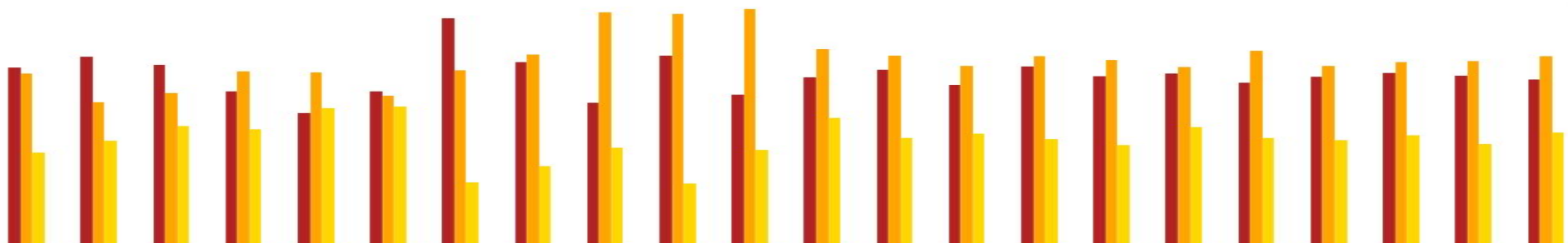




Lapage



Rapport statistique des ventes et autres chiffres clés.

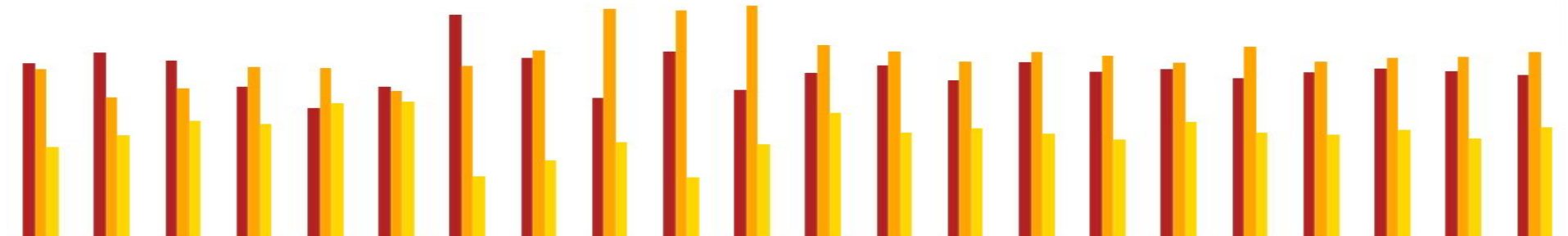


Table des matières :

⇒ En résumé

⇒ 1 - Importation des librairies et modules

⇒ 2 - Importation des documents

⇒ 3 - Nettoyage des documents

⇒ 3.1 - Fichier "customers"

⇒ 3.2 - Fichier "products"

⇒ 3.3 - Fichier "transactions"

⇒ 4 - Analyses préliminaires des documents

⇒ 4.1 - Fichier "customers"

⇒ 4.2 - Fichier "products"

⇒ 4.3 - Fichier "transactions"

⇒ 5 - Jointure des fichiers

⇒ 6 - Etude du chiffre d'affaires et des indicateurs de ventes

⇒ 6.1 - Chiffre d'affaires

- ⇒ Total
- ⇒ Par an
- ⇒ Par mois
- ⇒ Par jour

⇒ 6.2 - TOP

⇒ 6.3 - FLOP

⇒ 6.4 - Catégories

⇒ 6.5 - Autres chiffres clés

- ⇒ Courbe de Lorenz sur la répartition du CA entre les produits
- ⇒ Recherche des meilleurs clients
- ⇒ Achat et montant moyen par client
- ⇒ Courbe de Lorenz sur la répartition du CA entre les clients
- ⇒ Nombre moyen d'achats par sessions
- ⇒ Répartition du genre parmi les clients
- ⇒ Dépenses par genre

⇒ 7 - Corrélations :

⇒ 7.1 - Entre le genre et la catégorie d'achat

⇒ 7.2 - Entre l'âge et le montant

⇒ 7.3 - Entre l'âge et les catégories

⇒ 7.4 - Entre l'âge et la fréquence d'achat

⇒ 7.5 - Entre l'âge et le panier moyen

⇒ En conclusion

En résumé :

Des données allant du :

01/02/2021 au **28/02/2023**

8621 clients enregistrés et **3286** références

Un chiffre d'affaires cumulé sur cette période de
12 028 458.38€

pour un total de **687 534** ventes.

Année	CA	Nbr de ventes
2021	4 944 760.98	286 671
2022	6 108 681.81	346 380
2023	974 220.31	54 483

+23,5%

de CA entre 2021 et 2022.
(2023 non complet)

+20,8%

de ventes sur la même période

1 - Importation des librairies et modules

Panda as :
pd

Matplotlib
as:
plt

Seaborn as :
sns

Numpy as :
np

Datetime as:
dt

collections

Scipy.stat
as :
st

sklearn :
chi2

statsmodels
as:
sm

statsmodels.
formula :
ols

2 - Importation des documents

Customers.csv

Products.csv

Transactions.csv

Trois fichiers au format “CSV” avec “;” en règle de séparation.

3 - Nettoyage des documents

Les documents donnés étaient propres, mais des règles de nettoyage basique s'appliquent toujours :

- Vérifier la présence de données manquantes
- Vérifier la présence de doublons
- Vérifier le typage des données
- Vérifier le contenu de la table
- Modifier les éventuelles erreurs
- Ajouter les informations pertinentes

3.1 - Fichier “customers”

- Aucun doublon
- Ajout d’une colonne âge, calculée à partir des dates de naissance pour faciliter l’exercice

3.2 - Fichier “products”

- Passage de la colonne catégorie en “object”.
- Attention à la norme (les identifiants ne respectent pas le 1NF)

3.3 - Fichier “transactions”

- Passage de date en “datetime”
- 848 doublons qui ne peuvent être éliminés (double achat ? cadeaux?)

Les 3 fichiers sont maintenant propres et utilisables pour l’exercice.

4 - Analyses préliminaires des documents

4.1 - Fichier
"customers"

4.2 - Fichier
"products"

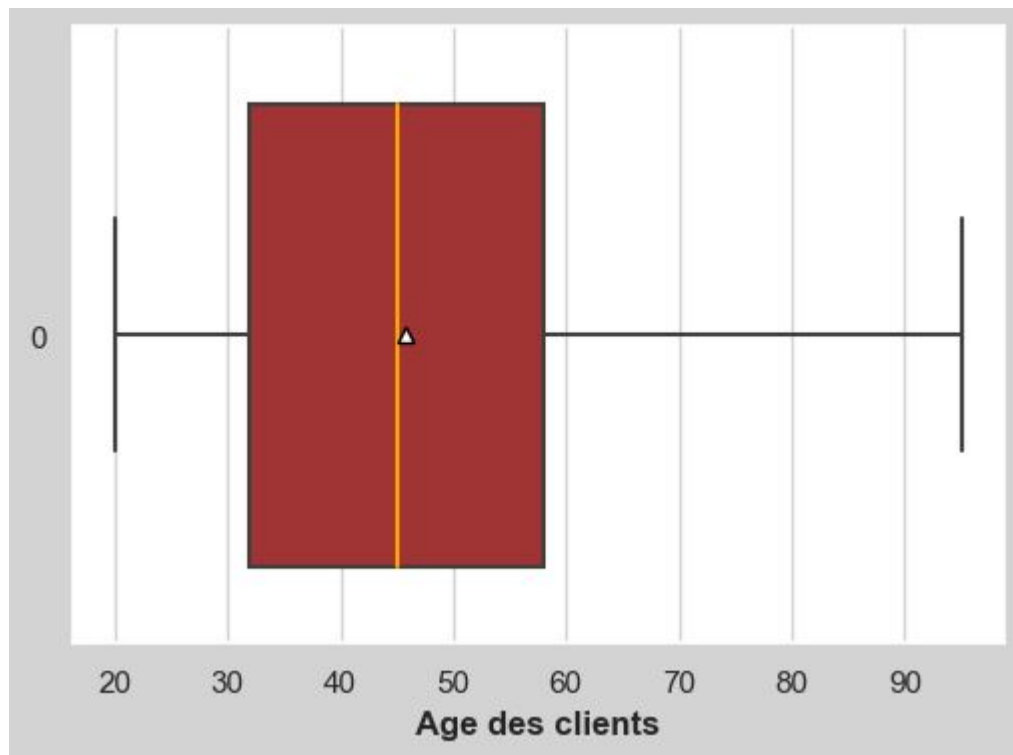
4.3 - Fichier
"transactions"

4.1 - Fichier “customers”

8621 clients présents dans la table,
dont les âges se répartissent entre 20
et 95 ans.

La moyenne d'âge se situe à 45 ans,
proche de l'âge médian.

Il y a 50% de notre clientèle entre 32
et 58 ans

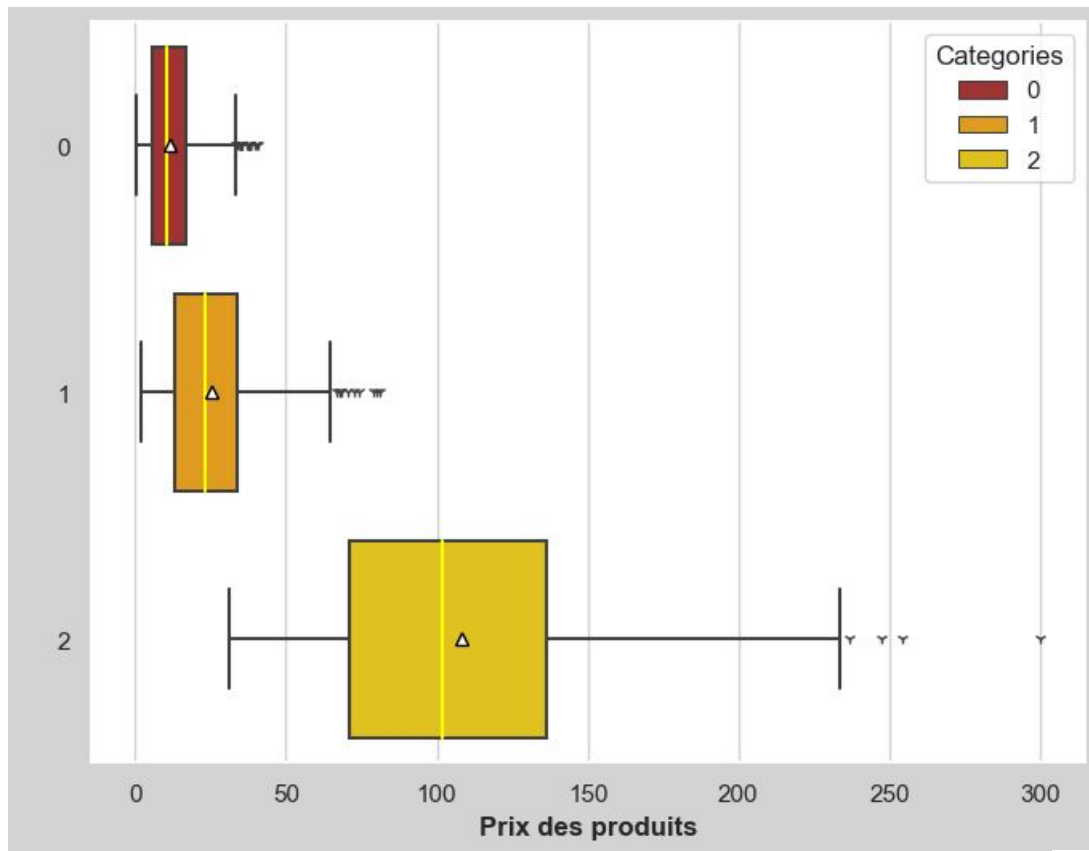


4.2 - Fichier “products”

3286 références présentes

La recherche des valeurs extrêmes, méthode “métier” ou interquartile, ne met pas en avant de valeurs aberrantes.

L’analyse des prix par catégorie montre des valeurs atypiques, mais non aberrantes.



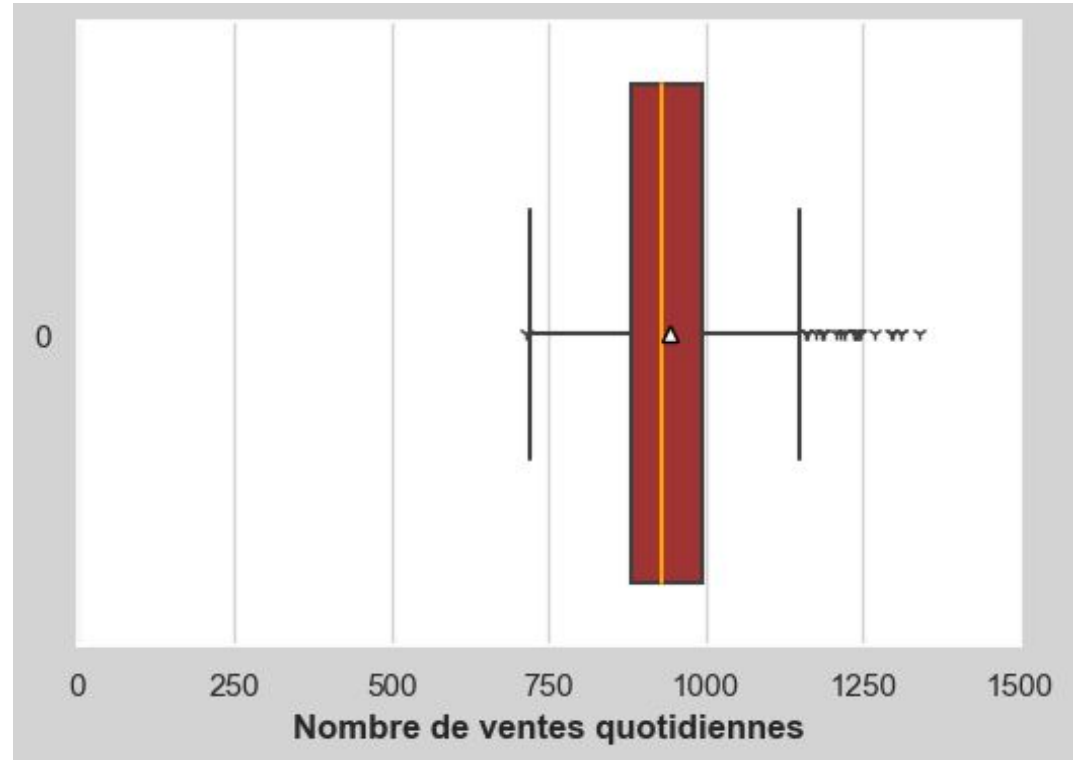
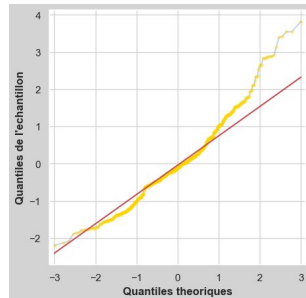
4.3 - Fichier “transactions”

8600 identifiants client sur 8621 :
21 clients n'ont jamais acheté sur le site.

3265 références produit sur 3286:
21 livres n'ont jamais été vendus.

Le nombre de ventes quotidienne
moyenne se situe aux alentours de
950 avec, visiblement, certains jours
meilleurs que d'autres.

Le test de Shapiro-Wilks
nous montre que la
distribution n'est pas
gaussienne :



5 - Jointure des fichiers



Jointure faite sur “outer” pour conserver les données comme les livres jamais vendus ou clients non-acheteurs.

Création d’un nouveau dataframe pour la suite de l’exercice.

6 - Etude du chiffre d'affaires et des indicateurs de vente

6.1 - Chiffre
d'affaire

6.2 - TOP

6.3 - FLOP

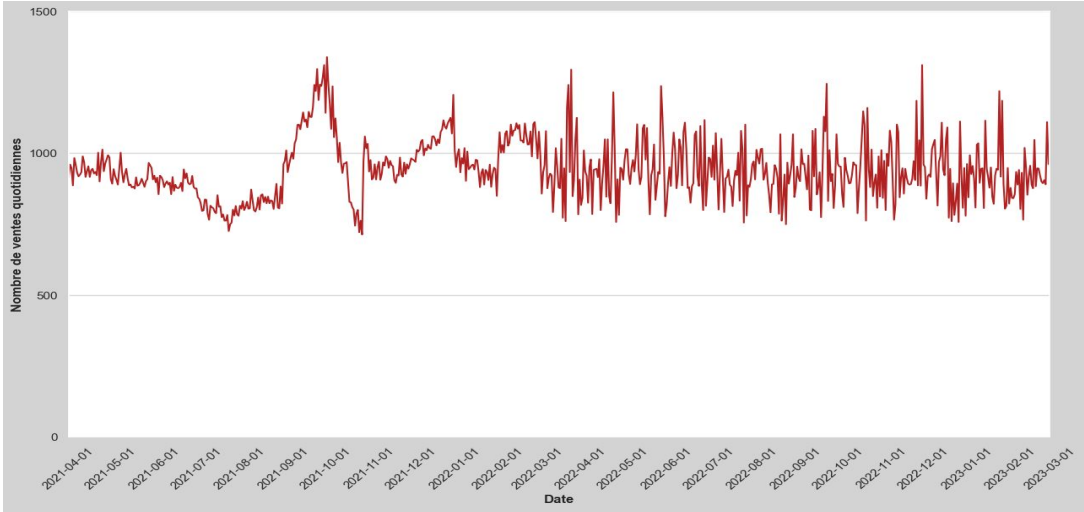
6.4 - Catégories

6.5 - Autres
chiffres clés

6.1 - Chiffre d'affaires :

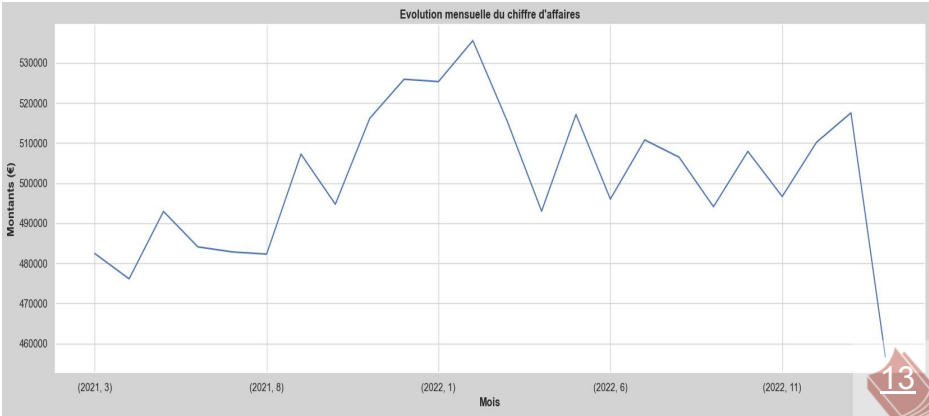
Total =
12 028 458.38€

de chiffre d'affaires cumulé sur la période
avec un total de **687 534** ventes.

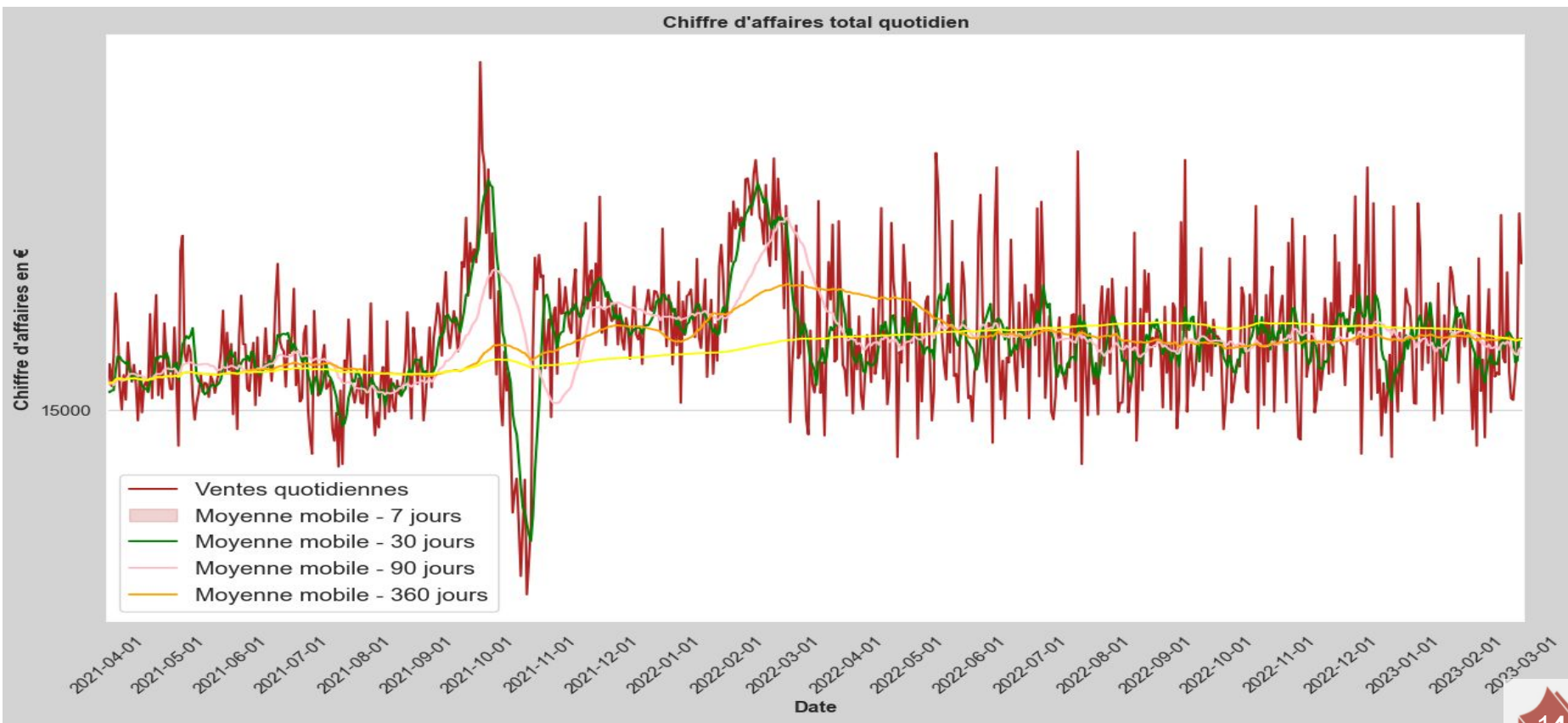


Par an :

Année	CA	Nbr de ventes
2021	4944760.98	286671
2022	6108681.81	346380
2023	974220.31	54483



Moyenne mobile journalière :



6.2 - TOP

Liste des 10 meilleures ventes :

Identifiant	Catégorie	Nombre de ventes :
1_369	1	2340
1_417	1	2269
1_414	1	2246
1_498	1	2202
1_425	1	2163
1_403	1	2040
1_413	1	2036
1_412	1	2014
1_406	1	2003
1_407	1	2001

6.3 - FLOP

Les 21 livres non vendus sont considérés comme les FLOP des ventes. Liste des identifiants :

	id_prod	price	categ
184	0_1016	35.06	0
279	0_1780	1.67	0
736	0_1062	20.08	0
793	0_1119	2.99	0
810	0_1014	1.15	0
845	1_0	31.82	1
1030	0_1318	20.92	0
1138	0_1800	22.05	0
1346	0_1645	2.99	0
1504	0_322	2.99	0
1529	0_1620	0.80	0

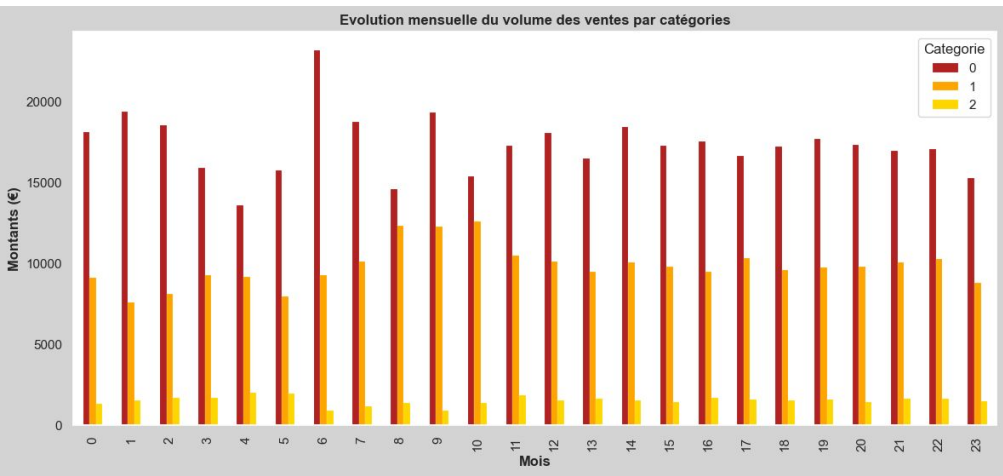
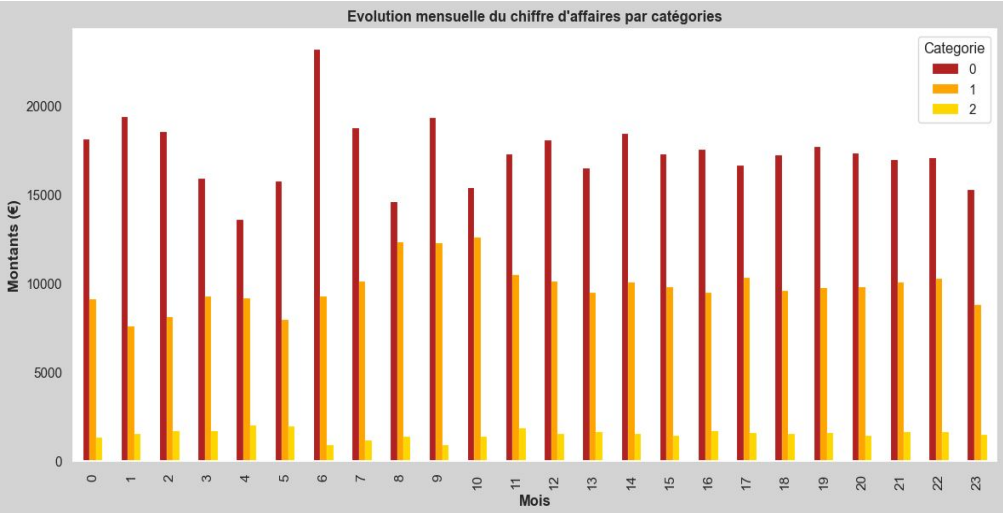
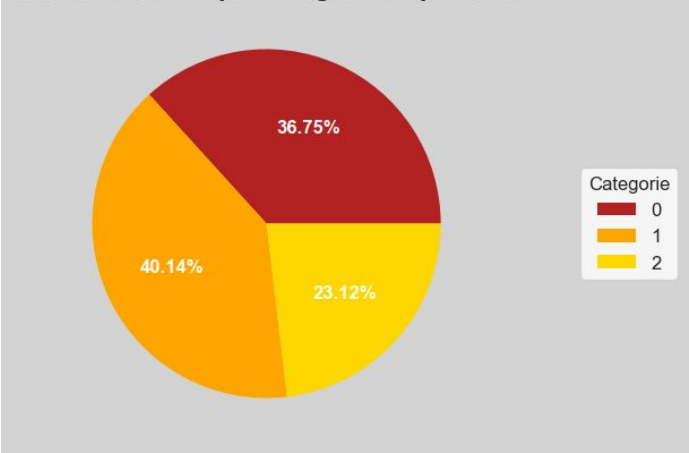
1542	0_1025	24.99	0
1708	2_87	220.99	2
1862	1_394	39.73	1
1945	2_72	141.32	2
2214	0_310	1.94	0
2407	0_1624	24.50	0
2524	2_86	132.36	2
2689	0_299	22.99	0
3030	0_510	23.66	0
3095	0_2308	20.28	0

6.4 - Catégories

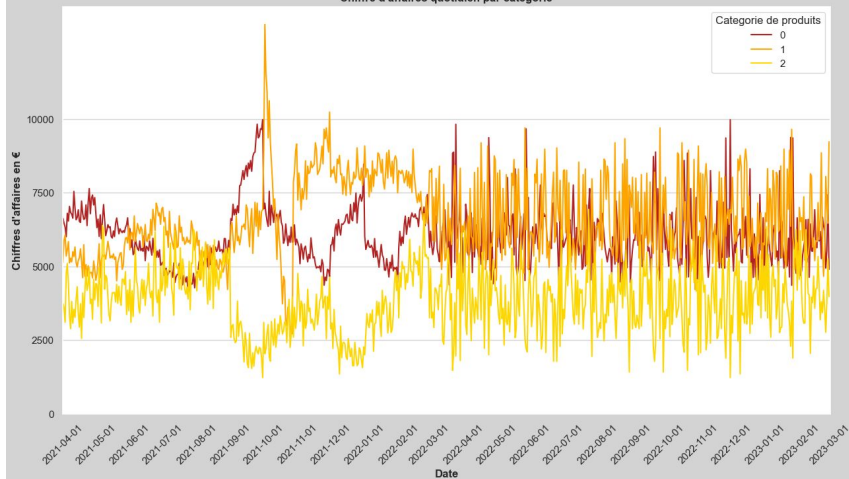
Nombre de livres par catégories :

Catégorie	Nombre
0	415459
1	235592
2	36483

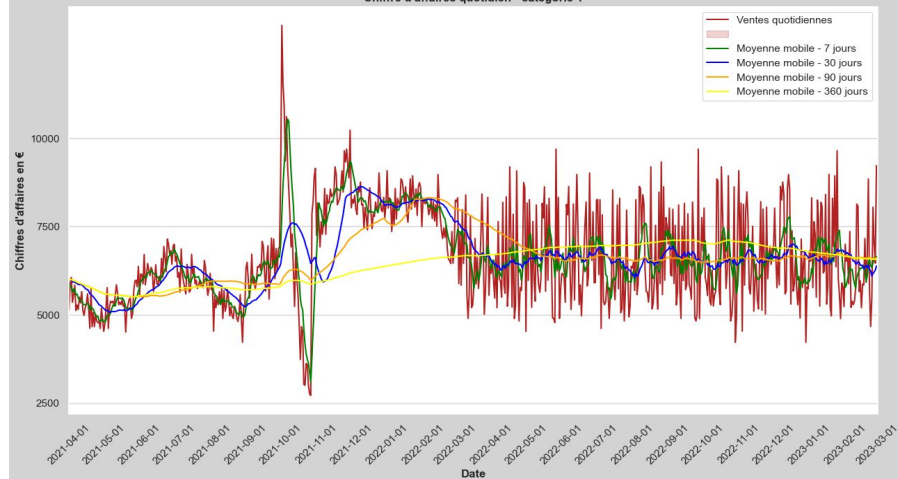
Chiffre d'affaires par categorie de produits



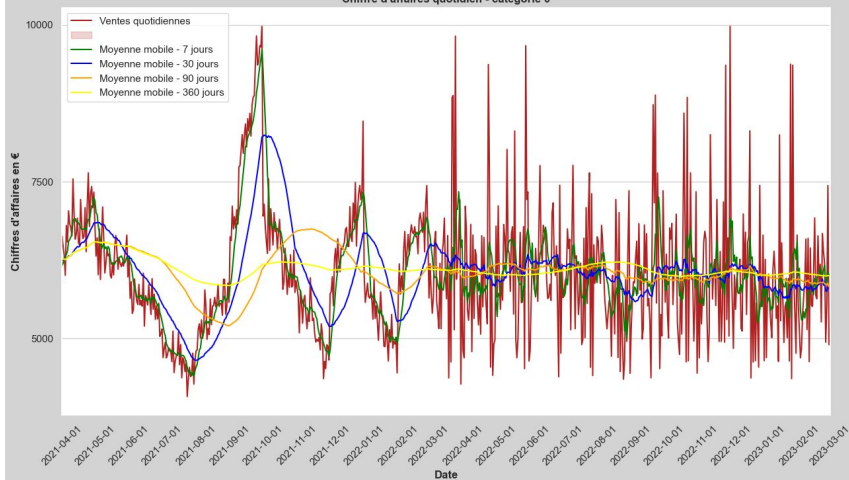
Chiffre d'affaires quotidien par categorie



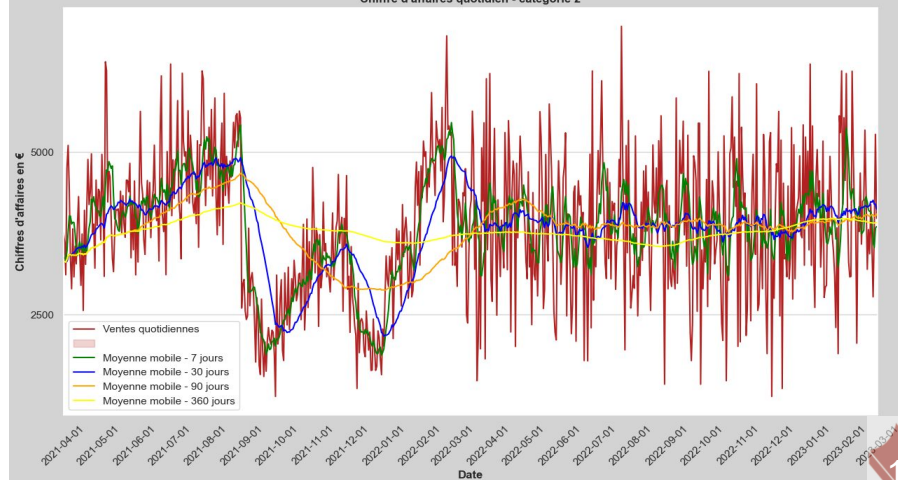
Chiffre d'affaires quotidien - categorie 1



Chiffre d'affaires quotidien - categorie 0



Chiffre d'affaires quotidien - categorie 2



6.5 - Autres chiffres clés

Répartition du CA
entre les produits

Recherche des
meilleurs clients

Achat et montant
moyen par client

Répartition du CA
entre les clients

Nombre moyen
d'achats par
session

Répartition du
genre parmi les
clients

Dépenses par
genre

Répartition du CA entre les produits : courbe de Lorenz

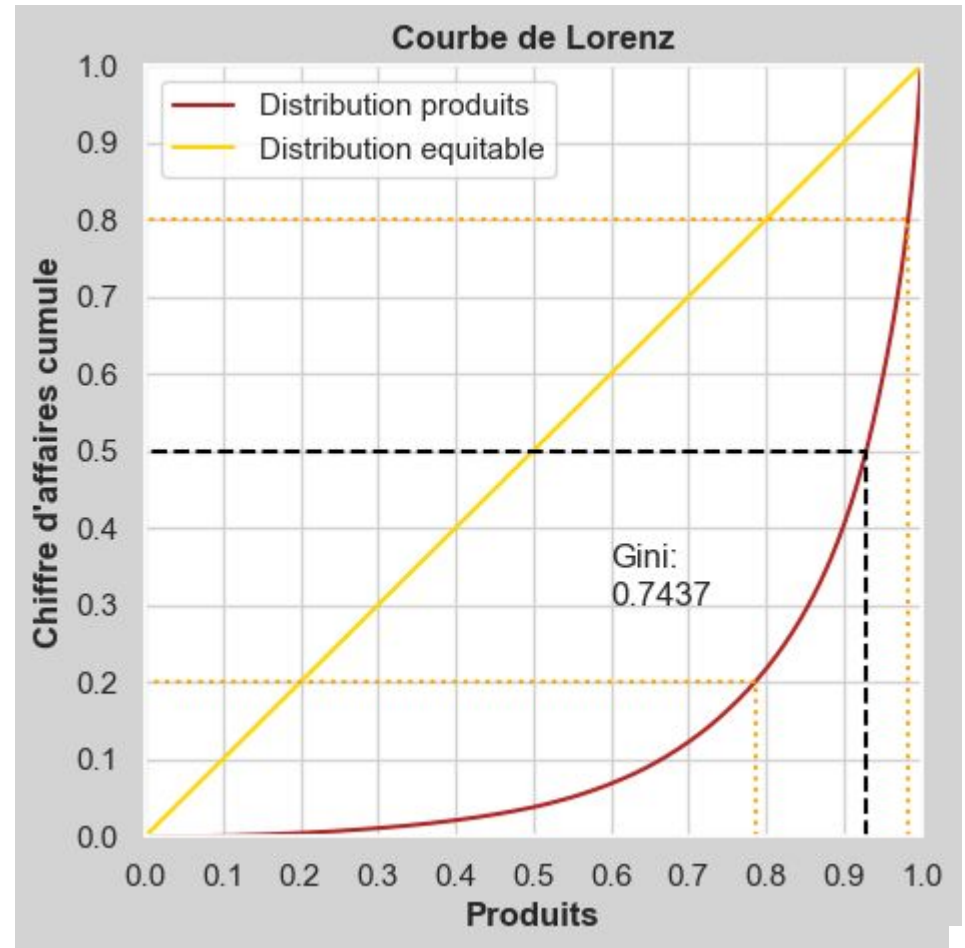
Indice de Gini éloigné de 0, il n'y a donc pas de répartition équitable des prix entre les différents produits.

20% du CA = \approx 80% des produits

50% du CA = \approx 91% des produits

80% du CA = \approx 98% des produits

Il faut seulement 10% des produits pour réaliser 50% du chiffre d'affaires.



Recherche des meilleurs clients :

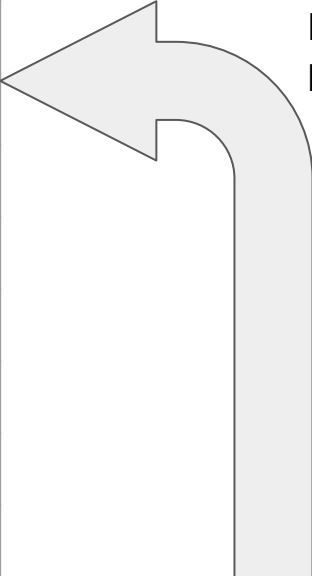
Liste :

Identifiant	Montant
c_1609	326039.89
c_4958	290227.03
c_6714	153918.60
c_3454	114110.57
c_1570	5285.82
c_3263	5276.87
c_2140	5260.18
c_2899	5214.05
c_7319	5155.77
c_7959	5135.75

Achat et montant moyen par client :

Le nombre moyen d'achat par client depuis l'ouverture de notre site est de 79.75

Le montant moyen d'achat par client depuis l'ouverture de notre site est de 1 398.57€



4 clients majeurs, sans doute professionnels.

Répartition du CA entre les clients : courbe de Lorenz

Indice de Gini éloigné de 0, il n'y a donc pas de répartition équitable des prix entre les différents clients.

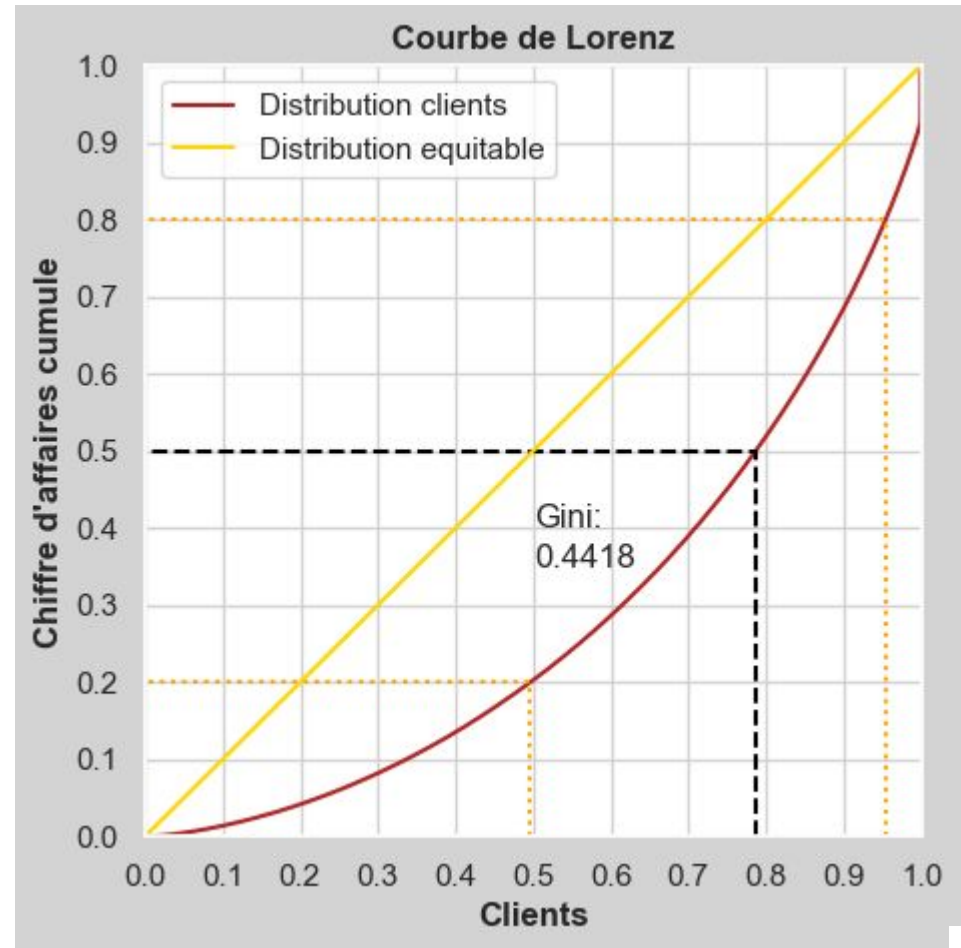
Identification des clients pro en fin de courbe.

20% du CA = \approx 50% des clients

50% du CA = \approx 80% des clients

80% du CA = \approx 95% des clients

Il faut seulement 5% des acheteurs pour réaliser 80% du chiffre d'affaires.

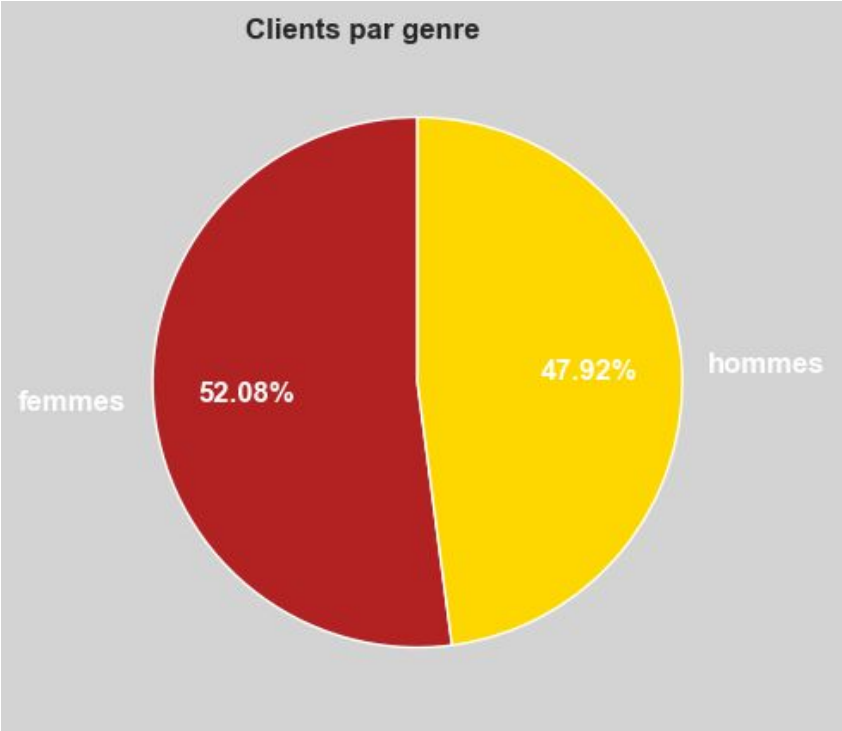


Nombre moyen d'achat par session :

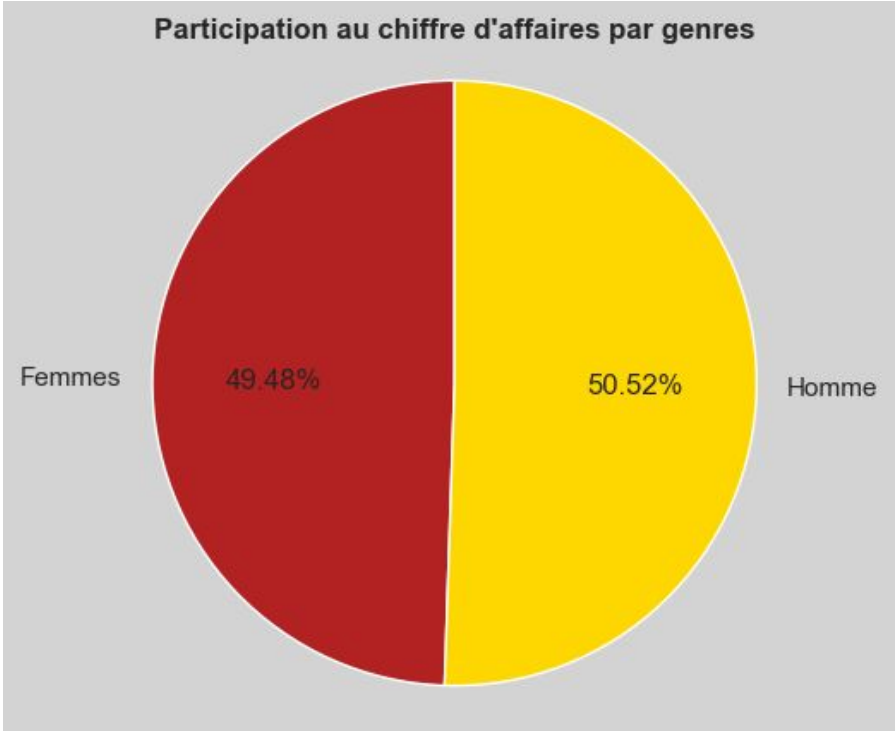
1,99

Sur 345 505 sessions depuis l'ouverture du site internet,
on aboutit, en moyenne à 1,99 achat par session.

Répartition du genre parmi les clients :



Dépenses par genres :



7 - Corrélations

7.1 - Entre le
genre et la
catégorie d'achat

7.2 - Entre l'âge et
le montant

7.3 - Entre l'âge et
les catégories

7.4 - Entre l'âge et
la fréquence
d'achat

7.5 - Entre l'âge et
le panier moyen

7.1 - Entre le genre et la catégorie d'achat :

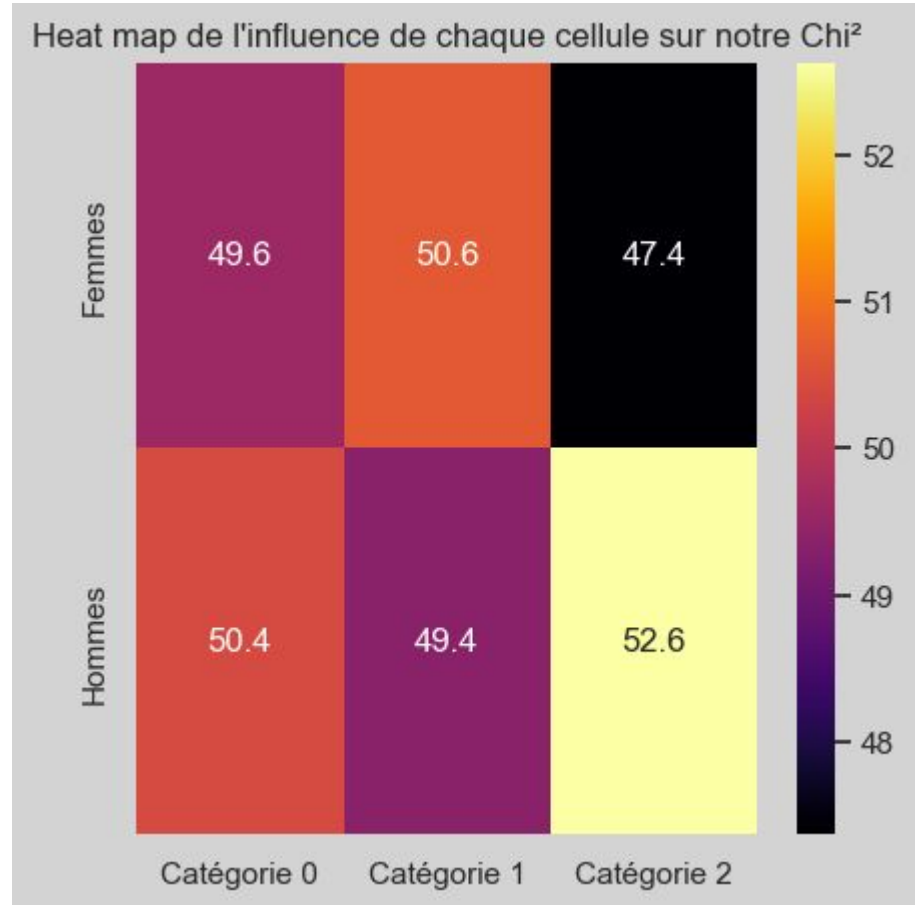
Variables qualitatives :
Test de χ^2

H0 : indépendance des catégories achetées et genre.

H1 : dépendance entre les deux.

P-value < au seuil significatif.

On peut rejeter l'hypothèse 0 : il y a une dépendance (faible) entre le genre et les catégories achetées.



7.2 - Entre l'âge et le montant :

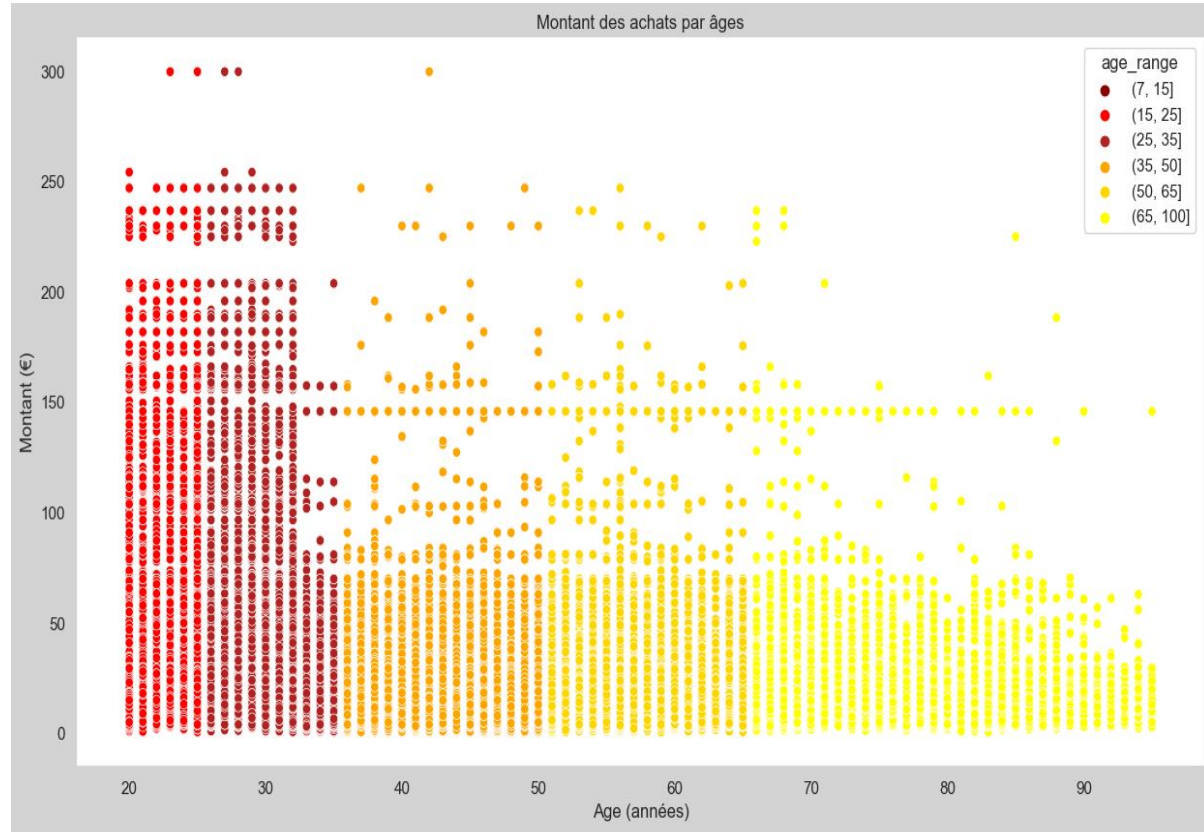
Variables qualitatives et quantitatives:
test ANOVA

H0 : indépendance des catégories âge et
montant.

H1 : dépendance entre les deux.

P-value < au seuil significatif.

On peut rejeter l'hypothèse 0 : il y a une
dépendance entre l'âge et le montant des
achats.



7.3 - Entre l'âge et les catégories :

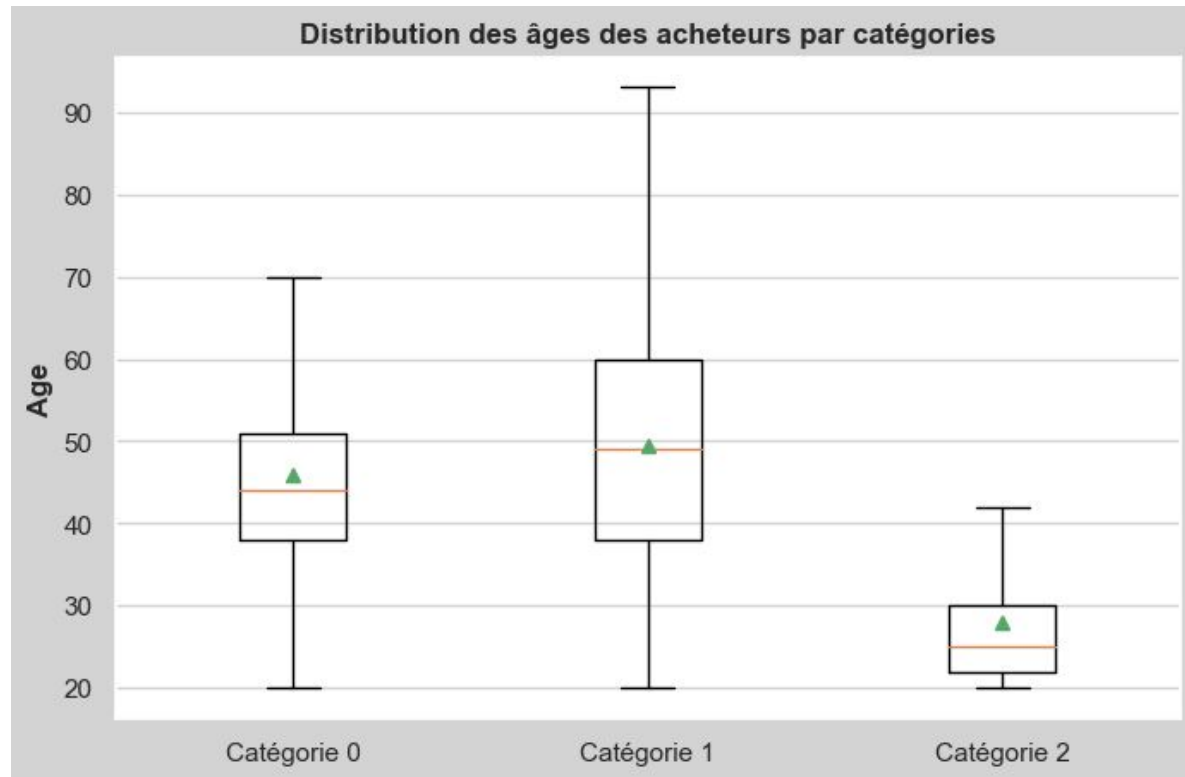
Variables quantitatives et qualitatives :
test Kruskal-Willis

H0 : ils ont tous la même tendance centrale

H1 : différences entre eux.

P-value < au seuil significatif.

On peut rejeter l'hypothèse 0 : il y a une différence entre les âges médians par catégorie.



7.4 - Entre l'âge et la fréquence d'achat :

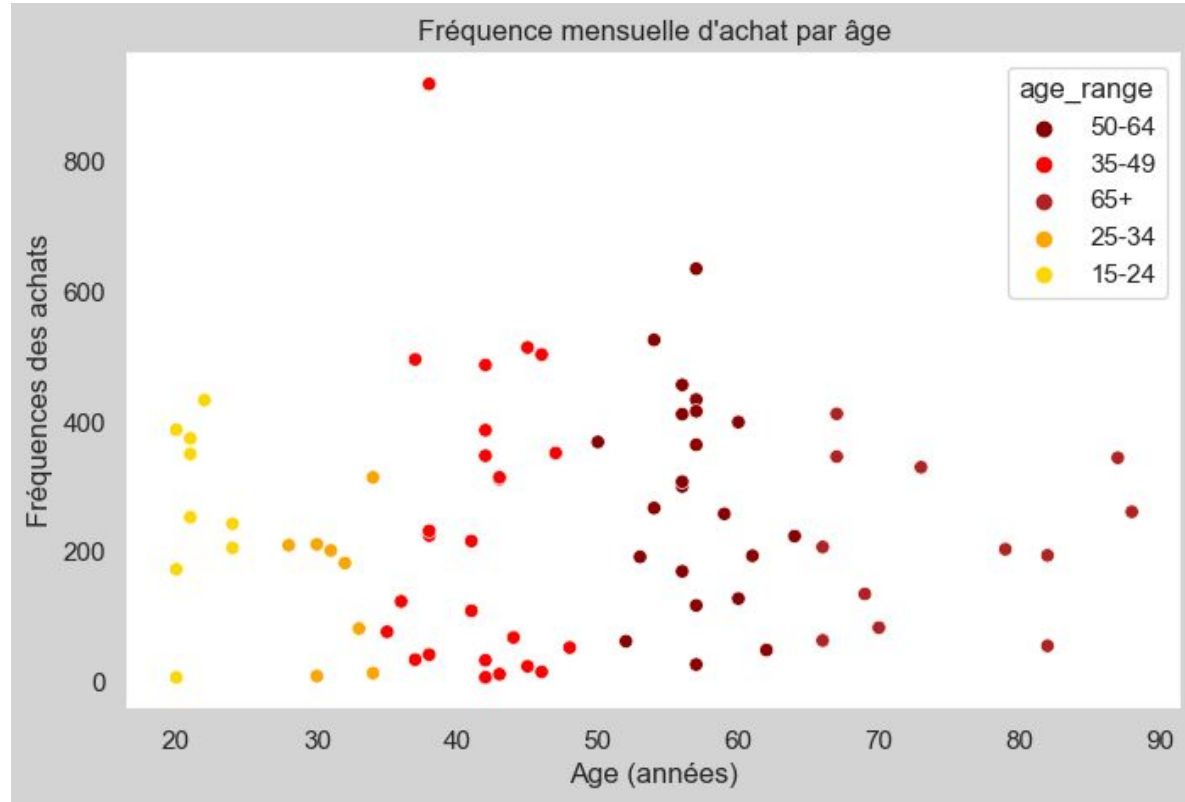
Variables quantitatives :

Test de Spearman :

H0 : Pas de dépendance entre les âges et la fréquence d'achat

H1 : dépendance entre les deux.

RS proche de 0, il n'y a pas de relation monotone entre les deux.



7.5 - Entre l'âge et le panier moyen :

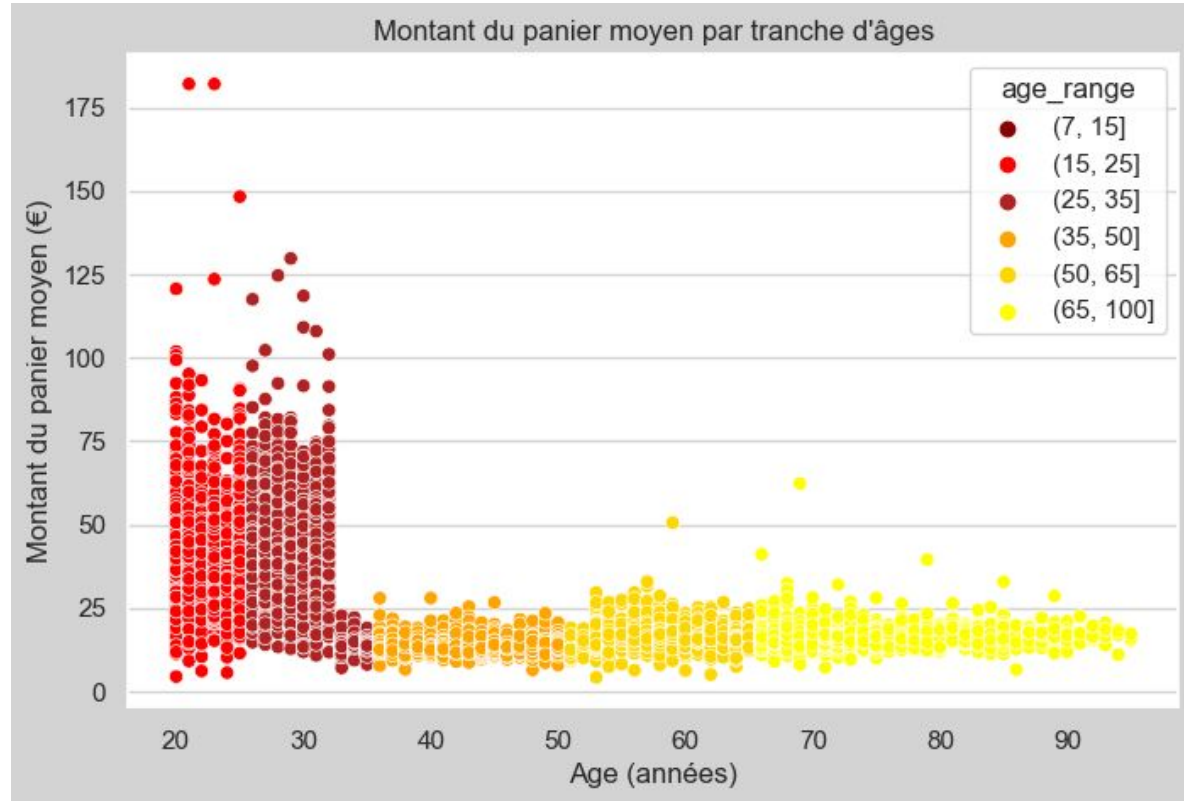
Variables quantitatives :

Test de Spearman :

H0 : Pas de dépendance entre les âges et le panier moyen

H1 : dépendance entre les deux.

RS basculant vers -1, il y a une relation négative entre les deux. Plus les clients sont jeunes, plus le montant du panier est élevé.



En conclusion :

- Problèmes à corriger dans l'acquisition des données : clients achetant de multiples exemplaires du même ouvrage (erreurs ? cadeaux ?), stockage des données non conforme à la 1NF
- Suivi individualisé à mettre en place pour les clients professionnels et établir des bases distinctes pour un calcul des chiffres plus approfondi.
- Historique à compléter : 2 années ne suffisent pas pour permettre la mise en place d'une analyse ouvrant la voie à l'établissement de probabilités.
- Offre produits à revoir : offre de catégorie 2 à élargir (top des ventes - moins de références), offre de catégorie 0 à repenser (beaucoup ont peu de succès), recommandations possibles grâce à une analyse des corrélations entre âge, genre et catégories.

Sur un produit comme le livre, il serait pertinent d'enrichir les informations avec des catégories détaillées par genre littéraire, les prix gagnés, les auteurs, les dates de sorties ou l'existence d'une suite. Ce sont des informations importantes pour pouvoir mieux cibler notre clientèle et faire des offres adaptées.