ELODIE MENDES

Analyse des ventes d’une librairie en ligne

Librairie présente en physique depuis des années.

Il y a 2 ans, elle décide de se lancer dans la vente en ligne

ANALYSE DES VENTES ET DU COMPORTEMENT CLIENT

Fichier clients

```
RangeIndex: 8621 entries, 0 to 8620
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   client_id  8621 non-null   object
 1   sex        8621 non-null   object
 2   birth      8621 non-null   int64
dtypes: int64(1), object(2)
```

	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

Fichier produits

```
RangeIndex: 3286 entries, 0 to 3285
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   id_prod  3286 non-null   object
 1   price    3286 non-null   float64
 2   categ    3286 non-null   int64
dtypes: float64(1), int64(1), object(1)
```

	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

Fichier transactions

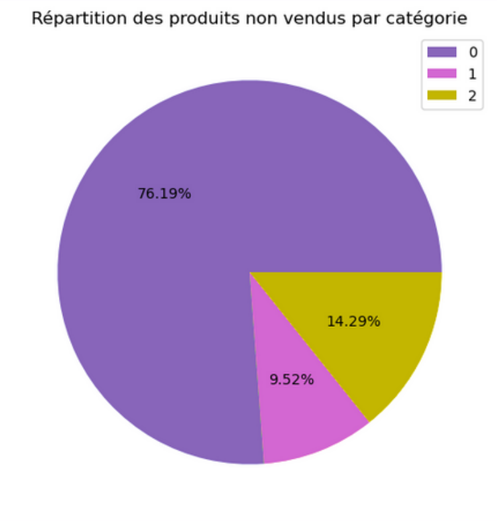
```
RangeIndex: 687534 entries, 0 to 687533
Data columns (total 4 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   id_prod     687534 non-null  object
 1   date        687534 non-null  object
 2   session_id  687534 non-null  object
 3   client_id   687534 non-null  object
dtypes: object(4)
```

	id_prod	date	session_id	client_id
0	0_1259	2021-03-01 00:01:07.843138	s_1	c_329
1	0_1390	2021-03-01 00:02:26.047414	s_2	c_664
2	0_1352	2021-03-01 00:02:38.311413	s_3	c_580
3	0_1458	2021-03-01 00:04:54.559692	s_4	c_7912
4	0_1358	2021-03-01 00:05:18.801198	s_5	c_2033

Jointure produits et transactions

```
df_ventes = pd.merge(produits,transactions, on= "id_prod", how= "outer"
df_ventes.info()
```

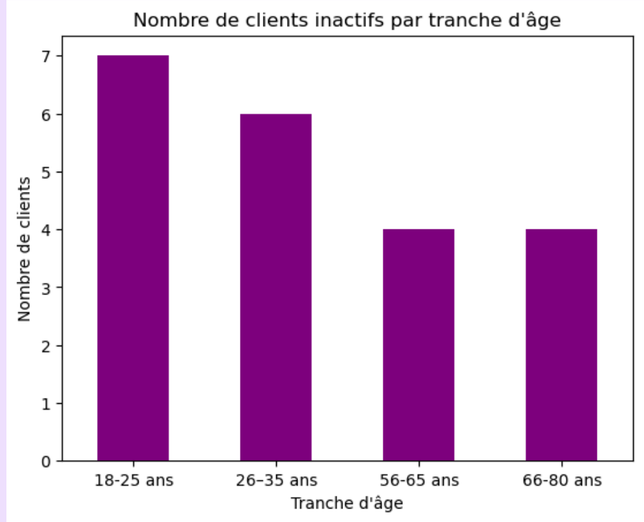
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 687555 entries, 0 to 687554
Data columns (total 6 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   id_prod     687555 non-null  object
 1   price       687555 non-null  float64
 2   categ       687555 non-null  int64
 3   date        687534 non-null  datetime64[ns]
 4   session_id  687534 non-null  object
 5   client_id   687534 non-null  object
```



Jointure transactions et clients

```
#Fusion du fichier 'transactions' et le fichier 'clients'
df_achat_client = pd.merge(transactions,clients, on= "client_id", how= "outer")
df_achat_client.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 687555 entries, 0 to 687554
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   id_prod      687534 non-null  object
 1   date         687534 non-null  datetime64[ns]
 2   session_id   687534 non-null  object
 3   client_id    687555 non-null  object
 4   sex          687555 non-null  object
 5   birth        687555 non-null  int64
 6   âge          687555 non-null  int64
 7   tranche_age  687555 non-null  category
```



Jointure des 3 fichiers (sans BtoB)

```
Index: 640734 entries, 0 to 687533
Data columns (total 12 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   id_prod         640734 non-null  object
 1   price           640734 non-null  float64
 2   categ           640734 non-null  int64
 3   date            640734 non-null  datetime64[ns]
 4   session_id      640734 non-null  object
 5   client_id       640734 non-null  object
 6   mois            640734 non-null  period[M]
 7   sex             640734 non-null  object
 8   birth           640734 non-null  int64
 9   âge             640734 non-null  int64
 10  tranche_age     640734 non-null  category
 11  transaction_id  640734 non-null  int64
```

8596

NOMBRE DE CLIENTS
QUI ONT COMMANDE

+11M€

CHIFFRE D'AFFAIRES
DEPUIS LE LANCEMENT
DU SITE

640734

NOMBRE DE VENTES

35€

PANIER MOYEN
PAR COMMANDE

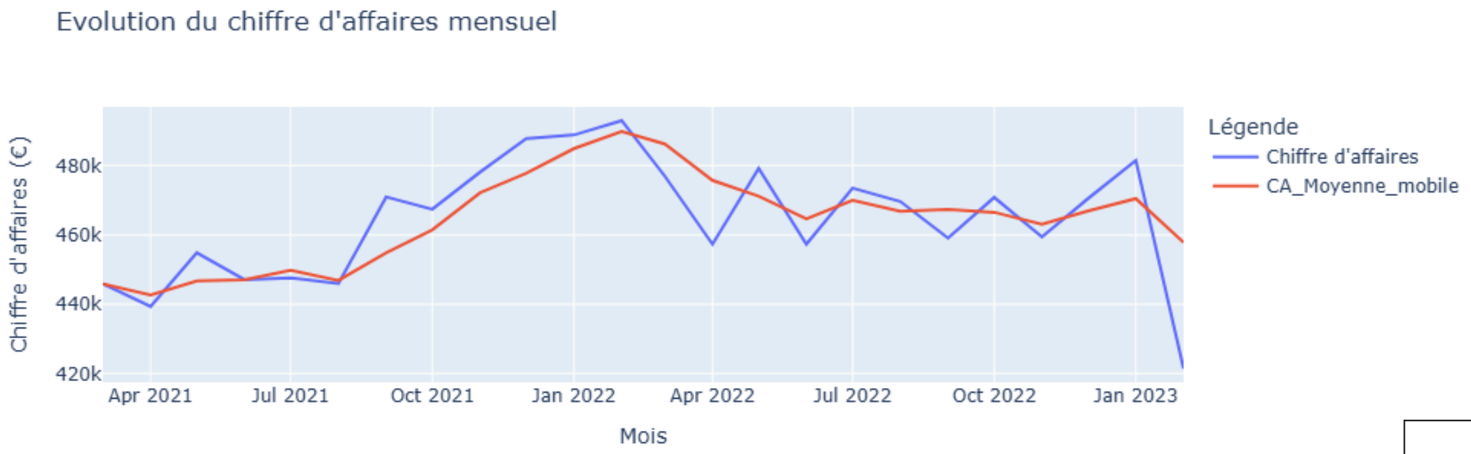
x2

FREQUENCE D'ACHAT
MENSUELLE

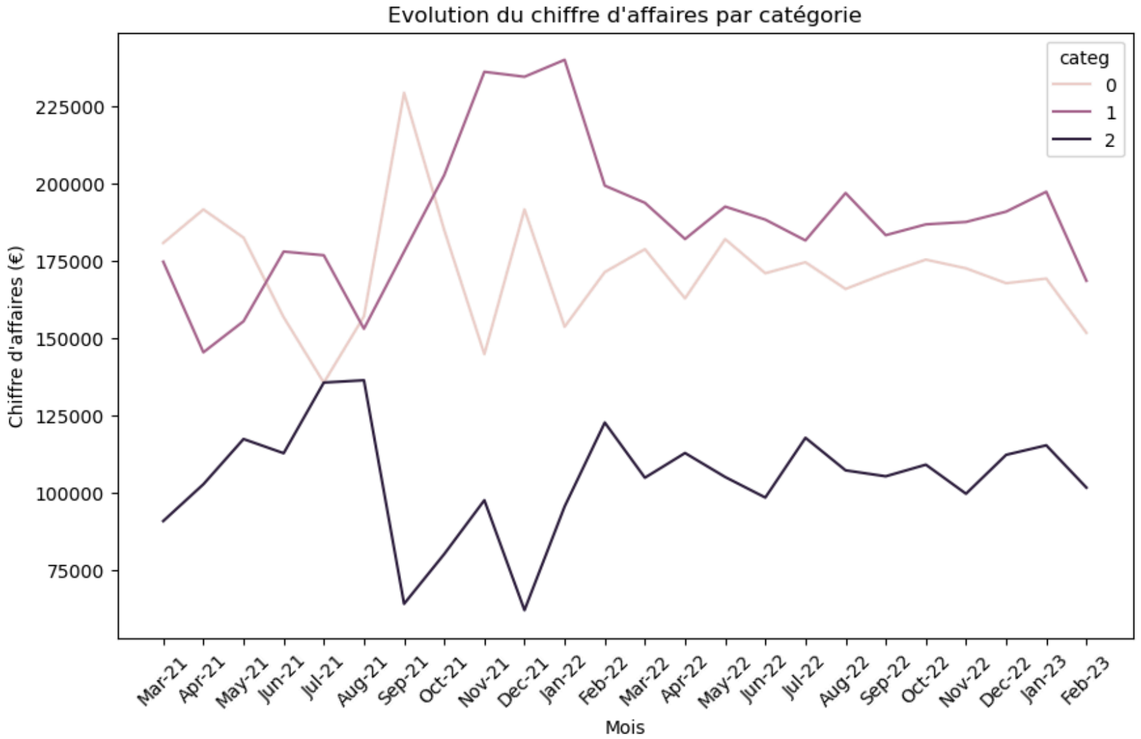
2

MOYENNE DE LIVRE
ACHETE PAR
COMMANDE

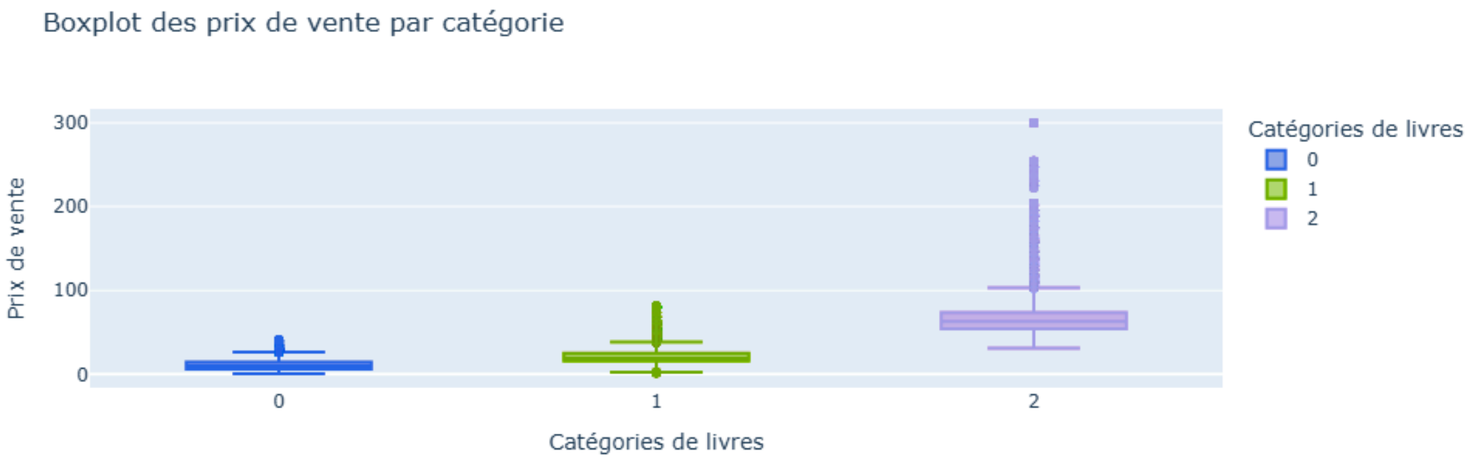
Un chiffre d’affaires contant au fil des mois

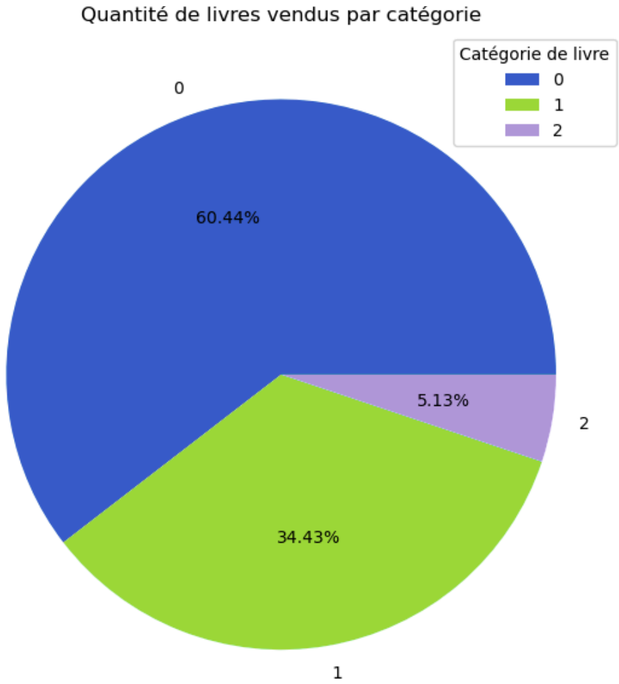


Il est porté par une catégorie



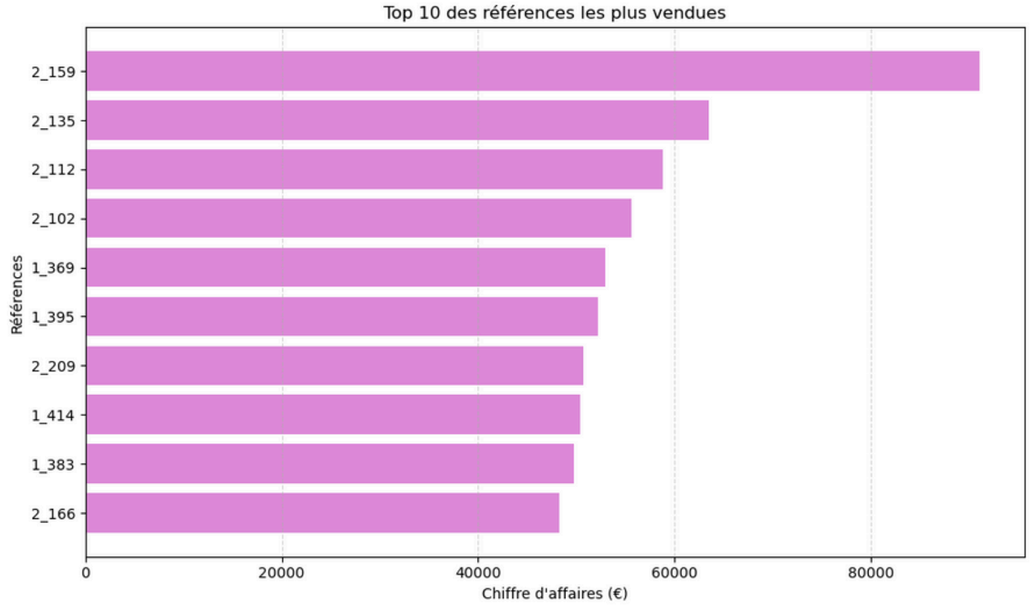
Petits prix, grandes ventes !





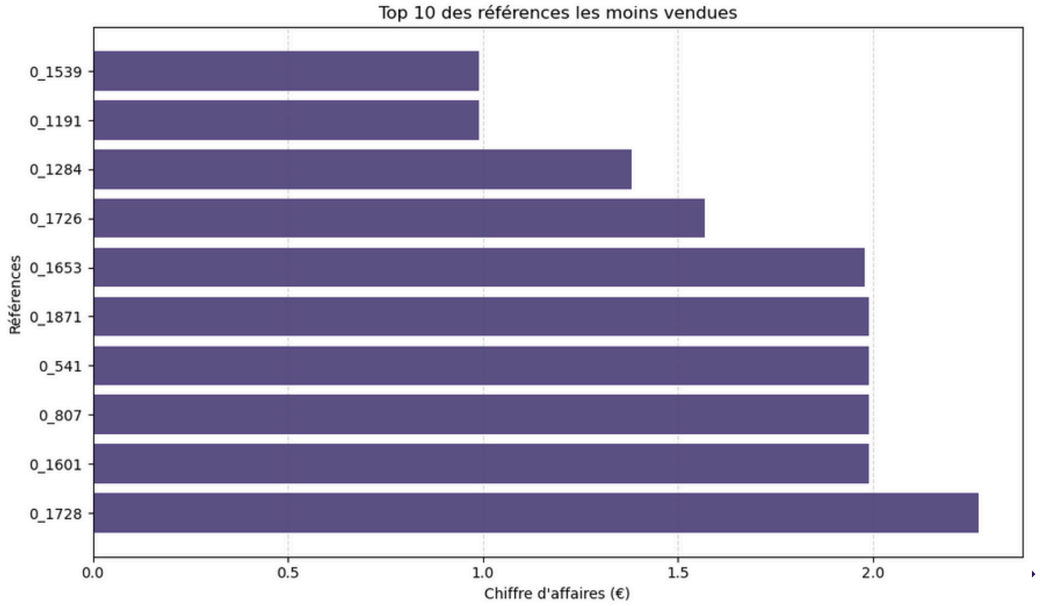
Top 10 des meilleures ventes

categ	CA
1	205203.41
2	368056.35

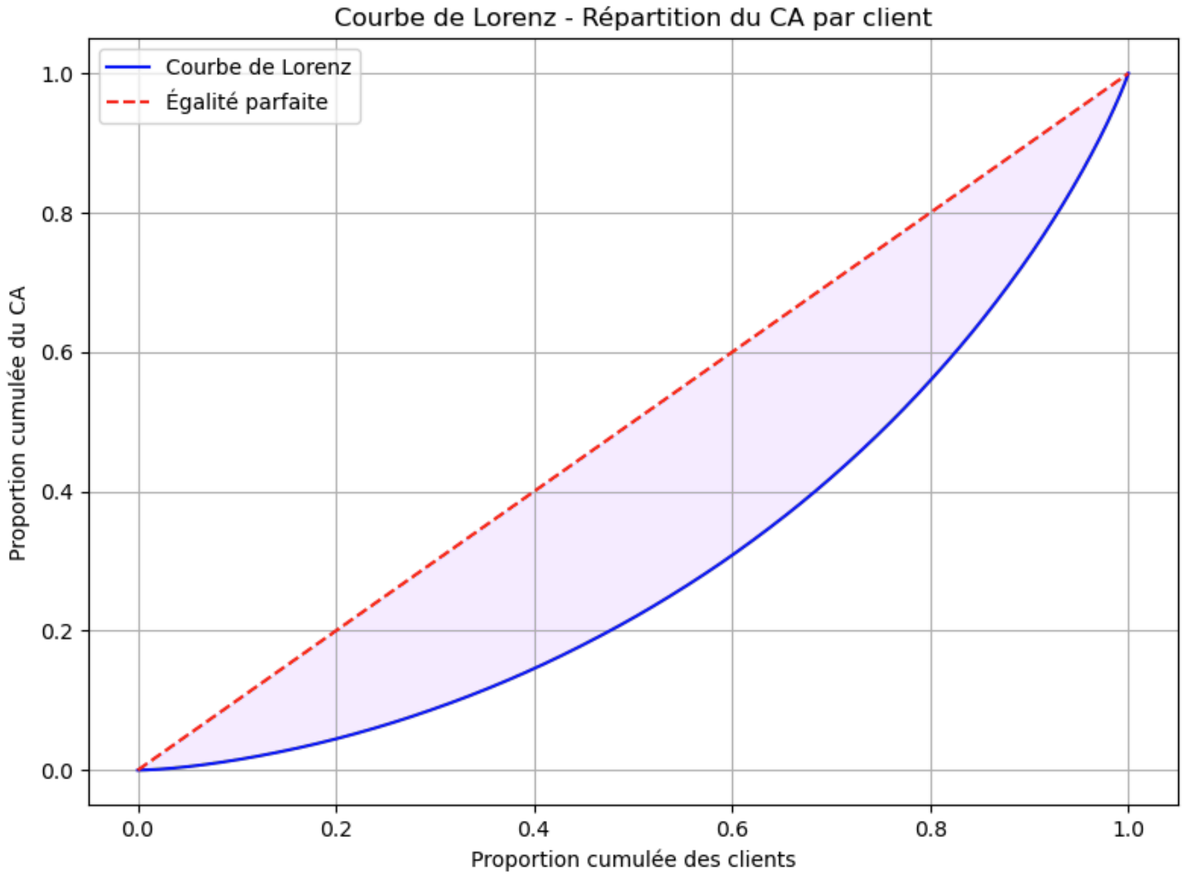


Flop 10 des mauvaises ventes

categ	CA
0	17.14



20% des clients génèrent 45% du chiffres d'affaires

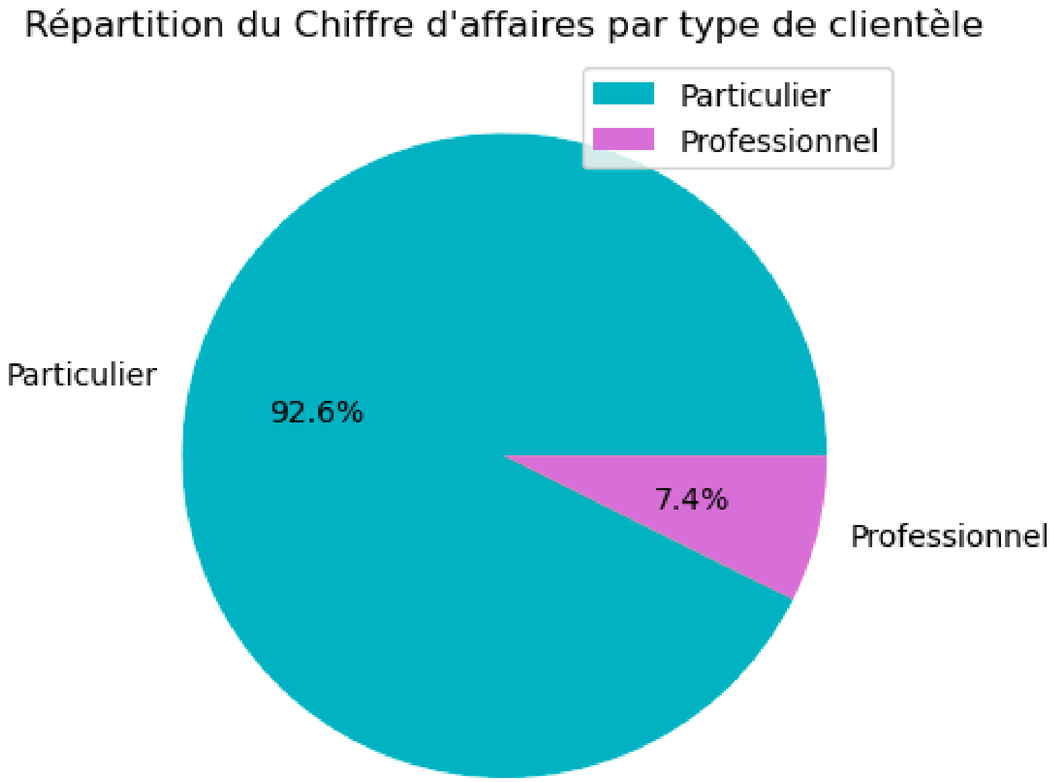


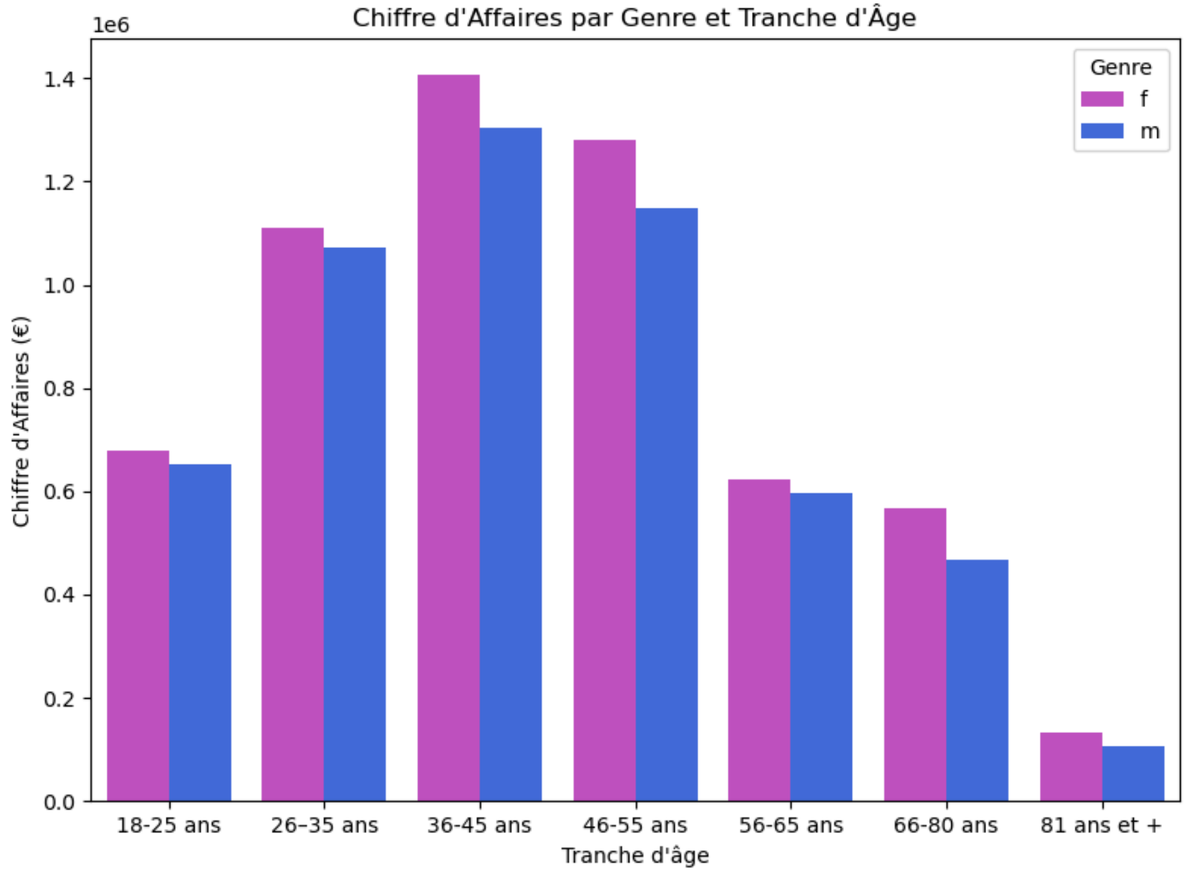

```
# Calcul de l'indice de Gini
gini = 1 - 2 * np.trapezoid(cumul_ca,cumul_client)
print(f"Indice de Gini : {round(gini,3)}")
```

```
Indice de Gini : 0.398
```

- L'indice de Gini mesure l'écart entre la diagonale et la courbe.
- Indice de gini proche de 1 --> fortement inégalitaire*
- Indice de gini proche de 0 --> égalitaire*

Les particuliers entre 36-55 ans boostent le chiffre d’affaires



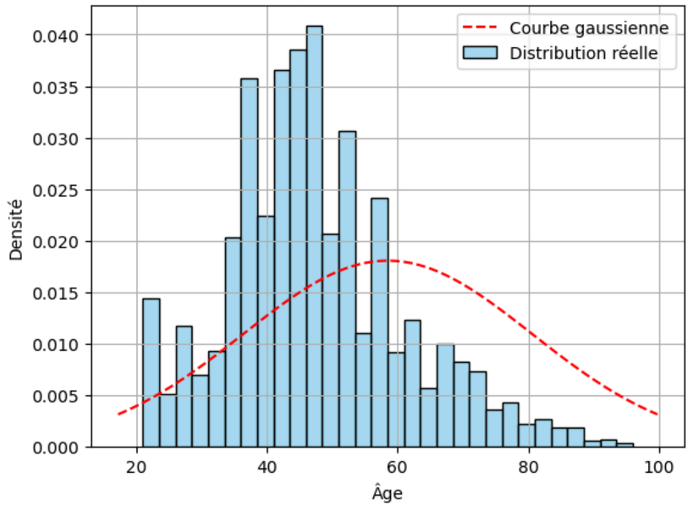


Il y a t-il une corrélation entre l'âge du client et le montant des achats ?

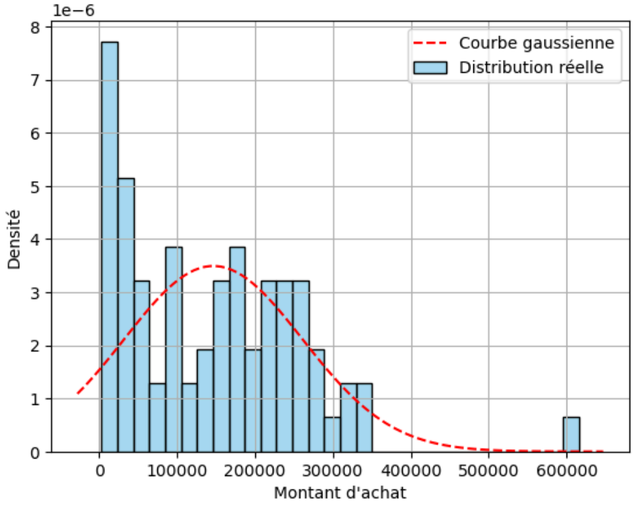
1. Test de Shapiro-Wilk pour vérifier la normalité de la distribution de variables quantitatives

Formulation des hypothèses:

- Hypothèse nulle (H0) : Les données suivent une distribution normale.
- Hypothèse alternative (H1) : Les données ne suivent pas une distribution normale.




```
Statistique du test de Shapiro-Wilk : 0.9549230726696206
Valeur p : 0.008752885621051844
La distribution ne suit pas une loi normale
```




```
Statistique du test de Shapiro-Wilk : 0.9124016052922391
Valeur p : 6.473609171793057e-05
La distribution ne suit pas une loi normale
```

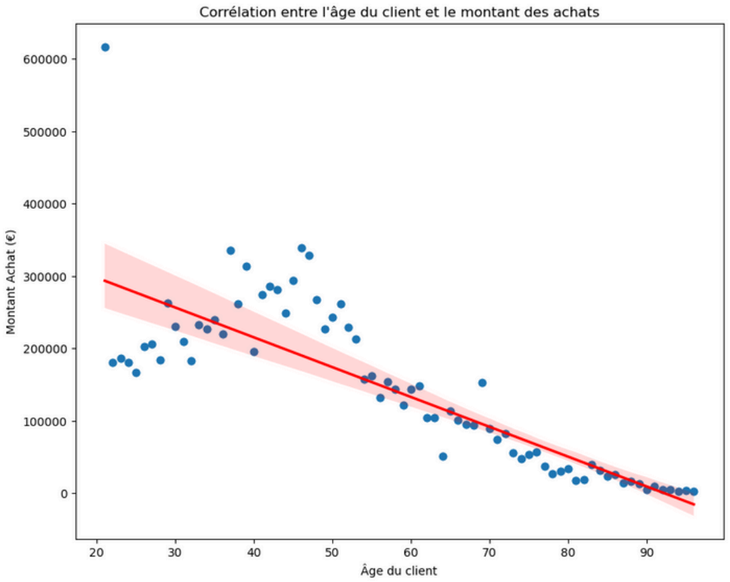
Test de shapiro-wilk (W):

- si W est proche de 1 = parfaite adéquation avec une distribution normale.
- si W est proche de 0 = mauvaise adéquation avec une distribution normale.

Valeur p:

- si p > 0.05 = pas de raison de rejeter H0 (la normalité de la distribution)
- si p < 0.05 = rejetter l'hypothèse H0

2. Test de Spearman



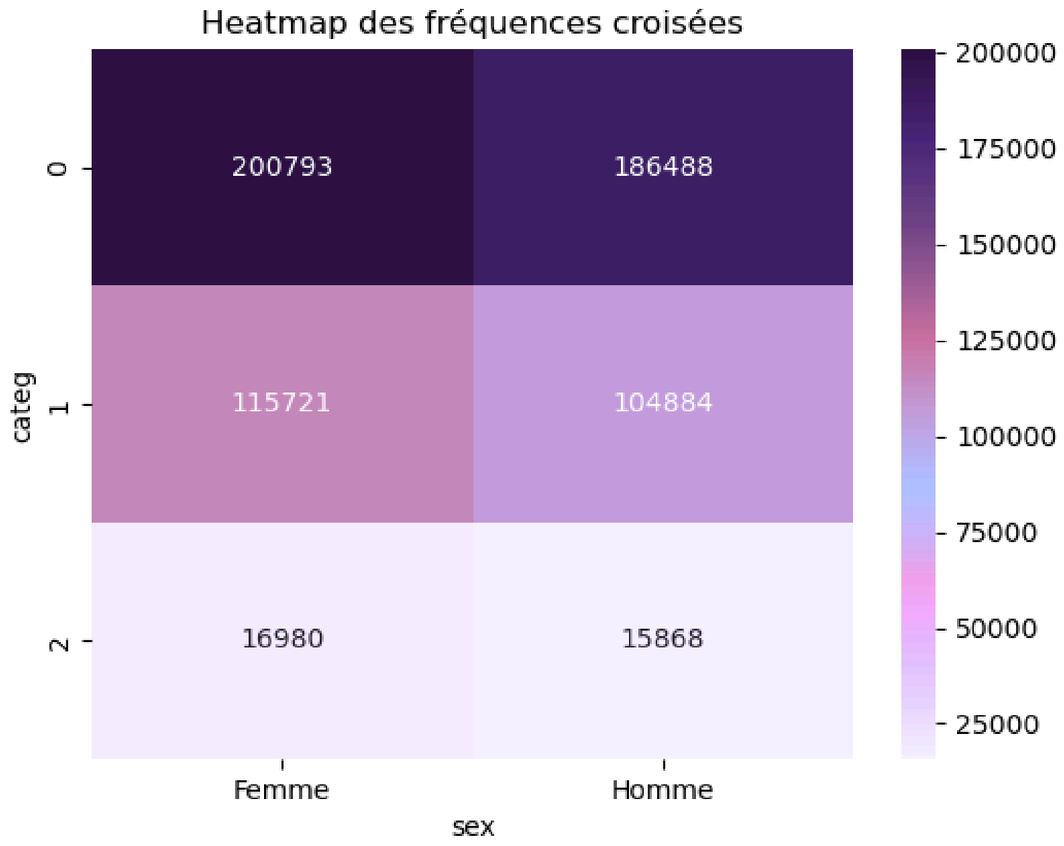

```
Coefficient de corrélation de Spearman: -0.8744497607655503
Valeur p: 5.956077505475151e-25
```

La valeur p du test :

- valeur p < 0.05 = corrélation réelle.
- valeur p > 0.05 = corrélation due au hasard.

Il y a t-il un lien entre le genre d'un client et le type de livre acheté ?

1. Visualisation des fréquences croisées pour les données catégorielles



2. Test de Chi-2

Formulation des hypothèses:

- H0 : Le type de livre acheté est indépendant du genre du client.
- H1 : Le type de livre acheté dépend du genre du client.

```
Statistique Chi-2: 22.66856665178056
Valeur p: 1.1955928116587024e-05
Degrés de liberté: 2
Fréquences attendues:
[[201574.89662481 185706.10337519]
 [114822.13191434 105782.86808566]
 [ 17096.97146086  15751.02853914]]
```

La valeur p est utilisée pour décider si nous rejetons l'hypothèse nulle H0

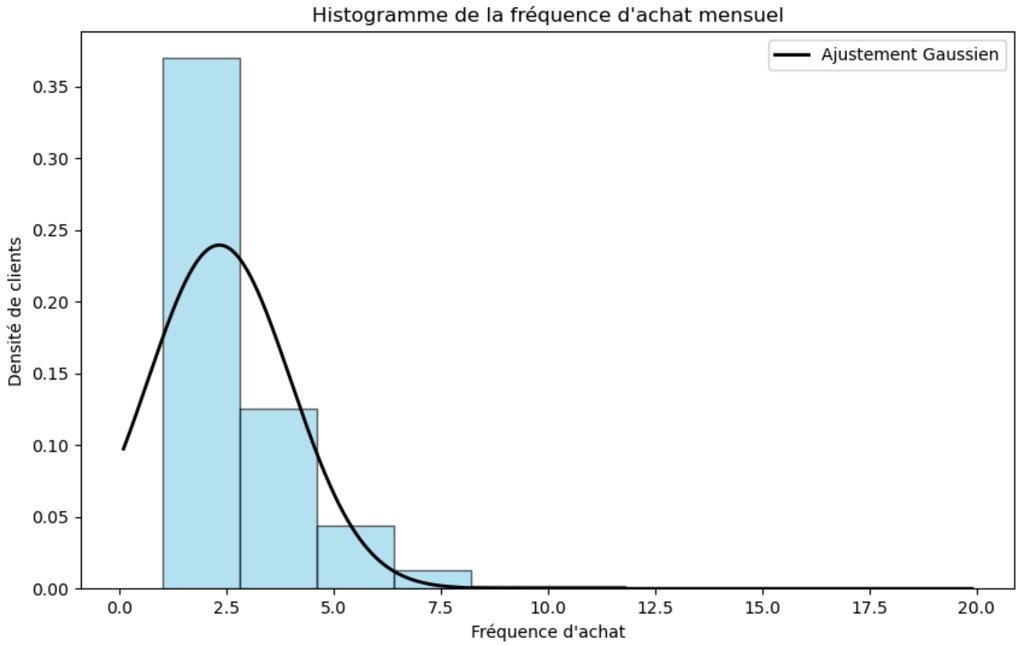
Une valeur p :

valeur p < 0.05 = corrélation réelle.

valeur p > 0.05 = corrélation due au hasard.

Il y a t-il une corrélation entre l’âge du client et la fréquence d’achat ?

1. Test de Shapiro-Wilk pour vérifier la normalité de la distribution de la fréquence d’achat




```
Statistique du test de Shapiro-Wilk : 0.6859536837525061
Valeur p : 2.0441181046246753e-70
Les données ne suivent pas une loi normale
```

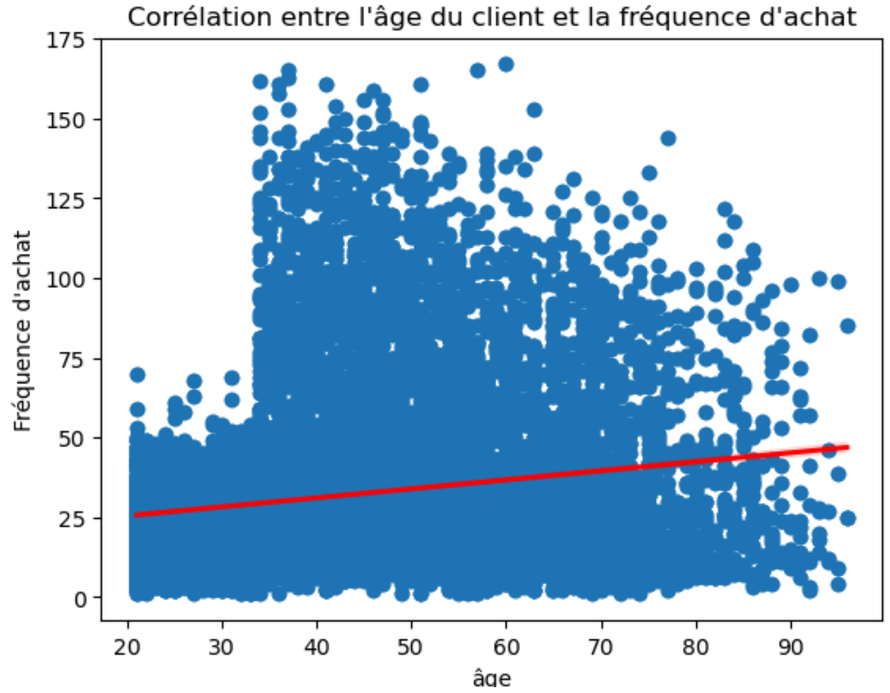
Formulation des hypothèses:

- Hypothèse nulle (H0) : Les données suivent une distribution normale.
- Hypothèse alternative (H1) : Les données ne suivent pas une distribution normale.

Valeur p:

- si p > 0.05 = pas de raison de rejeter H0 (la normalité de la distribution)
- si p < 0.05 = rejetter l'hypothèse H0

2. Test de Spearman



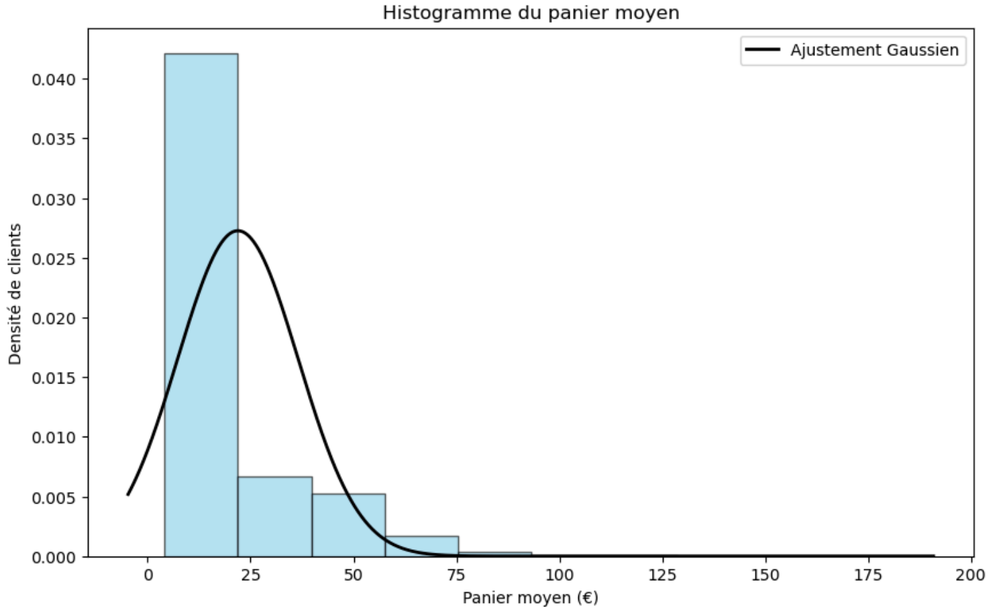

```
Coefficient de corrélation de Spearman: 0.21196373259671872
Valeur p: 6.629168433162815e-88
```

La valeur p du test :

- valeur p < 0.05 = corrélation réelle.
- valeur p > 0.05 = corrélation due au hasard.

Il y a t-il une corrélation entre l'âge du client et la taille du panier moyen ?

1. Test de Shapiro-Wilk pour vérifier la normalité de la distribution




```
Statistique du test de Shapiro-Wilk : 0.6831193605882802
Valeur p : 1.3445527402834681e-70
Les données ne suivent pas une loi normale
```

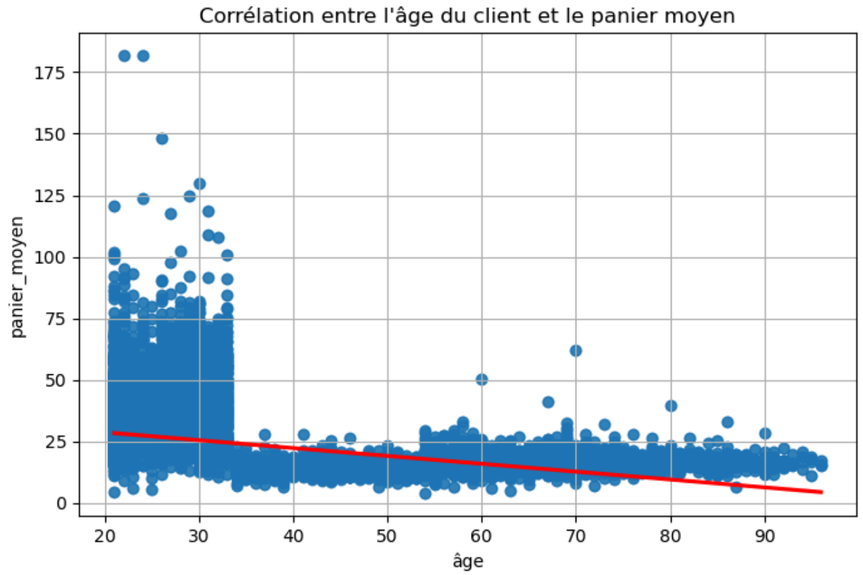
Valeur p:

- si p > 0.05 = pas de raison de rejeter H0 (la normalité de la distribution)
- si p < 0.05 = rejetter l'hypothèse H0

Formulation des hypothèses:

- Hypothèse nulle (H0) : Les données suivent une distribution normale.
- Hypothèse alternative (H1) : Les données ne suivent pas une distribution normale.

2. Test de Spearman



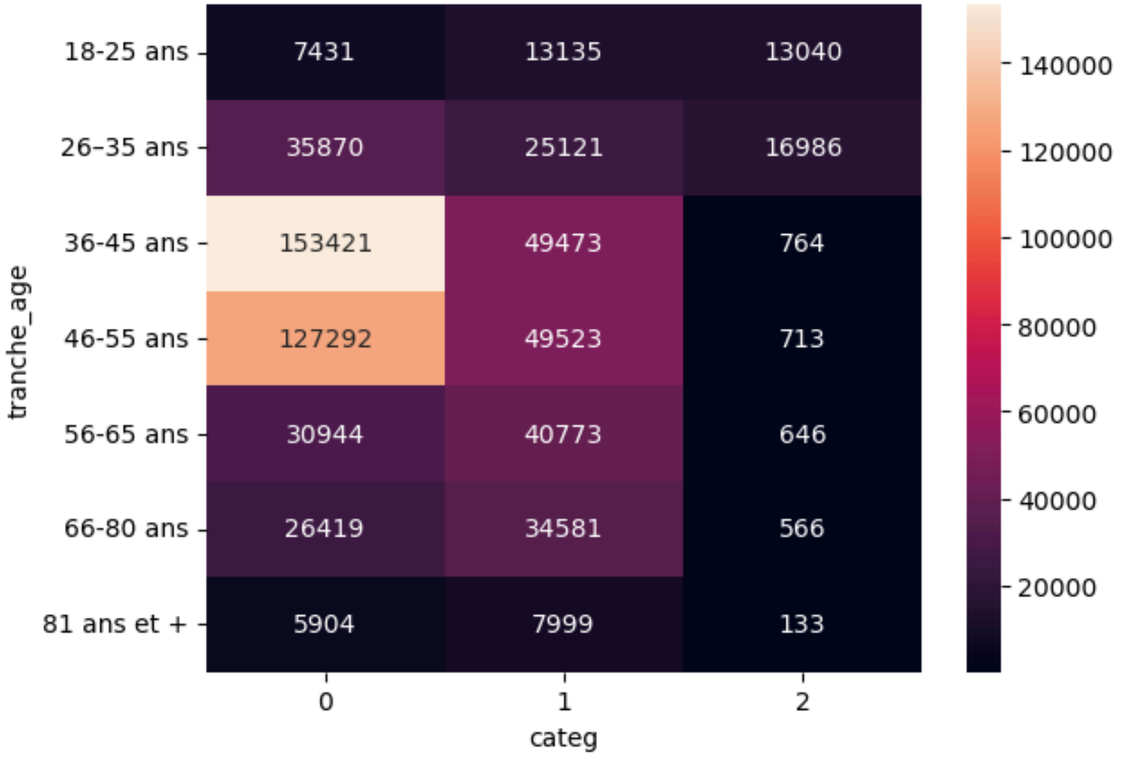

```
Coefficient de corrélation de Spearman: -0.32587401420827466
Valeur p: 8.202257290884015e-212
```

La valeur p du test :

- valeur p < 0.05 = corrélation réelle.
- valeur p > 0.05 = corrélation due au hasard.

Il y a t-il un lien entre l'âge du client et la catégorie de livre achetés ?

1. Visualisation des fréquences croisées pour les données catégorielles



2. Test de Kruskal Wallis

- H0 (hypothèse nulle) : les distributions d'âge sont identiques entre les catégories.
- H1 (hypothèse alternative) : au moins une catégorie a une distribution d'âges différente.

```
#Création d'une liste d'âge pour chaque catégorie de livre acheté pour le test de Kruskal
cat_0 = df_merge_filtré[df_merge_filtré['categ'] == '0']['âge']
cat_1 = df_merge_filtré[df_merge_filtré['categ'] == '1']['âge']
cat_2 = df_merge_filtré[df_merge_filtré['categ'] == '2']['âge']
```

```
# Test de Kruskal-Wallis
h_stat, p_value = kruskal(cat_0, cat_1, cat_2)

print(f"Statistique du test de Kruskal:{h_stat}")
print(f"Valeur p:{p_value}")
```

```
Statistique du test de Kruskal:71359.73412120914
Valeur p:0.0
```

La valeur p est utilisée pour décider si nous rejetons l'hypothèse nulle H0

Une valeur p :

valeur p < 0.05 = corrélation réelle.

valeur p > 0.05 = corrélation due au hasard.

LES PLUS JEUNES ONT UN PANIER MOYEN PLUS ÉLEVÉ QUE LES PLUS ÂGÉS :

- POUR LES CLIENTS DE 18 À 35 ANS , PROPOSER DES PACKS DE LIVRES PREMIUMS.
- RÉCOMPENSER LEUR ACHAT EN DONNANT DES AVANTAGES SUR LES NOUVEAUTÉS OU SUR LES LIVRES DE CATÉGORIE 2 QUI SONT LES PLUS CHERS DE L'OFFRE.

LES CLIENTS PLUS AGÉS COMMANDENT PLUS RÉGULIÈREMENT QUE LES PLUS JEUNES:

- CE SONT DES CLIENTS FIDÈLES À QUI ONT PEUT PROPOSER UN ABONNEMENT MENSUEL OU ANNUEL.
- DES POINTS DE FIDÉLITÉ PEUVENT LEUR ÊTRE OFFERT À CHAQUE ACHAT EFFECTUÉ.

LES HOMMES ET LES FEMMES ONT LEURS CATÉGORIES DE LIVRES PRÉFÉRÉES :

- SELON LE GENRE DU CLIENT, PROPOSER LES BESTSELLERS OU NOUVEAUTÉS DANS LA CATÉGORIE QUI L'INTÉRESSE (VIA DES NEWSLETTERS OU SUR LA PAGE D'ACCUEIL).