SEGMENTATION

MARKETING

OpenClassrooms - Projet 5

olist

Marketplace e-commerce

Mise en relation Acheteur / Vendeur

Objectifs

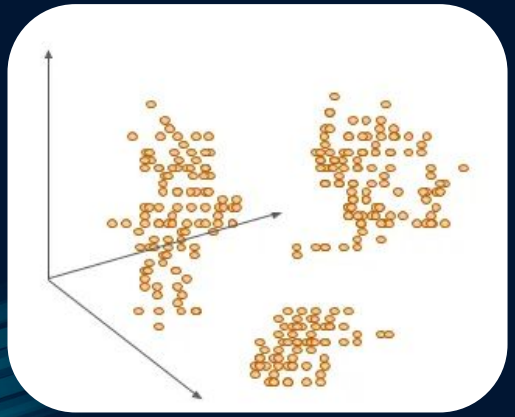
Segmentation clientèle

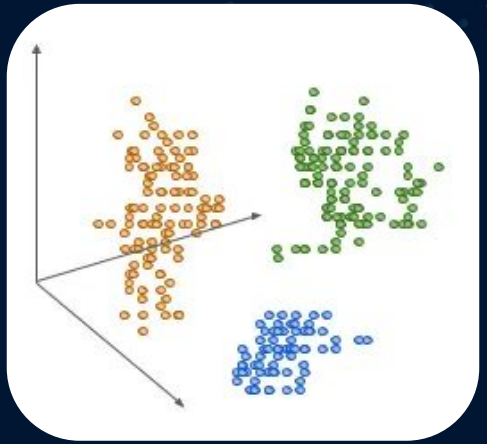
Comprendre la clientèle

Segmentation actionnable

Identifier maintenance

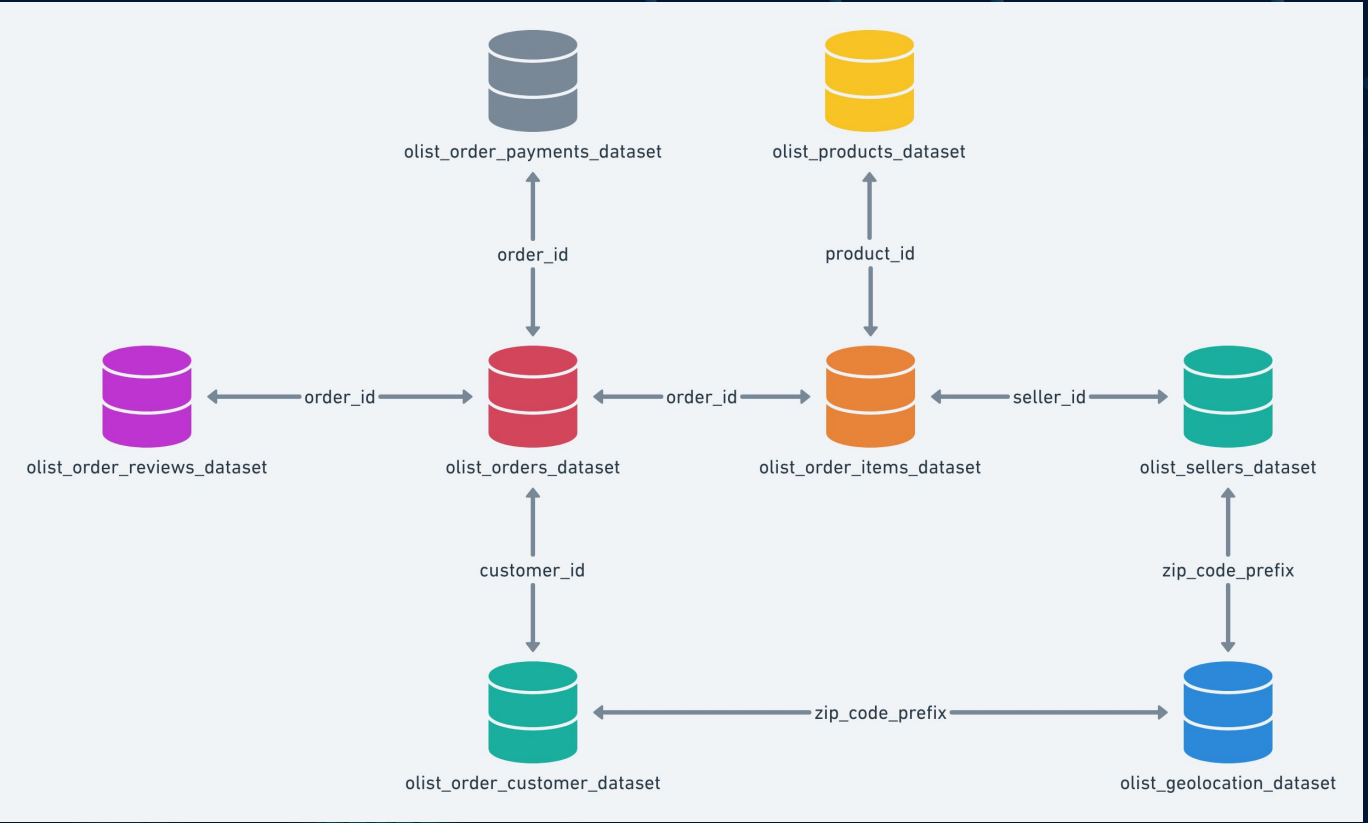
Classification non supervisée





Données

Données


olist_order_payments_dataset
olist_products_dataset
order_id
product_id
olist_order_reviews_dataset
order_id
olist_orders_dataset
order_id
olist_order_items_dataset
seller_id
olist_sellers_dataset
customer_id
zip_code_prefix
olist_order_customer_dataset
zip_code_prefix
olist_geolocation_dataset

Traitements effectués

Agrégation

Somme
Moyenne
Nombre occurrence

Engineering

Création de nouvelles variables

Nettoyage

Valeurs aberrantes
Valeurs manquantes

Scaling

Mise à l'échelle
Normalisation

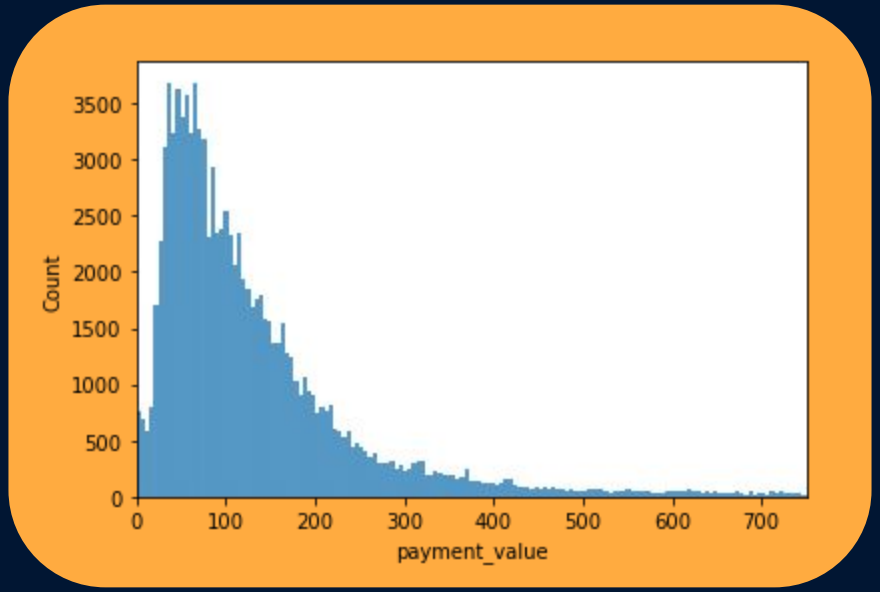
Vue générale

99440
commandes

16 MR$
vente

27
Mois d'activité

Panier



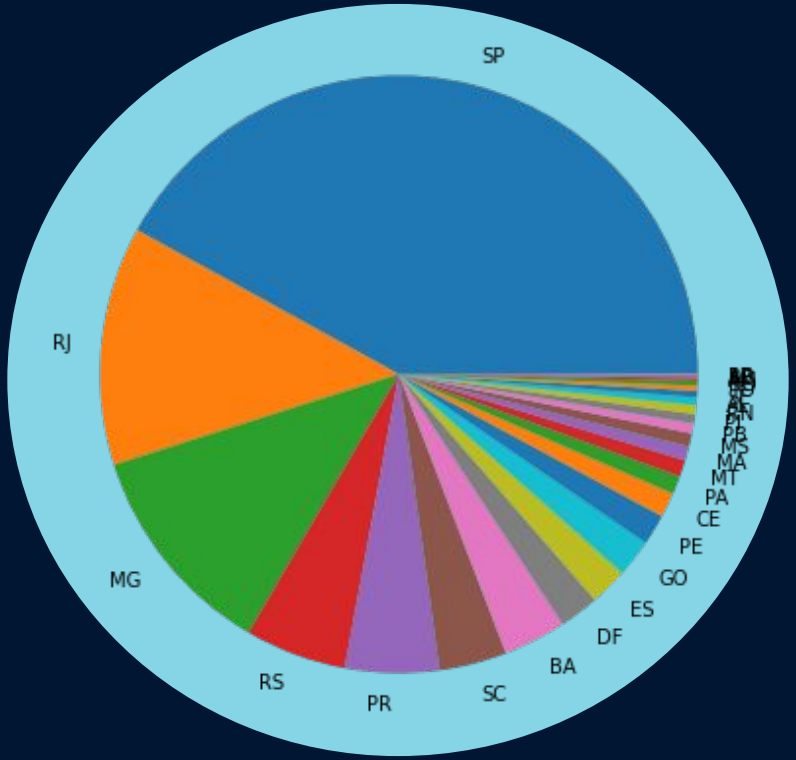
154 R$

Panier moyen

75%

Inférieur à 171 R$

Localisation



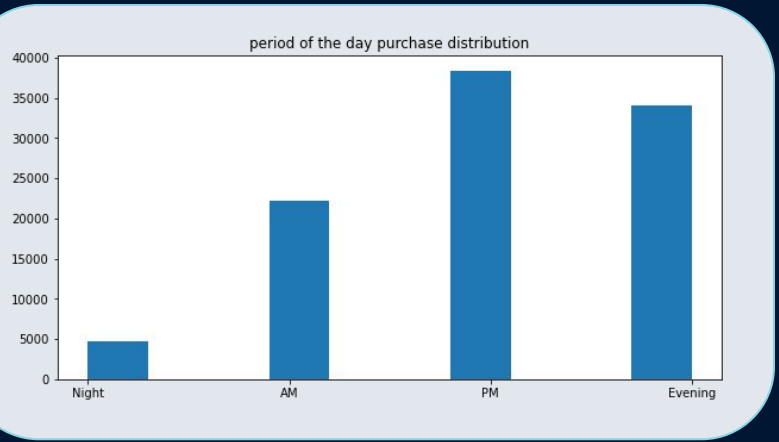
42%

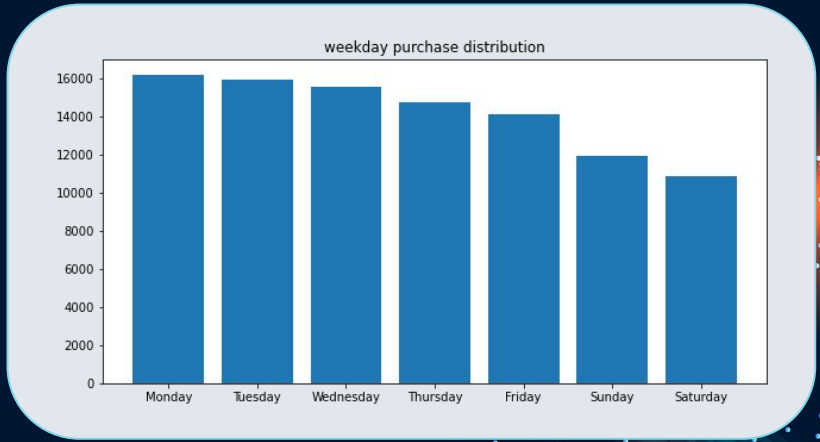
Etat de São Paulo

66%

Sur 3 états

Temporalité





Fréquence

94 %
Achat unique

5 %
Achat double

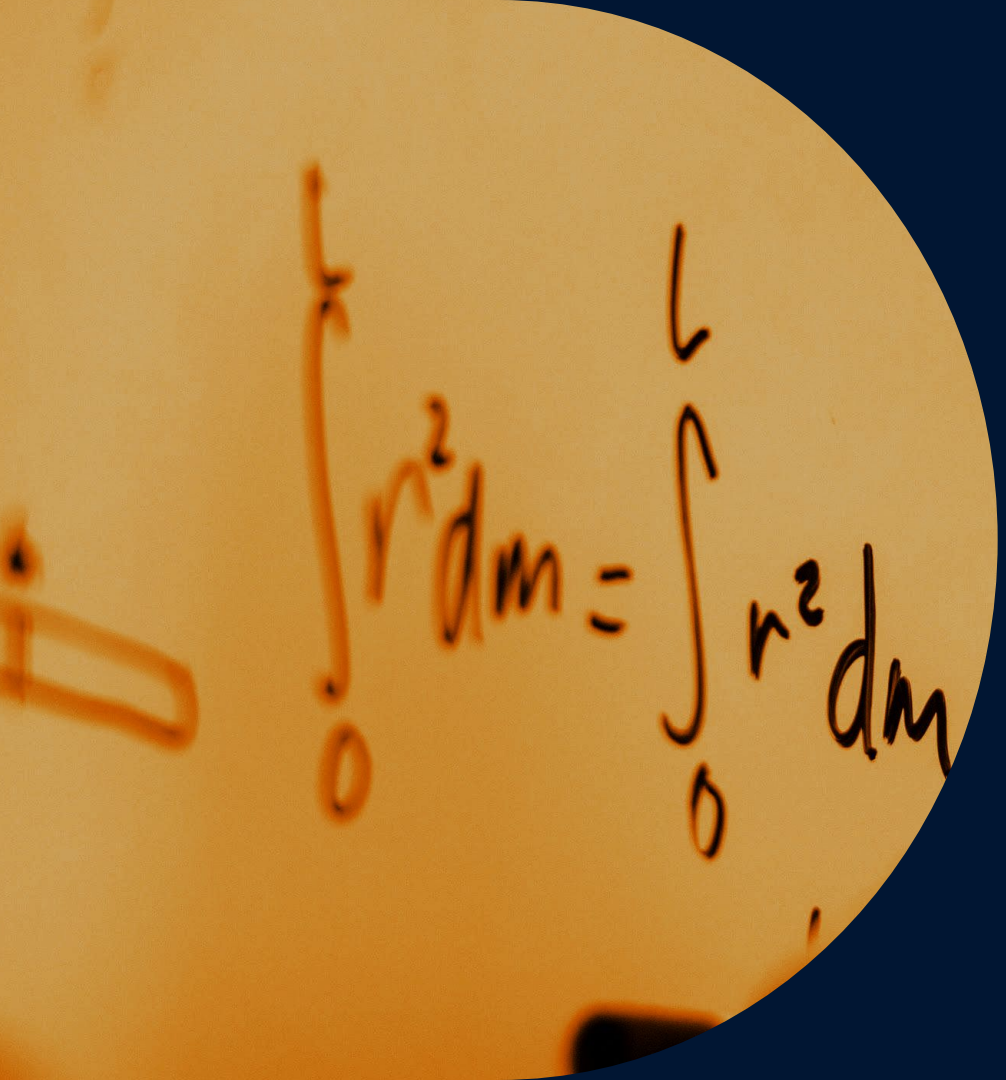
Avis

76 %

Avis supérieur à 4

15 %

Avis inférieur à 2



Modèles

Segmentation RFM

Nombre de cluster choisi à priori

K-means

Variable 1

Variable 2

Variable n

Quantitatives

TRANSFORMATION

SCALING

QuantileTransformer

StandardScaler

Cluster 1

Cluster 2

Cluster k

Minimum locaux
plusieurs itérations

Methodologie

RFM Clustering

Méthode ‘Manuelle’ 3 segments

ML 3 segments

K-means RFM

K-means RFM

Optimisation

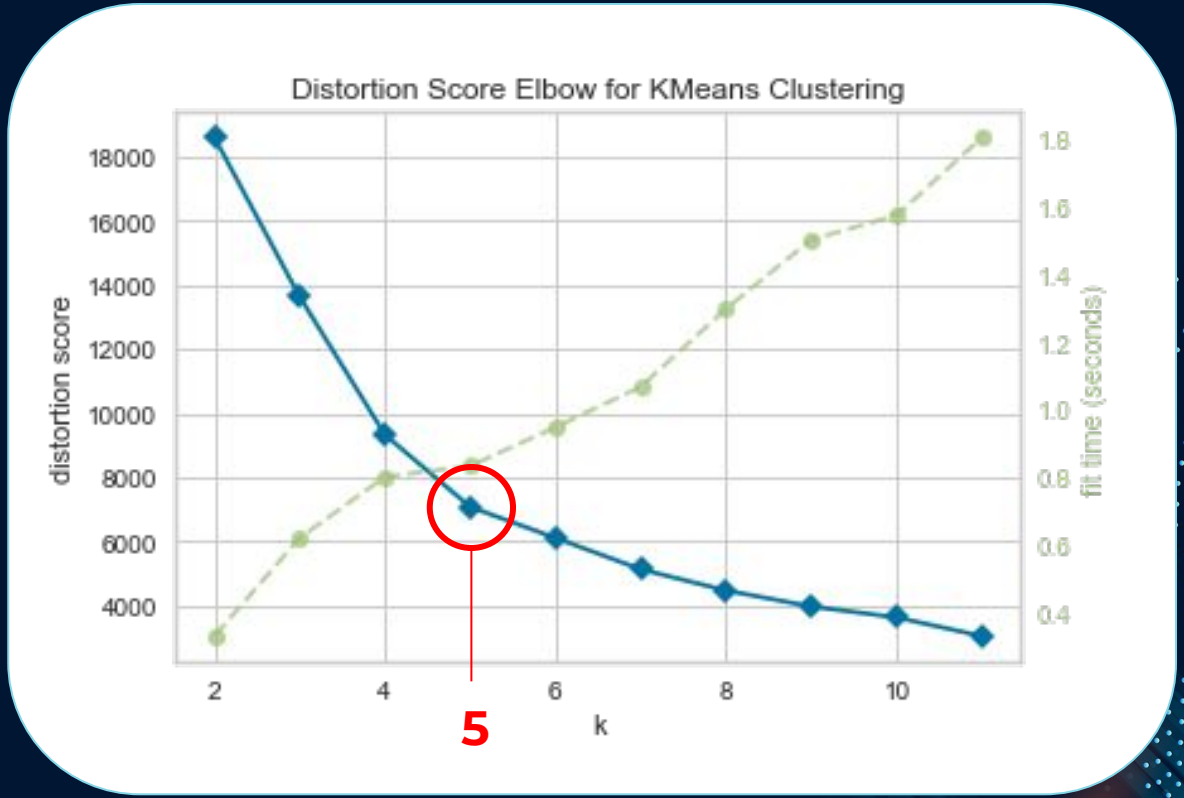
ACP Ajout de variable

K-means alternatif

Comparison Selection

K-means

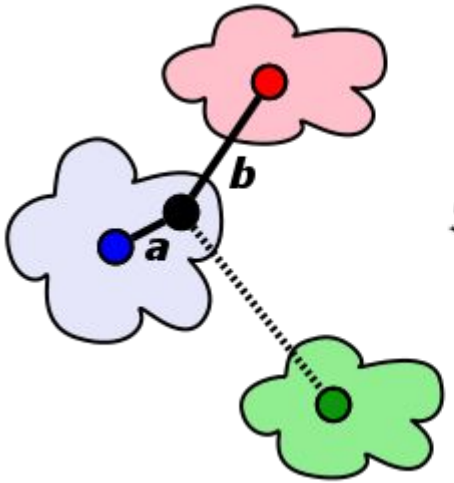
Optimisation



Evaluation

Score de silhouette

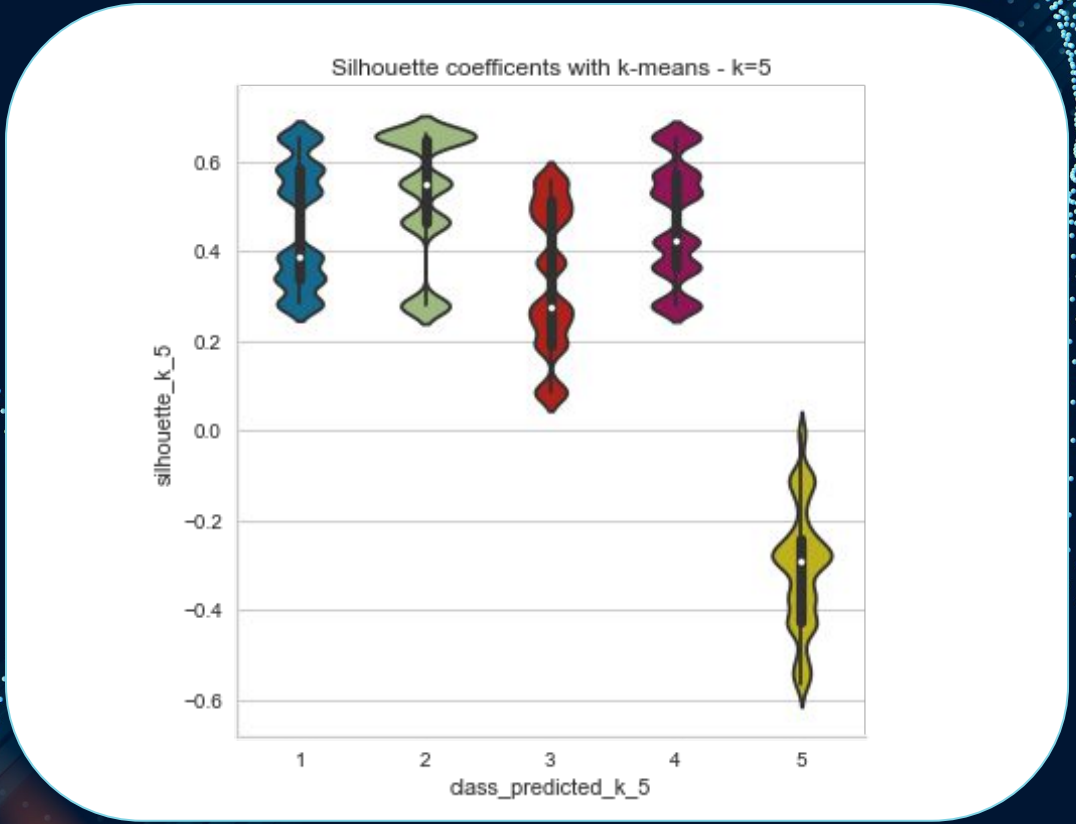
Performance brut



$$SSI_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

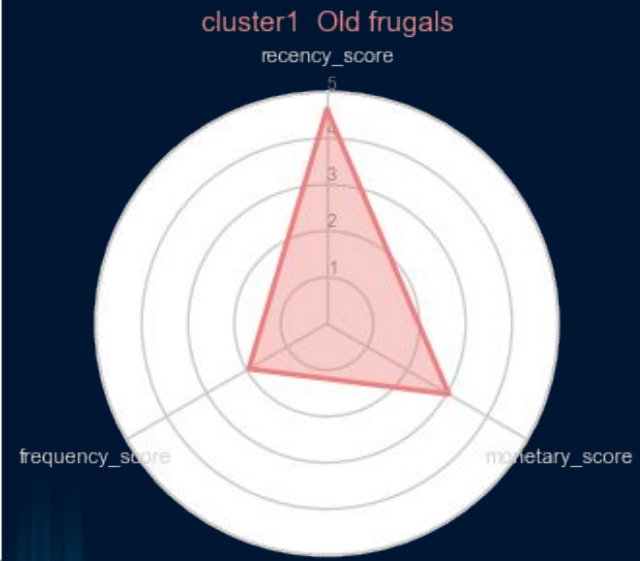
Score de silhouette par cluster

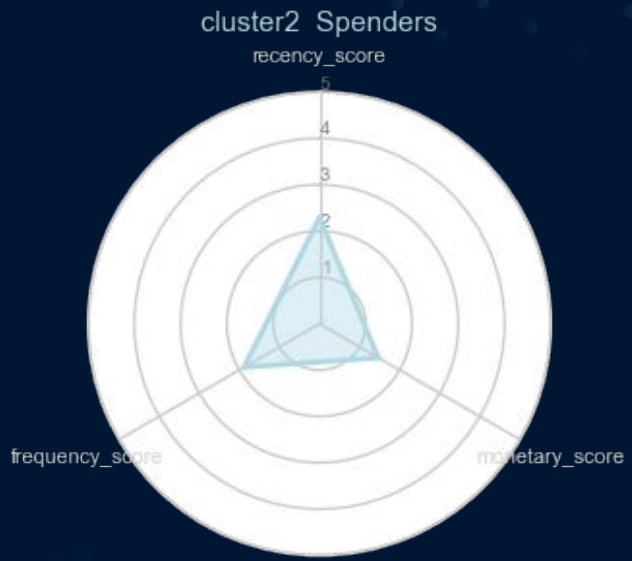
Homogénéité

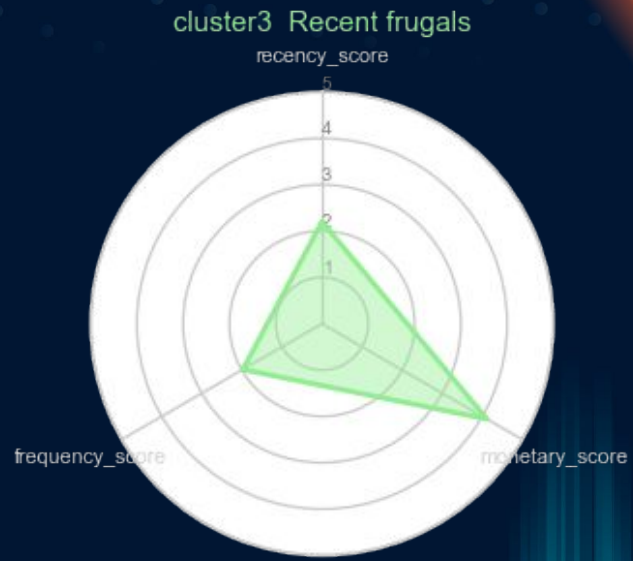


Profil type

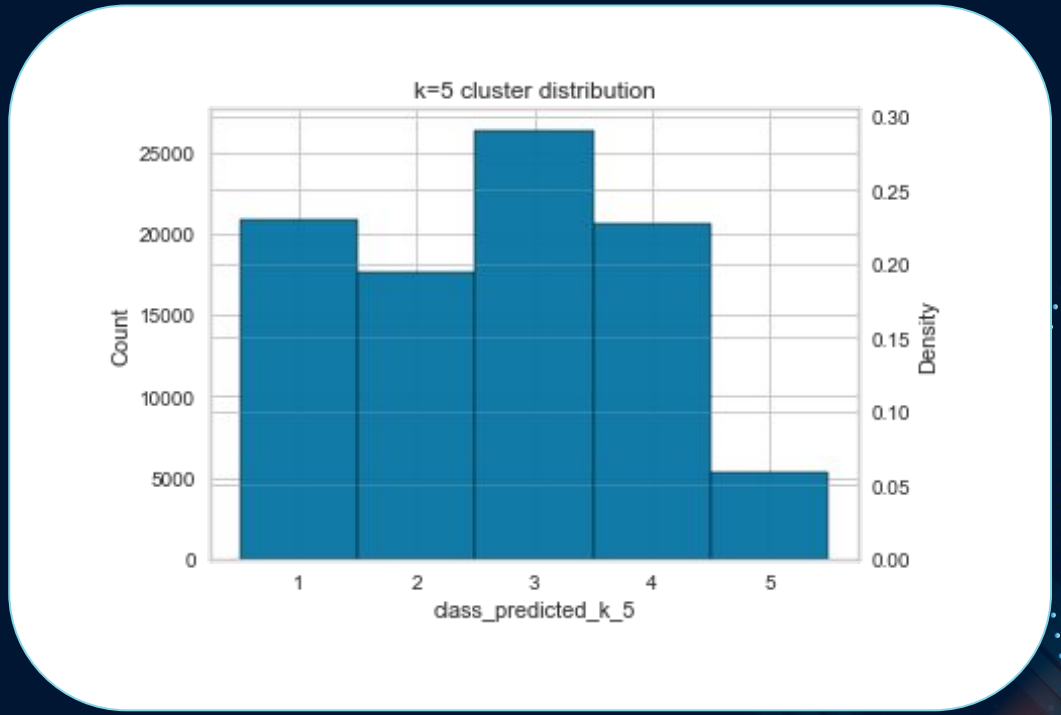
Interprétabilité / actionnabilité



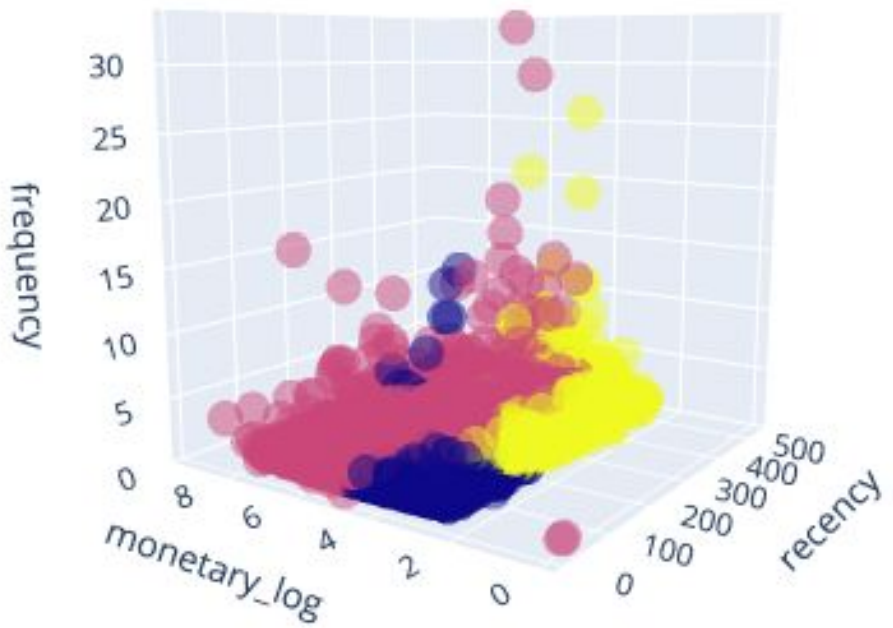




Distribution des effectifs



Visualisation 3D

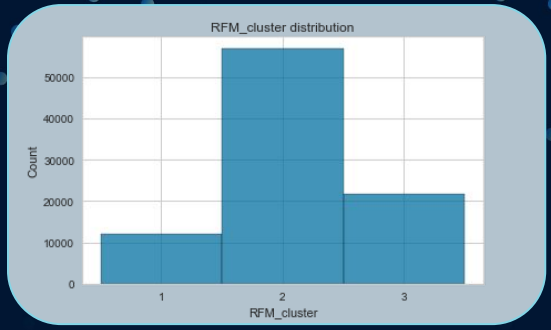


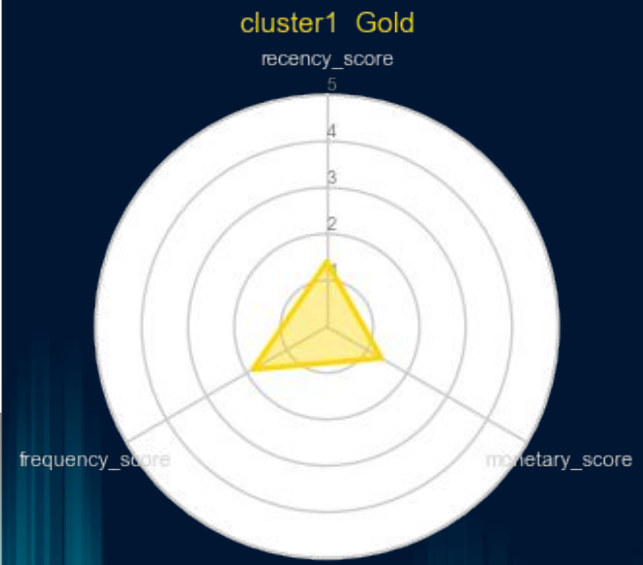

25.876,12
J J A S O N D J F M A M
vom Vortag, Veränderung in Prozent zum Vortag
2874,4 -0,07%
16.431,0 -0,08%
2882,3 -2,48%
9659,1 -0,02%
3140,1 -0,21%
-0,51%
USA
China

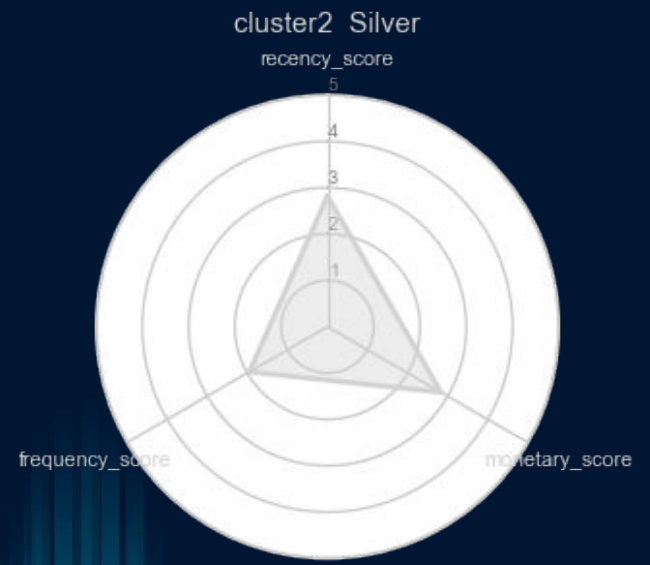
Résultats

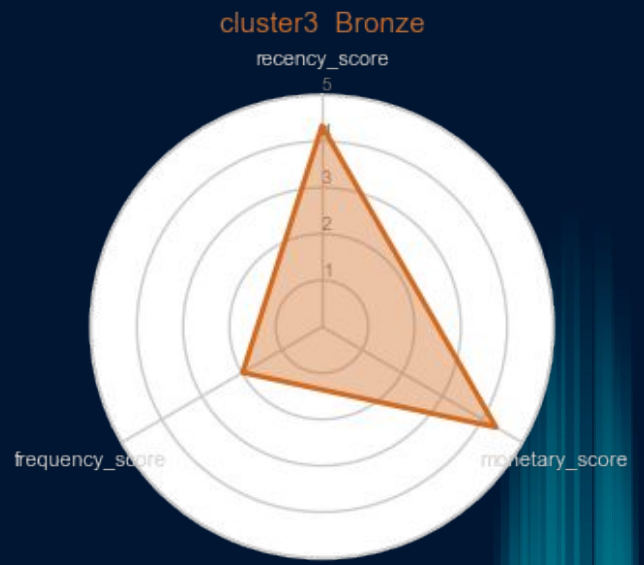
Résultats RFM

S = 0.21



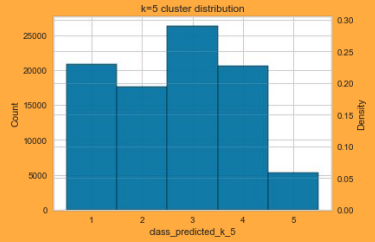


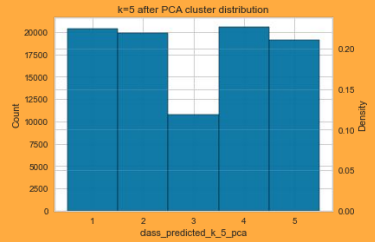


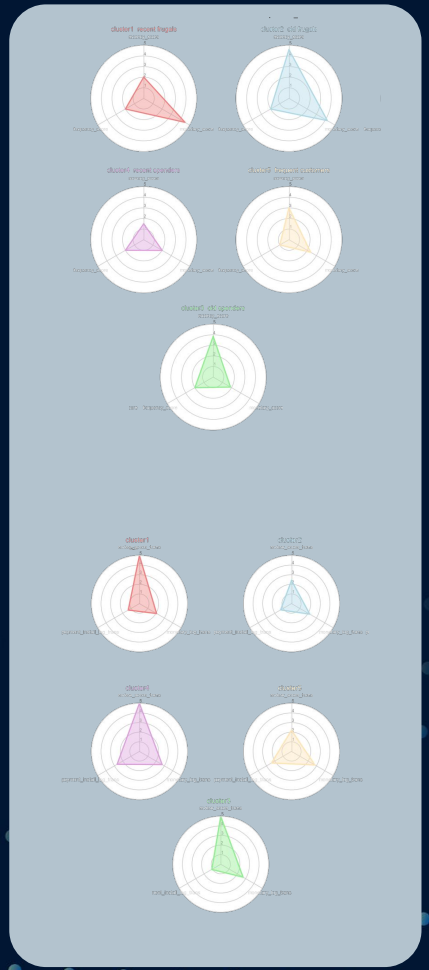


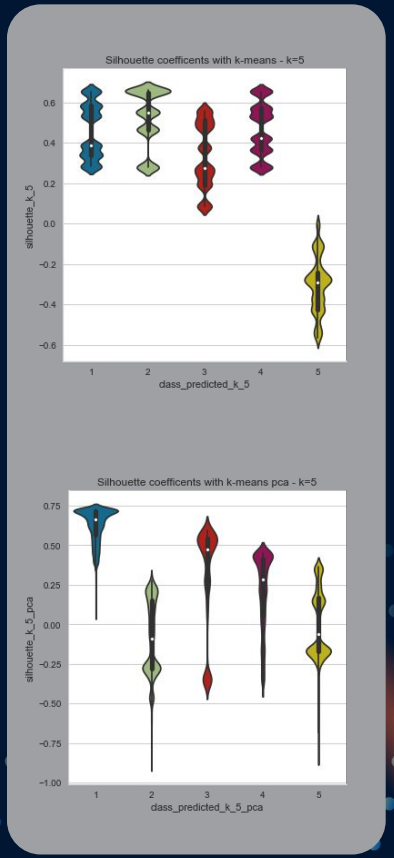
Choix du modèle

	silhouette coefficient
Manual RFM	0.207953
kmeans k=3 SS	0.268259
kmeans k=3 QT	0.382697
kmeans k=5	0.394105
kmeans k=5 pca	0.422509
kmeans k=5 reviews	0.226595





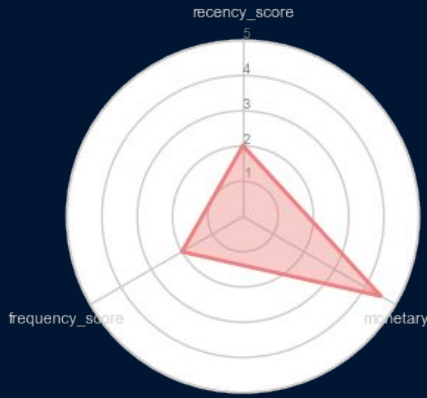




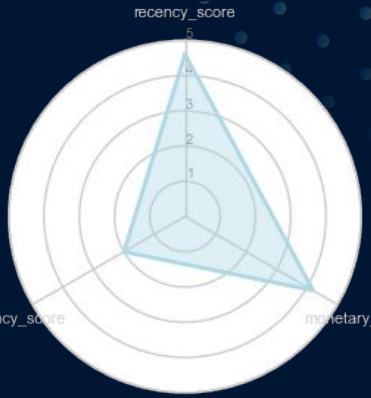
Résultats K-means k = 5

S = 0.39

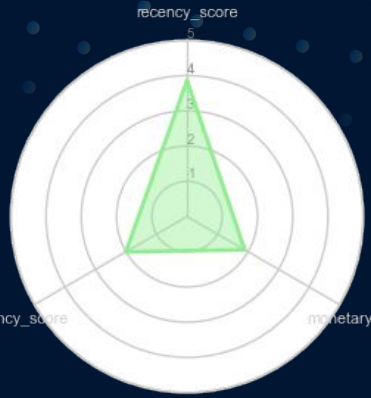
Economes récents



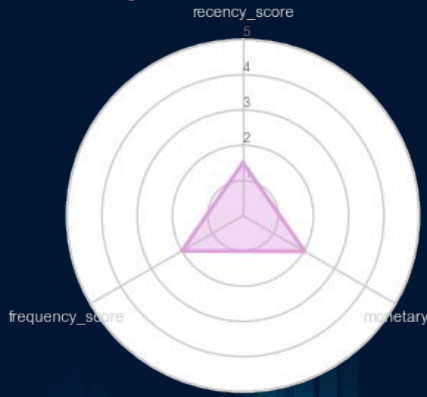
Economes anciens



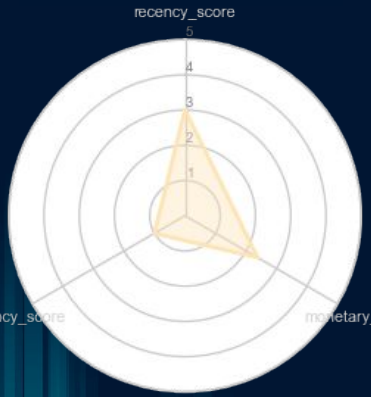
Dépensiers anciens

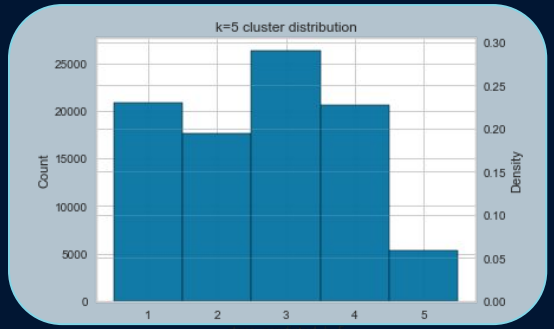


Dépensiers récents

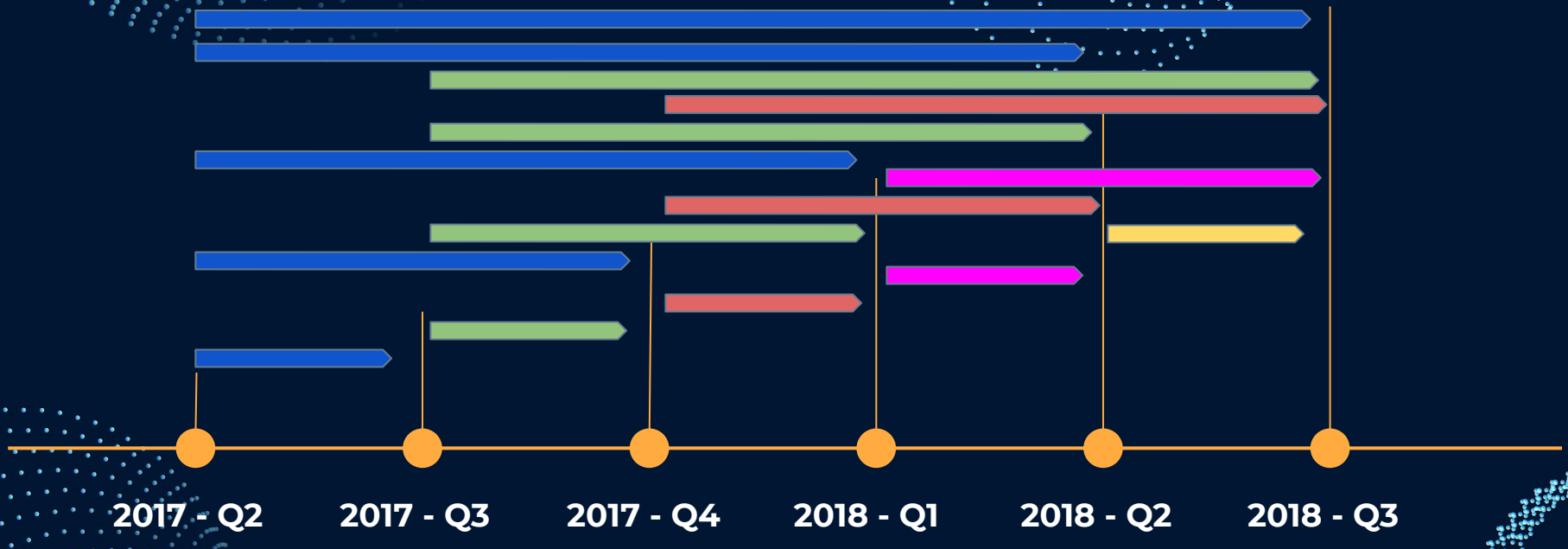


Fidèles



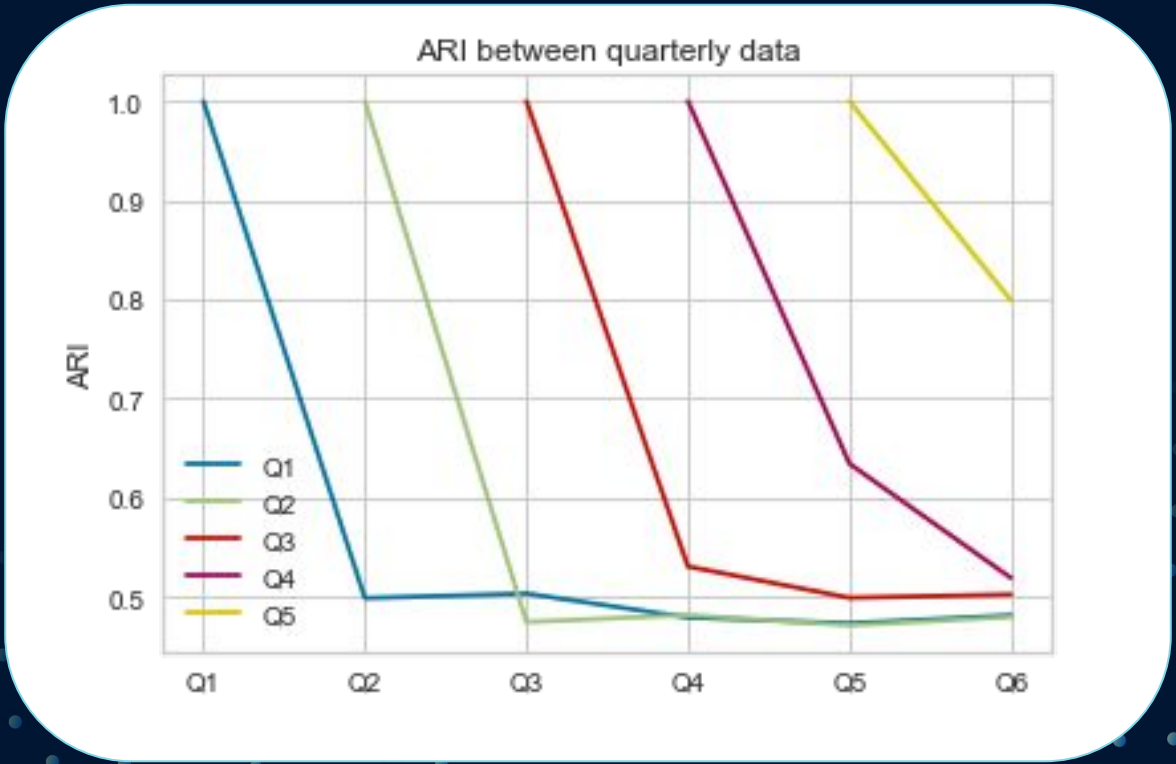


Stabilité dans le temps


2017 - Q2
2017 - Q3
2017 - Q4
2018 - Q1
2018 - Q2
2018 - Q3

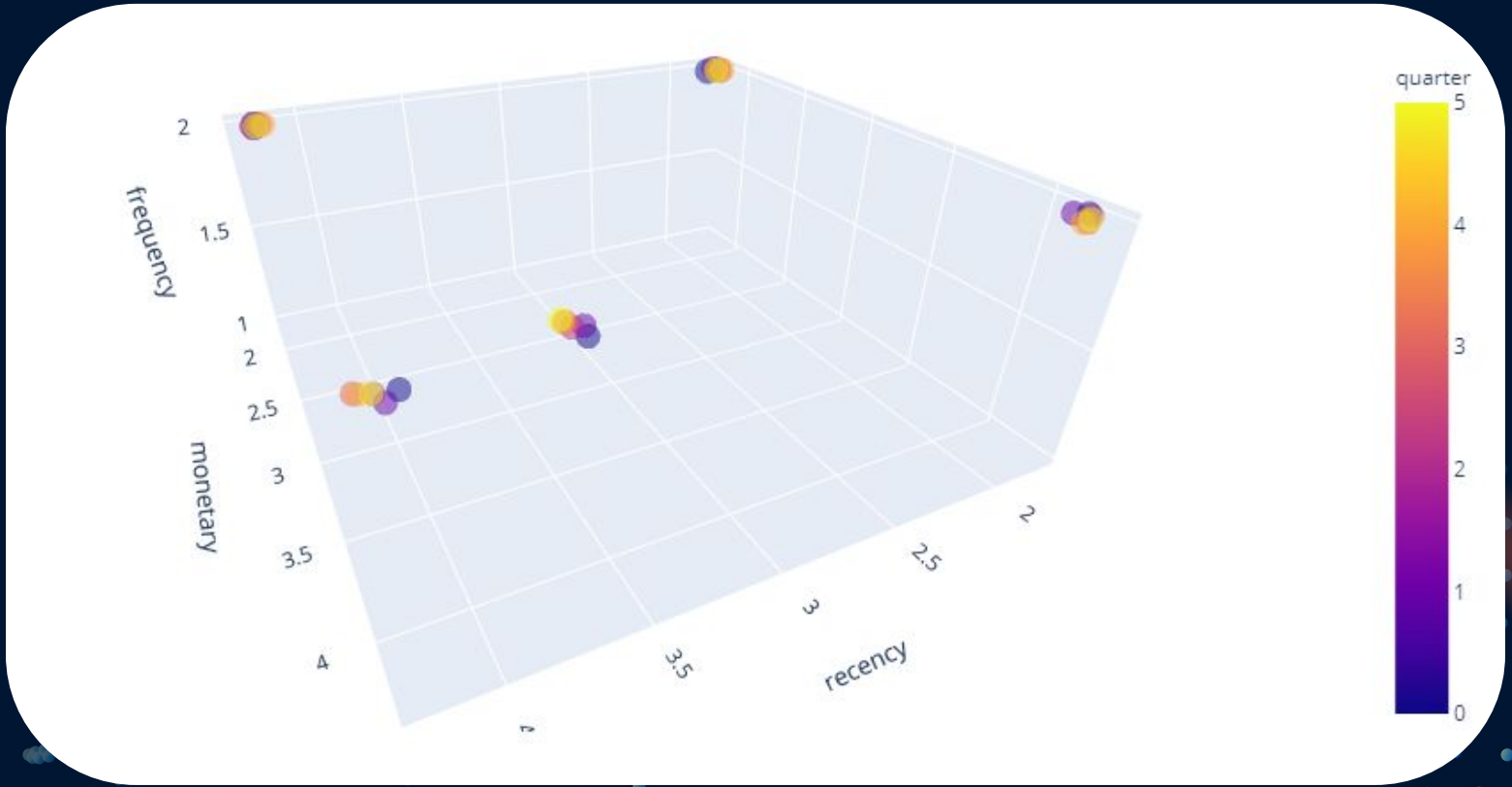
Stabilité des clusters

Indice de Rand



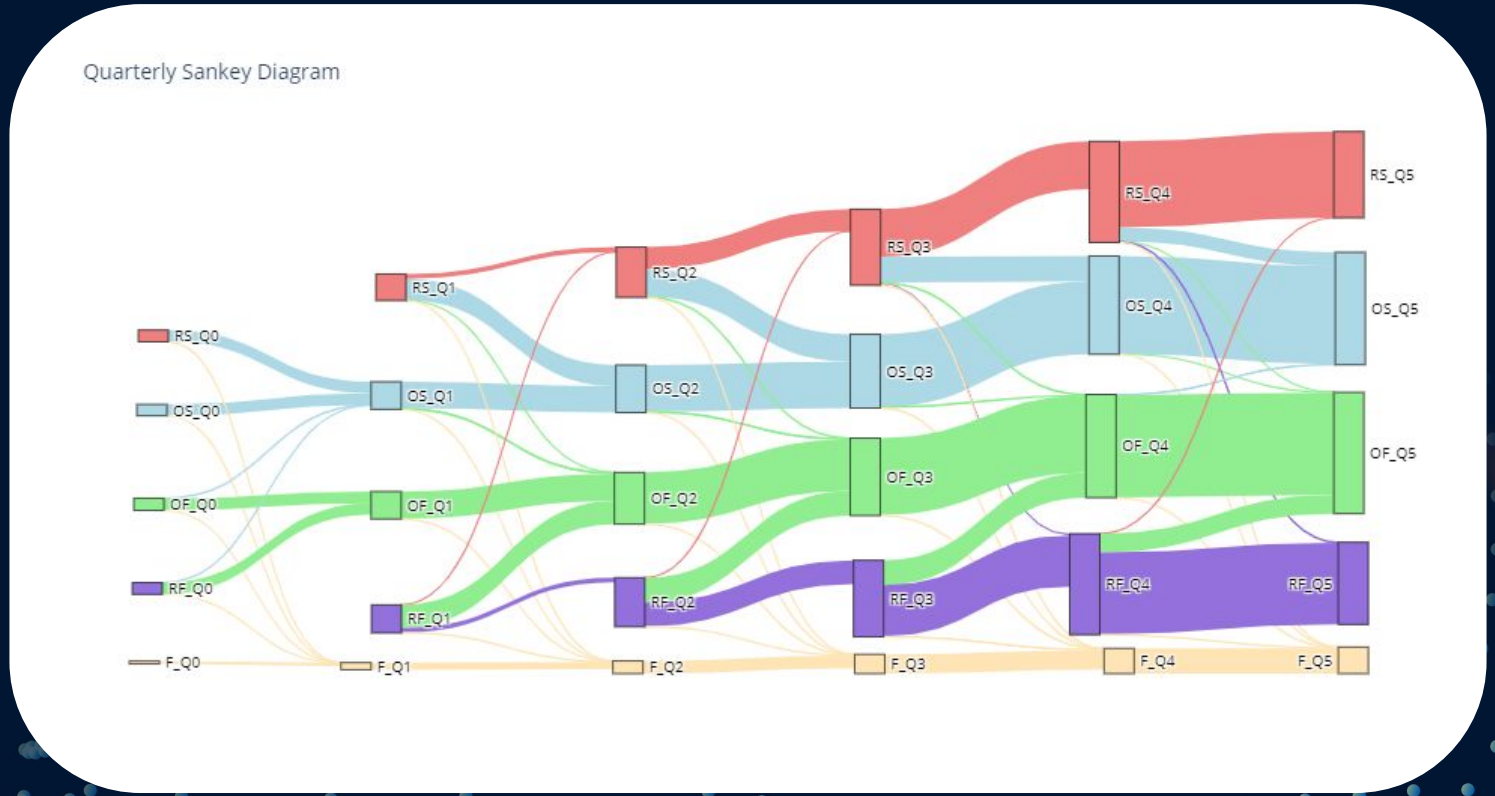
Stabilité des clusters

Mobilité des centroids



Stabilité des clusters

Mobilité des individus semestrielle



Conclusion

Retour sur objectifs

Comprendre la clientèle

Segmentation actionnable

Identifier maintenance

PEP - 8

Suggestions

Données riches pour l’analyse business:

Analyse de groupes type jour de semaine

Analyse des clients des différents états

Amélioration des capacités de segmentation ML

Continuer la collecte de données

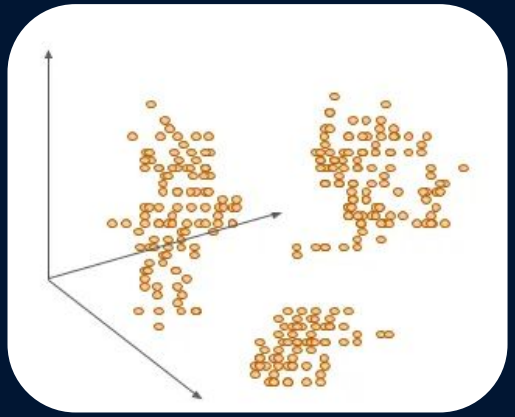
Analyse cognitive des avis

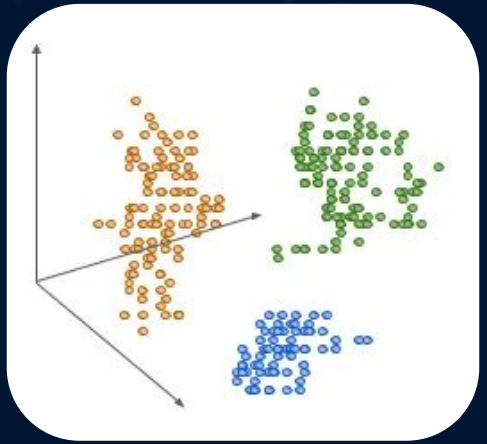
Données concurrentielles

Questions

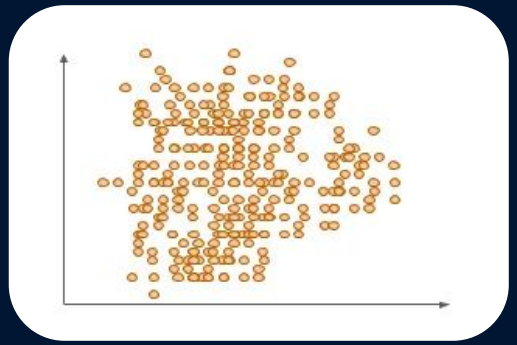
Annexes

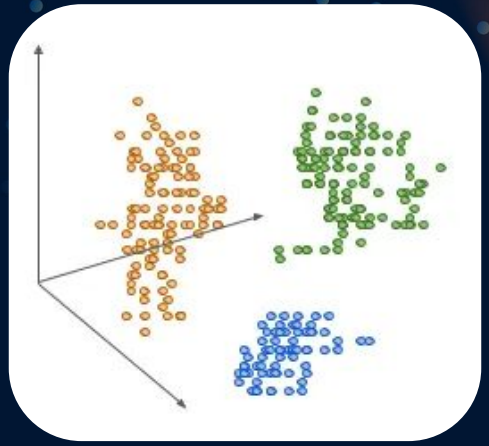
ACP



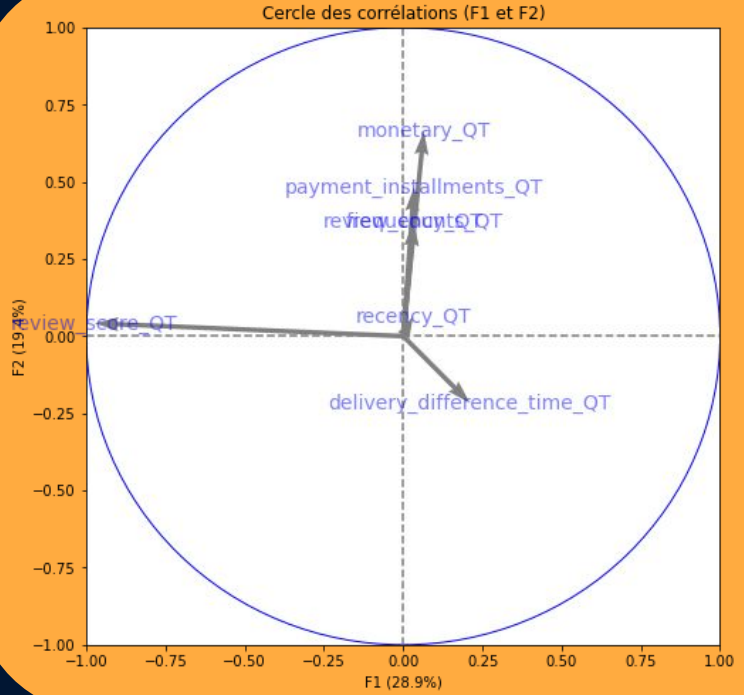


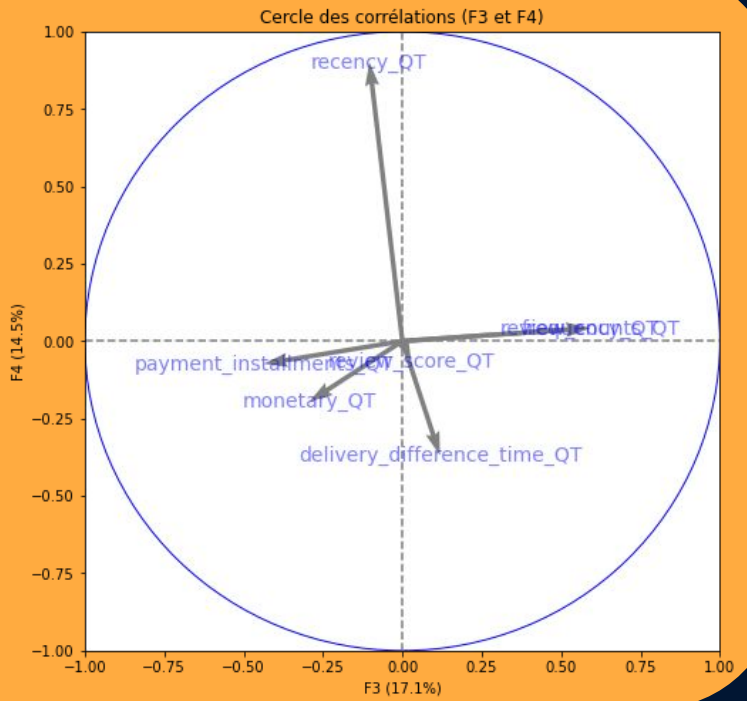
ACP





ACP





Variables représentant au mieux les données

Kmeans et ACP : nombre de variables

Indice de Davies-Bouldin

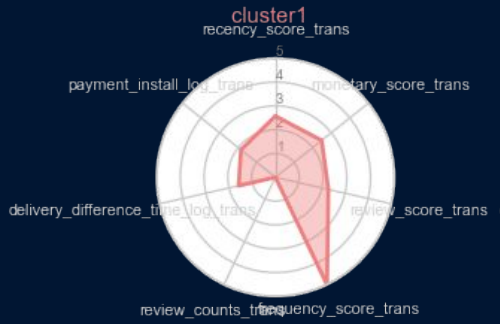
nb_features	k=2	k=3	k=4	k=5	k=6
2	0.679326	0.503614	0.626078	0.602093	0.725103
3	0.968973	0.871085	0.983333	0.844148	0.913325
4	1.203367	1.269512	1.409783	1.274790	1.252891
5	1.293888	1.389311	1.208515	1.317374	1.219004
6	1.353425	1.052996	1.140758	1.268431	1.179592
7	1.353425	1.052996	1.140322	1.268402	1.179235

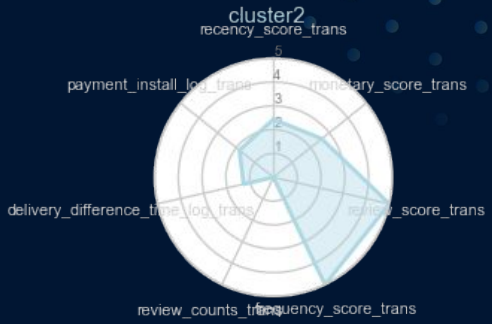
Pas de résultats pertinents

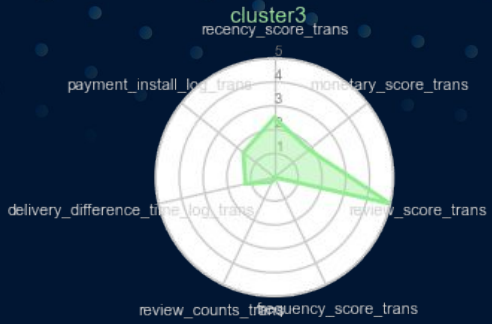
DBscan

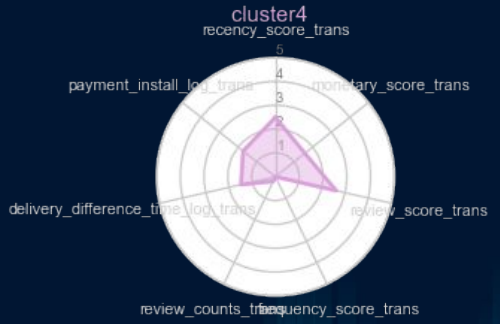
S = 0.28

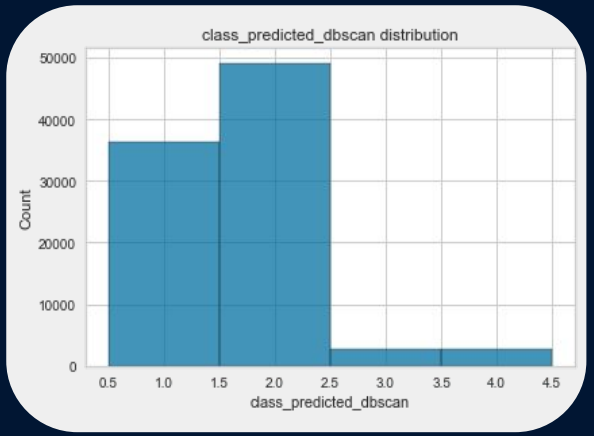
Cluster profile DBscan



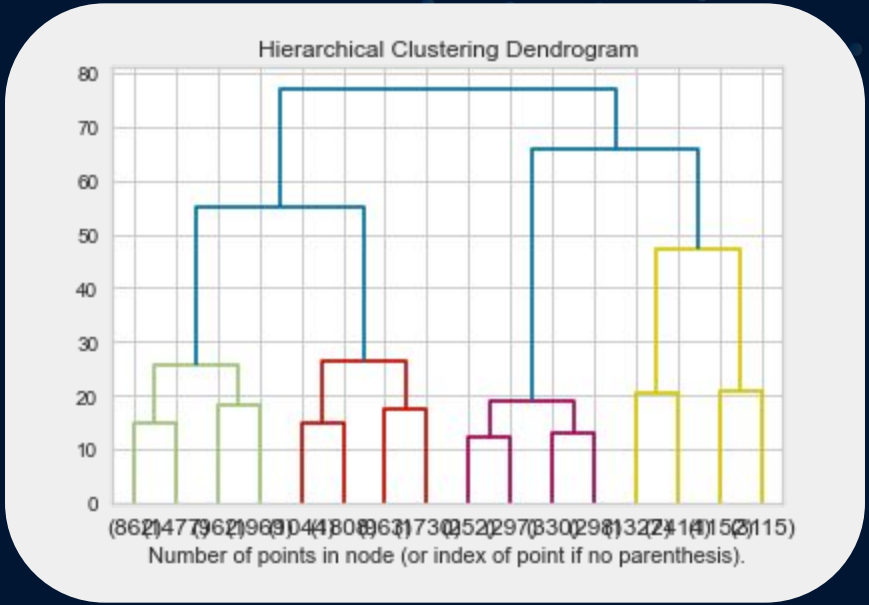








CAH


Hierarchical Clustering Dendrogram
80
70
60
50
40
30
20
10
0
Number of points in node (or index of point if no parenthesis).

CAH

Cluster profile CAH