

Segmentez des clients d'un site e-commerce

Jeux de données :

<https://www.kaggle.com/olistbr/brazilian-ecommerce>

Présentation par Hortense Monnard

Problématique :

Fournir une segmentation des clients aux équipes d'e-commerce pour un usage quotidien lors de campagnes de communication

Objectifs :

- **Comprendre les différents types d'utilisateurs** grâce à leur comportement et à leurs données personnelles ;
- **Fournir une description actionnable** de la segmentation des clients, avec sa logique sous-jacente pour une utilisation optimale ;
- **Proposer un contrat de maintenance** basé sur une analyse de la stabilité des segments au cours du temps.

PLAN

- 1. Présentation des Jeux de Données**
- 2. Analyses Exploratoires**
- 3. Création de Nouvelles Variables**
- 4. Sélection des Variables d'Intérêt et Découpage en Périodes Temporelles**
- 5. Clustering et Analyse Temporelle**
- 6. Conclusions et Perspectives**

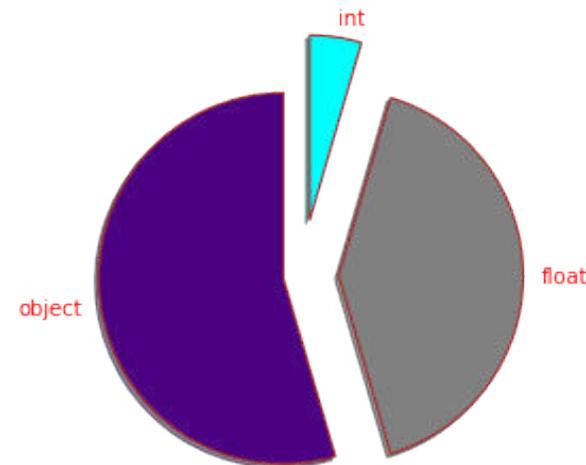
1. Présentation des Jeux de Données

1. Présentation des Jeux de Données

- Avant nettoyage -

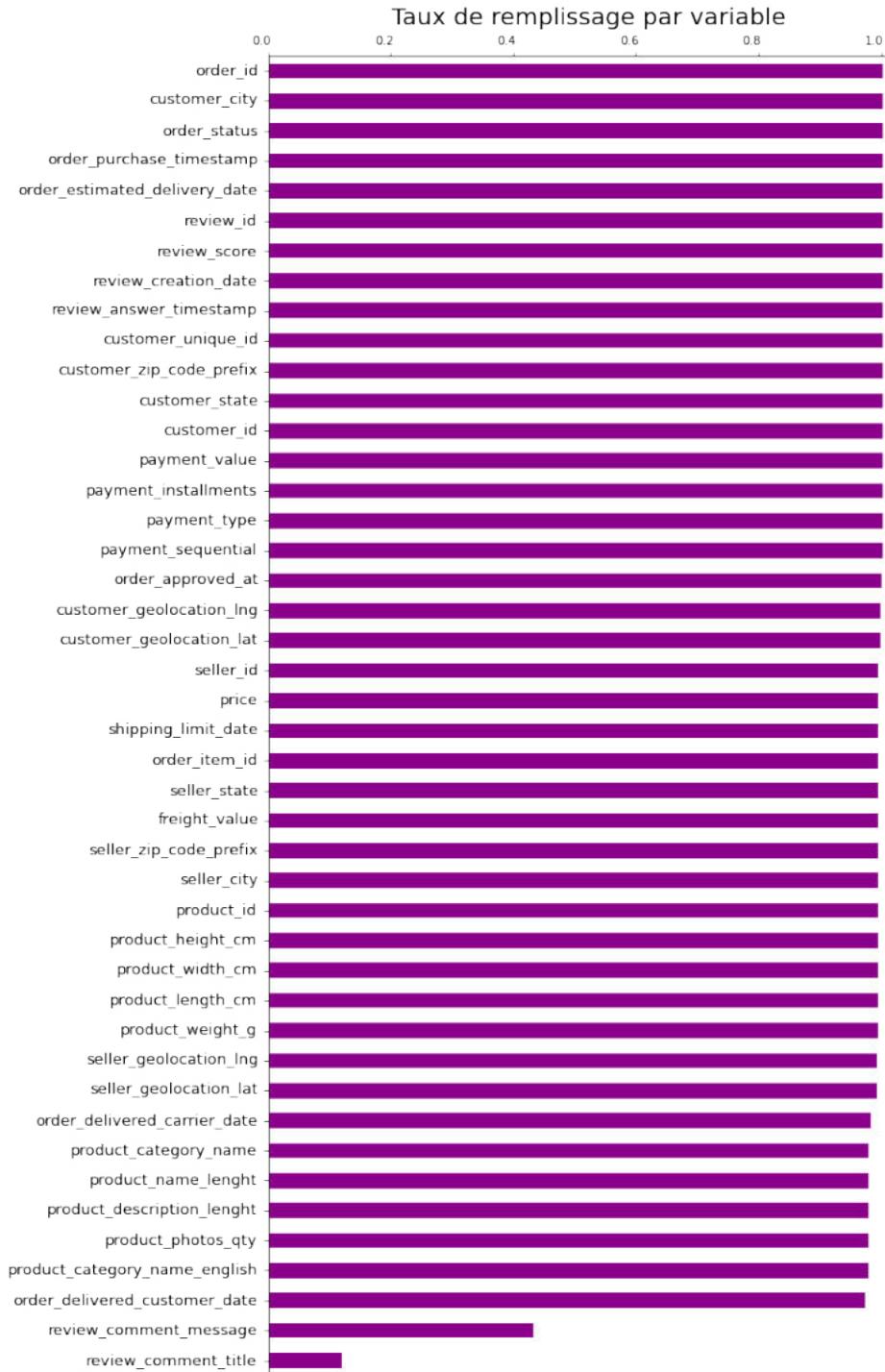
9 data sets à rassembler en 1 seul data set :

df_customers,
df_items,
df_reviews,
df_payments,
df_orders,
df_products,
df_sellers,
df_products_name,
df_geolocalisation.



→ Formation d'un data set de taille : 119151, 44

Taux de remplissage des variables	
order_id	1.000000
customer_city	1.000000
order_status	1.000000
order_purchase_timestamp	1.000000
order_estimated_delivery_date	1.000000
review_id	1.000000
review_score	1.000000
review_creation_date	1.000000
review_answer_timestamp	1.000000
customer_unique_id	1.000000
customer_zip_code_prefix	1.000000
customer_state	1.000000
customer_id	1.000000
payment_value	0.999975
payment_installments	0.999975
payment_type	0.999975
payment_sequential	0.999975
order_approved_at	0.998514
customer_geolocation_lng	0.997298
customer_geolocation_lat	0.997298
seller_id	0.993009
price	0.993009
shipping_limit_date	0.993009
order_item_id	0.993009
seller_state	0.993009
freight_value	0.993009
seller_zip_code_prefix	0.993009
seller_city	0.993009
product_id	0.993009
product_height_cm	0.992841
product_width_cm	0.992841
product_length_cm	0.992841
product_weight_g	0.992841
seller_geolocation_lng	0.990785
seller_geolocation_lat	0.990785
order_delivered_carrier_date	0.982493
product_category_name	0.978666
product_name_lenght	0.978666
product_description_lenght	0.978666
product_photos_qty	0.978666
product_category_name_english	0.978456
order_delivered_customer_date	0.971289
review_comment_message	0.430126
review_comment_title	0.119084



1. Présentation des Jeux de Données

- Nettoyage -

1. Créer des variables plus pertinentes
(ex: distance client/acheteur) ;
2. Filtrer les valeurs négatives pour : les prix des articles, les frais de port, les valeurs des commandes, le nombre de séquences de paiement, le nombre de paiements et le nombre d'article commandés ;
3. Changer le format des dates et extraire des variables temporelles ;
4. Supprimer les lignes où il y a des données manquantes pour les variables sélectionnées.

→ Data set final de taille : 116006, 12
(97% des lignes du data set initial)

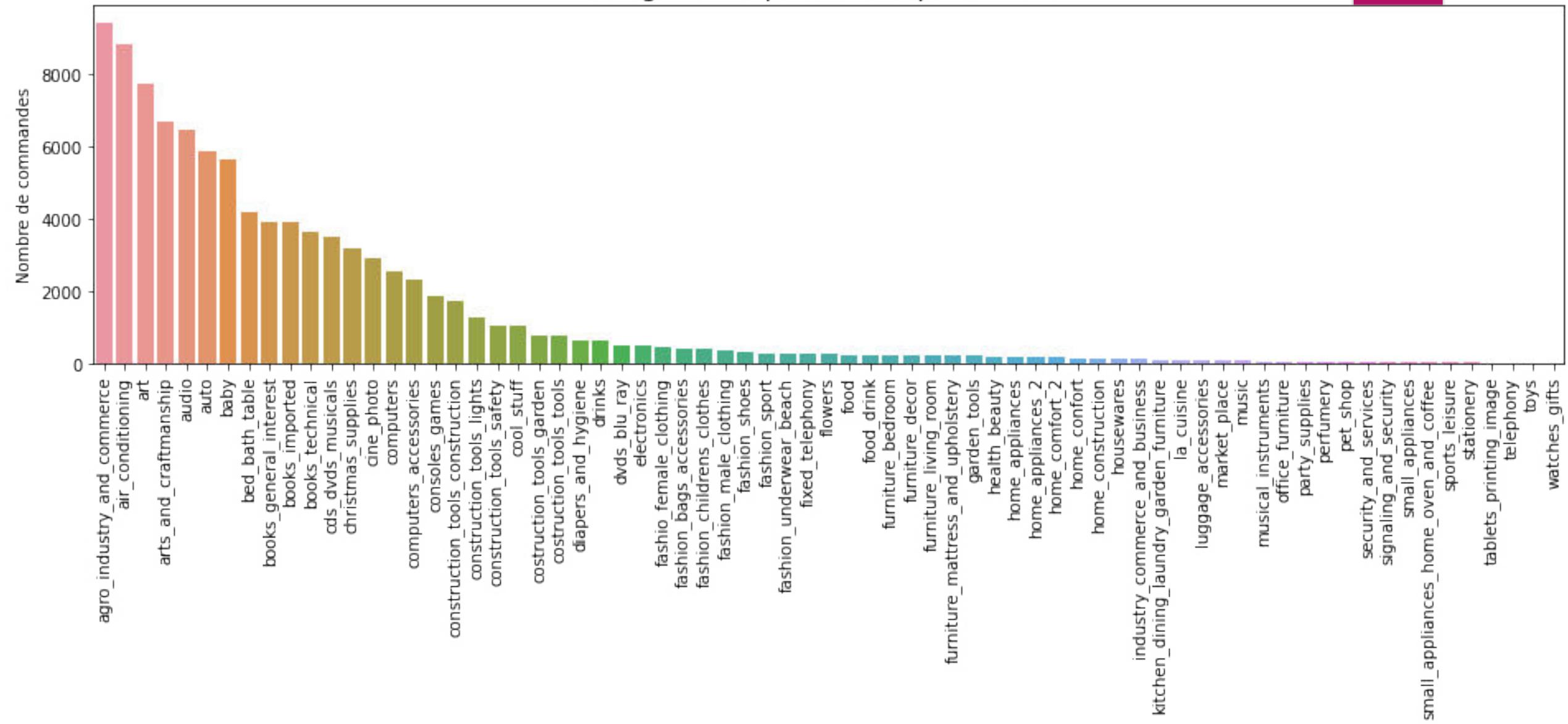
2. Analyses Exploratoires

2. Analyses Exploratoires

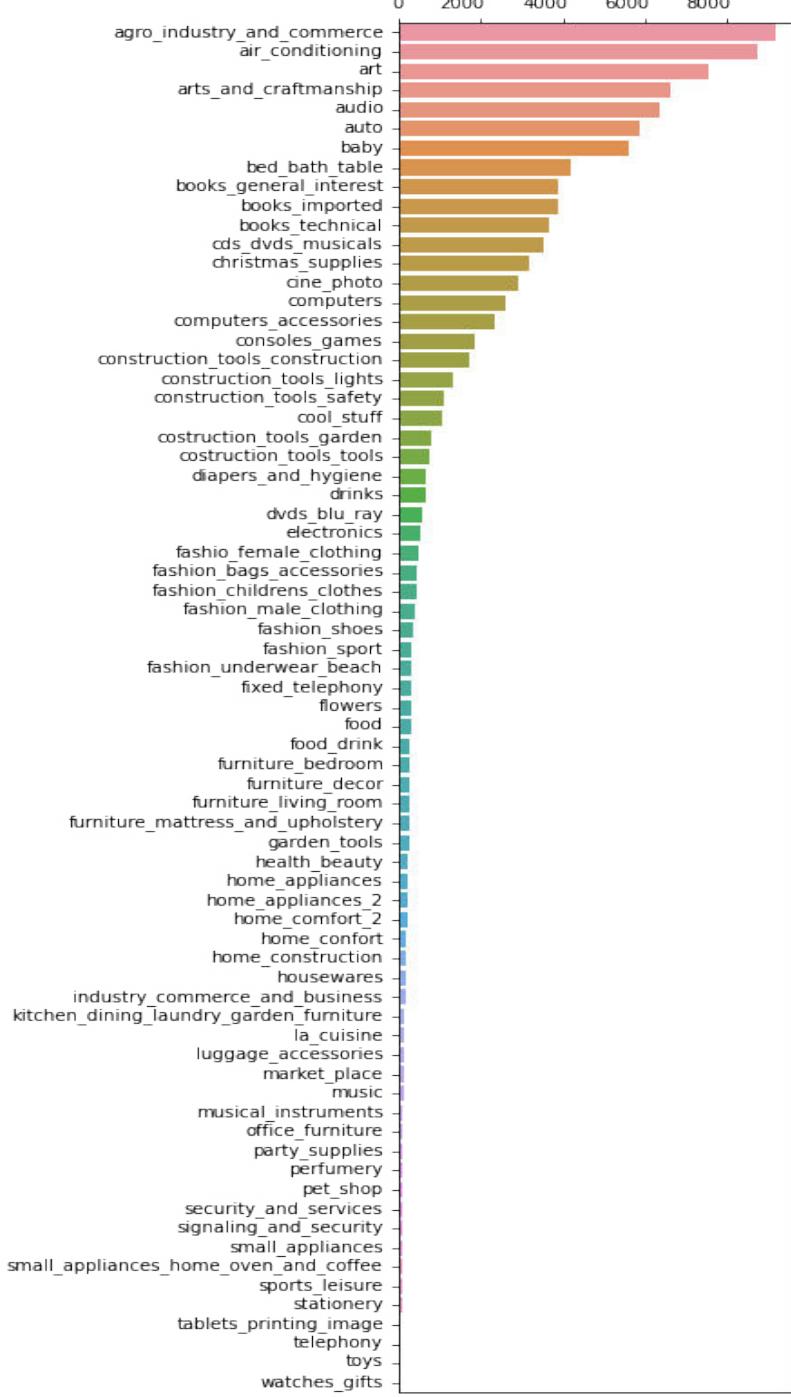
Informations générales sur les commandes :

- La date des commandes va du 2016-09-04 au 2018-10-17
- Le nombre de transactions différentes est : 99441
- Le nombre de clients différents est : 96096
- Le nombre de produits différents est : 32951
- Le nombre total de transactions approuvées est : 90733
- La valeur totale des transactions est : 20581109.62 réal brésilien
- Le prix des commandes va de 0.0 à 13664.08 réal brésilien
- Le prix des produits va de 0.85 à 6735.0 réal brésilien
- Le prix des livraisons va de 0.0 à 409.68 réal brésilien
- La distance entre un client et un vendeur va de 0.0 à 409.68 km

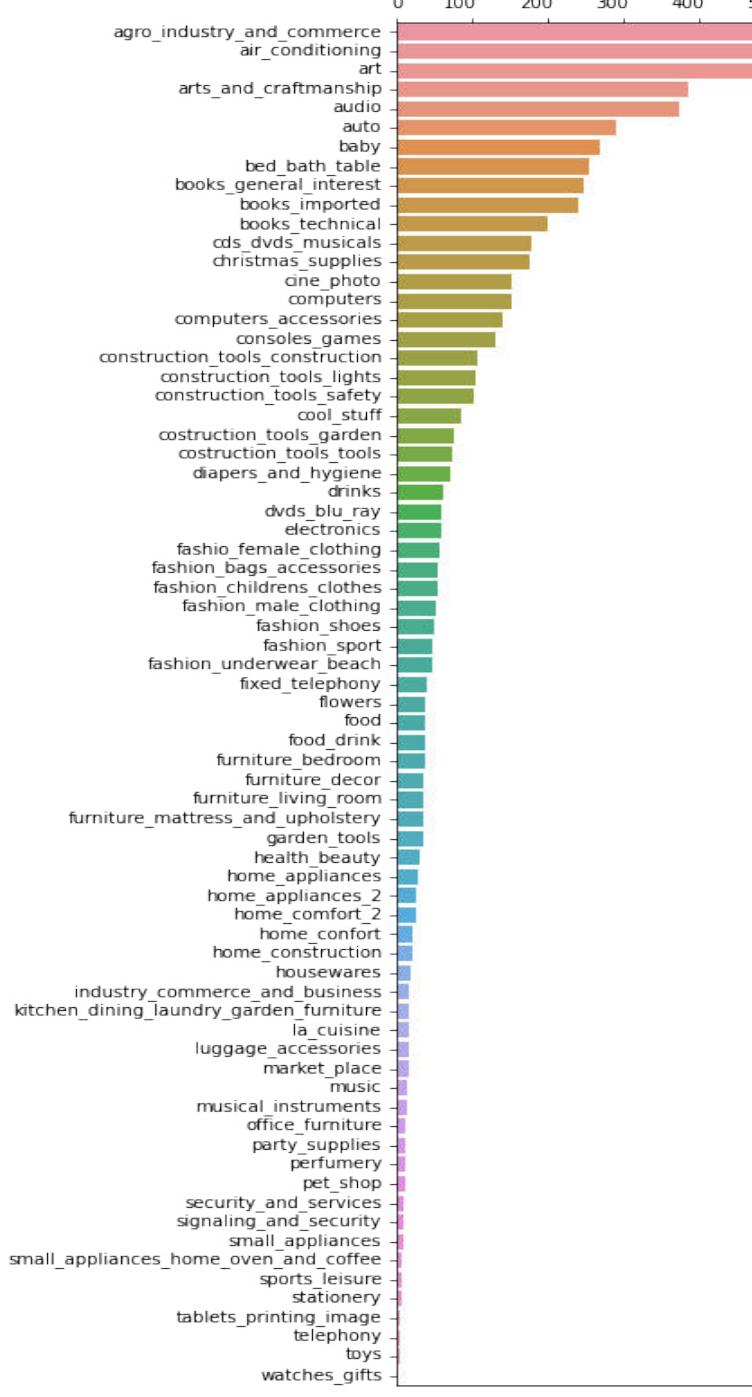
Catégories de produits les plus achetés



Nombre d'acheteurs par catégorie de produits



Nombre de vendeurs par catégorie de produits

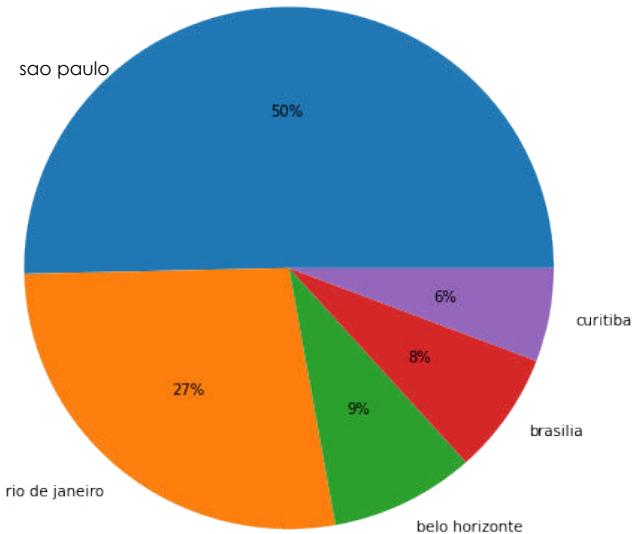


Villes des acheteurs

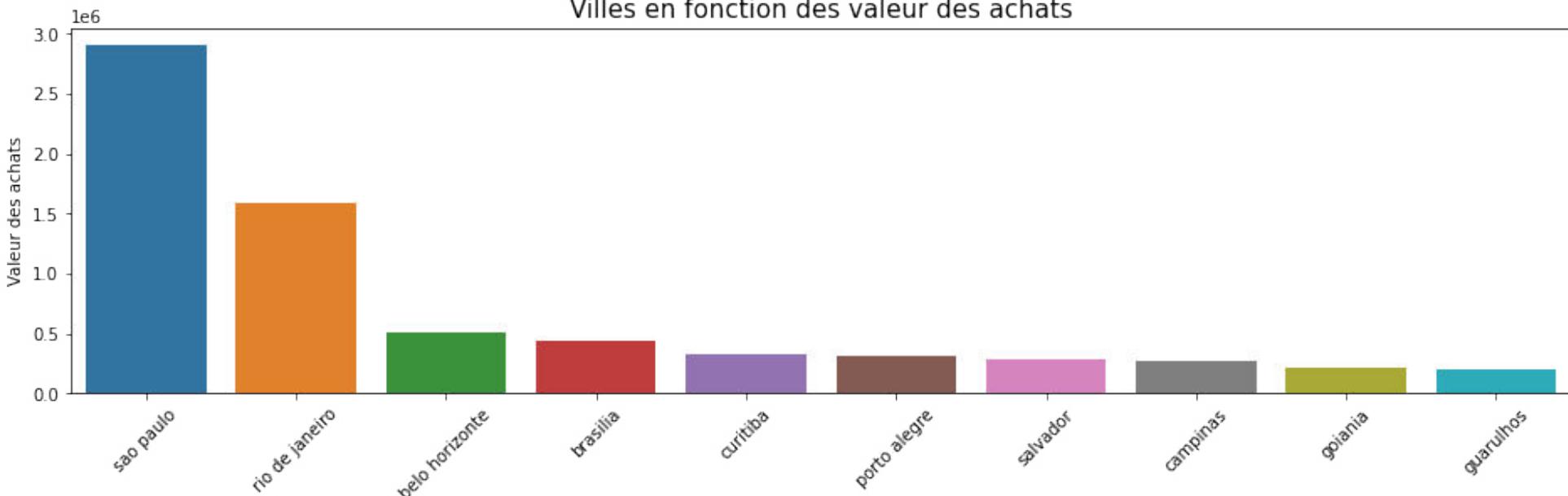
La moitié de la valeur des achats provient de clients vivants à São Paulo.

Cette grande ville semble représenter un marché important.

Top 5 des villes par valeur des achats



Villes en fonction des valeur des achats



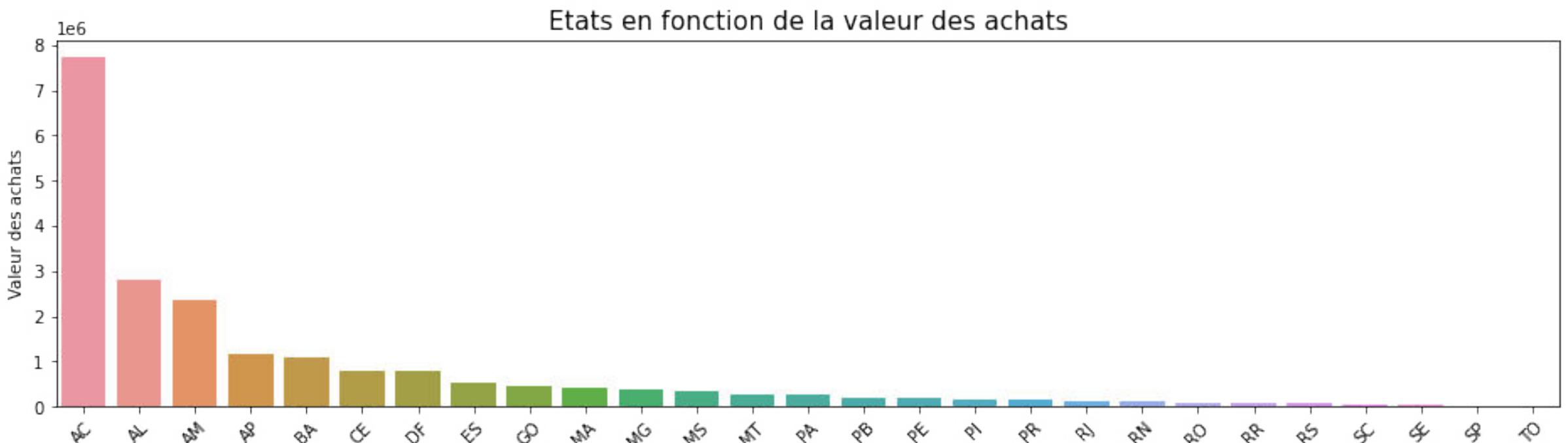
Etats des acheteurs

La majorité des achats s'effectue dans les états AC, AL, AM.

AC: Acre

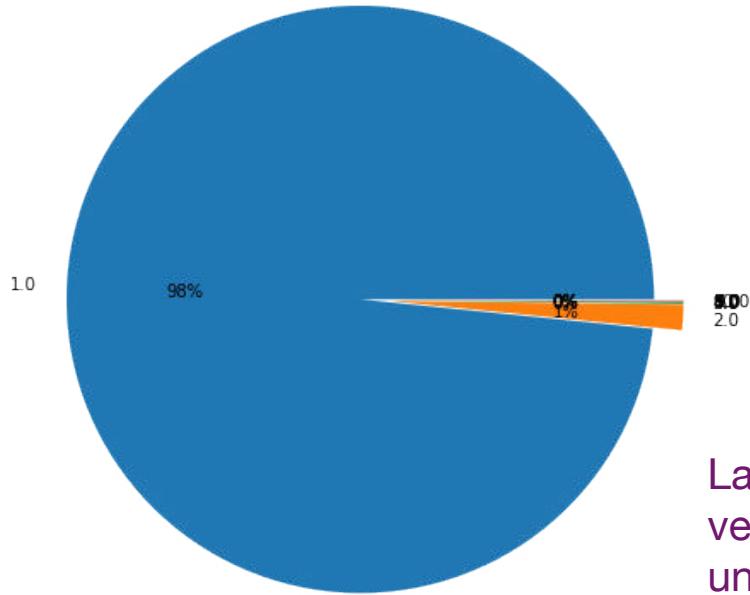
AL: Alagoas

AM: Amazonas



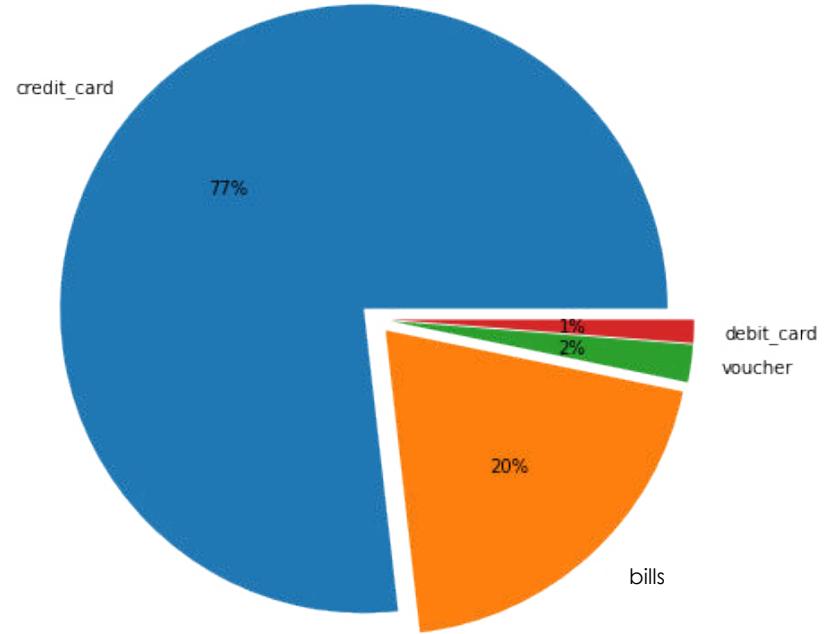
Types de paiement et nombre de séquences des transactions

Nombre de séquence de paiement par transactions



La majorité des ventes est payée en une seule séquence.

Types de paiement par valeurs de transactions



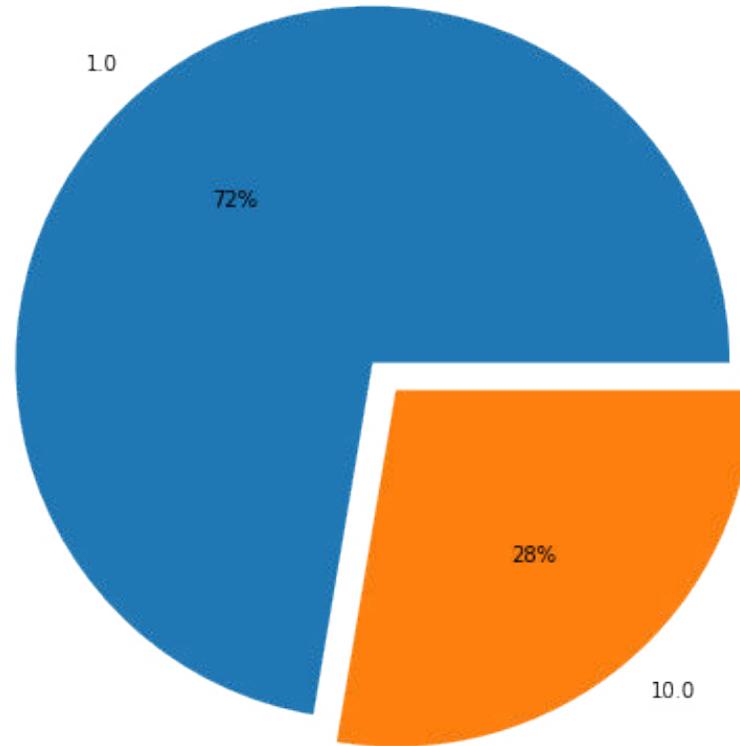
Les ¾ des paiements s'effectuent par carte de crédit.



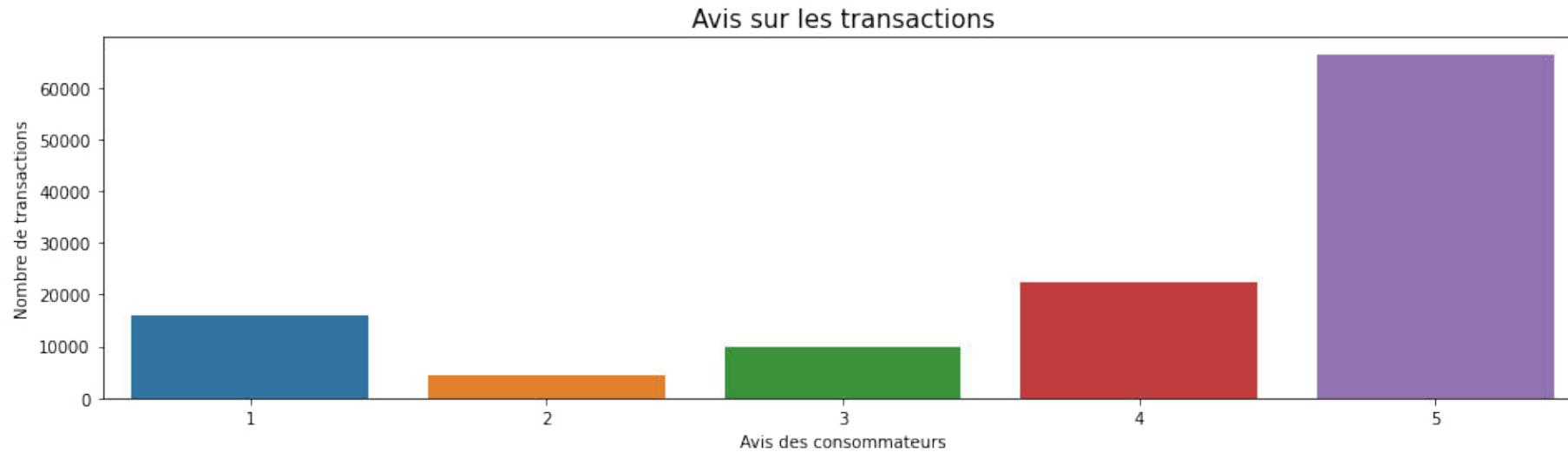
Paiement en plusieurs fois des transactions

Même si une majorité des transactions s'effectue avec un paiement en 1 seule fois, le paiement en 10 fois représente 28% de la valeur des transactions.

Paiement en plusieurs fois par transactions

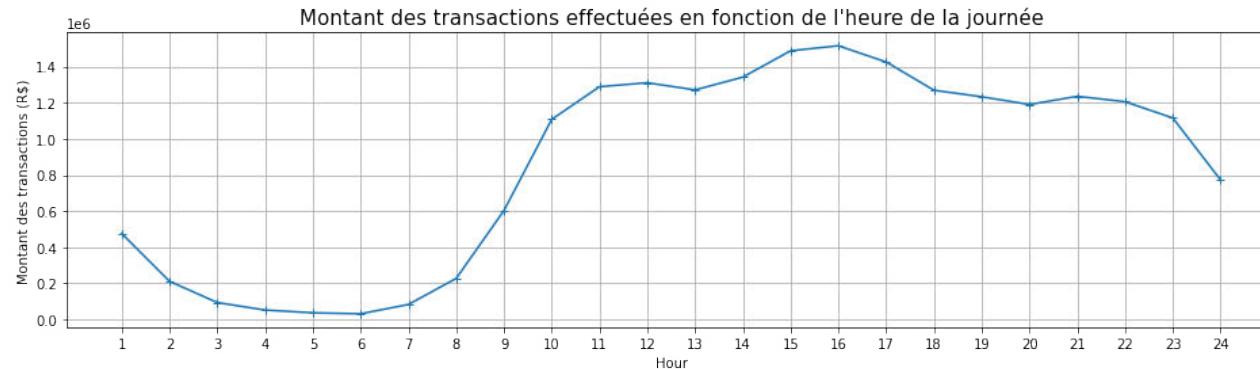


Satisfaction des clients

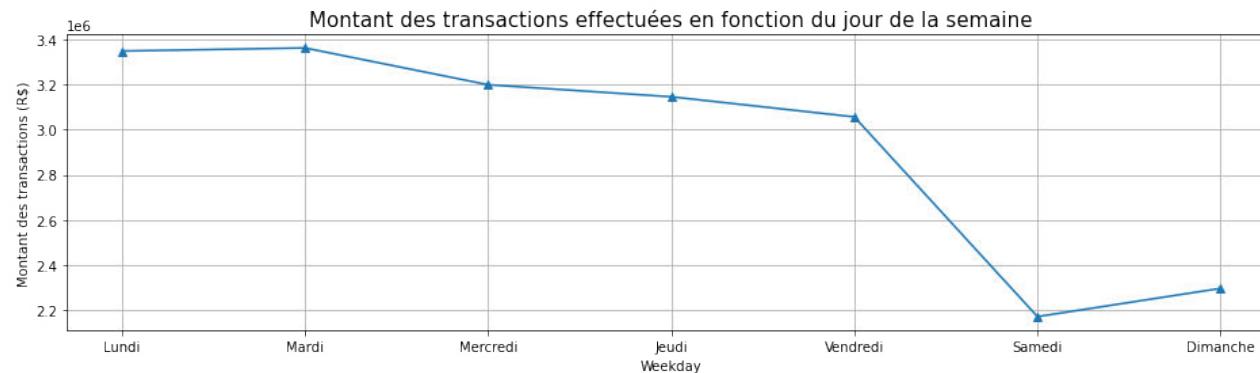


De la même manière, la grande majorité des catégories de produits ont leur plus grand nombre d'avis consommateurs à 5.

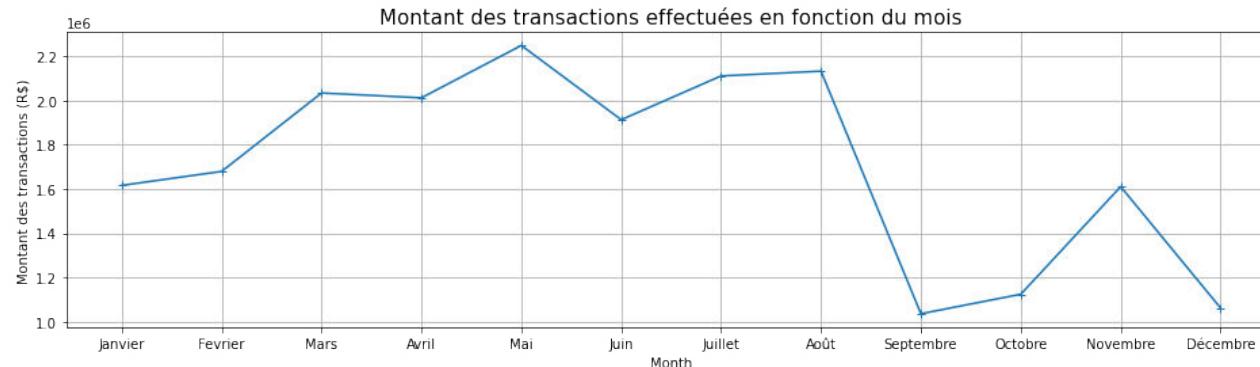
Moments d'achats



Les ventes sont meilleures en journée, entre 10h et 22h.



Les ventes sont meilleures en début de semaine et baissent durant le weekend.



Les ventes sont meilleures dans les mois allant de mars à août. Elles baissent drastiquement en septembre.

3. Cr ation de Nouvelles Variables

3. Création de Nouvelles Variables

Pour chaque client :

- Jours_Depuis_Dernier_Achat,
- Achats_Totaux,
- Avis_Moyen,
- Nb_Séquences_Paiement_Moyen,
- Nb_Categories_Produits,
- Nb_Transactions,
- Nombre_Vendeurs_Par_Commande,
- Distance_Sel_Cus,
- Panier_Moyen,
- Ratio_Frais_Panier_Moyen.

4. Sélection des Variables d'Intérêt et Découpage en Périodes Temporelles

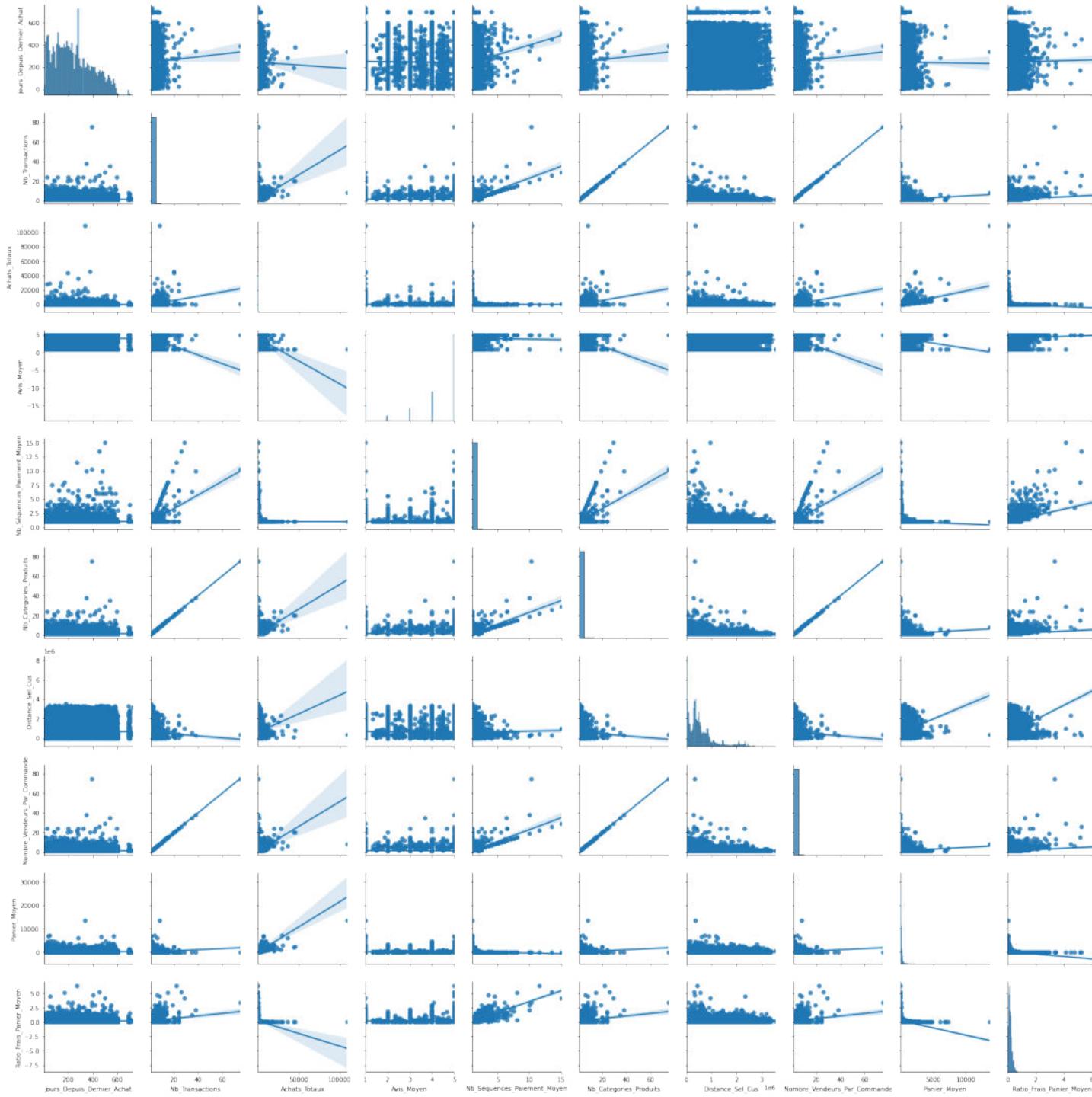
Observation de la corrélation des variables entre elles

3 variables corrélées :

- Nb de catégories de produits
- Nb de transactions
- Nb de vendeurs par commande

3 variables corrélées :

- Jours depuis dernier achat
- Avis Moyen
- Distance client/vendeur



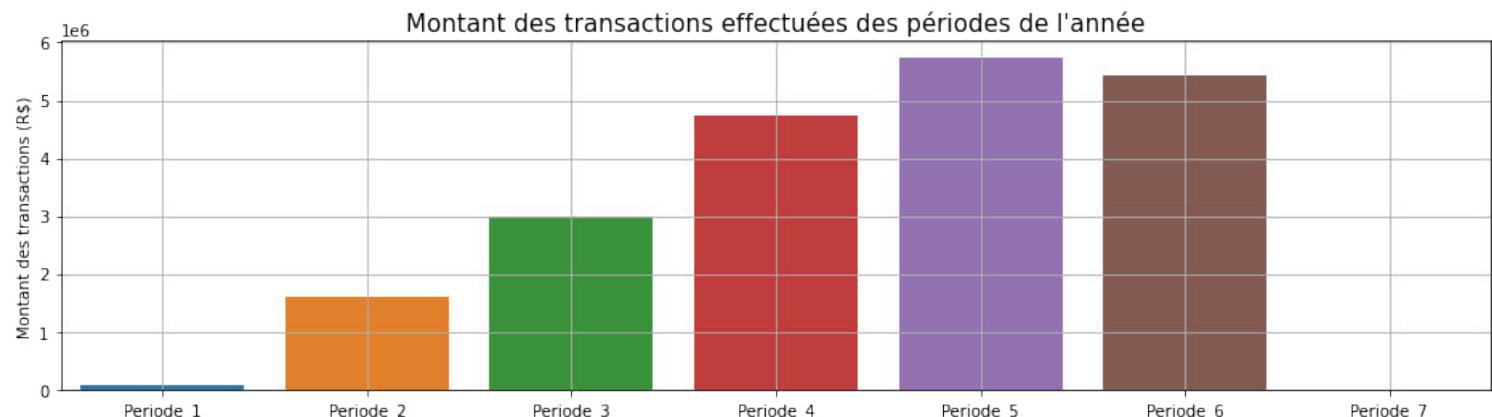
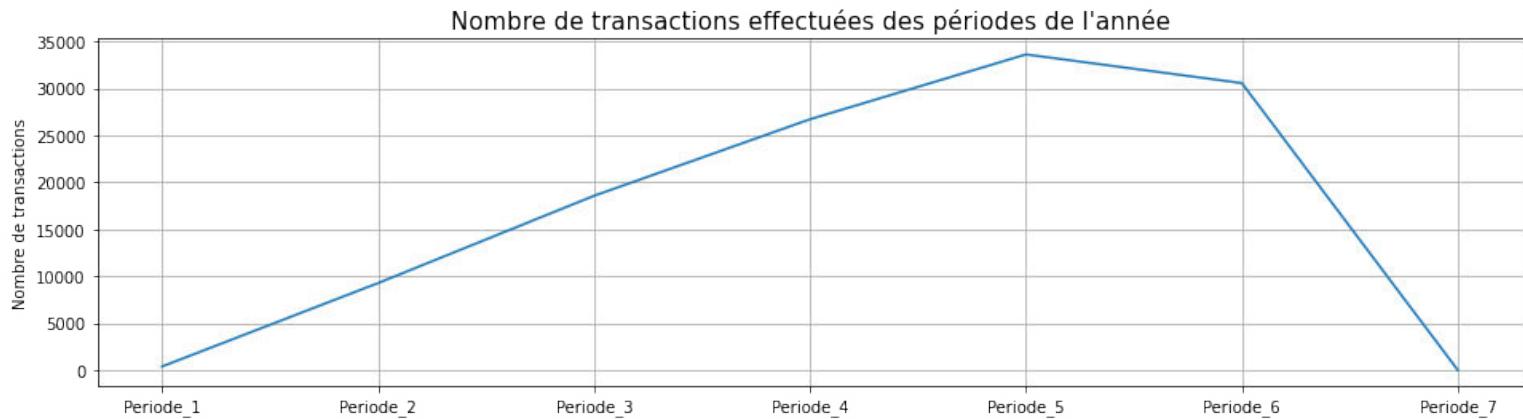
4. Sélection des Variables d'Intérêt

Variables sélectionnées

- ❖ 'Jours_Depuis_Dernier_Achat',
- ❖ 'Achats_Totaux',
- ❖ 'Panier_Moyen',
- ❖ 'Nb_Séquences_Paiement_Moyen',
- ❖ 'Nb_Categories_Produits',
- ❖ 'Ratio_Frais_Panier_Moyen'.

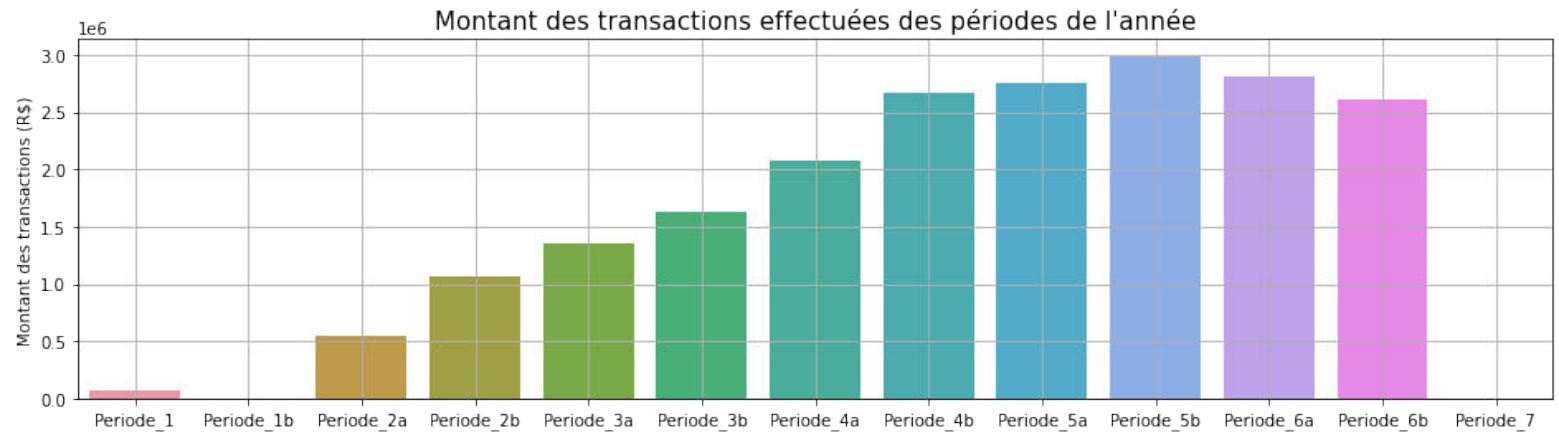
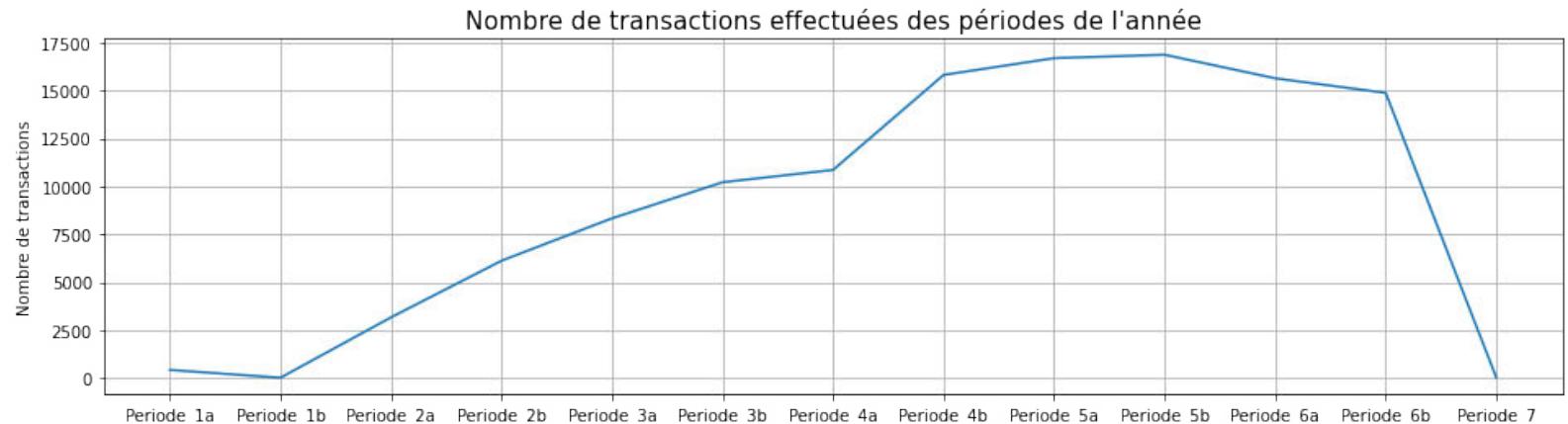
4. Découpage en Périodes Temporelles - Par périodes de 4 mois

- Periode_1 = 2016-09-04 au 2016-12-31
- Periode_2 = 2017-01-01 au 2017-04-31
- Periode_3 = 2017-05-01 au 2017-08-31
- Periode_4 = 2017-09-01 au 2017-12-31
- Periode_5 = 2018-01-01 au 2018-04-31
- Periode_6 = 2018-05-01 au 2018-08-31
- Periode_7 = 2018-09-01 au 2018-10-17



4. Découpage en Périodes Temporelles - Par périodes de 2 mois

- Periode_1a = 2016-09-04 to 2016-10-31
- Periode_1b = 2016-11-04 to 2016-12-31
- Periode_2a = 2017-01-01 to 2017-02-28
- Periode_2b = 2017-03-01 to 2017-04-30
- Periode_3a = 2017-05-01 to 2017-06-30
- Periode_3b = 2017-07-01 to 2017-08-31
- Periode_4a = 2017-09-01 to 2017-10-31
- Periode_4b = 2017-11-01 to 2017-12-31
- Periode_5a = 2018-01-01 to 2018-02-28
- Periode_5b = 2018-03-01 to 2018-04-30
- Periode_6a = 2018-05-01 to 2018-06-30
- Periode_6b = 2018-07-01 to 2018-08-31
- Periode_7 = 2018-09-01 to 2018-10-17

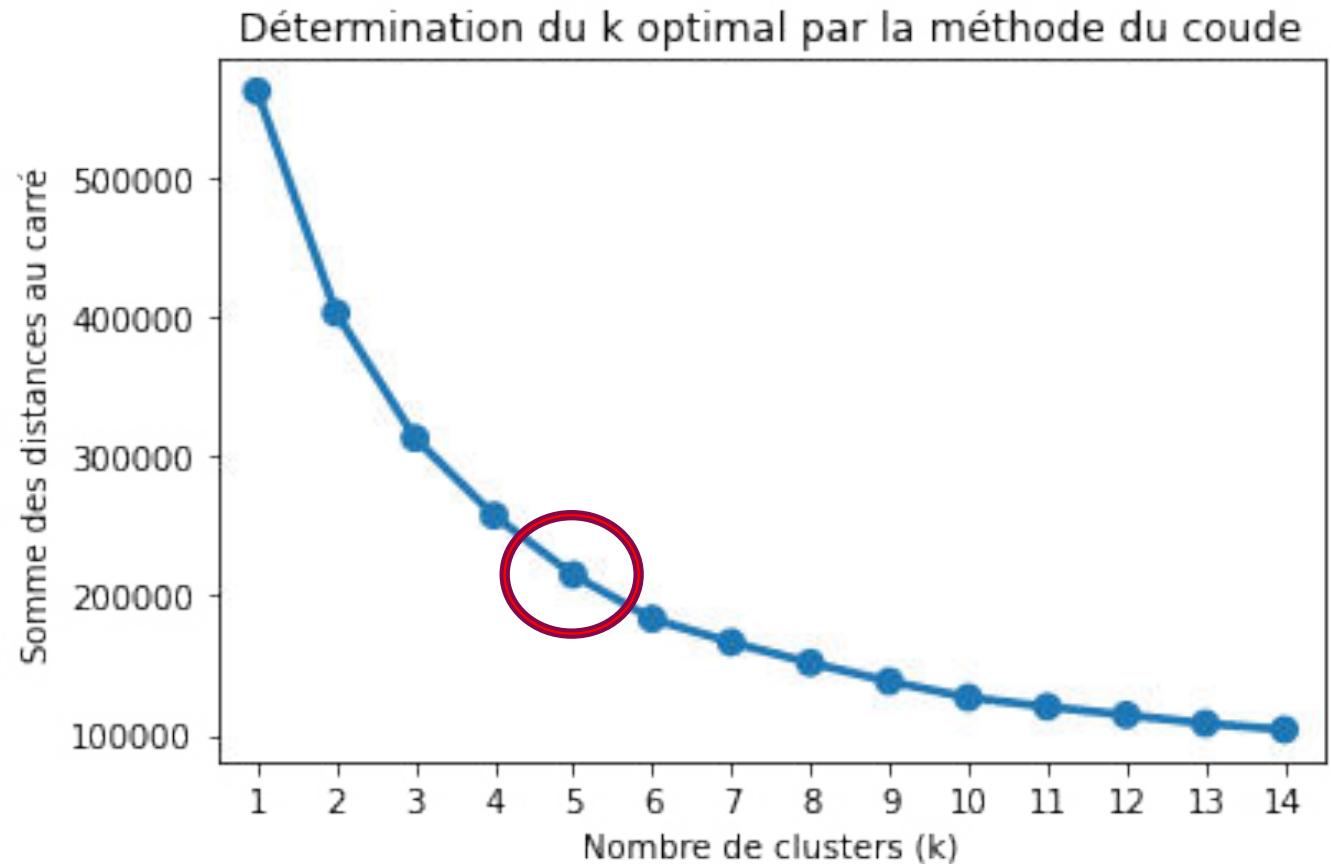


5. Clustering et Analyse Temporelle

5. Clustering

On choisit de découper notre panel de clients en 5 clusters

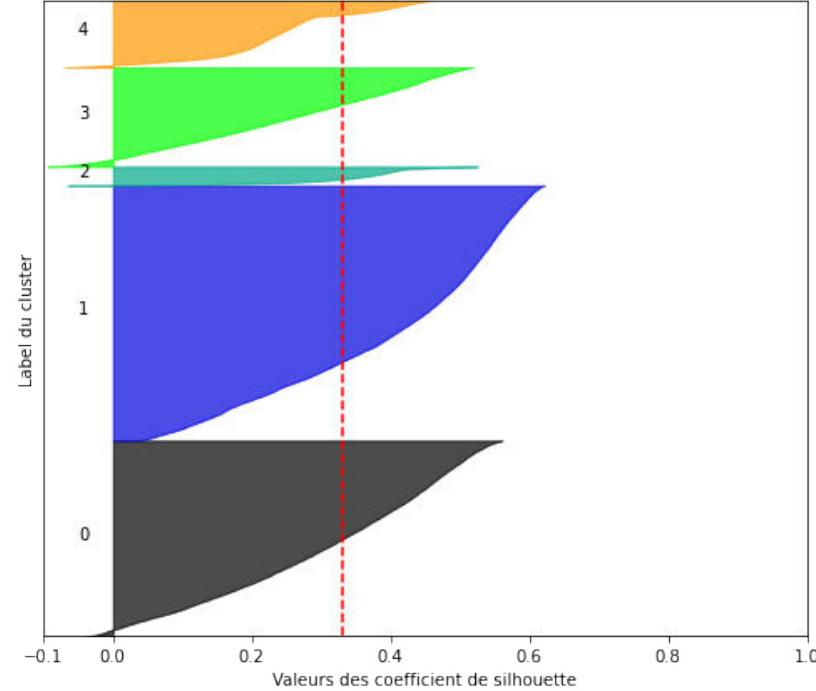
Pour n_clusters = 2 , le silhouette_score moyen est : 0.3073769887605267
Pour n_clusters = 3 , le silhouette_score moyen est : 0.3339132433257318
Pour n_clusters = 4 , le silhouette_score moyen est : 0.32576080192810963
Pour n_clusters = 5 , le silhouette_score moyen est : 0.33151151197621437
Pour n_clusters = 6 , le silhouette_score moyen est : 0.3175268470144766
Pour n_clusters = 7 , le silhouette_score moyen est : 0.321339628930385
Pour n_clusters = 8 , le silhouette_score moyen est : 0.32543679728595604



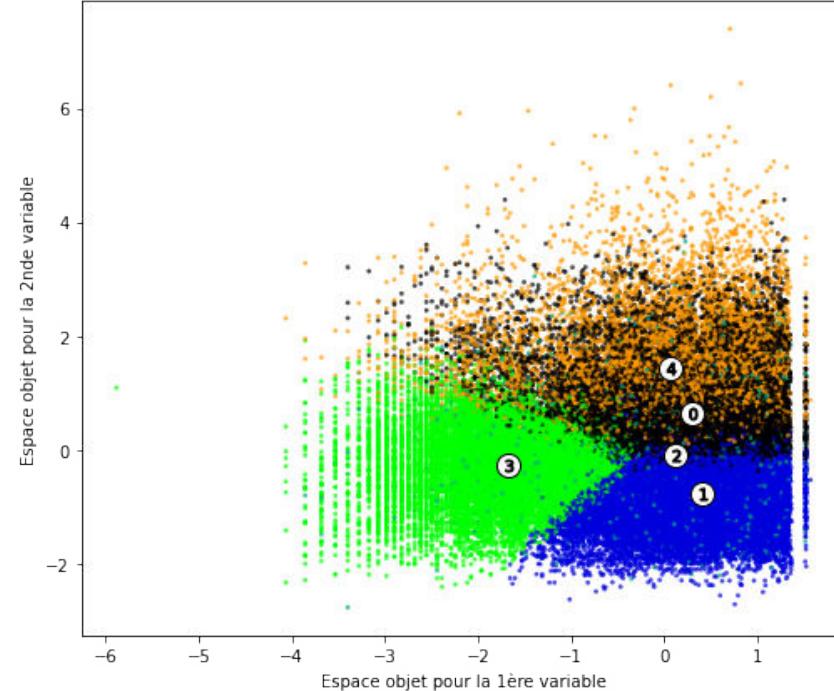
5. Clustering

Analyse de silhouette avec un clustering par la méthode des KMeans avec $n_clusters = 5$

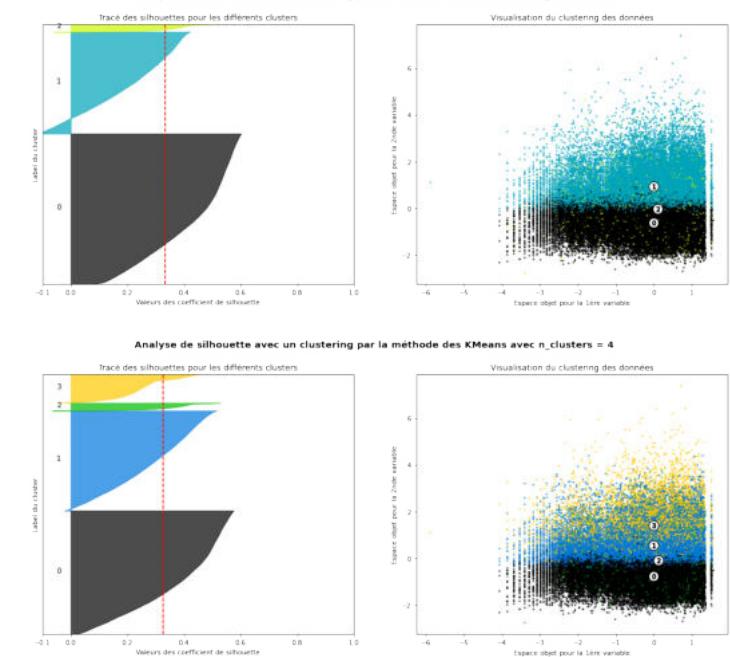
Tracé des silhouettes pour les différents clusters



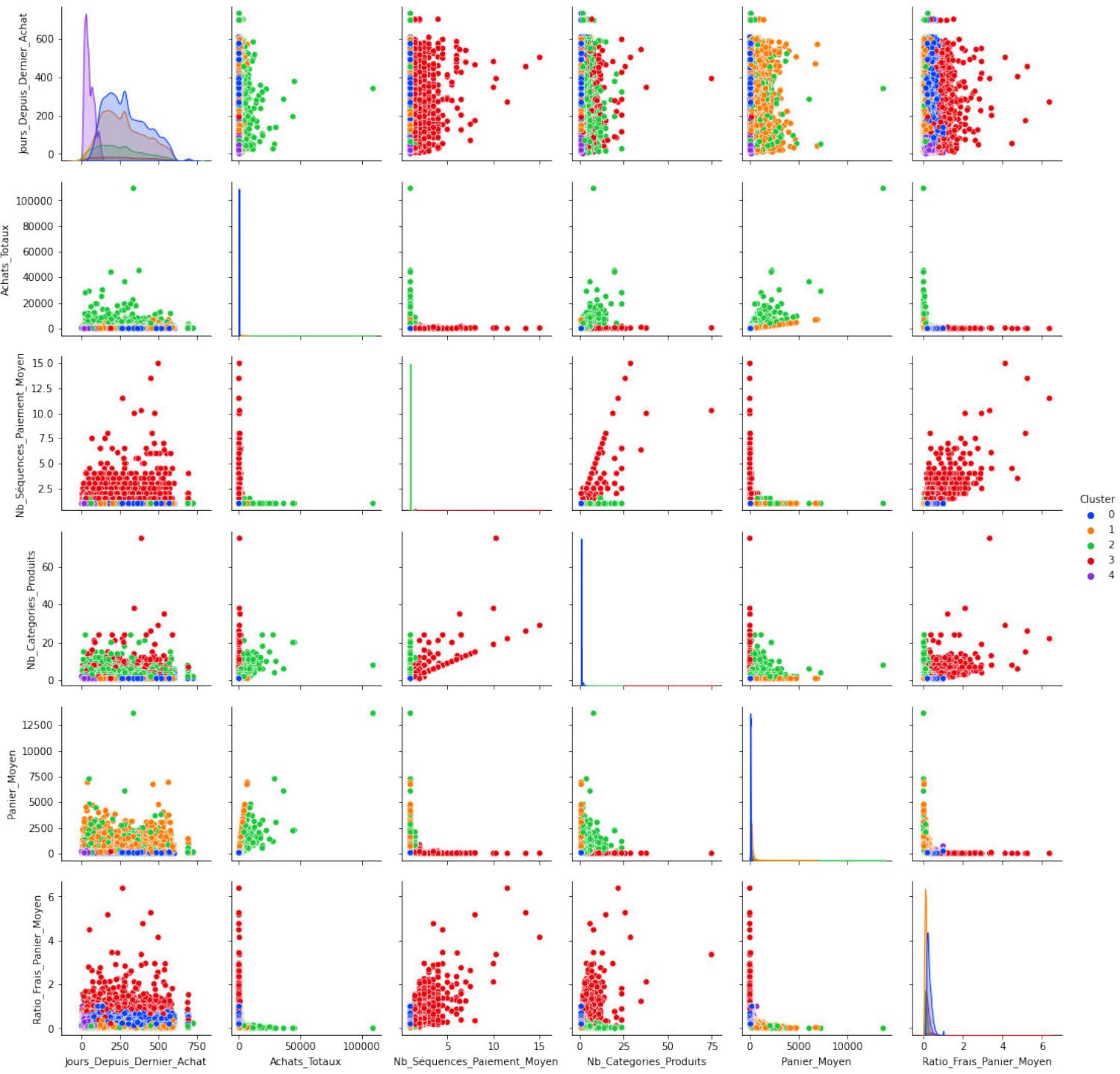
Visualisation du clustering des données



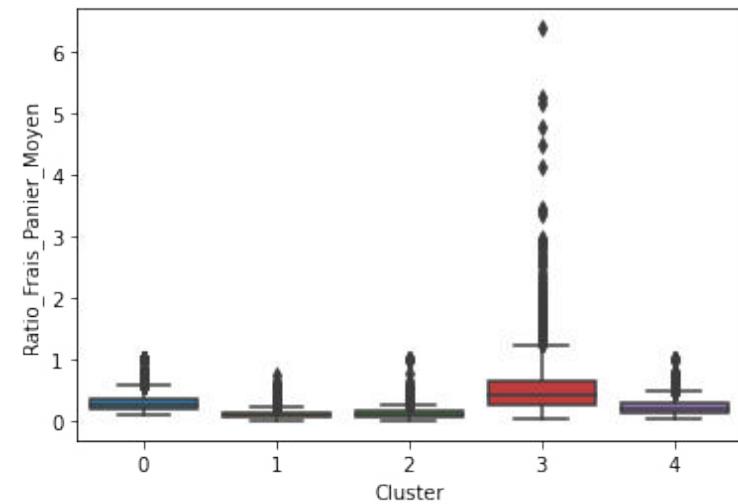
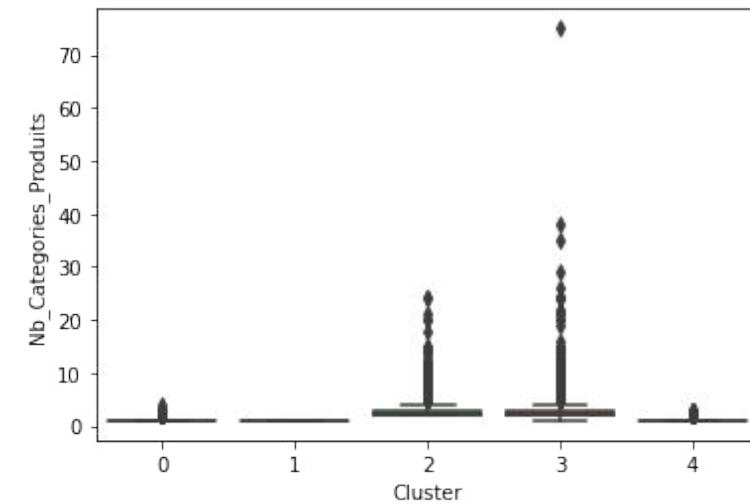
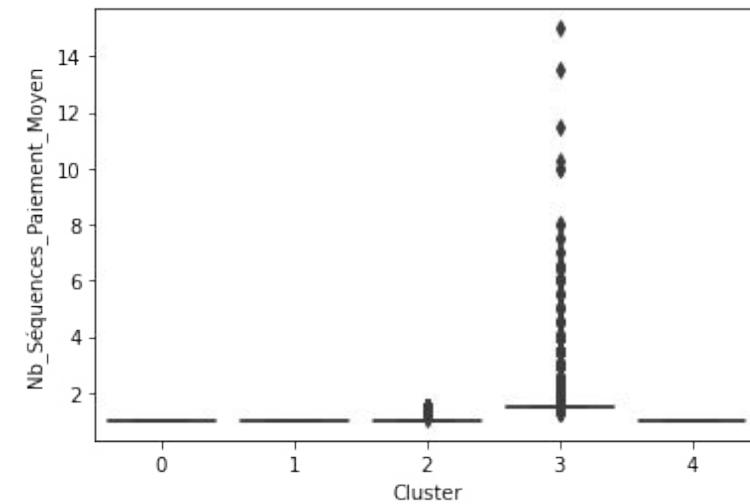
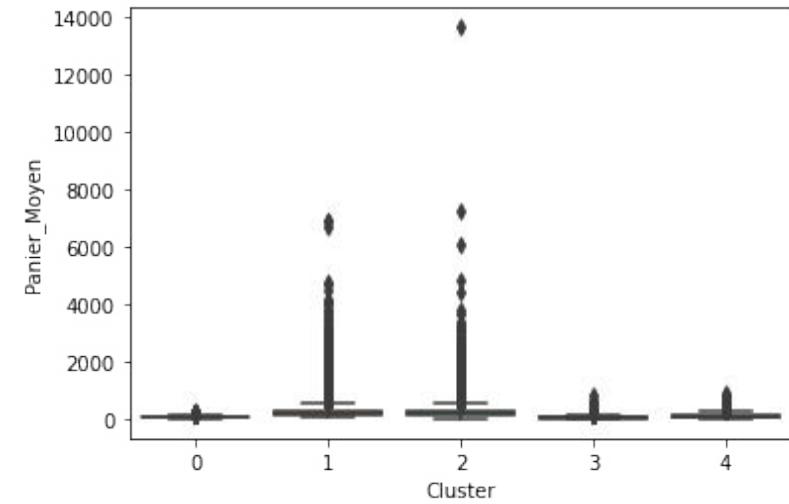
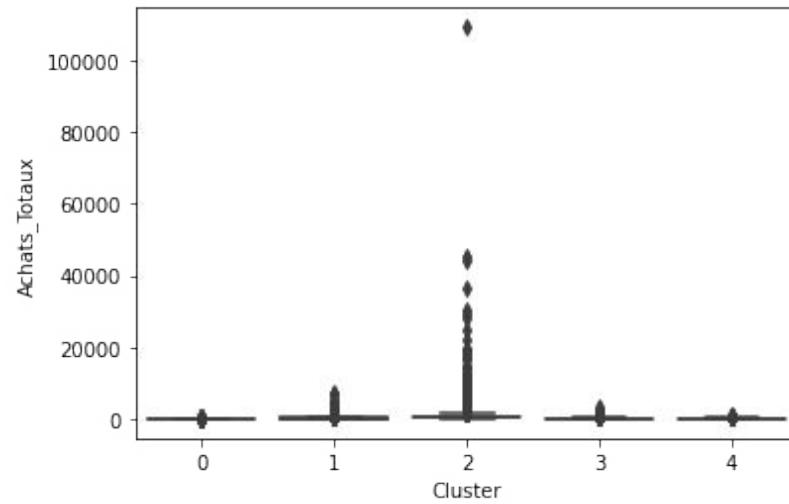
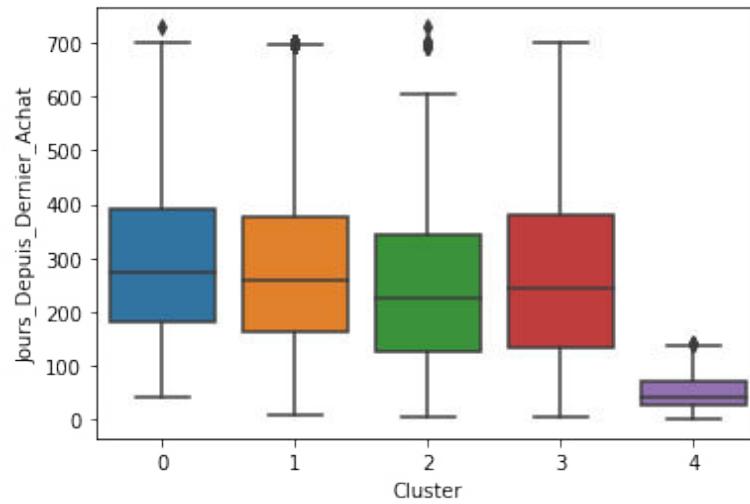
Analyse de silhouette avec un clustering par la méthode des KMeans avec $n_clusters = 3$



Observation de la distribution des clusters en fonction des variables



Distribution des variables en fonction des clusters



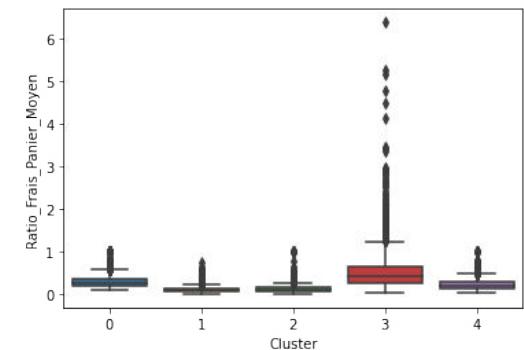
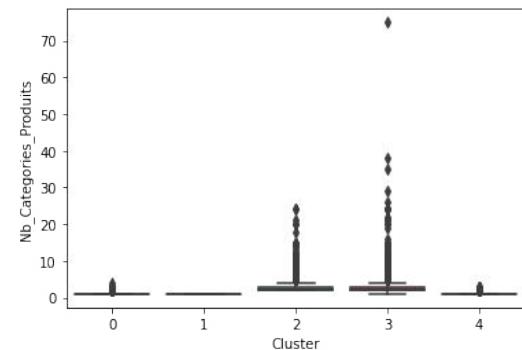
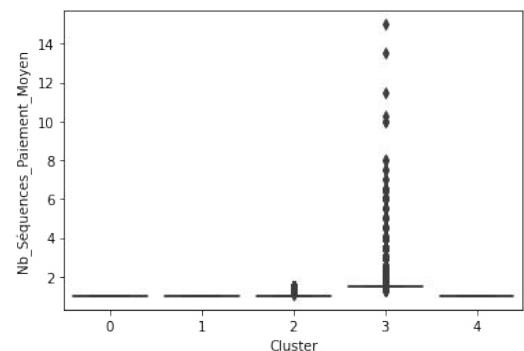
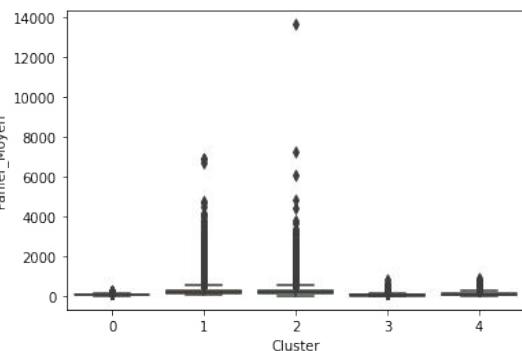
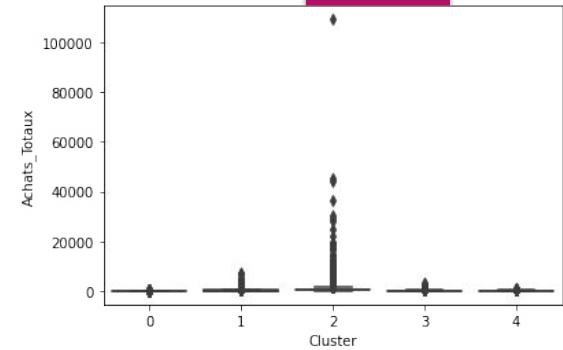
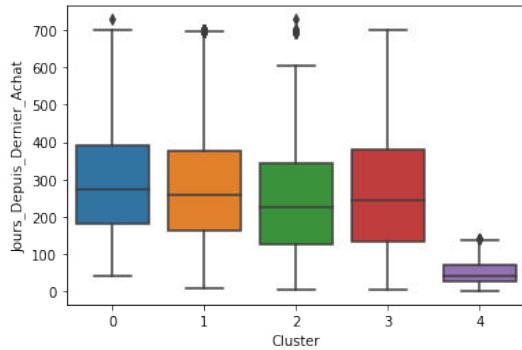
Cluster 0 : Ratio frais/panier un peu plus élevé : commande à faible valeur(avec PM et AT faibles)

Cluster 1 : Panier moyen important

Cluster 2 : Achat totaux et panier moyen très importants, et nombre élevé de catégories de produits

Cluster 3 : Nombre de catégorie de produits, nombre de séquences de paiement et ratio frais/panier plus élevés

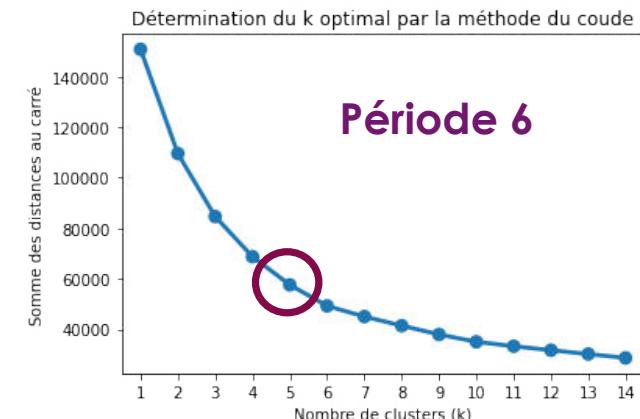
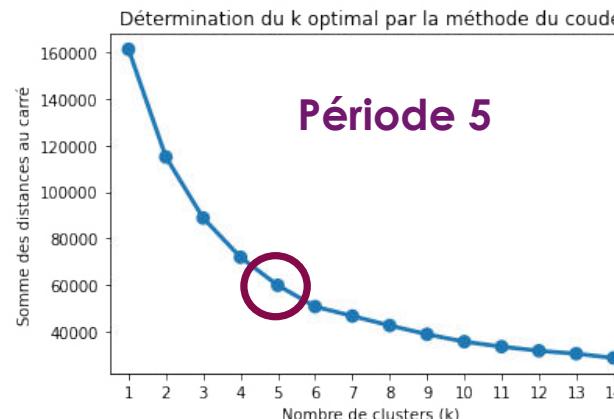
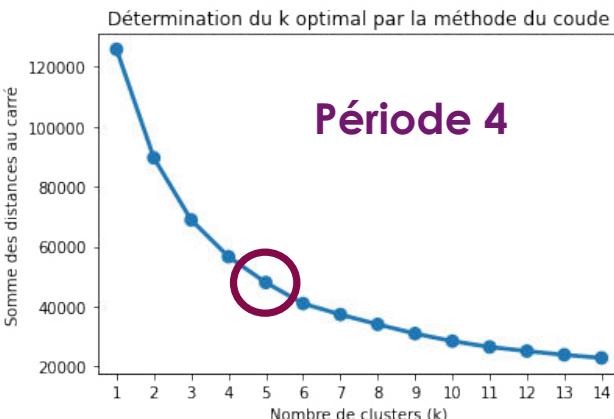
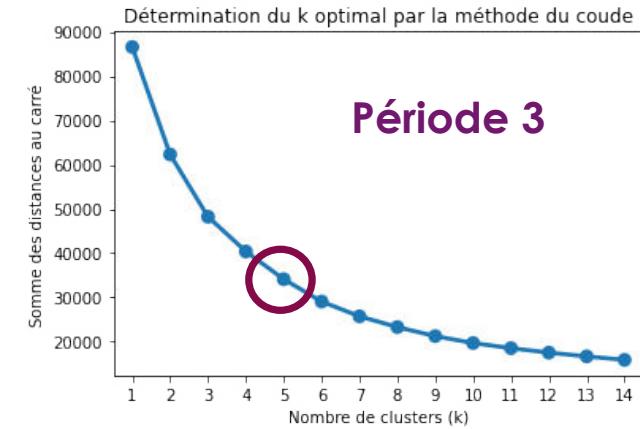
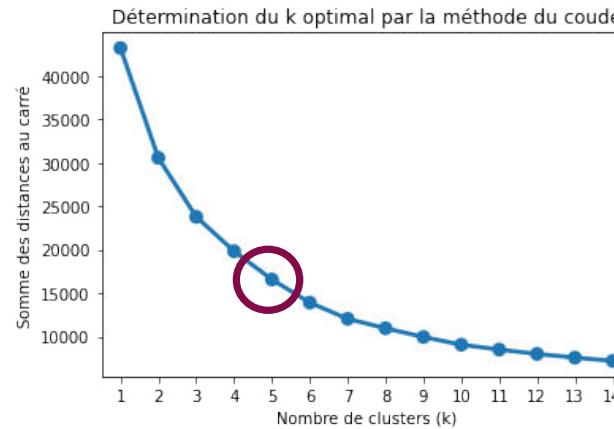
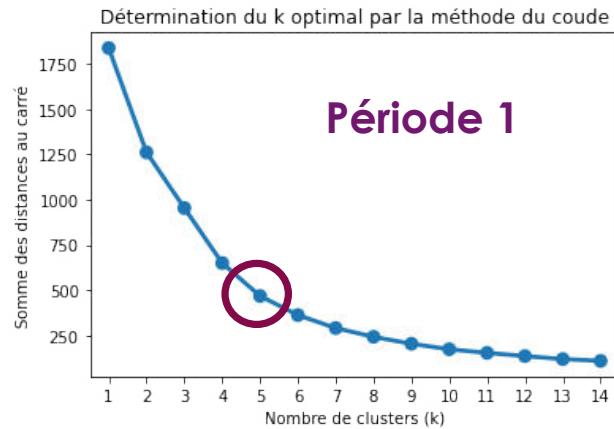
Cluster 4 : Dernier achat récent



5. Clustering et Analyse Temporelle - Périodes de 4 mois

Les courbes de détermination du nombre de clusters optimal sont similaires

Nombre de clusters optimal = 5



5. Clustering et Analyse Temporelle - Périodes de 4 mois

Analyse Temporelle :

- Observer l'évolution des clusters et de leurs caractéristiques sur les 6 périodes de 4 mois identifiées
- Déterminer la validité de ces clusters

Période 1

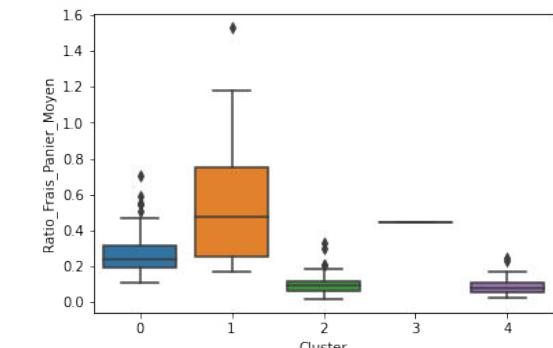
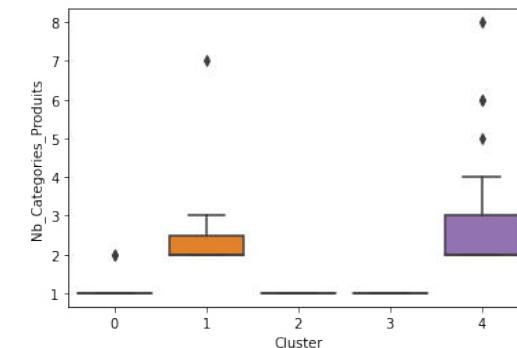
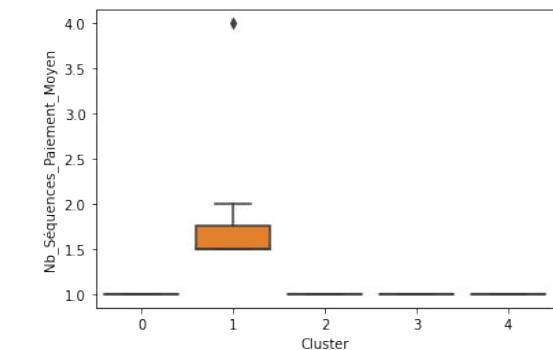
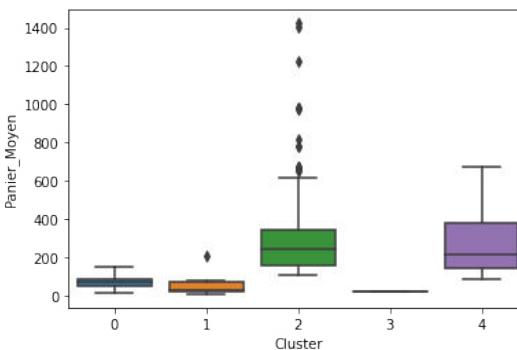
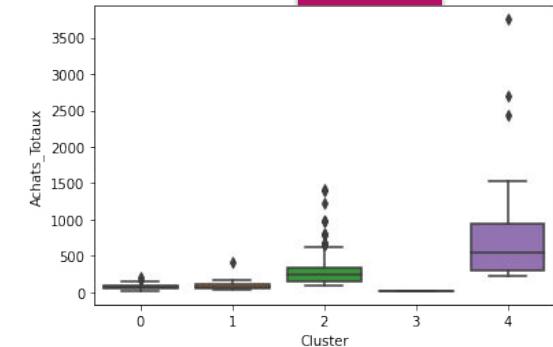
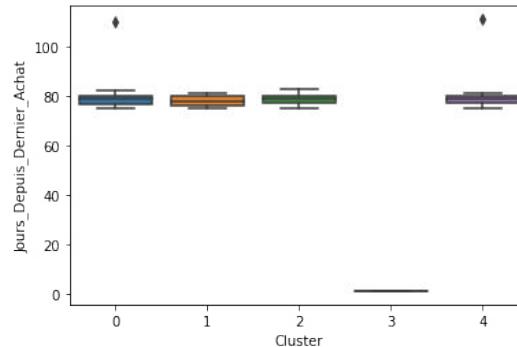
Cluster **0** Ratio frais/panier un peu plus élevé : commande à faible valeur (avec PM et AT faibles)

Cluster **2** Panier moyen important

Cluster **4** Achat totaux et panier moyen très importants, et nombre élevé de catégories de produits

Cluster **1** Nombre de catégorie de produits, nombre de séquences de paiement et ratio frais/panier plus élevés

Cluster **3** Dernier achat récent



Période 2

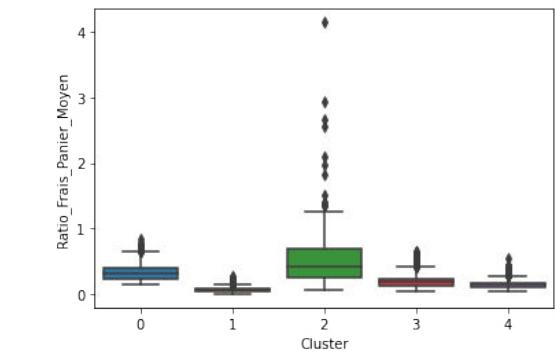
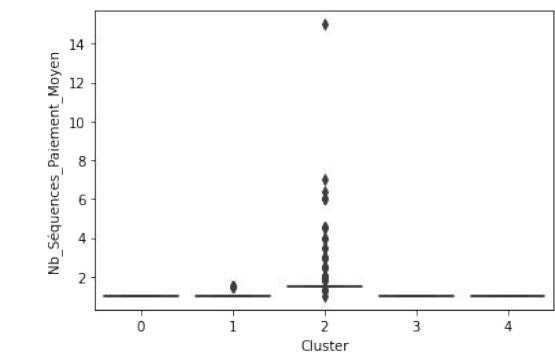
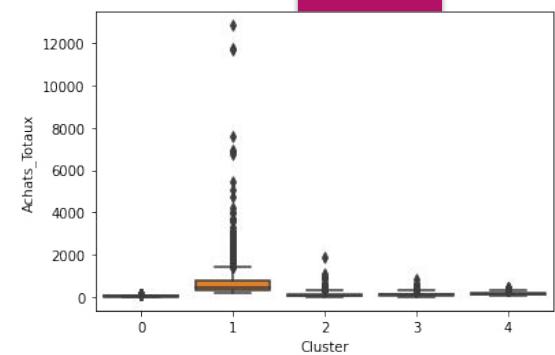
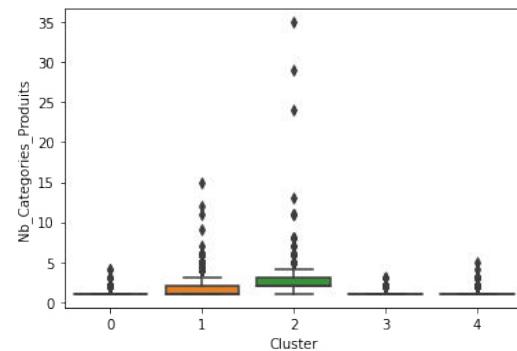
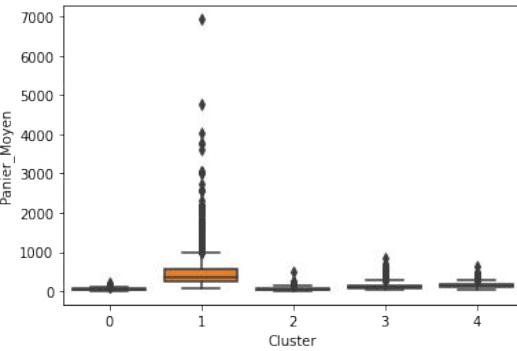
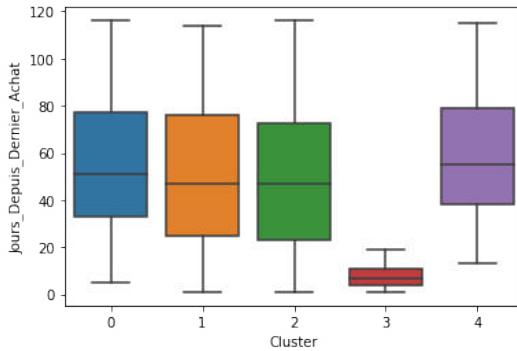
Cluster **0** Ratio frais/panier un peu plus élevé : commande à faible valeur (avec PM et AT faibles)

Cluster **4?** panier moyen important

Cluster **1** Achat totaux et panier moyen très importants, et nombre élevé de catégories de produits

Cluster **2** Nombre de catégorie de produits, nombre de séquences de paiement et ratio frais/panier plus élevés

Cluster **3** Dernier achat récent



Période 3

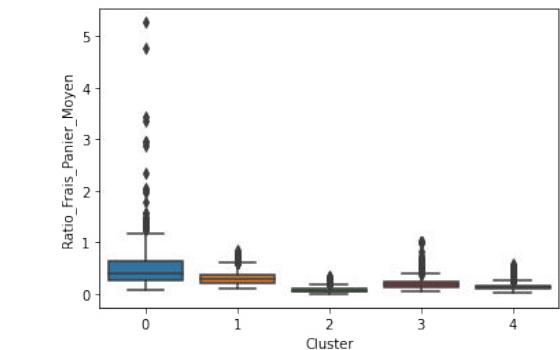
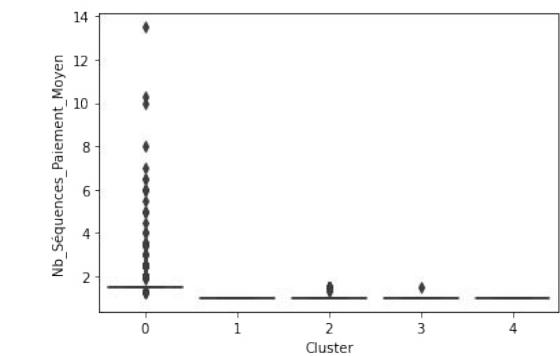
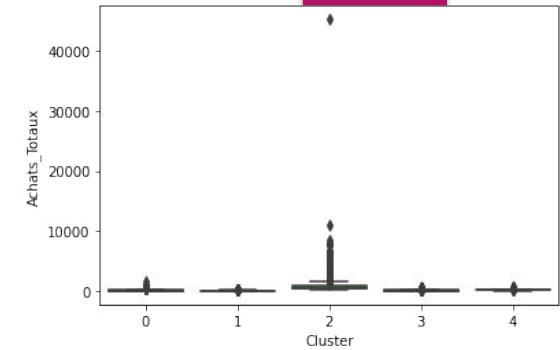
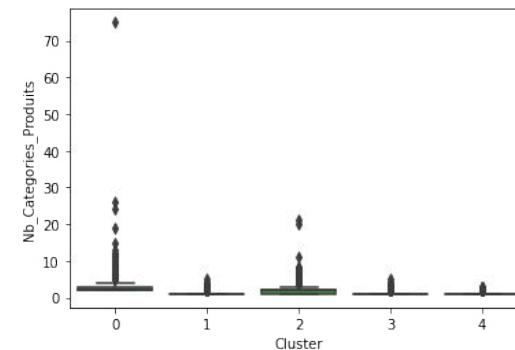
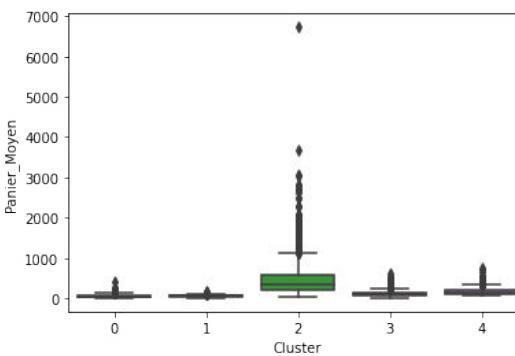
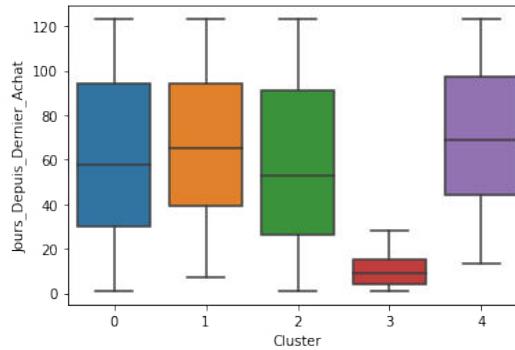
Cluster **1** Ratio frais/panier un peu plus élevé : commande à faible valeur(avec PM et AT faibles)

Cluster **4** Panier moyen important

Cluster **2** Achat totaux et panier moyen très importants, et nombre élevé de catégories de produits

Cluster **0** Nombre de catégorie de produits, nombre de séquences de paiement et ratio frais/panier plus élevés

Cluster **3** Dernier achat récent



Période 4

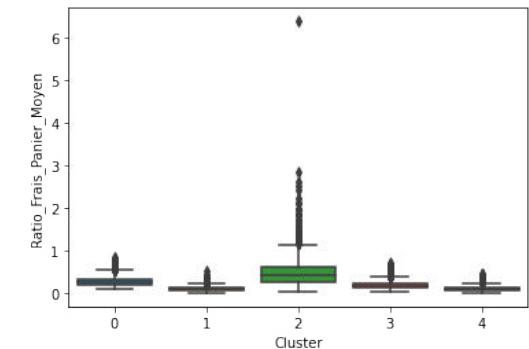
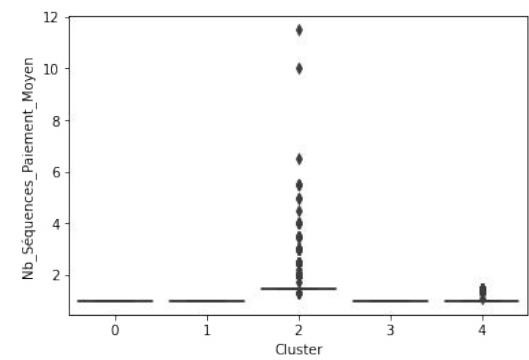
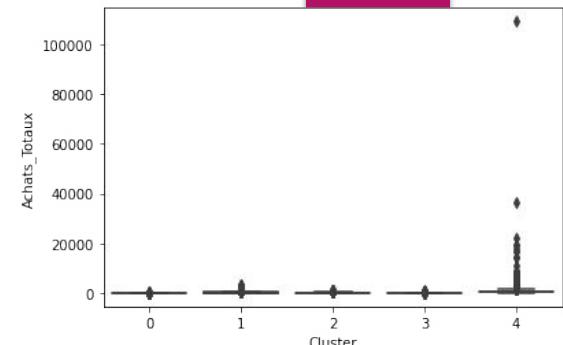
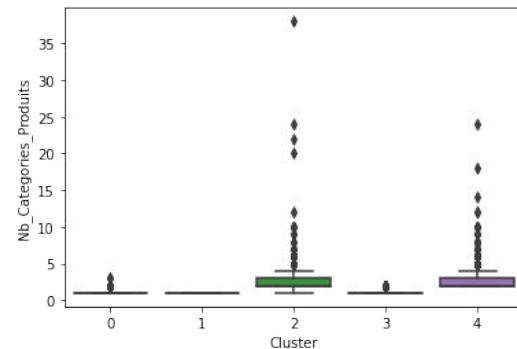
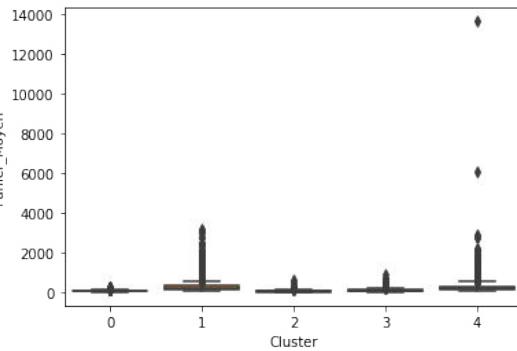
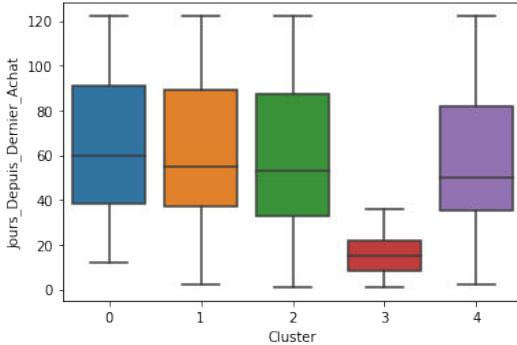
Cluster **0** Ratio frais/panier un peu plus élevé : commande à faible valeur(avec PM et AT faibles)

Cluster **1** Panier moyen important

Cluster **4** Achat totaux et panier moyen très importants, et nombre élevé de catégories de produits

Cluster **2** Nombre de catégorie de produits, nombre de séquences de paiement et ratio frais/panier plus élevés

Cluster **3** Dernier achat récent



Période 5

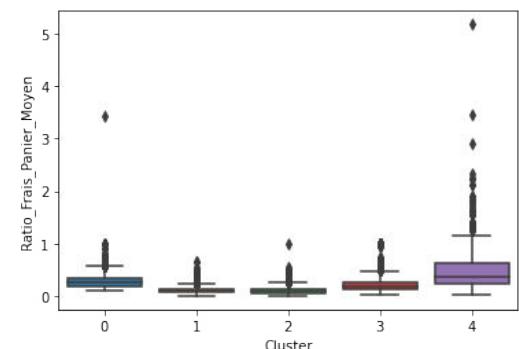
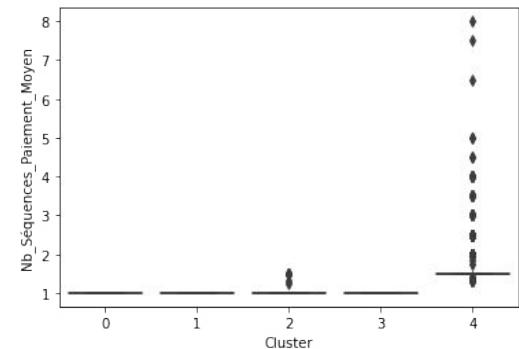
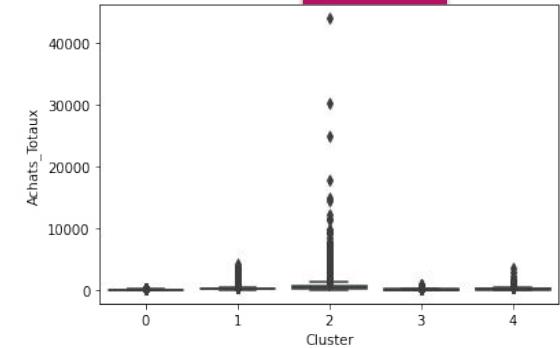
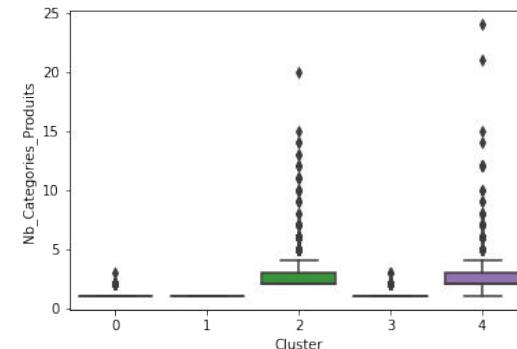
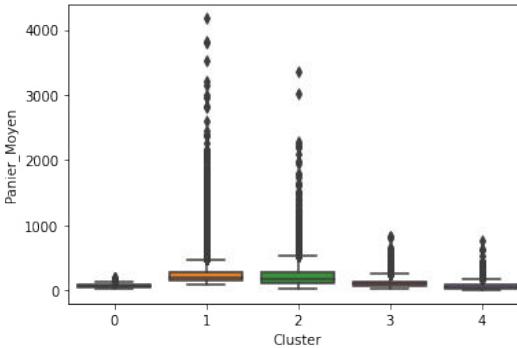
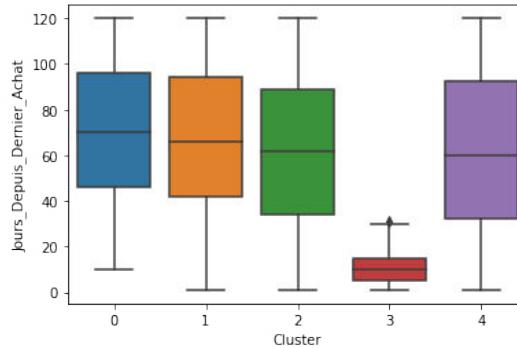
Cluster **0** Ratio frais/panier un peu plus élevé : commande à faible valeur(avec PM et AT faibles)

Cluster **1** Panier moyen important

Cluster **2** Achat totaux et panier moyen très importants, et nombre élevé de catégories de produits

Cluster **4** Nombre de catégorie de produits, nombre de séquences de paiement et ratio frais/panier plus élevés

Cluster **3** Dernier achat récent



Période 6

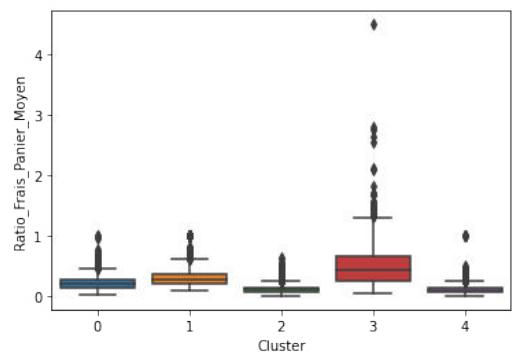
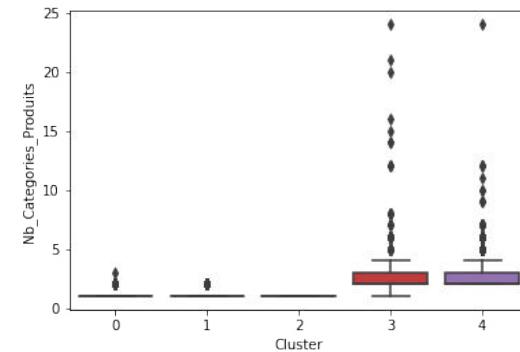
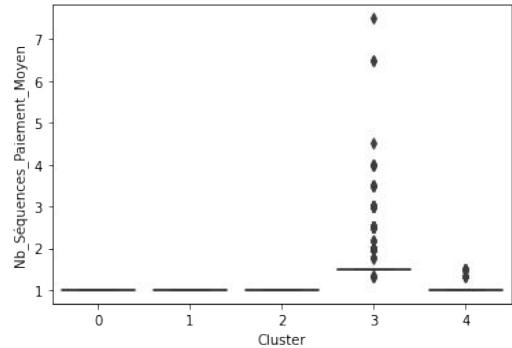
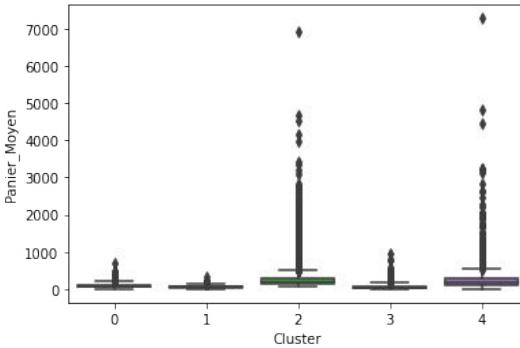
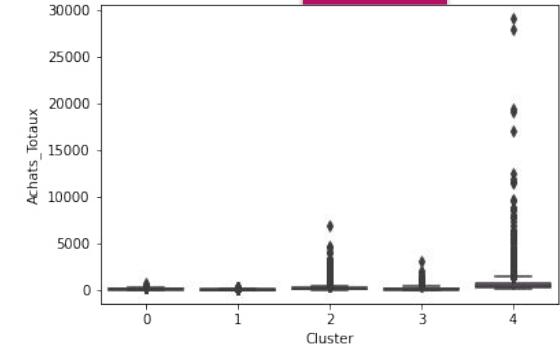
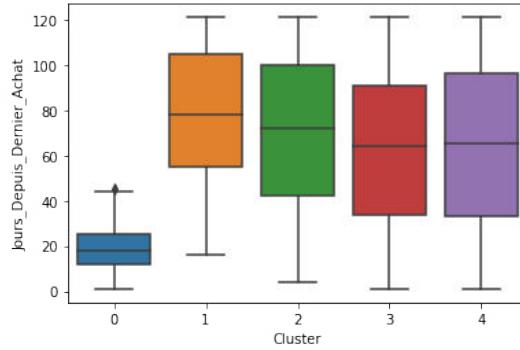
Cluster **1** Ratio frais/panier un peu plus élevé : commande à faible valeur(avec PM et AT faibles)

Cluster **2** Panier moyen important

Cluster **4** Achat totaux et panier moyen très importants, et nombre élevé de catégories de produits

Cluster **3** Nombre de catégorie de produits, nombre de séquences de paiement et ratio frais/panier plus élevés

Cluster **0** Dernier achat récent



5. Clustering et Analyse Temporelle - périodes de 4 mois

Les clusters semblent être stables



6. Conclusions et Perspectives

6. Conclusions et Perspectives

Conclusions :

- ❖ **5 clusters identifiés** avec des comportements caractéristiques en fonction de **6 variables d'intérêts** ;
- ❖ Les clusters sont **stables** sur la période étudiée ;
- ❖ Continuer des études sur des **périodes de 4 mois** semble approprié, tout en conservant les données déjà collectées.

6. Conclusions et Perspectives

Perspectives :

- ❖ Continuer de vérifier la stabilité des clusters ;
- ❖ Effectuer plus de feature engineering :
 - En fonction des catégories de produits,
 - En fonction de la ruralité des clients,
 - En fonction du paiement en crédit.



MERCI POUR VOTRE ATTENTION !