Librairie et vente en ligne

Rapport sur les ventes en ligne

08/11/2023

Processing

Fichier clients
- Contient l'Id_client, le sexe et l'année de naissance du client
- L'année de naissance nous permet de déterminer l'âge du client

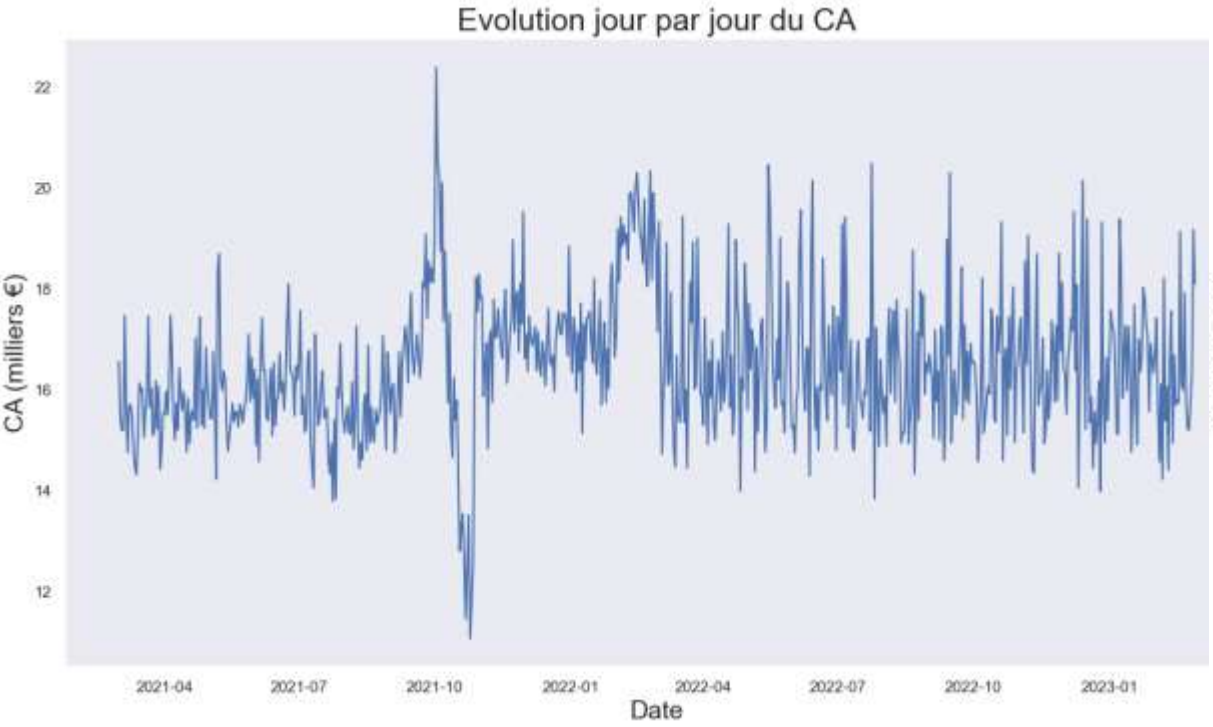
Fichier Products
- Contient l'Id_produit, le prix et la catégorie.
- 29 produits sont à moins d'1 euro

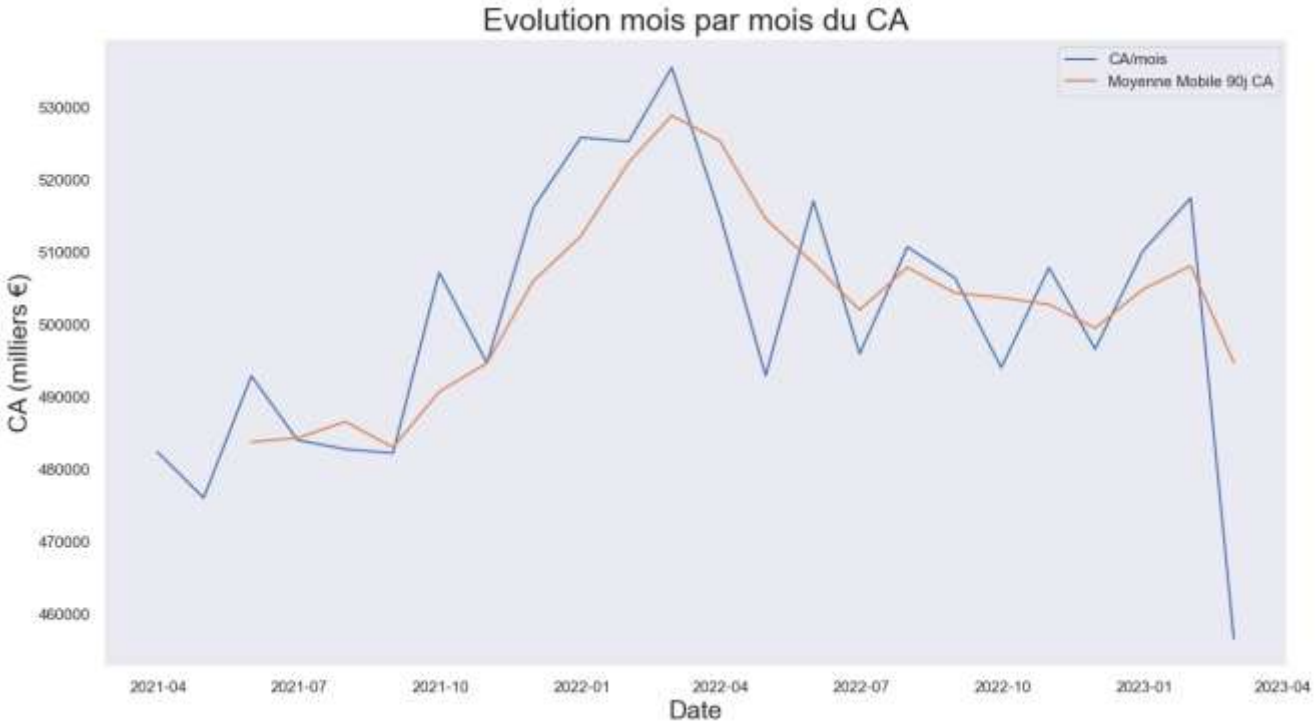
Fichier Transactions
- Contient l'Id_produit, l'id_client, l'id_session et la date.
- 236 lignes ont une date mal implémentée (24 au lieu de 00). Nous résolvons le problème en mettant le bon format

Jointure des fichiers
- Jointure sur le fichier transactions avec Id_prod et Id_client
- 21 produits n'ont pas de correspondance dans les transactions. On peut penser qu'ils n'ont jamais été vendus. Besoin de remonter l'information auprès du marketing pour comprendre pourquoi ils ne sont pas vendus (problèmes référencement ? Prix ?..)
- 21 clients ne sont pas présents dans la table Transactions. Un client n'est dans la base que s'il a effectué une transaction. Besoin de creuser ce problème avec l'équipe IT sur ces clients pour voir de qui il s'agit.

Fichier final pour analyse
- Nous choisissons d'écarter ces 42 lignes pour continuer notre analyse.
- 690 000 transactions, 8600 clients et 3265 produits à analyser

Evolution du CA



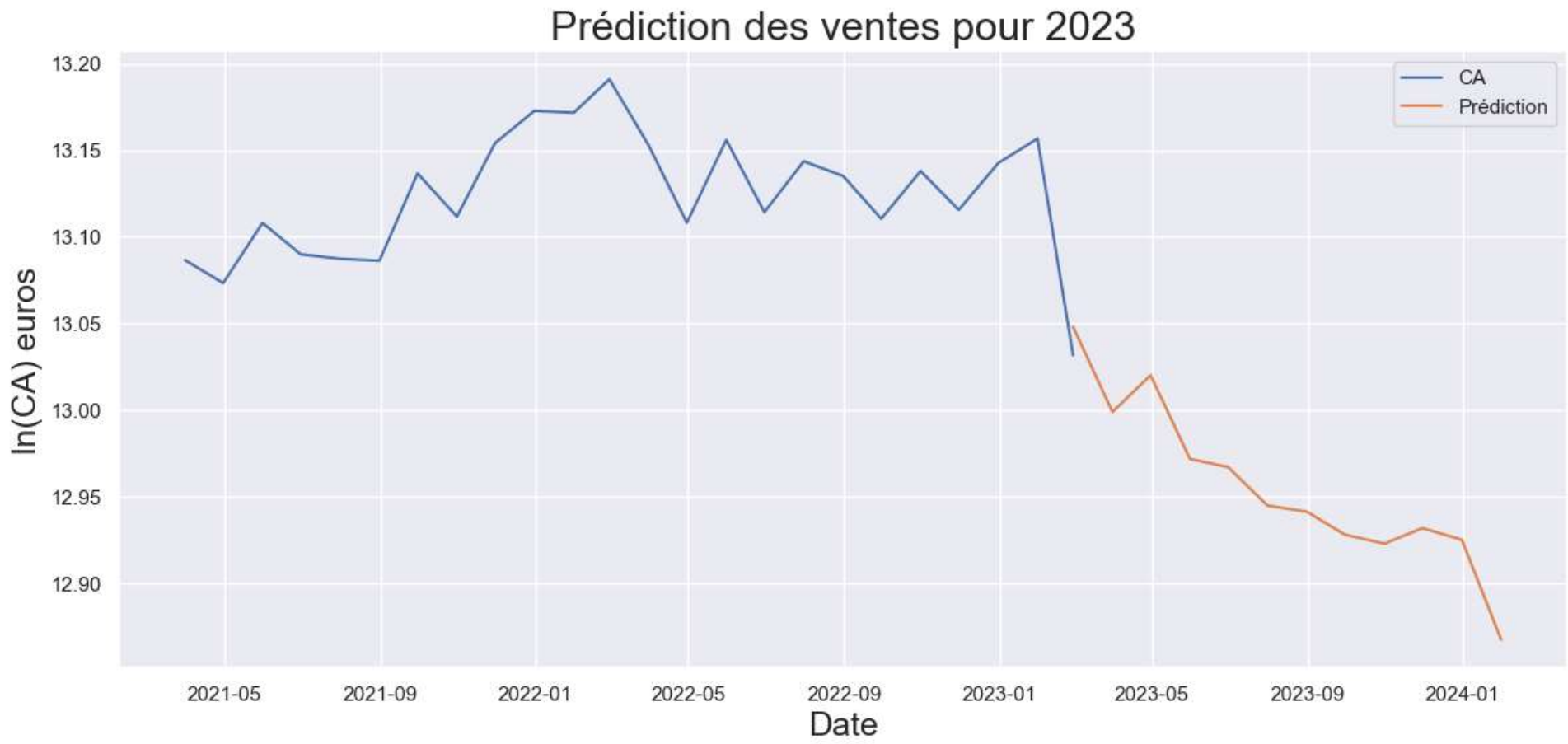


CA en ligne de 12,03 millions d'euros

- Hausse et chute brutale au mois d'octobre 2021
- Données à analyser pour comprendre les pics
- N'influence pas notre jeu de données car moyenne mois octobre équivalente aux autres mois

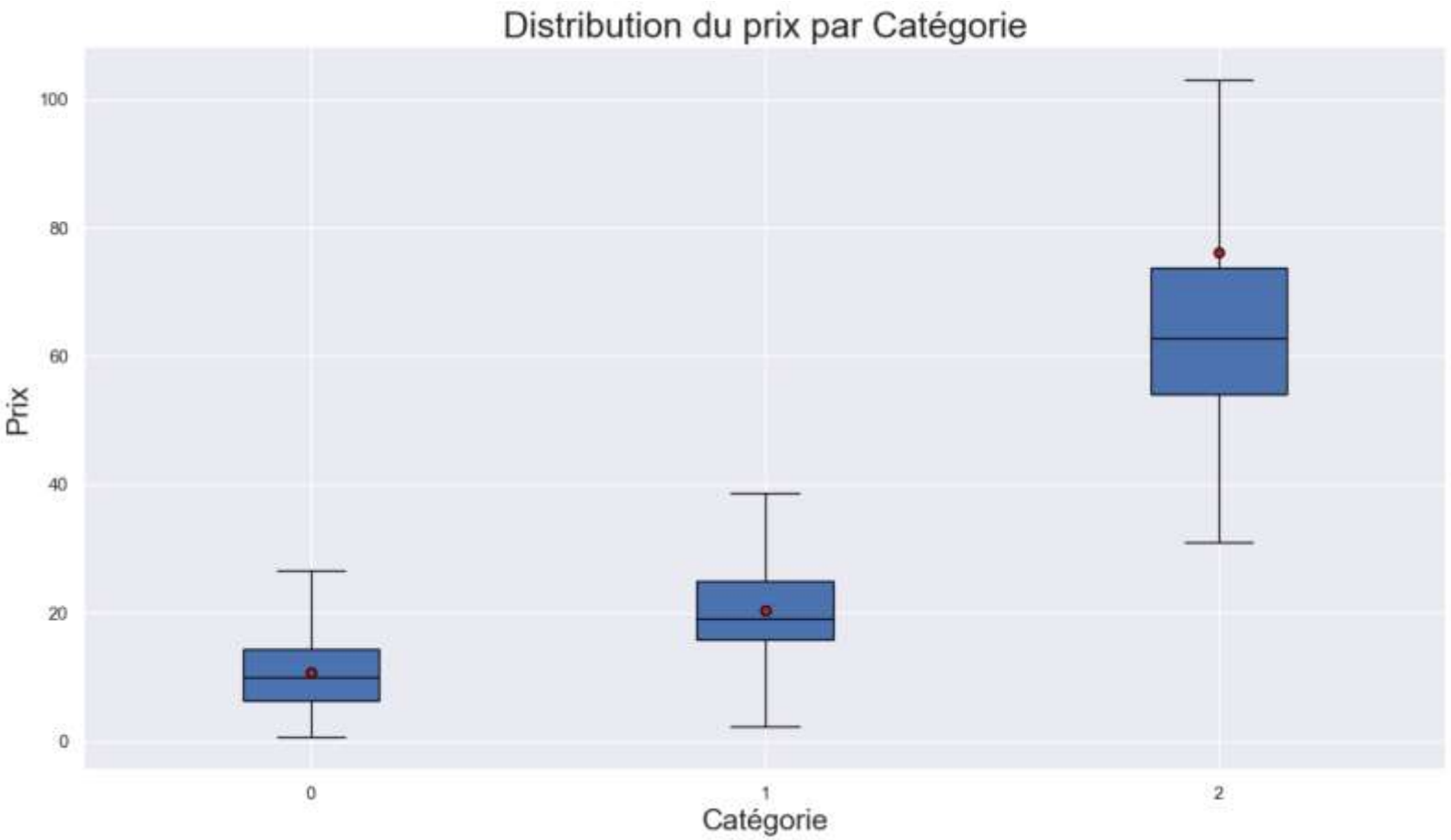
- La moyenne mobile permet de désaisonnaliser notre CA

Projection du CA sur 2023



- Utilisation du modèle de projection Holt-Winters
- Pour des projections plus fiables, possibilité d'utiliser des modèles plus performants

Distribution du prix par catégorie

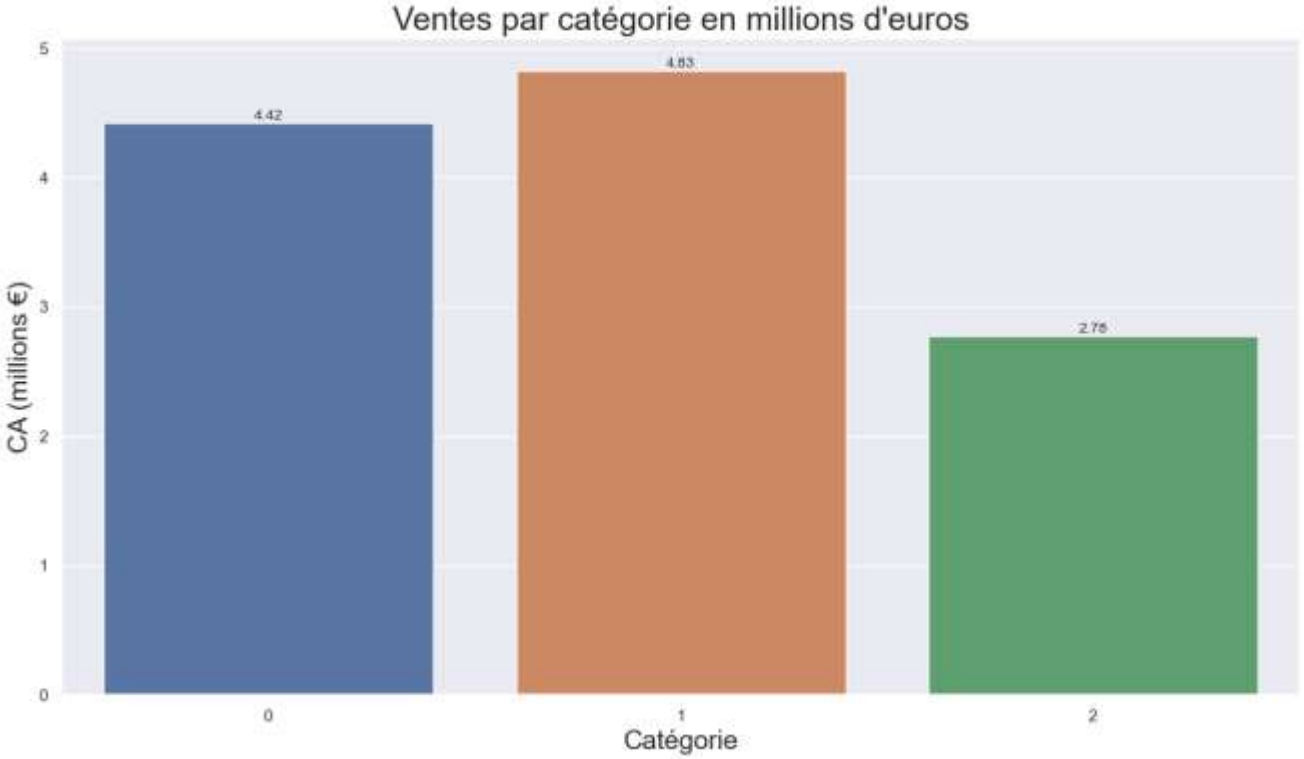


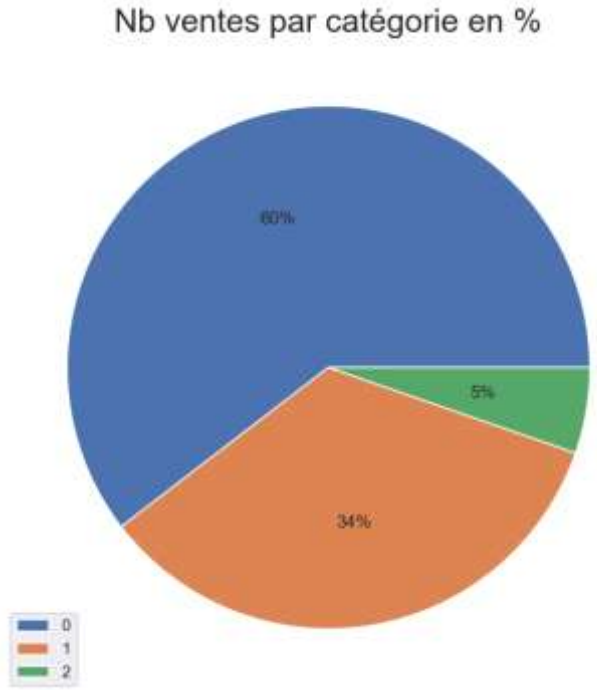
Catégorie 0 : low price

Catégorie 1 : middle price

Catégorie 2 : high price

Analyse des catégories



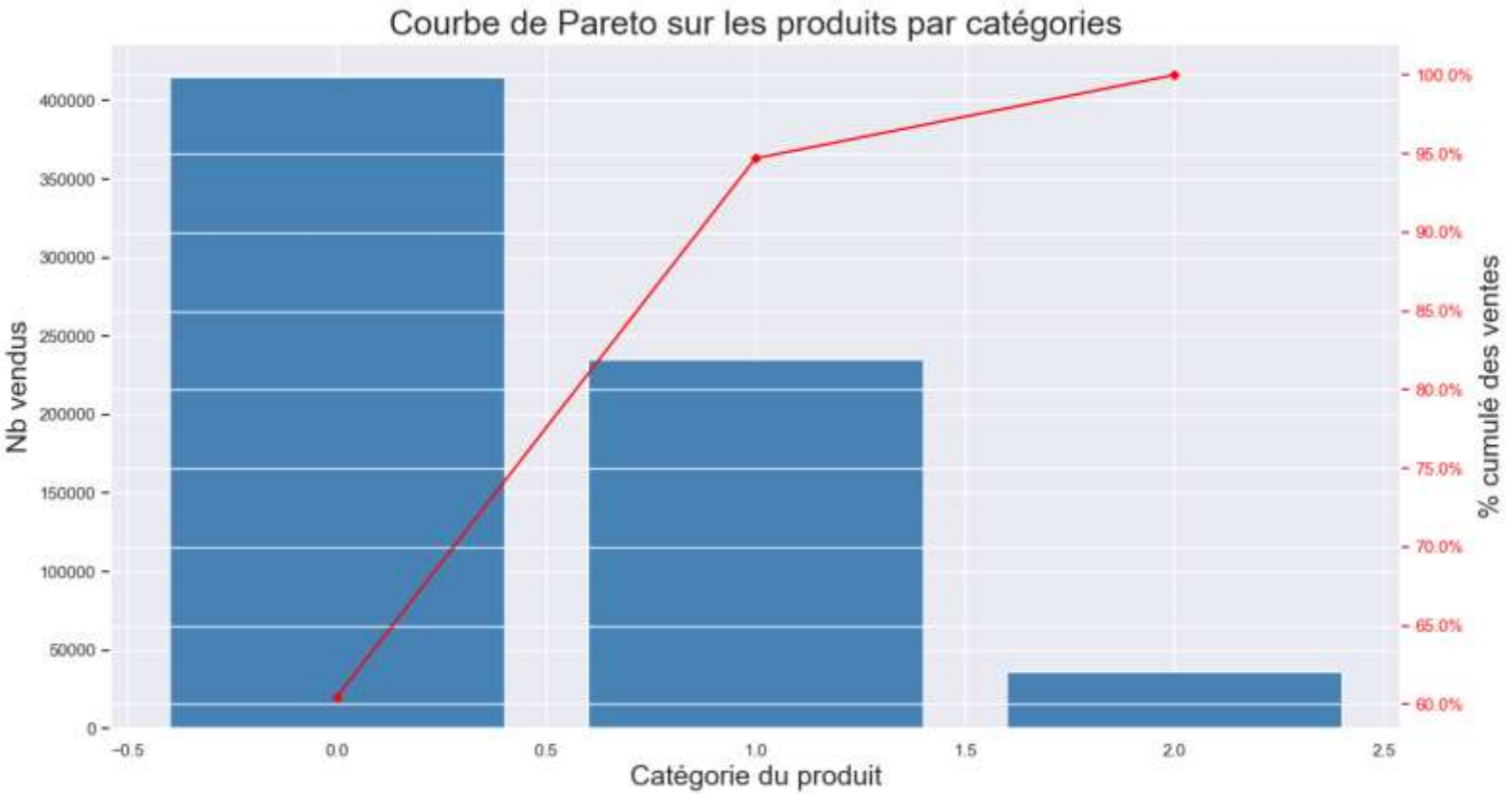


La catégorie 0 représente 37% du CA mais 60% de nos articles vendus

La catégorie 1 représente 40% du CA mais 34% de nos articles vendus

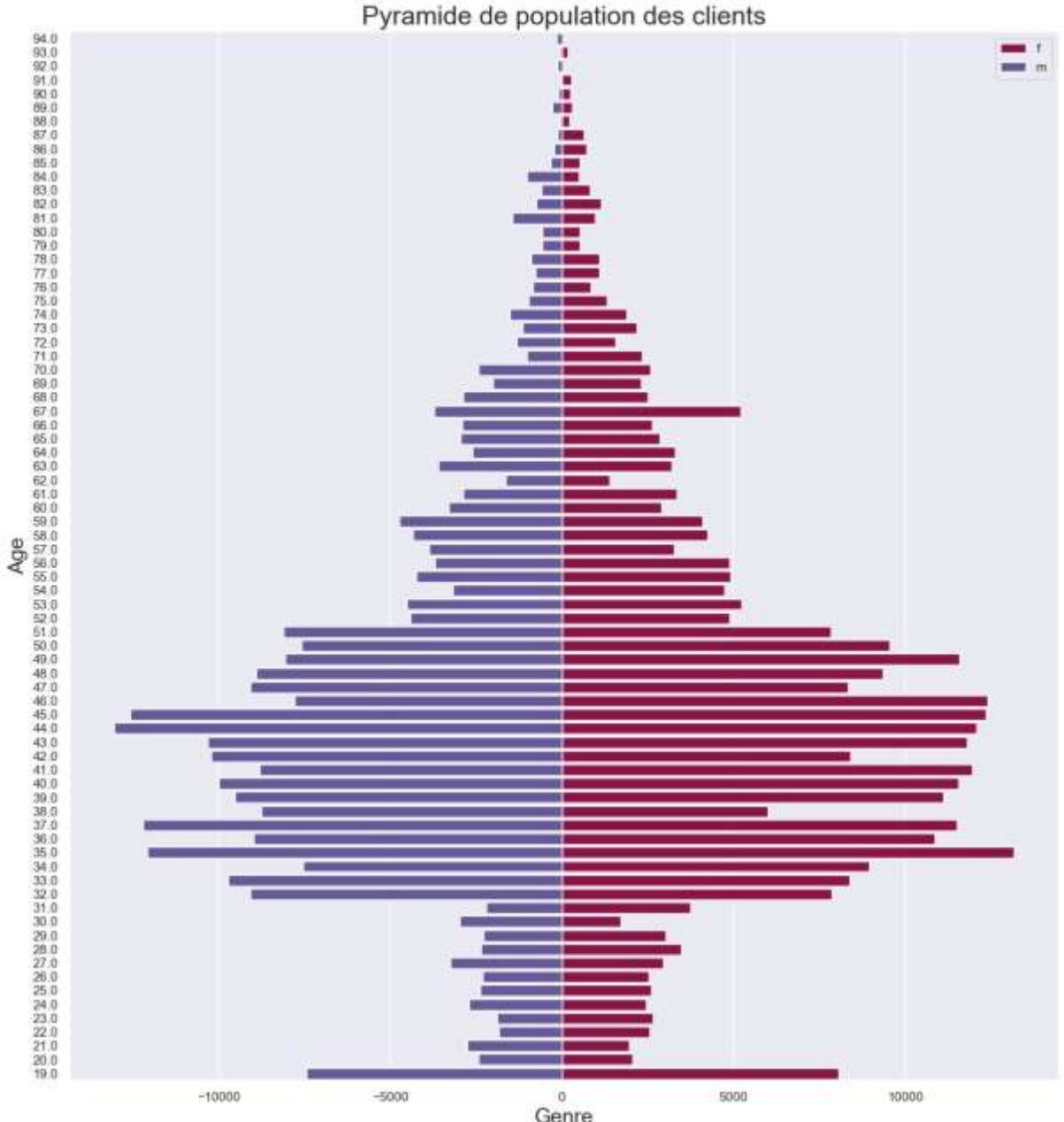
La catégorie 2 représente 23% du CA mais 5% de nos articles vendus

Courbe Pareto des produits



95 % des ventes en nombre se font avec la catégorie 1 et 2

Pyramide des âges

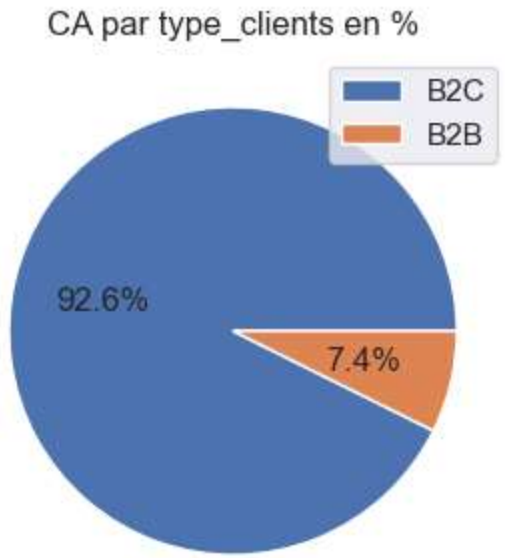


- Population 18 ans surreprésentée
- Hypothèse : une date de naissance minimale par défaut en 2003

Ventes et clients

Top 10 clients en CA

client_id	price
c_1609	326039.89
c_4958	290227.03
c_6714	153918.60
c_3454	114110.57
c_1570	5285.82
c_3263	5276.87
c_2140	5260.18
c_2899	5214.05
c_7319	5155.77
c_7959	5135.75

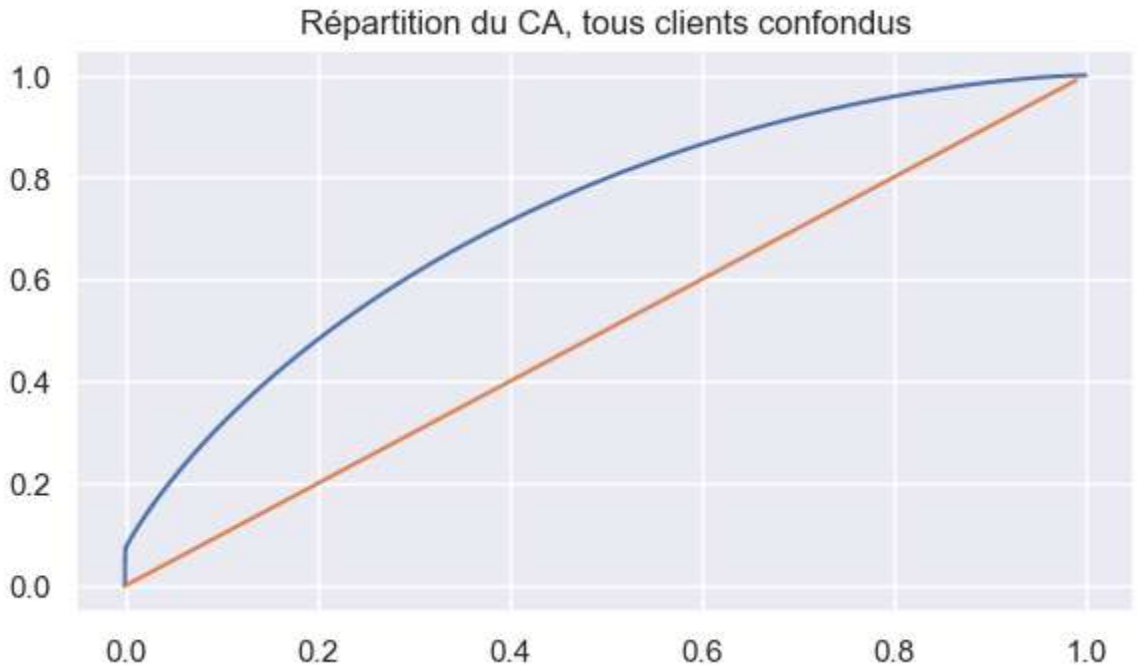


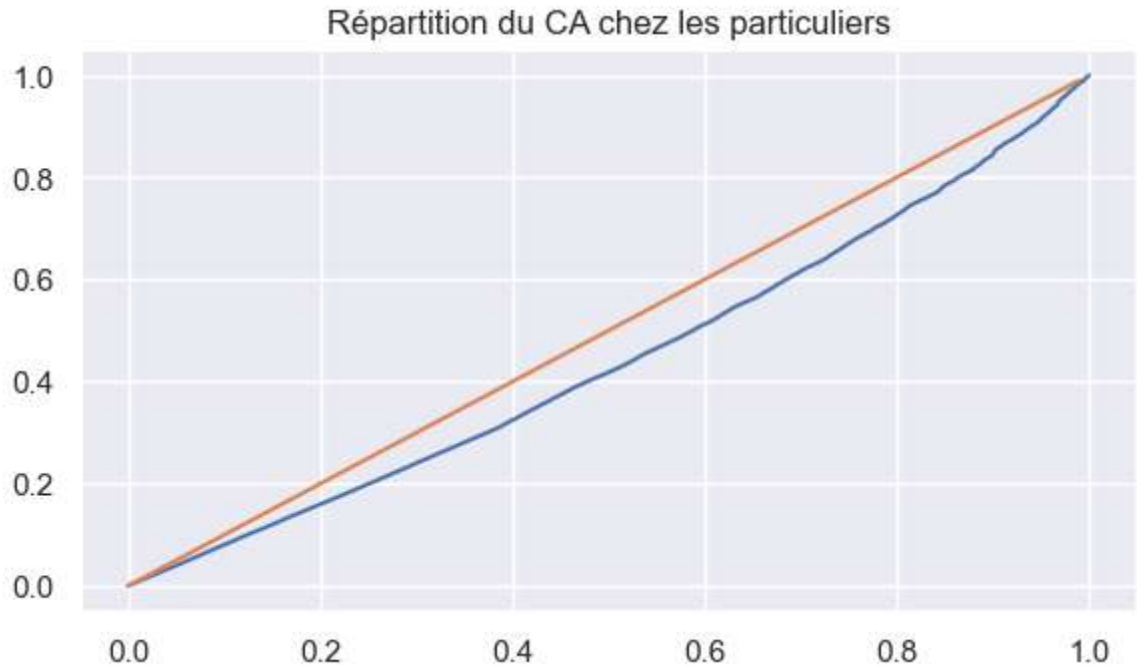
1400 euros par clients en moyenne

Nos identifions nos 4 premiers clients comme des clients professionnels

Nos clients particuliers réalisent 93% de notre CA.

Comportement clients





Tous nos clients :

- la distribution du CA par client est inégalitaire
- 50% de nos clients cumulent 80% du CA

Clients particuliers uniquement:
La répartition des achats est plutôt égalitaire

Pour la suite des analyses nous nous concentrons sur l'analyse des clients particuliers uniquement

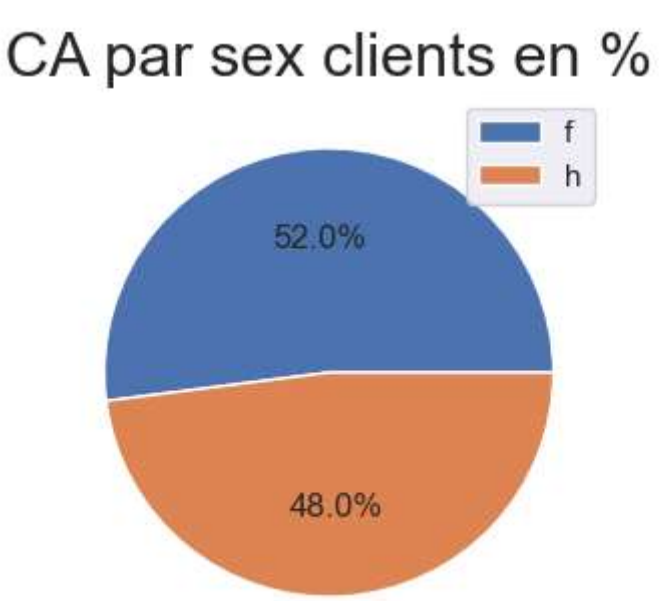
Méthodologie

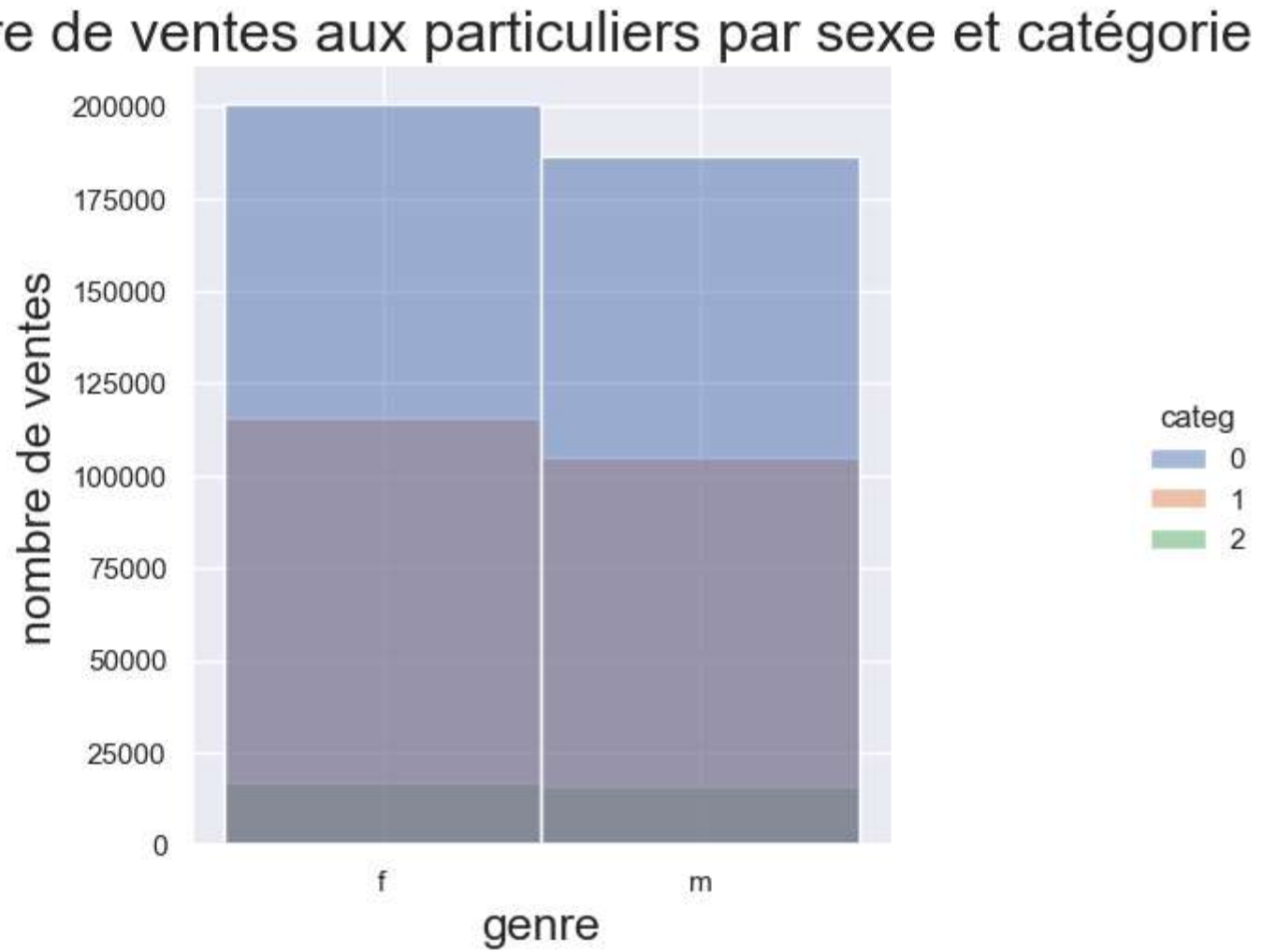
- Etape 1 : Poser la question : Existe-t-il une corrélation entre ces 2 variables ?
- Etape 2 : Définir si les variables étudiées sont de type qualitatives ou quantitatives
- Etape 3 : Choix du test paramétrique à utiliser (comparaison / association)
- Etape 4 : Effectuer un test paramétrique ou non paramétrique

Si distribution des données suit loi normale = Test paramétrique

- Etape 5 : Analyse du résultat donné par le test choisi et de la p value de celui-ci

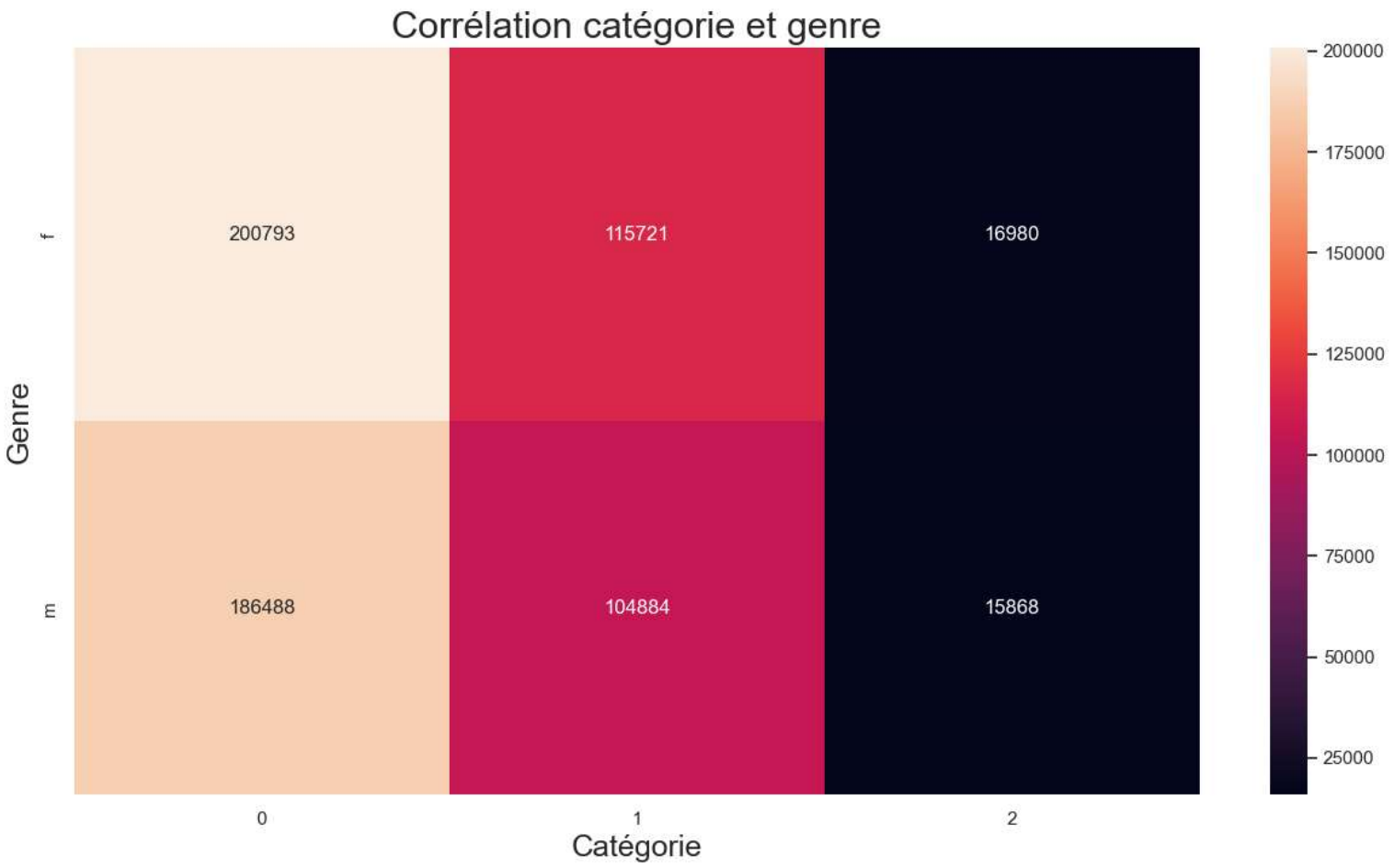
Corrélation genres et catégories





Aucune particularité sur ces premières réflexions

Corrélation genres et catégories



Etude du lien entre 2 valeurs qualitative (genre et catégories)

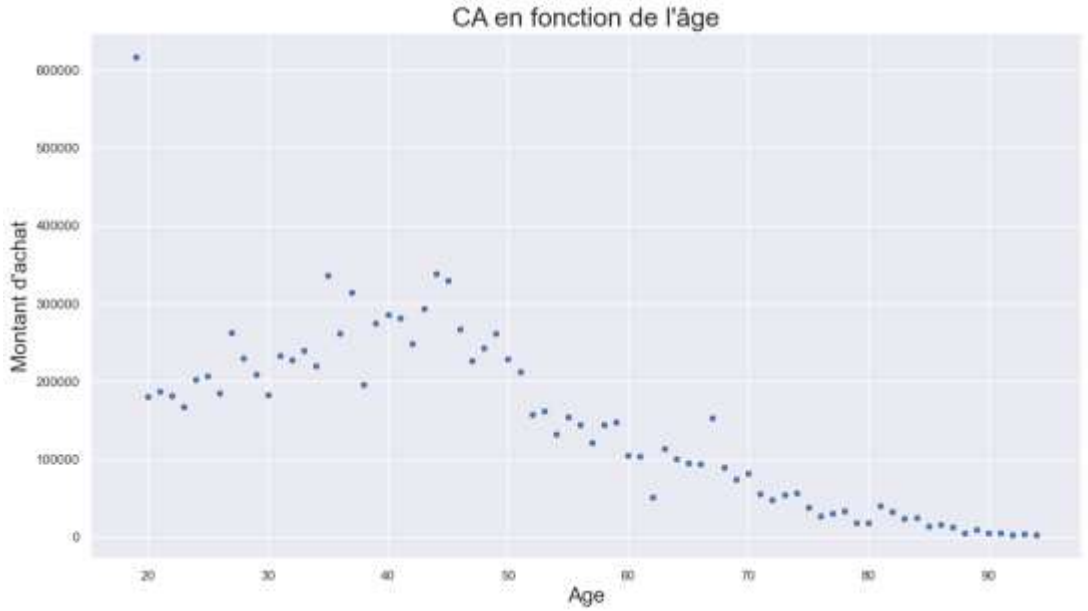
Réalisation d'un tableau de contingence puis d'un test de Chi2

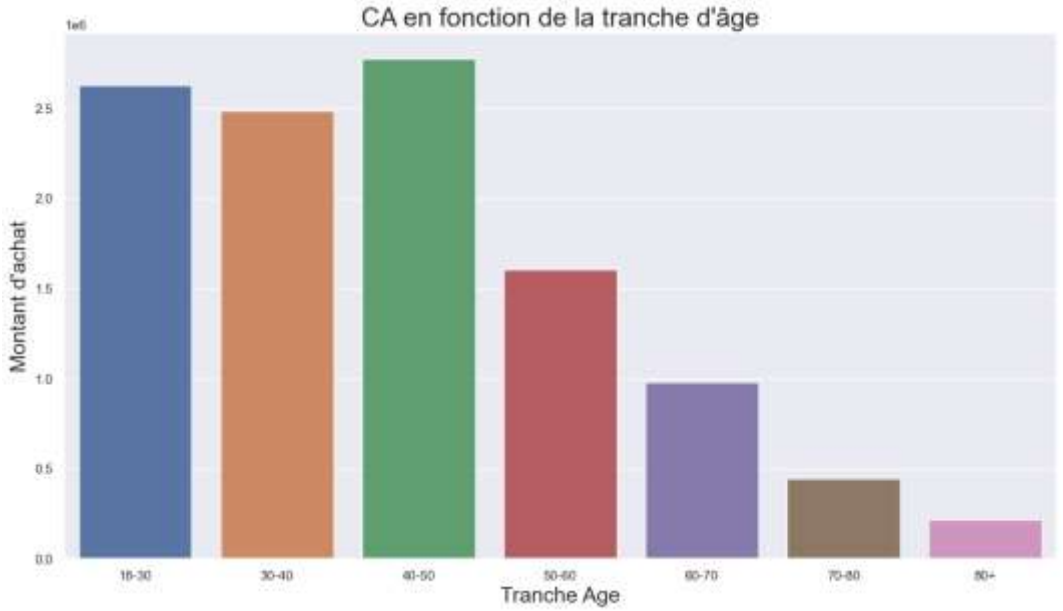
La Pvalue < à 5% : association statistiquement significative entre les variables catégorie et genre.

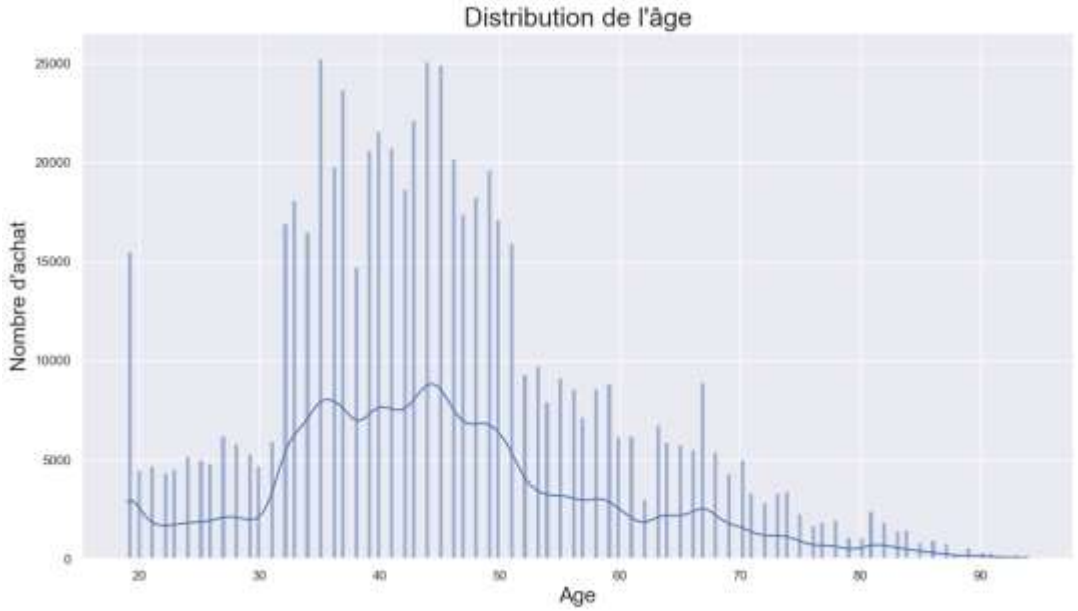
Conditions sur l'indépendance et sur la présence de +5 valeurs sont acquises

Représentation sous forme de heatmap de cette corrélation

Corrélation âge et montant d'achat

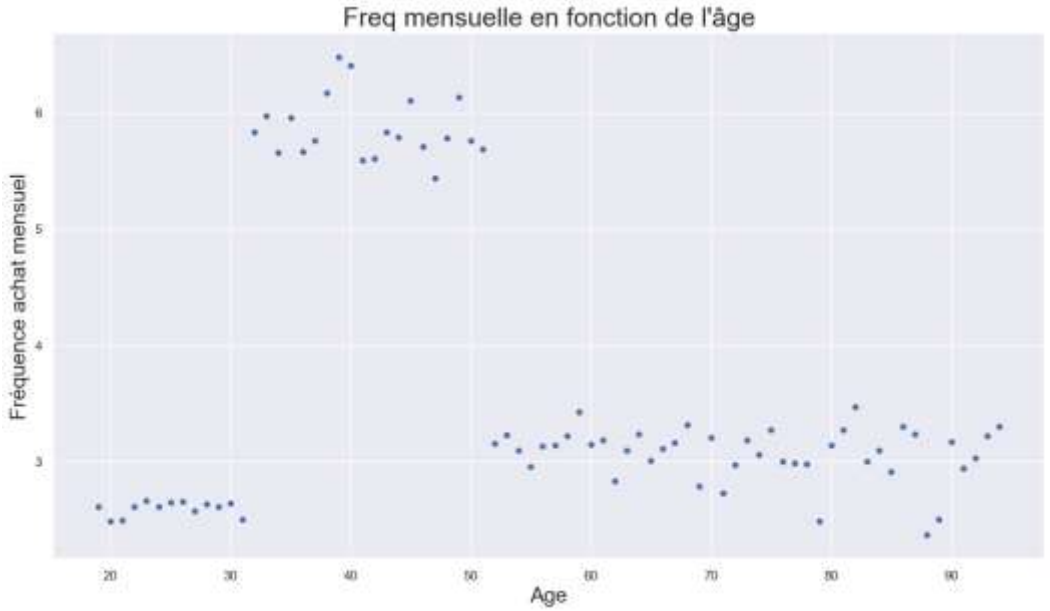


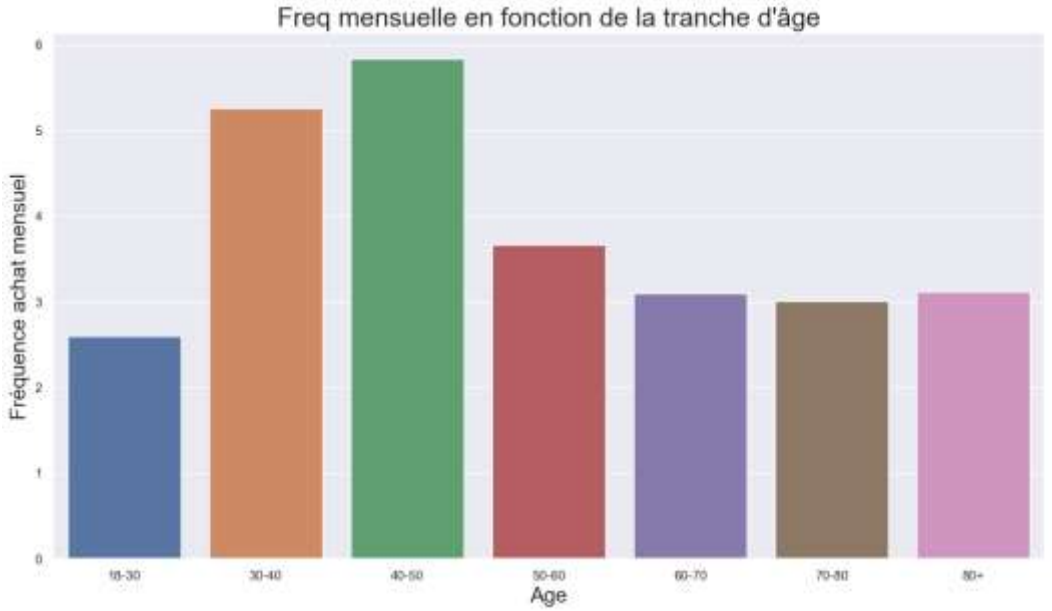


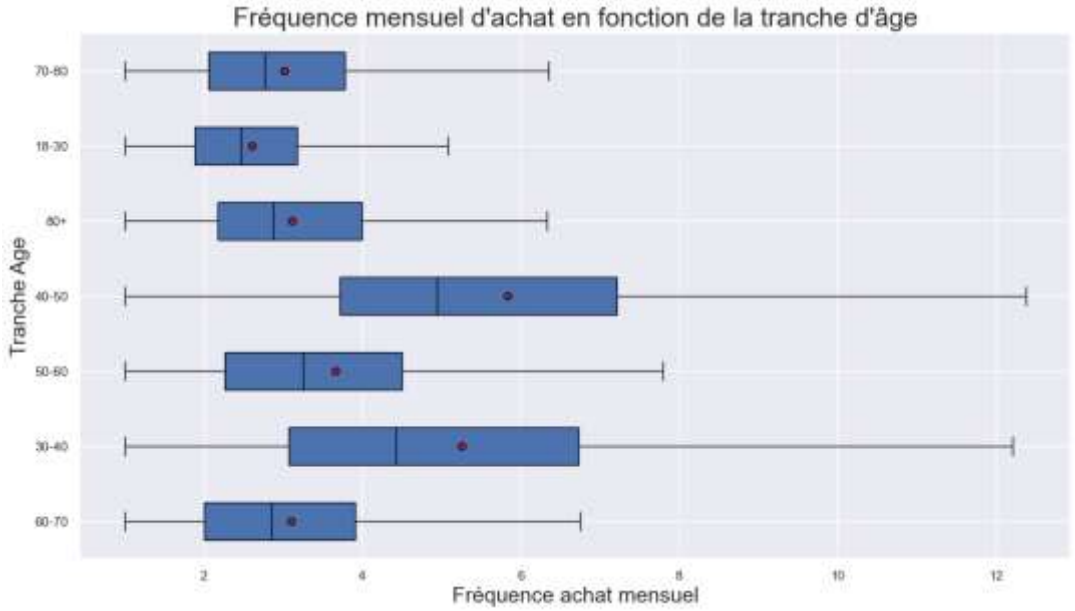


- Nos données âge et montant ne semble pas suivre une loi normale
- Le test de Kolmogorov-Smirnov (KS) nous montre une P value < 5% ce qui confirme notre impression
- Le test de Spearman nous donne une corrélation négative entre âge et montant total. La pvalue est inférieur à 5% et montre que cette corrélation est statistiquement significative

Corrélation âge et fréquence d'achat

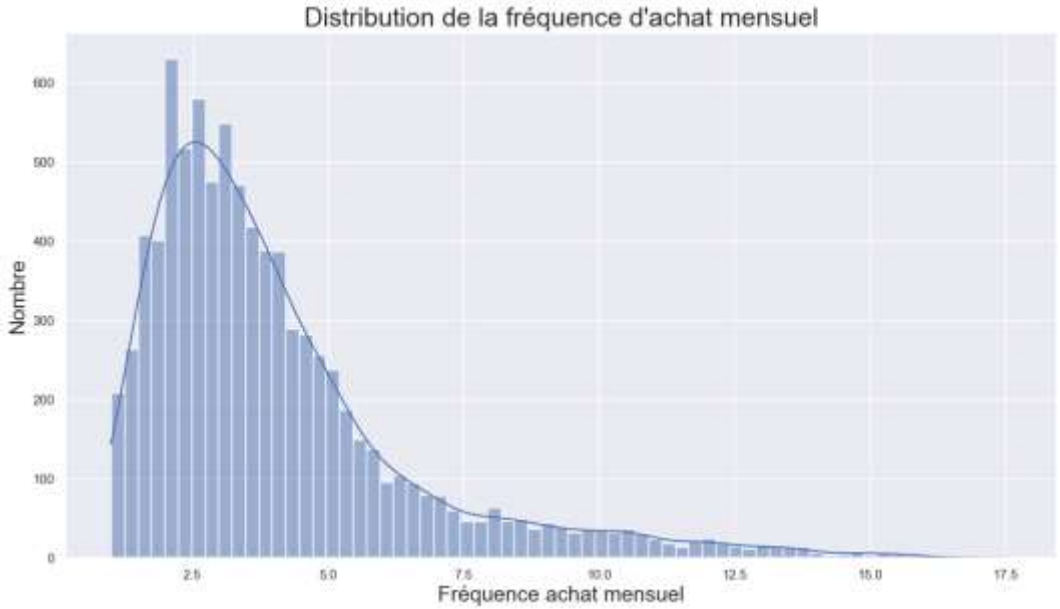


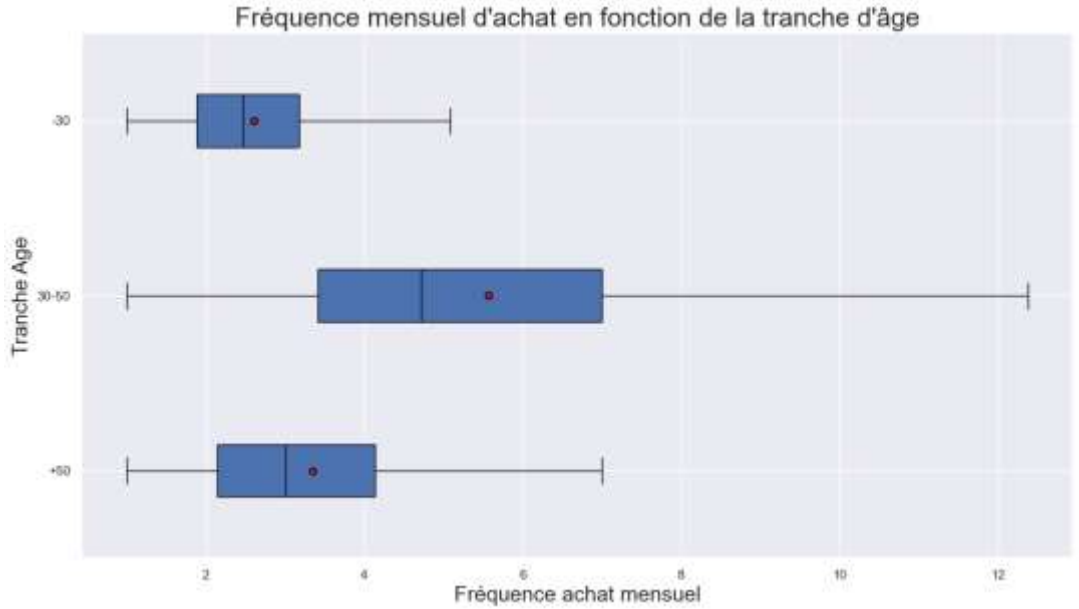


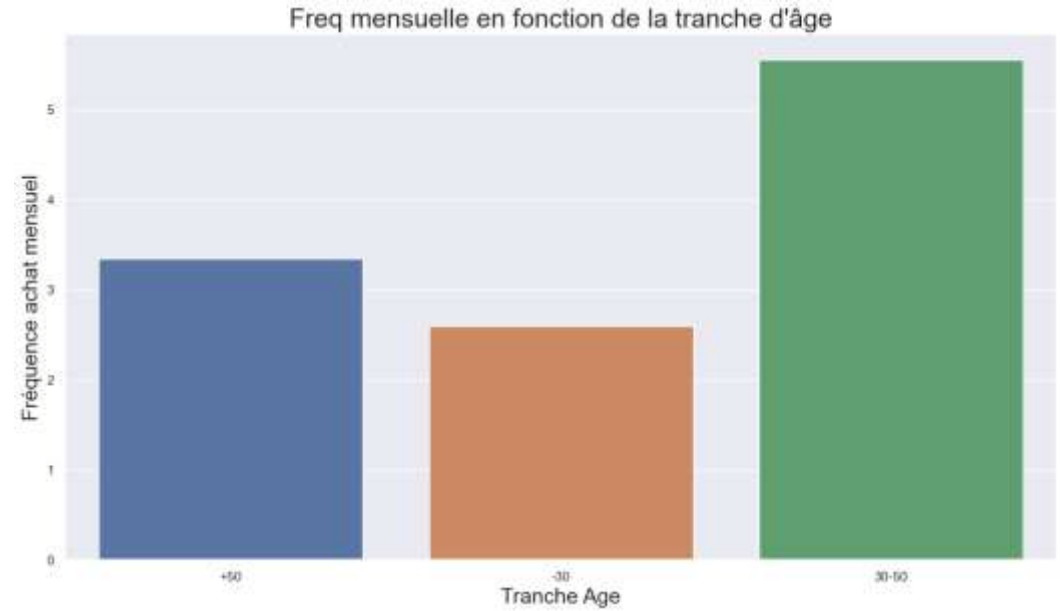


- Nos données âge et fréquence d'achat ne semble pas suivre une loi normale
- Le test de Kolmogorov-Smirnov (KS) nous montre une P value < 5% ce qui confirme notre impression
- Le test de Spearman nous indique qu'il n'y a pas de corrélation entre âge et freq d'achat.

Corrélation âge et fréquence d'achat



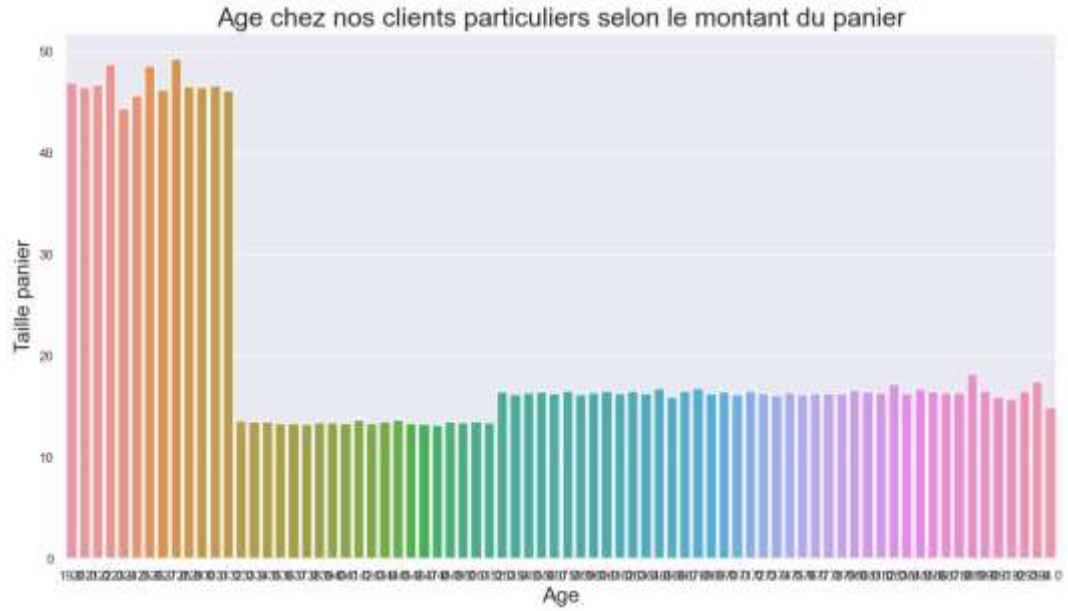




- Nos données fréquence d'achat ne semble pas suivre une loi normale ce que confirme le test de Kolmogorov-Smirnov (KS)

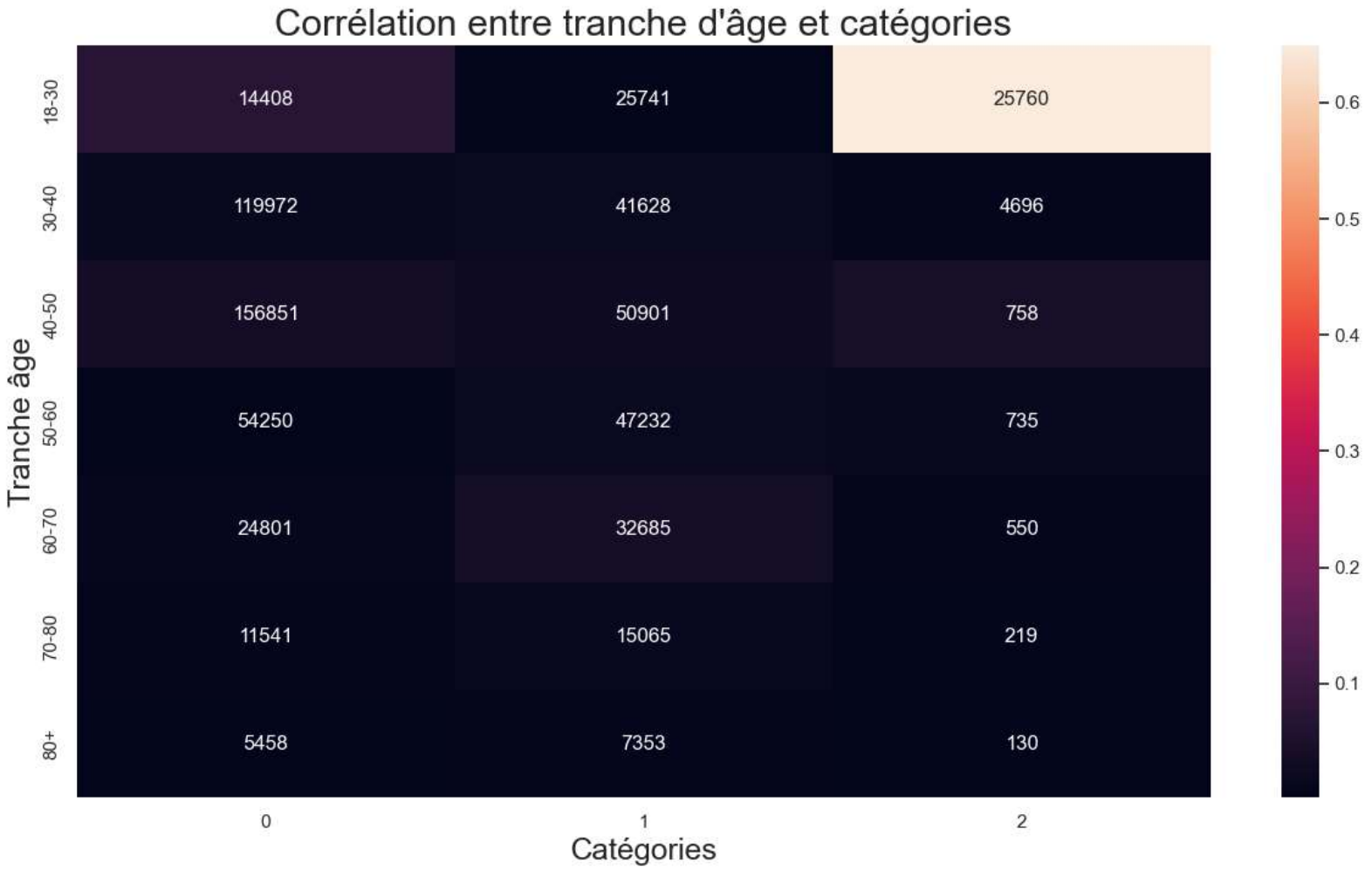
- Le test de Kruskal-Wallis nous indique qu'il n'y a pas de corrélation entre âge et freq d'achat.
- Revoir l'hypothèse et la p value

Corrélation âge et montant du panier



- Nos données âge et montant du panier ne semble pas suivre une loi normale
- Le test de Kolmogrov-Smirnov (KS) nous montre une P value < 5% ce qui confirme notre impression
- Le test de Spearman nous indique qu'il n'y a pas de corrélation entre âge et montant du panier.

Corrélation âge et catégorie



Etude du lien entre 1 valeur qualitative (catégorie) et 1 quantitative (âge)

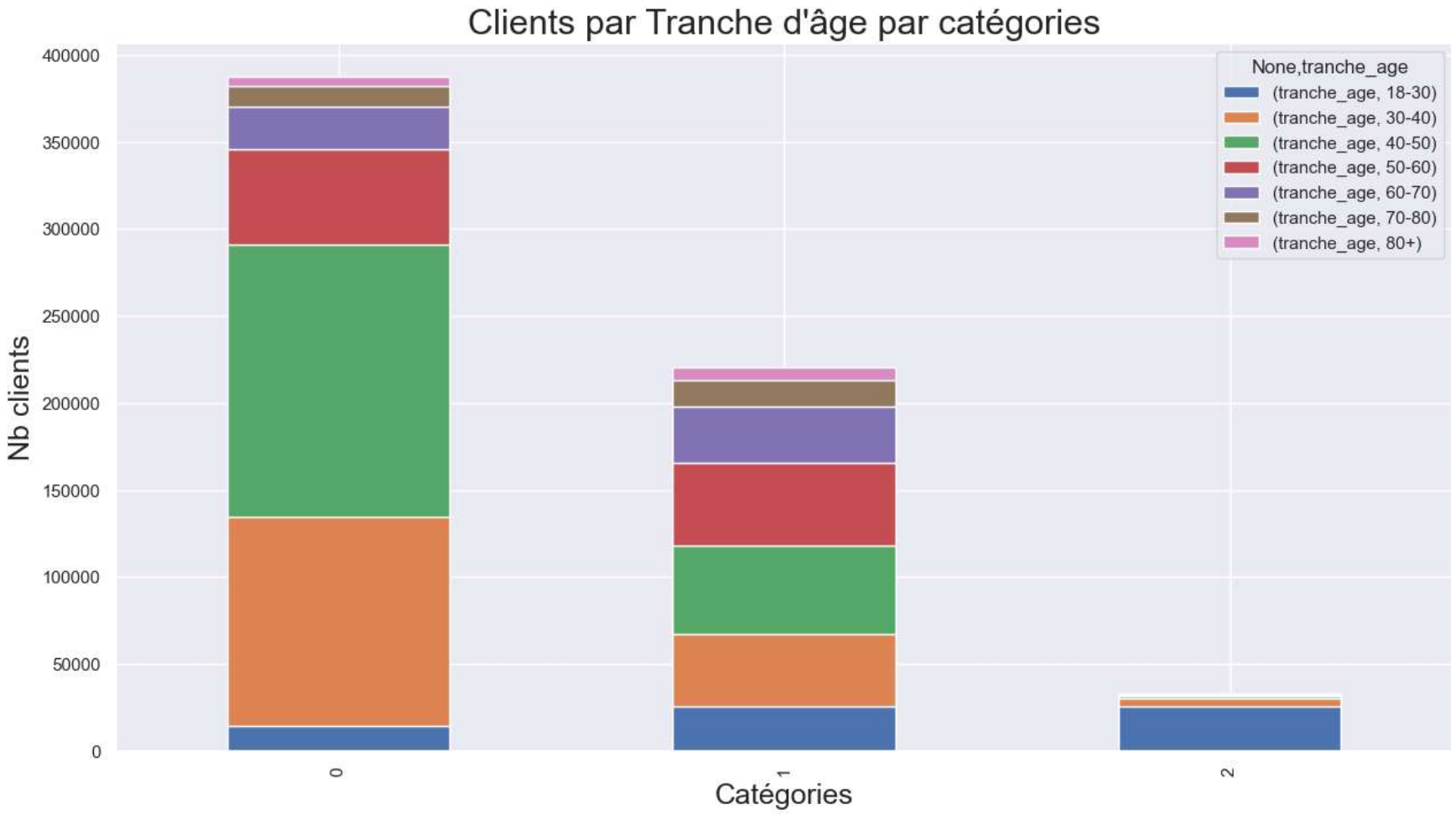
Pour étudier la corrélation nous allons étudier la variance.
Les catégories 0 et 1 ont des moyennes proches

Réalisation d'un tableau de contingence puis d'un test ANOVA

La Pvalue < à 5% : pas de corrélation

En considérant nos données comme non paramétric : Test de Kruskal-Wallis
Ce test confirme l'absence de corrélation entre l'âge et la catégorie.

Cette absence peut être visualiser via une heatmap


Clients par Tranche d'âge par catégories
Nb clients
400000
350000
300000
250000
200000
150000
100000
50000
0
0
1
2
Catégories
None,tranche_age
(tranche_age, 18-30)
(tranche_age, 30-40)
(tranche_age, 40-50)
(tranche_age, 50-60)
(tranche_age, 60-70)
(tranche_age, 70-80)
(tranche_age, 80+)

Conclusion

- Une moyenne du CA par mois assez stable comprise entre 490 000 et 520 000 euros
- Cat 0 : 60% des ventes, 37% du CA | Cat 1 : **34** % des ventes, **40** % du CA | Cat 2 : 5 % des ventes, 23 % du CA
- Corrélation entre Sex client et catégorie achetée
- Corrélation entre âge et montant d'achat
- Pas de corrélation entre l'âge et la fréquence d'achat
- Pas de corrélation entre l'âge et le montant du panier
- Absence de lien entre l'âge et la catégorie