Projet-4

Analysez les ventes de votre entreprise

Présenté par-
Manu Sharma
(Data Analyst)

Objectif du Projet

Analyser les données de vente d'une grande chaîne de librairies selon une approche de haut en bas

- ***Définir la mission***
- ***Collecter les données***
- ***Nettoyer les données***
- ***Analyser les données***
- ***Visualiser et partager les résultats***

Description du projet

Les prérequis pour ce projet sont de connaître R ou Python, donc je vais travailler sur python, et de savoir manipuler des Data frames (disponibles via la librairie Pandas en Python).

Il faut aussi connaître les bases de la statistique descriptive (moyenne, médiane, variance, représentations graphiques, tests de corrélation, analyse bivariée, etc).

Donc, je vais prendre quelques leçons à travers des vidéos sur le dossier open Classroom pour me rafraîchir la mémoire

Scénario

Je suis analyste de données pour une grande chaîne de librairies, fraîchement embauché depuis une semaine ! J'ai rencontré vos collègues, votre nouveau bureau.

Maintenant, je dois faire quelques missions

Le département informatique m'a donné accès à la base de données des ventes. Je dois maintenant me familiariser avec les données et les analyser.

Maintenant, j'ai la base de données qui a été donnée dans le projet.

Les données

Voici les fichiers CSV à votre disposition :

les ventes (appellees "Transactions") ;

la liste des clients ;

la liste des produits.

Mission du projet

Mission 1

- Il faut faire un peu de ménage ! Par exemple, je vais devoir faire des choix concernant le traitement des valeurs manquantes et des valeurs aberrantes.

Mission 2

- Pour mieux comprendre les ventes.
- Je vais utiliser des indicateurs de tendance centrale et de dispersion
- l'analyse de la concentration, via une courbe de Lorenz et un indice de Gini
- des représentations graphiques, dont au moins un histogramme, une représentation en " boîte et moustaches " et une représentation en série chronologique (c'est-à-dire un graphique dont l'axe des x représente des dates) ;
- analyses bivariées

Mission 3

- Voici quelques questions supplémentaires

- Existe-t-il une corrélation entre le sexe des clients et les catégories de produits achetés ?
- Existe-t-il une corrélation entre l'âge des clients et
- Le montant total des achats
- La fréquence d'achat (c'est-à-dire le nombre d'achats par mois, par exemple)
- La taille moyenne du panier (nombre d'articles)
- Les catégories de produits achetés.

Données téléchargées

les ventes (appelées "Transactions")

- (id_prod: Product_id, date, session_id, client_id)
- *Id_prod: 337016 (count)* → *3266 unique valuer*
- *trois cent trente-sept mille seize* → *trois mille deux cent soixante-six*

Original Records	2022	2021	test
337016	58883	277933	200

la liste des clients

- *(client_id, sex, birth)*
- *Client_id: 8623 (Count), Birth: 1929-2004 (75 entrée birth)*
- *2 types : 8621 'c_' et 2 'ct_'*

la liste des produits

- *(id_prod, price, categ)*
- *Product_id: 3287*
- *1455 entrées 'price'*
- Categ: 0,1,2

Mission n° 1: Nettoyage des données

a. Description des données

df_transactions

	id_prod	date	session_id	client_id
0	0_1483	4/10/2021	s_18746	c_4450
1	2_226	2/3/2022	s_159142	c_277
2	1_374	9/23/2021	s_94290	c_4270
3	0_2186	10/17/2021	s_105936	c_4597
4	0_1351	7/17/2021	s_63642	c_1242

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 337016 entries, 0 to 337015
Data columns (total 4 columns):
id_prod       337016 non-null object
date          337016 non-null object
session_id    337016 non-null object
client_id     337016 non-null object
dtypes: object(4)
memory usage: 10.3+ MB
```

df_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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8623 entries, 0 to 8622
Data columns (total 3 columns):
client_id    8623 non-null object
sex          8623 non-null category
birth        8623 non-null int64
dtypes: category(1), int64(1), object(1)
memory usage: 143.3+ KB
```

df_produits

	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3287 entries, 0 to 3286
Data columns (total 3 columns):
id_prod    3287 non-null object
price      3287 non-null float64
categ      3287 non-null category
dtypes: category(1), float64(1), object(1)
memory usage: 54.7+ KB
```

Cont.

b. Ajouter la colonne " Age " dans les données

Maintenant, dans les données, nous devons insérer une colonne "âge" dans le tableau des colonnes pour mieux analyser les données.

```
date = dt.datetime.now()
df_customers['age']= date.year-df_customers.birth.astype('int')
```

	client_id	sex	birth	age
0	c_4410	f	1967	54
1	c_7839	f	1975	46
2	c_1699	f	1984	37
3	c_5961	f	1962	59
4	c_5320	m	1943	78

Joindre les 3 tables sur les identifiants du produit et du client

Et créer une nouvelle variable ventes

```
ventes=df_transactions.merge(df_produits,
how='left',on='id_prod').merge(df_customers,how='left',on='client_id')

ventes.head()
```

	id_prod	date	session_id	client_id	price	categ	sex	birth	age
0	0_1483	4/10/2021	s_18746	c_4450	4.99	0	f	1977	44
1	2_226	2/3/2022	s_159142	c_277	65.75	2	f	2000	21
2	1_374	9/23/2021	s_94290	c_4270	10.71	1	f	1979	42
3	0_2186	10/17/2021	s_105936	c_4597	4.20	0	m	1963	58
4	0_1351	7/17/2021	s_63642	c_1242	8.99	0	f	1980	41

Cont.

c. Supprimer " t_0=-1 " des données et identifier les valeurs NaN

Maintenant, on trouve dans les données que :

nous n'avons qu'une seule valeur de t_0=-1 dans id_produits, donc nous la supprimons

ventes_1=ventes.drop(ventes[ventes.id_prod=='T_0'].index)

Maintenant le compte est 33681

For better analysis sortPour une meilleure analyse, sort les données :

the data :

ventes_1.sort_values(by = 'price', ascending = False).tail()

```
id_prod        0
date           0
session_id     0
client_id      0
price        103
categ        103
sex            0
birth          0
age            0
dtype: int64
```

Trouvé une valeur NaN dans les données avec product_id "0_2245"

	id_prod	date	session_id	client_id	price	categ	sex	birth	age
322710	0_2245	4/6/2021	s_16936	c_4167	NaN	NaN	f	1979	42
329417	0_2245	3/30/2021	s_13738	c_7790	NaN	NaN	f	1983	38
330490	0_2245	12/3/2021	s_128815	c_6189	NaN	NaN	f	1984	37
335531	0_2245	4/27/2021	s_26624	c_1595	NaN	NaN	f	1973	48
336220	0_2245	5/1/2021	s_28235	c_5714	NaN	NaN	f	1972	49

Cont.

d. Remplir les valeurs NaN avec les valeurs médian de la catégorie 0

- Le produit 0_2245 n'a pas de prix et pas de catégorie associés.
- Ce produit a été vendu 103 fois dans l'année.
- La catégorie 0 au produit et le prix médian des produits vendus entre 80 et 120 fois dans l'année

```
size=ventes_1.groupby('id_prod').size()
produit=size[(size>80)&(size<120)].index.values
prix=df_produits[df_produits.id_prod.isin(produit)]['price'].median()
ventes_1.loc[ventes_1.id_prod=='0_2245','categ']='0'
ventes_1.loc[ventes_1.id_prod=='0_2245','price']=prix
```

	id_prod	date	session_id	client_id	price	categ	sex	birth	age
6235	0_2245	6/17/2021	s_49705	c_1533	12.99	0	m	1972	49
10802	0_2245	6/16/2021	s_49323	c_7954	12.99	0	m	1973	48
14051	0_2245	11/24/2021	s_124474	c_5120	12.99	0	f	1975	46
17486	0_2245	2/28/2022	s_172304	c_4964	12.99	0	f	1982	39
21078	0_2245	3/1/2021	s_3	c_580	12.99	0	m	1988	33

Mission n° 2: l'analyse des données

Pour l'analyse univariée

a. Les caractéristiques des ventes (min, moyenne, max, écart-type)

ventes_1.describe()

Count: *Nombre d'observations dans la variable*

Mean: *La valeur moyenne de chaque variable*

Std: *Mesure de la quantité de variation ou de dispersion de chaque variable*

Min: *Valeur minimale dans l'ensemble de données*

25% : *Premier quartile : le nombre intermédiaire entre le plus petit nombre (et non le "minimum") et la médiane de l'ensemble de données.*

50% : *Médiane la valeur moyenne de l'ensemble de données*

75% *Troisième quartile : la valeur intermédiaire entre la médiane et la valeur la plus élevée (pas le "maximum") de l'ensemble de données.*

Max: *Valeur maximale de la variable*

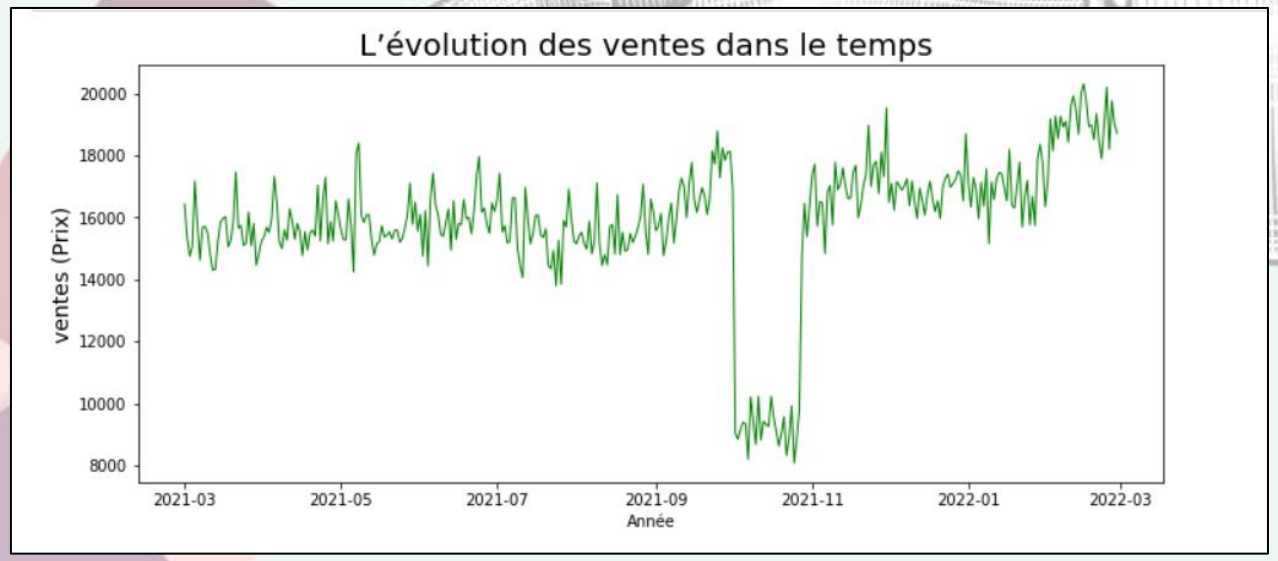
	categ	price	age
count	336816.000000	336816.000000	336816.000000
mean	0.430024	17.213896	43.176604
std	0.591039	17.852868	13.523923
min	0.000000	0.620000	17.000000
25%	0.000000	8.610000	34.000000
50%	0.000000	13.900000	41.000000
75%	1.000000	18.990000	50.000000
max	2.000000	300.000000	92.000000

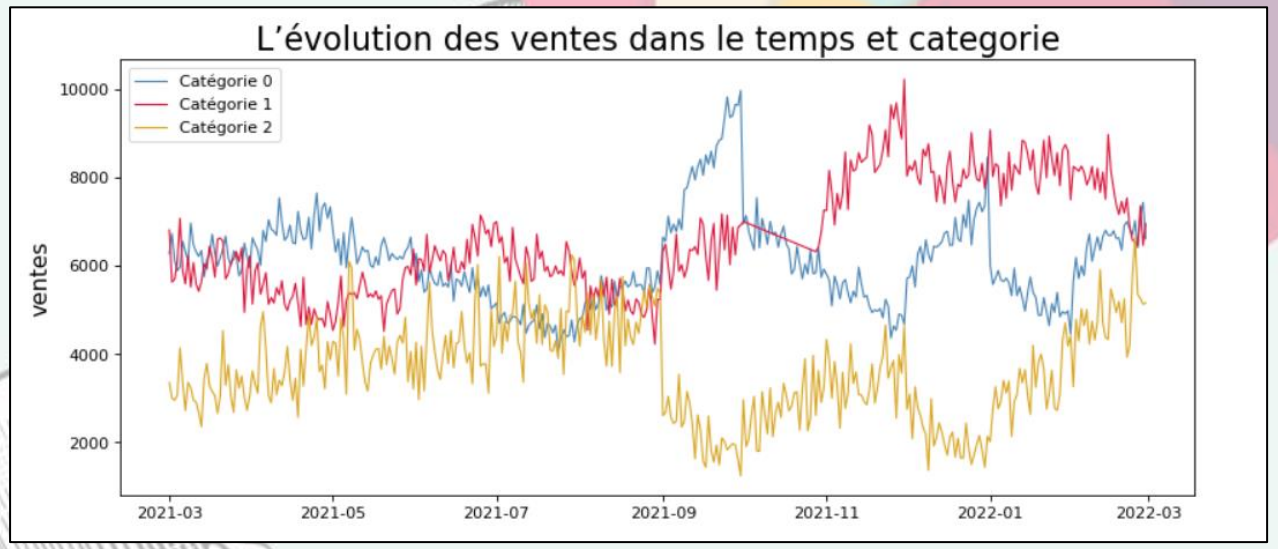
Cont.

b. L’évolution des ventes dans le temps

Observons d'abord l'évolution du ventes de mars 2021 à février 2022.

On peut constater une tendance à la hausse de cette évolution tout au long de l'année, avec interruption entre septembre et novembre avec une forte baisse des ventes



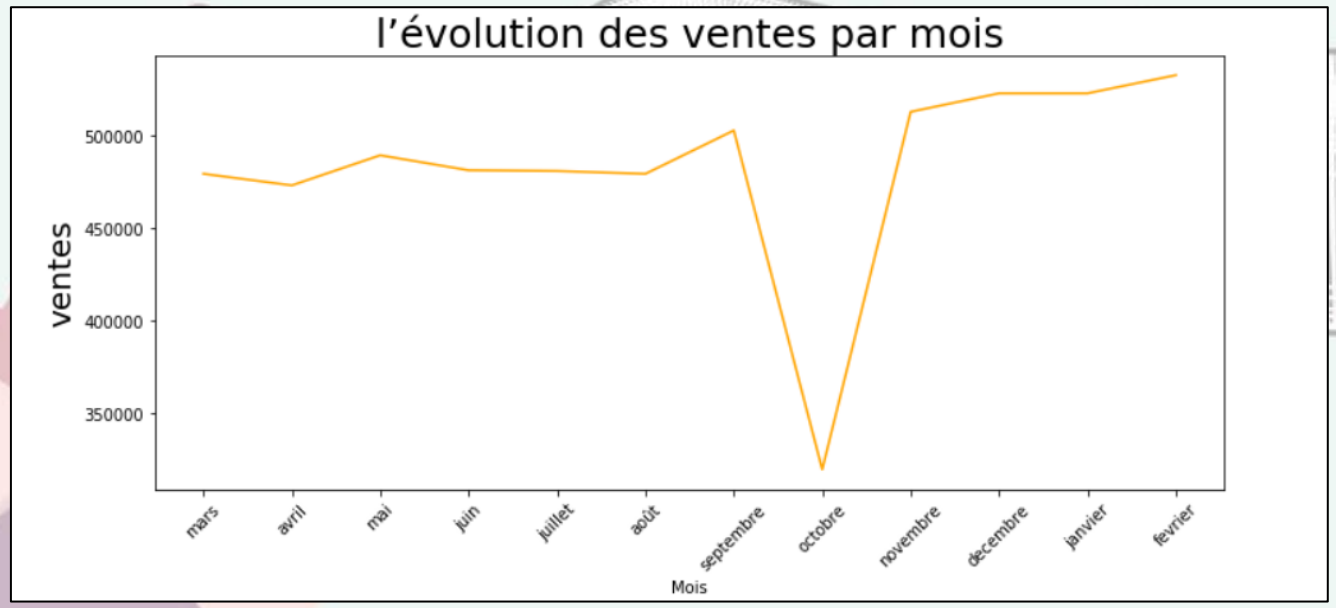


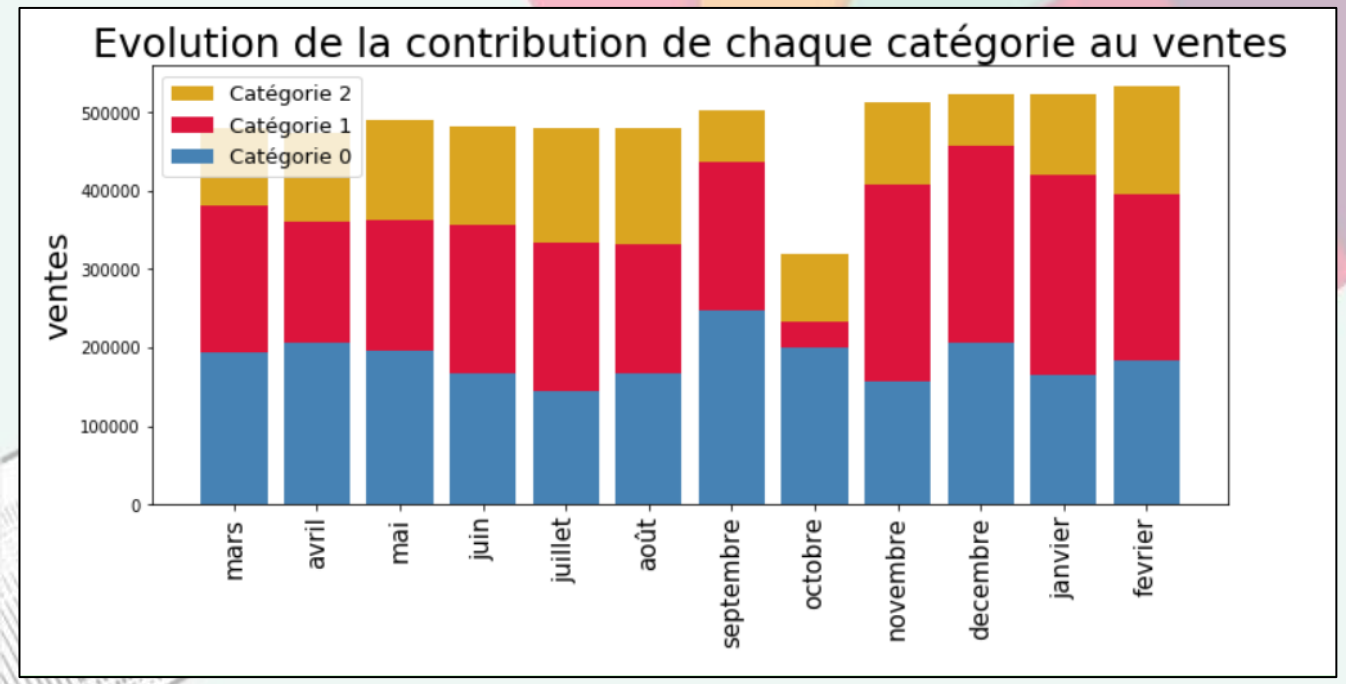
Ici, observe ventes plus faible pour la catégorie 2.
Mais la chute des ventes pour les trois catégories de livres autour de novembre.

Cont.

c. L'évolution des ventes par mois

C'est là que l'on peut constater la chute drastique du chiffre d'affaires en octobre.



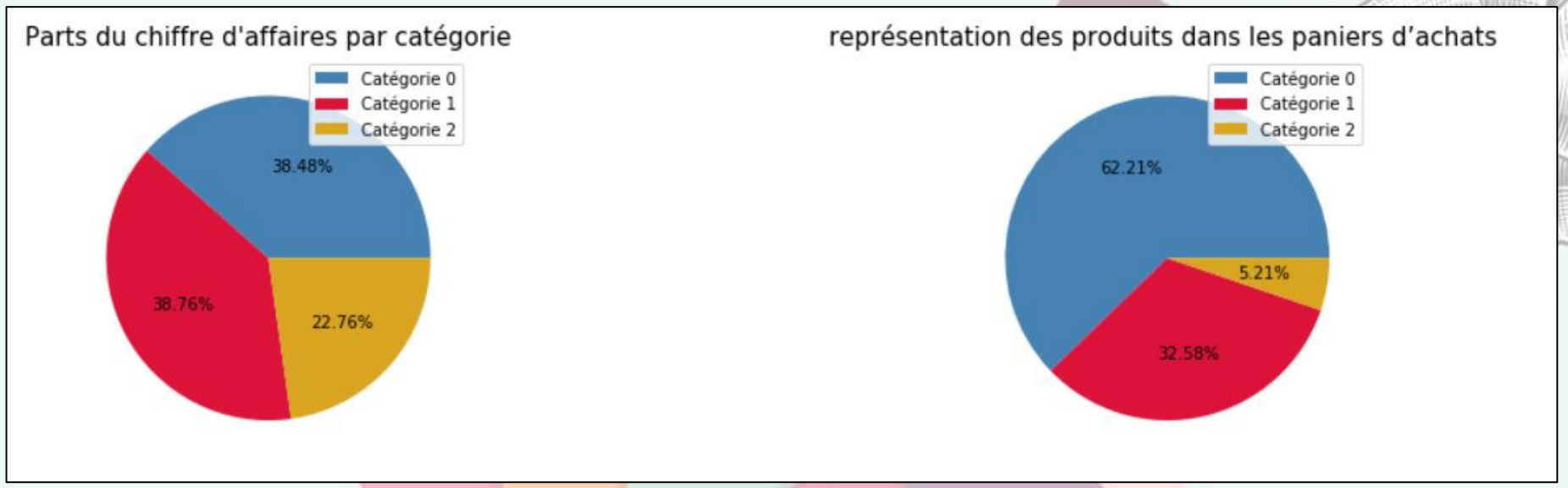


d. L'évolution de la contribution et la contribution de chaque catégorie aux ventes

Tout d'abord, il serait intéressant d'analyser cet histogramme selon les 3 catégories de livres.
La baisse du CA au mois d'octobre est due à la baisse du CA du la catégorie 1

Cont.

e. La contribution de chaque catégorie aux ventes

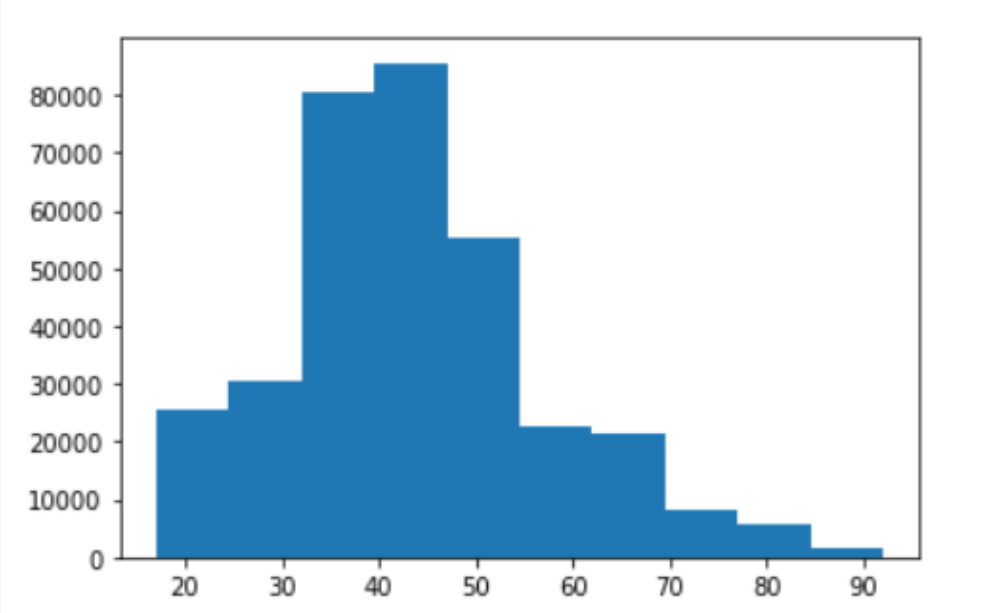


Comparaison des deux diagrammes circulaires :

- ***catégorie 0*** *se vendent bien et représentent le plus grand nombre de ventes (62,21%).*
- ***catégorie 1*** *sont les plus vendus (plus forte contribution au Chiffre d'affaires), malgré un nombre de ventes largement inférieure à celui de la catégorie 0.*
- ***catégorie 2*** *moins solliciter (5,21% des paniers) et contribuent le moins aux ventes.*
- *la comparaison entre le taux de vente et la participation aux ventes suggère que les livres de catégorie 2 sont globalement les livres les plus chers de notre inventaire.*

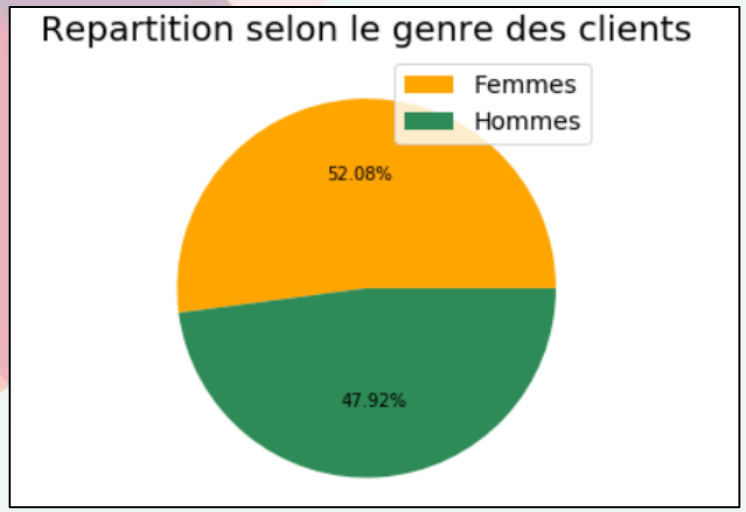
Cont.

f. La distribution des âges des clients



- *Nous pouvons constater que l'âge médian des clients, hommes et femmes, se situe entre 40 et 45 ans.*

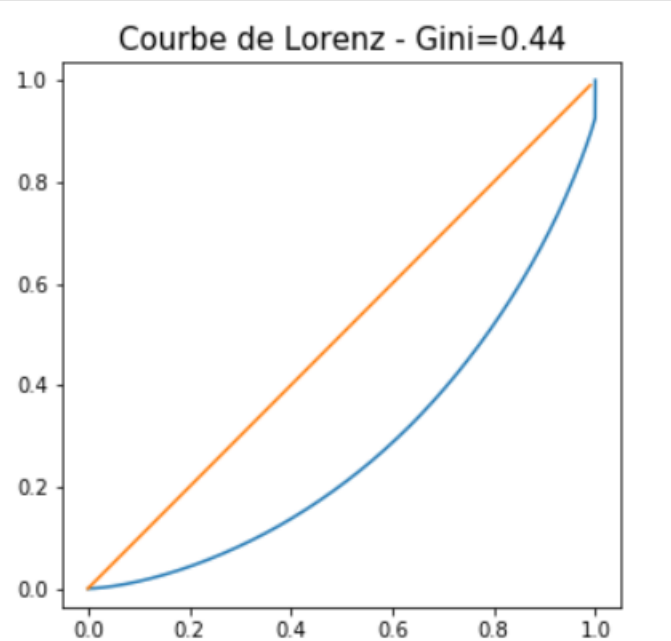
g. La répartition des clients par sexe



- *La proportion des clients âgés de 18 ans est la plus élevée et pourrait s'expliquer par une contrainte d'accès au site.*
- *Ce diagramme circulaire montre clairement que les clients sont majoritairement constituées de femmes (52,08%).*

Cont.

h. Analyse de la distribution du nombre d’achats par client à travers la courbe de Lorenz



On observe alors sur la courbe de Lorenz, avec un indice de Gini proche de 0,44 que distribution du nombre d’achats par client n'est pas égalitaire.

Ainsi, les achats ne sont pas distribués de manière égale entre les clients ; les gros achats effectués par quelques clients ne sont pas distribués à tous les clients.
Si les achats effectués par chaque client et dépensent chaque montant d'argent.
Alors la valeur de Gini serait égale à zéro.

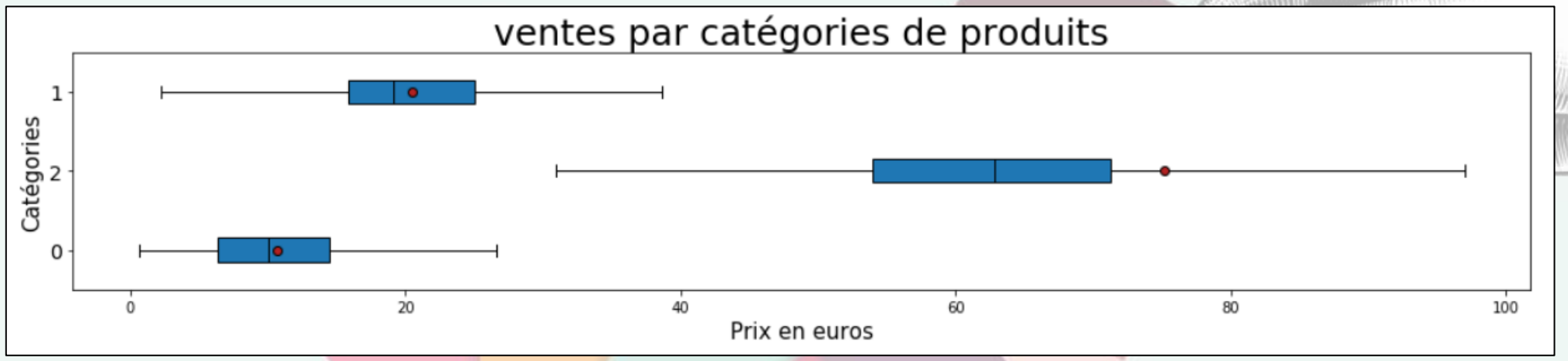
Le coefficient de Gini est une mesure de l'écart par rapport à l'égalité parfaite. Plus une courbe de Lorenz s'écarte de la ligne droite parfaitement égale (qui représente un coefficient de Gini de 0), plus le coefficient de Gini est élevé et moins la société est égale.

Mission n° 2: l’analyse des données

Analyse descriptive bivariée

Cont.

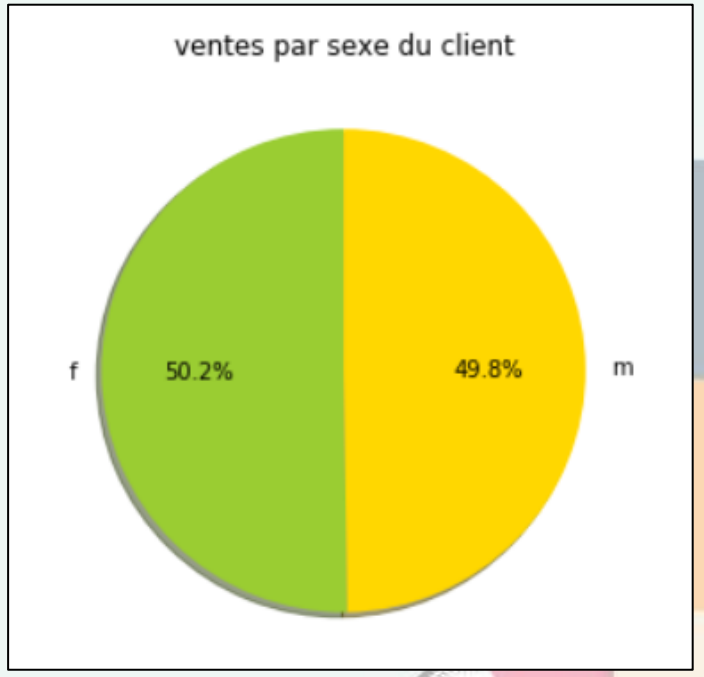
a. Analyse des ventes par catégories de produits



- ***CATÉGORIE 0 :*** *Les prix varient de 0 à 30 euros. En moyenne, ces livres coûtent environ 10 euros et la plupart d'entre eux se vendent à bas prix euros. D'après le graphique ci-dessus, il s'agit de la catégorie "la moins chère".*
- ***CATÉGORIE 1*** *: Les prix vont de plus de 0 à 40 euros. Cette catégorie semble convenir à tous les portefeuilles.*
- ***CATÉGORIE 2 :*** *Les prix vont de 30 euros à 100 euros, c'est la catégorie "la plus chère". C'est aussi la catégorie de livres qui offre le plus grand choix en termes de prix.*

Cont.

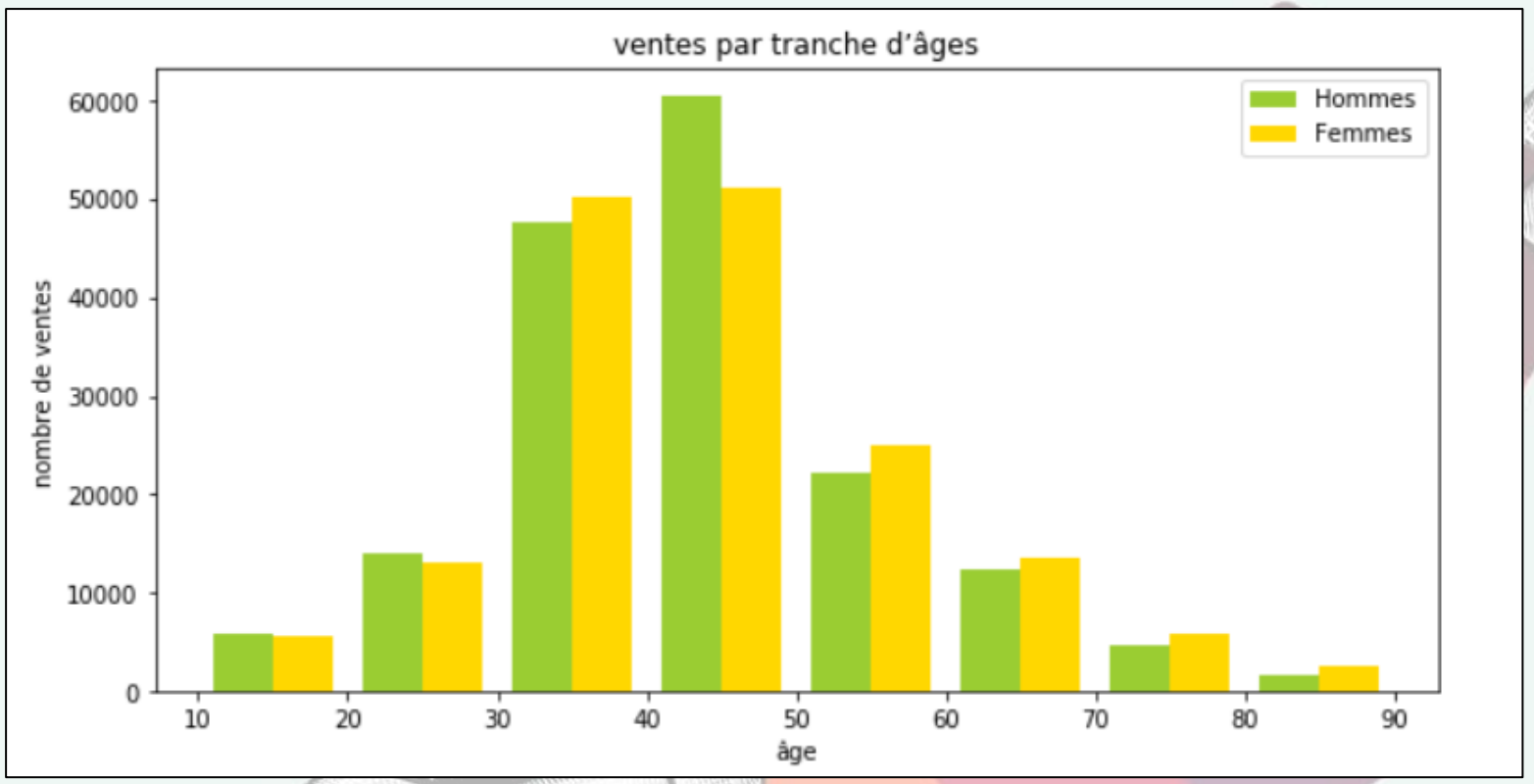
b. Analyse des ventes par sexe du client



- *Dans cette analyse des ventes par sexe du client, on peut observer que la contribution des femmes (50,2%) est un peu plus élevée que celle des hommes (49,8%).*
- *Donc les ventes des femmes sont plus nombreuses que celles des hommes*

Cont.

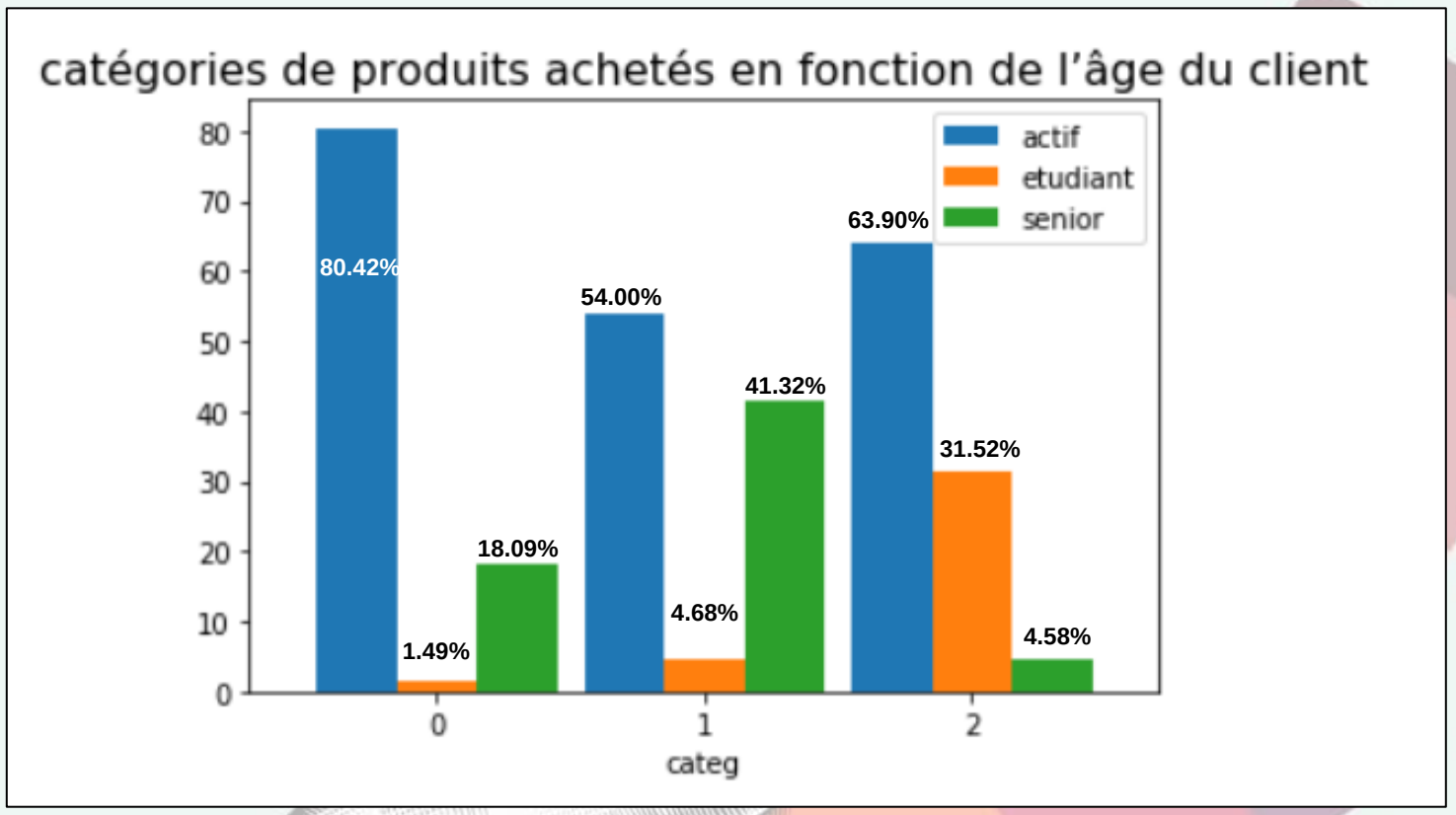
c. Analyse des ventes par tranche d'âges



- *Dans cette analyse des ventes, nous pouvons observer que l'âge inférieur à 20 ans montre très peu de ventes car le site est réservé aux 18 ans et plus.*
- *Les ventes augmentent donc avec l'âge de 20 ans et sont très élevées dans la tranche d'âge 30-50 ans.*
- *Ensuite, lorsque l'âge augmente, les ventes diminuent.*
- *Nous concluons donc que les jeunes de 30 à 50 ans consacrent beaucoup de temps aux ventes et que leurs habitudes d'achat sont également très fréquentes.*
- *De plus, à mesure que l'âge augmente, les personnes âgées sont moins intéressées par les achats et dépensent moins d'argent sur le site.*

Cont.

d. Analyse des catégories de produits achetés en fonction de l'âge du client

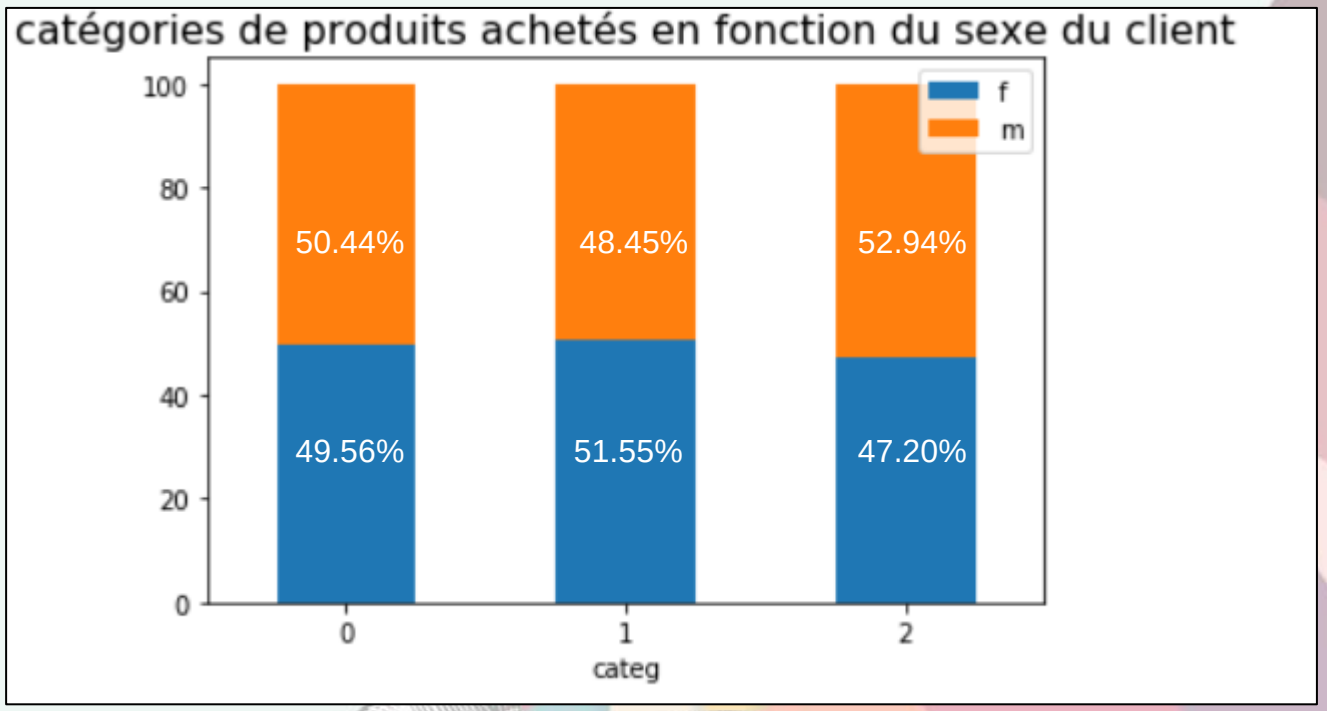


Age <=20 ("etudiant")
Age <=50 ("actif")
Age > 50 ("senior")

- *Nous pouvons clairement voir avec ce graphique que les clients actifs ont acheté plus de produits dans chaque catégorie.*
- *Ensuite, les étudiants ont également acheté un bon pourcentage de produits de la catégorie 2 qui sont chers par rapport aux deux autres produits.*
- *Ainsi, il est clair que les étudiants ne font pas attention au coût, ils achètent s'ils en ont besoin. De plus, un bon pourcentage de livres de catégorie 1 est également acheté par des clients seniors.*
- *Les seniors ne dépensent donc pas d'argent pour des livres chers, ils optent pour des produits de moyenne gamme*

Cont.

e. Analyse des catégories de produits achetés en fonction du sexe du client



- ***CATÉGORIE 0 :*** *Les achats de produits par les femmes (49,56%) sont un peu moins élevés que ceux des hommes (50,44%).*
- ***CATÉGORIE 1 :*** *Les achats de produits par les femmes (51,55%) sont un peu plus élevés que ceux des hommes (48,45%).*
- ***CATÉGORIE 2 :*** *Les produits achetés par les femmes (47,20%) sont un peu moins que ceux achetés par les hommes (52,94%).*

Mission n° 3: : Corrélation

a. Corrélation entre le sexe du client et l'achat de catégories de produits

Measure du chi-2	0	1	2
sex			
m	2.0	14.0	25.0
f	2.0	14.0	26.0

```
fig, ax = plt.subplots(figsize=(5, 3))
sns.heatmap(chi2, annot=True, vmin=0, vmax=5)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1f2080ac4c8>
```

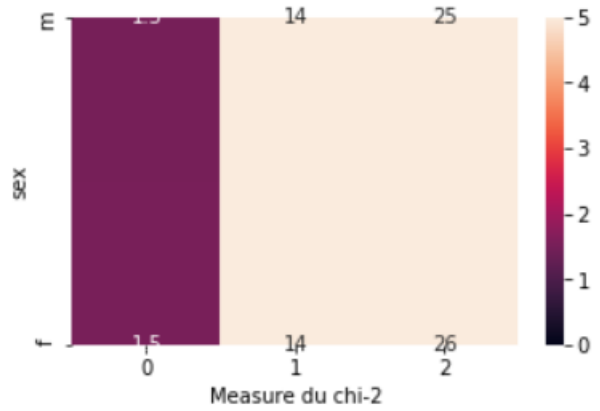


Table	0	1	2
sex			
f	103846.0	55469.0	8260.0
m	105683.0	54266.0	9292.0

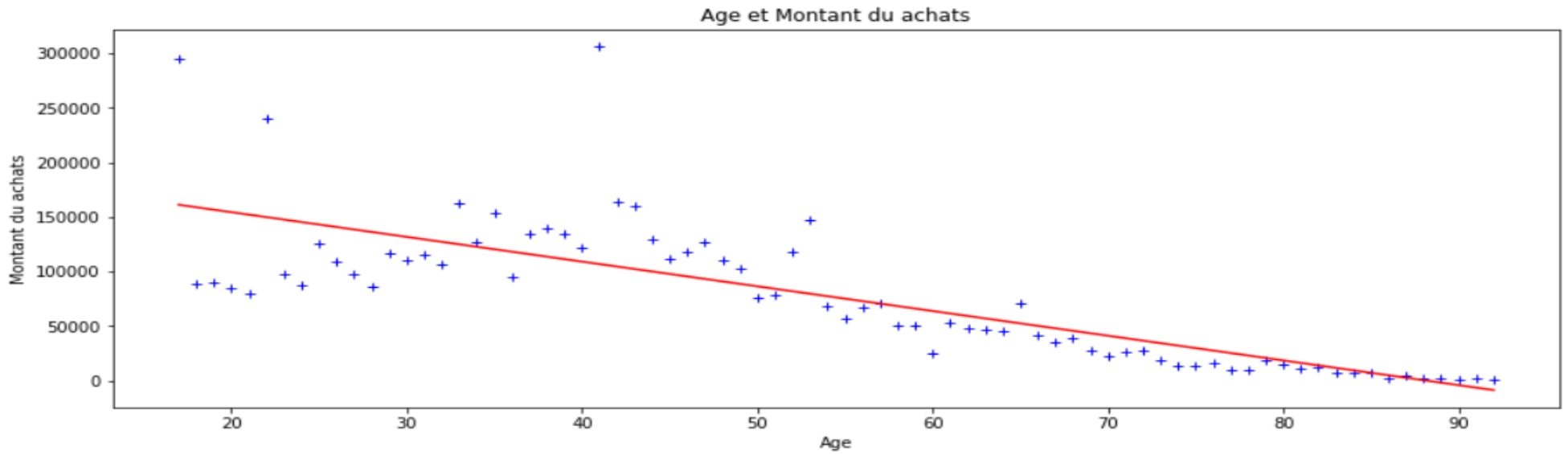
Le sexe et la catégorie du produit sont des données catégorielles.
Pour calculer la corrélation entre deux données catégorielles, nous utilisons le test du chi-2.

- ***Catégorie 0*** *: Les hommes semblent être plus intéressés par cette catégorie de livres que les femmes.*
- ***Catégorie 1 :*** *On constate ici que les femmes sont beaucoup plus intéressées par cette catégorie que les hommes.*
- ***Catégorie 2 :*** *Les hommes sont aussi souvent intéressés par cette catégorie de livres que les femmes.*

Cont.

b. Corrélation entre âge et montant des achats

```
le modèle linéaire est de la forme : y = ax + b
et a = -2265.262050307615
avec b = 199745.1467417662
et un R² de  0.6000541961960478
```

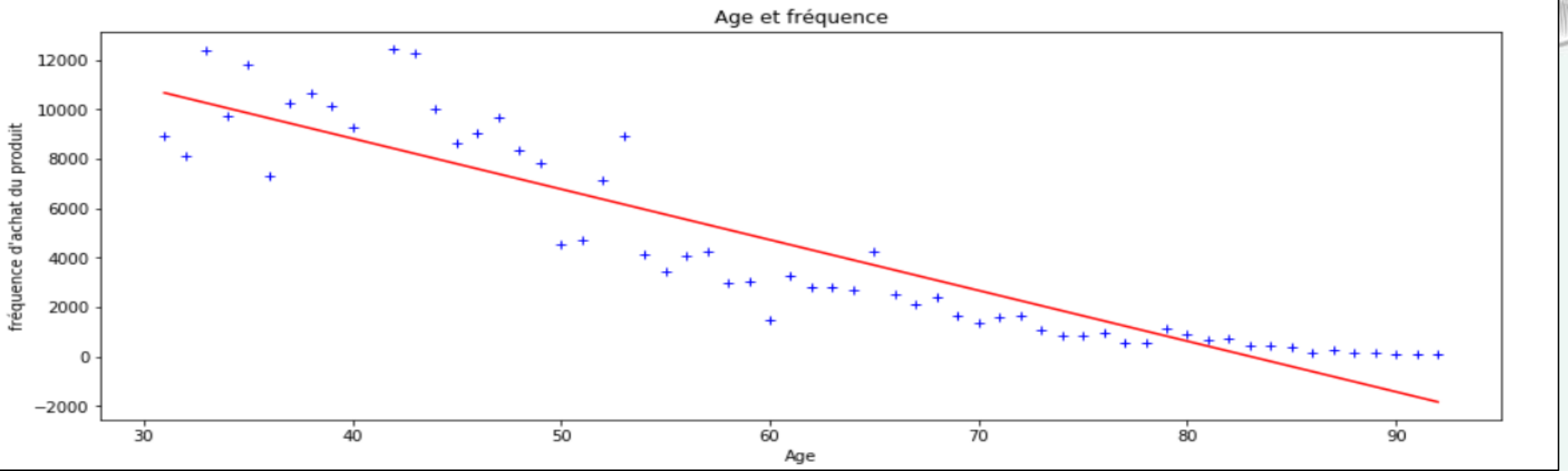


- *La valeur de R-2 est prédite 0,60, ce qui signifie 60 %, ce qui constitue un bon modèle de régression linéaire négative.*
- *Les clients les plus âgés dépensent moins en un an que les jeunes clients.*
- *Nous observons également que les clients âgés de 31 à 50 ans semblent dépenser plus que ceux âgés de moins de 30 ans.*

Cont.

c. Corrélation entre âge et fréquence d'achats

```
le modèle linéaire est de la forme : y = ax + b
et a = -204.90443748787052
avec b = 17023.35308531553
et un R² de  0.8407693764163361
```

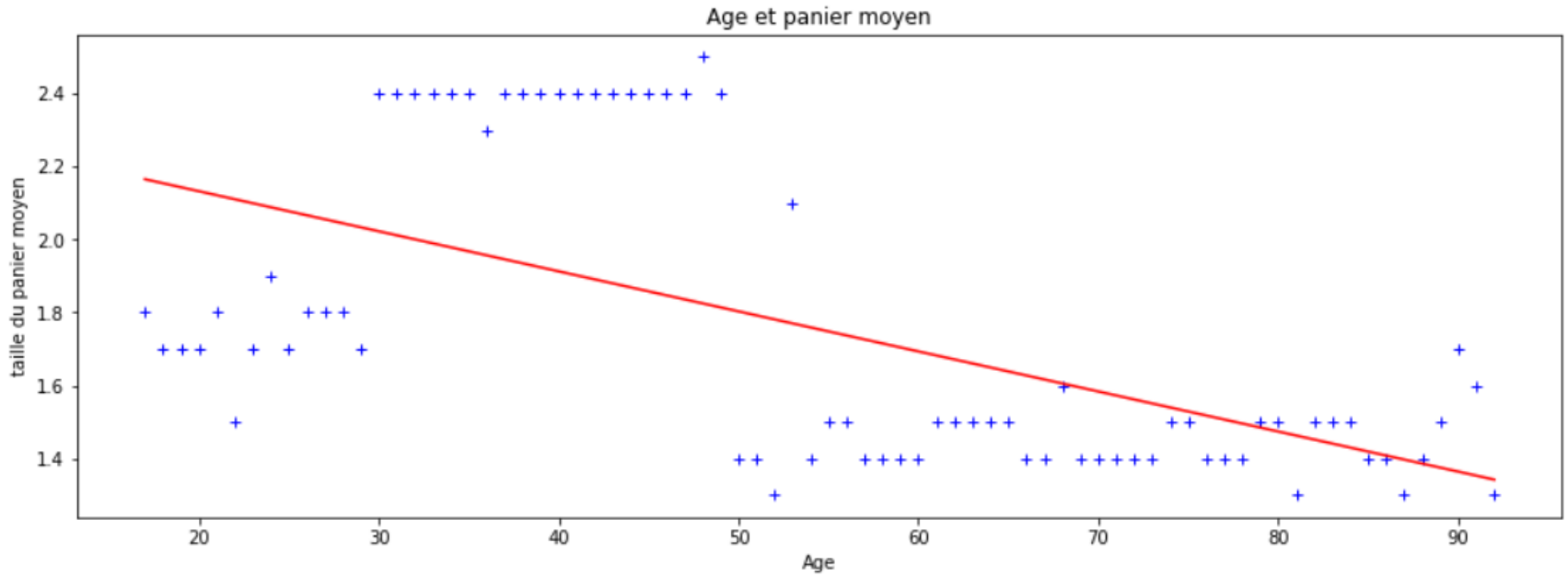


- *La valeur de R-2 est prédite 0,84, ce qui signifie 80 %, ce qui constitue un bon modèle de régression linéaire négative.*
- *La fréquence d'achat des clients plus âgés est moins élevée que celle des jeunes clients.*
- *Les clients âgés utilisent et achètent uniquement des choses spécifiques.*

Cont.

d. Corrélation entre âge et panier moyen

```
le modèle linéaire est de la forme : y = ax + b
et a = -0.010962406015037594
avec b = 2.351398496240601
et un R² de  0.3394423677957023
```

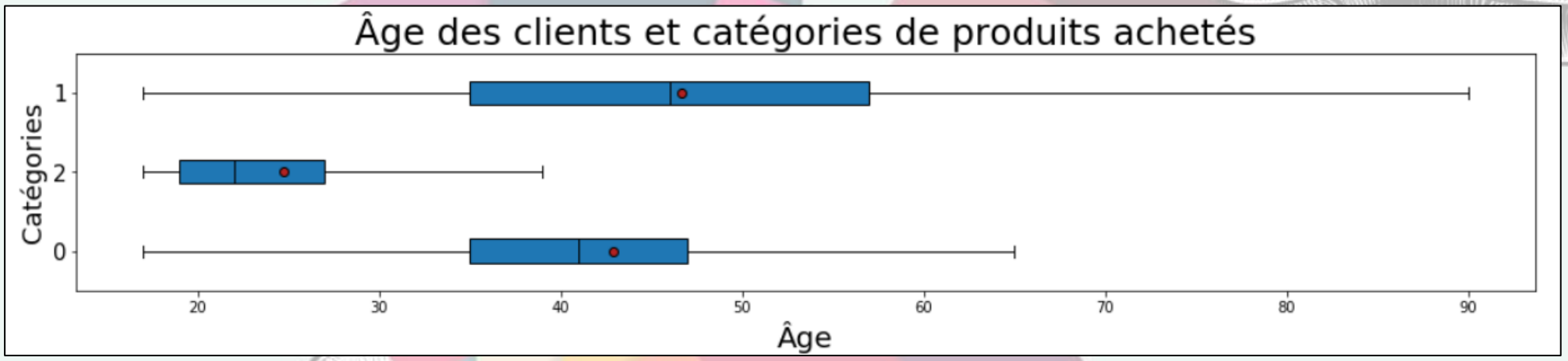


- *La valeur de R-2 est prédite 0,33, ce qui signifie 33 %, ce qui constitue un faible modèle de régression linéaire négative.*
- *Il y a pas une corrélation entre l'âge et le panier moyen.*

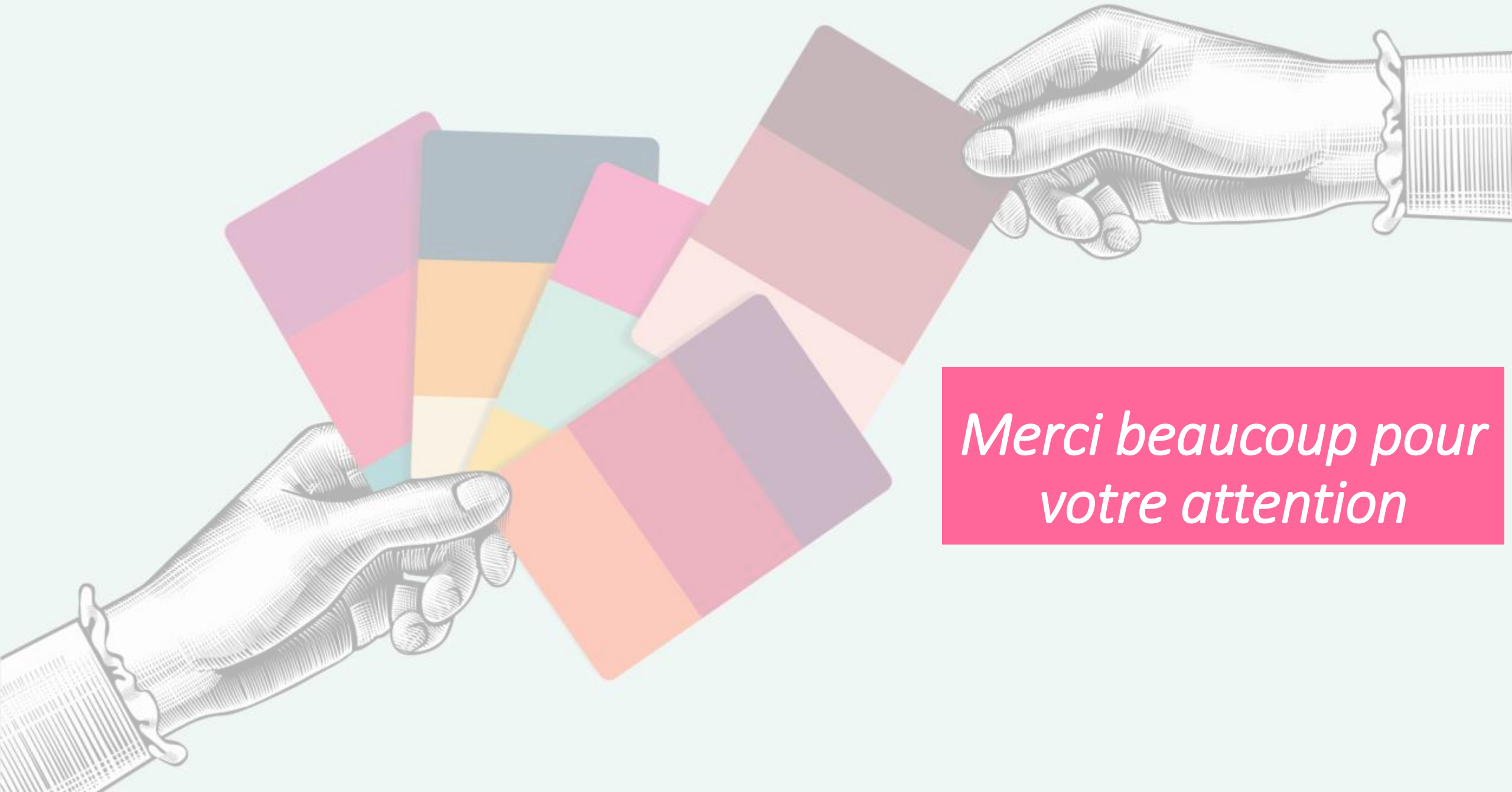
Cont.

e. Corrélation entre âge et catégorie de produit

nous prenons notre Data Frame initial « ventes »" sur lequel nous analysons les corrélations possibles entre les âges des clients et les catégories de produits achetés



- *Rapport de corrélation: 11.89 %*
- *Avec un taux de corrélation d'environ 12%, nous pouvons observer qu'il existe une corrélation entre l'âge des clients et les catégories de produits achetés :*
- *on remarque que la boîte à moustaches de catégorie 2 se distingue des deux autres. La catégorie 2 s'adresse généralement à une clientèle plus jeune, entre 18 et 25 ans, alors que les moyennes d'âge des catégories 0 et 1 semblent plus homogènes, et intéressent apparemment tous les types de clients, quel que soit leur âge.*



Merci beaucoup pour votre attention