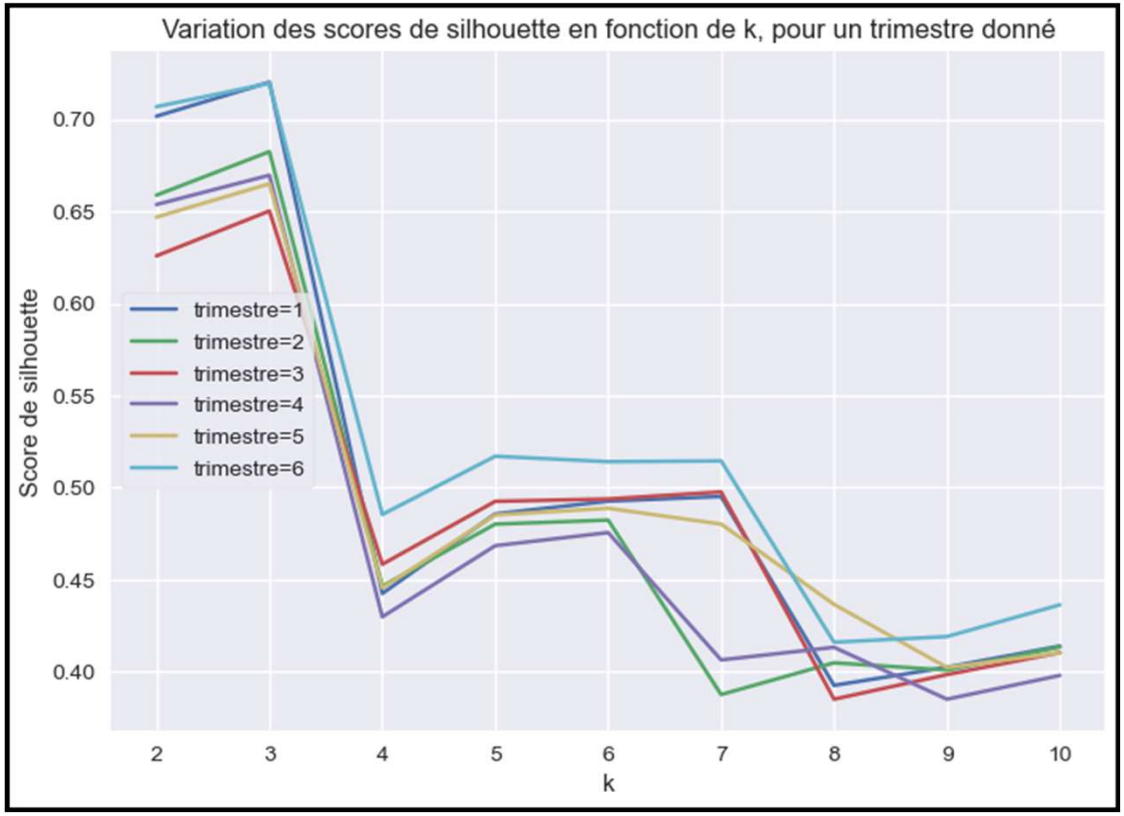
5) description des personas

6) représentation des segments sur un pair plot

7) stabilité du k-means par rapport à l'initialisation

Apprentissage non supervisé – k-means

Scores de silhouette :



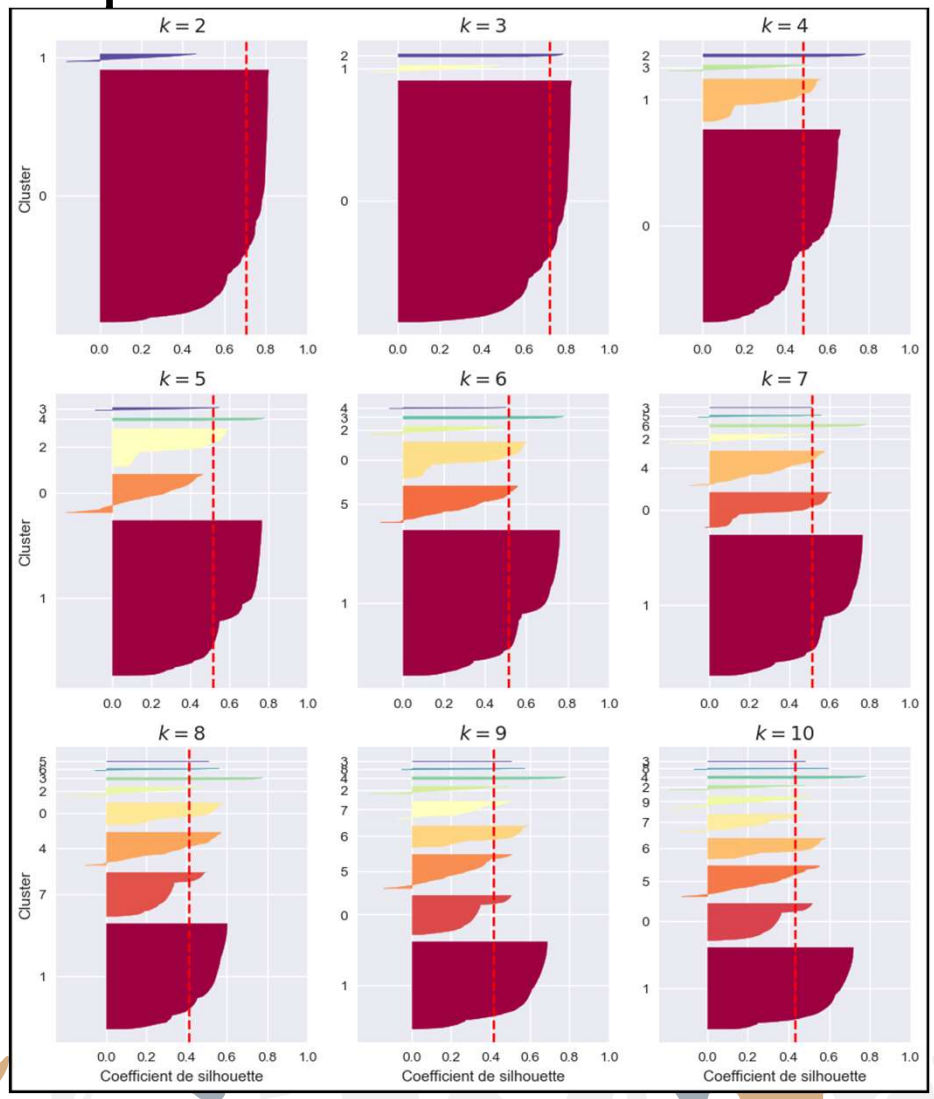
→ Qualité des partitions : k=3 > k=2 >> k=6 > k=5 > k=4 > …

Apprentissage non supervisé – k-means

Silhouette (exemple représentatif : trimestre 6) :

k=2 et k=3 :
qualité du partitionnement
mais **mauvaise répartition** (partitions déséquilibrées)

k=5 et k=6 :
compromis entre la qualité du partitionnement
et la répartition



Apprentissage non supervisé – k-means

Stabilité inter-trimestres des clusters :

Objectif : déterminer si un cluster observé le trimestre n-1 est de nouveau identifié au trimestre n par l'algorithme k-means.

Méthodologie retenue :

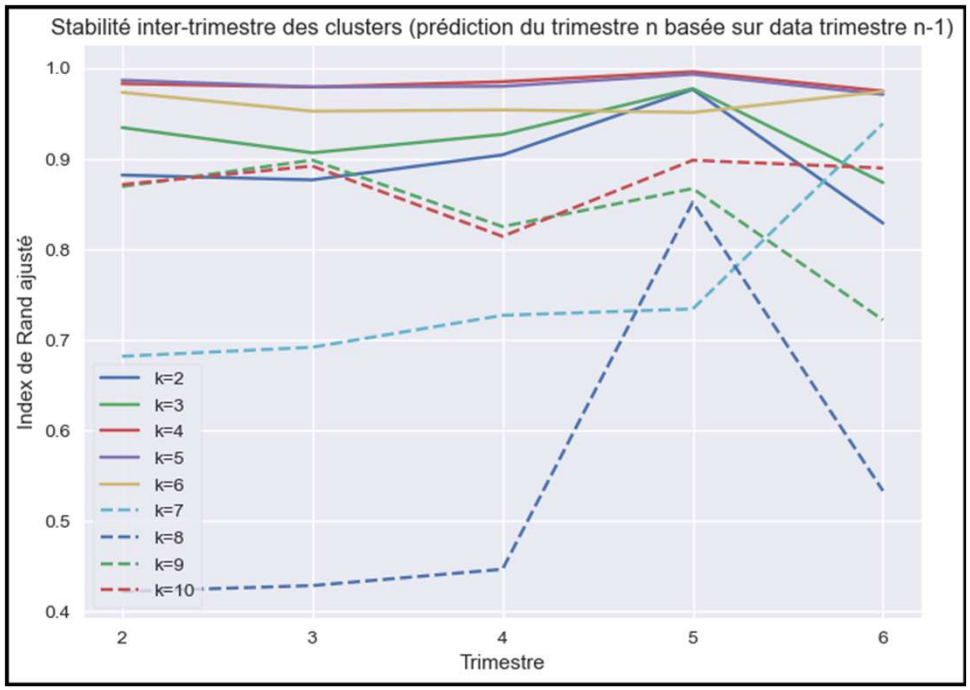
1) entrainer l'algorithme avec les données du trimestre n-1,
2) prédiction des labels des data du trimestre n avec le modèle de clustering du trimestre n-1
3) partitionner séparément les données du trimestre n (avec hyperparamètres identiques)
4) comparer (par ARI) les 2 partitionnements obtenus (pour trimestre n).

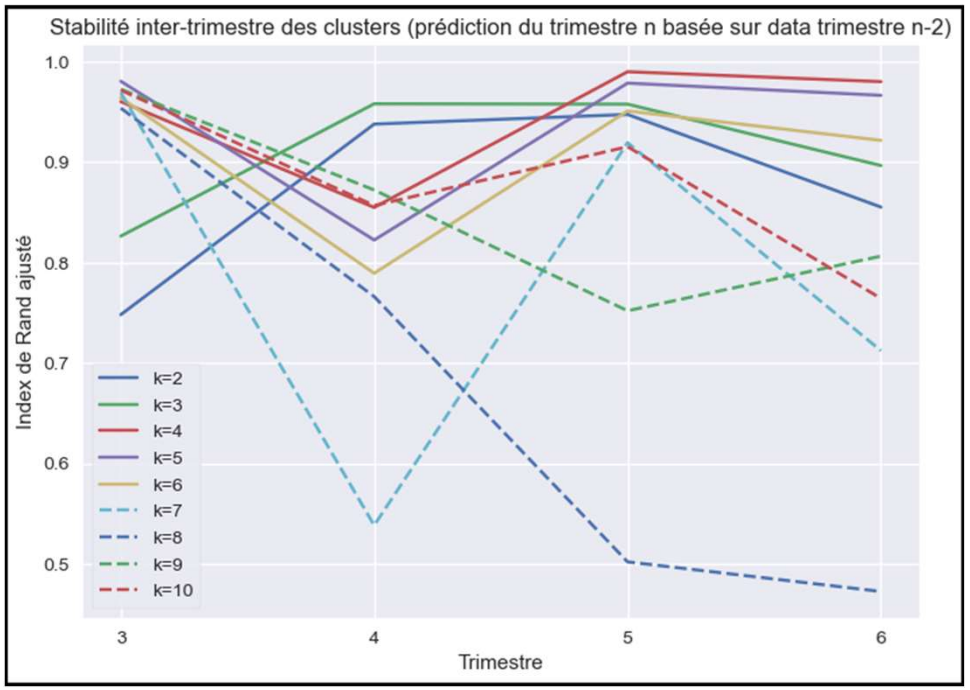
Exemple de prédictions pour k=6 :

labels du trimestre 2 prédits avec un modèle entrainé sur les data du trimestre 1 : ARI = 0.97

Apprentissage non supervisé – k-means

Stabilité inter-trimestres des clusters :





→ stabilité d'un trimestre au suivant très bonnes (ARI ≥ 0.95) pour k = 4, k = 5 et k = 6.
→ stabilité insuffisante si prédiction réalisée avec partitionnement du trimestre n-2

⇒ **Proposition de contrat de maintenance trimestrielle**

Apprentissage non supervisé – k-means

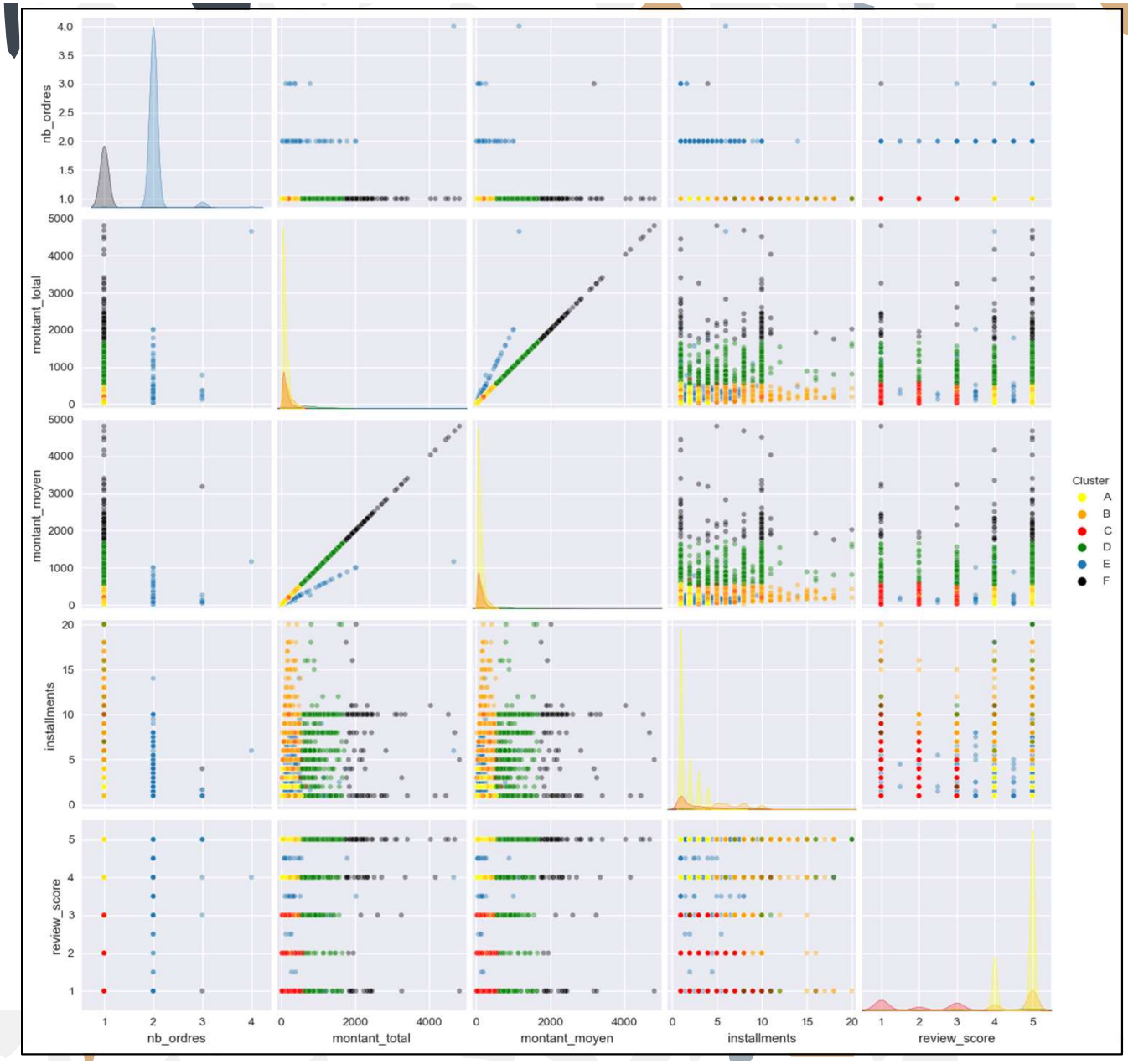
Nombre de partitions retenues :

Meilleurs partitionnements obtenus avec k=5 et k=6 (stabilité inter-trimestres, scores de silhouette)

⇒ **on choisit k=6** (plus de clusters que k=5, et score de silhouette supérieur)

Pair plot (trimestre 6) des clients dans les plans définis par les variables utilisées

Cluster	Descriptif persona
A	Client typique (petit montant, peu d'installments, très satisfait)
B	A payé en un grand nombre de fois
C	Insatisfait
D	A passé une grosse commande
E	A passé 2 commandes ou plus durant le trimestre
F	A passé une très grosse commande



Apprentissage non supervisé – k-means

Personas (les valeurs numériques du tableau correspondent au trimestre 6) :

Résultats du partitionnement			Variables utilisées pour le partitionnement / centroïdes obtenus				
Cluster	Descriptif client	Proportion (%)	Nb de commandes	Montant total	Montant moyen	Install-ments	Review score
A	Client typique (petit montant, peu d'installments, très satisfait)	63.2	1.00	112	112	1.7	4.8
B	Paie en un grand nombre de fois	16.0	1.00	199	199	7.2	4.6
C	Insatisfait	15.8	1.00	133	133	2.4	1.9
D	A passé une grosse commande	3.4	1.00	835	835	6.0	4.1
E	A passé plus de 2 commandes durant le trimestre	1.2	2.04	331	158	2.8	4.2
F	A passé une très grosse commande	0.5	1.02	2571	2496	6.9	3.8

→ valeurs des centroïdes et proportions différentes d'un trimestre à l'autre, personas strictement identiques.
→ segmentation permettant de réaliser un marketing différencié.

Apprentissage non supervisé – k-means

Explication de l'omission de la récence :

	Avec récence	**Sans** récence
Score de silhouette	0.32	0.48
Stabilité inter-trimestre	92%	96%

Conditions : k-means, k=6, moyenne des 6 trimestres.

⇒ Amélioration du score de silhouette et de la stabilité inter-trimestres par omission de la récence

Apprentissage non supervisé – k-means

Stabilité du k-means par rapport à l'initialisation :

Objectif : déterminer si l'algorithme k-means converge vers les mêmes clusters avec des initialisations différentes.

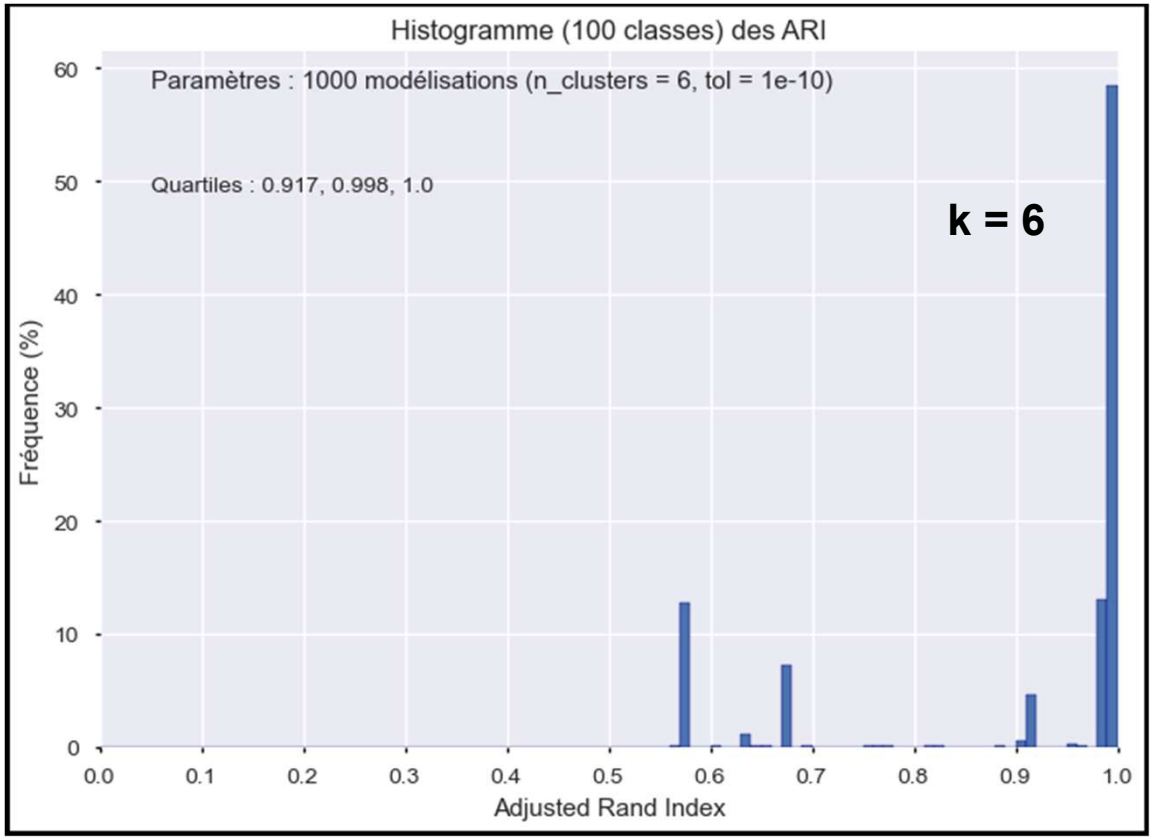
Méthodologie retenue :

1) Utilisation des conditions d'initialisation par défaut de KMeans() de sklearn (avec *init='k-means++'*), mais avec *n_init=1* (un seul run).
2) Réalisation de 1000 runs de KMeans (initialisations aléatoires).
3) Run étalon = run de plus basse inertie.
4) Calcul des ARI des partitionnements pour chaque run par rapport au run étalon.
5) Tracé de l'histogramme des ARI.

Apprentissage non supervisé – k-means

Stabilité du k-means par rapport à l'initialisation :

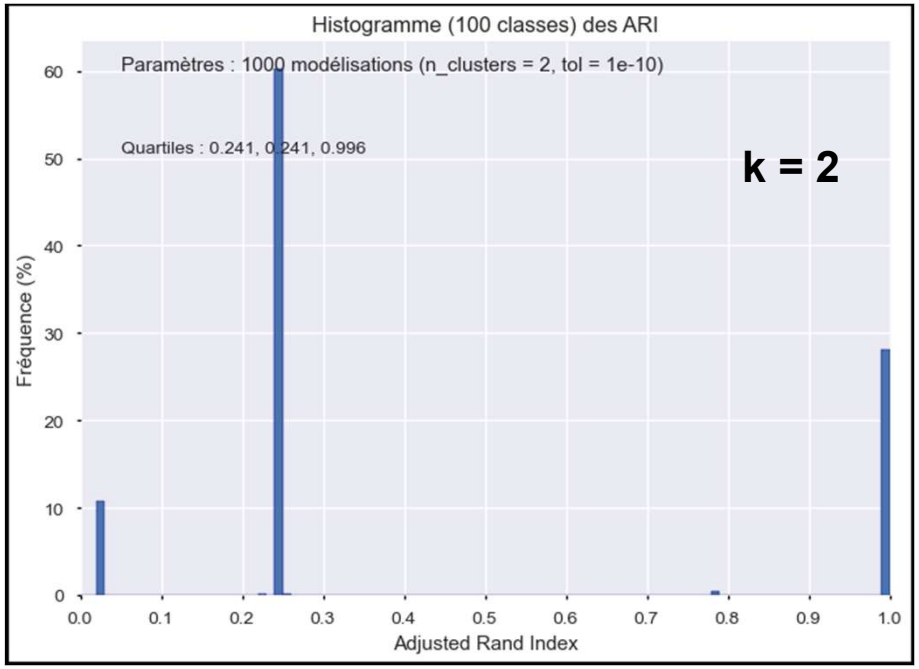
Exemple (trimestre 6) :

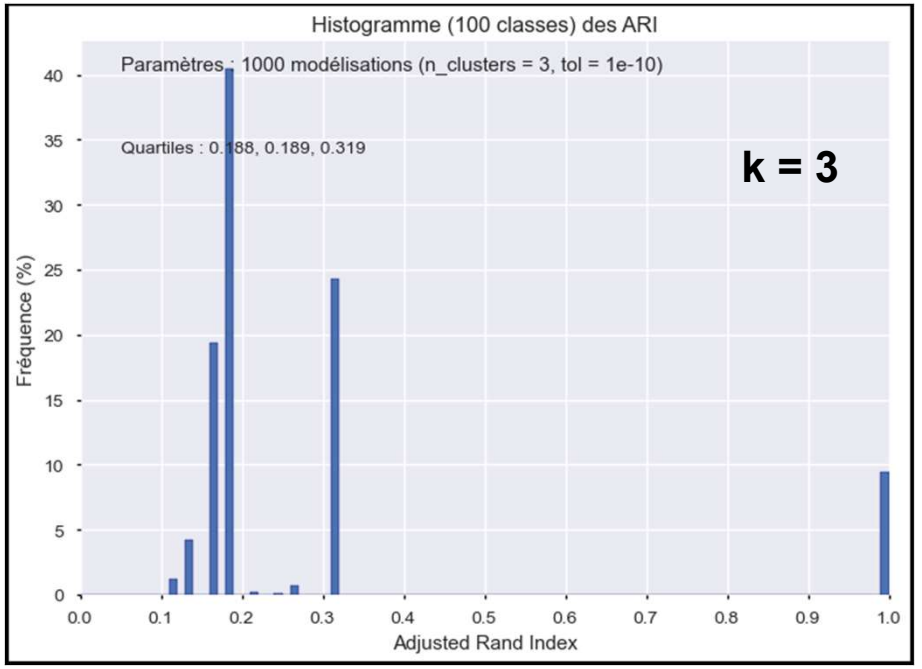


→ Pour k = 6 : convergence fréquente vers le partitionnement de plus basse inertie.

Apprentissage non supervisé – k-means

Stabilité du k-means par rapport à l'initialisation :
Données du trimestre 6 (résultats similaires pour autres trimestres) :





k=2 et k=3 → convergence moins fréquente vers le partitionnement de plus basse inertie.
k entre 4 et 10 → convergence fréquente vers le partitionnement de plus basse inertie.
Solution retenue pour s'assurer d'avoir le partitionnement de plus basse inertie :
toutes les modélisations k-means ont été réalisées avec **n_iter=100** (100 runs).

Bilan de l’apprentissage non supervisé

5 variables utilisées :
- nb de commandes moyen,
- montant total,
- montant moyen,
- nb d’installments moyen,
- review score moyen.

Modèles testés mais non retenus :
- Classification ascendante hiérarchique
- Affinity propagation
- Mean Shift
- BIRCH

Modèle retenu et ses hyperparamètres :
- k-means
- k = 6
- hyperparamètres par défaut de sci-kit learn (sauf : n_init = 100, tol = 1e-10)

Conclusion

Dataset Olist :

Données de qualité (absence de valeurs manquantes ou aberrantes)

Feature engineering nécessaire pour extraire des variables métier pertinentes pour une analyse marketing

Taille du dataset suffisante pour :

- mettre au point un modèle d'apprentissage supervisé.
- tester la stabilité du modèle au cours du temps.

5 modèles testés. Modèle retenu : k-means.

Partitionnement en 6 segments :

Segment	Descriptif persona
A	Client typique (petit montant, peu d'installments, très satisfait)
B	A payé en un grand nombre de fois
C	Insatisfait
D	A passé une grosse commande
E	A passé 2 commandes ou plus durant le trimestre
F	A passé une très grosse commande

Modèle de partitionnement stable d'un trimestre à l'autre mais pas au-delà

⇒ **Proposition de contrat de maintenance trimestrielle**