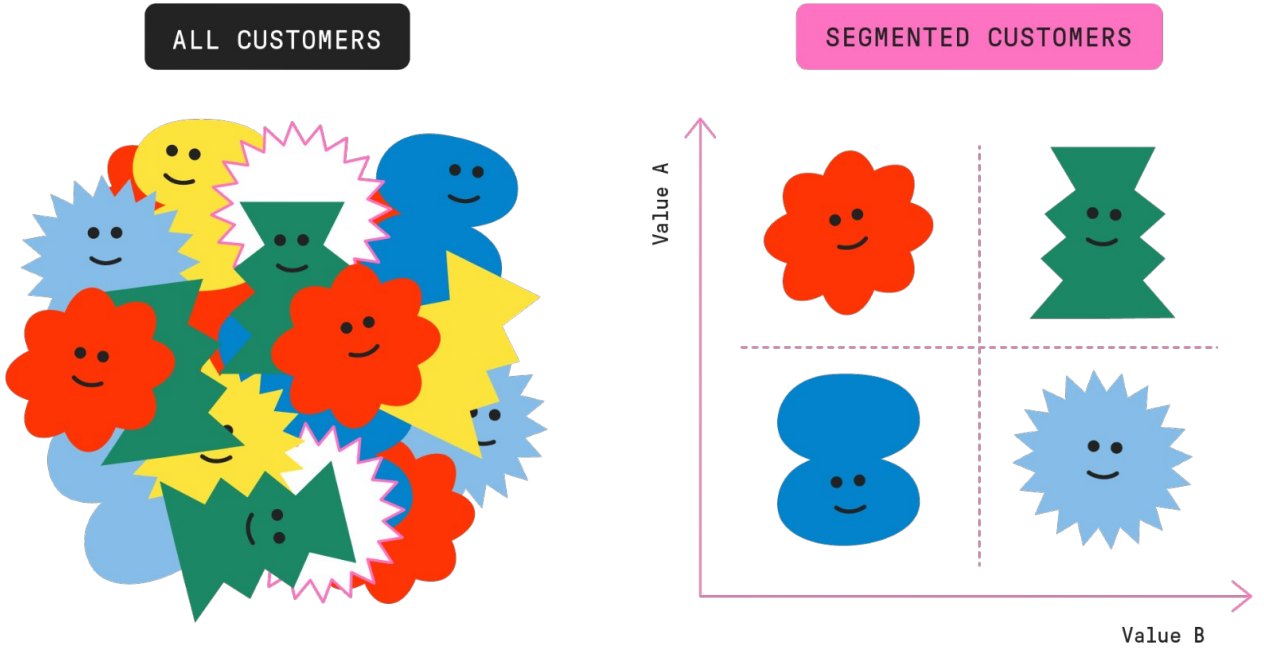
# P5 : Segmentez des clients d’un site e-commerce

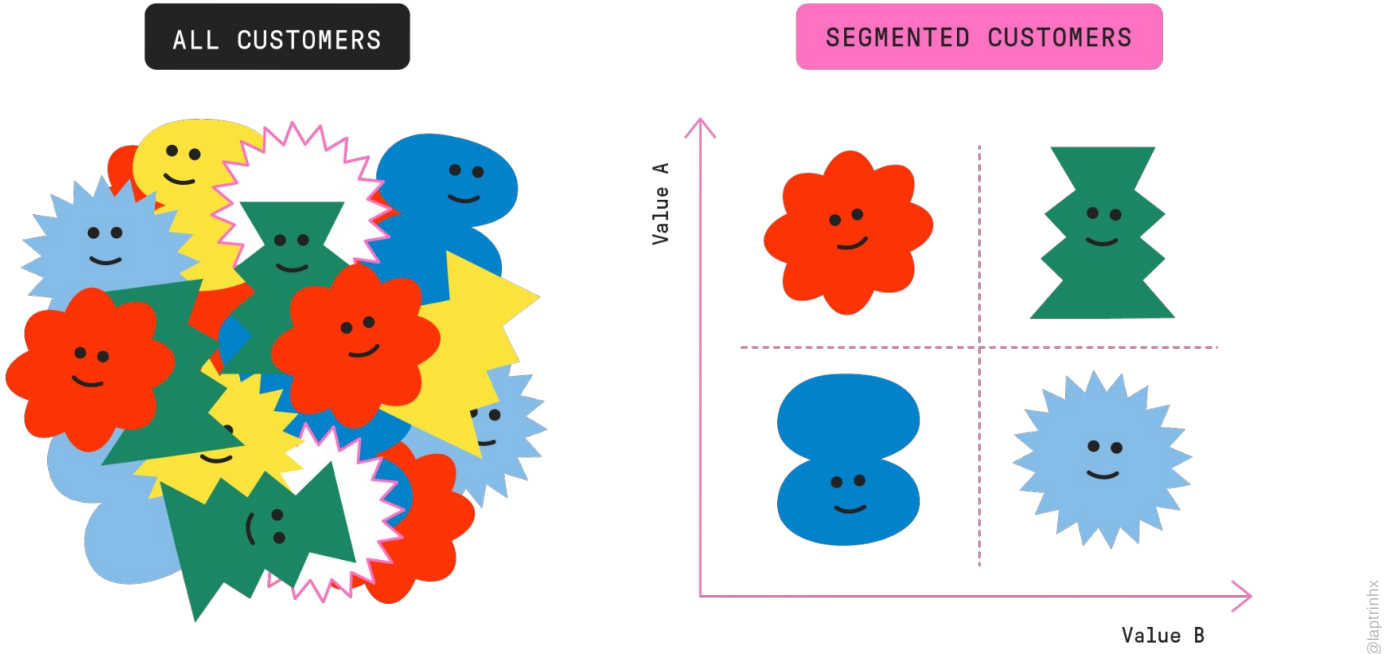
*Defense*

Formation : Data Scientist
Etudiante : Rocio Isorna
Evaluatrice : Souhail Toumdi

# *Client segmentation*



# *Client segmentation*




*Comprendre les différents types des clients*

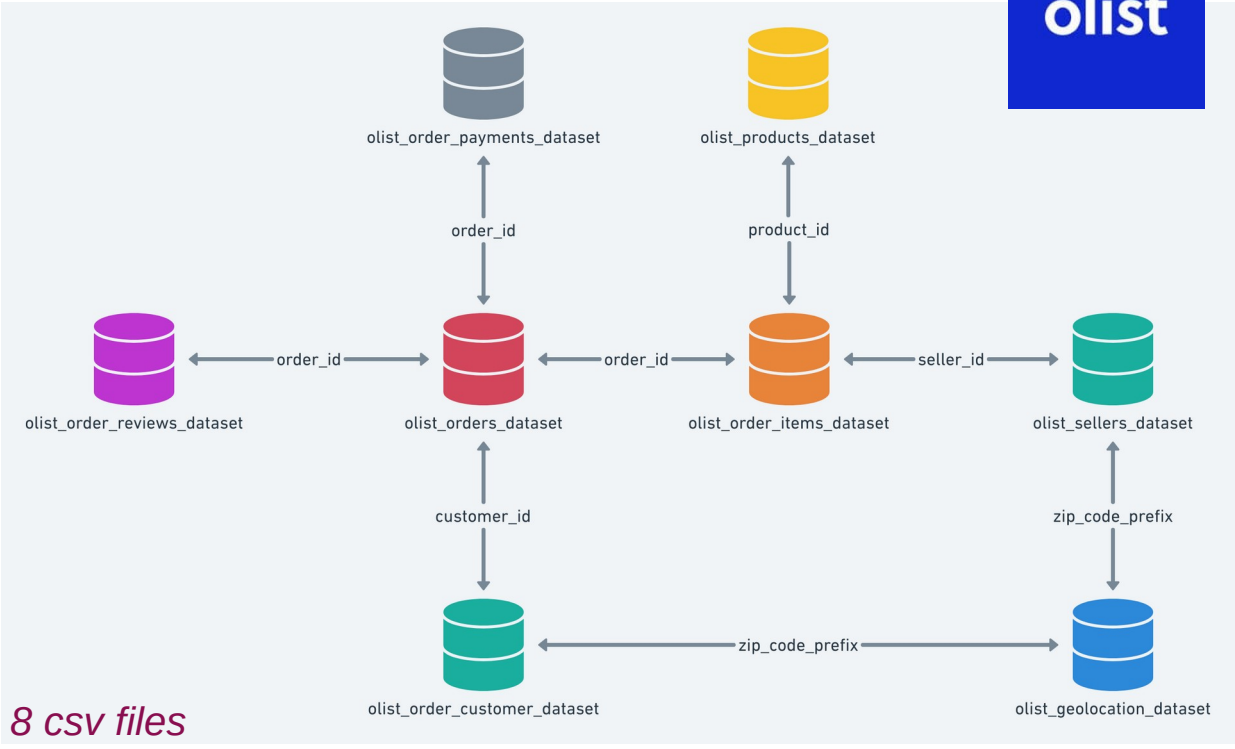
*Fournir à l'équipe marketing une description actionnable*

*Proposition de contrat de maintenance*

# *Segmentation des clients*

## ***Data set :*** *(99441 obs → achat)*

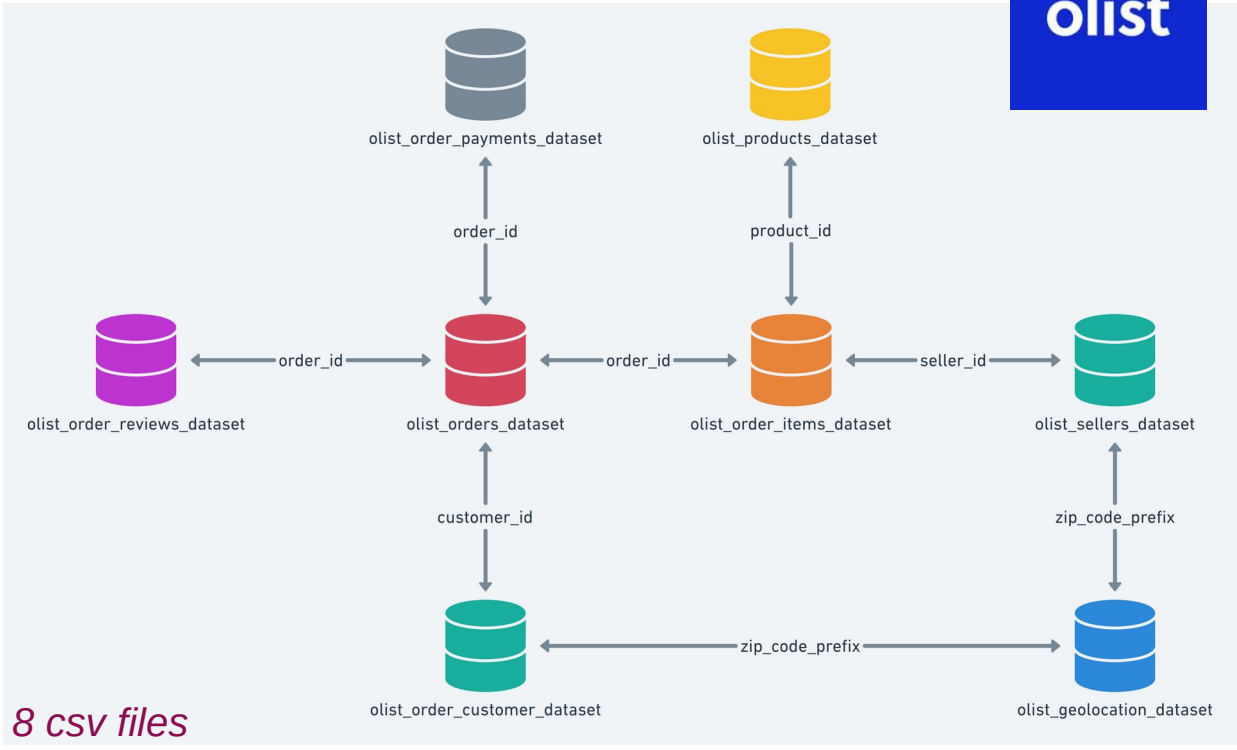
- ***Clients*** : id et zip-code
- ***Localisation***: villes
- ***Items commandé*** : id, item, vendeurs, prix (article et livraison)
- ***Commandes*** : id, statuts, date, livraison
- ***Paiements*** : type de paiement et montant
- ***Reviews*** : retour des clients
- ***Vendeurs*** : information des vendeurs



# *Segmentation des clients*

## ***Data set :*** *(99441 obs → achat)*

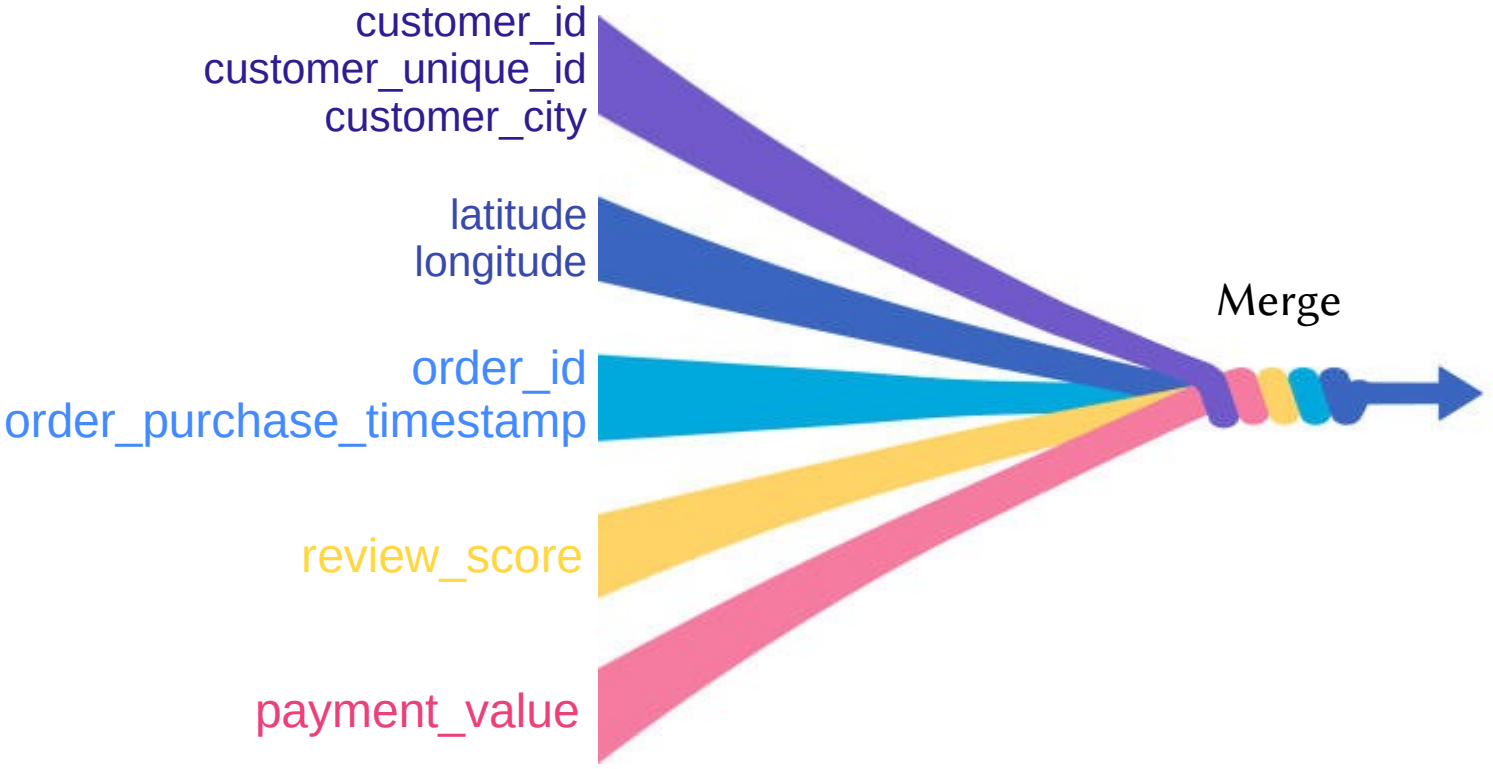
- ***Clients*** : id et zip-code
- ***Localisation***: villes
- ***Items commandé*** : id, item, vendeurs, prix (article et livraison)
- ***Commandes*** : id, statuts, date, livraison
- ***Paiements*** : type de paiement et montant
- ***Reviews*** : retour des clients
- ***Vendeurs*** : information des vendeurs



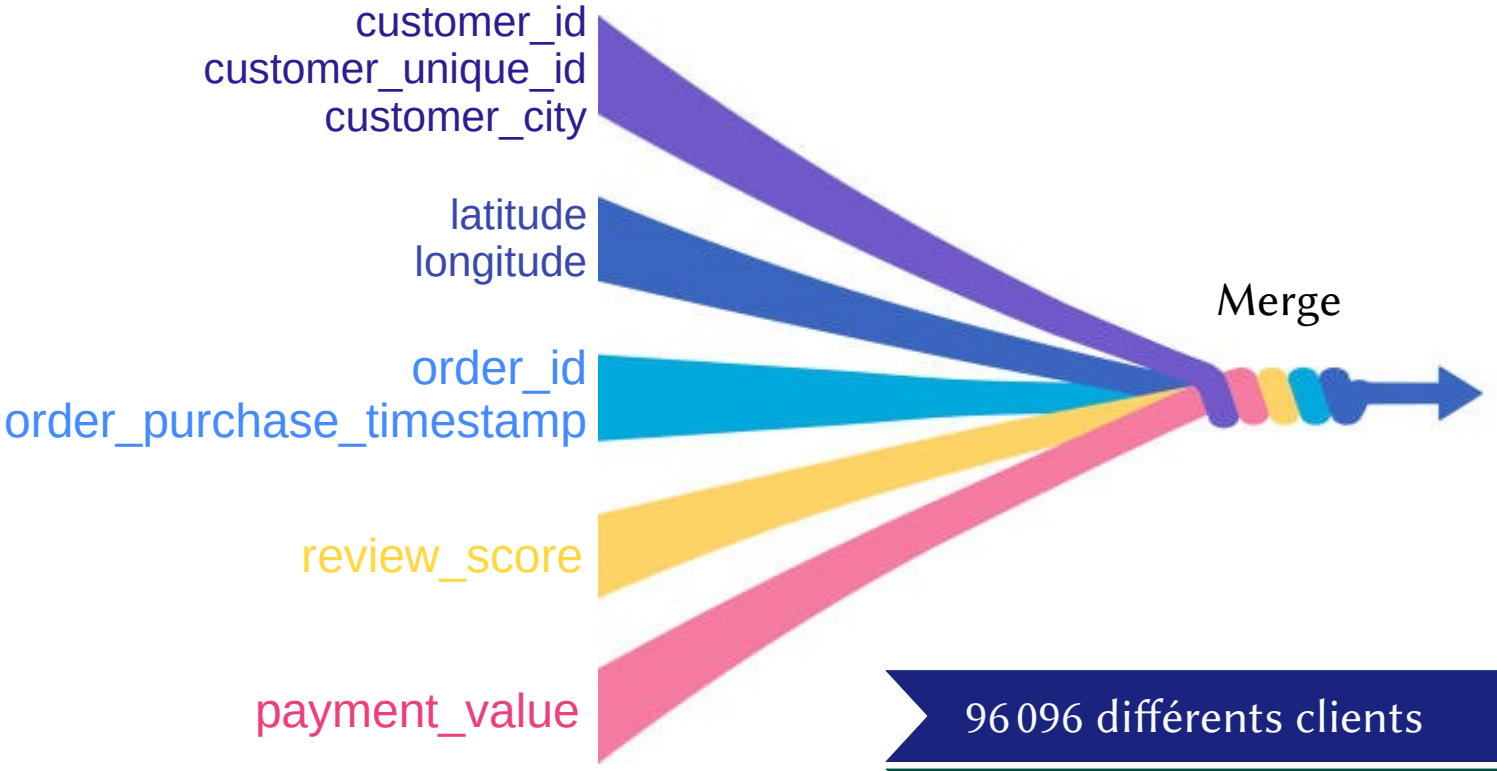

*8 csv files*

*Customer_id vs* ***Customer_unique_id***

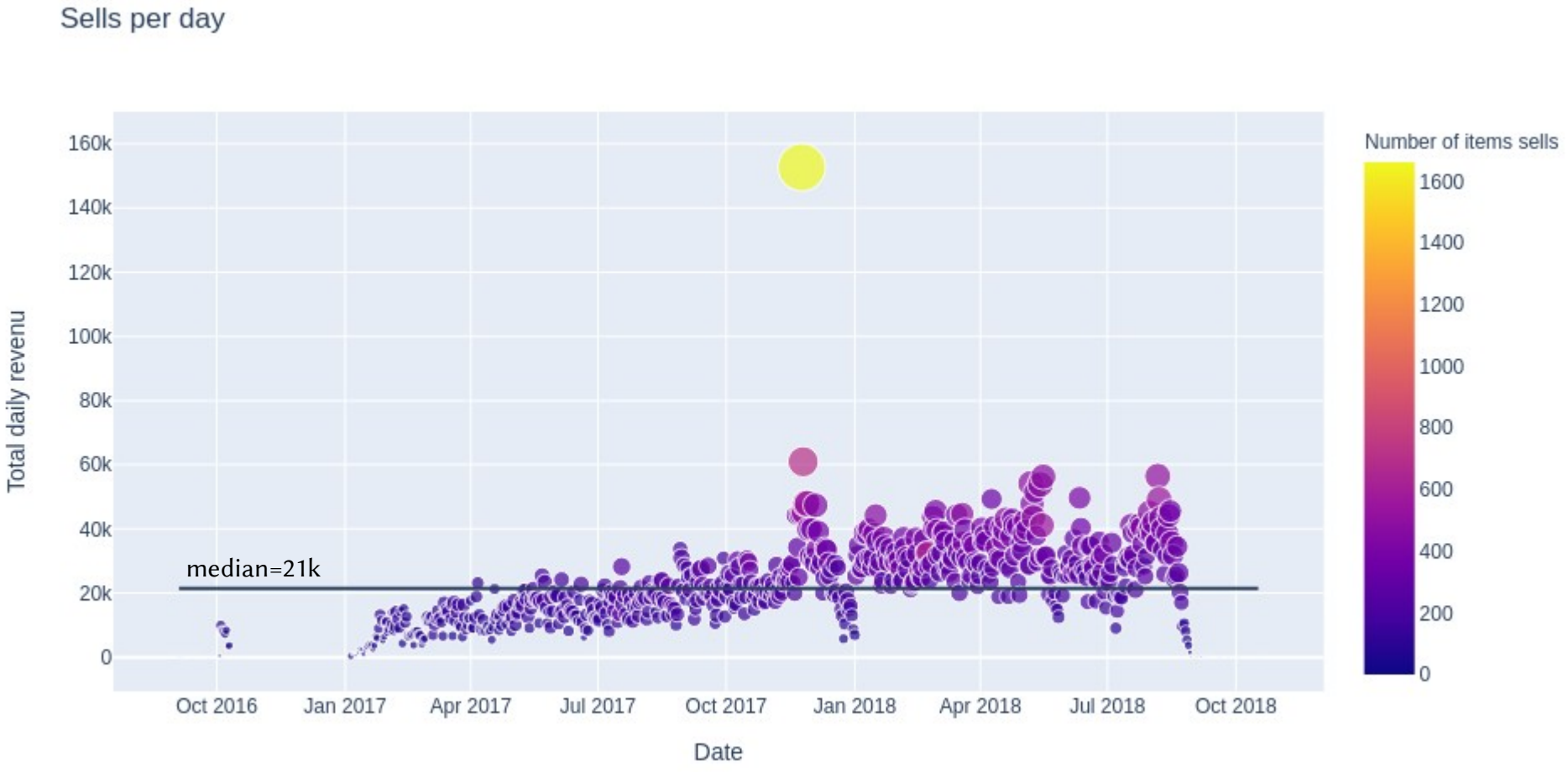
## *Sélection des features à partir du Data-set*



## *Sélection des features à partir du Data-set*



# *Analyse exploratoire*

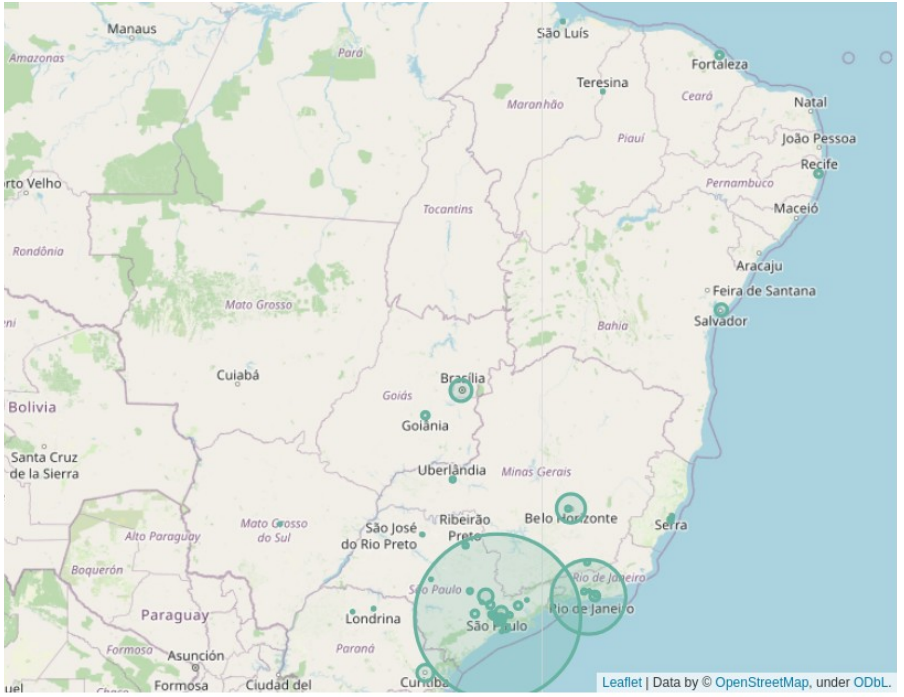


# *Analyse exploratoire*

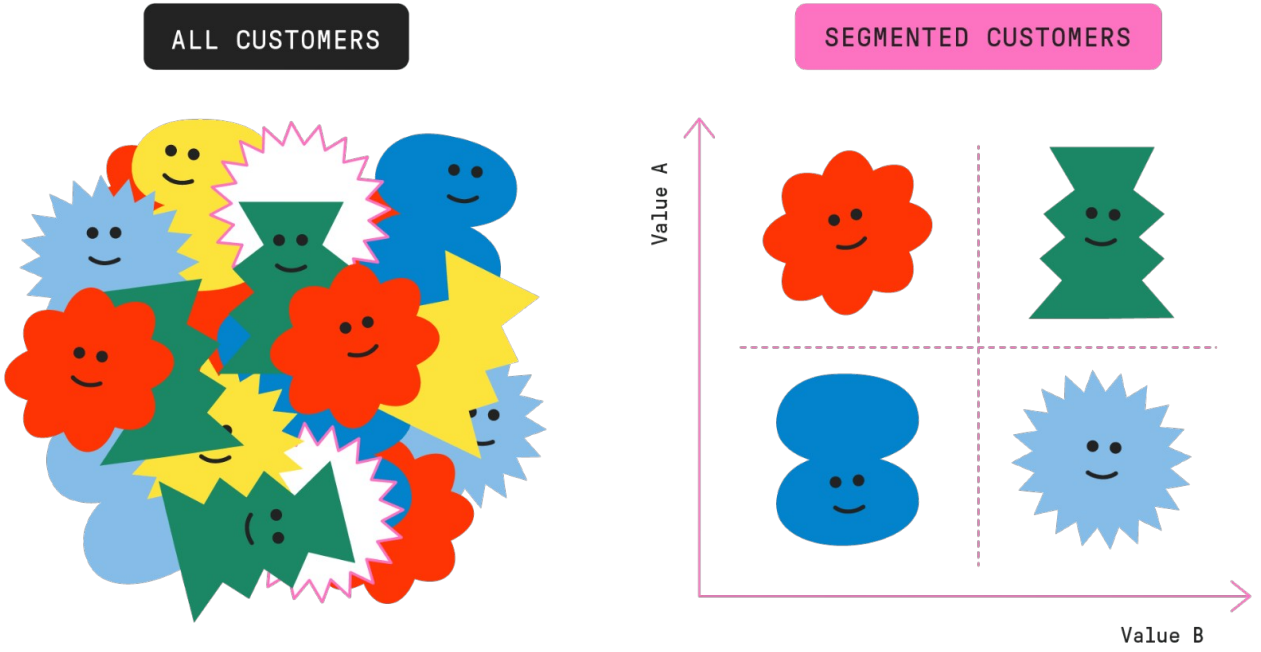
Localisation des acheteurs


4069 villes
50 % < 3 commandes

50 principales villes




Sao Paulo ⟶ 15k
Rio de Janeiro ⟶ 7k




# Évaluation des différentes méthodes de clustering

# Méthode analytique

**R** **F** **M** **Score**

Recency

Frequency

Monetary

Méthode analytique

**R** ecency **F** requency **M** onetary **Score**


Méthodes de apprentissage automatique non-supervisées

- K-Mean
- DBSCAN
- Agglomerative approach

Variables de classification

Méthode analytique

**R**ecency **F**requency **M**onetary **Score**


Méthodes de apprentissage automatique non-supervisées

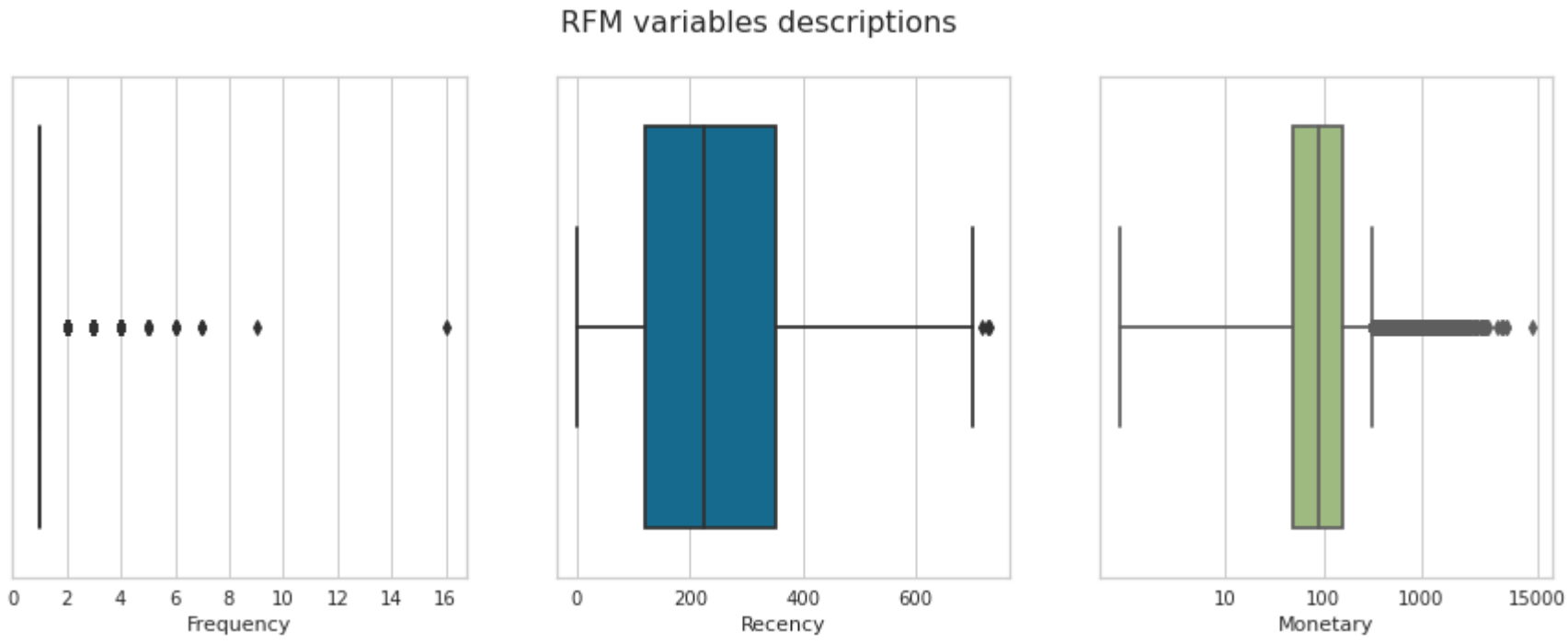
- K-Mean
- DBSCAN
- Agglomerative approach

Variables de classification

Fenêtre temporelle : Septembre 2016 à Septembre 2018 (100 % du data-set)

## *RFM score*

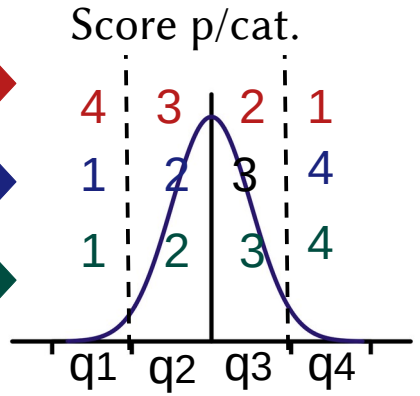
95 419 clients



## *RFM score*

95 419 clients

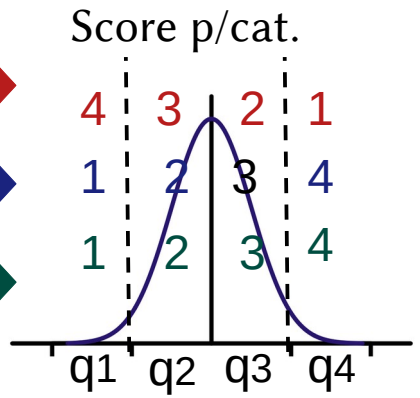
- Recency
- Frequency
- Monetary



## *RFM score*

95 419 clients

- Recency
- Frequency
- Monetary

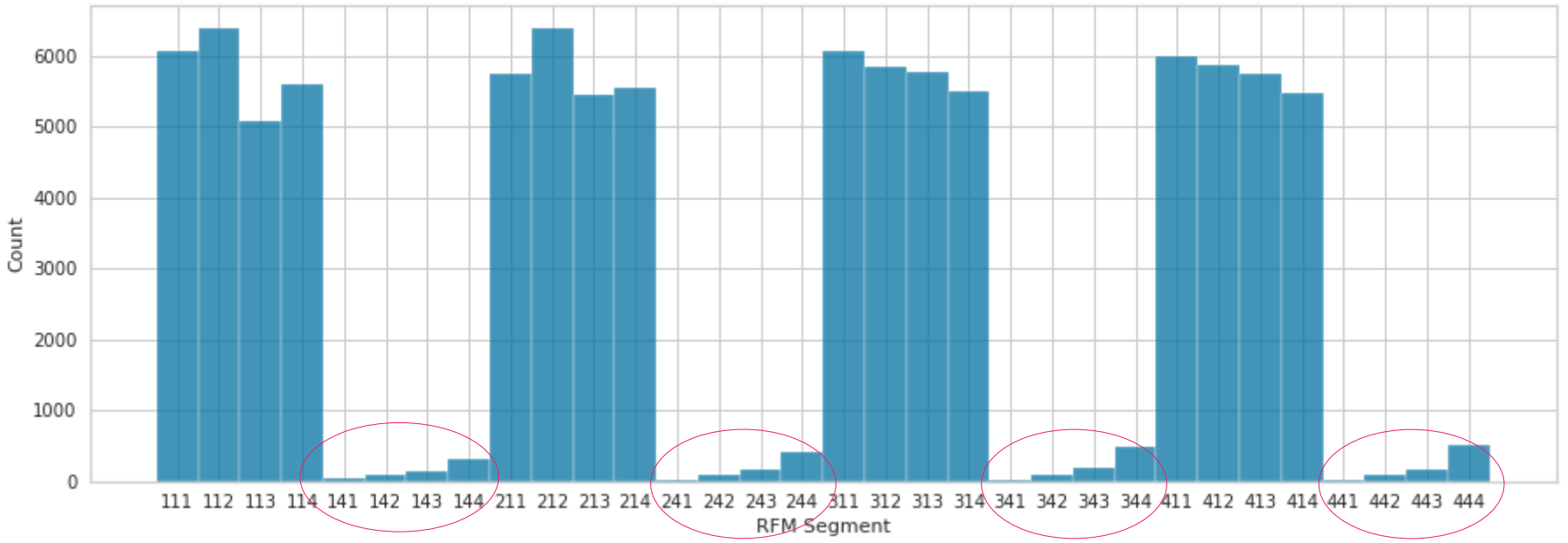



Note :
Frequency > 1 ⟶ score = 4

32 types de clients différents

RFM segment score

Clients avec frequency > 1



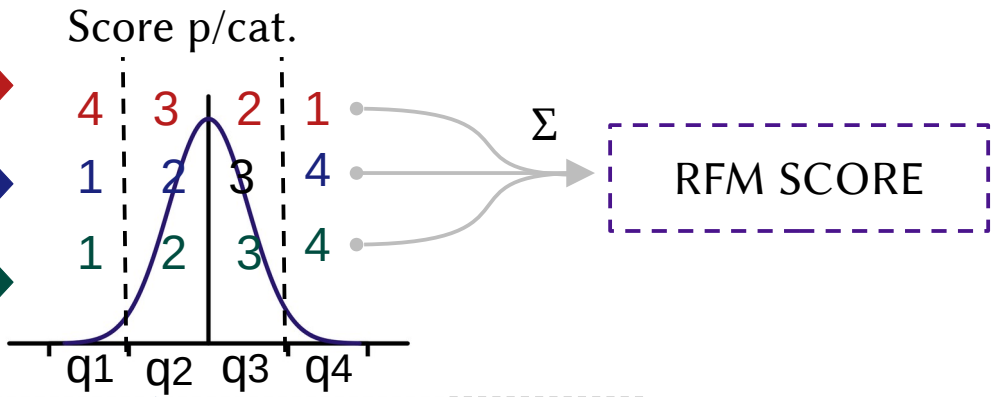
## *RFM score*

95 419 clients

Recency

Frequency

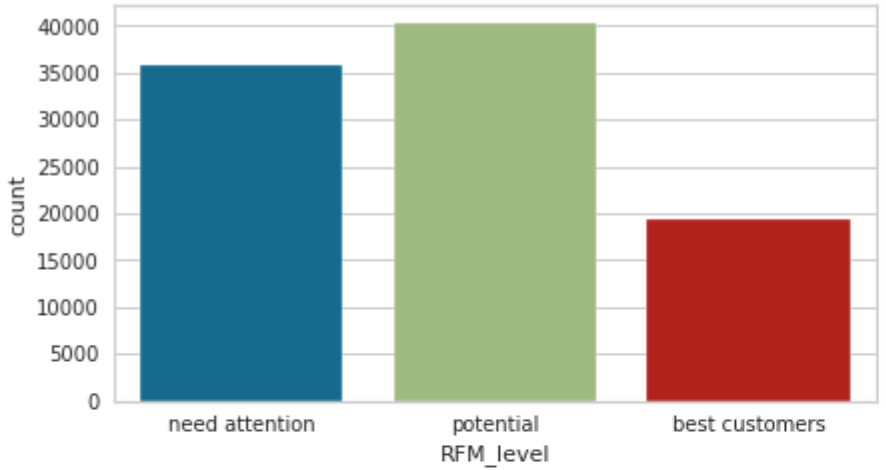
Monetary

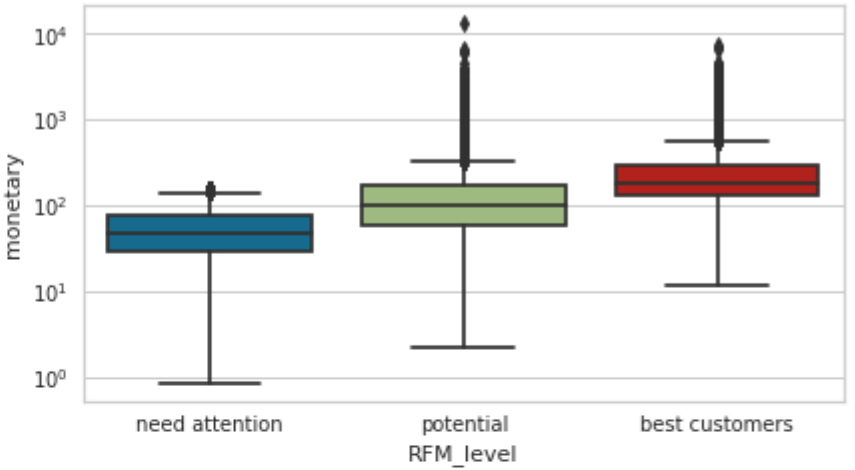


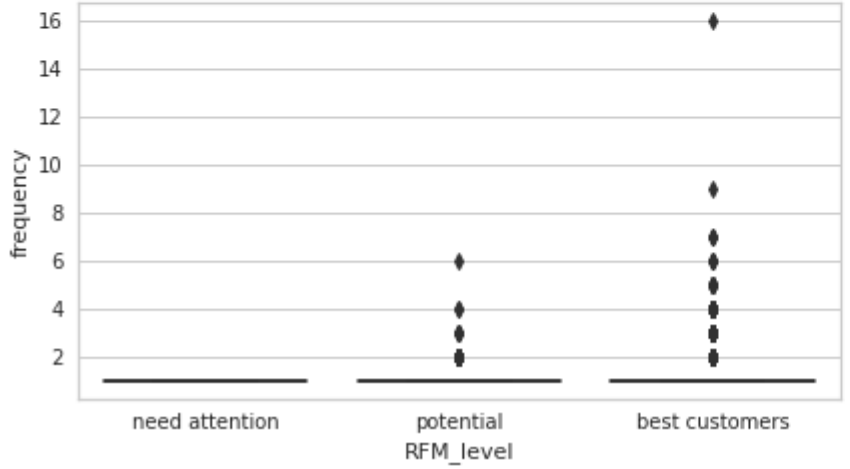


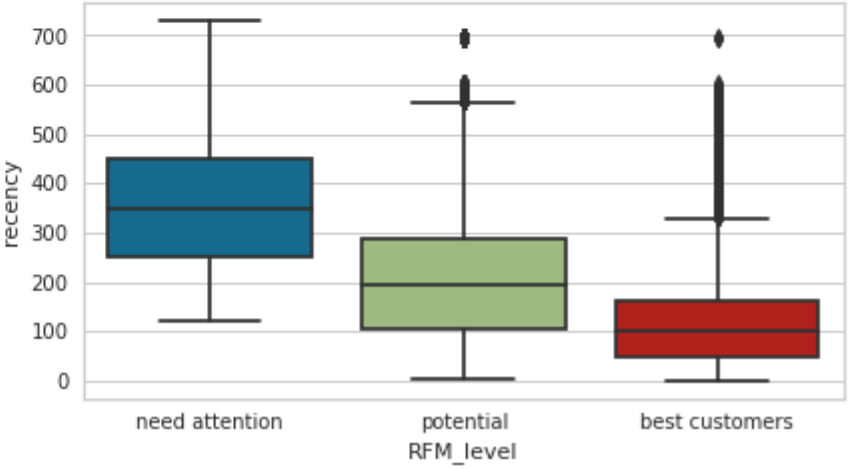
- Need attention: [min, Q1);
- Potential: [Q1, Q3);
- Best customers: [Q3, max]

# *RFM score*


count
40000
35000
30000
25000
20000
15000
10000
5000
0
need attention
potential
best customers
RFM_level


monetary
$10^4$
$10^3$
$10^2$
$10^1$
$10^0$
need attention
potential
best customers
RFM_level


frequency
16
14
12
10
8
6
4
2
need attention
potential
best customers
RFM_level


recency
700
600
500
400
300
200
100
0
need attention
potential
best customers
RFM_level

# ***Apprentissage non-supervisé***

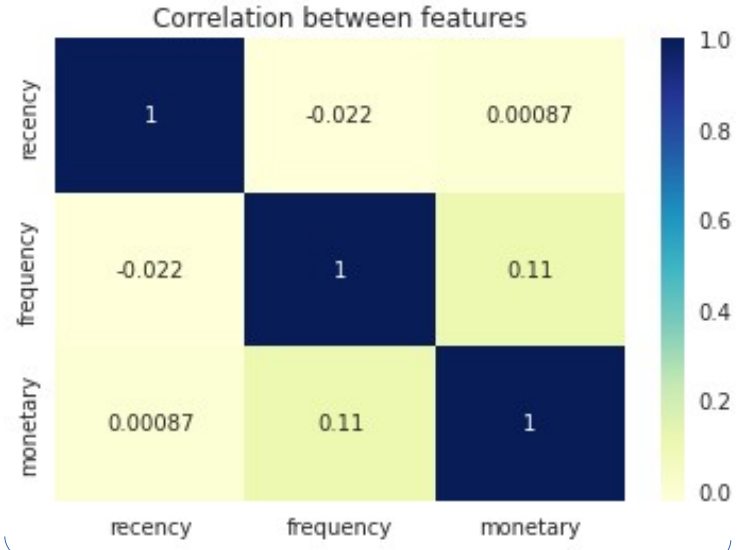
K-Mean

95 419 clients

DBSCAN et Agglomerative clustering

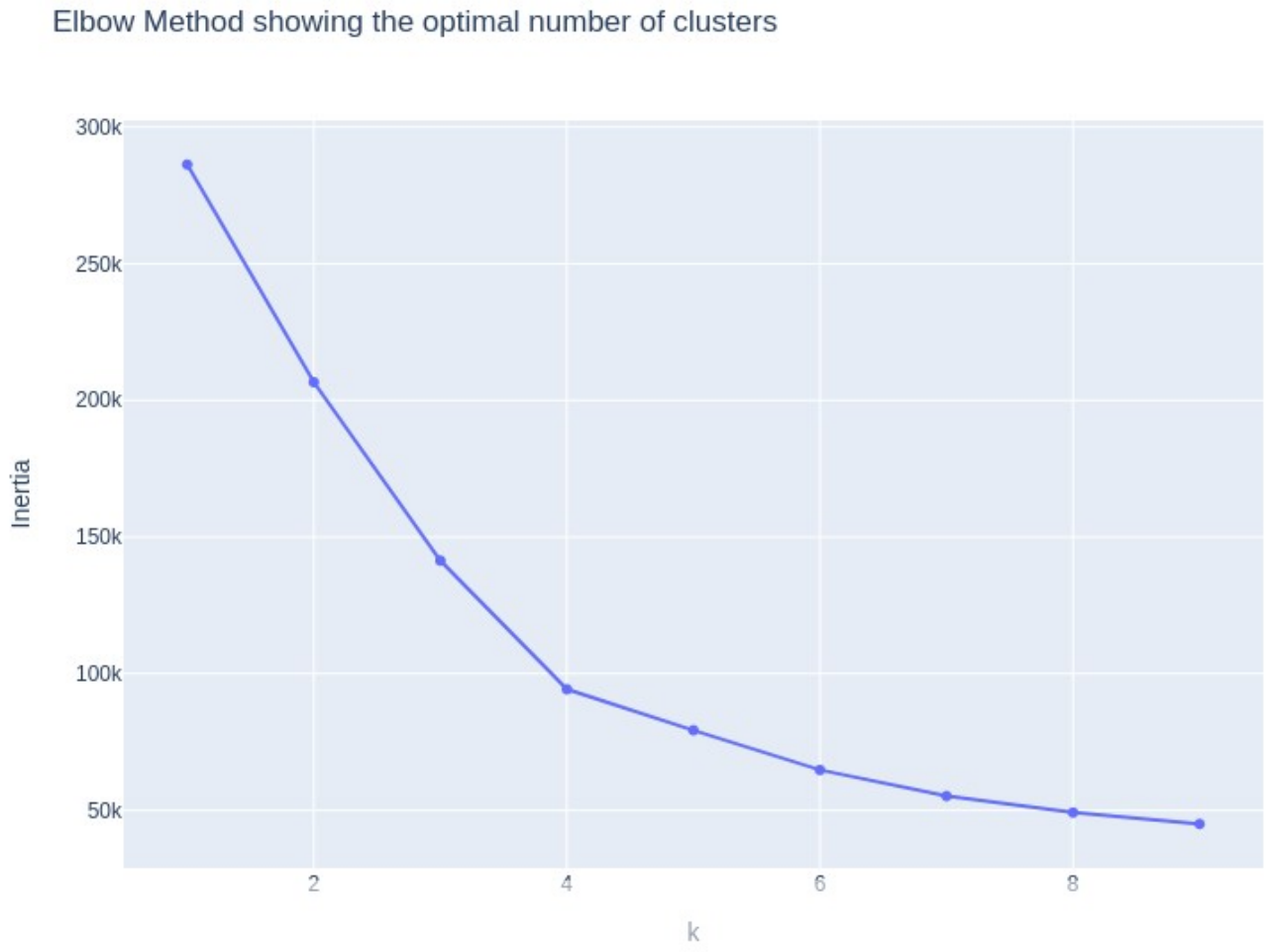
10 000 clients random

Features : RFM



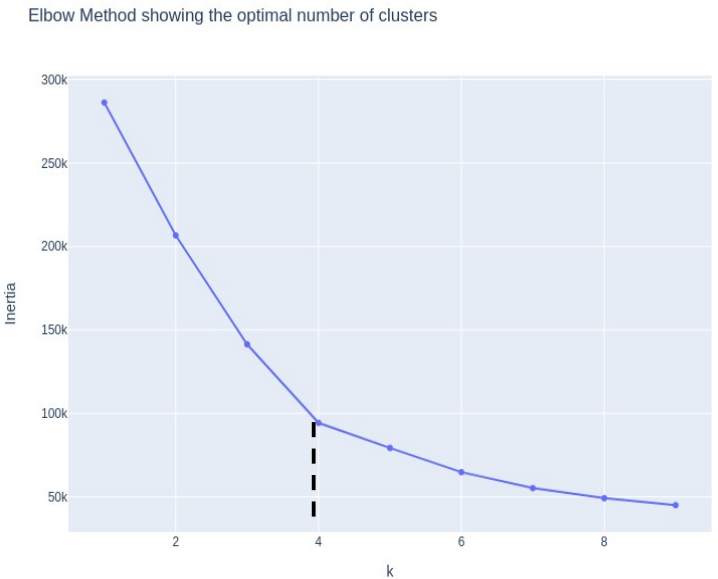

StandardScaler()

# Paramètre K : Méthode du coude

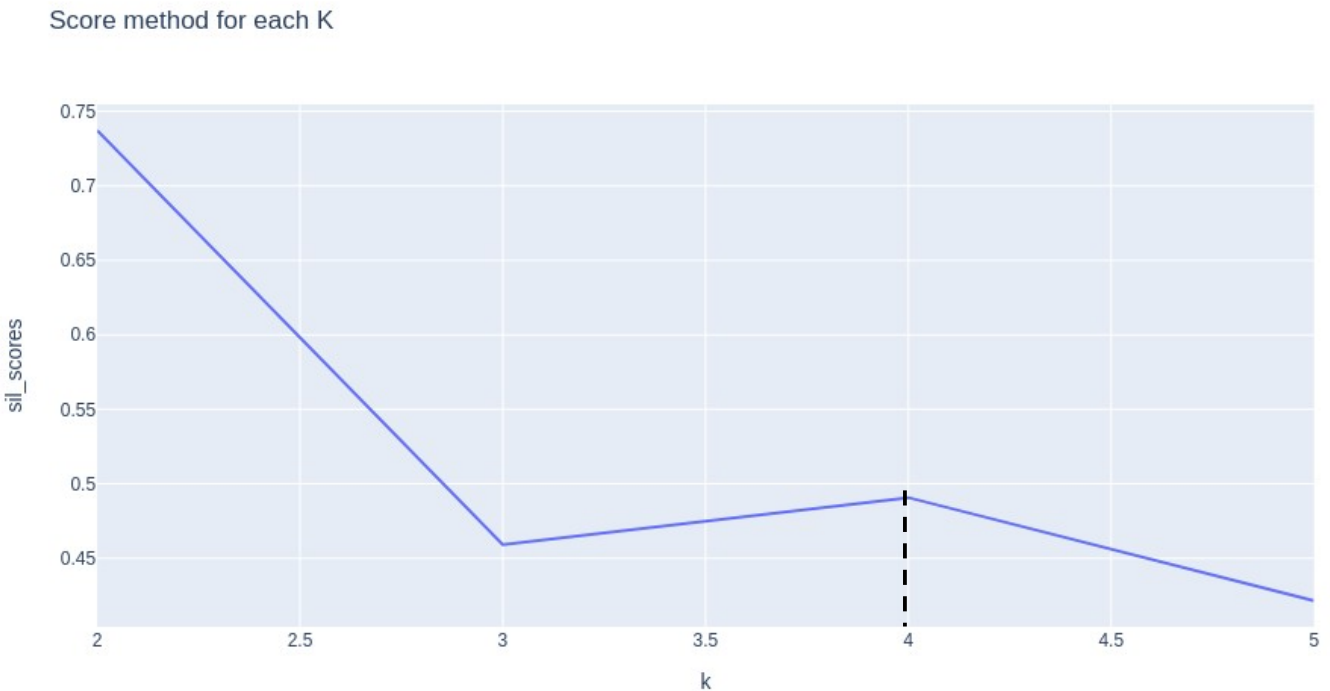


# K-Mean

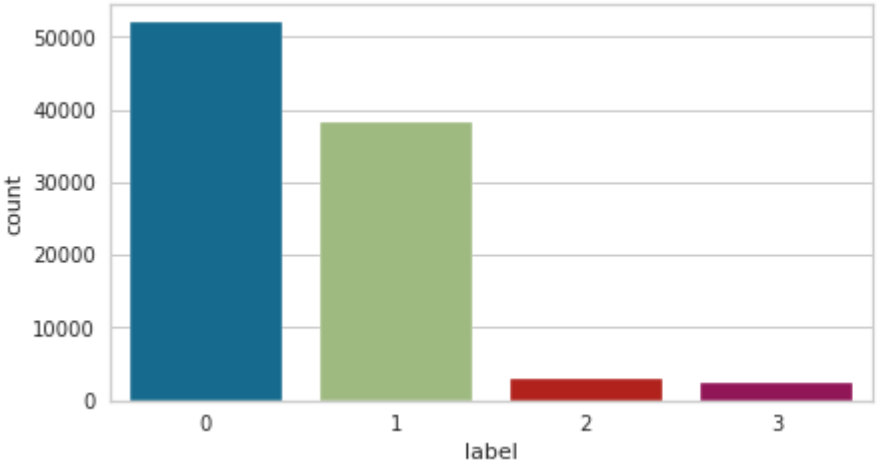
## Paramètre K : Méthode du coude

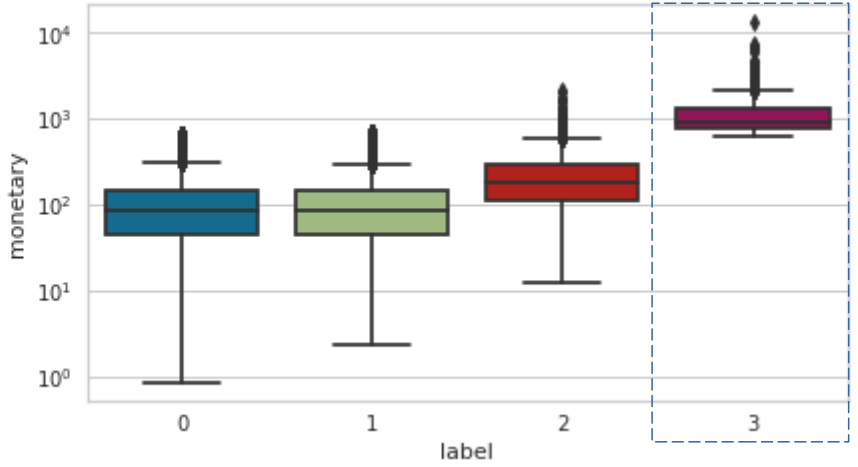


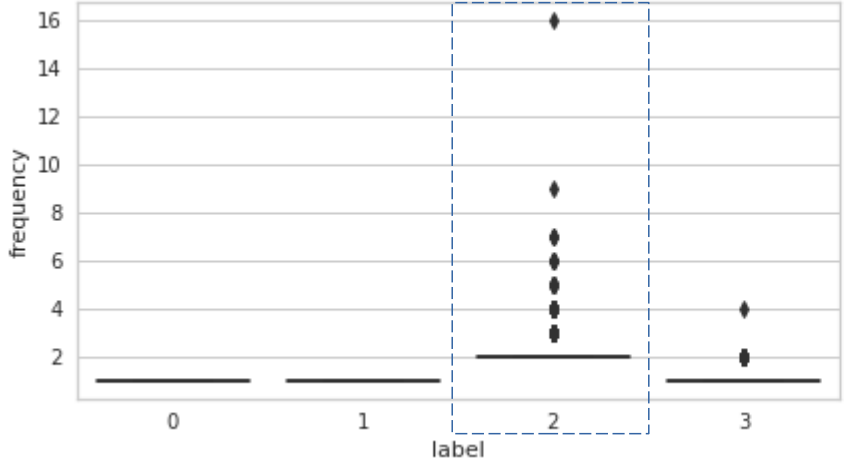

## Paramètre K : Méthode du score

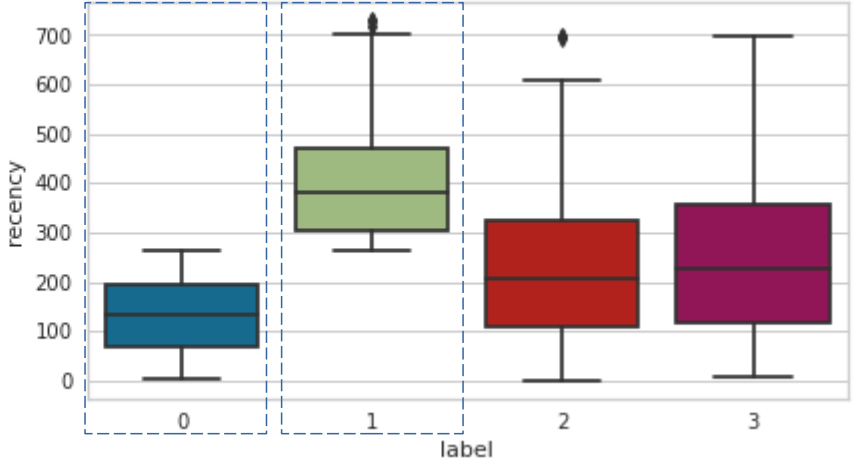


# K-Mean clusters



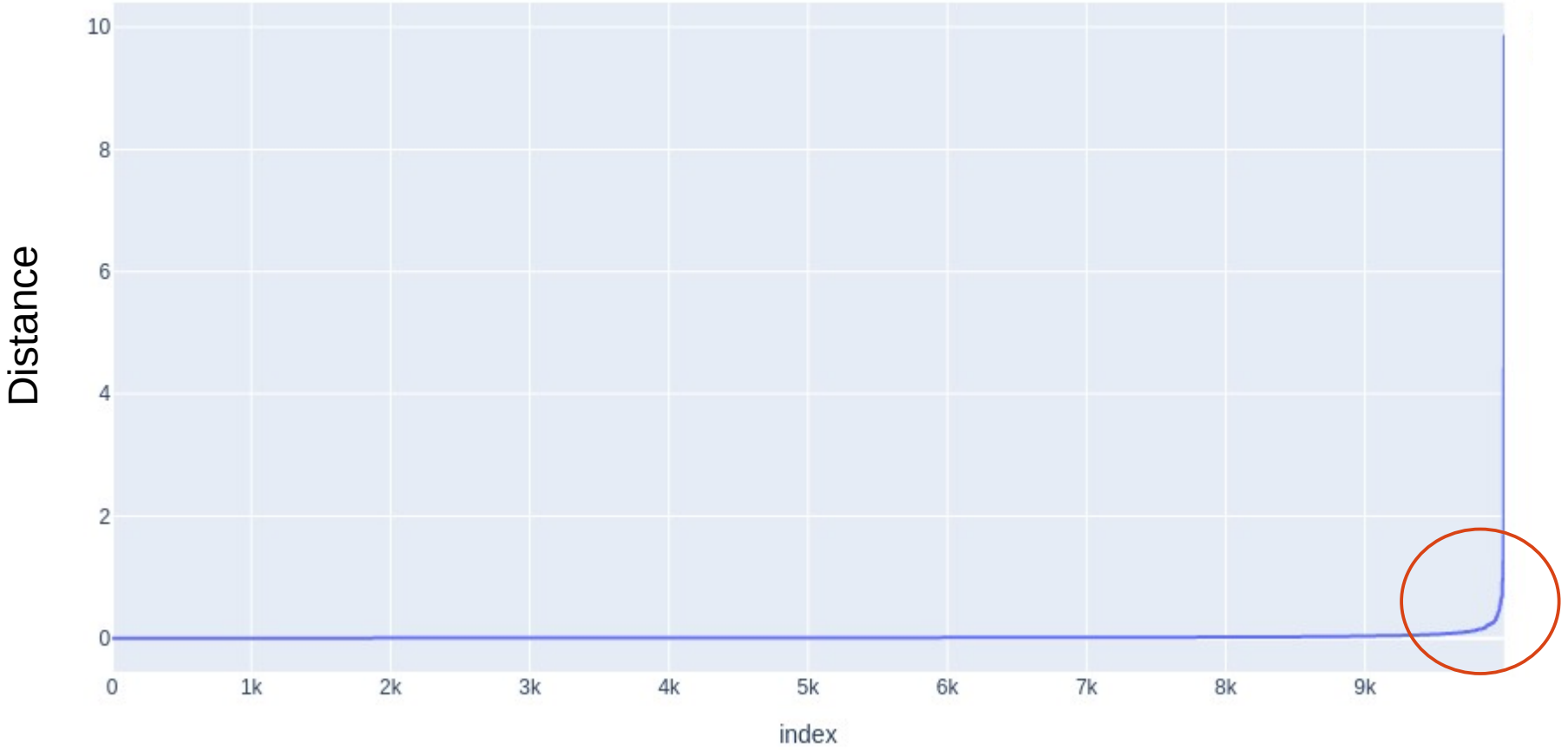



# DBSCAN : min_sample et epsilon



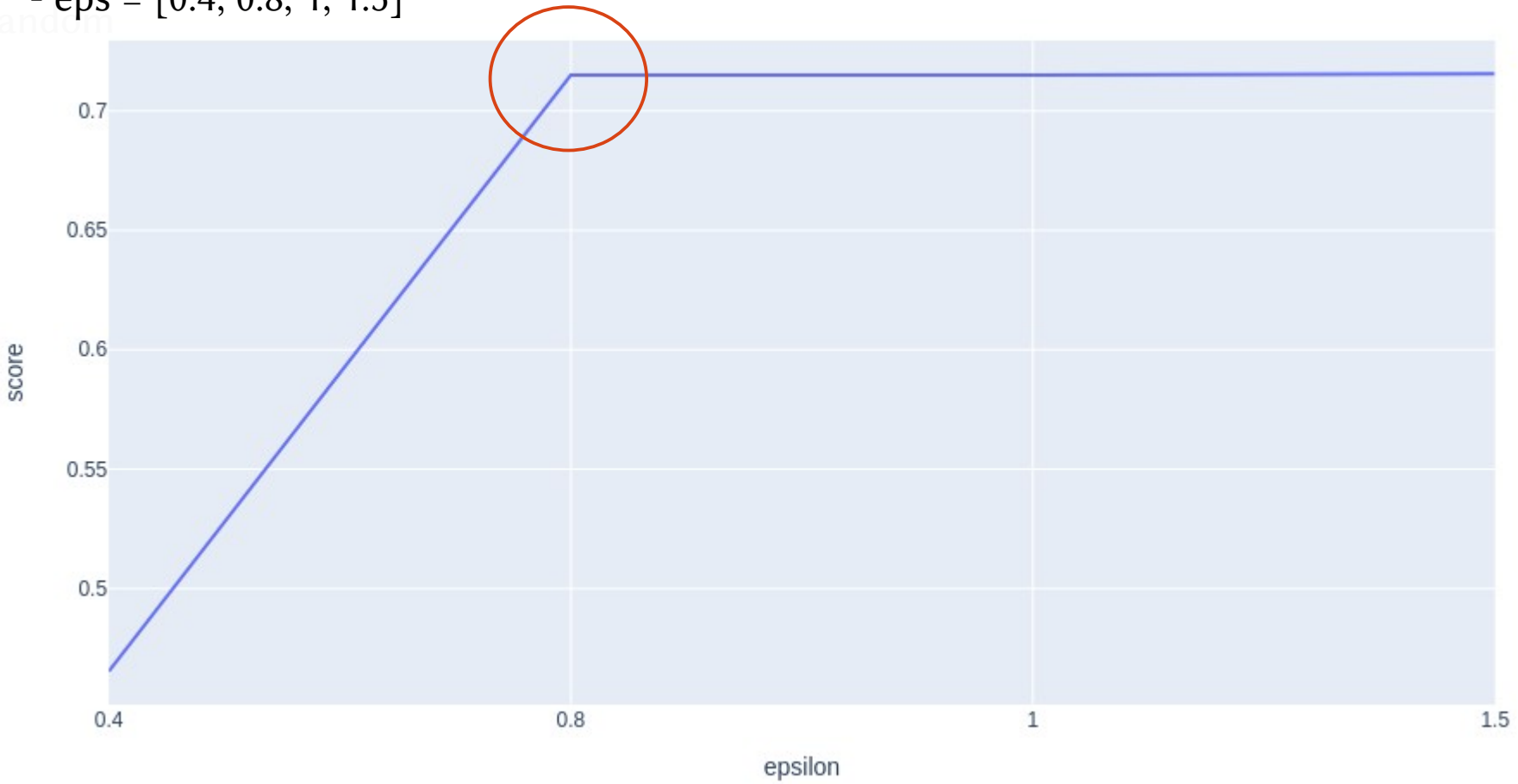
- min_samples = 2 * n_features
- Calcul de distance euclidien entre un point et les min_samples les plus proches



# DBSCAN : min_sample et epsilon

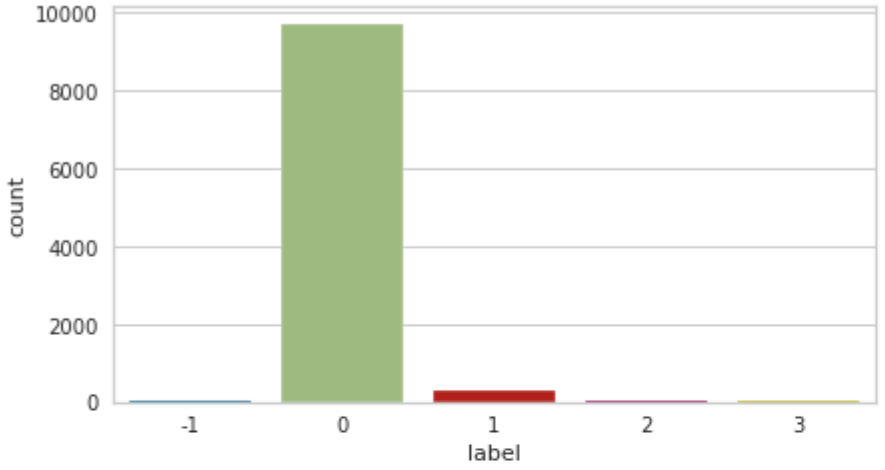
Biblio

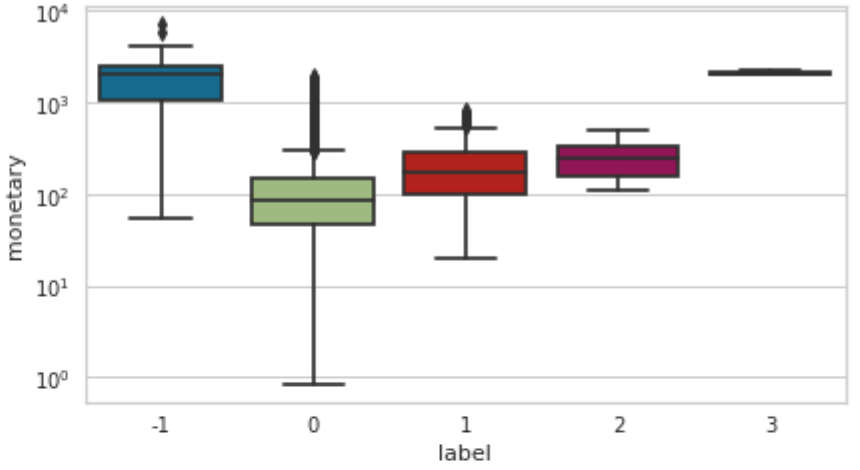
- min_samples = 2 * n_features
- eps = [0.4, 0.8, 1, 1.5]

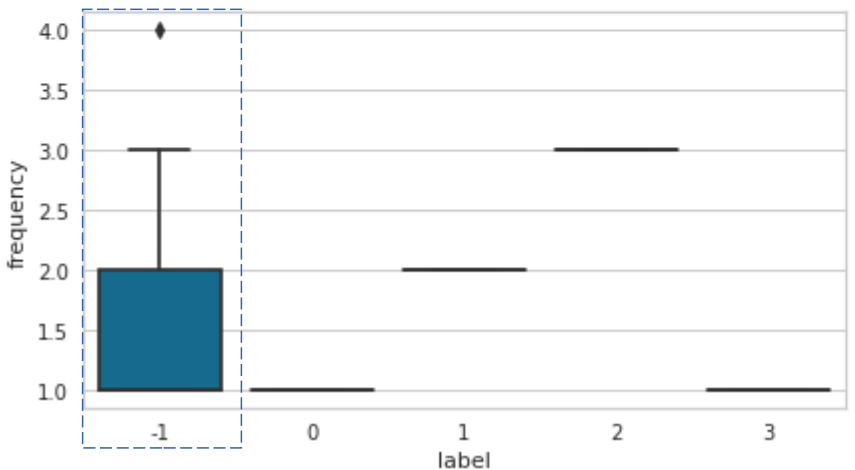


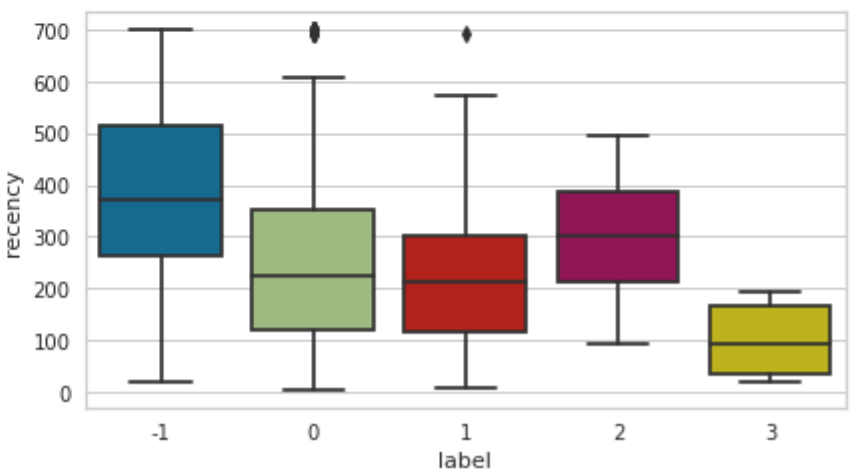
# DBSCAN : clusters



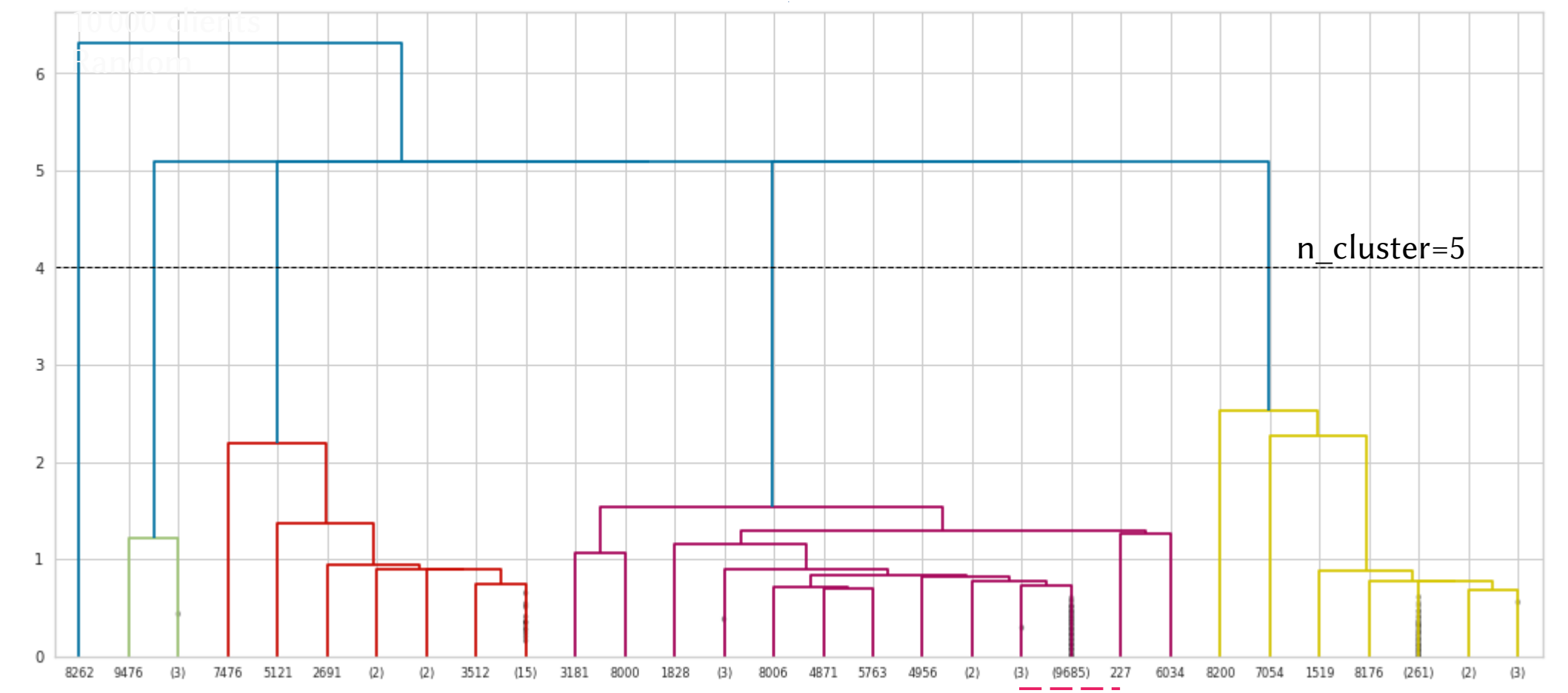


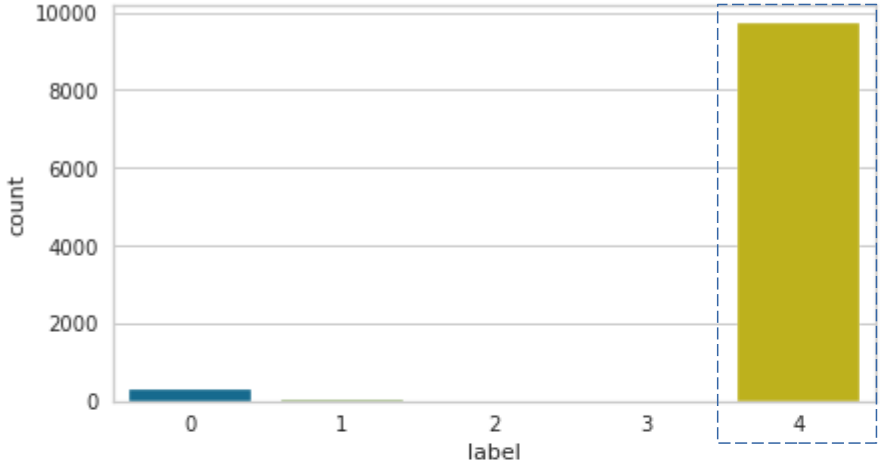



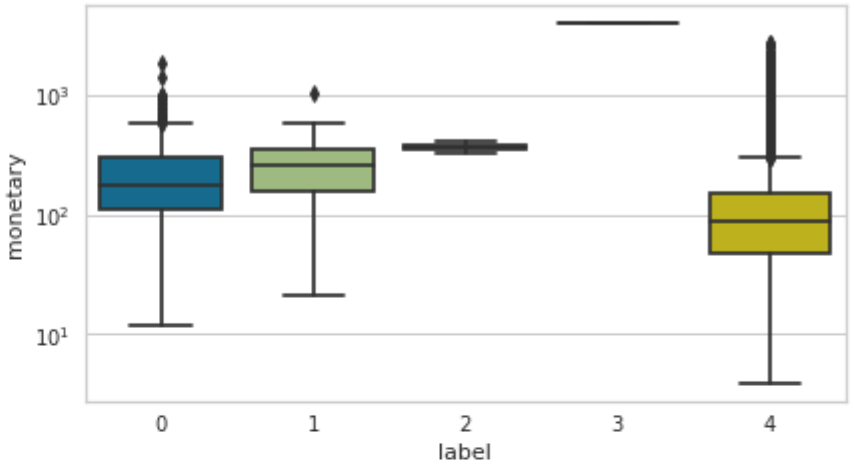



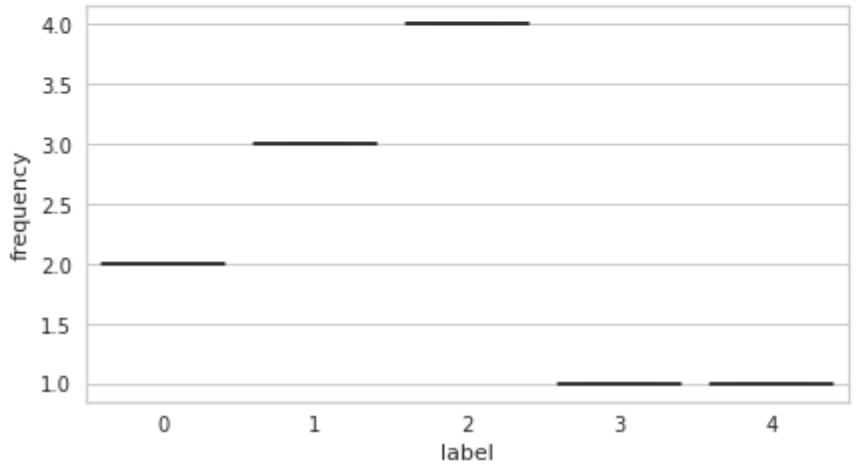
# Agglomerative clustering : dendrogramme

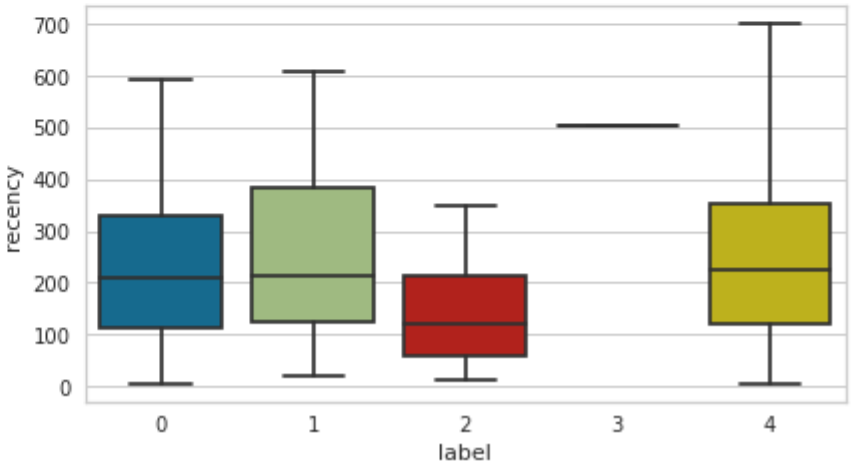


# Agglomerative clustering : clusters



# ***Conclusions***

- La méthode K-Mean donnée des clusters plus équilibrés avec des définition métier plus marquées.

- En terme de mémoire et temps de calcule, la méthode K-Mean est plus performante

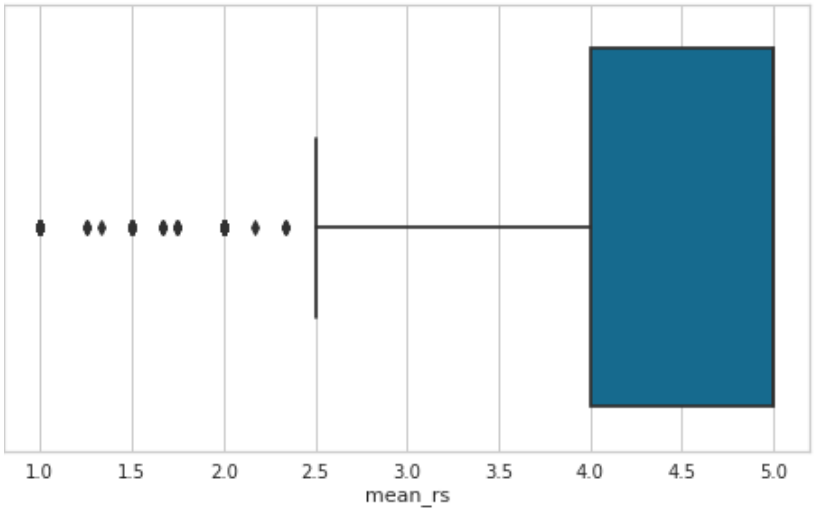
- La méthode Agglomerative n’as pas de fonction predict

# Modèle final, fonction utilisateur et maintenance

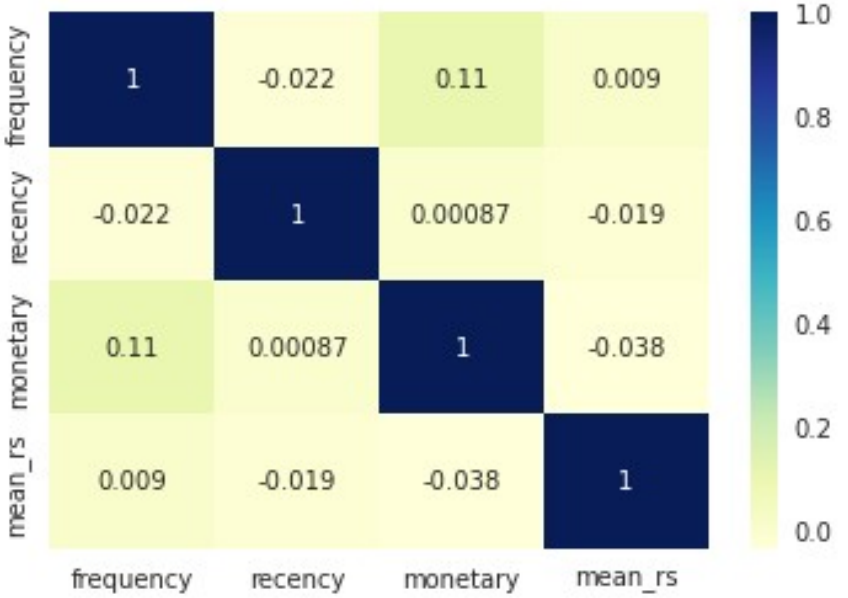
# ***K-Mean final***

95 419 clients

Features : RFM + Mean review score




Features : correlation



# ***K-Mean final***

Features

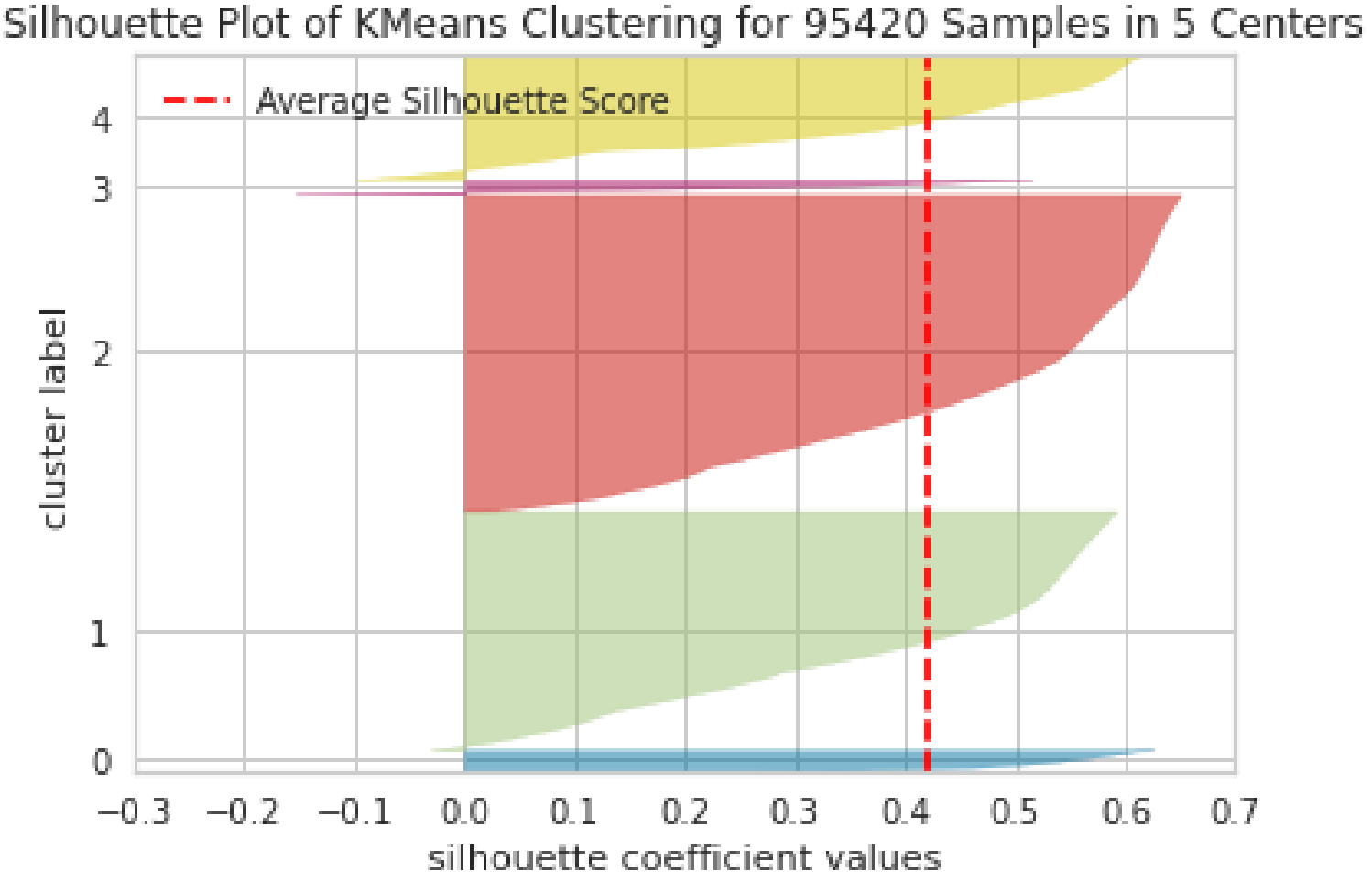
StandardScaler()

Elbow method

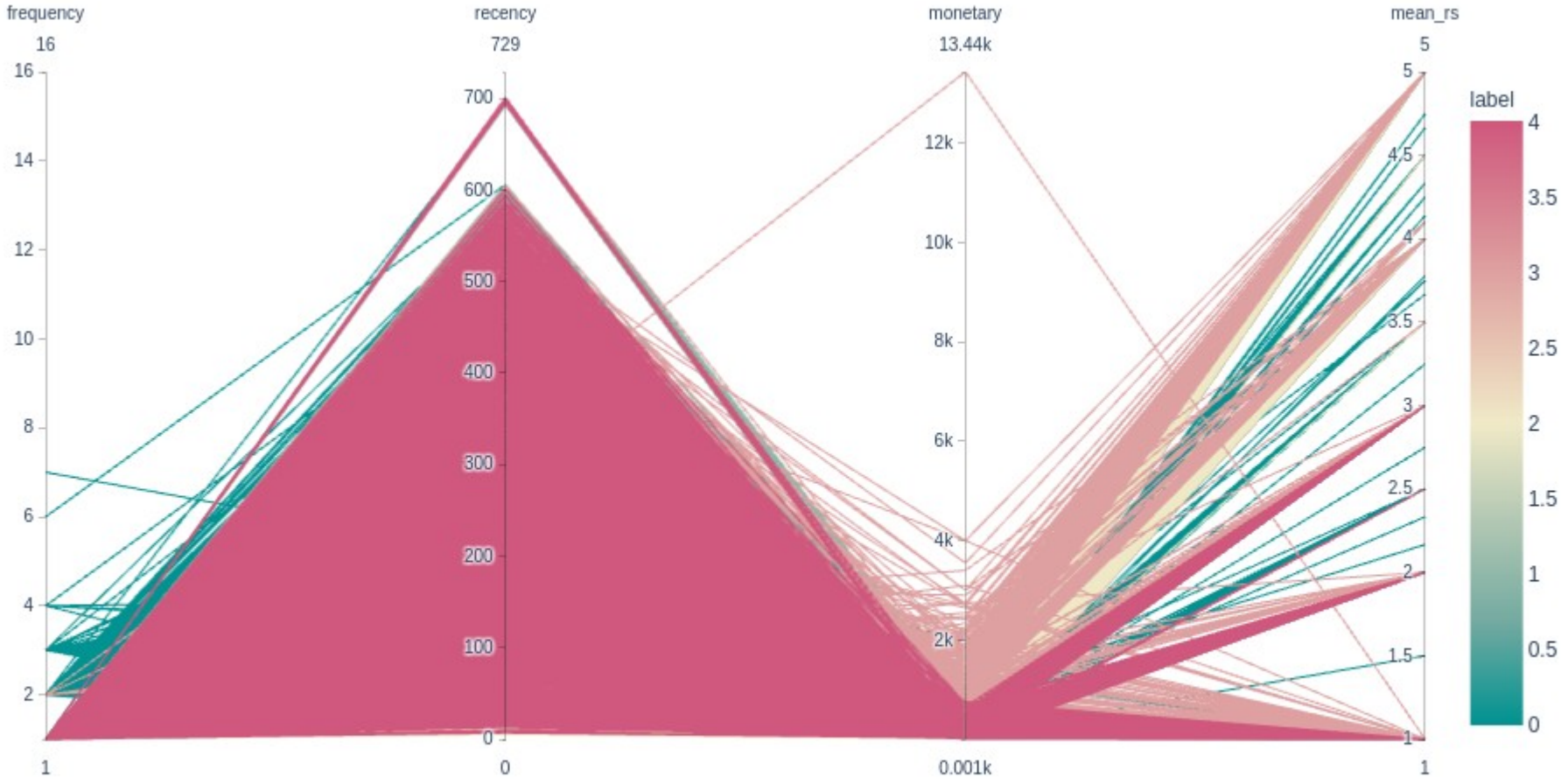
Score method

Paramètre

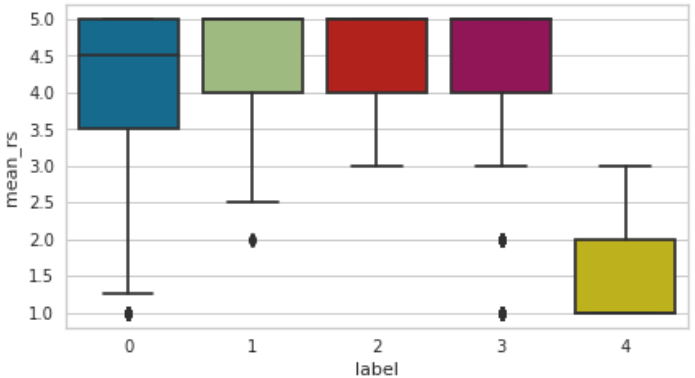
K = 5

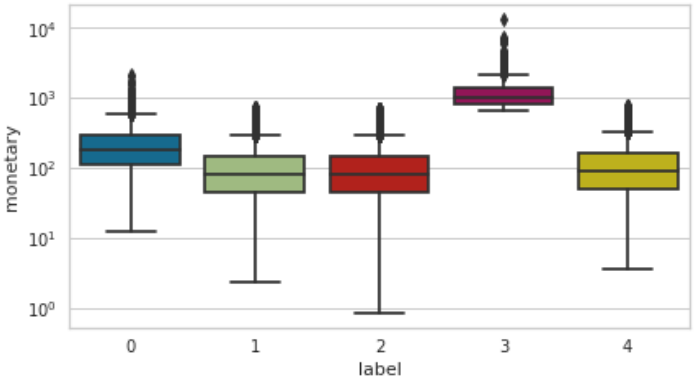


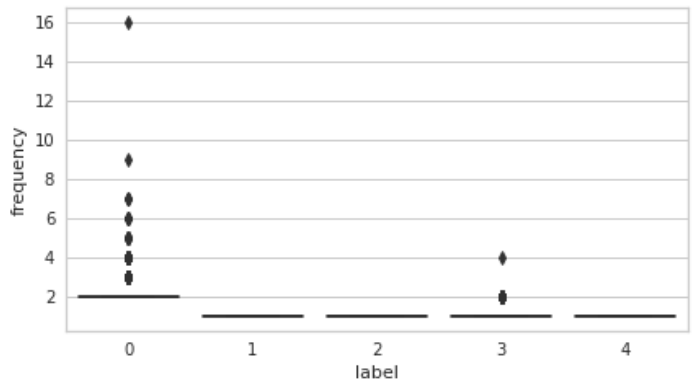
# ***K-Mean final***

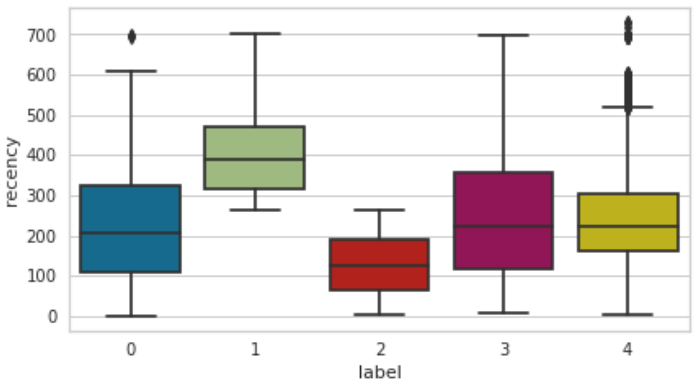

frequency
16
16
14
12
10
8
6
4
2
1
recency
729
700
600
500
400
300
200
100
0
0
monetary
13.44k
12k
10k
8k
6k
4k
2k
0.001k
mean_rs
5
5
4.5
4
3.5
3
2.5
2
1.5
1
label
4
3.5
3
2.5
2
1.5
1
0.5
0

# ***Clusters et « Persona »***




Jean:
- Achète fréquemment

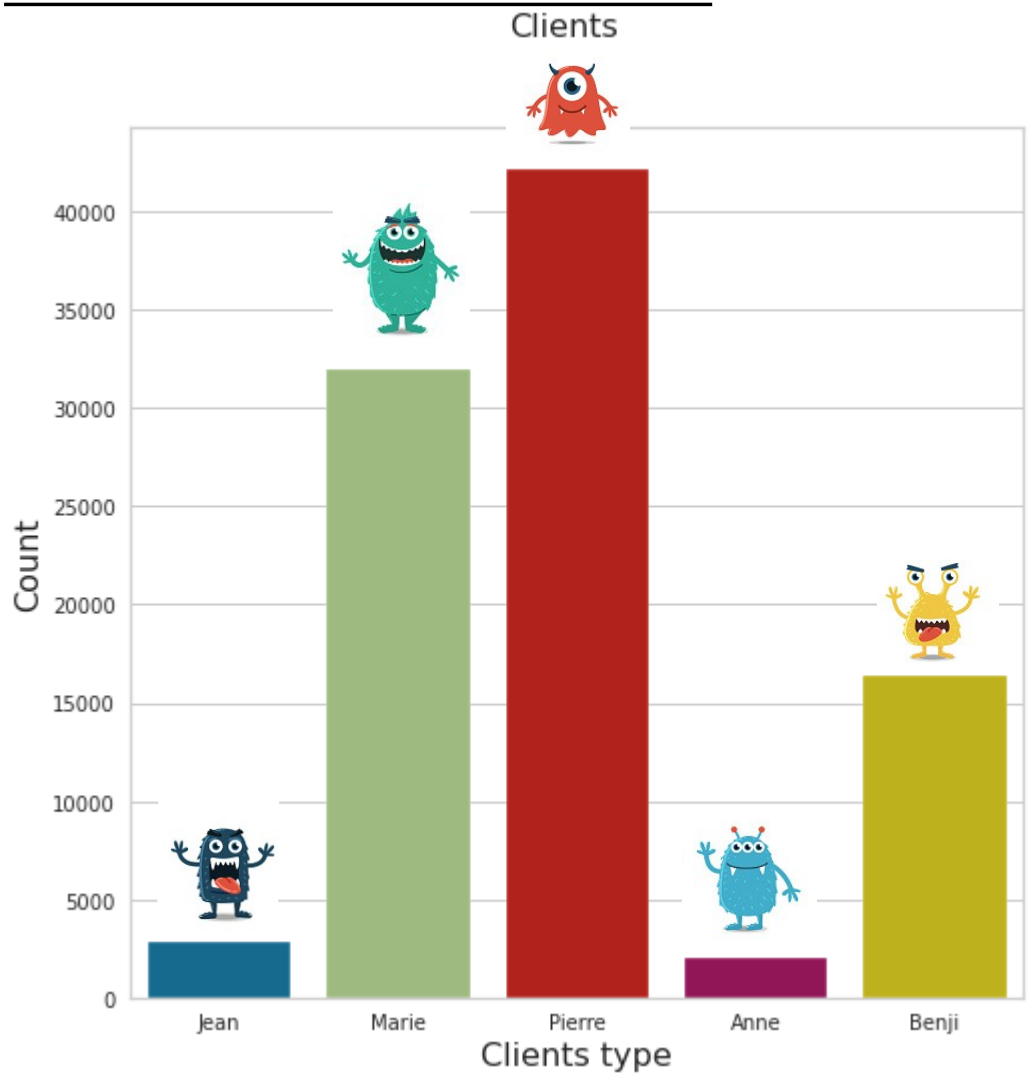
Marie:
- 1 achat il y + 1 an

Pierre:
- Client récent (1 an) pas dépensier

Anne:
- Ce qu'on aime $$$

Benji:
- Pas content

# Clusters et « Persona »




Jean:
- Achète fréquemment

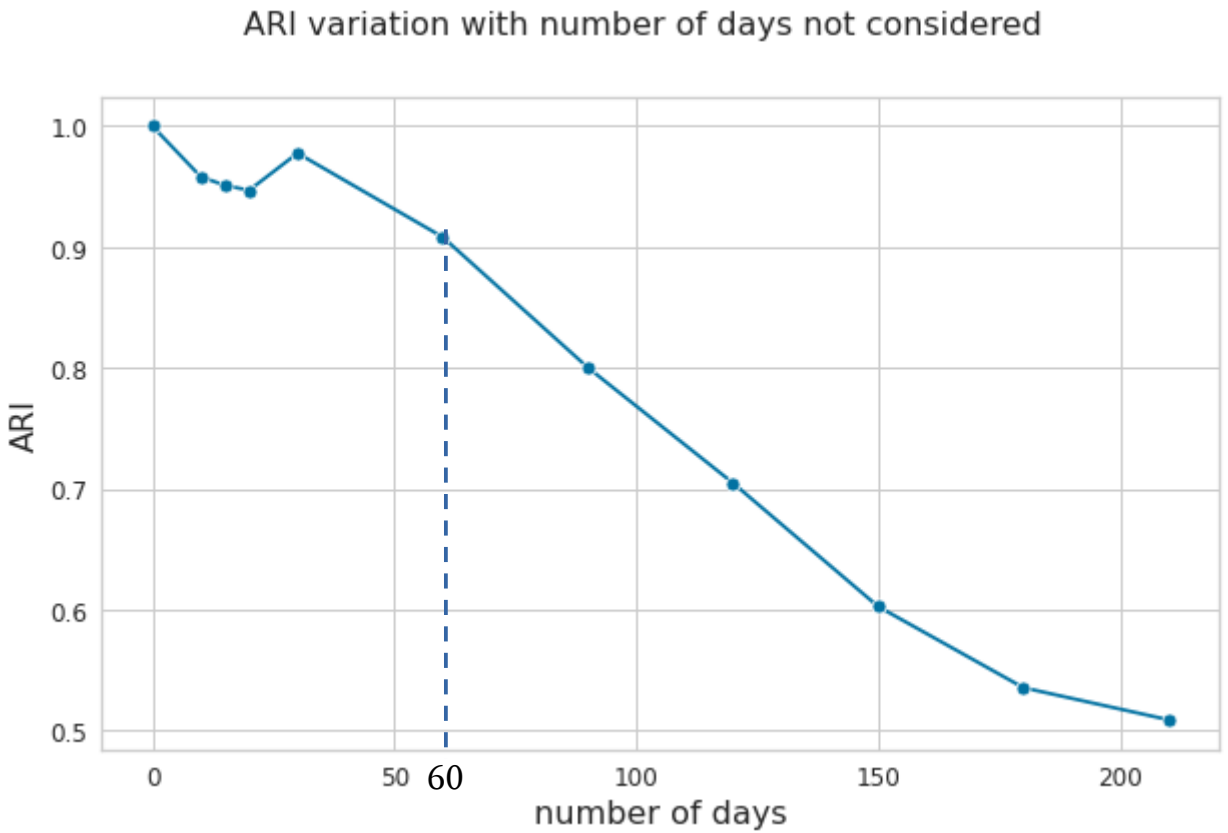
Marie:
- 1 achat il y + 1 an

Pierre:
- Client récent (1 an) pas dépensier

Anne:
- Ce qu'on aime $$$

Benji:
- Pas content

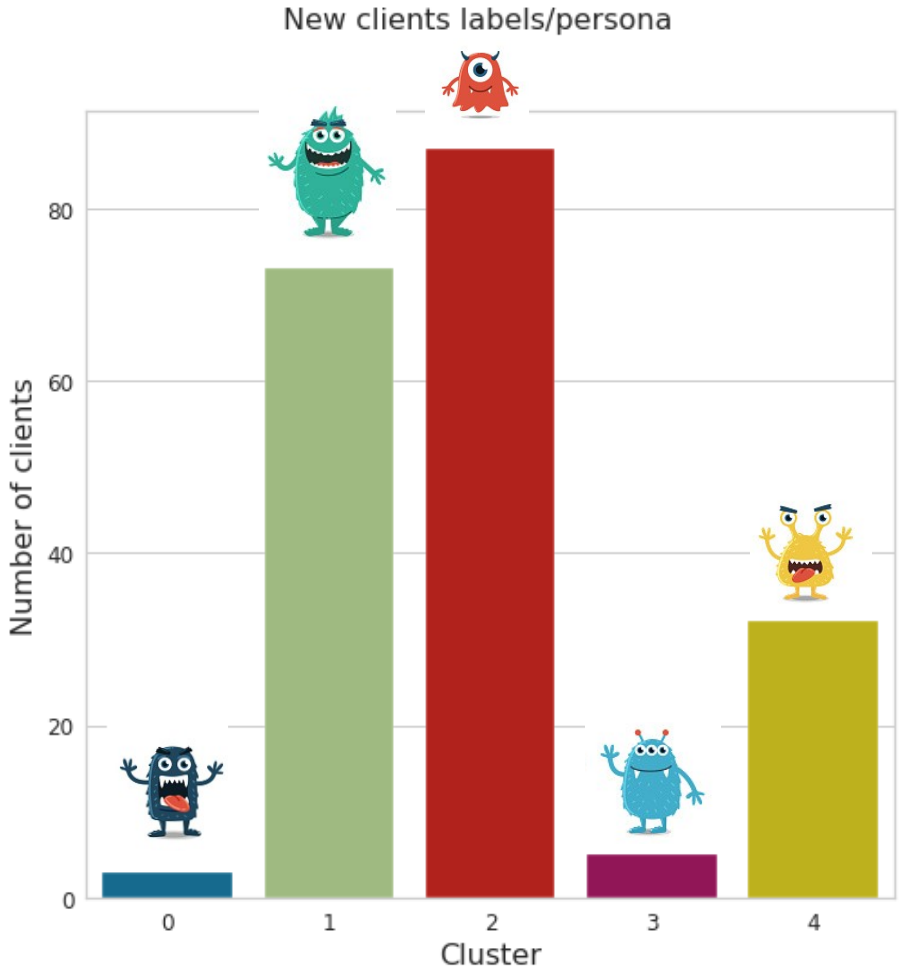
# ***Maintenance :***

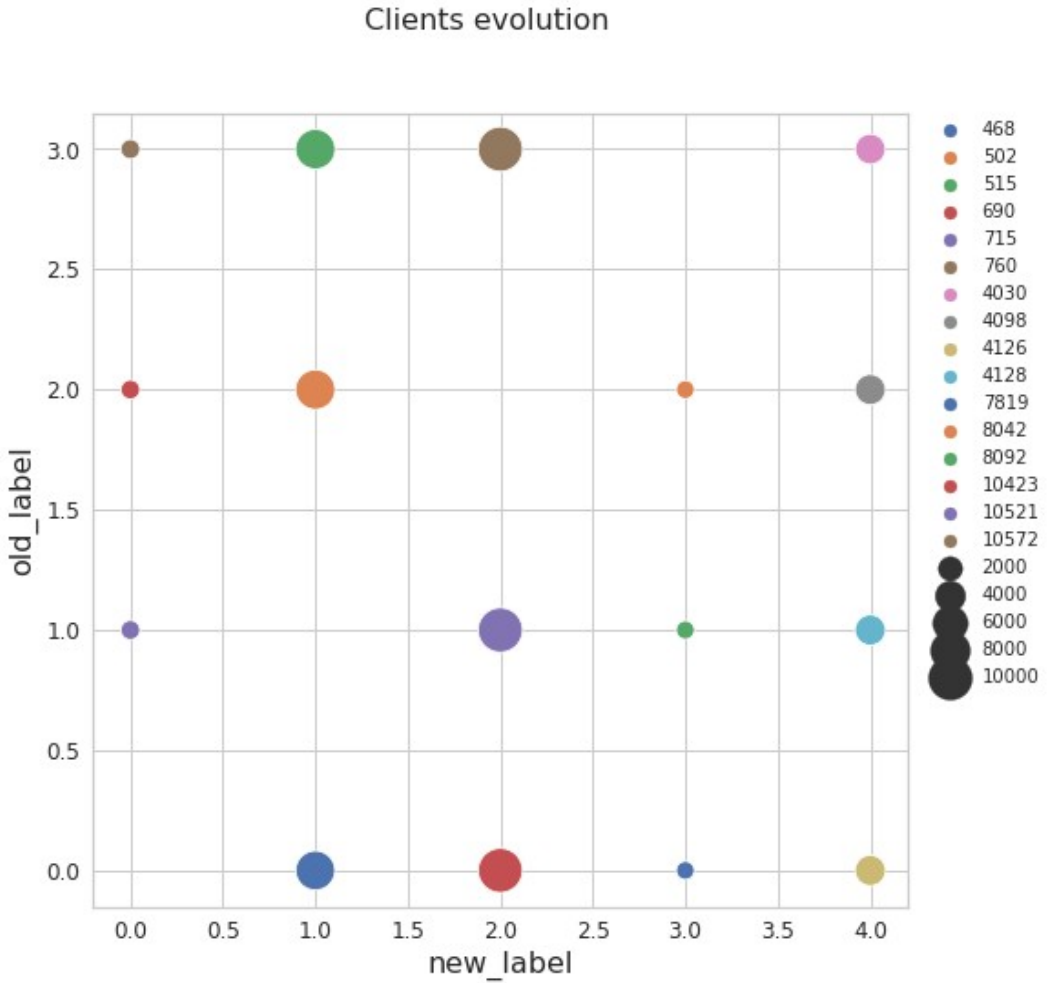



La maintenance est conseillé tout les 2 mois

# ***Fonction pour rajouter nouveaux clients:*** Example

Données d'entrée : base de données de la **dernier mise à jours** et **base des données à jour**




Rocio

# ***Conclusion:***

- Cinq clusters ont été établie pour identifier le comportement des clients

- Il a été établie qu'il est pertinent de faire une maintenance de l'algorithme de clustering tout les 2 mois

- Un fonction est développe pour voir l'évolution des clients dans le temps et identifier à quel cluster appartiens les nouveaux clients

Merci de votre attention