Projet 5: Segmentez des clients d'un site de e-commerce

QUENTIN STEPNIEWSKI

OPENCLASSROOMS

24 AOÛT 2020

Sommaire

1. Introduction – Présentation de la problématique

Consultant pour Olist, une solution de vente en ligne implantée en Amérique du Sud

Problématique principale

- Aider les équipes d'Olist à comprendre les différents types d'utilisateurs

Objectifs de l'étude

- Proposer une solution de segmentation clients clé en main
- Evaluer la fréquence de mise à jour de cette segmentation en vue d'établir un contrat de maintenance

olist

suas vendas estão no olist

1. Introduction – Présentation de la problématique

Interprétation de la problématique :

Exploration et choix de features adaptées :

Mise en place d'un pipeline d'extraction des données

↓

Mise en place d'une segmentation clients :

Tests et sélection d'algorithme de classification non-supervisée

↓

Compréhension des différents types de clients :

Analyse des caractéristiques des différents clusters identifiés

↓

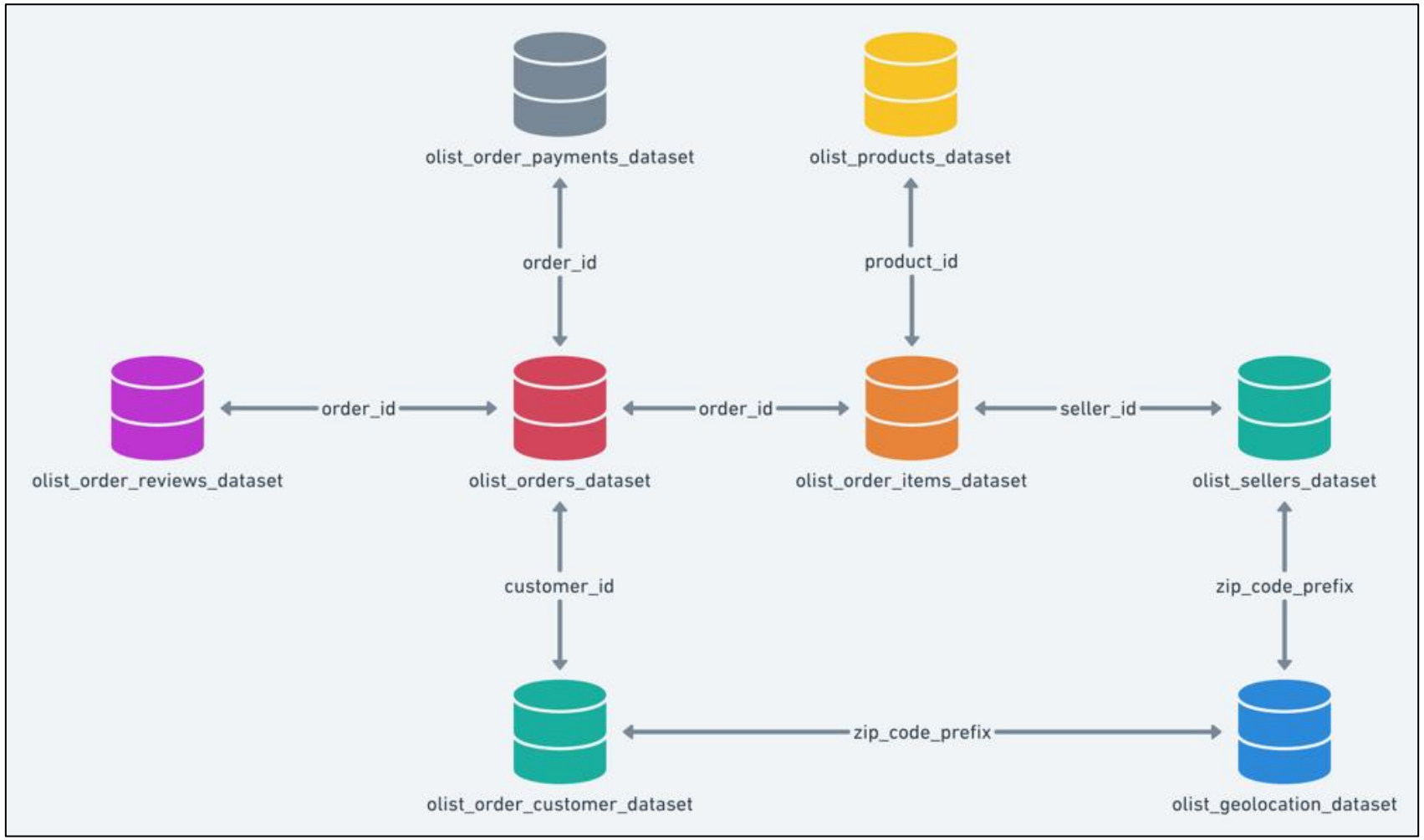
Etude d'une fréquence de mise à jour :

Etude du comportement des clusters dans le temps

2. Présentation des données utilisées

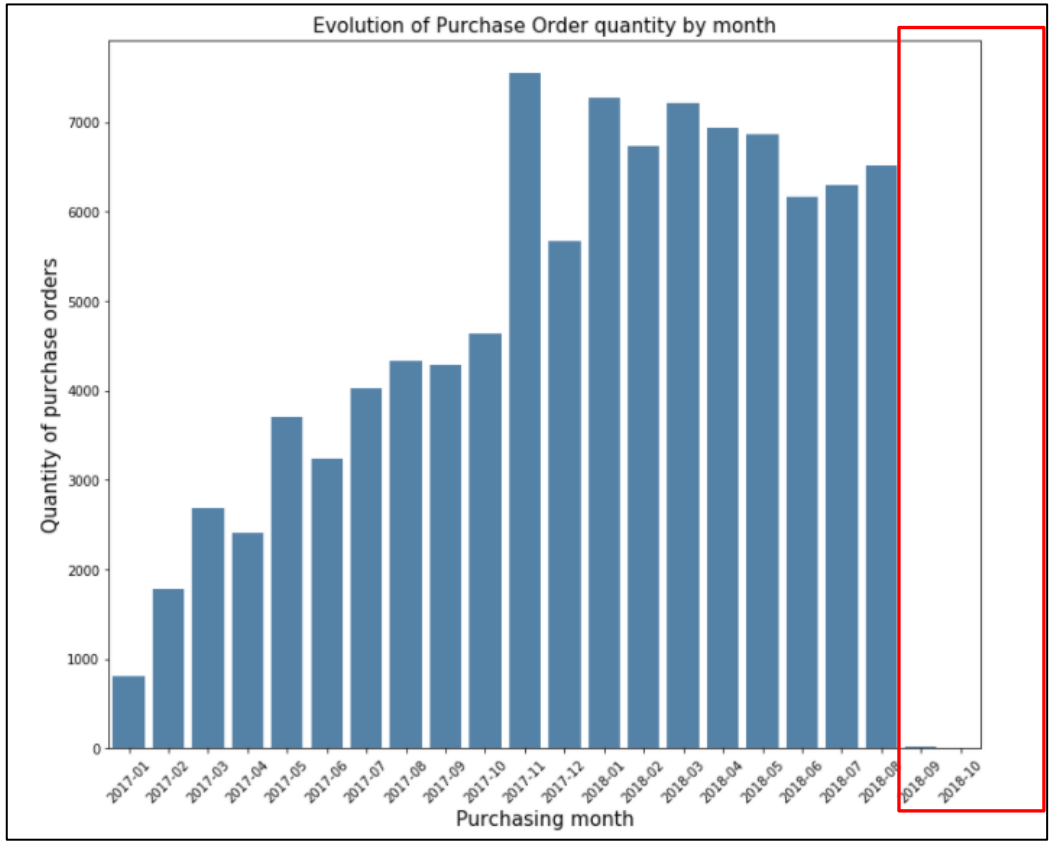
Base de données composée de 8 tables (+ une table de traduction des catégories de produits) :

- Clients
- Géolocalisation
- Commandes
- Produits
- Vendeurs
- Paiements
- Satisfaction



2. Présentation des données utilisées

Exploration :

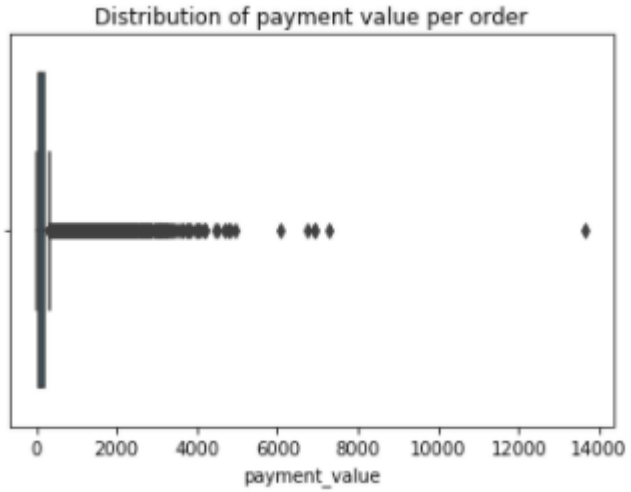


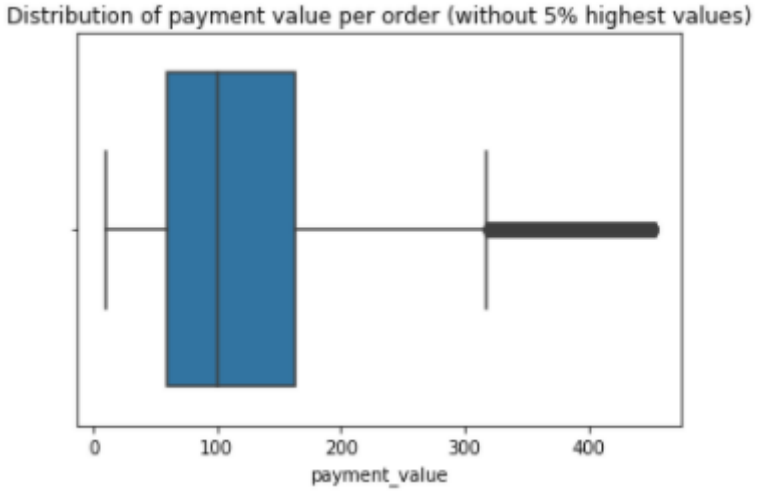
Moins de deux années représentées (moins de 2 cycles)

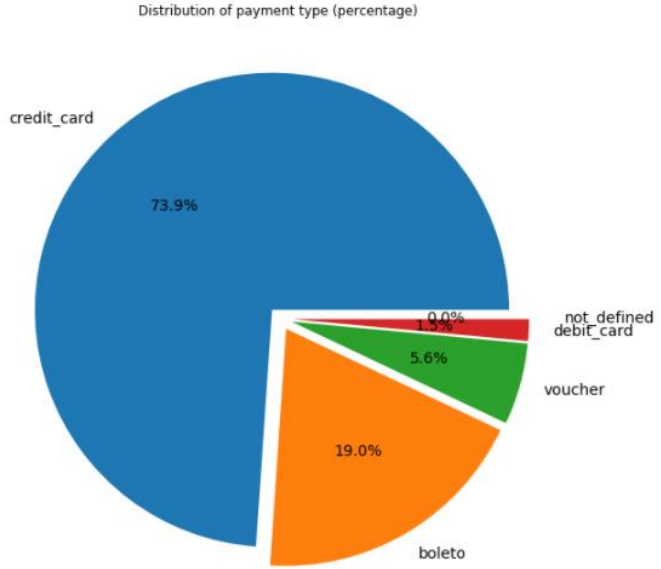
Plus de 95% des clients n'ont passé qu'une commande

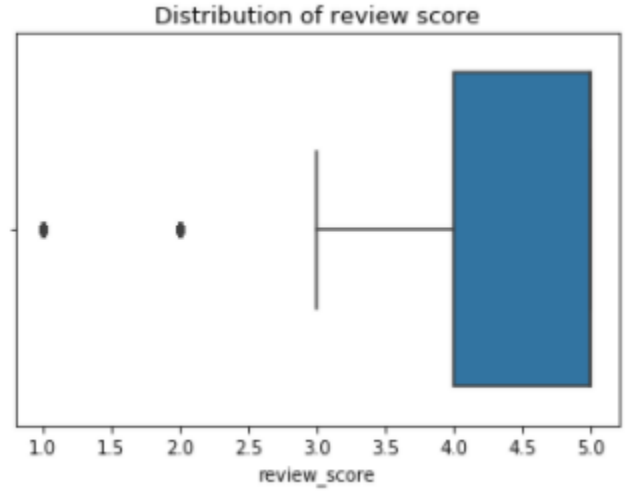
2. Présentation des données utilisées

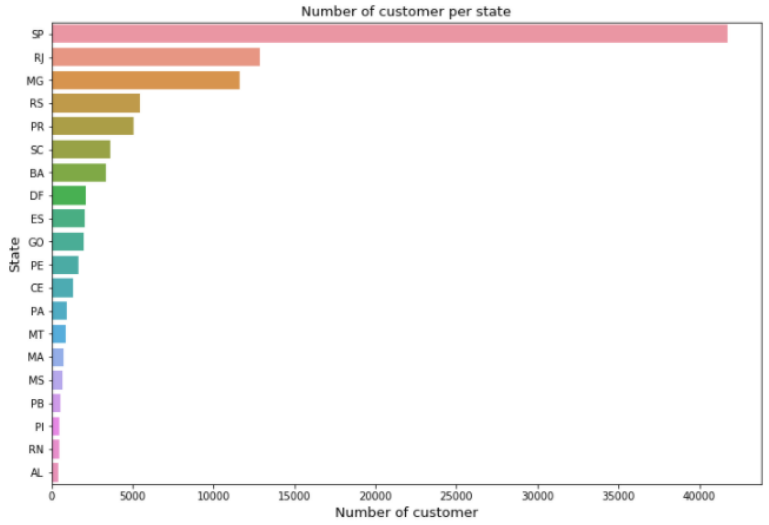
Exploration :

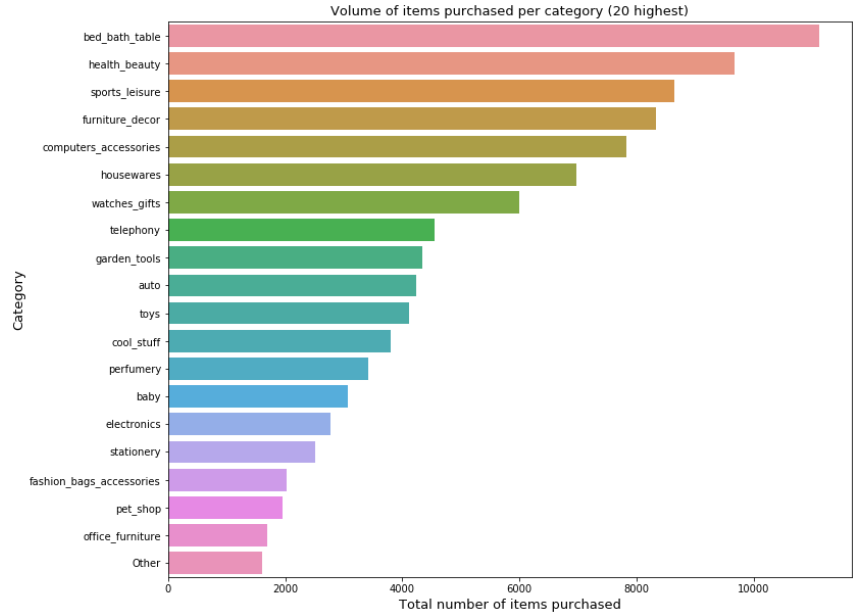












3. Nettoyage de la base de données

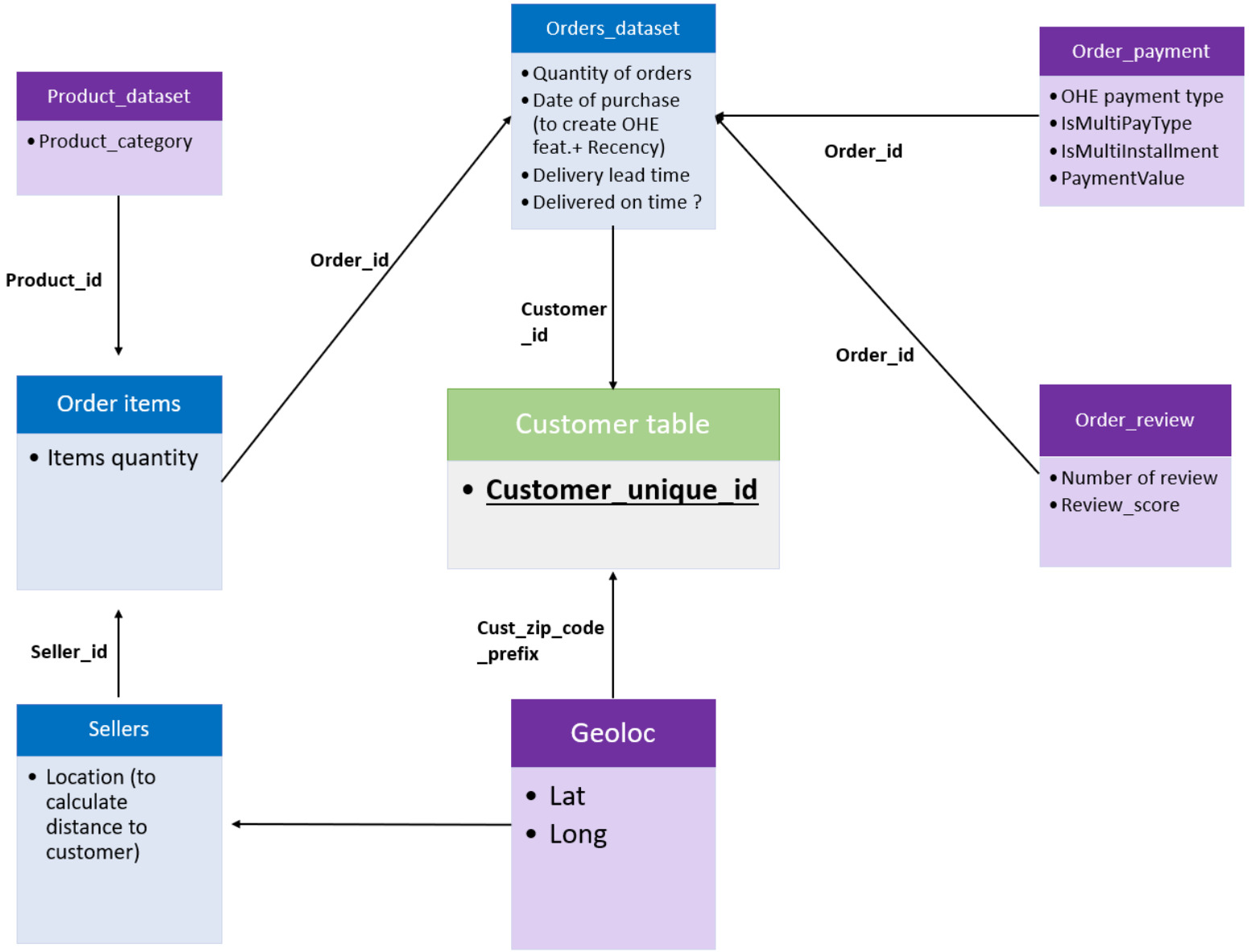
Sélection des features importantes pour notre étude :

Mise en place d'une 'map' représentant les features retenues pour chaque table

3.]

Sélec

Mise


Orders_dataset
• Quantity of orders
• Date of purchase (to create OHE feat.+ Recency)
• Delivery lead time
• Delivered on time ?
Order_payment
• OHE payment type
• IsMultiPayType
• IsMultiInstallment
• PaymentValue
Order_id
Product_dataset
• Product_category
Product_id
Order_id
Customer _id
Order_id
Order items
• Items quantity
Customer table
• Customer_unique_id
Order_review
• Number of review
• Review_score
Seller_id
Cust_zip_code _prefix
Sellers
• Location (to calculate distance to customer)
Geoloc
• Lat
• Long

3. Nettoyage de la base de données

Création d'une fonction de nettoyage/fusion des données
Observation et suppression des valeurs aberrantes
Gestion des types
Réduction du nombre de catégories produit (74 -> 13)
Fusion des 8 tables en une table ayant pour index le client_unique_id
OneHotEncoding des variables catégorielles
Création de nouvelles features

Création d'une fonction réalisant ces différents tâches sur une période temporelle à entrer en paramètre

3. Nettoyage de la base de données

Satisfaction

- Note moyenne
- Nombre de reviews
- Nombre de livraisons en retard

Informations générales de commandes

- Nombre de commandes
- Nombre de produits commandés
- Catégorie de produit
- Montant total dépensé

Localisation

- Latitude + Longitude
- Distance Client - Vendeur
- Temps moyen de livraison

Informations client

Paiement

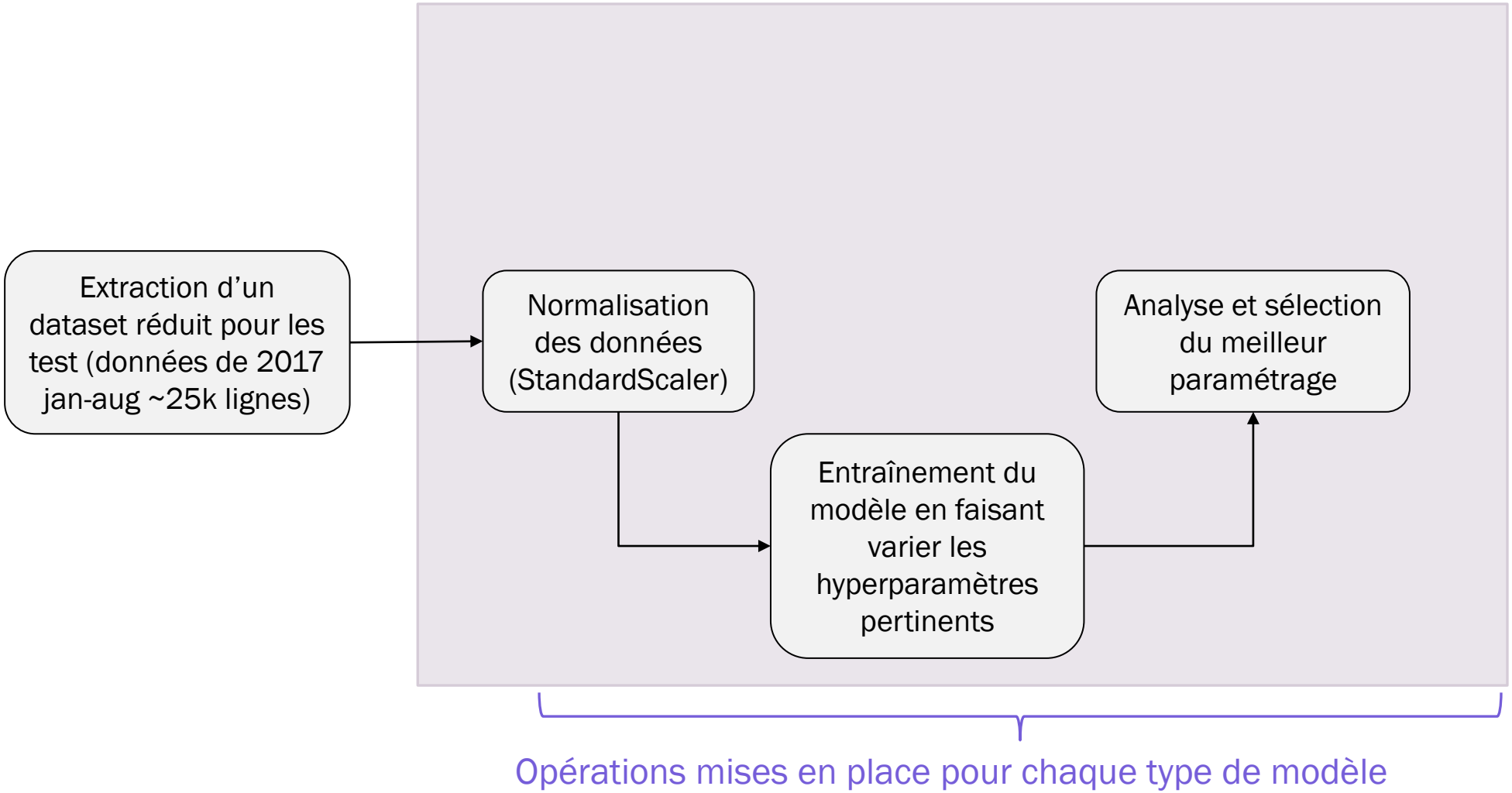
- Fréquence des moyens de paiement
- Demande d'étalement de paiement

Données temporelles

- Récence du dernier achat
- Mois d'achat
- Client de Noël

4. Mise en place et sélection des algorithmes de clustering

Principe de sélection mis en place pour les algorithmes de clustering :



4. Mise en place et [illegible] rithmes de clustering

Critères de sélection

« Forme » des clusters

Stabilité

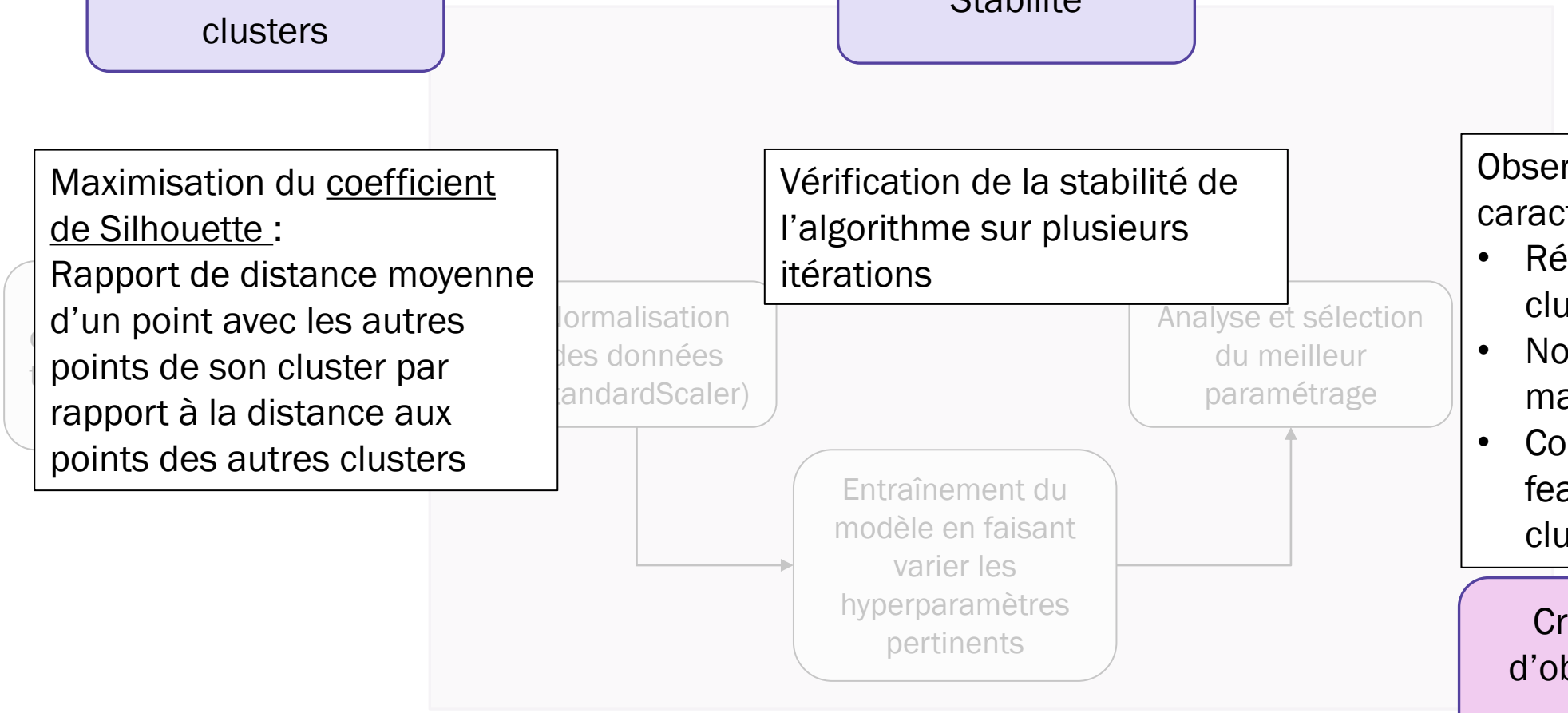
Pertinence des clusters

Maximisation du coefficient de Silhouette :
Rapport de distance moyenne d'un point avec les autres points de son cluster par rapport à la distance aux points des autres clusters

Vérification de la stabilité de l'algorithme sur plusieurs itérations

Observation des caractéristiques des clusters:
- Répartition des clients par cluster
- Nombre de clusters (10 max)
- Comportement des features pour chaque cluster

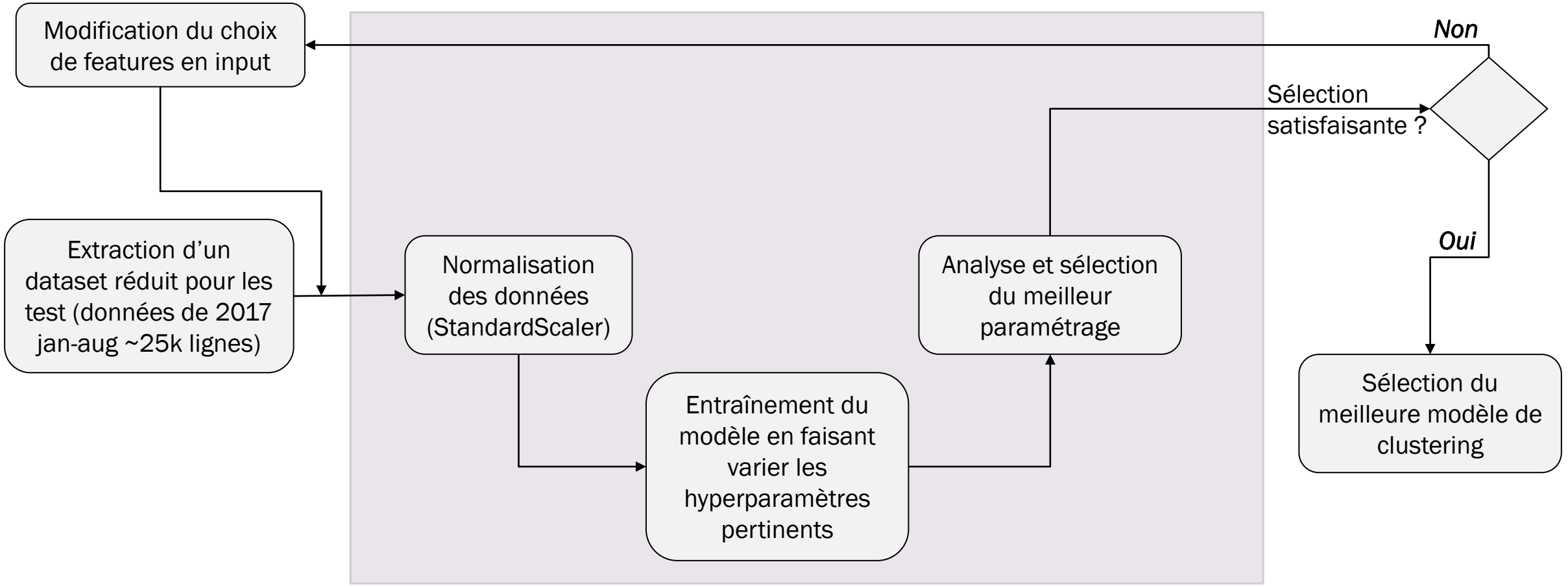
Création d'une fonction d'observation basique des clusters



Opérations mises en place pour chaque type de modèle

4. Mise en place et sélection des algorithmes de clustering

Principe de sélection mis en place pour les algorithmes de clustering :



4. Mise en place et sélection des algorithmes de clustering

Liste des différents algorithmes testés :

- Kmeans
- Gaussian Mixture
- AgglomerativeClustering (clustering hiérarchique)
- DBSCAN

Features écartées après plusieurs itérations :

- Features concernant le mois d'achat

(moins de deux années représentées sur le jeu de données)

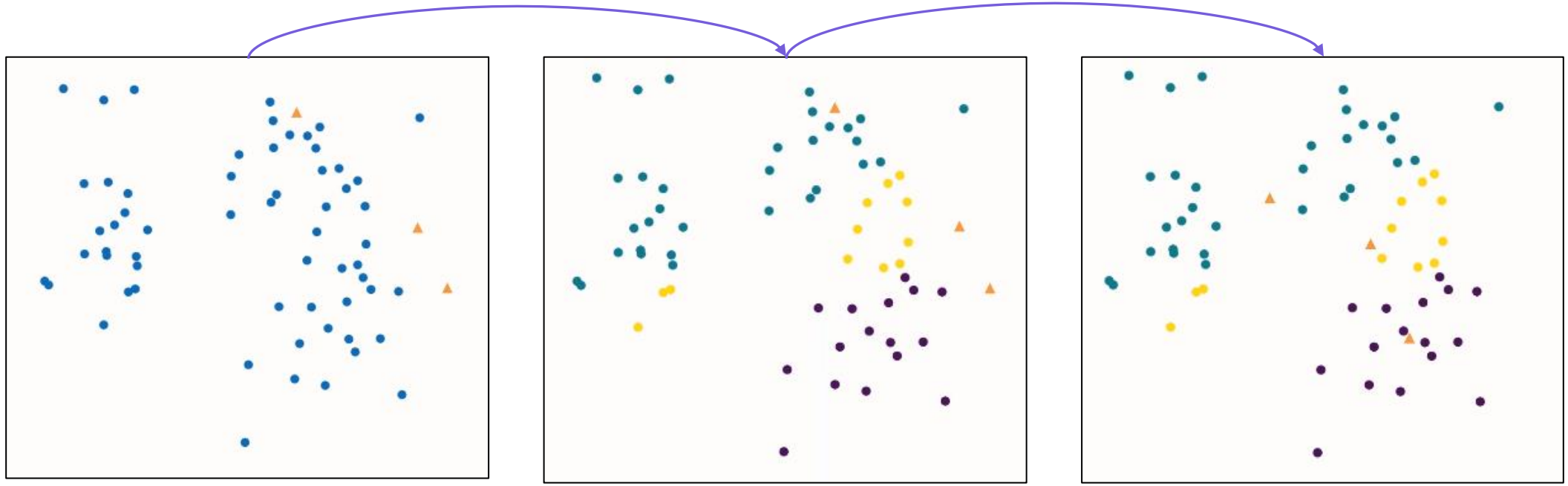
- Feature de catégories

(jeu de données très homogène, l'algorithme prend trop en compte les catégories)

4. Mise en place et sélection des algorithmes de clustering

k-means – Principe :

Exemple pour 3 clusters :



1) Mise en place aléatoire de 3 centroïdes

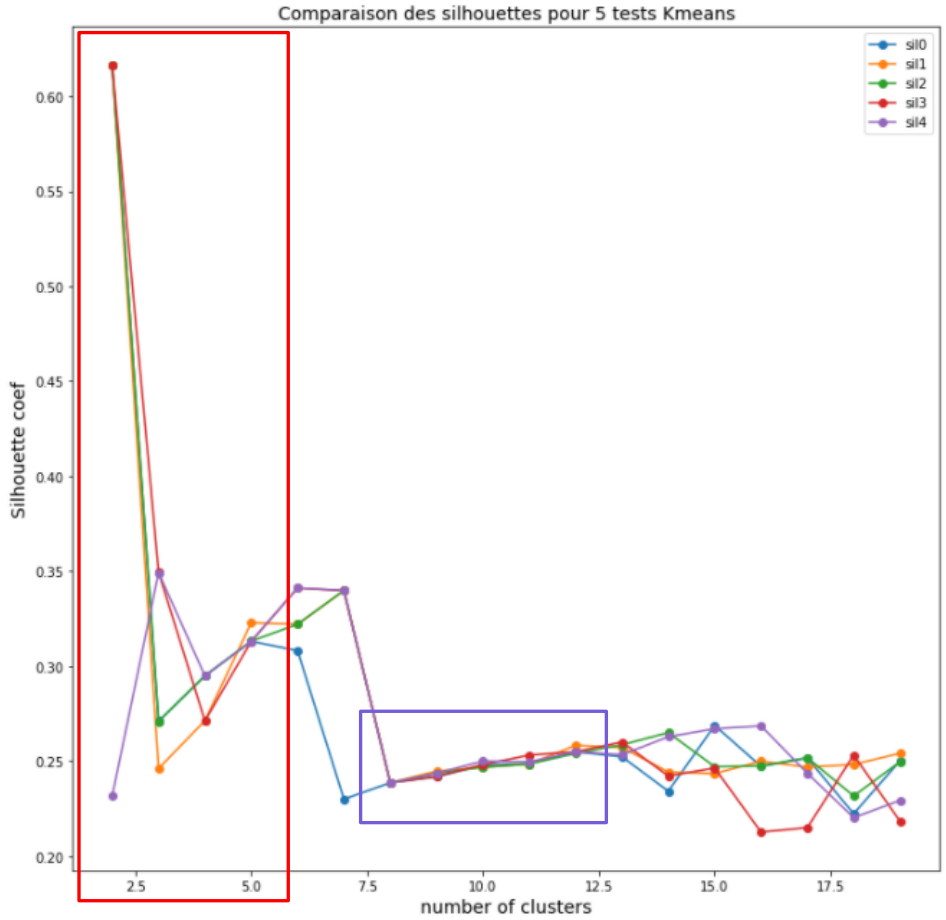
2) Assigner chaque point au cluster dont le centroïde est le plus proche

3) Recalculer les centroïdes des clusters ainsi créés

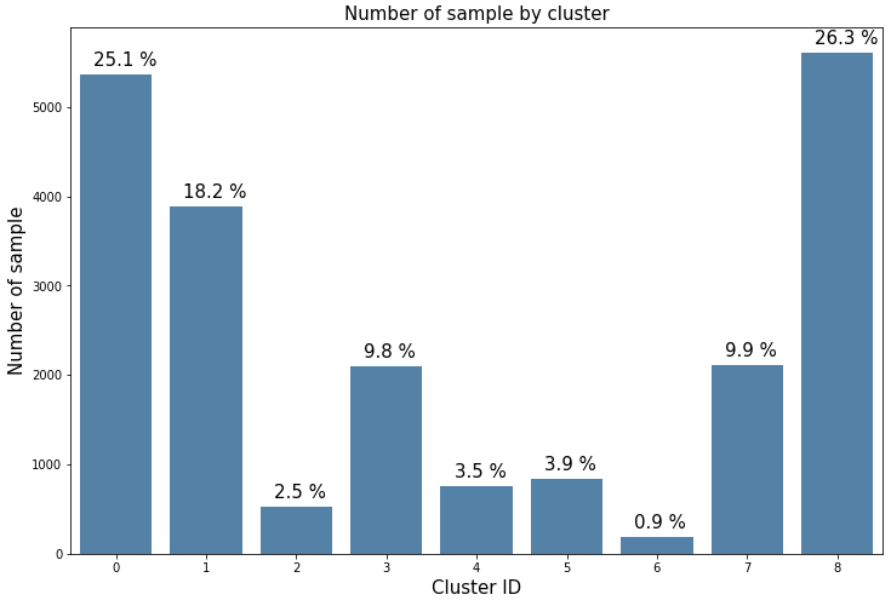
Réitérer 2) et 3) jusqu'à convergence des centroïdes

4. Mise en place et sélection des algorithmes de clustering

k-means :



- Silhouette intéressante mais clusters non-pertinents (mauvaise répartition)

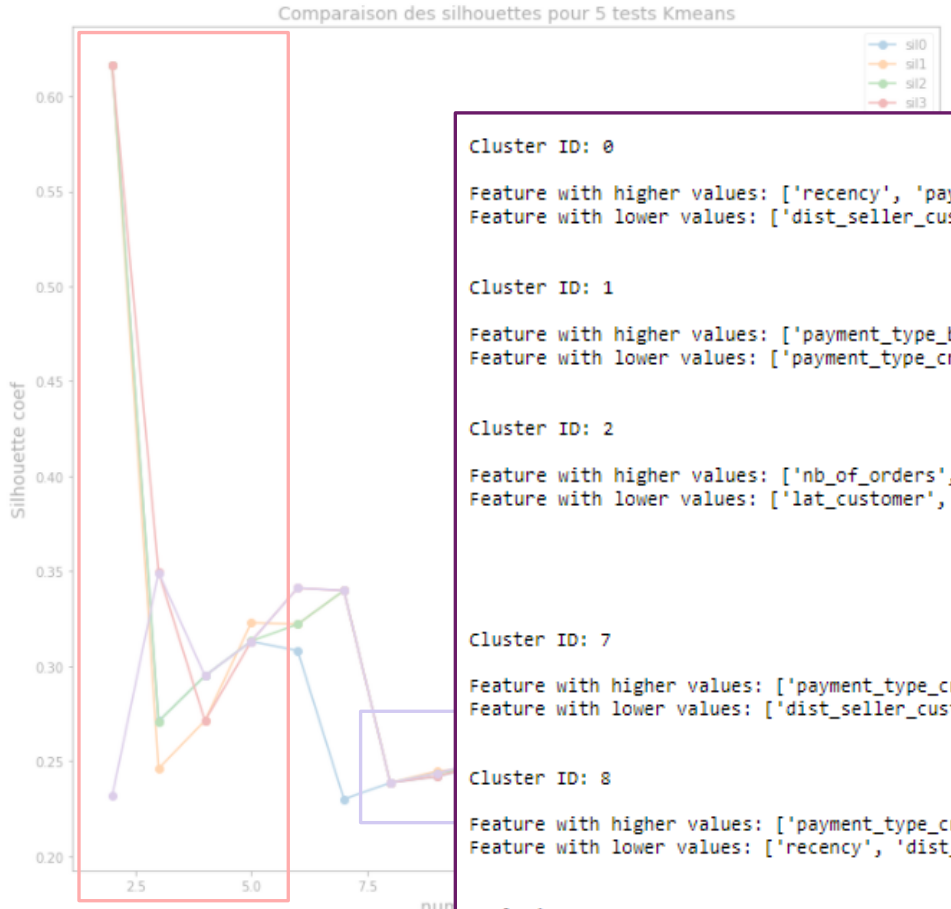


- Zone plus stable

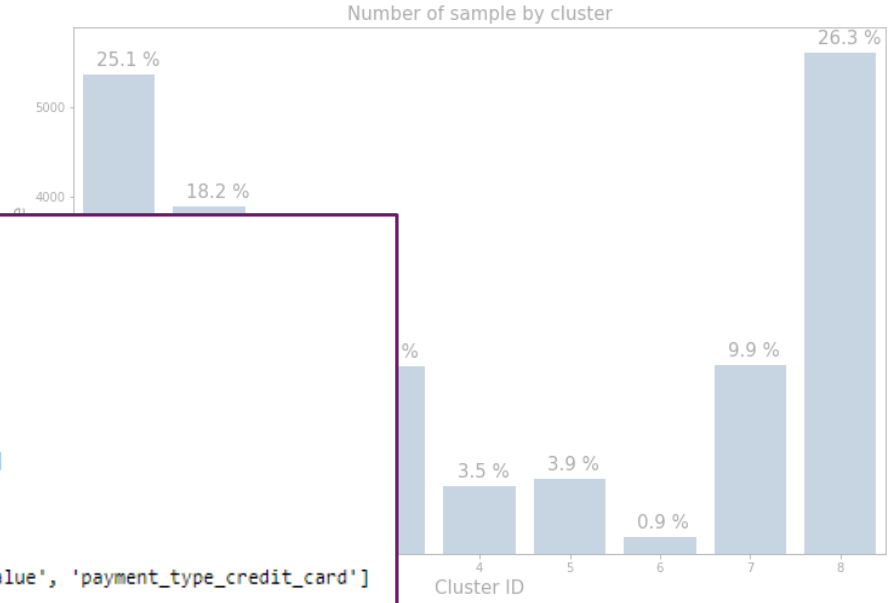
```
Cluster ID: 0

Feature with higher values: ['recency', 'payment_type_credit_card']
Feature with lower values: ['dist_seller_customer', 'lat_customer', 'long_customer']


Cluster ID: 1

Feature with higher values: ['payment_type_boleto']
Feature with lower values: ['payment_type_credit_card', 'lat_customer', 'long_customer']


Cluster ID: 2

Feature with higher values: ['nb_of_orders', 'nb_of_items', 'nb_of_reviews', 'payment_value', 'payment_type_credit_card']
Feature with lower values: ['lat_customer', 'long_customer']

[...]

Cluster ID: 7

Feature with higher values: ['payment_type_credit_card']
Feature with lower values: ['dist_seller_customer', 'review_mean_score', 'lat_customer', 'long_customer']


Cluster ID: 8

Feature with higher values: ['payment_type_credit_card']
Feature with lower values: ['recency', 'dist_seller_customer', 'lat_customer', 'long_customer']


Analysis Done
```

4. Mise en place et sélection des algorithmes de clustering

k-means :



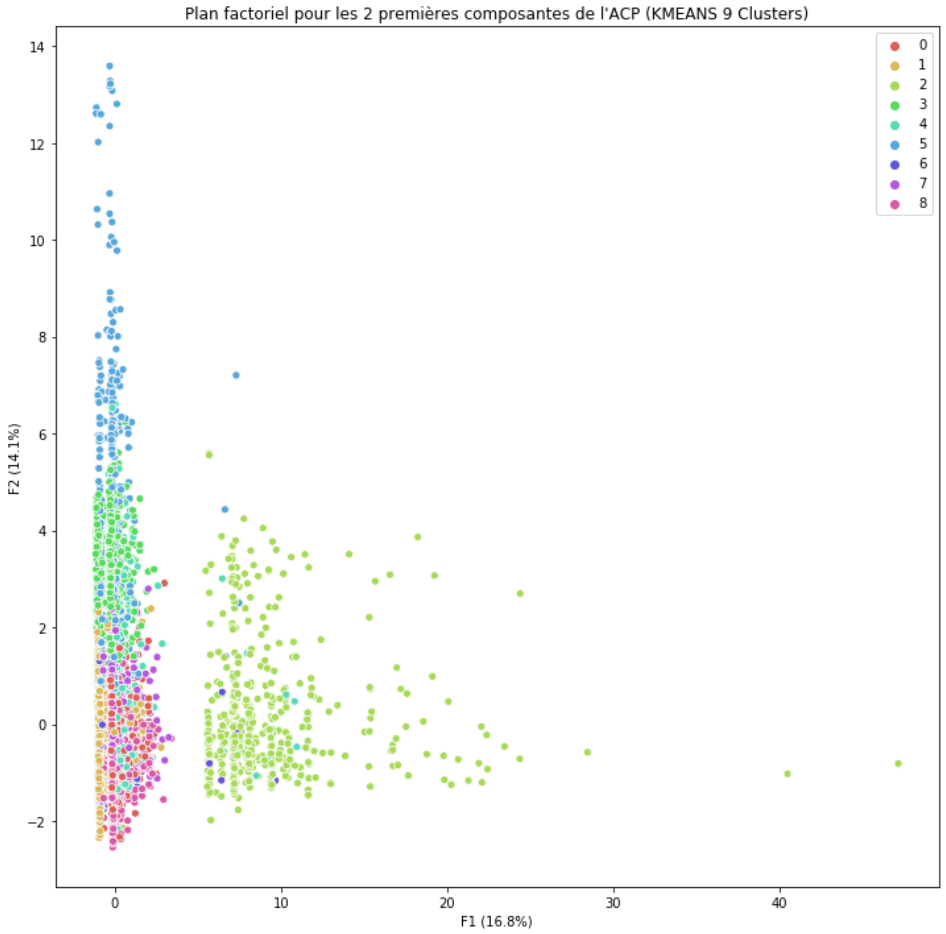

```
Cluster ID: 0

Feature with higher values: ['recency', 'payment_type_credit_card']
Feature with lower values: ['dist_seller_customer', 'lat_customer', 'long_customer']


Cluster ID: 1

Feature with higher values: ['payment_type_boleto']
Feature with lower values: ['payment_type_credit_card', 'lat_customer', 'long_customer']


Cluster ID: 2

Feature with higher values: ['nb_of_orders', 'nb_of_items', 'nb_of_reviews', 'payment_value', 'payment_type_credit_card']
Feature with lower values: ['lat_customer', 'long_customer']

[...]


Cluster ID: 7

Feature with higher values: ['payment_type_credit_card']
Feature with lower values: ['dist_seller_customer', 'review_mean_score', 'lat_customer', 'long_customer']


Cluster ID: 8

Feature with higher values: ['payment_type_credit_card']
Feature with lower values: ['recency', 'dist_seller_customer', 'lat_customer', 'long_customer']


Analysis Done
```

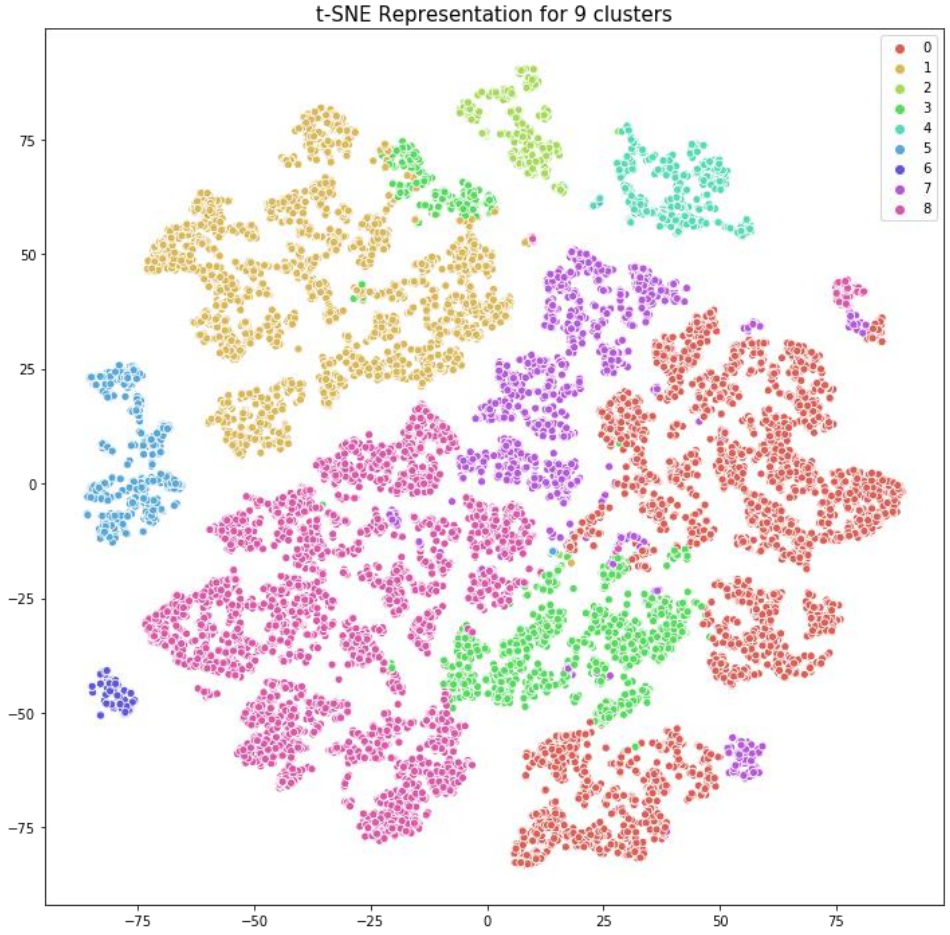


- Silhouette intéressante mais clusters non-pertinents (mauvaise répartition)
- Zone plus stable

4. Mise en place et sélection des algorithmes de clustering

k-means – 9 clusters :

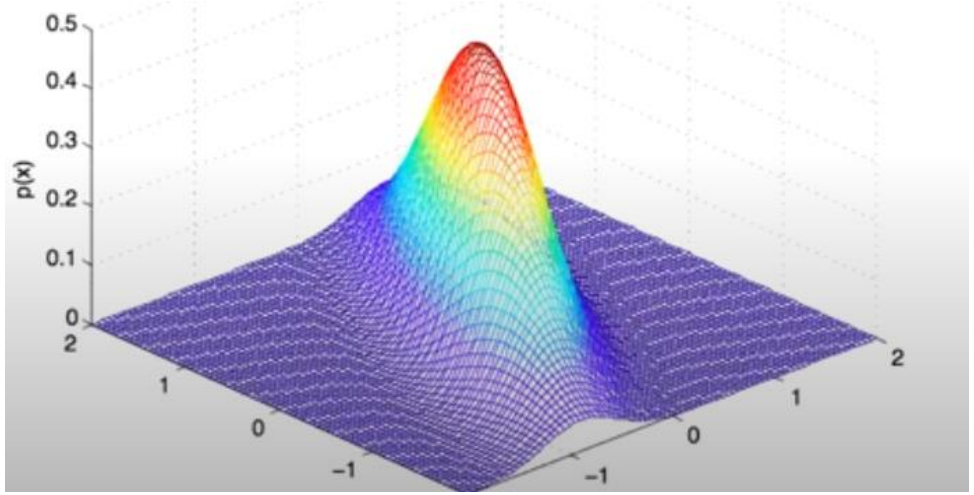




4. Mise en place et sélection des algorithmes de clustering

Gaussian Mixture Model – Principe:

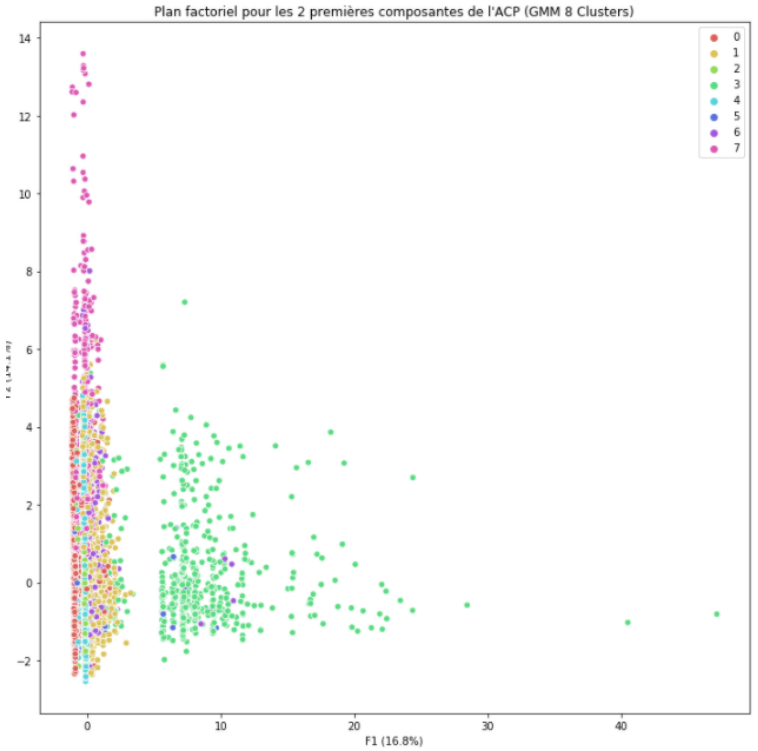
Méthode assez similaire au k-means, à la différence qu'ici on va considérer que les clusters sont distribués en suivant une loi normale (gaussiennes)
=> On ne va donc plus construire nos clusters via la distance euclidienne des points au centroïde mais en se basant sur le principe de maximum de vraisemblance

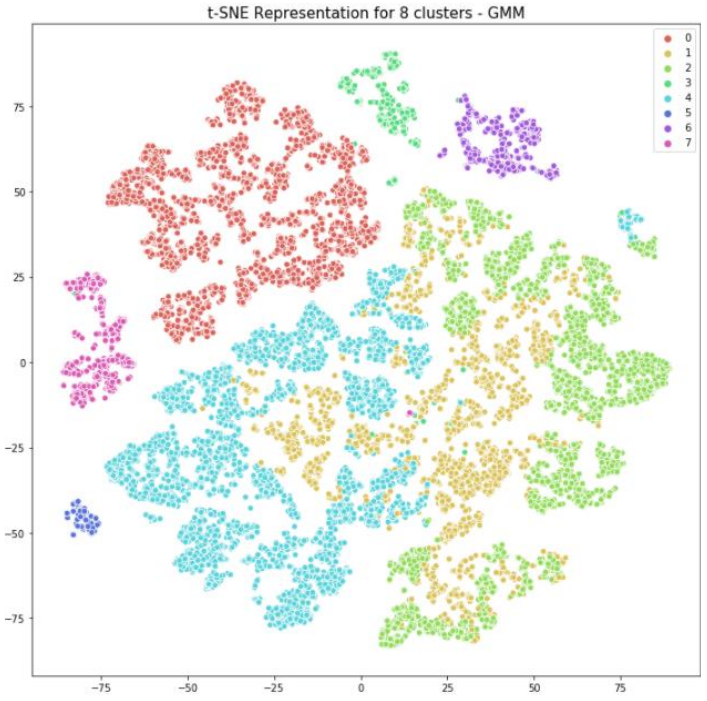


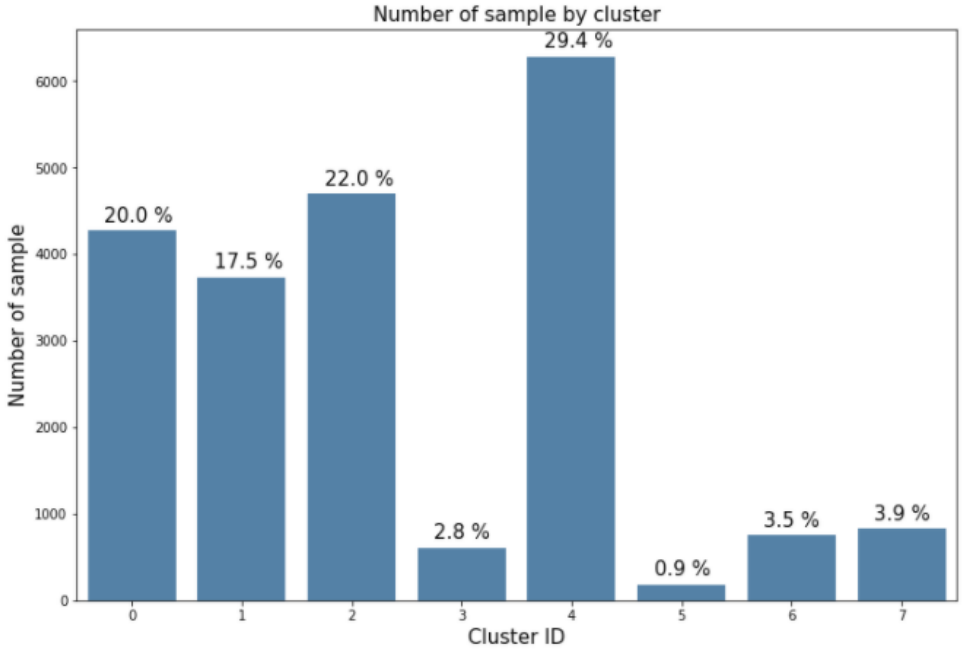
Grâce à cette méthode, on ne se limite donc plus à des clusters sphériques !

4. Mise en place et sélection des algorithmes de clustering

Gaussian Mixture Model – 8 clusters :

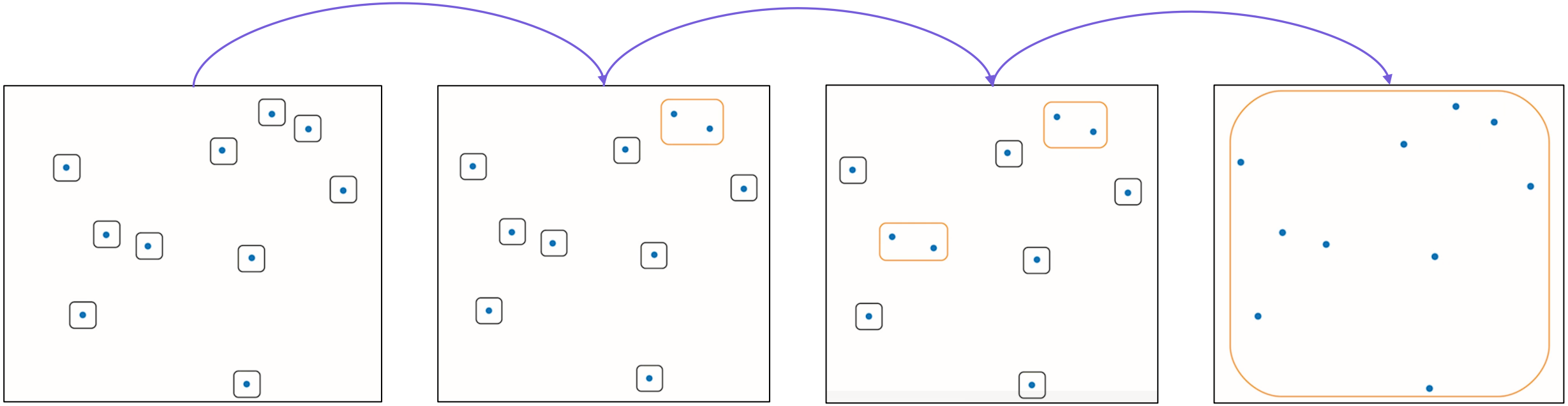






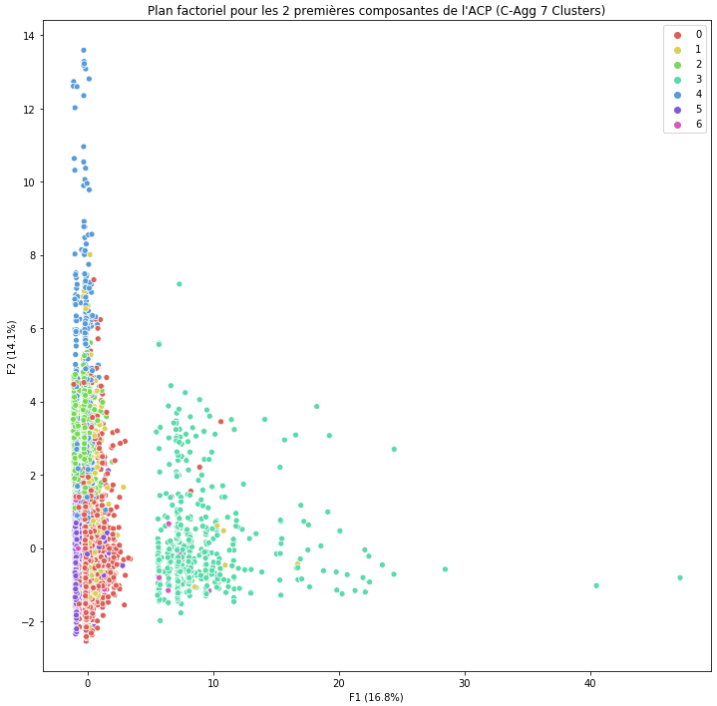
4. Mise en place et sélection des algorithmes de clustering

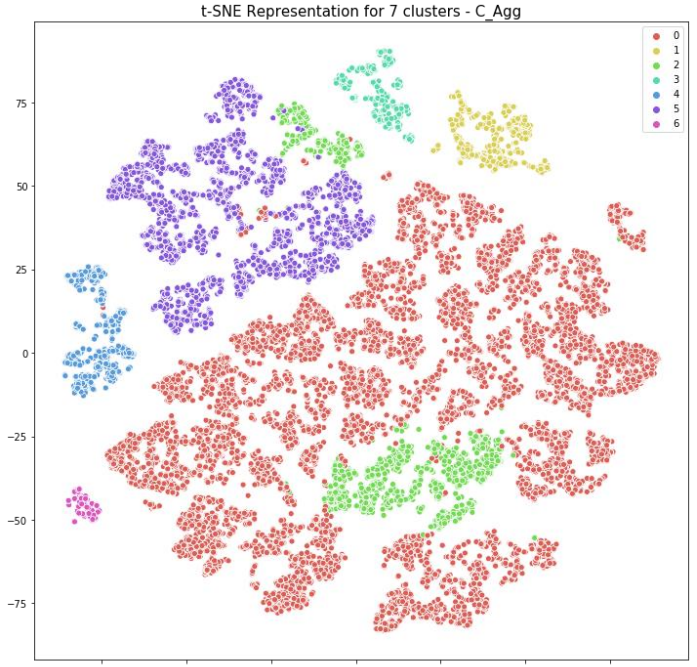
Agglomerative Clustering – Principe :

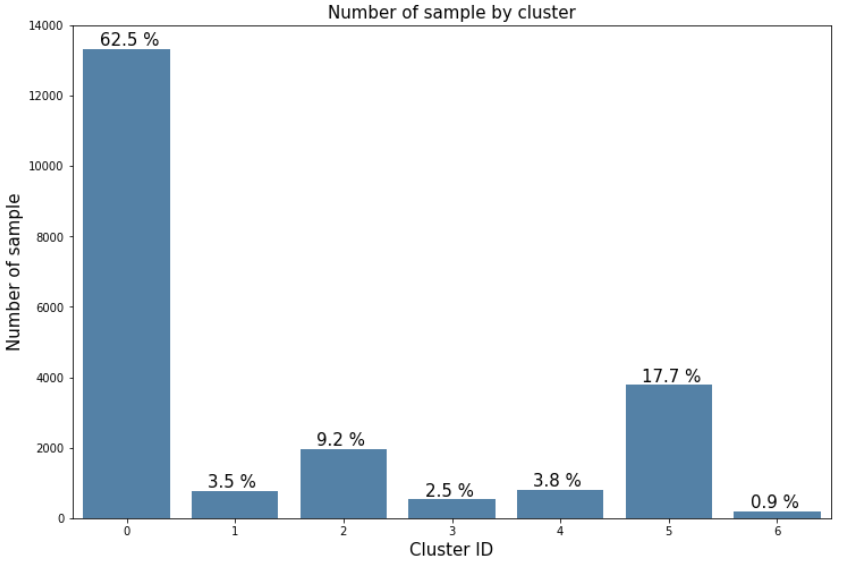


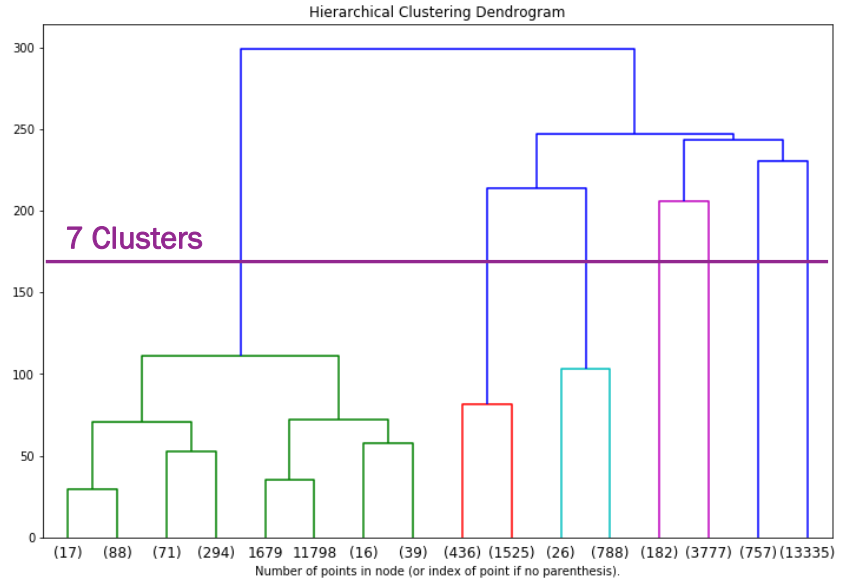
4. Mise en place et sélection des algorithmes de clustering

Agglomerative Clustering– 7 clusters :



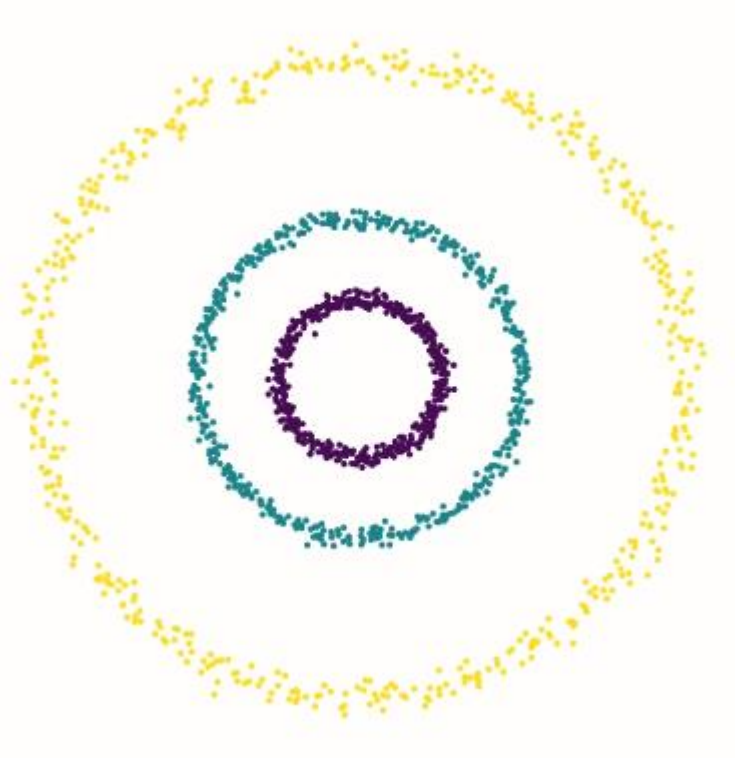


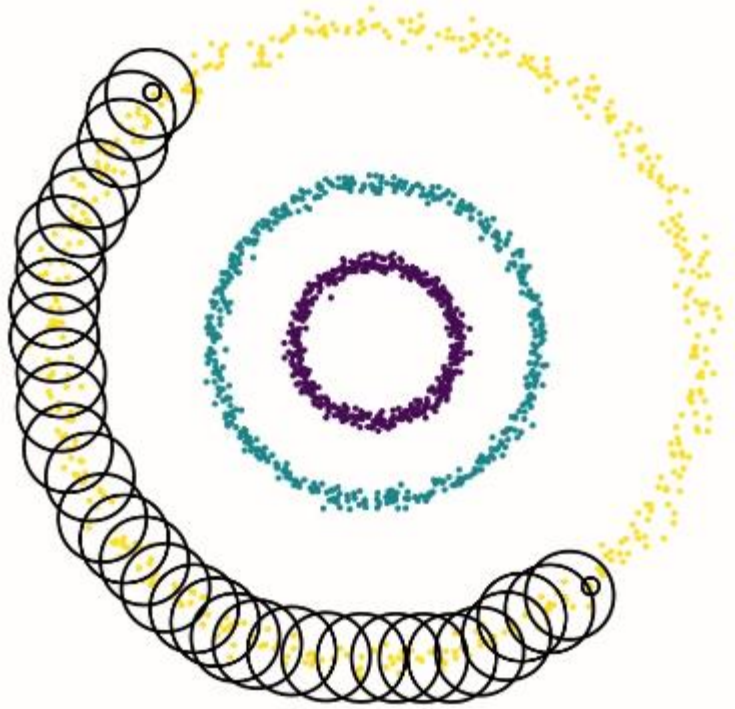




4. Mise en place et sélection des algorithmes de clustering

DBSCAN – Principe :



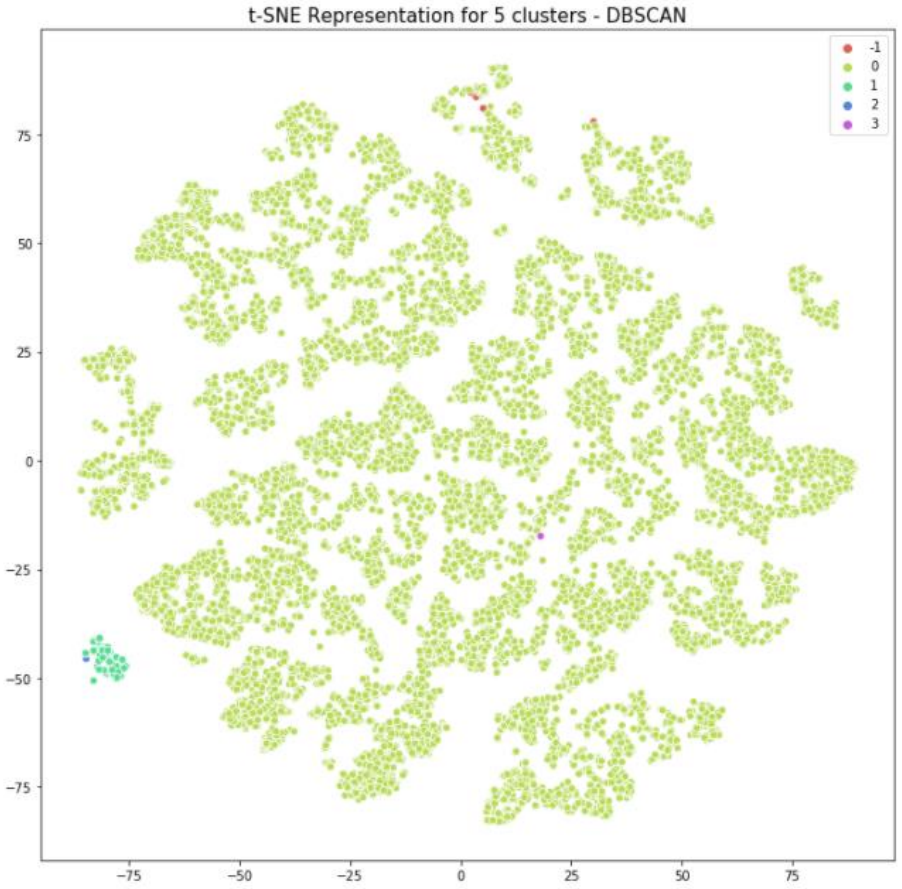


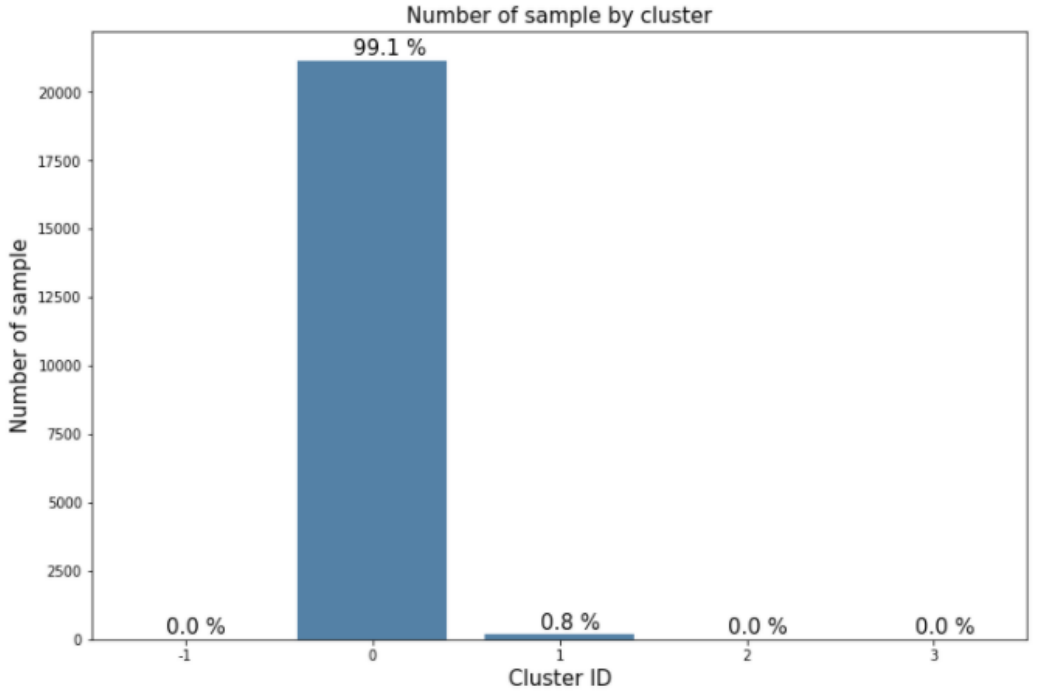
Epsilon Voisinage :
=> Pour chaque point d'un cluster, on doit pouvoir trouver n-voisins à une distance epsilon du point observé

4. Mise en place et sélection des algorithmes de clustering

DBSCAN :

	eps	min_samples	sil	nb_clusters
17	7	5	0.616222	3
16	7	3	0.615799	4
19	7	9	0.615428	3
18	7	7	0.615428	3
21	9	3	0.61525	5
24	9	9	0.61518	3
23	9	7	0.615117	3
22	9	5	0.615055	3
20	9	1	0.61368	13
15	7	1	0.608472	25
11	5	3	0.473123	10
12	5	5	0.472512	8
13	5	7	0.471993	7
14	5	9	0.471819	7
10	5	1	0.46604	64
8	2	7	0.231853	19
7	2	5	0.230956	23
9	2	9	0.228972	17
6	2	3	0.222736	34
5	2	1	0.140804	515
2	1	5	0.0354588	60
3	1	7	0.0336575	44
4	1	9	0.0308664	35
1	1	3	0.00543526	113
0	1	1	-0.163533	2226

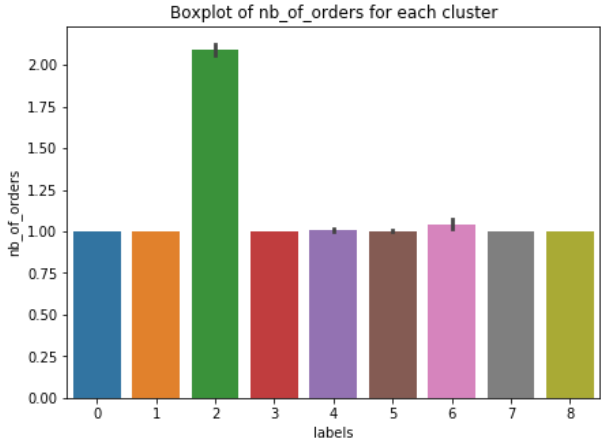


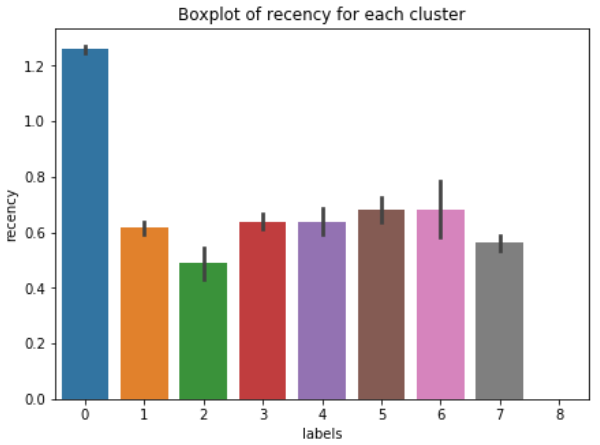


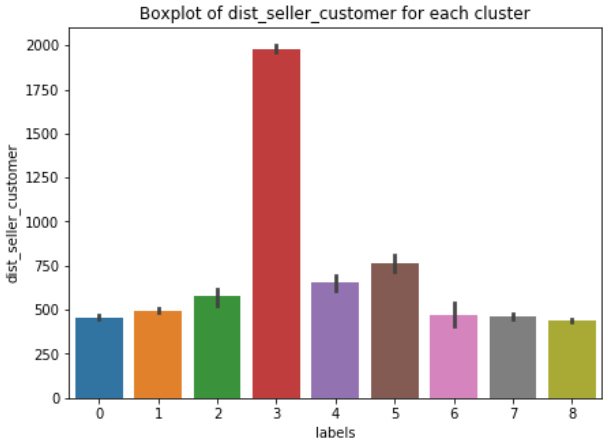
5. Analyse du modèle retenu

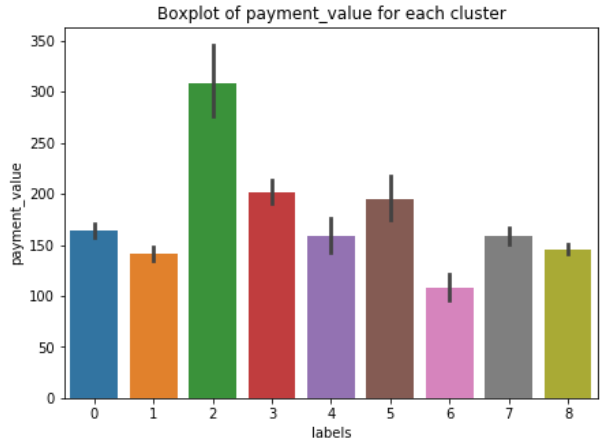
Solution retenue : k-means – 9 clusters

Cluster ID	Description
0	Client non-récent
1	Client utilisant Boleto
2	Client à haut volume de commande
3	Client à longue distance / haut temps de livraison
4	Client utilisant carte de crédit et vouchers
5	Client non satisfait et livré en retard
6	Client utilisant une carte de débit
7	Client généralement non satisfait
8	Client récent





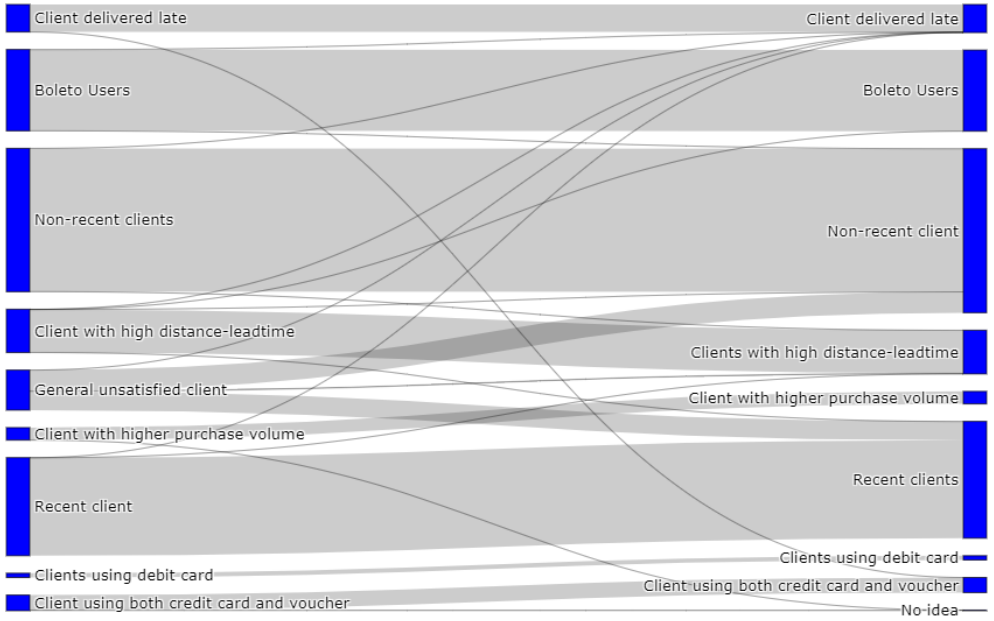




5. Analyse du modèle retenu

Etude temporelle:

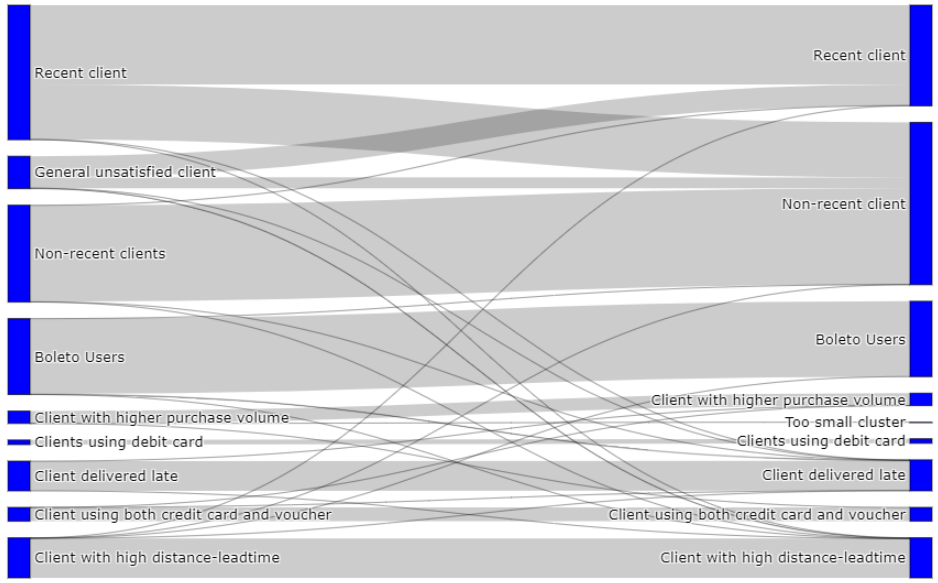
Transferts entre clusters prédits et cluster fittés sur les données jusque Nov 2017 (3 mois d'écart avec données de base)




```
print('Adjusted Rand Score: {}'.format(metrics.adjusted_rand_score(test_cls9.labels_,test_labels)))
```

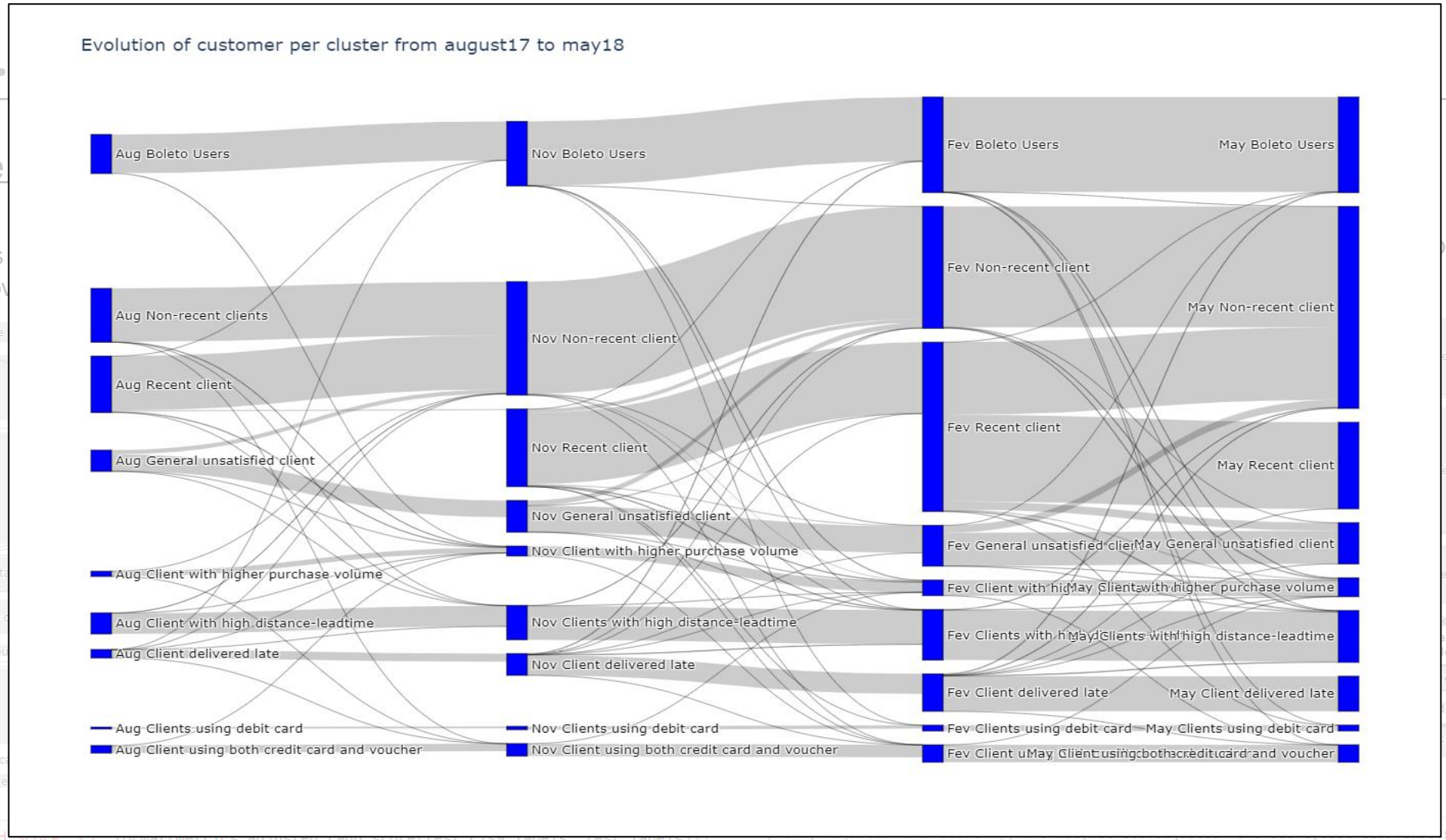
```
Adjusted Rand Score: 0.8513350114694752
```

Transferts entre clusters prédits et cluster fittés sur les données jusque Fév 2018 (6 mois d'écart avec données de base)




```
print('Adjusted Rand Score: {}'.format(metrics.adjusted_rand_score(test_cls9.labels_,test_labels_fev)))
```

```
Adjusted Rand Score: 0.5837009151575708
```



Le modèle semble être assez stable dans le temps

Proposition d'update tous les 3 mois (pour aligner l'update avec la variation de la feature de récence)

6. Conclusion sur les modèles et la problématique

- Modèle sélectionné semble pertinent, première base à appliquer par l'équipe marketing :
  - ❑ Gestion des clients mécontents
  - ❑ Proposition de livraison express pour les clients éloignés
  - ❑ Relance des clients non-récents
  - ❑ Etc …

- Comparer avec un modèle de clustering manuel orienté métier (via système de scoring) :
  - ❑ Score de review faible
  - ❑ Clients récents
  - ❑ Clients avec le plus de dépenses
  - ❑ Clients éloignés

- Le modèle pourra être amélioré lorsque la base de données sera plus consistante :
  - ❑ Ré-implantation de la composante temporelle lorsque plusieurs cycles annuels seront disponibles
  - ❑ Augmentation des clients fidèles (clients avec plusieurs commandes)
  - ❑ Ajout d'informations (âge, sexe, …)

Merci de votre attention