



RESTER LIVRE

Analyse des ventes

OPENCLASSROOMS

ENSAE-ENSAI
Formation continue
[Cepe]



1. Introduction

Contexte et objectifs.....p 4

2. Préparation et exploration des données

MISSION1: Gestion des doublons et valeurs atypiques.....p 6

Gestion de la variable qualitative ordinale date.....p 7

Gestion des valeurs manquantes.....p 8

Gestion des valeurs aberrantes ou outliers.....P 10

3. Modélisation

MISSION2: La clientèle.....p 12

Les catégories et les produits.....p 13

Fréquentation du site web.....p 14

Les ventes..... p 15

MISSION3: Questions du manager.....p 19

4. Synthèse des résultats et recommandations

Synthèse.....p 25

Recommandations.....p 26

1. Introduction

Contexte et objectif de l'analyse

Notre boutique en ligne rencontrant un franc succès, cette analyse effectuée à l'aide d'outils issus de la data science, tend à rendre compte de notre activité de vente en ligne et de produire des recommandations afin d'optimiser les performances de ventes en ligne.

2. Préparation et exploration des données

Mission 1

Gestion des doublons et valeurs atypiques

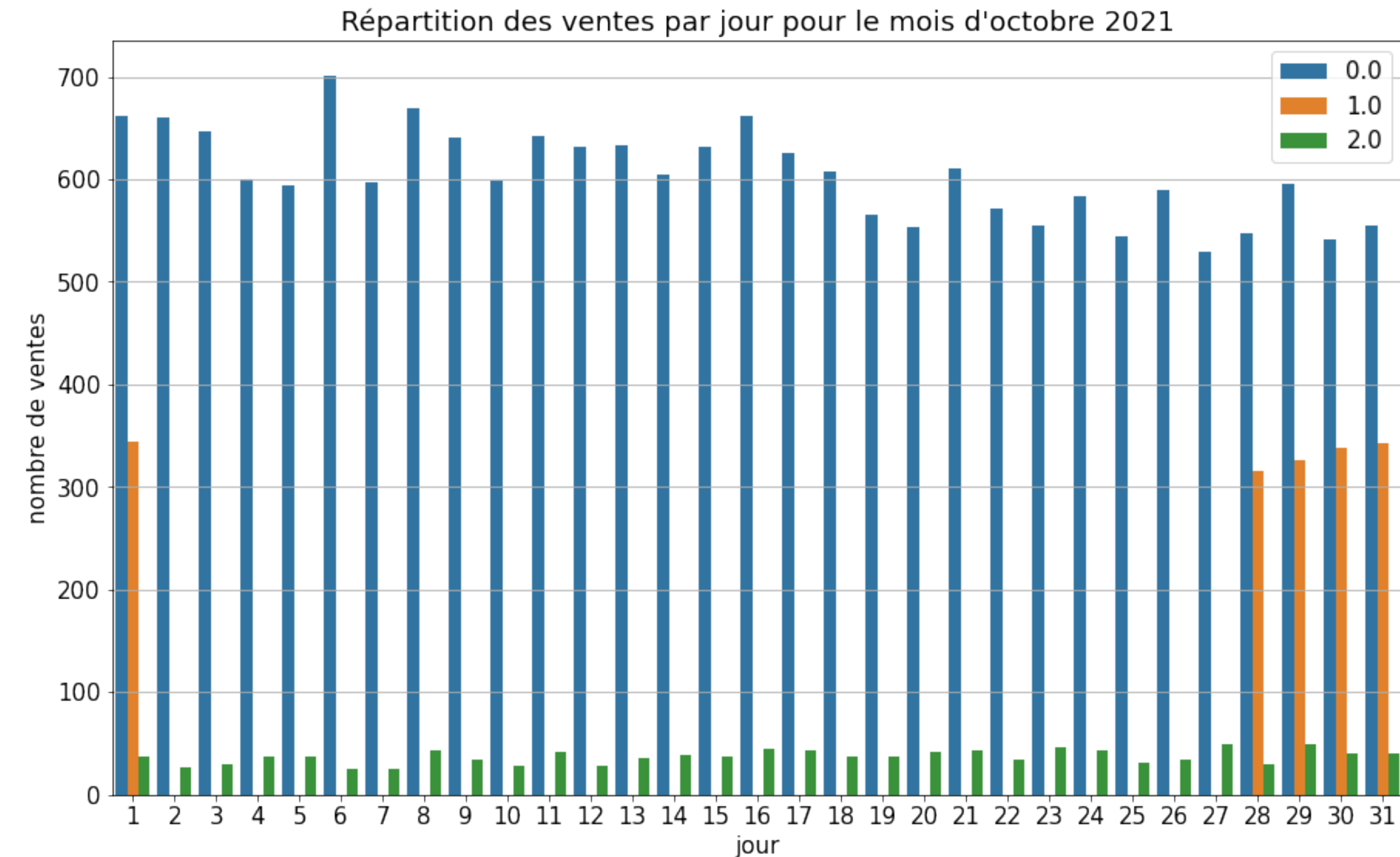
 Recherche de doublons.

 Découverte de valeurs atypiques dans date sur 1 session et concernant 1 produit.

 Regroupement dans un échantillon pour conserver les données de test.

 ~~Suppression totale de l'échantillon du dataframe principale.~~

Gestion de la variable qualitative ordinale date



🔍 Exploration du comportement de la variable date.

✅ Découverte d'une absence de vente categ 1.

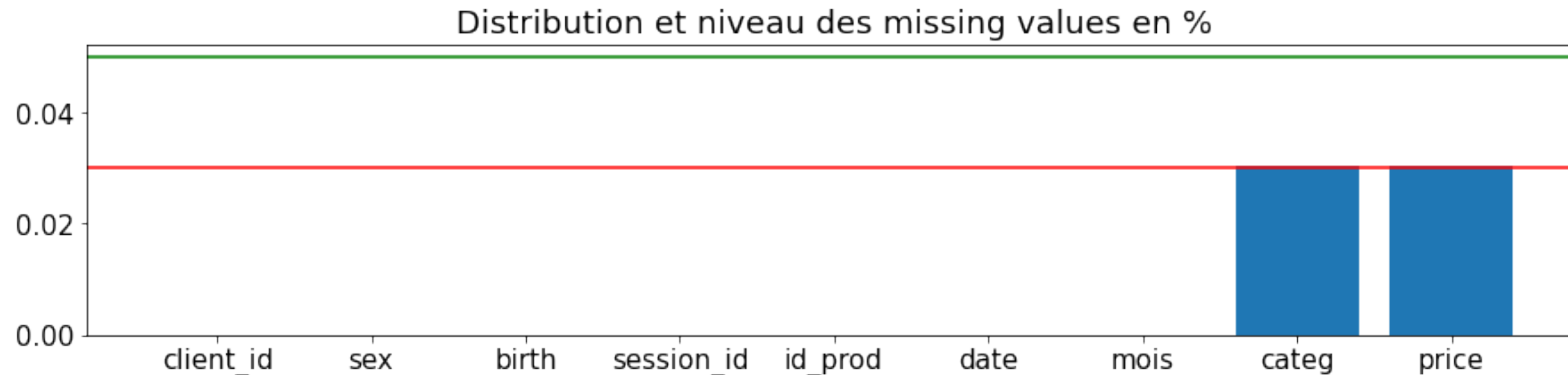
🎓 ou 🏠 ? Hypothèse rupture de stock (période de rentrée scolaire / universitaire) ou perte d'information (mouvements bancaires données ont été supprimées de la BDD).

🧩 ou 🧩❌ IMPUTER ou IGNORER ?



🧩❌ Démarche de moindre mal, IGNORER.


Gestion des valeurs manquantes

- 🔍 Exploration de la cohérence des modalités des variables qualitatives nominales et quantitatives continues.
- ✅ Découverte des 206 valeurs manquantes dans price et categ dû au produit '0_2245' .

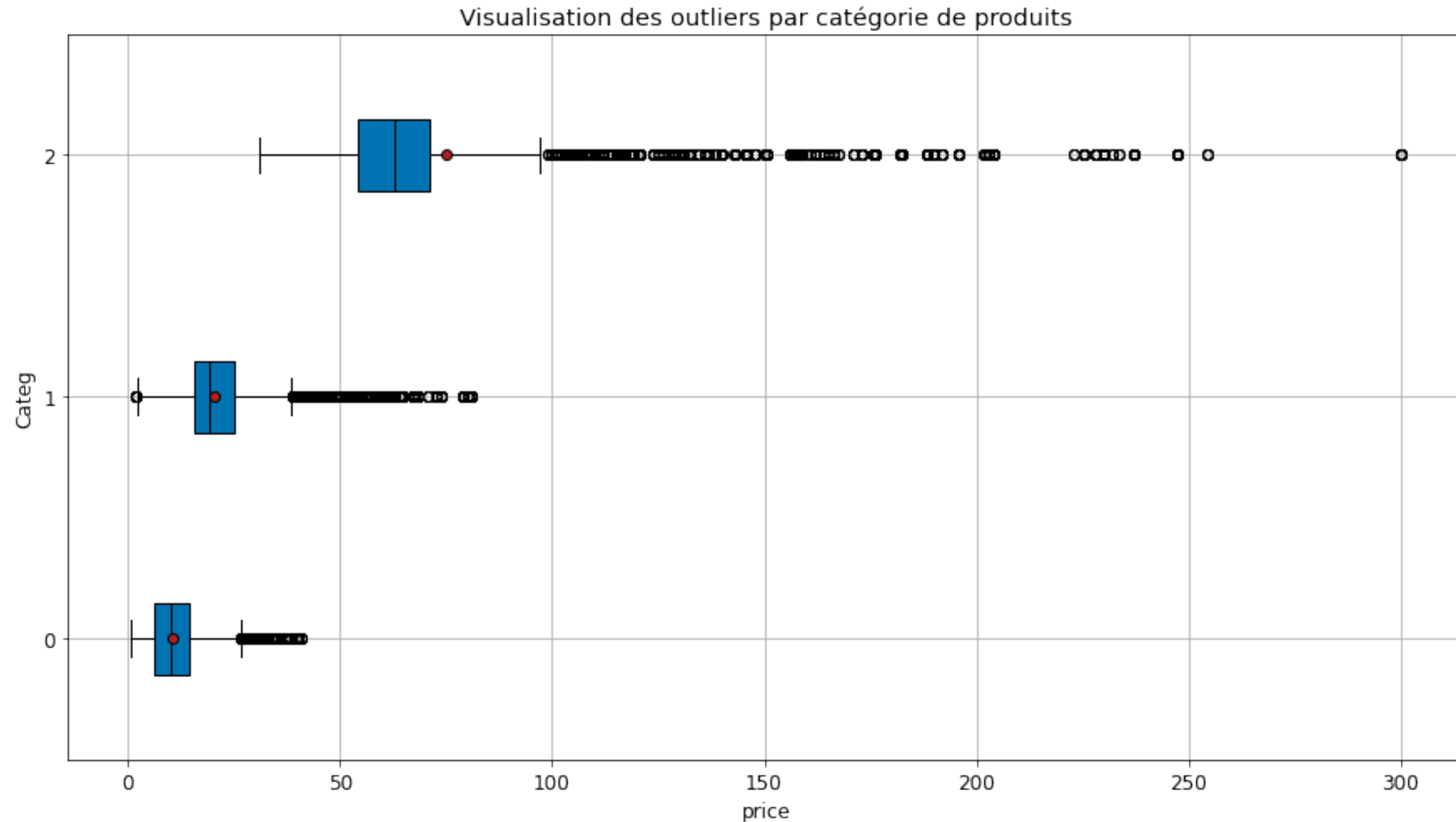


- 🔬 Analyse factorielle des correspondances (AFC) variables categ et price, puis avec les autres variables pour déterminer le type de manquant .
- 🚩 Résultat manquants type MAR car la probabilité d'absence est liée à plusieurs autres variables observées .

  Eu égard à la distribution asymétrique de la variable price , imputation algorithmique
 KNNImputer de la librairie SckitLearn.

  Eu égard au type de variable de categ (qualitative) Imputation algorithmique SimpleImputer
 de la librairie SkiLearn avec une stratégie « most_frequent ».

Gestion des valeurs aberrantes ou outliers



👁 Visualisation de la dispersion de price par categ.

📈 Comportement propre au secteur d'activité.

✅ Conserver tous les outliers afin de ne pas affecter la représentativité des données.

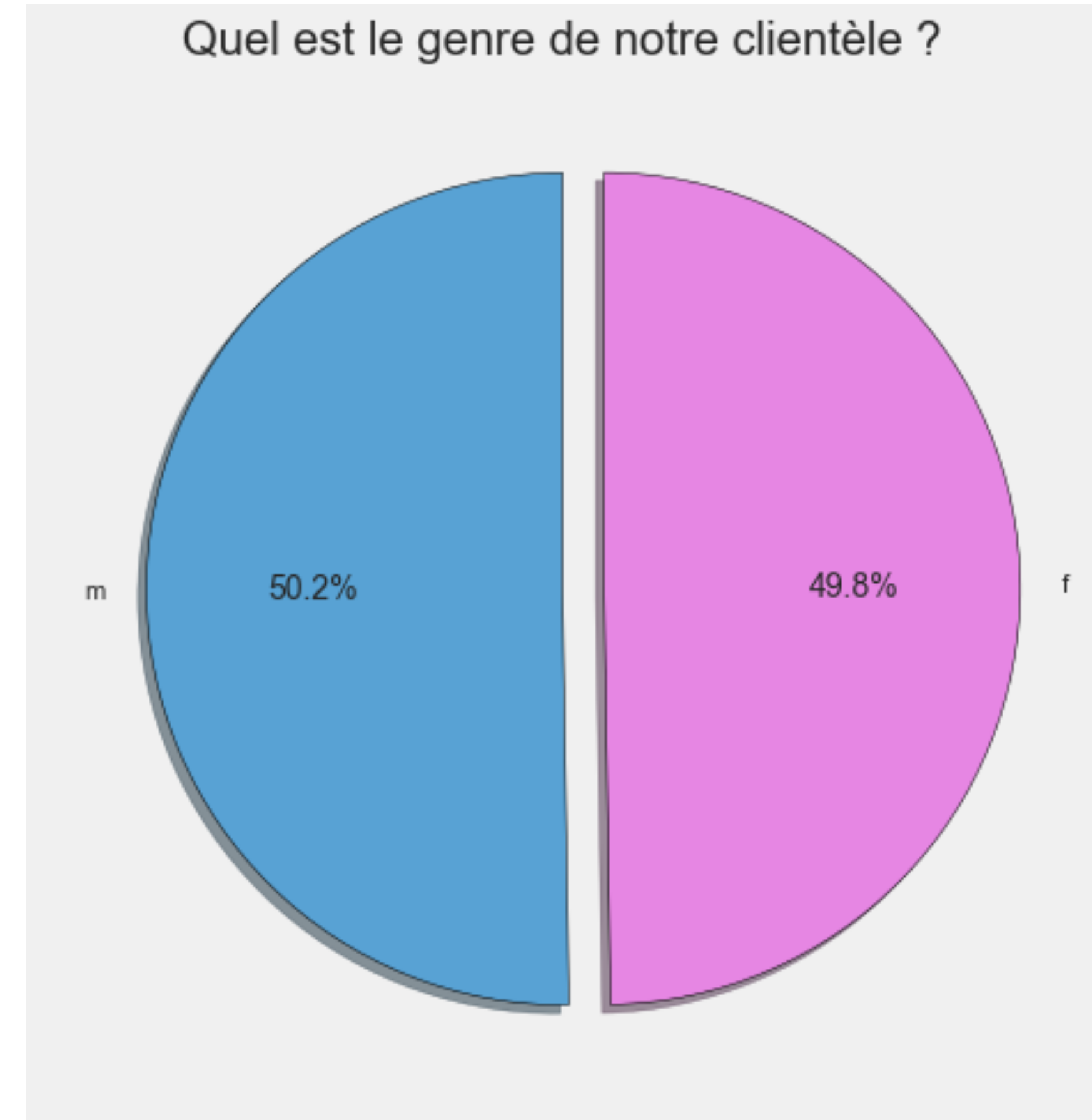
⚠ Prudence quant à l'utilisation d'outils et raisonnement conçus pour une distribution normale.

3. Modélisation

Mission 2

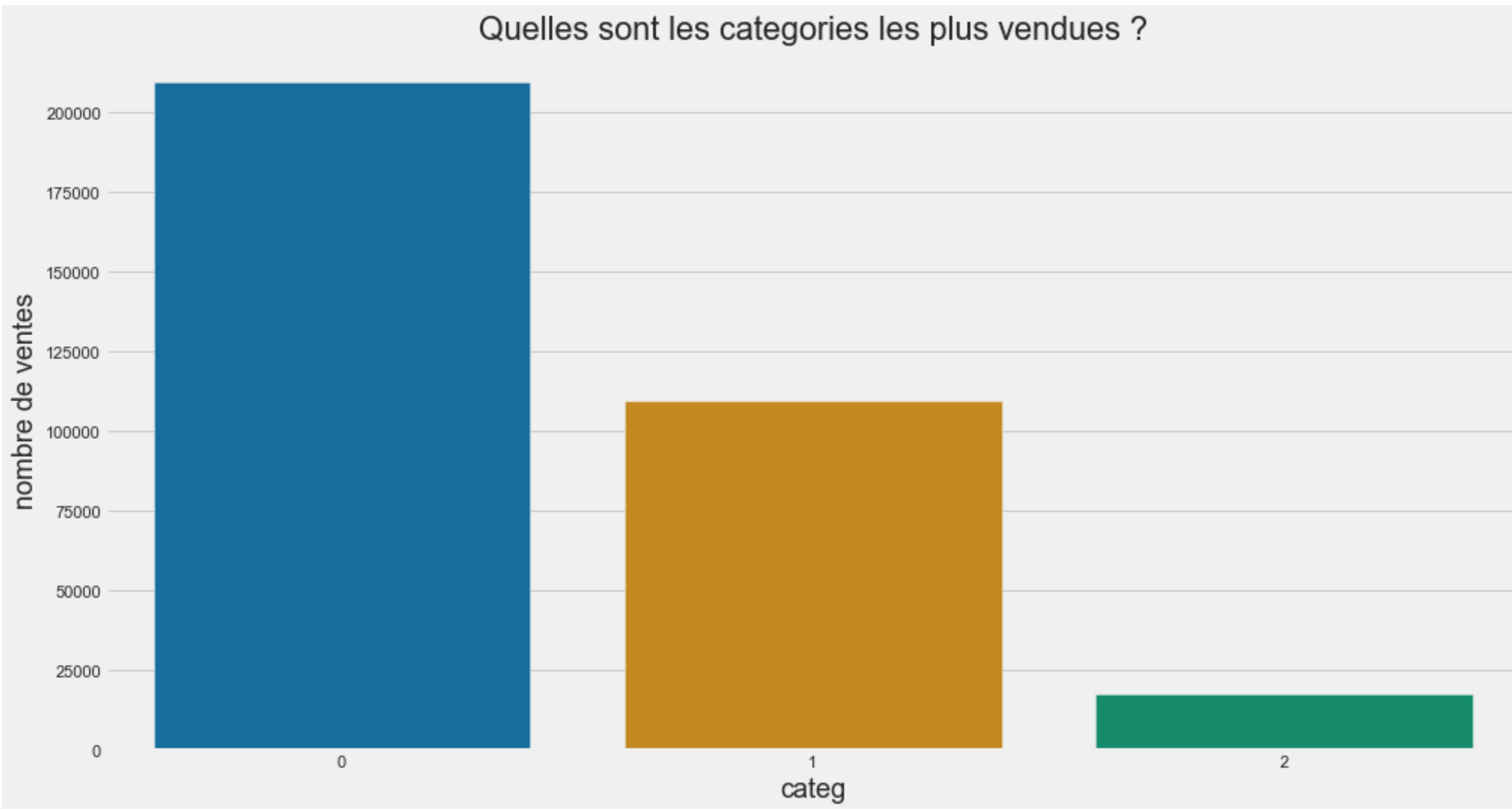
La clientèle

- ❖ L'entreprise compte 8600 clients, l'âge moyen de l'ensemble de la clientèle est de 43,17 ans.
- ❖ Une clientèle mixte 50,2% de m et 49,8% de f avec une répartition équilibrée par catégorie de produits et la moyennes d'âge est de 44,74 ans pour les femmes et 43,62 pour les hommes.
- ❖ Une segmentation des âges par catégorie de produits, on a pour $\text{categ0} = 43,90$ ans, pour $\text{categ1} = 47,66$ ans et pour $\text{categ2} = 25,73$ ans.



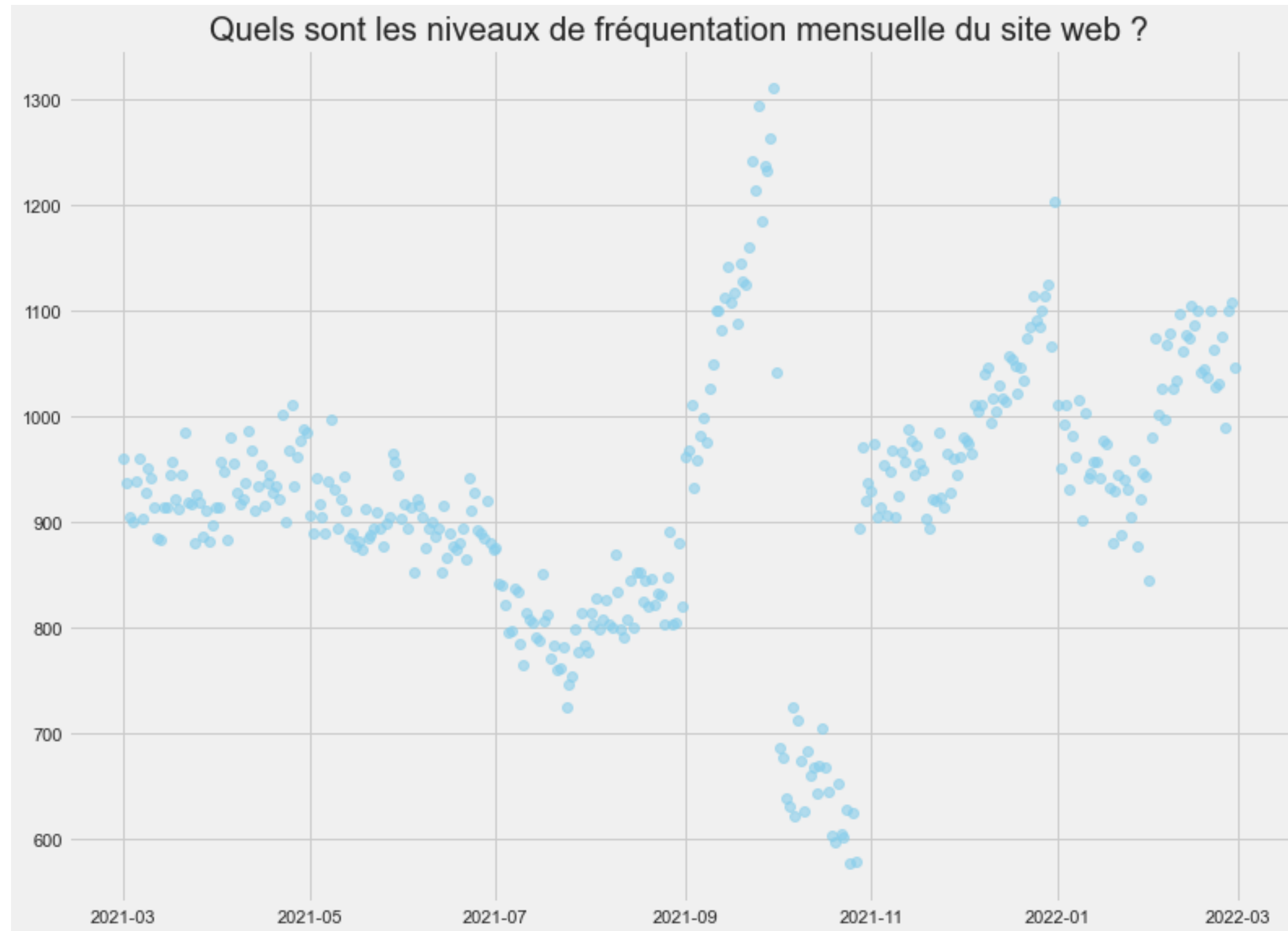
les catégories et les produits 📖

Quelles sont les categories les plus vendues ?



- ❖ Le site de e-commerce propose 3 265 produits .
- ❖ C'est aussi 3 catégories de produits, catégorie 0 = plus de 200 000 ventes, catégorie 1 = plus de 100 000 ventes et la catégorie 2 = moins de 25 000 ventes.
- ❖ Des produits vendus plus de 1000 fois tandis que d'autres n'enregistre qu'1 ou 2 ventes.

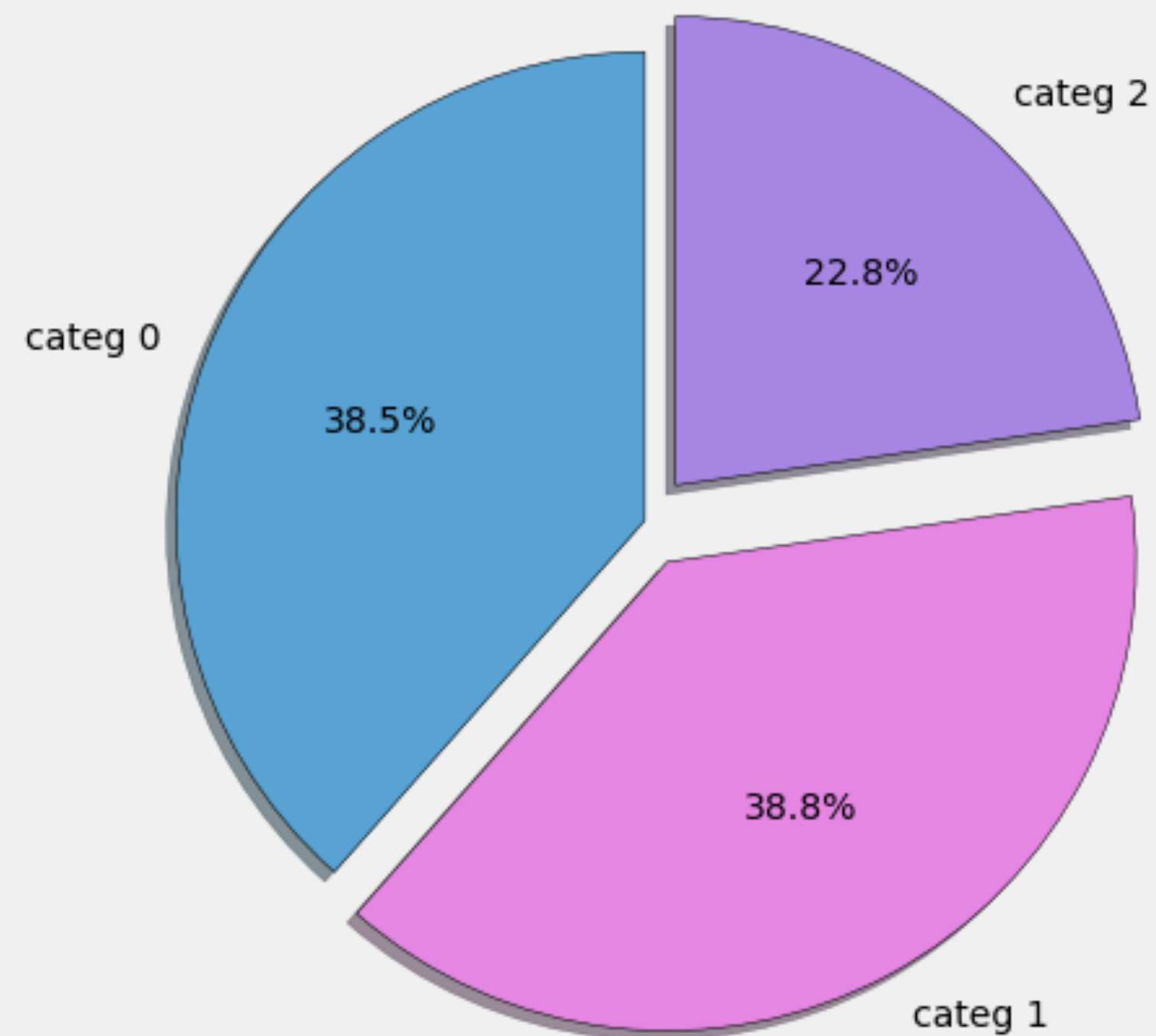
Fréquentation du site web @



- ❖ fréquentation mensuelle site de e-commerce comprise entre 800 et 1100 visites par jour.
- ❖ Forte affluence en septembre 2021 qui coïncide avec les rentrées scolaires / universitaires.
- ❖ Très basses périodes d'affluence juste après le pic de la rentrée à savoir octobre 2021.

Concernant les ventes

Quelle est le niveau de participation par catégorie à la formation du CA ?



- ❖ L'ensemble des ventes est de 5 798 350,93 €, soit pour categ 0 = 2 231 496 €, categ 1 = 2 247 384 € et categ 2 = 1 319 471 €.
- ❖ Les participations au CA (chiffre d'affaire) : pour la categ 0: 38,5%, categ 1: 38,8% et 22,8% pour la categ 2 .
- ❖ La concentration des transactions n'est pas négligeable car elles sont réparties de manière peu égalitaire. L'indice de Gini nous indique que 39% d'individus de la variable price concentre les ventes.

Les mesures de distribution et de dispersion des ventes :

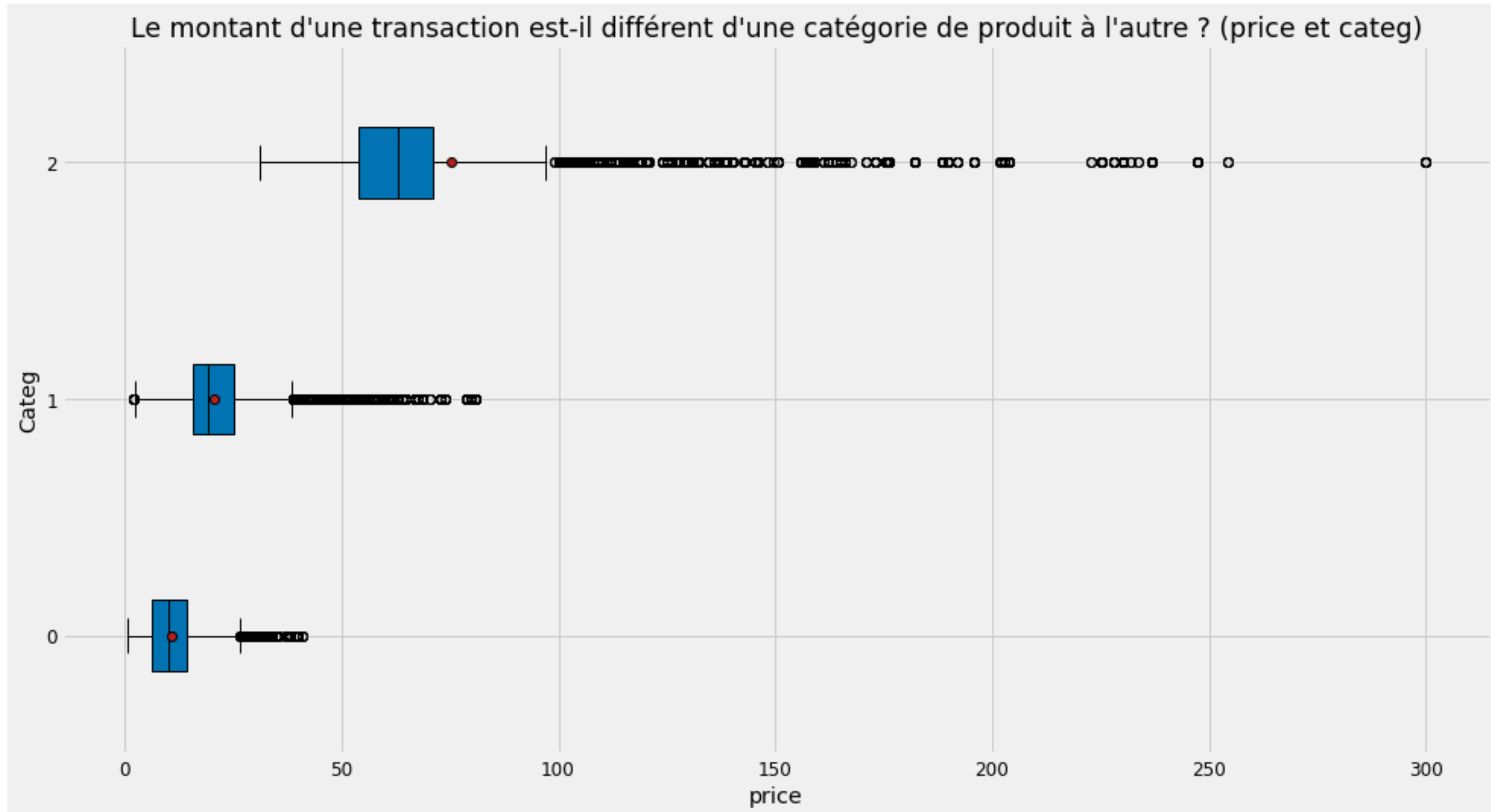
- ❖ Les 3 catégories inclinaison positive, «biaisées à droite», kurtosis leptokurtic pour les catégories 1 et 2.
- ❖ Une asymétrie le mode est inférieur à la médiane et elle même inférieure à la moyenne $Mode < Med < \bar{x}$.
- ❖ Forte dispersion catégorie 2 , modéré pour la catégorie 1 puis légère catégorie 0. Ces résultat sont dues aux outliers.
- ❖ La p-value est hautement significative, on rejette l'hypothèse H_0 , les catégories n'ont pas des variances égales.
- ❖ La p-value est hautement significative on rejette l'hypothèse H_0 , les catégories n'ont pas de moyennes égales.



Interprétations des résultats :

- ❖ Les ventes par catégories de produits sont hétérogènes.
- ❖ Les catégories de produits segmentés par niveaux de prix.
- ❖ Le point commun des catégories, les ventes pour ces 3 catégories sont en dessous de la moyenne.
- ❖ Performances liées aux ventes, la catégorie 1 affiche un résultat quasi similaire à la catégorie 0 malgré un volume de vente deux fois inférieur à la catégorie 0 et la catégorie 2 faible résultat de vente/volume

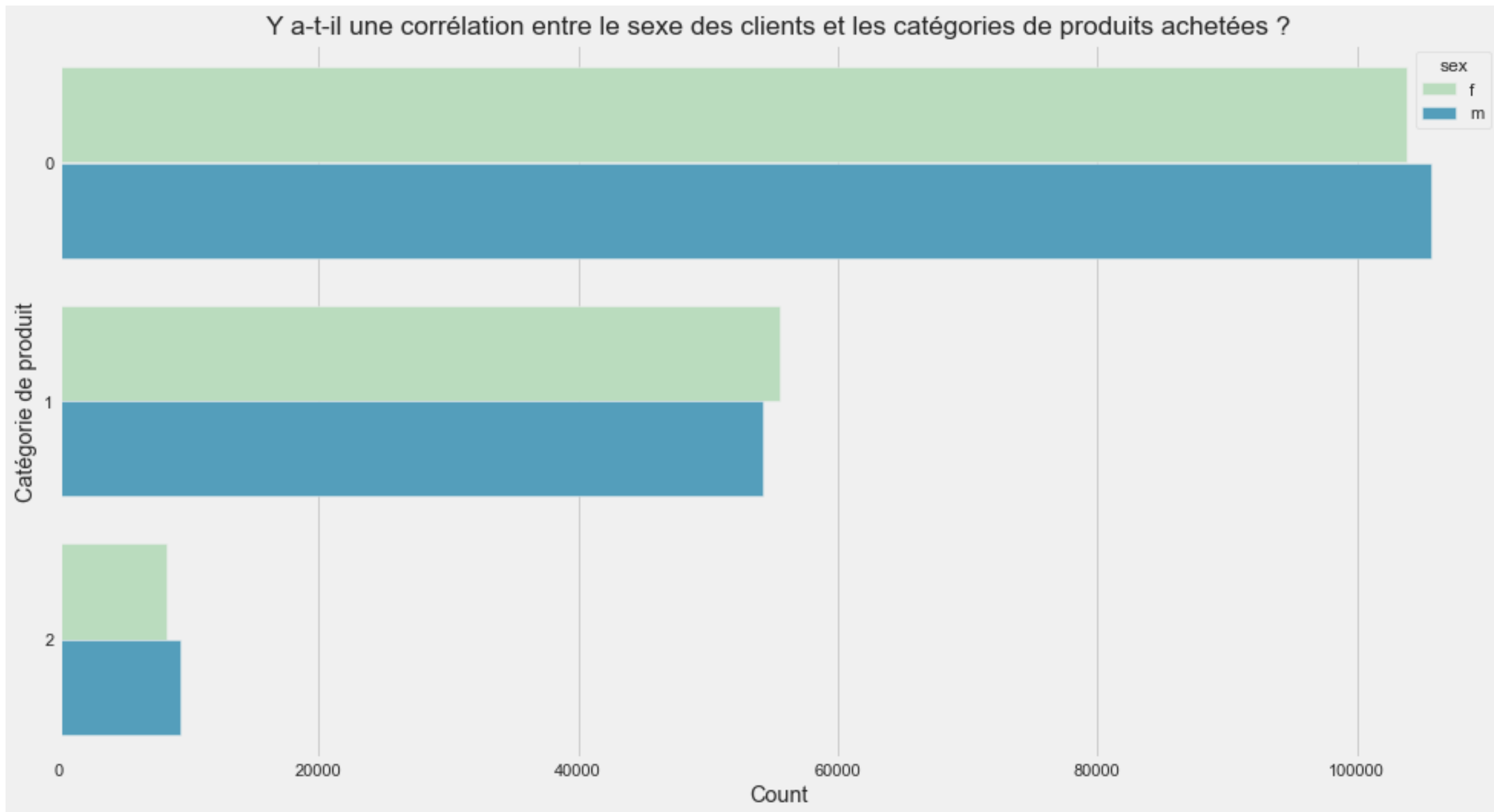
Le montant d'une transaction est-il différent d'une catégorie de produit à l'autre ?



On a obtenu $\eta^2_{Y/X} = 0,64$, cela signifie que les moyennes sont différentes, donc à priori il existe une corrélation entre la variable categ et la variable price.

? Question du manager

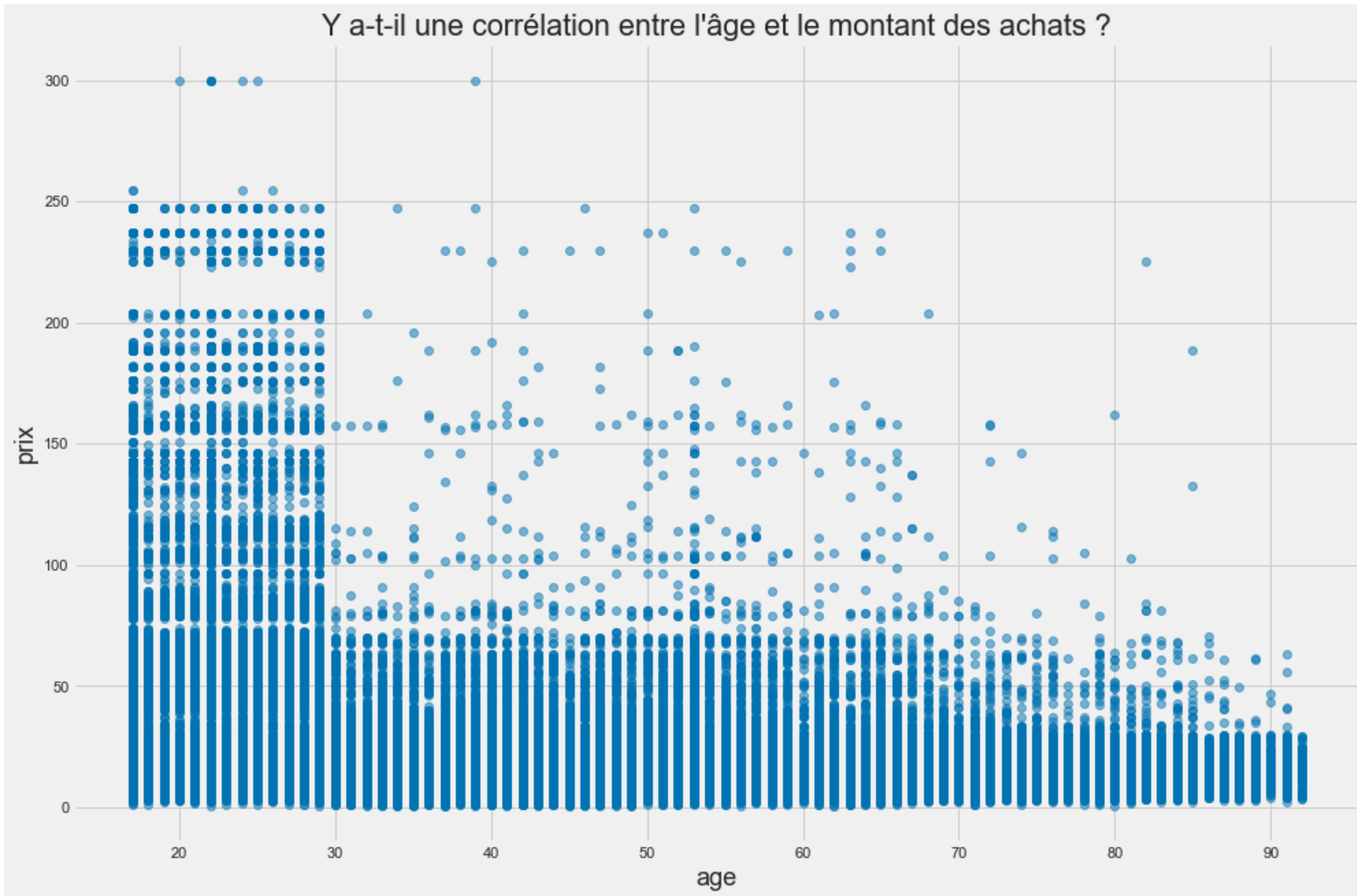
Y a-t-il une corrélation entre le sexe des clients et les catégories de produits achetés ?



Il apparait un très faible lien entre la variable sex et categ, il est hautement significatif du point de vu de la P valeur, aussi, on va rejeter l'hypothèse H0.

En somme, il y a une très faible corrélation entre le sexe des clients et les catégories de produits achetés.

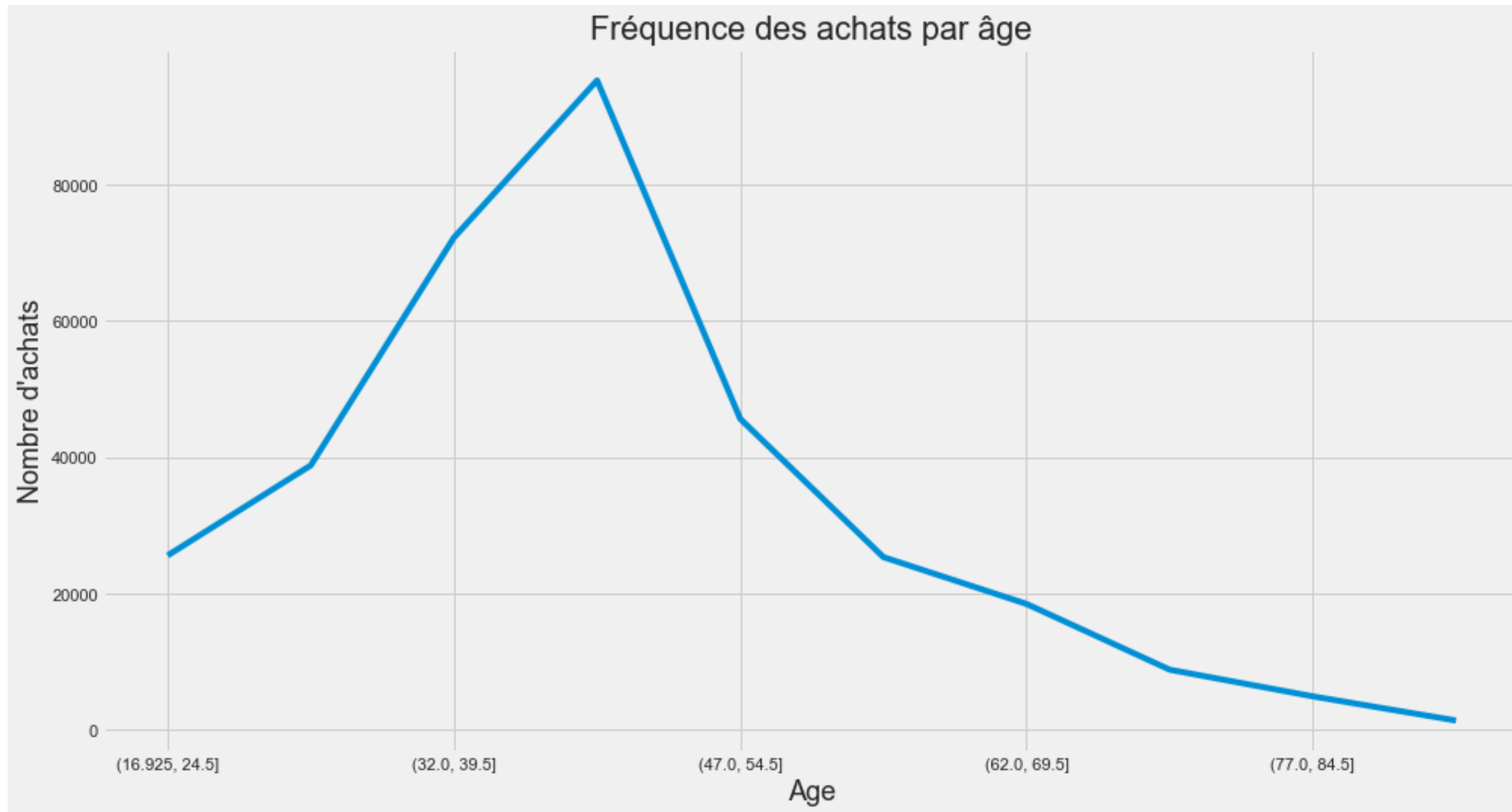
Y a-t-il une corrélation entre l'âge des clients et la montant total des achats ?



Le coefficient corrélation est faible $\rho = -0,211$ car sa forme n'est pas linéaire, cependant, on rejette H_0 .

En somme, il y a une corrélation entre l'âge des clients et le montant total des achats.

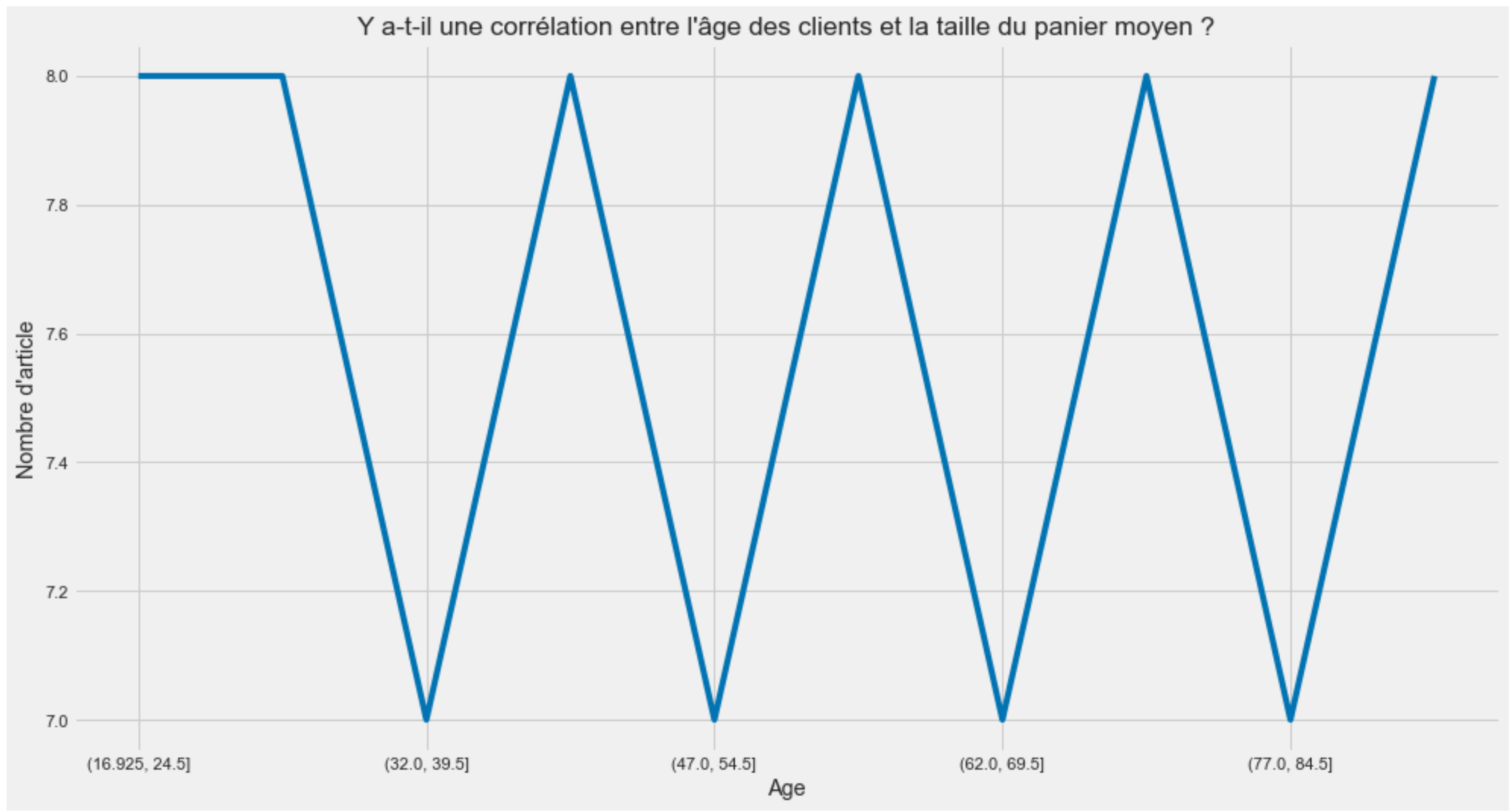
Y a-t-il une corrélation entre l'âge des clients et la fréquence des achats ?



Il apparait un très faible lien entre la variable âge et mois, il est hautement significatif du point de vu de la P valeur, aussi, on va rejeter l'hypothèse H0.

En somme, il y à une corrélation entre l'âge des clients et la fréquence des achats même si celles-ci est très faible.

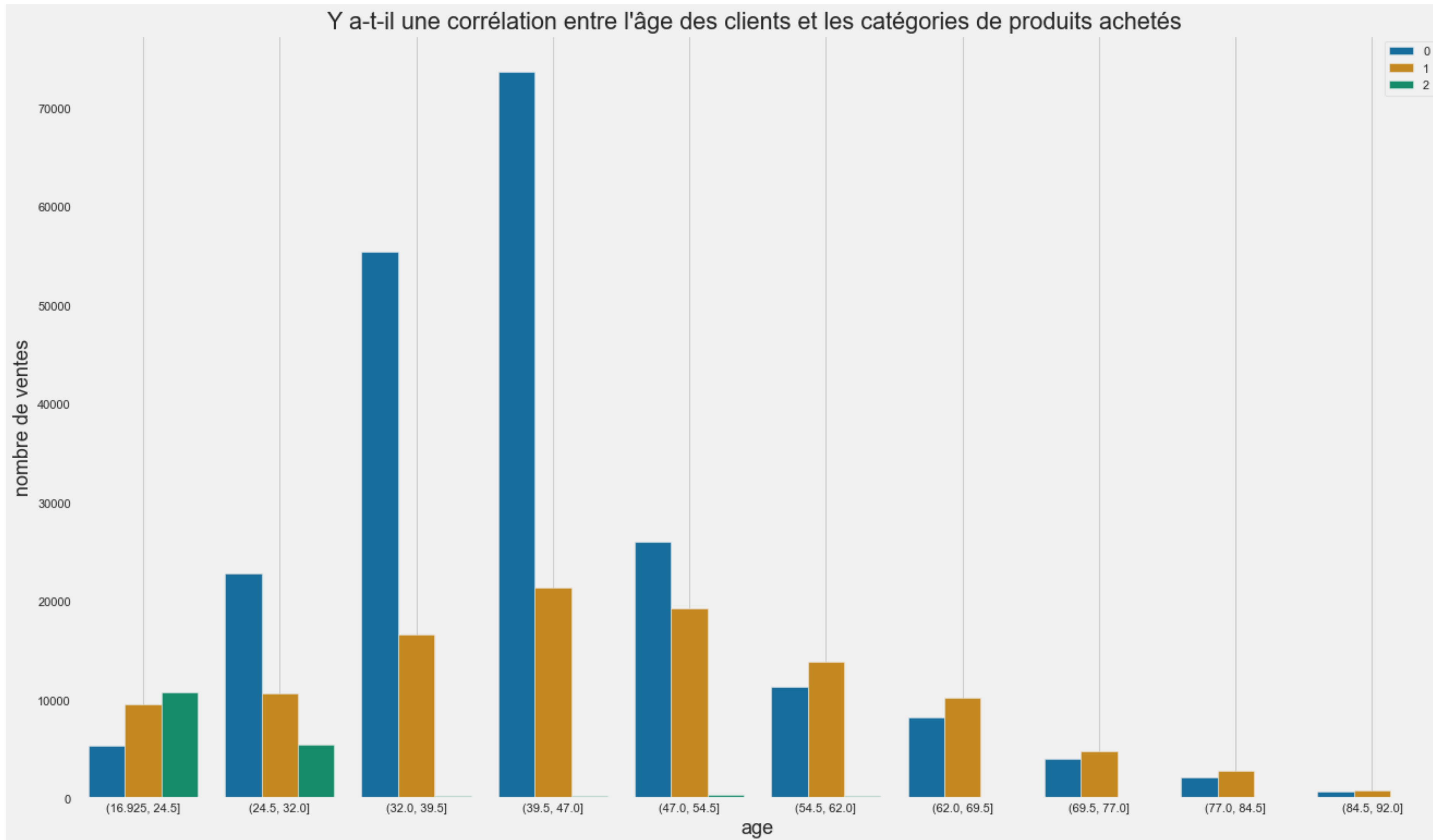
Y a-t-il une corrélation entre l'âge des clients et la taille du panier moyen (en nombre d'article) ?



Il apparait un très fort lien entre l'âge des clients et leurs paniers moyens en nombre d'articles et il est significatif du point de vu de la P valeur , aussi, on rejette l'hypothèse H0.

En somme, il y a une très forte corrélation entre l'âge des clients et la taille du panier moyen (en nombre d'article).

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



Il apparait un très fort lien entre l'âge des clients et les catégories de produits achetées. Il est hautement significatif du point de vu de la P valeur, aussi, on va rejeter l'hypothèse H_0 .

En somme, il y une corrélation entre l'âge des clients et les catégories de produits achetés.

4. Synthèse des résultats et recommandations

Synthèse :

Forces

- Clientèle à fort potentiel d'achat
- Une partie de la clientèle est fidèle
- Un beau volume de vente pour la categ 0
- Diversité des âges

Opportunités

- Se reapproprier la clientèle juniors et seniors
- Améliorer le résultat du CA
- Communiquer davantage sur l'étendue des catalogues de livres

Faiblesses

- Des segments de la clientèle peu exploités
- Un faible volume de vente pour la categ 1 et 2 par rapport à categ 0
- Rupture approvisionnement durant la période critique des ventes (rentrée)

Menaces

- Baisse résultat du CA
- E-réputation du site concernant la qualité du service (disponibilité produits...) et l'image de l'entreprise qui peut-être cantonnée à un type de clientèle
- Rupture Supply-chain

Recommandations :

- ❖ Opérer une campagne de recrutement sur les segments de clientèle sous exploités à savoir les juniors et les seniors(ex: pour seniors réfléchir à un design de site web favorisant un accès facilité, pour juniors ex: proposer cadeau si parrainage nouveau client...).
- ❖ Opérer des offres promotionnelles (ventes flash...) sur les catalogues de produits de la categ 1 et 2 qui enregistrent des volumes de ventes faibles par rapport à la categ 0.
- ❖ Conforter la visibilité du catalogue de produits de la categ 0 avec une veille sur les performances car les ventes sont en dessous de la moyenne.
- ❖ Communiquer davantage sur l'exhaustivité des catalogues de l'entreprise du junior au senior avec un slogan (ex:« tous livre »).