Analyse de ventes et profils clients

Sommaire

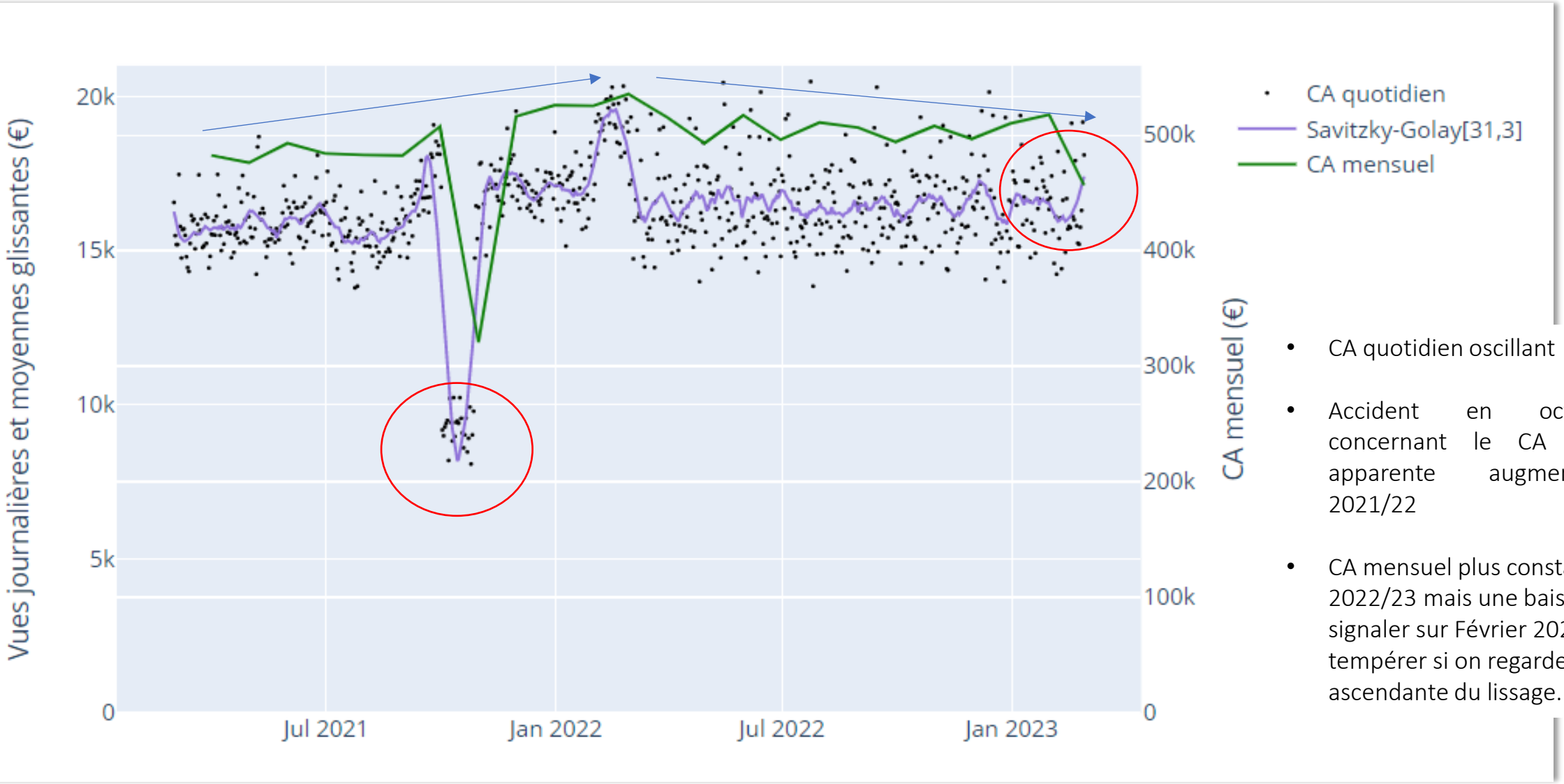
Nettoyage des données

Les différentes étapes

- Repérage des valeurs N/A
- Vérification de l'uniformité des types de données pour chaque dataset (customers, products, transactions)
- Repérage des valeurs "tests" dans 'products' (id_prod = T_0, price = -1)
- Repérage puis suppression des valeurs "tests" dans 'transactions' (T_0 test_2021-03-01 02:30:02.237419 s_0 ct_0)
- Vérification dans 'transactions' du format des dates
- Jointure à gauche sur transactions/products : une référence est manquante et est supprimée de "transactions"
- Pour la suite de l'analyse
  - une jointure interne transactions/products et transactions/customers sur un fichier 'transactions' propre permet de s'assurer que les informations rajoutées sont bien présentes.

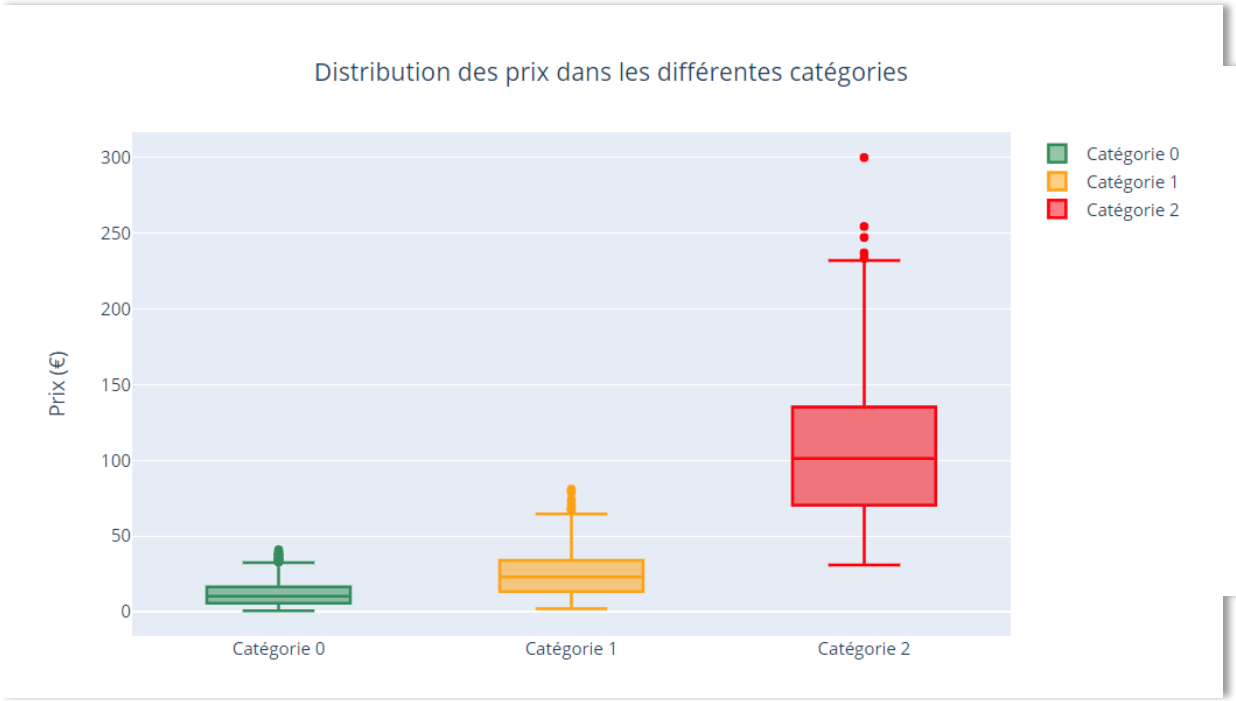
Le Chiffre d’Affaire

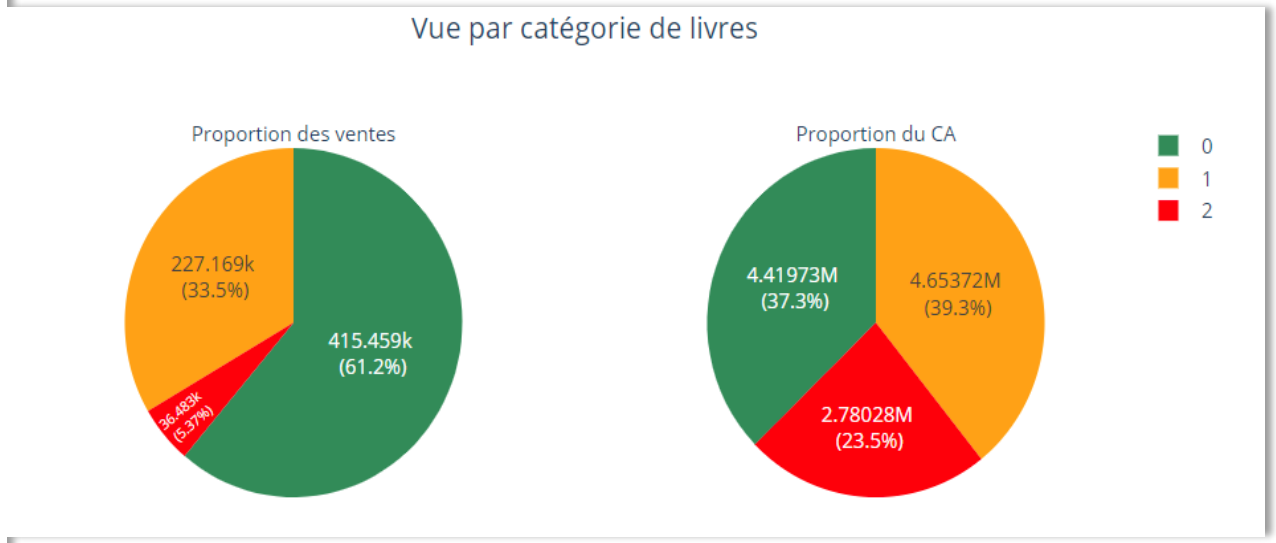
Evolution du chiffre d'affaire



- CA quotidien oscillant
- Accident en octobre 2021 concernant le CA mensuel, en apparente augmentation sur 2021/22
- CA mensuel plus constant sur 2022/23 mais une baisse est à signaler sur Février 2023 – à tempérer si on regarde la pente ascendante du lissage.

Part des livres par catégorie dans le chiffre d'affaire





Livres de catégorie 0

- Les moins chers
- 61% des ventes
- 37% du chiffre d'affaire

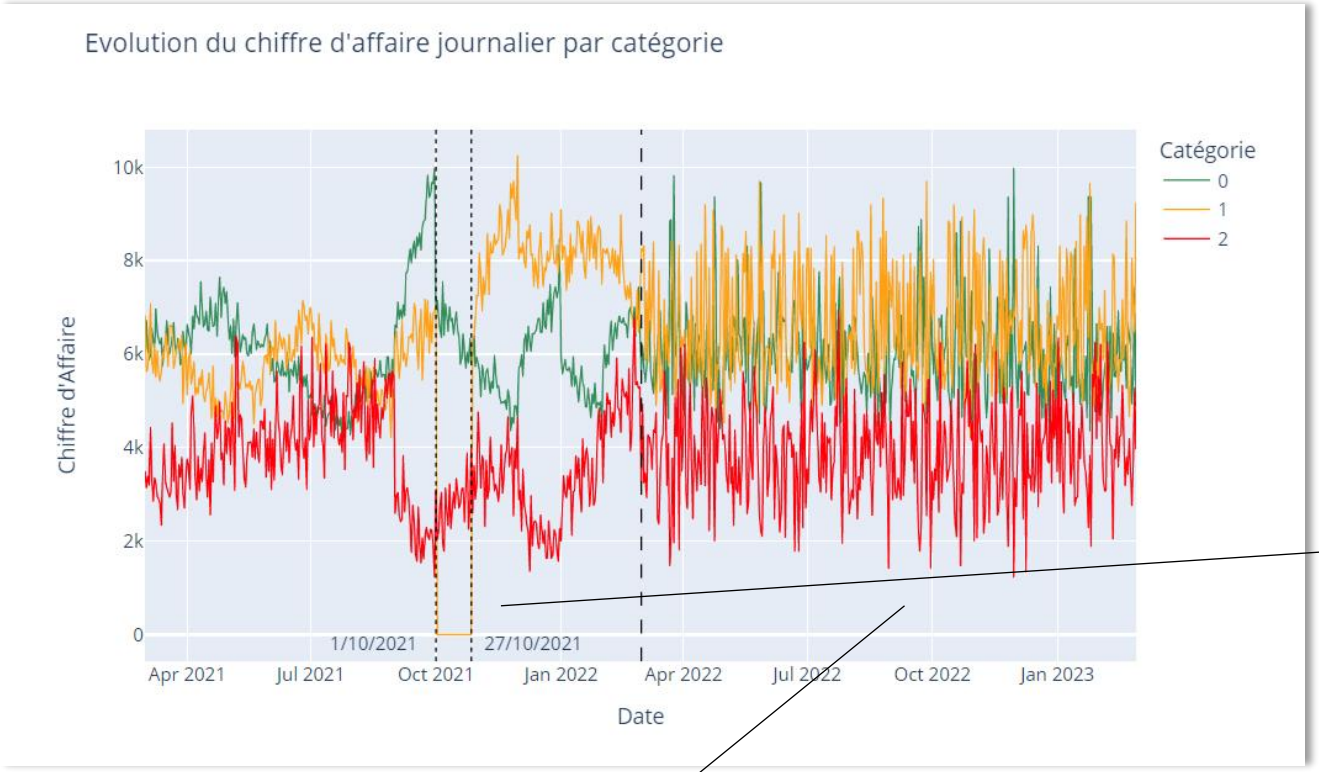
Livres de catégorie 1

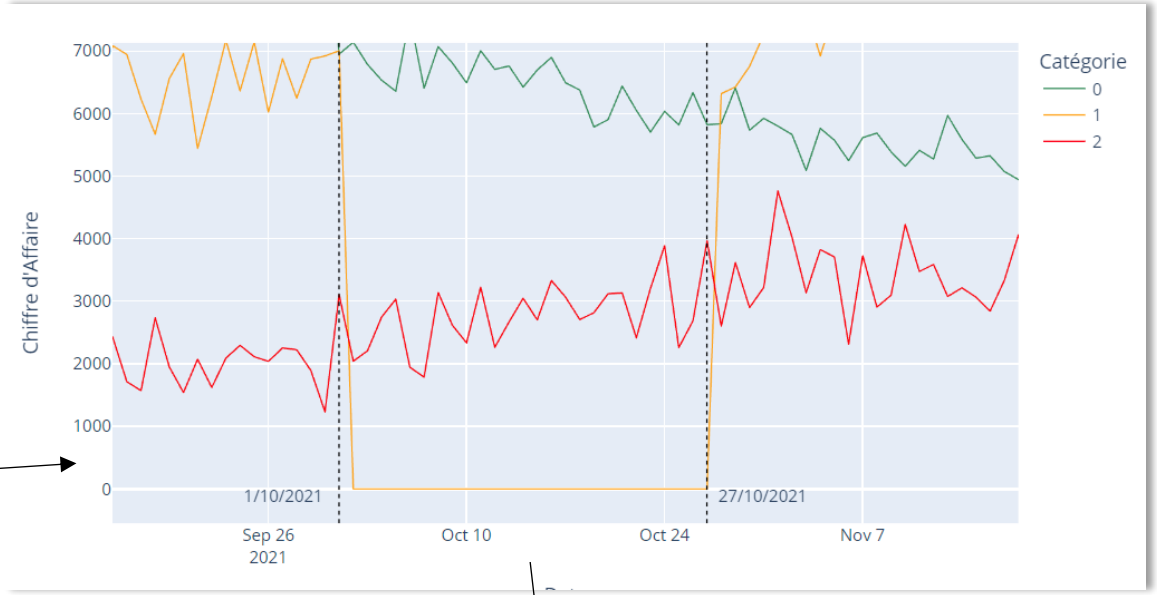
- Prix modérés
- Un tiers des ventes
- 23% du chiffre d'affaire

Livres de catégorie 2

- Les plus chers
- 5% des ventes
- 23% du chiffre d'affaire

Mois d’octobre 2021 et changement de comportement en 2022

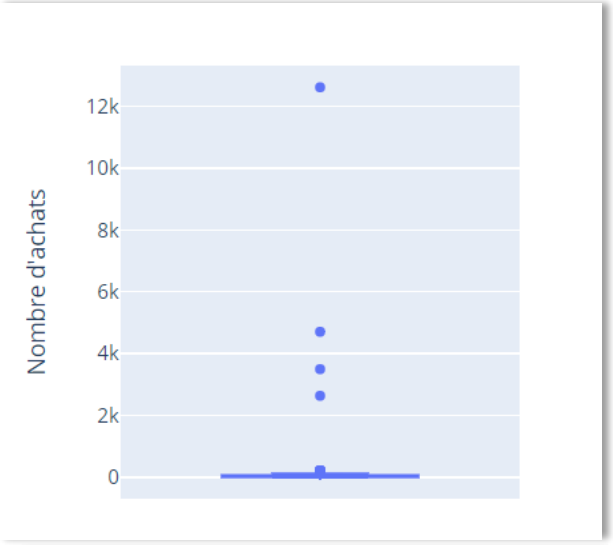


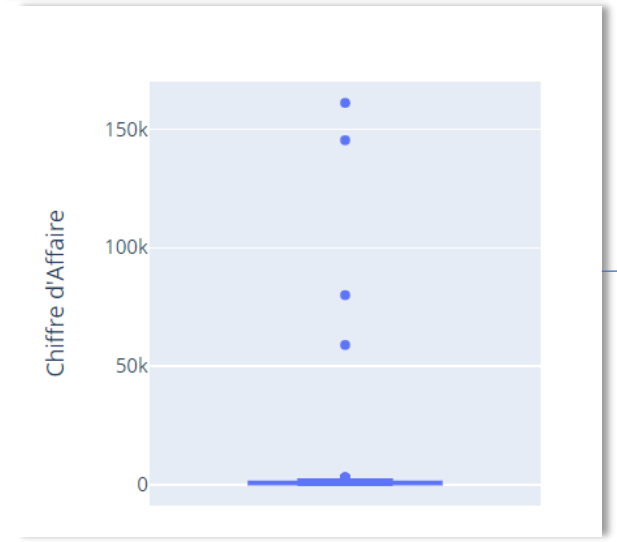


Changement de comportement à partir du 1/03/2022

Aucun achat de catégorie 0 n’est enregistré

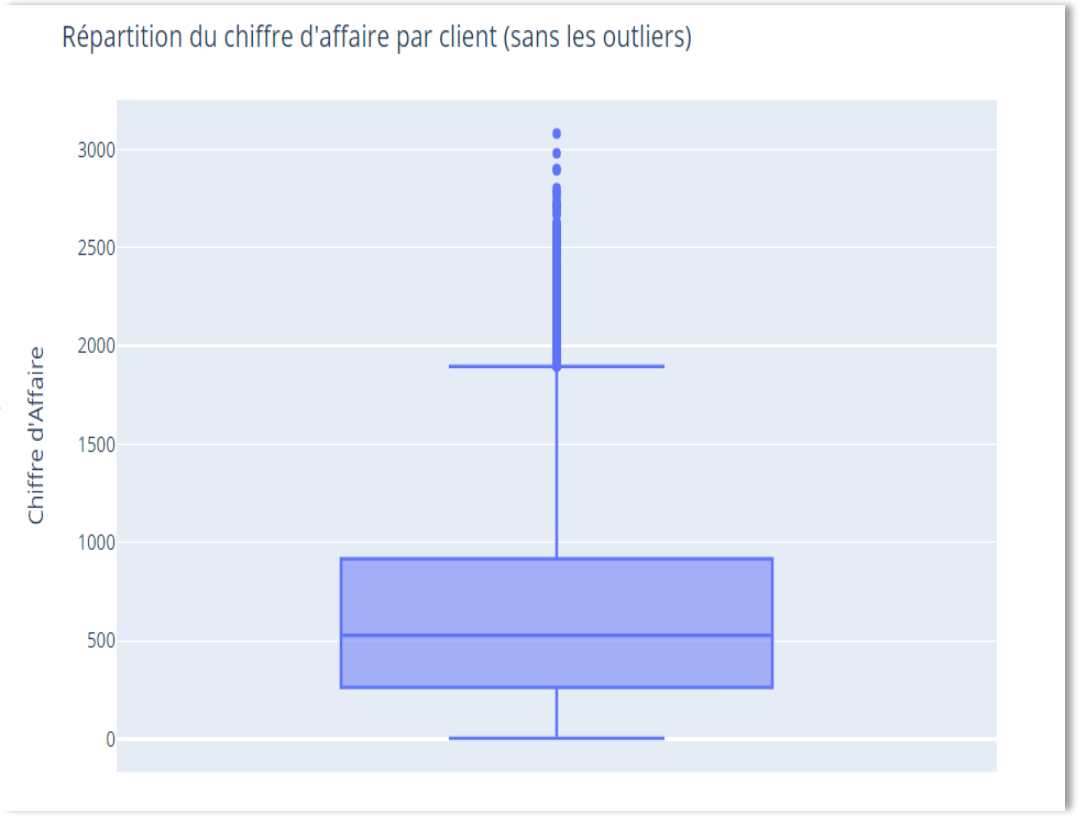
Recherche des plus gros clients (outliers)





4 clients (sans doute des organisations)
c_1609 c_4958 c_6714 c_3454
7.4 % du chiffre d'affaire
6.86 % des achats

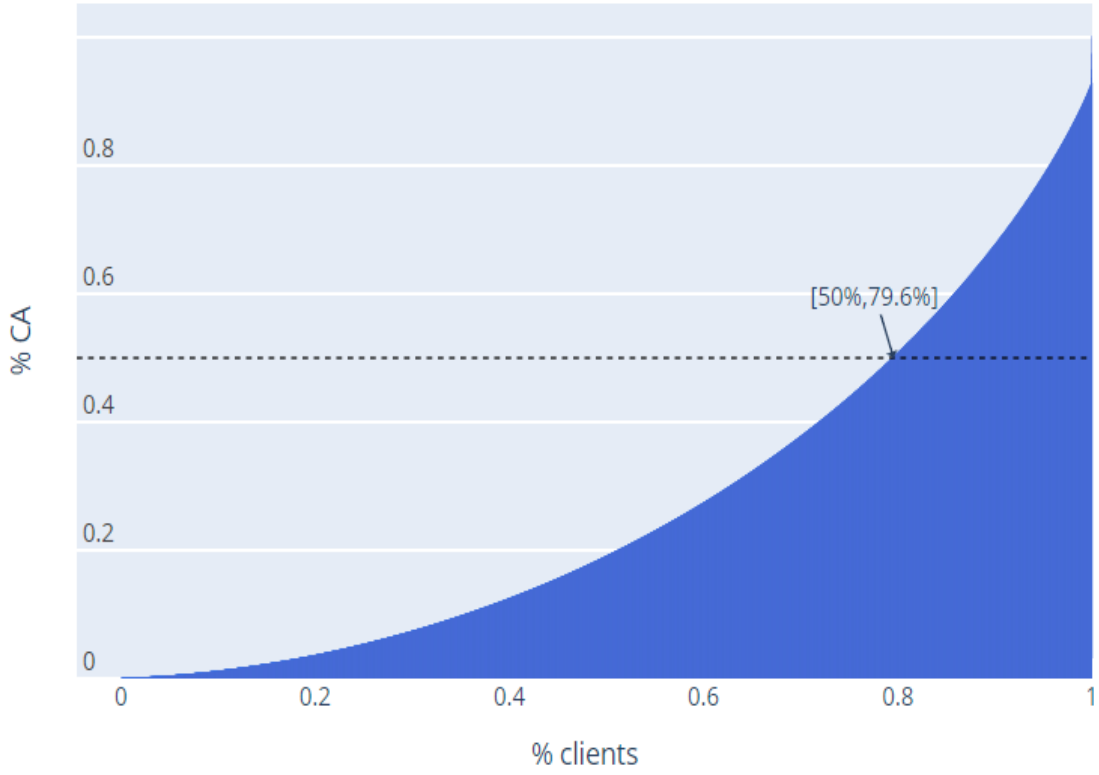
Sans les 4 plus gros clients



Répartition du Chiffre d'Affaire parmi les clients

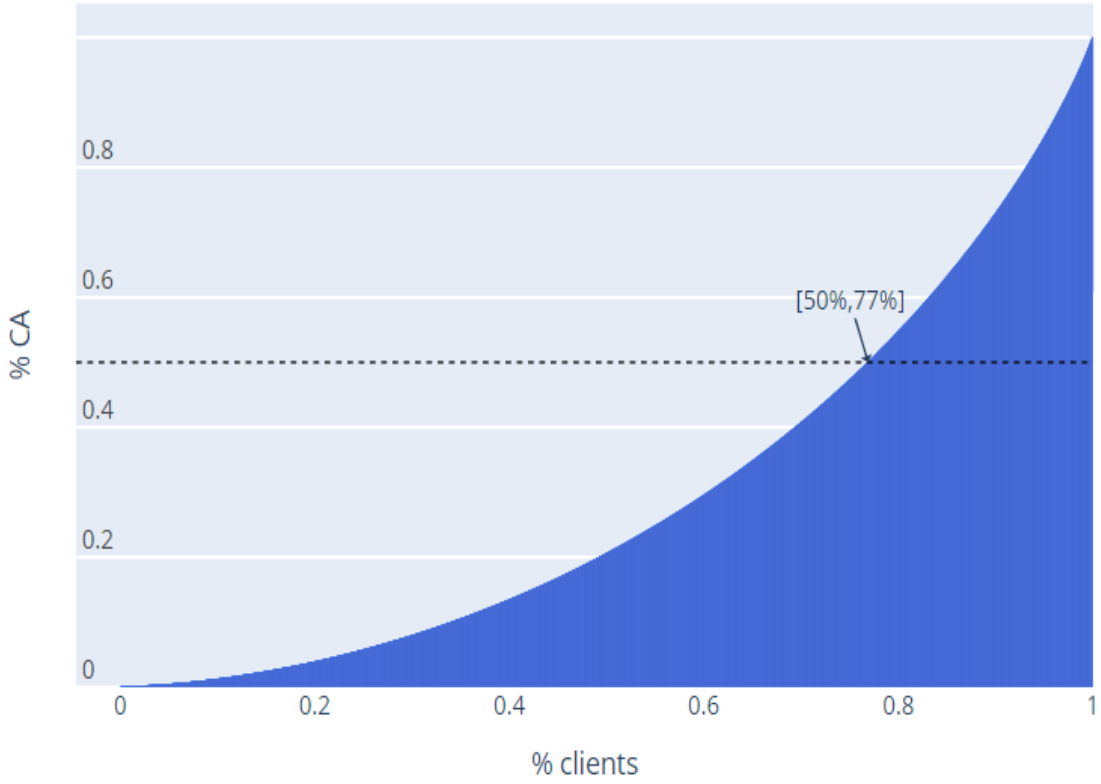
Répartition du CA (courbe de Lorenz)

Indice de gini : 0.46



Répartition du CA (courbe de Lorenz) - sans les outliers

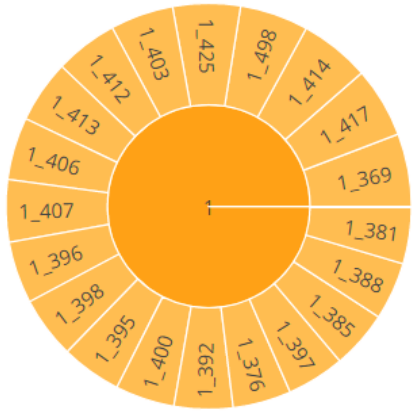
Indice de gini : 0.418



Tops et flops

Top 20 et Flop 20

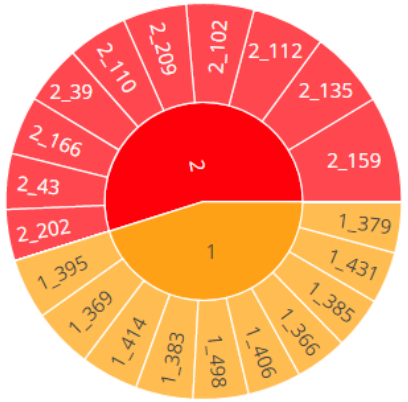
Livres les plus vendus



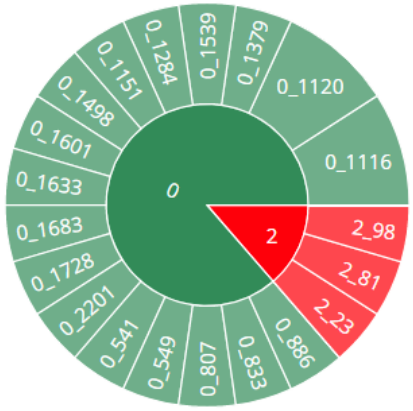
Références des ouvrages

1_369
1_414
1_498
1_406
1_395
1_385

Livres qui rapportent le plus

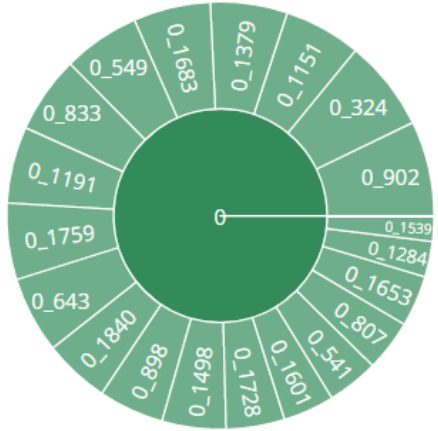


Livres les moins vendus

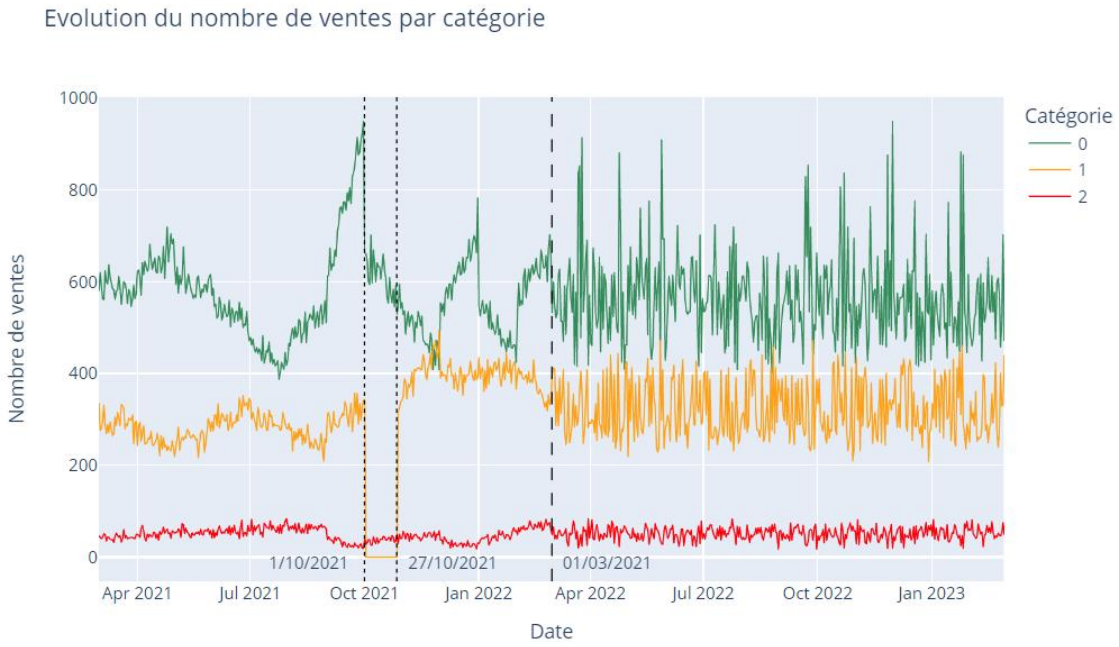


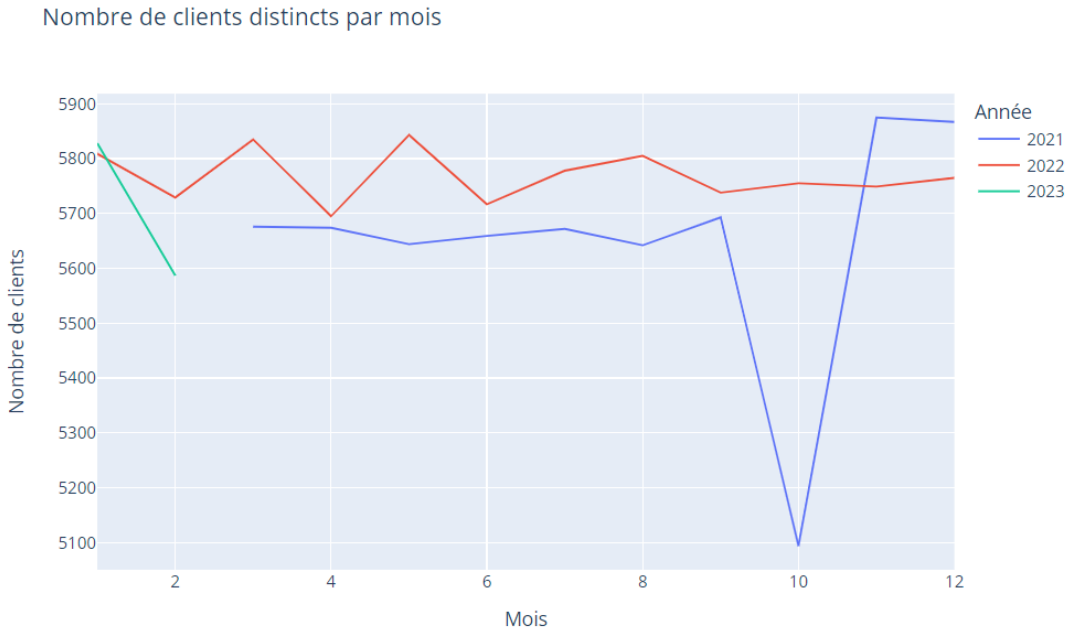
0_1151
0_1284
0_1379
0_1498
0_1539
0_1601
0_1683
0_1728
0_541
0_549
0_807
0_833

Livres qui rapportent le moins



Sélection des données étudiées par la suite





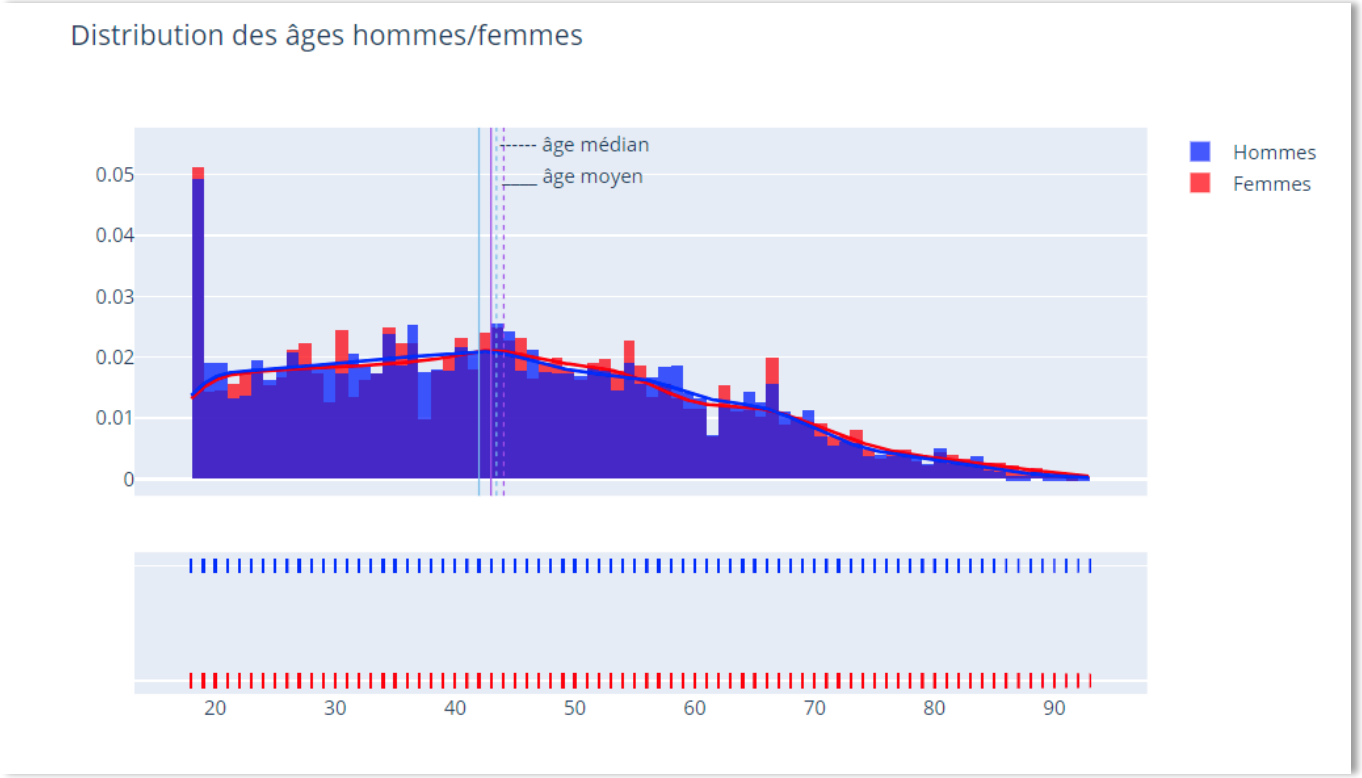
- Augmentation du nombre de clients distincts en 2022 à l'exception du mois de février 2023,
- Changement de comportement dès mars 2022 (changement de stratégie commerciale ?)
- Les quatre plus gros clients, probablement des organisations, enregistrées comme des clients individuels risquer de biaiser le résultat de certaines analyses

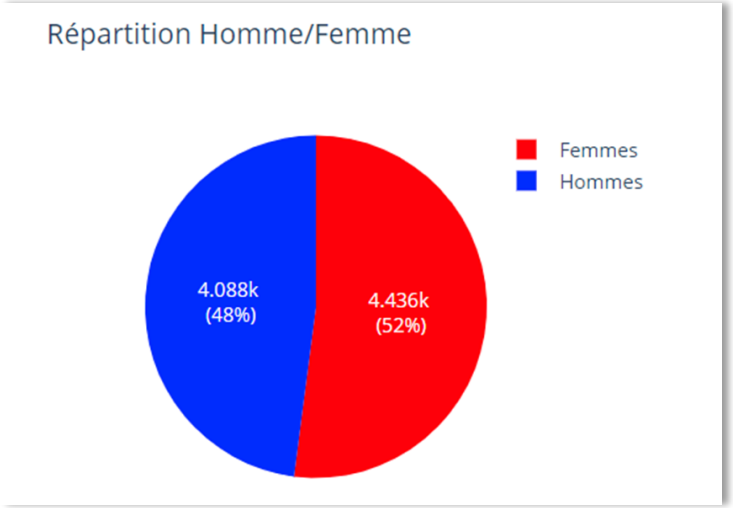
Aussi dans la suite, je travaille sur les données de la 2ème année, desquelles j'ai supprimé les outliers.

Profil des clients

Genre et âge des clients

Tests de normalité (Shapiro-Wilks) et de comparaison





- Test de Shapiro-Wilkes

On **peut rejeter** l'hypothèse de normalité de la distribution des âges des femmes avec **0.005 % de risque**
On **peut rejeter** l'hypothèse de normalité de la distribution des âges des hommes **avec 0.0007 % de risque**

- **Mann-Whitney** : on ne peut pas rejeter l'hypothèse de proximité des médianes avec p-value = 0.1757
- **Levene** : On ne peut pas rejeter l'hypothèse d'égalité des variances avec p-value = 0.2456

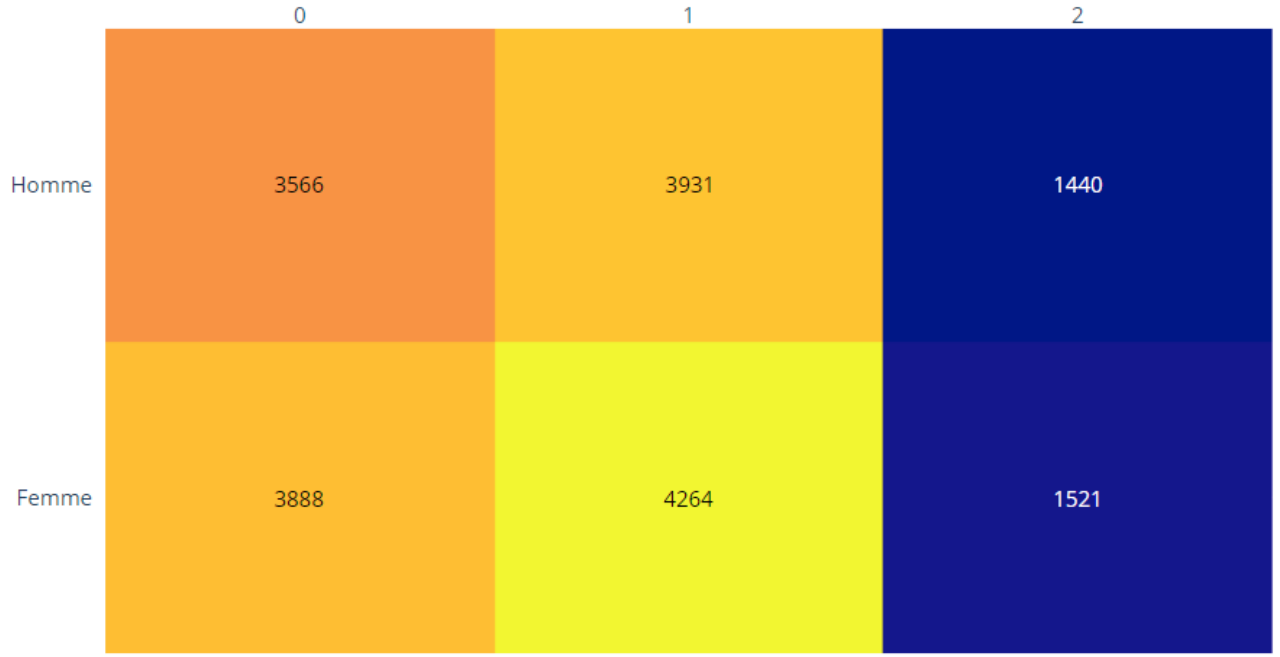
- 52% de femmes, 48% d'hommes
- Répartition des âges similaire chez les hommes et les femmes

Analyse du comportement des clients

Genre des clients et catégorie de livres achetés

Nombre de clients distincts par genre et par catégorie : Test d'association (Khi2)

	Catégorie 0	Catégorie 1	Catégorie 2
Homme	3566	3931	1440
Femmes	3888	4264	1521



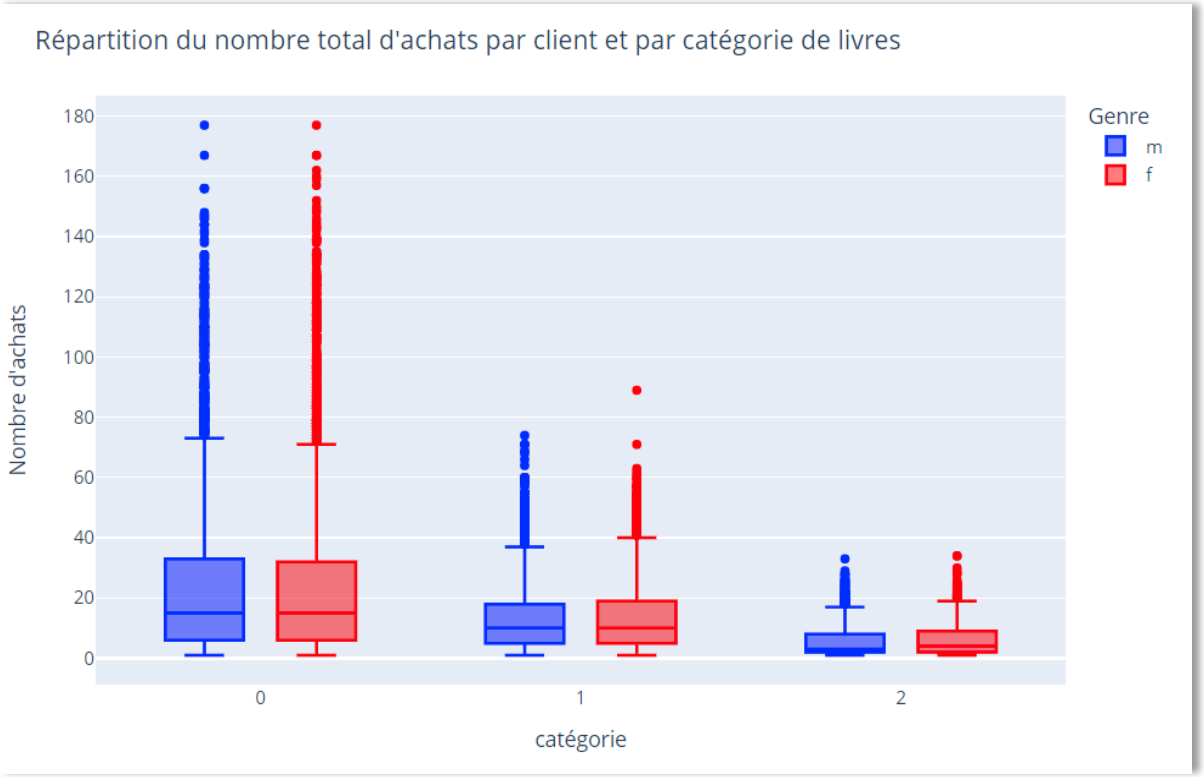
Résultat

On ne peut pas rejeter l'hypothèse H0 d'indépendance avec p-value = 0.7595

Il existe une très forte probabibilité que le genre du client n'influe pas sur la catégorie de livre qu'il achète

Genre des clients et catégorie de livres achetés

Tests de comparaison (Mann-Whitney, Levene)



Catégorie 0

- Mann-Whitney : p-value = 0.6834
- Levene : p-value = 0.6112

Catégorie 1

- Mann-Whitney : p-value = 0.6953
- Levene : p-value = 0.121

Catégorie 2

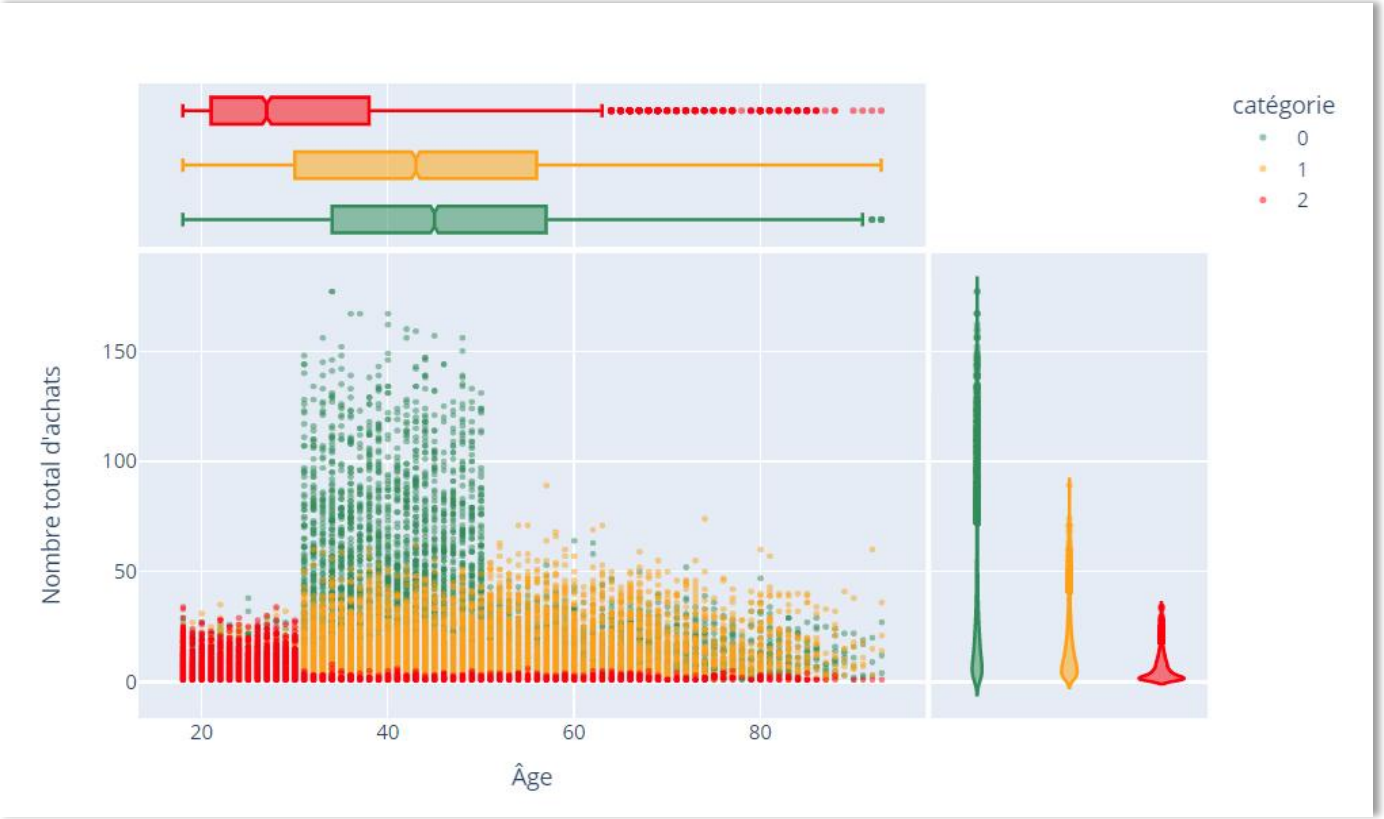
- Mann-Whitney : p-value = 0.205
- Levene : p-value = 0.992

Résultat

Les hypothèses de proximité des médianes et d'égalité des variances ne sont rejetées pour aucune catégorie.

Il existe une très forte probabibilité que le genre du client n'influe pas sur la catégorie de livre qu'il achète (résultat corrélé avec le précédent)

Âge des clients et catégories de livres achetés

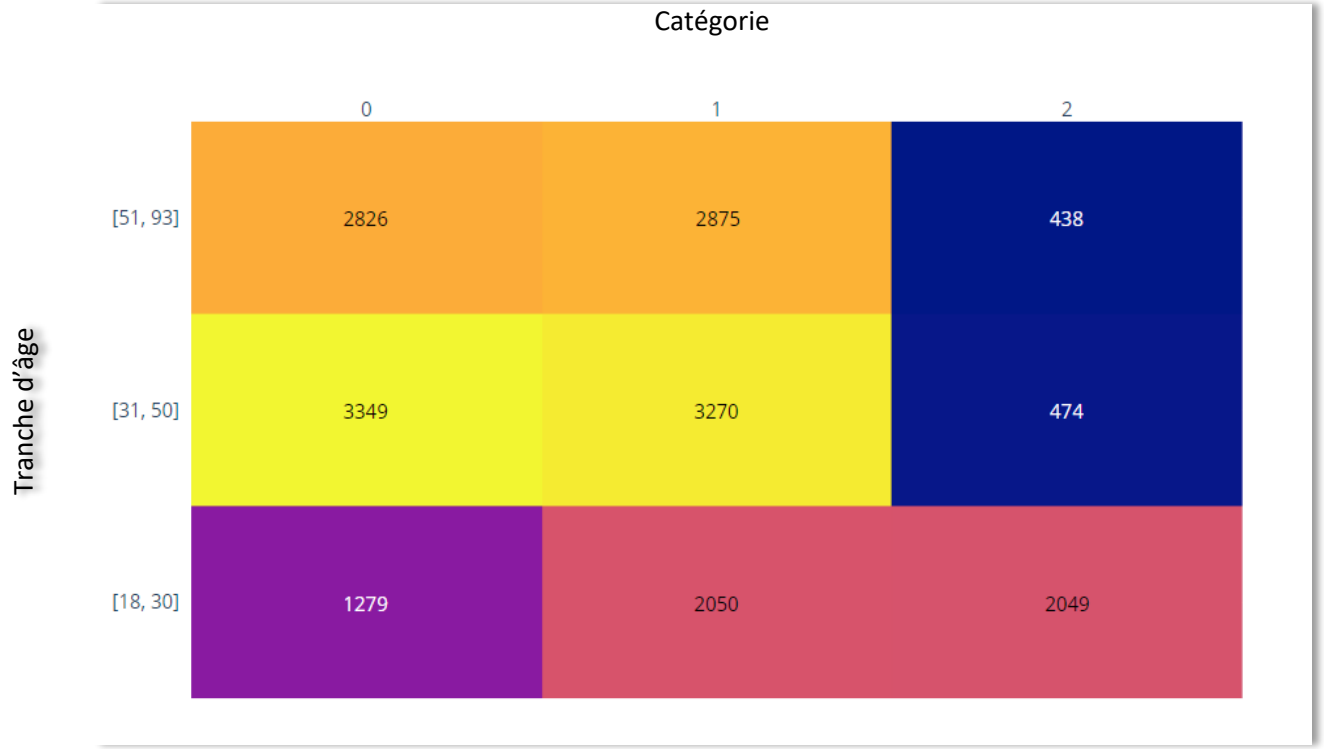


- Détermination de tranches d'âge de façon empirique : [18,30][31,50][51,93]
- Utilisation des tranches d'âge en tant que variable qualitative

Âge des clients et catégories de livres achetés

Nombre de clients distincts par tranche d'âge et par catégorie : Test d'association (Khi2)

Catégorie Tranche d'âge	Catégorie 0	Catégorie 1	Catégorie 2
[51, 93]	2826	2875	438
[31, 50]	3349	3270	474
[18, 30]	1279	2050	2049



Résultat

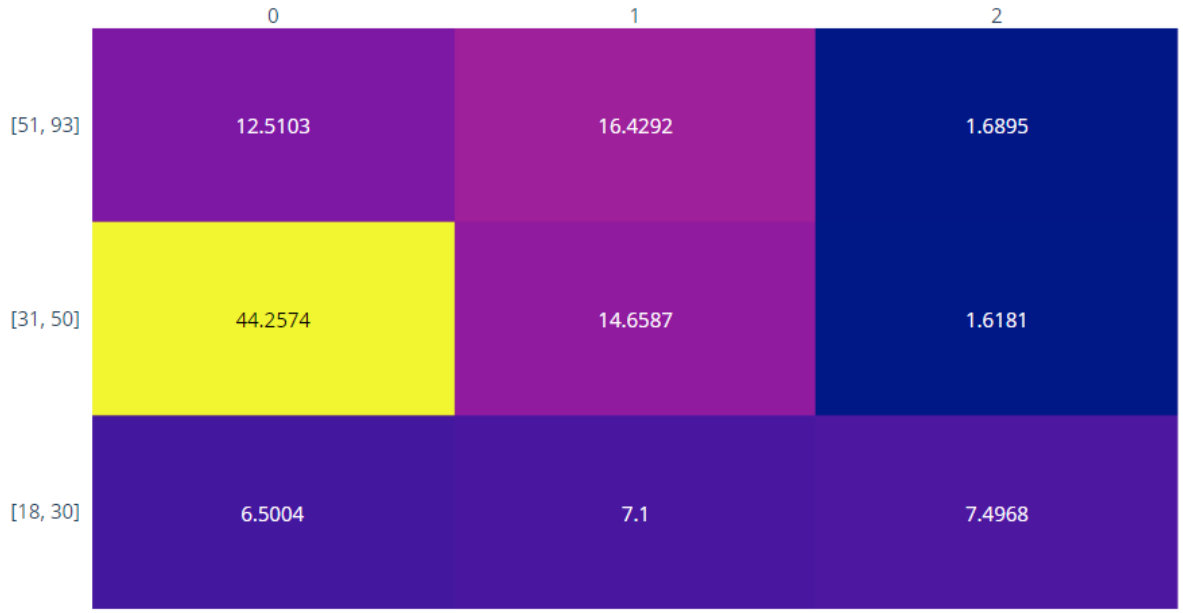
On peut rejeter l'hypothèse d'indépendance entre tranche d'âge et catégorie d'achat avec 0.0 % de risque

Il existe une très probabibilité que la tranche d'âge à laquelle le client appartient influe sur la catégorie de livre qu'il achète

Âge des clients et catégories de livres achetés

Nombre moyen d'achat par catégorie et tranche d'âge : Double-check

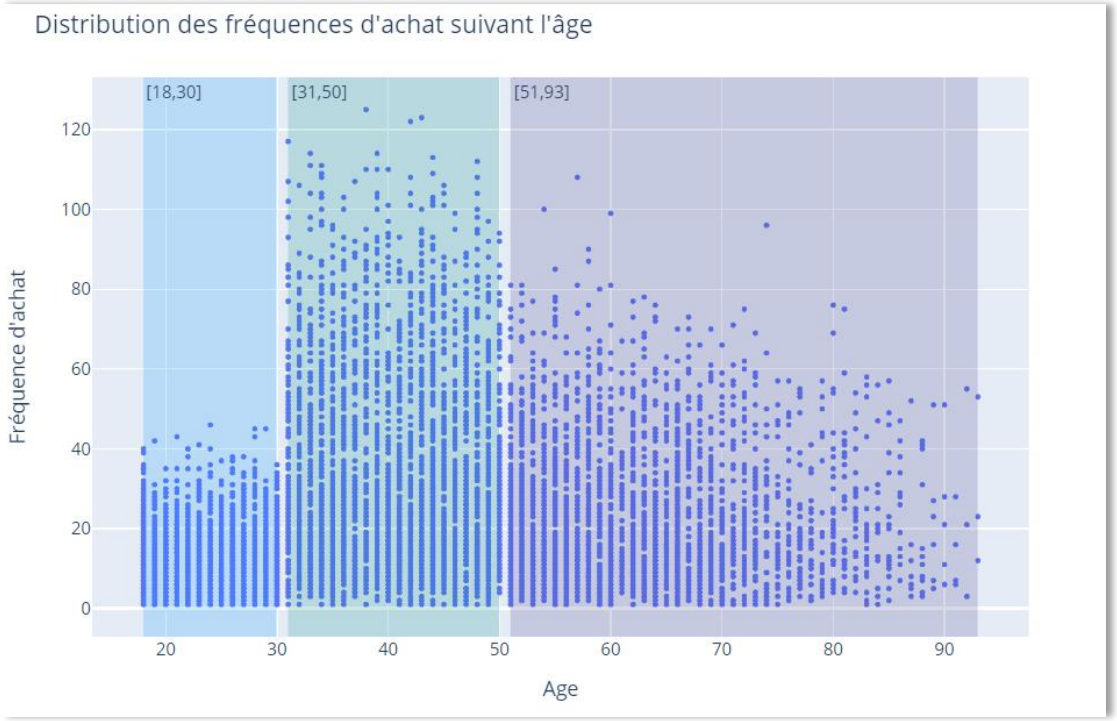
Catégorie Tranche d'âge	Catégorie 0	Catégorie 1	Catégorie 2
[51, 93]	12.5103	16.4292	1.6895
[31, 50]	44.2574	14.6587	1.6181
[18, 30]	6.5004	7.1000	7.4968



On peut observer les très fortes disparités dans les choix de catégorie de livres suivant la tranche d'âge du client (résultat corrélé avec le précédent)

Âge des clients et fréquence d’achat

Test de comparaison : Kruskal-Wallis



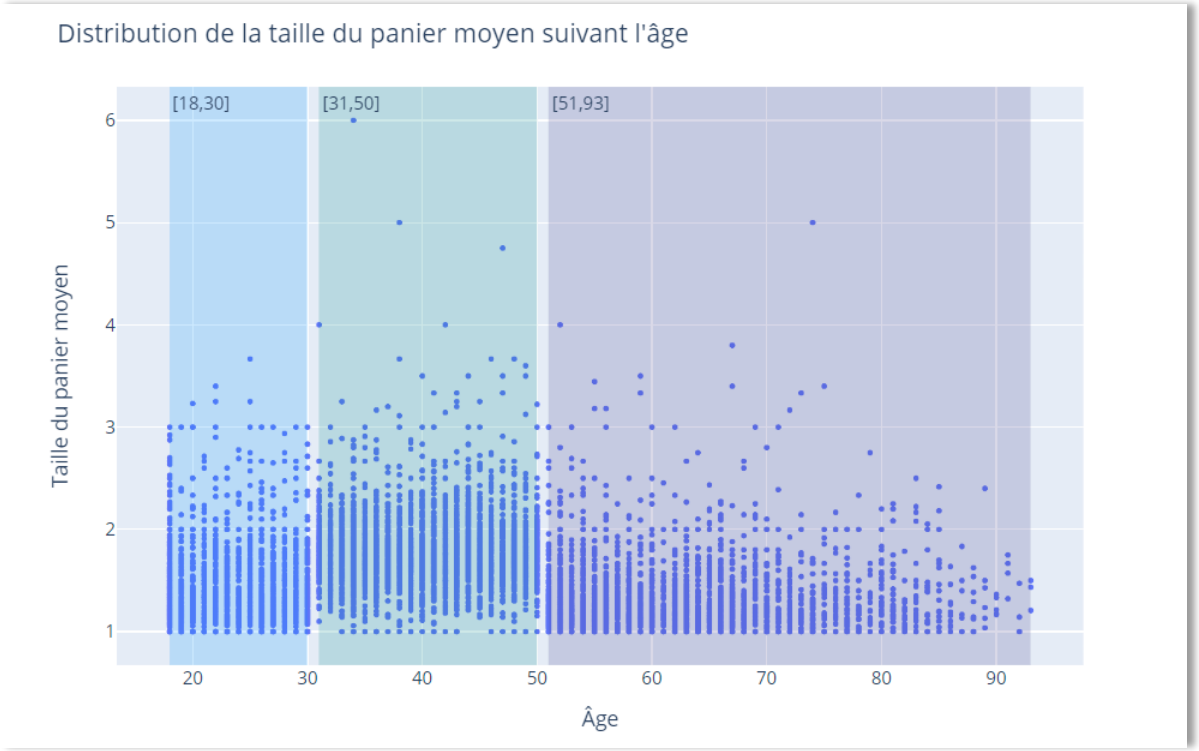
Résultat

On peut rejeter l'hypothèse H0 de similarité des distributions avec 0.0 % de risque

Il existe une très forte probabibilité que la tranche d’âge à laquelle le client appartient soit corrélé avec sa fréquence d’achat

Âge des clients et taille du panier moyen

Test de comparaison : Kruskal-Wallis



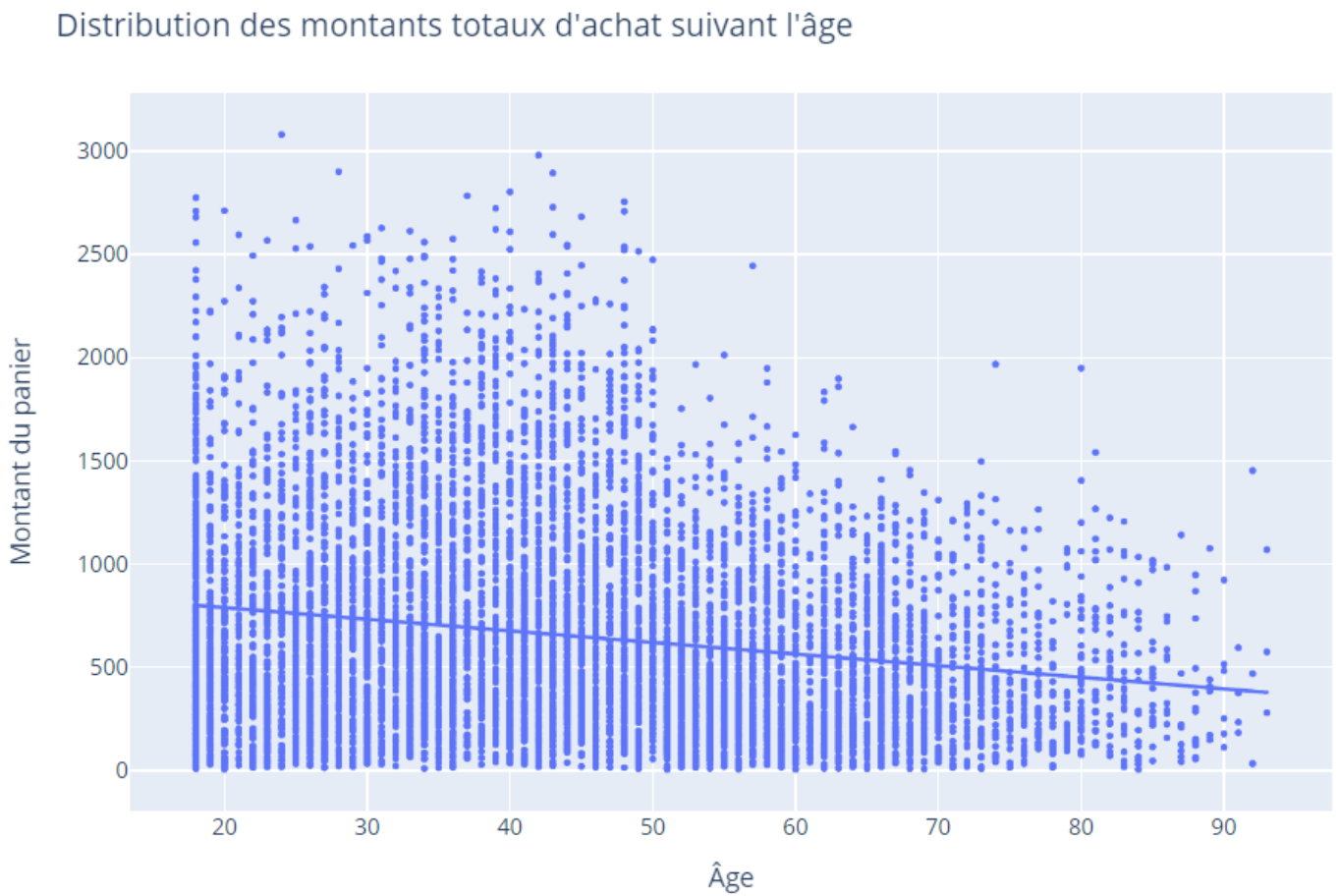
Résultat

On peut rejeter l'hypothèse H0 de similarité des distributions avec 0.0 % de risque

Il existe une très forte probabibilité que la tranche d'âge à laquelle le client appartient soit corrélé avec la taille de son panier moyen

Âge des clients et montant total des achats

Coefficient de corrélation (Spearman)



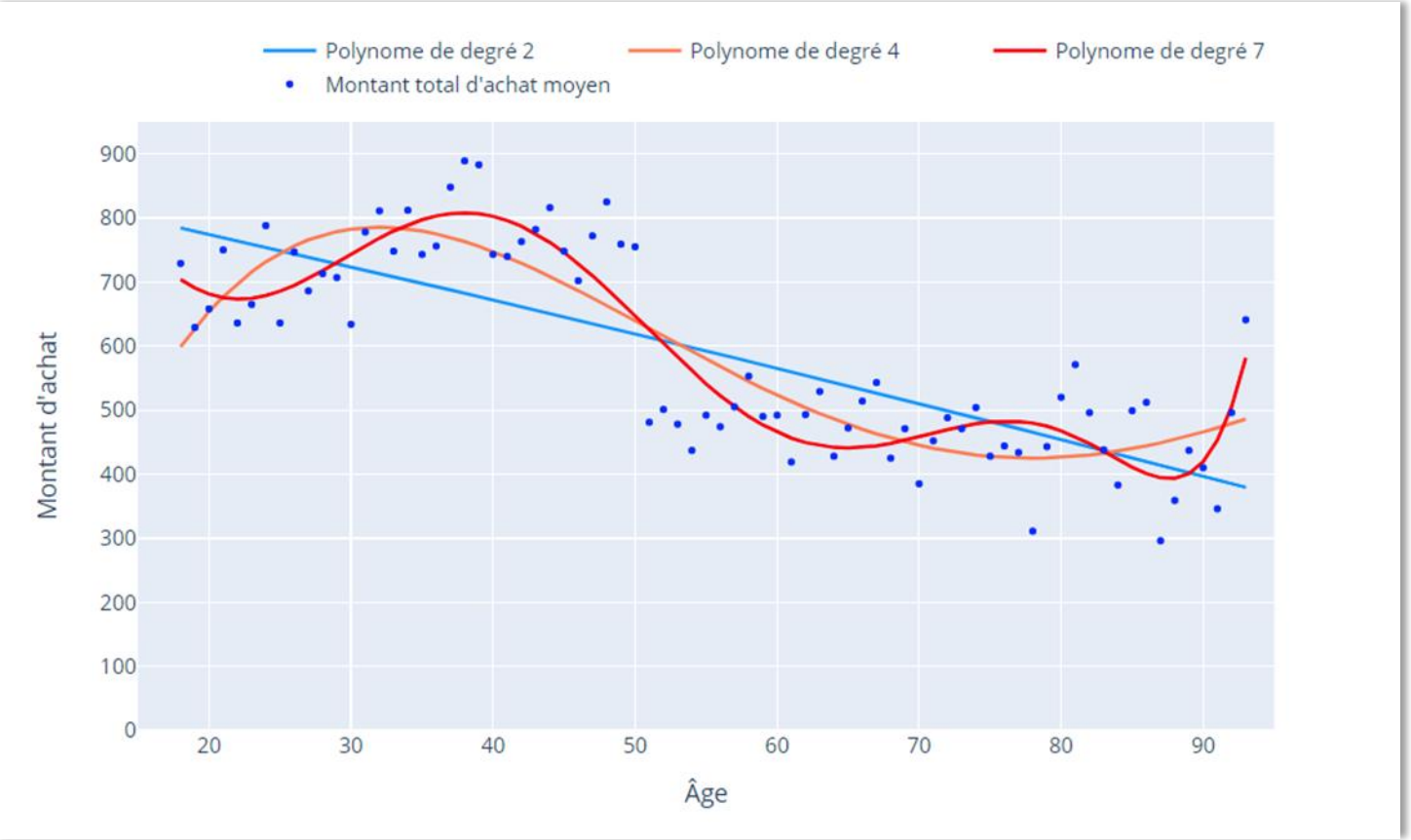
- Utilisation de l'âge comme variable quantitative

- Autre variable quantitative : la moyenne des montants des achats par âge

Coefficient de Spearman : -0.8734

Âge des clients et montant total des achats

Régression linéaire et polynomiale



- Polynome de degré 2

Root Mean Square Error : 99.88
Coefficient de corrélation r2 : 0.584

- Polynome de degré 4

Root Mean Square Error : 78.63
Coefficient de corrélation r2 : 0.742

- Polynome de degré 7

Root Mean Square Error : 65.05
Coefficient de corrélation r2 : 0.823

Probabilité qu'un client achète la référence 0_525 s'il a acheté la référence 2_159

Méthode utilisée

Pour chaque client, je ne conserve que la date du premier achat pour chaque référence.

Je m'intéresse :

- au nombre de clients qui ont acheté la référence 0_525, le même jour ou après la référence 2_159
- au nombre de clients qui ont acheté la référence 2_159 d'une façon générale (1er achat)

Résultat

La probabilité qu'un client achète la référence 0_525 sachant qu'il a acheté la référence 2_159 est de : **0.8557**
autrement dit, si un client achète la référence 2_159, il y a 85.57 % de chances qu'il achète la référence 0_525

MERCI DE VOTRE ATTENTION ☺

Bonne journée !