OPENCLASSROOMS

Projet n°4

Segmentez des clients d'un site e-commerce : Olist

Fayz El Razaz

Parcours Ingénieur machine learning

Plan de la présentation

I - Introduction et présentation des tables de données

II - Analyse exploratoire, cleaning & préparation des données

III - Analyse non supervisée

IV - Contrat de maintenance

V - Synthèse

I - Introduction

Travail effectué pour Olist :

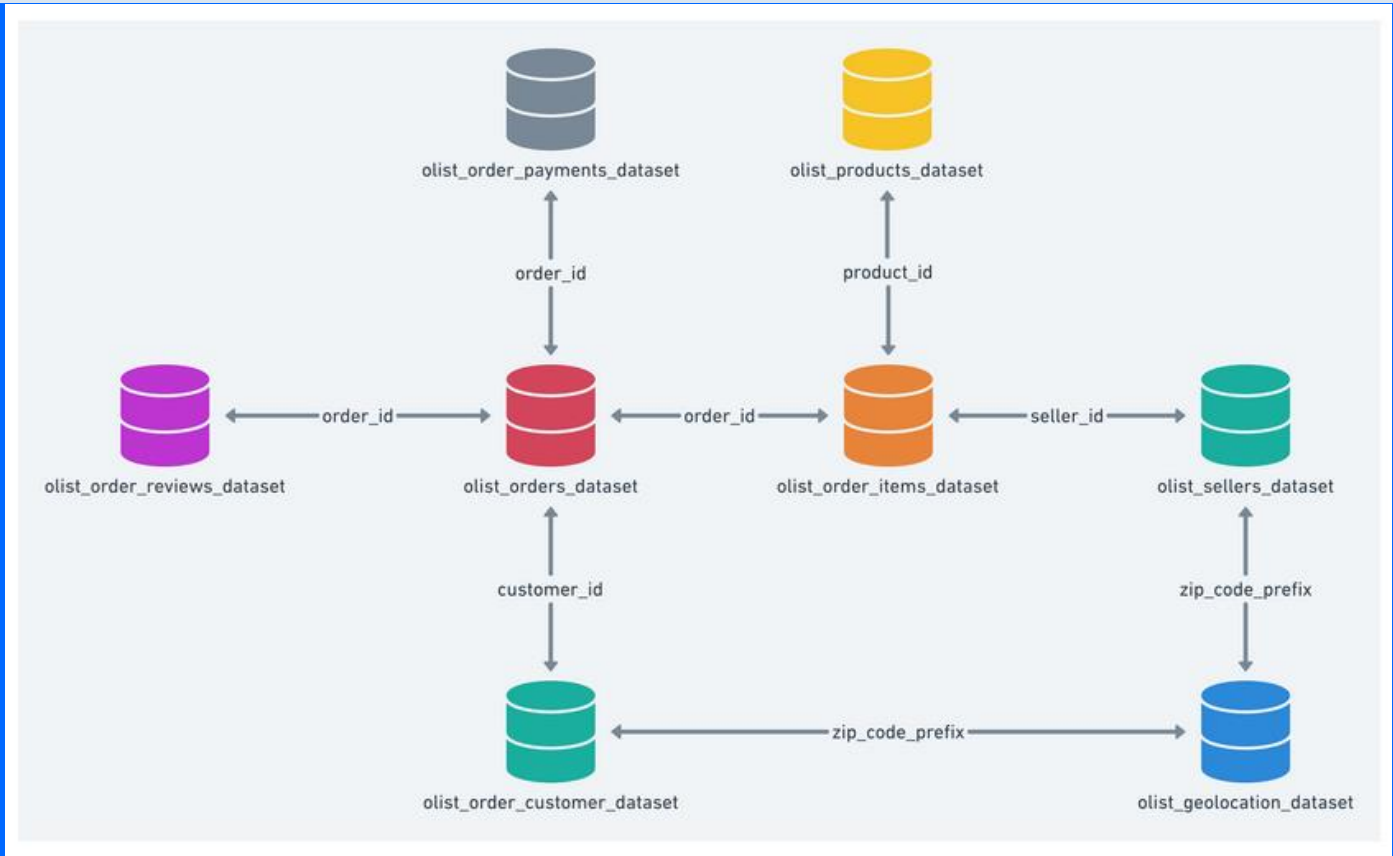
- Entreprise brésilienne
- Solution de vente sur les marketplaces

Souhait de l'entreprise :

- Fournir à ses équipes d'e-commerce une segmentation des clients.
- Fournir à l'équipe marketing un contrat de maintenance pour une utilisation optimale de la segmentation proposée

I – Ensembles de tables

Pour ce faire :



I - Tables

```
1 customers.head()
```

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
0	06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
1	18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	9790	sao bernardo do campo	SP
2	4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2b5e	1151	sao paulo	SP
3	b2b6027bc5c5109e529d4dc6358b12c3	259dac757896d24d7702b9acbbff3f3c	8775	mogi das cruzes	SP
4	4f2d8ab171c80ec8364f7c12e35b23ad	345ecd01c38d18a9036ed96c73b8d066	13056	campinas	SP

```
1 payments.head()
```

	order_id	payment_sequential	payment_type	payment_installments	payment_value
0	b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	8	99.33
1	a9810da82917af2d9aefd1278f1dcfa0	1	credit_card	1	24.39
2	25e8ea4e93396b6fa0d3dd708e76c1bd	1	credit_card	1	65.71
3	ba78997921bbcdc1373bb41e913ab953	1	credit_card	8	107.78
4	42fdf880ba16b47b59251dd489d4441a	1	credit_card	2	128.45

Données structurées

I - Tables

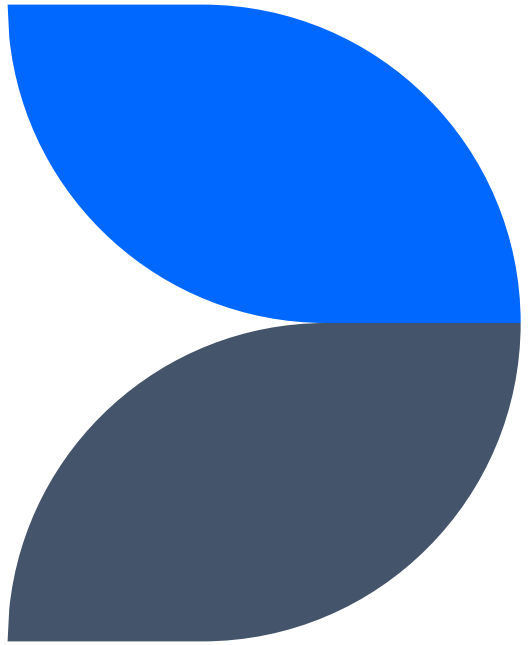
	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state	order_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date
0	06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP	00e7ee1b050b8499577073aeb2a297a1	delivered	2017-05-16 15:05:35	2017-05-16 15:22:12	2017-05-23 10:47:57
1	8912fc0c3bbf1e2fbf35819e21706718	9eae34bbd3a474ec5d07949ca7de67c0	68030	santarem	PA	c1d2b34febe9cd269e378117d6681172	delivered	2017-11-09 00:50:13	2017-11-10 00:47:48	2017-11-22 01:43:37
2	8912fc0c3bbf1e2fbf35819e21706718	9eae34bbd3a474ec5d07949ca7de67c0	68030	santarem	PA	c1d2b34febe9cd269e378117d6681172	delivered	2017-11-09 00:50:13	2017-11-10 00:47:48	2017-11-22 01:43:37
3	f0ac8e5a239118859b1734e1087cbb1f	3c799d181c34d51f6d44bbbc563024db	92480	nova santa rita	RS	b1a5d5365d330d10485e0203d54ab9e8	delivered	2017-05-07 20:11:26	2017-05-08 22:22:56	2017-05-19 20:16:31
4	6bc8d08963a135220ed6c6d098831f84	23397e992b09769faf5e66f9e171a241	25931	mage	RJ	2e604b3614664aa66867856dba7e61b7	delivered	2018-02-03 19:45:40	2018-02-04 22:29:19	2018-02-19 18:21:47

Dimensions :
112372 lignes
32 colonnes

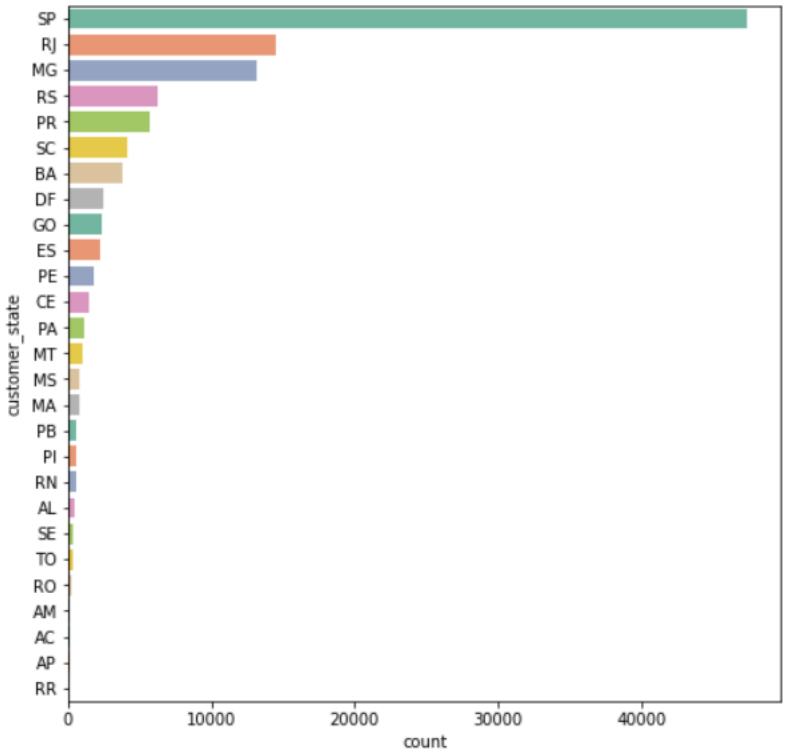
OPENCLASSROOMS

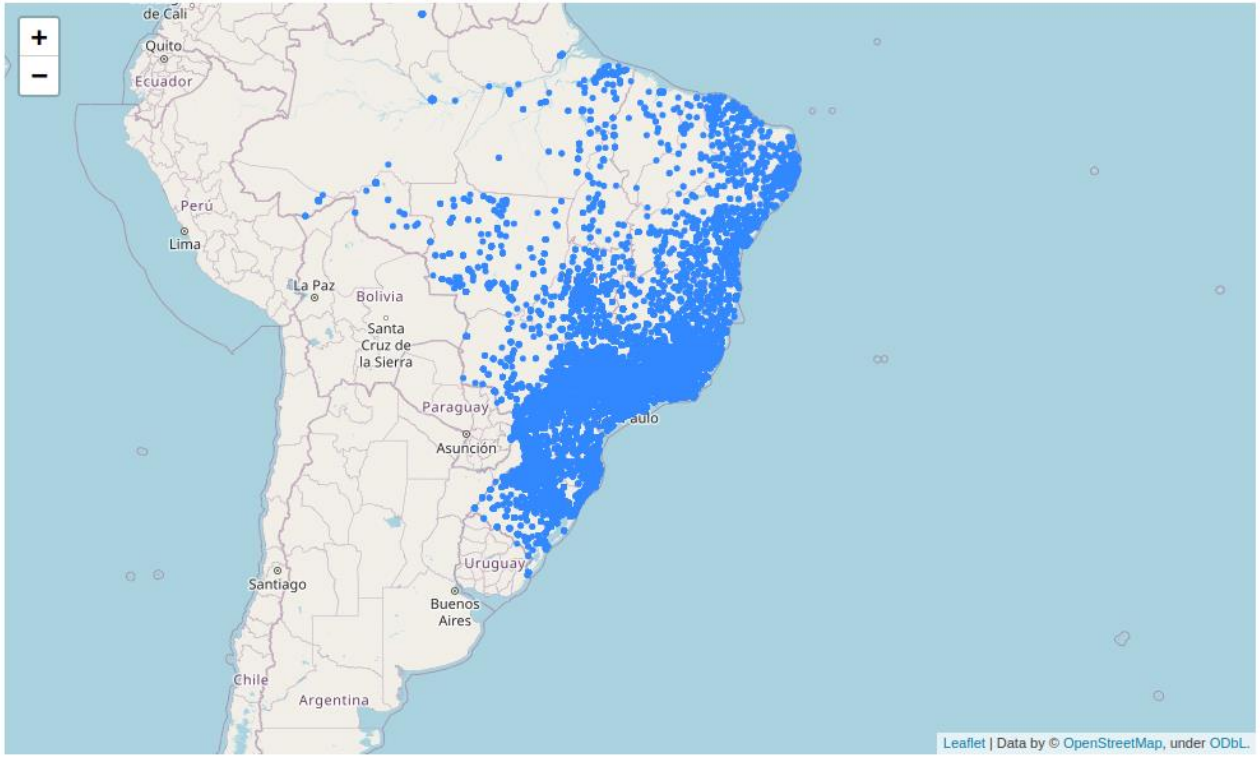
II – Analyse exploratoire, cleaning & préparation des données

Analyse univariée
Feature engineering
Analyse bivariée

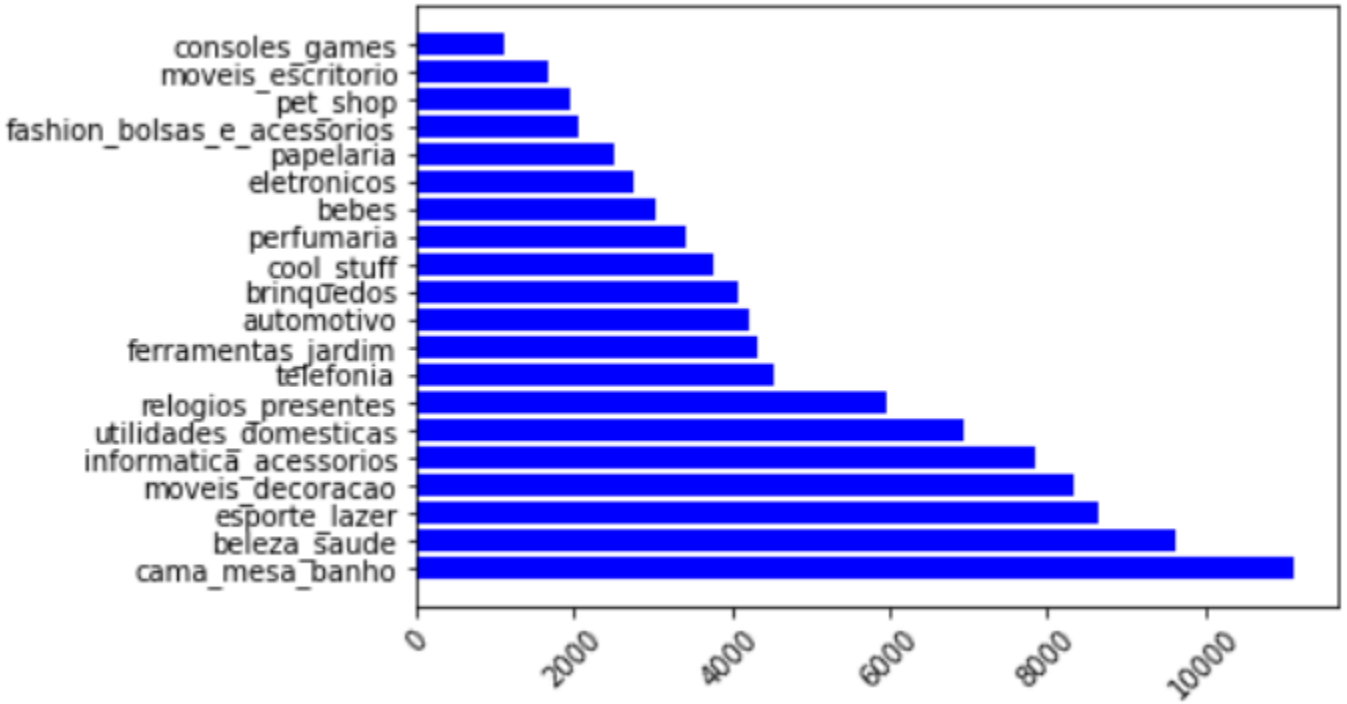


Répartition des clients par région





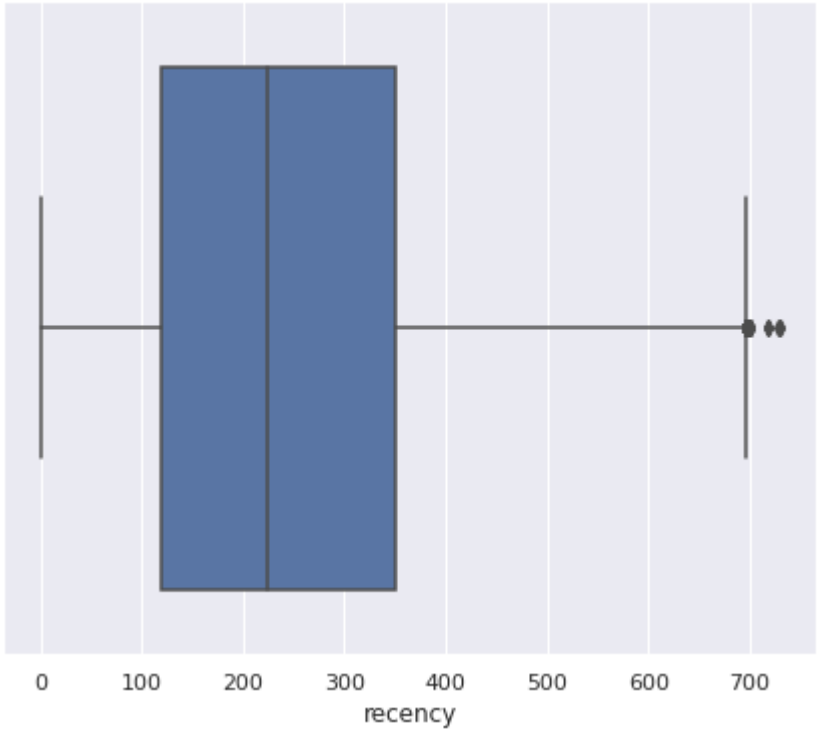
Répartition des clients par type de produits

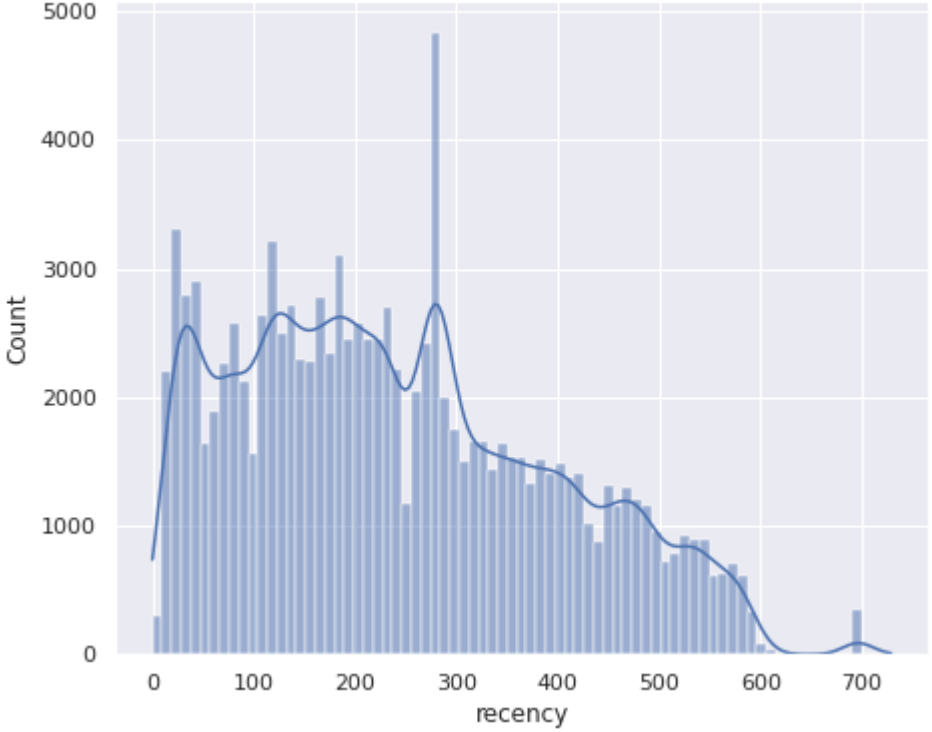


Création des variables RFM

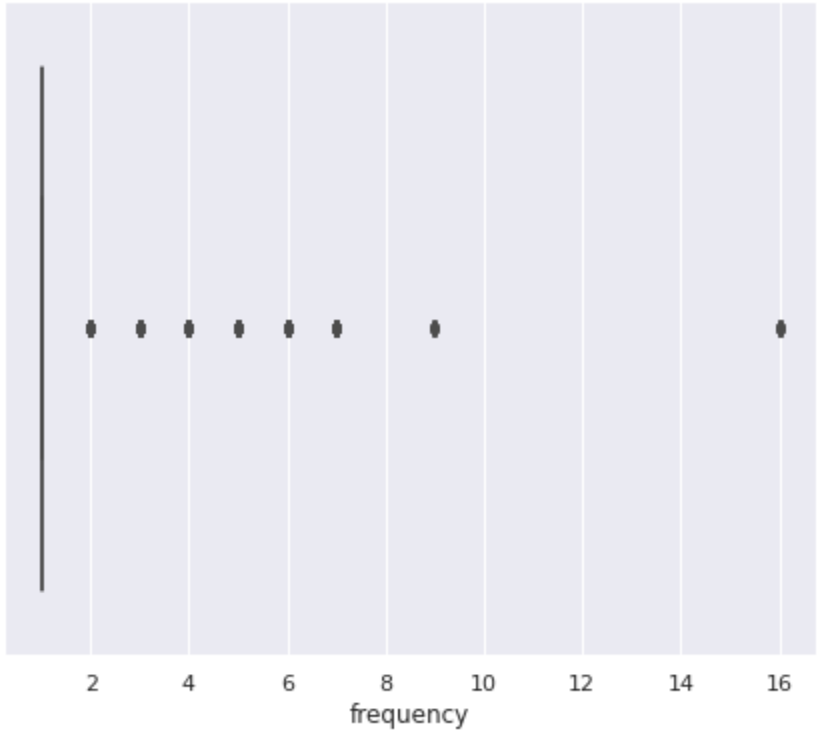
R : Récence (en jours)
F : Fréquence (nombre de retour sur le site)
M : Montant (en real brésilien)

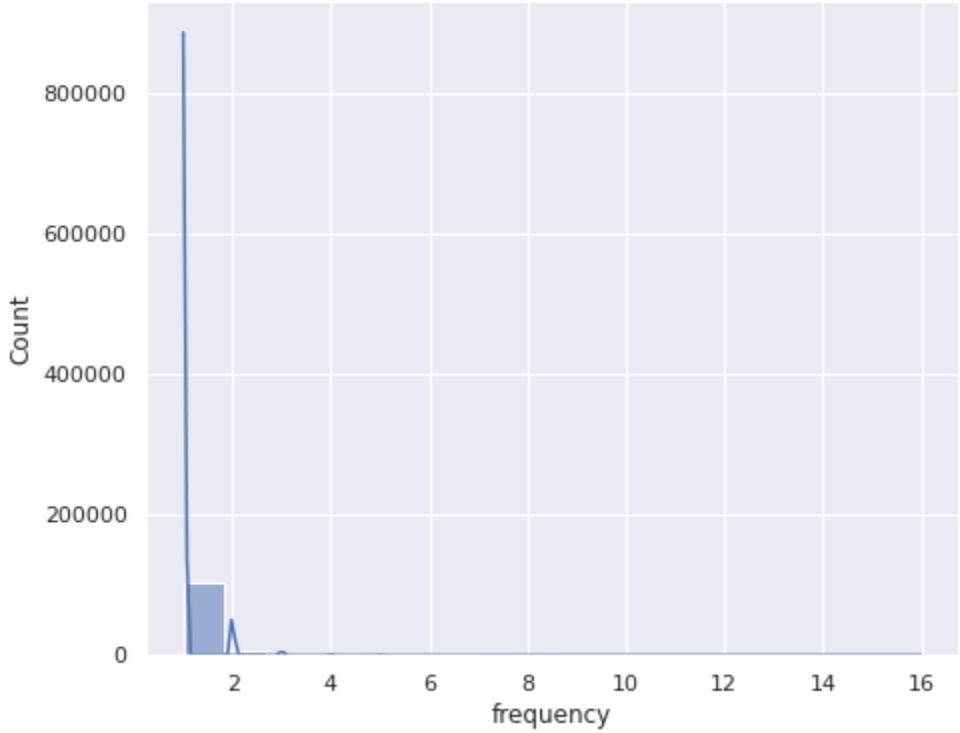
Création des variables RFM



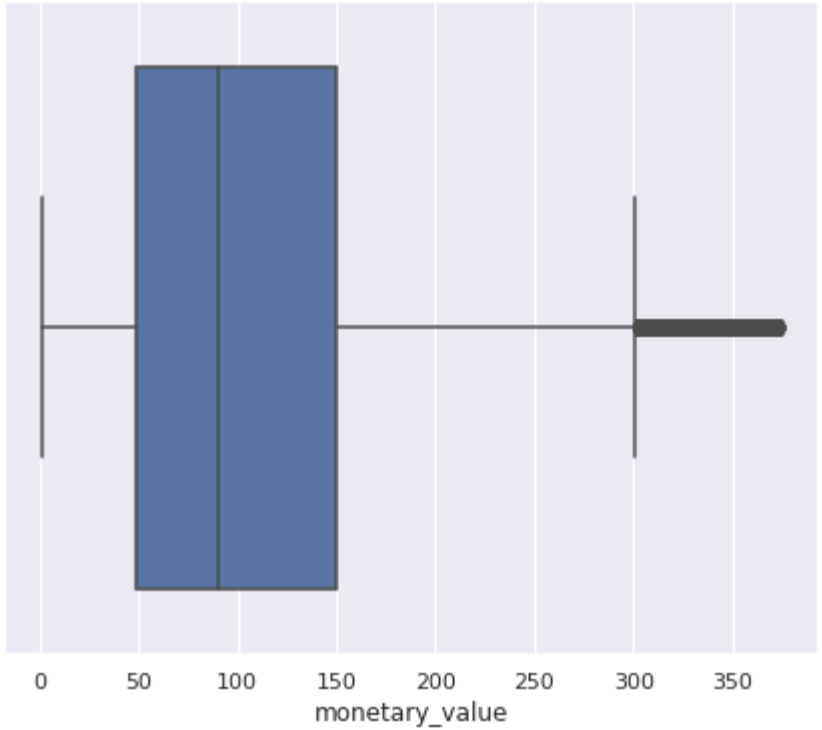


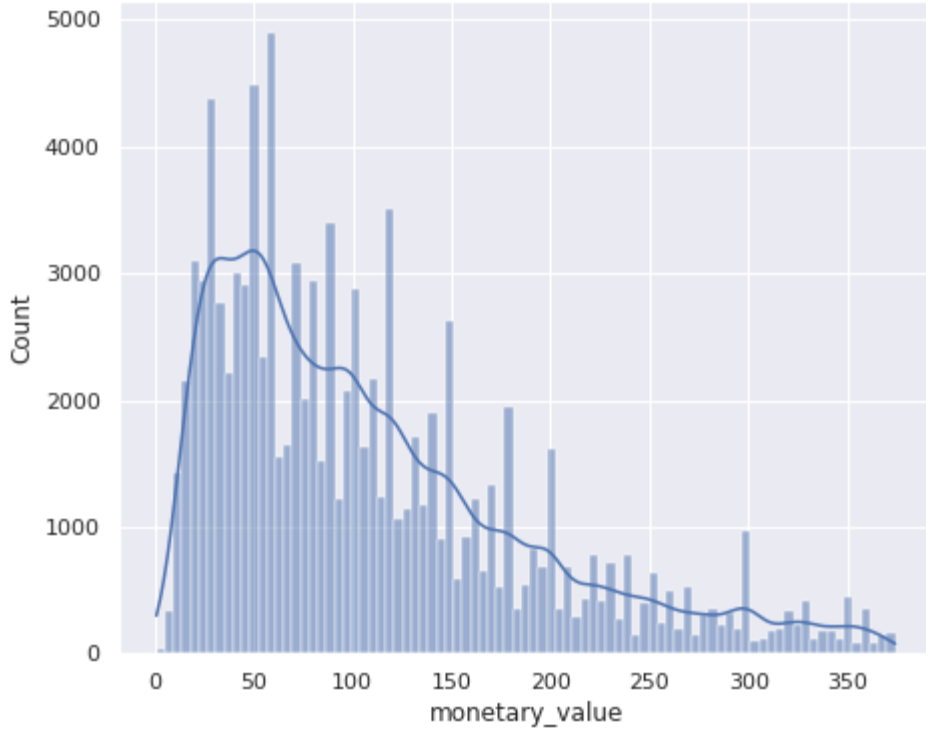
Création des variables RFM



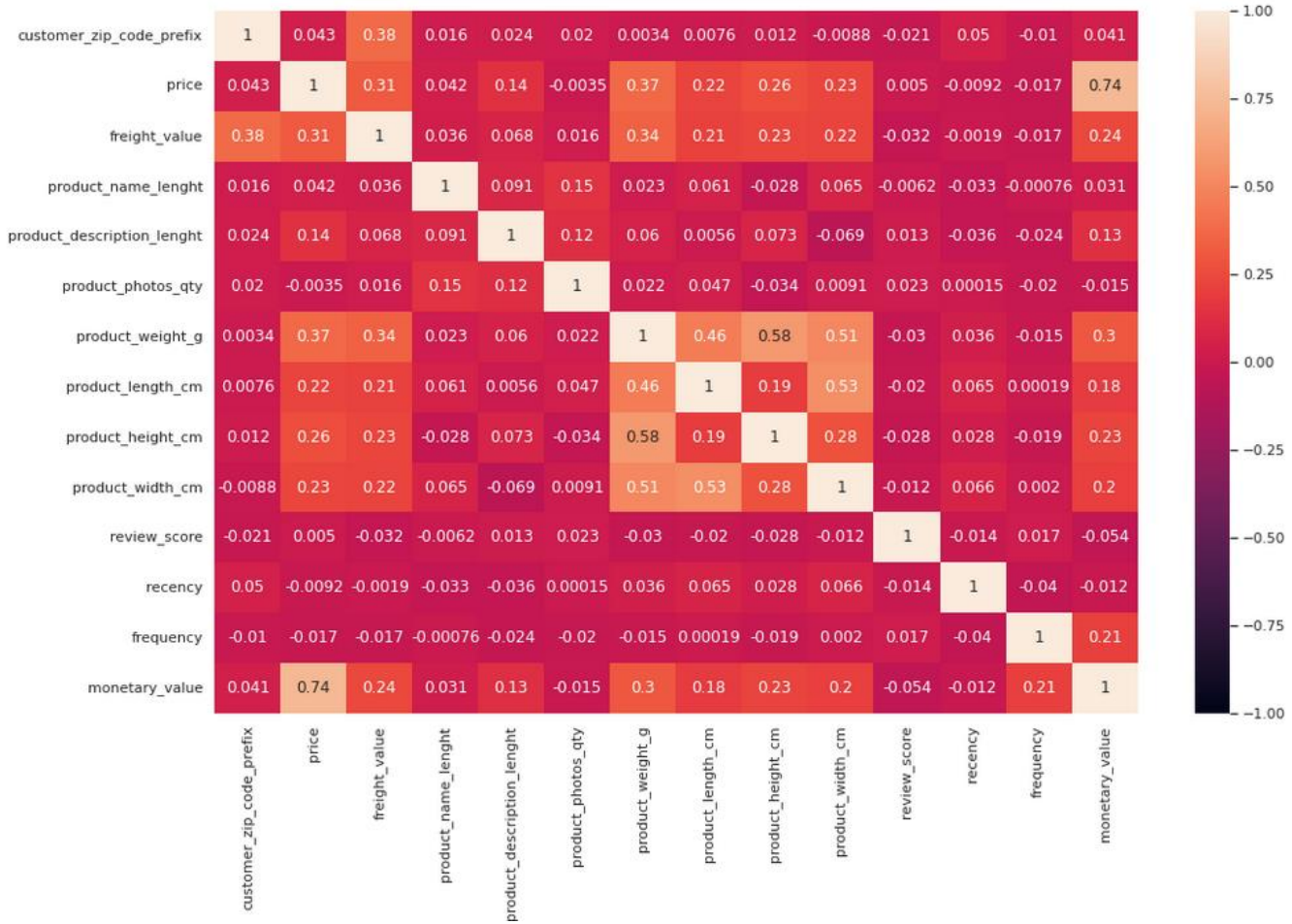


Création des variables RFM





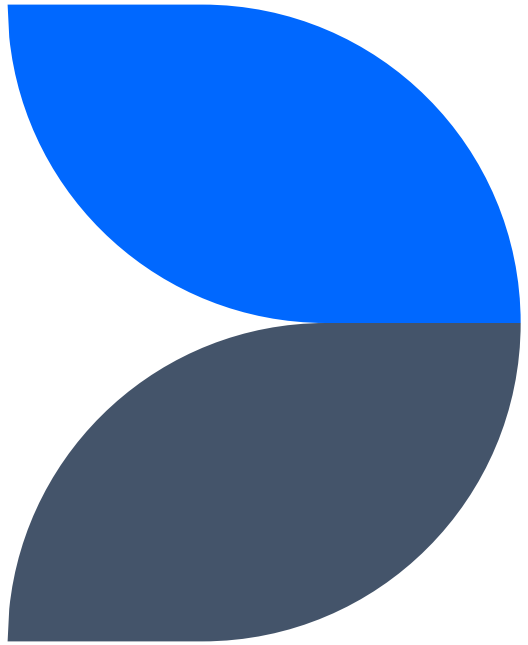
Analyse des corrélations



OPENCLASSROOMS

III – Analyse non supervisée

Rappels sur le K-means
K-means avec les variables RFM
K-means avec ajout de variables
K-means utilisant l'ensemble des données
DBSCAN
T-SNE sur ACP
Agglomerative clustering



Rappel sur le K-means

Objectif : Partitionnement de données

Méthode : Regrouper au sein de k groupes, les n données du dataset étudié

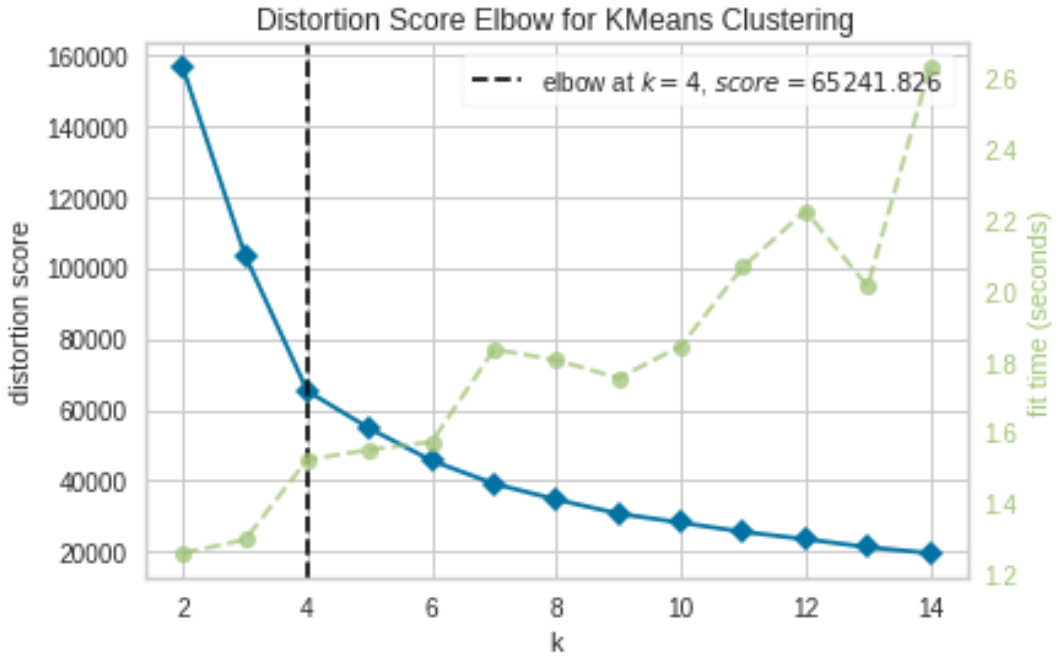
Difficulté : Obtention du nombre k optimal qui permettra :

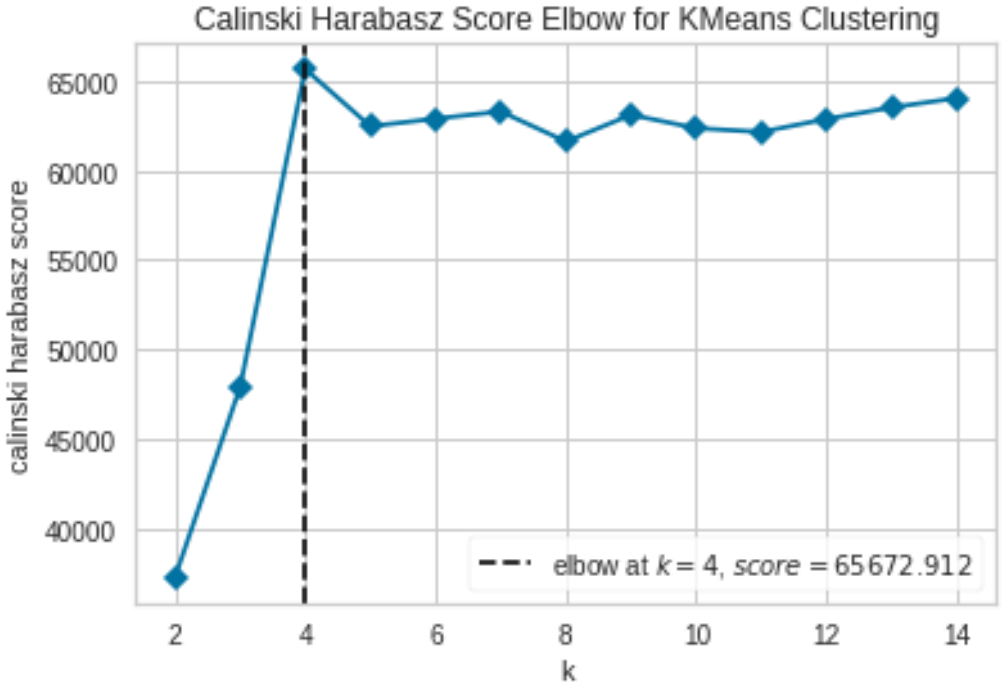
- la plus grande homogénéité au sein des k groupes
- la plus grande hétérogénéité entre les k groupes

Obtention des meilleurs centroïdes

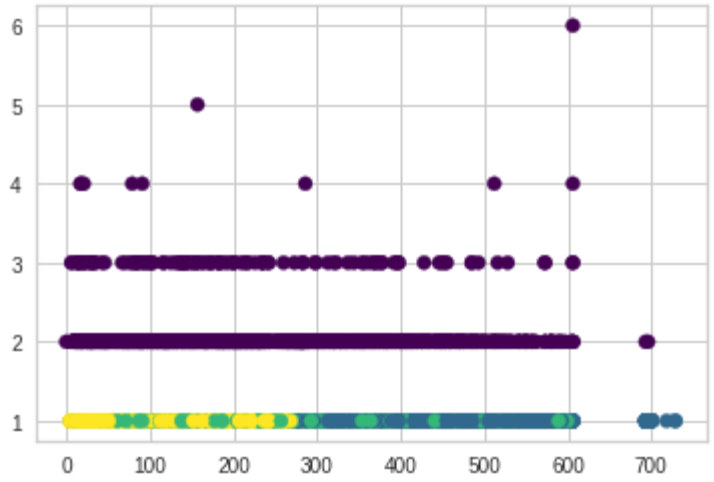
K-means sur les données RFM

Recherche de la meilleur quantité de clusters

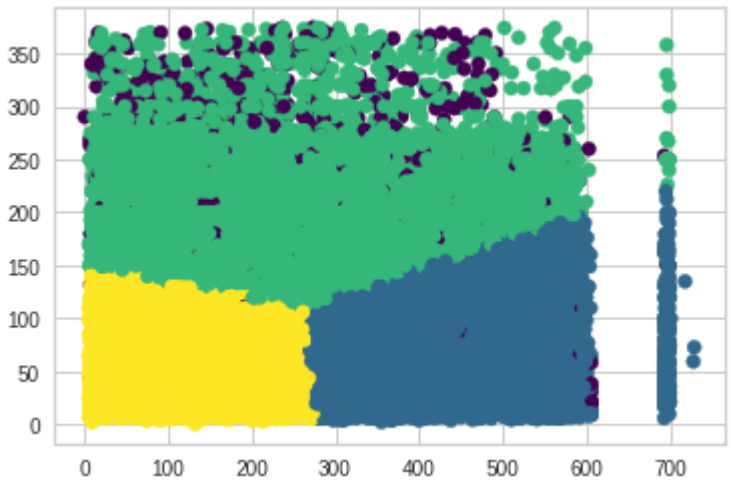




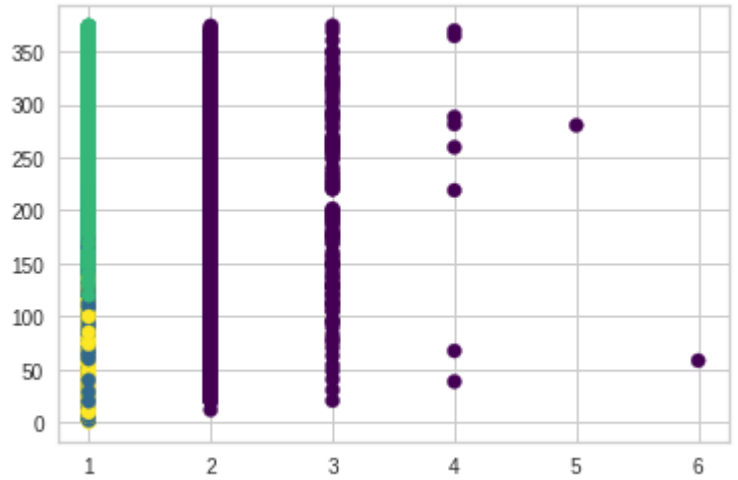
Valeur optimale : k = 4



Récence - Fréquence

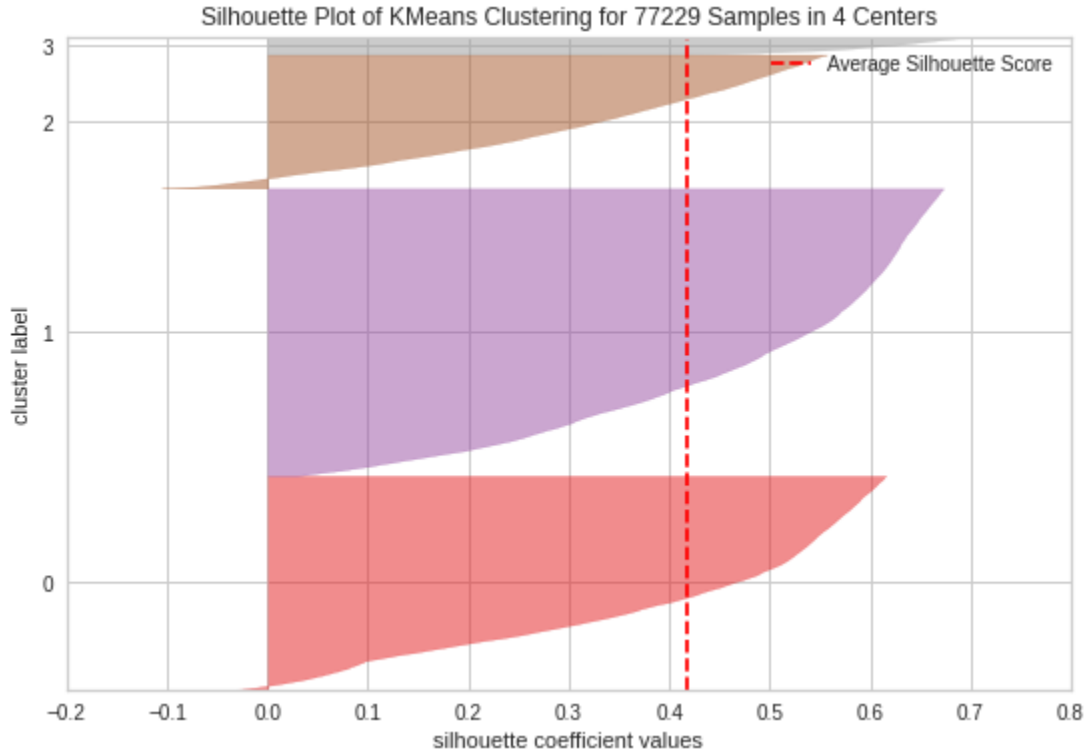


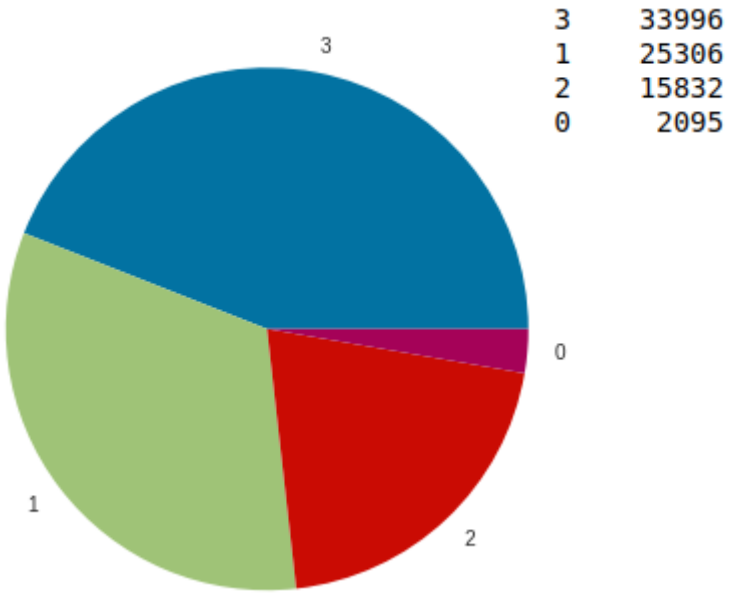
Monnaie - Récence



Fréquence - Monnaie

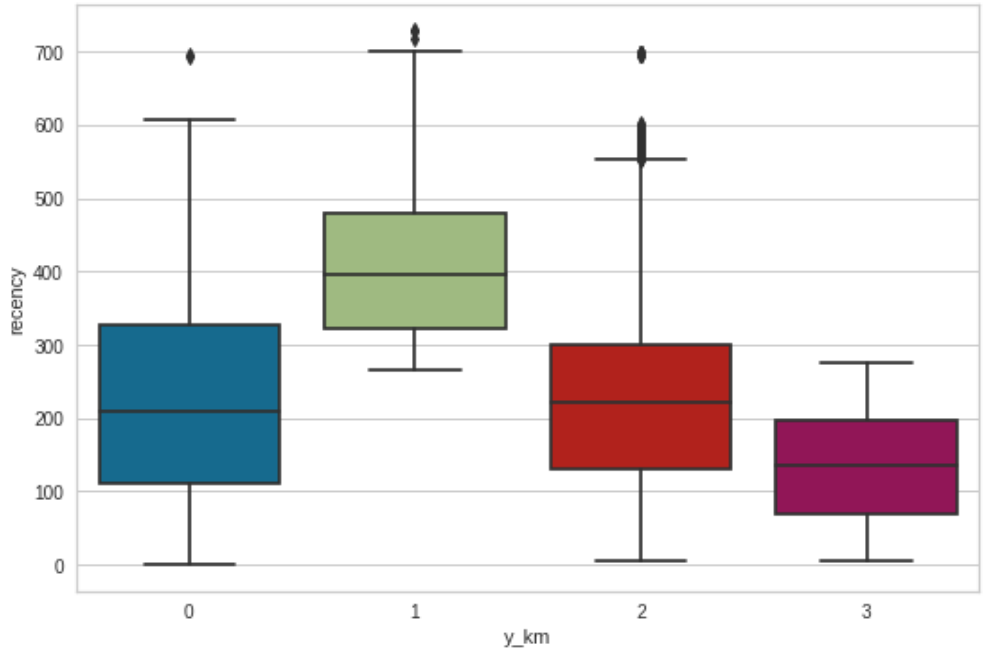
Valeur optimale : k = 4

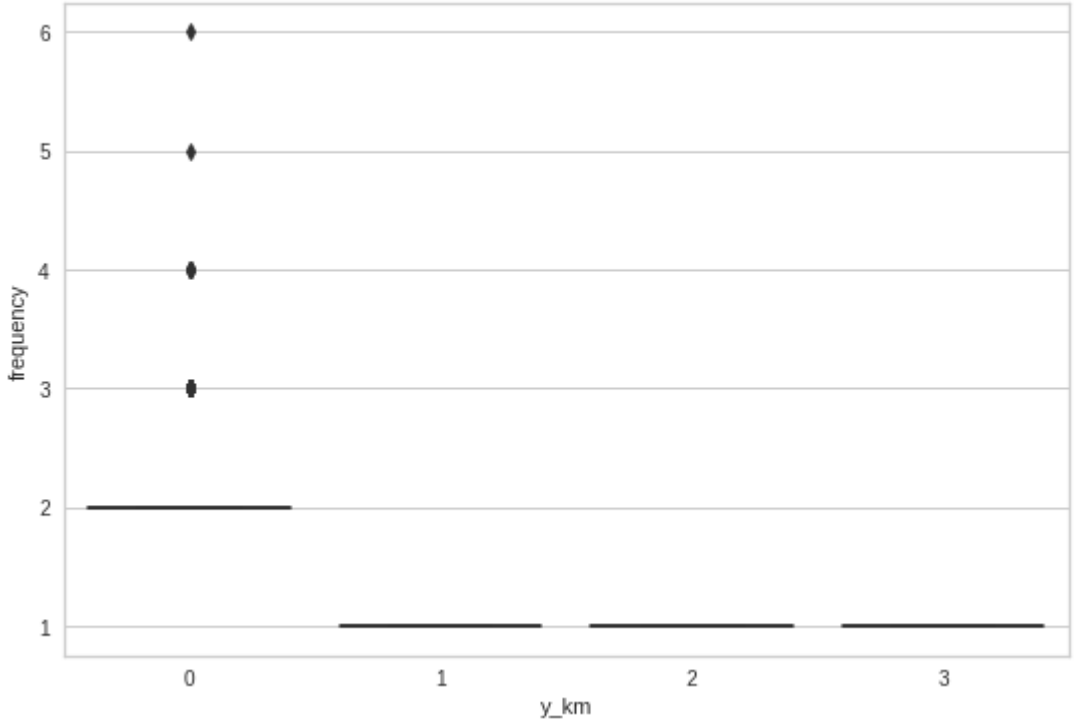




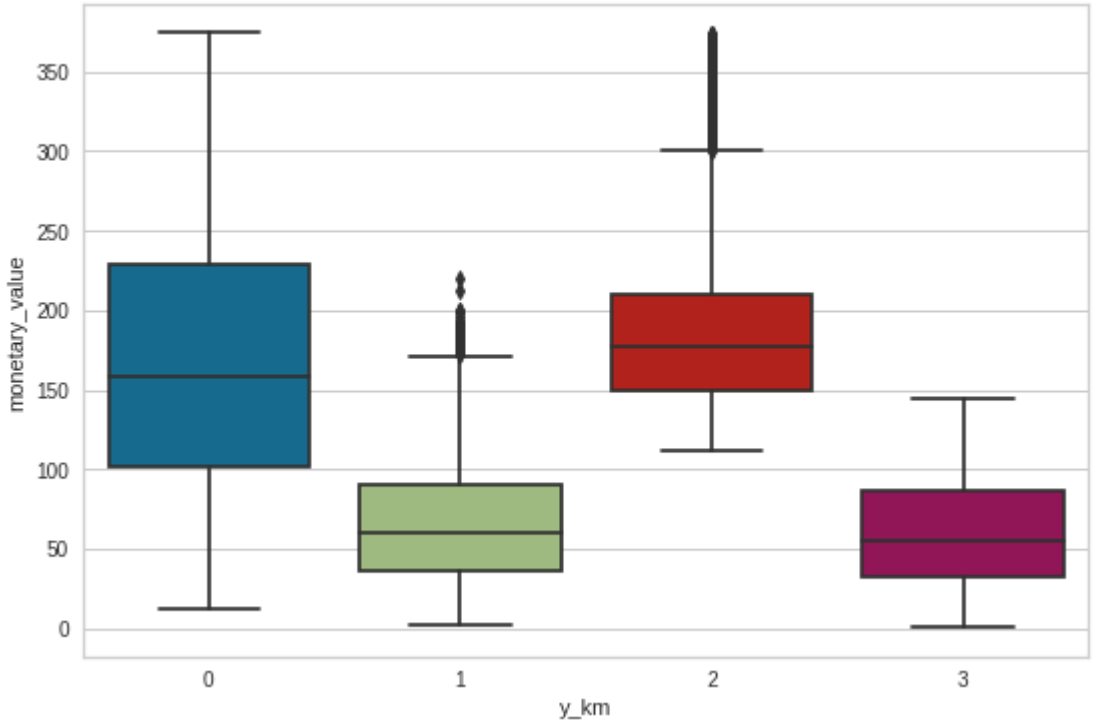
3	33996
1	25306
2	15832
0	2095

Valeur optimale : k = 4



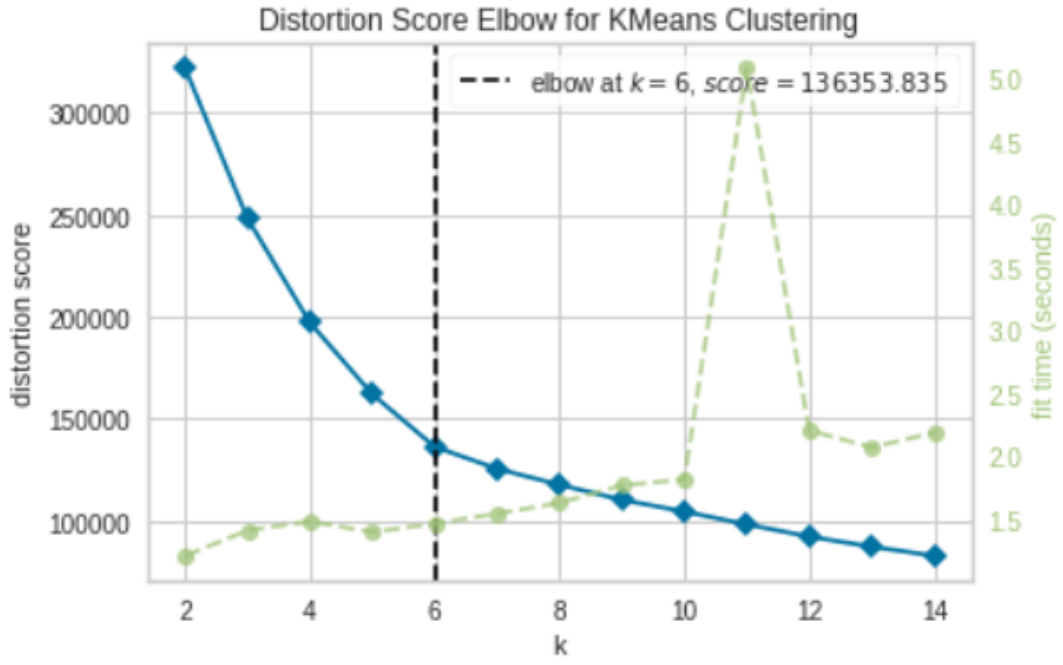


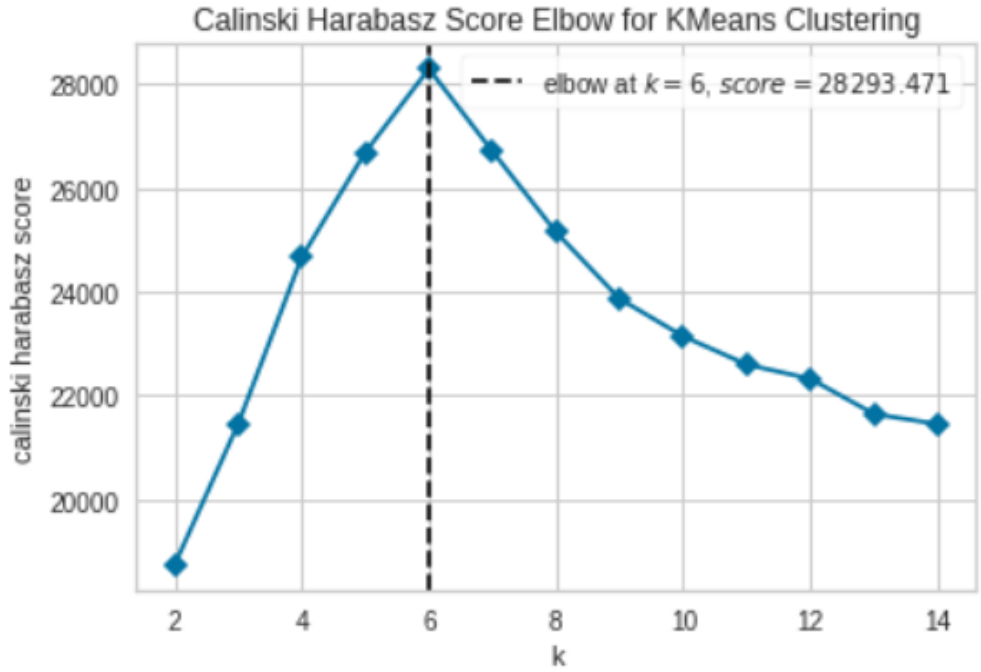
Valeur optimale : k = 4



K-means sur RFM étendu

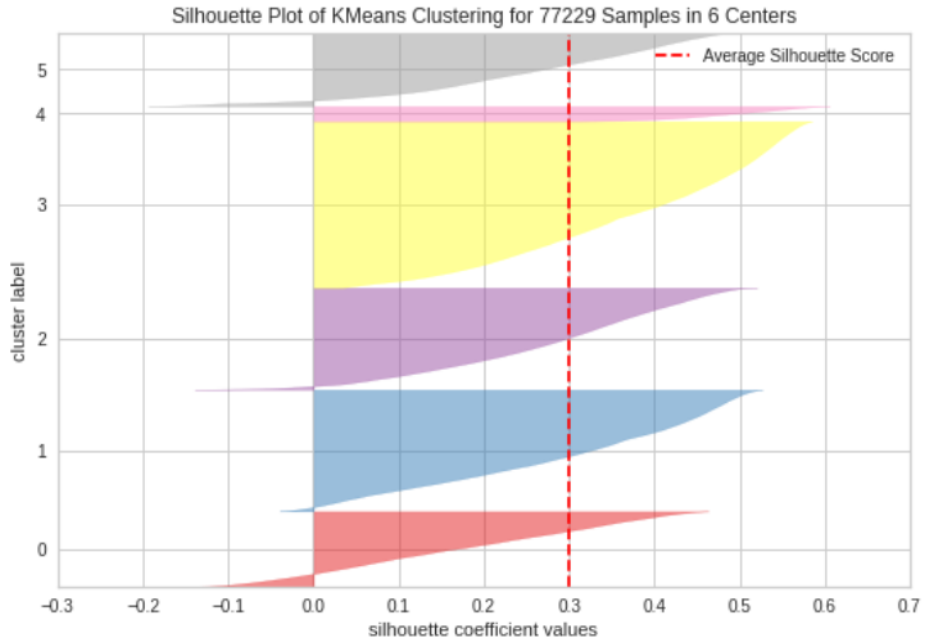
Ajout de la note et de la localisation du client – k = 6

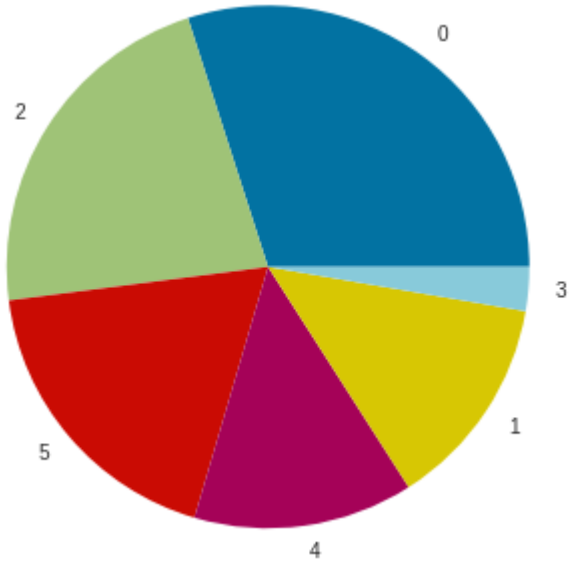




K-means sur RFM étendu

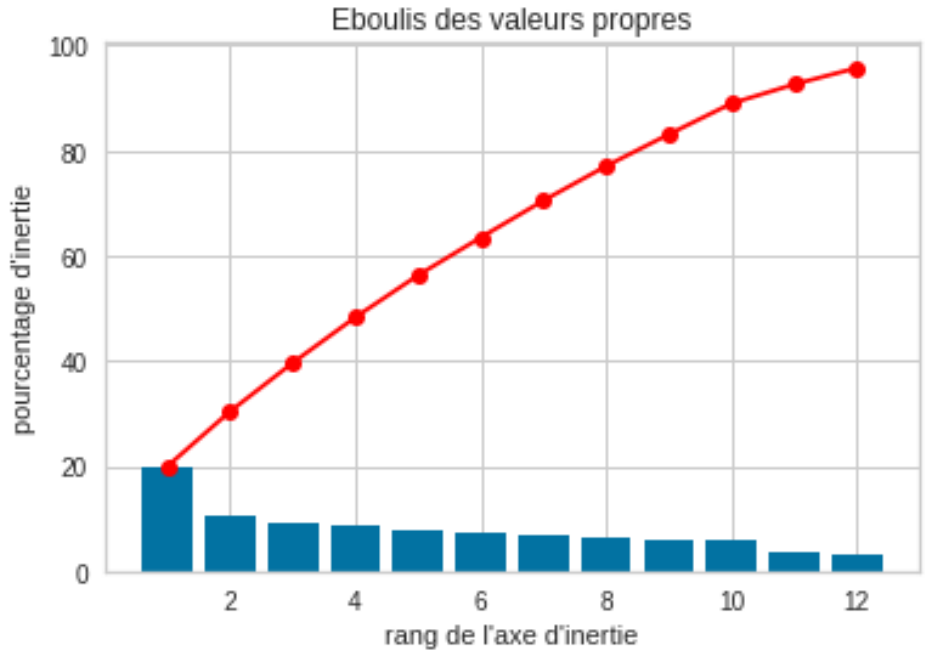
Ajout de la note et de la localisation du client – k = 6

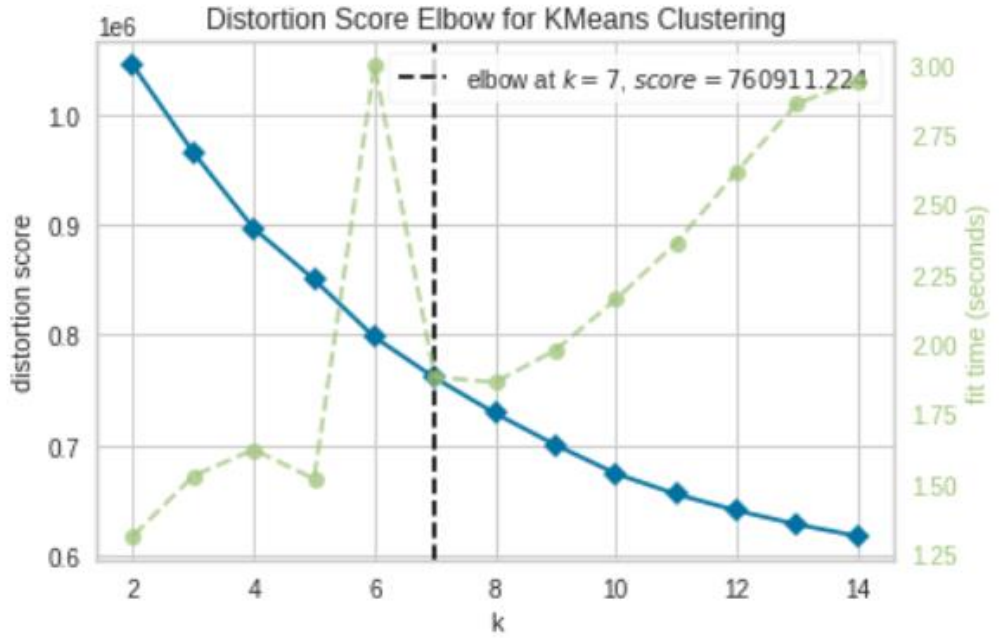




K-means sur tout le dataset

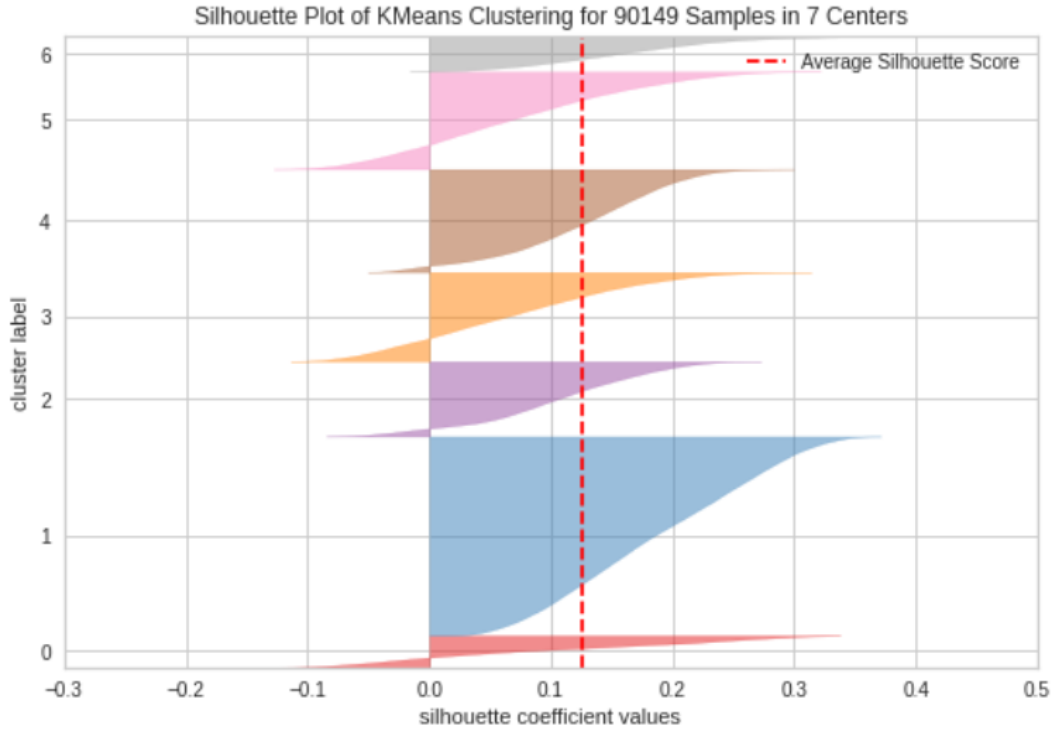
Analyse en composantes principales





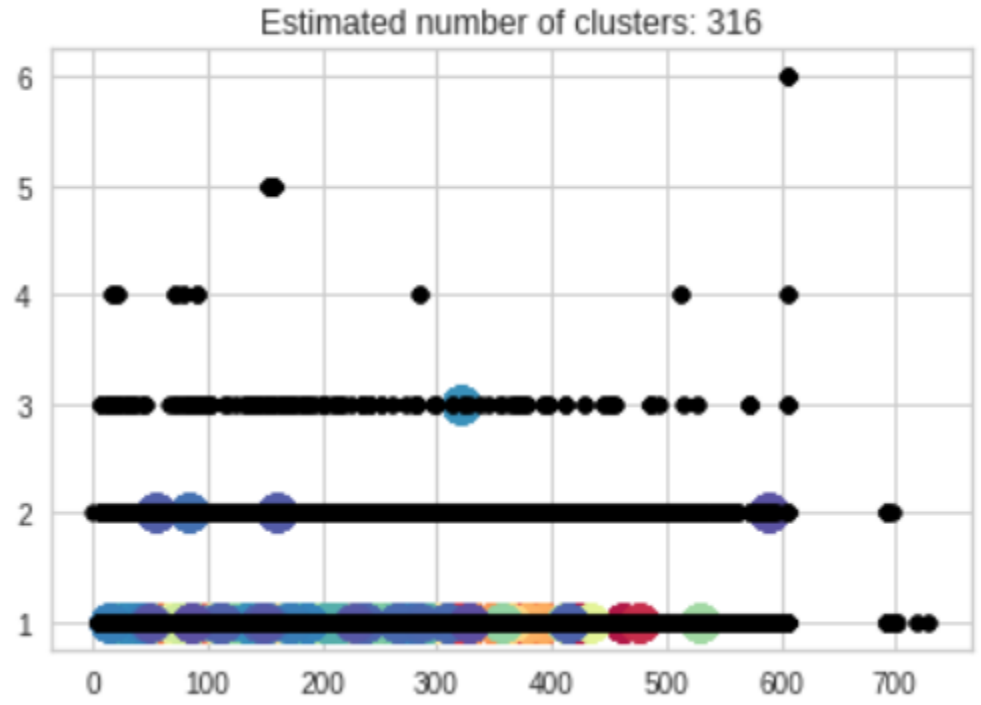
K-means sur tout le dataset

Analyse en composantes principales

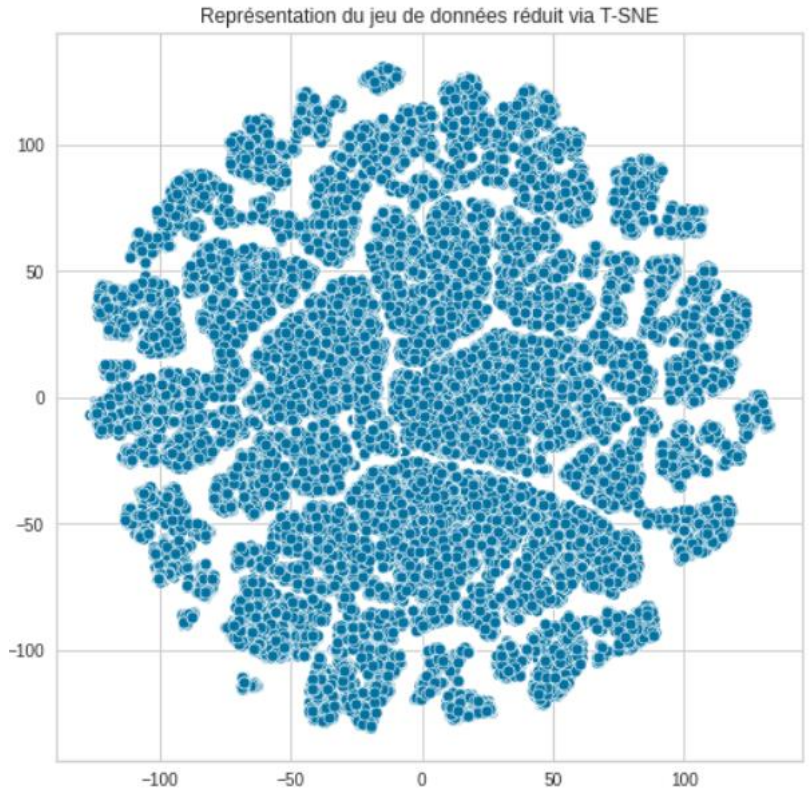


DBSCAN

Non approprié au dataset : 316 clusters obtenus



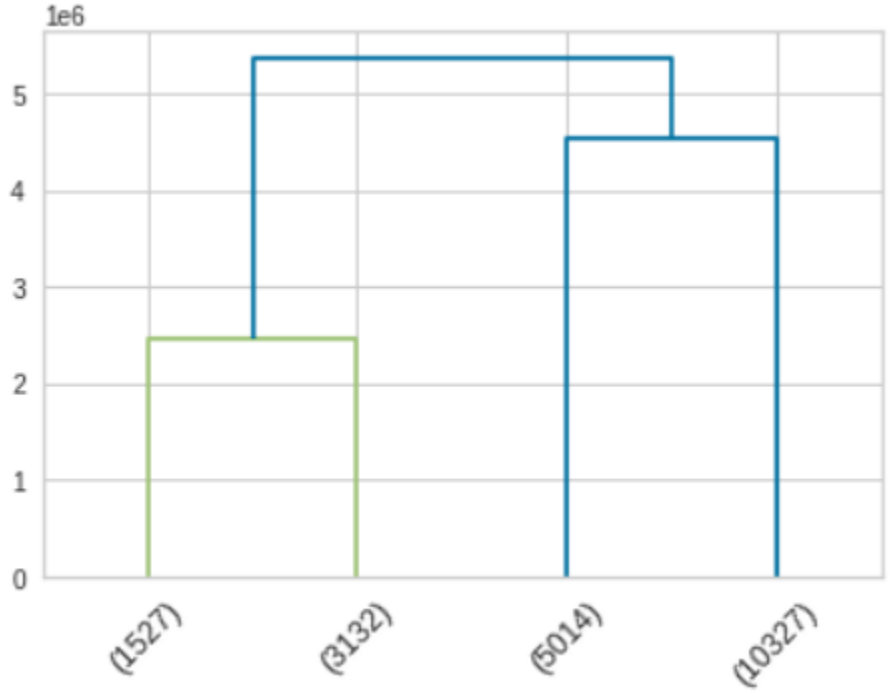
T-SNE sur ACP



Clusters non identifiables clairements

Si on s'intéresse aux ilôts, clusters trop nombreux

Agglomerative clustering



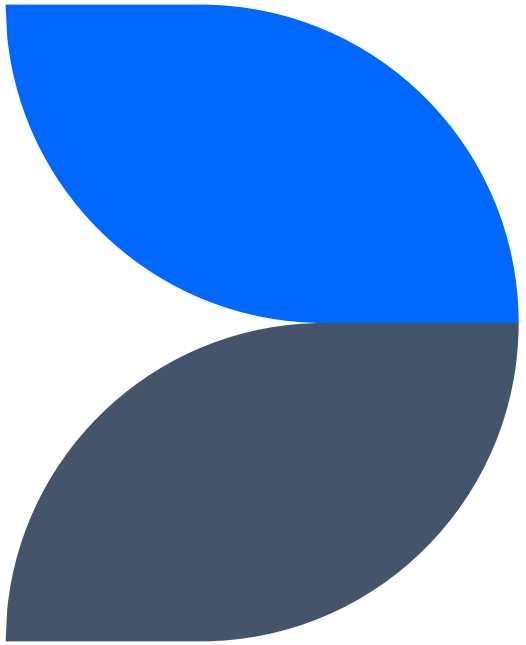
Effectué sur un sample de 20 000 individus

OPENCLASSROOMS

IV – Contrat de maintenance

Sur les données RFM (4 clusters)

Sur les données RFM étendues (6 clusters)



Méthodologie

Evolution de l'ARI (adjust rand index)

- On retire 6 mois du jeu de données
- On crée un modèle sur ces données (privées de 6 mois)
- On ajoute itérativement une semaine de données
- On compare via l'ARI :
  - La prédiction du modèle sur les nouvelles données
  - Le fit d'un nouveau modèle basé sur les nouvelles données

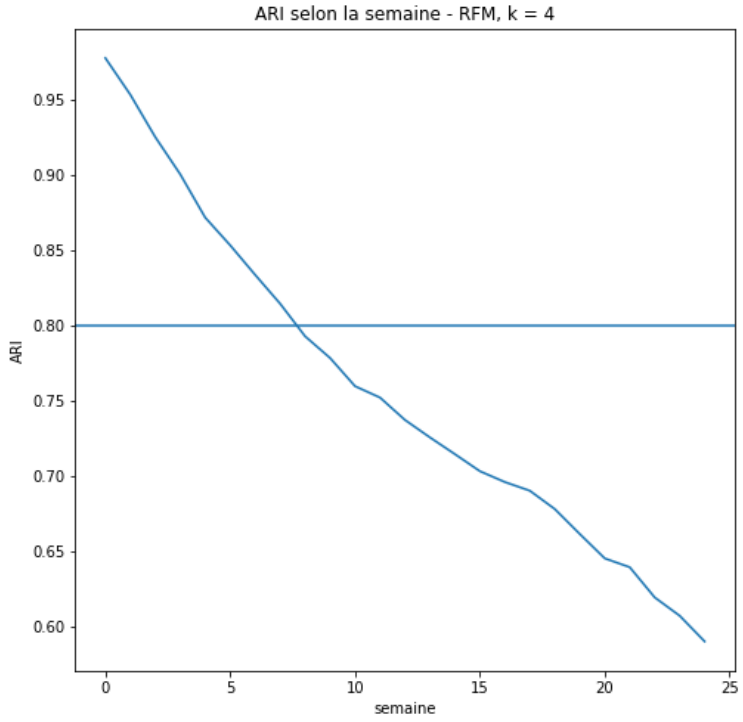
Etude de la robustesse

Après obtention d'un nombre optimal de semaines

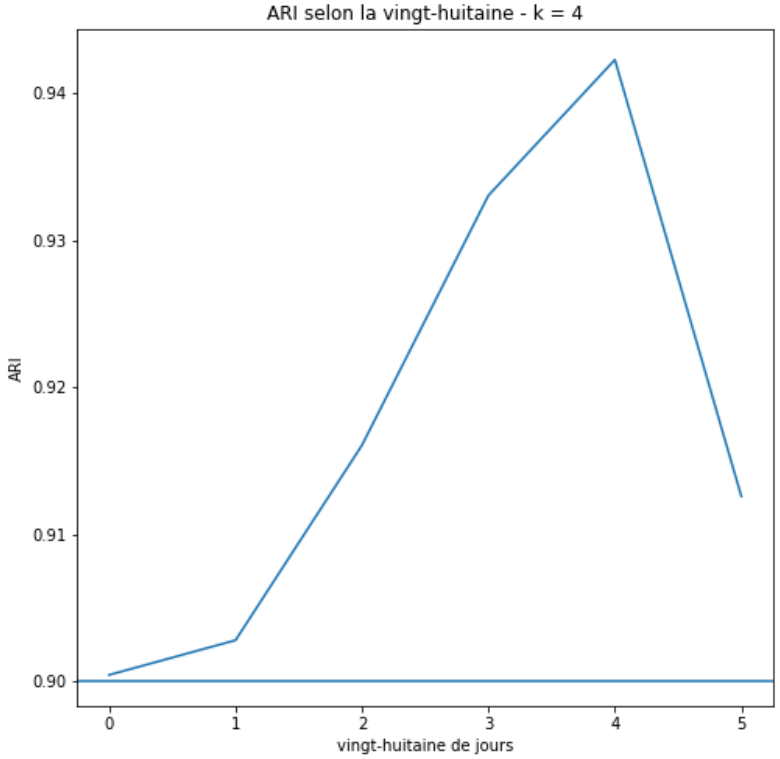
Suivi de l'évolution de l'ARI avec réapprentissage régulier

Sur les données RFM

Evolution de l'ARI – k = 4

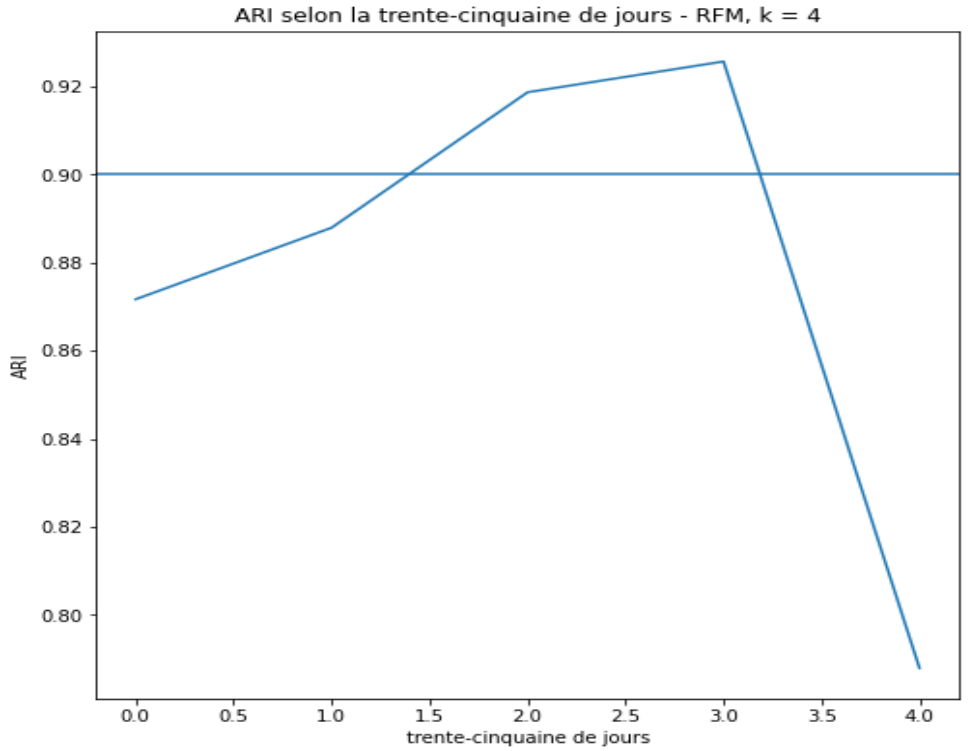


Robustesse pour 4 semaines



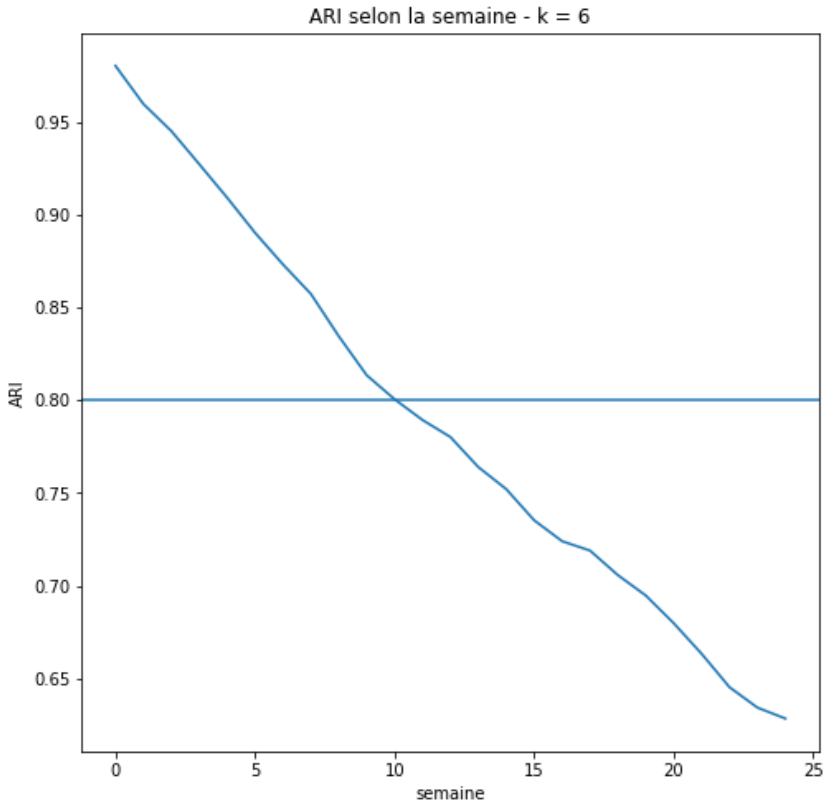
Sur les données RFM

Robustesse pour 5 semaines

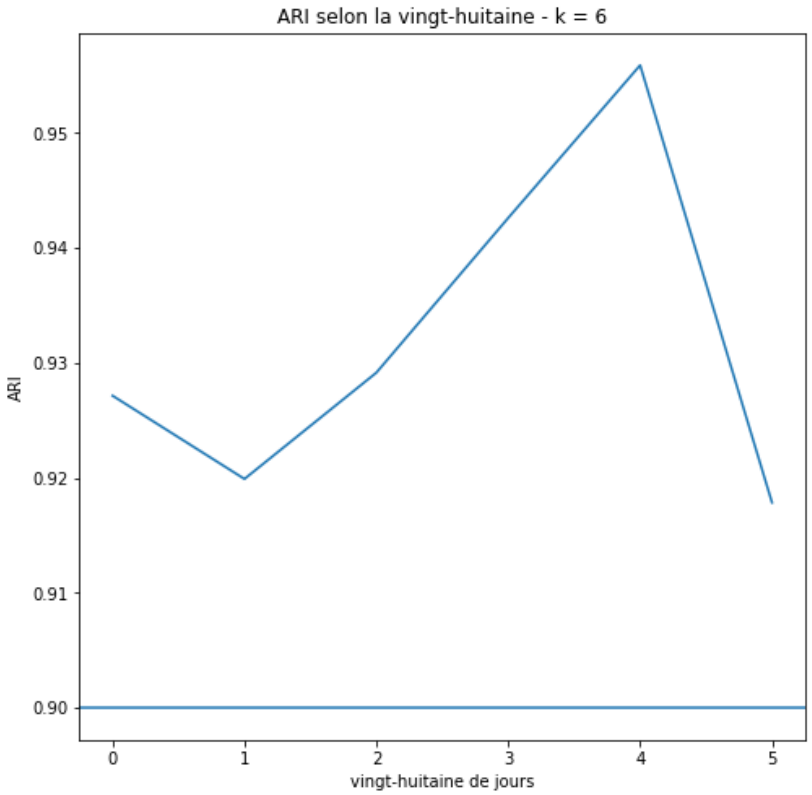


Sur les données RFM étendues

Evolution de l'ARI – k = 6

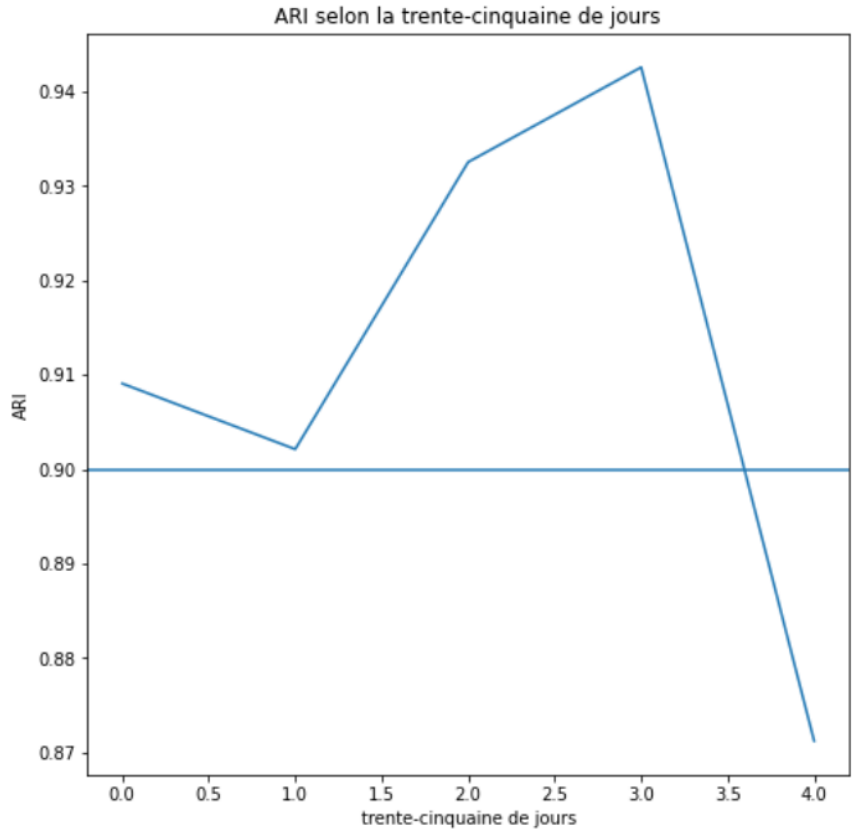


Robustesse pour 4 semaines



Sur les données RFM étendues

Robustesse pour 5 semaines



V - Synthèse

Deux modèles envisageables : 4 clusters ou 6 clusters

Mise à jour conseillée tous les mois, qu'importe le modèle validé

Notre recommandation : modèle RFM à 4 clusters

OPENCLASSROOMS

Merci

Fayz El Razaz