SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE

02/07/2024

Khoty WOLIE

SOMMAIRE

1. Problématique
2. Présentation des données
3. Nettoyage, étude des données et élaboration de nouvelles variables
4. Segmentation des clients
5. Maintenance du modèle

1. Problématique

Contexte : En tant que consultant pour Olist , une entreprise brésilienne spécialisée dans les solutions de vente sur les marketplaces en ligne, je suis chargé de fournir à ses équipes d'e-commerce une segmentation des clients qu'elles pourront utiliser au quotidien pour leurs campagnes de communication et publicitaires.

But : Fournir une segmentation précise des clients afin de mieux cibler les campagnes de communication et publicitaires.

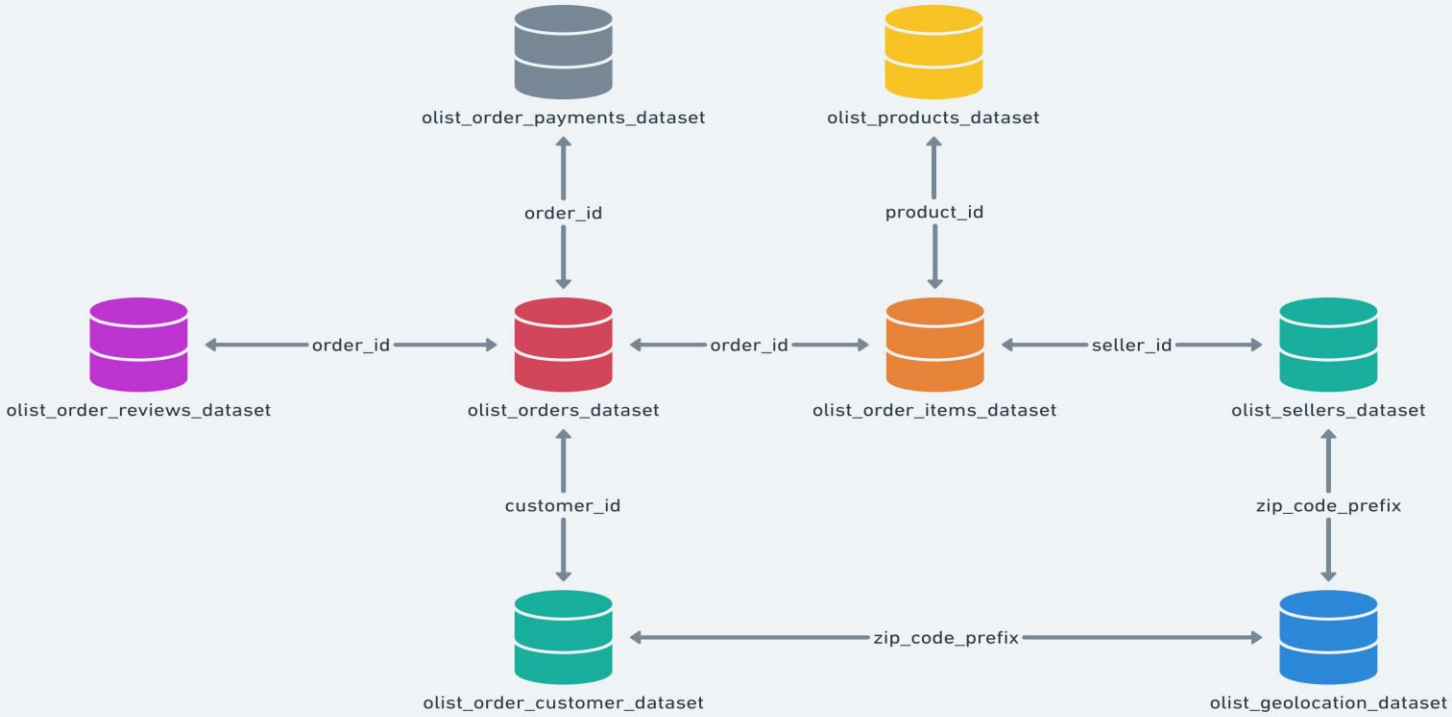
Objectif :

- Comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.
- Proposer une segmentation des clients adaptée aux campagnes publicitaires quotidiennes.
- Établir un contrat de maintenance pour assurer la mise à jour et la pertinence continue de cette segmentation.

2. Présentation des données

Les données : [Kaggle - Brazilian E-commerce Datasets](https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce)

- Données anonymisées couvrant la période d'octobre 2016 à août 2018.
- Olist met à disposition 8 DataFrames contenant diverses informations sur les clients, les vendeurs et les produits.



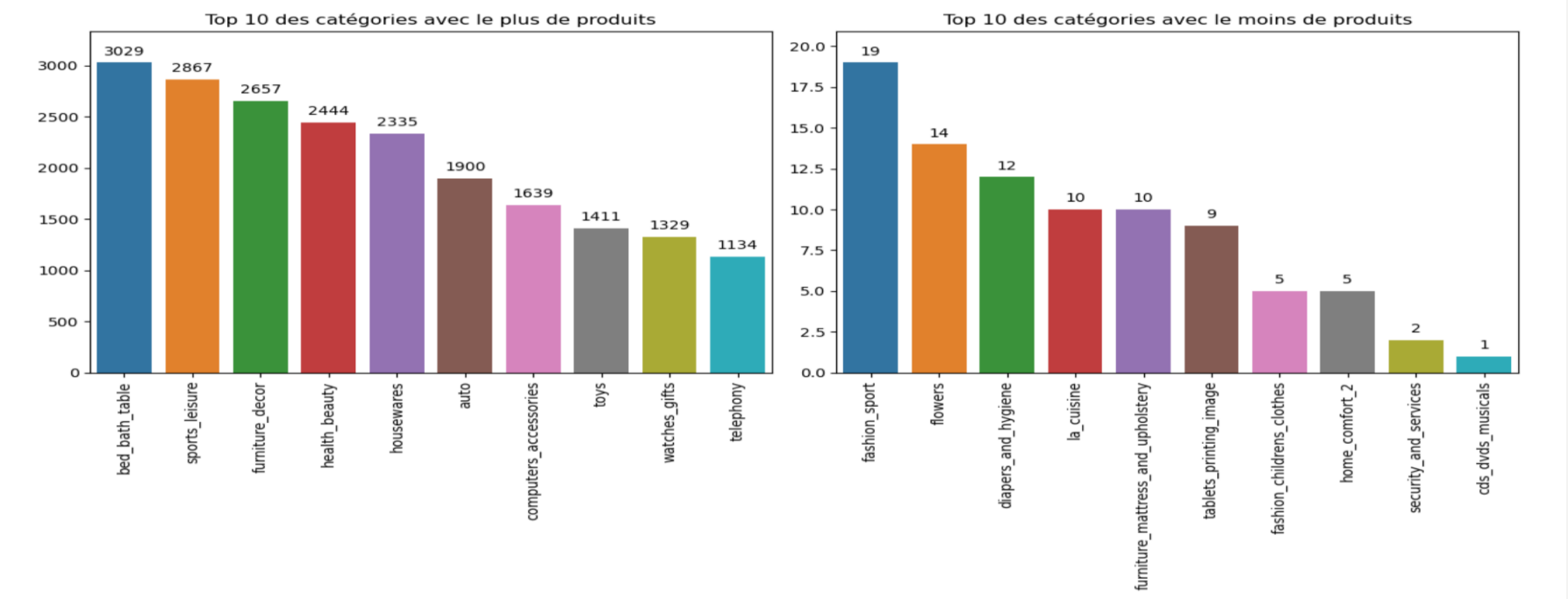
- ❑ 32 951 produits vendus
- ❑ 73 catégories de produits
- ❑ 99 441 commandes distinctes
- ❑ 99 224 commentaires
- ❑ 3 095 vendeurs
- ❑ 96 000 clients uniques

3. Nettoyage, étude des données et élaboration de nouvelles variables

3.1 Table sur les produits et la production :

Variables principales :

- ❑ Description détaillée de chaque produit
- ❑ Prix et frais de transport de chaque produit
- ❑ Catégorie de chaque produit
- ❑ Dimensions et poids des produits

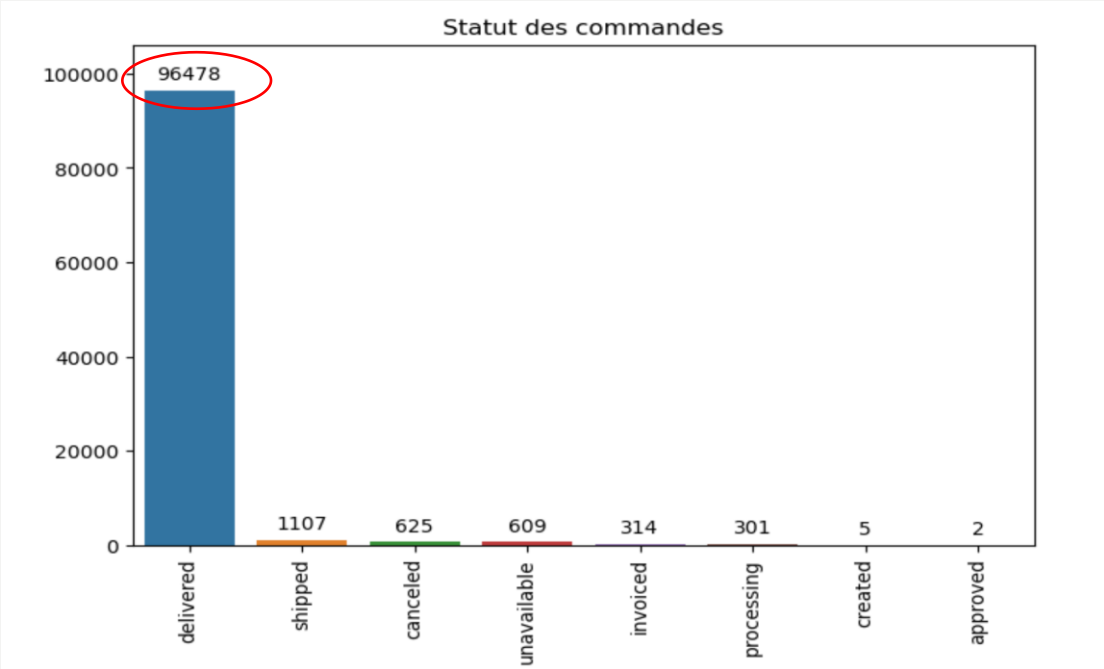


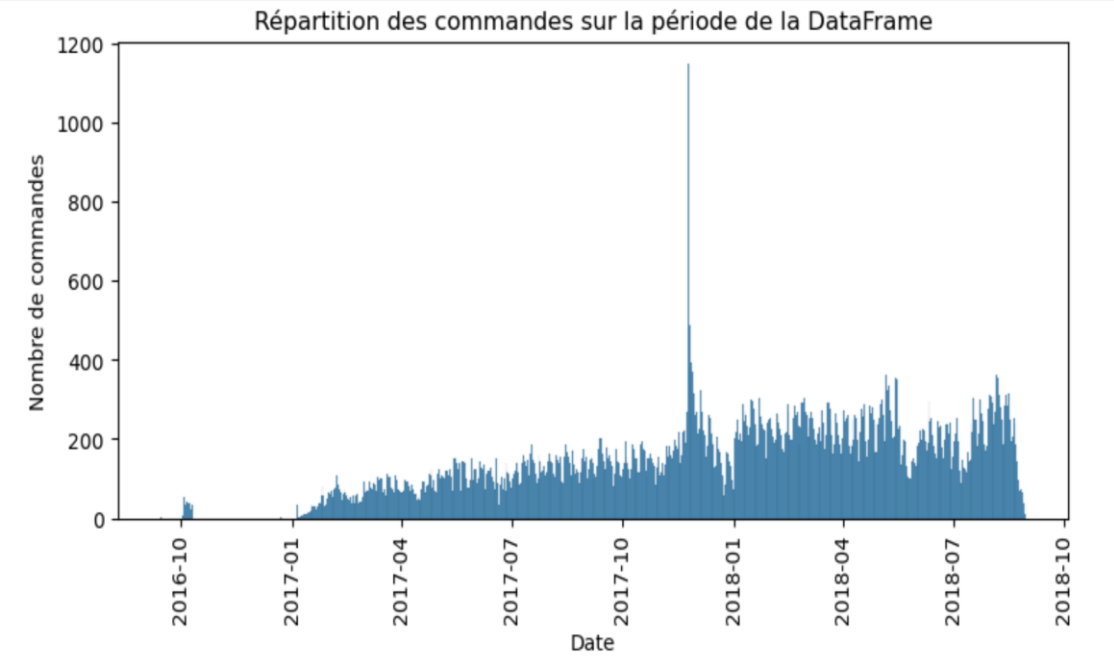
3.2 Table sur les commandes :

Variables principales :

- Statut des commandes, avec un focus sur les commandes livrées.

- Les dates de commande ainsi que les dates estimées et réelles de réception présentent un pic significatif observé en novembre et décembre 2017 en raison des fêtes de fin d'année.





- Élaboration de différentes variables temporelles :
  - Intervalle entre la commande et la réception en jours (temps de livraison).
  - Écart entre la date prévue de réception et la réception effective (retard de livraison).
  - Détermination de la date du dernier achat (en référence à 2019).
  - Diverses autres variables non retenues (heure, mois saison, jour).

jour_commande	1	2	3	4	5	6	7
mois_commande							
1	107	140	97	155	101	70	80
2	252	266	253	246	222	200	214
3	418	361	443	431	364	280	249
4	320	346	369	318	290	331	329
5	676	609	607	457	426	317	453
6	482	479	401	553	518	306	396
7	713	611	548	531	504	453	512
8	632	763	739	717	530	399	413
9	631	674	643	544	650	516	492
10	805	881	635	586	522	403	646
11	993	931	957	1046	1629	886	846
12	875	858	816	701	902	692	669

jour_commande	1	2	3	4	5	6	7
heure_commande							
0	145	141	162	151	214	174	123
1	62	61	89	82	125	78	77
2	28	45	40	36	36	33	36
3	11	15	14	12	23	27	20
4	8	15	15	15	25	12	13
5	5	10	14	15	17	13	17
6	31	28	33	30	40	24	14
7	78	88	84	90	88	46	35
8	190	223	211	204	224	103	95
9	300	344	349	297	351	191	156
10	444	391	395	410	454	282	225
11	444	475	435	412	453	336	301
12	402	412	377	396	387	308	314
13	446	467	440	473	473	280	307
14	468	460	442	409	454	304	306
15	460	464	426	397	428	320	326
16	460	477	444	397	424	324	338
17	401	426	399	398	366	320	361
18	381	360	361	328	342	278	381
19	404	394	346	348	398	338	395
20	428	411	401	370	323	331	417
21	491	451	406	384	368	288	371
22	454	439	353	370	365	245	392
23	363	322	272	261	280	198	279

3.3 Table sur les commentaires :

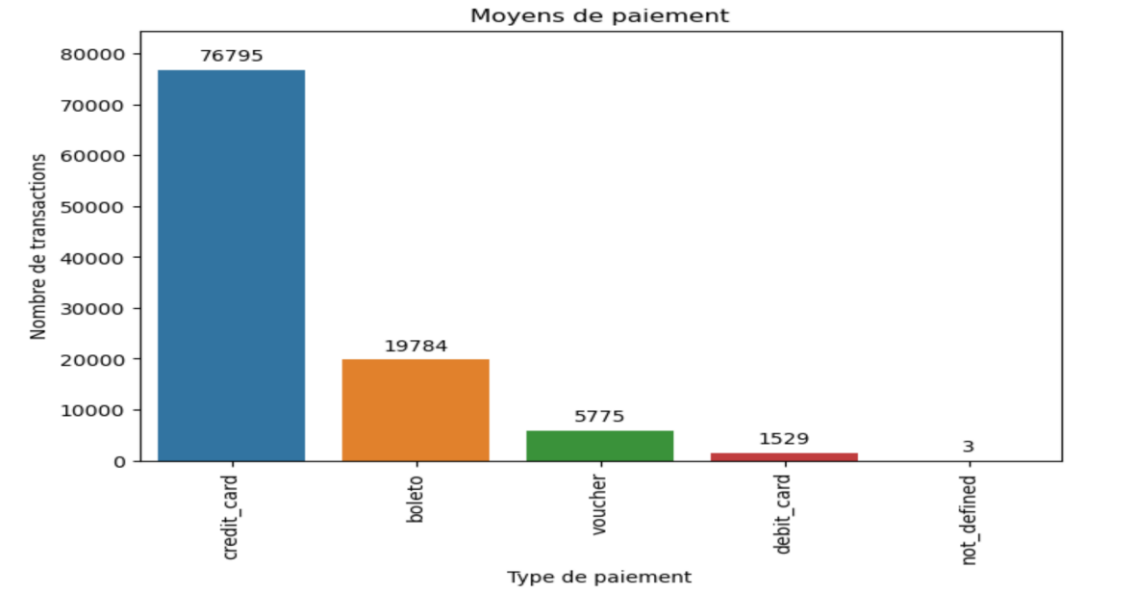
Variables principales :

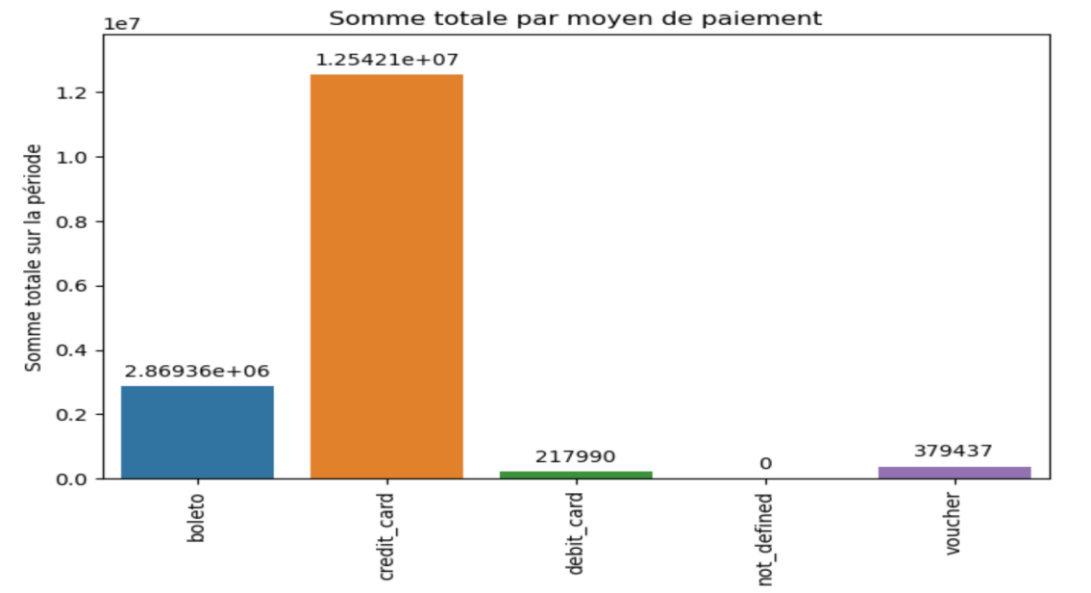
- ❑ Les scores commentaire attribuées par clients

3.4 Table sur les paiements :

Variables principales :

- ❑ Le nombre de paiements
- ❑ Le montant de chaque paiement
- ❑ Le mode de paiement (carte bancaire, billets, bons d'achat)

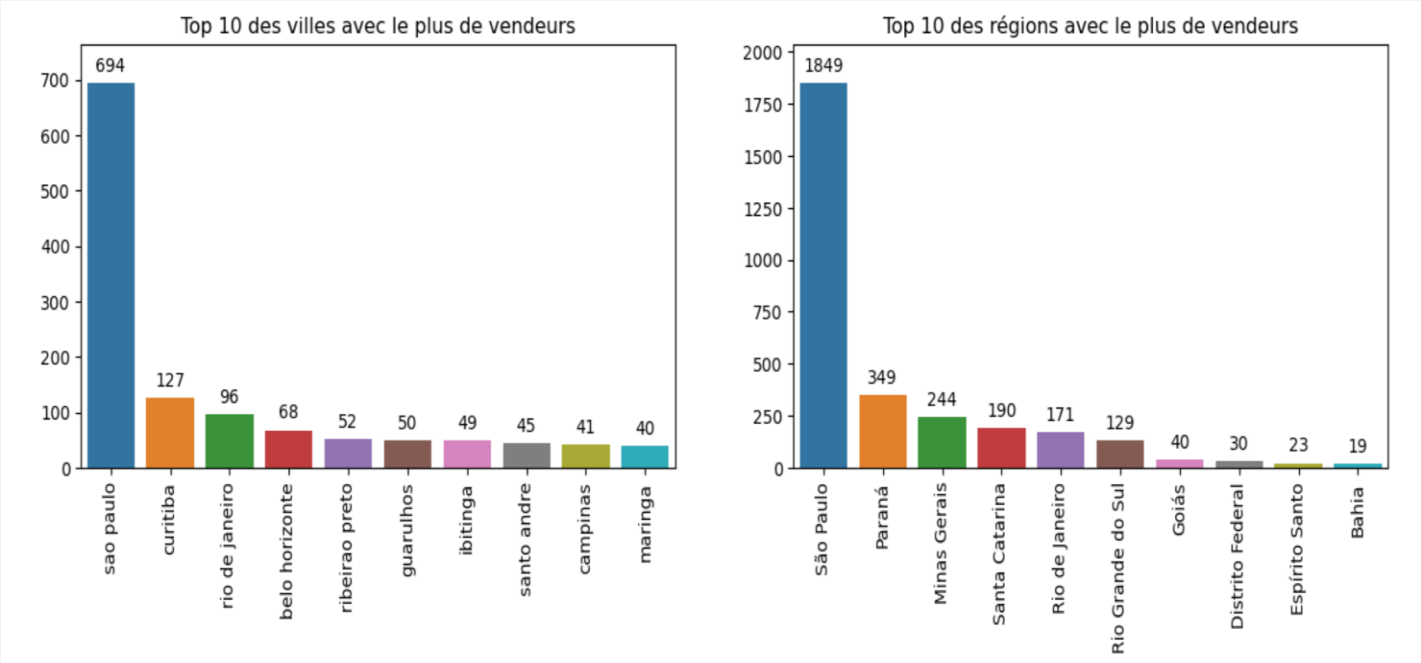


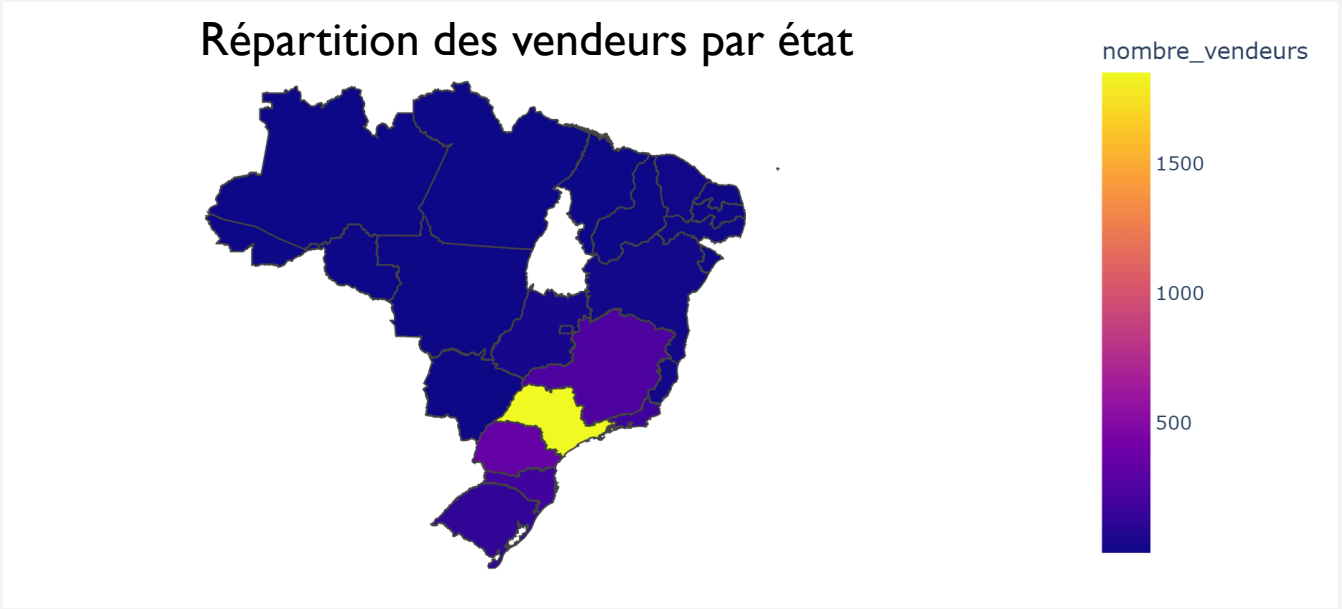


3.5 Table sur les vendeurs :

Variables principales :

- ❑ La ville des vendeurs
- ❑ La région des vendeurs





3.6 Table sur la géolocalisation :

Variables principales :

- ❑ Les coordonnées géographiques
- ❑ Les villes
- ❑ Les états

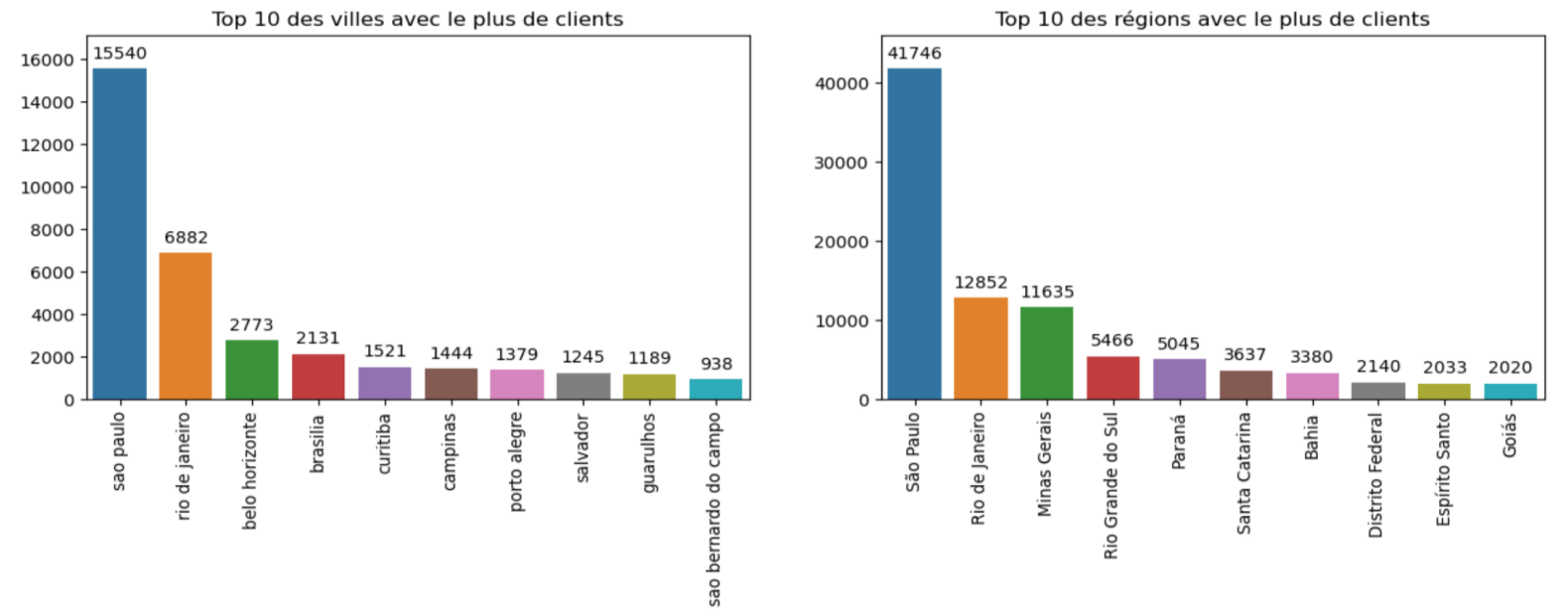
lat_min, lat_max (Brésil) = -33.7500, 5.2725

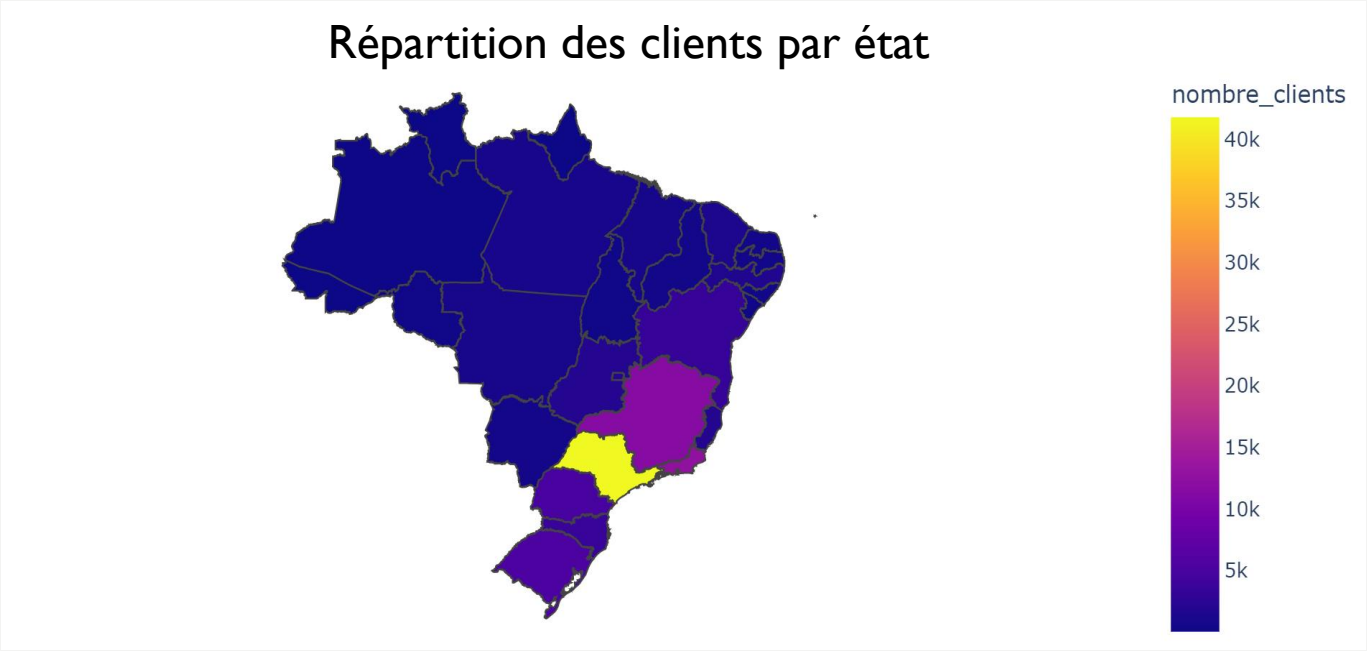
lng_min, lng_max (Brésil) = -73.9831, -34.7939

3.7 Table sur les clients :

Variables principales :

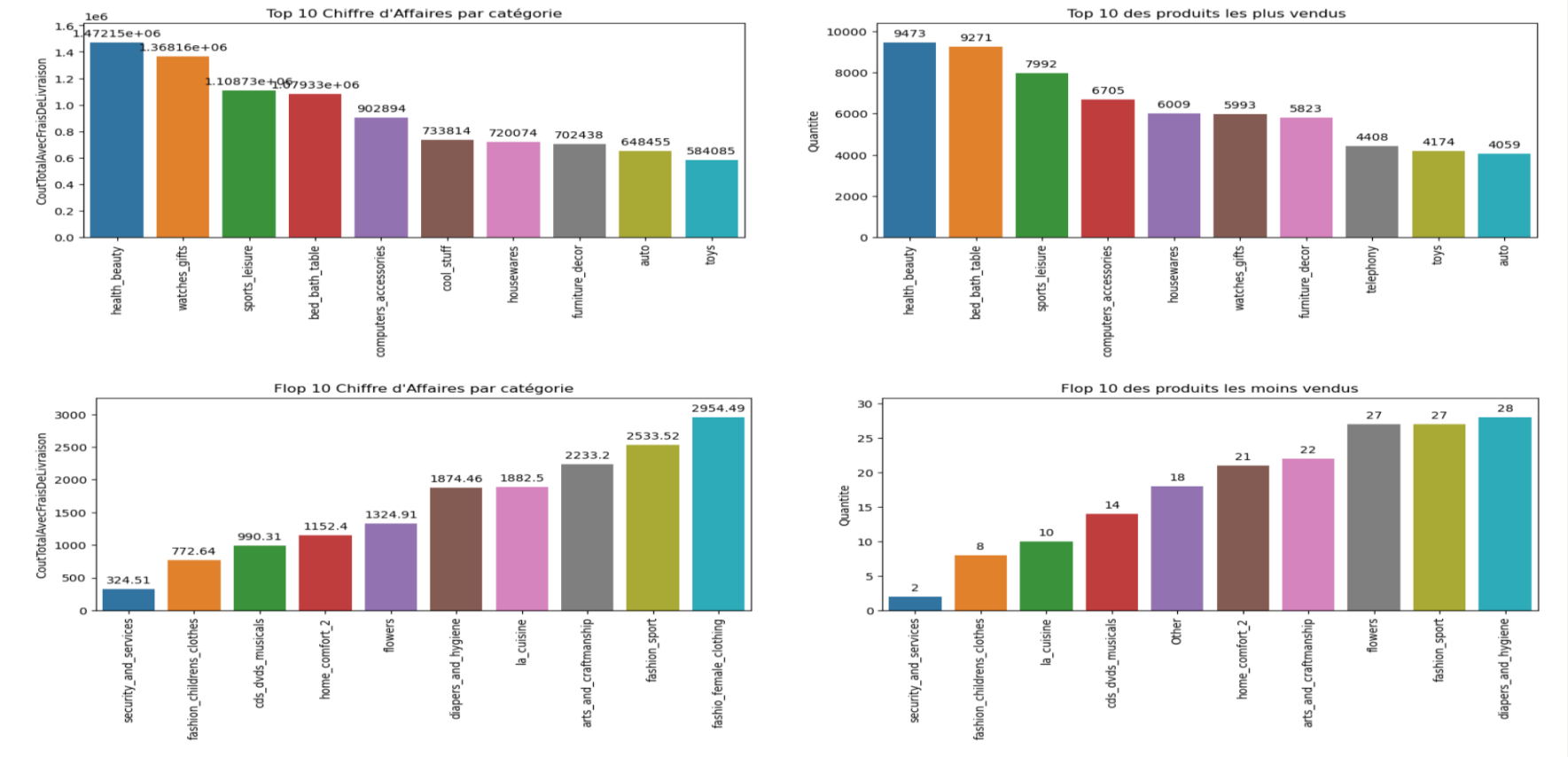
- ❑ La ville des clients
- ❑ La région des clients



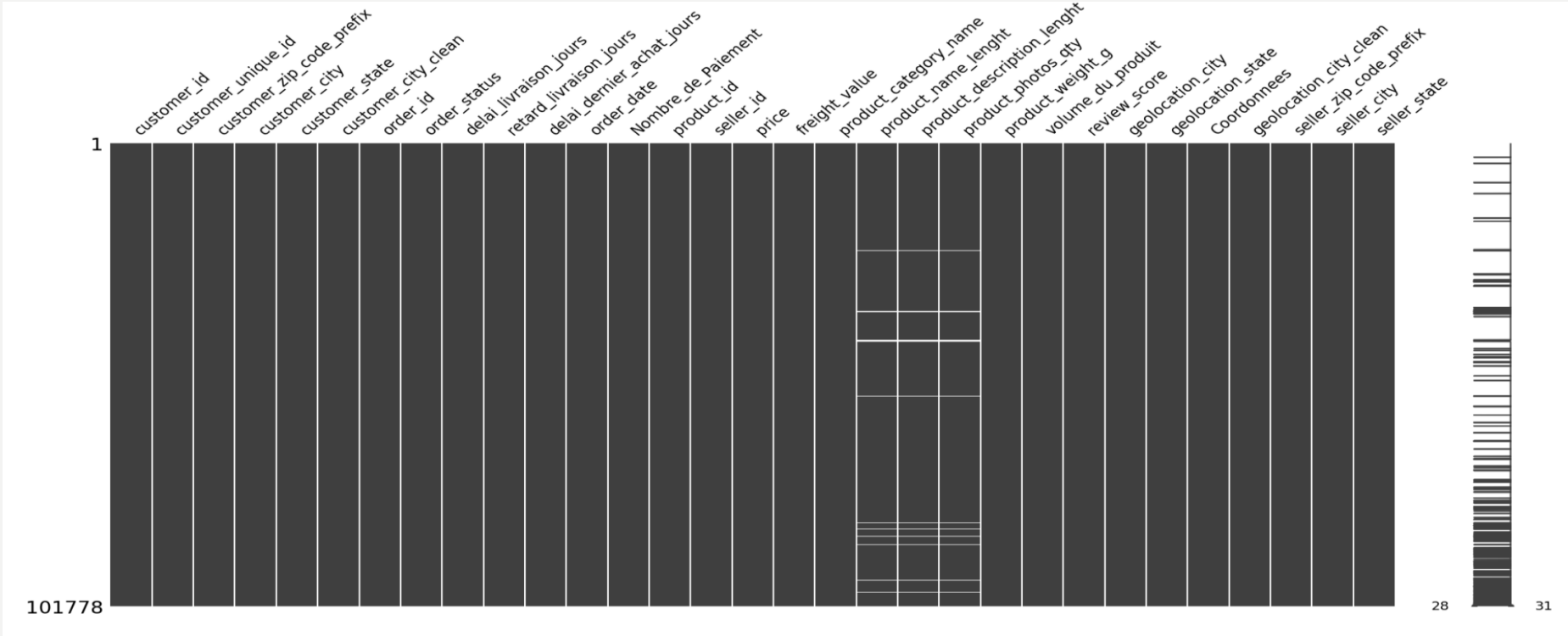


3.8 Information complémentaire

- Les catégories **health_beauty**, **watches_gifts**, et **sports_leisure** se distinguent nettement par leur chiffre d'affaires élevé et leurs volumes de ventes importants. Cela indique une forte demande et une performance commerciale robuste dans ces segments.
- En revanche, des catégories comme **security_and_services**, **fashion_childrens_clothes**, et **cds_dvds_musicals** affichent des chiffres d'affaires et des volumes de ventes très bas, indiquant soit une faible demande soit une faible performance commerciale.



3.9 Jointure des différentes tables :



Processus finals :

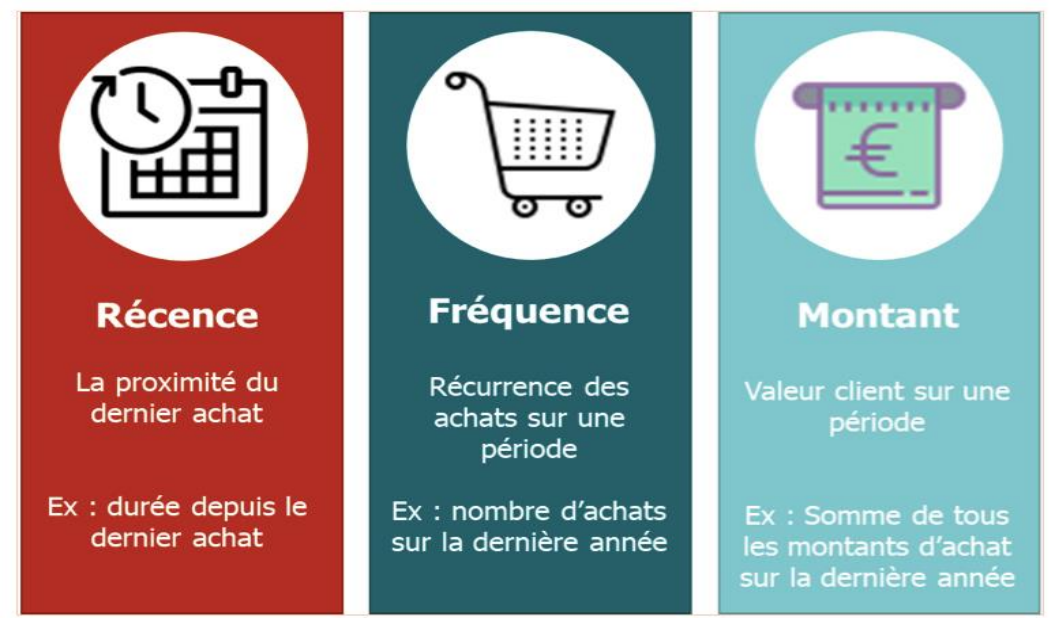
- Suppression des lignes contenant des valeurs manquantes (**NaNs**).
- Élimination des variables superflues (noms de villes, descriptions des produits, coordonnées géographiques, etc.).
- Ajout d'une colonne **`CoutTotalAvecFraisDeLivraison`** pour le coût total incluant les frais de livraison.
- Agrégation par `Customer_Unique_Id` pour garantir un seul client par ligne.
- Transformation logarithmique (log(x+1)) de la variable **`CoutTotalAvecFraisDeLivraison`** pour réduire l'étalement des données.

4. Segmentation des clients

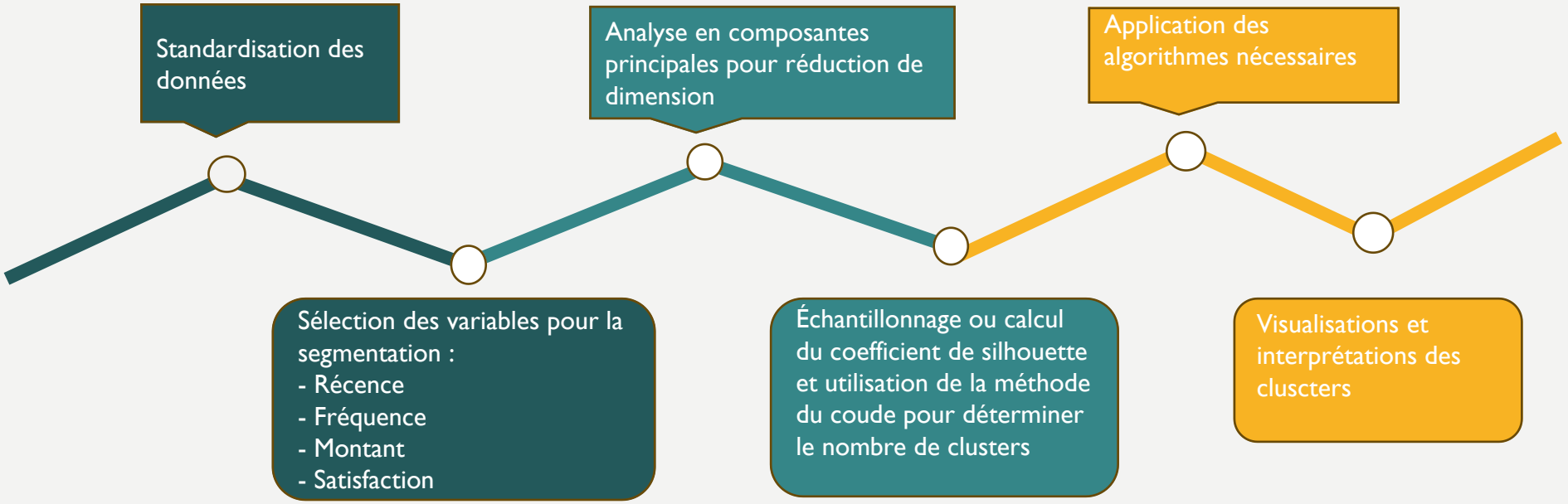
4.1 Modélisation RFM

Variables retenues :

- `**Delais_dernier_achat_jours**` : Récence (R)
- `**Nombre_de_commande**` : Fréquence (F)
- `**CoutTotalAvecFraisDeLivraison(log)**` : Montant (M)

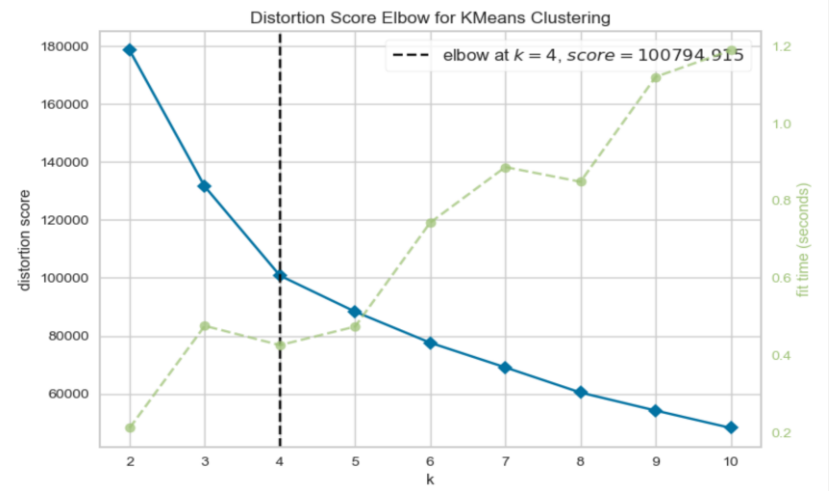


Processus de segmentation



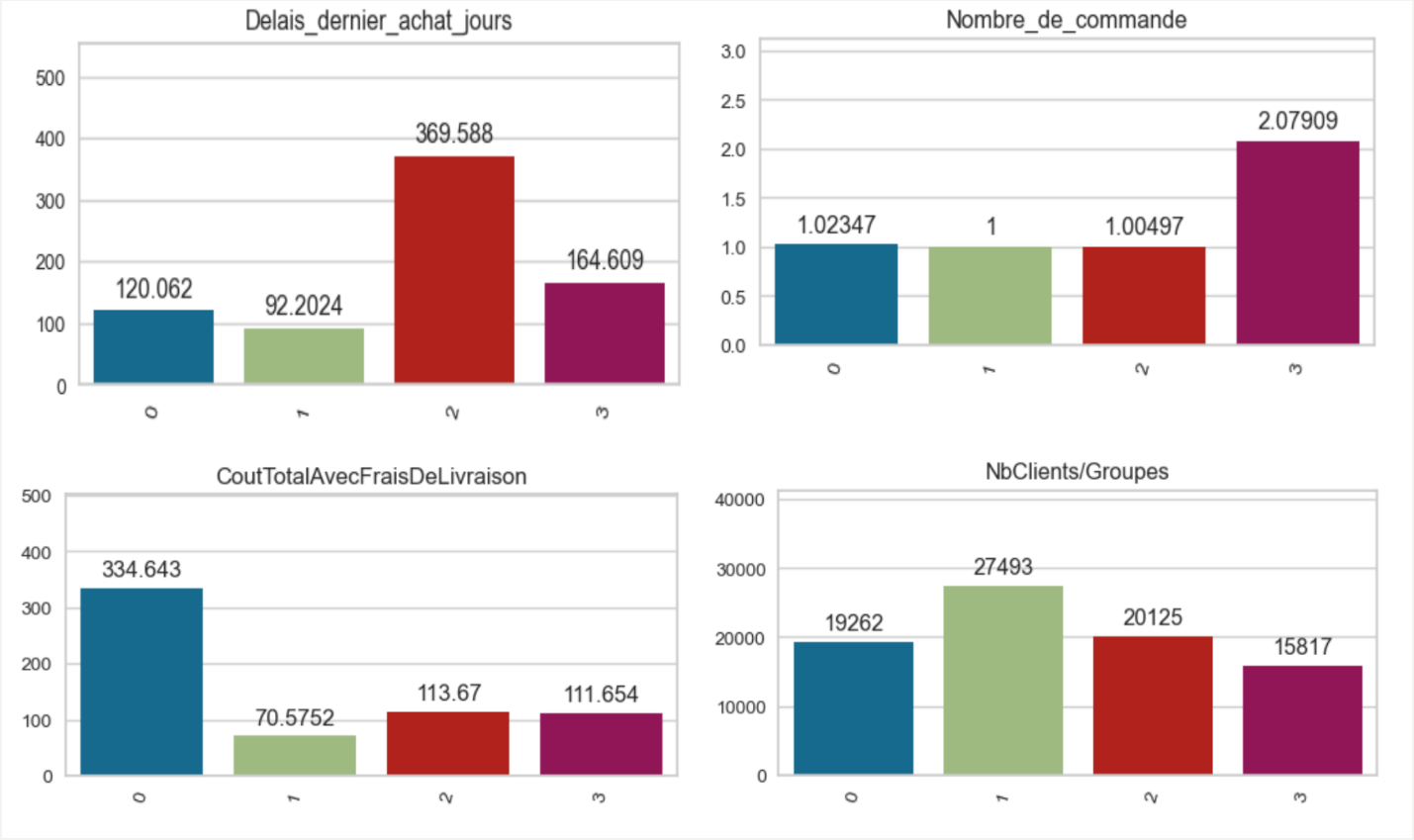
Résultats :

- Nombre de clusters : 4
- Statistiques :



Résultats du test statistique de Tukey :
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	-0.6688	0.0	-0.6816	-0.6561	True
0	2	0.0976	0.0	0.0839	0.1113	True
0	3	0.3842	0.0	0.3696	0.3987	True
1	2	0.7664	0.0	0.7538	0.779	True
1	3	1.053	0.0	1.0395	1.0665	True
2	3	0.2866	0.0	0.2722	0.301	True



Observation:

Groupe 0 : Clients qui dépensent beaucoup mais achètent rarement.

Groupe 1 : Clients récents avec des dépenses modérées et un nombre élevé.

Groupe 2 : Clients anciens qui n'ont pas effectué d'achat récent et ont des dépenses moyennes.

Groupe 3 : Clients fréquents acheteurs avec des dépenses moyennes.

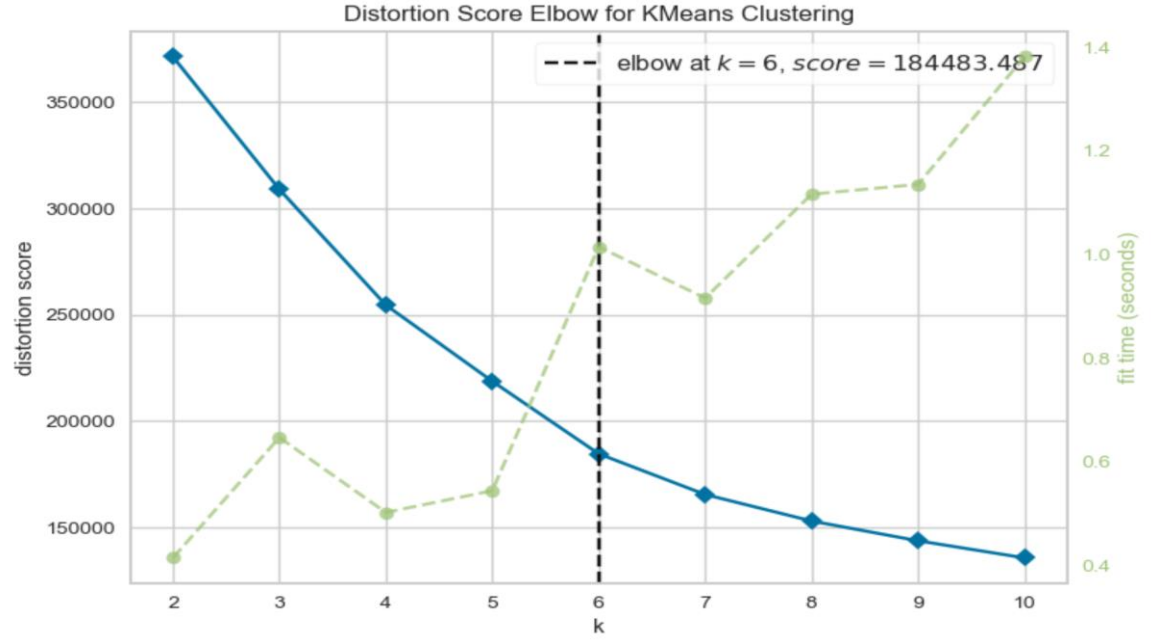
4.2 Modélisation RFM + Score + Retard livraison + Nombre de paiement

Variables retenues :

- **`Delais_dernier_achat_jours`** : Récence (R)
- **`Nombre_de_commande`** : Fréquence (F)
- **`CoutTotalAvecFraisDeLivraison(log)`** : Montant (M)
- **`ScoreCommentaireMoyen`**
- **`Retard_livraison_jours`**
- **`Nombre_de_Paiement`**

Résultats du test statistique de Tukey :
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	0.4685	0.0	0.4557	0.4813	True
0	2	0.1904	0.0	0.178	0.2027	True
0	3	-0.2891	0.0	-0.2994	-0.2787	True
0	4	-0.3514	0.0	-0.3657	-0.3371	True
0	5	1.0834	0.0	1.0379	1.129	True
1	2	-0.2781	0.0	-0.292	-0.2642	True
1	3	-0.7576	0.0	-0.7697	-0.7455	True
1	4	-0.8199	0.0	-0.8355	-0.8043	True
1	5	0.6149	0.0	0.5689	0.6609	True
2	3	-0.4794	0.0	-0.4911	-0.4678	True
2	4	-0.5418	0.0	-0.557	-0.5265	True
2	5	0.893	0.0	0.8472	0.9389	True
3	4	-0.0623	0.0	-0.076	-0.0487	True
3	5	1.3725	0.0	1.3271	1.4178	True
4	5	1.4348	0.0	1.3884	1.4812	True



Le point de coude à (k = 6) indique qu'au-delà, la réduction du score de distorsion devient marginale. Ainsi, 6 clusters sont considérés comme optimaux pour ce jeu de données.

Observations :

Groupe 0 : Clients les plus lointains qui n'ont pas effectué d'achat récemment, mais satisfaits.

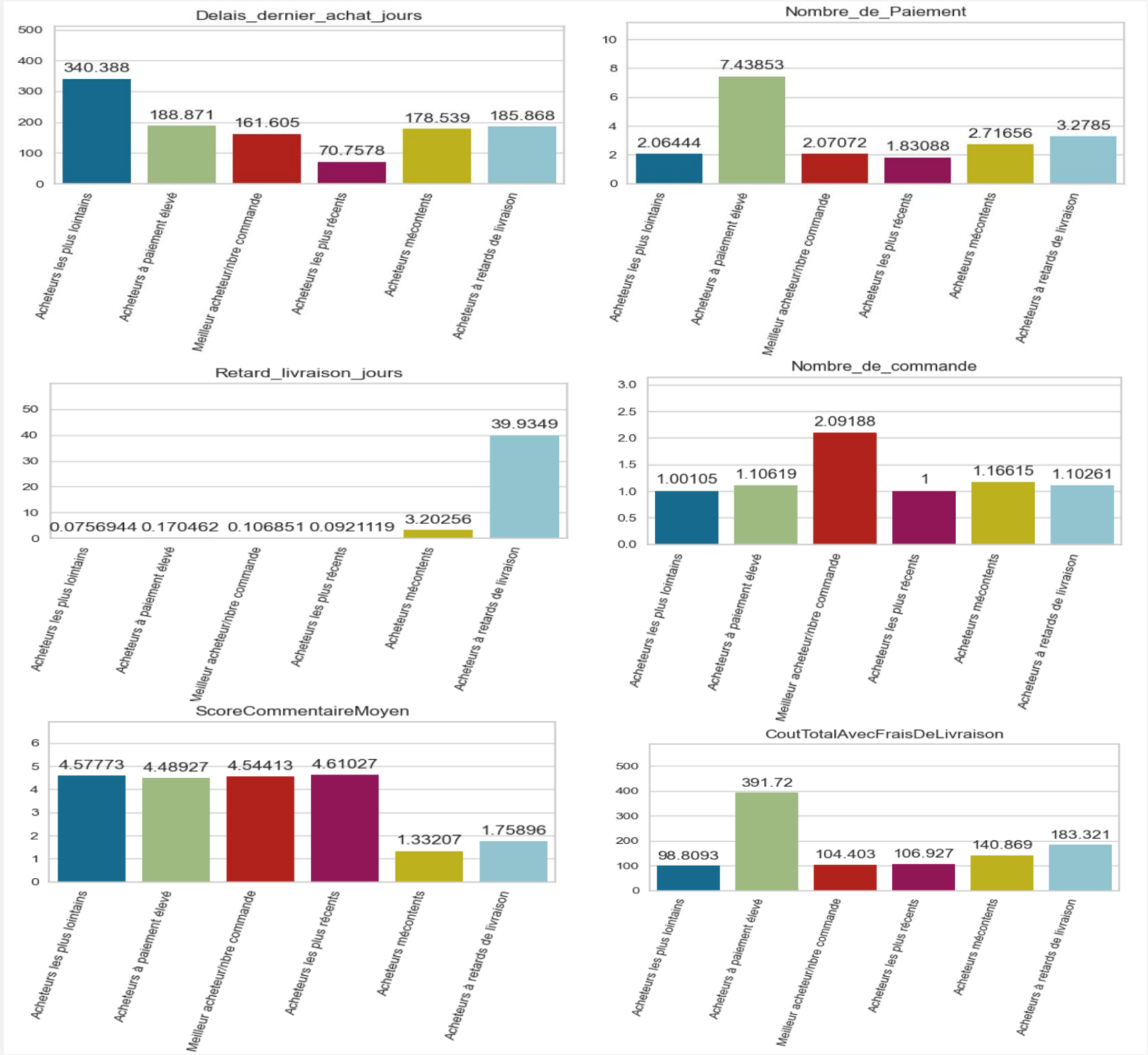
Groupe 1 : Clients à paiement élevé, avec un retard de livraison très faible et une bonne satisfaction.

Groupe 2 : Meilleurs acheteurs par nombre de commandes, avec une forte satisfaction.

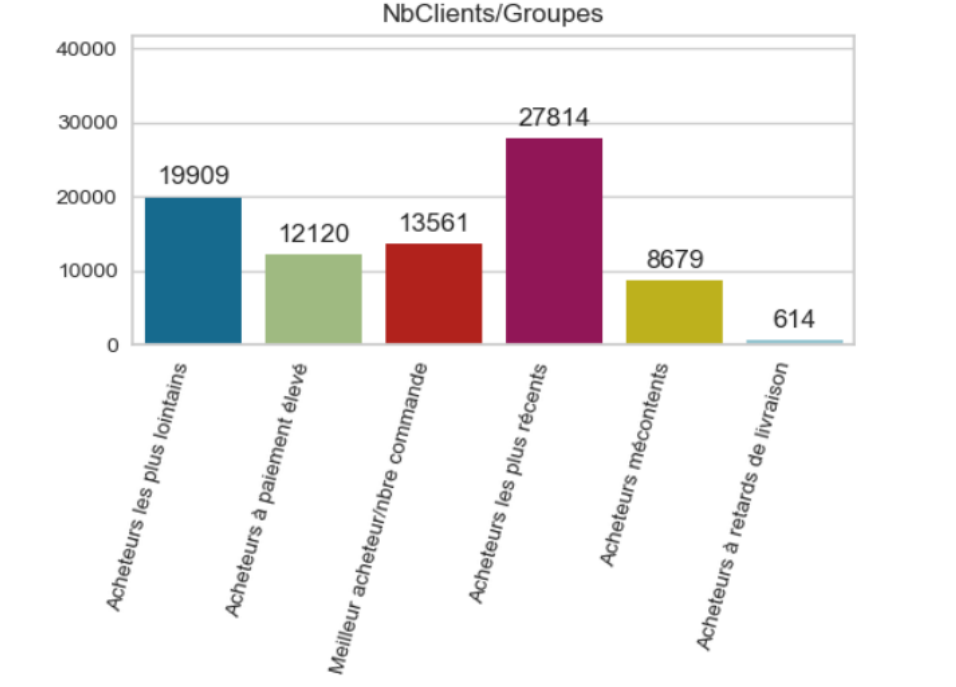
Groupe 3 : Acheteurs les plus récents, avec la plus haute satisfaction.

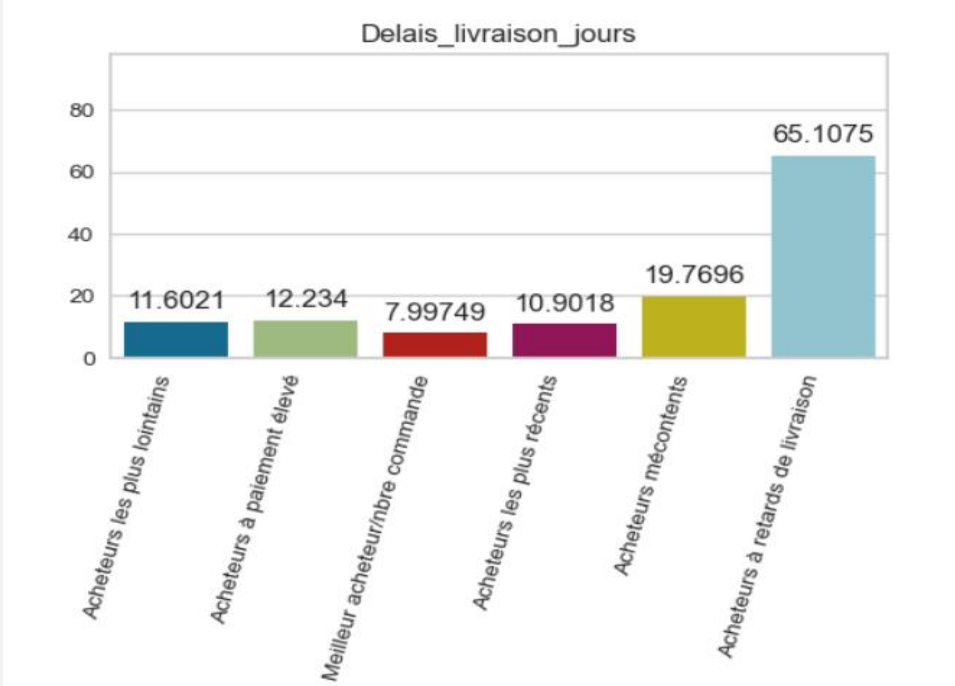
Groupe 4 : Acheteurs insatisfaits avec un retard de livraison relativement élevé.

Groupe 5 : Acheteurs à retards de livraison, avec une satisfaction faible et un coût total élevé.



Information complémentaire





5. Maintenance du modèle

5.1 Procédure pour établir la fréquence de maintenance du modèle

Modèle Initial M0

- **Période T0** : Les 12 premiers mois de données
- **Variables** : 4 (R, F, M, S)
- **Standardisation** : Application de StandardScaler
- **Clustering** : Algorithme K-means
- **Évaluation** : Indice de Rand Ajusté (ARI)

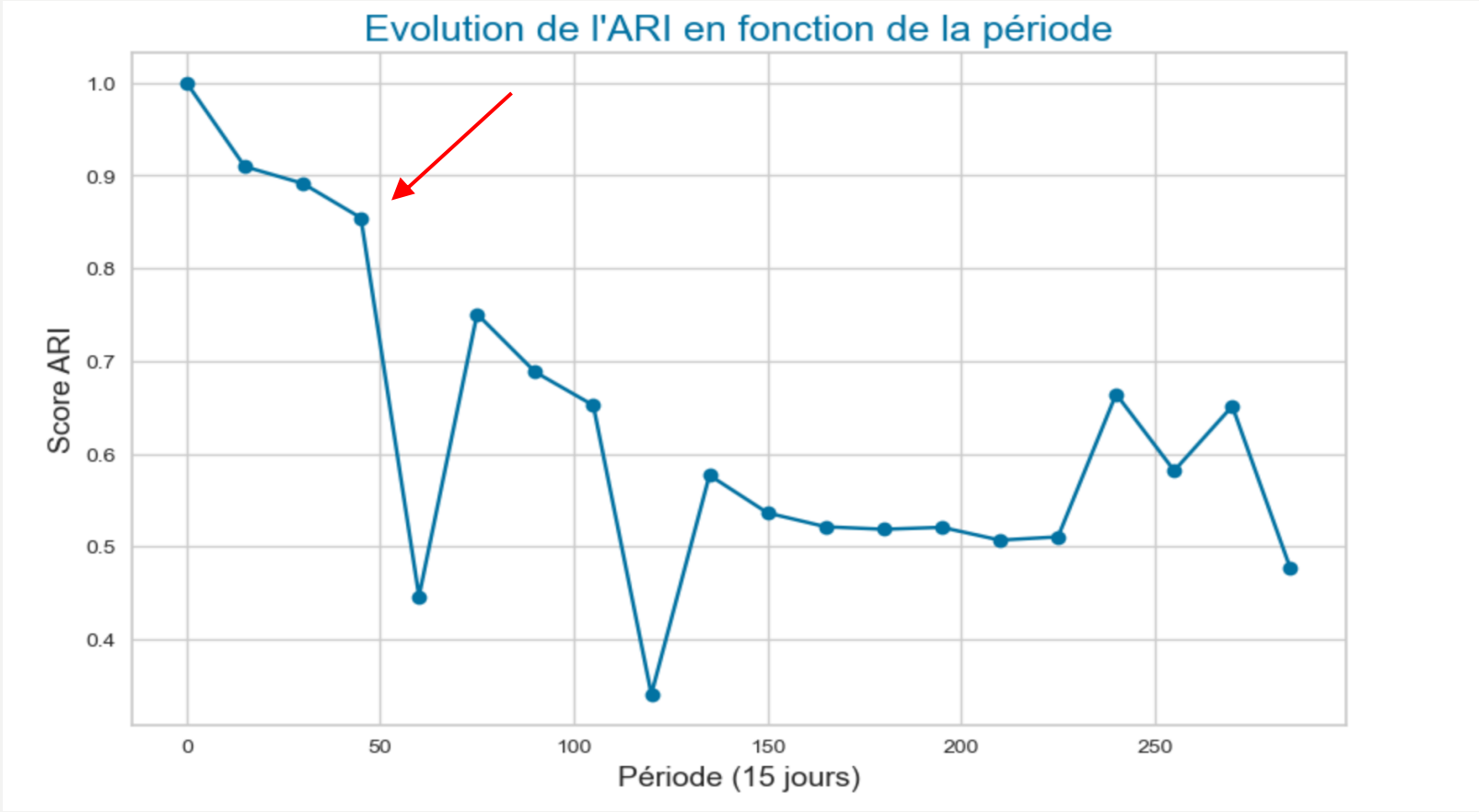
Calcul de Segmentation

- **Fréquence** : Tous les 15 jours
- **Période F1** : 12 mois avec des intervalles de 15 jours de données, etc.
- **Variables** : 4 (R, F, M, S)
- **Standardisation** : Application de StandardScaler
- **Modèle** : Utilisation du modèle 0 sur F1
- **Clustering** : Algorithme K-means sur F1 pour les périodes M0 et M1

Comparaison : Indice de Rand

- Évaluation des résultats de M0 sur F1 par rapport à ceux de M1 sur F1, et ainsi de suite.

5.2 Quelle est la durée de stabilité de la segmentation ?



Il serait intéressant de proposer une nouvelle segmentation au client après 45 jours.

Observation

On remarque une stabilité du score ARI jusqu'à environ 45 jours, suivie d'une chute marquée. Après environ 50 jours, il y a une légère remontée, mais la tendance générale est à la baisse continue avec des fluctuations autour de 150 jours. Vers 225 jours, on observe un rebond, suivi de nouvelles fluctuations, mais la tendance générale continue de baisser jusqu'à 300 jours. Cette diminution constante pourrait s'expliquer par les cycles de commande observés dans l'analyse exploratoire.

Recommandations

- **Surveillance Continue :** Continuez à surveiller le score ARI pour détecter toute tendance à la baisse ou fluctuation anormale.
- **Réévaluation Périodique :** Réévaluez le modèle régulièrement et envisagez de le réentraîner si des baisses significatives du score ARI sont observées sur plusieurs périodes consécutives.
- **Analyse des Cycles :** Effectuez une analyse approfondie des cycles de commande pour comprendre comment ils affectent les performances du modèle et ajustez les paramètres de clustering en conséquence.
- **Amélioration des Données :** Travaillez à améliorer la qualité des données en nettoyant régulièrement les anomalies et les erreurs potentielles.