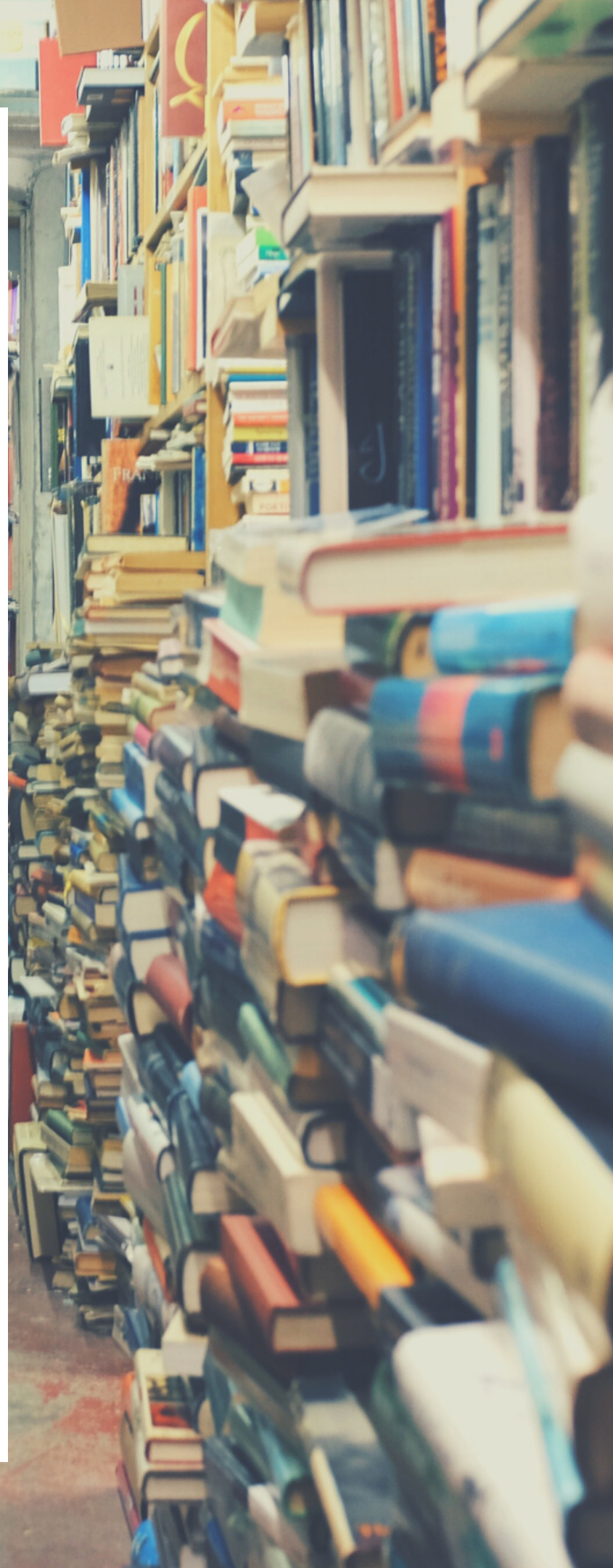
Lapage

ANALYSE DES VENTES

Mars 2021 - Fevrier 2023



Sommaire

Descriptions des données

1.

1.1 Trois sources de données

- **Les Transactions**

3267 produits distincts répertoriés

Clés primaire : Id_prod

200 valeurs aberrantes sur les dates

Pas de valeurs manquantes

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

- **Les Produits**

Les produits sont partagés en 3 categories (0, 1 et 2)

731 produits avec un prix aberrant (valeur negative)

Pas de valeurs manquantes

id_prod	price	categ
0_1421	19.99	0
0_724	21.78	0
0_1467	4.99	0
0_1076	25.11	0
0_2211	9.99	0

- **Les clients**

8623 clients Hommes et femmes de 17 à 92 ans

Clés primaire : client_id

Pas de valeurs manquantes

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

1.2 Jointures

Jointure table Transactions - Produits

21 produits invendus

d_prod	date	session_id	client_id	price	categ
0_310	NaT	NaN	NaN	1.94	0.0
0_322	NaT	NaN	NaN	2.99	0.0
0_1620	NaT	NaN	NaN	0.80	0.0
0_1025	NaT	NaN	NaN	24.99	0.0
0_510	NaT	NaN	NaN	23.66	0.0

Produit 0_2245 : pas de prix
pas de categorie

id_prod	session_id	client_id	price	categ
0_2245	s_272266	c_4746	NaN	NaN
0_2245	s_242482	c_6713	NaN	NaN
0_2245	s_306338	c_5108	NaN	NaN
0_2245	s_76493	c_1391	NaN	NaN
0_2245	s_239078	c_7954	NaN	NaN

Jointure 3 tables

21 clients inactifs

id_prod	date	session_id	client_id	price	categ	sex	birth
NaN	NaT	NaN	c_8253	NaN	NaN	f	2001
NaN	NaT	NaN	c_3789	NaN	NaN	f	1997
NaN	NaT	NaN	c_4406	NaN	NaN	f	1998
NaN	NaT	NaN	c_2706	NaN	NaN	f	1967
NaN	NaT	NaN	c_3443	NaN	NaN	m	1959

Suivi de la qualité des données

2.

2.1 Les données dupliquées

200 sessions de tests à supprimer du dataset

id_prod	date	session_id	client_id
T_0	test_2021-03-01 02:30:02.237437	s_0	ct_1
T_0	test_2021-03-01 02:30:02.237419	s_0	ct_0
T_0	test_2021-03-01 02:30:02.237412	s_0	ct_1
T_0	test_2021-03-01 02:30:02.237419	s_0	ct_0
T_0	test_2021-03-01 02:30:02.237443	s_0	ct_1

- 'date' : prefixe test_
- 'Id_prod' : T_0
- 'session_id' : s_0
- 'client_id' : ct_0 , ct_1
- 'price' : -1

2.2 Les données manquantes

Prix et categorie manquantes pour le produits 0_2245

d_prod	date	session_id	client_id	price	categ
0_2245	2022-09-23 07:22:38.636773	s_272266	c_4746	10.63	0.0
0_2245	2022-07-23 09:24:14.133889	s_242482	c_6713	10.63	0.0
0_2245	2022-12-03 03:26:35.696673	s_306338	c_5108	10.63	0.0
0_2245	2021-08-16 11:33:25.481411	s_76493	c_1391	10.63	0.0
0_2245	2022-07-16 05:53:01.627491	s_239078	c_7954	10.63	0.0

- 221 lignes => 0,03 % du dataset
- Valeurs de remplacement
  categorie : 0 (prefixe du prodiut)
  prix : 10,63 (moyenne de la categorie 0)

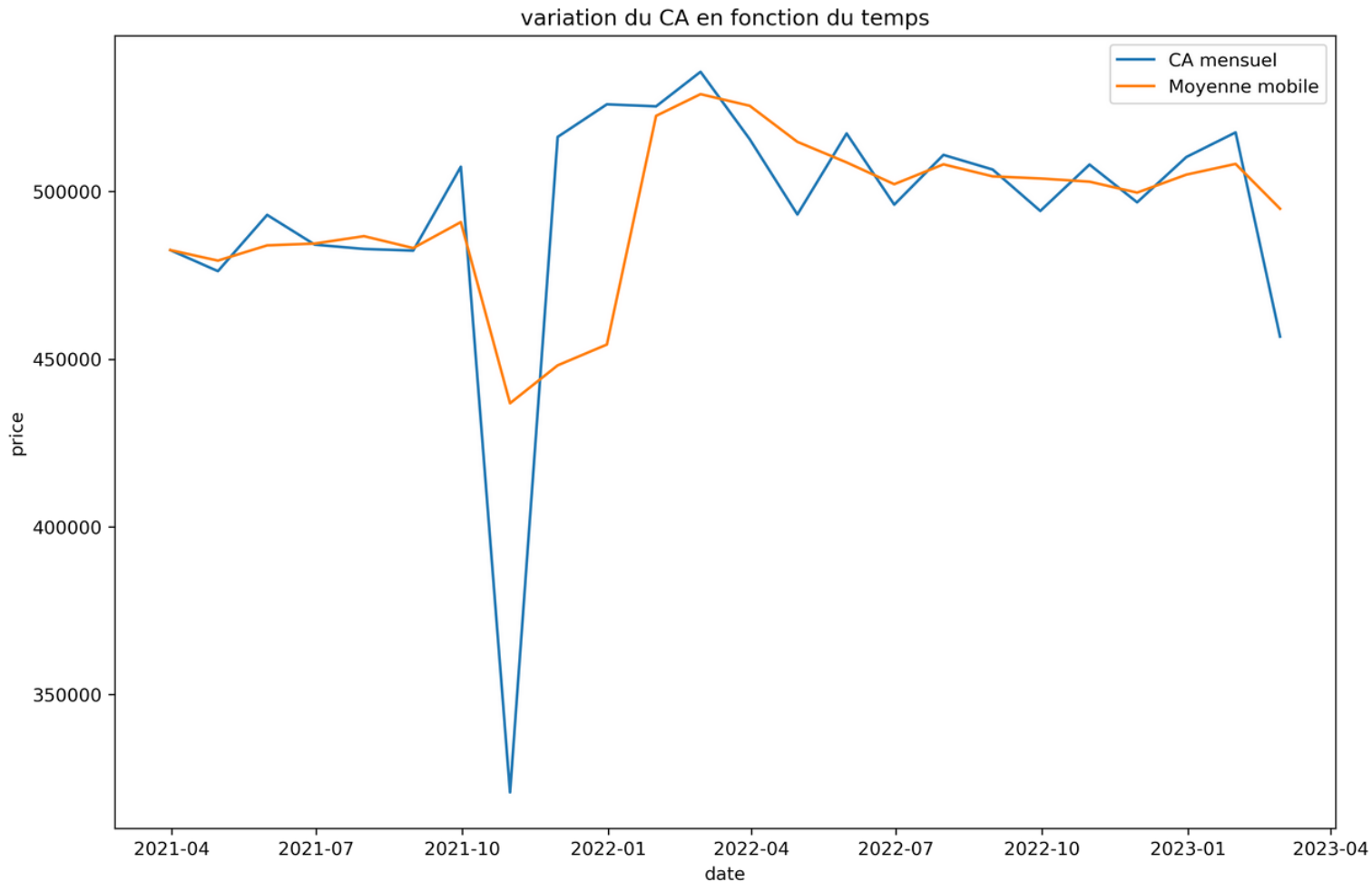
Les données des clients inactifs et des produits invendus sont supprimés du dataset aussi

Analyse des données

3.

3.1 Analyse autour du CA

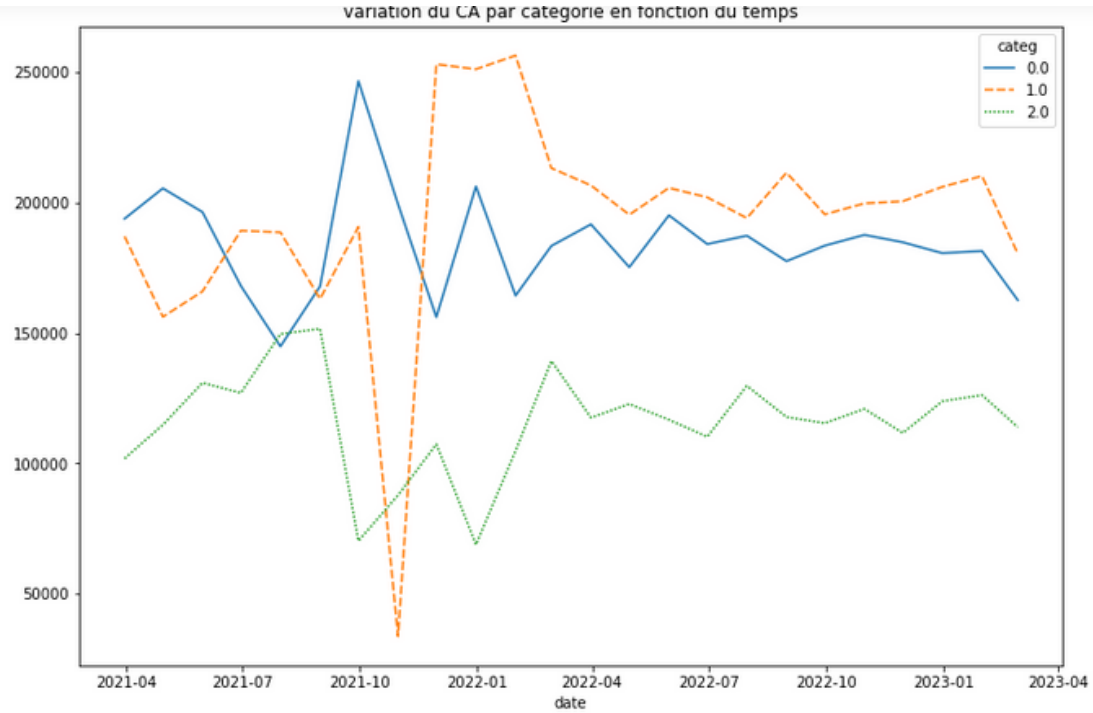
La période de vente est du 2021-03-01 au 2023-02-28

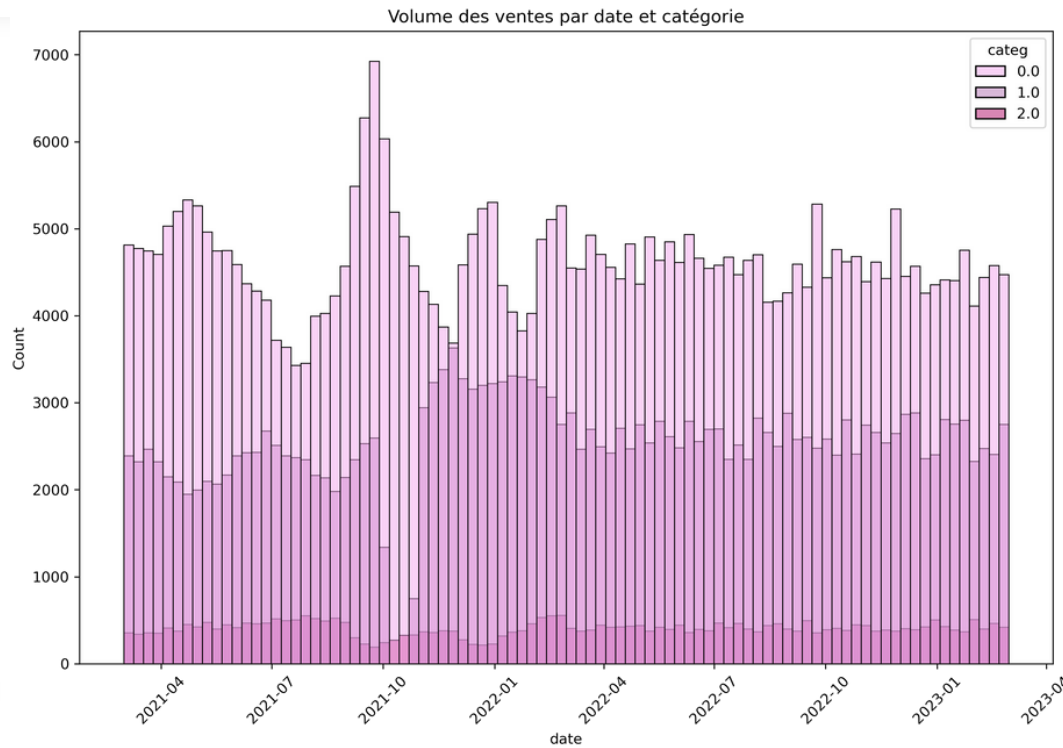


- Nous avons la courbe d'évolution du CA au cours du temps ainsi que la courbe de la moyenne mobile à 3 valeurs permettant de souligner les tendances à long terme.
- Forte baisse du chiffre d'affaire est constatée durant le mois d'octobre 2021

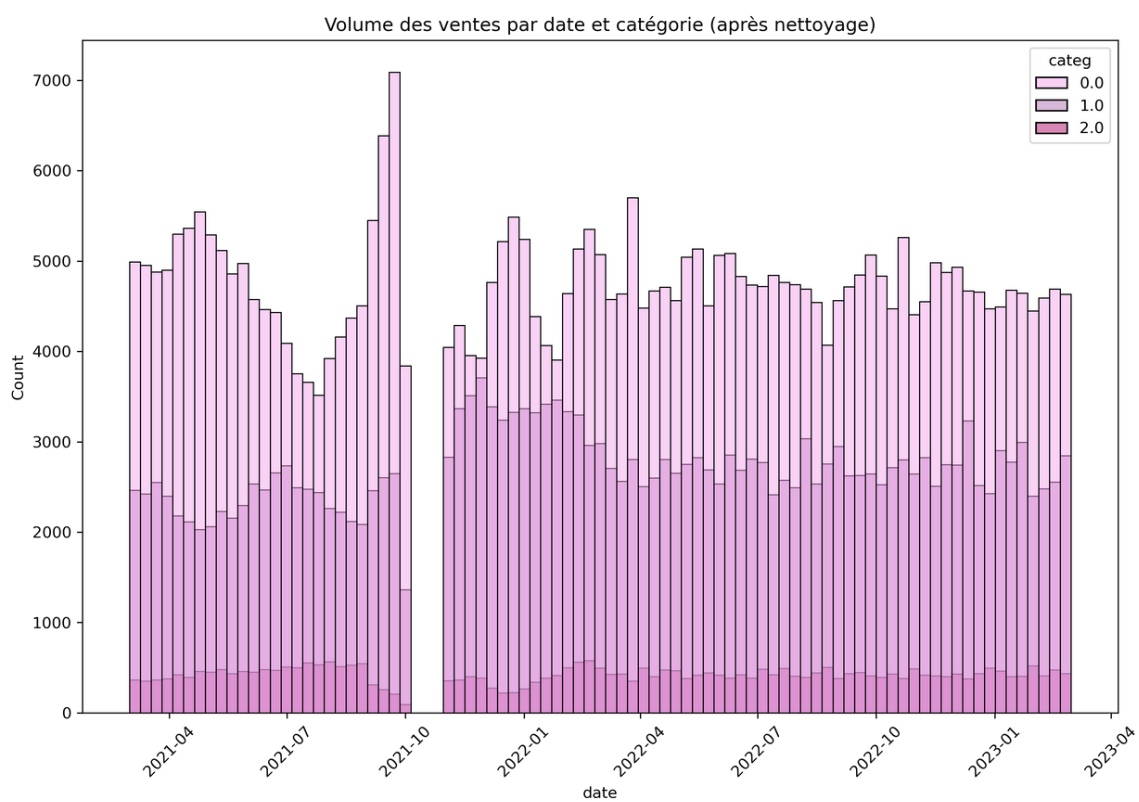
Regardons en détails les ventes par date et catégorie pour comprendre cette variation du CA.

Données manquantes



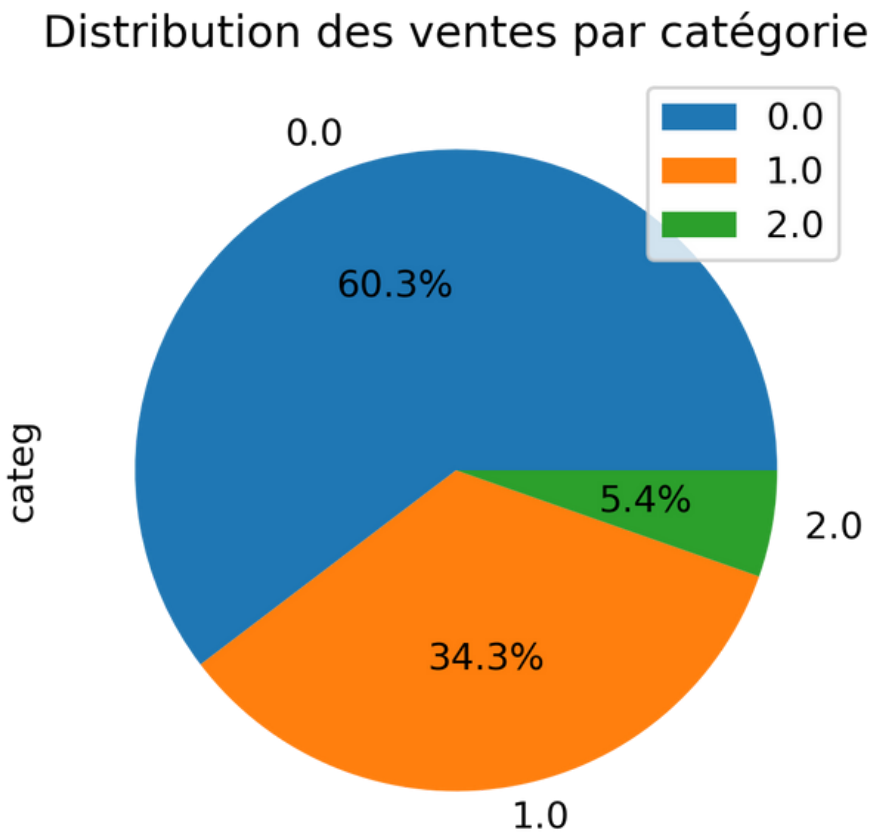


- Anomalie au mois d'octobre 2021.
- Données catégorie 1 manquantes
- Ces données représentent 3,04% du dataset



- Suppression des données du mois d'octobre 2021 pour ne pas fausser l'analyse

Distribution des ventes par catégorie



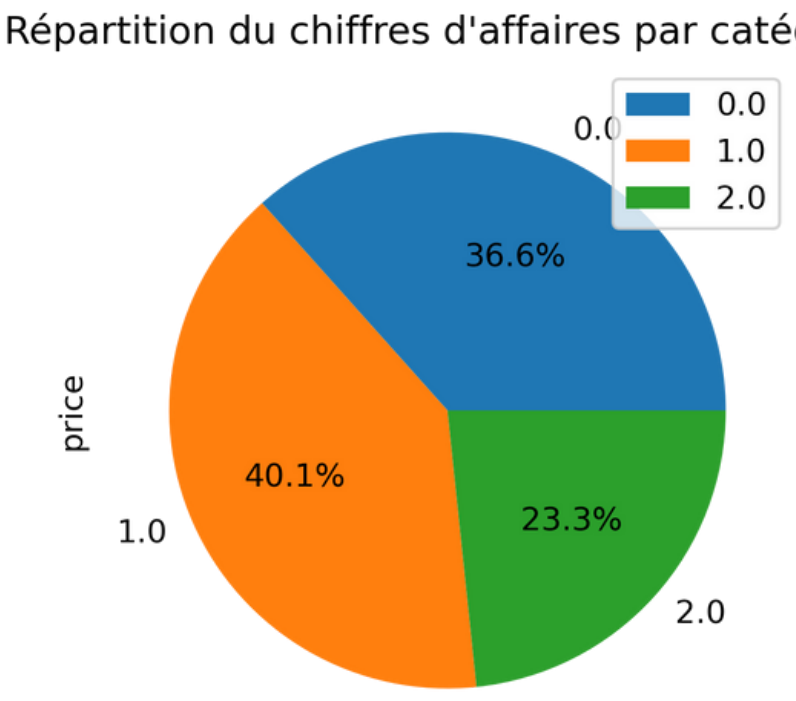
- La catégorie 0 représente 60% des ventes, la catégorie 2 et 1 seulement 5% et 34%.

- Malgré tout, nous remarquons que les catégories 1 et 2 ont une influence sur le chiffre d'affaires annuel.

Catégorie	0	1	2
Prix minimum	0.62	2.0	30.99
Prix médian	9.99	19.08	62.83
Prix maximum	40.9	80.99	300.00

Nous remarquons que les catégories 2 et 1 sont plus chères que la catégorie 0.

Distribution du CA par catégorie



La catégorie 0 malgré un grand nombre de ventes (61%), ne représente qu'un tiers du chiffre d'affaires contrairement à la categorie 1 qui représente presque 40% du chiffre d'affaires pour une volume de ventes de 34%.

Cette première analyse nous permet de dire que le chiffre d'affaires est corrélé à la categorie de livre.

Les Tops et les Flops

Les 10 tops

	id_prod	categ	CA Total
0	2_159	2.0	92265.68
1	2_135	2.0	67472.22
2	2_112	2.0	62840.10
3	2_102	2.0	59080.86
4	2_209	2.0	55502.07
5	1_395	1.0	54095.34
6	1_369	1.0	53761.59
7	2_110	2.0	51916.50
8	1_414	1.0	51711.10
9	1_383	1.0	51225.33

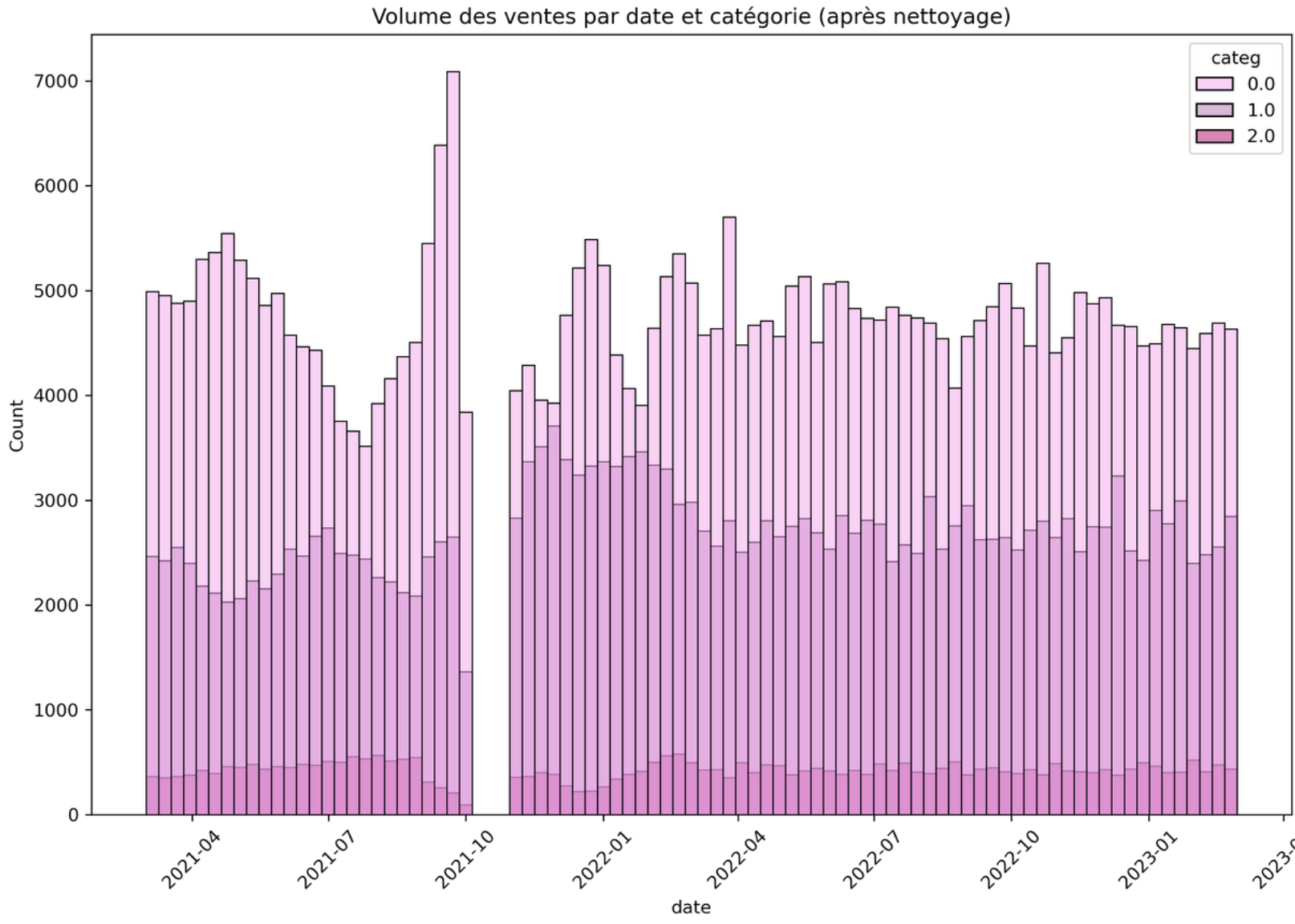
	id_prod	categ	nbre ventes Total
0	1_369	1.0	2241
1	1_417	1.0	2174
2	1_414	1.0	2170
3	1_498	1.0	2119
4	1_425	1.0	2087
5	1_403	1.0	1952
6	1_412	1.0	1942
7	1_406	1.0	1934
8	1_413	1.0	1933
9	1_407	1.0	1925

Les 10 flops

	id_prod	categ	CA Total
0	0_1539	0.0	0.99
1	0_898	0.0	1.27
2	0_1284	0.0	1.38
3	0_1653	0.0	1.98
4	0_643	0.0	1.98
5	0_1601	0.0	1.99
6	0_541	0.0	1.99
7	0_807	0.0	1.99
8	0_1728	0.0	2.27
9	0_324	0.0	2.36

Les 10 livres avec le plus gros chiffre d'affaires annuel appartiennent à la catégorie 1 et 2.
La categorie 1 compte plus de ventes.

la periodicité des ventes par catégorie



La périodicité des ventes apparaît corrélée à la catégorie

- **Catégorie 0 : meilleurs vente (Sept-Ocotbre) correspond à la rentrée scolaire de septembre**
- **Catégorie 1 : meilleurs vente (Dec-Janvier) correspond à la periode des fêtes de fin d'année**
- **Catégorie 2 : meilleurs vente (Juillet-aout) pendant l'été et au mois de février**

3.2 Typologie des clients

Pour faciliter notre l'analyse, quelques variables supplémentaires sont créées dans le dataset :

total_ventes	ventes_mensuelles	taille_panier_moyen	panier_moyen	total_achats	age	classe_age
181	8.0	2.479452	11.978051	2118.62	36	30-40
181	8.0	2.479452	11.978051	2118.62	36	30-40
181	8.0	2.479452	11.978051	2118.62	36	30-40
181	8.0	2.479452	11.978051	2118.62	36	30-40
181	8.0	2.479452	11.978051	2118.62	36	30-40

- La frequence d'achats mensuelle par client
- Le nombre de ventes total par client sur les 2 années
- Le panier moyen
- La taille du panier moyen pour chaque client
- Le chiffre d'affaires total par client sur les 2 années
- L'age du client
- La classe d'age

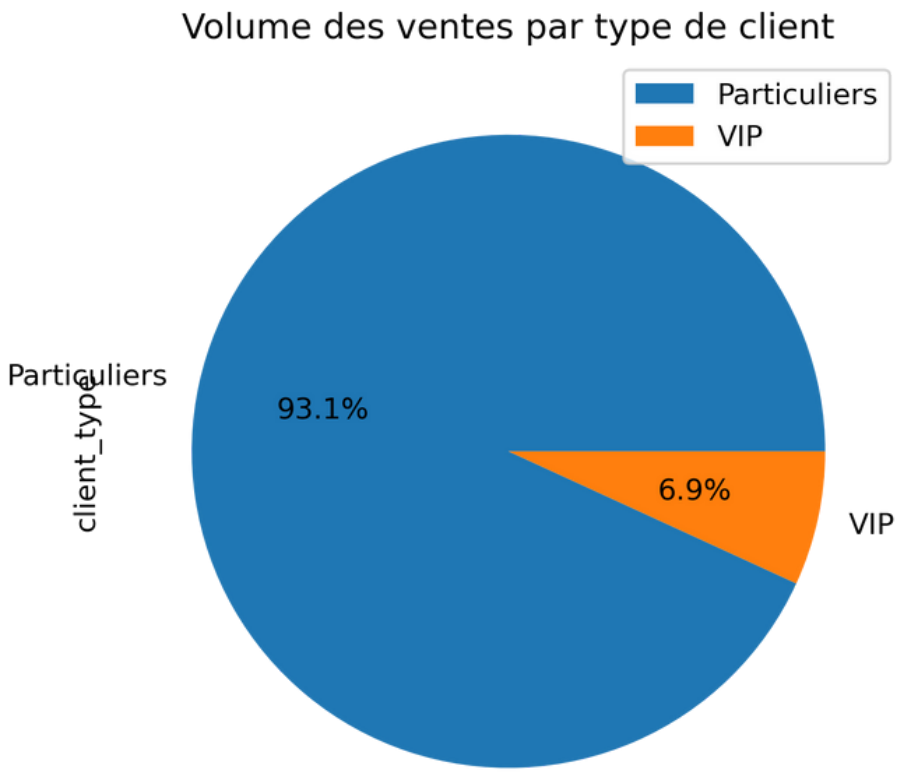
Les différents type de clients

client_id	panier_moyen	taille_panier_moyen	total_achats	total_ventes	ventes_mensuelles
c_1609	12.770388	34.991429	313044.92	24494	1065.0
c_4958	55.461723	7.321839	282959.96	5096	222.0
c_6714	16.855326	13.195556	149938.74	8907	387.0
c_3454	16.680603	9.487143	111903.05	6641	289.0
c_2899	55.754703	1.640625	5214.05	105	5.0
c_1570	14.958710	2.679104	5166.45	359	16.0
c_3263	13.237727	3.015385	5129.89	392	17.0
c_7319	13.472283	2.746269	5120.55	368	16.0
c_5263	58.905308	1.507692	5006.85	98	4.0
c_8026	13.747661	2.830769	4980.04	368	16.0

- 4 clients qui se détachent du lot. Leur nombre d'achats est largement plus élevé que celui des autres clients : on a sûrement affaire à des clients VIP.

Nous allons regarder la répartition du chiffre d'affaires entre ces différents clients

3.3 Répartition du CA entre les clients

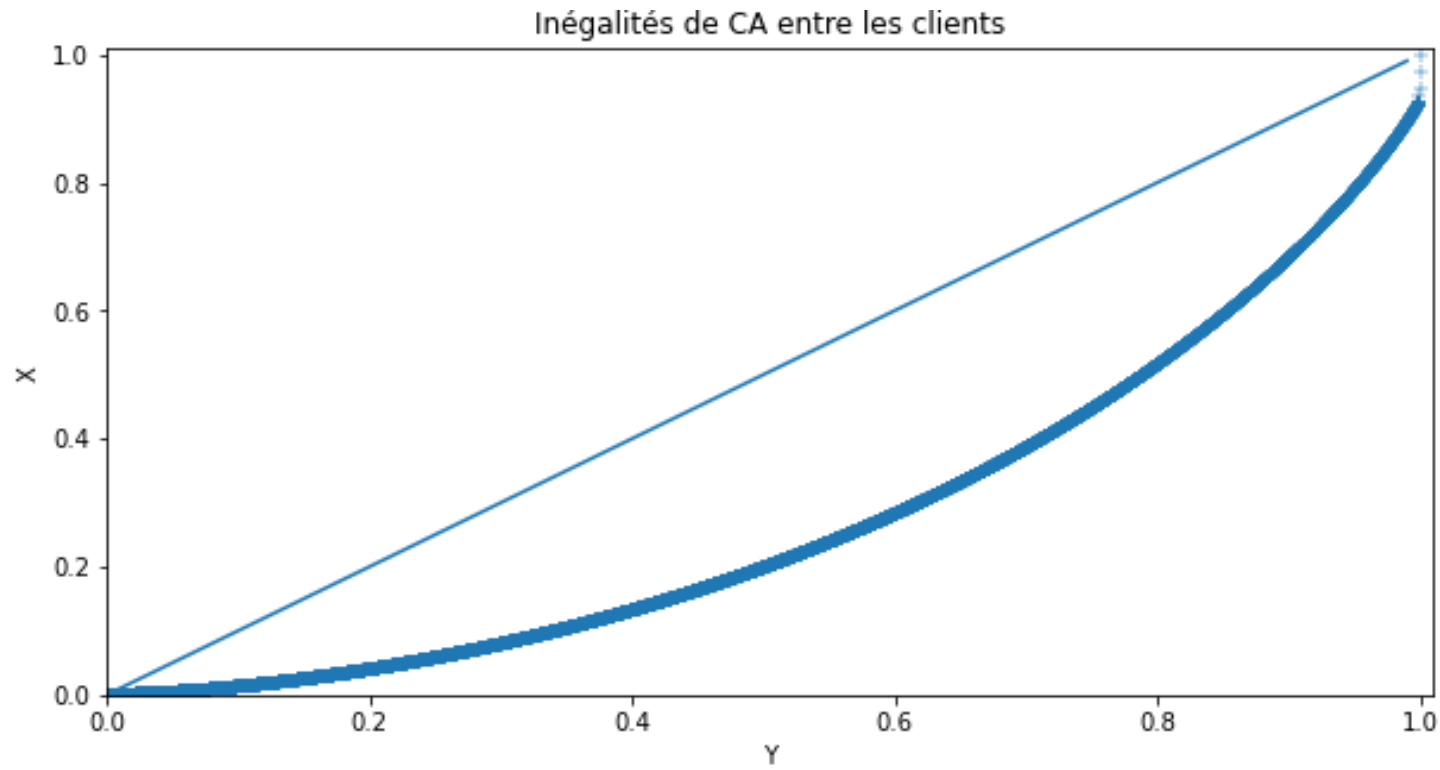


- Les clients VIP
  - 6.9% des ventes
  - 7.43% du CA

- Les clients particuliers
  - 93% des ventes
  - 92.57% du CA

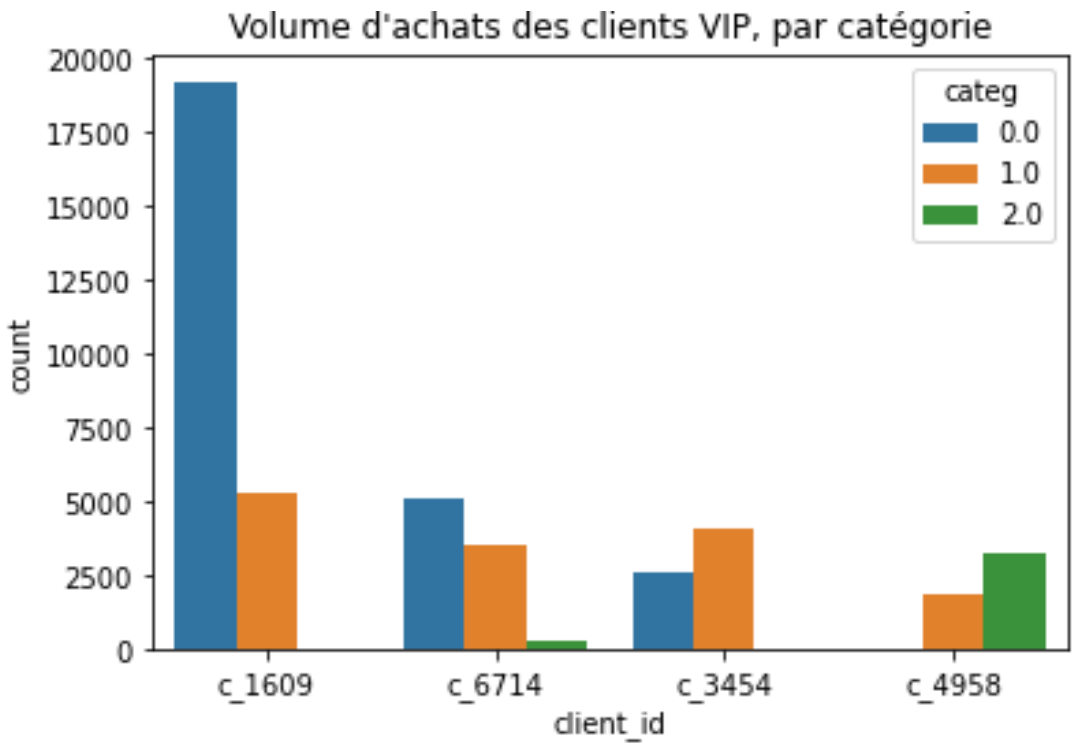
Nous constatons une inégalité du chiffre d'affaires confirmer par la courbe de Lorenz avec une indice de Gini de 0.44

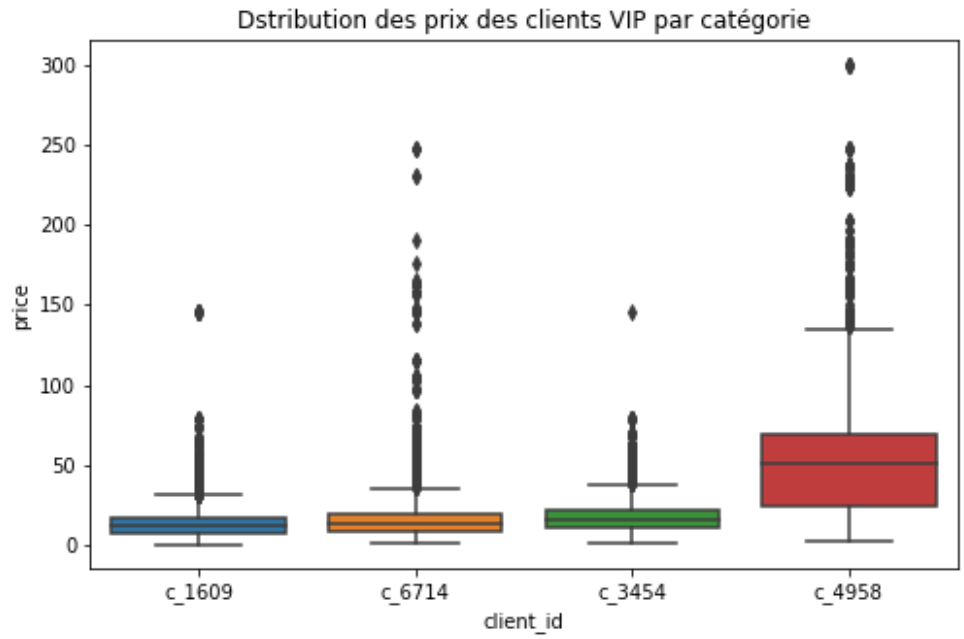
Incide de Gini : 0.447



Zoom sur les clients VIP

client_id	sex	age	birth	categ	mois	panier_moyen	price	taille_panier_moyen	total_achats	total_ventes	ventes_mensuelles
c_1609	m	42	1980	0.216584	6.459541	12.770388	12.780474	34.991429	313044.92	24494	1065.0
c_3454	m	53	1969	0.612860	6.405812	16.680603	16.850331	9.487143	111903.05	6641	289.0
c_4958	m	23	1999	1.630887	6.331633	55.461723	55.525895	7.321839	282959.96	5096	222.0
c_6714	f	54	1968	0.463231	6.388908	16.855326	16.833809	13.195556	149938.74	8907	387.0



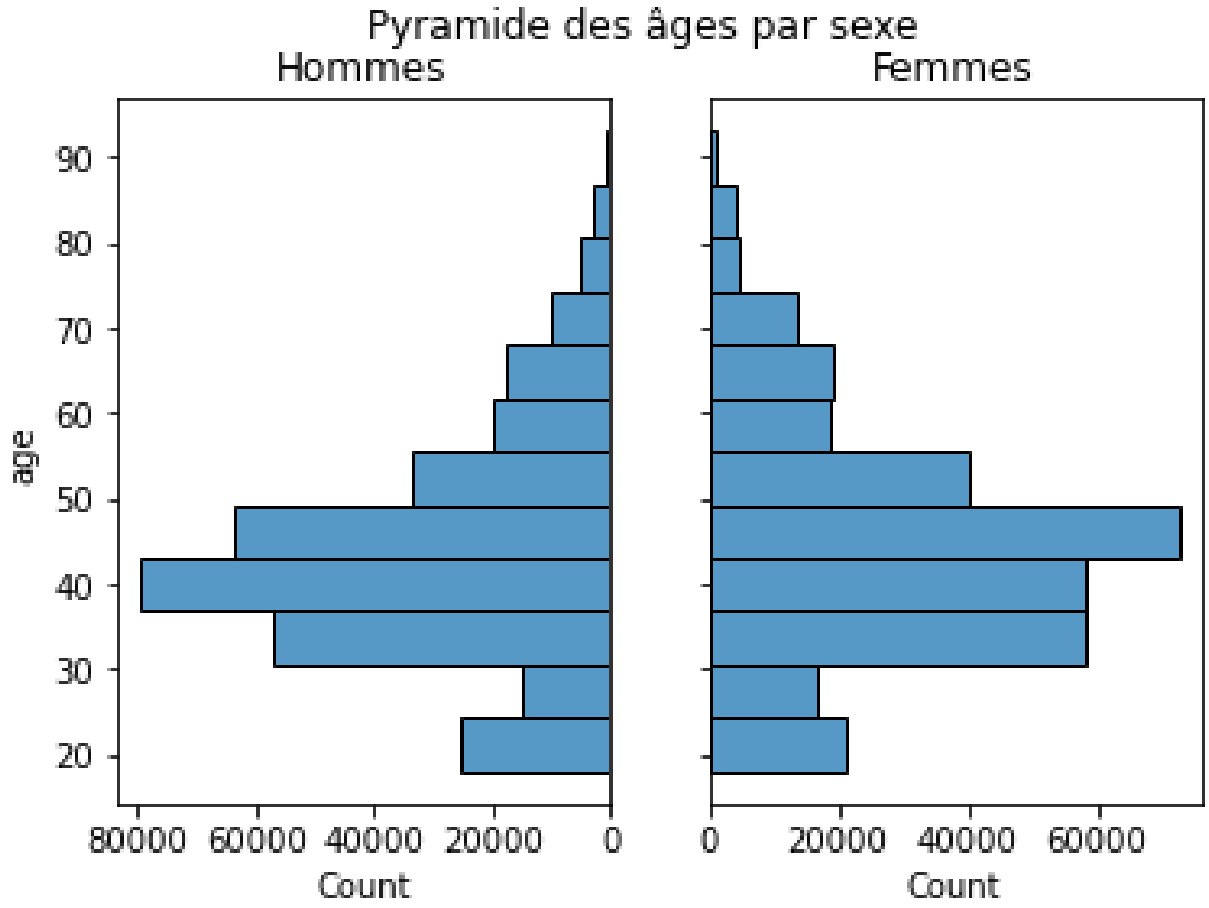


Le volume d'achats par client et la distribution des prix par catégorie de livre nous permet de confirmer que le client c_4958, principal consommateur de la catégorie 2, achète des livres plus chérs que les autres.

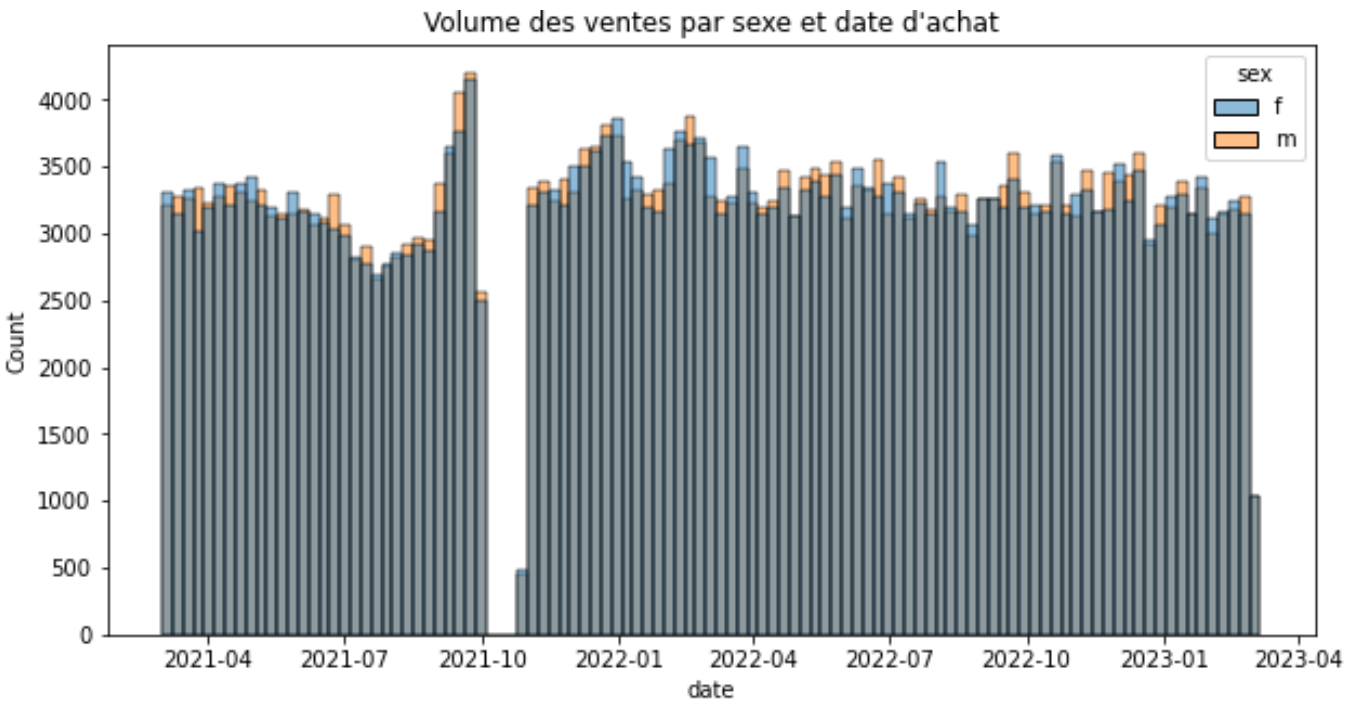
Le prix est corrélé à la catégorie des livres achetées :

- **Catégorie 0 : prix et panier moyen faibles**
- **Catégorie 2 : prix et panier moyen élevés**

3.4 Liens entre genre, âge et catégorie



- Meme proportion d'hommes et de femmes pour chaque d'age



- La périodicité des ventes est la même pour les hommes et les femmes

Liens entre genre et catégorie

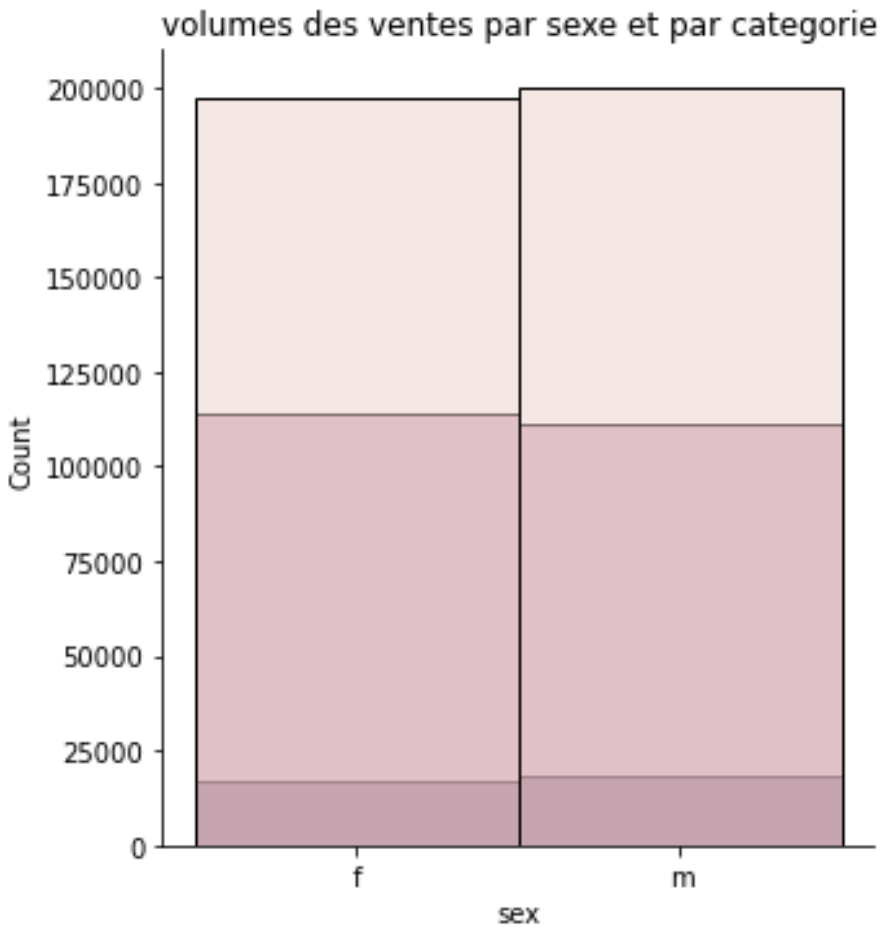


Table de contingence

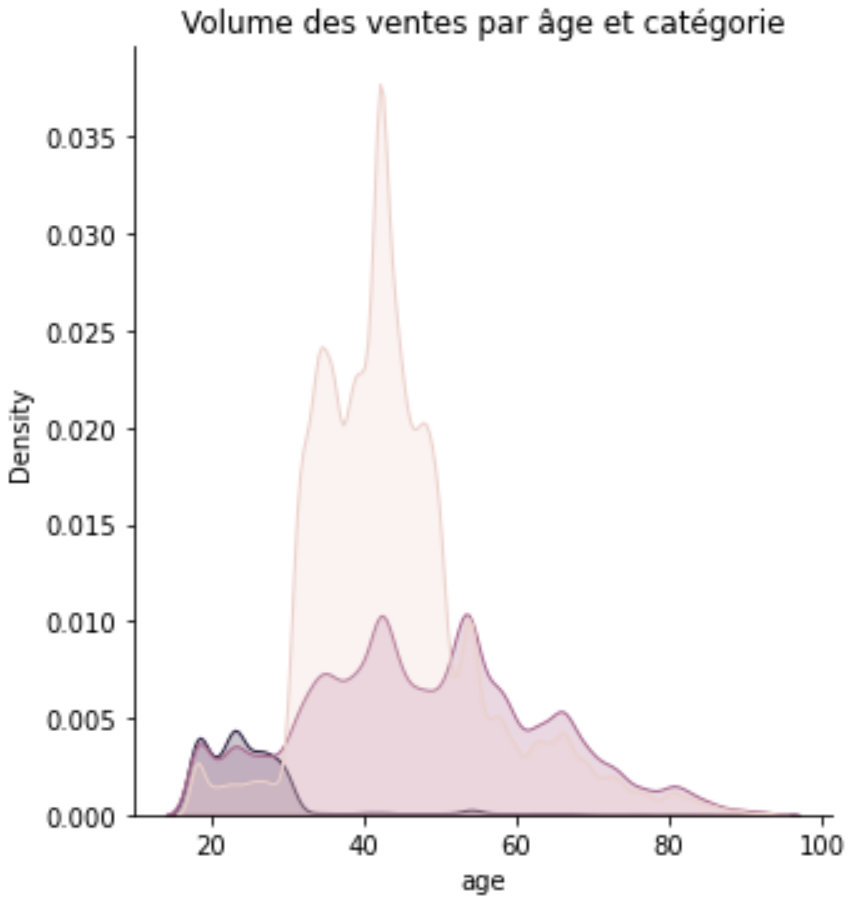
sex	f	m	total
categ			
0.0	197385	200081	397466
1.0	114258	111587	225845
2.0	16746	18607	35353
total	328389	330275	658664

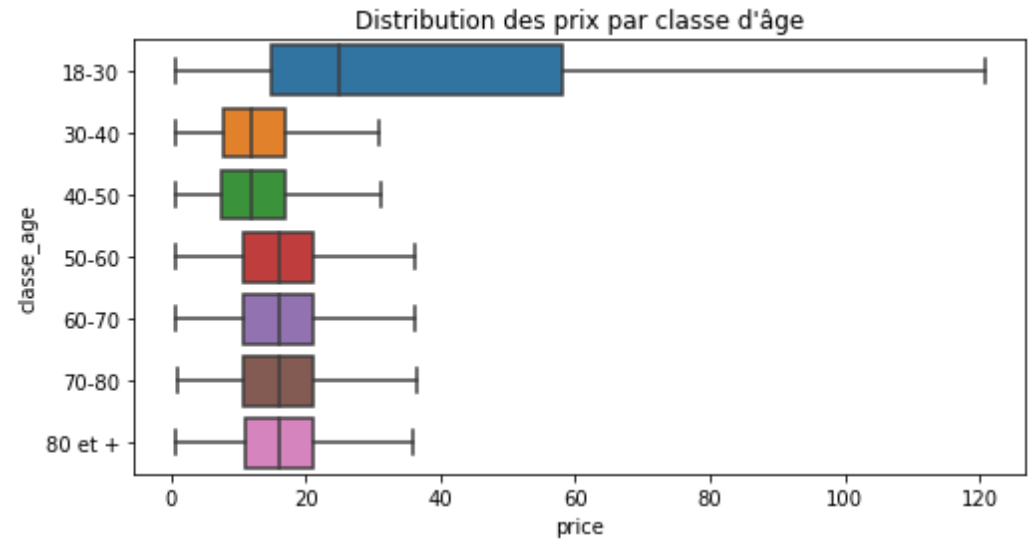
Répartion des ventes en %

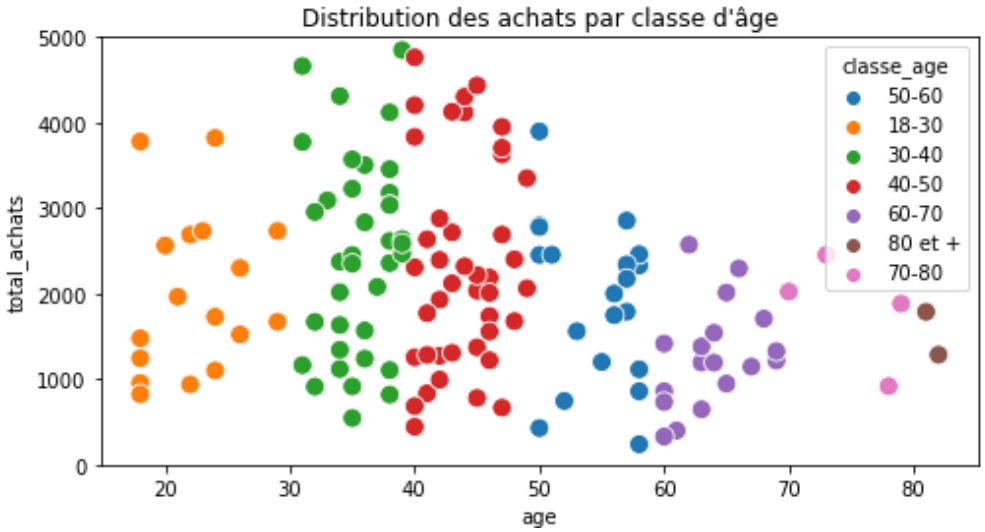
sex	f	m
categ		
0.0	49.66	50.34
1.0	50.59	49.41
2.0	47.37	52.63

Nous constatons quelques différences sur les volumes de ventes par catégorie des hommes par rapport aux volume de ventes des femmes surtout pour la catégorie 2 (environ 5%) qui est la catégorie la plus chére.

Liens entre age et catégorie



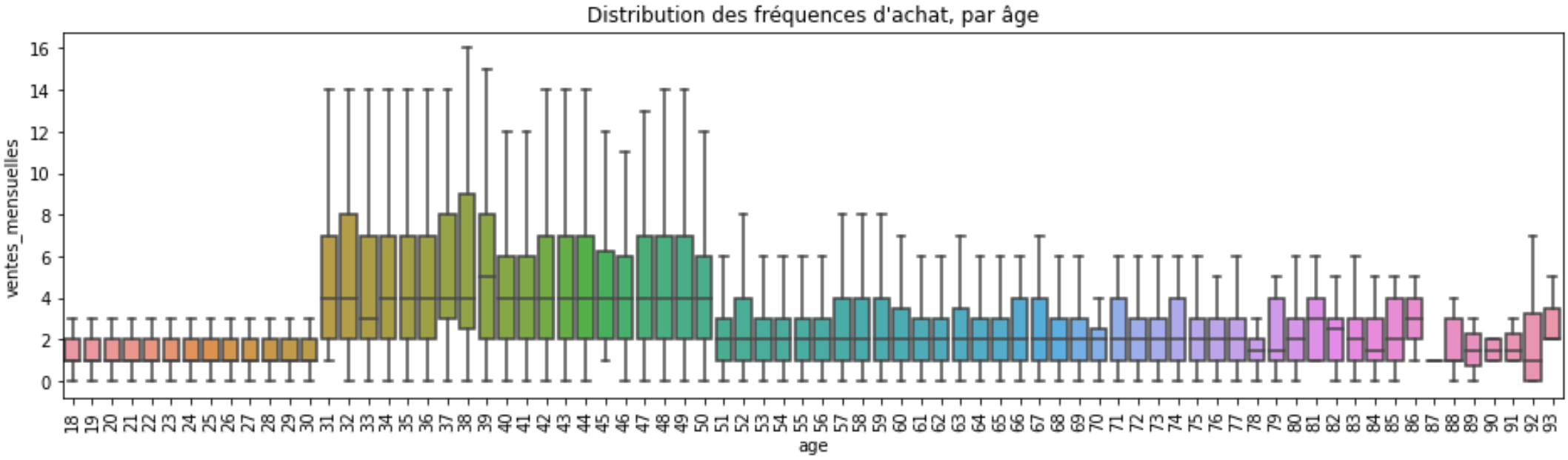




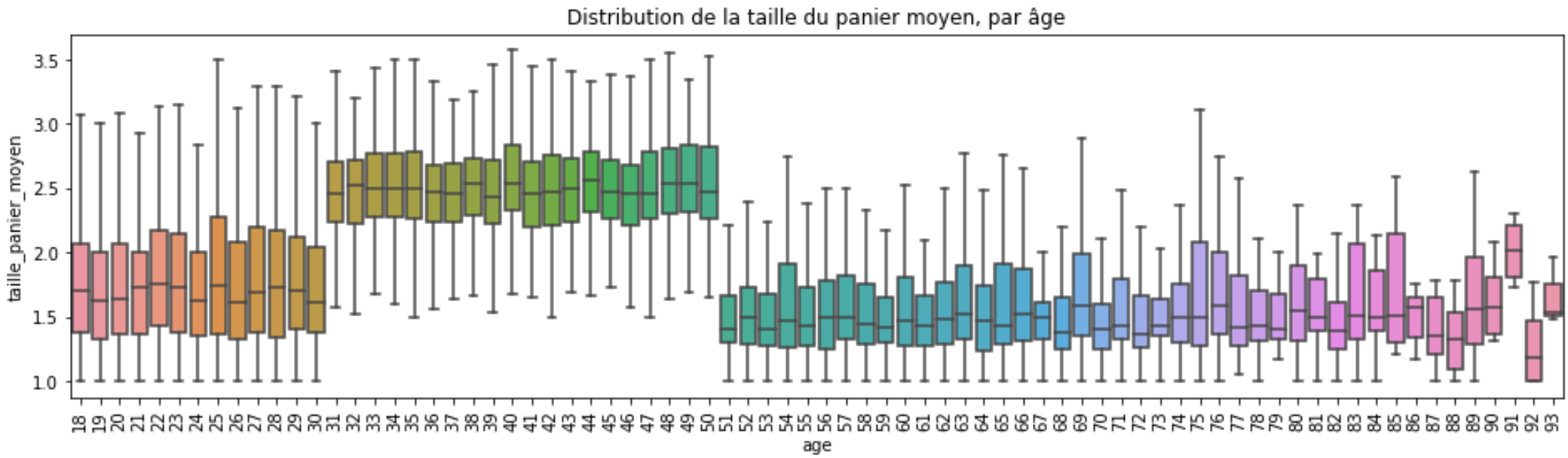
3 groupes de personnes

- Moins de 30 ans
  Consommateur principal de la categ2
  Prix d'achat des livres plus hauts
- 30-50 ans
  Consommateur principal de la categ0
  Prix d'achat des livres plus bas
  Plus grosse volume d'achats
- +50 ans
  Habitudes plus variées

Confirmation des 3 groupes

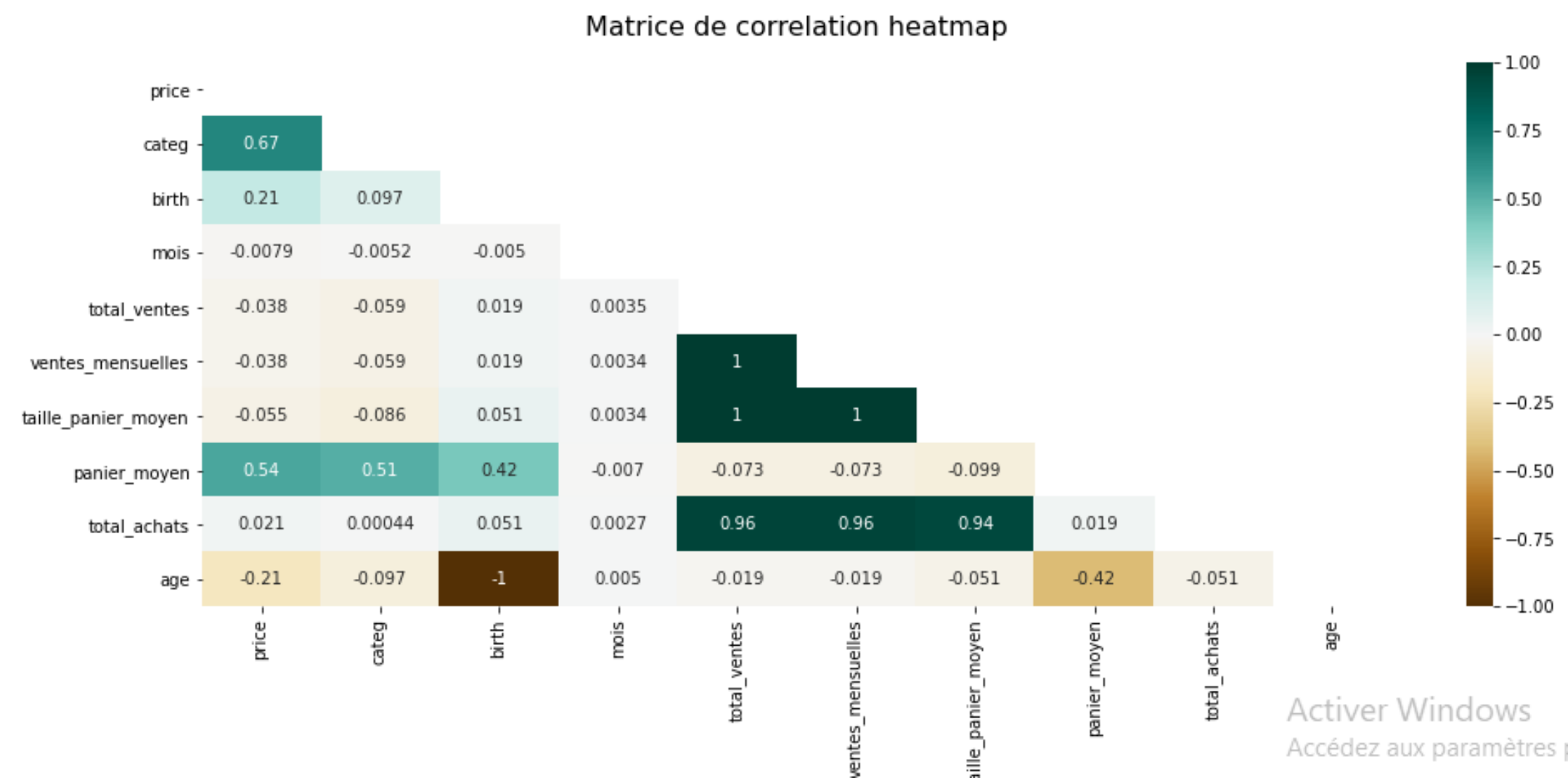


- -30 ans achètent le même nombre de livres (pas plus de 2 livres par mois)
- 30-50 ans : la plus grande partie achète plus de 6 livres par mois.
- +50 ans ont des habitudes plus variées, mais achètent rarement plus de 5 livres



- Les moins de 30 ans comptent en moyenne moins de 2 livres par panier
- Les 30-50 ans achètent 2 à 3 livres par commande
- Les habitudes sont aléatoires chez les plus de 50 ans

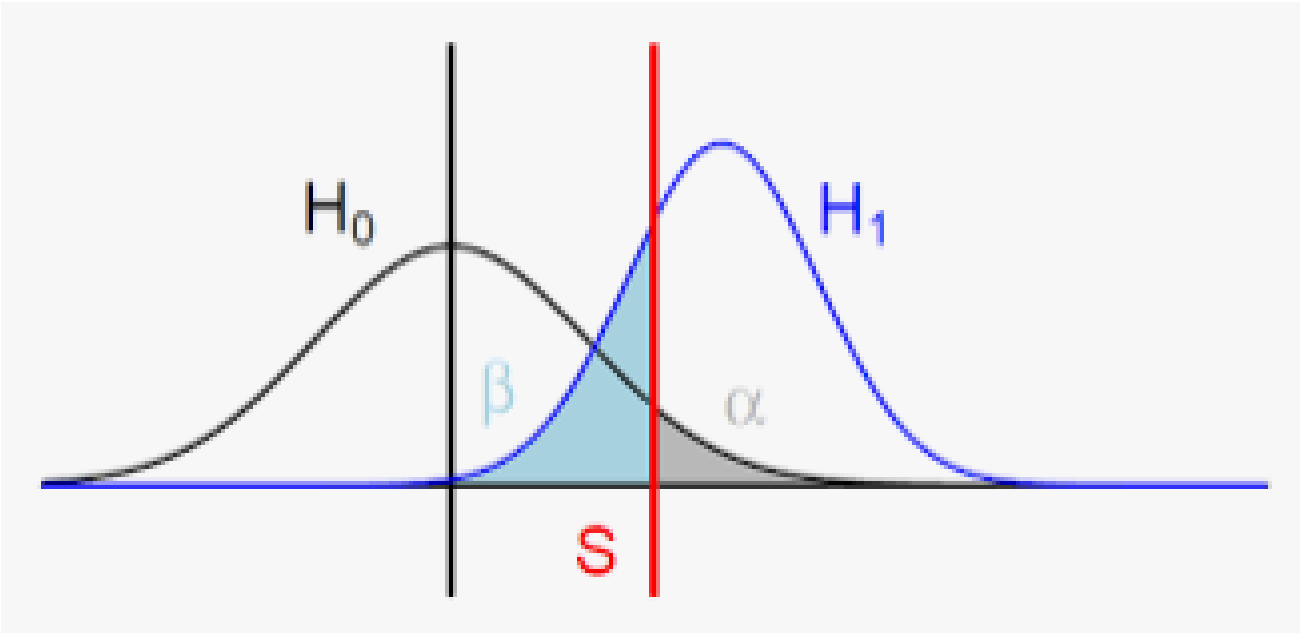
Matrice de corrélation



le coéfficient de corrélation linéaire donne une mesure de l'intensité et du sens de la relation linéaire entre 2 variables

- **Une forte corrélation entre l'âge du client, le prix du livre et le panier moyen. La corrélation est négative parce que plus le prix est haut moins l'age du client est importante.**
- **Corrélation entre l'âge et la catégorie de livre achetée**
- **Corrélation entre la catégorie du livre achetée et le prix**

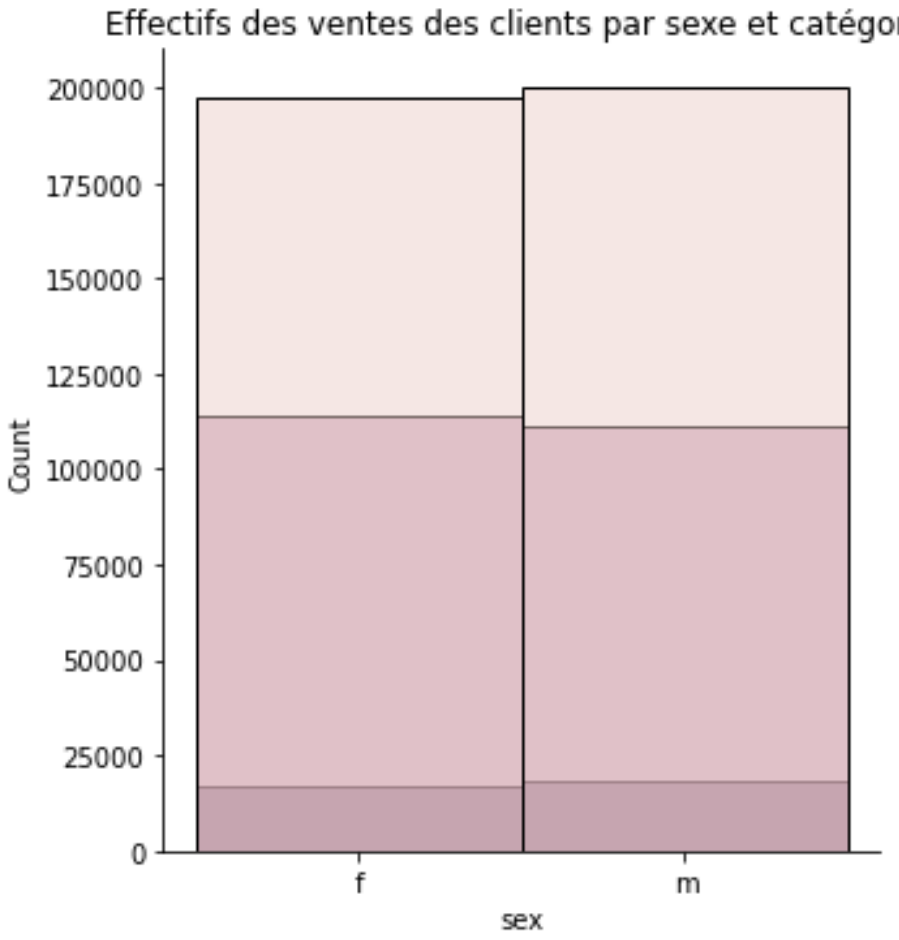
Tests Statistiques

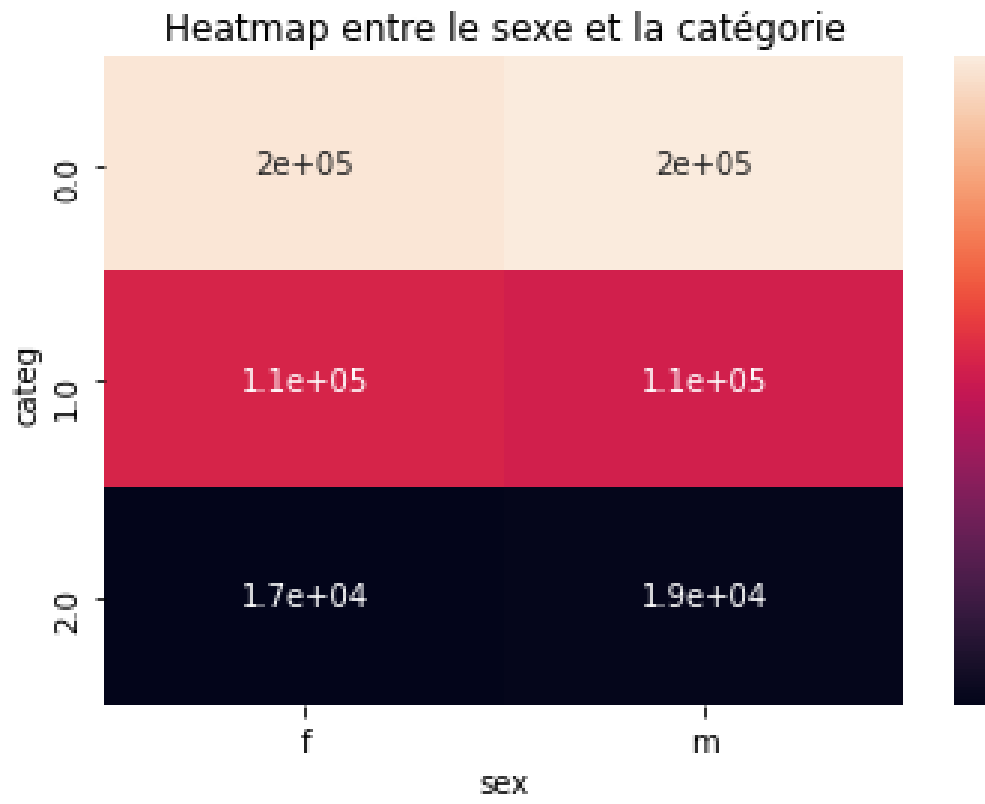


4.

Corrélation entre catégorie et sexe

La statistique du khi-deux est particulièrement adaptée pour les observations qualitatives (2 variables qualitatives)





Hypothèses :

- **H0 = les 2 variables ne sont pas corrélées**
- **H1 = les 2 variables sont corrélées**
- **Seuil de test fixé à 5%**

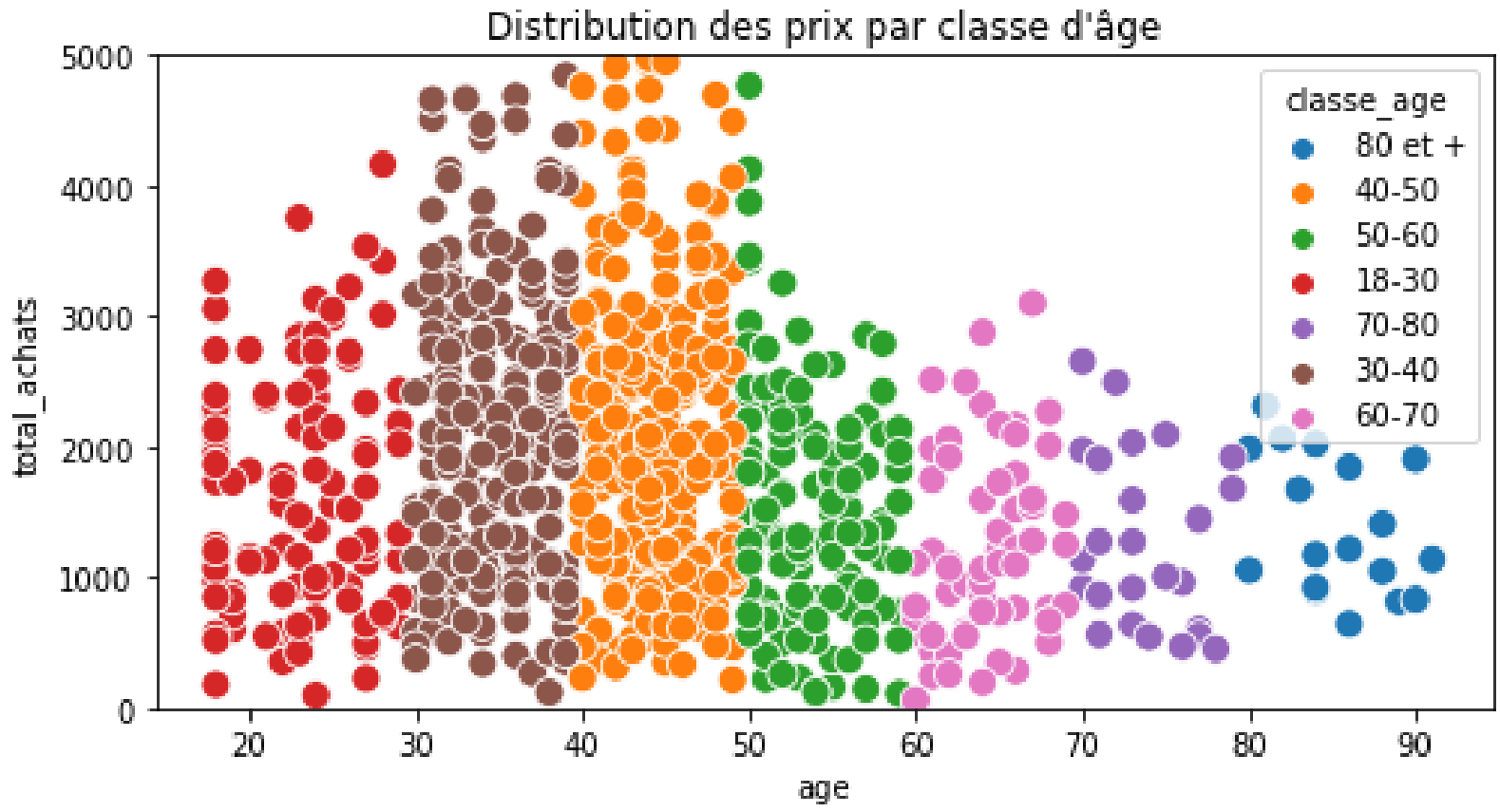
Test de Chi 2 (2 catégorielles)
Stat = 142.441
p-value = 0.11e-29 (< 0,05)
Nombre de degrés de liberté= 2
H1 : les 2 variables sont corrélées.

Corrélation entre âge et CA

Le test statistique de Pearson est utilisée pour voir la corrélation entre 2 variables quantitatives.

Conditions d'application du test de Pearson :

- Ne peut pas être appliqué aux variables ordinales.
- La taille de l'échantillon doit être modérée pour une bonne estimation.
- Les valeurs aberrantes peuvent conduire à des valeurs trompeuses, ce qui signifie non robuste avec les valeurs aberrantes.



Hypothèses :

- **H0 = les 2 variables ne sont pas corrélées**
- **H1 = les 2 variables sont corrélées**
- **Seuil de test fixé à 5%**

Test de Pearsonr

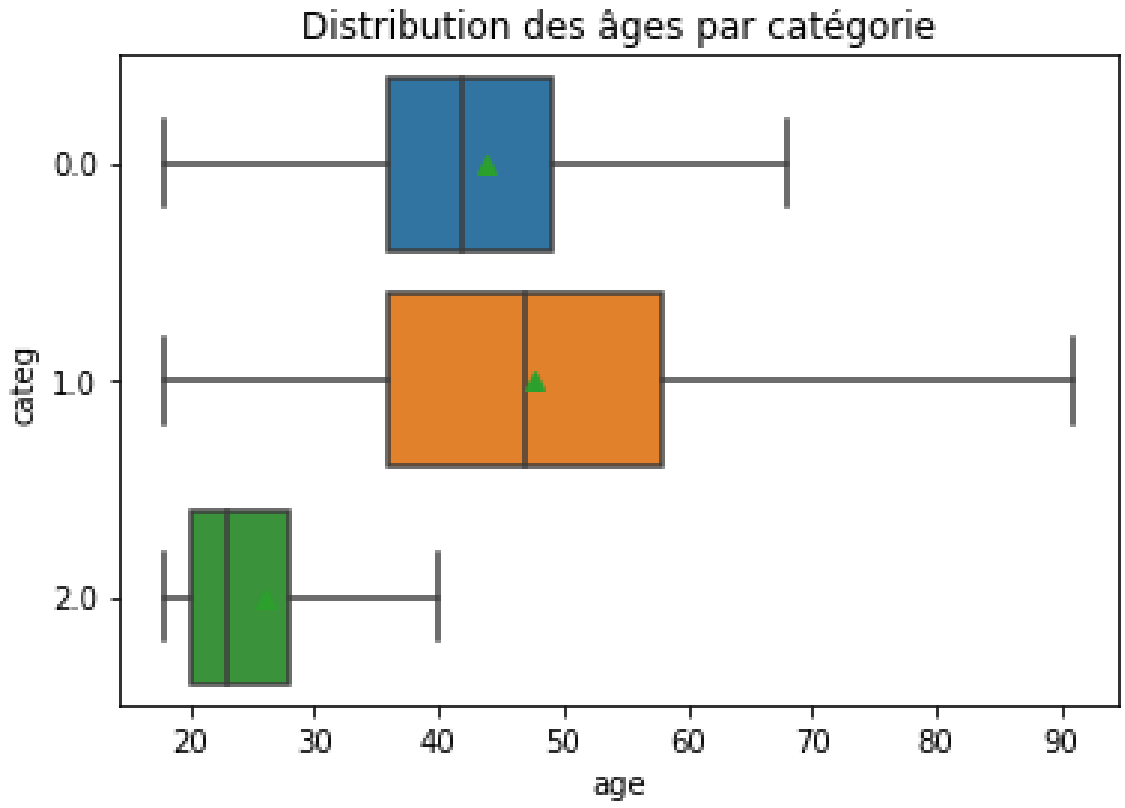
p-value : 9.26e-05

stat : -0.055253605020773 64

H1 : les 2 variables sont corrélées.

Corrélation entre âge et catégorie

Le test statistique ANOVA est utilisé pour comparer deux ou plusieurs moyennes d'un ensemble. Elle est utile pour vérifier la corrélation entre une qualitative (avec plus de deux modalités) et une quantitative.



categ	
0.0	43.902580
1.0	47.649193
2.0	25.940005

La pertinence du test ANOVA repose sur la validation de plusieurs hypothèses

- L'indépendance entre les échantillons de chaque groupe
- L'égalité des variances.
- La normalité des résidus.

Hypothèses :

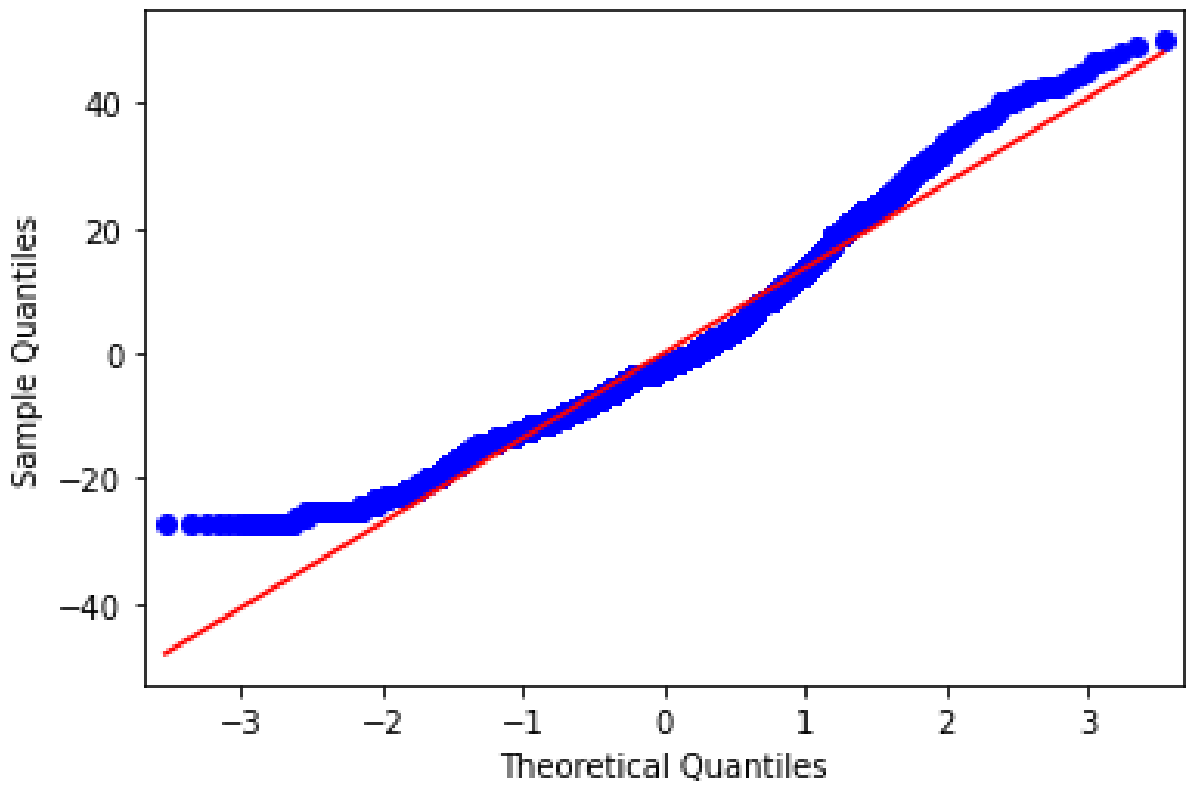
- **H0: Les moyennes de chaque groupe sont égales si p-value > 5%**
- **H1: Les moyennes ne sont pas toutes égales si p-value < 5%**
- **Seuil de test fixé à 5%**

Indépendance des échantillons

On a 3 categorires de livre (categorie0, categorie1, categorie2), on peut dire que les echantillons sont indépendants.

Normalité des résidus

L'objectif est de s'assurer que les résidus suivent une loi normale.
On utilise le test de Shapiro-Wilk pour tester la normalité des résidus.



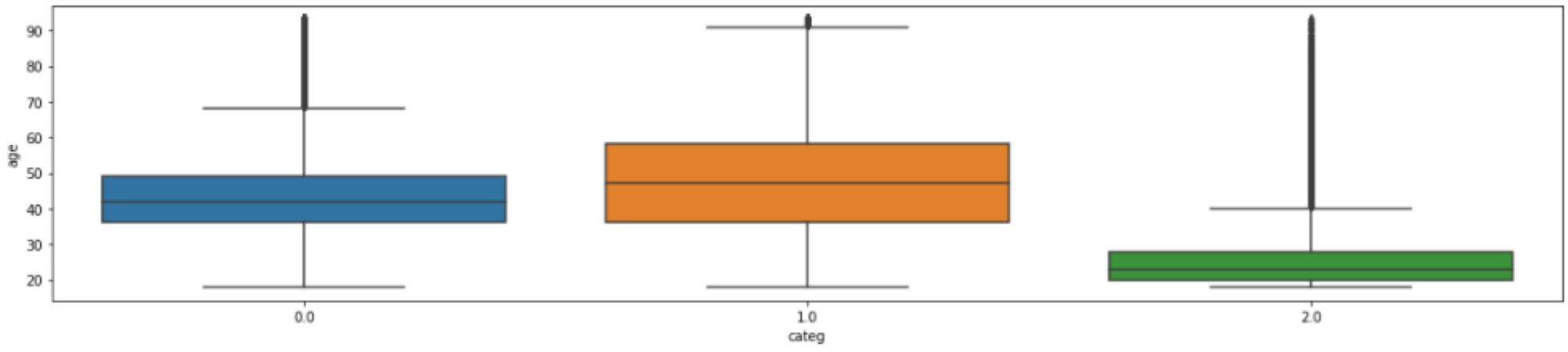
LHypothése :

- **H0: Les résidus suivent une loi normale si p-value > 5%**
- **H1: Les résidus ne suivent pas une loi normale si p-value < 5%**

Normalité des résidus (Shapiro)
stats : 0.9651251435279846
p-value : 3.899125414595082e-33
H0 rejetée : distribution probablement pas normale.

Egalité des variances

On cherche à démontrer que les variances de chaque categorie sont égales



les boites à moustaches montrent une inégalité des variances, mais le test de Bartlett ou Levene permet de tester si les variances sont significativement différentes.

Hypothèses :

- **H0 : Les variances de chaque groupe sont égales si p-value > 5%**
- **H1 : Les variances de chaque groupe ne sont pas toutes égales < 5%**

Egalité des variances (Levene)
stats : 901.0244701169303
p-value : 0.0
H1: les variances ne sont pas égales

Egalité des variances (Bartlett)
stats : 1172.3672530080573
p-value : 2.652e-255
H1: les variances ne sont pas égales

Test de Welch

Dans le cas où la condition d'homoscédasticité n'est pas remplie, on peut utiliser le test de Welch

Hypothèses :

- **H0: les moyennes des échantillons sont égales**
- **'H1: une ou plus des moyennes des échantillons sont inégales**

Test de Welch (si absence d'homoscédasticité)
p-value : 0.0
stat : 227.89736043782594
H1: une ou plus des moyennes des échantillons sont inégales

Conclusion

- **2 profils de clients classés en 3 groupes d'âge**
  Les -30 ans consommateurs des livres les plus chers
  Les 30-50 ans les grands lecteurs
  Les +50 ans aux habitudes plus vairées

- **3 catégories de livre ordonnées par prix et par période de vente**
  Catégorie 0 : moins chère, meilleure vente à la rentrée scolaire
  Catégorie 1 : prix moyen, meilleure vente à la fin de l'année
  Catégorie 2 : plus chère, meilleure vente pendant l'été

- **Corrélation entres les variables**
  Le sexe du client est corrélé à la catégorie du livre.
  L'âge du client est corrélé à la catégorie du livre e et au CA.
  La catégorie du livre est corrélé au prix.

Recommandations:

- Créer un espace en ligne et des offres réservés au client VIP
- Promouvoir la vente de chaque catégorie en fonction de la meilleure période.