# RESTER LIVRES

Analyse des données et des transactions e-commerce du 01/03/2021 au 28/02/2022

*Projet 4*

*Marc Sellam.*

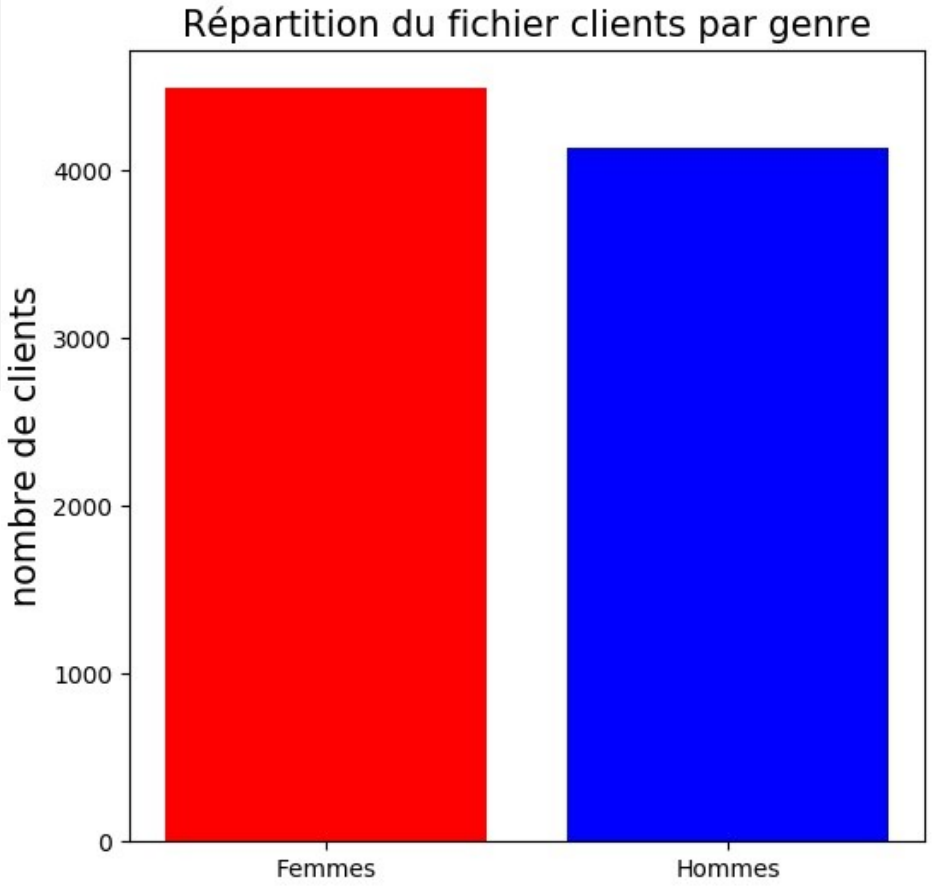
# Sommaire:

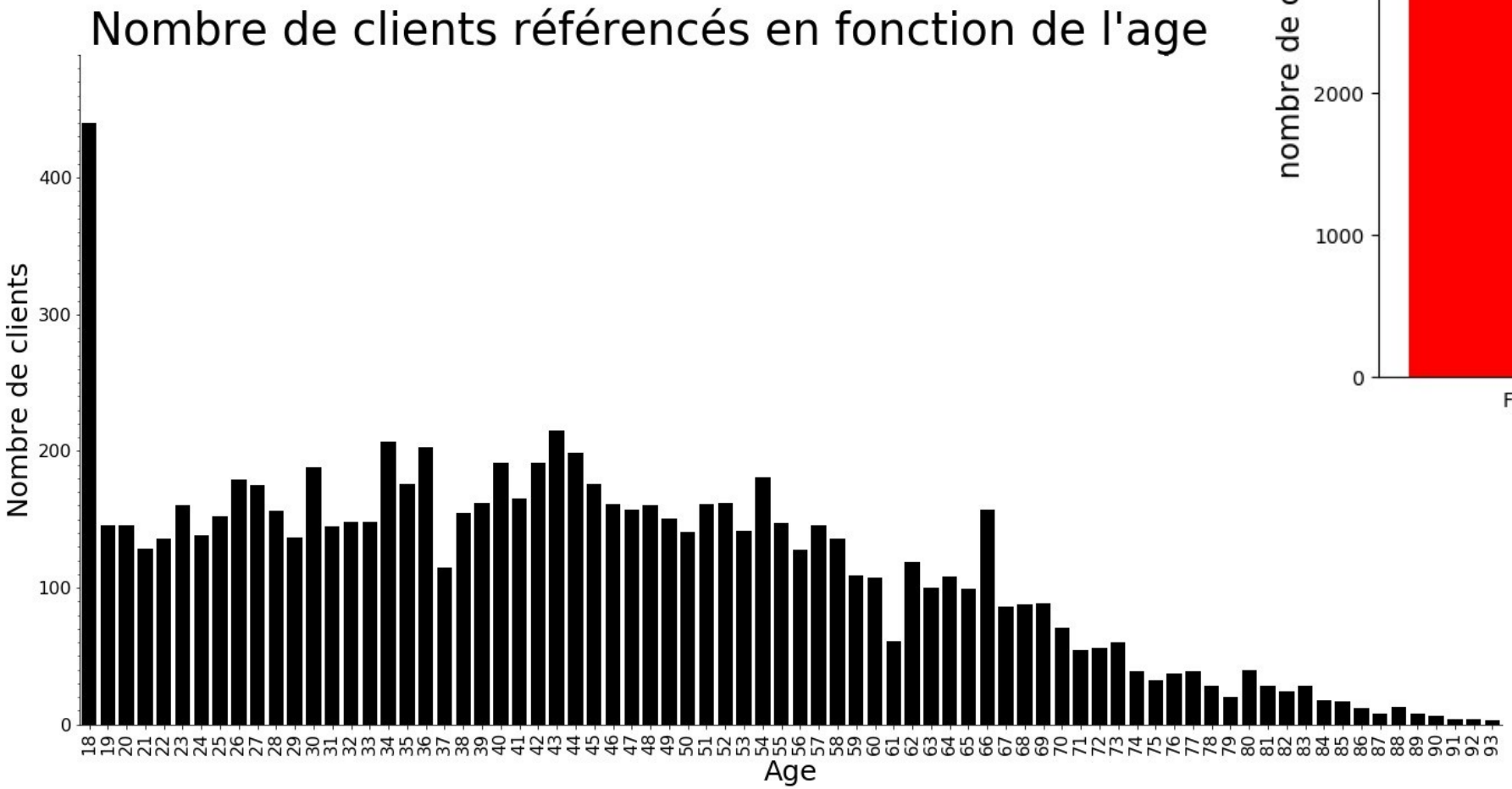


# Les clients

*Le fichier clients ne présente pas d'incohérences.*

*Nombre important de clients âgés de 18 ans.*



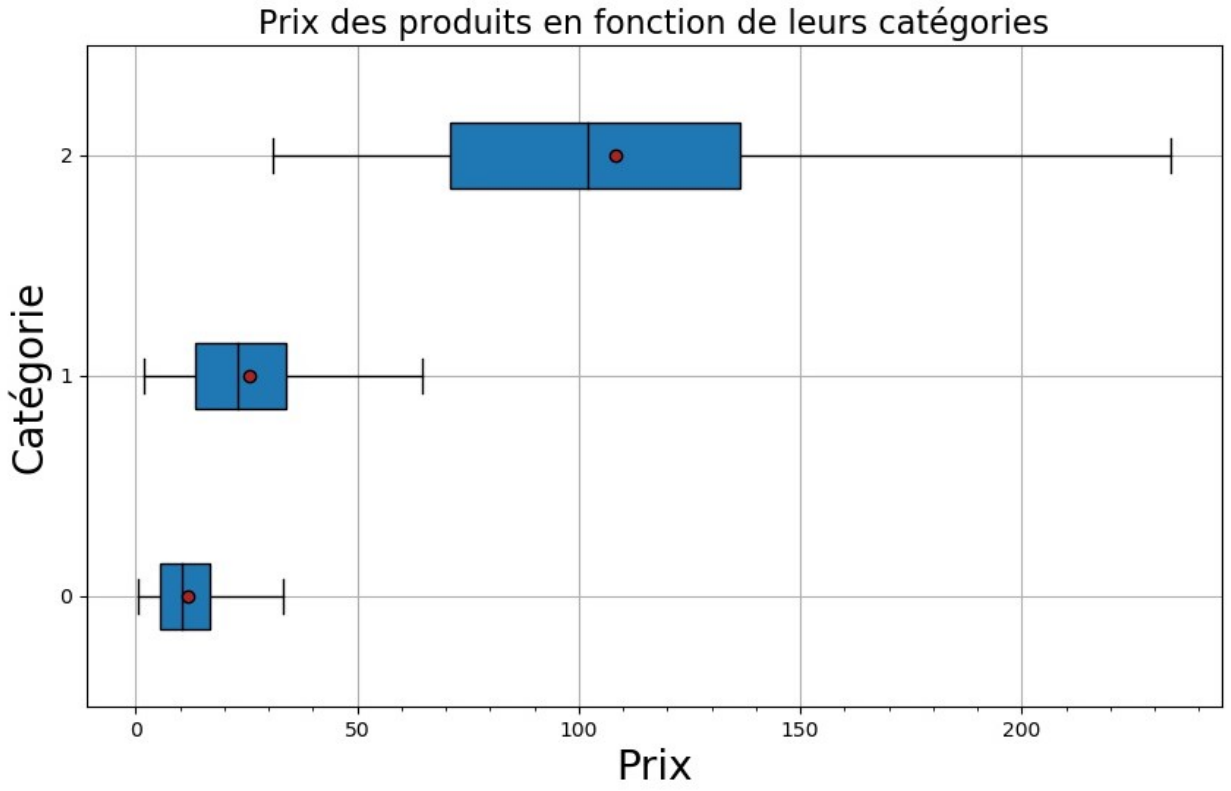



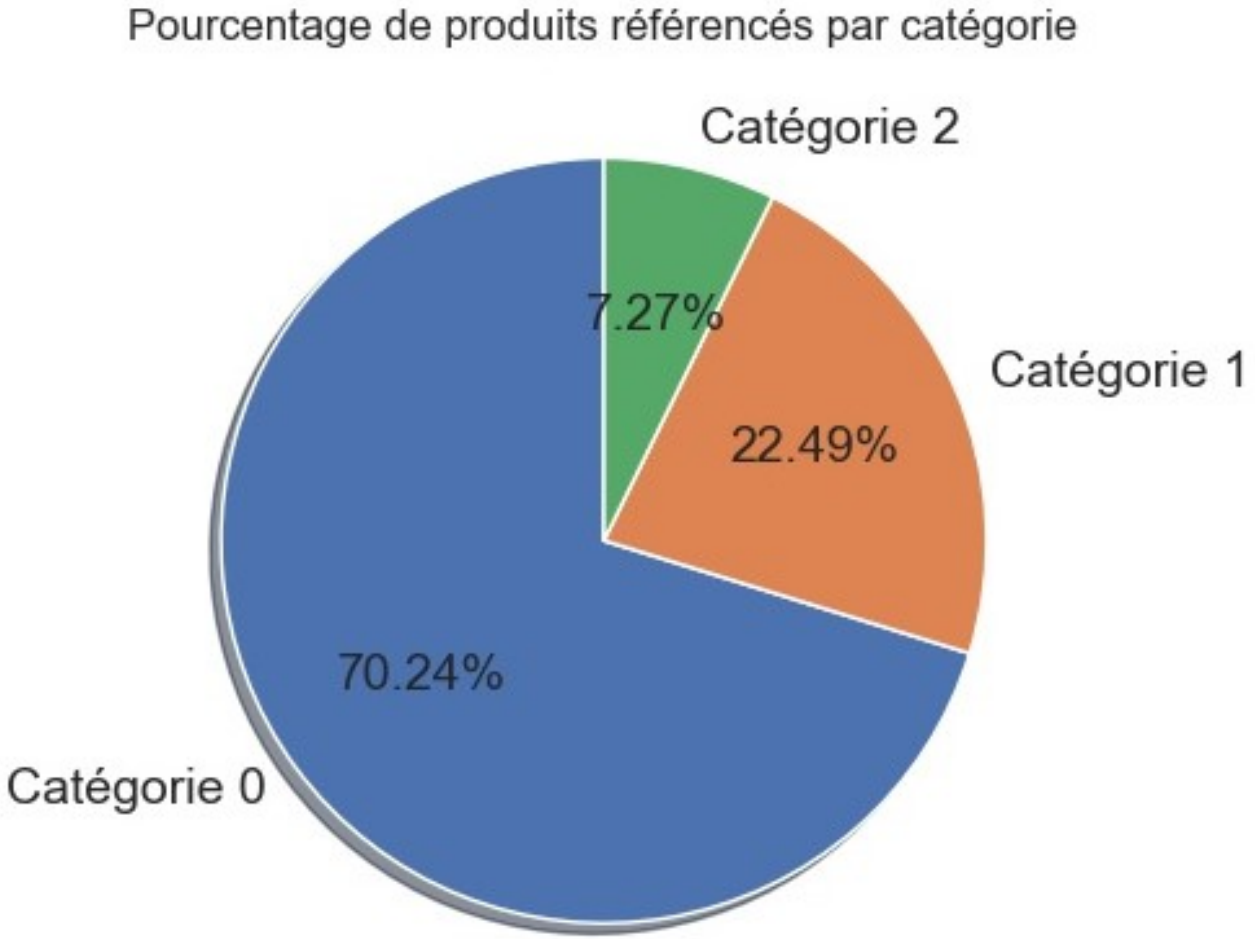
# Les Produits

*Un produit référencé « T_0 » a une valeur négative, nous le retirons de la liste produits.*

| | id_prod | price | categ |
|---|---|---|---|
| 947 | 2_2 | 300.00 | 2 |
| 724 | 2_76 | 254.44 | 2 |
| 394 | 2_158 | 247.22 | 2 |
| 1436 | 2_167 | 236.99 | 2 |
| 2779 | 2_30 | 233.54 | 2 |
| ... | ... | ... | ... |
| 1211 | 0_1844 | 0.77 | 0 |
| 370 | 0_120 | 0.66 | 0 |
| 2272 | 0_528 | 0.62 | 0 |
| 2355 | 0_202 | 0.62 | 0 |
| 731 | T_0 | -1.00 | 0 |

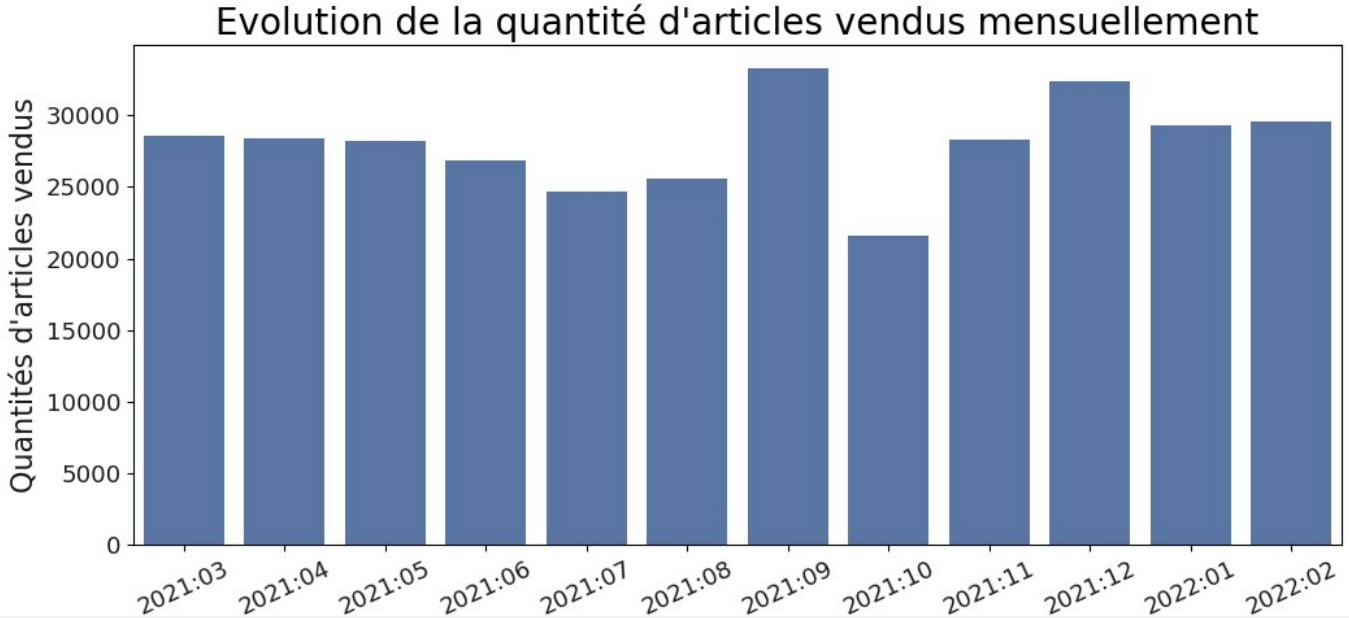
3287 rows × 3 columns



# Les transactions

*200 transactions de test retirées*




| | id_prod | date | session_id | client_id |
|---|---|---|---|---|
| 1431 | T_0 | test_2021-03-01 02:30:02.237420 | s_0 | ct_1 |
| 2365 | T_0 | test_2021-03-01 02:30:02.237446 | s_0 | ct_1 |
| 2895 | T_0 | test_2021-03-01 02:30:02.237414 | s_0 | ct_1 |
| 5955 | T_0 | test_2021-03-01 02:30:02.237441 | s_0 | ct_0 |
| 7283 | T_0 | test_2021-03-01 02:30:02.237434 | s_0 | ct_1 |
| ... | ... | ... | ... | ... |
| 332594 | T_0 | test_2021-03-01 02:30:02.237445 | s_0 | ct_0 |
| 332705 | T_0 | test_2021-03-01 02:30:02.237423 | s_0 | ct_1 |
| 332730 | T_0 | test_2021-03-01 02:30:02.237421 | s_0 | ct_1 |
| 333442 | T_0 | test_2021-03-01 02:30:02.237431 | s_0 | ct_1 |
| 335279 | T_0 | test_2021-03-01 02:30:02.237430 | s_0 | ct_0 |

200 rows × 4 columns

```
la transaction la plus ancienne date du  2021-03-01 00:01:07.843138
la transaction la plus récente date du  2022-02-28 23:59:58.040472
nous avons 336816 transactions distinctes
```

# Rassemblement et nettoyage des données.

*148 transactions retirées:*

*-le produit 0_2245 n'a ni prix ni catégorie.*

*-22 produits référencés ne figurent pas dans les transactions et n'ont donc pas été achetés.*

*-23 client qui n'ont pas acheté cette année ne figurent donc pas dans les transactions.*

| | id_prod | fulldate | session_id | client_id | jour | heure | mois | price | categ | sex | age |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 336713 | NaN | NaN | NaN | c_8253 | NaN | NaN | NaN | NaN | NaN | f | 21 |
| 336714 | NaN | NaN | NaN | c_3789 | NaN | NaN | NaN | NaN | NaN | f | 25 |
| 336715 | NaN | NaN | NaN | c_4406 | NaN | NaN | NaN | NaN | NaN | f | 24 |
| 336716 | NaN | NaN | NaN | ct_0 | NaN | NaN | NaN | NaN | NaN | f | 21 |
| 336717 | NaN | NaN | NaN | c_2706 | NaN | NaN | NaN | NaN | NaN | f | 55 |
| 336718 | NaN | NaN | NaN | c_3443 | NaN | NaN | NaN | NaN | NaN | m | 63 |
| 336719 | NaN | NaN | NaN | c_4447 | NaN | NaN | NaN | NaN | NaN | m | 66 |
| 336720 | NaN | NaN | NaN | c_3017 | NaN | NaN | NaN | NaN | NaN | f | 30 |
| 336721 | NaN | NaN | NaN | c_4086 | NaN | NaN | NaN | NaN | NaN | f | 30 |
| 336722 | NaN | NaN | NaN | c_6930 | NaN | NaN | NaN | NaN | NaN | m | 18 |
| 336723 | NaN | NaN | NaN | c_4358 | NaN | NaN | NaN | NaN | NaN | m | 23 |
| 336724 | NaN | NaN | NaN | c_8381 | NaN | NaN | NaN | NaN | NaN | f | 57 |
| 336725 | NaN | NaN | NaN | c_1223 | NaN | NaN | NaN | NaN | NaN | m | 59 |
| 336726 | NaN | NaN | NaN | c_6862 | NaN | NaN | NaN | NaN | NaN | f | 20 |
| 336727 | NaN | NaN | NaN | c_5245 | NaN | NaN | NaN | NaN | NaN | f | 18 |
| 336728 | NaN | NaN | NaN | c_5223 | NaN | NaN | NaN | NaN | NaN | m | 19 |
| 336729 | NaN | NaN | NaN | c_6735 | NaN | NaN | NaN | NaN | NaN | m | 18 |
| 336730 | NaN | NaN | NaN | c_862 | NaN | NaN | NaN | NaN | NaN | f | 66 |
| 336731 | NaN | NaN | NaN | c_7584 | NaN | NaN | NaN | NaN | NaN | f | 62 |
| 336732 | NaN | NaN | NaN | c_90 | NaN | NaN | NaN | NaN | NaN | m | 21 |
| 336733 | NaN | NaN | NaN | c_587 | NaN | NaN | NaN | NaN | NaN | m | 29 |
| 336734 | NaN | NaN | NaN | ct_1 | NaN | NaN | NaN | NaN | NaN | m | 21 |
| 336735 | NaN | NaN | NaN | c_3526 | NaN | NaN | NaN | NaN | NaN | m | 66 |

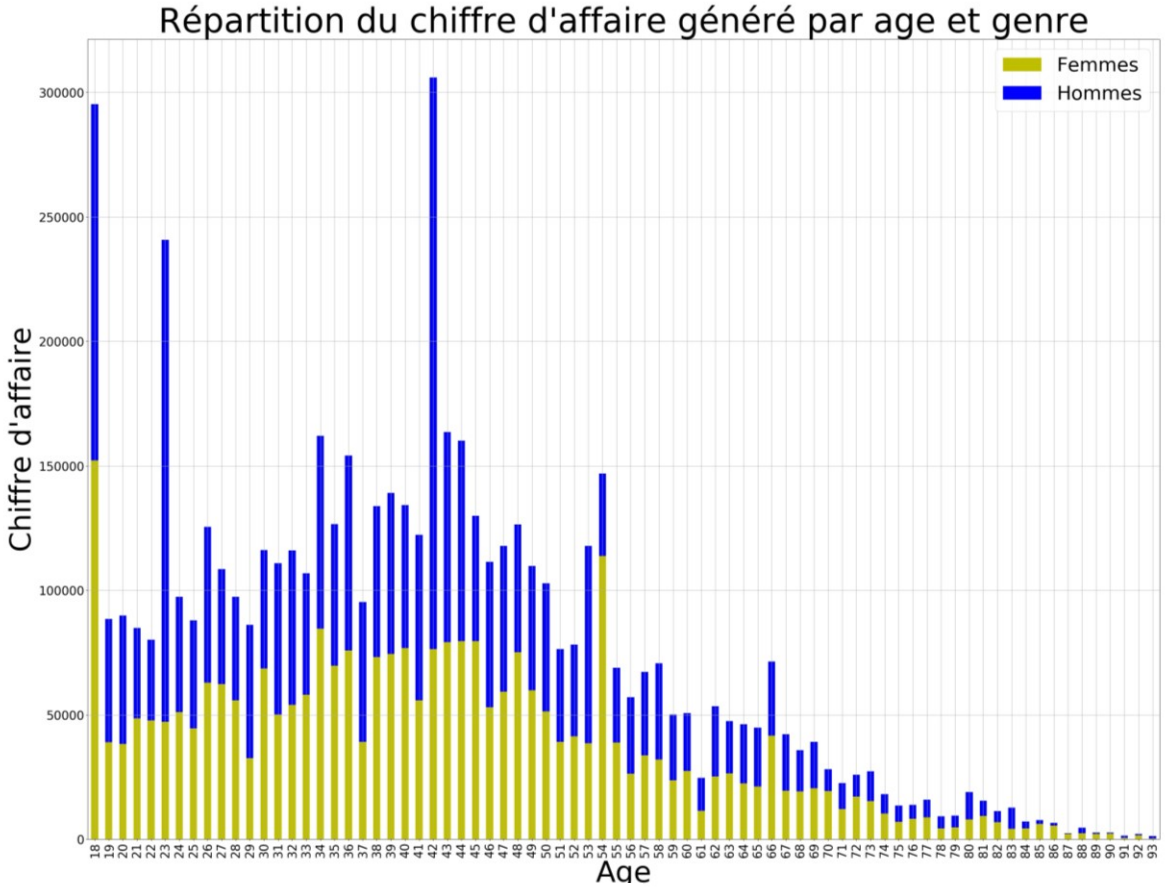
```
data = pd.merge(transactions, products, on=["id_prod"],how="outer", ).copy()
print("nombre de NaN sur DATA:")
print(data.isna().sum())
data[data.price.isna()]
```

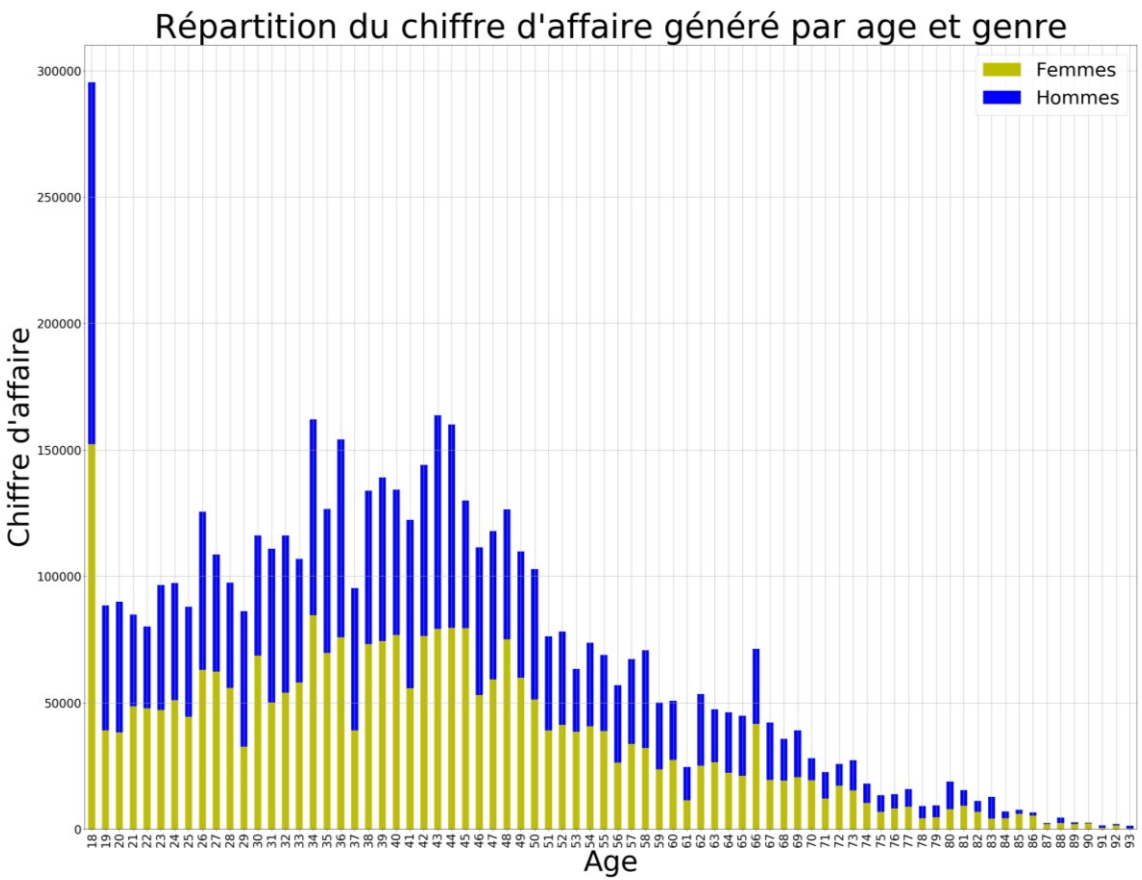
```
nombre de NaN sur DATA:
id_prod         0
fulldate       22
session_id     22
client_id      22
jour           22
heure          22
mois           22
price         103
categ         103
dtype: int64
```

| | id_prod | fulldate | session_id | client_id | jour | heure | mois | price | categ |
|---|---|---|---|---|---|---|---|---|---|
| 55760 | 0_2245 | 2021-03-01 00:09:29.301897 | s_3 | c_580 | 2021:03:01 | 00:09 | 2021:03 | NaN | NaN |
| 55761 | 0_2245 | 2021-03-02 15:55:49.728644 | s_783 | c_8060 | 2021:03:02 | 15:55 | 2021:03 | NaN | NaN |
| 55762 | 0_2245 | 2021-03-07 20:00:01.525768 | s_3159 | c_1189 | 2021:03:07 | 20:00 | 2021:03 | NaN | NaN |
| 55763 | 0_2245 | 2021-03-23 15:57:44.266387 | s_10427 | c_5869 | 2021:03:23 | 15:57 | 2021:03 | NaN | NaN |
| 55764 | 0_2245 | 2021-03-28 10:46:04.804784 | s_12567 | c_966 | 2021:03:28 | 10:46 | 2021:03 | NaN | NaN |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 55858 | 0_2245 | 2022-02-20 03:50:42.646180 | s_167926 | c_7287 | 2022:02:20 | 03:50 | 2022:02 | NaN | NaN |
| 55859 | 0_2245 | 2022-02-24 04:18:45.054650 | s_169978 | c_5953 | 2022:02:24 | 04:18 | 2022:02 | NaN | NaN |
| 55860 | 0_2245 | 2022-02-25 00:08:08.736068 | s_170426 | c_6236 | 2022:02:25 | 00:08 | 2022:02 | NaN | NaN |
| 55861 | 0_2245 | 2022-02-25 06:36:49.557289 | s_170548 | c_2242 | 2022:02:25 | 06:36 | 2022:02 | NaN | NaN |
| 55862 | 0_2245 | 2022-02-28 18:08:49.875709 | s_172304 | c_4964 | 2022:02:28 | 18:08 | 2022:02 | NaN | NaN |

103 rows × 9 columns

# Analyse graphique des données rassemblées




| | age | sex | client_id | price |
|---|---|---|---|---|
| **677** | 42 | m | c_1609 | 162007.34 |
| **4388** | 23 | m | c_4958 | 144257.21 |
| **6337** | 54 | f | c_6714 | 73197.34 |
| **2724** | 53 | m | c_3454 | 54442.92 |

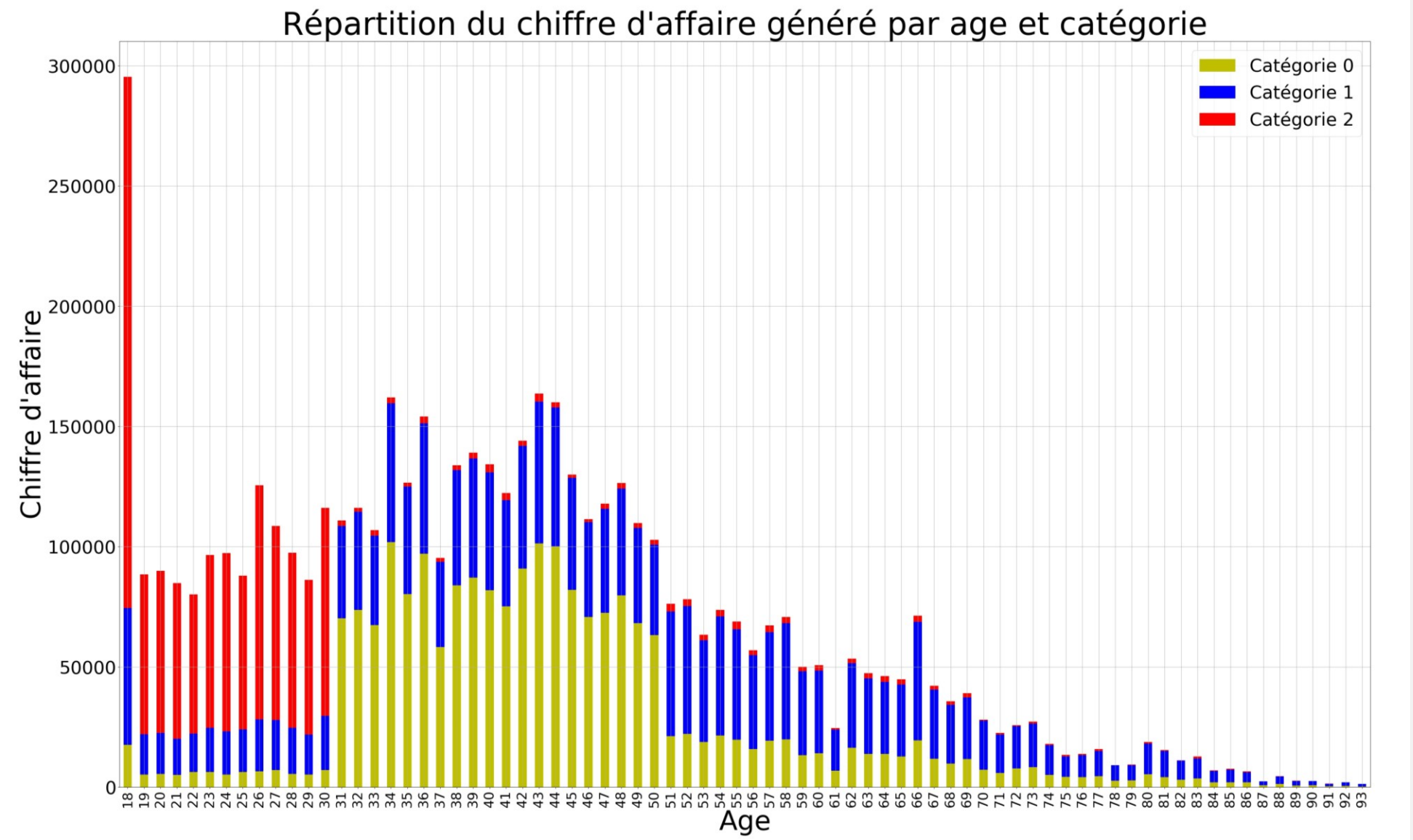
*On remarque un CA en adéquation avec le reste du graphique à 23 ans, 42 ans, 53 ans et 54 ans.*

*Ces 4 clients qui représentent 7.49 % des ventes sont des achats en gros: retirés des données.*

*Pour ce qui est des clients âgés de 18 ans on suppose qu'il s agit de mineurs.*

RESTER LIVRES

# Analyse graphique des données rassemblées

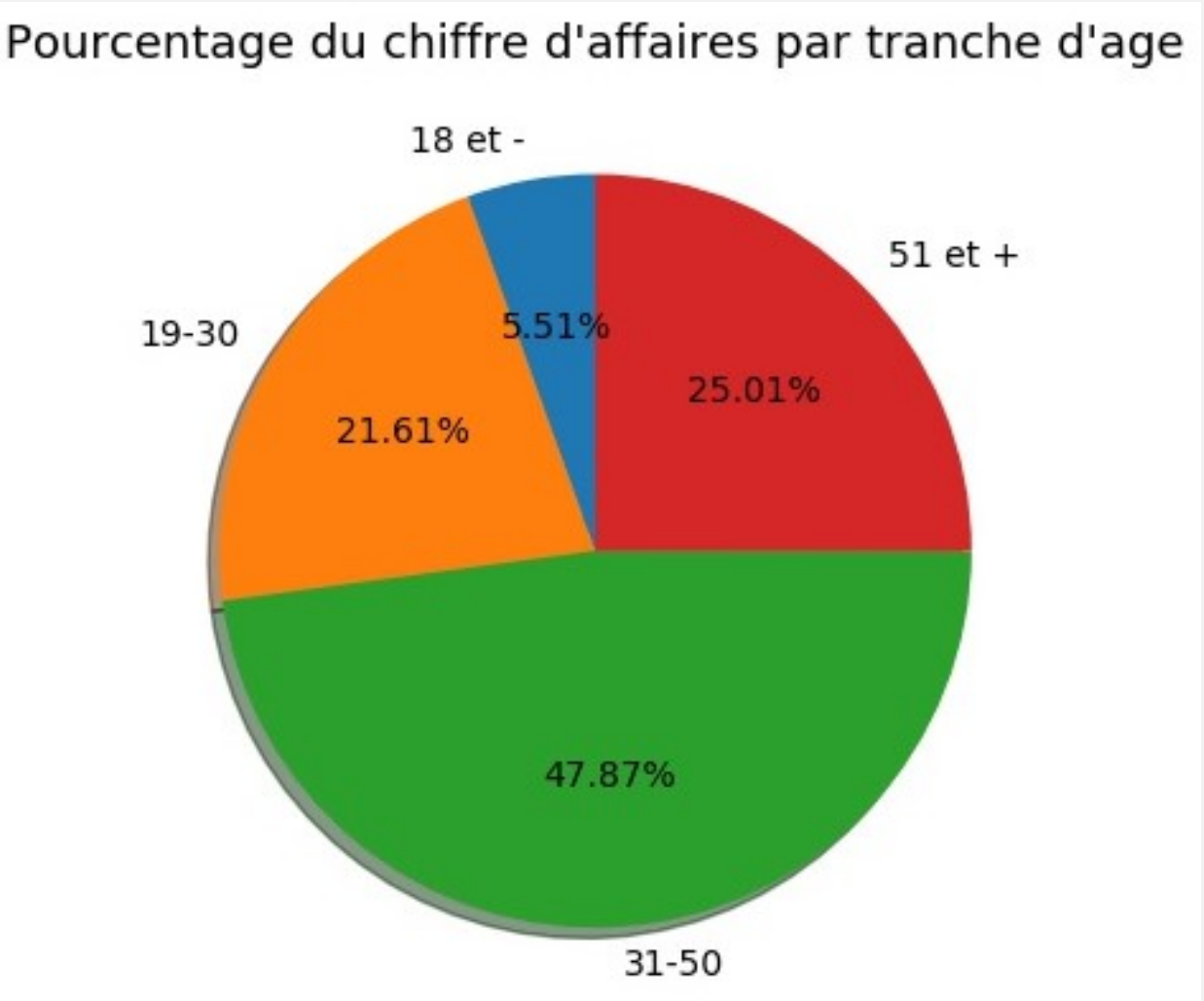


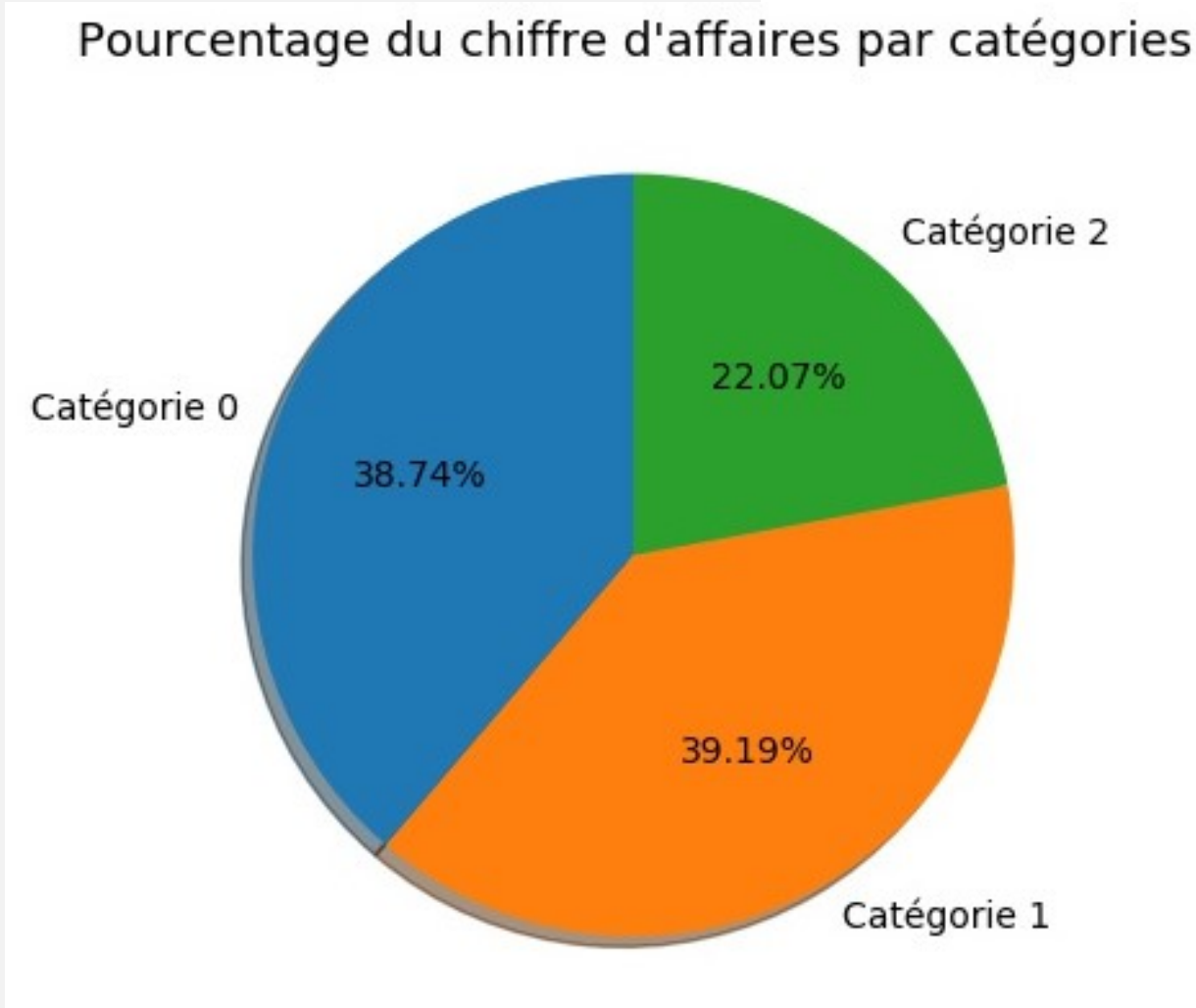

*Une clientèle divisible en 4*

*(18 ans, 19-30 ans, 31-50 ans et 51-93 ans)*

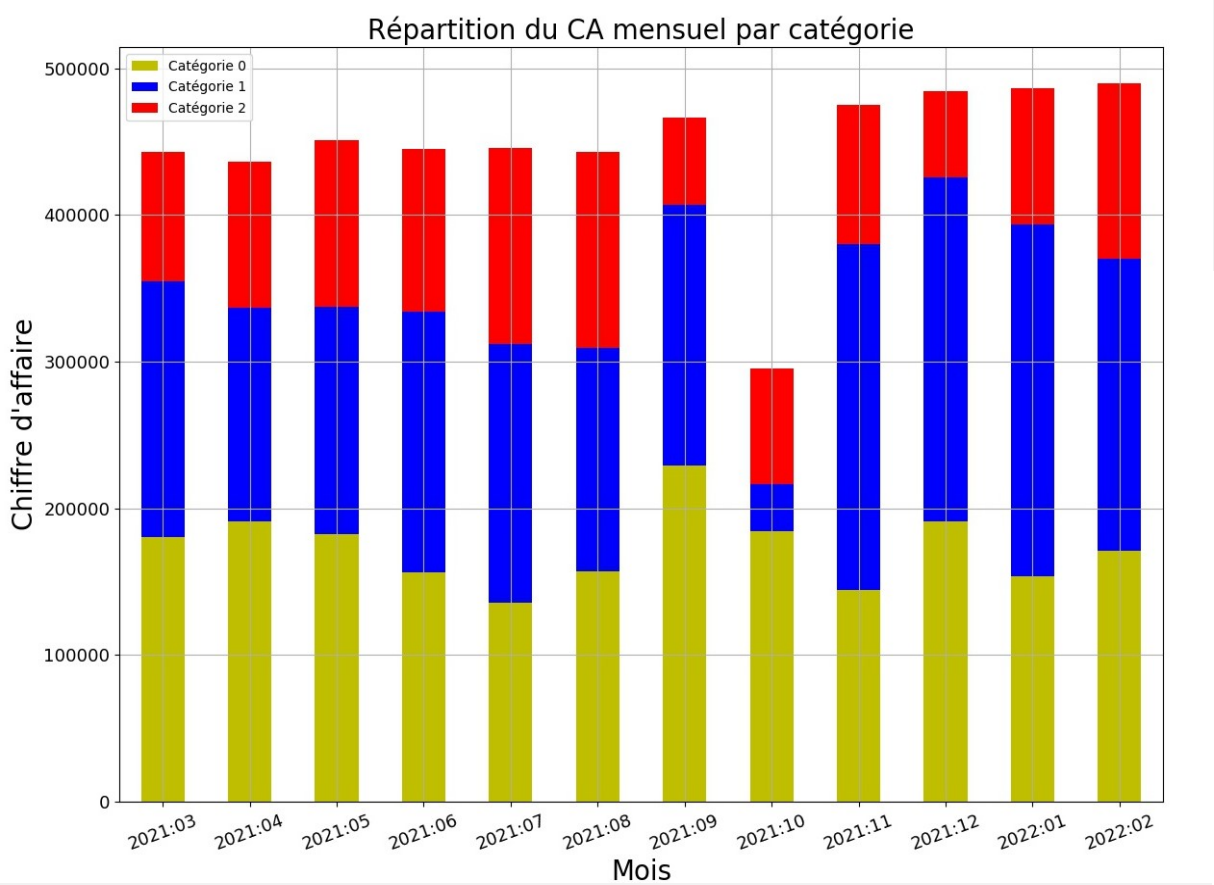
RESTER LIVRES

# Analyse graphique des données rassemblées



# Analyse graphique des données rassemblées



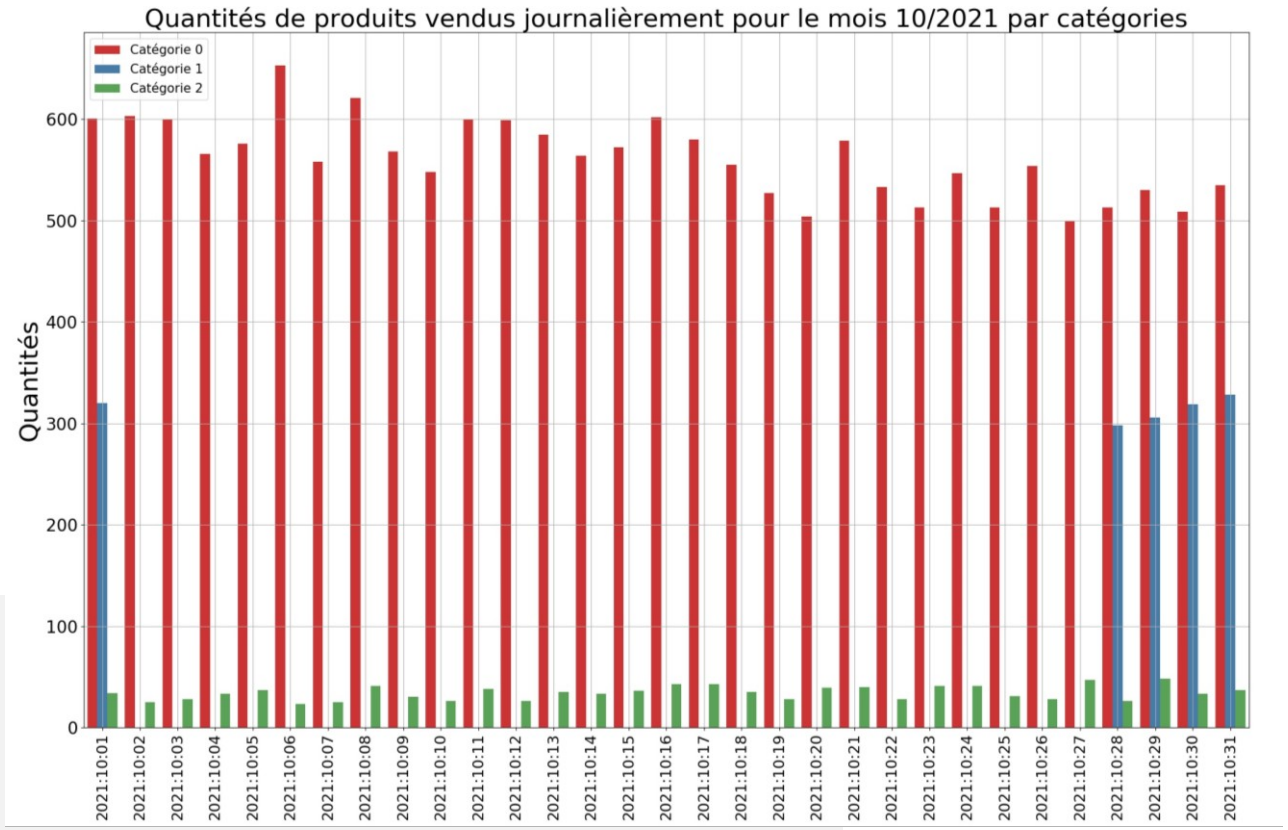

*Les produits de catégorie 1 ont cessés d'etre vendus*

*à partir du*

*2021-10-01 23:50:02.475137*

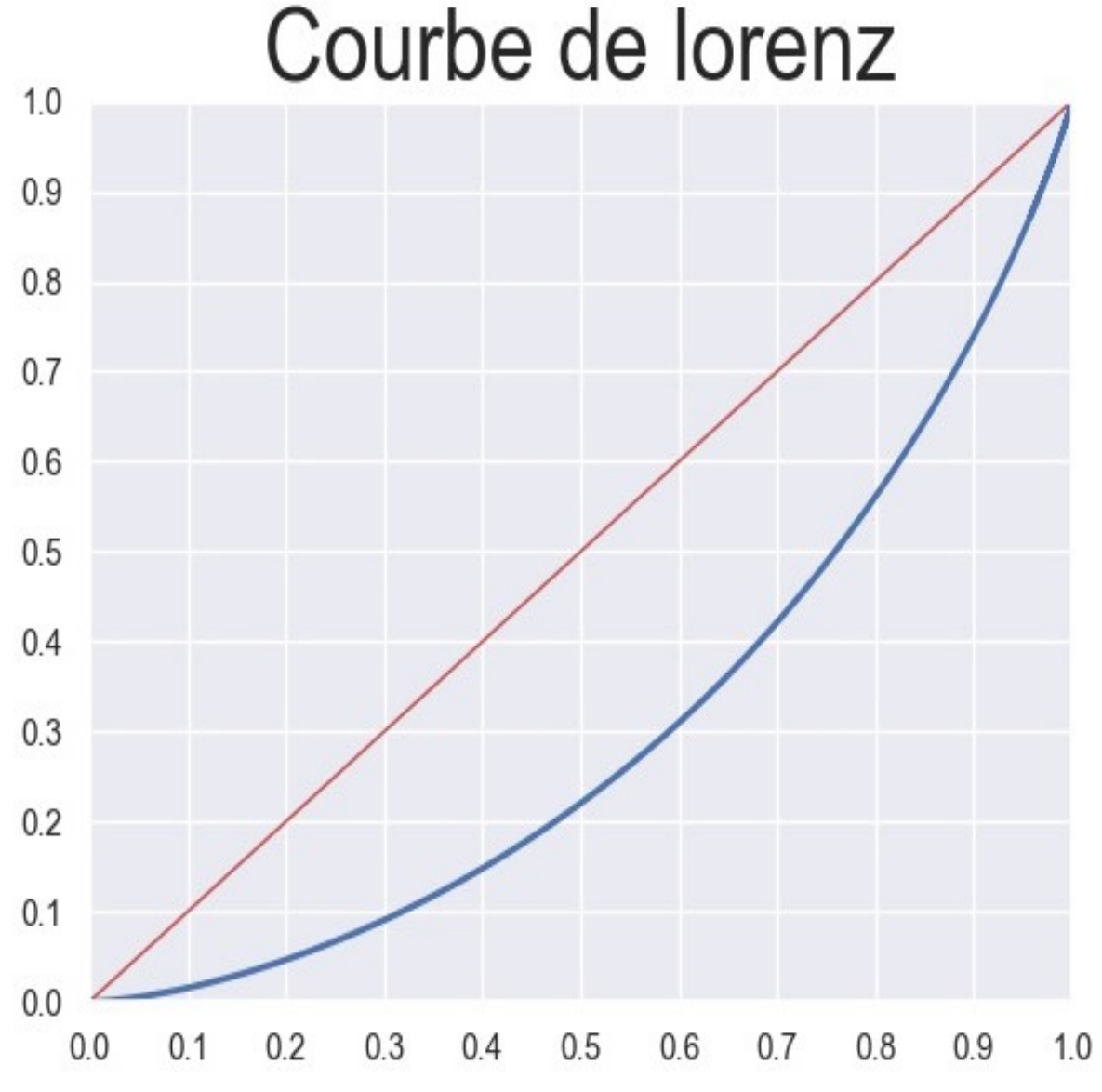
*et leur vente a repris le*

*2021-10-28 00:02:54.903656*




En octobre 2021 on remarque une grosse baisse de CA pour les produits de catégorie 1.

# Analyse graphique des données rassemblées




*Inégalités des dépenses clients existante car l'indice de gini est à*

0.395053650417579 1

*tend plus vers 0 que 1.*

*Répartition des dépenses clients:*

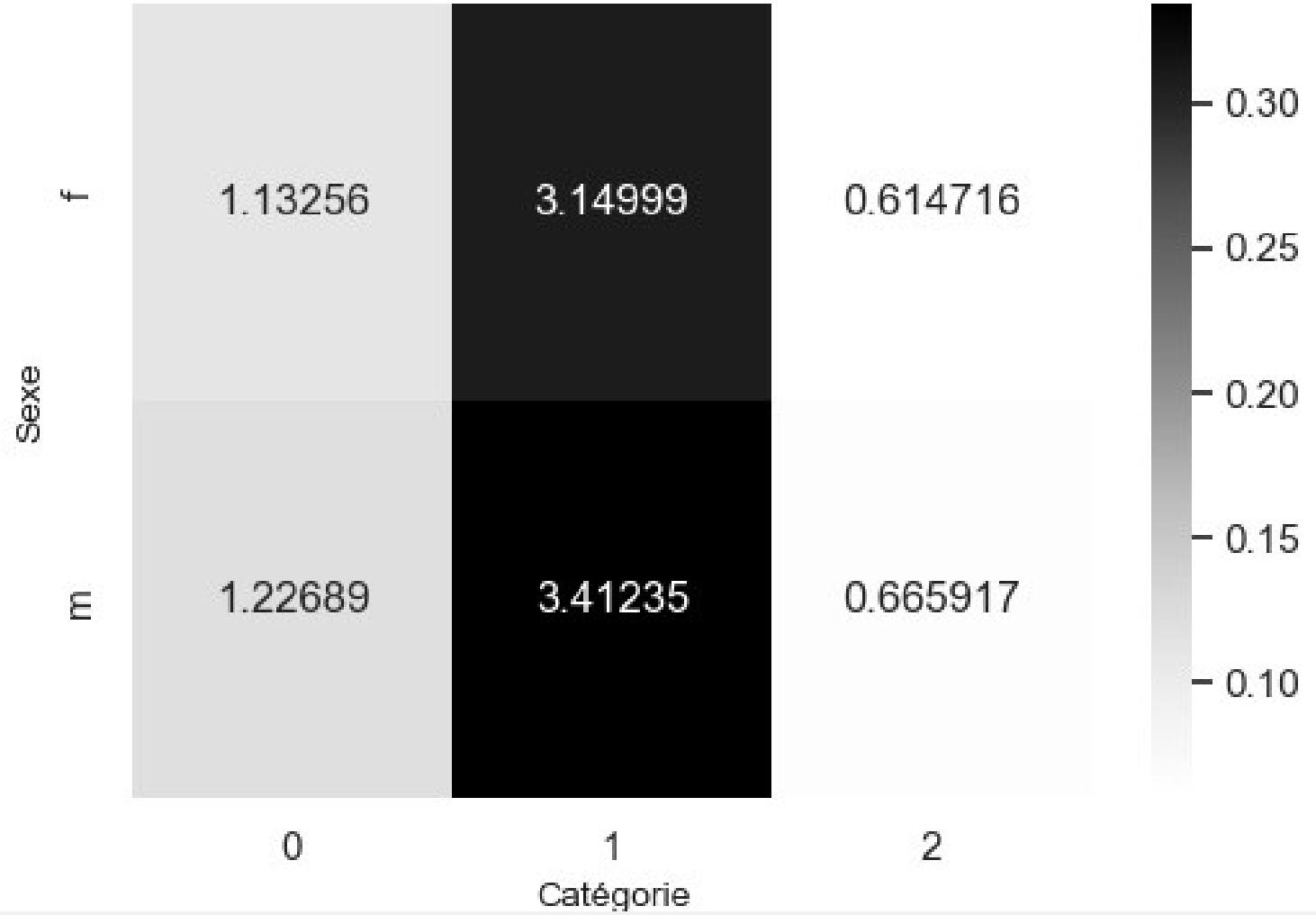
*y=pourcentage cumulé des dépenses*

*x=pourcentage cumulé de clients*

*50% des achats = 25% de la clientèle.*

*50% de la clientèle = 20% des achats .*

# Y a-t-il une corrélation entre le sexe des clients et les catégories de produits achetés ?




xi_n

10.20241727727317 4

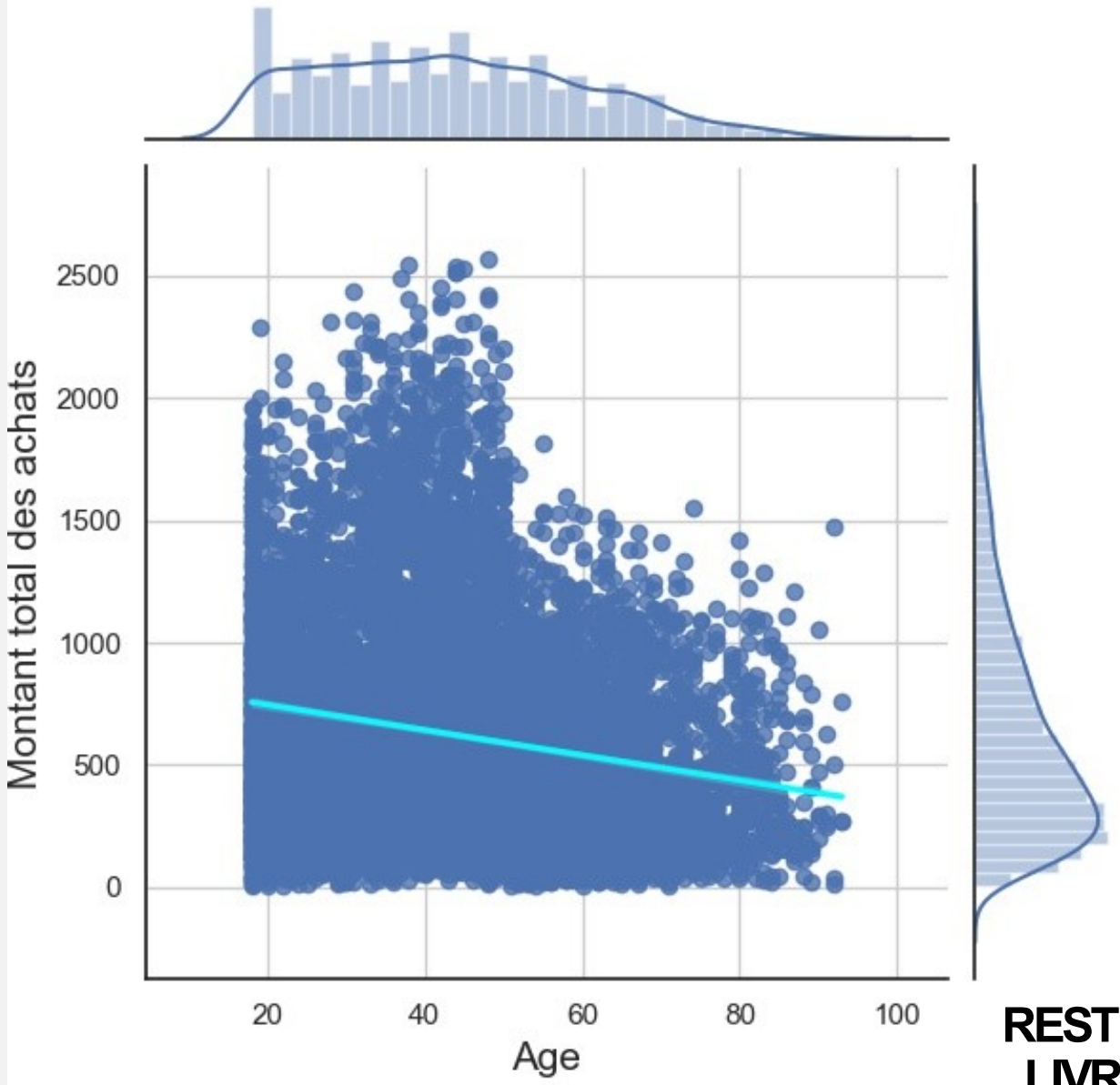
Pas de corrélation entre le sexe et la catégorie de produit acheté.

# Y a-t-il une corrélation entre l'âge des clients et le montant total des achats?

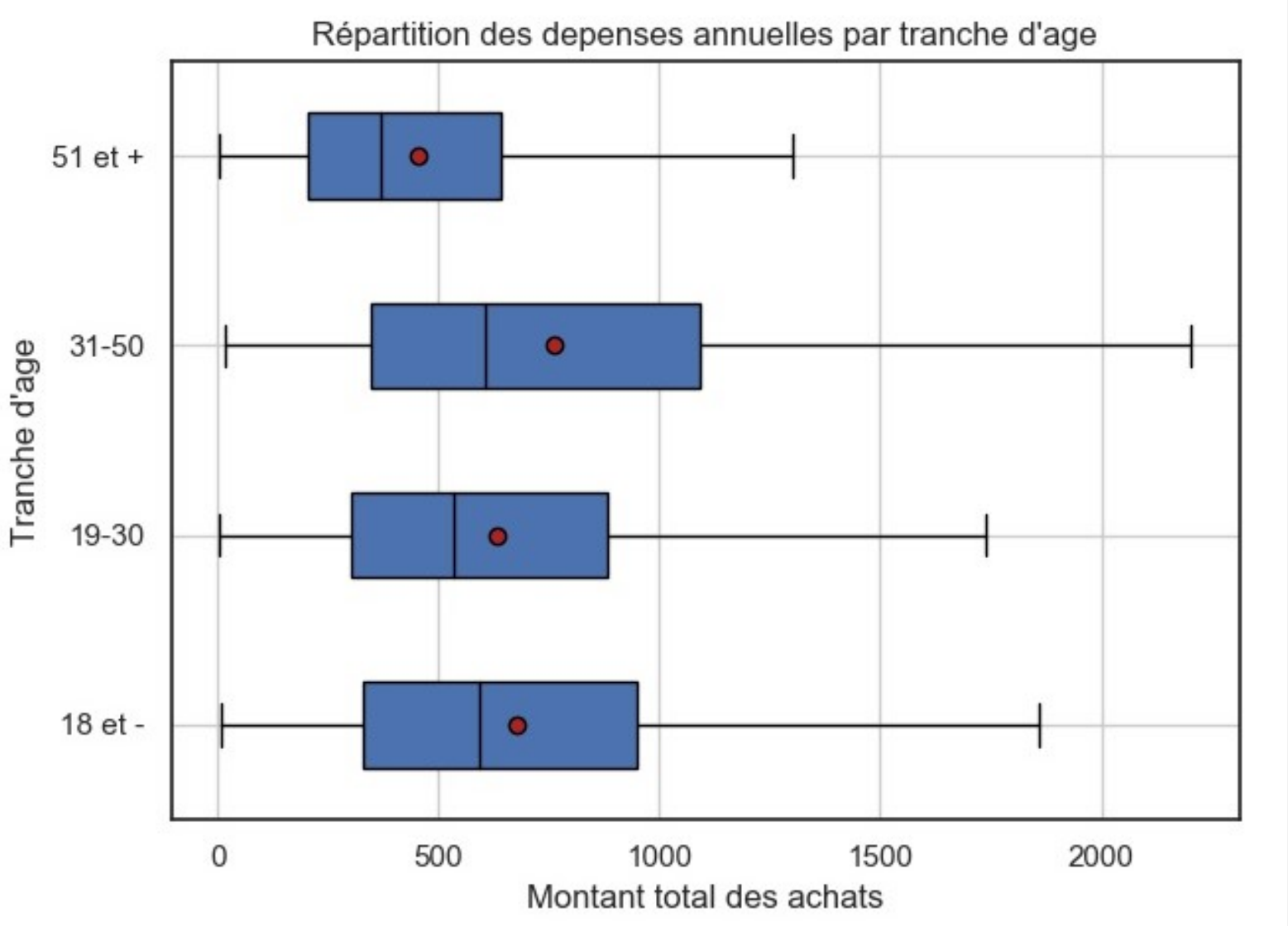
Coefficient de détermination ($R^2$):
0.0362609800913991 1

Pas de corrélation

Et par tranche d'âge ?



# Y a-t-il une corrélation entre l'âge des clients et le montant total des achats?




Pour les 18 et -
achat moyenne: 675.95
achat médian 591.58
Mode: 8.82
Ecart-type: 442.48

Pour les 19-30
achat moyenne: 633.64
achat médian 533.2
Mode: 118.98
Ecart-type: 423.16

Pour les 31-50
achat moyenne: 762.96
achat médian 607.77
Mode: 131.1
Ecart-type: 524.28

Pour les 51 et
achat moyenne: 452.28
achat médian 369.91
Mode: 74.95
Ecart-type: 319.95

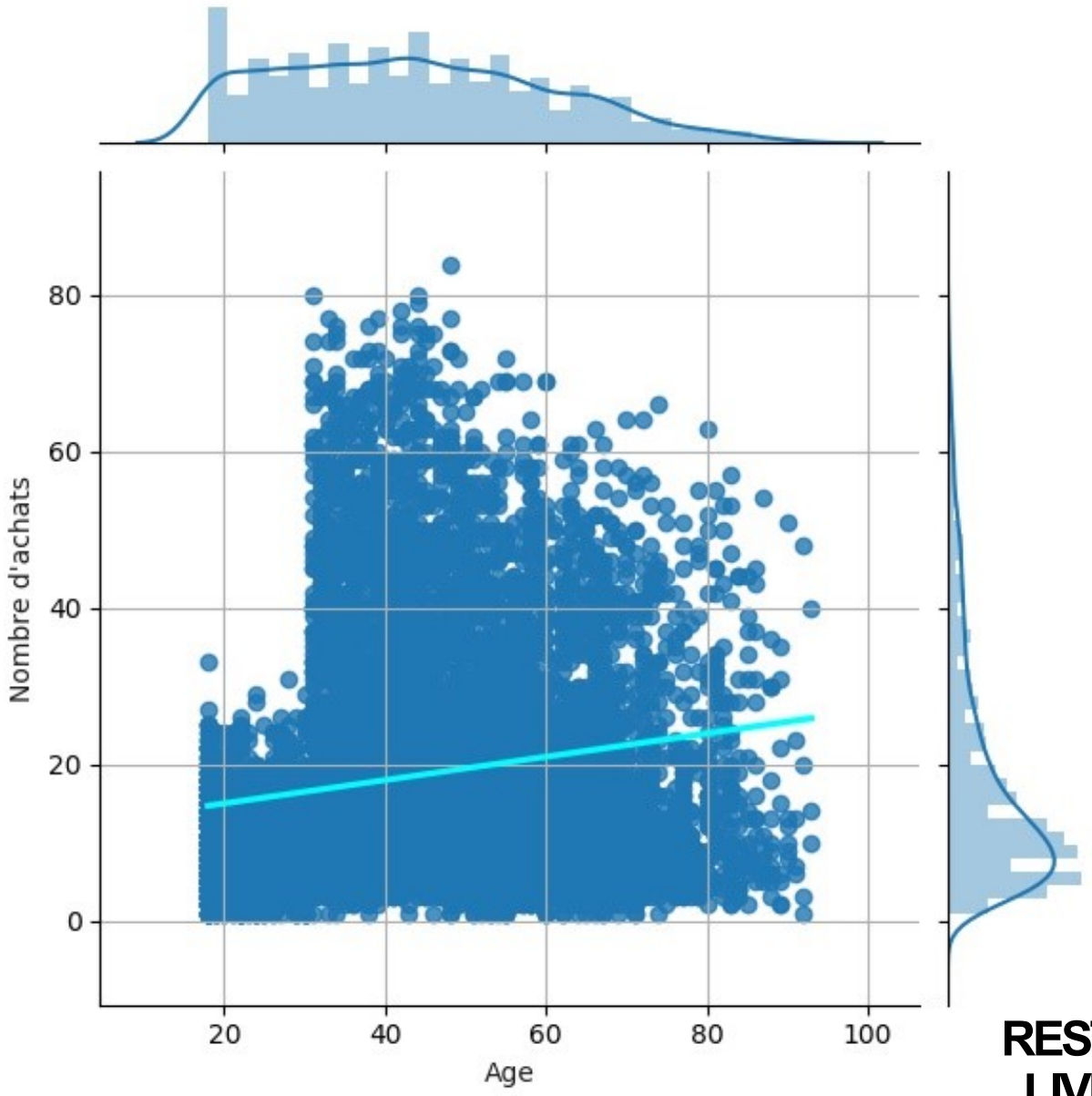
L'éta-carré (η2) est de 0.0856465377948988 1

# Y a-t-il une corrélation entre l'âge des clients et la fréquence d'achat par an?

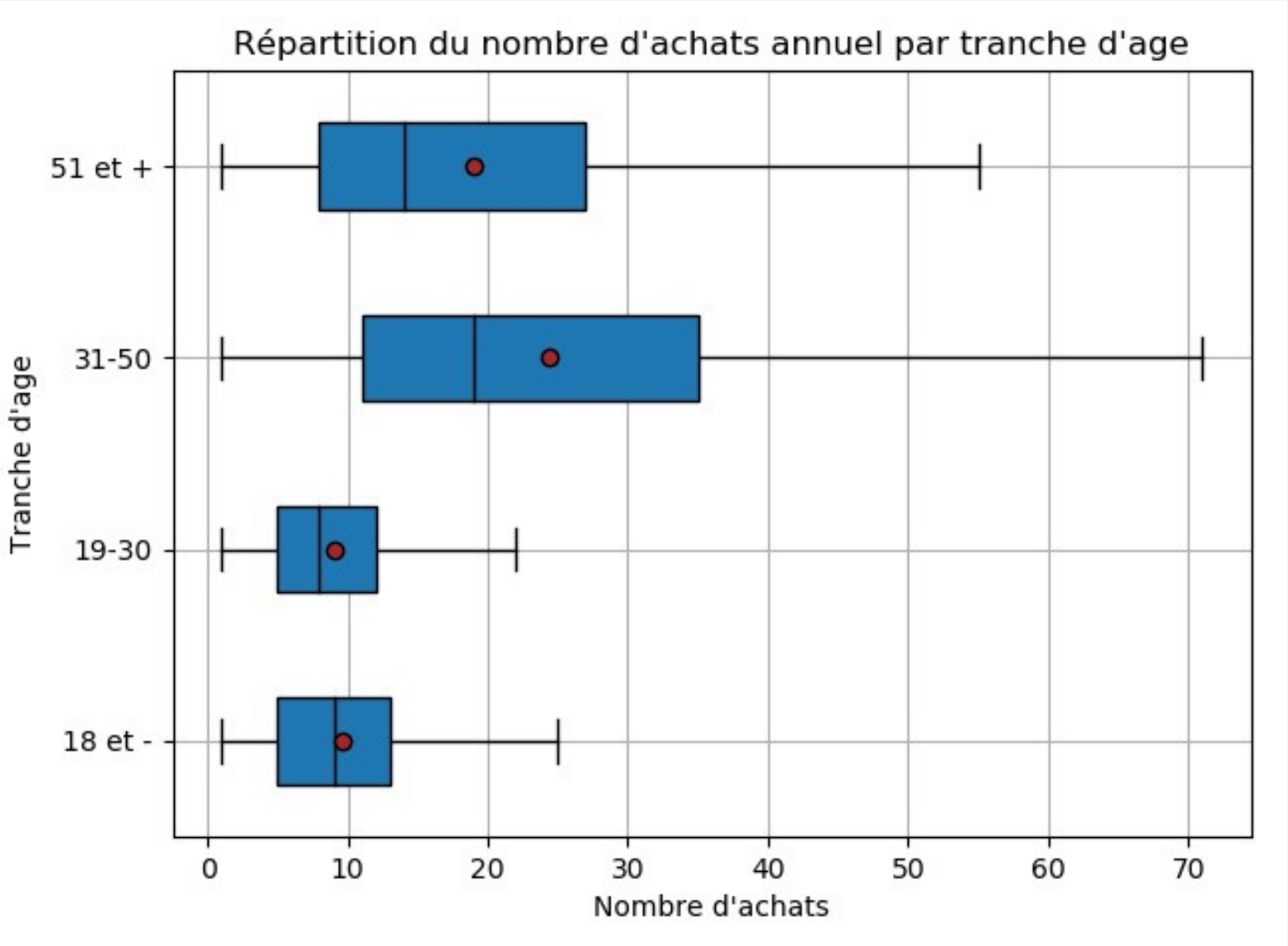
Coefficient de détermination ($R^2$):
0.027093535682130996

Pas de corrélation

Et par tranche d'âge ?



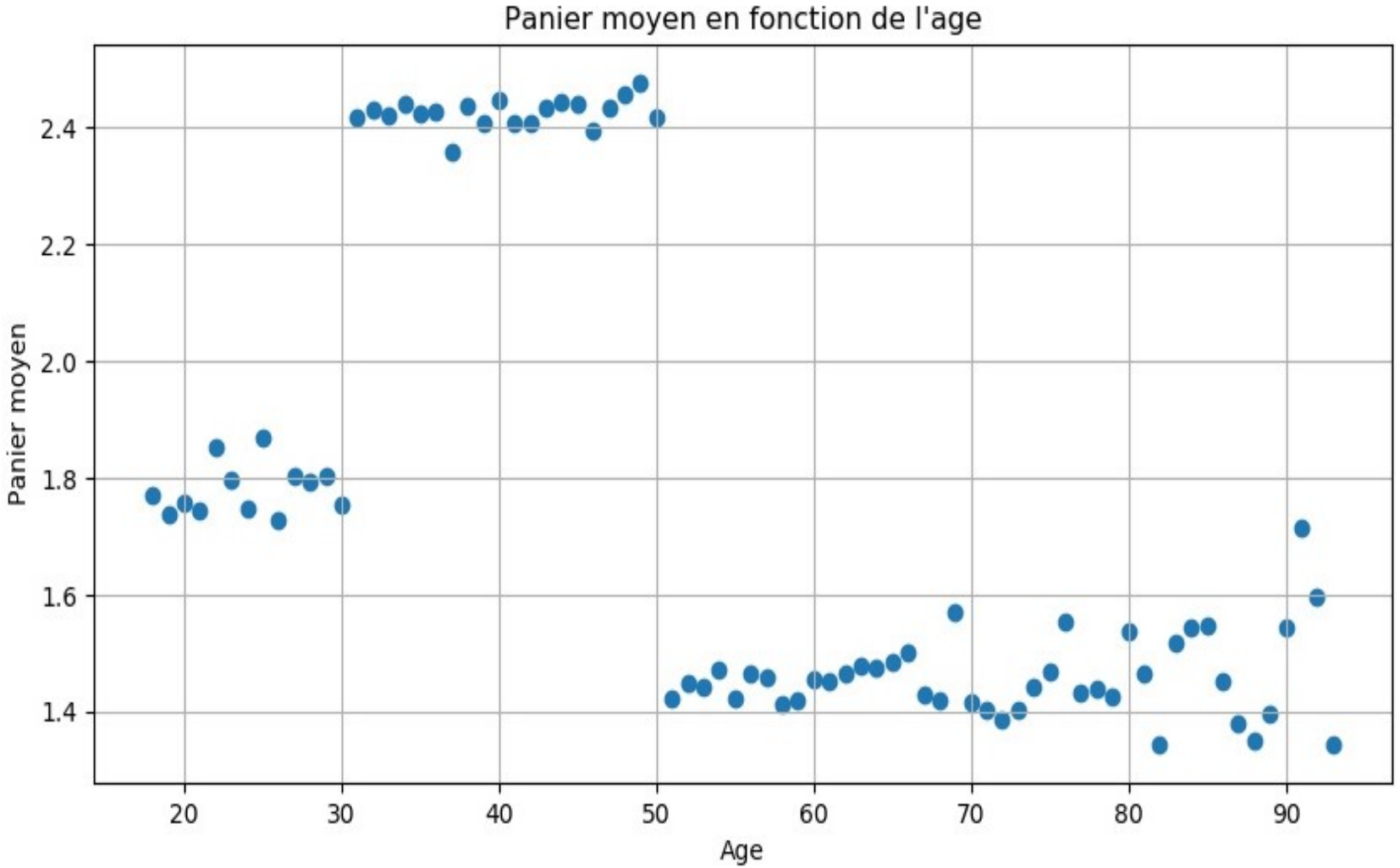
# Y a-t-il une corrélation entre l'âge des clients et la fréquence d'achat par an?




L'éta-carré (η2) est de 0.37014990886538435

Achats pour les 18 et -
achat moyenne: 9.68
achat médian 9.0
Mode: 5
Ecart-type: 5.68
---
Achats pour les 19-30
achat moyenne: 9.07
achat médian 8.0
Mode: 5
Ecart-type: 5.49
---
Achats pour les 31-50
achat moyenne: 24.45
achat médian 19.0
Mode: 11
Ecart-type: 17.37
---
Achats pour les 51 et +
achat moyenne: 18.95
achat médian 14.0
Mode: 6
Ecart-type: 14.47

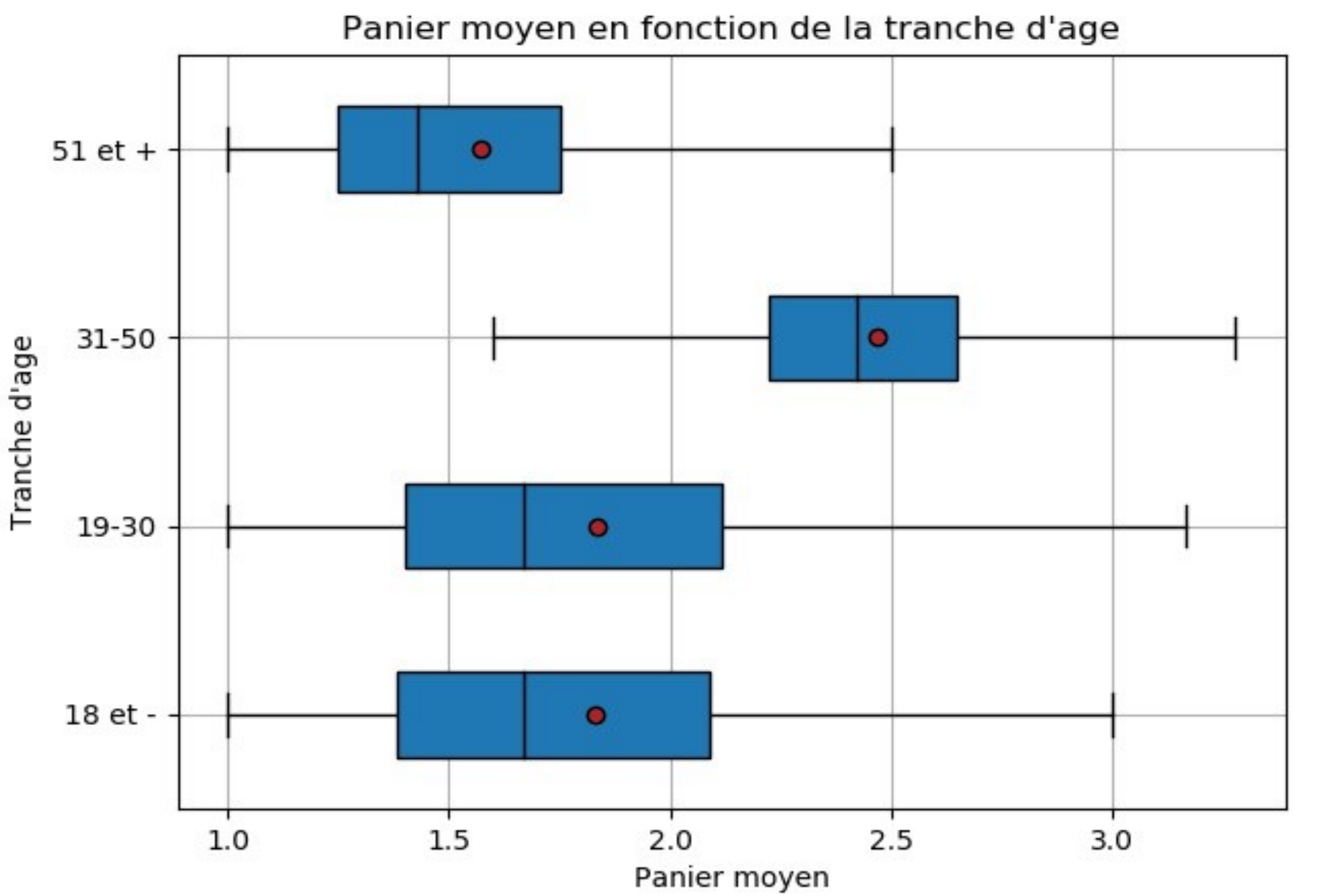
# Y a-t-il une corrélation entre l'âge des clients et La taille du panier moyen ?




Coefficient de détermination ($R^2$): 0.05168913800813791 6

Pas de corrélation entre l'âge et la taille du panier moyen, et par tranche d'âge ?

# Y a-t-il une corrélation entre l'âge des clients et La taille du panier moyen ?




18 et -
qté moyenne: 1.83
qté médiane 1.67
Mode: 2.0
Ecart-type: 0.66

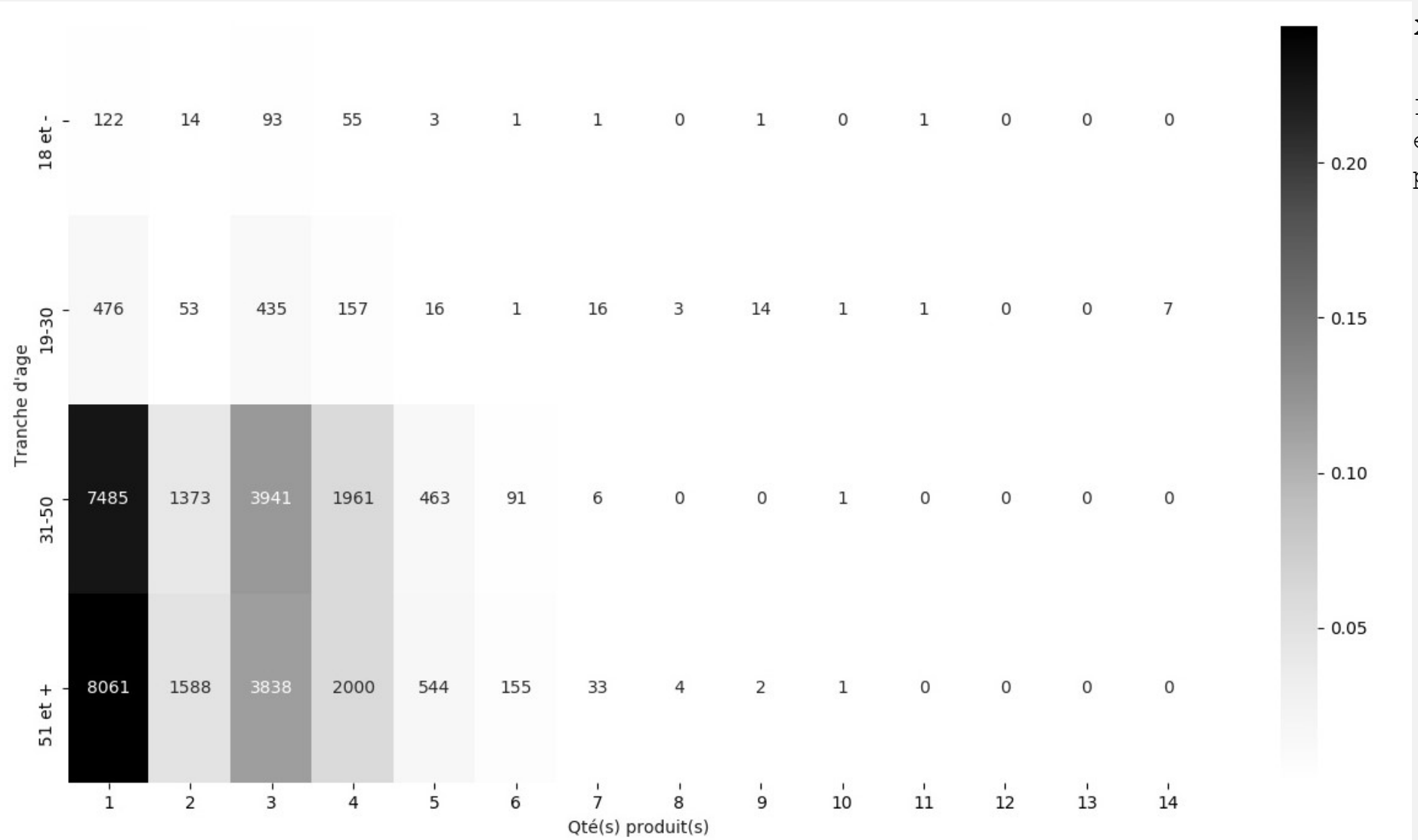
19-30
qté moyenne: 1.84
qté médiane 1.67
Mode: 2.0
Ecart-type: 0.66

31-50
qté moyenne: 2.47
qté médiane 2.42
Mode: 2.0
Ecart-type: 0.4

51 et +
qté moyenne: 1.57
qté médiane 1.43
Mode: 1.0
Ecart-type: 0.5

L'éta-carré (η2) est de 0.37014990886538435

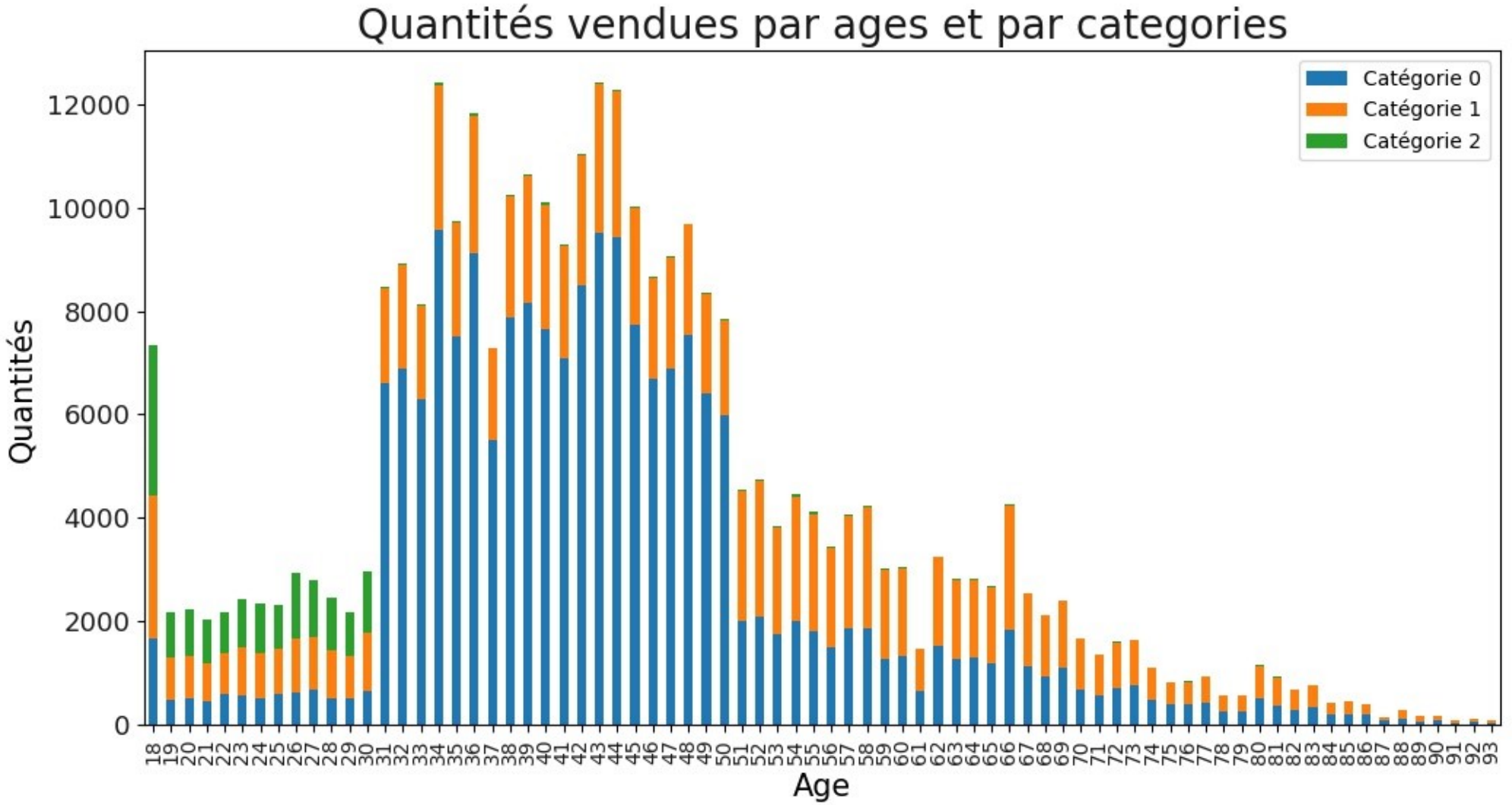
# Y a-t-il une corrélation entre l'âge des clients et La taille du panier moyen ?



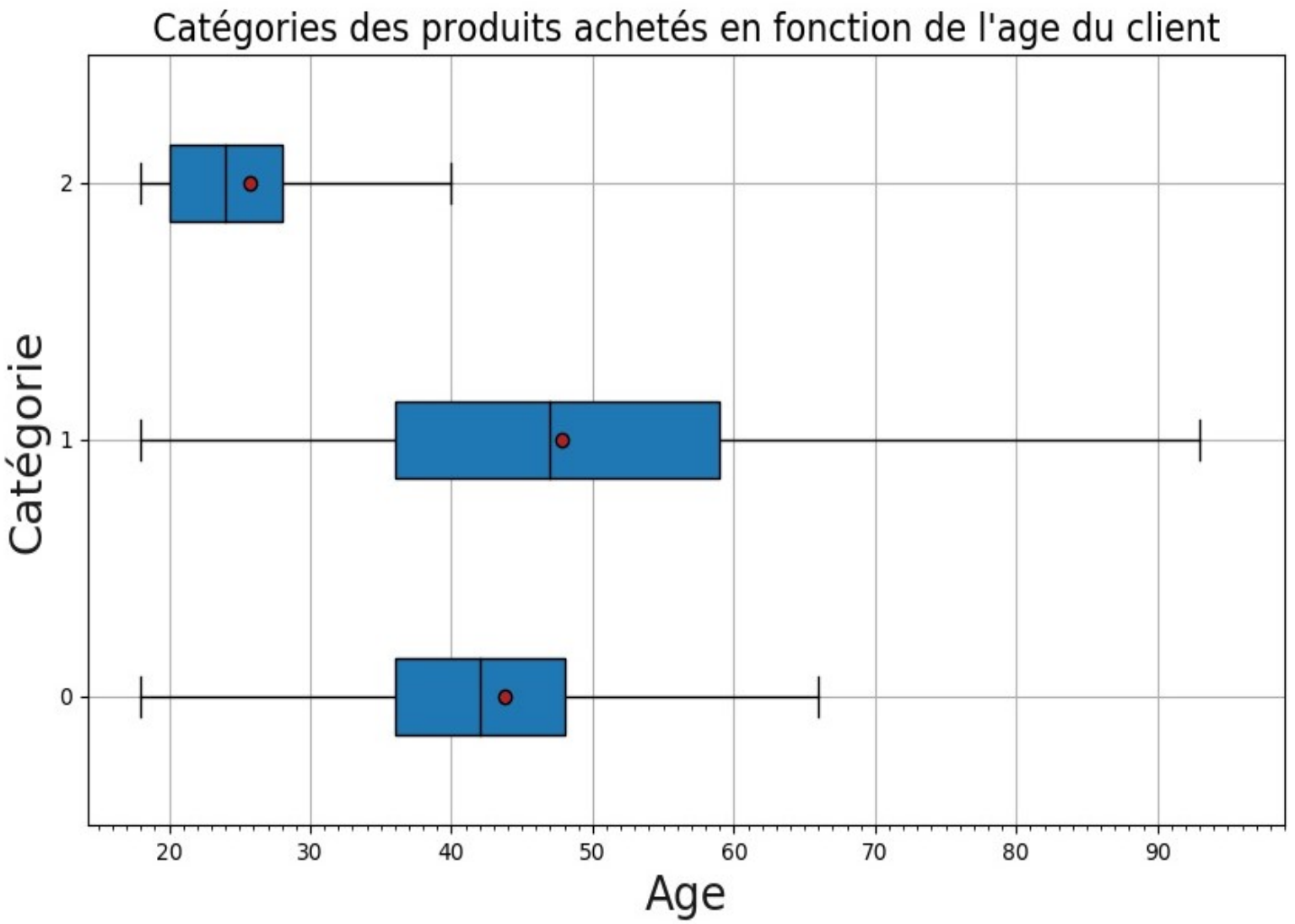

xi_n : 33041

Il existe un lien fort entre le nombre d'articles panier et la tranche d'âge.

# Y a-t-il une corrélation entre l'âge des clients et les catégories de produits achetés?



# Y a-t-il une corrélation entre l'âge des clients et les catégories de produits achetés?




Catégorie 0
Age moyen: 43.8
Age médian: 42.0
Mode: 34

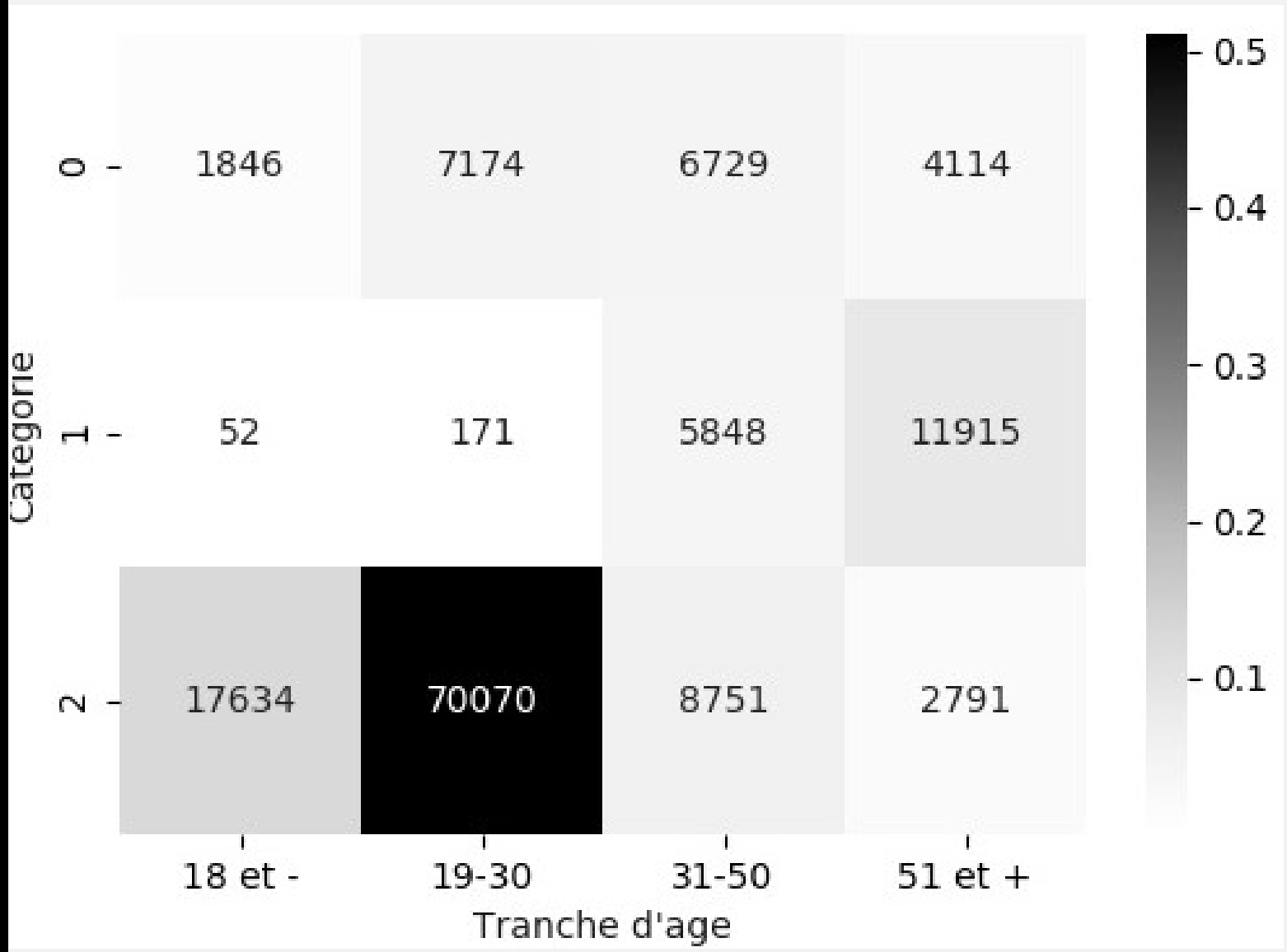
Catégorie 1
Age moyen: 47.81
Age médian: 47.0
Mode: 43

Catégorie 2
Age moyen: 25.75
Age médian: 24.0
Mode: 18

L'éta-carré (η2) est de 0.1128169615984857 8.

Pas de corrélation avec l'âge,
et par tranche d'âge ?

# Y a-t-il une corrélation entre l'âge des clients et les catégories de produits achetés?




xi_n : 137100

Il existe un lien fort entre la tranche d'âge des clients et la catégorie de produits achetés.

# Conclusions

La clientèle peut être divisé en quatre tranche d'âge dont les habitudes d'achats diffèrent
*(18 ans, 19-30 ans, 31-50 ans et 51-93 ans).*

-La catégorie 2 est achetée principalement par les clients de 18 à 30 ans et représente 22% du CA.

-La tranche d'âge 31-50 ans représente la moitié du CA et achète principalement des produits de catégorie 0 et 1.

-Les catégorie 0 et 1 représentent a part égales 78% du CA.

-La tranche d'âge 51-93 ans achète principalement des produits de catégorie 0 et 1, beaucoup moins que les 31-50 ans ,avec une baisse du CA au fur et à mesure que l'âge augmente.