P5 - SEGMENTEZ DES CLIENTS D'UN SITE DE E-COMMERCE

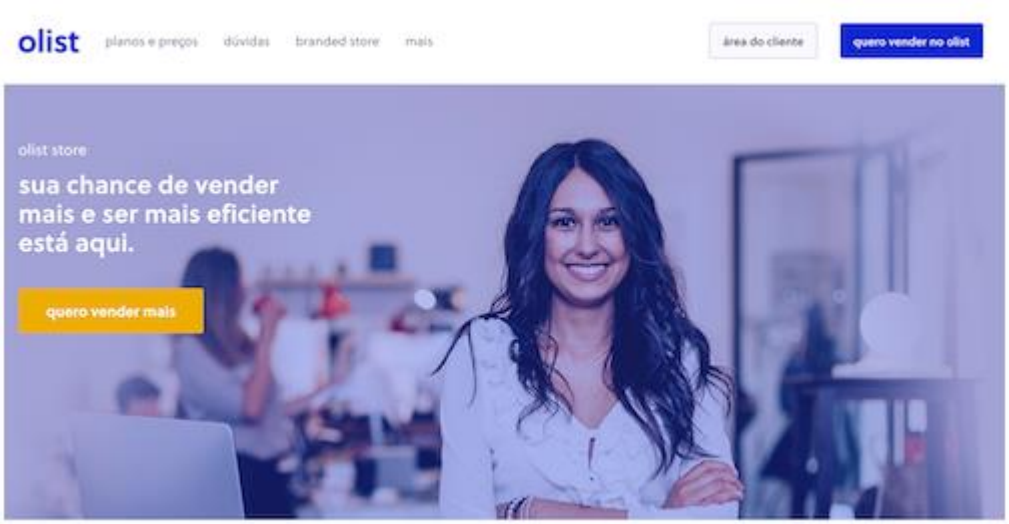
22/07/2021

Etudiant : Luc Rogers
Mentor : Etienne Sanchez

Sommaire

Problématique

Problématique



- Segmentation des clients d'une plateforme de e-commerce

- Objectifs :
  - Fournir à l'équipe marketing une segmentation clients actionnable
  - Proposition de contrat de maintenance (analyse de la stabilité des segments au cours du temps)

Base de données open source :
https://www.kaggle.com/olistbr/brazilian-ecommerce

Nettoyage

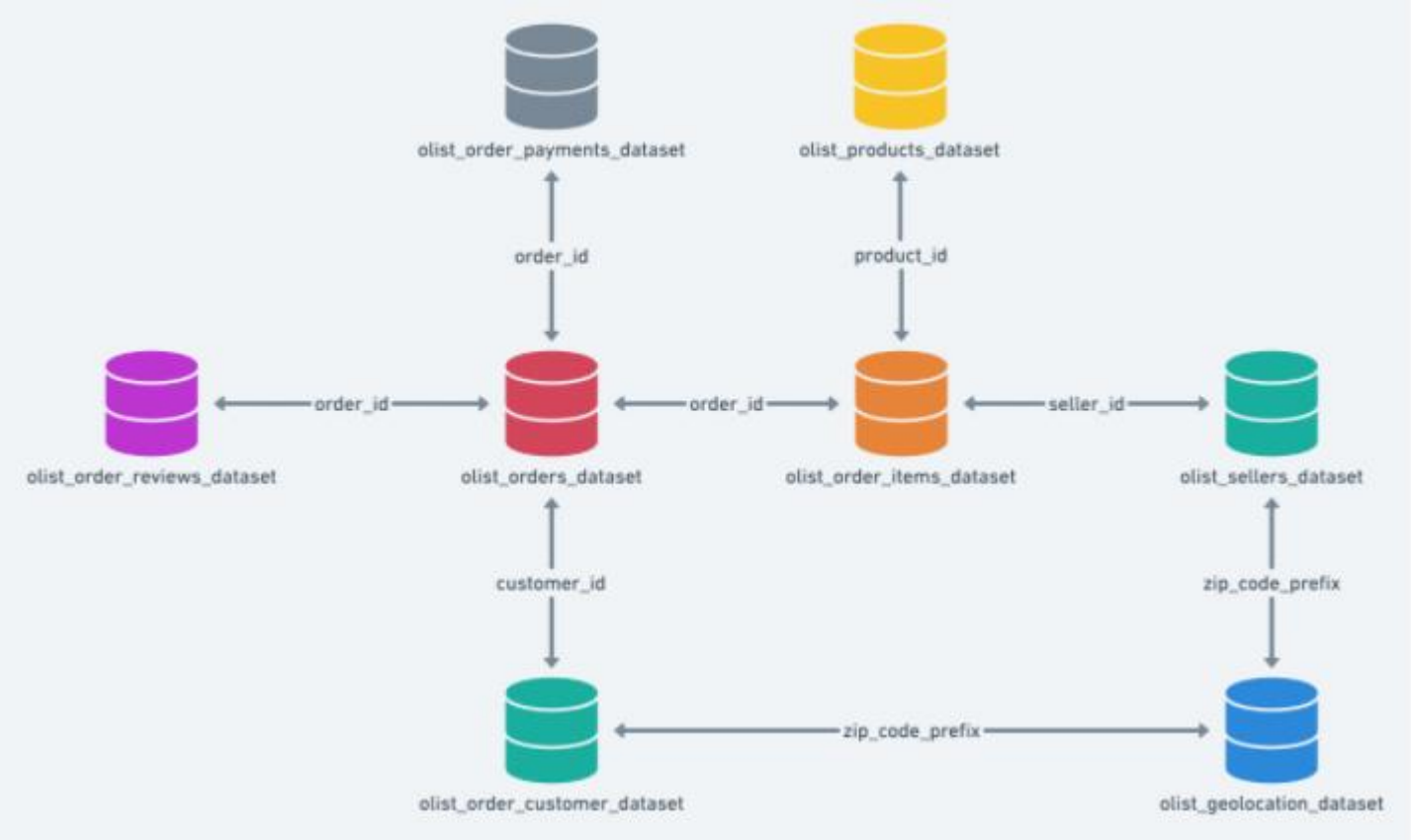
Nettoyage

- Vérification des doublons ➔ 0 doublons
  - Orders, 'order_id'
  - Customers, 'customer_unique_id'
- On ne garde que les ordres ayant abouti (livraison effectuée, 'order-status' = delivered)
- Suppression des produits dont la catégorie n'est pas renseignée
- Supression des ordres avec bug sur la date d'approbation/de livraison
- Supression des quelques lignes avec valeurs manquantes
- Création à la main de catégories plus larges

➔ ***Jeu de données très propre, très peu de nettoyage à réaliser***

Feature engineering

Feature engineering



Feature engineering

- Order reviews:
  - order_id: ID de la commande
  - review_id: ID de l'avis utilisateur
  - review_creation_date: date de création de l'avis
  - review_score: note attribuée par l'utilisateur sur sa commande
- Order payments:
  - order_id
  - payment_installments : en combien de fois l'achat a été effectuée
  - payment_value: montant de l'achat
- Products:
  - product_category_name: catégorie du produit
  - product_category_name_translation: traduction en anglais
- Order items:
  - order_id

Feature engineering

- Customers:
  - customer_id: ID du client par commande (similaire à order_id?)
  - customer_unique_id: ID du client
  - customer_state: état de résidence du client
- Orders:
  - order_id
  - order_purchase_timestamp: data d'achat
  - order_delivered_customer_date: date de livraison
  - order_estimated_delivery_date: date de livraison estimée
- Products:
  - product_category_name: catégorie du produit
  - product_category_name_translation: traduction en anglais

Feature engineering

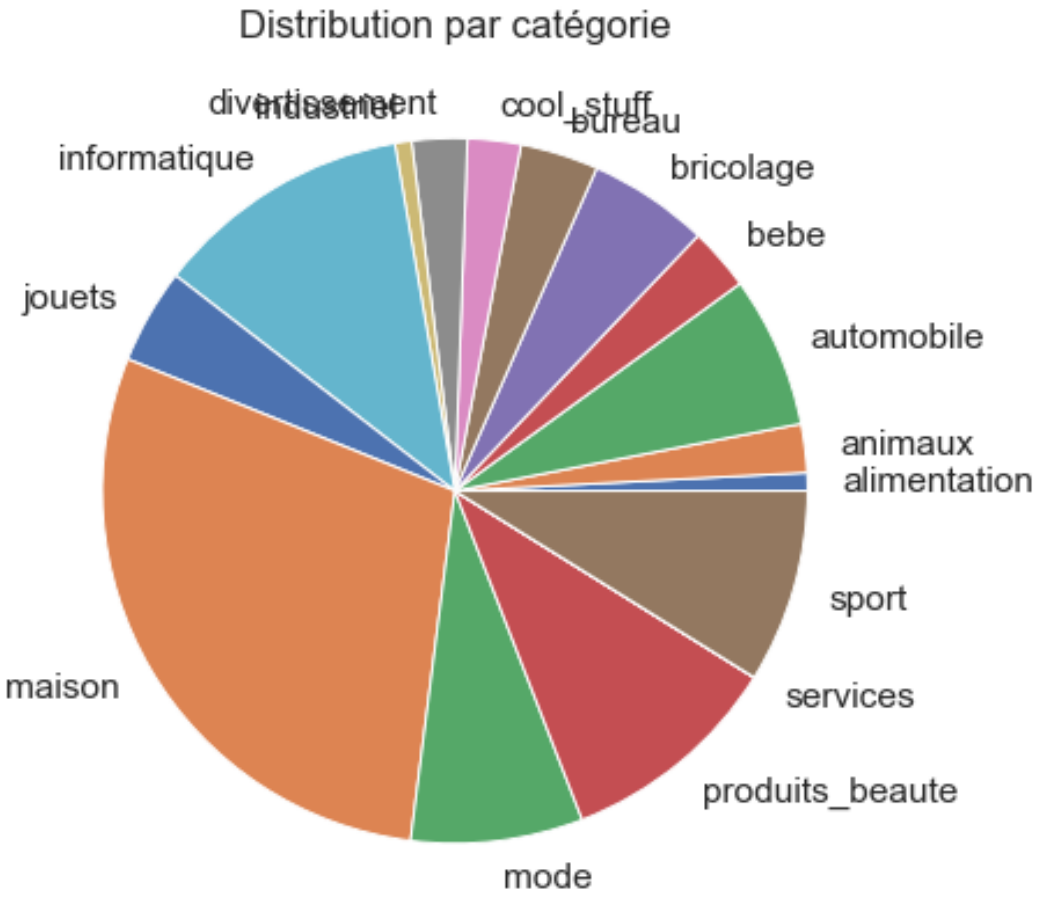
- On réalise un ‘merge’ sur la variable ‘order_id’ des df suivants :
  - orders
  - order_reviews
  - order_payments
  - order_items
- On merge ensuite sur la variable ‘customer_id’ du df:
  - customers
- On réalise enfin un groupby sur la variable ‘customer_unique_id’:
  - Mean:
    - délai_livraison
    - retard_livraison
    - review_score
    - payment_installment
  - Sum:
    - payment_value
    - categories (variables encodées en one hot encoding)
  - First:
    - customer_states (variables encodées en one hot encoding)

Feature engineering

- Variables du jeu de données cleané:
  - Récence (nombre de jours depuis le dernier achat)
  - Fréquence (nombre d'achat sur la période étudiée)
  - Montant (montant total des achats)
  - Délai de livraison
  - Retard de livraison (par rapport à la date de livraison estimée)
  - Review score
  - Payment installments (nombre de paiements)
  - Catégories des produits achetés (one hot encoding)
  - Etats de résidence (one hot encoding)

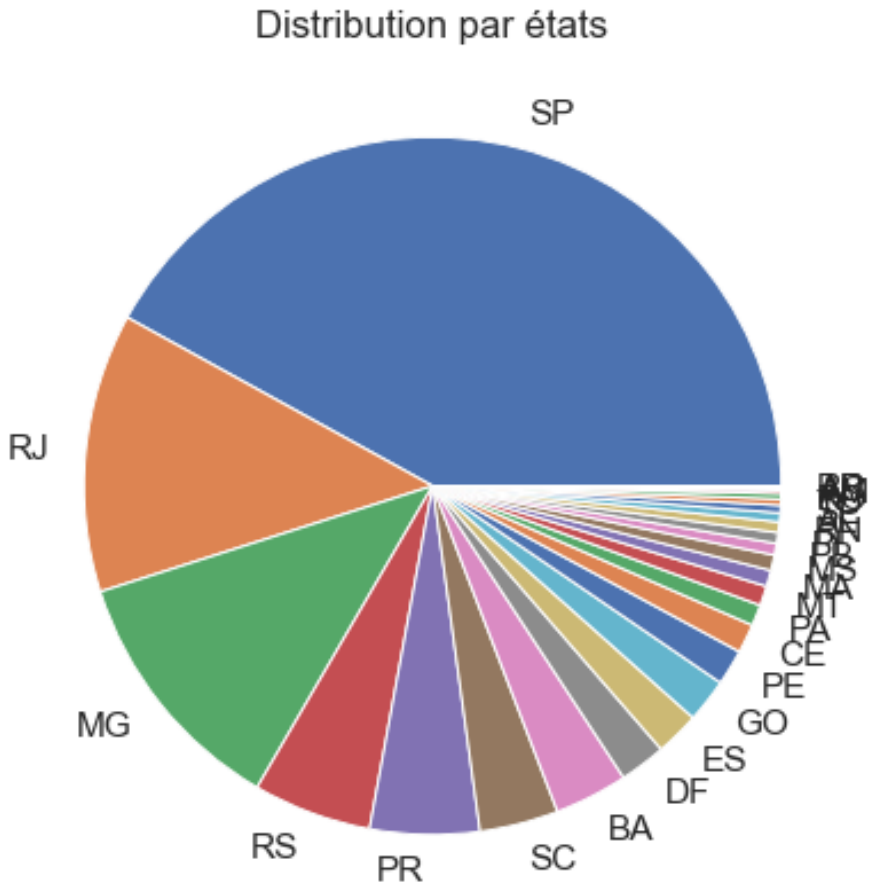
Exploration

Exploration



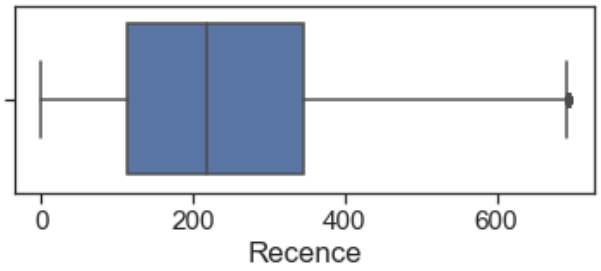
- Des catégories relativement bien distribuées
- On passe de 73 à 16 catégories

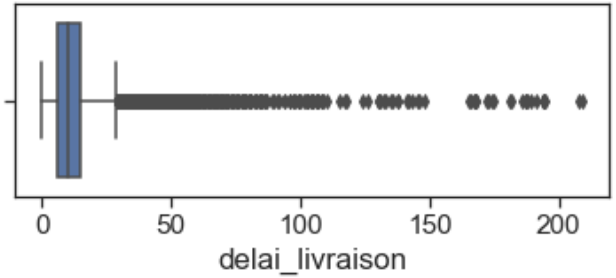
Exploration

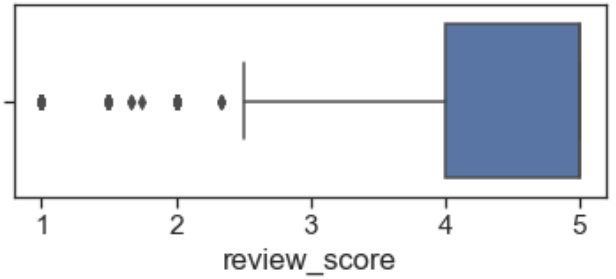


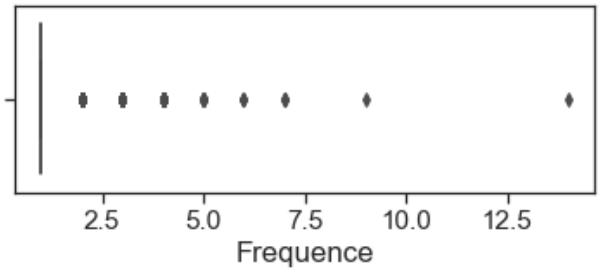
- Distribution majoritaire vers Sao Paulo (SP)

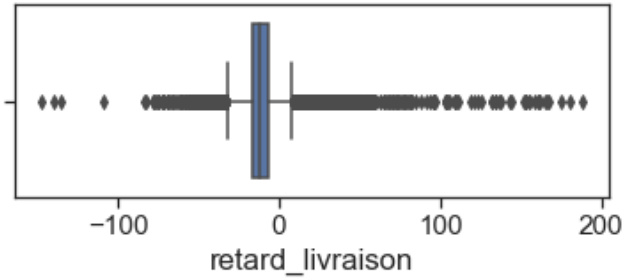
Exploration

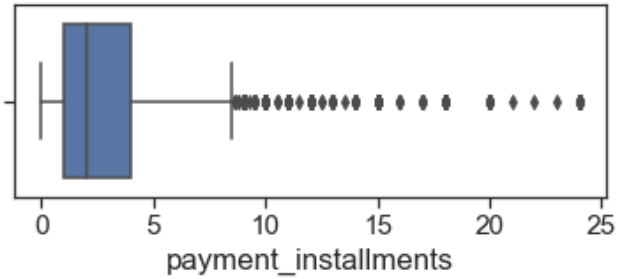


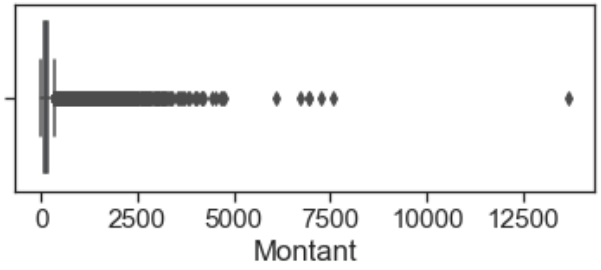




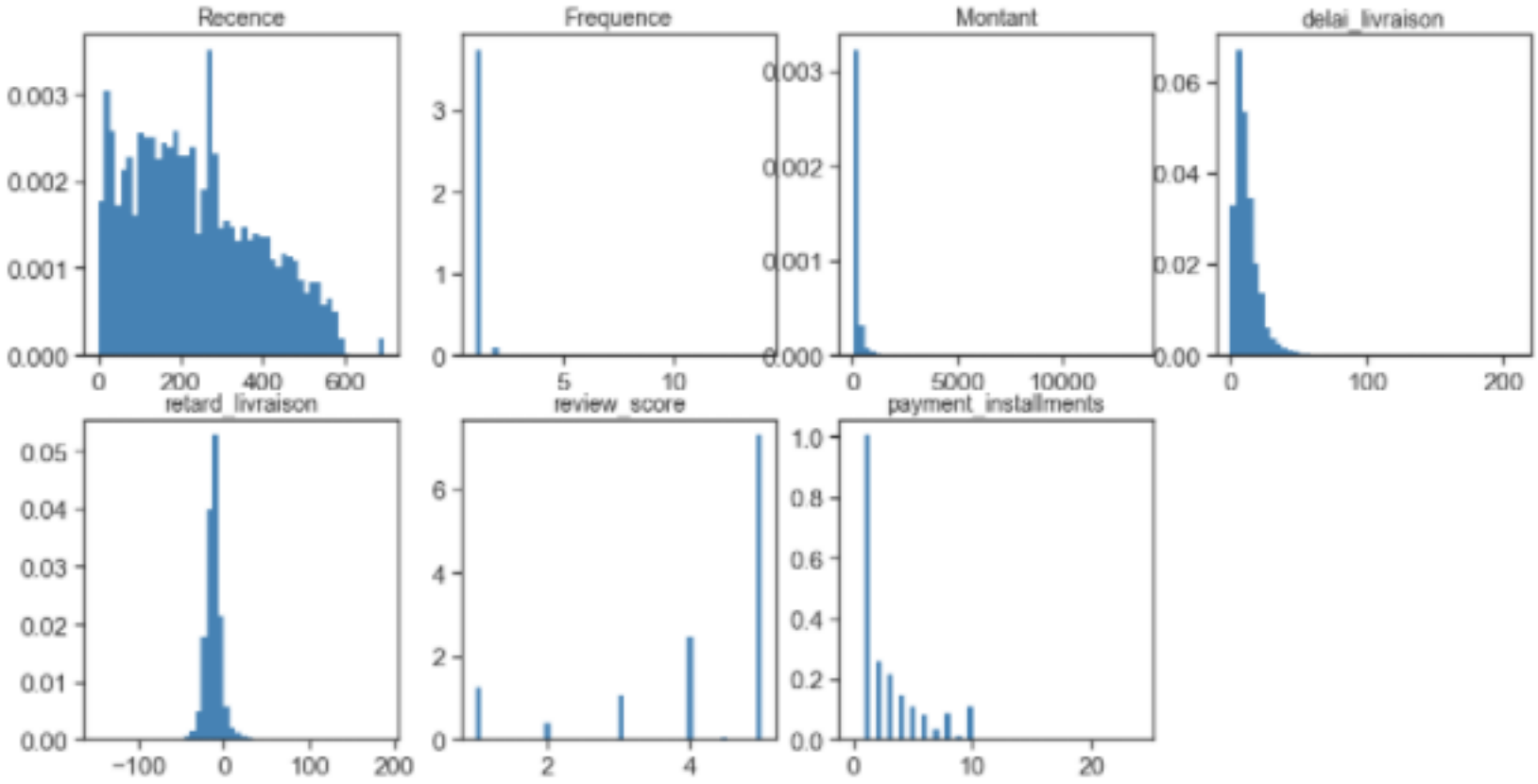




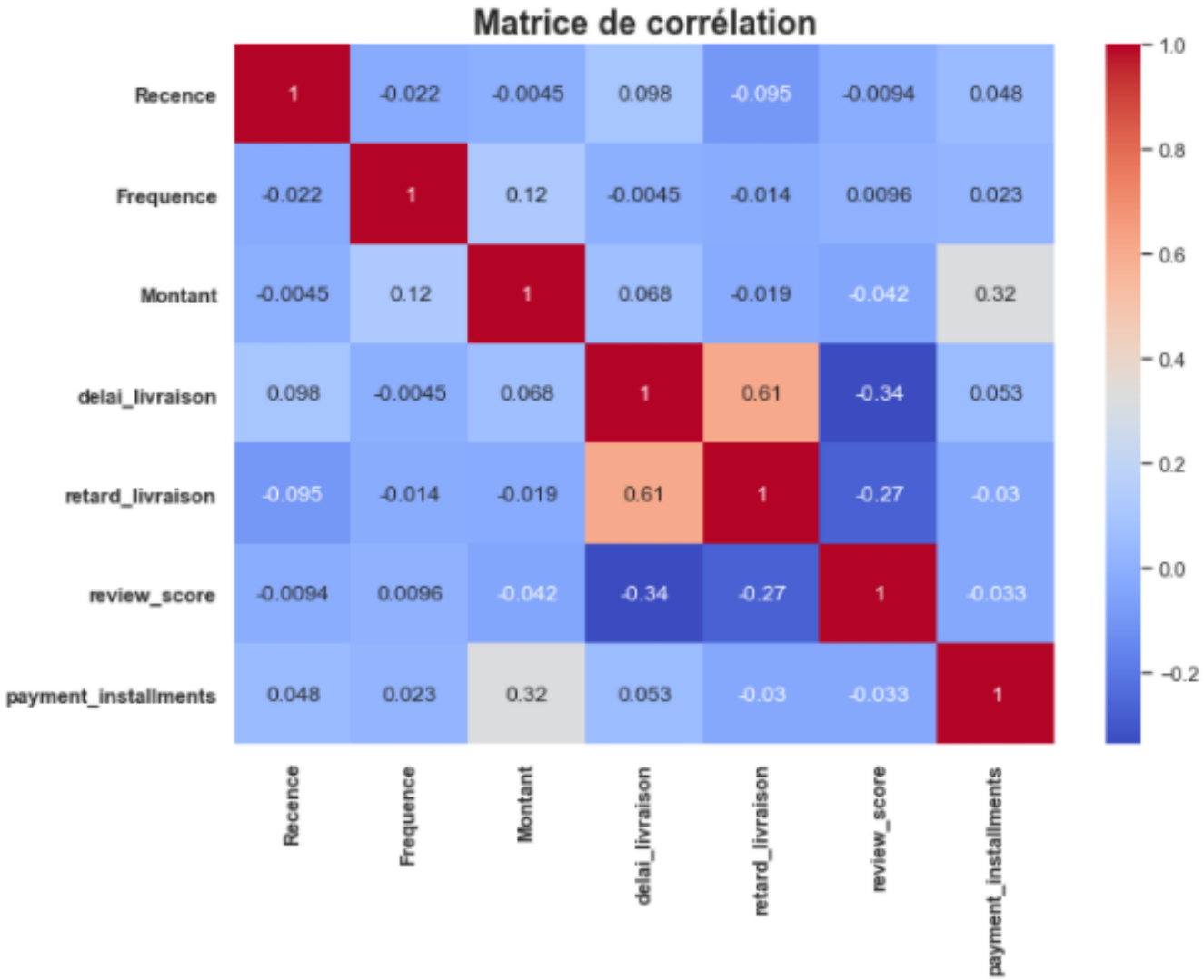




Exploration



Exploration

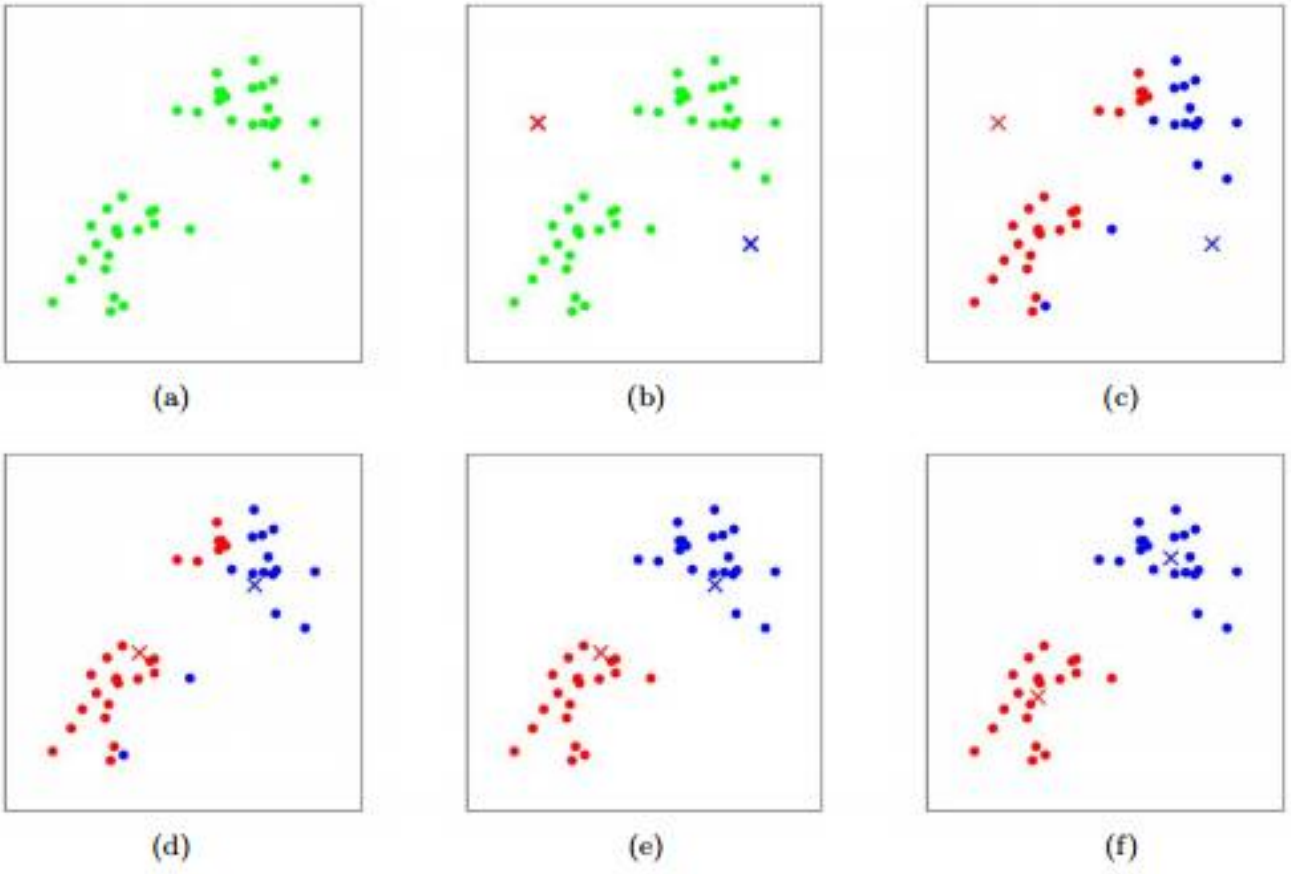


Modélisations

- KMeans
- DBSCAN
- RFM

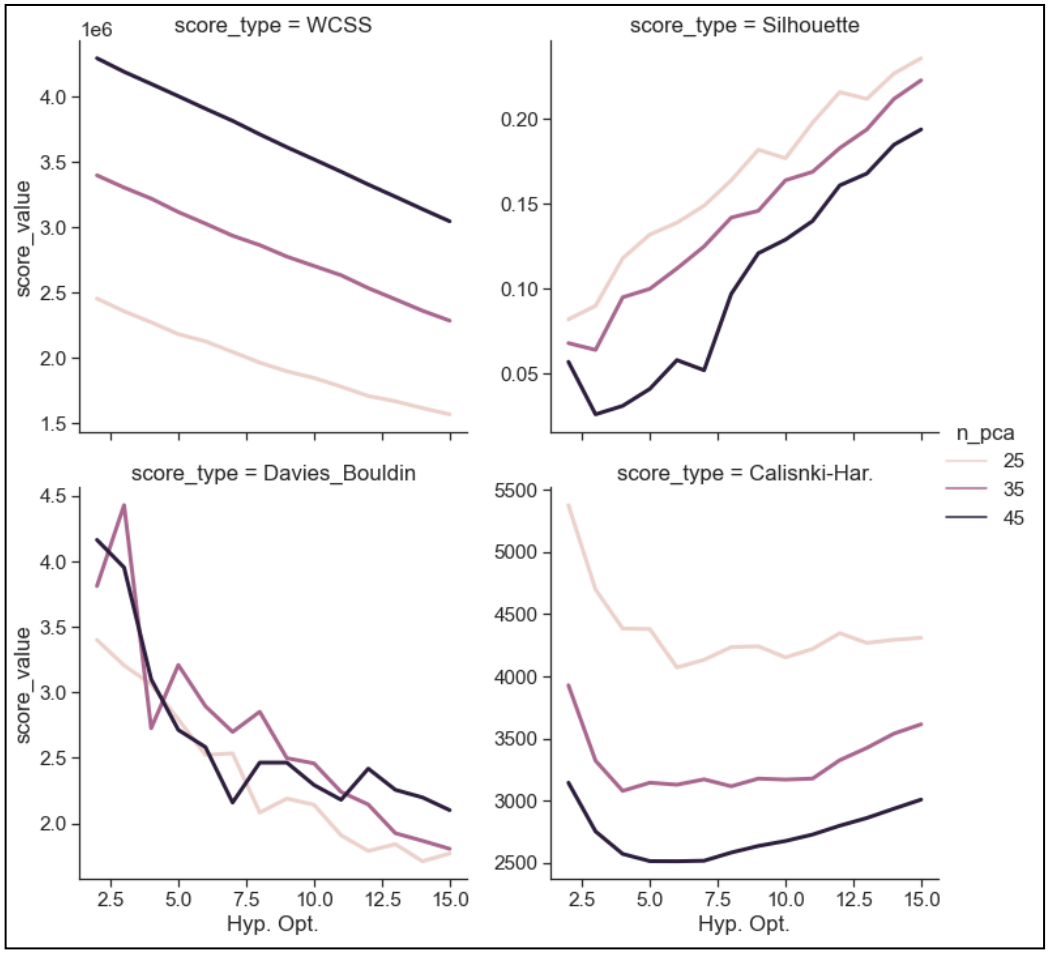
Modélisations – k-means

- Hyperparamètre: k, le nombre de clusters
- Illustration du concept:



Modélisations – k-means

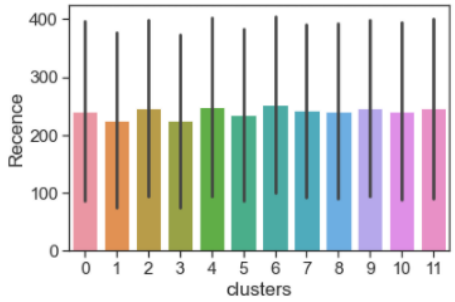
- 1er essai avec toutes les variables

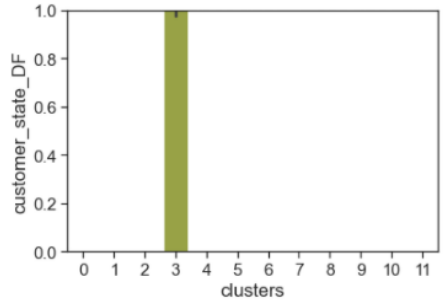


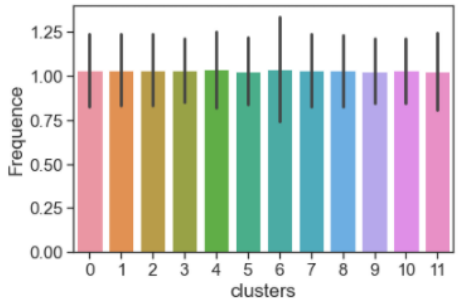
➔ Les scores obtenus ne permettent pas d'identifier un nombre de clusters optimal

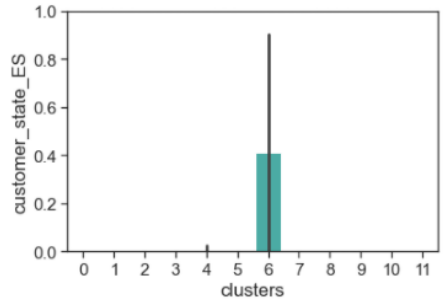
Modélisations – k-means

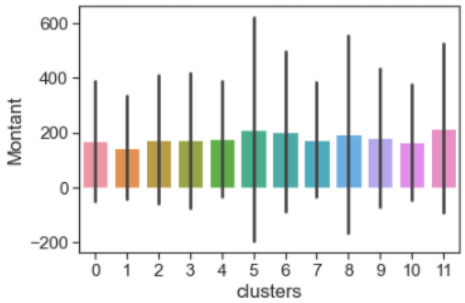
- 1er essai avec toutes les variables, exemple avec k=12

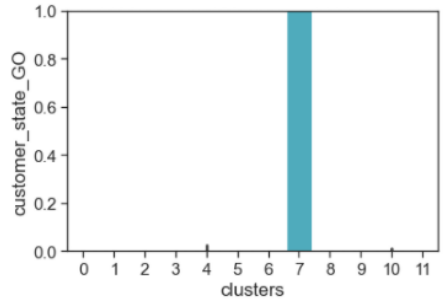








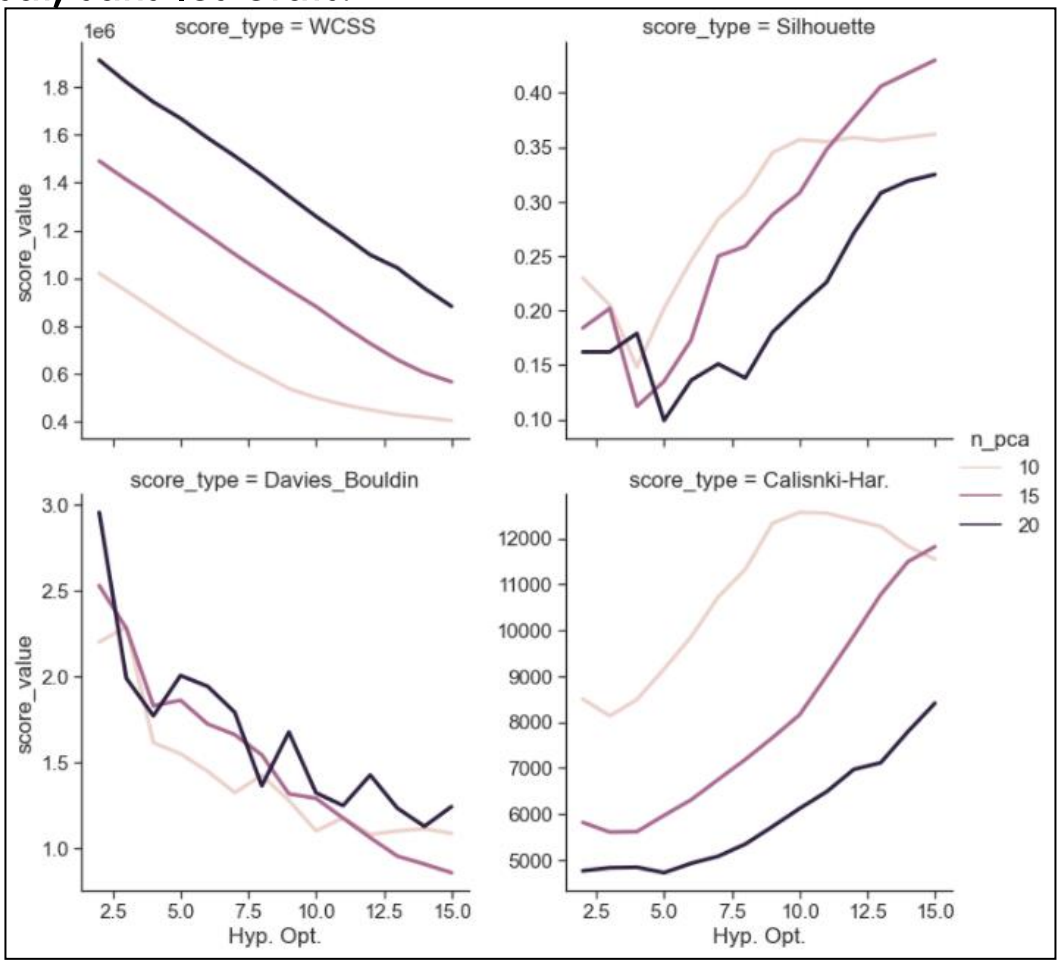




➔ Les clusters sont trop corrélés avec les états

Modélisations – k-means

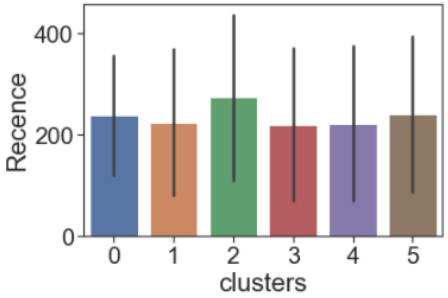
- 2ème essai, sans les états:

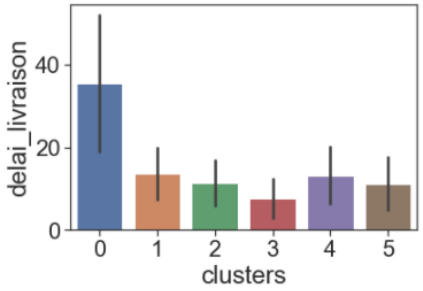


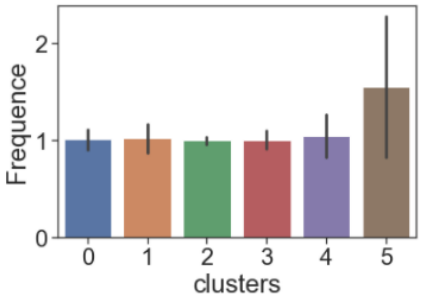
➔ Les scores obtenus ne permettent pas d'identifier un nombre de clusters optimal

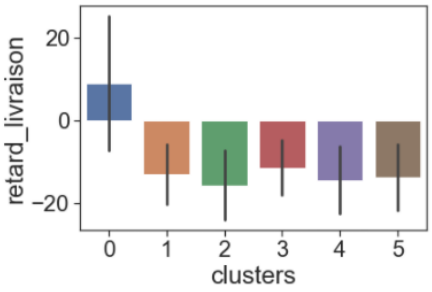
Modélisations – k-means

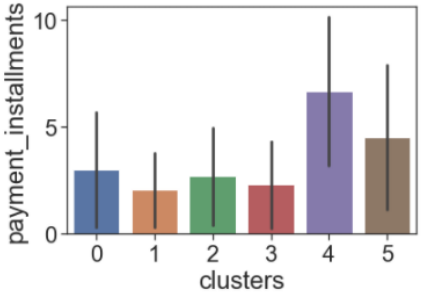
- 2nd essai sans les états, exemple avec k=6

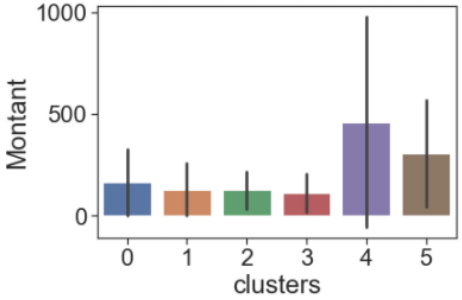


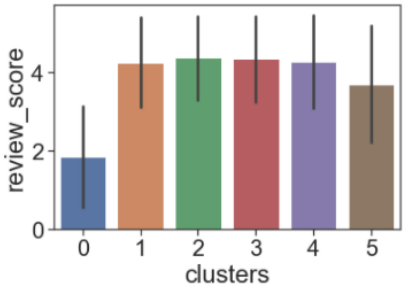








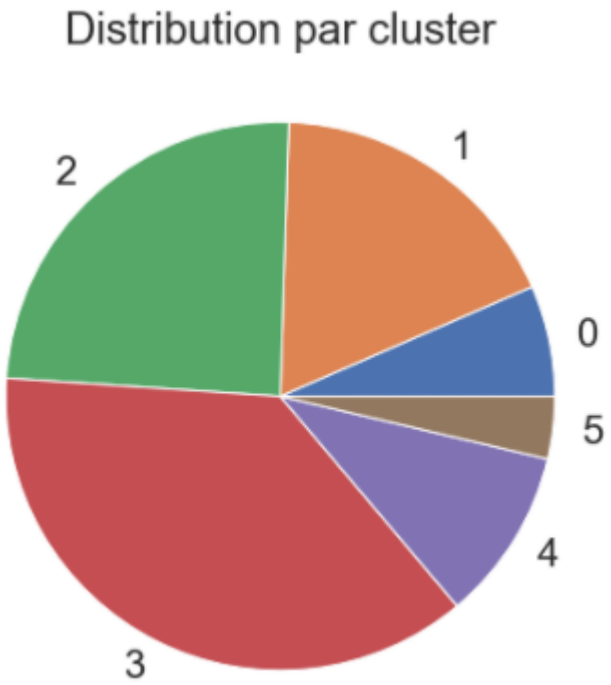




➔ Groupes 4 et 5 à prioriser !

Modélisations – k-means

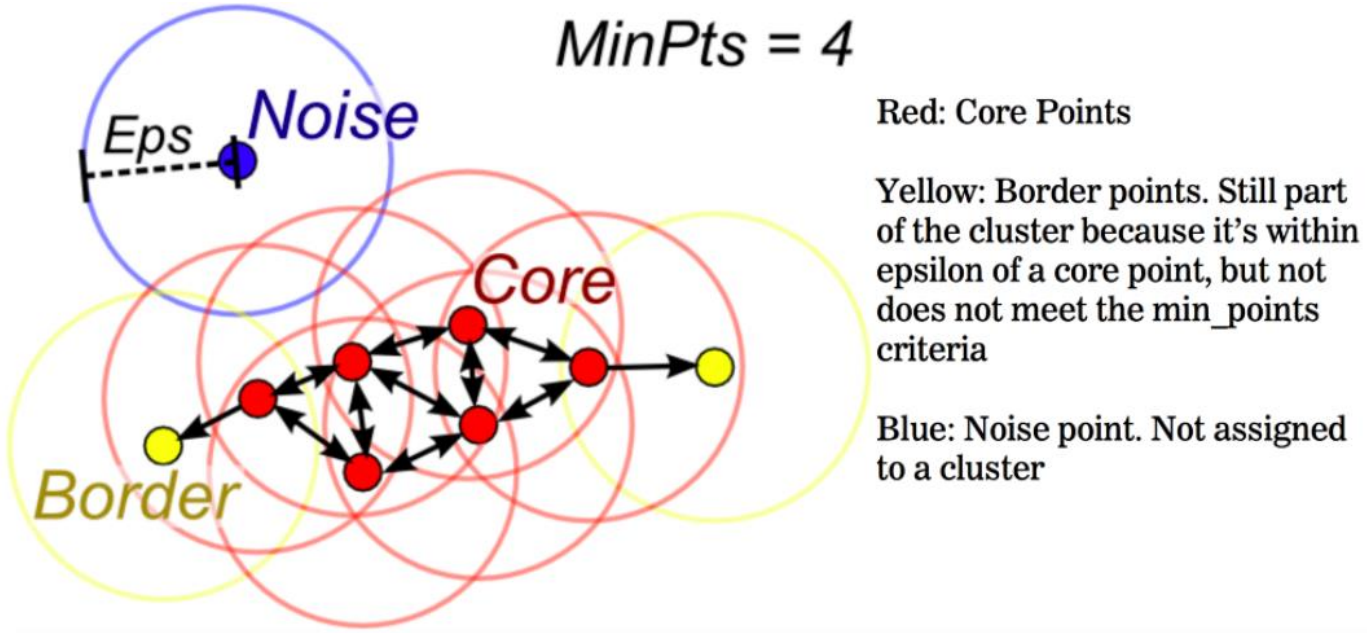
- 2nd essai sans les états, exemple avec k=6



➔ Groupes 4 et 5 à prioriser !

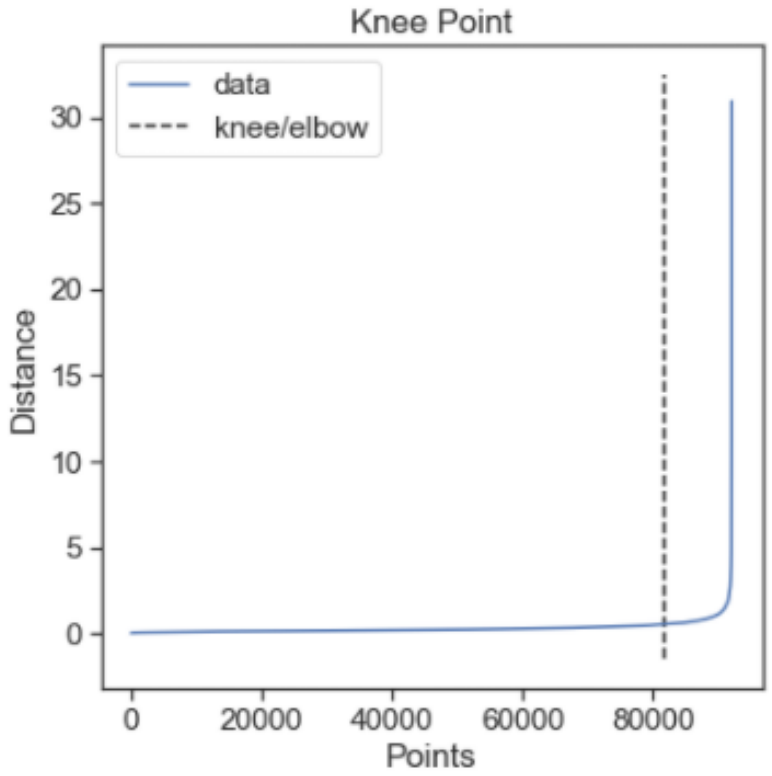
Modélisations – DBSCAN

- Hyperparamètre: epsilon, min_samples
- Illustration du concept:

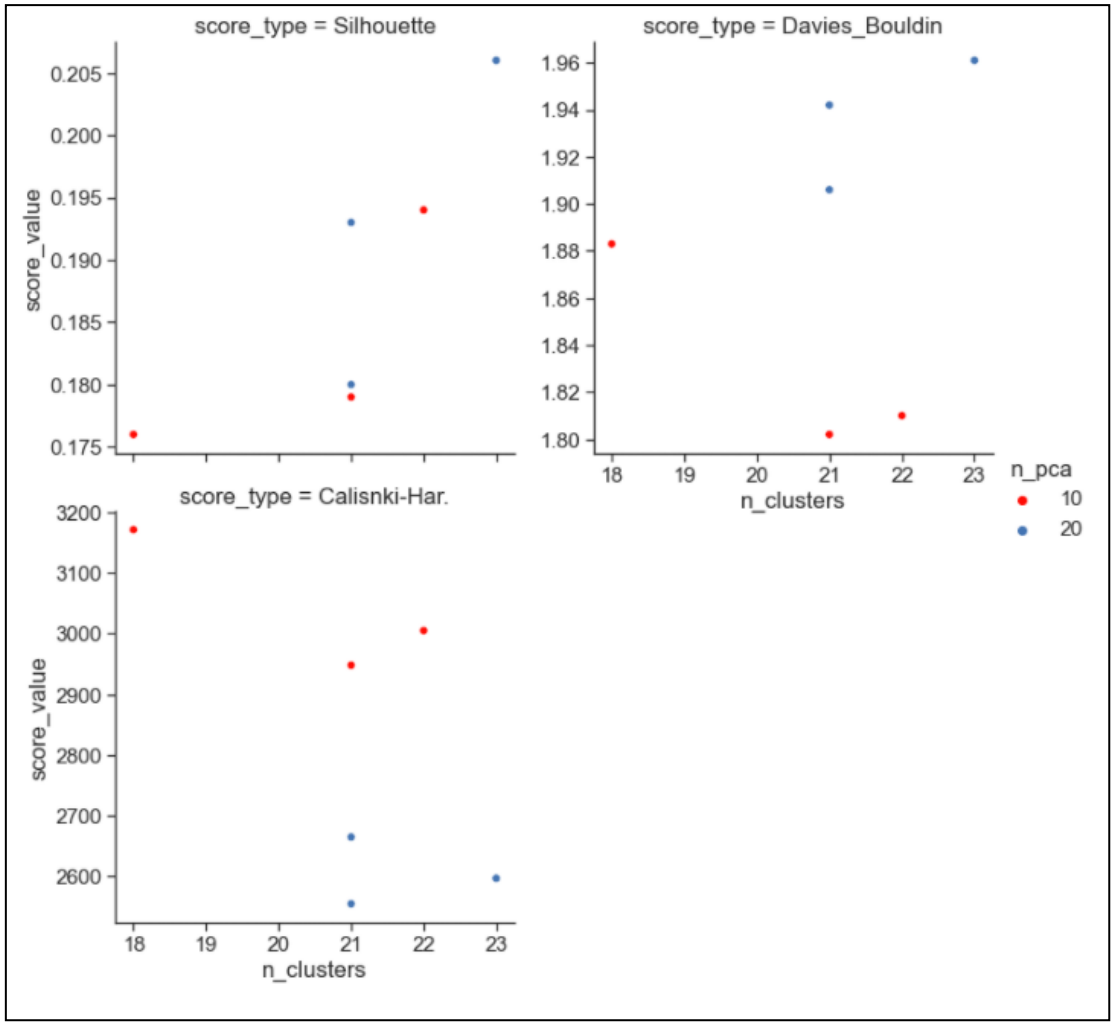


Modélisations – DBSCAN

- Estimation d'epsilon en regardant la distance aux voisins de chaque point
- On fixe epsilon de façon à ce que 90% des points aient au moins un voisin



Modélisations – DBSCAN



min_samples

= [60, 80, 100]

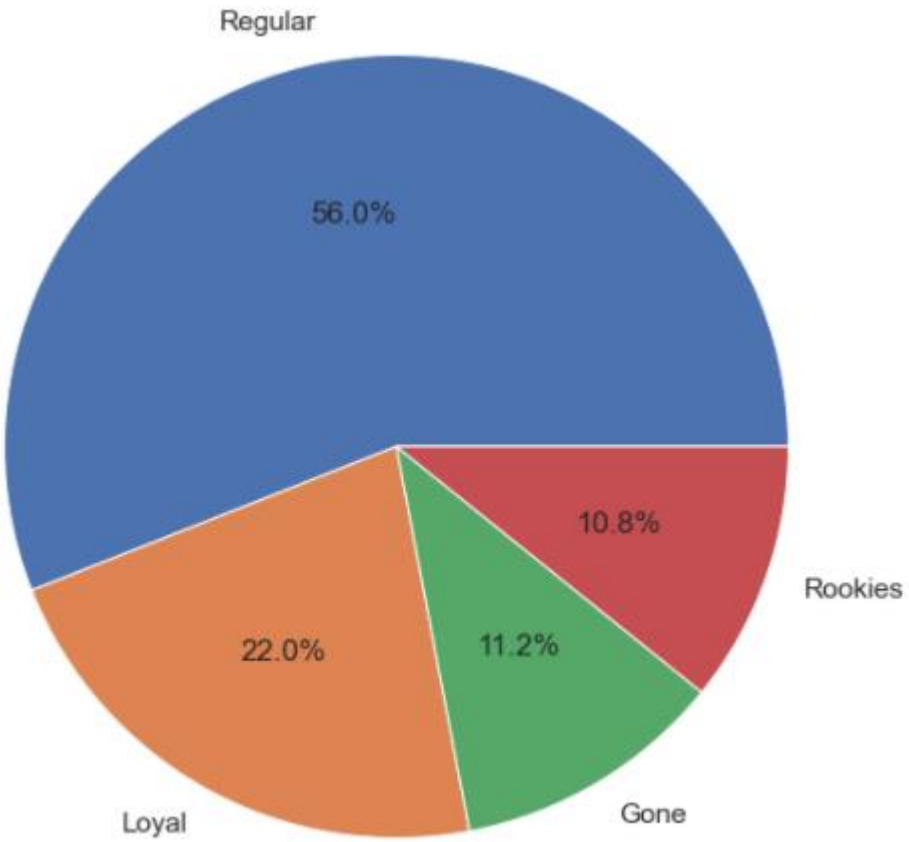
eps = [0.67, 1.04]

Modélisations – Segmentation RFM

- Segmentation manuelle: attribution d'un score en fonction des variables Récence-Fréquence-Montant:
  - CORE - '123' - most recent, frequent, revenue generating - core customers that should be considered as most valuable clients
  - GONE - '311', '312', '313' - gone, one-timers - those clients are probably gone
  - ROOKIE - '111', '112', '113' - just have joined - new clients that have joined recently
  - WHALES - '323', '213', '223 - most revenue generating - whales that generate revenue
  - LOYAL - '221', '222', '321', '322' - loyal users
  - REGULAR - '121', '122', '211', '212' - average users - just regular customers that don't stand out

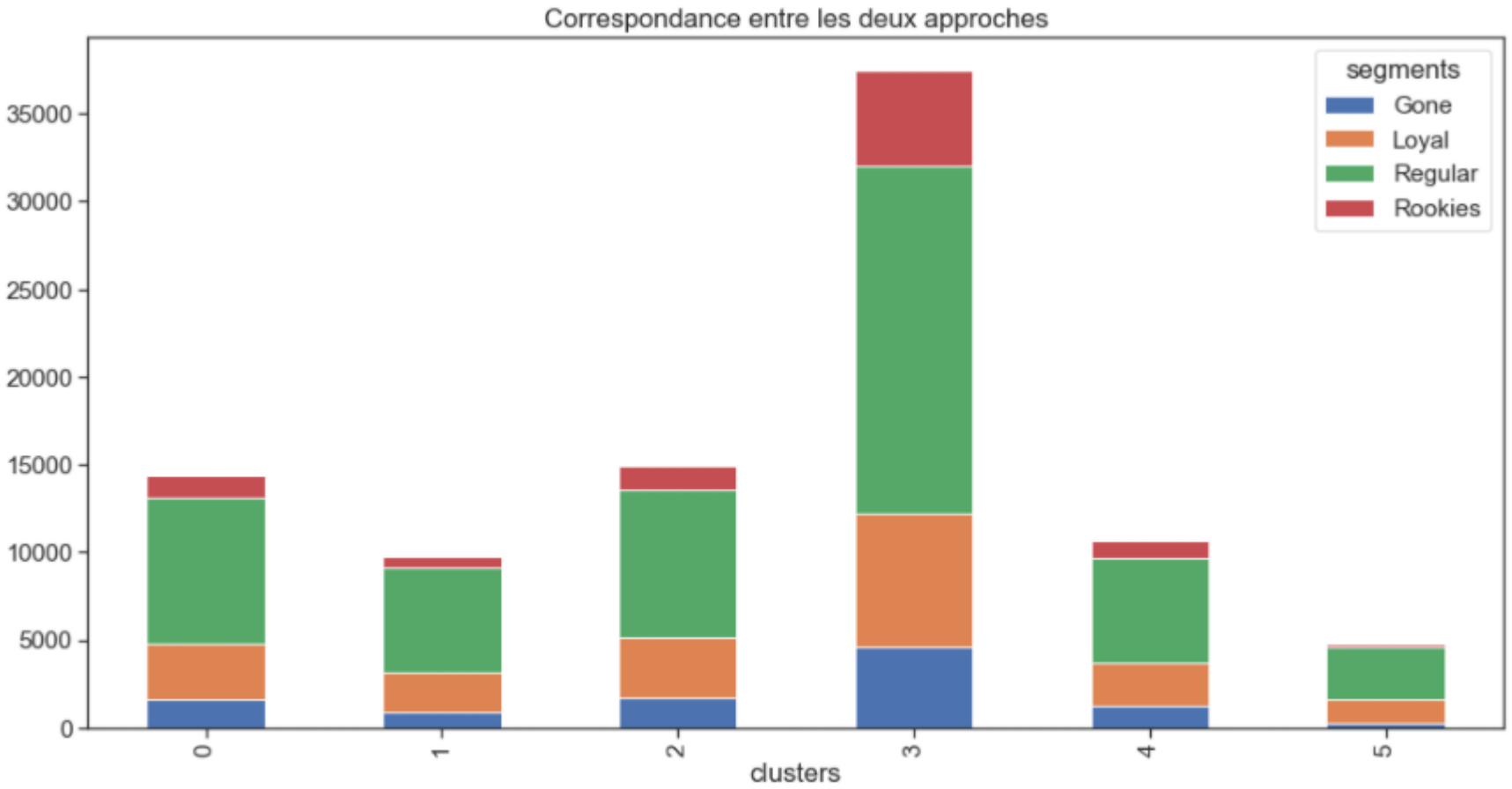
Modélisations – Segmentation RFM

- Distribution des segments



Modélisations – Segmentation RFM

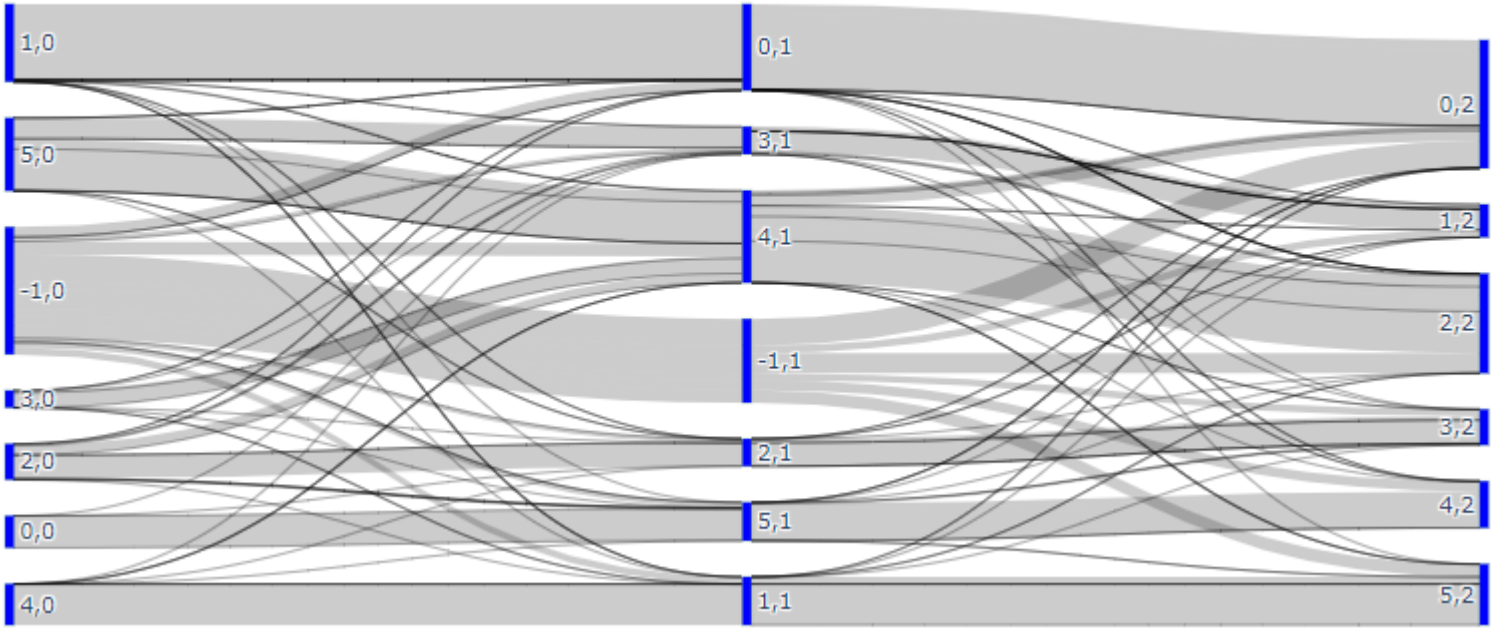
- Correspondance avec KMeans



Stabilité dans le temps

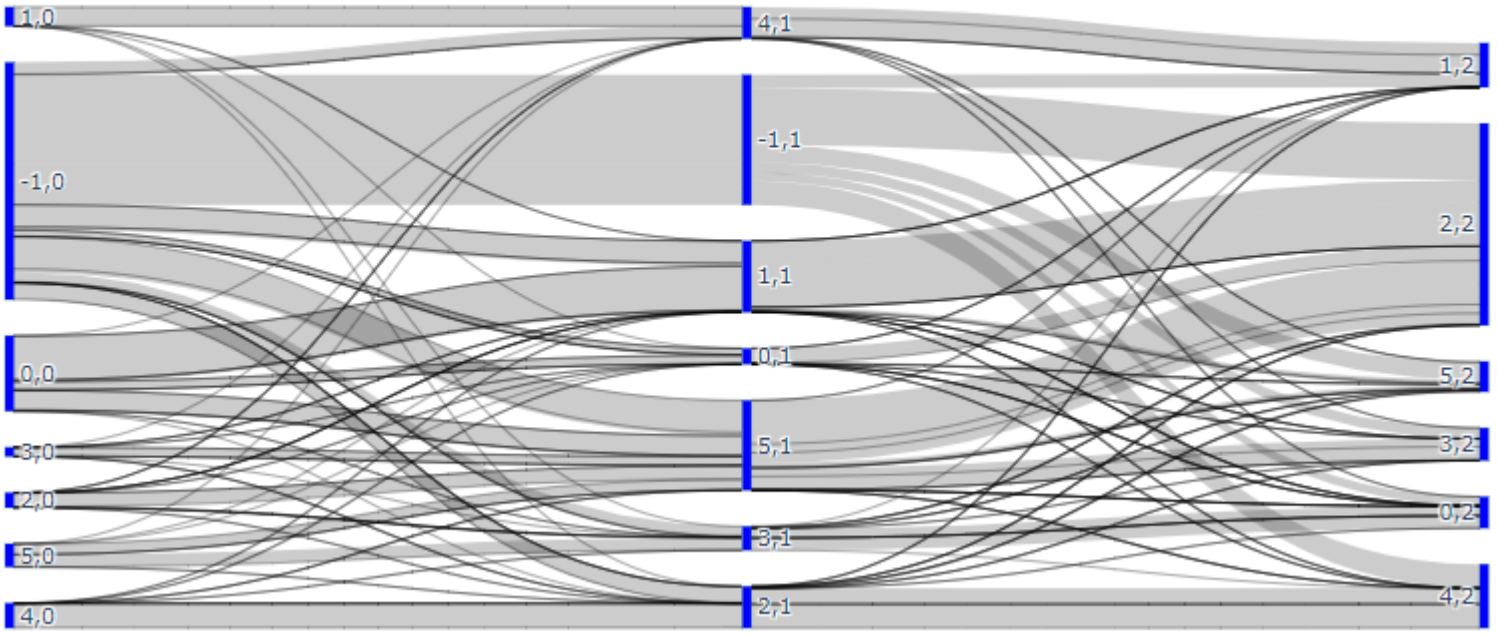
Stabilité dans le temps

Evolution des flux clients pour 6 clusters tous les 1 mois



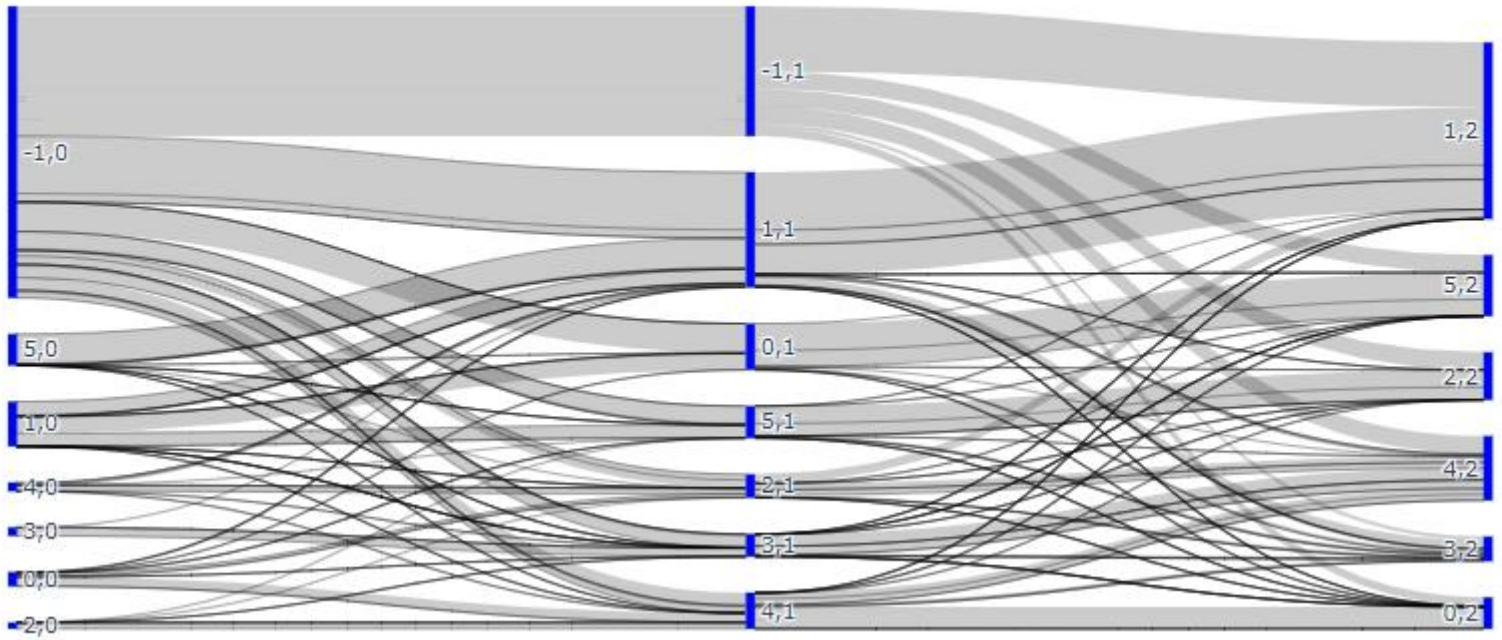
Stabilité dans le temps

Evolution des flux clients pour 6 clusters tous les 3 mois



Stabilité dans le temps

Evolution des flux clients pour 6 clusters tous les 6 mois



Conclusion

- Algorithmes non supervisés assez peu performants globalements sur ce jeu de données
- Pas de cluster qui se détache distinctement à part celui des retards de livraison, qui engendrent de mauvais avis
- La plupart des consommateurs n'achètent qu'une fois (97% de la base de données)
- Contrat de maintenance tous les 3 mois

Merci de votre attention

Annexe