



Librairie en ligne

Rester livres

Glossaire

categ : catégorie de produits regroupement selon la même classe tarifaire

client_id : id du client (un seul id par client)

id : identifiant

prod_id : id du produit (le premier chiffre correspondant à la catégorie)

session_id : id de la connexion (un ou plusieurs achats possible pour une session_id)

Discrétisation : découpage en classes

éta² : mesure l'intensité de la liaison entre une variable quantitative et une variable qualitative

Implémenter : deviner une valeur manquante

Indice de Gini : mesure le niveau d'inégalité de la répartition d'une variable dans la population

Null : pointeur sans cible ou variable sans valeur

p_value : plus petite valeur d'alpha pour que le test statistique montre une différence significative

R² : coefficient de détermination

Présentation

Rester livres

Grande **chaîne de librairie** qui s'est d'abord développée dans une grande ville de France, avec **plusieurs magasins**.

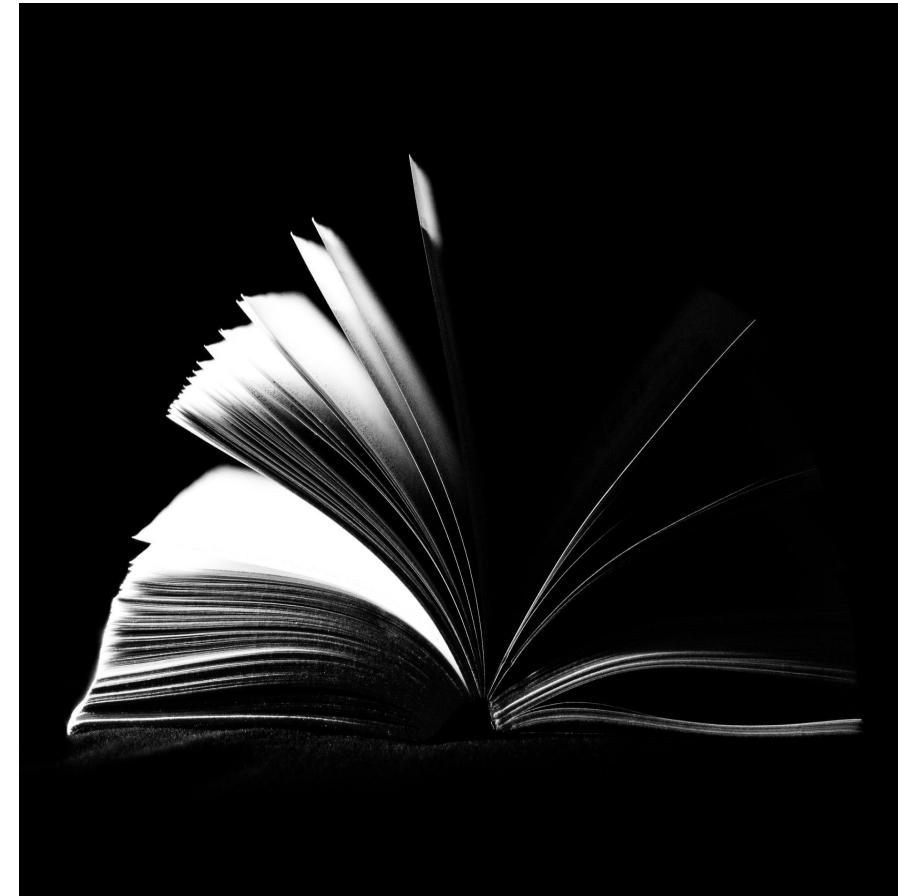
“**Rester livres**” a décidé d'ouvrir une boutique en ligne et d'avoir recours à des **algorithmes de recommandation**.

Sommaire

1- Nettoyage des données

2- Analyse des données

3- Réponses au manager



Nettoyage de données



Les trois fichiers sont propres

	client_id	sex	birth
0	c_4410	f	1967
1	c_7839	f	1975

8623 L x 3 C

	id_prod	price	categ
0	0_1421	19.99	0
1	0_1368	5.13	0

3287 L x 3 C

	id_prod	date	session_id	client_id
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450
1	2_226	2022-02-03 01:55:53.276402	s_159142	c_277

337016 L x 4 C

Importation et
lecture du fichier

```
df_fichier = pd.read_csv(fichier.csv,  
encoding=ENCODAGE)
```

Aucune valeur
dupliquée

```
df_fichier.duplicated(subset=  
['client_id']).sum()
```

Aucune valeur null

```
df_fichier.isnull().sum().sort_value  
s(ascending=False)
```

Fichier transactions

	id_prod	date	session_id	client_id
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450
1	2_226	2022-02-03 01:55:53.276402	s_159142	c_277

Suppression des valeurs test

	id_prod	date	session_id	client_id
57755	T_0	test_2021-03-01 02:30:02.237413	s_0	ct_1
59043	T_0	test_2021-03-01 02:30:02.237413	s_0	ct_0

```
df_transactions.drop(df_transactions[df_transactions['id_prod'] == 'T_0'].index, axis=0)
```

Conversion de la colonne date
au format “datetime”

```
df_transactionsSansTest['date'] =  
pd.to_datetime(df_transactionsSansTest['date'])
```

```
id_prod          object  
date    datetime64[ns]  
session_id      object  
client_id      object  
dtype: object
```

Export du fichier transactions

```
df_transactionsSansTest.to_csv("transactions_clean  
.csv", index = False, encoding='UTF-8')
```

Fichier clients

	client_id	sex	birth
0	c_4410	f	1967
1	c_7839	f	1975

Valeurs distinctes

```
df_customers.nunique()
```

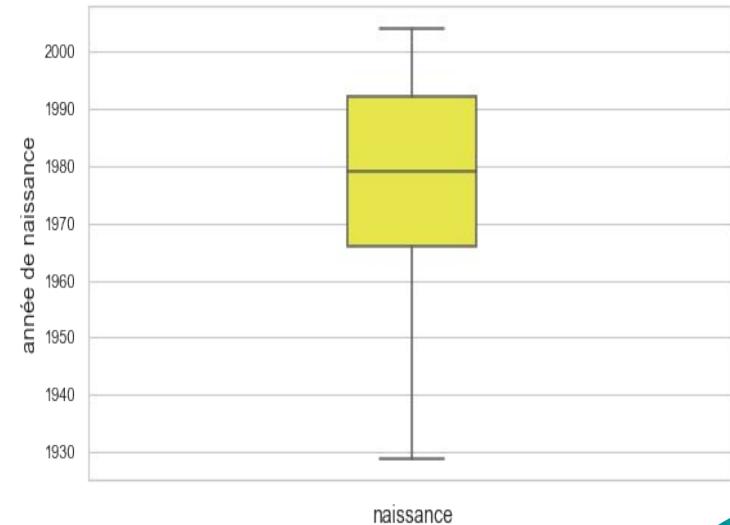
```
client_id      8623  
sex             2  
birth           76  
dtype: int64
```

	birth
count	8623.000000
mean	1978.280877
std	16.919535
min	1929.000000
25%	1966.000000
50%	1979.000000
75%	1992.000000
max	2004.000000

```
df_products.describe()
```

Les années de naissance se distribuent entre 1929 et 2004. **Aucune valeur aberrante.**

Répartition des années de naissance des clients



Modification du fichier clients

Suppression des valeurs test (ct_0, ct_1)

```
df_customers.drop(df_customers[(df_customers['client_id']=='ct_0')].index, axis=0)
```

Changement de casse de la colonne "sex"

```
df_customersSansTest['sex'] =  
df_customersSansTest['sex'].str.upper()
```

Nommage des colonnes

```
df_customersSansTest.rename({'sex' : 'sex',  
'client_id':'client_id', 'birth' : 'year_of_birth'}, axis=1)
```

Exportation du fichier

```
df_customersSansTest.to_csv("customers_clean.csv",  
index = False, encoding='UTF-8')
```

	client_id	sex	year_of_birth
0	c_4410	F	1967
1	c_7839	F	1975

Fichier products

	id_prod	price	categ
0	0_1421	19.99	0
1	0_1368	5.13	0

Valeurs distinctes

```
df_products.nunique()
```

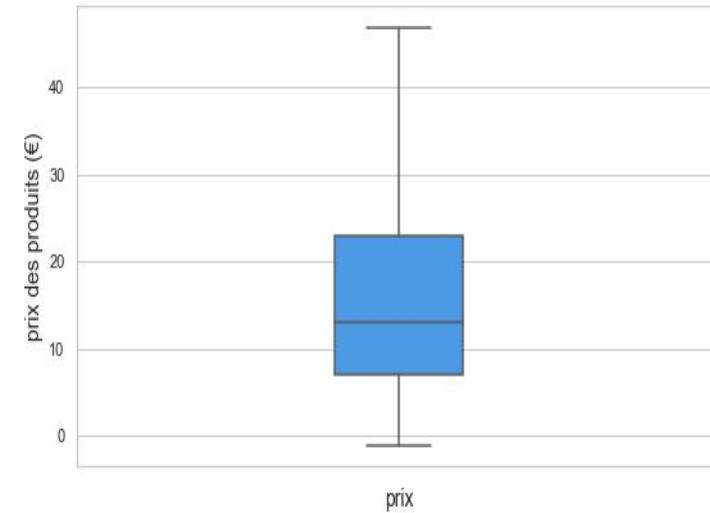
```
id_prod      3287  
price       1455  
categ         3  
dtype: int64
```

	price	categ
count	3287.000000	3287.000000
mean	21.856641	0.370246
std	29.847908	0.615387
min	-1.000000	0.000000
25%	6.990000	0.000000
50%	13.060000	0.000000
75%	22.990000	1.000000
max	300.000000	2.000000

```
df_products.describe()
```

Masquage des outliers. Existence de prix négatifs
(valeur aberrante)

Répartition des prix des produits



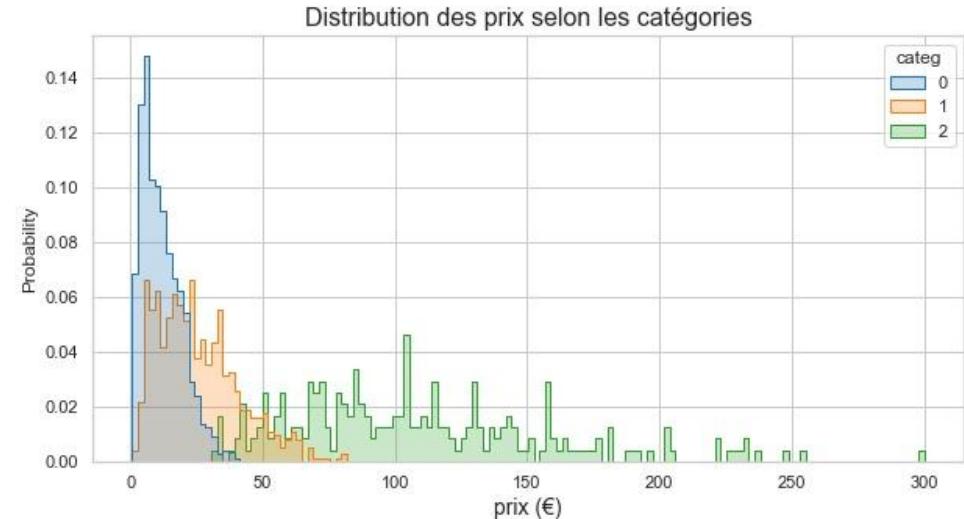
Fichier products

Suppression des valeurs test (T_0) qui correspondent à des prix **négatifs**

```
df_products[~df_products.id_prod.isin(['T_0'])]
```

	<u>id_prod</u>	<u>price</u>	<u>categ</u>
731	T_0	-1.0	0

Les **catégories** regroupent des produits de **même classe tarifaire**.



Eta² = 0.69. Corrélation entre les prix et les catégories

Catégorie et distribution des prix

Catégorie 0

	price	categ
count	2308.000000	2308.0
mean	11.732795	0.0
std	7.565755	0.0
min	0.620000	0.0
25%	5.587500	0.0
50%	10.320000	0.0
75%	16.655000	0.0
max	40.990000	0.0

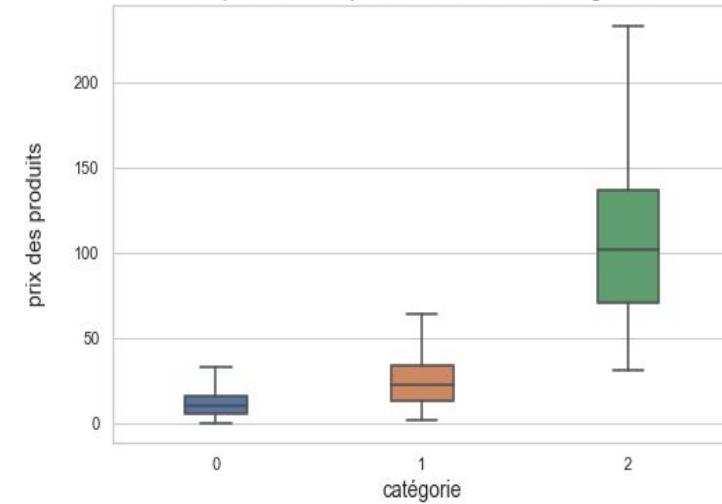
Catégorie 1

	price	categ
count	739.000000	739.0
mean	25.531421	1.0
std	15.425162	0.0
min	2.000000	1.0
25%	13.390000	1.0
50%	22.990000	1.0
75%	33.990000	1.0
max	80.990000	1.0

Catégorie 2

	price	categ
count	239.000000	239.0
mean	108.354686	2.0
std	49.561431	0.0
min	30.990000	2.0
25%	71.065000	2.0
50%	101.990000	2.0
75%	136.530000	2.0
max	300.000000	2.0

Répartition des prix en fonction des catégories



Un produit non référencé

Jointure de “transactions” avec
“products”

```
pd.merge(df_transactionsSansTest, df_products,  
on='id_prod', how='left')
```

Recherche valeur null

```
df_transSansTestProduct.isnull().sum()
```

**103 valeurs null : 103 transactions pointent sur un
produit non référencé dans le fichier “products”**

Le produit “**0_2245**” n'est **pas référencé**
dans le fichier “product” (présence des
NaN)

	id_prod	date	session_id	client_id	price_euros	categ
6231	0_2245	2021-06-17 03:03:12.668129	s_49705	c_1533	NaN	NaN
10797	0_2245	2021-06-16 05:53:01.627491	s_49323	c_7954	NaN	NaN

Mais il est toujours au catalogue !!

```
Entrée [79]: 1 df_transSansTestProduct.date.min()  
Out[79]: Timestamp('2021-03-01 00:09:29.301897')  
  
Entrée [80]: 1 df_transSansTestProduct.date.max()  
Out[80]: Timestamp('2022-02-28 18:08:49.875709')
```

Imputation de la valeur du produit "0_2245"

Imputation en prenant la **moyenne** de la valeur du prix des produits de la **catégorie 0**

```
11.73= df_prodCatZero.price_euros.mean()
```

Intégration du produit "0_2245" dans le fichier "products"

```
df_products = df_products.append({'id_prod' : '0_2245', 'price_euros' : 11.73, 'categ' : 0 }, ignore_index=True)
```

Nommage des colonnes

```
df_products.rename({'id_prod':'id_prod', 'price' : 'price_euros', 'categ' : 'categ'}, axis=1)
```

Exportation du fichier

```
df_products.to_csv("products_clean.csv", index = False, encoding='UTF-8')
```

id_prod	price_euros	categ
3286	0_2245	11.73

fichier “products” et “transactions”

Jointure droite de “df_products”
avec “df_transactions”

(pour récupérer TOUTES les TRANSACTIONS)

```
pd.merge(df_products, df_transactions,  
on='id_prod', how='right')
```

Ajout des colonnes

annee : df_transactions['annee'] =
df_transactions.date.dt.year

mois : df_transactions['mois']=
df_transactions.date.dt.month

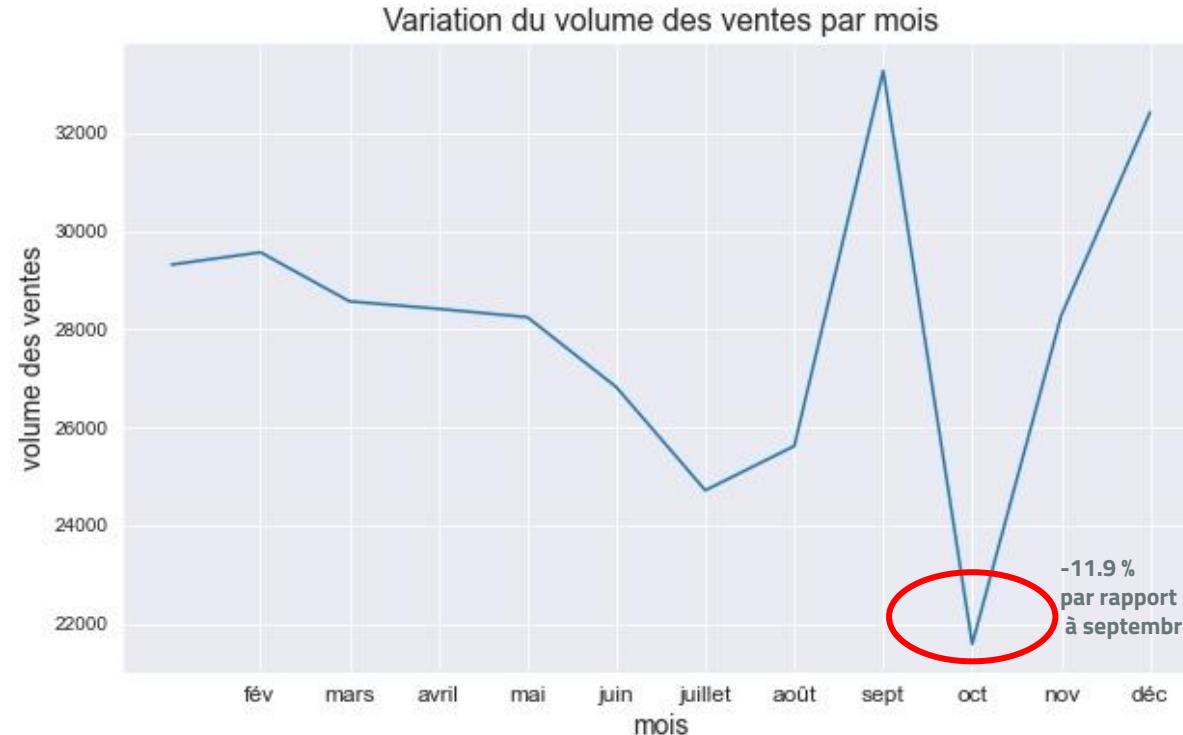
jour : df_transactions['jour'] =
df_transactions.date.dt.day

	id_prod	price_euros	categ		date	session_id	client_id	annee	mois	jour
0	0_1483	4.99	0	2021-04-10 18:37:28.723910	s_18746	c_4450	2021	4	10	
1	0_1483	4.99	0	2021-10-18 19:16:14.767807	s_106741	c_1576	2021	10	18	

A photograph of a person in a dynamic pose on a snowy mountain peak. They are wearing dark ski pants and a dark jacket, and are holding two skis above their head with their arms bent. The background shows a vast, snow-covered mountain range under a clear blue sky.

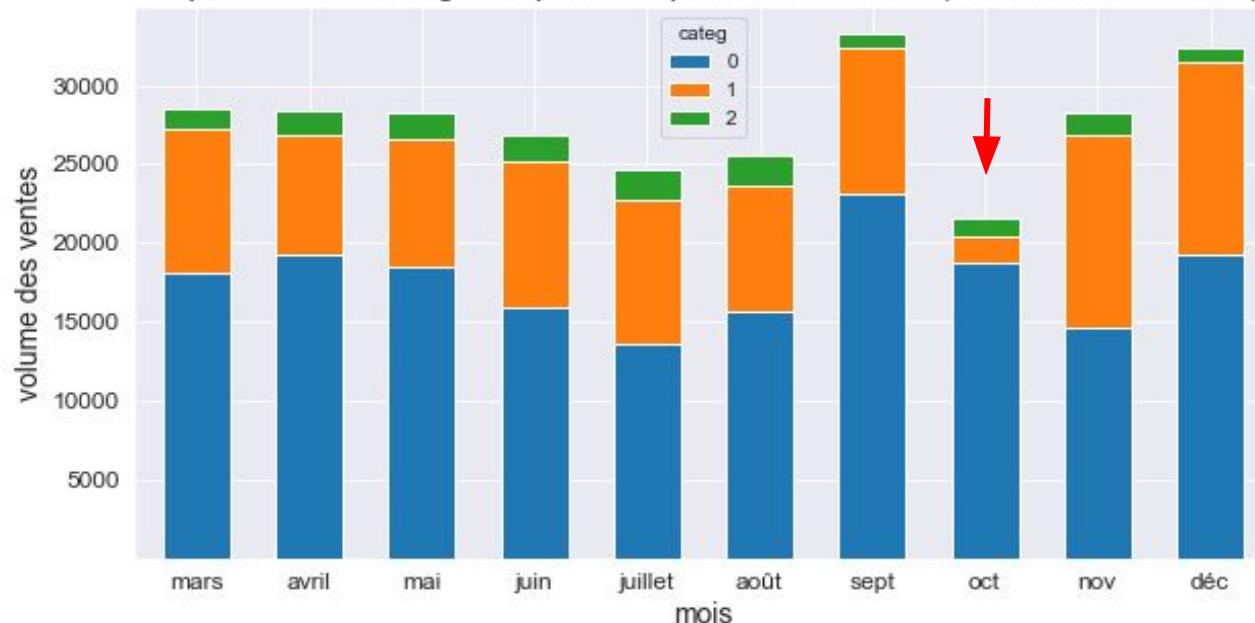
Chute des ventes du mois d'octobre

Forte baisse du volume des ventes en octobre

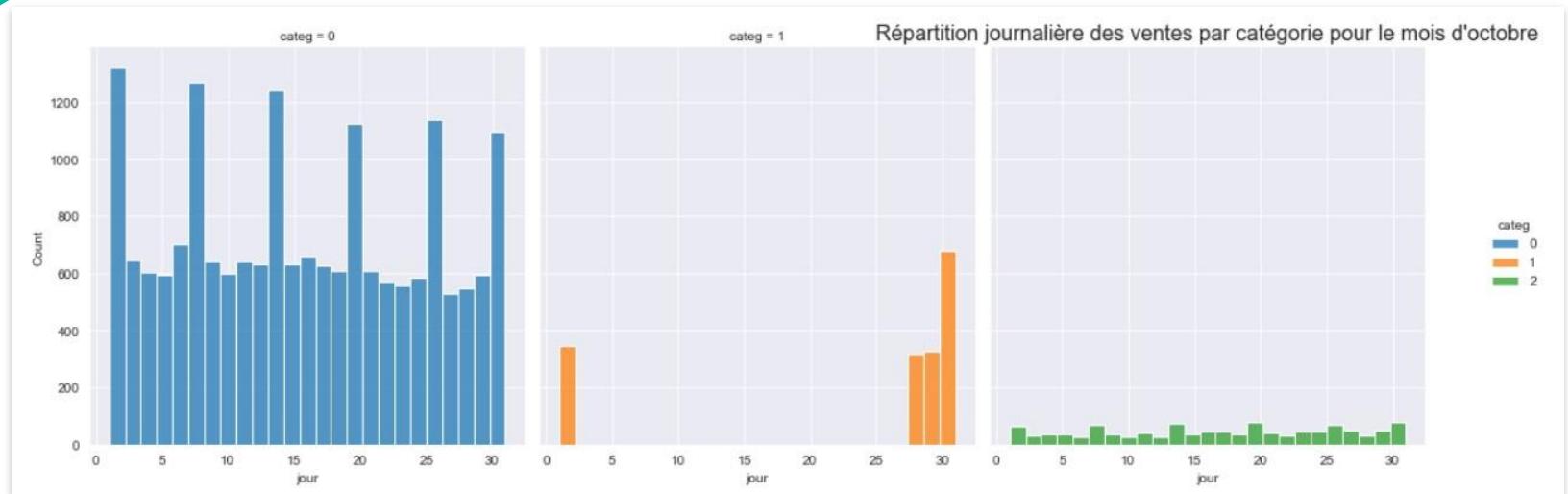


Forte diminution du volume de vente de la catégorie 1

Répartition des catégories par mois pour l'année 2021 (de mars à décembre)



Du 02 au 28 octobre absence de données de la catégorie 1



Rupture de stock ou problème d'enregistrement des données ?

Imputation ou suppression d'octobre ?



6 %

du chiffre d'affaire annuel de "Rester livres"

6.4 %

des transactions annuelles

8.3 %

chiffre d'affaire moyen par mois



Suppression d'octobre

Analyse de données

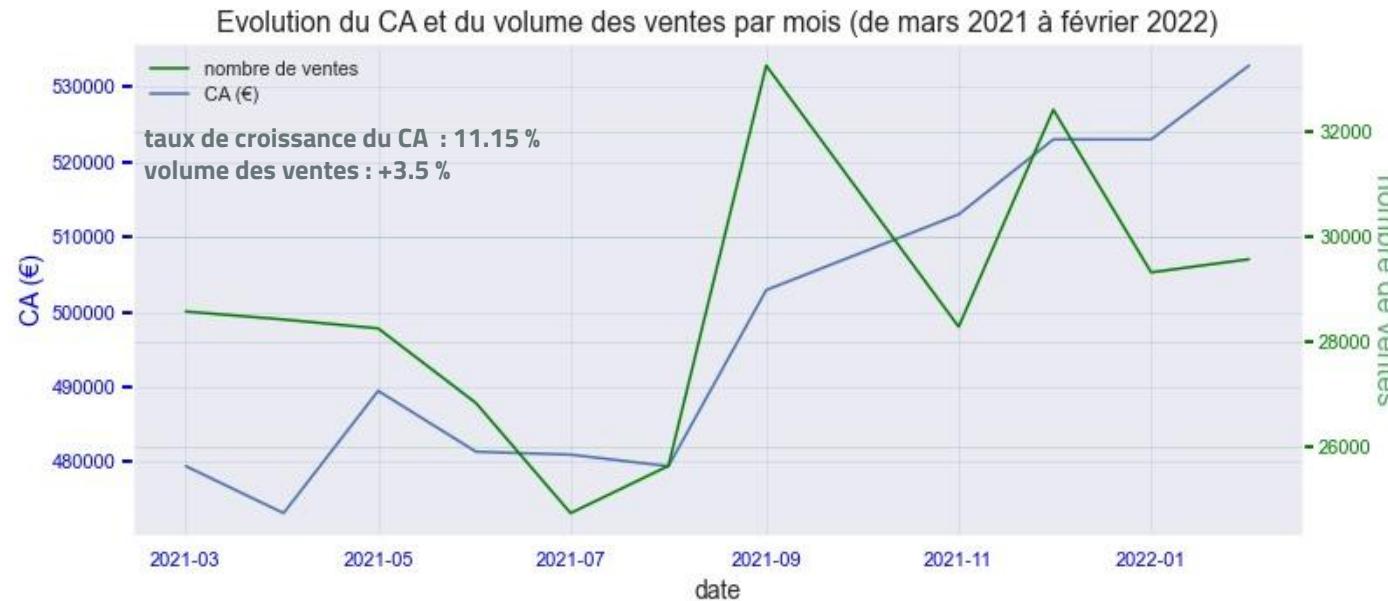




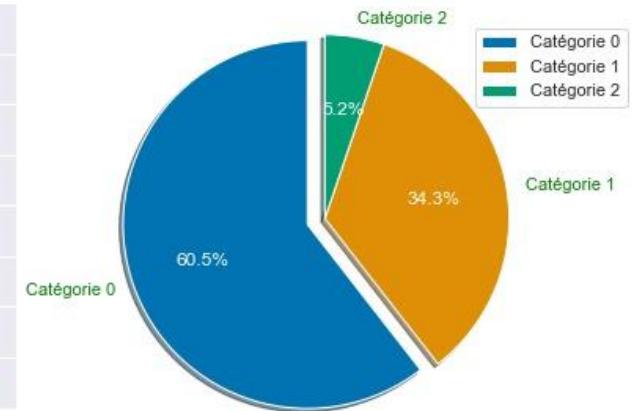
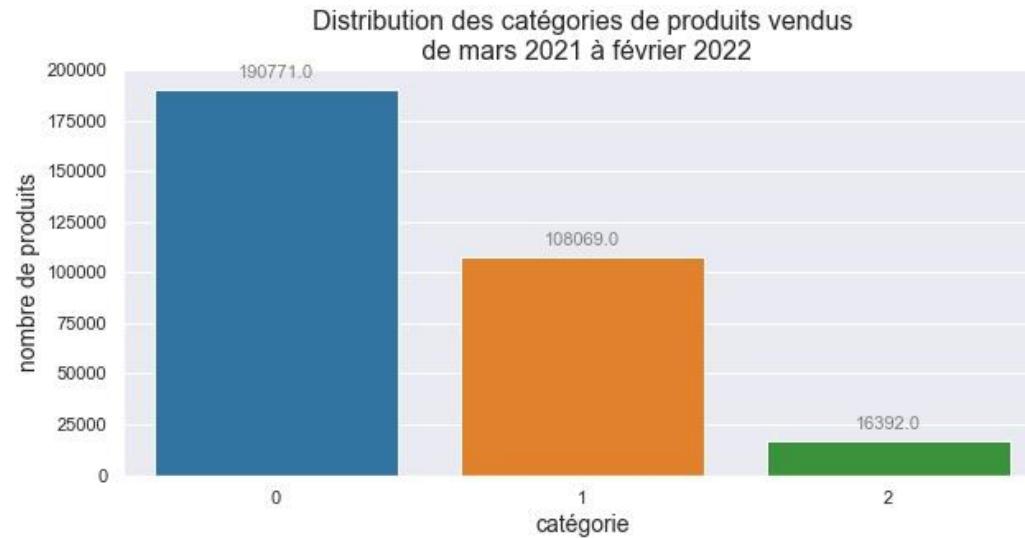
Etude des ventes



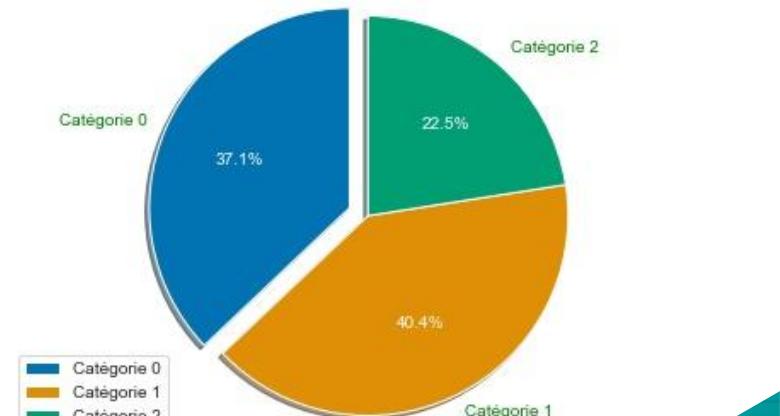
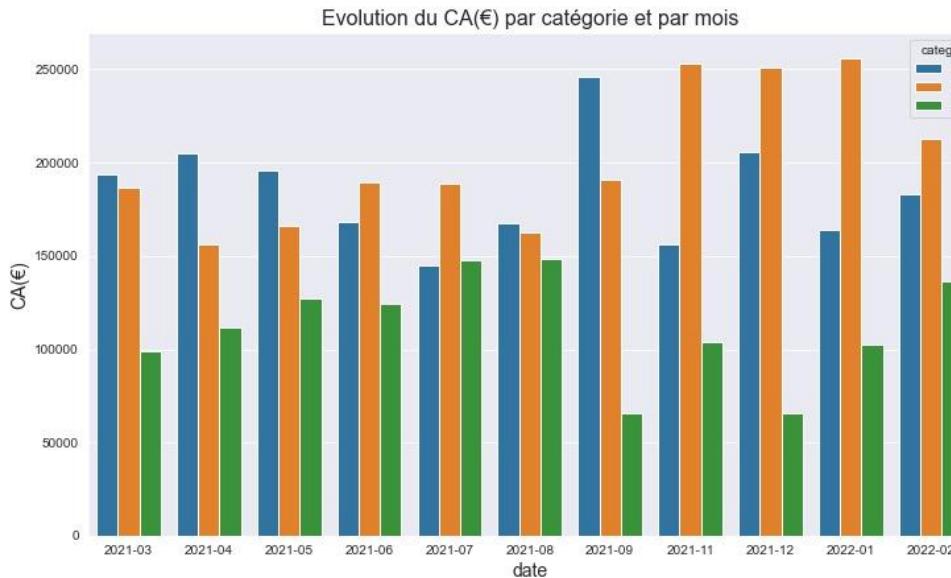
Augmentation du chiffre d'affaire et du volume des ventes



Catégorie 0, la plus vendue

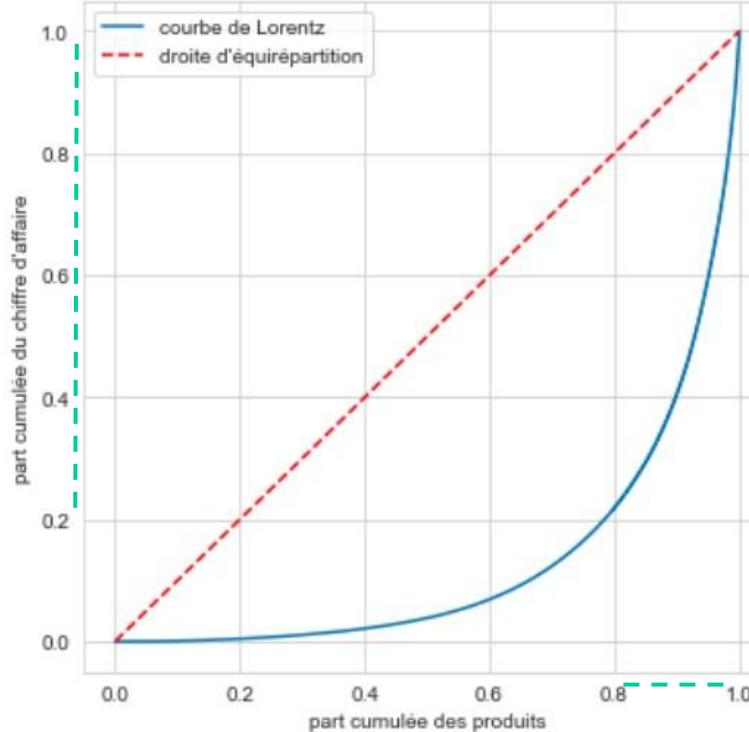


Catégorie 1, le plus haut CA



Répartition inégalitaire entre le CA et les produits

Distribution du chiffre d'affaire en fonction des produits



Indice de gini : 0.74

20 % des produits
contribuent
à 80 % du CA

Modeste corrélation entre le nombre et le montant des achats par session

Répartition du chiffre d'affaire (€) par session en fonction du nombre d'achats



var ou ecart type

Stabilité du panier moyen mensuel

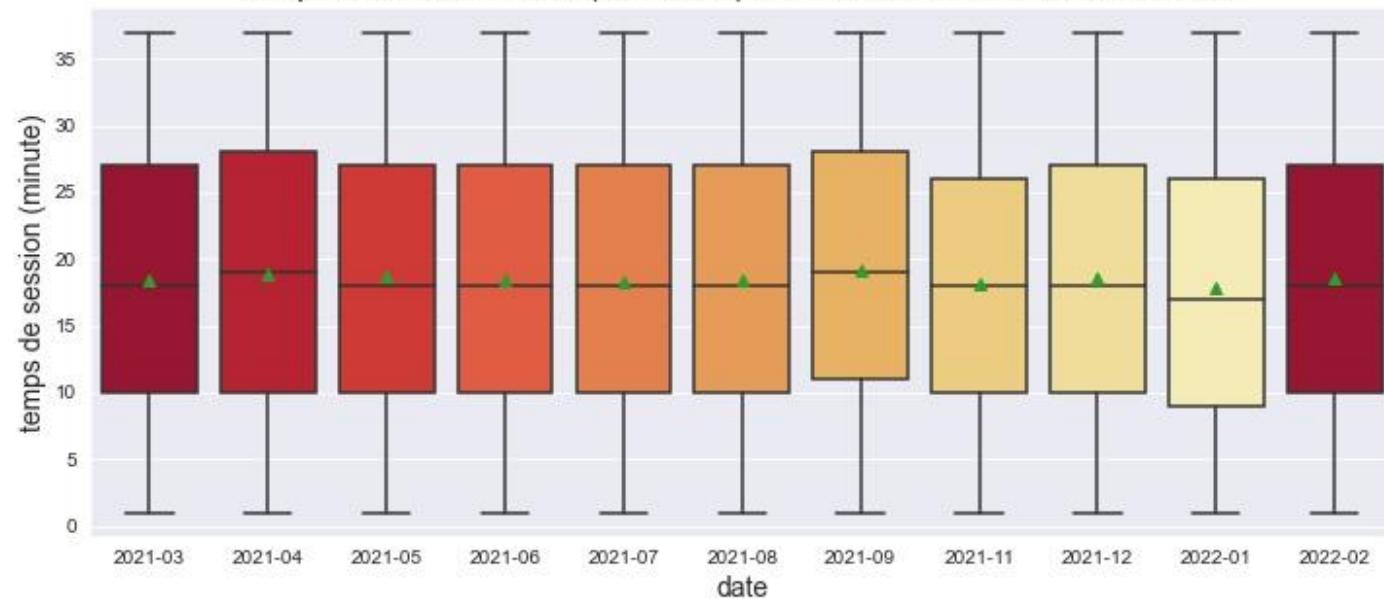




Rester livres, un site internet...

Temps élevé de consultation mensuel

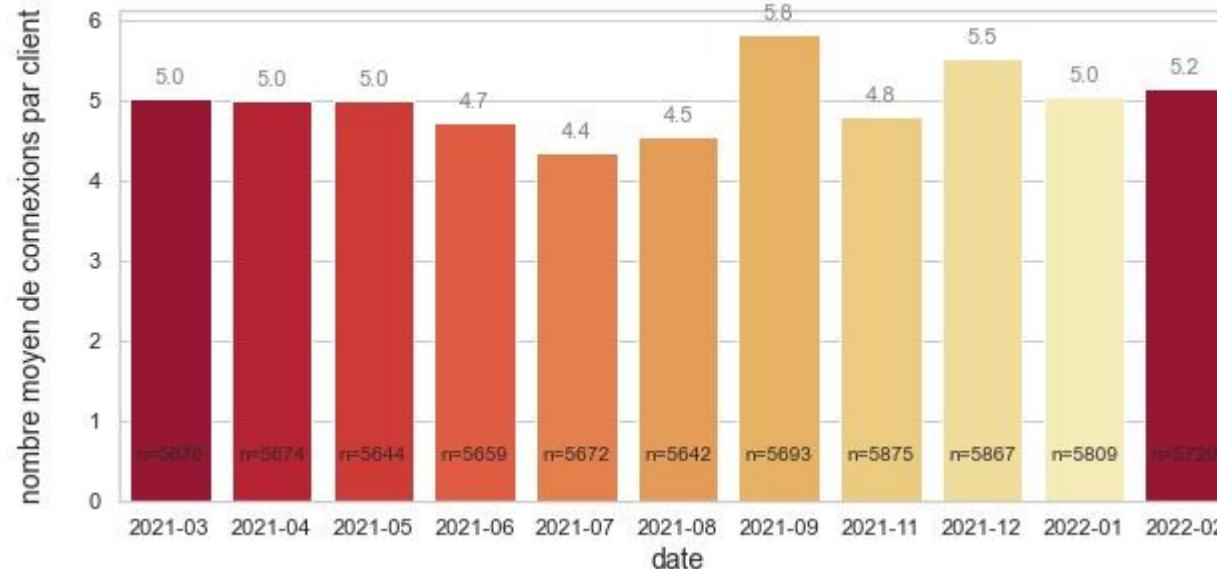
Temps de visite mensuel (en minute) du site de mars 2021 à février 2022



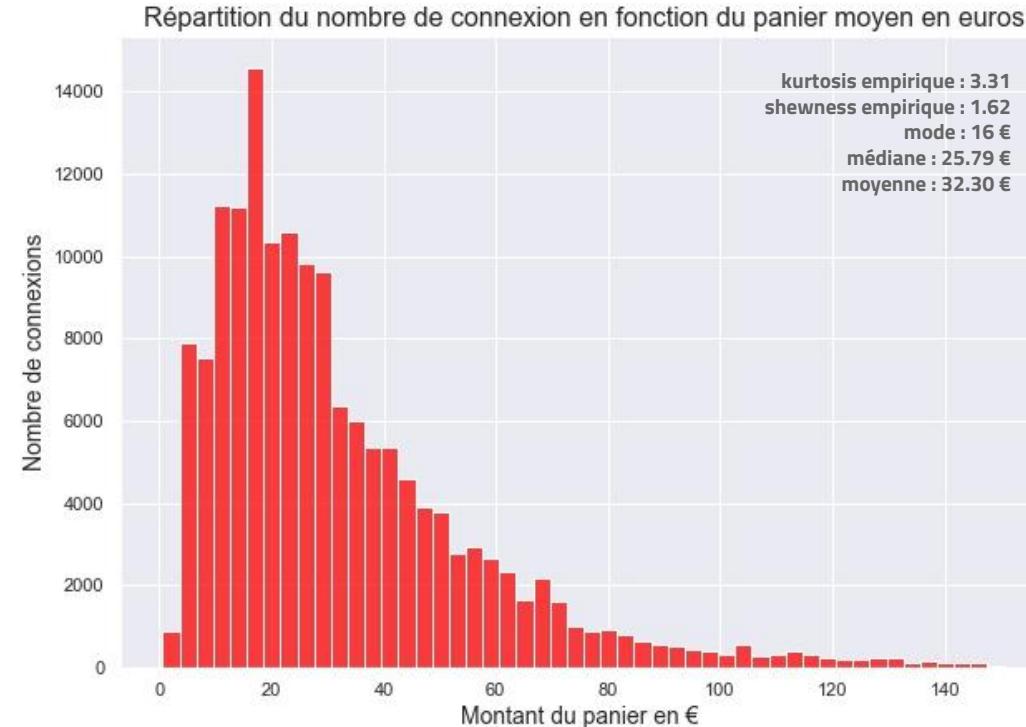
corr entre date et nb de connexion

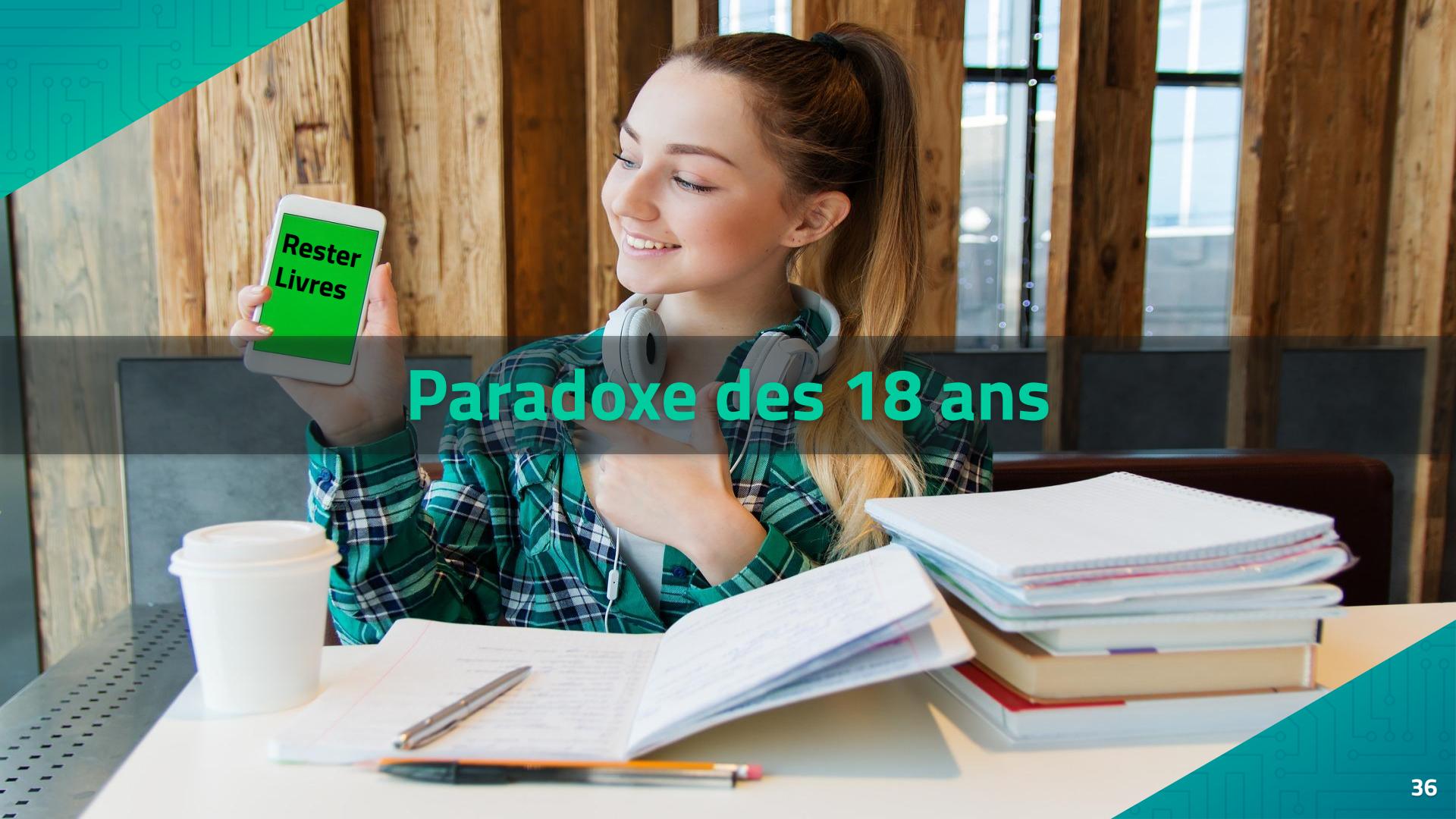
Des clients fidèles (5 connexions mensuelles en moyenne)

Nombre moyen de connexions mensuelles par client



Connexions à 16 €



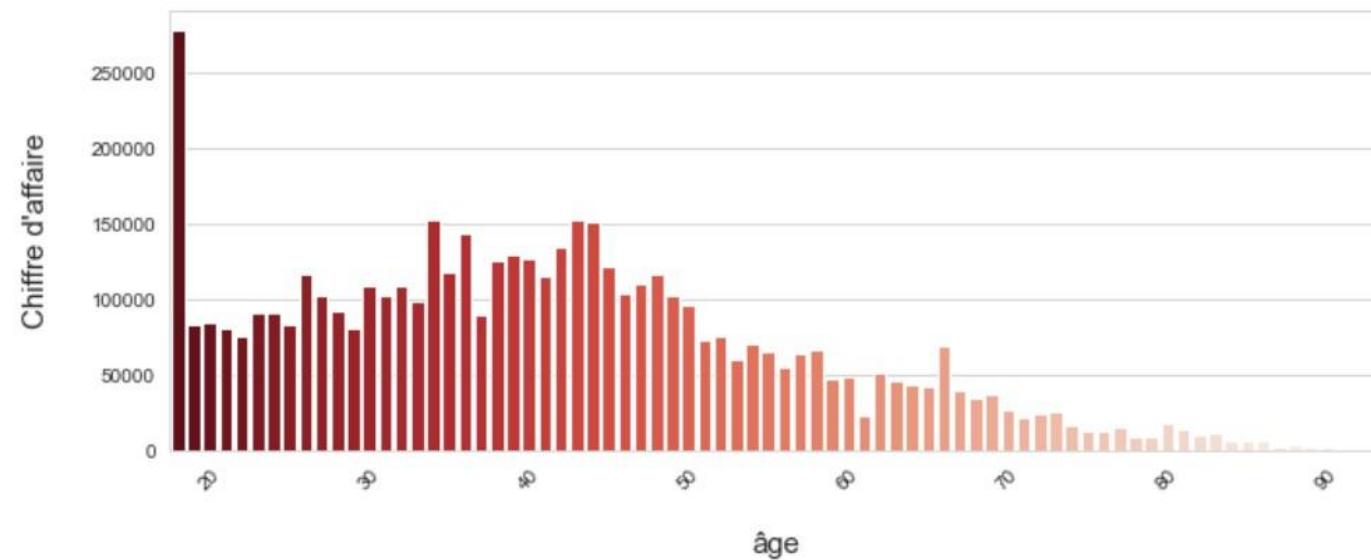


Paradoxe des 18 ans



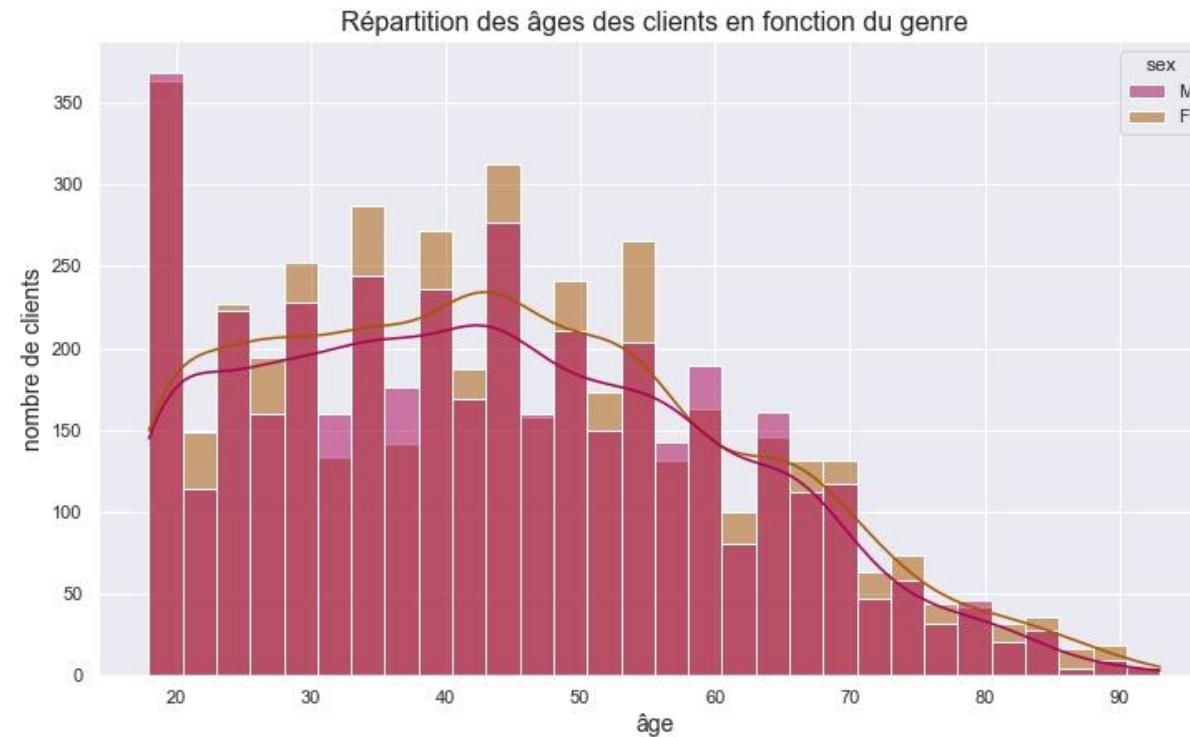
18 ans, l'âge le plus “acheteur” ?

Chiffre d'affaire en fonction de l'âge du client (de mars 2021 à février 2022)



Des clients jeunes

(compte de mineurs ou âge par défaut sur le site ?)

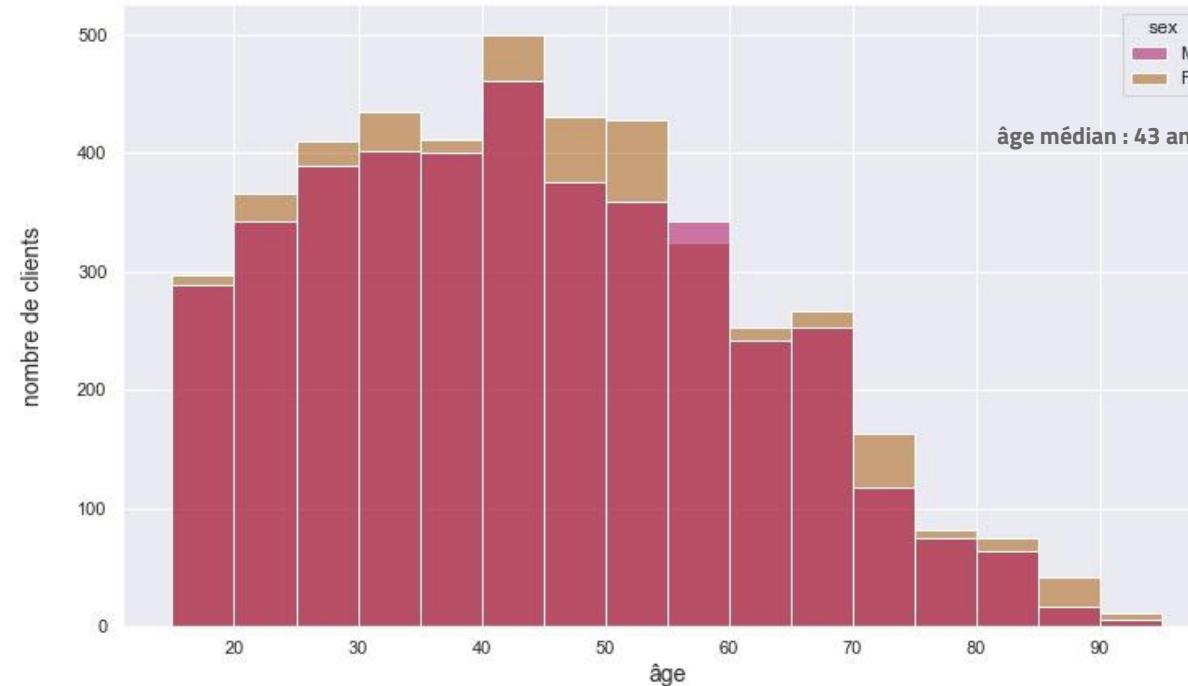




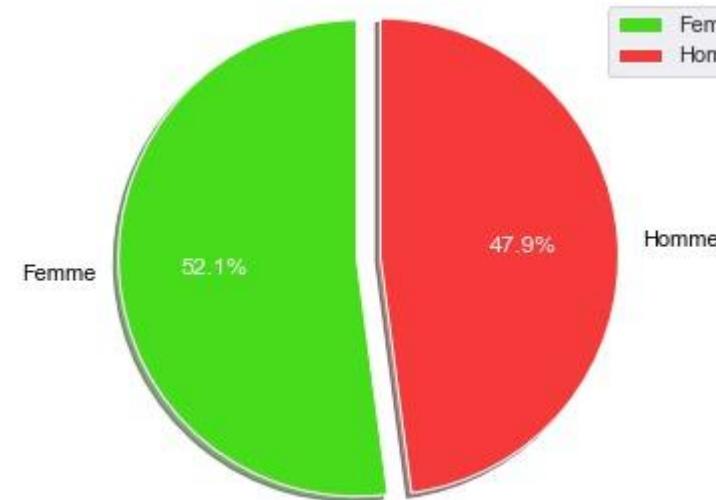
Profil des clients

Les 35 et 45 ans les plus actifs

Répartition des clients par classes d'âge et par genre



Une faible majorité de femmes



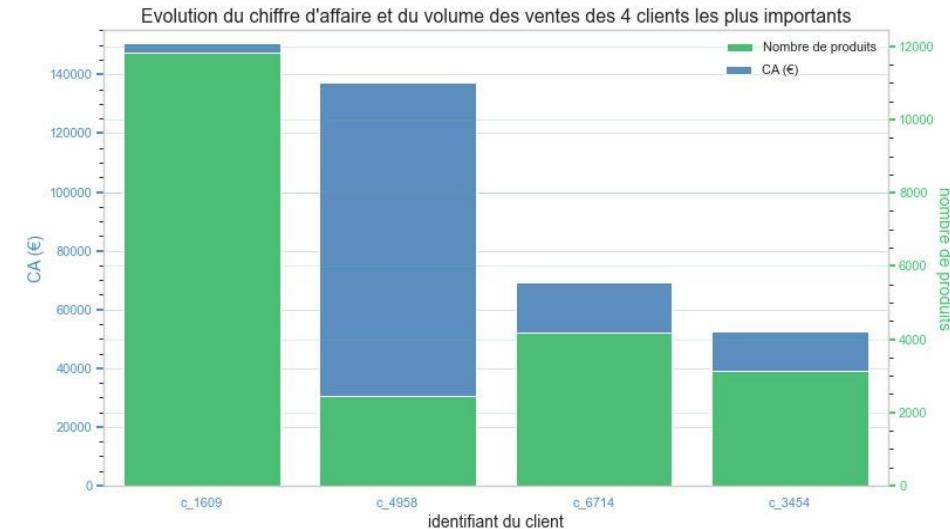
4 principaux clients

Ils représentent **7.48%** du chiffre d'affaire annuel (indice de Gini : 0.44)

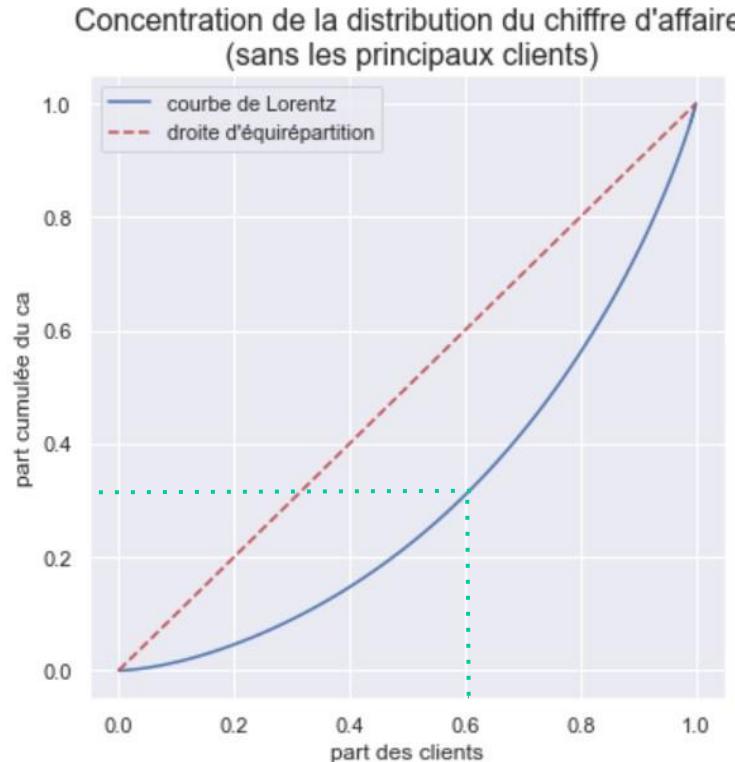
client_id	sex	age	montant_achat_total	nb_produits
0	c_1609	M	42	150729.07
1	c_4958	M	23	137151.48
2	c_6714	F	54	69405.63
3	c_3454	M	53	52744.14

Volume de vente de c_3454 : 17 fois plus élevé que le 5ème client

Suppression des ces **4 outliers** pour une étude plus fine des ventes et du CA.



Répartition inégalitaire entre le CA et les clients



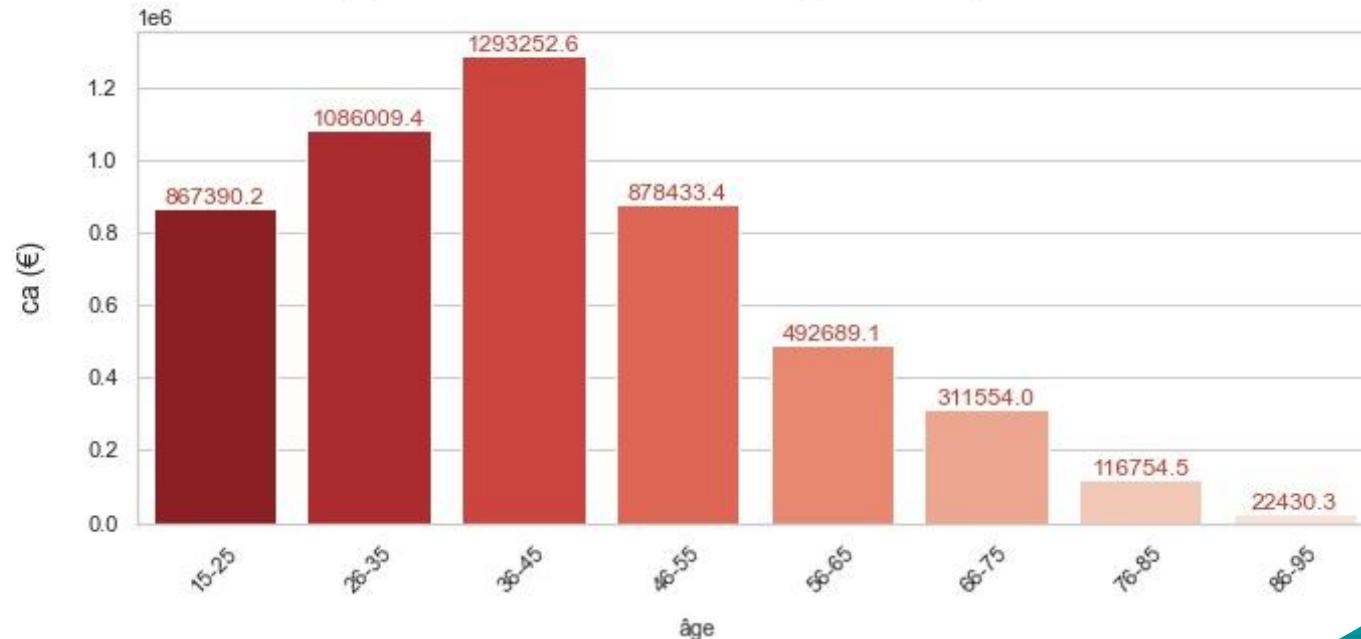
indice de Gini : **0.396**

60% des clients
participent pour

30% du CA

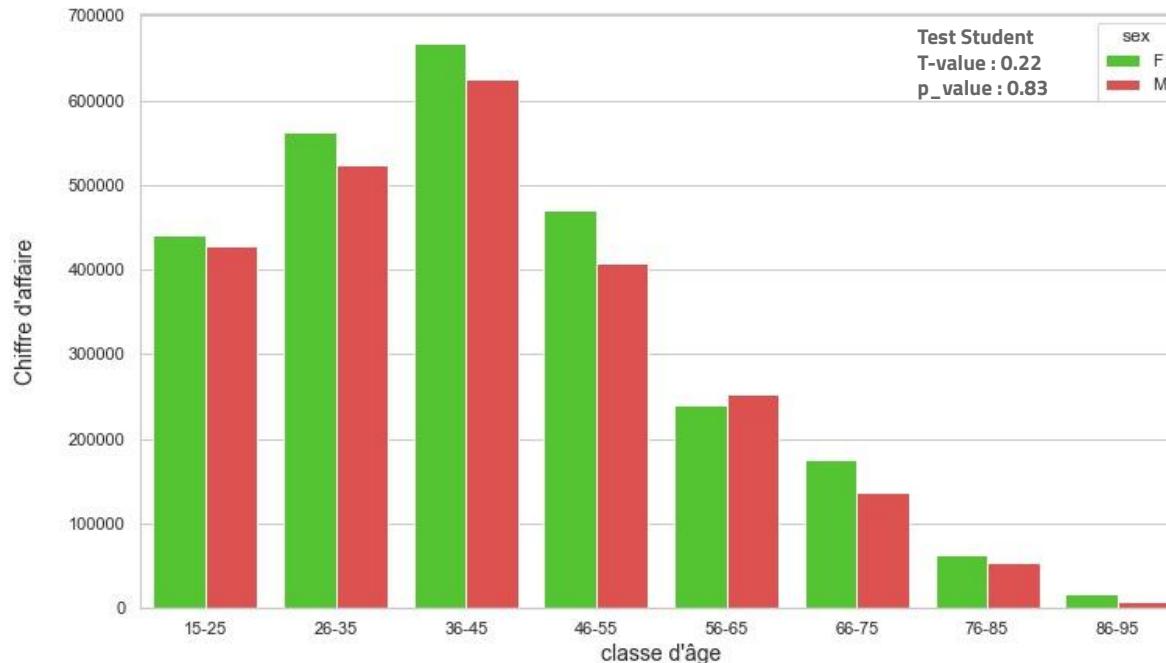
Les 36-45 ans, les meilleurs clients

Chiffre d'affaire (€) en fonction de la tranche d'âge du client(de mars 2021 à février 2022)



Pas de différence de dépense selon le genre

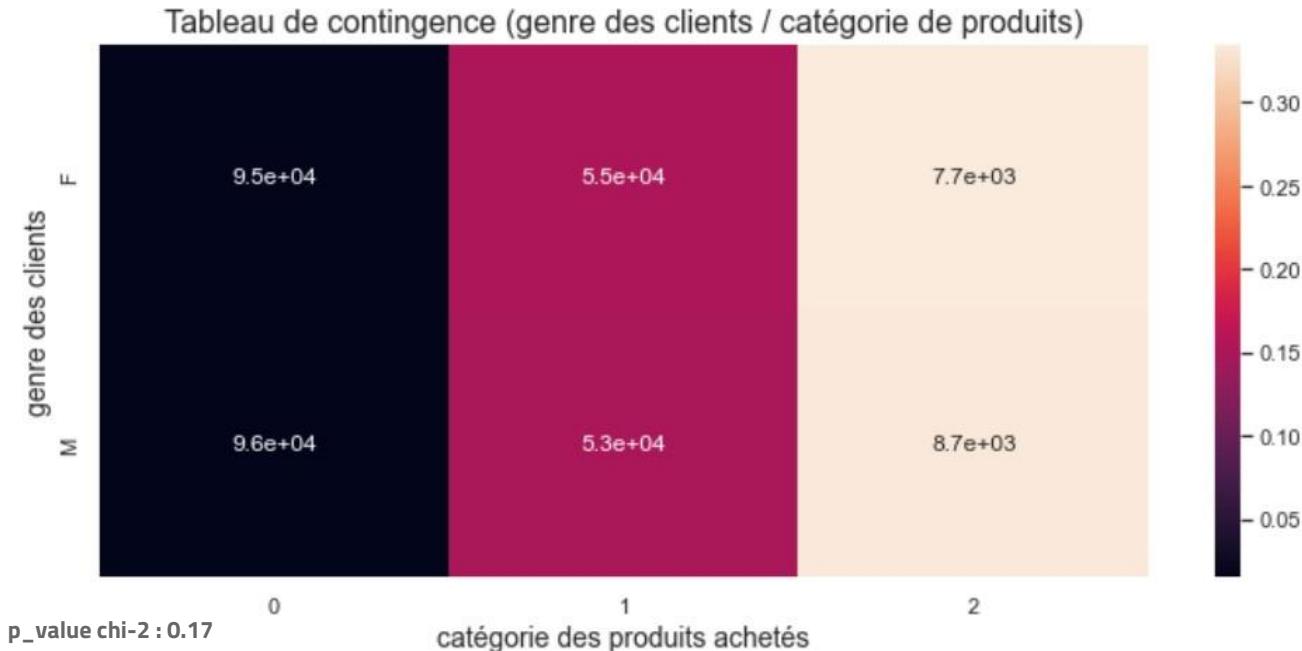
Répartition du chiffre d'affaire en fonction de la tranche d'âge et du genre (de mars 2021 à février 2022)





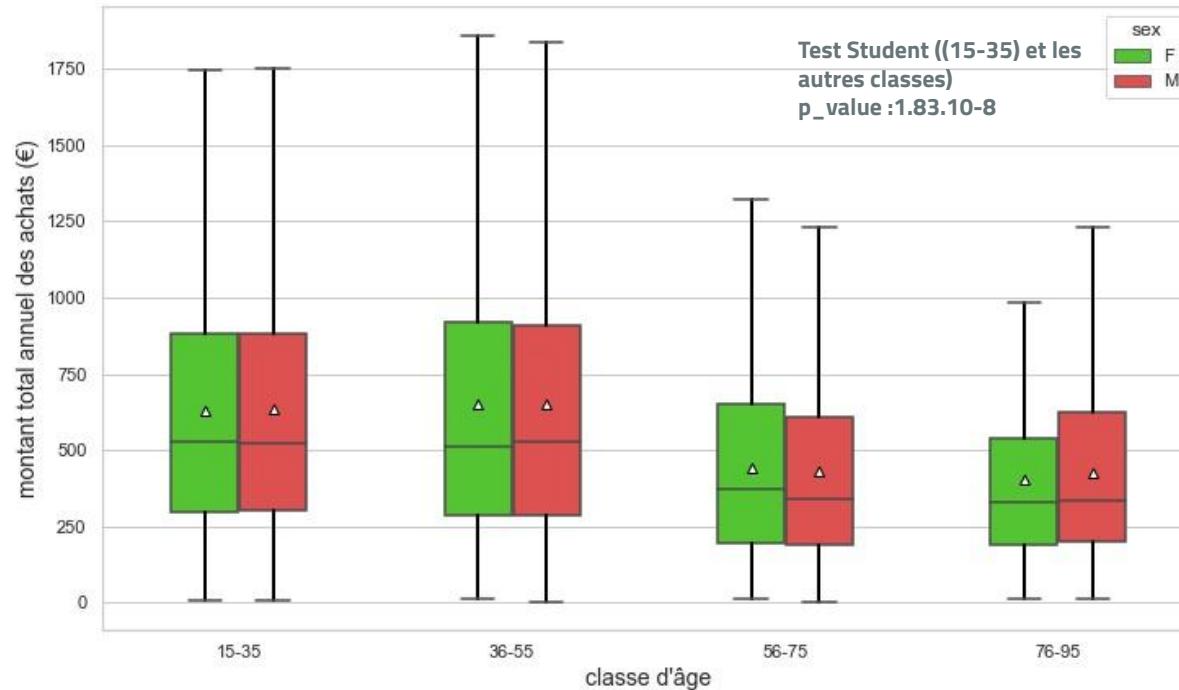
Etude des corrélations

Pas de corrélation entre le genre des clients et les catégories de produits



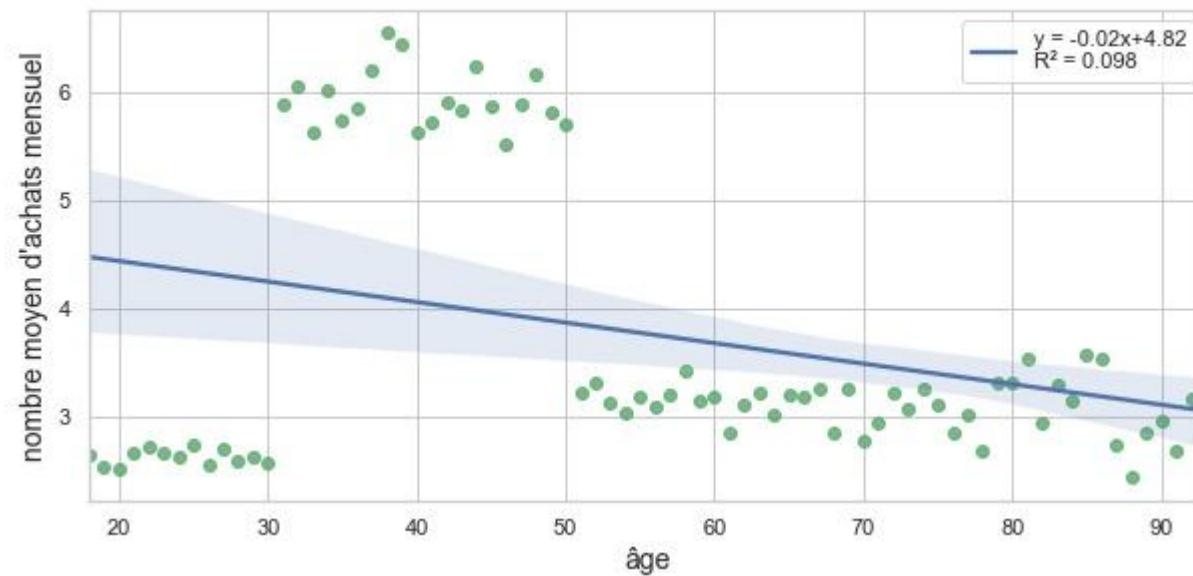
Corrélation entre l'âge des clients et le montant total des achats

Répartition du montant total annuel des achats (€) en fonction de l'âge des clients

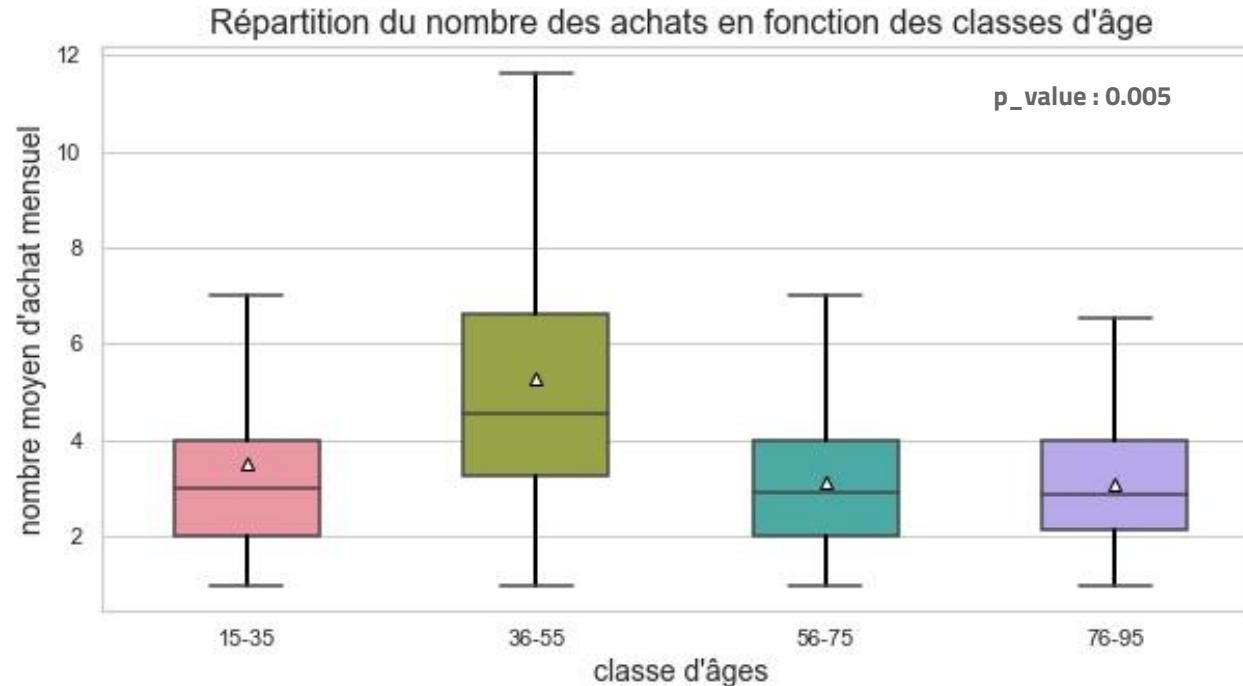


Faible corrélation entre l'âge des clients et la fréquence d'achat

Répartition du nombre moyen d'achats mensuel en fonction de l'âge des clients

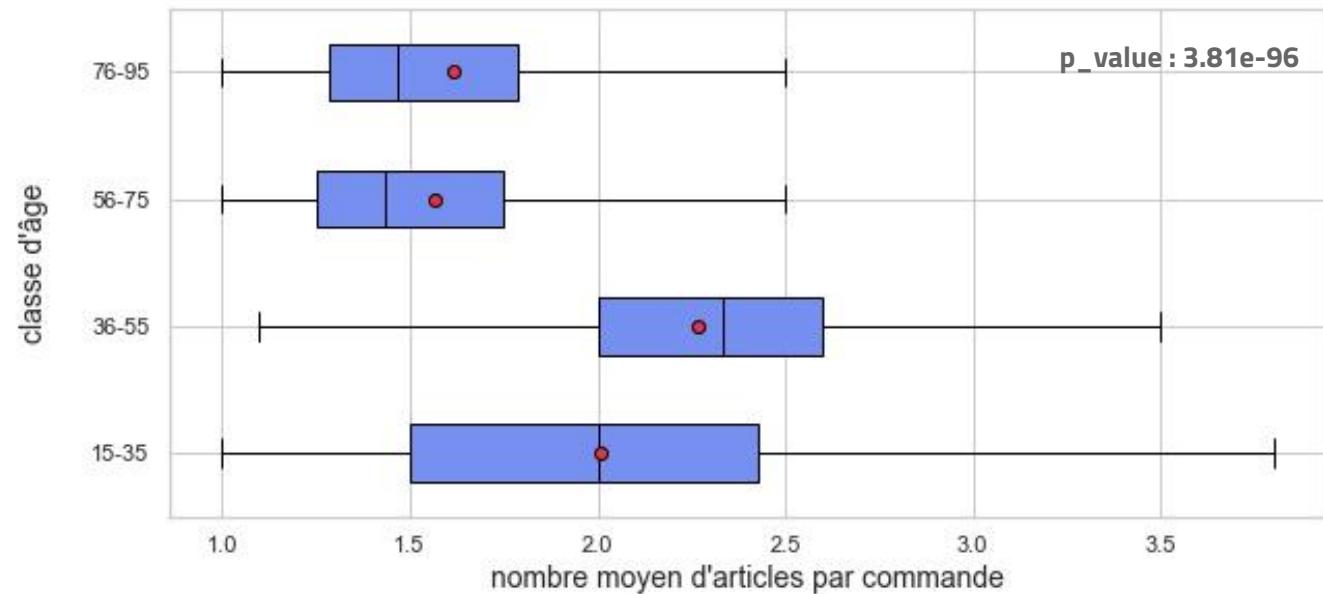


Faible corrélation entre l'âge des clients et la fréquence d'achat



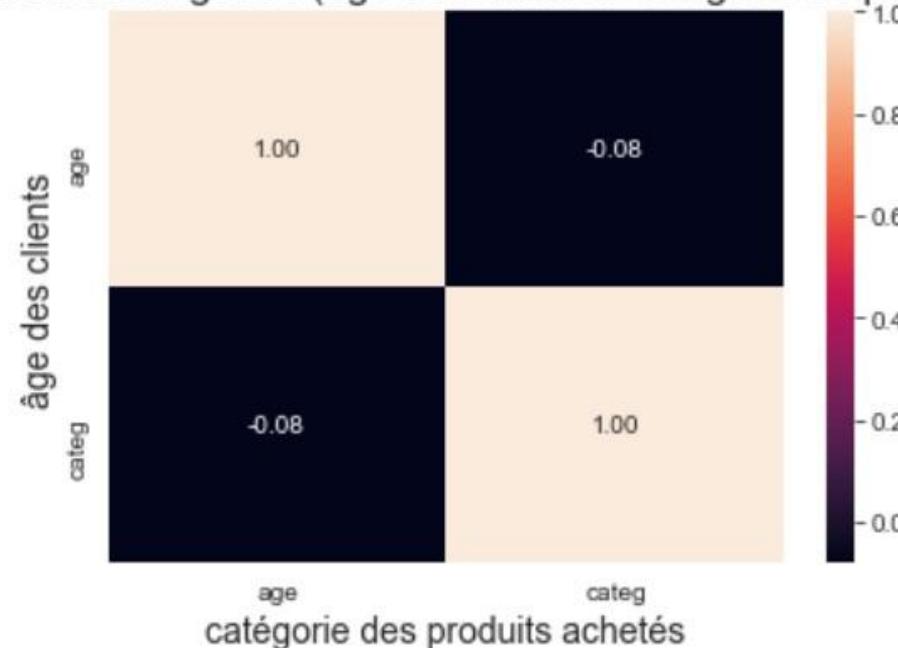
Corrélation entre l'âge des clients et la taille du panier moyen

Distribution du nombre d'article moyen par commande en fonction de l'âge



Pas de corrélation entre l'âge des clients et les catégories de produits achetés

Tableau de contingence (âge des clients / catégorie de produits)



Merci!

Des questions?

Contactez - moi:



www.linkedin.com/in/isabelle-barbier