Segmentation de clients pour le site d’e-commerce Olist

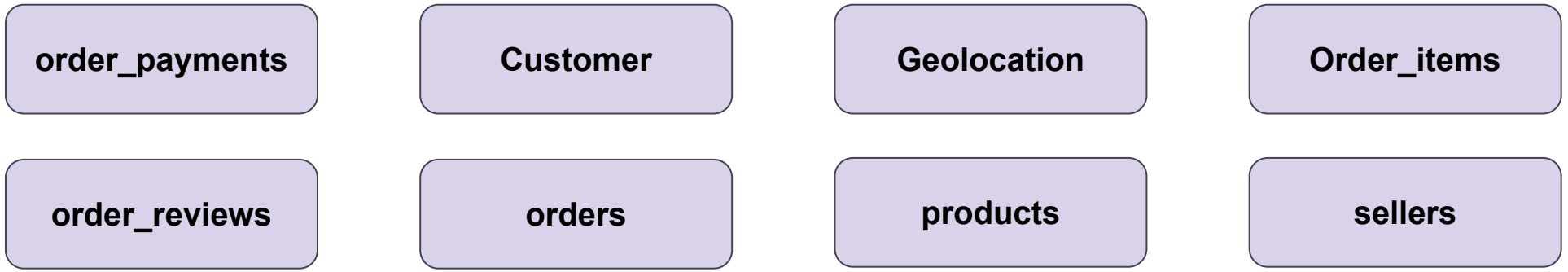
-

Open Classrooms parcours Data Science - projet 5

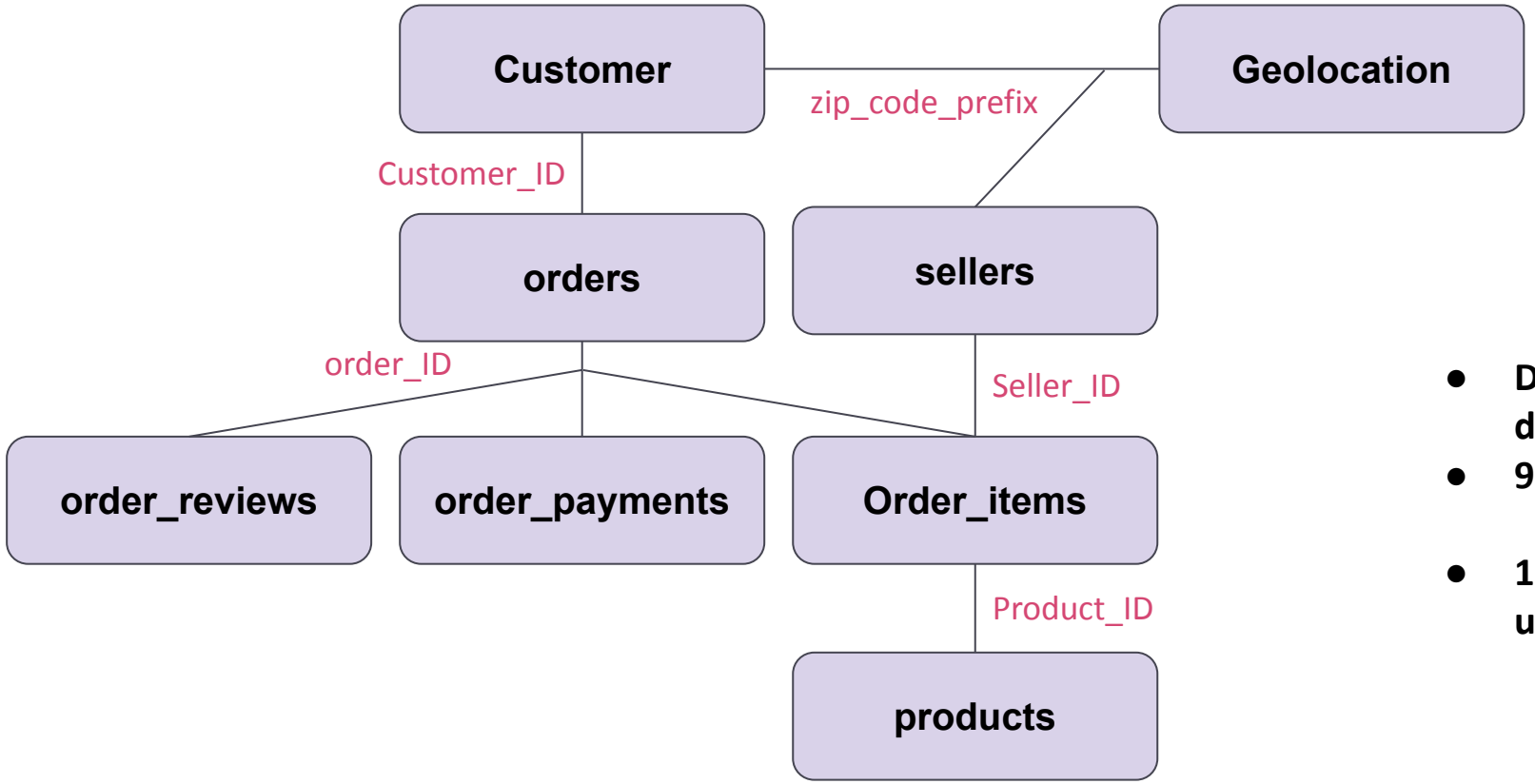
Camille Besançon

Données d'origine

- Olist : Site d'e-commerce Brésilien
- Informations de commandes de clients
  - Répertorient les identifiants de commandes, clients uniques, montants de commande, modes de paiement, nombre d'objets, catégories, satisfaction du client, …
  - Données depuis Sept. 2016 jusqu'à Sept.2018
- Objectif : Fournir aux équipes d'Olist une segmentation des clients du site grâce à leurs données. Décrire les "profils types" associés à ces groupes pour permettre aux équipes marketing de définir les prochaines actions à réaliser.



Organisation des données

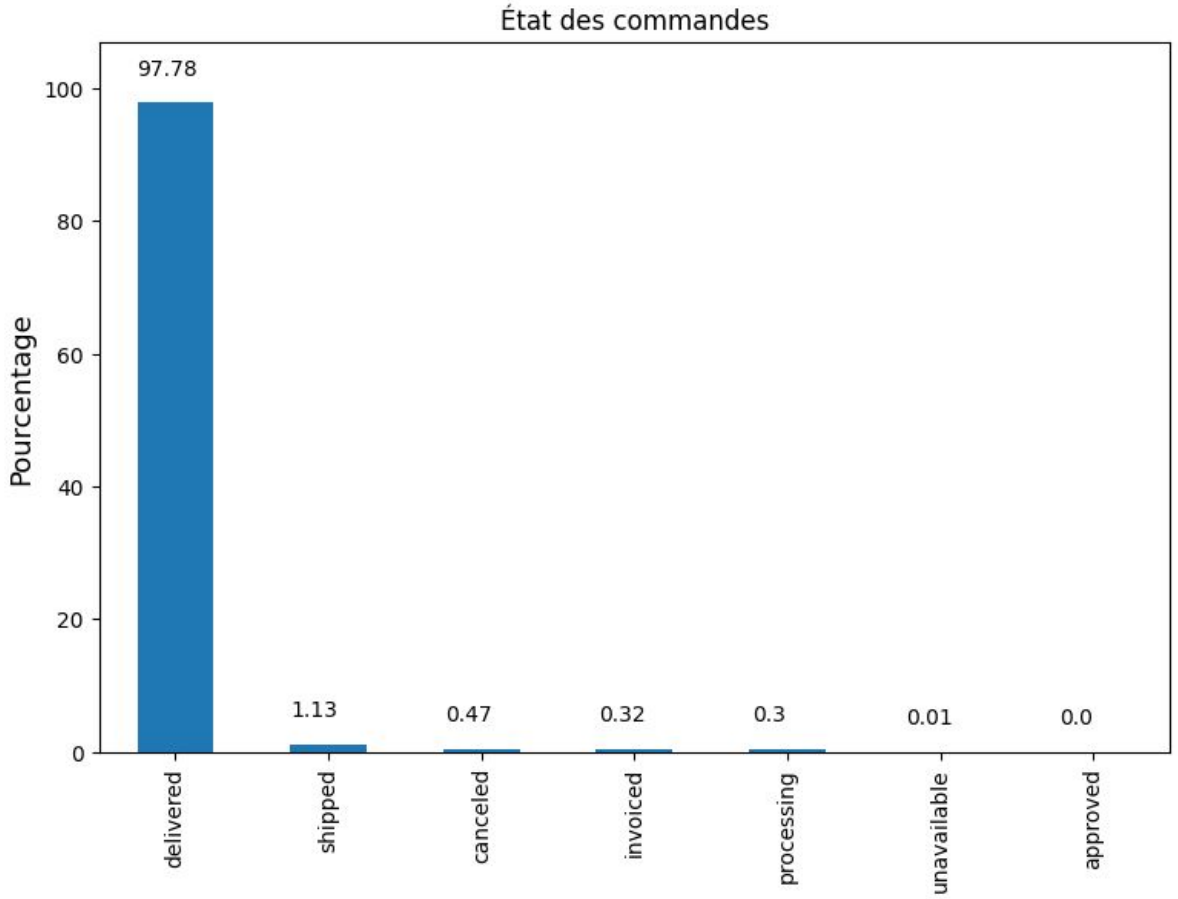


- **Données finales : 99.441 entrées de commandes uniques**
- **96.096 clients uniques**

- **19 variables sélectionnées dans un premier temps**

Description des données

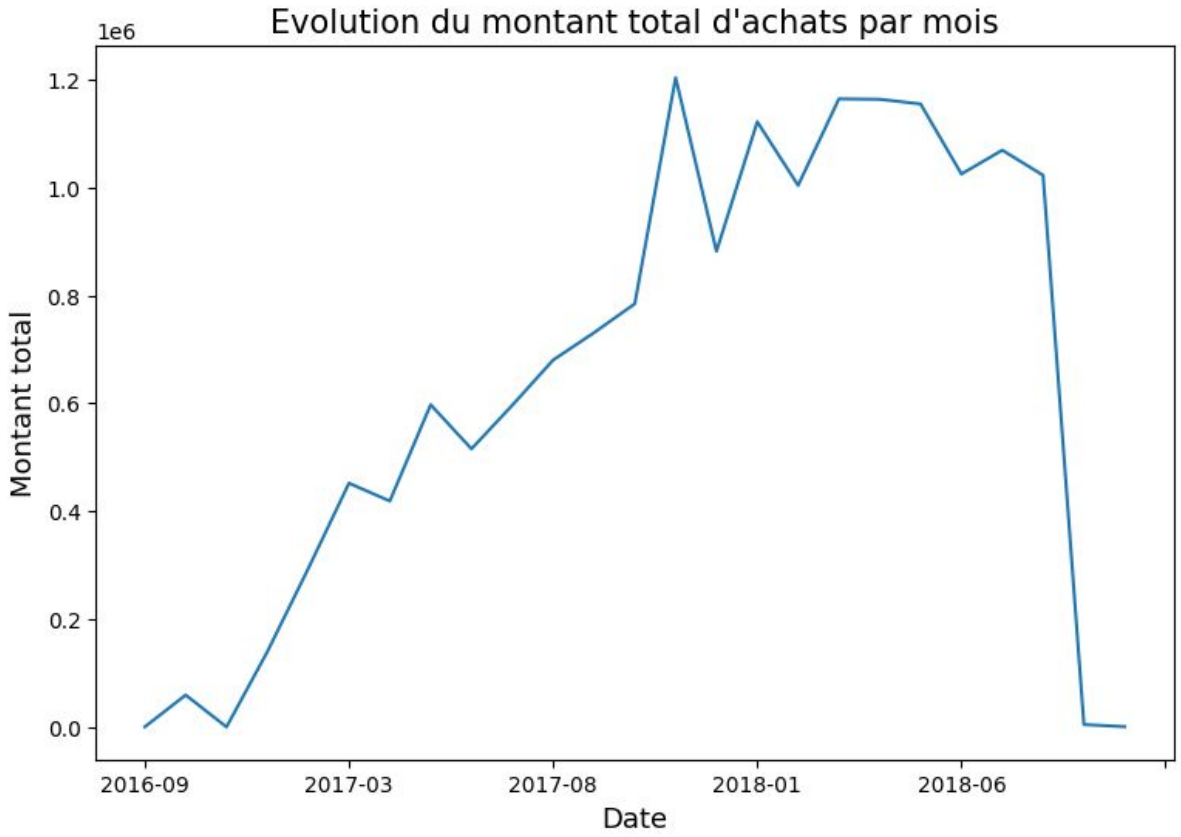
- Etat des commandes :
  - Presque 98% des commandes ont le statut “delivered”.
- On pourra travailler sur ces commandes uniquement :
  - On ne supprimera que peu de données
  - Les commandes annulées ou encore en cours n’auront pas les informations de satisfaction des clients



Description des données

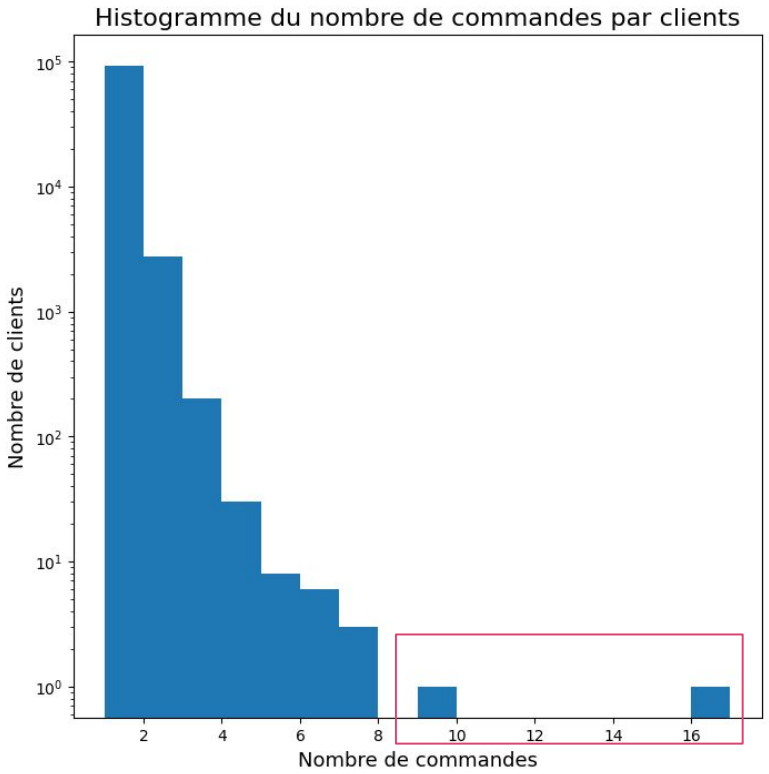
- Etat des commandes :
  - Presque 98% des commandes ont le statut “delivered”.
- On pourra travailler sur ces commandes uniquement :
  - On ne supprimera que peu de données
  - Les commandes annulées ou encore en cours n’auront pas les informations de satisfaction des clients

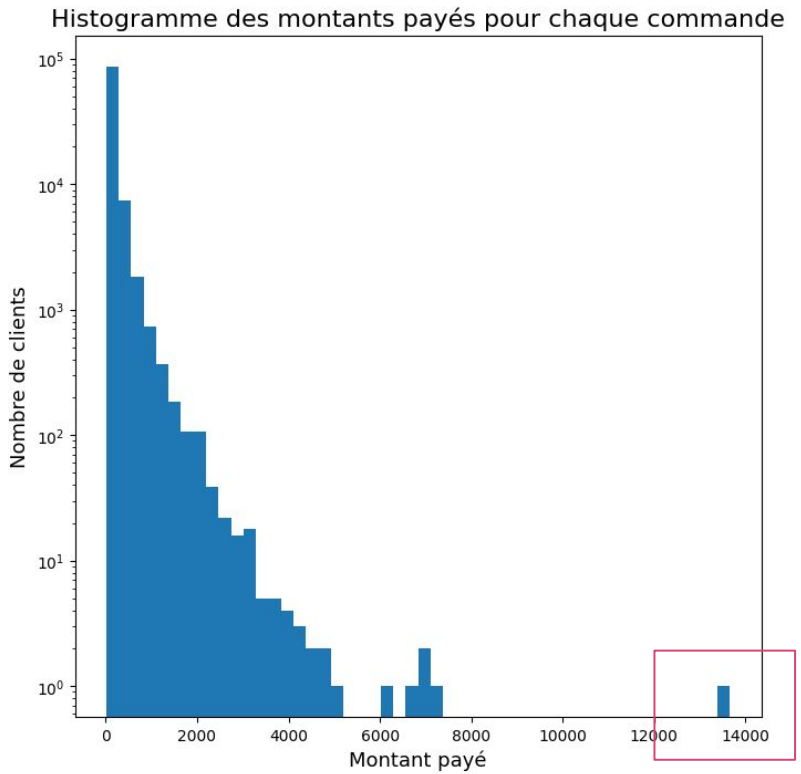
- Montant total d’achat par mois en hausse sur la période

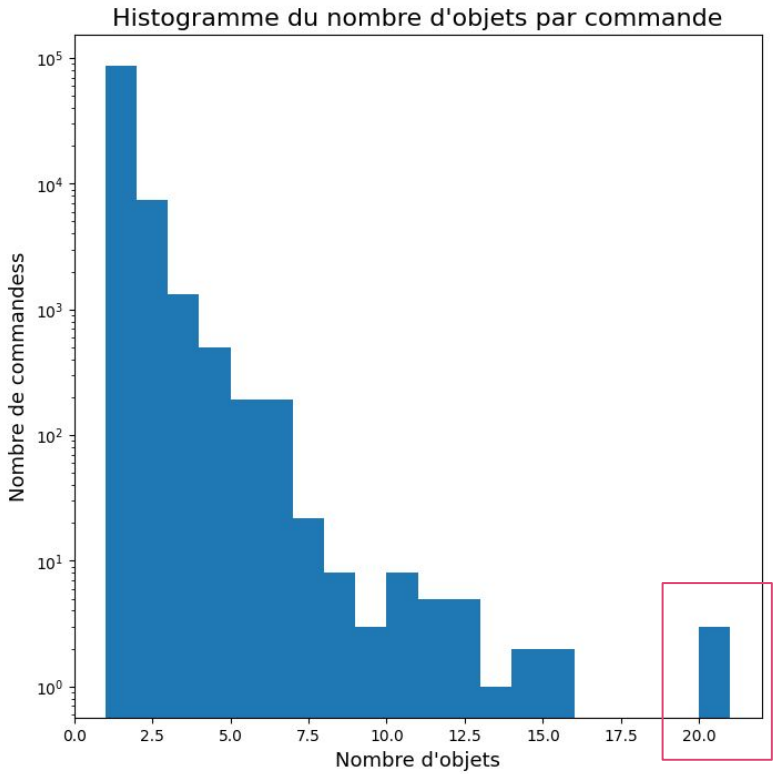


Description des données

- Quelques valeurs “extrêmes” sont présentes
  - Clients “importants” avec gros volumes de commandes, commandes très fréquentes
  - Données sont supprimées de l’analyse : Risque de biais + clients déjà fidélisés

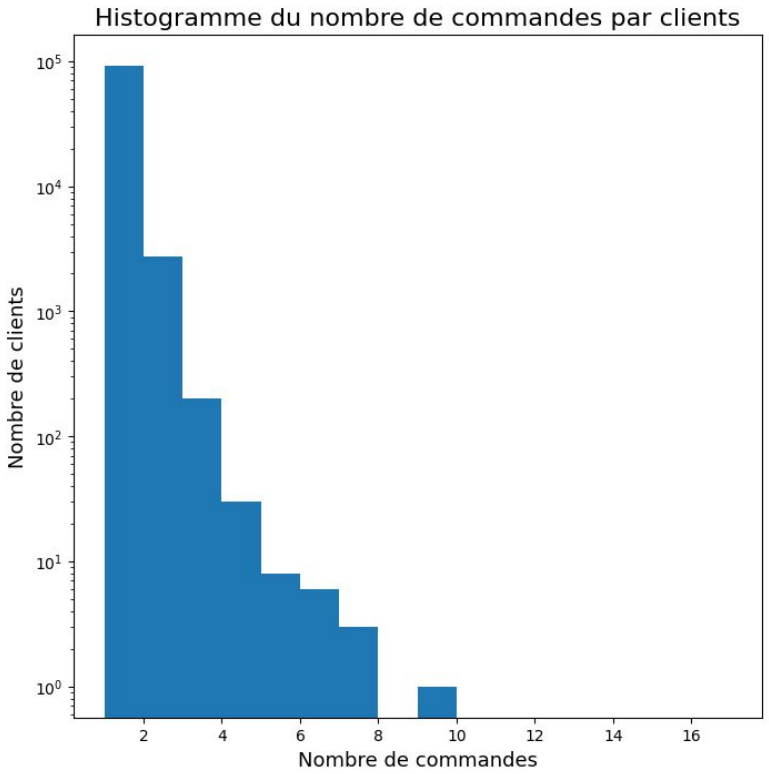


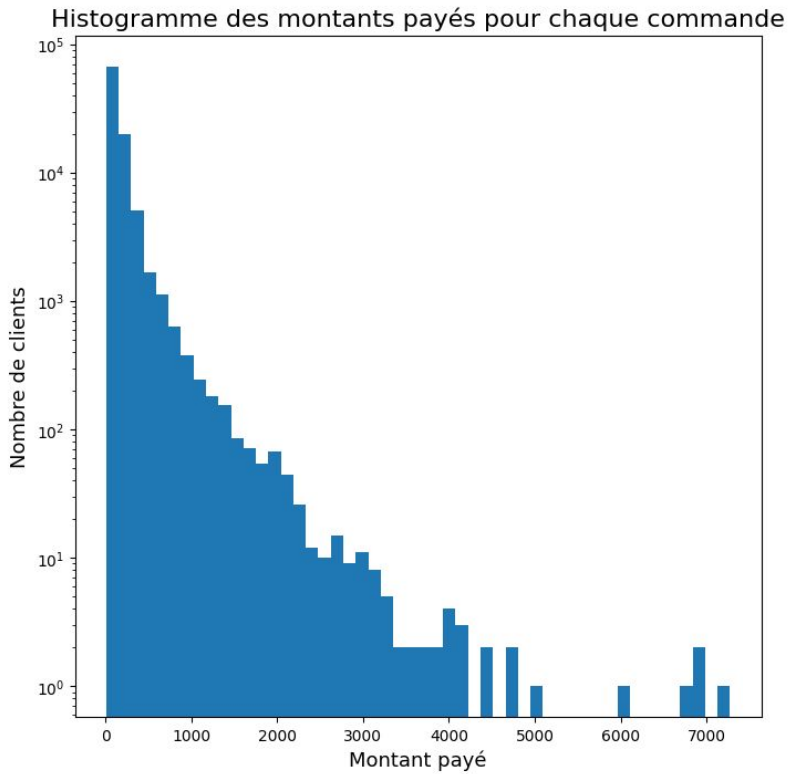


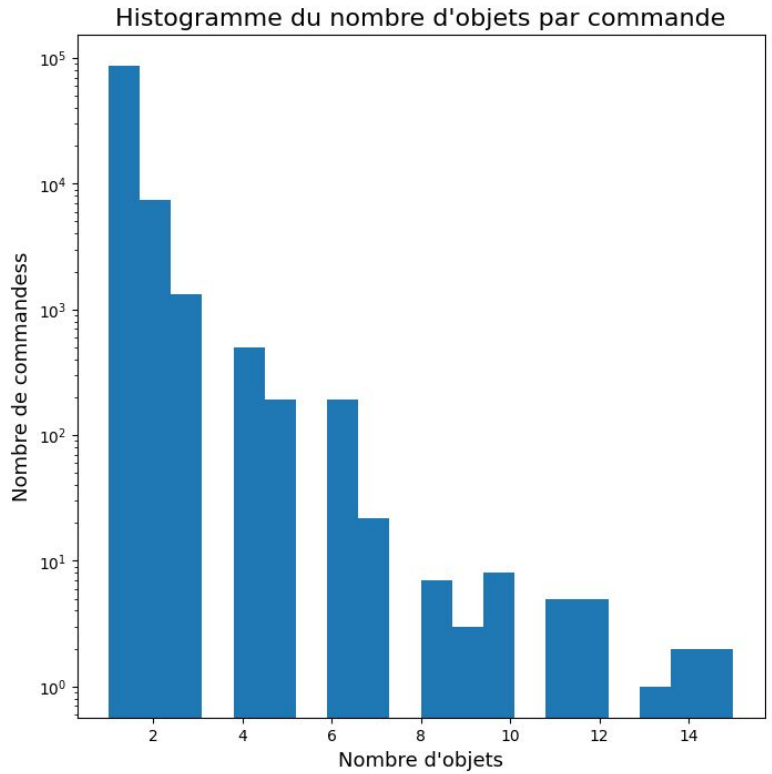


Description des données

- Quelques valeurs “extrêmes” sont présentes
  - Clients “importants” avec gros volumes de commandes, commandes très fréquentes
  - Données sont supprimées de l’analyse : Risque de biais + clients déjà fidélisés

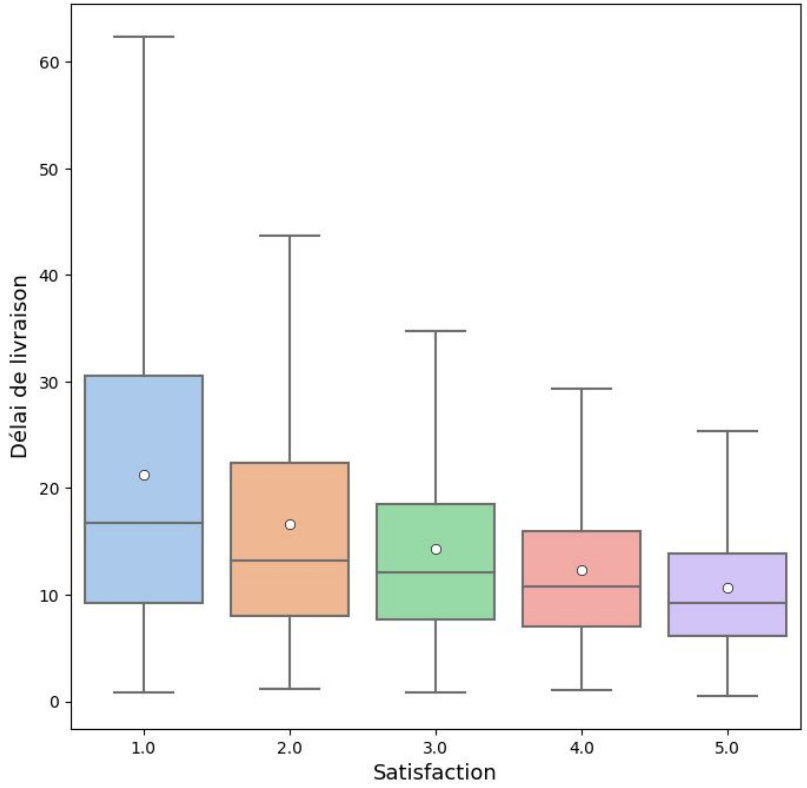


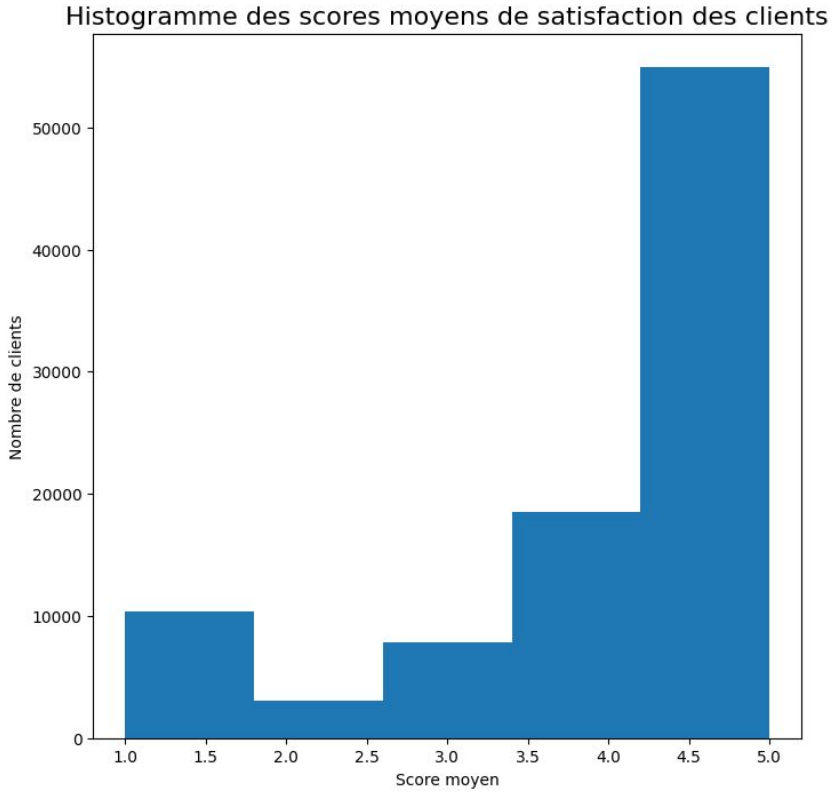




Satisfaction des clients

- Semble corrélée au délai de livraison
- Bonne satisfaction générale des clients





Méthode d'analyse

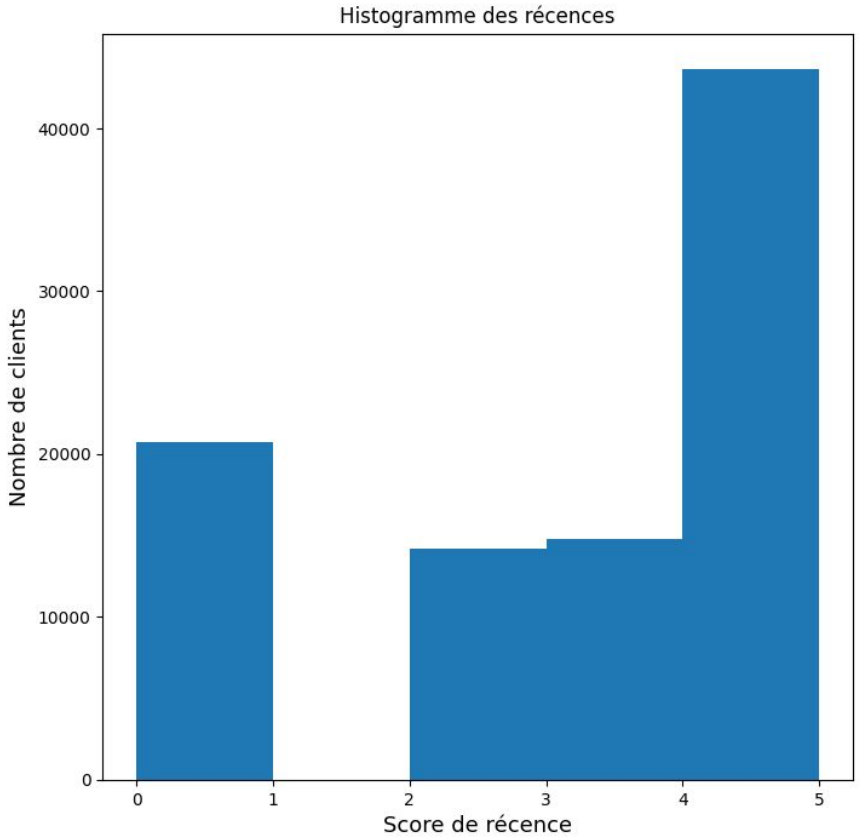
- Analyse RFM (Récence, Fréquence, Montant)
  - Type de segmentation de clients plutôt courant, permet de visualiser les comportements des clients (sont-ils plutôt des utilisateurs ponctuels ? Gros acheteurs ? …)
  - Identification des profils de clients en fonction de ces 3 critères + leur satisfaction (évaluée à partir de la notation moyenne laissée sur le site)
    - Un client avec un haut score de récence et de fréquence sera un client déjà bien fidélisé
    - A l'inverse, un client peu investi, avec un score de fréquence et/ou de récence faible devra être ciblé par les offres définies par le marketing
    - La proportion de ces clients est également une information importante : Si la grande majorité des clients sont déjà fidélisés, il faudra se concentrer sur la recherche de nouveaux clients
- Définition des périodes :
  - Données réparties sur 2 ans : Intervalles de 4 mois

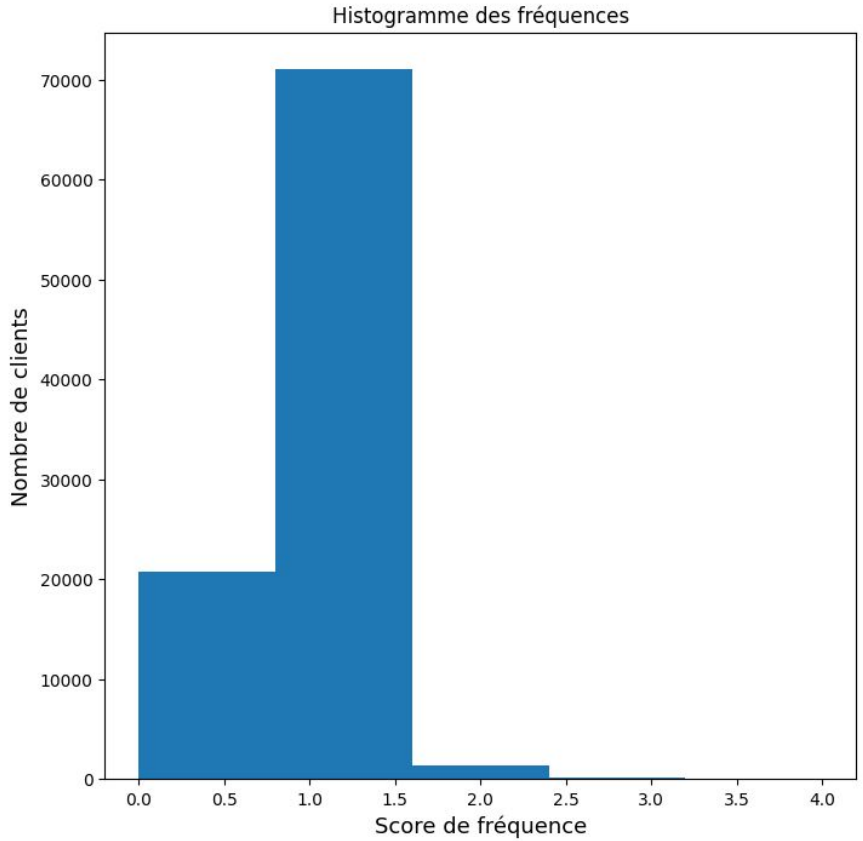
ID unique client	N (période la + récente)	N-1	N-2	N-3
1	0	0	1	1
2	1	1	1	1
3	1	0	0	0

ID unique client	Récence	Fréquence
1	3	2
2	5	4
3	5	1

Score de récence et fréquences

- Beaucoup de clients arrivés récemment, mais peu de commandes fréquences
  - Données de départ : 3% des clients ont fait plusieurs commandes. Arrivée récente de clients ?





Segmentation : Quel algorithme ?

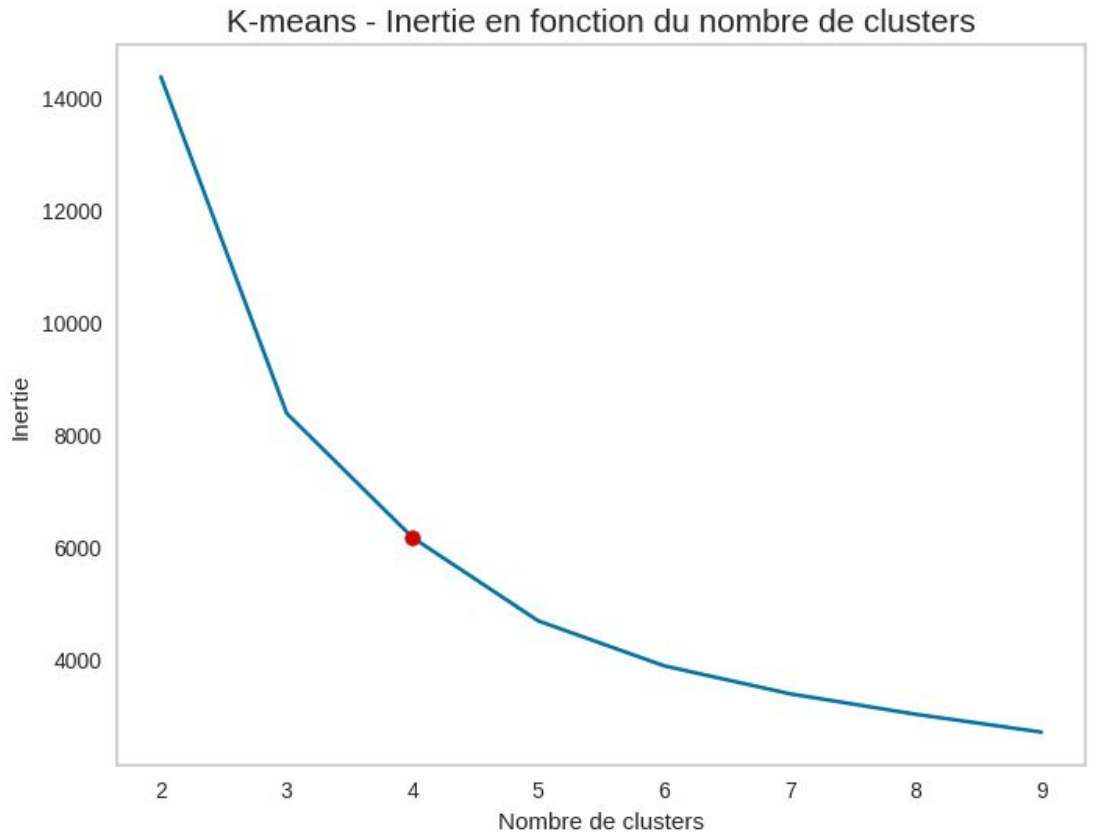
- Test de 3 algorithmes différents
  - DBScan
  - Kmeans
  - Agglomerative clustering
- Evaluation des algorithmes sur leurs performances : Kmeans est le plus efficace

k-means	2.22s
DBscan (knn)	2.50s
DBscan (manuel)	12.69s
Agg. clustering	16.24

De plus : DBScan est difficile à paramétrer (pas de choix du nombre de clusters; le paramétrage avec la méthode Knn donne plusieurs dizaines de clusters et le paramétrage ‘manuel’ est trop long

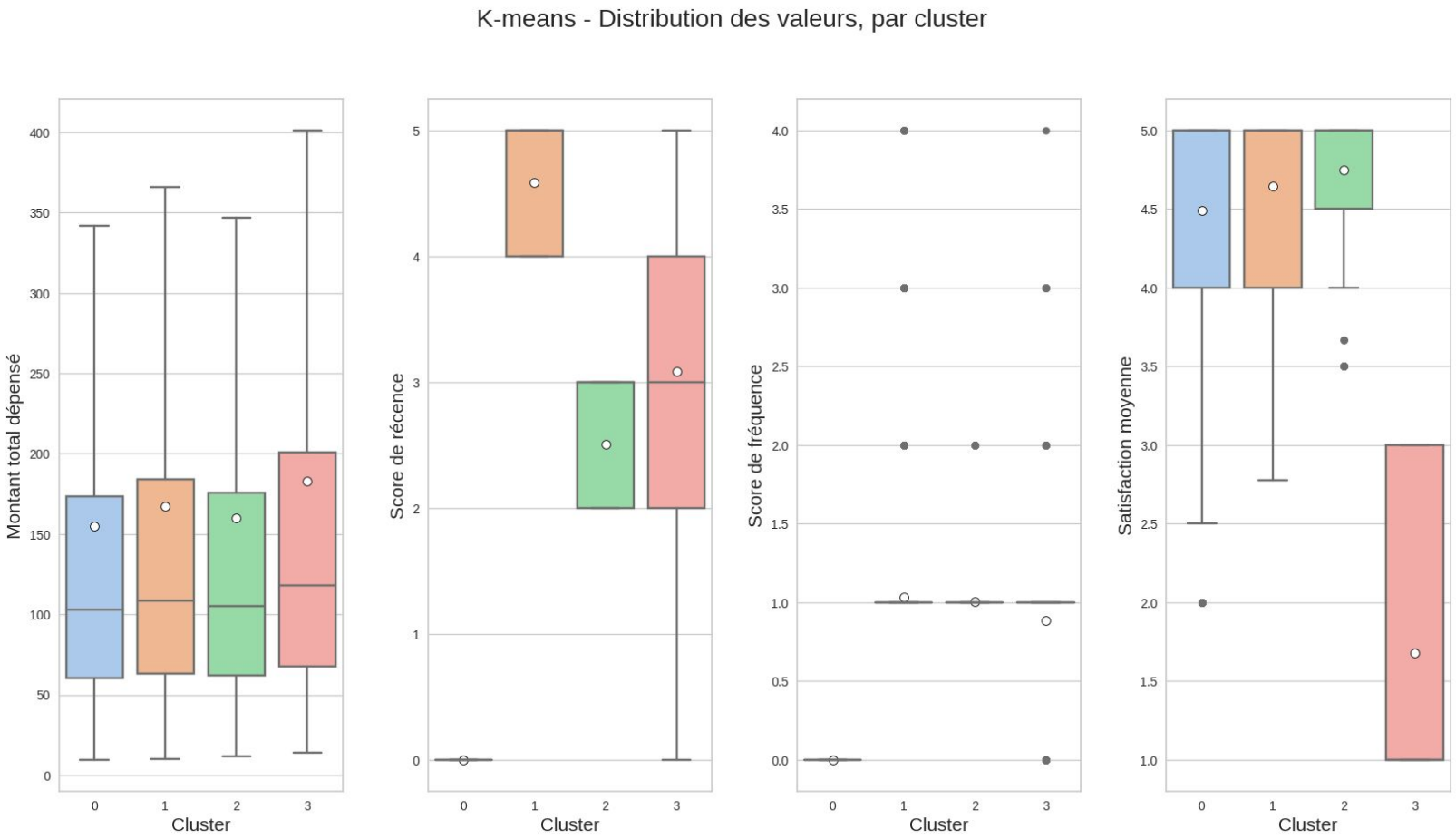
K-means - étude sur toutes les données

- Choix du nombre de clusters : “Elbow method”
  - Compromis entre le coût de calcul / la minimisation de la variance dans les clusters



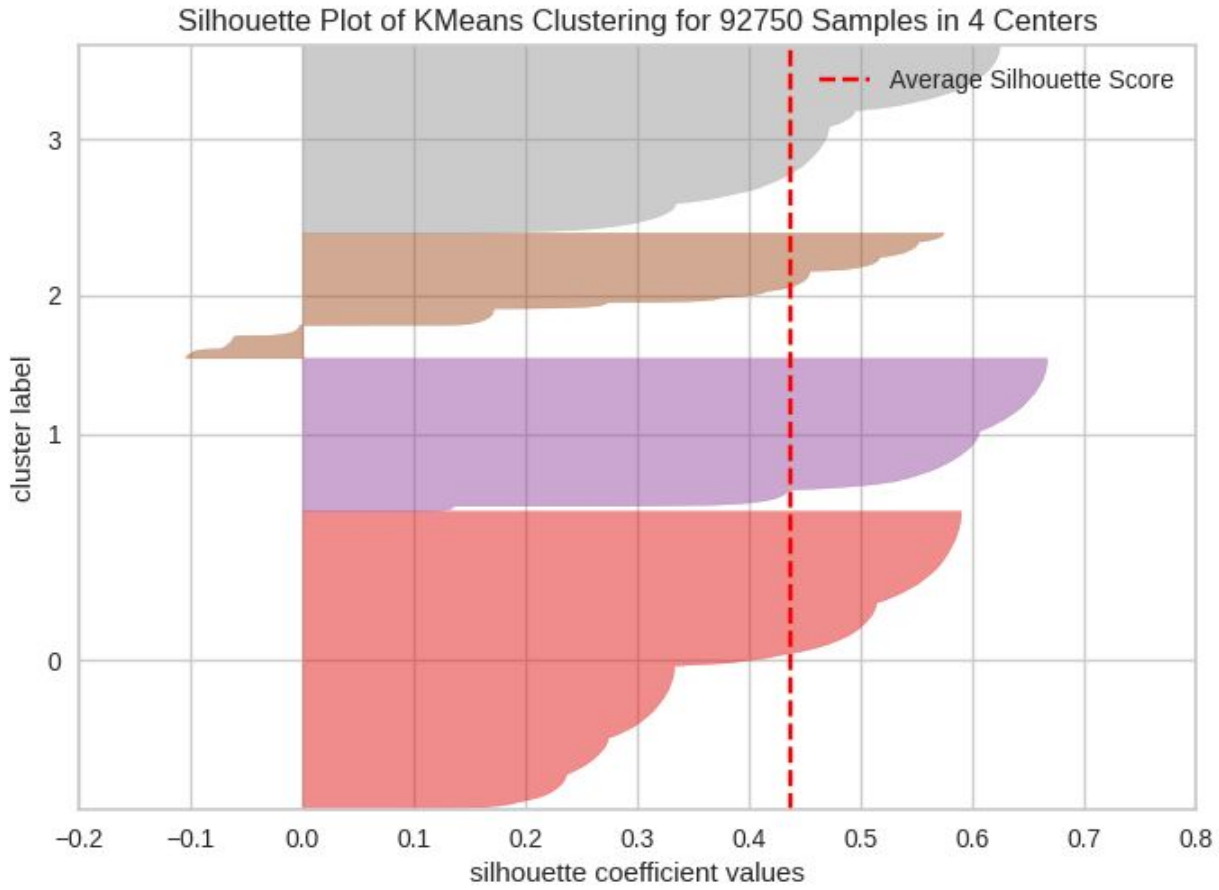
K-means - étude sur toutes les données

- 4 clusters
  - Cluster 0 : Clients plutôt satisfaits mais commandes anciennes
  - Cluster 1 et 2 : Bon clients (récence et fréquence), satisfaits. Clients du cluster 2 n'ont pas commandé depuis quelques temps
  - Cluster 3 : Clients les moins satisfaits
- Le montant dépensé n'est pas le facteur discriminant !



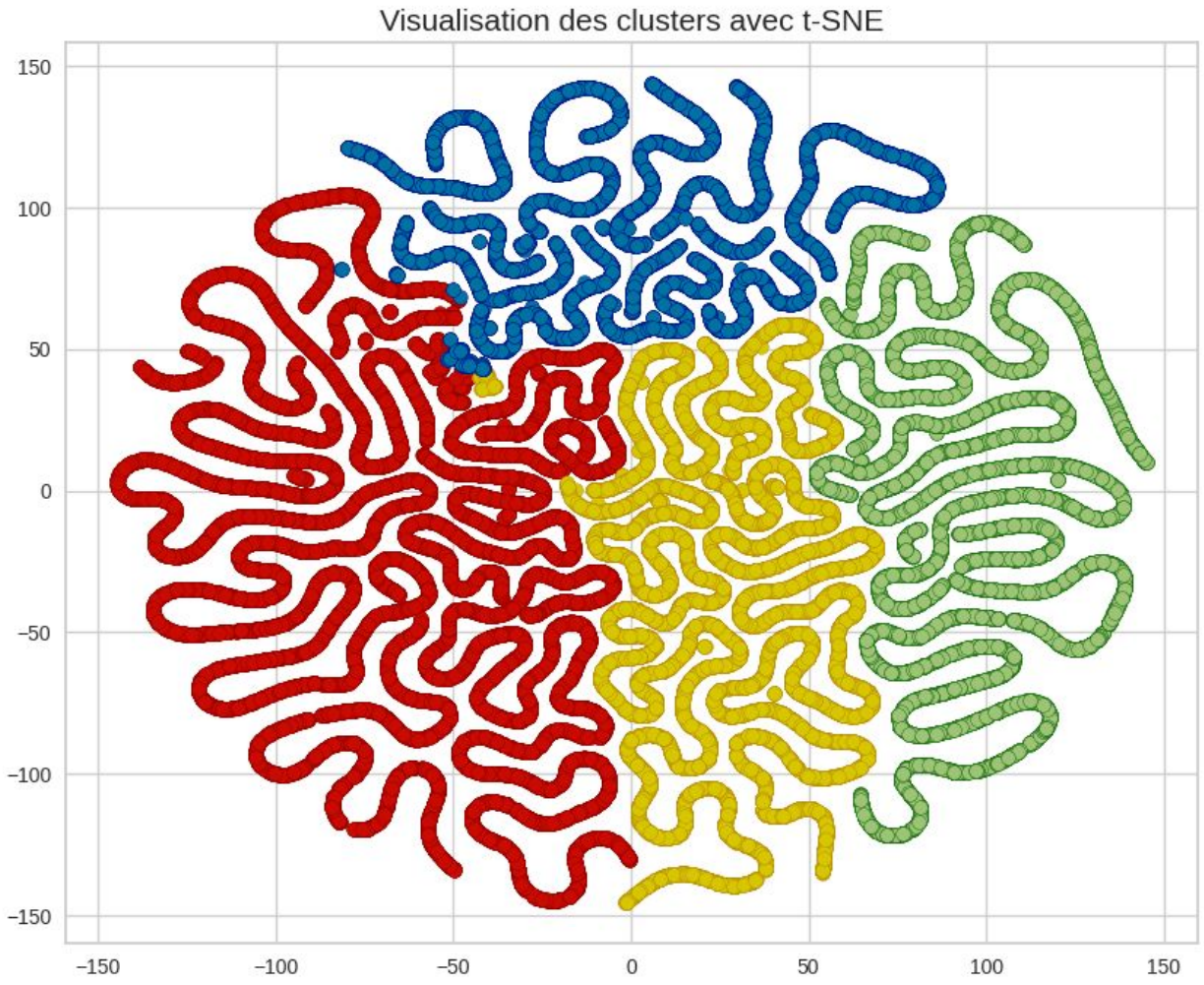
Kmeans - Qualité des clusters

- Score de silhouette : Peu de points mal classifiés.
- Le cluster 0 est le plus important. Travail à faire pour ramener ces clients vers le site ?
- Cluster 3, clients peu satisfaits : Assez important également.



Kmeans - Qualité des clusters

- Méthode de représentation des données : t-SNE
  - Méthode de réduction de dimension
  - Favorise la topologie locale : les points proches / éloignés le resteront dans l'espace en 2 dimensions
- Scores de récence / fréquences : valeurs entières, d'où l'aspect 'linéraire'
- Les clusters sont bien séparés



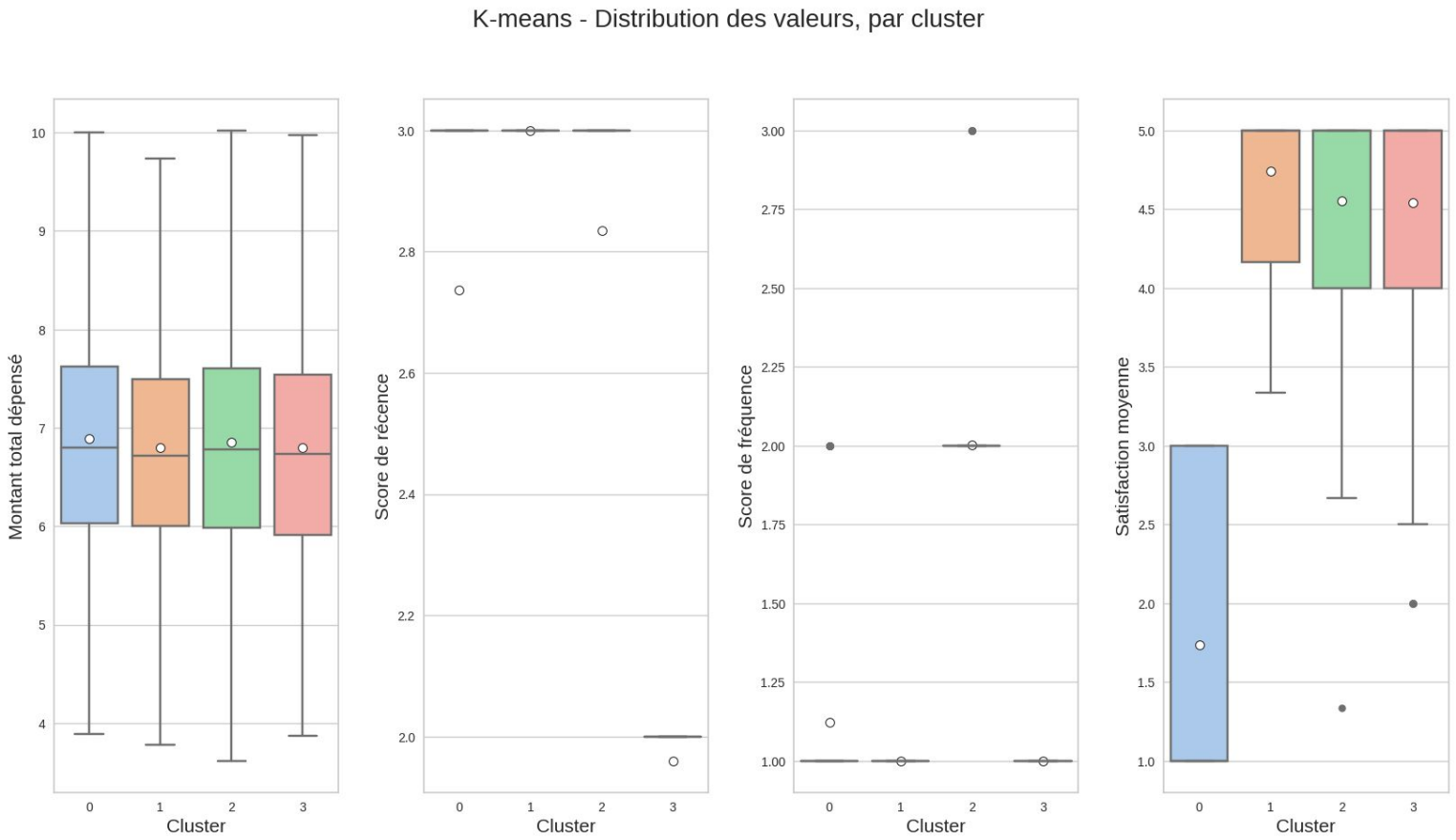
Proposition d’un contrat de maintenance

- Utilisation de l’Adjusted Rand Index (ARI)
  - Métrique qui permet d’évaluer la similarité entre deux expériences de clustering. 1 signifie que les résultats sont identiques.

- Méthodologie :
  - Définition d’une période de référence : 1 an (Sept. 2016 à Sept. 2017)
  - Clustering sur cette période de référence
  - Clustering sur [période de ref. + période supplémentaire]
    - La période ajoutée aux données est incrémentée pour mesurer la variation de l’ARI au cours du temps
  - Calcul de l’ARI entre les deux résultats
  - Détermination de la fréquence de maintenance

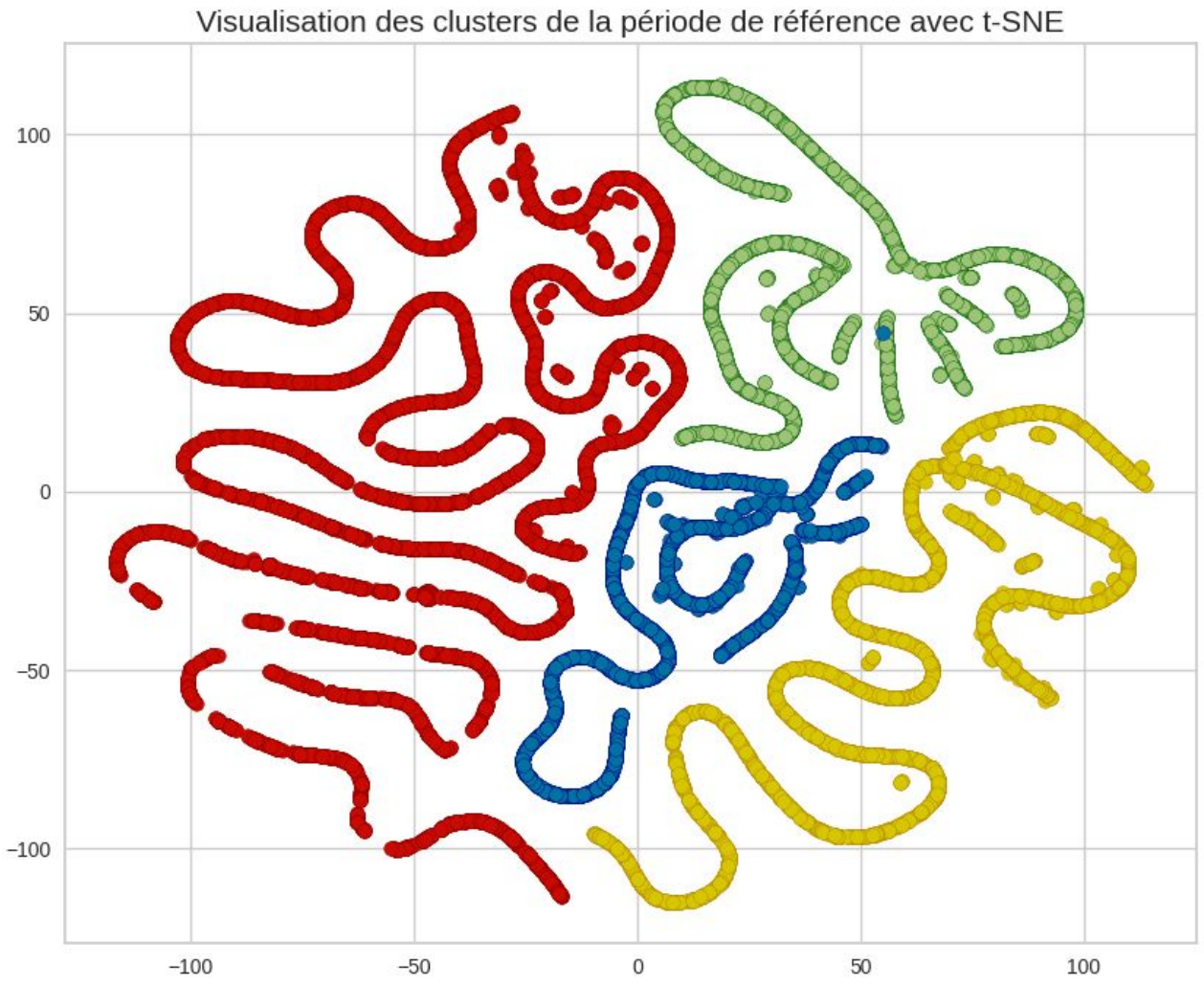
Clusters sur la période de référence

- Clusters similaires :
  - Cluster 3, clients satisfaits mais commandes anciennes
  - Cluster 0, clients peu satisfaits
- Peu d'influence du montant dépensé



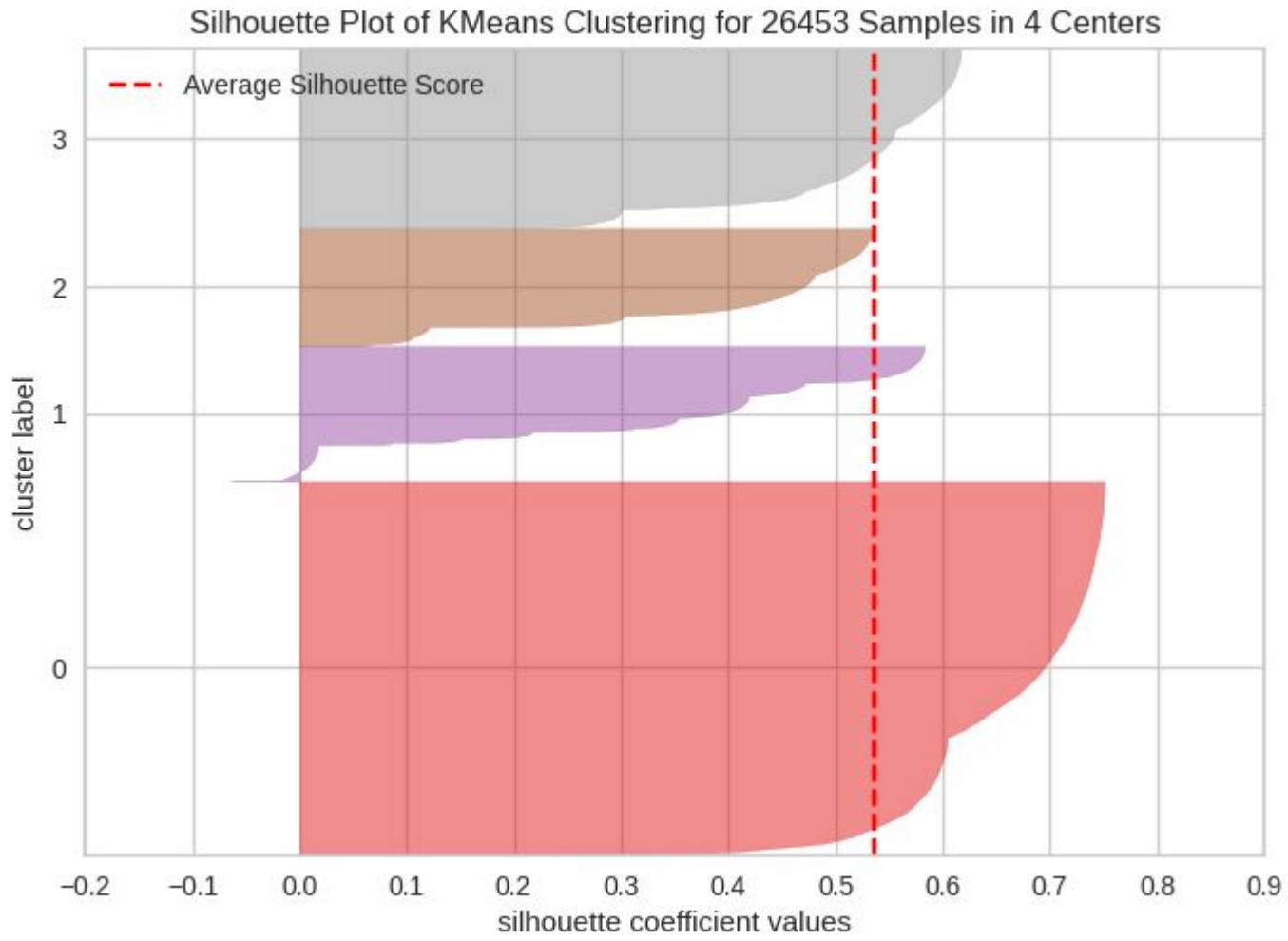
Clusters sur la période de référence

- Clusters similaires :
  - Cluster 3, clients satisfaits mais commandes anciennes
  - Cluster 0, clients peu satisfaits
- Peu d'influence du montant dépensé
- t-SNE : Clusters bien définis



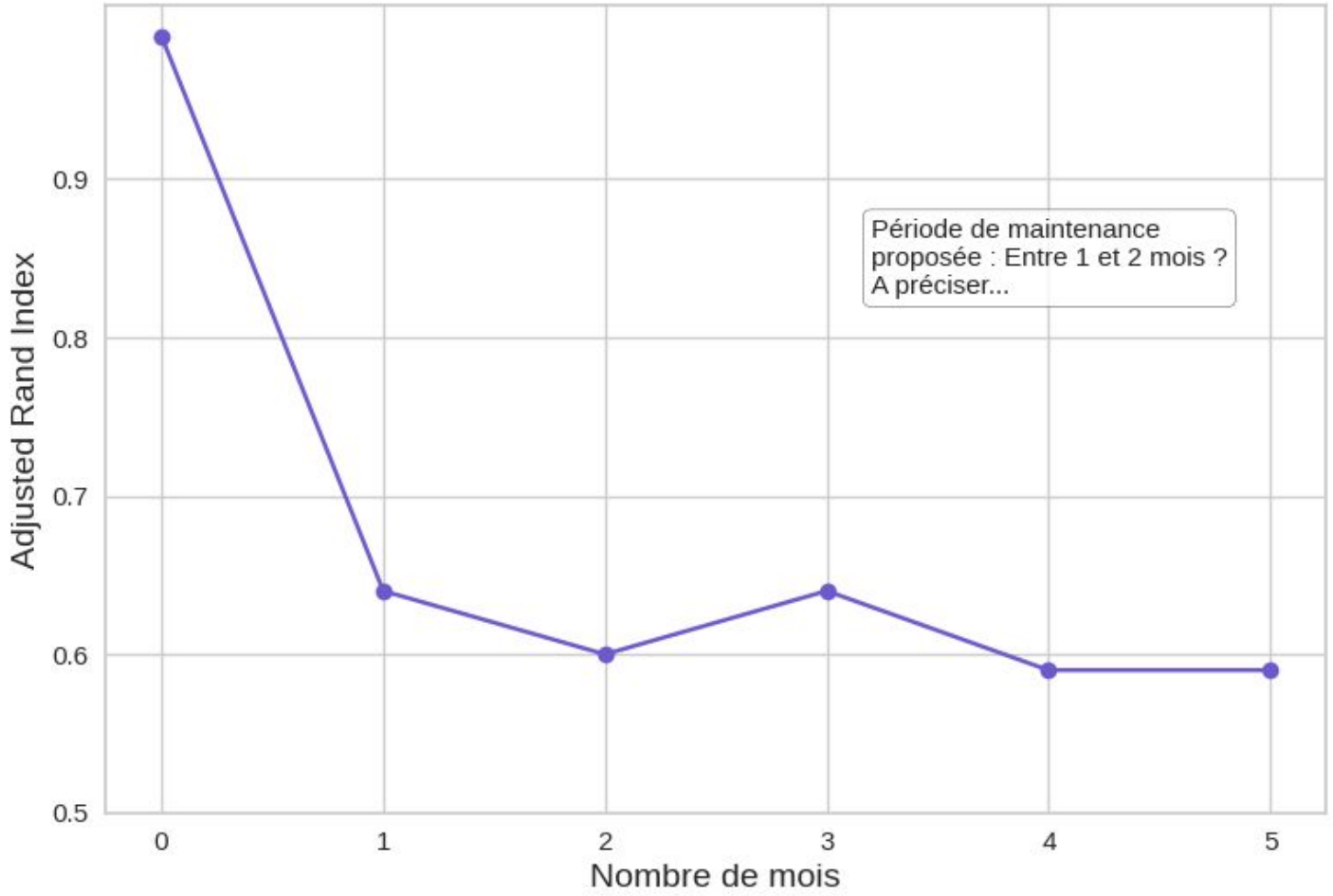
Clusters sur la période de référence

- Clusters similaires :
  - Cluster 3, clients satisfaits mais commandes anciennes
  - Cluster 0, clients peu satisfaits
- Peu d'influence du montant dépensé
- t-SNE : Clusters bien définis
- Score de silhouette : Proportions différentes : les clients au faible taux de satisfaction sont les plus nombreux



Evolution de l'ARI

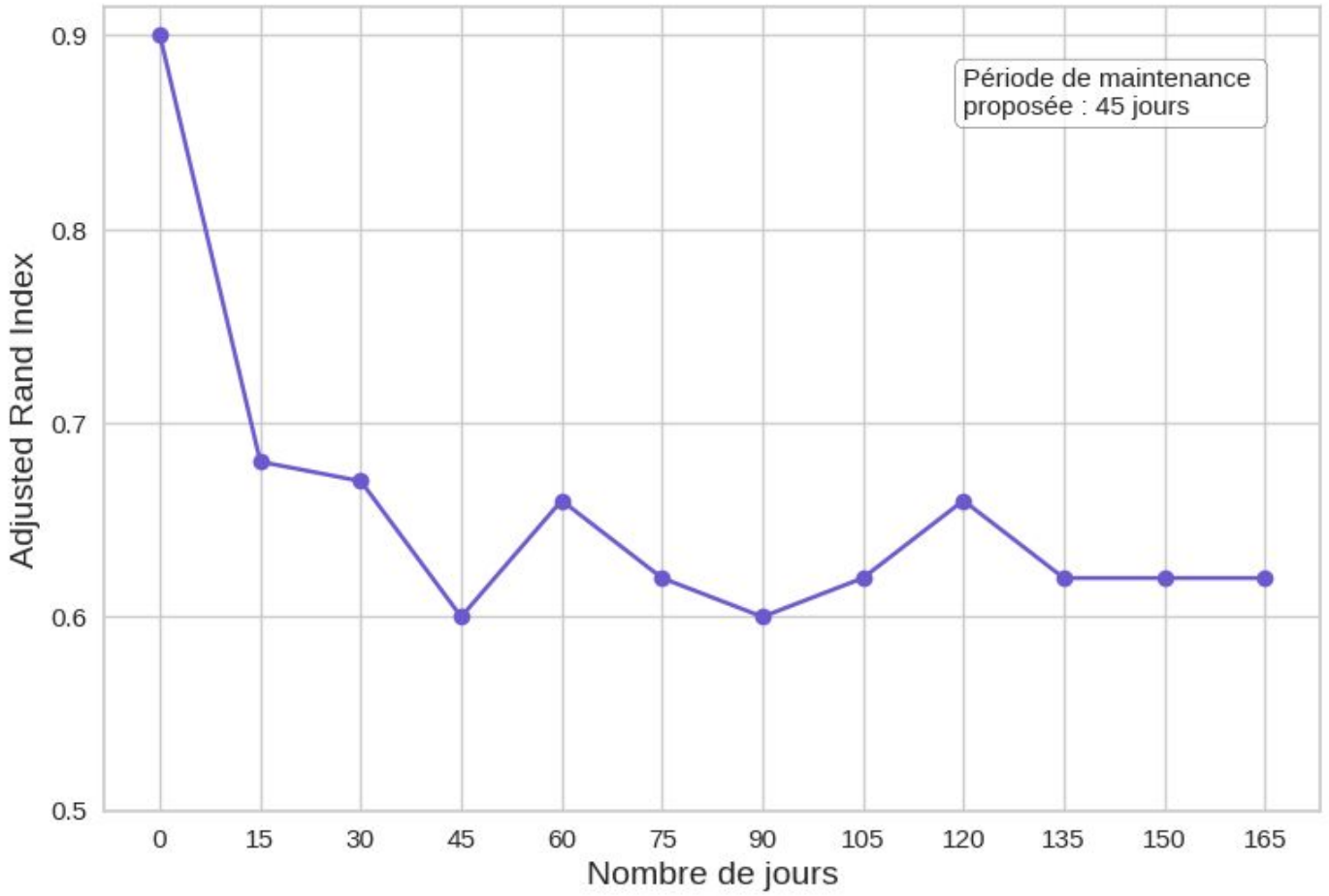
- Diminution très rapide du score
- En appliquant une limite à 0.6, il faudrait une période de maintenance aux alentours de 2 mois



Evolution de l'ARI

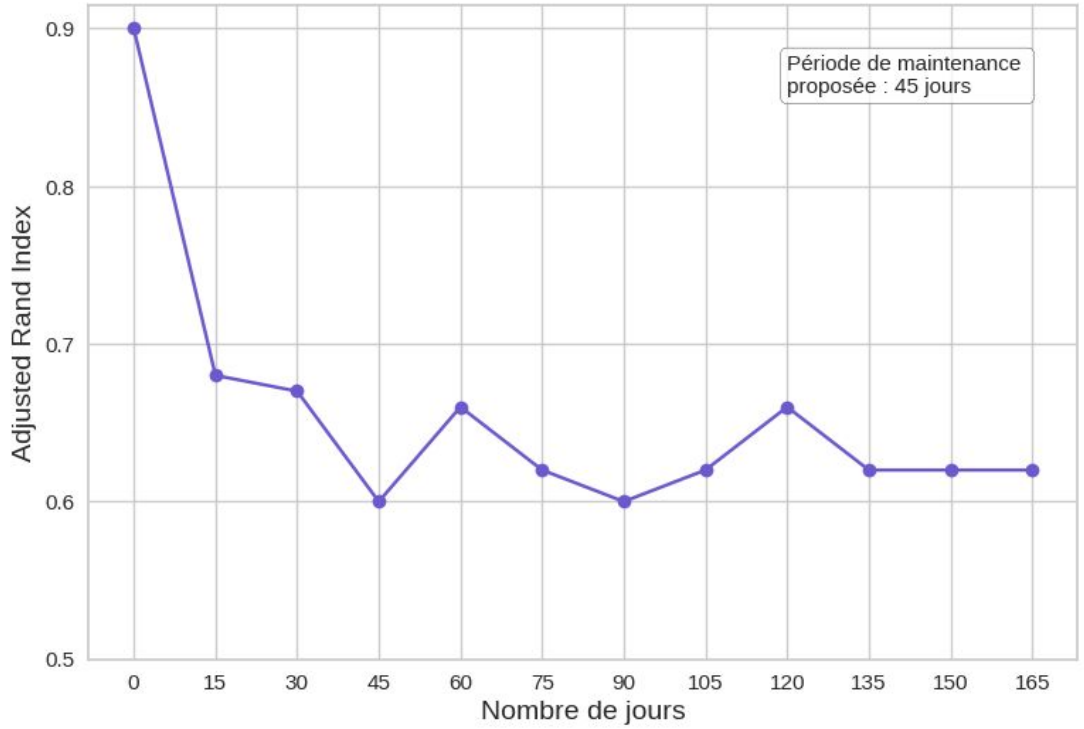
- Diminution très rapide du score
- En appliquant une limite à 0.6, il faudrait une période de maintenance aux alentours de 2 mois

- En recalculant le score sur des périodes de 15 jours : Possibilité de préciser un peu plus la période de maintenance proposée.



Période de maintenance : 45 jours

- Pourrait être réduite à 15 ou 30 jours; mais coût trop important ?
- Cette durée sera amenée à varier : Les efforts des équipes marketing devraient modifier les comportements des clients et en amener de nouveaux



Conclusions

- Problématique : Segmentation de clients pour le site d'e-commerce Olist et proposition d'une période de maintenance
- Algorithme choisi : kmeans. Efficace, coup de calcul modéré.
- Période de maintenance proposée : 45 jours dans un premier temps
- 4 types de clients ont été définis :
  - Clients peu satisfaits
  - Bons clients, satisfaits, mais scores de récence / fréquence parfois plus faibles.
  - Clients satisfaits qui n'ont pas commandé depuis longtemps.
- Pistes possibles pour les équipes marketing :
  - Fidéliser les clients avec les scores intermédiaires de récence / fréquence
  - Ramener les clients qui n'ont pas commandé depuis longtemps

Segmentation de clients pour le site d’e-commerce Olist

-

Merci pour votre attention

Open Classrooms parcours Data Science - projet 4

Camille Besançon