Segmentez des clients d'un site e-commerce

Projet 5 du parcours
« Data Scientist » d'OpenClassrooms

Mark Creasey

Sommaire

Segmentation des clients d’un site d’e-commerce

01 Présentation de la problématique

Mission - Segmentation des clients

Pour améliorer les actions du marketing

- comprendre **les différents types d'utilisateurs**
- fournir à l'équipe marketing **une description actionnable**
- une proposition de **contrat de maintenance** (fréquence à laquelle la segmentation doit être mise à jour pour rester pertinente)

Contraintes

- Source de données:
  - anonymisées
  - période limitée de 18 mois
  - 9 fichiers CSV à intégrer
- Format des livrables
  - Suivre PEP8

olist

[olist] - solution de vente en ligne

Interprétation du problématique

La plateforme Olist

customers

payments

orders

products

sellers

Les démarches

- Nettoyage fusion de **données par client**
- Analyse exploratoire des **dimensions disponibles**
- Feature engineering: Sélection / création d'**indicateurs de comportement**
- Modélisation : **Segmentation des clients**
- Interprétation: **actions à prendre**
- Evaluation de la **stabilité** de segmentation

02 Nettoyage et analyse exploratoire

Les données

La plateforme Olist - tables

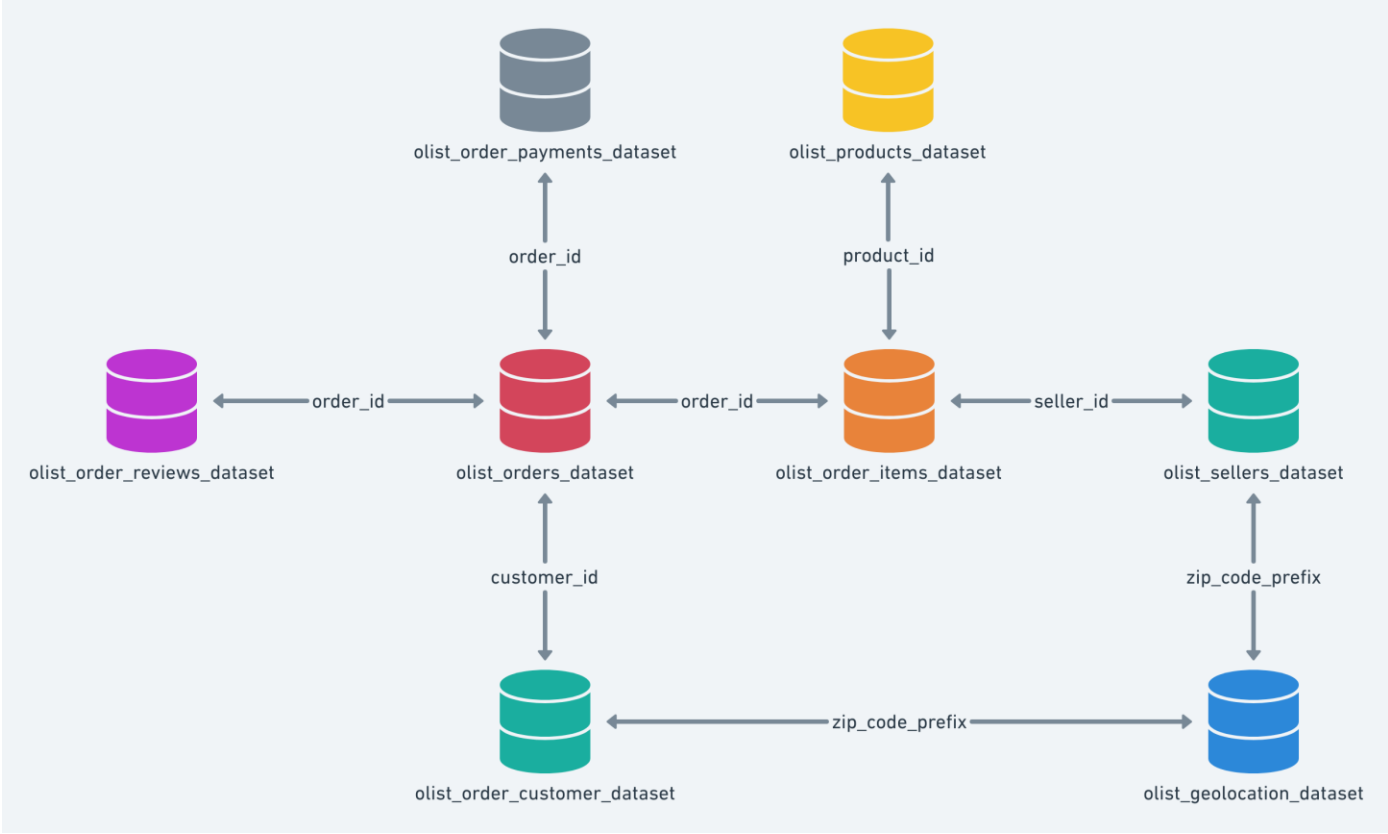
customers
geolocation
orders (delivery)
order reviews
order payments
order items
products
categories
sellers
geolocation

clients

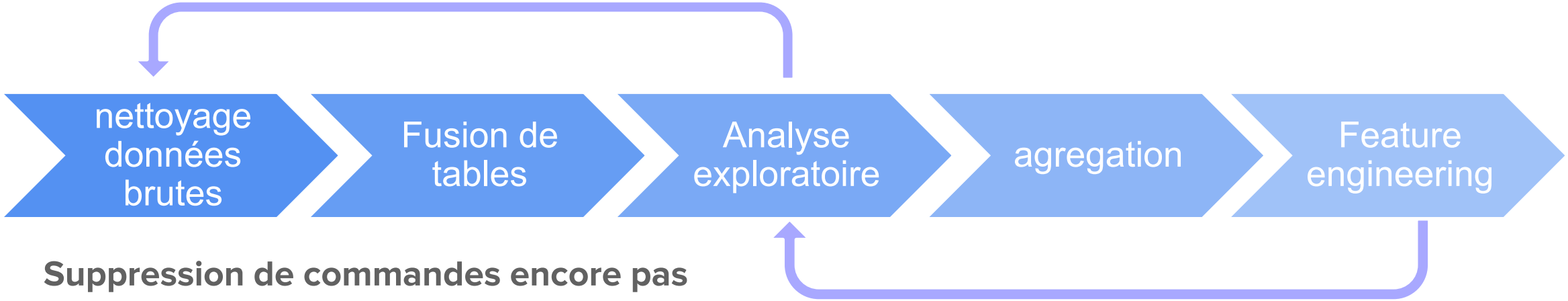
commandes

commande items

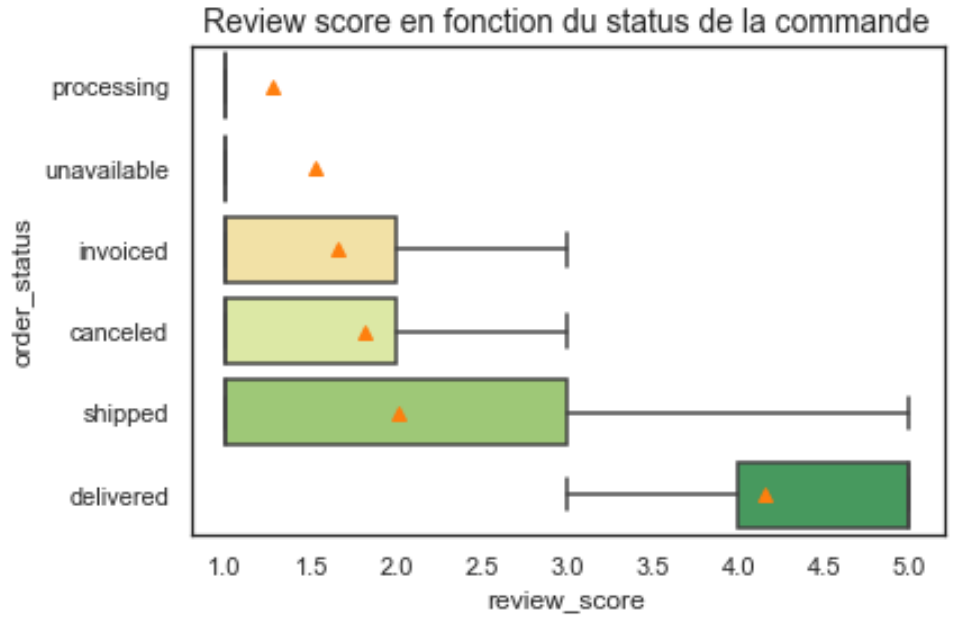
- Téléchargeable sur https://www.kaggle.com/olistbr/brazilian-ecommerce

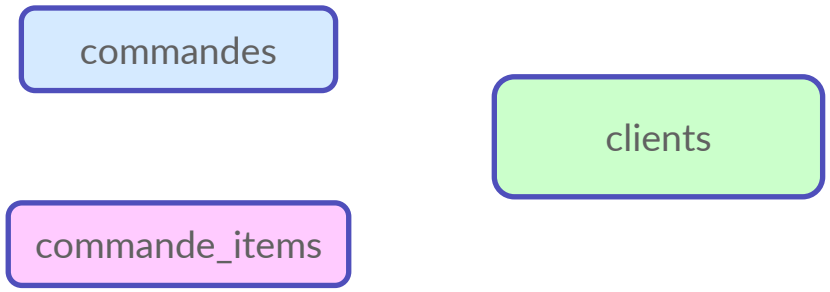


Préparation des données



Suppression de commandes encore pas livrés





Clients - group by unique_customer_id

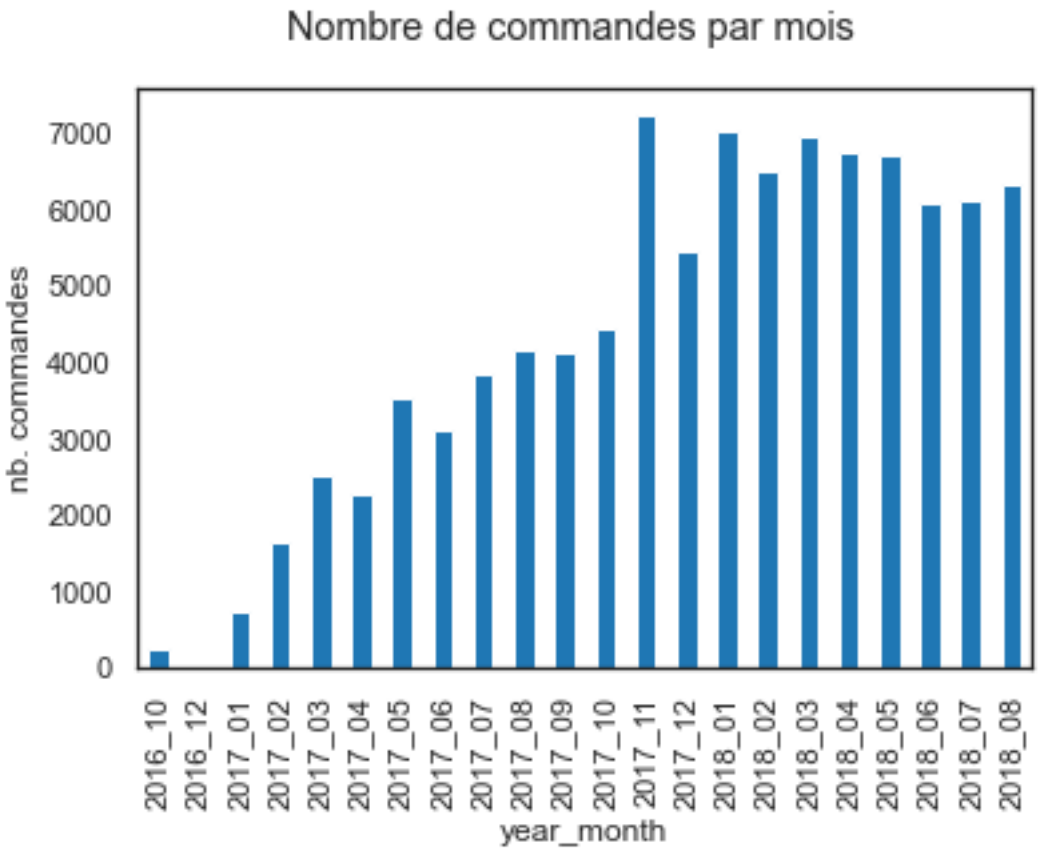
Feature engineering

- Confort financier
  - Nombre de commandes
  - Montant total dépensé
- Moyen de paiements
  - Mode de paiement préféré
- Satisfaction
  - Note moyenne des avis postés
- Accessibilité
  - Délai de livraison (date prévisionnelle vs. date de livraison)
- Produits :
  - Simplification de catégorie (segmentation produit)
  - Catégorie préférée

Analyse financier: nombre de commandes croissant

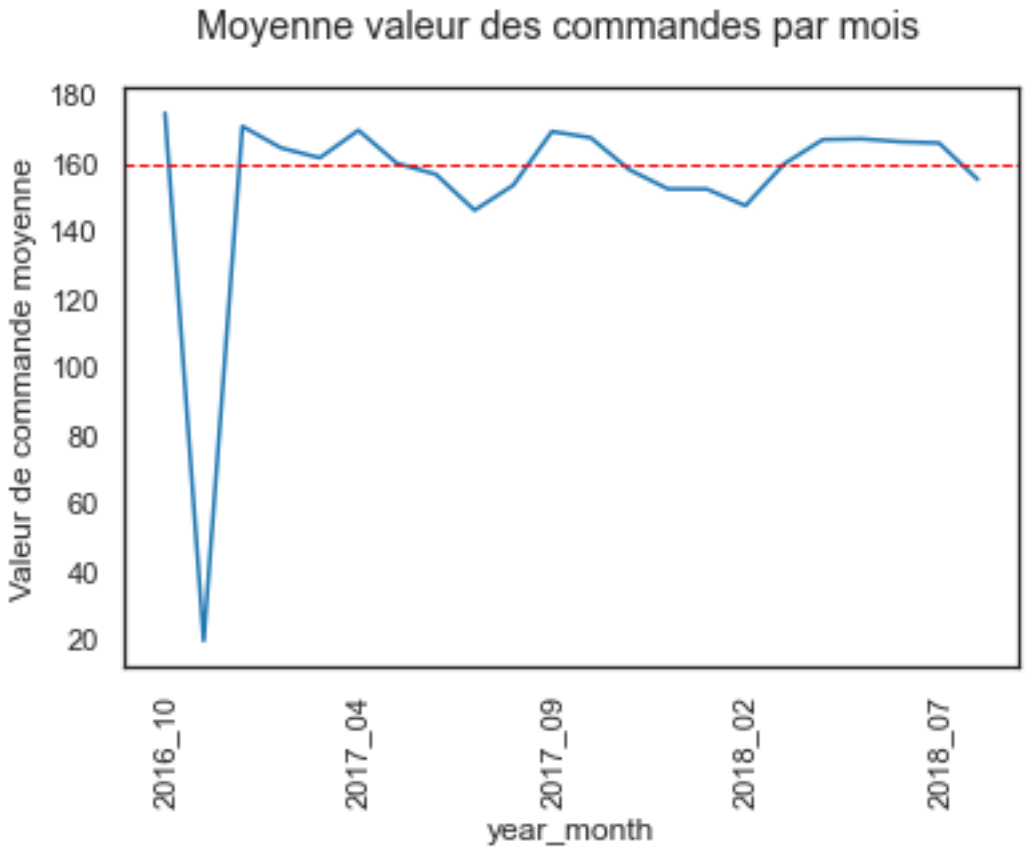
Nombre de commandes

- Croissant pendant 1 an, puis stable
- Stabilité des segments ???



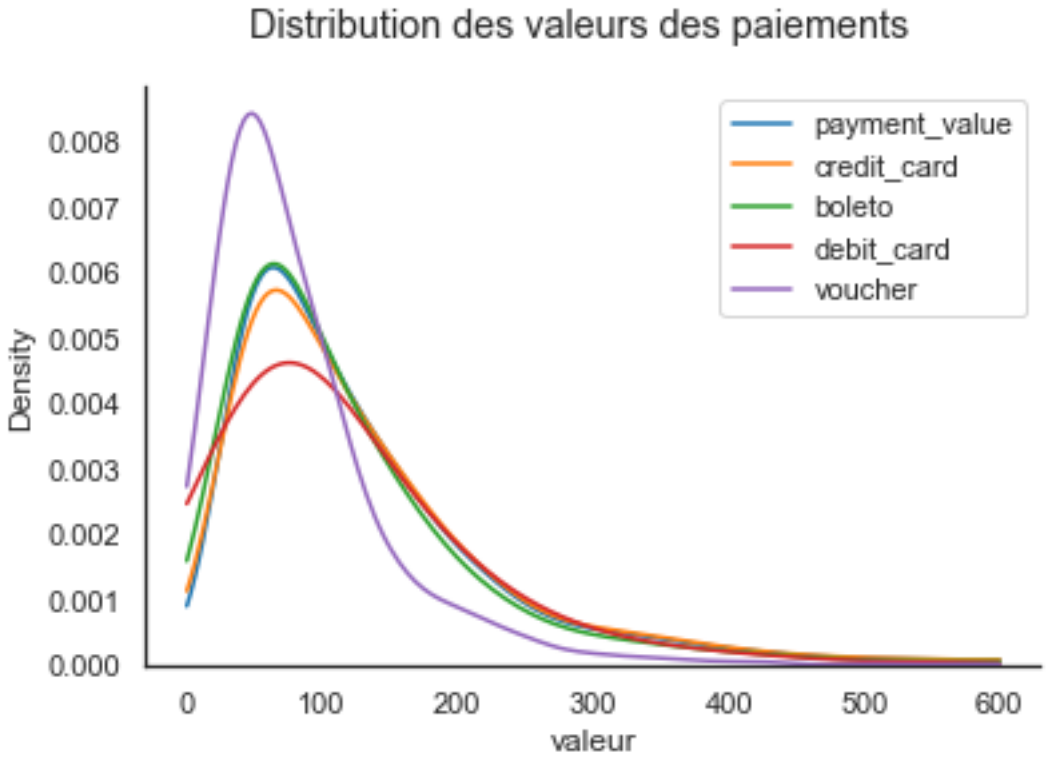
Montant dépensé

- 97% des clients ont fait seulement un commande

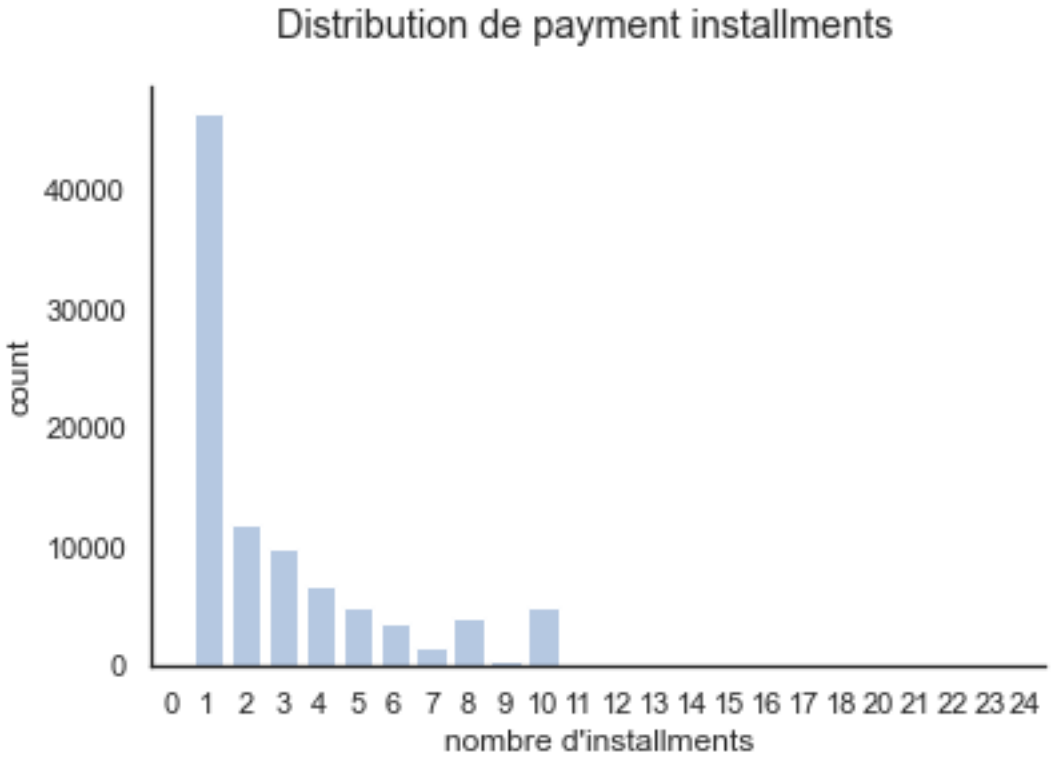


Analyse financier : Moyen et nombre de paiements

- Moyenne de paiement indépendant du valeur

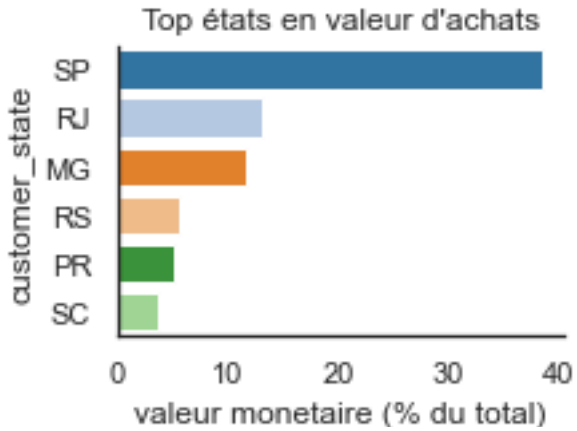


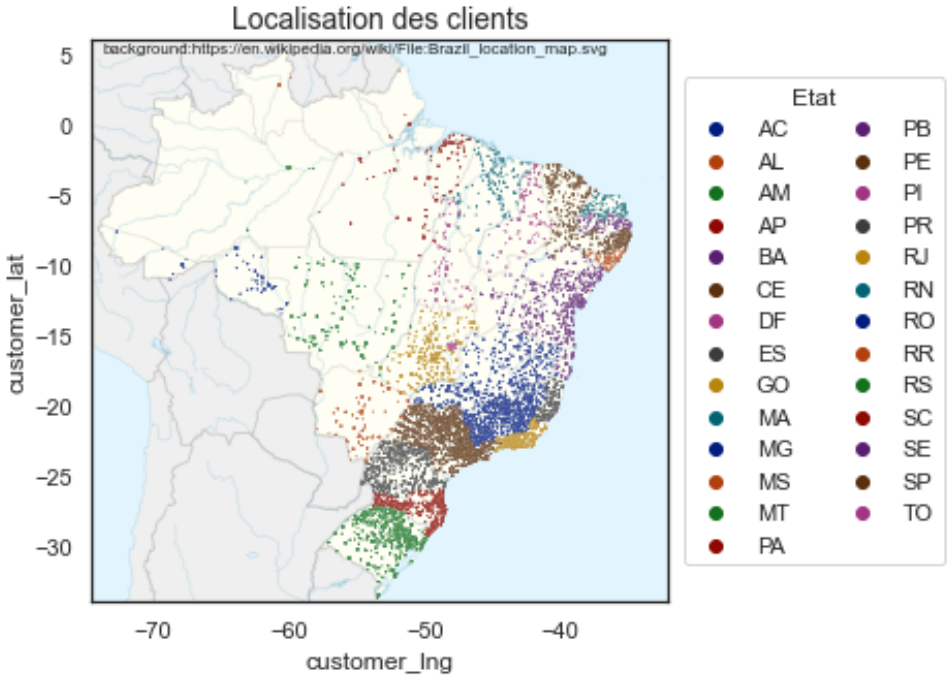
- 55% des commandes sont payés en plusieurs fois



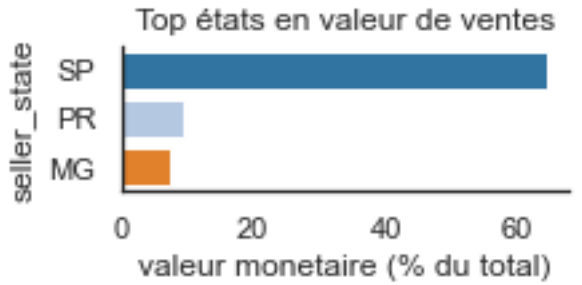
Analyse accessibilité : 80% du commerce en 6 états

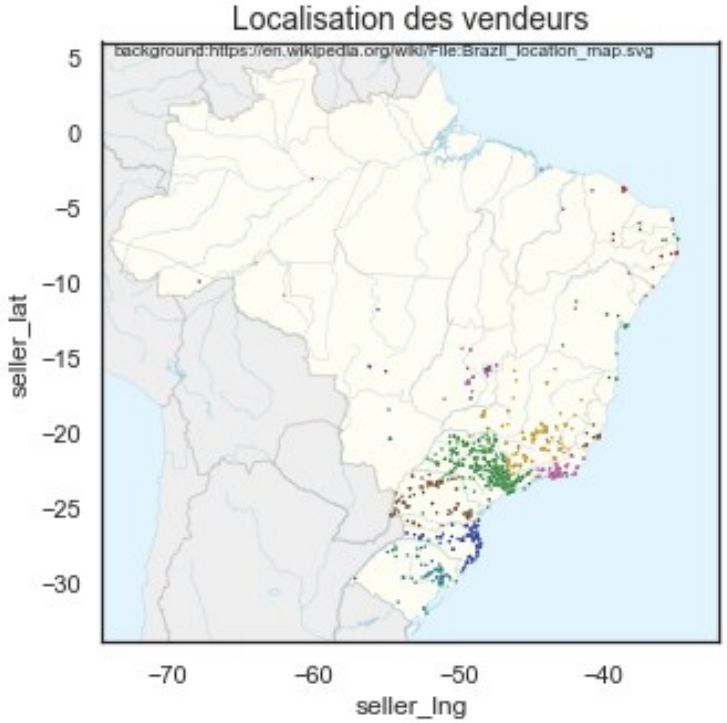
80% achats en 6 états du sud



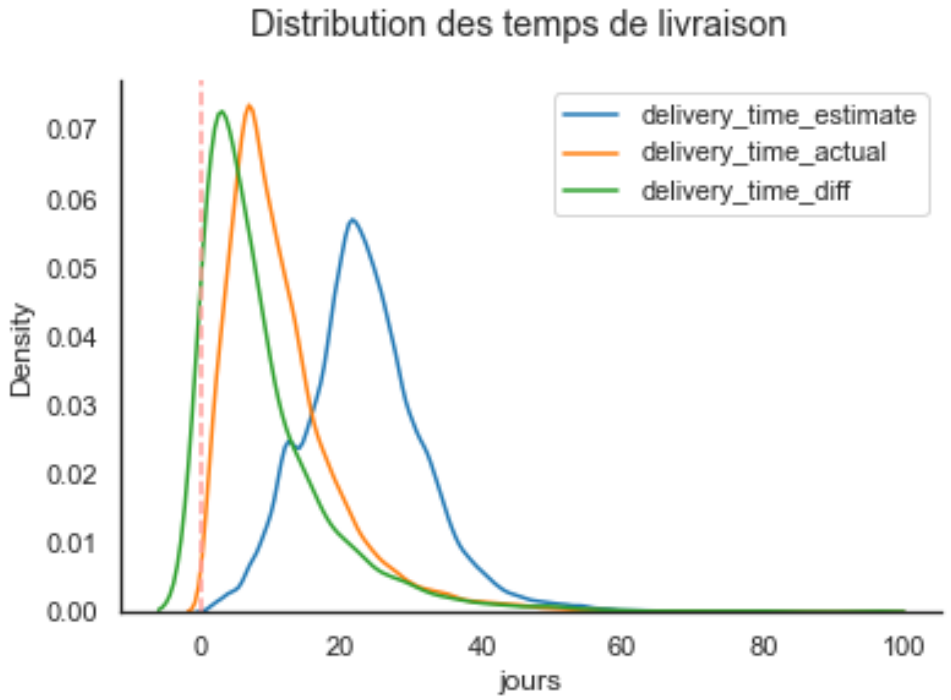


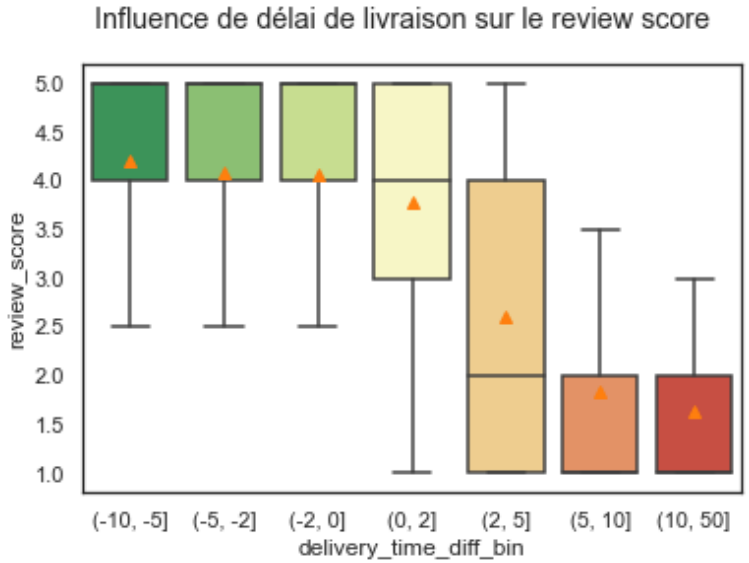
- 80% des ventes de 3 états (SP, PR, MG)

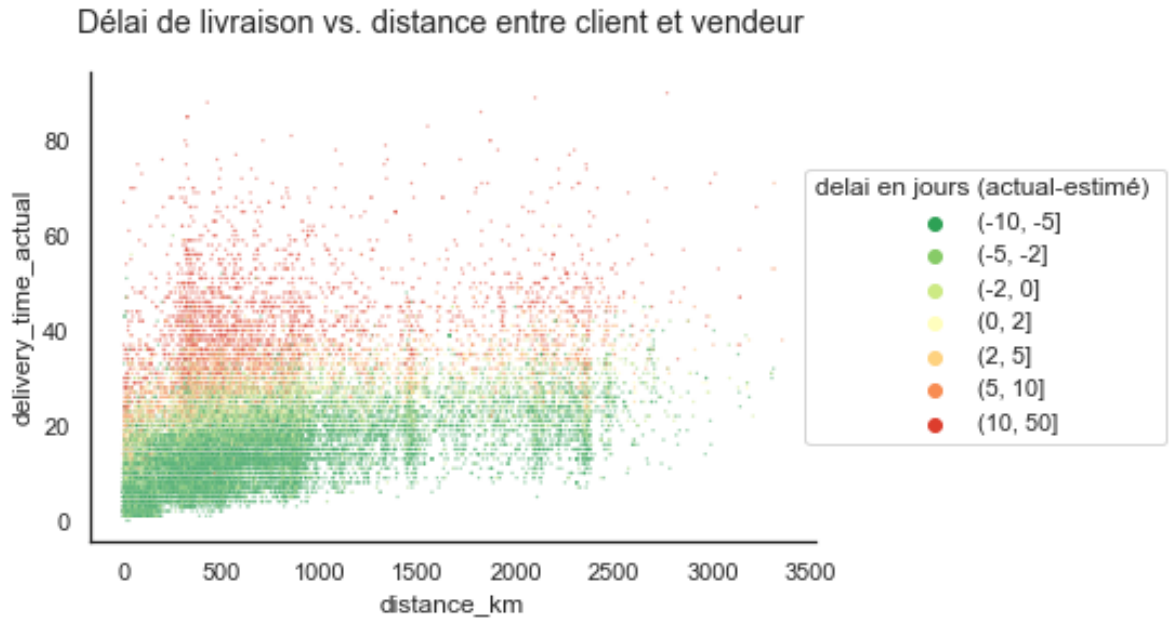




Analyse accessibilité : Délai de livraison

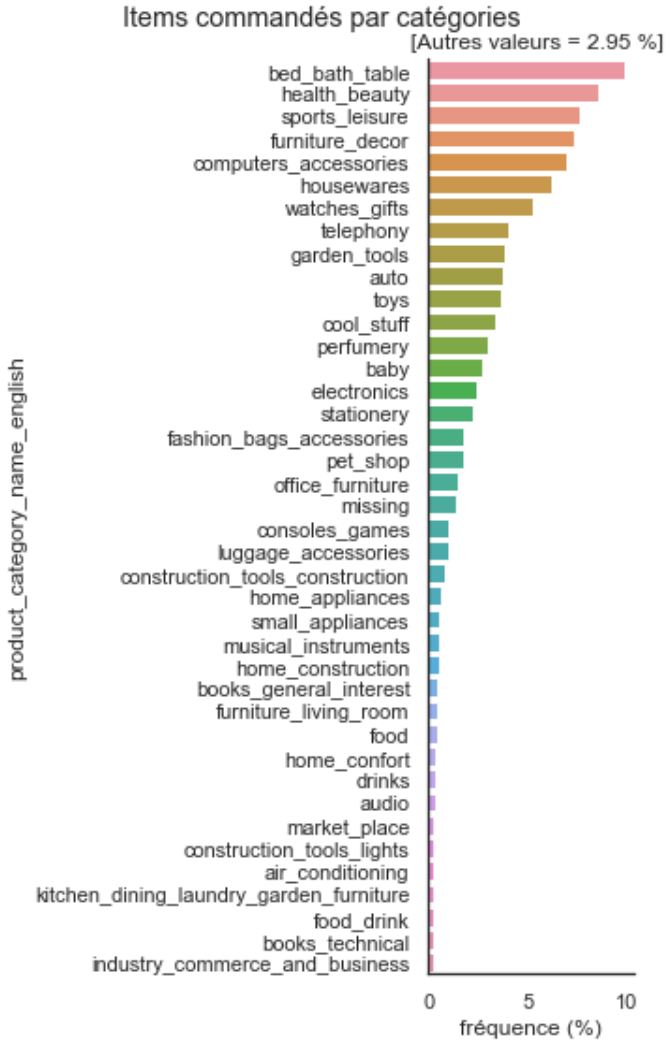




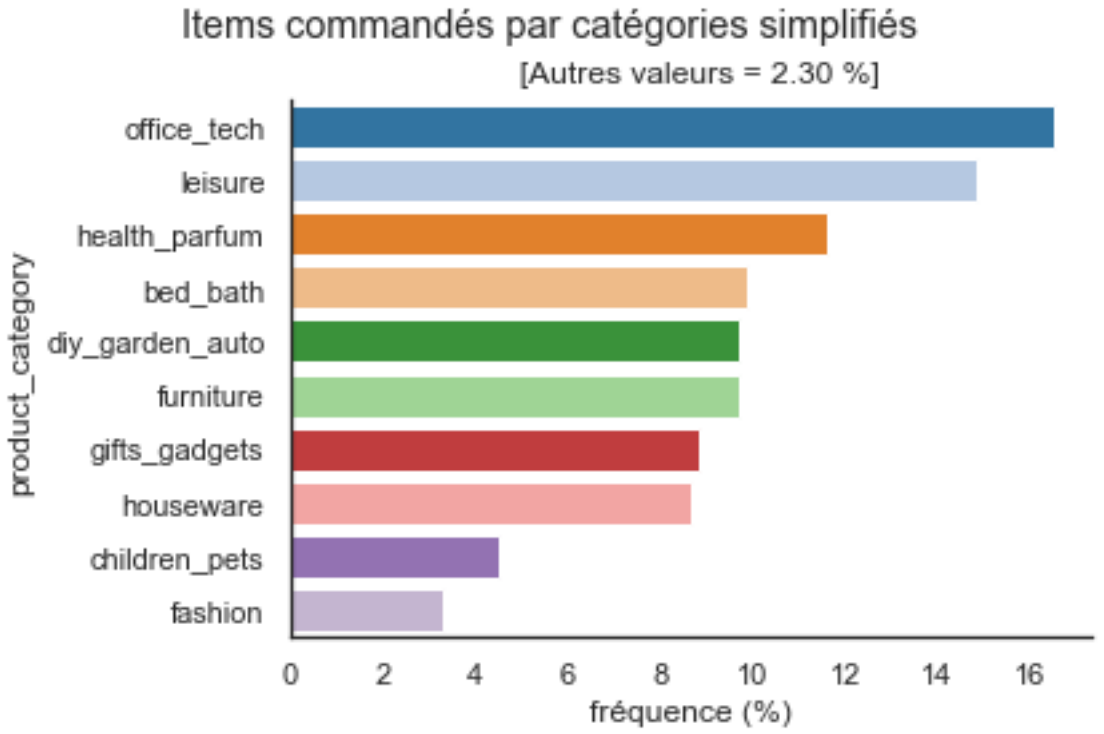


Analyse produits : simplification de catégories

- > 56 catégories

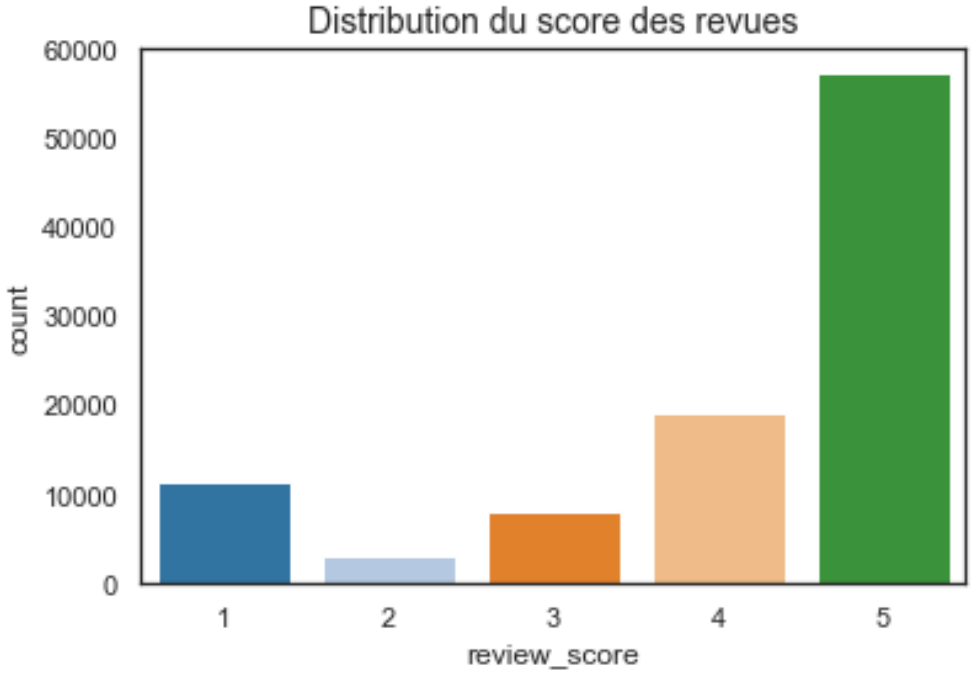


- Groupement manuelle en 10 catégories

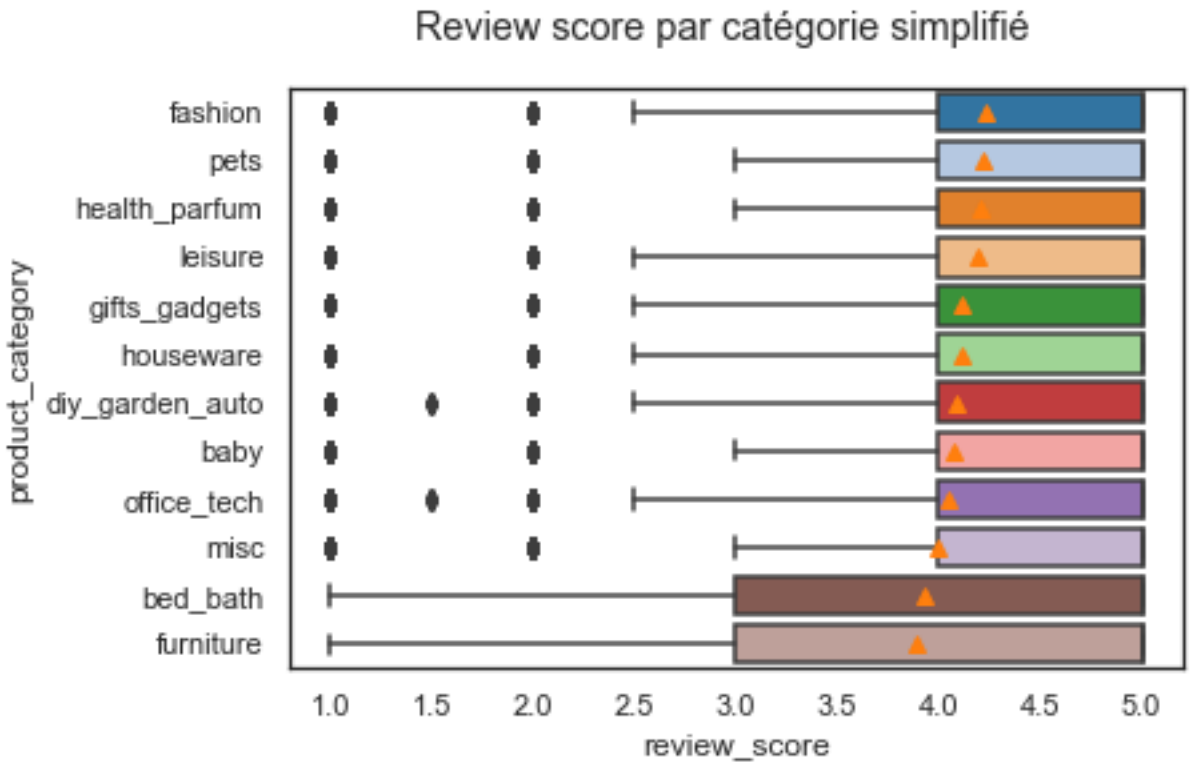


Analyse satisfactions : Review Score

- La plupart des review scores sont 5/5



- Review score varie par catégorie
  - Temps de livraison ?
  - Qualité / localisation vendeur ?
  - Expectations client ?

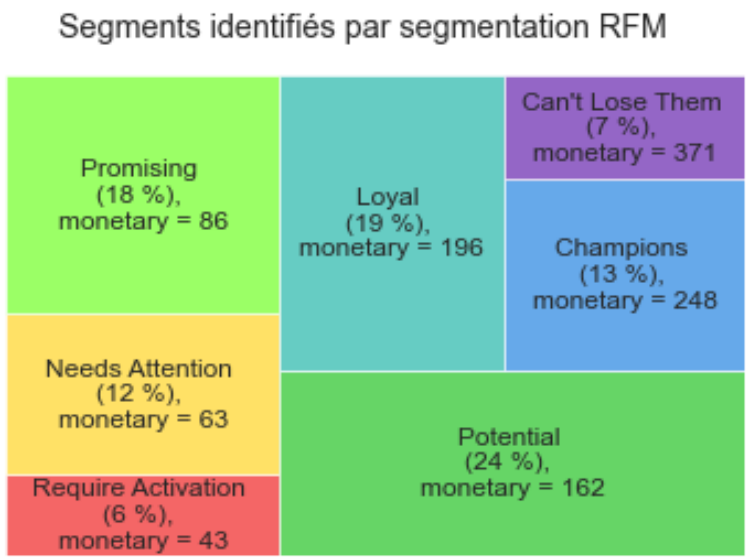


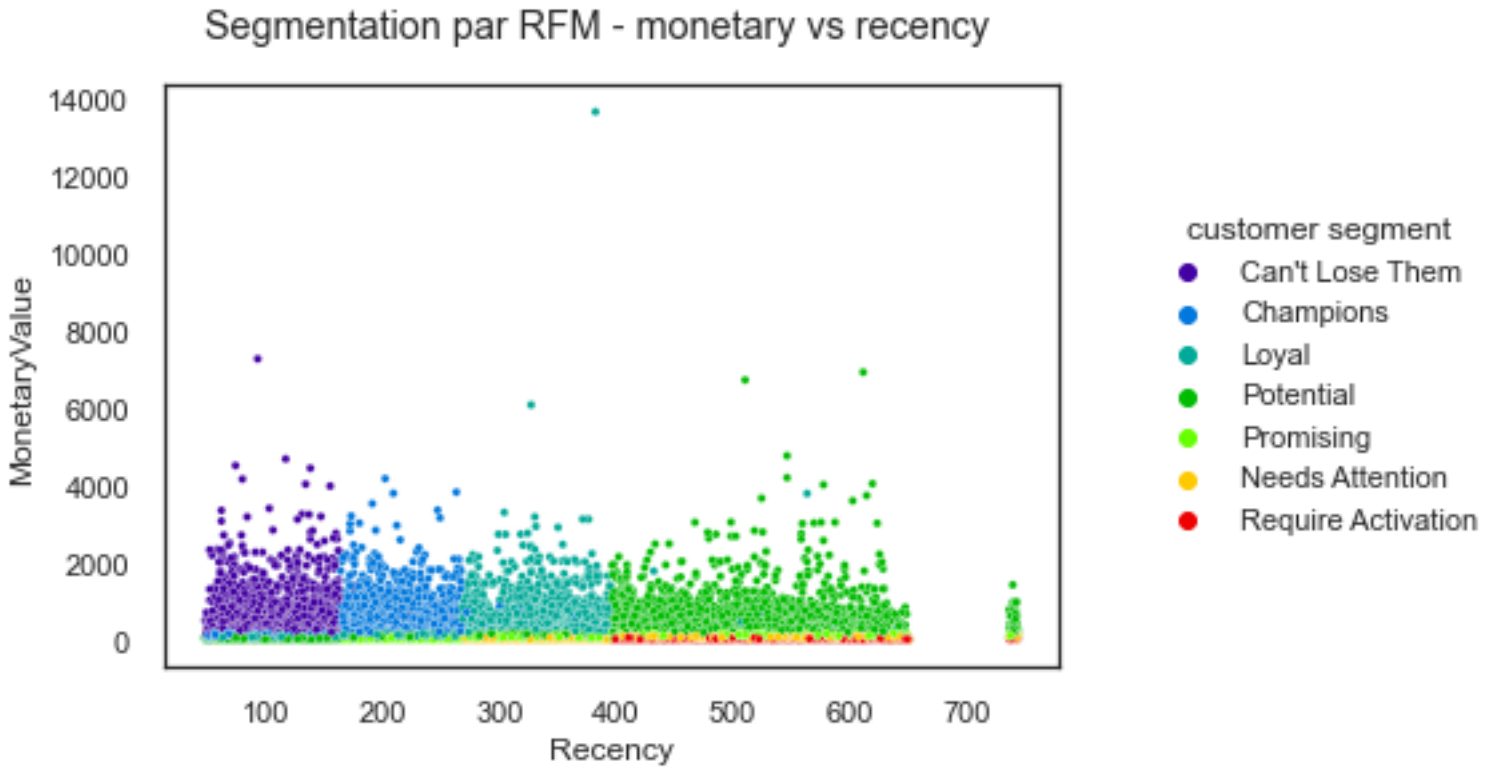
03. Modélisations effectuées

1. Segmentation RFM (Récence, Fréquence, Monétaire)

- Scoring RFM
  - R = Récence en 4 bins (quantiles)
  - F = Fréquence en 3 bins (1,2, plus)
  - M = Montant moyenne en 4 bins (quantiles)
- Score = R + F + M

- Incite à faire du marketing au plus récent (97% clients ont fait une seule commande)

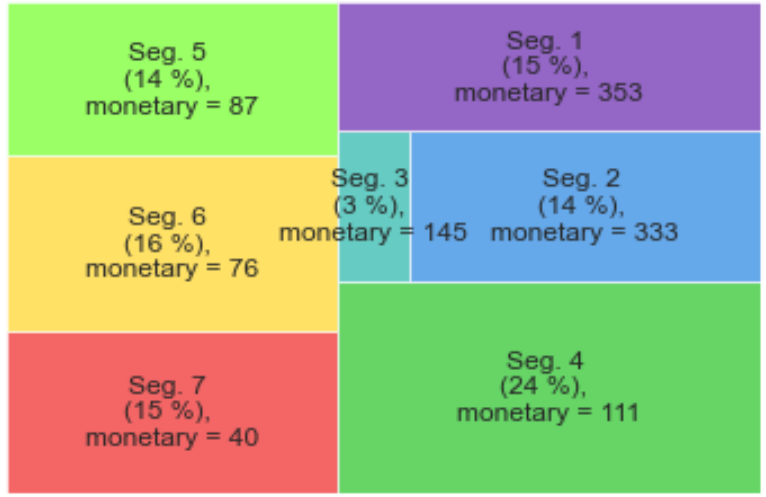


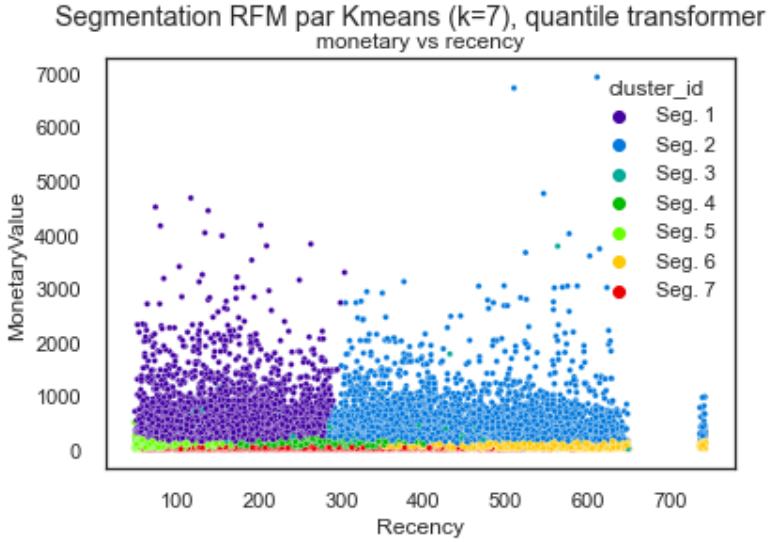


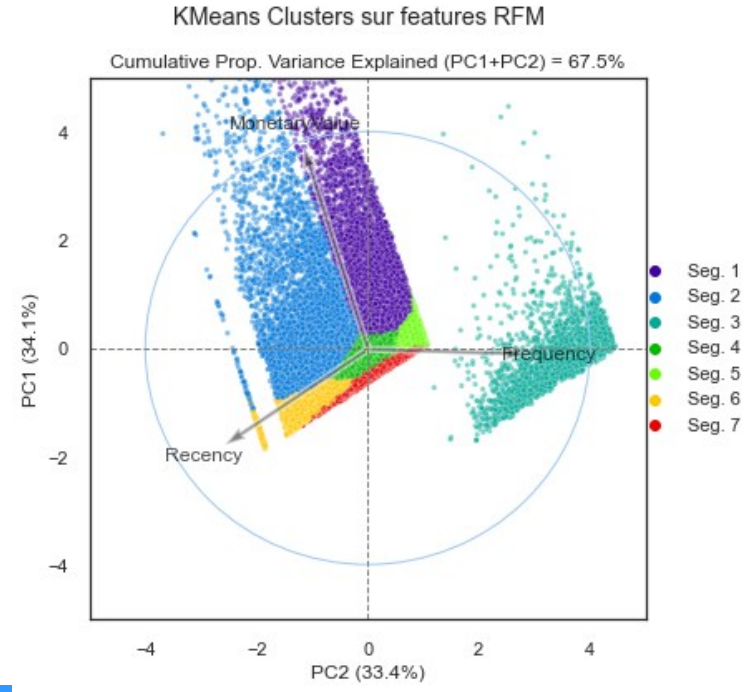
2. Segmentation Kmeans (Récence, Fréquence, Monétaire)

- Segmentation par Kmeans (K=7)

Segments identifiés par segmentation KMeans (k=7) sur features RFM

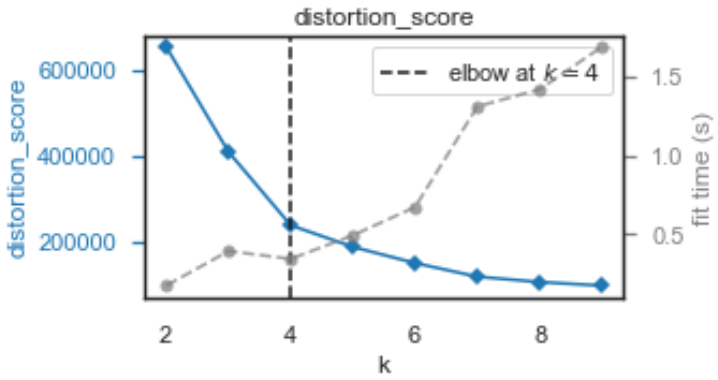


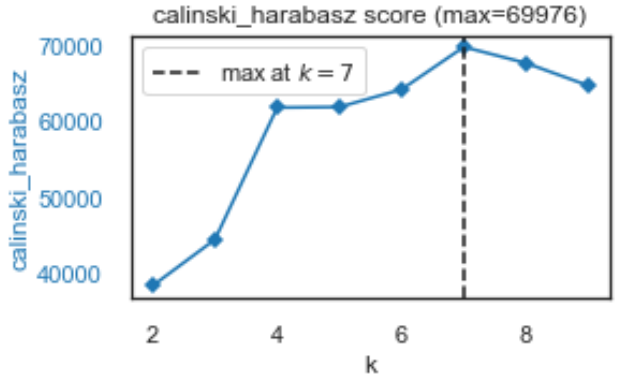


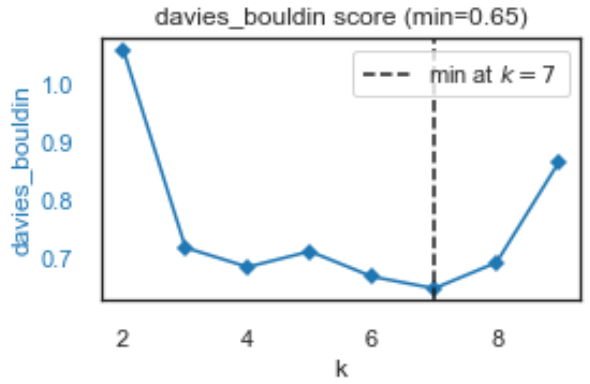


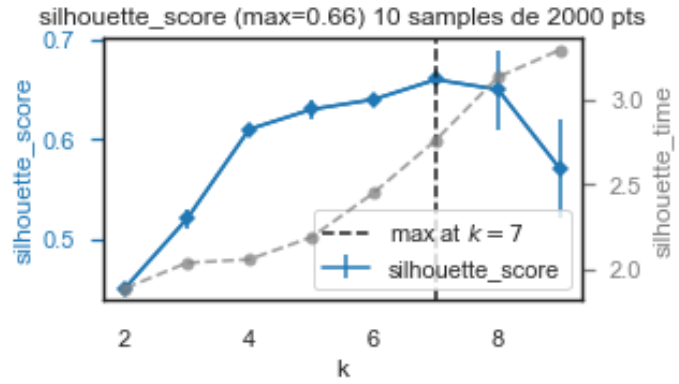
3. Segmentation KMeans – choix du nombre de clusters (k)

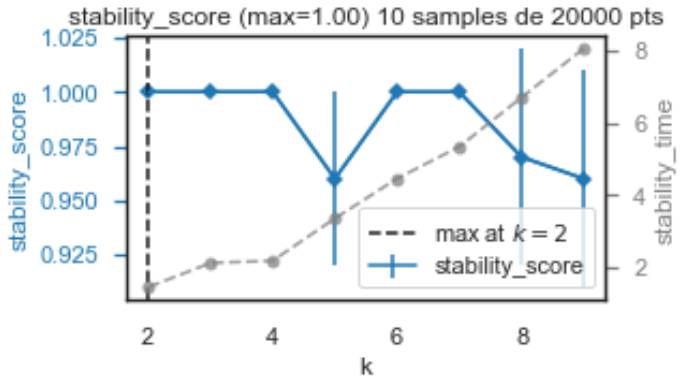
- Cohérence
  - Inertia (distortion)
  - Calinski Harabasz
  - Davies_Bouldin
- Forme
  - Silhouette
- Stabilité
  - Mesure de la reproductibilité des clusters sur échantillons aléatoires





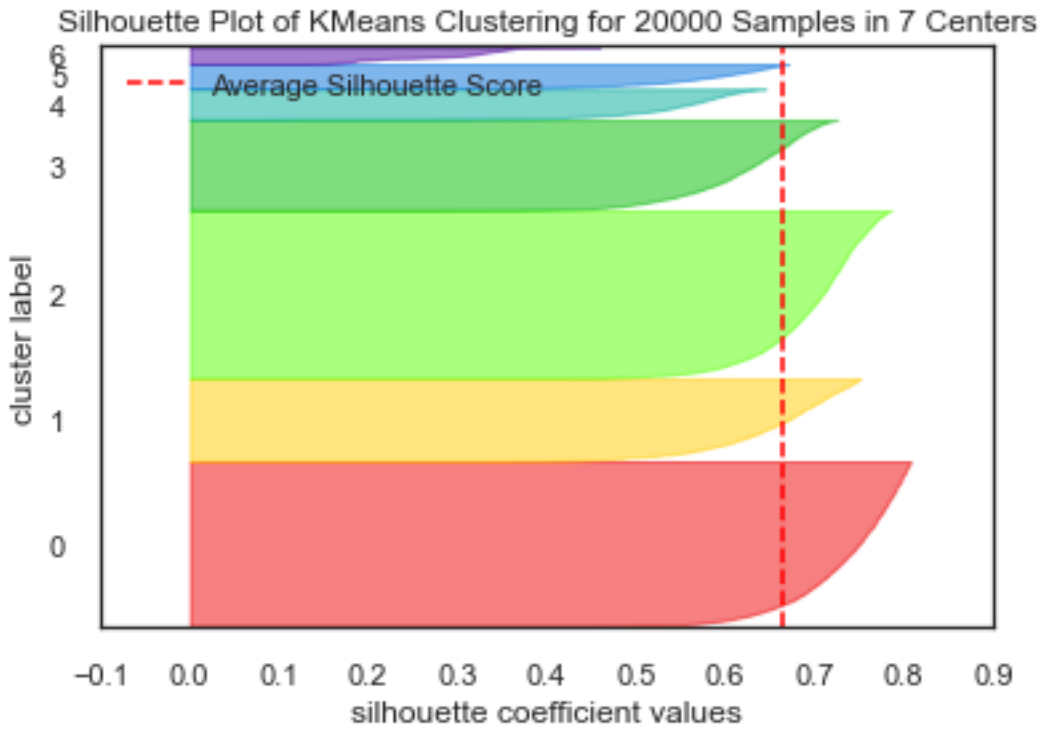







```
FEATURES =
['MonetaryValue',
'Frequency',
'review_score',
'mean_nb_payments',
'delivery_delay']
```

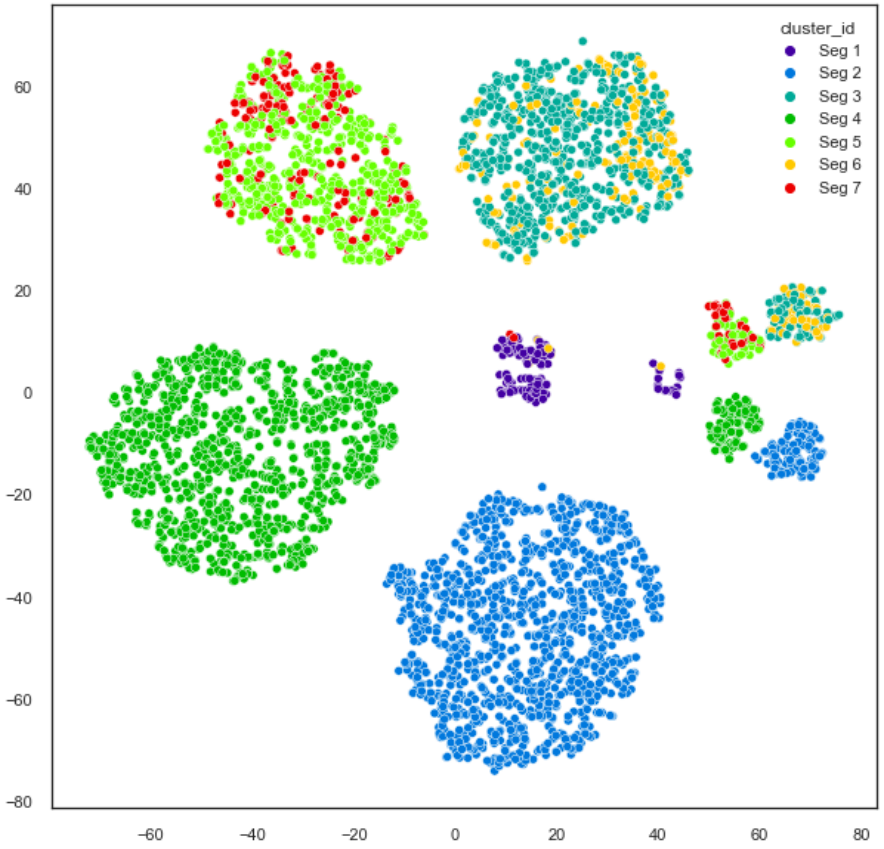
Kmeans – silhouette scores



- `FEATURES = ['MonetaryValue', 'Frequency', 'review_score', 'mean_nb_payments' ,'delivery_delay']`

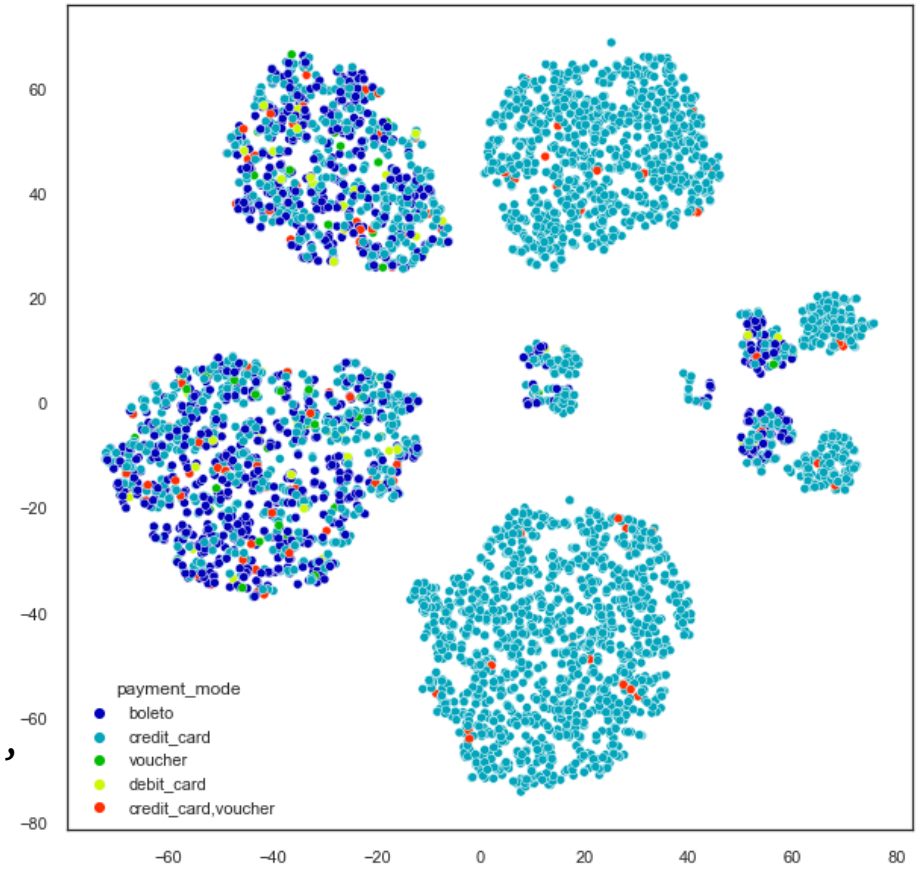
Kmeans (k=7) – Visualisation t-SNE

- Fréquent acheteurs au milieu
- Gros dépensiers à droite
- Petits achats à gauche



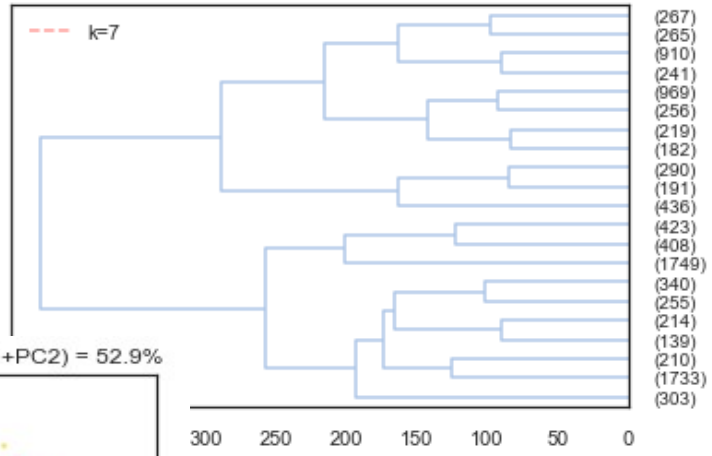

```
FEATURES =
['MonetaryValue',
'Frequency',
'review_score',
'mean_nb_payments',
'delivery_delay']
```

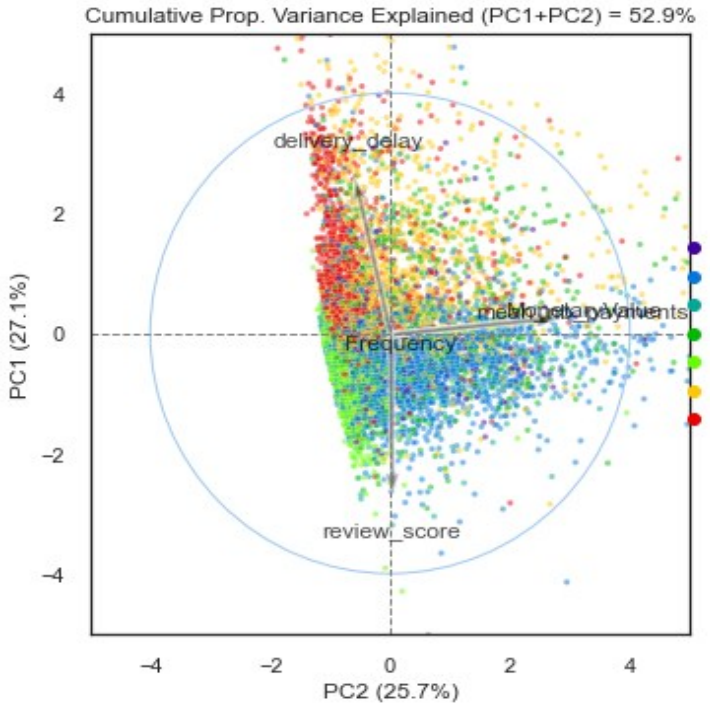
- La moyenne de paiement est aussi discriminatoire



Clustering Agglomerative

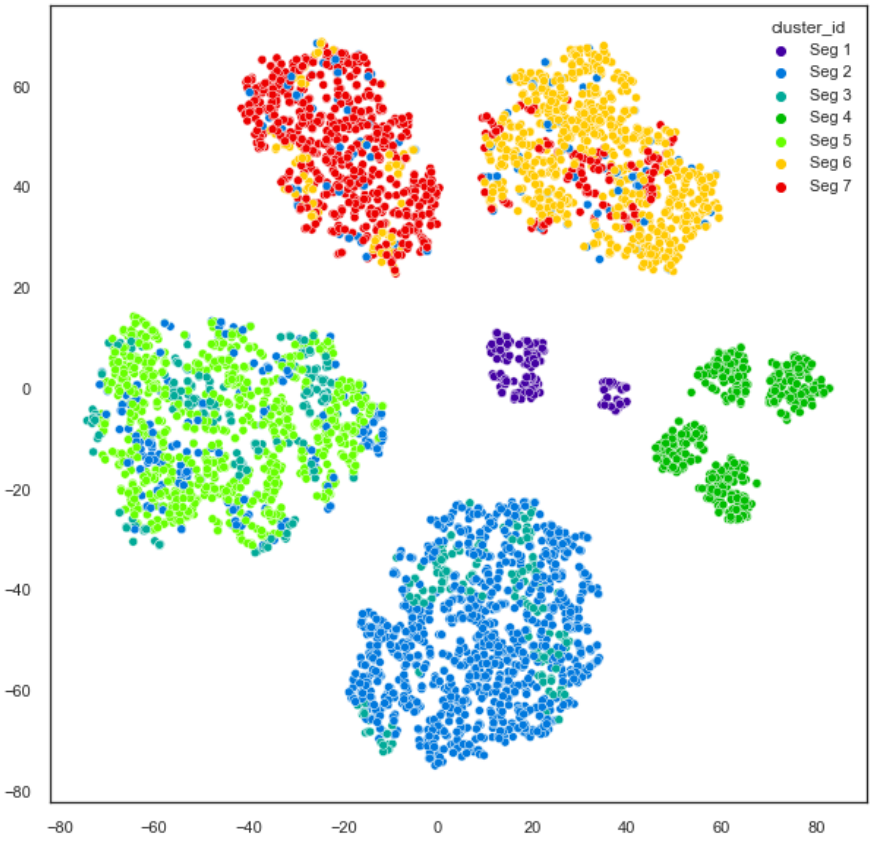
- Meilleurs scores avec linkage Ward, k=7





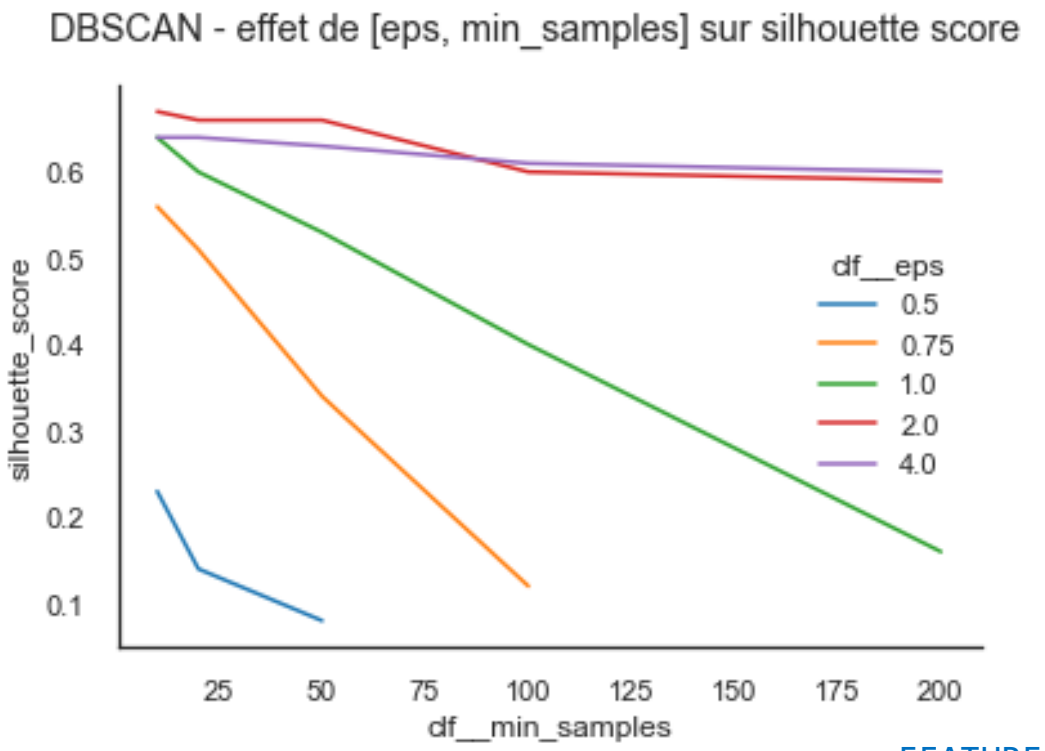

```
FEATURES =
['MonetaryValue',
'Frequency',
'review_score',
'mean_nb_payments',
'delivery_delay']
```

- Segmentation similaire à Kmeans
- Très chronophage



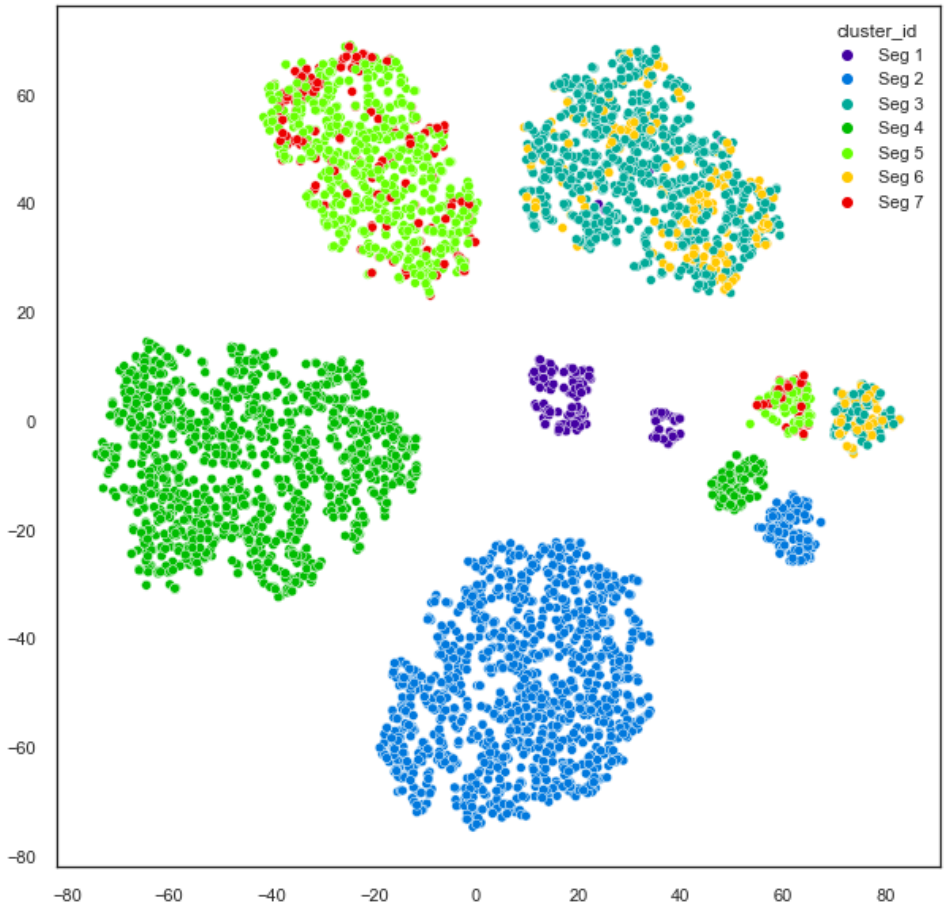
DBSCAN

- Les résultats sont très sensibles au choix d'hyperparamètres




```
FEATURES =
['MonetaryValue',
'Frequency',
'review_score',
'mean_nb_payments',
'delivery_delay']
```

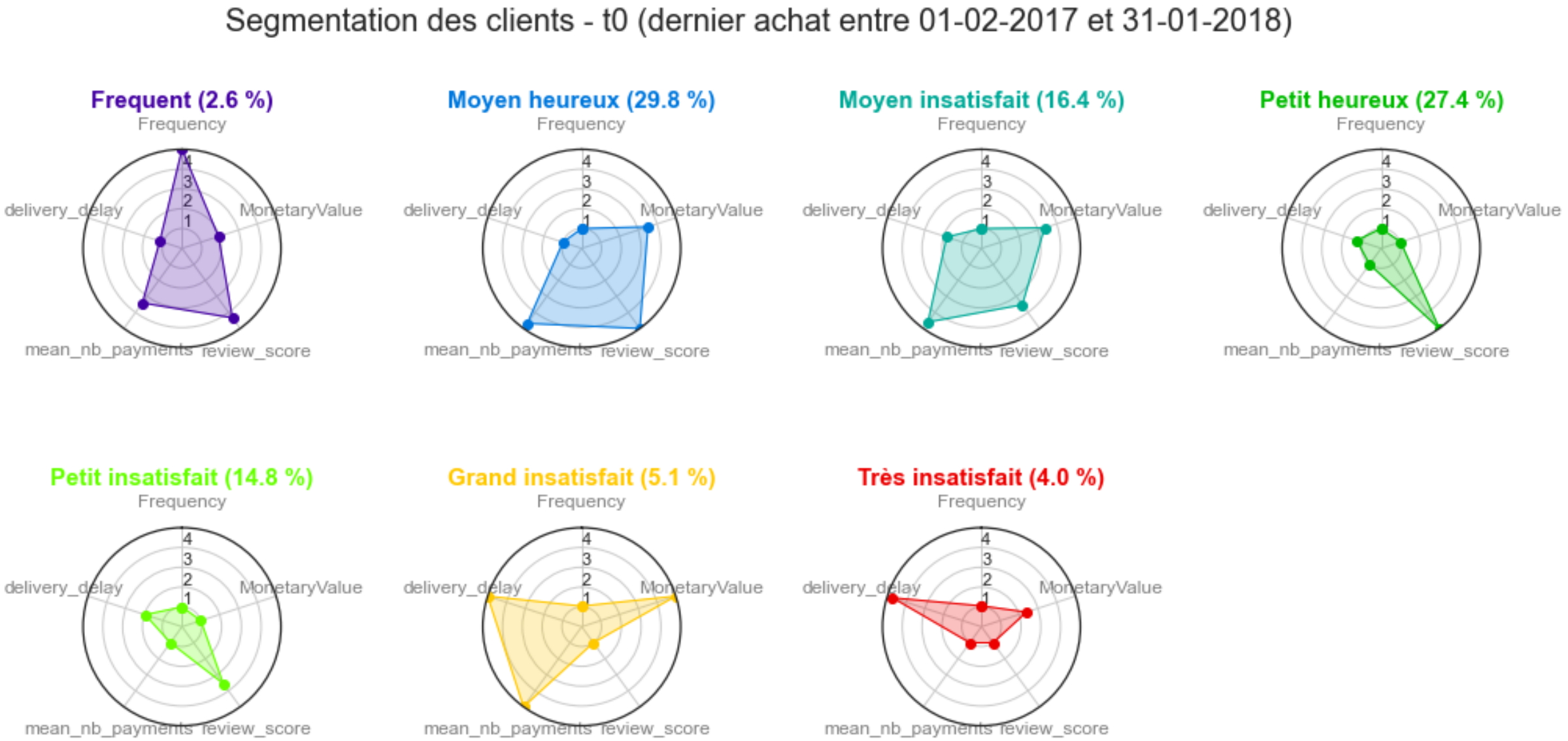
- Meilleurs résultats presque identiques au Kmeans (k=7)



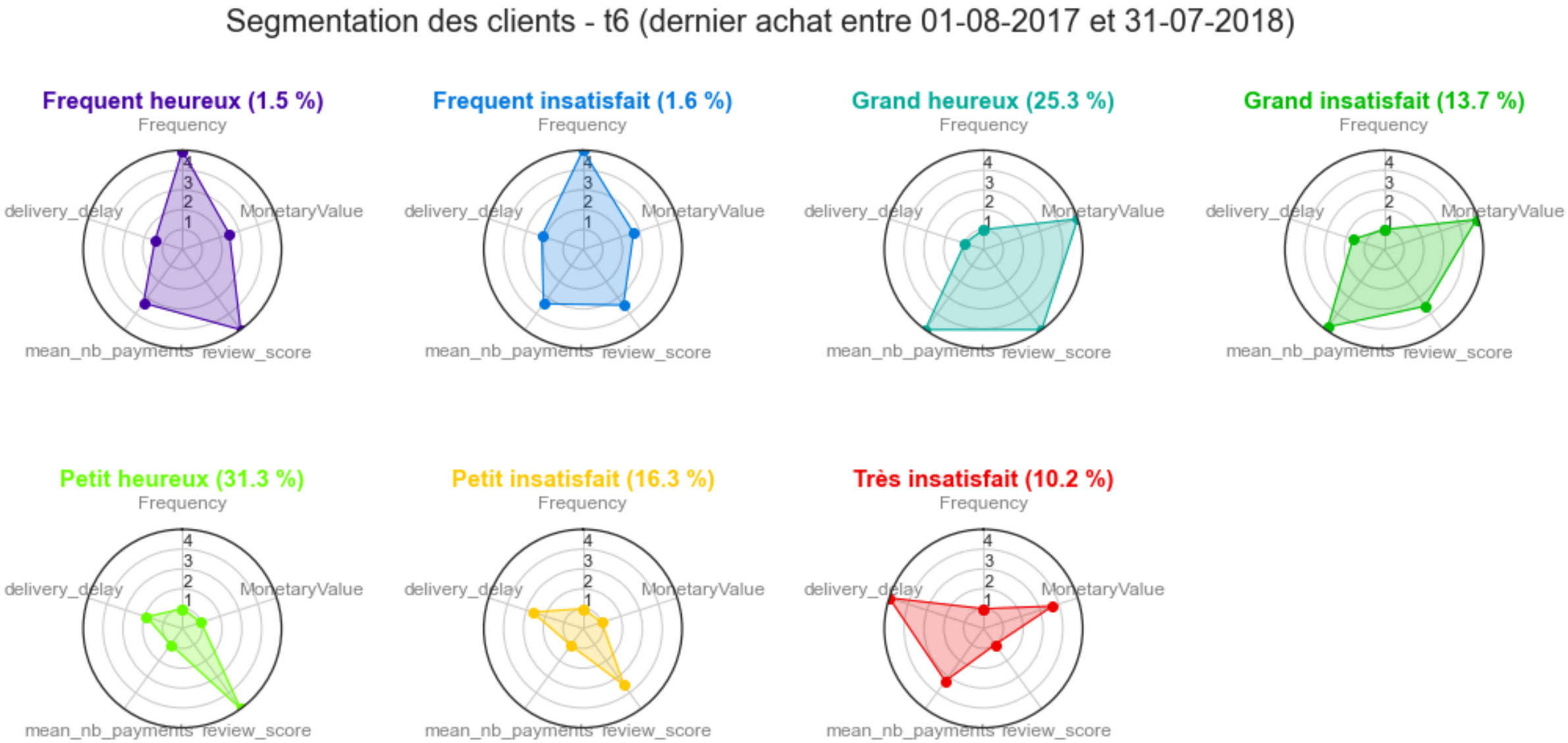
04 Le modèle final sélectionné

05 Simulation de stabilité

Segments par Kmeans (k=7) (année février 2017-8)



Segments change pendant le temps

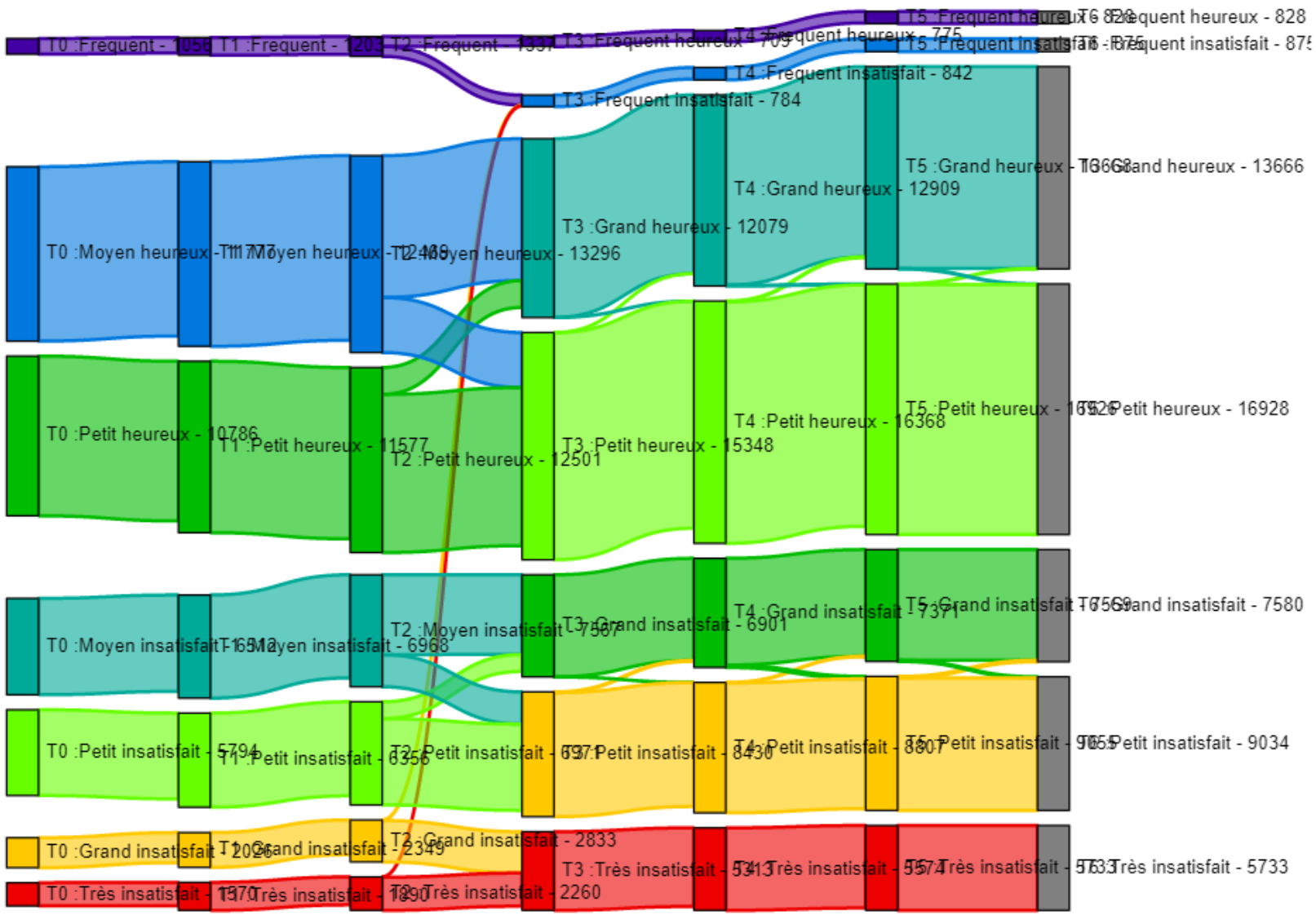


Evolution des segments dans le temps

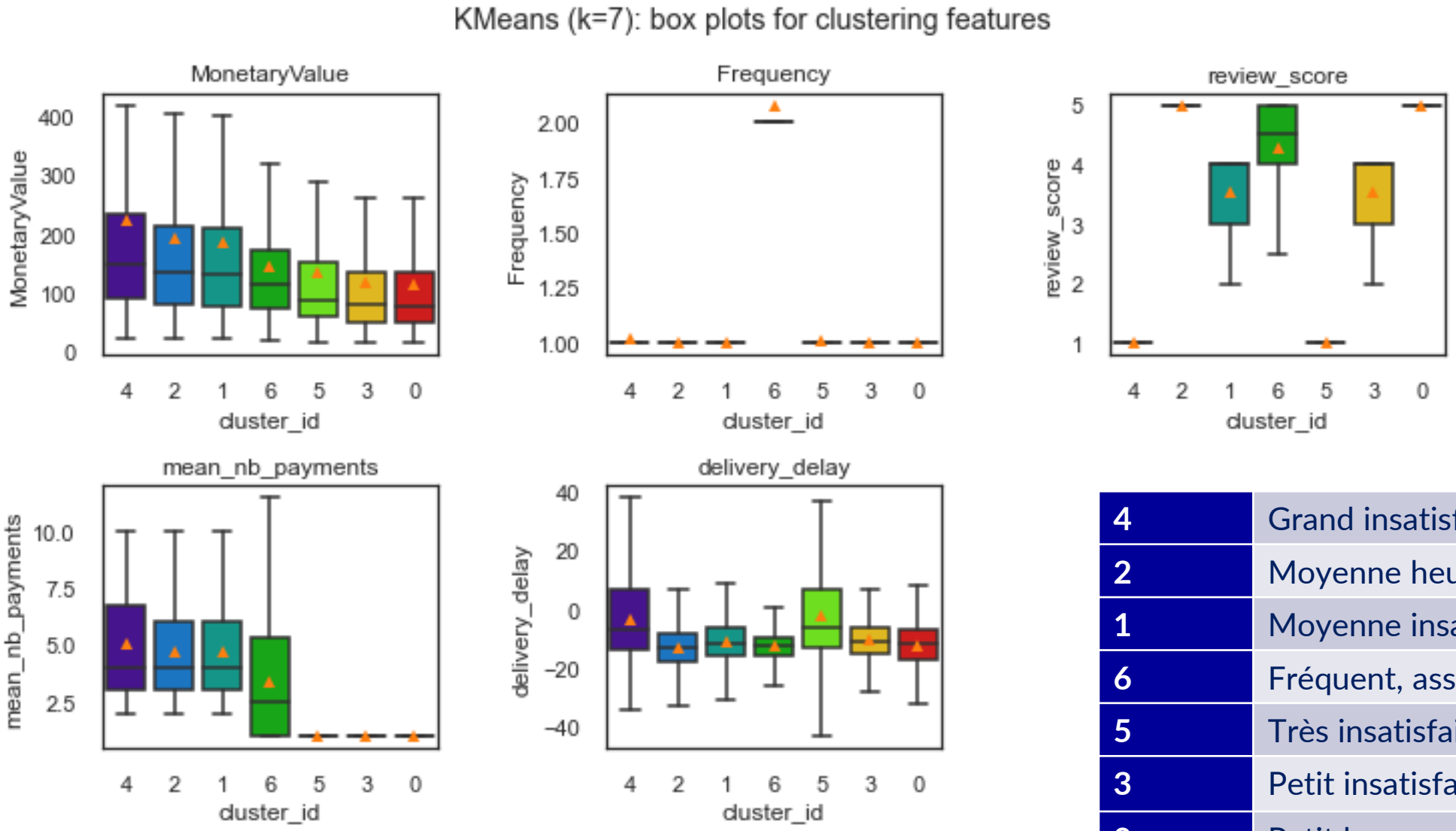
Le segment ‘fréquent’ divise en 2 segments (satisfait/insatisfait)

Les segments très insatisfait et grand insatisfait devient 1 segment seul

```
FEATURES =
['MonetaryValue',
'Frequency',
'review_score',
'mean_nb_payments',
'delivery_delay']
```



Profil des segments (février 2017-8)



4	Grand insatisfait
2	Moyenne heureux
1	Moyenne insatisfait
6	Fréquent, assez satisfait
5	Très insatisfait
3	Petit insatisfait
0	Petit heureux

Profil des segments – Moyens de paiement

Fréquent (seg 1.)

- plutôt carte de credit

Grand/moyenne dépense (Seg. 2,3,6)

- toujours carte de crédit

Petits dépensiers (seg 4,5,7)

- beaucoup de transfert bancaire (boleto)
- Beaucoup de vouchers

Seg 1	Fréquent
Seg 2	Moyenne heureux
Seg 3	Moyenne insatisfait
Seg 4	Petit heureux
Seg 5	Petit insatisfait
Seg 6	Grand insatisfait
Seg 7	Petit heureux

Profil des segments – catégories préférées

Fréquent (Seg 1)

- Parfum, santé, chambre et salle de bains

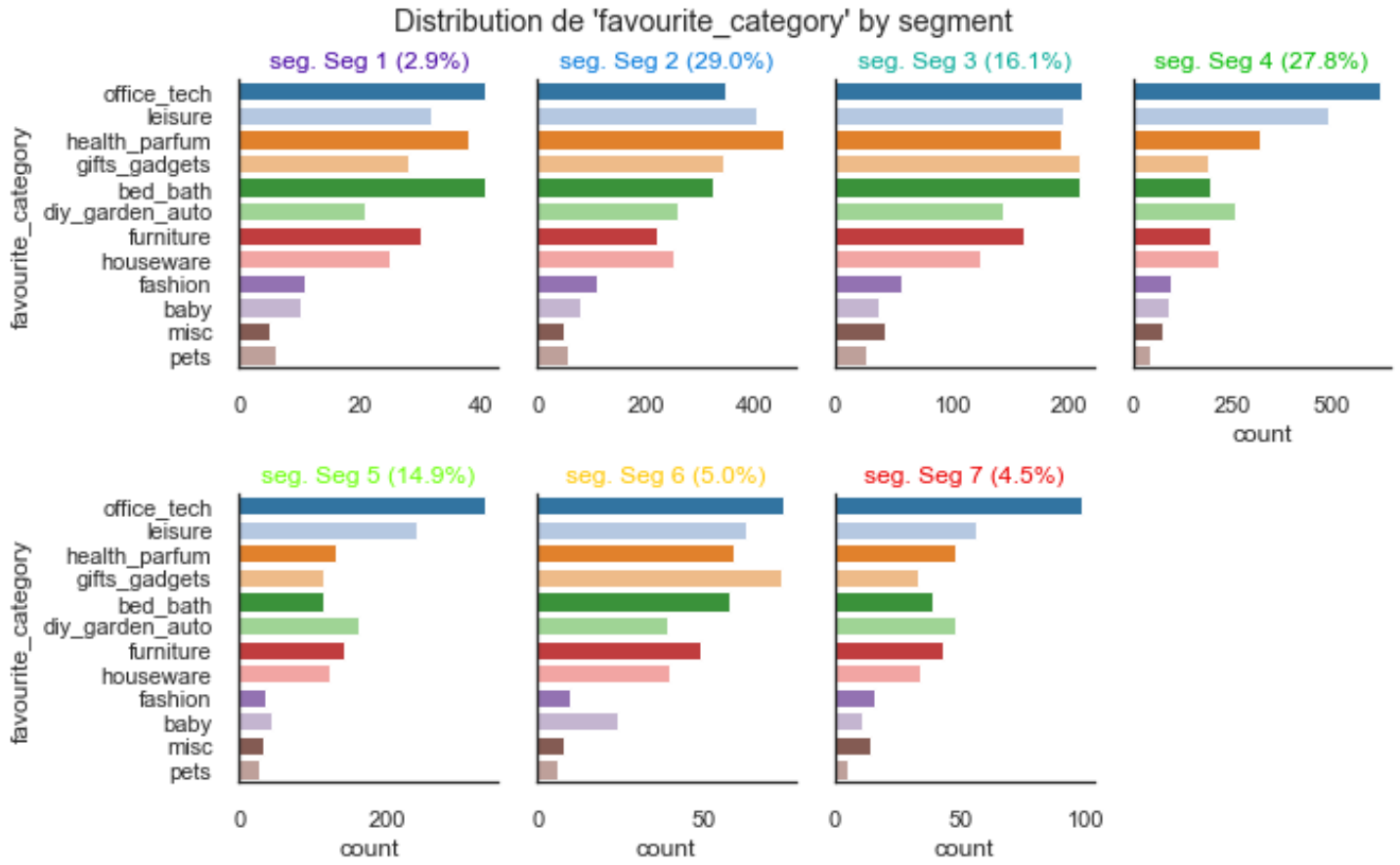
Grand/moyenne (Seg. 2,3,6)

- Parfum/santé (satisfait - 2)
- meubles (insatisfait 3 et 6)

Petits dépensiers (seg 4,5,7)

- Office tech

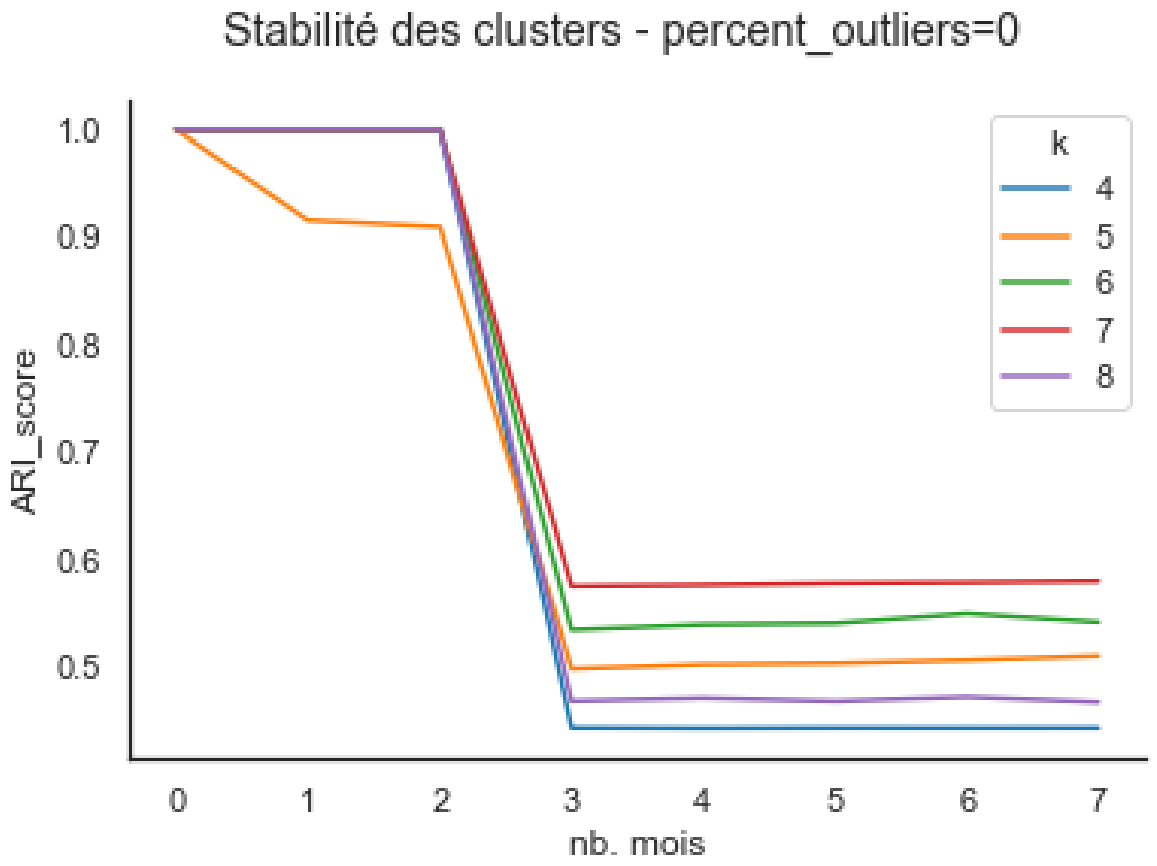
Seg 1	Fréquent
Seg 2	Moyenne heureux
Seg 3	Moyenne insatisfait
Seg 4	Petit heureux
Seg 5	Petit insatisfait
Seg 6	Grand insatisfait
Seg 7	Petit heureux



Stabilité des segments

La segmentation reste stable pendant 2 mois

- Meilleur stabilité pour 7 segments



- ARI Score = Adjusted Rand Index

Indique le similarité de la segmentation entre 2 partitions

06 Conclusion et améliorations à faire

Améliorations à faire

- Comprendre mieux pourquoi les clients sont insatisfaits.
- Ajouter des variables catégoriques aux segmentation
  - Kprototype avec les catégories préférées one-hot encoded
- Analyse NLP des commentaires et titres des ‘customer reviews’
- Segmentation par Kmeans semble très instable
  - Vérifier la stabilité des segments pour des dates de début différents

Questions

images: Mark Creasey

- mrcreasey@gmail.com

- Merci !