Lapage

Analyse des ventes

Pascal Brochart – Juin 2024

Contexte

- Analyser les différents indicateurs de vente de livres de la librairie Lapage après l'ouverture d'un site de vente en ligne
- Cibler le comportement des clients en ligne afin de comparer avec la connaissance acquise via les librairies physiques
- Analyser la relation entre 2 variables avec des tests statistiques

Analyse exploratoire des données

- Fichier de transactions:

id_prod	string	Clé étrangère de produits
date	datetime64[ns]	
session_id	string	Clé de sessions
client_id	string	Clé étrangère de clients

```
#Identifier les lignes vides
df_transactions.isnull().sum(axis = 0)
```

```
id_prod       361041
date          361041
session_id    361041
client_id     361041
dtype: int64
```

```
#Suppression des lignes vides
df_transactions.dropna(inplace=True)
```

Le fichier contient 687534 lignes non vides

Analyse exploratoire des données

- Fichier de produits:

```
id_prod      string     Clé de produits
price       float64
categ         int64
```

```
#Identifier les lignes vides
df_products.isnull().sum(axis = 0)
```

```
id_prod    0
price      0
categ      0
dtype: int64
```

```
#Vérifier si il y a les lignes en doublons dans la colonne id_prod
df_products.duplicated(subset='id_prod').sum()
```

```
0
```

Le fichier contient 3286 lignes non vides

Analyse exploratoire des données

- Fichier de clients:

```
client_id    string       Clé de clients
sex          string
birth         int64
```

```
#Identifier les lignes vides
df_customers.isnull().sum(axis = 0)
```

```
client_id    0
sex          0
birth        0
dtype: int64
```

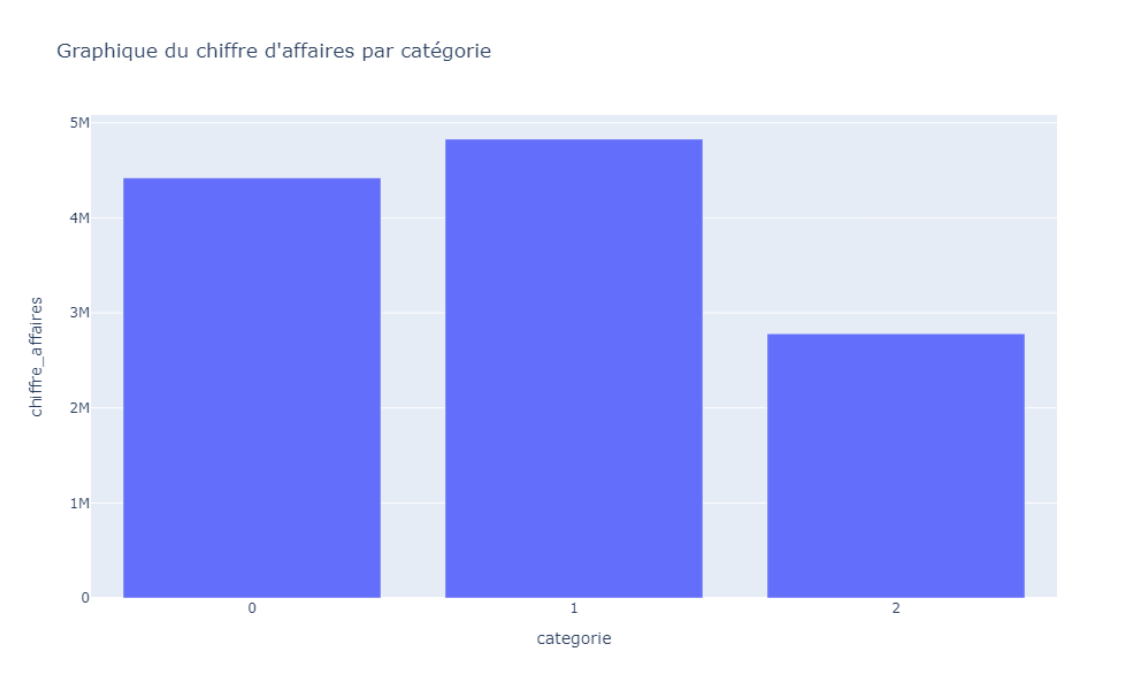
```
#Vérifier si il y a les lignes en doublons dans la colonne client_id
df_customers.duplicated(subset='client_id').sum()
```

```
0
```

Le fichier contient 8621 lignes non vides

Analyse autour du chiffre d'affaires

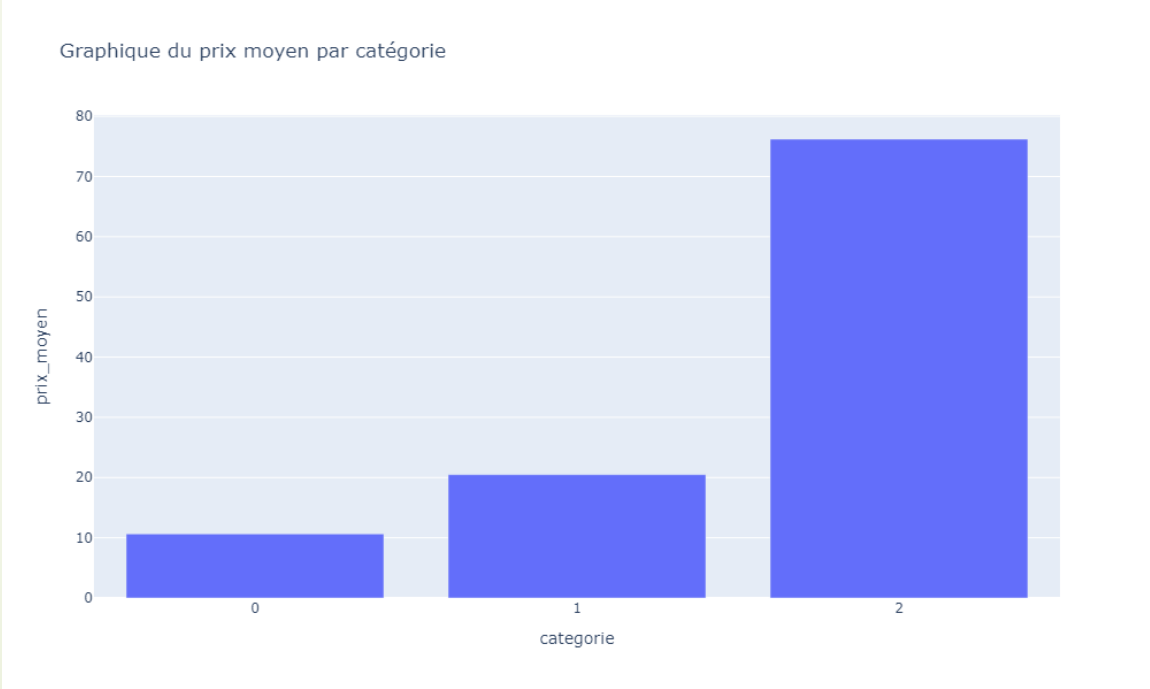
- Chiffre d'affaires par catégorie



Les produits de catégorie 2 ont un chiffre d'affaires de presque moitié moins que les produits de catégorie 1

Analyse autour du chiffre d'affaires

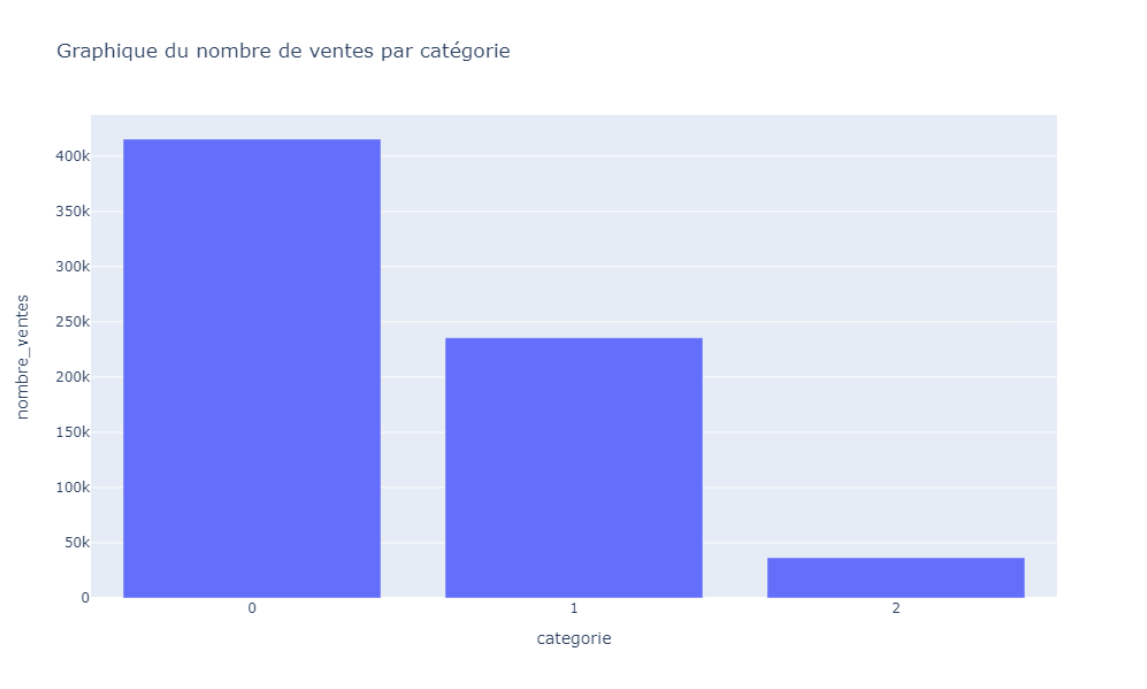
- Prix moyen par catégorie



En revanche le prix moyen des produits de catégorie 2 est près de 4 fois supérieur à celui de des produits de catégorie 1

Analyse autour du chiffre d'affaires

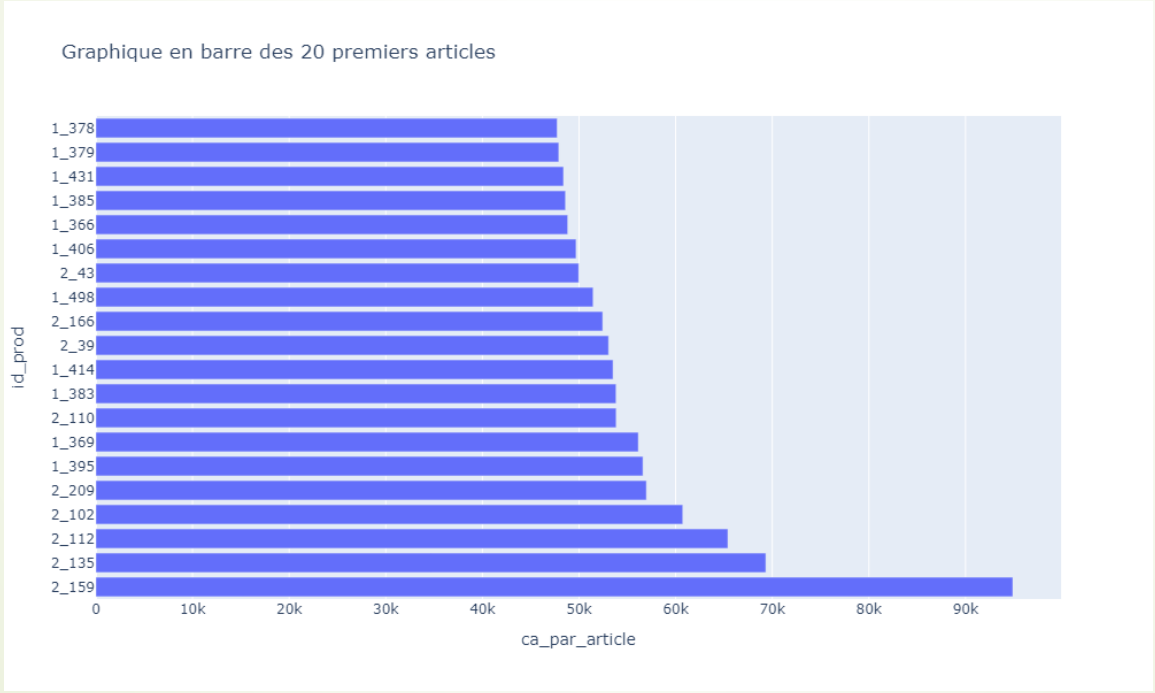
- Nombre de ventes par catégorie



Le nombre de ventes par catégorie est globalement à l'opposé du prix moyen par catégorie, plus le prix moyen est bas et plus le produit se vend

Analyse autour du chiffre d'affaires

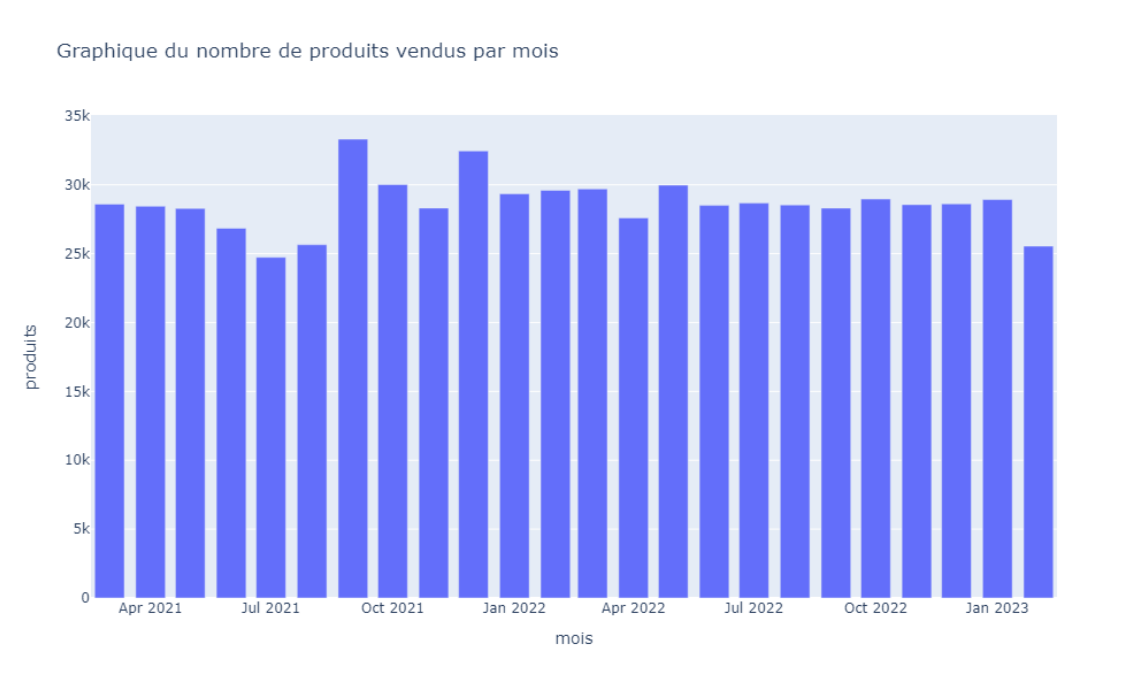
- Top 20 des articles en chiffre d'affaires



Les 4 premiers articles sont de catégorie 2 avec un prix unitaire assez élevé comme le produit 2_159 qui est vendu à 145€99

Analyse autour du chiffre d'affaires

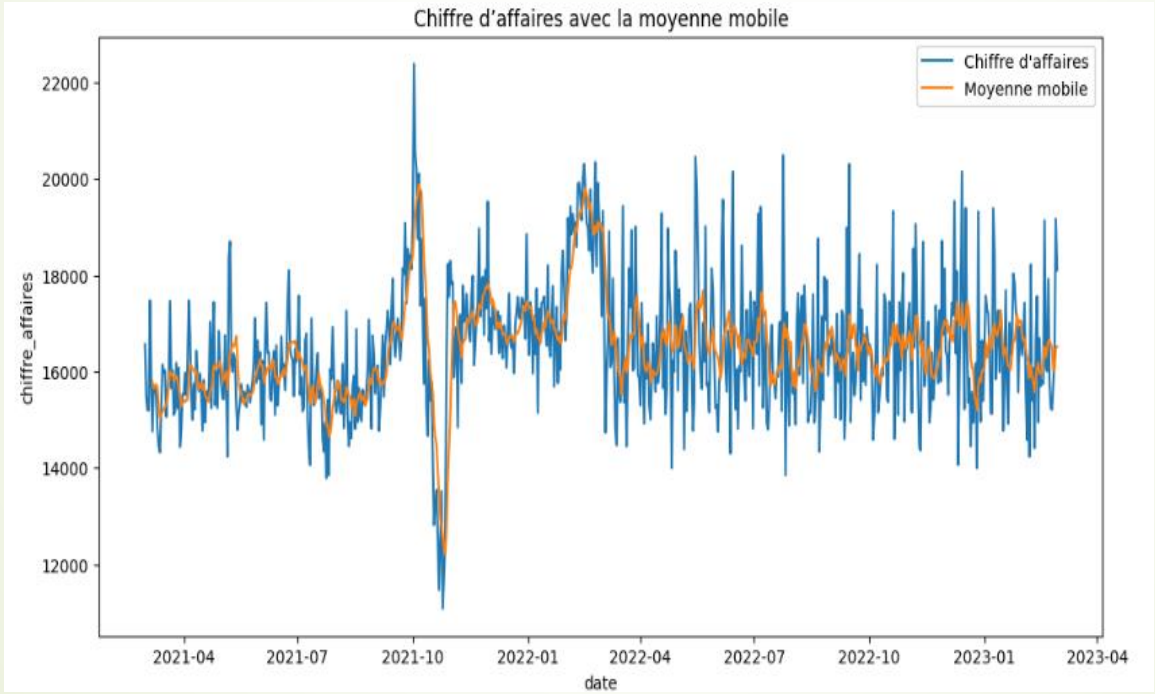
- Nombre de produits vendus par mois



Le nombre de produits vendus est constant sauf en période d'été 2021 avec une légère baisse puis une hausse à la rentrée et durant les fêtes de fin d'année

Analyse autour du chiffre d'affaires

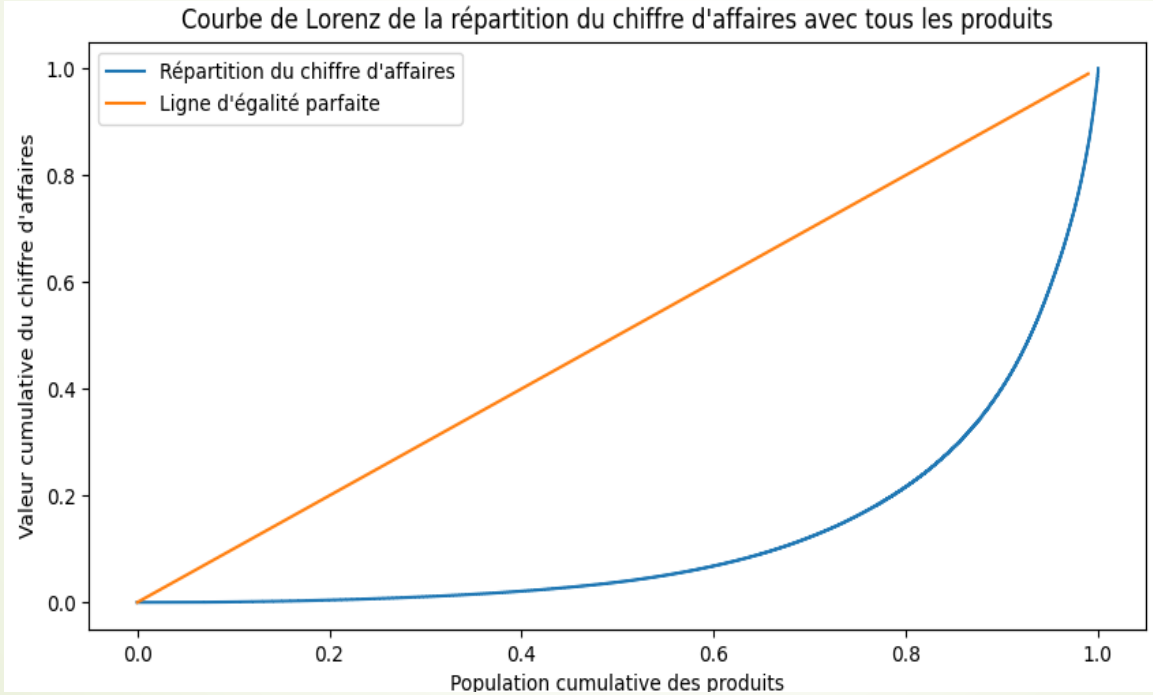
- Chiffre d'affaires avec la moyenne mobile



La moyenne mobile avec un intervalle de temps de 7 jours nous permet d'évaluer plus facilement les tendances avec les pics d'octobre 2021 et de février 2022

Analyse autour du chiffre d'affaires

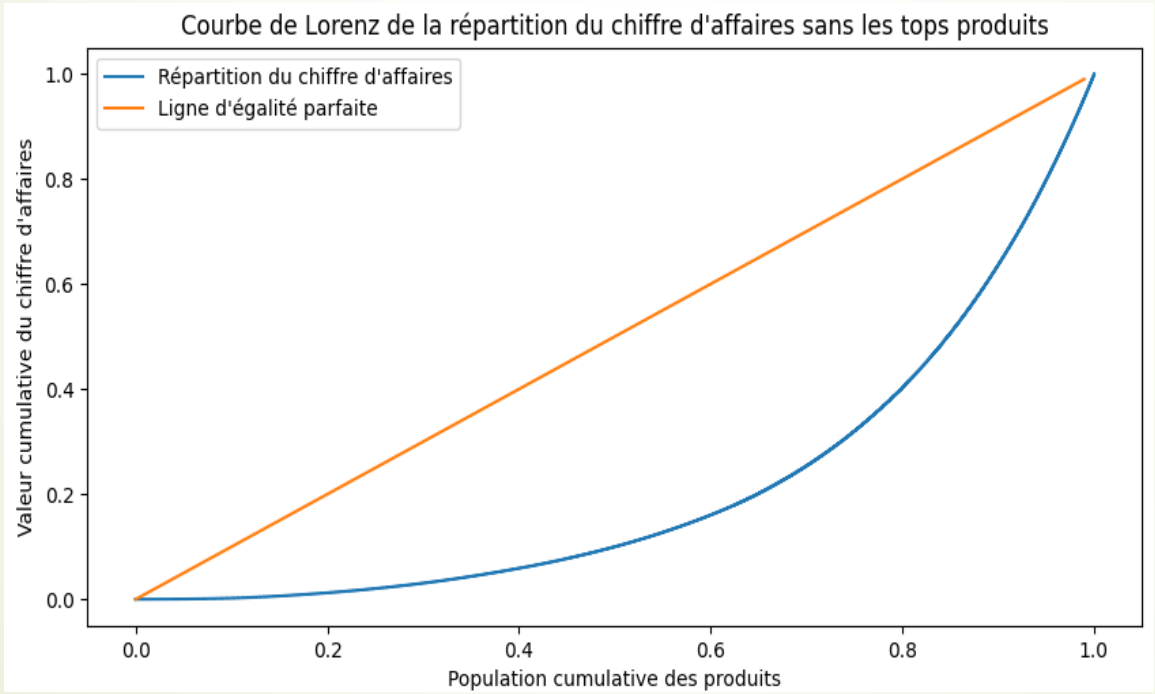
- Courbe de Lorenz avec tous les produits



On remarque une courbe bleue assez éloignée de la ligne d'égalité parfaite
Le coefficient de Gini est de 0.743997 ce qui confirme une forte inégalité des répartitions

Analyse autour du chiffre d'affaires

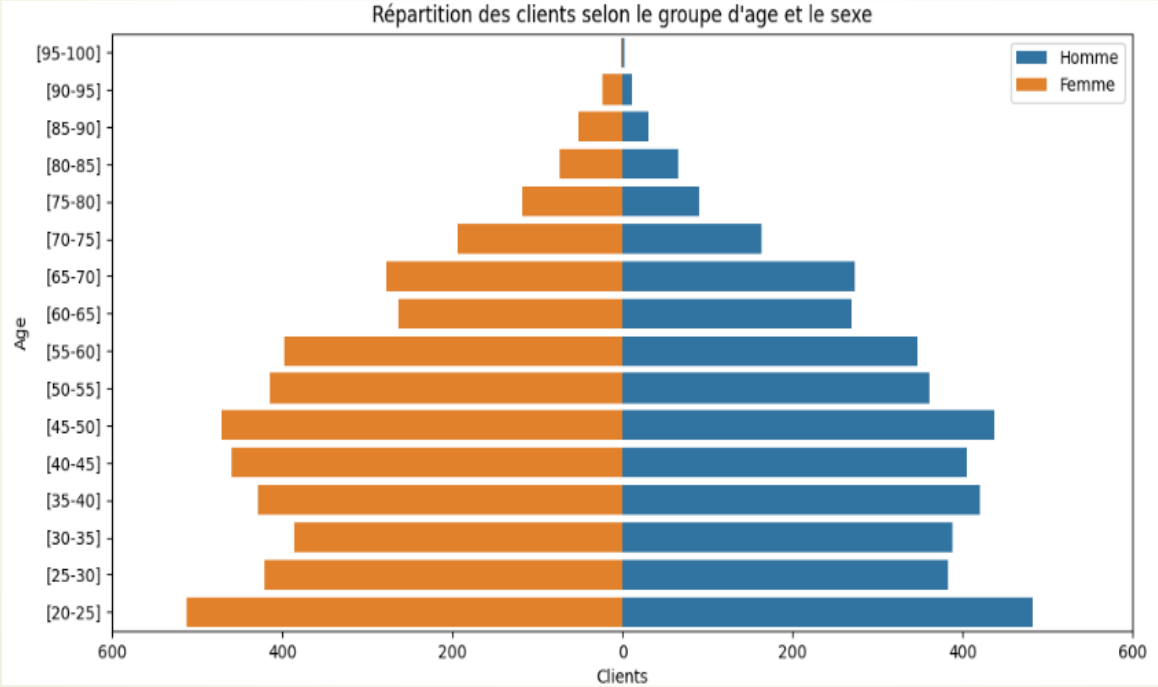
- Courbe de Lorenz sans les tops produits



La courbe bleue est beaucoup plus proche de l'égalité parfaite sans les tops produits (80/20)
Le coefficient de Gini est de 0.575621 ce qui confirme une meilleur égalité des répartitions

Analyse autour des clients

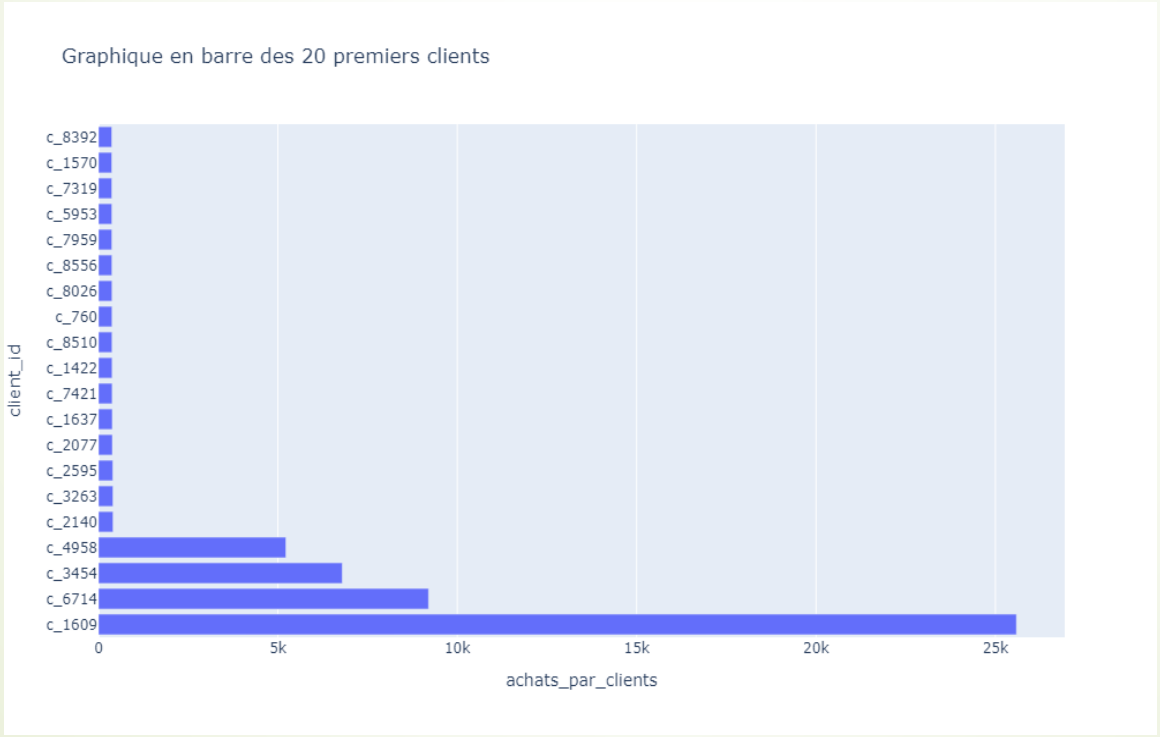
- Répartition des clients selon le groupe d'âge et le sexe



La tranche d'âge des 20-25 ans est la plus importante avec les 45-50 ans et la répartition par sexe est globalement égale

Analyse autour des clients

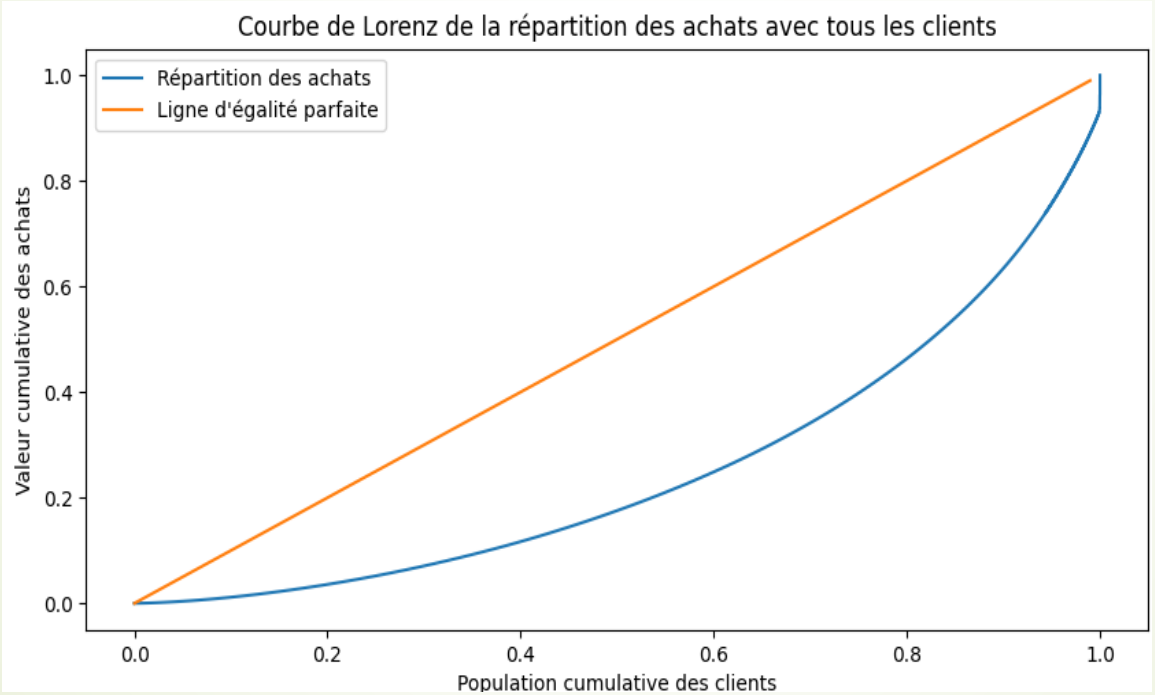
- Top 20 des clients par achats



4 clients se démarquent très largement des autres en terme d'achats et l'on peut considérer qu'il s'agit de clients professionnels

Analyse autour des clients

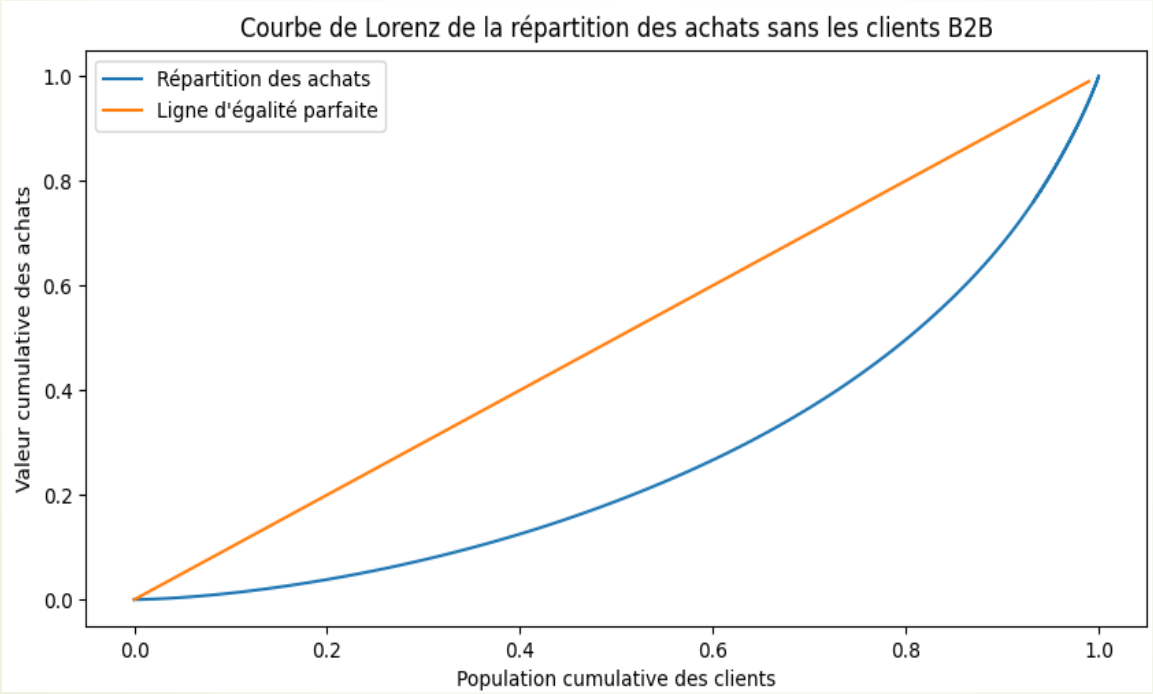
- Courbe de Lorenz avec tous les clients



On remarque un décrochage en fin de courbe correspondant à l'inégalité de la répartition des achats provoqués par les clients professionnels

Analyse autour des clients

- Courbe de Lorenz sans les clients B2B



La suppression des clients professionnels permet de rétablir une meilleure égalité de la répartition des achats

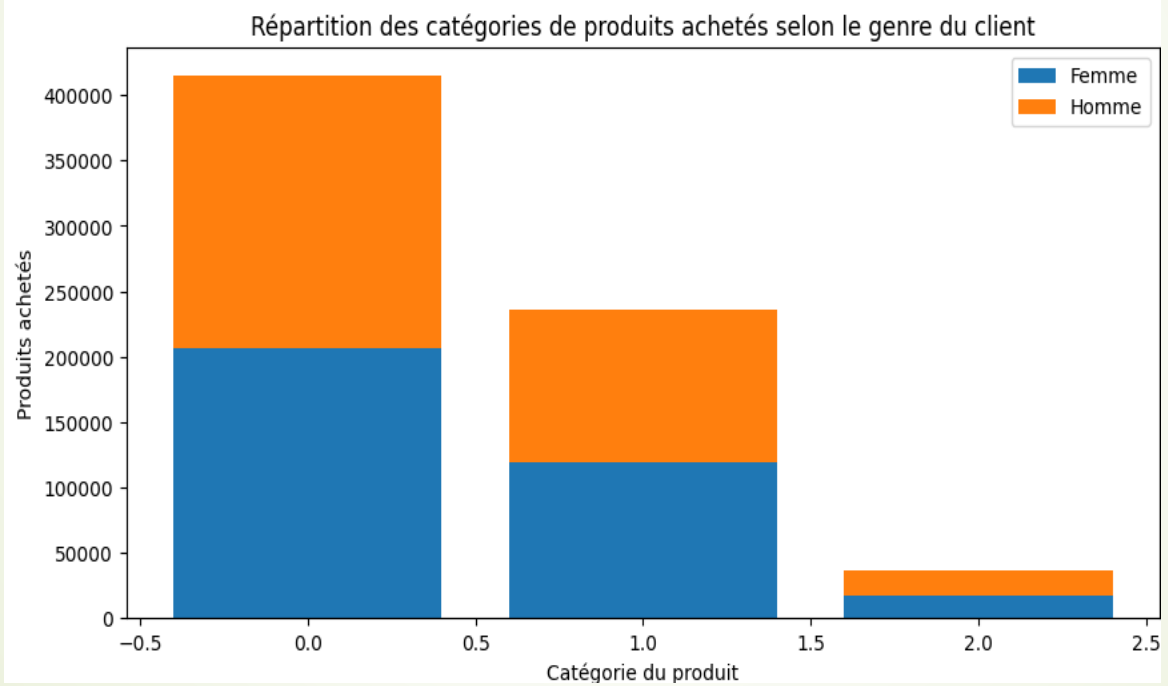
Analyse autour des clients

- Coefficient de Gini
  - Mesure de l'éloignement entre les courbes orange et bleue
  - Plus la courbe bleue se rapproche de l'égalité parfaite et plus l'indice (ou le coefficient) de Gini va se rapprocher de 0
  - A l'inverse, plus la répartition des achats est inégalitaire et plus la courbe bleue va s'éloigner de la courbe orange et l'indice de Gini va augmenter
  - Les résultats ci-dessous confirment cette analyse:

- Le coefficient de Gini avec tous les clients est de: 0.493856
- Le coefficient de Gini sans les clients B2B est de: 0.457587

Analyse bivariée et tests statistiques

- Répartition des catégorie de produits achetés et le genre du client

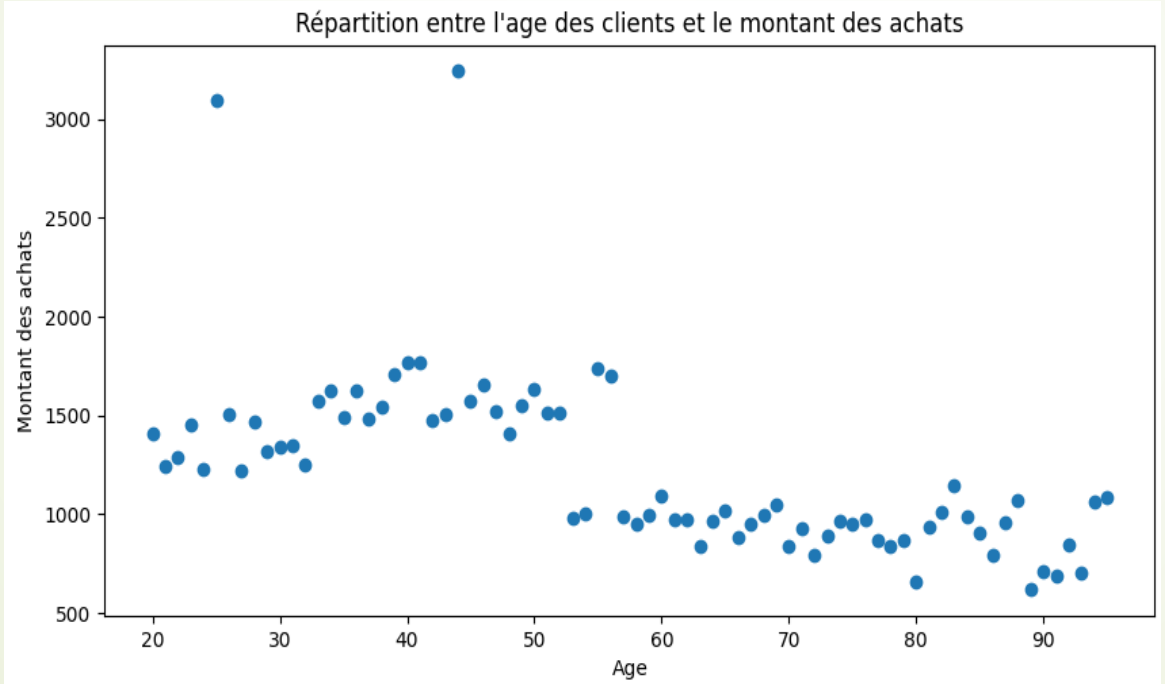


Le graphique ne semble pas montrer une corrélation entre les 2 variables qualitatives
Mais le test Khi2 donne le résultat contraire:

```
La valeur p pour le test Khi2 est de: 4.3205822283997063e-35
H1 : Variables non indépendantes si p-value < 5%
```

Analyse bivariée et tests statistiques

- Répartition entre l'âge des clients et le montant des achats

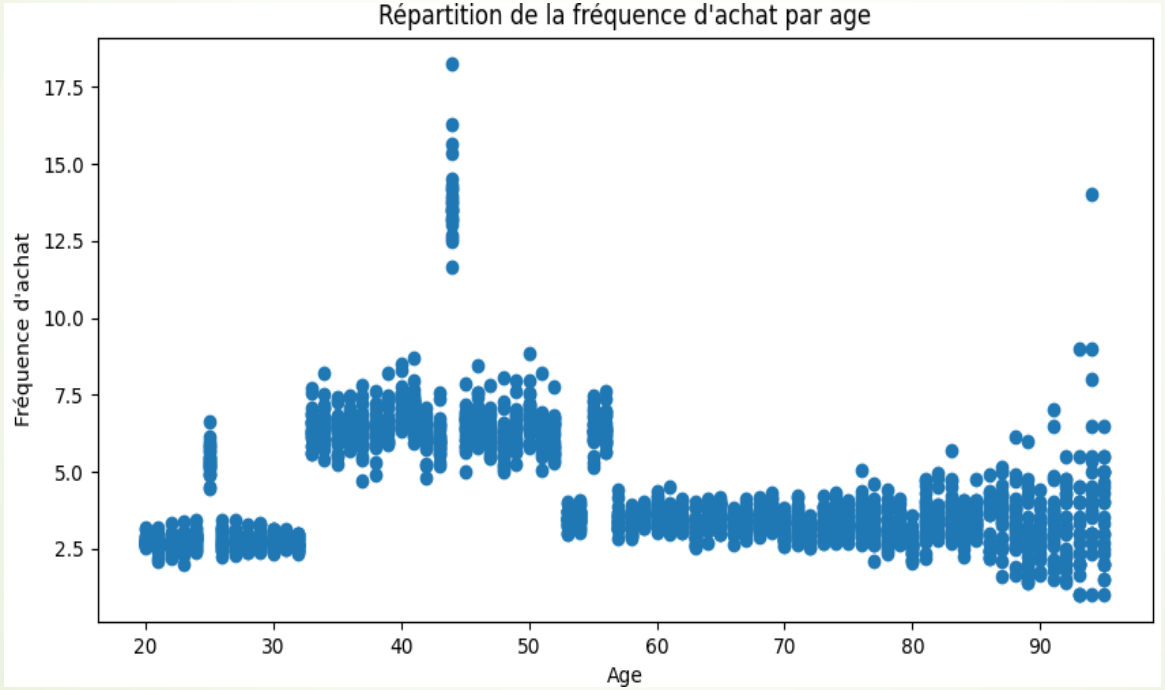


Le graphique montre une corrélation entre les 2 variables quantitatives
Après l'échec du test de normalité Shapiro
Le test Spearman confirme cette observation:

```
La valeur du coefficient de corrélation rs pour le test Spearman est de: -0.1843794612560274
La valeur p pour le test Spearman est de: 1.2370988979980202e-66
H1 : Variables non indépendantes si p-value < 5%
```

Analyse bivariée et tests statistiques

- Répartition de la fréquence d'achats par âge

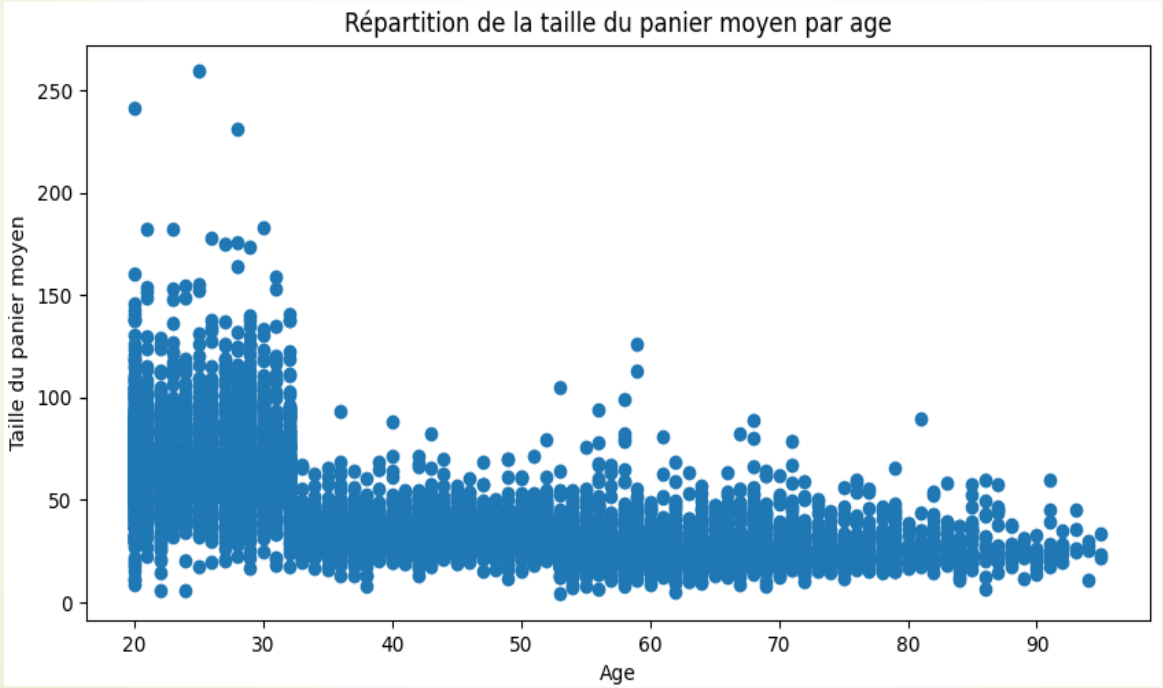


Le graphique montre une corrélation entre les 2 variables quantitatives
Après l'échec du test de normalité Shapiro
Le test Spearman confirme cette observation:

```
La valeur du coefficient de corrélation rs pour le test Spearman est de: -0.14711768830331576
La valeur p pour le test Spearman est de: 2.732867581398287e-10
H1 : Variables non indépendantes si p-value < 5%
```

Analyse bivariée et tests statistiques

- Répartition de la taille du panier moyen par âge

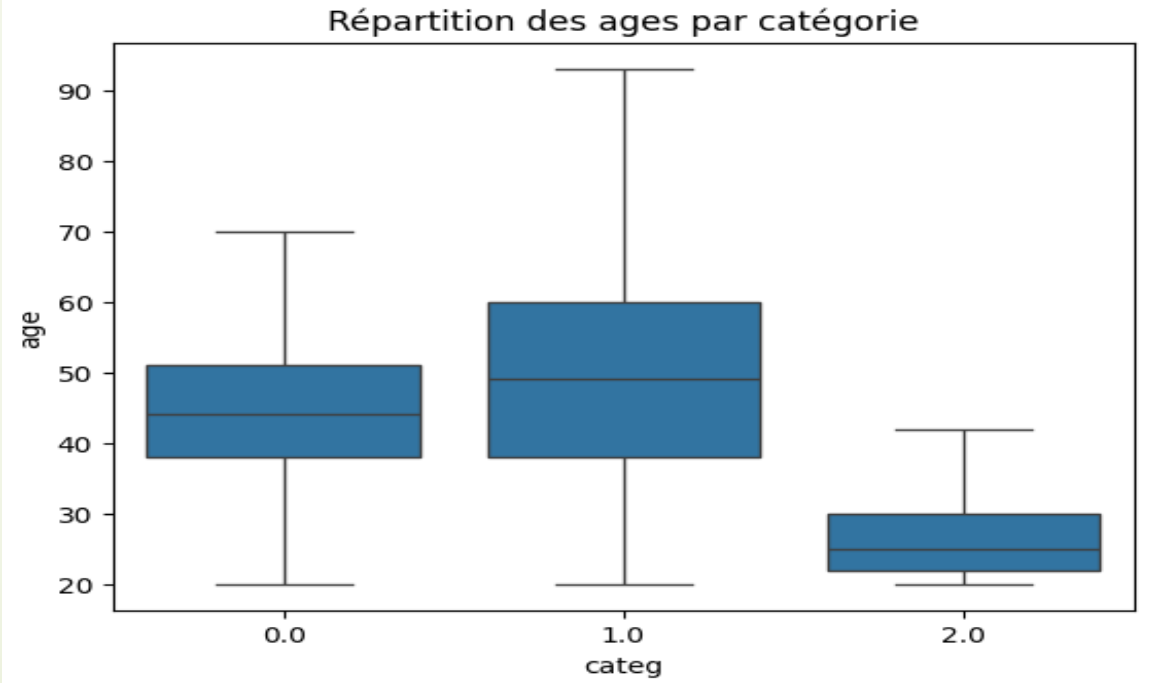


Le graphique montre une corrélation entre les 2 variables quantitatives
Après l'échec du test de normalité Shapiro
Le test Spearman confirme cette observation:

```
La valeur du coefficient de corrélation rs pour le test Spearman est de: -0.7004399996375055
La valeur p pour le test Spearman est de: 0.0
H1 : Variables non indépendantes si p-value < 5%
```

Analyse bivariée et tests statistiques

- Répartition des âges par catégorie



Nous allons tenter une analyse de la variance (ANOVA) entre une variable quantitative et qualitative mais des conditions préalables doivent être remplies

Analyse bivariée et tests statistiques

- ✓ Test normalité des résidus Shapiro

```
La valeur p pour le test Shapiro est de: 1.0
H0 : Les résidus suivent une loi normale si p-value > 5%
```

- × Test homoscédasticité Barlett et Leven

```
La valeur p pour le test Barlett est de: 5.334304290023777e-50
H1 : Les variances de chaque groupe ne sont pas toutes égales < 5%
```

```
La valeur p pour le test Levene est de: 1.6560622800601952e-72
H1 : Les variances de chaque groupe ne sont pas toutes égales < 5%
```

- ➡ Conditions non remplies pour une ANOVA
- × Test Kruskall-Wallis

```
La valeur p pour le test Kruskal-Wallis est de: 0.0
H1 : Au moins deux groupes ont des rangs moyens différents si p-value < 5%
```

Le test final de Kruskall-Wallis n'est pas satisfait dans le sens où des catégories ont des tranches d'âges moyens différents