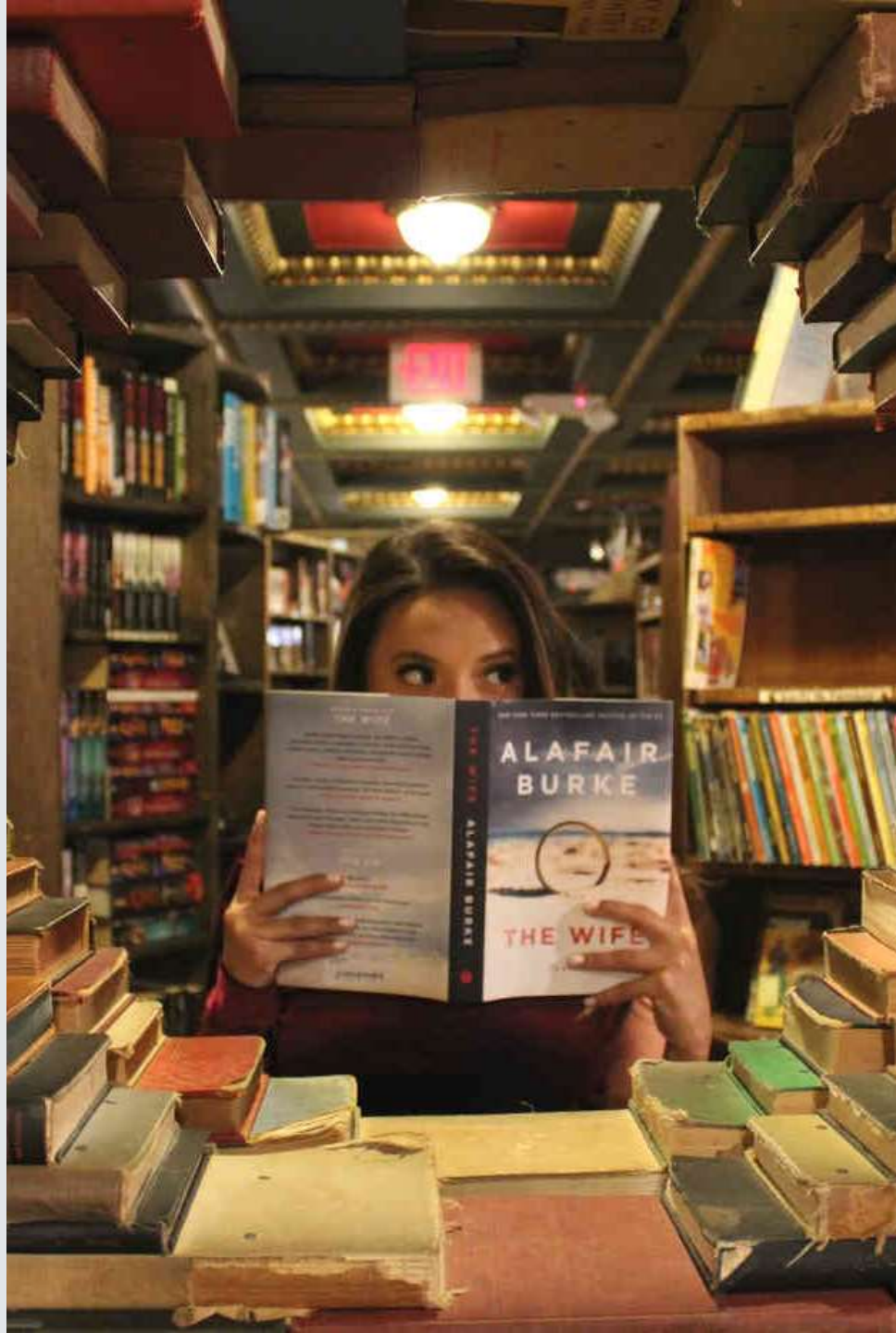
ANALYSEZ LES VENTE D'UNE LIBRAIRIE AVEC R OU PYTHON

Par Claude Olukoya



ANALYSES EXPLORATOIRES DES DONNÉES DE DÉPART

	df_Customers	df_Products	df_Transactions
Lignes :	8,621	3,286	1,048,575
Colonnes :	3	3	4
Clé primaire :	Client_id	Id_prod	Session_id

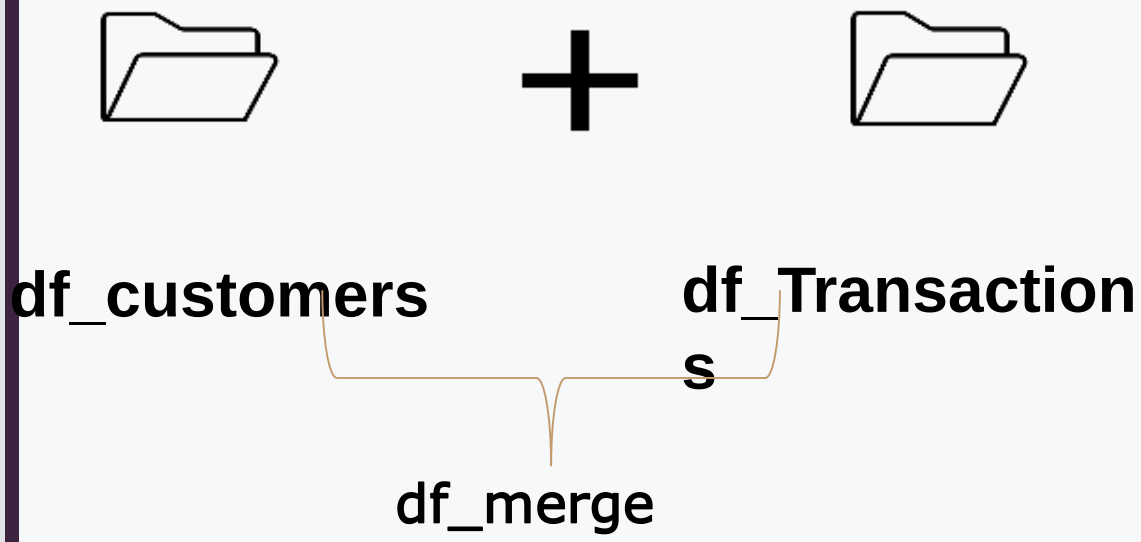
CARACTÉRISTIQUES DU DATASET :

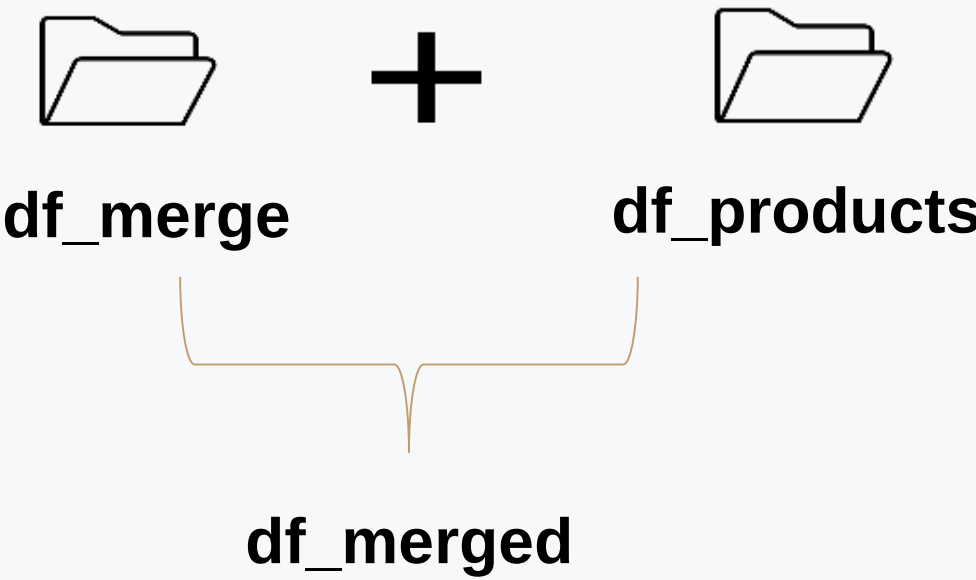
• **Customers.csv :** L’extraction du id des clients, sexe, leur date de naissance.

• **Products.csv :** L’extraction des références produits, le prix et sa catégorie.

• **Transaction.csv :** Une table de transactions qui permet de lier les références clients, produits vendus et la session d’achat.

JONCTION DES FICHIERS





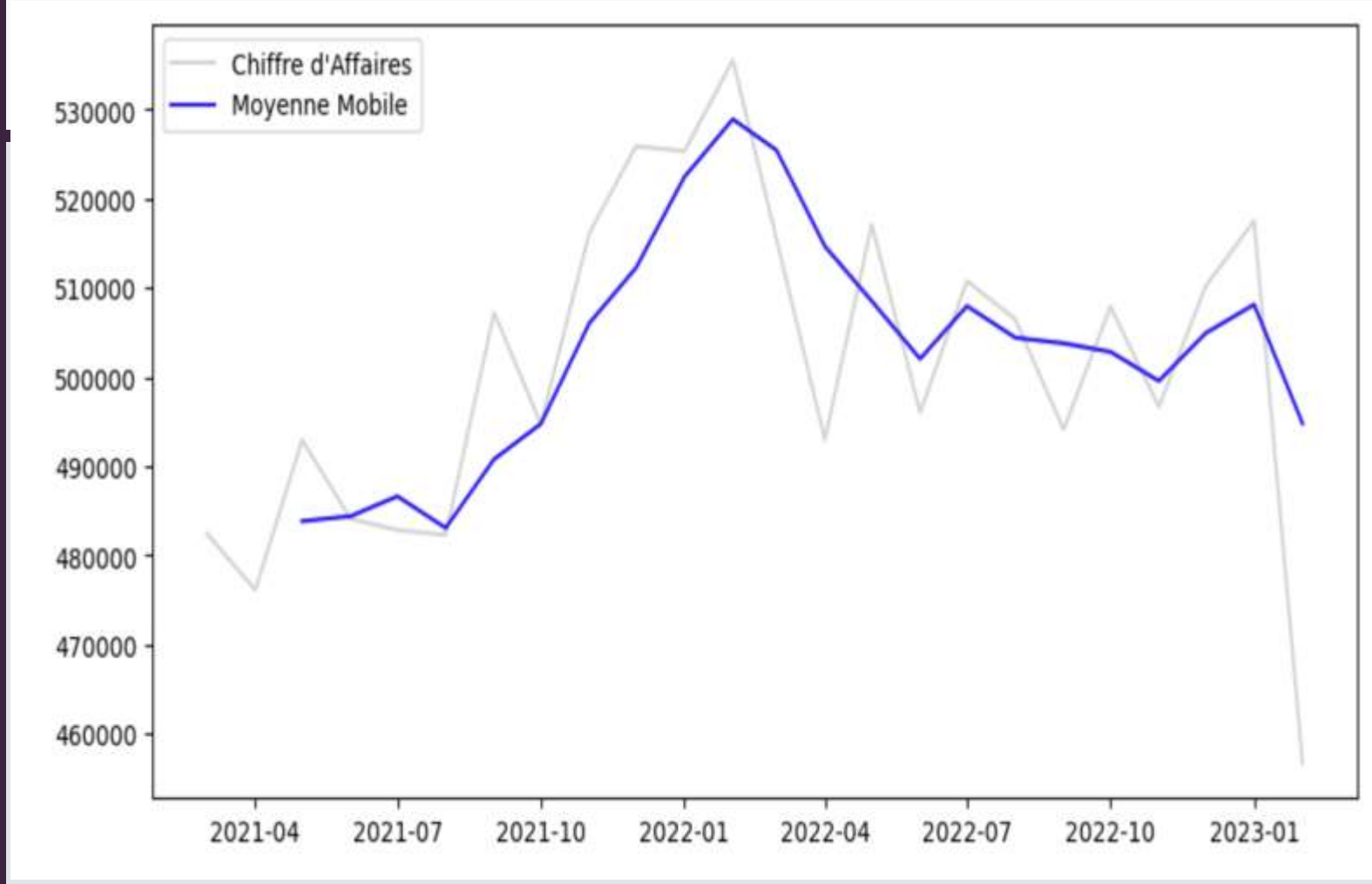
NETTOYAGE DU DATASET :

- Vérification de la présence des doublons dans chaque colonne des fichiers
- Vérification de la présence de valeurs négatives
- Repérage des outliers dans les 3 fichiers
- Conversion de la colonne ‘date’ (objet) en DateTime dans transaction.csv
- Vérification du code de la codification
- 360,000 de lignes de valeurs manquantes supprimées dans le fichier Transactions.csv

LES DIFFERENTS INDICATEURS DE VENTES

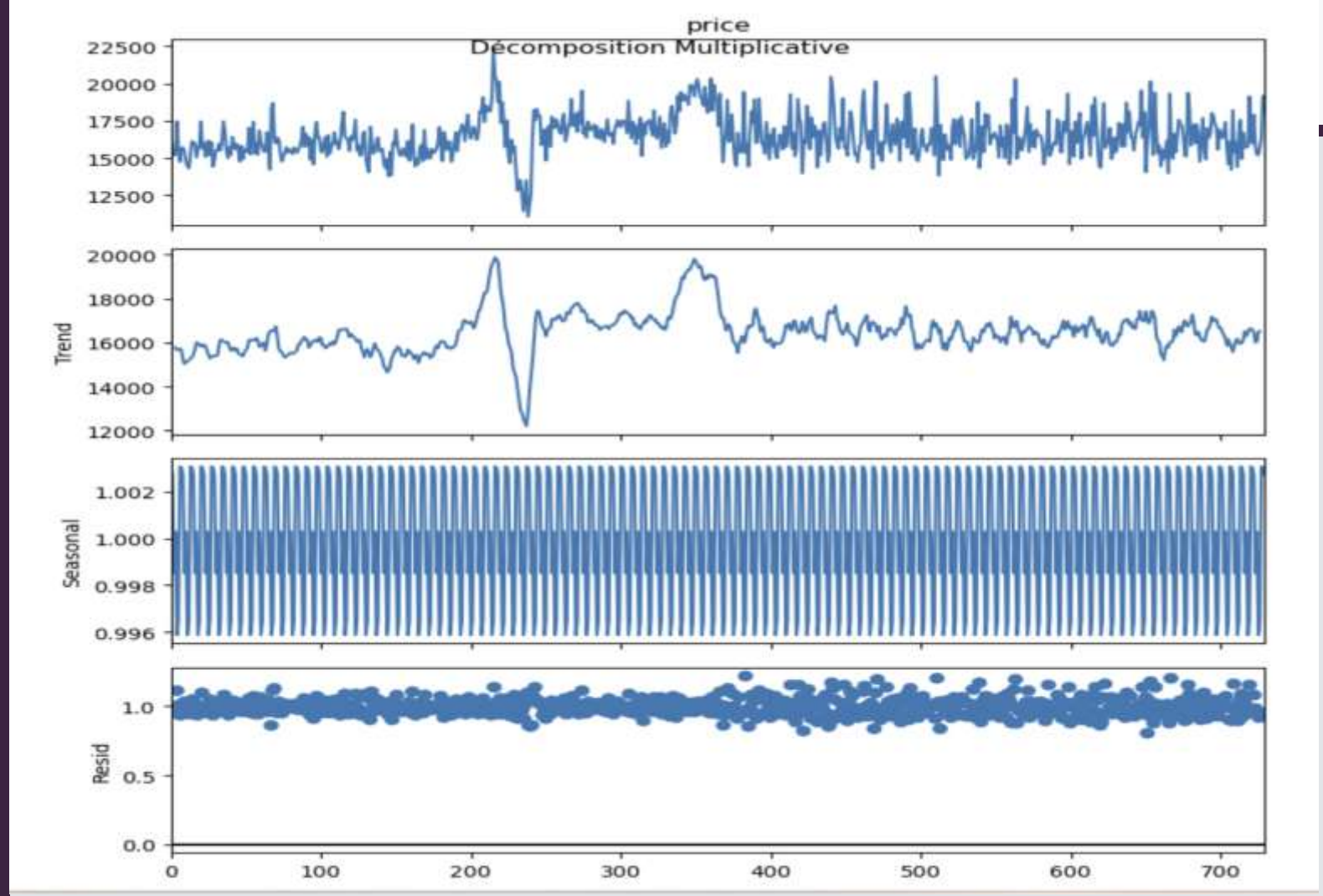
CHIFFRE D’AFFAIRES PAR MOIS AVEC LA MOYENNE MOBILE :

permet de lisser une série de valeurs exprimées en fonction du temps (série chronologique). Elle permet d'éliminer les fluctuations les moins significatives



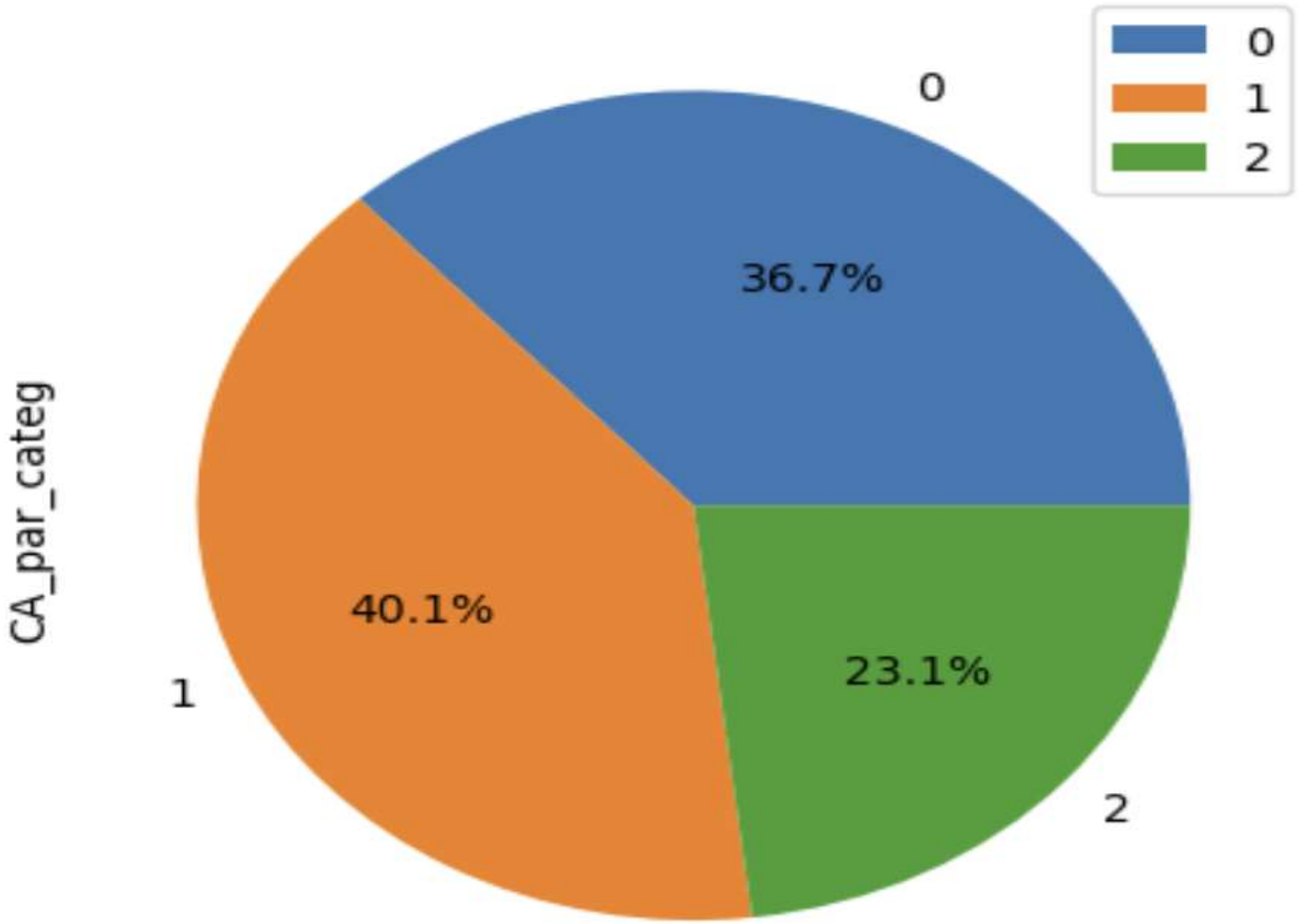
DECOMPOSITION DE LA TIME SERIES (par semaine) de la variable PRIX :

aide à comprendre les modèles, les tendances, les cycles et la saisonnalité des données, et à établir des prévisions basées sur des données historiques.



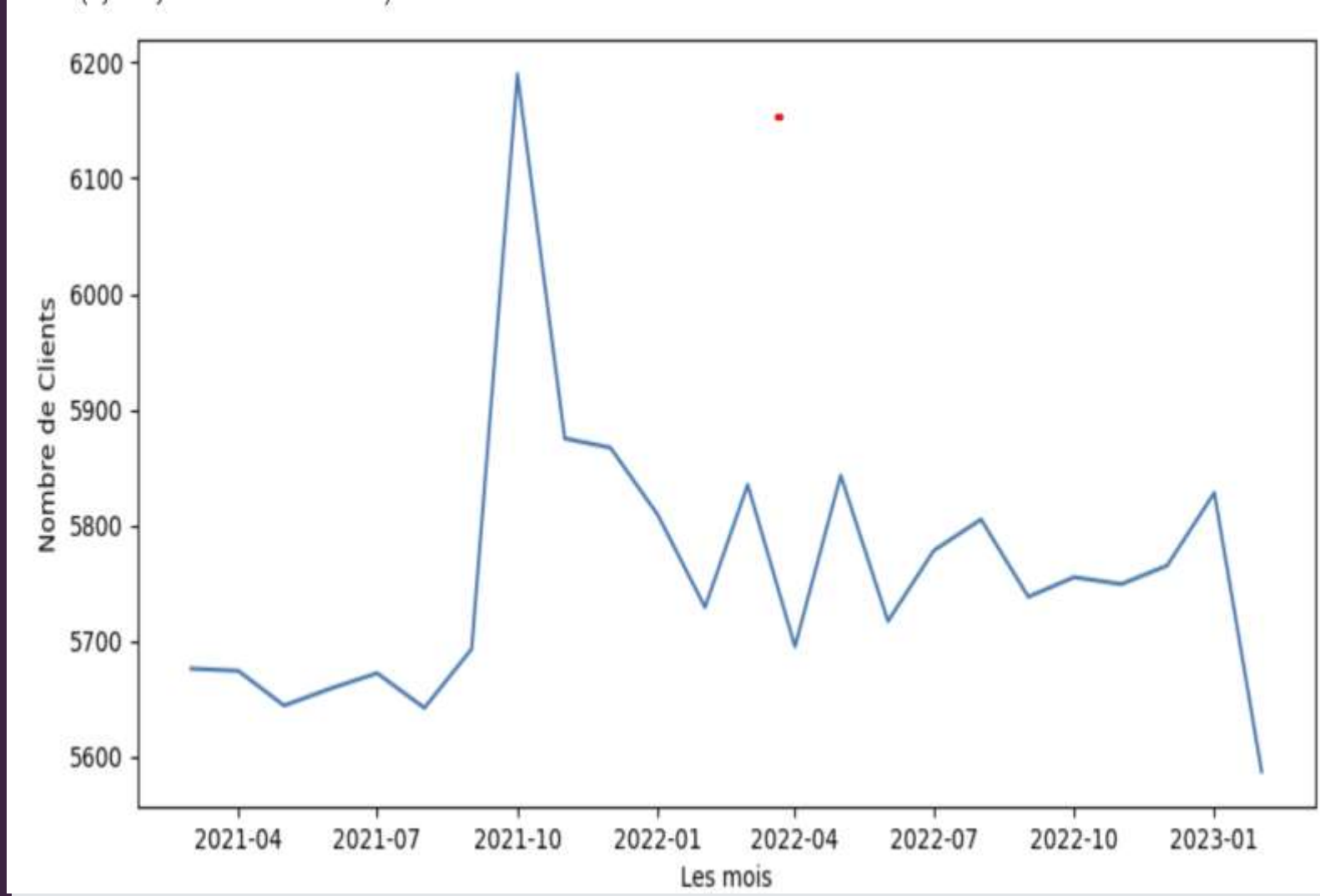
CHIFFRE D'AFFAIRES PAR CATÉGORIE:

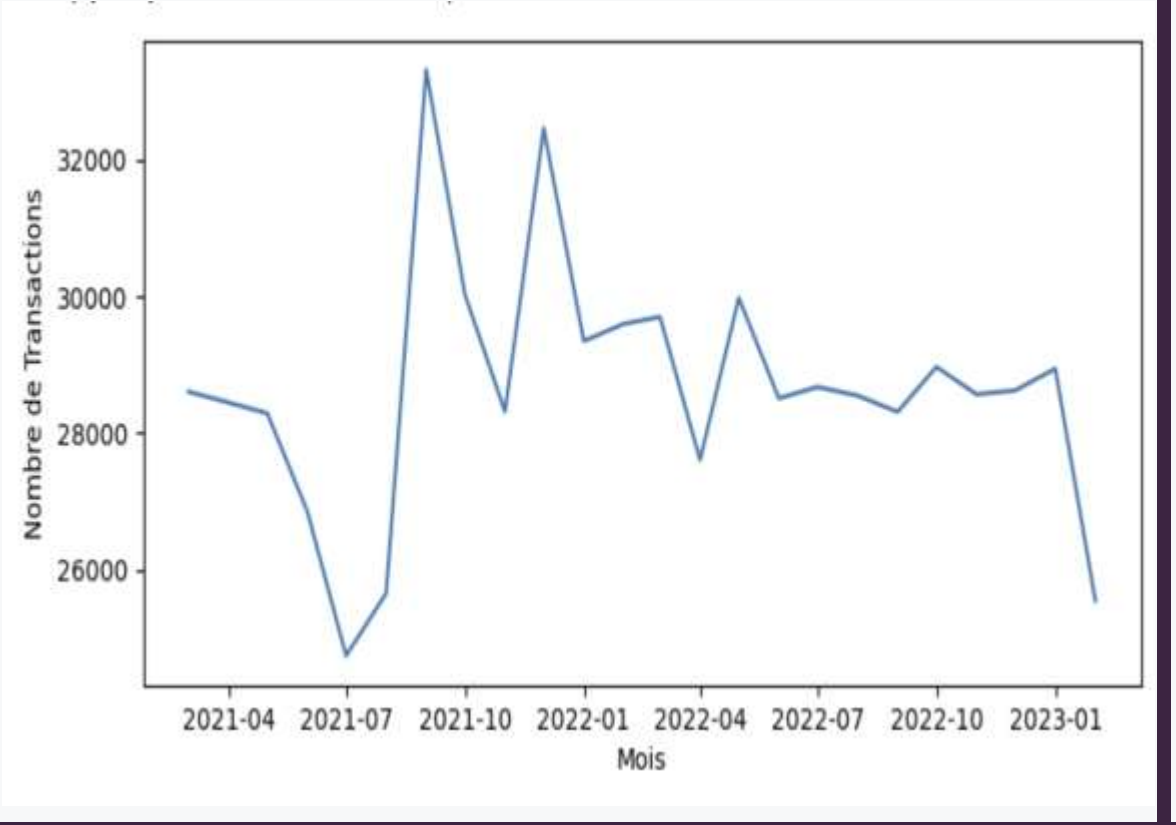
Nous constatons que les livres de la catégorie 1 arrivent en tête du chiffre affaires de la librairie suivis des livres de la catégorie 0. Les livres de la catégorie 2 arrivent en dernier.

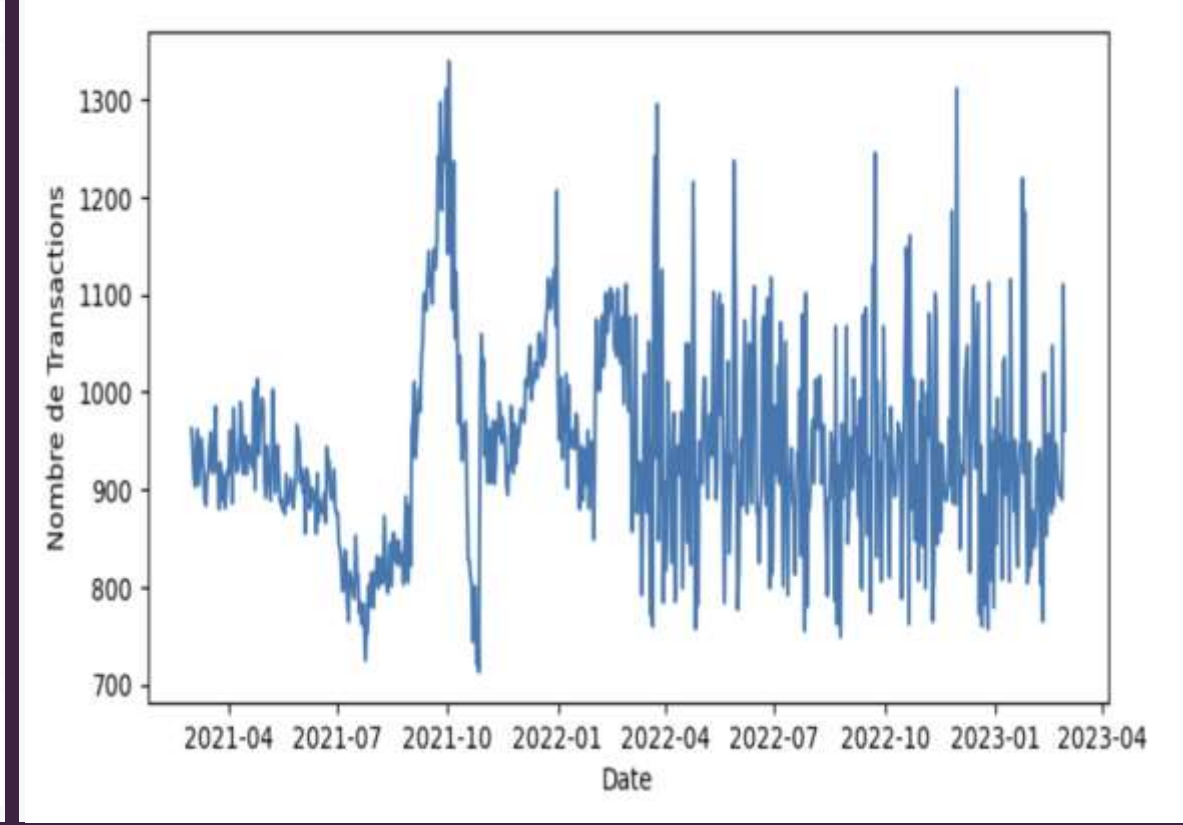


NOMBRE DE CLIENTS PAR MOIS :

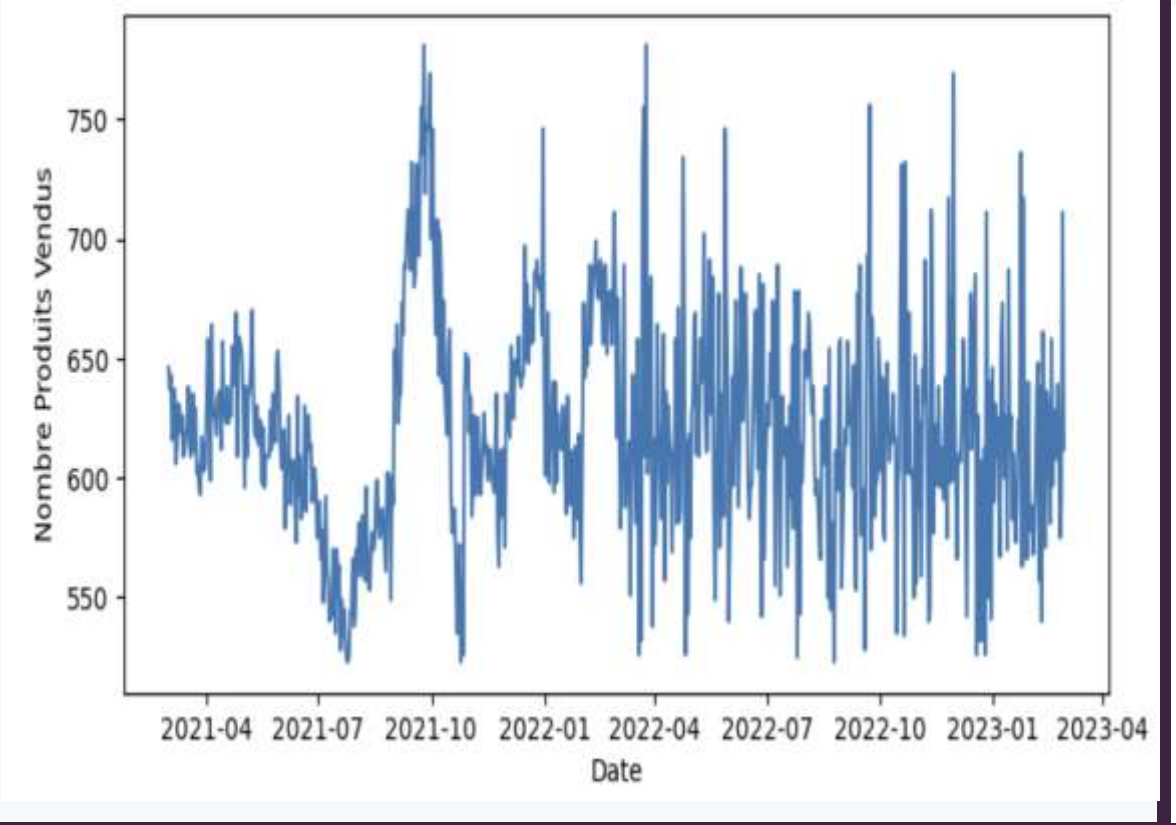
La fonction ‘nunique’ a été utilisée pour ne pas compter un client plus d’une fois

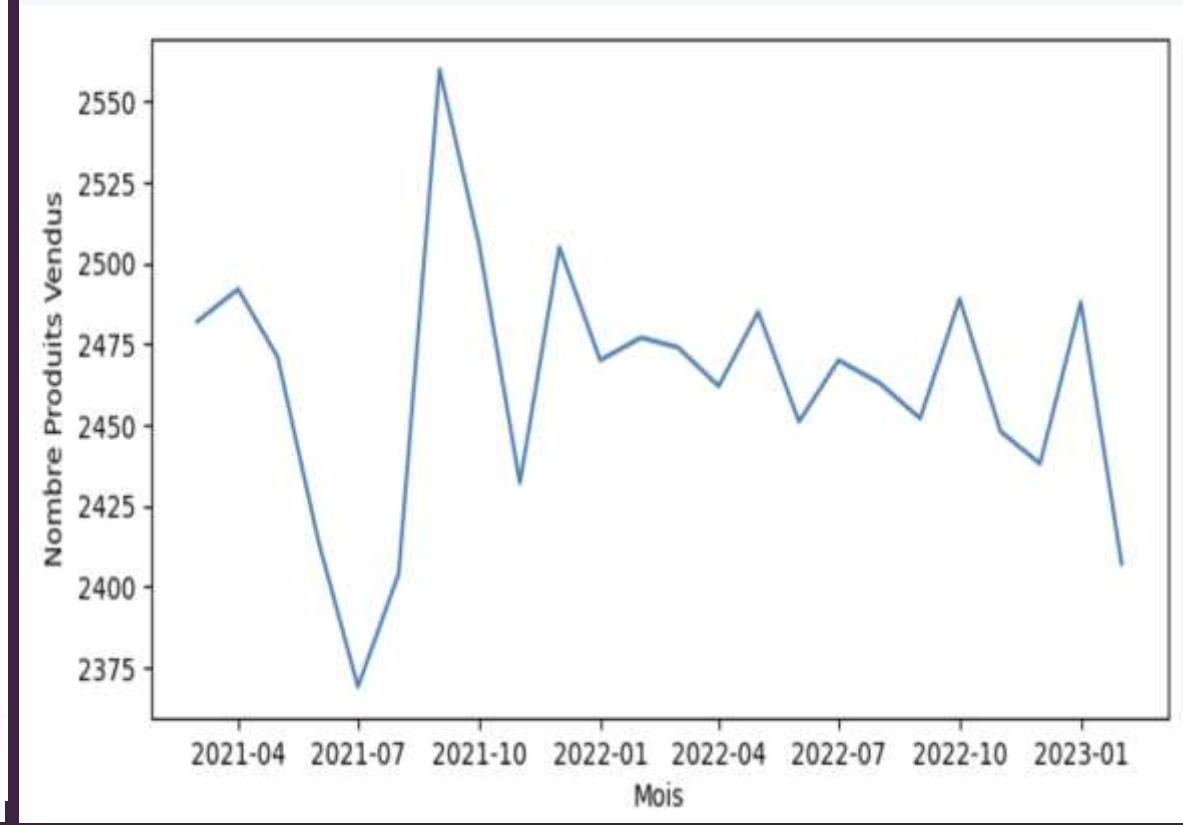






NOMBRE DE TRANSACTIONS PAR JOUR ET PAR MOIS

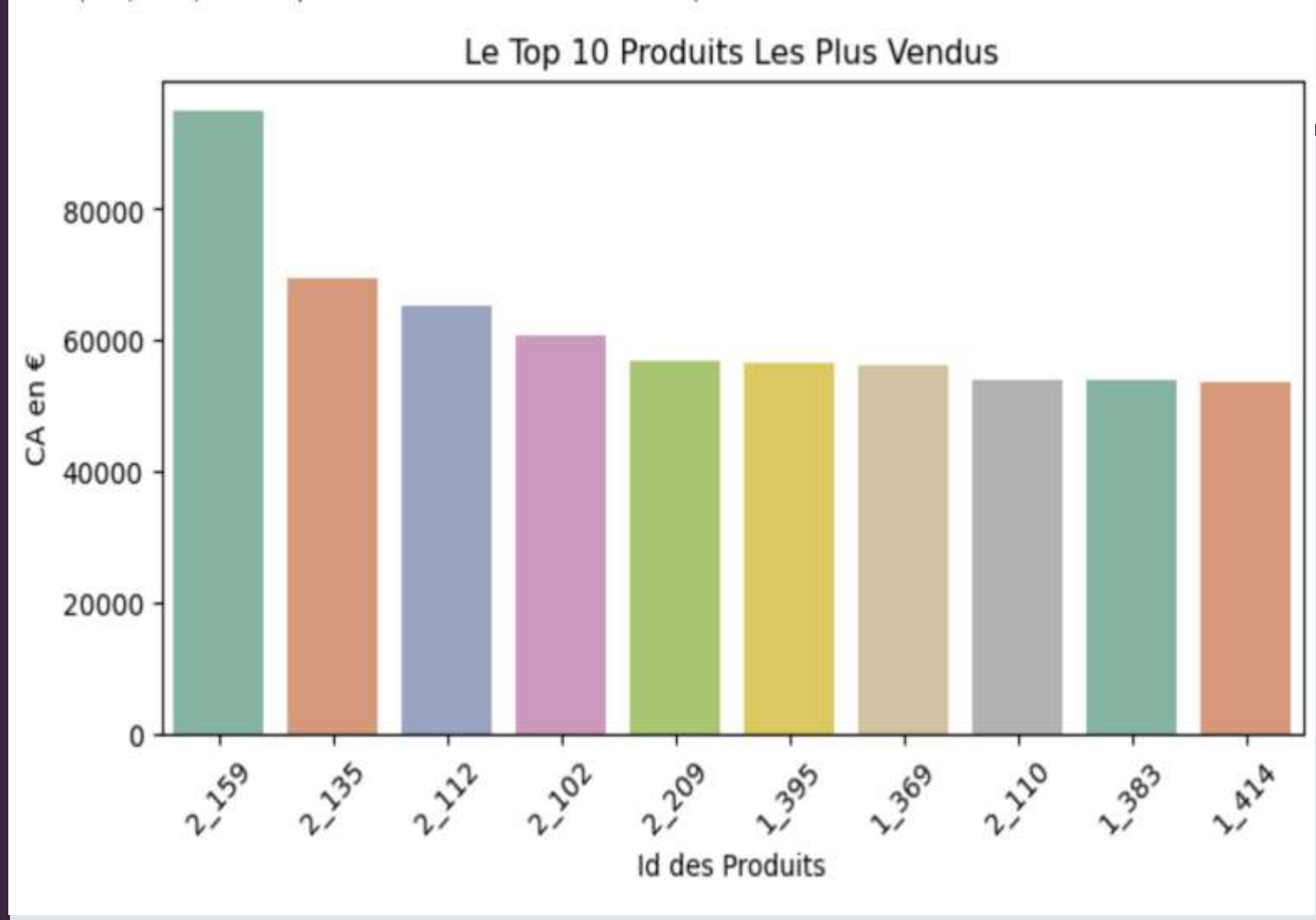




NOMBRE DE PRODUITS (UNIQUES) VENDUS PAR JOUR ET PAR MOIS

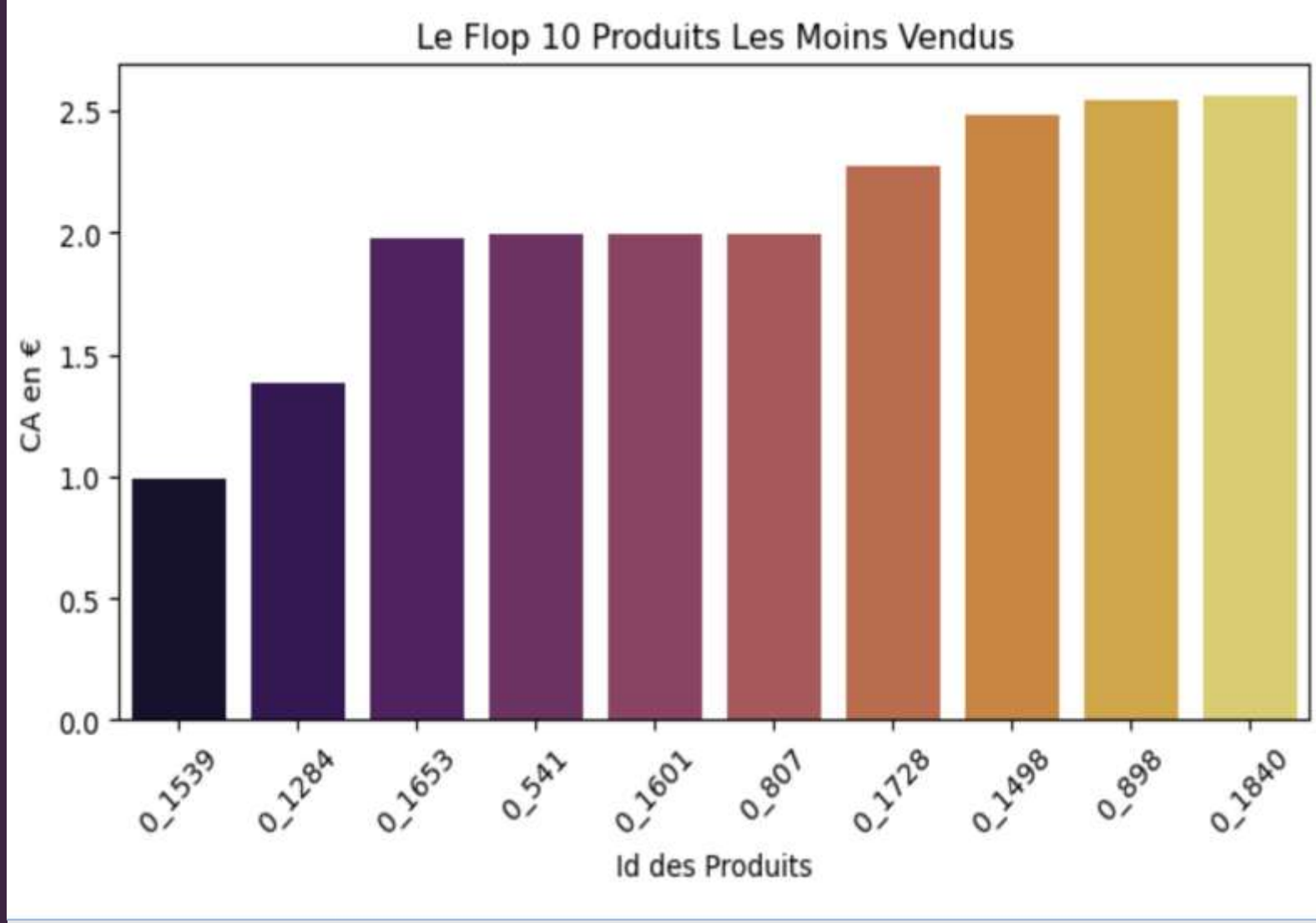
TOP 10 PRODUITS LES PLUS VENDUS:

Le produit qui rapporte le plus à la bibliothèque est le ‘id prod 2159’.



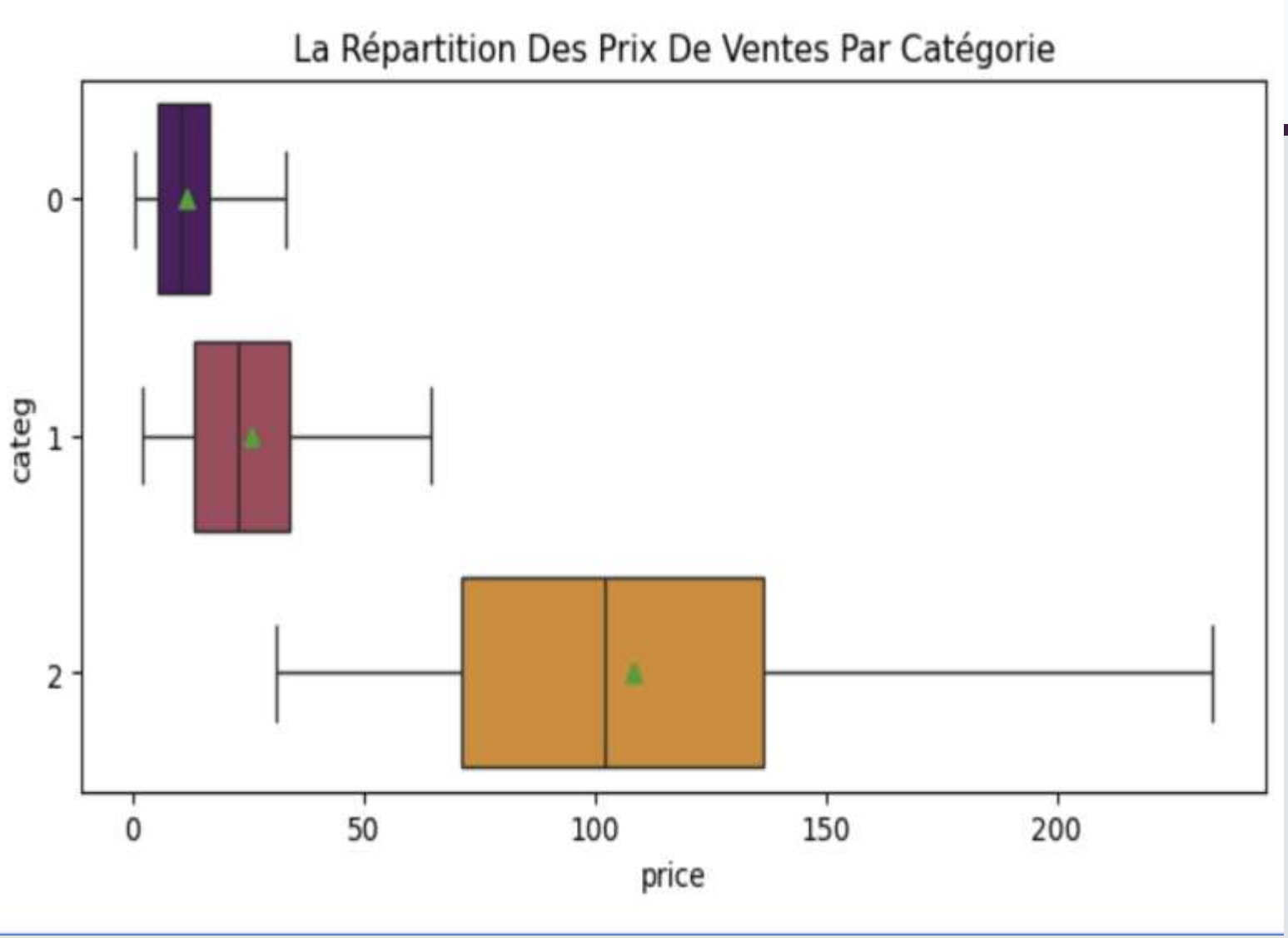
FLOP 10 PRODUITS LES PLUS VENDUS:

Le flop produit qui rapporte le moins à la bibliothèque est le ‘id prod 1539.



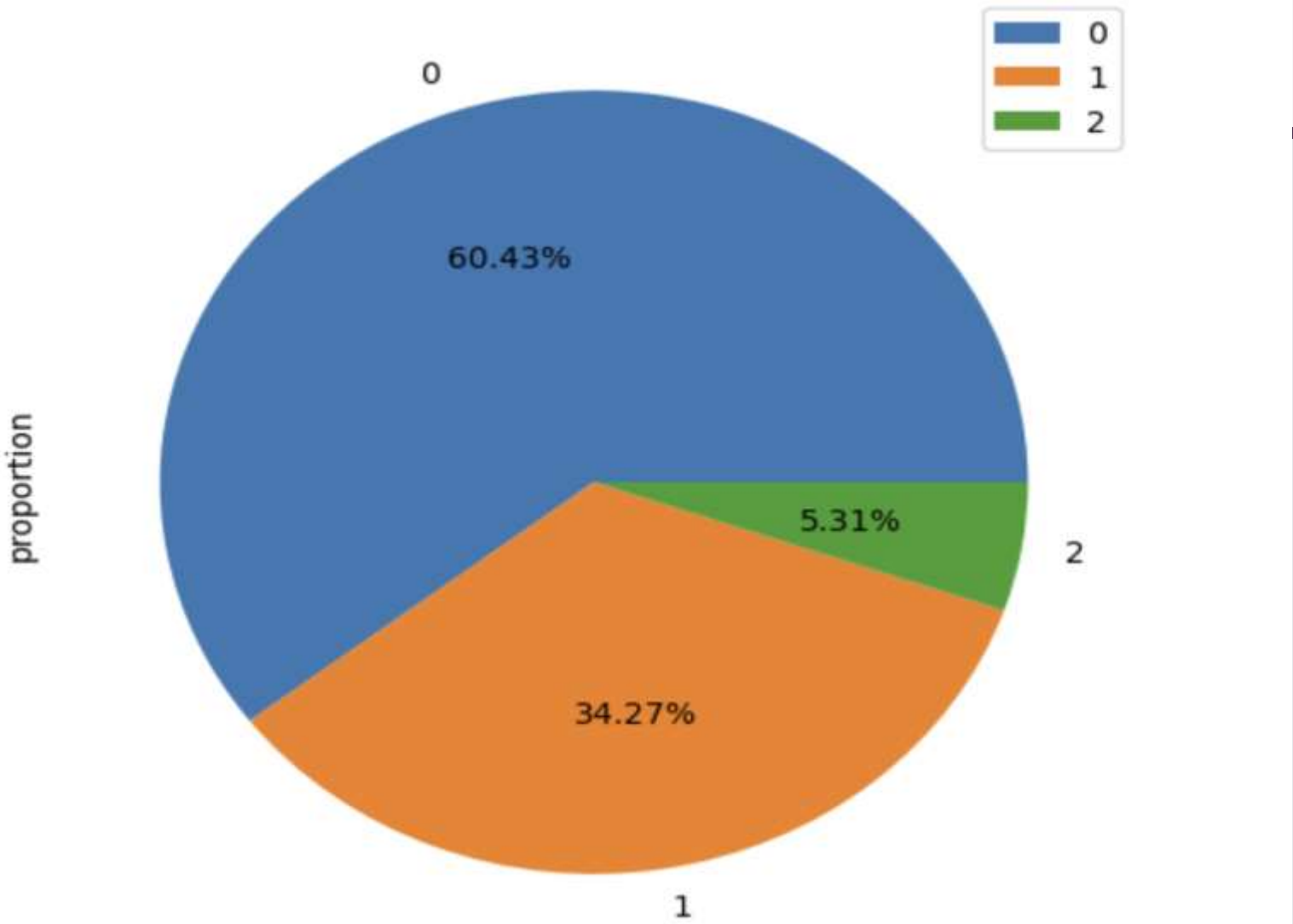
LA RÉPARTITIONON DES PRIX DE VENTES PAR CATEGORIE :

Les livres de la catégorie 2 sont le plus élevés des trois catégories



LA RÉPARTITION DU VOLUME DES VENTES PAR CATÉGORIE :

Les livres de catégorie 0 se vendent le plus suivis des livres de la catégorie 1



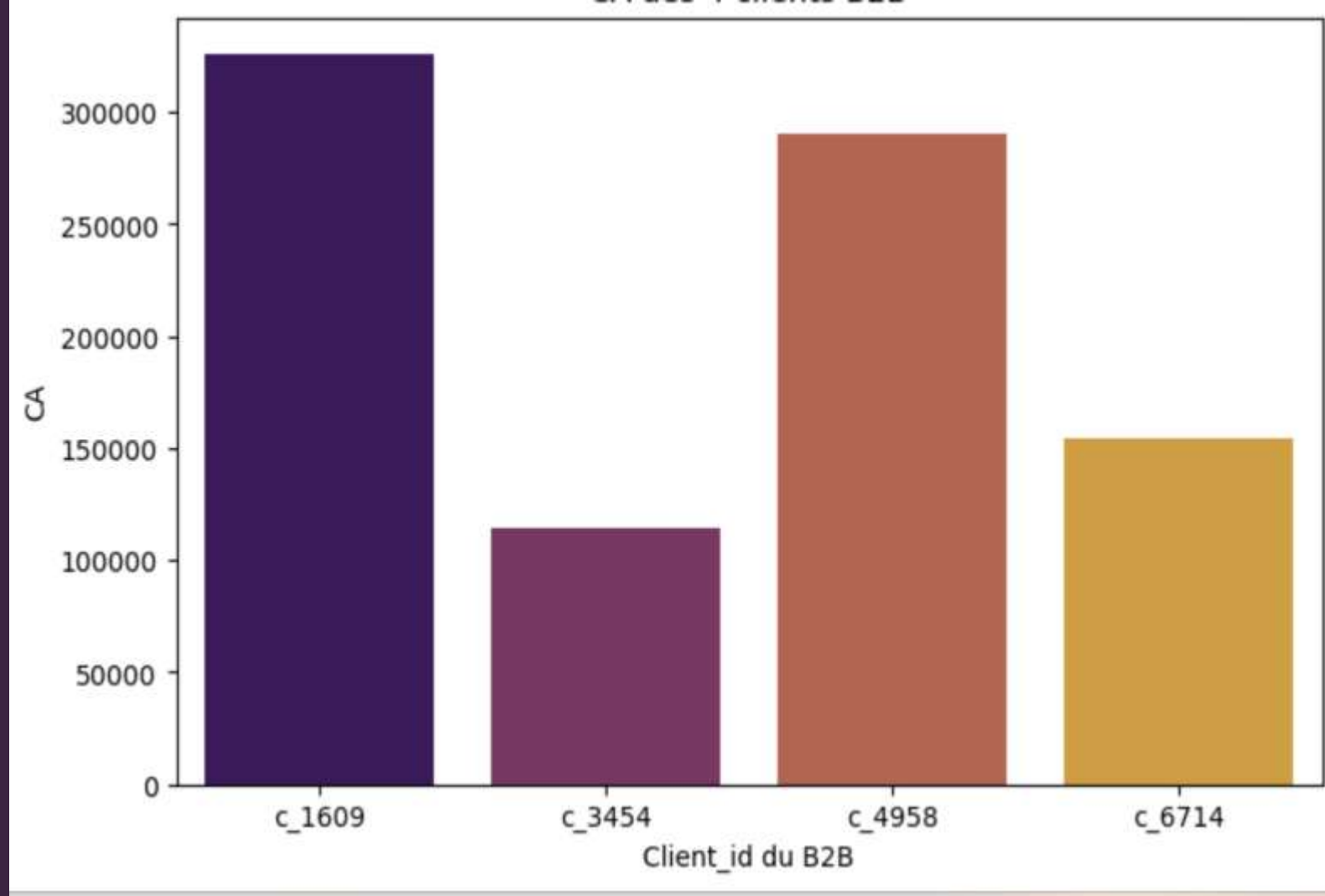
INFORMATIONS SUR LE PROFIL DES CLIENTS

DEUX TYPES DE CLIENTS :

• **B2B :** En calculant le top dix plus gros chiffres d'affaires de la librairie, nous remarquons une anomalie. Le 4 premiers (probablement des sociétés) ont des chiffres d'affaires qui se distinguent des autres. Nous mettons ces 4 clients à part (dans un nouveau DataFrame).

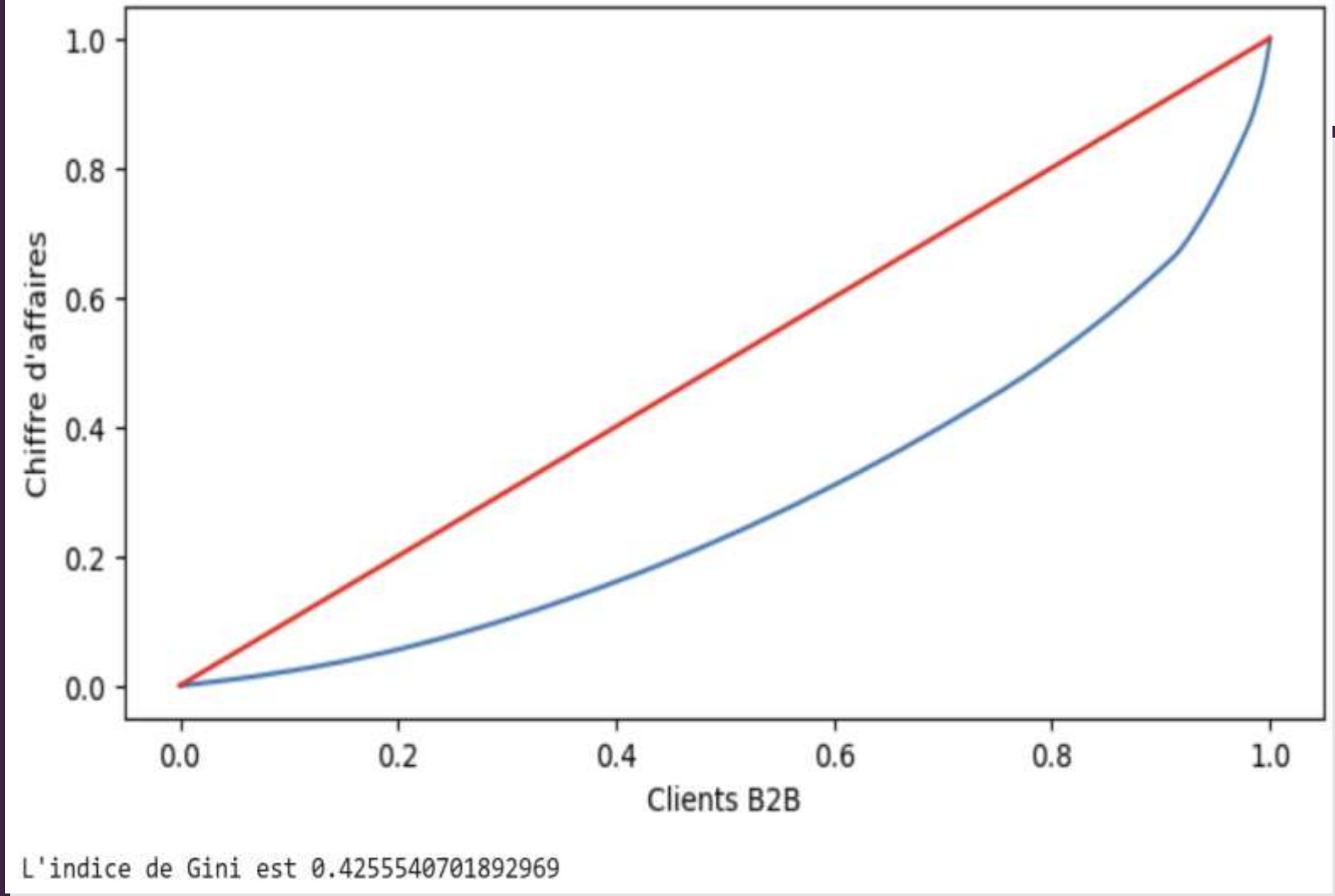
• **PARTICULIERS :** Nous mettons le 2ème type de clients dans un nouveau DataFrame qu'on nommera 'Particuliers'.

RÉPARTITION DU CHIFFRE D'AFFAIRES PAR LES CLIENTS B2B



COURBE DE LORENZ POUR LES CLIENTS B2B

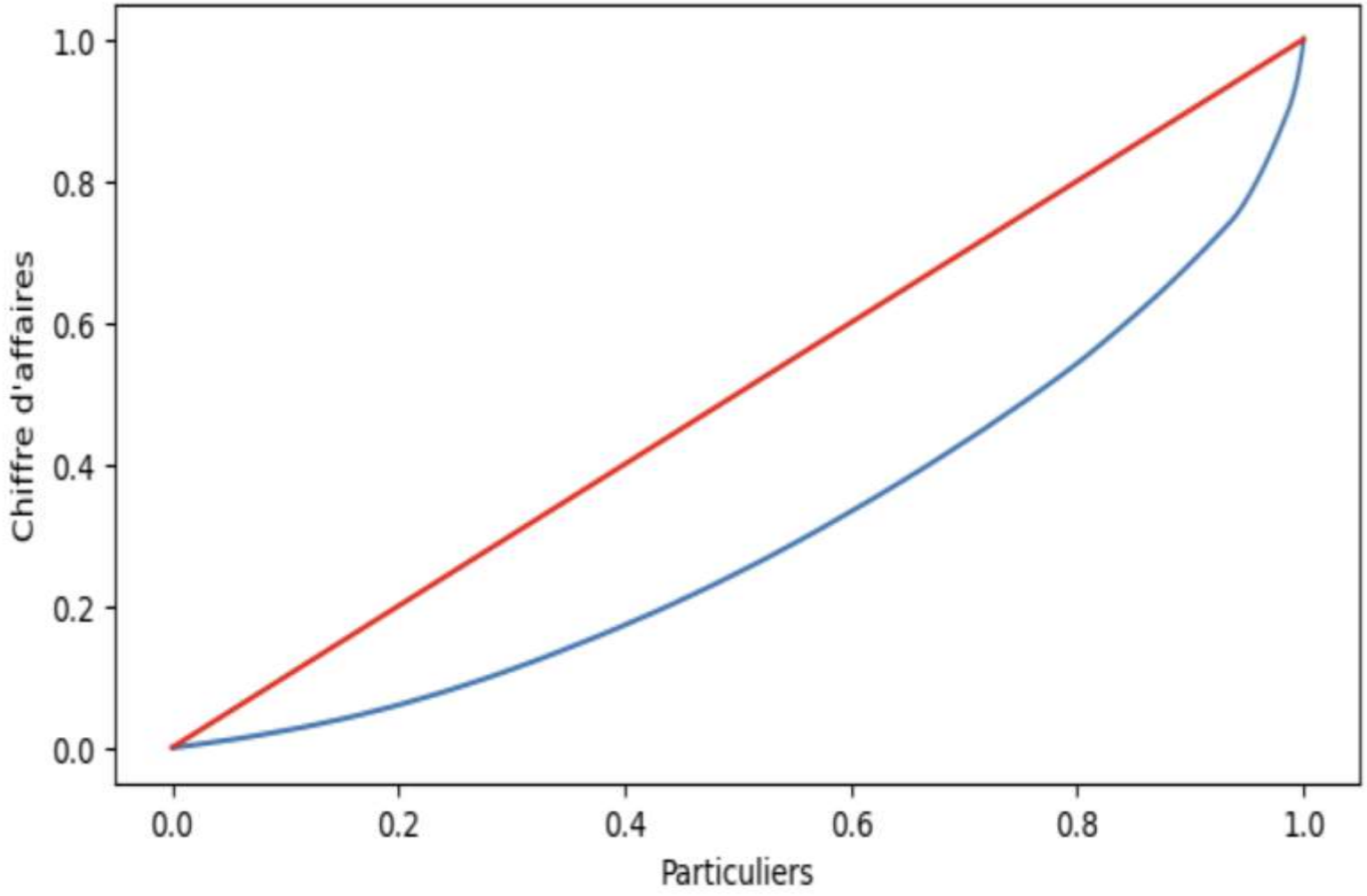
Nous constatons que la courbe de Lorenz n’est pas loin de la ligne d’égalité parfaite, ce qui signifie que la distribution du CA entre les 4 gros clients B2B est moyennement partagée entre eux.



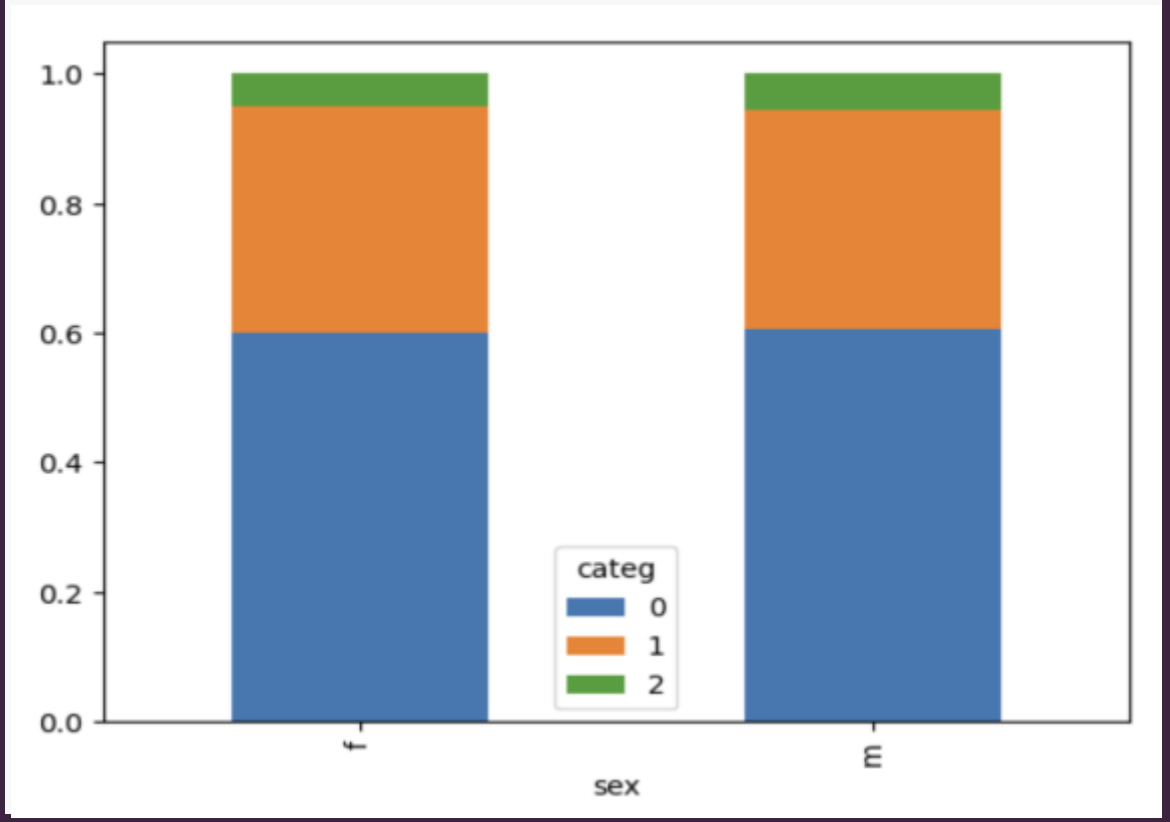
L'indice de Gini est 0.4255540701892969

COURBE DE LORENZ POUR LES PARTICULIERS :

L’indice de Gini à 0.39 est très proche de 0 pour les Particuliers ce qui signifie que la distribution du CA est bien distribuée entre eux.



L'indice de Gini est 0.39085002982852024



Chi-Squared test

```
# Détermine s'il existe une association entre des variables catégorielles (c'est-à-dire si les variables sont indépendantes ou liées).
# Il s'agit d'un test non paramétrique

# Il y a trois assumptions :

# 1) Les observations dans chaque echantillons sont indépendantes et distribuées identiquement
# 2) La valeur attendue des cellules doit être de 5 ou plus dans au moins 80 % des cellules
# 3) Les deux variables doivent être catégoriques

# Hypothèse :

# H0 : Les variables sont indépendantes si la p_value est > 0.05

# Ha : Il y a une dépendance entre les deux variables si la p_value est < 0.05
```

```
d = chi2_contingency(df_sexe_cat_CA)
d
```

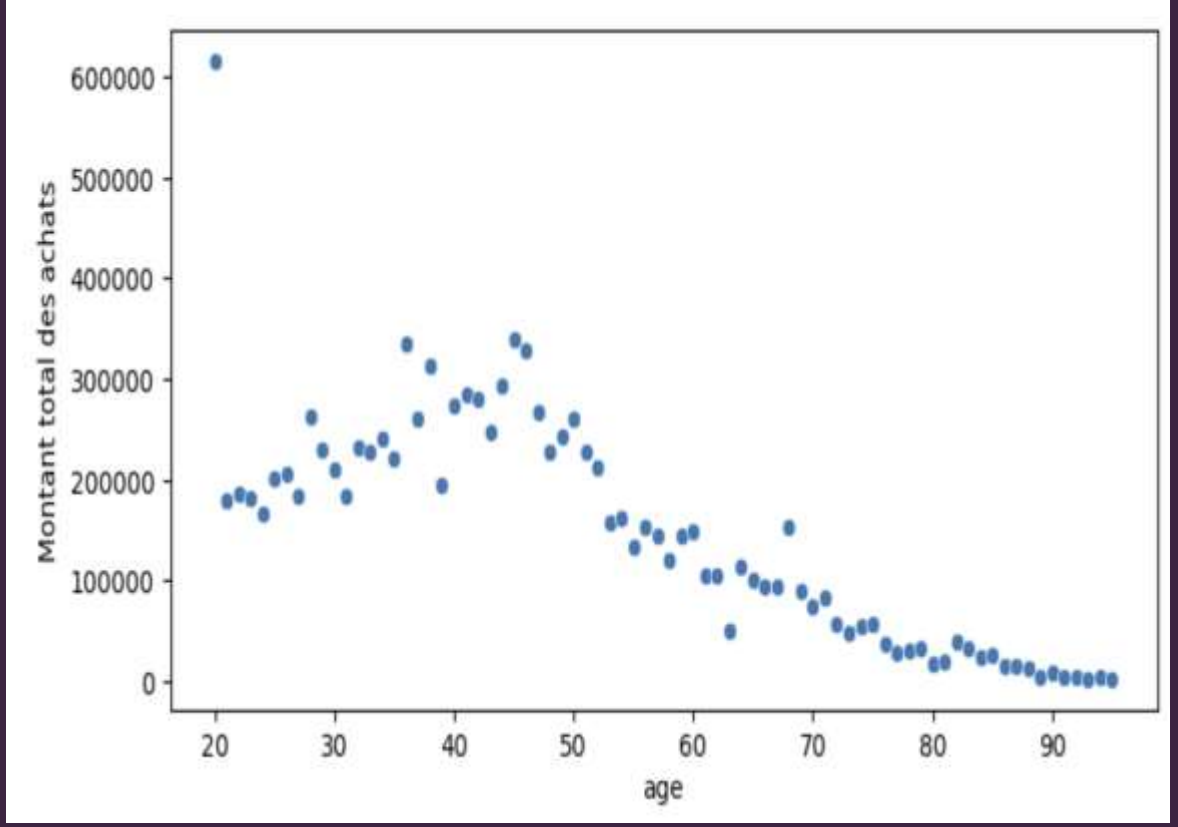
```
Chi2ContingencyResult(statistic=22.66856665178056, pvalue=1.1955928116587024e-05, dof=2, expected_freq=array([[201574.89662481, 114822.13191434,  17096.97146086],
       [185706.10337519, 105782.86808566,  15751.02853914]]))
```

```
#Unpacking

stats, p_value, a,b = d
```

```
alpha = 0.05
if p_value > alpha:
    print(f"p_value est {p_value,10} on peut conclure avec 95% certitude que les variables sont indépendantes")
else:
    print(f"p_value est {p_value} on peut conclure avec 95% certitude qu'il y a une forte dépendance entre les deux variables")
```

```
p_value est 1.1955928116587024e-05 on peut conclure avec 95% certitude qu'il y a une forte dépendance entre les deux variables
```

LIEN : LE GENRE D'UN CLIENT ET LA CATÉGORIE DES LIVRES ACHETÉS



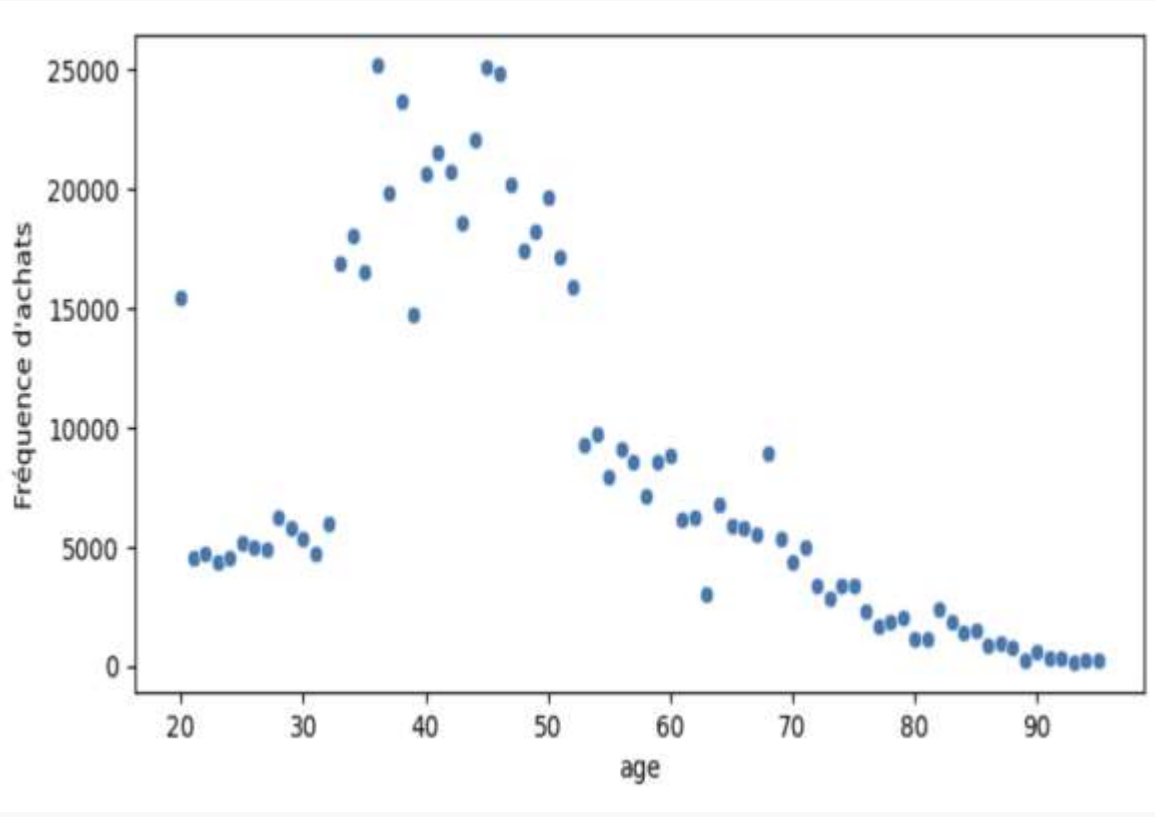
Spearman corrélation

```
# Hypothèse :

# H0 : Il n'y a pas de corrélation entre les deux variables :  rs = 0

# Ha : Il y a une corrélation entre les deux variables : rs != 0

# la corrélation de Spearman entre deux variables est égale à la corrélation de Pearson entre les "valeurs de rang" de ces deux variables
# Elle mesure la force mais aussi la direction de la relation entre deux variables continues

from scipy.stats import pearsonr,spearmanr

# rs et p_value de Spearman

spearmanr_coefficient, p_value = spearmanr(df_age_CA["age"],df_age_CA["CA"])
print("Le coffecient de Spearman est %0.3f"% (spearmanr_coefficient))
print("La p_value est : ", p_value)
```

```
le coffecient de Spearman est -0.874
la p_value est :  5.956077505475151e-25
```

```
# la (rs) Spearman Corrélation rejète l'hypothèse nulle et le test révèle une correlation négative très forte.
```

LIEN : AGE DES CLIENTS ET LE MONTANT TOTAL DES ACHATS



Spearman Corrélation Test

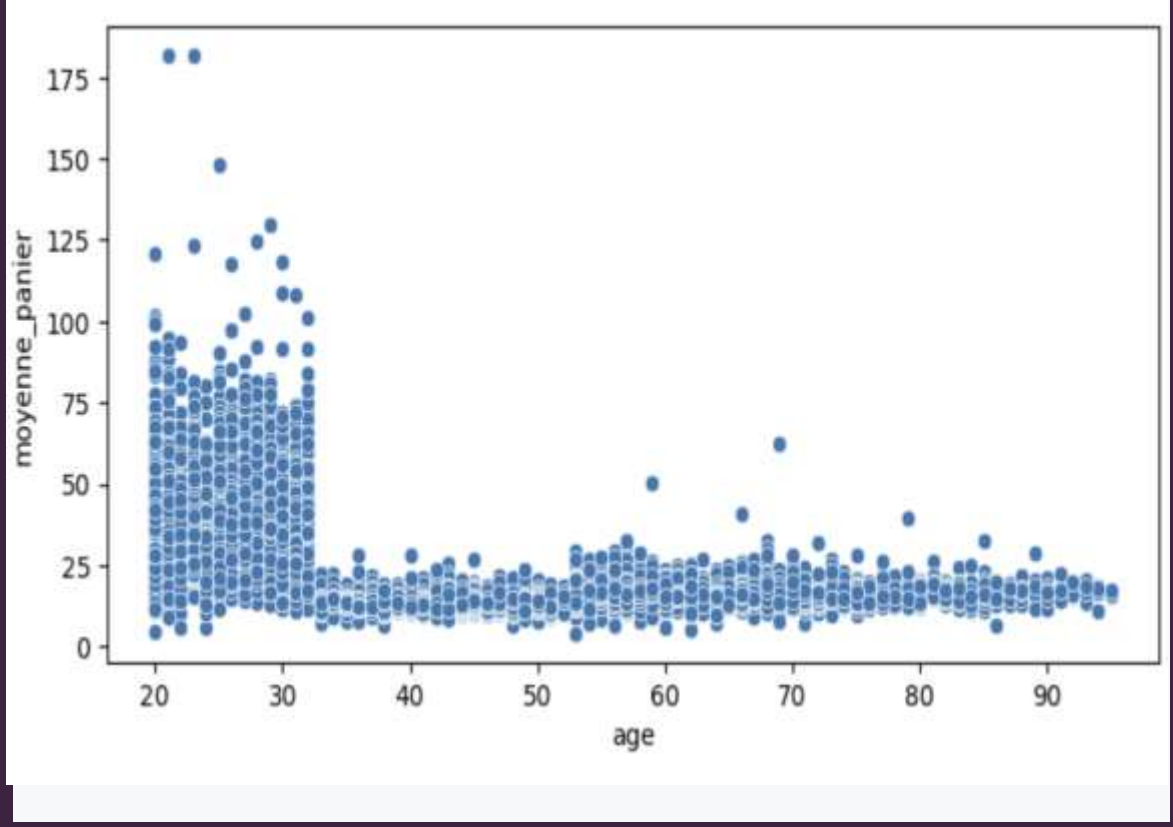
```
# Hypothèse :

# H0 : Il n'y a pas de corrélation entre les deux variables :  rs = 0

# Ha : Il y a une correlation entre les deux variables : rs != 0
```

```
# rs et p_value de Spearman

spearmanr_coefficient, p_value = spearmanr(df_age_freq["age"],df_age_freq["freq"])
print("Le coffecient de Spearman est %0.3f"% (spearmanr_coefficient))
print("La p_value est : ", p_value)
```

```
Le coffecient de Spearman est -0.676
La p_value est :  2.146632000345534e-11
```

```
# La Corrélation Spearman rejète l'hypothèse nulle et révèle une corrélation forte entre âge des clients et la fréquence de leurs acahats.
```

LIEN : AGE DES CLIENTS ET LA FREQUENCE D’ACHATS



Mann Whitney U Test

```
# Hypoythèse :

# H0 : Il n'y a pas de différence significative dans la moyenne de dépenses de tous les groupes (p_value est > 0.05)

# Ha : Il y a une différence significative dans la moyenne de dépenses de tous les groupes (p_value est < 0.05)
```

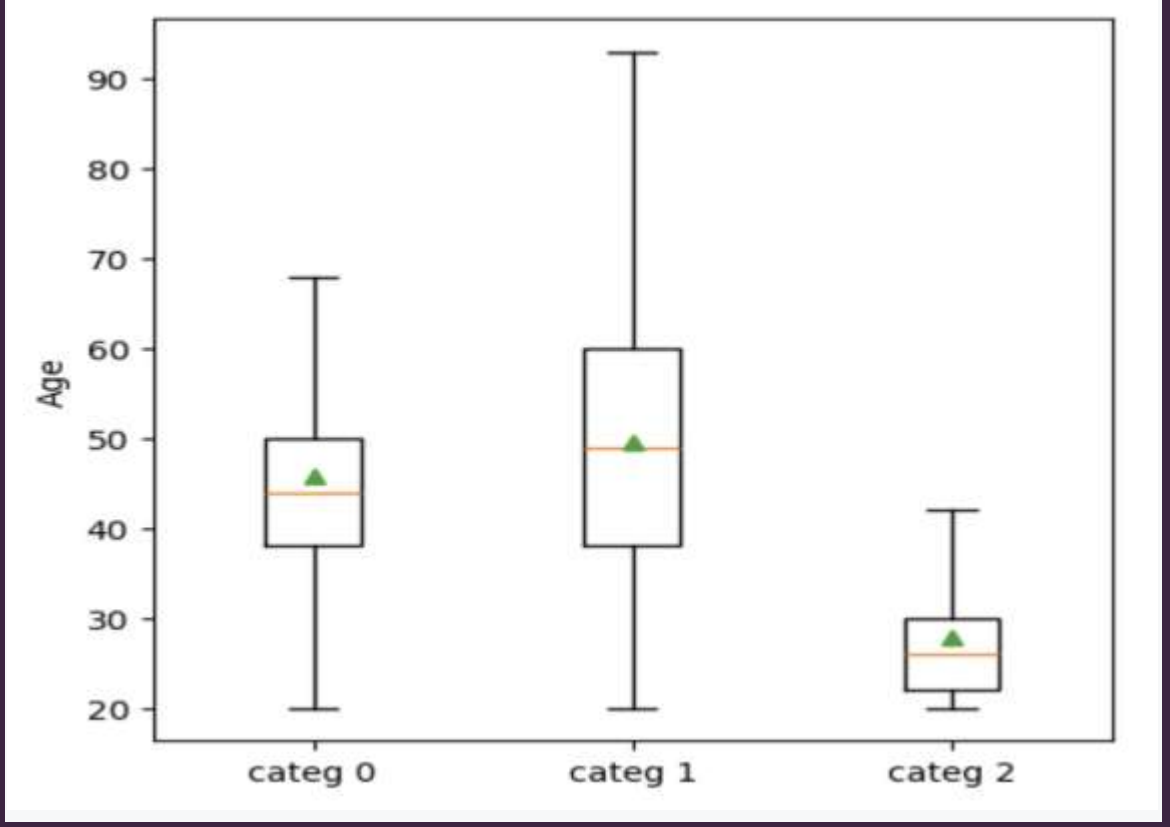
```
from scipy.stats import mannwhitneyu
```

```
# Unpacking

statistic, p_val = mannwhitneyu(df_panier_moy["price"], df_panier_moy["age"], use_continuity=True)

print(statistic)
print(p_val)
```

```
8387698.0
0.0
```

```
# P_value est < 0.05, donc on peut dire qu'il y a une différence significative dans le panier moyen et âge clients
```

LIEN : AGE DES CLIENTS ET LA TAILLE DU PANIER MOYEN



Kruskal–Wallis Test

```
# C'est un test statistique non paramétrique utilisé pour comparer trois groupes indépendants ou plus afin de
# déterminer s'il existe des différences statistiquement significatives entre eux.
# C'est une alternative à Anova Test
# Il est utilisé pour comparer deux ou plusieurs échantillons indépendants de tailles égales ou différentes
```

```
# Hypoythèse :

# H0 : Il n'y a pas de différence statistiquement significative dans les groups (p_value est > 0.05)

# Ha : Il y a une différence statistiquement significative dans les groups (p_value es < 0.05)
```

```
from scipy.stats import kruskal

kruskal(categ_age[0], categ_age[1], categ_age[2])
```

```
KruskalResult(statistic=71359.73412120914, pvalue=0.0)
```

```
stats, p = kruskal(categ_age[0], categ_age[1], categ_age[2])

if p_value > 0.05:
    print("On peut conclure quIl n'y a pas de différence statistiquement significative dans les groups")
else:
    print("On peut conclure qu'il a y une différence statistiquement significative dans les groups")
```

```
On peut conclure qu'il a y une différence statistiquement significative dans les groups
```

LIEN : AGE DES CLIENTS ET LA CATEGORIE DES LIVRES ACHETES