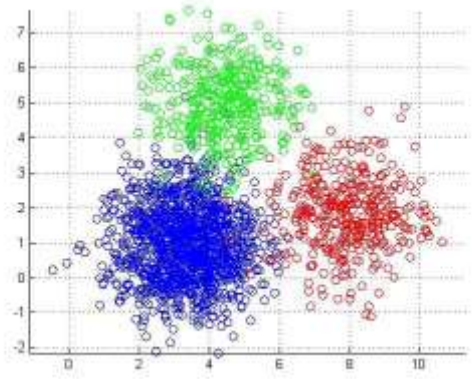
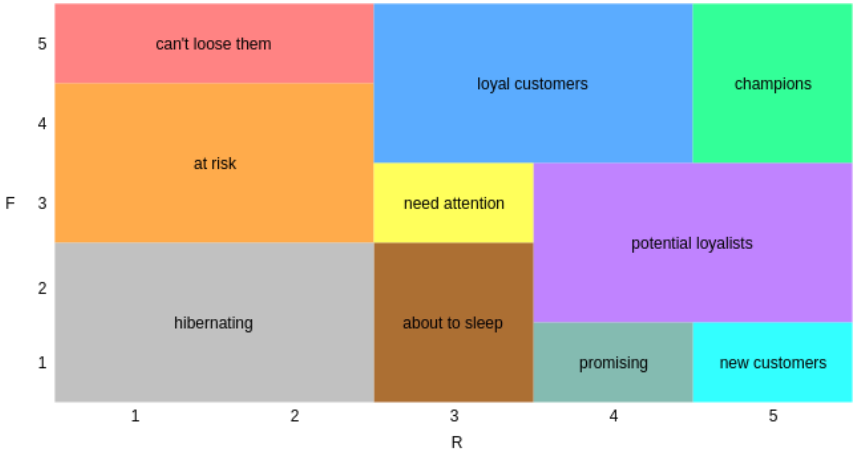
Segmentation clients d’un site e-commerce

CLUSTERING



SEGMENTATION



Sommaire

Sommaire

Démarche

Mission:

Délivrer aux équipes Marketing de l'entreprise Olist (site e-commerce) une **segmentation des clients** utilisables dans leurs campagnes de communication

Objectif:

- **Analyser** les différents types de clients (comportements sur le site, données personnelles, …)
- **Proposer** un contrat d'évaluation de la fréquence (maintenance) axée sur des segments au cours d'une durée d'observation

Compréhension du site Olist

Olist:

Plateforme vitrine du e-commerce au Brésil (2016).

Sa particularité s'exprime par: - mise en liaison entre les acheteurs et les vendeurs

- gestion des commandes, paiement, suivi de livraison
- notation et avis sur la commande

Exposition de la problématique

Segmentation clients:

Problématique autour de la segmentation traditionnelle RFM
Problématique classification non supervisée

Analyse sur les clients, avec 9 datasets:
Fusion ? , variables pertinentes ? , Feature engineering ?
Algos ? Métriques ?

Compréhension clients:

- Interprétation des catégories:
  Critères ? Métriques associées ?
Description des actions à effectuer ?

Contrat maintenance:

- Evaluation de la stabilité ?
- Maintenance sur différents mois ?

Méthodologie

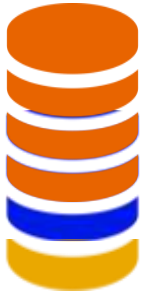
METIER

- 9 datasets séparées
- Population hétérogène

Fusion, encodage, Transfo log

ANALYSE TRANSFORMATION

- Fusion en 1 dataset

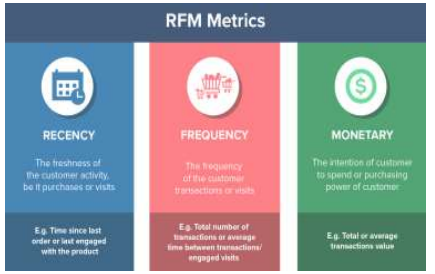


SEGMENTATION RFM

ALGO APPRENTISSAGE NON SUPERVISEE

SEGMENTATION CLIENTS

- Modèle sur les nouveaux clients
- Recherche sous-ensembles homogènes clients



INTERPRETATION

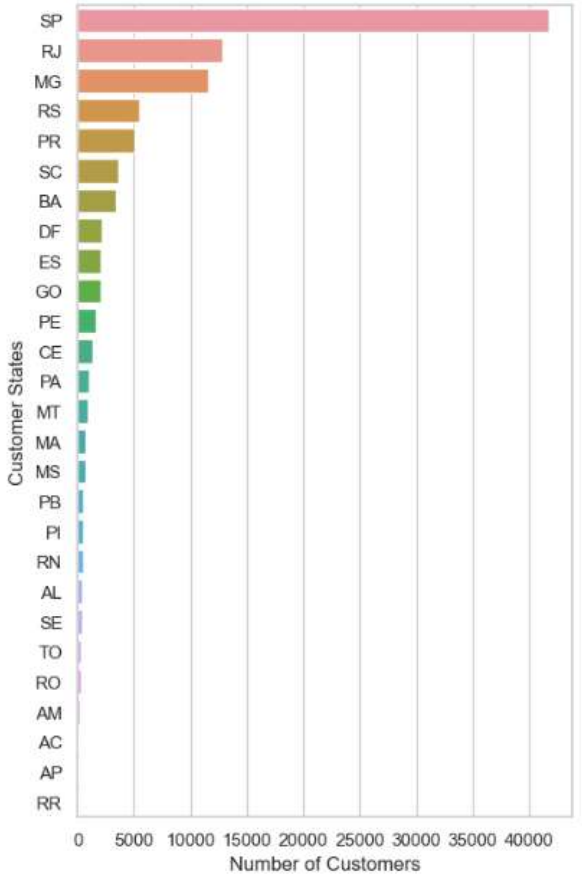
- Amélioration recommendation client
- Segmentation plus ciblée
- Analyse des profils

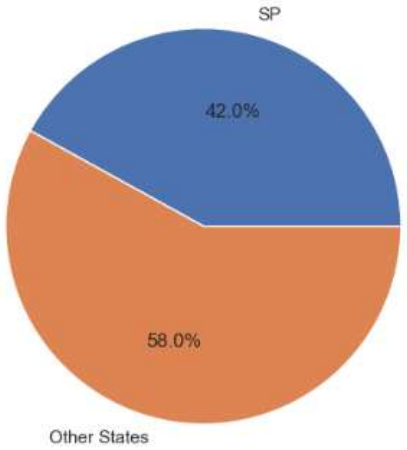
Données

1. Problématique
2. Données
3. Modélisations
4. Conclusions

Données – Analyse exploratoire

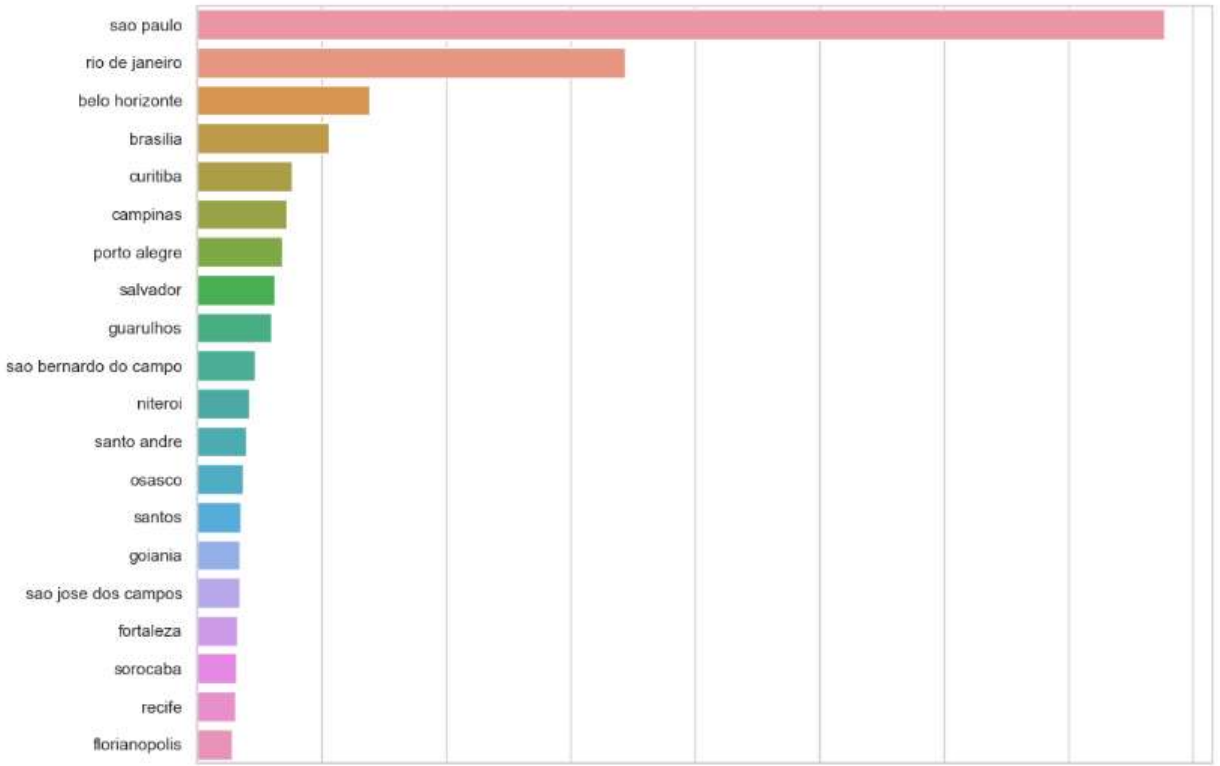
Localisation des clients



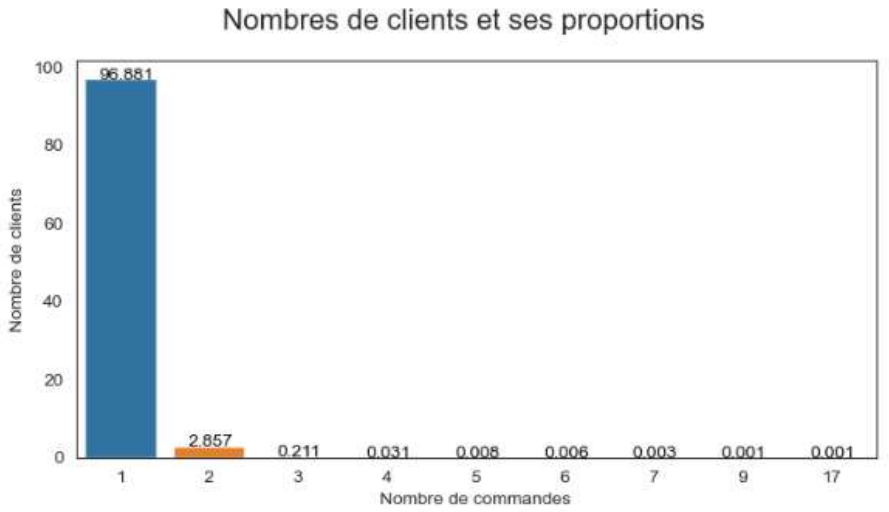


Données – Analyse exploratoire

Commandes

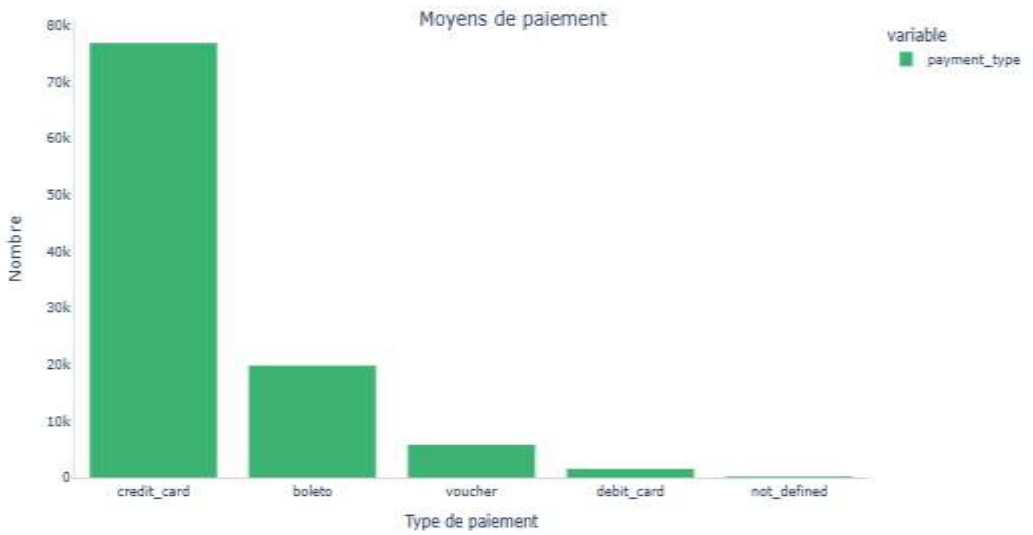


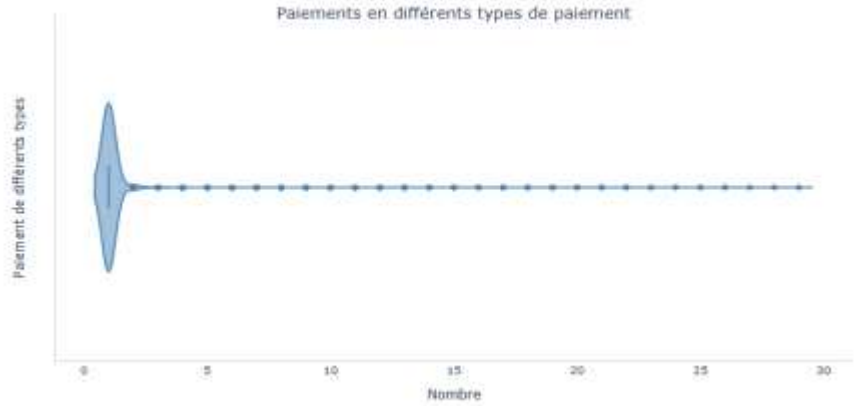
Fréquence des commandes

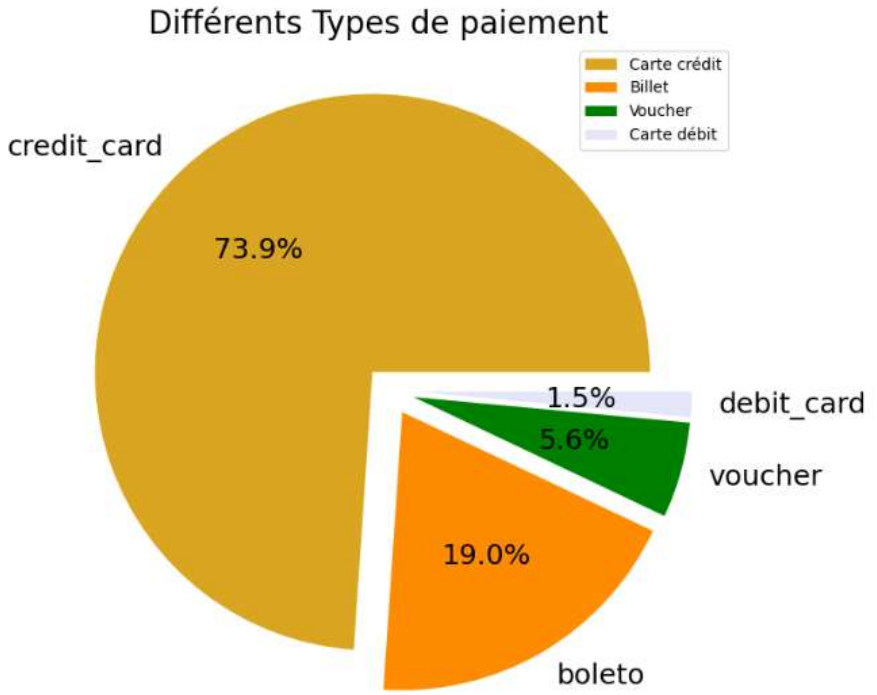


Données – Analyse exploratoire

Moyen de paiements

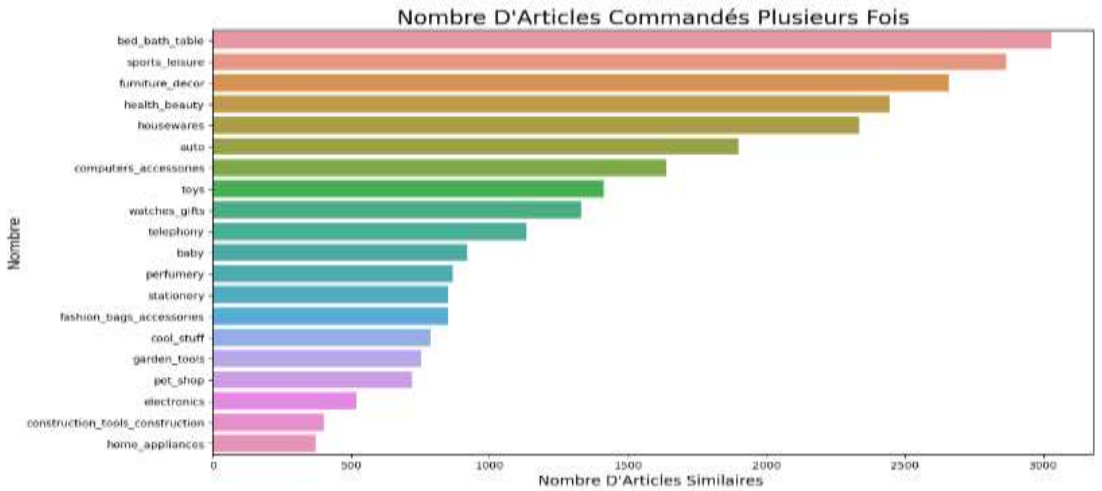


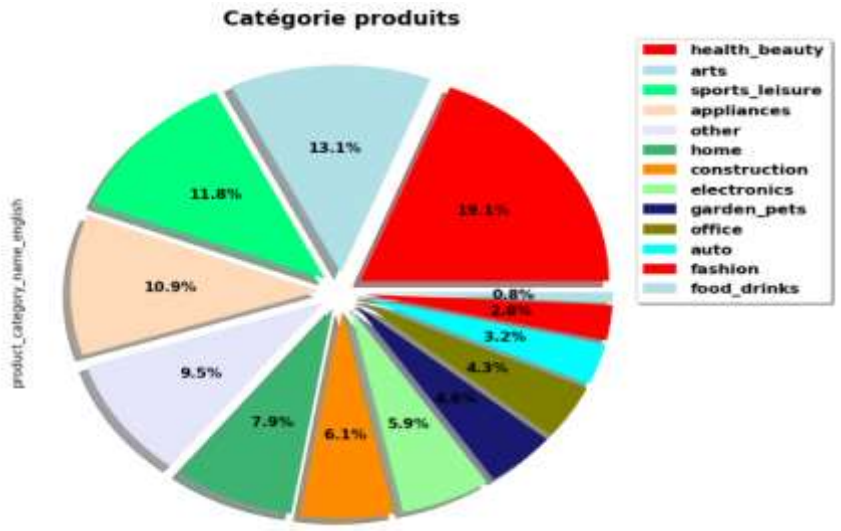


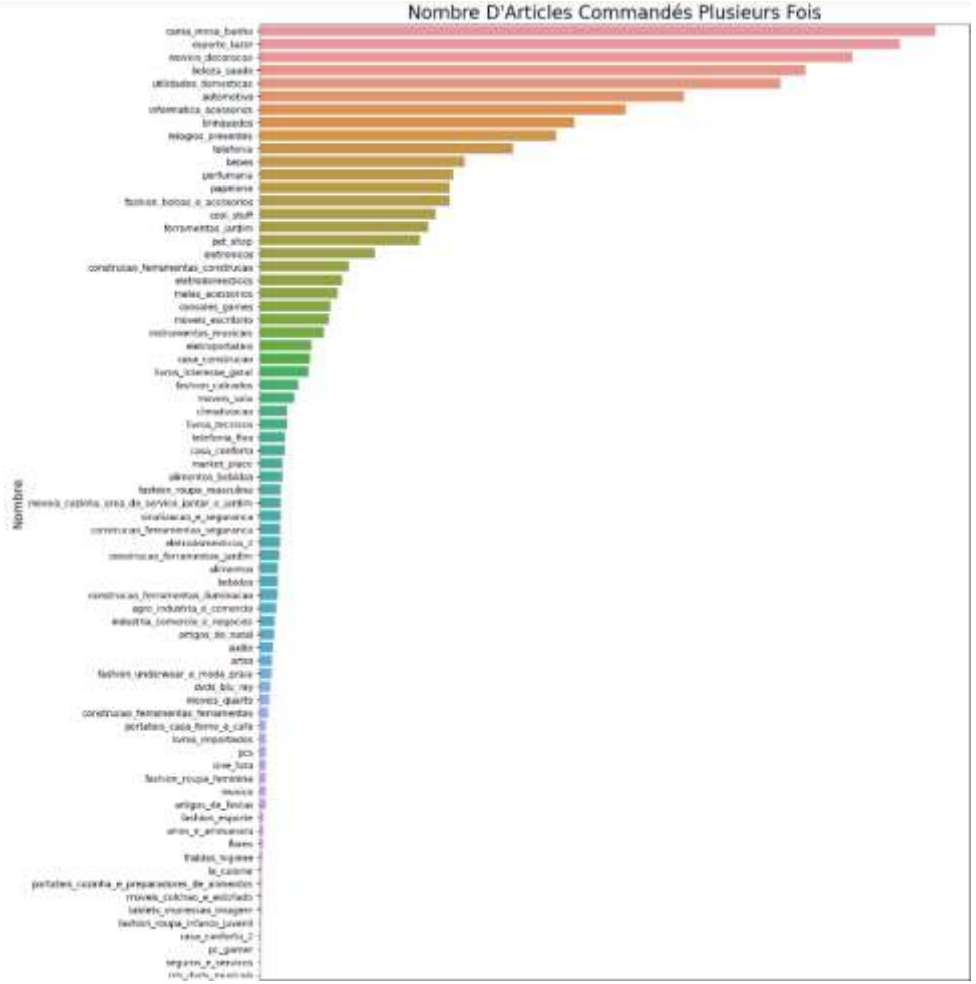


Données – Analyse exploratoire

Produits

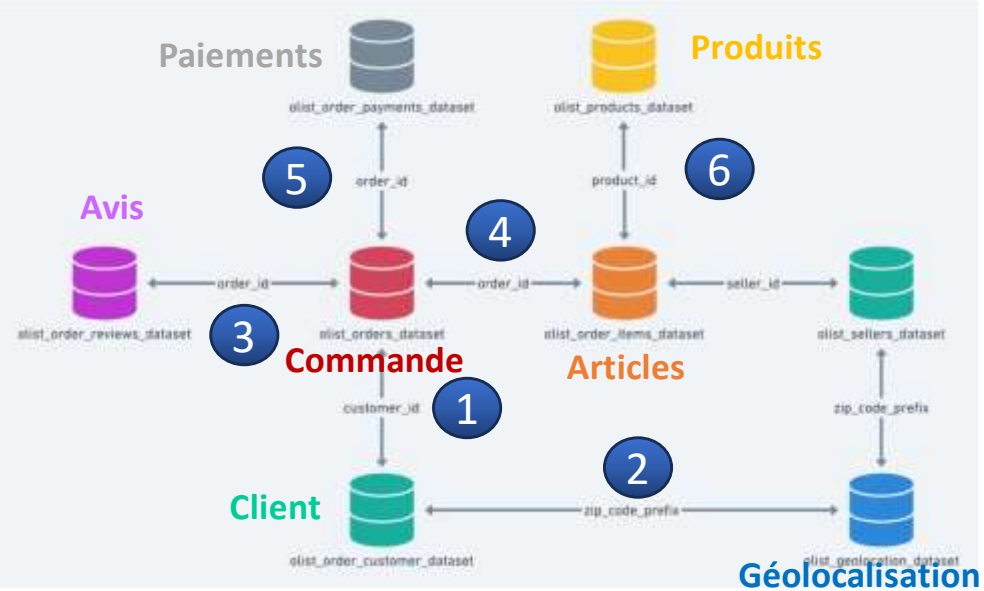






Données - Fusion des données – Nettoyage

1 - FUSION – Clés des datasets et leurs assemblages



2 - Nettoyage

- Client/géolocalisation: variable **code postal** transformée en int64

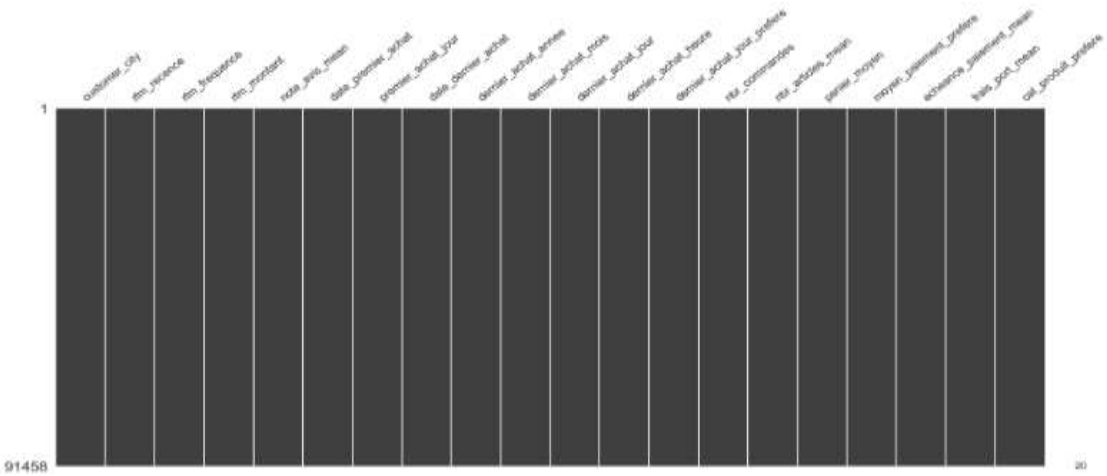
- Suppressions de variables inutiles après fusion

- Valeurs manquantes: peu (0.48 %) **->** dropna()

- Traitement des valeurs aberrantes: gestion des dates entre le traitement de la commande, et la date de la livraison

- **Filtre** sélectionné sur les commandes livrées

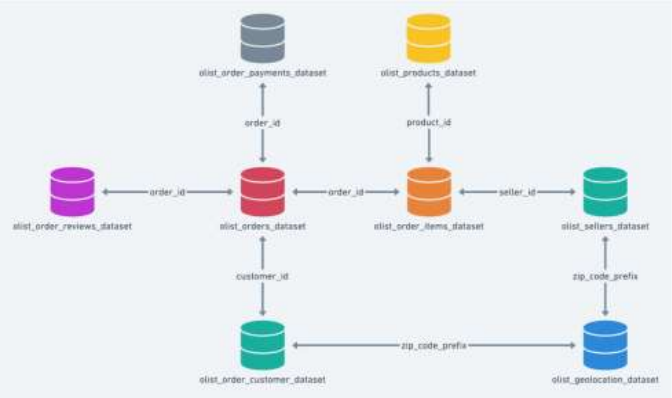
3 –Dataset Final



Données – Ajout variables

KPI clients pertinents qui déterminent une segmentation ?

1 **Datamining sur les datasets, et compréhension métier**



2 **Dataset fusionné**

3 **Datasets KPI CLIENTS**

SEGMENTATION RFM

4 **CLASSIFICATION NON SUPERVISEE**

Données – Indicateurs clients

GEOGRAPHIQUES

Jour avec une forte fréquence de commandes

Localisation:
ville résidence
état résidence
Géolocalisation:
latitude
longitude

COMPORTEMENTAUX

R

Récence de l'activité client (achats, visites)

F

Fréquence d'achat

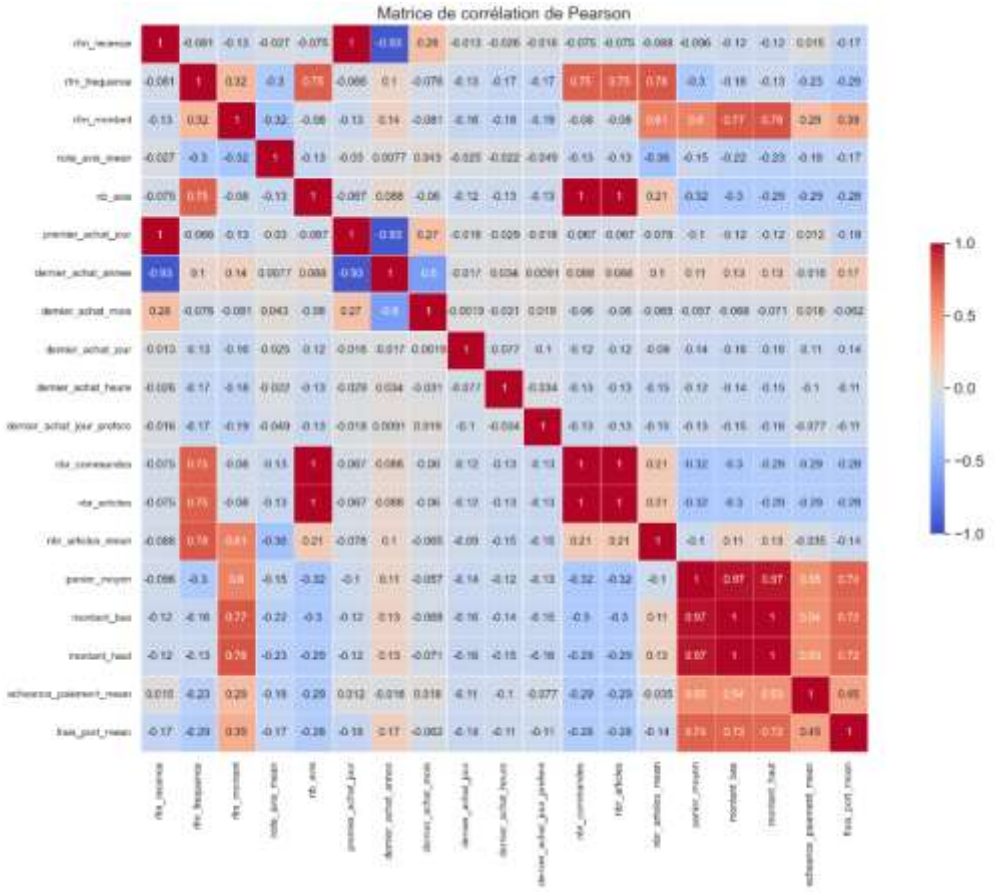
M

Montant panier moyen

PSYCHOLOGIQUES

Note moyenne de satisfaction, **avis**

Données – Analyse multivariée



2 datasets finaux pour la segmentation

Segmentation RFM

3 variables

Algorithmes de classification non supervisée

20 variables

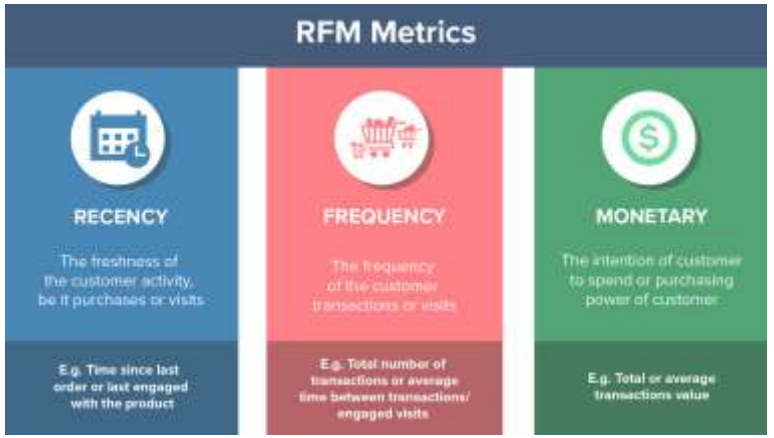
Modélisations

1. Problématique
2. Données
3. **Modélisations**
4. Conclusions

Modélisations – Segmentation

1 **Segmentation RFM**

3 variables



Segmentation CLIENTS

2 **Algorithmes de classification non supervisée**

20 variables

- KMeans
- Kmeans + ACP
- KPrototype

X Clusters possibles

Modélisations

1. Problématique
2. Données
3. Modélisations – Segmentations RFM
4. Conclusions

RFM – Scores et Clients

R F M

customer_unique_id	rfm_recence	rfm_frequence	rfm_montant	R	F	M	RFM_Segment	RFM Score	RFM_Score	Segment
0000366f3b9a7992bf8c76cfdf3221e2	112	1	141.90	4	1	4	414	414	9	Recent Customers
0000b849f77a49e4a4ce2b2a4ca5be3f	115	1	27.19	4	1	1	411	411	6	Potential Loyalist
0000f46a3911fa3c0805444483337064	537	1	86.22	1	1	2	112	112	4	Lost customers
0000f6ccb0745a6a4b88665a16c9f078	321	1	43.62	2	1	1	211	211	4	About to Sleep
0004aac84e0df4da2b147fca70cf8255	288	1	196.89	2	1	4	214	214	7	Others

1 → 2 → 3 → 4

R: récence, F: fréquence,
M: montant
séparés en 4 quartiles

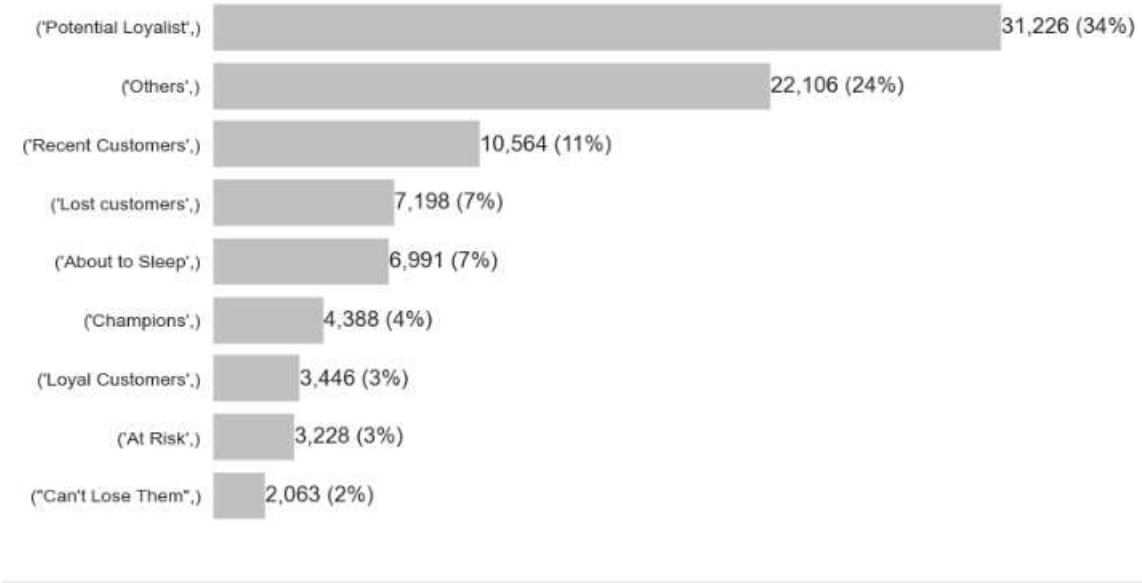
Valeurs Codées:
444 -> Meilleur client
111 -> Pire client

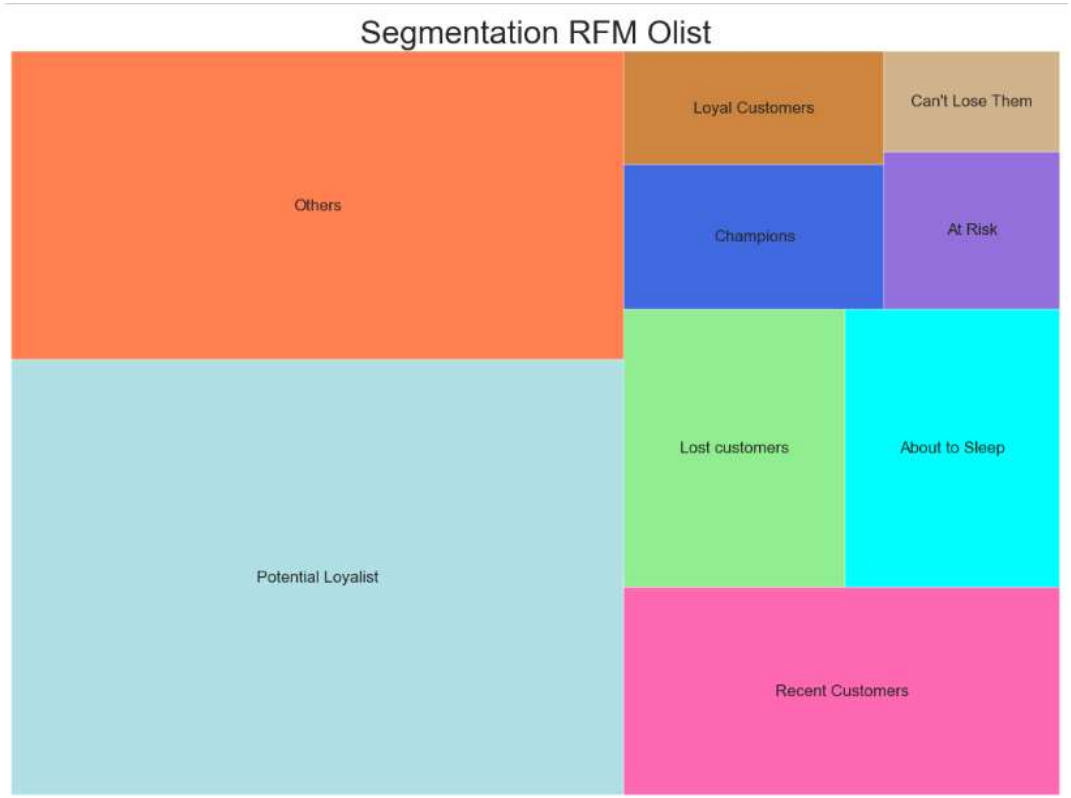
Scoring Client

Dénomination Clientèle

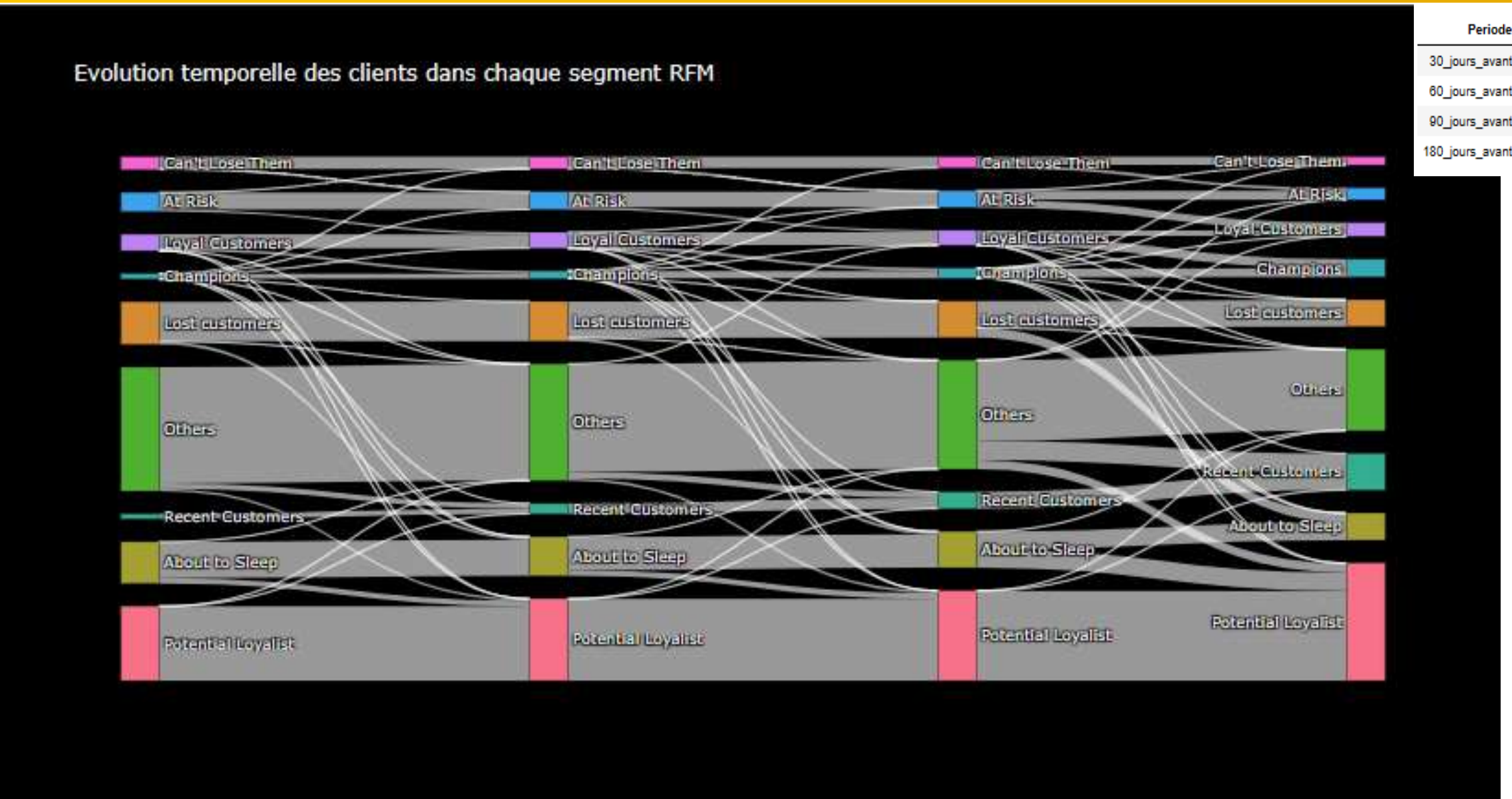
	R	F	M
Champions	4/5	4/5	4/5
Loyal customers	3/4/5	3/4/5	2/3/4/5
Potential loyalist	3/4/5	1/2/3	1/2/3
Recent customers	4/5	1/2	1/2/3/4/5
Promising	3/4	1	1
Need Attention	2/3	2 /3	2/3
About to Sleep	2/3	1/2	1/2
Can't lose them	1/2	4 /5	4 /5
At risk	1/2	2/3/4/5	2/3/4/5
Lost customers	1/2	1/2	1/2
Others	1/2/3/4/5	1/5	1/2/3/4/5

RFM – Visualisation Segmentation





RFM – Stabilité segments(Quantile)



Periode	ARI_rfm (%)	ARI_clusters (%)
30_jours_avant	68.703000	91.264000
60_jours_avant	58.174000	75.589000
90_jours_avant	61.258000	67.088000
180_jours_avant	57.385000	67.489000

RFM – Stabilité segments Clustering

t0+30 j — t0+60 j — t0+90 j — t0+180 j

Rang I: Clients VIP Champions

Rang II: Clients VIP à ne pas perdre

Rang III: Clients fidèles

Rang IV: Clients en dangers

PROPOSITION CONTRAT MAINTENANCE

Mise à jour trimestrielle:

- Bonne stabilité de l'ensemble des clients sur 1 mois (t0 +30 j)
- Bonne stabilité des clients VIP sur 3 mois (t0 + 90 j)
- (1/3) des clients VIP Champions (R I) se transforment en clients VIP à ne pas perdre (R II)
- Plus de (2/ 3) des clients VIP à ne pas perdre'(R II) se transforment en clients fidèles (R III)

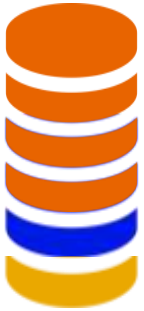
⟹ Surveillance, actions à effectuer au-delà des 3 mois

Modélisations

1. Problématique
2. Données
3. **Modélisations – Apprentissage non supervisée**
4. Conclusions

Clustering - fonctionnement

KPI Clients



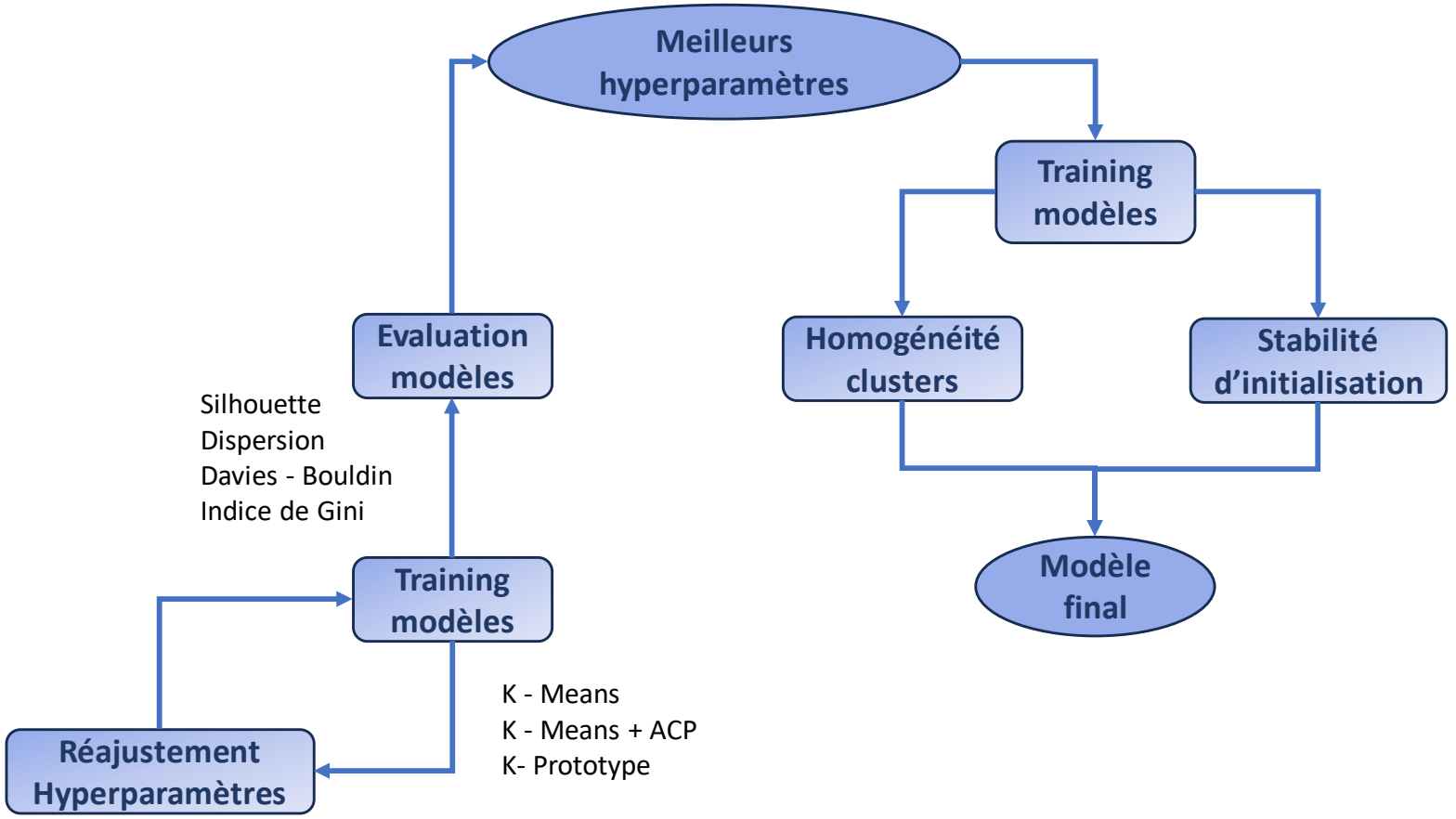
20 variables

Pré-processing

Standardisation,
Transfo. Log
Encodage

Techniques réduction dimension

ACP, TSNE

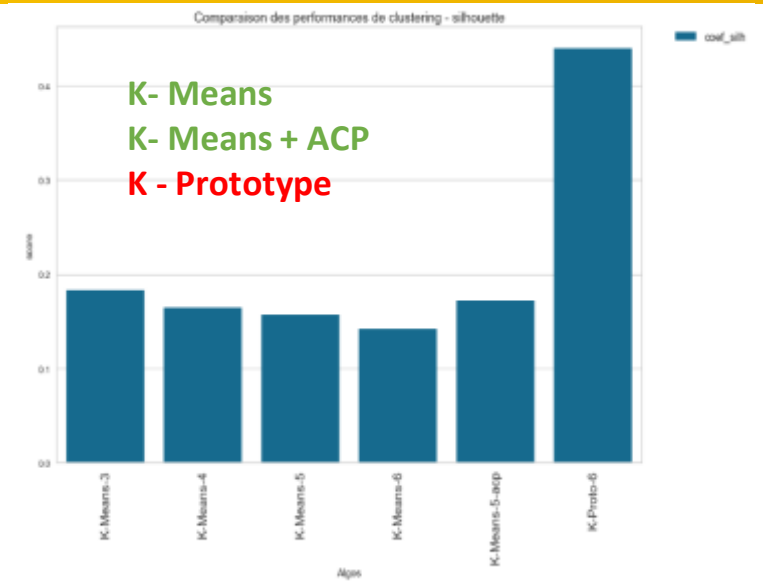


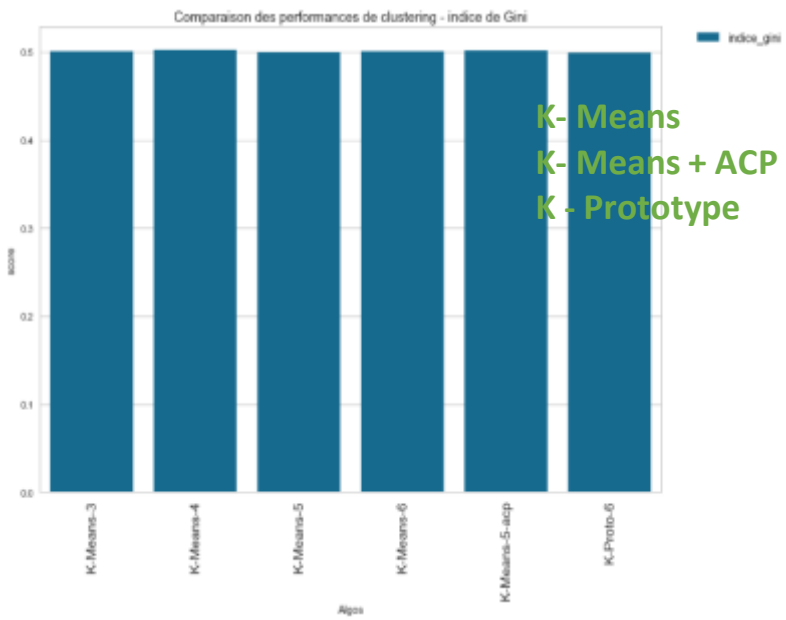
Clustering – Performances

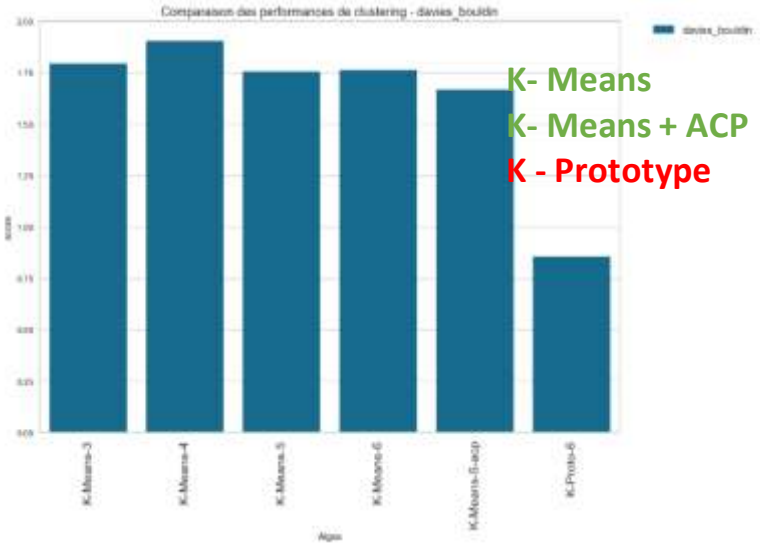
Comparaison sur 15 000 Clients

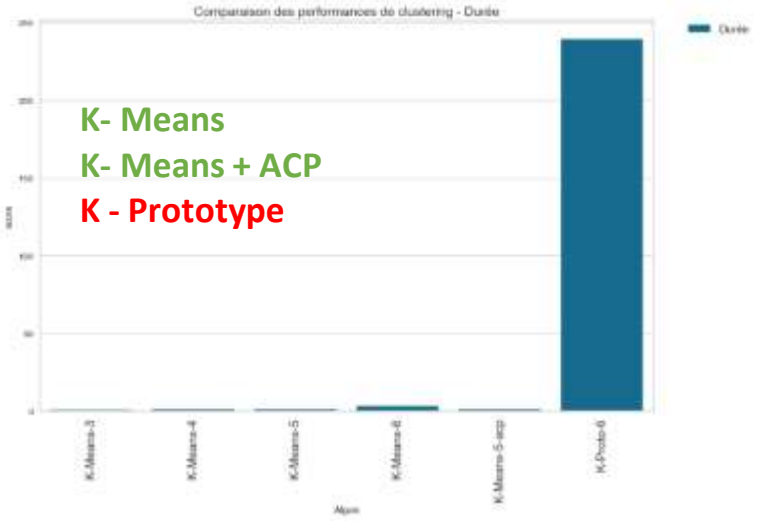
Algos	Nb_clusters	coef_silh	indice_gini	davies_bouldin	Durée
K-Means-3	3	0.183388	0.498198	1.795041	1.293138
K-Means-4	4	0.164147	0.502767	1.906098	1.381447
K-Means-5	5	0.157271	0.500433	1.752424	1.247740
K-Means-6	6	0.141841	0.500773	1.765386	3.157904
K-Means-5-acp	5	0.172155	0.500433	1.666656	1.495468
K-Proto-6	6	0.440042	0.500852	0.856480	215.145733

Algorithme le plus performant et rapide:
K-Means



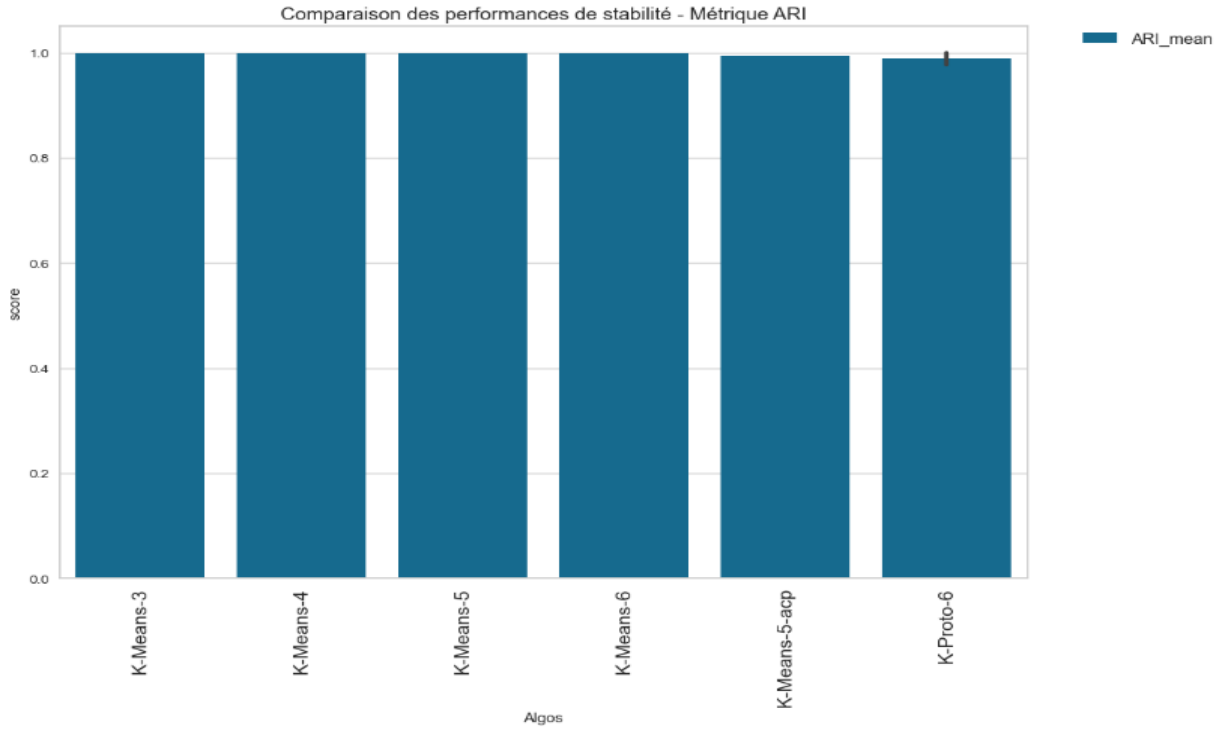






Clustering – Stabilité partitions

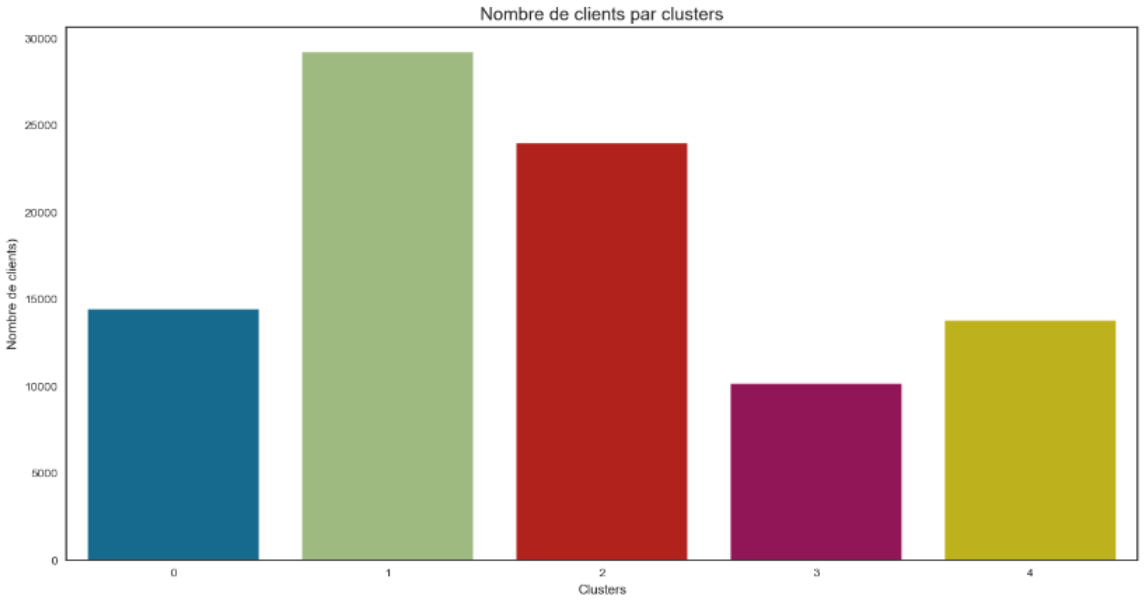
Algos	ARI_mean	ARI_std
K-Means-3	1.000000	0.000000
K-Means-4	1.000000	0.000000
K-Means-5	1.000000	0.000000
K-Means-6	1.000000	0.000000
K-Means-5-acp	0.994410	0.005584
K-Proto-6	0.976857	0.029878



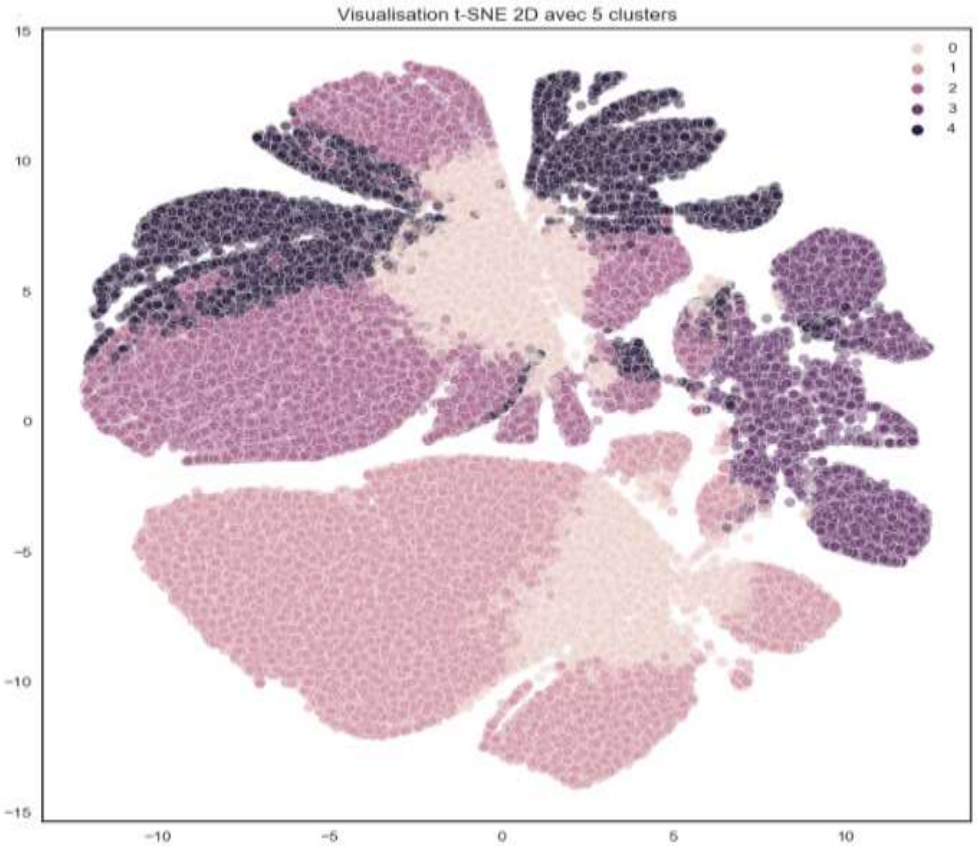
Tous les algorithmes sont stables:

K – Means/ K – Means +ACP / K - Prototypes stables

Clustering – K - Means (exemple k=5)

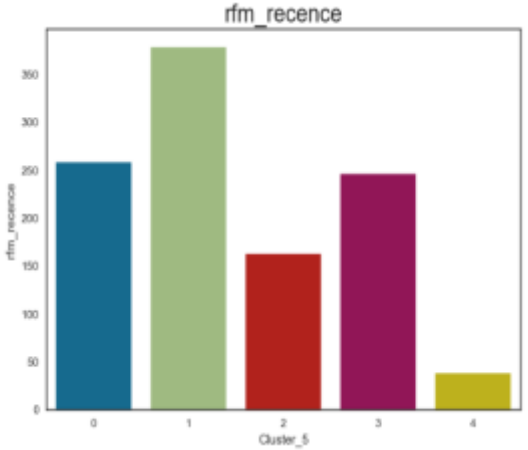


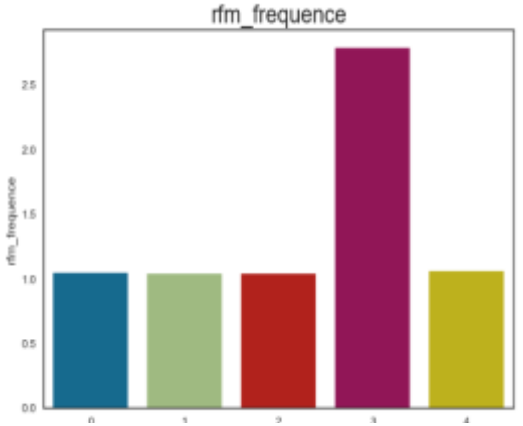
Clusterisation assez stable
des segments

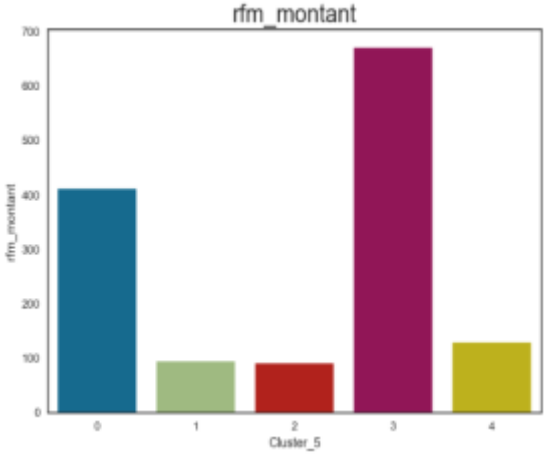


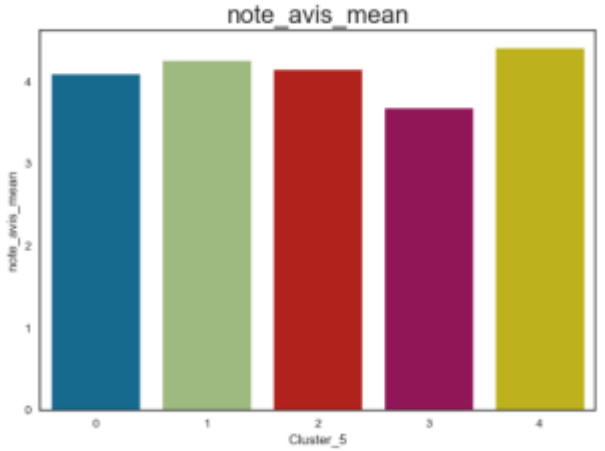
Clusters distinguables
entre eux

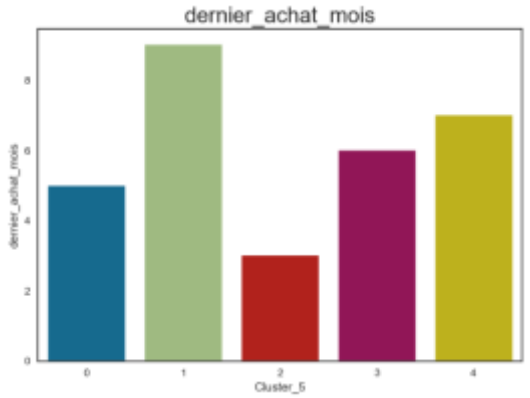
Clustering – K - Means (exemple k=5)

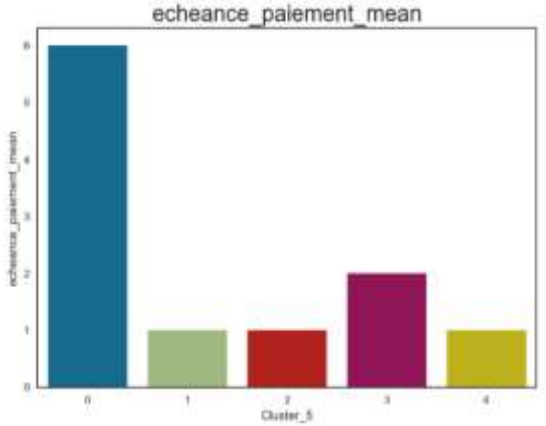






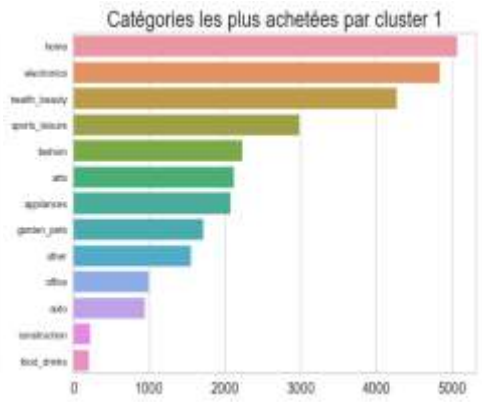


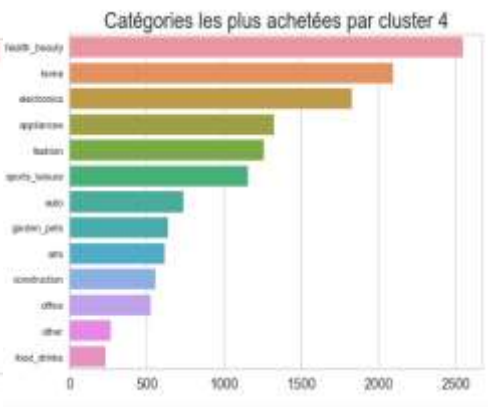




Cluster 0: meilleurs clients utilisant des facilités de paiement.
Cluster 1: clients presque perdus.
Cluster 2: meilleurs clients fidèles
Cluster 3: Bons clients sur le point de dormir
Cluster 4: nouveaux clients.

Clustering – K - Means (exemple k=5)





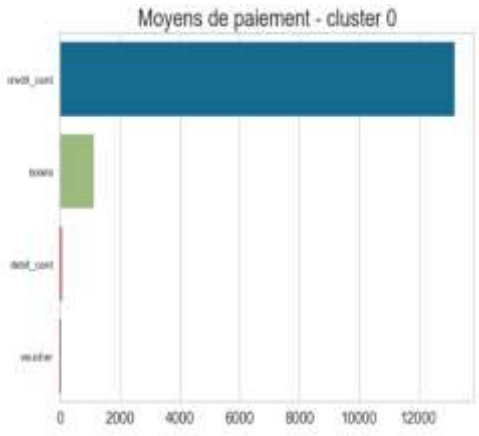
Cluster 0: meilleurs clients utilisant des facilités de paiement.

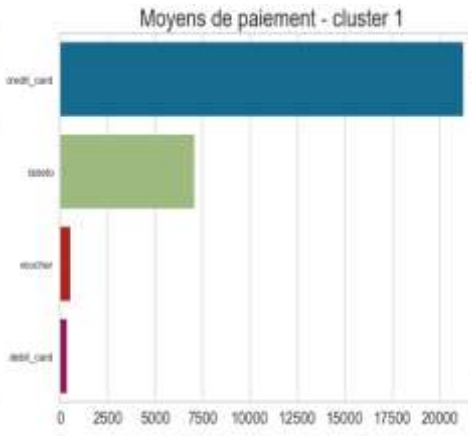
Cluster 1: clients presque perdus.

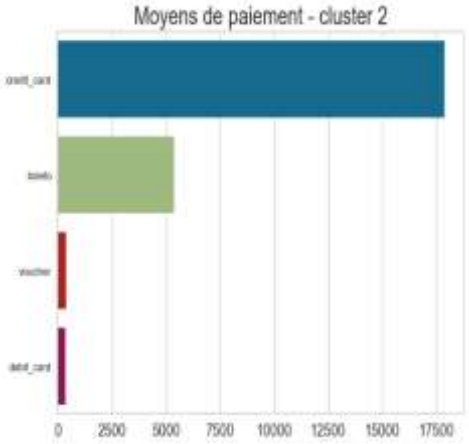
Cluster 2: meilleurs clients fidèles

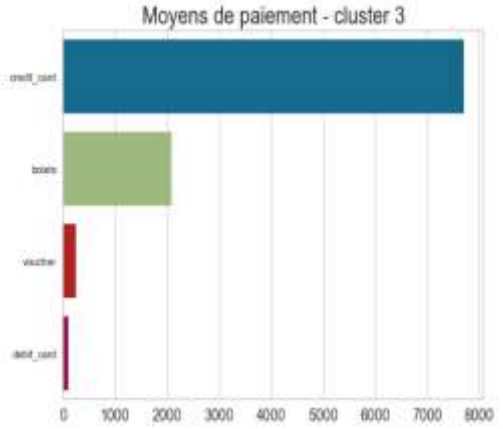
Cluster 3: Bons clients sur le point de dormir

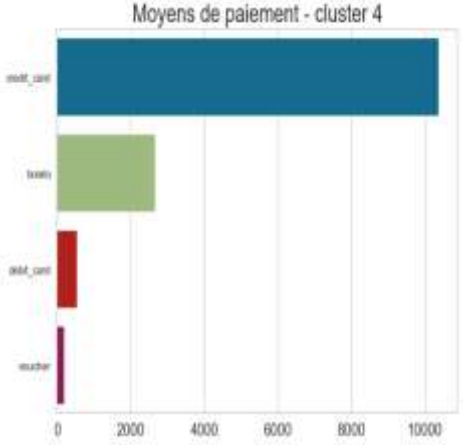
Cluster 4: nouveaux clients.





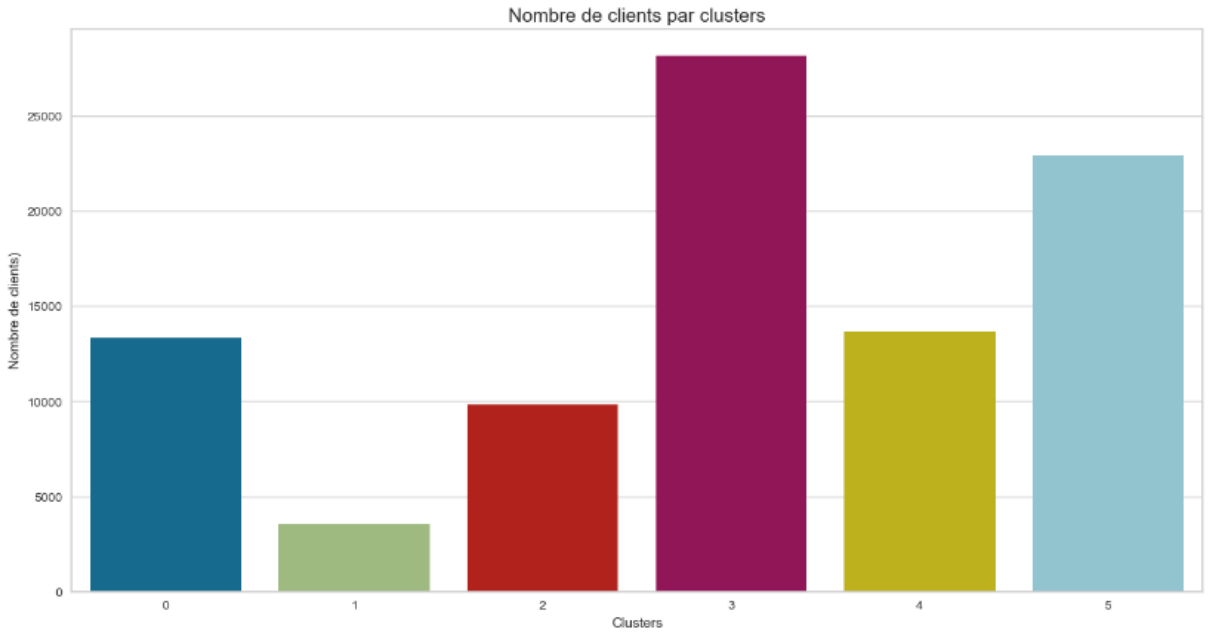






Catégories de moyens de paiement par cluster

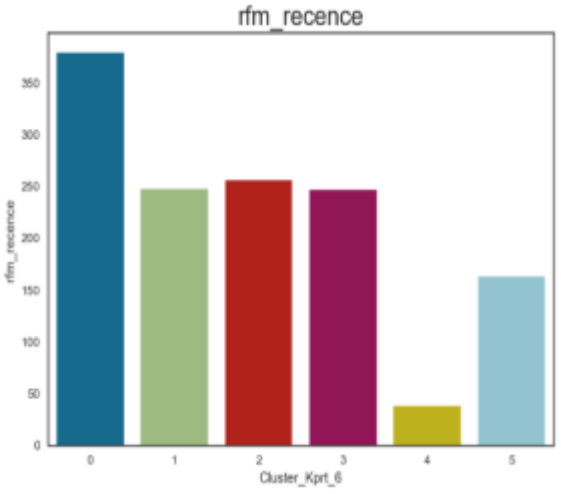
Clustering – K – Prototype (exemple k=6)

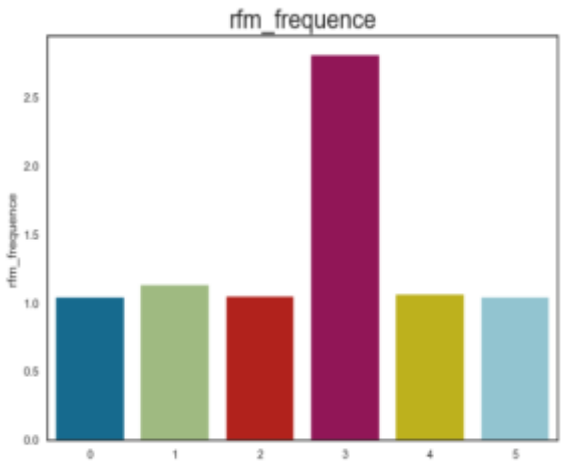


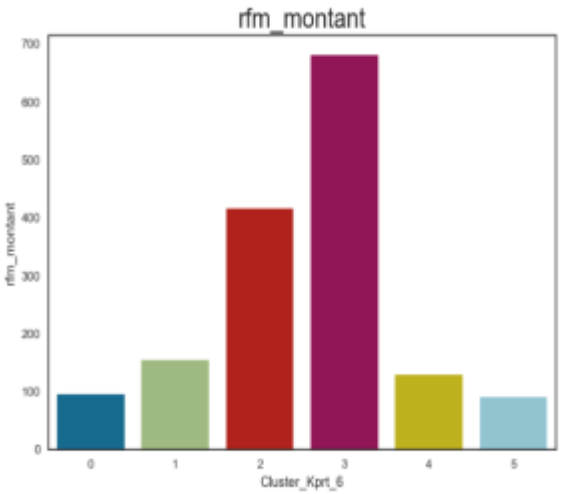
Clusterisation assez stable
des segments

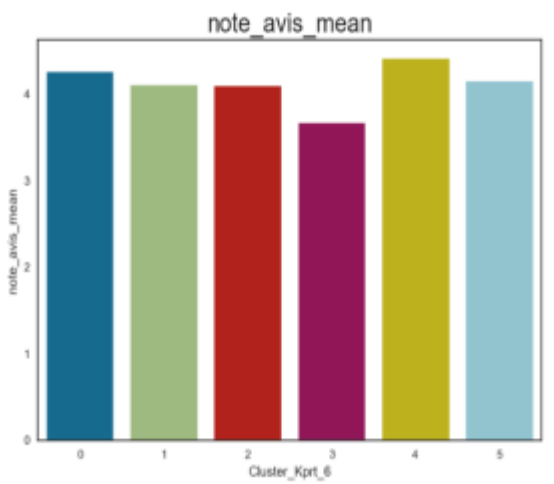
Clusters distinguables
entre eux

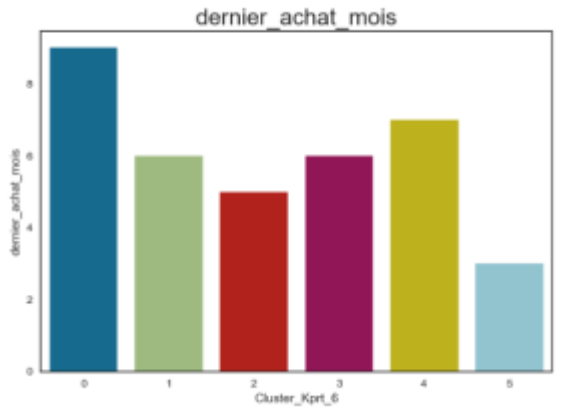
Clustering – K – Prototype (exemple k=6)

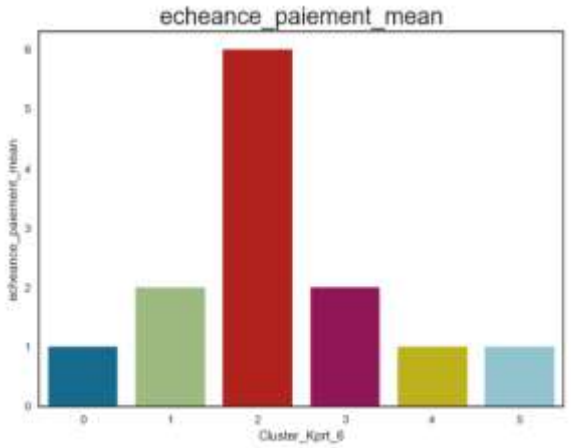






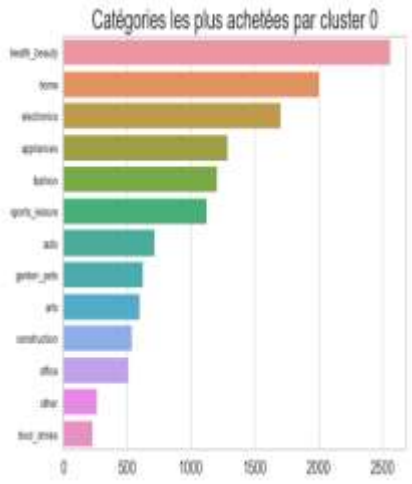




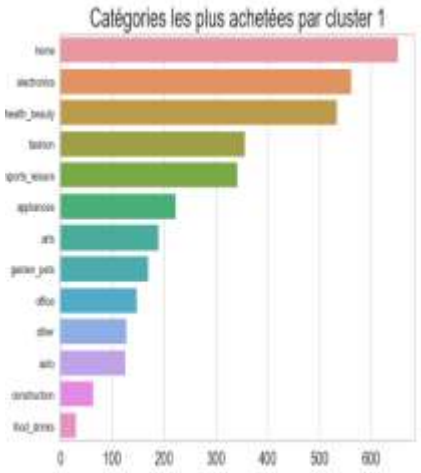


Cluster 0: clients perdus
Cluster 1: Meilleurs clients utilisant les facilités de paiement ayant besoin d'attention.
Cluster 2: Clients fidèles (hors Nord).
Cluster 3: Meilleurs clients ayant besoin d'attention.
Cluster 4: nouveaux clients.
Cluster 5 : Clients fidèles du Nord du Brésil achetant la catégorie 'fashion' ayant besoin d'attention.

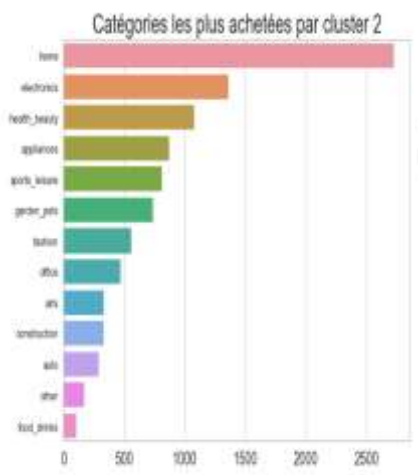
Clustering – K – Prototype (exemple k=6)



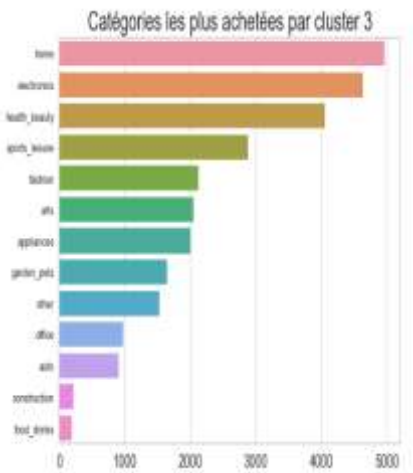
Cluster 0: clients perdus



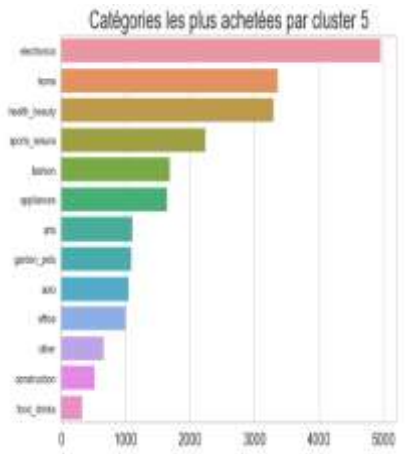
Cluster 1: Meilleurs clients utilisant les facilités de paiement ayant besoin d'attention.



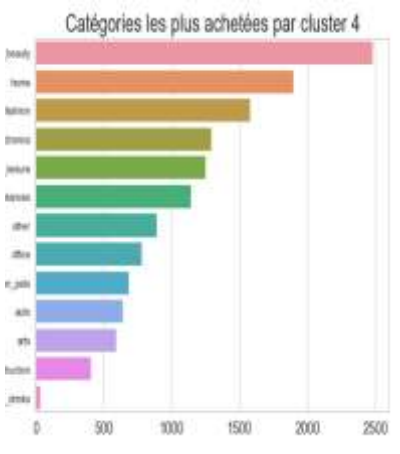
Cluster 2: Clients fidèles (hors Nord).



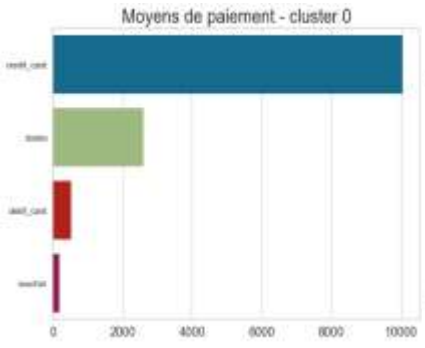
Cluster 3: Meilleurs clients ayant besoin d'attention.

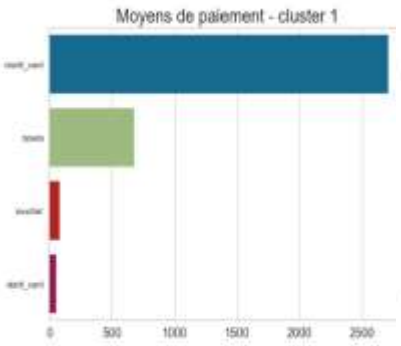


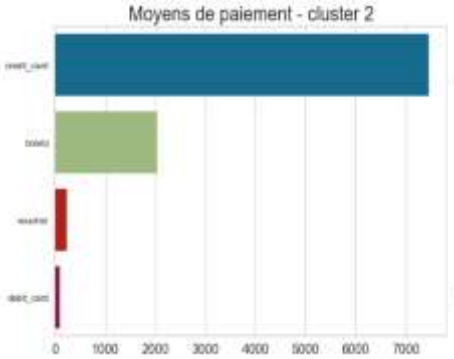
Cluster 4: nouveaux clients.

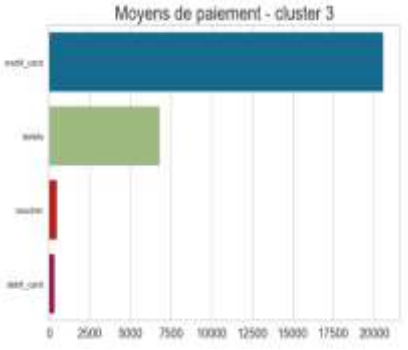


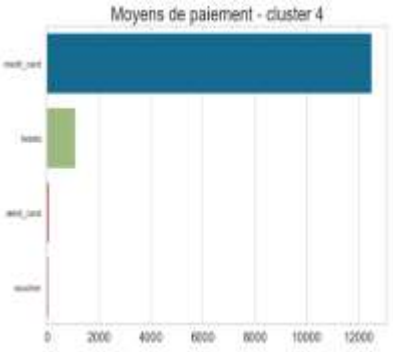
Cluster 5 : Clients fidèles du Nord du Brésil achetant la catégorie 'fashion' ayant besoin d'attention.

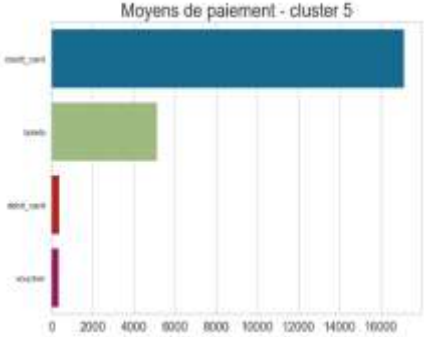






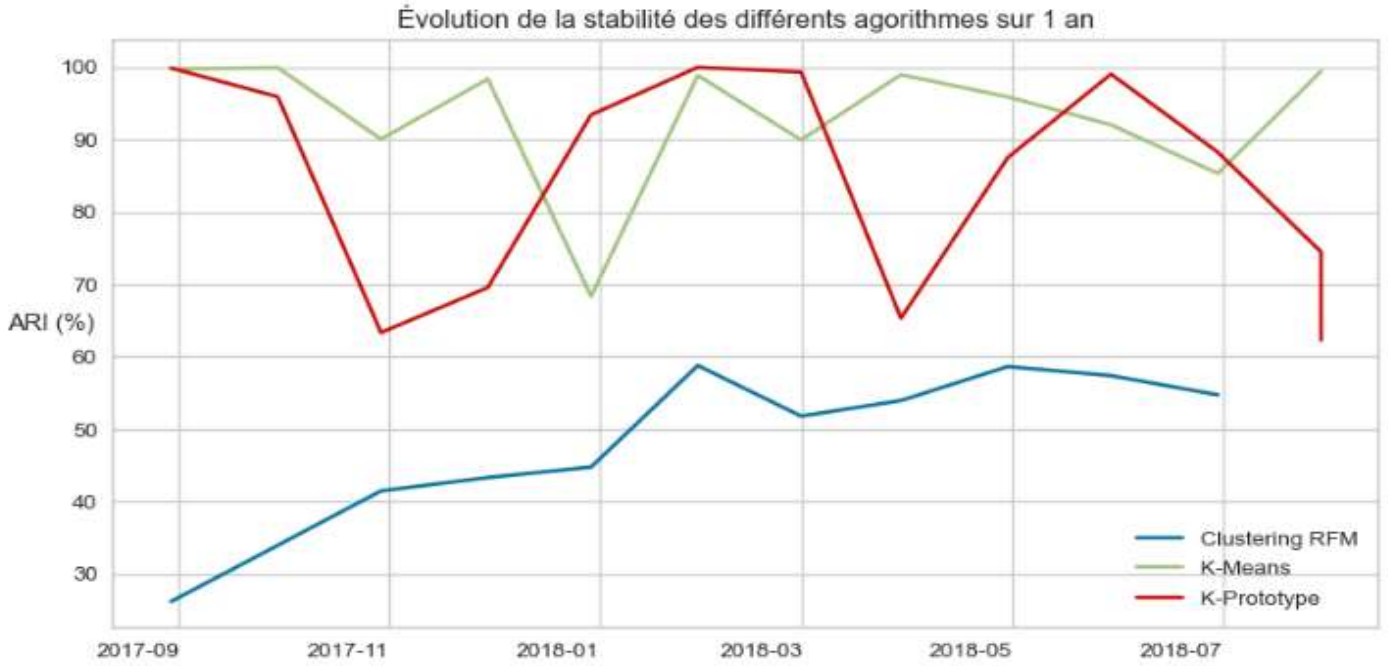






Catégories de moyens de paiement par cluster

Clustering – Stabilité dans le temps



- L'algorithme le plus stable est le K-Means, par rapport au K-Prototype et celui du clustering RFM traditionnel
- Proposition d'un contrat de maintenance, avec une mise à jour des fréquences trimestrielle

Conclusions

1. Problématique
2. Données
3. Modélisations
4. Conclusions

Conclusions – Modèle final

Modèle	Avantages	Inconvénients
Segmentation RFM	Très simple à déployer Très simple à interpréter Segmente de manière nette Marketing traditionnel	Une segmentation trop pauvre pour être actionnable, avec seulement 3 variables entrant en jeu
K - Means	Indicateurs numériques, catégorielles encodées intégrées. Groupes homogènes, stables sur 3 mois ou plus, ainsi que l'initialisation. Prédictions possibles, avec intégration de nouveaux clients	La difficulté d'interprétabilité des variables qualitatives, le choix du paramètre k, au départ.
K-Prototype	L'ensemble des indicateurs mixtes prises en compte (mélange qualitatif, quantitatif), segments interprétables, homogènes, et actionnables avec un cycle de stabilité de 3 mois	Temps d'exécution très lent, et paramétrage assez complexe, sans connaissances, et recherches.

Modèle final:

K - Means avec 6 clusters, ainsi que d'un contrat de révision de la stabilité des segments de façon trimestrielle (**3 mois**).

Conclusions – Eventuelles améliorations

Datasets:

- Problème fréquence des commandes: une majorité effectuent une commande unique
- Manque de facteurs externes, de renseignement supplémentaire sur les tranches de famille (célibataires,…)

Segmentation:

- Affiner encore les segments actuels, en fonction de la demande métier, marketing