OPENCLASSROOMS CentraleSupélec

Segmentez des clients d'un site e-commerce

Formation Data Scientist - Janvier / Novembre 2022

Auteur: Eric TREGOAT

Mentor: Benjamin TARDY

Evaluateur: Florian GUILLET

Ordre du jour de la soutenance

❑ Présentation

- Présentation de la problématique, du cleaning effectué, du feature engineering et de l'exploration
- Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné
- Présentation de la simulation pour définir le délai de maintenance du modèle (contrat de maintenance)

❑ Discussion

❑ Débriefing

Problématique et démarche

❑ Problématique

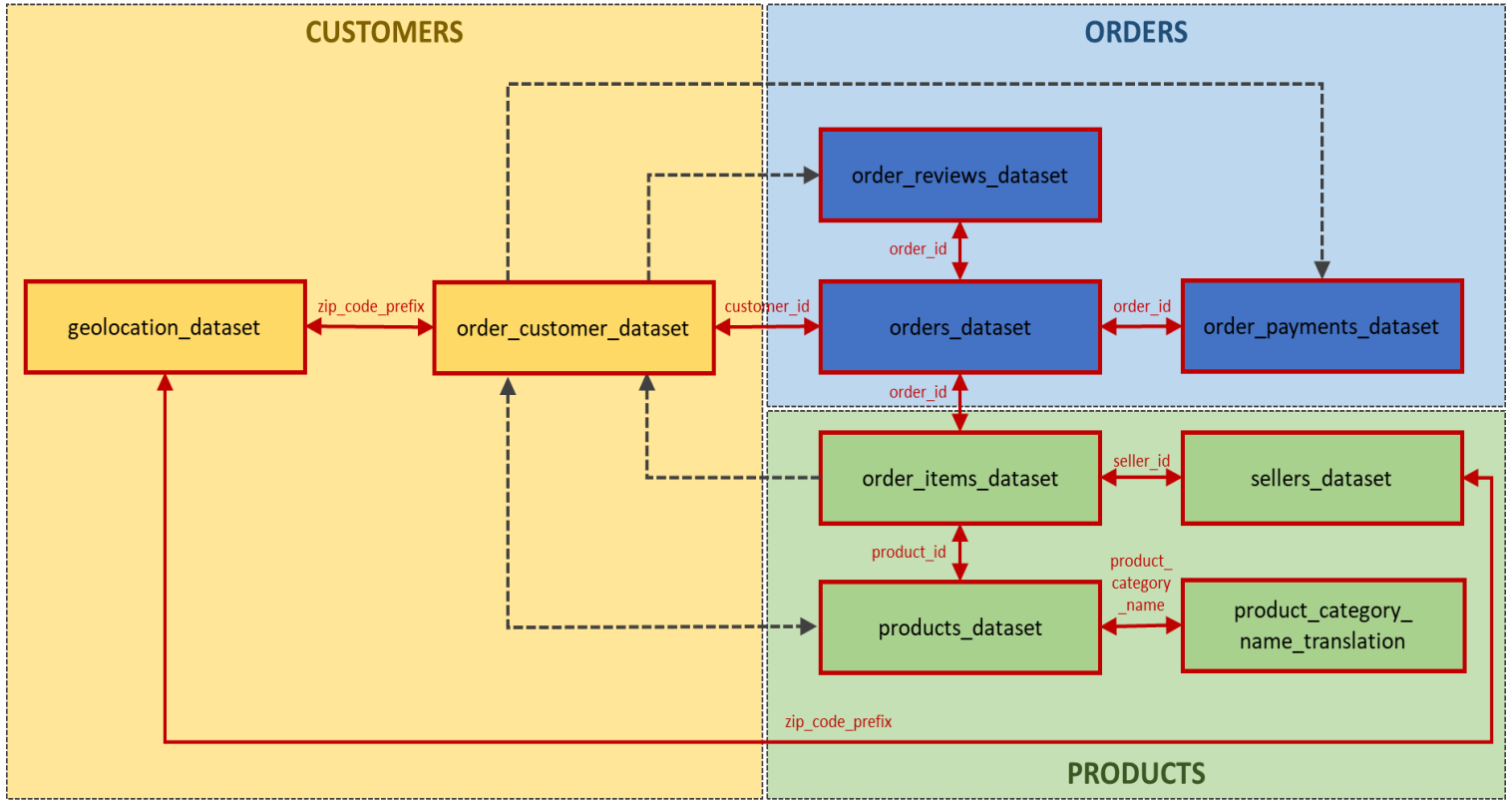
- **Objectif**: réaliser une segmentation clients à partir des données d'un site de commerce électronique
- **Données de base** : 9 fichiers relatifs aux commandes effectuées sur une période de 2 ans

❑ Démarche

1. Constituer un jeu de données clients nettoyé
2. Analyser le jeu pour en acquérir la meilleure compréhension
3. Identifier les caractéristiques utiles à la segmentation
4. Préparer les données pour le machine learning : transformation et réduction de la dimensionnalité afin de permettre la visualisation
5. Effectuer des apprentissages non supervisés avec différents algorithmes pour déterminer le clustering le plus pertinent
6. Effectuer l'analyse métier du clustering pour en valider l'usage par le marketing de l'entreprise
7. Estimer la pérennité du modèle de clustering dans le temps afin de donner de la visibilité sur son temps d'utilisation comparé au délai de réalisation des actions marketing.

Assemblage, nettoyage et filtrage du jeu de données

- **9 fichiers de données**
  - → Données relatives aux commandes, pas d'autre information (ex: socio-démographique)
- **Valeurs manquantes**
  - → Date de livraison: fixée à sa valeur estimée
  - → Commentaires de revue client: non utilisé
  - → Dimension produit: fixé à la valeur médiane mais non utilisé
- **Variables retenues initialement**
  - → Pour la constitution du RFM
  - → Informations potentiellement utiles, relatives à la **localisation** des clients, à leur **satisfaction**, aux modes de **paiement** des commandes et aux **produits** achetés
- **Fusion des fichiers**
  - → Chargement du customers dataset dans 'data'
  - → Couples de variables charnières entre paires de fichier
  - → Fusions successives avec chaque fichier (df) :
  - → `data.merge(df, how='inner', left_on=var[0], right_on=var[1])`



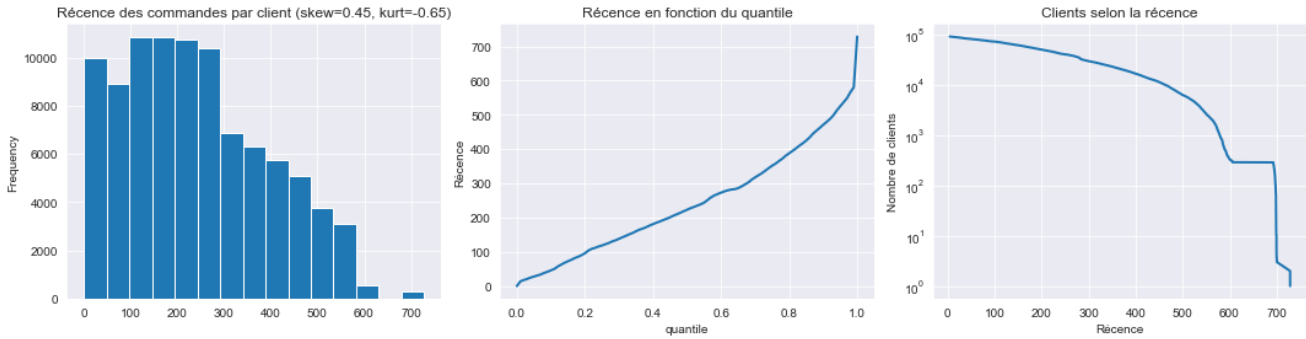
Constitution du jeu de données clients

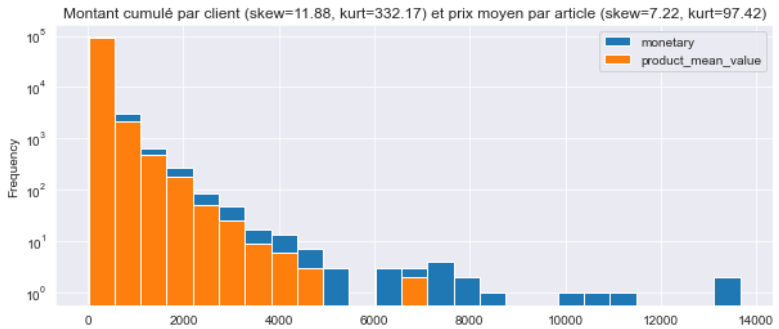
Démarche basée sur l'utilisation de 'groupby' avec 'customer_unique_id' combinée avec la variable cible et une méthode ('mean', 'first', 'count',...)

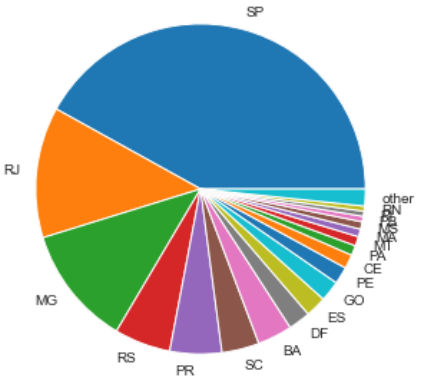
OPENCLASSROOMS CentraleSupélec

Analyse univariée

- La récence est élevée (moyenne 242 j) et marque 2 groupes:
  - Ceux qui ont commandé au cours des 600 derniers
  - Ceux qui n'ont pas effectué de commande depuis plus de 690 jours : population de 334 clients (0.36%)
- Seulement 3% des clients ont passé plus d'une commande
- Le montant par client varie fortement autour de 174 reals
  - Coefficient de dispersion de 1.51
  - Moins de 2.86% ont dépensé plus de 700 reals, quelques outliers au-delà de 6000 réals
- 20% des catégories de produit (15 sur 71) représentent plus de 80% des commandes, pas de modes sur ces catégories
- 77% des clients sont très satisfaits (score ≥ 4) et un peu moins de 15% sont insatisfaits (score < 3)
- Délais de livraison
  - 80% des commandes sont livrées sous 18 jours
  - Le délai prévisionnel est surestimé (80% des cas) et non maitrisé pour les délais longs
- Paiements
  - 3.1% des clients utilisent plusieurs moyens de paiement par commande
  - Plus de la moitié des clients paient en plusieurs fois
- Les clients sont localisés dans 27 Etats, parmi lesquels :
  - plus de 40% des clients sont localisés dans l'Etat de Sao Paulo,
  - 2/3 des clients sont localisés dans 3 Etats (SP, RJ, MG),
  - plus de 80% des clients sont localisés dans 6 Etats



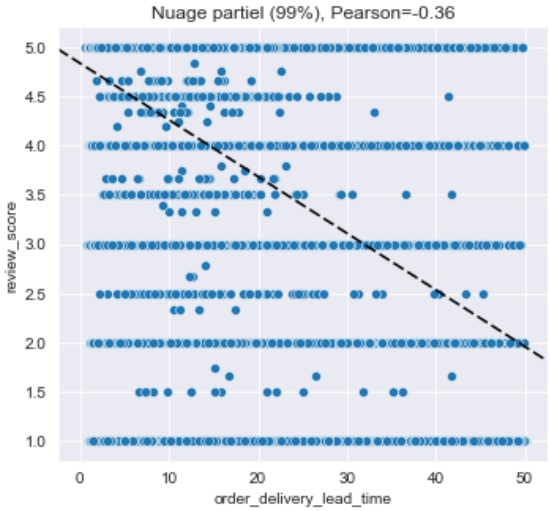


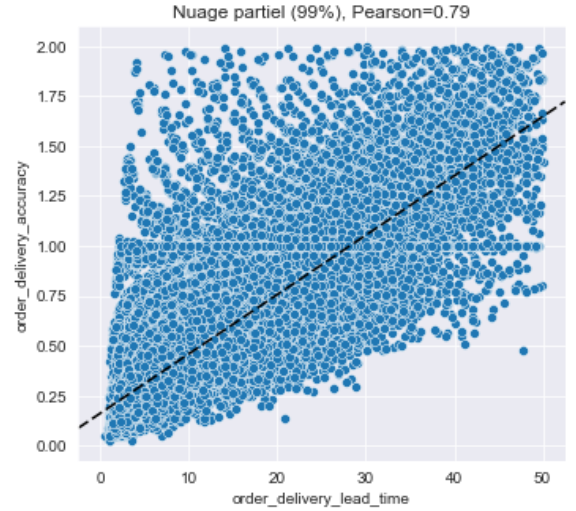


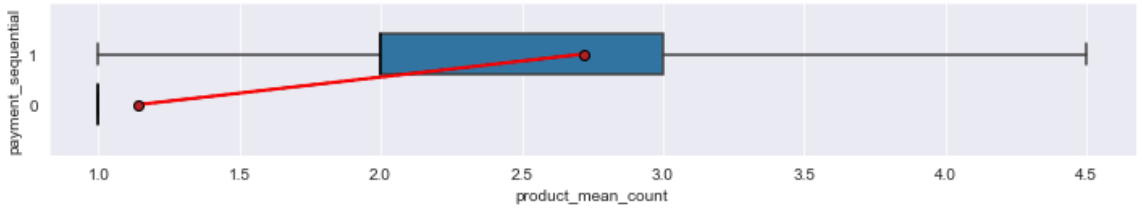
OPENCLASSROOMS
CentraleSupélec

Analyse multivariée

- Permet de réduire le nombre de variables en identifiant les variables liées et retenant celles qui sont les plus pertinentes :
  - Le nombre moyen de produits par commande et leur valeur moyenne sont abandonnées car apportent peu par rapport à la variable de montant
  - Nous conservons la variable de fréquence annuelle des commande au détriment du nombre de commandes
  - Le score de revue client est la variable qui explique le mieux la satisfaction des clients. Elle est préférée au délai moyen de livraison, qui est lui-même préféré à la précision de la prévision de livraison et au coût de livraison.
- Apporte des explications complémentaires
  - Le recours à plusieurs moyens de paiements s'effectue pour les commandes de plusieurs articles
  - Il y a une différence significative sur les coûts et délais de livraison selon les Etats, les mieux servis bénéficiant du coût le plus faible et réciproquement.







Variables conservées

→ RFM (numériques)
→ Customer State (3 valeurs + 1 pour les autres Etats)
→ Review score (numérique)
→ Payment sequential (binaire)
→ Payment installments (binaire)

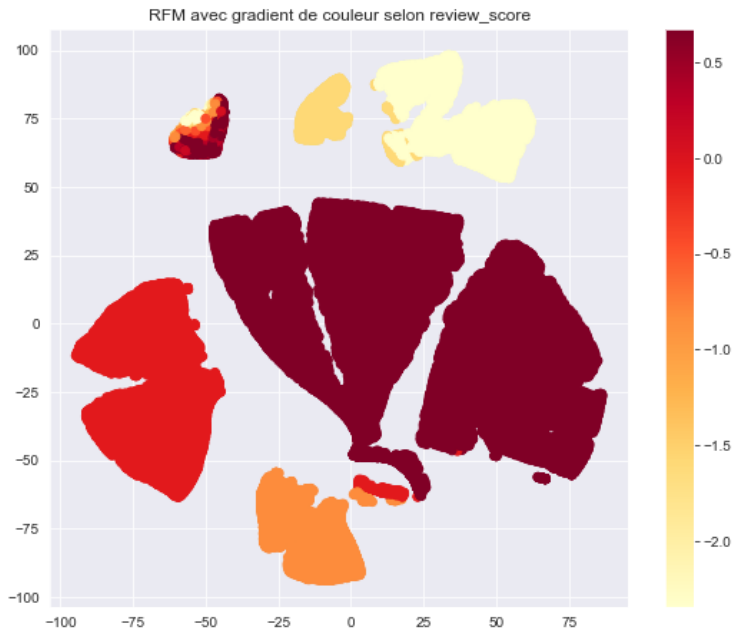
Préparation au machine learning

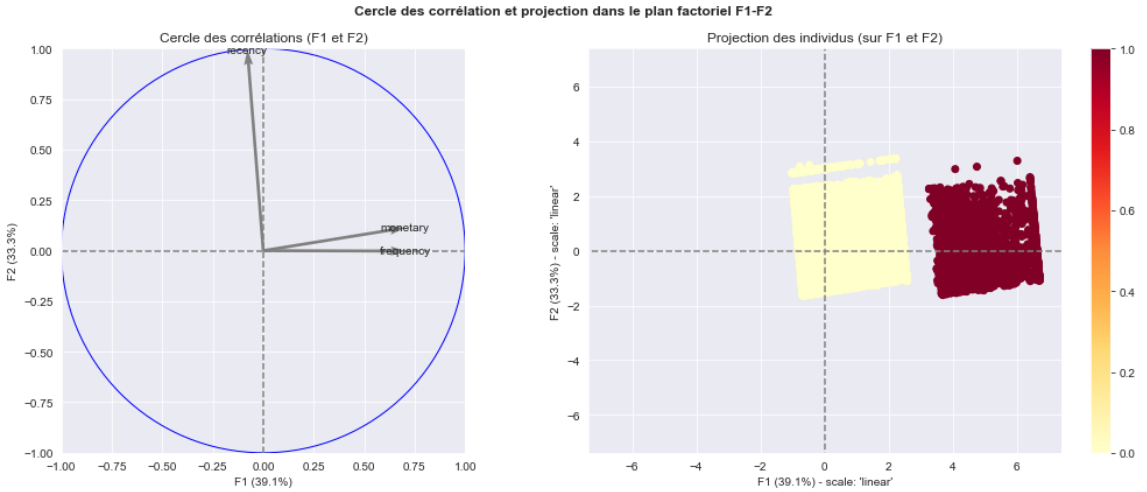
❑ Transformation des features

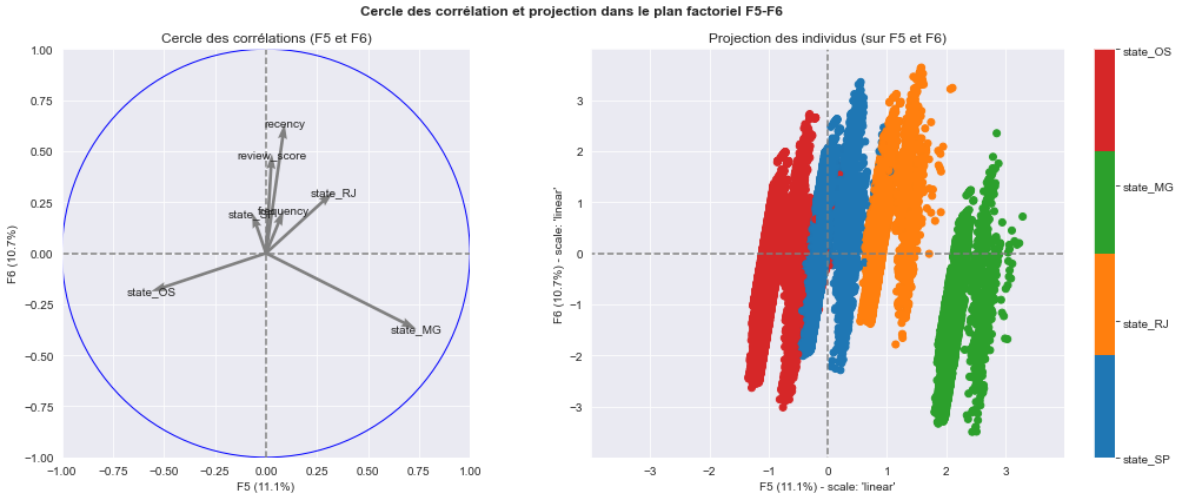
- Catégorielles :
  - → One-hot encoding → variables binaires par catégorie
  - → Division par la racine carrée de la fréquence de 1 (FAMD)
    Source: [FAMD: How to generalize PCA to categorical and numerical data]
  - → Centrage
- Numériques :
  - → Ecrêtage de la fréquence à 1 (plusieurs commandes) et du montant à 700
  - → StandardScaler

❑ Visualisation de la population

- ACP linéaire sur variables numériques
- ACP linéaire mixte (FAMD)
- t-SNE



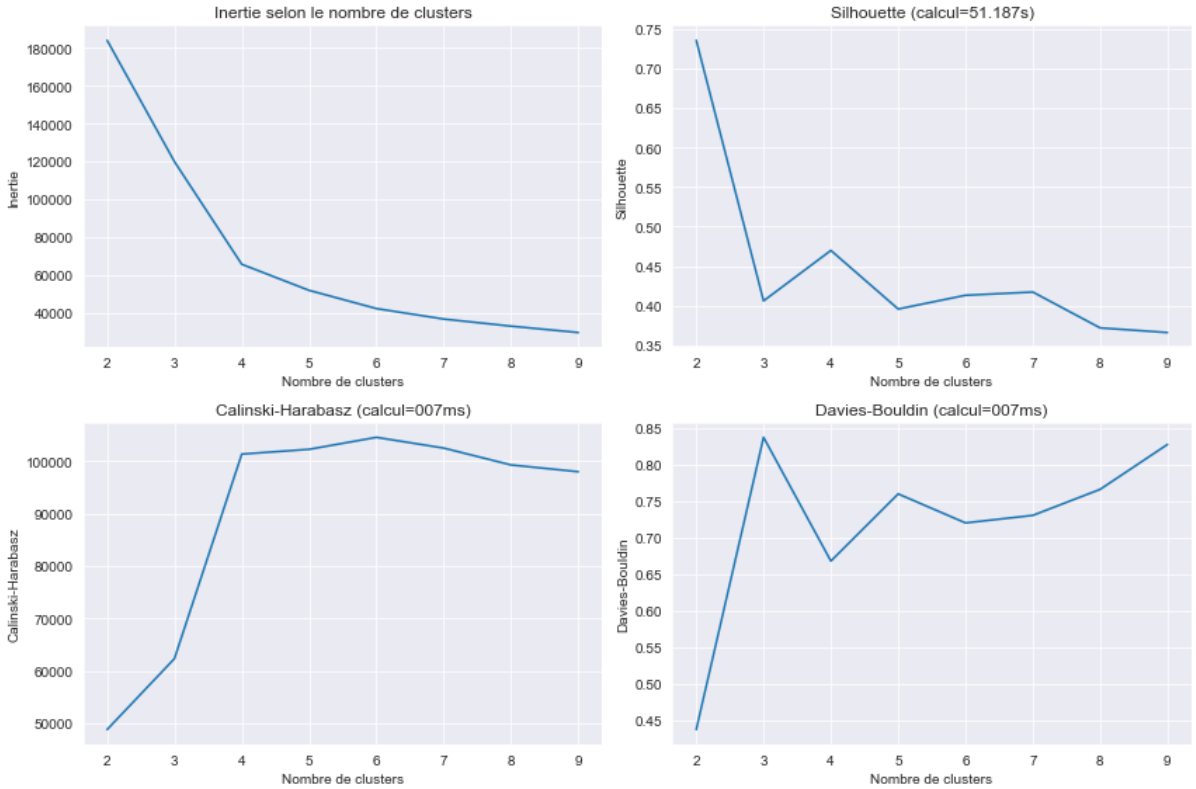


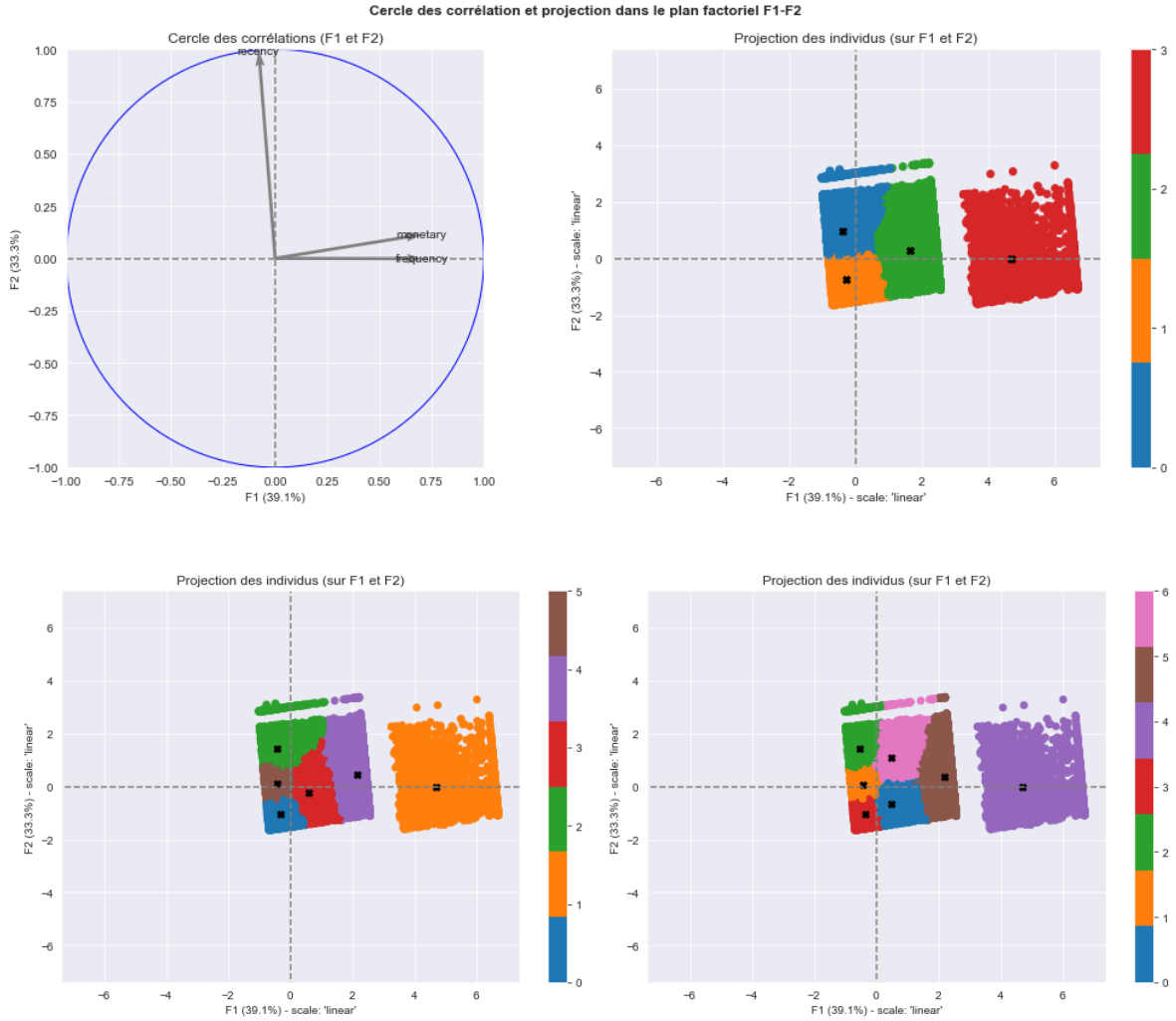


OPENCLASSROOMS
CentraleSupélec

Test de l'algorithme k-Means

- L'algorithme qui donne les **meilleurs résultats**
- Meilleurs clusterings pour **k = 4**, 6 et 7
- Possibilité d'enrichir avec des features supplémentaires :
  - → Ajout de 'customer_review' (5 clusters)
  - → Ajout de 'payment_installments' (10 clusters)



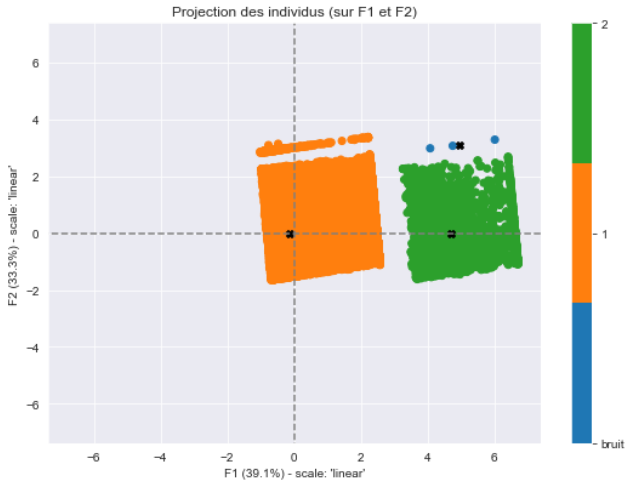


OPENCLASSROOMS
CentraleSupélec

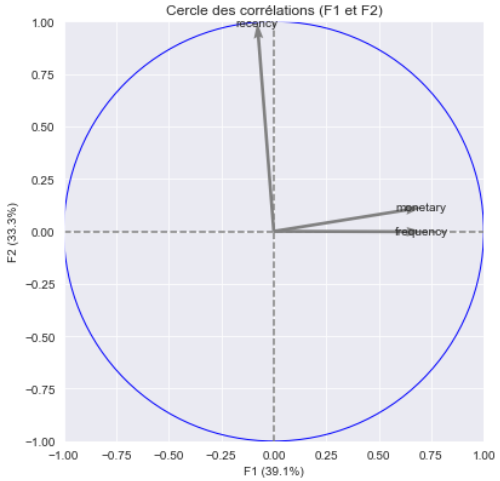
Test de l'algorithme DBSCAN

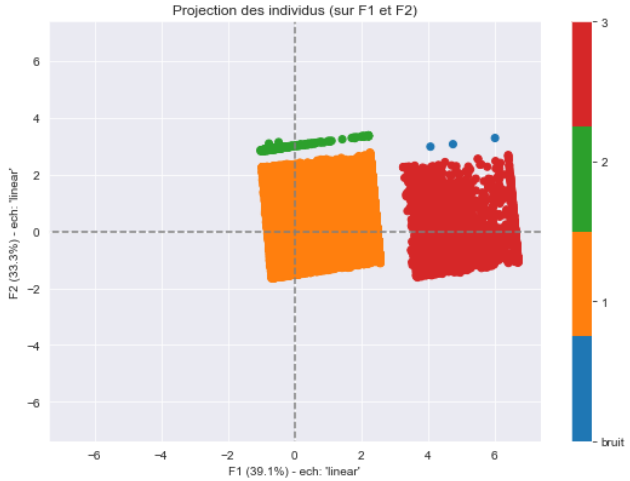
Résultat non probant

→ Variations de densité de population selon le montant et la récence

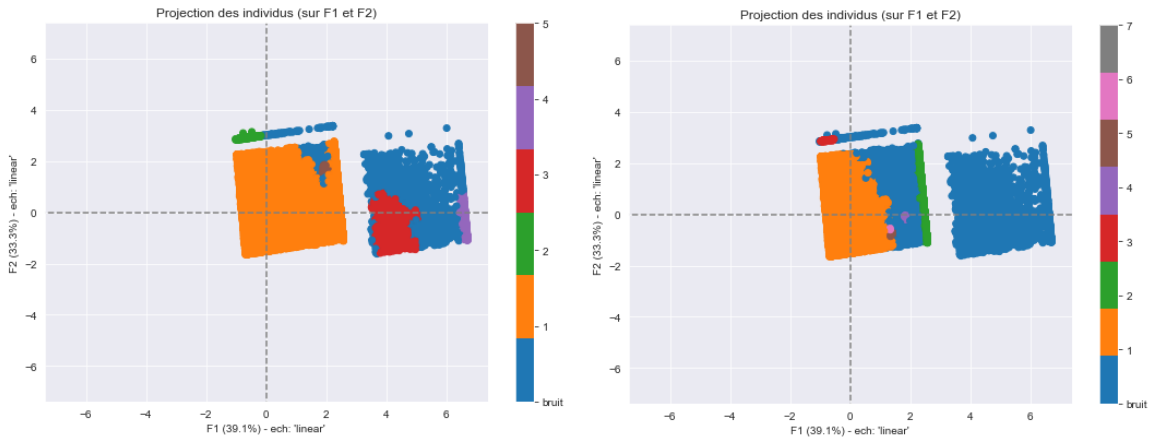


Cercle des corrélation et projection dans le plan factoriel F1-F2

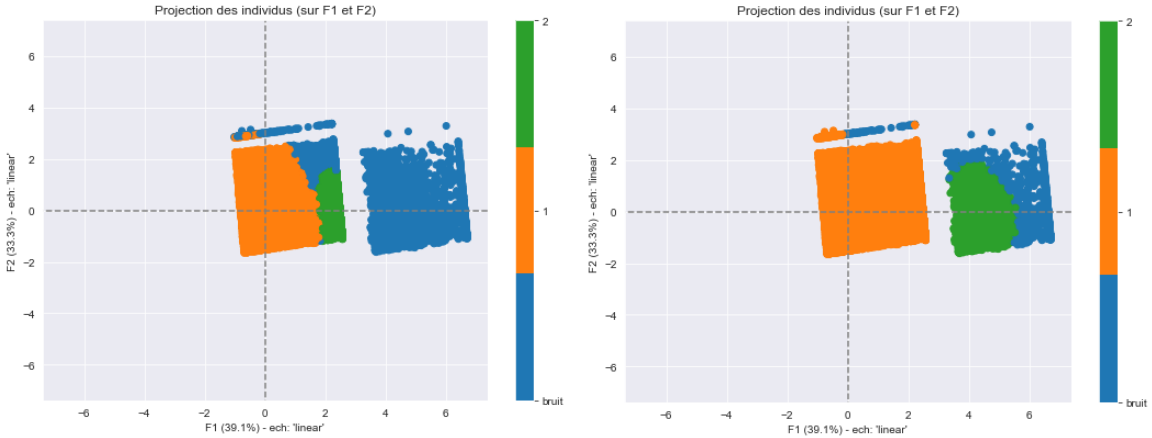




Effet de la diminution de epsilon



Effet de l'augmentation de n_samples

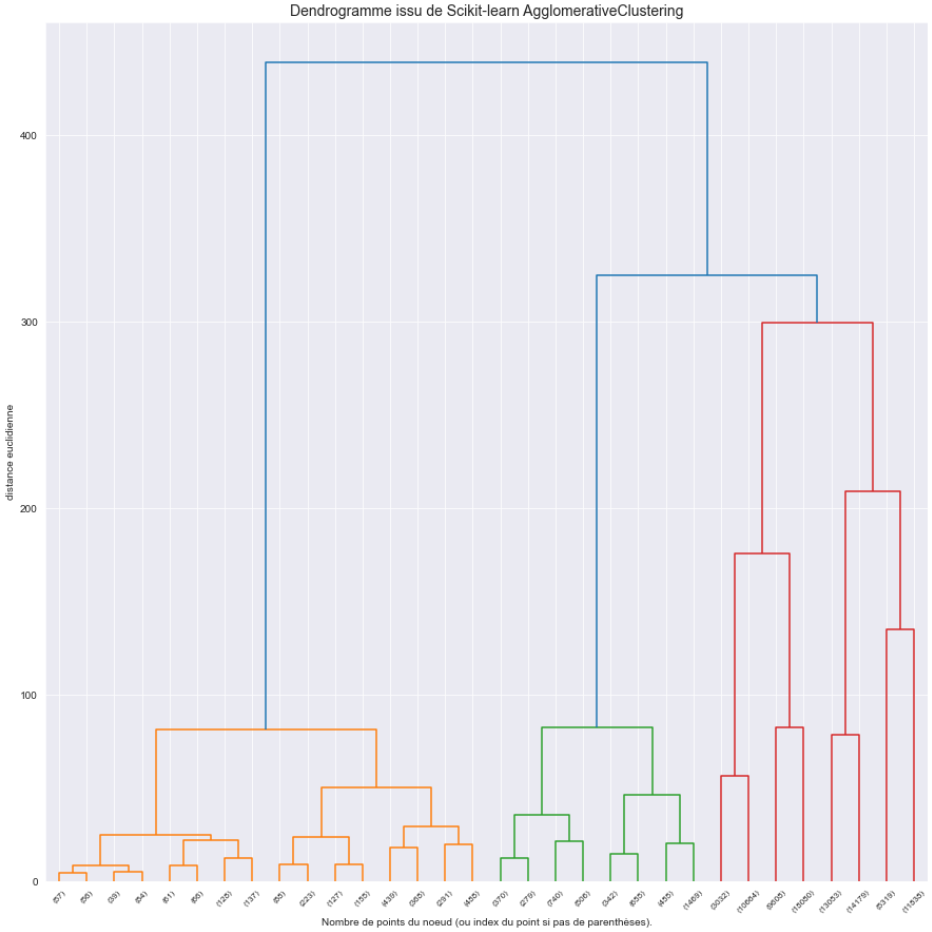


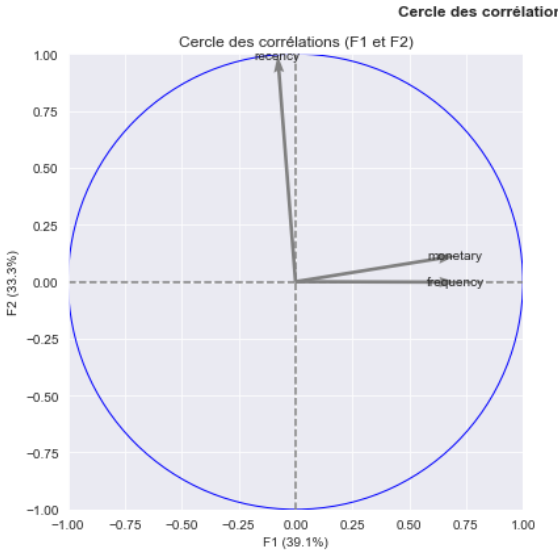
OPENCLASSROOMS
CentraleSupélec

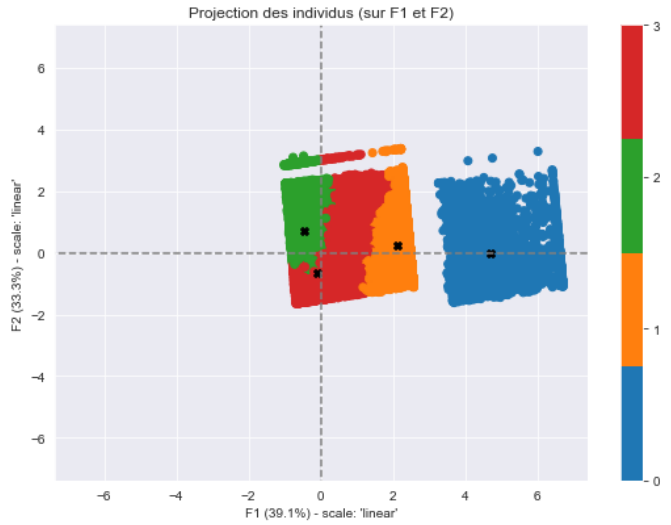
Test d'algorithmes hiérarchiques agglomératifs

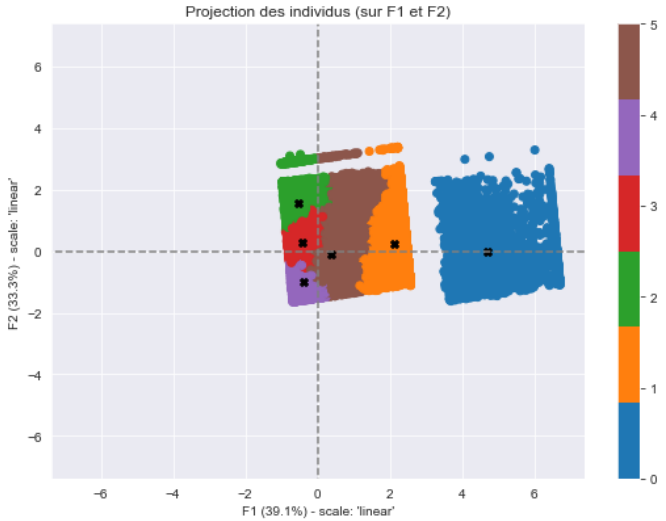
Résultat intéressant mais indices moins bon que pour k-Means

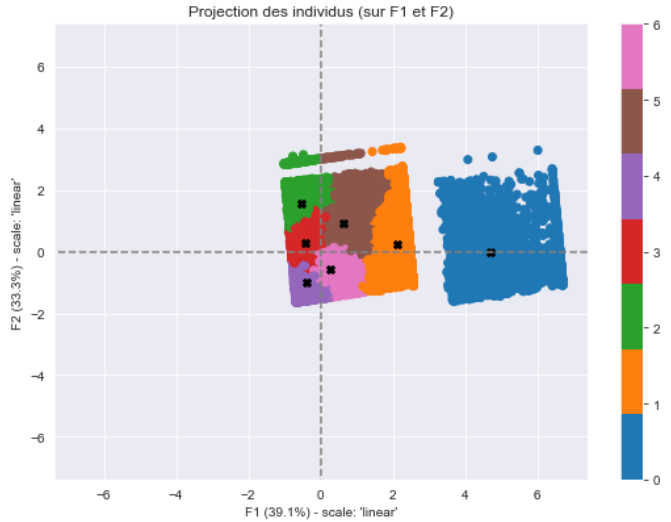
- → Clustering préférentiel selon indices: 7 > 4 > 6
- → Le dendrogramme permet de guider le choix
- → Clusters moins faciles à interpréter pour la segmentation









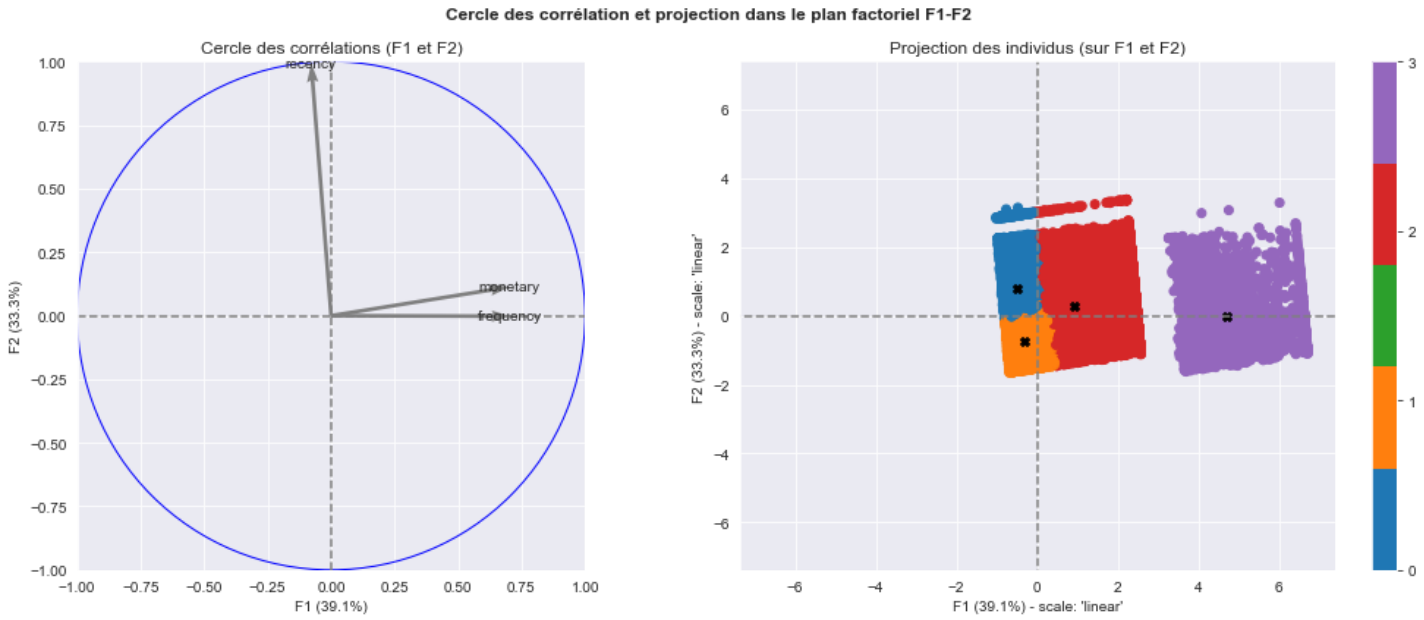


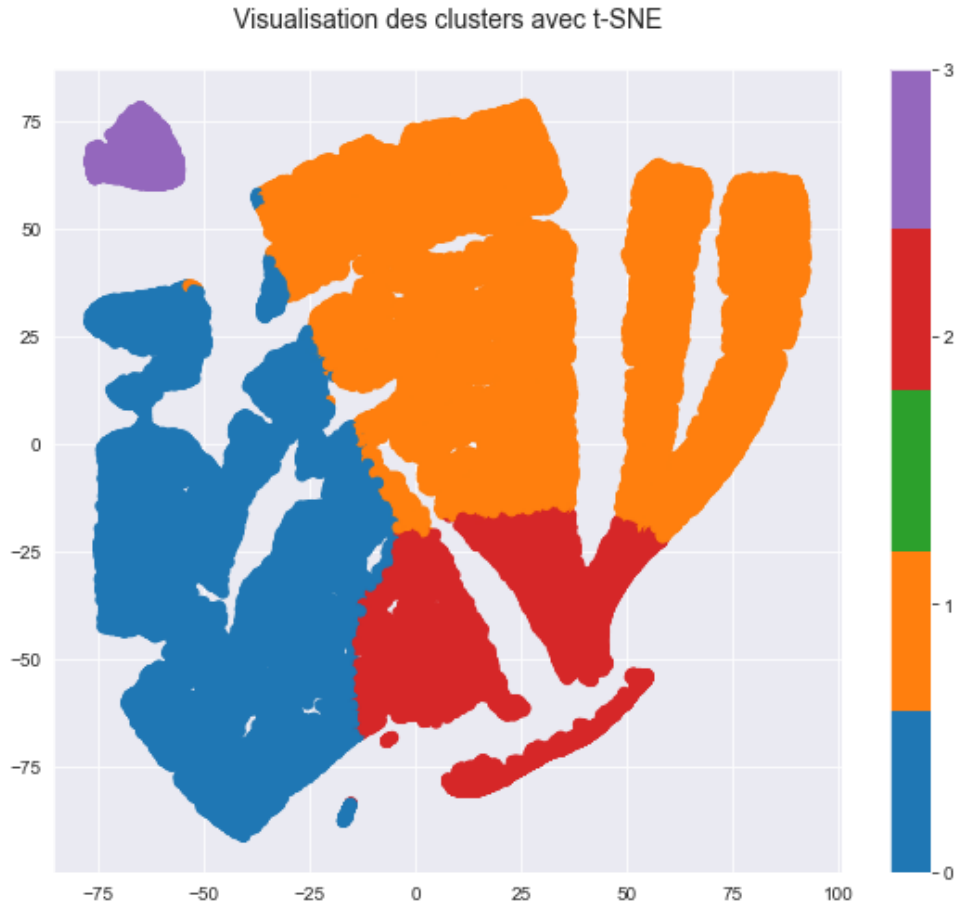
OPENCLASSROOMS
CentraleSupélec

Test de l’algorithme GMM

Résultat intéressant avec des indices légèrement moins bon que pour k-Means

- → Principe de découpage similaire à k-Means avec un seuil de montant et un seuil de récence (visible sur ACP et t-SNE)
- → Seuils différents de ceux du k-Means





OPENCLASSROOMS
CentraleSupélec

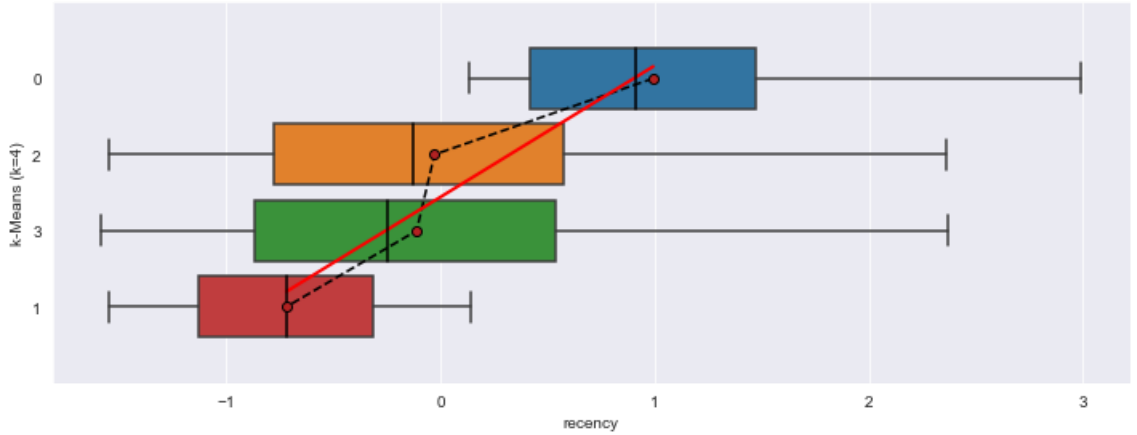
Analyse métier avec le clustering de k-Means pour k=4

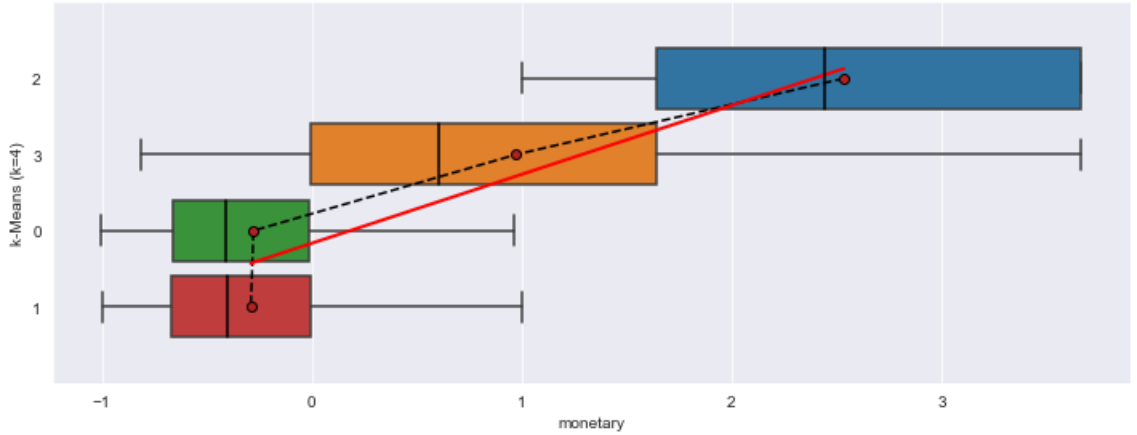
Mesures et analyses orientées métiers :

- Calcul des données caractéristiques des clusters : population, profil moyen et min / max
- ANOVA pour visualiser les clusters en fonctions des features
- Explicitation des segments en fonction des clusters
- Appréciation des segments pour évaluer la pertinence de la segmentation

k = 4				
Cluster	Taille	Fréquence	Récence	Montant
0	37.4%	1	Faible	Faible
1	50.9%	1	Elevé	Faible
2	8.8%	1	Tous	Elevé
3	3.0%	2+	Tous	Tous

cluster	size	min_recency	centroid_recency	max_recency	min_frequency	centroid_frequency	max_frequency	min_monetary	centroid_monetary	max_monetary
0	34905	262.607	394.752	728.494	0.5	0.5	0.5	10.07	117.051	455.01
1	47503	4.754	132.381	264.052	0.5	0.5	0.5	10.89	115.209	365.38
2	8181	4.784	238.105	698.822	0.5	0.5	0.5	306.64	532.516	700.00
3	2807	0.000	225.403	696.027	1.0	1.0	1.0	37.34	302.214	700.00





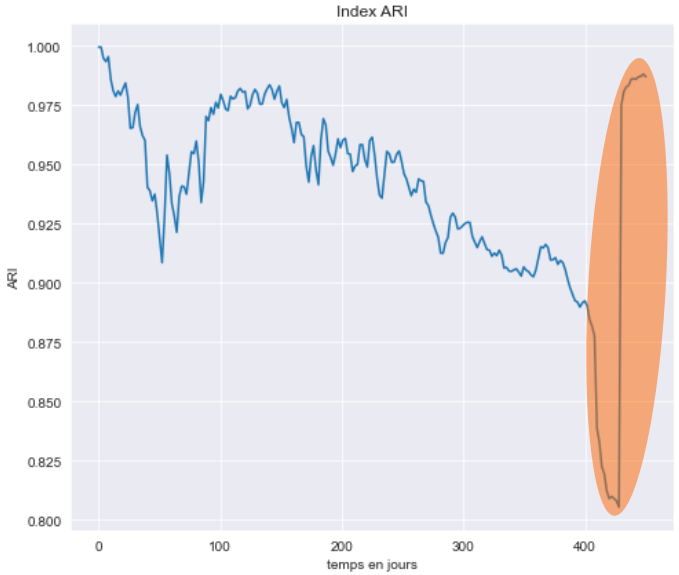
OPENCLASSROOMS CentraleSupélec

Choix de l'algorithme et du clustering

- **Choix de l'algorithme k-Means**
  - k-Means donne les meilleurs résultats en termes d'indice (coef de silhouette, indices de Calinski-Harabasz et Davies-Bouldin) et temps de calcul
  - Les clustering de k-Means permettent l'analyse métier
- **Choix de 4 clusters avec les features RFM** :
  - Le découpage en 6 et 7 clusters ne produisent pas des segmentations nécessairement plus fines (cf les couleurs des segments).
  - Il serait en revanche intéressant de rechercher une segmentation plus fine avec d'autres features, tout d'abord la satisfaction client avec review score, puis les moyens de paiement et éventuellement l'Etat de localisation. Notons que les informations client disponibles sont limitées à celles concernant les commandes et qu'il serait utile de rechercher d'autres données, en particulier à caractère socio-démographique (voire comportement sur le site d'achat), pour compléter la segmentation.
- **Les features RFM constituent donc une bonne base qui pourra être enrichie dans le cadre d'un échange avec l'équipe marketing du client**

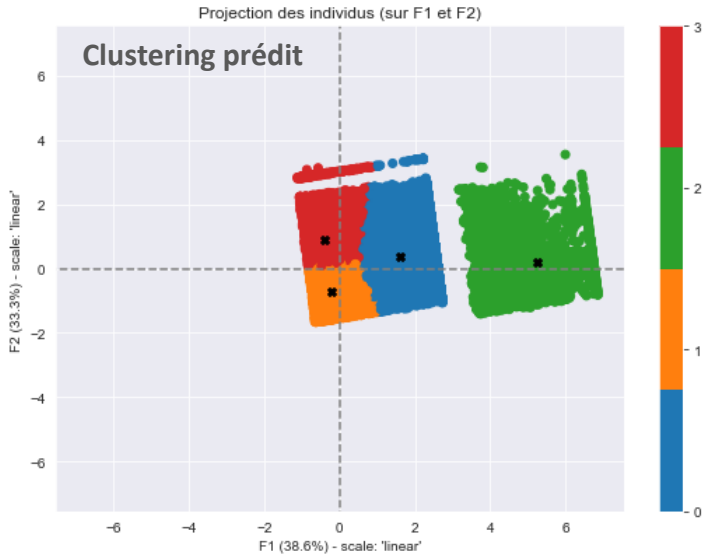
Estimation de la pérennité dans le temps du modèle de clustering

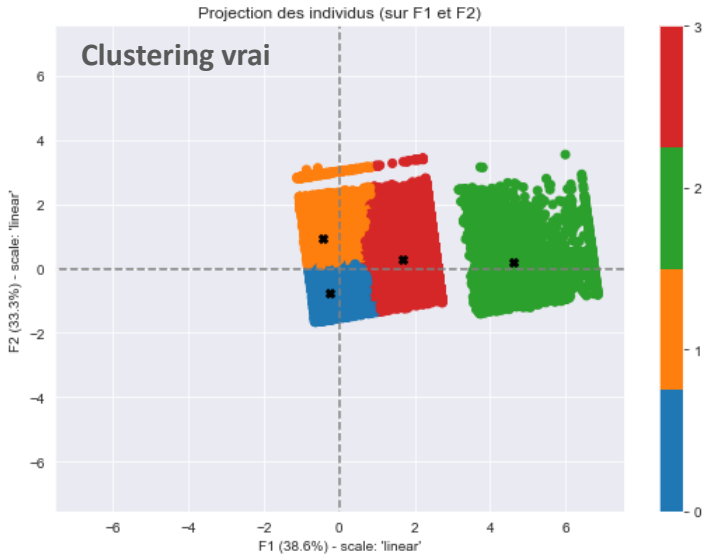
1. Estimation basée sur le modèle k-Means avec k=4 et features RFM
2. Création d'une fonction de sélection des données à une date donnée pour constituer le jeu de données clients
3. Choix de la date du modèle de base
   - Première date permettant de constituer un modèle à 4 clusters de forme similaire au modèle déterminé précédemment et coefficient de silhouette et indices CH et DB de même grandeur
   - Etabli sur la base des 280 premiers jours du jeu de données avec 12,8% du jeu complet
4. Calcul de l'index de Rand ajusté (ARI) de la date du modèle de base à celle du jeu complet
   - Label de cluster prédit par le modèle de base
   - Label de cluster vrai calculé par le clustering du jeu à la date de calcul de l'index
5. Représentation graphique du clustering prédit et clustering vrai
6. Examen de l'écart de positionnement entre les centroïdes prédits et vrais
7. Conclusion: 400 jours minimum de pérennité



Ecart de positionnement des centroïdes

	recency	frequency	monetary
0	-99.48	0.00	-2.28
1	-280.24	0.00	-3.30
2	-166.42	-0.00	-22.50
3	-168.39	0.00	-16.72





OPENCLASSROOMS
CentraleSupélec

Conclusion

- Les données du site de e-commerce Olist permettent d'effectuer une segmentation des clients en fonction seulement de leurs commandes

- Les données qui apparaissent les plus exploitables pour cette segmentation sont:
  - En priorité les données RFM
  - Puis les données de satisfaction client (review score), de mode de paiement (un ou plusieurs moyens, paiement ou pas en plusieurs fois), et d'Etat de localisation.

- Le modèle le plus intéressant pour la segmentation est k-Means, qui permet de proposer une segmentation en 4 groupes avec les features RFM, avec la possibilité de l'enrichir avec les autres features → point de discussion avec le marketing

- Le modèle proposé a une pérennité d'au moins un an, sachant que l'évolution porte essentiellement sur la récence (compte tenu du faible taux de commandes multiples)

Echanges avec l’évaluateur

- **Discussion**
- **Débriefing**

OPENCLASSROOMS CentraleSupélec

Contact: Eric TREGOAT

eric.tregoat@gmail.com

06 49 99 79 59

https://www.linkedin.com/in/erictregoat/