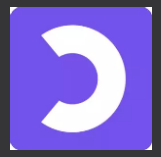
Lapage

Librairie en ligne
2 ans d'activités

Projet P6 - Analysez les ventes d'une librairie



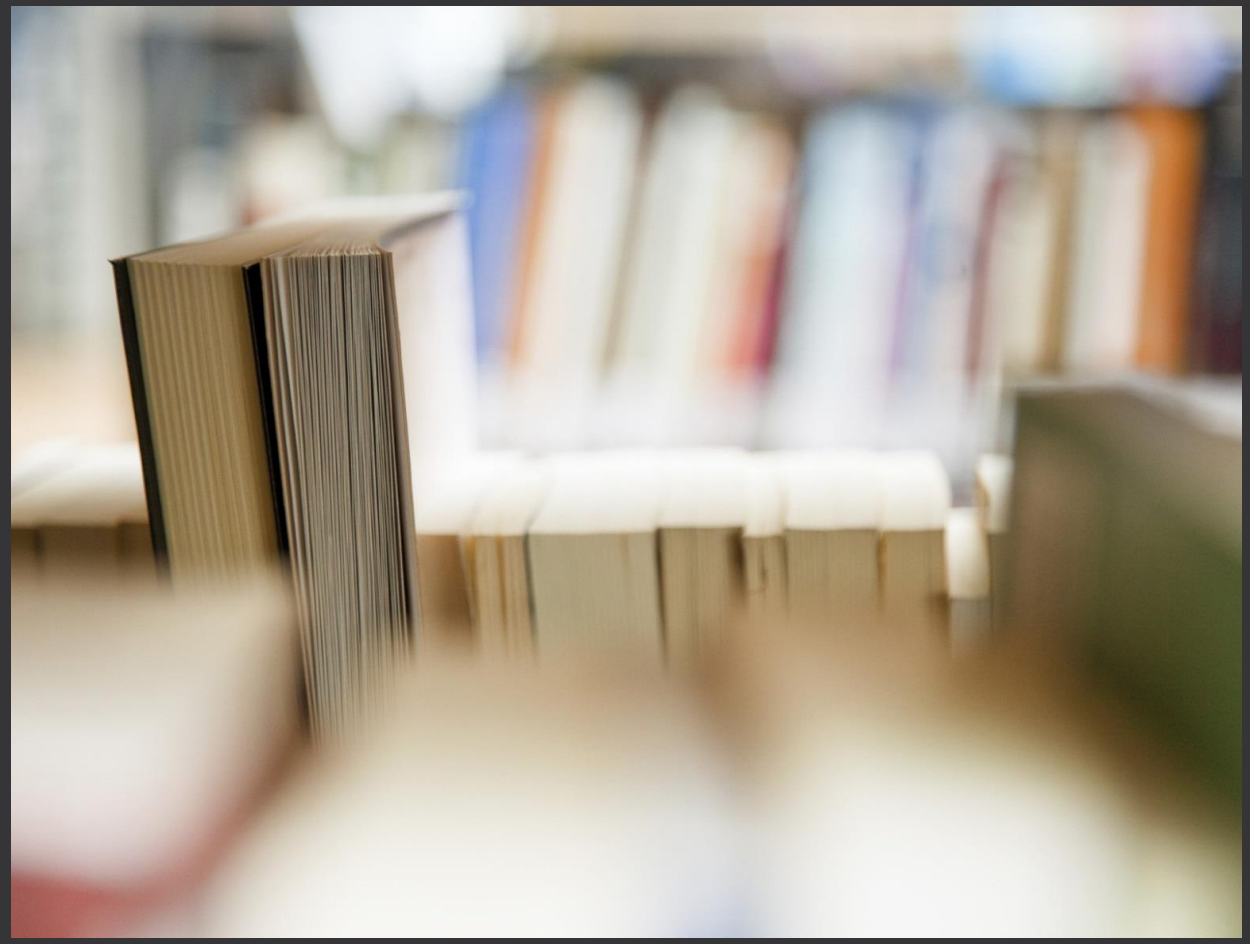
PARCOURS DATA ANALYST_V2

ADELINE LE RAY

Sommaire

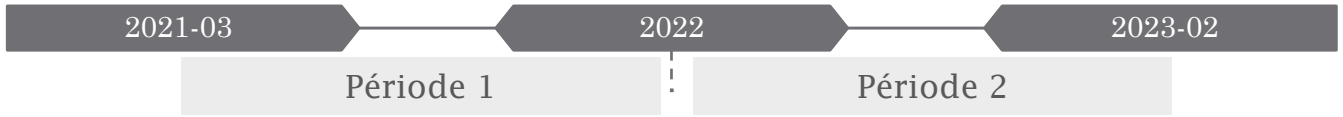
Lapage

Chiffres clés

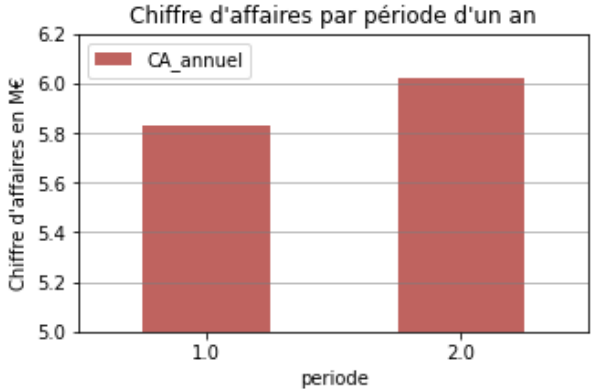


Lapage en ligne : chiffres en hausse

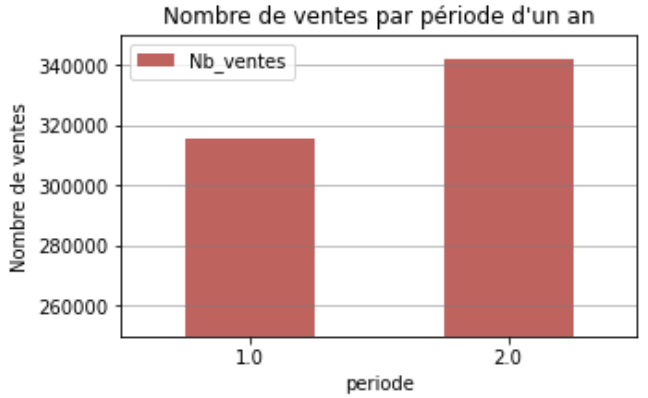
Lancement de la librairie en ligne — Aujourd'hui



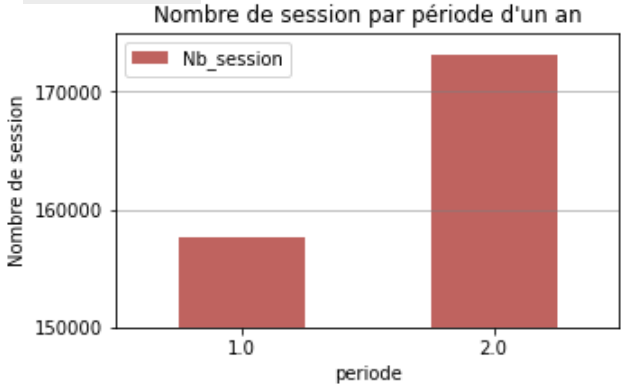
+3% *11,9 M€ sur 2 ans*



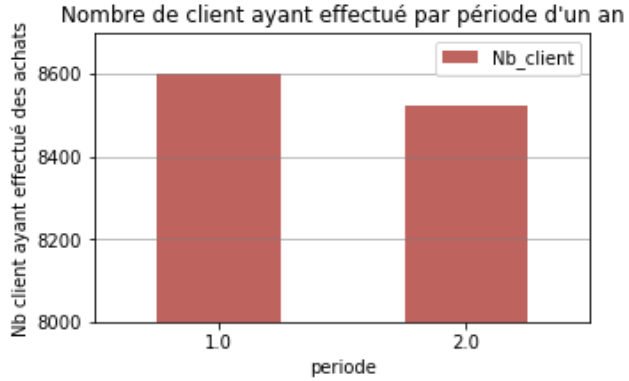
+8% *657 726 ventes sur 2 ans*



+10% *330 769 sessions sur 2 ans*



-1% *8 598 clients*



En bref

Meilleurs clients : les 30-50 ans

Le plus d'achat, presque 50% du chiffre d'affaire

99% des clients fidélisés

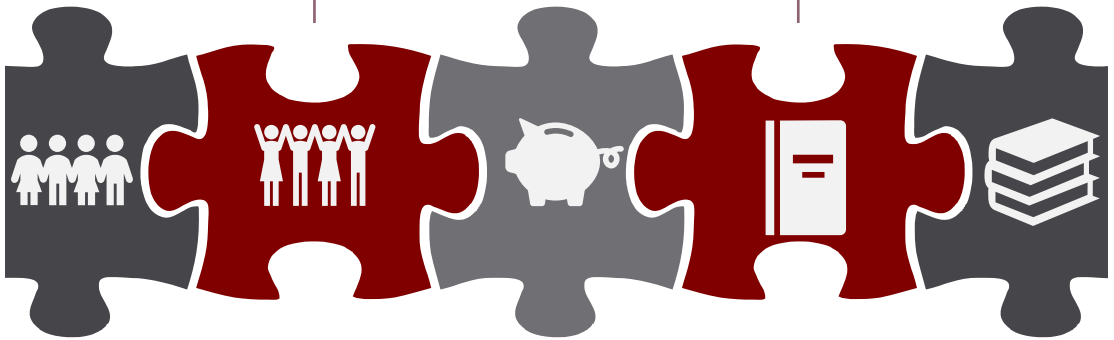
Les clients reviennent mais pas de nouveau client

Produits phares : la catégorie 1

Top des ventes, 40% du CA, achetés par 98% des clients

Catégorie 2 : cible < moins de 30 ans

Les principaux acheteurs de la catégorie 2 ont moins de 32 ans



Chiffre d'affaires stabilisé

16 500 € / jour

Lapage

Nettoyage

Principales modifications apportées au jeu de données lors du nettoyage

Données de tests

200 transactions tests

	id_prod	date	session_id	client_id
3019	T_0	test_2021-03-01 02:30:02.237419	s_0	ct_0
52424	T_0	test_2021-03-01 02:30:02.237419	s_0	ct_0
130188	T_0	test_2021-03-01 02:30:02.237419	s_0	ct_0
168341	T_0	test_2021-03-01 02:30:02.237419	s_0	ct_0
185962	T_0	test_2021-03-01 02:30:02.237419	s_0	ct_1
311604	T_0	test_2021-03-01 02:30:02.237419	s_0	ct_1

Les données de tests, transactions et produit, ont été supprimées.

Produit T_0 avec prix négatif

	id_prod	price	categ
731	T_0	-1.0	0

Un produit sans prix de ventes

Livre 0_2245 : 221 transactions mais absent du listing ‘products’

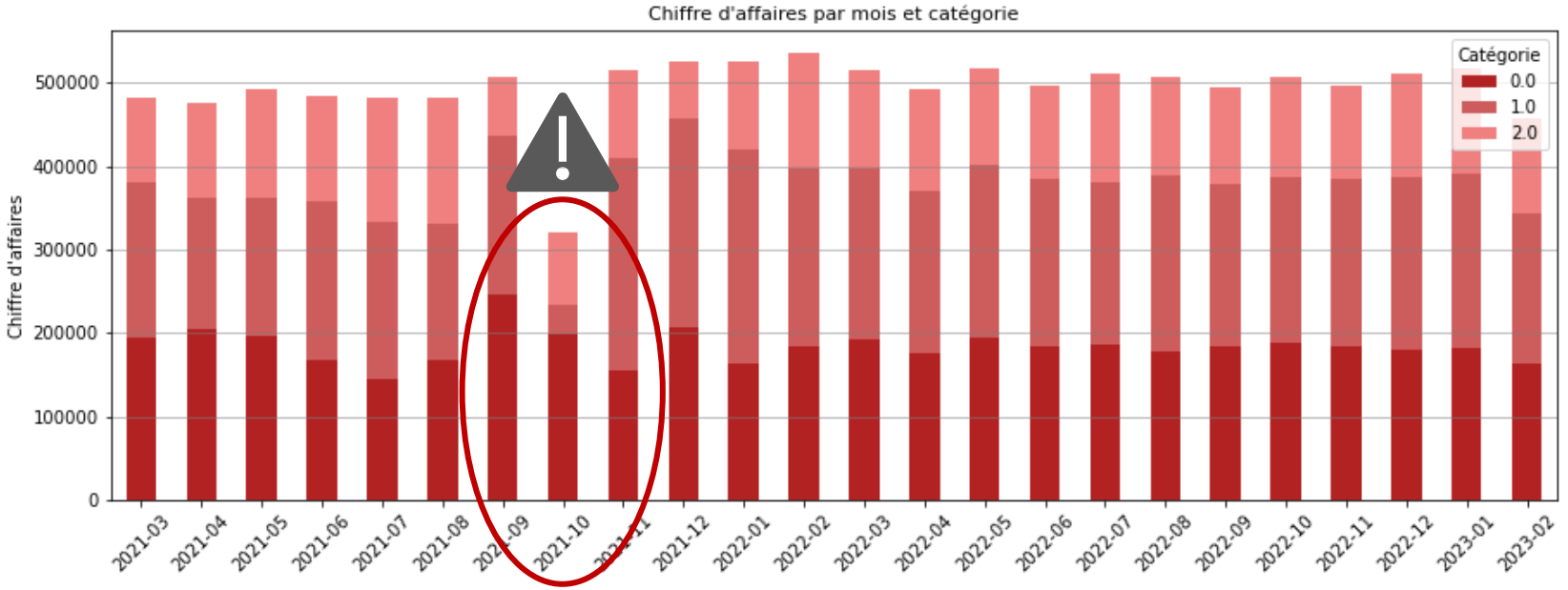
	id_prod	date	session_id	client_id
2633	0_2245	2022-09-23 07:22:38.636773	s_272266	c_4746
10106	0_2245	2022-07-23 09:24:14.133889	s_242482	c_6713
11727	0_2245	2022-12-03 03:26:35.696673	s_306338	c_5108
15675	0_2245	2021-08-16 11:33:25.481411	s_76493	c_1391
16377	0_2245	2022-07-16 05:53:01.627491	s_239078	c_7954
...	...	...	...	...
669730	0_2245	2021-08-25 09:06:03.504061	s_80395	c_131
670682	0_2245	2022-03-06 19:59:19.462288	s_175311	c_4167
671286	0_2245	2022-05-16 11:35:20.319501	s_209381	c_4453
675679	0_2245	2022-02-11 09:05:43.952857	s_163405	c_1098
677996	0_2245	2021-12-14 22:34:54.589921	s_134446	c_4854

221 rows × 4 columns

Produit 0_2245 :
- Appartient à la catégorie 0
- Imputation de la médiane * des prix de la catégorie 0

** : La médiane est plus robuste aux valeurs extrêmes.*

Octobre 2021 :
Bug informatique ou rupture de stock ?



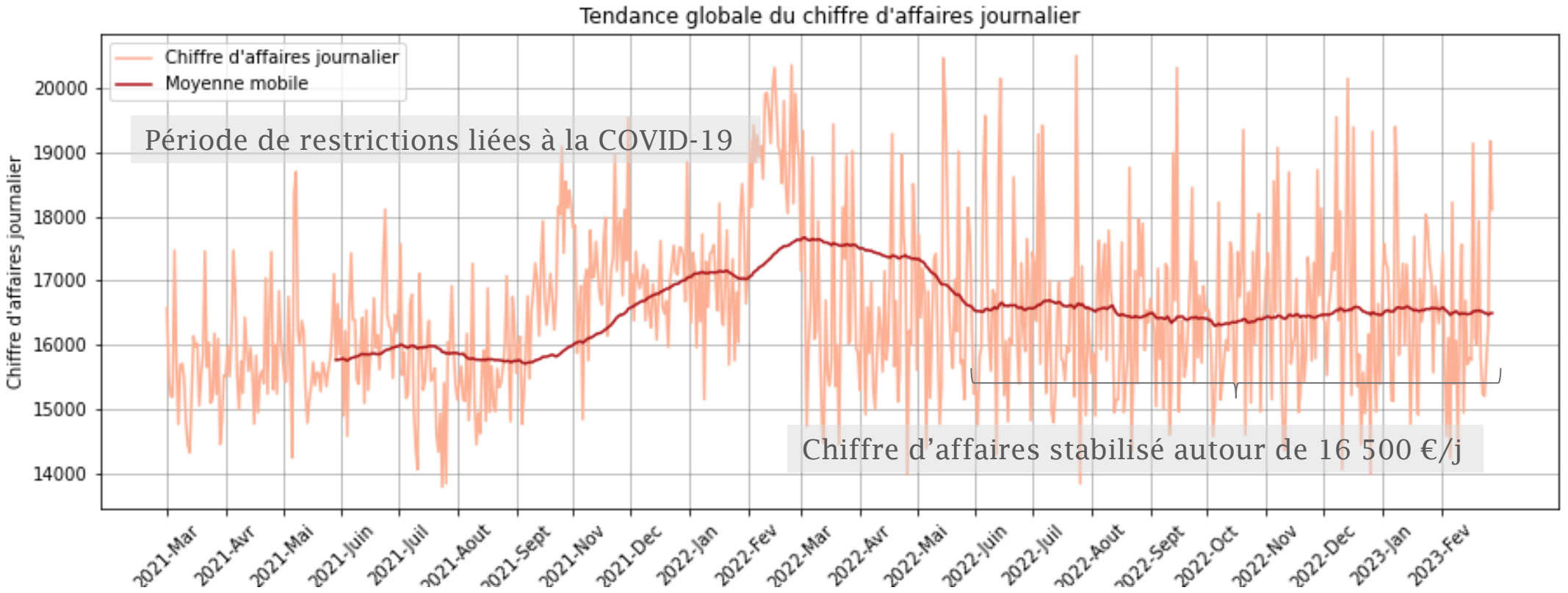
Données d'octobre 2021 non retenues pour l'analyse

Lapage

Bilan de l’activité

Chiffre d’affaires, analyse des ventes et profils des clients

Chiffre d’affaires stabilisé

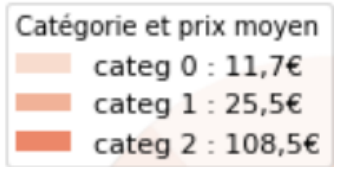


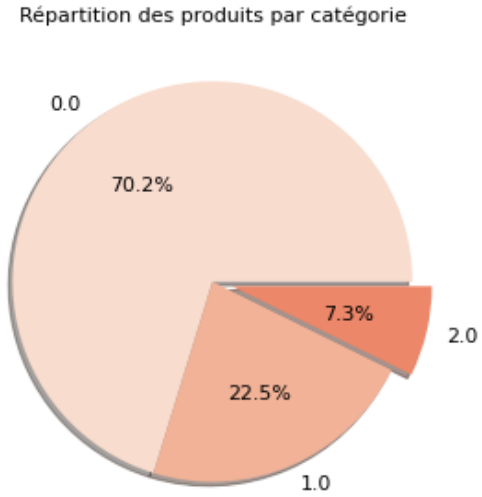
Période de restrictions liées à la COVID-19

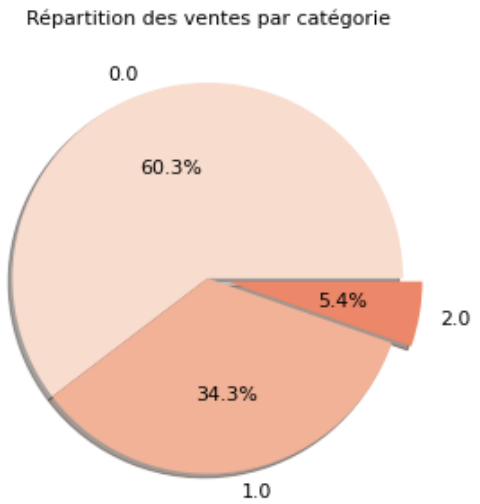
Chiffre d’affaires stabilisé autour de 16 500 €/j

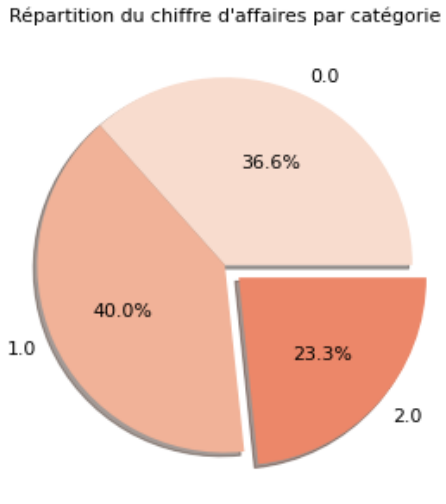
Pas de périodicité observée, en attente des données de 2023-2024

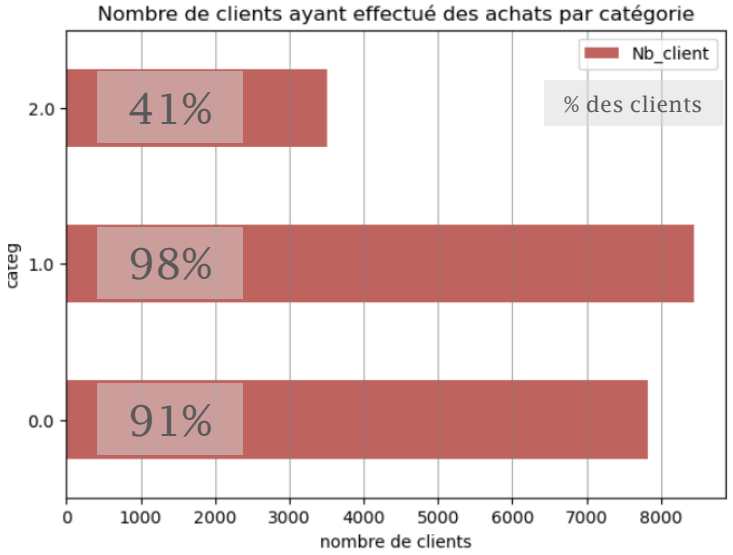
La catégorie 1 : 40% du CA, achetée par 98% des clients







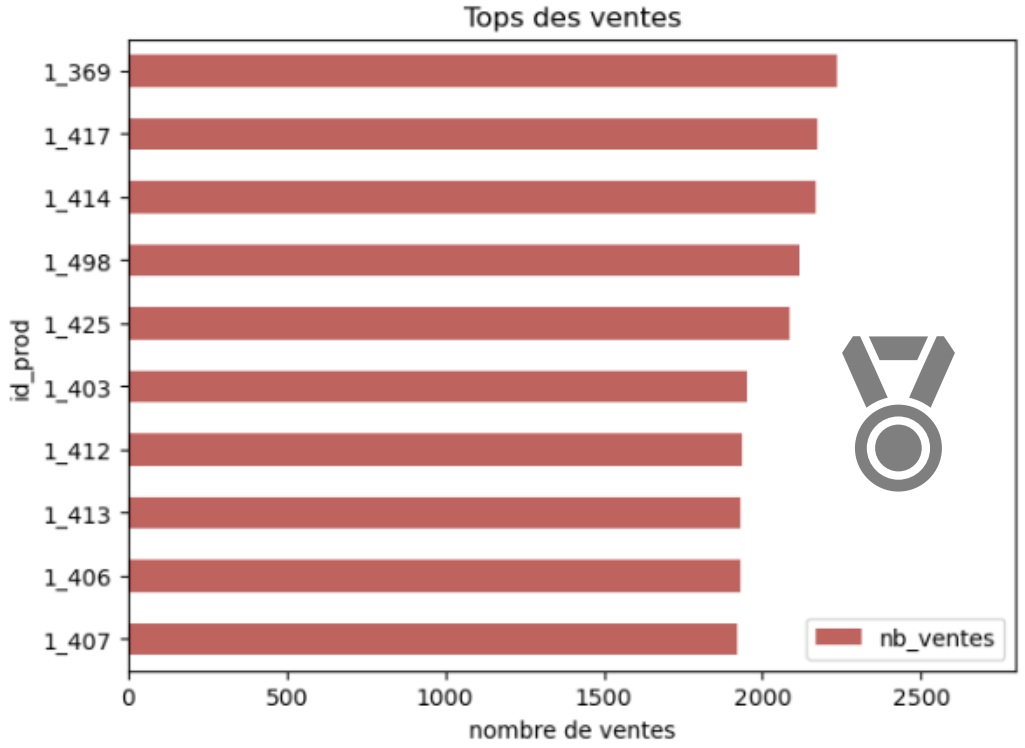




Catégorie 2 : peu de produits, peu de clients mais presque ¼ du chiffre d'affaires

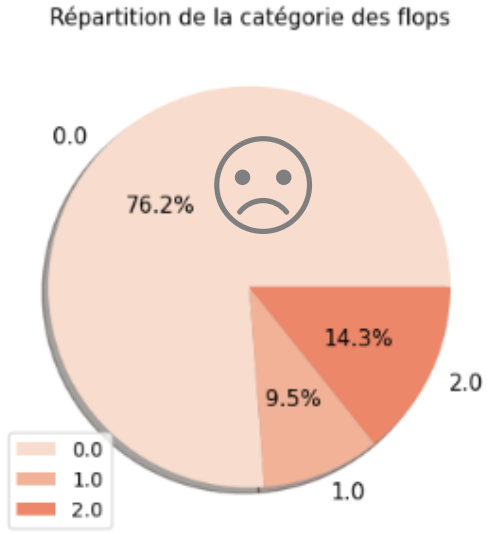
Tops et Flops des ventes

Tops des ventes : Catégorie 1
1900 ventes et +



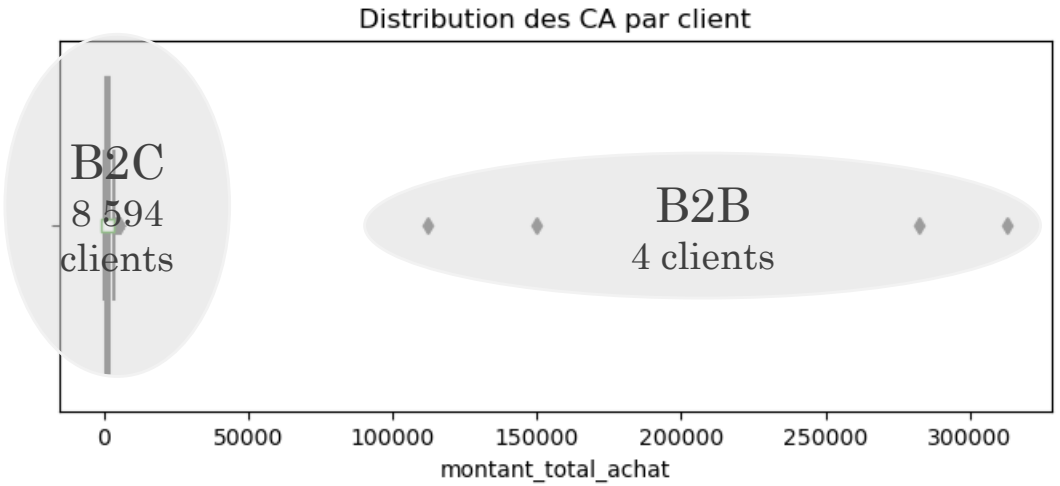
Flops des ventes : des livres de la catégorie 0 en majorité

id_prod	categ	price	nb_ventes
0_1620	0.0	0.80	0
0_1014	0.0	1.15	0
0_1780	0.0	1.67	0
0_310	0.0	1.94	0
0_1119	0.0	2.99	0
0_1645	0.0	2.99	0
0_322	0.0	2.99	0
0_1062	0.0	20.08	0
0_2308	0.0	20.28	0
0_1318	0.0	20.92	0
0_1800	0.0	22.05	0
0_299	0.0	22.99	0
0_510	0.0	23.66	0
0_1624	0.0	24.50	0
0_1025	0.0	24.99	0
1_0	1.0	31.82	0
0_1016	0.0	35.06	0
1_394	1.0	39.73	0
2_86	2.0	132.36	0
2_72	2.0	141.32	0
2_87	2.0	220.99	0

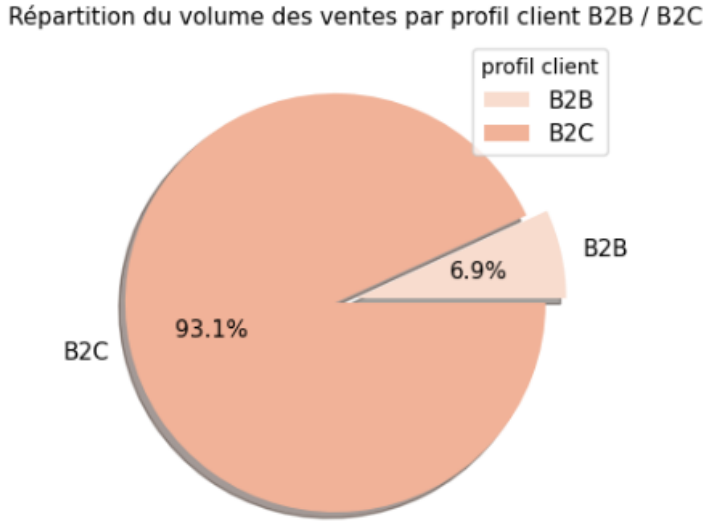


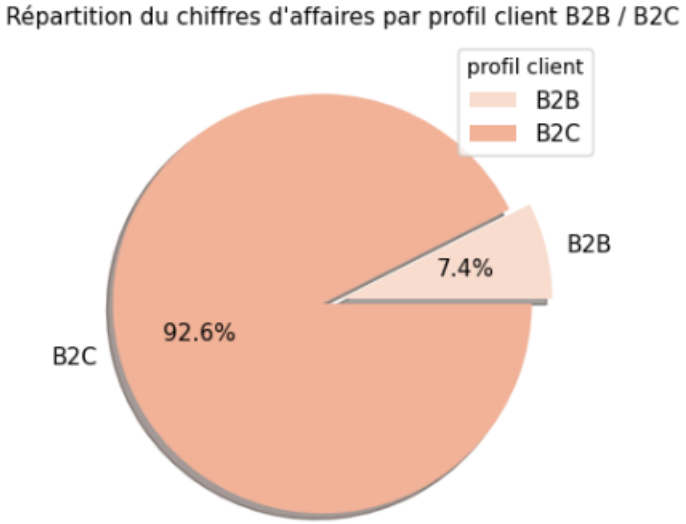
2 types de clients B2B et B2C

B2C = Business to Customer , particuliers
B2B = Business to Business, professionnels

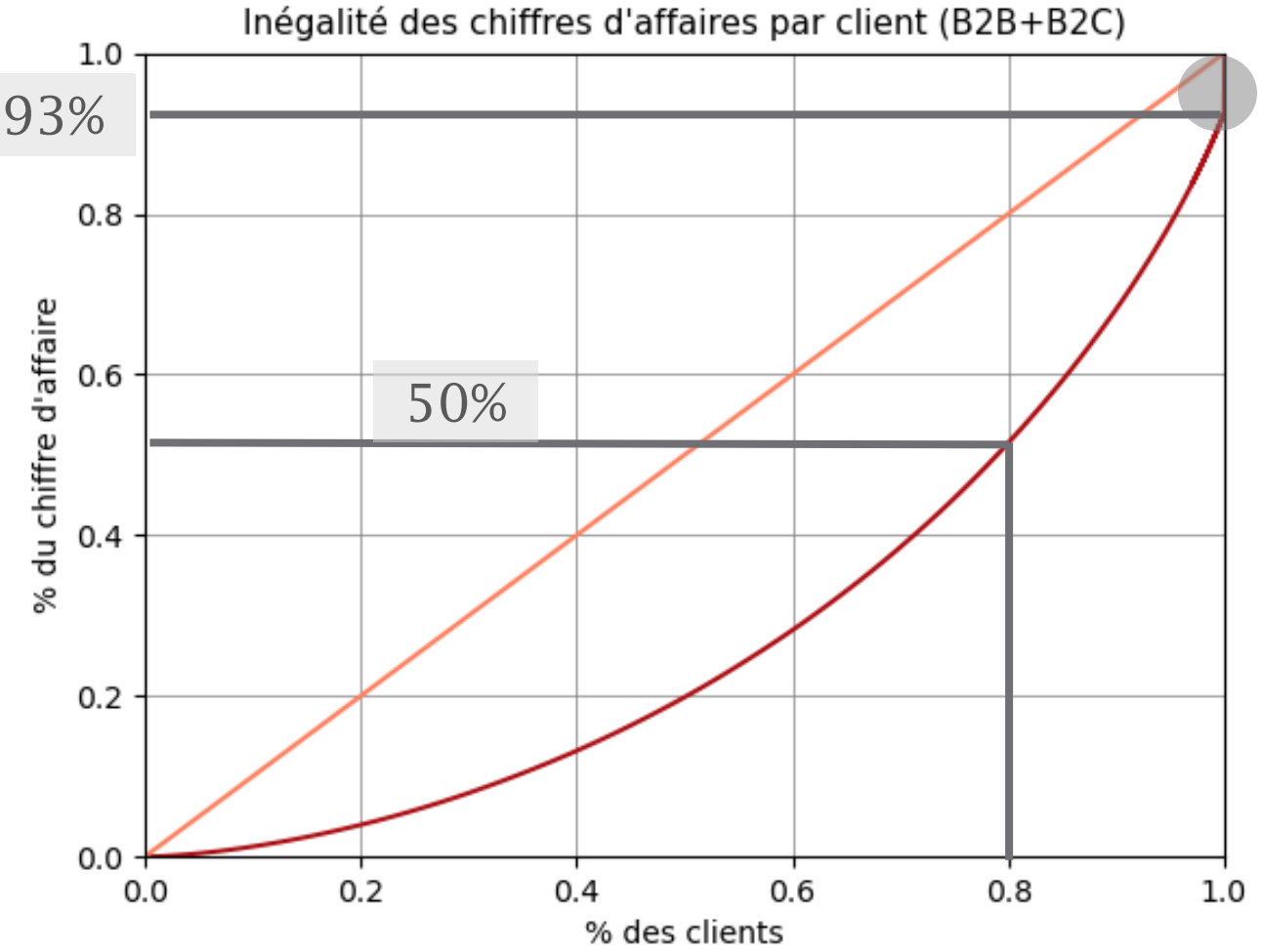


	client_id	sex	age	montant_total_achat
678	c_1609	m	43.0	312755.08
4397	c_4958	m	24.0	282654.61
6349	c_6714	f	55.0	149845.67
2727	c_3454	m	54.0	111797.67
2110	c_2899	f	29.0	5214.05
...	...	...	...	...





50% du chiffre d'affaires réalisé par 20% des clients



Clients B2B :
4 clients, 7% du CA

Comportements clients

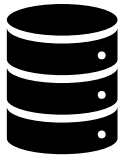
Recherche des liens entre :

- Genre et catégorie de livres
- Age et montant total des achats
- Age et fréquence des achats
- Age et taille du panier moyen
- Age et catégorie de livres achetés

Hypothèses de tests

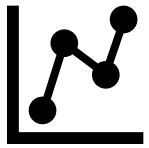
Données utilisées

Clients retenus pour l'analyse statistique :
- Particuliers, B2C



Taille de l'échantillon : 8 594 clients

Tests statistiques

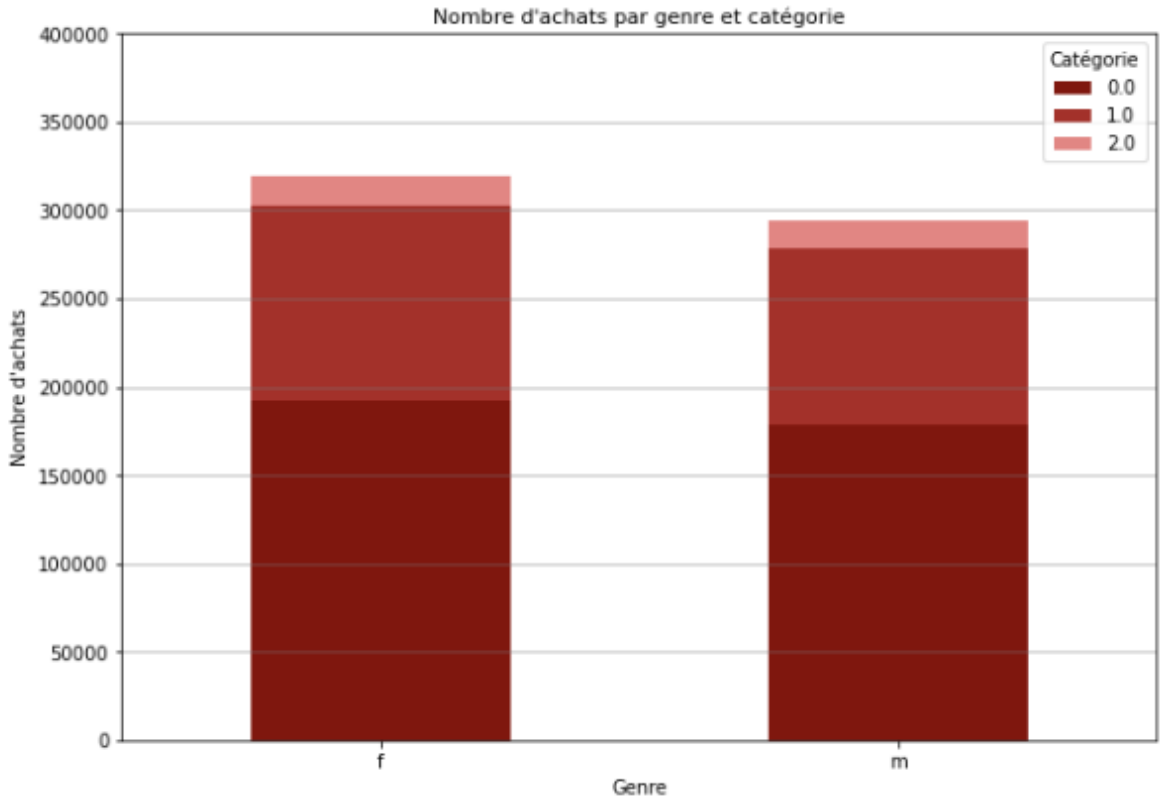


- Le risque d'erreurs de première espèce, alpha, est pris à 0.05, soit 5%
- Les échantillons sont > 30 donc selon le théorème central limite, la distribution de l'échantillon tend à suivre une loi normale.

Genre du client et catégorie de livres

Test statistique :

- Type de variables : 2 variables qualitatives
- H0 : les variables sont indépendantes
- Test applicable : Chi-2 d'indépendance
- Conditions d'application du test :
  - ✓ Variables collectées indépendamment
  - ✓ Aucune valeur attendue = 0
  - ✓ Valeurs observées et attendues > 5
- Résultats:
  - Chi² = 18,75 , degré de liberté = 2
  - p-value < 0,05 -> H0 rejetée
  - V de Cramer = 0,006 -> lien faible



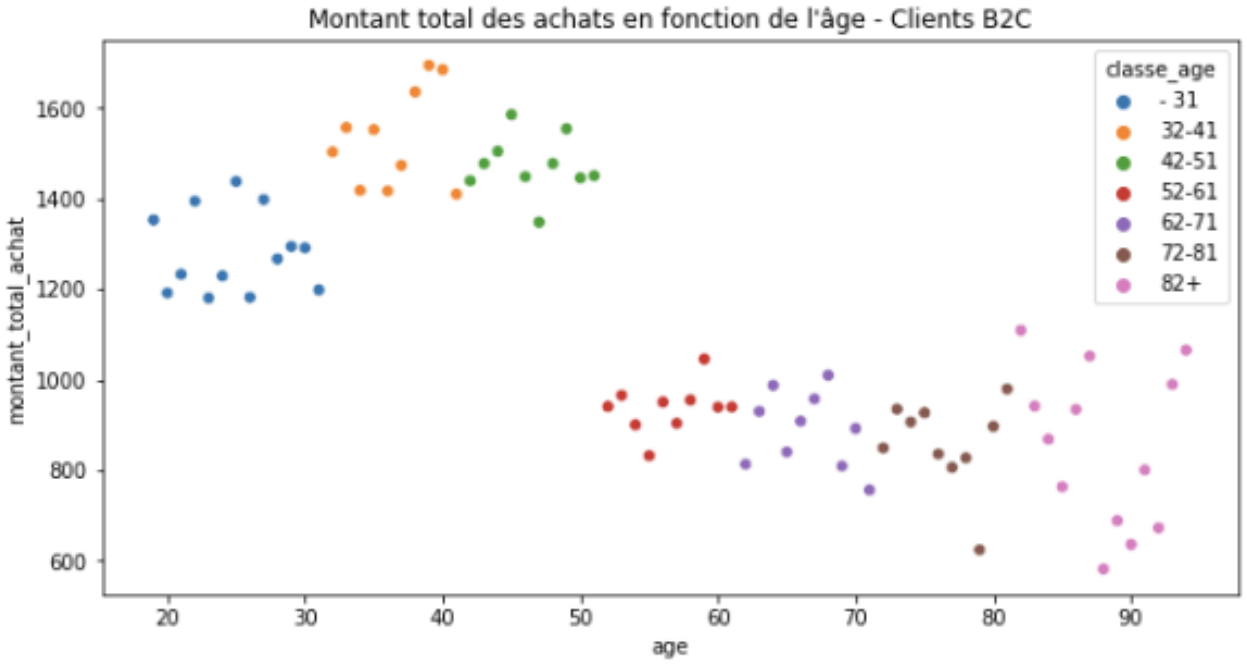
Observations :

La répartition des achats par catégorie est similaire pour les hommes et les femmes.

Interprétation du test statistique :

La corrélation entre le genre du client et la catégorie de livres n'est pas significative.

Age des clients et montant total des achats



Test statistique :

- Type de variables : 2 variables quantitatives
- H0 : Il n'y a pas de corrélation linéaire monotone entre les variables
- Test applicable : Coefficient de corrélation de Pearson (test paramétrique)
- Conditions d'application du test :
  - ✓ Distribution suivant une loi normale
- Résultats:
  - r = - 0,18

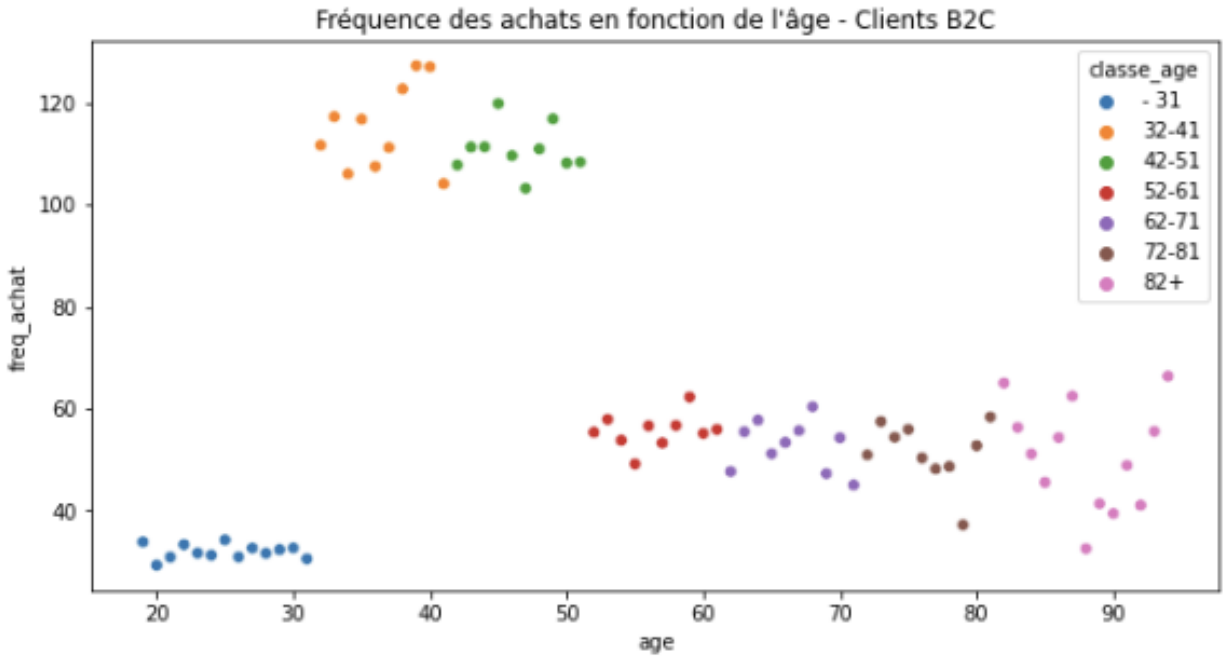
Observations :

Pas de relation linéaire forte entre les variables

Résultat du test statistique :

Pas de corrélation linéaire significative entre les variables

Age des clients et fréquence des achats



Test statistique :

- Type de variables : 2 variables quantitatives
- H0 : Il n'y a pas de corrélation linéaire monotone entre les variables
- Test applicable : Coefficient de corrélation de Pearson (test paramétrique)
- Conditions d'application du test :
  - ✓ Distribution suivant une loi normale
- Résultats:
  - r = 0,03

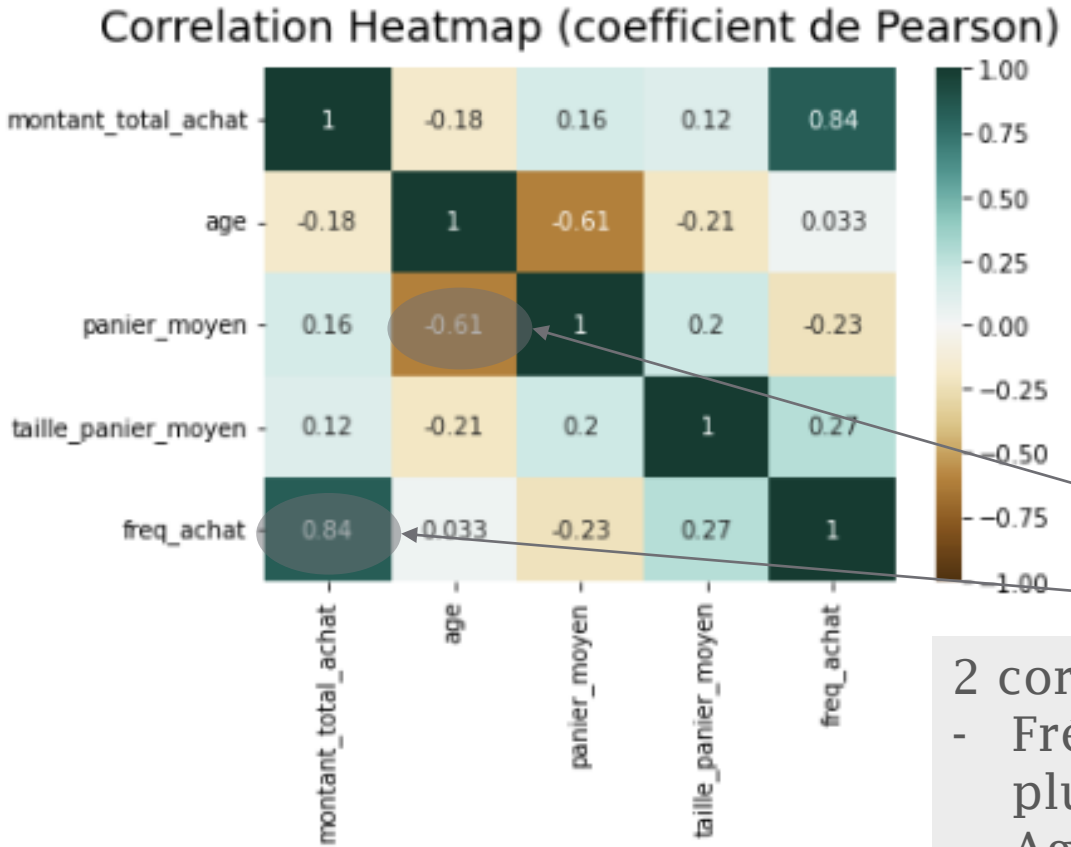
Observations :

Pas de relation linéaire forte entre les variables

Résultat du test statistique :

Pas de corrélation linéaire entre les variables

Vue globale des corrélations linéaires entre les variables quantitatives

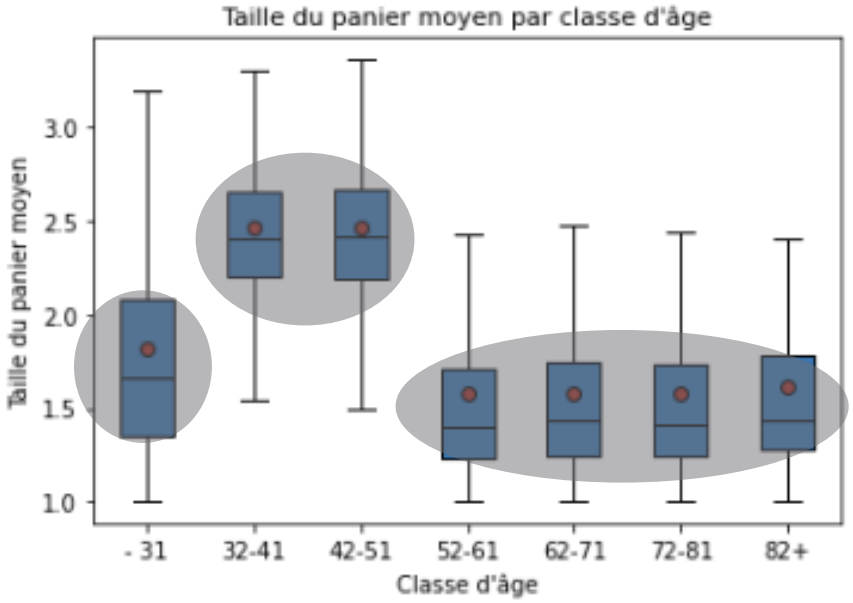


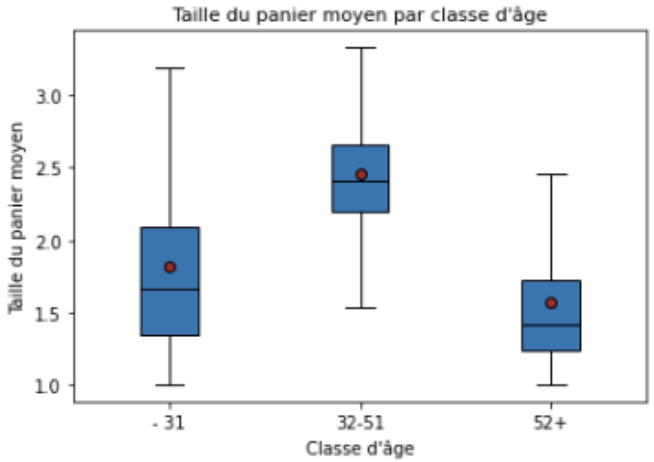
2 corrélations fortes (r > 0,5) :

- Fréquence des achats et montant total des achats : plus on achète, plus le montant est élevé (logique).
- Age et panier moyen : plus l'âge augmente, plus le montant du panier moyen est faible.

Voir analyses en annexe

Age des clients et taille du panier moyen





Test statistique :

- Type de variables : 1 variable qualitative et 1 variable quantitative
- H0 : Toutes les moyennes sont égales.
- Test applicable : Test Welch-ANOVA + post-hoc Games-Howell
- Conditions d'application du test :
  - ✓ Distribution suivant une loi normale
  - ✓ Les variances ne sont pas égales
- Résultats:
  - P-value = 0

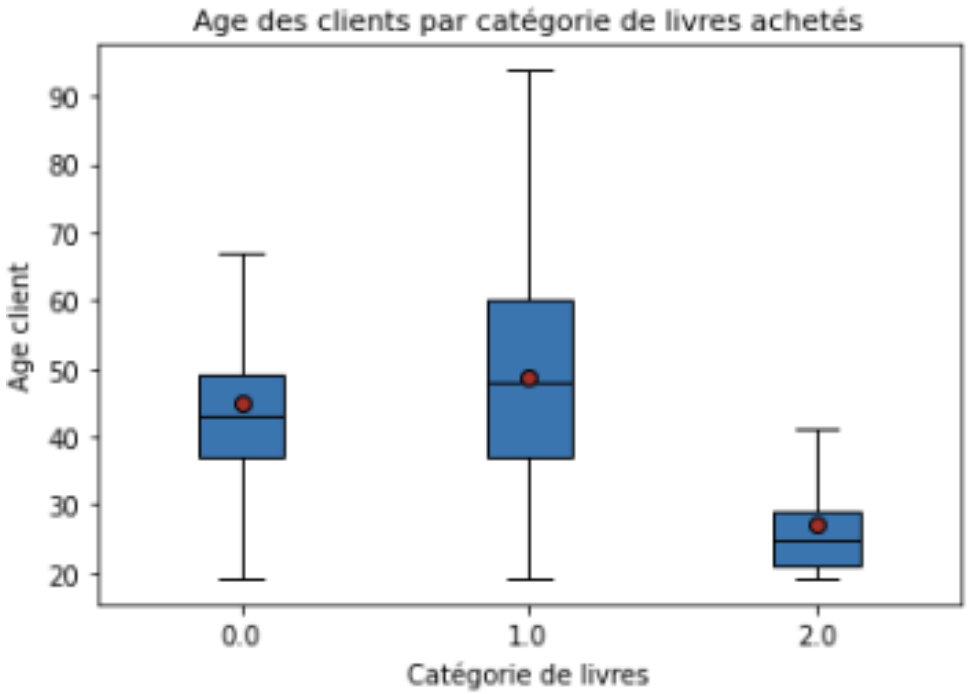
Observations :

3 groupes distincts: -31 ans, 32-51 ans, 52+ ans

Résultat du test statistique :

Les moyennes des 3 groupes ne sont pas égales.
L'âge influe sur la taille du panier.

Age des clients et catégorie de livres achetés



Test statistique :

- Type de variables : 1 variable qualitative et 1 variable quantitative
- H0 : Toutes les moyennes sont égales.
- Test applicable : Test Welch-ANOVA + post-hoc Games-Howell
- Conditions d'application du test :
  - ✓ Distribution suivant une loi normale
  - ✓ Les variances ne sont pas égales
- Résultats:
  - P-value = 0

Observations :

- Catégorie 0 achetée par les clients < 70 ans
- Catégorie 1 achetée par tous les clients
- Catégorie 2 achetée par les clients < 40 ans

Résultat du test statistique :

Les moyennes des 3 groupes ne sont pas égales.
L'âge influe sur la catégorie de livres achetés.

Conclusion des recherches de corrélations

Variable 1	Variable 2	Conclusion
Genre	Catégorie de livres	Pas de corrélation significative
Age	Montant total des achats	Pas de corrélation linéaire significative
Age	Fréquence des achats	Pas de corrélation linéaire significative
Age	Taille du panier moyen	Age influe sur la taille du panier
Age	Catégorie de livres	Age influe sur la catégorie de livres achetés

3 groupes d'âges aux comportements différents pour montant total des achats, fréquence des achats, taille du panier moyen :

- < 32 ans
- 32 – 51 ans
- > 52 ans

Merci de votre attention. Des questions ?

Pour aller plus loin, consultez les annexes.

Lapage

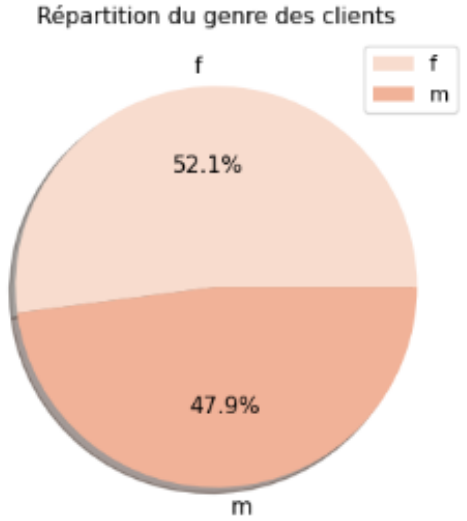
ANNEXES

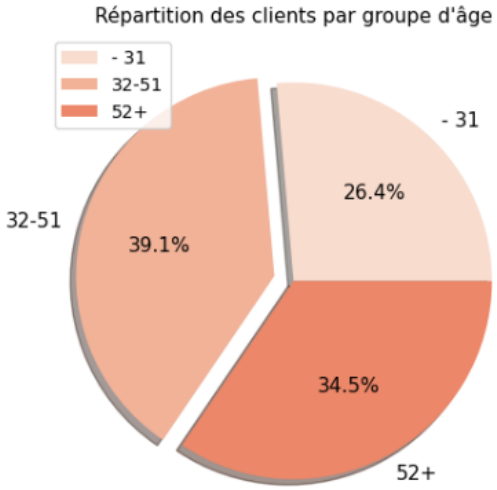
Profils des clients

- Genre et âge des clients
- Les clients Business to business
- Les clients B2C < 32 ans

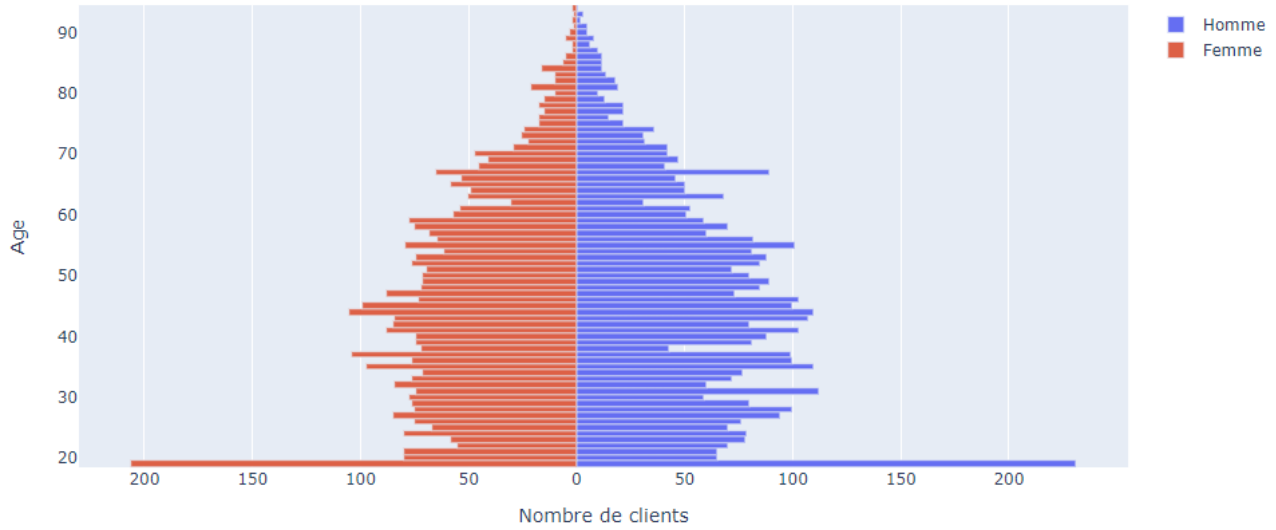
Genre et âge des clients

Parité homme / femme





Pyramide équilibrée entre homme et femme



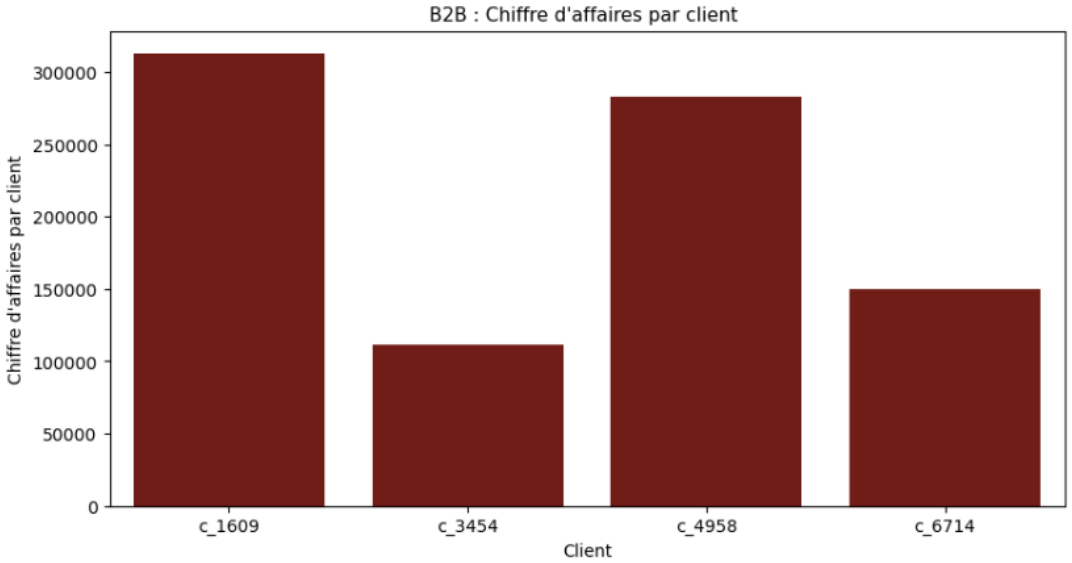
Pic à 19 ans : limite d'âge pour les comptes de la librairie en ligne ?

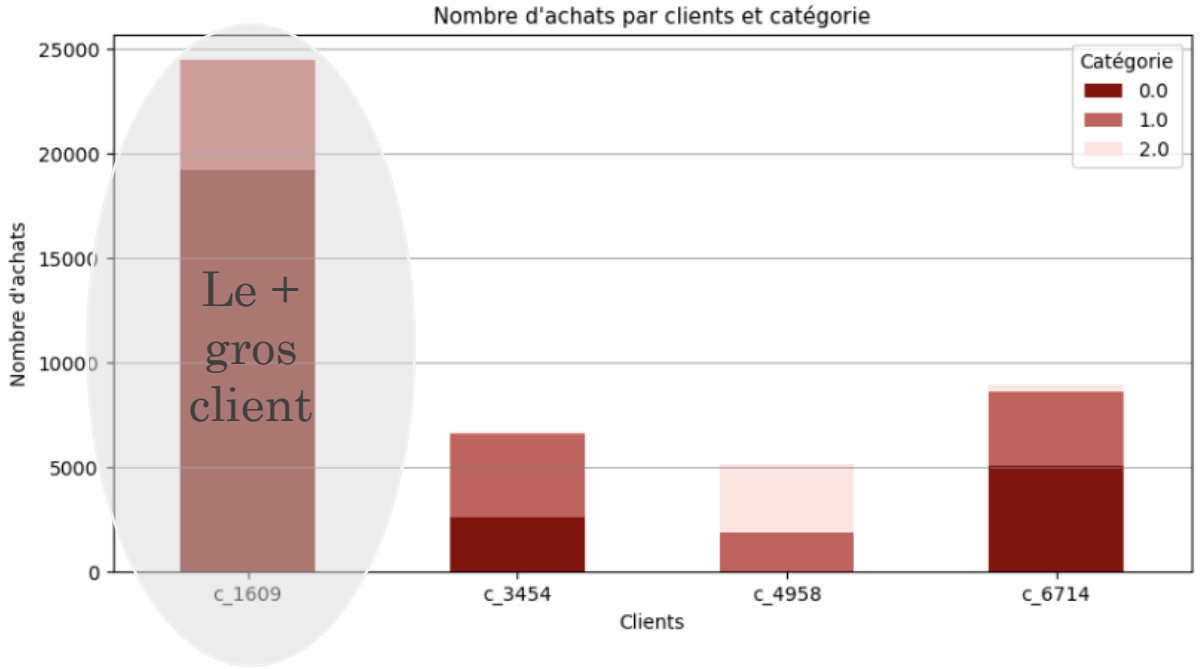
Majeur partie des clients < 50 ans

Les clients Business to Business

Les 4 professionnels achètent pour des librairies différentes : réparation par catégorie non homogène

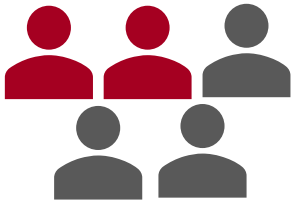
Fréquence d'achat : 7 à 35 achats / jour
Petit panier : 1 à 4 articles



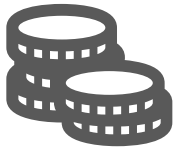


Le client c_1609 est le plus gros client (ventes et chiffre d'affaires)

Les + gros clients particuliers = les 30-50 ans



2 clients sur 5 ont entre 30 et 50 ans

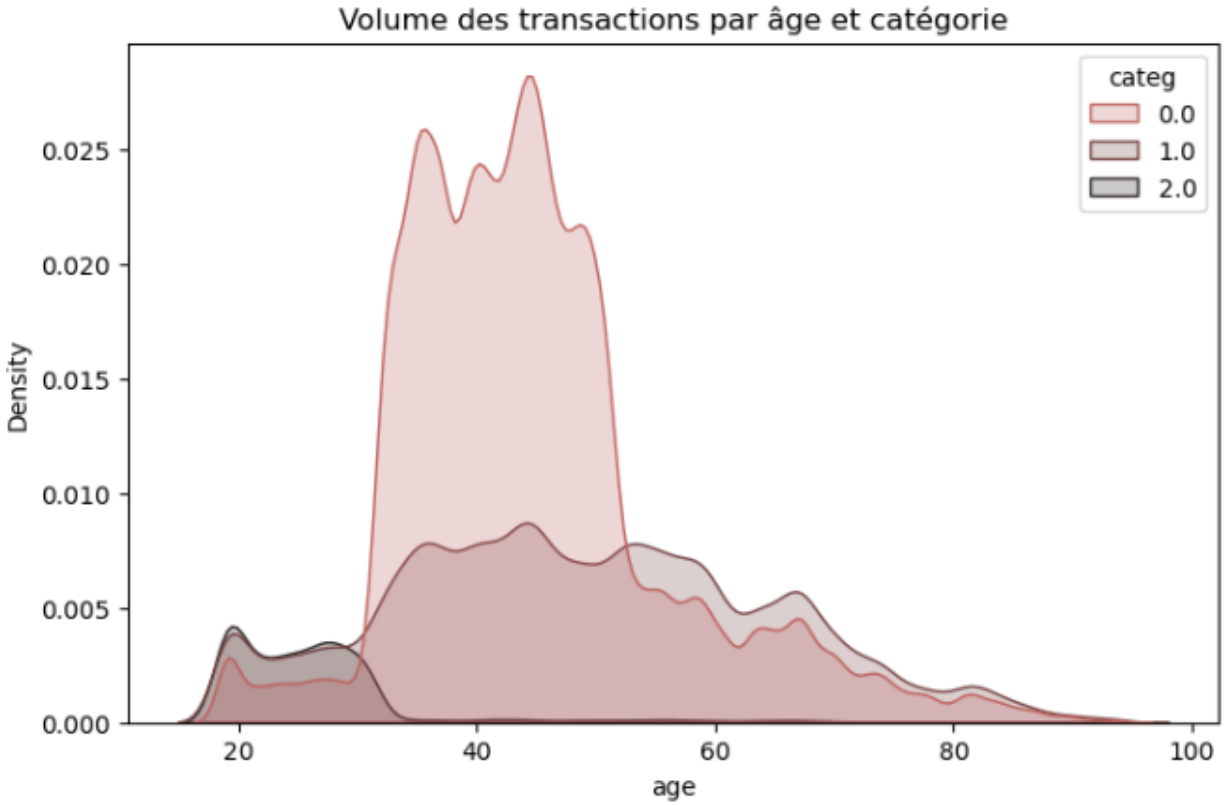


Presque la moitié du chiffre d'affaire (47%)

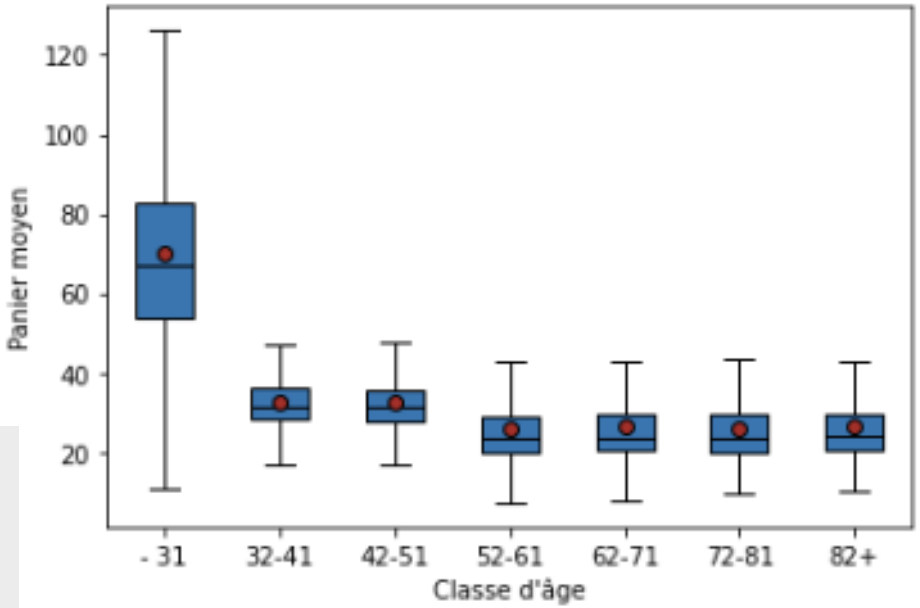
6 achats sur 10 sont réalisés par des 30-50 ans

Les plus grands consommateurs de catégorie 0

Les clients Business to Customer < 32 ans



Les moins de 32ans :
- Achètent le plus de catégorie 2
- Ont les paniers les plus élevés



0,3% des clients ont créé un compte sans effectué d'achats, plus de la moitié ont moins de 32 ans

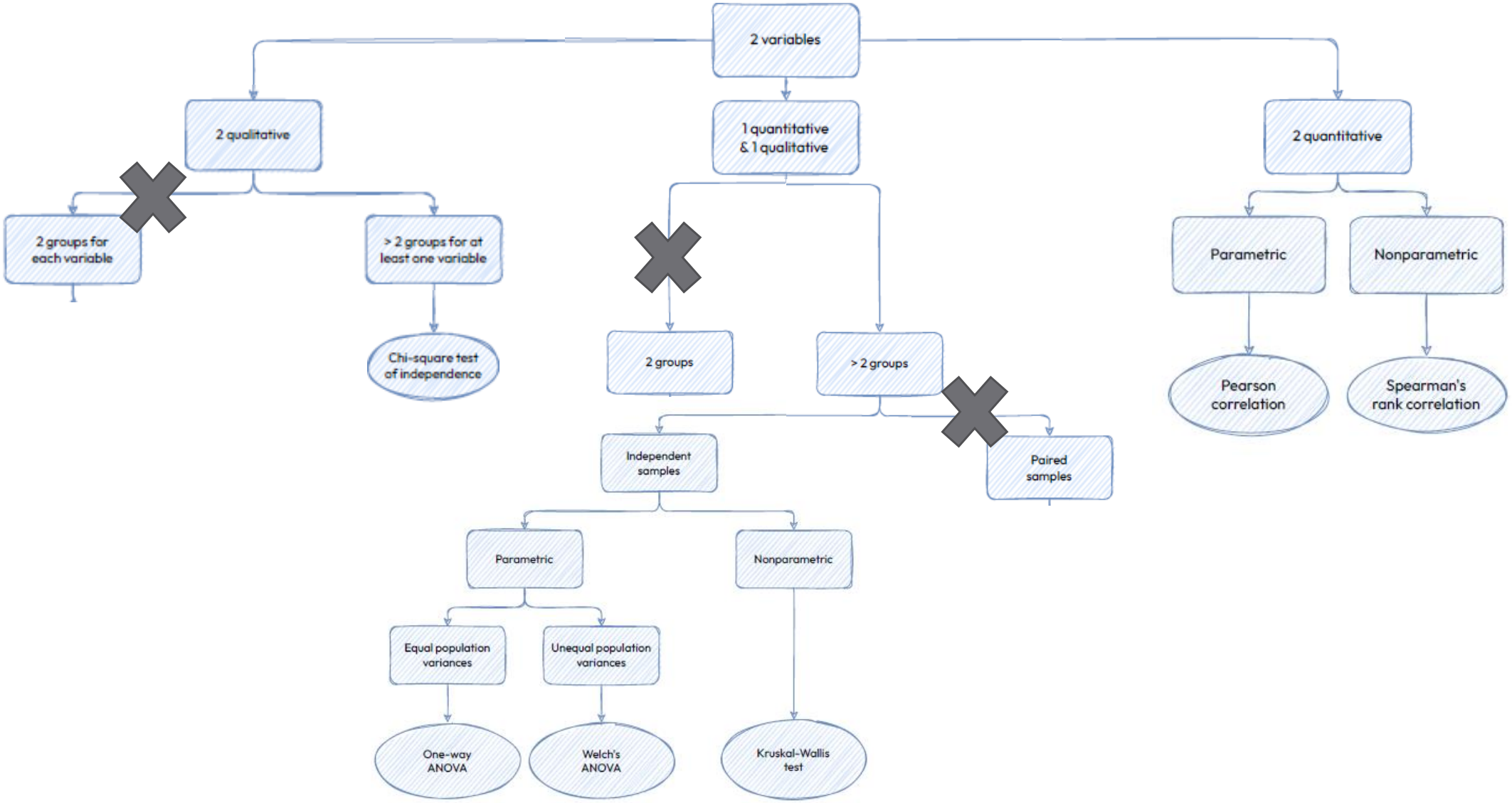
- Démarche et choix des tests statistiques
- Relation entre la fréquence d'achat et le montant total d'achats
- Relation entre l'âge et le panier moyen

Comportements clients

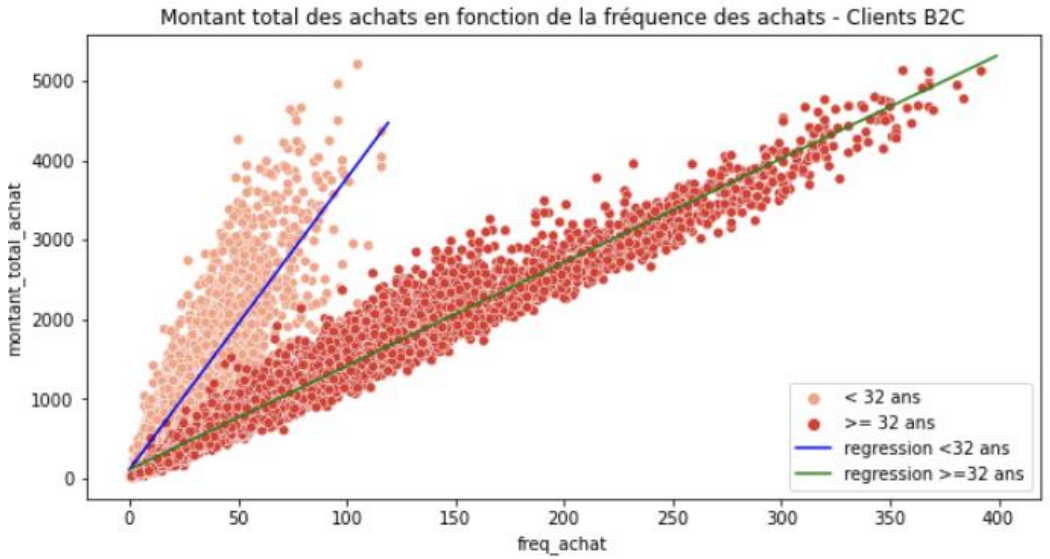
Démarche de test statistique

Etape 1	Construire les hypothèses de test : H0 et H1
Etape 2	Définir les risques d'erreur : - α : risque de première espèce - β : risque de seconde espèce
Etape 3	Choix du test statistique : vérification des conditions d'applications
Etape 4	Réaliser le test et interpréter le résultat

Choix des tests statistiques



Modélisation de la relation fréquence des achats et montant total



Observations :
2 relations linéaires positives en fonction de l'âge : les moins de 32 ans et les plus de 32 ans.

Nota : les moins de 32 ans achètent le plus de catégorie 2, la plus chère, et ont donc un montant total d'achats plus élevé.

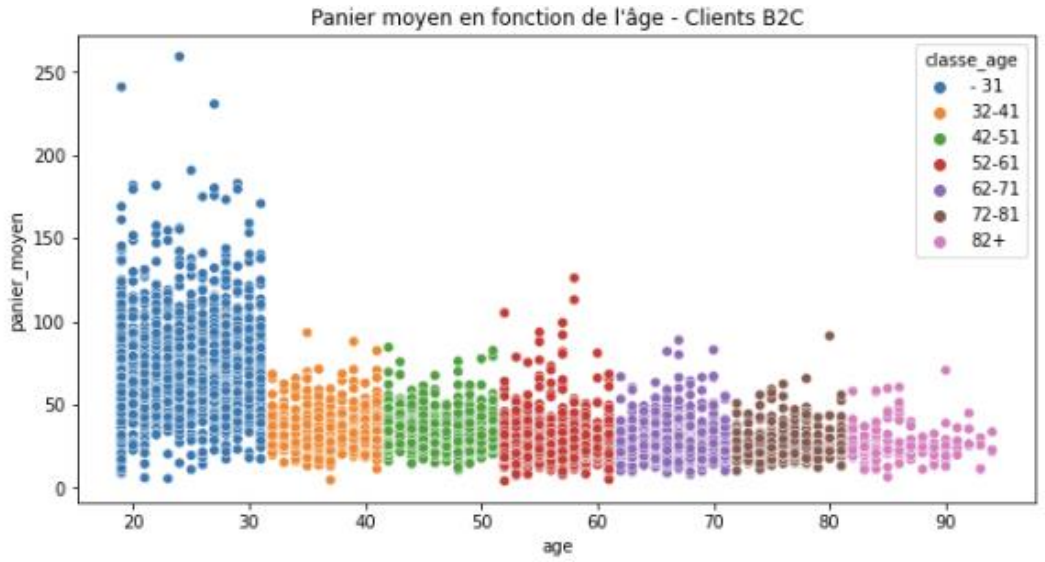
Méthode : Régression linéaire : méthode des moindres carrés ordinaires, y' = a*x +b

Echantillonnage stratifié pour composer les échantillons d'entraînement et de test

Coefficient de détermination R^2: modèle performant

- < 32 ans : R^2=0,7, 70% de la variabilité des points est expliqué par le modèle
- > 32 ans : R^2=0,96, 96% de la variabilité des points est expliqué par le modèle

Modélisation de la relation âge et panier moyen



Observations :
Relation linéaire négative entre l'âge et le panier moyen

Nota : les moins de 32 ans achètent le plus de catégorie 2, la plus chère, et ont donc un panier moyen plus élevé.

Méthode : Régression linéaire : méthode des moindres carrés ordinaires, y' = a*x +b

Echantillonnage stratifié pour composer les échantillons d'entraînement et de test

Coefficient de détermination R^2: 0,38
38% de la variabilité des points est expliqué par le modèle
-> le modèle n'est pas performant.