Segmentation des clients d'un site e-commerce

Campagne de communication - Olist

Besoins

- Classification non-supervisée
- Segmentation exploitable, facile d’utilisation
- Contrat de maintenance
- Suivre la norme PEP8

Problématique

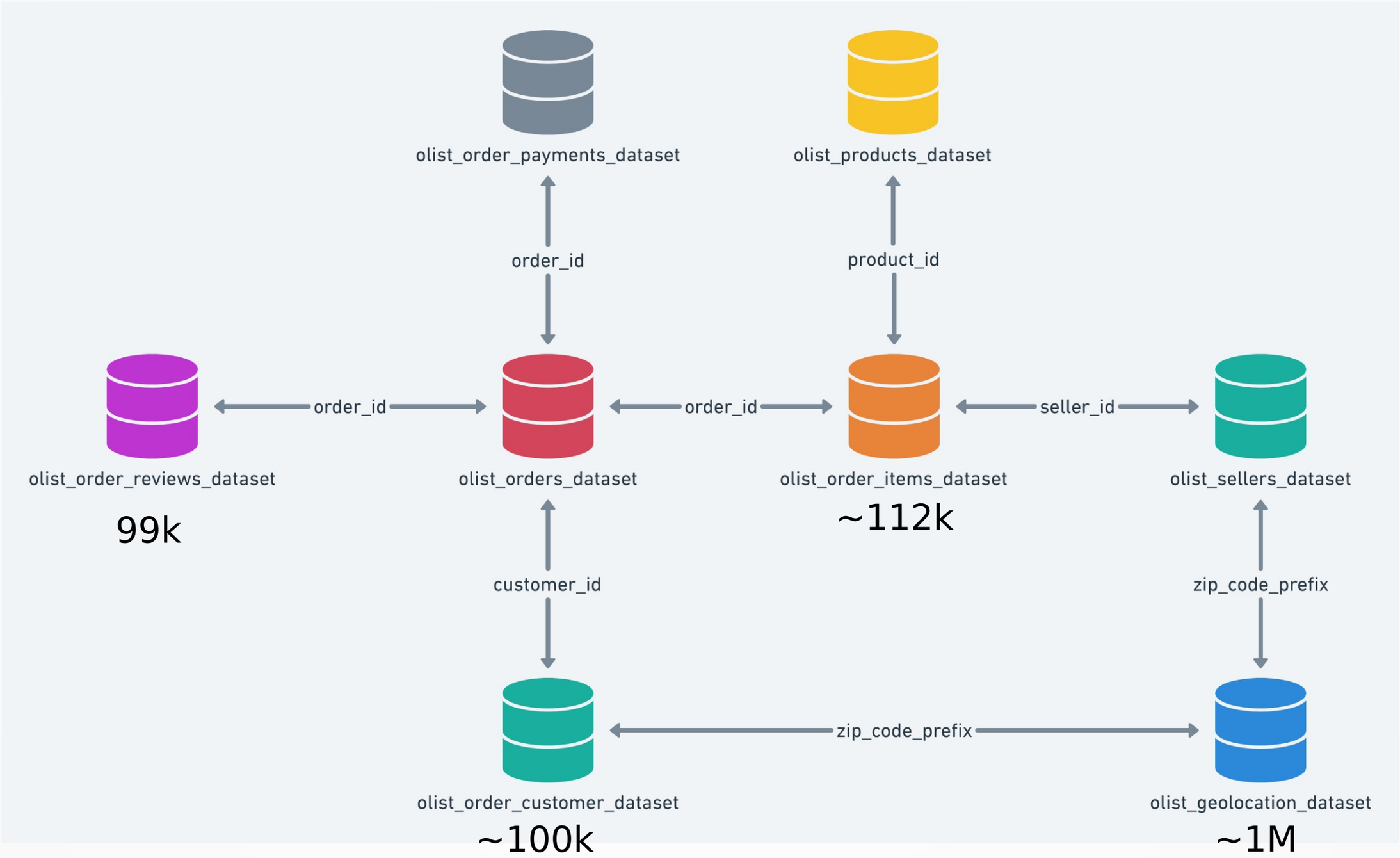
- Quelles sont les variables pertinentes ? Faut-il les transformer ?
- Quel algorithme de classification sera le plus efficace ?
- Comment évaluer la qualité de la classification ?
- Comment évaluer sa stabilité ?
- Quel outil pour formater le code ?

Réponse

- Essayer différentes entrées et visualiser les données
- Tester plusieurs classificateurs et les comparer
- Utiliser l’indice de silhouette et d’autres indices
- Vérifier la stabilité :
  - Sur plusieurs itérations
  - Sur l’ajout de données
- Utiliser l’extension auto-pep8

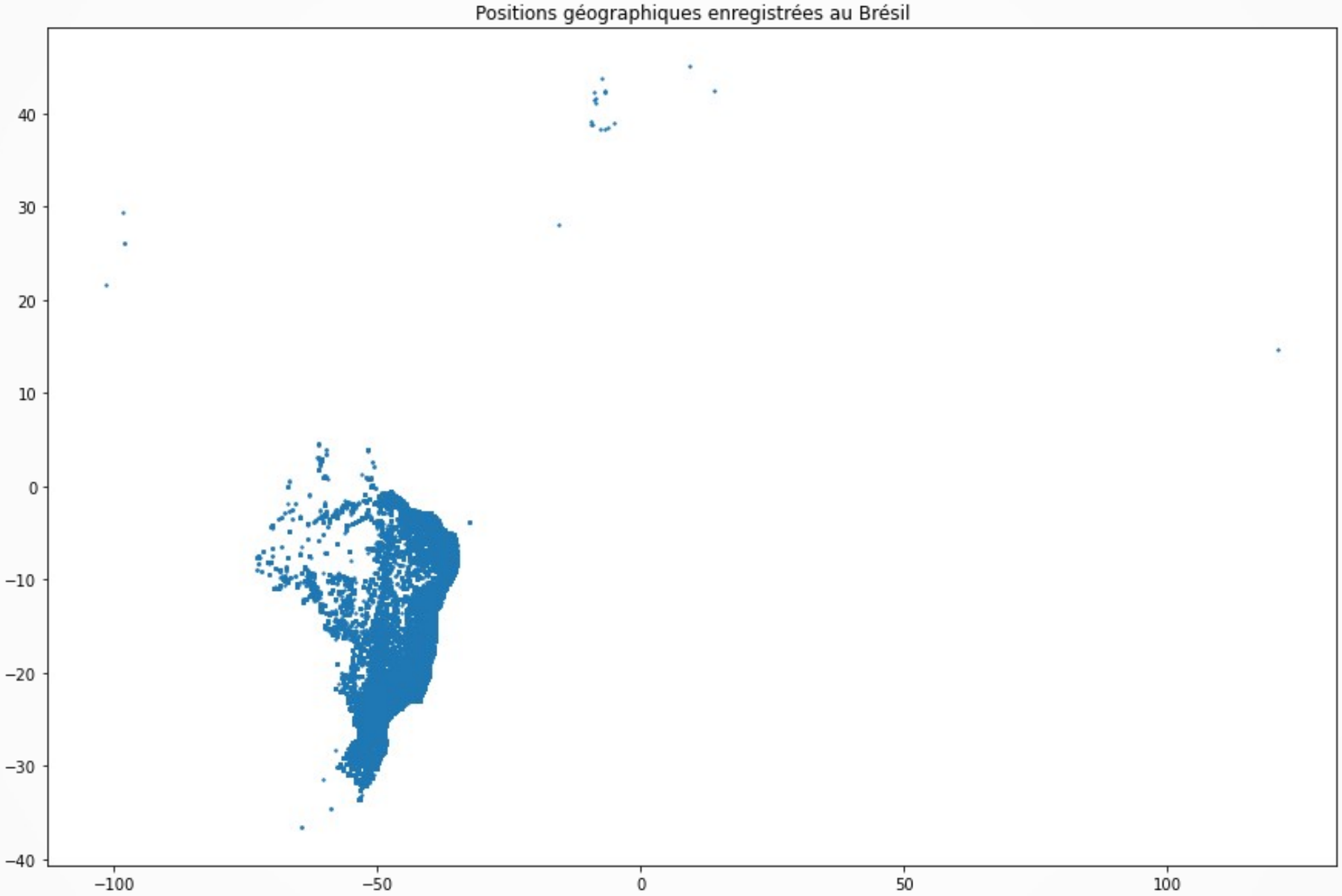
I- La base de données

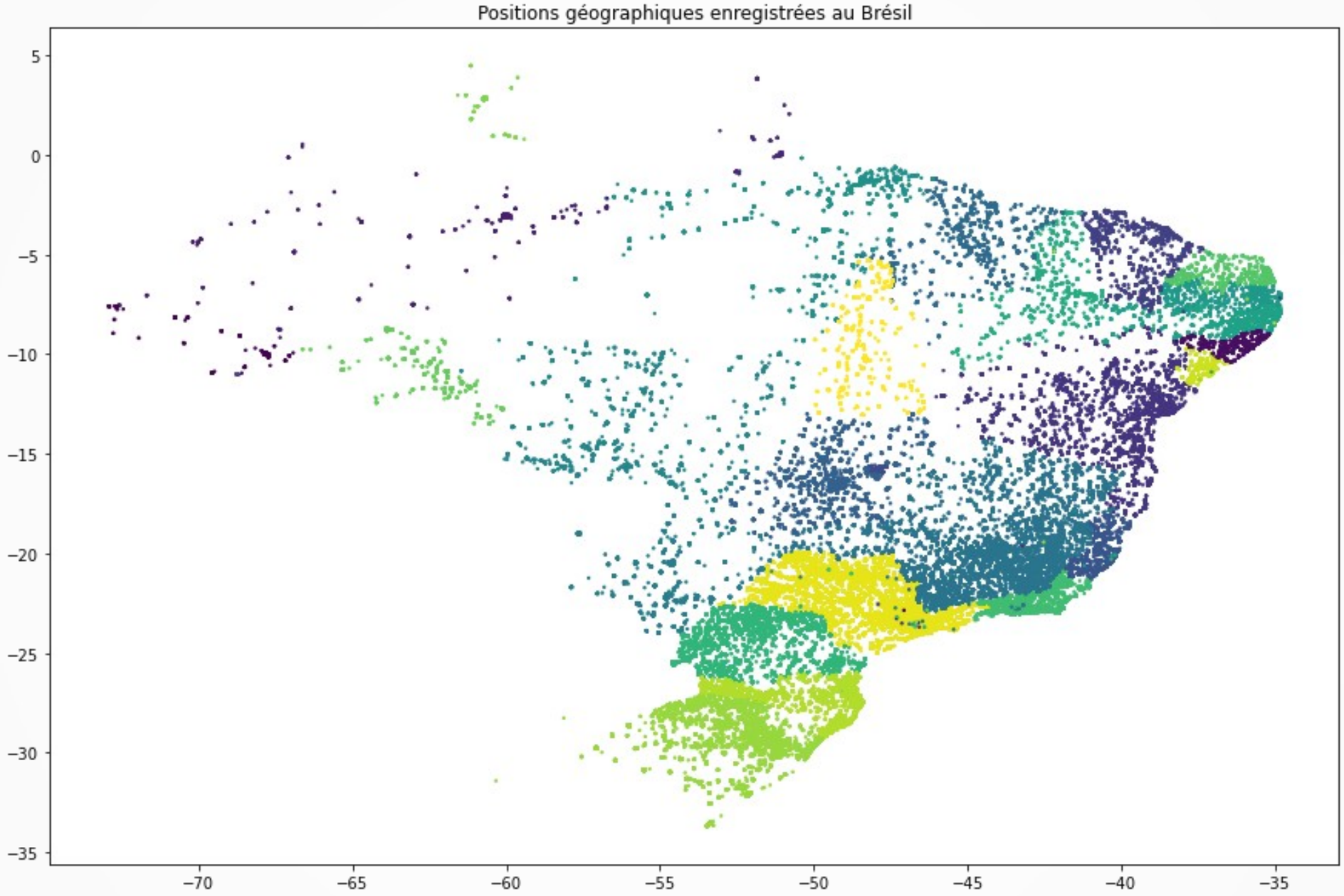
- Fournie par le site d’e-commerce Olist
- Anonymisée
- Contient plusieurs tables, liées par des clés étrangères
  - Clients
  - Géolocalisations
  - Produits
  - Commandes
  - Items des commandes
  - Paiements


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
99k
~112k
customer_id
zip_code_prefix
zip_code_prefix
olist_order_customer_dataset
olist_geolocation_dataset
~100k
~1M

II- Nettoyage – Valeurs aberrantes

- Suppression des doublons
- Suppression des valeurs extrêmes
- Supprimer les valeurs incohérentes
- Vérification de la cohérence entre les tables


Positions géographiques enregistrées au Brésil
40
30
20
10
0
−10
−20
−30
−40
−100
−50
0
50
100


Positions géographiques enregistrées au Brésil
5
0
−5
−10
−15
−20
−25
−30
−35
−70
−65
−60
−55
−50
−45
−40
−35

II- Nettoyage – Valeurs aberrantes

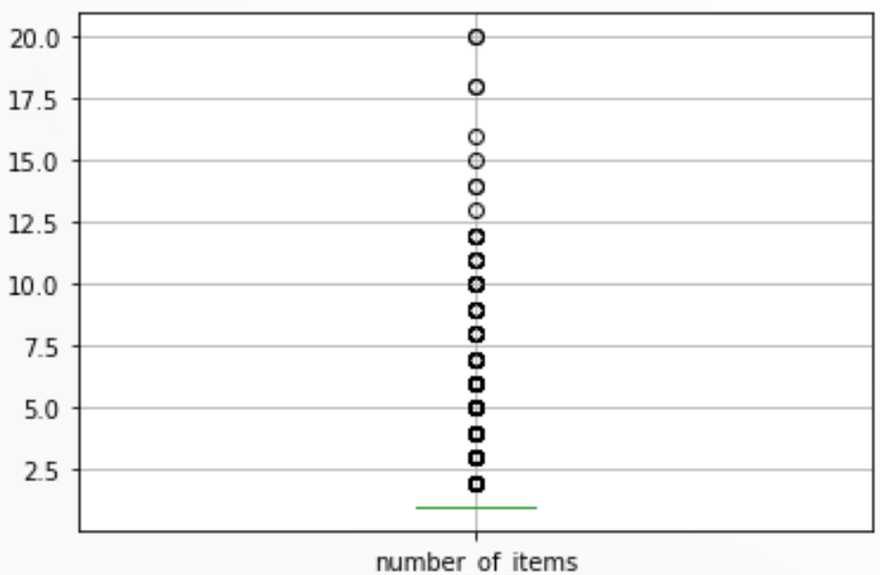
- Catégories : ‘cool_stuff’
- Catégories non-référencées
- Cohérence entre les prix et les paiements

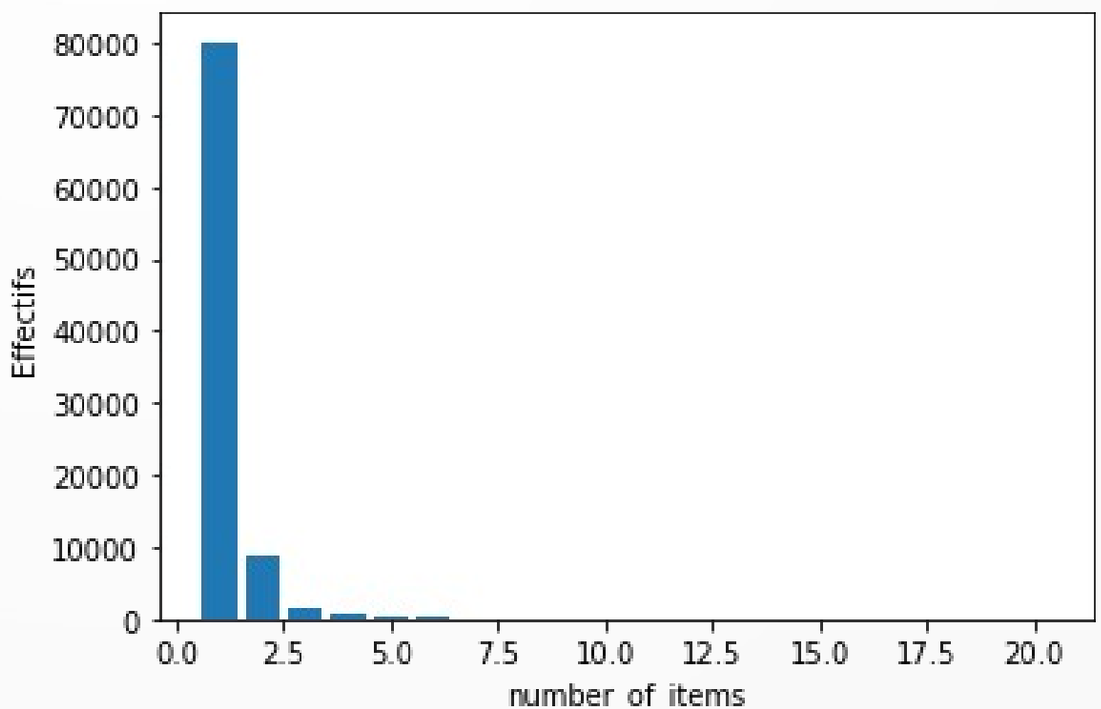
II- Nettoyage - Imputation

- Taux de remplissage minimal : 97 %
- Mis à part les commentaires avec les notes
- On ne fait pas d’imputation

III- Exploration

- Nombre d’items achetés





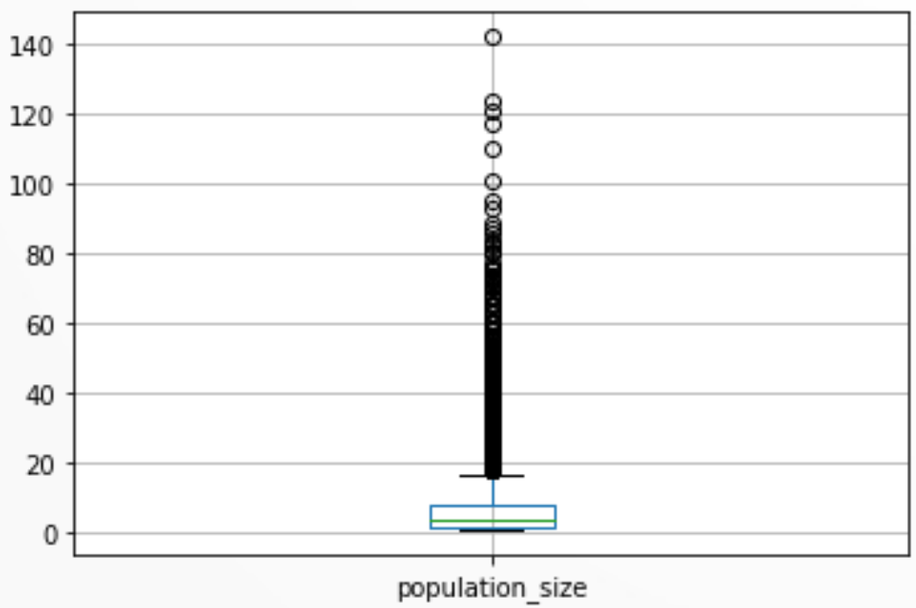
III- Exploration

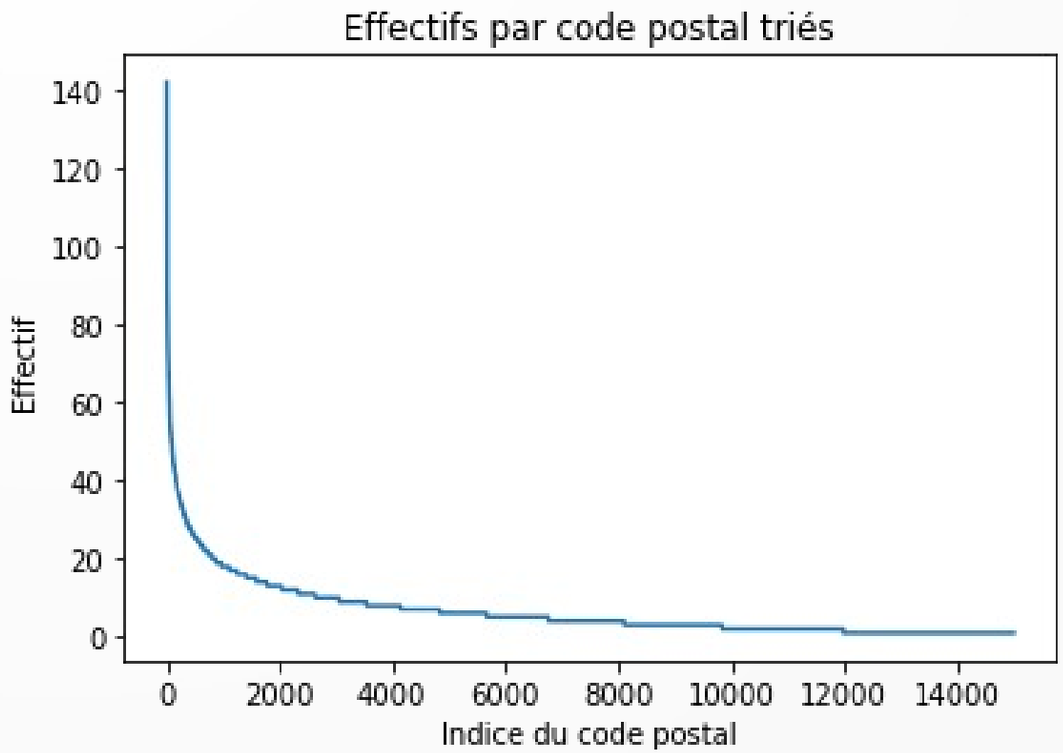
- Analyse factorielle non-négative
  - Coordonnées d’un axe

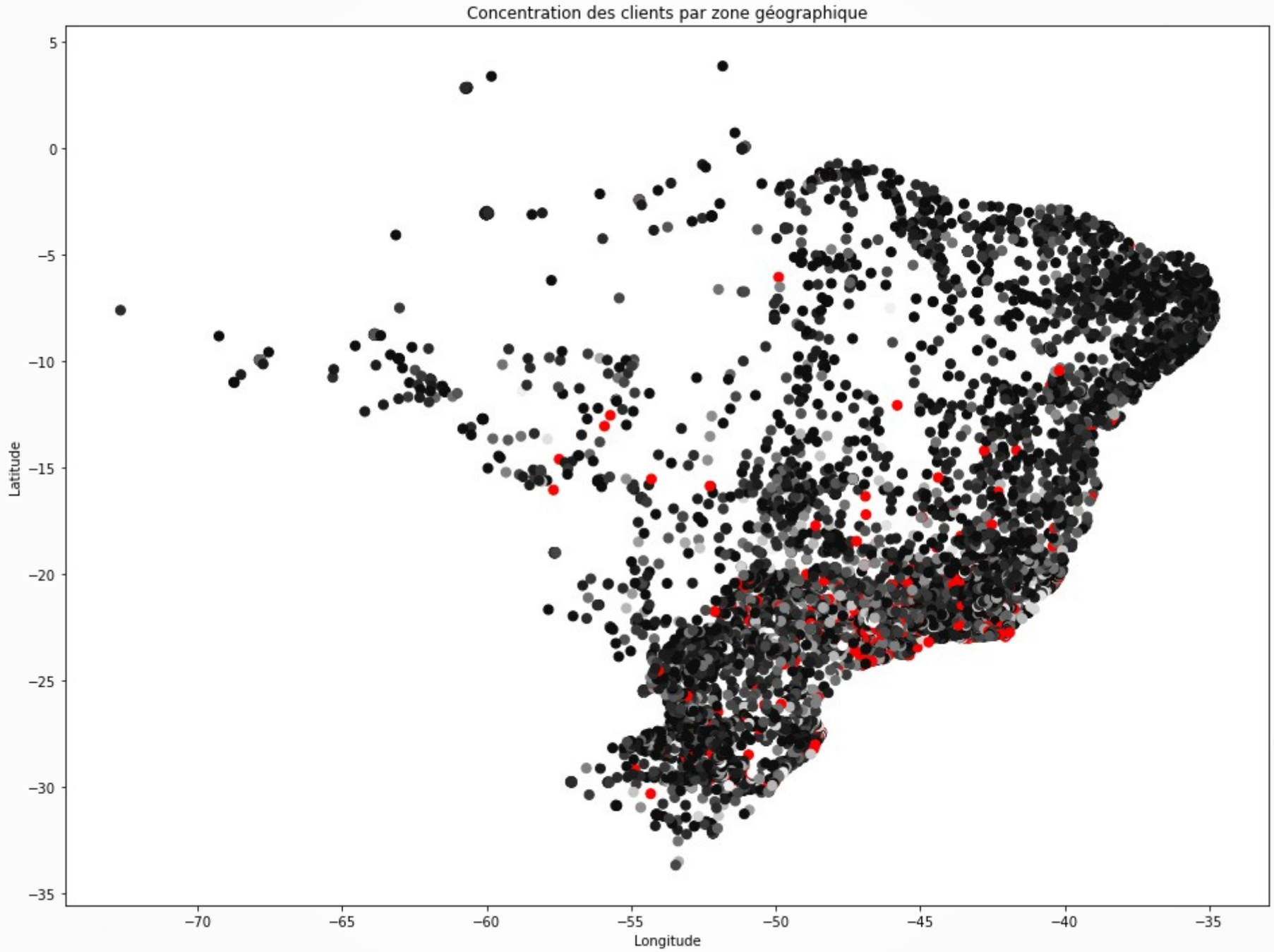
moveis_decoracao	10.176362
ferramentas_jardim	0.111993
beleza_saude	0.060263
relogios_presentes	0.035623
casa_construcao	0.034349

III- Exploration

- Effectifs par code postal





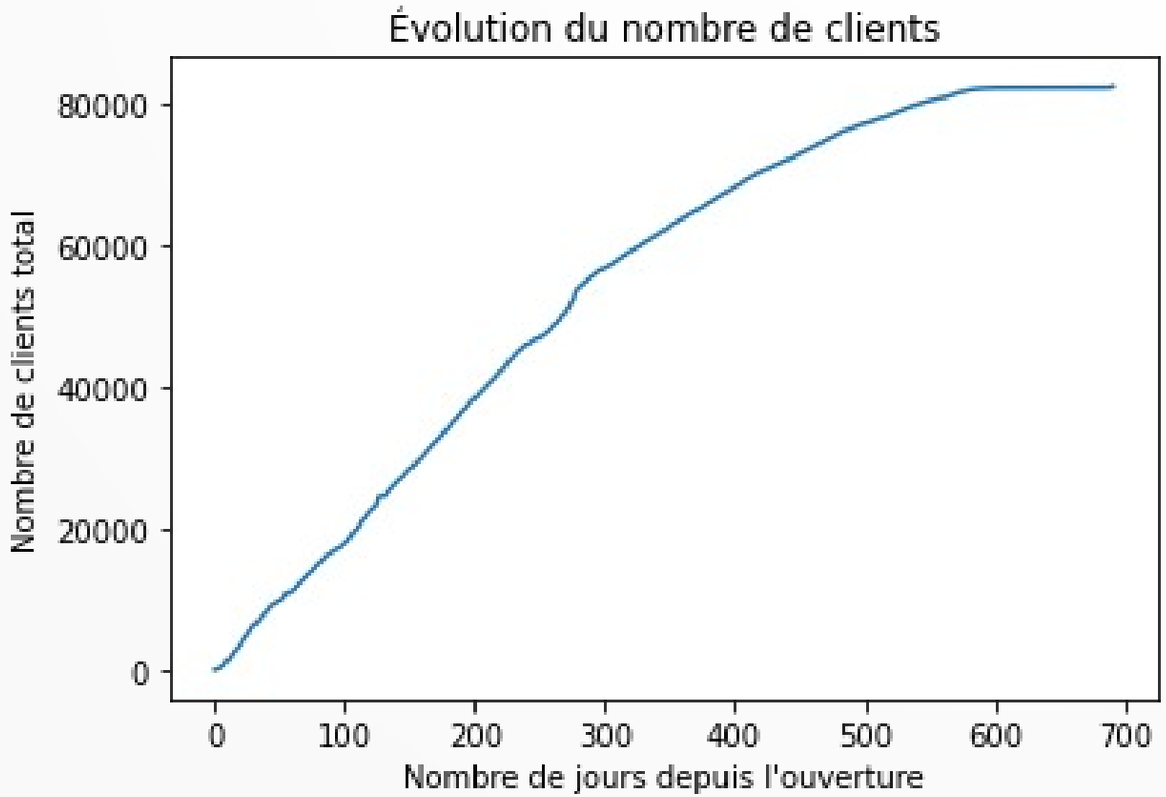

Concentration des clients par zone géographique
Latitude
Longitude
5
0
−5
−10
−15
−20
−25
−30
−35
−70
−65
−60
−55
−50
−45
−40
−35

IV- Construction des profils

- Variables prises en compte :
  - RFM
  - Dimensions moyennes des produits
  - Longueur de la description du produit
  - Notes moyennes données
  - Proximité à la ville

V- Segmentation

- Évolution du nombre de clients

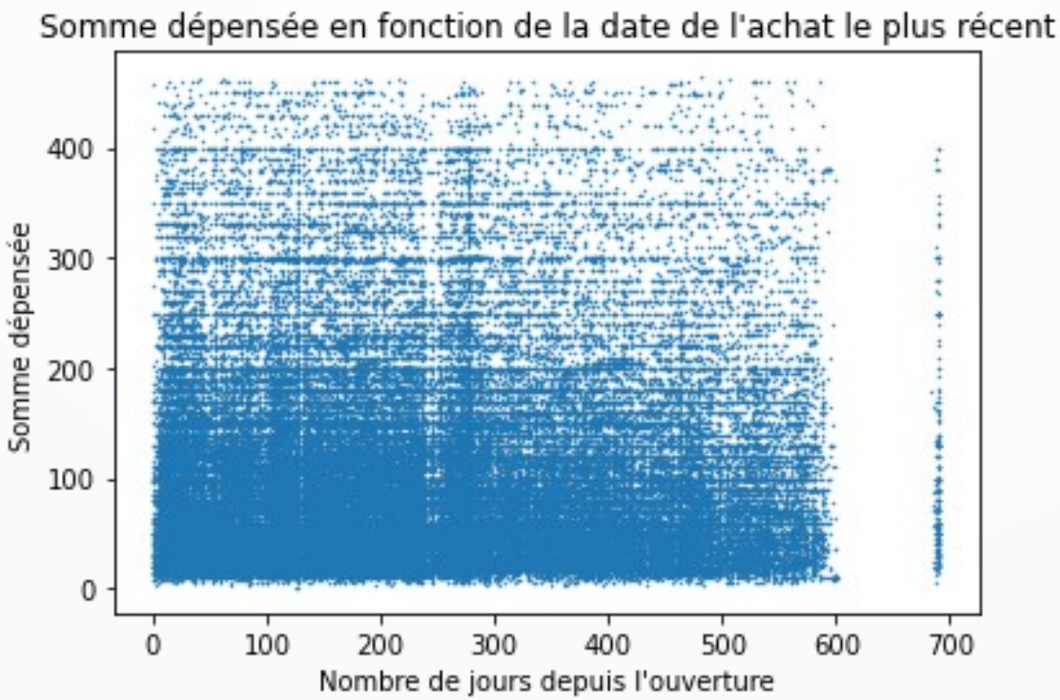


Pente : 143.43
R2 = 0.97

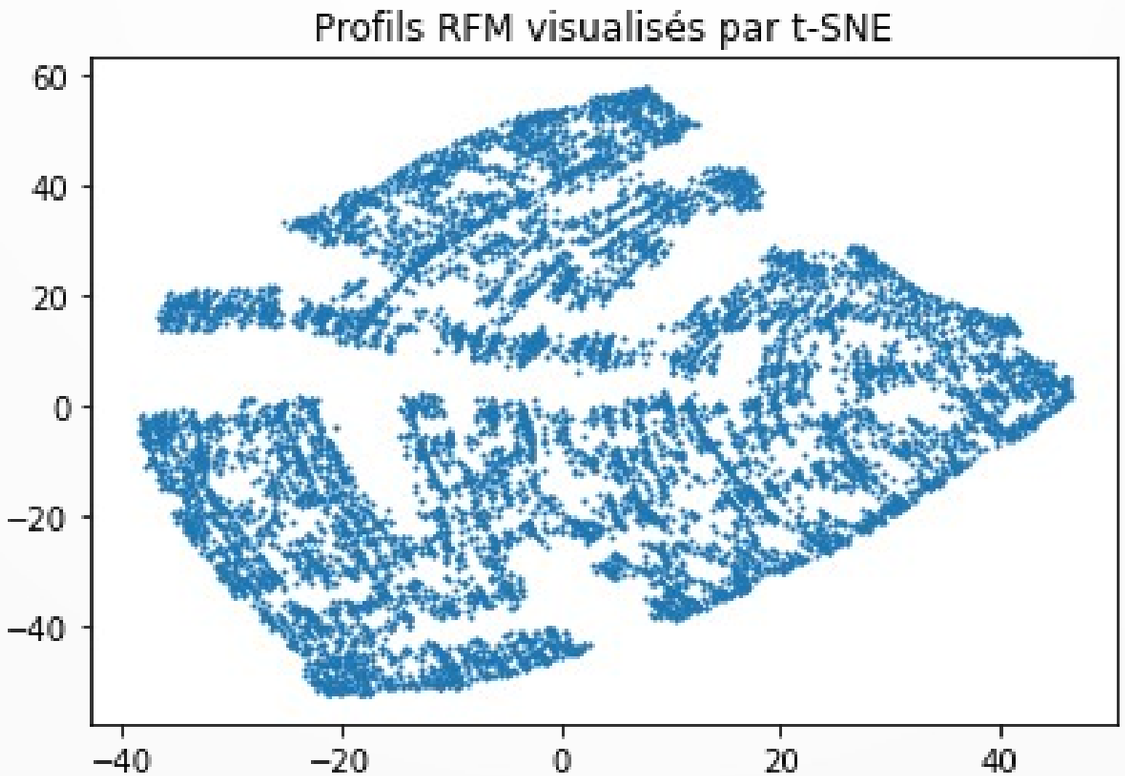
→ Échantillon : 73886

V- Classification RFM

- La fréquence est la même pour tous, on ne l’utilise pas

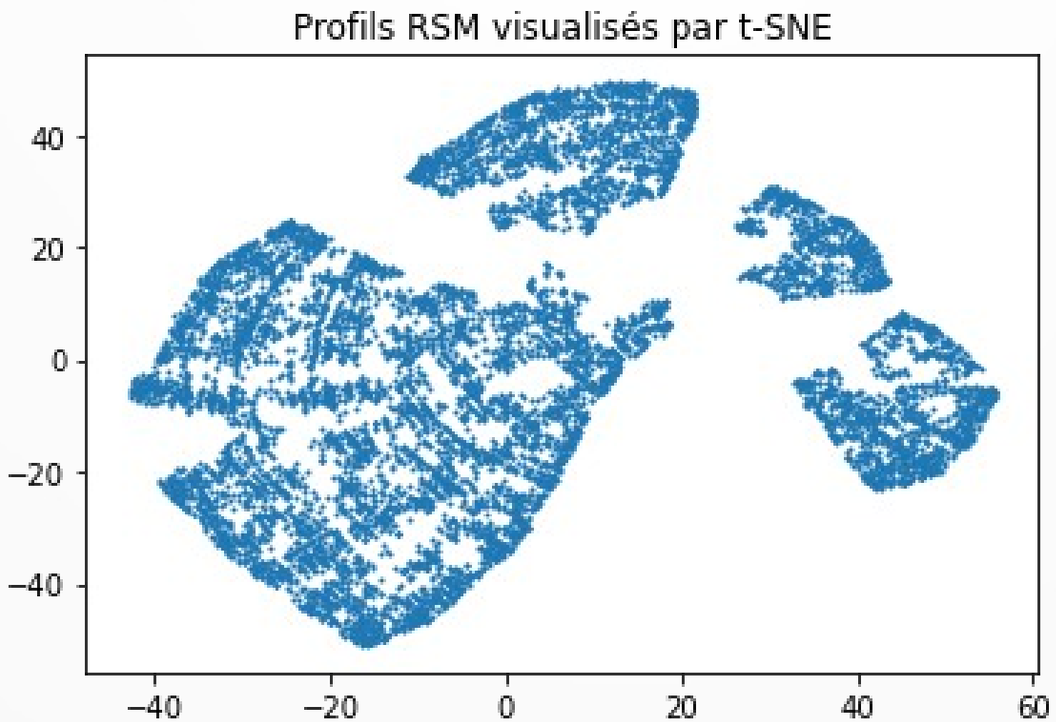


V- Classification RFM



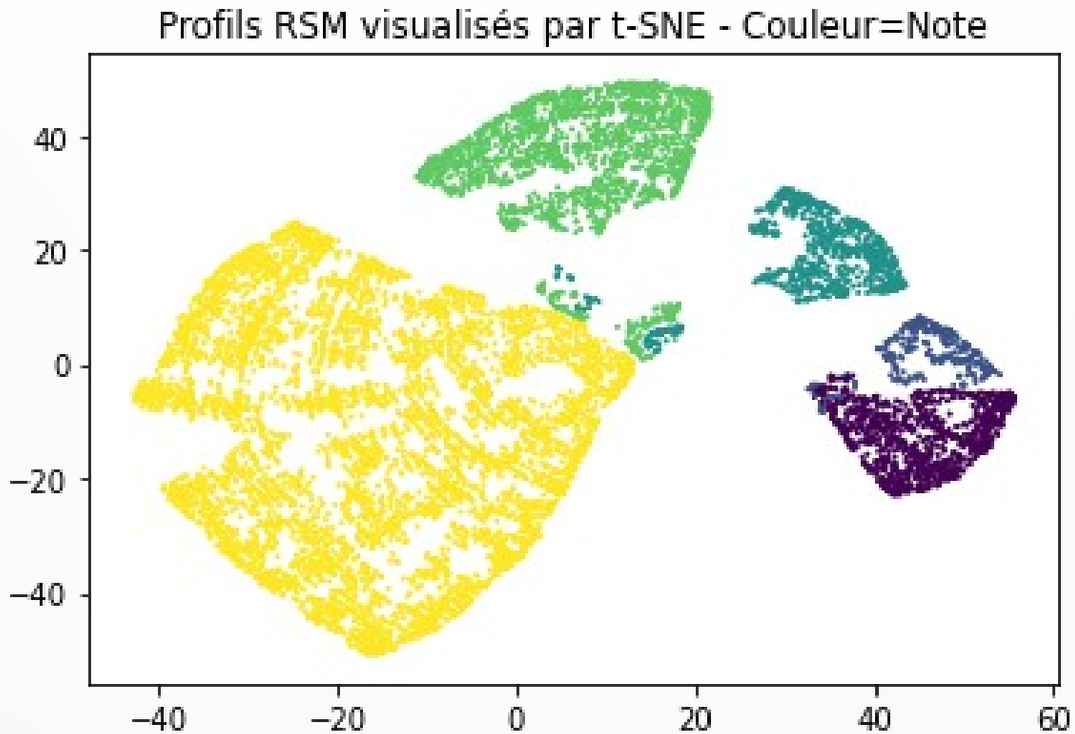
V- Classification RSM

- Récence, Score, Montant



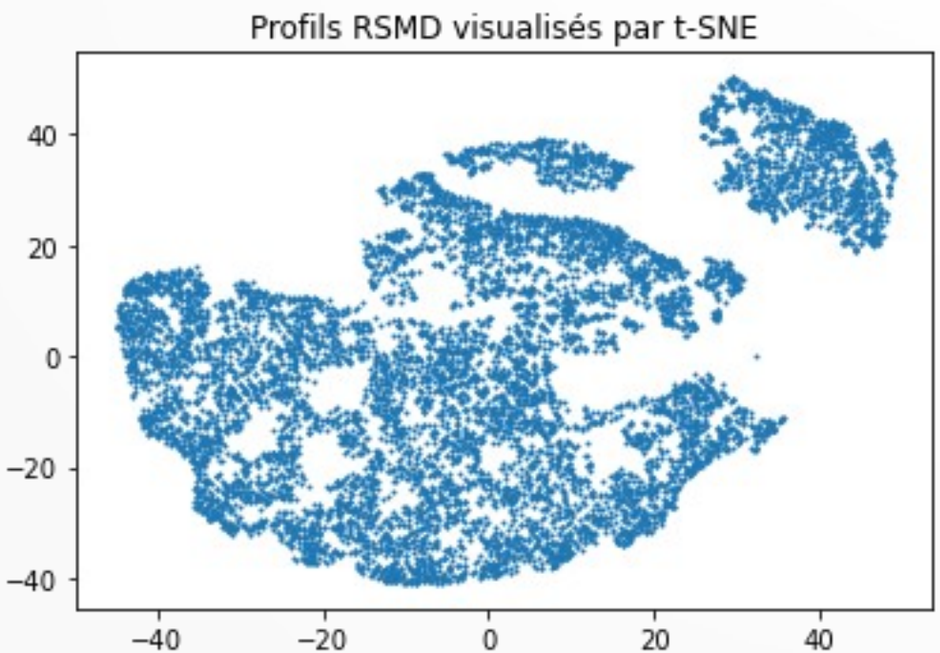
V- Classification RSM

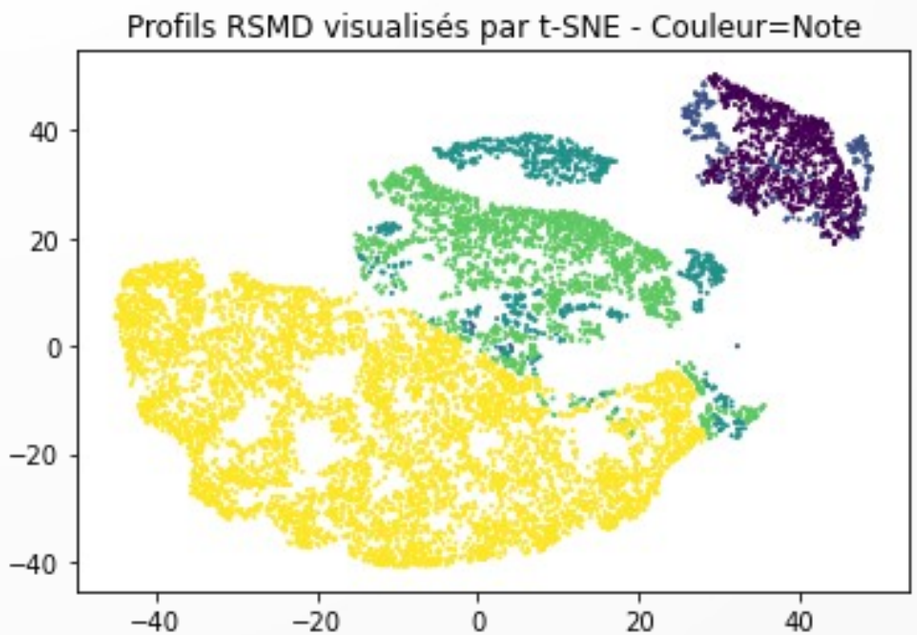
- Si on met les notes en couleur :



V- Classification RSMD

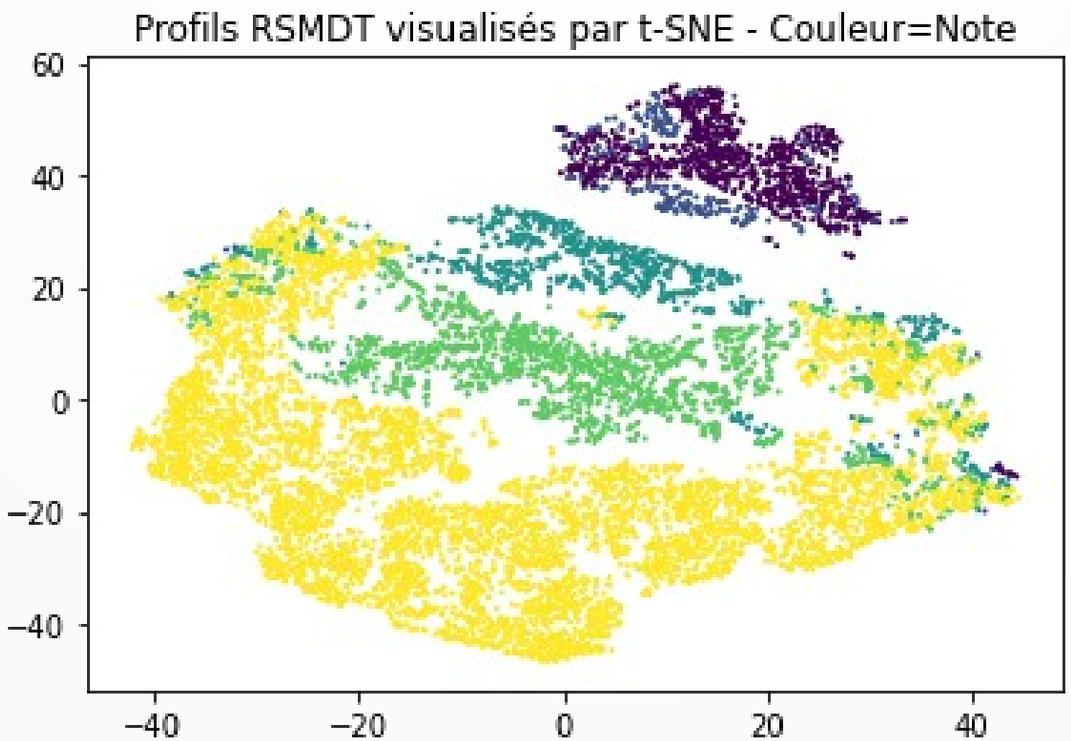
- Récence, Score, Montant, Distance



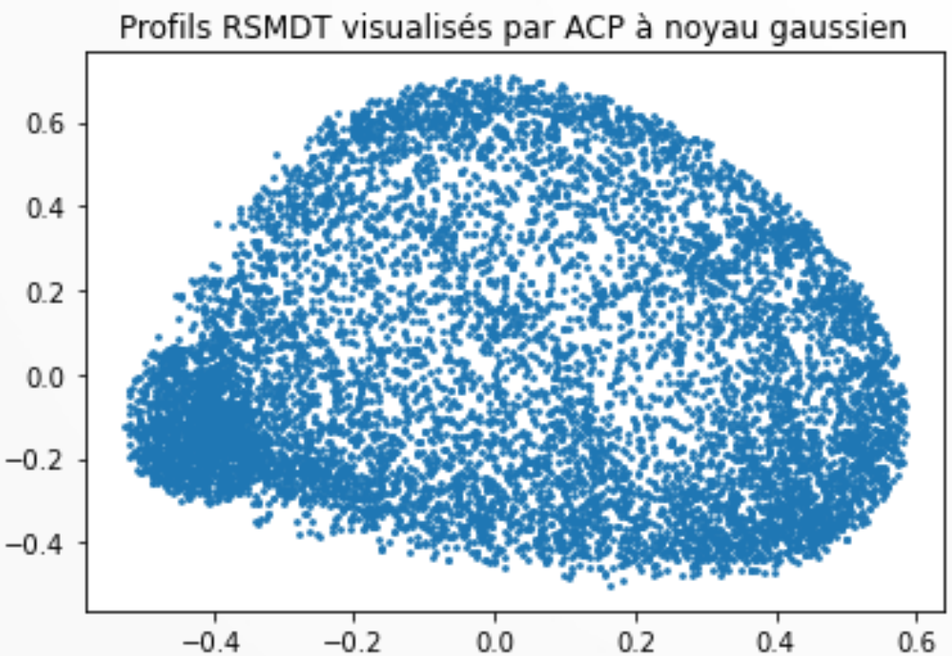


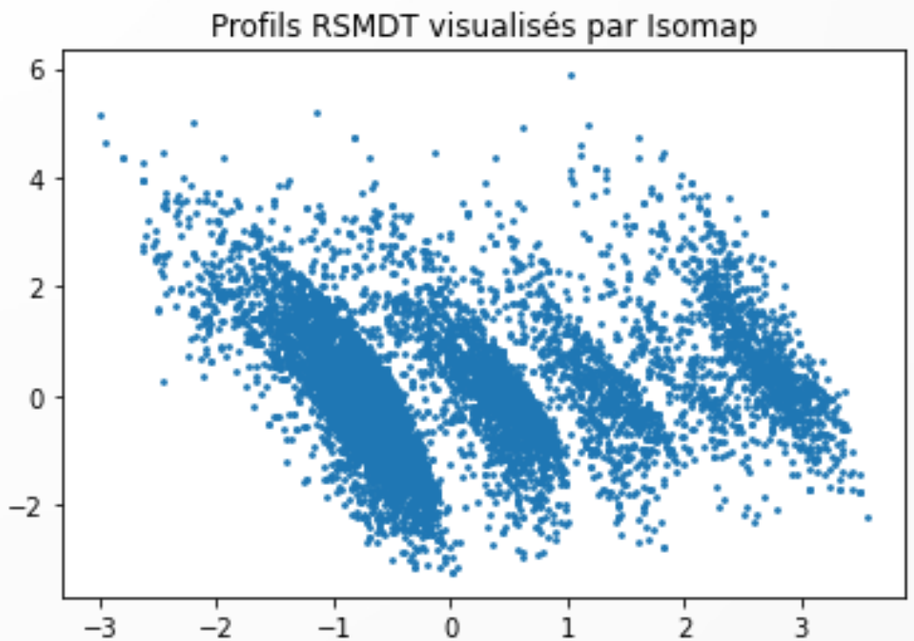
V- Classification RSMDT

- Récence, Score, Montant, Distance, Texte



V- Classification RSMDT





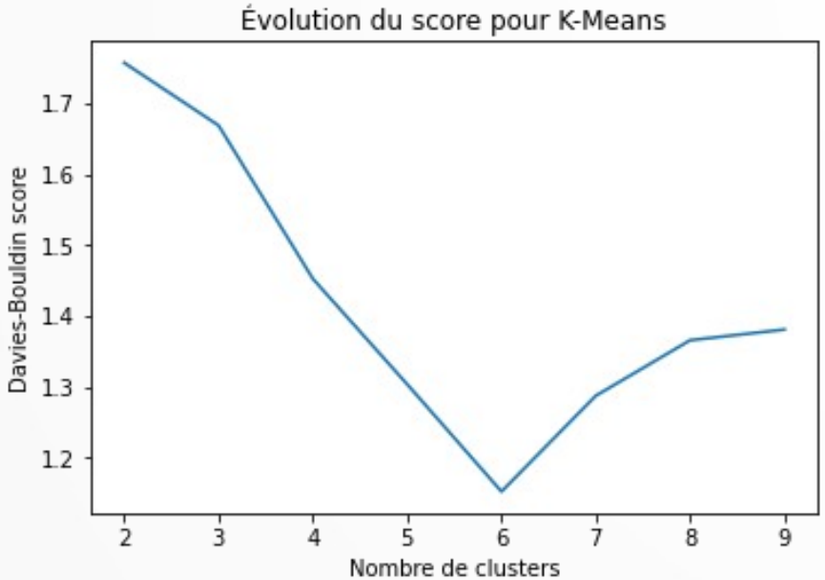
V- Classification RSMDT

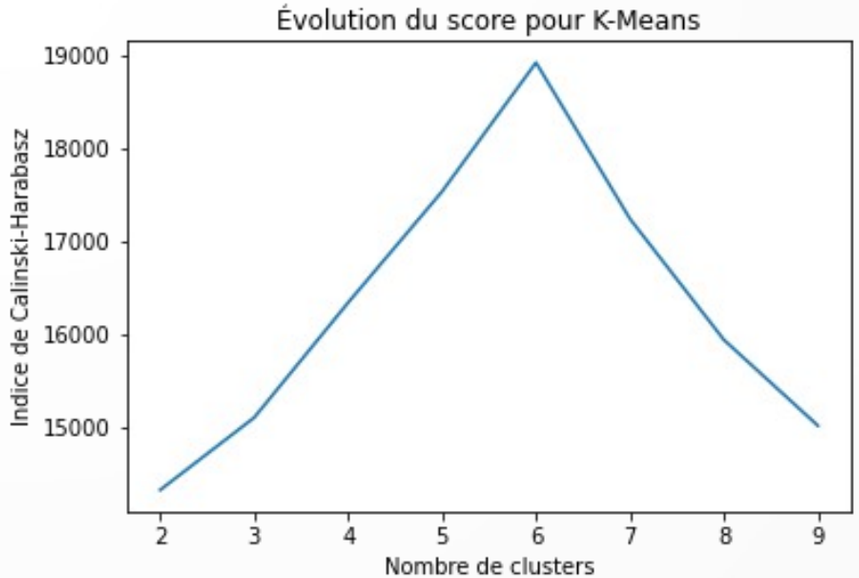
- Critères d’évaluation des segmentations
  - Indice de Davies-Bouldin
  - Indice de silhouette
  - Indice de Calinski-Harabasz
- Critères supplémentaires
  - Interprétabilité
  - Stabilité

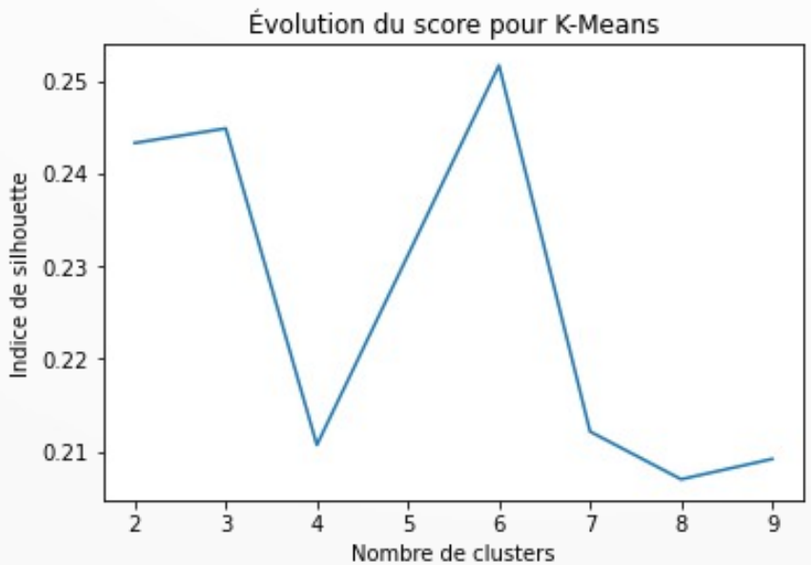
V- Algorithmes testés

- K-Means
- DBSCAN
- Classification agglomérative

V- K-Means : Nombre de clusters







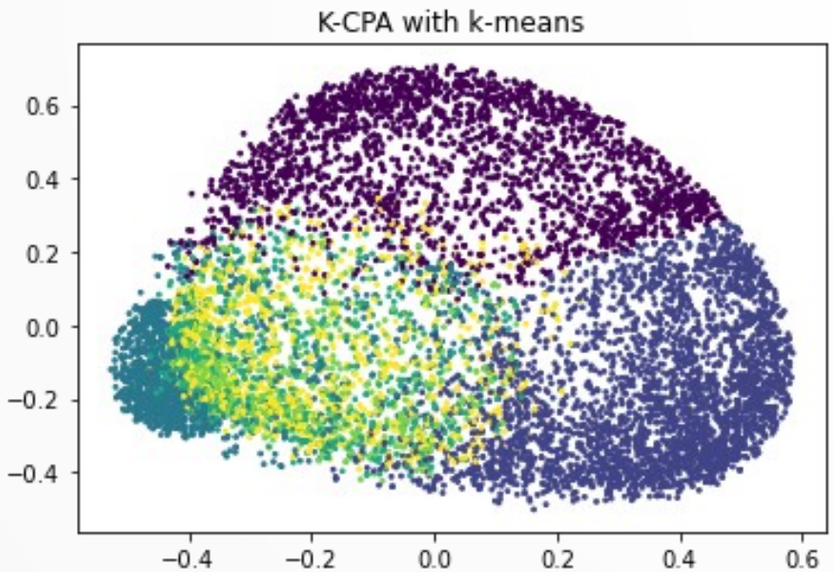
V- K-Means : Performances

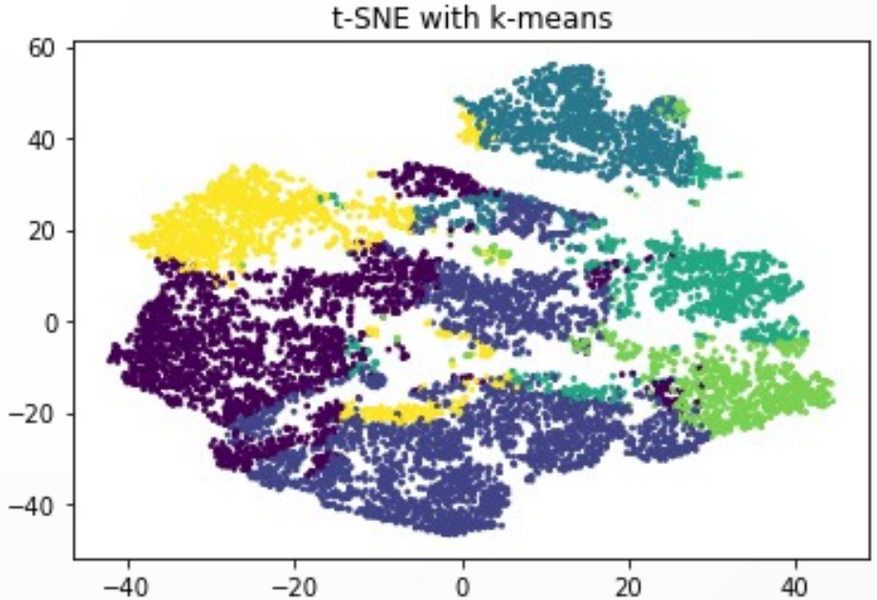
- Score de sihouette: 0.25
- Score de Davis-Bouldin: 1.15
- Score de calinski-harabasz: 19080.11

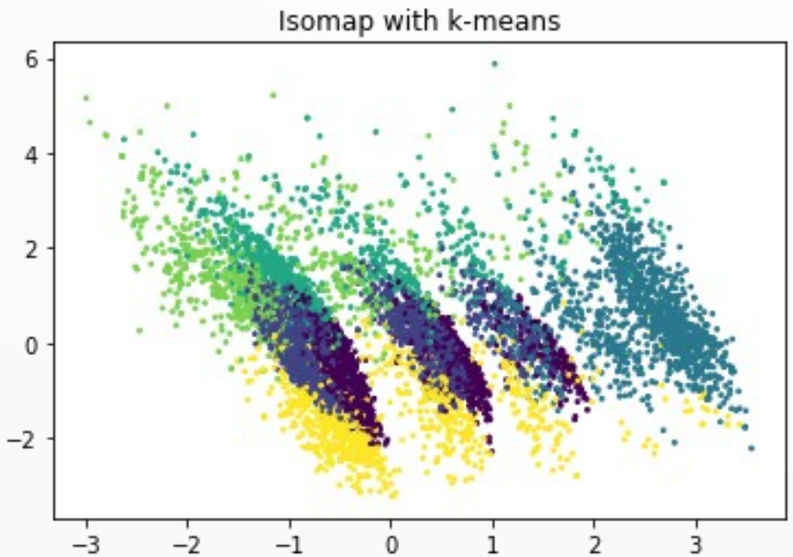
V- K-Means : Profils Moyen

class_kmeans	recency	total_spent	review_score	distance_city	product_description_lenght	proportion
0	238.939851	96.414424	1.426272	0.044878	649.616291	13.495500
1	220.334383	85.372200	4.357561	0.155019	656.352629	12.064776
2	233.421942	293.568459	4.314093	0.041238	793.130124	9.477528
3	396.100886	76.008131	4.613352	0.033036	583.693016	24.663020
4	216.567258	124.639651	4.329035	0.045247	2432.331990	7.578493
5	119.096038	75.402231	4.639038	0.032946	619.123522	32.720683

V- K-means : Visualisations







V- DBSAN : Paramétrage

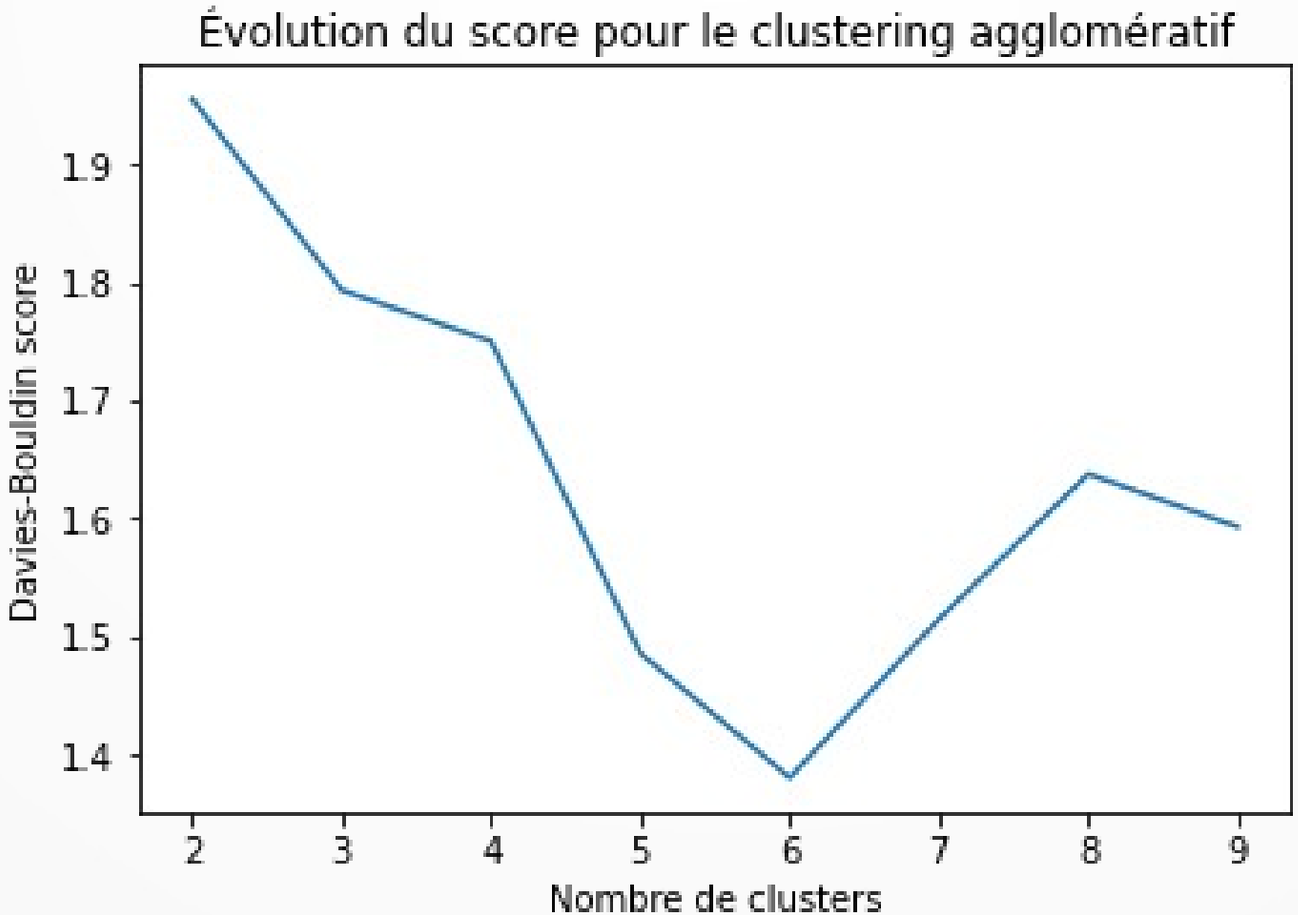
- Davies-Bouldin
- Calinski-Harabasz
- Silhouette

Davies-Bouldin	1.05
Calinski-Harabasz	176.53
Silhouette	0.42

Davies-Bouldin	3.64
Calinski-Harabasz	3776.04
Silhouette	0.02

Davies-Bouldin	1.05
Calinski-Harabasz	176.64
Silhouette	0.44

V- Classification agglomérative : Paramétrage



V- Classification agglomérative : Paramétrage

- Silhouette: 0.14
- Davis-Bouldin: 1.38
- Calinski-Harabasz: 3155.99

V- Choix de l'algorithme

- K-Means
  - Meilleurs résultats
- DBSCAN
  - Classes disproportionnées
  - Classes selon la note
- Agglomératif
  - Nécessite beaucoup de mémoire

V- Stabilité de la classification

- Sur plusieurs itérations :
  - Les clusters sont stables à 99.9 %
- Sur ajout de données (~60 jours) :
  - Les clusters sont stables à 99.25 %

Conclusion

- Classification avec K-Means
- Interprétable
- Performante : Score au minimum 200x supérieur
- Clusters stables
- Maintenance : Tous les 2 mois c’est suffisant

Limites

- Anonymisation des données
- Nombre de clusters limités
- Fréquence des achats faible