



Lapage

Librairie en ligne
Analyse des ventes

Exercices comptables 2021-2022 & 2022-2023

Sommaire

Notes sur la méthodologie Executive summary

1 - Téléchargement et vérification des données

1.1 - Import & vérification du typage des colonnes

1.2 - Étude de la qualité des données, duplicates, imputations, suppressions d'indicateurs

 1.2.1 - Identification, analyse & suppression des doublons

 1.2.1.1 - Lignes dupliquées

 1.2.1.2 - Autres informations dupliquées

 1.2.2 - Identification des valeurs manquantes

 1.2.3 - Recherche d'outliers

 1.2.3.1 - Table clients

 1.2.3.2 - Table books

 1.2.3.3 - Tables sales

Sommaire

1.2.4 - Vérification des clés primaires

1.2.5 - Ajout d'indicateurs

2 - Indicateurs de vente

2.1 - Analyse du chiffre d'affaires

2.2 - Offre produits

2.3 - Profils clients

3 - Comportements d'achat

3.1 - Genre et categories de livres achetées

3.2 - Influence de l'âge...

3.2.1 - ...sur le montant total des achats

3.2.2 - ...sur la fréquence d'achat

3.2.3 - ...sur la taille du panier moyen

3.2.4 - ...sur les categories de livres achetées

3.3 - Autres indicateurs

Sommaire

Conclusions & recommandations

Annexes

Annexe 1 - tests de normalité

Annexe 2 - produits non vendus

Annexe 3 - clients non acheteurs

Annexe 4 - chiffre d'affaires en moyenne glissante
par catégorie de produits

Notes sur la méthodologie

- ❖ Source des comparaisons avec la moyenne nationale :

Baromètre IPSOS / CNL "les Francais et la lecture 2021"
<https://centrenationaldulivre.fr/>

- ❖ Tests statistiques:

Non-paramétriques sauf mention contraire (2 exceptions), la normalité de la distribution des données n'étant pas vérifiée (cf. Annexe 1 sur les différents tests de normalité)

Executive summary

Des résultats solides malgré une conjoncture défavorable

+ 3.26%

sur le chiffre d'affaires

Moyenne nationale : -2.7%

+ 1.41%

sur le nombre d'exemplaires vendus

Moyenne nationale : -3.4%

-0.88%

sur le nombre d'acheteurs

Moyenne nationale : -14.0%

+ 21.03 €

sur les dépenses annuelles moyennes des clients particuliers

Moyenne nationale : +12 €

Un panier moyen en hausse

+1.9%

chez les clients particuliers

+0.48%

chez les clients professionnels

1 - Téléchargement et vérification des données

1.1 - Import & vérification du typage des colonnes

- ❖ Import des 3 fichiers customers.csv, products.csv & transactions.csv
- ❖ Vérification que le contenu des colonnes correspond bien à leur type de données
- ❖ Conversion des "objects" en "strings"
- ❖ Renommage des tables en clients, books et sales respectivement
- ❖ Renommage des colonnes pour faire référence à leur contenu et leur table de provenance selon le modèle contenu_colonne_nomTable (ex: la colonne id_prod de la table books devient product_id_books)
- ❖ La conversion de la colonne date de la table transactions d' "object" à "datetime" retourne un message d'erreur <Unknown string format> => corrigé après élimination des doublons (cf. page suivante)

1 - Téléchargement et vérification des données

1.2 - Étude de la qualité des données, duplicates, imputations, suppressions d'indicateurs

1.2.1 - Identification, analyse & suppression des doublons

1.2.1.1 - Lignes dupliquées

- ❖ La recherche de doublons ne retourne aucun résultat dans les tables clients et books
- ❖ Elle retourne 126 lignes dans la table sales, toutes caractérisées par la présence du produit T_0, une référence de session s_0, des clients ct_0 et ct_1 et une date de forme string commençant par "test_"
 - On élimine donc ces 126 lignes de données de test du fichier sales (renommée sales_trim) et on élimine le produit T_0 (dont le prix est par ailleurs négatif) de la table books et les clients ct_0 et ct_1 de la table clients
 - Suite à cette suppression dans la table sales_trim, le casting de la colonne date_sales en format datetime devient possible (plus de message d'erreur) et la table ne contient plus de doublons

1 - Téléchargement et vérification des données

1.2 - Étude de la qualité des données, duplicates, imputations, suppressions d'indicateurs

1.2.1 - Identification, analyse & suppression des doublons

1.2.1.2 - Autres informations dupliquées

- ❖ La table books contient une colonne catégorie qui est en fait identique au premier caractère de l'identifiant du produit
- ❖ Cette colonne a été conservée pour notre analyse, mais il faudrait en pratique reformatter les références produits dans les données stockées car le format actuel ne respecte pas la 1NF (référence produit non atomique et information dupliquée entre le produit et la catégorie à laquelle il appartient) par exemple en utilisant le numéro ISBN en lieu et place de la référence produit actuelle

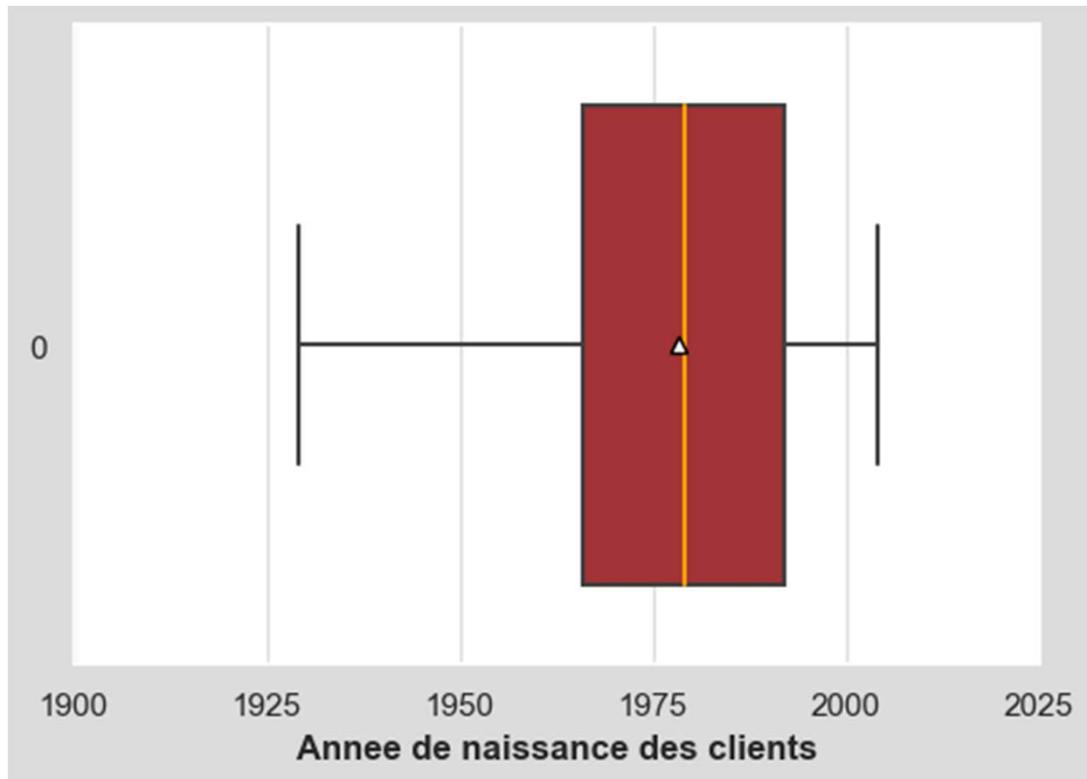
1.2 - Étude de la qualité des données, duplicates, imputations, suppressions d'indicateurs

1.2.2 - Identification des valeurs manquantes

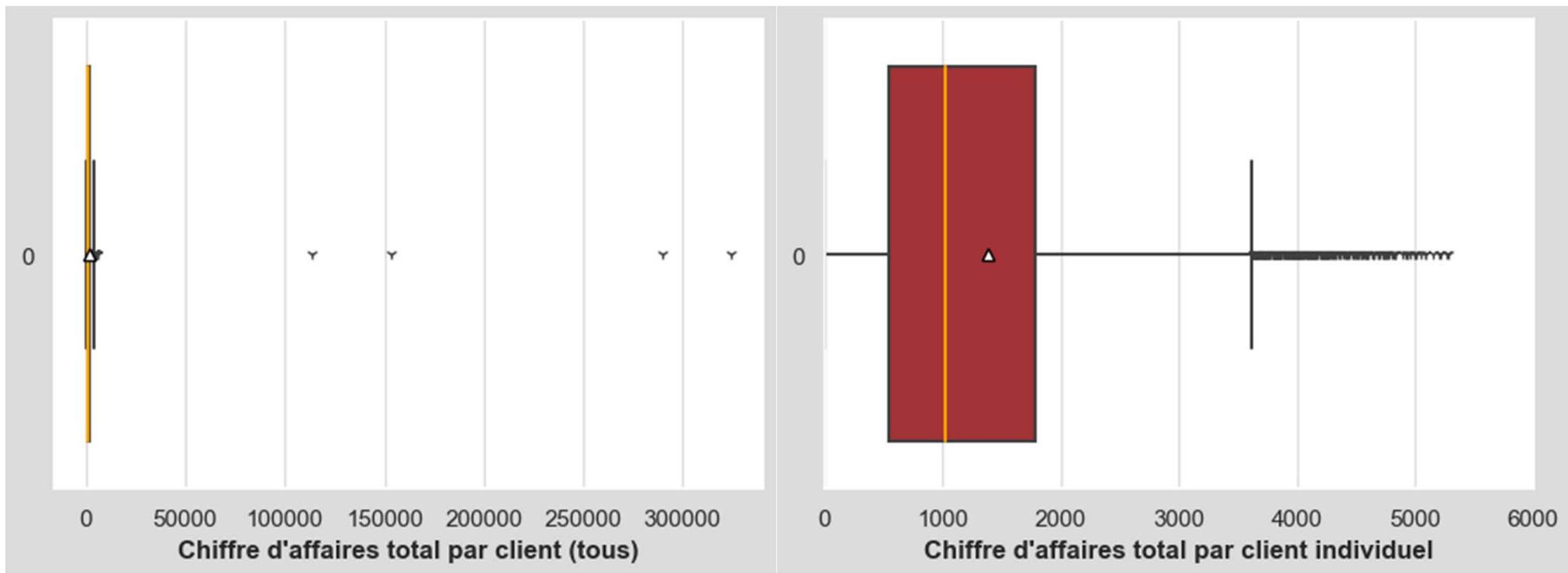
- ❖ L'utilisation de la librairie python missingno ne révèle aucune valeur manquante dans les données
- ❖ Le croisement de la table clients avec la table sales_trim fait apparaître 21 clients n'ayant jamais acheté en ligne (cf. liste en Annexe 3)
- ❖ Le croisement de la table sales_trim avec la table books fait apparaître 21 produits jamais vendus en ligne (cf. liste en Annexe 2), ainsi que 221 transactions sur le produit 0_2245 qui est absent du référentiel books
 - Ce produit est ajouté à la table books, on lui impute la catégorie indiquée par le premier caractère de sa référence et le prix médian des produits de cette catégorie
- ❖ La recherche d'outliers dans la table sales_trim fera en outre apparaître des données manquantes en octobre 2021 (cf. 1.2.3 ci-dessous)

1.2 – Étude de la qualité des données, duplicitas, imputations, suppressions d'indicateurs

1.2.3 – Recherche d'outliers – table clients



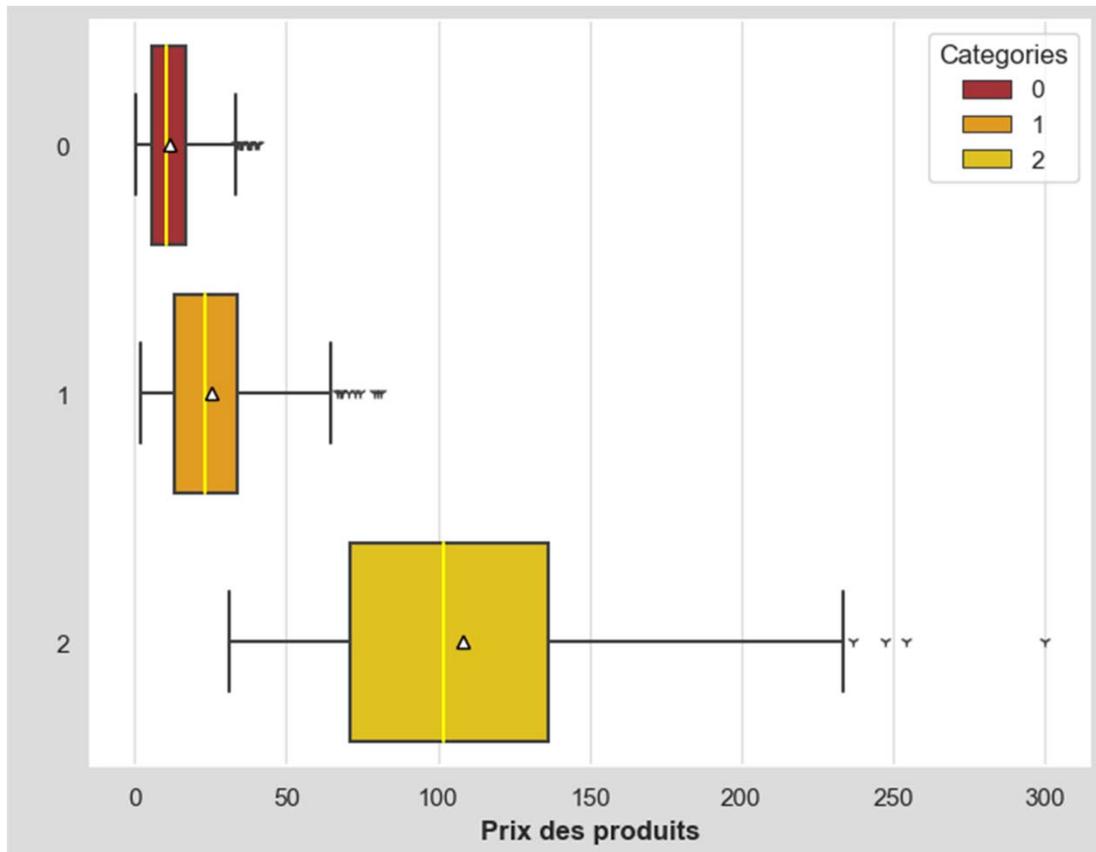
- ❖ Les années de naissance des clients s' étalent de 1929 (94 ans) à 2004 (19 ans)
- ❖ Age moyen 45 ans
- ❖ Pas de valeurs atypiques ou aberrantes détectées



- ❖ 4 clients ont un chiffre d'affaires individuel supérieur à la moustache haute

1.2 - Étude de la qualité des données, duplicates, imputations, suppressions d'indicateurs

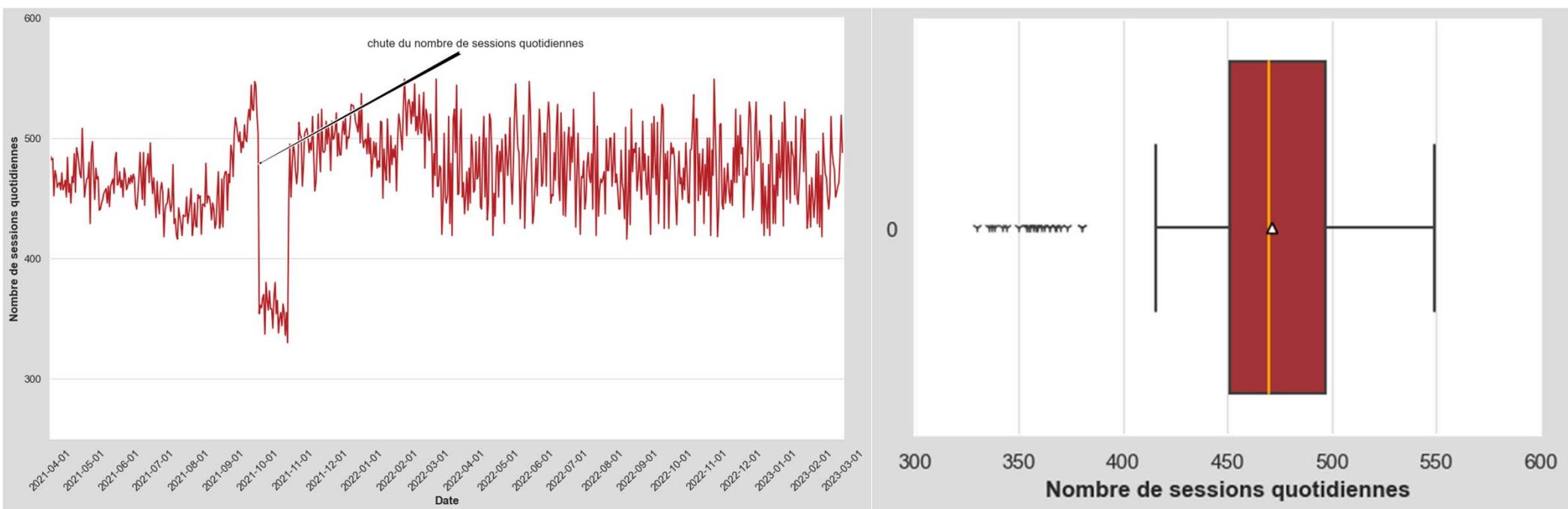
1.2.3 - Recherche d'outliers - table books



- ❖ 22 livres de catégorie 0, 10 de catégorie 1 et 4 de catégorie 2 ont des prix supérieurs à la moustache haute de leurs catégories respectives dans la table books
- ❖ En l'absence d'informations détaillées sur ces produits, il n'est pas possible de déterminer si leurs prix sont aberrants ou juste atypiques, ils ont donc tous été conservés dans la table books pour l'analyse
- ❖ 3 gammes de prix différentes pour ces 3 catégories de produits

1.2 - Étude de la qualité des données, duplicates, imputations, suppressions d'indicateurs

1.2.3 - Recherche d'outliers - table sales



- ❖ 25 jours dans la table sales_trim ont un nombre de sessions quotidiennes inférieur à la moustache basse

1.2 - Étude de la qualité des données, duplicates, imputations, suppressions d'indicateurs

1.2.3 - Recherche d'outliers - table sales

date_only_sales	session_id_sales
02/10/2021	689
03/10/2021	679
04/10/2021	641
05/10/2021	632
07/10/2021	623
08/10/2021	713
09/10/2021	675
10/10/2021	629
11/10/2021	684
12/10/2021	663
13/10/2021	669
14/10/2021	646
15/10/2021	672
16/10/2021	706
17/10/2021	669
18/10/2021	645
19/10/2021	604
20/10/2021	597
21/10/2021	653
22/10/2021	607
23/10/2021	602
24/10/2021	628
25/10/2021	577
26/10/2021	628
27/10/2021	579

- ❖ Aucune vente de catégorie 1 n'a été enregistrée pendant 25 jours du mois d'octobre 2021
- ❖ En l'absence d'un historique suffisamment long, aucune imputation n'a été effectuée pour combler ce manque de données
 - En pratique, ici encore il conviendrait de vérifier en amont l'intégrité des données fournies par le site de vente en ligne

date_only_sales	category_sales	session_id_sales	date_only_sales	category_sales	session_id_sales
02/10/2021	0	661	16/10/2021	0	661
02/10/2021	2	28	16/10/2021	2	45
03/10/2021	0	648	17/10/2021	0	625
03/10/2021	2	31	17/10/2021	2	44
04/10/2021	0	603	18/10/2021	0	608
04/10/2021	2	38	18/10/2021	2	37
05/10/2021	0	594	19/10/2021	0	567
05/10/2021	2	38	19/10/2021	2	37
07/10/2021	0	597	20/10/2021	0	555
07/10/2021	2	26	20/10/2021	2	42
08/10/2021	0	669	21/10/2021	0	610
08/10/2021	2	44	21/10/2021	2	43
09/10/2021	0	640	22/10/2021	0	572
09/10/2021	2	35	22/10/2021	2	35
10/10/2021	0	600	23/10/2021	0	555
10/10/2021	2	29	23/10/2021	2	47
11/10/2021	0	642	24/10/2021	0	584
11/10/2021	2	42	24/10/2021	2	44
12/10/2021	0	633	25/10/2021	0	545
12/10/2021	2	30	25/10/2021	2	32
13/10/2021	0	633	26/10/2021	0	592
13/10/2021	2	36	26/10/2021	2	36
14/10/2021	0	606	27/10/2021	0	530
14/10/2021	2	40	27/10/2021	2	49
15/10/2021	0	634			
15/10/2021	2	38			

1.2 – Étude de la qualité des données, duplicates, imputations, suppressions d'indicateurs

1.2.4 – Vérification des clés primaires

- ❖ L'identifiant client (client_id_clients) est une clé primaire pour la table clients
- ❖ La référence produit (product_id_books) est une clé primaire pour la table books
- ❖ Aucune des colonnes de la table sales ne peut être considérée individuellement comme une clé candidate, sauf date_sales qui est l'horodatage de la transaction à la nanoseconde près (ce qui a peu d'utilité pour notre analyse)
 - Les colonnes product_id_sales et client_id_sales sont des clés étrangères pour la table sales_trim vers les tables books et clients respectivement

1.2 - Étude de la qualité des données, duplicates, imputations, suppressions d'indicateurs

1.2.5 - Ajout d'indicateurs

- ❖ Table clients:

- Ajout d'une colonne d'âge
 - Ajout d'une colonne de classe d'âge : 15-24 / 25-34 / 35-49 / 50-64 / 65+

- ❖ Table sales:

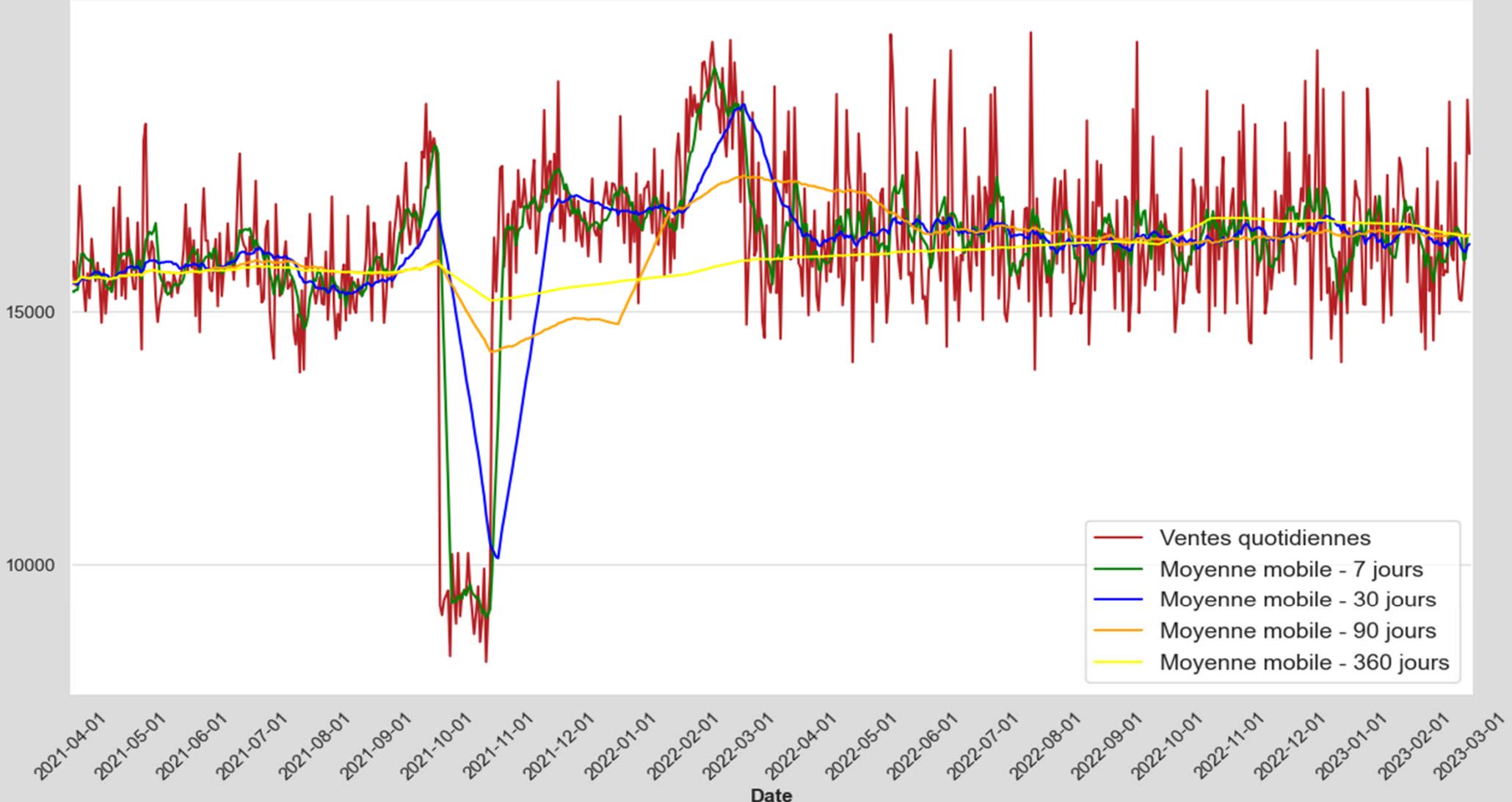
- Ajout de colonnes pour le jour, le mois, le trimestre et l'année calendaire des ventes
 - Ajout d'une colonne identifiant le premier exercice comptable du second (année financière)

2 - Indicateurs de vente

2.1 - Analyse du chiffre d'affaires

Chiffre d'affaires total quotidien

Chiffre d'affaires en €



- ❖ CA réalisé sur 2 exercices: 11,856,009.40 €
 - Année 1 : 5,832,800.01 €
dont clients professionnels : 435,733.85 €
dont clients particuliers : 5,397,066.16 €
 - Année 2 : 6,023,209.39 €
dont clients professionnels : 445,389.57 €
dont clients particuliers : 5,577,819.82 €
- ❖ Nombre de ventes réalisées sur 2 exercices: 679,332
 - Année 1 : 337,288 ventes
dont clients professionnels : 23,189 ventes
dont clients particuliers : 314,099 ventes
 - Année 2 : 342,044 ventes
dont clients professionnels : 23,454 ventes
dont clients particuliers : 318,590 ventes
- ❖ Evolution totale du chiffre d'affaires : + 3.26 % en valeur
 - dont clients particuliers + 3.35 %
 - dont clients professionnels + 2.22 %
- ❖ Evolution totale des ventes : + 1.41 % en volume
 - dont clients particuliers + 1.42 %
 - dont clients professionnels + 1.14 %

❖ Répartition des ventes par catégorie

- Clients professionnels

Année 1

category_books	product_id_sales
0	14271
1	7110
2	1808

Année 2

category_books	product_id_sales
0	13916
1	7711
2	1827

➤ Les clients professionnels comme particuliers achètent majoritairement des produits de catégorie 0

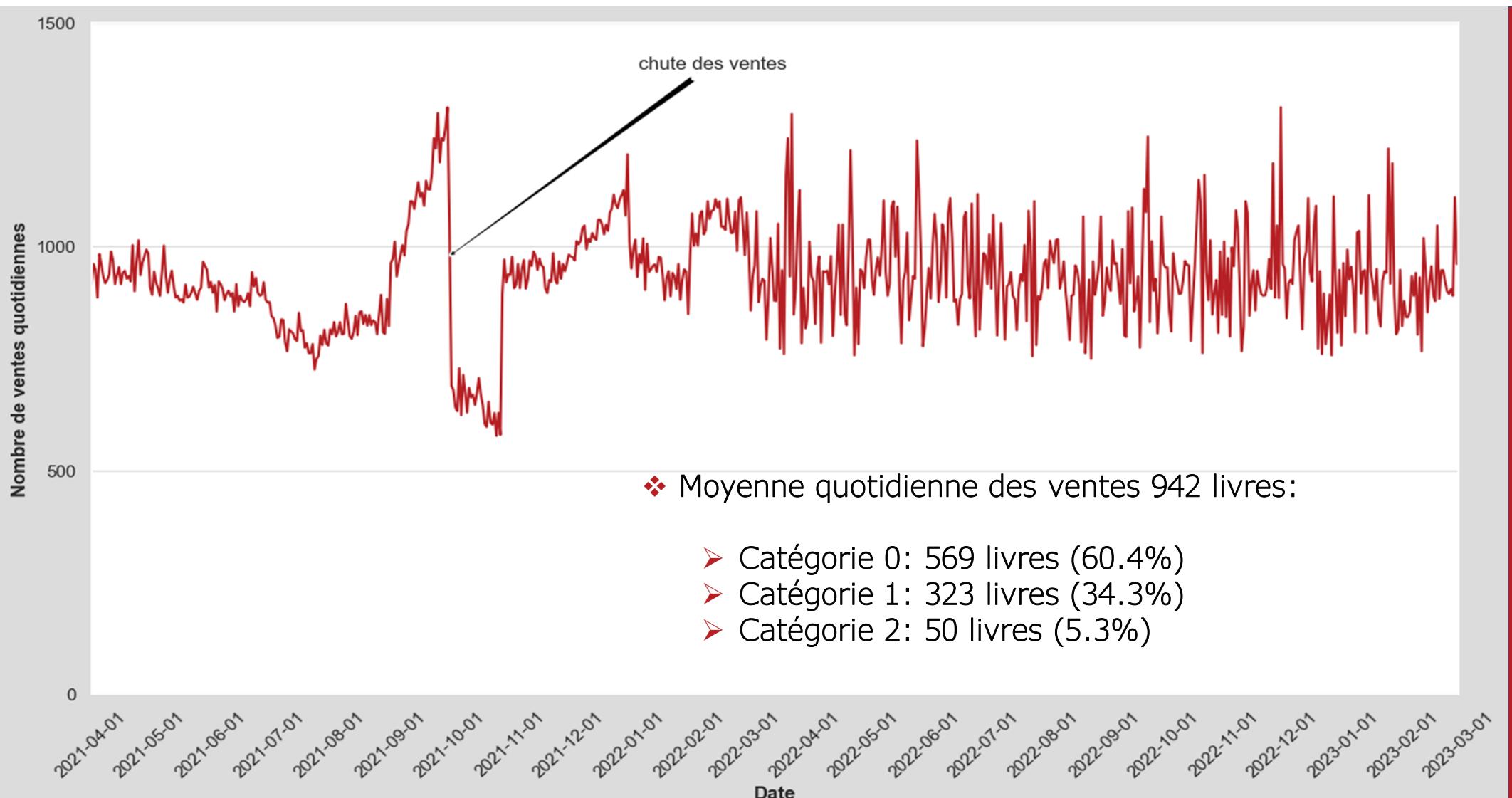
- Clients particuliers

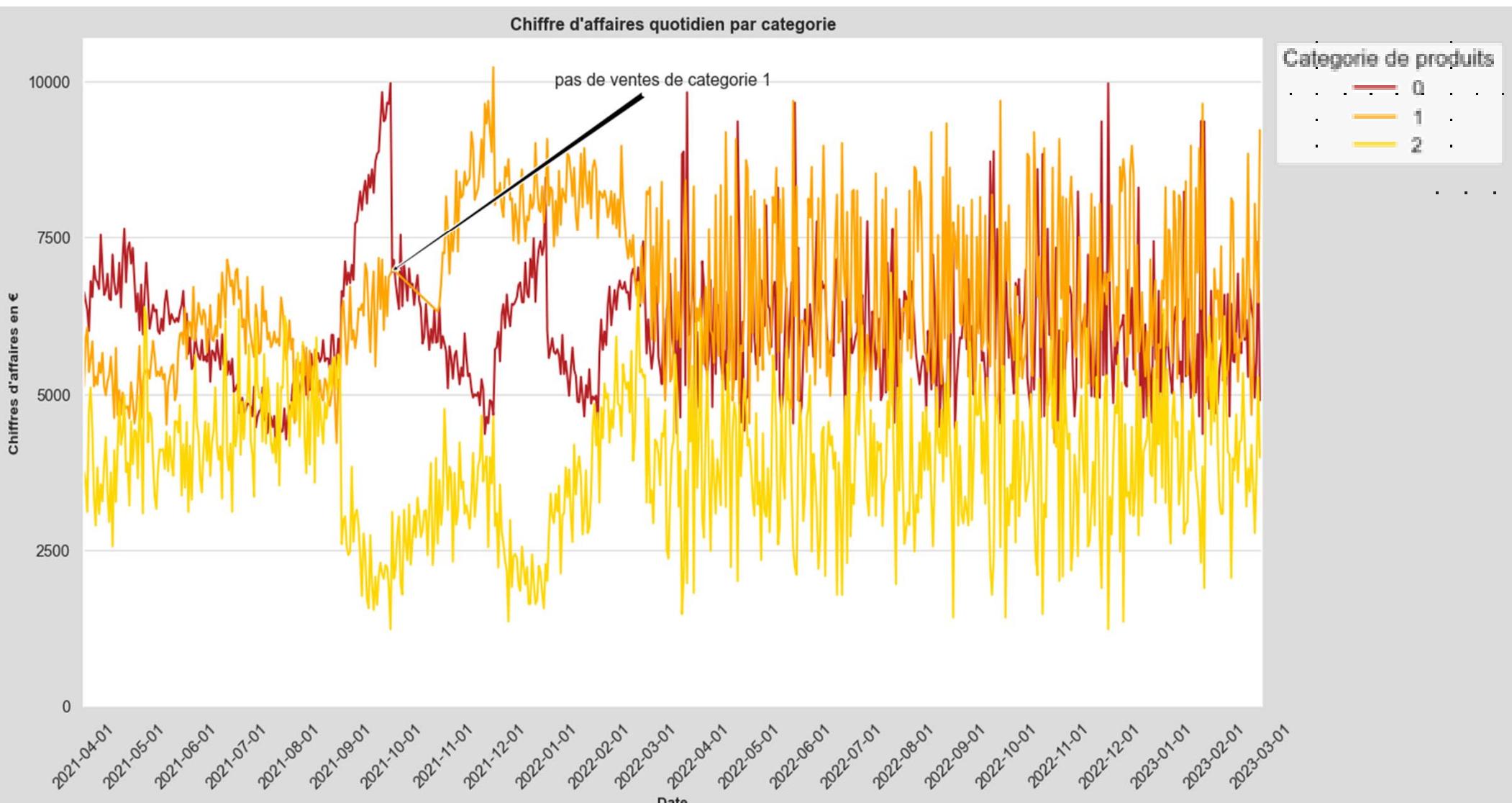
Année 1

category_books	product_id_sales
0	195494
1	102625
2	15980

Année 2

category_books	product_id_sales
0	191999
1	109723
2	16868

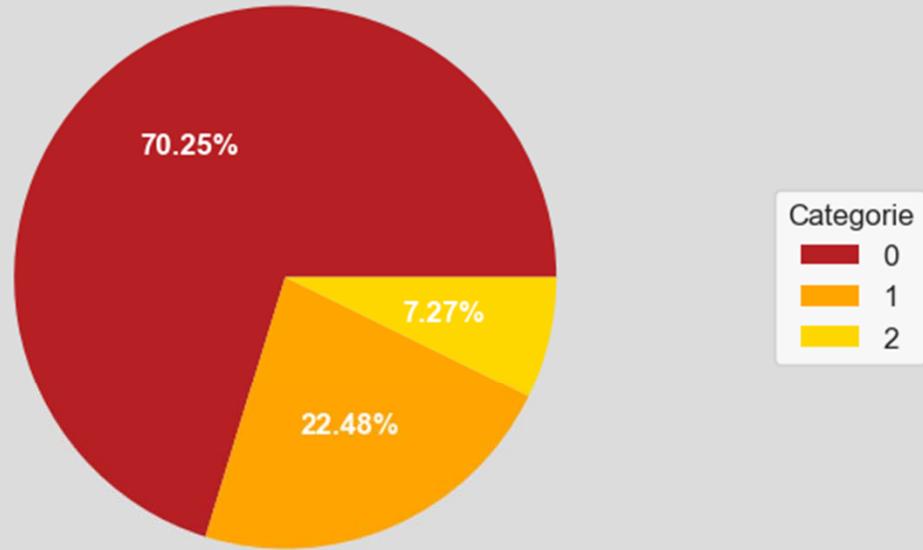




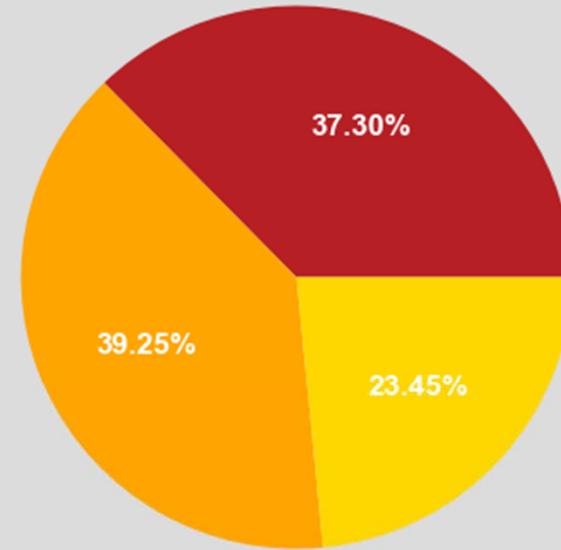
2 - Indicateurs de vente

2.2 - Offre produits

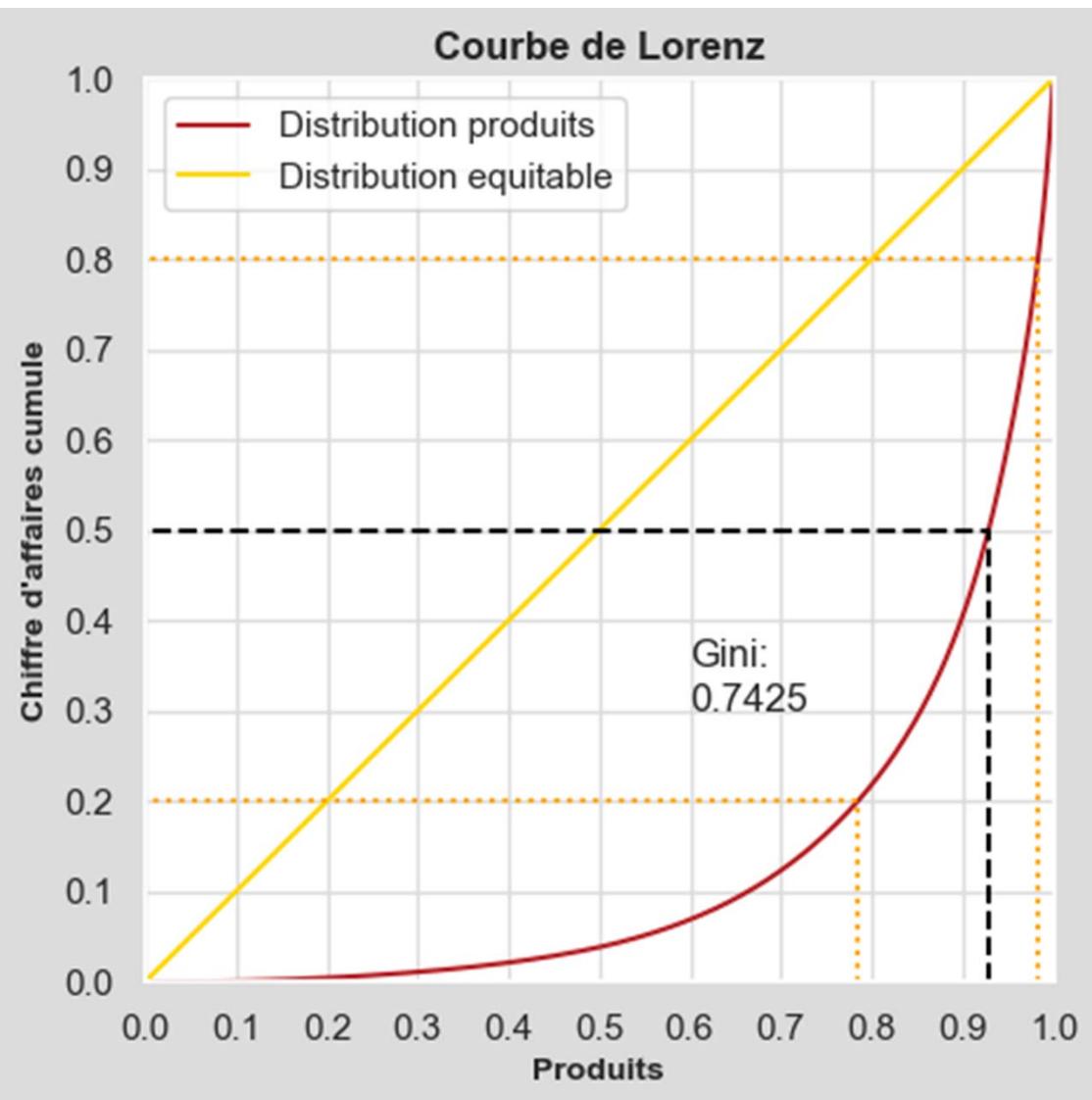
References dans le catalogue produits



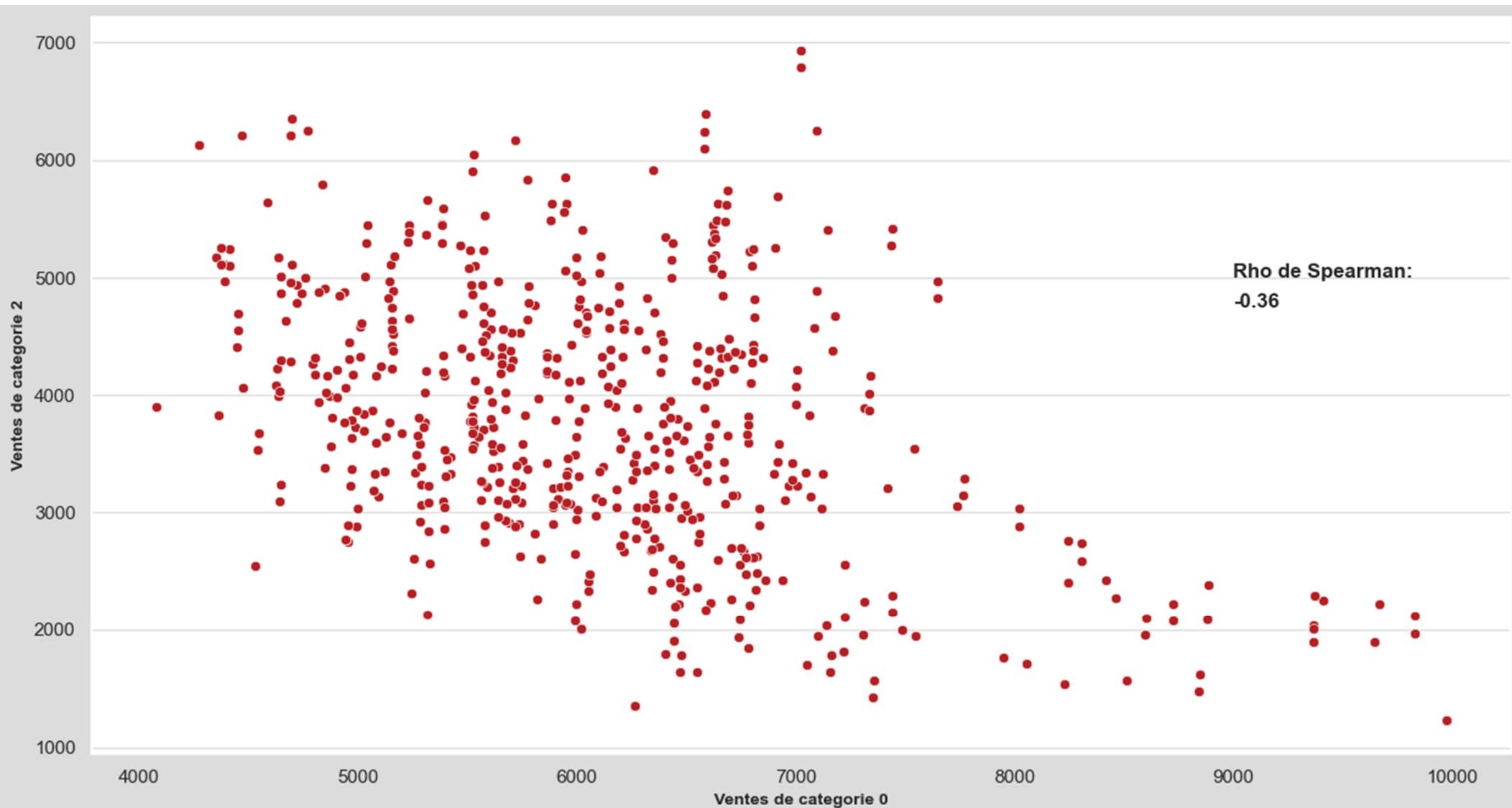
Chiffre d'affaires par categorie de produits



- ❖ Les contributions respectives des catégories de produits au CA sont très inégales :
 - les produits de catégorie 0 représentent plus de 70% de notre catalogue mais à peine plus de 37% du CA
 - Les produits de catégorie 2 représentent moins de 8% de notre catalogue mais plus de 23% du CA



- ❖ Indice de Gini: 0.7425
- ❖ Confirme l'inégalité de contribution des produits au CA
 - 78.50% de nos produits contribuent 20% du CA
 - CA médial atteint avec 92.9% des produits
 - 98.34% des produits contribuent 80% du CA



❖ Top 10 des ventes en volume

product_id_sales	category_books	CA_total	CA_total	nb_sold
		sum	mean	count
1_369		1 54025.48	23.99	2252
1_417		1 45947.11	20.99	2189
1_414		1 51949.4	23.83	2180
1_498		1 49731.36	23.37	2128
1_425		1 35611.04	16.99	2096
1_403		1 35260.4	17.99	1960
1_412		1 32484.15	16.65	1951
1_413		1 34990.55	17.99	1945
1_406		1 48106.59	24.81	1939
1_407		1 30940.65	15.99	1935

- Les meilleures ventes en volume sont toutes des produits de catégorie 1

❖ Top 10 des ventes en valeur

product_id_sales	category_books	CA_total	nb_sold
		sum	count
2_159		2 94893.5	650
2_135		2 69334.95	1005
2_112		2 65407.76	968
2_102		2 60736.78	1027
2_209		2 56971.86	814
1_395		1 54356.25	1875
1_369		1 54025.48	2252
2_110		2 53846.25	865
2_39		2 53060.85	915
2_166		2 52449.12	228

- 8 des meilleures ventes en valeur sont des produits de catégorie 2

❖ Flop 10 des ventes en volume

product_id_sales	category_books	CA_total	CA_total	nb_sold	sum	mean	count
					sum	mean	count
0_2201		0	20.99	20.99	1		
0_1151		0	2.99	2.99	1		
0_1728		0	2.27	2.27	1		
2_81		2	86.99	86.99	1		
0_1539		0	0.99	0.99	1		
0_1284		0	1.38	1.38	1		
0_549		0	2.99	2.99	1		
0_1498		0	2.48	2.48	1		
0_541		0	1.99	1.99	1		
0_886		0	21.82	21.82	1		

❖ Flop 10 des ventes en valeur

product_id_sales	category_books	CA_total	nb_sold	sum	count
				sum	count
0_1840		0	2.56	2	
0_898		0	2.54	2	
0_1498		0	2.48	1	
0_1728		0	2.27	1	
0_541		0	1.99	1	
0_1601		0	1.99	1	
0_807		0	1.99	1	
0_1653		0	1.98	2	
0_1284		0	1.38	1	
0_1539		0	0.99	1	

- Les pires ventes en volume et en valeur sont majoritairement des produits de catégorie 0...
- ... tout comme les 21 références qui n'ont jamais été vendues (cf. liste en Annexe 2)
- 1,066 produits sur 3,266 références ont été vendus moins de 24 fois : plus de 33% de notre catalogue se vend très peu (moins d'1 exemplaire par mois)

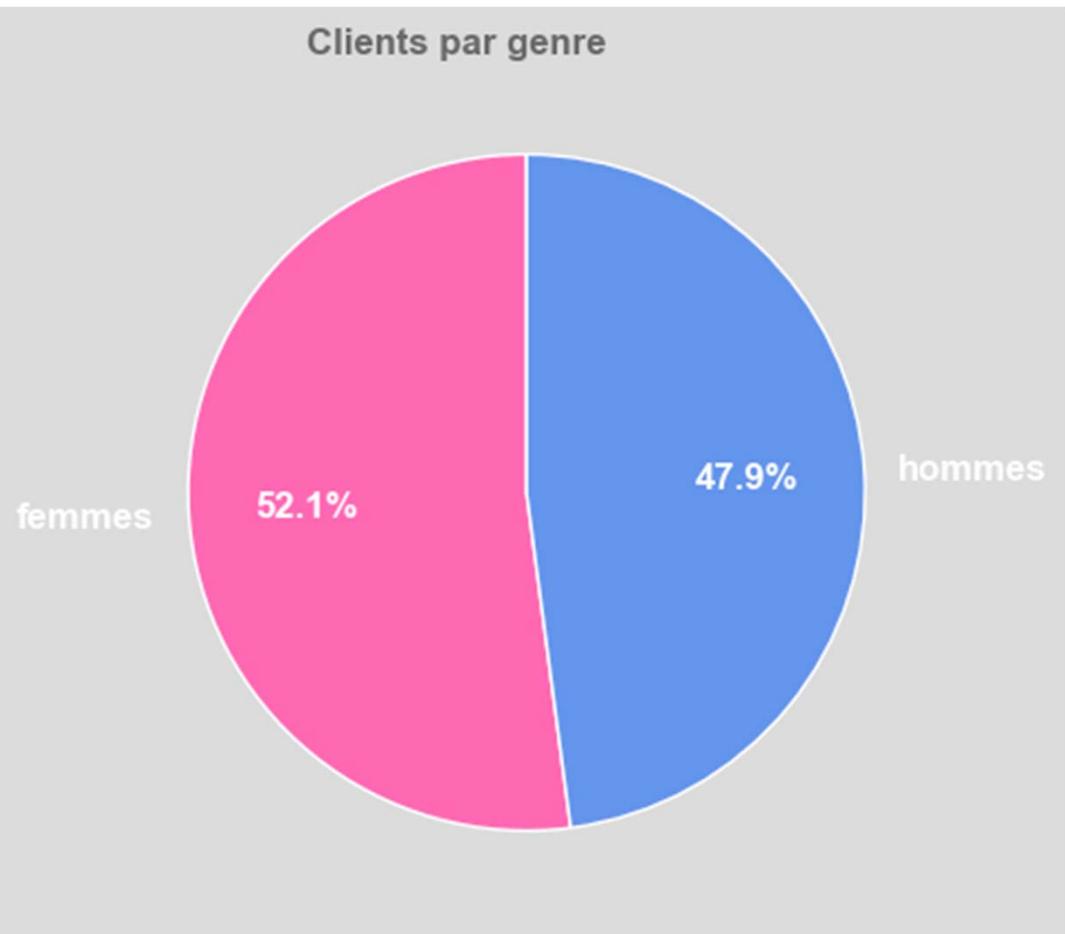
❖ Analyse des associations

Products from top 10	Frequently bought with
1_414	['1_434', '0_1621', '0_1593']
1_412	['1_400']
1_369	['1_277', '0_1059', '0_1205', '1_141']
1_406	['0_1467']
1_425	['0_1525', '0_1641', '0_2231', '1_504', '1_370']
1_417	['0_1525', '0_1641', '0_2231', '1_504', '1_370', '1_320', '1_396']
1_407	['1_433', '0_1205', '1_434', '1_459']
1_498	['0_1047', '0_1596', '0_1616', '1_456', '1_277', '0_1059', '1_141']
1_413	['0_1047', '0_1596', '0_1616', '1_456', '1_441']

- Les produits du top 10 en volume sont principalement associés aux 14 produits 1_434, 1_277, 0_1059, 0_1205, 1_141, 0_1525, 0_1641, 0_2231, 1_504, 1_370, 0_1047, 0_1596, 0_1616 et 1_456 dans le panier des clients, et dans une moindre mesure avec les 9 produits 0_1621, 0_1593, 1_400, 0_1467, 1_320, 1_396, 1_433, 1_459, 1_441
- Information à utiliser pour les recommandations de cross-selling sur la page de validation du panier
- Le produit 1_403 ne figure pas dans cette liste, bien que dans le top 10, car il est vendu seul dans 99.8% des cas et n'est que très rarement associé à d'autres produits que lui-même dans le panier des clients

2 - Indicateurs de vente

2.3 - Profils clients



* source: <https://www.insee.fr/fr/statistiques/4238375?sommaire=4238781>

❖ Test de proportion:

On veut tester si nos clients ont plus de probabilité p_0 d'être des femmes que dans la population française au total (51.6% en 2019 d'après l'INSEE*).

$$H_0 : p_0 \leq 51.6\%$$

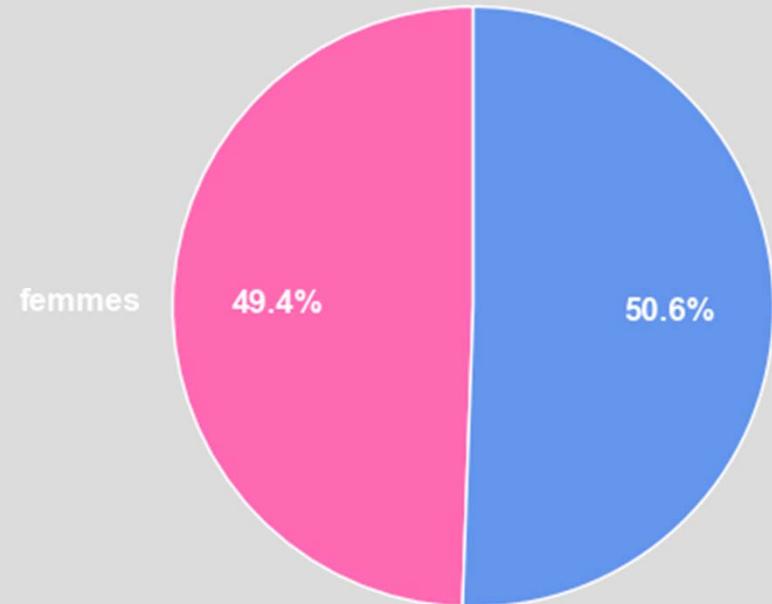
$$H_1 : p_0 > 51.6\%$$

$$\alpha = 5\%$$

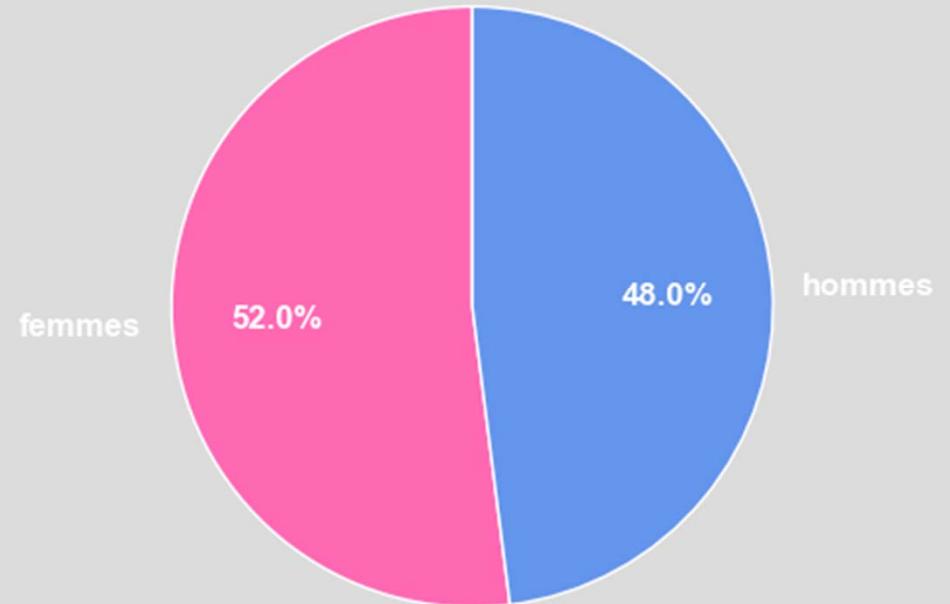
❖ $Z = 0.8961$ et $p\text{-value} = 0.1851$

- au seuil de risque donné, nous n'avons pas suffisamment d'informations pour rejeter H_0 et considérons que la différence entre les proportions de femmes dans nos clients et dans la population française en général n'est pas statistiquement significative mais due au hasard de l'échantillonnage.

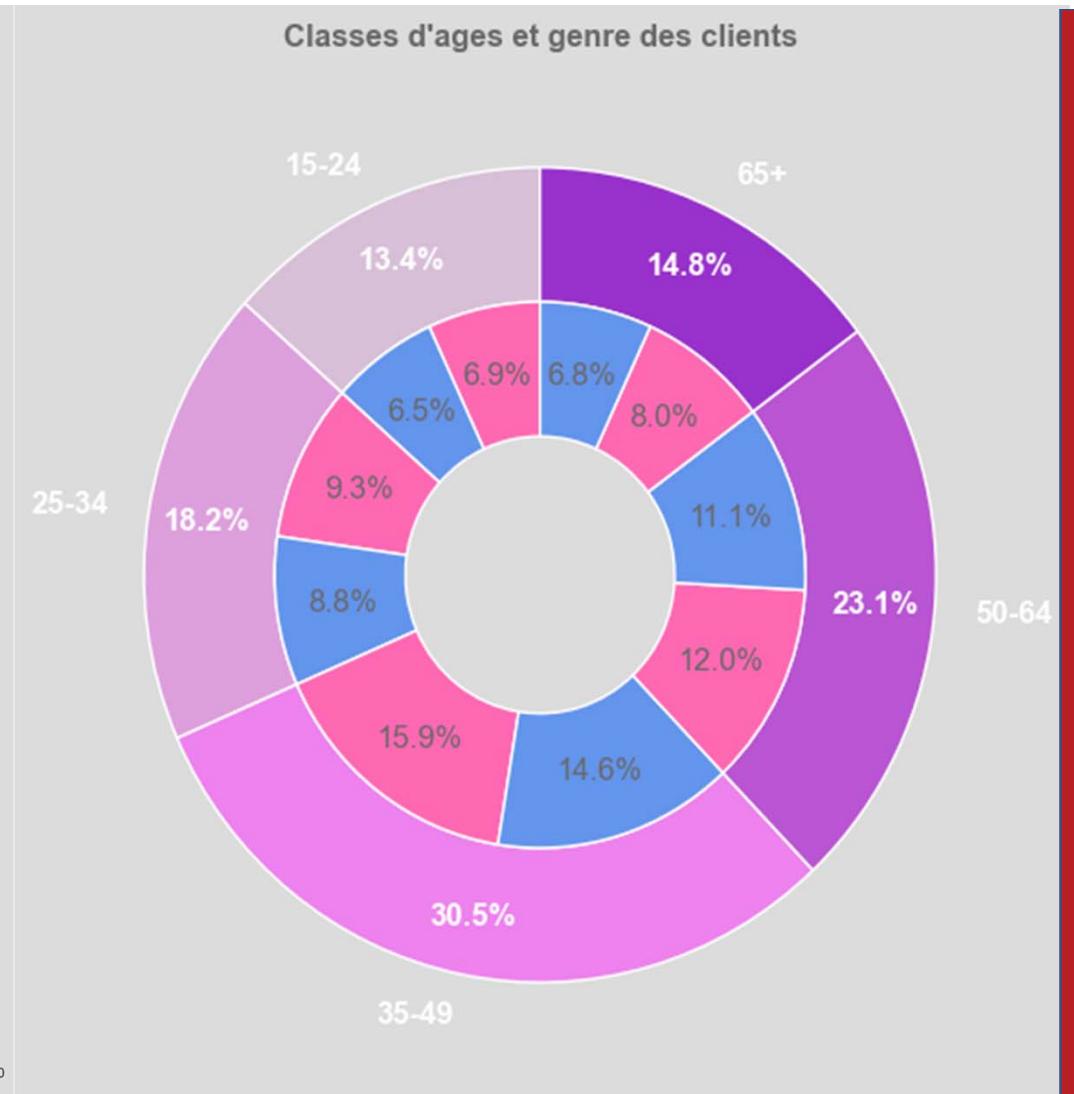
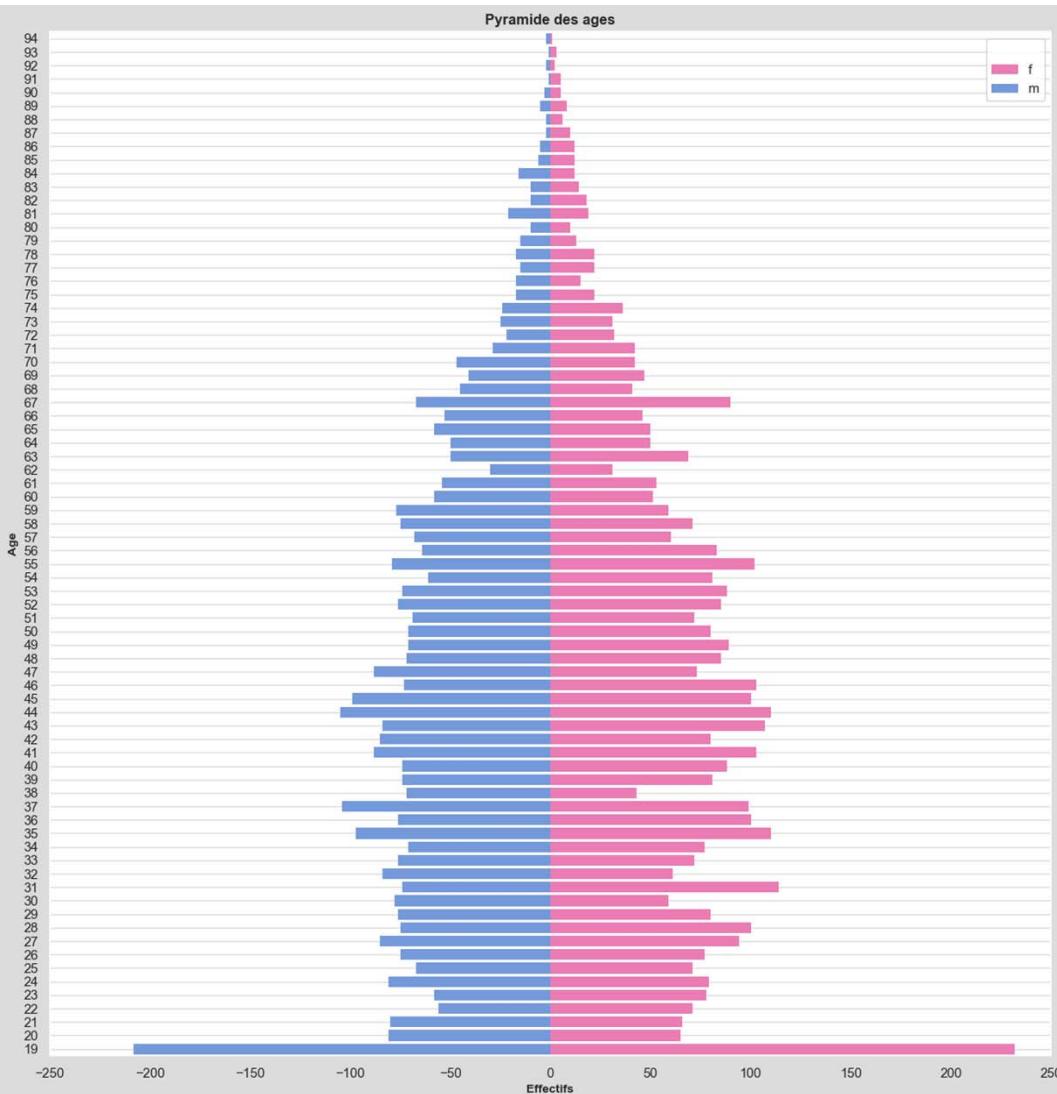
Chiffre d'affaires par genre (clients professionnels inclus)



Chiffre d'affaires par genre (hors clients professionnels)



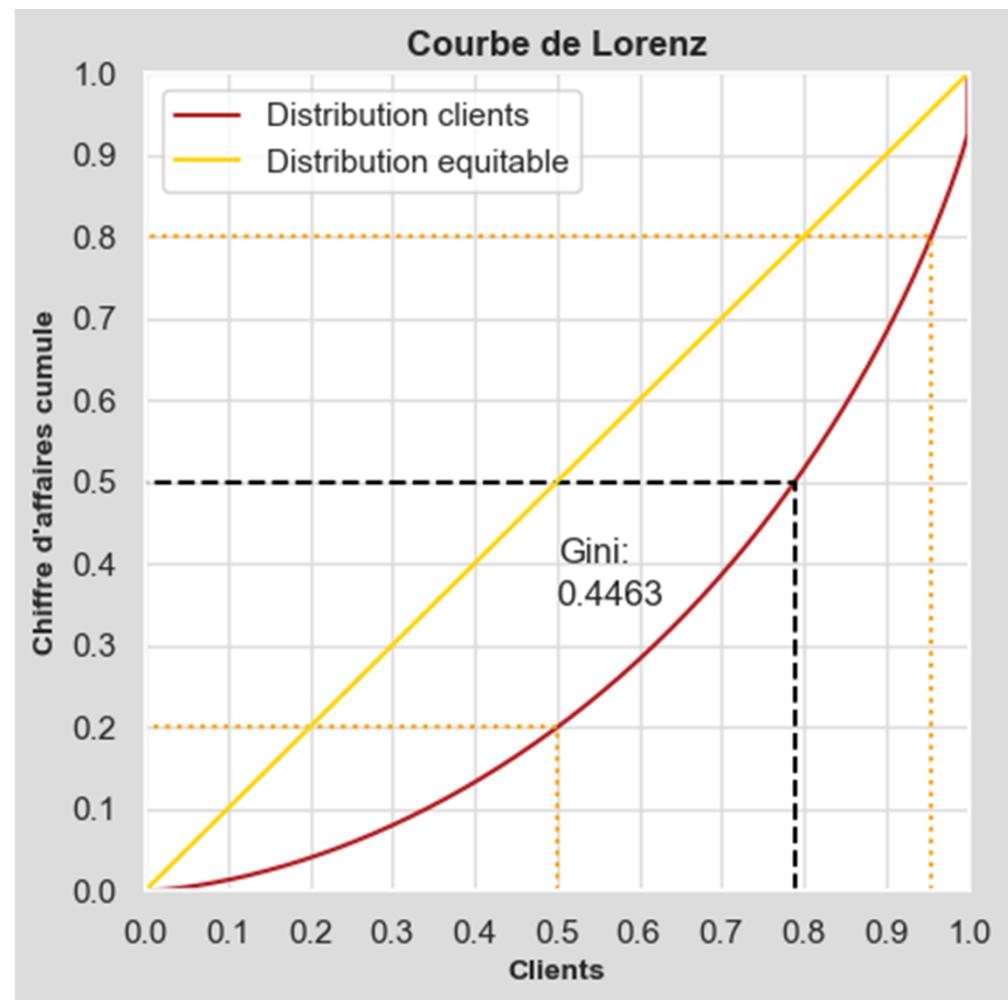
- ❖ Les femmes représentent 52% du CA parmi les clients particuliers et 49.4% du CA tous clients confondus
- ❖ Chez les clients particuliers:
 - Panier moyen des hommes: 19.04 €
 - Panier moyen des femmes: 18.95 €

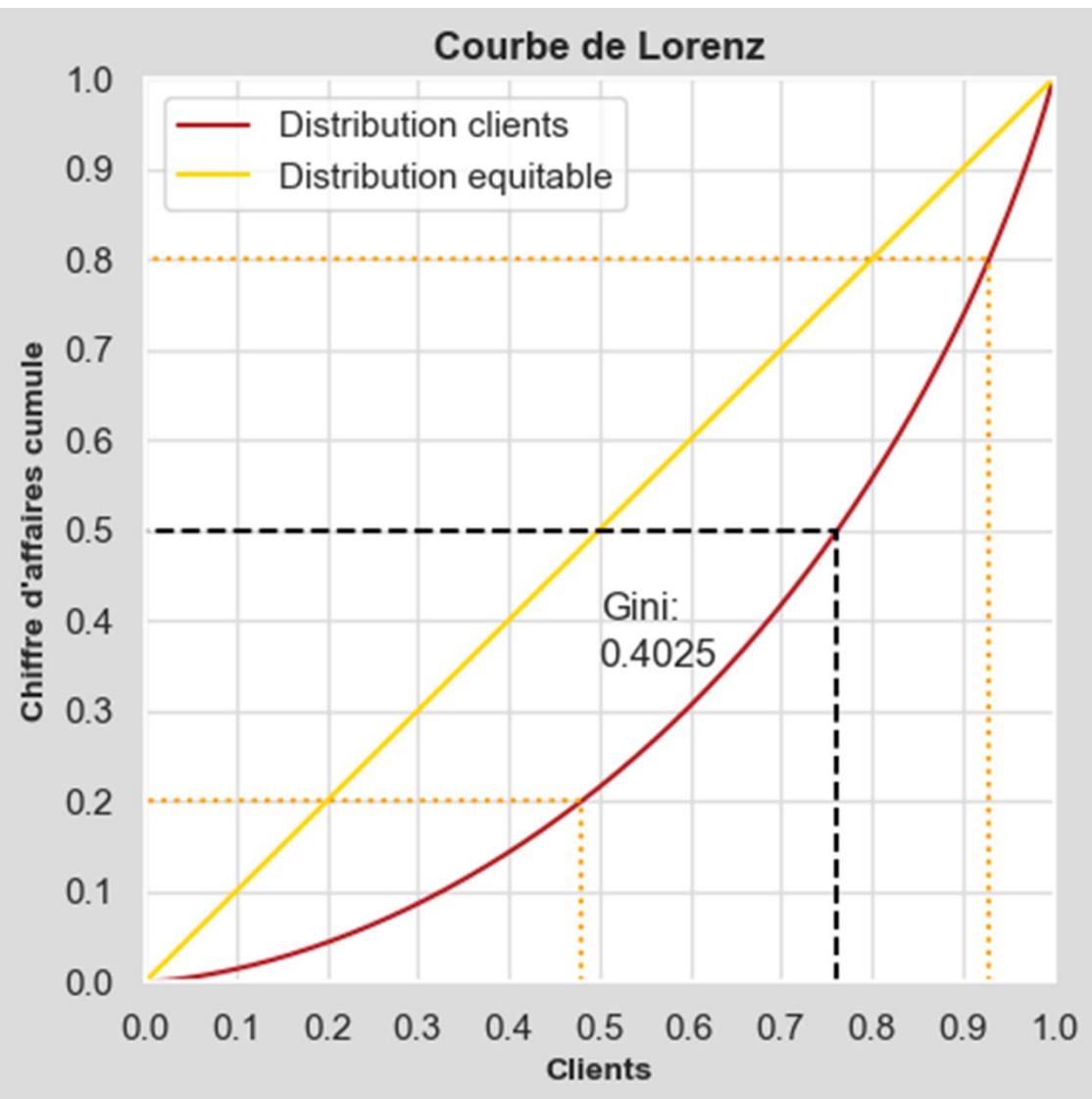


- ❖ Top 10 des clients par chiffre d'affaires individuel

client_id_sales	amount_spent	books_purchased
	sum	count
c_1609	324033.35	25488
c_6714	153660.84	9187
c_3454	113668.89	6773
c_4958	289760.34	5195
c_3263	5276.87	403
c_2140	5208.82	402
c_2595	4959.66	398
c_2077	4816.78	384
c_1637	4698.87	380
c_7421	5050.2	379

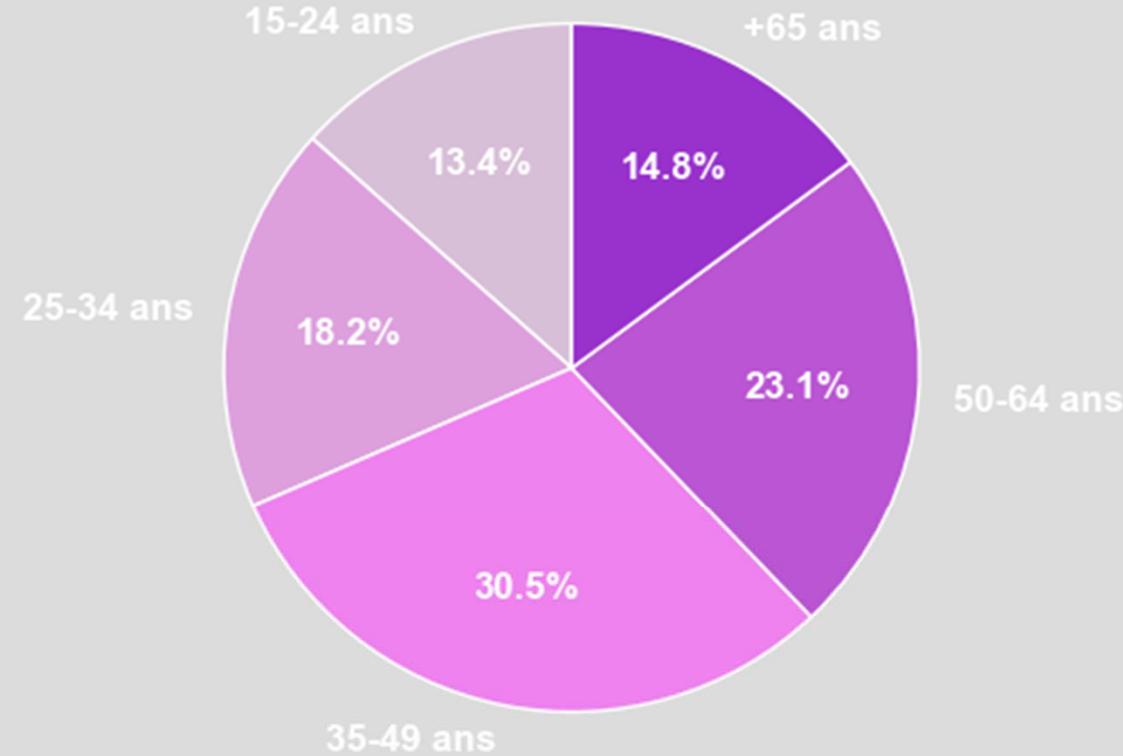
- 4 clients professionnels représentent 7.43% du CA



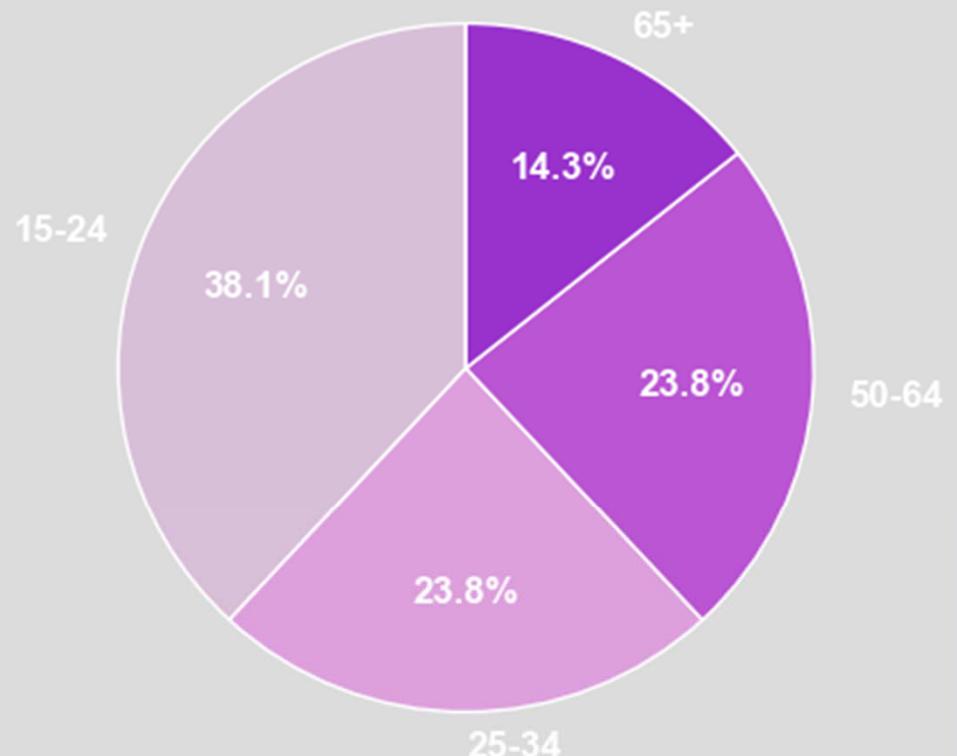


- ❖ Pour les clients individuels (i.e. excluant les 4 clients professionnels) :
 - Indice de Gini: 0.4025
 - Chiffre d'affaires relativement concentré sur certains clients:
 - 20% du CA réalisé avec 48.07% des clients
 - 80% du CA réalisé avec 92.92% des clients
 - CA médial atteint avec 76.21% des clients

Clients acheteurs par classe d'âge

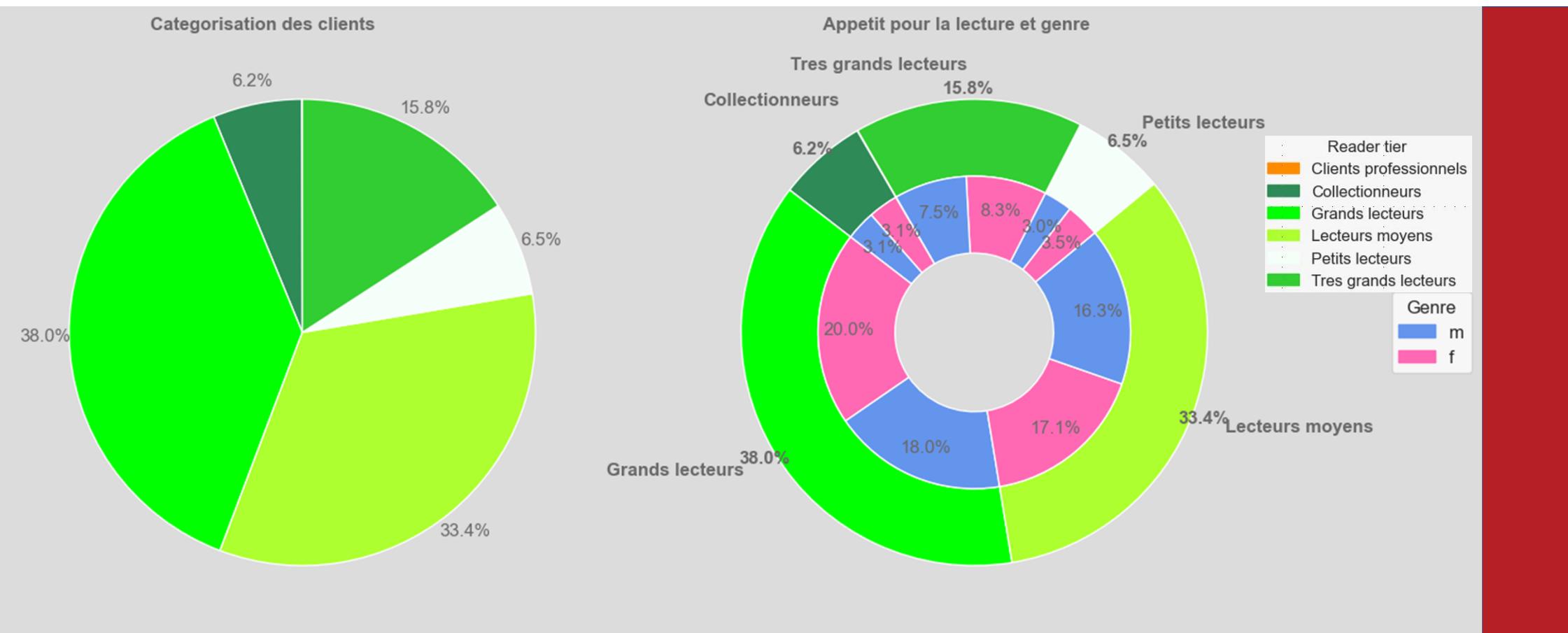


Clients non-acheteurs par classe d'âge



❖ Cf. liste des 20 clients non-acheteurs en Annexe 3

- ❖ Les 15-24 ans sont la classe d'âge majoritaire parmi les clients non-acheteurs
- ❖ La classe des 35-49 ans n'est pas représentée parmi les clients non-acheteurs

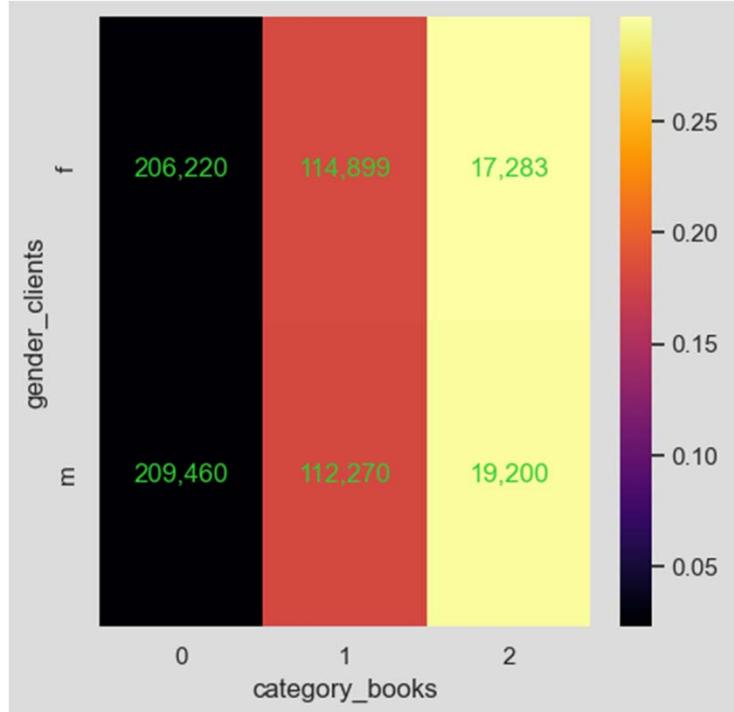


- ❖ Plus 60% de nos clients sont des grands ou très grands lecteurs ou des collectionneurs (plus de 52 livres / an)
- ❖ Près de 40% de nos clients sont composés de petits ou de moyens lecteurs
 - Nos actions marketing devraient probablement se concentrer sur ces derniers

3 – Comportements d'achat

3.1 – Genre et catégories de livres achetées

❖ Heatmap & effectifs réels:



❖ Effectifs théoriques:

		category_books		
		0	1	2
gender_clients	m	207,066.56	113,161.82	18,173.62
	f	208,613.44	114,007.18	18,309.38

❖ Test du chi-2 de contingence: on pose

- H_0 : la distribution des observations est identique pour les 3 catégories de livres aux fréquences théoriques attendues, i.e. le genre du client n'affecte pas les catégories de livres achetées;
- H_1 : la proportion de livres achetés dans chaque catégorie est influencée par du genre du client, i.e. les fréquences observées diffèrent significativement des fréquences théoriques attendues.
- $\alpha = 5\%$

❖ On trouve:

- chi2 calculé : 147.0
- p-value 1.1989607410166063e-32
- Nombre de degrés de liberté: 2
- chi2 théorique : 0.1026
- On rejette H_0 . Les variables ne sont pas indépendantes

❖ V de Cramer : 0.0147

- la dépendance entre ces deux variables est en fait très faible et les livres de catégorie 2 contribuent le plus à cette très faible non-indépendance.

3 – Comportements d'achat

3.2 – Influence de l'âge ...

3.2.1 – ... sur le montant total des achats

❖ Montant total des ventes par classe d'âge

age_range_clients	price_books
15-24	1,514,681.30
25-34	2,169,260.51
35-49	4,099,959.07
50-64	2,044,163.09
65+	1,146,822.01

❖ Nombre de clients par classe d'âge

age_range_clients	client_count
15-24	1,155
25-34	1,566
35-49	2,633
50-64	1,991
65+	1,276

❖ Nombre de ventes par classe d'âge

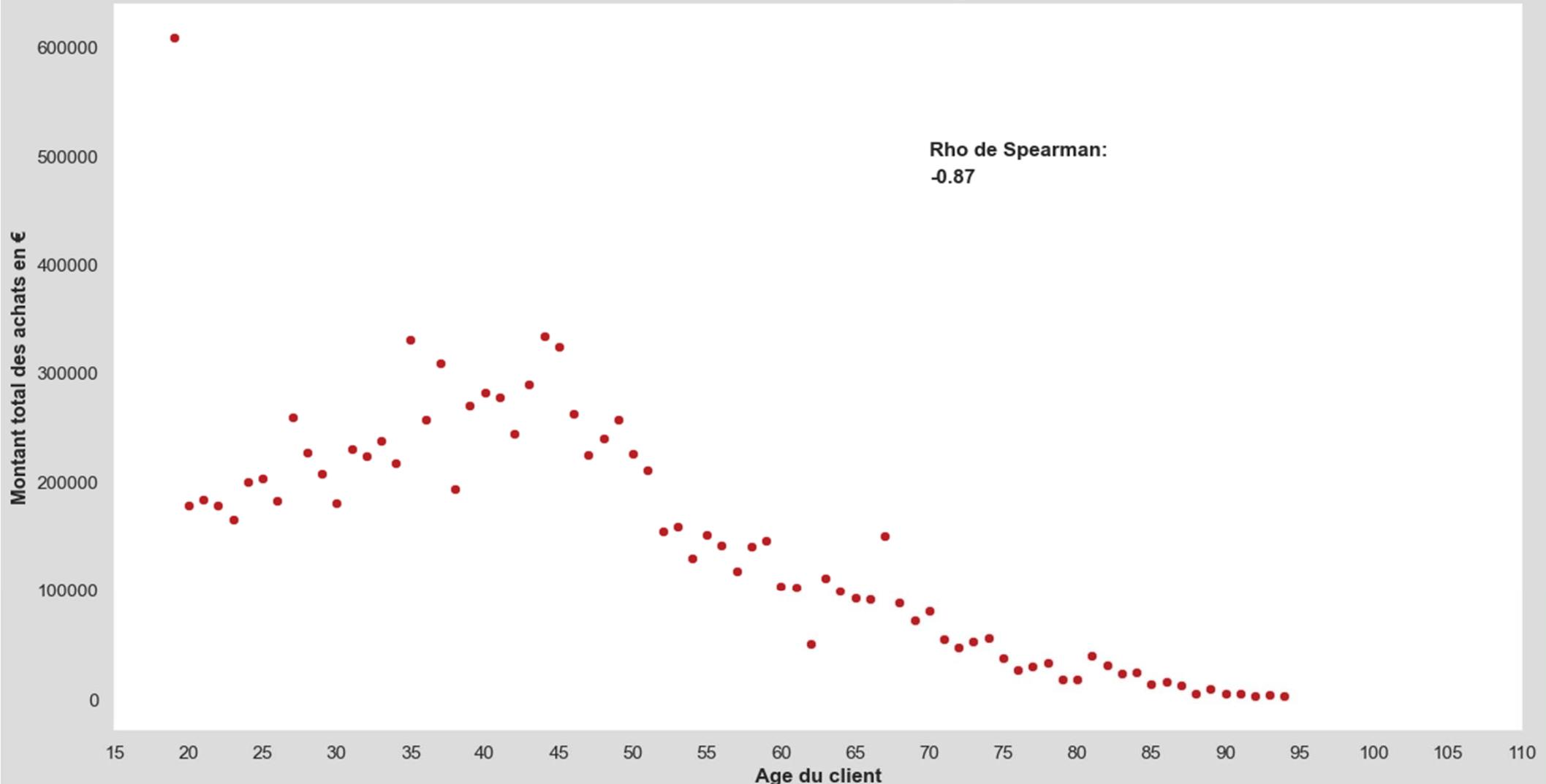
age_range_clients	product_id_sales
15-24	37,820.00
25-34	87,911.00
35-49	309,876.00
50-64	128,536.00
65+	68,546.00

❖ Panier moyen par classe d'âge

age_range_clients	price_books
15-24	40.05
25-34	24.68
35-49	13.23
50-64	15.90
65+	16.73

- Les 15-24 ans et les 25-34 ans dépensent le plus par achat mais ils réalisent moins de transactions que les 35-49 ans et les 50-64 ans, probablement car leurs effectifs totaux dans notre pool de clients sont moindres que ces 2 classes.
- Il pourrait donc être pertinent d'organiser des campagnes marketing ciblant les 2 classes de moins de 35 ans pour augmenter leur proportion dans le total de nos clients.

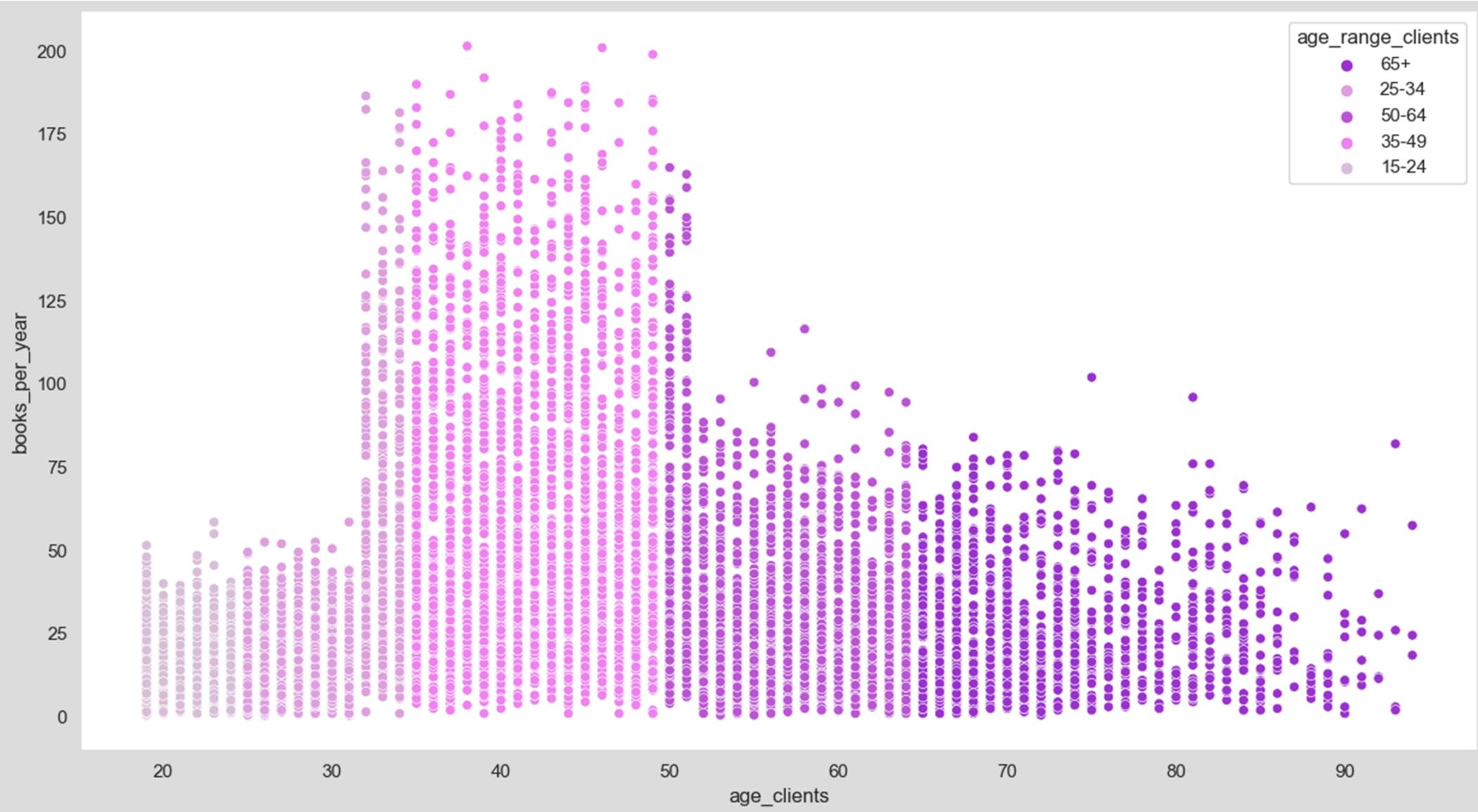
Montant total des achats en fonction de l'age des clients

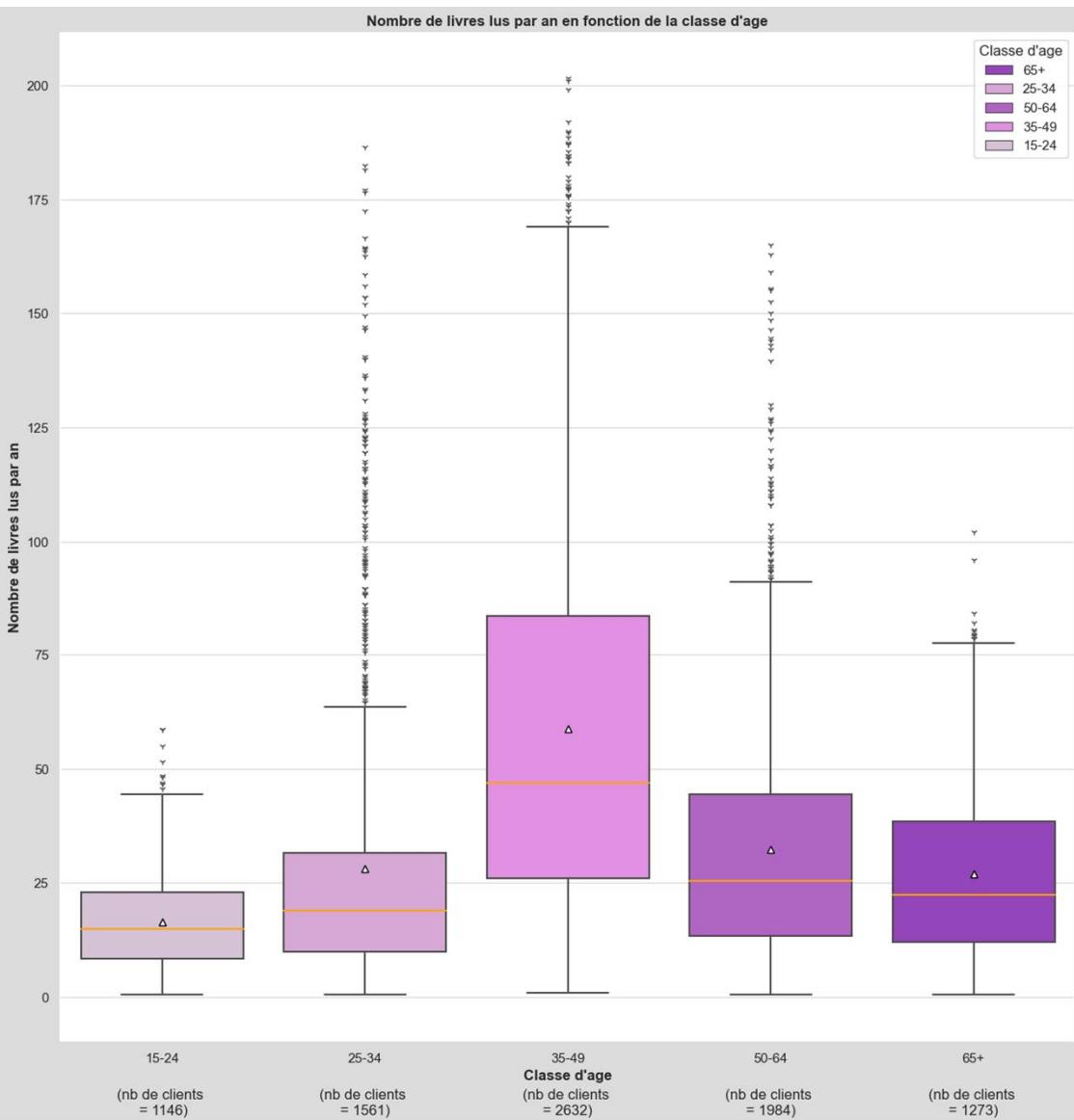


3 - Comportements d'achat

3.2 - Influence de l'âge ...

3.2.2 - ... sur la fréquence d'achat





❖ Test de Kruskal-Wallis:

- H_0 : le nombre médian de livres lus par an est le même pour toutes les classes d'âge
- H_1 : le nombre médian de livres d'au moins une classe d'âge est différent de celui
- $\alpha = 5\%$.

❖ On trouve:

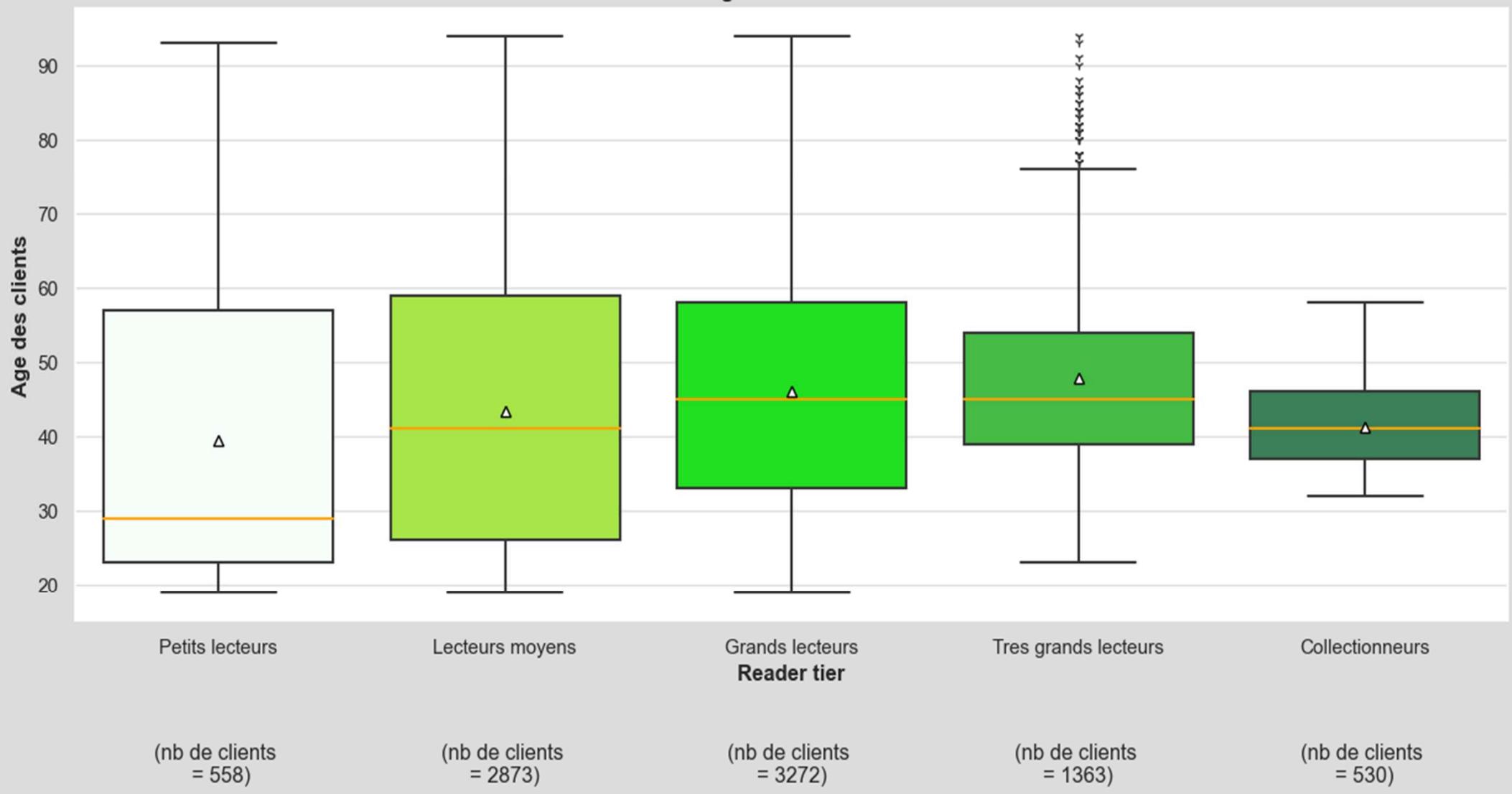
- Statistique de test: 1834.9450615796056
- p-value: 0.000000e+00
- On rejette donc l'hypothèse nulle: au moins une classe d'âge a une médiane du nombre de livres lus par an différente des 4 autres.

❖ Post-hoc test: test de Dunn:

	1	2	3	4	5
1	1	2.38E-20	1.06E-307	8.05E-70	6.54E-34
2	2.38E-20	1	1.88E-201	2.29E-18	0.00036
3	1.06E-307	1.88E-201	1	5.27E-113	2.08E-131
4	8.05E-70	2.29E-18	5.27E-113	1	7.31E-06
5	6.54E-34	0.00036	2.08E-131	7.31E-06	1

- Le test de Dunn donnant des p-values pour toutes les différences entre les classes d'âge très proches de zéro, nous pouvons en conclure que toutes nos classes d'âge ont des médianes du nombre de livres lus par an différentes au seuil $\alpha = 5\%$.

Age et reader tier



- ❖ Test de Kruskal-Wallis:
 - H₀ : l'âge médian est le même pour tous les reader tiers
 - H₁ : l'âge médian d'au moins un reader tier est différent de celui des 2 autres
 - $\alpha = 5\%$.
- ❖ On trouve:
 - Statistique de test: 211.1079212441607
 - p-value: 1.534854e-44
 - On rejette donc l'hypothèse nulle, au moins un reader tier a un âge médian d'achat différent des 4 autres.

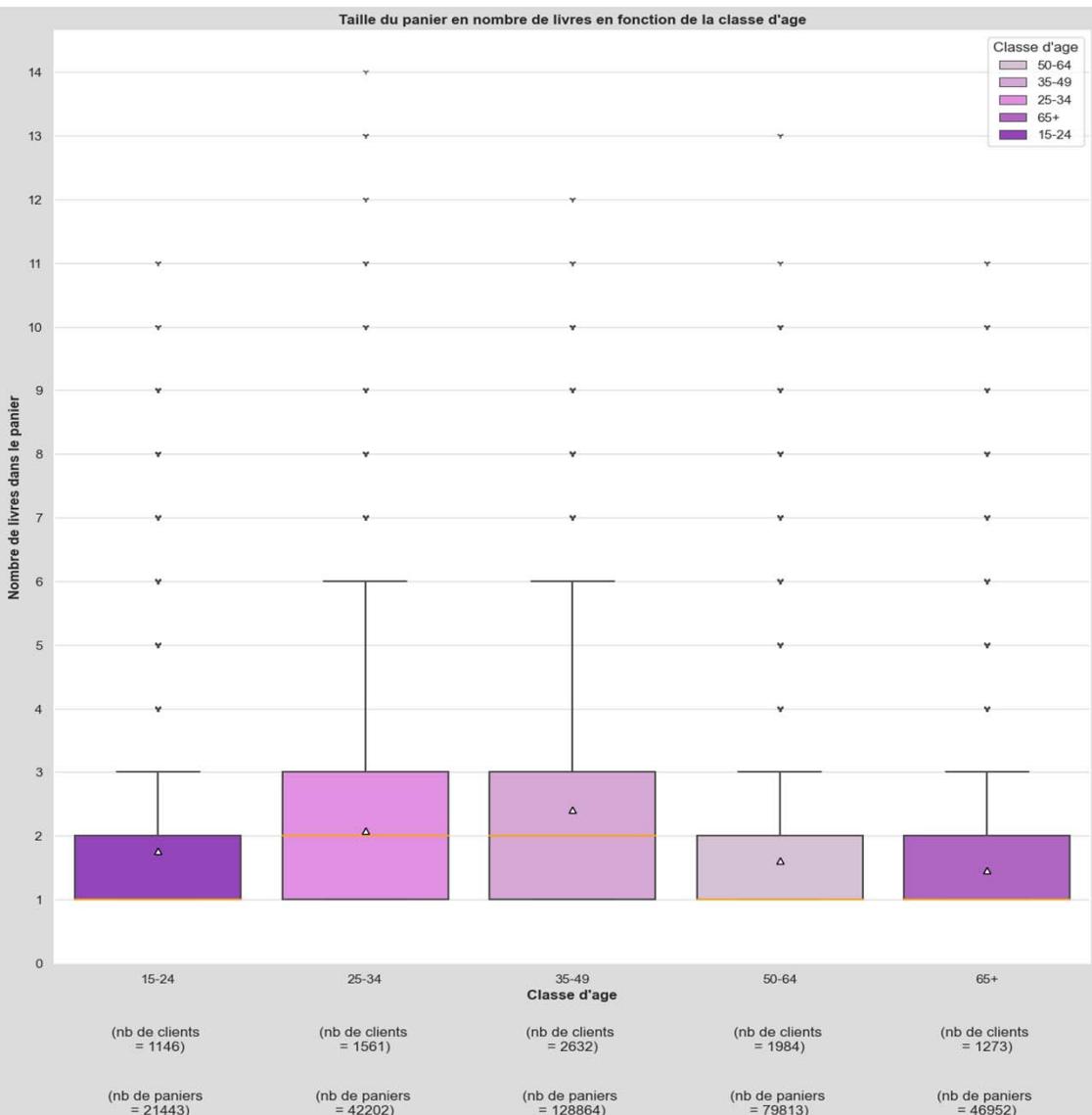
	1	2	3	4	5
1	1	2.41E-07	0.317312	0.000781	1.28E-13
2	2.41E-07	1	2.75E-14	2.27E-22	2.02E-05
3	0.317312	2.75E-14	1	5.75E-08	5.89E-24
4	0.000781	2.27E-22	5.75E-08	1	4.09E-31
5	1.28E-13	2.02E-05	5.89E-24	4.09E-31	1

- ❖ Post-hoc test: test de Dunn
 - Le test de Dunn donnant des p-values pour toutes les différences entre les reader tiers très proches de zéro sauf entre les catégories 1 et 3, nous pouvons en conclure que tous nos reader tiers ont des médianes différentes au seuil $\alpha = 5\%$ sauf les grands et les très grands lecteurs

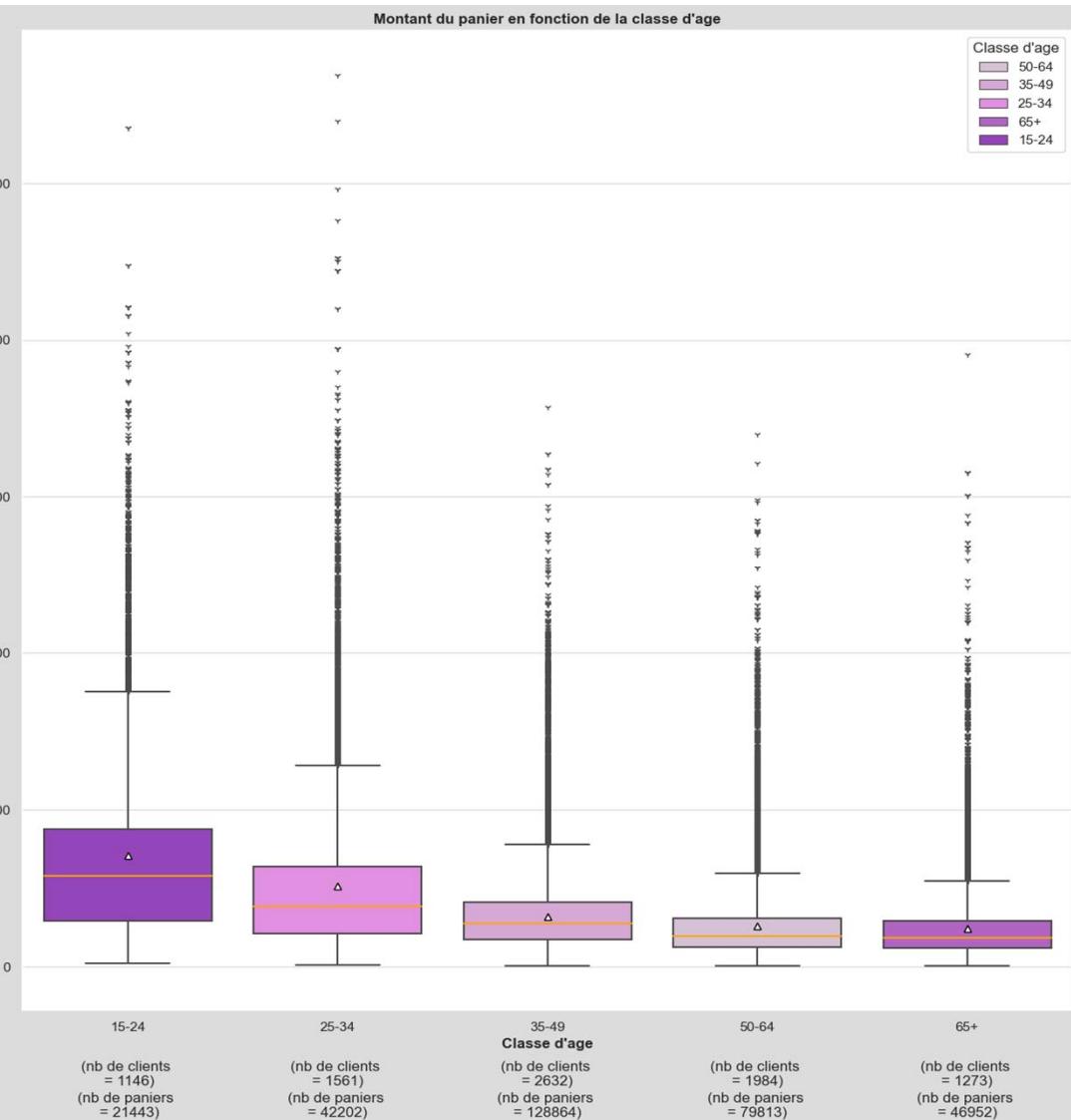
3 - Comportements d'achat

3.2 - Influence de l'âge ...

3.2.3 - ... sur la taille du panier moyen

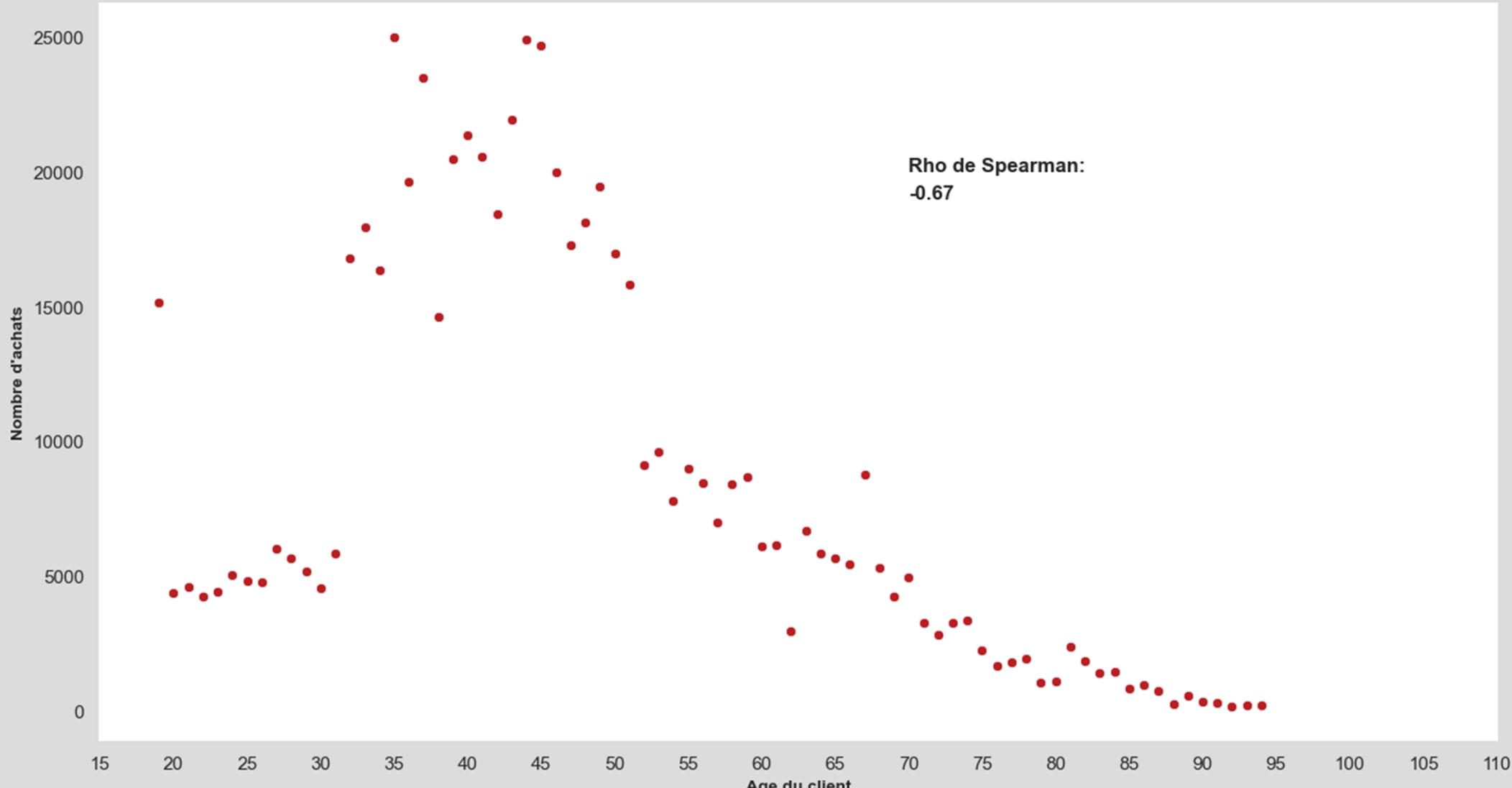


- ❖ Rho de Spearman: -0.23
- ❖ p-value: 0.00
- ❖ Aucune tendance claire ne se dégage des boxplots, et le coefficient de Spearman reste relativement proche de zéro la p-value est significative, donc nous pouvons conclure qu'il existe une relation monotone décroissante faible entre l'âge des clients et le nombre de livres dans leur panier.



- ❖ Le montant du panier médian semble décroître avec la classe d'âge.
- ❖ Rho de Spearman:
-0.34
- ❖ p-value: 0.00
- ❖ La p-value est significative, donc on peut ici conclure à une relation monotone décroissante faible entre l'âge d'un client et le montant de son panier médian.

Nombre d'achats en fonction de l'age des clients

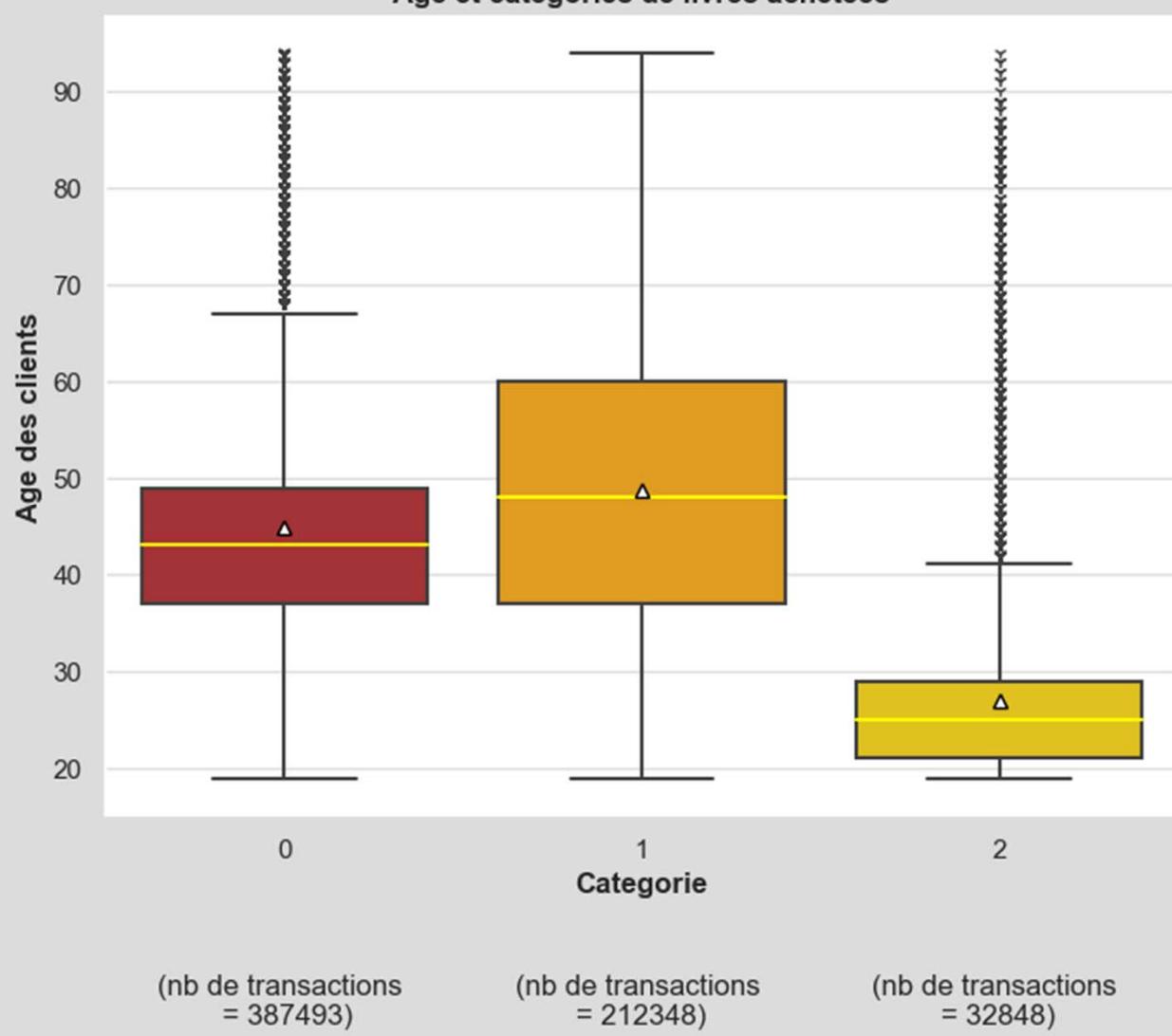


3 – Comportments d'achat

3.2 – Influence de l'âge ...

3.2.4 – ... sur les catégories de livres achetées

Age et categories de livres achetees



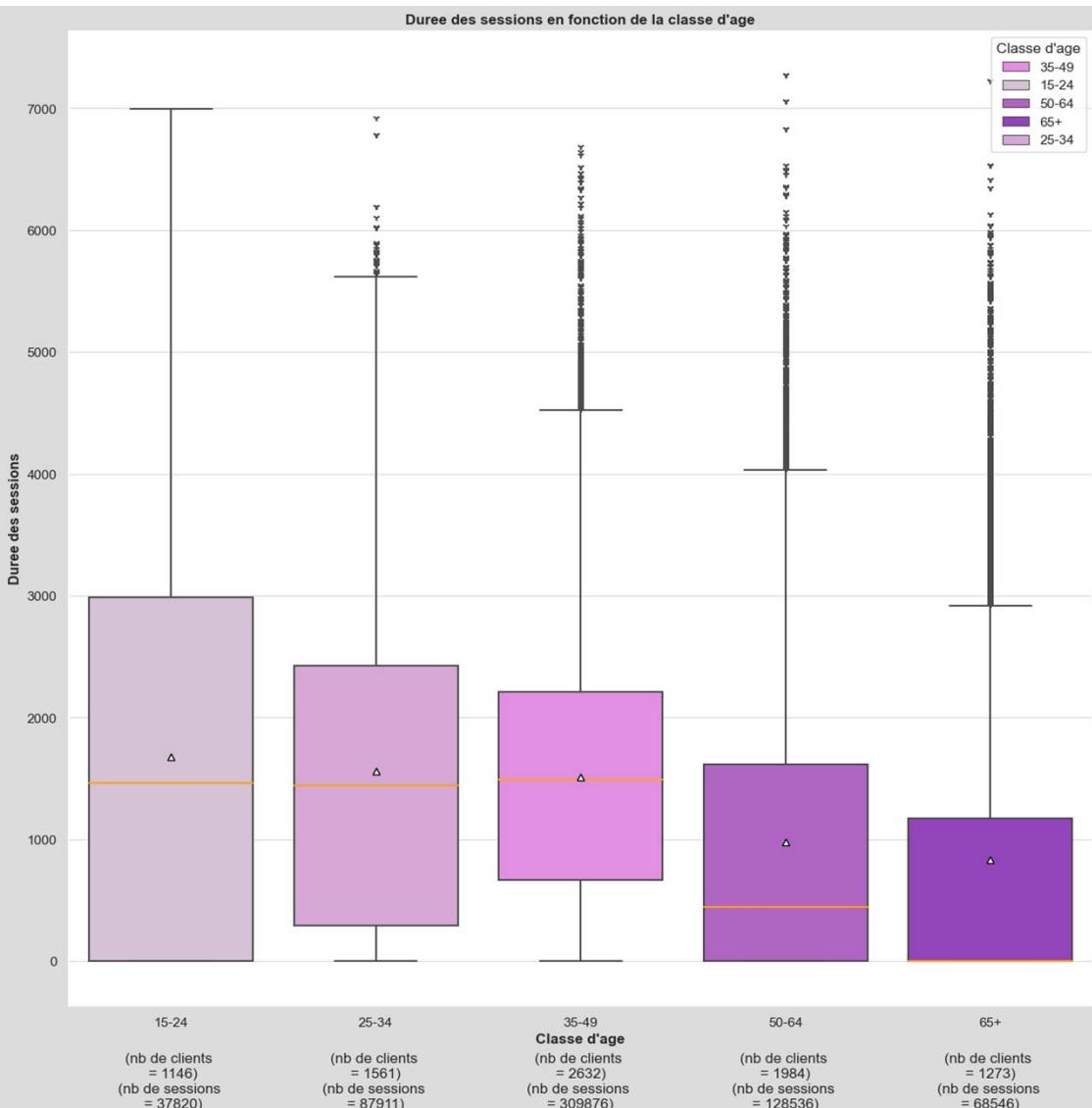
❖ Test de Kruskal-Wallis:

- H_0 : l'âge médian est le même pour toutes les catégories de livres
- H_1 : l'âge médian d'au moins une catégorie de livres est différent de celui des 2 autres
- $\alpha = 5\%$

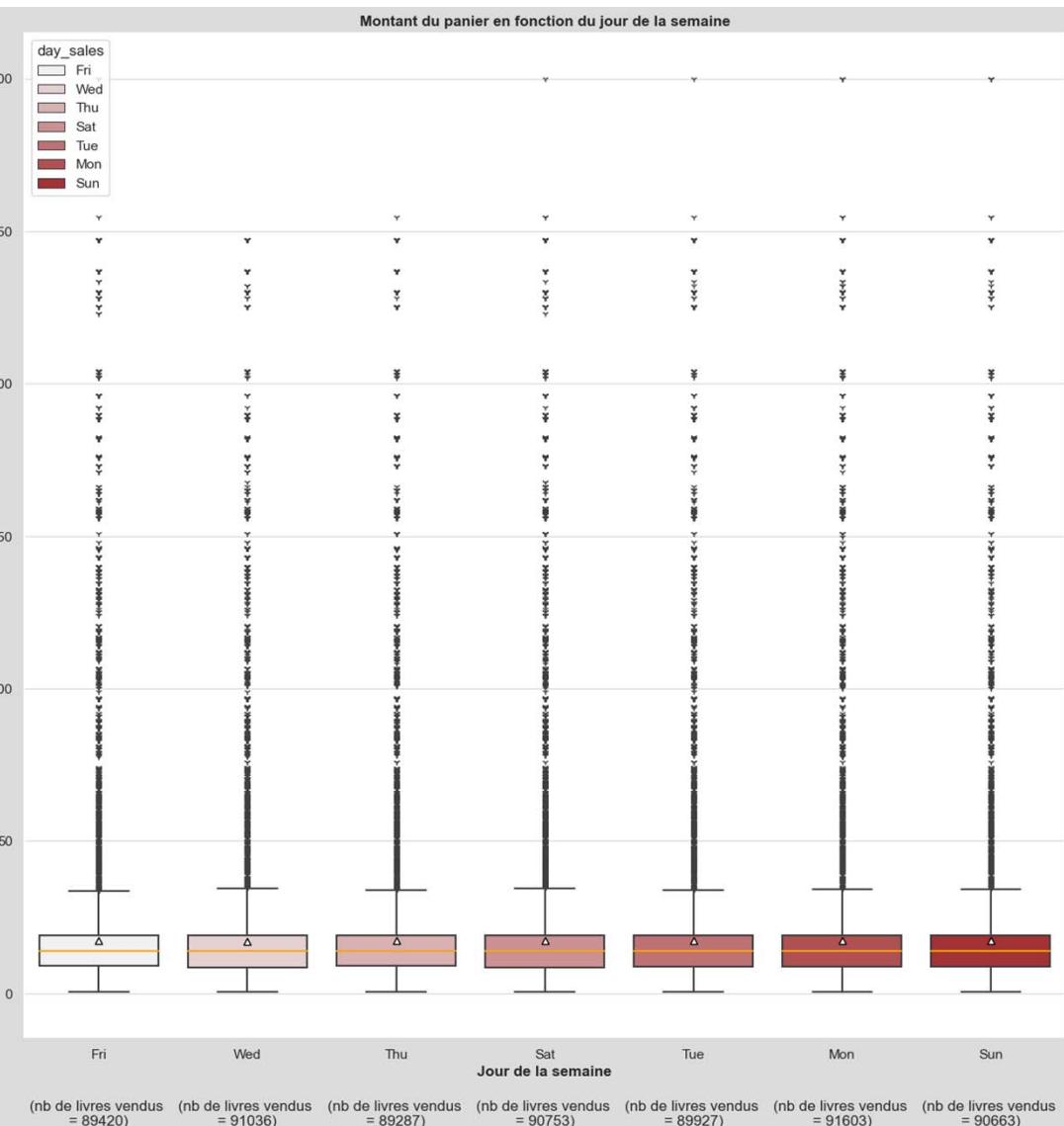
❖ On trouve:

- Statistique de test: 72214.83433330593
- p-value: 0.000000e+00
- On rejette donc l'hypothèse nulle, au moins une catégorie de livres a un âge médian d'achat différent des 2 autres.

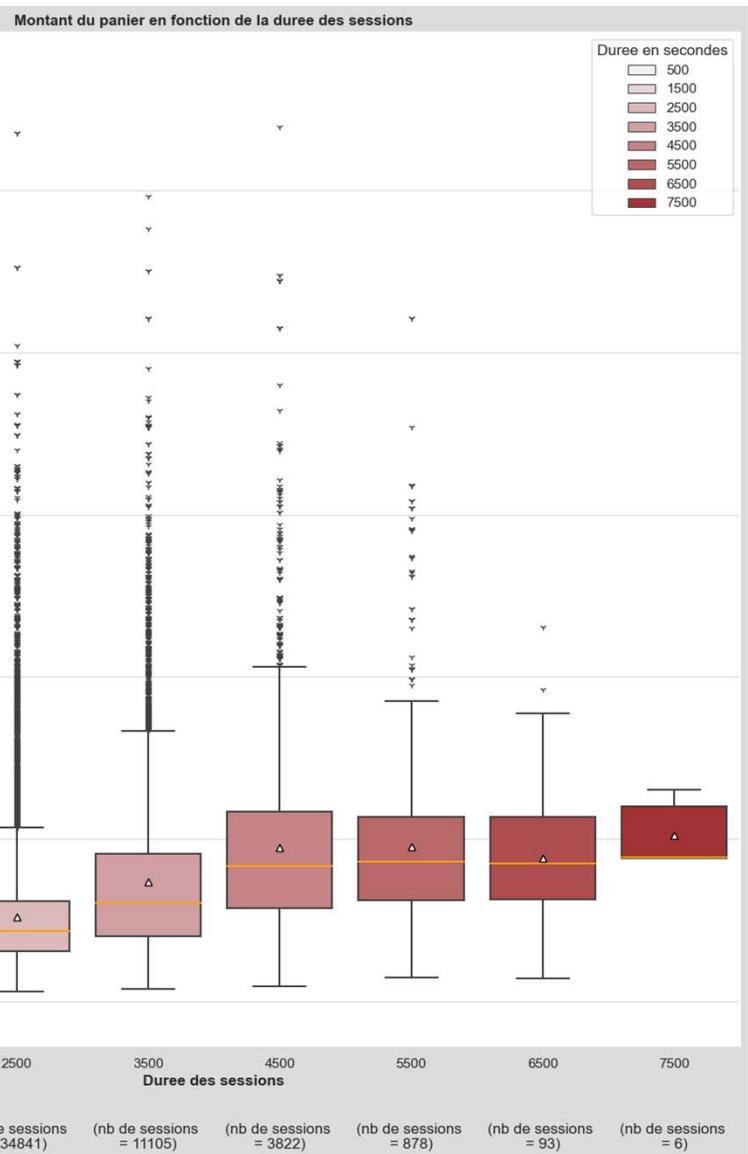
3.3 - Autres indicateurs



- ❖ Le temps médian passé en ligne est relativement proche pour toutes les classes d'âges de moins de 50 ans et s'établit autour de 1,500 secondes (environ 25 minutes)
- ❖ Ce résultat est biaisé par le fait que notre système de vente en ligne ne permet de calculer une durée de session que pour les paniers comportant plus d'un article
- ❖ Il conviendrait de changer les modalités d'horodatage de notre magasin en ligne afin d'enregistrer la durée de toutes les sessions pour pouvoir raffiner cette analyse ainsi que l'analyse sur le lien entre la durée des sessions et l'âge



- ❖ Il ne semble pas exister de lien entre le jour de la semaine et le montant du panier; les paniers moyens et médian sont extrêmement proches pour tous les jours de la semaine



- ❖ Les montants moyen et médian du panier semblent croître avec la durée des sessions
- ❖ Ici encore, analyse faussee par le système d'horodatage

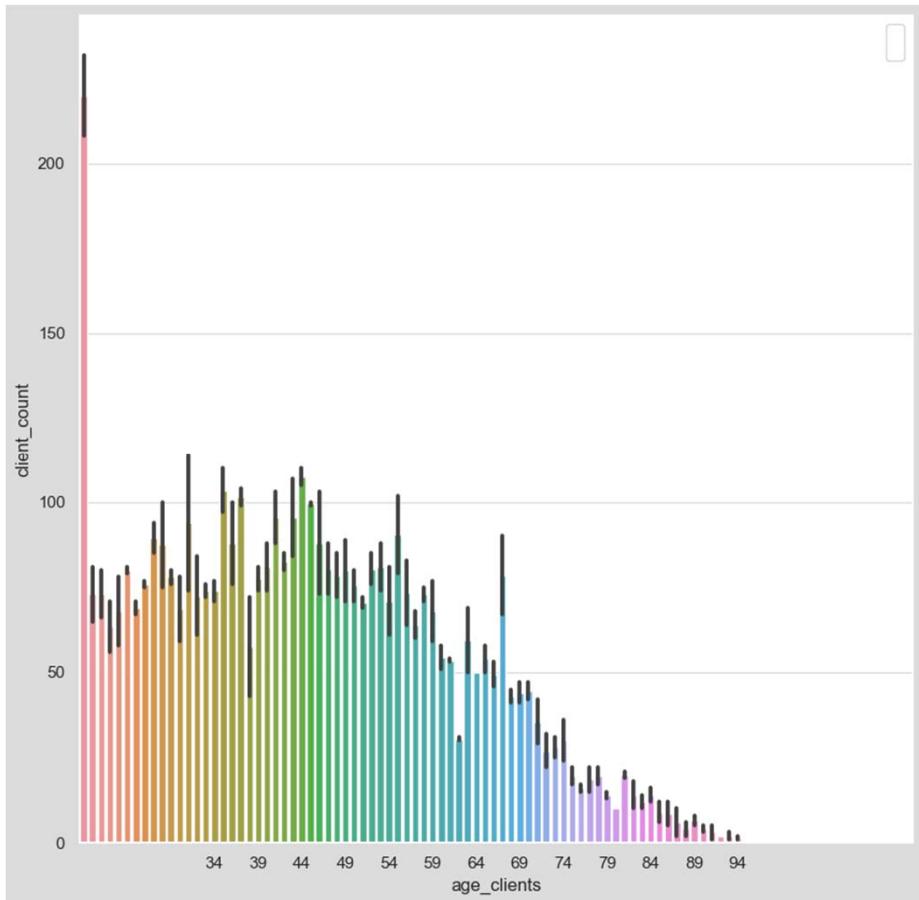
Conclusions & recommandations

- ❖ Problèmes à corriger dans l'acquisition des données : ventes non enregistrées, produits absents du référentiel, données de test dans une base de prod, clients achetant de multiples exemplaires du même ouvrage (cf. analyse des associations), stockage des données non conforme à la 1NF
- ❖ Information client à enrichir pour affiner l'analyse : niveau d'études, CSP, loisirs favoris hors lecture, zone géographique, fréquentation de nos magasins physiques
- ❖ Suivi individualisé : à mettre en place pour les clients professionnels
- ❖ Historique à compléter : 2 exercices comptables ne suffisent pas pour permettre la mise en place d'une analyse RFM (à supposer qu'on juge une telle analyse pertinente pour un bien culturel comme le livre) ni pour évaluer le coût d'acquisition de nouveaux clients, le taux d'attrition et la durée de vie d'un client de façon significative
- ❖ Offre produits à revoir : près de 33% se vendent peu, offre de catégorie 2 à élargir (up-selling), offre de catégorie 0 à repenser (beaucoup ont peu de succès), analyse des associations à utiliser pour faire des recommandations (cross-selling)
- ❖ Information produit à enrichir : analyser séparément les e-books et les livres papier, enrichir le référentiel produit (catégories détaillées par genre littéraire, existence d'une suite ou série, auteur, éditeur, date de sortie, prix littéraires) pour affiner l'analyse et améliorer les recommandations aux clients

Annexes

Annexe 1 - Tests de normalité

❖ âge des clients



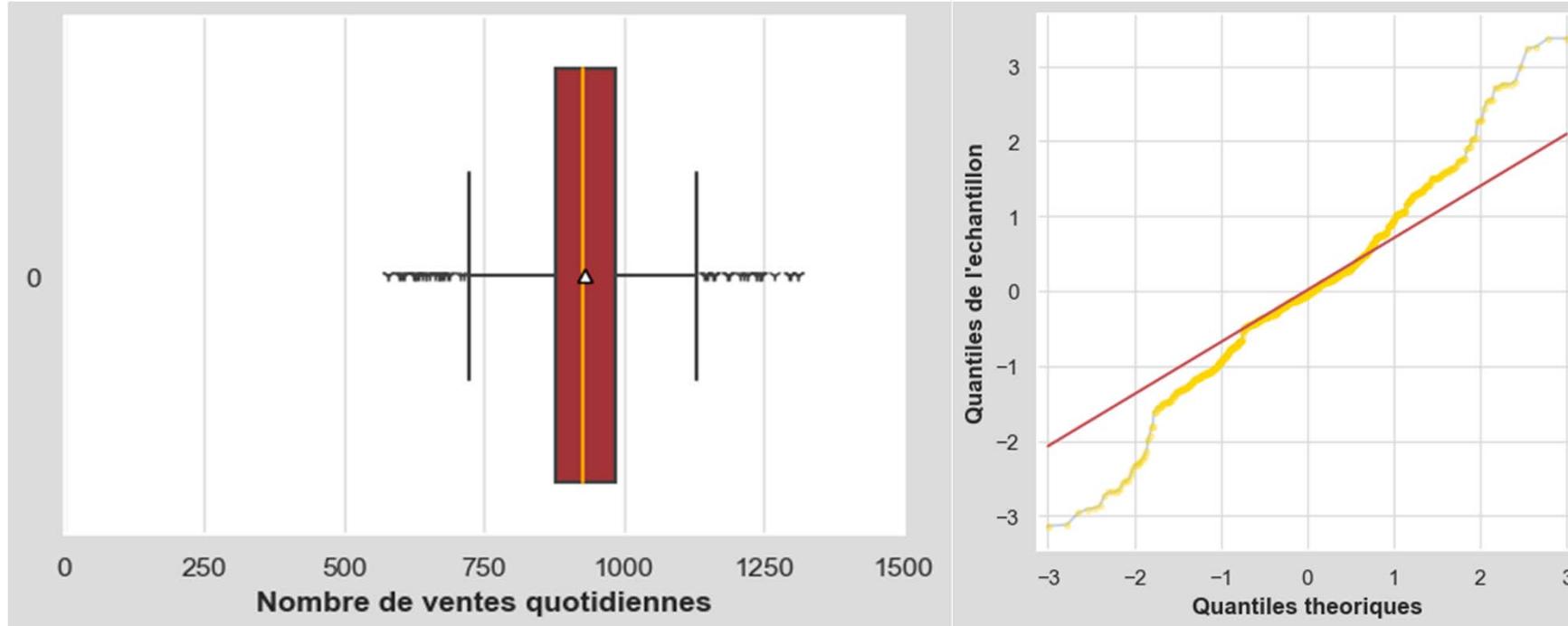
Test d'Anderson-Darling:

```
statistic=1.6715963597803238,  
critical_values=array([0.562, 0.64 , 0.768,  
0.895, 1.065])  
significance_level=array([15. , 10. , 5. ,  
2.5, 1. ])
```

La statistique de test vaut 1.6716. Elle est supérieure à toutes les valeurs critiques pour tous les niveaux de risque testés.

Les résultats du test sont donc significatifs jusqu'au seuil de risque $\alpha = 1\%$, et nous avons donc suffisamment d'informations pour rejeter H_0 et conclure que l'âge de nos clients n'est pas distribué selon une loi $N(\mu, \sigma)$.

❖ ventes quotidiennes

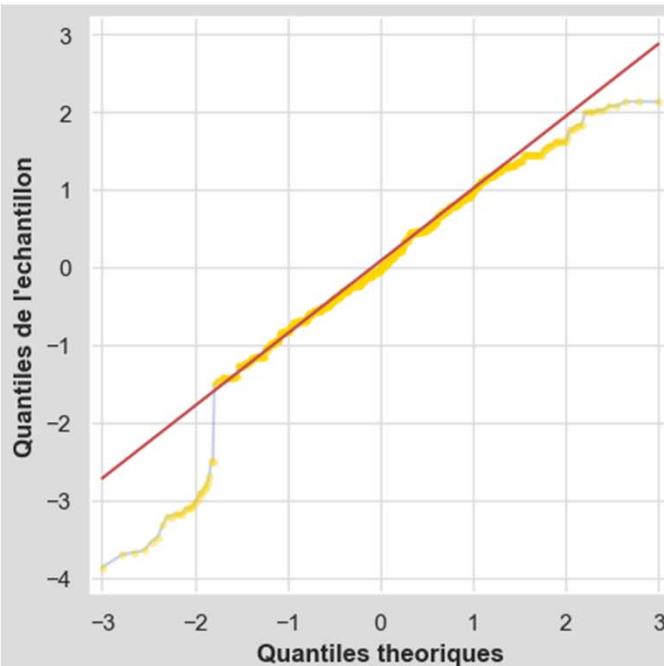
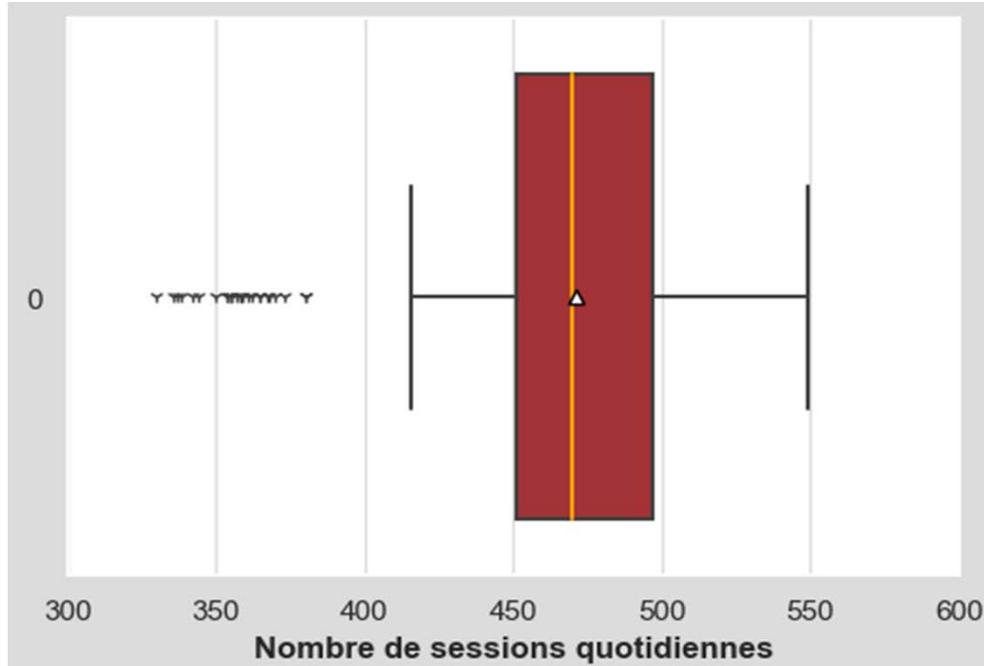


Test de Shapiro-Wilk: Statistique de test: 0.9786 / p-value: 0.0

Test d'Anderson-Darling: statistic=5.3202
critical_values=array([0.573, 0.652, 0.783, 0.913, 1.086])
significance_level=array([15., 10., 5., 2.5, 1.])

Dans les 2 cas, on rejette H_0 au seuil $\alpha = 5\%$. La distribution du nombre de ventes quotidiennes ne suit pas une loi Normale.

❖ sessions quotidiennes

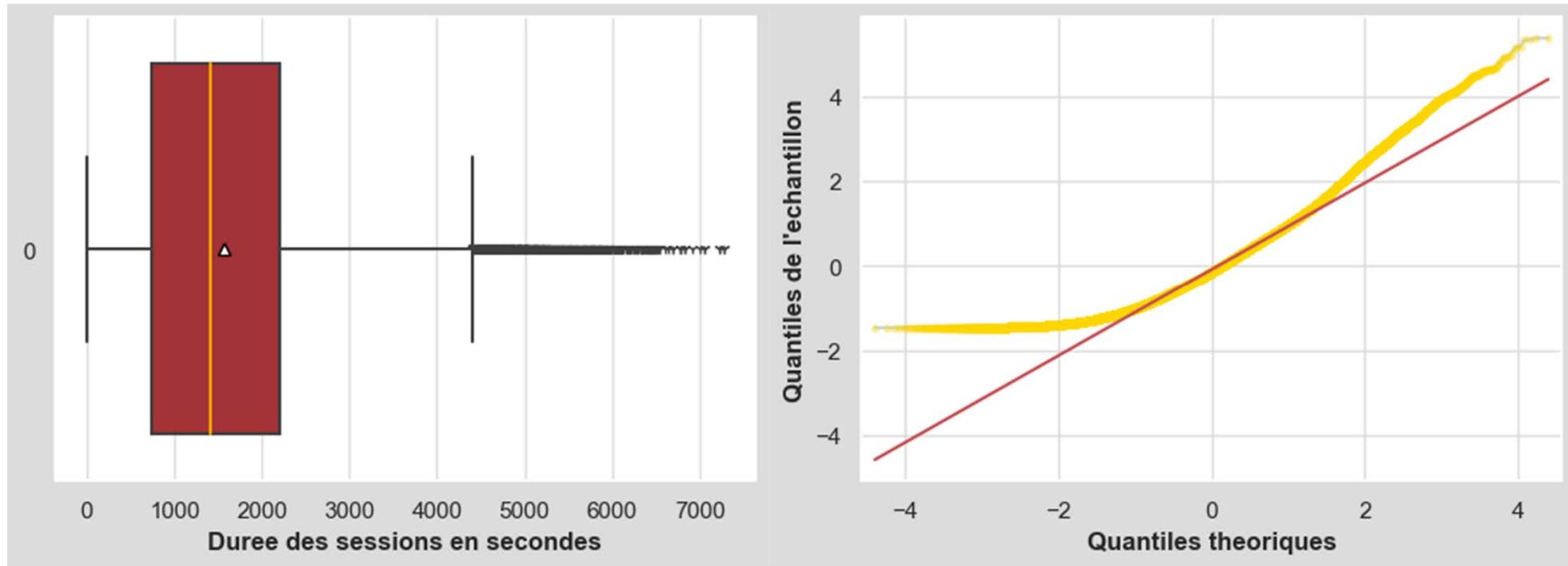


Test de Shapiro-Wilk : Statistique de test: 0.9516 / p-value: 0.0

Test d'Anderson-Darling: statistic=5.025177509684568
critical_values=array([0.573, 0.652, 0.783, 0.913, 1.086])
significance_level=array([15. , 10. , 5. , 2.5, 1.])

Dans les 2 cas, on rejette H_0 au seuil $\alpha = 5\%$. La distribution du nombre de sessions quotidiennes ne suit pas une loi Normale.

❖ durée des sessions



Test de Shapiro-Wilk
Statistique de test: 0.9475
p-value: 0.0

} On rejette H_0 au seuil $\alpha = 5\%$. La distribution de la durée des sessions ne suit pas une loi Normale.

Annexe 2 – Produits non vendus

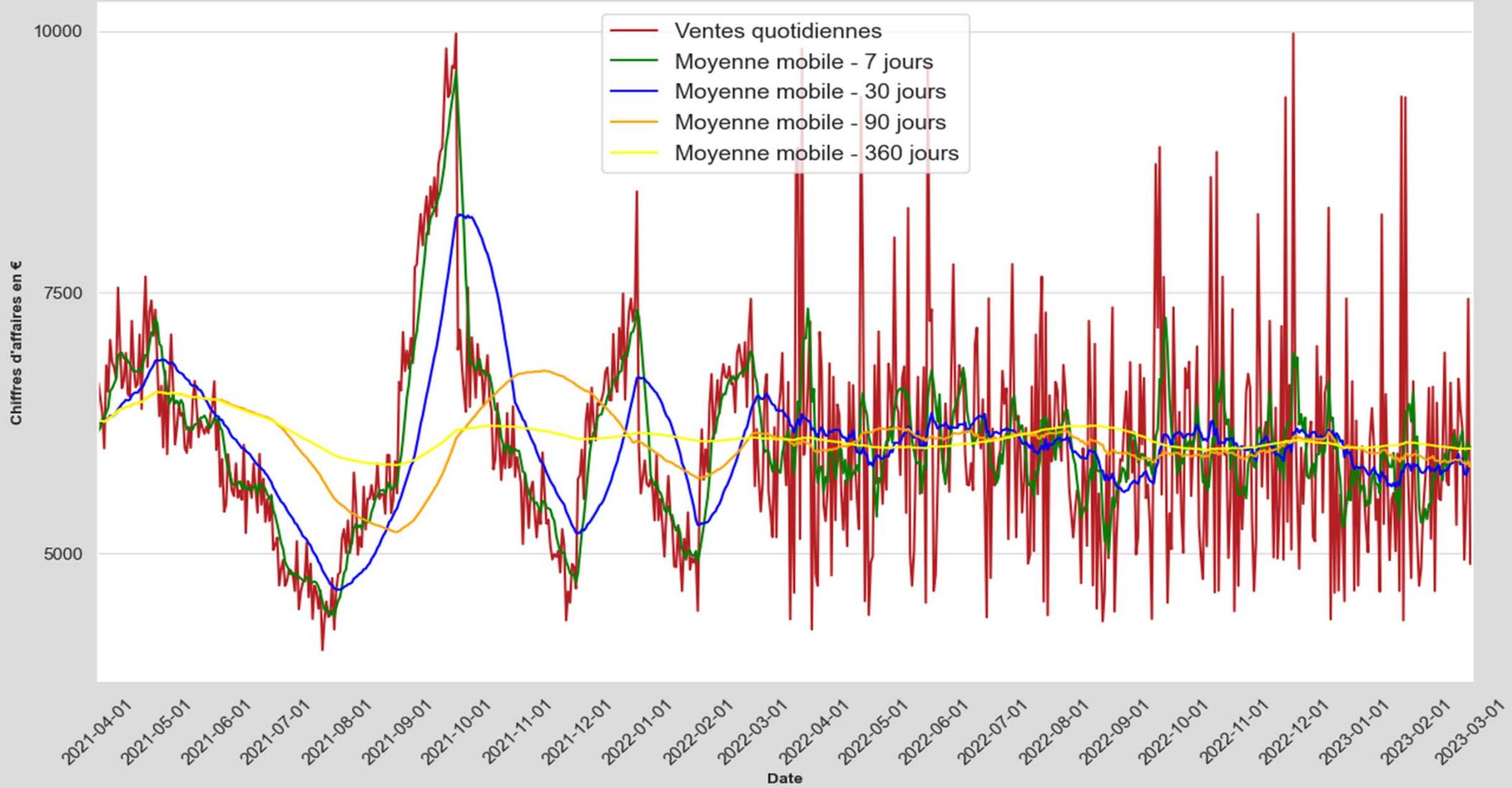
product_id_books	price_books	category_books
0_1016	35.06	0
0_1780	1.67	0
0_1062	20.08	0
0_1119	2.99	0
0_1014	1.15	0
1_0	31.82	1
0_1318	20.92	0
0_1800	22.05	0
0_1645	2.99	0
0_322	2.99	0
0_1620	0.8	0
0_1025	24.99	0
2_87	220.99	2
1_394	39.73	1
2_72	141.32	2
0_310	1.94	0
0_1624	24.5	0
2_86	132.36	2
0_299	22.99	0
0_510	23.66	0
0_2308	20.28	0

Annexe 3 – Clients non acheteurs

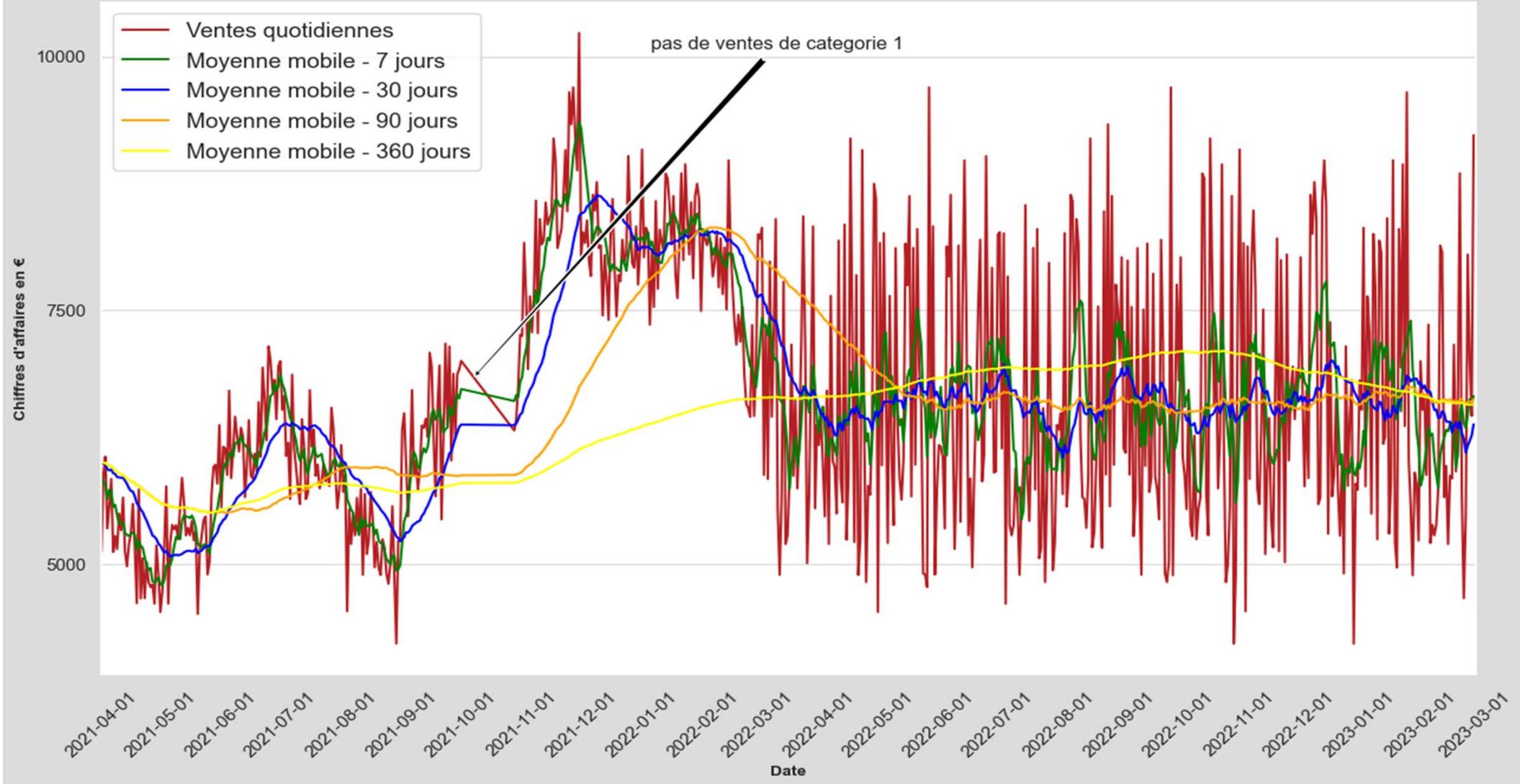
client_id_clients	gender_clients	birth_year_clients
c_8253	f	2001
c_3789	f	1997
c_4406	f	1998
c_2706	f	1967
c_3443	m	1959
c_4447	m	1956
c_3017	f	1992
c_4086	f	1992
c_6930	m	2004
c_4358	m	1999
c_8381	f	1965
c_1223	m	1963
c_6862	f	2002
c_5245	f	2004
c_5223	m	2003
c_6735	m	2004
c_862	f	1956
c_7584	f	1960
c_90	m	2001
c_587	m	1993
c_3526	m	1956

Annexe 4 - Chiffre d'affaires en moyenne glissante par catégorie de produits

Chiffre d'affaires quotidien - catégorie 0



Chiffre d'affaires quotidien - catégorie 1



Chiffre d'affaires quotidien - catégorie 2

