Projet 5: Segmentez des clients d'un site e-commerce

Maodo FALL

OpenClassrooms

Soutenance du projet

22 octobre 2024

Sommaire

Problématique et jeu de données

Problématique :

- Fournir aux équipes d'e-commerce d'Olist une segmentation des clients qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

But :

- Comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.

Olist

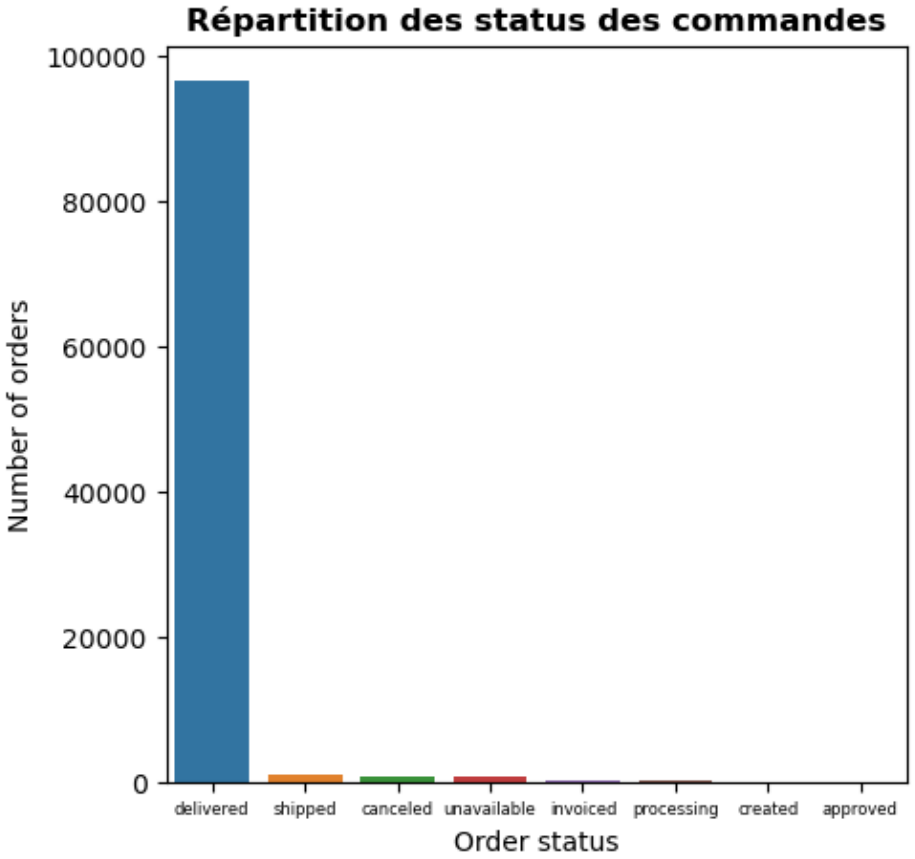
Nettoyage des données

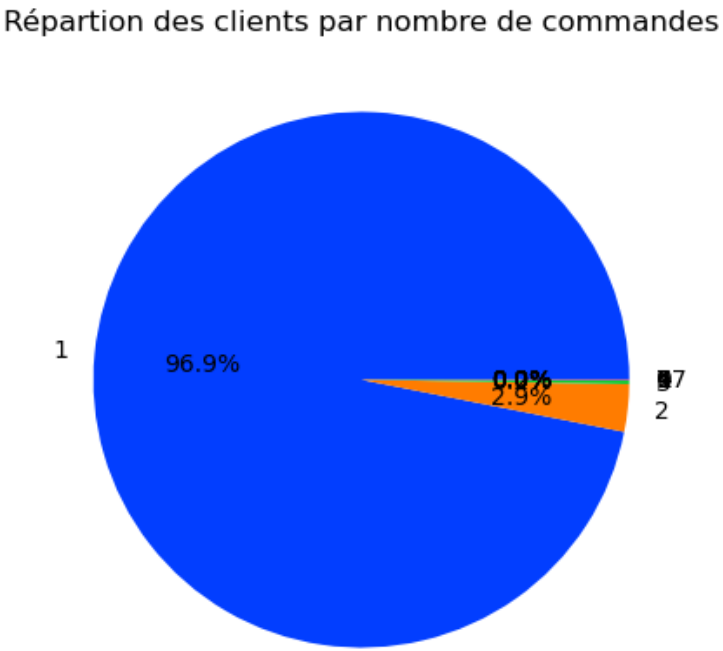
La base de données exploitée contient 09 jeux de données et sa documentation est disponible sur kaggle.

Nous avons :

- exploré les 09 jeux de données.
- représenté graphiquement certaines caractéristiques.
- joint les jeux de données pour obtenir un jeu de données des clients.
- sélectionné les variables pertinentes pour l'étude.
- conservé uniquement les commandes livrées.

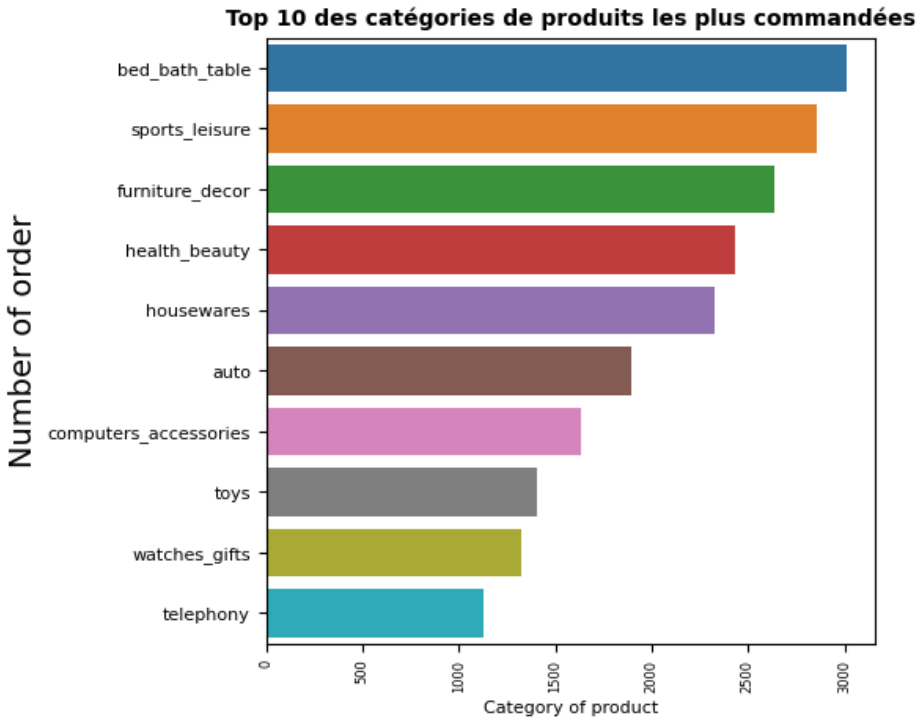
Status des commandes - Nombre de commandes par client

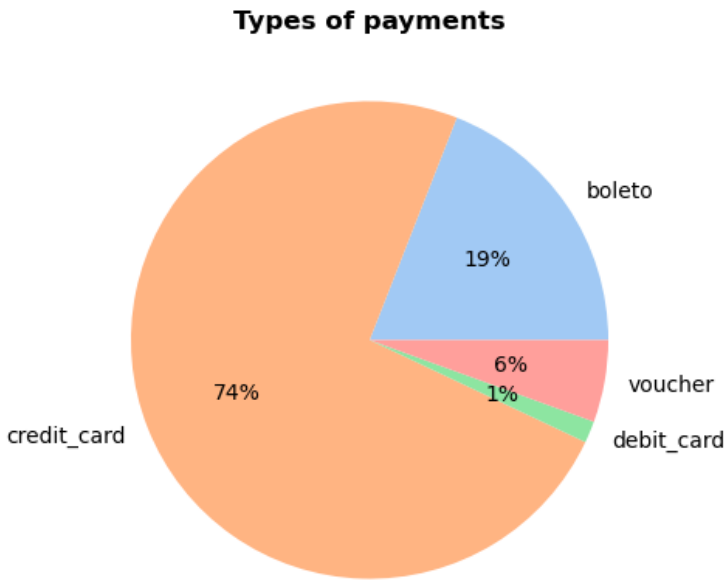




- Quasiment toutes les commandes sont livrées.
- 97% des clients n'ont commandé q'une seule fois.

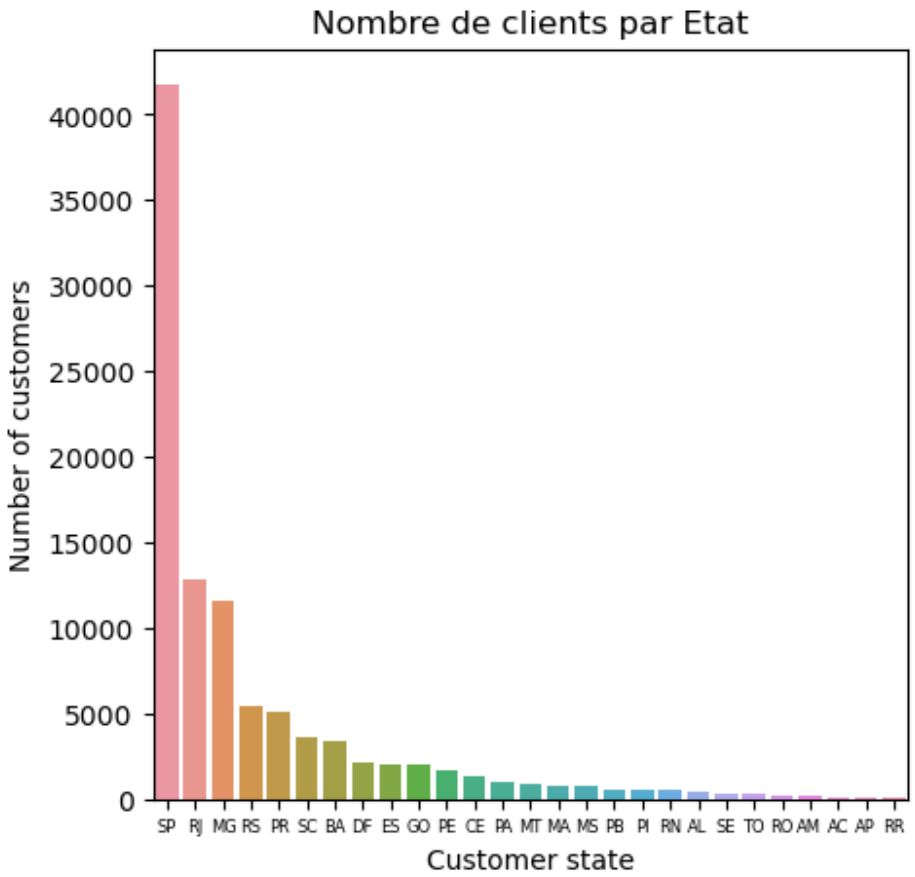
Produits commandés - Types de payements





- La majorité des payements se font par carte bancaire.

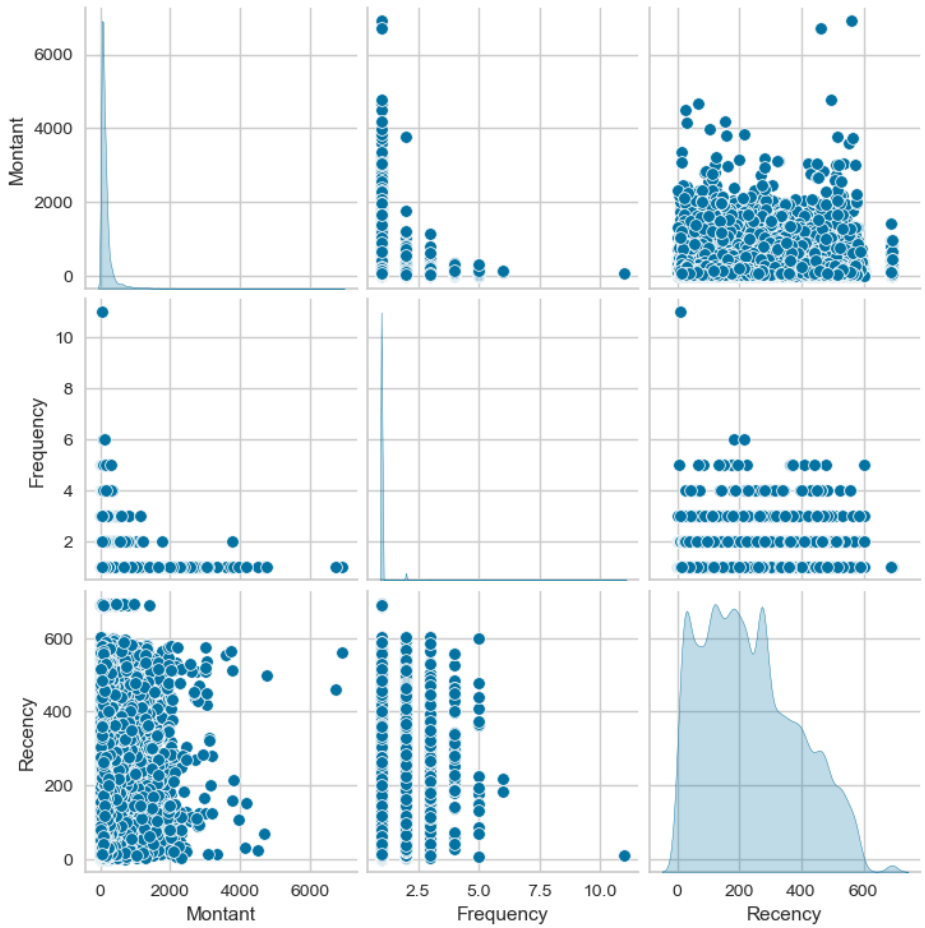
Nombre de clients par commandes



- La majorité des clients vivent à São Paulo.

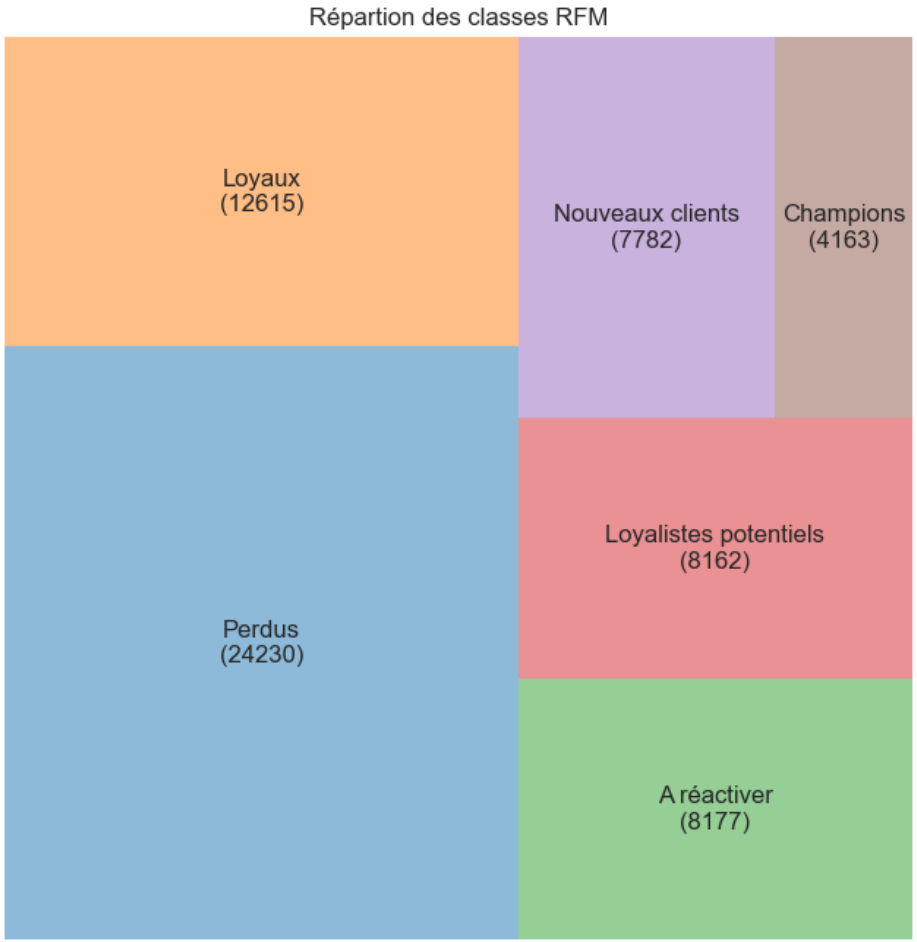
1. Analyse exploratoire des données
2. Segmentation RFM et essais de modèles de clustering
   - Segmentation RFM
   - Clustering : Essai de modèles
3. Contrat de maintenance
4. Conclusion

Représentation des données RFM



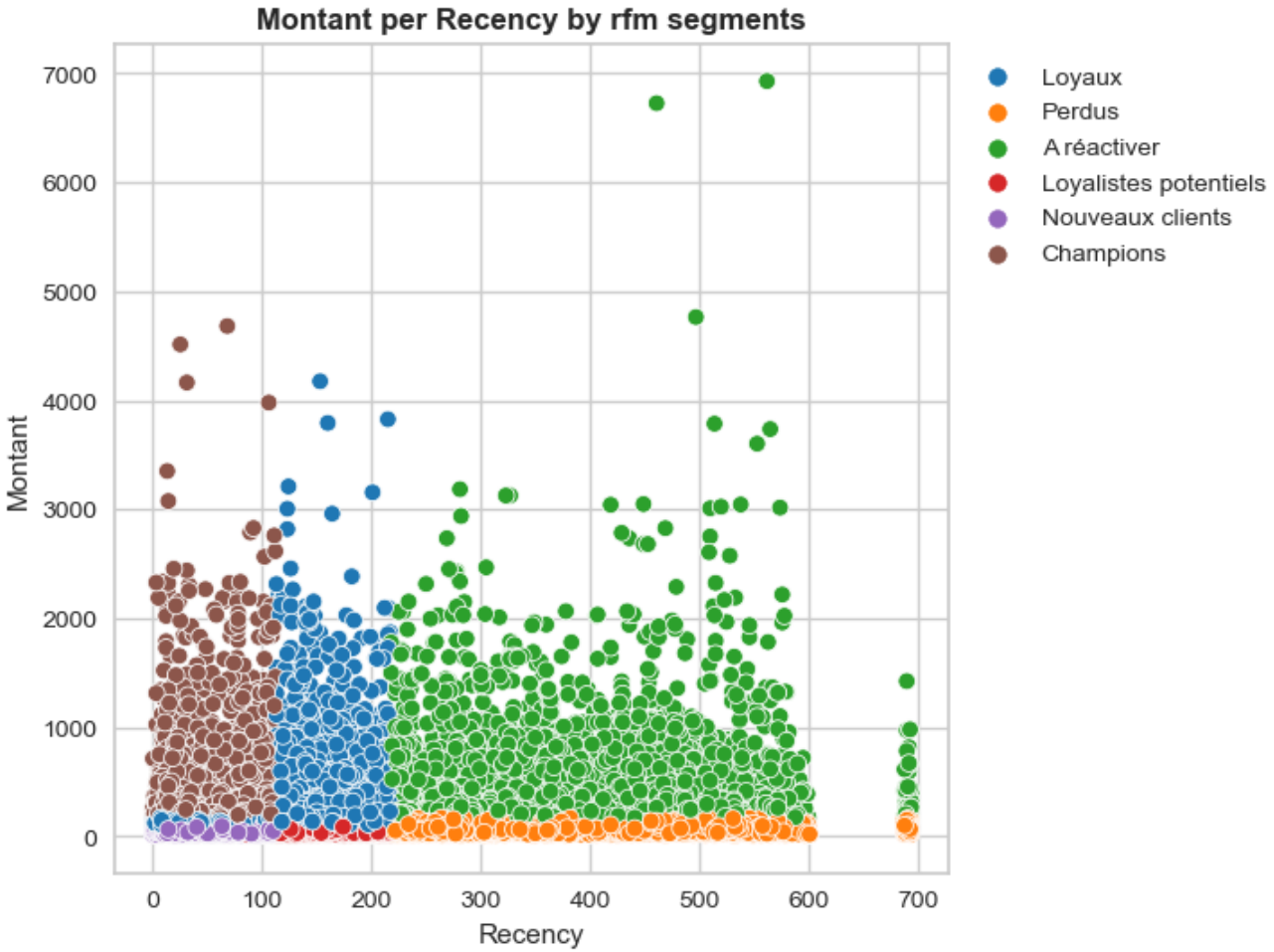
- Il n'y a pas de lien linéaire entre les variables RFM.
- La caractéristique fréquence ne varie quasiment pas.

Segments RFM - TreeMap des segments



- Les clients champions constituent le plus petit segment.
- Les clients perdus constituent le plus grand segment.

Segments RFM - Récence/Montant



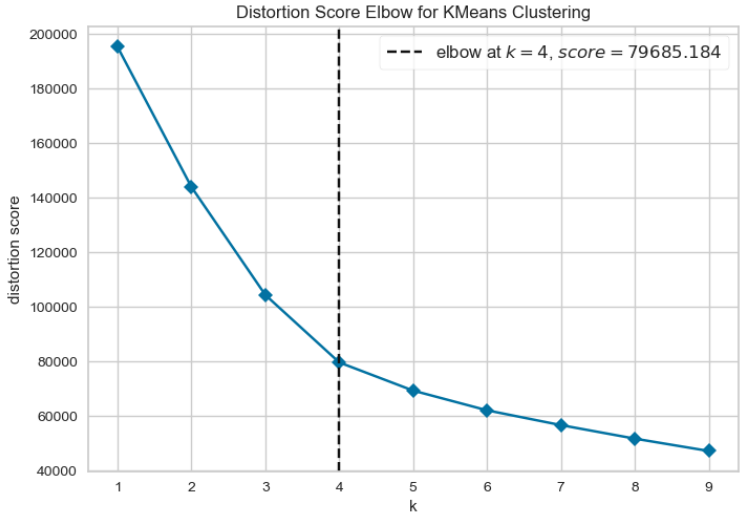
- On voit mieux la séparation des segments RFM via ce nuage de points entre la récence et le montant.

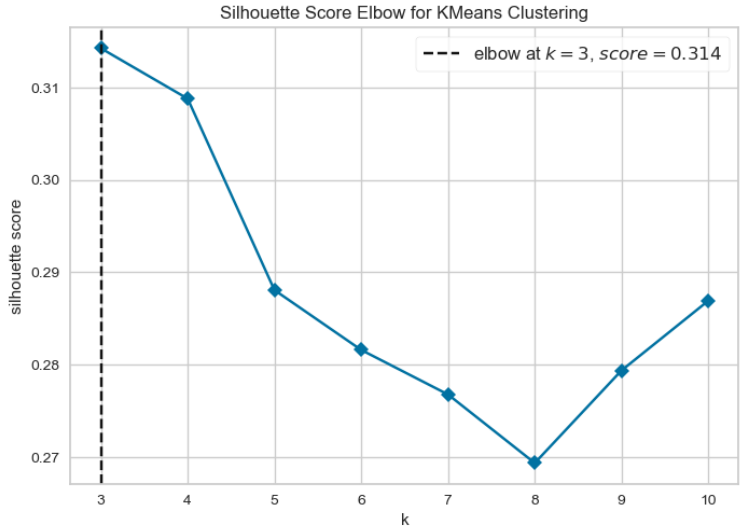
La segmentation avec les algorithmes de clustering

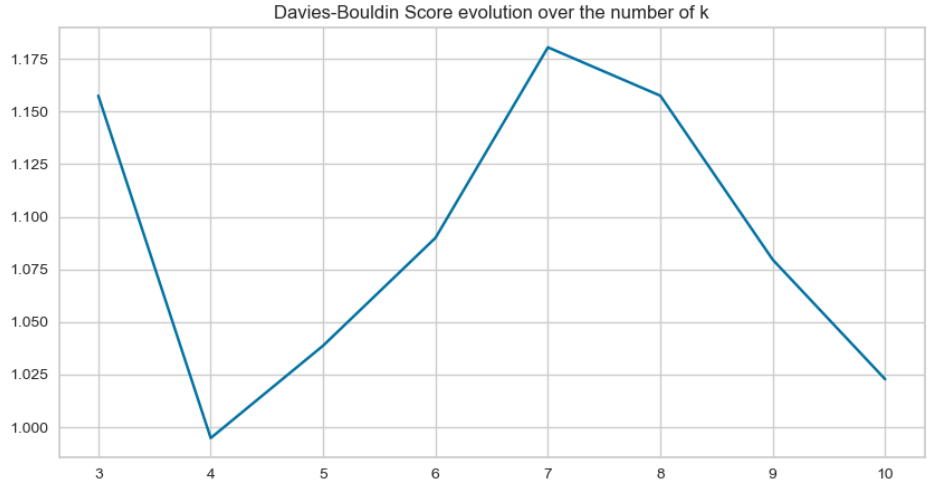
Nous avons :

- mis à l'échelle les données.
- essayé deux modèles KMeans (avec $k = 3$ et $k = 4$).
- essayé deux modèles DBSCAN (avec le paramètre epsilon $eps = 0.2$ et $eps = 0.4$).

Essais KMeans - Choix de k





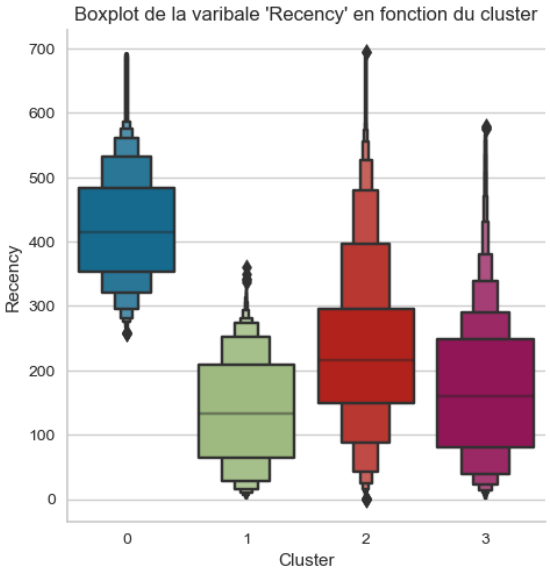


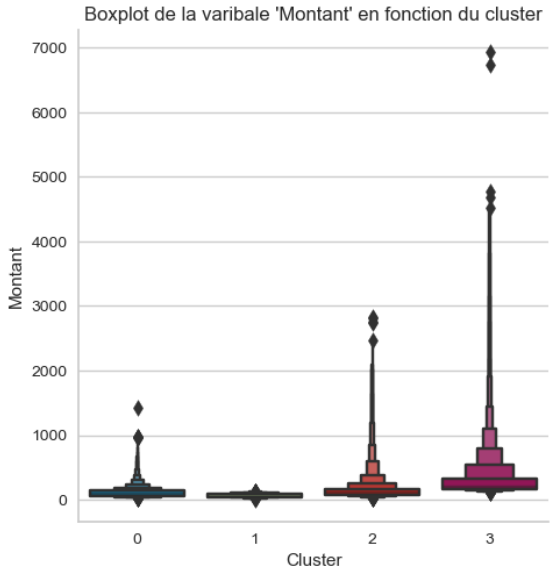
Essai KMeans avec $k = 3$

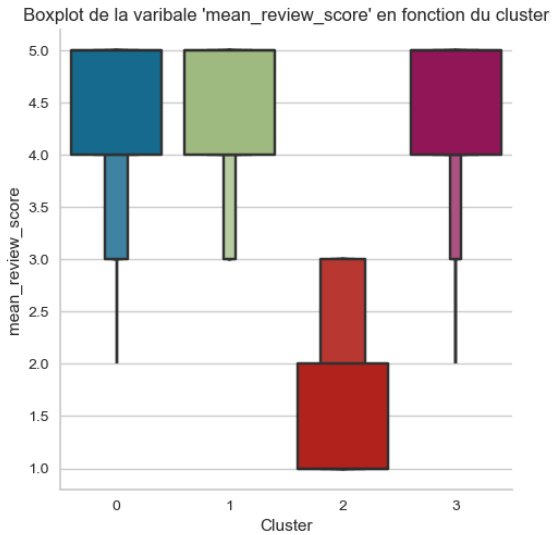
Cluster	Recency	Montant	mean_review_score
0	389.921984	153.387419	4.645626
1	227.354893	166.308591	1.823107
2	120.504363	153.438596	4.721309

- Ce clustering semble assez contradictoire avec des classes qui dépensent presque de la même façon et dont les notations des commandent se contredisent.

KMeans avec $k = 4$ - Interprétation des clusters







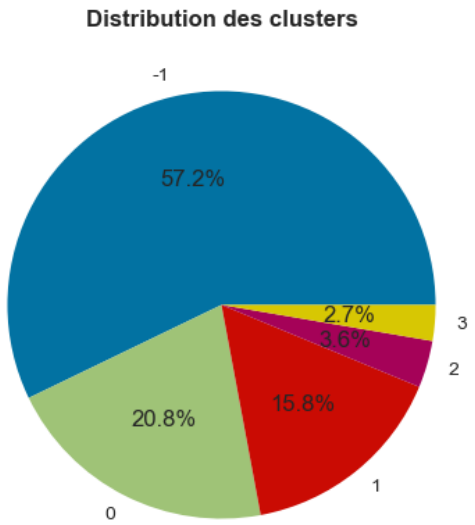
Cluster	Recency	Montant	mean_review_score
0	421.559379	111.674373	4.607041
1	138.247886	66.984588	4.656266
2	233.717736	157.891441	1.633977
3	169.664418	316.878178	4.660296

KMeans avec $k = 4$ - Interprétation des clusters

- Cluster 0 : Clients à réactiver.
- Cluster 1 : Clients loyalistes potentiels et des nouveaux clients.
- Cluster 2 : Clients perdus.
- Cluster 3 : Clients loyaux et les champions.

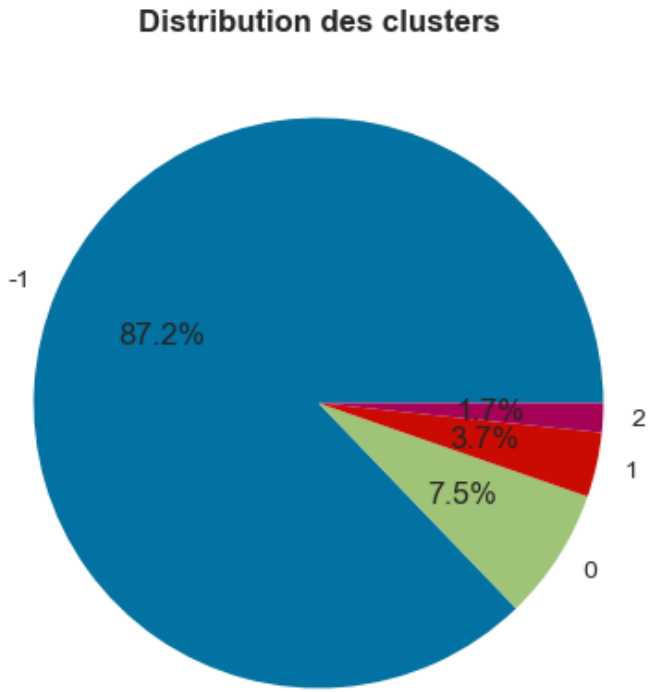
Essai DBSCAN avec $eps = 0.2$

Cluster	Recency	Montant	mean_review_score
-1	287.175427	293.047454	2.921113
0	227.349538	123.702740	4.999996
1	210.842760	108.427379	4.000000
2	202.720017	115.980305	1.000000
3	196.454030	100.271761	2.999905



- Les clusters ont une récence assez similaire, dépensent relativement de la même façon mais notent différemment.
- Du point de vu métier, ce clustering n'a pas trop de sens.
- De plus, les clusters 2 et 3 sont trop petits, ce qui confirme que cette segmentation ne semble pas pertinente.

Essai DBSCAN avec $eps = 0.4$



- Les clusters 0, 1 et 2 sont trop petits et presque toutes les données sont localisées au cluster -1.
- Cette segmentation DBSCAN ne semble pas non plus pertinente.

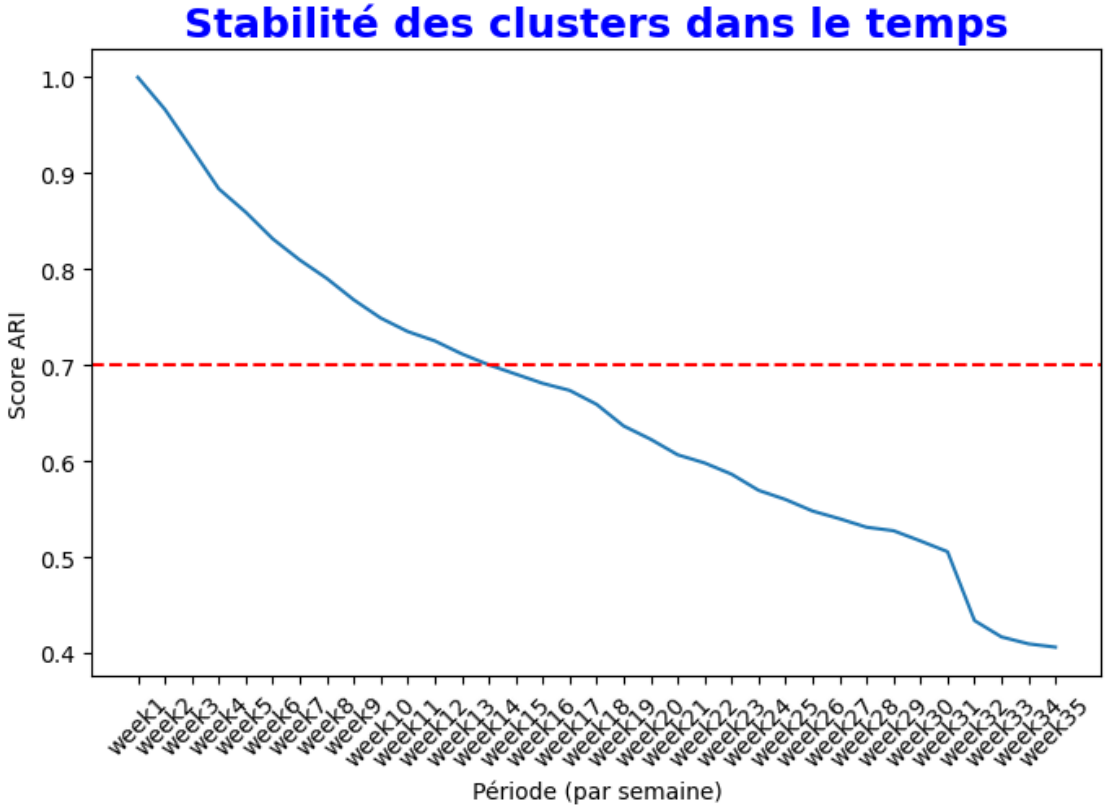
1. Analyse exploratoire des données
2. Segmentation RFM et essais de modèles de clustering
   - Segmentation RFM
   - Clustering : Essai de modèles
3. **Contrat de maintenance**
4. Conclusion

Stabilité des clusters dans le temps

- Notre meilleur modèle est le KMeans avec k=4.

- L'étude est effectué sur une période de presque 2 ans(Janvier 2017 à fin Août 2018).
  - Nous avons entraîné le modèle initial sur les données de 2017.
  - Ensuite nous avons itéré la modélisation sur chaque nouvelle semaine et calculé le score ARI pour comparer les clusters.

Stabilité des clusters dans le temps



- Le score ARI décroît fortement et atteint une valeur inférieur à 0.7 au bout de 13 semaines de mise en production.
- Il faudrait réaliser une maintenance du modèle après 13 semaines de son déploiement pour mettre à jour les clusters afin de les garder stables.

1. Analyse exploratoire des données
2. Segmentation RFM et essais de modèles de clustering
   - Segmentation RFM
   - Clustering : Essai de modèles
3. Contrat de maintenance
4. **Conclusion**

Conclusion

- Avec la segmentation RFM, nous avons réussi à sortir des segments assez cohérents.
- Le clustering non supervisé nous donne cependant une approche plus rigoureuse et des clusters mieux contrôlables dans le temps.
- En effet, avec le contrat de maintenance de notre modèle, nous parvenons désormais à détecter au bout de combien de temps il est utile de de revoir le comportement de nos clients. Ce qui nous permettra de gagner en efficacité dans nos besoins de campagnes de communication.

MERCI BEAUCOUP !