



**olist**  
empowering commerce

# Segmentation des clients e-commerce

Katrin-Misel Ponomarjova

Formation Ingénieur IA

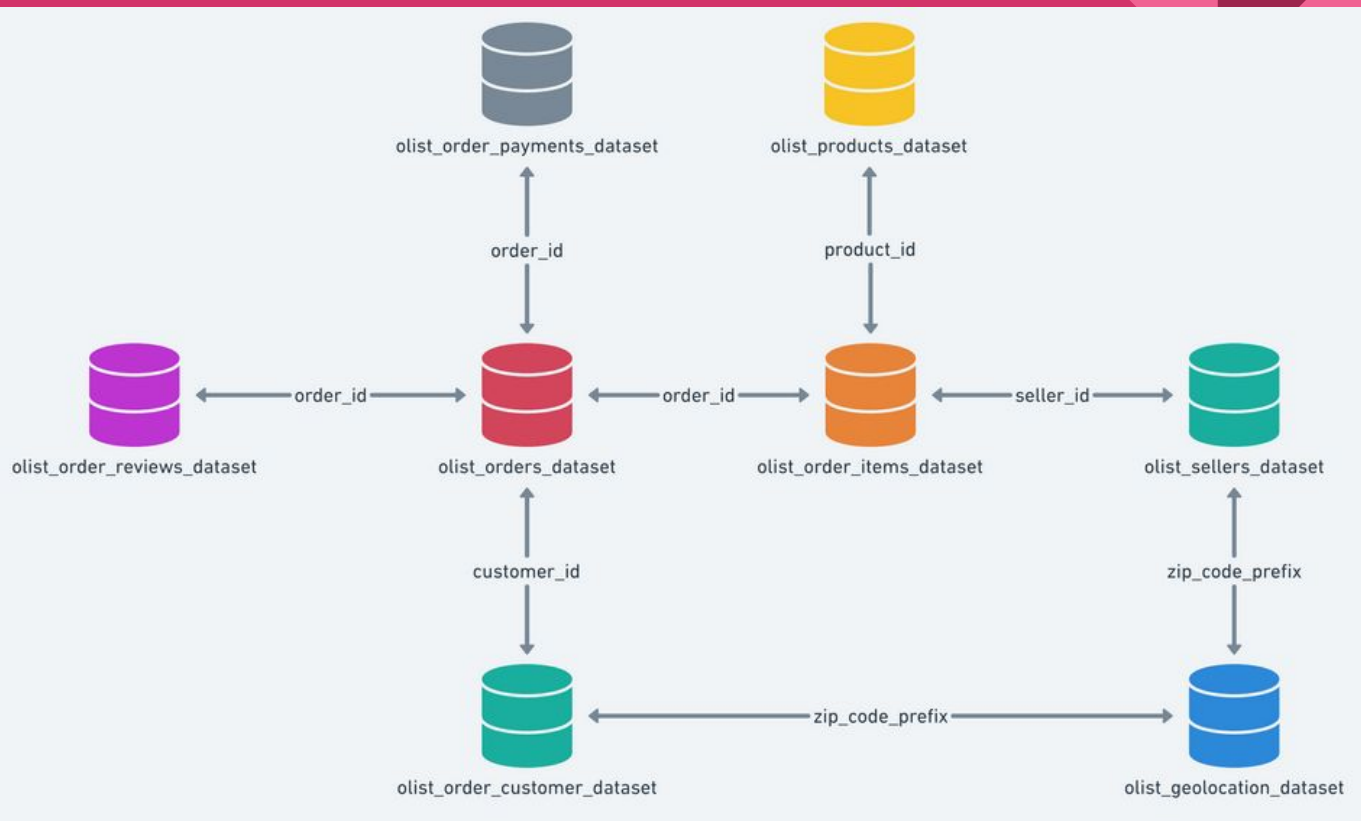
# Nettoyage et exploration

# Problématique : comprendre les clients Olist

La segmentation permet une meilleure compréhension de la clientèle

1. Améliorer l'efficacité de la communication
2. Comprendre les tendances d'achat de chaque groupe
3. Augmenter la probabilité d'achat des clients
4. Cibler les clients importants



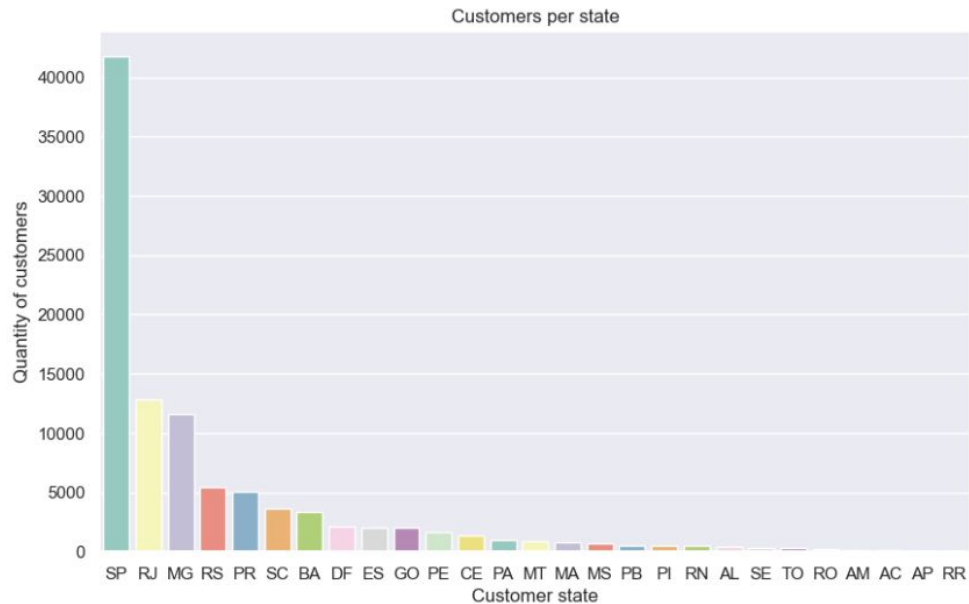


# Dataset clients

99441 commandes

96096 clients uniques

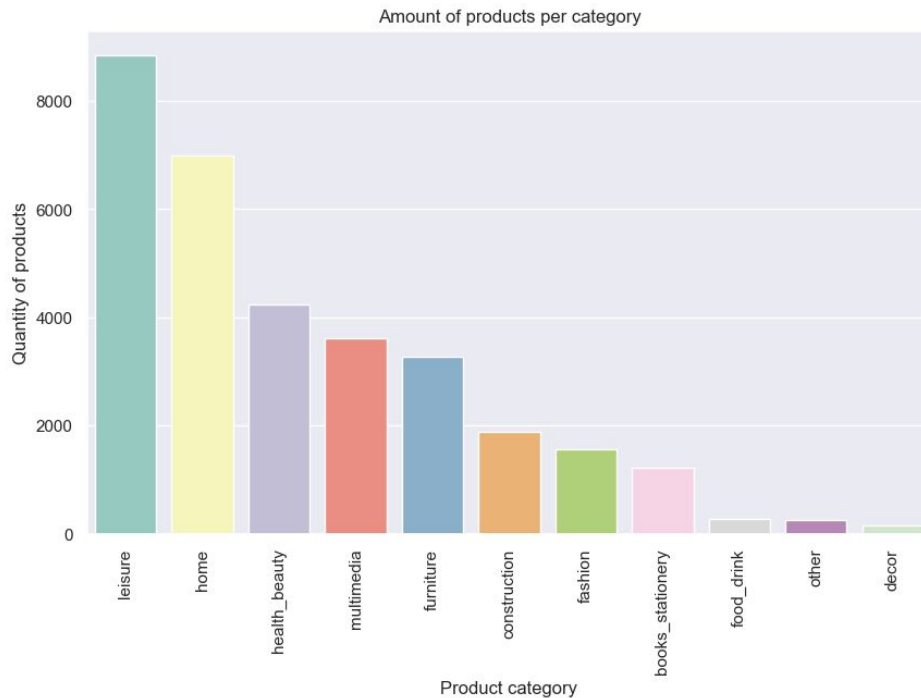
3.36% des clients qui reviennent



# Dataset produits

32951 produits

- Traduction des noms des catégories
- Regroupement sémantique des 71 catégories en 11



# Dataset commandes

**99441 commandes**

**2 jours**, délai moyen pour transmettre la commande au livreur

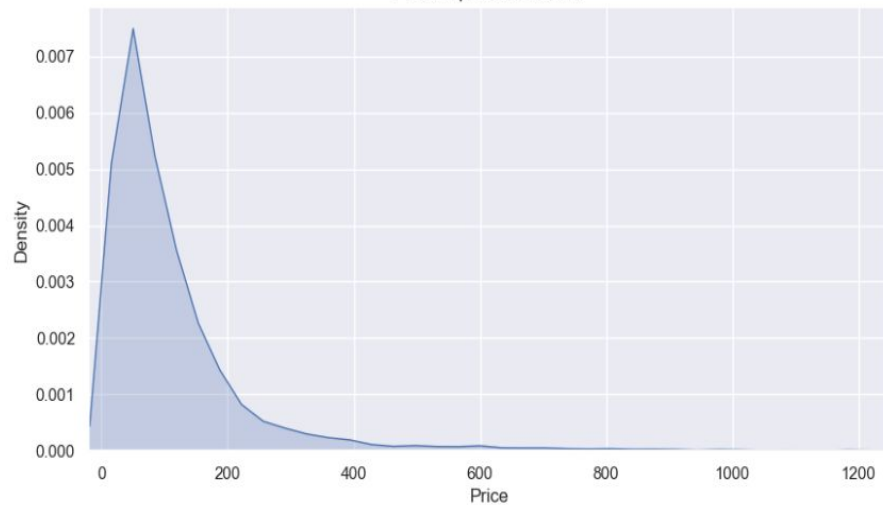
**9 jours**, délai moyen de livraison

**11 jours**, différence moyenne entre la date de livraison estimée et réelle

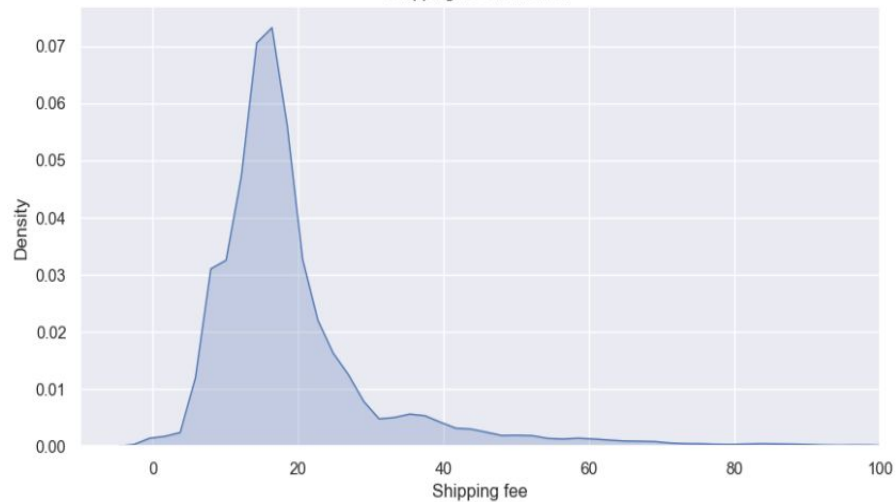
Feature engineering : mois de commande, délai de livraison, delta date estimée et réelle

# Dataset order items

Product price distribution

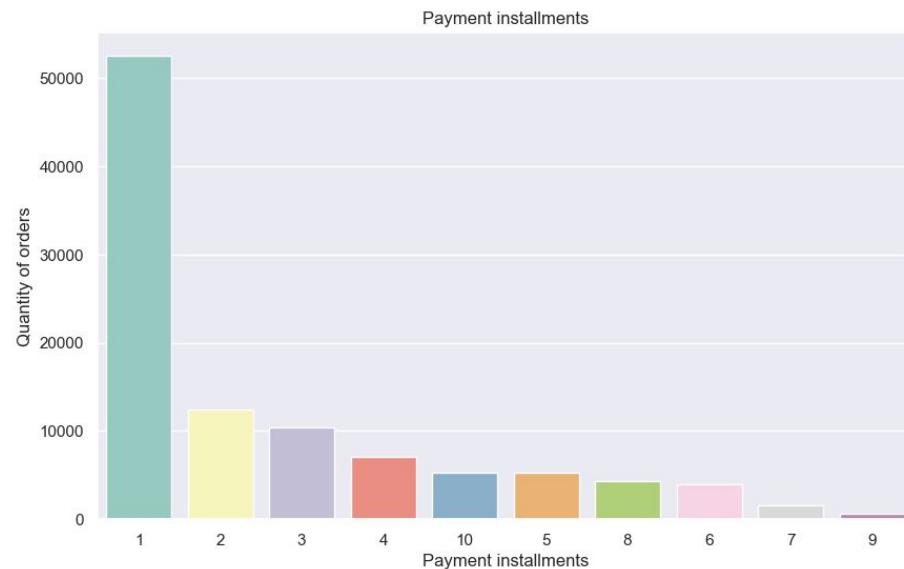
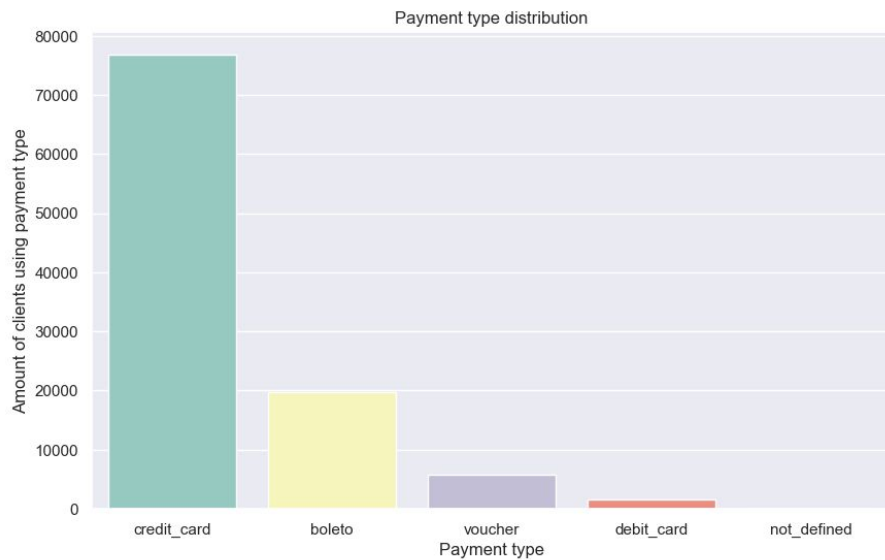


Shipping fee distribution





# Dataset paiement



# Dataset reviews

Review score de 5 corrélé avec :

Une baisse de délai de livraison

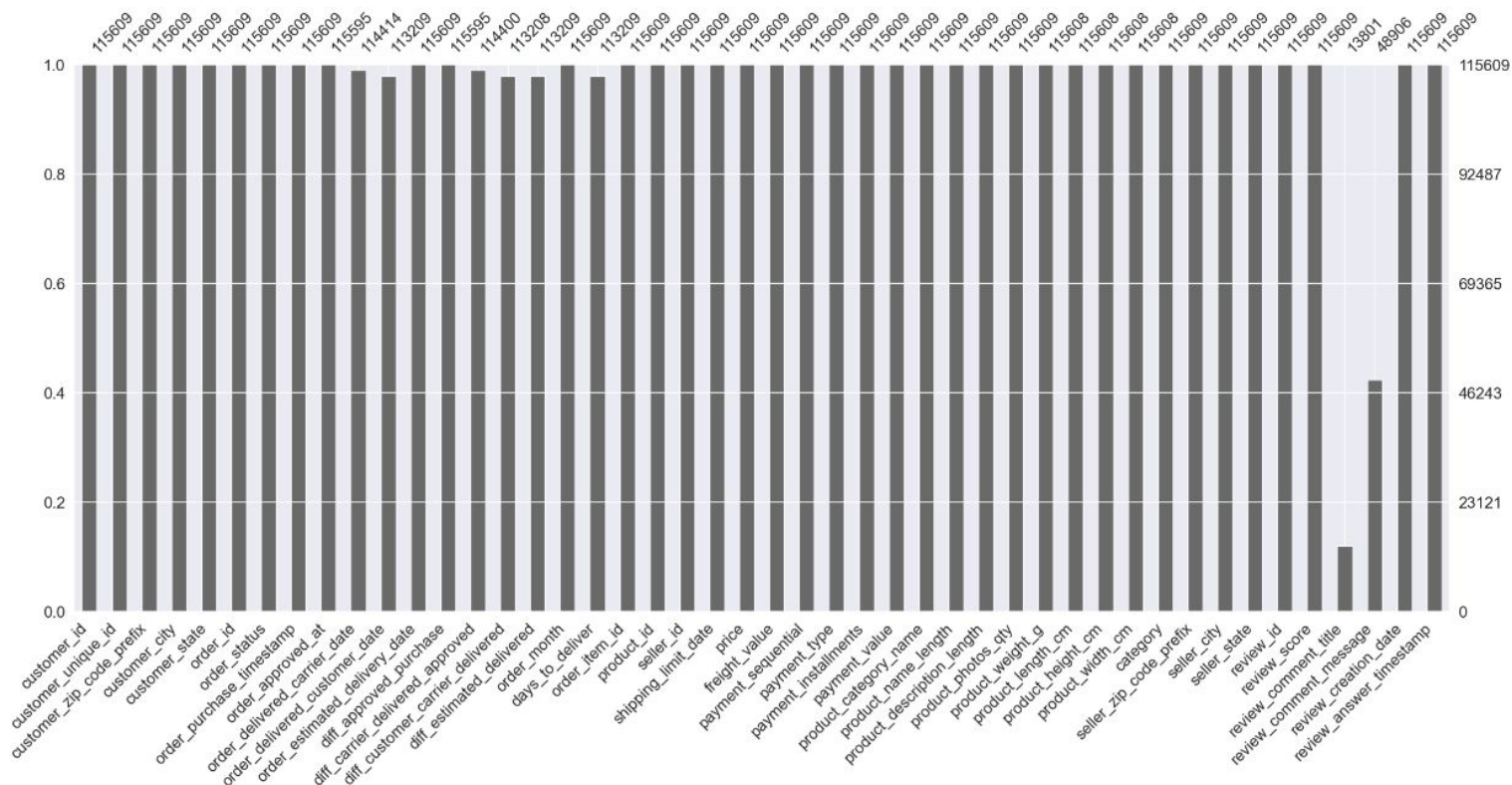
Une baisse de frais de livraison

Une commande moins chère

Une commande moins volumineuse



## Cleaning and merging



# Modélisation



**RFM :**

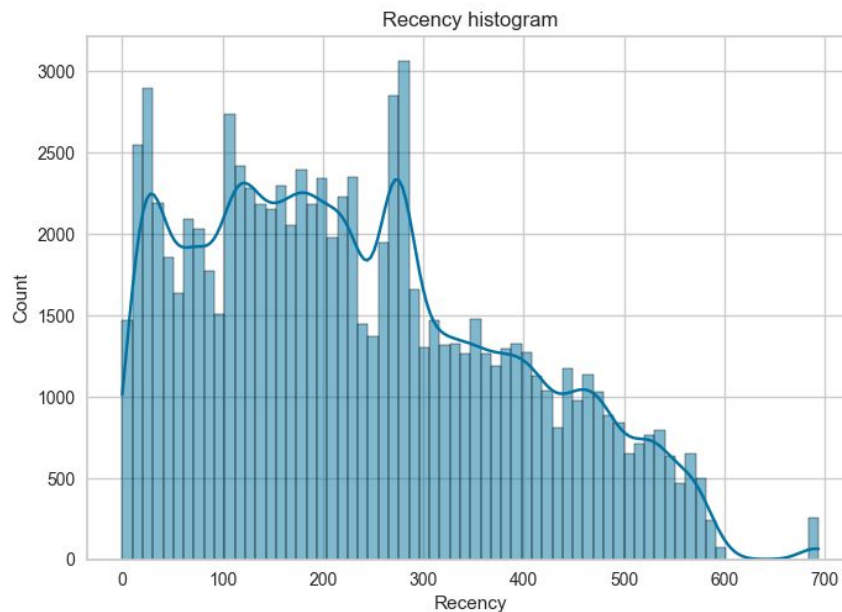
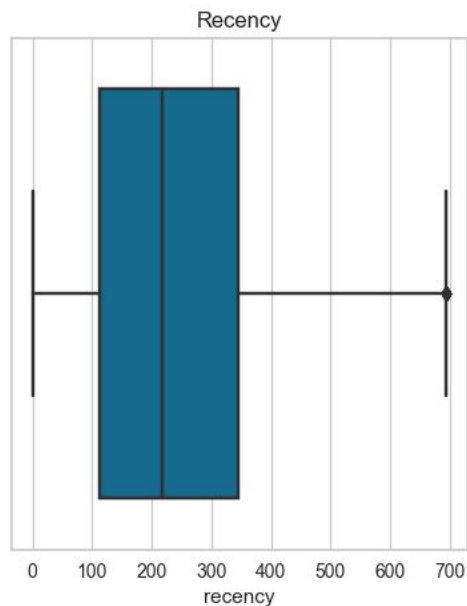
Recency

Frequency

Monetary value

# Recency

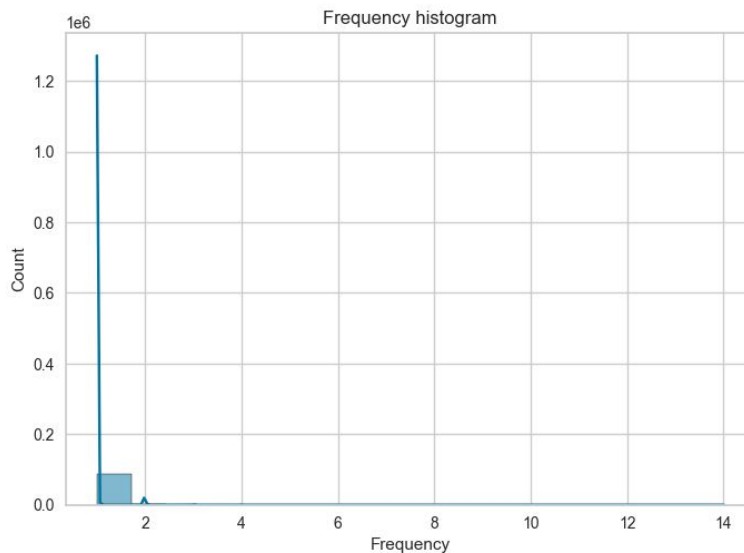
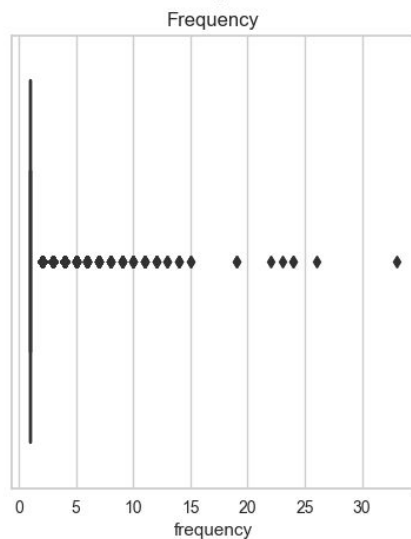
Un pic de commandes à 300 j depuis la dernière commande



# Frequency

Environ 3% des clients reviennent acheter sur Olist

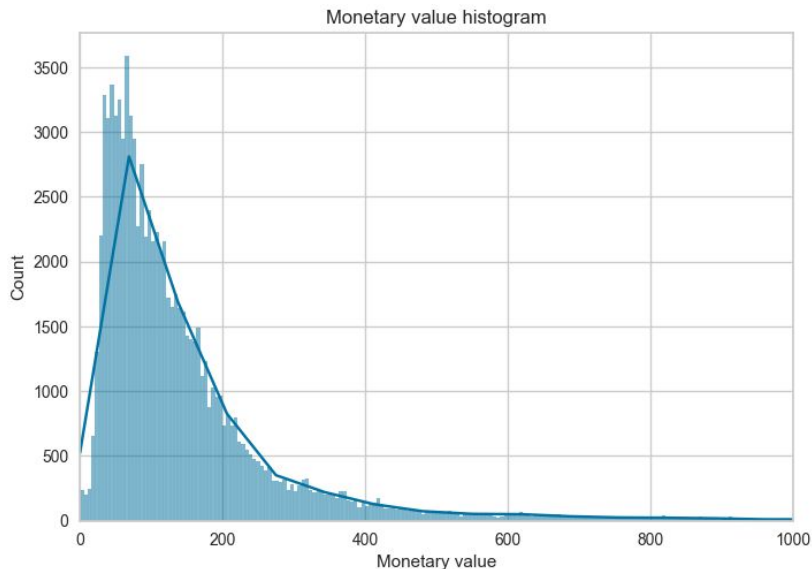
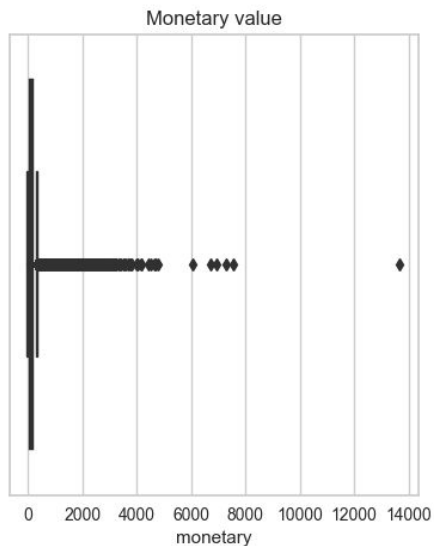
Le maximum est de 33 commandes



# Monetary value

La valeur d'une commande moyenne est de **162 BRL**

La commande la plus chère était de **13664 BRL**





# RFM segments



New : 31.4%

Gone : 31.2%

Regular : 22.2%

Spender : 12%

Loyal : 1.9%

Main : 1%

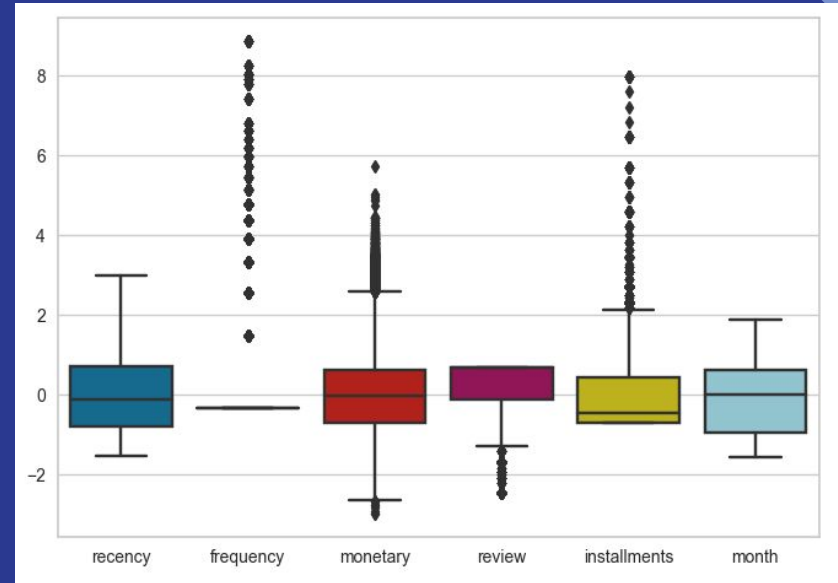
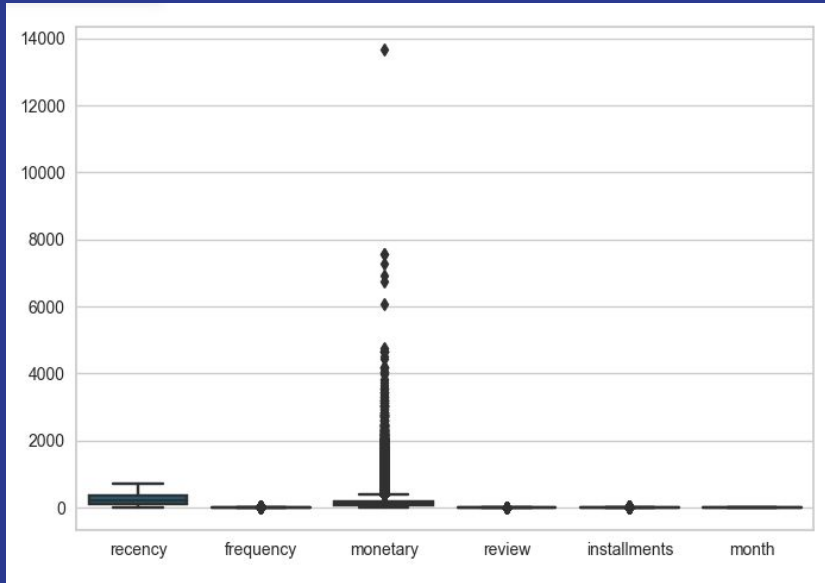
# KMeans

Variables prises en compte dans le modèle :

- recency
- frequency
- monetary
- review
- installments
- month

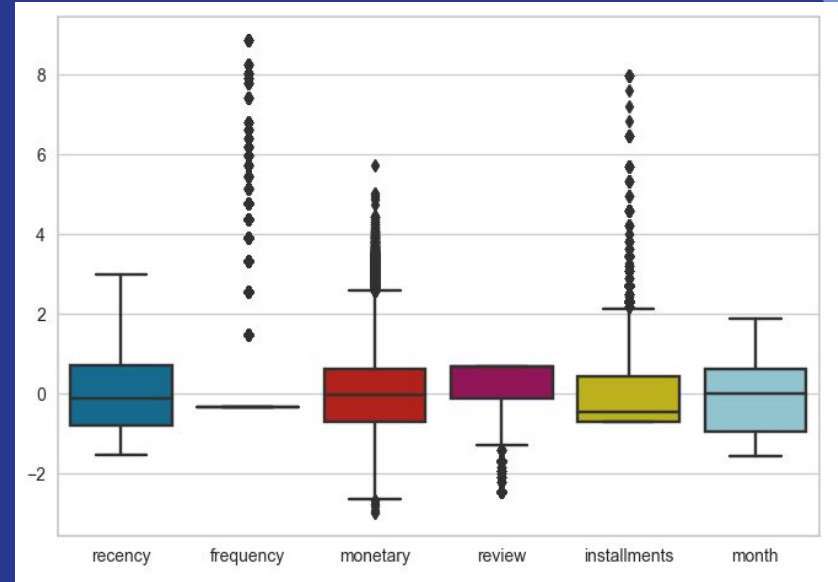
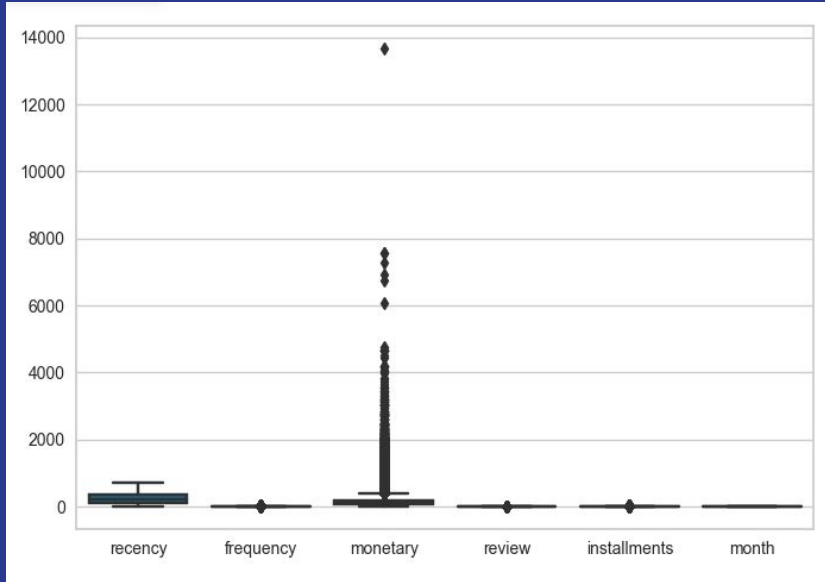


# Distributions des variables avant et après preprocessing

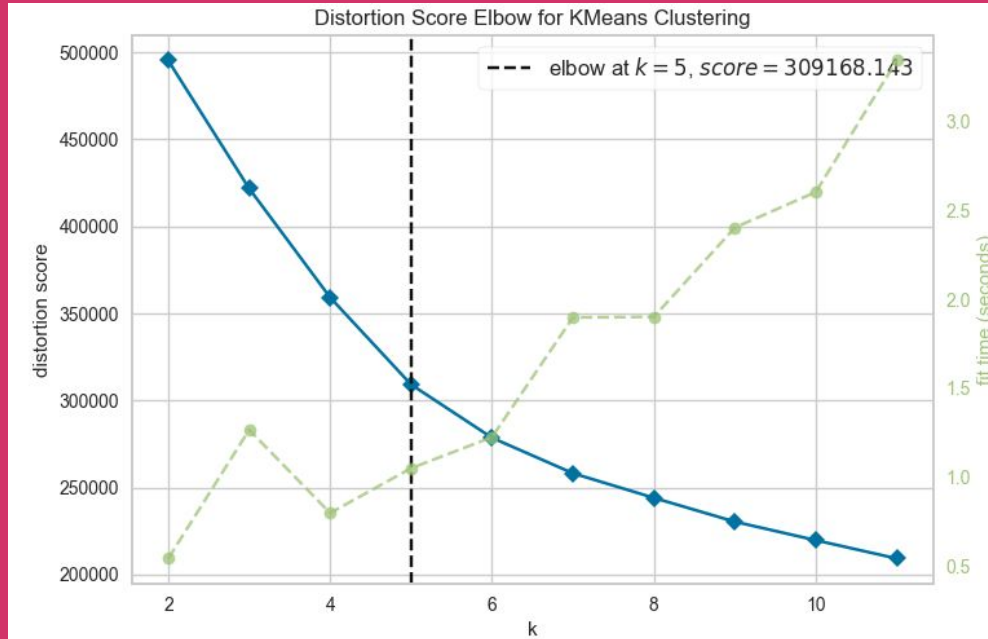


# Distributions des variables avant et après preprocessing

Transformation logarithmique et StandardScaler()



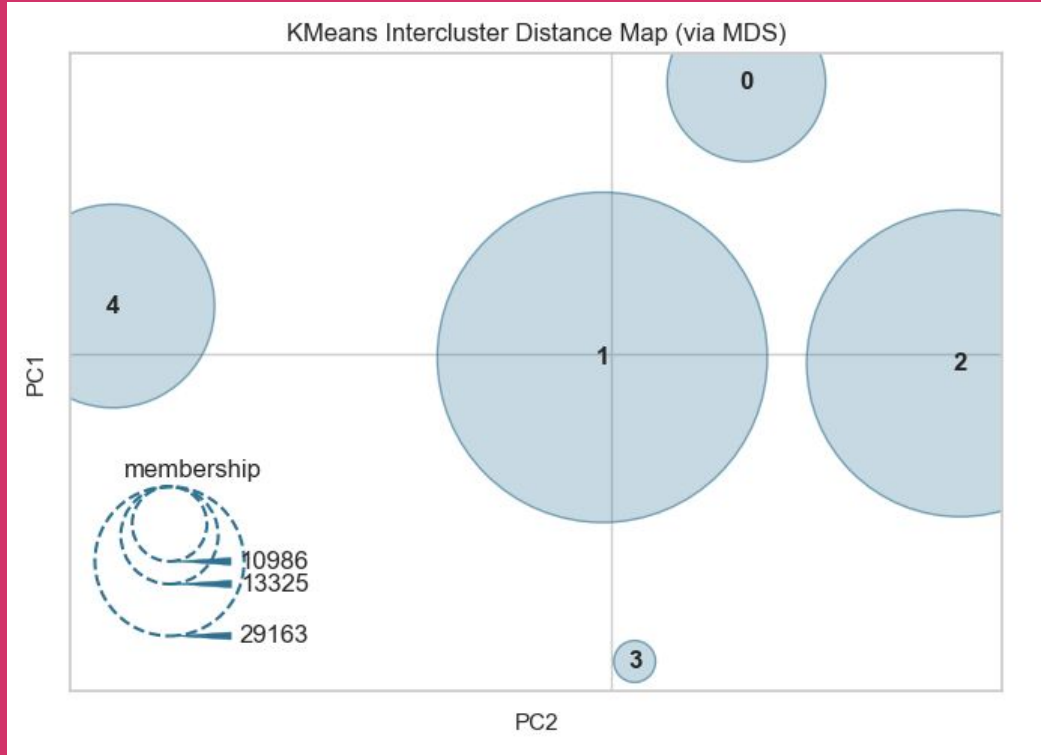
# Choix de quantité de clusters



**Distortion score :**

la moyenne des **distances**  
**au carré** depuis les centres  
de cluster des clusters  
respectifs

# Intercluster distance



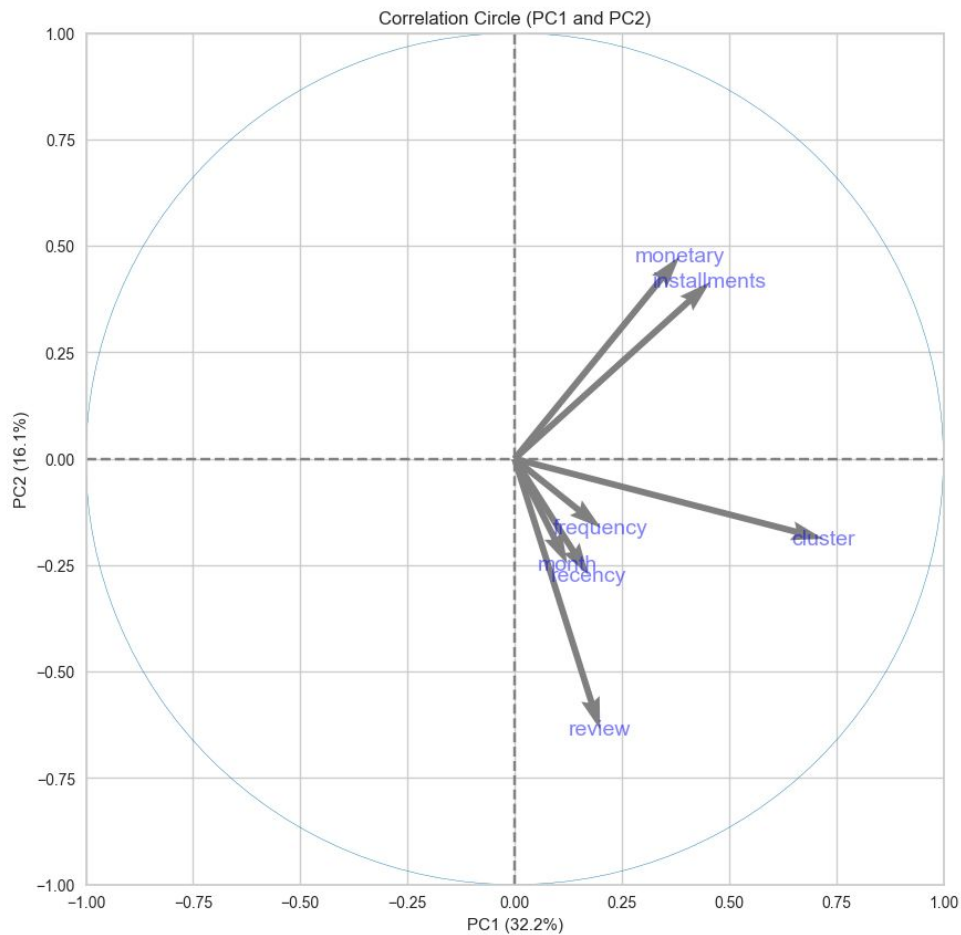
# Comparaison avec DBSCAN

KMeans silhouette score : la qualité des clusters, de -1 à 1

Différence entre la distance moyenne avec les points du même cluster (cohésion) et la distance moyenne avec les points des autres clusters (séparation).

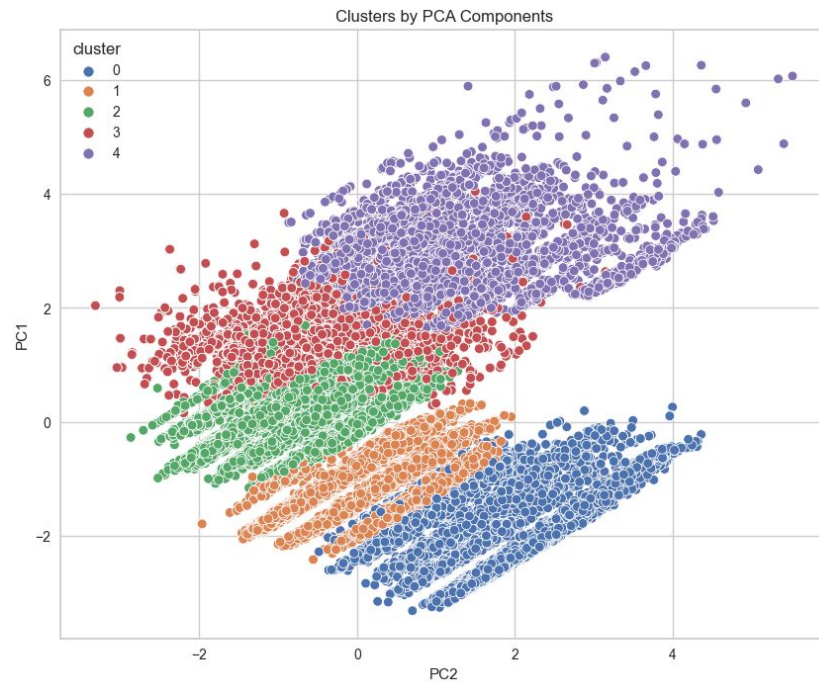
KMeans : silhouette score 0.26, 5 clusters

DBSCAN : silhouette score -0.1, 1506 clusters estimés

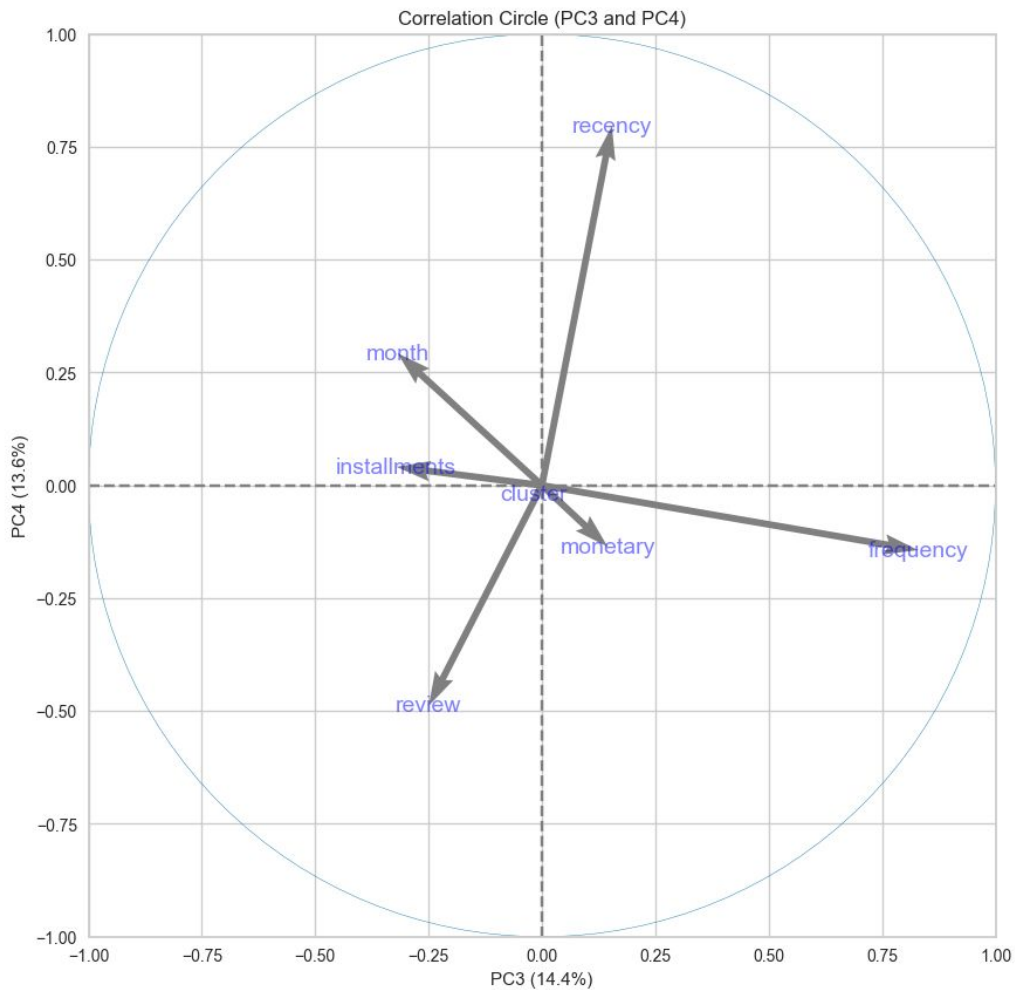


PC1 : high monetary value, many installments

PC2 : lower reviews, older customers, bought once

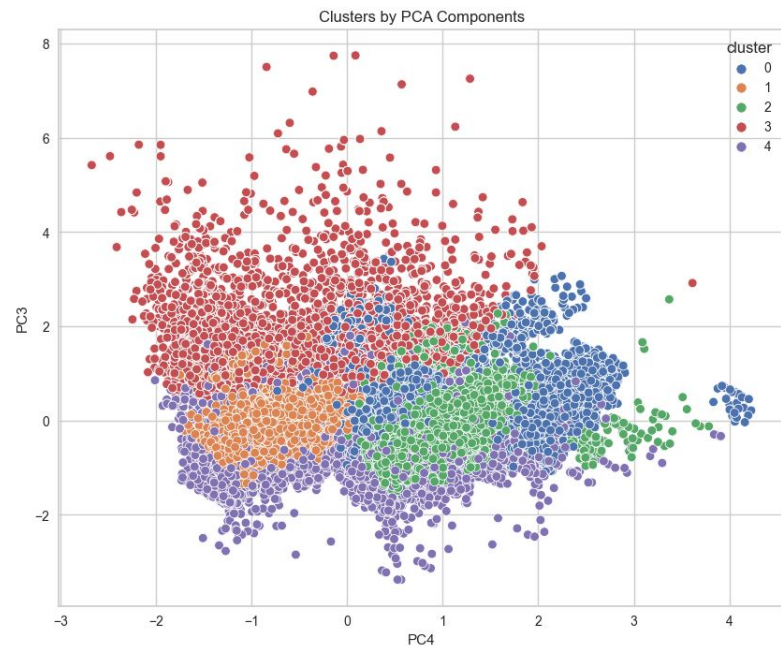




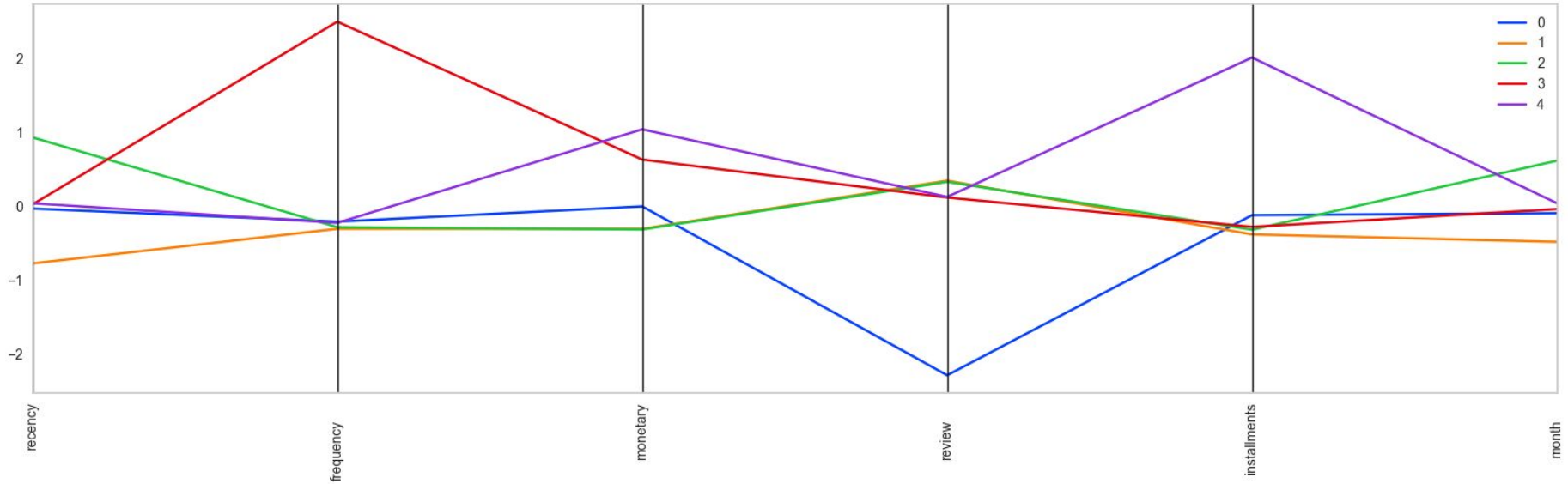


PC3 : frequent customers

PC4 : recent customers, low reviews



Parallel Coordinates plot for the Centroids



Cluster 0 : low reviews

Cluster 1 : oldest, pay in less installments, buy earlier in the year

Cluster 2 : newest customers, buy later in the year

Cluster 3 : frequent buyers, spend more than average

Cluster 4 : spend the most, pay in many installments

*Le mécontent*

**11%**

**Cluster 1**

Dernier achat il y a  
8 mois

A acheté 1 fois

A dépensé 165  
BRL

Payé en 2.5 fois

Achète en  
mai-juin

Review 1/5

*L'actif*

**36%**

**Cluster 2**

Dernier achat il y a  
4 mois

A acheté 1 fois

A dépensé 116  
BRL

Payé en 2 fois

Achète en  
avri-mai

Review 4.6/5

*Le perdu*

**29%**

**Cluster 3**

Dernier achat il y a  
12 mois

A acheté 1 fois

A dépensé 117  
BRL

Payé en 2 fois

Achète en août

Review 4.6/5

*Le fidèle*

**10%**

**Cluster 4**

Dernier achat il y a  
8 mois

A acheté 3.5 fois

A dépensé 273  
BRL

Payé en 2 fois

Achète en  
mai-juin

Review 4/5

*Le dépensier*

**13%**

**Cluster 5**

Dernier achat il y a  
8 mois

A acheté 1 fois

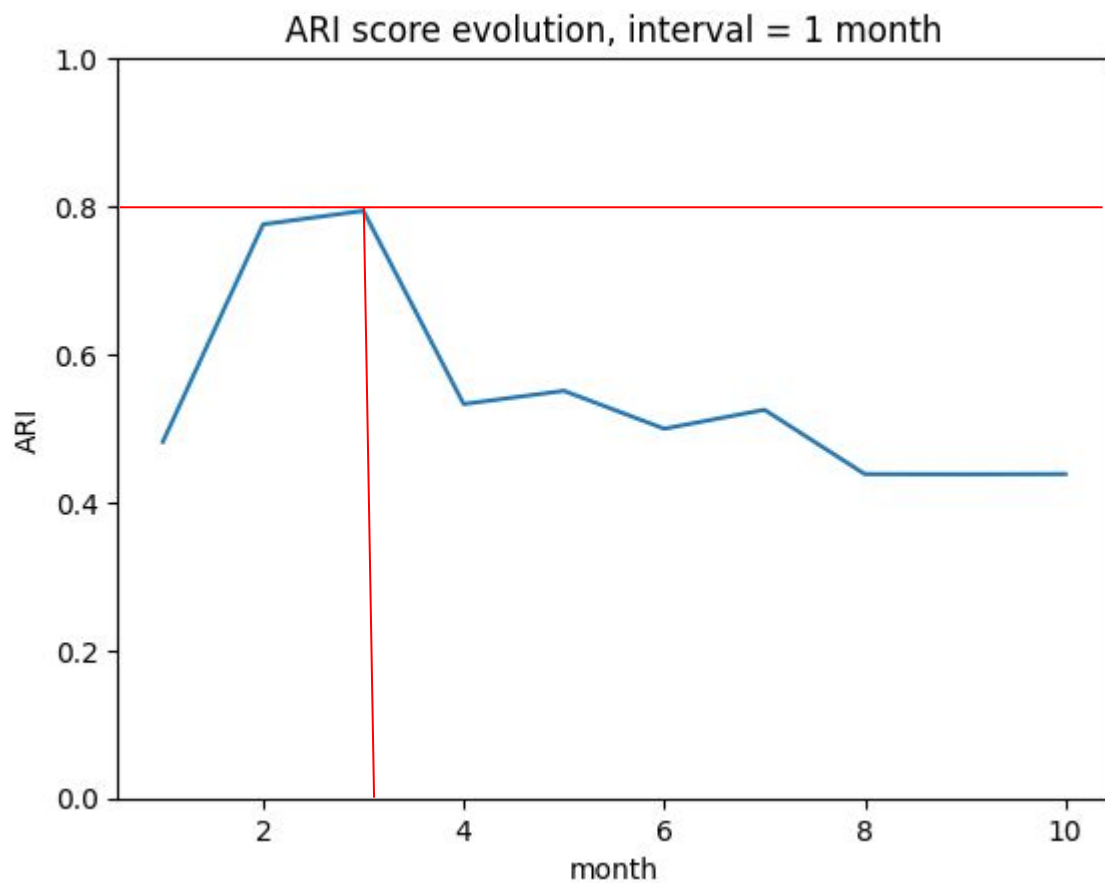
A dépensé 384  
BRL

Payé en 8 fois

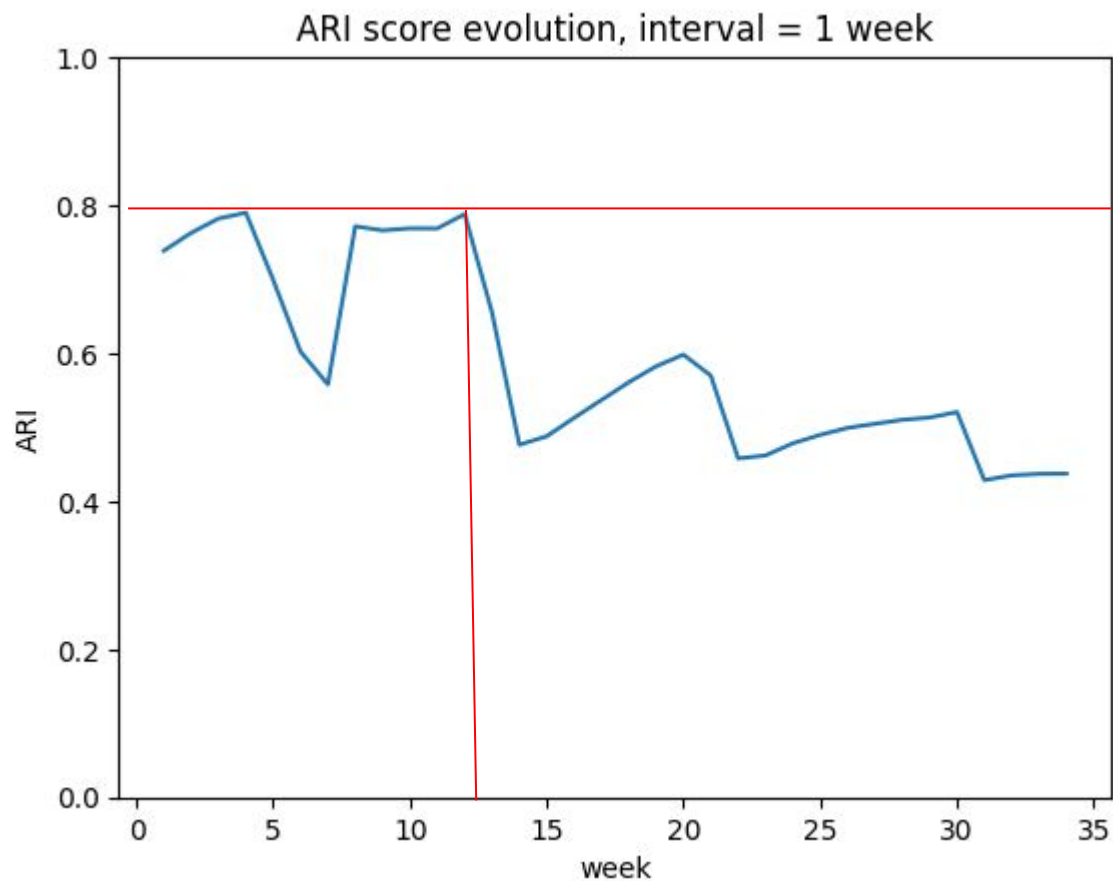
Achète en juin

Review 4/5

# Stabilité du modèle



month		ARI
0	1	0.481578
1	2	0.775614
2	3	0.793945
3	4	0.532816
4	5	0.550655
5	6	0.499465
6	7	0.525058
7	8	0.438124
8	9	0.437925
9	10	0.438172



week		ARI
0	1	0.738616
1	2	0.762719
2	3	0.782074
3	4	0.790059
4	5	0.699201
5	6	0.602255
6	7	0.557990
7	8	0.771625
8	9	0.766095
9	10	0.768901
10	11	0.768762
11	12	0.788142
12	13	0.655346
13	14	0.476402
14	15	0.487930
15	16	0.512729
16	17	0.536880
17	18	0.560867
18	19	0.582462
19	20	0.598265
20	21	0.569865
21	22	0.457967

PEP-8 compliant with Black



# Discussion