SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE

« ANALYSE EXPLORATOIRE ET MODÉLISATION »

Samira MAHJOUB _ 06/2024

SOMMAIRE

PROBLÉMATIQUE - CONTEXTE - OBJECTIFS

- **Moyens mis à disposition:**
  Base de données anonymisée avec historique des commandes, produits achetés, commentaires de satisfaction, et localisation des clients (9 fichiers CSV)

- **Contexte :** *Olist souhaite exploiter les données clients pour améliorer la satisfaction et l'efficacité des équipes Marketing.*

- **Problématique:** *Identifier et segmenter les différents types de clients pour optimiser les campagnes de communication.*

- **Objectif d'Olist :**

- ***Fournir une segmentation des clients basée sur leurs comportements et données personnelles***

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

- **Missions:** *Analyser les données, réaliser des segmentations clients exploitables et proposer un contrat de maintenance pour la mise à jour régulière de la segmentation.*

Apprentissage non supervisé

Consiste à :

- *Inférer des connaissances sur les données sur la seule base des échantillons d'apprentissage.*
- *Pas de cible, recherche de structures naturelles dans les données.*

JEU DE DONNÉES

Tableau récapitulatif basé sur les données

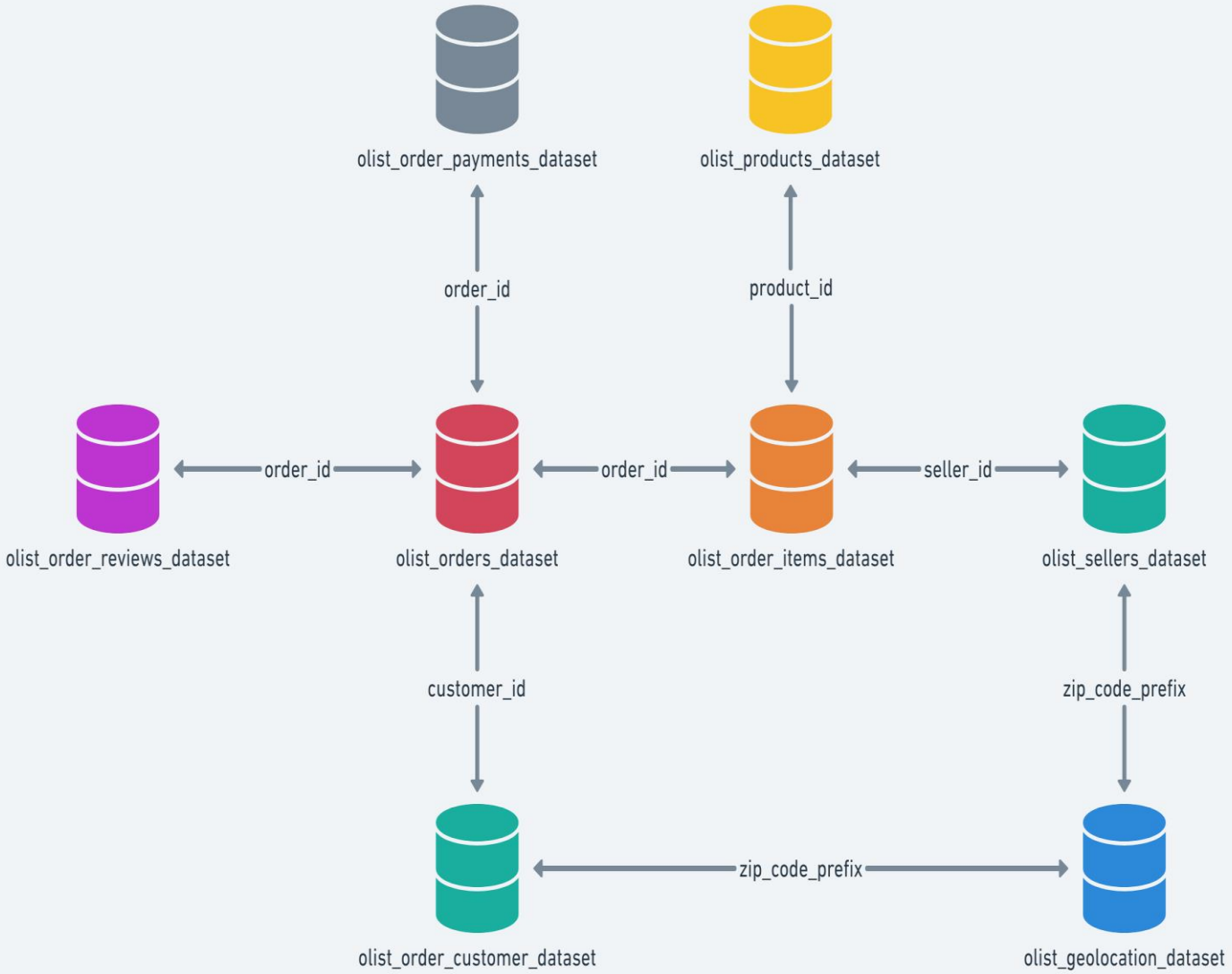
Nombre de fichiers : **9 fichiers de données brutes**

Données structurelles : nombre de commandes, détails des produits, avis des clients, ...

Données de localisation : adresse, quartier, coordonnées Lat./Long., code postal, ...

Données financières : détails des paiements, prix des produits, frais de port, ...

Données des utilisateurs : informations clients, préférences, catégories de produits, ...



SQL POUR L'EXTRACTION ET L'ANALYSE DES DONNÉES

Pourquoi SQL ?

- *Extraction des données brutes depuis la base de données.*
- *Nettoyage et transformation des données pour l'analyse.*
- *Accès à des informations spécifiques rapidement (comme les produits les plus vendus).*

Ce tableau résume les résultats des requêtes SQL effectuées pour analyser les commandes récentes avec retard, les vendeurs ayant généré un chiffre d'affaires élevé, les nouveaux vendeurs engagés et les codes postaux avec les scores moyens les plus bas.

Ces informations sont essentielles pour la segmentation des clients et l'optimisation des stratégies marketing.

Requêtes et Résultats Essentiels :

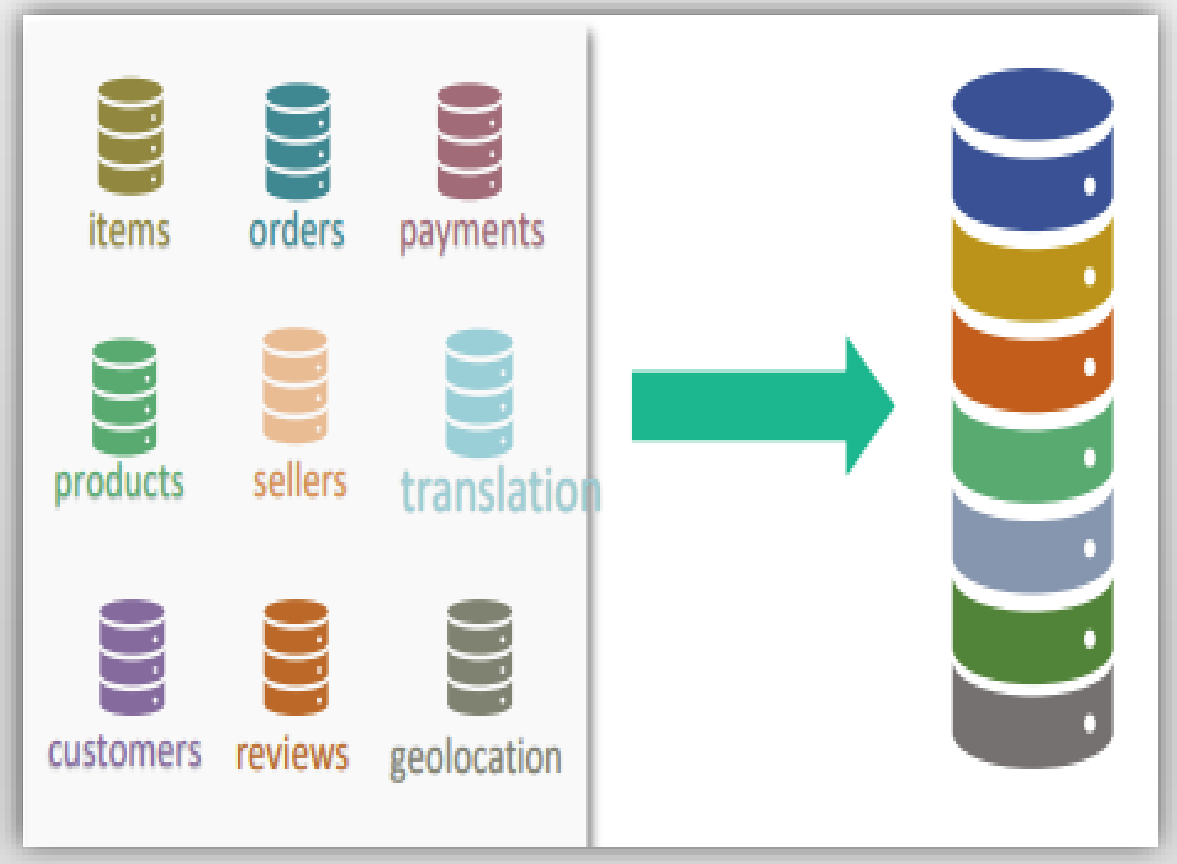
Question	Résultat
Commandes récentes avec retard	order_id
	123456
	789012
Vendeurs ayant généré un chiffre d'affaires de plus de 100K	seller_id
	seller123
	seller456
Nouveaux vendeurs engagés	seller_id
	seller789
	seller012
Codes postaux avec les scores moyens les plus bas	customer_zip_code_prefix
	22753
	22770
	13056
	22793
	21321

DÉMARCHE A SUIVRE

Étape	Description
Métier	- 9 jeux de données séparés - Plusieurs lignes par client - Population hétérogène
Analyse et Nettoyage	- **Fusion des données** - **Valeurs Manquantes** : Suppression des valeurs nulles - **Doublons** : Pas de doublons identifiés - **Outliers** : Vérification des valeurs aberrantes et suppression si nécessaire.
Feature Engineering	- Standardisation des variables quantitatives - Encodage des variables catégorielles - **Création de nouvelles variables** - **Transformation des variables existantes** - **Transformation logarithmique**
Modélisation	- Algorithmes d'apprentissage non supervisé - Segmentation RFM - Validation des modèles - Optimisation des clusters
Interprétation	- Connaissance client améliorée - Segmentation intelligente - Analyse descriptive des profils
Rapport	- Génération de rapports détaillés - Visualisation des résultats - Interprétation des clusters - Recommandations

PRÉPARATION DES DONNÉES

1- FUSION - Clés et ordre d'assemblage



2- Nettoyage

- Suppressions des variables inutiles après fusion ou à l'analyse.
- Valeurs Manquantes ➔ dropna()
- Doublons ➔ Pas de doublons
- Valeurs aberrantes : contrôle des différentes dates
- Transformer variable date de Object en datetime

3- Feature Engineering

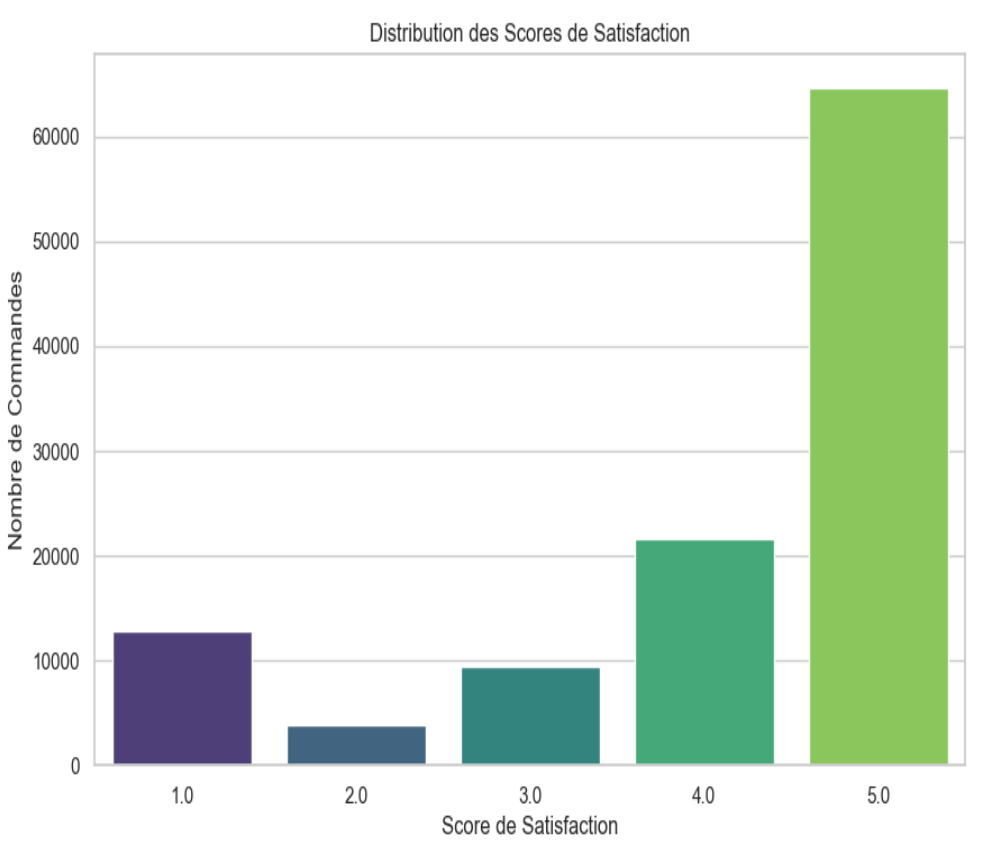
Quels sont les indicateurs clients pertinents permettant d'effectuer une segmentation ?

- Standardisation des variables quantitatives/
- Encodage des variables catégorielles
- Création de nouvelles variables : Recency ,Frequency et Monetary.
- Transformation des variables existantes: démographiques, géographiques ,comportementales.

ANALYSE EXPLORATOIRE DES DONNÉES

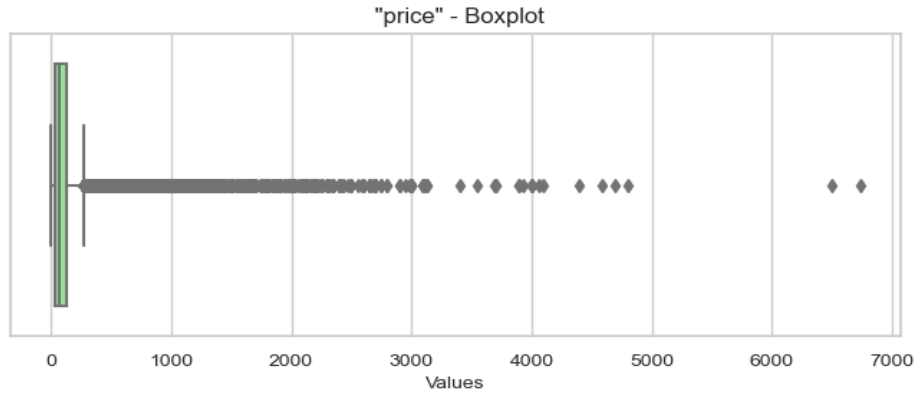
➢ Analyse Univariée: *Analyse de la Satisfaction et des Prix*

Distribution des Scores de Satisfaction



La majorité des clients ont donné des notes élevées, indiquant une expérience globalement positive.

Distribution des Prix

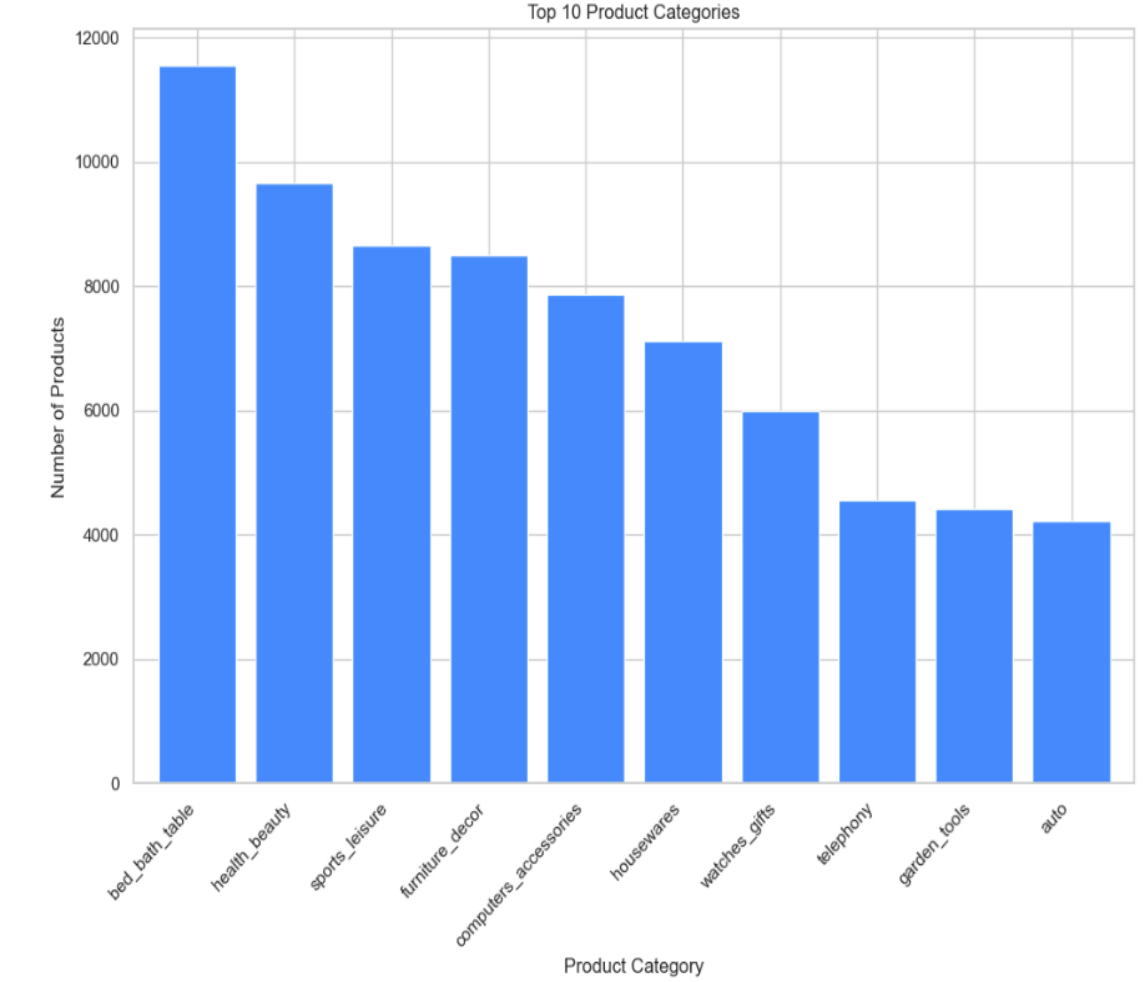


La majorité des produits sont vendus à des prix inférieurs à 1000, avec quelques produits de luxe dépassant les 4000.

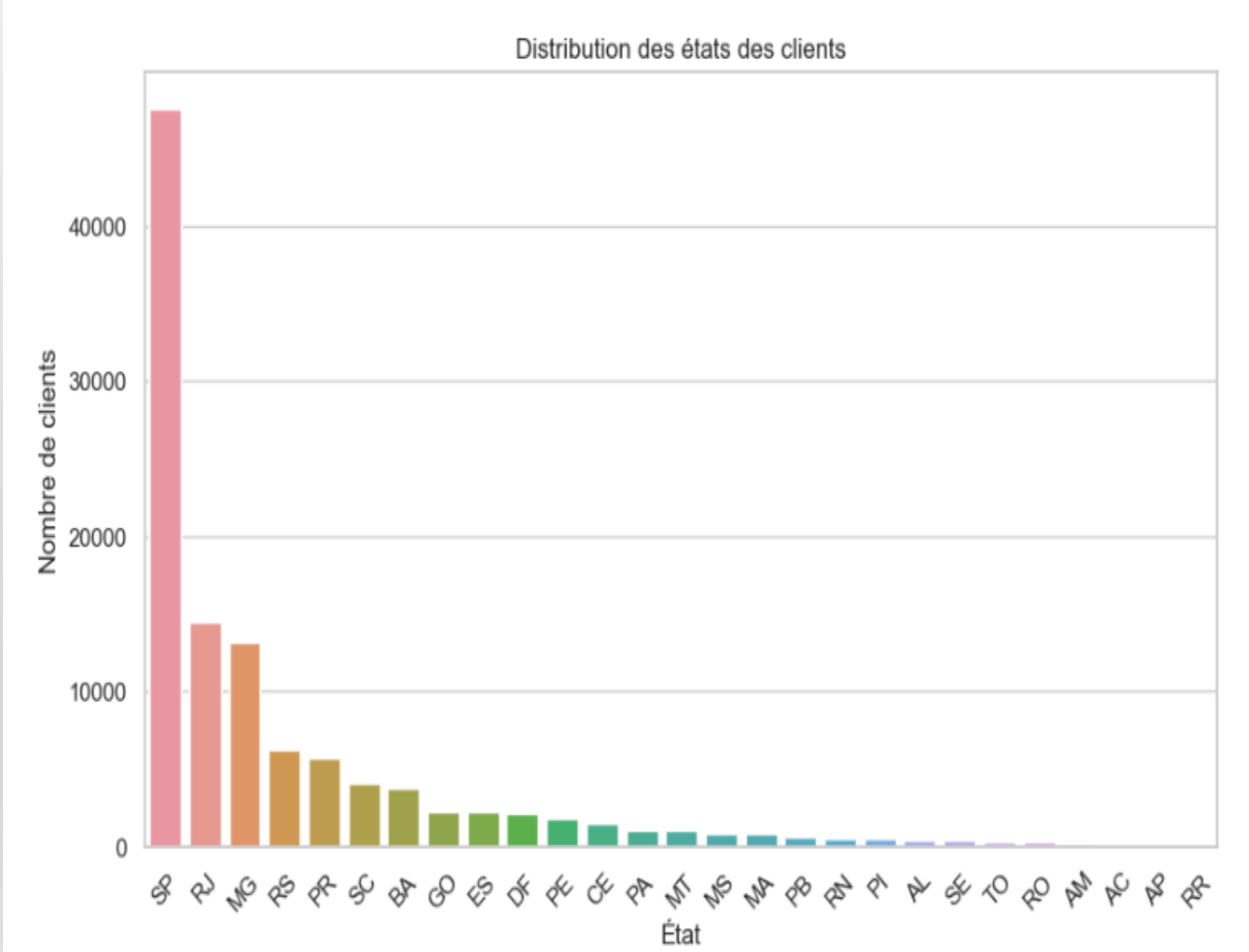
ANALYSE EXPLORATOIRE DES DONNÉES

➢ *Segmentation des Clients*

71 CATÉGORIES DE PRODUITS

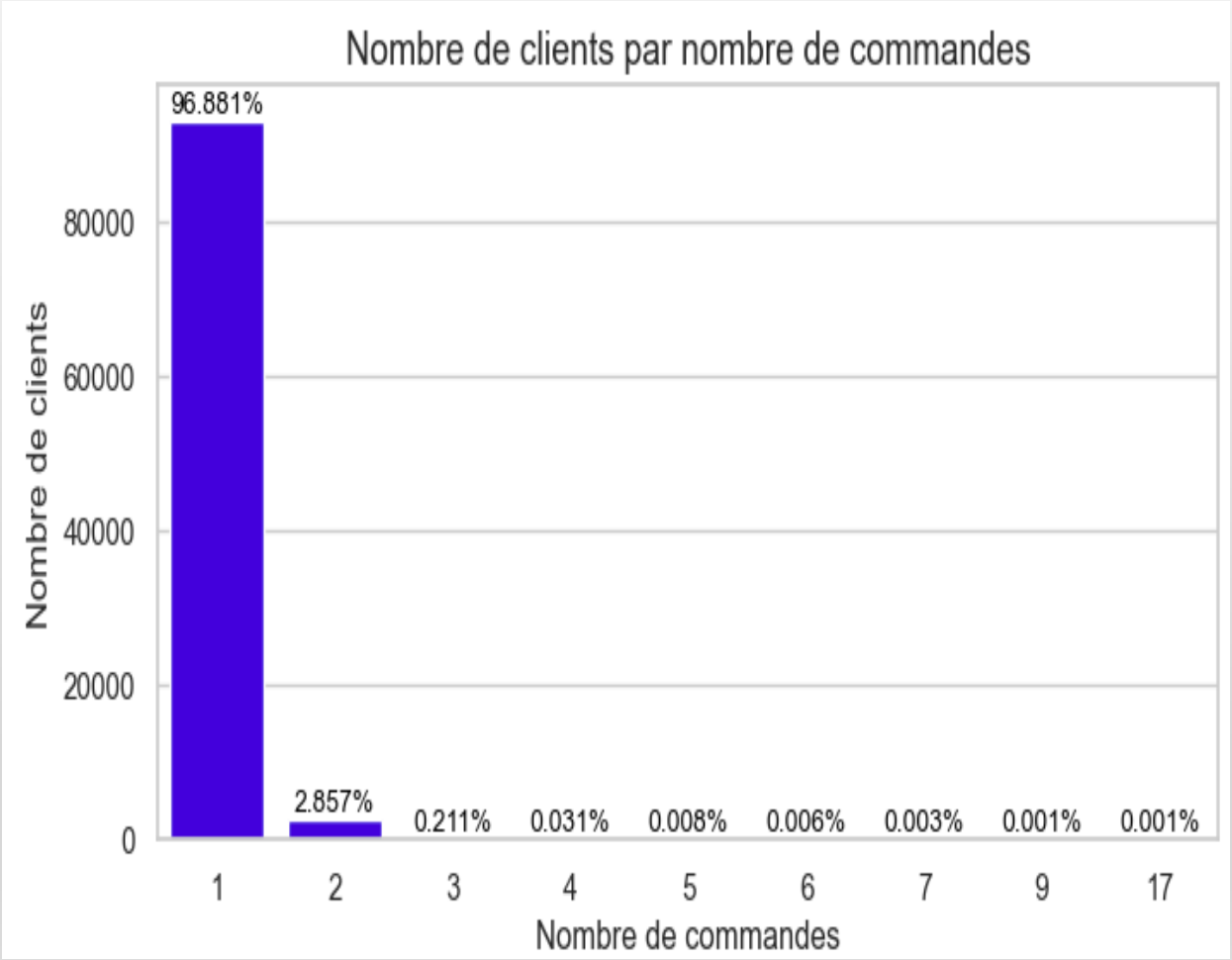


LOCALISATION DES CLIENTS

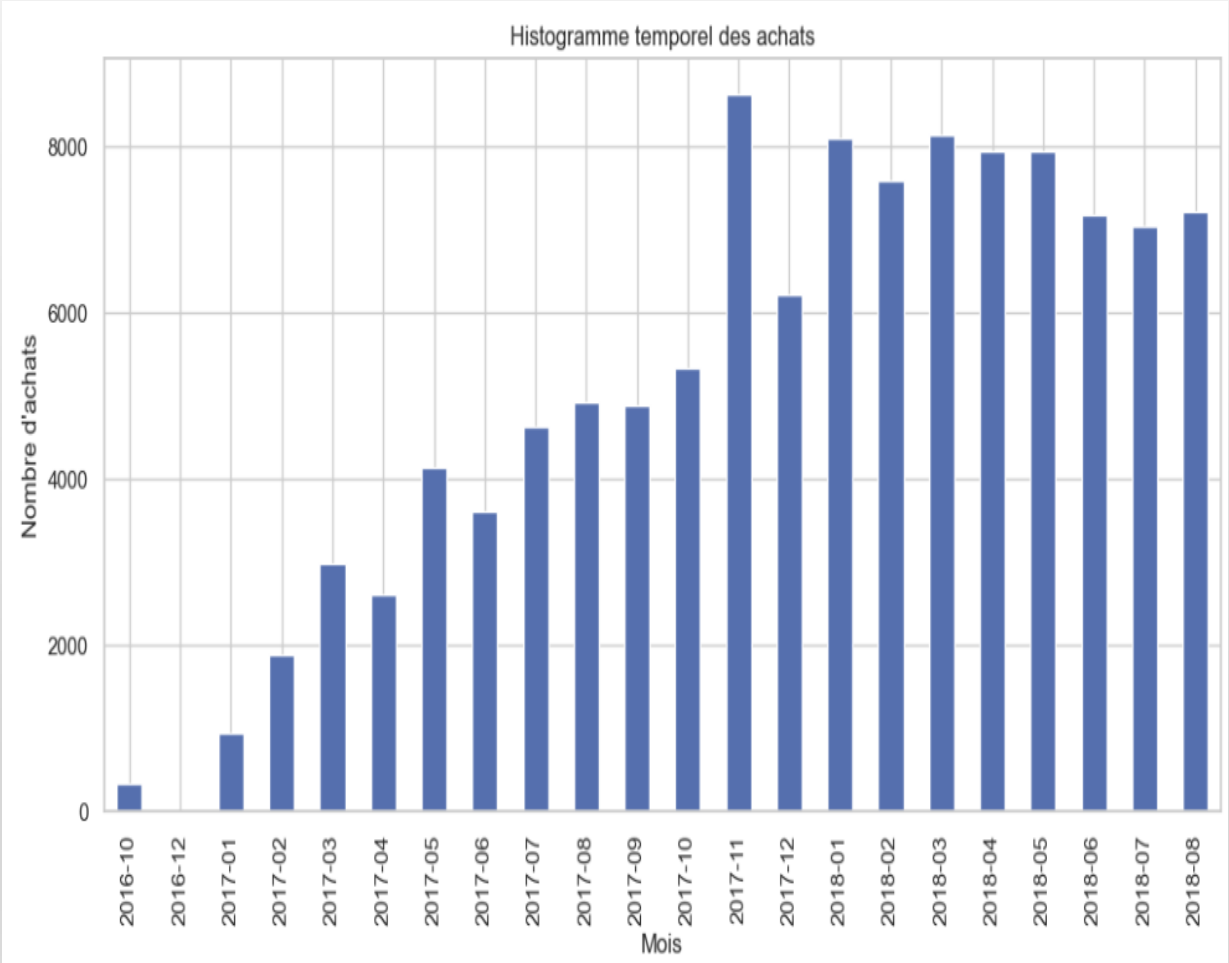


ANALYSE EXPLORATOIRE DES DONNÉES

➢ *Comportement d'Achat des Clients*



Grande majorité des clients passent une seule commande

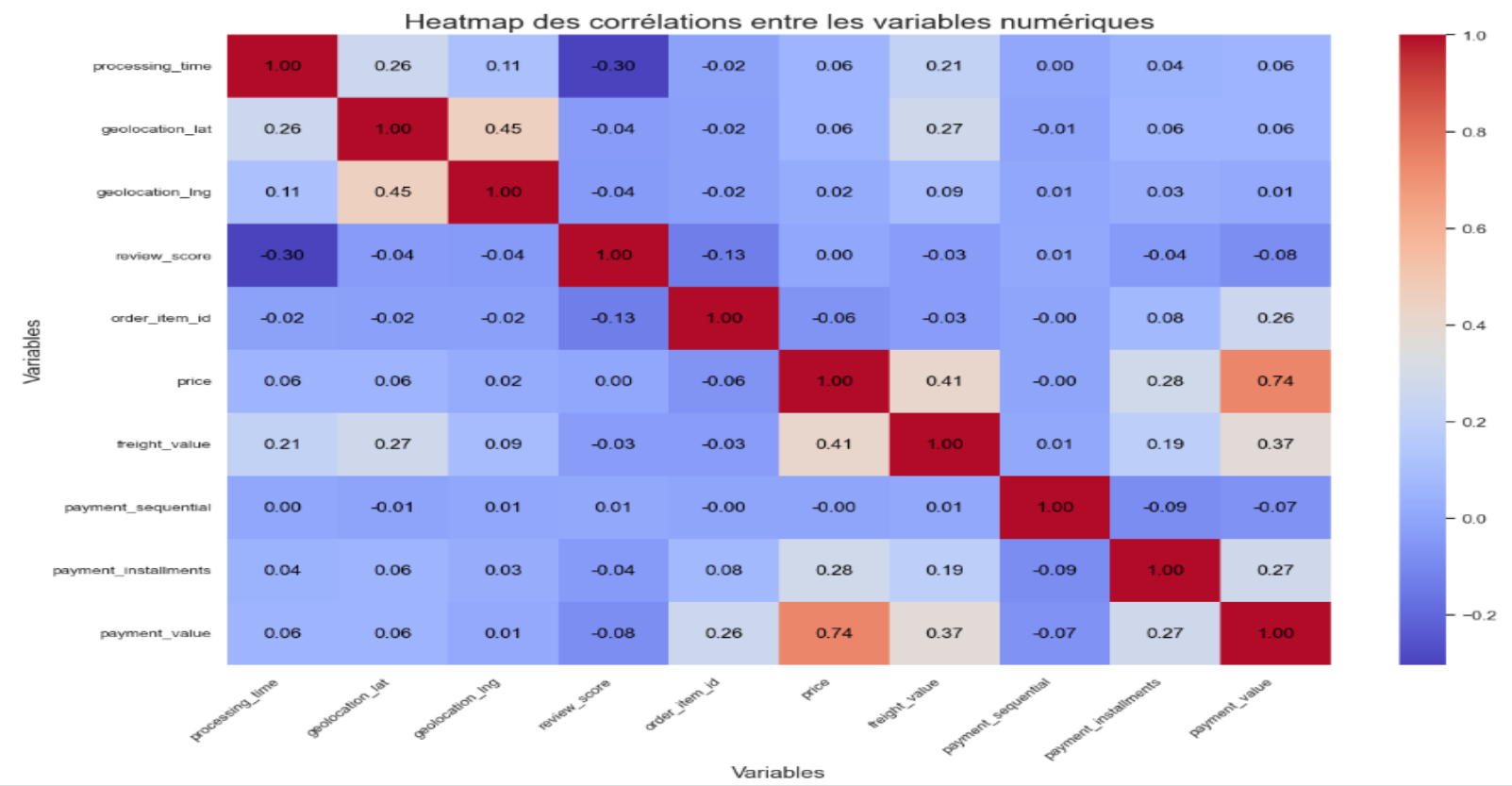


Le graphique montre une augmentation progressive du nombre d'achats mensuels de 2016 à 2018

ANALYSE EXPLORATOIRE DES DONNÉES

➢ Analyse Multivariée

- ❑ **Corrélation entre Variables Quantitatives** : Heatmap des corrélations
- ❑ **Corrélation entre Variables Quantitative et Qualitative** : Test ANOVA
- ❑ **Relation entre Variables Qualitatives** : Test du Chi-carré



Test ANOVA

•**Objectif :** Examiner la relation entre les scores de satisfaction (review_score) et les états des clients (customer_state).

•**Résultats :** F-value de 34.12 et P-value de 1.55e-169.

Test du Chi-carré

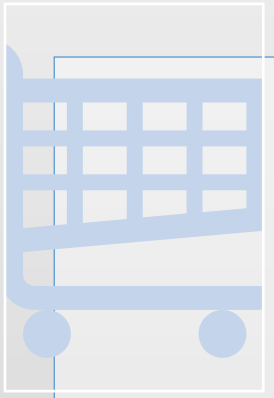
•**Objectif :** Vérifier l'indépendance entre les types de paiement (payment_type) et les catégories de produits (product_category_name_english).

•**Résultats :** Chi2 Stat de 1486.10, P-value de 1.55e-184, et Degrees of Freedom de 216.

La heatmap montre une corrélation positive notable entre le prix et la payment_value (0.74).

Feature Engineering

- Créer et transformer des caractéristiques pour la modélisation
- Quels sont les indicateurs clients pertinents permettant d'effectuer une segmentation ?



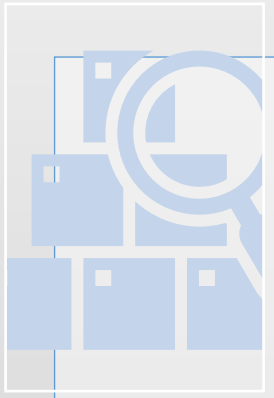
Création des features RFM

Récence - Fréquence – Montant

Recency : Temps écoulé.

Frequency : Nombre d'achats.

Monetary : Montant dépensé.



Transformation des variables existantes

Les indicateurs de segmentation comportementale

Nombre d'articles commandés

Moyen de paiement préféré

Echéances de paiement

Produit le plus acheté,…

Les indicateurs de segmentation géographique

Ville/Etat de résidence

SEGMENTATION RFM

➢ Comparaison des Approches de Segmentation RFM

➢ **Objectif de comparaison : Évaluer les performances et l'adaptabilité des deux algorithmes pour la segmentation des clients.**

Critère	K-Means	DBSCAN
Objectif	Segmentation en groupes homogènes	Clustering basé sur la densité
Nombre optimal de clusters	4	4 (excluant les points de bruit)
Normalisation des données	Oui	Oui
Transformation logarithmique	Oui (Frequency et Monetary)	Oui
Scores de Validation	**Silhouette Score :** 0.3798	**Silhouette Score :** 0.278
	Calinski-Harabasz Score : 16741.4224	**Calinski-Harabasz Score :** 3217.9874
	Davies-Bouldin Score : 1.2313	**Davies-Bouldin Score :** 1.8925
Méthode de détermination des clusters	Méthode du coude	Optimisation des paramètres
Paramètres optimaux	K = 4	eps = 0.5, min_samples = 15
Nombre de points de bruit	N/A	587
Avantages	- Clusters distincts et bien définis - Moins de points de bruit	- Identification de clusters de formes variées - Aucun besoin de spécifier le nombre de clusters à l'avance
Inconvénients	- Sensible à l'initialisation - Nécessite de spécifier le nombre de clusters à l'avance	- Sensible aux paramètres eps et min_samples - Beaucoup de points de bruit
Recommandation	**Utiliser pour segmentation client en raison de sa performance et de la clarté des résultats**	**Considérer pour des données avec des formes de clusters variées, mais moins adapté dans ce cas précis**

SEGMENTATION RFM

❑ Caractérisation des clusters

- Les graphiques montrent la distribution de la récence, de la fréquence et du montant par cluster, ce qui permet de caractériser chaque segment de clients.

- **Graphique 1 : Distribution de la Récence par Cluster**

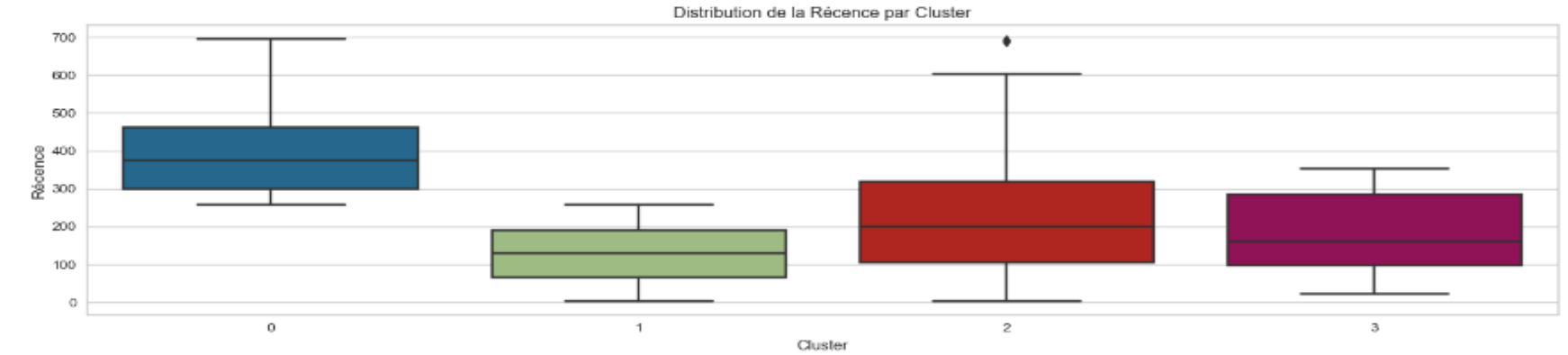
Cluster 0 : Ce cluster a une récence plus élevée, ce qui signifie que ces clients ont effectué des achats plus récemment.

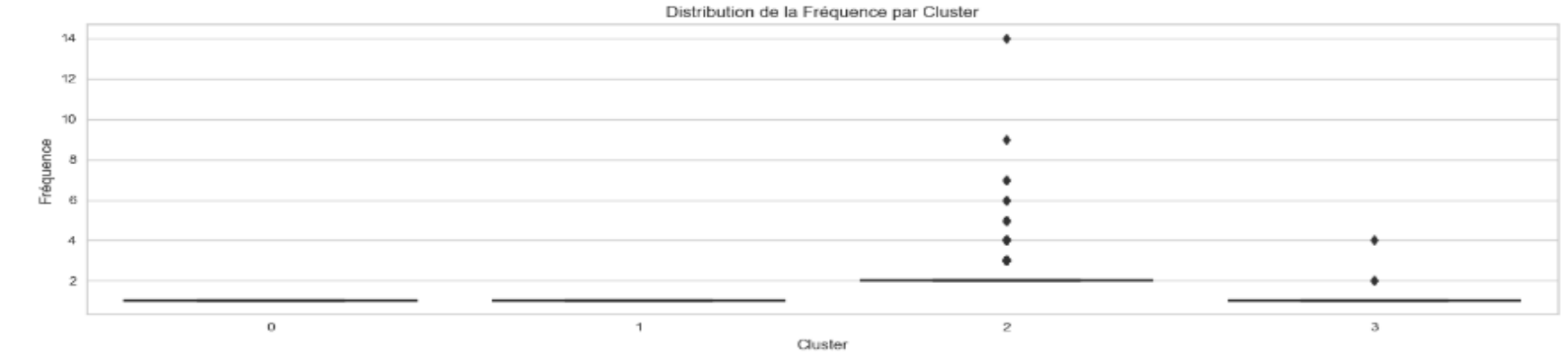
- **Graphique 2 : Distribution de la Fréquence par Cluster**

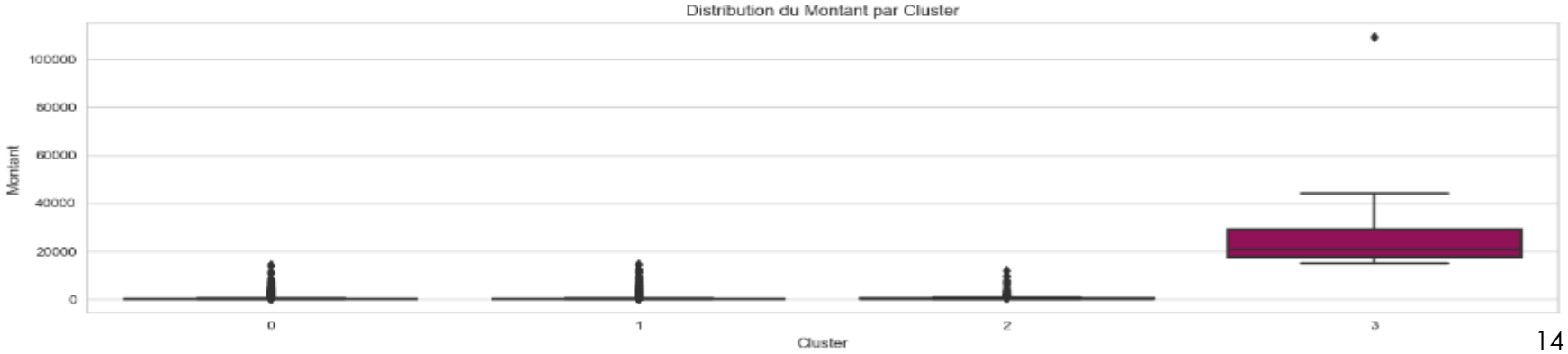
Cluster 0 : La fréquence d'achat est faible, généralement autour de 1.

- **Graphique 3 : Distribution du Montant par Cluster**

Cluster 3 : Les clients du Cluster 3 ont un montant d'achat beaucoup plus élevé, indiquant un panier moyen plus élevé.

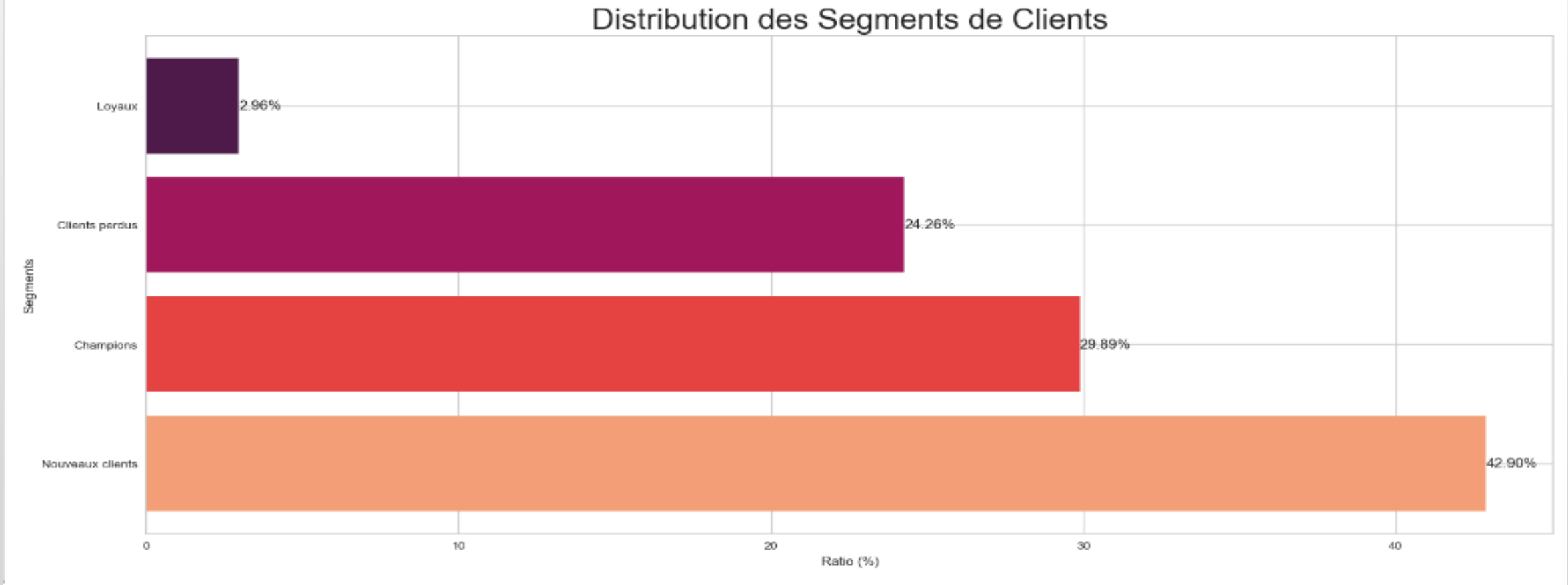






SEGMENTATION RFM

- **Segmentation RFM par K-MEANS**



SEGMENTATION DES CLIENTS PAR K-MEANS

❑ Étapes du Clustering KMeans

Préparation des données

Transformation logarithmique.
Normalisation avec StandardScaler

Détermination du Nombre de Clusters

Utilisation de la méthode du coude.

Évaluation avec des métriques supplémentaires.

Application et Évaluation

Appliquer KMeans avec le k optimal.

Étiquetage et évaluation des clusters

Interprétation et Visualisation

Analyse des segments de clients.

Visualisation avec scatter plot et treemaps.

K-MEANS

❑ K-MEANS

Caractéristiques Principales de KMeans

Partitionnement en K clusters : Les données sont réparties en K groupes basés sur la distance aux centroïdes.

Nombre de clusters prédéfini : Le nombre de clusters doit être spécifié à l'avance.

Avantages

Simplicité : Facile à comprendre et à implémenter.

Efficacité : Rapide pour des petits datasets.

Inconvénients

Sensibilité aux Outliers : Les valeurs extrêmes peuvent influencer les résultats.

Hypothèse de clusters sphériques : Suppose que les clusters ont une forme sphérique et des tailles similaires

Mesures de Performance

Score de Silhouette :

Définition : Évalue la similarité des points au sein des clusters et la différence entre les clusters.

Interprétation : Un score plus élevé indique une meilleure qualité de clustering.

Score Calinski-Harabasz :

Définition : Quantifie la séparation et la cohésion des clusters en maximisant la variance inter-clusters et minimisant la variance intra-cluster.

Interprétation : Un score plus élevé indique une meilleure séparation des clusters.

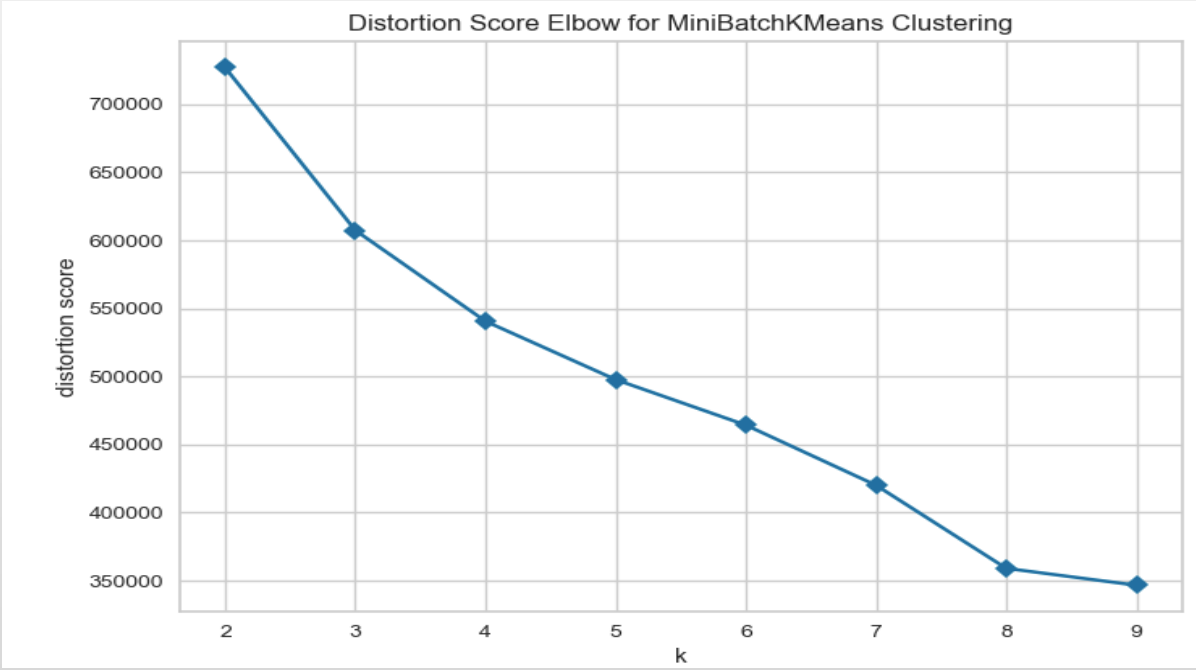
Indice Davies-Bouldin :

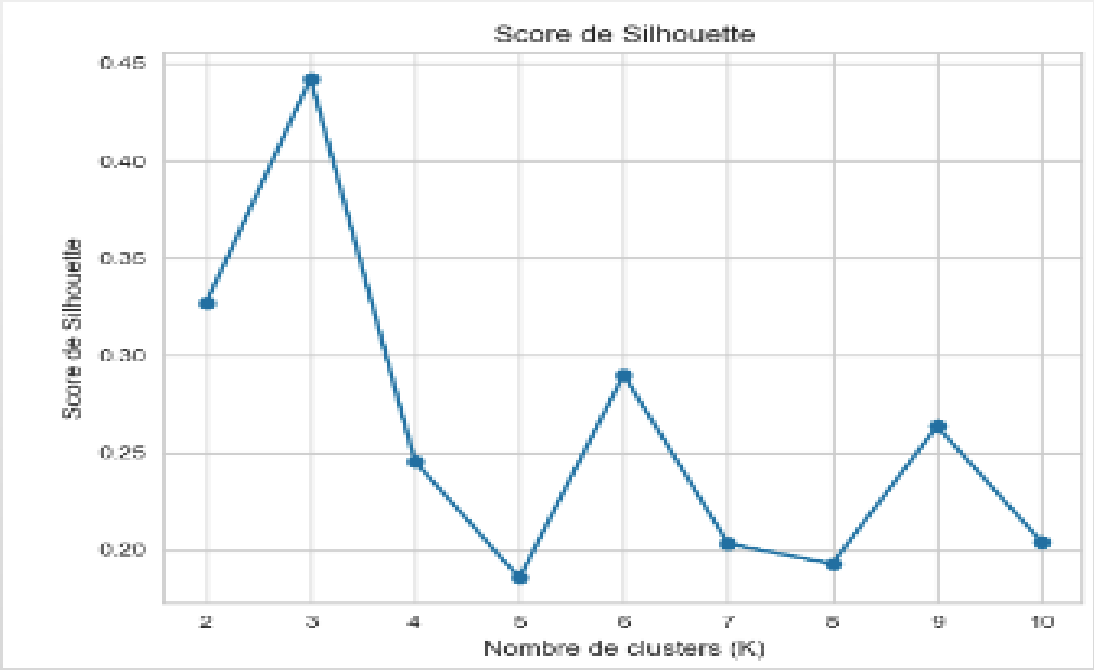
Définition : Mesure la similarité moyenne entre chaque cluster et son cluster voisin le plus proche.

Interprétation : Un score inférieur indique une meilleure séparation des clusters.

K-MEANS

❑ K-MEANS : Nombre Optimal de Clusters

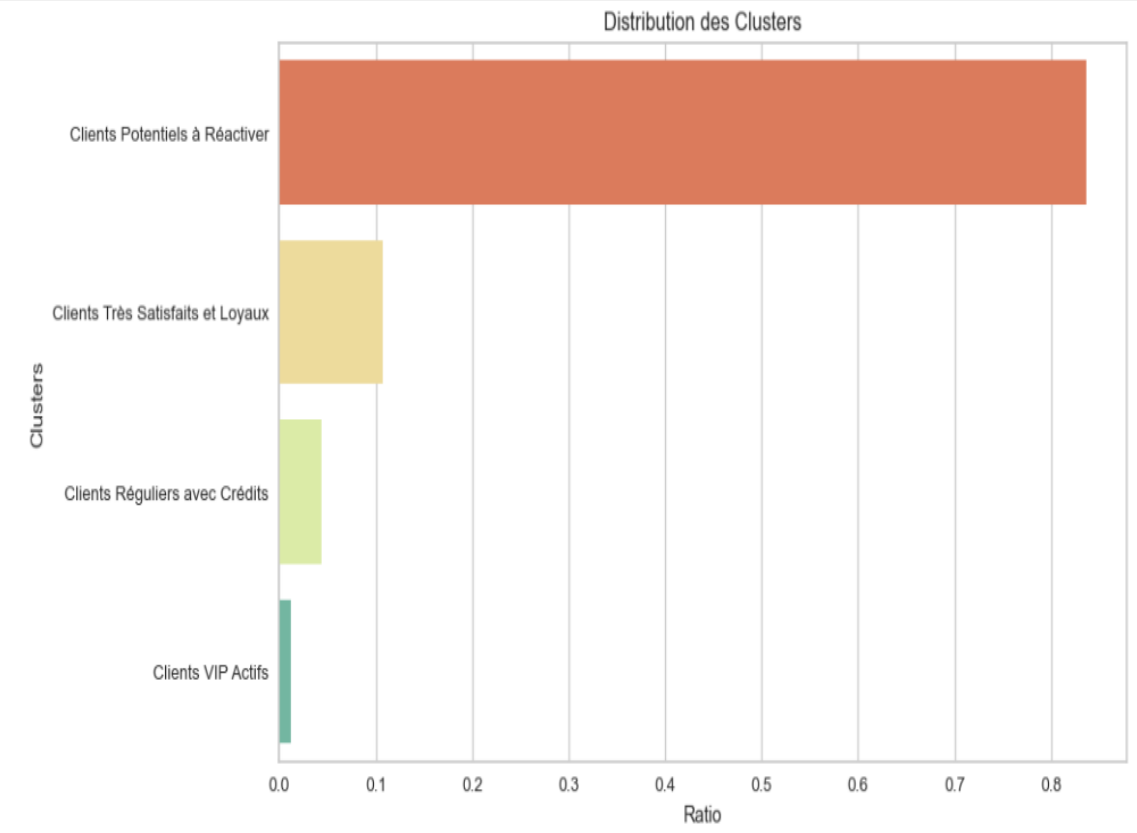


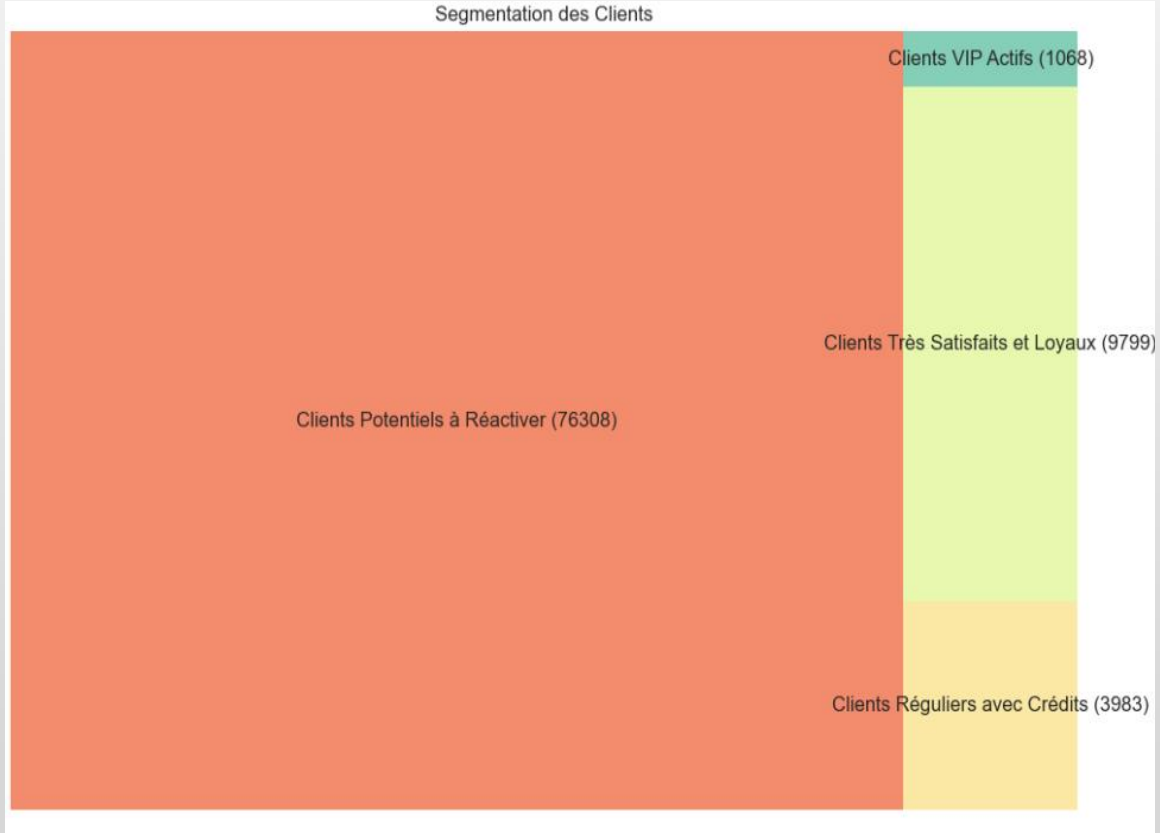


K = 4

K-MEANS

❑ Segmentation des Clients par K-Means





Les graphiques illustrent une majorité de clients potentiels à réactiver, avec des segments distincts de clients très loyaux, réguliers avec crédits, et VIP actifs, permettant des stratégies marketing ciblées pour chaque groupe.

MAINTENANCE DU MODÈLE

❑ Évaluation de la Stabilité Temporelle de la Segmentation par K-Means

```
Semaine 0 - Score pour la période du 2016-10-03 au 2017-08-03: 1.0
Semaine 1 - Score pour la période du 2016-10-03 au 2017-08-10: 0.5889972761926491
Semaine 2 - Score pour la période du 2016-10-03 au 2017-08-17: 0.3966799643144724
Semaine 3 - Score pour la période du 2016-10-03 au 2017-08-24: 0.5405596090707863
Semaine 4 - Score pour la période du 2016-10-03 au 2017-08-31: 0.5308465793093103
Semaine 5 - Score pour la période du 2016-10-03 au 2017-09-07: 0.5223182732591013
Semaine 6 - Score pour la période du 2016-10-03 au 2017-09-14: 0.7515929903571672
Semaine 7 - Score pour la période du 2016-10-03 au 2017-09-21: 0.24417578678557936
Semaine 8 - Score pour la période du 2016-10-03 au 2017-09-28: 0.4912347174778657
Semaine 9 - Score pour la période du 2016-10-03 au 2017-10-05: 0.2592325196675725
Semaine 10 - Score pour la période du 2016-10-03 au 2017-10-12: 0.45692564985477196
Semaine 11 - Score pour la période du 2016-10-03 au 2017-10-19: 0.4385449191008515
Semaine 12 - Score pour la période du 2016-10-03 au 2017-10-26: 0.41752504437991156
Semaine 13 - Score pour la période du 2016-10-03 au 2017-11-02: 0.3914917832329332
Semaine 14 - Score pour la période du 2016-10-03 au 2017-11-09: 0.5693670850367284
Semaine 15 - Score pour la période du 2016-10-03 au 2017-11-16: 0.349943488939137
Semaine 16 - Score pour la période du 2016-10-03 au 2017-11-23: 0.33957220473633815
Semaine 17 - Score pour la période du 2016-10-03 au 2017-11-30: 0.3620156212253497
Semaine 18 - Score pour la période du 2016-10-03 au 2017-12-07: 0.36803657169964704
Semaine 19 - Score pour la période du 2016-10-03 au 2017-12-14: 0.5323627465137227
Semaine 20 - Score pour la période du 2016-10-03 au 2017-12-21: 0.2502670142342228
Semaine 21 - Score pour la période du 2016-10-03 au 2017-12-28: 0.3275184931962192
Semaine 22 - Score pour la période du 2016-10-03 au 2018-01-04: 0.3106193485068281
Semaine 23 - Score pour la période du 2016-10-03 au 2018-01-11: 0.4677922276459362
Semaine 24 - Score pour la période du 2016-10-03 au 2018-01-18: 0.3103676266720149
Semaine 25 - Score pour la période du 2016-10-03 au 2018-01-25: 0.4554932548667202
Semaine 26 - Score pour la période du 2016-10-03 au 2018-02-01: 0.47832025593309757
Semaine 27 - Score pour la période du 2016-10-03 au 2018-02-08: 0.3010641880405808
Semaine 28 - Score pour la période du 2016-10-03 au 2018-02-15: 0.2922996148958716
Semaine 29 - Score pour la période du 2016-10-03 au 2018-02-22: 0.455384442769322
Semaine 30 - Score pour la période du 2016-10-03 au 2018-03-01: 0.17249720309983252
Semaine 31 - Score pour la période du 2016-10-03 au 2018-03-08: 0.35738274003260356
Semaine 32 - Score pour la période du 2016-10-03 au 2018-03-15: 0.280923474983111
Semaine 33 - Score pour la période du 2016-10-03 au 2018-03-22: 0.29703594292458624
```

- ***Adjusted Rand Index (ARI)** : Indique la similarité entre les clusters prédits et les clusters réels, ajustée pour le hasard. Les scores très élevés (souvent de 1) montrent une forte stabilité et cohérence des clusters.*

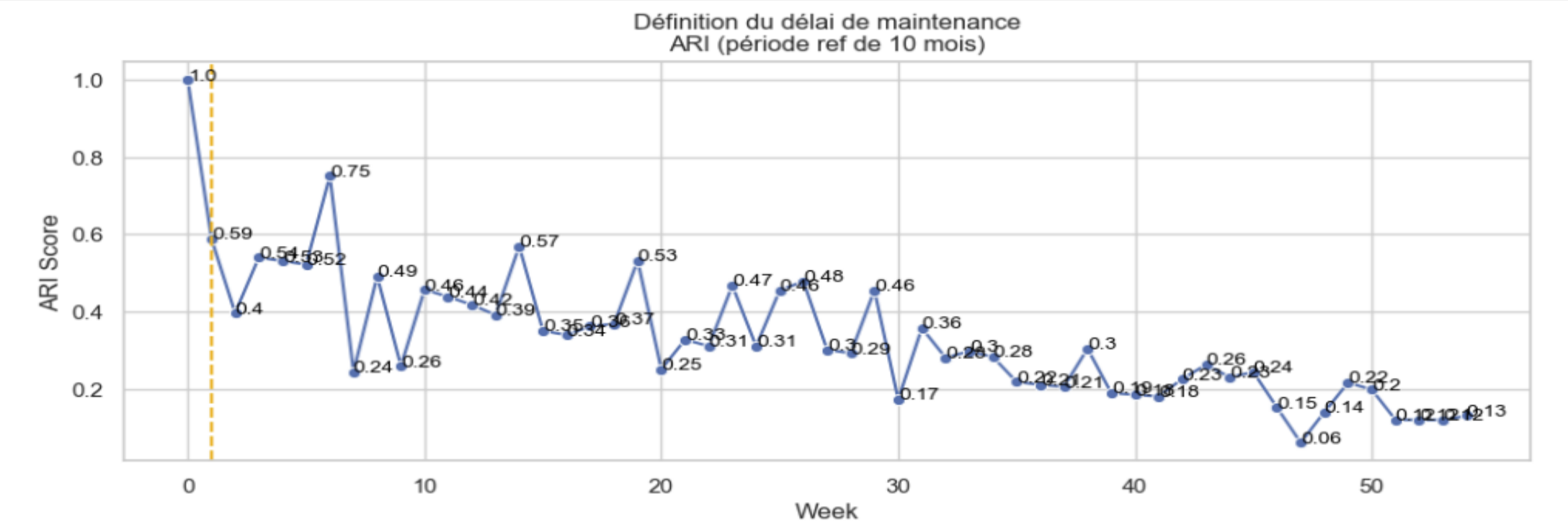
- ***Adjusted Rand Index (ARI) : Les scores ARI montrent une bonne correspondance initiale avec la période de référence, mais chutent rapidement après la première semaine.***

- ***Recommandation:***

Il est conseillé de réentraîner le modèle lorsque l'ARI descend en dessous de 0.8 pour assurer des prédictions de haute qualité et des segments pertinents.

MAINTENANCE DU MODÈLE

❑ Résultats de la Simulation et Recommandations



- ***Fréquence de Réentraînement du Modèle***
- ***Le graphique montre que les scores ARI diminuent rapidement après la période de référence de 10 mois, indiquant une perte de pertinence du modèle.***

Réentraînement Nécessaire : ***Pour maintenir une segmentation fiable, un réentraînement du modèle devrait être effectué toutes les 6 semaines, car c'est le point où les scores ARI commencent à chuter de manière significative en dessous de 0.8.***

CONCLUSION

Contrat de maintenance tous les 4 à 6 mois.

Clustering Final : Le modèle K-Means a identifié plusieurs segments clients distincts selon les indicateurs RFM (Recency, Frequency, Monetary), démontrant une homogénéité et une cohérence élevées.

Segments Identifiés et Actions Marketing :

- **Clients Fidèles et Rentables :** Programmes de fidélité, offres personnalisées, événements VIP.
- **Nouveaux Clients Satisfaits :** Offres de bienvenue, recommandations de produits, promotions spéciales.
- **Clients à Réactiver :** Campagnes de réactivation, enquêtes de satisfaction, offres limitées.
- **Acheteurs Occasionnels :** Offres saisonnières, communications sur les nouveautés, programmes de fidélité.
- **Clients à Fort Potentiel :** Programmes de parrainage, packages de produits, notifications push.

Recommandations :

- **Personnalisation des Communications :** Adapter les messages marketing pour chaque segment.
- **Suivi et Évaluation Continus :** Évaluer régulièrement les segments pour ajuster les stratégies.
- **Ré-entrainement du Modèle :** Ré-entrainer le modèle tous les 6 semaines ou dès que l'ARI descend en dessous de 0.8.

Intégrer des données externes (réseaux sociaux,…) et explorer des techniques avancées (clustering hiérarchique) pour des segments plus précis et des stratégies marketing optimisées.

MERCI !

Des questions ?