olist

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

Laurent Cagniart

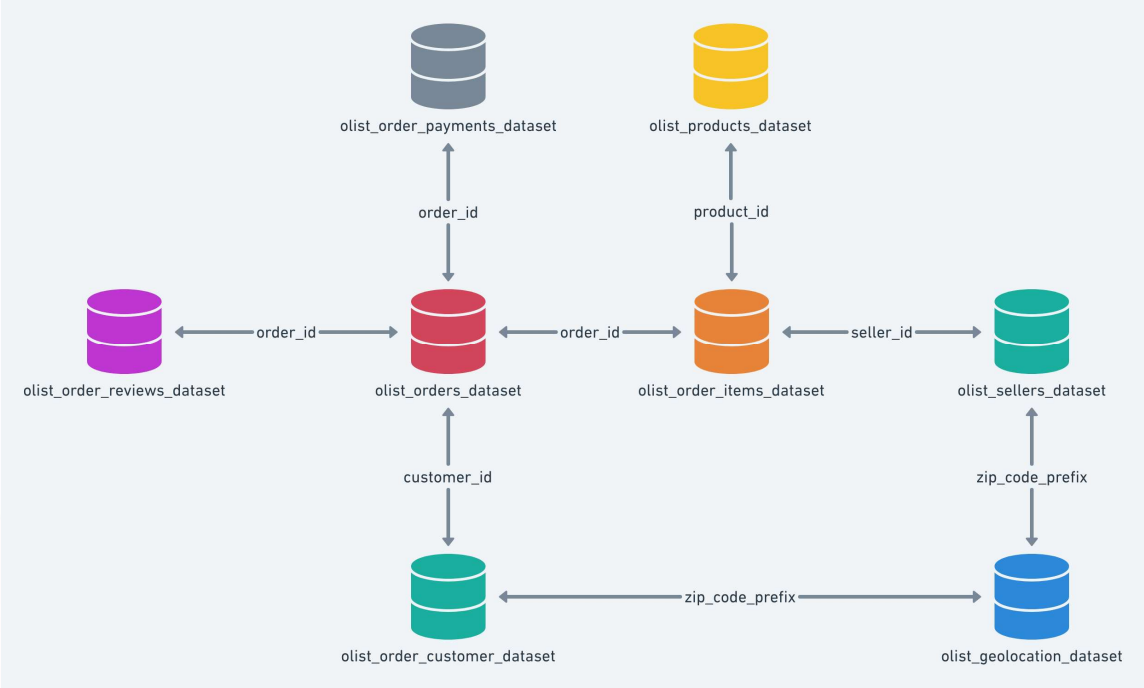
Problématique

Olist, entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne

- Objectifs :

1) Vous aider à optimiser vos campagnes de communication grâce à une **meilleure connaissance de vos clients** et de leur **segmentation**
2) Vous fournir une **proposition de contrat de maintenance** basée sur une analyse de la stabilité des segments au cours du temps.

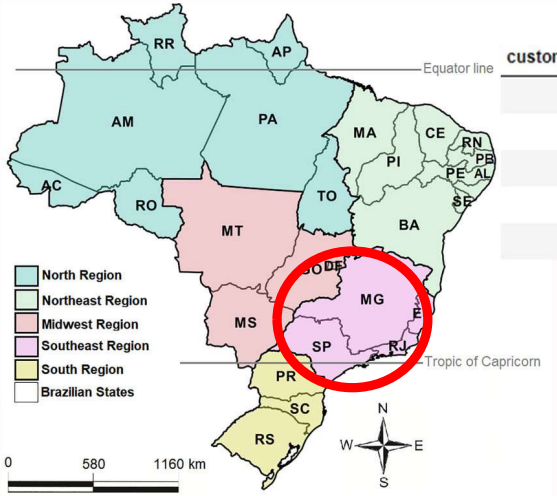
Jeu de données : cleaning



- Création d'un dataset global à partir des 8 fichiers mis à disposition

Analyse exploratoire (1/5)

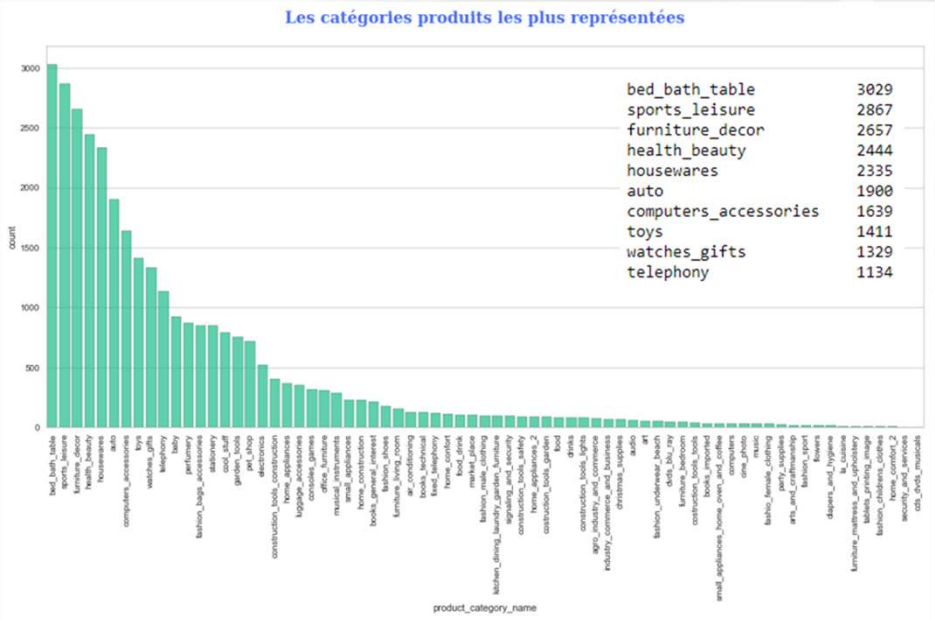
- Des clients concentrés sur 3 états (66%)



customer_state	customer_id
SP	41746
RJ	12852
MG	11635
RS	5466
PR	5045

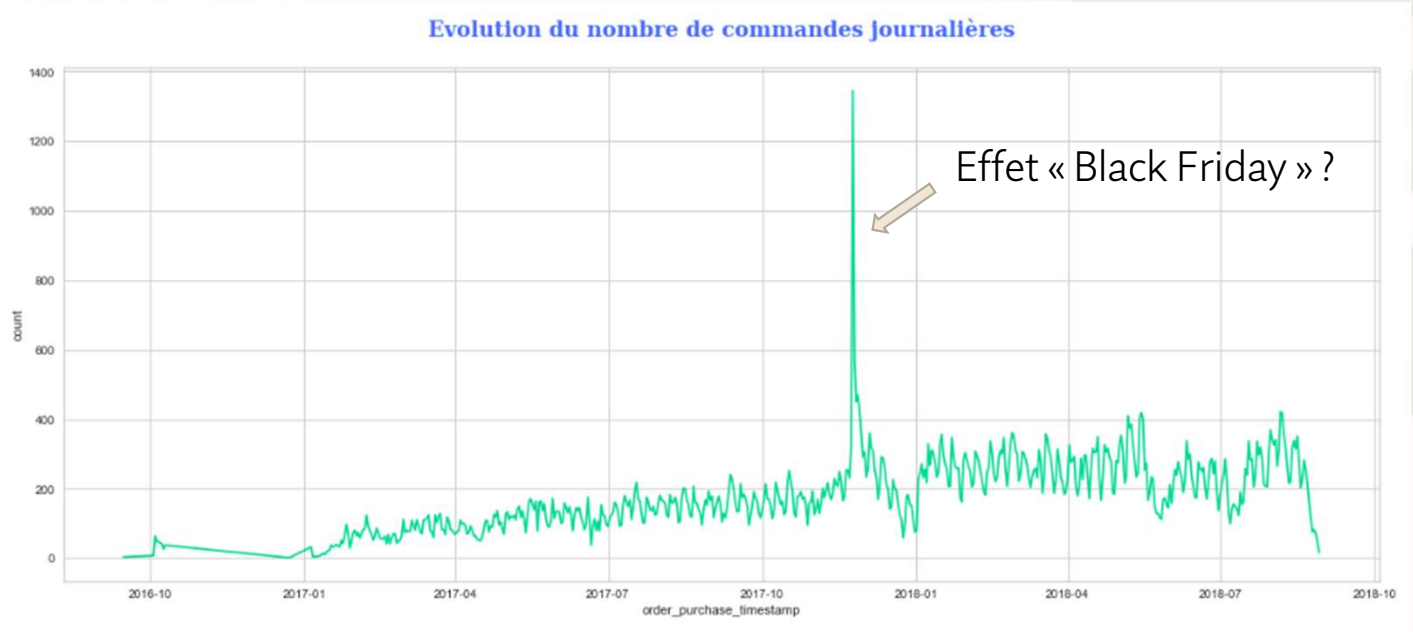
- Les catégories de produits les plus représentées :

  Mobilier/maison, Sport, Beauté et Tech



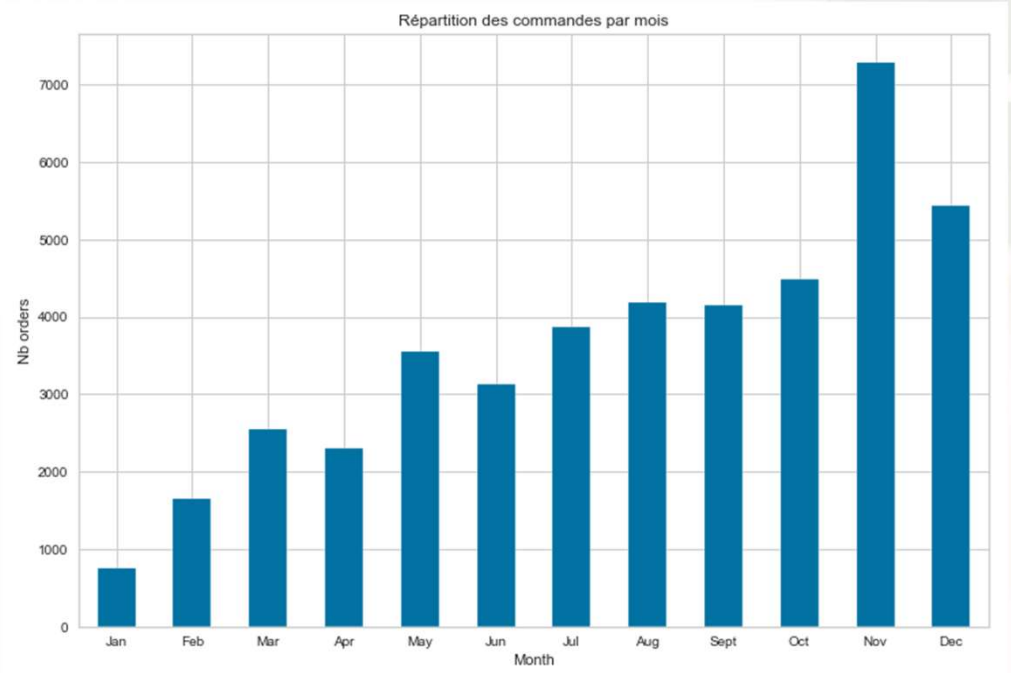
Analyse exploratoire (2/5)

- Des commandes quotidiennes en **croissance moyenne régulière** avec un **pic en novembre 2017**



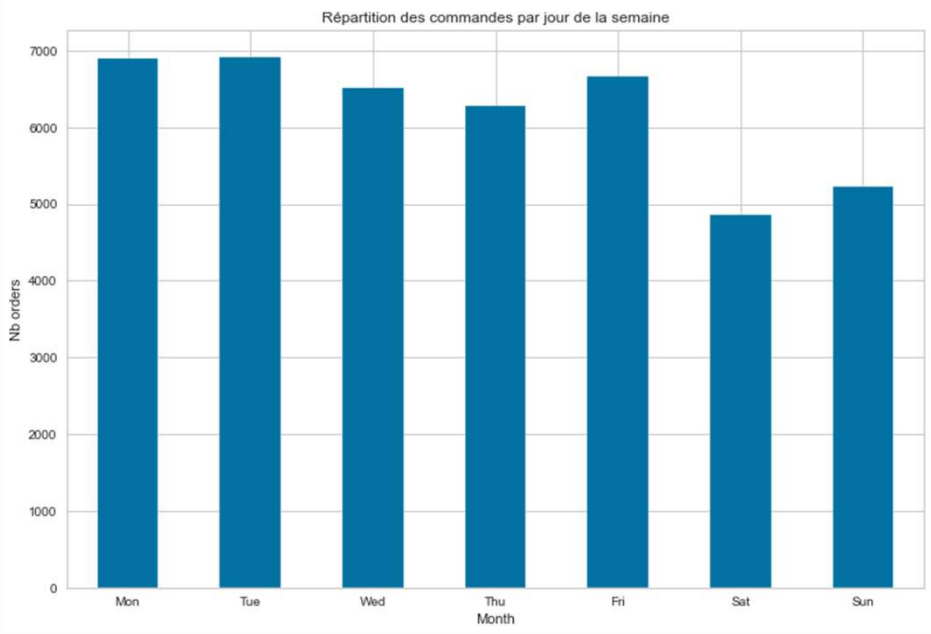
Analyse exploratoire (3/5)

- **Pas de saisonnalité marquée** en 2017

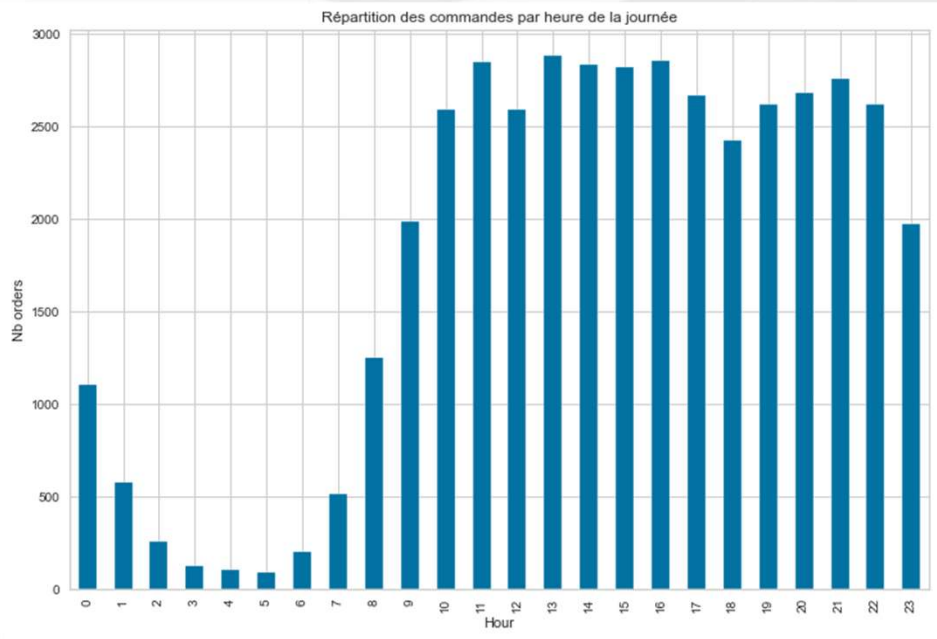


Analyse exploratoire (4/5)

- Les **jours ouvrés** de la semaine (surtout, **lundi et mardi**) plus propices aux ventes

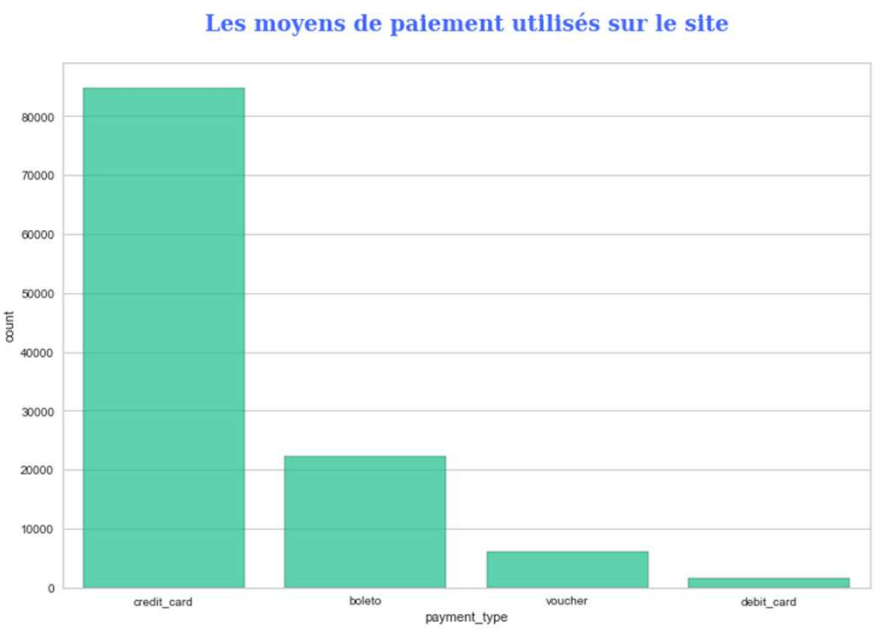


- Des ventes réparties essentiellement sur les horaires **10h-22h** (début d'après-midi) avec des reflux à 12h et 18h

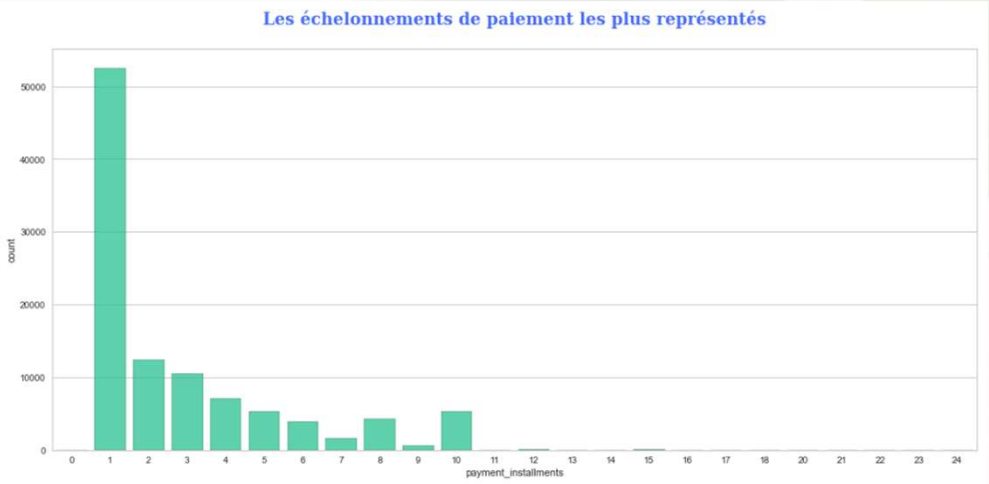


Analyse exploratoire (5/5)

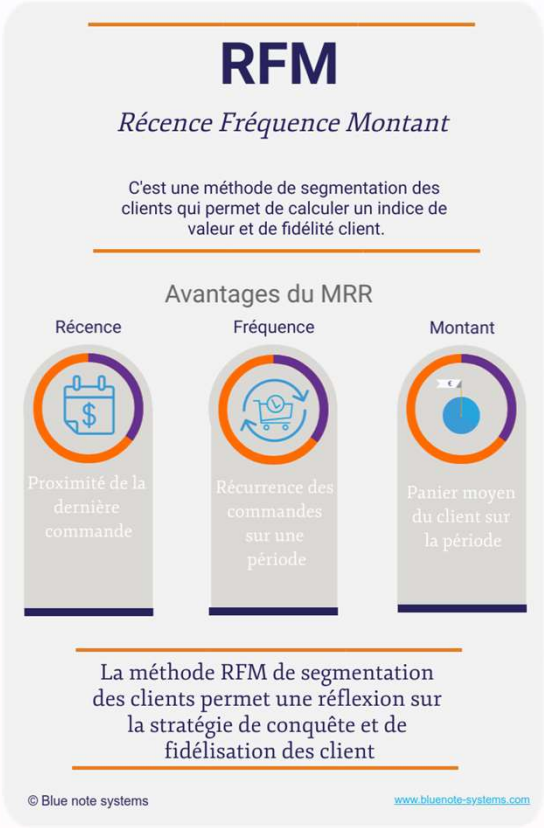
- Environ **75% des paiements** sont réalisés par **credit_card**; suivent ensuite 2 moyens de paiement électroniques "prépayés" ('boleto' et 'voucher')



- **50% des paiements se font en une fois** et 75% en 4 fois ou moins



Feature engineering (1/2)



Nous allons nous appuyer dans un 1er temps sur une segmentation « classique » pour les acteurs de la vente à distance, la **segmentation RFM**

À partir des données de notre dataset global, nous allons devoir créer les 3 features suivantes :

- **Récence**
- **Fréquence**
- **Montant**

Feature engineering (2/2)

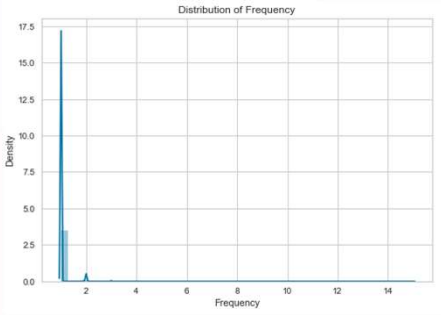
Afin de **compléter la segmentation RFM**, nous allons créer 3 nouvelles features liées aux délais de livraison et avis laissés par les clients :

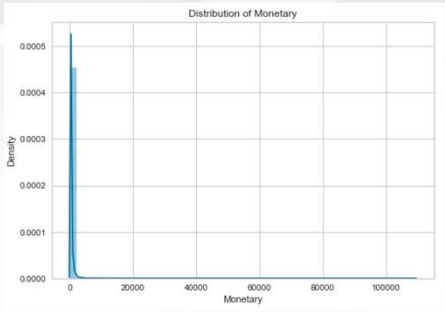
- **Délai de livraison** : différence entre date d'achat et date de livraison chez le client
- **Retard de livraison** : différence entre date prévue de livraison et date de livraison effective chez le client
- **Review score moyen :** moyenne des avis « notés » par les clients

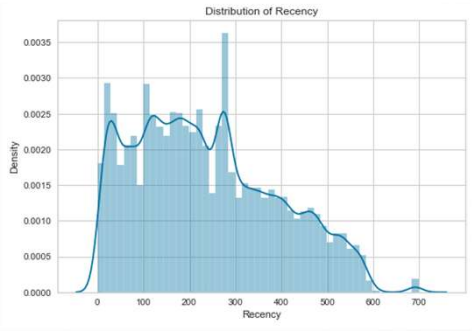
Modélisation(1/8)

Des achats en moyenne peu fréquents pour un montant d'environ 40€

	Recency	Frequency	Monetary
count	93358.000000	93358.000000	93358.000000
mean	237.478877	1.033420	212.964557
std	152.595054	0.209097	646.223866
min	0.000000	1.000000	0.000000
25%	114.000000	1.000000	63.830000
50%	218.000000	1.000000	113.140000
75%	346.000000	1.000000	202.637500
max	713.000000	15.000000	109312.640000



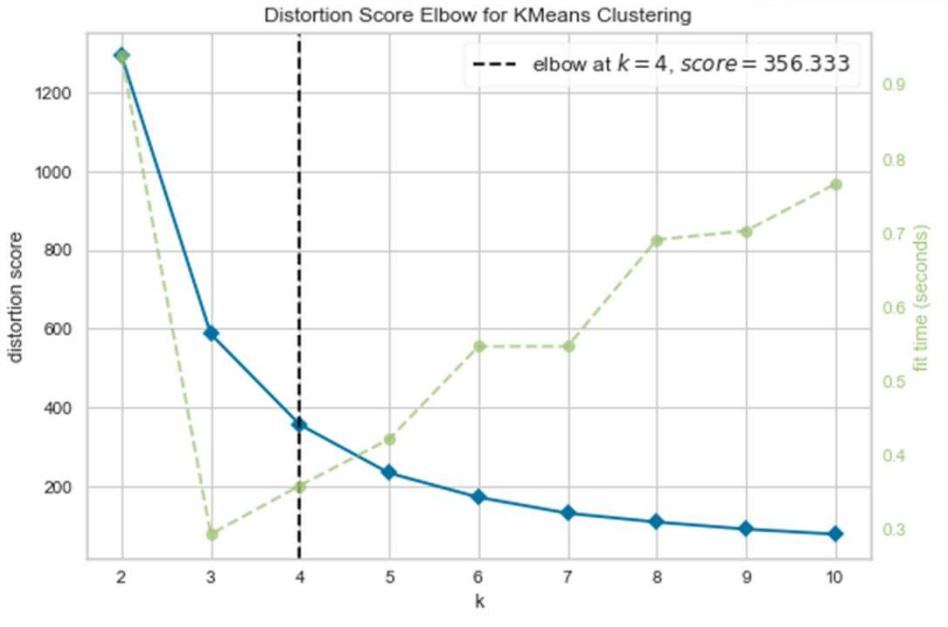




Modélisation(2/8)

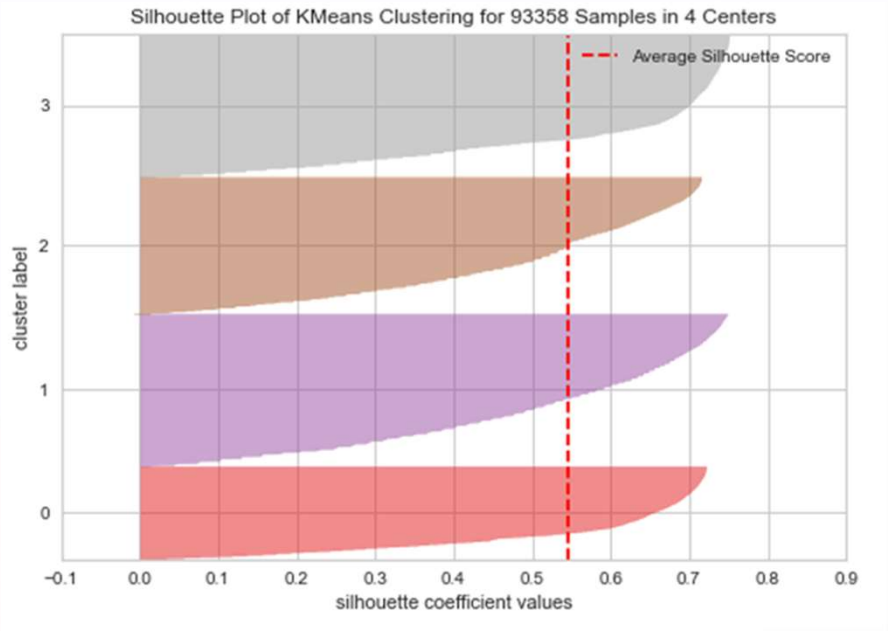
Segmentation RFM, modélisation par la **méthode K-means**

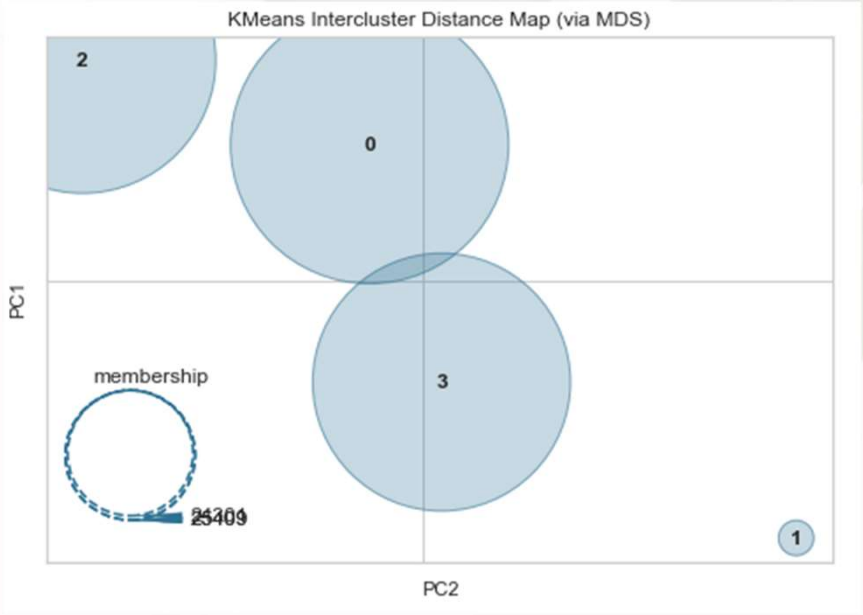
K = 4, valeur optimale de K



Modélisation(3/8)

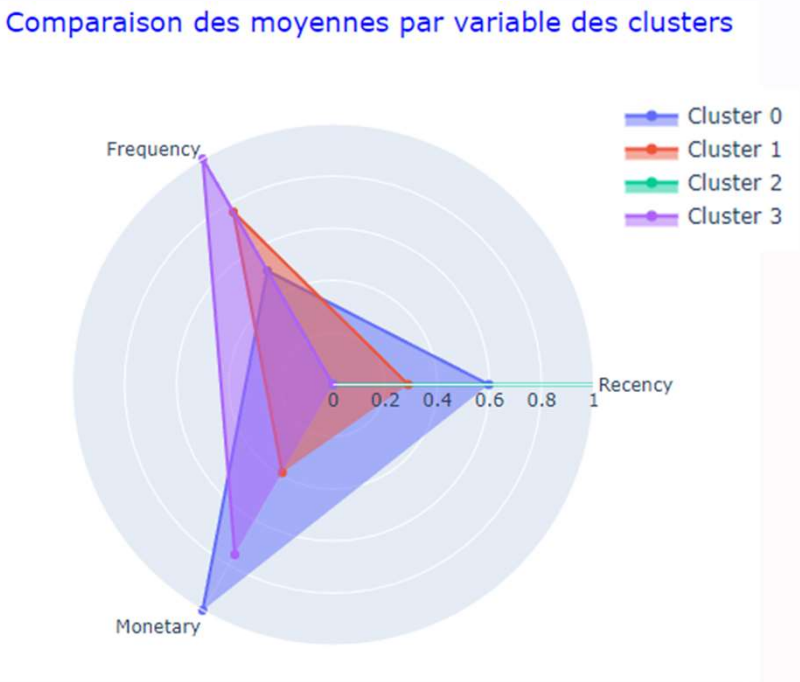
Clustering satisfaisant (score de silhouette)





Modélisation(4/8)

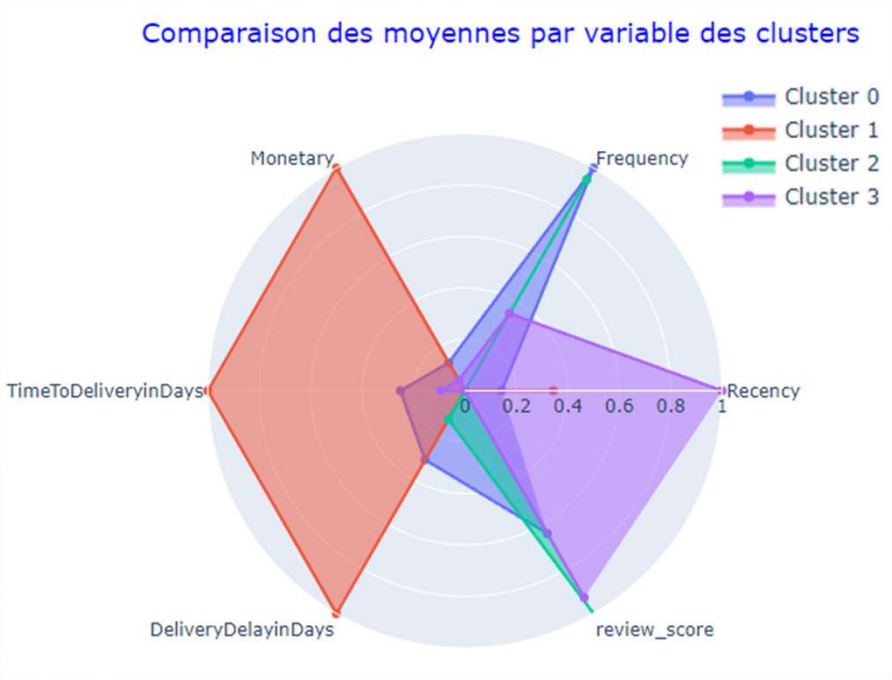
Segmentation RFM, modélisation par la **méthode K-means**



- Cluster 0 : Des clients à forte valeur mais avec des fréquence et récence moyennes
- Cluster 1 : Des clients dits « loyaux » avec des commandes de montant et de récence moyens mais plus fréquents que la moyenne
- Cluster 2 : Des clients dits oubliés qui n'ont pas commandé depuis longtemps (et avec faible valeur)
- Cluster 3 : Des clients récents sur lesquels capitaliser (fréquence élevée et montants supérieurs à la moyenne)

Modélisation(5/8)

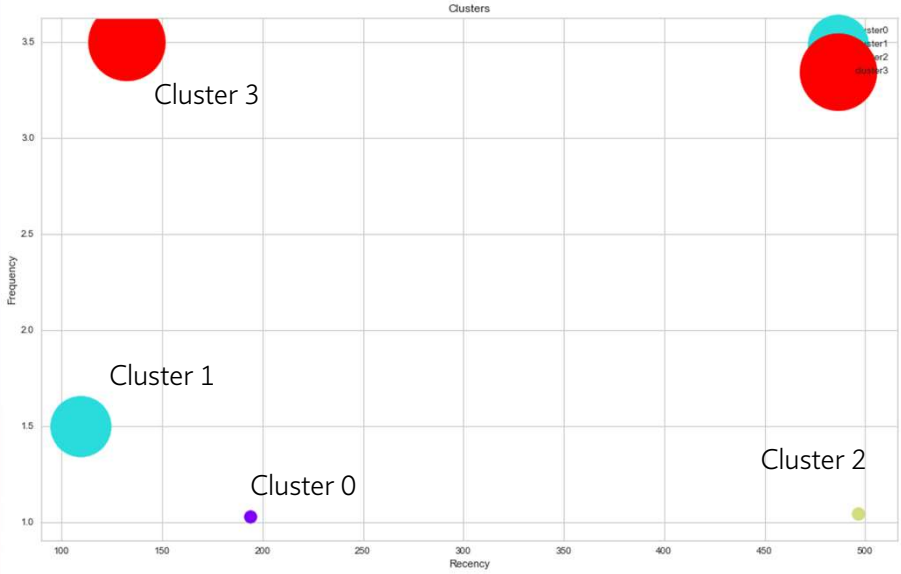
Segmentation RFM « élargie », modélisation par la **méthode K-means**



- Cluster 0 : Des clients ayant commandé plusieurs fois pour un panier moyen et plutôt récemment pour des délais légèrement plus élevés que la moyenne et des avis plutôt positifs
- Cluster 1 : Des clients à forte valeur ayant commandé une seule fois, pas récemment avec des avis négatifs et des délais de livraison élevés
- Cluster 2 : Des clients ayant commandé récemment pour une valeur faible mais des avis très positifs
- Cluster 3 : Des clients anciens avec des composantes moyennes mais des avis positifs

Modélisation(6/8)

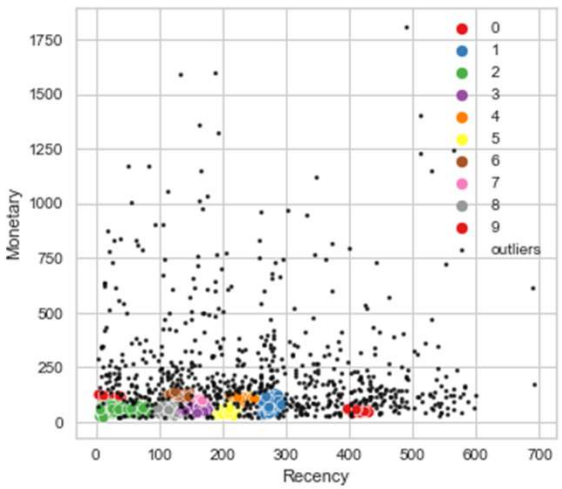
Segmentation RFM, modélisation par la **méthode Clustering hiérarchique**

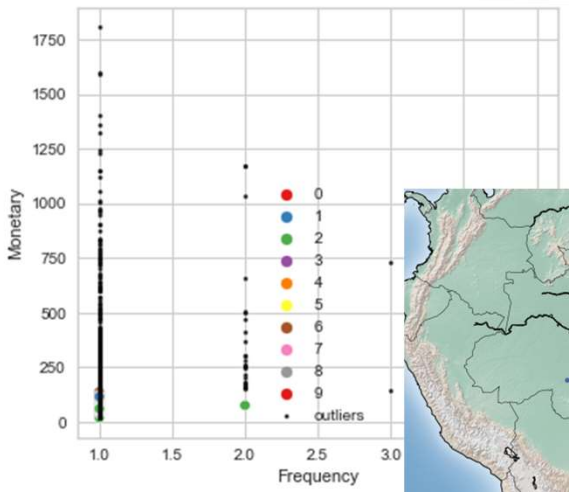


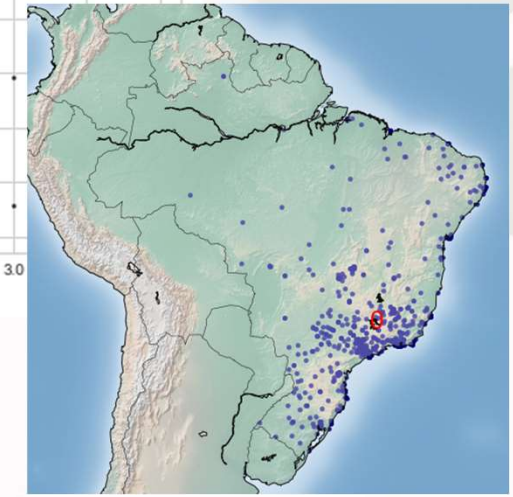
- Cluster 0 : Des clients ayant commandé une fois non récemment pour un panier de valeur peu élevée
- Cluster 1 : Des clients « loyaux» à forte valeur ayant commandé peu de fois et récemment
- Cluster 2 : Des clients « perdus » ayant commandé il y a longtemps pour une valeur faible
- Cluster 3 : Des clients « big spenders » récents ayant commandé plusieurs fois pour un panier moyen à valeur élevée

Modélisation(7/8)

Segmentation RFM, modélisation par la **méthode DBScan**







- **Résultats non concluants** (malgré testing sur hyperparamètres variés)

Modélisation(8/8)

Choix de modélisation

- Malgré des résultats similaires, nous privilégierons la **méthode k-Means** au clustering hiérarchique

Scores de stabilité à l'initialisation

Iteration	FitTime	Inertia	Homo	ARI	AMI
Iter 0	0.065s	6582	0.662	0.565	0.770
Iter 1	0.057s	2869	1.000	1.000	1.000
Iter 2	0.032s	2869	1.000	1.000	1.000
Iter 3	0.056s	2869	1.000	1.000	1.000
Iter 4	0.048s	2869	1.000	1.000	1.000
Iter 5	0.053s	2869	1.000	1.000	1.000
Iter 6	0.057s	2869	1.000	1.000	1.000
Iter 7	0.048s	2869	1.000	1.000	1.000
Iter 8	0.051s	2869	1.000	1.000	1.000
Iter 9	0.048s	2869	1.000	1.000	1.000

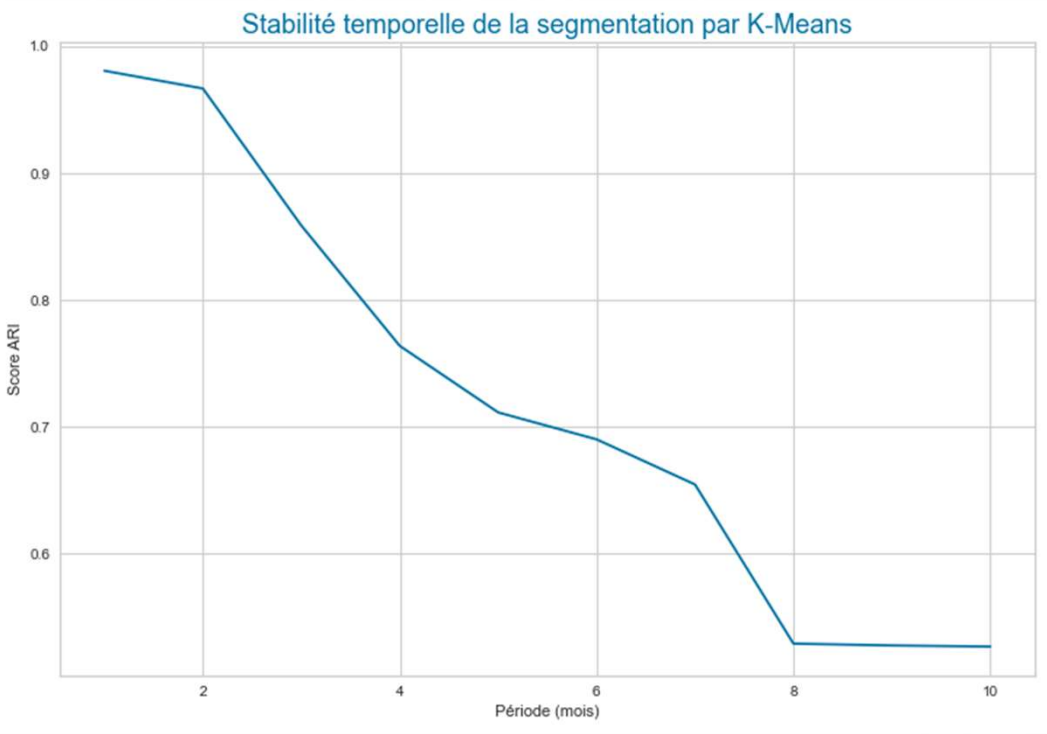
- **Simplicité d'utilisation** et visualisation plus aisée
- **Pertinence du résultat** avec des actions concrètes à déterminer et à mener selon les clusters
- **Stabilité** du modèle

Maintenance(1/2)

- Dans le but d'établir un **contrat de maintenance** de l'algorithme de segmentation client, nous devons tester sa **stabilité dans le temps** et voir, par exemple, à quel moment les clients changent de Cluster.

- Pour déterminer le moment où les clients changent de cluster, nous allons itérer le K-Means sur toute la période avec des **deltas de 1 mois** et calculer le **score ARI**, en prenant garde à bien comparer les mêmes clients (ceux des 12 mois initiaux).

Maintenance(2/2)



- Sur ce graphique des scores ARI obtenus sur les itérations par période de 1 mois, on remarque une **forte inflexion après 2-4 mois** sur les clients initiaux.

- Il faudra donc prévoir la maintenance du programme de segmentation a minima tous les trimestres dans un premier temps puis re-tester cette stabilité temporelle au fil du temps afin de l'affiner. Il sera donc nécessaire de redéfinir les segments clients à chaque maintenance.

Conclusion

- Détermination de **4 clusters** de clients grâce à une **segmentation RFM élargie** (RFM + délais de livraison et scoring avis client)

- Contrat de **maintenance trimestrielle**

Evolutions possibles :

- Avec des données historiques plus complètes, la feature fréquence pourra prendre plus d'importance et dissocier davantage les clusters
- Il sera aussi possible d'élargir les features à faire entrer dans le modèle. Ex. données géographiques (internationalisation…), modes de paiement…

Merci

Laurent Cagniart