



FORMATION DATA SCIENTIST : PROJET 5



Segmentez des clients d'un site e-commerce



Sommaire

- ▶ Introduction
- ▶ Données et sources
- ▶ Extraction des données
- ▶ Analyse exploratoire et prétraitement des données
- ▶ Segmentation : Approches testées
- ▶ Modèle final
- ▶ Profils
- ▶ Stratégie de mise à jour et maintenance
- ▶ Conclusion et Recommandations



1. Introduction

- Contexte :

Olist, le pont entre vendeurs et marketplaces.

Objectif : Innover avec une équipe Data dédiée

- Problématique :

Segmentation client : Clé d'une communication marketing sur-mesure

- But :

Décrypter les comportements pour des actions ciblées. Une segmentation client précise et utile.



2. Données et Sources

- Base de Données Riches

Accès à un trésor d'informations : commandes, satisfaction, localisation etc.

De janvier 2017 à octobre 2018.

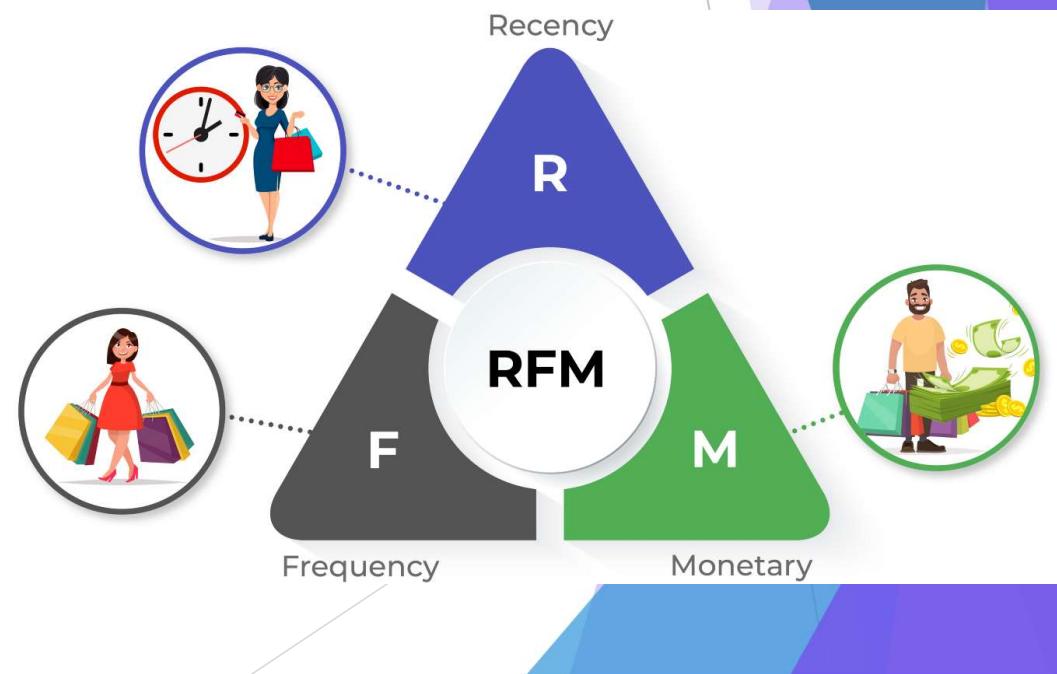


- Diversité des Données

Un panorama complet : Récence, Fréquence, Montant et bien plus.

- Analyse en Profondeur

Des données prêtes pour l'exploration : Identifier les tendances, comportements et préférences.



3. Extraction des données

- Base de données SQL

9 tables

translation
123 index
ABC product_category_name
ABC product_category_name_english

sellers
123 index
ABC seller_id
123 seller_zip_code_prefix
ABC seller_city
ABC seller_state

customers
123 index
ABC customer_id
ABC customer_unique_id
123 customer_zip_code_prefix
ABC customer_city
ABC customer_state

geoloc
123 index
123 geolocation_zip_code_prefix
123 geolocation_lat
123 geolocation_lng
ABC geolocation_city
ABC geolocation_state

order_pymts
123 index
ABC order_id
123 payment_sequential
ABC payment_type
123 payment_installments
123 payment_value

order_items
123 index
ABC order_id
123 order_item_id
ABC product_id
ABC seller_id
ABC shipping_limit_date
123 price
123 freight_value

order_reviews
123 index
ABC review_id
ABC order_id
123 review_score
ABC review_comment_title
ABC review_comment_message
ABC review_creation_date
ABC review_answer_timestamp

orders
123 index
ABC order_id
ABC customer_id
ABC order_status
ABC order_purchase_timestamp
ABC order_approved_at
ABC order_delivered_carrier_date
ABC order_delivered_customer_date
ABC order_estimated_delivery_date

products
123 index
ABC product_id
ABC product_category_name
123 product_name_lenght
123 product_description_lenght
123 product_photos_qty
123 product_weight_g
123 product_length_cm
123 product_height_cm
123 product_width_cm

3. Extraction des données

- Sélection et pré traitement

3 requêtes pour sortir les données pertinentes de la base

Extract .csv	Variables brutes	Features engineering
Customers	<u>customer_unique_id</u> customer_city customer_state geolocation_lat geolocation_lng	FirstOrderDate, LastOrderDate, TotalOrders, TotalSpent, TotalFreight, AvgItems, nb_item, ActualDeliveryTime, DeliveryDelay, AverageReviewScore, NumberOfReviews, NumberOfCommentTitles, NumberOfComments, DifferentCategories, AvgWeight, AvgVolume
Category	<u>customer_unique_id</u> product_category_name_english	CategoryCount TotalSpentPerCategory
Payment	<u>customer_unique_id</u> payment_type	PaymentCount TotalInstallments, TotalPaymentValue

- Merge, Nettoyage et features engineering

92411 clients, 1 dataframe

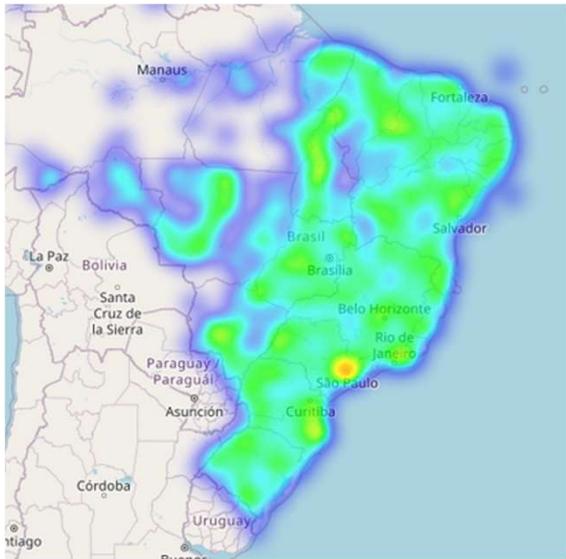
TotalOrders
nb_item
ActualDeliveryTime
DeliveryDelay
AverageReviewScore
DifferentCategories
AvgWeight
AvgVolume
TotalInstallments

boleto_pct
credit_card_pct
debit_card_pct
voucher_pct
Fashion_Beauty_pct
Home_pct
Food_drink_pct
Miscellaneous_pct
Books_pct
Office_pct
Construction_pct
Sports_Leisure_pct
Technology_Electronics_pct

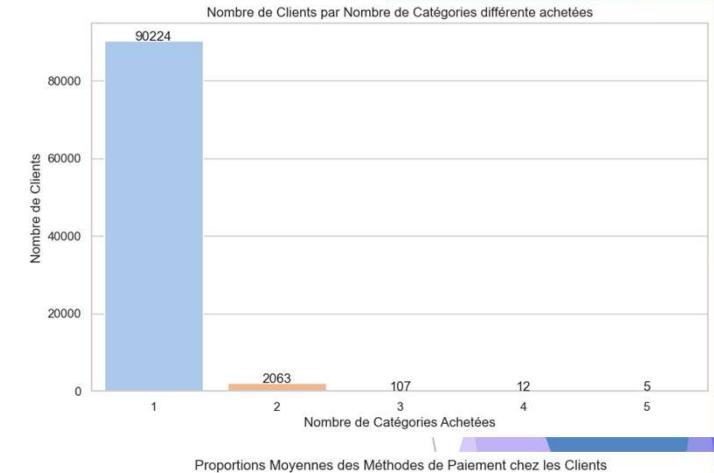
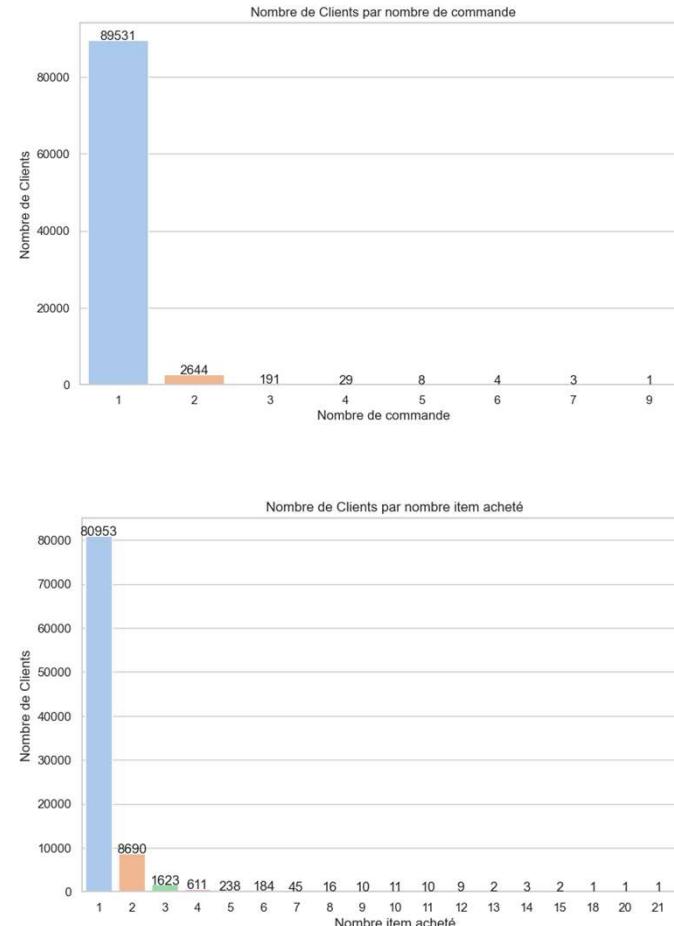
Recency
Frequency
TotalFreightPct
AvgBasket
EngagementIndex
Region

4. Analyse exploratoire et prétraitement des données

- Position géographique des clients d'olist

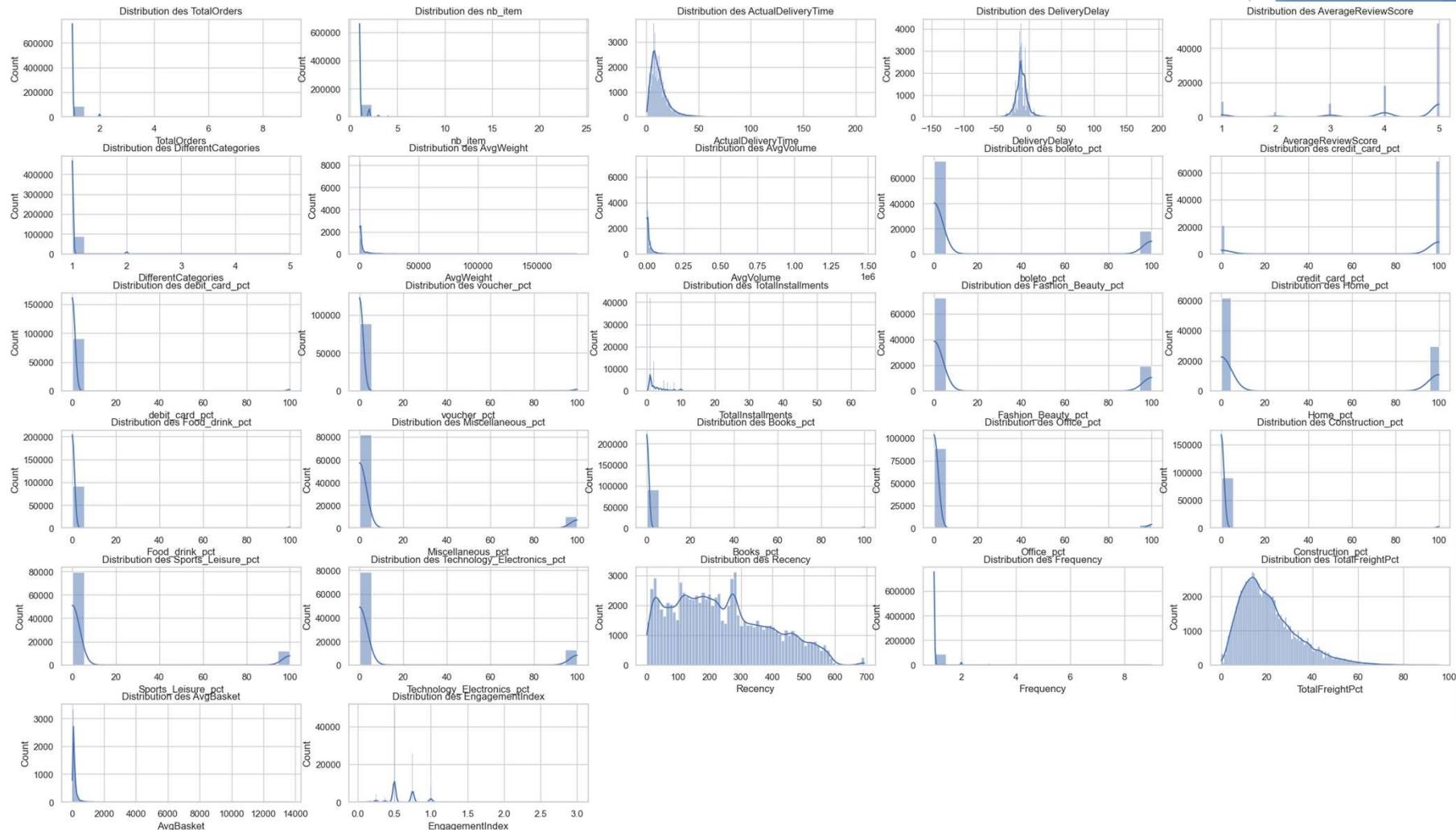


- Statistiques notables



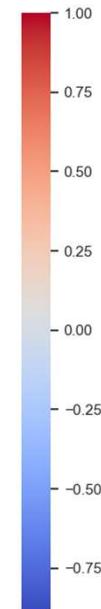
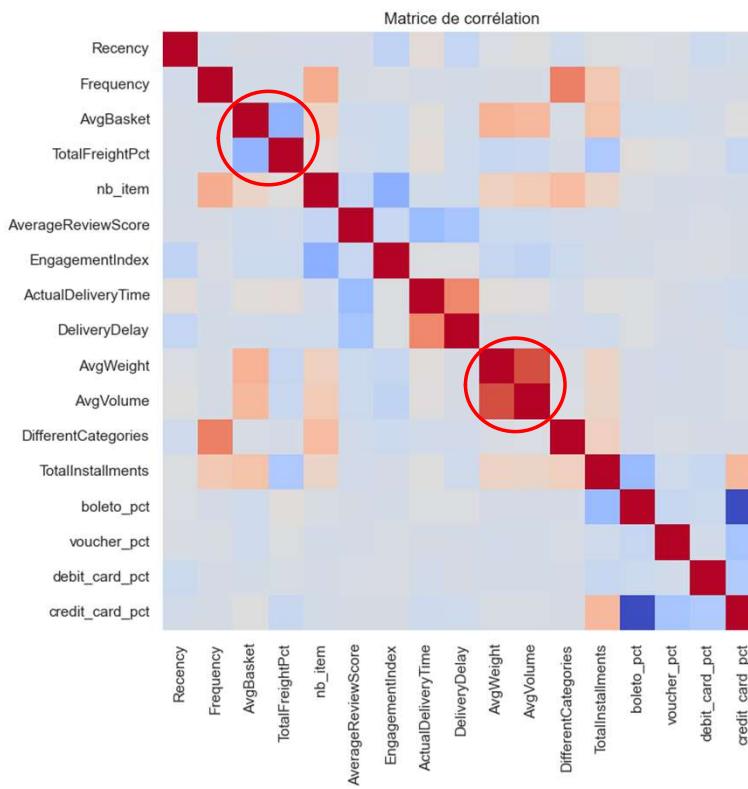
4. Analyse exploratoire et prétraitement des données

- Distribution des variables numériques



4. Analyse exploratoire et prétraitement des données

- Analyse des corrélations, ACP et sélection finale des indicateurs

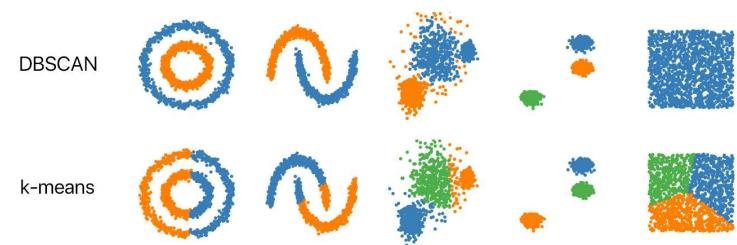


Matrice de corrélation

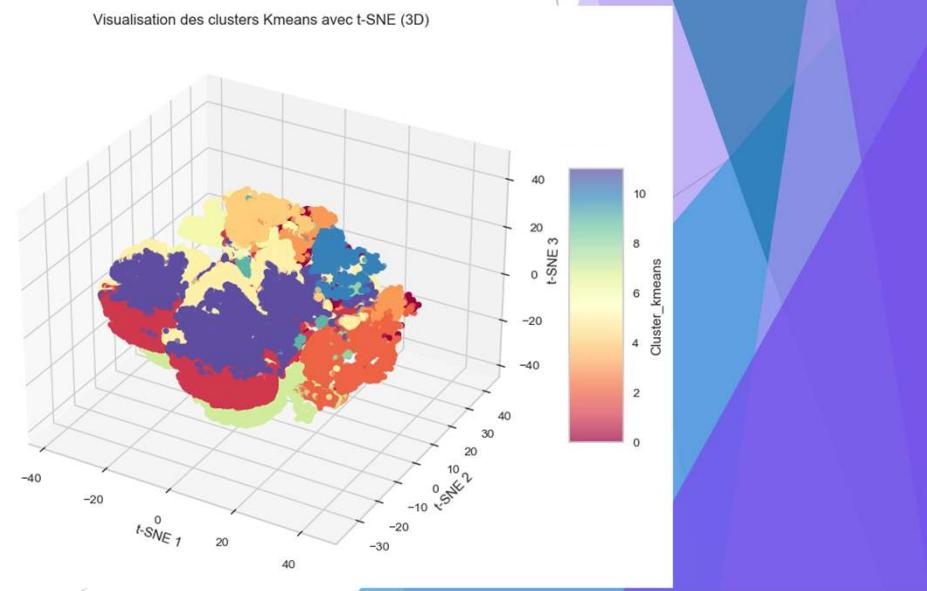
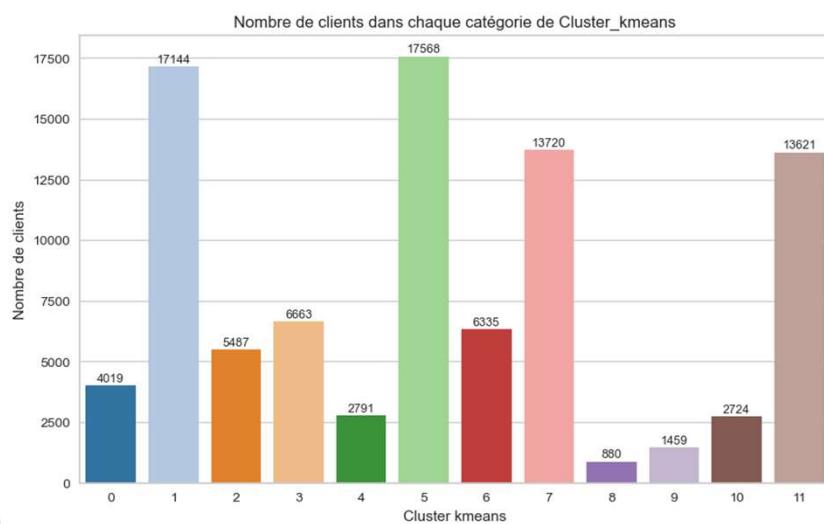
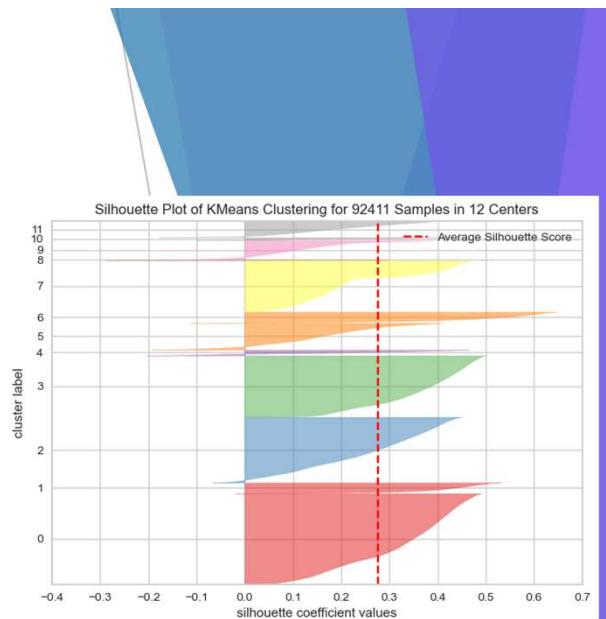
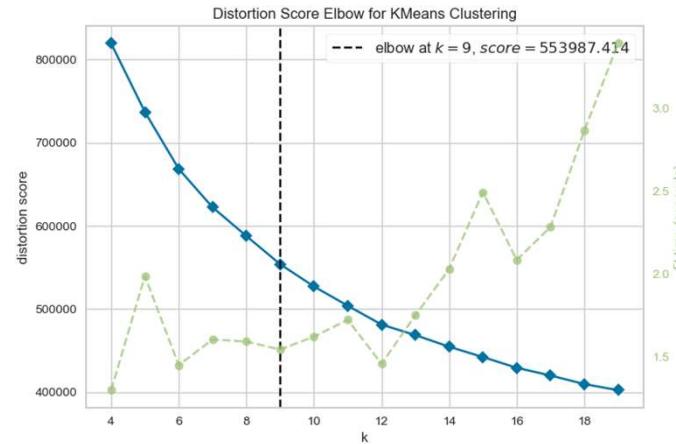
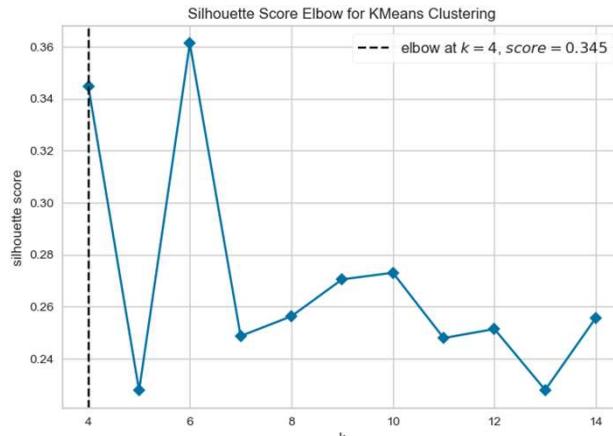
	Recency	Frequency	AvgBasket	nb_item	AverageReviewScore	EngagementIndex	ActualDeliveryTime	DeliveryDelay	AvgWeight	AvgVolume	DifferentCategories	TotalInstallments	boleto_pct	voucher_pct	debit_card_pct	credit_card_pct
Recency	1.00	-0.02	-0.00	-0.01	-0.01	-0.13	0.10	-0.10	0.03	0.04	0.03	0.02				
Frequency	-0.02	1.00	-0.01	0.43	0.00	0.02	-0.00	-0.01	-0.00	0.26	-0.00	0.01				
AvgBasket	-0.00	-0.01	1.00	0.16	-0.04	-0.05	0.07	-0.02	0.40	0.30	-0.04	-0.04				
nb_item	-0.01	0.43	0.16	1.00	-0.11	-0.41	-0.02	-0.03	0.20	0.17	0.02	-0.01				
AverageReviewScore	-0.01	0.00	-0.04	-0.11	1.00	-0.08	-0.34	-0.27	-0.06	-0.03	0.00	0.00				
EngagementIndex	-0.13	0.02	-0.05	-0.41	-0.08	1.00	0.04	0.04	-0.09	-0.00	-0.01	0.01				
ActualDeliveryTime	0.10	-0.00	0.07	-0.02	-0.34	0.04	1.00	0.60	0.07	0.05	0.05	0.00				
DeliveryDelay	-0.10	-0.01	-0.02	-0.03	-0.27	0.04	0.60	1.00	0.00	-0.03	0.04	-0.00				
AvgWeight	0.03	-0.00	0.40	0.20	-0.06	-0.09	0.07	0.00	1.00	0.18	-0.01	-0.02				
AvgVolume	0.04	0.26	0.30	0.17	-0.03	-0.00	0.05	-0.03	0.18	1.00	-0.34	-0.04				
DifferentCategories	0.03	-0.00	-0.04	0.02	0.00	-0.01	0.05	0.04	-0.01	-0.34	1.00	-0.09				
TotalInstallments	0.02	0.01	-0.04	-0.01	-0.00	0.01	-0.00	-0.00	-0.02	-0.04	-0.09	1.00				
boleto_pct																
voucher_pct																

5. Segmentation : Approches testées

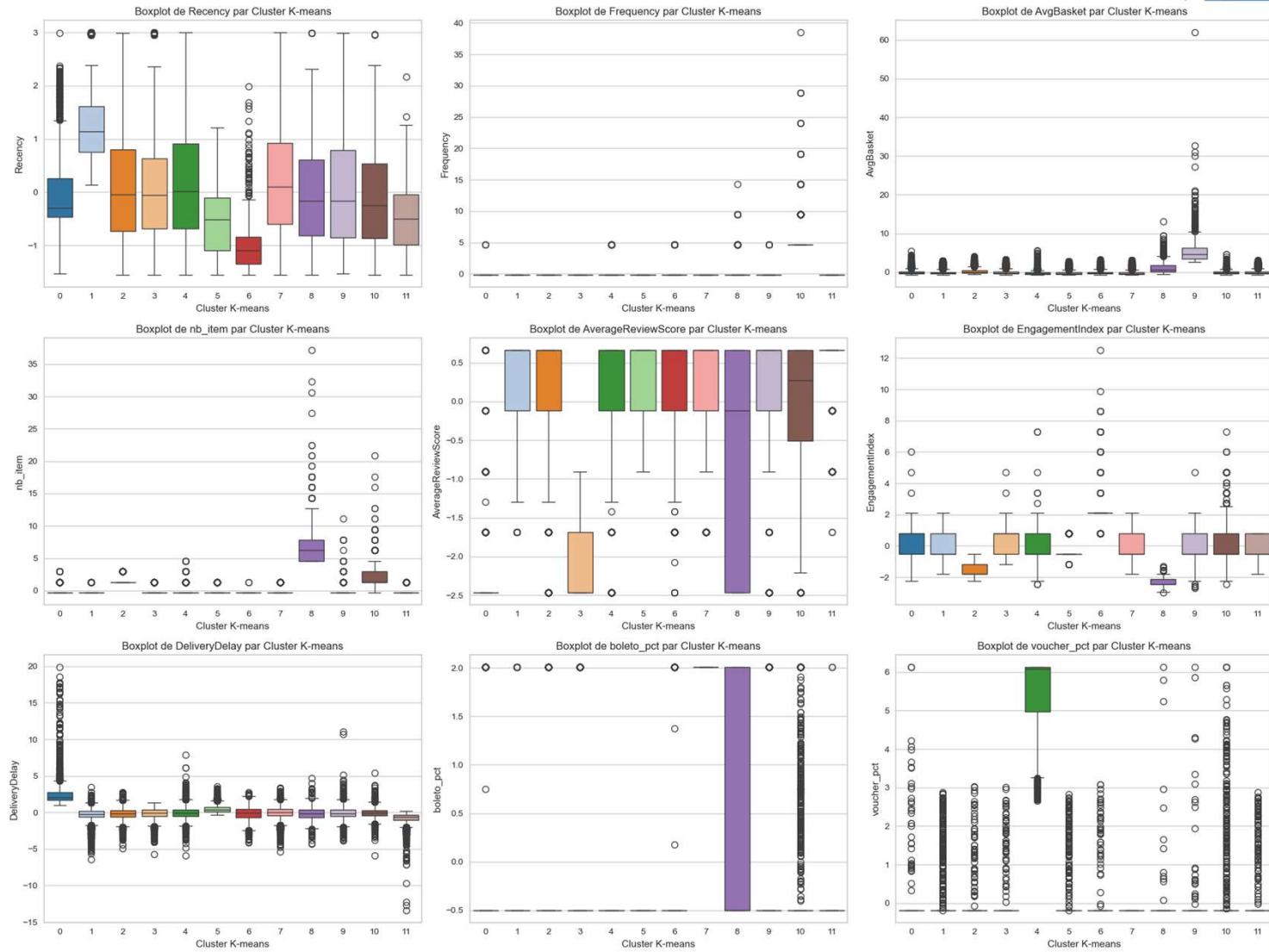
- 7 indicateurs
 - 'Recency',
 - 'Frequency',
 - 'AvgBasket',
 - 'nb_item',
 - 'AverageReviewScore',
 - 'EngagementIndex',
 - 'DeliveryDelay'
- 9 indicateurs : 7 indicateurs +
 - 'boleto_pct',
 - 'voucher_pct'
- 12 indicateurs : 9 indicateurs +
 - 'AvgWeight',
 - 'ActualDeliveryTime',
 - 'TotalInstallments'
- 3 algorithmes testés
 - K-means++
 - DBScan
 - K-means + Hiérarchique
- Métriques
 - Silhouette score
 - Distortion score
 - Score de Davies-Bouldin
 - Taille des clusters
 - **Pertinence des profils**



6. Modèle final



7. Profils

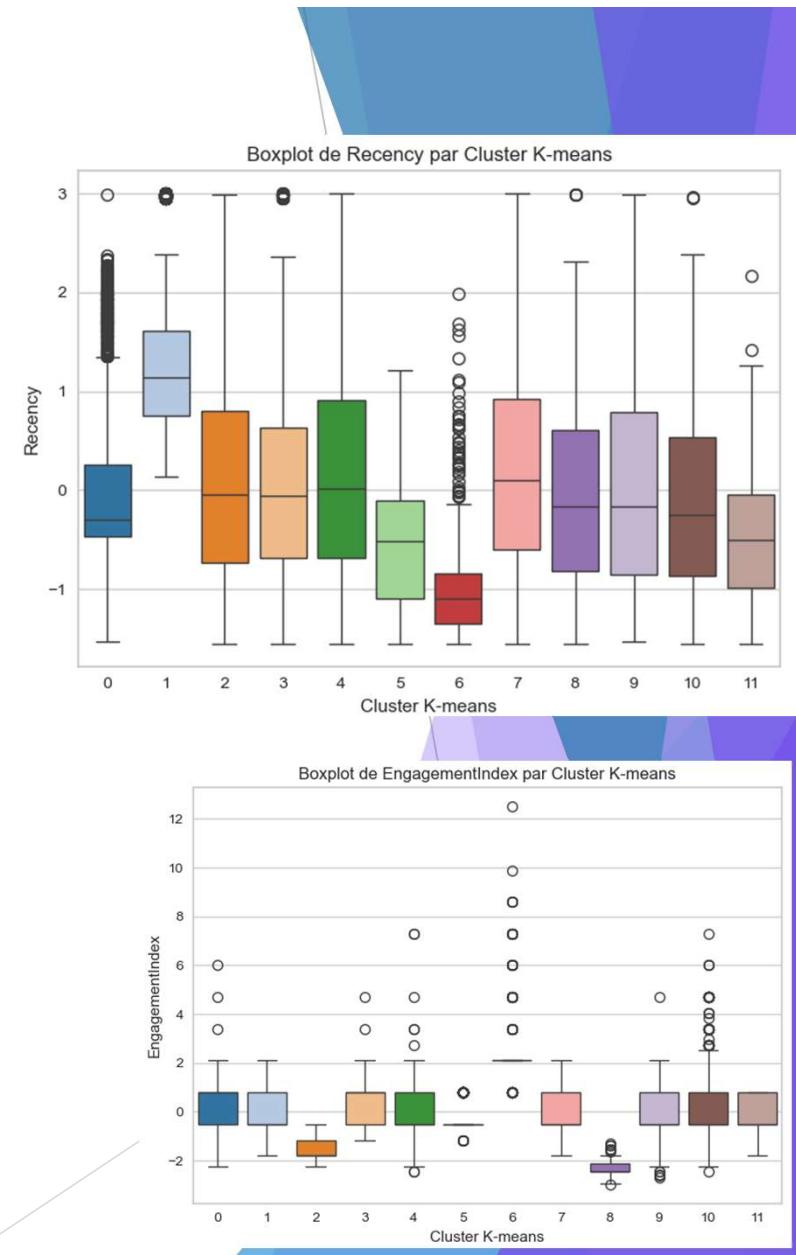
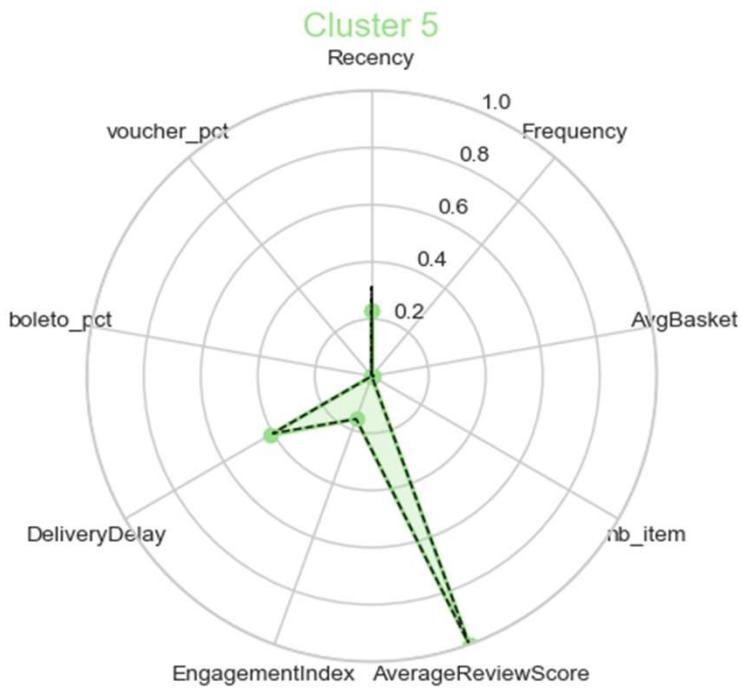


7. Profils

Clients Actifs Récemment

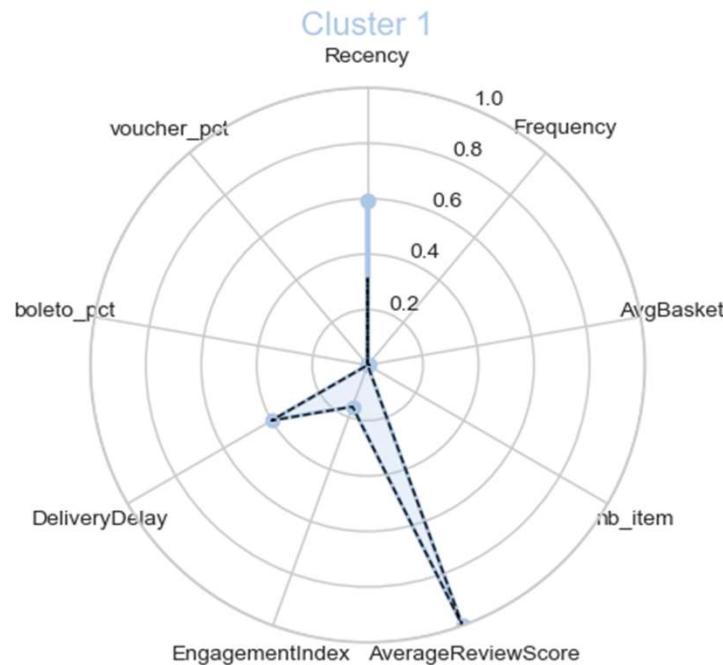
- Particularités :
 - Commande récente

- Proportion :
 - 17568 clients
 - 19 %



7. Profils

Clients Inactifs

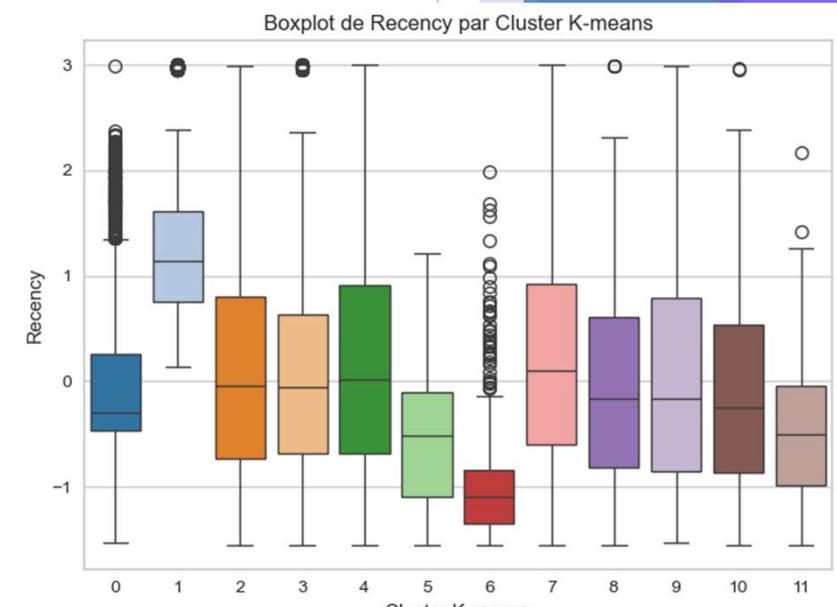


- Particularités :

- Pas de commandes depuis longtemps

- Proportion :

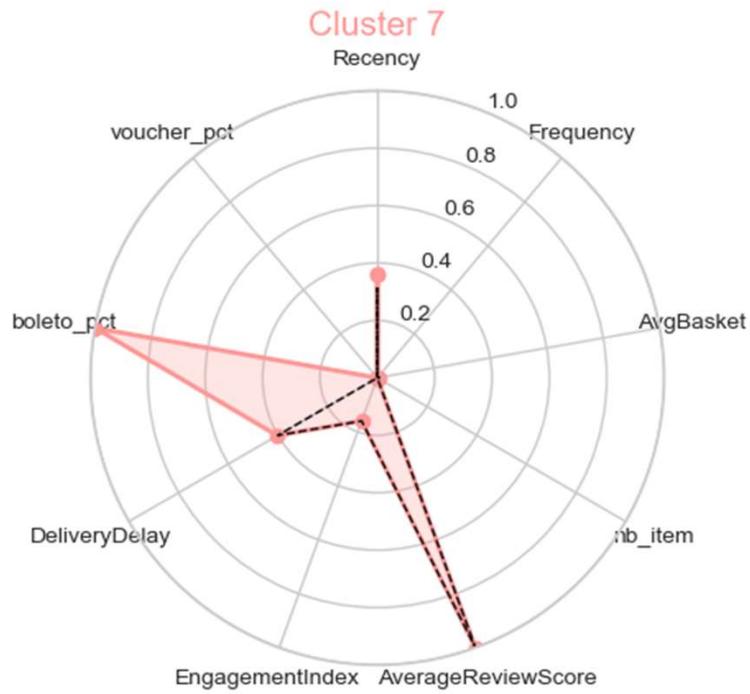
- 17144 clients
- 18,6 %



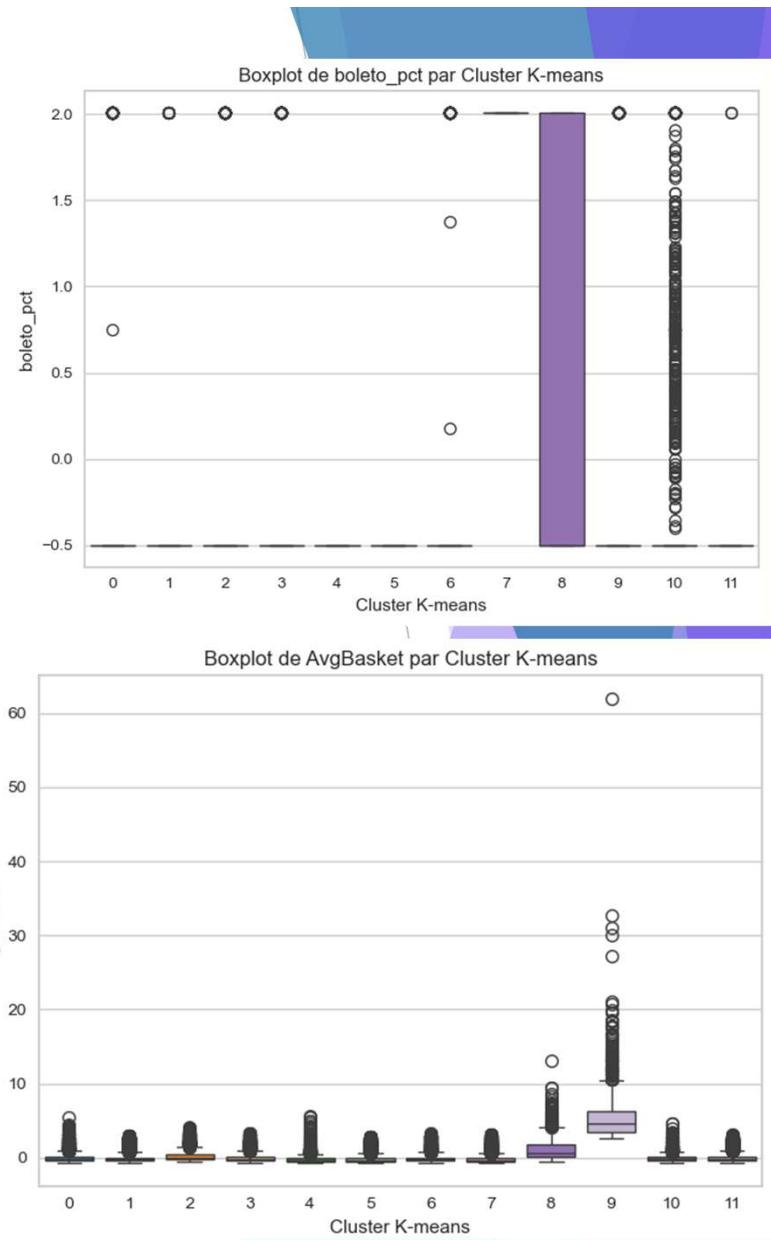
7. Profils

Les économiques

- Particularités :
 - Paye en Boletos



- Proportion :
 - 13720 clients
 - 14,8 %



7. Profils

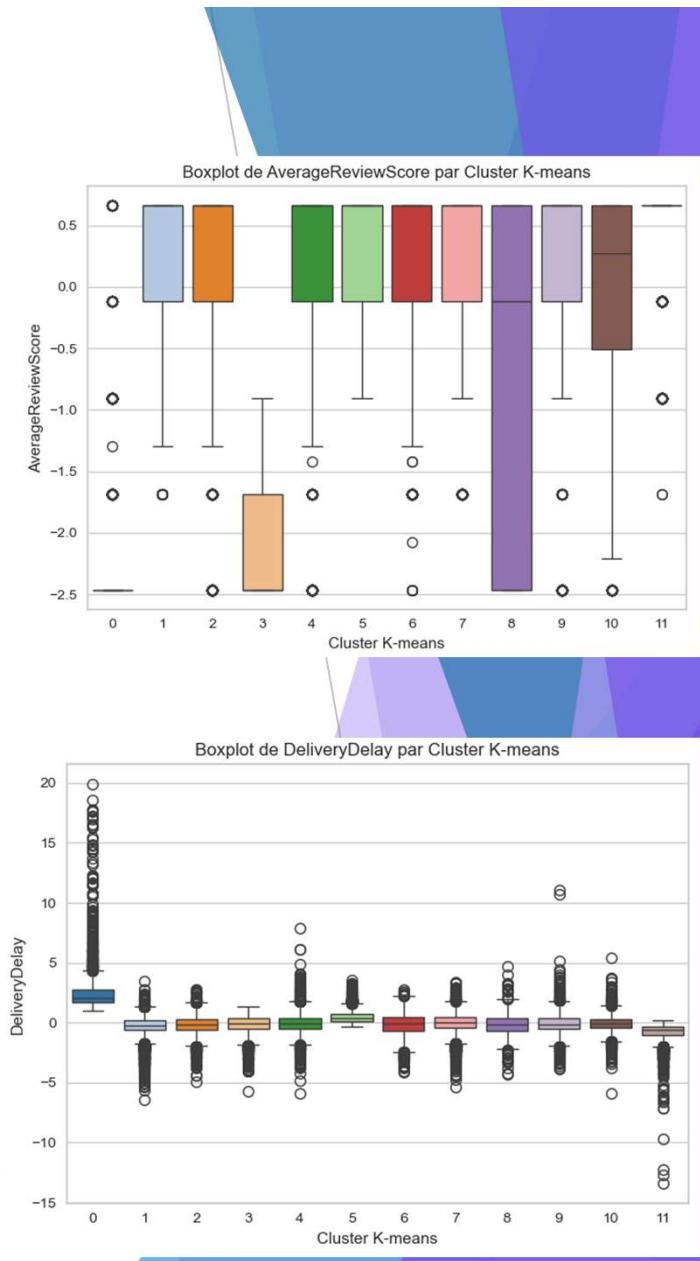
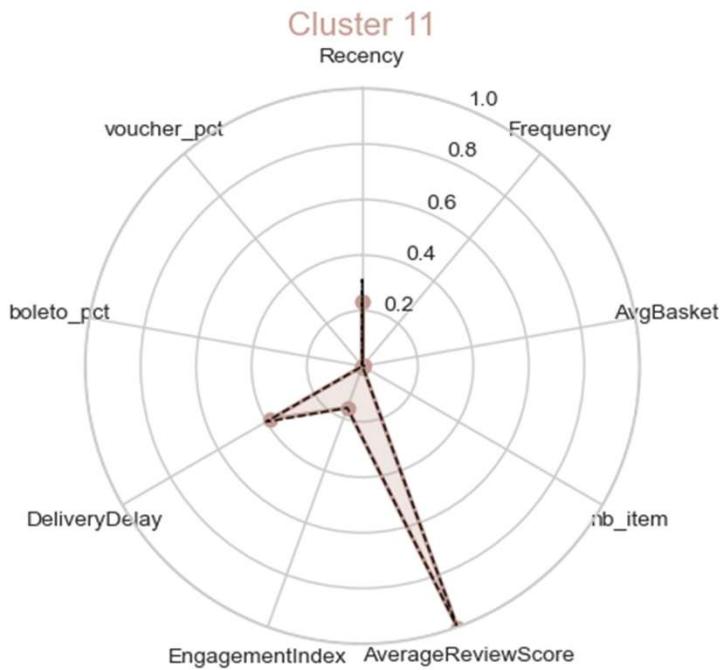
Les Satisfaits Privilégiés

- Particularités :

- Satisfaction très élevée
- Livraisons en avance

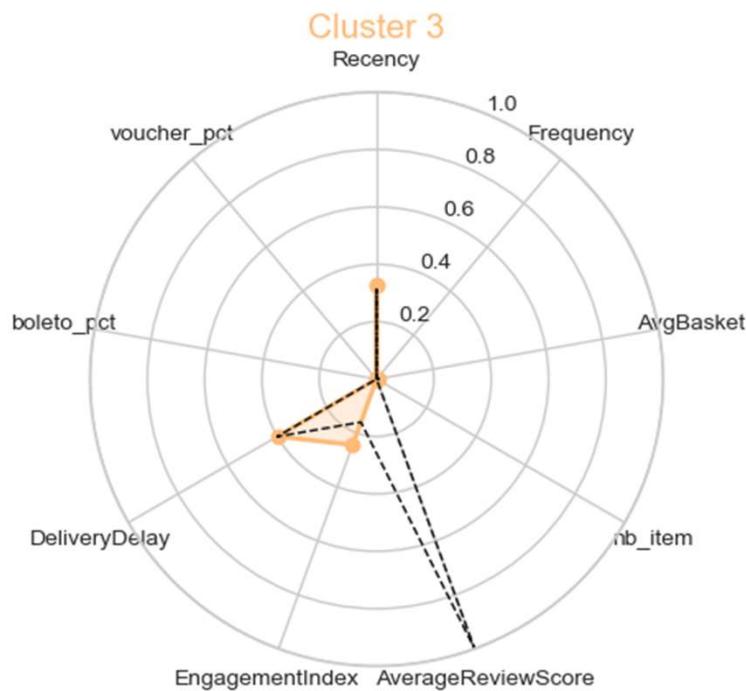
- Proportion :

- 13621 clients
- 14,7 %



7. Profils

Commentateurs Actifs Non Satisfaits

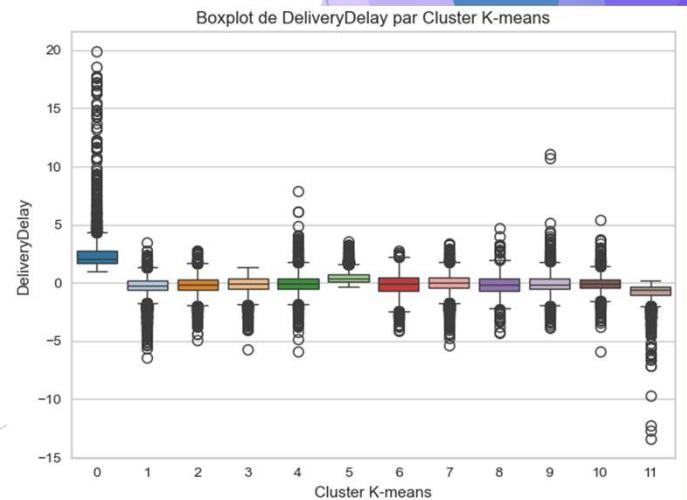
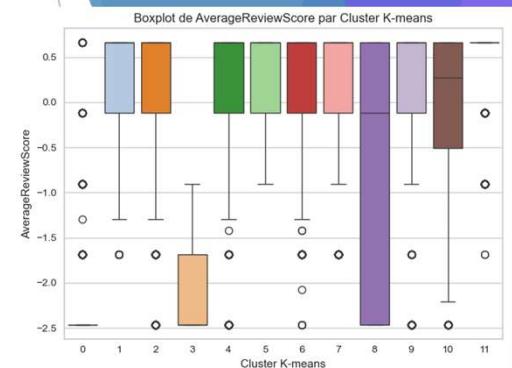


- Particularités :

- Clients mécontents
- Pas de retards de livraison
- Laisse plus de commentaires que la médiane

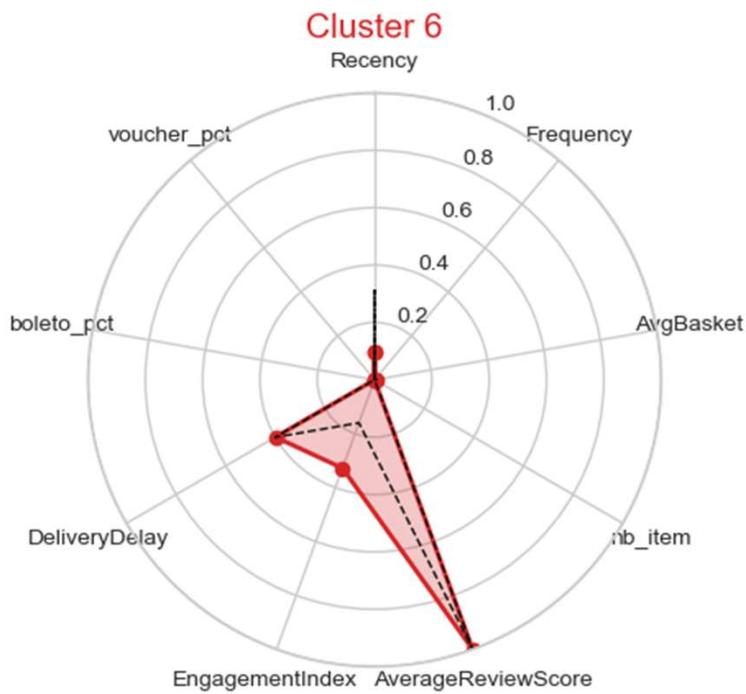
- Proportion :

- 6663 clients
- 7,2 %



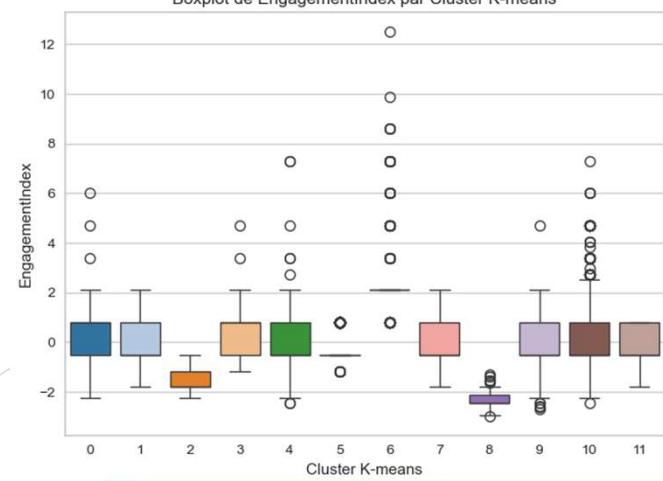
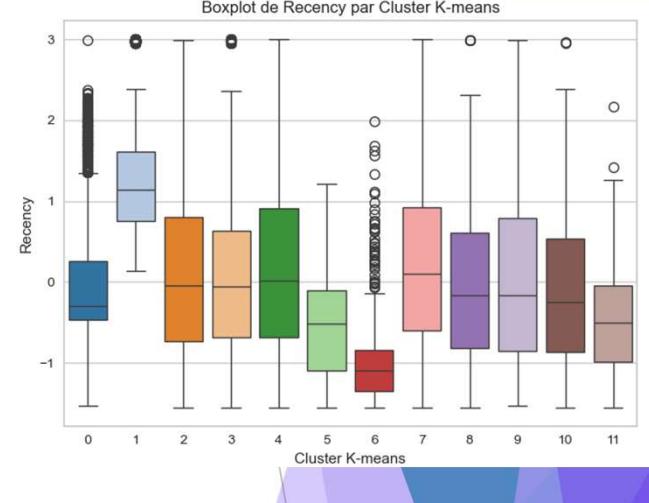
7. Profils

Commentateurs Actifs Récents



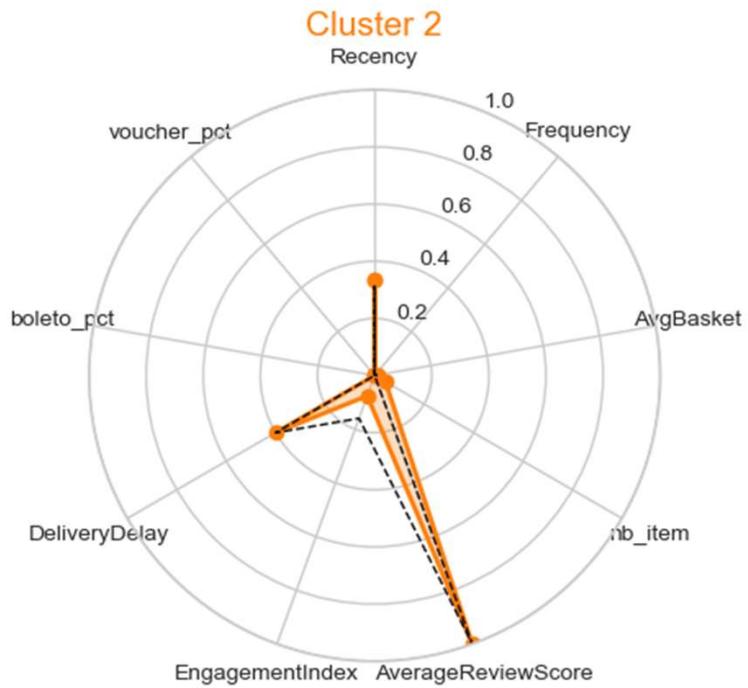
- Particularités :
 - Commande très récente
 - Laisse beaucoup de commentaires

- Proportion :
 - 6335 clients
 - 6,9 %



7. Profils

Acheteurs Silencieux

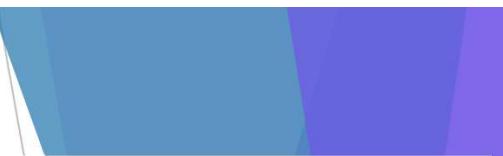


- Particularités :

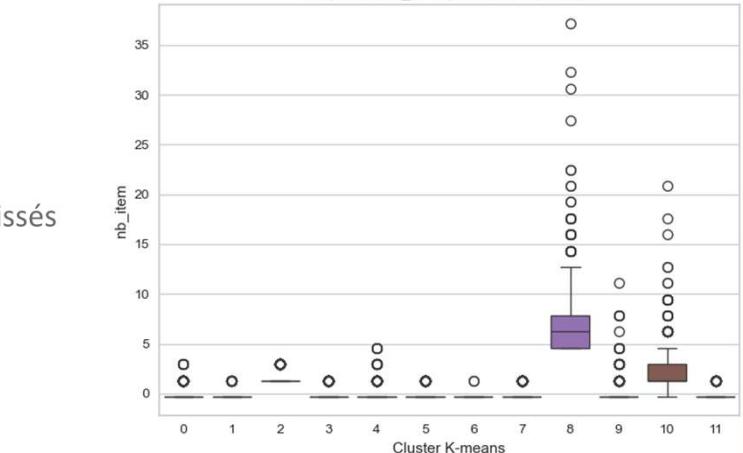
- Plus d'items que la médiane
- Très peu de commentaires laissés

- Proportion :

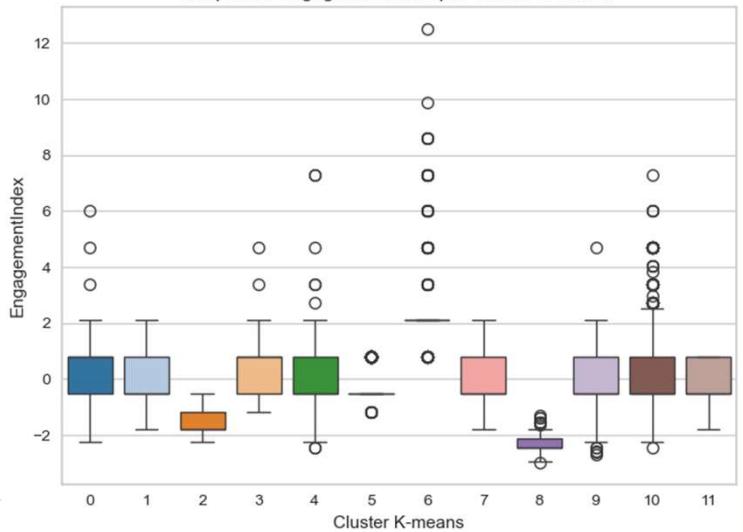
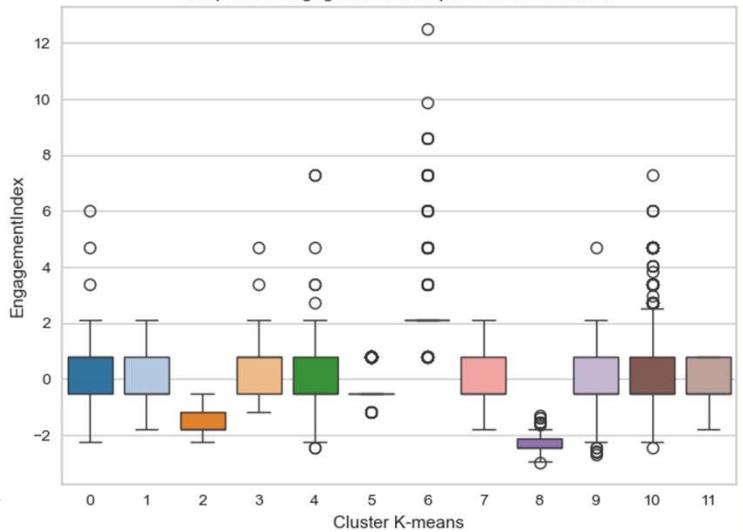
- 5487 clients
- 5,9 %



Boxplot de nb_item par Cluster K-means

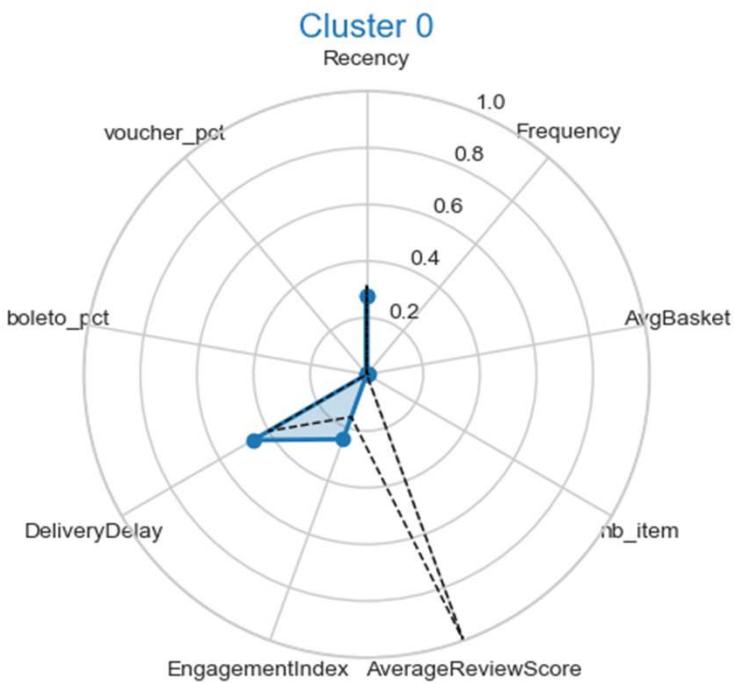


Boxplot de EngagementIndex par Cluster K-means



7. Profils

Clients Insatisfaits et Expressifs

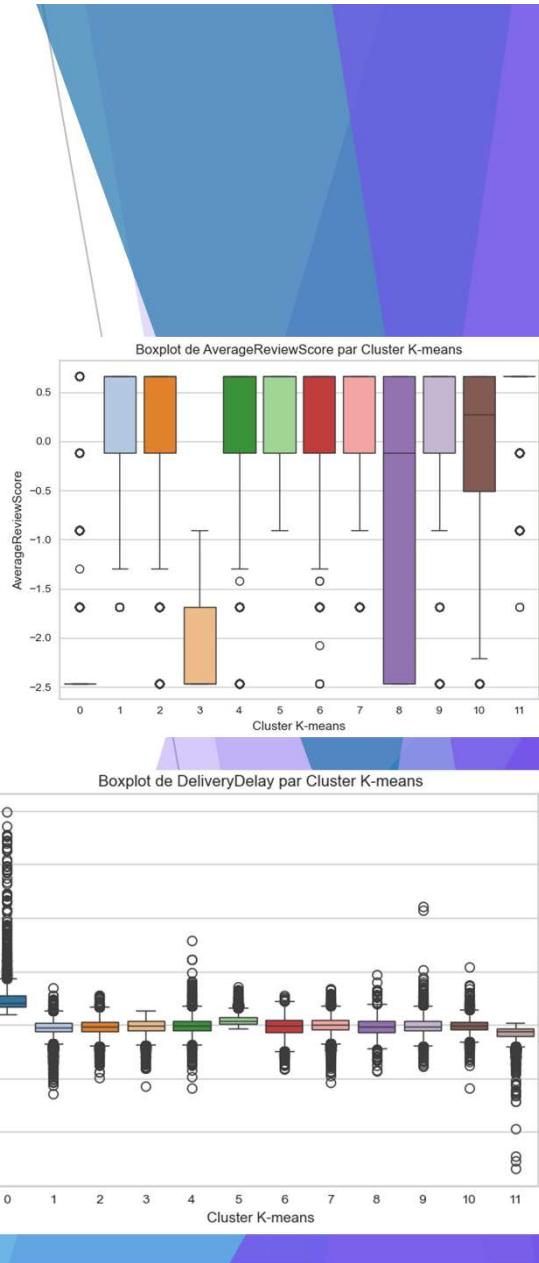


- Particularités :

- Satisfaction très faible
- Gros retards de livraison
- Laisse plus de commentaires que la médiane

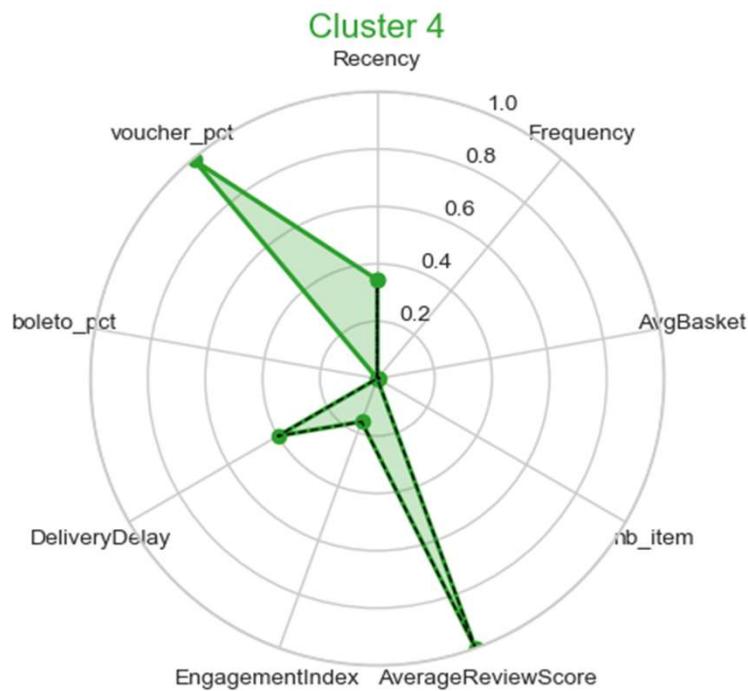
- Proportion :

- 4019 clients
- 4,3 %



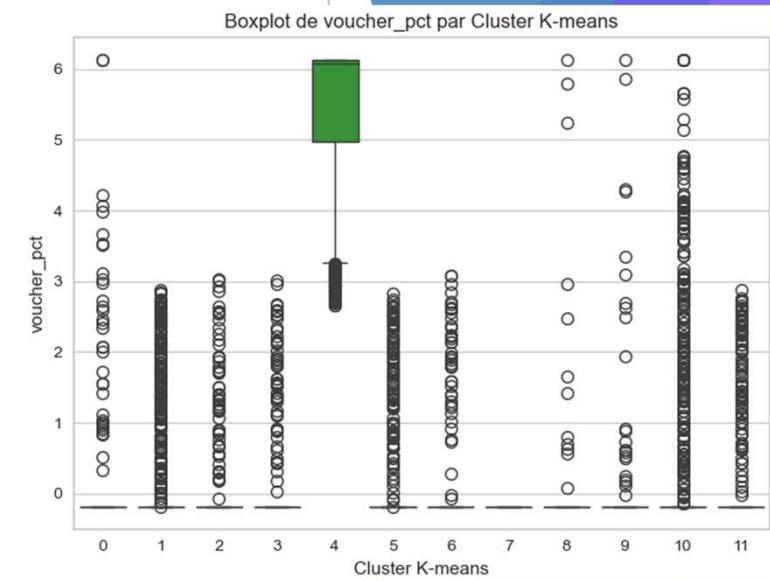
7. Profils

Voucher



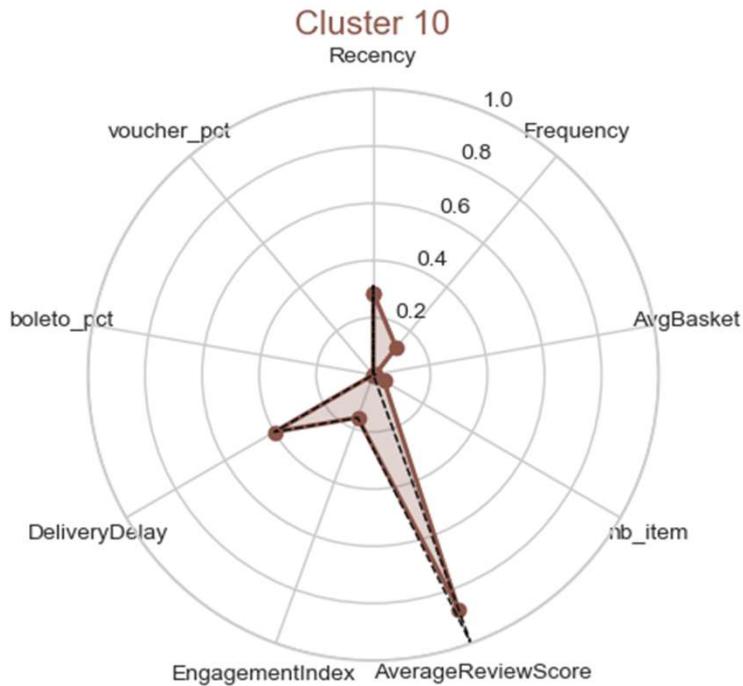
- Particularités :
 - Payent en vouchers

- Proportion :
 - 2791 clients
 - 3 %



7. Profils

Clients Fidèles

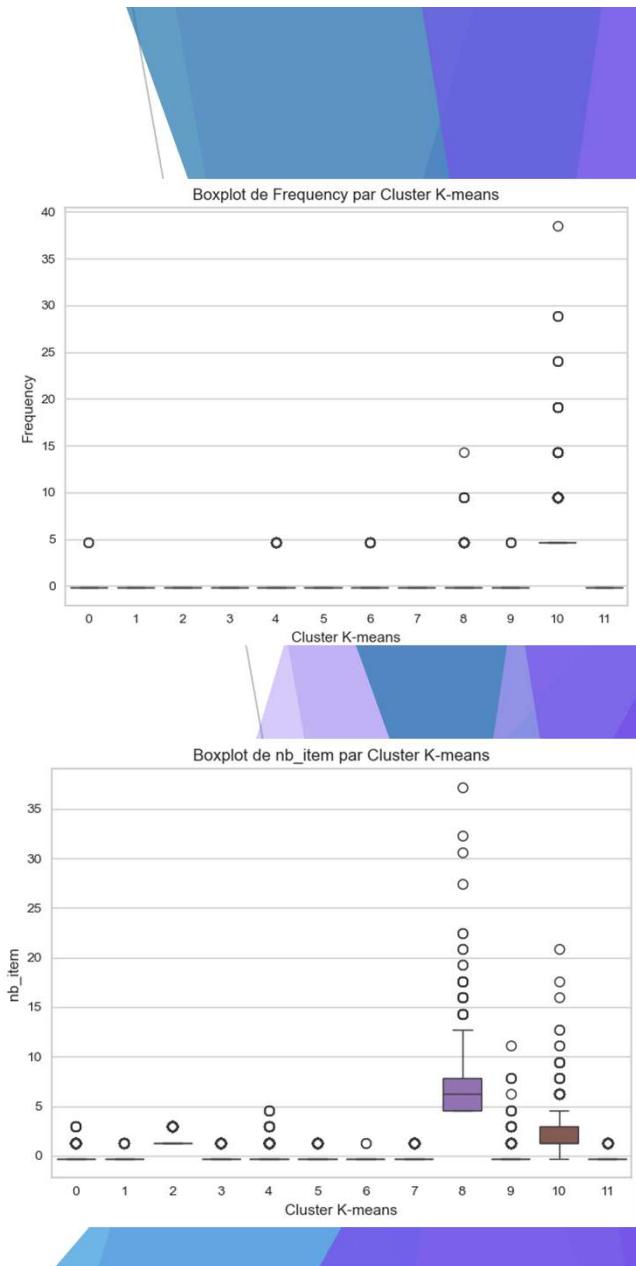


- Particularités :

- Achats fréquents
- Plus d'items que la médiane

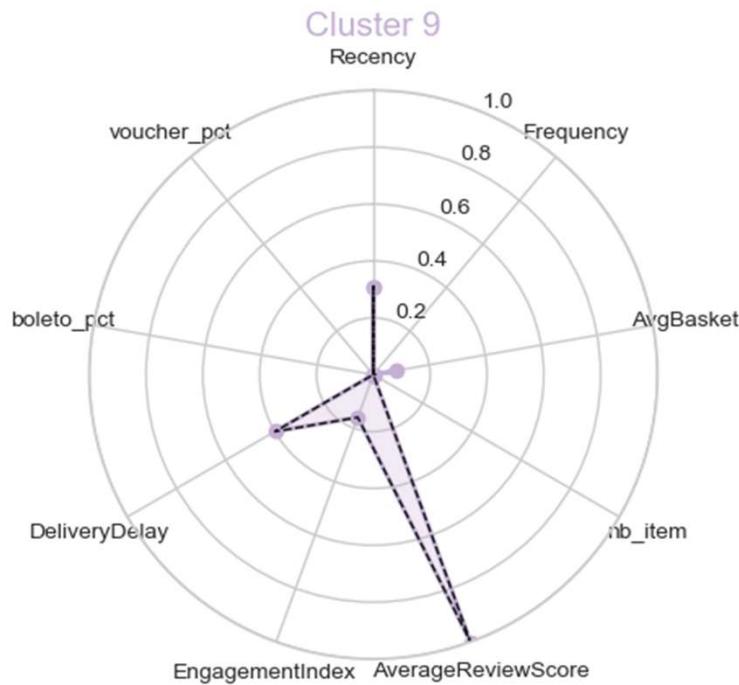
- Proportion :

- 2724 clients
- 2,9 %



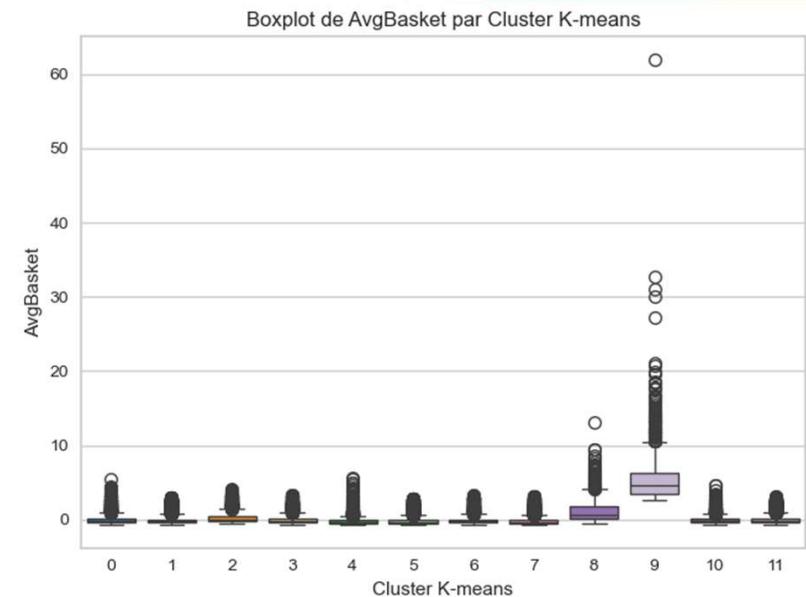
7. Profils

Clients à Forte Valeur



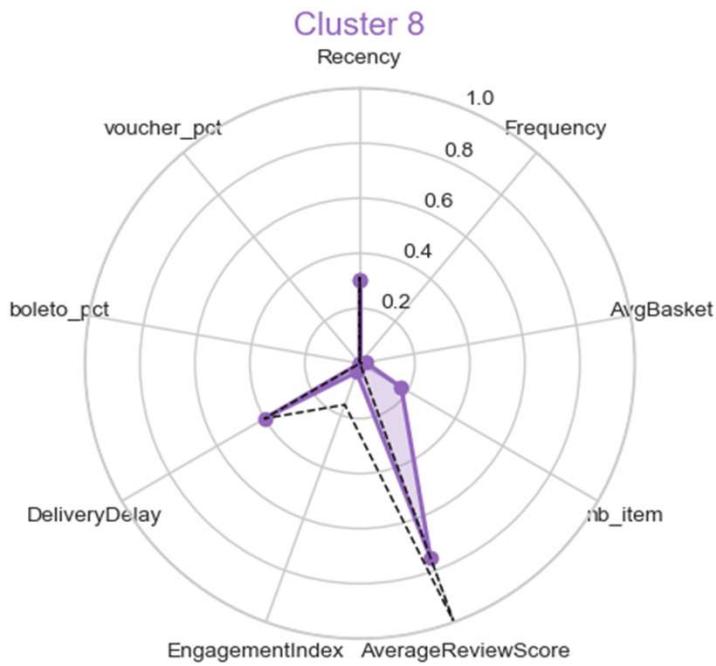
- Particularités :
 - Panier moyen beaucoup plus élevé

- Proportion :
 - 1459 clients
 - 1,6 %



7. Profils

Grands Acheteurs Discrets

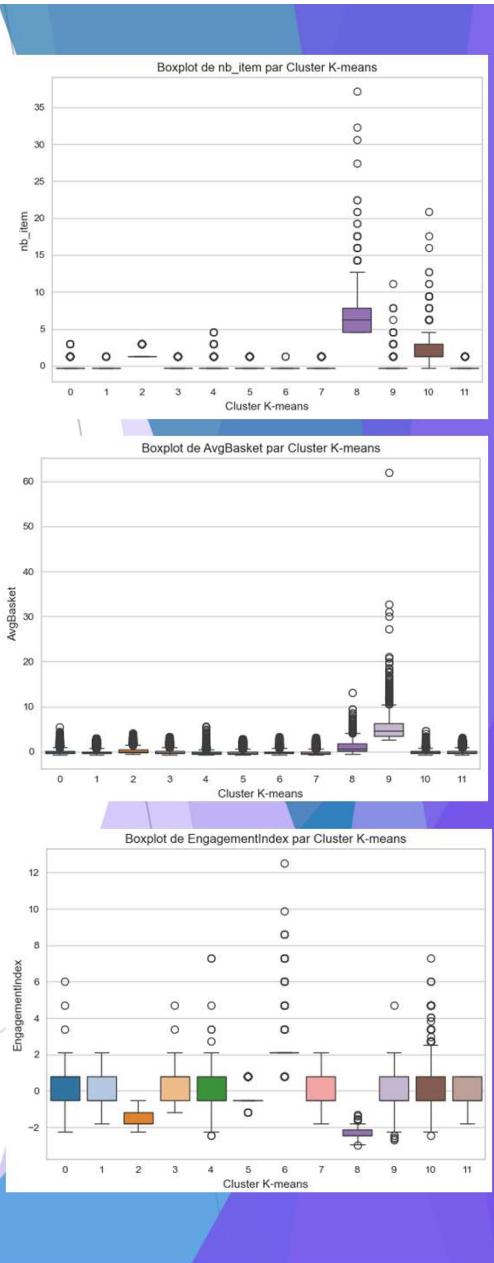


- Particularités :

- Achète beaucoup d'item
- Panier moyen plus élevé
- Laisse très peu de commentaires

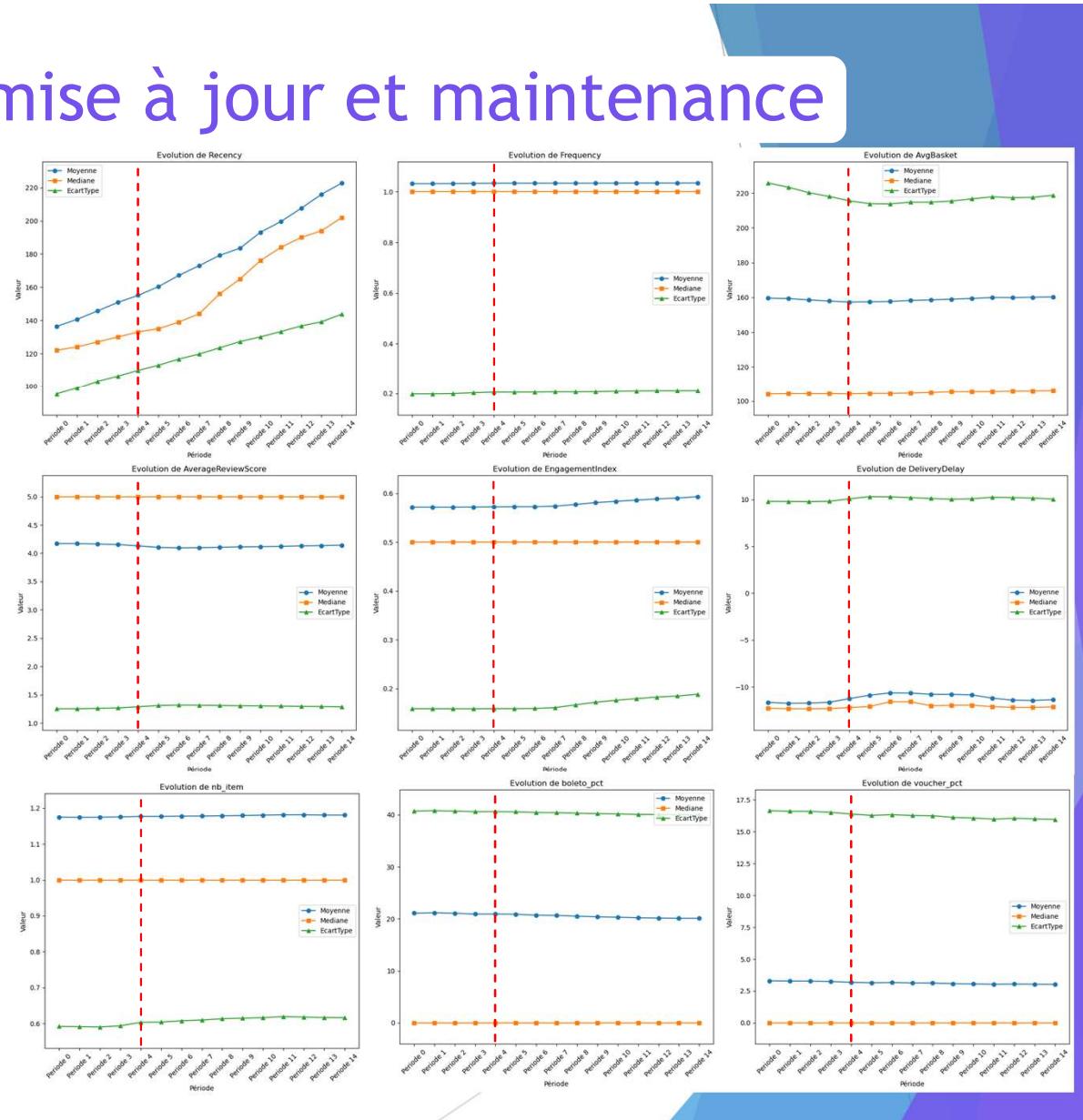
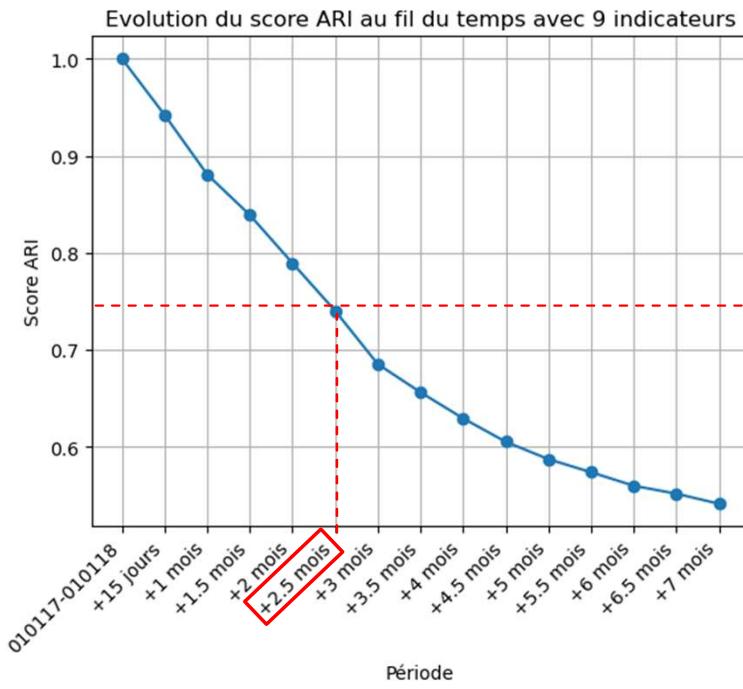
- Proportion :

- 880 clients
- 1 %



8. Stratégie de mise à jour et maintenance

- Ajout de nouveaux clients
 - Pipeline de traitement des données
 - Application du modèle
- Mise à jour tous les 2 mois et demi recommandée



Conclusion et Recommendations

• Synthèse

- **Segmentation Éclairée** : Identification de 12 clusters significatifs via le modèle K-means++, combinant pertinence statistique et insights métier.
- **Intelligence Client Accrue** : Profilage détaillé pour une approche marketing affinée et personnalisée.

• Recommandations

- **Personnalisation Marketing** : Adapter les communications en fonction des clusters pour augmenter l'engagement.
- **Réévaluation Régulière** : Revisiter le modèle tous les 2.5 mois maximums pour maintenir la pertinence des segments.
- **Intégration Continue** : Intégrer systématiquement les nouveaux clients dans les segments existants.
- **Surveillance des Métriques** : Monitorer les indicateurs clés pour une optimisation continue.
- **Analyse Approfondie** : Étudier les segments exceptionnels pour des actions ciblées.
- **Réaction Proactive** : Investiguer et répondre activement au feedback des segments critiques.

• Prochaines Étapes

- Utilisation des insights pour des campagnes marketing ciblées.
- Exploration d'opportunités de croissance basées sur la segmentation client.



Merci pour votre attention

Des questions ?

