# Projet 9

## Analysez les ventes d'une librairie avec R

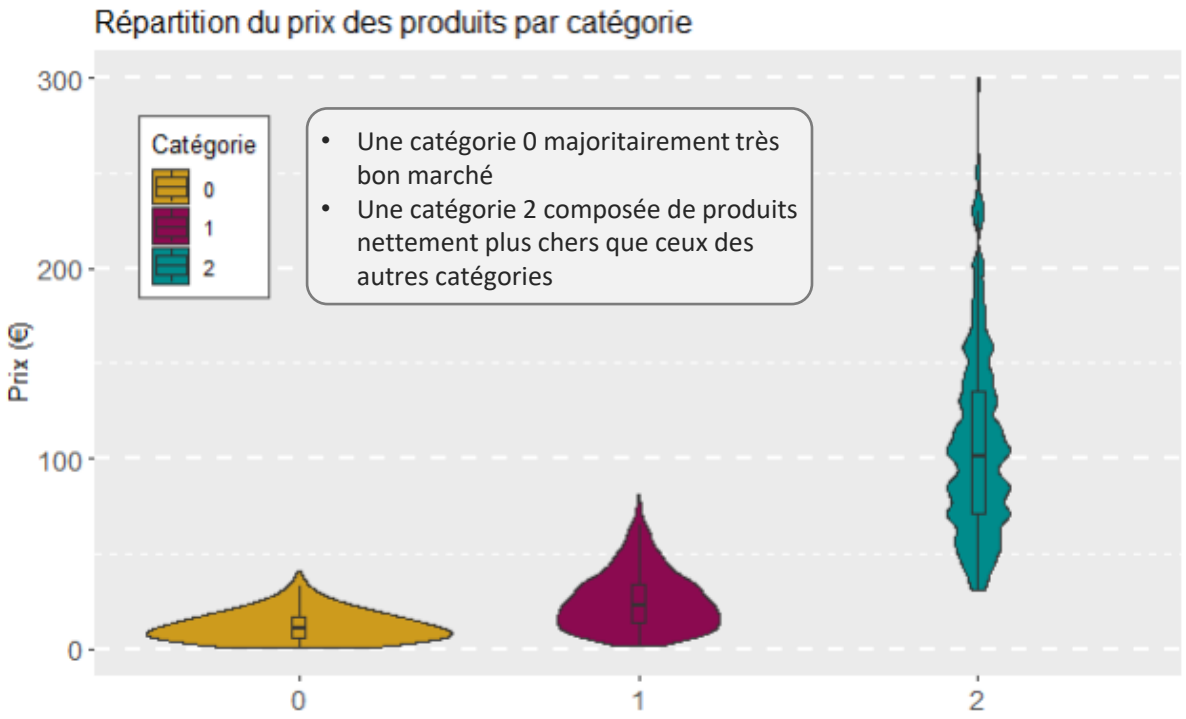
**Marie G.**

**Parcours Data Analyst - 19/02/2025**

1. Typologie des produits, ventes et clients
2. Une légère progression des ventes marquée par des perturbations
3. Des corrélations marquées entre l'âge des clients et la typologie de leurs achats

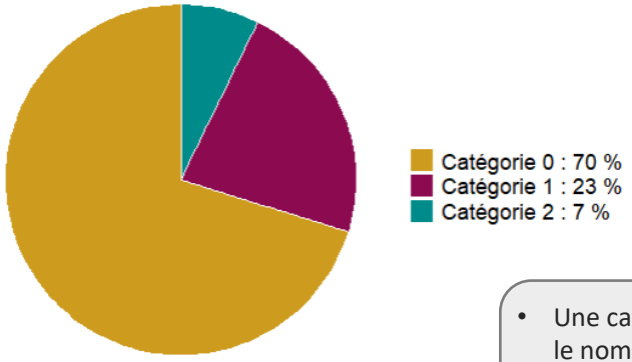
# 1. Typologie des produits, ventes et clients

## 1.1 Des catégories marquées par leur répartition des prix

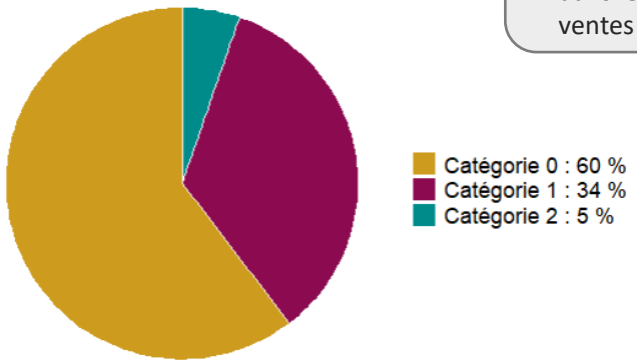



- Une catégorie 0 majoritairement très bon marché
- Une catégorie 2 composée de produits nettement plus chers que ceux des autres catégories

**Répartition par catégorie des produits du catalogue**




**Répartition des ventes par catégorie**



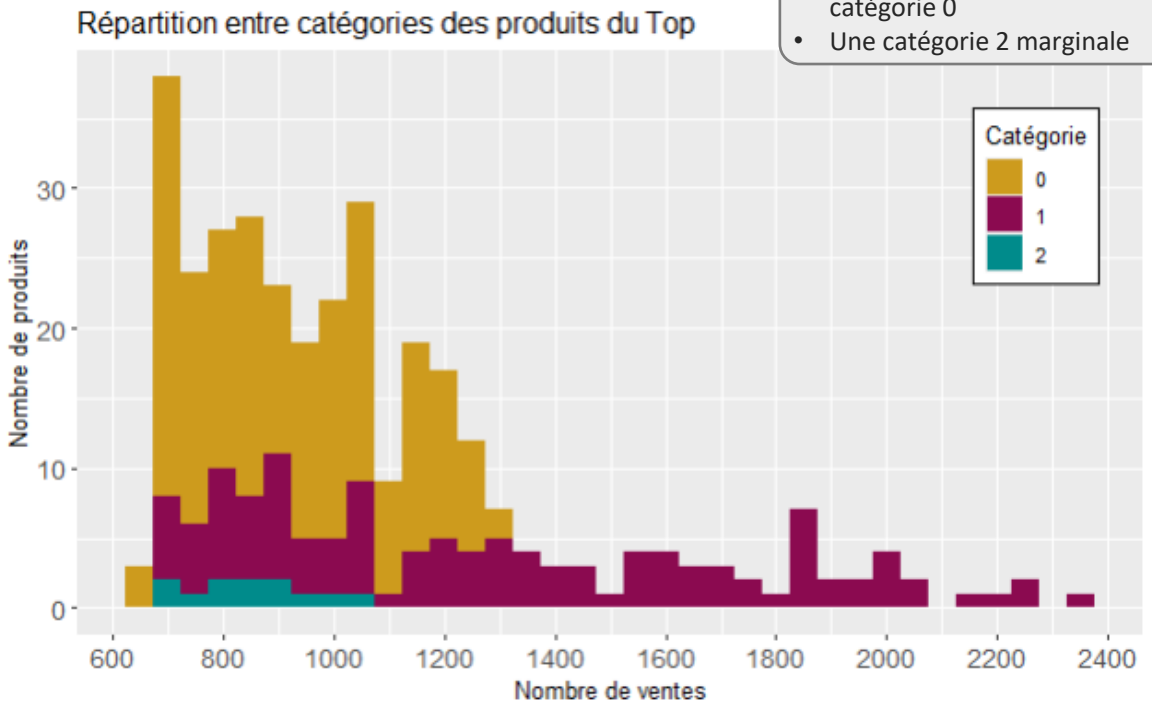

- Une catégorie 0 sur-représentée dans le nombre de références comme dans le nombre de ventes
- Une catégorie 1 mieux vendue en proportion du nombre de références
- Une catégorie 2 sous-représentée dans le catalogue comme dans les ventes

# 1. Typologie des produits, ventes et clients

## 1.2 Une catégorie 1 très présente dans le Top des ventes

On définit ici le Top des ventes par le 10ème décile des produits en nombre de ventes sur les 2 ans

- Un top monopolisé par la catégorie 1 à partir de 1300 ventes
- Une bonne représentation de la catégorie 0
- Une catégorie 2 marginale



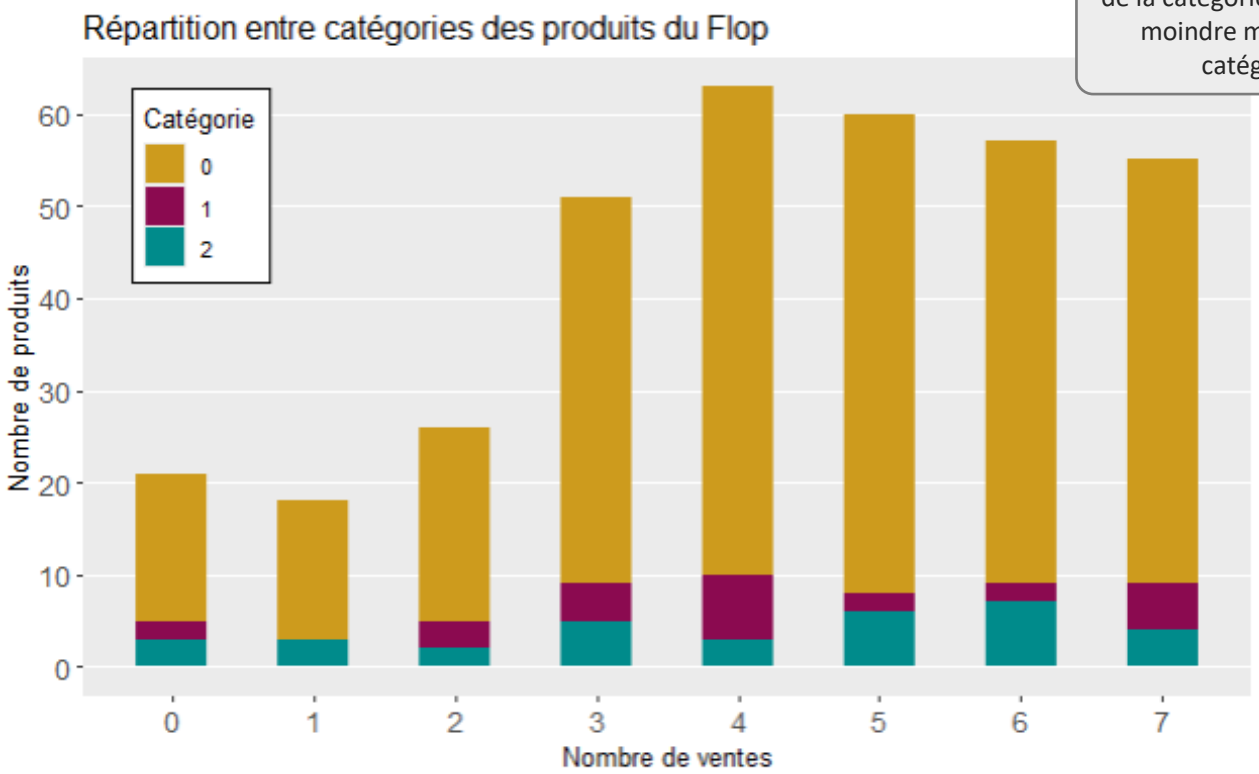

**Top 10 des produits en nombre de ventes sur 2 ans**

| | Nombre de ventes | Id produit | Prix (€) | Catégorie | Chiffre d'affaires (€) |
|---|---|---|---|---|---|
| 1 | 2 340 | 1_369 | 23,99 | 1 | 56 137 |
| 2 | 2 269 | 1_417 | 20,99 | 1 | 47 626 |
| 3 | 2 246 | 1_414 | 23,83 | 1 | 53 522 |
| 4 | 2 202 | 1_498 | 23,37 | 1 | 51 461 |
| 5 | 2 163 | 1_425 | 16,99 | 1 | 36 749 |
| 6 | 2 040 | 1_403 | 17,99 | 1 | 36 700 |
| 7 | 2 036 | 1_413 | 17,99 | 1 | 36 628 |
| 8 | 2 014 | 1_412 | 16,65 | 1 | 33 533 |
| 9 | 2 003 | 1_406 | 24,81 | 1 | 49 694 |
| 10 | 2 001 | 1_407 | 15,99 | 1 | 31 996 |

# 1. Typologie des produits, ventes et clients

## 1.3 Un Flop des ventes constitué majoritairement de produits de la catégorie 0

On définit ici le Flop des ventes par le 1er décile des produits en nombre de ventes sur les 2 ans



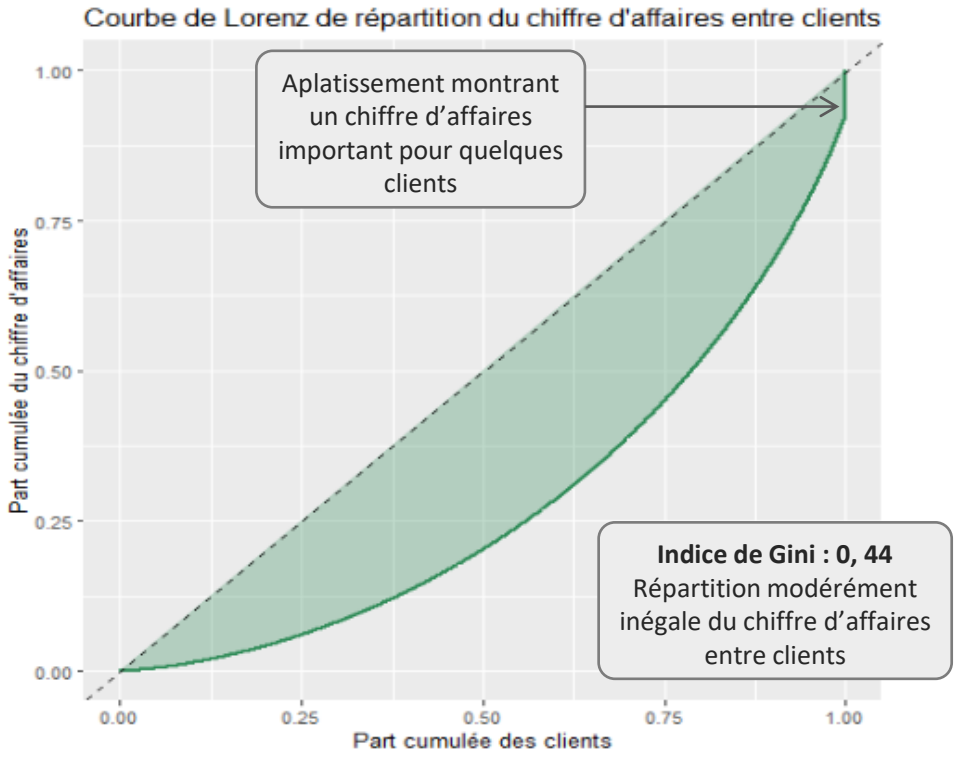

Une nette sur-représentation de la catégorie 0 et dans une moindre mesure de la catégorie 2

**Flop des produits non vendus en 2 ans**

| | Id produit | Prix (€) | Catégorie |
|---|---|---|---|
| 1 | 0_1014 | 1.15 | 0 |
| 2 | 0_1016 | 35.06 | 0 |
| 3 | 0_1025 | 24.99 | 0 |
| 4 | 0_1062 | 20.08 | 0 |
| 5 | 0_1119 | 2.99 | 0 |
| 6 | 0_1318 | 20.92 | 0 |
| 7 | 0_1620 | 0.80 | 0 |
| 8 | 0_1624 | 24.50 | 0 |
| 9 | 0_1645 | 2.99 | 0 |
| 10 | 0_1780 | 1.67 | 0 |
| 11 | 0_1800 | 22.05 | 0 |
| 12 | 0_2308 | 20.28 | 0 |
| 13 | 0_299 | 22.99 | 0 |
| 14 | 0_310 | 1.94 | 0 |
| 15 | 0_322 | 2.99 | 0 |
| 16 | 0_510 | 23.66 | 0 |
| 17 | 1_0 | 31.82 | 1 |
| 18 | 1_394 | 39.73 | 1 |
| 19 | 2_72 | 141.32 | 2 |
| 20 | 2_86 | 132.36 | 2 |
| 21 | 2_87 | 220.99 | 2 |

# 1. Typologie des produits, ventes et clients

## 1.4 4 clients représentant chacun plus de 100 000 € d'achats sur 2 ans




Aplatissement montrant un chiffre d'affaires important pour quelques clients

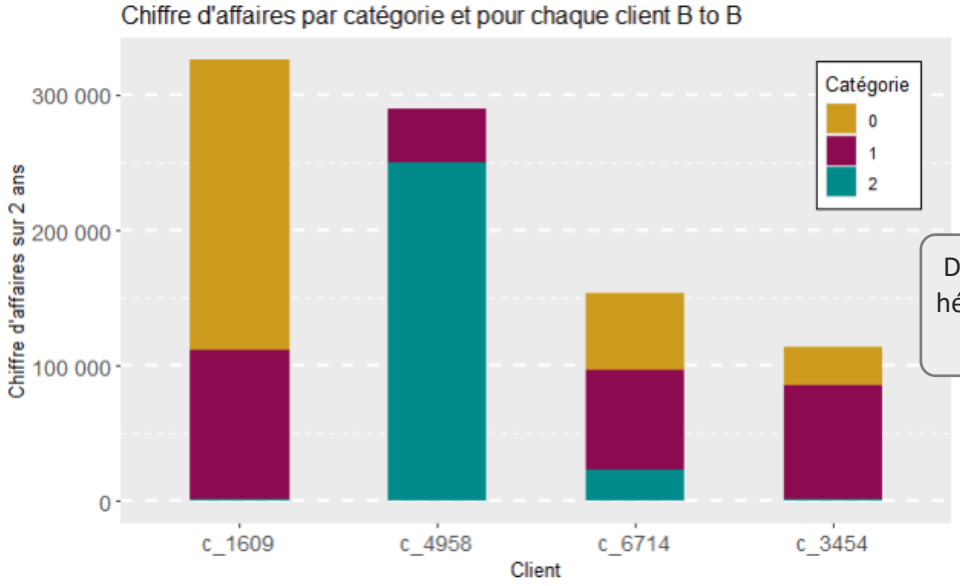
**Indice de Gini : 0, 44**
Répartition modérément inégale du chiffre d'affaires entre clients

**8 600 clients actifs**

**8 596 clients « B to C »**
Achats < 6 000 €

**4 clients « B to B »**
Achats > 100 000 €
7% du chiffre d'affaires

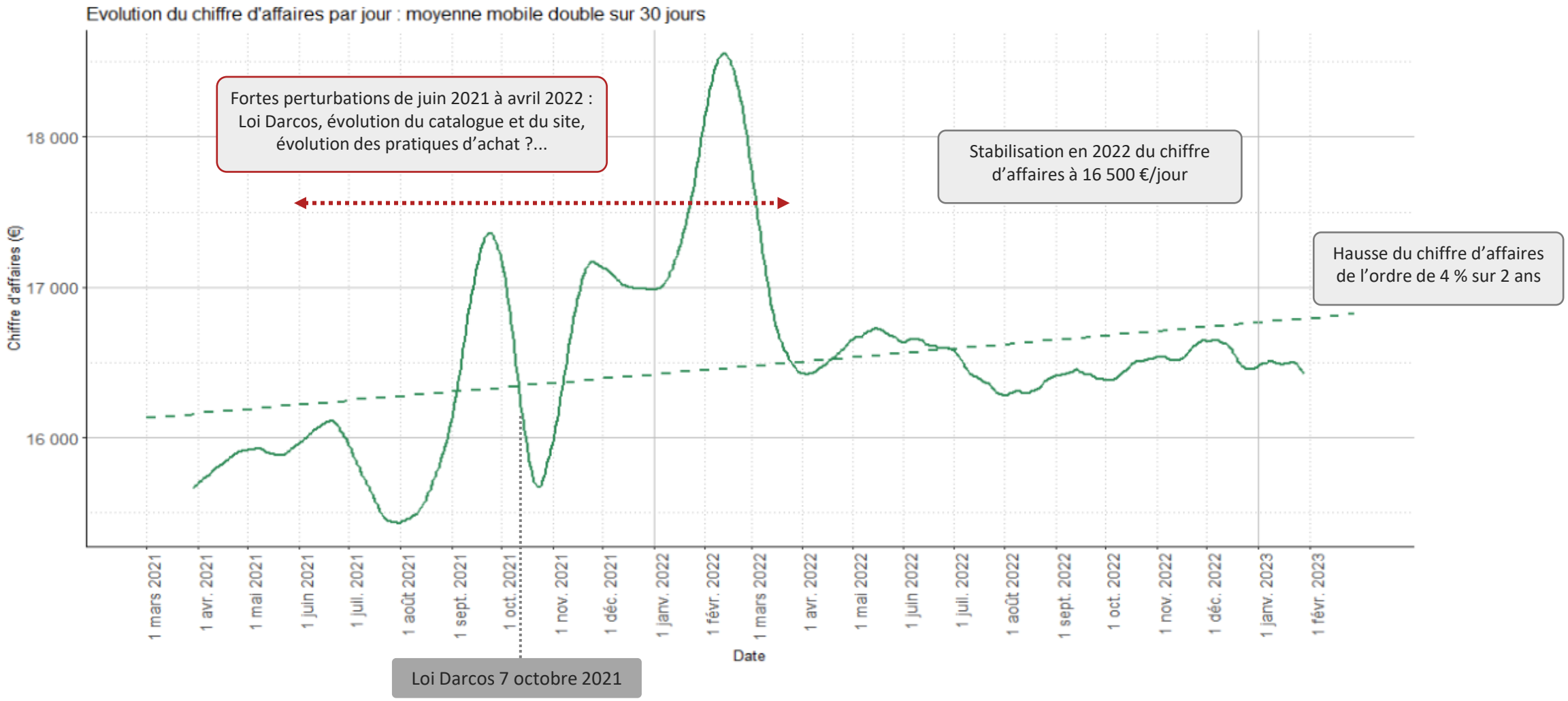
! Pas d'information sur la nature des clients : certains clients ici identifiés comme « B to C » pourraient représenter des entreprises




Des typologies d'achats très hétérogènes concentrées sur certaines catégories

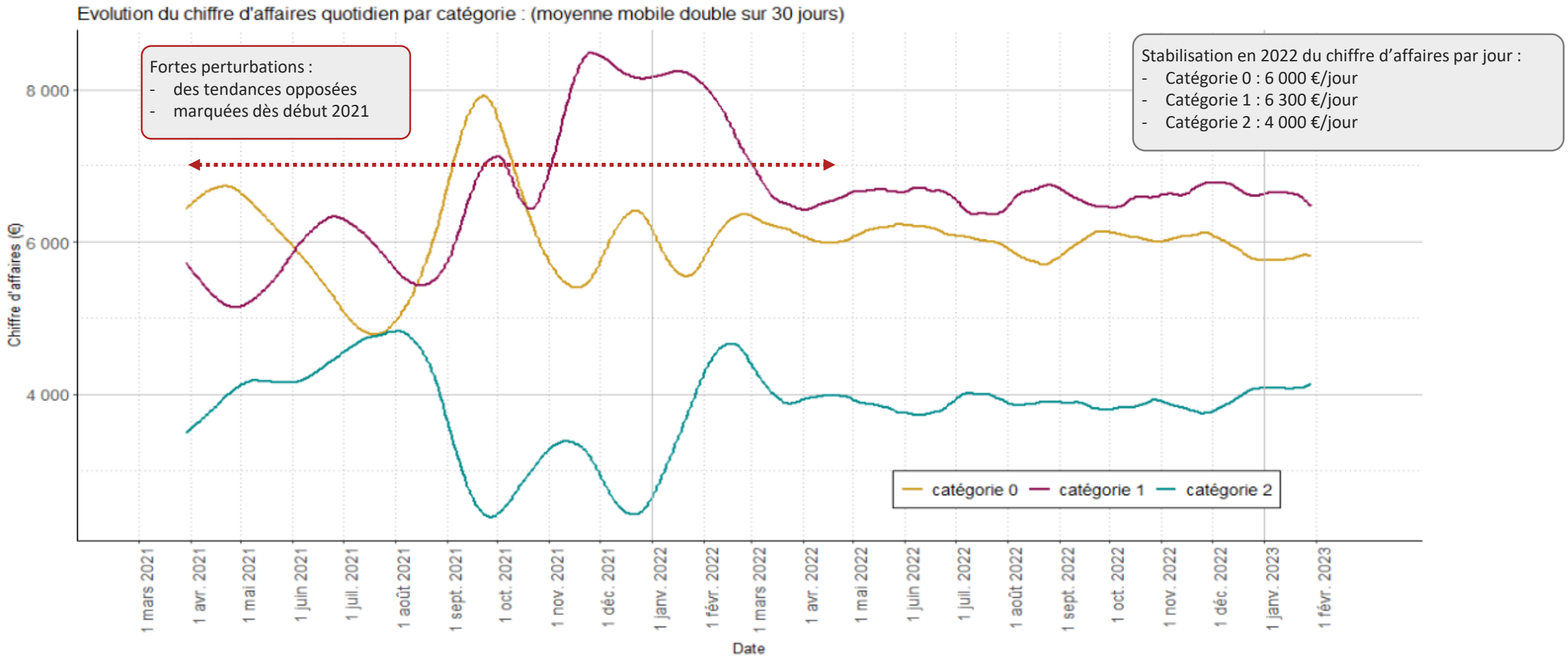
# 2. Une légère progression des ventes marquée par des perturbations

## 2.1 Un chiffre d'affaires stabilisé, en légère hausse par rapport à 2021



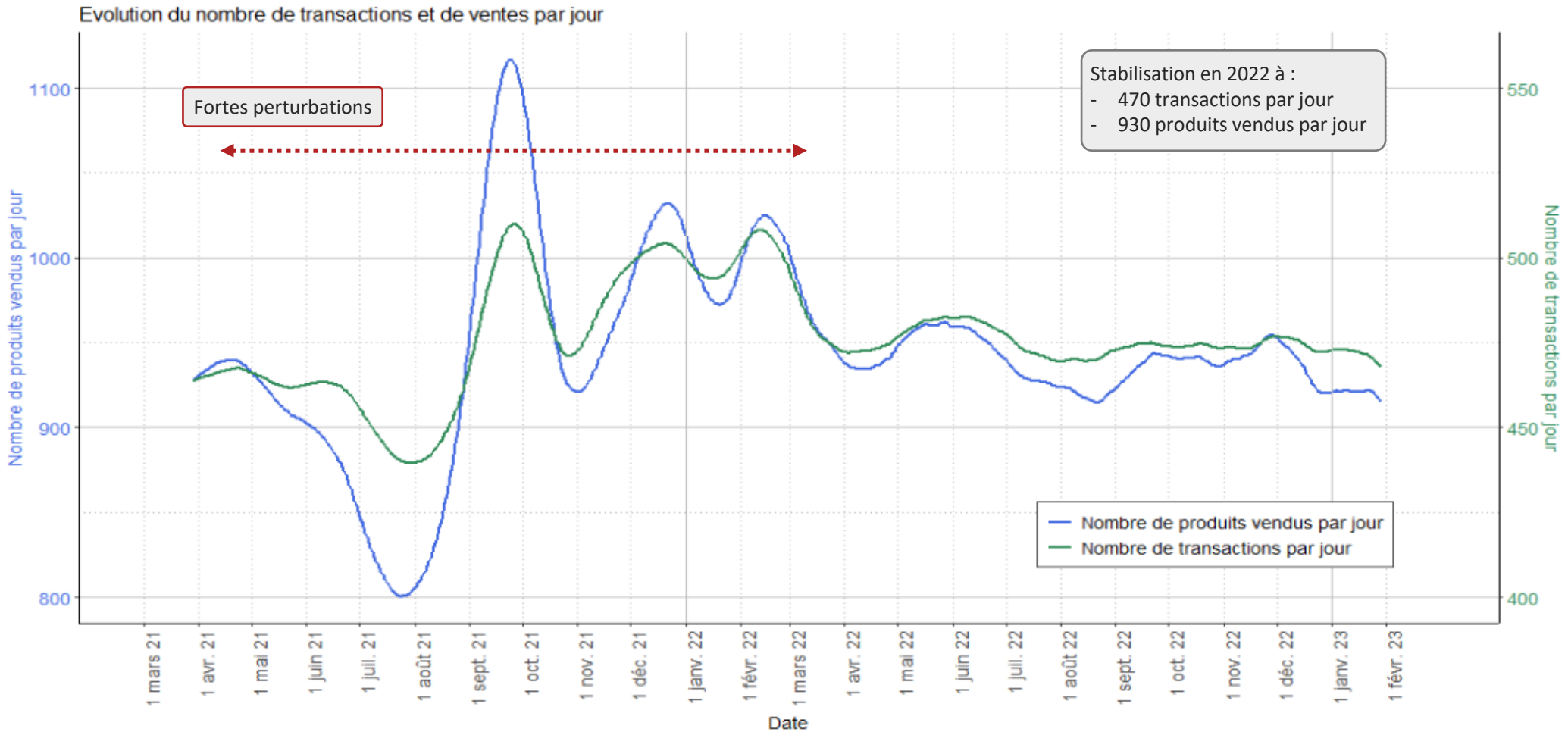
# 2. Une légère progression des ventes marquée par des perturbations

## 2.2 Lors des perturbations, des tendances opposées entre les catégories



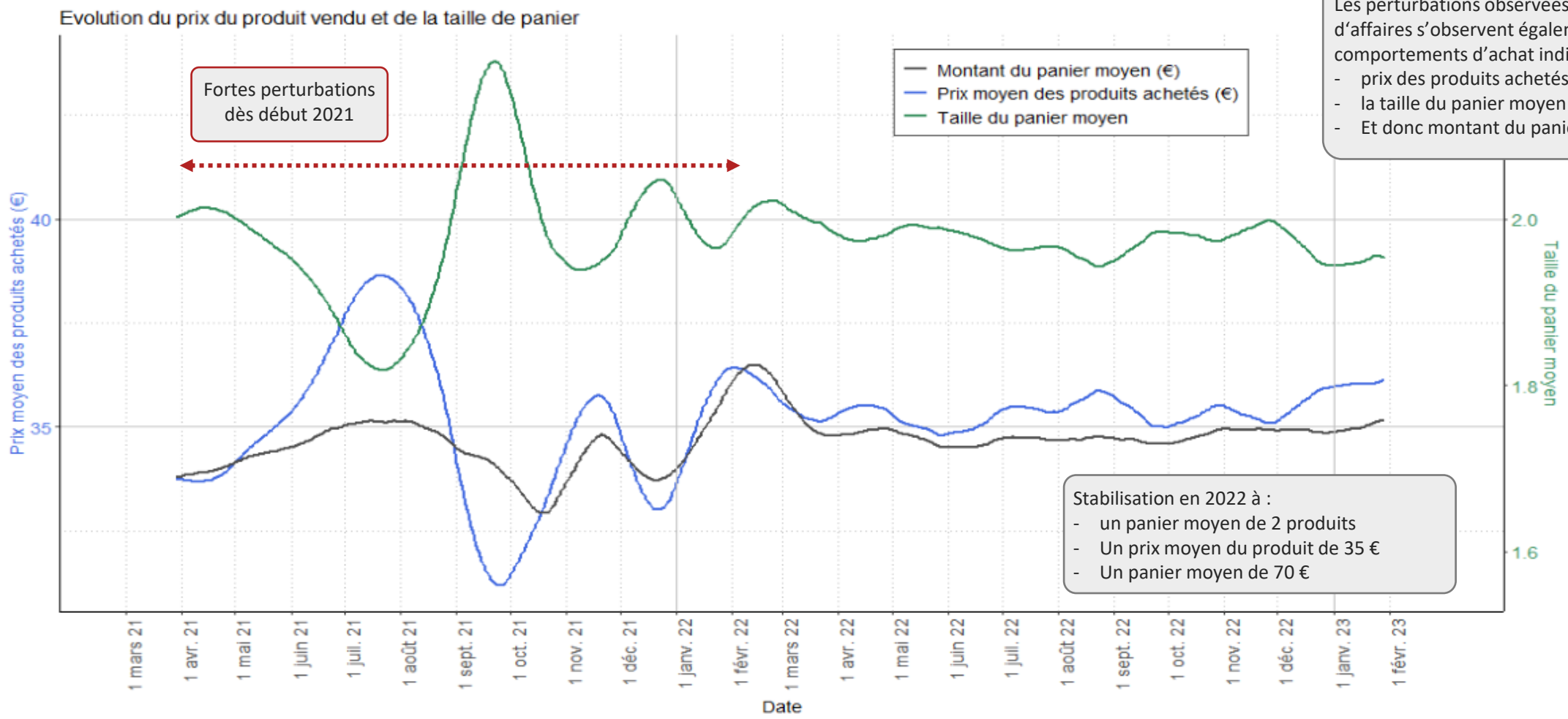
# 2. Une légère progression des ventes marquée par des perturbations

## 2.3 Evolution du nombre de transactions et de produits vendus par jour



# 2. Une légère progression des ventes marquée par des perturbations

## 2.4 Une évolution des comportements d'achat avec une stabilisation en 2022

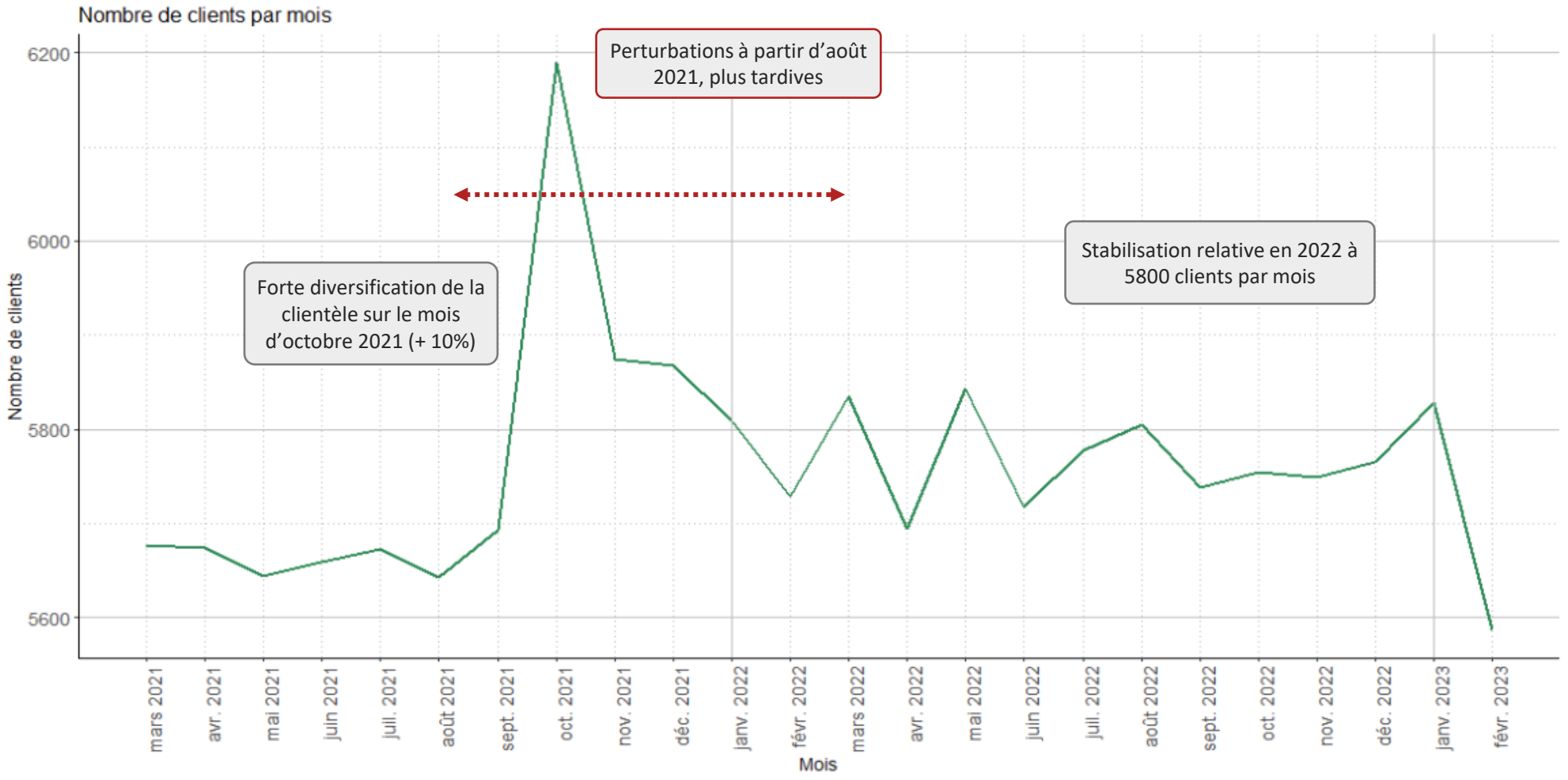



Les perturbations observées sur le chiffre d'affaires s'observent également sur les comportements d'achat individuels :
- prix des produits achetés
- la taille du panier moyen
- Et donc montant du panier moyen

Stabilisation en 2022 à :
- un panier moyen de 2 produits
- Un prix moyen du produit de 35 €
- Un panier moyen de 70 €

# 2. Une légère progression des ventes marquée par des perturbations

## 2.5 Des perturbations visibles également sur le nombre de clients différents chaque mois



# 3. Des corrélations marquées entre l'âge des clients et la typologie de leurs achats

## 3.1 Objectifs

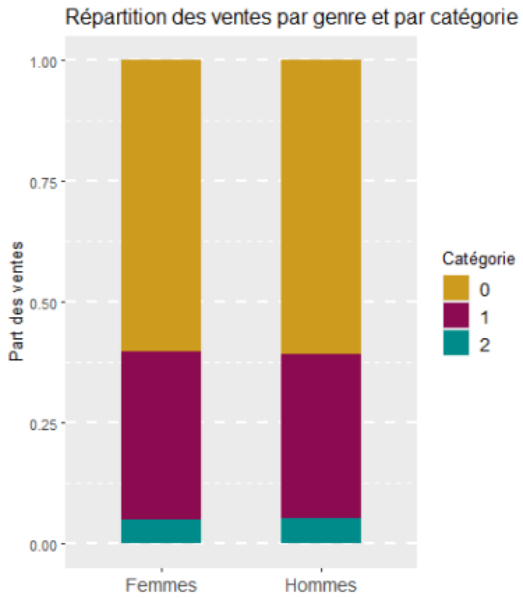
Les recherches de corrélation permettent de **mettre en évidence des pratiques d'achat spécifiques des clients en fonction de leur profil**, et donc de mieux cibler les clients.
On analysera dans cette partie les corrélations suivantes :

| | Age | Genre |
|---|---|---|
| Catégorie de produits achetés | x | x |
| Montant total d'achats | x | |
| Fréquence d'achat | x | |
| Taille du panier | x | |

On ne prendra pas en compte les ventes des clients « B to B » considérés comme outliers.

Ces analyses de corrélations sont d'autant plus pertinentes que le nombre d'observations reste très important : 641 000 produits vendus en 2 ans

## 3.2 Des catégories très peu genrées




**Analyse de la corrélation des variables « Catégorie » et « Genre » par un test du Khi-deux sur une table de contingence**

$\chi^2 = 23 \Rightarrow p_{valeur} = 10^{-5} << 5\%$

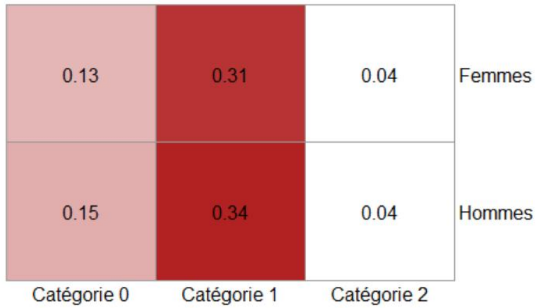
- $\chi^2$ : somme des variations entre les observations et les proportions théoriques
- $p_{valeur}$ : probabilité de se tromper en rejetant l'hypothèse d'indépendance des variables « Catégorie » et « Genre »

On peut donc rejeter l'hypothèse d'indépendance des variables

Coefficient V de Cramer = 0,006 ⇒ très faible corrélation

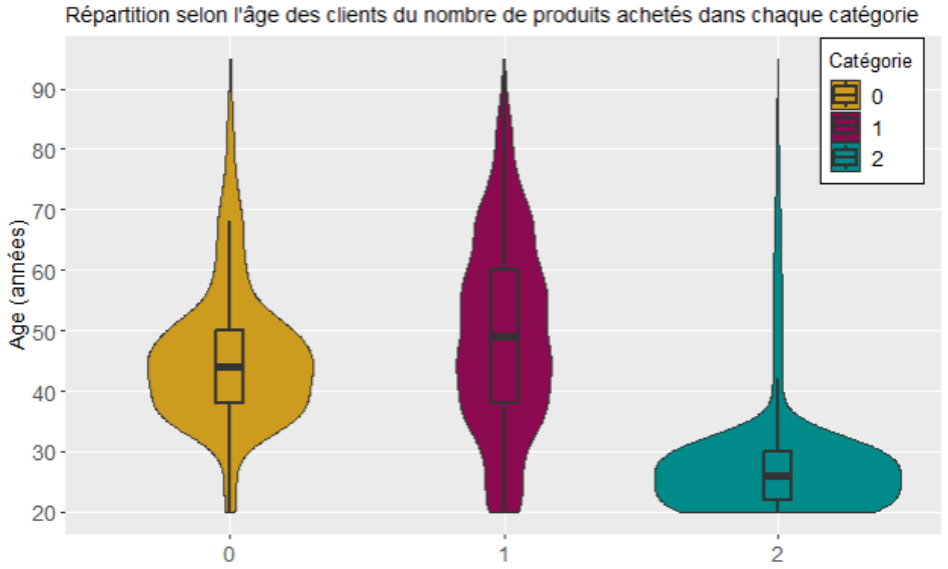
On observe une légère sur-représentation des femmes dans les ventes de catégorie 1 : + 900 sur 200 000 ventes totales de catégorie 1

**Carte de chaleur**
Participation de chaque groupe à la corrélation
entre les variables « Catégorie » et « Genre »



# 3. Des corrélations marquées entre l'âge des clients et la typologie de leurs achats

## 3.2 Une catégorie 2 très prisée par les plus jeunes

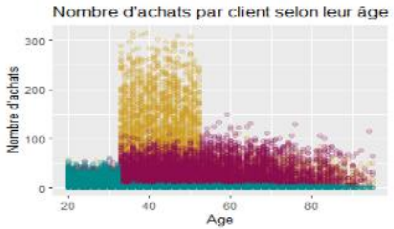



**1. Analyse de la répartition des âges par catégorie par un test de Kruskal-Wallis**

Les conditions de distribution normale et d'égalité des variances ne sont pas réunies pour un test ANOVA

$\chi^2$= 71 360 ⇒ $p_{valeur} < 10^{-15} << 5\%$
On peut donc rejeter l'hypothèse d'indépendance des variables

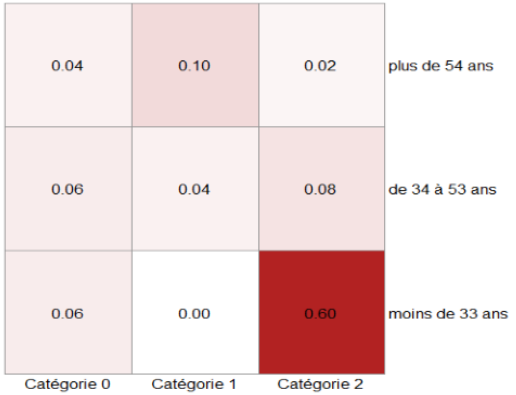
**2. Analyse de l'indépendance des variables par un test du Khi-2 après discrétisation par tranche d'âge**




On note très nettement à la lecture des différents nuages de points une rupture à 33 et à 53 ans dans les comportements d'achat : on utilise donc ces valeurs pour définir nos tranches

La table de contingence donne un coefficient V de Cramer = 0,43
⇒ corrélation assez marquée

Cette corrélation est principalement due à une forte présence des moins de 33 ans dans les achats de catégorie 2



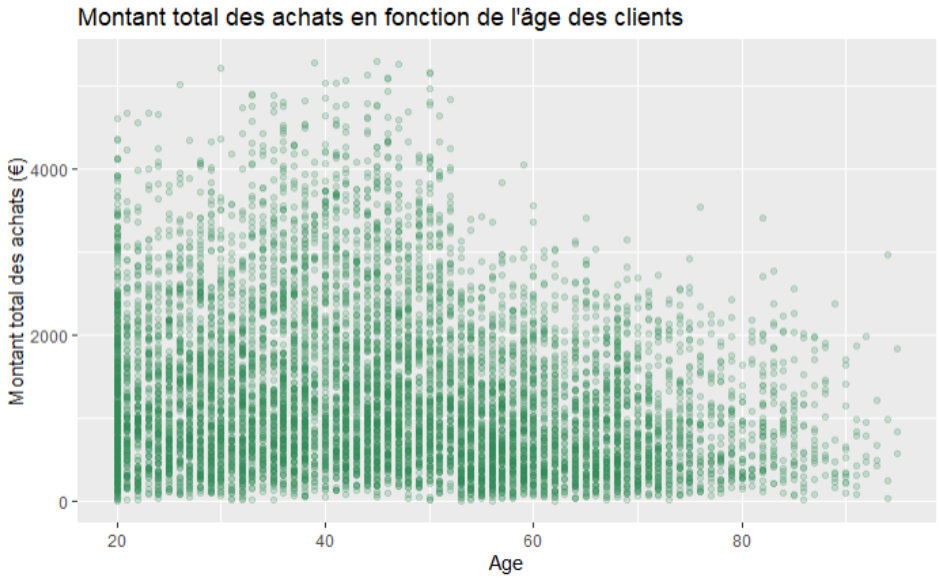

**Carte de chaleur**
Participation de chaque groupe à la corrélation entre les variables « Catégorie » et « Age »

**Interprétation**

La catégorie 2, plus chère que les autres et donc constituée d'ouvrages particuliers, pourrait correspondre à des livres destinés aux étudiants, et/ou à des albums illustrés

# 3. Des corrélations marquées entre l'âge des clients et la typologie de leurs achats

## 3.3 Le montant des achats décroît avec l'âge




**1. Analyse de la corrélation par le coefficient de Spearman**

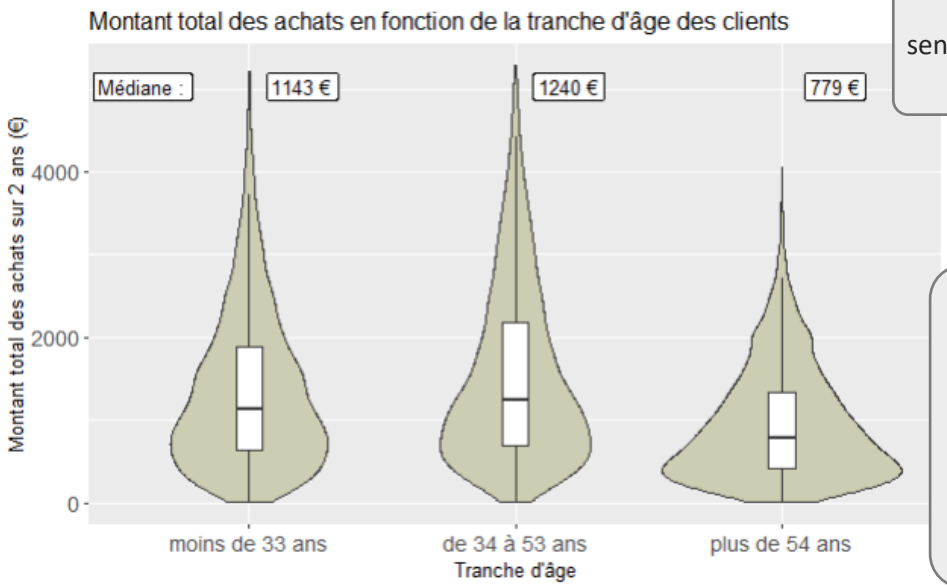
Deux variables quantitatives de corrélation non linéaire
Coefficient de corrélation de Spearman : - 0,18 (faible corrélation négative)

**2. Analyse de l'indépendance des variables par un test de Kruskal-Wallis après discrétisation par tranche d'âge**

$\chi^2 = 558 \Rightarrow p_{valeur} < 10^{-15} << 5\%$

On a donc une probabilité nulle de se tromper en rejetant l'hypothèse d'indépendance des variables

La table de contingence après discrétisation de la variable « montant total des achats » donne un coefficient V de Cramer de 0,16 ⇒ corrélation assez faible



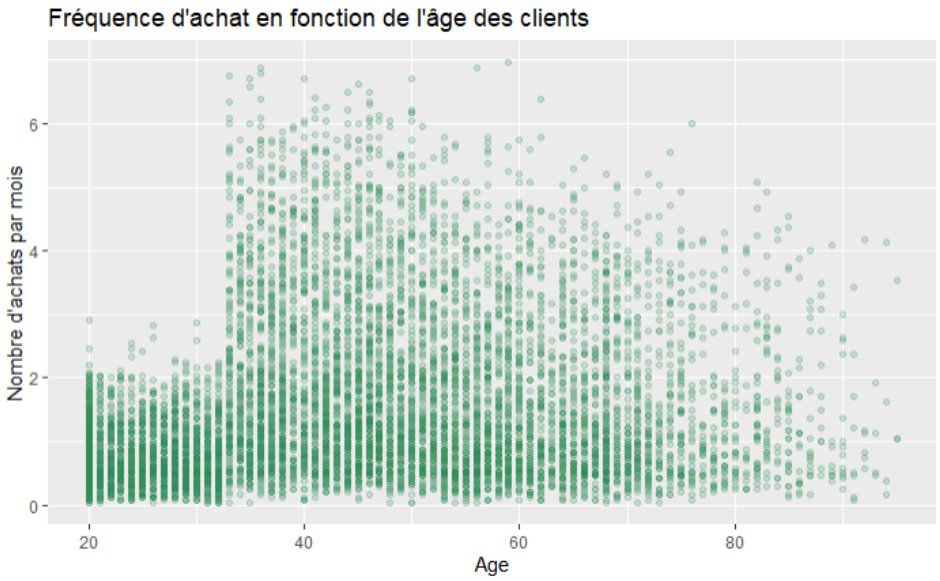

Un montant des achats sensiblement plus faible après 54 ans

**Interprétation**

Cette tendance pourrait s'expliquer par une utilisation moindre d'internet chez les plus âgés, qui n'utiliseraient la vente en ligne que pour des achats spécifiques

# 3. Des corrélations marquées entre l'âge des clients et la typologie de leurs achats

## 3.4 Des achats sensiblement moins fréquents chez les plus jeunes




**1. Analyse de la corrélation par le coefficient de Spearman**

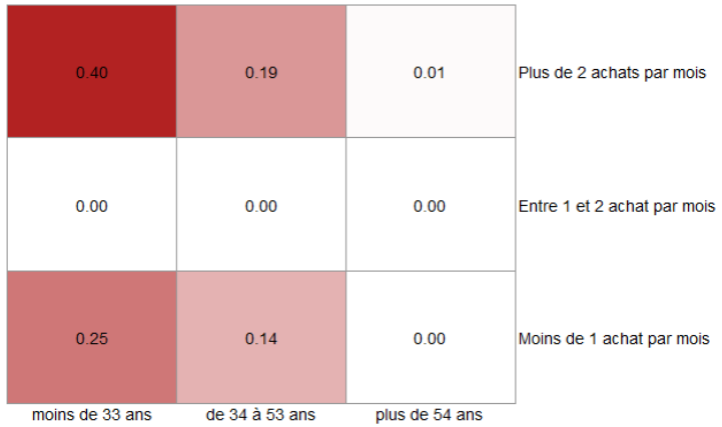
Deux variables quantitatives de corrélation non linéaire
Coefficient de Spearman : 0,21 (faible corrélation positive)

**2. Analyse de l'indépendance des variables par un test du Khi-2 après discrétisation par tranche d'âge et de fréquence d'achat**

- $\chi^2 = 1\,141 \Rightarrow p_{valeur} < 10^{-15} << 5\%$ ⇒ pas d'indépendance des variables
- Coefficient V de Cramer = 0,26 ⇒ corrélation peu marquée

La corrélation est principalement due à :

- **une faible présence des moins de 30 ans dans les fréquences supérieures à $2/_{mois}$**
- une forte présence des moins de 30 ans dans les fréquences inférieures à $1/_{mois}$
- une tendance inverse et plus faible chez les 34-53 ans



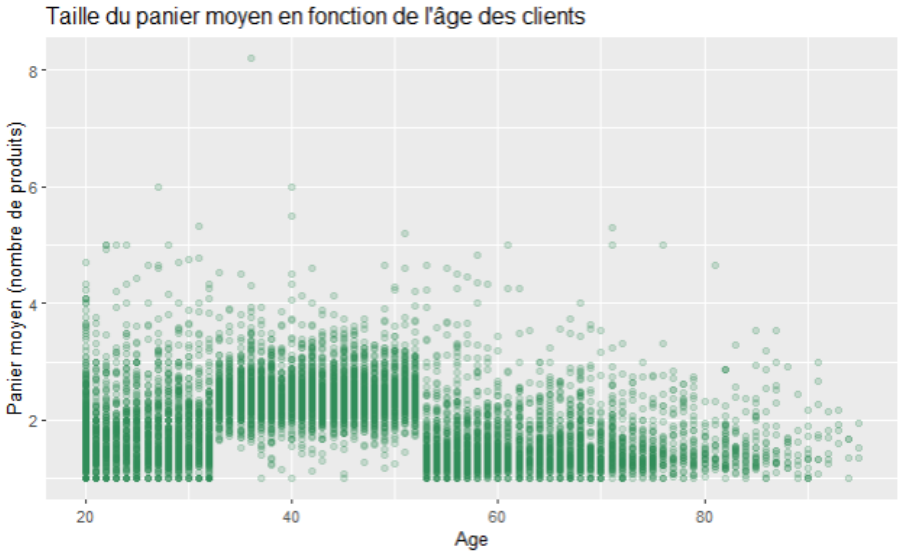

**Carte de chaleur**
Participation de chaque groupe à la corrélation entre les variables « Age » et « Fréquence d'achat »

**Interprétation**

- Un nombre de transactions supérieures à 2 par mois peut laisser supposer un client de type business, même petit, ce qui est effectivement plus probable dans la tranche 34-53 ans
- Les clients les plus jeunes se tournent probablement plus facilement vers la vente en ligne malgré des achats peu fréquents. Les personnes plus âgées s'y tourneraient au contraire par commodité lorsqu'ils réalisent beaucoup d'achats

# 3. Des corrélations marquées entre l'âge des clients et la typologie de leurs achats

## 3.5 Age et taille du panier




**1. Analyse de la corrélation par le coefficient de Spearman**

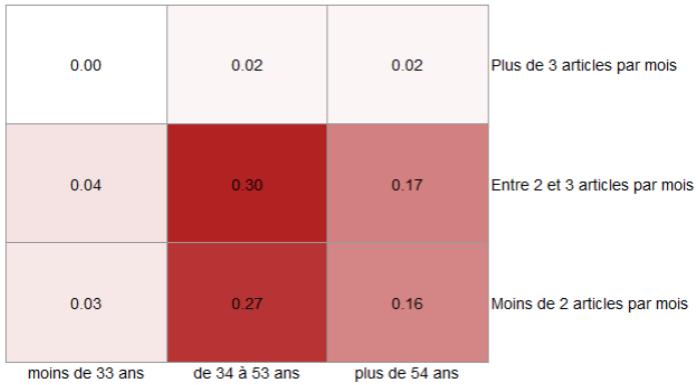
Deux variables quantitatives de corrélation non linéaire
Coefficient de corrélation de Spearman = - 0,20 (faible corrélation négative)

**2. Analyse de l'indépendance des variables par un test du Khi-deux après discrétisation par tranche d'âge et taille de panier moyen**

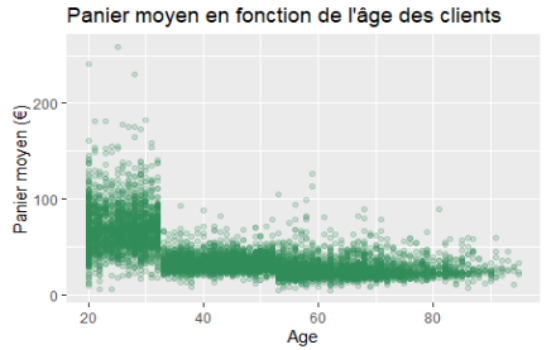
- $\chi^2$= 3 505 ⇒ $p_{valeur} < 10^{-15} << 5\%$
  ⇒ pas d'indépendance des variables
- Coefficient V de Cramer = 0,45
  ⇒ corrélation assez marquée

La corrélation est principalement due à :

- **une forte présence des 34-53 ans dans les paniers de 2 à 3 articles**
- **une faible présence des 34-53 ans dans les paniers de moins de 2 articles**
- une tendance inverse et plus faible chez les plus de 54 ans




**Carte de chaleur**
Participation de chaque groupe à la corrélation entre les variables « Age » et « Taille du panier moyen»



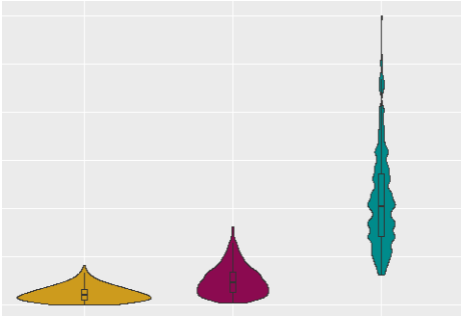

**Interprétation**

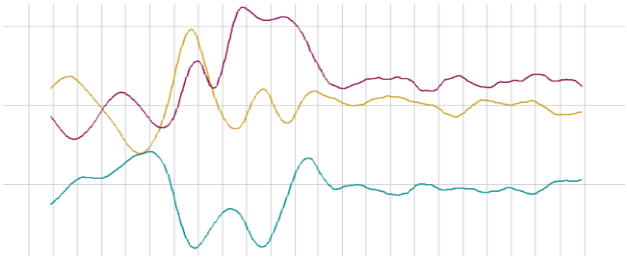
Cet écart peut s'expliquer par

- Un meilleur pouvoir d'achat des 34-53 ans
- Une possible présence d'entreprises dans la tranche 34-53 ans
- Des achats plus anticipés et ciblés chez les plus de 54 ans

# Conclusion

- La catégorie 0, qui constitue 70% du catalogue, est bon marché et très présente dans le flop des ventes
- La catégorie 1 est très présente dans le top des ventes et très vendue en proportion
- La catégorie 2 est nettement plus chère et peu vendue, et contient peu de références





- Les ventes ont connu de fortes perturbations de mi-2021 à début 2022, avec une stabilisation en 2022
- Les perturbations ont probablement des origines multiples : législation, évolution des pratiques d'achat après le Covid, peut-être une évolution du site et du catalogue

- L'âge impacte principalement la fréquence d'achat et la taille du panier :
  - Les moins de 33 ans achètent rarement, peu d'articles et des produits plus chers. Ils sont les principaux clients de la catégorie 2
  - Les 34-53 ans sont les meilleurs clients en montant total d'achats. Ils mettent plus d'articles dans leur panier et achètent plus fréquemment. Certains représentent peut-être de petites entreprises
  - Les plus de 54 ans dépensent sensiblement moins par an, et mettent peu d'articles dans leur panier