Formation Data Scientist OpenClassrooms

Projet 5 – Livrable 3 – Support de présentation

Segmentez des clients d'un site e-commerce

(Project commencé avant le 1er Décembre 2012)

Etudiant : Monine Chan

Evaluateur : Alexandre Gazagnes

Vendredi 18 Mars 2022

PLAN

1. Présentation de la problématique, interprétation et pistes de recherche.

2. Nettoyage des données, feature engineering et exploration.

3. Présentation des segmentations effectuées.

4. Fréquence de mise à jour de la segmentation

PLAN

1. Présentation de la problématique, interprétation et pistes de recherche.

2. Nettoyage des données, feature engineering et exploration.

3. Présentation des segmentations effectuées.

4. Fréquence de mise à jour de la segmentation

1. PRÉSENTATION DE LA PROBLEMATIQUE

CONTEXTE

- L'entreprise de commerce en ligne Olist nous demande de créer une segmentation de leur client afin d'identifier les différents type d'utilisateurs et d'adapter leur campagnes de communication.

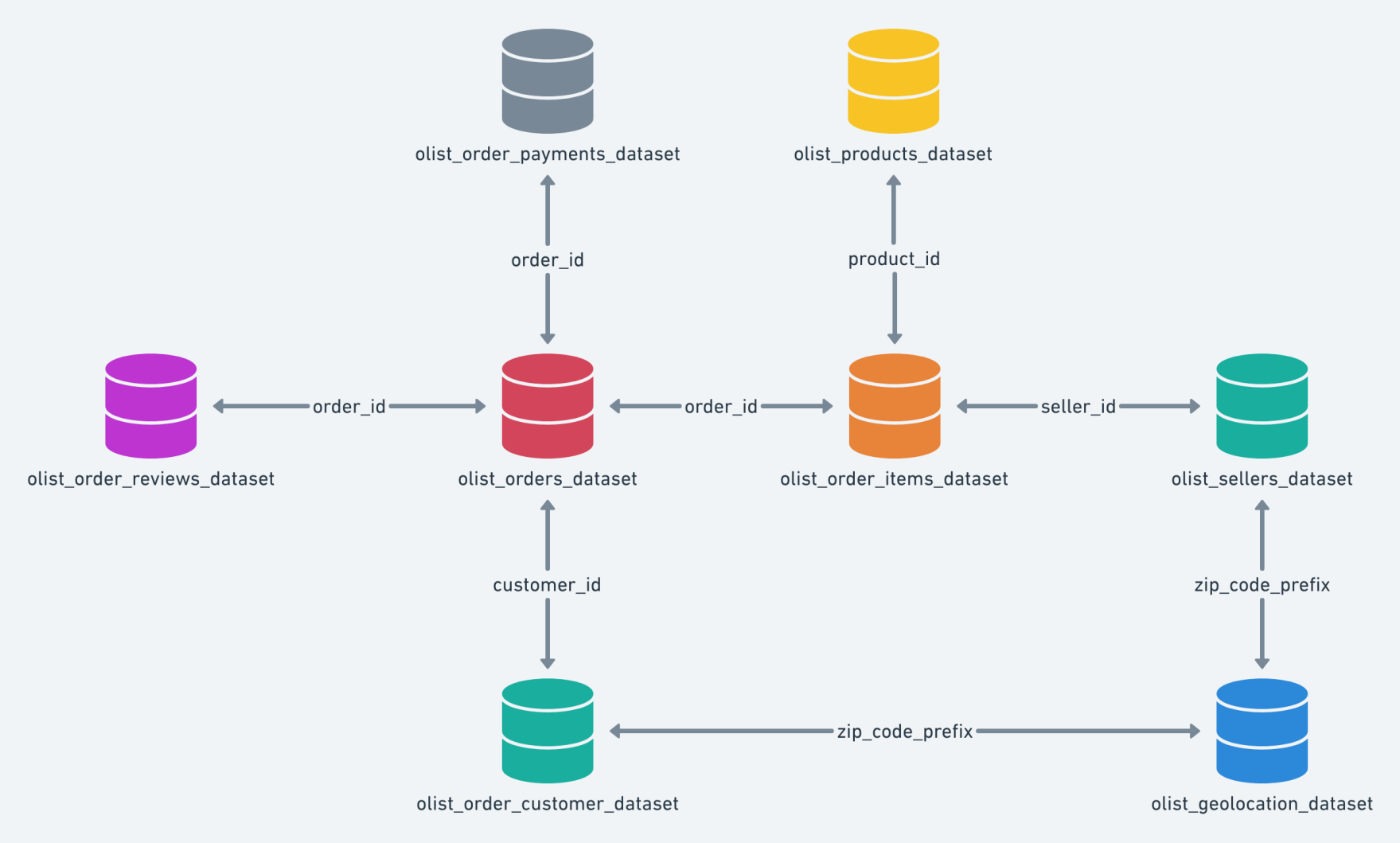
- Le but de ce projet est de:
  - Fournir à l'équipe marketing une description actionnable de la segmentation des clients,
  - Fournir une proposition de contrat maintenance de cette segmentation (i.e. étudier la stabilité de la segmentation au cours du temps).

- On mettra en place une segmentation RFM (Récence, Fréquence et Montant) dans un premier temps.

1. PRÉSENTATION DE LA PROBLEMATIQUE

SYNOPTIQUE DES DIFFERENTS JEUX DE DONNEES

- Nous allons faire une analyse exploratoire de ces différents jeux de données.



PLAN

1. Présentation de la problématique, interprétation et pistes de recherche.

2. Nettoyage des données, feature engineering et exploration.

3. Présentation des segmentations effectuées.

4. Fréquence de mise à jour de la segmentation

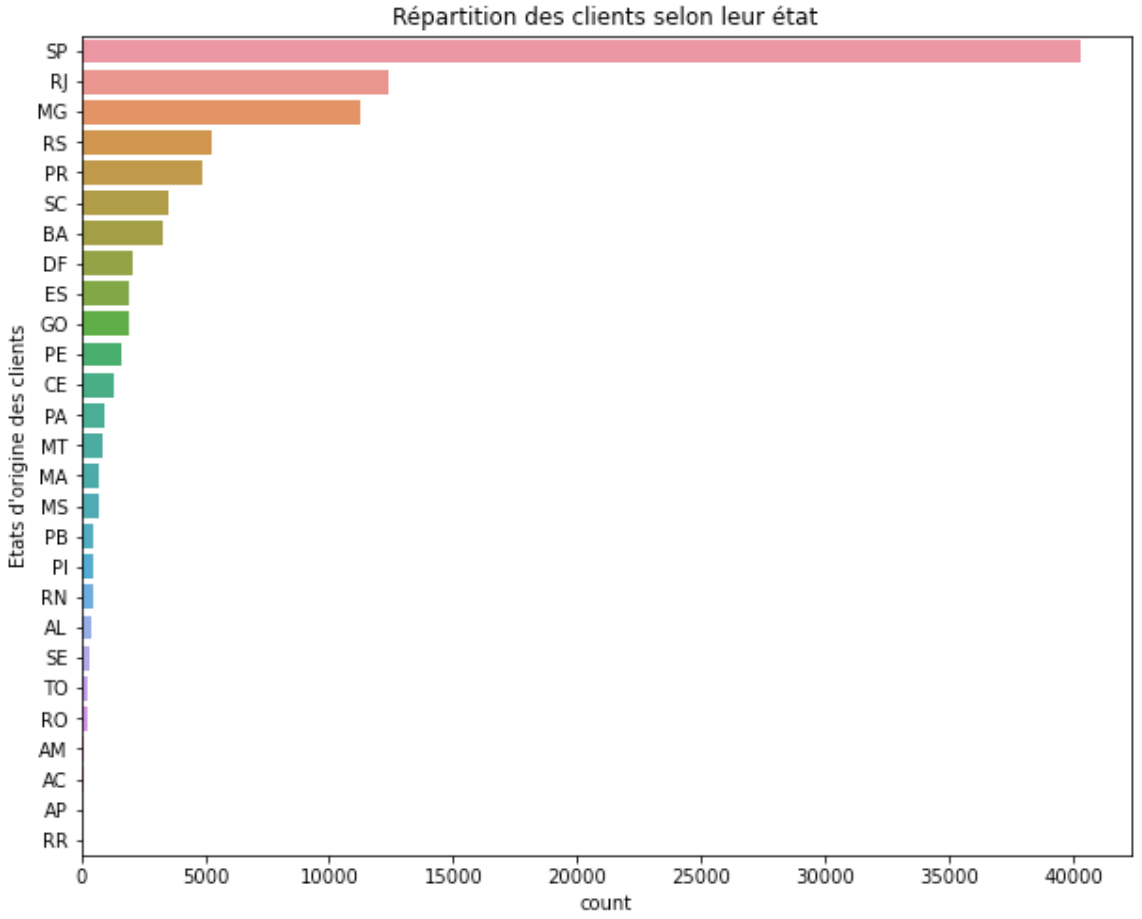
2. NETTOYAGE DE DONNEES ET EXPLORATION

DOUBLONS, ANALYSE EXPLORATOIRE

➢ Chaque client est identifié par la valeur de la variable customer_unique_id : on supprime tous les doublons liés au même customer_unique_id.

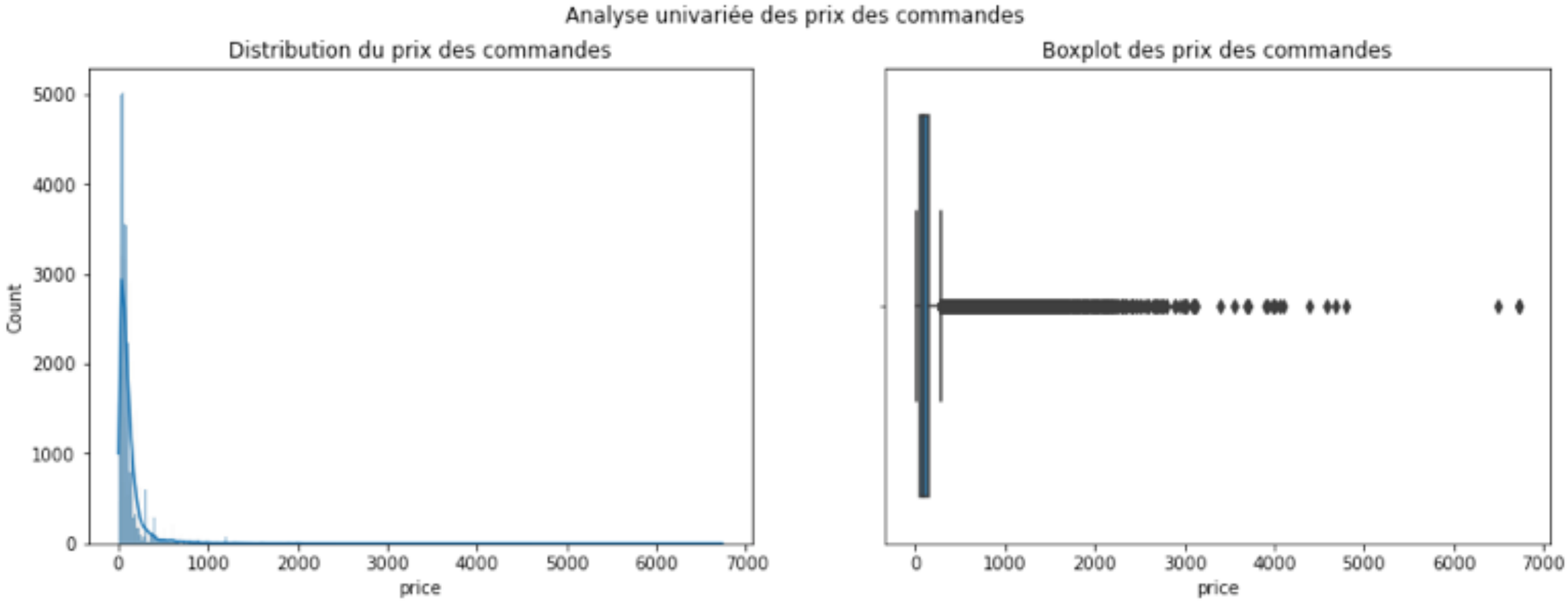
➢Les états où se trouvent le plus de clients :

- SP : Sao Paulo
- RJ : Rio de Janeiro
- MG : Minas Gerai
- RS : Rio Grande do Sul
- PR : Parana



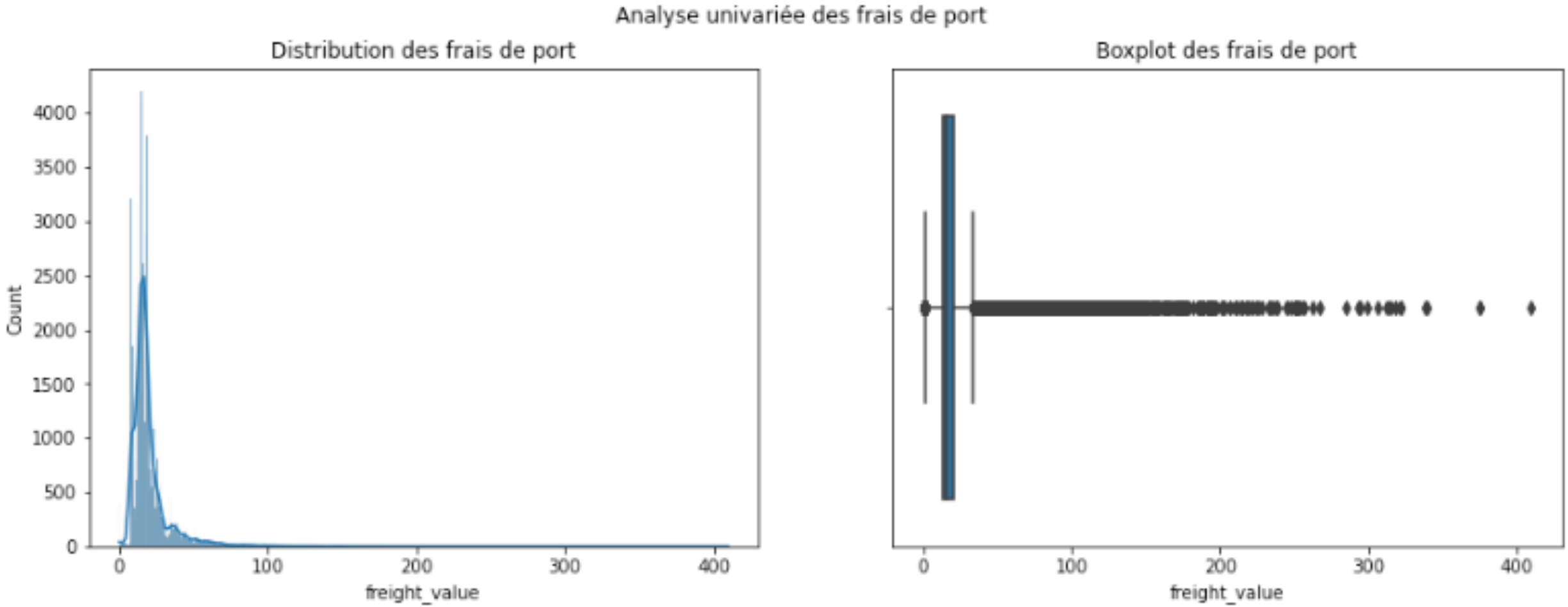
2. EXPLORATION

PRIX DES COMMANDES EN LIGNE



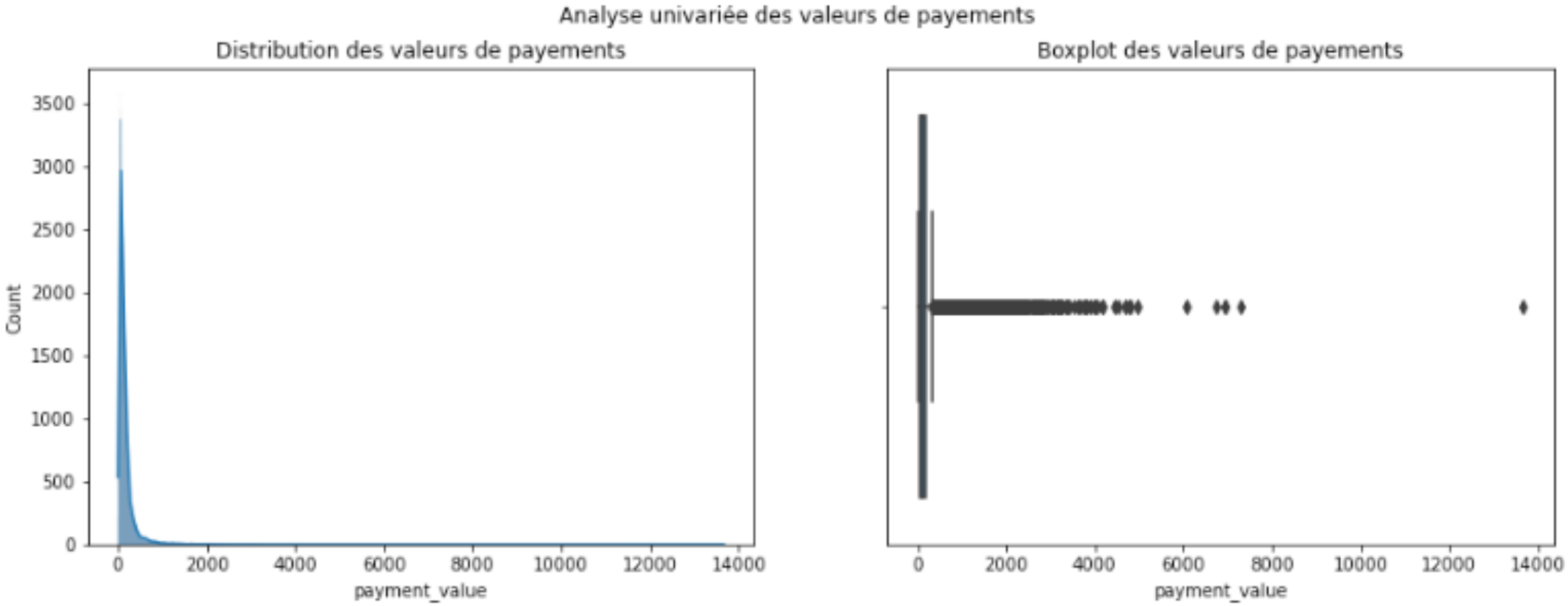
2. EXPLORATION

FRAIS DE PORT



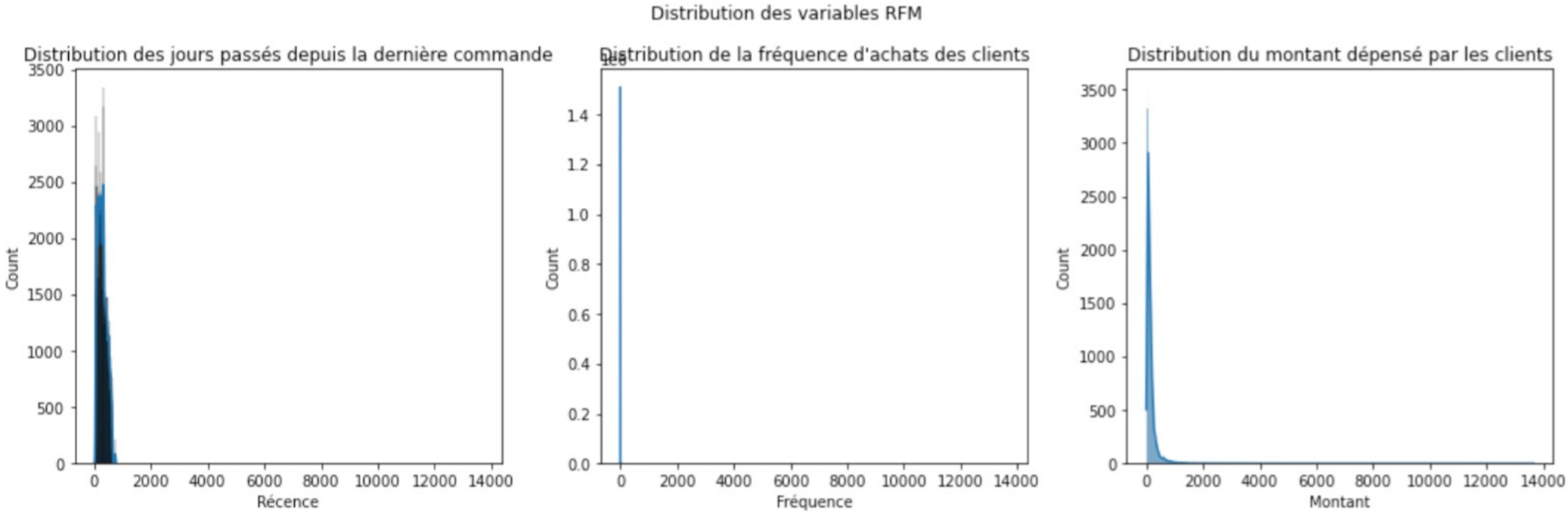
2. EXPLORATION

VALEURS DE PAYMENT



2. FEATURE ENGINEERING : CRÉATION DE NOUVELLES VARIABLES

RÉCENCE, FRÉQUENCE ET MONTANT



PLAN

1. Présentation de la problématique, interprétation et pistes de recherche.

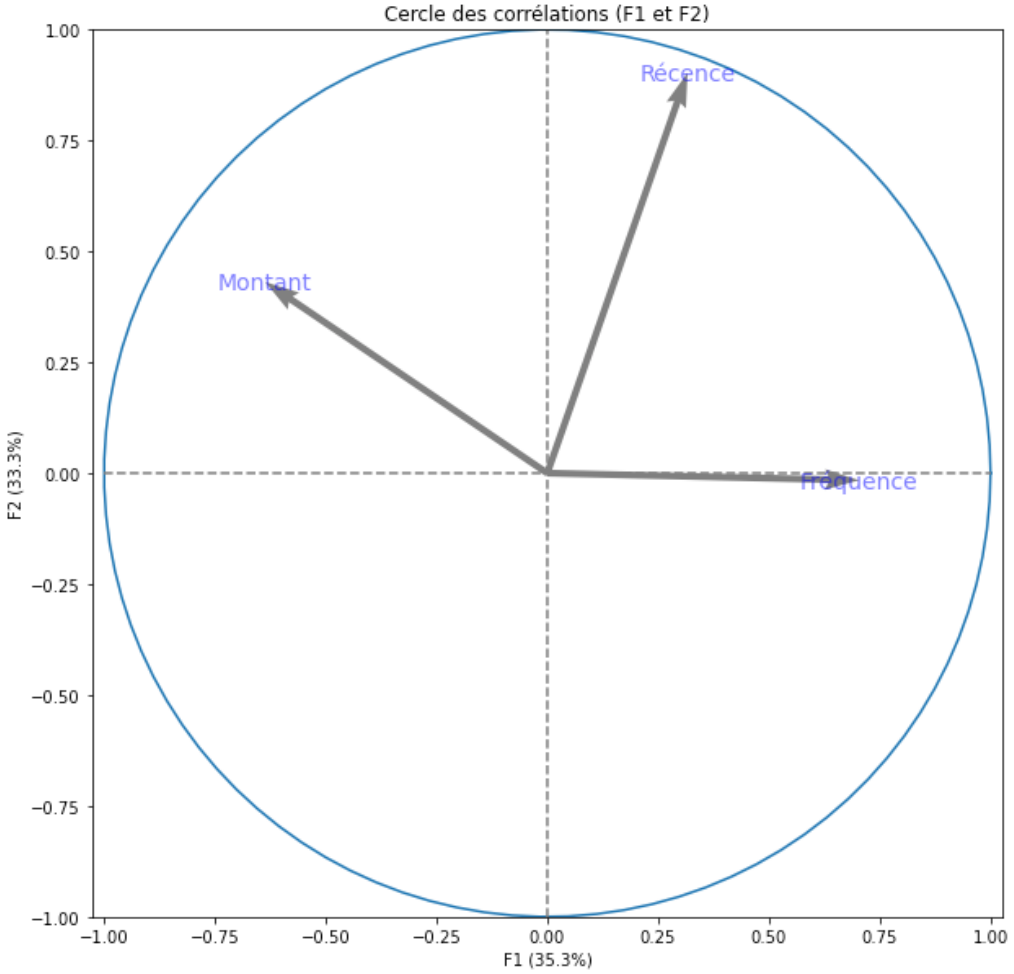
2. Nettoyage des données, feature engineering et exploration.

3. Présentation des segmentations effectuées.

4. Fréquence de mise à jour de la segmentation

3. SEGMENTATIONS EFFECTUEES

SEGMENTATION RFM AVEC K-MEANS

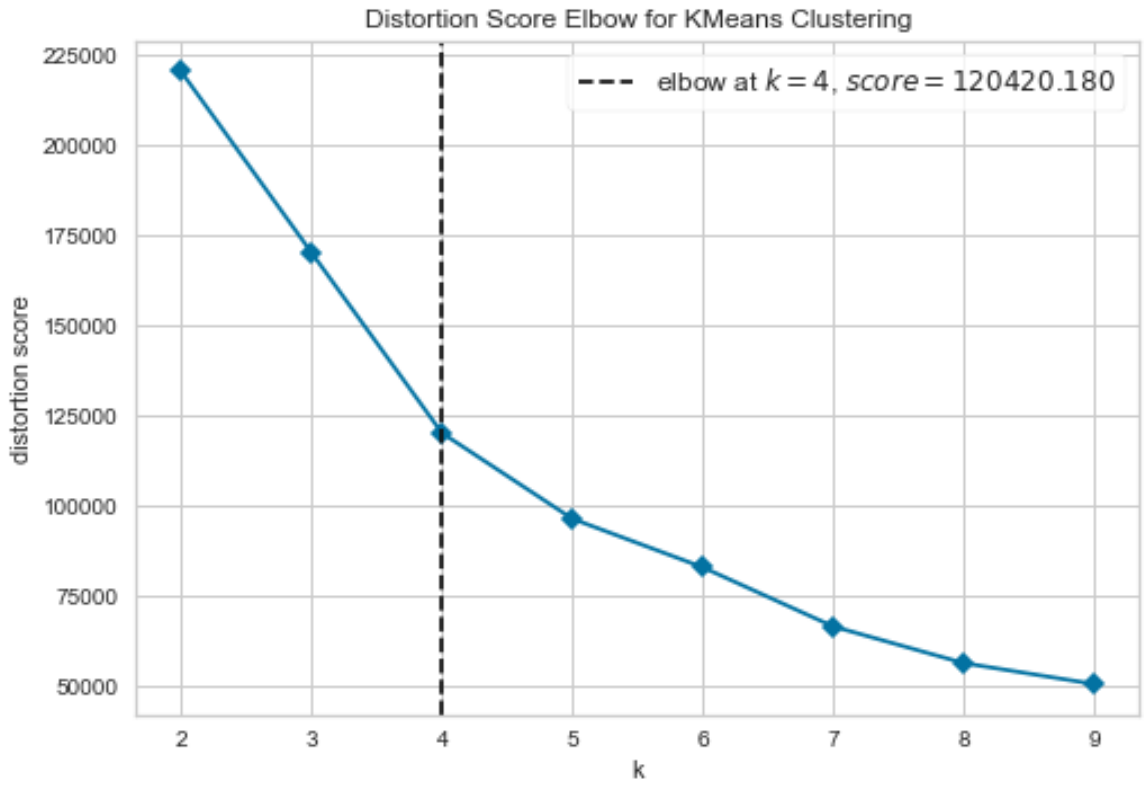


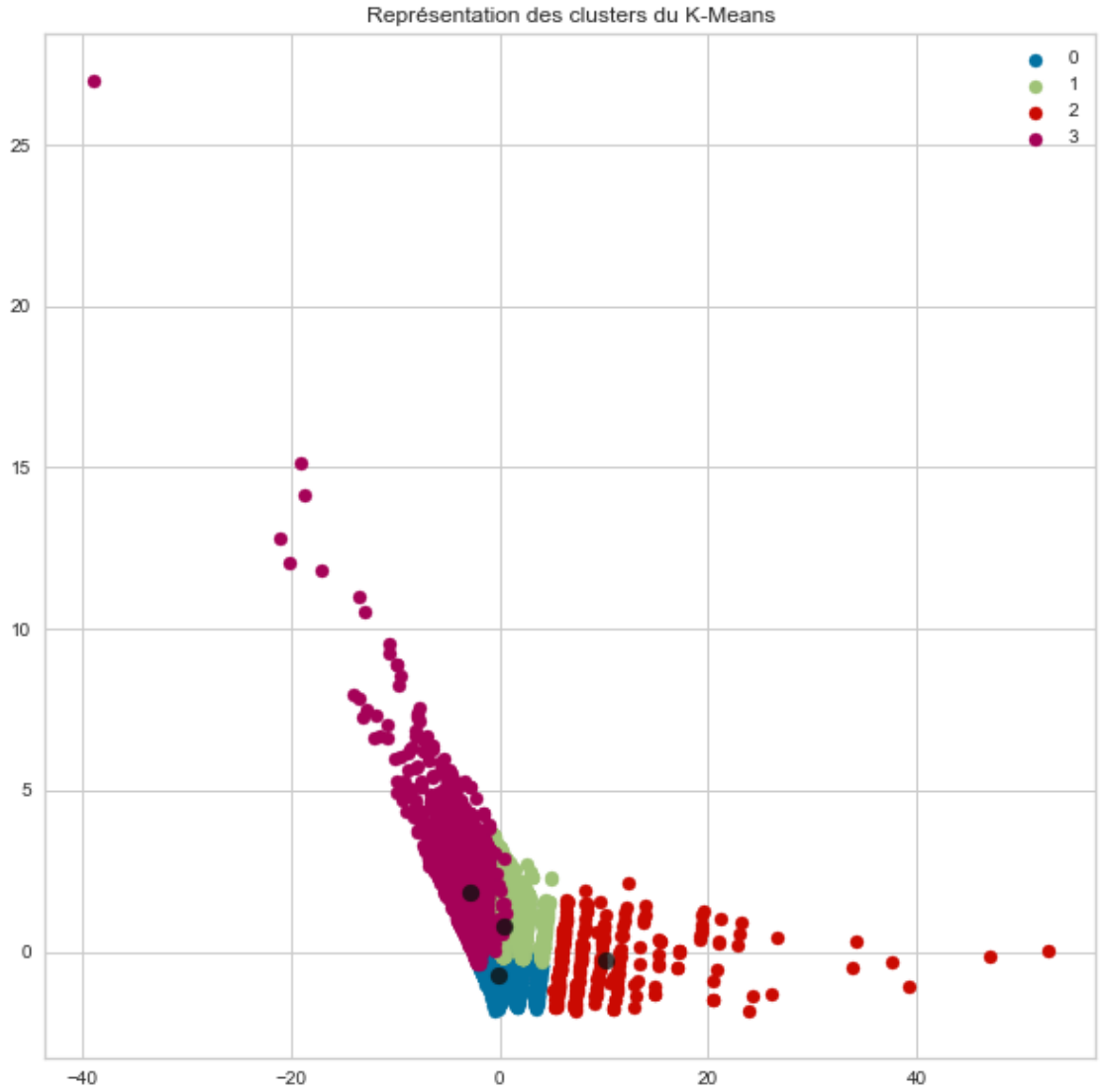
Les principaux axes d'inertie sont la fréquence d'achat et la récence.

- Segmentation RFM : Récence, Fréquence et Montant des achats
- Réduction de dimensions par ACP.
- Les axes principaux d'inertie sont la fréquence et la récence.

3. SEGMENTATIONS EFFECTUEES

SEGMENTATION RFM AVEC K-MEANS

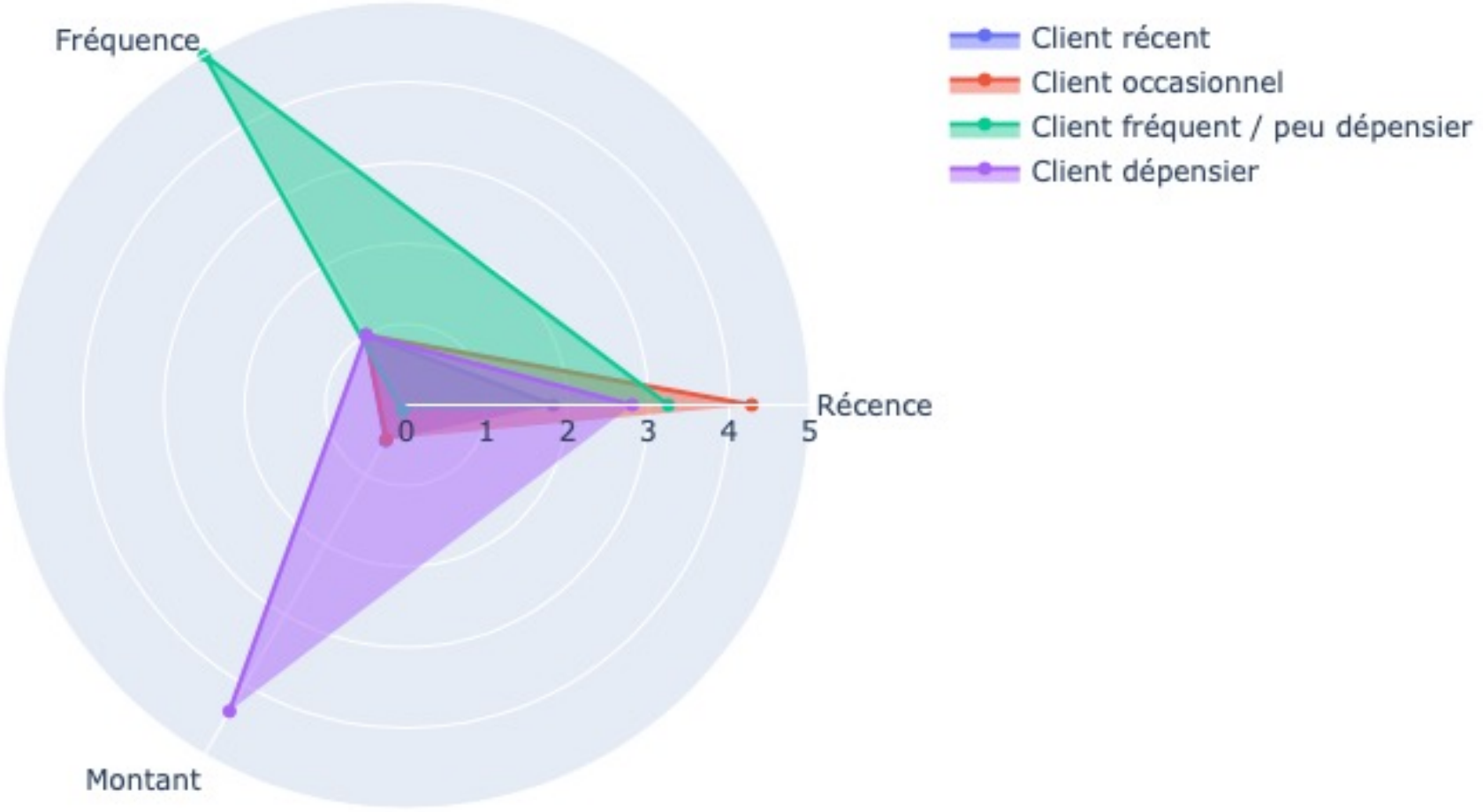




- On va segmenter la clientèle en 4 clusters d'après la méthode du coude.

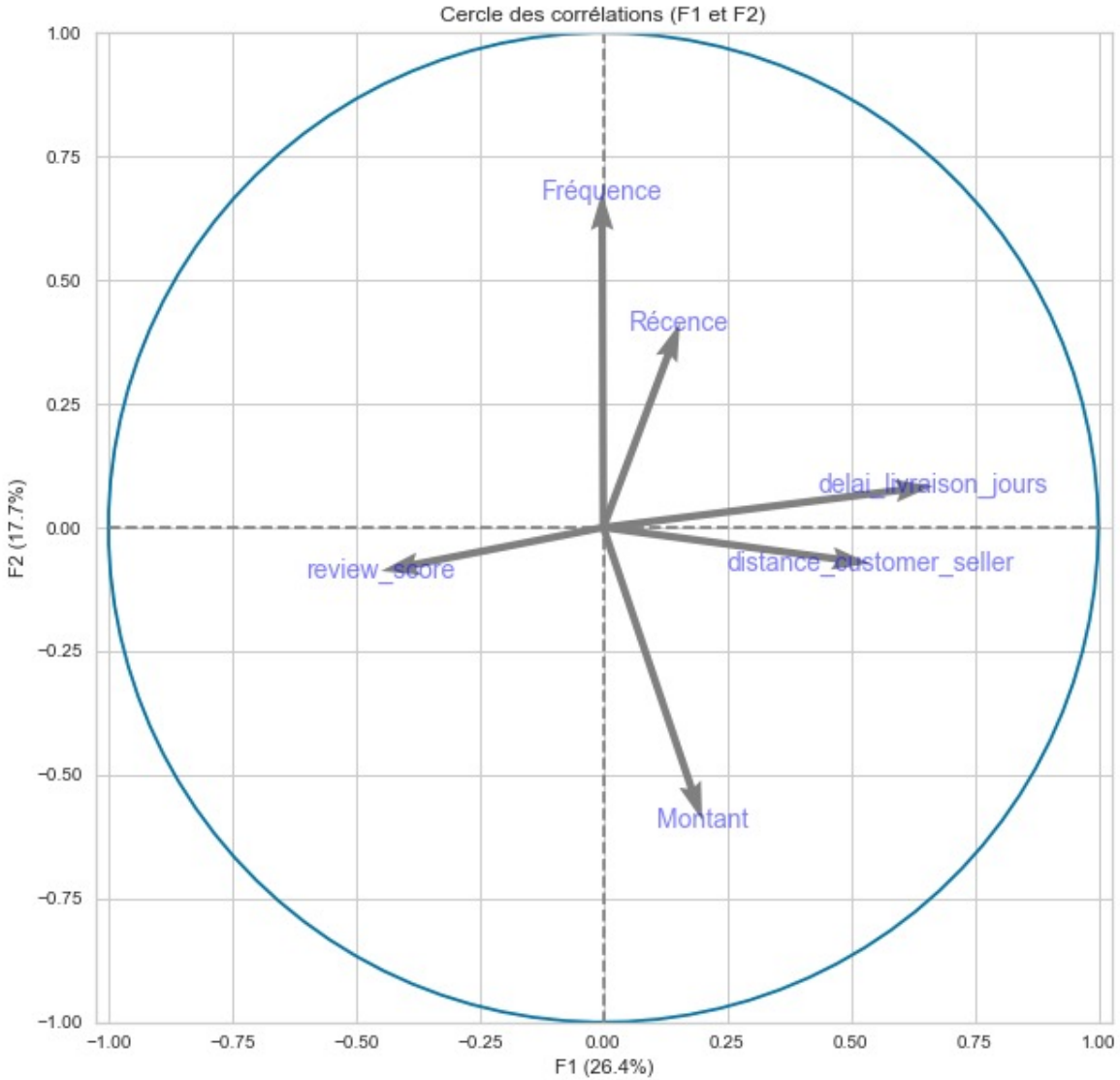
3. SEGMENTATIONS EFFECTUEES

SEGMENTATION RFM = RÉCENCE, FRÉQUENCE, MONTANT (K-MEANS)



3. SEGMENTATIONS EFFECTUEES

SEGMENTATION RFM + 3 VARIABLES QUANTITATIVES AVEC K_MEANS



- Réduction de dimensions par ACP.

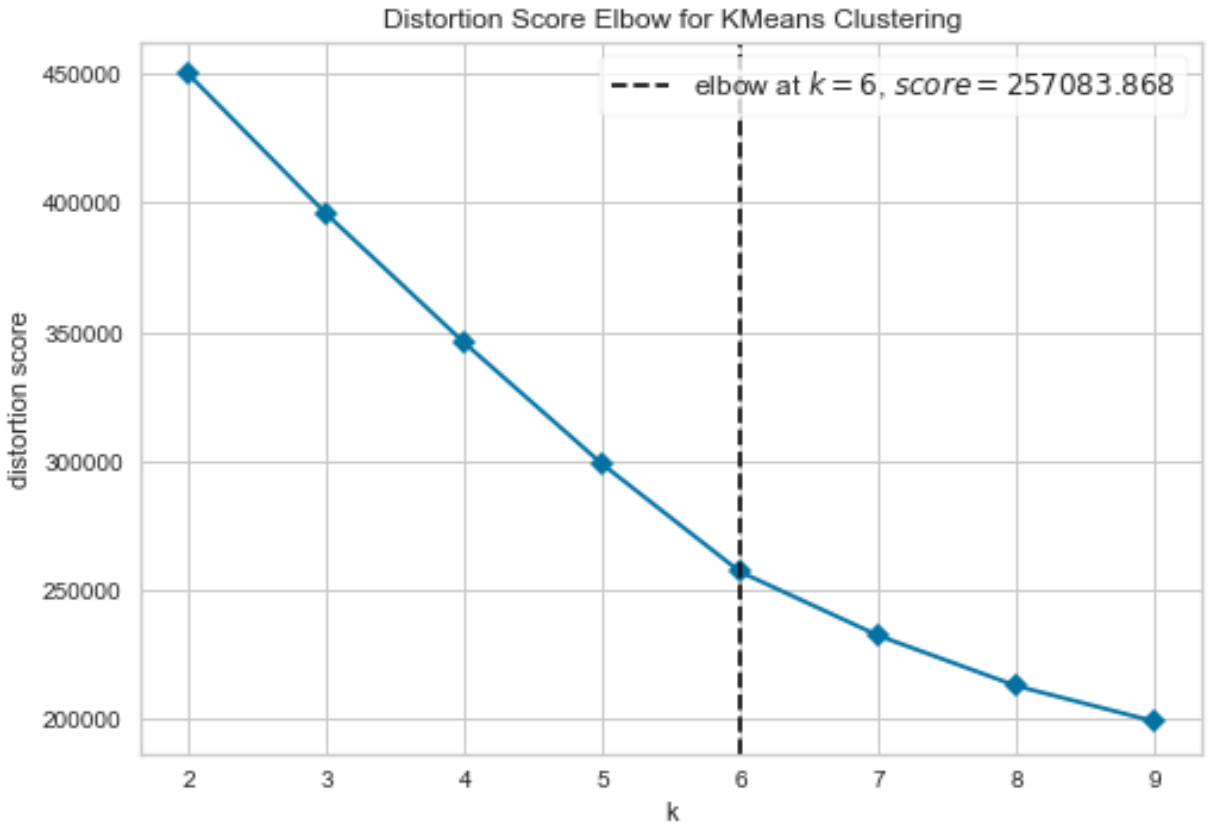
- Le 1er axe d'inertie peut être interprété dépendant du délai de livraison et de la distance client-vendeur.

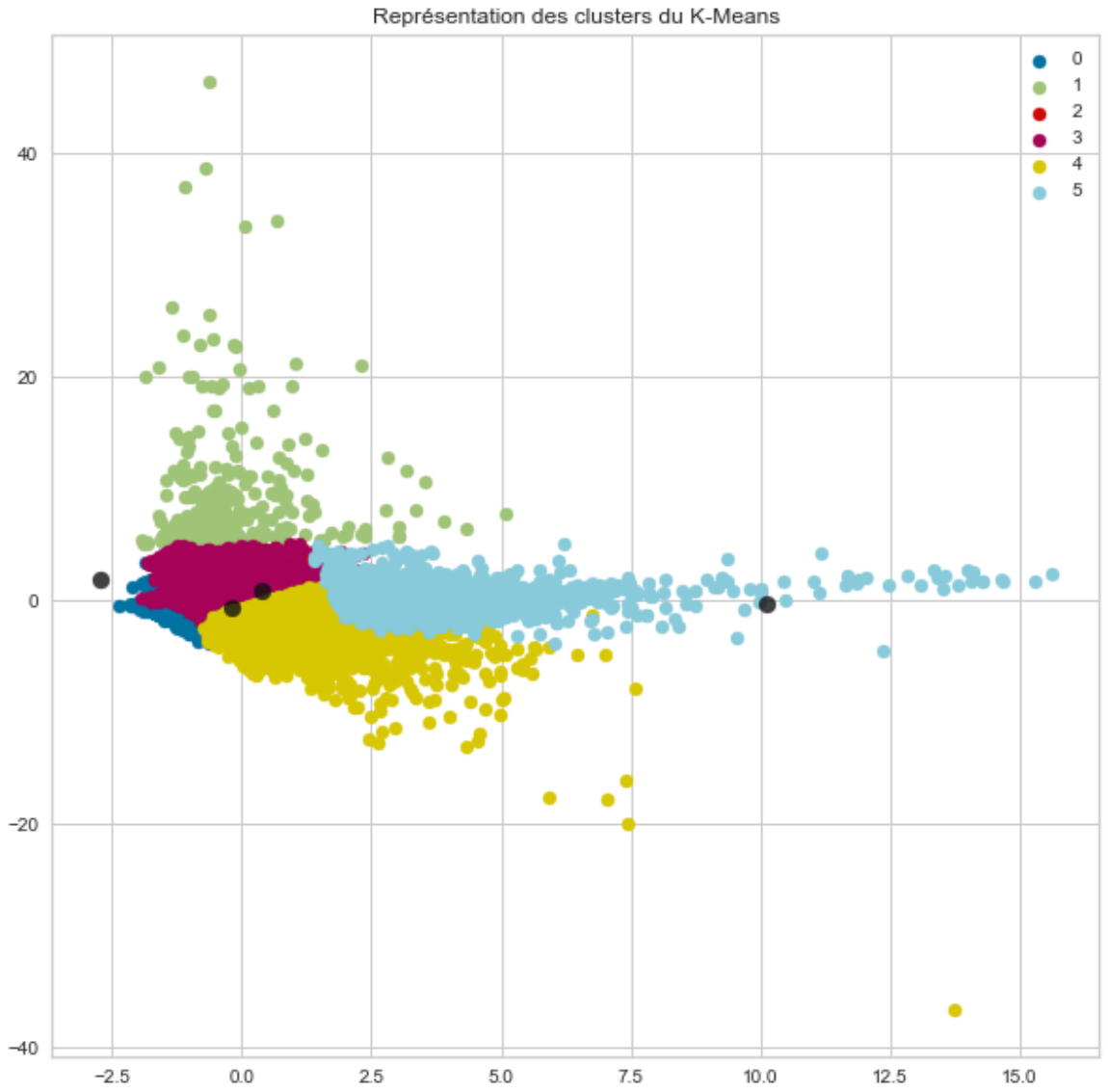
- Le 2nd axe d'inertie est lié à la fréquence d'achat principalement.

Le 1er axe d'inertie peut être interprété dépendant du délai de livraison et de la distance client-vendeur.
Le 2nd axe d'inertie est lié à la fréquence d'achat principalement.

3. SEGMENTATIONS EFFECTUEES

SEGMENTATION RFM + 3 VARIABLES QUANTITATIVES AVEC K-MEANS

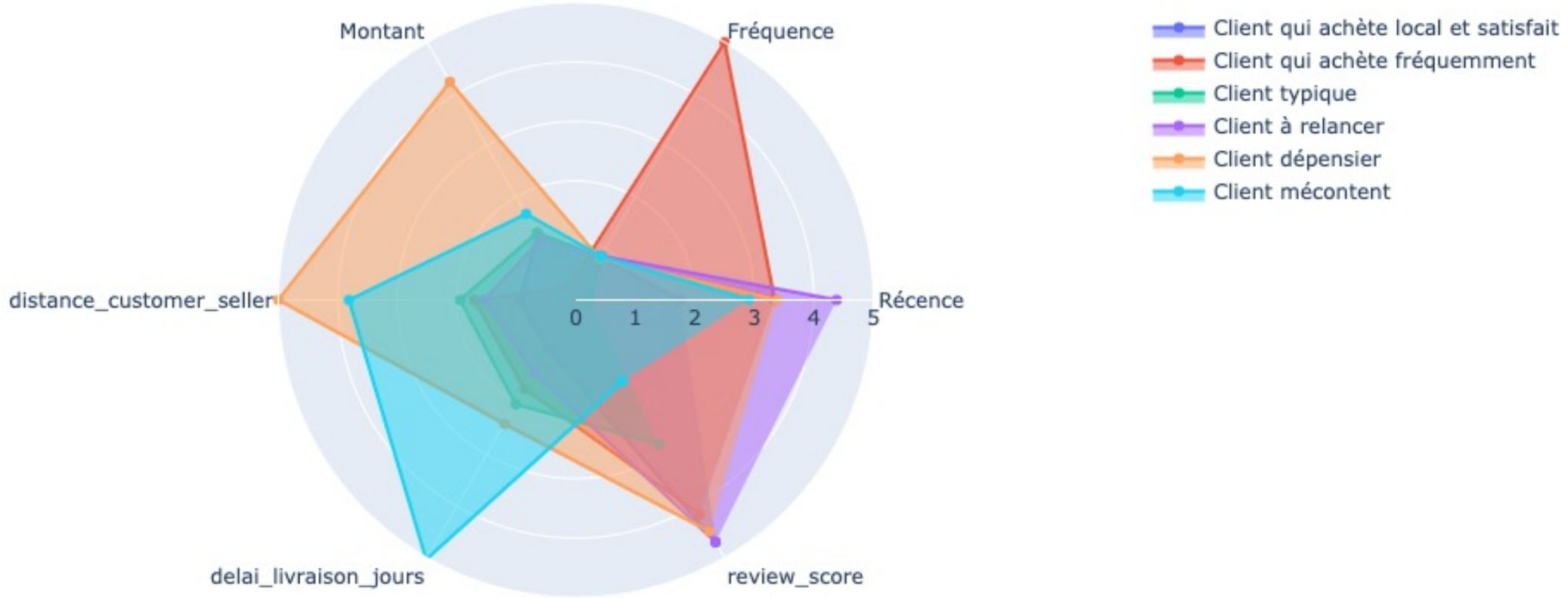




- On va segmenter la clientèle en 6 clusters d'après la méthode du coude.

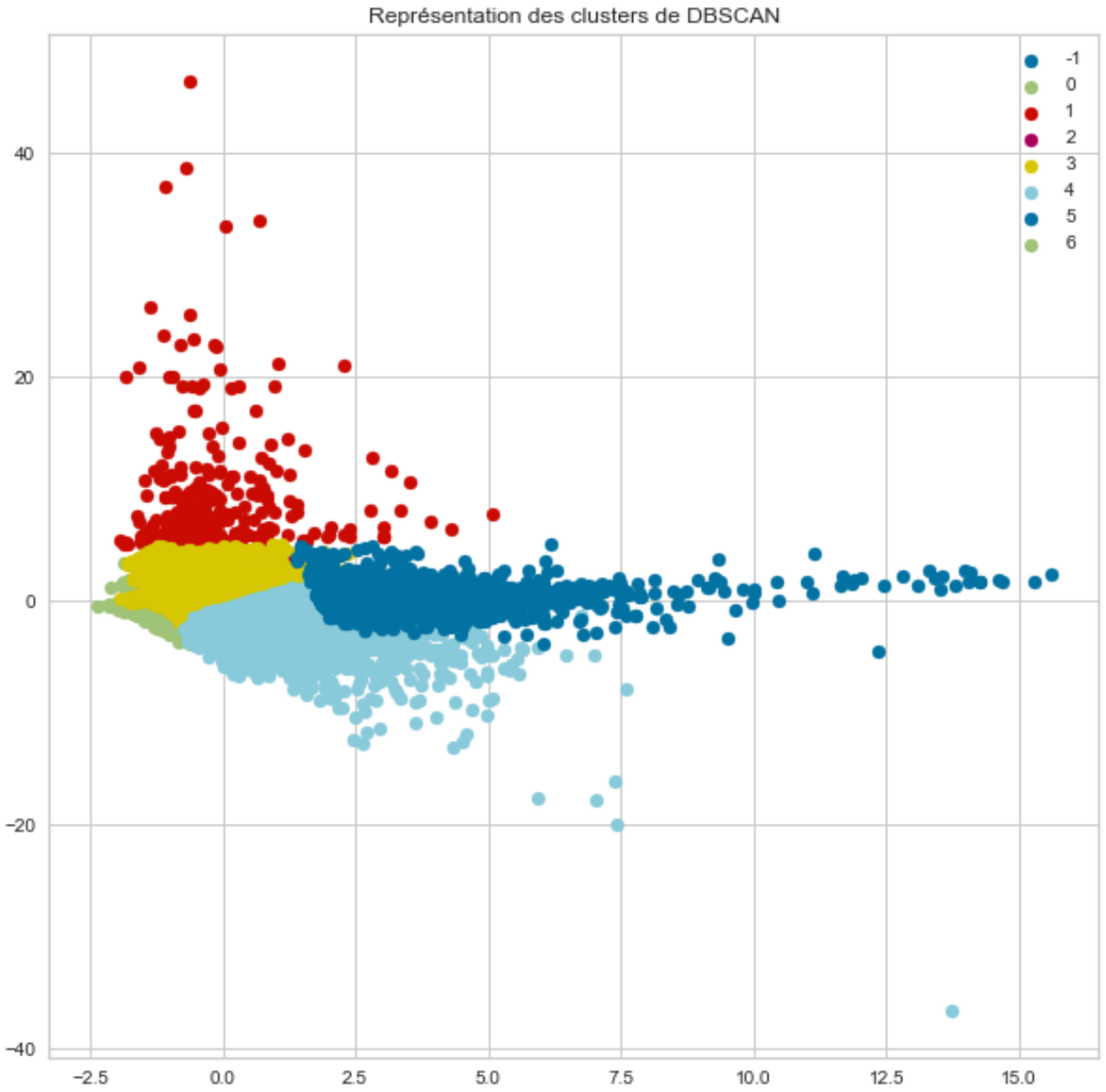
3. SEGMENTATIONS EFFECTUEES

SEGMENTATION RFM + 3 VARIABLES QUANTITATIVES



3. SEGMENTATIONS EFFECTUEES

SEGMENTATION + 3 VARIABLES QUANTITATIVES AVEC DBSCAN



- Hyper paramètres :
  - eps = 0,1
  - min_samples = 100

- Silhouette score = - 0,315

3. SEGMENTATIONS EFFECTUEES

SEGMENTATION RFM + 3 VARIABLES QUANTITATIVES AVEC DBSCAN

Type de modèle de Clustering	**DBSCAN** *eps = 0,5 min_samples = 100*	**DBSCAN** *eps = 0,5 min_samples = 50*	**DBSCAN** *eps = 0,1 min_samples = 100*	**DBSCAN** *eps = 0,2 min_samples = 100*	**DBSCAN** *eps = 0,35 min_samples = 100*
Nombre de variables quantitatives	*6*	*6*	*6*	*6*	*6*
Nombre de clusters	2	2	7	3	2
Silhouette	0.565	0.605	-0.315	0.250	0.455
Davies-Bouldin Index (DBI)	2.42	2.62	1.68	1.61	2.48

➢ Il n’est pas possible de trouver une segmentation qui conviennent.

PLAN

1. Présentation de la problématique, interprétation et pistes de recherche.

2. Nettoyage des données, feature engineering et exploration.

3. Présentation des segmentations effectuées.

4. Fréquence de mise à jour de la segmentation

4. FRÉQUENCE DE MISE À JOUR

STABILITE DES CLUSTERS DANS LE TEMPS : ALGORITHME

1. On part d'une base de référence B0 d'octobre 2016 à décembre 2017, on apprend dessus un clustering C0 (livré au client) → `C0_cluster.fit(B0_data)`

2. Maintenant, on créé les bases futures : B0 + un mois = B1, B0 + 2 mois = B2, etc.

3. Apprendre tous les clustering sur les bases artificielles B1, B2, etc → C1, C2, C3 (les clustering futurs) → `C1_cluster.fit(B1_data),` etc. pour tous les B de 1 à N

4. Segmenter B1, B2, B3 etc avec C0 → `C0_cluster.predict(B1_data)`, etc pour tous les B de 1 à N

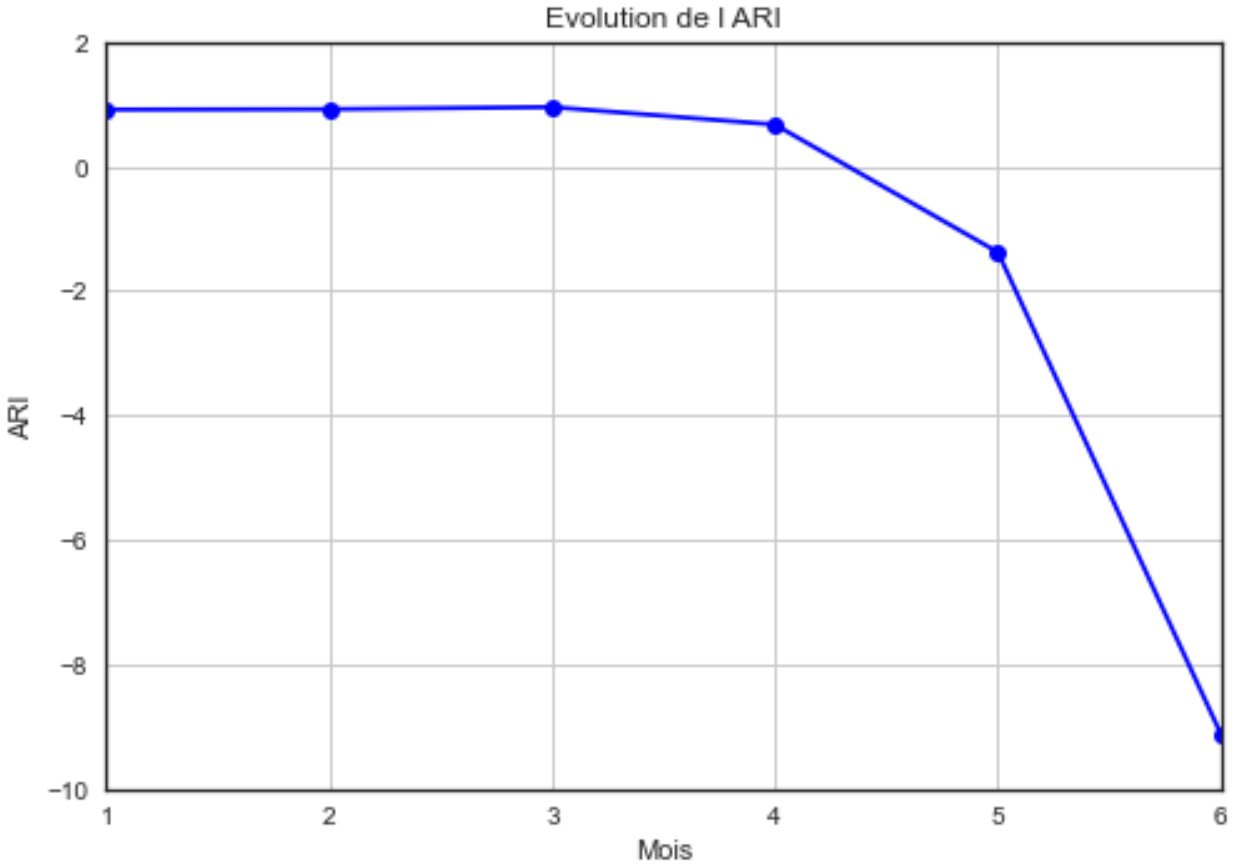
5. Segmenter B1, B2, B3 avec les clustering respectifs C1, C2 etc → `B1_by_C1 = C1_Cluster.predict(B1_data) =B1_by_C1, C2_Cluster.predict(B2_data)= B2_by_C2,` etc pour tous les B et C de 1 à N

6. Comparer la segmentation de B1 B2 B3 etc entre C0 (livré au client) et les autres (C1 etc) avec l'indice de Rand Ajusté `adjusted_rand_score(B1_by_C0, B1_by_C1), adjusted_rand_score(B2_by_C0, B2_by_C2),` etc pour tous les B de 1 à N

7. Représenter l'évolution de l'indice de rand ajusté.

4. FRÉQUENCE DE MISE À JOUR

STABILITE DES CLUSTERS DANS LE TEMPS : CALCUL DE L'ARI



ARI – prédiction Janvier 2018 à Juin 2018					
Janvier	Février	Mars	Avril	Mai	Juin
0.914	0.920	0.954	0.673	-1.376	-9.11

- On calcule l'ARI en comparant les clusters modélisés (« true labels ») sur 6 mois de janvier 2018 à juin 2018 par rapport à la prédiction d'un modèle de clustering basé uniquement sur les données de 2016-2017 (« predicted labels »)
- On observe que l'ARI se dégrade rapidement après 3 mois donc propose donc de faire **une mise à jour trimestrielle du clustering sur la base du modèle KMEANS avec segmentation RFM + 3 variables quantitatives**

CONCLUSION

CONCLUSION
RECAPITULATIF

Type de modèle de Clustering	K-Means *Segmentation RFM*	K-Means *Segmentation RFM + ajout de 3 var. quantitatives*
Nombre de variables quantitatives	*3*	*6*
Nombre de clusters	4	6
Silhouette	0.456	0.315
Davies-Bouldin Index (DBI)	0.724	0.984

- On retient le modèle K-Means avec segmentation RFM + 3 variables quantitatives qui devra être mise à jour tous les 3 mois.

CONCLUSION

SEGMENTATION RFM + 3 VARIABLES QUANTITATIVES

PERSONAE

Achète local & satisfait

- 1 achat
- 126 Real
- **300km**
- **4.7/5**

Achète compulsivement

- **6 achats**
- 25 Real
- 547km
- 4.16/5

Client « typique »

- 1 achat
- 130 Real
- 623km
- 2.80/5

Client à relancer

- Pas d'achat depuis plus d'1 an

Client dépensier

- 1 achat
- **423 Real**
- 1600km
- 4.49/5

Client mécontent

- 1 achat
- 291 Real
- 1224km
- **1.6/5**