olist

Segmentation des clients d’un site de e-commerce

Sommaire

Contexte

Analyse exploratoire

Segmentation & maintenance

Conclusion & Perspectives

Qui est Olist?

Segmentation des clients => campagnes de communication ciblées

Comprendre les **différents types d'utilisateurs** grâce à leur comportement et à leurs données personnelles

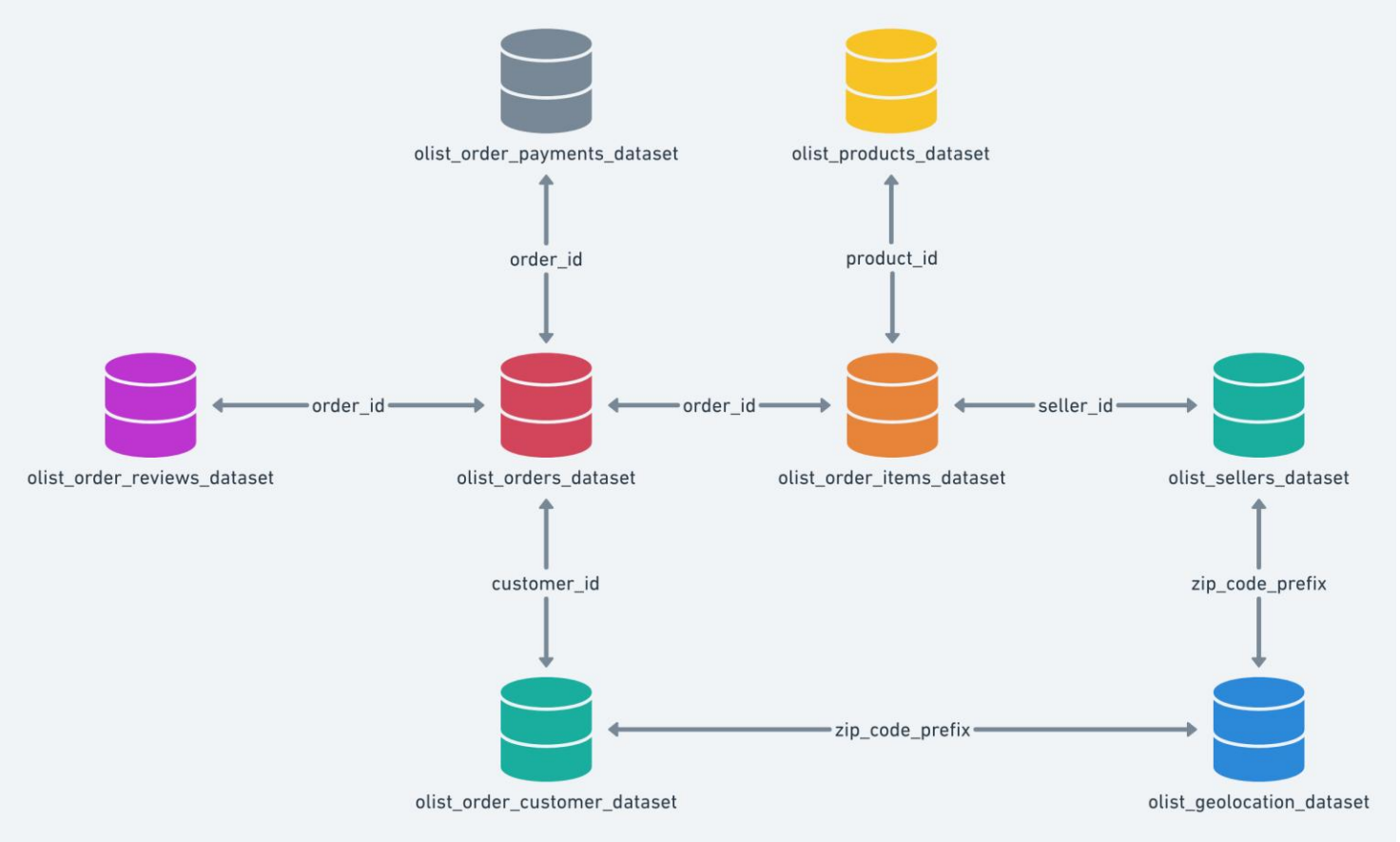
Différencier **les bons et moins bons clients** en termes de **commandes** et de **satisfaction**

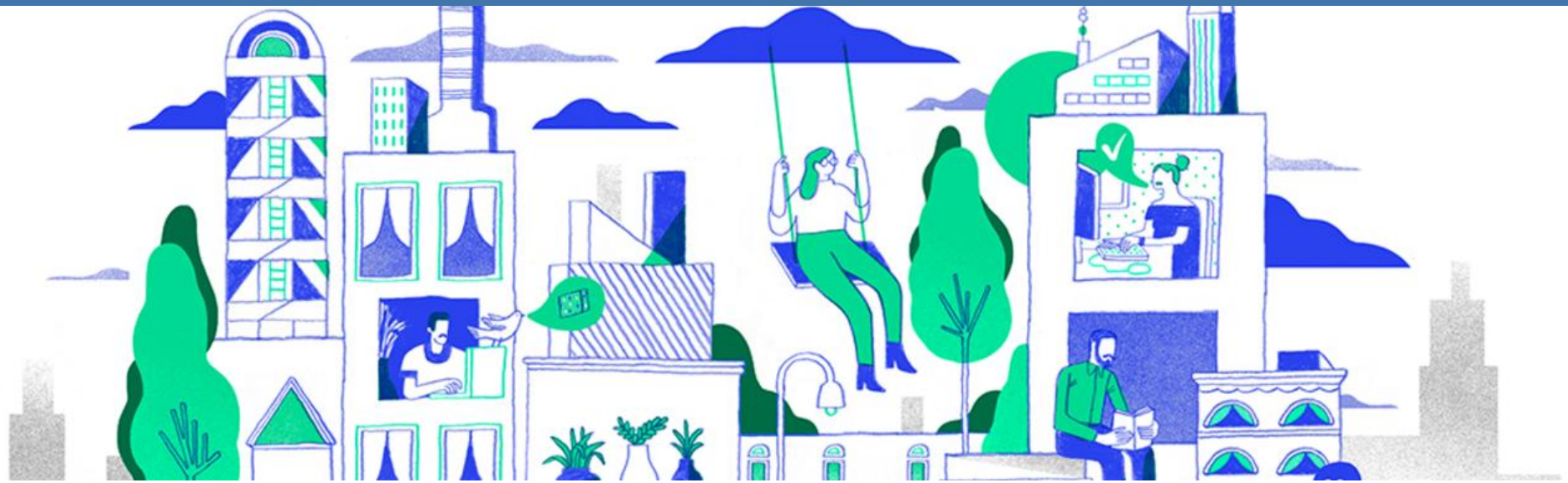
Fournir une **description actionnable** de la segmentation et de sa logique sous-jacente pour une utilisation optimale

Proposition de **contrat de maintenance :** analyse de la **stabilité des segments** au cours du temps

Les données disponibles

- Base de données SQL
- 9 tables
- Données anonymisées
- Commandes de 2016 à 2018





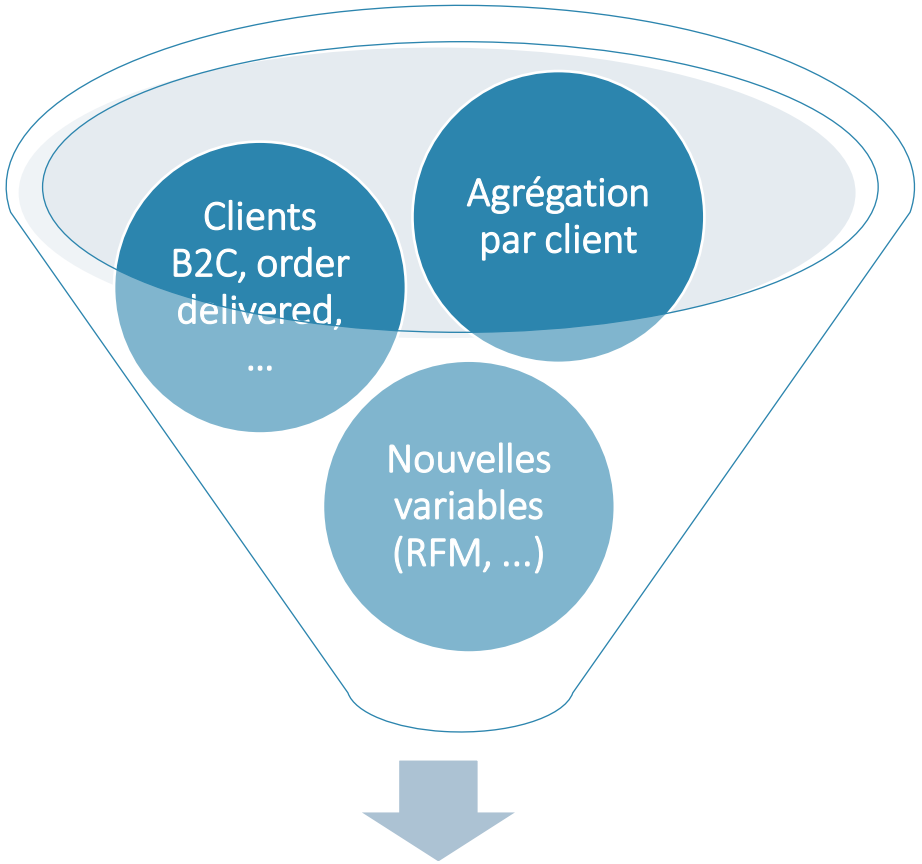
olist

empowering commerce

Analyse exploratoire

Nettoyage des données

96096 clients (99441 commandes), 51 variables



93042 clients (96203 commandes), 86 variables

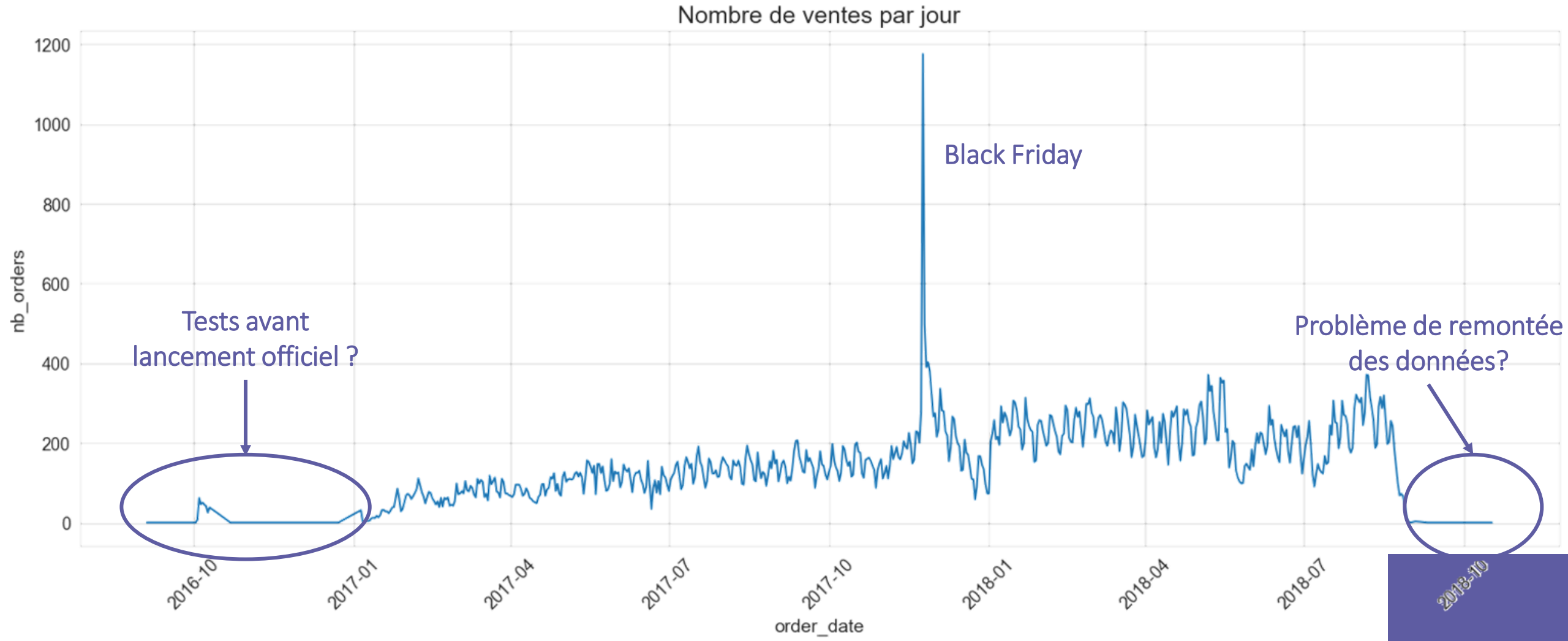
Filtres sur les données :

- order_status = 'delivered' (97%)
- Janvier 2017< Période retenue < Septembre 2018
- Ecarter les clients B2B (6 clients)

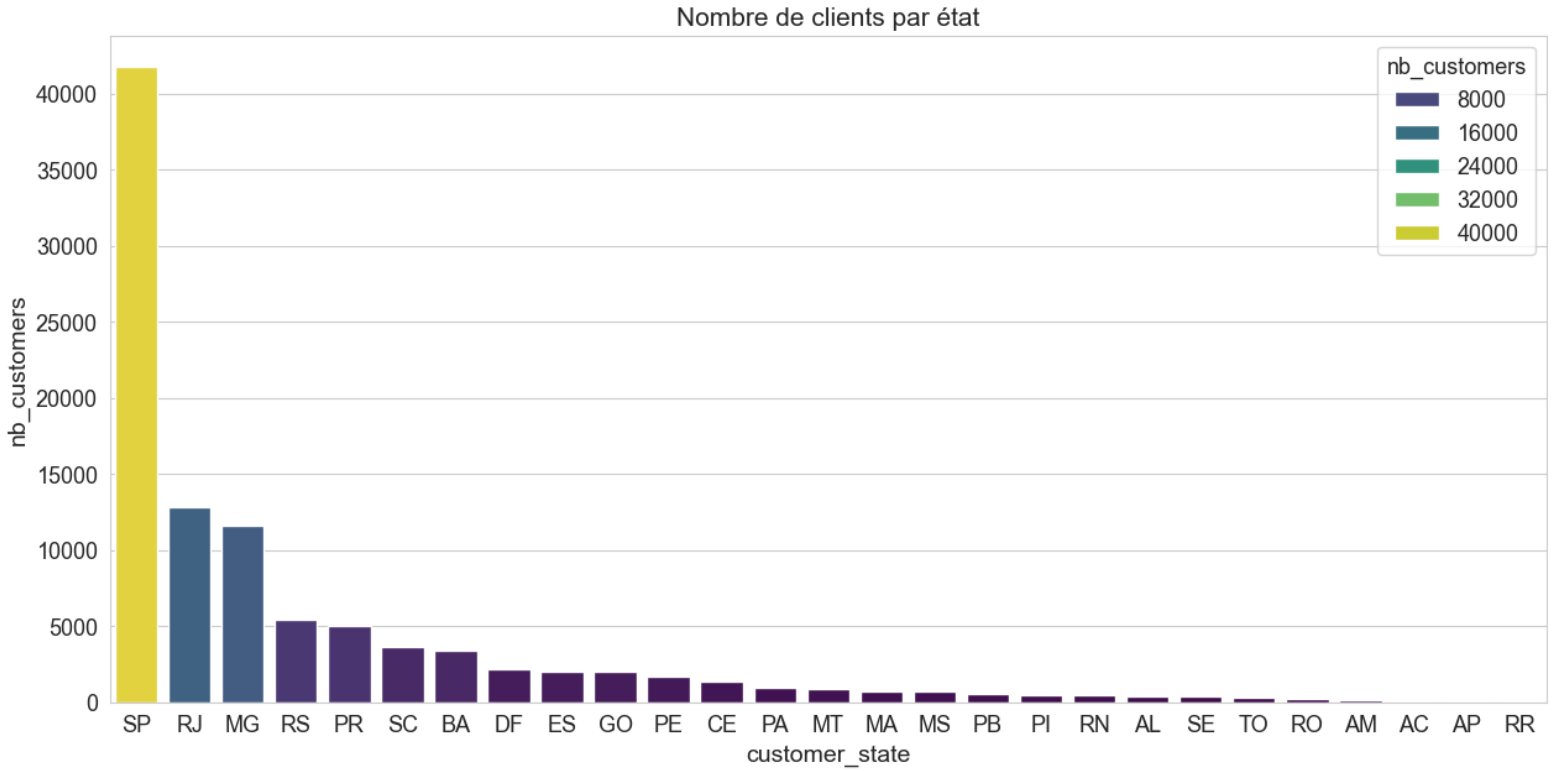
Nouvelles variables : agrégation par client

- RFM : Récence, Fréquence, Montant total
- Satisfaction clients
- Délai de livraison moyen
- Panier moyen : montant, nombre d'articles
- Nombre d'articles achetés par catégorie et total

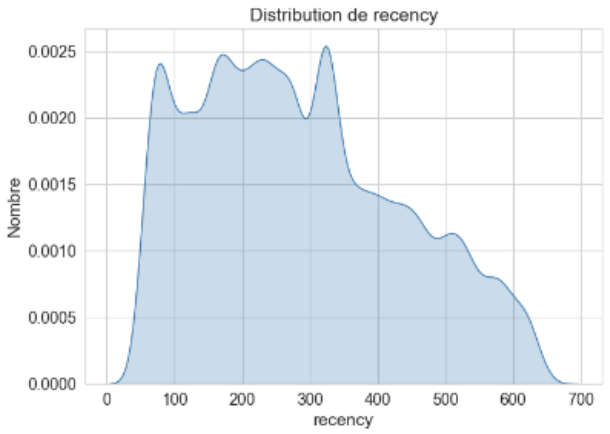
Pic de ventes au moment du Black Friday

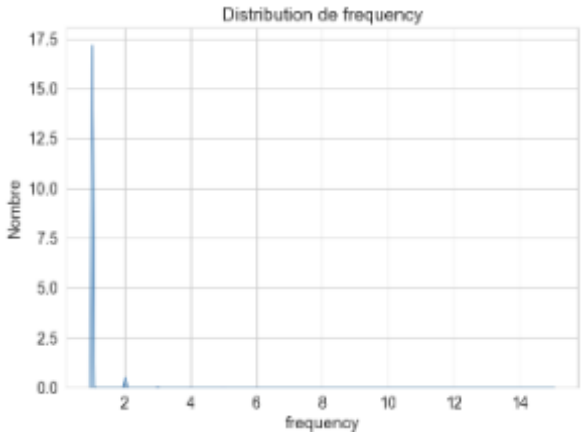


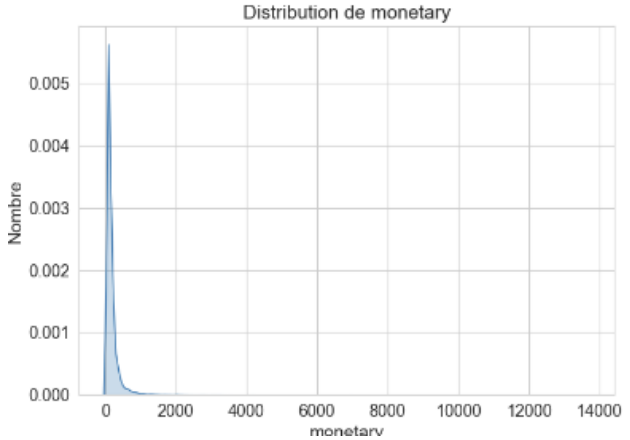
2 clients sur 5 habitent dans l’état dans Sao Paulo

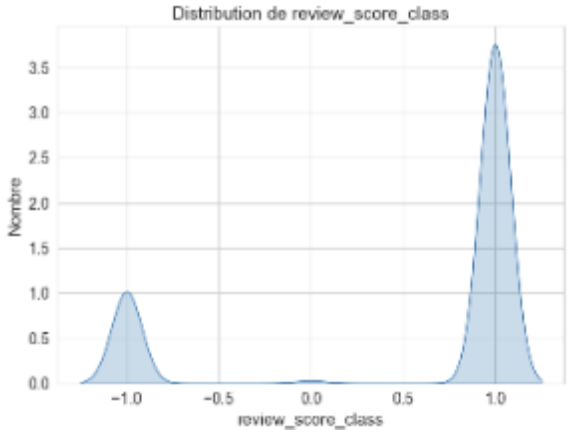


Seulement 3% des clients ont effectué plusieurs achats









Récence

1 client sur 2 a effectué un achat dans les 9 derniers mois.

Fréquence

3 clients sur 100 ont effectué plusieurs achats

3% des clients = 7% des commandes

Monétaire

3 clients sur 4 ont dépensé moins de 200 € au total sur le site

10% des clients ~ 38% du CA

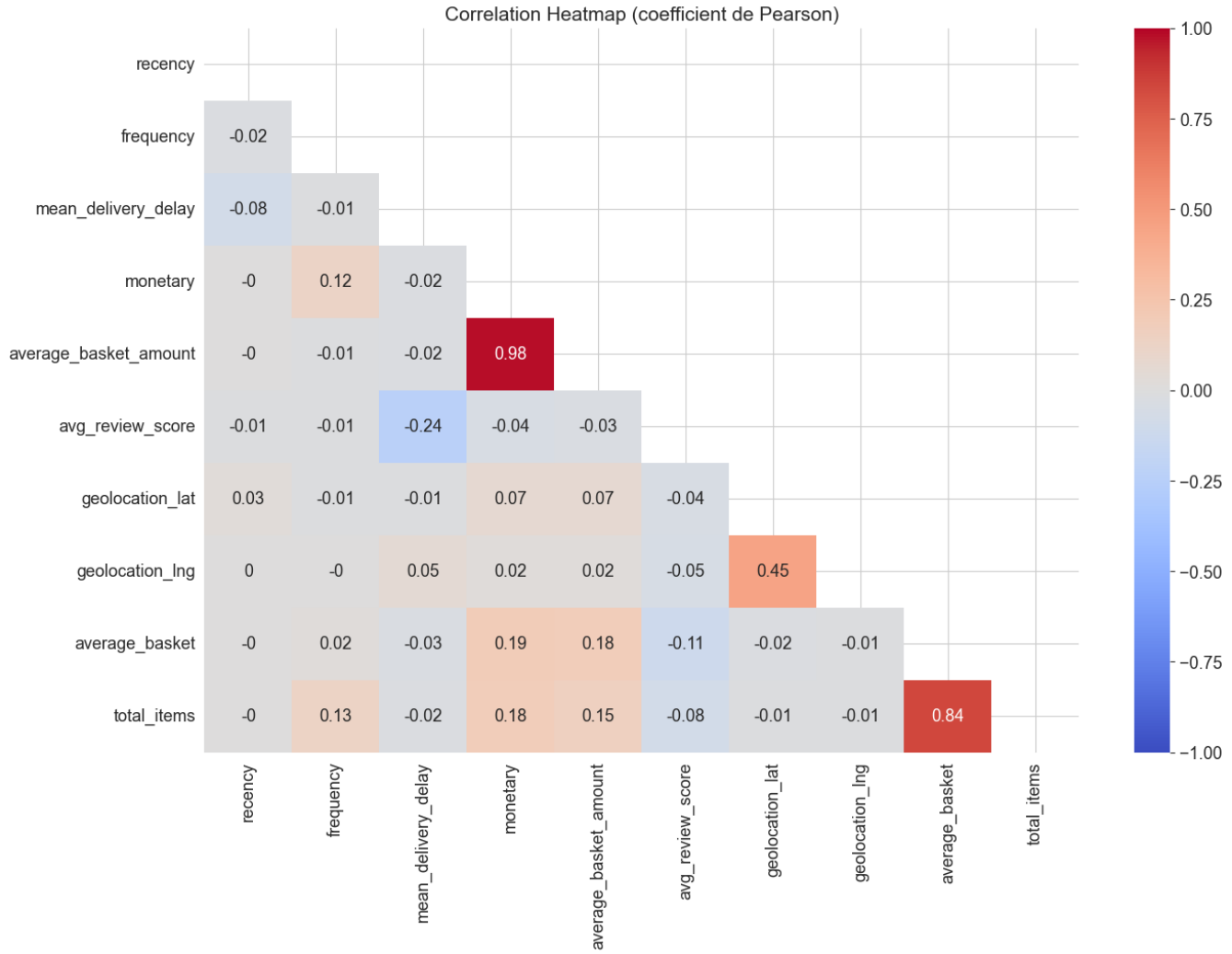
Satisfaction

8 clients sur 10 satisfaits après la livraison de leur commande

Corrélation entre les variables

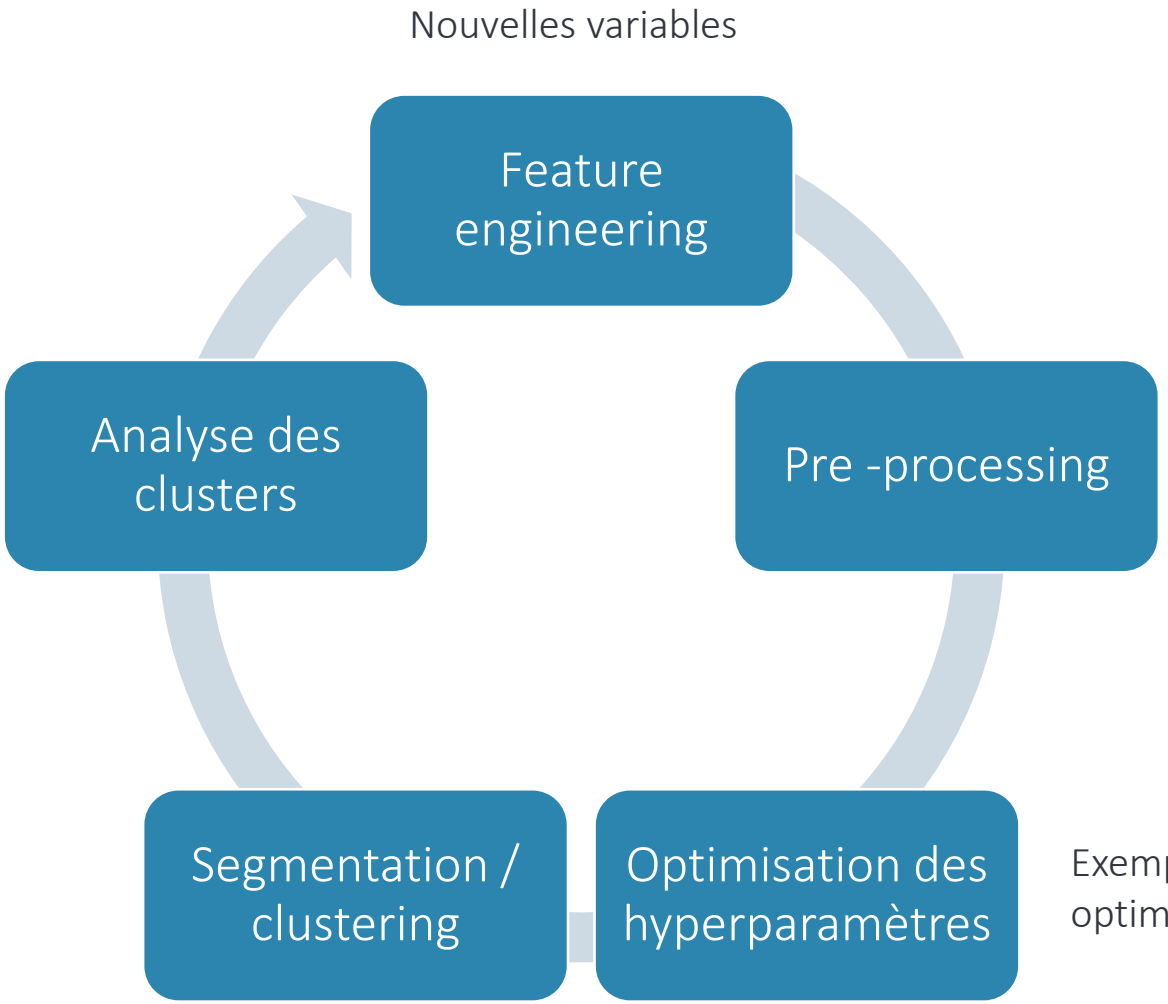
Peu de clients ont effectué plusieurs achats => corrélation entre:

- average_basket_amount et monetary
- average_basket et total_items



Segmentation des clients

La démarche



Standardisation : mise à l'échelle des données et centrage autour de 0
Encodage : données catégorielles -> données numériques

Exemple : Nombre optimal de clusters k

Les pistes explorées

Baseline

RFM Marketing

Variables RFM

- **K-Means**
- **CAH**
- **DBScan**

Méthodes non supervisées

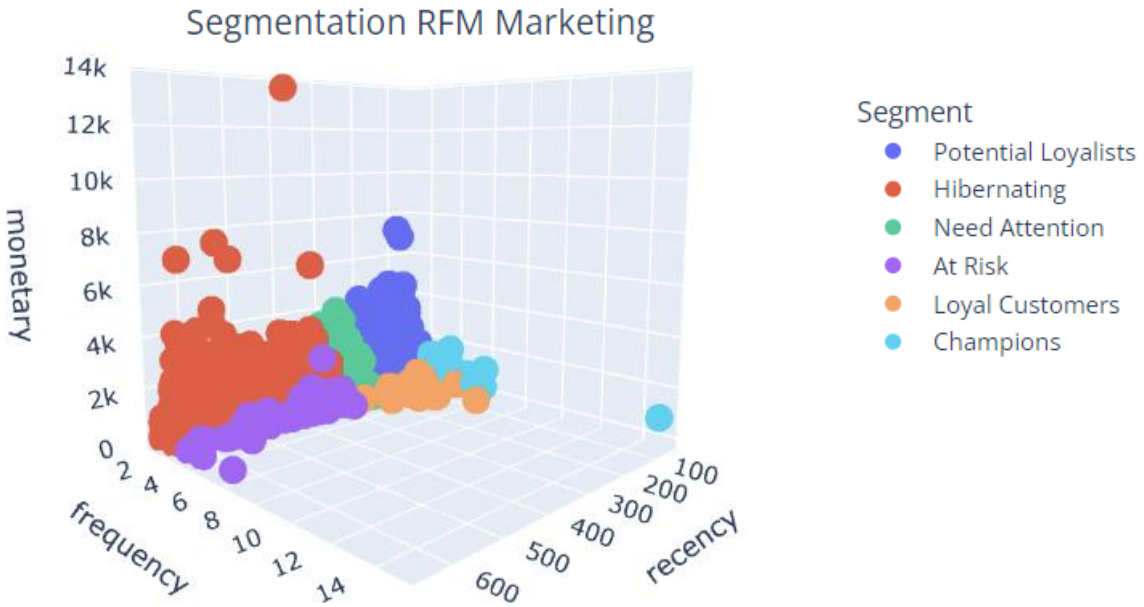
K-Means

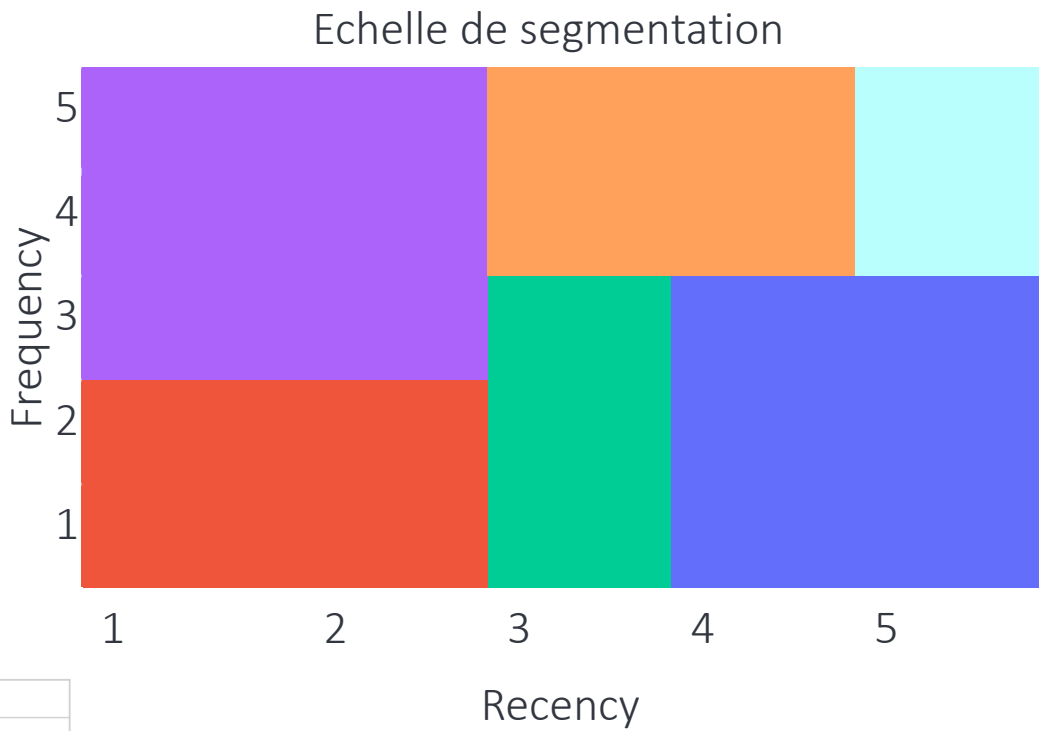
Variables RFM + satisfaction

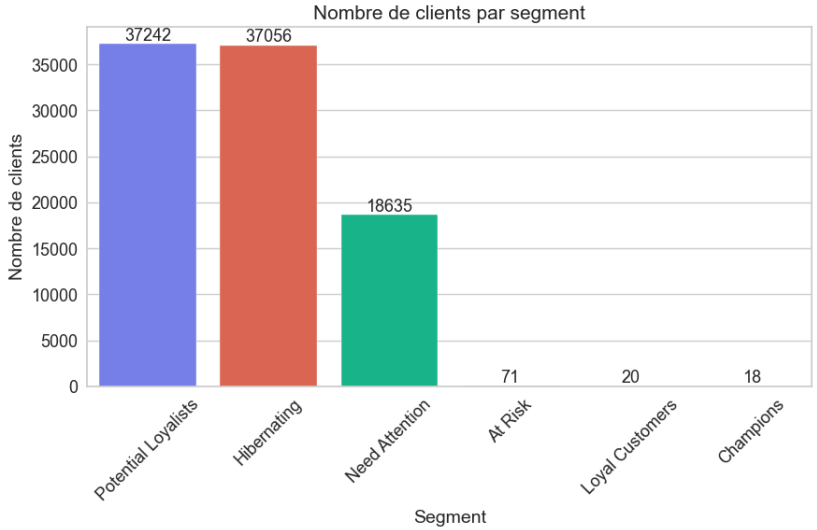
Plus de variables

ACP + K-Means

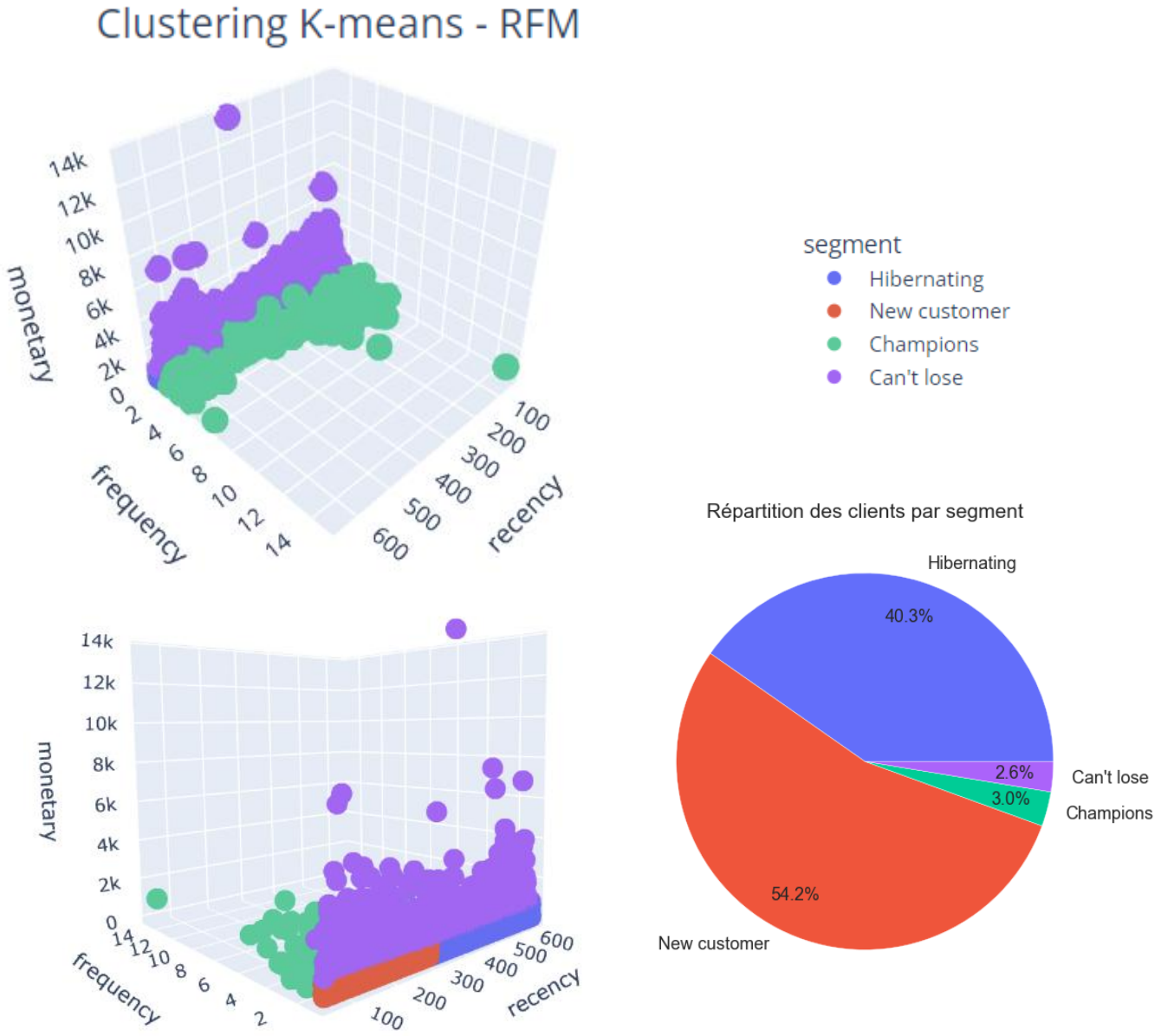
RFM Marketing : démarche simple mais limitée

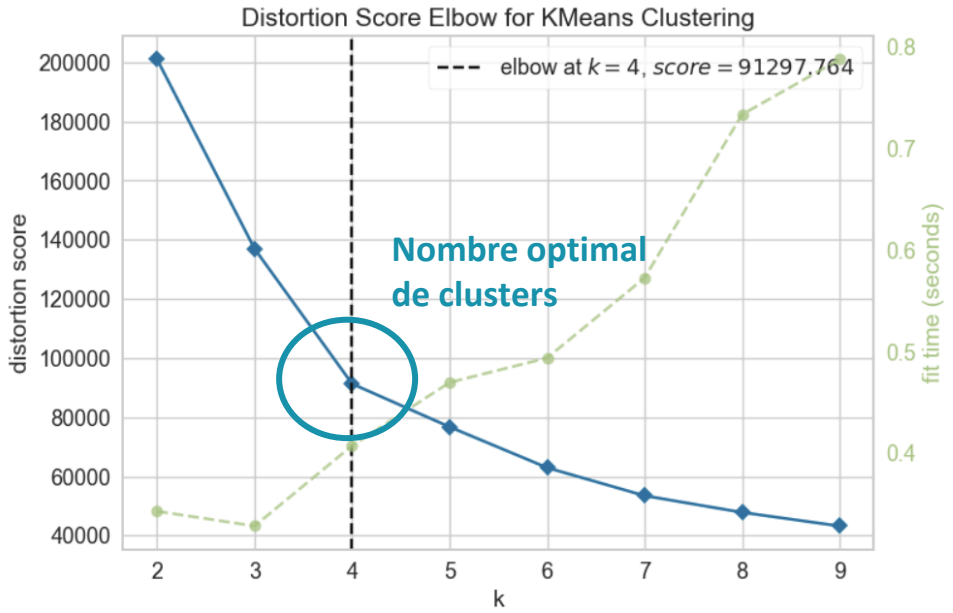






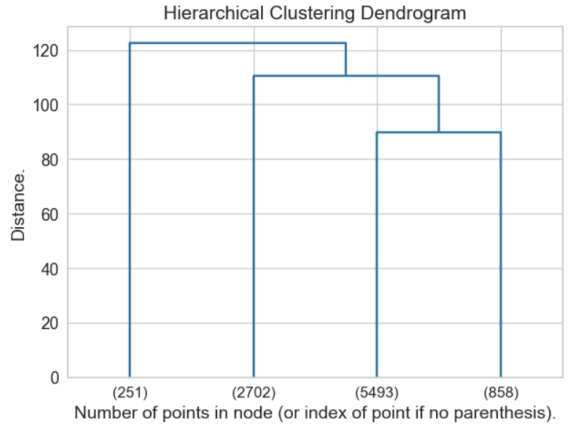
RFM | k-means, 4 clusters





RFM | Autres pistes explorées

CAH : Dataset trop important

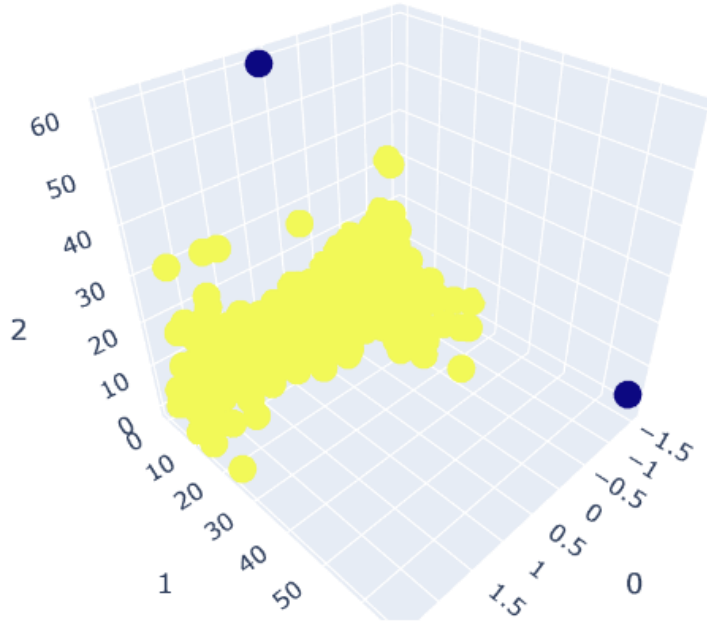


Test sur un échantillon du dataset

Alternative : k-means puis CAH

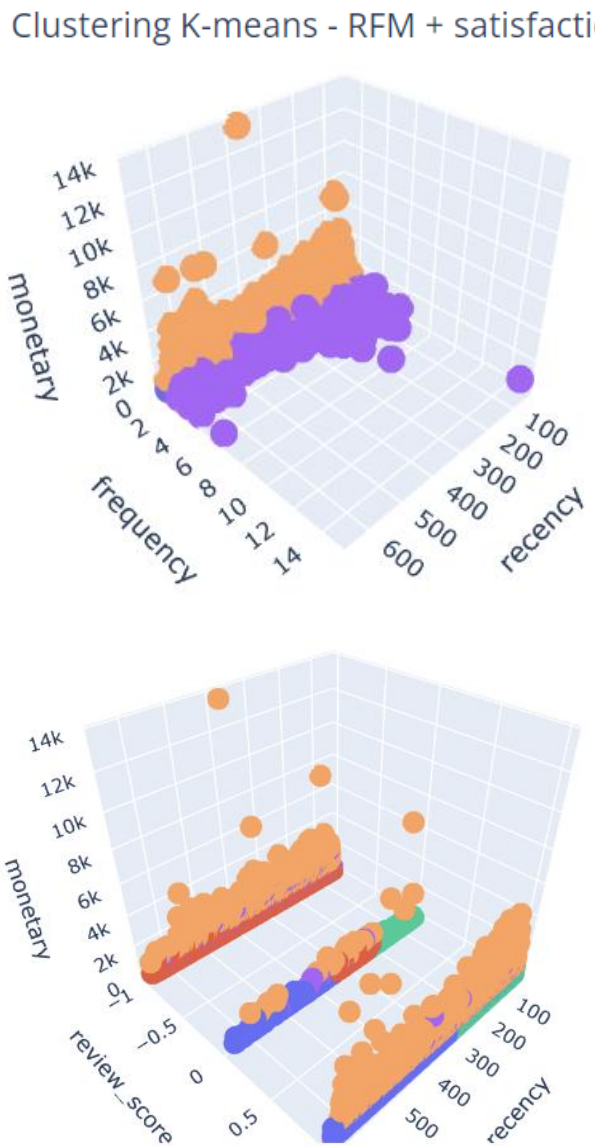
DBScan : Dataset trop important et nuage de points trop dense

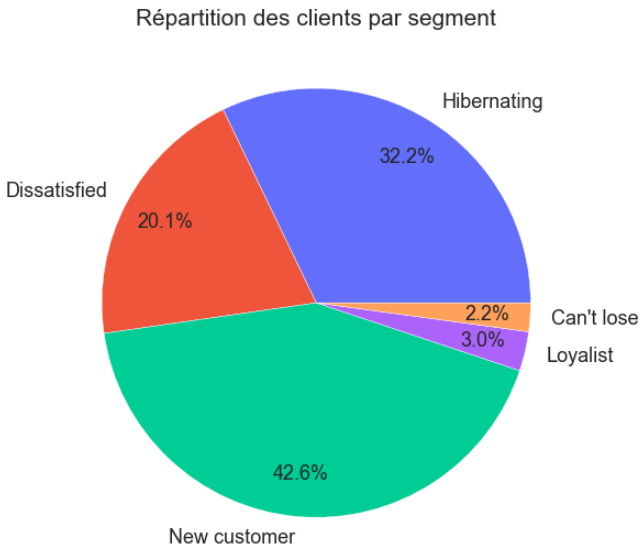
Clustering DBScan avec les variables RFM

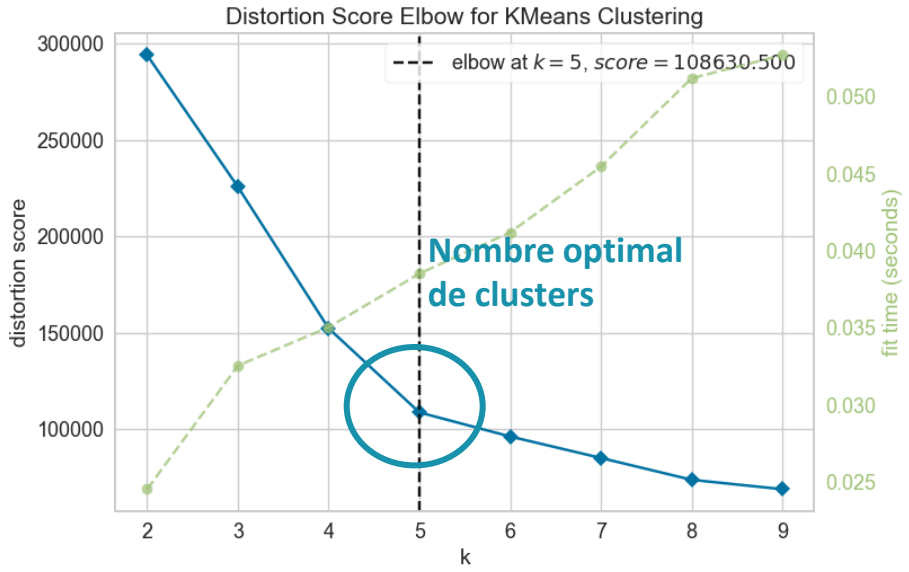


Test sur un échantillon du dataset

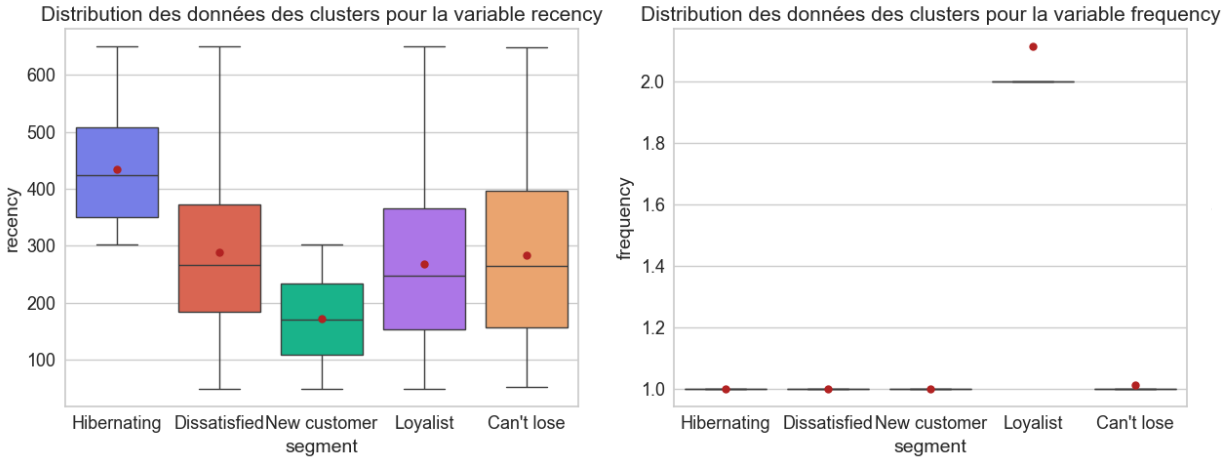
RFM + Satisfaction | K-Means

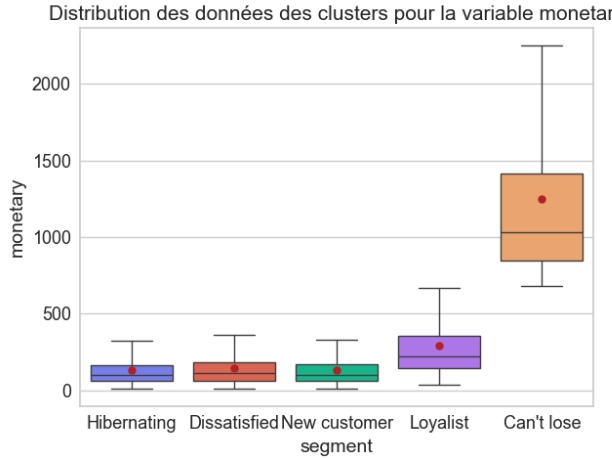


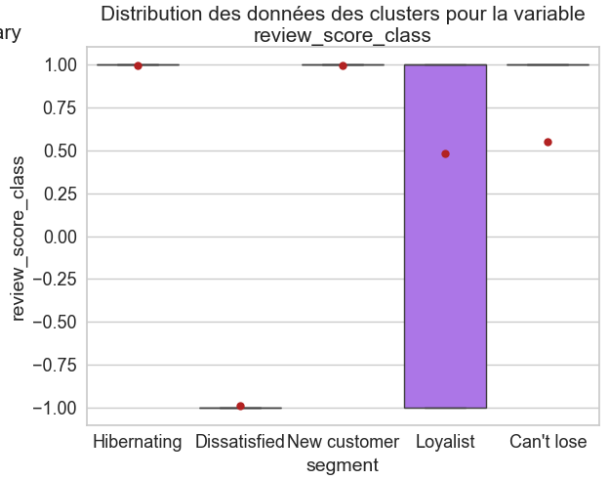




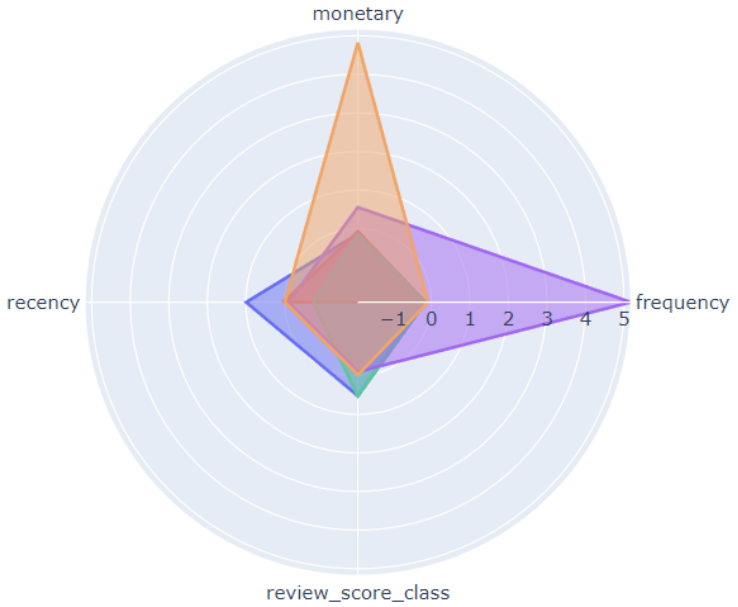
RFM + Satisfaction | K-Means



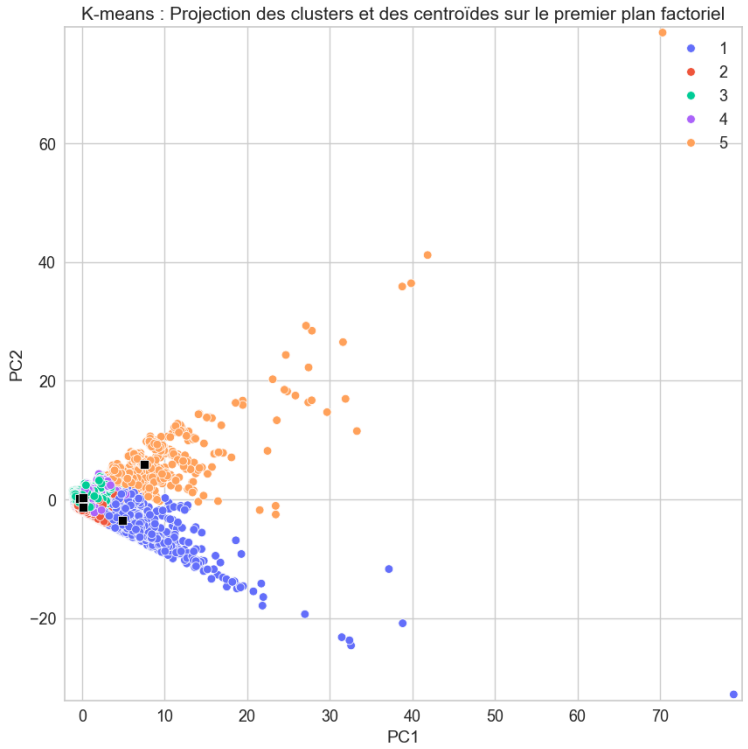


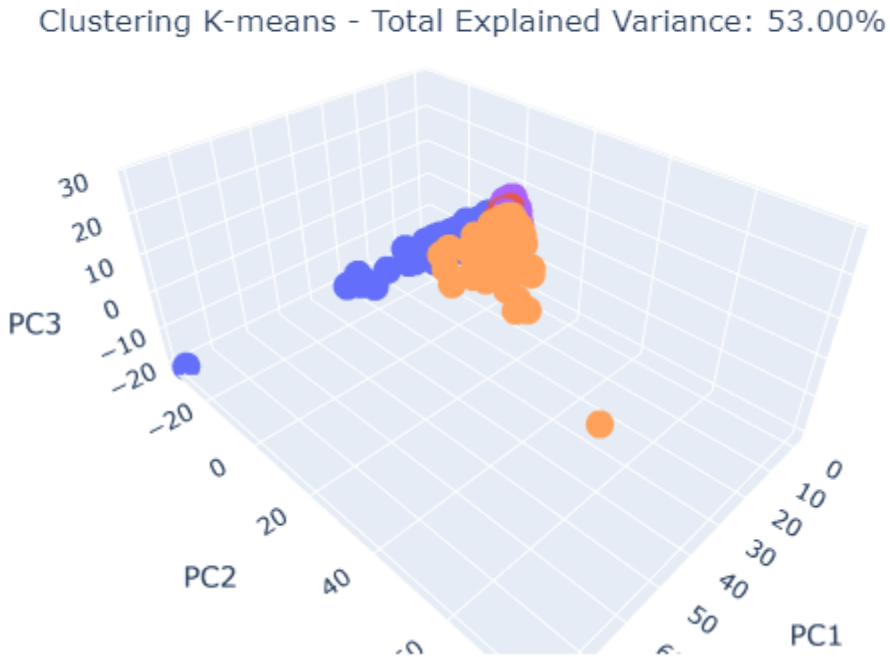


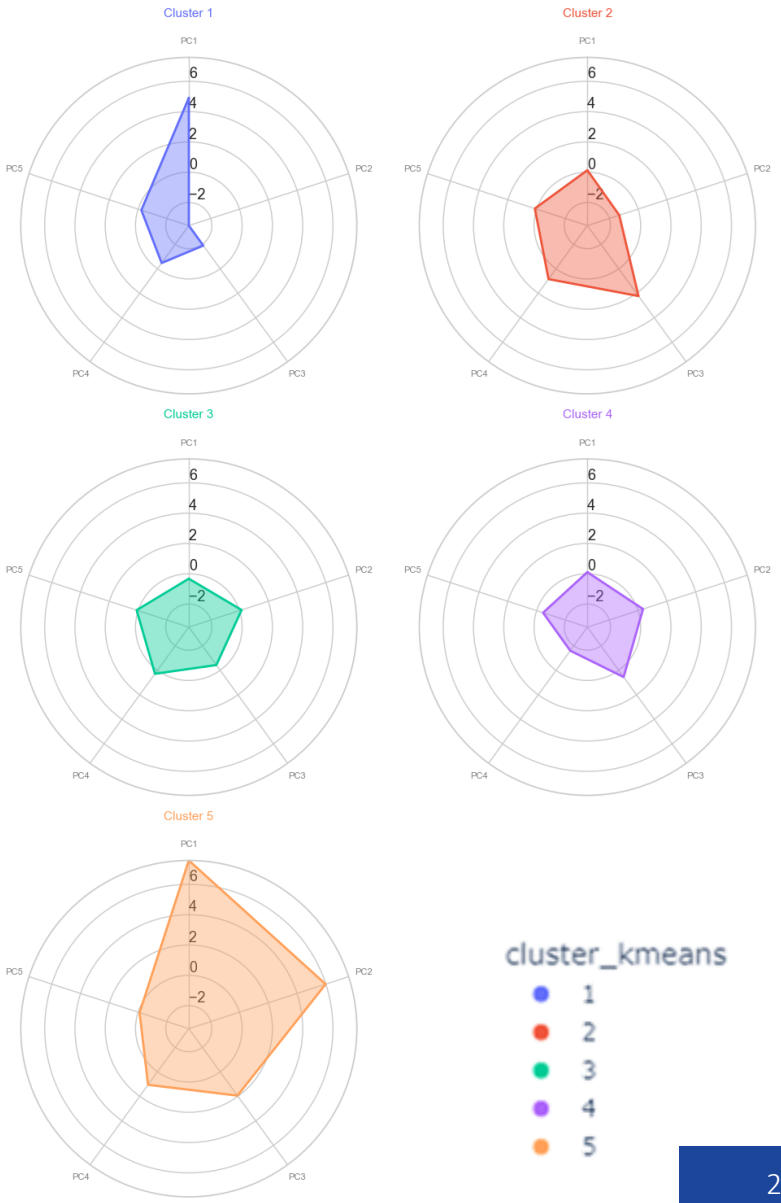
Segment	Définition
Hibernating	High value recency
Dissatisfied	Low value review score
New customer	Low value recency
Loyalist	High value frequency
Can’t lose	High value monetary



ACP + K-Means : définition des clusters non actionnables

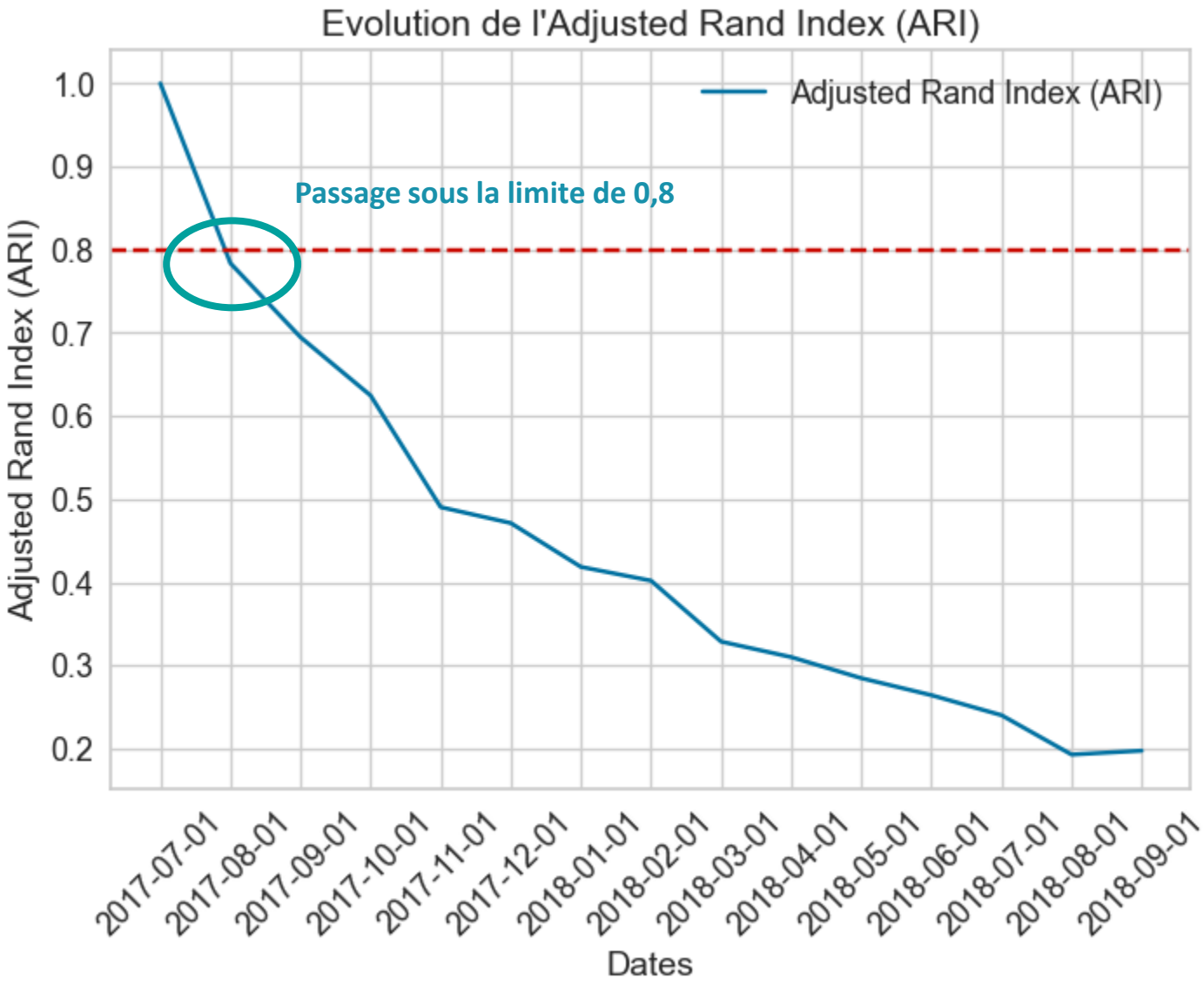






Simulation Maintenance

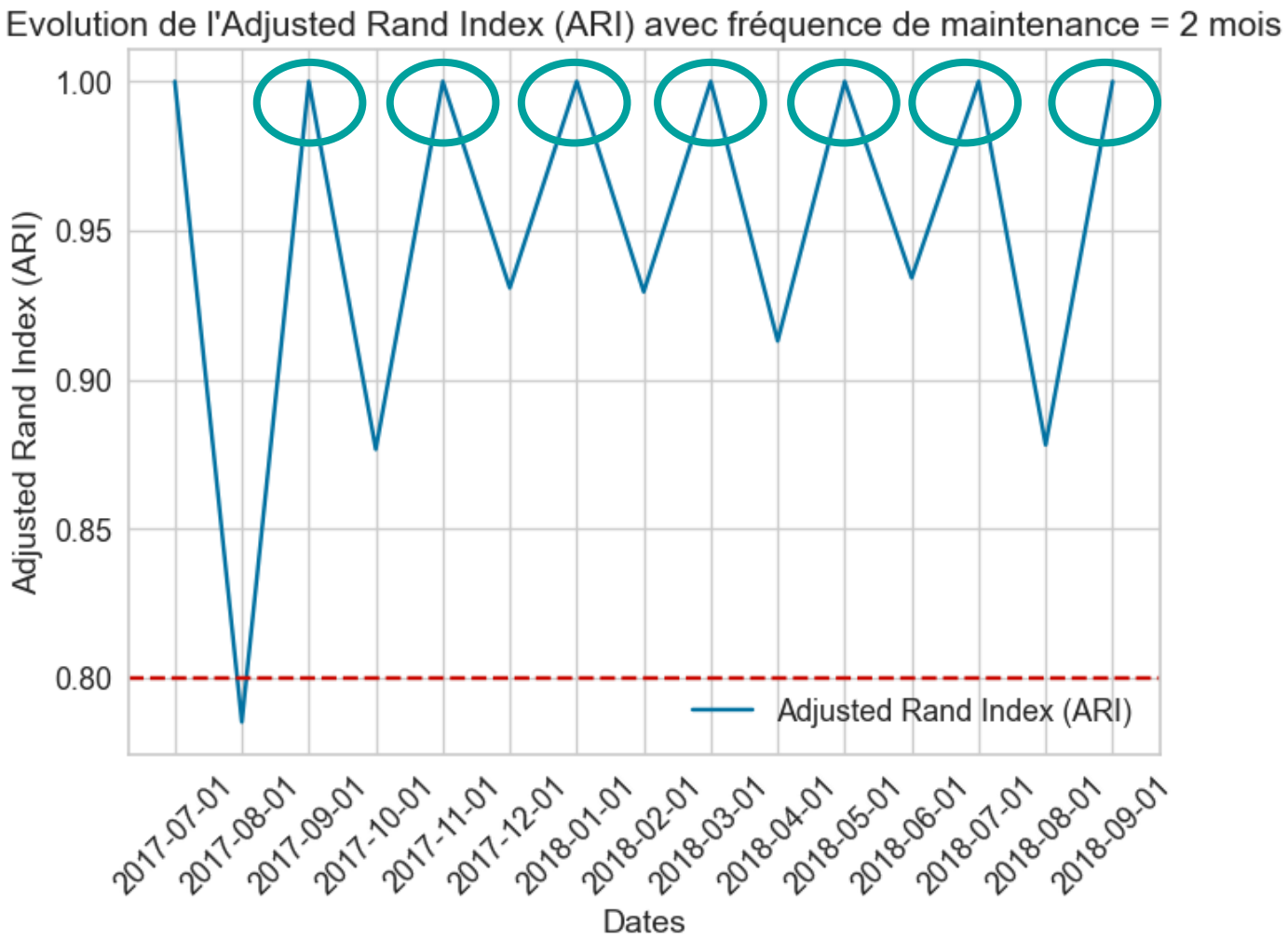
Détermination de la période de maintenance : ARI score



- Echantillon d'entraînement : 6 mois
- Incrément de calcul de l'ARI : 1 mois

- Fréquence de maintenance : 2 mois

Simulation de la maintenance tous les 2 mois



Réentraînement du modèle

- Un réentraînement du modèle tous les 2 mois permet de rester au-dessus de la limite de 0,8.
- Possible d'espacer les maintenances au bout d'un certain temps.

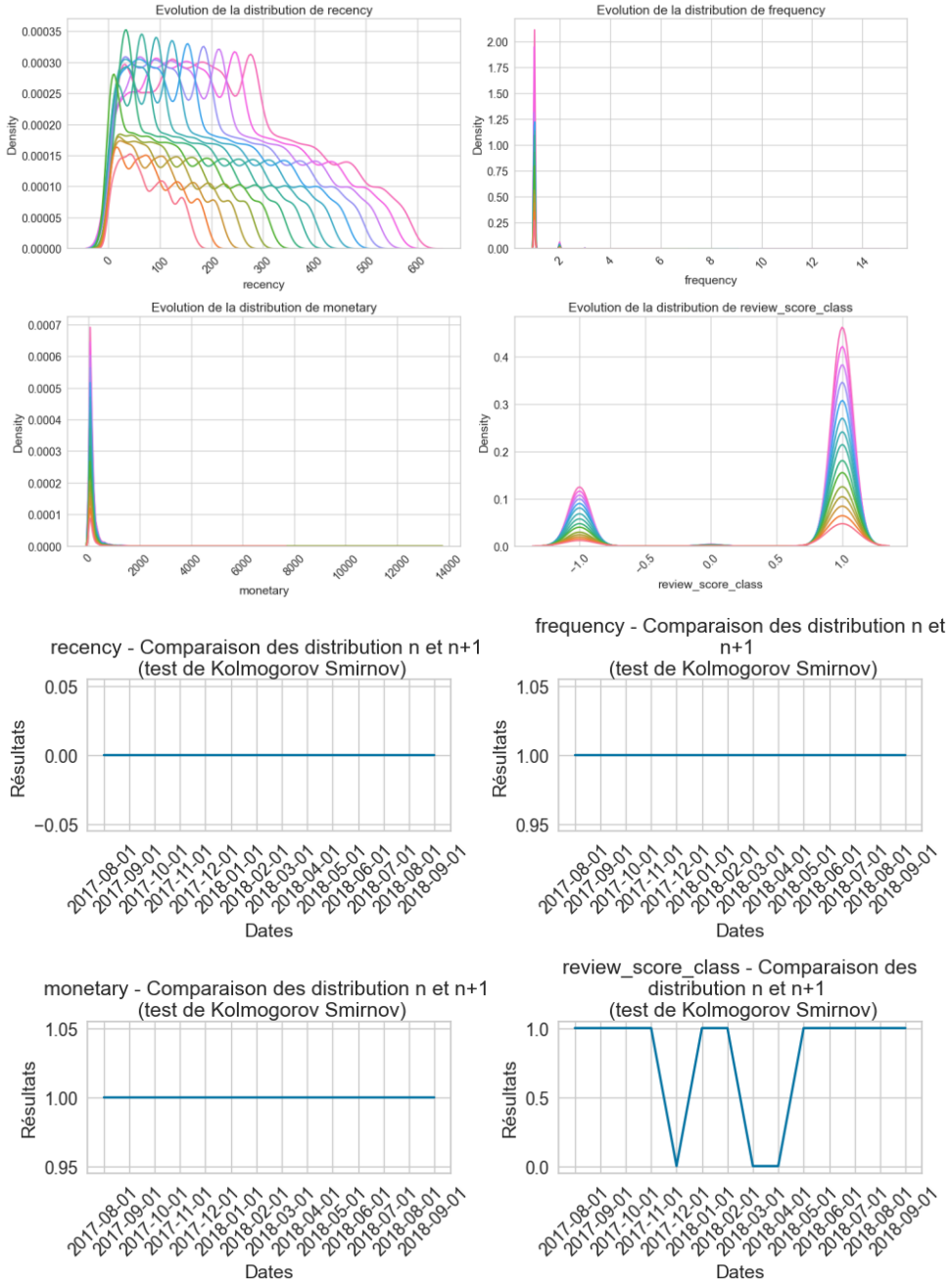
Points de vigilance

Paramètres à surveiller :

- Nombre optimal de clusters k
- Tests statistiques (ex. Kolmogorov Smirnov)

Changement des données d'entrée :

- Echelle du review_score
- Suppression de catégorie
- Ventes exceptionnelles (Black Friday)



Test de Kolmogorov Smirnov

H0 : les distributions sont identiques.

- 1 = H0 non rejetée
- 0 = H0 rejetée

Conclusion & Perspectives

Conclusion

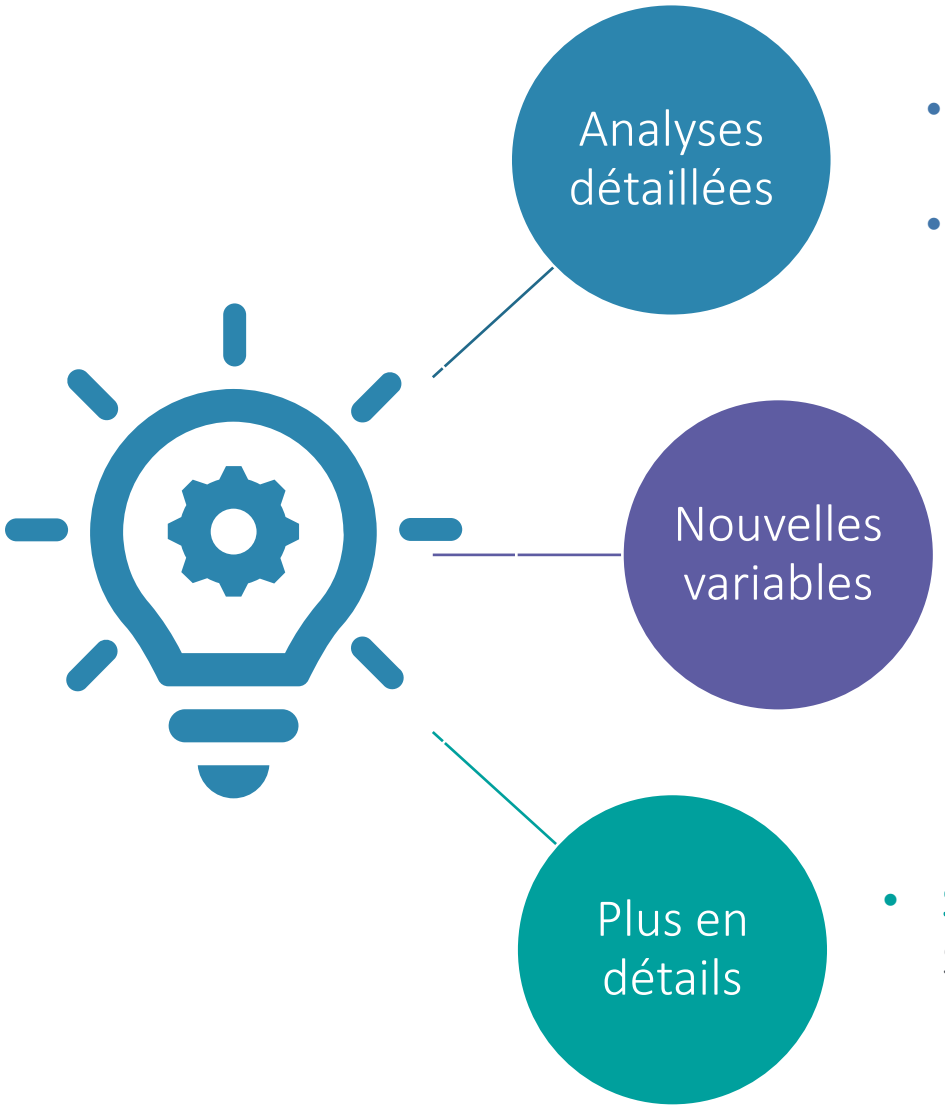
Méthode retenue : **K-Means**

Variables : **Récence, Fréquence, Montant, Satisfaction clients**

5 clusters : New customer, Hibernating, Dissatisfied, Loyalist, Can’t Lose

Maintenance tous les 2 mois dans un premier temps, surveillance de paramètres (k, distribution des variables, …)

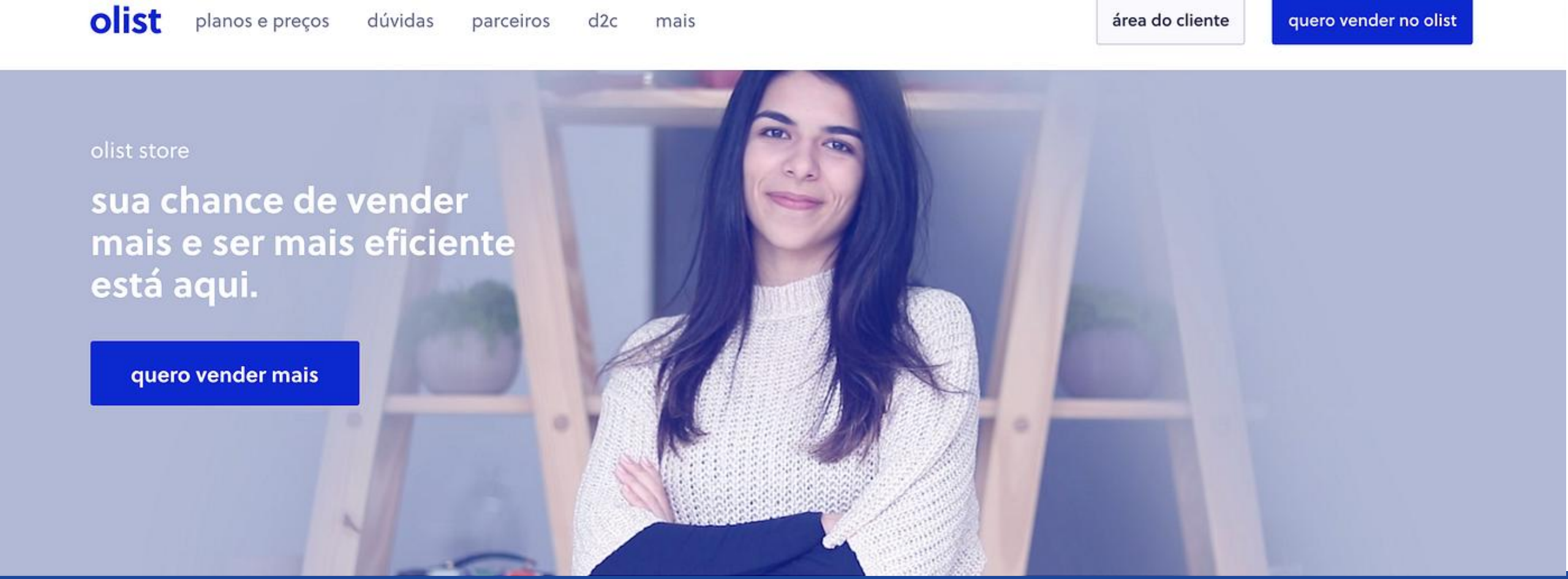
Perspectives



- **Avis des clients :** analyse de sentiment, longueur de l'avis, étapes concernées, vendeurs
- **Délai de livraison :** distance client-vendeur, respect de la date limite de livraison produit, …

- Recueillir **plus d'informations sur les clients** : Age, Genre, …
- Ajouter la **richesse du quartier** pour l'utilisation des données de géolocalisation

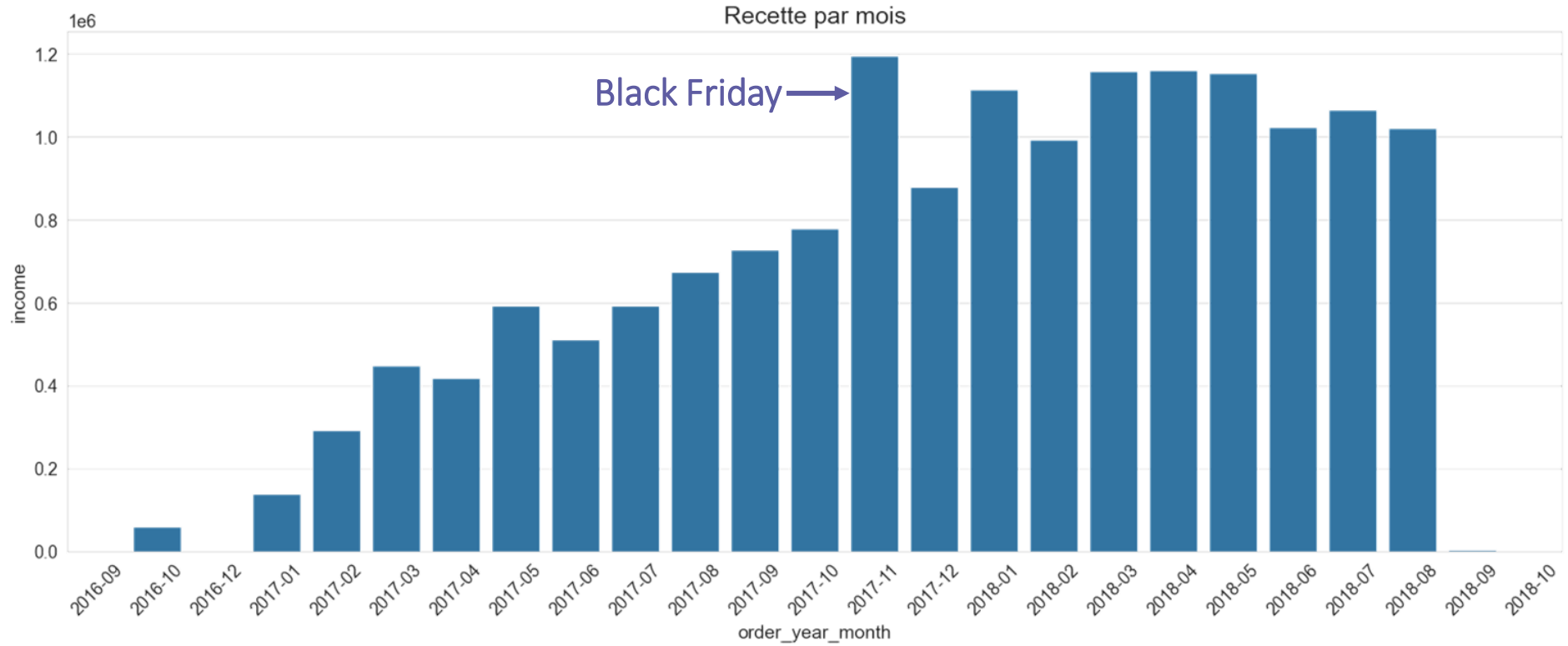
- **Segmentation clients pour les évènements micro** (Black Friday, Soldes, …) : définir des campagnes ciblées pour les clients qui pourraient revenir



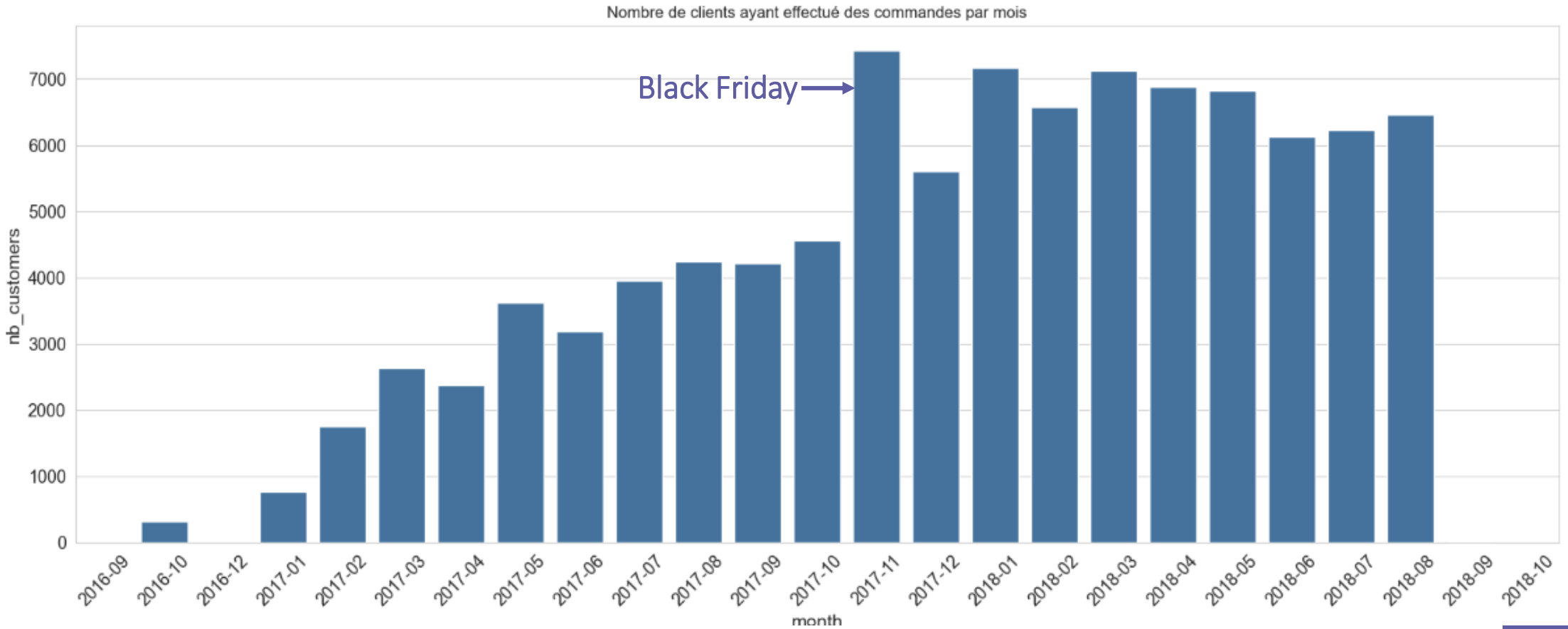
Merci de votre attention

Back-up slides

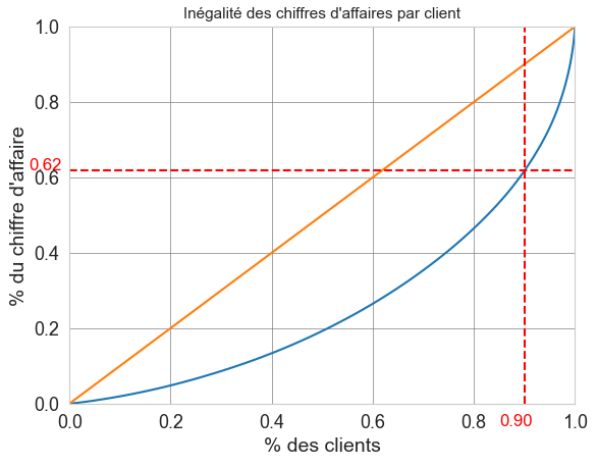
Un chiffre d'affaires en croissance

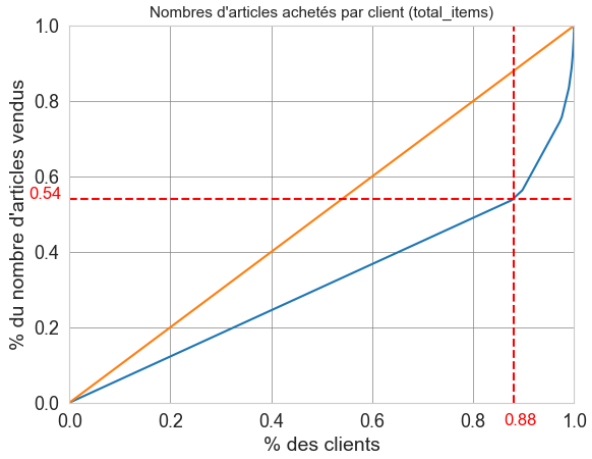


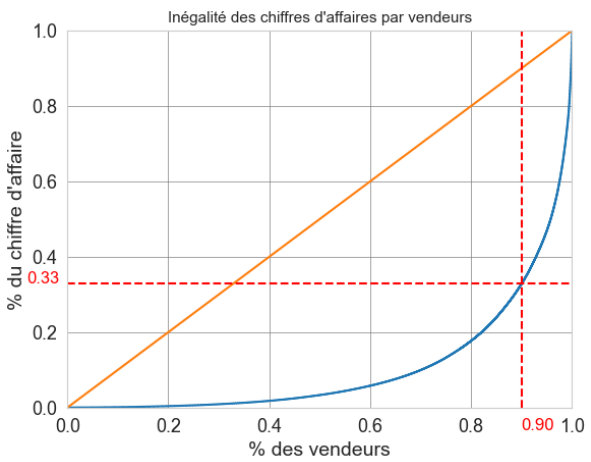
Nombre de clients en augmentation



10% des clients représentent 38% du chiffres d'affaires







Fréquence

3% des clients =

7% des commandes

Monétaire

10% des clients =

38% du CA

Items

12% des clients =

46% des articles vendus

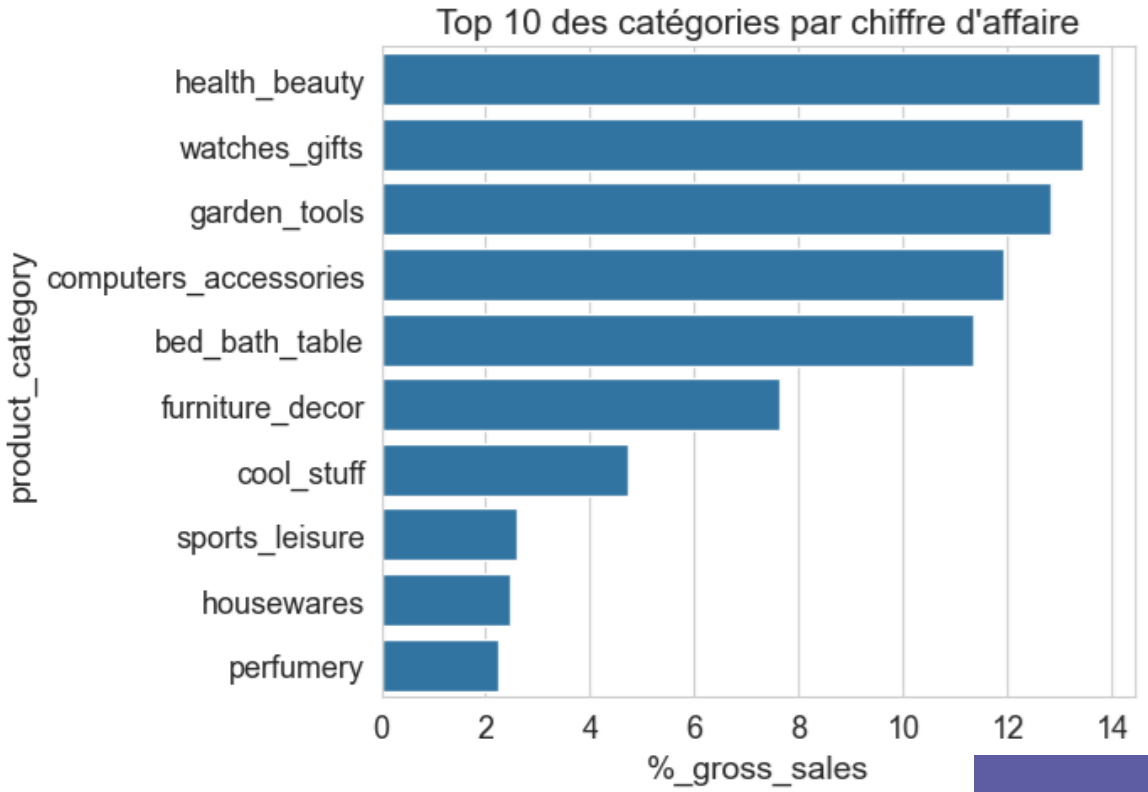
CA par vendeur

10% des vendeurs =

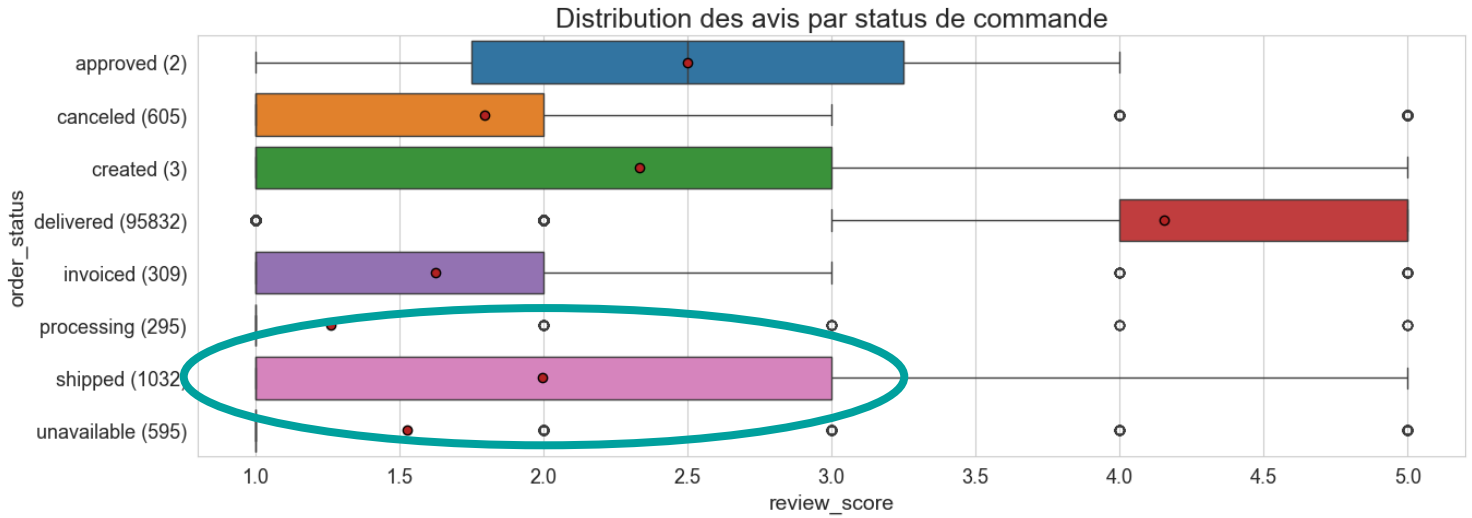
67% du chiffre d'affaires

Health beauty: la catégorie de produit phare

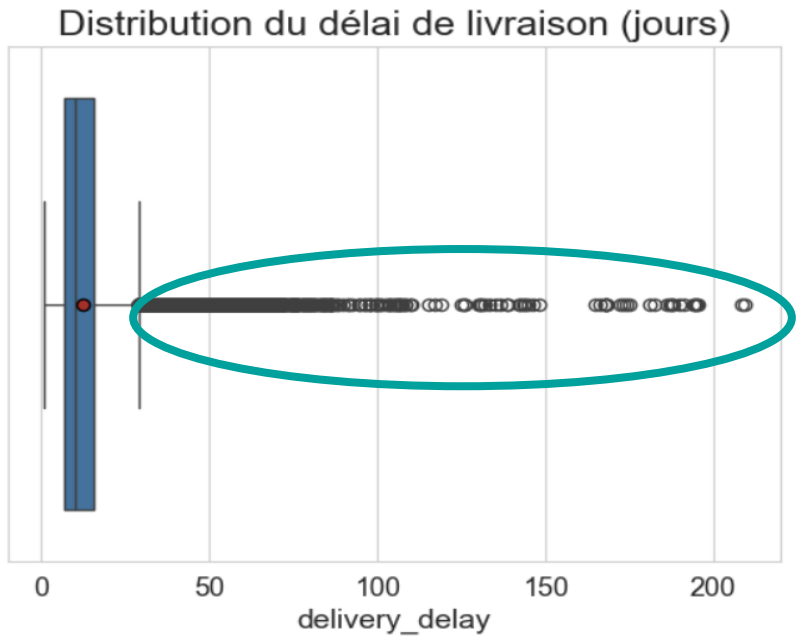
12% des ventes et 14% du chiffre d'affaires



Des clients insatisfaits de la livraison



25% des délais de livraison > 15j



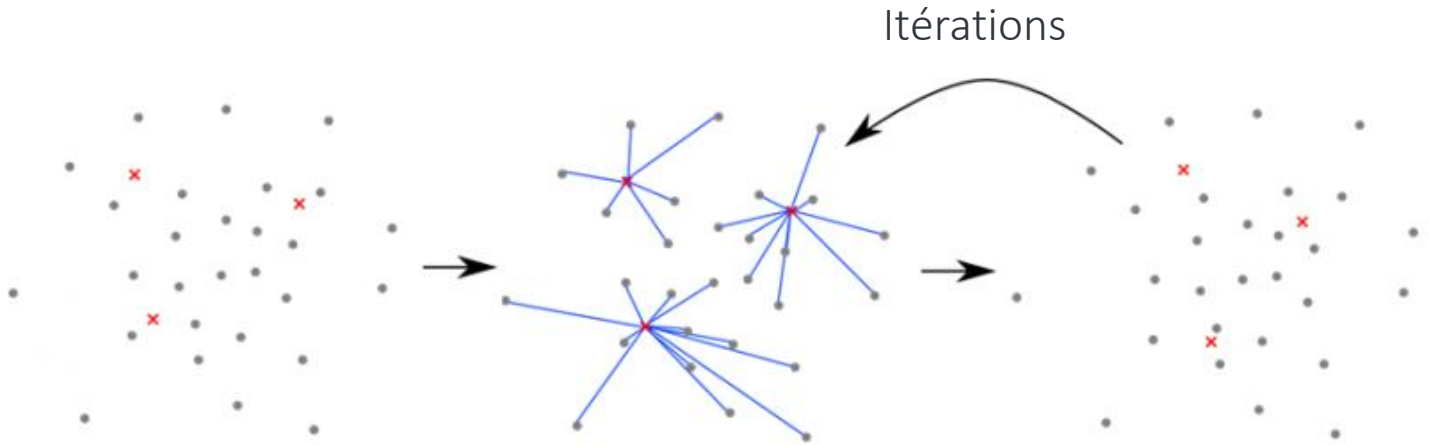
Méthode des k-means

Choix du nb de clusters : k

Initialisation : k centres placés aléatoirement au sein du nuage de points

Itérations :

- Association des points au centre le plus proche,
- Calcul des nouveaux centres de gravité
- Association des points aux nouveaux centres
- Etc… jusqu'à ce que les centres ne bougent plus (= convergence)



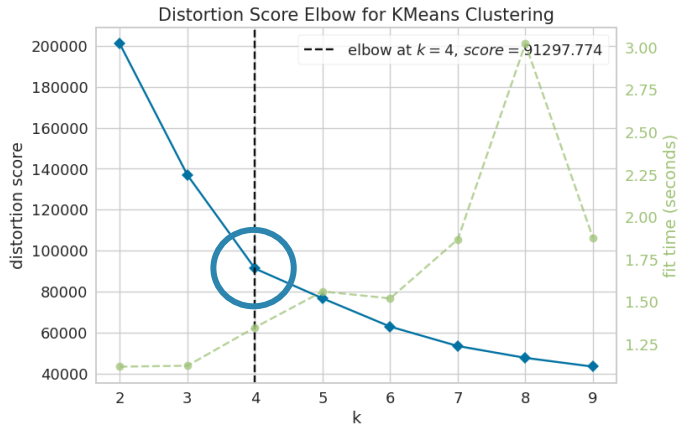
1- Initialisation, k centres

2- Association

3- Redéfinition des centres

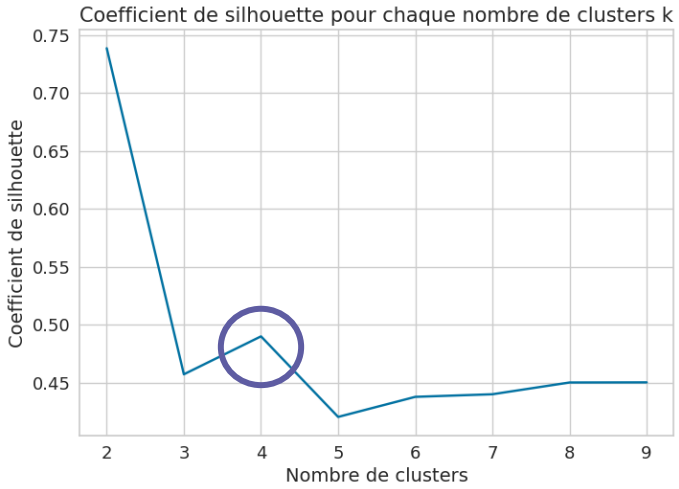
Méthodes pour déterminer le nombre optimal de clusters k

Distortion score



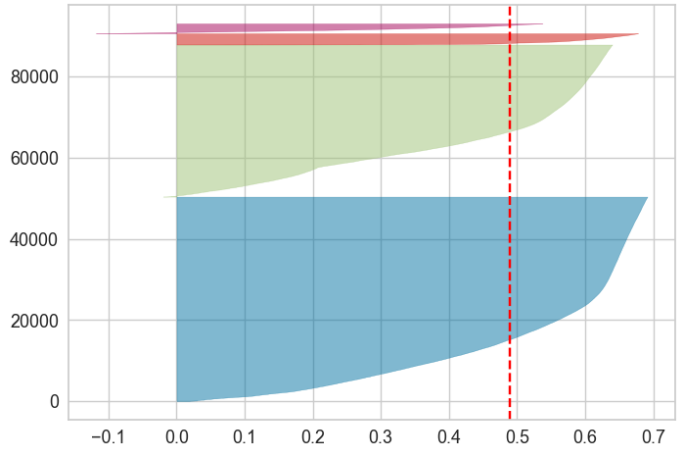
Méthode du coude

Silhouette score

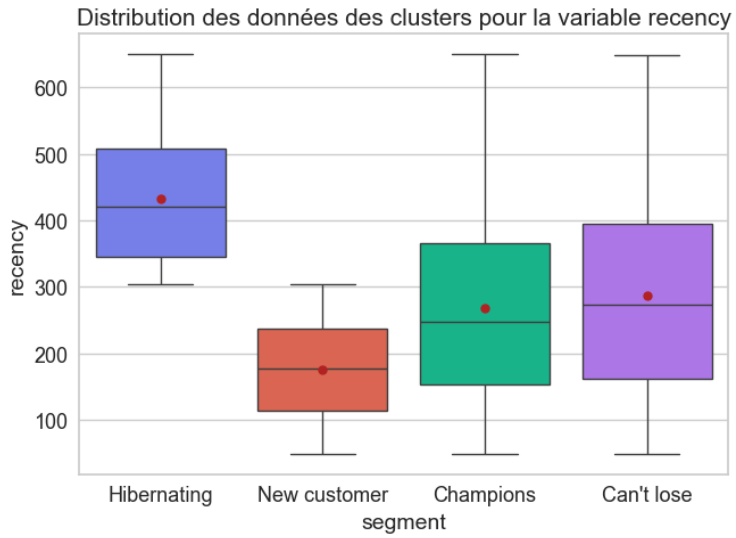


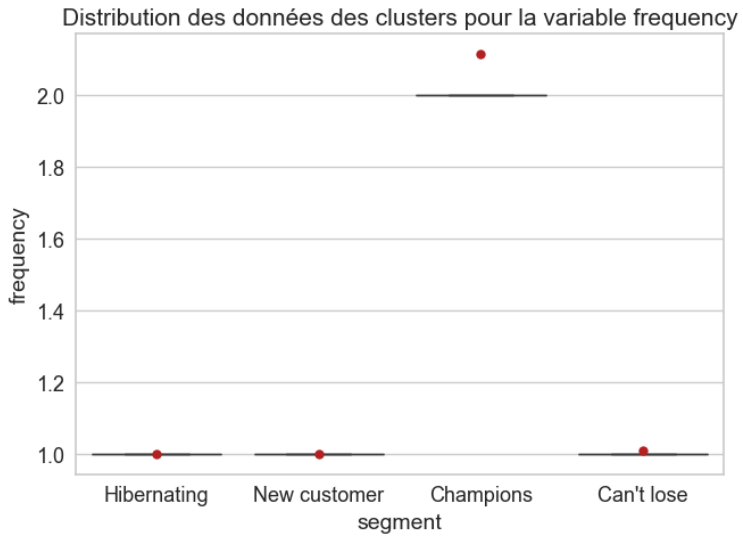
Valeur la plus élevée

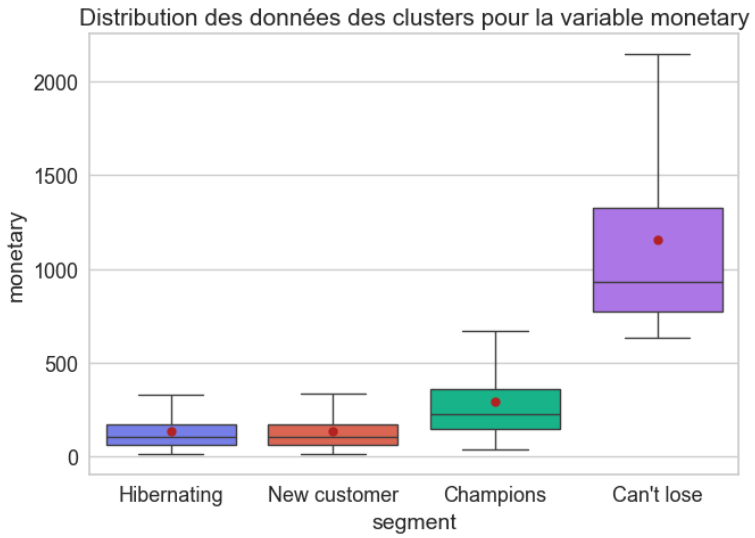
Silhouette plot

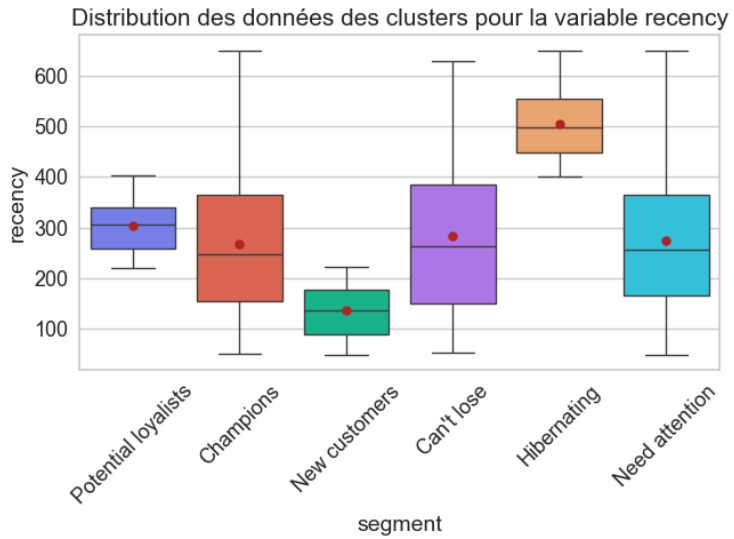


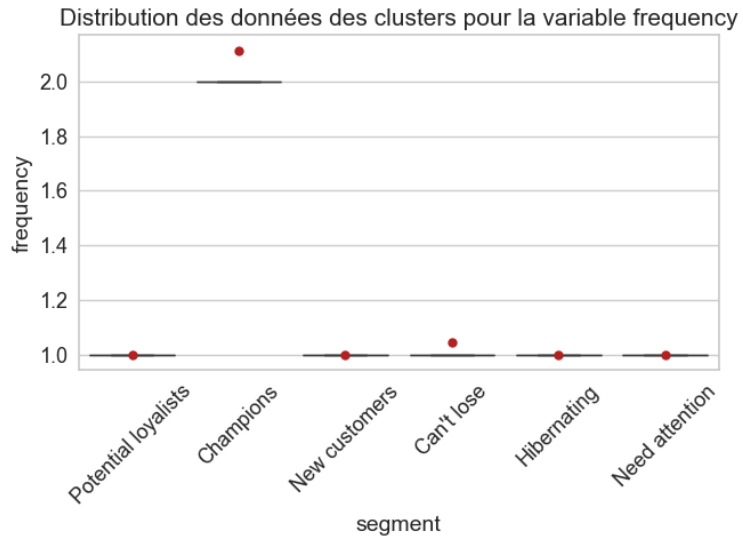
RFM k-means : comparaison 4 et 6 clusters

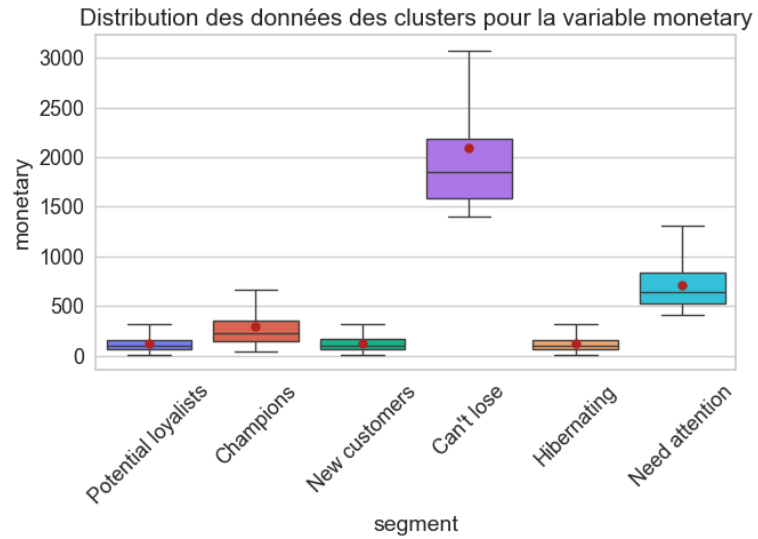








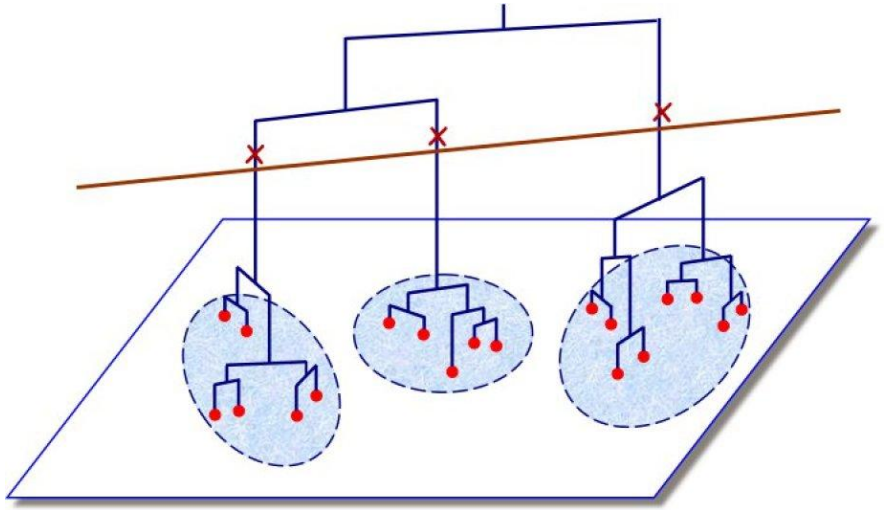




Classification ascendante hiérarchique (CAH)

Initialisation : 1 individu = 1 cluster

Itérations : Regroupement avec les individus les + proches en distances (les plus similaires)



Un dendrogramme représente la hiérarchie sous la forme d'un arbre :

- Individus = feuilles
- Branches = liaisons entre les individus

La distance entre les individus est représentée par la longueur de la branche : plus la branche est courte, plus les individus sont similaires et inversement.

Les classes, ou *clusters*, sont obtenues en coupant l'arbre à une distance choisie *(souvent à l'endroit où les branches s'allongent).*

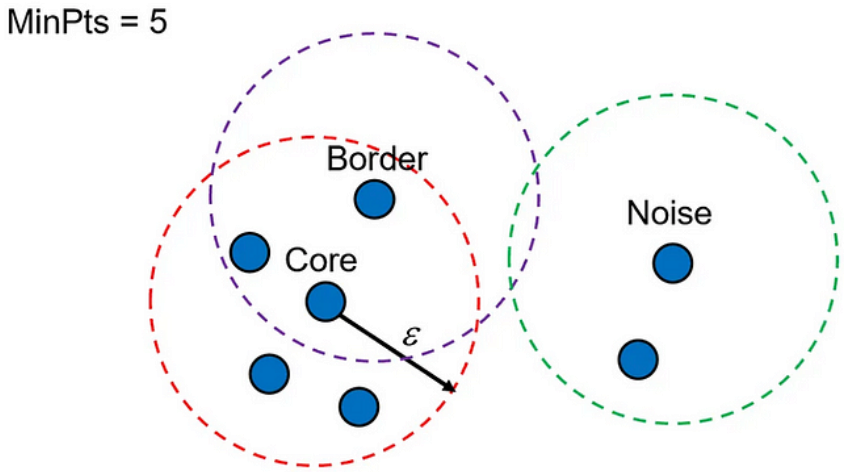
Density based Scan (DBScan)

Hyperparamètres :

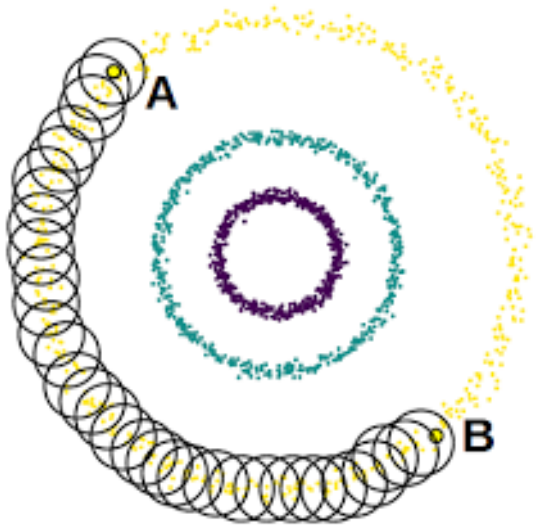
- **Epsilon** : rayon autour d'un point de données
- **Minpts** : nombre minimum de points nécessaires pour former une région dense (=cluster)

Initialisation : 1 individu = 1 cluster

Itérations : Regroupement des individus par densité



Core, border, and noise points (image by author)



Analyse en Composantes Principales

Objectif : Réduction de dimension

\+ : Visualisation des données, Améliorer l'apprentissage, Gain en stockage et temps de traitement par la suite

\- : Perte d'informations

- Recherche de la projection permettant de visualiser au mieux les données
- Composante principale = combinaison linéaire des variables initiales