

# Segmenter des clients d'un site e-commerce

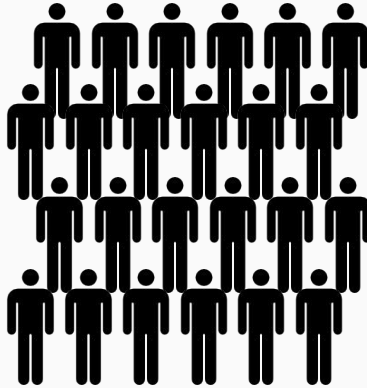
20/10/2021 - Parcours Data Scientist  
Sébastien Bourgeois

# Sommaire

1. Problématique
2. Nettoyage & exploration
3. Pistes de modélisation
4. Modèle final sélectionné

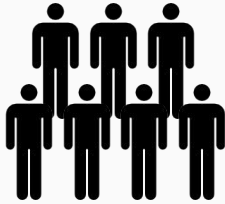
# 1. Problématique

## *Problématique*



# 1. Problématique

## *Problématique*



# 1. Problématique

## *Problématique*



# 1. Problématique

*Interprétation*

Identification de groupes de clients similaires



Problème de clustering

# 1. Problématique

*Pistes de recherche envisagées*

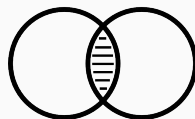
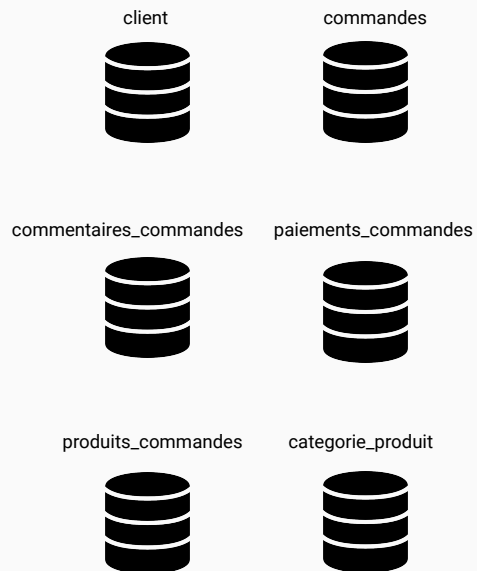
Segmentation RFM

k-means

k-prototypes

## 2. Nettoyage & exploration

### *Nettoyage des données*

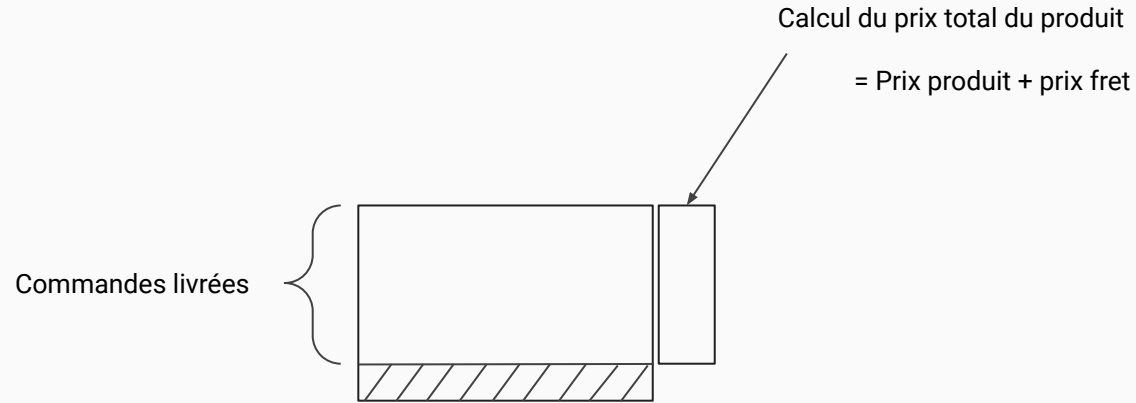


Produits commandés par les clients



## 2. Nettoyage & exploration

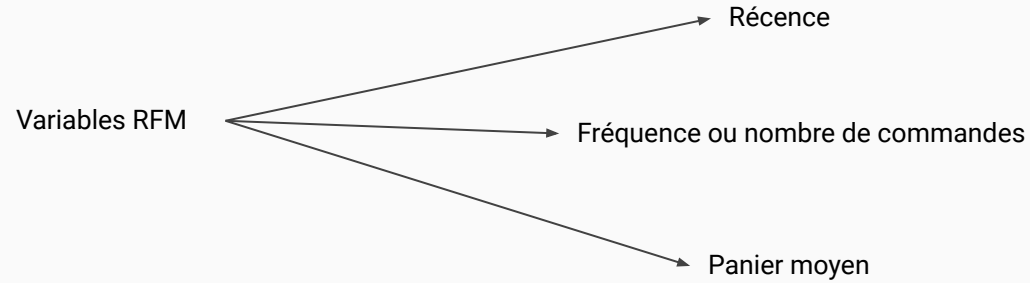
### *Nettoyage des données*



## 2. Nettoyage & exploration

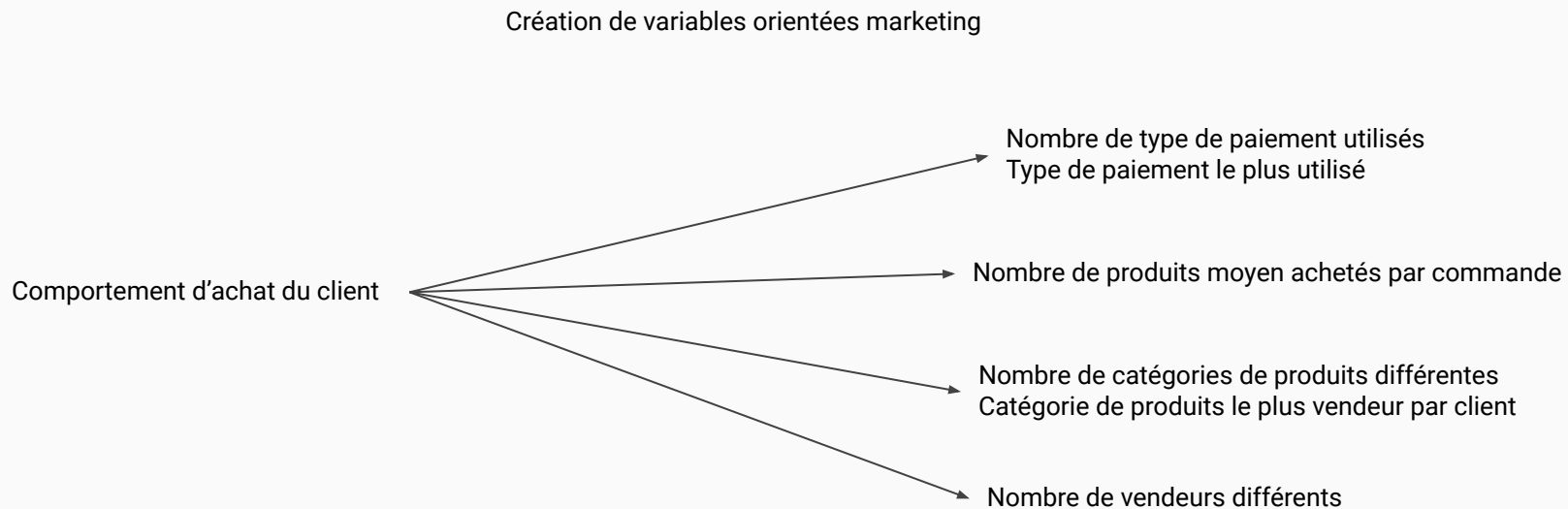
### *Features engineering*

Création de variables orientées marketing



## 2. Nettoyage & exploration

### *Features engineering*

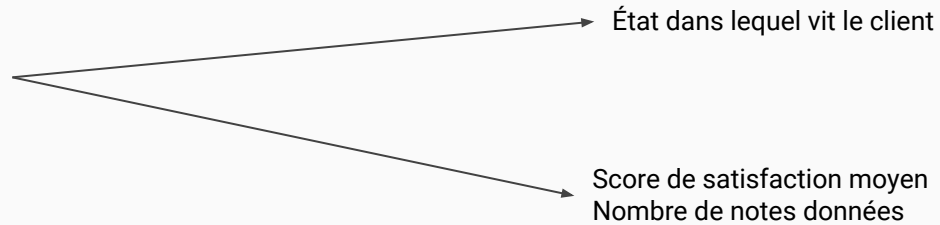


## 2. Nettoyage & exploration

### *Features engineering*

Création de variables orientées marketing

Autres variables sur le client



## 2. Nettoyage & exploration

### *Exploration*

*top\_category*

bed_bath_table
NULL
NULL
telephony
baby
auto
health_beauty
auto

## 2. Nettoyage & exploration

### *Exploration*

*top\_category*

bed_bath_table
NULL
NULL
telephony
baby
auto
health_beauty
auto

## 2. Nettoyage & exploration

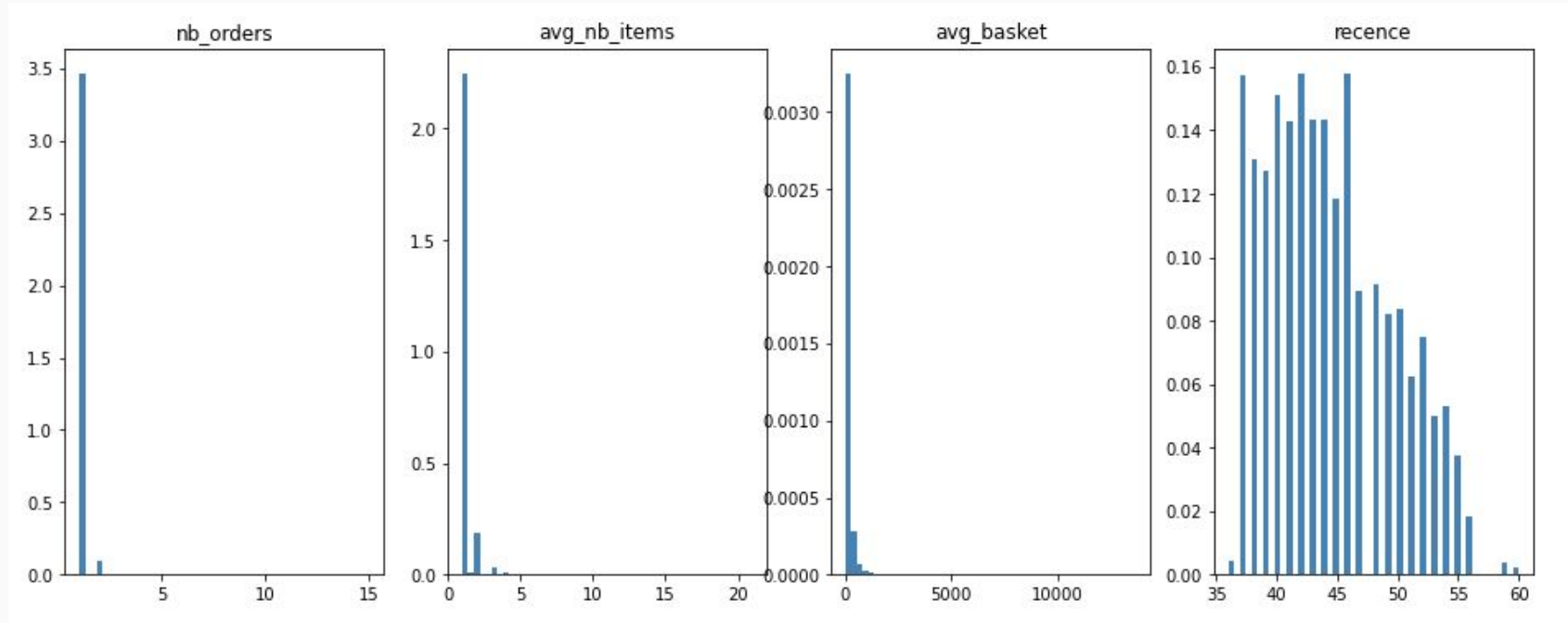
### *Exploration*

*top\_category*

bed_bath_table	
NULL	→ unknown
NULL	→ unknown
telephony	
baby	
auto	
health_beauty	
auto	

## 2. Nettoyage & exploration

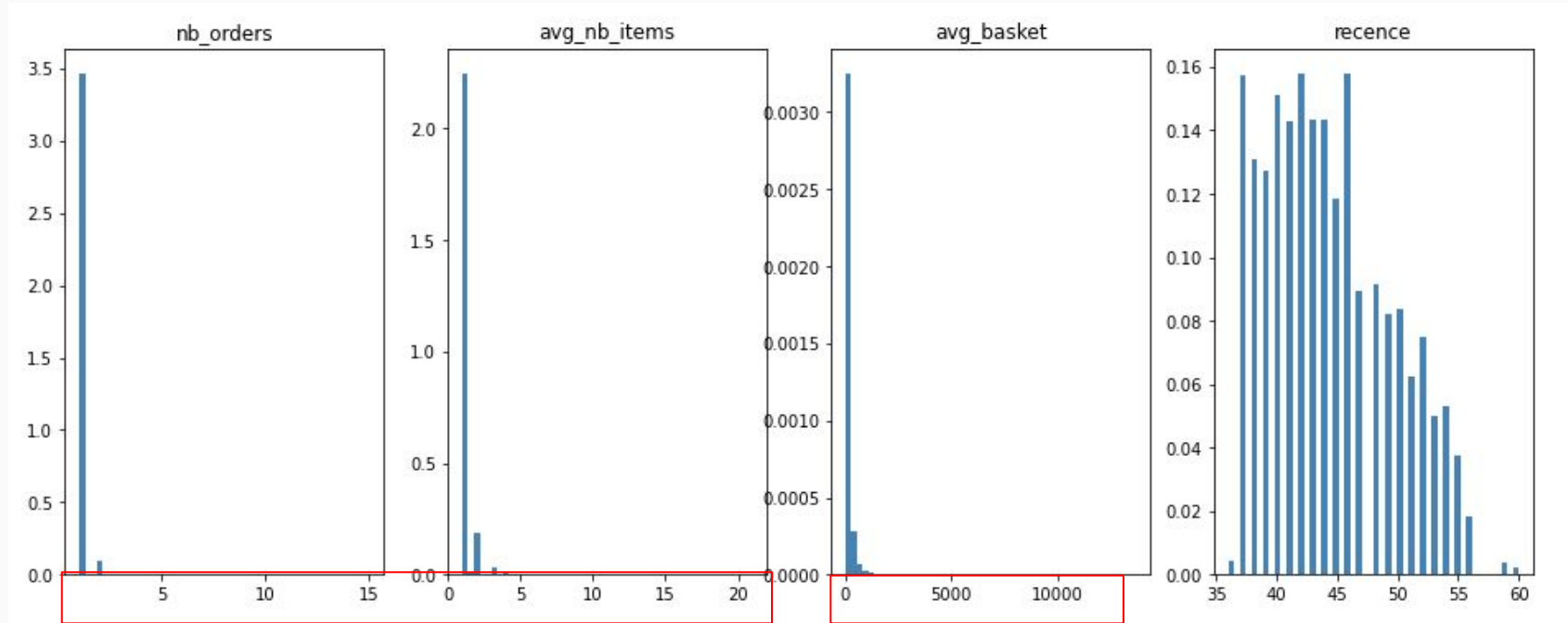
### Exploration





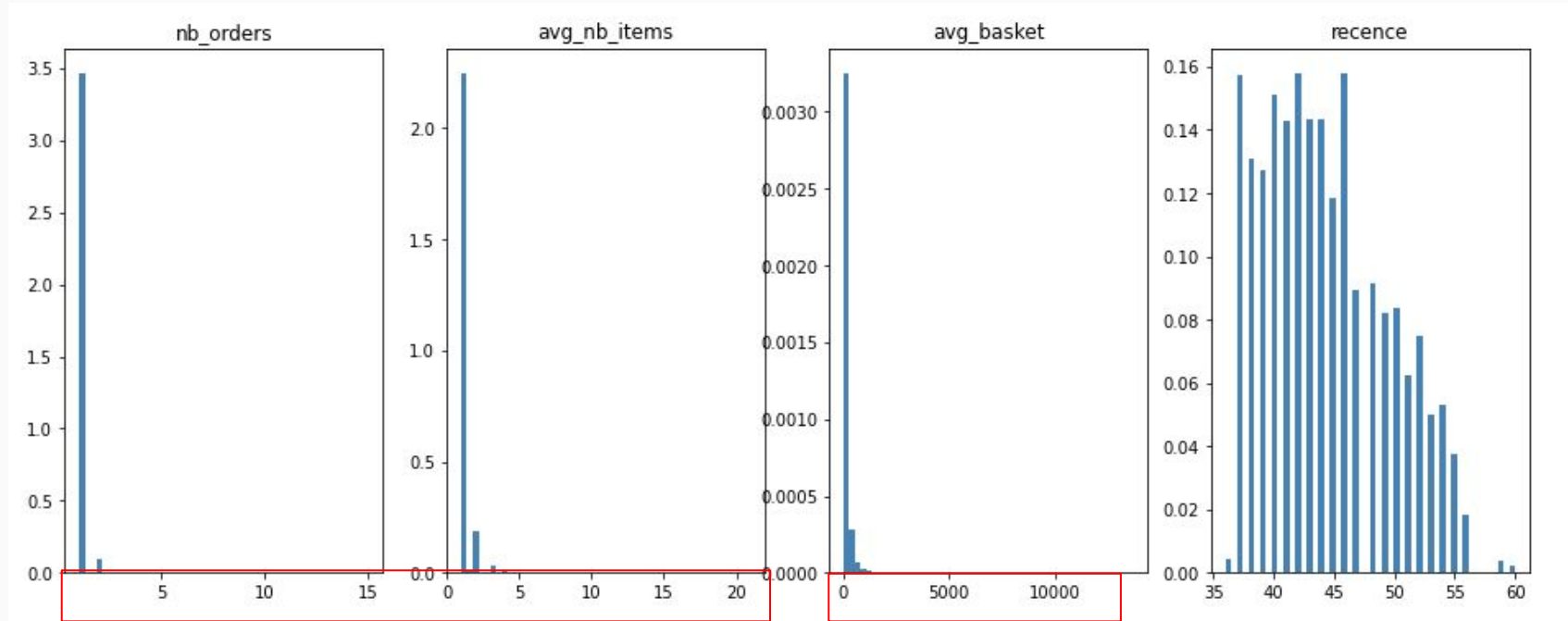
## 2. Nettoyage & exploration

### Exploration



## 2. Nettoyage & exploration

### Exploration



=> Standardisation des variables numériques

## 2. Nettoyage & exploration

### *Exploration*



1 achat

1 produit

1 commentaire

1 vendeur

Bonne satisfaction

## 2. Nettoyage & exploration

### *Exploration*



1 achat  
1 produit  
1 commentaire  
1 vendeur  
Bonne satisfaction

En moyenne

## 2. Nettoyage & exploration

### *Exploration*



1 achat

1 produit

1 commentaire

1 vendeur

Bonne satisfaction

75% des clients

## 2. Nettoyage & exploration

### *Exploration*

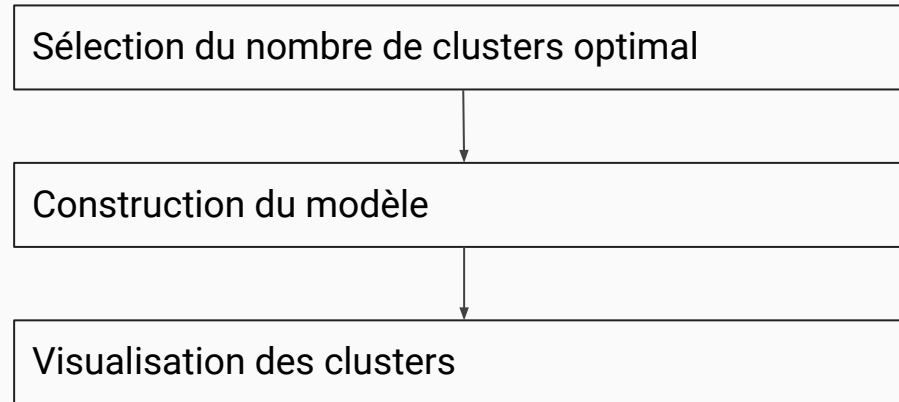


São Paulo

Rio de Janeiro

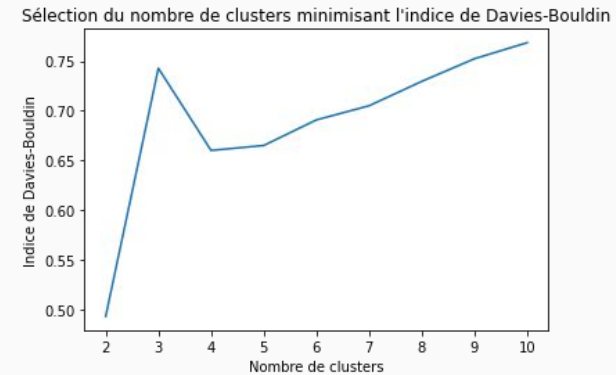
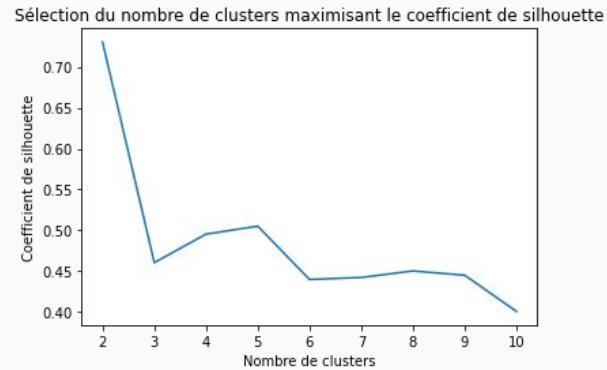
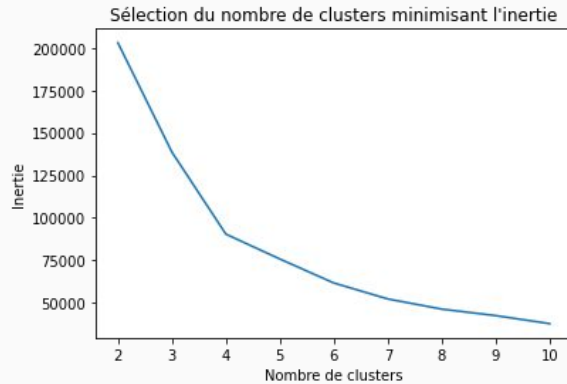
### 3. Pistes de segmentation des clients

#### *Méthodologie*



### 3. Pistes de segmentation des clients

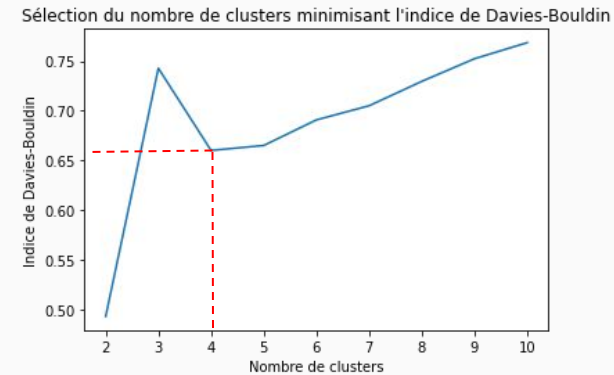
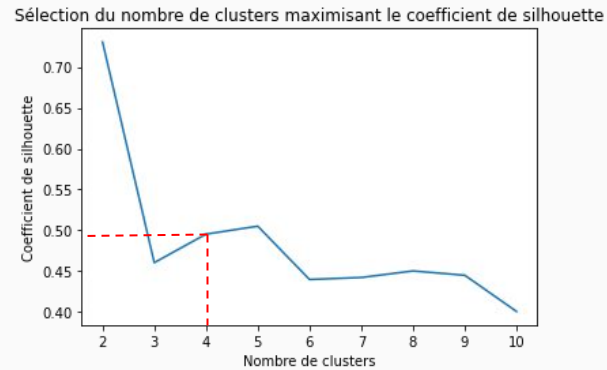
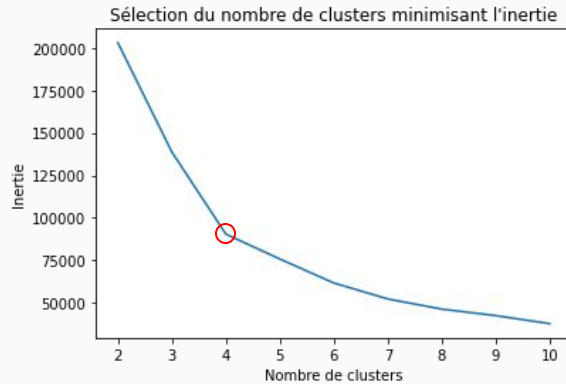
*Baseline : segmentation RFM*





### 3. Pistes de segmentation des clients

*Baseline : segmentation RFM*

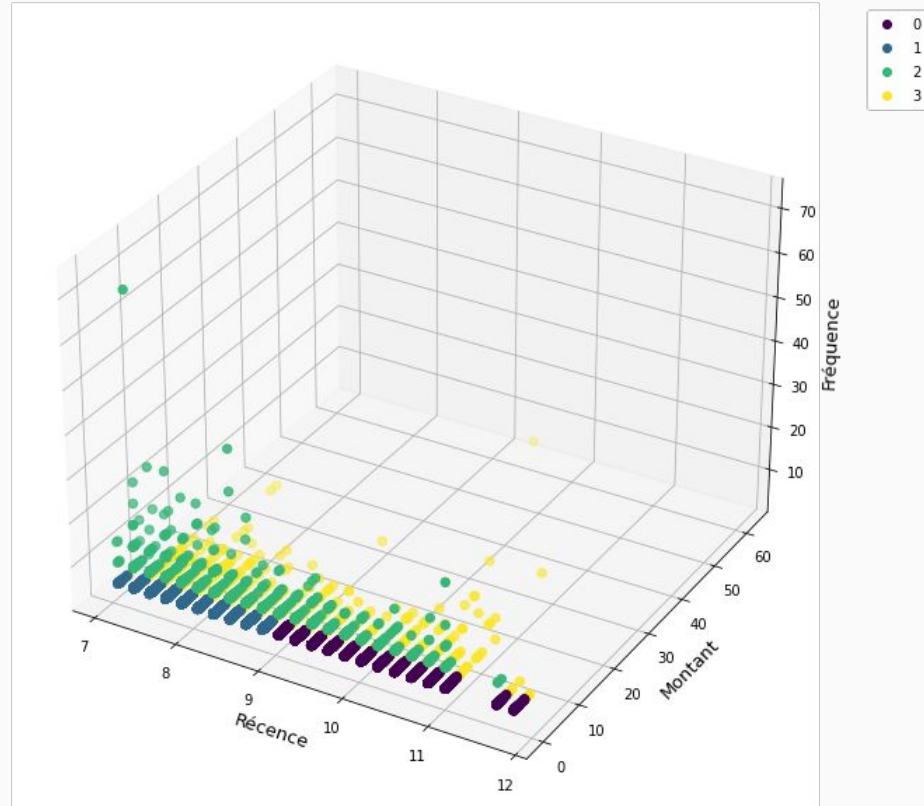


k-means avec 4 clusters avec les variables RFM

### 3. Pistes de segmentation des clients

*Baseline : segmentation RFM*

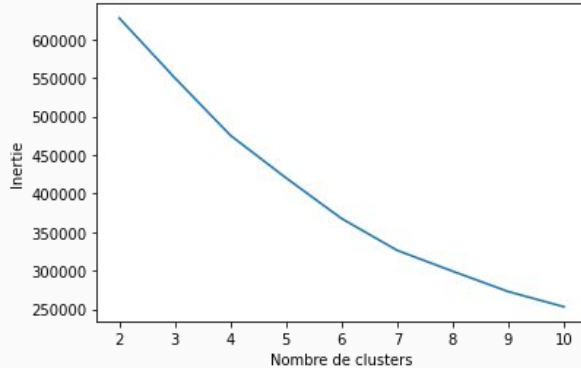
Visualisation des 93357 clients selon leur RFM



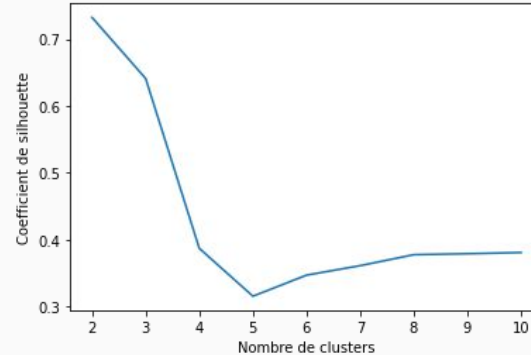
### 3. Pistes de segmentation des clients

#### *k*-means

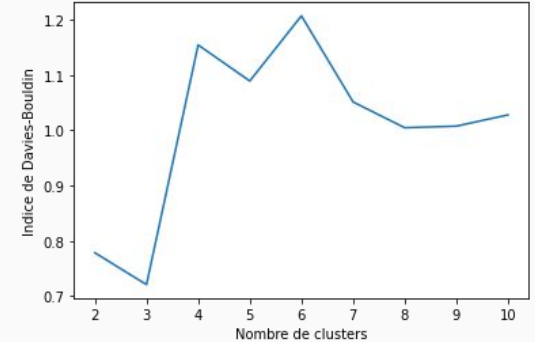
Sélection du nombre de clusters minimisant l'inertie



Sélection du nombre de clusters maximisant le coefficient de silhouette



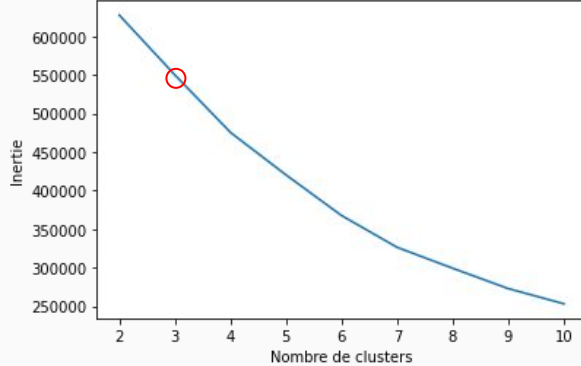
Sélection du nombre de clusters minimisant l'indice de Davies-Bouldin



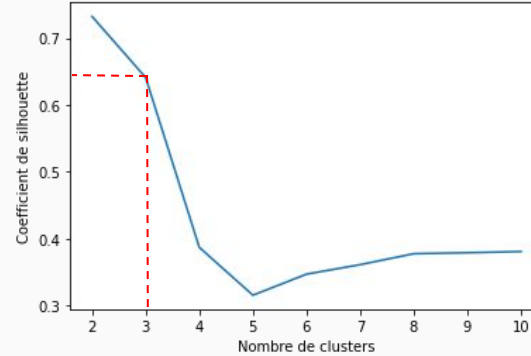
### 3. Pistes de segmentation des clients

#### *k-means*

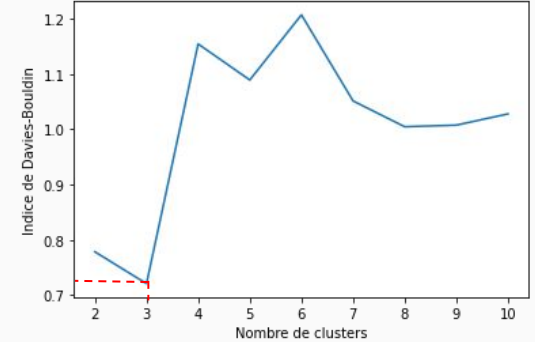
Sélection du nombre de clusters minimisant l'inertie



Sélection du nombre de clusters maximisant le coefficient de silhouette



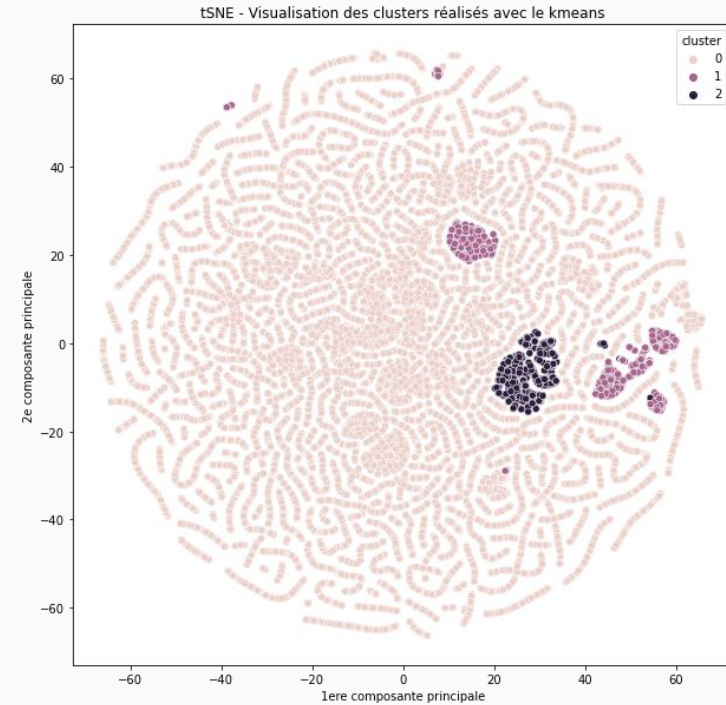
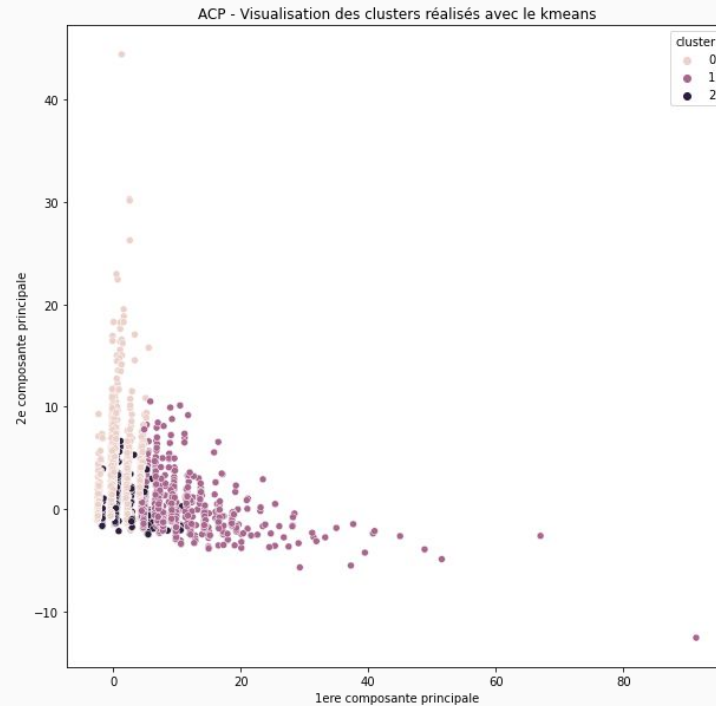
Sélection du nombre de clusters minimisant l'indice de Davies-Bouldin



k-means avec 3 clusters sur toutes les variables numériques

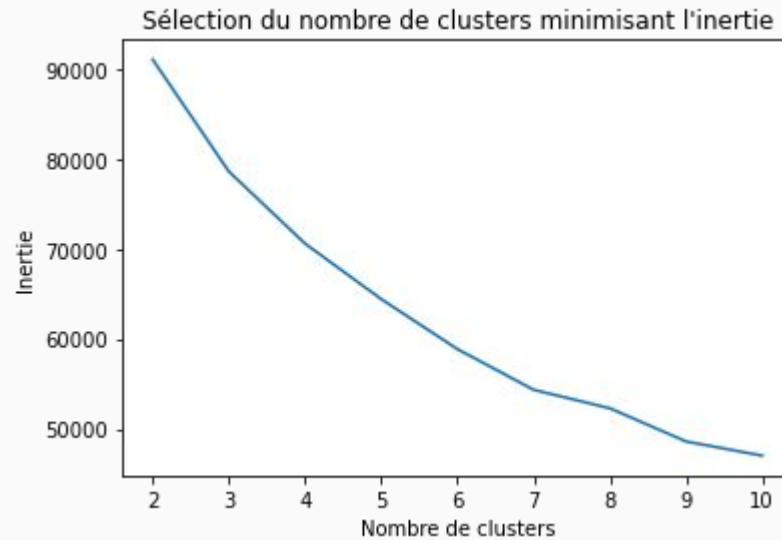
### 3. Pistes de segmentation des clients

#### *k*-means



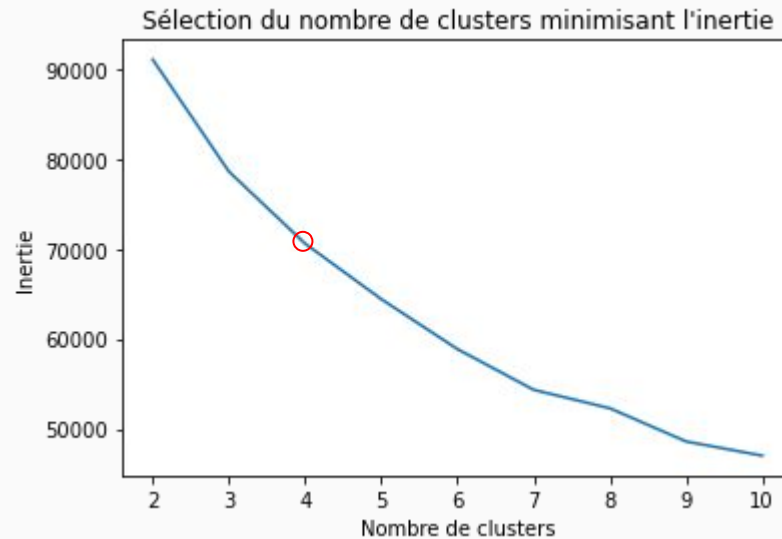
### 3. Pistes de segmentation des clients

*k*-prototypes



### 3. Pistes de segmentation des clients

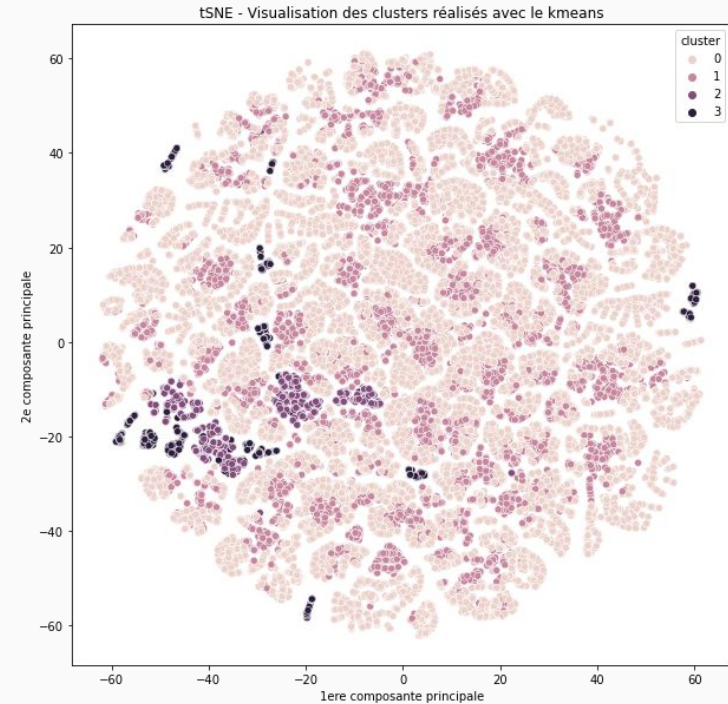
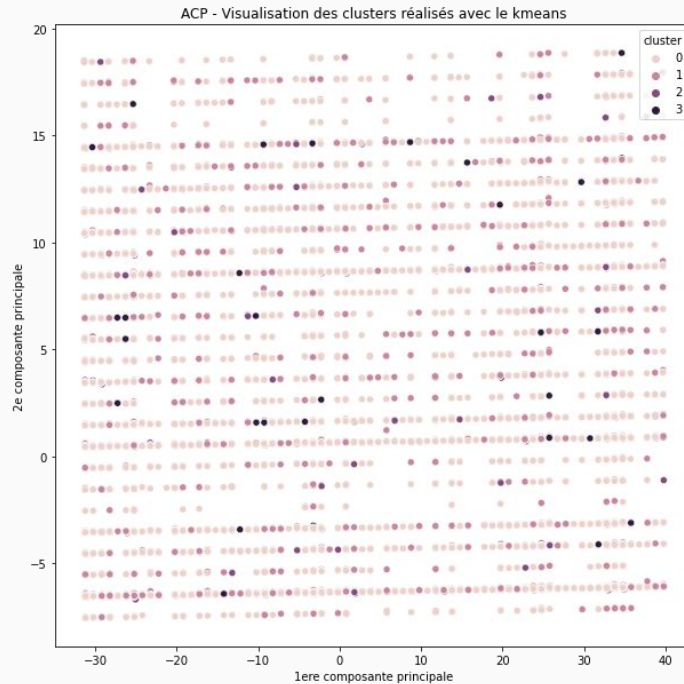
*k*-prototypes



k-prototypes avec 4 clusters sur toutes les variables

### 3. Pistes de segmentation des clients

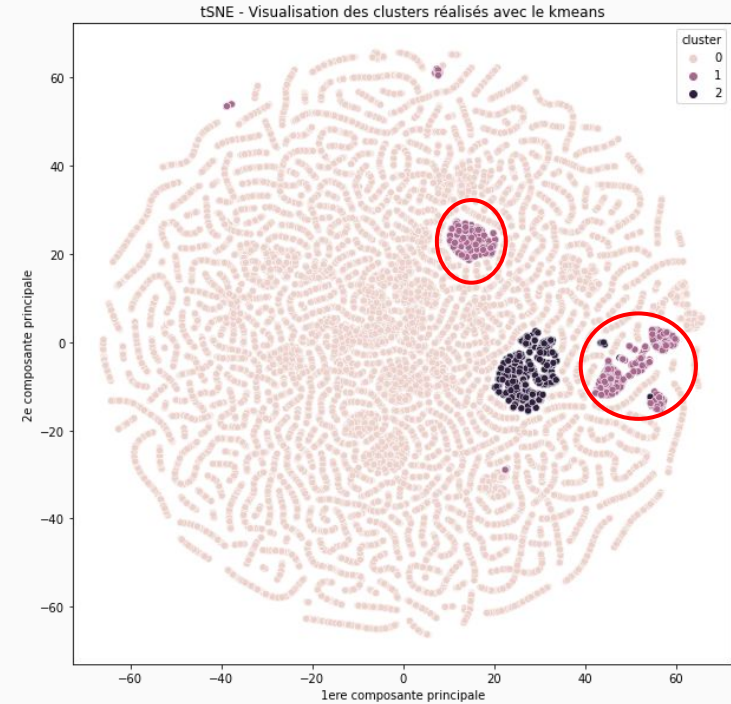
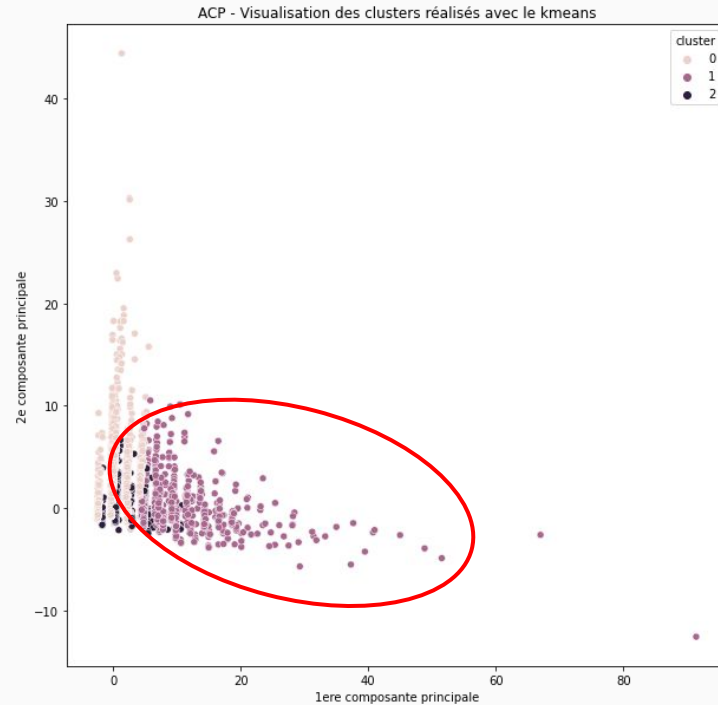
*k*-prototypes





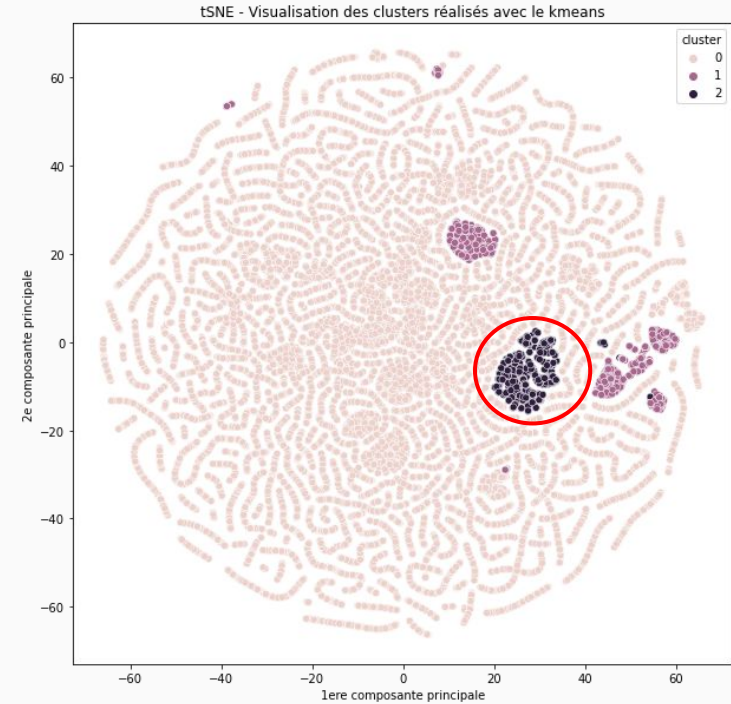
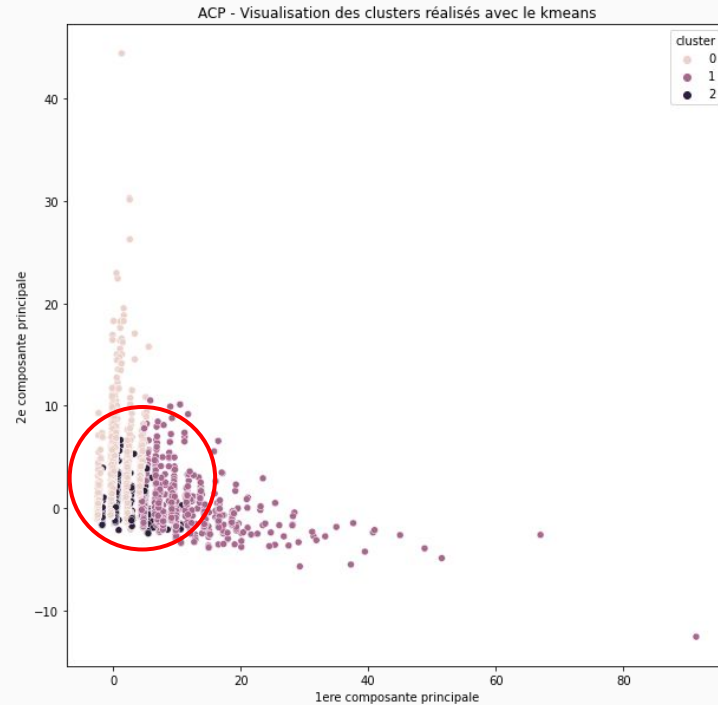
## 4. Modèle final sélectionné

### *k-means sur les variables numériques*



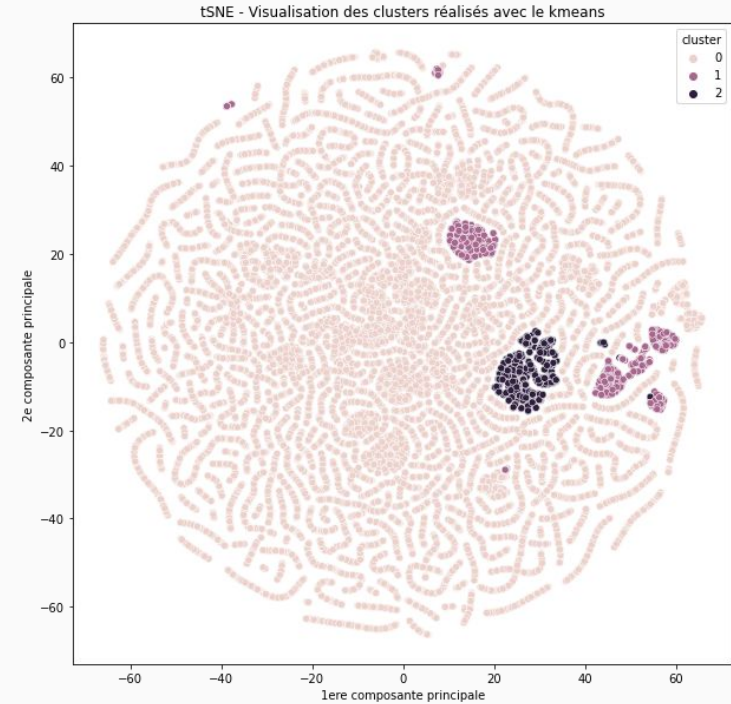
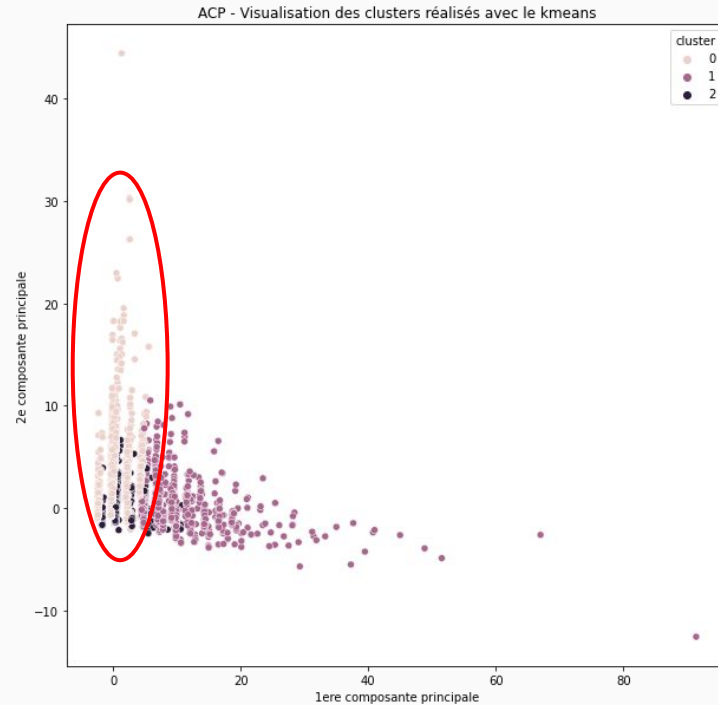
## 4. Modèle final sélectionné

### *k-means sur les variables numériques*



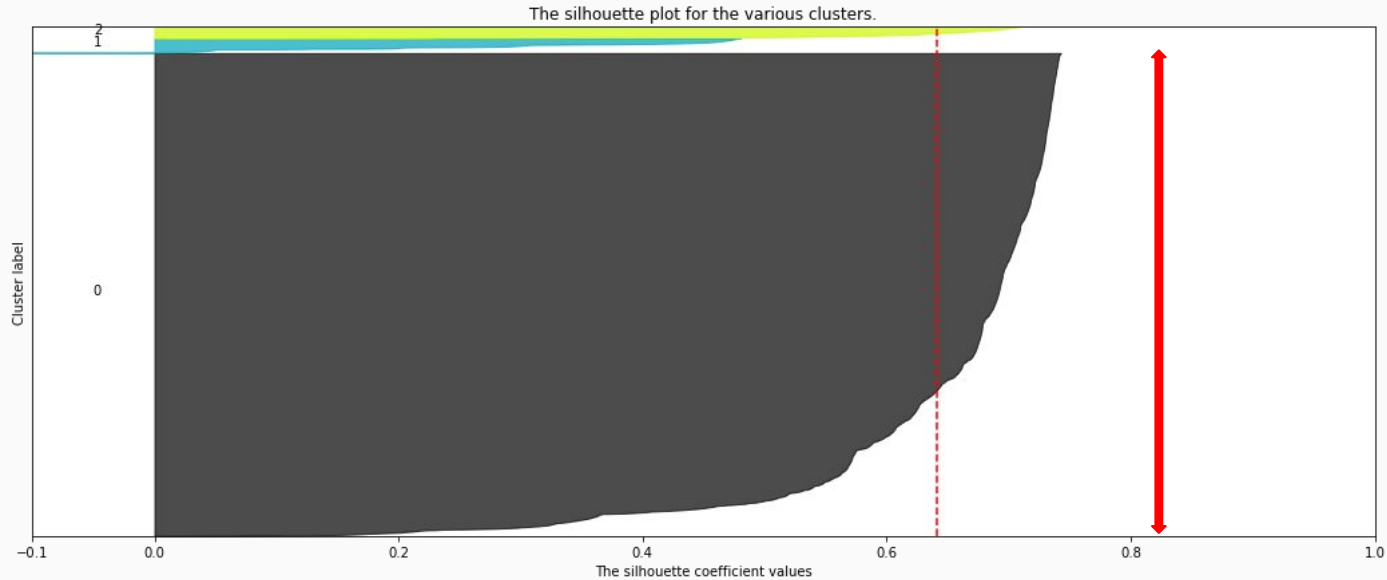
## 4. Modèle final sélectionné

### *k*-means sur les variables numériques



## 4. Modèle final sélectionné

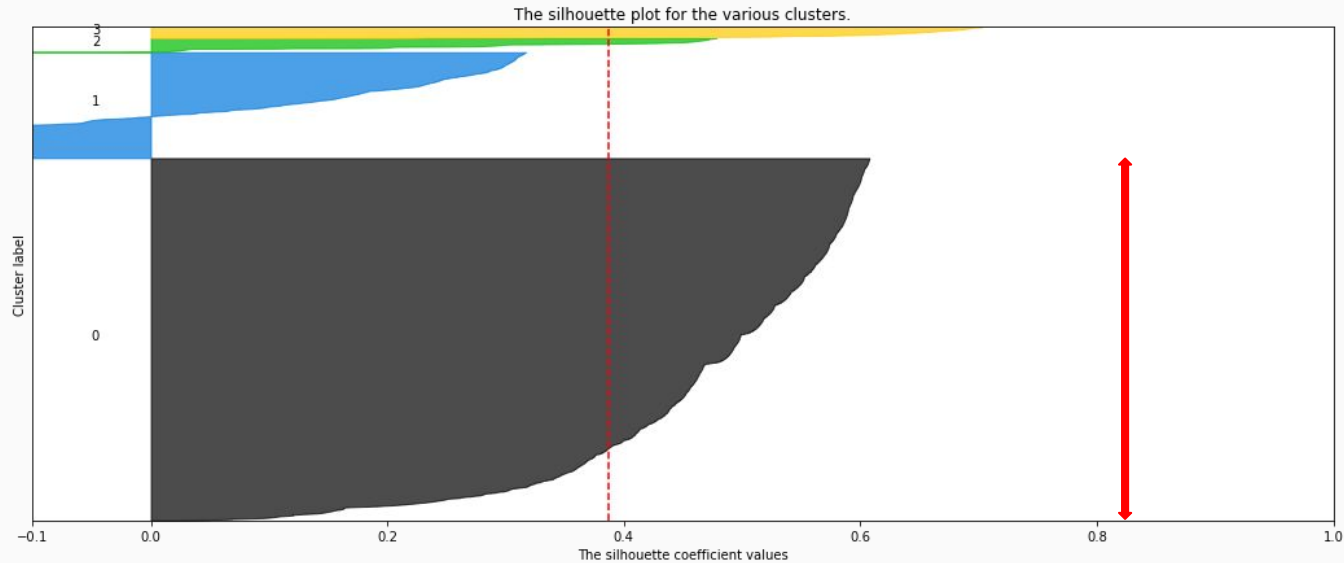
*k-means sur les variables numériques*



Surreprésentation

## 4. Modèle final sélectionné

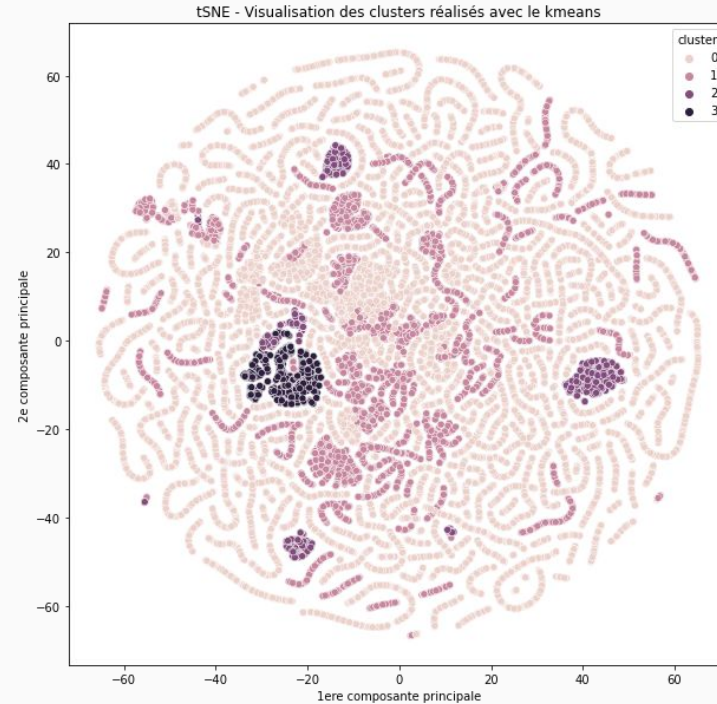
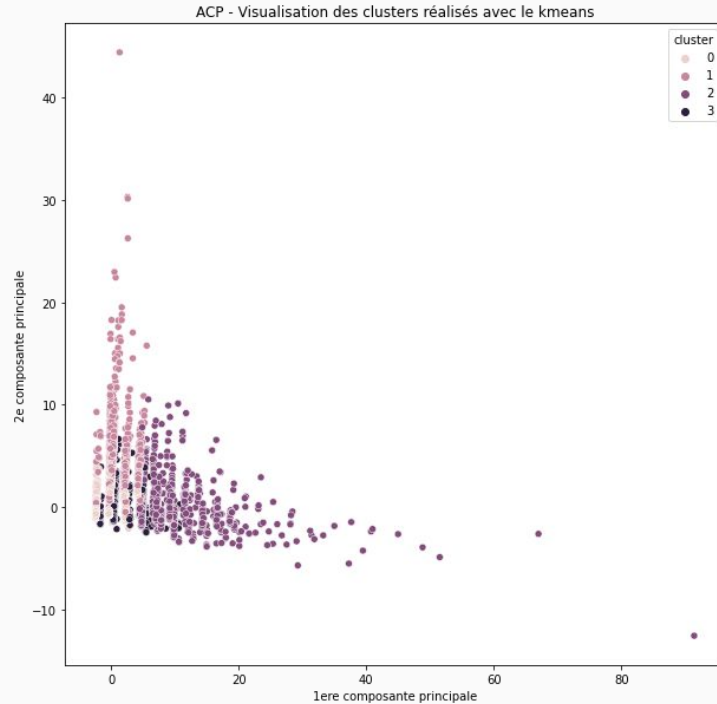
*k-means sur les variables numériques*



Meilleur équilibre

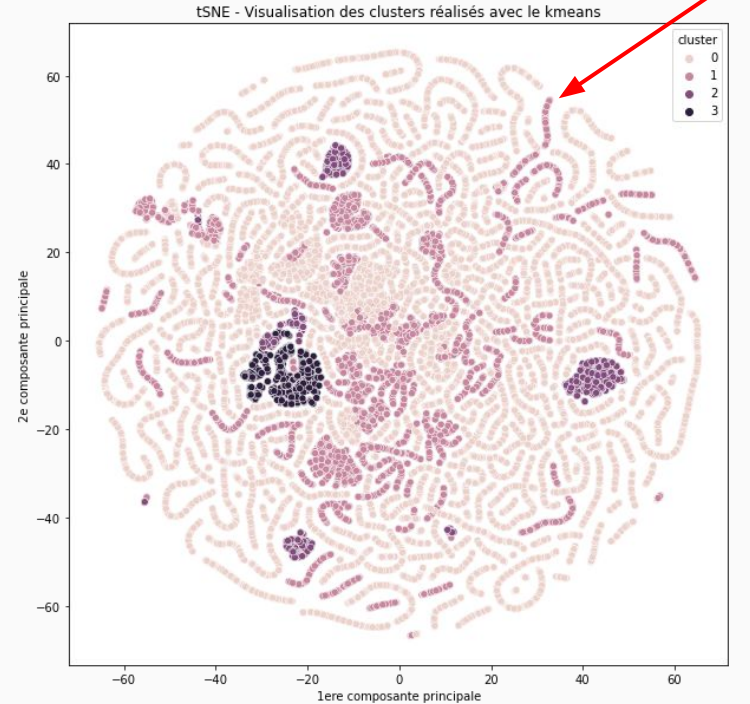
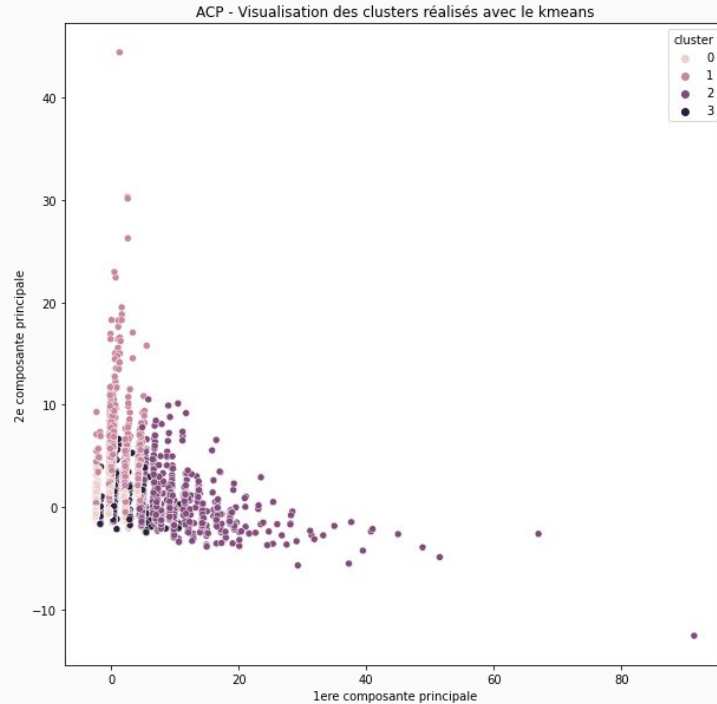
## 4. Modèle final sélectionné

### *k*-means sur les variables numériques



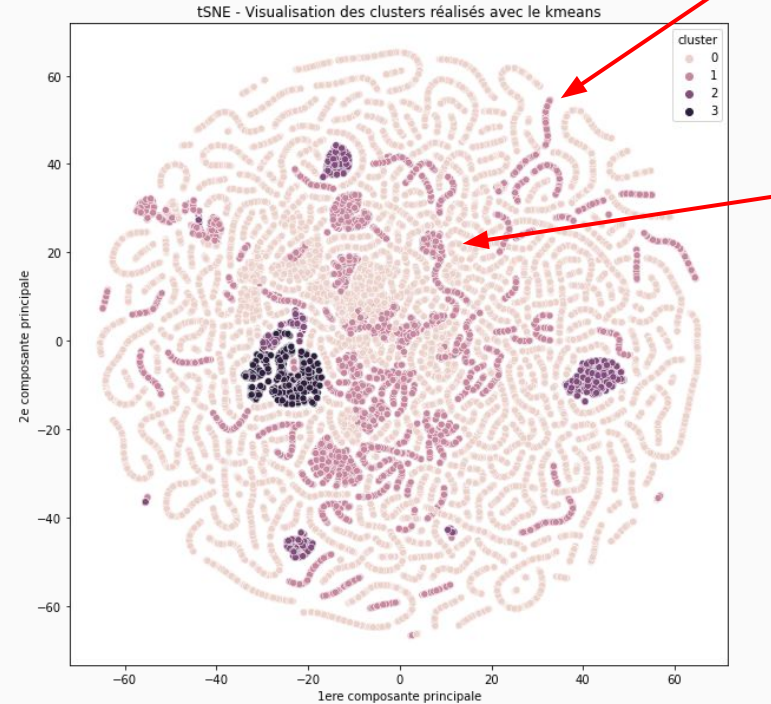
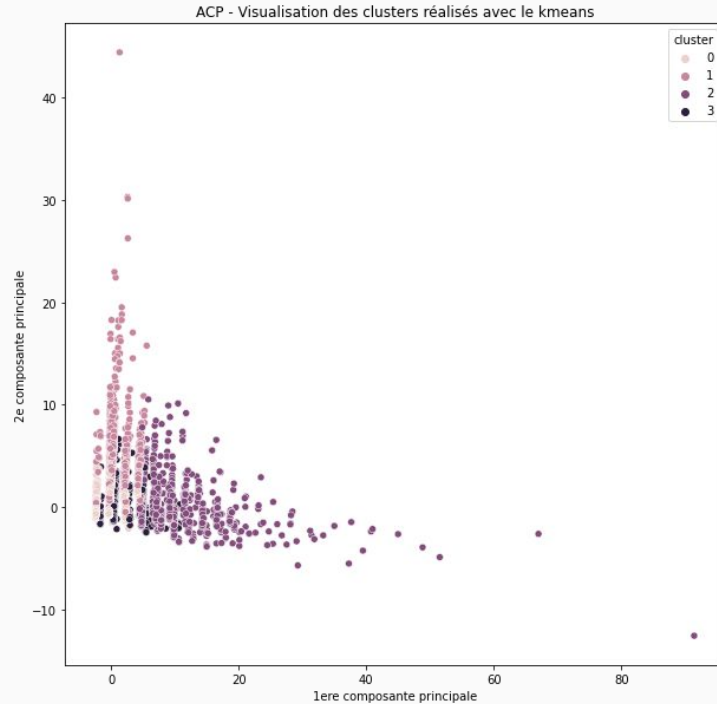
## 4. Modèle final sélectionné

### *k*-means sur les variables numériques



## 4. Modèle final sélectionné

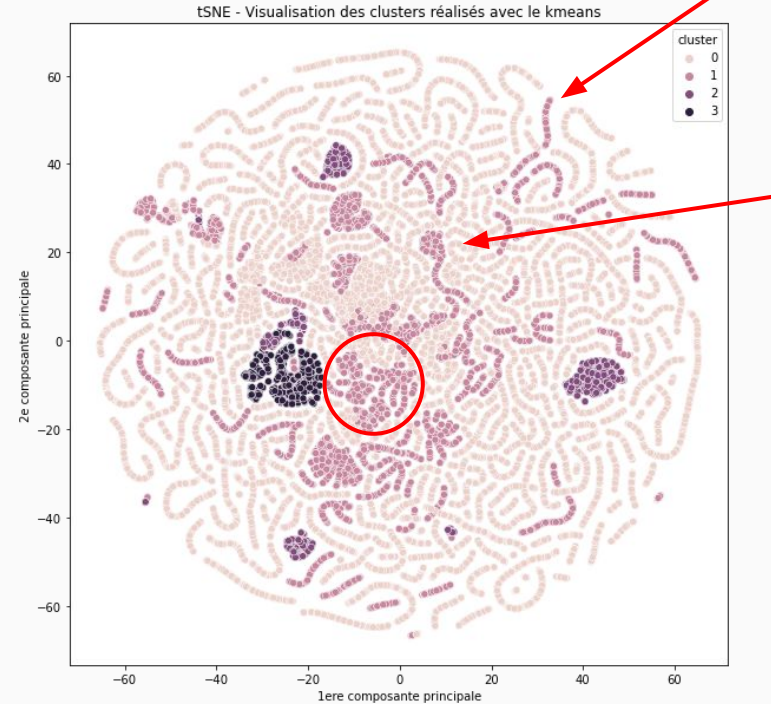
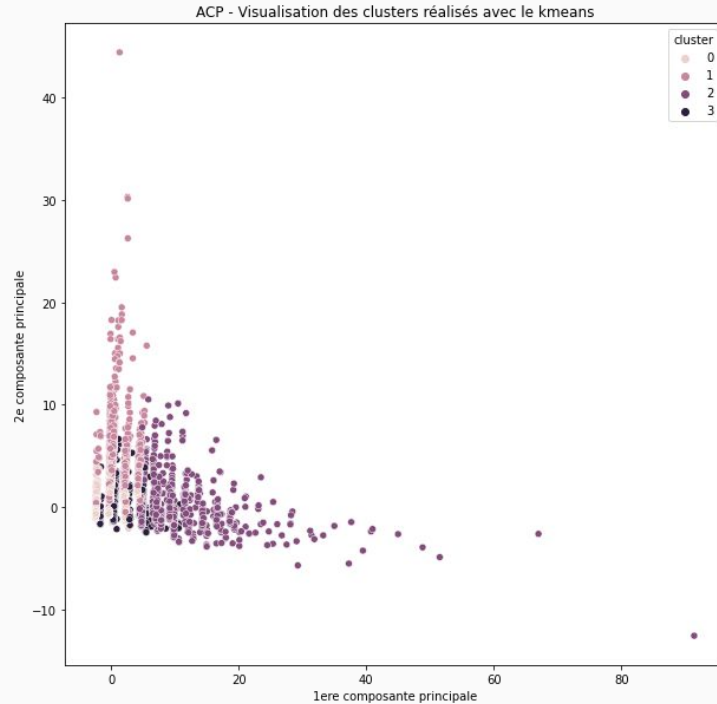
### *k-means sur les variables numériques*





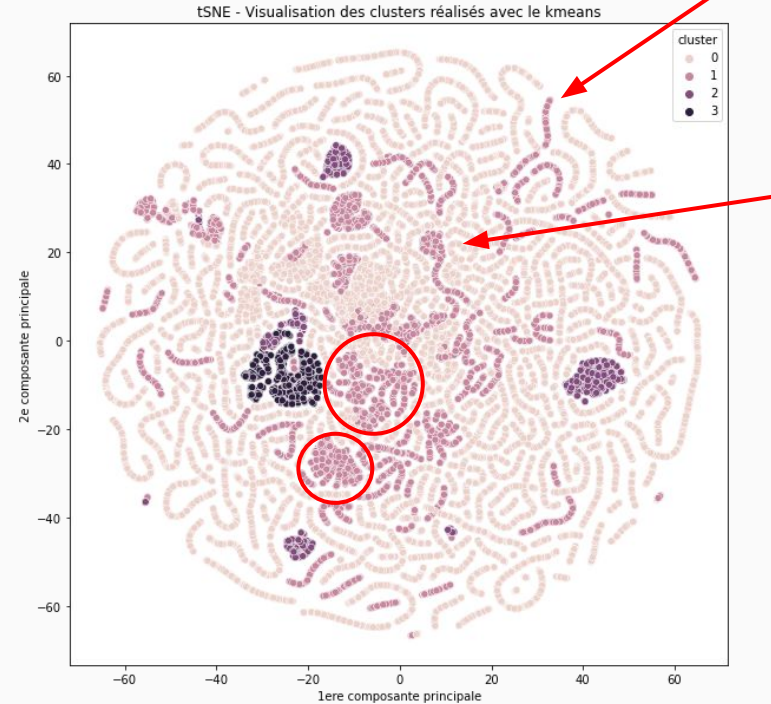
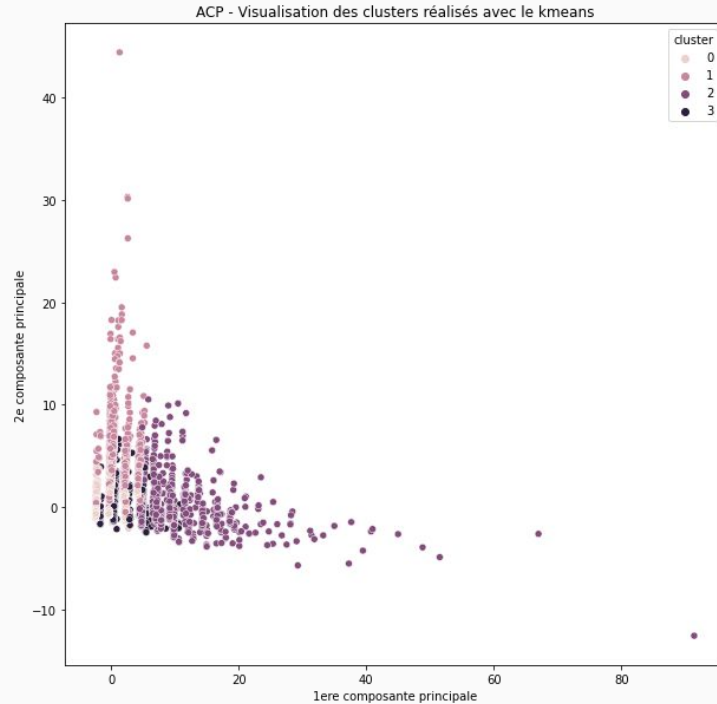
## 4. Modèle final sélectionné

### *k*-means sur les variables numériques



## 4. Modèle final sélectionné

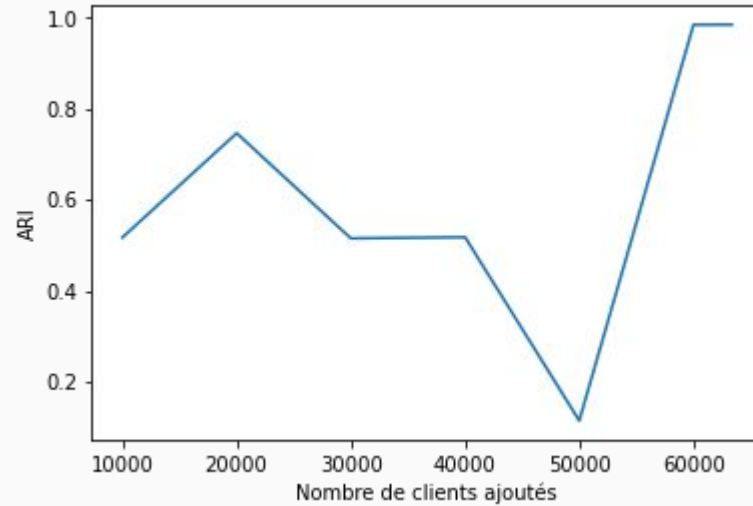
### *k*-means sur les variables numériques



## 4. Modèle final sélectionné

*k-means sur les variables numériques*

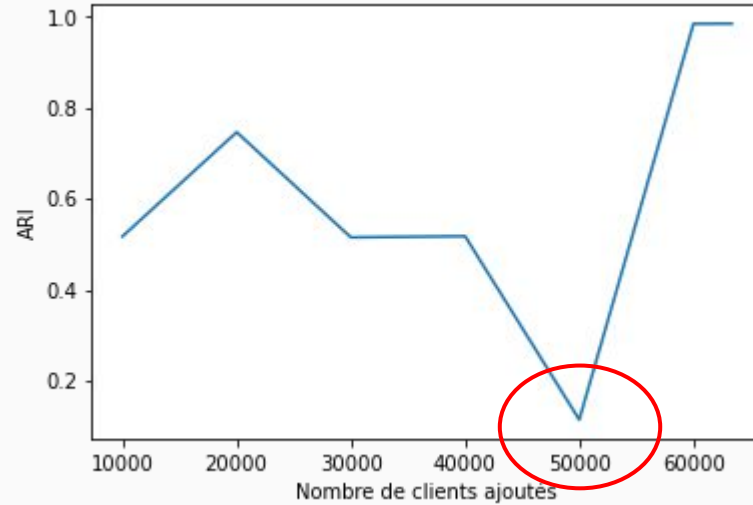
Evolution de l'ARI selon le nombre de clients ajoutés dans la segmentation



## 4. Modèle final sélectionné

*k-means sur les variables numériques*

Evolution de l'ARI selon le nombre de clients ajoutés dans la segmentation



# Questions/Réponses

Fin.