Analysez les ventes de la librairie

- **I) Nettoyage des données**
- **II) Analyse univariée :**

  indicateurs statistiques et graphiques
- **III) Analyse bivariée :**

  graphiques et tests

Nettoyage des données : Table clients

- ⇒Import du fichier customers.csv
- ⇒Informations sur la table()
- ⇒Pas de valeurs manquantes
- ⇒Pas de doublons

```
len(data_customers['client_id'].unique())
```

```
8623
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8623 entries, 0 to 8622
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   client_id  8623 non-null   object
 1   sex        8623 non-null   object
 2   birth      8623 non-null   int64 
dtypes: int64(1), object(2)
memory usage: 202.2+ KB
```

Nettoyage des données : Table transactions

- ⇒**Import du fichier transactions.csv**
- ⇒**Informations sur la table()**
- ⇒**Pas de valeurs manquantes**

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

- **Modification de type de variable :**

La variable ‘date’ : d’object en datetime

Nettoyage des données : Table produits

- ⇒**Import du fichier products.csv**
- ⇒**Informations sur la table()**
- ⇒**Pas de valeurs manquantes**
-
- ⇒**Pas de doublons**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3287 entries, 0 to 3286
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   id_prod  3287 non-null   object 
 1   price    3287 non-null   float64
 2   categ    3287 non-null   int64  
dtypes: float64(1), int64(1), object(1)
memory usage: 77.2+ KB
```

```
len(data_products['id_prod'].unique())
```

```
3287
```

Rechercher les valeurs aberrantes :

- **Dans la table produits :**

```
data_products['price'].min()
```

```
-1.0
```

```
data_products[data_products['price']==-1]
```

	id_prod	price	categ
731	T_0	-1.0	0

Dans la table transactions

	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

Supprimer ces valeurs (test) dans toutes les tables

Vérifier les tables entre elles : par jointure

Rechercher les valeurs manquantes :


```
data_trans_prod1['price'].isnull().sum()
```

```
103
```

```
data_trans_prod1.loc[data_trans_prod1['price'].isnull()]
```

	id_prod	date	session_id	client_id	price
6231	0_2245	2021-06-17 03:03:12.668129	s_49705	c_1533	NaN
10797	0_2245	2021-06-16 05:53:01.627491	s_49323	c_7954	NaN
14045	0_2245	2021-11-24 17:35:59.911427	s_124474	c_5120	NaN
17480	0_2245	2022-02-28 18:08:49.875709	s_172304	c_4964	NaN
21071	0_2245	2021-03-01 00:09:29.301897	s_3	c_580	NaN
...	...	...	...	...	...

Le produit 0_2245 de catégorie 0 : 2/2

- **Attribuer le prix médian de la catégorie 0,corriger dans la table transaction et introduire dans la table produits**

```
data_trans_prod.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 336816 entries, 0 to 336815
Data columns (total 6 columns):
 #   Column      Non-Null Count   Dtype         
---  ------      --------------   -----         
 0   id_prod     336816 non-null  object        
 1   date        336816 non-null  datetime64[ns]
 2   session_id  336816 non-null  object        
 3   client_id   336816 non-null  object        
 4   price       336816 non-null  float64       
 5   categ       336816 non-null  int64         
dtypes: datetime64[ns](1), float64(1), int64(1), object(3)
memory usage: 18.0+ MB
```

Analyse uni-variée :

- **Mesures de tendance centrale:**

  La moyenne

  La médiane

  Le mode

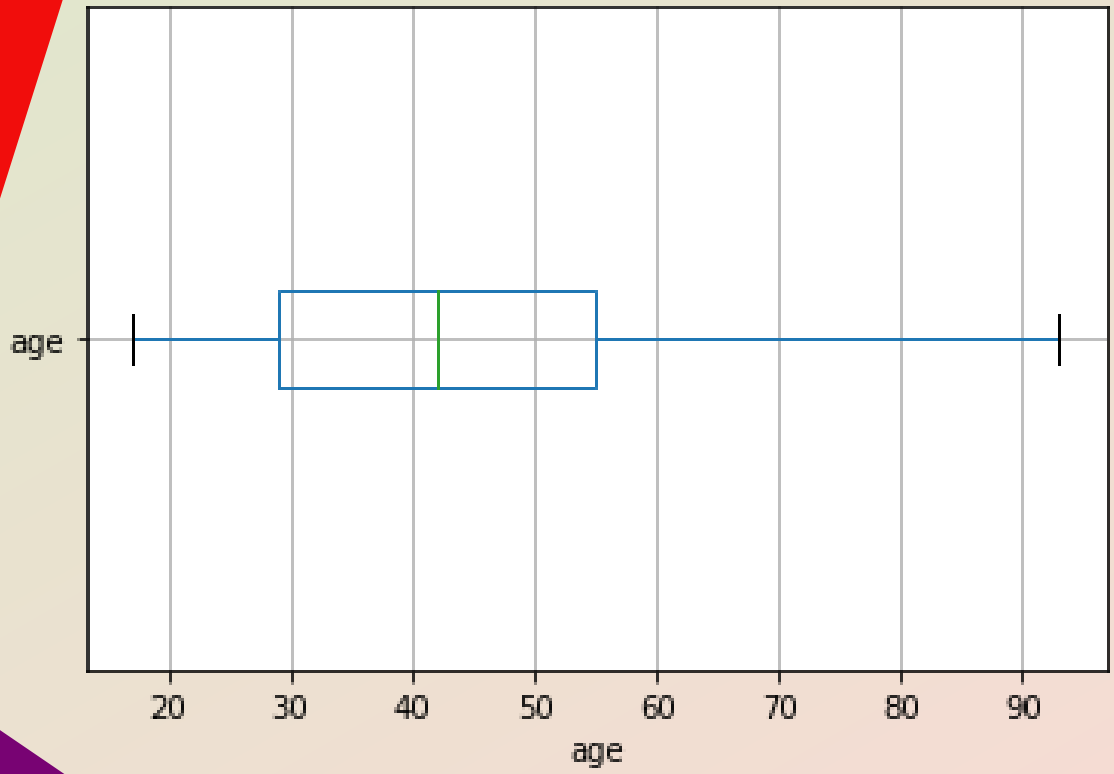
- **Mesures de dispersion¶:**

  La variance empirique :

  Écart type empirique :

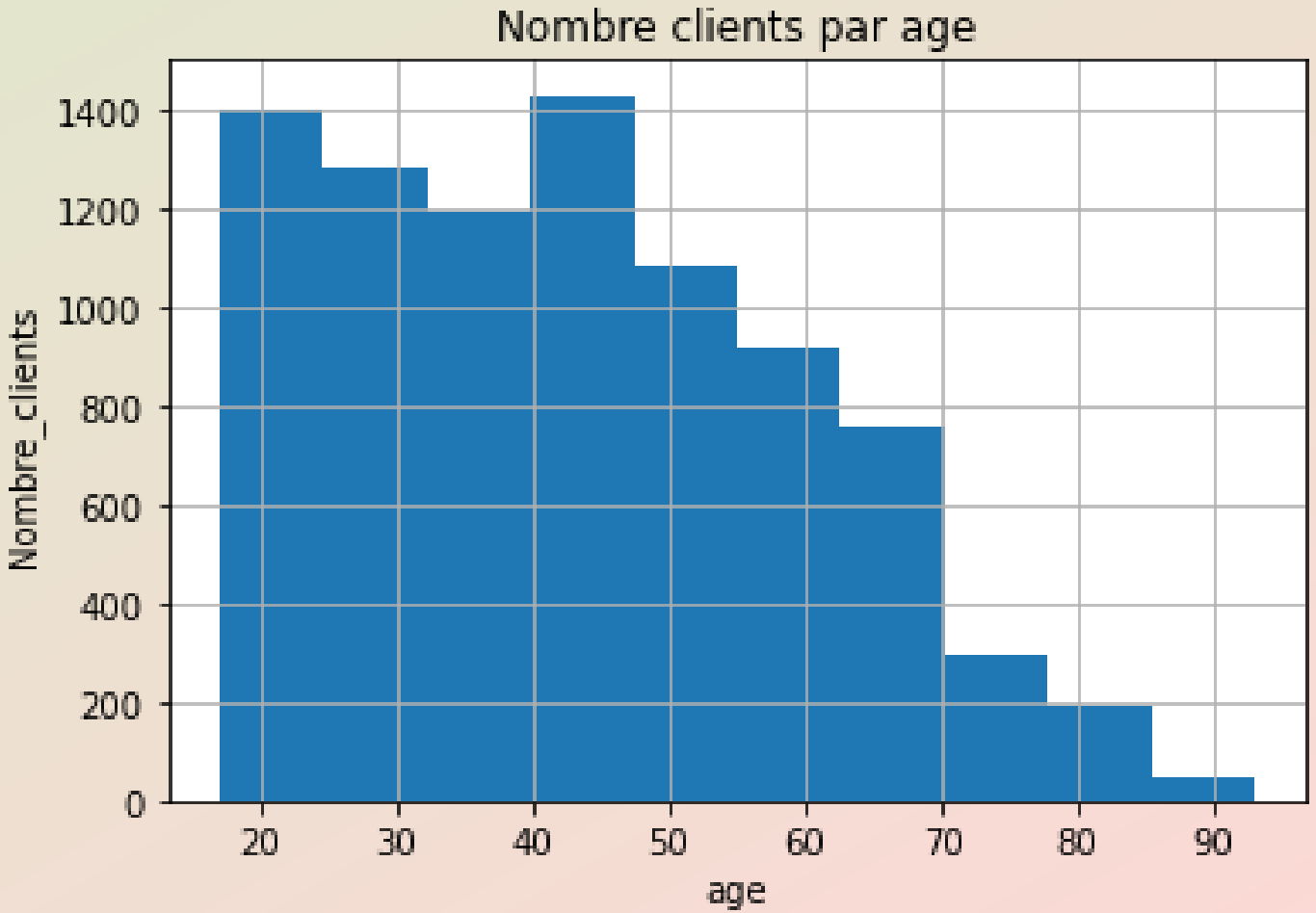
  1er quartile et 3ème quartile

Age clients :

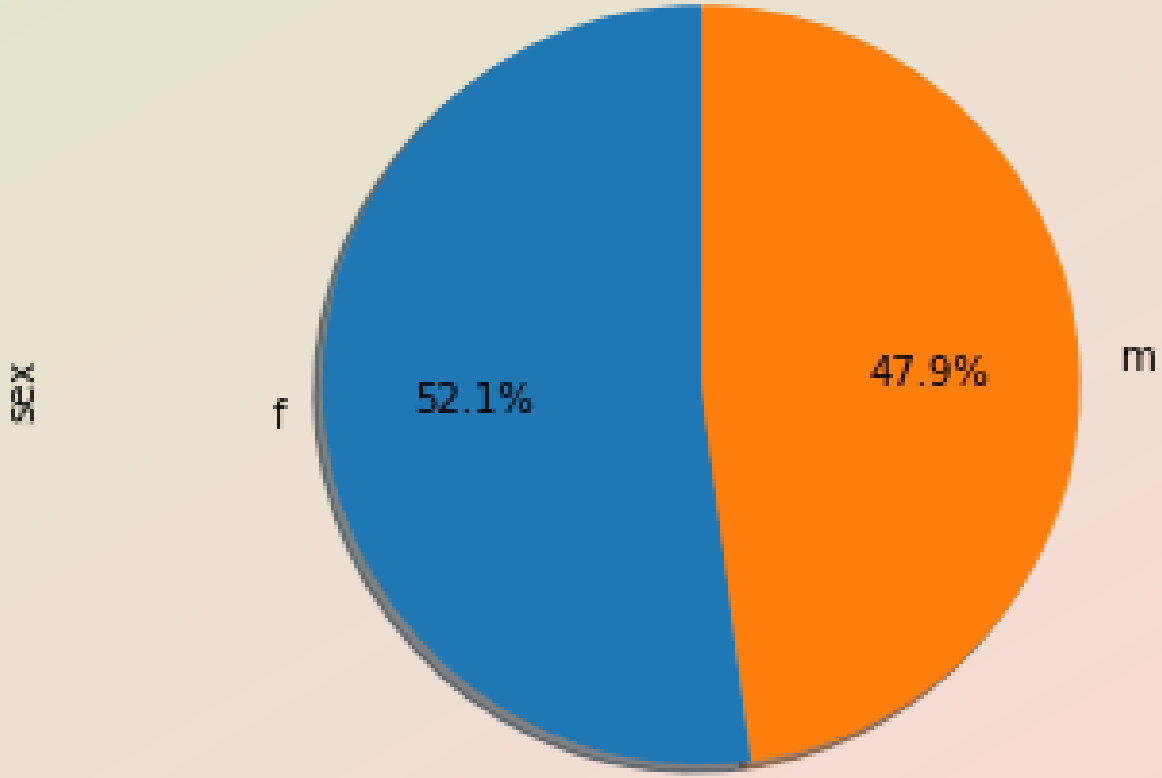


Mesures	Valeurs
Moyenne	42,9
médiane	42
mode	17
min	17
max	93
1er quartile	29
3ème quartile	55
Écart type	16,91
variance	286,04

Age : 17ans à 93ans

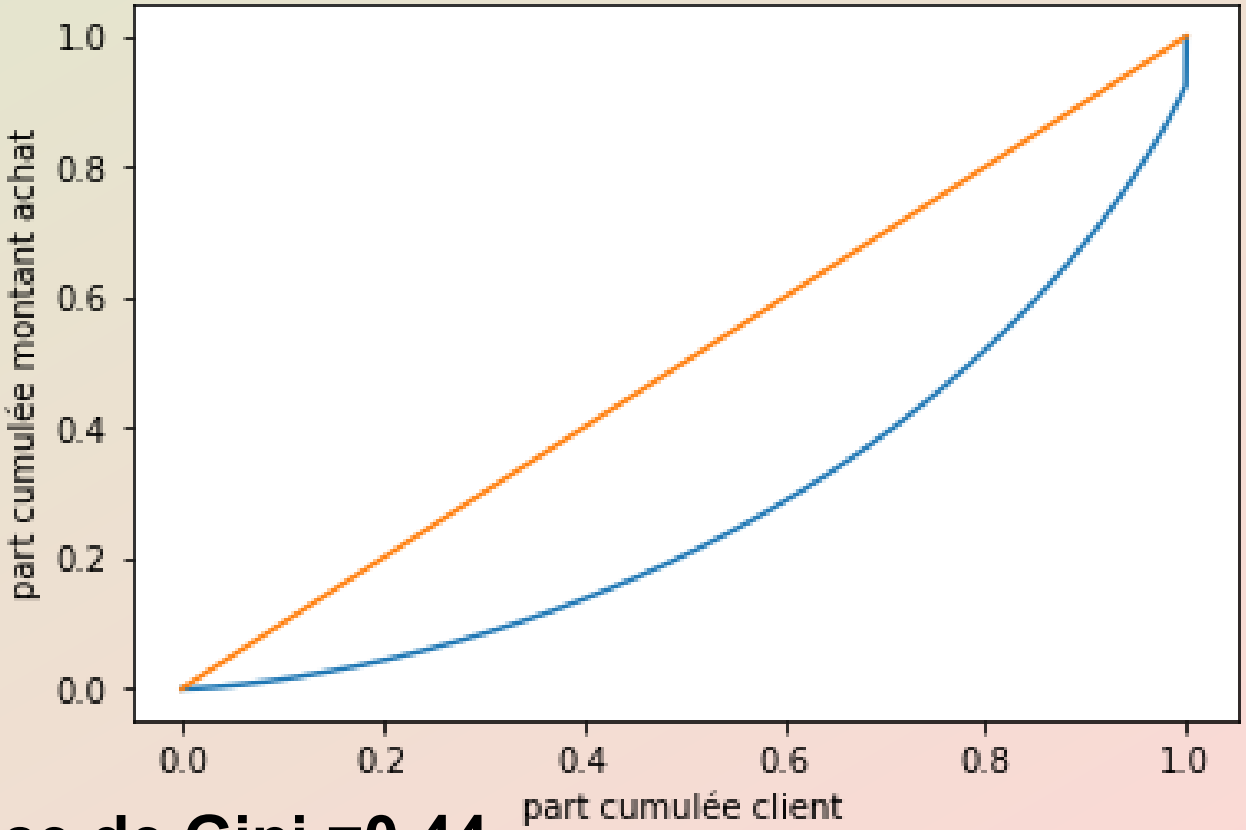


Pourcentage de la clientèle par sexe



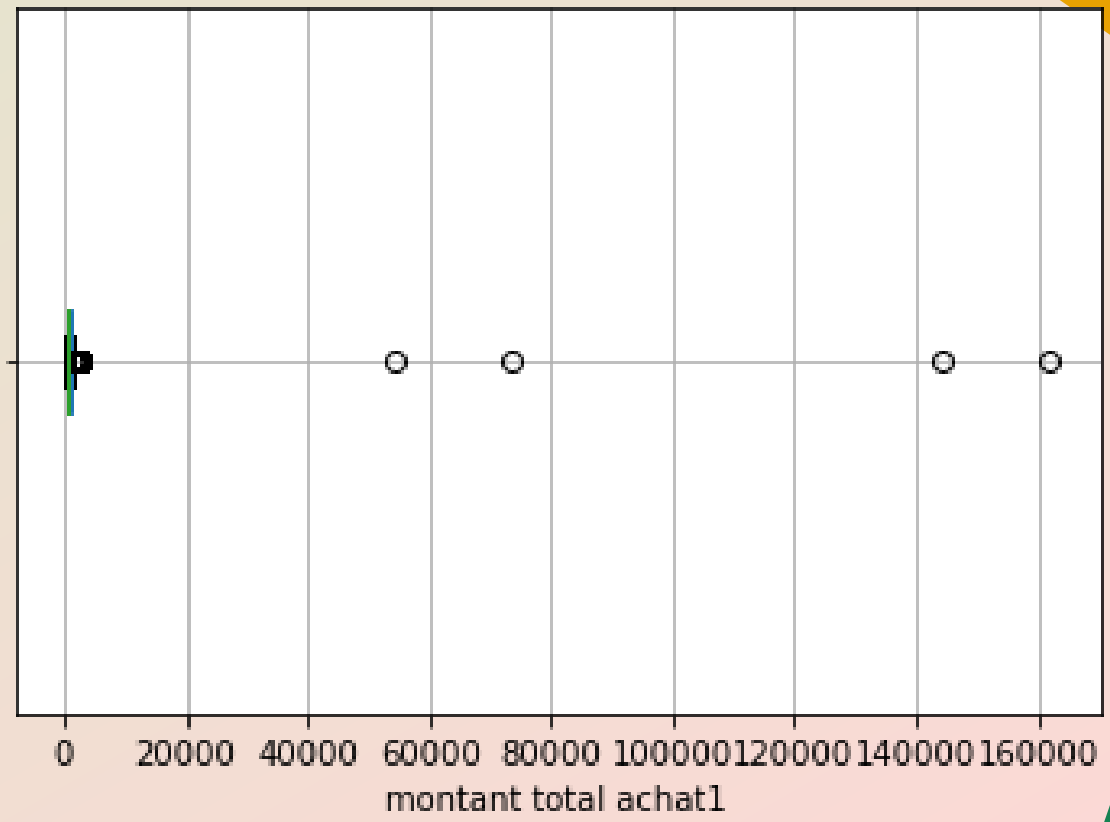
Analyse de concentration
Montant d'achats et nombre de clients

- **Courbe de Lorenz**

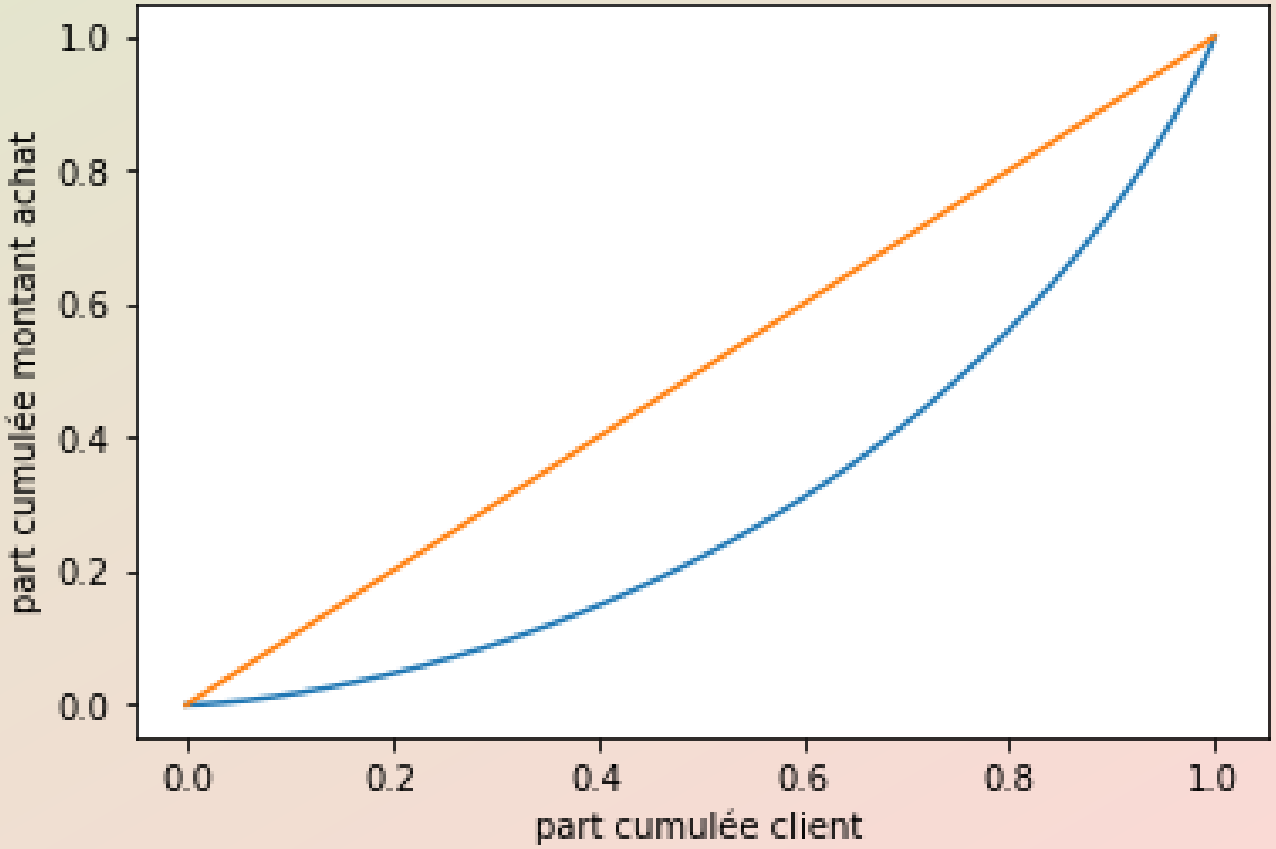


- **Indice de Gini =0,44**

	client_id	montant total achat1
7918	c_8140	4.15
7889	c_8114	4.99
750	c_1675	5.57
8480	c_890	6.08
8151	c_8351	6.31
...	...	...
7715	c_7959	2564.25
2724	c_3454	54463.56
6337	c_6714	73217.98
4388	c_4958	144257.21
677	c_1609	162007.34

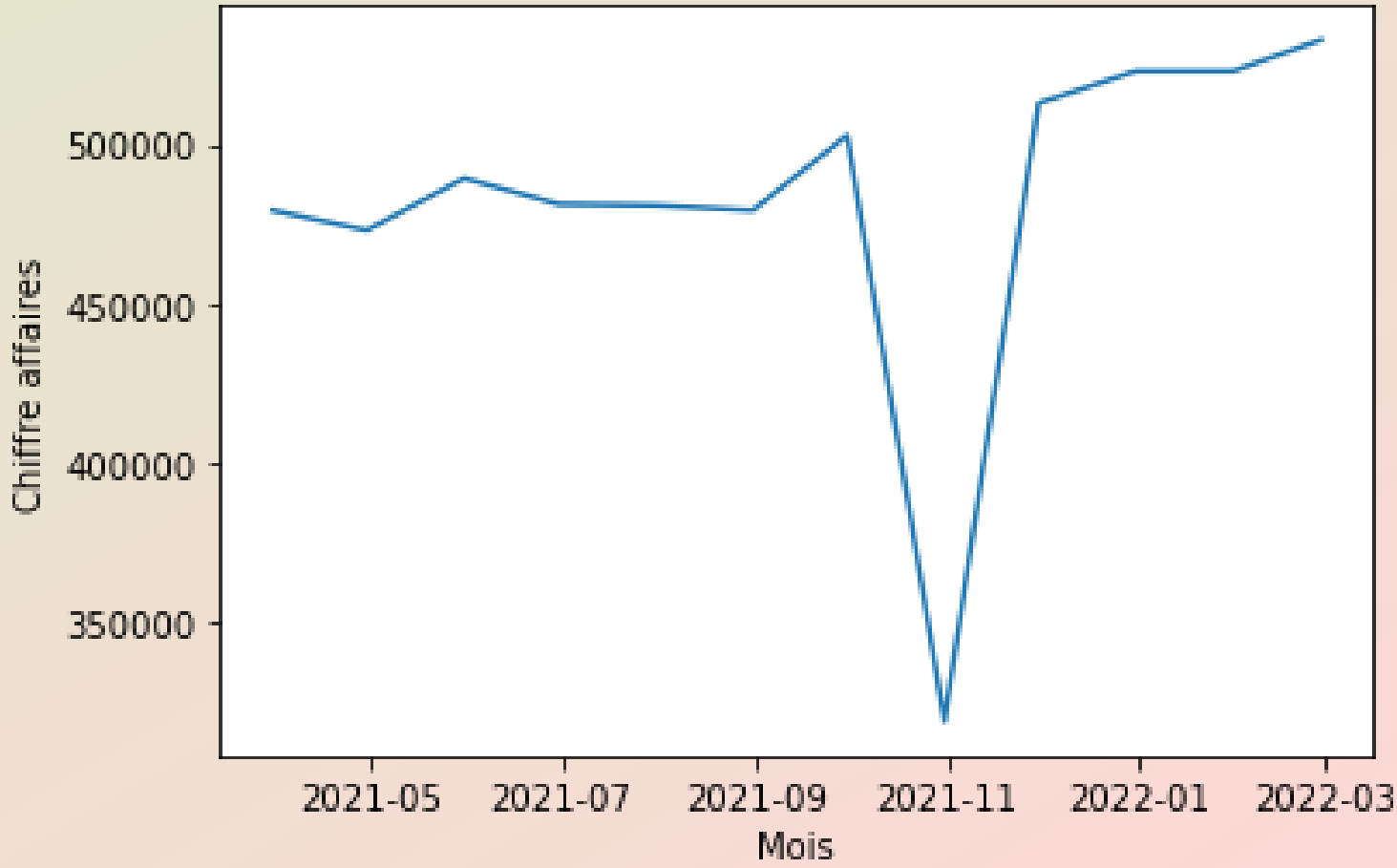


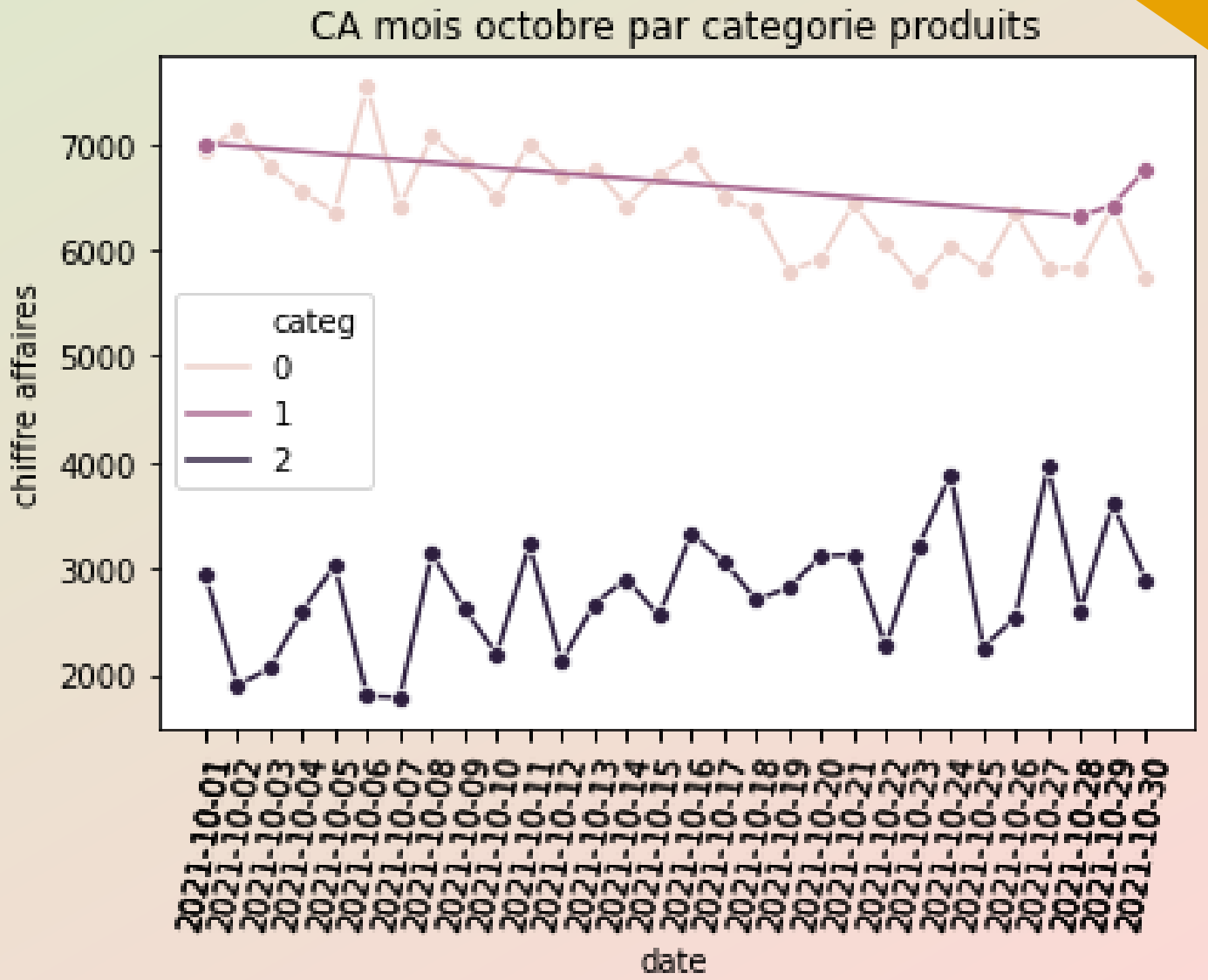
Sans les montants d’achats importants



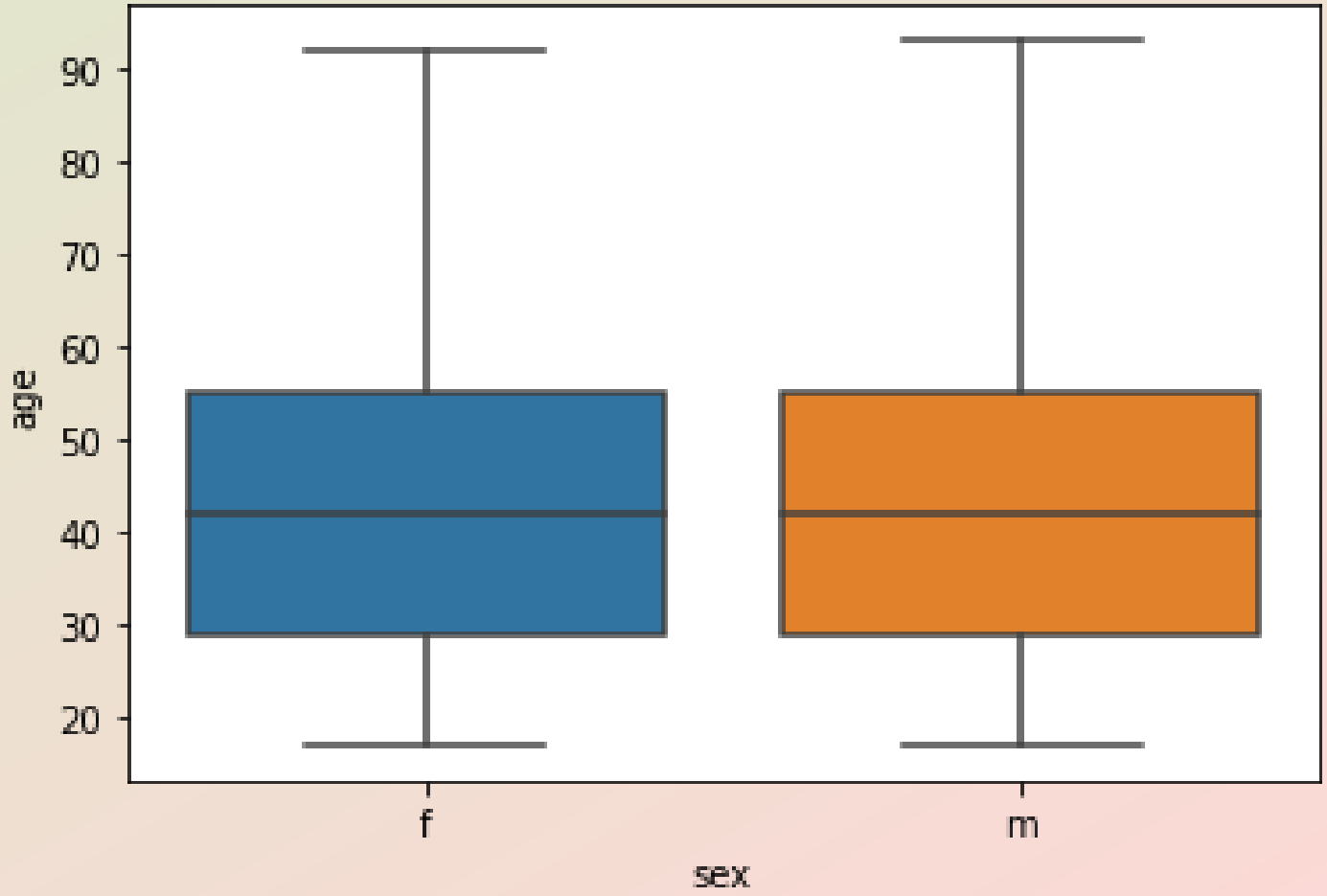
Indice de Gini:0,395

Analyse bivariée
Chiffre d’affaires par mois

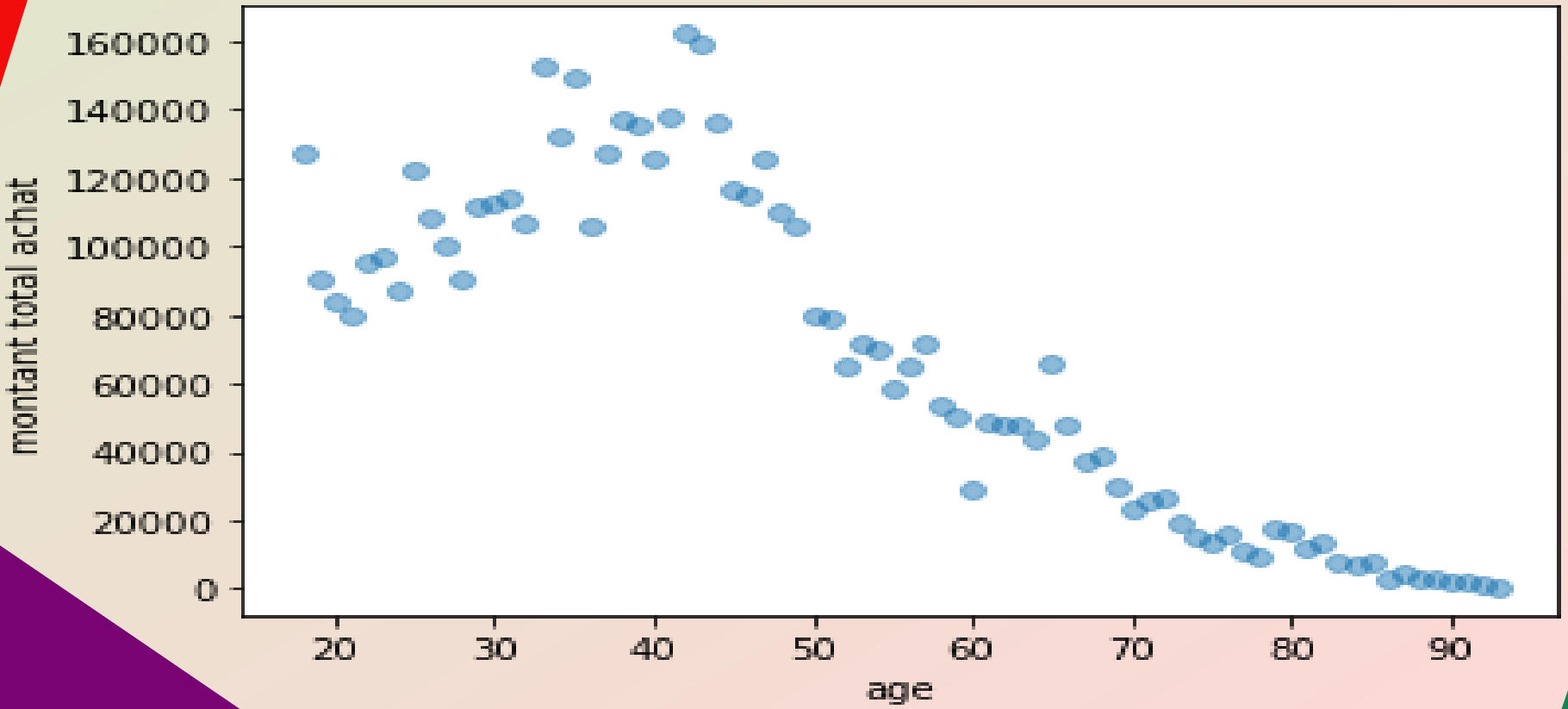



CA mois octobre par categorie produits
chiffre affaires
7000
6000
5000
4000
3000
2000
categ
0
1
2
date

Répartition clientèle par âge et sexe



Étude des corrélations
Âge et montant total achats
Deux variables quantitatives



Indicateurs numériques

- **Le coefficient de corrélation linéaire :**

$$r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$$

$$s_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x - \overline{x})^2$$

- **$r_{X,Y}$ proche de 1 ou de -1 : forte corrélation linéaire**

  rx/y< 0 les deux variables ont des sens de variation inversé

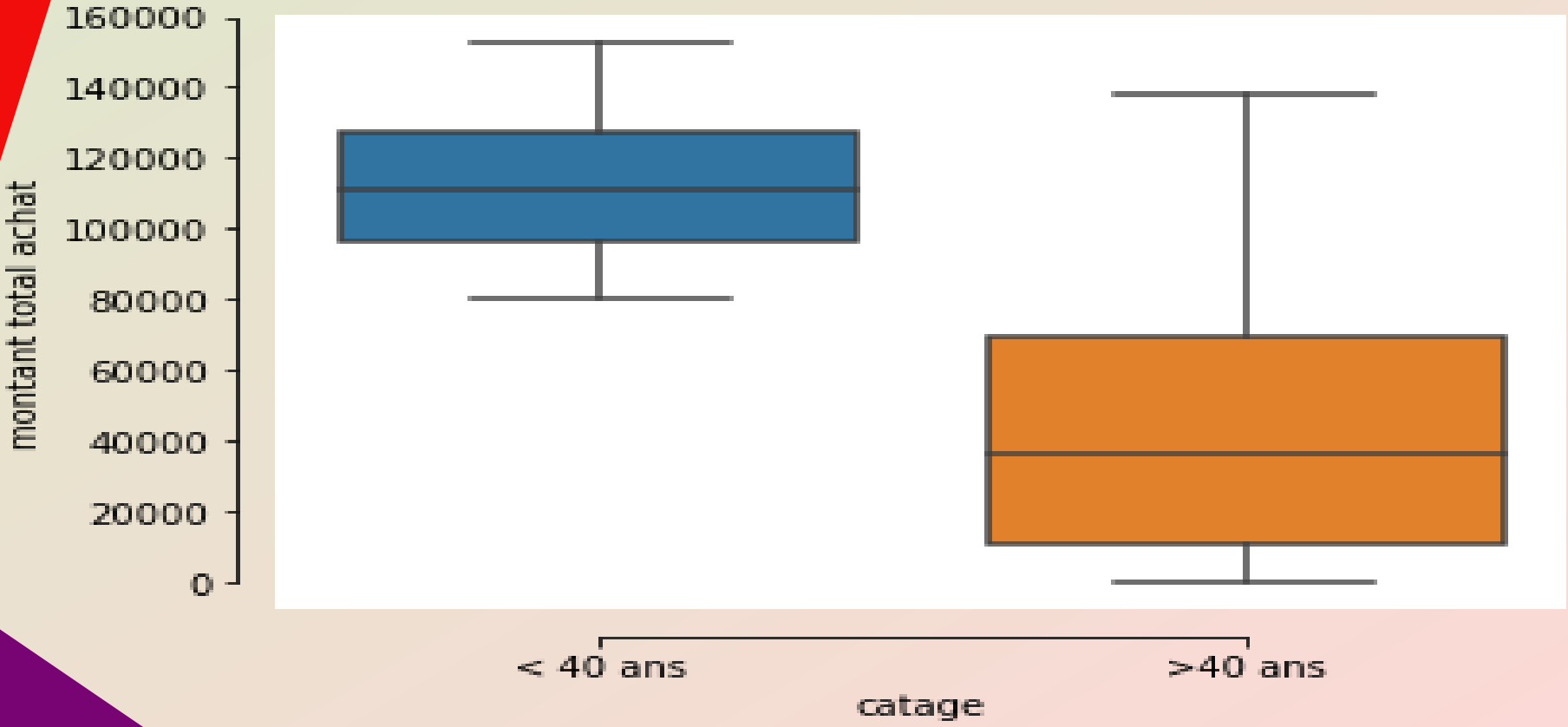
  rX/Y>0 les deux variables sont de même sens de variation

  $r_{X,Y}$ proche de 0 la corrélation n'est pas linéaire

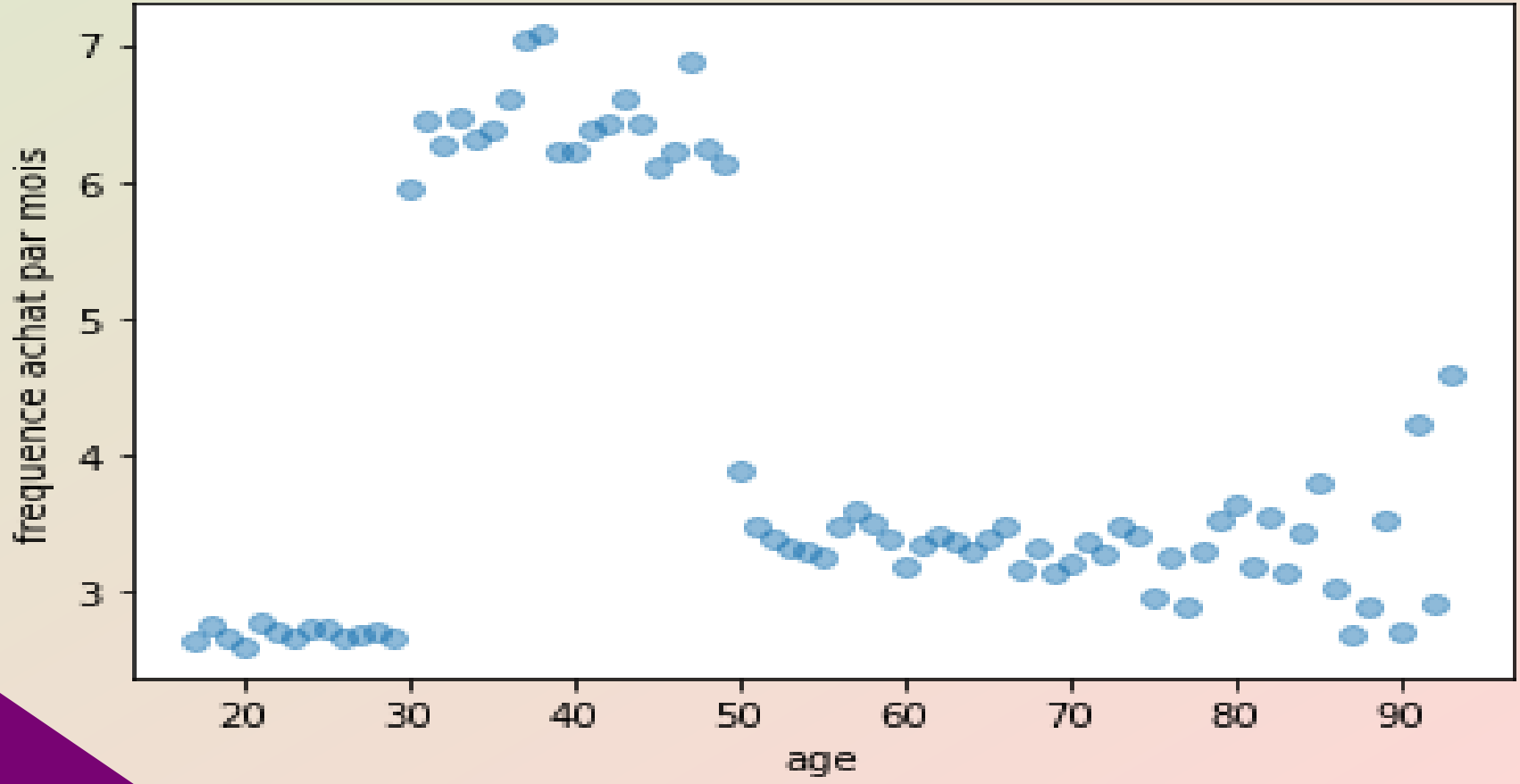
Coefficient de corrélation linéaire : $r_{X,Y}$ =-0,86

Il y’a une forte corrélation linéaire négative entre l’âge et le montant des achats

Par tranche d’âge



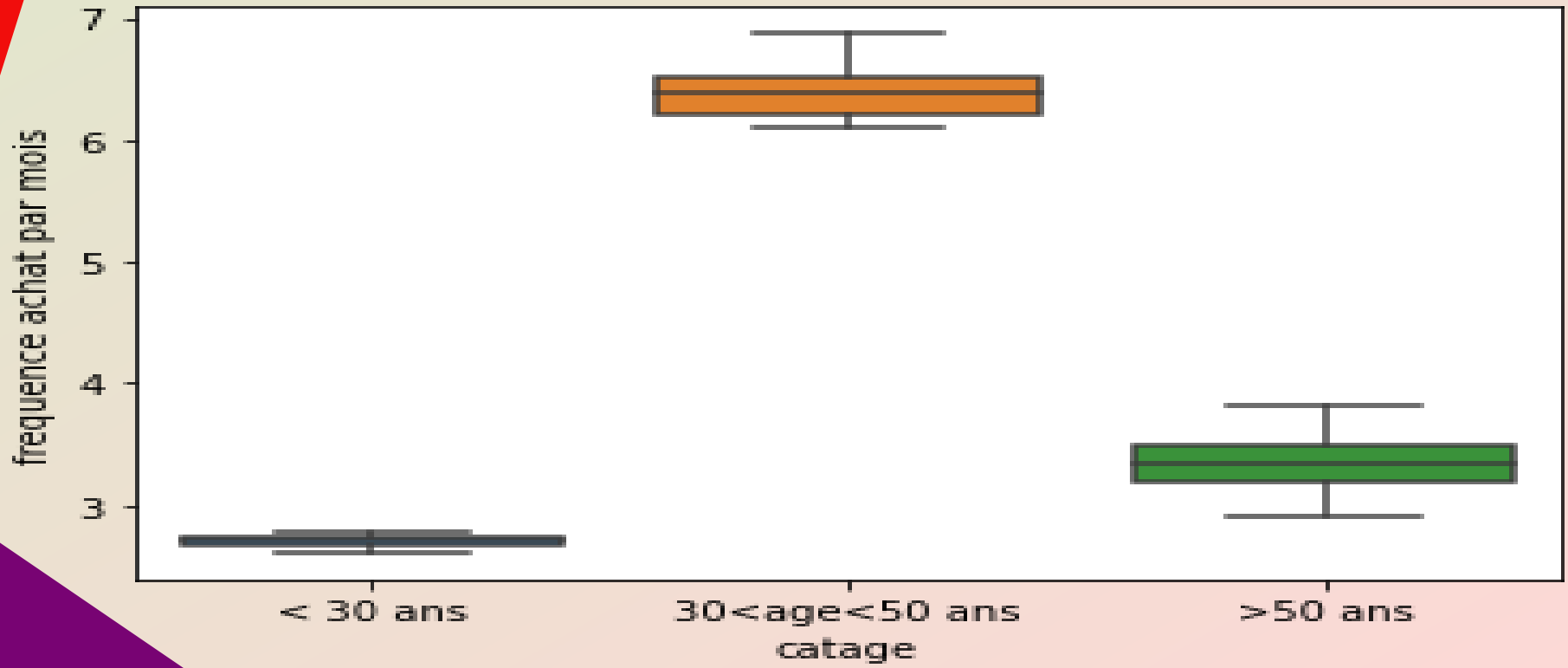
Âge client et fréquence d'achat
Deux variables quantitatives



- **Coefficient de corrélation linéaire : $r_{X,Y}$ = -0,27**

Une faible corrélation linéaire entre l'âge et la fréquence des achats

Par tranche d’age :
Une variable quantitative : Fréquence d’achat
Une variable qualitative : tranche d’age



Analyse de la variance : d'ANOVA

Rapport de corrélation :

$$\eta^2_{Y/X} = \frac{V_{interclasses}}{V_{totale}}$$

Variation totale :

$$SCT = \sum_{j=1}^{n} (y_j - \bar{y})^2$$

Variation interclasse :

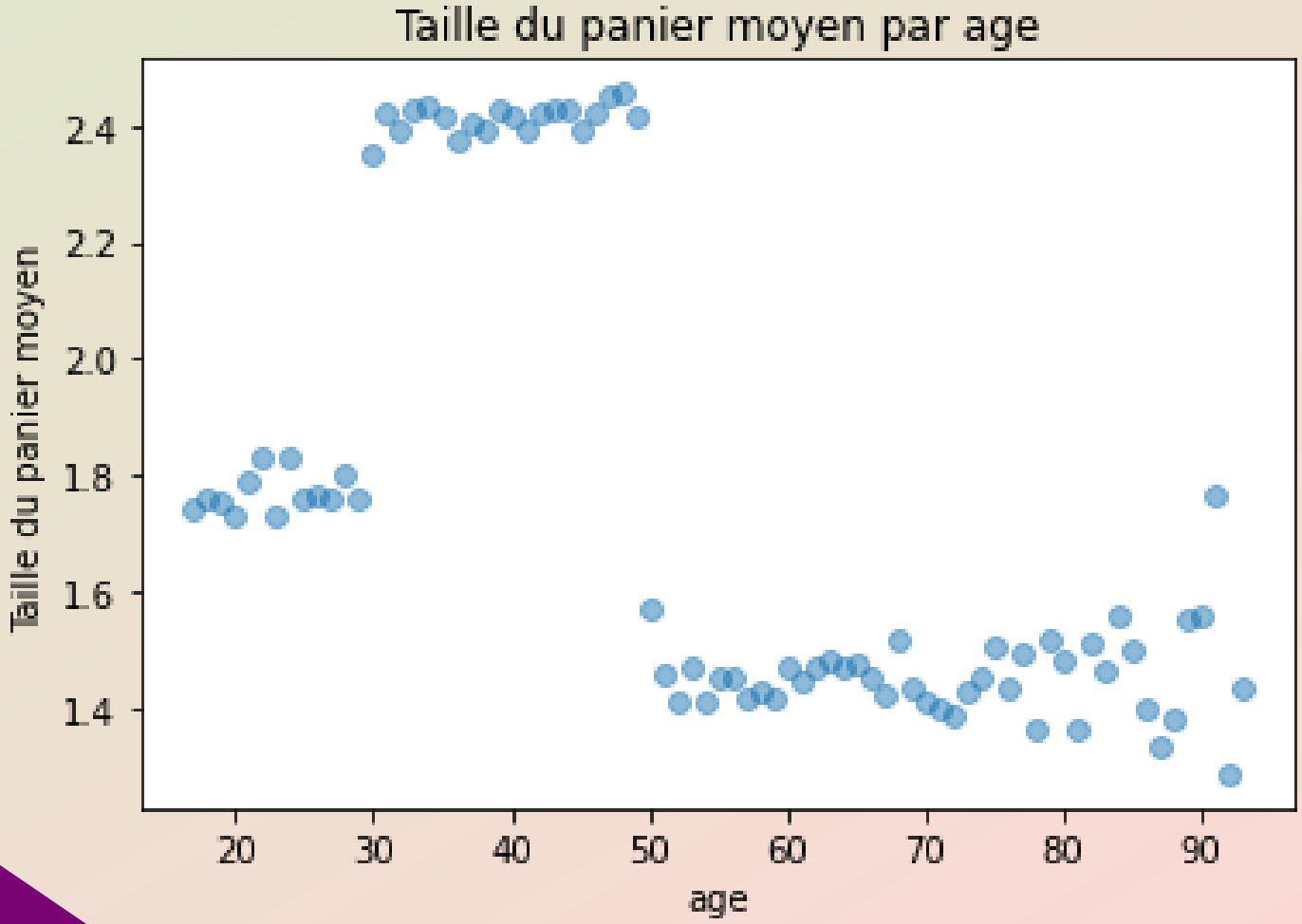
$$\sum_{i=1}^{k} n_i (\overline{y_i} - \bar{y})^2$$

- **Si $\eta^2_{Y/X}$ proche de 0 Il n'y a pas de relation entre les variables X et Y.**
- **Si $\eta^2_{Y/X}$ proche de 1 il y'a une relation entre les variables X et Y**

Rapport de corrélation : $\eta^2_{Y/X}$ =0,86

Par tranche d’âge , il y’a une corrélation avec la fréquence d’achat

Age client et la taille du panier moyen
Deux variables quantitatives



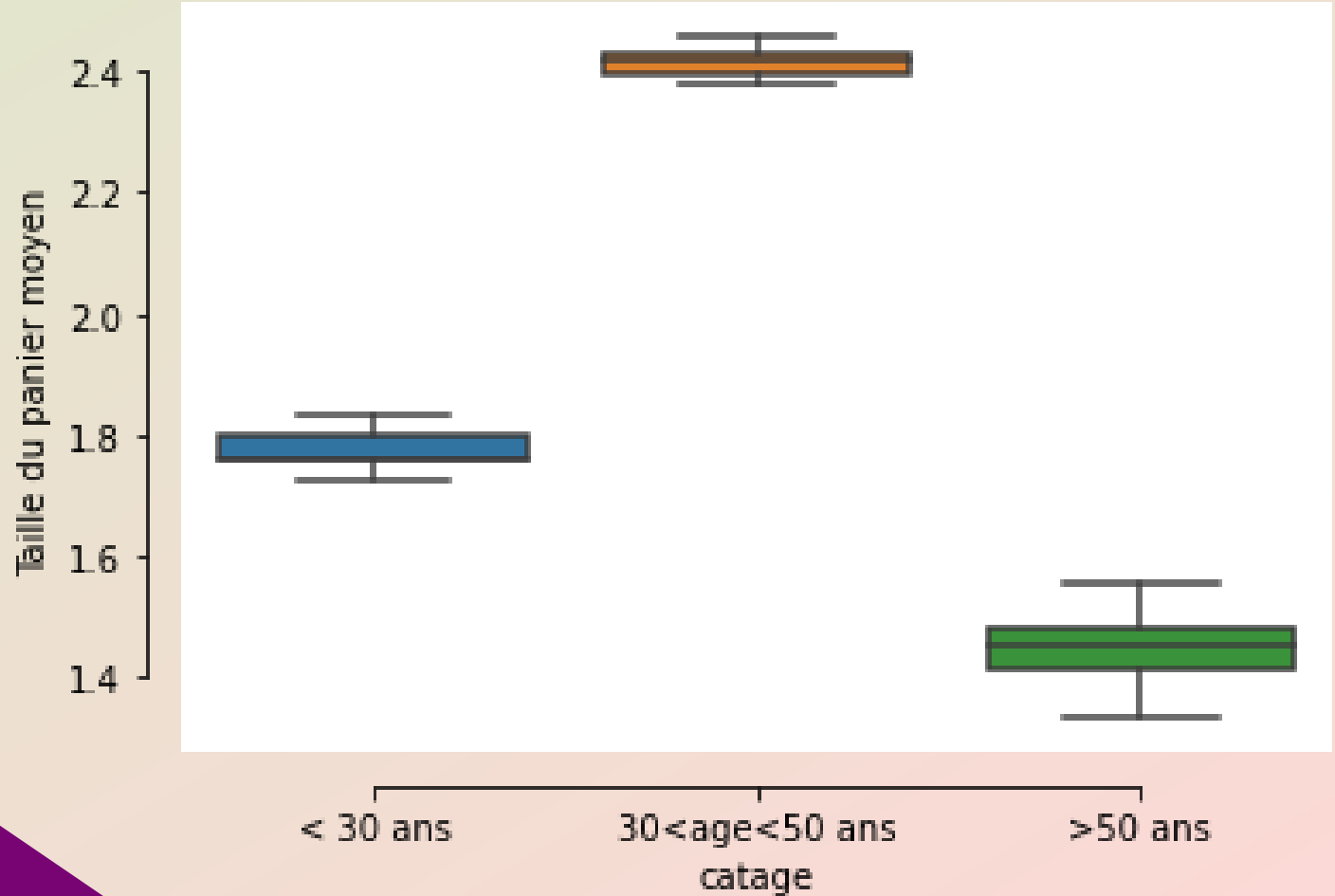
- **Coefficient de corrélation linéaire : $r_{X,Y} = -0,6$**

Les deux variables sont moyennement corrélées linéairement et négativement

Par tranche d’âge

Une variable quantitative : Taille du panier moyen

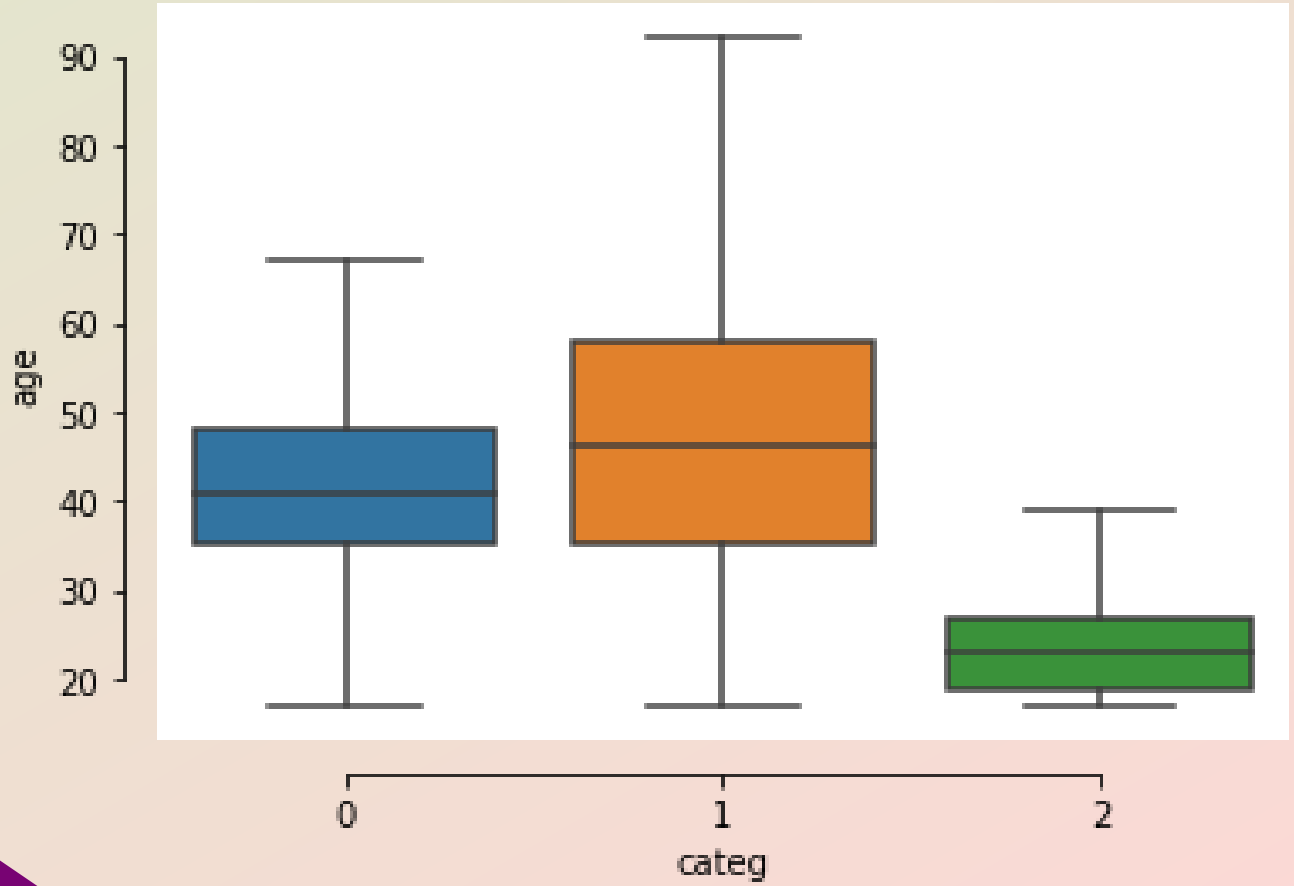
Une variable qualitative : tranche d’âge



Rapport de corrélation : $\eta^2_{Y/X}$ =0,90

Par tranche d’âge et la taille du panier moyen sont fortement corrélés

Âge et catégories produits
Une variable quantitative et une variable qualitative



Rapport de corrélation : $\eta^2_{Y/X}$ =0,11

$\eta^2_{Y/X}$ est très faible , les deux variables ne sont pas corrélées

Sexe et catégorie produits
Deux variables qualitatives Test de CHI2

Tableau de contingence : Chaque case contient un effectif conjoint nij (i ligne, et j colonne)

dans notre cas i: sexe(femme,masculin)

J :catégorie produit(0,1,2)

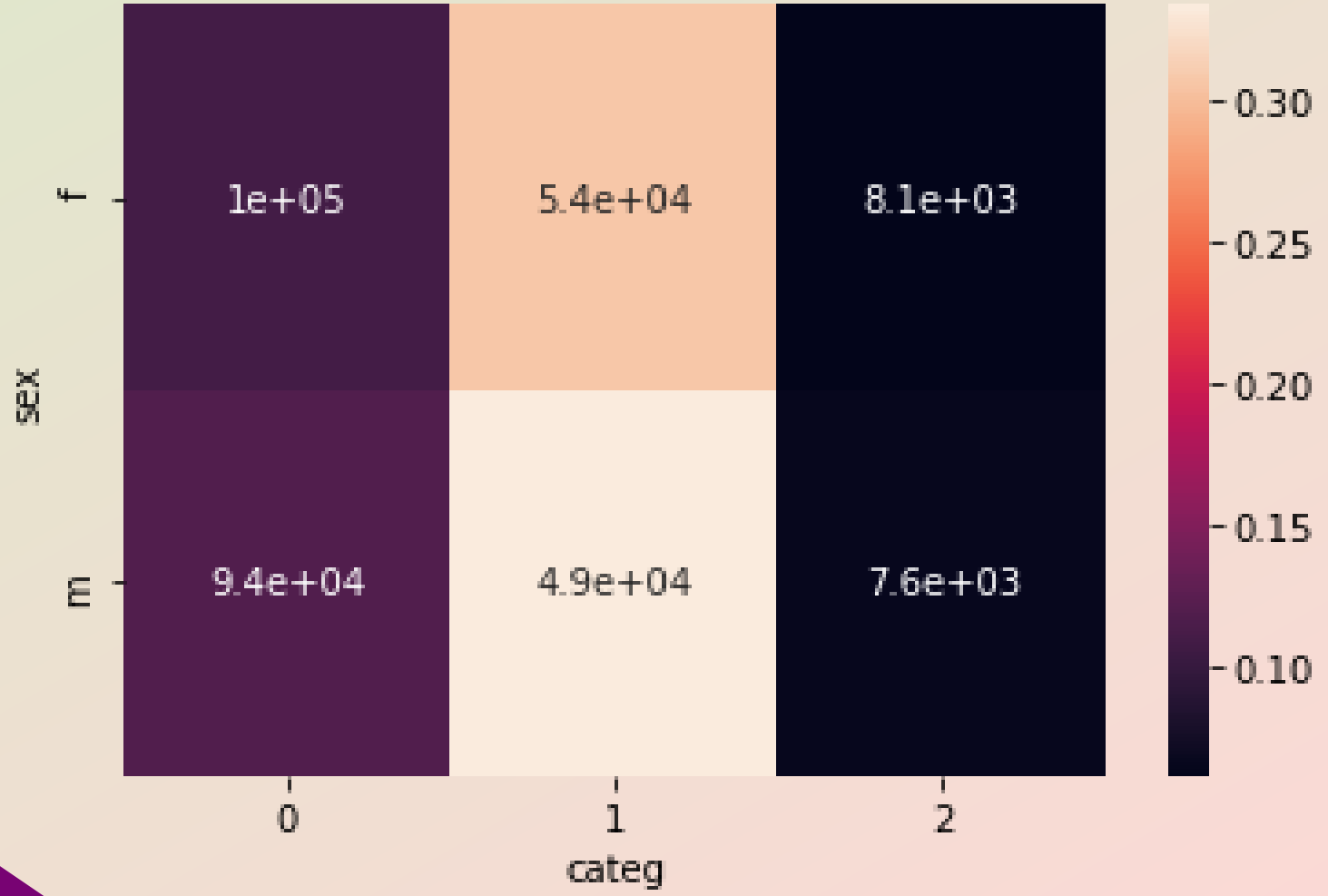
- **à chaque case du tableau un ξij**

$$\xi_{ij} = \frac{(n_{ij} - \frac{n_i.n_j}{n})^2}{\frac{n_i.n_j}{n}}$$

La contribution :ξij /ξn

$$\xi_n = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - \frac{n_i.n_j}{n})^2}{\frac{n_i.n_j}{n}}$$

$\Rightarrow$ **Plus la contribution est grande et plus l'hypothèse d'indépendance est rejetée ;**


sex
f
m
1e+05
5.4e+04
8.1e+03
9.4e+04
4.9e+04
7.6e+03
0
1
2
categ
0.30
0.25
0.20
0.15
0.10

categ sex	0	1	2	Total
f	0.110333	0.308298	0.061358	0.0
m	0.119533	0.334005	0.066474	0.0
Total	0.000000	0.000000	0.000000	0.0

La contribution varie entre 0,06 et 0,33,

elle est plus élevée entre la catégorie1 et masculin

et plus faible entre catégorie 2 et féminin

Conclusion 1/2

Pas de valeurs manquantes dans les tables importées

Des valeurs manquantes par jointure : un produit manquant dans la table des produits

Valeurs aberrantes : prix =-1 dans la table produit
Des données test du logiciel à supprimer

Le nombre de clients :age <55 ans est plus importants

Le chiffre d'affaires augmente : le mois d'octobre à vérifier

⇒Âge et montant des achats corrélation linéaire négative,

⇒Par tranche d'âge, il y'a une corrélation avec la fréquence des achats,

⇒Âge et la taille du panier moyen sont moyennement corrélées linéairement et négativement,

⇒Par tranche d'âge et la taille du panier moyen sont fortement corrélés,

⇒Âge et catégories produits ne sont pas corrélés,

⇒Catégories produits et sexes des clients très faiblement corrélés