Lapage

Analyse des ventes

Mars 2022

Présentation du projet

Lapage est une librairie généraliste, possédant plusieurs points de ventes.
Depuis 2 ans elle exploite également un site vente en ligne.

Mission
Faire un bilan de l'activité de vente en ligne

Interlocuteurs :

Annabelle
Responsable marketing

Antoine
Chargé de produit

Julie
Chargée d'étude

Demandes:

Analyse des indicateurs de ventes :
- Analyse du CA
- Analyse de références

Analyse du profil des clients

Corrélations entre :
- L'âge et
  - le montant total des achats
  - le panier
  - les catégories de livres
  - la fréquence d'achat
- Le genre d'un client et catégorie de livre

Plan

- Préparation du jeu de données
- Analyse du jeu de données
- Analyse des ventes
- Analyse du comportement des clients
- Analyse des corrélations
- Test de normalité

Préparation du jeu de données

Nous avons 3 bases de données à notre disposition :

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

products

	id_prod	prix	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

transactions

	id_prod	date	session_id	client_id
0	0_1518	2022-05-20	s_211425	c_103
1	1_251	2022-02-02	s_158752	c_8534
2	0_1277	2022-06-18	s_225667	c_6714
3	2_209	2021-06-24	s_52962	c_6941
4	0_1509	2023-01-11	s_325227	c_4232

Nettoyage

- Les colonnes sont renommées
- Une colonne ‘age’ est ajoutée

- Les colonnes sont renommées
- Suppression de la référence correspondant à la ligne de test

- Suppression de 200 lignes de tests
- Changement du format de la variable ‘date’

Analyse du jeu de données 1/3

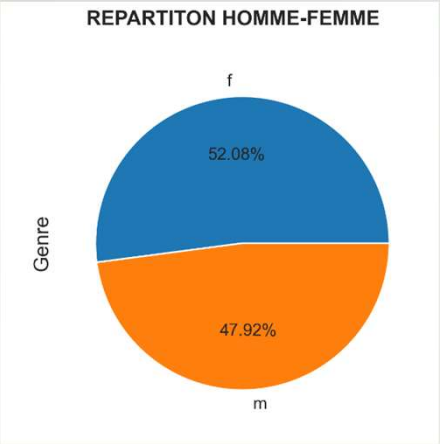
Fichier clients :

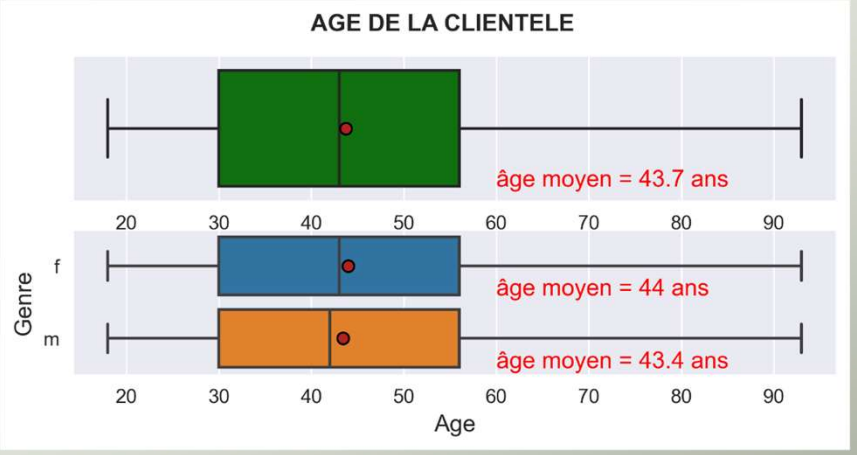
8 623 clients inscrits – 4 494 femmes
- 4 132 hommes

Agés de 18 à 93 ans

Moyenne d'âge 43 ans

50% entre 18 et 43 ans
50% entre 43 et 93 ans





	client_id	genre	naissance	age
0	c_4410	f	1967	55
1	c_7839	f	1975	47
2	c_1699	f	1984	38
3	c_5961	f	1962	60

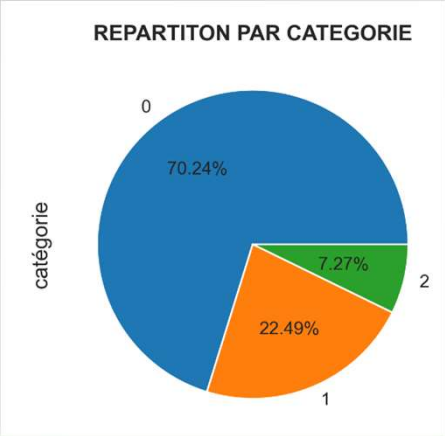
Analyse du jeu de données 2/3

Fichier produits :

3 287 références classées en 3 catégories :

catégorie 0 : 2 308 réfs, prix de 0.62 à 40.99
catégorie 1 : 739 refs, prix de 2.00 à 80.99
catégorie 2 : 239 refs, prix de 30.99 à 300.00

prix moyen 21.85, (0.62 à 300),
75% entre 0.62 et 23.

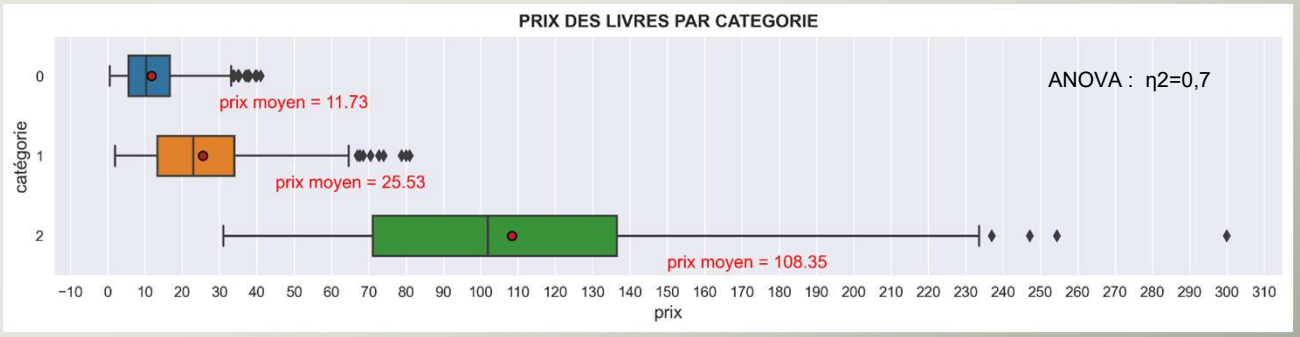


Y-a-t-il une lien entre l'appartenance à une catégorie et le prix d'un livre ?

- Analyse de la variance : ANOVA
$\eta 2 = V_{interclasse}/V_{totale} = 0,7$
$\eta 2$ est proche de 1, le lien est avéré.

- Représentation graphique :
Le boxplot ci-dessous, confirme le calcul

	id_prod	prix	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



Analyse du jeu de données 3/3

Fichier transactions :

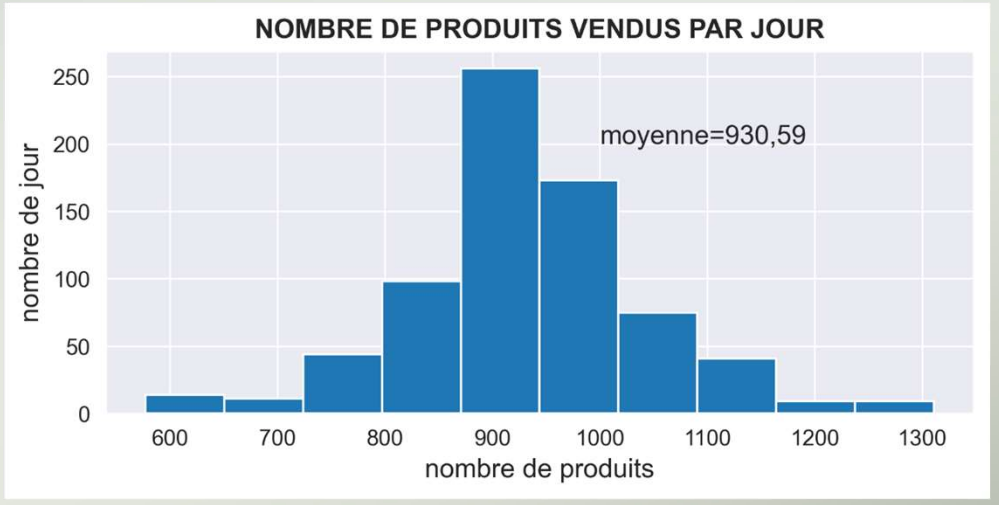
Sur 2 ans, 679 332 transactions ont été réalisées avec 3 266 produits différents(sur 3 287 référencés).
Il y a donc 21 produits invendus.

Le best-seller a été vendu 2 252 fois
(id :1_369, catégorie 1).

Chaque jour, il se vend en moyenne 930 produits

Ces ventes ont été réalisées par 8 600 clients sur 8 623 inscrits, il y a donc 23 clients inactifs

	id_prod	date	session_id	client_id
0	0_1518	2022-05-20	s_211425	c_103
1	1_251	2022-02-02	s_158752	c_8534
2	0_1277	2022-06-18	s_225667	c_6714
3	2_209	2021-06-24	s_52962	c_6941
4	0_1509	2023-01-11	s_325227	c_4232



Analyse des ventes 1/7

Analyse des indicateurs de ventes :
- Analyse du CA
- Analyse de références

Analyse du profil des clients

fusion des df transactions et products

	id_prod	prix	categ	date	session_id	client_id
0	0_1518	4.18	0.0	2022-05-20	s_211425	c_103
1	1_251	15.99	1.0	2022-02-02	s_158752	c_8534
2	0_1277	7.99	0.0	2022-06-18	s_225667	c_6714
3	2_209	69.99	2.0	2021-06-24	s_52962	c_6941
4	0_1509	4.99	0.0	2023-01-11	s_325227	c_4232

Nettoyage

- la fusion fait apparaître 221 valeurs manquantes qui correspondent à un seul produit id = 0_2245, non référencé. Je lui attribue un prix = le prix moyen et une catégorie, categ=1,
- Je modifie le type de la variable 'categ' de float à int64,
- J'ajoute une colonne 'periode' : mois-année

Ventes

	id_prod	prix	categ	date	session_id	client_id	periode
0	0_1518	4.18	0	2022-05-20	s_211425	c_103	2022-05
1	1_251	15.99	1	2022-02-02	s_158752	c_8534	2022-02
2	0_1277	7.99	0	2022-06-18	s_225667	c_6714	2022-06
3	2_209	69.99	2	2021-06-24	s_52962	c_6941	2021-06
4	0_1509	4.99	0	2023-01-11	s_325227	c_4232	2023-01

Analyse des ventes 2/7

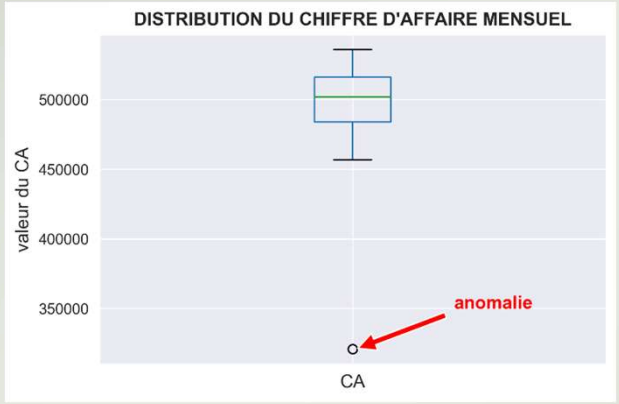
1ères explorations du df ventes :

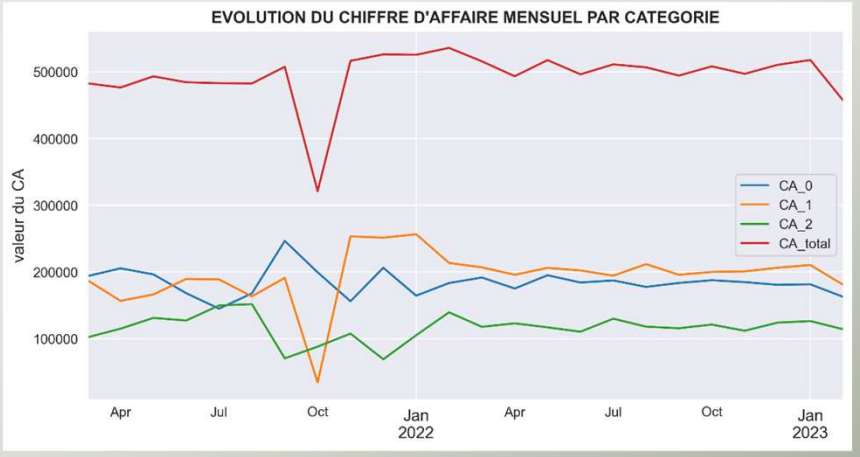
Mise en évidence d'une anomalie :

Le boxplot réalisé sur le CA mensuel, révèle un outlier.

La visualisation temporelle des CA mensuels, explique l'anomalie par l'absence données concernant la catégorie 1, du 04 au 27-10-2021.

Après évaluation de la perte d'information, je neutralise de mois d'octobre 2021.





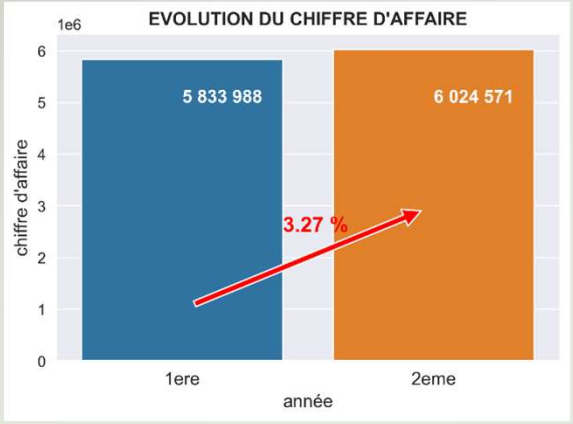
Analyse des ventes 3/7

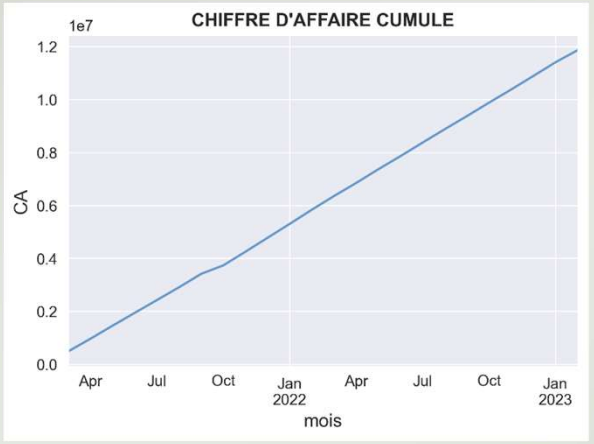
Le chiffre d'affaire:

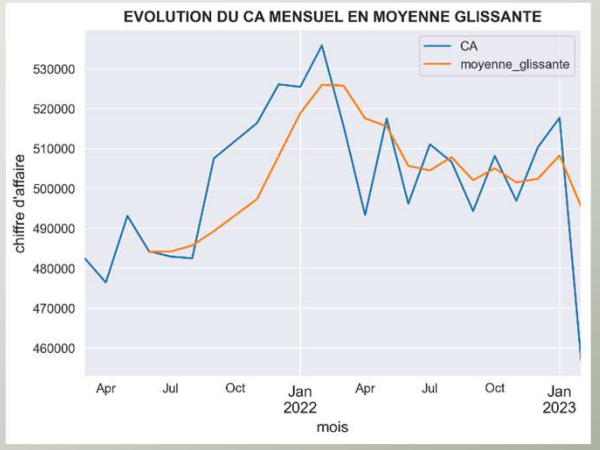
Augmente du CA de 3,27% sur 2 ans

Le CA est régulier d'un mois sur l'autre.

Ebauche d'un saisonnalité, avec une légère hausse sur le 1er trimestre 2022, à confirmer l'année suivante.







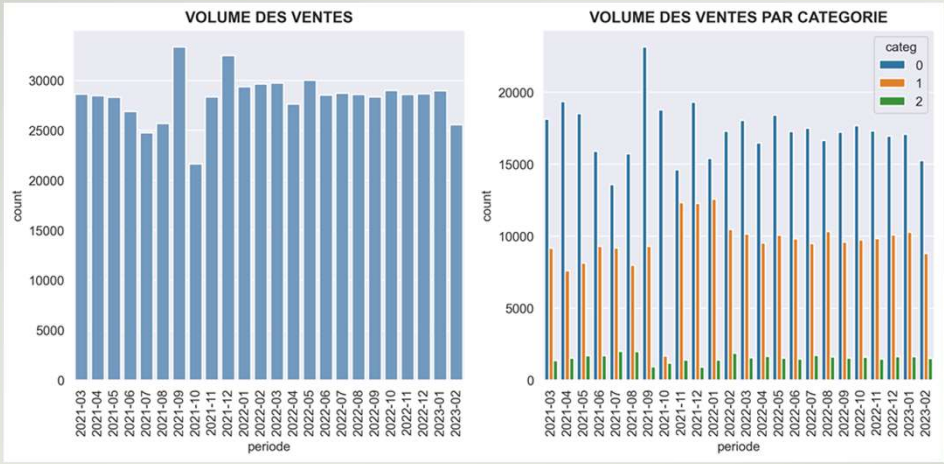
Analyse des ventes 4/7

Volume des ventes :

Le volume des ventes est très variable en fonction de la catégorie.

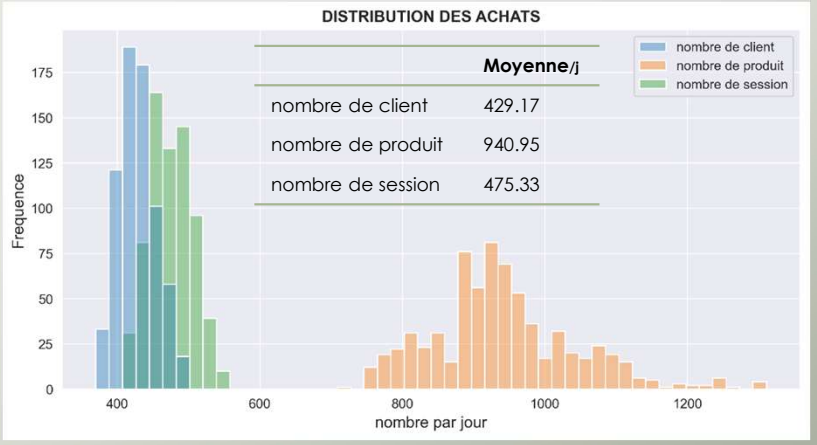
La distribution du nombre de client et du nombre de session est très concentrée contrairement à celle du nombre de produit.

Etonnement, certains clients ont effectué plusieurs sessions d'achat sur une journée.



Pivot sur df ventes : df ventes_jour

date	nombre de client	nombre de produit	CA	nombre de session
2021-03-01	438	963	16587.08	487
2021-03-02	428	940	15508.31	471
2021-03-03	395	911	15198.69	437
2021-03-04	406	903	15196.07	449
2021-03-05	440	943	17471.37	496



Analyse des ventes 5/7

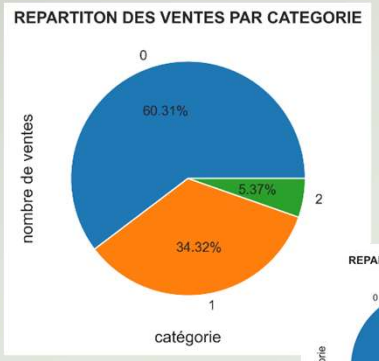
Répartition des ventes par produits :

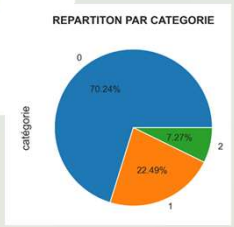
Les produits les moins chers (0) représentent :
60.31% des ventes - 36.58% du CA.

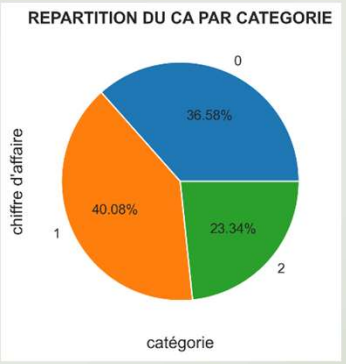
Les produits les plus chers (2) représentent :
5.37% des ventes et 22.34% du CA.

Les produits intermédiaires (1) représentent :
34.32% des ventes et 40.08% du CA.

Les articles constituant le top 5, ont fait l'objet de plus de 2000 ventes.







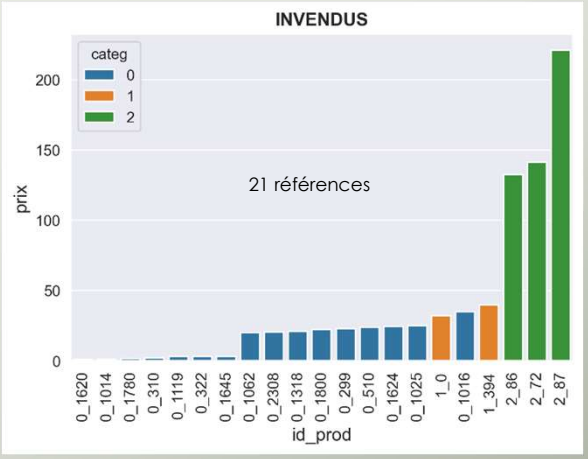
Pivot sur df ventes : df produits

Top 5

id_prod	categ	Nombre de vente
1_369	1	2237
1_417	1	2173
1_414	1	2166
1_498	1	2117
1_425	1	2084

Flop 5

id_prod	categ	Nombre de vente
2_23	2	1
0_1633	0	1
0_1601	0	1
0_1595	0	1
0_1683	0	1

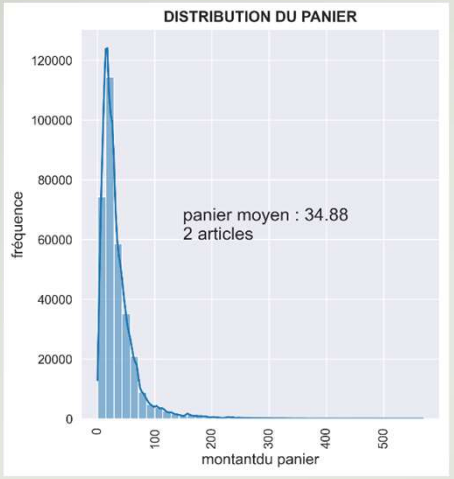


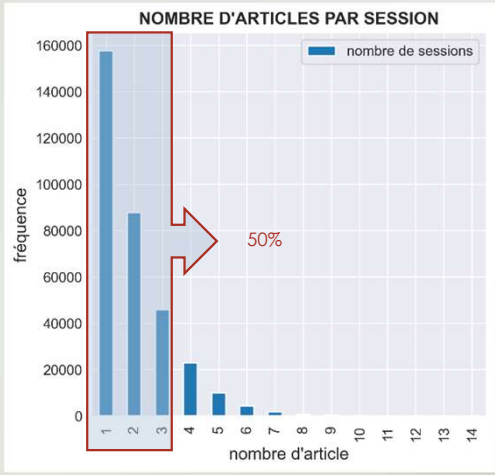
Analyse des ventes 6/7

Répartition des ventes par session :

Contenu : 1à 14 articles, 2 en moyenne
1 à 3 pour 50% des sessions

Panier : 0,62 à 568,88 (€), en moyenne 34,88 (€)
15,81 à 43,14 (€) pour 50% des sessions





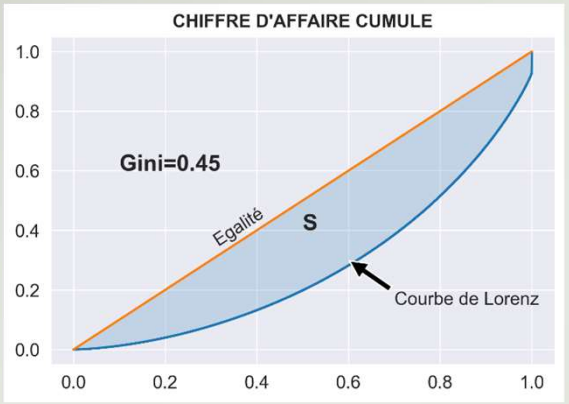
Pivot sur df ventes : df analyse_sessions

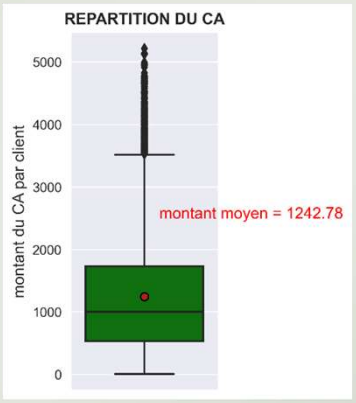
	session_id	nombre d'article	panier
0	s_1	1	11.99
1	s_10	1	26.99
2	s_100	2	33.72
3	s_1000	4	39.22
4	s_10000	3	41.49

Analyse des ventes 7/7

Analyse des indicateurs de ventes :
- Analyse du CA
- Analyse de références

Analyse du profil des clients





Pivot sur df ventes : df ventes_clients

TOP 10

	client_id	nombre de jour d'achat	nombre de produit	CA	nombre de session
677	c_1609	699	24472	312755.08	10538
4387	c_4958	695	5090	282654.61	3764
6336	c_6714	674	8903	149914.91	2511
2723	c_3454	699	6635	111832.29	5444
2108	c_2899	64	105	5214.05	69
634	c_1570	133	356	5136.14	151
2513	c_3263	130	392	5129.89	138
7005	c_7319	134	368	5120.55	142
7790	c_8026	130	368	4991.27	140
4725	c_5263	64	96	4964.87	67

Répartition des ventes par client :

Il y a 4 gros clients, CA>111832.

Pour les autres clients :
- CA de 8,3 à 5214 (€), 1243,78 en moyenne
- Nombre de session de 1 à 164, 36 en moyenne
- Nombre de produits de 1à 392, 71 en moyenne

Il y a des petits et des gros consommateurs mais le CA est régulièrement répartit sur l'ensemble la clientèle.

21 clients inscrits, mais inactifs pourront être relancés.

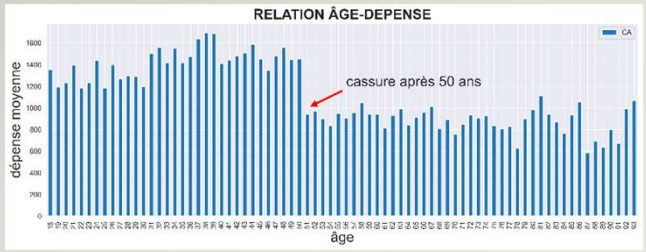
Analyse du comportement des clients 1/6

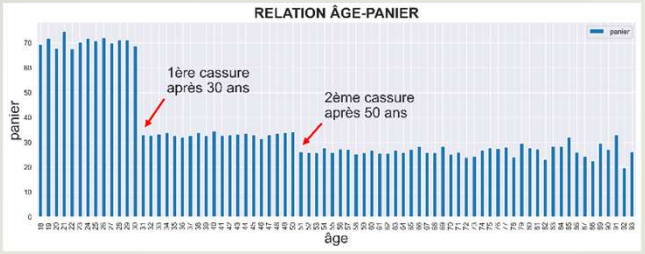
Corrélations entre :

- L'âge et
  - le montant total des achats
  - le panier
  - les catégories de livres
  - la fréquence d'achat
- Le genre d'un client et catégorie de livre

fusion des df ventes_rec et customers : df analyse_clients

	client_id	nombre de jour d'achat	nombre de produit	CA	nombre de session	genre	naissance	age	panier
0	c_1	32	38	550.19	32	m	1955	67	17.19
1	c_10	34	58	1353.60	34	m	1956	66	39.81
2	c_100	5	8	254.85	5	m	1992	30	50.97
3	c_1000	84	122	2209.92	91	f	1966	56	24.28
4	c_1001	41	96	1720.08	44	m	1982	40	39.09

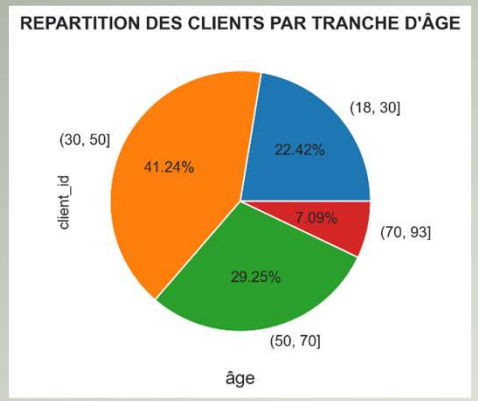




Définition de tranches d'âge : 18 - 30 - 50 - 70 - 93

	client_id	nombre de jour d'achat	nombre de produit	CA	nombre de session	genre	naissance	age	panier	tranche d'âge
0	c_1	32	38	550.19	32	m	1955	67	17.19	50-70
1	c_10	34	58	1353.60	34	m	1956	66	39.81	50-70
2	c_100	5	8	254.85	5	m	1992	30	50.97	18-30
3	c_1000	84	122	2209.92	91	f	1966	56	24.28	50-70
4	c_1001	41	96	1720.08	44	m	1982	40	39.09	30-50

Analyse du comportement des clients 2/6

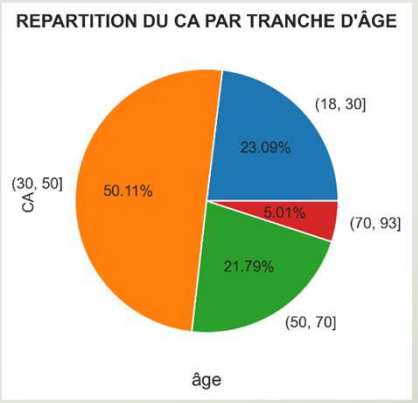


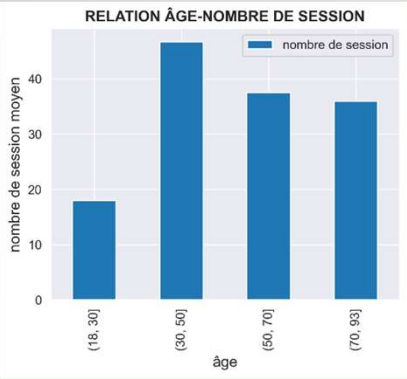
Analyse des ventes en fonction de l'âge :

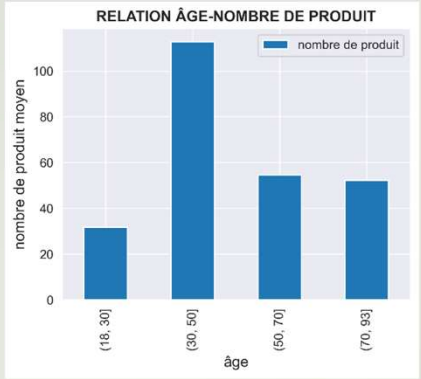
Les 30–50 ans
représentent 41,24% de la clientèle et 50,11% du CA

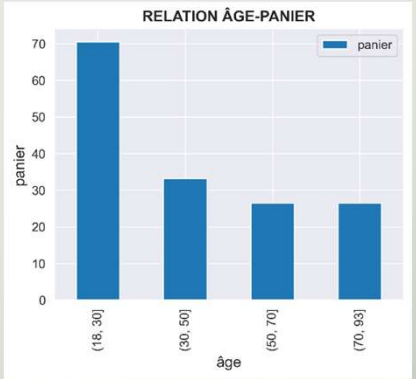
Les 18–30 ans
effectuent moins de sessions d'achat, mais ont le panier moyen le plus élevé, d'où une dépense moyenne importante

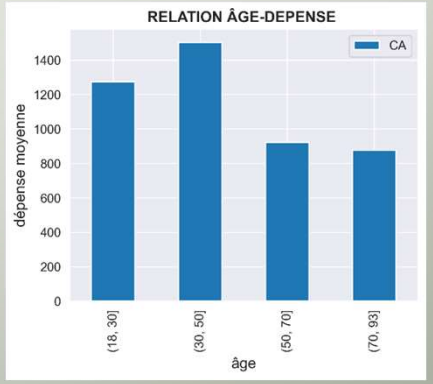
Les 50–70 et 70–93 ont des comportements d'achat similaires



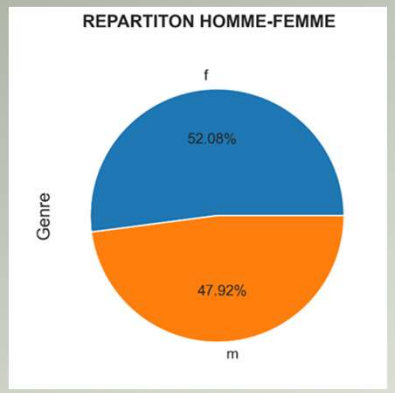






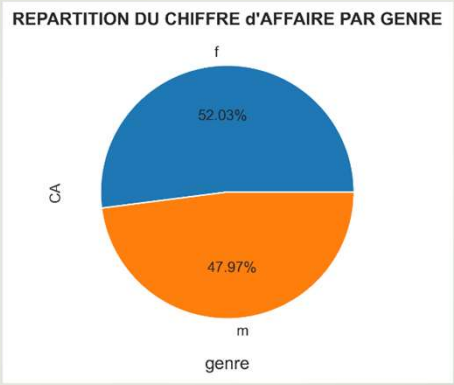


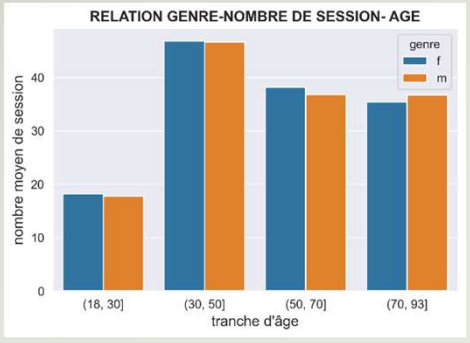
Analyse du comportement des clients 3/6

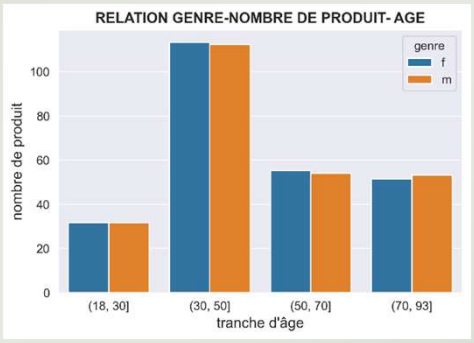


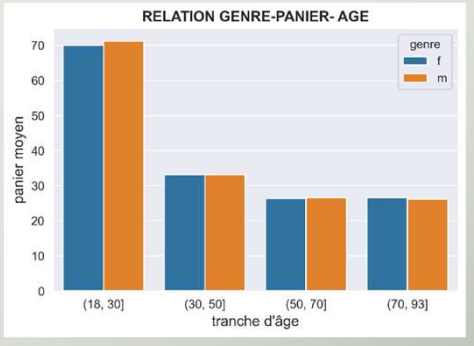
Analyse des ventes en fonction du genre:

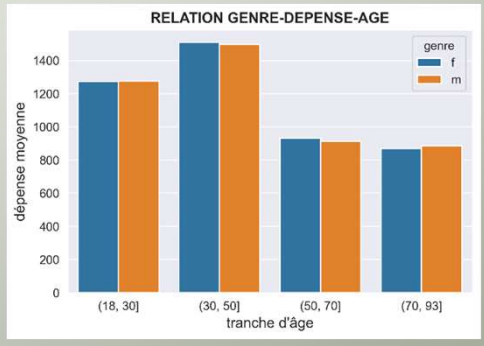
Pas d'influence du genre sur les volumes d'achat, quelque soit l'âge.











Analyse du comportement des clients 4/6

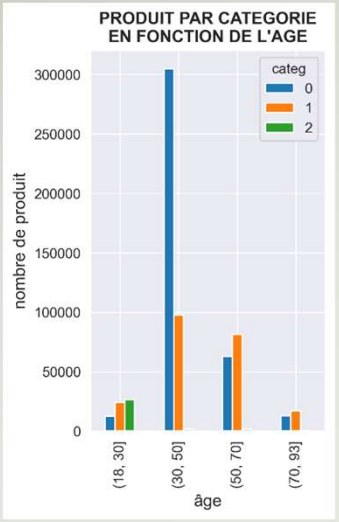
Corrélations entre :

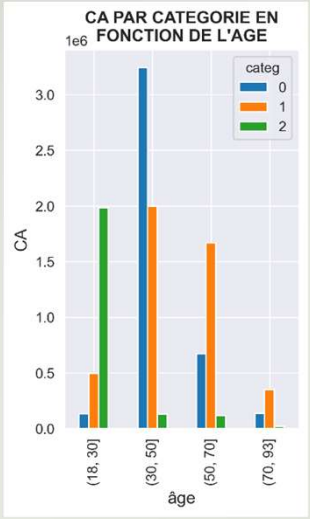
- L'âge et
  - le montant total des achats
  - le panier
  - les catégories de livres
  - la fréquence d'achat

- Le genre d'un client et catégorie de livre

fusion des df ventes_rec et customers : df analyse_ventes

	id_prod	prix	categ	date	session_id	client_id	periode	genre	naissance	age	tranche d'âge
0	0_1518	4.18	0	2022-05-20	s_211425	c_103	2022-05	f	1986	36	30-50
1	1_251	15.99	1	2022-02-02	s_158752	c_8534	2022-02	m	1988	34	30-50
2	0_1277	7.99	0	2022-06-18	s_225667	c_6714	2022-06	f	1968	54	50-70
3	2_209	69.99	2	2021-06-24	s_52962	c_6941	2021-06	m	2000	22	18-30

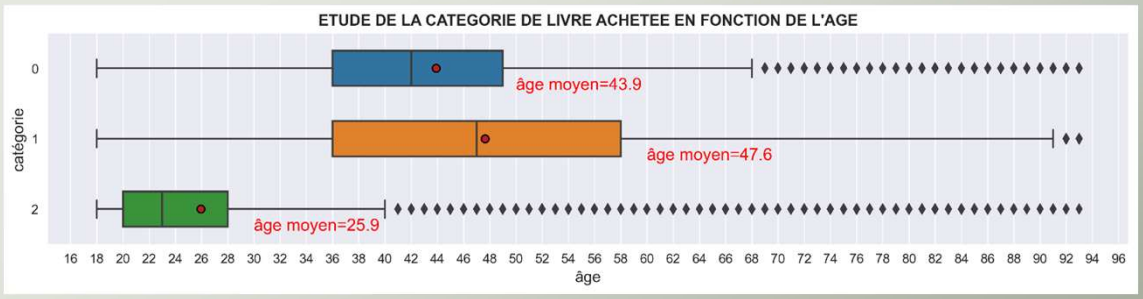




Les 18-30 ans
achètent les 3 catégories mais sont les principaux acheteurs de catégorie 2.

Les 30-50 ans
achètent 3 fois plus de catégorie 0 que de catégorie 1.

Les 50-70 et 70-93 ans
achètent plus de catégorie 1 que de catégorie 0.



Analyse du comportement des clients 5/6

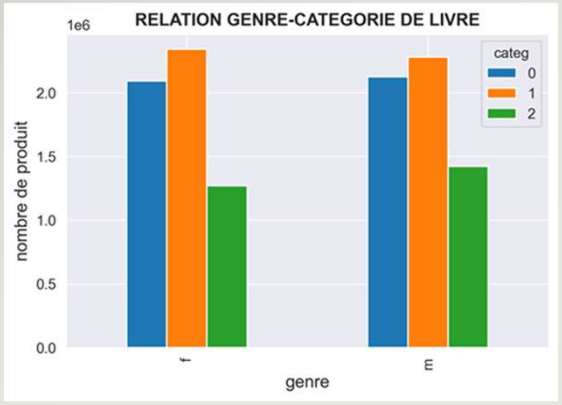
Corrélations entre :

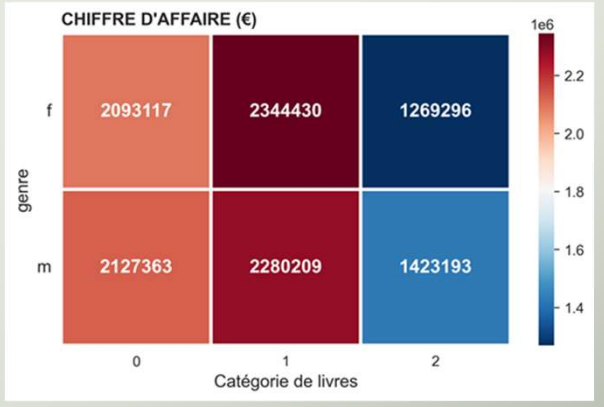
- L'âge et
  - le montant total des achats
  - le panier
  - les catégories de livres
  - la fréquence d'achat

- Le genre d'un client et catégorie de livre

Il n'y a peu d'influence du genre du client sur la catégorie de livre acheté.

Le heatmap fait ressortir que les hommes seraient plus nombreux que les femmes à acheter des catégories 0 et 2, et les femmes plutôt des catégorie 1, (rappelons que les femmes sont légèrement majoritaire au sein de la clientèle).



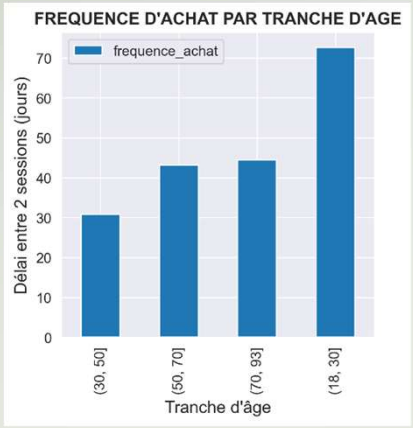


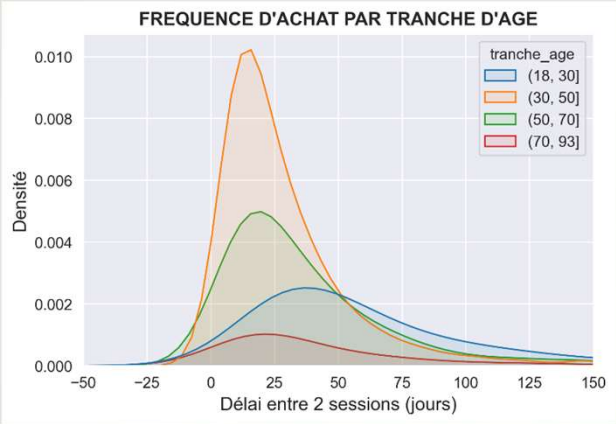
Analyse du comportement des clients 6/6

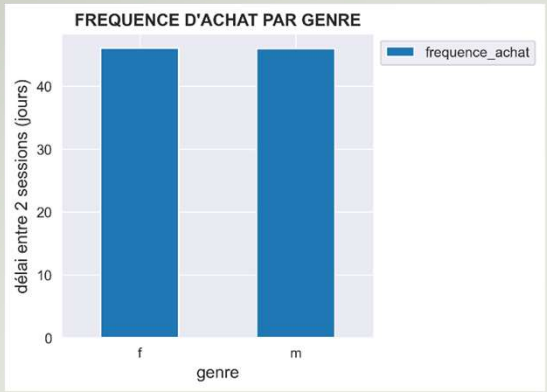
Corrélations entre :

- L'âge et
  - le montant total des achats
  - le panier
  - les catégories de livres
  - la fréquence d'achat
- Le genre d'un client et catégorie de livre

La fréquence d'achat est mesurée par l'attente entre 2 achats.







Les 30-50 ans achètent le plus souvent (tous les 30 jours)
Les 18-30 ans achètent moins souvent (tous les 72 jours)

Le genre n'influence pas la fréquence d'achat.

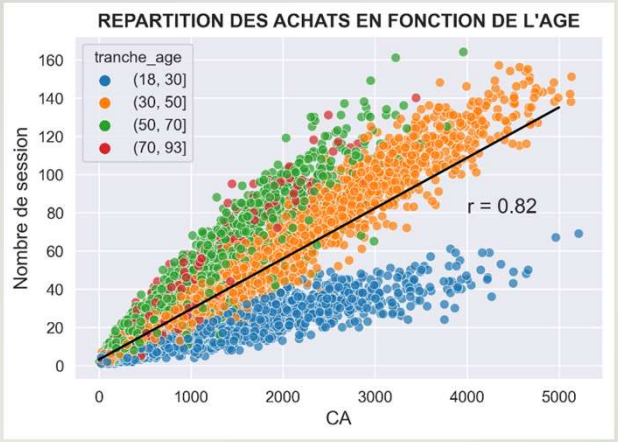
Analyse des corrélations 1/2

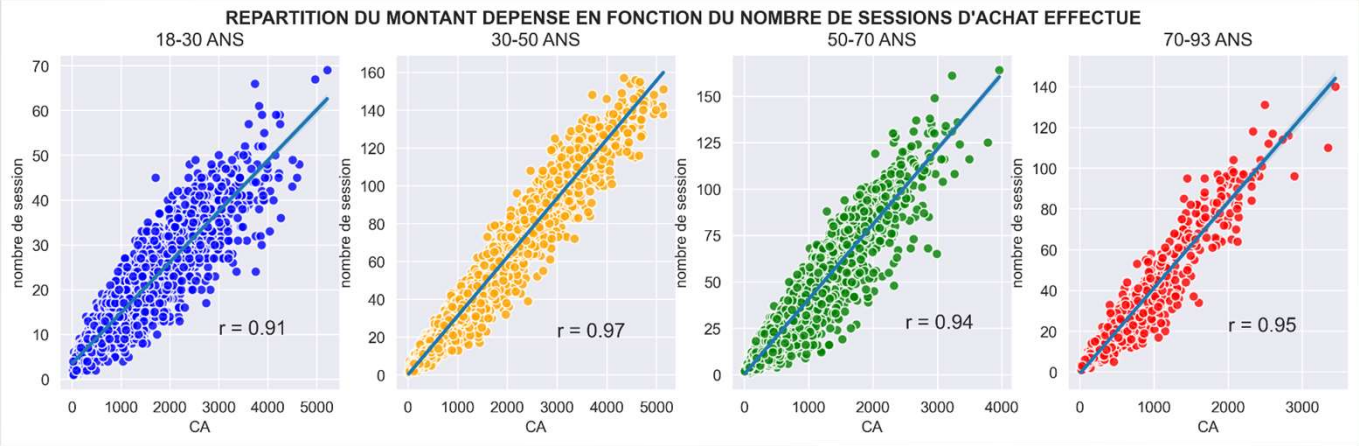
Corrélation CA – nombre de session :

Ce sont 2 variables quantitatives, je vais donc visualiser leur corrélation par un scatterplot complété d'une régression linéaire et du calcul du coefficient de Pearson.

Il y a une relation entre 2 variables :

Les 18-30 ans fractionnent moins leurs achats que les 30-50 ans et que les 50-93 ans.





Analyse des corrélations 2/2

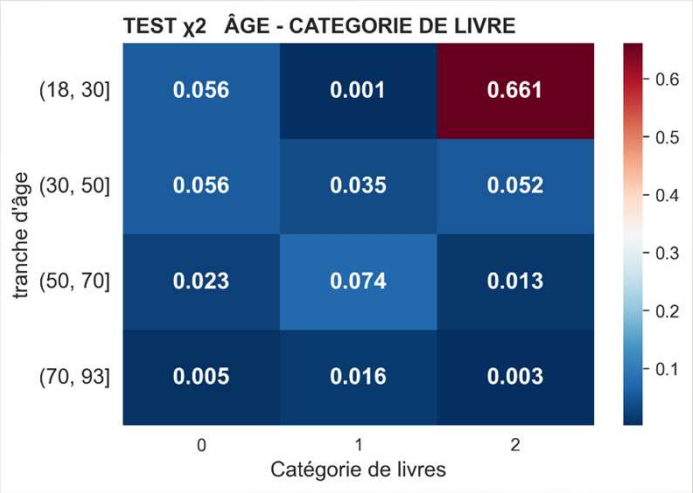
Corrélation tranche d'âge – catégorie :

Ce sont 2 variables qualitatives, je vais donc visualiser leur corrélation par une heatmap complétée d'un test d'indépendance de chi2 ($\chi 2$).

Hypothèses du test :
H0 : pas d'influence d'une variable sur l'autre.
H1 : influence d'une variable sur l'autre.

p value < 0.05, donc H0 est rejetée, les variables ne sont pas indépendantes l'une de l'autre.

Le heatmap nous indique que l'influence la plus forte entre tranche d'âge et catégorie de livre correspond à l'achat de catégorie 2 par les 18 -30 ans.



Test de normalité

sur l'âge de la clientèle suivant le genre :

Skewness : distribution asymétrique, s'étalant vers les âges élevés, il y a de plus une limite d'âge à 18 ans.

Kurtosis : distribution étalée

Test de Bartlett sur les variances : p value = 0,14
p values > 0.05 , les variances des 2 populations sont égales, on peut faire le test de Student sur les moyennes.

Test de Student sur les moyennes : p value = 0,10

Hypothèses du test :

H0 = les 2 moyennes sont égales, il n'y pas d'influence du genre sur la distribution de l'âge des 2 populations.
H1 = les 2 moyennes sont différentes, il y a influence du genre sur la distribution de l'âge des 2 populations.

p values > 0.05, donc on ne rejette pas H0.

Il n'y a donc pas d'influence du genre sur la distribution de l'âge des 2 populations.