OPENCLASSROOMS

Analysez les ventes de votre entreprise

Mentor : Claire Della Nova

Agenda

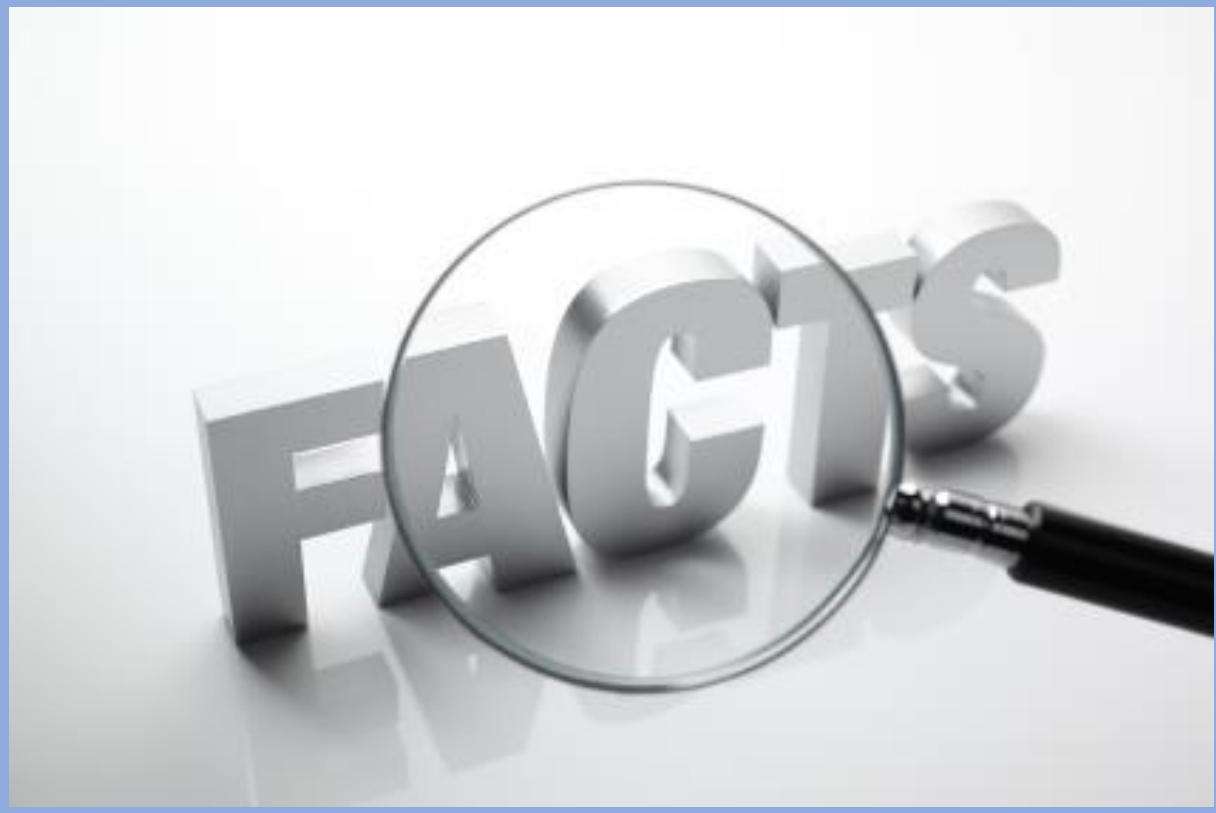
Question 1: Nettoyage, préparation, investigation préléminaire
- Identification des valeurs étrangères
- Identfication des clés
- Nettoyage des 3 fichiers
- Merge des tables et traitement des NaN
- Contrôle de la cohérence des variables
- Recherche temporelle

Question 2 : Analyse des données
- Les types de variables
- Âge et price
- Courbe de Lorenz et indice Gini sur les ventes
- Catégories par âge, prix et sexe
- Analyse globale
- Analyse par catégories
- Boxplot produits achetés par catégorie

Question 3 : Tests statistiques
- Corrélation entre genre et catégories de produits achetés
- Corrélation entre l'âge des clients et le montant des achats
- Corrélation entre l'âge des clients et le nombre d'achats
- Corrélation entre l'âge des clients et la taille du panier moyen
- Corrélation entre l'âge des clients et les catégories de produits achetés

Question 1

Nettoyage, préparation, investigation préléminaire



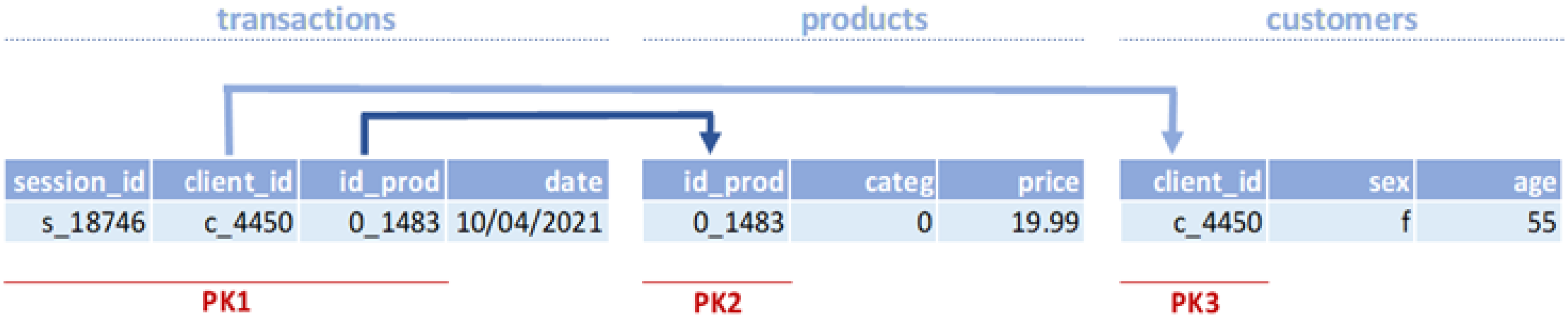
Identification des valeurs étrangères

- L'objectif premier est d'**analyser les ventes** (fichier 'transactions')

	session_id	client_id	id_prod	date
1431	s_0	ct_1	T_0	test_2021-03-01 02:30:02.237420

- Nous avons **3 clés potentielles** : session_id, client_id, id_prod
- date, qui devrait être un timestamp, contient des valeurs avec la chaîne de caractères 'test_'
- En sélectionnant ces valeurs, nous en avons 200 contenant le mot 'test_' qui donnent pour les **3 clés potentielles** les valeurs suivantes :
  - id_prod = T_0
  - session_id = s_0
  - client_id = ct_0, ct_1
- **Ces valeurs doivent être exclues** non seulement dans 'transactions', mais aussi dans les fichiers 'products' et 'customers' s'ils contiennent une de ces clés

Identfication des clés



Nous nous intéressons avant tout aux transactions

- La clé primaire de **transactions** est la résultante de la jonction des trois clés
- S'il n'y a pas de **transaction**, 'client_id' dans le df **customers** est inutile pour l'étude
- En revanche, il faut vérifier que la variable 'id_prod' de **products** est aussi renseignée dans **transactions** afin d'optimiser le dataset
- Pour joindre les trois fichiers, nous utiliserons donc un **left join** entre **transactions** et **customers** puis un **outer join** sur **products**

Nettoyage des 3 fichiers

Transactions

session_id	client_id	id_prod	timestamp	timeline	date	year	month	day	hour
s_12716	c_6624	1_264	2021-03-28 17:16:26.325707	2021-03	28-03-2021	2021	3	Sunday	17

- Transformation de la colonne date en timestamp
- Création de valeurs de temps

Nous avons

- 12 mois d'activité de mars 2021 à février 2022 pour 'timeline‘
- 2 valeurs pour 'year'
- 365 valeurs pour 'date‘
- 12 valeurs pour 'month‘
- 31 valeurs pour 'day'
- 24 valeurs pour 'hour‘

Customers

	client_id	sex	age
0	c_4410	f	55

- Pas de doublons dans client_id
- Transformation de date de naissance en âge (nous sommes en mars 2022)
- L’âge va de 18 à 93 ans

Products

	id_prod	price	categ
0	0_1421	19.99	0

- Pas de doublons dans id_prod
- Il y a 3 catégories
- Les prix vont de 0,62€ à 300€
- 75% des Prix sont inférieurs à 23€

Merge des tables et traitement des NaN

	session_id	client_id	sex	age	id_prod	price	categ	timeline	date	year	month	day	hour	timestamp
0	s_12716	c_6624	m	69.0	1_264	16.07	1.0	2021-03	28-03-2021	2021.0	3.0	Sunday	17.0	2021-03-28 17:16:26.325707

```
session_id      0
client_id       0
id_prod        22
categ          22
price          22
dtype: int64
```

```
session_id     103
client_id      103
id_prod        103
categ            0
price            0
dtype: int64
```

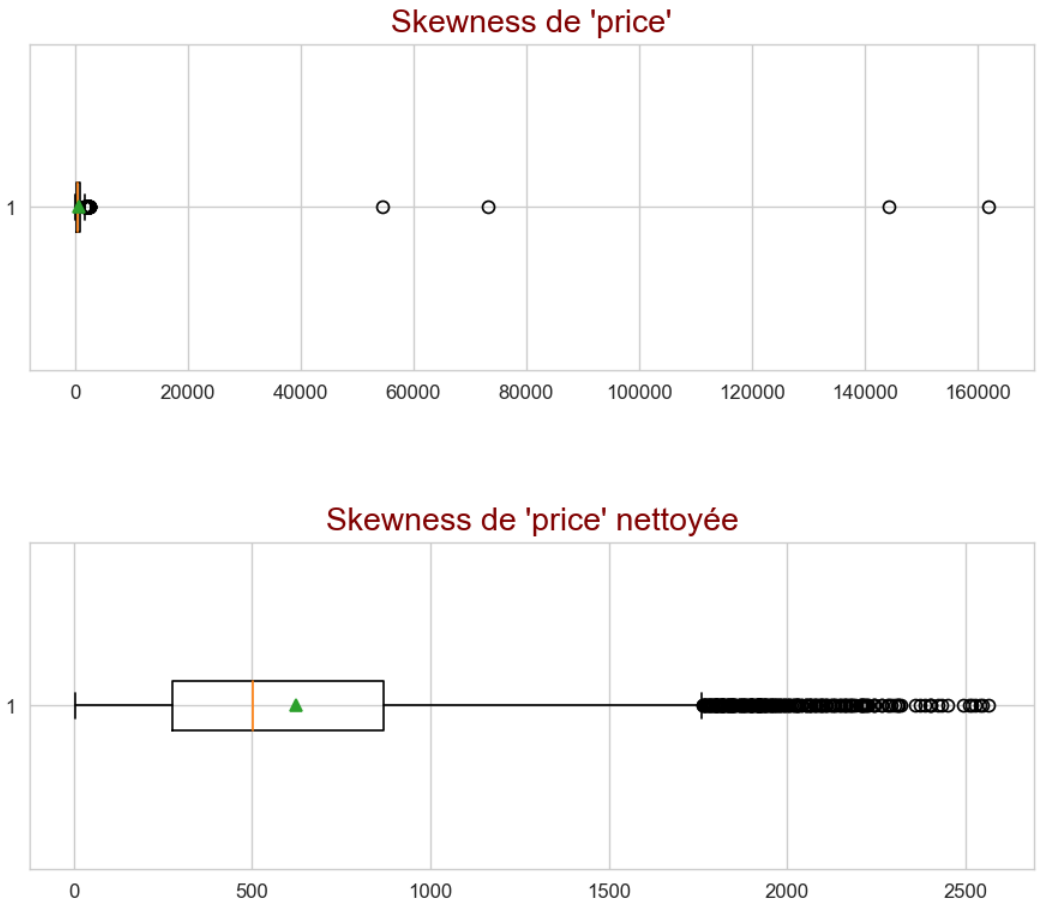
NaN dans session_id

- En cherchant dans session_id nous trouvons 22 NaN qui n'ont pas non plus de client_id
- Sans session_id, pas de transaction. Nous suprimons donc ces **22 valeurs**

NaN dans price

- En cherchant dans price nous trouvons **103 NaN** qui n'ont pas non plus de catégorie
- Toutes concernent le id_prod 0_2245
- Grâce à la nomenclature de id_prod nous en déduisons que ce produit est dans la catégorie 0
- Nous remplaçons les NaN de price par la médiane de la catégorie 0 soit **9,99€**
- A l'issue de cette opération, la somme des prix remplacés ne représente que 0,018% du total des ventes ce qui est négligeable et ne modifiera pas l'observation
- Nous aurions donc pu tout aussi bien suprimer également ces NaN

Contrôle de la cohérence des variables



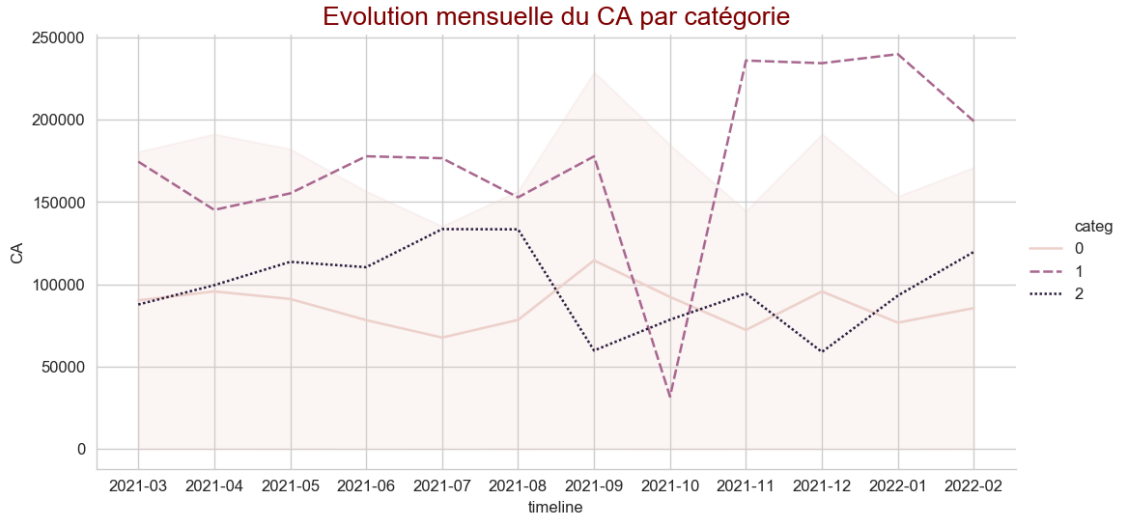
La somme des achats offre un montant maximum de 162.007€ alors que le 3e quartile n'est qu'à 871€

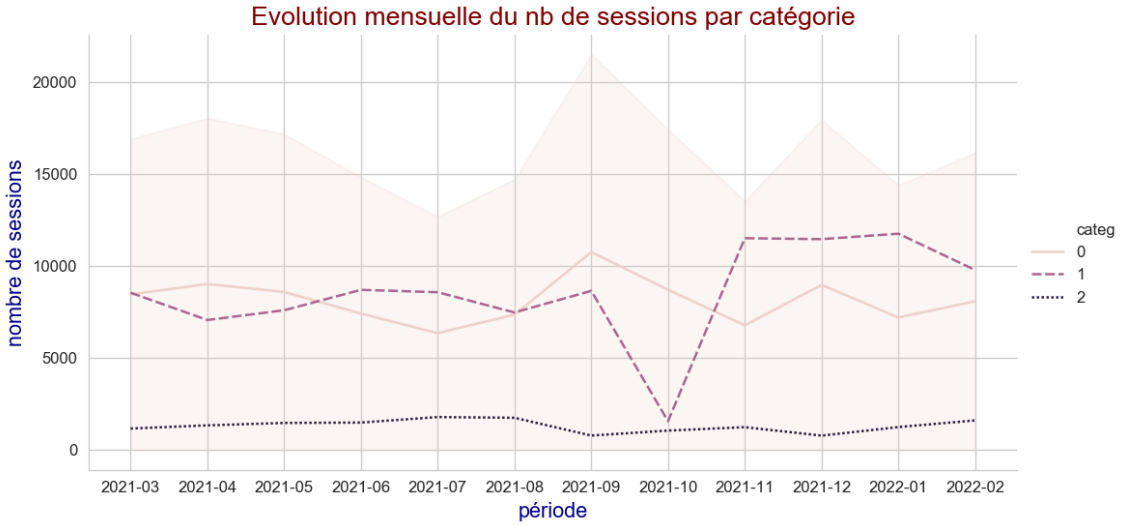
Nous trouvons **4 clients** avec des des différences notables

Nous supposons que ce sont des clients institutionnels et qu'il faut les exclure si l'on veut étudier le client lambda

Recherche temporelle

- En observant l'évolution mensuelle en nombre de sessions et par CA par catégories, nous observons un plongeon en **octobre** pour la catégorie 1
- Après une recherche sur octobre pour la cat1, nous voyons qu'il n'y a **pas de données entre le 2 et le 27 octobre**
- Les remplacement des valeurs manquantes n'est pas envisageable
- Pour les observations futures, il faudra décider
  - Retirer toutes les valeurs d'octobre et faire l'étude sur 11 mois ?
  - Laisser les valeurs en l'état en gardant en mémoire la situation ?
- Je choisis cette dernière solution, nous faisons une analyse des ventes, et les ventes peuvent être soumis à des aléas, comme des ruptures de stock



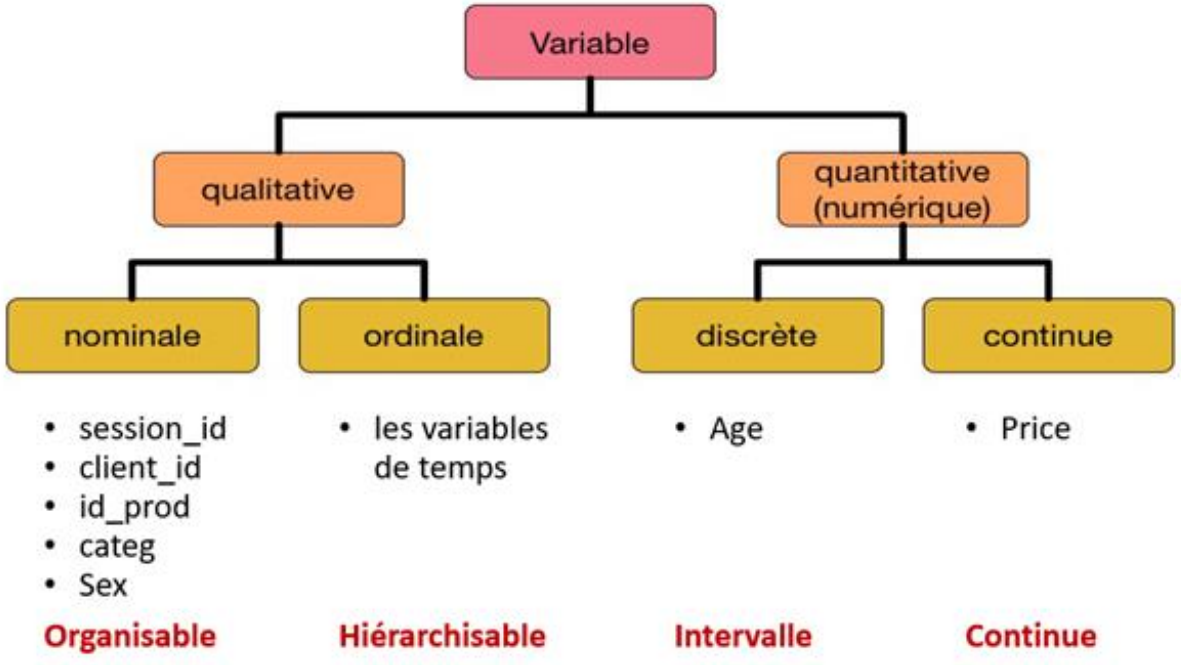


Question 2

Analyse des données

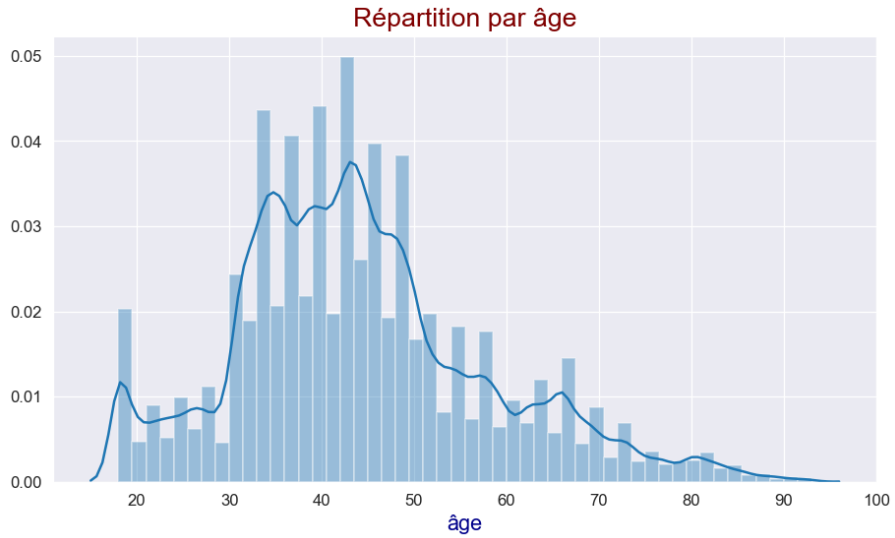
Les types de variables

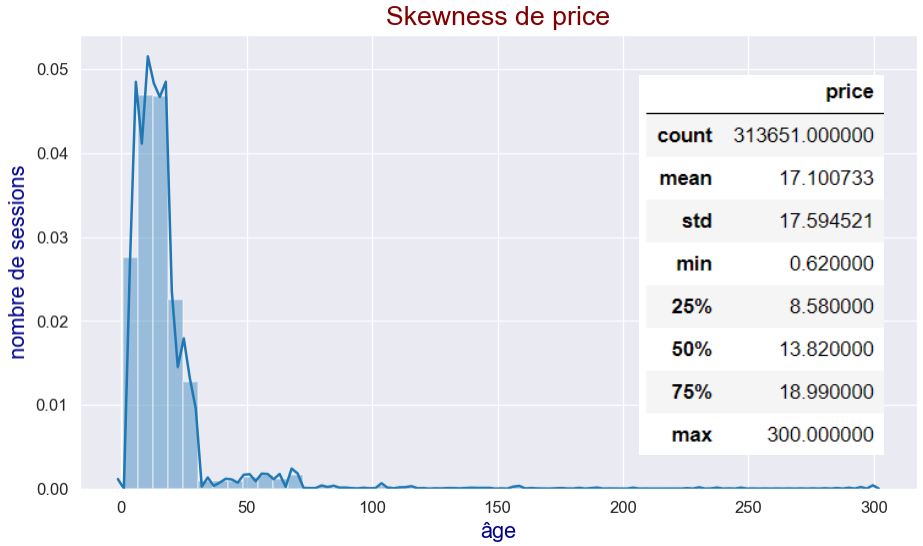
	index	session_id	client_id	sex	age	id_prod	price	categ	timeline	date	year	month	day	hour	timestamp
0	0	s_12716	c_6624	m	69	1_264	16.07	1	2021-03	28-03-2021	2021	3	Sunday	17	2021-03-28 17:16:26.325707



- Nous ne disposons de 2 valeurs **quantitatives**, **age** et **price**
- Les variables de temps sont **qualitatives ordinales**
- Les sessions, les clients, les produits, les catégories et le sexe sont des variables **quantitatives nominales**

Âge et price





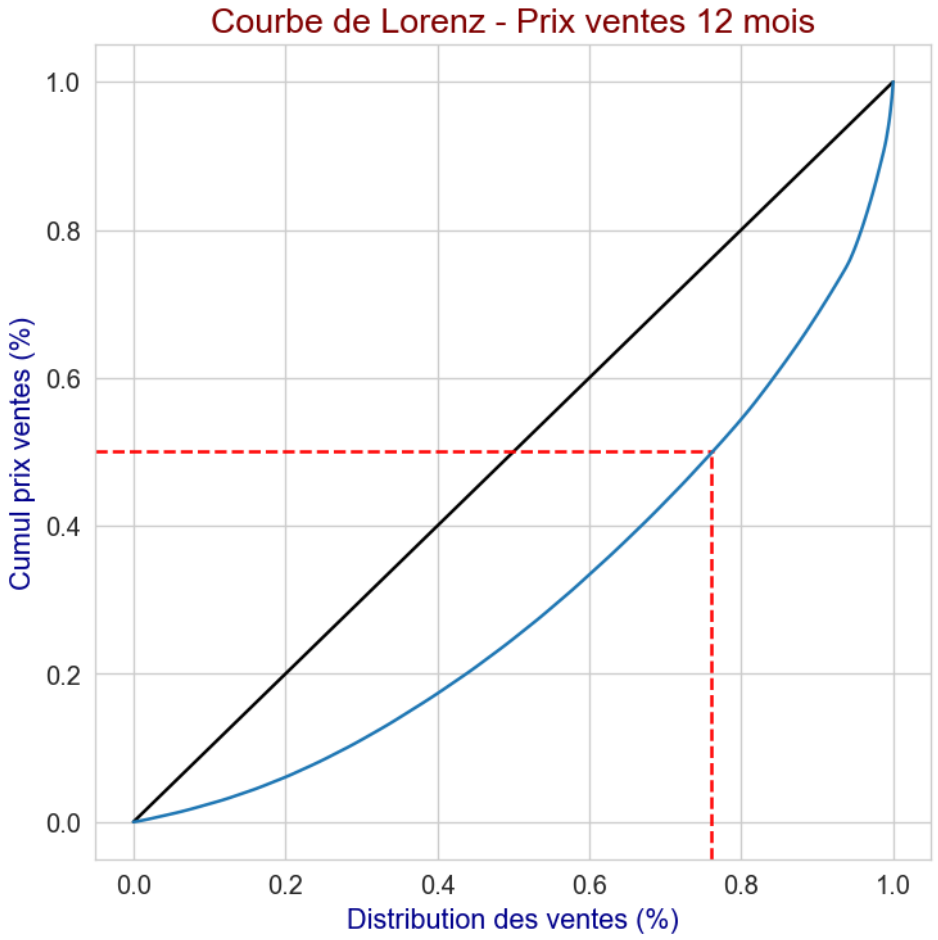
Âge

- Il y a des écarts suspects entre les tranches d'une même classe d'âge. Sans doute les barres pourraient être lissées si nous disposions également du mois et du jour de naissance
- Pour 18 ans, c'est un oulier : peut-être qu'on a mis dans cette catégorie des clients mineurs. Ces données **devront être exclues statistiques qui utilisent la variable 'age'**
- La tranche 30-50 ans est clairement sur-représentée
- Les tranches 18-30 et 50-70 sont très grossièrement égales en termes de répartition
- Au-delà de 70 ans, le nombre décroît en pente douce

Price

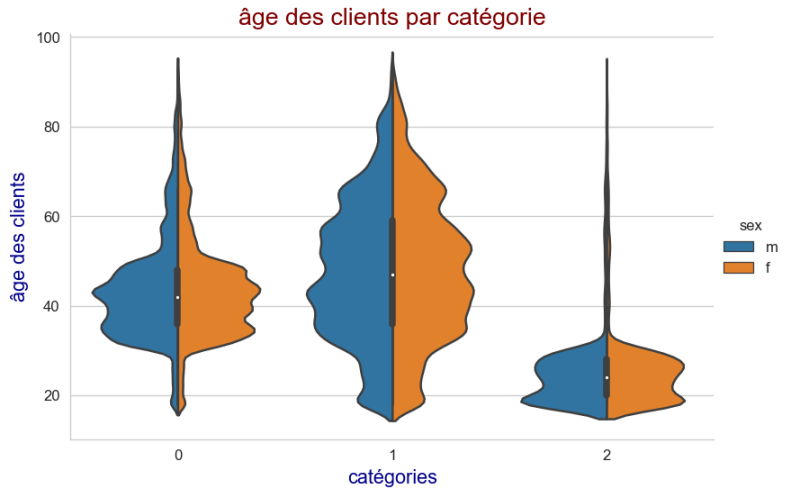
- 75% des prix des produits est inférieur à 19€
- La moyenne est plus proche du 3e quartile que de la médiane du fait du skewness à droite
- Graphiquement on voit que c'est très concentré entre 0 et 30, qu'il reste une petite portion entre 30 et 70 et que ce qui reste est en très faible quantité

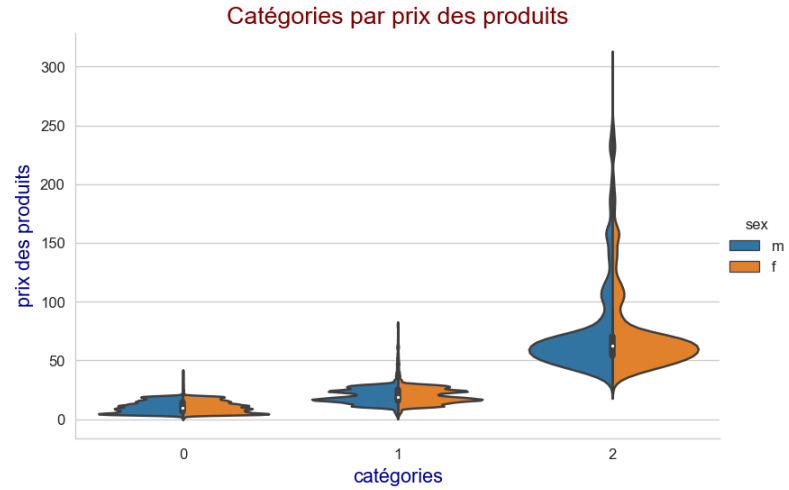
Courbe de Lorenz et indice Gini sur les ventes



- Avec un **indice Gini = 39%** (même plus proche de 0 que de 1), on commence à sentir une certaine inégalité dans la répartition des ventes
- Si l'on s'en tient à la **médiale**,
  - **50% du cumul** des ventes représente environ 76% de la distribution des ventes les plus basses
  - **50% du cumul** des ventes représente environ 24% de la distribution des ventes les plus hautes

Catégories par âge, prix et sexe





Par âge

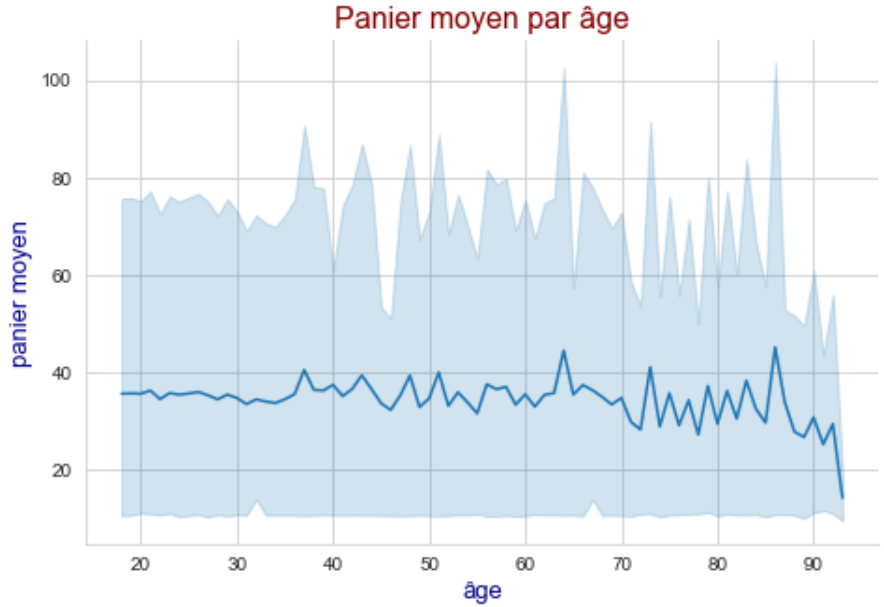
- La répartition par âge et catégories donne une **image quasi symétrique par sexe en fonction de l'âge**
- Une concentration de la **catégorie 0** pour les 30-50 ans, 1e quartile à 35 ans et 3e quartile à 50 ans
- Une répartition plus diffuse de la **catégorie 1** malgré un étrécissement pour les moins de 30 ans, 1e quartile à 35 ans et 3e quartile à 60 ans
- En revanche, elle nous indique une très forte concentration des 18-30 ans pour la **catégorie 2**, 1e quartile à 18 ans et 3e quartile à 30 ans

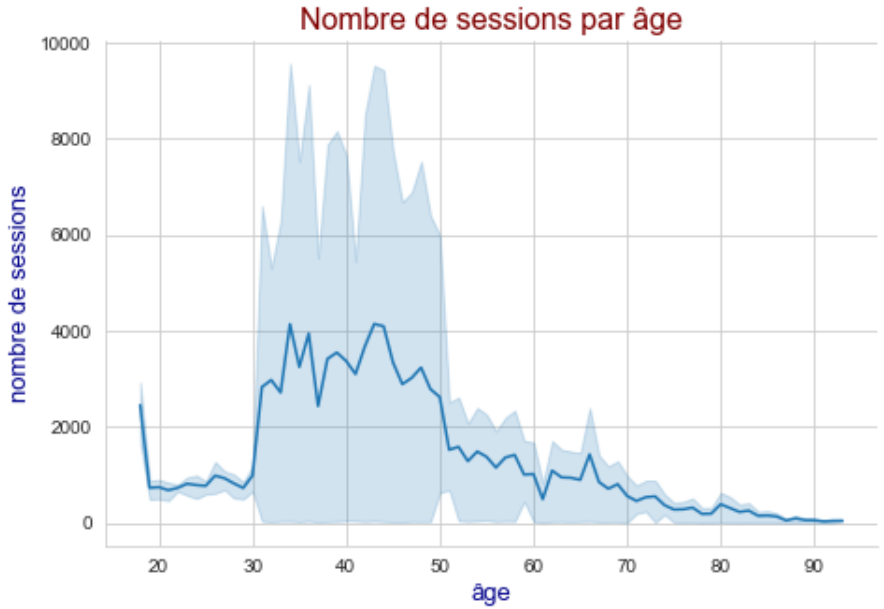
Par prix

- Les catégories semblent être **réparties symétriquement entre hommes et femmes.**
- Les violons sont plutôt concentrés, à part pour la catégorie 2 dont le skew est très allongé
- En regardant les espaces interquartiles, on peut observer une possible correlation entre le prix et la catégorie

Analyse globale

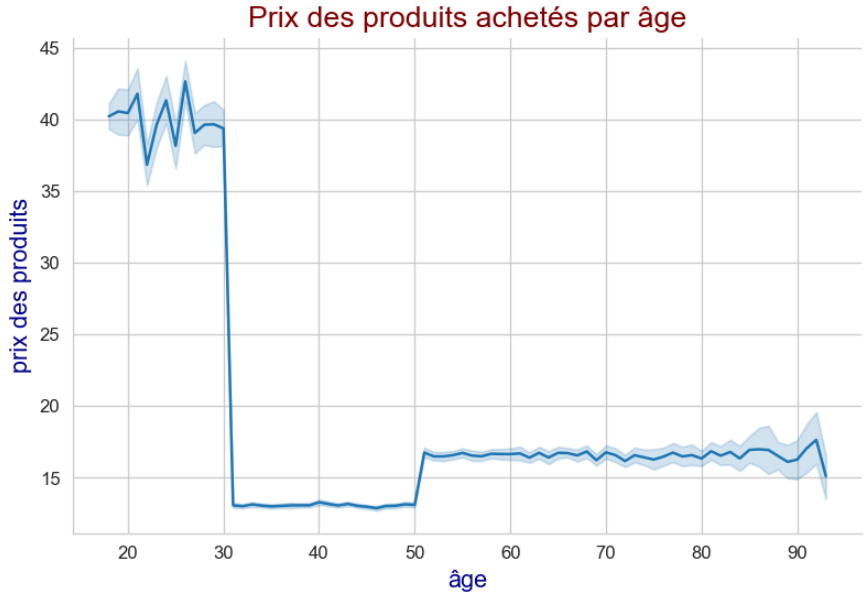
- La répartition du **panier moyen** est relativement homogène quel que soit l'âge aux alentours de 38€ avec une amplitude de 7 à 75€
- Le **nombre de sessions** est également tributaire de l'âge : base et homogène pour les 19-30 ans, haute avec une grand variation pour les 31-50 ans, moyenne et décroisssante pour les plus de 50 ans

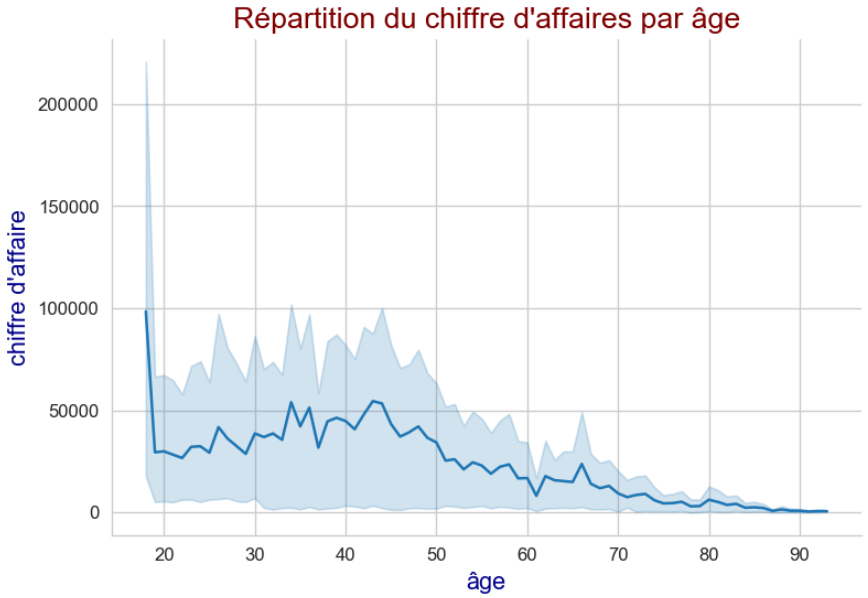




Analyse globale

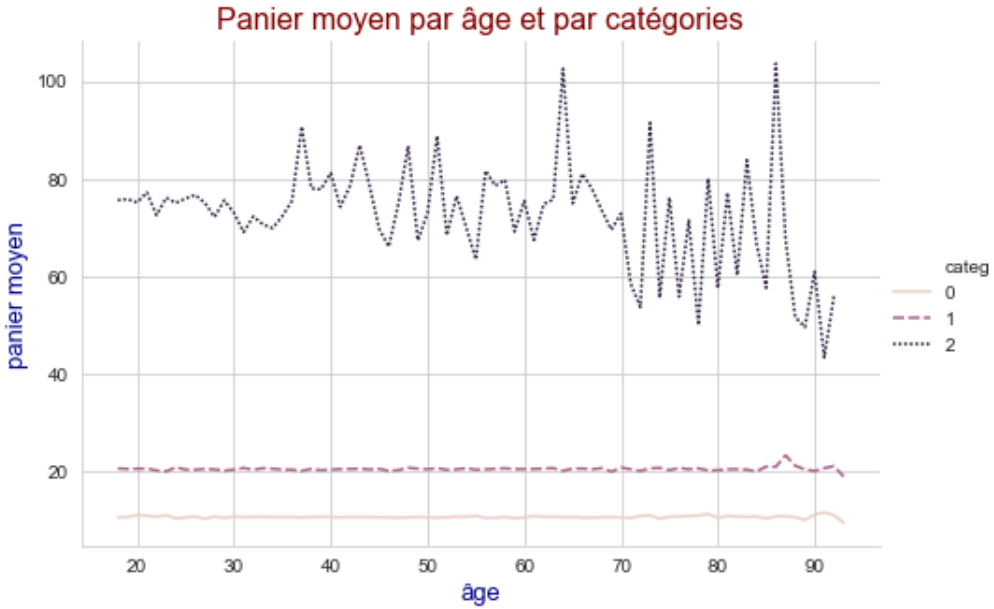
- Le prix des **produits achetés** est caractérisé par l'âge environ 40€ pour les moins de 30 ans, environ 12€ pour les 31-50 ans et 17€ pour les plus de 50 ans
- La répartition **du CA** est finalement plutôt homogène de 19 à 50 ans. En effet, si les 31-50 ans achètent des produits à des prix nettement plus bas que les 19-30 ans, leur nombre de sessions est nettement plus élevé

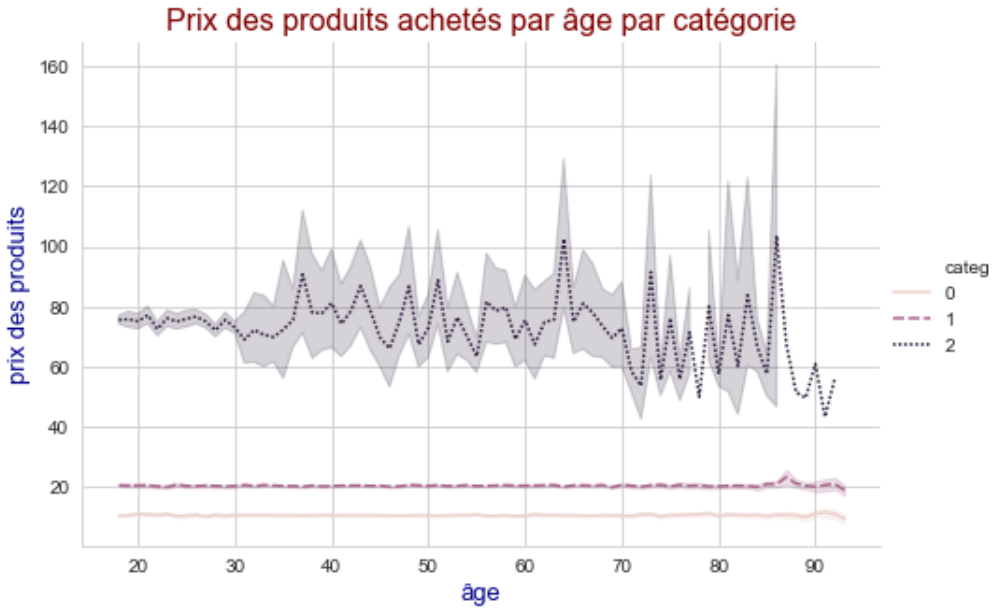




Analyse par catégories

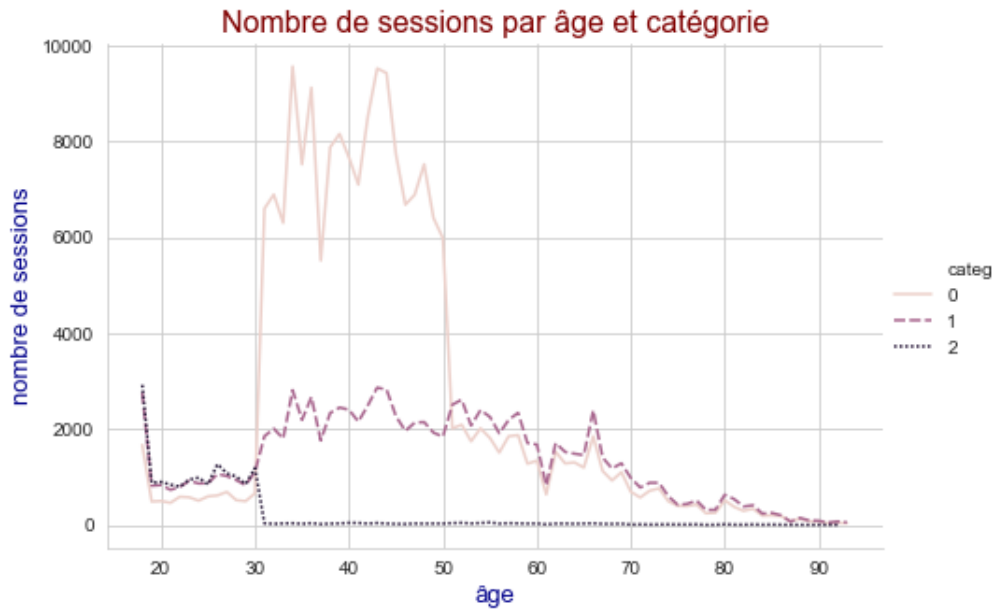
- On voit, tant pour le **panier moyen** que pour le Prix des **produits achetés** que les catégories 0 (environ 9€) et 2 (environ 20€) sont très homogènes et donc très predictibles.
- La courbe pour la catégorie 2 est beaucoup plus ératique mais se situe en moyenne aux alentours de 75€

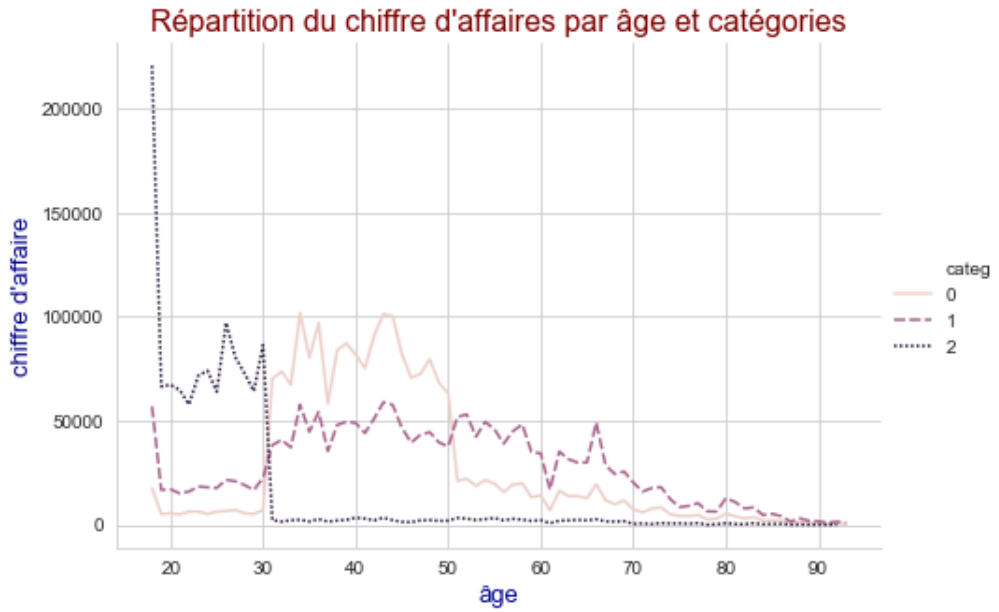




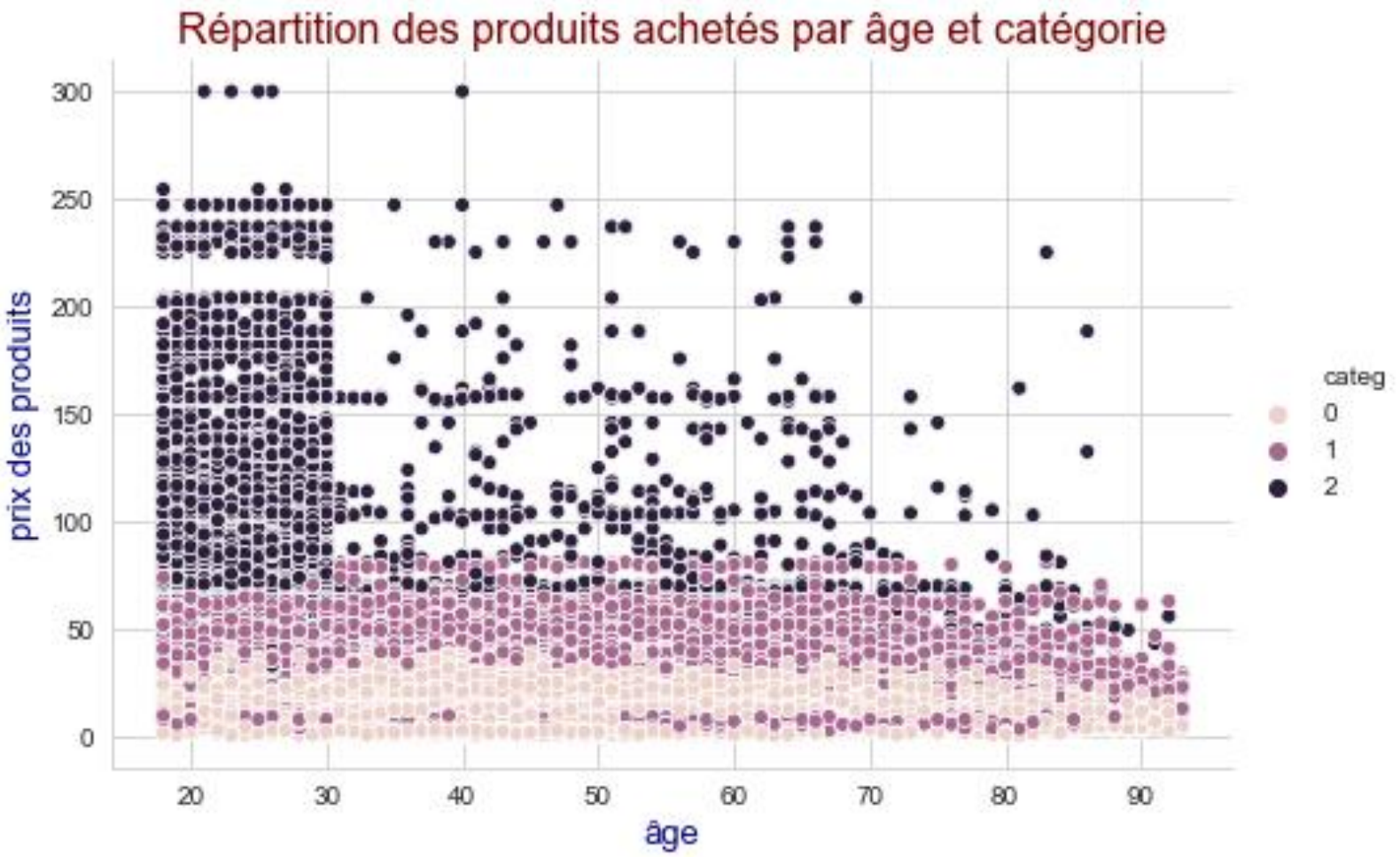
Analyse par catégories

- Pour le **nombre de sessions** et la **répartition du CA**, les 31-50 ans sont les plus actifs, mais achètent en majorité des produits moins chers (cat0).
- Les 19-30 ans réalisent moins de sessions mais achètent des produits de la cat2 aux prix plus élevés
- La catégorie 1 est mieux répartie avec 50% des ventes entre 30 et 60 ans





Boxplot produits achetés par catégorie



Ce boxplot par prix des produits, âge et catégories résume en un seul graphique ce que nous avons découvert précédemment

Question 3

Tests statistiques

Corrélation entre genre et catégories de produits achetés

Hypothèse nulle (H0) : la catégorie de produits achetés est indépendante du sexe des clients

Hypothèse alternative (H1) : la catégorie de produits achetés est dépendante du sexe des clients

- Le khi2 total est de 10.111
- Calcul du **degré de liberté** = (nb de lignes - 1) * (nb de colonnes - 1) **DDL = 2**
- Recherche de la valeur de seuil et de la position du Chi2 calculé dans la Table de la loi du khi-deux
  - **α =0.05**
  - **Valeur de seuil = 5.99**
- Le **khi2 de 10.111** et supérieur à la valeur de seuil de 5.99
- La **pvalue** (la probabilité) que l'hypothèse H0 se réalise est inférieure à 1%
- On donc peut rejeter l'hypothèse H0

Contingences observées

sex	0_obs	1_obs	2_obs	tot_obs
f	101206	53774	8122	163102
m	94064	48851	7634	150549
total	195270	102625	15756	313651

Contingences attendues

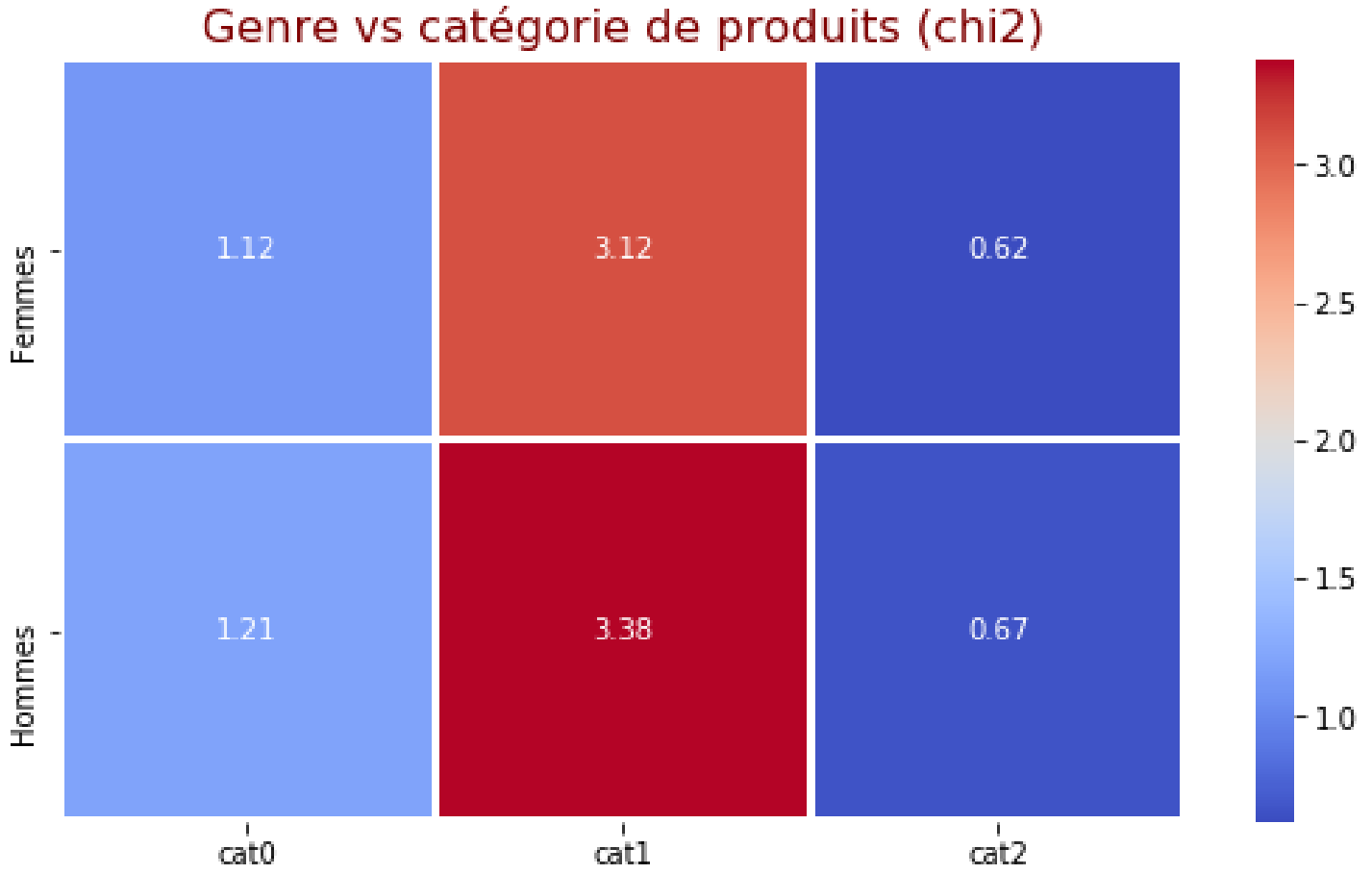
0_att	1_att	2_att
101543.0	53366.0	8193.0
93727.0	49259.0	7563.0
195270.0	102625.0	15756.0

Khi2 genre/catégories

khi2_0	khi2_1	khi2_2
1.118433	3.119289	0.615281
1.211700	3.379362	0.666534
0.000000	0.000000	0.000000

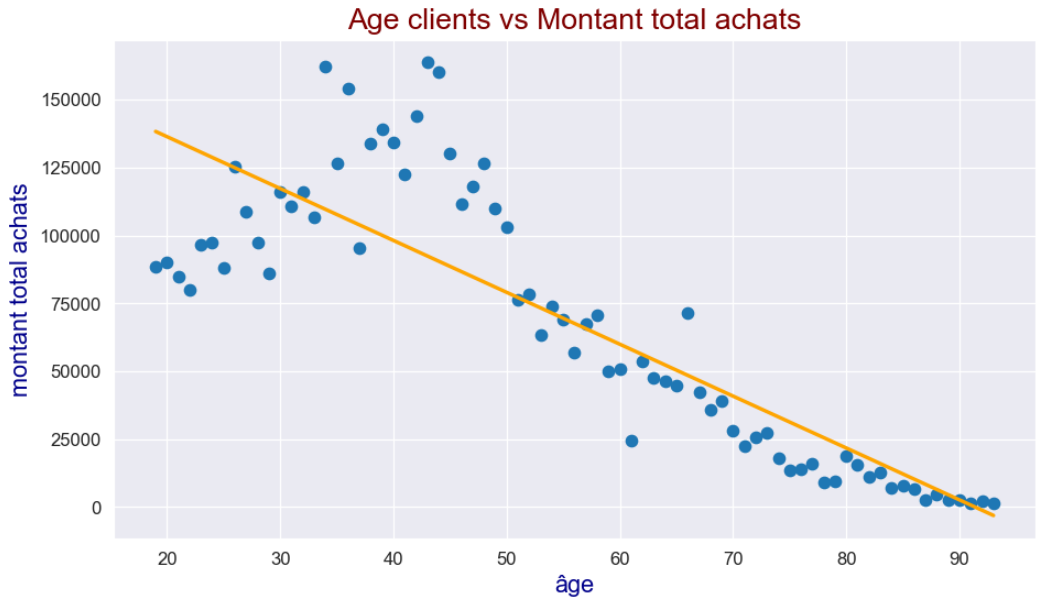
Corrélation entre genre et catégories de produits achetés

- On peut constater avec cette hitmap que, si nous sommes dans l'hypothèse H1, la différence entre sexe n'est pas spectaculaire
- La différence la plus marquée est dans la catégorie 1 oú l'écart est de 0.26
- **Pour affiner le résultat**, nous pourrions créer des catégories par sexe et par âge en utilisant les 3 populations précédemment identifiées : **<=30 | 31-50 | >50**, soit 3 autres tableaux de 6 cellules



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

- age et montant des achats sont des variable quantitatives basées sur la **somme** de price
- l'âge entrant en ligne de compte, nous exclurons l'outlier **18 ans**
- Pour ce type de variables nous utilisons le coefficient de corrélation linéaire de type **y =ax + b + epsilon**
- Nous utiliserons une régression linéaire et la méthode OLS (Ordinary Least Square) ou **méthode de moindre carrés**



Coefficient de corrélation de Pearson = -085

On pourrait dire qu'il y a une corrélation linéaire négative assez élevée pour cet ensemble de données, mais le graphique le dément partiellement

Par soucis de cohérence avec les observations précédentes et futures nous allons séparer en 3 groupes d'âge

- de 19 à 30 ans, corrélation positive moyenne
- de 31 à 50 ans, pas de corrélation
- à partir de 51 ans une forte corrélation négative

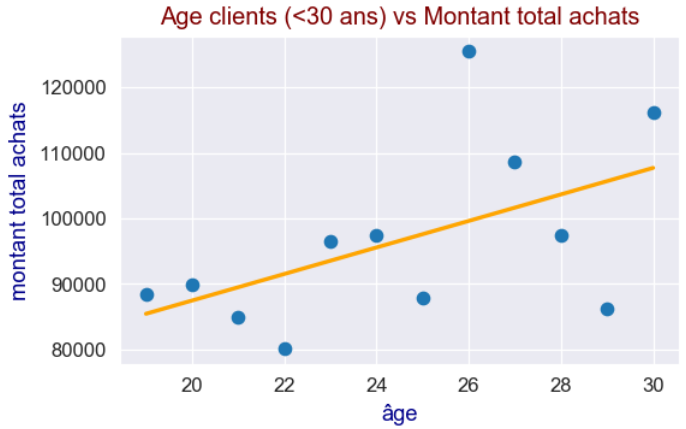
Corrélation entre l'âge des clients et le montant total des achats

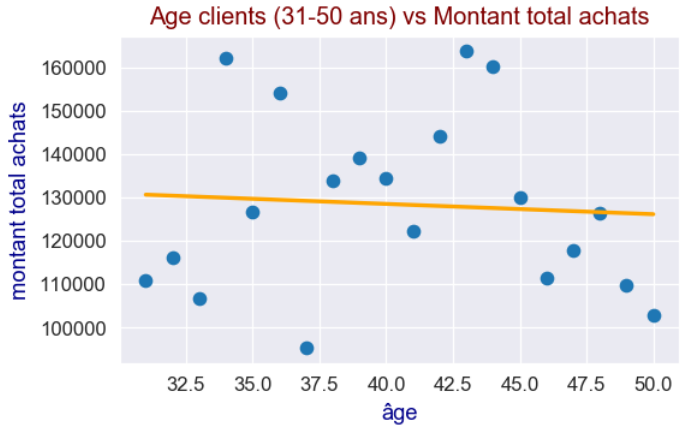
Conclusions

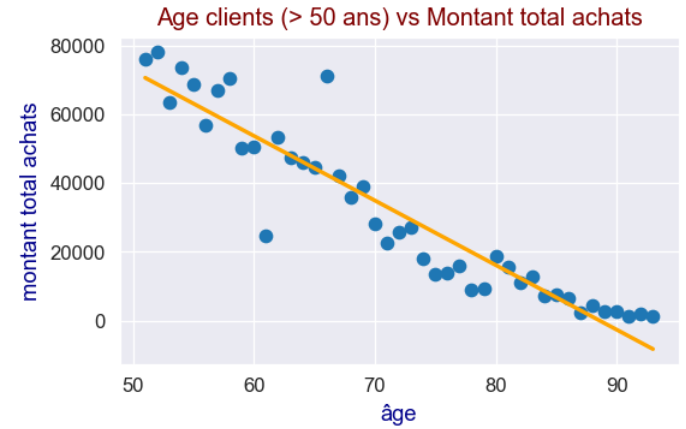
- Le coefficient de corrélation pour les **moins de 30 ans = 0.532**
  - la courbe est ascendante, la corrélation moyenne
  - le modèle semble peu solide du fait de la grande dispertion des valeurs

- Le coefficient de corrélation pour les **31 - 50 ans =-0.069**
  - pas de corrélation

- Le coefficient de corrélation pour les **plus de 50 ans = -0.939**
  - la courbe est descendante, la corrélation forte
  - le modèle semble solide du fait de la concentration des valeurs mis à par 2 outliers

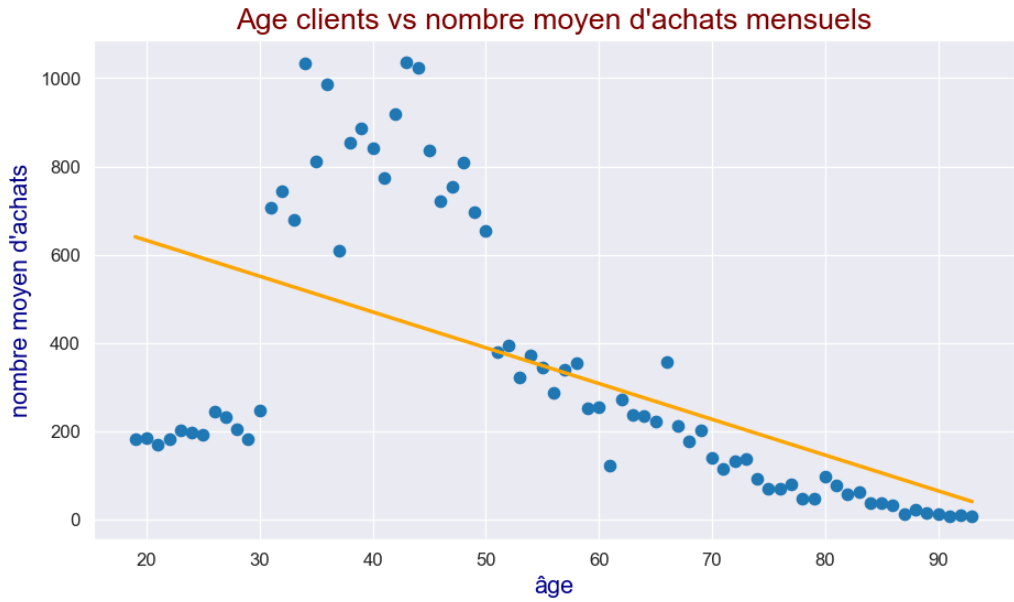






Corrélation entre l'âge des clients et le nombre d'achats

- Nous sommes encore dans un exemple avec deux valeurs quantitatives, mais cette fois-ci entre l'âge et la **fréquence moyenne mensuelle d'achats**
- L'âge entrant en ligne de compte, nous exclurons l'outlier **18 ans**
- Pour ce type de variables nous utilisons le coefficient de corrélation linéaire avec
  - **y =ax + b + epsilon**
- Nous utiliserons une régression linéaire et la méthode OLS (Ordinary Least Square) ou **méthode de moindre carrés**



coefficient de corrélation Pearson = -0.585

Nous avons une droite de régression décroissante est assez peu corrélée sauf au-delà de 50 ans

Nous pouvons découper la courbe en 3 zones : 19-30 ans, 31-50 ans, plus de 50 ans

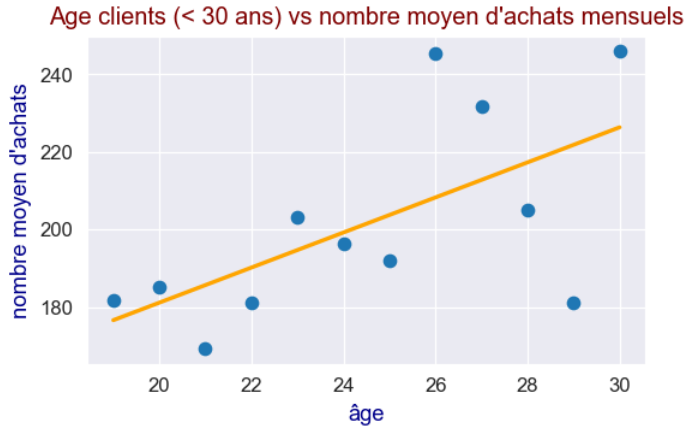
Corrélation entre l'âge des clients et le nombre d'achats

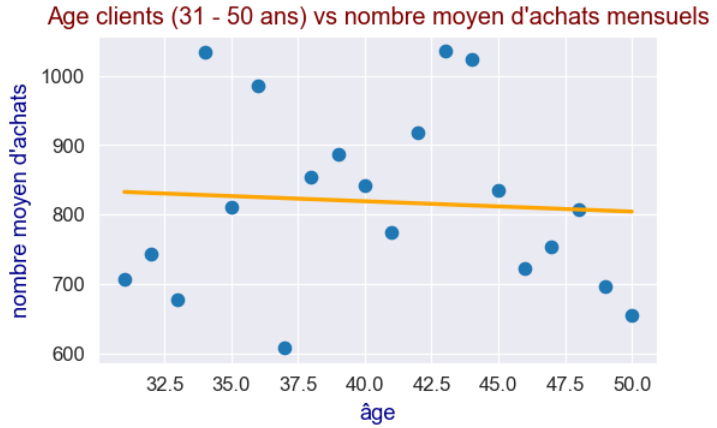
Conclusions

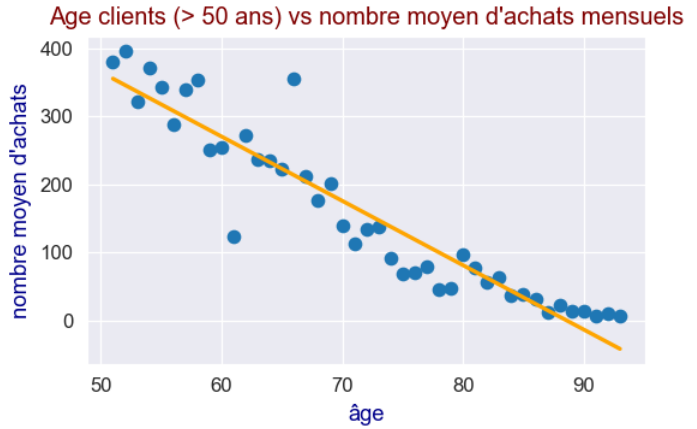
- Le coéficient de corrélation des **moins de 30 ans = 0.626**
  - la courbe est ascendante, la corrélation moyenne
  - le modèle semble peu solide du fait de la grande dispertion des valeurs

- Le coéficient de corrélation pour les **31- 50 ans =-0.068**
  - pas de corrélation

- Le coéficient de corrélation des **plus de 50 ans = -0.939**
  - la courbe est descendante, la corrélation forte
  - le modèle semble solide du fait de la concentration des valeurs mis à par 2 outliers







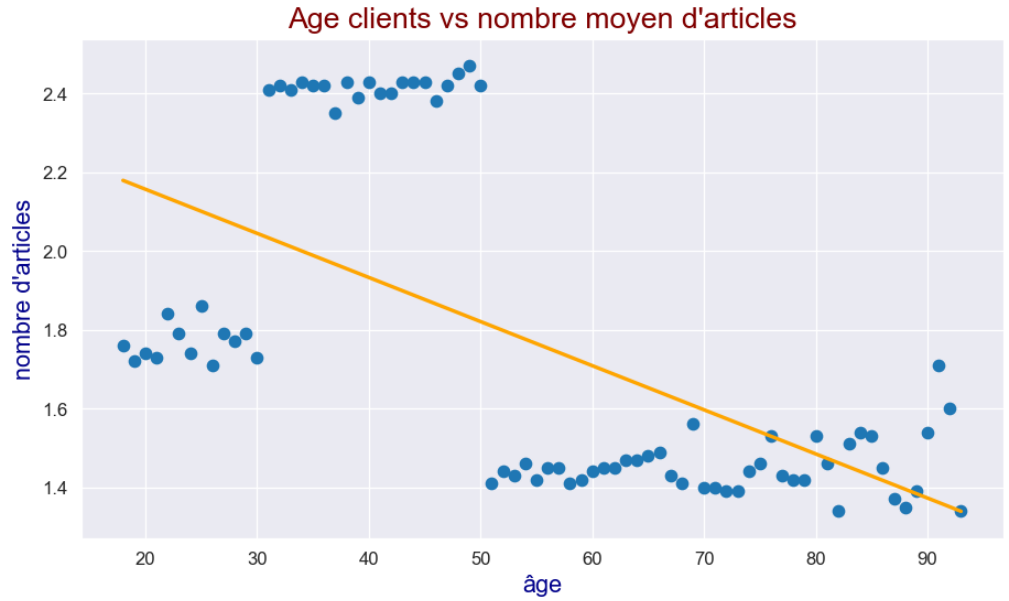
Corrélation entre l'âge des clients et la taille du panier moyen (en nombre d'articles)

- Nous sommes dans un exemple avec deux valeurs quantitatives, mais cette fois-ci entre l'âge et la **taille du panier moyen**
- L'âge entrant en ligne de compte, nous excluerons l'outlier **18 ans**

Pour ce type de variables nous utilisons le coefficient de corrélation linéaire avec

  - **y =ax + b + epsilon**

- Nous utiliserons une régression linéaire et la méthode OLS (Ordinary Least Square) ou **méthode de moindre carrés**



coefficient de corrélation Pearson = -0,594

Les données sont assez peu corrélées, ce qui est confirmé par le graphique

On distingue clairement 3 zones qui que l'on pourrait calculer individuellement : 18-30 ans, 31-50 ans, 51-90 ans. Au delà, il y a peu de clients et ce sont des outliers

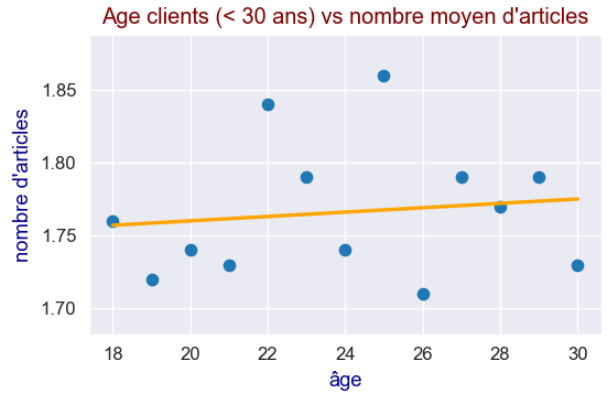
Corrélation entre l'âge des clients et la taille du panier moyen

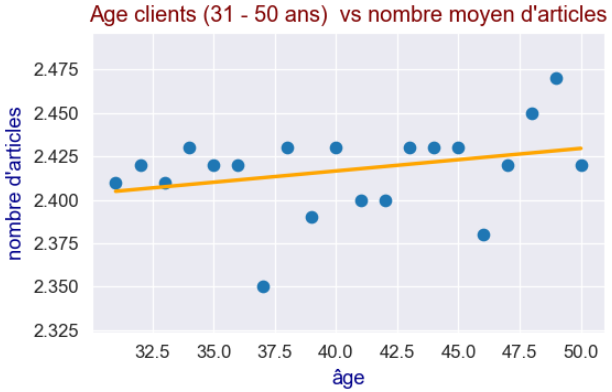
Conclusions

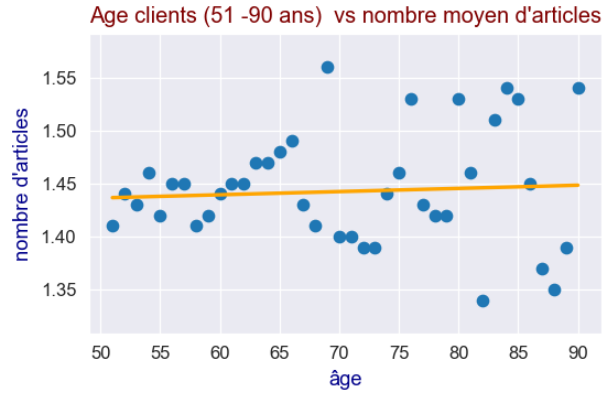
- Le coéficient de corrélation des **moins de 30 ans = 0.131**
  - pas de corrélation

- Le coéficient de corrélation pour les **31- 50 ans = 0.302**
  - pas de corrélation

- Le coéficient de corrélation des **plus de 50 ans = 0.077**
  - Pas de corrélation

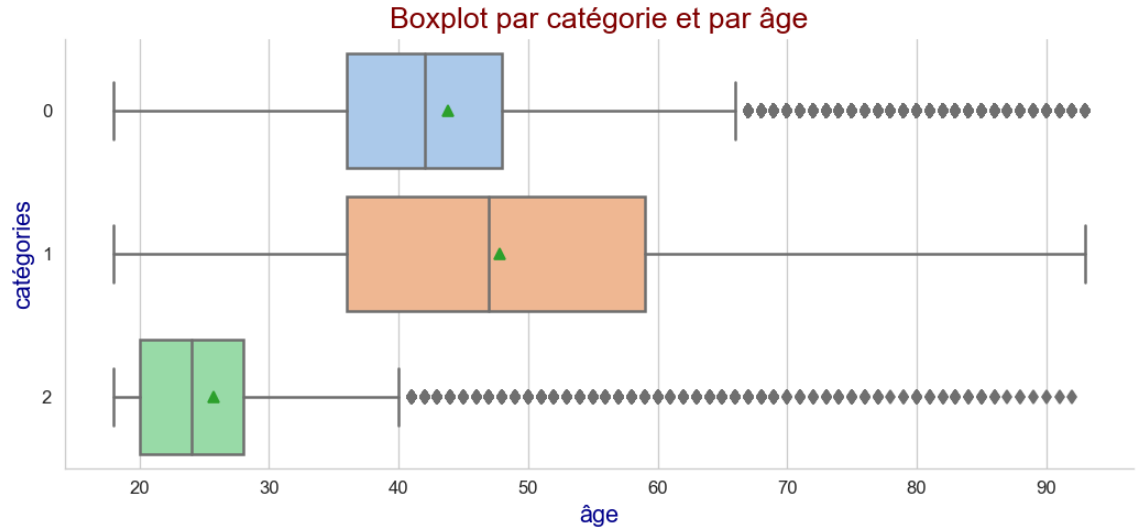






Corrélation entre l'âge des clients et les catégories de produits achetés

- Nous avons une variable quantitative discrète (l'âge) et une qualitative nominale (la catégorie)
- Nous utiliserons la méthode **ANOVA**. Nous allons déterminer la somme des carrés intra-classes et inter-classes
- Il faudra utiliser le test de Fisher pour la vérification des hypothèses



Conclusions

- Nous avons un skewness à droite très allongé pour les catégories 0 et 2
- Nous avons un léger skewness à droite pour la catégorie 1

- 50% des acheteurs de la catégorie 0 a **entre 36 et 48 ans**
- 50% des acheteurs de la catégorie 1 a **entre 36 et 59 ans**
- 50% des acheteurs de la catégorie 2 a **entre 20 et 28 ans**

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

ANOVA

- La Somme des Carrés Totaux **SCT** est la somme des distances au carré entre chaque valeur observée et la moyenne globale ou **Grand Mean**
- Elle peut être décomposée en deux éléments :
  - La somme des Carrés Factoriels **SCF** imputable aux modalités de la variable étudiée (les catégories)
  - La Somme des Carrés Résiduels **SCR** est la somme des distances au carré entre chaque valeur observée et la moyenne de chaque catégorie

SCT = SCF + SCR
variation totale = variation interclasse + variation intraclasse

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

Dispersion totale des données

- **SCT** = 59705564.51

Somme des Carrés Factoriels

- **SCF0** = 32297.85
- **SCF1** = 1335411.37
- **SCF2** = 5366499.27
- **SCF totale** = 6734208.49

Somme des Carrés Résiduels

- **SCR0** = 25881944.2
- **SCR1** = 25627629.1
- **SCR2** = 1461782.73
- **SCR totale** = 52971356.02

Variance

- **CMF** = 3367104.25
- **CMR** = 168.89

Test statistique de Fisher (F-test)

- **F** = CMF/CMR = 19936.92

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

Seuil de tolérance α : 0.05 (5%) pour l'hypothèse nulle

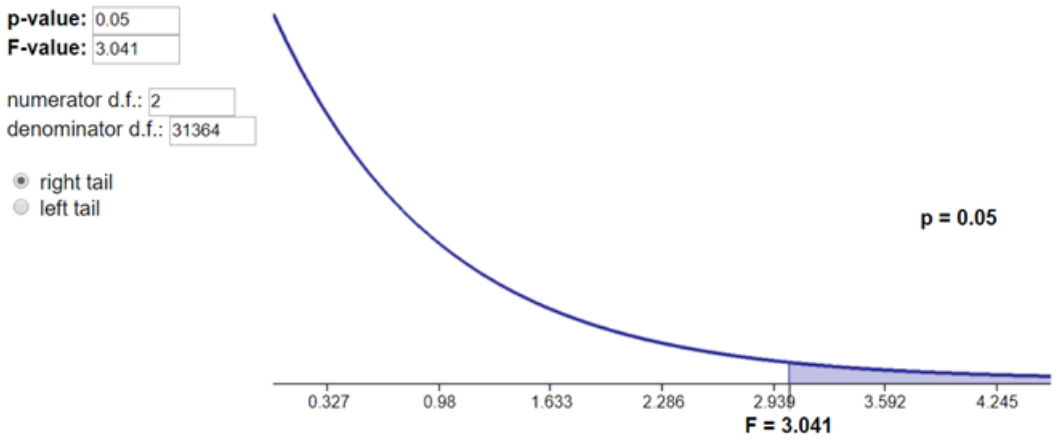
Test d'hypothèse

- **H0** = Les catégories nont pas d'impact (les moyennes des différentes catégories sont égales)
- **H1** = Les catégories ont un impact (aux moins 2 moyennes sont différentes)

Conclusions

- Notre valeur de F calculée (19936.92) est très largement supérieure à la valeur F critique (3.041)
- L'hypothèse H0 est à rejeter
- Au moins deux catégories ont un impact très sigificatif

Calcul de la valeur F critique pour **α** = 0.05



Des questions ?

lgourdon@gmail.com (+34 659 34 05 15)