Rapport d‘analyse Lapage
(données du 01-03-2021 au 28-02-2023)

Nettoyage des données: remarques et arbitrages

- Suppression des lignes de tests datées du 01-03-2021, liées à des clients fictifs supprimés eux aussi.

- Un produit (réf.: 0_2245) de la table 'transactions' n'a pas de correspondance dans 'products'.

Ce produit a été vendu 221 fois.

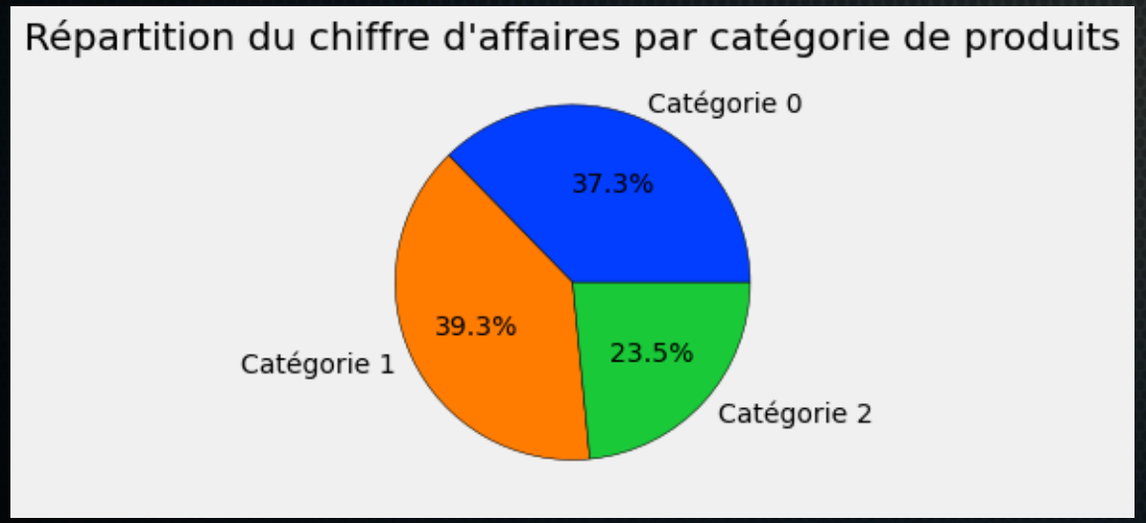
Son prix étant inconnu, il a été décidé d‘imputer la valeur médiane de sa catégorie (cat. 0).

- 21 références produits n‘ont aucune correspondance dans la table 'transactions‘.

Autour du chiffre d‘affaires

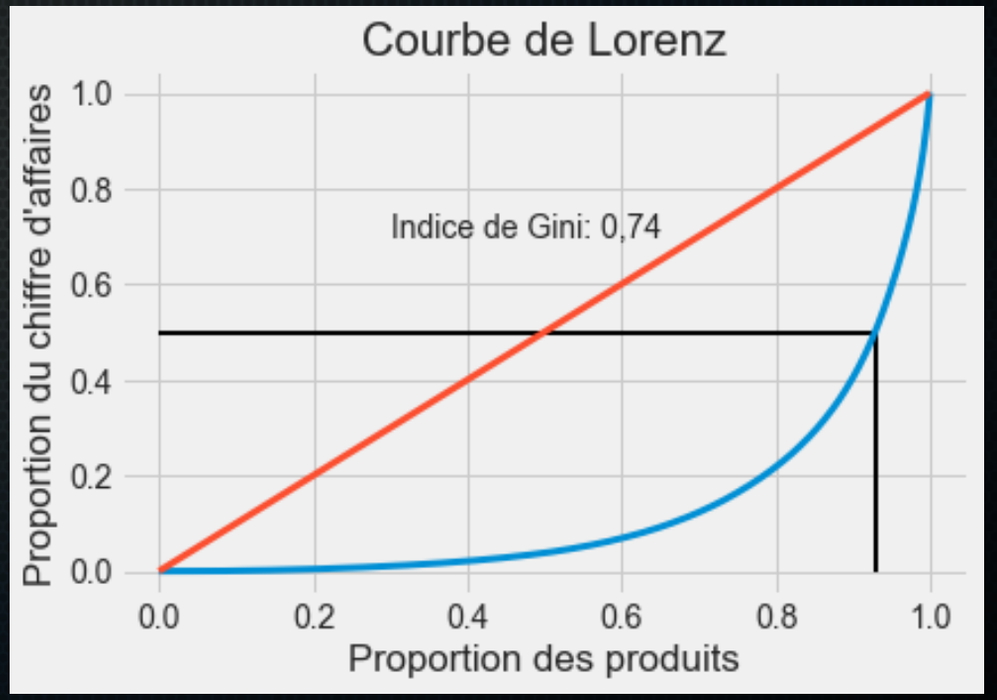
Le chiffre d‘affaires pour la période s‘établit à 11 856 009 €:

- 4 422 011 € pour les produits de la catégorie 0
- 4 653 722 € pour les produits de la catégorie 1
- 2 780 275 € pour les produits de la catégorie 2

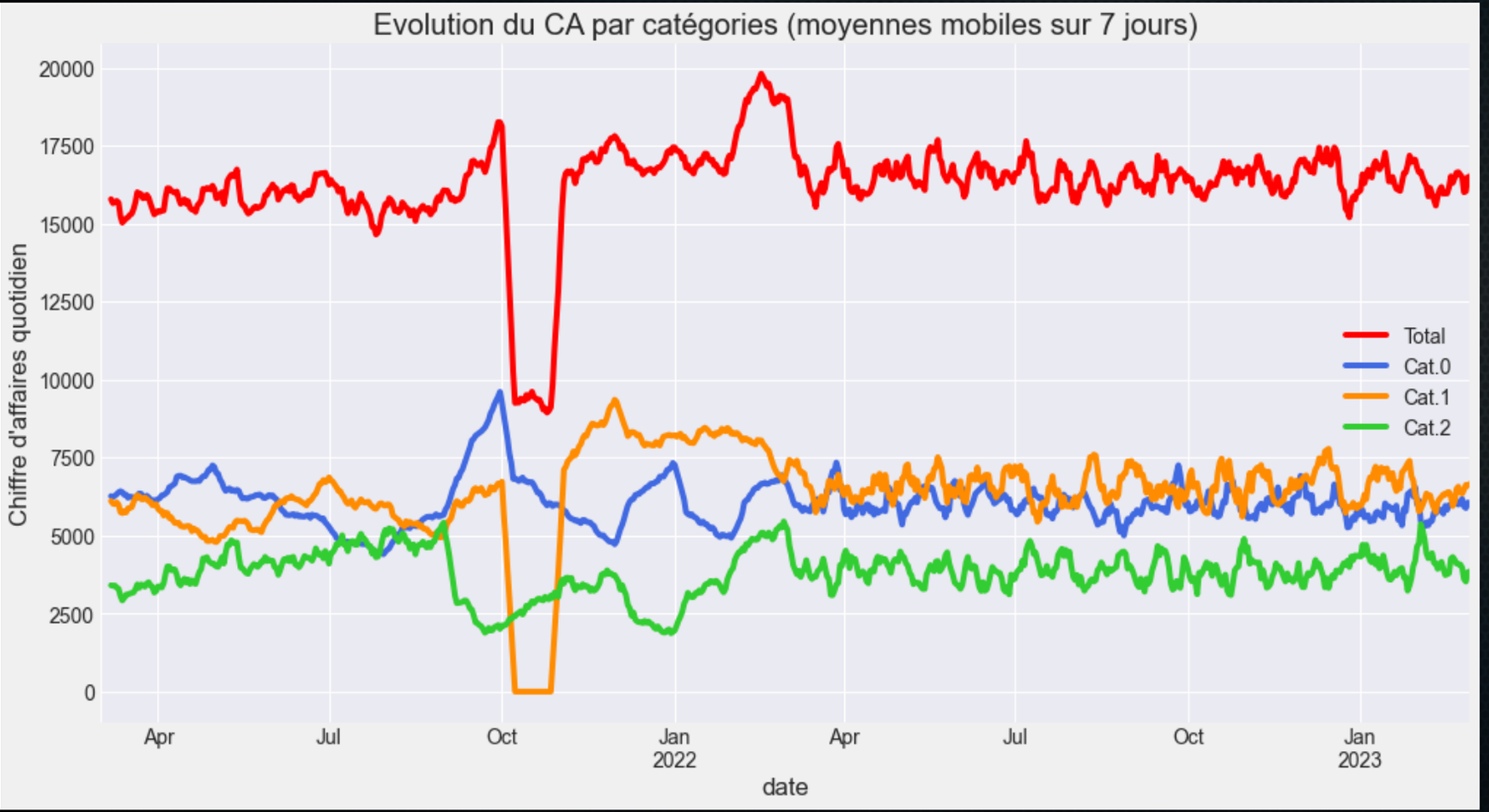


La répartition du CA selon les produits montre une forte disparité:

- 50% du CA est apporté par 8% des produits
- L‘indice de Gini est de 0,74



Evolution dans le temps

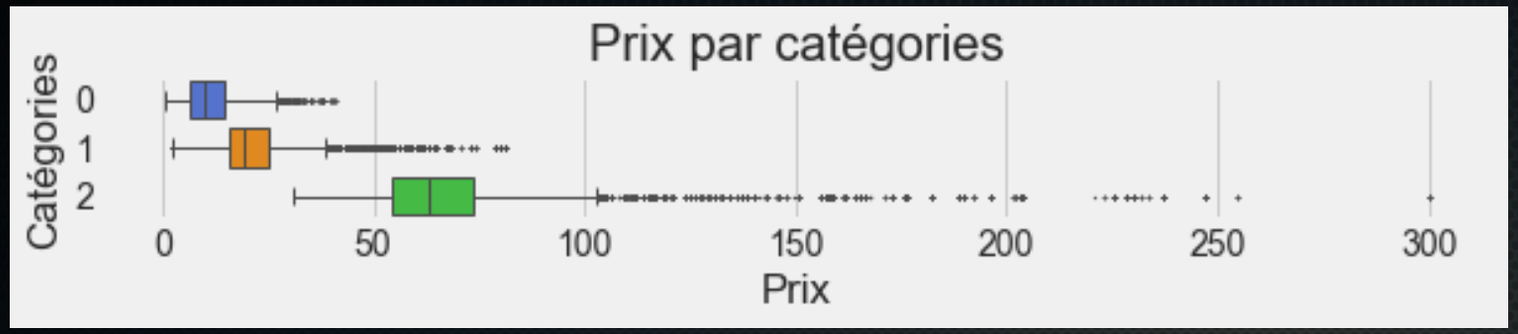


Sur la période étudiée, la tendance du chiffre d‘affaires est globalement stable

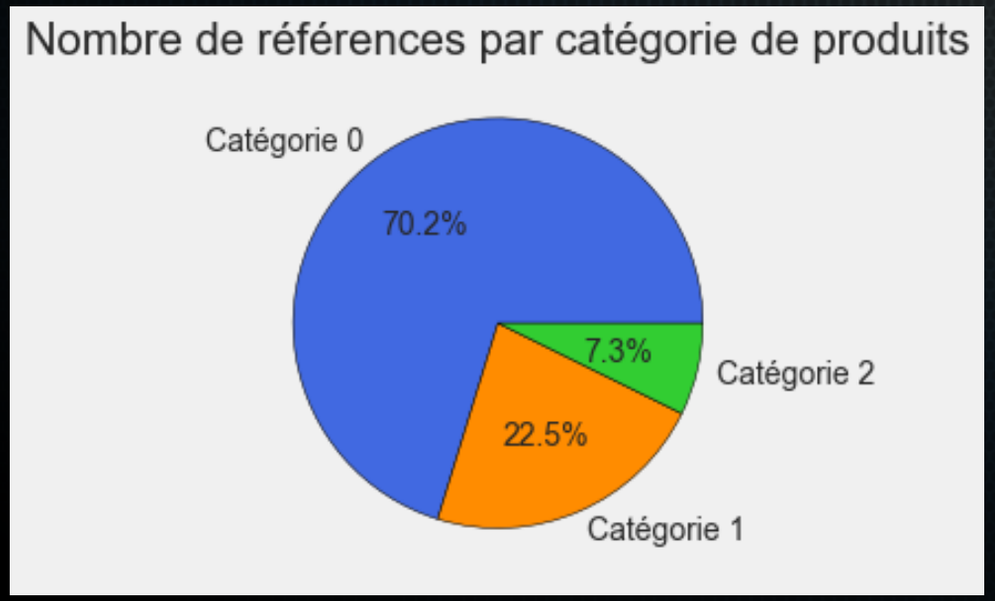
L‘analyse a permis de mettre en évidence un arrêt des ventes pour les produits de catégorie 1 entre le 2 et le 28 octobre 2021 (défaillance de la récupération des données, rupture de stocks, …?)

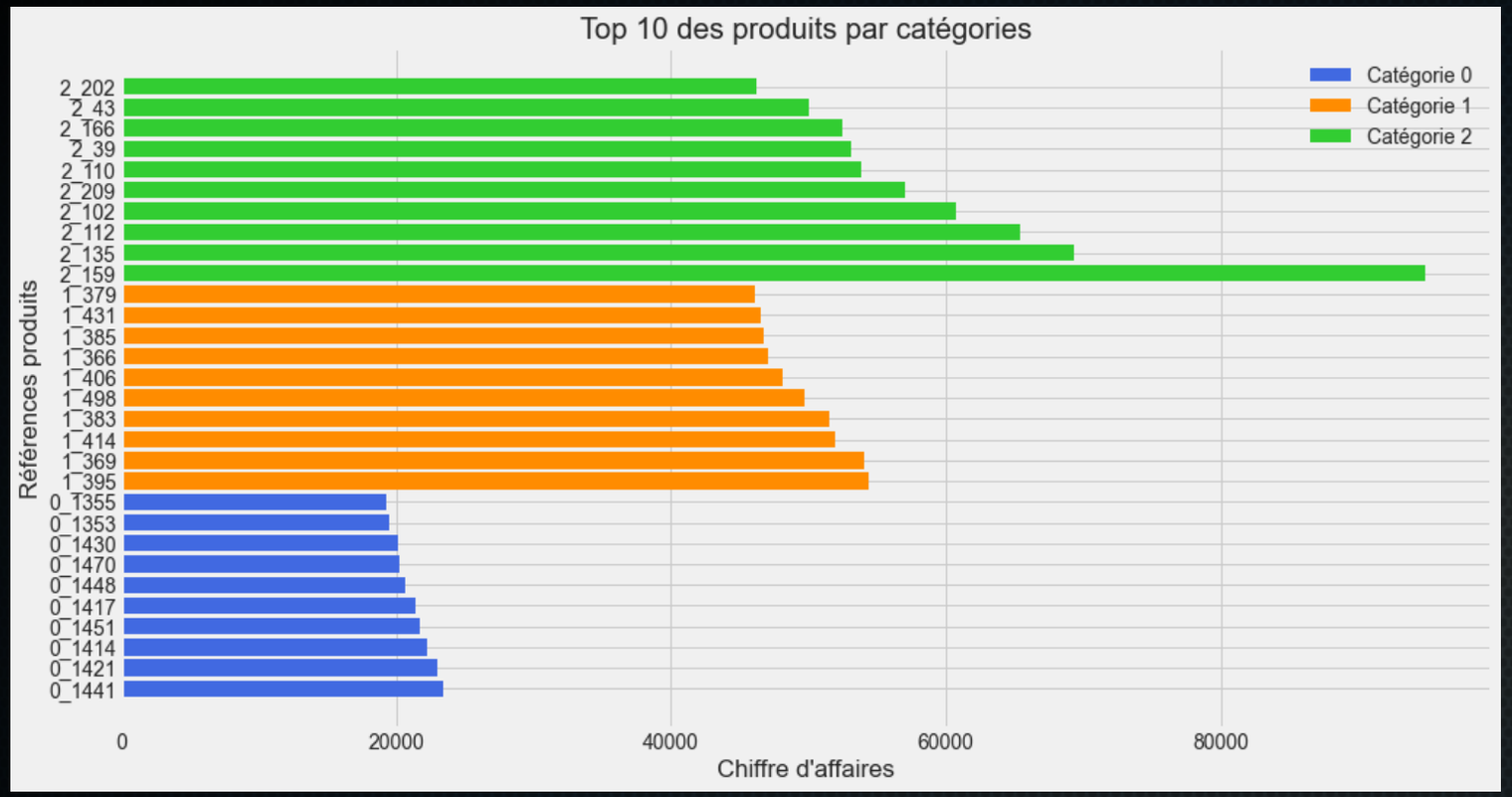
Zoom sur les références

Des gammes de prix très différentes selon les catégories:



Les produits de catégorie 0 constituent l'essentiel de nos références:



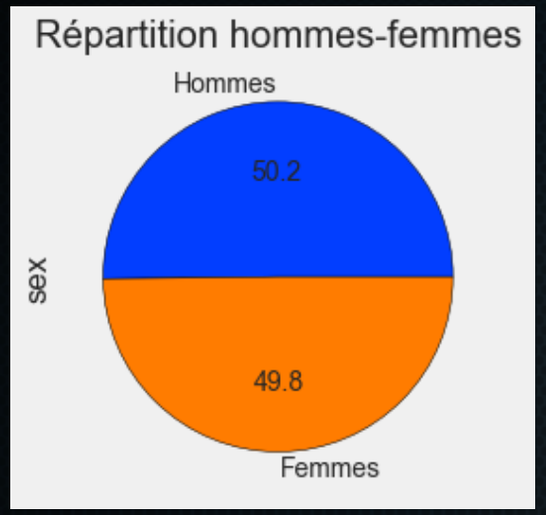


Références ne recensant aucune vente sur la période:

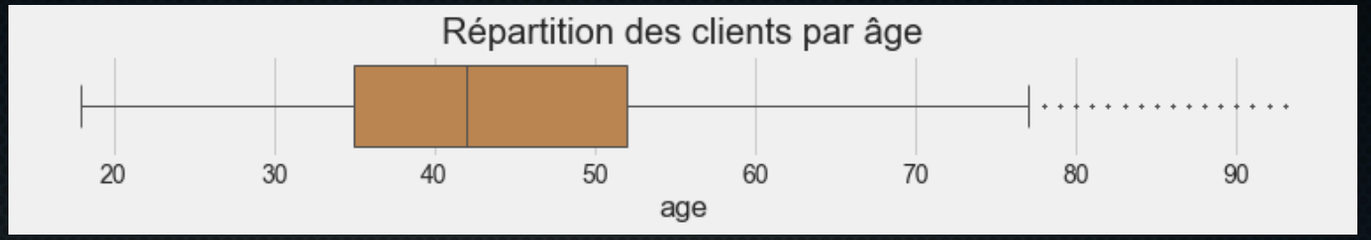
id_prod	price	categ
0_1016	35.06	0
0_299	22.99	0
0_1624	24.50	0
0_310	1.94	0
0_1025	24.99	0
0_510	23.66	0
0_322	2.99	0
0_1645	2.99	0
0_1620	0.80	0
0_1318	20.92	0
0_1014	1.15	0
0_1119	2.99	0
0_1062	20.08	0
0_1780	1.67	0
0_1800	22.05	0
0_2308	20.28	0
1_0	31.82	1
1_394	39.73	1
2_87	220.99	2
2_72	141.32	2
2_86	132.36	2

Profil des clients

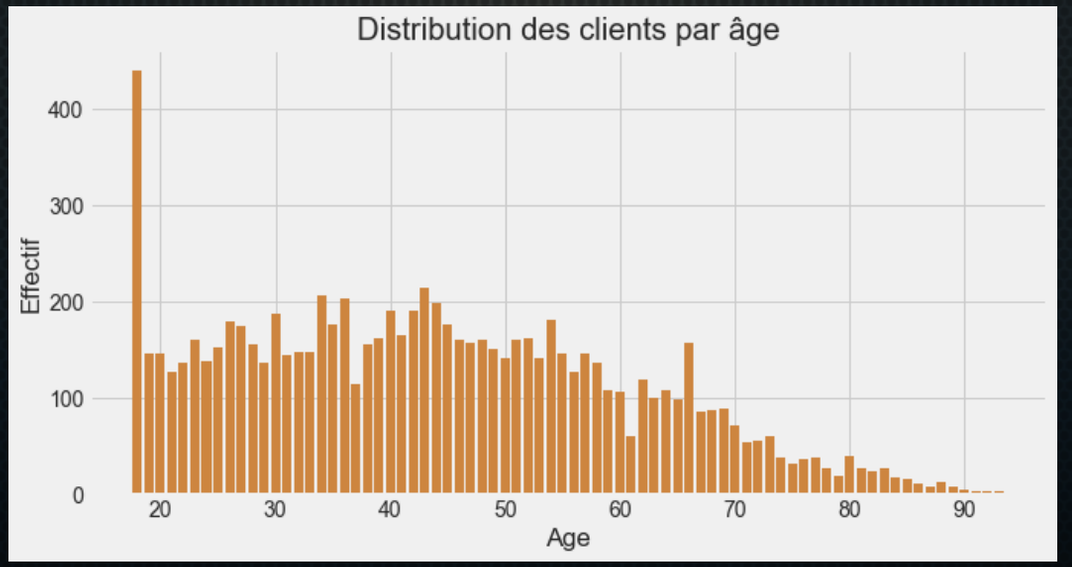
Une répartition par sexe de nos clients plutôt égalitaire:



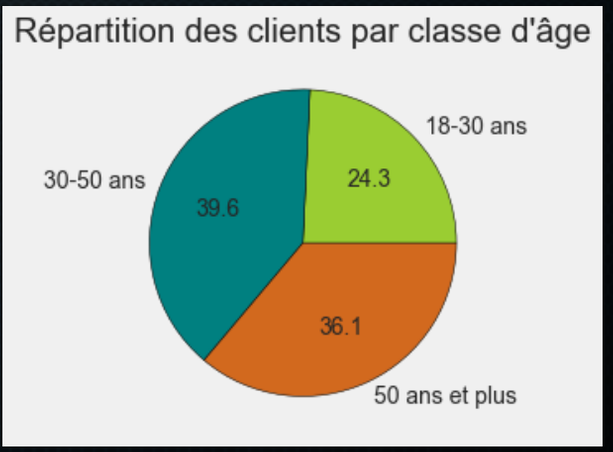
L‘âge moyen est de 44 ans, l‘âge médian de 42 ans



L‘effectif des clients ayant 18 ans est anormalement élevé:



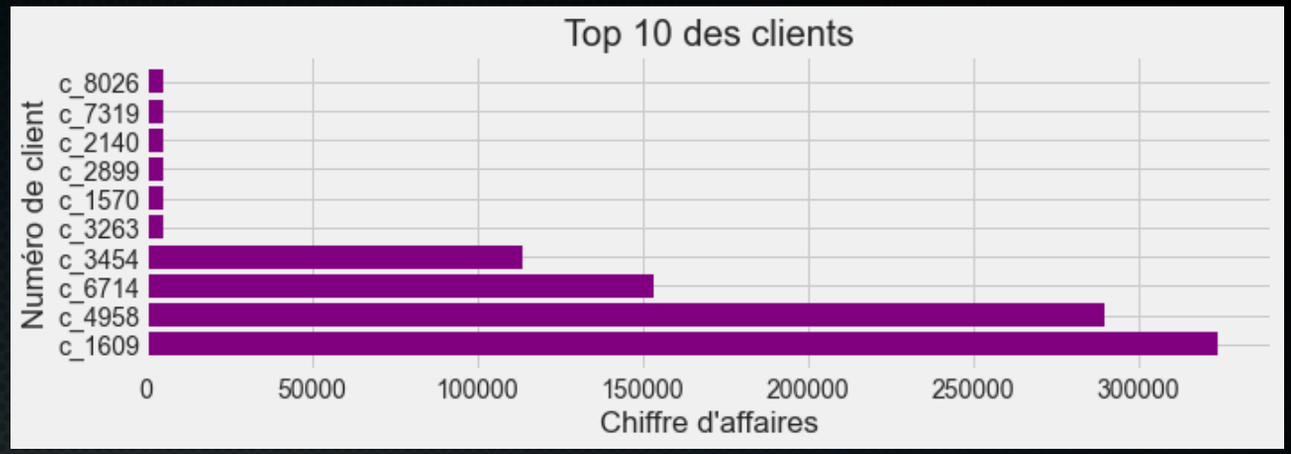
Après définition de classes d‘âge adaptées, la répartition s‘établit ainsi:



4 clients ont des volumes d‘achats très supérieurs aux autres: plus de 100 000 à 300 000 € chacun en 2 ans.

Ils représentent 7,5% du CA.

L‘analyse de leurs achats montre qu‘il s‘agit de professionnels (produits achetés en plusieurs dizaines d‘exemplaires).

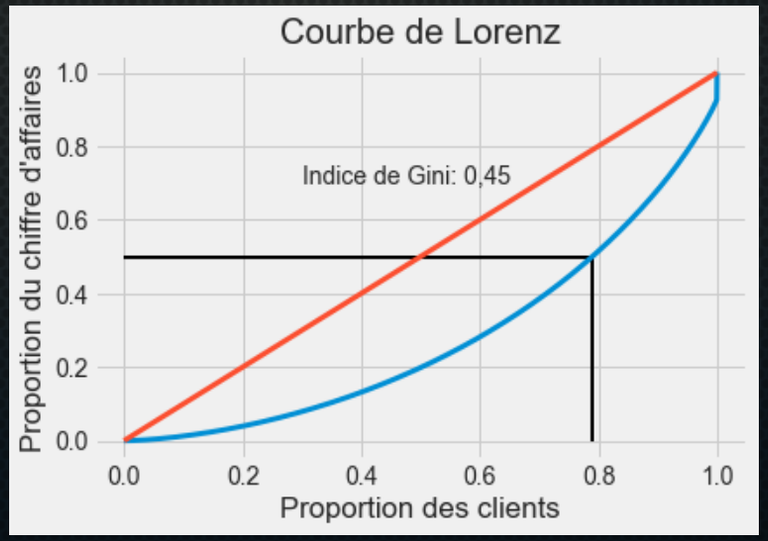


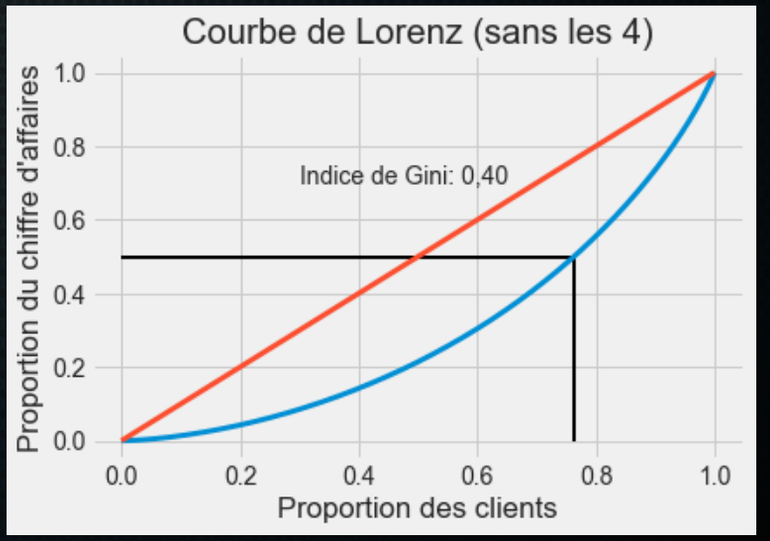
La répartition du CA entre les clients affiche des disparités plus ou moins grandes selon que l‘on garde ou que l‘on exclut les 4 professionnels:

- indice de Gini de 0,45, puis de 0,40

- la médiale se déplace sensiblement vers la gauche

- forme de l‘extrémité haute modifiée





Arbitrages sur les données avant l‘analyse des corrélations

4 clients sont manifestement des professionnels et seront sortis des données pour les analyses suivantes.

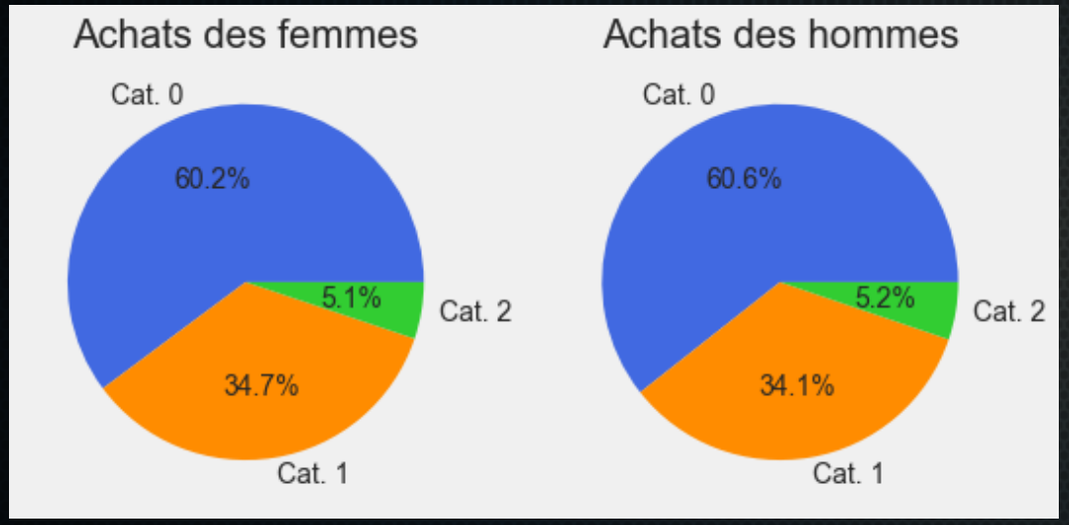
Les données liées à ces clients fausseraient les analyses des préférences d‘achats liées au genre et à l‘âge.

Concernant le problème de l‘absence de ventes des produits de catégorie 1 entre le 2 et le 27 octobre 2021, il a été décidé de sortir toutes les données des ventes du mois d‘octobre 2021 pour toutes les catégories.

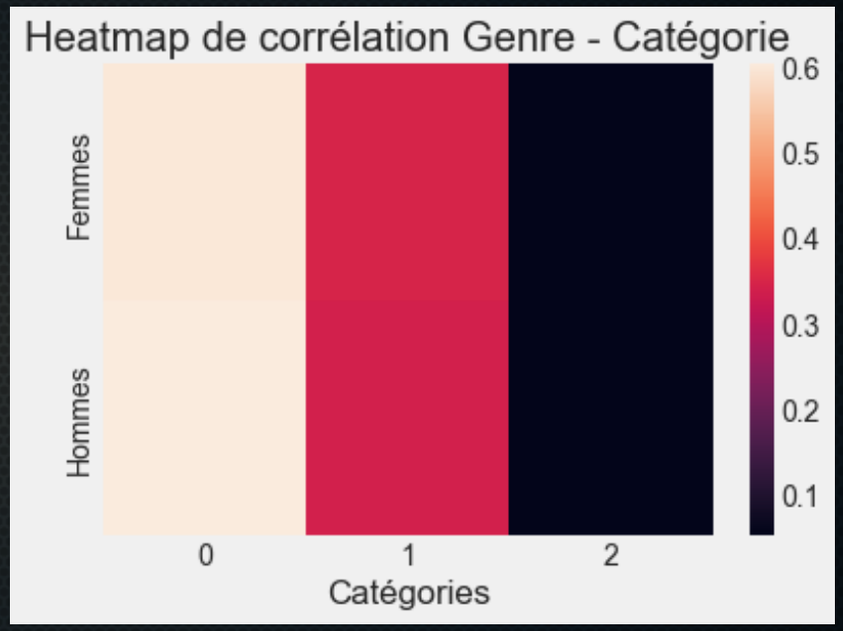
Pour les analyses impliquant l‘âge des clients, un découpage en 3 classes (18-30 ans, 30-50 ans, 50 ans et plus) sera retenu quand cela s‘avérera nécessaire.

Lien entre le genre du client et la catégorie des livres achetés

La représentation des achats par sexe montre une très faible différence en ce qui concerne la catégorie de livres achetée:



Cette faible différence est confirmée par la heatmap Genre - Catégorie :

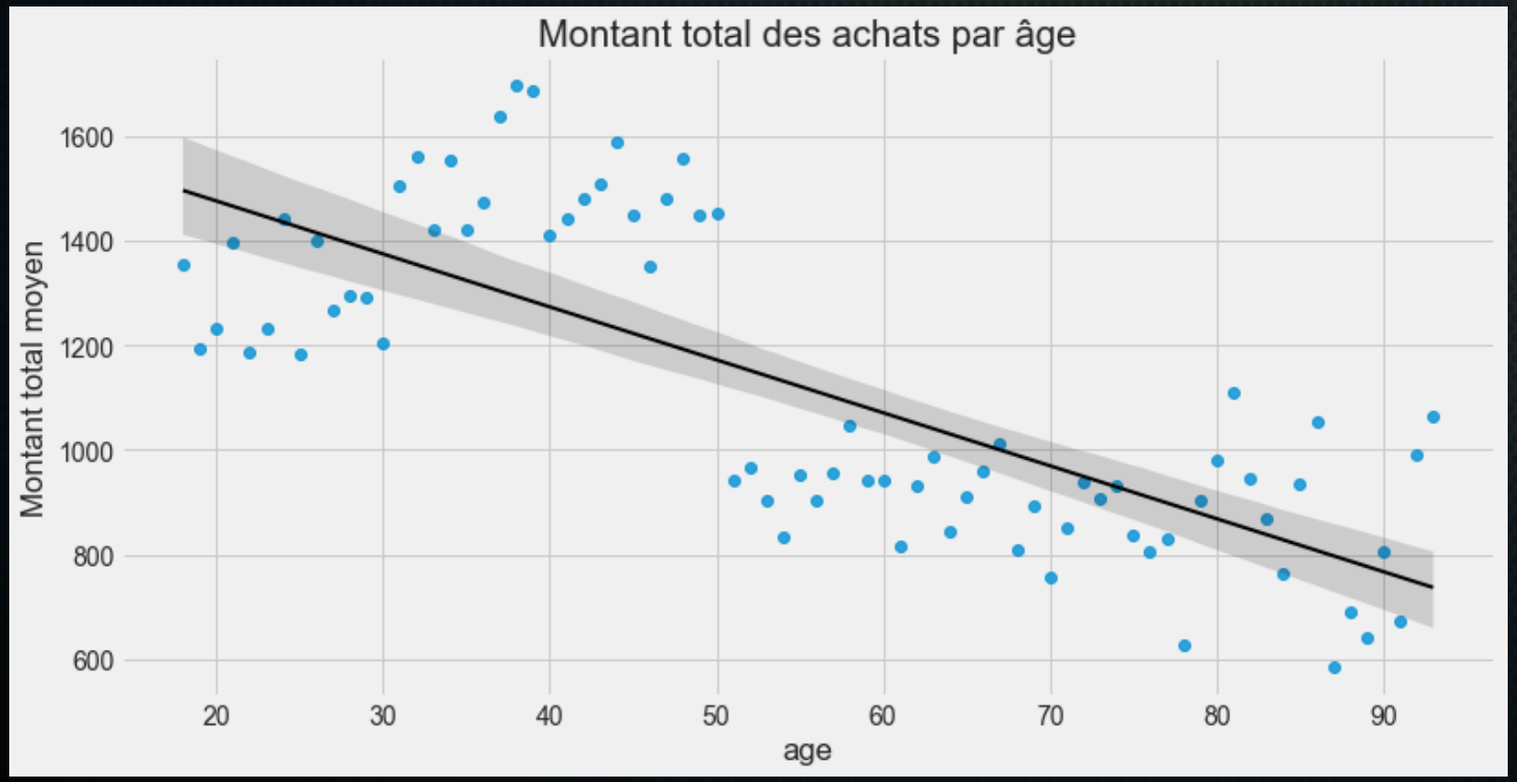


Le calcul du chi² renvoie une p-value>0,05.

L'hypothèse nulle d'indépendance des 2 variables ne peut être rejetée.

Lien entre l‘âge du client et le montant total des achats

Le nuage de points et la droite de régression semblent indiquer une corrélation entre l‘âge du client et le montant total de ses achats



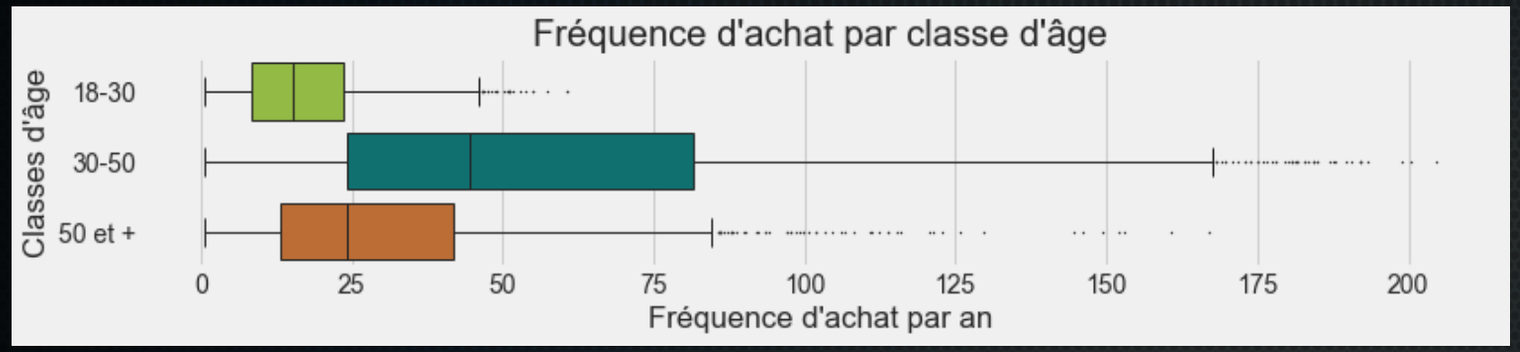
Le coefficient de corrélation de Spearman (test non paramétrique) pour ces deux variables est de -0,72, la p-value <0,05.

L‘hypothèse nulle ("Les 2 variables sont indépendantes") peut donc être rejetée.

Il y a une corrélation négative entre les variables "Âge" et "Montant total des achats".

Lien entre l‘âge du client et la fréquence d‘achat

Fortes disparités de fréquences d‘achat selon la classe d‘âge:

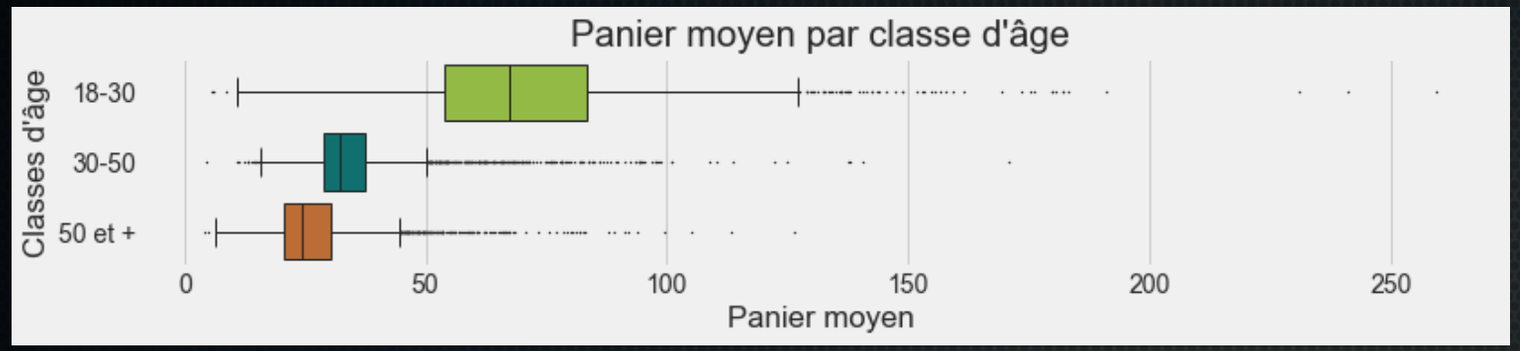


La distribution de la variable "Fréquence d‘achat" n‘étant pas de forme normale (cf. test de Kormogorov-Smirnov) et les 3 échantillons n‘étant pas de même variance (cf. test de Levene), le test de corrélation retenu ici est non-paramétrique: Kruskal-Wallis.

Le test nous indique que les valeurs médianes des 3 groupes sont différentes.

Lien entre l‘âge du client et le panier moyen

Fortes disparités du panier moyen selon la classe d‘âge:

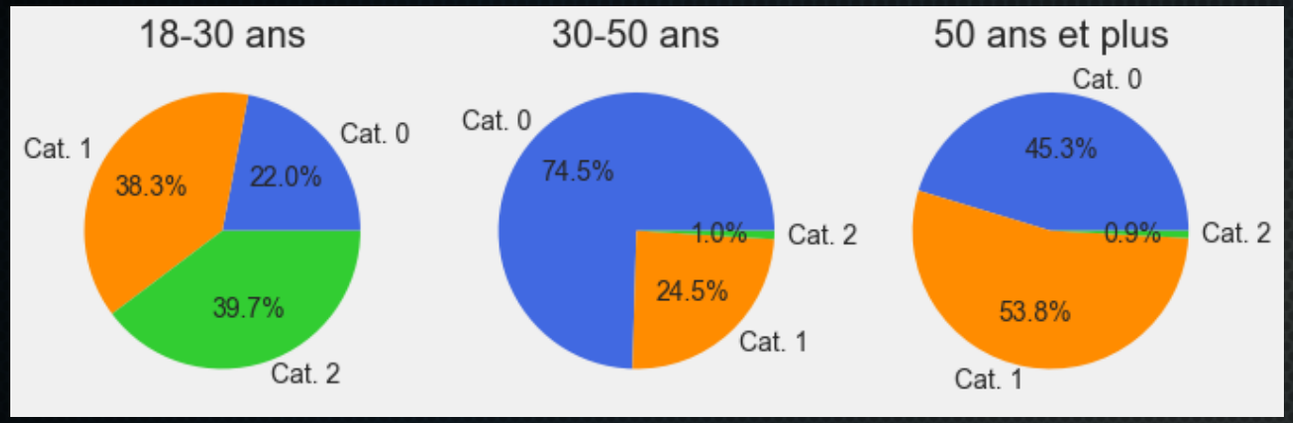


La distribution de la variable "Panier moyen" n‘étant pas de forme normale (cf. test de Kormogorov-Smirnov) et les 3 échantillons n‘étant pas de même variance (cf. test de Levene), le test de corrélation retenu ici est non-paramétrique: Kruskal-Wallis.

Le test nous indique que les valeurs médianes des 3 groupes sont différentes.

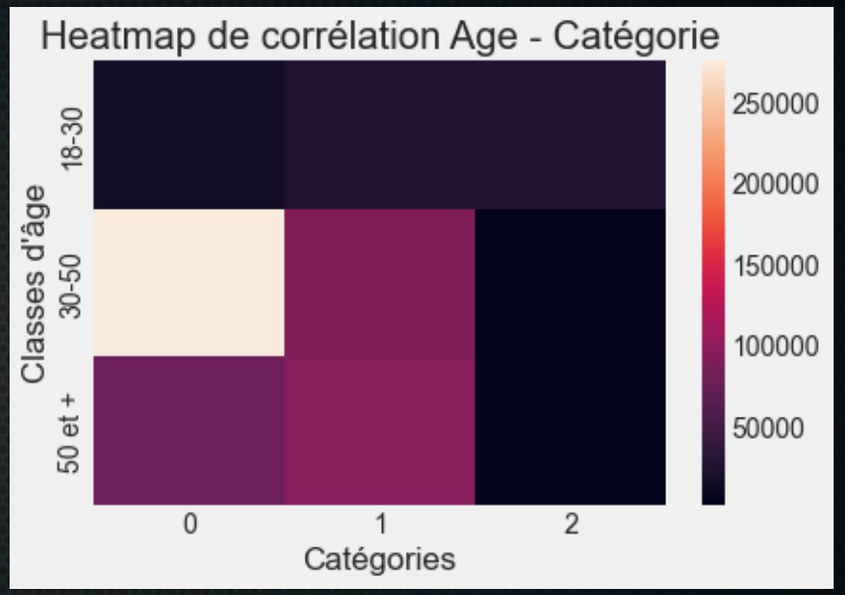
Lien entre l‘âge du client et la catégorie achetée

La répartition des achats par catégorie selon l‘âge du client fait apparaître de grandes disparités:



classe d'âge	categ 0	1	2
18-30	14753	25721	26674
30-50	276121	90648	3642
50 et +	79730	94736	1501

Ces différences de comportement sont corroborées par la heatmap Âge-Catégorie:



Le calcul du chi² (p-value<0,05) permet de rejeter l‘hypothèse nulle (“Les variables sont indépendantes“). Il existe un lien entre l‘âge et la catégorie de livres achetée.