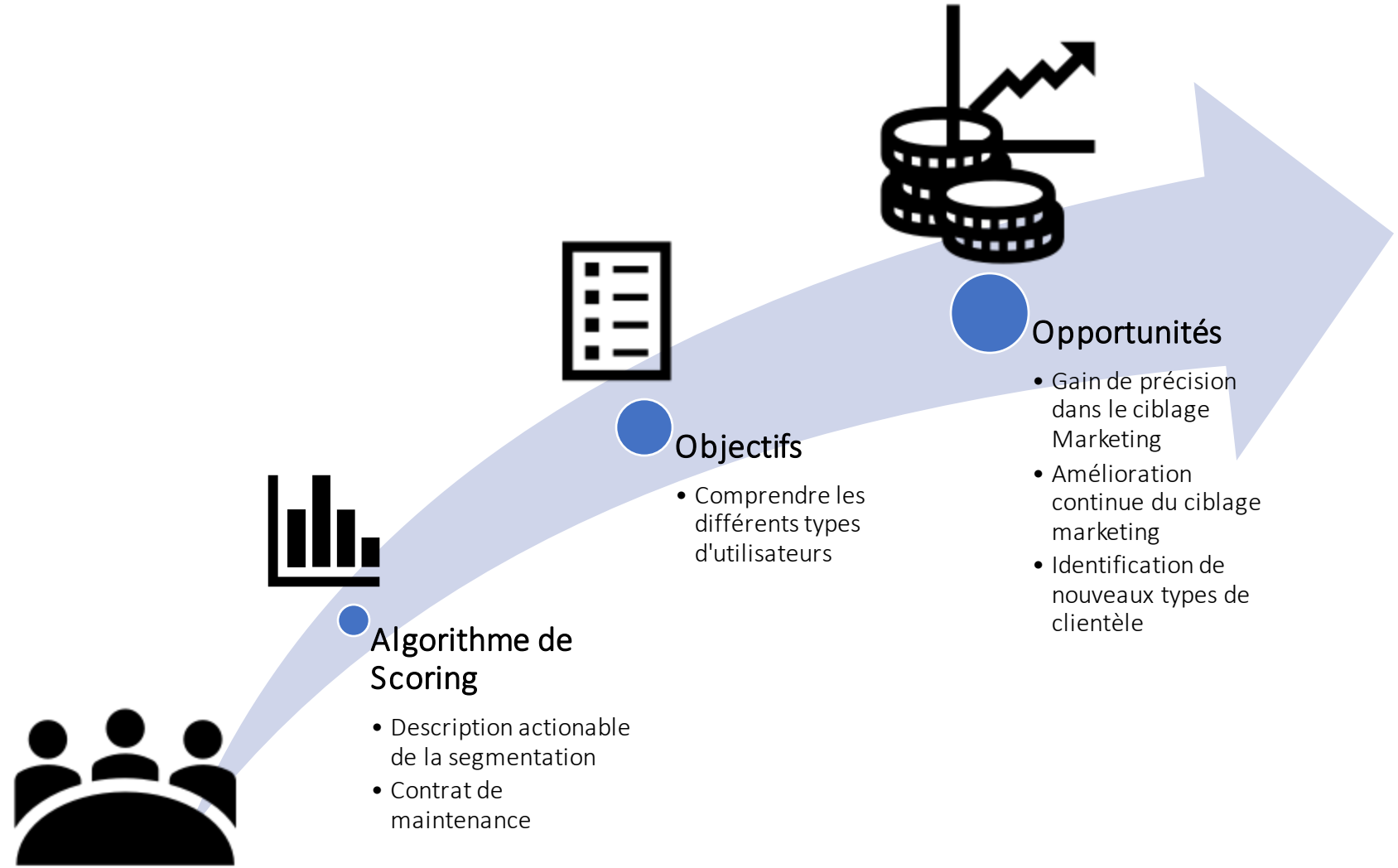
Segmentation de la clientèle

Société OLIST

Bailly DIOUNOU, Ingénieur IA – 09/02/2021

Contexte & périmètre du projet



Présentation de la problématique

Interprétation & pistes de recherche

Comprendre les différents types d'utilisateurs

Problème de "riches"

Croissance... de la base de clients

- Augmentation et diversification des flux de communication, de richesses, etc.
- Limites de la relation client/Marketing de proximité (ou ultra spécialisée)

Comprendre les différents types d'utilisateurs

Axes de réponse

Croissance... de la base de clients

- Travail: augmenter les heures, diversifier la qualité de son métier
- Capital humain: investir dans de nouvelles compétences humaines
- Progrès technique: investir dans de nouvelles technologies

Comprendre les différents types d'utilisateurs

Nouvelles technologies

Croissance... de la base de clients

- Automatisation de tout ou partie du traitement de la donnée générée
- **Economies d'échelle**

Exemple: La *boulangerie de Briis-sous-Forge* face au nouveau *parc d'activités* (Institut supérieur, Entrepôt du secours populaire, Call center)

Profilage commercial de OLIST

- C.A. réalisé sur la transaction
  - Facteurs déterminants: Fidélité du client, Montant de la transaction
- OLIST: place de marché, intermédiation et transaction électronique en retail
- Base de clientèle à diversité potentiellement infinie, corrélé à la gamme des produits proposés par les vendeurs

Profilage commercial de OLIST

- C.A. réalisé sur la transaction
  - Facteurs déterminants: Fidélité du client, Montant de la transaction
- OLIST, place de marché, intermédiation et transaction électronique en retail
- Base de clientèle à diversité potentiellement infinie, dépendant des produits des vendeurs

- Croissance rapide de la base clientèle
  - Opportunité Marketing
  - **Connaitre sa base client**

Segmentation de la base de clientèle

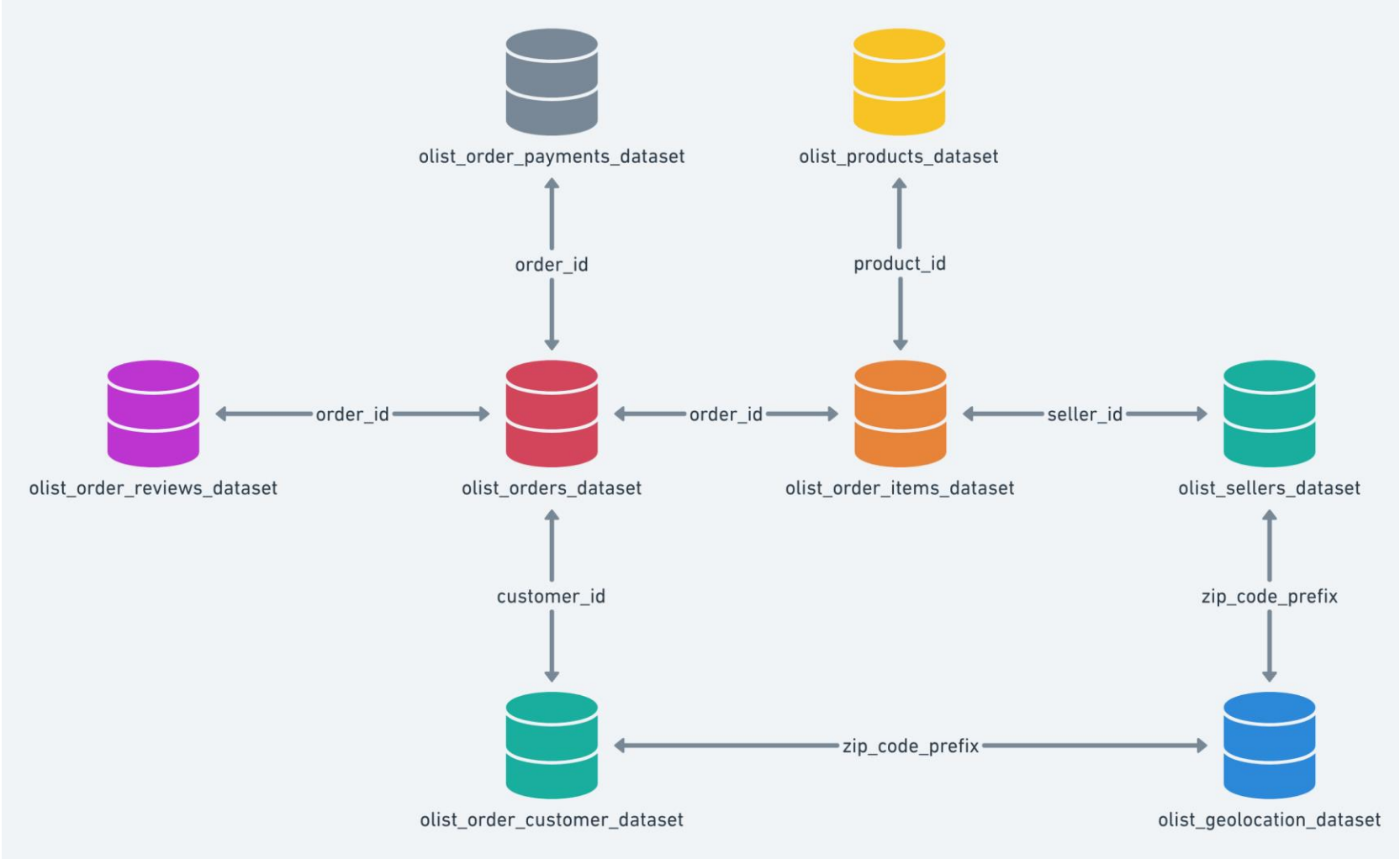
- Segmentation manuelle sur la base des déterminants du CA: Fidélité & Montant des transactions
  - Segmentation RFM
- Segmentation automatique par modèles d'Intelligence Artificielle
  - Segmentation par K-Means & DBScan

Nettoyage & Analyse exploratoire des données

Présentation générale du jeu de données

Données statistiques de base - Données brutes

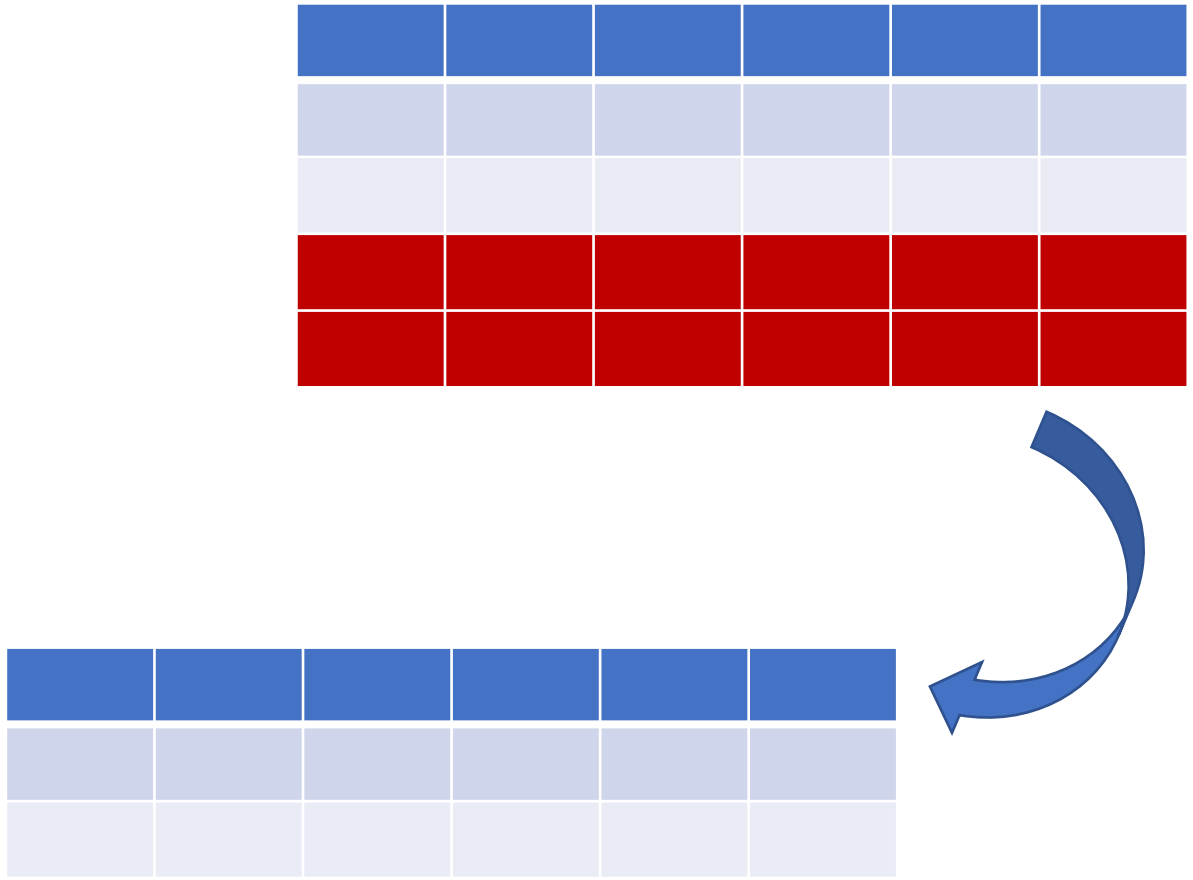
- Données relatives aux commandes (clients, produits, paiements et frêt)
- 100k commandes entre 2016 et 2018 (~2ans)
- Variables catégorielles **(nominales et ordinales)** majoritaires
- Relativement peu de valeurs manquantes (sauf sur les commentaires de satisfaction et les commandes)



Nettoyage des données

Commandes

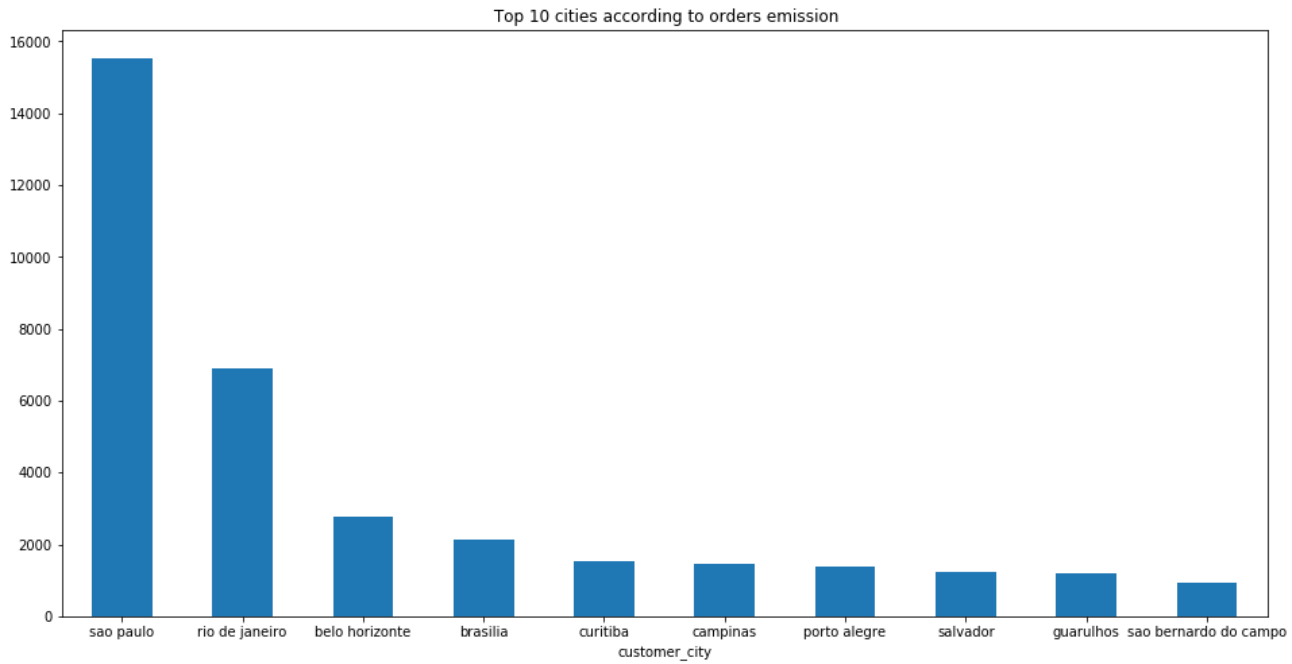
- Suppression des 3% de commandes non-livrées.
- Les données affectées: *orders, customers, items, payments*



Données sur les clients

Géographie

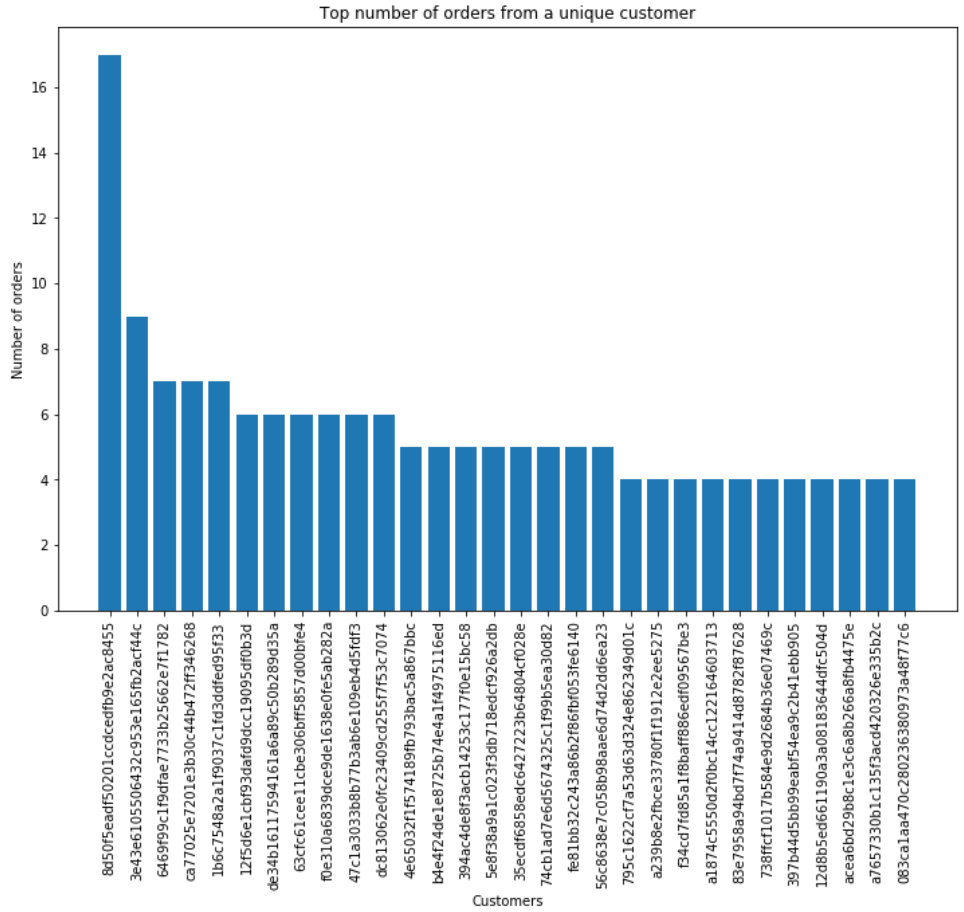
- 35% des commandes proviennent de 10 villes (0.24%)
- Champ de cibles pertinent pour des projets pilotes



Données sur les clients

Nombre de commandes

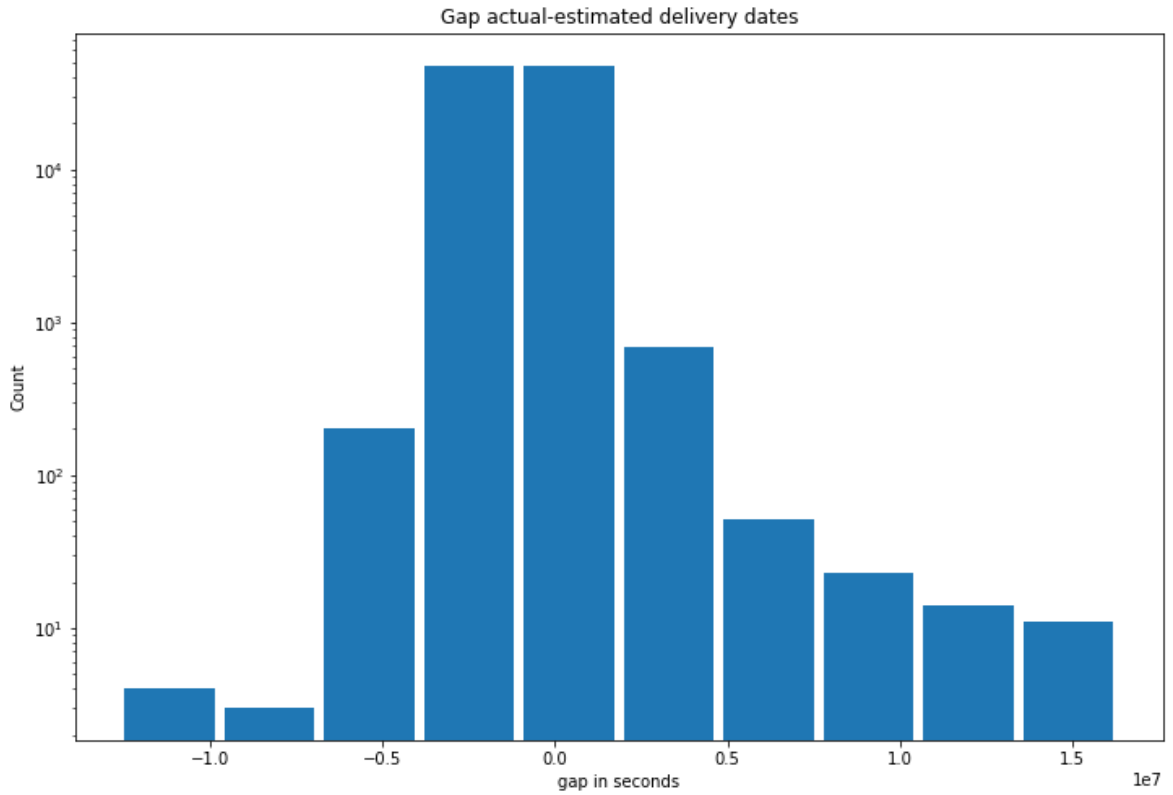
- Seulement 20 clients sur 96k ont plus de 4 commandes ordonnées



Données sur les commandes

Délais de livraison

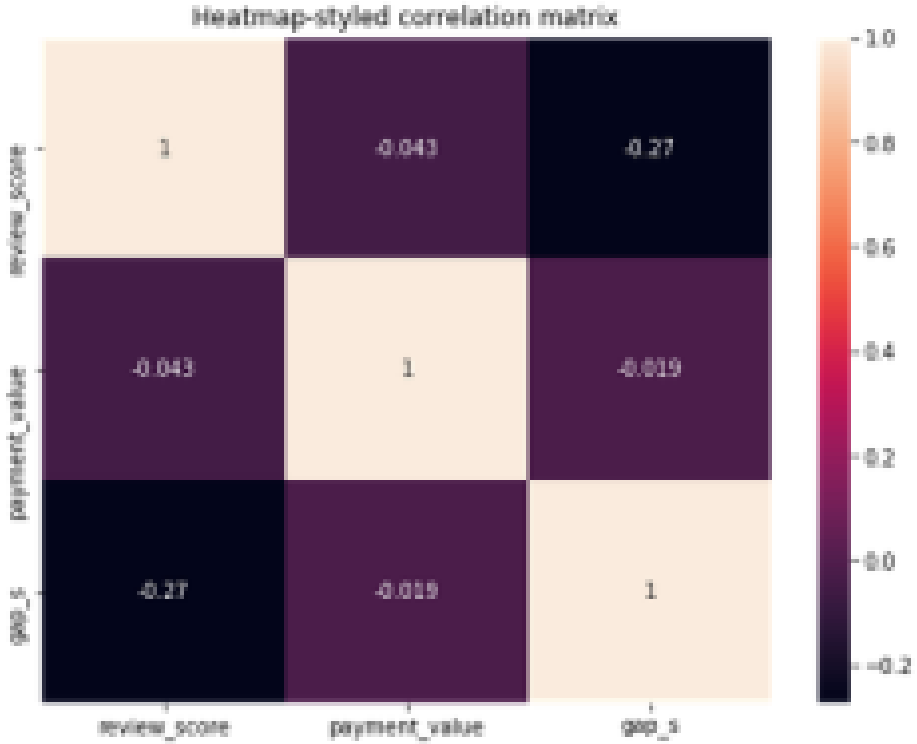
- Distribution asymétrique du délai de livraison, avec plus de retard que d'avance



Données sur les commandes

Délais de livraison

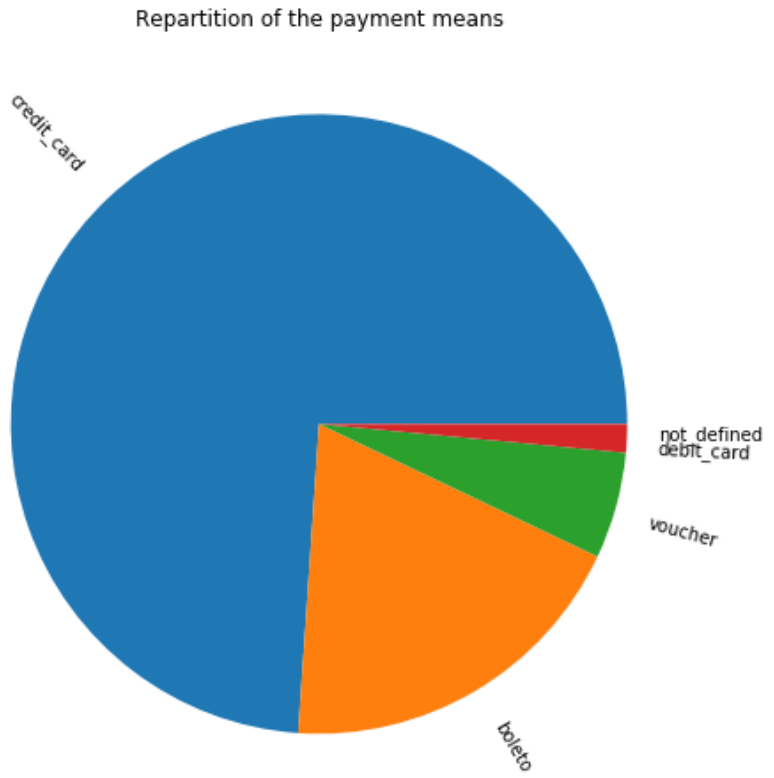
- Distribution asymétrique du délai de livraison, avec plus de retard que d'avance.
- Note de satisfaction *review_score* bien mieux corrélée avec le délai de livraison *gap* qu'avec le chiffre d'affaires des partenaires vendeurs.



Données sur les paiements

moyens de paiement

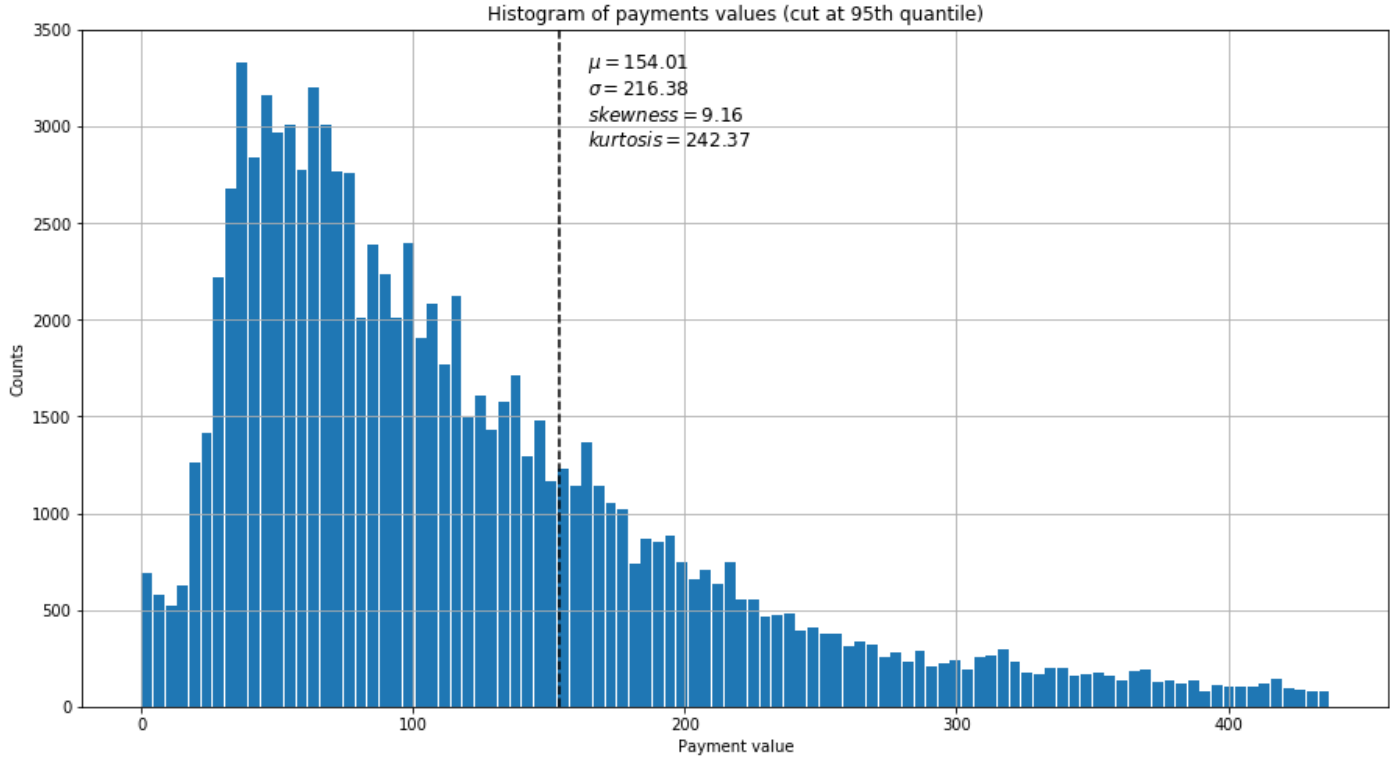
- Près de ¾ de cartes de crédit, et relativement peu de carte de débit
    - Propension à la consommation
- 15% de Boleto
    - Taux de non-bancarisation des clients non-négligeable



Données sur les paiements

valeurs des paiements

- La grande majorité des paiements < C.A. moyen

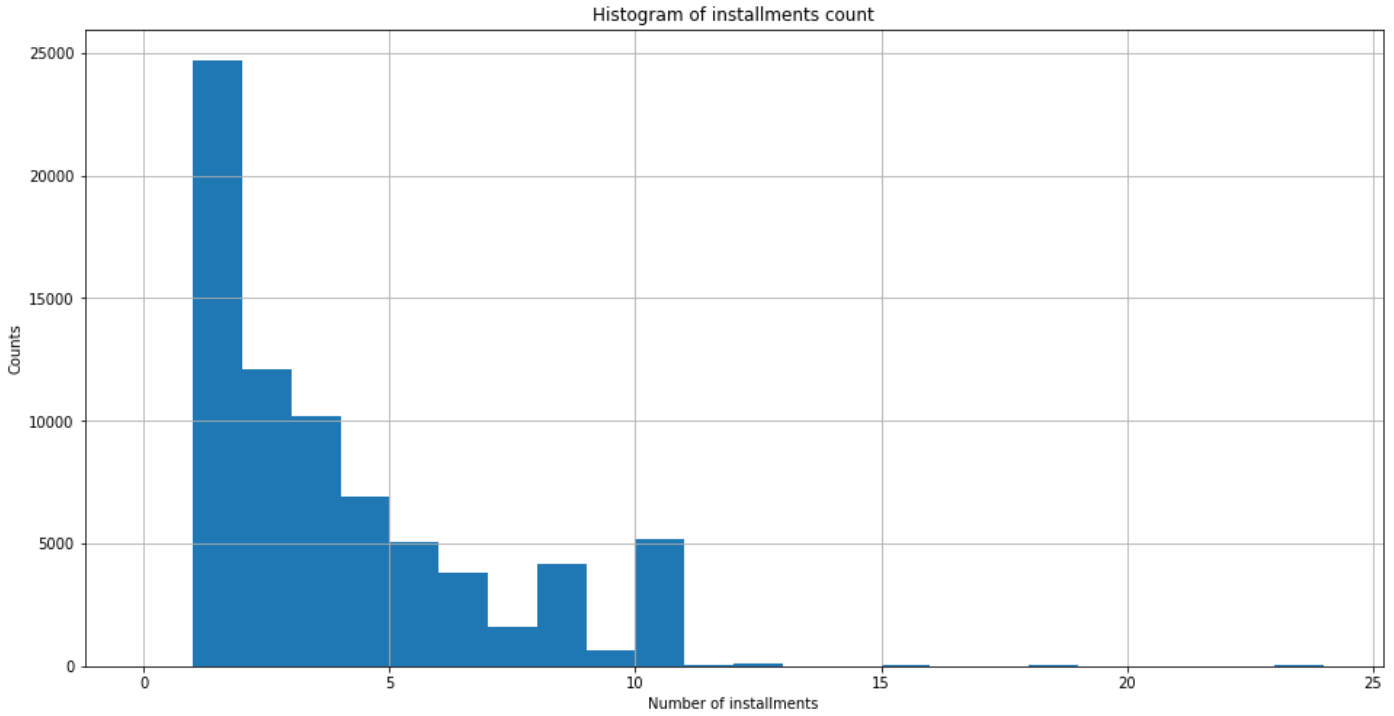


Données sur les paiements

nombre de paiements/commande

- La grande majorité des paiements < C.A. moyen
- Recours important aux facilités de paiement

-> Politique spécifique sur les coûts fixes des transactions (throughputs des serveurs, card scheme fees)



Segmentation Manuelle

RFM Score

Modélisation RFM

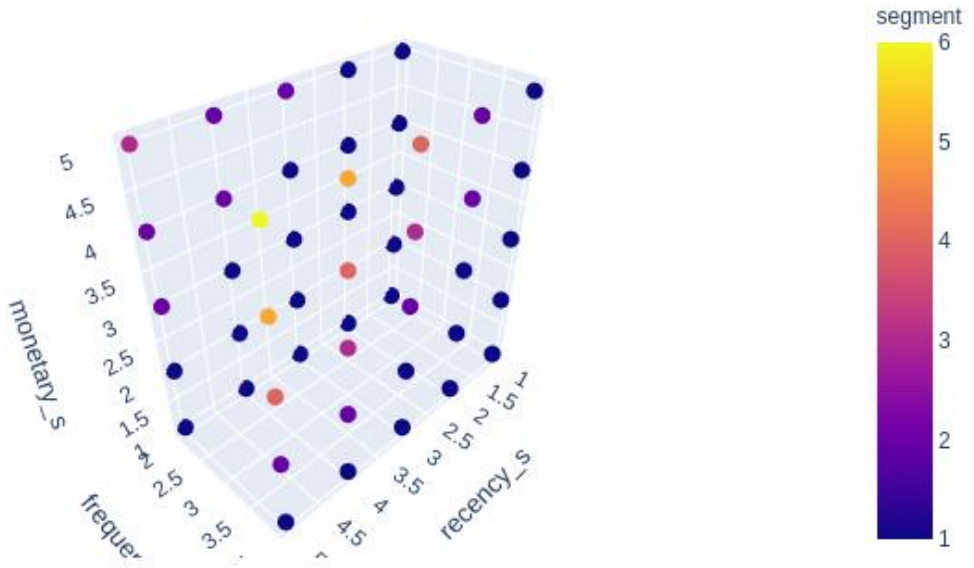
Nature

- R pour Récence du client
  - Âge de la dernière transaction -> valeur prospective et stratégique
- F pour Fréquence du client
  - Nombre de transactions, indicateur objectif de la fidélité -> influence directe sur le C.A.
- M pour Monétaire (valeur)
  - Valeur des transactions -> influence directe sur le C.A.
  - Calculée comme la somme de toutes les transactions

Modélisation RFM

Constitution

- Score agrégé = produit des variables scores de chaque client
    - Perte de l'information "spatiale" de chaque facteur.
    - Appréciation issue de ce score ne peut se faire que sur une échelle unique globale.
- Segment = Niveau global du potentiel de chiffre d'affaire



Modélisation RFM

Segments

Segment 1: **l'étincelle**: rfm_score in [1,16]

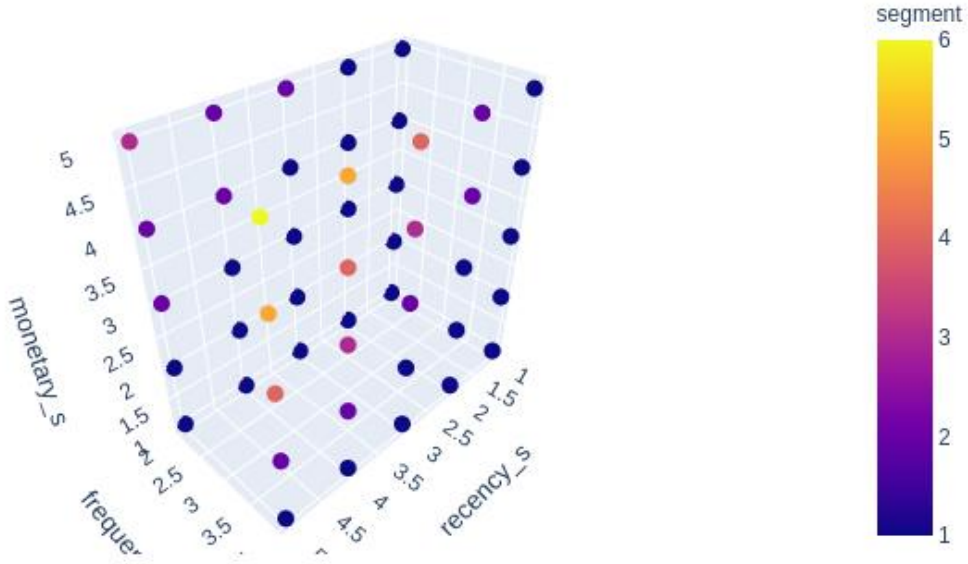
Segment 2: **la buchette [d'allumette]**: rfm_score in [17,32]

Segment 3: **le briquet**: rfm_score in [33,48]

Segment 4: **le flambeau**: rfm_score in [49,64]

Segment 5: **le feu de camp**: rfm_score in [65,80]

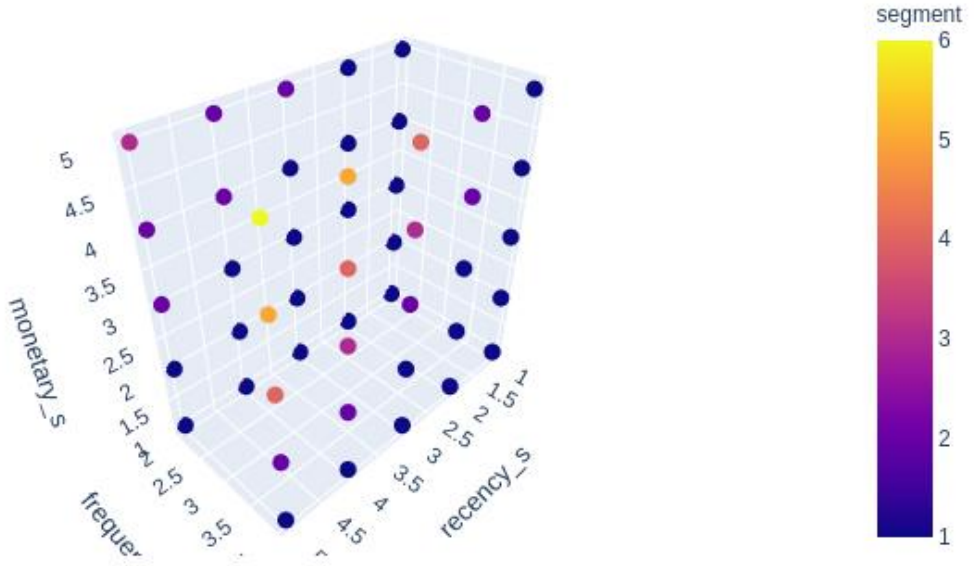
Segment 6: **le dragon**: rfm_score in [81, 100]



Modélisation RFM

Limitations de la modélisation

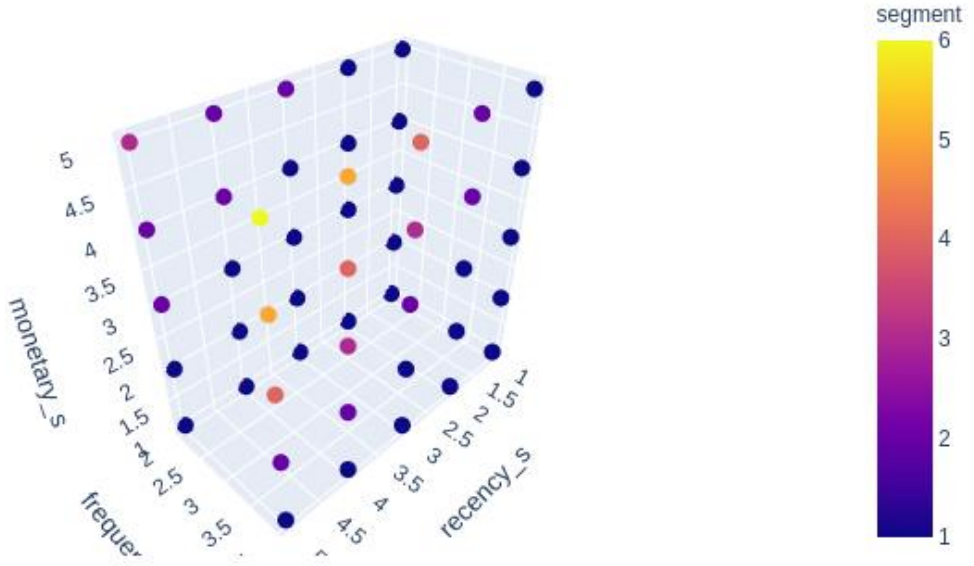
- Perte d'information "spatiale" (agrégation par produit)
- Fréquence inexploitable
  - Distribution quasi unimodale vs Effectif de 9 modalités



Modélisation RFM

Limitations de la modélisation

- Perte d'information "spatiale" (agrégation par produit)
- **Fréquence inexploitable**
  - Distribution quasi unimodale vs Effectif de 9 modalités
- **Segmentation non-pertinente !**



Feature engineering

Variables additionnelles injectées

- Score de satisfaction
  - moyennée sur toutes les commandes du client
- Nombre d'objets achetés
  - calculée comme la somme de tous les objets achetés

Segmentation automatique

KMeans , DBScan

Segmentation KMeans

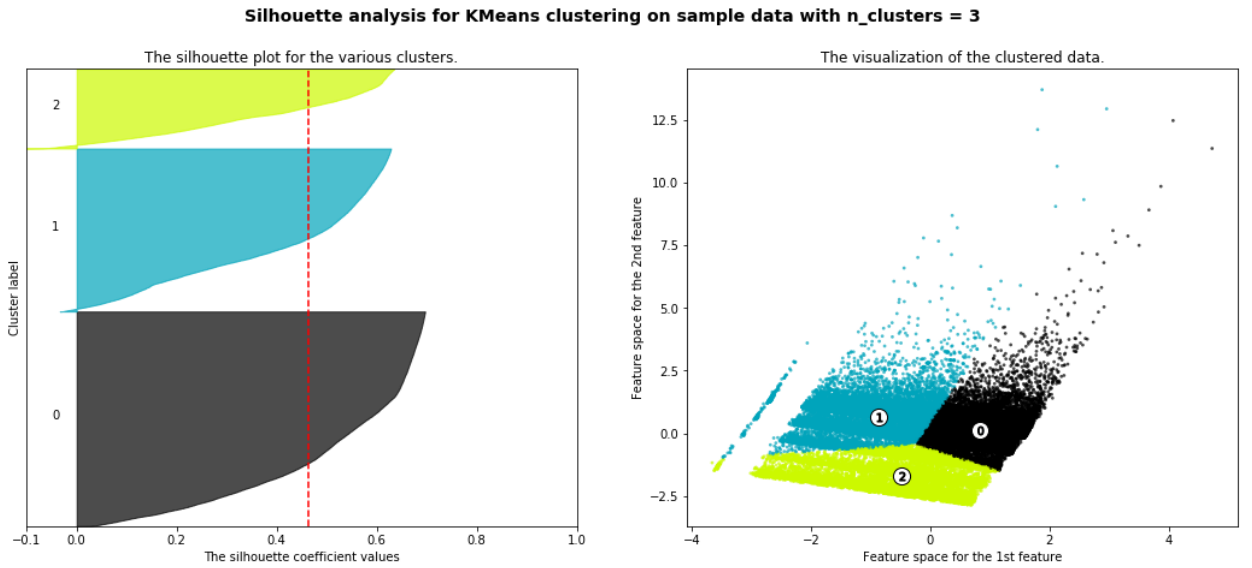
Constitution

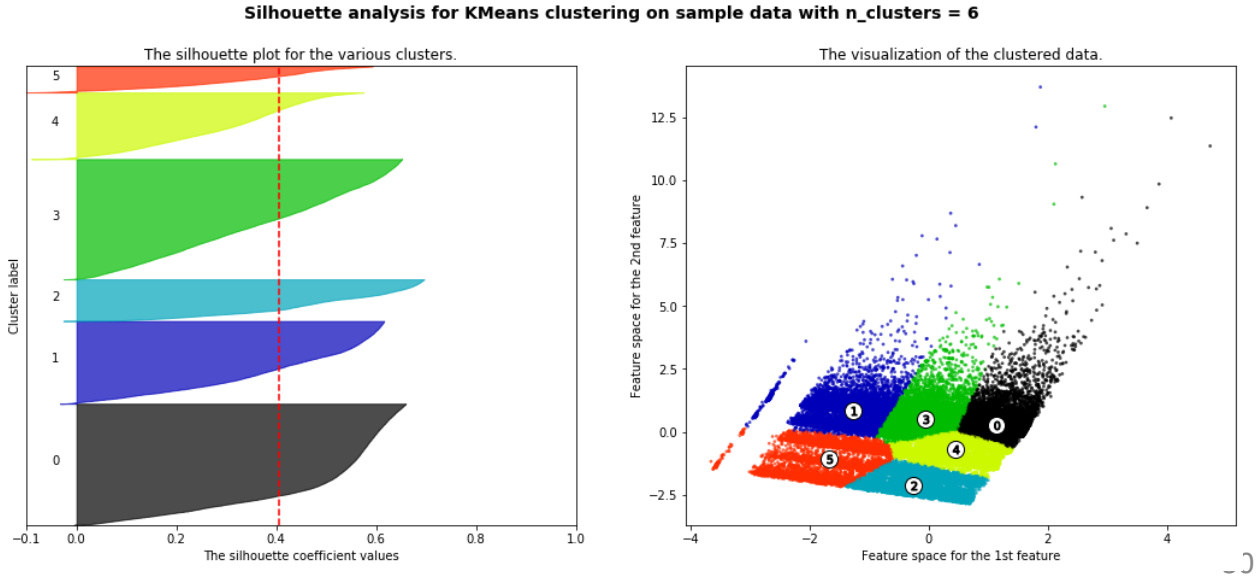
- Définition: proximité par rapport à un point représentatif parmi K points représentatifs
- K dans {2,3,4,5,6}
- Evaluation du K optimal par le score silouhette (quantification relative de l'appartenance effective à son cluster)

Segmentation KMeans

Résultats

- Appliqué sur un espace à 4 dimensions (récence, monetaire, review_score, items)
- Les meilleures scores silouhette:
  - K=3 ; silouhette score = 0.464
  - K=6; silouhette score = 0.407
- La valeur K=6 est retenue
  - Faible fluctuation de la taille des silouhettes
  - Possibilité d'interprétation plus diverse





Segmentation KMeans

Segments

Segment 0: le dragon déchainé

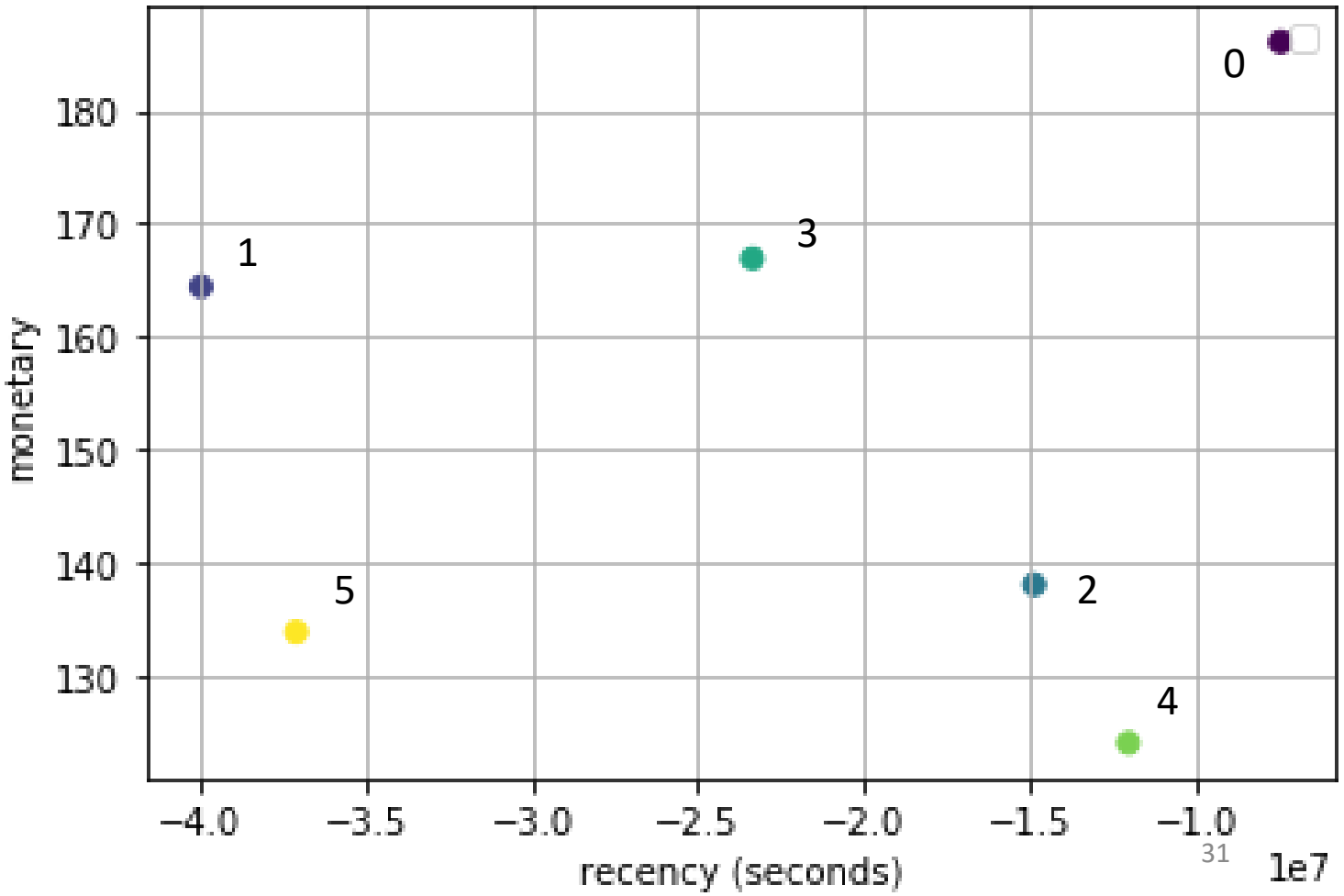
Segment 1: le phoenix endormi

Segment 2: la braise fraîche

Segment 3: le feu de camp interrompu

Segment 4: l'étincelle

Segment 5: l'allumette fumante



Segmentation DBScan

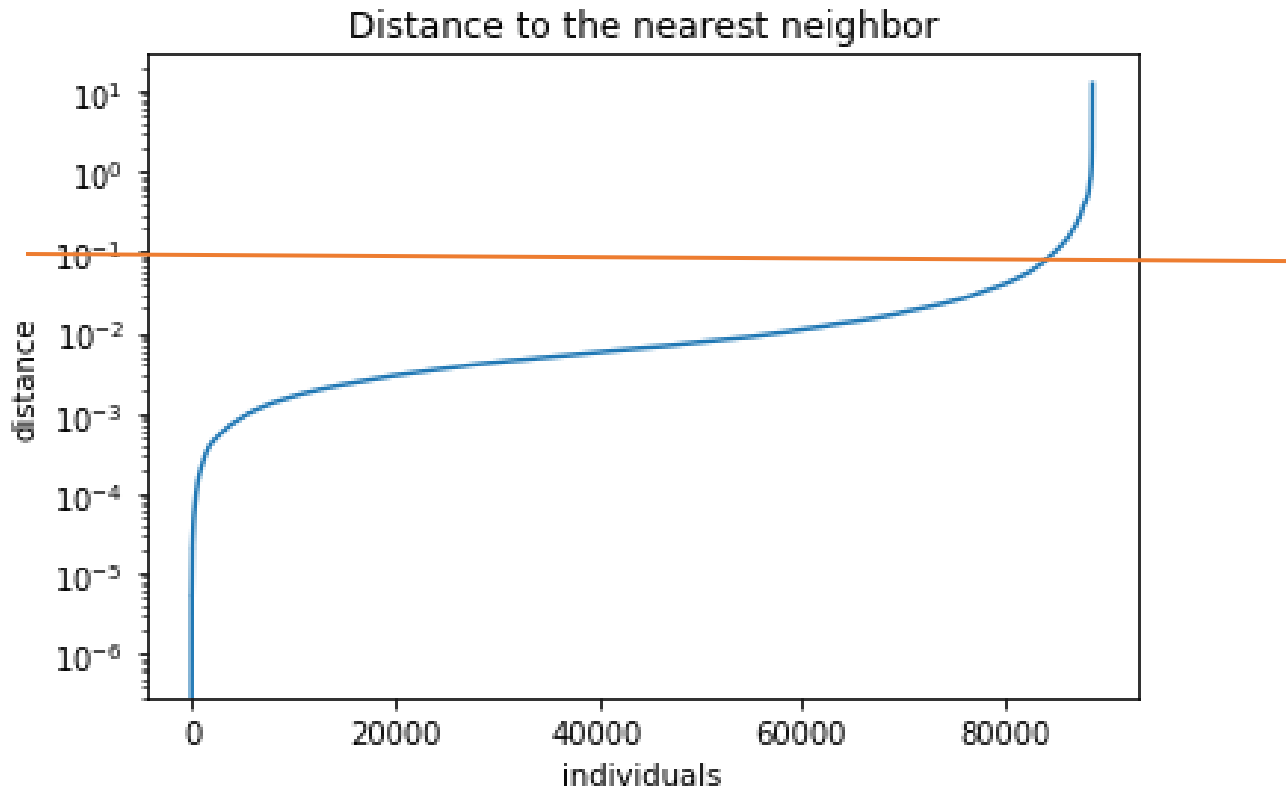
Constitution

- **Définition: proximité avec d'autres individus, séparation en densités locales**
- Pré-calcul du rayon de boule optimal

Segmentation DBScan

Constitution

- Définition: proximité avec d'autres individus, séparation en densités locales
- **Pré-calcul du rayon de boule optimal**



Segmentation DBScan

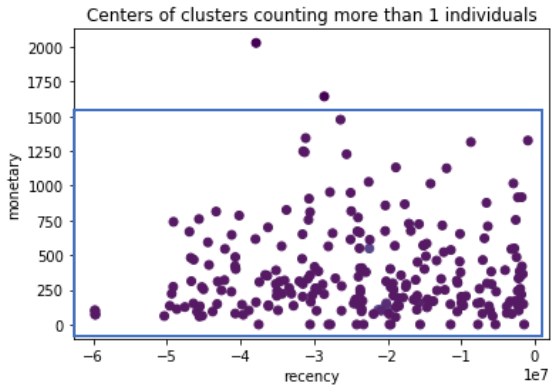
Calcul

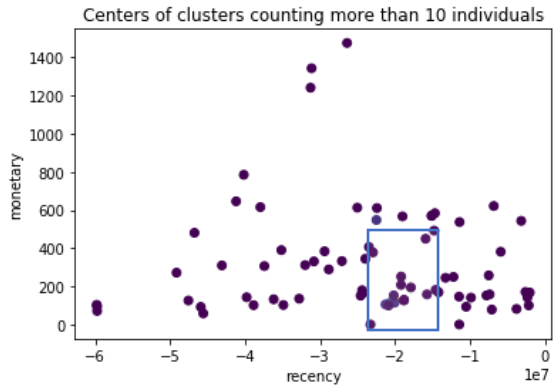
- Appliqué sur un espace à 4 dimensions (récence, monétaire, review_score, items)
- epsilon = 0.1, min_samples = 5

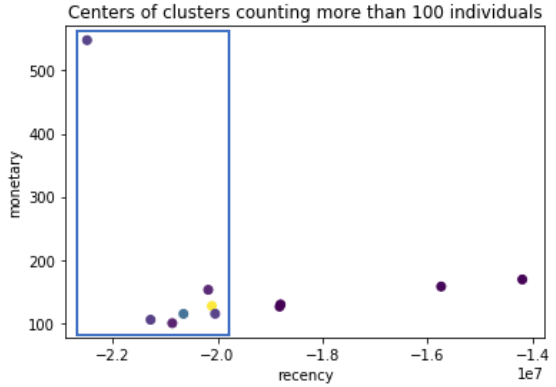
Segmentation DBScan

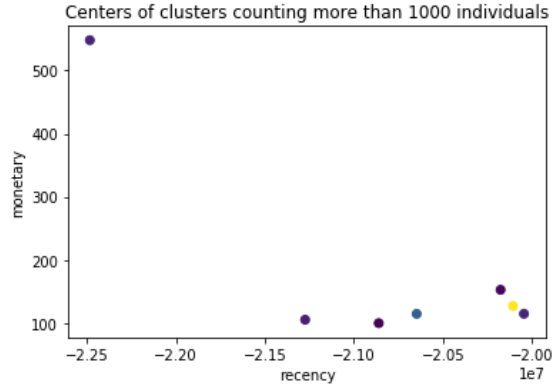
Calcul

- Appliqué sur un espace à 4 dimensions (récence, monétaire, review_score, items)
- epsilon = 0.1, min_samples = 5
- Densité trop importante du nuage





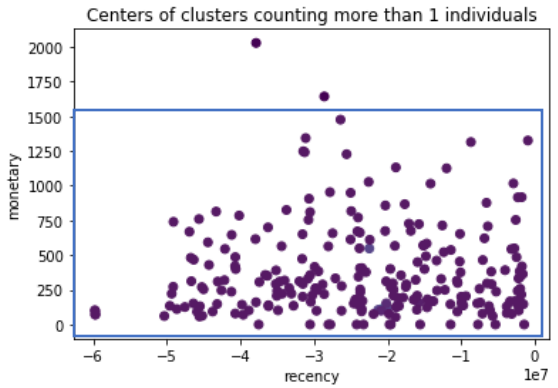


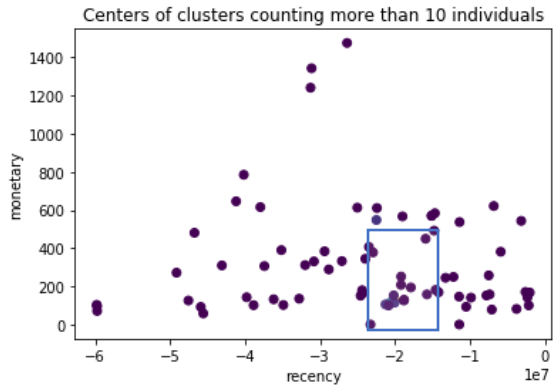


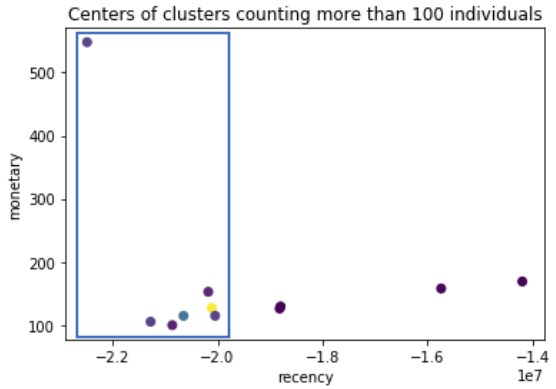
Segmentation DBScan

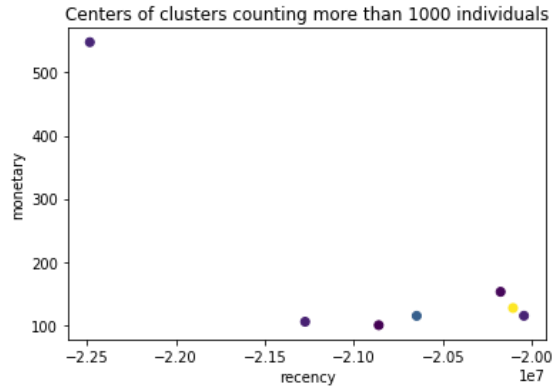
Calcul

- Appliqué sur un espace à 4 dimensions (récence, monétaire, review_score, items)
- epsilon = 0.1, min_samples = 5
- **Densité trop importante du nuage**
- Modélisation incompatible avec la structure des données!









Bilan des essais de segmentation

- Segmentation manuelle RFM
- Segmentation KMeans
- Segmentation DBScan

Bilan des essais de segmentation

- Segmentation manuelle RFM: Fréquence inexploitable
- Segmentation KMeans (6 clusters)
- Segmentation DBScan: Incompatible avec la structure des données

Stabilité des segments

KMeans

Stabilité des segments

Procédure

- Base initiale B_0; Clusterer[ing] intial C_0 : entrainement sur et segmentation de B_0.

- Bases futures {B_i} obtenues par extension incrémentales successives
- Clusterers futures {C_i}: entrainements sur {B_i}

- Clusterings prospectifs {C_0_i} : segmentation de B_i par C_0
- Clusterings futures {C_i_i} : segmentation de B_i par C_i

- Analyse dynamique de l'écart ({C_0_i} , {C_i_i})

Stabilité des segments

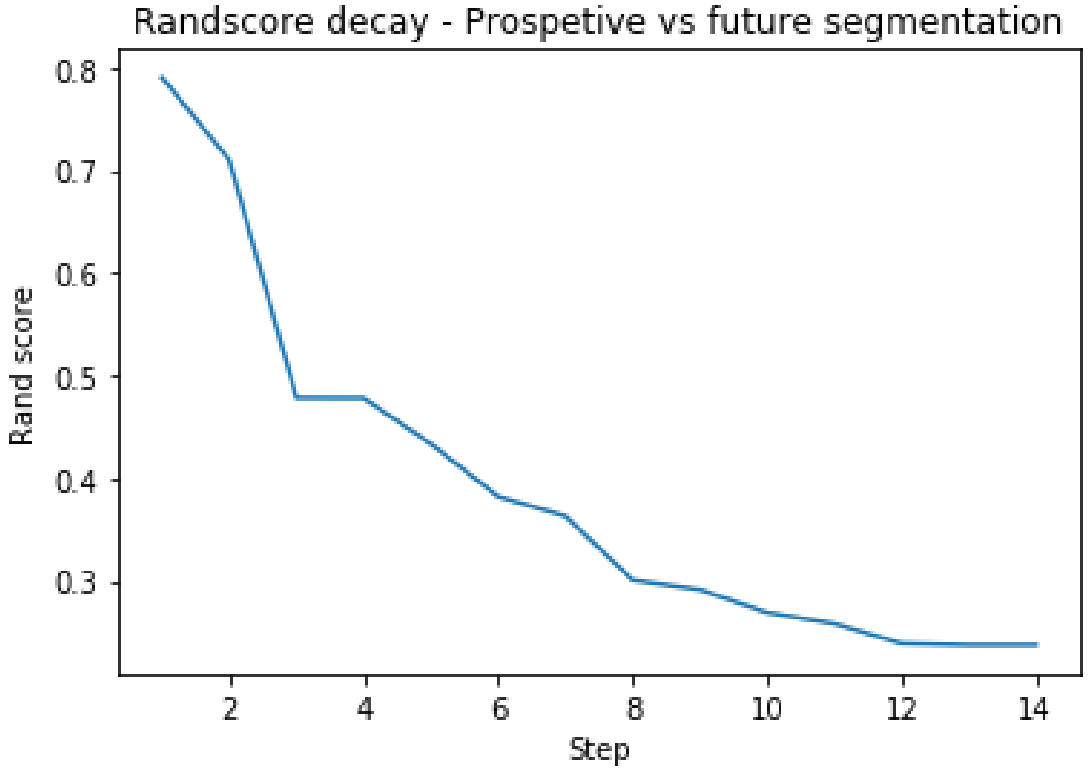
Résultat

- B_0, de taille 12 mois
- KMeans (n_clusters = 6)
- Incrément = 1 mois

Stabilité des segments

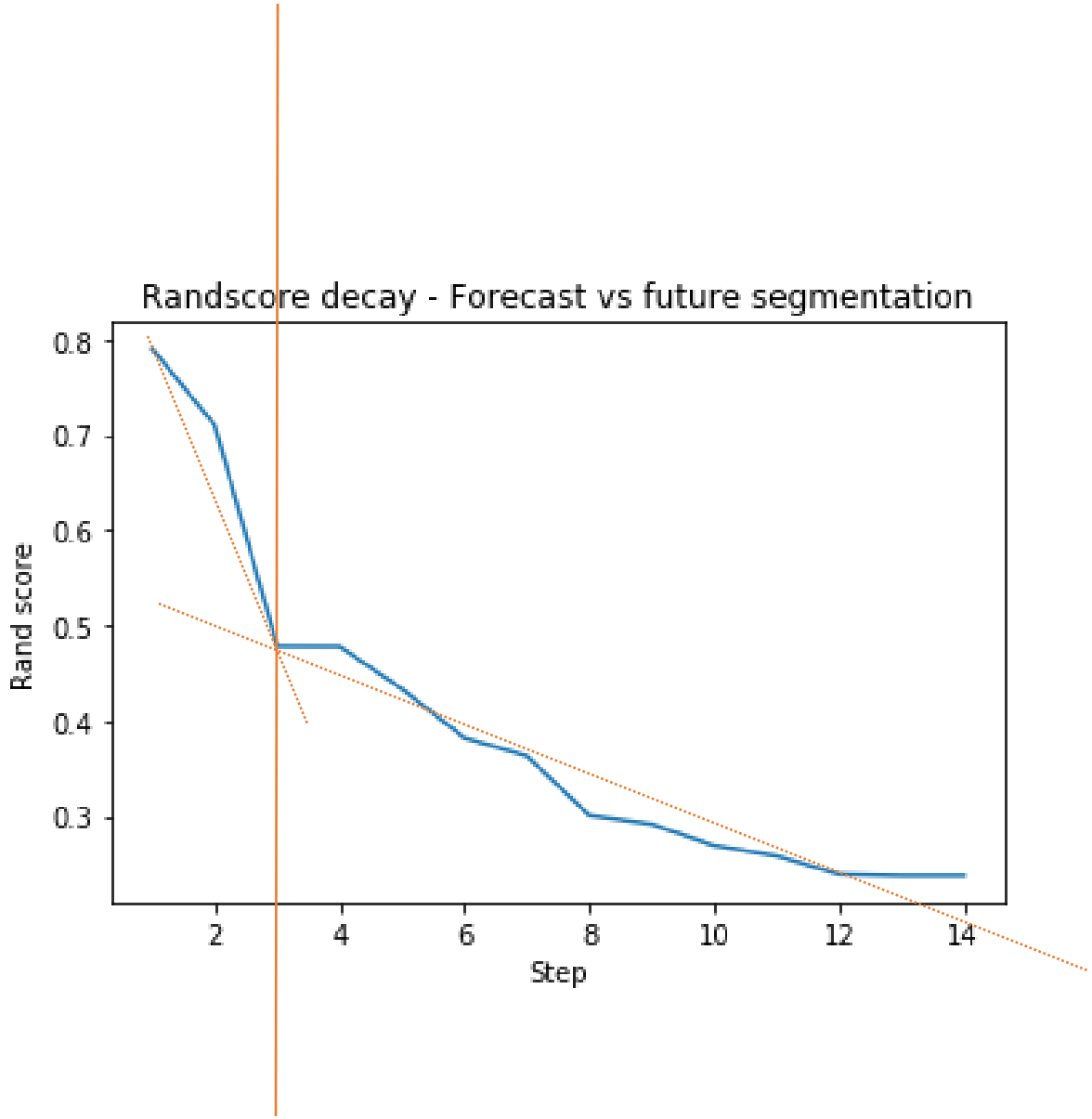
Résultat

- B_0, de taille 12 mois
- KMeans (n_clusters = 6)
- Incrément = 1 mois
- **Analyse comparative par *adjusted rand_score***



Contrat de Maintenance

- Rupture de similarité prospectif-futur
- Fréquence de mise à jour de **3 mois**



Merci pour votre attention

Disponible pour des questions/réponses