OLIST

Segmentation des clients d'un site de e-commerce

Marion Dedieu

12/2023

SOMMAIRE

01.

PROBLÉMATIQUE, CLEANING, EDA & FEATURE ENGINEERING

CONTEXTE

olist

Société brésilienne de vente en ligne

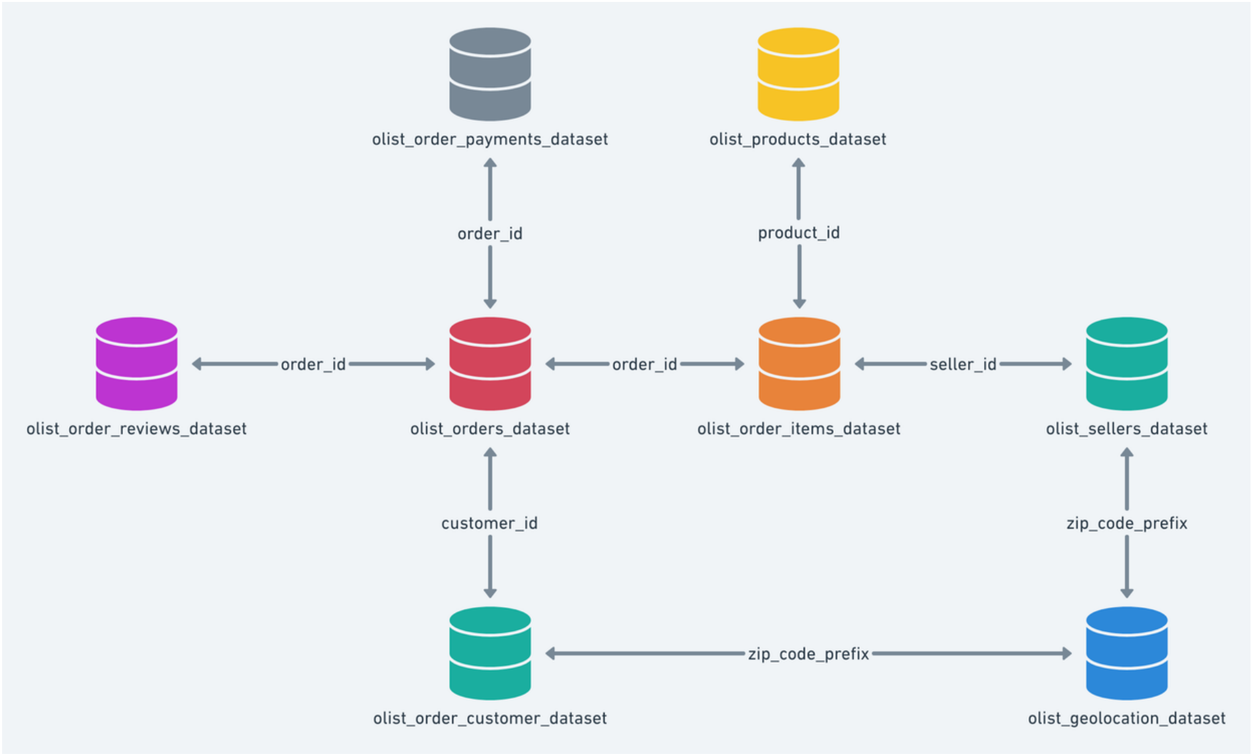
PROBLÉMATIQUE

- **Segmenter les clients** à l'aide de méthodes non supervisées pour optimiser les campagnes de communication de l'équipe Marketing

- Proposer un **contrat de maintenance** en analysant la stabilité des segments sur le temps pour déterminer la fréquence des mises à jour nécessaires.

JEU DE DONNÉES

- 9 fichiers de données brutes :
  - Vendeurs
  - Traduction des catégories
  - Commandes
  - Produits des commandes
  - Clients
  - Localisation
  - Paiements
  - Avis des clients
  - Produits



NETTOYAGE DES DONNÉES

Valeurs manquantes

Suppression des valeurs nulles (max 2.08% sur order_delivered_customer_date)

Doublons

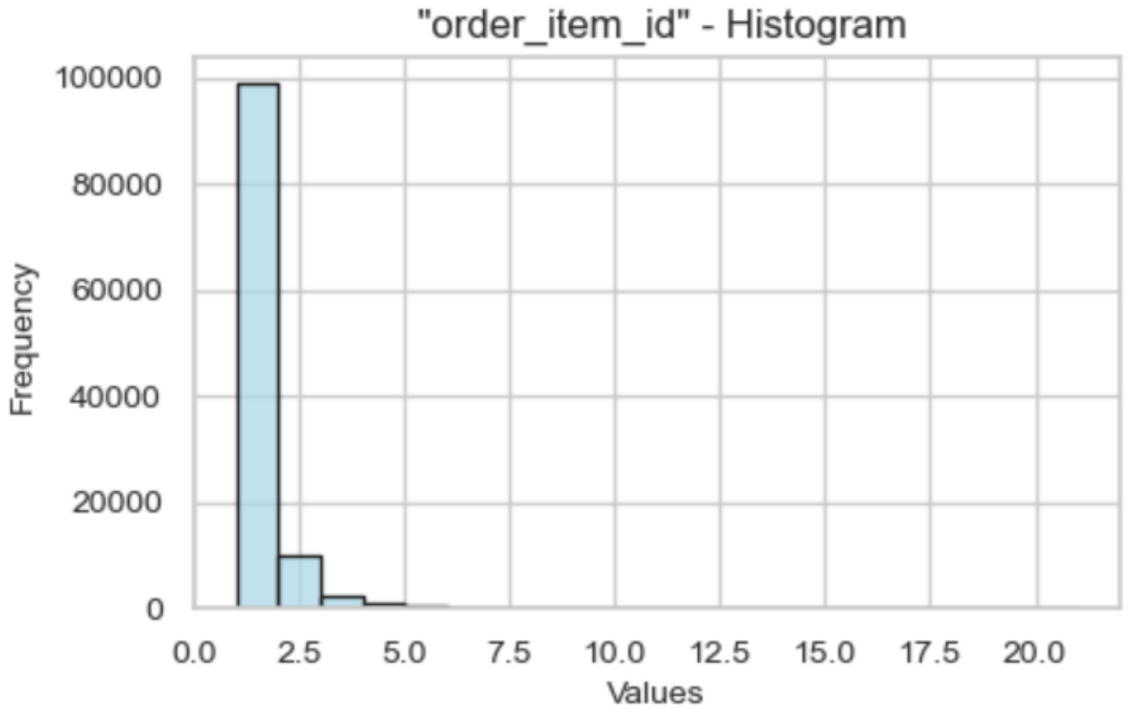
Pas de doublons

Outliers

Vérification des valeurs aberrantes et suppression si nécessaire
Ex : Le montant du paiement de 13 664,08 real s'explique par l'achat de 8 articles à 1680,01 real et des frais d'expédition de 28,01 chacun.

ANALYSE EXPLORATOIRE DE DONNÉES

NOMBRE D'ARTICLES PAR COMMANDE

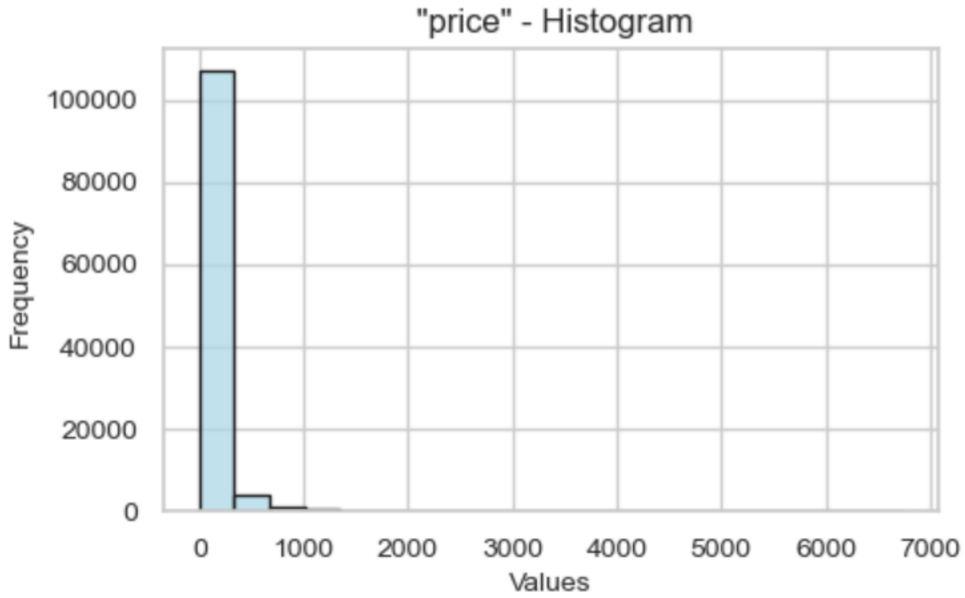


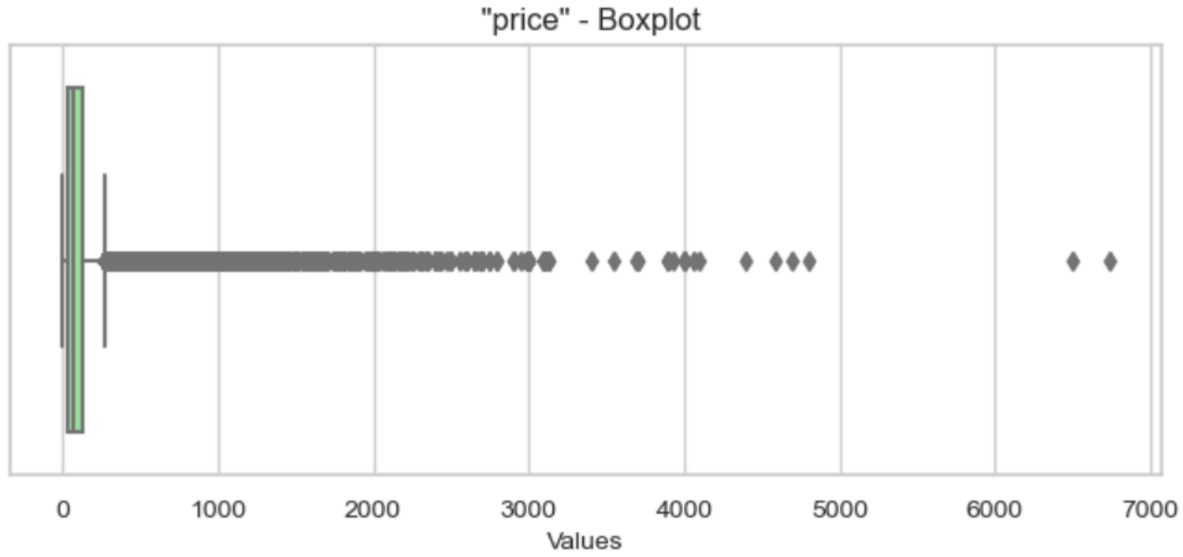
MÉDIANE : 1.00
MINIMUM: 1.00
MAXIMUM: 21.00

ANALYSE EXPLORATOIRE DE DONNÉES

PRIX

MÉDIANE : 74.90
MINIMUM: 0.85
MAXIMUM: 6735.00



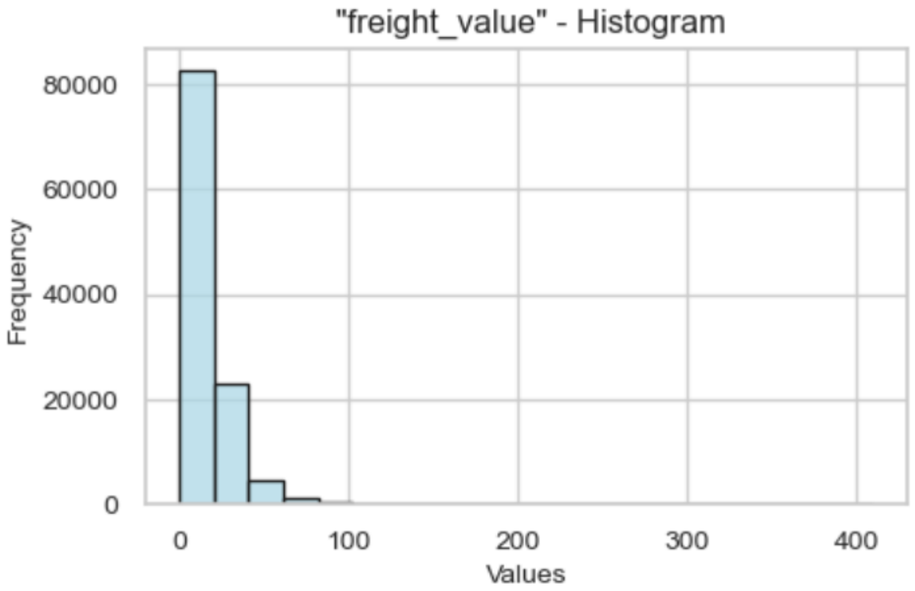


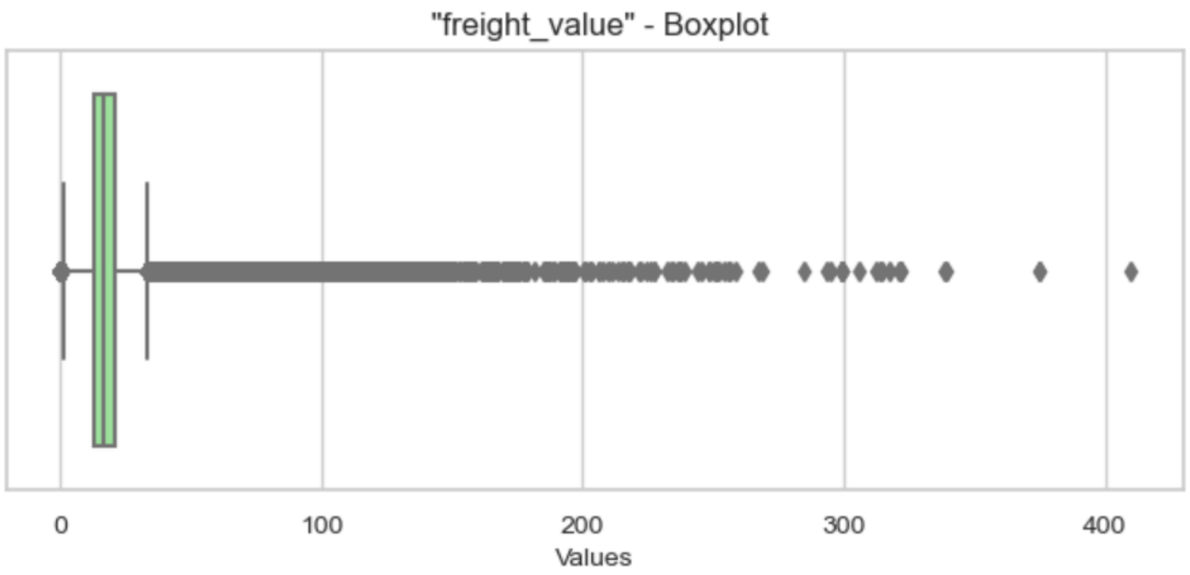
EN 2018 : 1 RÉAL BRÉSILIEN ≈ 0,23 €

ANALYSE EXPLORATOIRE DE DONNÉES

FRAIS DE PORT

MÉDIANE : 16.32
MINIMUM: 0.00
MAXIMUM: 409.68

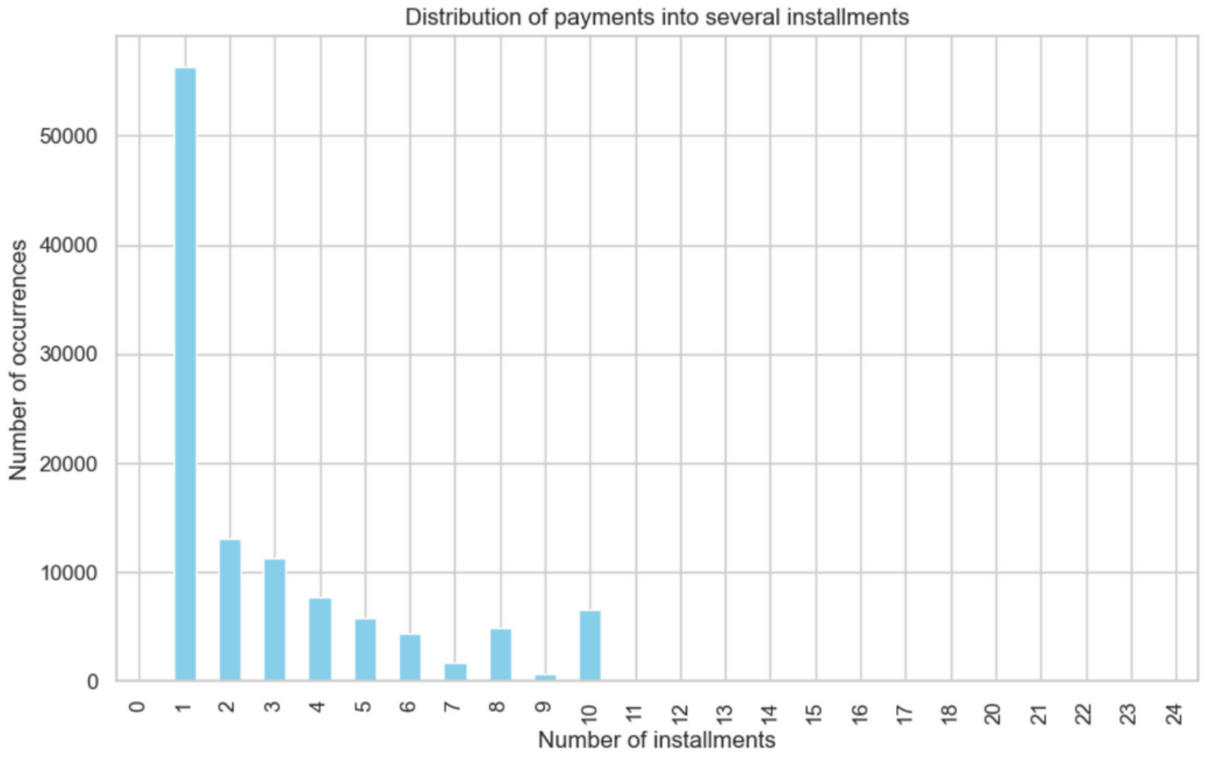


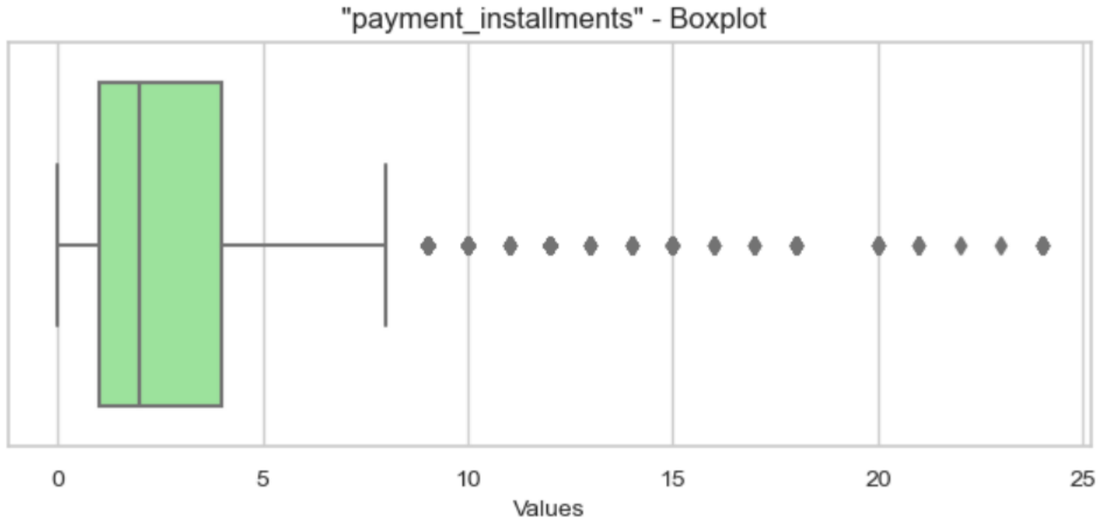


ANALYSE EXPLORATOIRE DE DONNÉES

NOMBRE DE VERSEMENTS

MÉDIANE : 2.00
MINIMUM: 0.00
MAXIMUM: 24.00



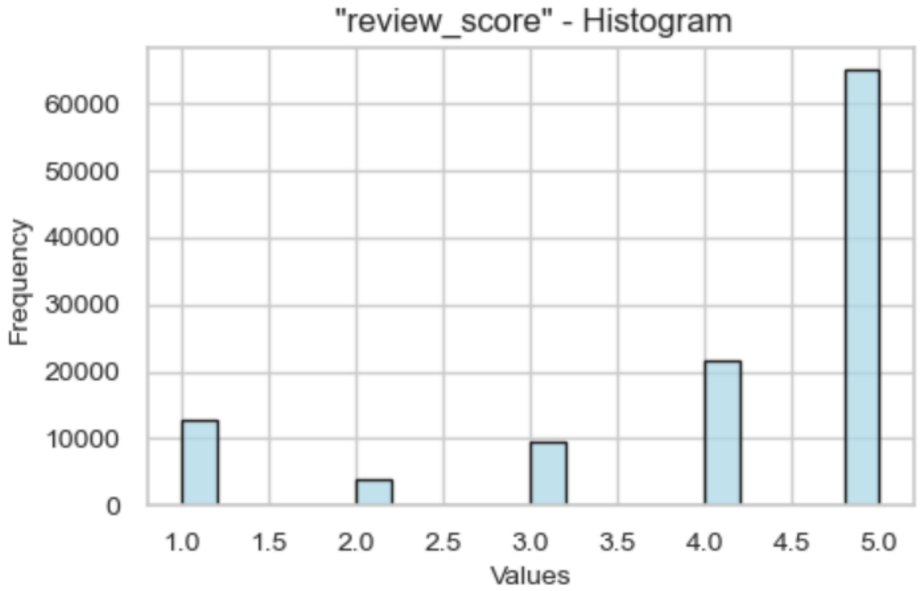


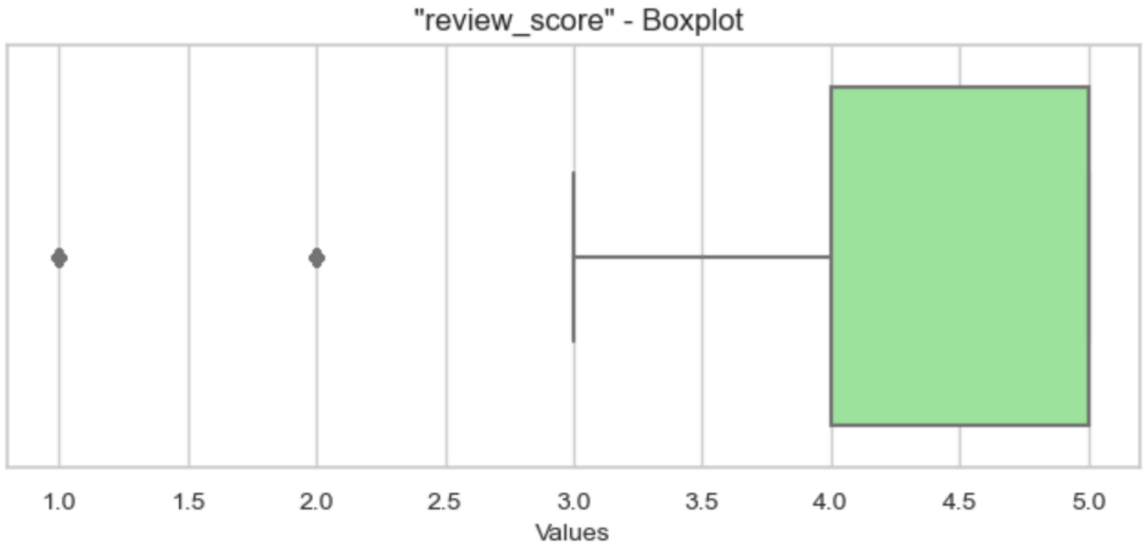
=> Suppression des versements à 0

ANALYSE EXPLORATOIRE DE DONNÉES

AVIS CLIENT

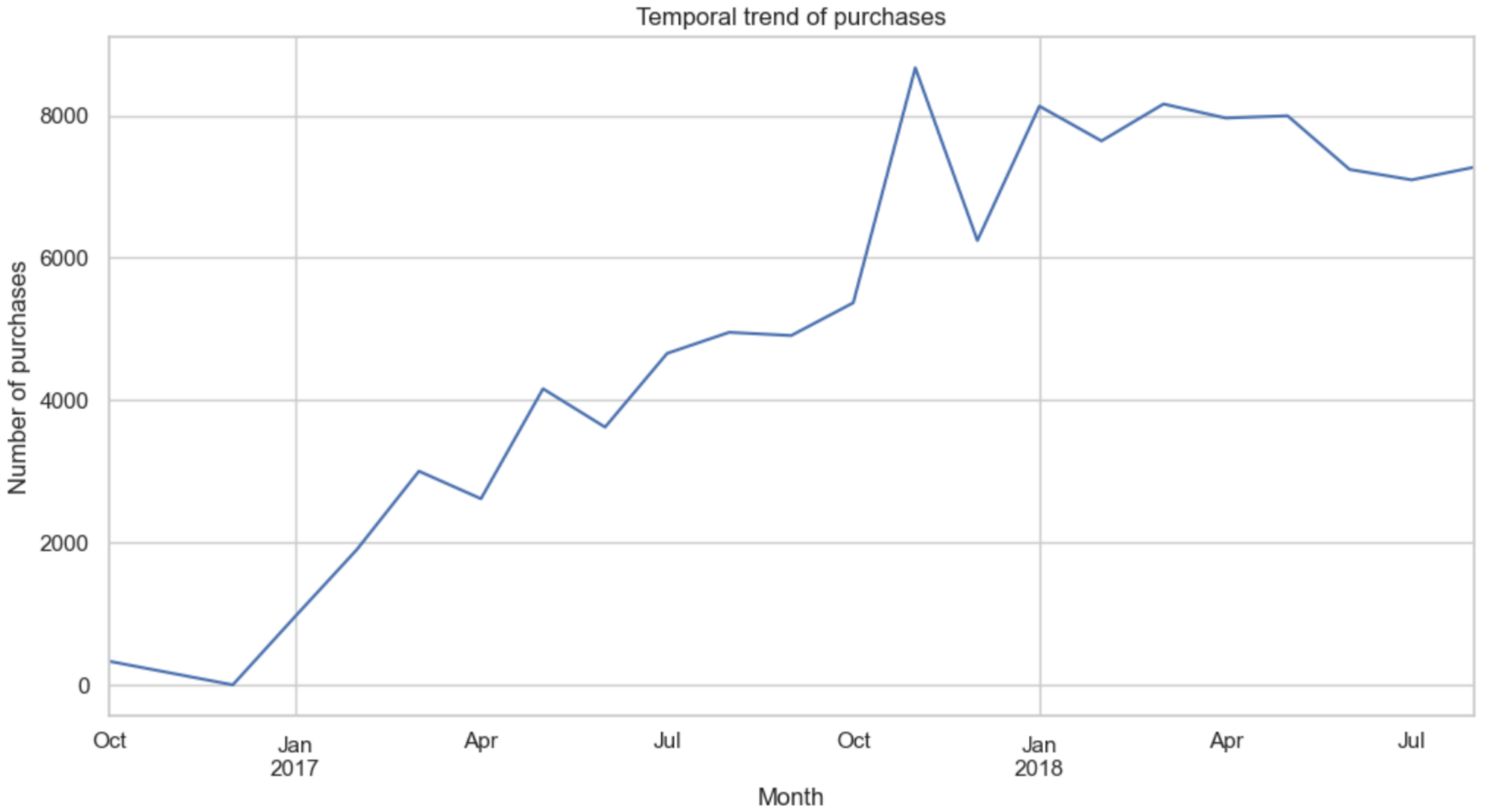
MOYENNE : 4.08
MÉDIANE : 5.00
MINIMUM: 1.00
MAXIMUM: 5.00





ANALYSE EXPLORATOIRE DE DONNÉES

NOMBRE DE COMMANDES PAR DATE

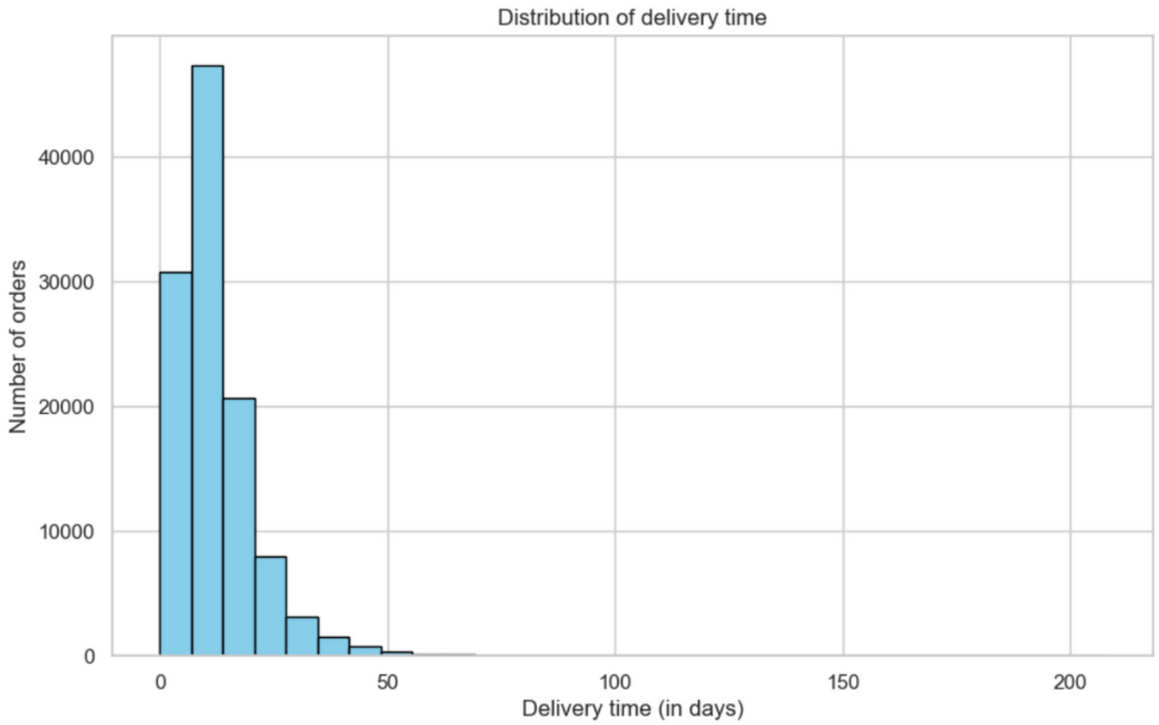


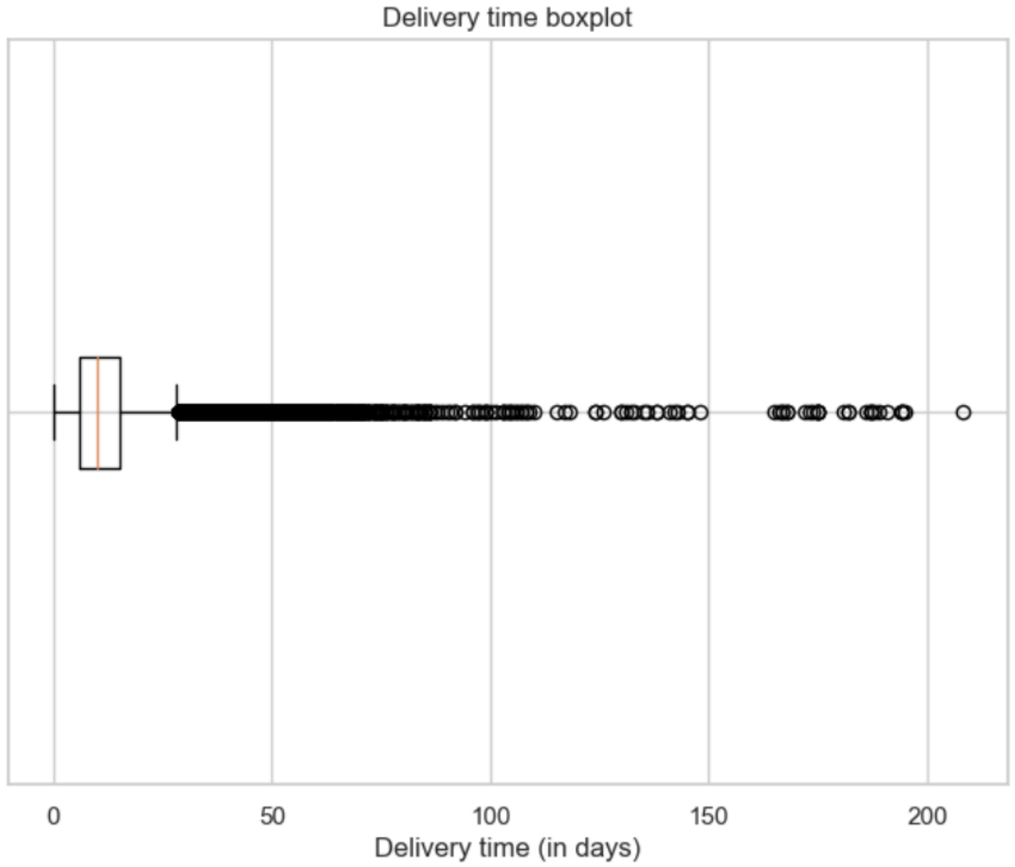
ANALYSE EXPLORATOIRE DE DONNÉES

DÉLAI DE LIVRAISON

Date de livraison au client - Date d'achat

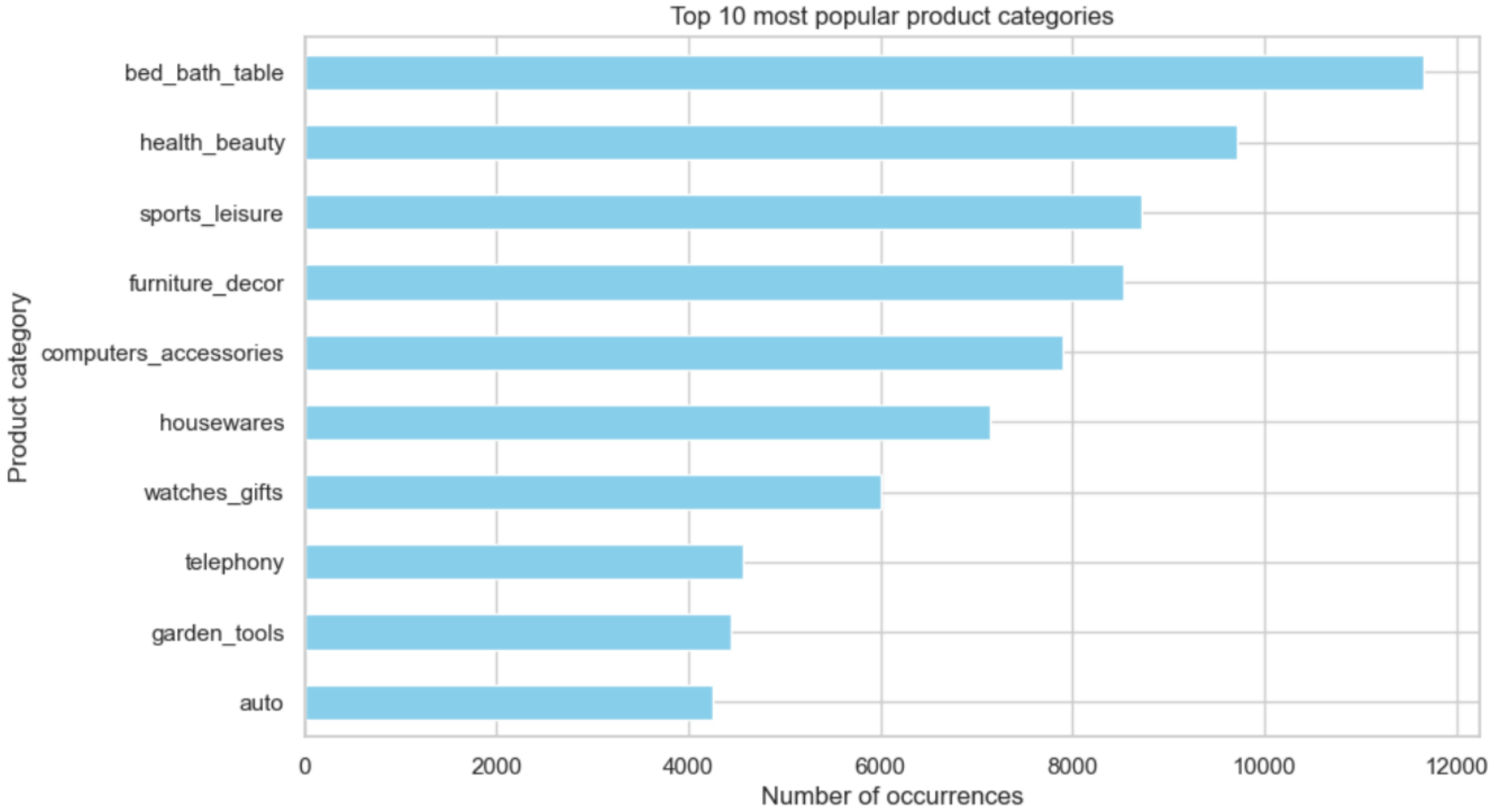
MOYENNE : 11.97
MINIMUM: 0.00
MAXIMUM: 208.00





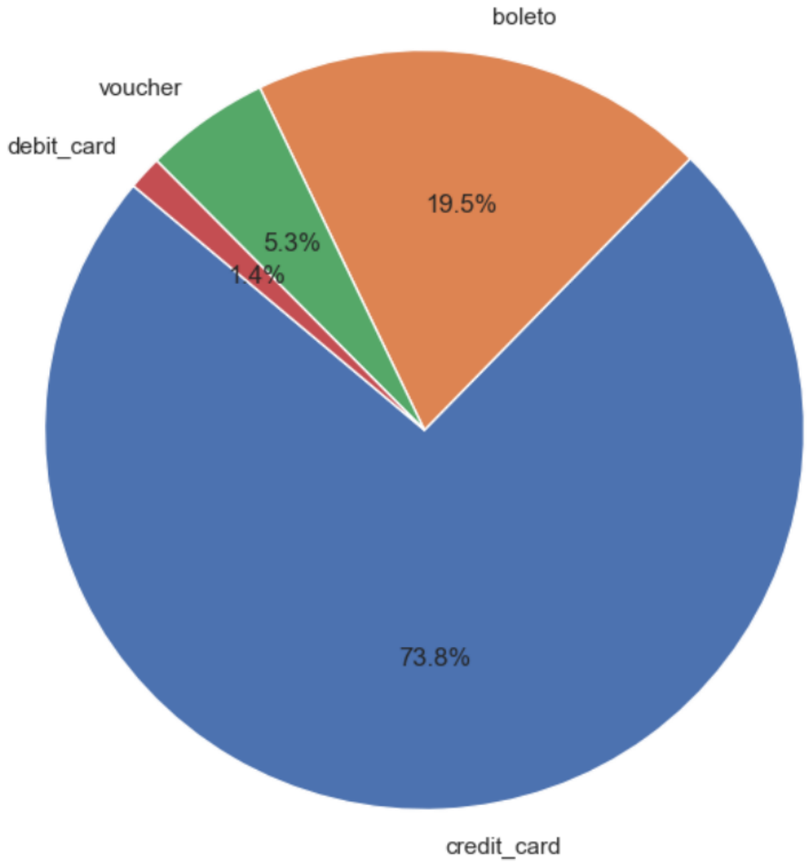
ANALYSE EXPLORATOIRE DE DONNÉES

71 CATÉGORIES DE PRODUITS



ANALYSE EXPLORATOIRE DE DONNÉES

TYPES DE PAIEMENTS



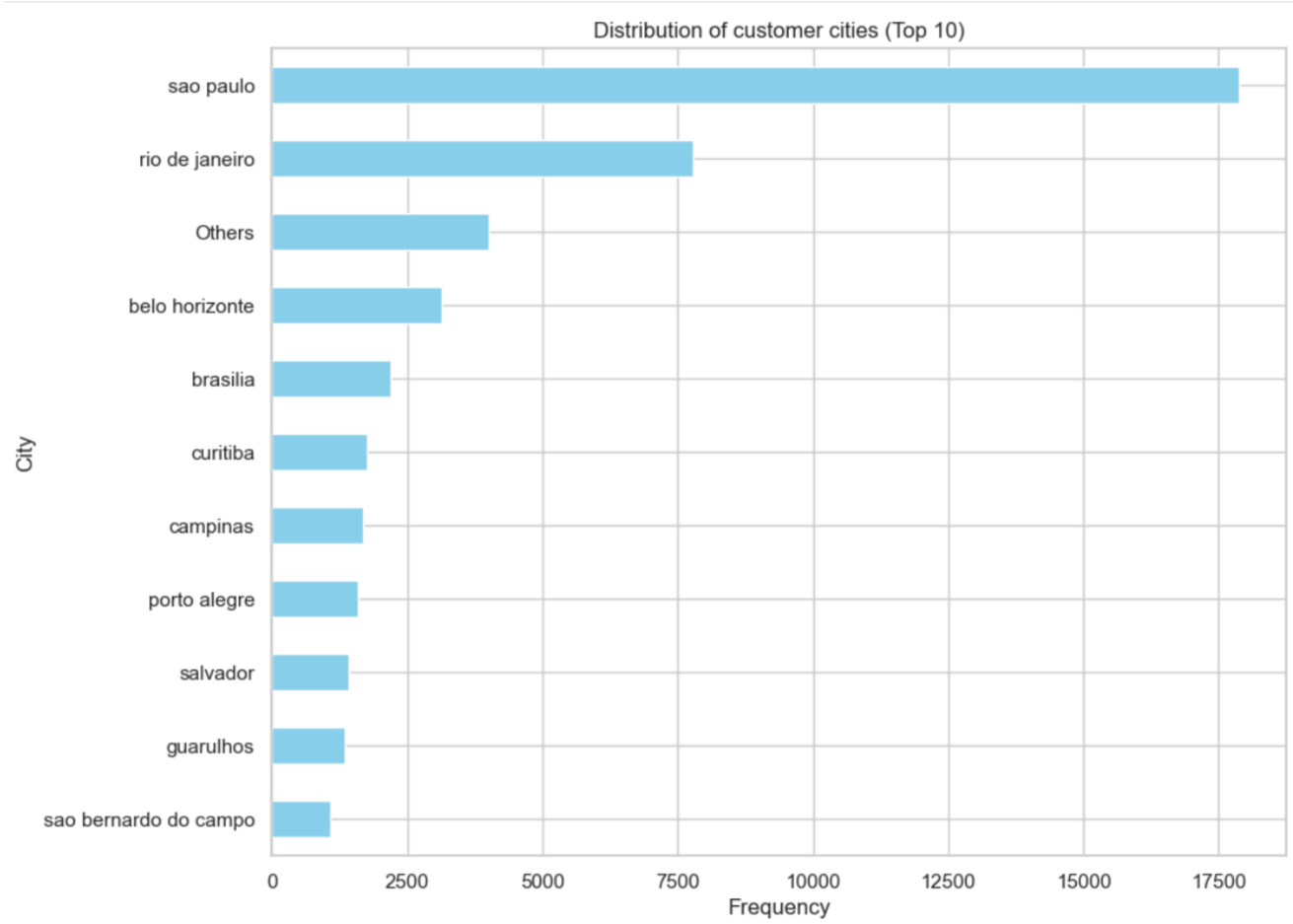
Le boleto bancário, ou boleto en abrégé, est le mode de paiement en espèces privilégié au Brésil.

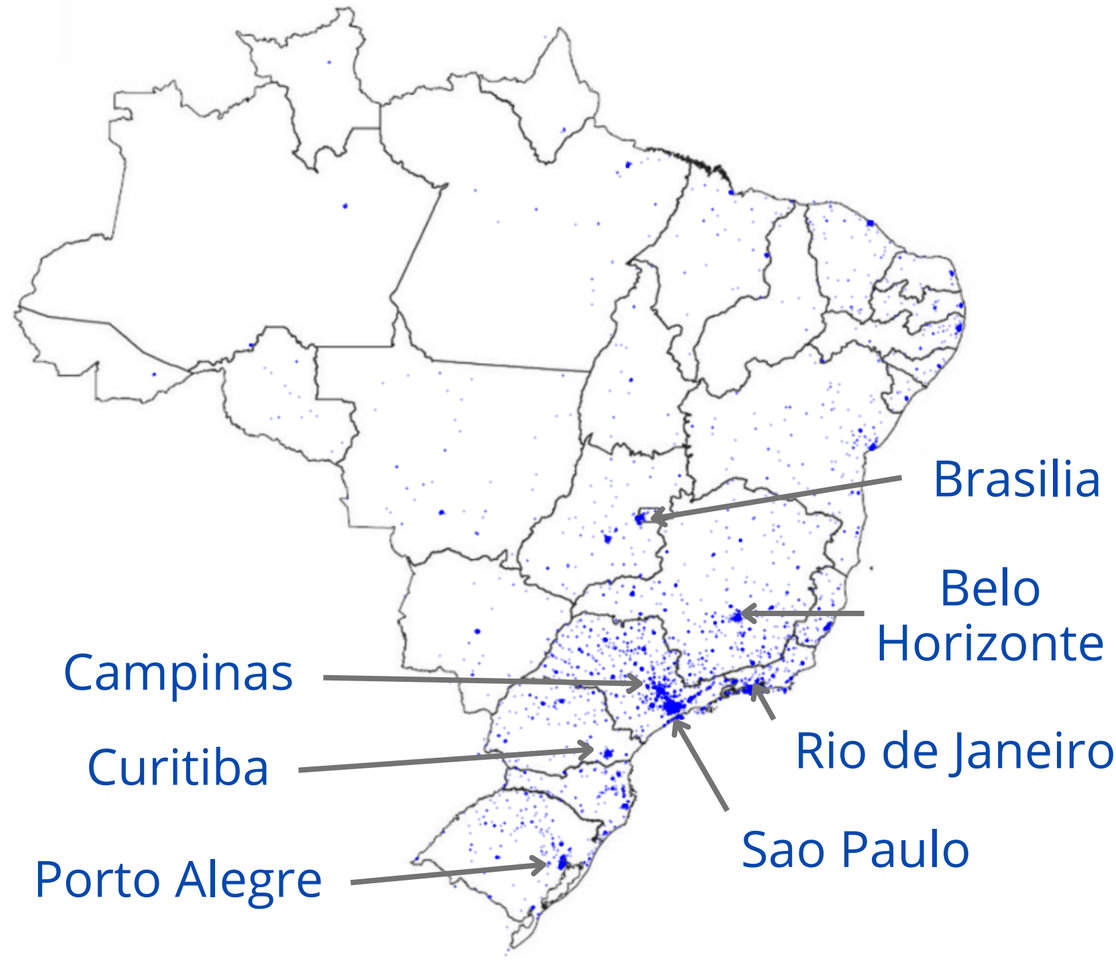
bon imprimé ou d'un bon virtuel (PDF ou image)

=> permet d'acheter en ligne sans devoir ouvrir un compte bancaire ou obtenir une carte de crédit.

ANALYSE EXPLORATOIRE DE DONNÉES

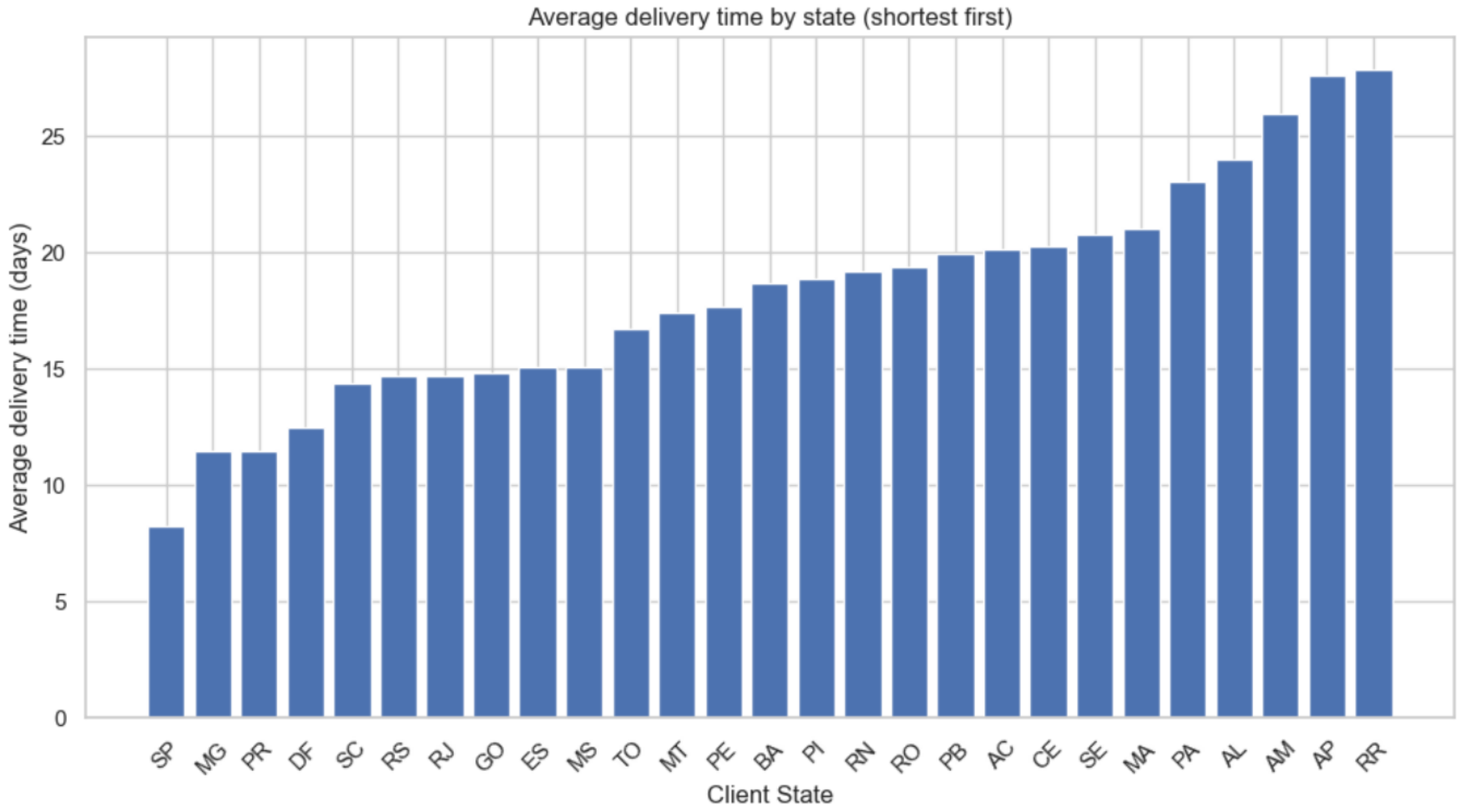
LOCALISATION DES CLIENTS





ANALYSE EXPLORATOIRE DE DONNÉES

DÉLAI DE LIVRAISON PAR ETAT



- Sao Paulo (SP)
- Minas Gerais (MG)
- Paraná (PR)
- District fédéral (DF) (contient Brasilia)
- Santa Catarina (SC)

Tous ces États sont les plus proches de São Paulo.

FEATURE ENGINEERING

CRÉATION DE FEATURES

- **Récence** : durée écoulée depuis la dernière commande d'un client
- **Fréquence** : mesure le nombre de commandes passées par un client
- **Monétaire** : mesure le montant cumulé des commandes passées par un client
- Nombre total **d'articles** par client
- **Note** moyenne donnée
- **Délai de livraison** moyen
- Nombre de **versements** moyen
- Ratio de **frais de port** : rapport entre les dépenses d'expédition et les dépenses totales pour chaque client

TRANSFORMATIONS

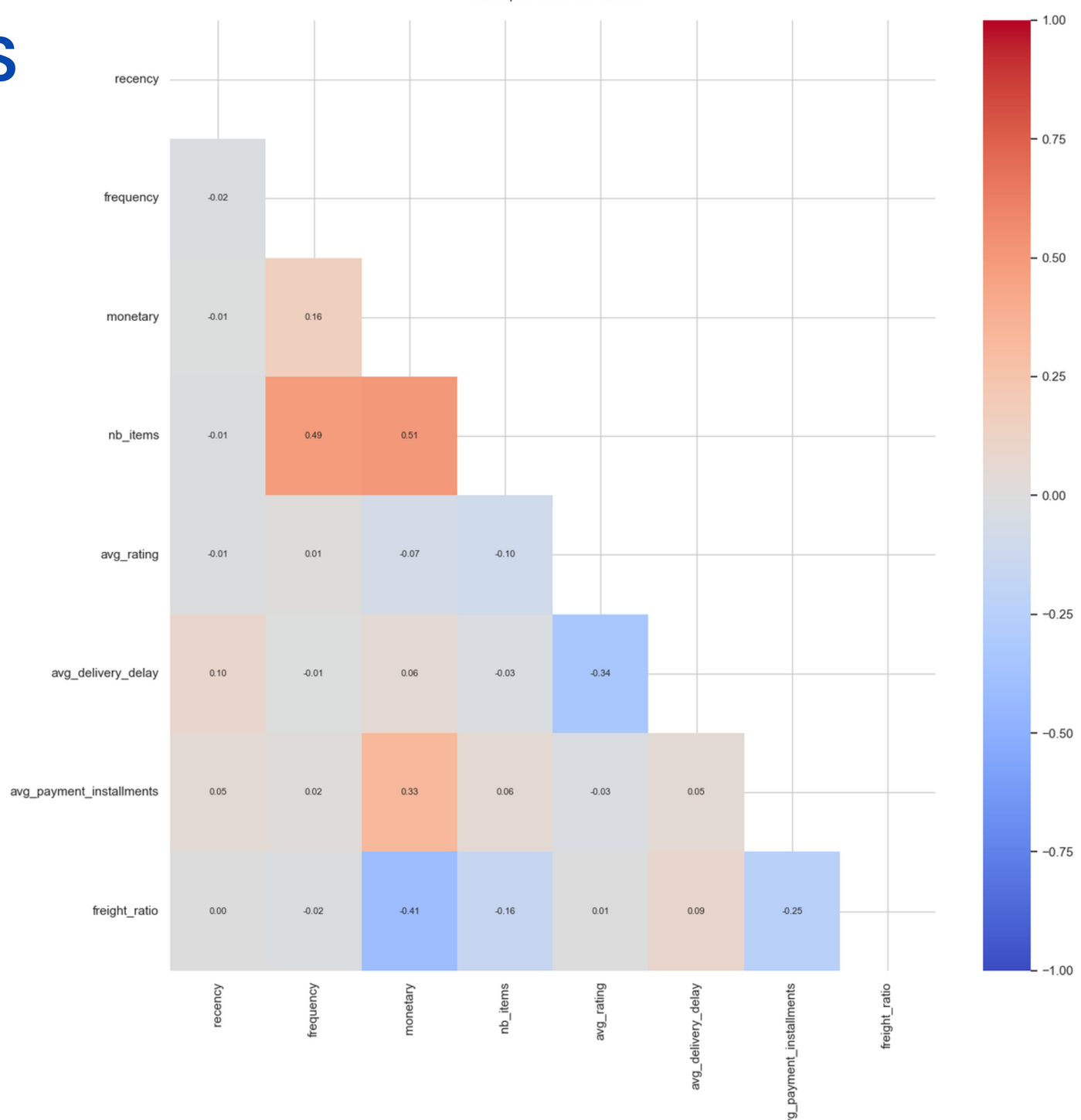
Pour les features/modèles nécessaires :

- **Normalisation :** StandardScaler : Ajuste la moyenne des données à 0 et l'écart-type à 1
- **Passage au logarithme :** Réduit l'effet des valeurs extrêmes et rend la distribution des données plus proche d'une forme normale

TEST AUTRES FEATURES

- Catégories de produits
- Mois d'achat favori
- Distance du client

HEATMAP DES NOUVELLES FEATURES



Corrélations positives :

- le montant des dépenses avec le nombre d'échéances,
- le nombre d'articles achetés avec le montant total dépensé et la fréquence d'achat.

Corrélations négatives :

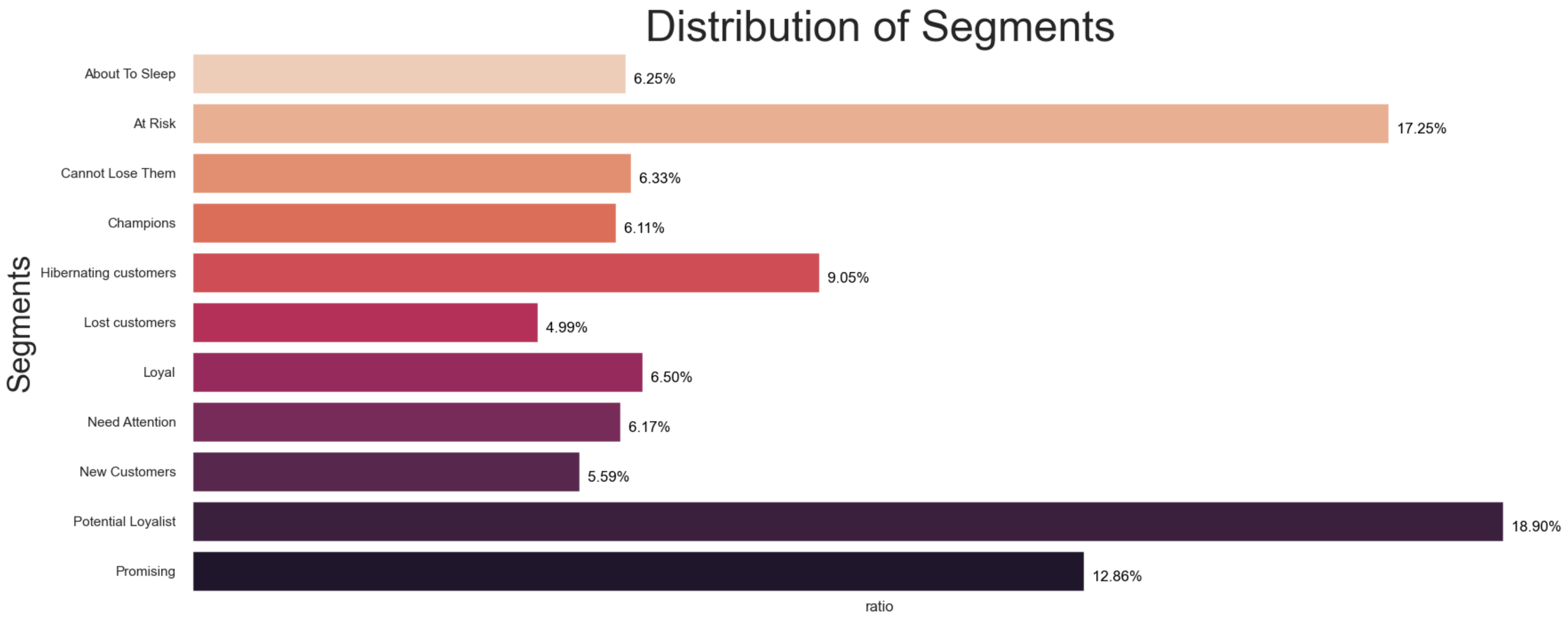
- la note d'avis est directement corrélée au délai de livraison, ce qui implique qu'il s'agit d'un critère de satisfaction,
- les frais de livraison diminuent lorsque le montant total de la commande augmente.

SEGMENTATION RFM CLASSIQUE

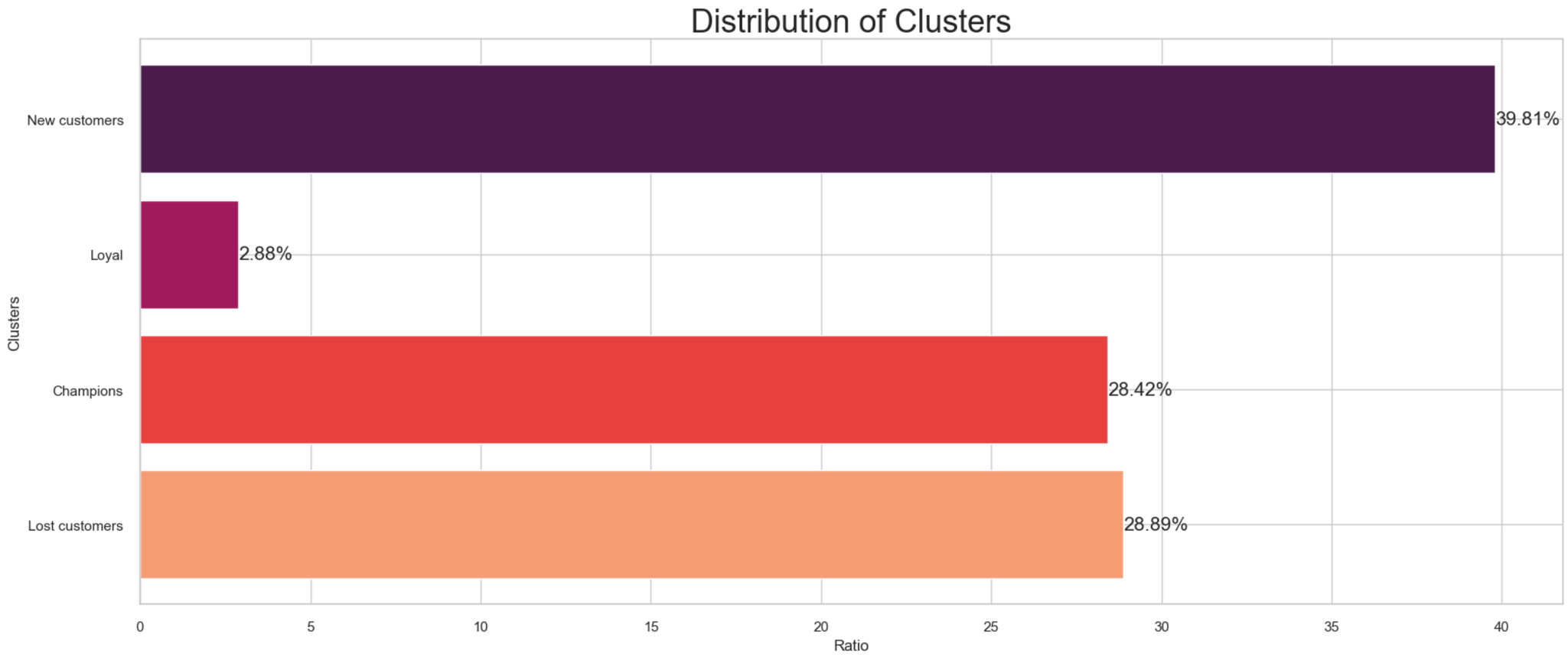
Segment	Scores
Champions	555, 554, 544, 545, 454, 455, 445
Loyal	543, 444, 435, 355, 354, 345, 344, 335
Potential Loyalists	553, 551, 552, 541, 542, 533, 532, 531, 452, 451, 442, 441, 431, 453, 433, 432, 423, 353, 352, 351, 342, 341, 333, 323
New Customers	512, 511, 422, 421 412, 411, 311
Promising	525, 524, 523, 522, 521, 515, 514, 513, 425,424, 413,414,415, 315, 314, 313
Need Attention	535, 534, 443, 434, 343, 334, 325, 324
About To Sleep	331, 321, 312, 221, 213, 231, 241, 251
Cannot Lose Them But Losing	155, 154, 144, 214,215,115, 114, 113
At Risk	255, 254, 245, 244, 253, 252, 243, 242, 235, 234, 225, 224, 153, 152, 145, 143, 142, 135, 134, 133, 125, 124
Hibernating Customers	332, 322, 233, 232, 223, 222, 132, 123, 122, 212, 211
Losing But Engaged	111, 112, 121, 131, 141, 151 Engagement: Last email campaign clicked in the last 180 days *OR* Last session_start in the last 90 days
Lost Customers	111, 112, 121, 131, 141, 151

Source : https://documentation.bloomreach.com/engagement/docs/rfm-segmentation

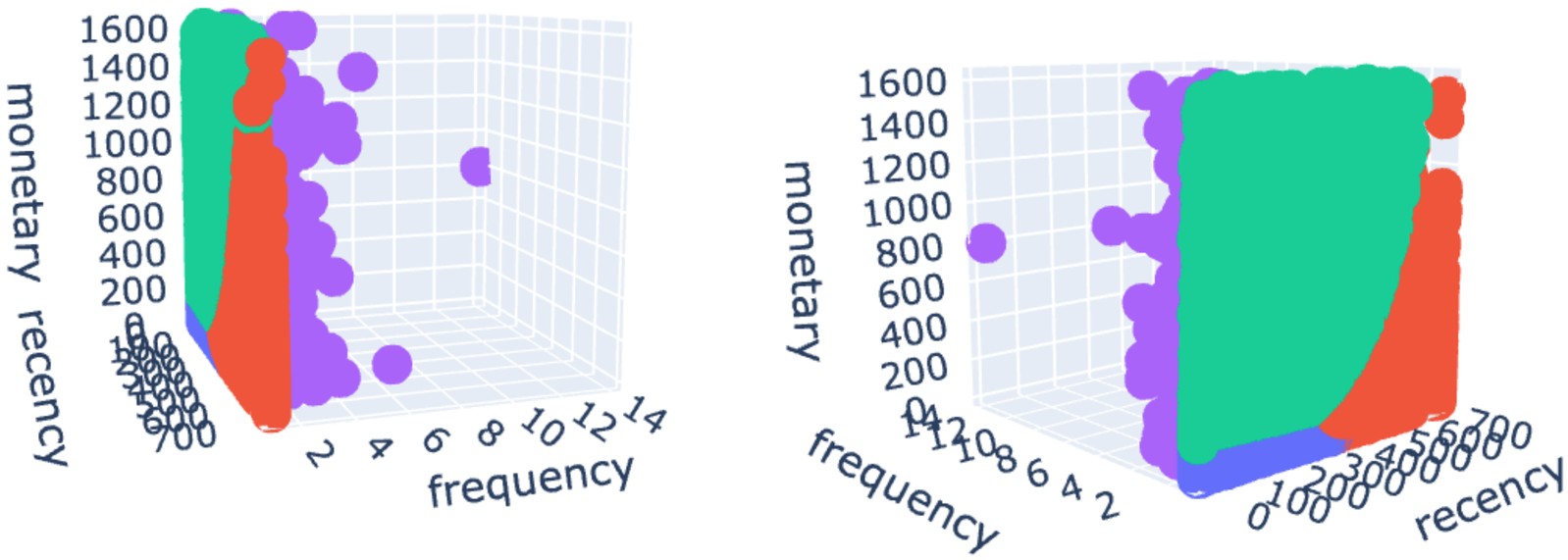
SEGMENTATION RFM CLASSIQUE



CLUSTERING RFM BASIQUE K-MEANS



CLUSTERING RFM BASIQUE K-MEANS



02.

APPROCHE DE MODÉLISATION

APPROCHES DE MODÉLISATION EXPLORÉES

BIRCH

DBSCAN

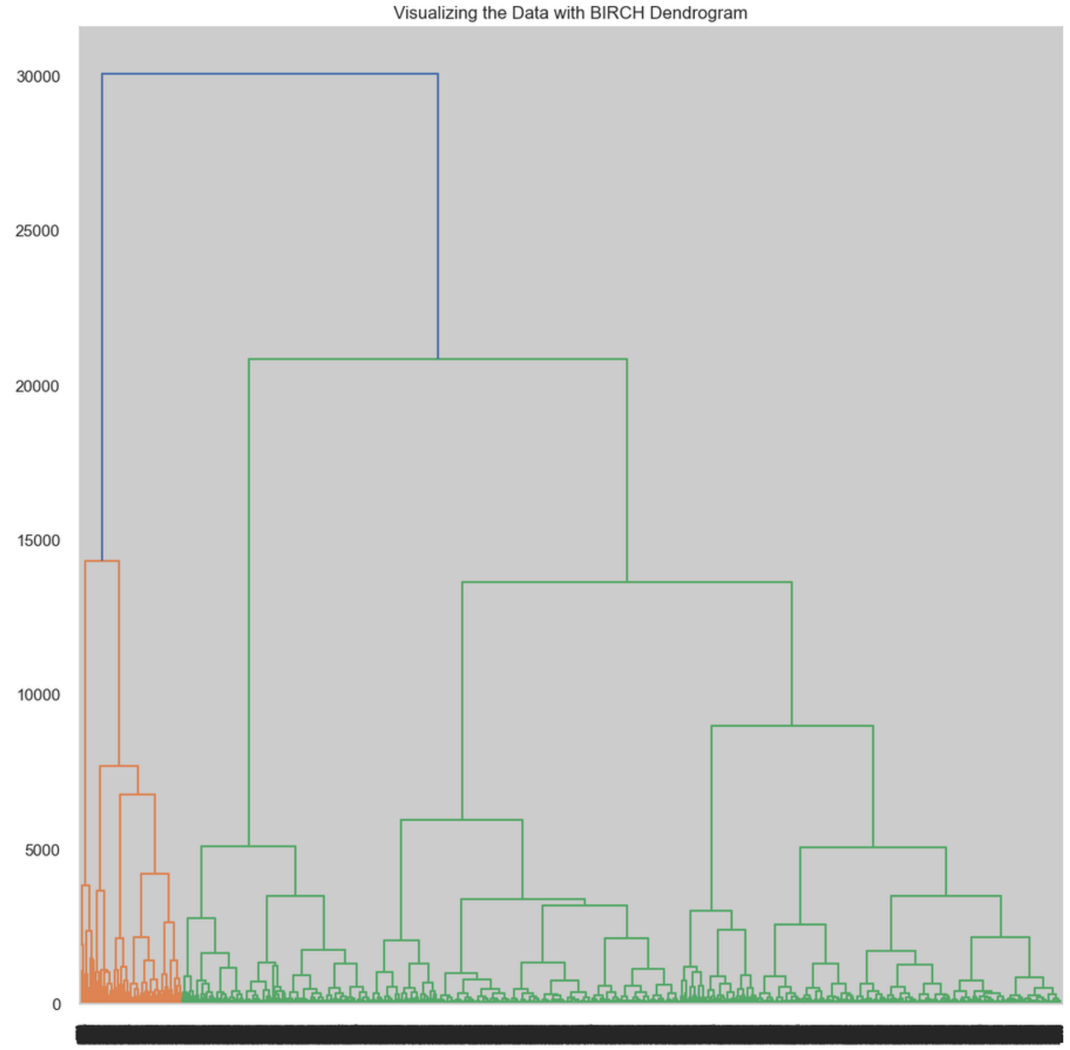
K-Means

BIRCH

(Balanced Iterative Reducing and Clustering using Hierarchies)

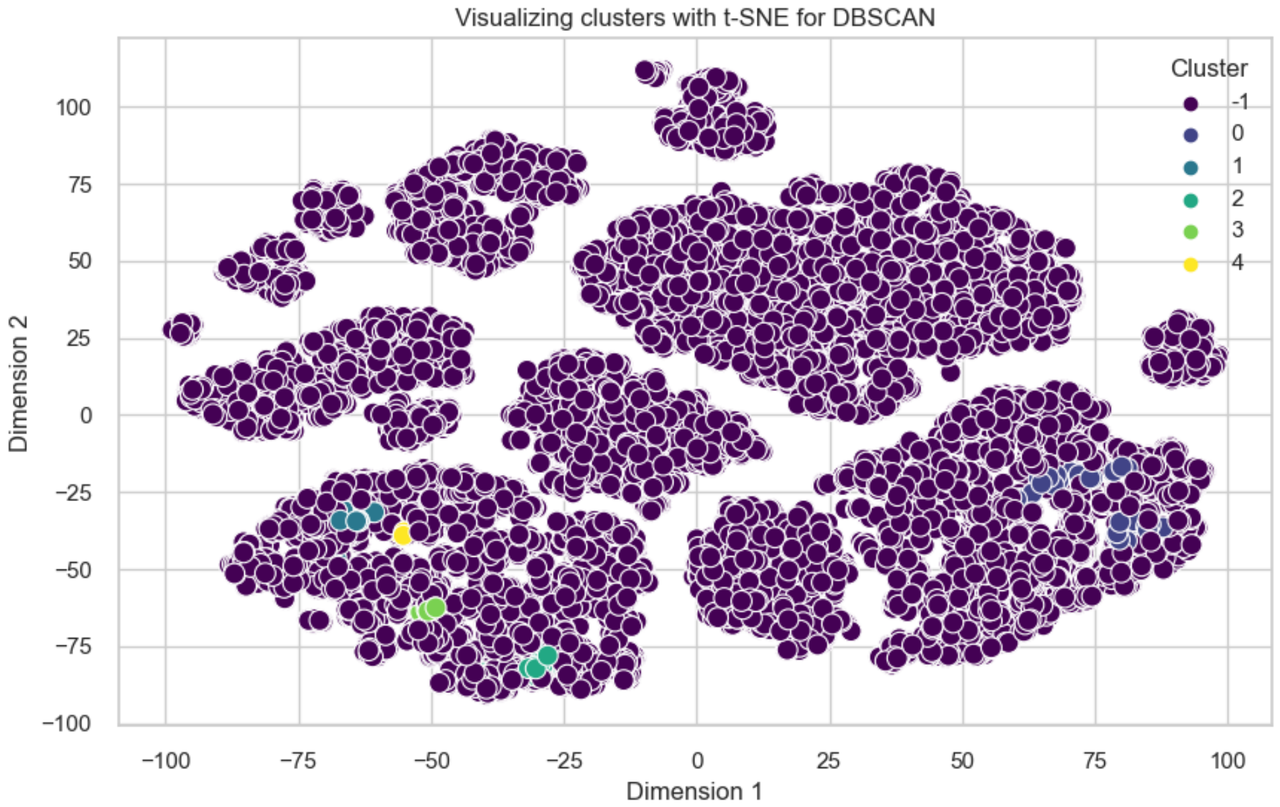
- **Caractéristiques Principales:**
  - Construit un arbre CF (Clustering Feature) pour condenser les données.
  - Efficace pour les grandes bases de données.
- **Avantages:**
  - Capable de traiter des ensembles de données volumineux.
  - Peut être utilisé de manière incrémentale.
- **Inconvénients:**
  - Moins efficace pour un nombre élevé de features.
  - Sensible aux données bruitées.

Tests effectués sur un échantillon de 20%



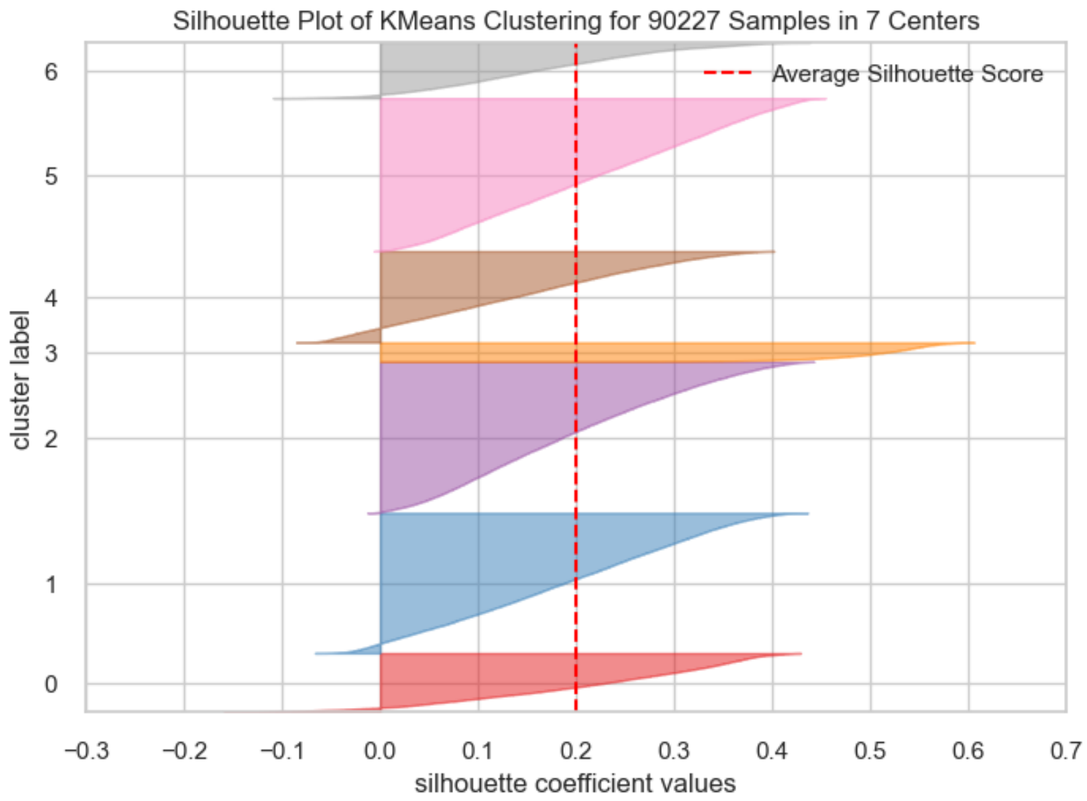
DBSCAN

- **Caractéristiques Principales:**
  - Identifie des clusters basés sur la densité des données.
  - Bon pour détecter des formes irrégulières de clusters et exclure les outliers.
- **Avantages:**
  - Ne nécessite pas de spécifier le nombre de clusters.
  - Découvre des formes de cluster complexes, robuste aux outliers.
- **Inconvénients:**
  - Sensible aux paramètres (eps et MinPts).
  - Peut ne pas bien fonctionner avec des différences de densité variées.



K-MEANS

- **Caractéristiques Principales:**
  - Partitionne les données en K clusters basés sur la distance aux centroïdes.
  - Nécessite de définir le nombre de clusters à l'avance.
- **Avantages:**
  - Simple et facile à comprendre.
  - Efficace en termes de calcul, particulièrement sur des petits datasets.
- **Inconvénients:**
  - Sensible aux outliers.
  - Suppose que les clusters sont sphériques et de taille relativement égale.



MESURES DE PERFORMANCE

- **Score de Silhouette :**
  - Identifie dans quelle mesure les points de données au sein des clusters sont similaires les uns aux autres et différents des clusters voisins
  - Un score plus élevé indique une meilleure qualité de clustering.

- **Score Calinski-Harabasz :**
  - Quantifie la séparation et la cohésion des clusters en maximisant la variance inter-clusters tout en minimisant la variance intra-cluster
  - Un score plus élevé indique une meilleure séparation des clusters.

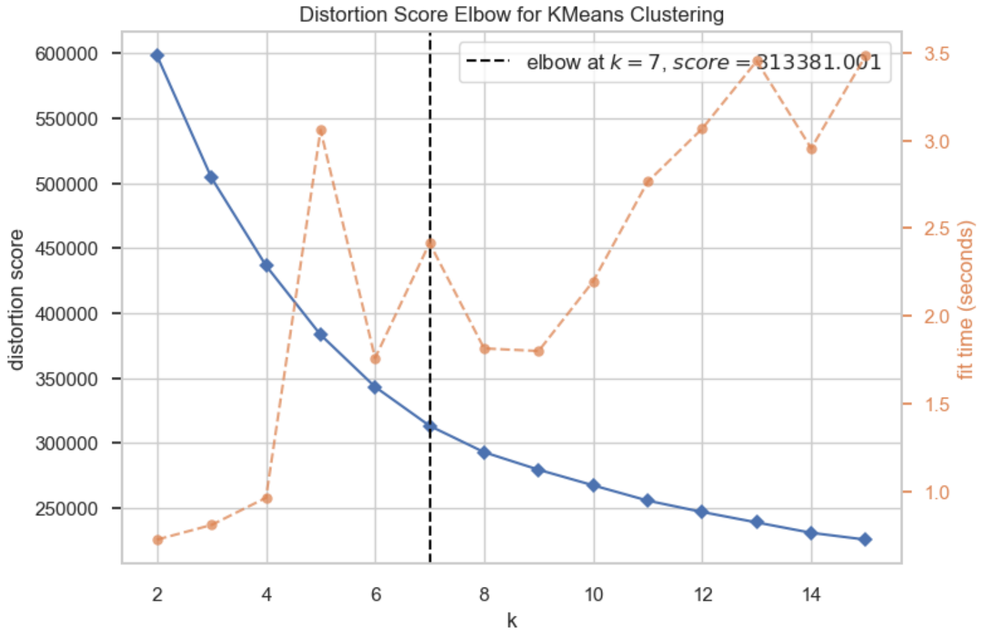
- **Indice Davies-Bouldin :**
  - Mesure la similarité moyenne entre chaque cluster et son cluster voisin le plus proche
  - Un score inférieur indique une meilleure séparation des clusters.

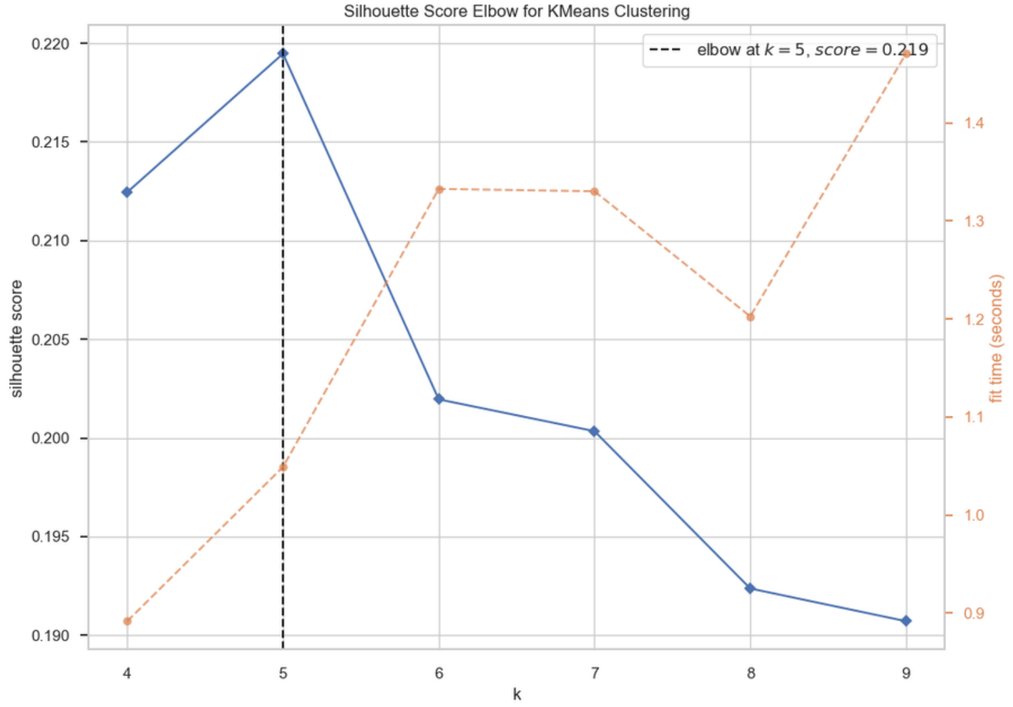
COMPARAISONS DES PERFORMANCES

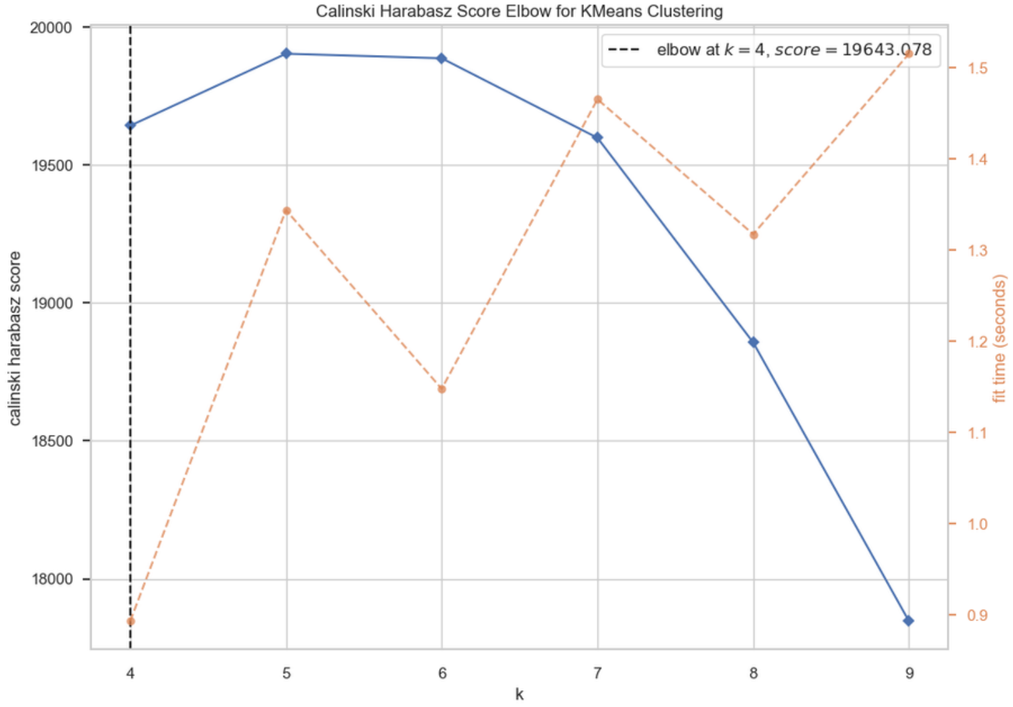
	BIRCH (échantillon) n_clusters = 7	DBSCAN eps : 0.3 min_samples : 20 Nombre de clusters : 5	K-MEANS k = 6
Silhouette	0.3229	-0.3894	0.2019
Calinski-Harabasz	14343.4042	45.9312	19886.9133
Davies-Bouldin	0.8872	1.4326	1.3540
Fit time	+++	++	+

K-MEANS : CHOIX DU NOMBRE OPTIMAL DE K

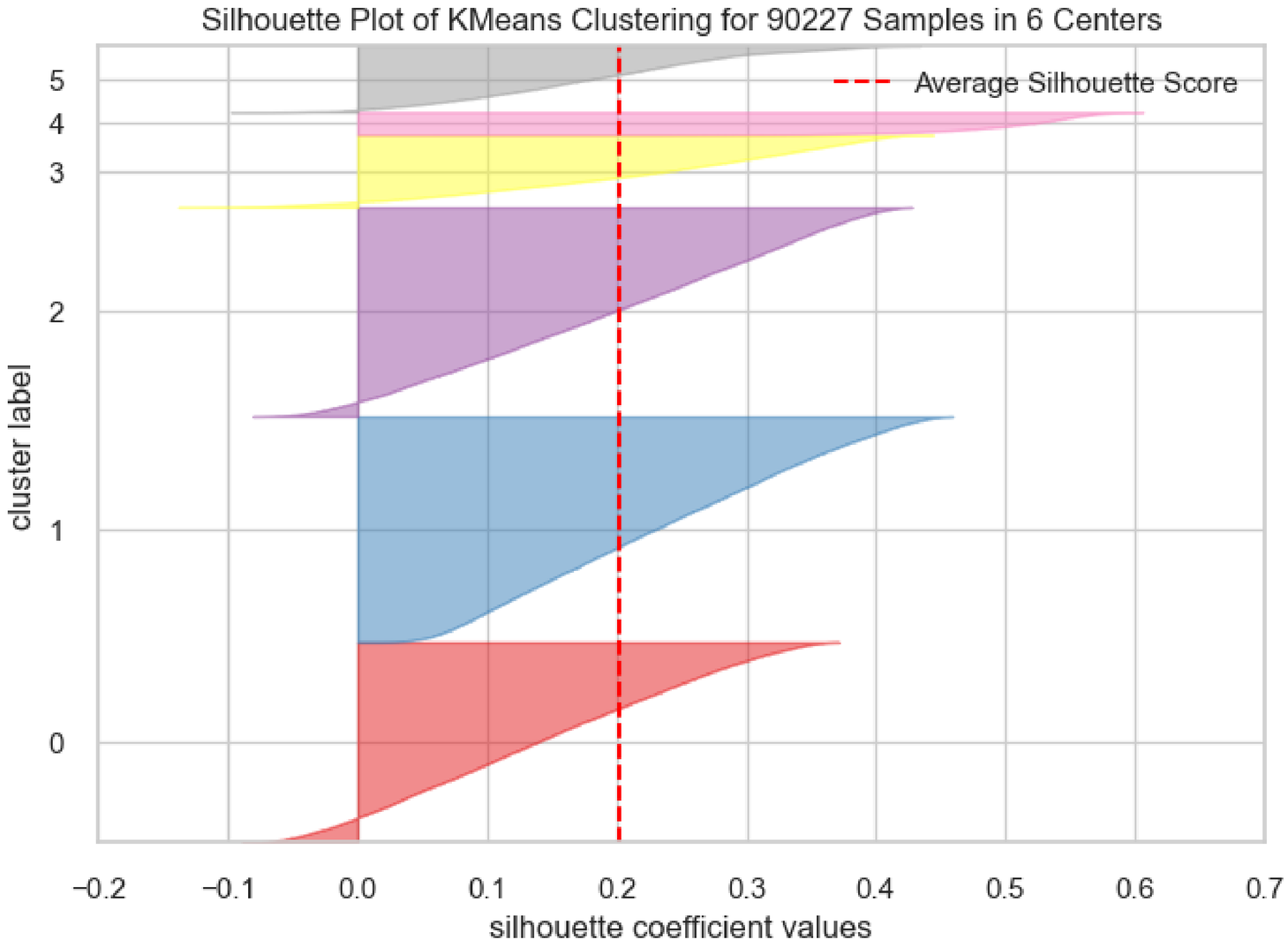
K = 6







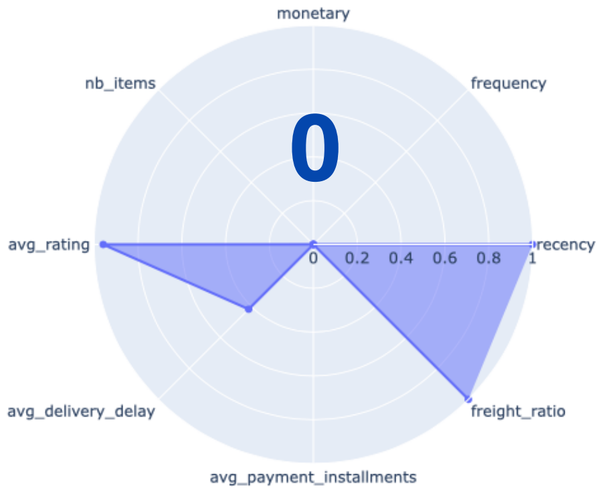
K-MEANS : SCORE DE SILHOUETTE



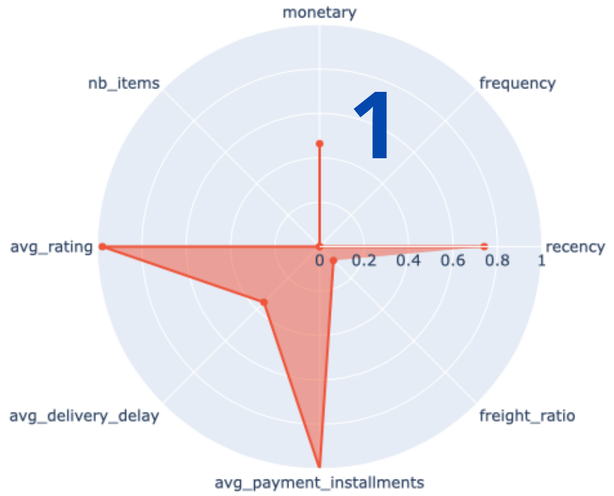
Test avec ACP :
non nécéssaire.

K-MEANS : ANALYSE DES CLUSTERS

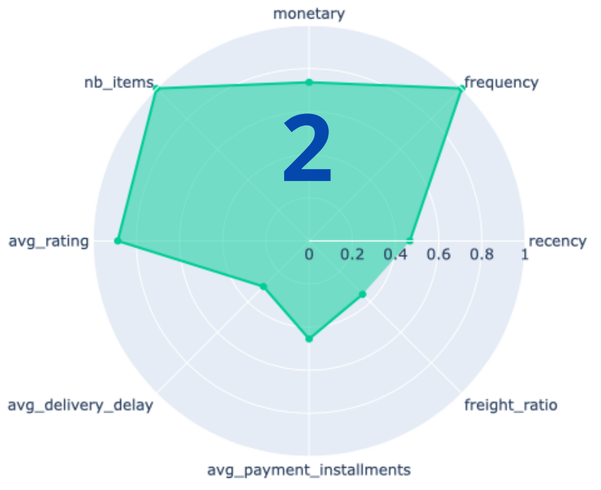
À RÉACTIVER SATISFAITS



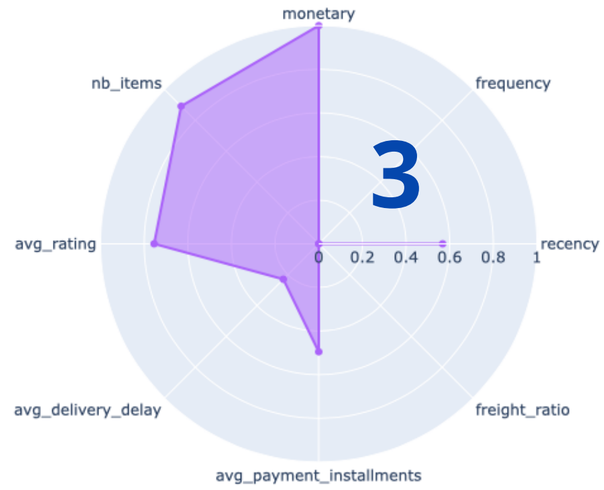
VERSEMENTS MULTIPLES SATISFAITS



LOYAUX SATISFAITS



VIP OCCASIONNELS

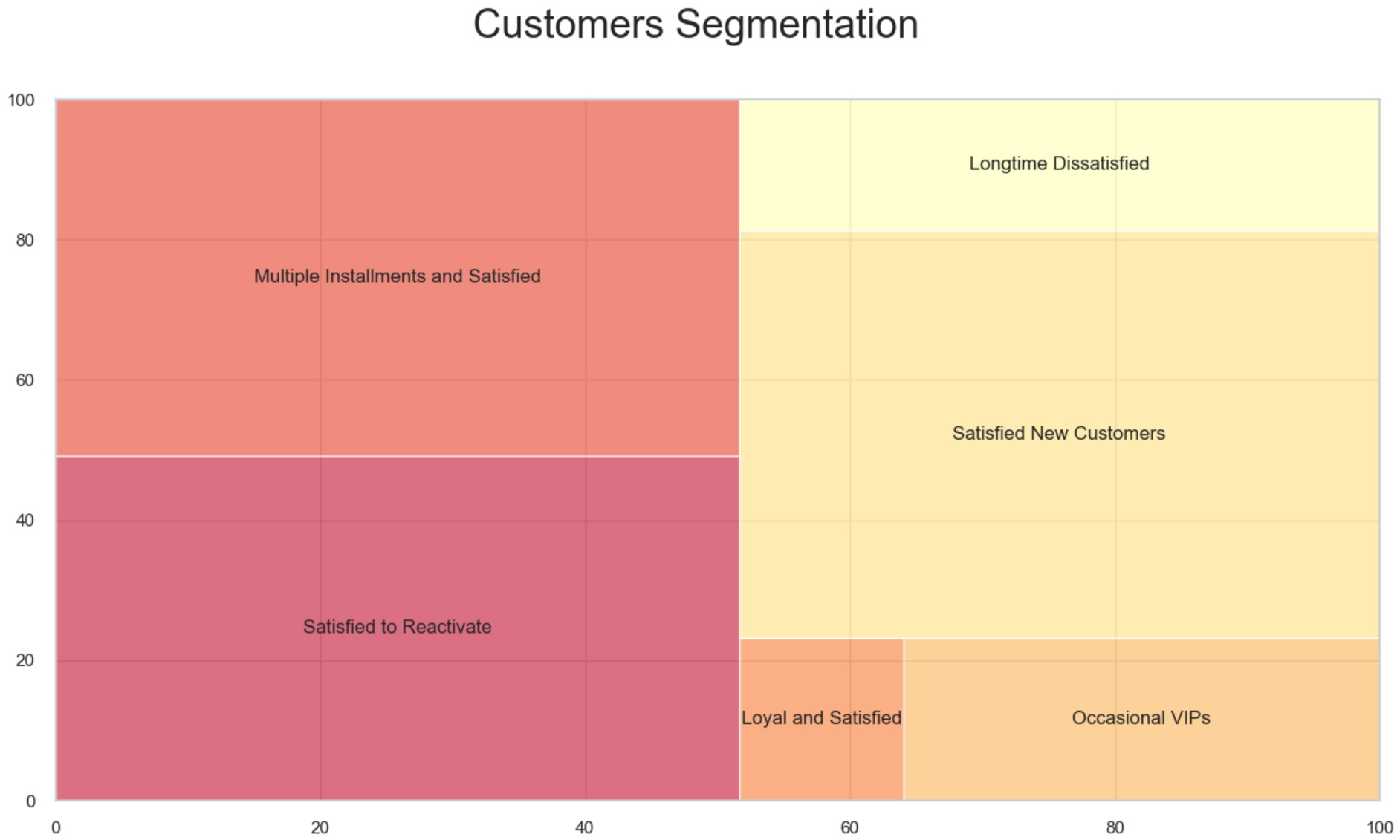


NOUVEAUX CLIENTS SATISFAITS

MÉCONTENTS ANCIENS

K-MEANS : ANALYSE DES CLUSTERS

K-MEANS : ANALYSE DES CLUSTERS



03.

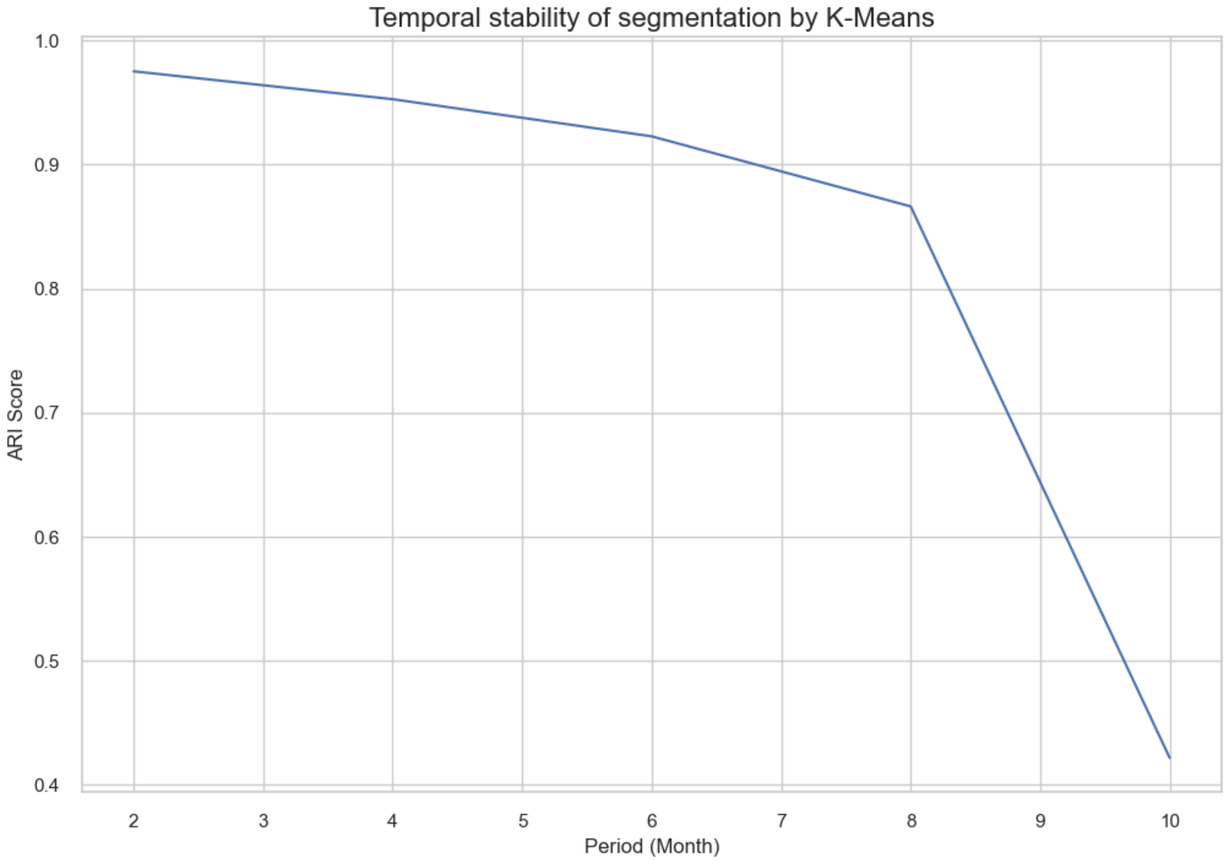
MAINTENANCE DU MODÈLE

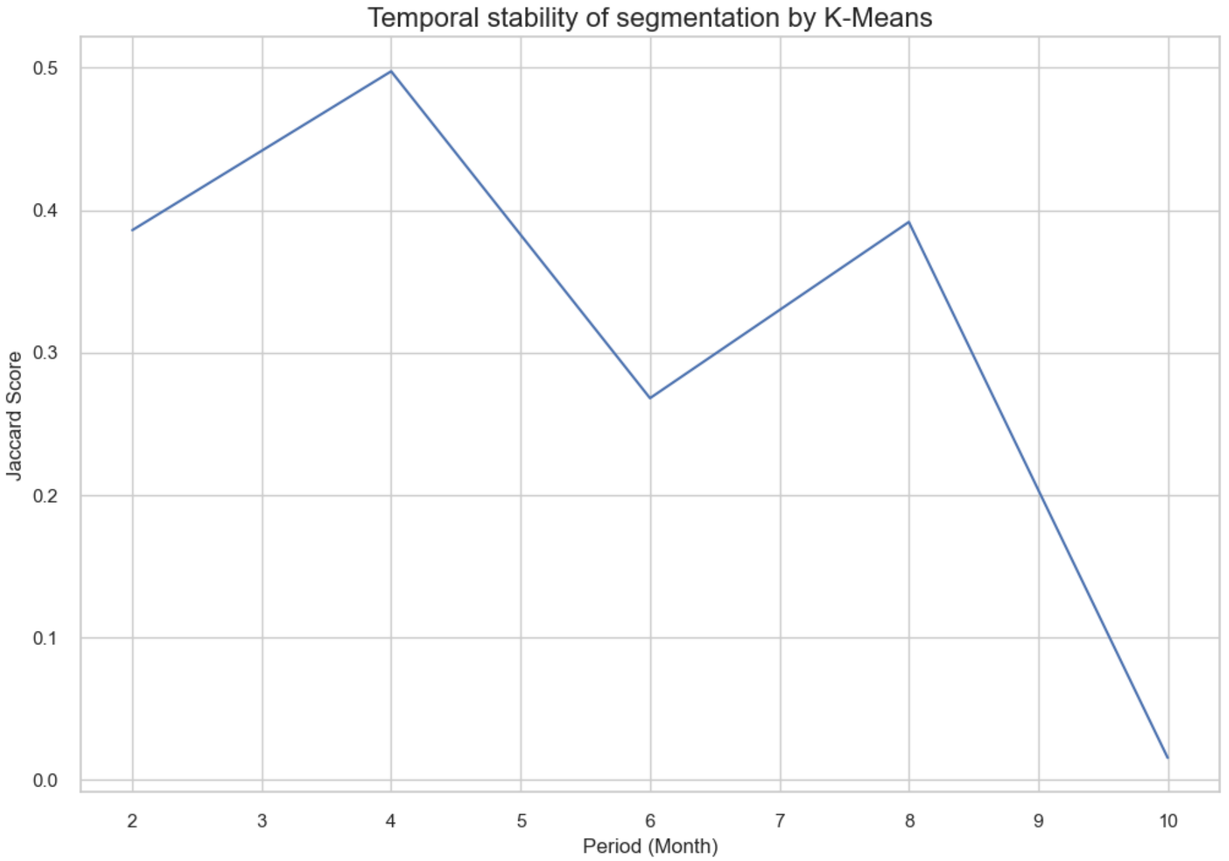
STABILITÉ À L'INITIALISATION

Initialization stability scores

Iteration	FitTime	Inertia	Homo	ARI	AMI
Iter 0	0.386s	343379	0.994	0.996	0.994
Iter 1	0.166s	343376	0.999	0.999	0.999
Iter 2	0.239s	343379	0.996	0.998	0.996
Iter 3	0.171s	362184	0.601	0.491	0.590
Iter 4	0.142s	343375	0.947	0.954	0.947
Iter 5	0.134s	343383	0.937	0.943	0.937
Iter 6	0.316s	343378	0.996	0.998	0.996
Iter 7	0.308s	343378	0.996	0.998	0.996
Iter 8	0.152s	343381	0.940	0.945	0.940
Iter 9	0.232s	343384	0.936	0.942	0.936

STABILITÉ DANS LE TEMPS





04.

CONCLUSION

CLUSTERING FINAL + ACTIONS MARKETING POSSIBLES

- **Cluster 0 : À réactiver satisfaits**
  - Réductions pour réactivation.
  - Recommandations personnalisées de produits.
- **Cluster 1 : Clients à versements multiples satisfaits**
  - Programme de fidélité.
  - Motivation par le parrainage.
- **Cluster 2 : Loyaux satisfaits**
  - Offres VIP exclusives.
  - Services d'abonnement.
- **Cluster 3 : VIP occasionnels**
  - Surclassement VIP.
  - Offres limitées dans le temps.
- **Cluster 4 : Nouveaux clients satisfaits**
  - Réduction de bienvenue.
  - Lots de produits.
- **Cluster 5 : Mécontents anciens**
  - Campagne de récupération.
  - Enquêtes de satisfaction client.

MAINTENANCE

Contrat tous les 8 mois.

MERCI !

Des questions ?