Projet n°5

Segmentation de clients d'un site de e-commerce

OLIST

OPENCLASSROOMS

Sommaire

INTRODUCTION

Mission et objectifs

L'objectif de l'entreprise OLIST est d'optimiser ses campagnes de communication marketing.
Notre mission ici, est de segmenter les clients en fonction de certains critères (RFM).

PROCÉDURE :

- Récupération des données (SQL)
- Analyse exploratoire des données
- Clustering
- Maintenance

État des lieux

Après récupération, notre jeu de données est un fichier `.csv`, que l'on nommera `data` et dont les caractéristiques sont les suivantes :

Information	**Valeur**
Nombre de lignes	94721
Nombre de colonnes	8
Nombre de colonnes `float`	2
Nombre de colonnes `object`	4
Nombre de colonnes `int`	2

Table: Résumé descriptif de `data`

ANALYSE EXPLORATOIRE DES DONNÉES

Description des données

- Chaque client est identidée par son **ID**
- Variables catégorielles
  - Ville
  - État
  - Date
- Variables numériques
  - Code postal
  - Montant
  - Nombre d'achats
  - Satisfaction
- Aucun doublon détecté dans le jeu de données `data`

Création d'une variable

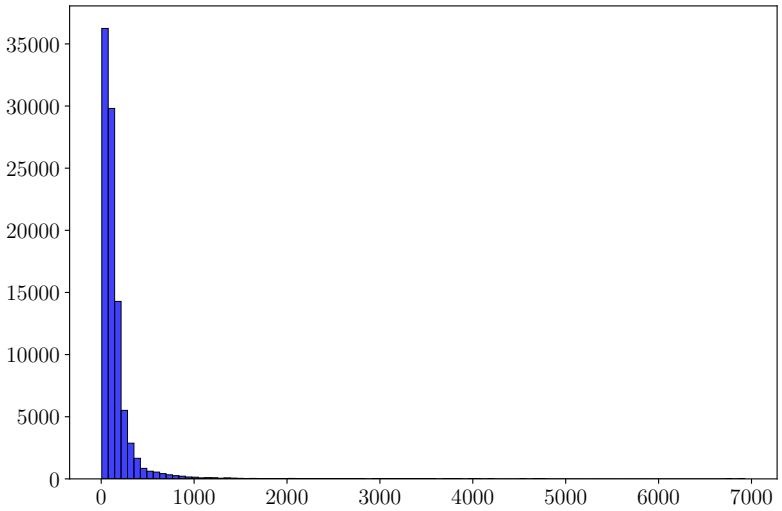
- Modification du type de la variable `date`
- Création d'une variable `timestamp`
- Création de la variable `recence`

$$\texttt{recence} = \frac{\max\left(\texttt{timestamp}\right) - \texttt{timestamp}}{24 \times 3600}$$

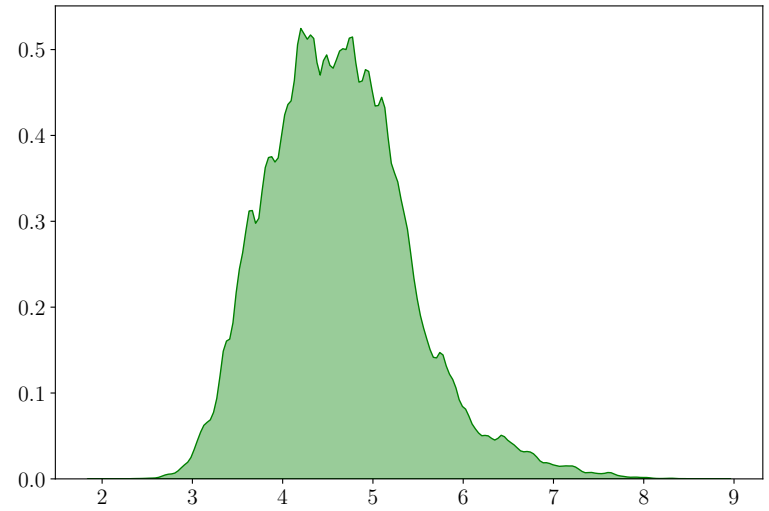
Notre variable `recence` est donc exprimée en jours, et est égale à 0 pour la date d'achat la plus récente.

Distributions (`float`)

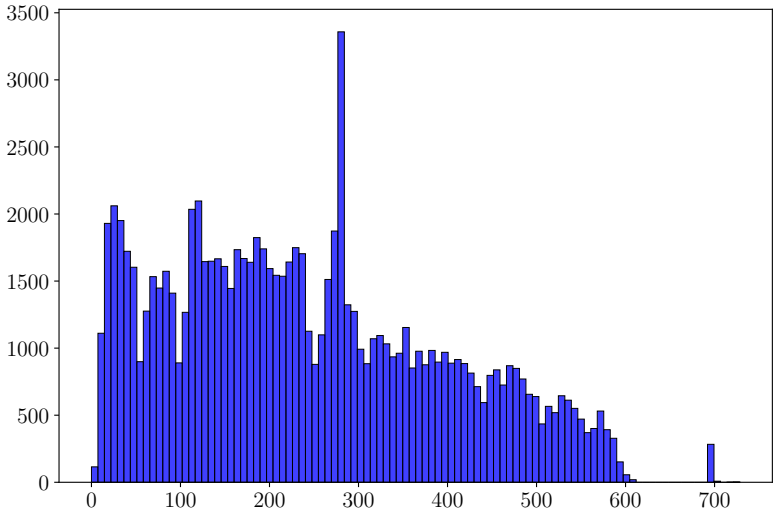
Montant – Normal



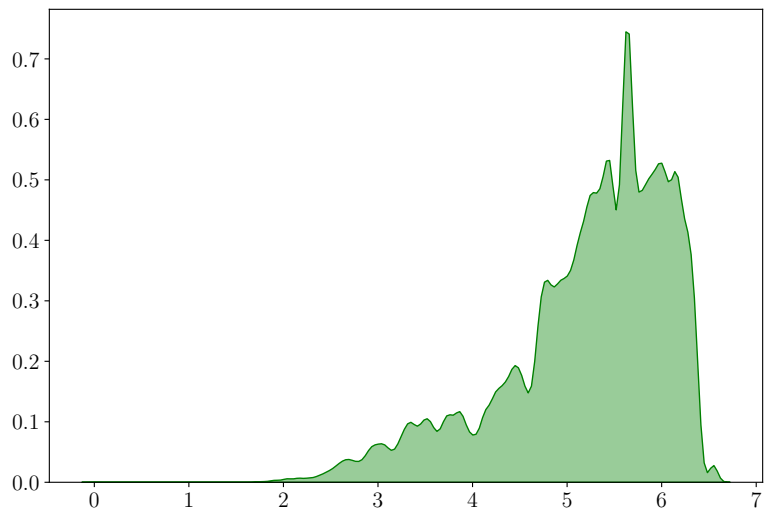
Montant – Logarithmique



Recence – Normal

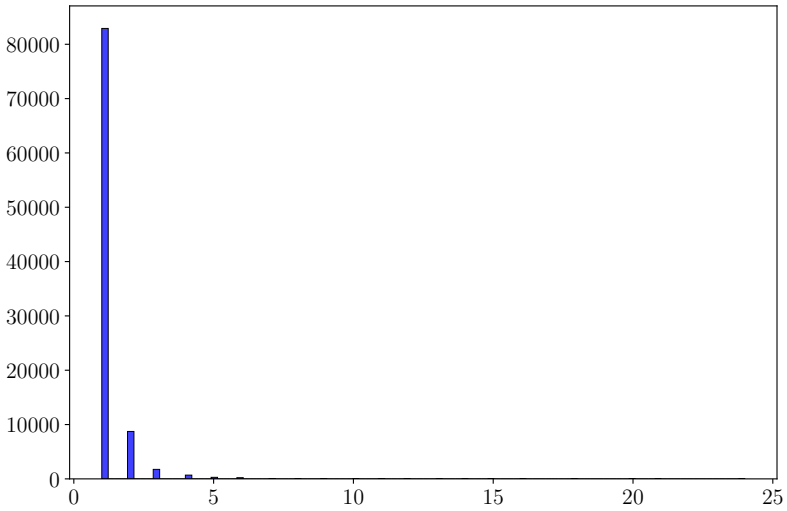


Recence – Logarithmique

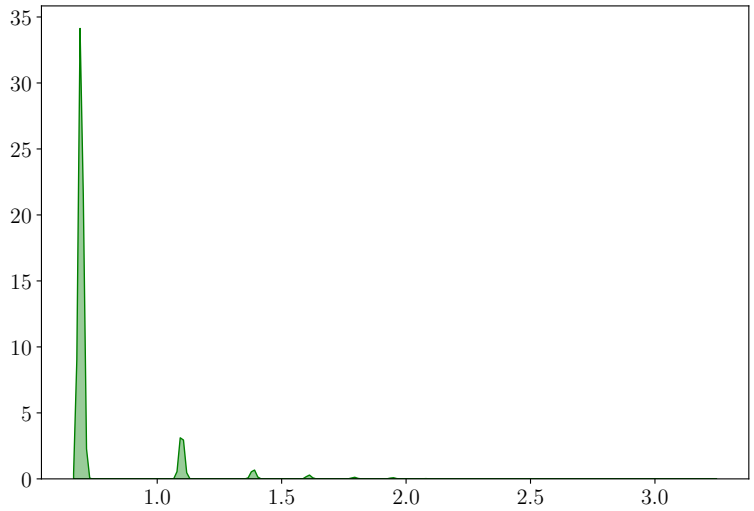


Distributions (`int`)

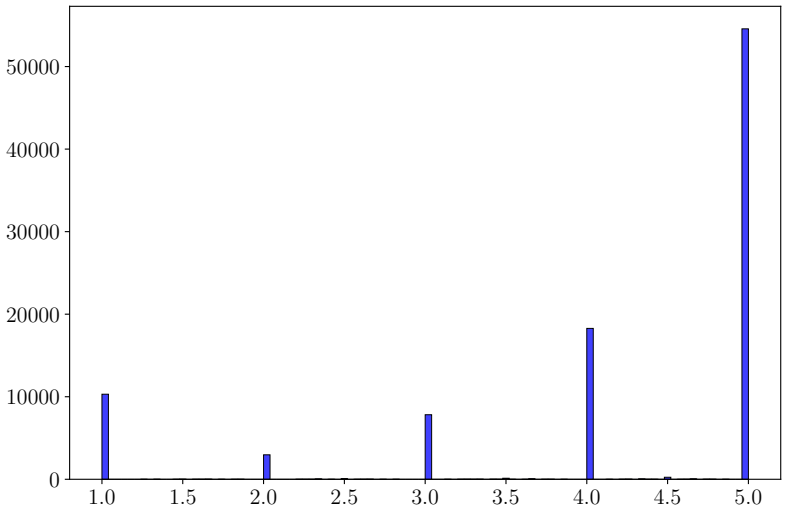
Nb achats – Normal



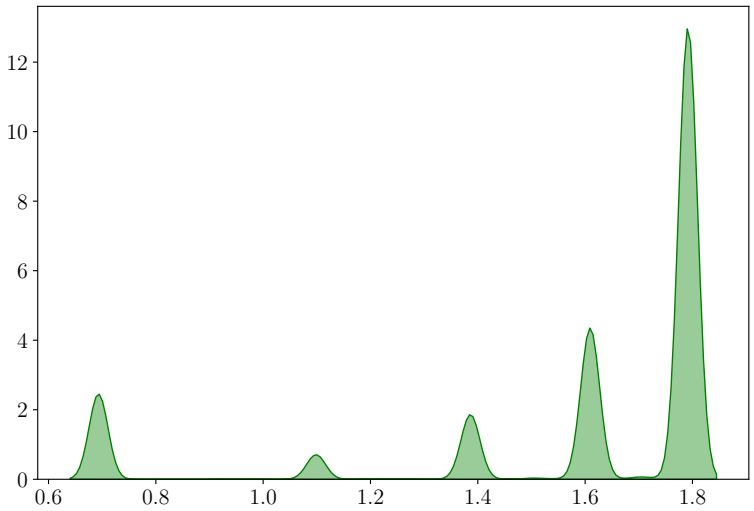
Nb achats – Logarithmique



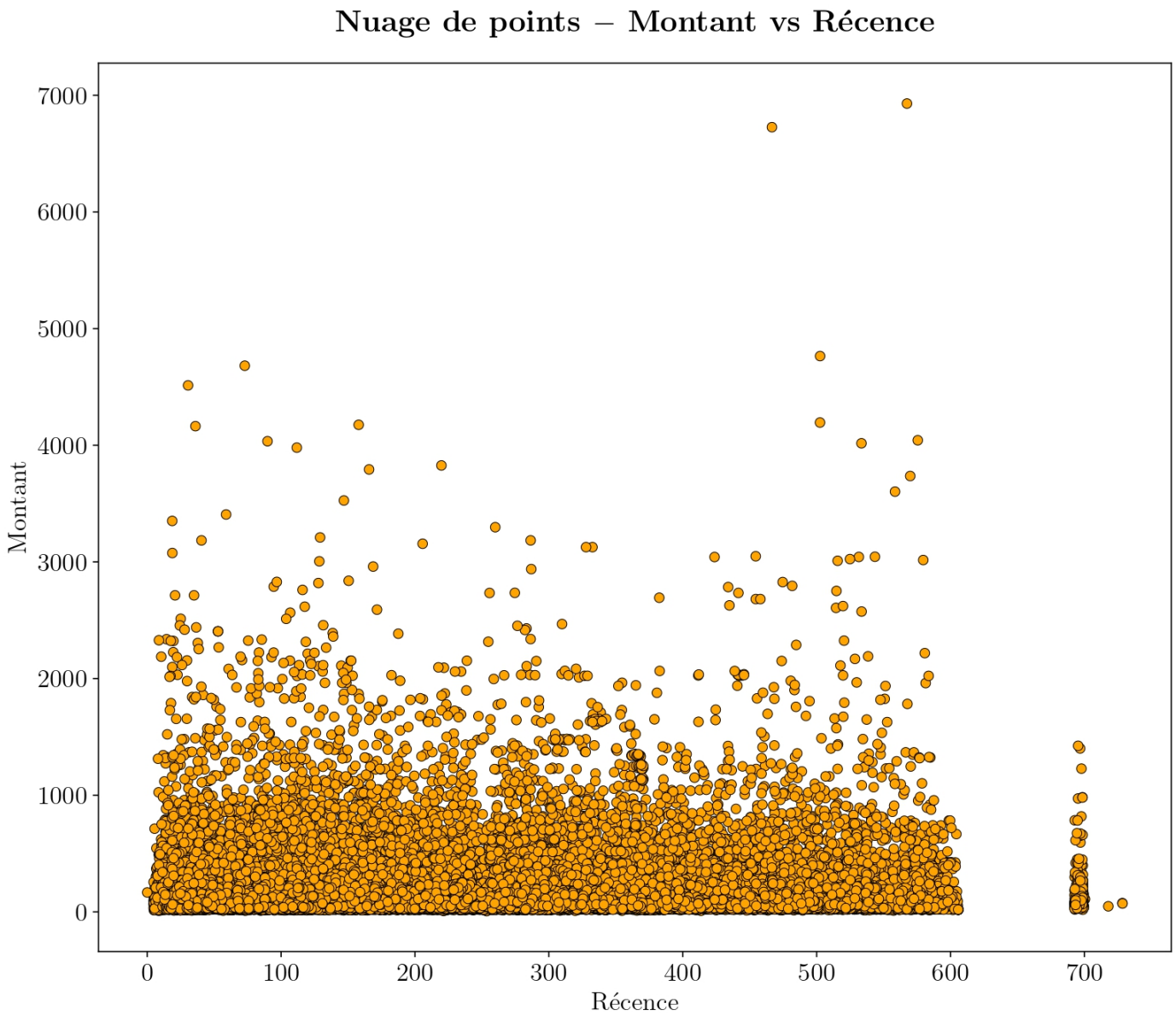
Satisfaction – Normal

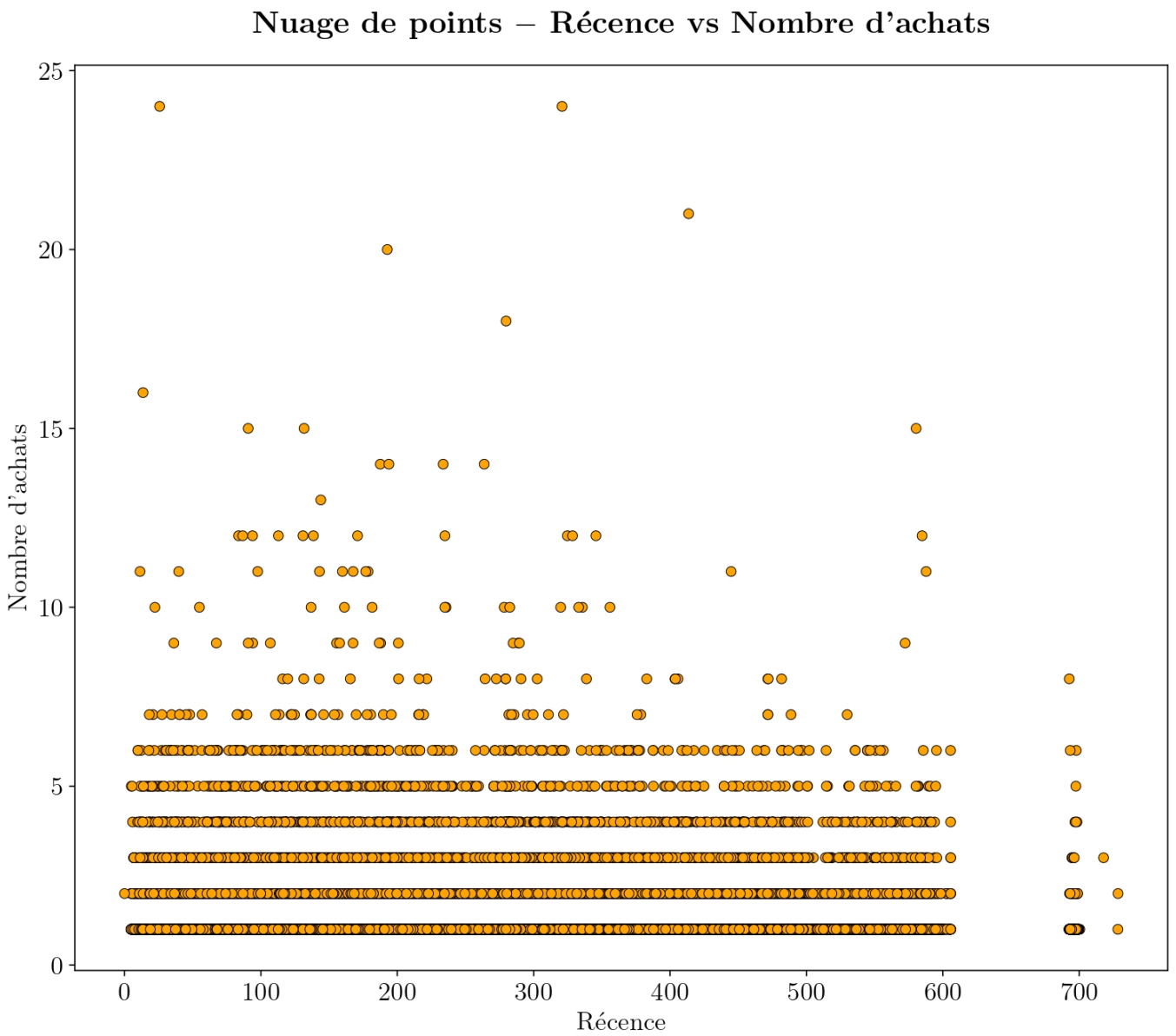


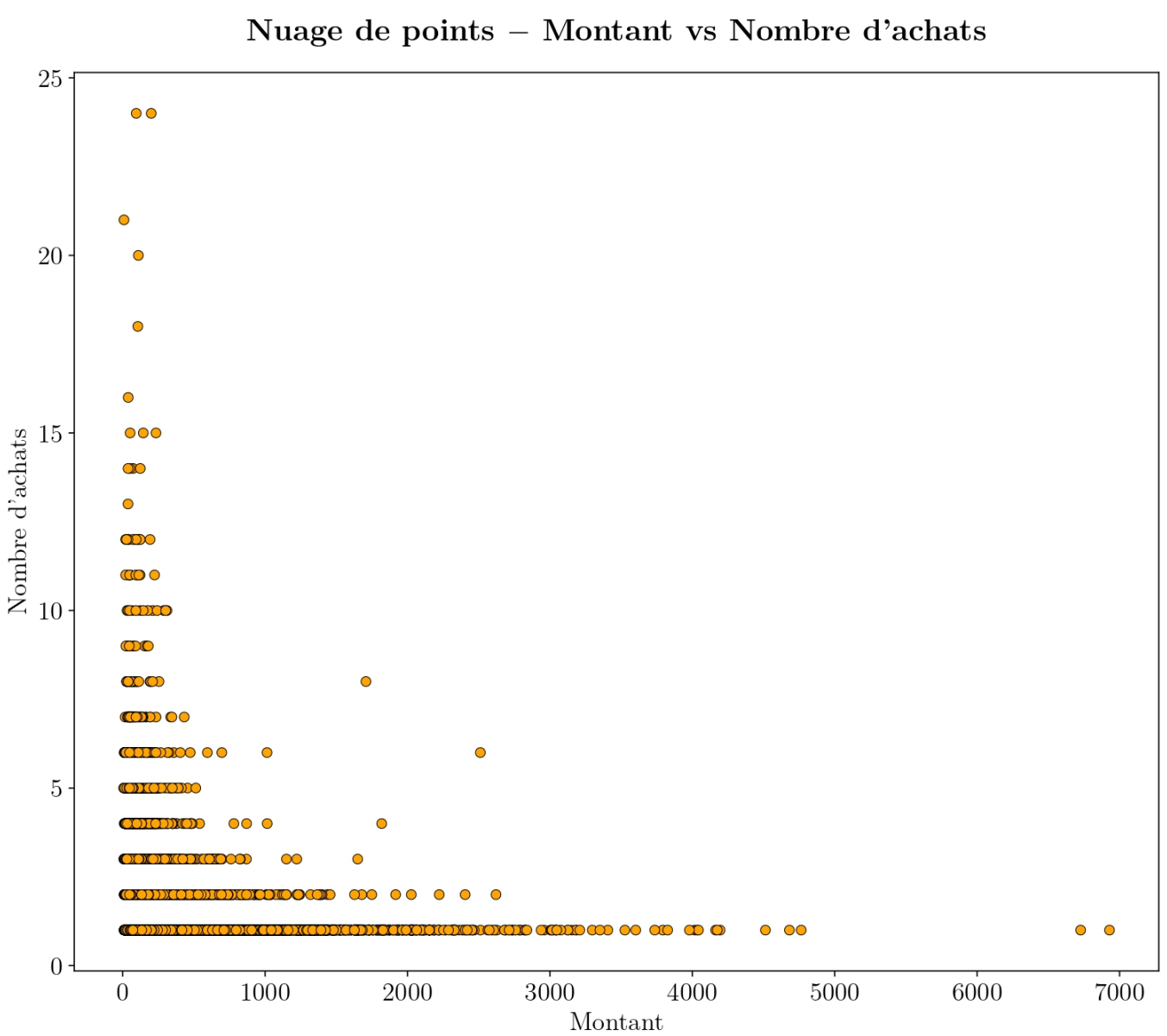
Satisfaction – Logarithmique



Analyse multivariée




Nuage de points – Récence vs Nombre d'achats
Nombre d'achats
Récence
0
5
10
15
20
25
0
100
200
300
400
500
600
700


Nuage de points – Montant vs Nombre d'achats
Nombre d'achats
Montant
0
5
10
15
20
25
0
1000
2000
3000
4000
5000
6000
7000

CLUSTERING N° 1
KMEANS

Feature Engeneering

- Création de la variable `prix_moyen_achat`

$$\texttt{prix\_moyen\_achat} = \frac{\texttt{montant}}{\texttt{nb\_achats}}$$

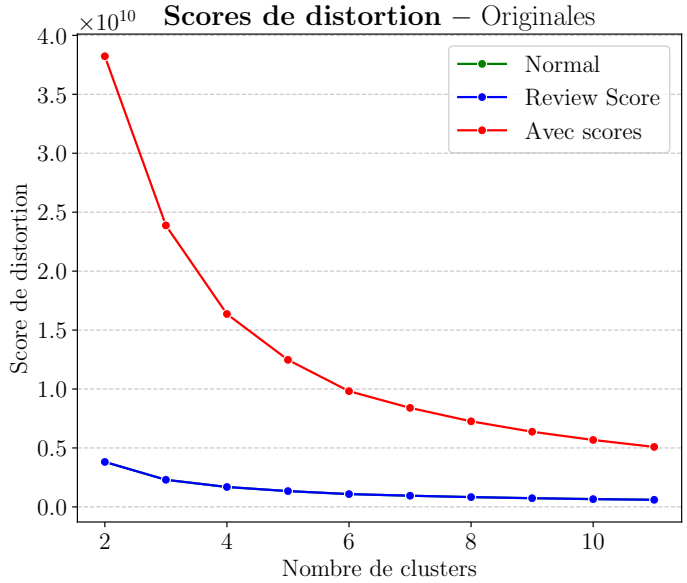
- Création de la variable `montant_satisfaction`

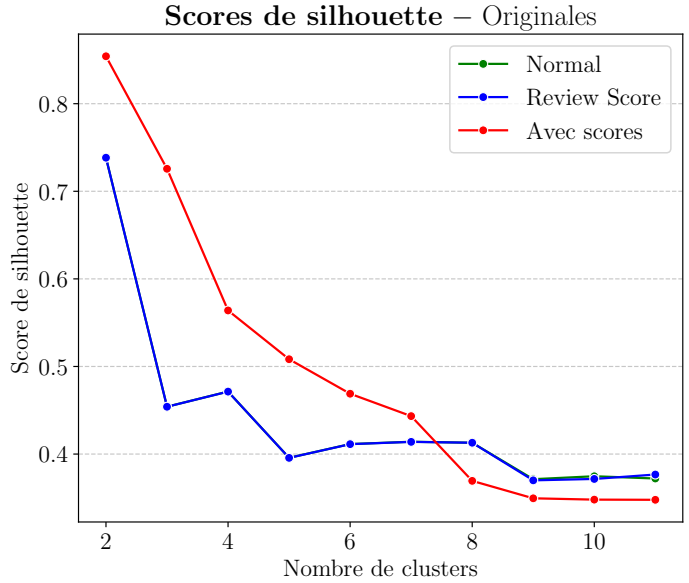
$$\texttt{montant\_satisfaction} = \texttt{montant} \times \texttt{satisfaction}$$

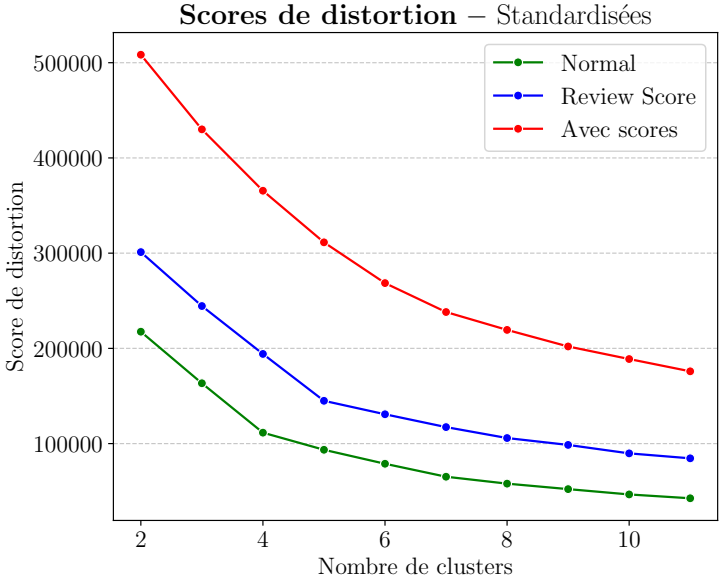
- Création de la variable `frequence_achats`

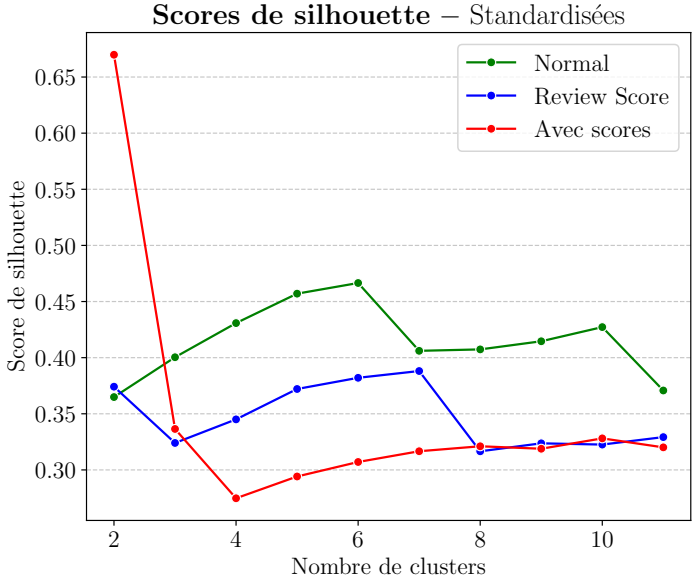
$$\texttt{frequence\_achats} = \frac{\texttt{nb\_achats}}{\texttt{recence} + 1}$$

Optimisation du clustering

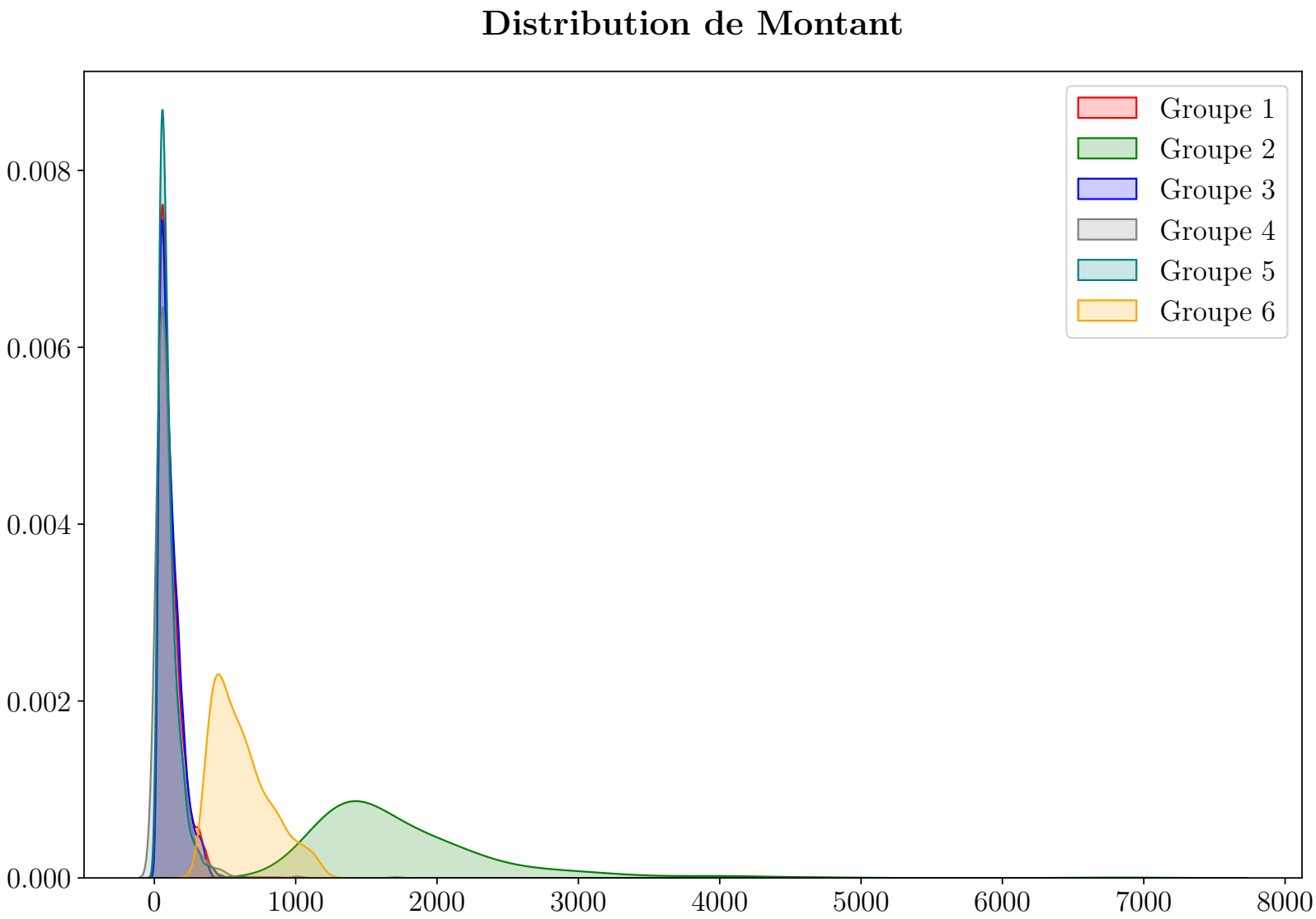




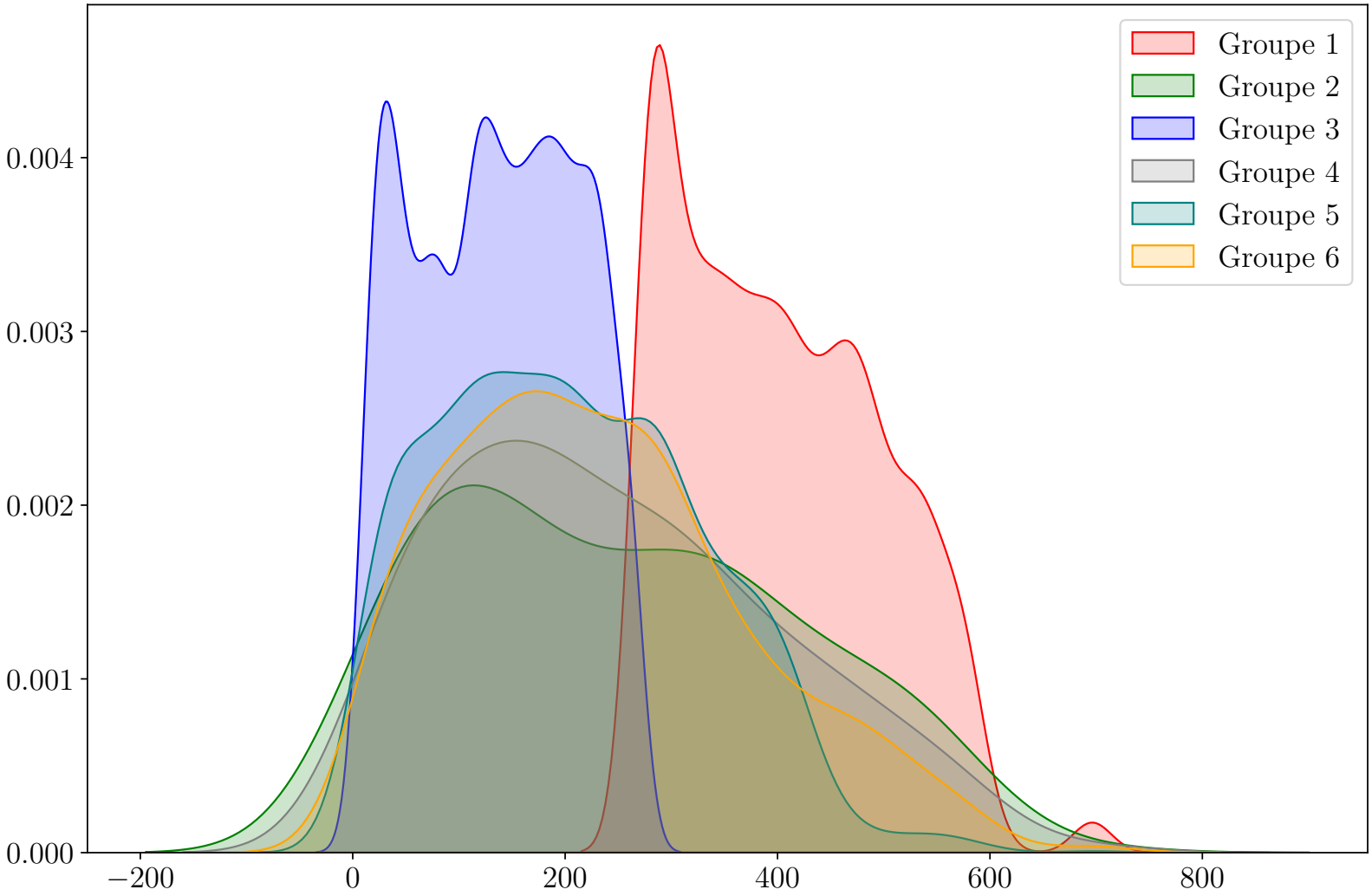


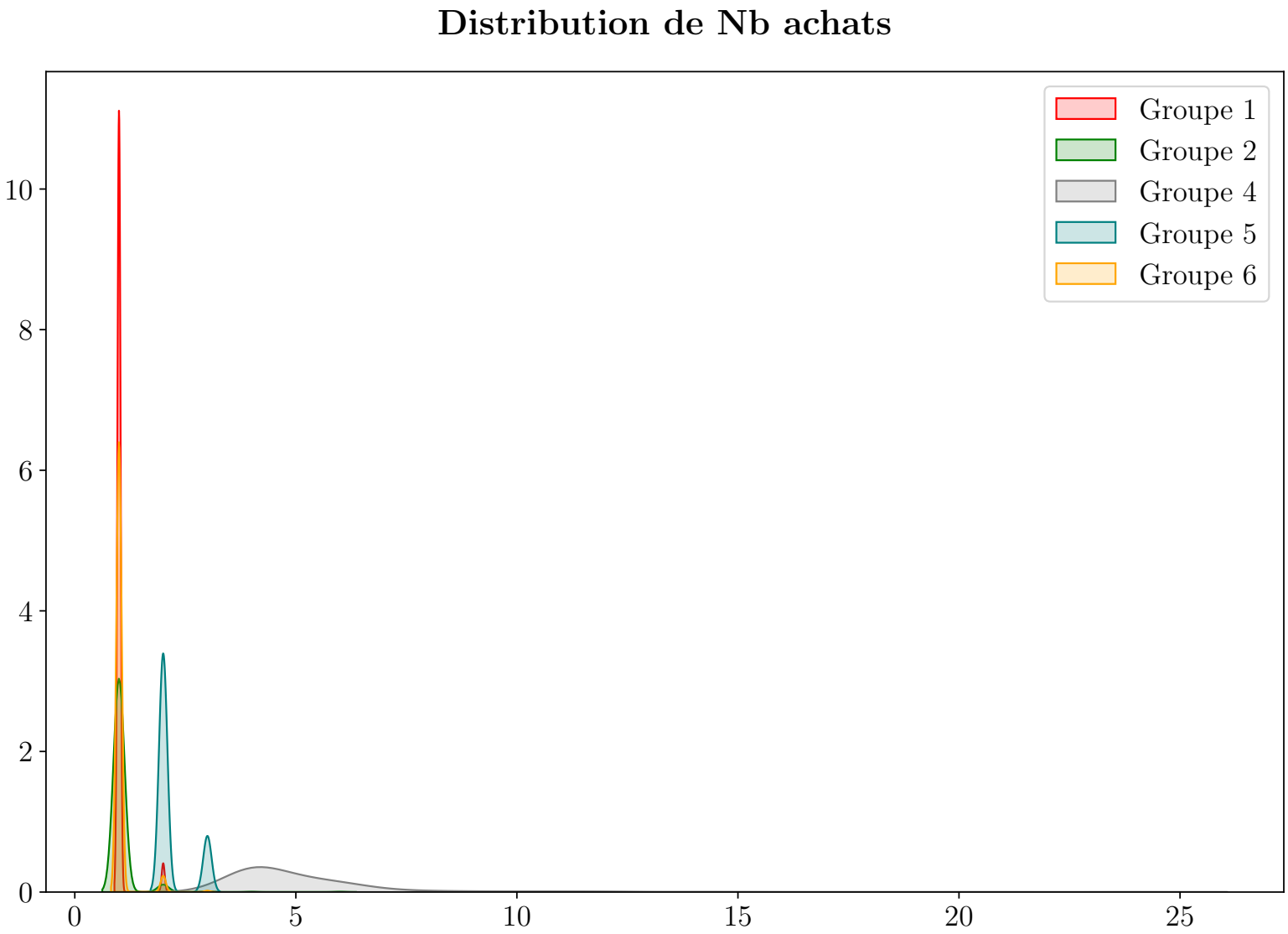


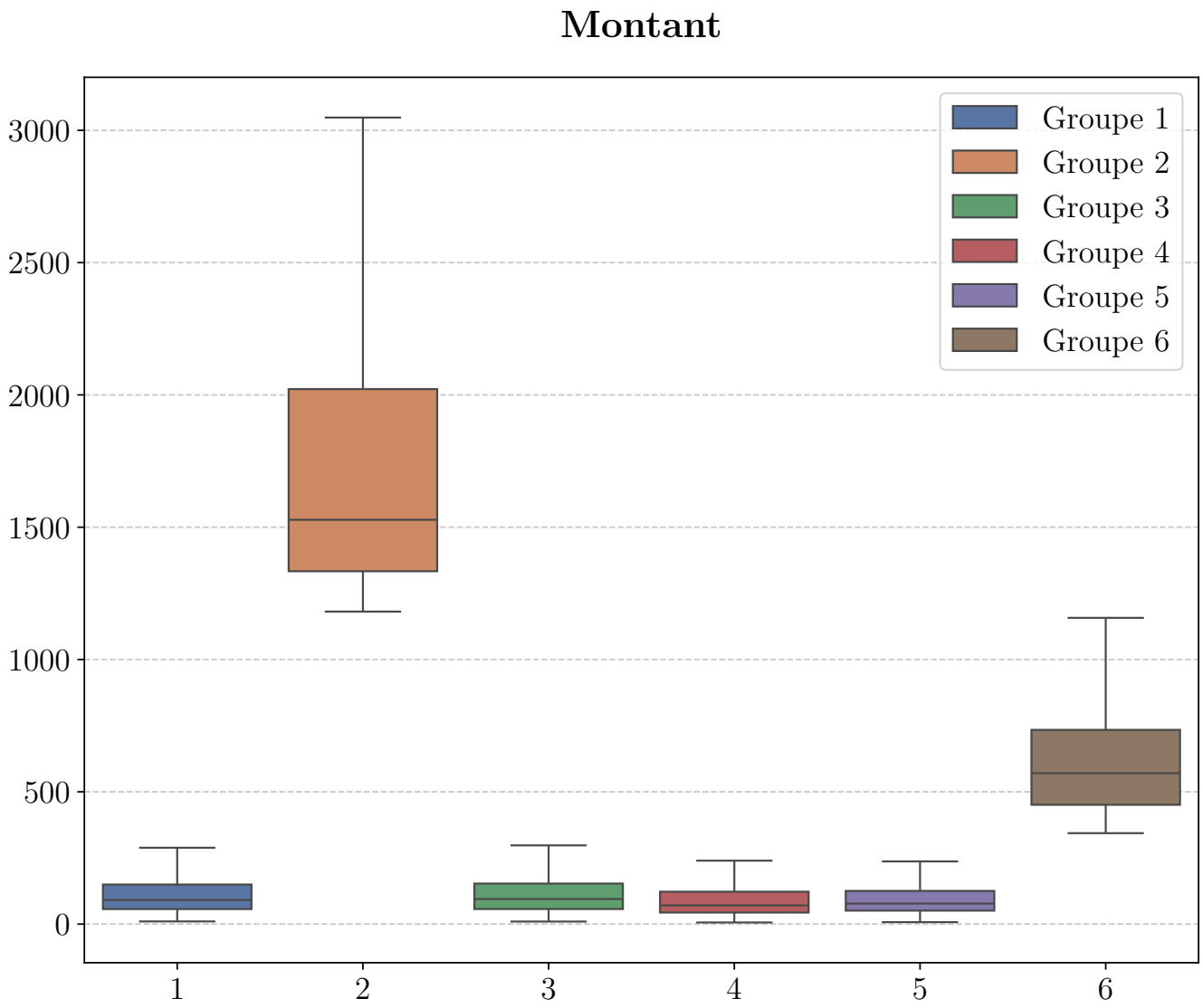
Analyse intra-cluster univariée

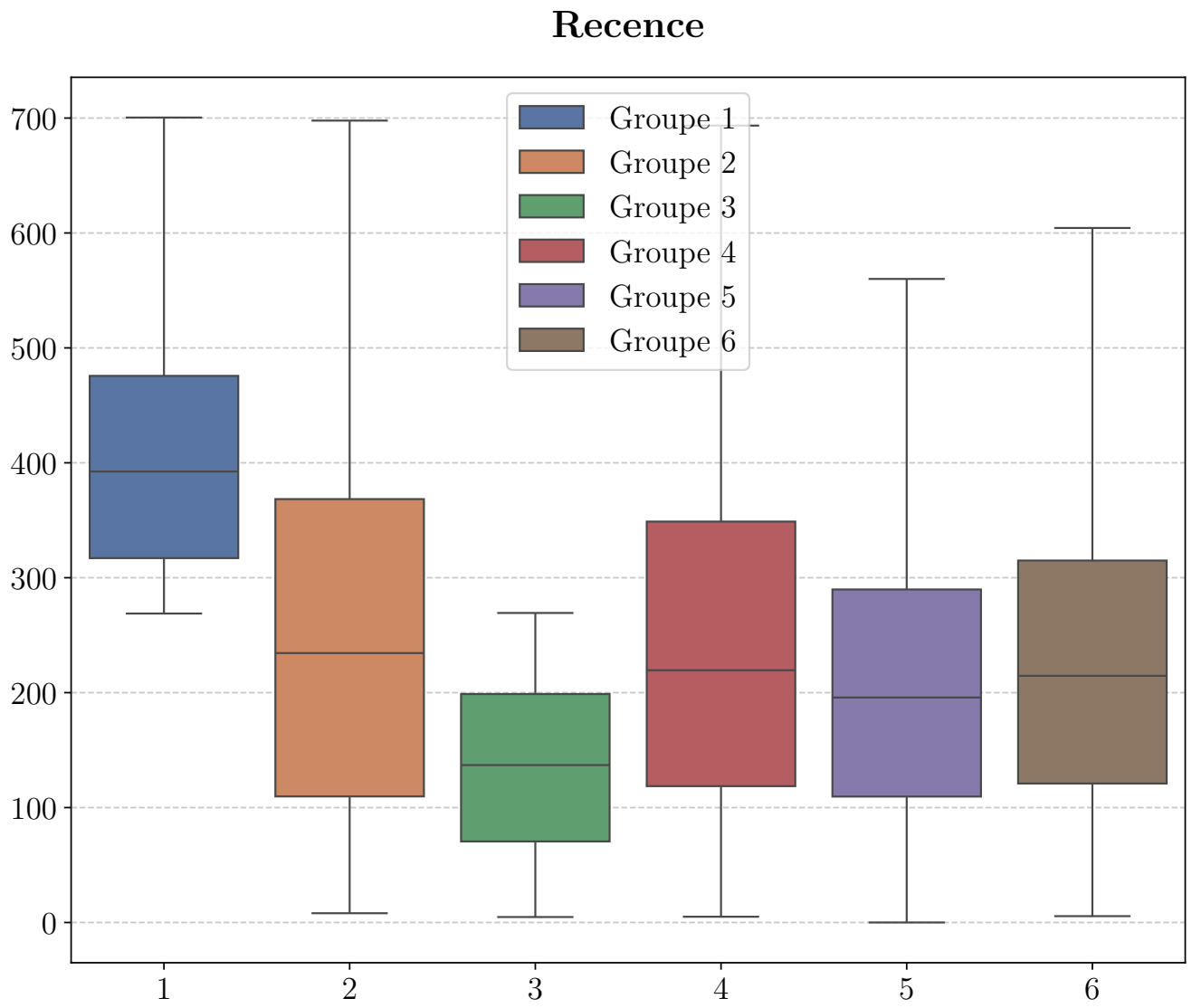


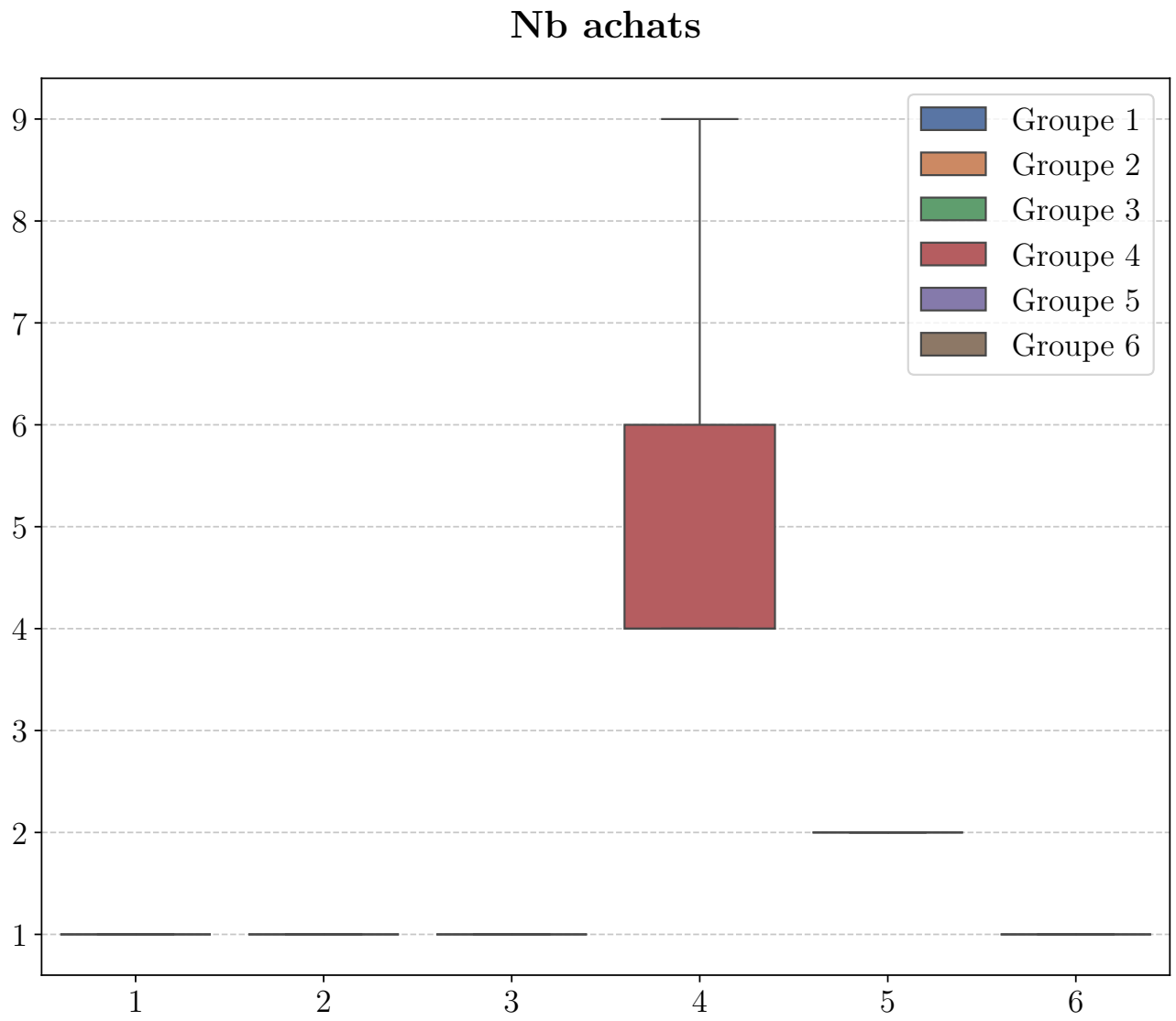
Distribution de Recence



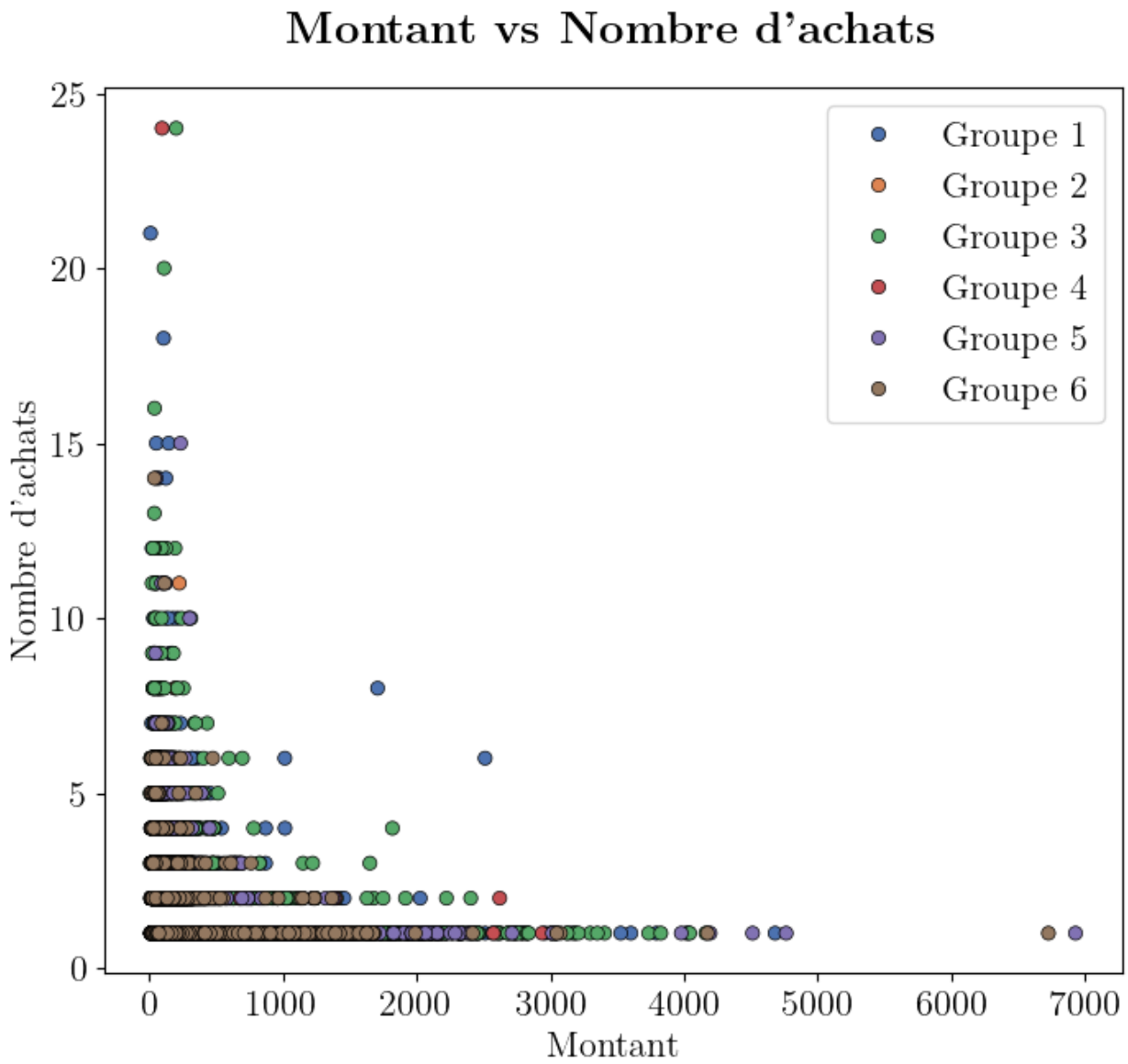

Distribution de Nb achats
Groupe 1
Groupe 2
Groupe 4
Groupe 5
Groupe 6
0
2
4
6
8
10
0
5
10
15
20
25

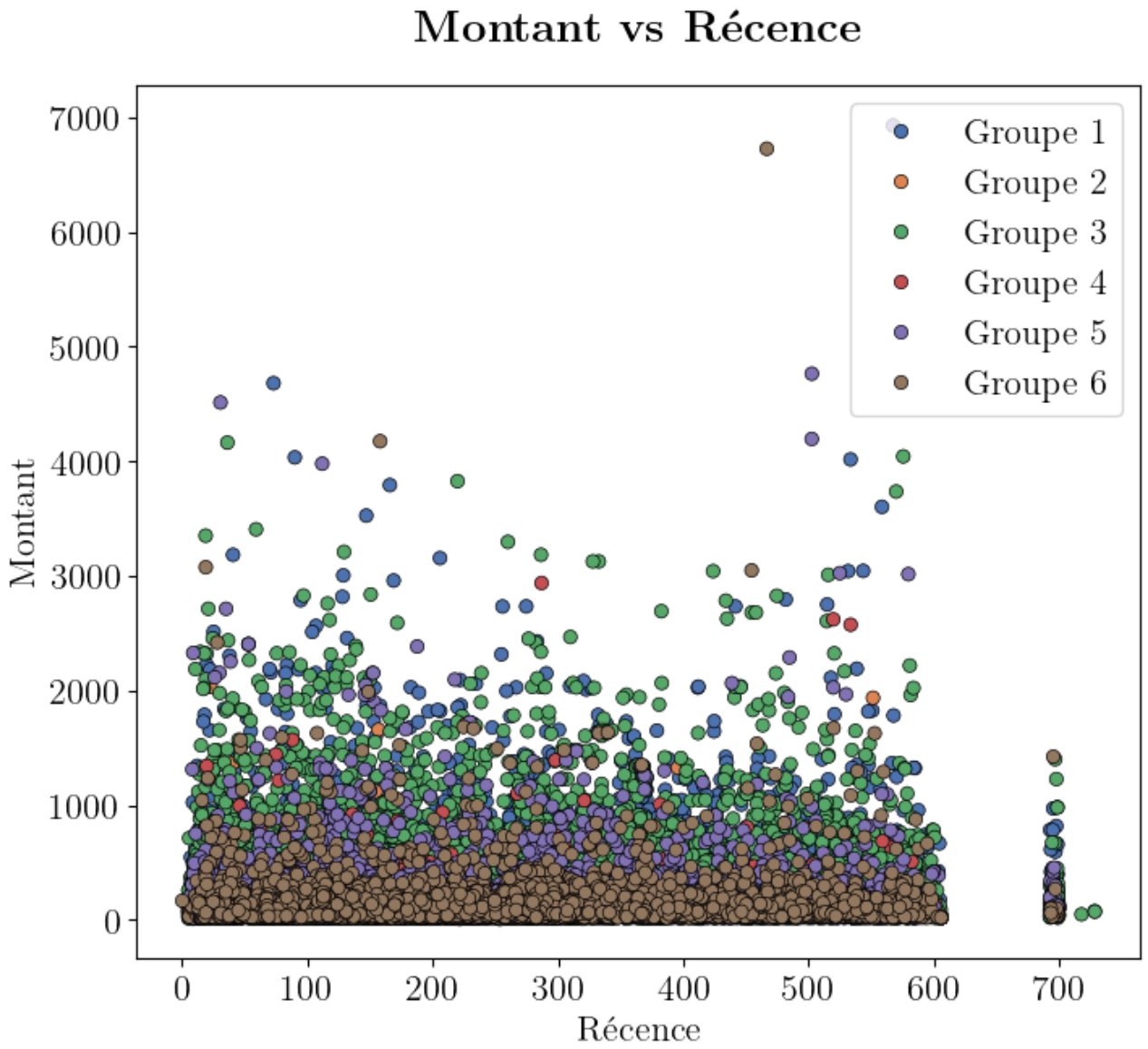

Montant
Groupe 1
Groupe 2
Groupe 3
Groupe 4
Groupe 5
Groupe 6
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6

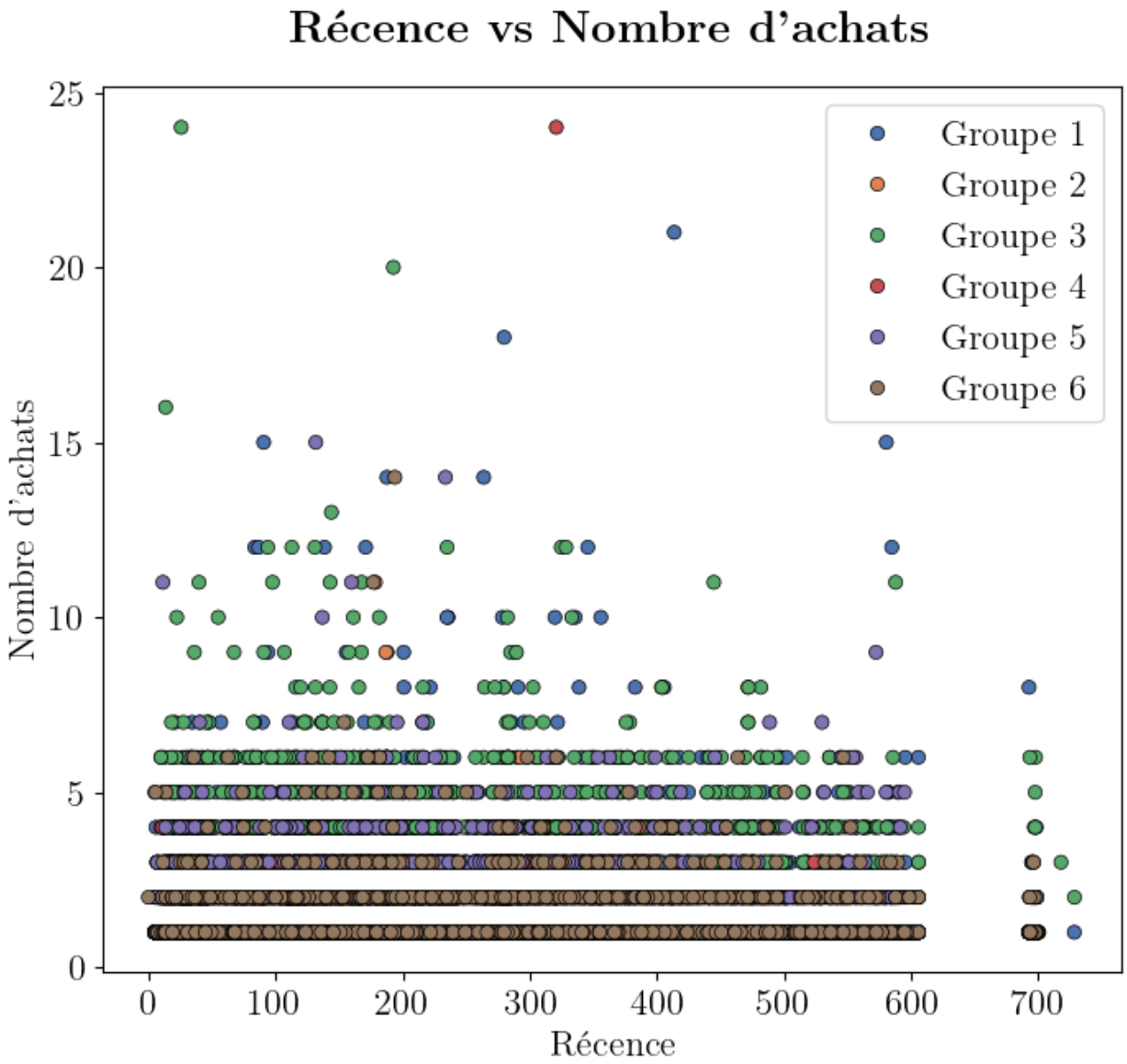

Recence
Groupe 1
Groupe 2
Groupe 3
Groupe 4
Groupe 5
Groupe 6
700
600
500
400
300
200
100
0
1
2
3
4
5
6


Nb achats
Groupe 1
Groupe 2
Groupe 3
Groupe 4
Groupe 5
Groupe 6
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6

Analyse intra-cluster bivariée




Montant vs Récence
Groupe 1
Groupe 2
Groupe 3
Groupe 4
Groupe 5
Groupe 6
Montant
Récence
0
1000
2000
3000
4000
5000
6000
7000
0
100
200
300
400
500
600
700

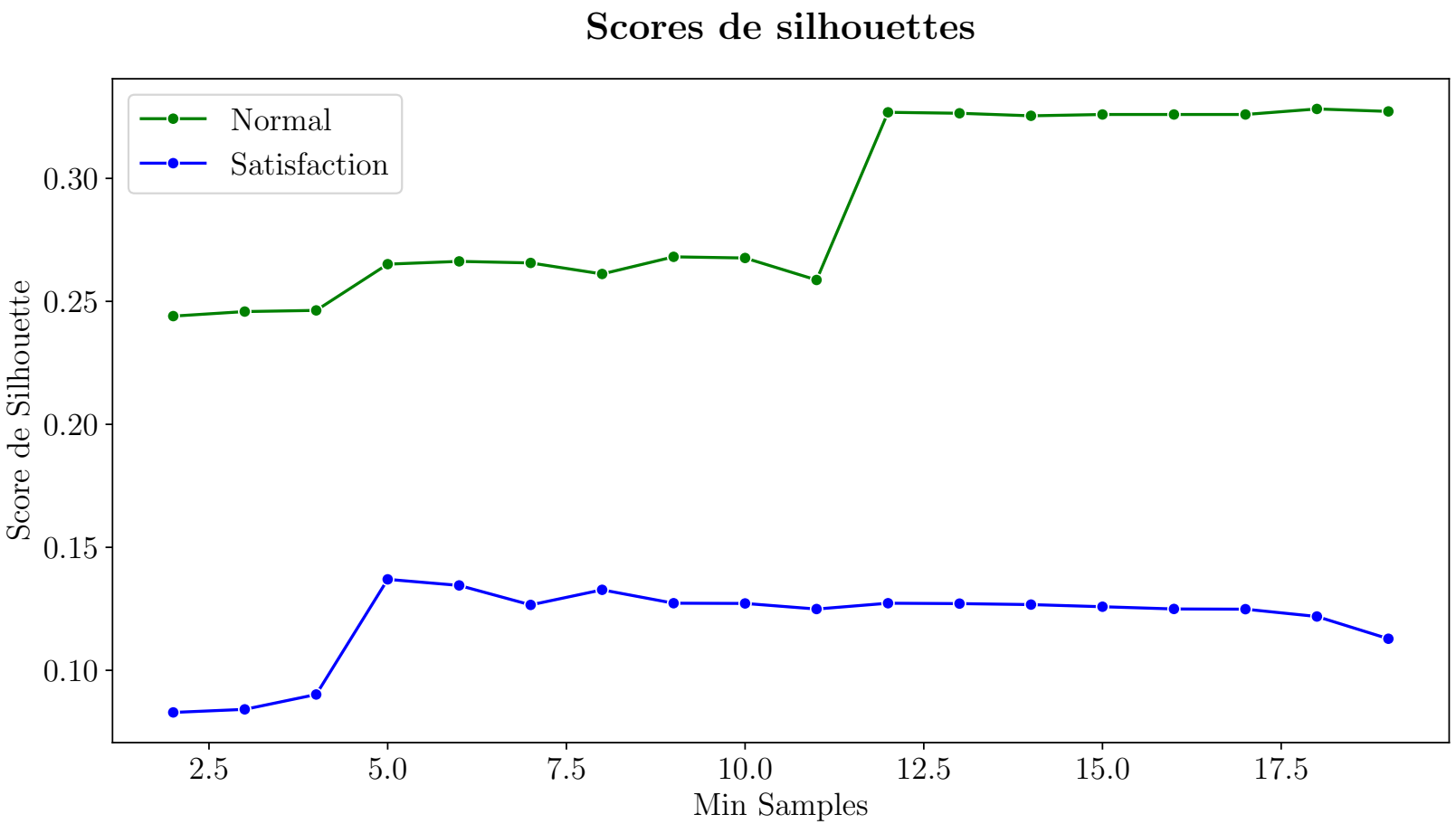

Récence vs Nombre d'achats
Nombre d'achats
Récence
0
5
10
15
20
25
0
100
200
300
400
500
600
700
Groupe 1
Groupe 2
Groupe 3
Groupe 4
Groupe 5
Groupe 6

CLUSTERING N°2
DBSCAN

Preprocessing

- Le temps de calcul avec DBSCAN étant très long pour notre jeu de données, on ne l'effectuera que sur un échantillon de taille 5000
- Pour avoir une meilleure sélection du rayon de sélection des noyaux, on standardise les données
- Pour la recherche d'hyperparamètres, on recherchera uniquement le `min samples` n, et ϵ sera fixé à $0,6$.
- La recherche d'hyperparamètres sera effectué sur les données sans et avec `satisfaction`.

Optimisation des paramètres

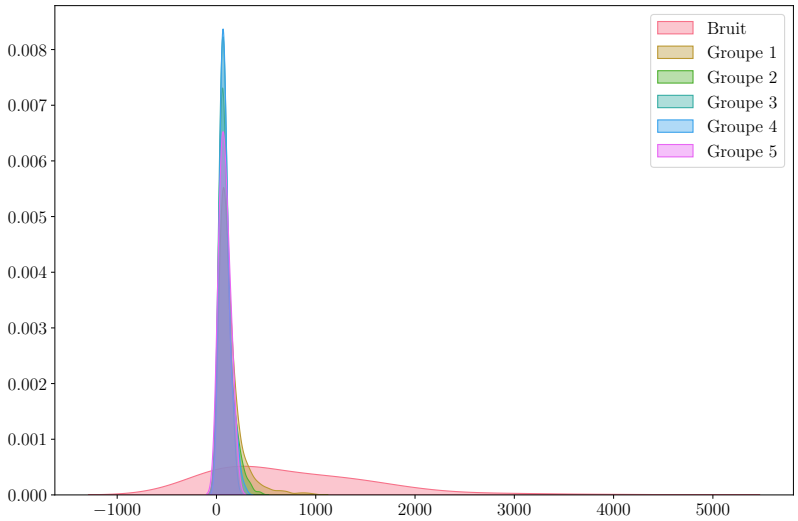


Résultats des évaluations

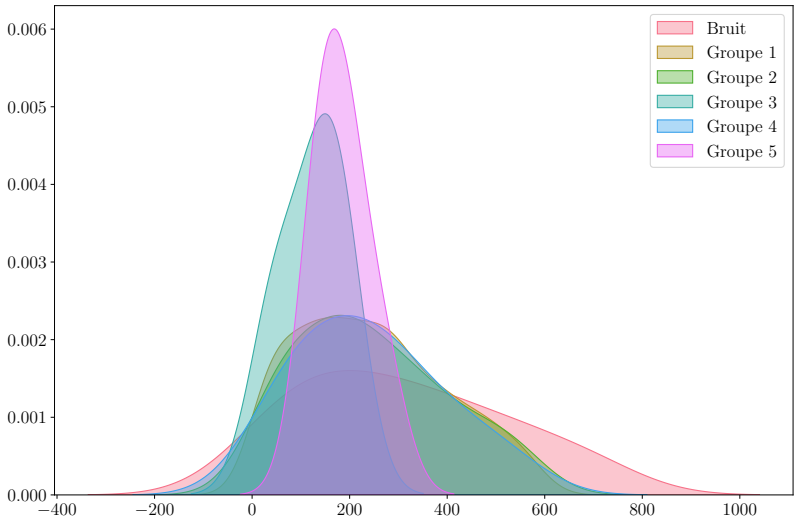
Min Samples	Clusters	Noise	Score
5	7	74	0,265
6	7	77	0,266
7	7	79	0,266
8	9	83	0,261
9	7	102	0,268
10	7	109	0,268
11	8	114	0,259
12	5	145	0,327
13	4	162	0,326
14	4	165	0,325
15	3	185	0,326

Analyse intra-cluster univariée

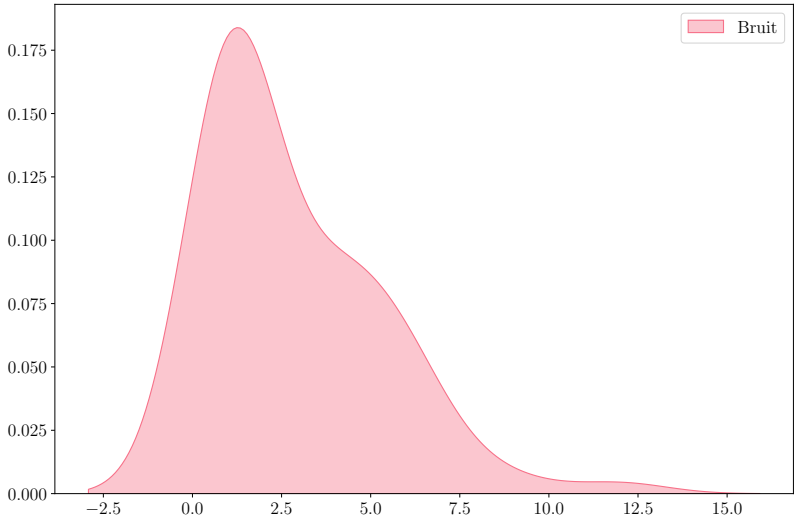
Distribution de Montant



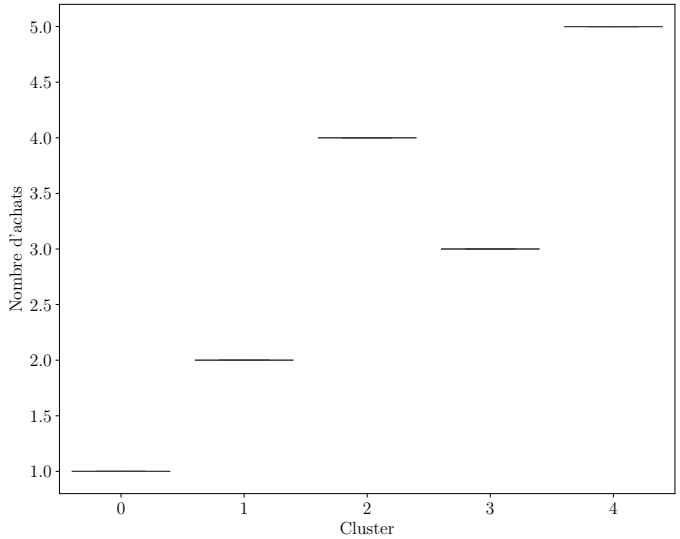
Distribution de Recence



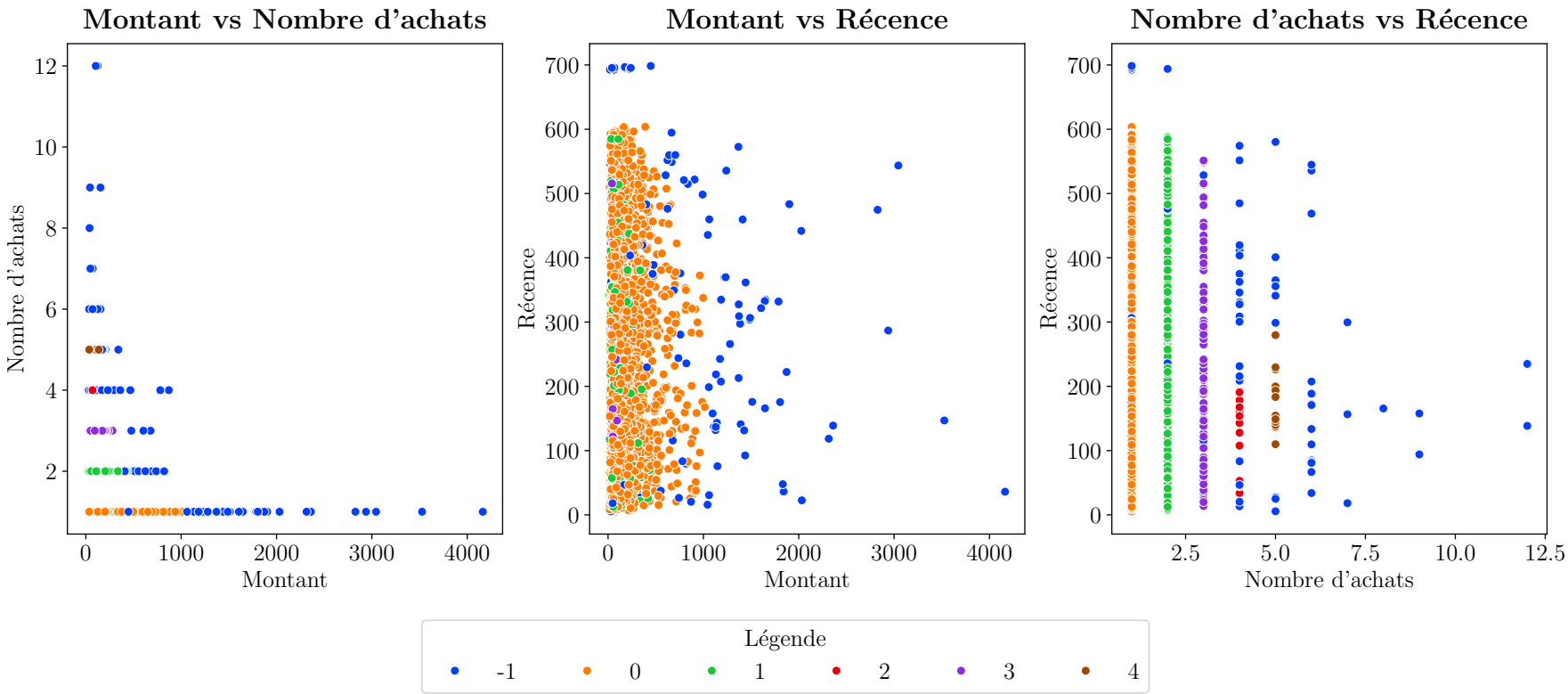
Distribution de Nb achats



Nombre d'achats par cluster



Analyse intra-cluster bivariée



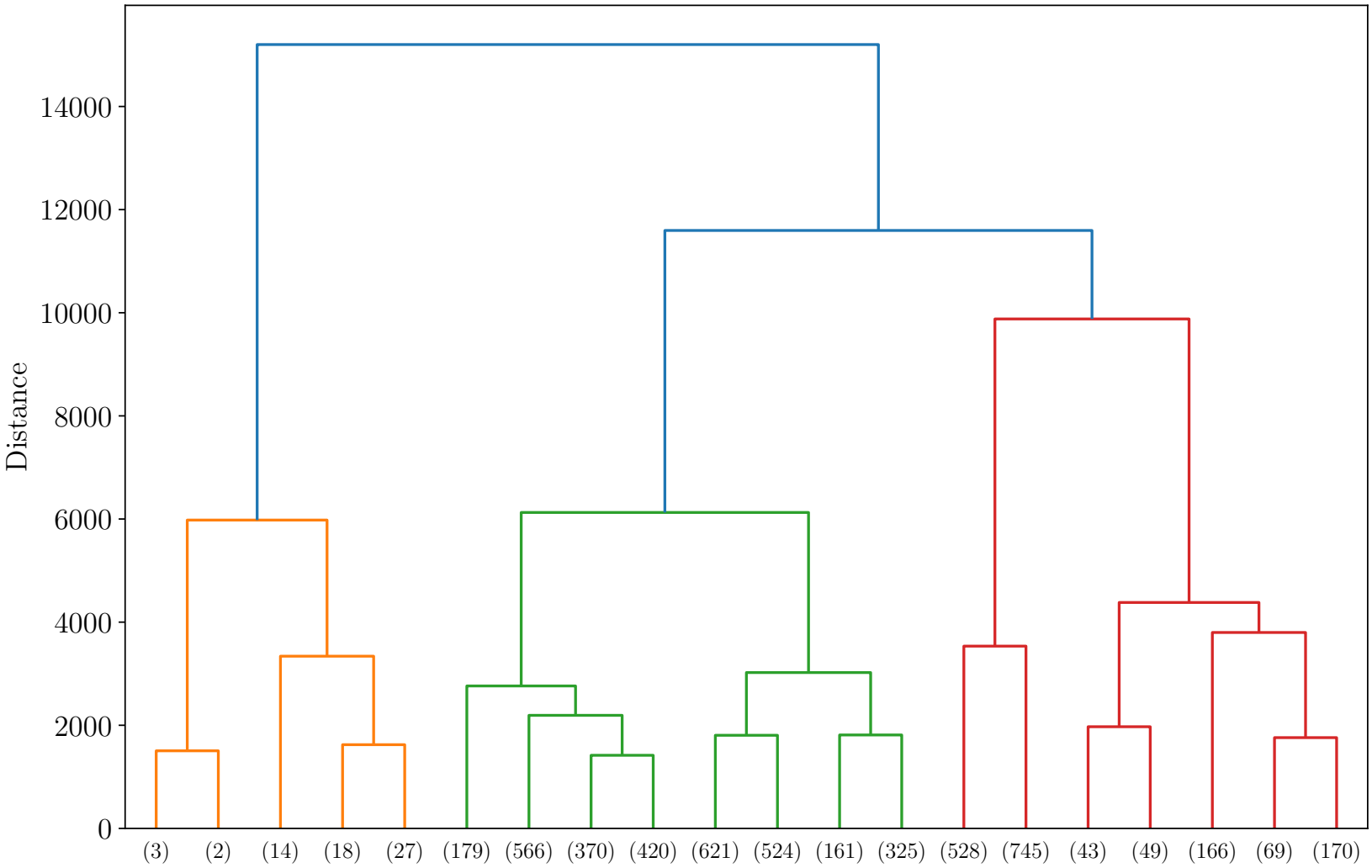
Le groupe −1 représente le bruit.

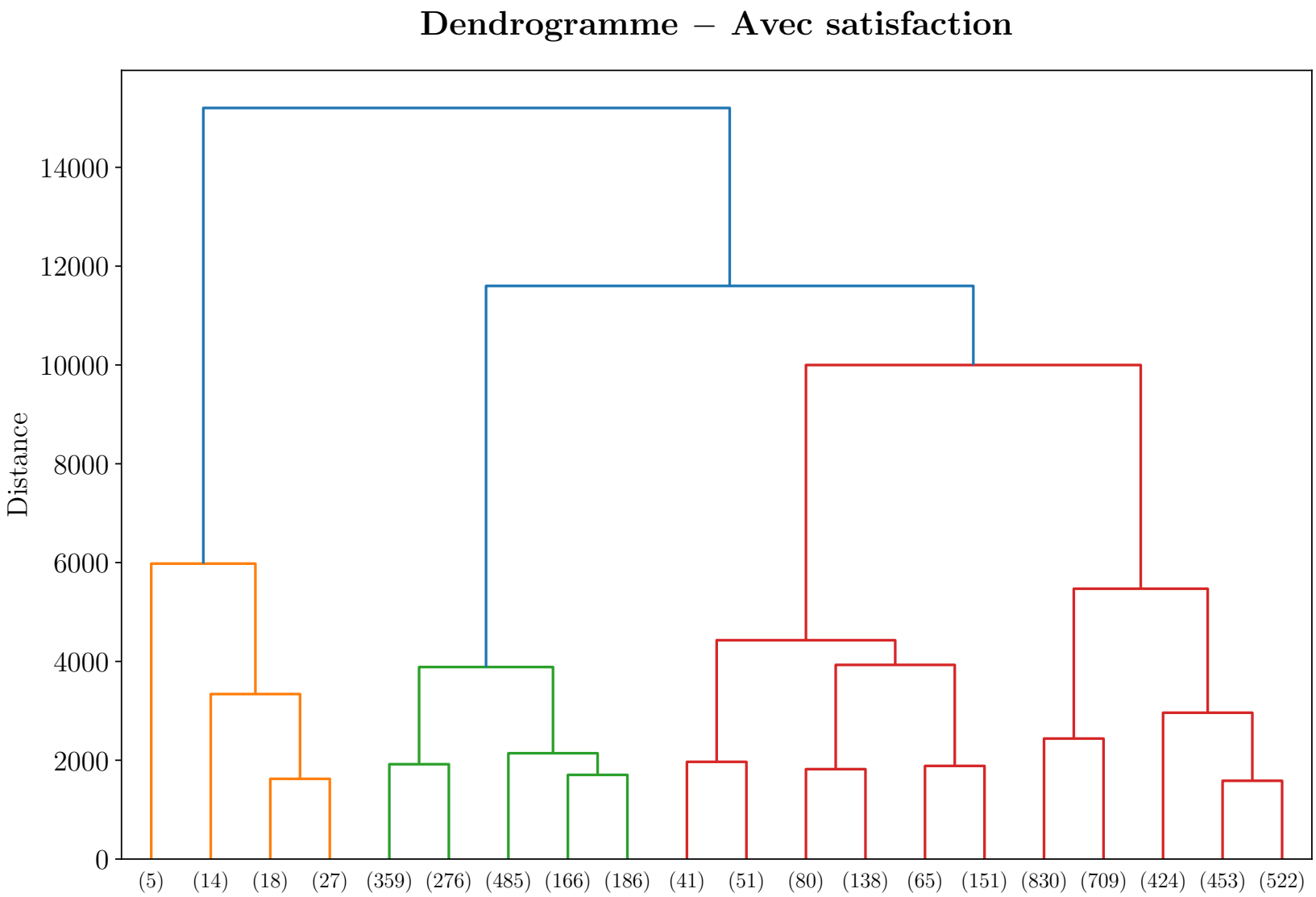
CLUSTERING N° 3
CLUSTERING HIÉRARCHIQUE

Optimisation des paramètres

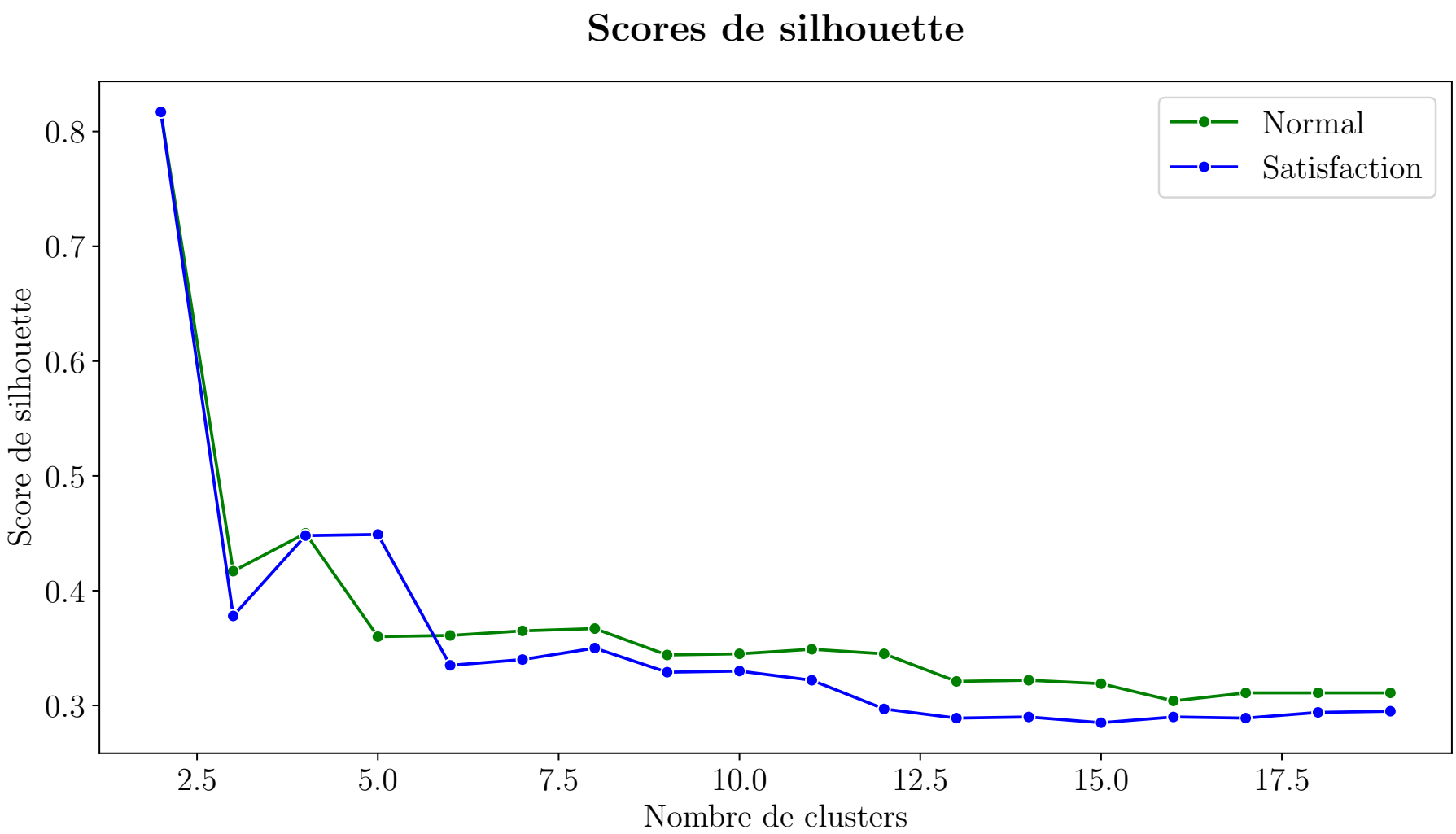
- De même que pour le DBSCAN, les recherches s'effectueront sur un échantillon aléatoires de 5000 individus.
- Affichage des dendrogrammes
- Évaluation du score de silhouette en fonction du nombre de clusters choisi
- Les évaluations se feront sans et avec la variable `satisfaction`

Dendrogramme – Sans satisfaction



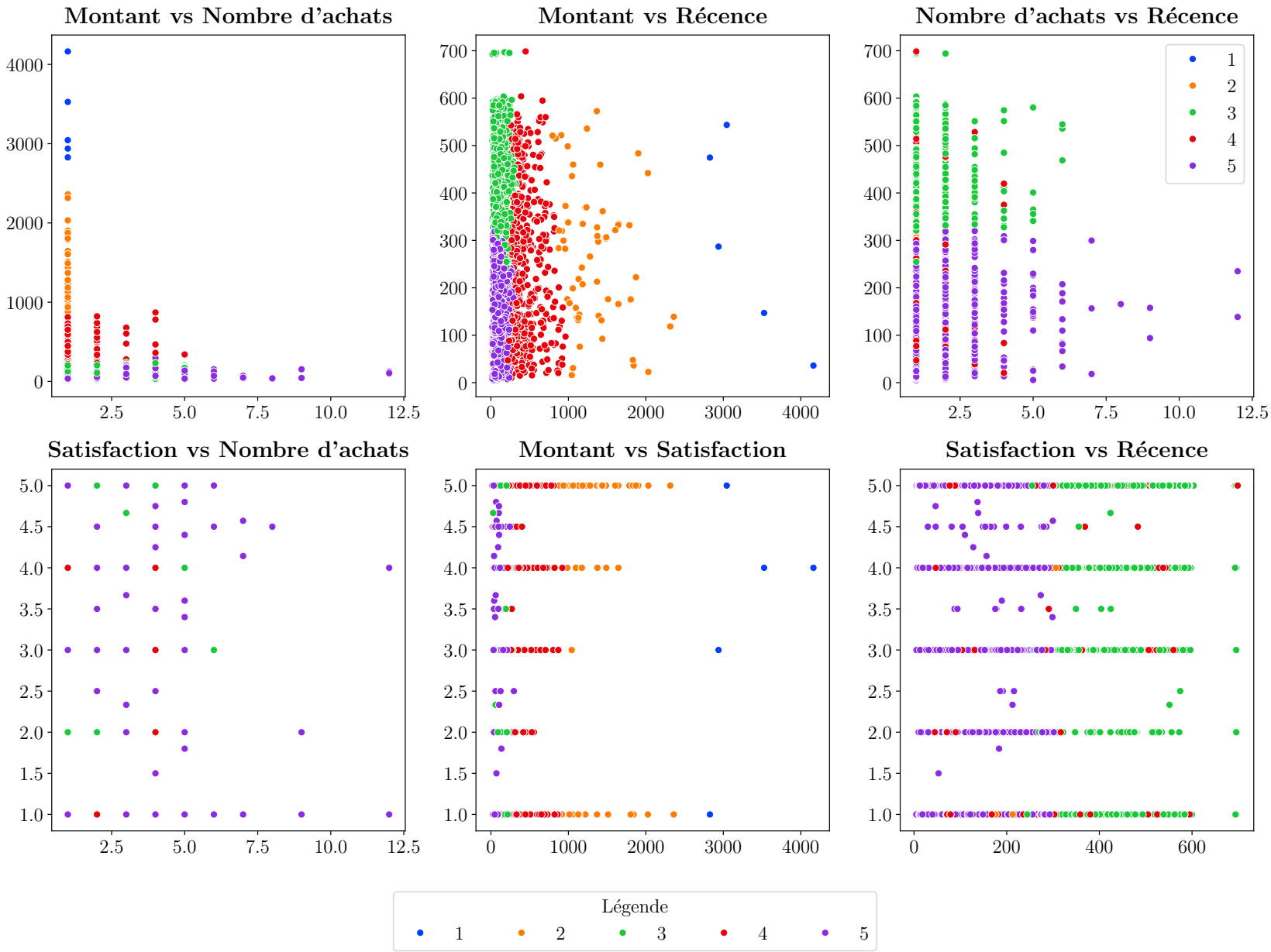

Dendrogramme – Avec satisfaction
Distance
0
2000
4000
6000
8000
10000
12000
14000
(5)
(14)
(18)
(27)
(359)
(276)
(485)
(166)
(186)
(41)
(51)
(80)
(138)
(65)
(151)
(830)
(709)
(424)
(453)
(522)

Évaluation du clustering



Résultats des évaluations

Nombre de clusters	Score — Normal	Score — Satis
2	0,817	0,817
3	0,417	0,378
4	0,450	0,448
5	0,360	0,449
6	0,361	0,335
7	0,365	0,340
8	0,367	0,350
9	0,344	0,329
10	0,345	0,330
11	0,349	0,322
12	0,345	0,297


Montant vs Nombre d'achats
Montant vs Récence
Nombre d'achats vs Récence
Satisfaction vs Nombre d'achats
Montant vs Satisfaction
Satisfaction vs Récence
1
2
3
4
5
Légende
1 2 3 4 5

MAINTENANCE
ET
ÉVOLUTION DU MODÈLE

Modèle optimal

Après les évaluations faites des différents types de segmentation, le modèle de clustering optimal est le KMeans.

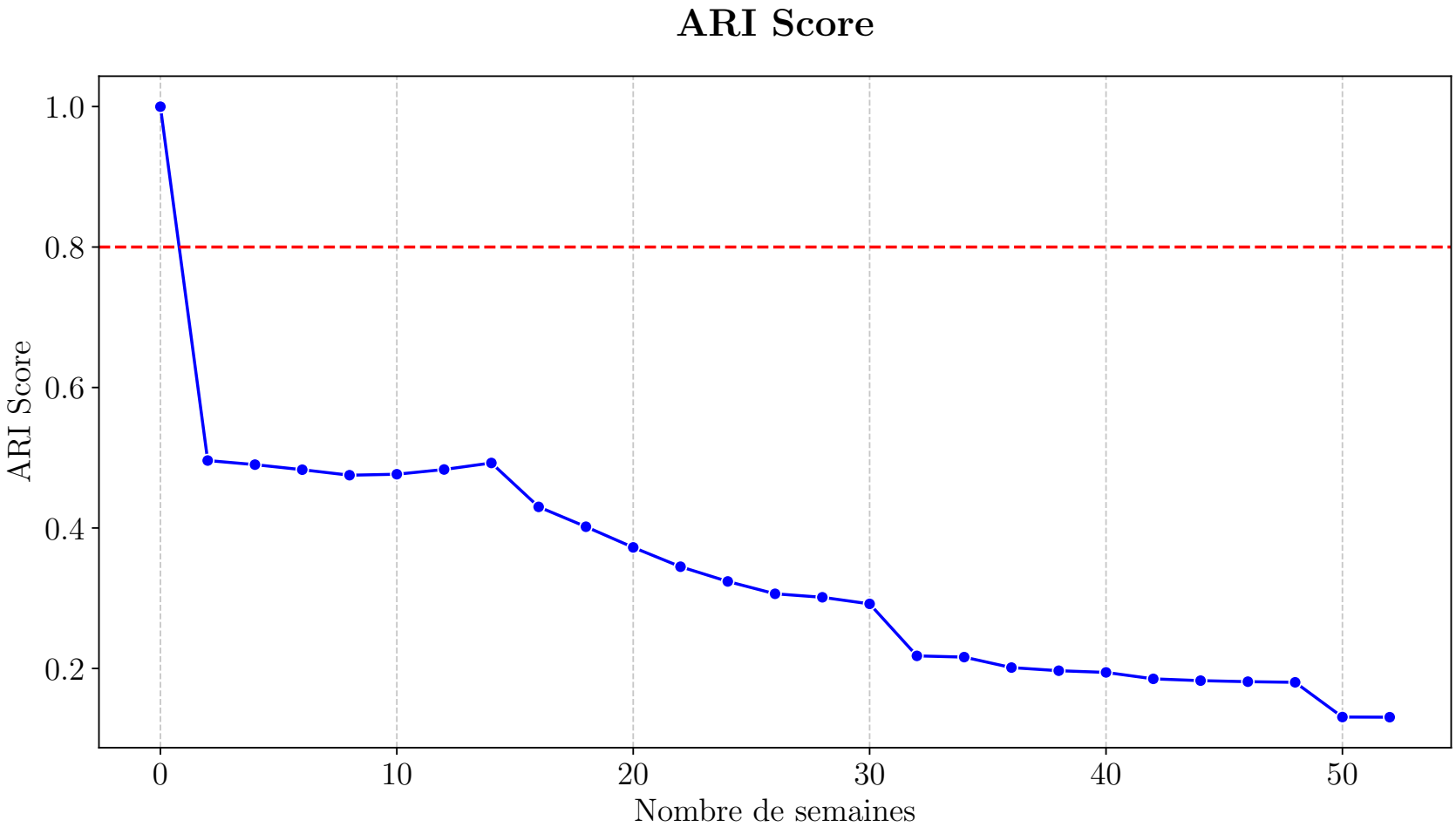
➢ Nombre de clusters	6
➢ Standardisation	OUI

On utilisera donc ce modèle pour en évaluer la maintenance

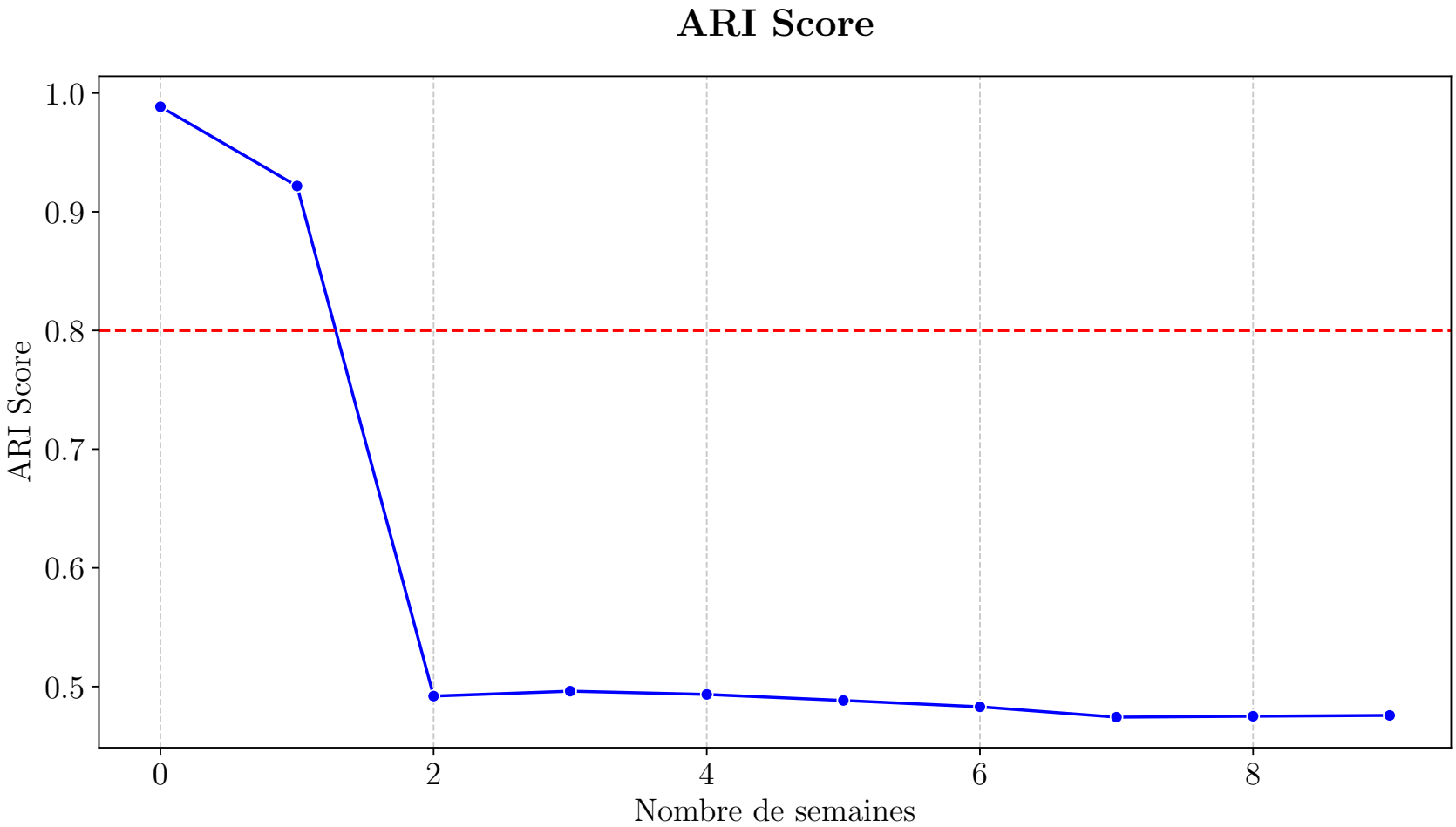
Procédure

- Reprise des données non groupées par client.
- Création d'une fonction qui prend en entrée :
  - Temps initial (en jours)
  - Pas ou incrémentation (en semaine)
  - Nombre de clusters
- On ne prend que les commandes passées avant le temps initial, on applique le clustering.
- On incrémente le temps initial du pas, puis on réapplique le clustering, jusqu'à arriver au temps maximal.
- À chaque incrémentation, on compare le clustering avec le clustering initial à l'aide de l'ARI Score

Temps initial : 6 mois et Pas : 2 semaines



Temps initial : 6 mois et Pas : 1 semaine



Résultat de l'évaluation

Après 2 semaines, l'ARI score descend en dessous de 0, 8.

On conclut de ces évaluations, qu'au bout de 2 semaines, le modèle de segmentation n'est plus à jour et doit être refait avec les données nouvelles.

CONCLUSION

Caractéristiques des groupes de clients

- Groupe 1 → Dépenses — Faibles (< 100); Nombre d'achats — Faible (≈ 1); Récence — Élevée (≈ 1 an)
- Groupe 2 → Dépenses — Élevées (> 1500); Nombre d'achats — Faible (≈ 1); Récence — Moyenne (≈ 9 mois)
- Groupe 3 → Dépenses — Faibles (< 100); Nombre d'achats — Faible (≈ 1); Récence — Faible (≈ 5 mois)

- Groupe 4 → Dépenses — Faibles (< 100); Nombre d'achats — Élevé (≈ 4); Récence — Moyenne (≈ 7 mois)
- Groupe 5 → Dépenses — Faibles (< 100); Nombre d'achats — Moyen (≈ 2); Récence — Plutôt basse (≈ 6 mois)
- Groupe 6 → Dépenses — Plutôt élevées (≈ 600); Nombre d'achats — Faible (≈ 1); Récence — Moyenne (≈ 7 mois)

— FIN —