SOUTENANCE DU PROJET 5:
SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE

Présenté par Check KOUTAME

PLAN

I/ Mission, description et observation des données du projet

- Mission & Description du projet
- Observations des données: Formes et qualités

II/ Analyse exploratoire & feature engineering

- Analyse exploratoire
- Feature engineering: segmentation RFM

III/ Clustering – modélisation

- K-mean
- DBSCAN
- Hiérarchique

IV/ Contrat de maintenance

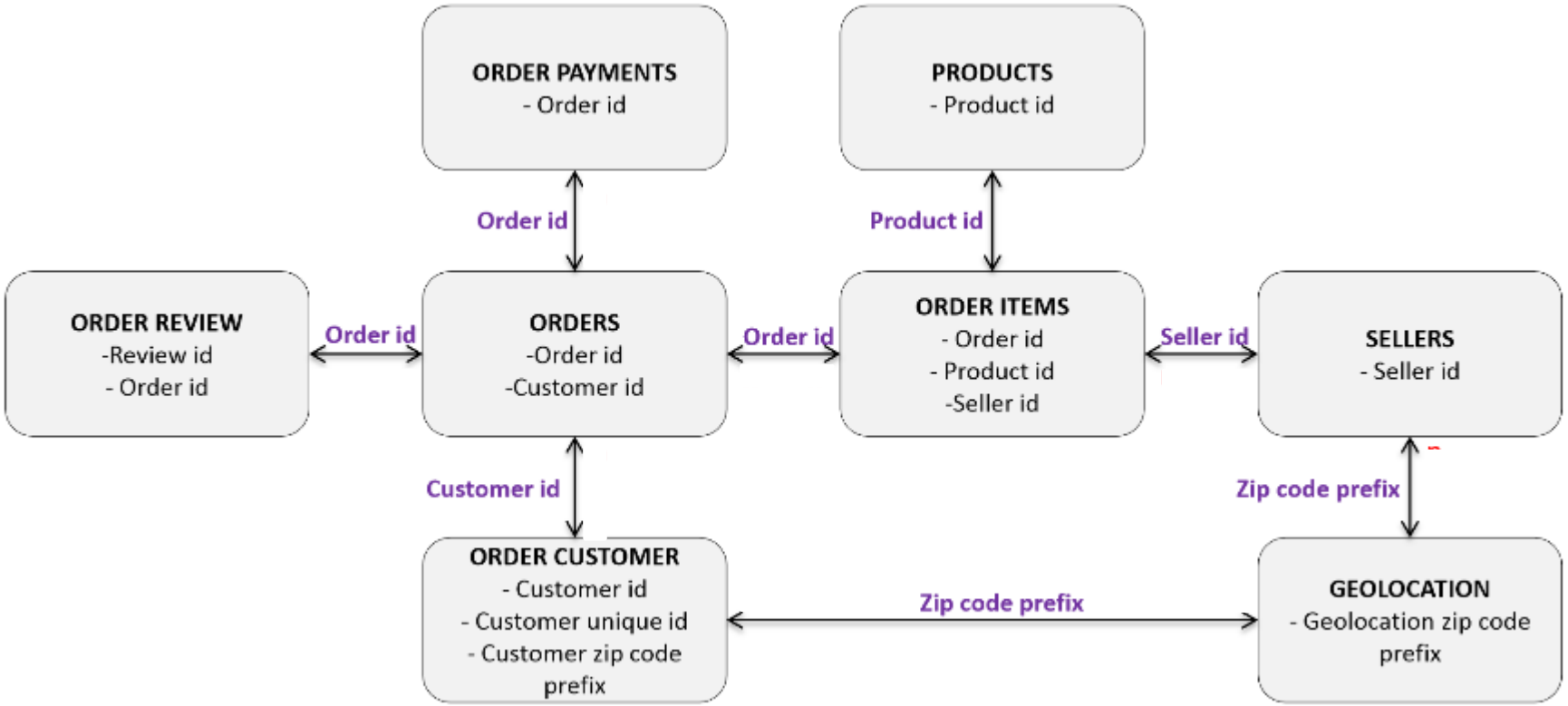
Conclusion

I. MISSION & DESCRIPTION DU PROJET:

- L'objectif pour Olist: Solution de vente sur les marketplaces en ligne
- **Problématique :**
  - Site e-commerce brésilien a mis à disposition une base de données avec des commandes passées entre 2016 et 2018
    - Explorations des données avec plusieurs variables: achat, type de carte d'achat, nombre d'achats…
    - Comment donc établir une segmentation du type de client en fonction de ces données ?
- **Mission :**
  - Etablir une segmentation avec des clusters interprétables pour l'équipe commerciale d'Olist qui pourrait l'utiliser pour un plan d'action:
    - Fidéliser les clients
    - Trouver de nouvelles pistes de publicités et autres démarche commerciales
  - Etablir une stabilité des clusters au fils du temps afin de proposer une maintenance

I. MISSION & DESCRIPTION DU PROJET: NOS DONNÉES: OBSERVATIONS

Données reparties en 8 tables : clients / géolocalisation / type de commandes/ commandes / paiements / produits / vendeurs / avis/ catégories de produits..



I. MISSION & DESCRIPTION DU PROJET: NOS DONNÉES: OBSERVATIONS

Données reparties en 8 tables : clients / géolocalisation / type de commandes/ commandes / paiements / produits / vendeurs / avis/ catégories de produits..

	Dimension	Nombre de variables	Nombre observations	Nombres de types de variables	Nombre de cellules manquantes	% de cellules manquantes	Nombre de lignes dupliquées	% de lignes dupliquées
customers	(99441, 5)	5	99441	object 4 int64 1 dtype: int64	0	0.000000	0	0.000000
geolocalisation,	(1000163, 5)	5	1000163	float64 2 object 2 int64 1 dtype: ...	0	0.000000	261831	0.261788
order_items	(112650, 7)	7	112650	object 4 float64 2 int64 1 dtype: ...	0	0.000000	0	0.000000
order_payments	(103886, 5)	5	103886	int64 2 object 2 float64 1 dtype: ...	0	0.000000	0	0.000000
order_reviews	(99224, 7)	7	99224	object 6 int64 1 dtype: int64	145903	0.210063	0	0.000000
Orders	(99441, 8)	8	99441	object 8 dtype: int64	4908	0.006169	0	0.000000
Products	(32951, 9)	9	32951	float64 7 object 2 dtype: int64	2448	0.008255	0	0.000000
Sellers	(3095, 4)	4	3095	object 3 int64 1 dtype: int64	0	0.000000	0	0.000000
product_category	(71, 2)	2	71	object 2 dtype: int64	0	0.000000	0	0.000000

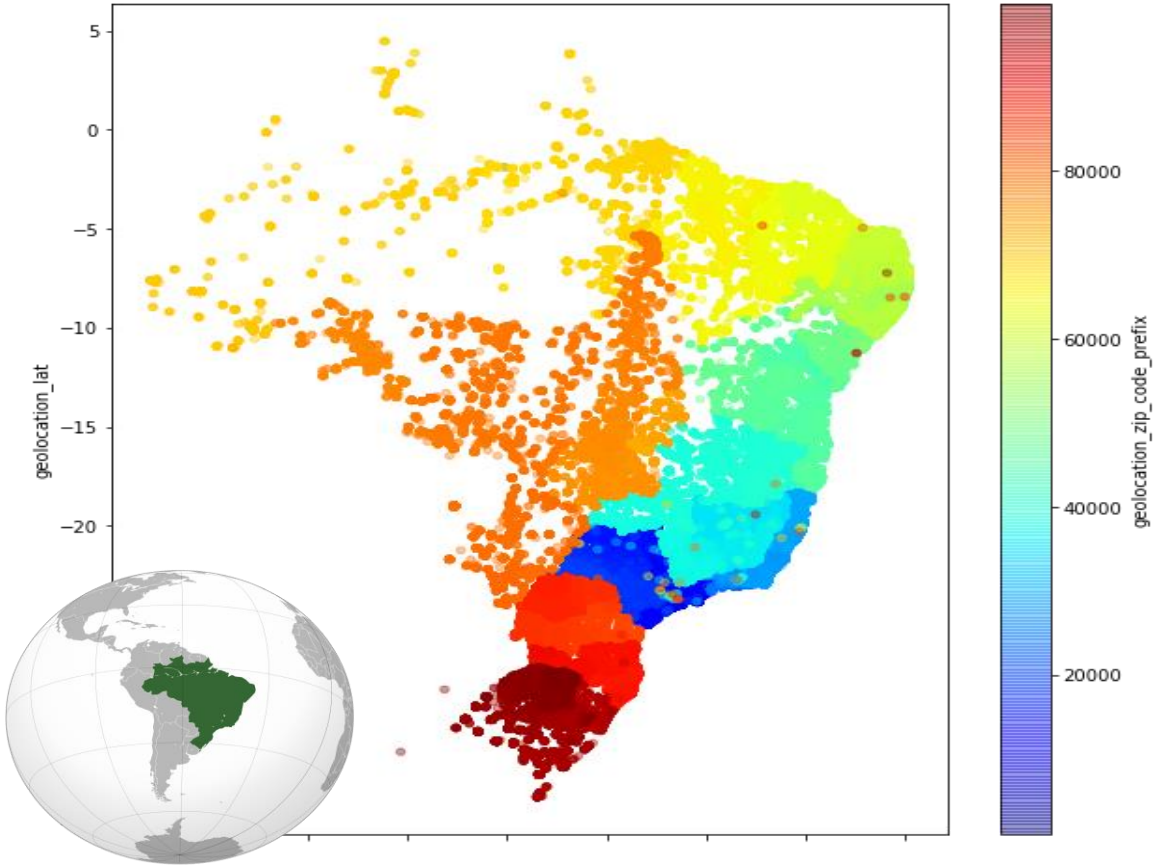
II/ ANALYSE EXPLORATOIRE & FEATURE ENGINEERING NETTOYAGES

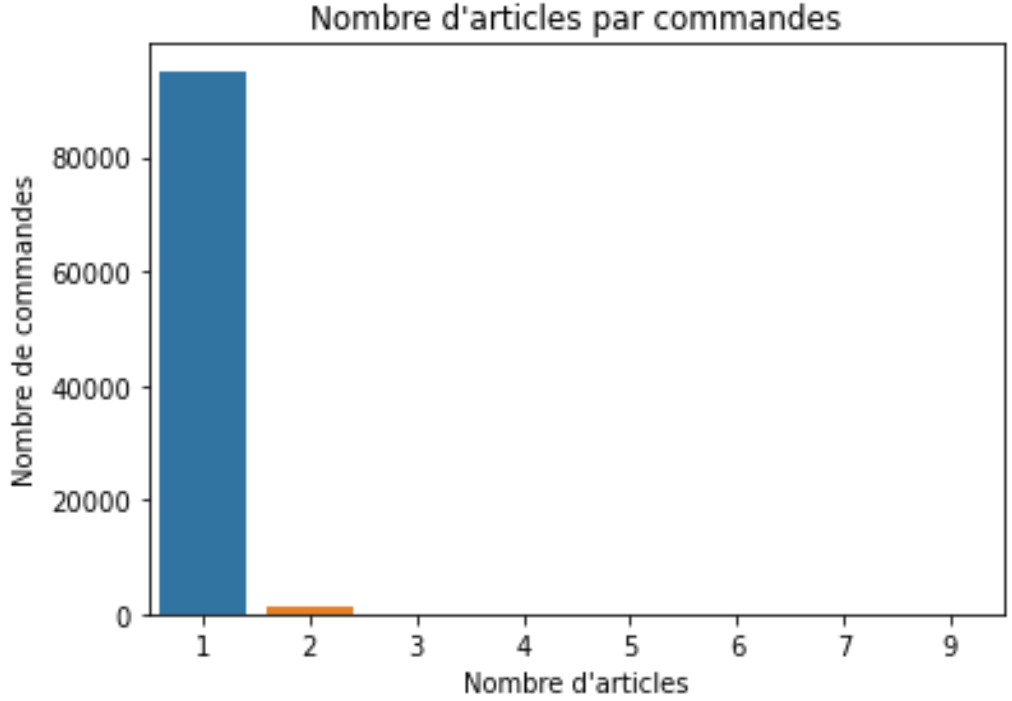
Données reparties en 8 tables : clients / géolocalisation / type de commandes/ commandes / paiements / produits / vendeurs / avis/ catégories de produits..

Pas trop d'absence de données

II/ ANALYSE EXPLORATOIRE & FEATURE ENGINEERING
ANALYSE EXPLORATOIRE

Données reparties en 8 tables : clients / **géolocalisation** / **type de commandes**/ **commandes** / paiements / produits / vendeurs / avis/ catégories de produits..

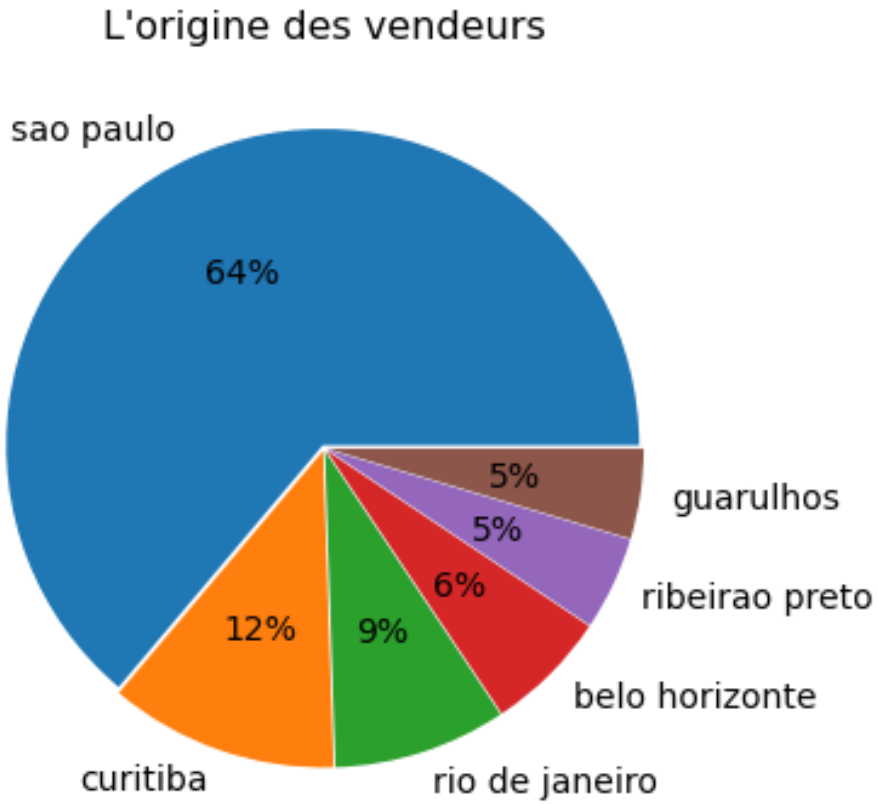


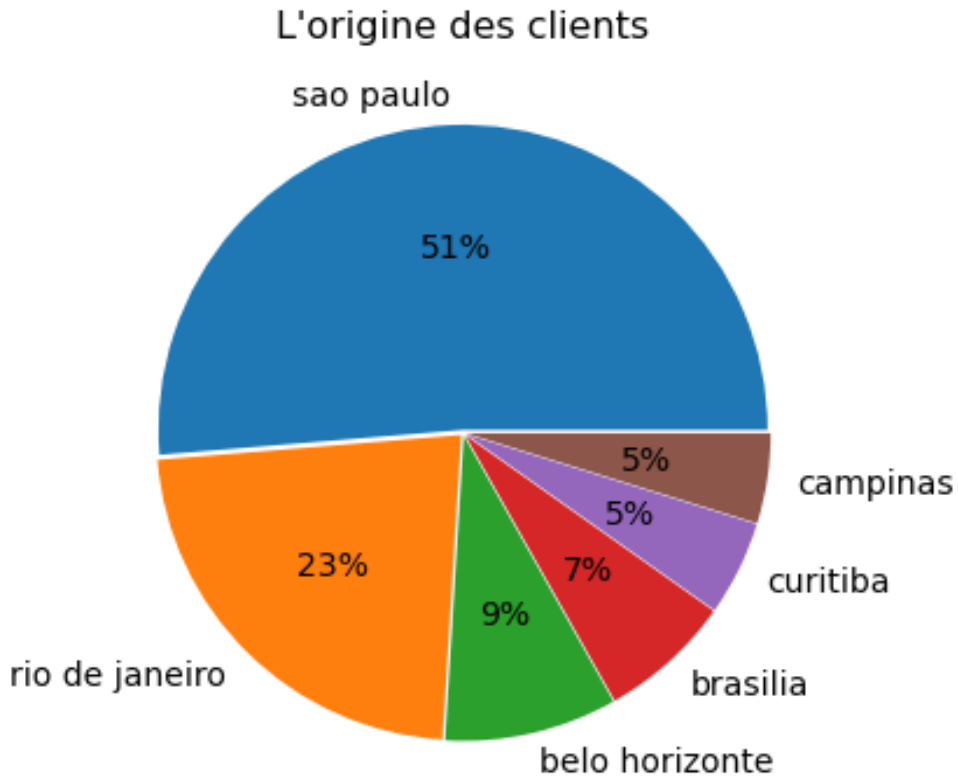


- **La plus grande majorité de commandes comporte un seul article**

II/ ANALYSE EXPLORATOIRE & FEATURE ENGINEERING
ANALYSE EXPLORATOIRE

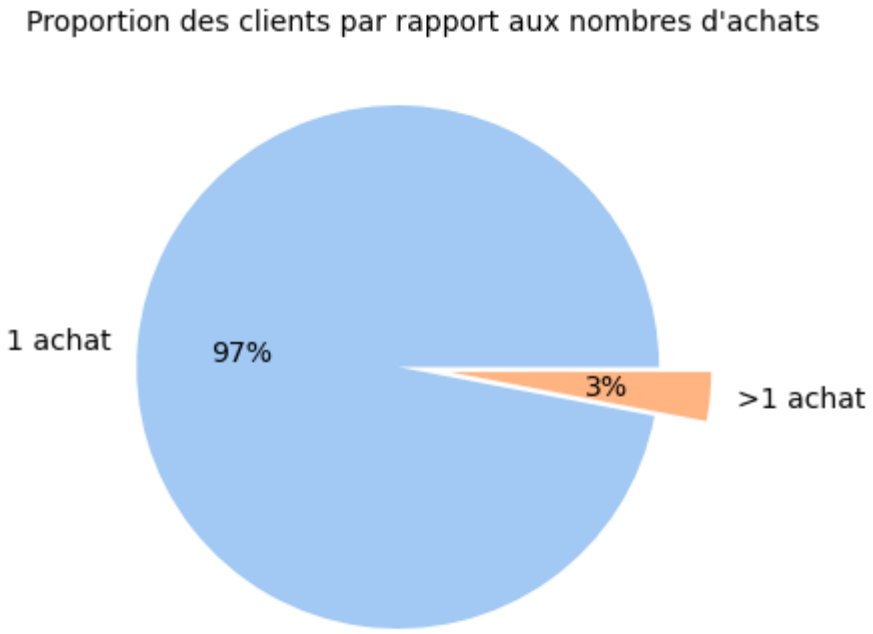
Données reparties en 8 tables : clients / géolocalisation / type de commandes/ commandes / paiements / produits **/ vendeurs** / avis/ catégories de produits..





II/ ANALYSE EXPLORATOIRE & FEATURE ENGINEERING
ANALYSE EXPLORATOIRE

Données reparties en 8 tables : **clients** / géolocalisation / type de commandes/ **commandes** / paiements / produits / vendeurs / avis/ catégories de produits..



Quelques infos utiles.	
Nombre de villes	4093
Nombre d'états	27
Nombre de clients	93404
Nombre de catégories d'articles	71

II/ ANALYSE EXPLORATOIRE & FEATURE ENGINEERING
FEATURE ENGINEERING: SEGMENTATION RFM

Dans ce projet, nous allons utiliser la segmentation RFM. RFM est un acronyme pour Récence, Fréquence et Montant.

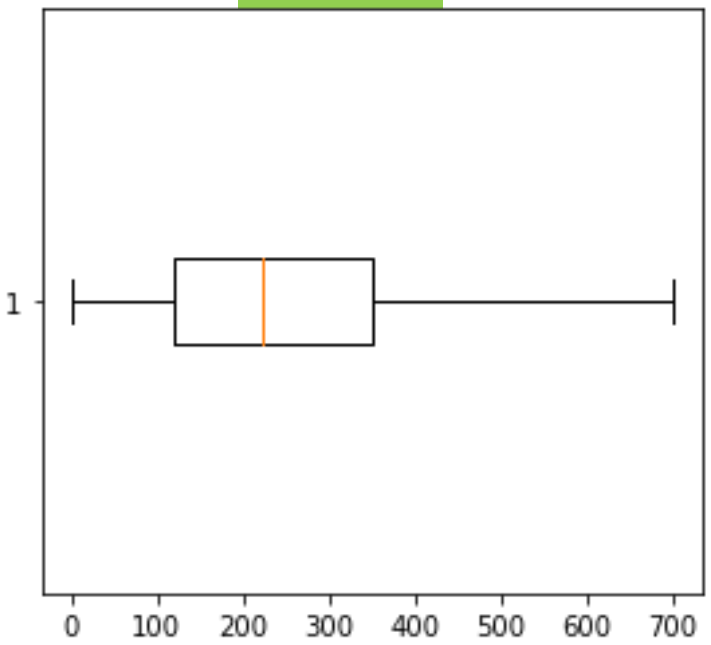
- **Récence** indique la date du dernier achat. Il s'agit du nombre de jours depuis la dernière commande d'un client.
- **Fréquence** est le nombre d'achat sur une période déterminé. Cela peut être 3 mois, 6 mois ou un 1 an. Elle indique la fidélité d'un client, plus sa valeur est élevé et plus le client est engagé
- **Montant** est la somme totale qu'un client dépense sur une période donnée

- Création d'une fonction permettant de calculer ces différentes variables

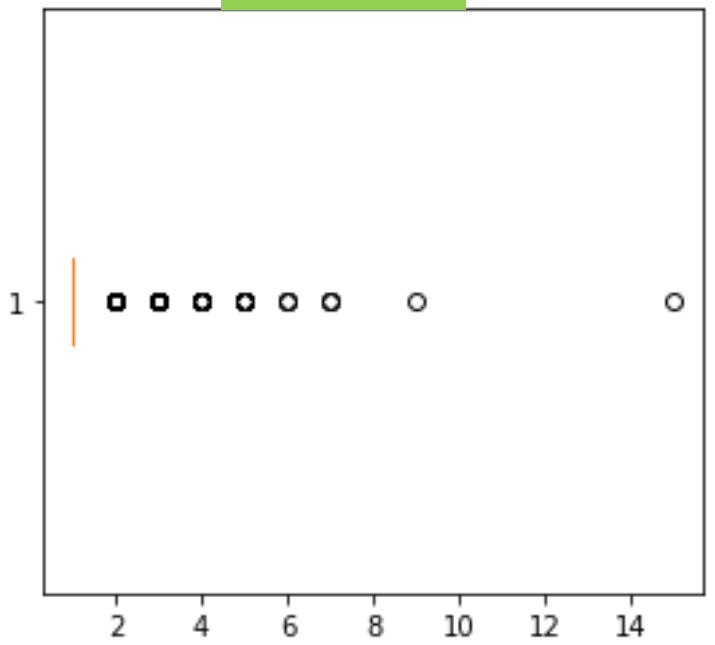
II/ ANALYSE EXPLORATOIRE & FEATURE ENGINEERING
FEATURE ENGINEERING: SEGMENTATION RFM

Box plot des variables RMF sur l'ensemble des données

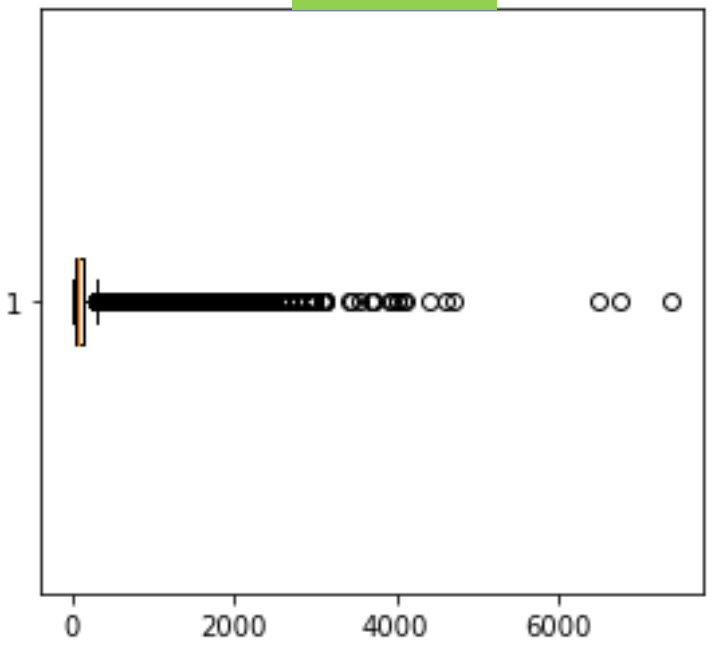
Récence



Fréquence

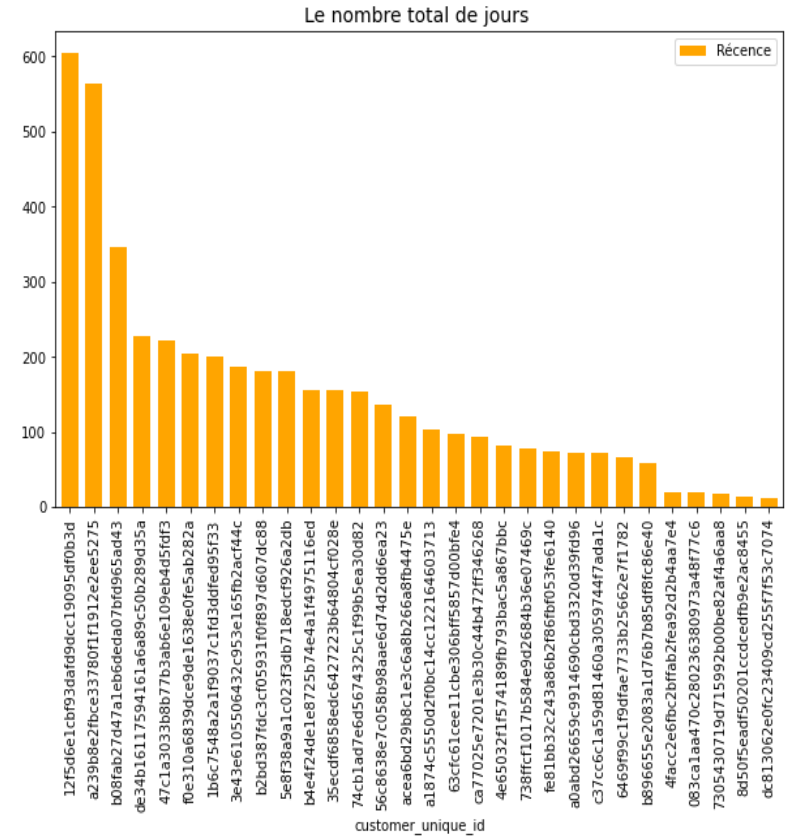


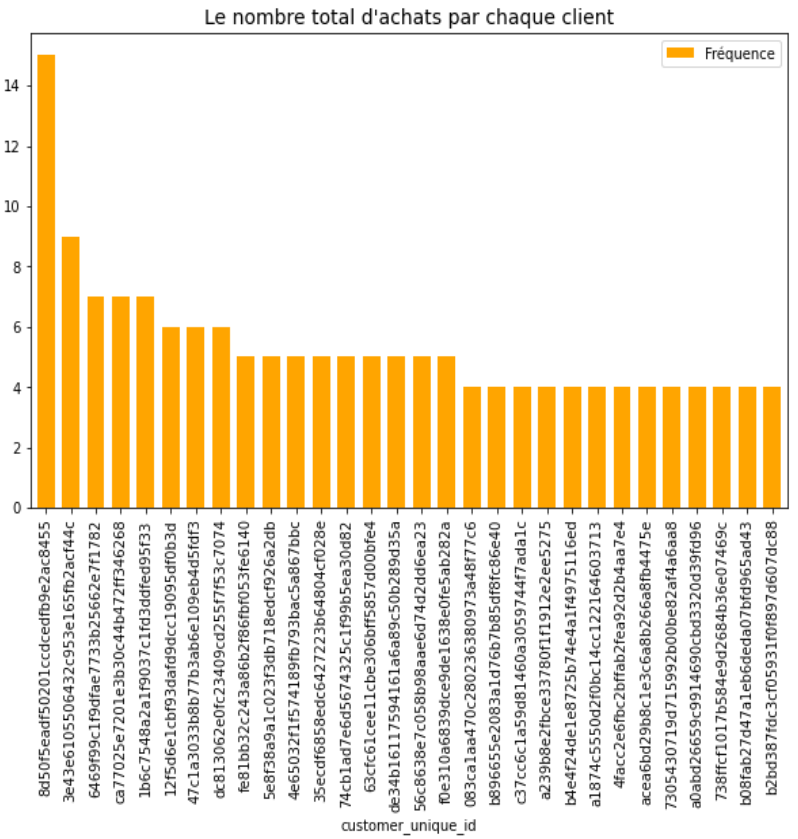
Montant

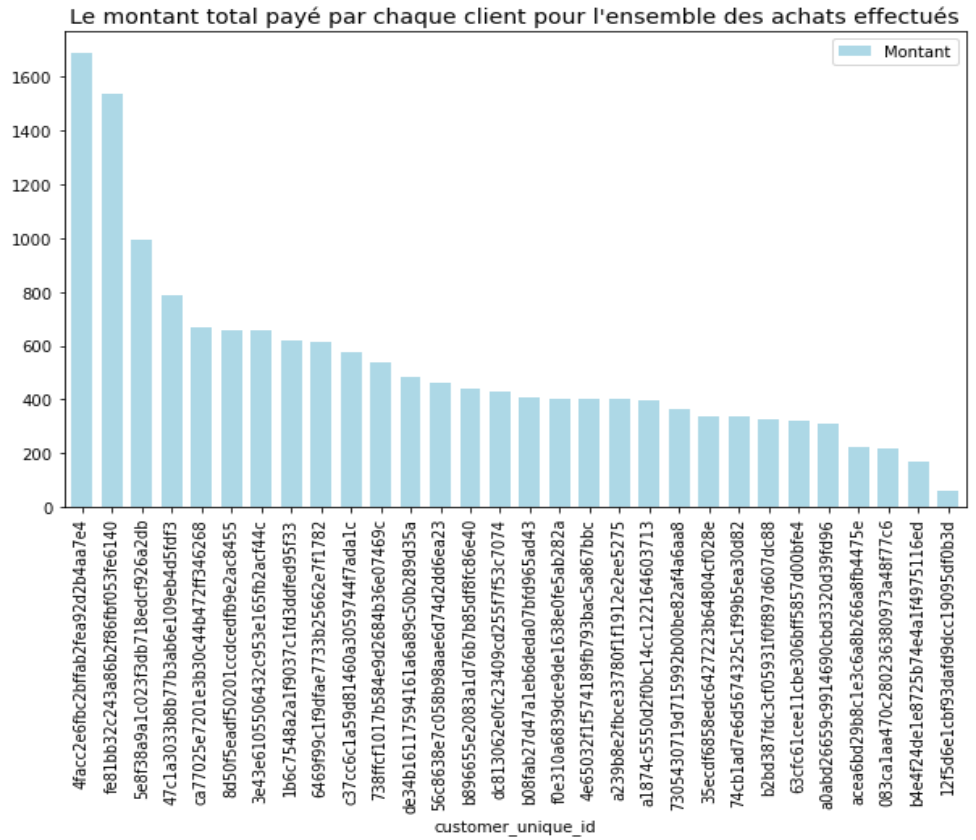


- **Data la plus ancienne:** 2016-10-04 09:43:32
- **Date la plus récente:** 2018-09-03 17:40:06

II/ ANALYSE EXPLORATOIRE & FEATURE ENGINEERING
FEATURE ENGINEERING: SEGMENTATION RFM

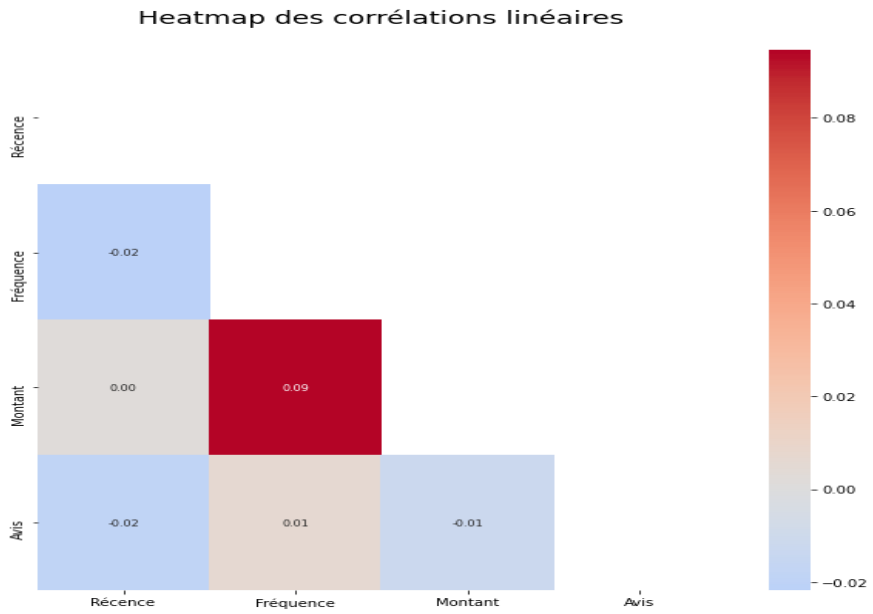


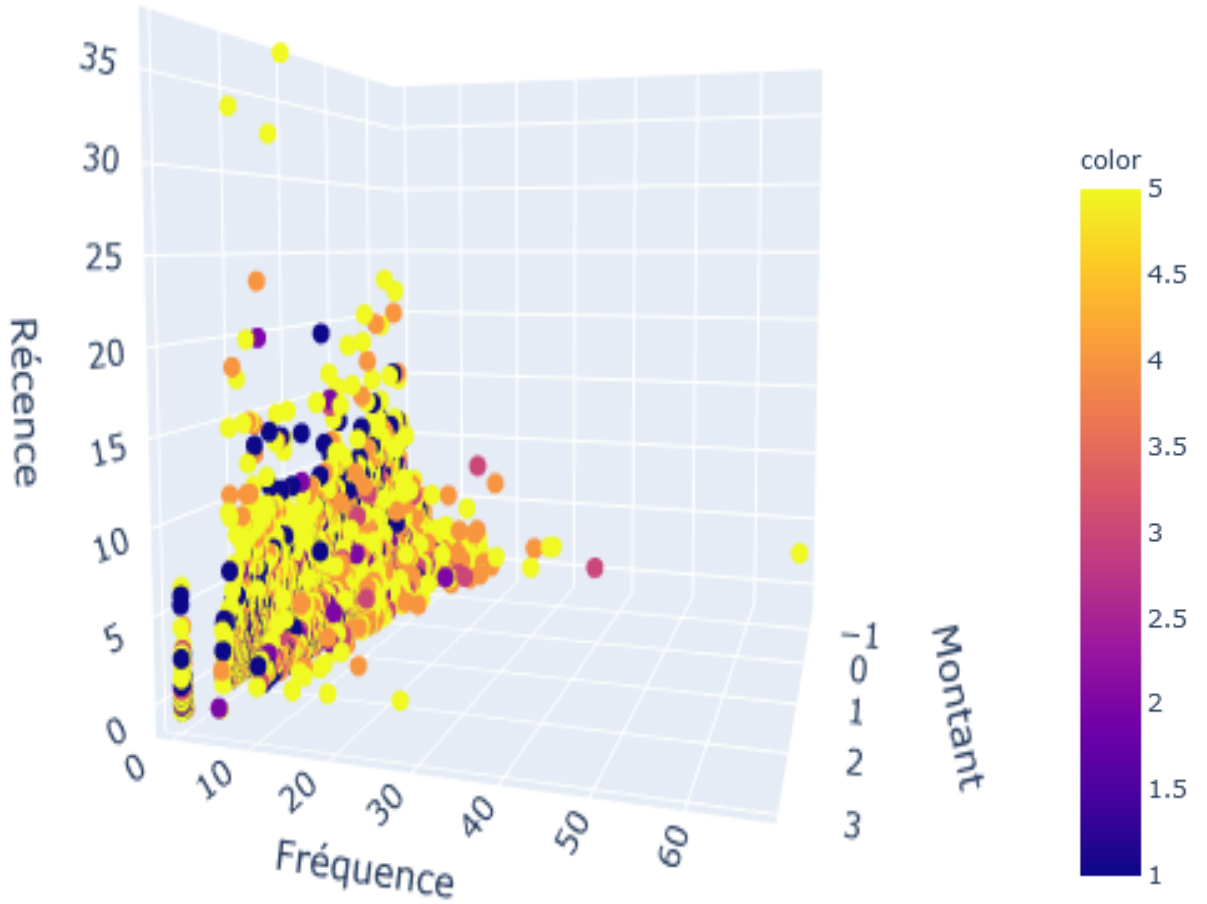




II. MODÉLISATION : CLUSTERING

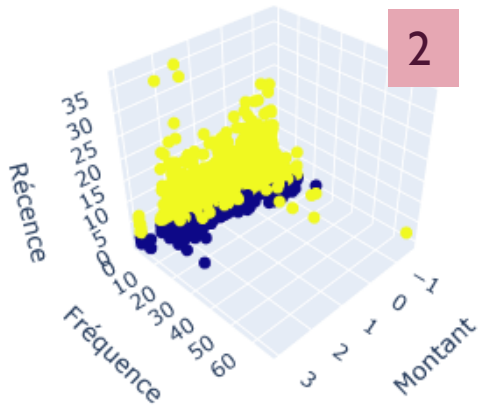
- **Base de données RFM + Avis**
- **Trois méthodes de clustering:**
  - **K-MEANS**
  - **DBSCAN**
  - **HIERARCHIQUE**

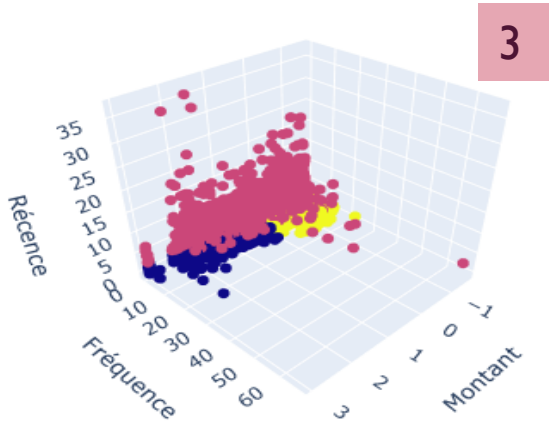


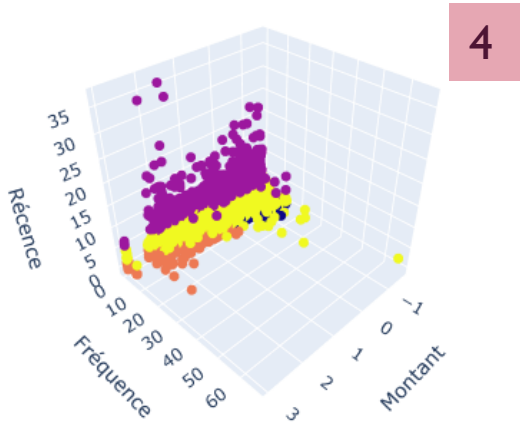


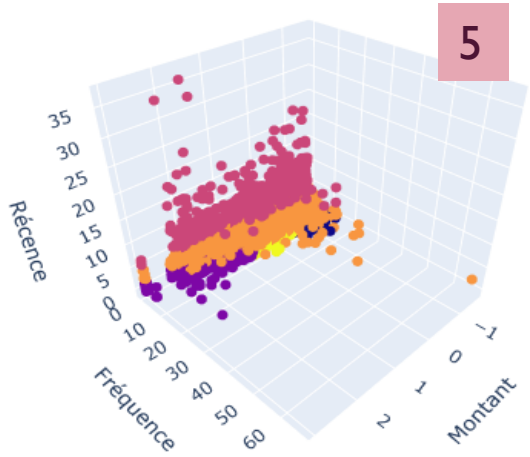
Affichage 3D des RFM + les avis

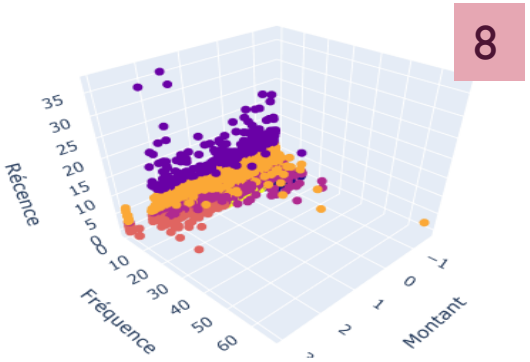
III. MODÉLISATION : K-MEANS
K-MEAN : RFM

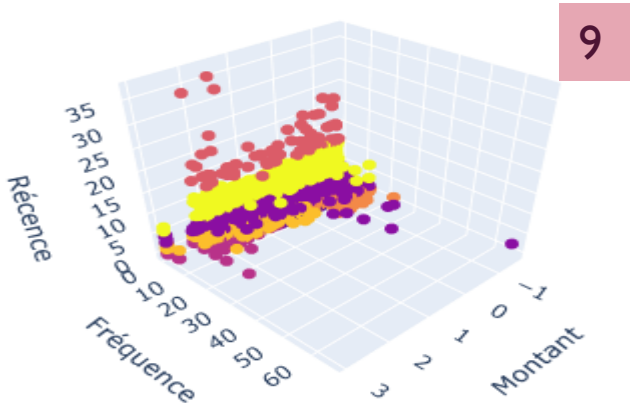


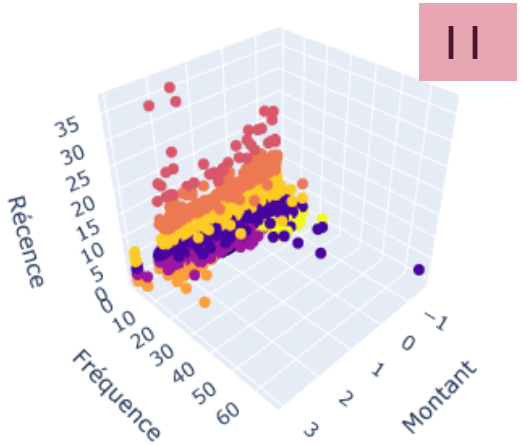








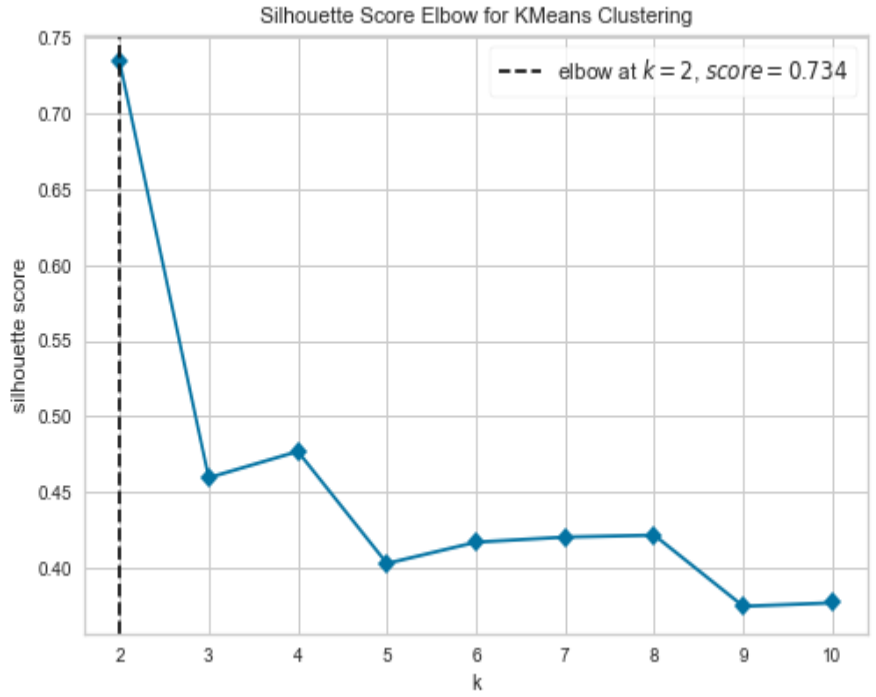


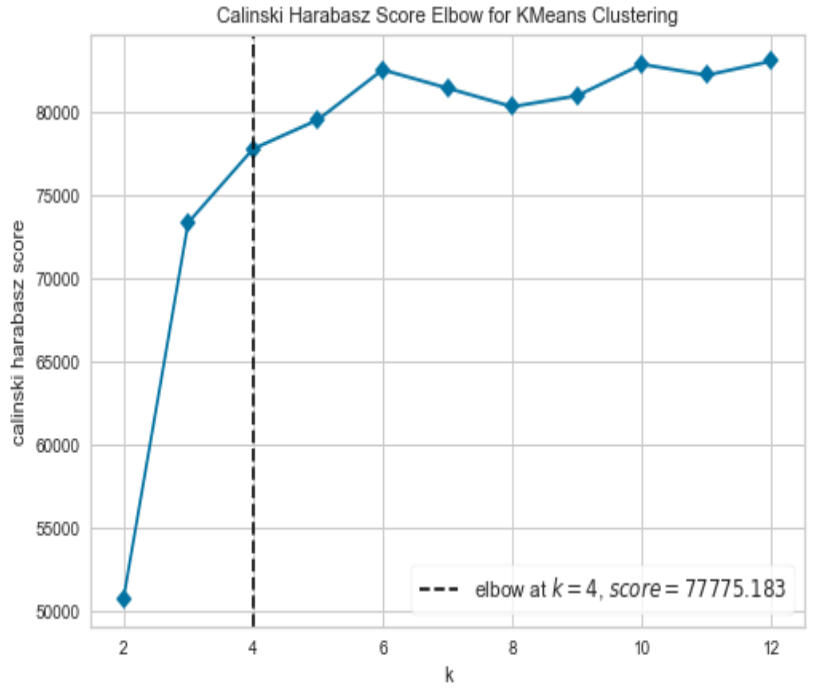


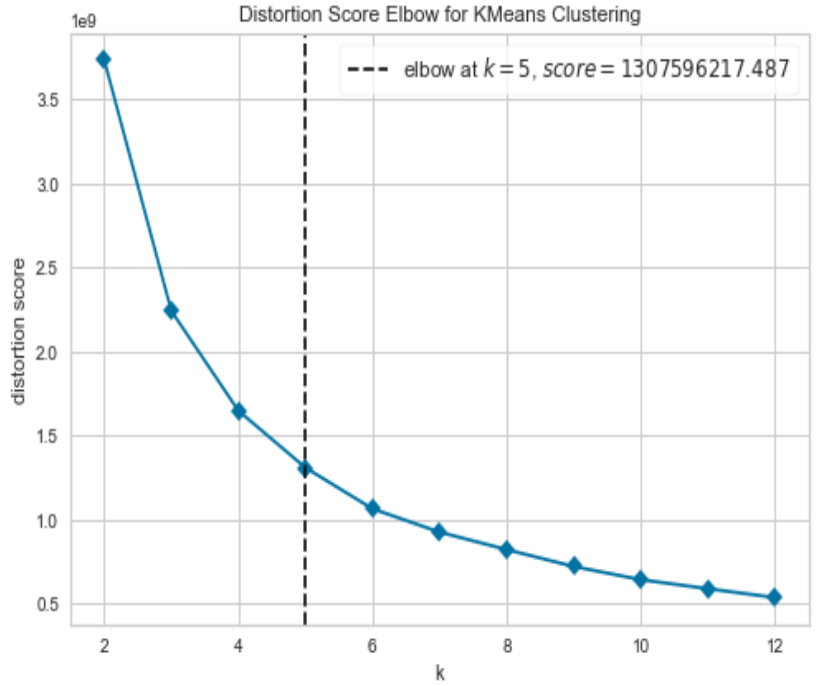
Affichage 3D des RFM en fonction du nombre de clusters

- **2 à 11 clusters**

III. MODÉLISATION : K-MEANS
RFM : K-MEAN - MÉTRIQUES

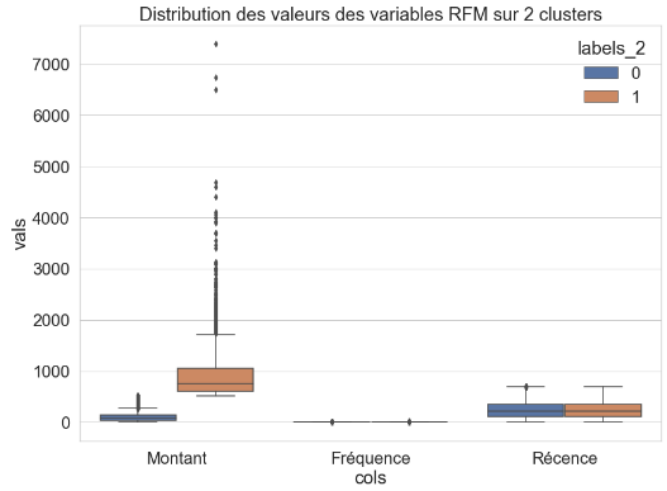


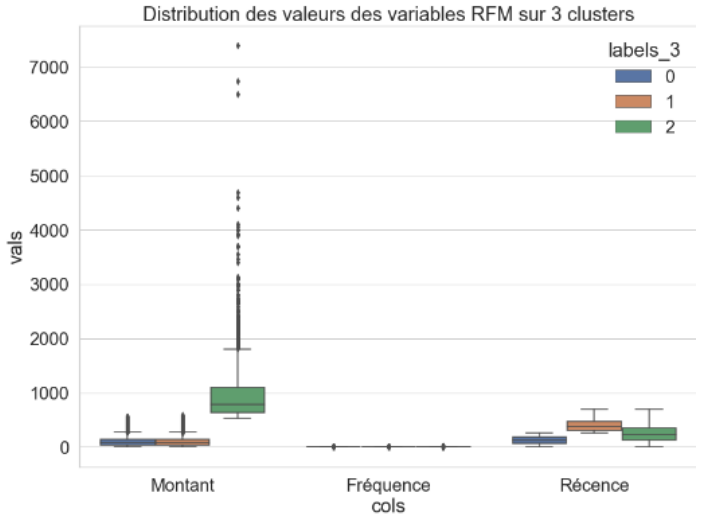


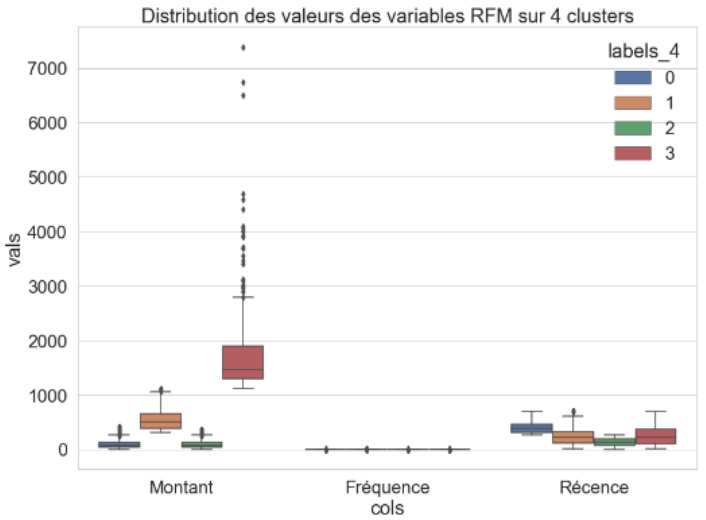


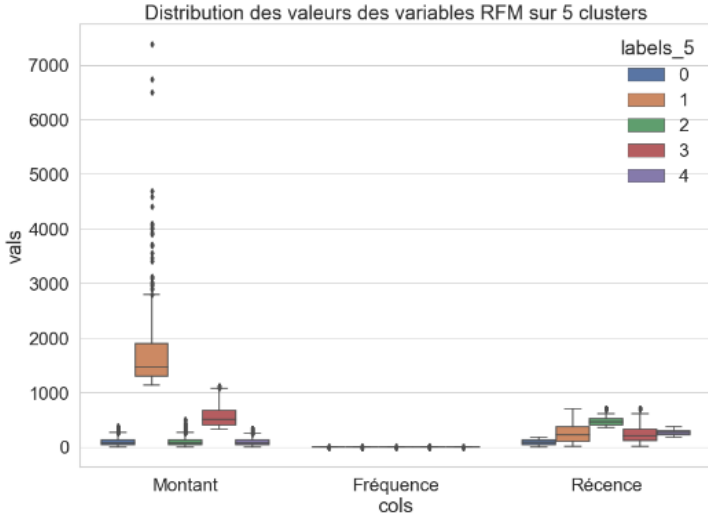
➢ **Les scores des métriques suggèrent un k compris entre 2 et 5**

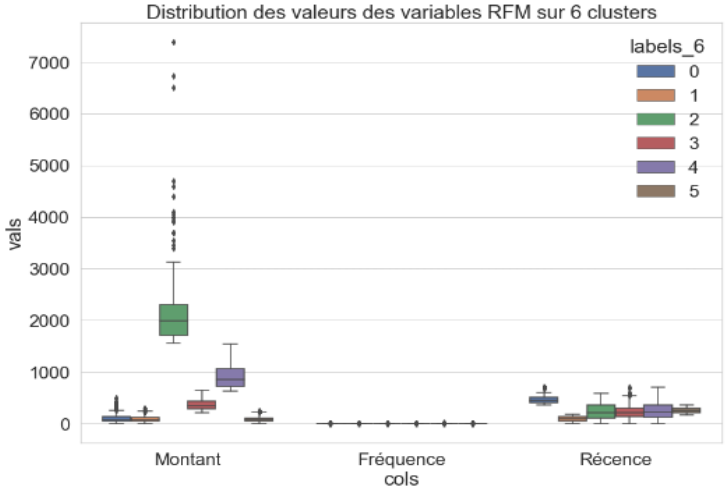
III. MODÉLISATION : K-MEANS
RFM: K-MEAN - MÉTRIQUES





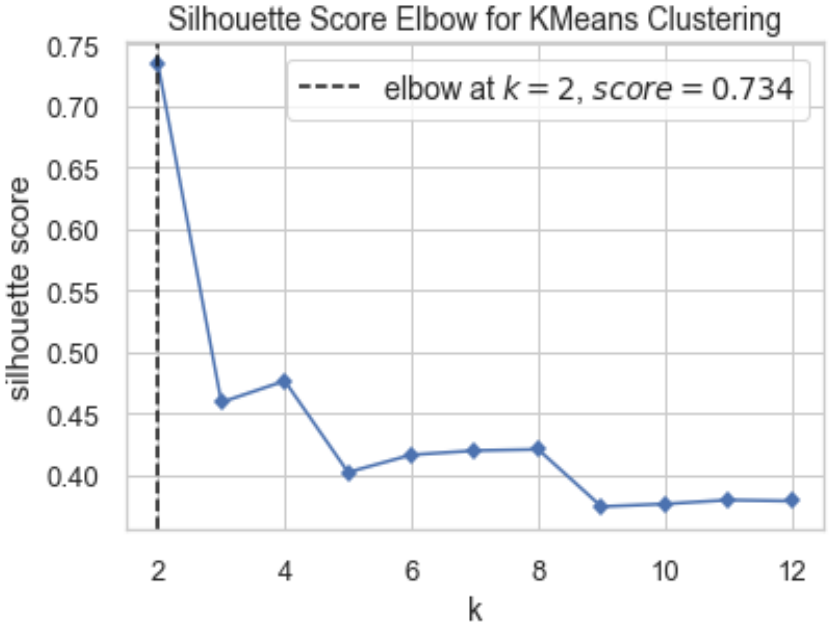


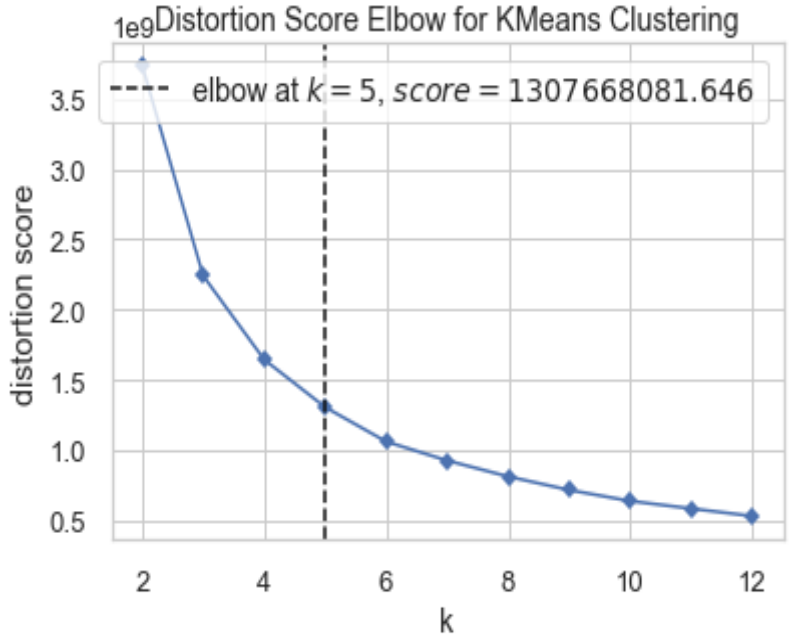


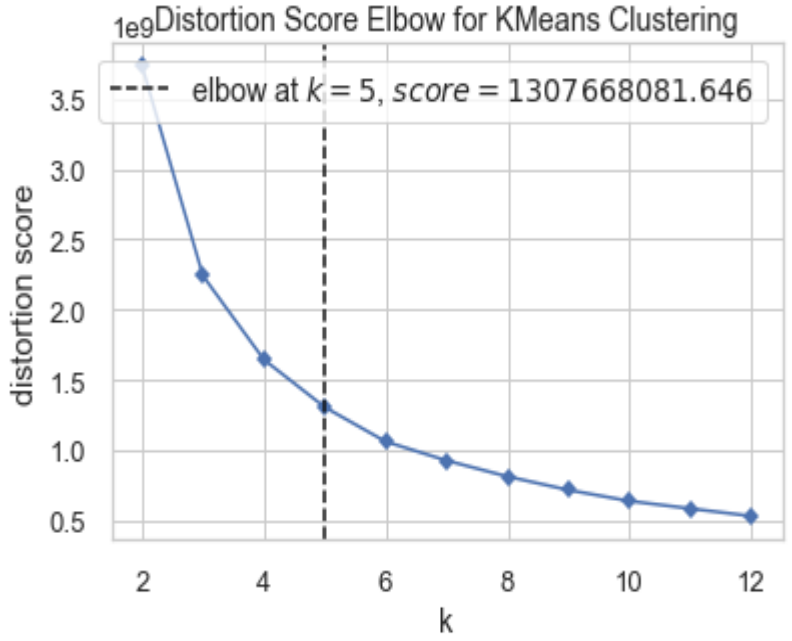


- Fréquences: variabilité faible
- Prendre en compte les clusters k=4 et k=5

III. MODÉLISATION : K-MEANS
RFM+AVIS :K-MEAN - MÉTRIQUES

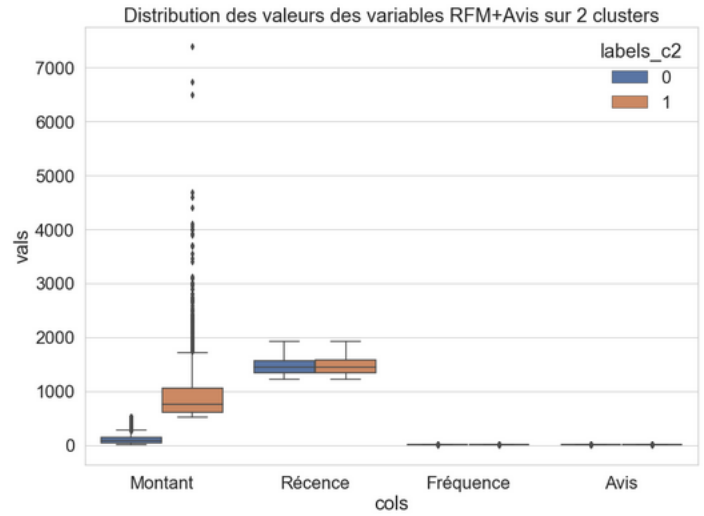


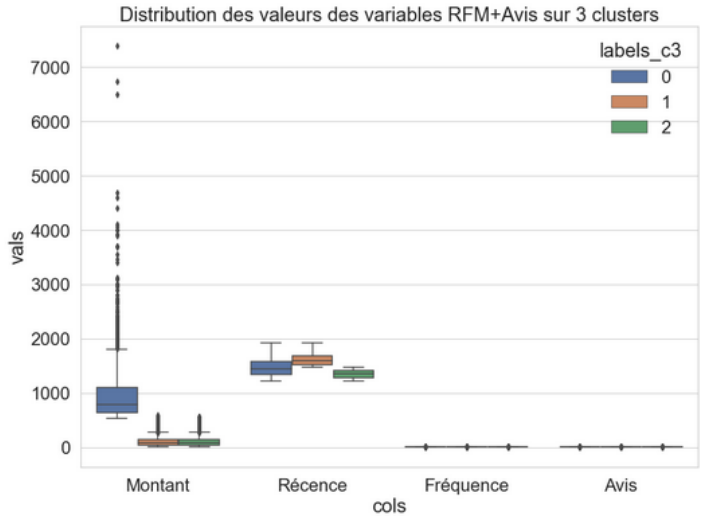


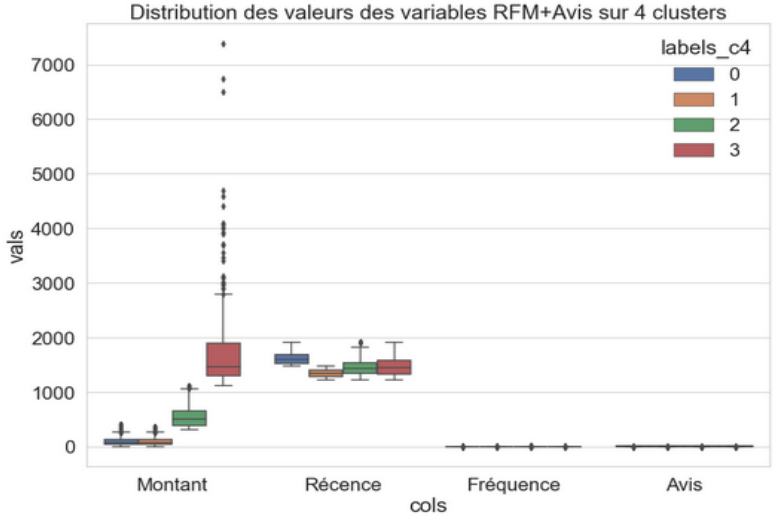


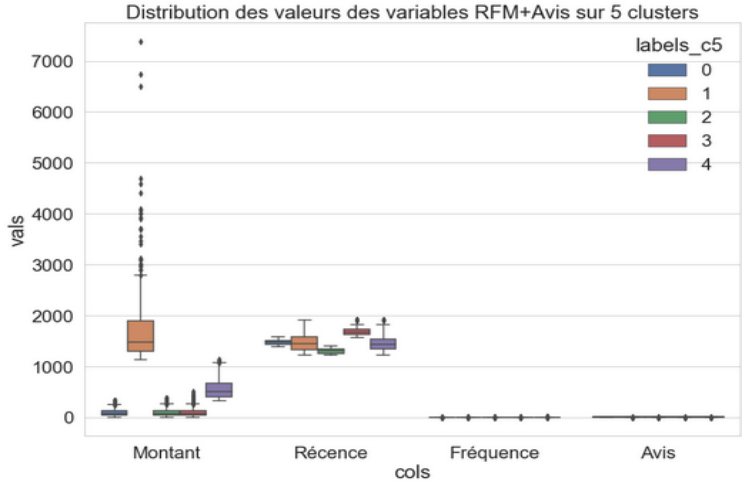
➢ **Les scores des métriques suggèrent un k compris entre 2 et 5**

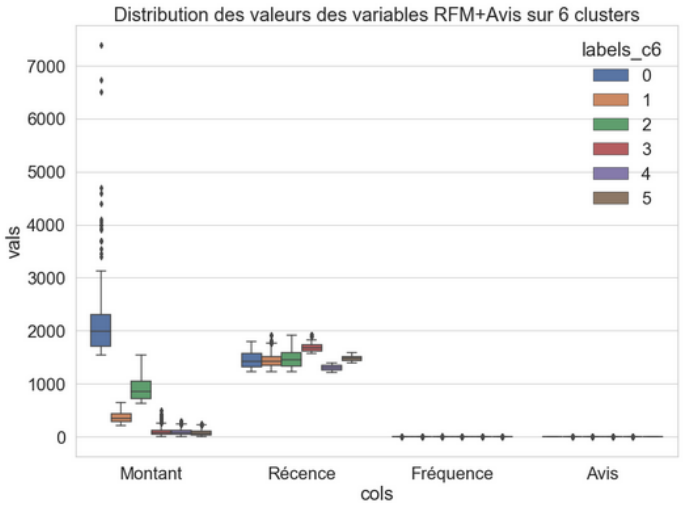
III. MODÉLISATION : K-MEANS
RFM+AVIS: K-MEAN - MÉTRIQUES





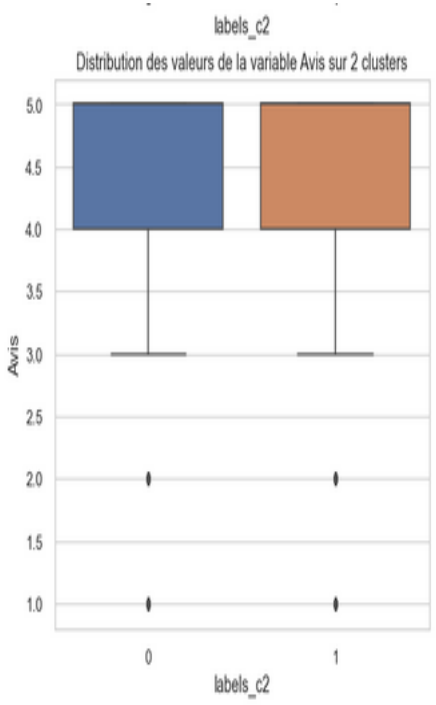


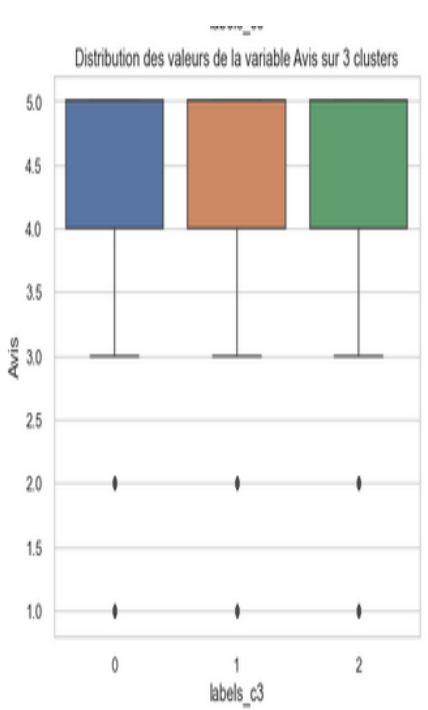


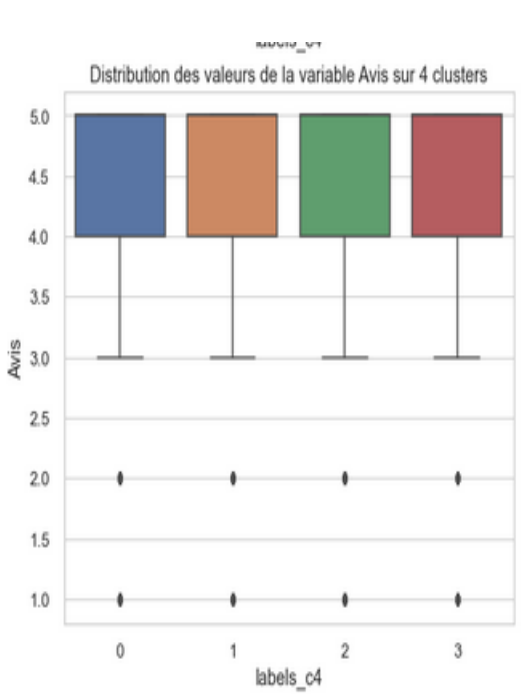


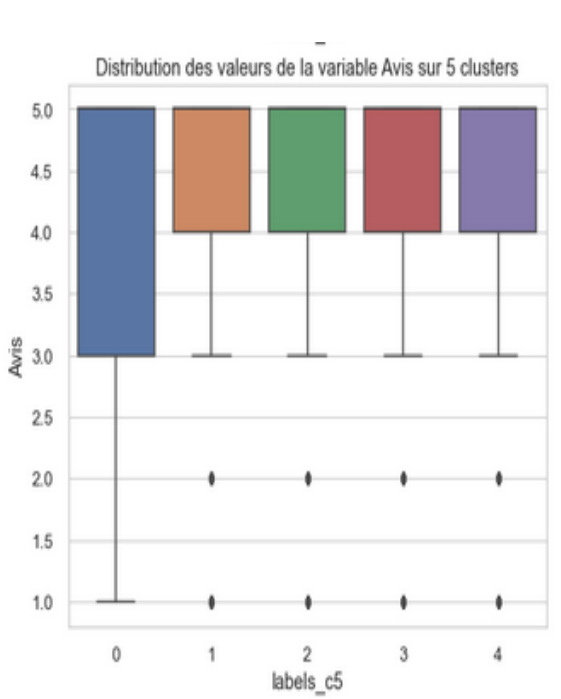
- Fréquences: variabilité faible
- Prendre en compte les clusters k=4 et k=5

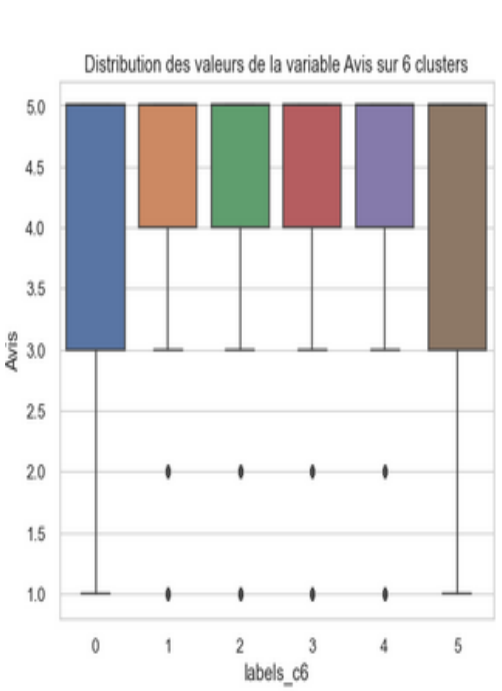
III. MODÉLISATION : K-MEANS
RFM+AVIS: K-MEAN - METRICS





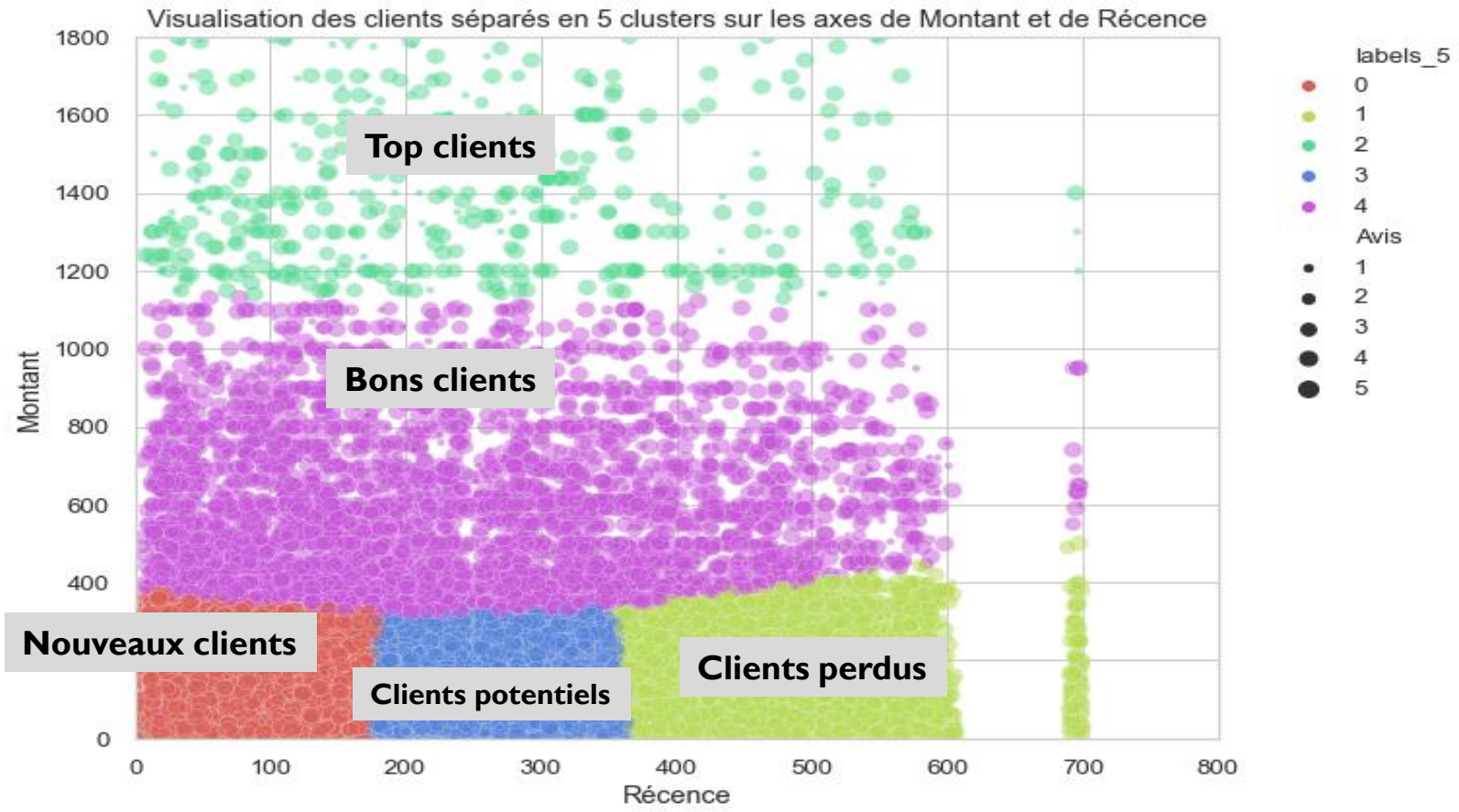


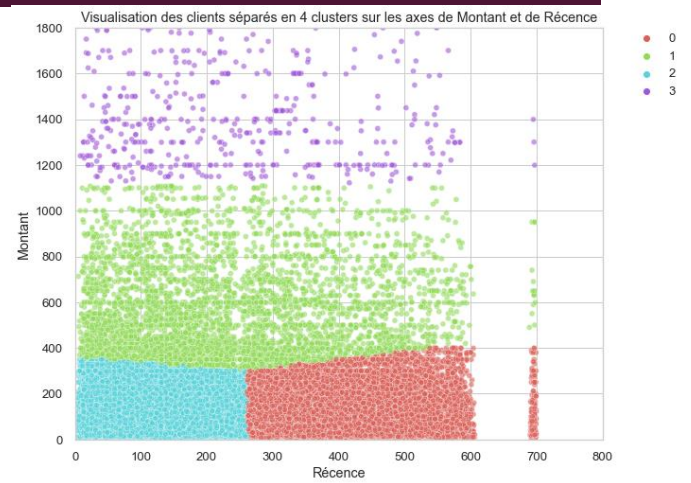




- Seul k=5 permet de mieux faire une différence entre les clusters

III. MODÉLISATION : K-MEANS RFM+AVIS: VISUALISATION



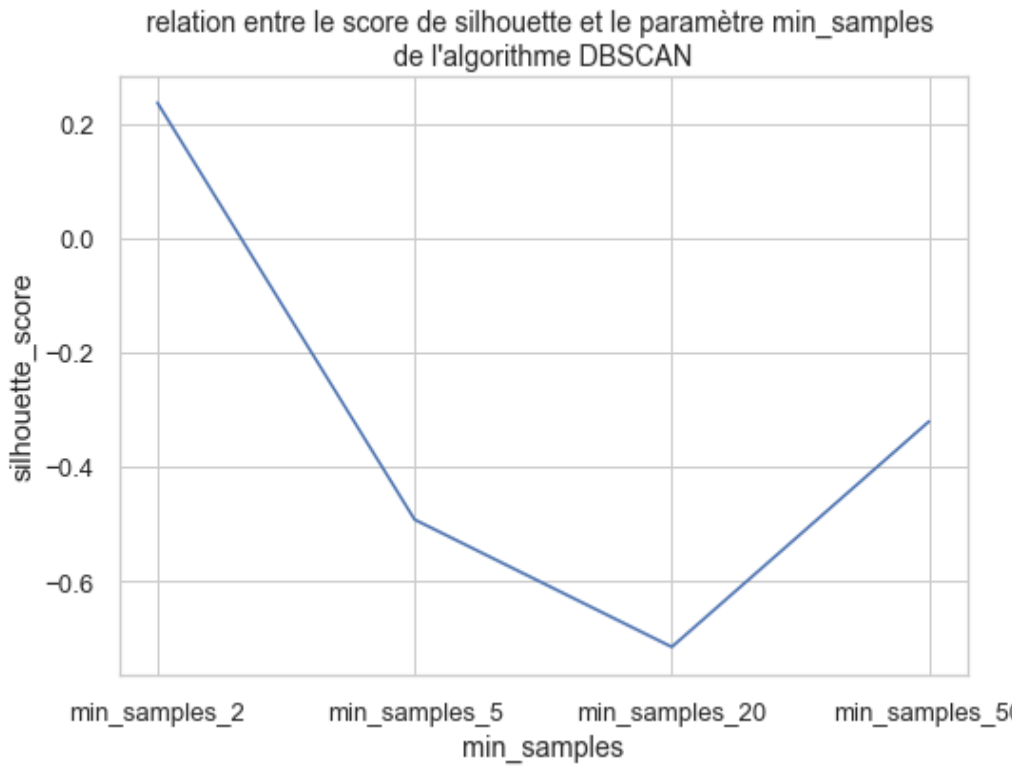


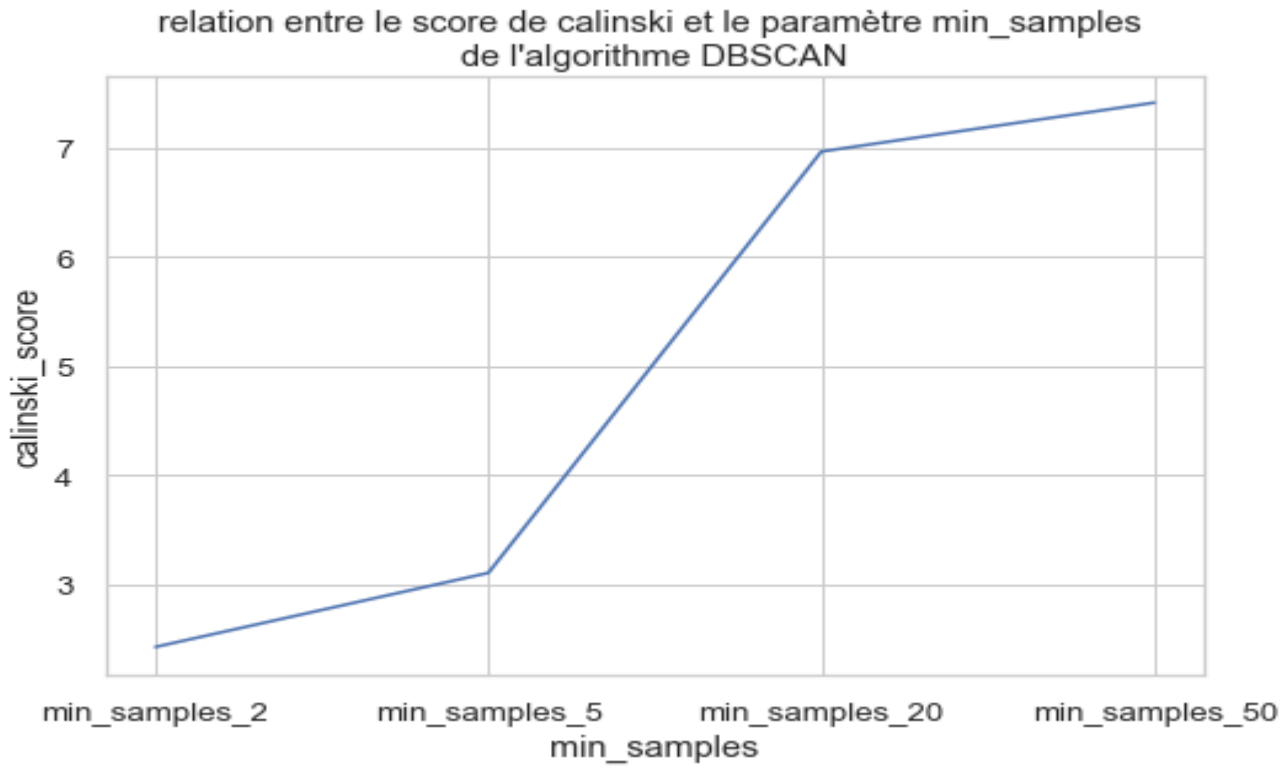
	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	total_clients
k2	90401	3003	NaN	NaN	NaN	NaN	93404
k3	38521	52061	2822.0	NaN	NaN	NaN	93404
k3	38521	52061	2822.0	NaN	NaN	NaN	93404
k4	37319	5146	50272.0	667.0	NaN	NaN	93404
k5	20726	34083	664.0	33022.0	4909.0	NaN	93404
k6	31259	1755	32630.0	20312.0	307.0	7141.0	93404

Nombres de clients par clusters

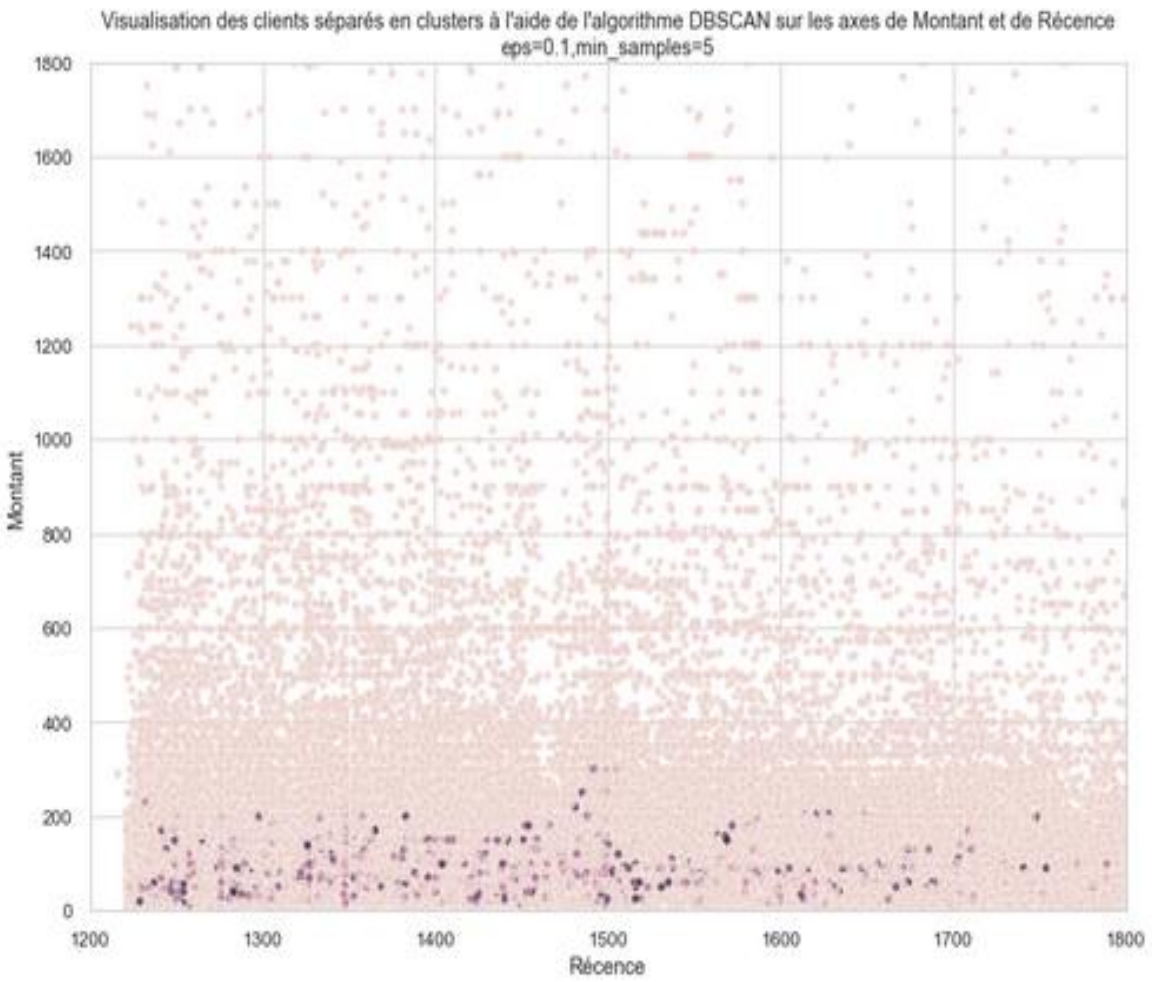
- Seul k=5 permet de mieux faire une différence entre les clusters

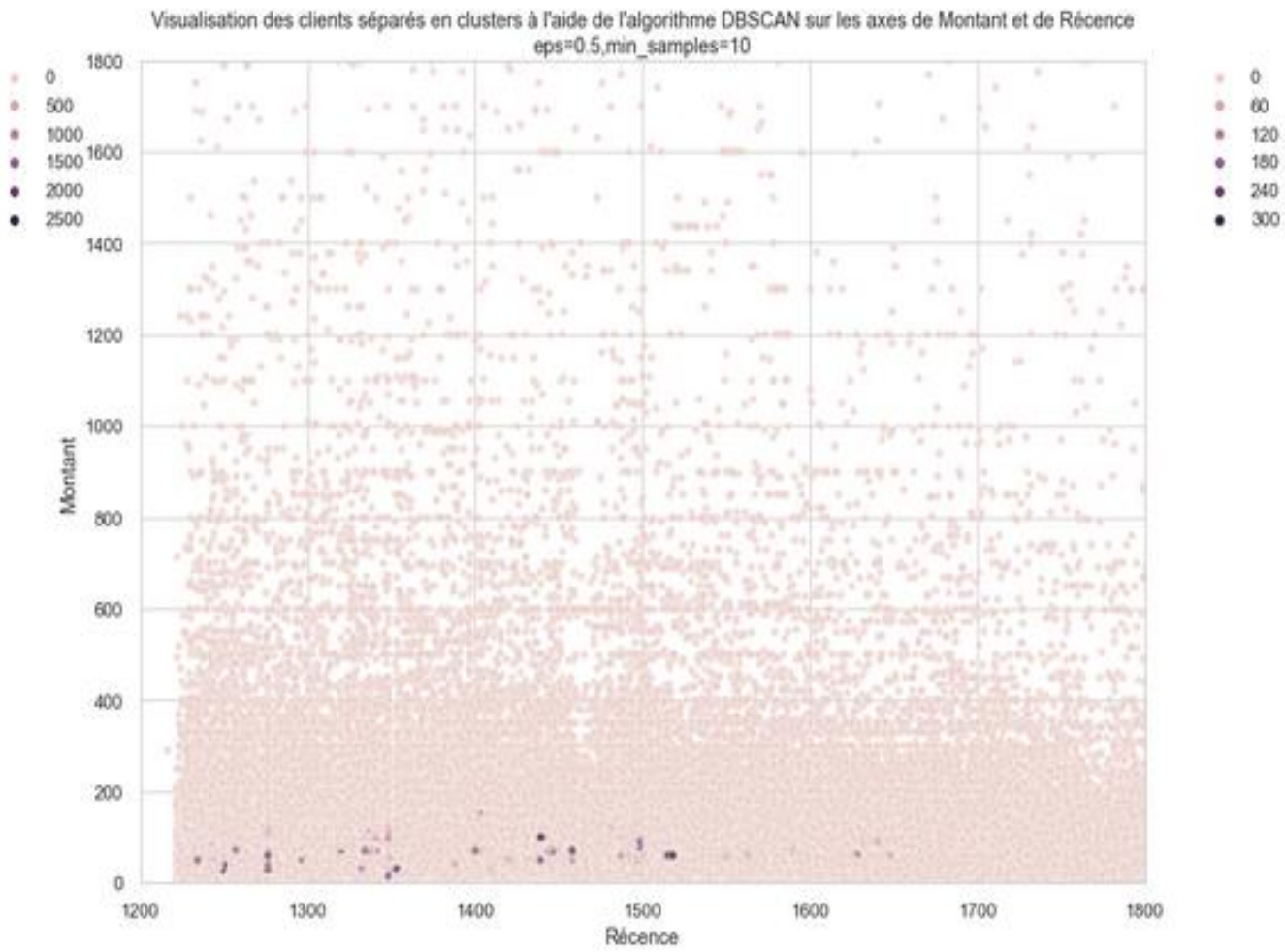
III. MODÉLISATION : DBSCAN
RFM: MODIFICATIONS DES HYPERPARAMÈTRES EPS ET MIN_S



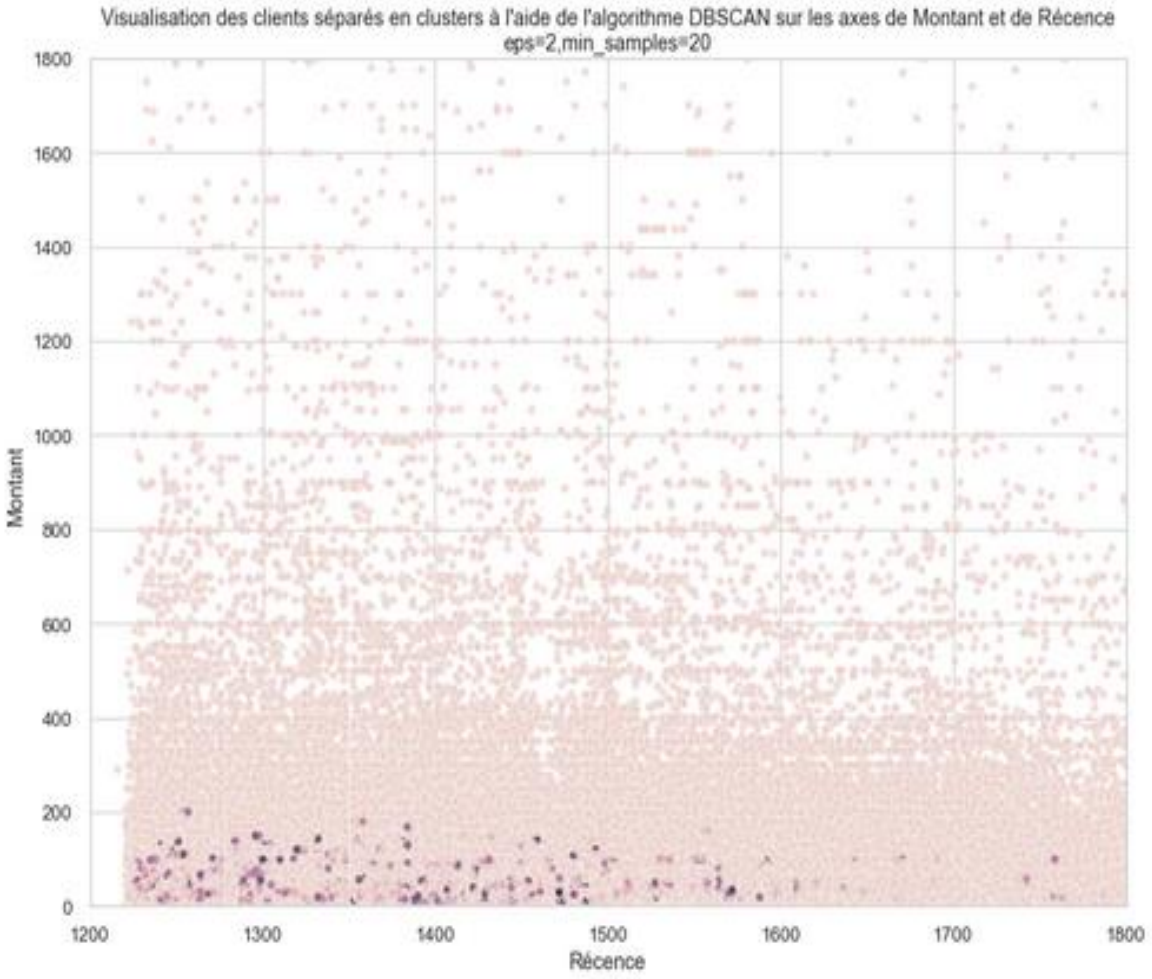


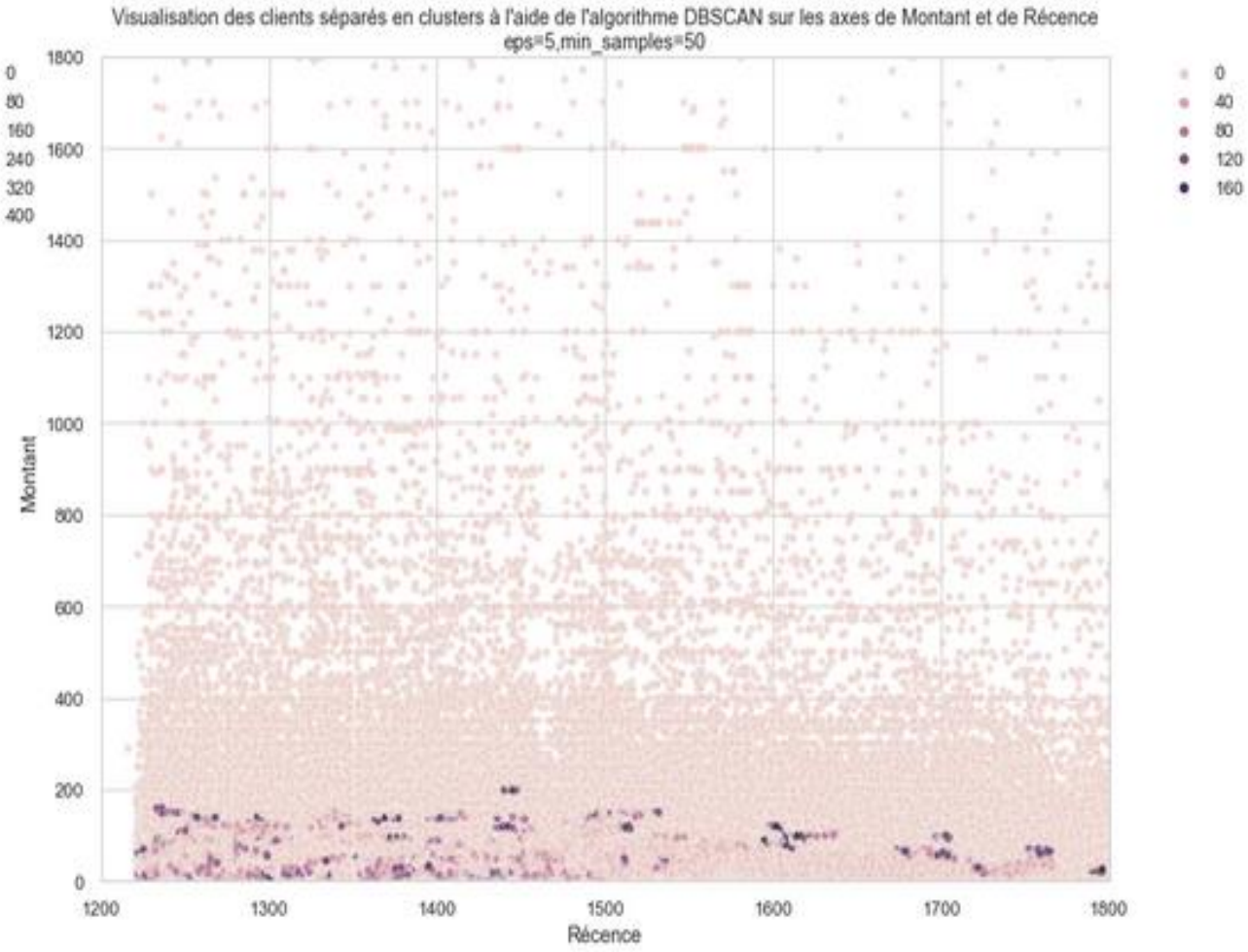
III. MODÉLISATION : DBSCAN
RFM : VISUALISATION



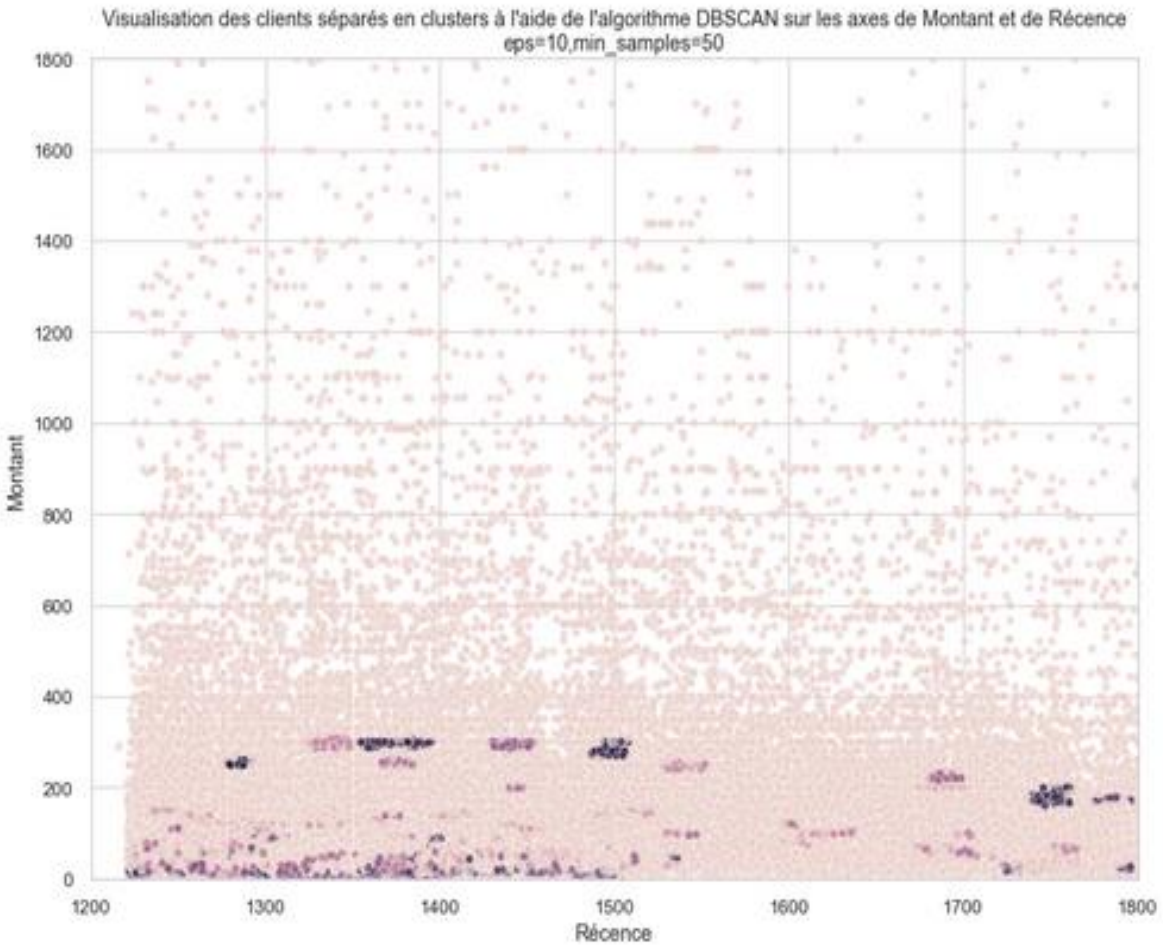


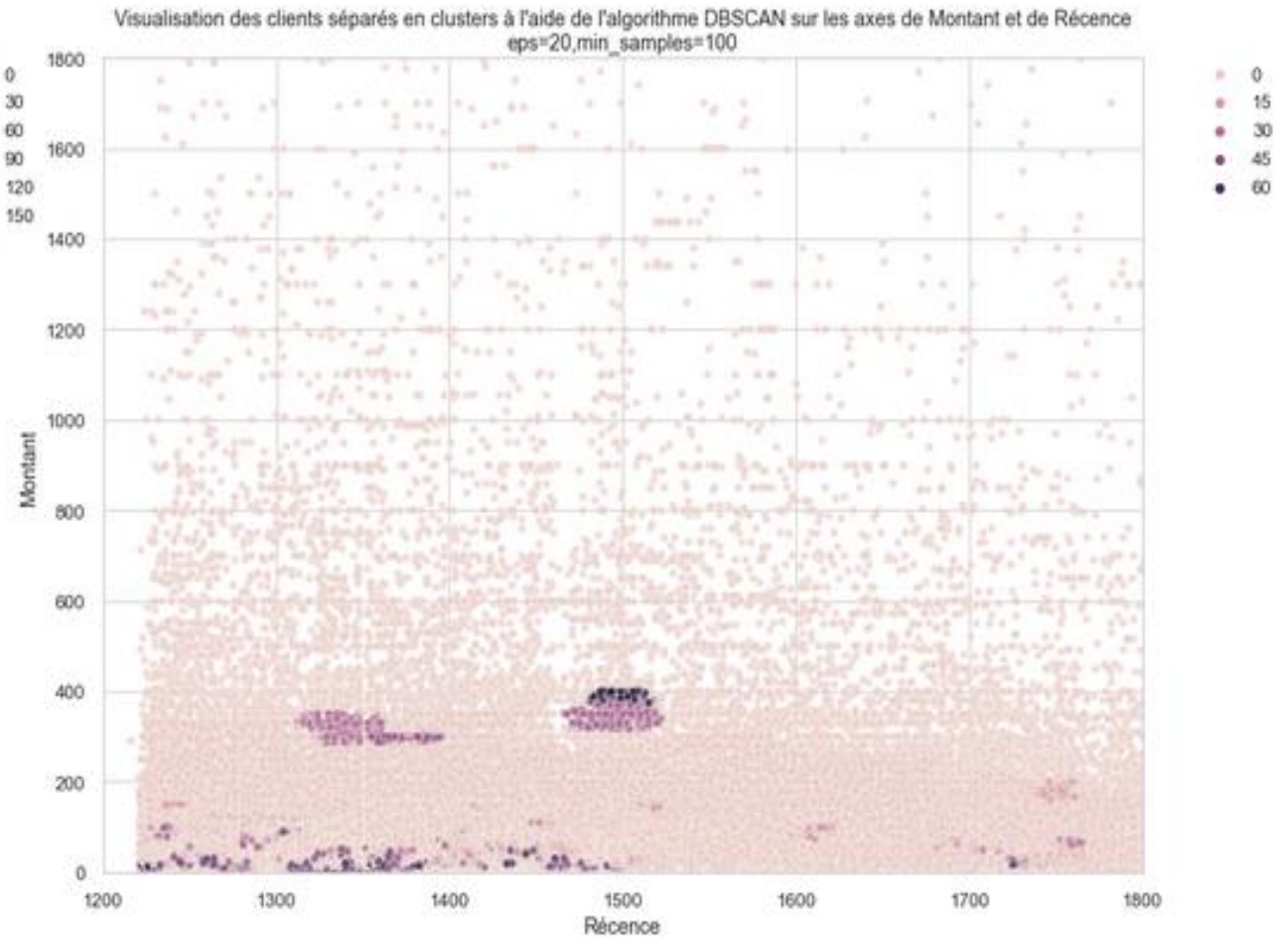
III. MODÉLISATION : DBSCAN
RFM : VISUALISATION





III. MODÉLISATION : DBSCAN
RFM : VISUALISATION

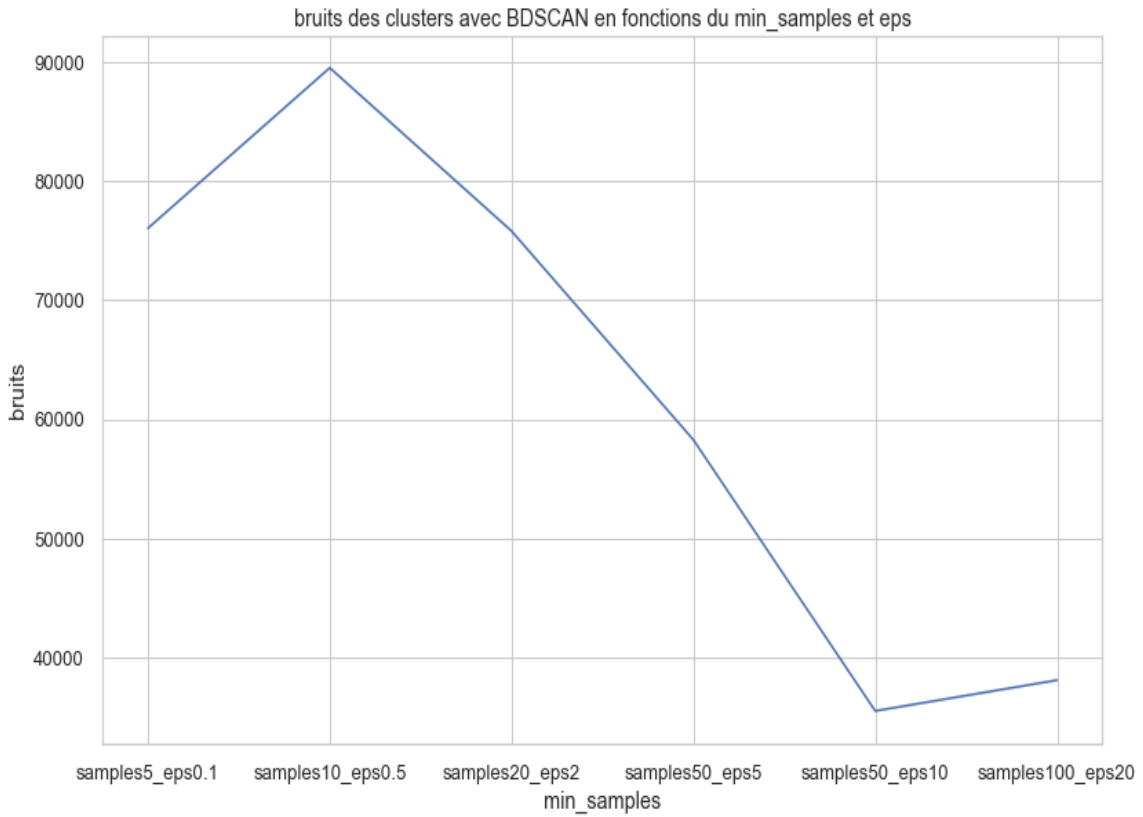




III. MODÉLISATION : DBSCAN RFM:VISUALISATION

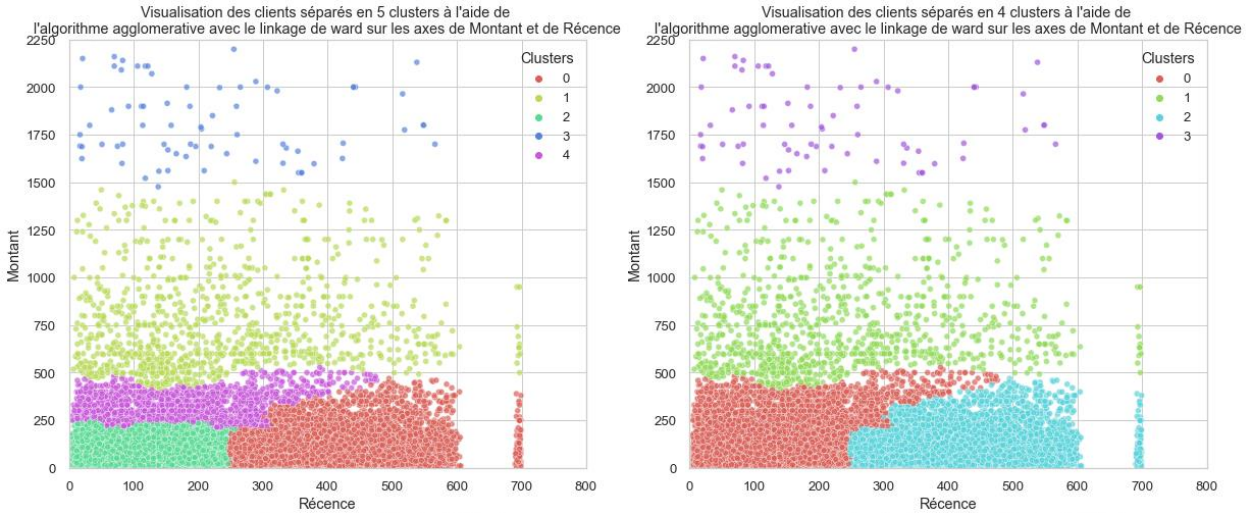
	DBSCAN1	DBSCAN2	DBSCAN3	DBSCAN4	DBSCAN5	DBSCAN6	Non_clustered
N1	17388.0	NaN	NaN	NaN	NaN	NaN	75993
N2	NaN	4041.0	NaN	NaN	NaN	NaN	89340
N3	NaN	NaN	18239.0	NaN	NaN	NaN	75142
N4	NaN	NaN	NaN	34121.0	NaN	NaN	59260
N5	NaN	NaN	NaN	NaN	56828.0	NaN	36553
N6	NaN	NaN	NaN	NaN	NaN	55533.0	37848

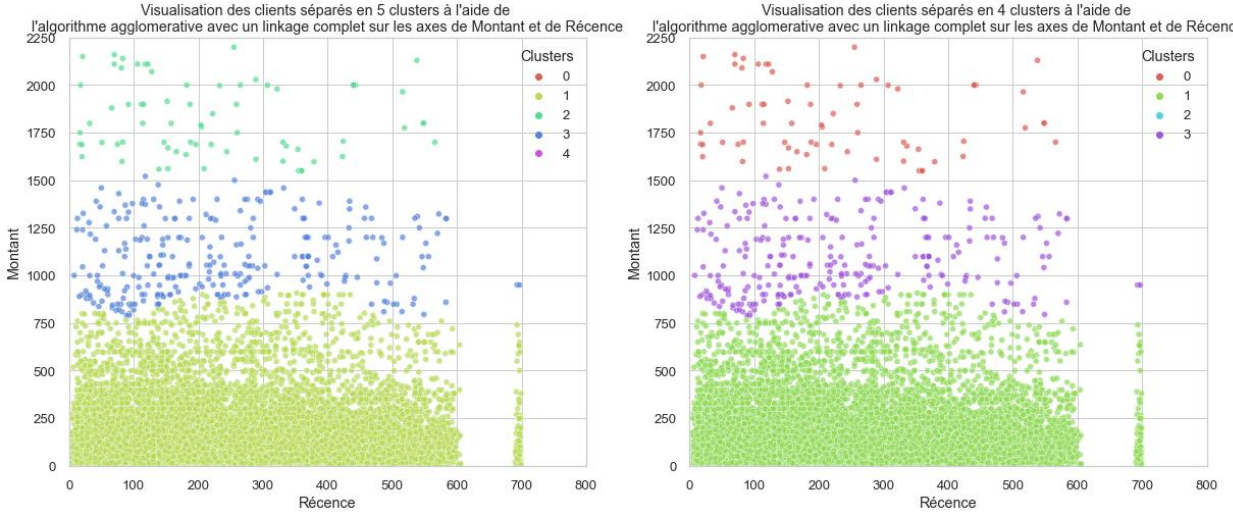
Nombres de clients par clusters



On voit clairement que l'algorithme DBSCAN n'est pas adapté à notre problème de segmentation puisqu'il permet d'exclure beaucoup de clients et le bruit est trop élevé.

III. MODÉLISATION : HIÉRARCHIQUES AGGLOMÉRATIF RFM:VISUALISATION





	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	total_clients
c4complete	27183	43	2768	6	NaN	30000
c5complete	2768	534	26649	6	43.0	30000
c4ward	24457	2373	2959	211	NaN	30000
c5ward	15518	2373	2959	211	8939.0	30000

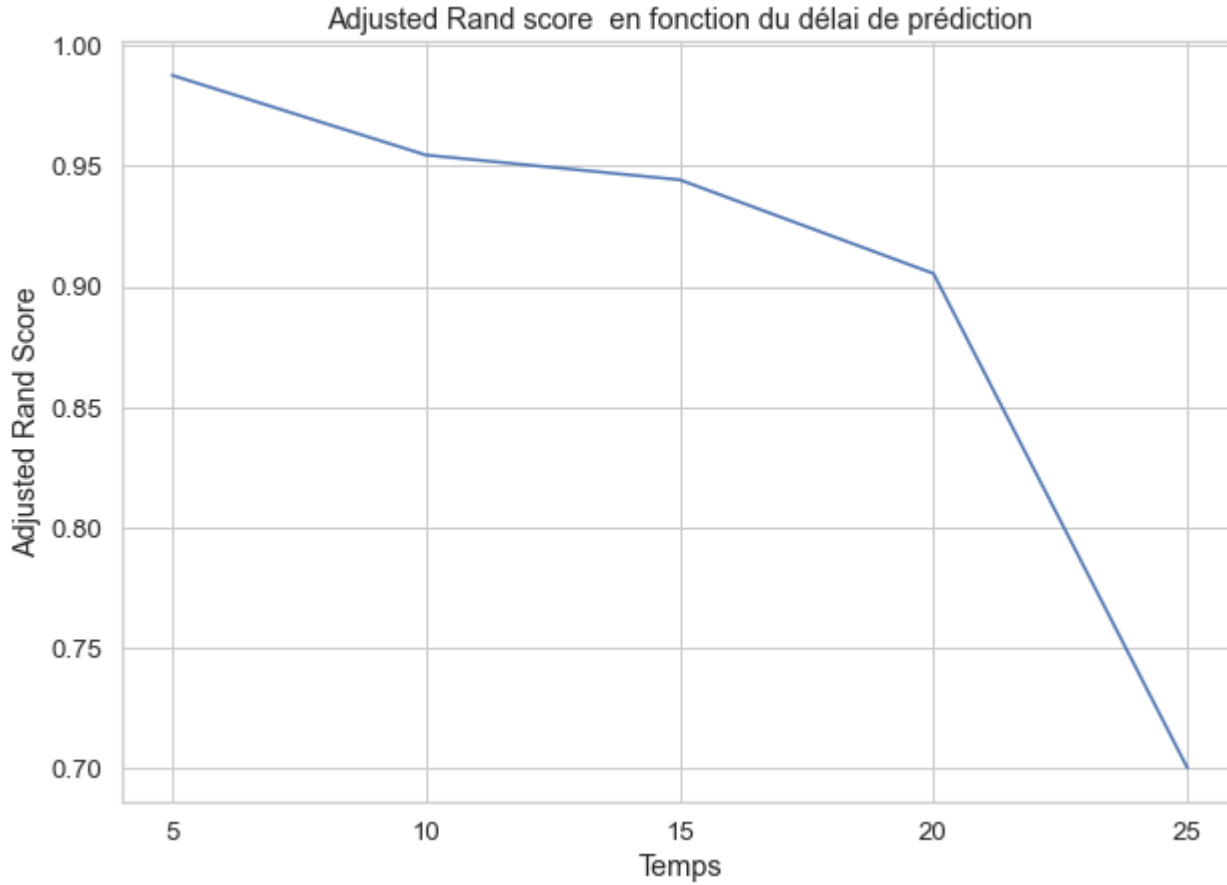
Nombres de clients par clusters

- Presque similaires aux résultats du K-Means.
- Sauf que nous avons obtenu des nombres de clients très déséquilibrés entre les clusters avec l'algorithme hiérarchique.
- Le complete-linkage clustering a concentré environ 99% des clients dans un seul cluster, tandis qu'avec la méthode de ward on trouve une meilleure dispersion des clients dans les clusters par rapport au complete-linkage clustering.

Les résultats de clustering obtenus avec le K-Means sont meilleurs que ceux obtenus avec le DBSCAN et les algorithmes hiérarchiques.

IV. CONTRAT DE MAINTENANCE
RFM: ARI

- La date d'achat la plus récente 2018-09-03 correspond à un nombre de jours écoulé = 0 jours et la date d'achat la plus ancienne 2016-10-04 correspond à un nombre de jours écoulé = 699 jours
- Séparation des données en « Baseline »: Création d'un fichier initial sans les 45 derniers jours.
- Injection de données par jours à partir du fichier initial
- Calcul d'indice de rand ajusté entres les clusters de Baseline et nouvelle segmentation
- Calcul approximatifs: seulement la dernière commande de chaque client est prise en compte



- **Conclusion**: le score ARI commence à devenir instable avec une tendance à la baisse à partir du 20 ième jours. Dans l'idéal, la maintenance de l'algorithme devrait être faite donc tous les 22 jours.

CONCLUSION

- Création d'une segmentation à l'aide d'algorithme de Machine Learning non supervisée à des fins d'utilisation en marketing: le k-means plus adapté dans notre étude.
- La segmentation peut être améliorer par:
  - Incorporation de nouvelles données vis-à-vis des promotions par exemple
  - Création d'une segmentation par type de produit
  - Analyse de l'évolution des segments en fonction du temps
- Concernant le contrat de maintenance : faut il le maintenir en fonction de l'ensemble des produits, ou en fonction du client…