

L'Entreprise Lapage

Analyse des Ventes d'une Librairie

avec R & Python

Rondeau Cécile — Data Analyst

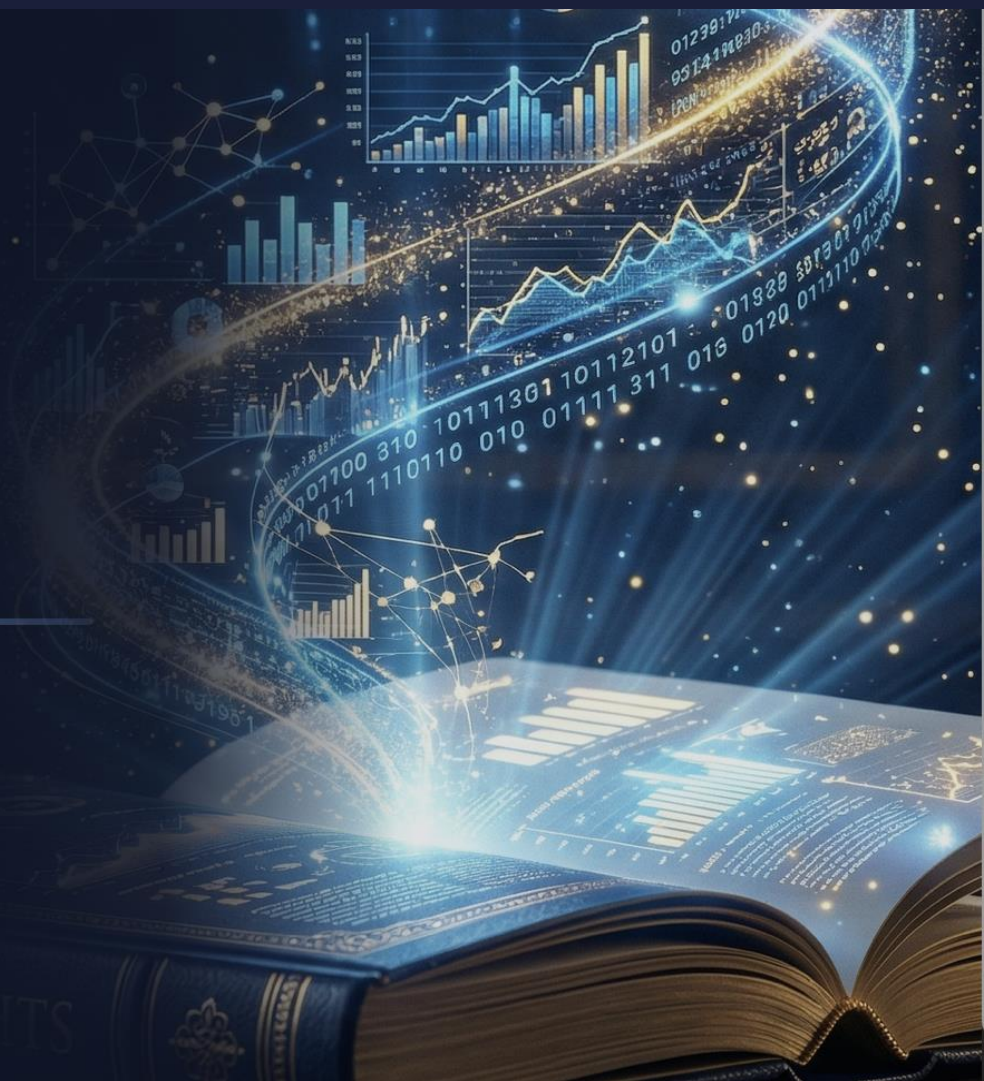
12 Janvier 2026 | Soutenance de projet d'analyse



01

Les Données

3 datasets ·



Exploration des données – Datasets

01 Clients – 8 621 Lignes

client_id	sex	birth
c_4410	f	1967
c_7839	f	1975
c_1699	f	1984
c_5961	f	1962
c_5320	m	1943

Âge moyen des clients : 44.72 ans
Âge minimum : 19
Âge maximum : 94
Médiane : 44.0

02 Livres – 3 286 Lignes

id_prod	price	categ
0_1421	19.99	0
0_1368	5.13	0
0_731	17.99	0
1_587	4.99	1
0_1507	3.99	0

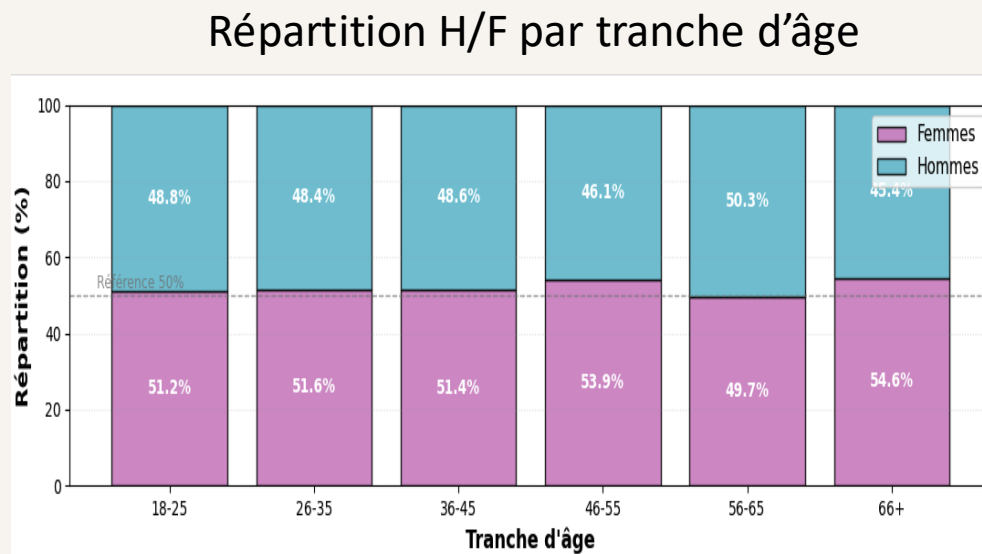
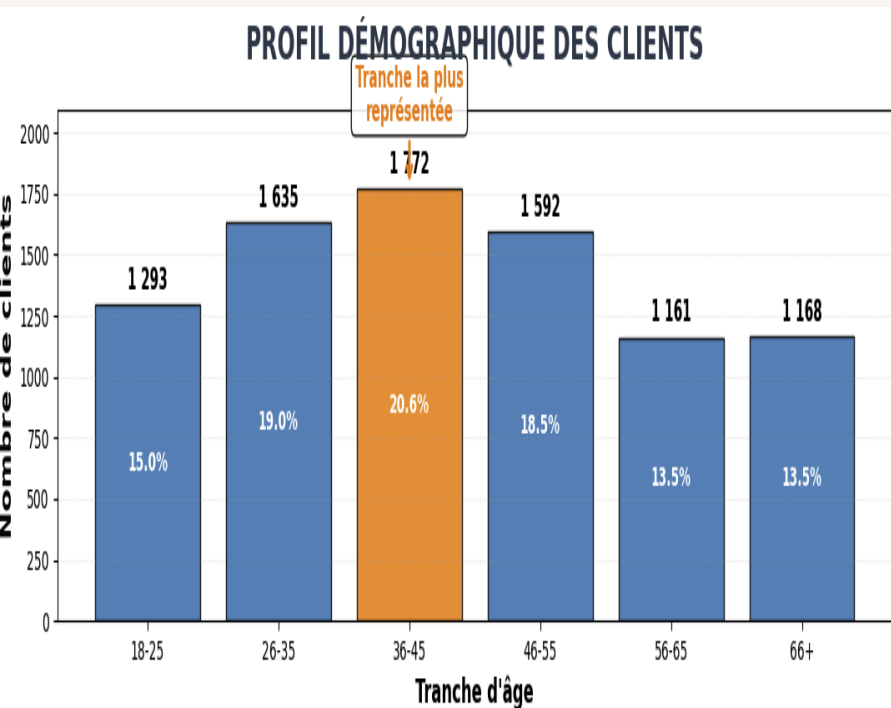
Répartition des catégories :
categ
0 2308
1 739
2 239

03 Transactions – 687 534 Lignes

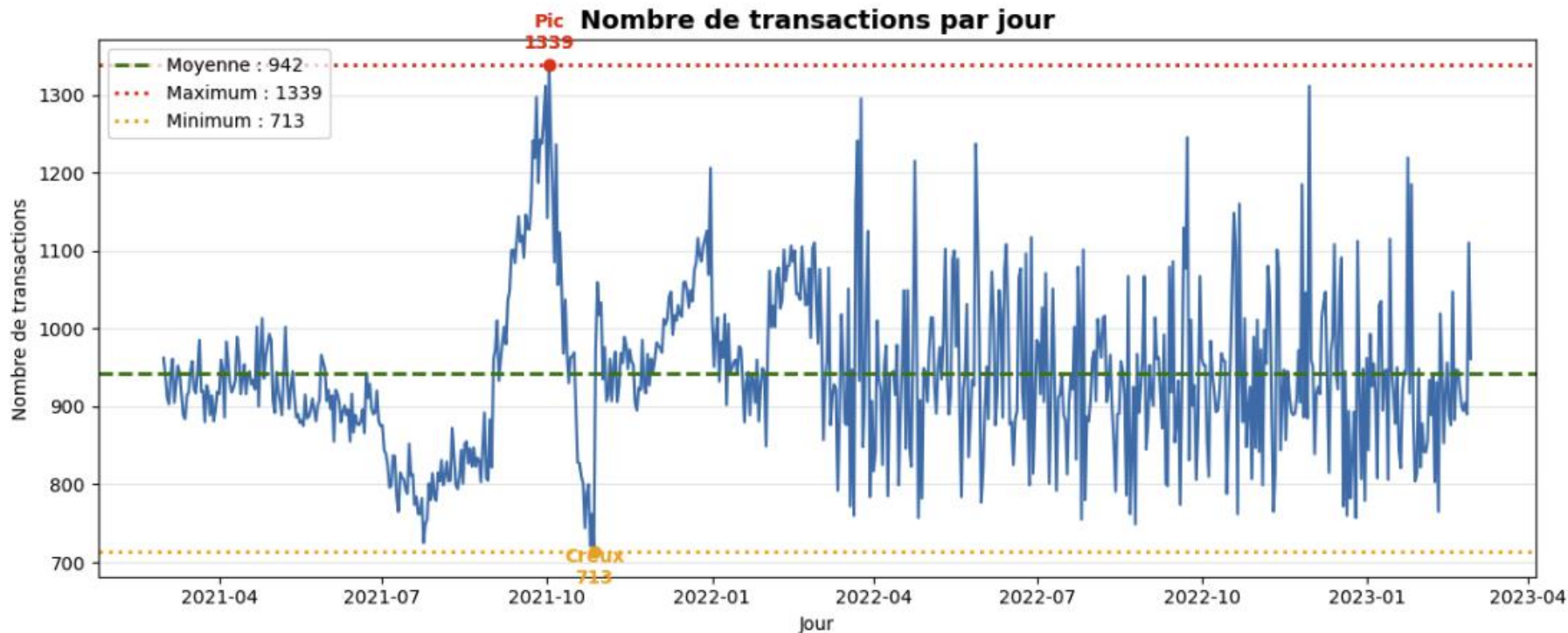
id_prod	date	session_id	client_id
0_1259	2021-03-01 00:01:07.843138	s_1	c_329
0_1390	2021-03-01 00:02:26.047414	s_2	c_664
0_1352	2021-03-01 00:02:38.311413	s_3	c_580
0_1458	2021-03-01 00:04:54.559692	s_4	c_7912
0_1358	2021-03-01 00:05:18.801198	s_5	c_2033

Dates non convertibles : 0
Min date : 2021-03-01 00:01:07.843138
Max date : 2023-02-28 23:58:30.792755

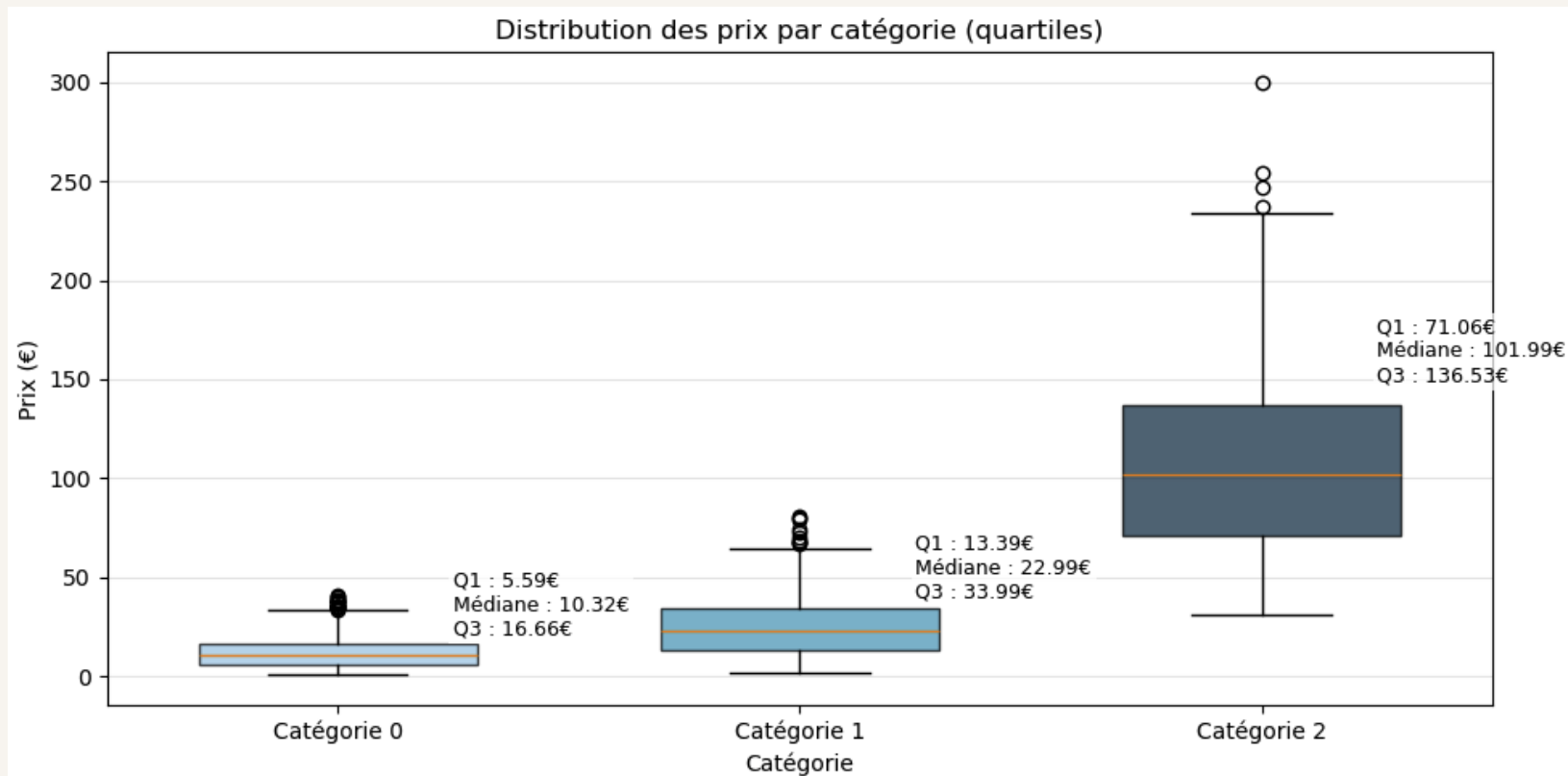
Analyse démographique de nos clients



Activité transactionnelle journalière sur la période 2021 – 2023

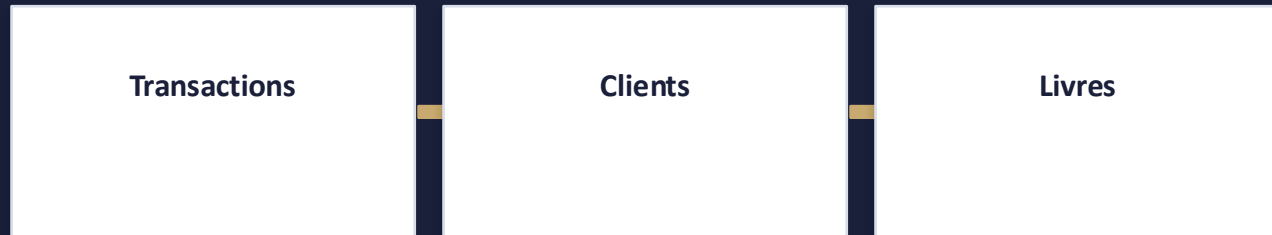


Prix des livres par categories



Construction du Master DataFrame

Pipeline de jonction



687 534

Transactions

8 600

Clients actifs

Ont passé ≥ 1 commande

3 265

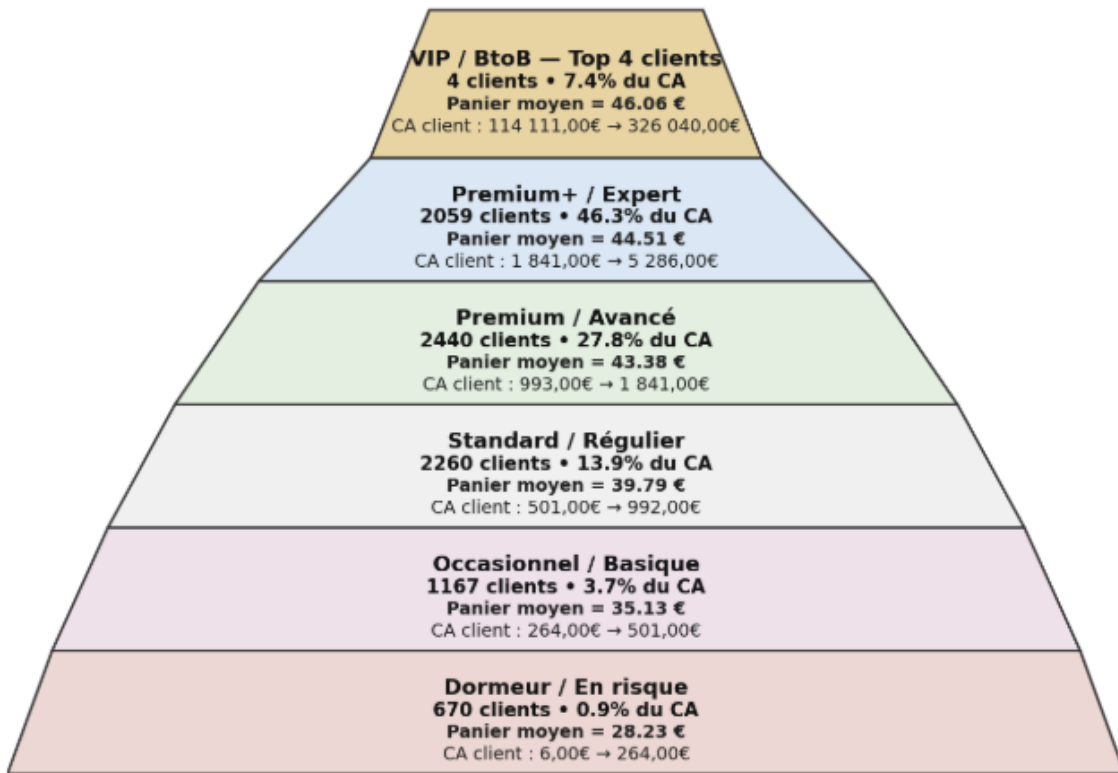
Produits vendus

Sur 3 286 référencés

Segmentation clients par contribution au CA

Segmentation clients — pyramide de contribution (CA)

VIP = Top 4 clients • Segments suivants = CA cumulé (statistique)



Statistiques globales

Total clients : 8600

CA total : 12 027 663,00€

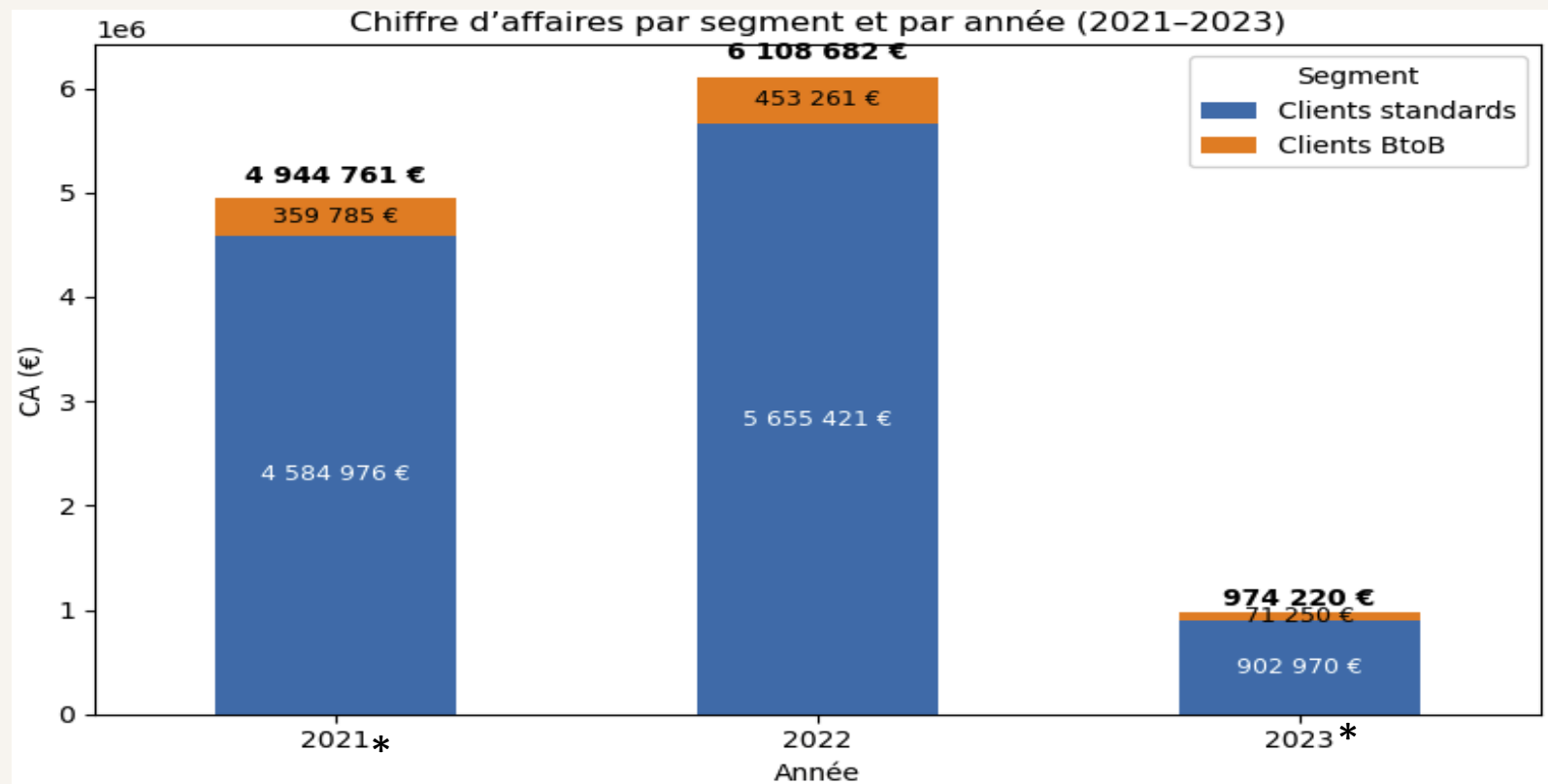
Panier moyen global : 40.41 €

Détails des 4 clients VIP

(Top 4 clients BtoB par contribution CA)

#1	c_1609	CA : 326 040,00€	Panier moyen : 29.65 €
		Part du CA total : 2.7%	Transactions : 10997
#2	c_4958	CA : 290 227,00€	Panier moyen : 75.36 €
		Part du CA total : 2.4%	Transactions : 3851
#3	c_6714	CA : 153 919,00€	Panier moyen : 58.75 €
		Part du CA total : 1.3%	Transactions : 2620
#4	c_3454	CA : 114 111,00€	Panier moyen : 20.48 €
		Part du CA total : 0.9%	Transactions : 5571

CA annuel par segment sur la période 2021–2023



2021* et 2023* - données partielles — 2021 : mars–déc. | 2023 : janv.–fév.

02

Analyses Descriptives

CA · Produits · Clients · Saisonnalité – Hors 4 clients VIP *BtoB*

Sur la période de mars 2021 à février 2023

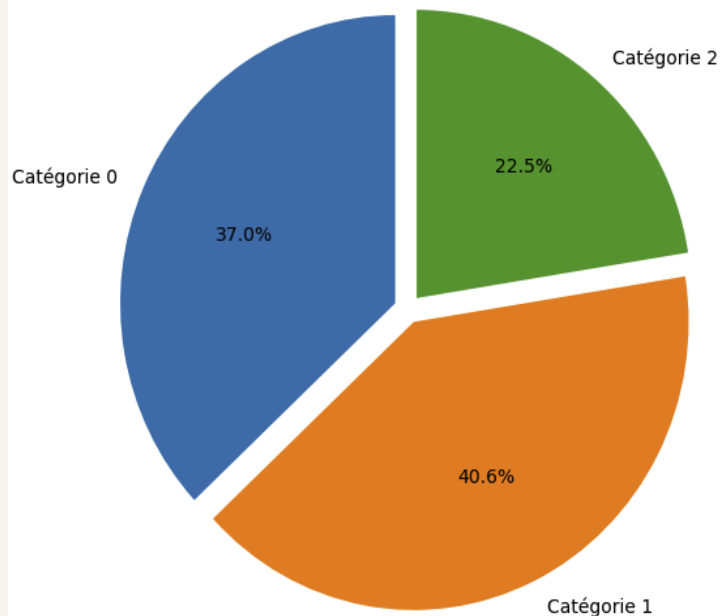
CA par catégorie de livres – clients standards-

Section 02 — Analyses descriptives

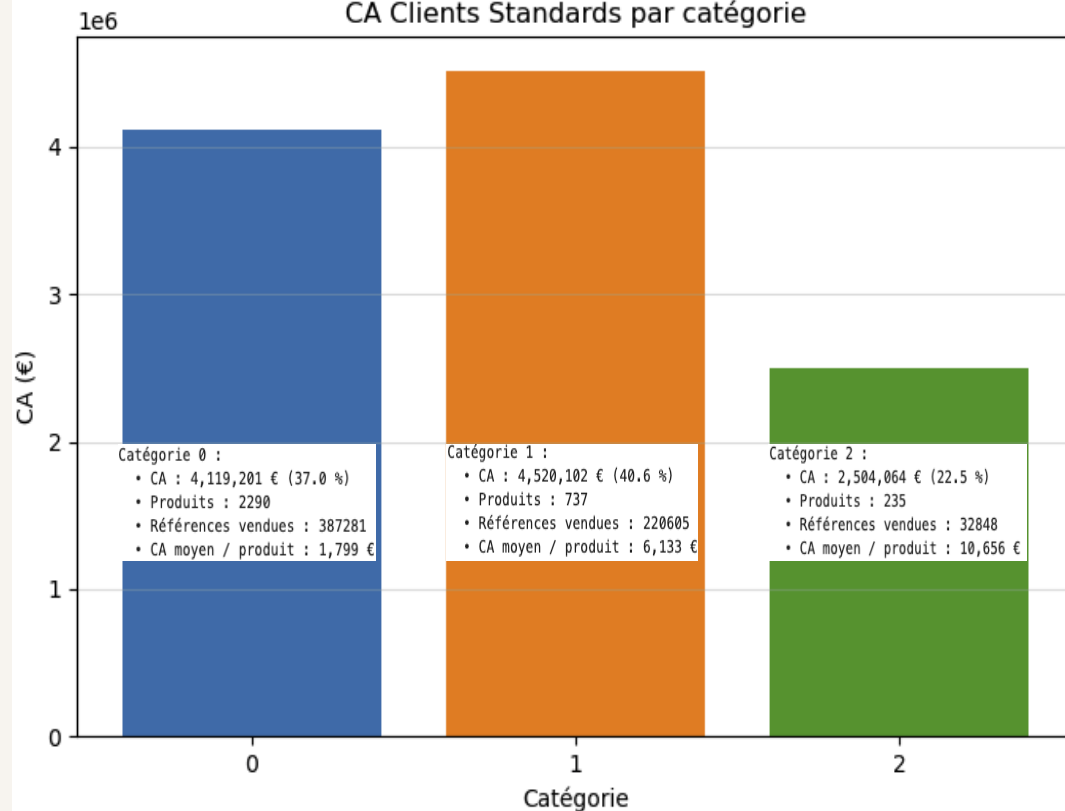
Chiffre d'affaires total- Clients standards

11 143 367 €

Répartition du CA par catégorie - Clients standards

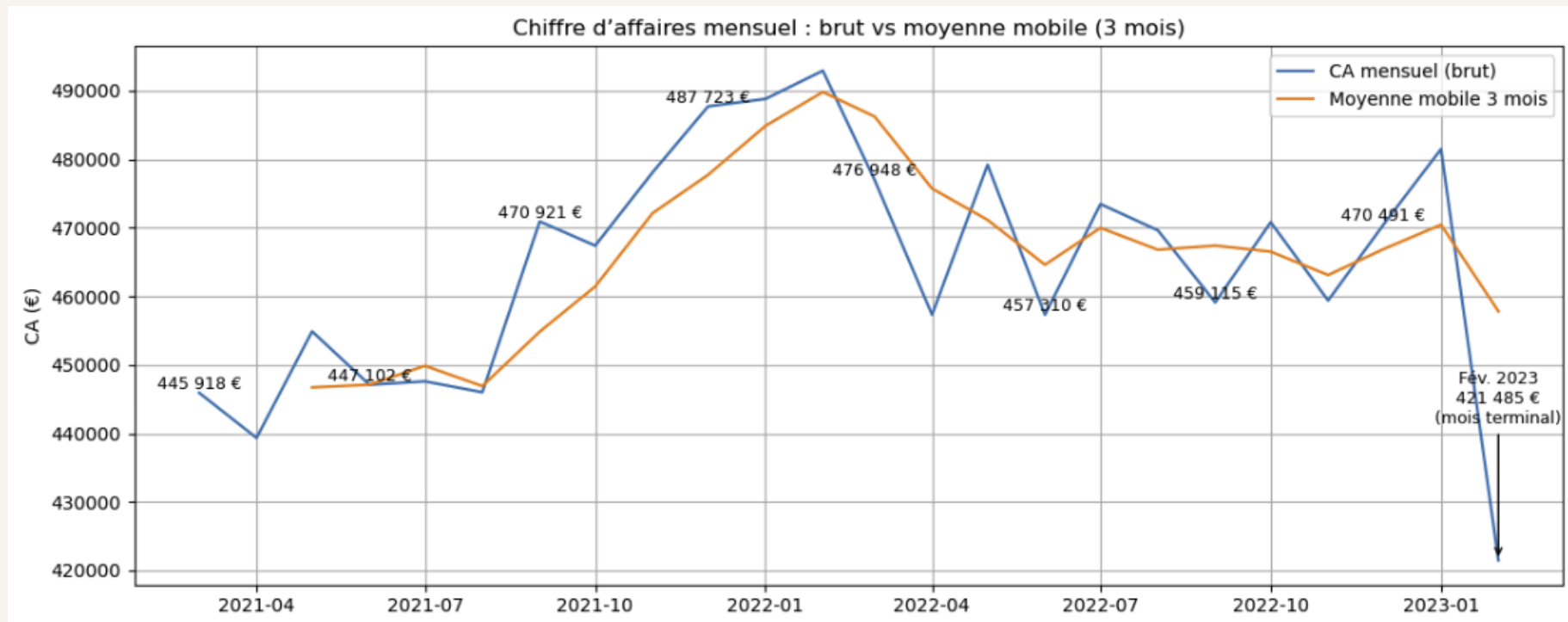


CA Clients Standards par catégorie



CA mensuel & tendance MM3 (2021 -2023)

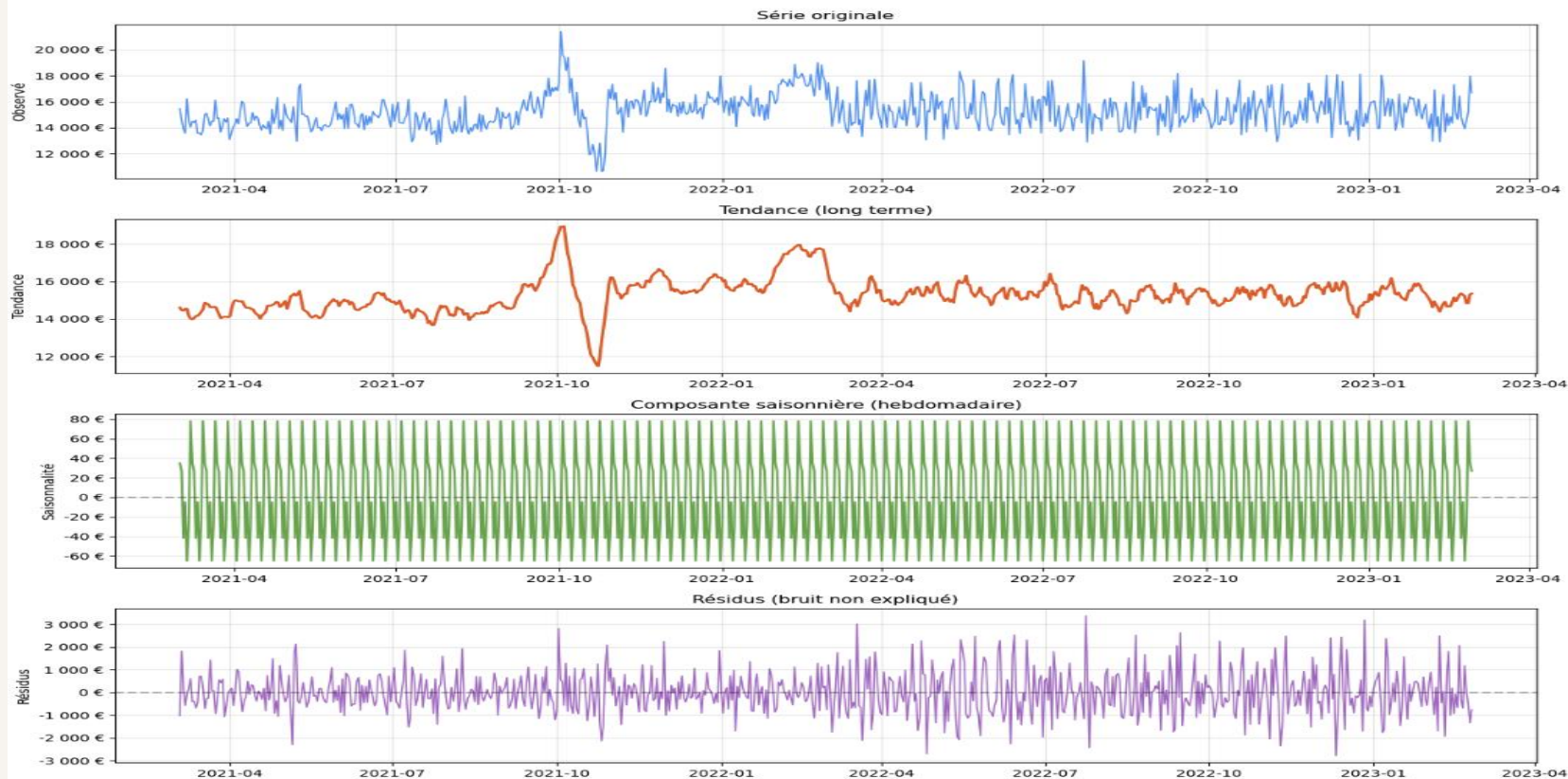
Section 02 — Analyses descriptives



Décomposition du CA journalier

Section 02 — Analyses descriptives

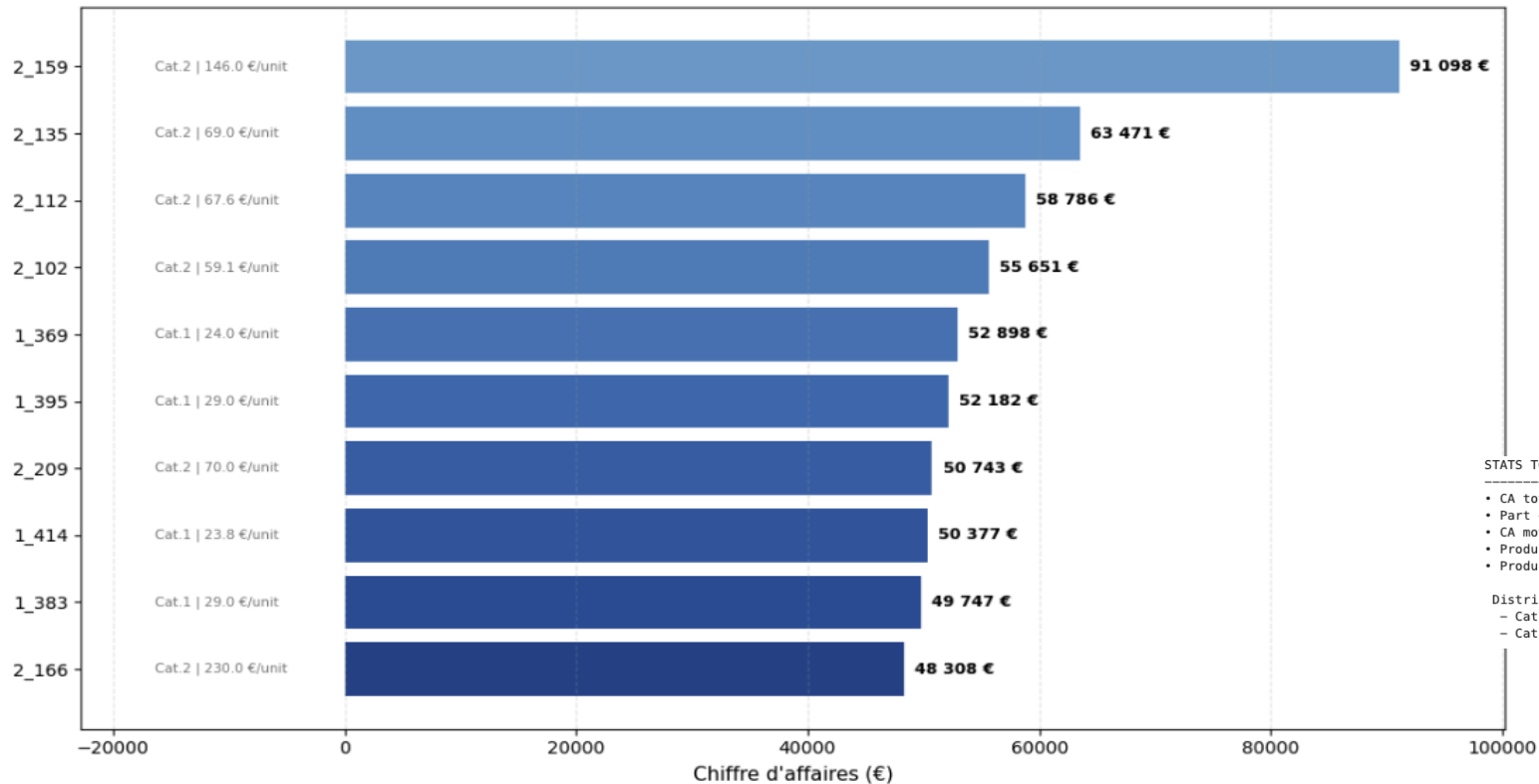
Décomposition du CA Journalier — Cycle hebdomadaire



TOP 10 produits par contribution au CA

Section 02 — Analyses descriptives

TOP 10 - Produits générant le plus de CA



STATS TOP 10 CA GLOBAL :

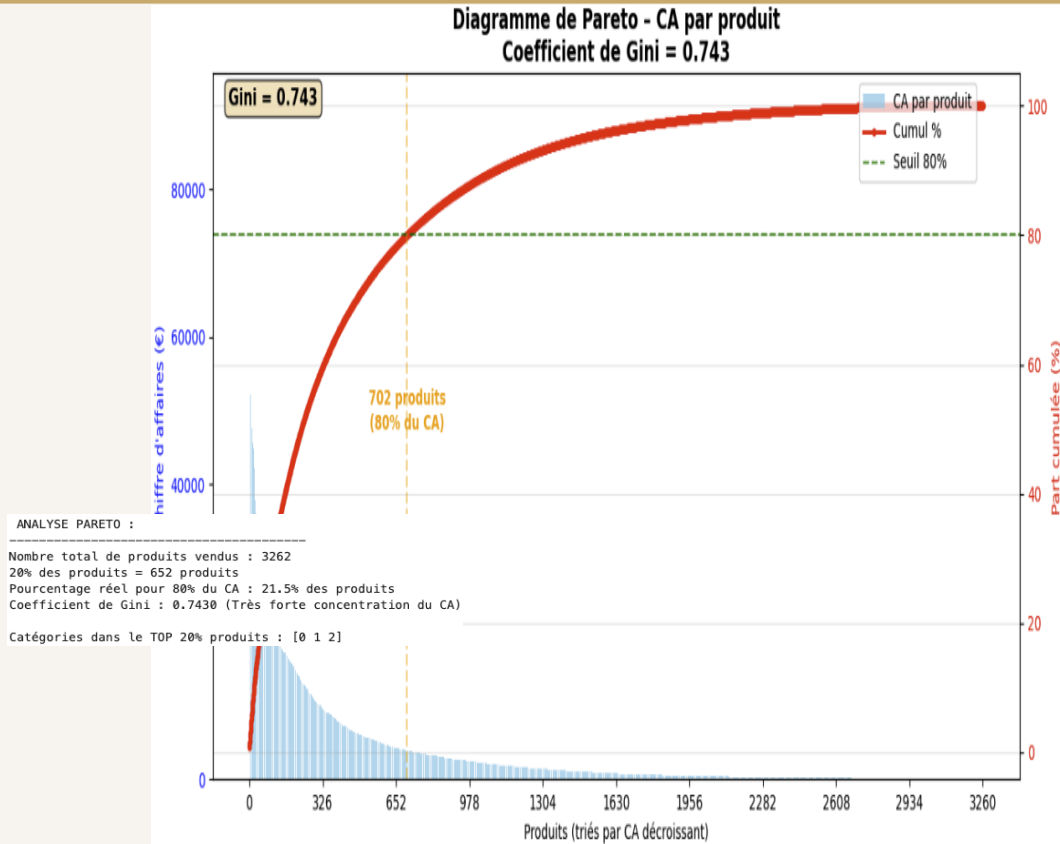
- CA total TOP 10 : 573 260 €
- Part du TOP 10 : 5.1% du CA total
- CA moyen/produit : 57 326 €
- Produit leader : 2_159 (91 098 €)
- Produit 10ème : 2_166 (48 308 €)

Distribution par catégorie dans le TOP 10 :

- Cat.2 : 6 produit(s)
- Cat.1 : 4 produit(s)

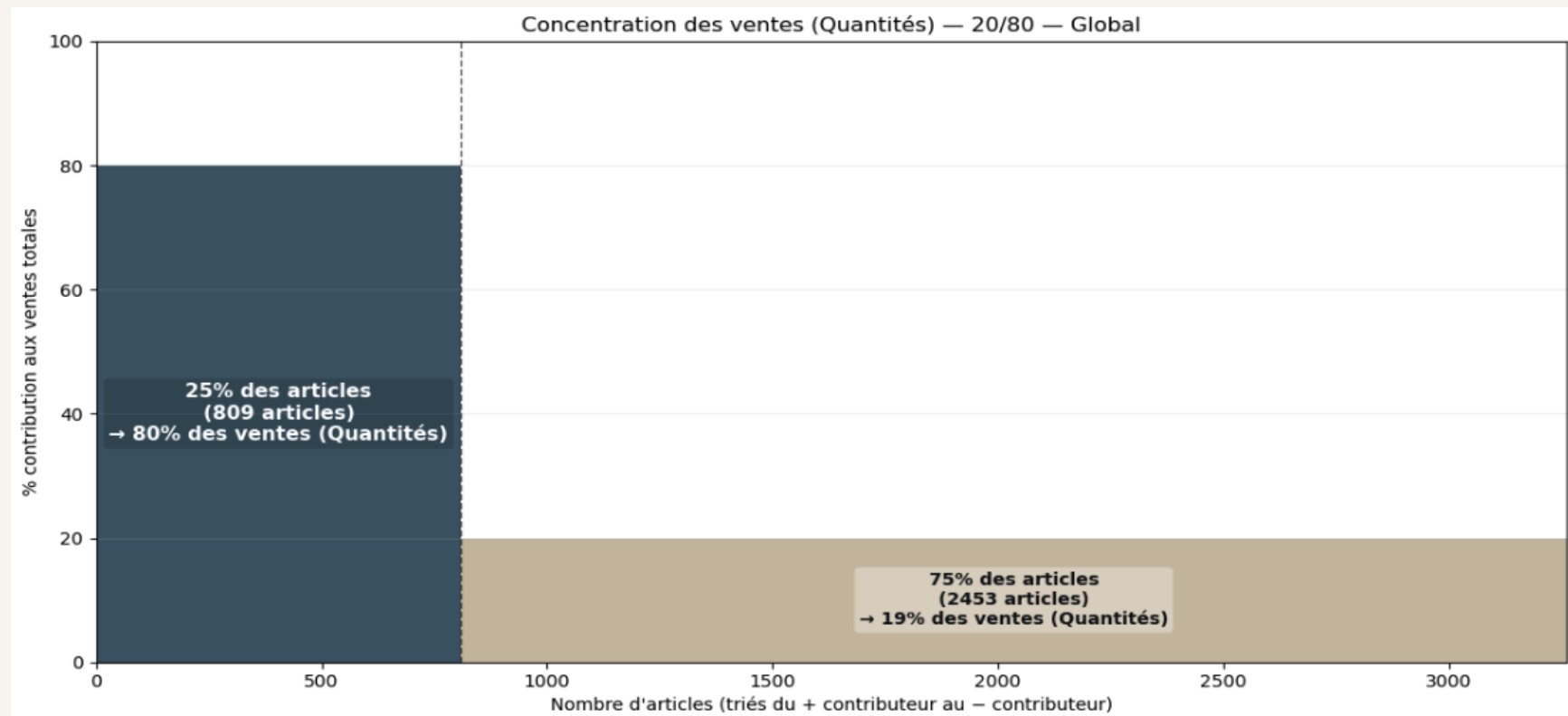
Concentration du CA par produit - Diagramme de Pareto

Section 02 — Analyses descriptives



Concentration des ventes par volume – 20/80

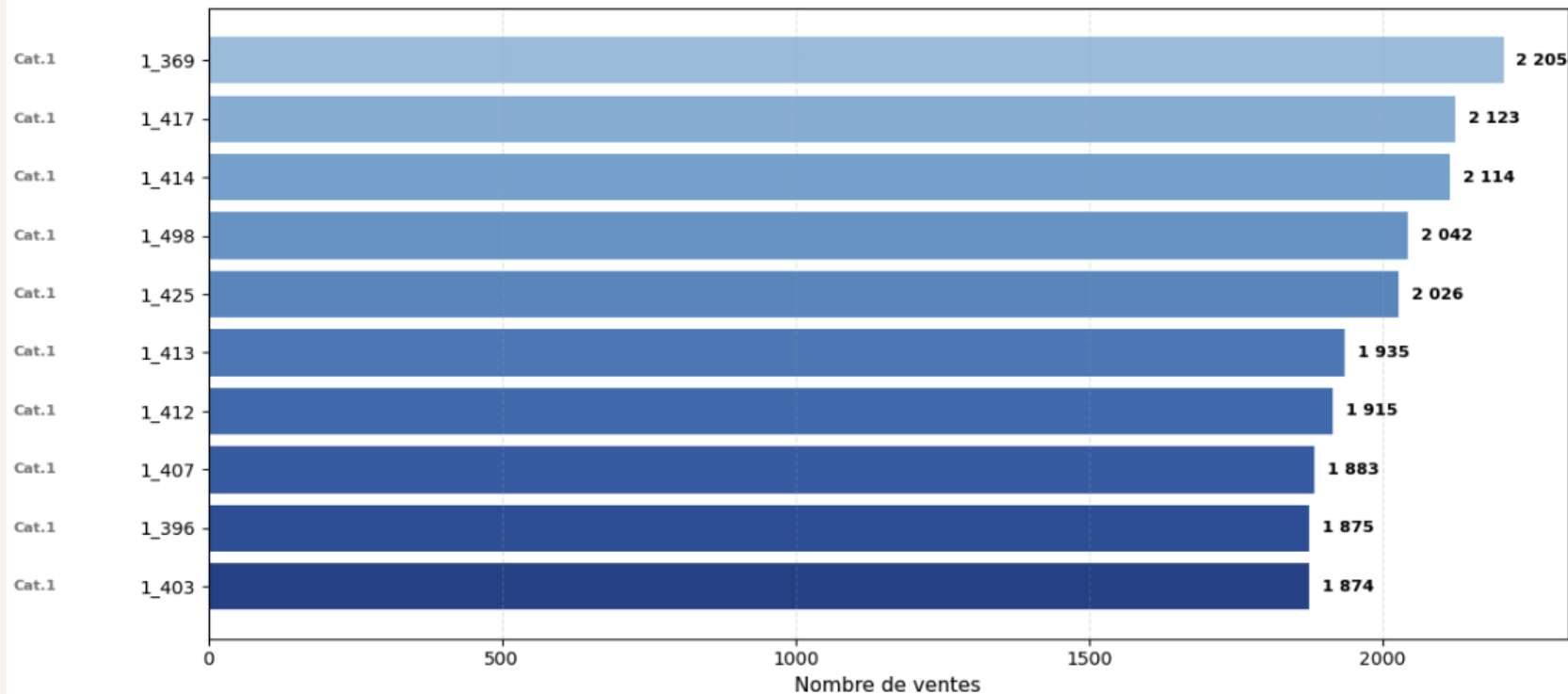
Section 02 — Analyses descriptives



Volume des ventes: Top 10

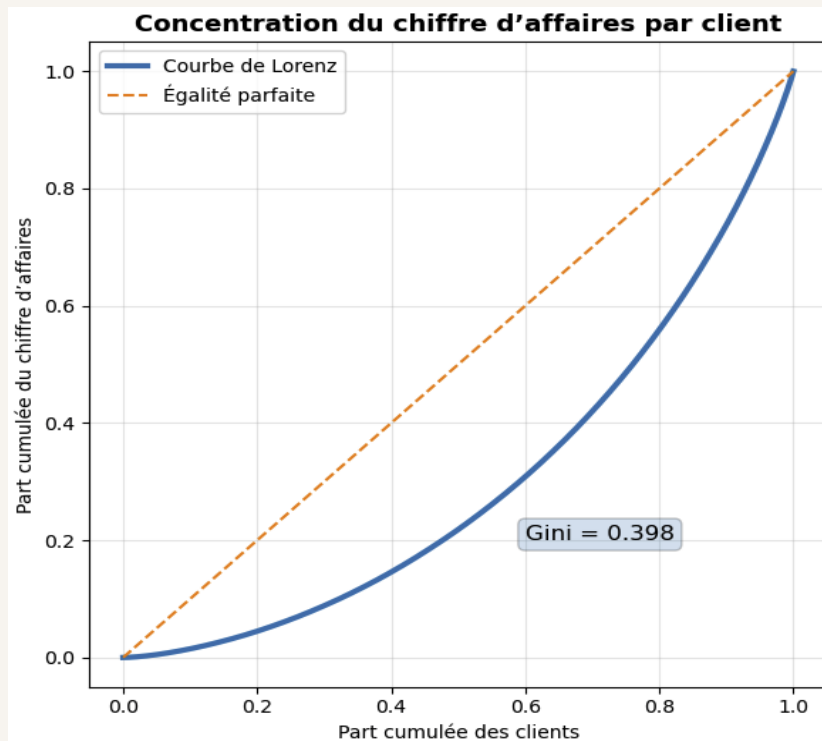
Section 02 — Analyses descriptives

TOP 10 - Produits les plus vendus (volume)



Concentration du CA par client — Courbe de Lorenz

Section 02 — Analyses descriptives



03

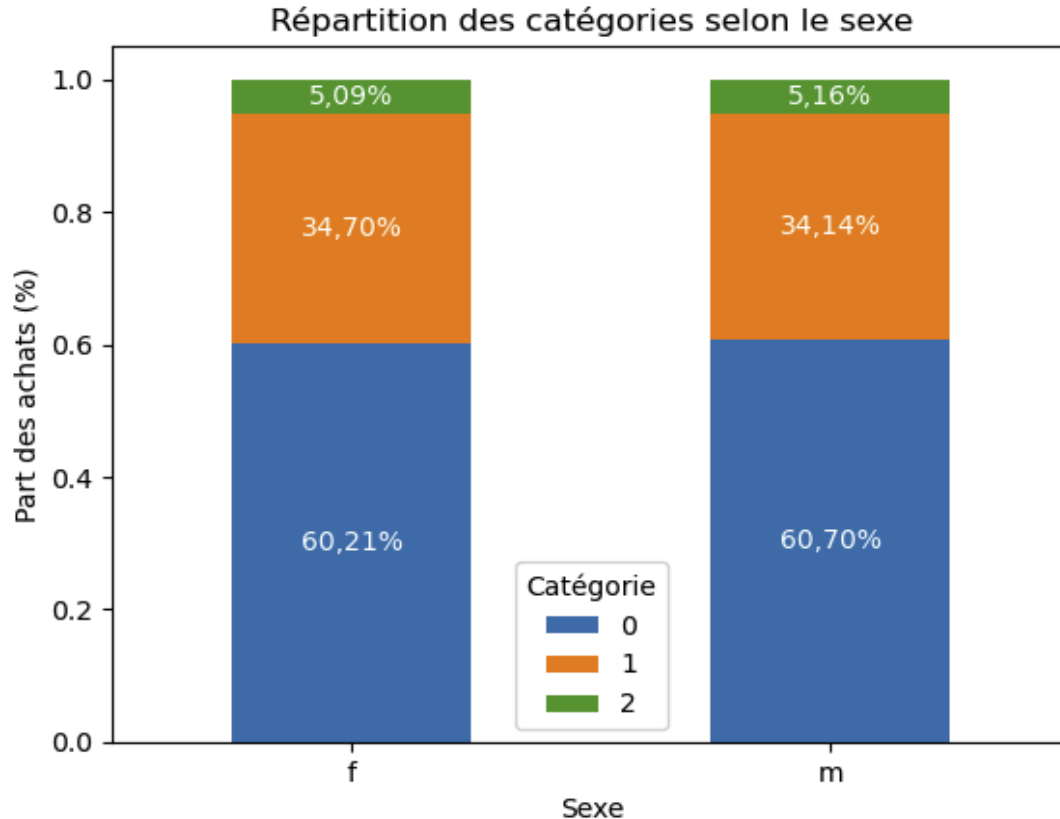
Analyses Statistiques

5 tests statistiques

Chi² · Kruskal-Wallis · Spearman · Mann-Whitney · ANOVA — test choisi · résultat · interprétation



Analyse 1 — Le Genre du client et les catégories achetées



Test Chi²

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

2 Variables Qualitative: Genre & Catégories

Grand échantillon

Plusieurs catégories

Fréquence ≥ 5 dans chaque case

Tableau de contingence 2x3

Statistique :

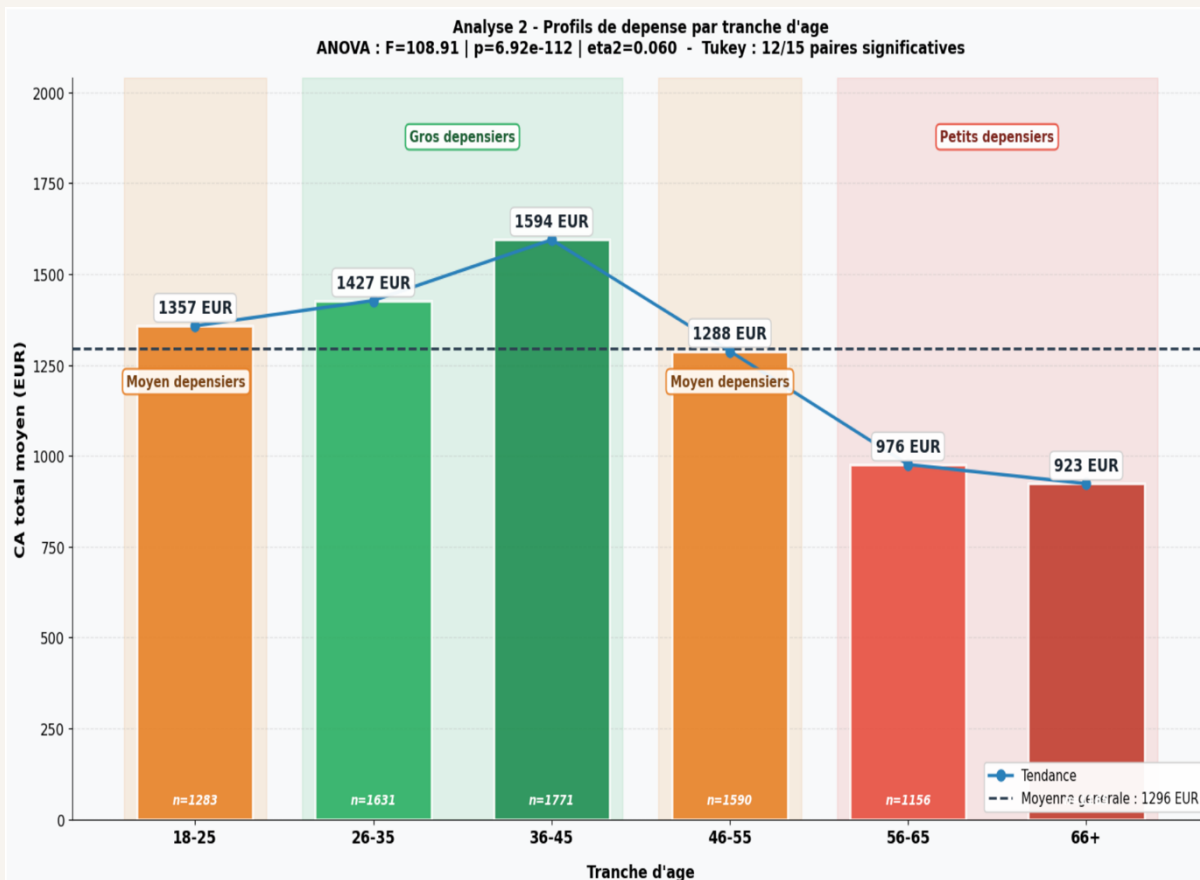
$\chi^2 : 22,67$ | DDI = 2 | $p = 0,000012$ |

V de Cramer = 0,01

Résultat $P < 0,05$ il y a un lien | 0,01= très Faible

Dépendance très faible entre le genre et la catégorie: impact négligeable

Analyse 2 — Le lien entre l'âge des clients et le montant total des achats



Test Anova

Données Mixtes

3 groupes ou plus

Groupes indépendants

Non Normale (p Shapiro $< 0,05$) détectée mais
Anova validée TCL $n > 1000$ par groupe (élevé)

Statistique :

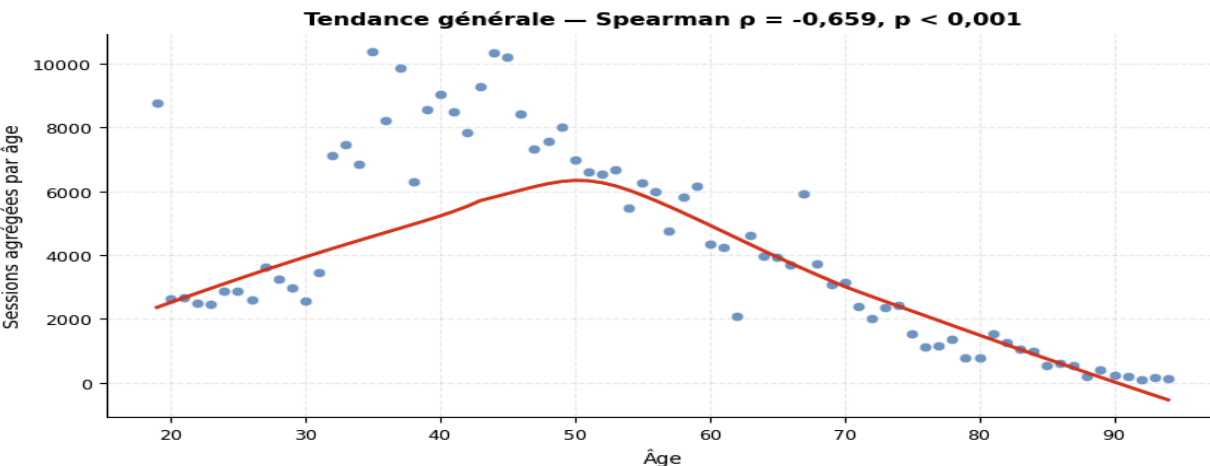
Anova $F: 108,91$ | p value : $6,92e112$ |
 $n^2 : 0,060$

Résultat: significative $p < 0,05$; Taille d'effet
modérée $n^2 = 0,06$; Test Post-hoc (Tukey) 12
sur 15 paires significatives

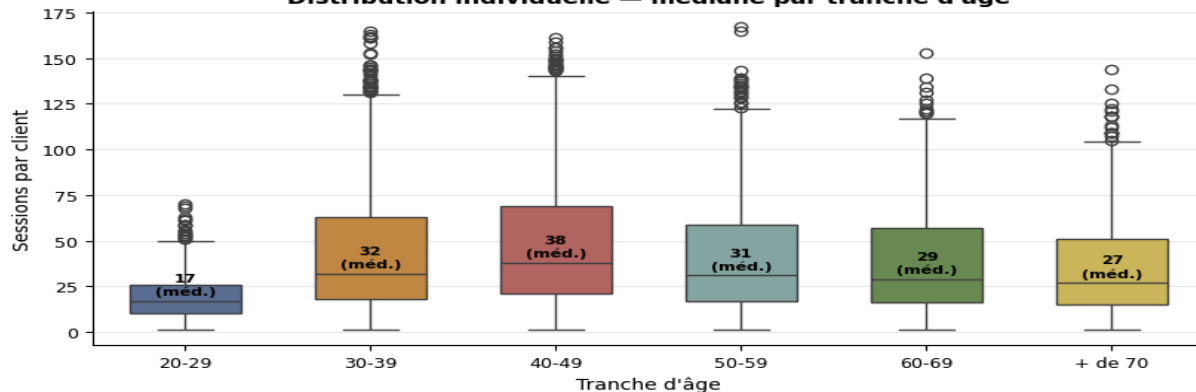
Il existe une différence significative entre les
tranches d'âge; l'âge explique environ 6% des
dépenses

Analyse 3 — Le lien entre l'âge des clients et la fréquence d'achats

Lien entre l'âge des clients et la fréquence d'achat



Distribution individuelle — médiane par tranche d'âge



Test Spearman

2 Variables Quantitative: Age & Fréquence d'achat (nb de Session par client)

Données NON NORMALES ($p < 0,05$)

Relation monotone globale mais non linéaire

Mesure Tendance Générale

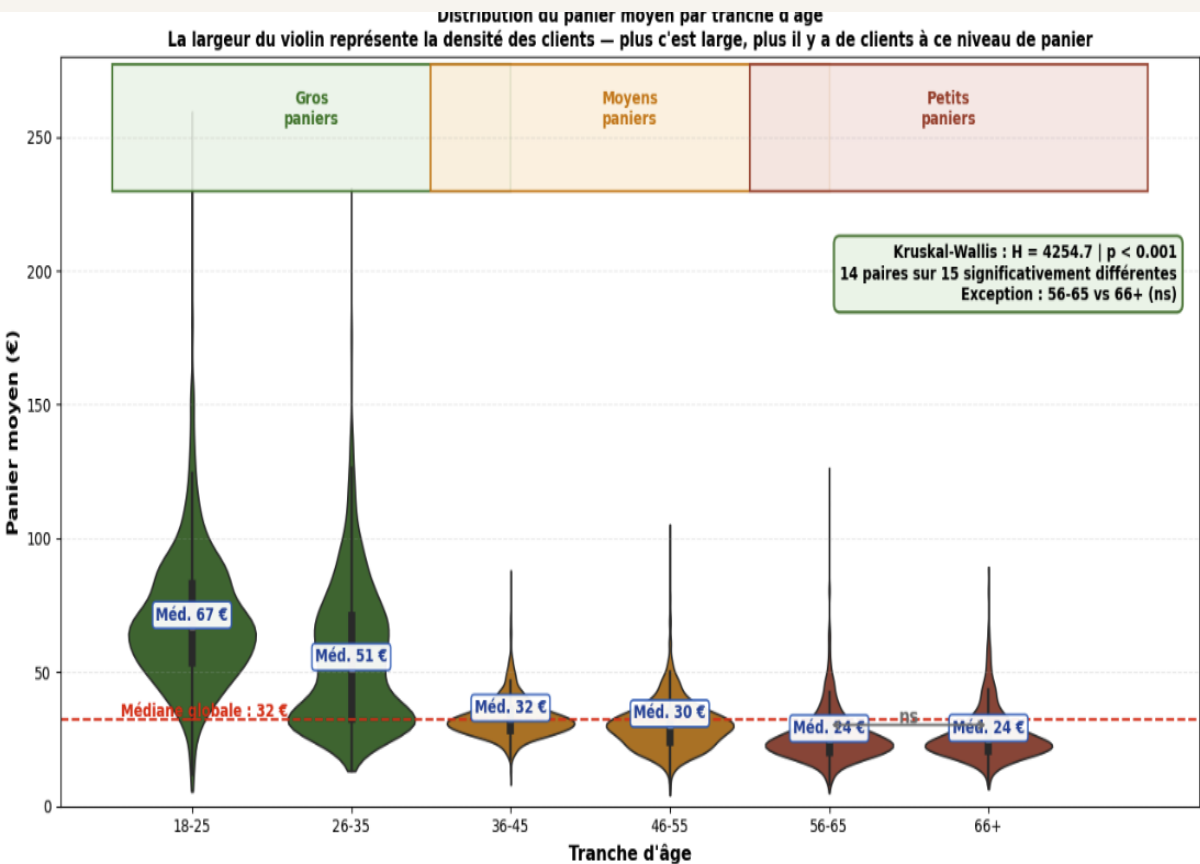
Statistique :

Spearman Rho: $-0,659$ | $p : 9.36e11$ | $n : 76$ (Valeur agrégées par âges)

Résultat : $p < 0,05$ corrélation significative ;
 $\rho = -0,659$ corrélation négative forte

Les clients plus âgés achètent moins fréquemment mais relation forte car effet fort

Analyse 4 — Le lien entre l'âge des clients et la taille du panier moyen



Test Kruskal-Wallis

Données Mixtes

Données Non Normale (p Shapiro $< 0,05$)

6 Groupes indépendants

Variable dépendante continue (panier moyen €)

Comparaison des médianes

Statistique :

Kruskal Wallis $H : 4254,7$ | $p < 0,01$

Résultat: $P < 0,01$ très petit ; $H : 4254,7$ très élevé ; Différence significative entre les groupes écart important entre les tranches d'âge

Post-Hoc: 14/15 paires significativement différentes (56_65 vs 66+ non significatif)

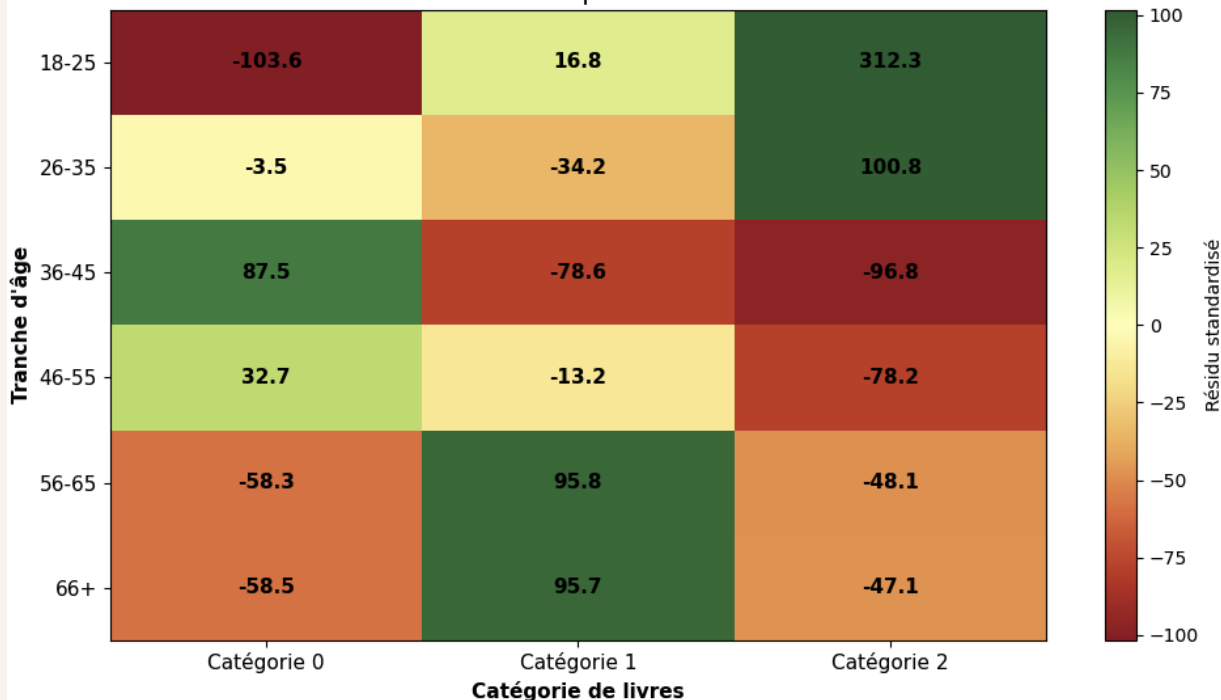
Le panier moyen diminue significativement avec l'âge

Analyse 5 — Le lien entre l'âge des clients et la catégorie des livres achetées

LECTURE DE LA HEATMAP :

- Vert : combinaison sur-représentée (plus d'achats que prévu)
- Rouge : combinaison sous-représentée (moins d'achats que prévu)
- Seuil : résidu > 2 ou < -2 = écart significatif
- V de Cramer = 0.3749 → association modérée entre âge et catégorie

Heatmap des résidus standardisés — Âge vs Catégorie
V de Cramer = 0.3749 | association modérée



$$\text{Test } \chi^2 \quad \chi^2 = \sum \frac{(O - E)^2}{E}$$

2 Variables Qualitative: Âge & Catégories

Grand échantillon

Plusieurs catégories

Fréquence >= 5 dans chaque case

Tableau de contingence 6x3

Statistique :

$\chi^2 : 180123,73$ | DDI = 10 | p value =
<0,001 | V de Cramer = 0,3749

Résultat P< 0,01 il y a un lien significatif |
0,3749 = association modérée

L'âge est associé à la la catégorie achetée

04

Conclusions



Merci pour votre attention.

Questions ?

