

Analyse des ventes



Projet 6 - Laurent May - 2021

OpenClassrooms ENSEA - ENSAI



SOMMAIRE

Les problématiques :

Evolution des ventes

Répartition des clients

Corrélations entre les clients et les produits vendus

PRESENTATION

- l'entreprise
- les données disponibles

NETTOYAGE

Identifier et corriger les données altérées, inexactes ou non pertinentes.

ANALYSES

Evolution des ventes, la typologie client et la répartition des prix de ventes.

CORRELATIONS

Comportement de nos clients en ligne.

The background image shows a dimly lit library or study room. On the left, there are tall wooden bookshelves filled with books. In the center-right, several vintage-style lightbulbs hang from the ceiling by wires. The overall atmosphere is quiet and scholarly.

Présentation et nettoyage des données



À Propos de "Lapage"

Lapage était originellement une librairie physique avec plusieurs points de vente.

Devant le succès de certains de nos produits et l'engouement de nos clients, Nous avons décidé depuis 2 ans d'ouvrir un site de vente en ligne.

Nettoyage des données

Customers

	client_id	sex	birth
0	c_4410	f	1967
1	c_7839	f	1975
2	c_1699	f	1984
3	c_5961	f	1962
4	c_5320	m	1943

- client_id : numéro d'identifiant du client
- sex : genre du client
- birth : année de naissance du client

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8623 entries, 0 to 8622
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype  
 ---  -- 
 0   client_id    8623 non-null   object 
 1   sex          8623 non-null   object 
 2   birth         8623 non-null   int64  
 dtypes: int64(1), object(2)
memory usage: 202.2+ KB
```

- 8623 clients (pas de doublons)
- Minimum plage année : 1929
- Maximum plage année : 2004
- client_id : clé primaire

Products

	id_prod	price	categ
0	0_1421	19.99	0
1	0_1368	5.13	0
2	0_731	17.99	0
3	1_587	4.99	1
4	0_1507	3.99	0

- id_prod : numéro d'identifiant de l'article
- price : prix de l'article
- categ : catégorie à laquelle fait partie l'article

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3287 entries, 0 to 3286
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype  
 ---  -- 
 0   id_prod   3287 non-null   object 
 1   price     3287 non-null   float64 
 2   categ     3287 non-null   int64  
 dtypes: float64(1), int64(1), object(1)
memory usage: 77.2+ KB
```

	price	categ
count	3.287.00	3.287.00
mean	21.86	0.37
std	29.85	0.62
min	-1.00	0.00
25%	6.99	0.00
50%	13.06	0.00
75%	22.99	1.00
max	300.00	2.00

	id_prod	price	categ
	731	T_0	-1.00

A supprimer au nettoyage

- 3287 produits (pas de doublons)
- 3 catégories : 0, 1 et 2
- id_prod : clé primaire

Transactions

	id_prod	date	session_id	client_id
0	0_1518	2022-05-20 13:21:29.043970	s_211425	c_103
1	1_251	2022-02-02 07:55:19.149409	s_158752	c_8534
2	0_1277	2022-06-18 15:44:33.155329	s_225667	c_6714
3	2_209	2021-06-24 04:19:29.835891	s_52962	c_6941
4	0_1509	2023-01-11 08:22:08.194479	s_325227	c_4232

- id_prod : numéro d'identifiant de l'article
- date : date de la transaction
- session_id : numéro de session
- client_id : numéro d'identifiant du client

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 679532 entries, 0 to 679531
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype  
 ---  -- 
 0   id_prod   679532 non-null   object 
 1   date      679532 non-null   object 
 2   session_id 679532 non-null   object 
 3   client_id  679532 non-null   object 
 dtypes: object(4)
memory usage: 20.7+ MB
```

Transactions Test à supprimer

	id_prod	date	session_id	client_id
T_0	test_2021-03-01 02:30:02.237419		s_0	ct_0
T_0	test_2021-03-01 02:30:02.237425		s_0	ct_0
T_0	test_2021-03-01 02:30:02.237437		s_0	ct_1

Clients et produits Test à supprimer :

- id_prod test : ['T_0']
- client_id test : ['ct_0' 'ct_1']

- 679 532 transactions produits (161 doublons)
- Variable 'date' au format string à convertir
- Produit '0_2245' ne fait pas partie du catalogue

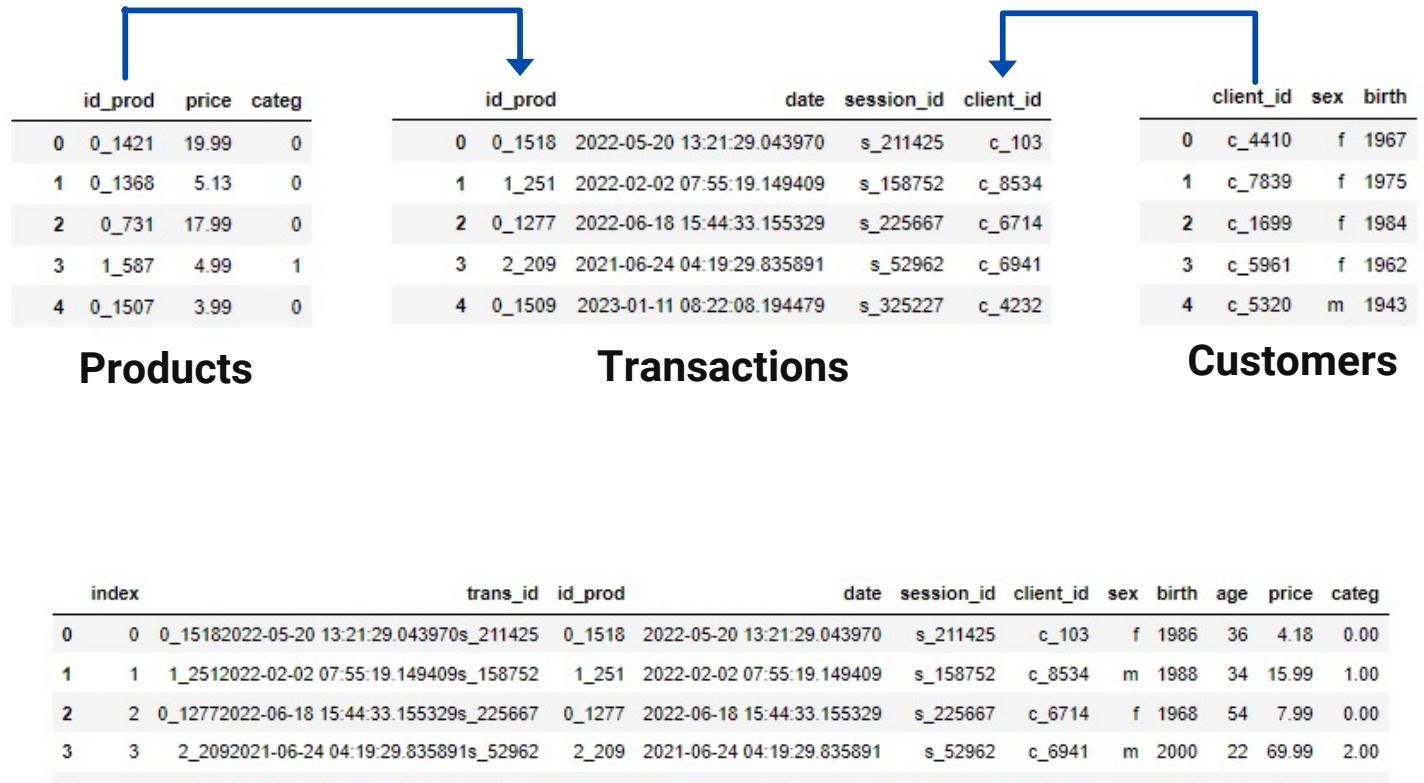
Jointures

Etapes de la jointure

- 2 jointures
- Ajout de la variable âge
- Conversion des dates

Après jointure, on observe des transactions qui concernent un produit qui n'est pas présent dans le catalogue : **0_2245**

Ces transactions représentent 0.03% des données. On décide de supprimer ces données



Products			Transactions				Customers		
id_prod	price	categ	id_prod	date	session_id	client_id	client_id	sex	birth
0 0_1421	19.99	0	0 0_1518	2022-05-20 13:21:29.043970	s_211425	c_103	0 c_4410	f 1967	
1 0_1368	5.13	0	1 1_251	2022-02-02 07:55:19.149409	s_158752	c_8534	1 c_7839	f 1975	
2 0_731	17.99	0	2 0_1277	2022-06-18 15:44:33.155329	s_225667	c_6714	2 c_1699	f 1984	
3 1_587	4.99	1	3 2_209	2021-06-24 04:19:29.835891	s_52962	c_6941	3 c_5961	f 1962	
4 0_1507	3.99	0	4 0_1509	2023-01-11 08:22:08.194479	s_325227	c_4232	4 c_5320	m 1943	

index	trans_id	id_prod	date	session_id	client_id	sex	birth	age	price	categ
0 0 0_15182022-05-20 13:21:29.043970s_211425	0_1518	2022-05-20 13:21:29.043970	s_211425	c_103	f 1986	36	4.18	0.00		
1 1 1_2512022-02-02 07:55:19.149409s_158752	1_251	2022-02-02 07:55:19.149409	s_158752	c_8534	m 1988	34	15.99	1.00		
2 2 0_12772022-06-18 15:44:33.155329s_225667	0_1277	2022-06-18 15:44:33.155329	s_225667	c_6714	f 1968	54	7.99	0.00		
3 3 2_2092021-06-24 04:19:29.835891s_52962	2_209	2021-06-24 04:19:29.835891	s_52962	c_6941	m 2000	22	69.99	2.00		
4 4 0_15092023-01-11 08:22:08.194479s_325227	0_1509	2023-01-11 08:22:08.194479	s_325227	c_4232	m 1980	42	4.99	0.00		

679 111 transactions

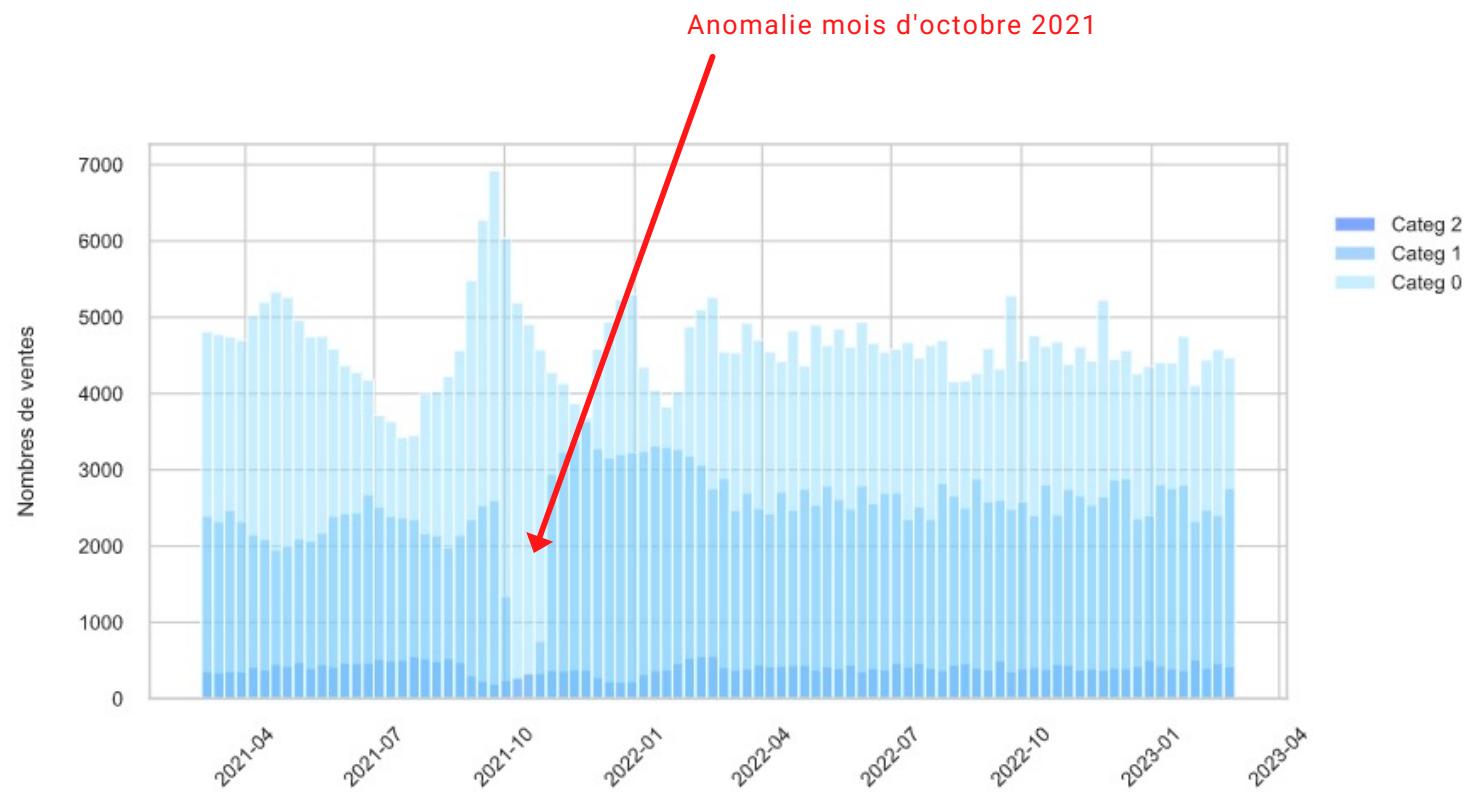
Plage de dates couvertes

- Début des transactions : 01-03-2021

- Fin des transactions : 28-02-2023



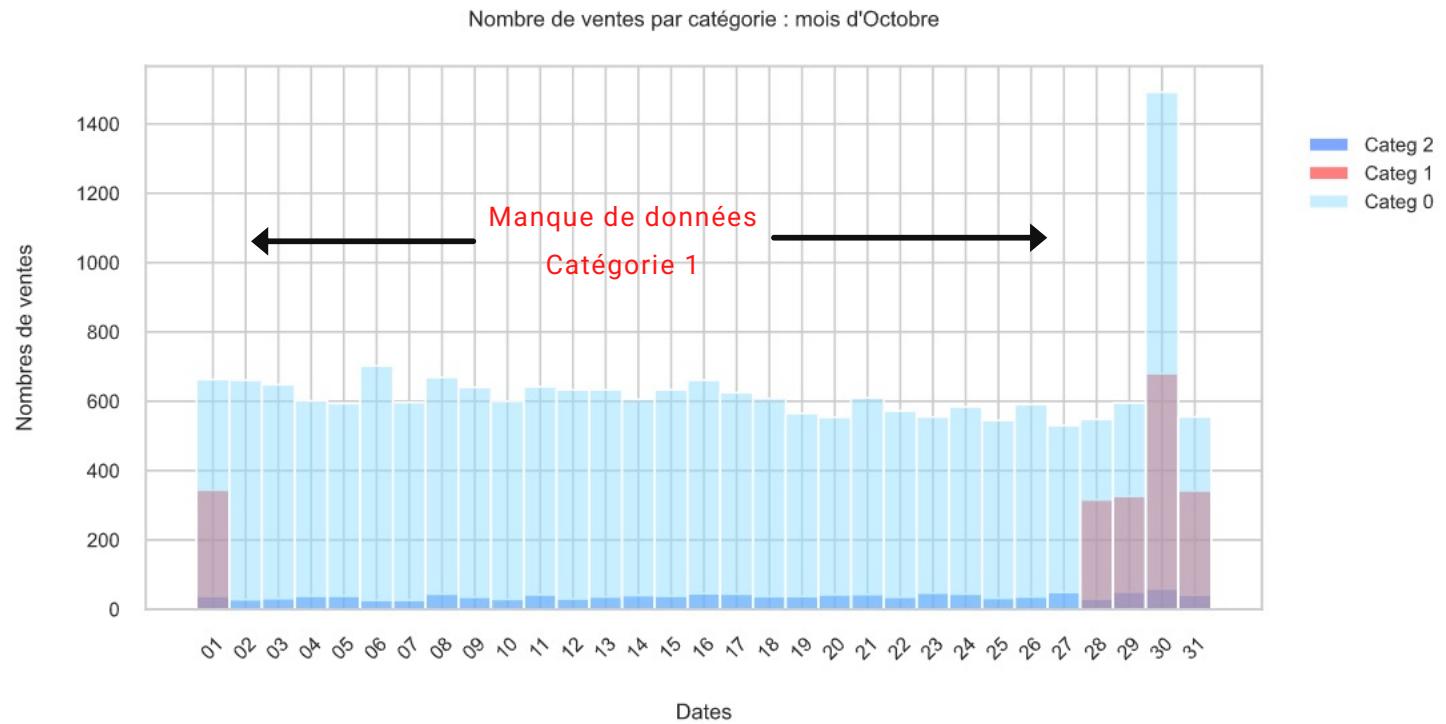
Couverture de
2 années fiscales



Plage de dates couvertes

Données manquante pour Categ 1 :
Du 02 au 27 octobre 2021

↓
Réprésente
3.37% du dataset



Cette absence de données peut être due à un problème de stock ou du site internet.

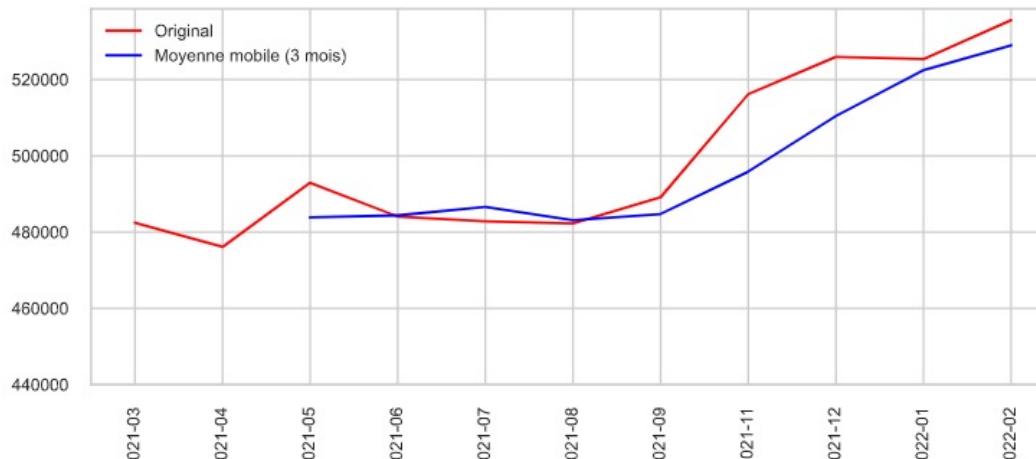
Pour éviter un biais dans l'analyse, on peut supprimer cette partie sans que cela n'affecte notre étude.

A photograph of a library interior at night. Bookshelves filled with books are visible in the background. Several hanging lightbulbs are suspended from the ceiling by wires, casting a warm glow. The scene is dimly lit, with the lightbulbs being the primary light source.

Chiffre d'affaire

Evolution du chiffre d'affaire

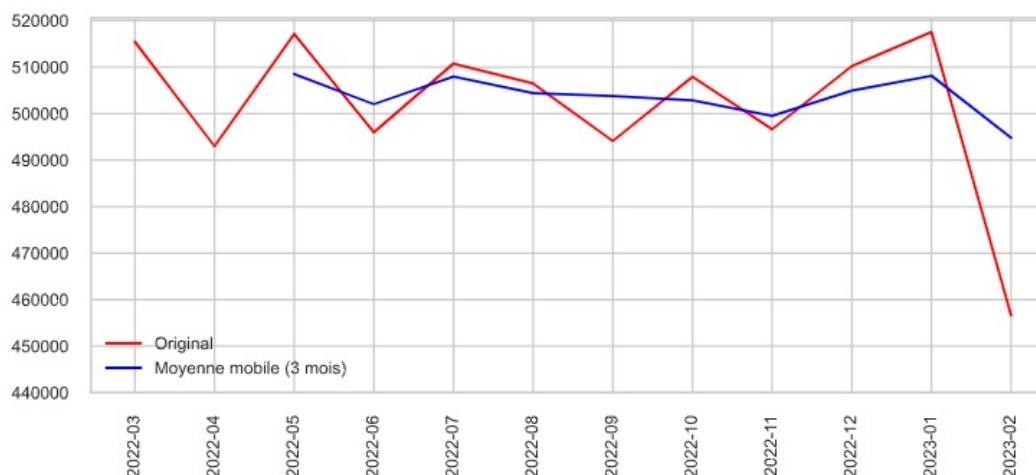
Chiffre d'affaire année fiscale : 2021



Année fiscal 2021 :

Le chiffre d'affaire a augmenté au cours de l'année. On observe une tendance haussière.

Chiffre d'affaire année fiscale : 2022



Année fiscale 2022 :

Le chiffre d'affaire est globalement constant durant l'année 2022.
On ne retrouve pas la même tendance haussière de 2021.

Baisse

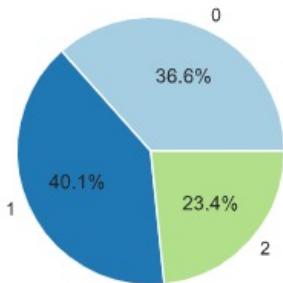
février 2023

?

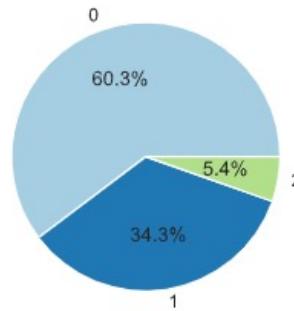
-60 861
(-13.33%)

Chiffre d'affaire par catégories

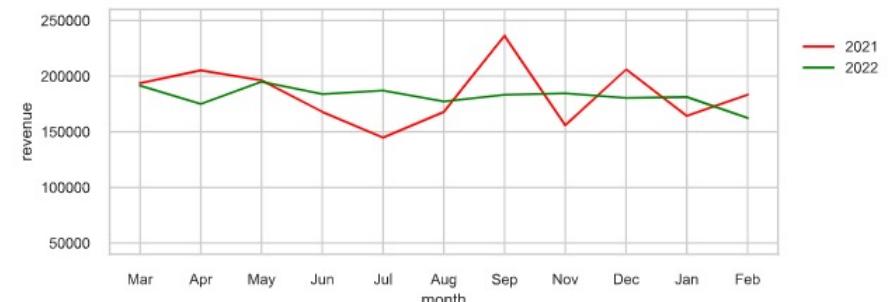
Chiffre d'affaire par catégorie



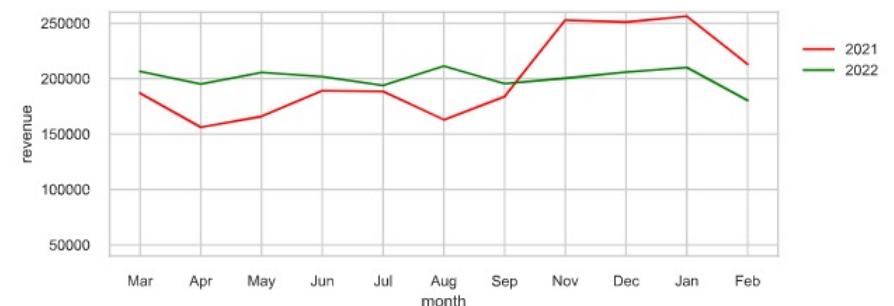
Volume des ventes par catégorie



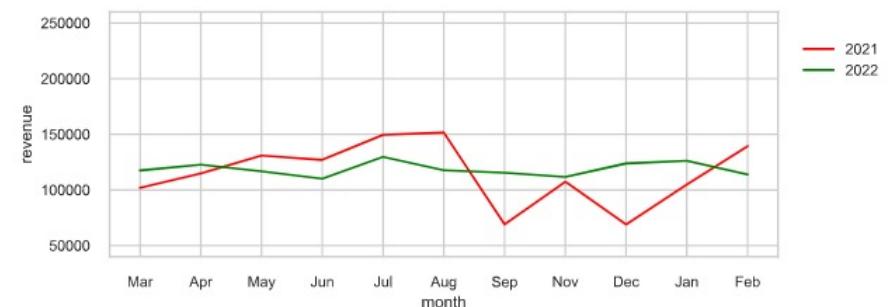
Chiffre d'affaire catégorie : 0



Chiffre d'affaire catégorie : 1



Chiffre d'affaire catégorie : 2

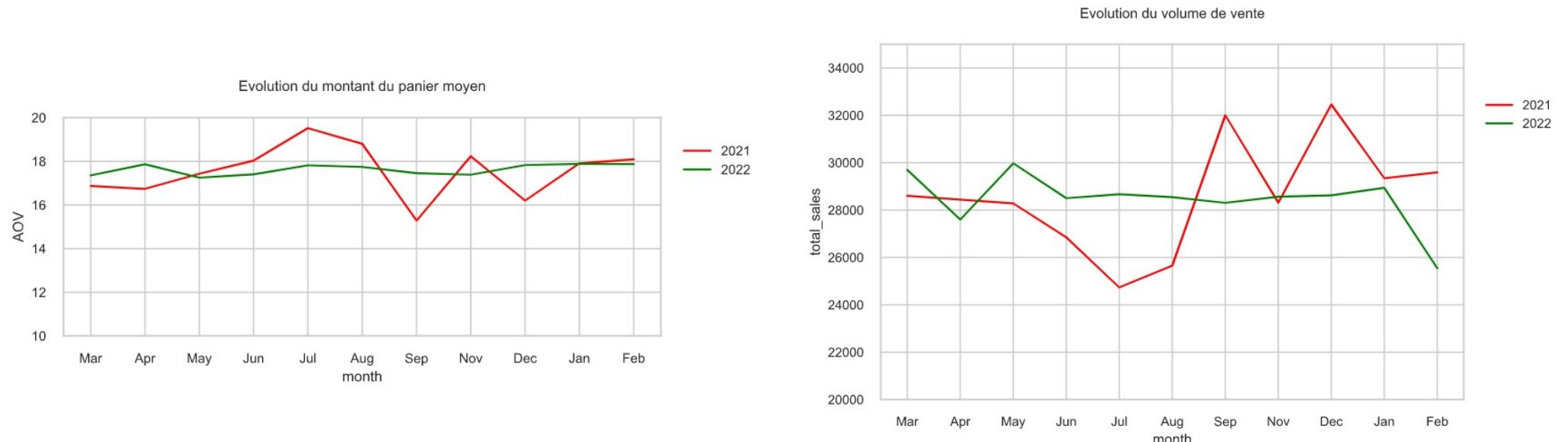


Les catégories 1 et 2 génèrent le plus de chiffre d'affaire.

La catégorie 0 est la plus vendue, suivie de la catégorie 1.

La baisse du chiffre d'affaire au mois de février
est généralisée sur l'ensemble des catégories.

Evolution du panier moyen



Le panier moyen est plutôt constant sur l'année 2022.

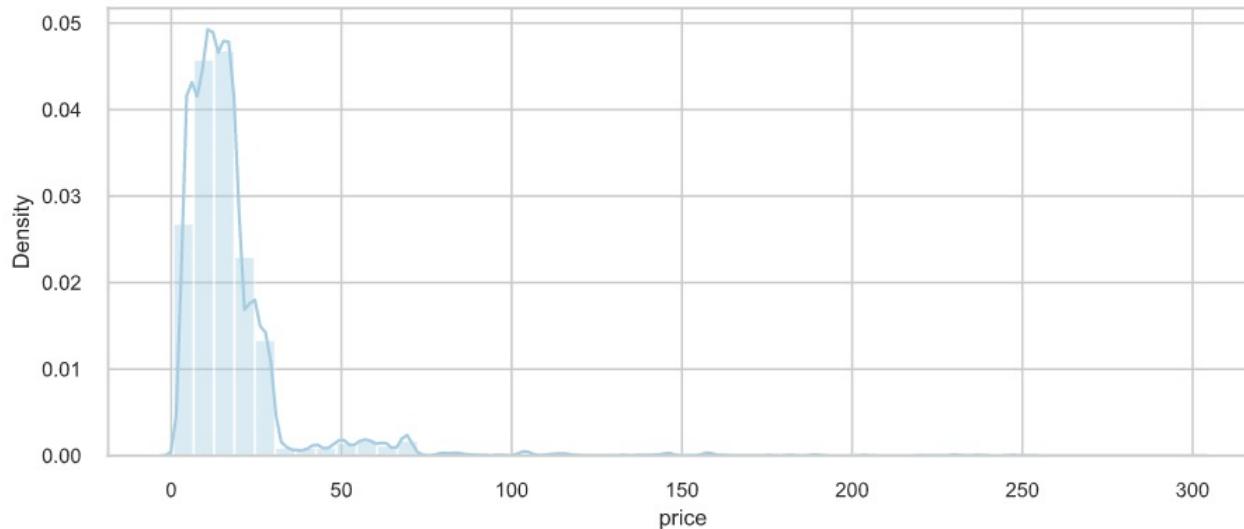
La baisse de chiffre d'affaire du mois de février 2023 est une résultante de la **baisse globale des volumes de ventes** pour ce mois.

The background image shows a dimly lit library or bookstore. Bookshelves filled with books are visible in the background. Several vintage-style lightbulbs hang from the ceiling by wires, casting a warm glow. The overall atmosphere is quiet and scholarly.

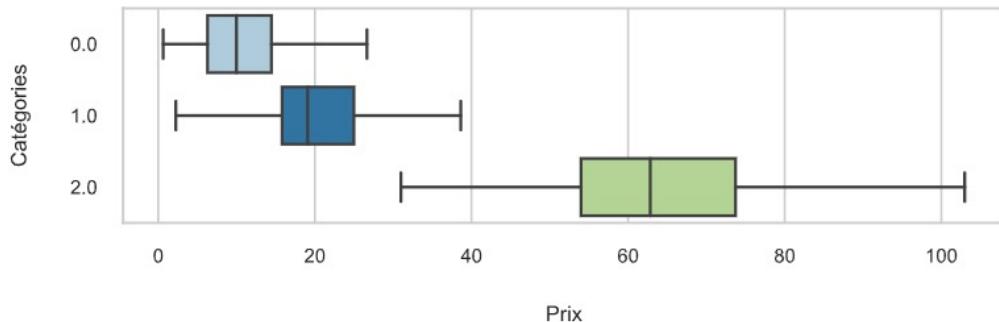
Produits

Plage de prix

Distribution des ventes par prix



Distribution des prix par catégorie



Les 3 catégories ont des plages de prix bien définies.
Le gros du volume des ventes se concentre
autour de 0 à 40 eur.

Classement des produits

Produits les plus vendus

Les références les plus vendues appartiennent à la catégorie 1.

Les références qui génèrent le plus de chiffre d'affaire appartiennent à la catégorie 2.

Par nombre de ventes :

id_prod	total_sales	revenue
1_369	2234	53,593.66
1_417	2169	45,527.31
1_414	2161	51,496.63
1_498	2117	49,474.29
1_425	2079	35,322.21

Par chiffre d'affaires :

id_prod	total_sales	revenue
2_159	632	92,265.68
2_135	977	67,403.23
2_112	929	62,772.53
2_102	997	58,962.58
2_209	790	55,292.10

Produits les moins vendus

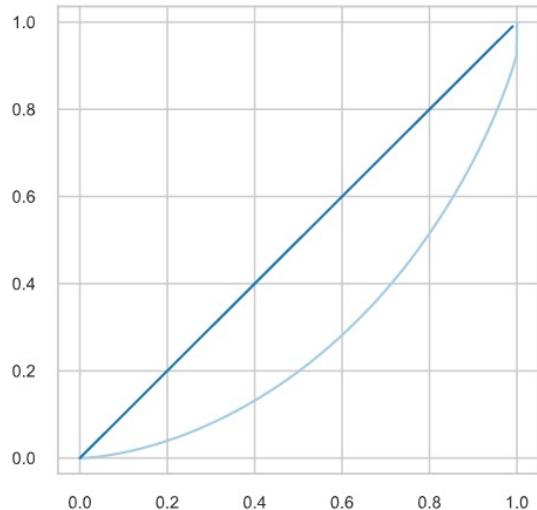
Les références les moins vendues et qui génèrent le moins de chiffre d'affaire font partie de la catégorie 0.

id_prod	total_sales	revenue
0_1233	1	21.99
0_1533	1	27.99
2_23	1	115.99
0_1539	1	0.99
0_1728	1	2.27

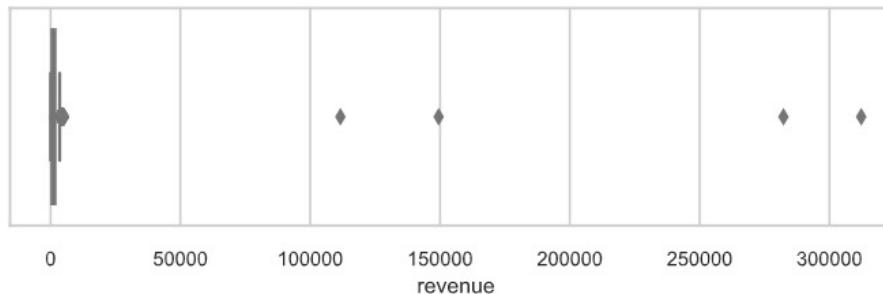
id_prod	total_sales	revenue
0_1539	1	0.99
0_898	1	1.27
0_1284	1	1.38
0_643	2	1.98
0_1653	2	1.98

Catégorisation des clients

Courbe de Lorenz : inégalité des revenus générés par les clients



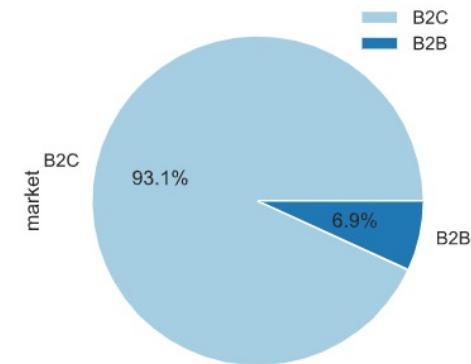
Distribution des chiffre d'affaire par clients



client_id	total_sales	revenue
c_1609	24427	312,247.61
c_4958	5085	282,289.70
c_6714	8875	149,484.49
c_3454	6623	111,638.84

Il y a 4 clients qui génèrent un chiffre d'affaire très supérieur aux autres. On les écarte pour ne pas impacter l'analyse sur les variables. Nous avons deux types de clients: B2B et B2C.

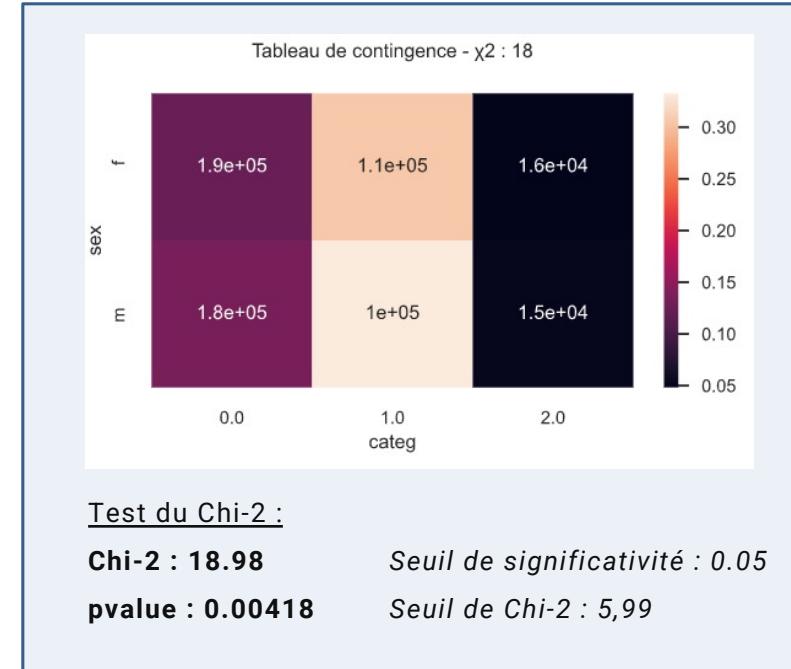
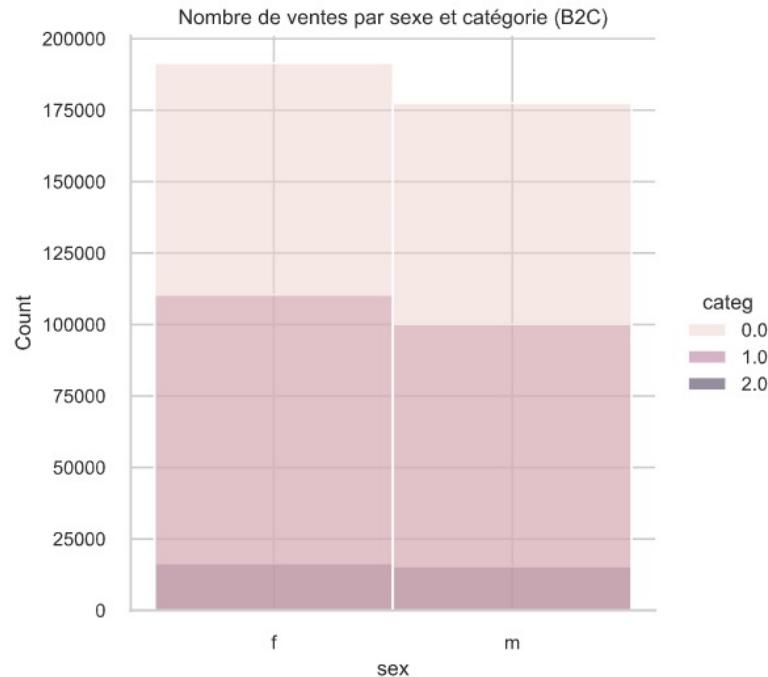
Volume des transactions par type de client



The background image shows a dimly lit library or bookstore. Bookshelves filled with books are visible in the background. Several hanging lightbulbs are suspended from the ceiling by wires, casting a warm glow. The overall atmosphere is quiet and scholarly.

Corrélations Client B2C

Genre et catégories des livres achetés



H0 Les catégories d'articles achetées ne dépendent pas du genre des clients.

H1 Les deux variables sont dépendantes.

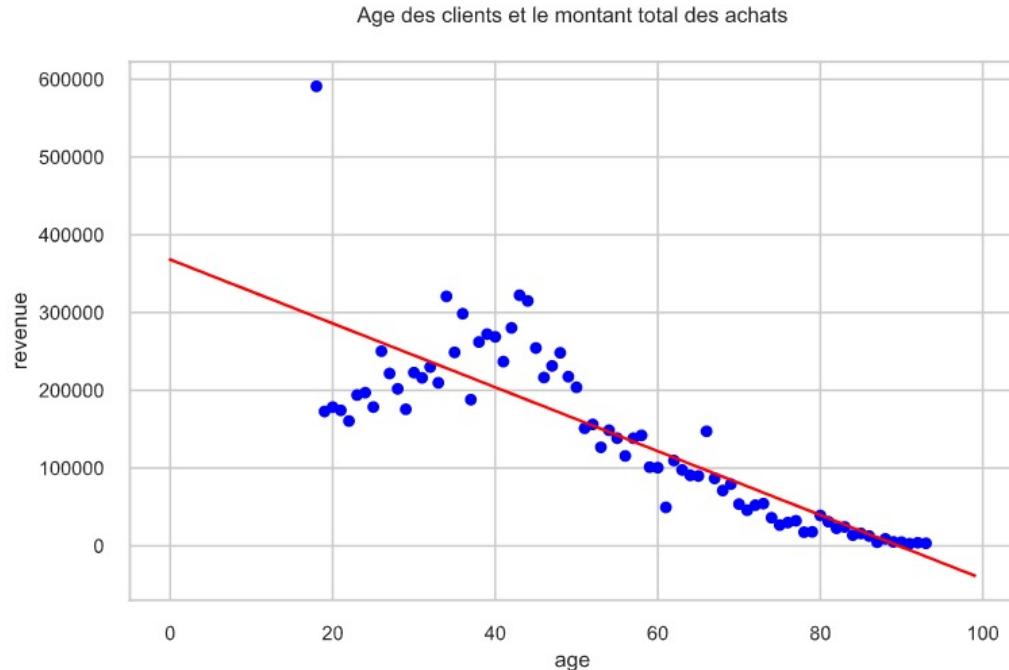
Résultat :

L'hypothèse H0 est rejetée en faveur de l'hypothèse H1.

L'association entre les variables est statistiquement significative.

Il existe donc un lien entre le sexe des clients et les catégories produits.

Age des clients et le montant total des achats



Résultat :

Coefficient de Pearson : -0.83

Covariance : -1976447.96

Modèle de régression linéaire :

r-square : 0.69

p-value : 1.4764e-20

Seuil de significativité : 0.05

R²: les points sont donc peu dispersés par rapport à la courbe.
p-value est faible, les résultats sont statistiquement significatifs.

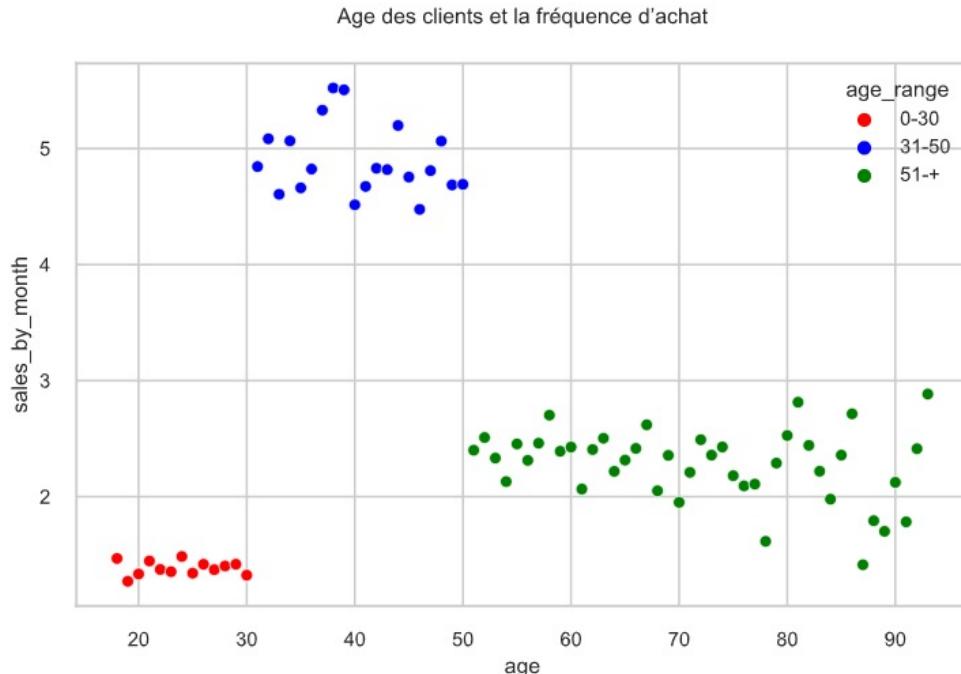
On rejette donc l'hypothèse H0.

H0 Les montants d'achats ne dépendent pas de l'âge des clients.

H1 Les deux variables sont dépendantes.

Les clients dont l'âge est inférieur à 50 ans ont des montants d'achats plus élevés sur notre site. Il est intéressant d'étoffer notre offre pour cette tranche d'âge pour augmenter notre chiffre d'affaire.

Age des clients et la fréquence d'achats



On observe 3 groupes d'individus

H0 La fréquence d'achats ne dépend pas de l'âge des clients.

H1 Les deux variables sont dépendantes.

Résultat ANOVA:

Eta squared : 0.95

p-value : 4,279e-50

Seuil de significativité : 0.05

Eta-2: proche de 1 :

Les moyennes par classes sont très différentes, au sein d'une même classe les valeurs sont peu dispersées.

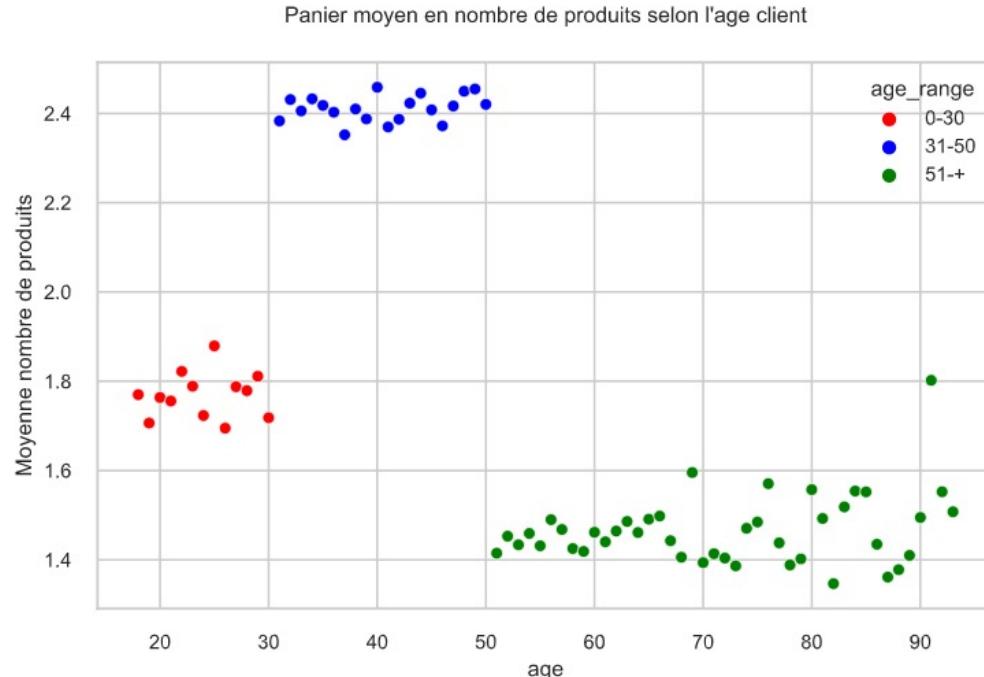
Il existe donc une relation entre les tranches d'âges et les fréquences d'achats.

On rejette donc l'hypothèse H0.

Il est important d'avoir une stratégie de relance ou promotionnelle pour la tranche d'âge 31-50 ans, en passant notamment par des newsletters ou un programme de fidélisation.

- Les moins de 30 ans -> 1 livre par mois.
- Les 30 à 50 ans -> 5 livres par mois.
- Les plus de 50 ans -> 2 à 3 livres par mois.

Age des clients et la taille du panier moyen



H0 La taille du panier moyen ne dépend pas de l'âge des clients.

H1 Les deux variables sont dépendantes.

Résultat ANOVA :

Eta squared : 0.97

p-value : 2,307e-59

Seuil de significativité : 0.05

Eta-2: proche de 1:

Les moyennes par classes sont très différentes.

Il existe une corrélation forte entre les tranches d'âges et la taille du panier moyen.

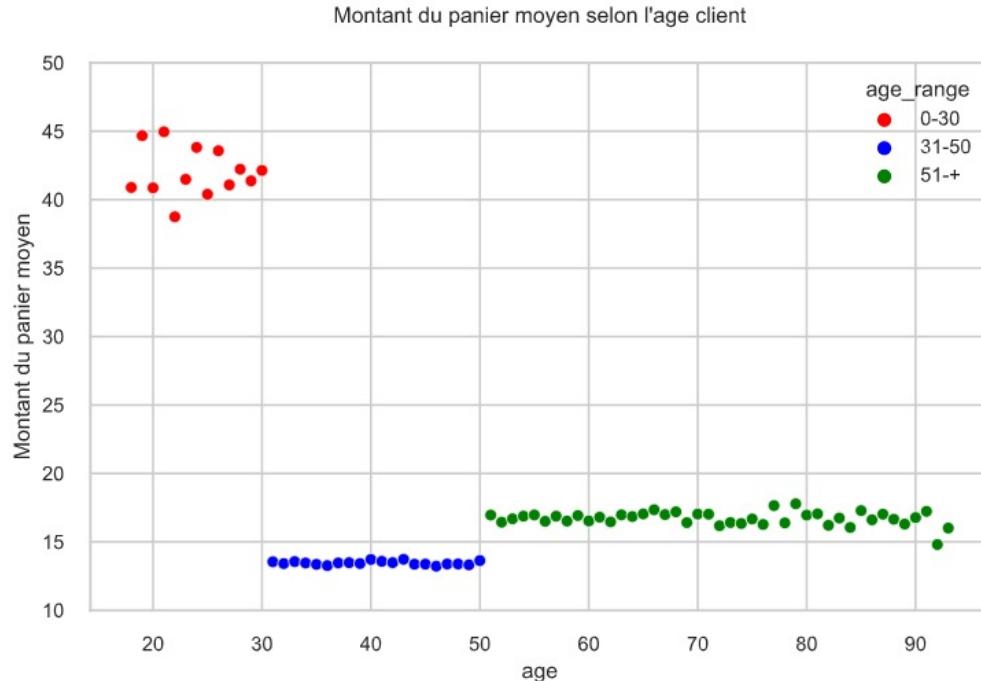
On rejette donc l'hypothèse H0.

Les clients de 31 à 50ans passent de plus grosses commandes.

Nous pouvons leurs faire des recommandation pendant leurs sessions en fonction de leurs panier.

- Les moins de 30 ans achètent en moyenne moins de 2 livres par session.
- Les 30 à 50 ans achètent 2 à 3 livres par session.
- Les plus de 50 ans

Age des clients et le montant du panier moyen



H0 Le montant du panier moyen ne dépend pas de l'âge des clients.

H1 Les deux variables sont dépendantes.

Résultat ANOVA:

Eta squared : 0.9935

p-value : 1,276e-80

Seuil de significativité : 0.05

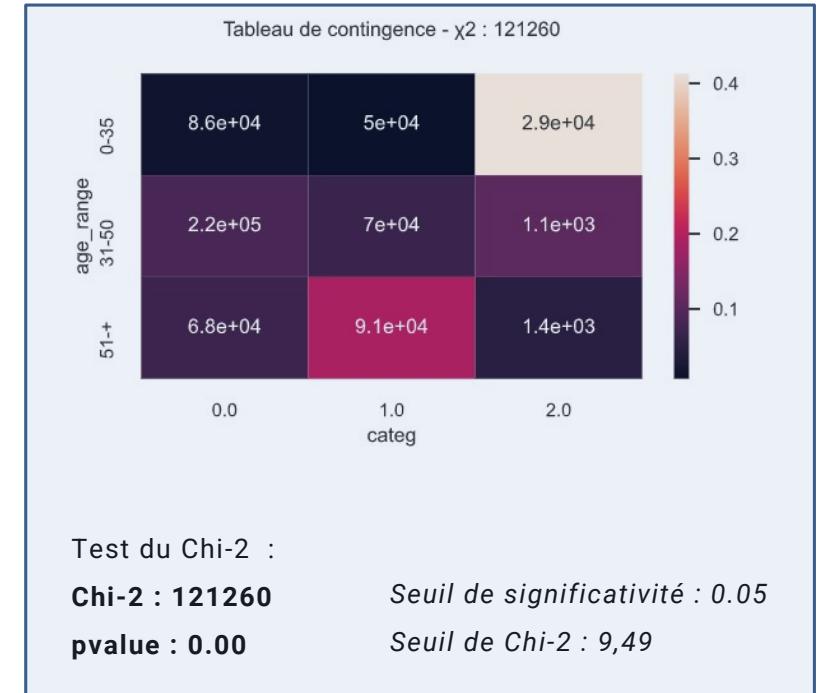
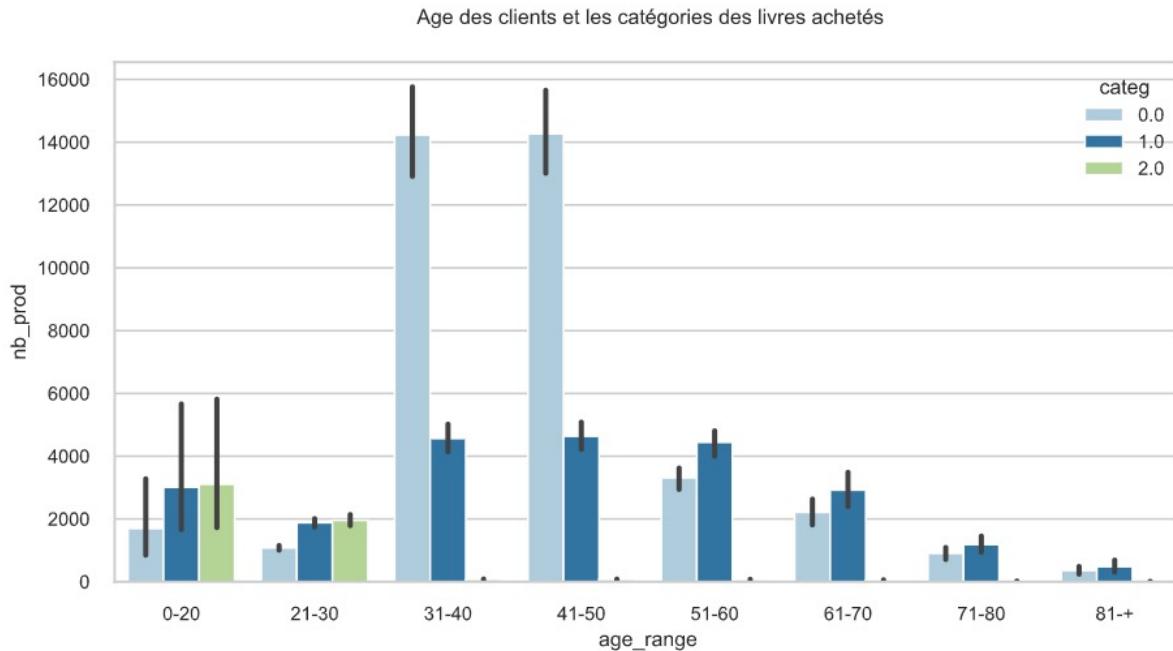
Eta-2: proche de 1:

Il existe une corrélation forte entre les tranches d'âges et le montant du panier moyen.

On rejette donc l'hypothèse H0.

Les clients de 31 à 50ans passent de plus grosses commandes. Le montant de leurs achats restent ce pendant plus bas que les autres tranches d'âges. Les produits qui leurs sont proposés doivent avoir des prix peu élevés.

Age des clients et les catégories des livres achetés



H0 Les catégories achetées sont liées à l'âge des clients.

L'hypothèse H0 est rejetée en faveur de l'hypothèse H1. L'association entre les variables est statistiquement significative.

H1 Les deux variables sont dépendantes.

Il existe donc un lien entre l'âge des clients et les catégories produits.

The background image shows a dimly lit library or bookstore. On the left, several tall wooden bookshelves are filled with books, their spines visible in the low light. In the center-right, three glowing incandescent lightbulbs hang from the ceiling by wires. The right side of the image is heavily blurred, creating a bokeh effect with bright, out-of-focus lights.

Relation entre deux ouvrages

Relation entre deux ouvrages



On cherche à connaître la probabilité qu'un client achète la référence 0_525 sachant qu'il a acheté la référence 2_159.

$$p_{B|A}(A) = p(A \cap B) / p(A)$$

Nombre de clients ayant achetés la référence 2_159 : 514

Nombre de clients ayant achetés la référence 0_525 : 450

Nombre de clients ayant achetés à la fois la référence 2_159 et la référence 0_525 : 450

Nombre total des clients : 8598

Le nombre de clients ayant achetés à la fois les deux références et la référence 0_525 sont les mêmes.

Les probabilités sont égales : $p(A \cap B) = p(A)$

On a donc : $p_{B|A}(A) = p(A) / p(A)$

• Probabilité qu'un client achète la référence 2_159 :

0.05978

• Probabilité qu'un client achète la référence 2_159 :

0.05233

Résultat :
87,54 %

CONCLUSION

CHIFFRE D'AFFAIRE :

- Chiffre d'affaire sur l'année fiscale 2022 : linéaire
- Baisse février 2023 (-13.3%) = baisse globale du nombre de vente.
- Top catégories ventes : 0 et 1.
- B2B : presque 7% du chiffre d'affaire.

CORRELATION :

- Genre : corrélation confirmée entre le sexe des clients et les catégories produits.
- Age : Les corrélations montrent que l'âge est un caractère qui influe sur les ventes en 3 tranches d'âge :

	Montant total des achats	Fréquence d'achat	Taille panier moyen	Montant panier moyen	Catégories achetées
< 31 ans	++	+	++	+++	1 et 2
31 à 50 ans	+++	++++	+++	+	0 et 1
> 50 ans	+	++	+	+	1

A group of students are gathered around a table, working together on a project. They are looking at papers and discussing their work. The scene is set in a classroom or workshop environment.

fin.

AVEZ-VOUS DES QUESTIONS ?