OpenClassrooms

Soutenance: Projet 4

Clustering Clients - RFM + Satisfaction + Categories

Edward Levavasseur

Updated: 2021/04/15

Problématique Olist

Votre objectif est de comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.

Vous devrez fournir à l'équipe marketing une description actionable de votre segmentation et de sa logique sous-jacente pour une utilisation optimale, ainsi qu'une proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps.

Plan

1. RFM + Satisfaction + Catégories
2. K-means : Recherche du nombre optimal de clusters
3. Stabilité des clusters dans le temps
4. Description des Clusters
5. Conclusion

RFM + Satisfaction + Catégories

Création d une base de données RFM

- Importation des données
- Suppression des Commandes doubles
- Transformation de la date des commandes en numérique:
  - Nombre de jours écoulés depuis le 01 Janvier 2015

- Pour chaque "customer unique id":
  - **R**ecency: dernière commande
  - **F**requency: nombre total de commandes
  - A**M**ount: Somme des dépenses

  - Satisfaction: Review score moyen

Ajout des Catégories

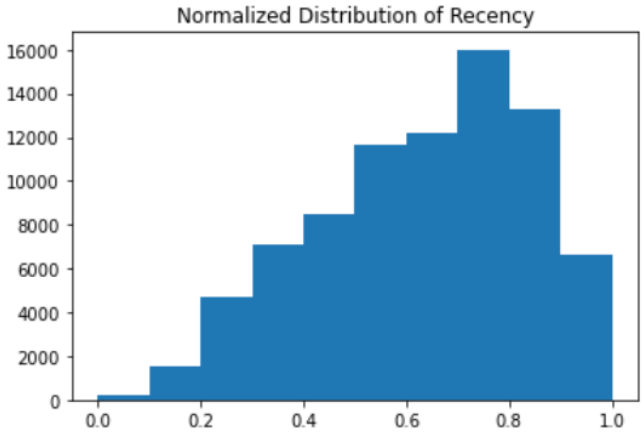
- Variables Catégoriques:
  - Ville
  - Etat
  - Type de Payment
  - Type de produit

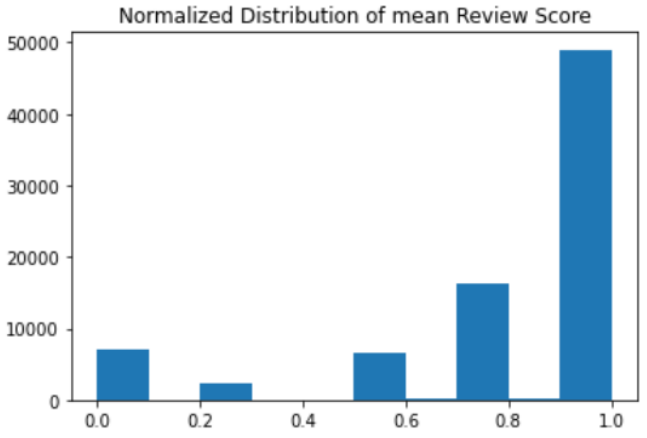
- Target Encoding:
  - 4 variables Target (R+F+M+Satisfaction)
  - 4 variables Catégoriques
  - $\Longrightarrow \quad 4 \times 4 = 16 \quad$ variables target encoding

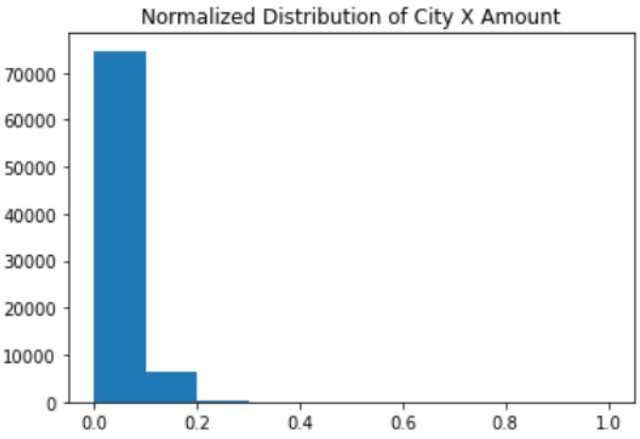
- Base de données RFM+Satisfaction+Catégories
  - 4+16 = 20 variables

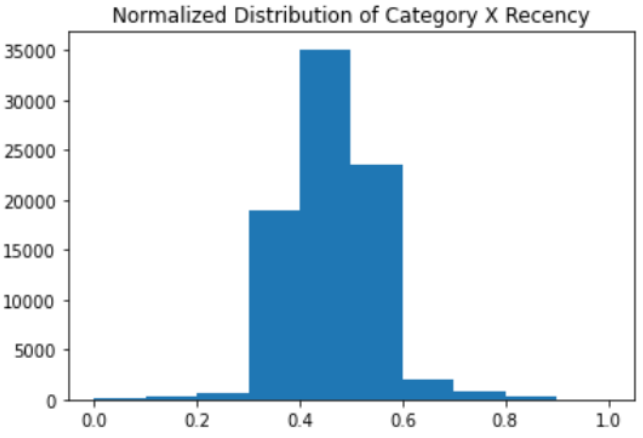
K-means : Recherche du nombre optimal de clusters

Normalization des variables

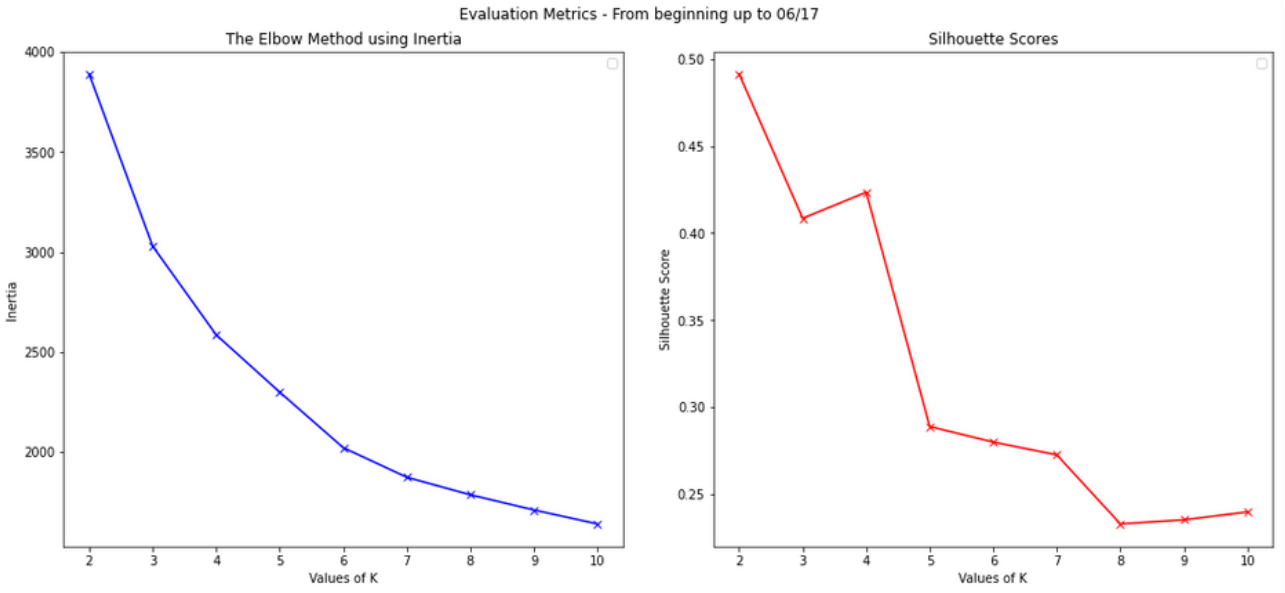


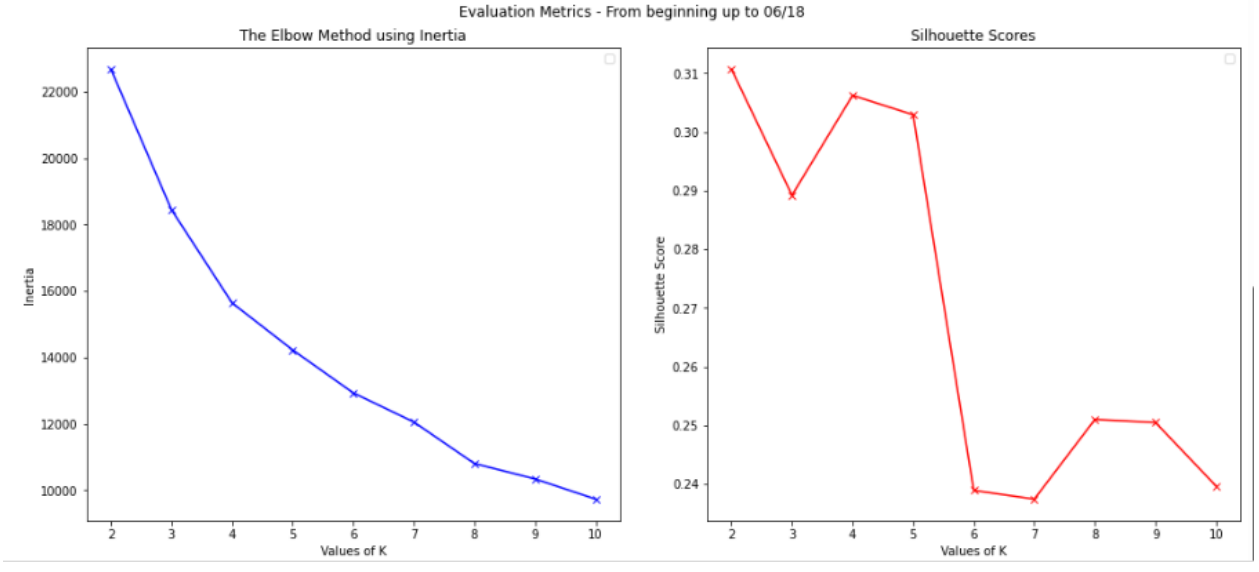




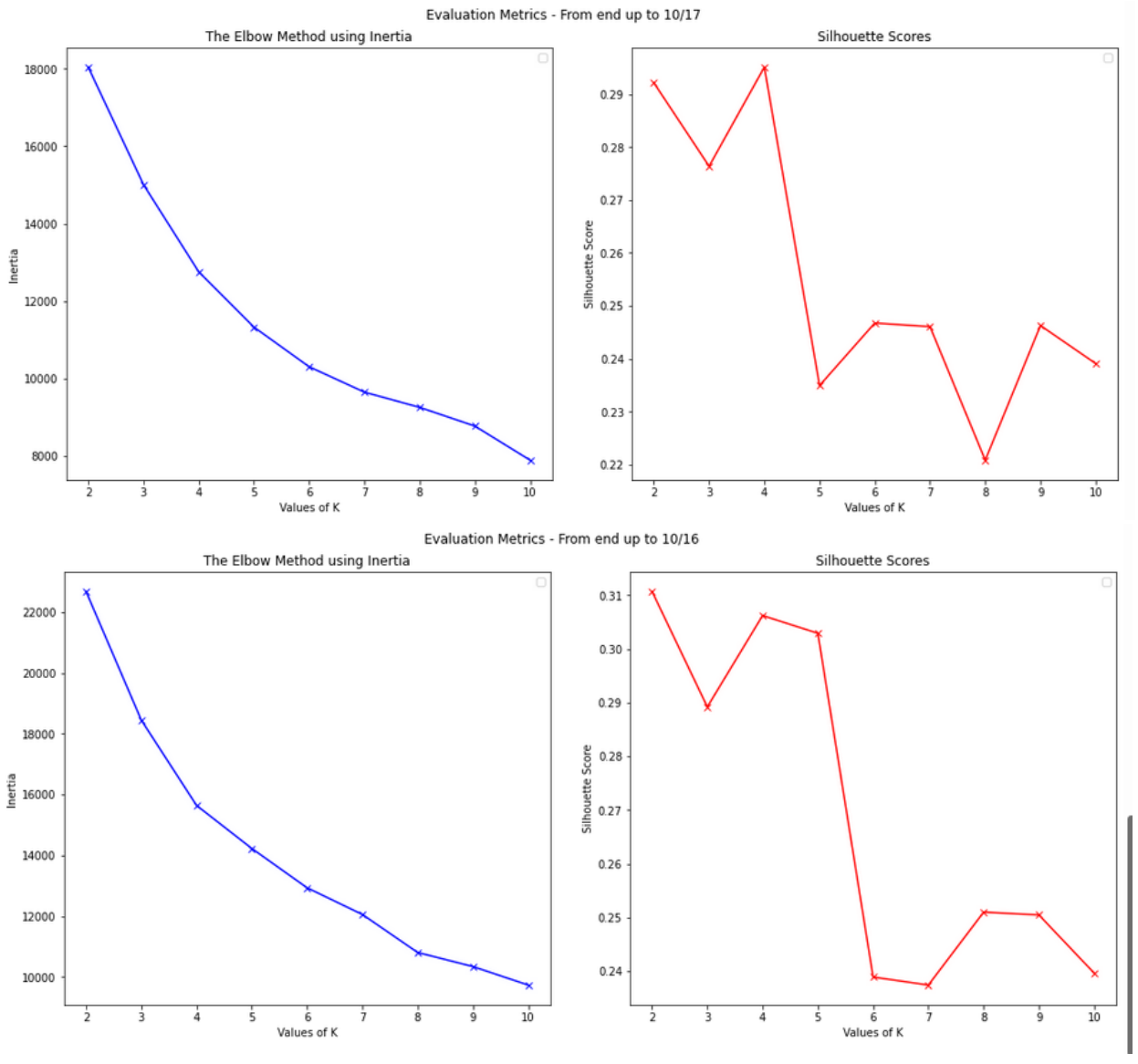


Clusters Optimaux Chronologique (1/2)





Clusters Optimaux Anti-Chronologique (2/2)



Décision

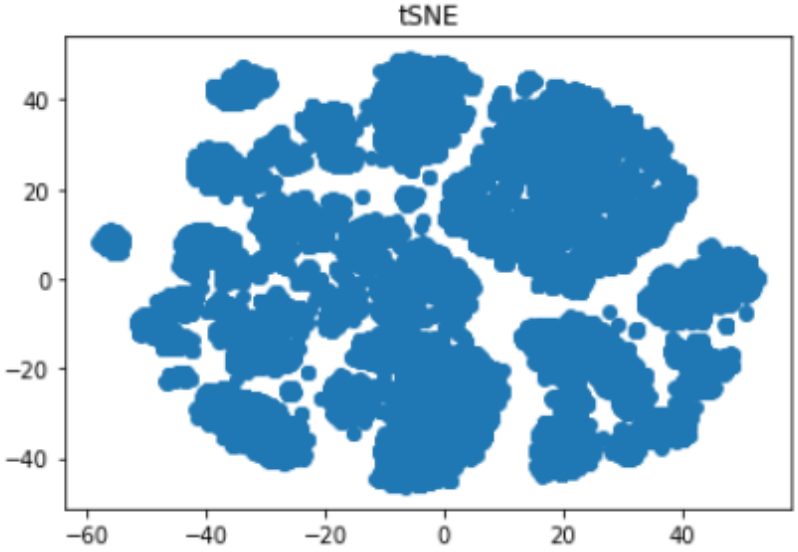
- Pas de "Elbow" très distinct
- Score de Silhouette souvent maximisé pour 4 clusters
- Décison $\Longrightarrow$ 4 Clusters

Représentation des Clusters sur tSNE

- tSNE:
  - Calcul de la distance euclidienne de chaque client, par rapport aux autres clients
  - Construction d'une matrice de similarité des clients entre eux
  - Recherche itérativement à rapprocher les clients similaires, et éloigner les clients différents sur un graphique de moindre dimensions

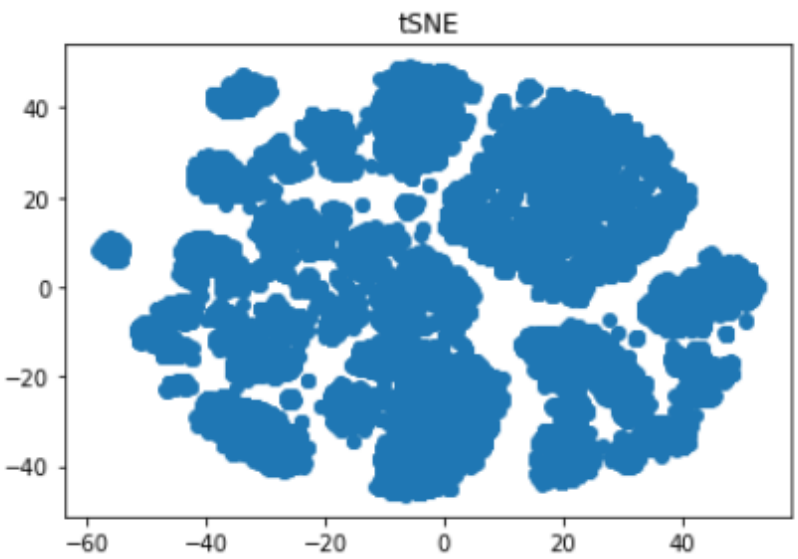
Représentation des Clusters sur tSNE

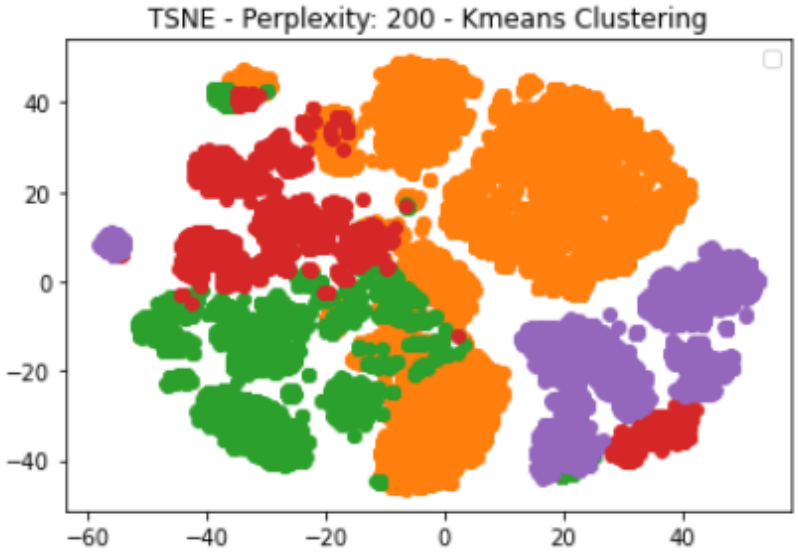
- tSNE:
  - Calcul de la distance euclidienne de chaque client, par rapport aux autres clients
  - Construction d'une matrice de similarité des clients entre eux
  - Recherche itérativement à rapprocher les clients similaires, et éloigner les clients différents sur un graphique de moindre dimensions



Représentation des Clusters sur tSNE

- tSNE:
  - Calcul de la distance euclidienne de chaque client, par rapport aux autres clients
  - Construction d'une matrice de similarité des clients entre eux
  - Recherche itérativement à rapprocher les clients similaires, et éloigner les clients différents sur un graphique de moindre dimensions





Stabilité des clusters dans le temps

Fitter le K-means sur des donnés plus anciennes

- Fitting du K-Means sur des données en $t - n$, et prédiction des clusters sur des données en t

Fitter le K-means sur des donnés plus anciennes

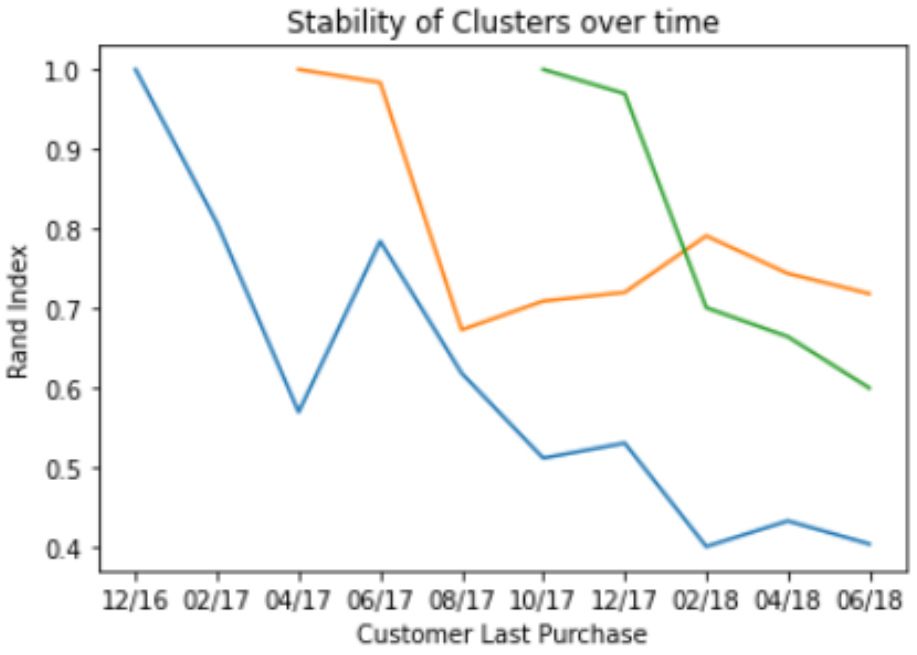
- Fitting du K-Means sur des données en $t - n$, et prédiction des clusters sur des données en t
- Fitting du K-means sur des données en t, et prédiction des clusters sur les donénes en t

Fitter le K-means sur des donnés plus anciennes

- Fitting du K-Means sur des données en $t - n$, et prédiction des clusters sur des données en t
- Fitting du K-means sur des données en t, et prédiction des clusters sur les donénes en t
- Comparaison des performances du fitting en $t - n$ et du fitting en t pour la prédiction des clusters en t

- Evaluation des erreurs de prédiction du "fitting" en $t - n$, en utilisant le Rand Index

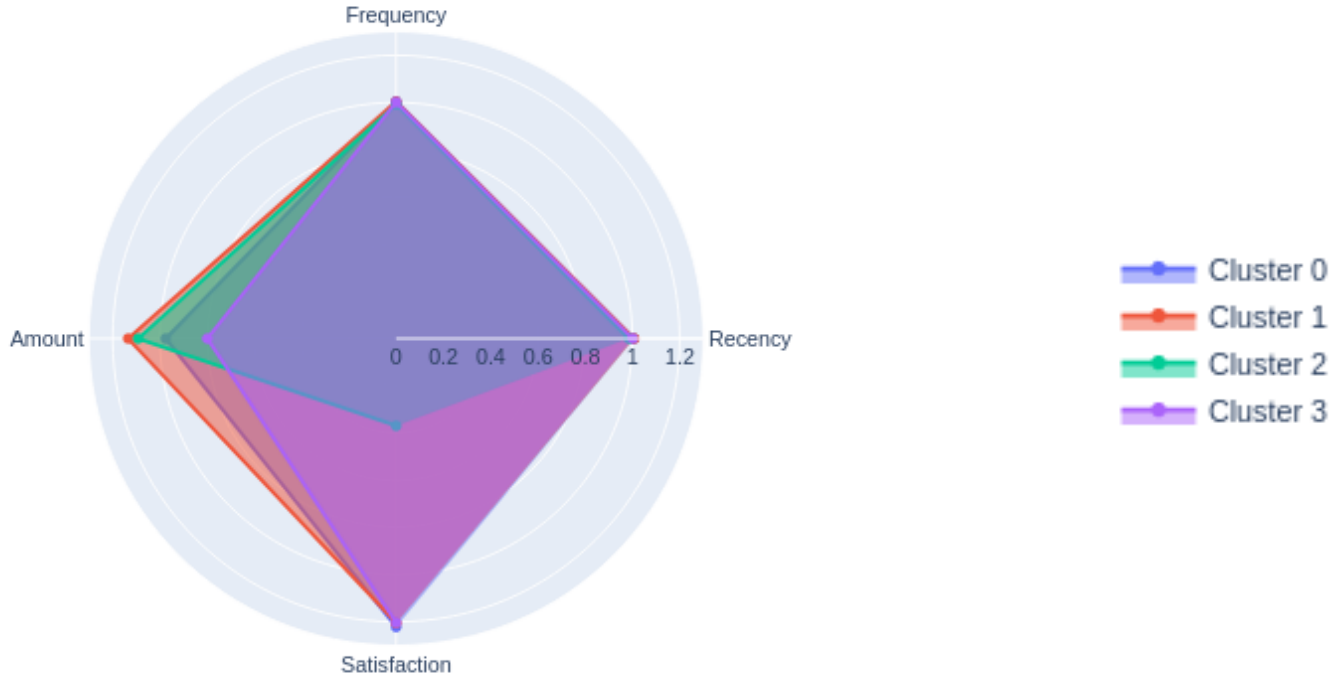
Rand Index - Pourcentage d'erreurs de prédiction



- Plus le "fitting" est ancien, plus le clustering est erronné
- Recommandation: Mettre à jour le clustering impérativement avant 4 mois, et avant 2 mois pour des performances optimales

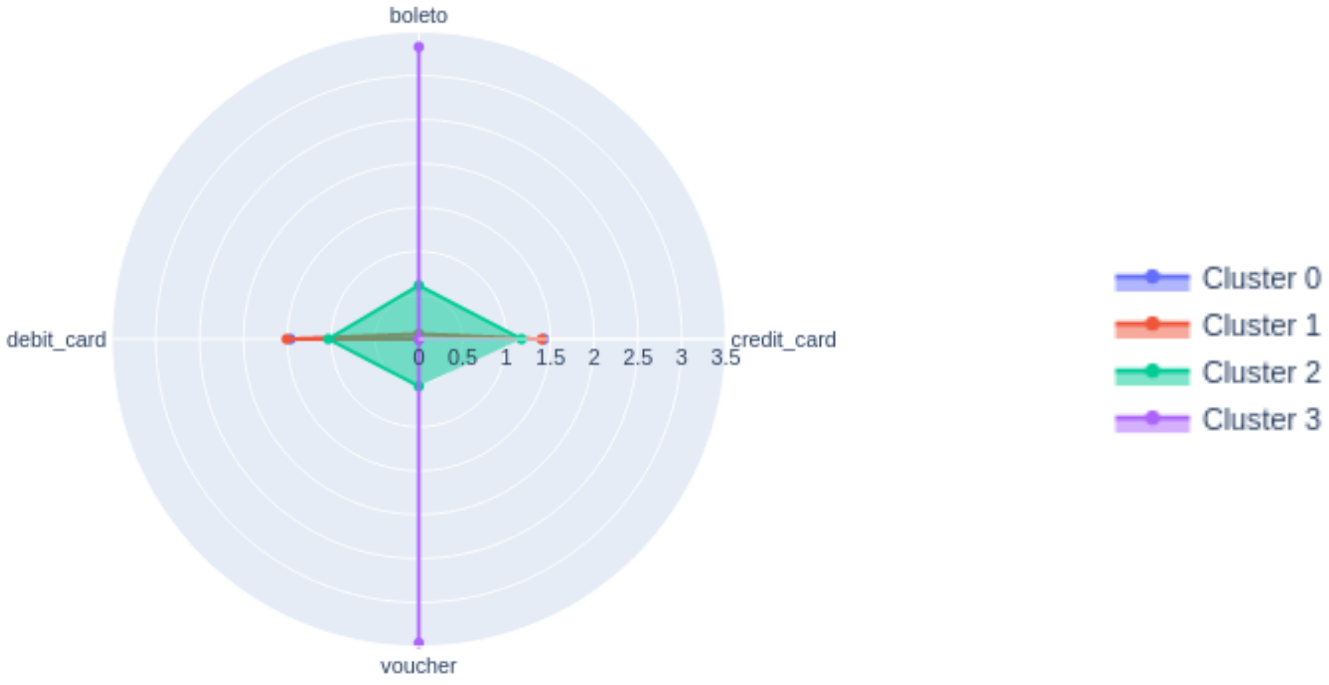
Description des Clusters

Description des Clusters: RFM + Satisfaction



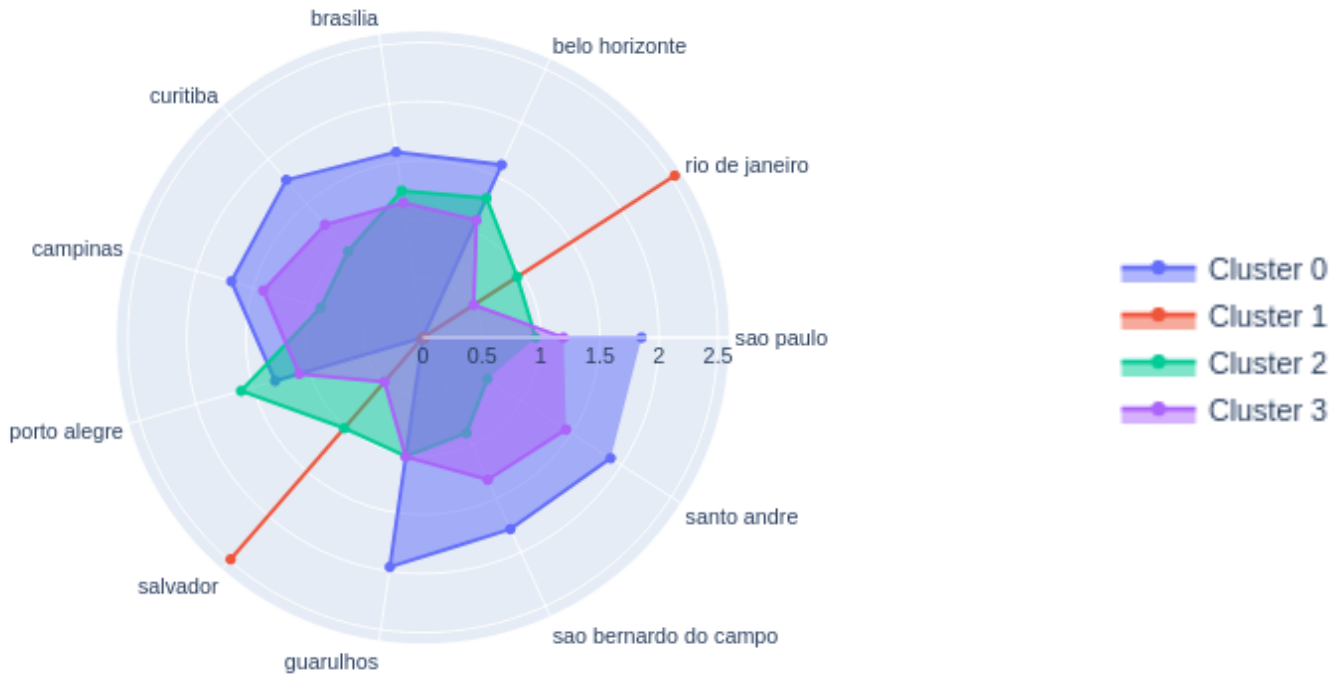
- Différence de Montant: Cluster1 > Clustero > Cluster3
- Insatisfaction: Cluster 2

Description des Clusters: Payment



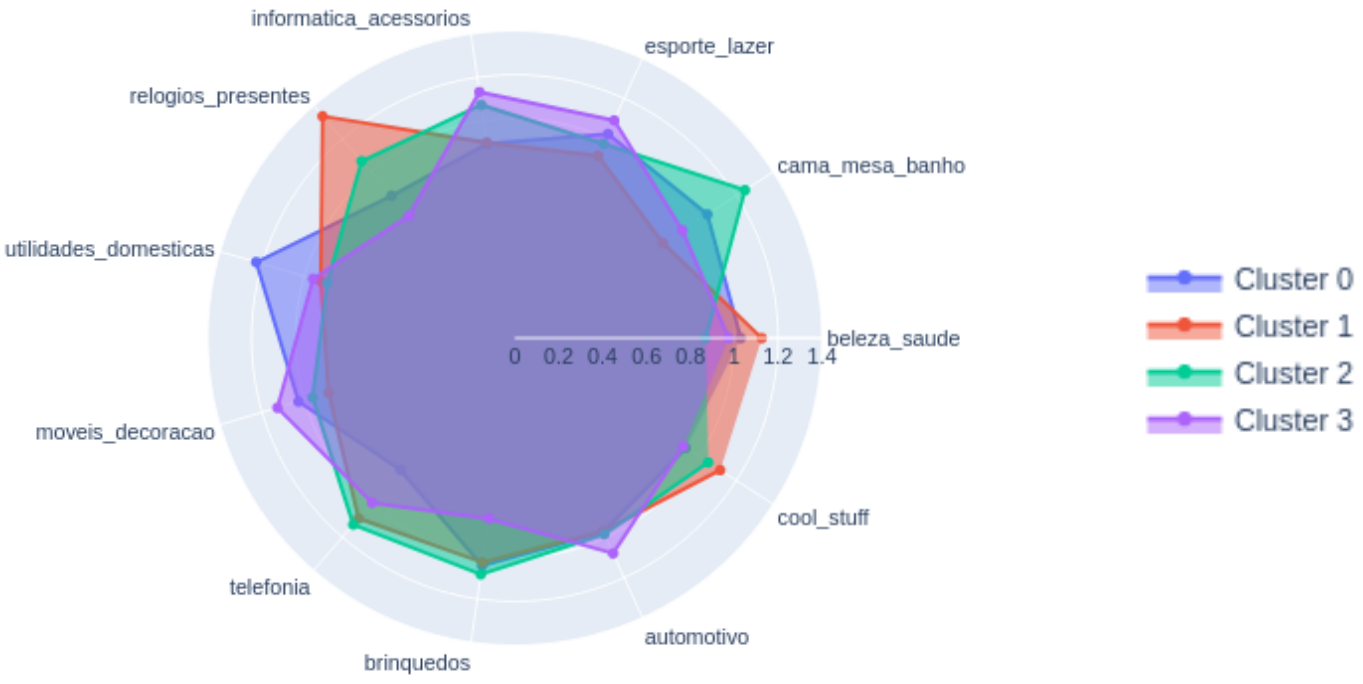
- Cluster 1 (Riche): Paye par carte crédit / débit
- Cluster 3 (Pauvre): Paye par moyen alternatif (Boletto /voucher)
- Cluster 2 (Insatisfait): utilise tous les moyens

Description des Clusters: Ville



- Cluster 1 (Riche): Rio de Janeiro + Salvador
- Cluster 3 (Pauvre): Partout, mais moins à Rio et Salvador
- Cluster 2 (Insatisfait): Sur-représentés à Porto Alegre

Description des Clusters: Types de produits



- Cluster 1 (Riche): Relogios presentes (Horlogerie)
- Cluster 3 (Pauvre): Un peu tout, mais peu d'horlogerie
- Cluster 2 (Insatisfait): Casa Mesa Banho (Maison, Table, Salle de Bain)

Conclusion

Conclusion

- Maintenance du clustering:
  - 2 mois: Fortement conseillé (< 20% d'erreurs)
  - 4 mois: Indispensable (≤ 45% d'erreurs)

- 4 Clusters:
  - Dépense beaucoup :
    - Rio de Janeiro, Salvador
    - Horlogerie
    - CB
  - Dépense peu :
    - Moins représentés à Rio et Salvador
    - Peu d'Horlogerie
    - Paye par Boleto / voucher
  - Insatisfaits :
    - Très présents à Porto Alegre
    - Maison, Table, Salle de bain

Recommandations

- Clients Dépensiers:
  - Continuer de proposer de l'horlogerie à la clientèle de Rio et Salvador
- Clients peu dépensiers:
  - Proposer des produtis peu chères au clients payant par boletto et voucher
- Insatisfaction:
  - Vérifier s'il n'y a pas un problème de livraison à Porto Allegre
  - Vérifier si les produits Maison, Table, Salle de Bain n'arrivent pas endommagés.