



Lapage

Analyse des ventes d'un site d'e-commerce

OpenClassrooms - Parcours Data Analyst V2
Projet 6

Valérie Chadeau



Lapage

Une librairie physique avec plusieurs points de vente qui, au vu de son succès, a décidé depuis 2 ans d'ouvrir un site de vente en ligne.

Ma mission

Après deux ans d'existence, la société souhaite faire l'analyse des ventes en ligne de son site : points forts, points faibles, les comportements clients, etc. Consultant data analyst rattaché au service marketing, je suis en charge de cette analyse.

Sommaire

PARTIE 1:Chargement, exploration et nettoyage des données

PARTIE 2: Analyse des données / Analyses univariées

CHIFFRES D'AFFAIRES : [Analyse globale (du 01/03/2022 au 28/02/2023)
Evolution par rapport à période 1 (du 01/03/2021 au 28/02/2022)
Détail évolution CA mensuel sur les 2 périodes (mars 2021 => février 2023)
Comparaison mensuelle entre les 2 périodes / (CA par catégorie)
Décomposition en Moyenne mobile d'ordre 3 entre mars 2021 et février 2023
Répartition entre clients - analyse de concentration
Répartition par catégorie sur les 2 périodes

Distribution des prix

Zoom sur les références littéraires

Profil des clients : [Clients professionnels : analyse globale
Clients non professionnels :
 Analyse par genre
 Analyse par âge

PARTIE 3:Analyse des données – Test statistiques

Corrélation genre client – catégories des livres achetés

Corrélation âge client – montant total des achats

Corrélation âge client – fréquence des achats

Corrélation âge client – taille panier moyen

Corrélation âge client – catégorie d'achat

Pré-requis : les données source

Trois fichiers de données fournis au format « csv » :

- customers.csv : le fichier des clients
- products.csv : le catalogue des produits
- transactions.csv : le fichier des transactions : 1 ligne correspond à 1 produit acheté au cours d'un achat

PARTIE 1

Chargement, exploration et nettoyage des données

Les étapes :

- Importation dans un dataframe et vérification cohérence des données
- Vérification de la présence de valeurs nulles dans le fichier
- Vérification de la présence de doublons dans la clé
- Analyse statistiques descriptives rapides des valeurs du dataframe
- Vérification du formalisme des variables

Fichier “customers.csv”

- pas de problème de formatage incohérent de donnée
- type de données cohérent par rapport au nom des colonnes
- pas de valeur nulle dans le jeu de données
- pas de doublons sur la clé « client_id »
- pas d'incohérence au niveau des minimum et maximum d'année de naissance
- 2 « client_id » dont le formalisme est incorrect : à vérifier s'il y a eu des commandes passées avec ces 2 identifiants
- pas d'anomalie détectée : il y a 2 valeurs possibles pour la donnée «sex » : M ou F

	client_id	sex	birth
0	c_4410	f	1967
1	c_7839	f	1975
2	c_1699	f	1984
3	c_5961	f	1962
4	c_5320	m	1943

(8623, 3)

```
client_id    object
sex          object
birth        int64
dtype: object
```

#Analyse statistiques descriptives rapides des valeurs de la table
`customers.describe().T`

	count	mean	std	min	25%	50%	75%	max
birth	8623.0	1978.280877	16.919535	1929.0	1966.0	1979.0	1992.0	2004.0

#Vérification du formalisme du champ client_id qui doit être de la forme 'c_9999' en utilisant une expression régulière
`customers.client_id.str.contains('c_[0-9]+', regex= True, na=False).sum()`
`wrongformat=customers[~customers.client_id.str.contains('c_[0-9]+', regex= True, na=False)]`
`wrongformat`

8621

	client_id	sex	birth
2735	ct_0	f	2001
8494	ct_1	m	2001

#Vérification valeur 'sexe'
`customers['sex'].unique()`

`array(['f', 'm'], dtype=object)`

Fichier “products.csv”

- pas de problème de formatage incohérent de donnée
- type de données cohérent par rapport au nom des colonnes
- pas de valeur nulle dans le jeu de données
- pas de doublons sur la clé « id_prod »
- des incohérences au niveau de la colonne prix (présence de prix négatifs)
- 75% des prix sont inférieurs à 22,99euros avec un maximum à 300 euros : valeur extrême mais pas aberrante
- 3 types de catégorie : 0, 1 et 2

	id_prod	price	categ
0	0_1421	19.99	0
1	0_1368	5.13	0
2	0_731	17.99	0
3	1_587	4.99	1
4	0_1507	3.99	0

(3287, 3)

```
id_prod    object
price      float64
categ      int64
dtype: object
```

```
#Analyse statistiques descriptives rapides des valeurs de la table
products.describe().T
```

	count	mean	std	min	25%	50%	75%	max
price	3287.0	21.856641	29.847908	-1.0	6.99	13.06	22.99	300.0
categ	3287.0	0.370246	0.615387	0.0	0.00	0.00	1.00	2.0

Fichier "products.csv" (suite)

→ 1 seul prix est négatif : à vérifier s'il est présent dans les commandes

→ 1 seul « id_prod » est mal formaté : c'est le même produit dont le prix est négatif : à vérifier s'il est présent dans les commandes

```
#Pour Les prix <0  
products.loc[products['price']<0]
```

	id_prod	price	categ
731	T_0	-1.0	0

```
#Vérification valeur 'categ'  
products['categ'].unique()
```

```
array([0, 1, 2], dtype=int64)
```

```
#Vérification du formalisme de la valeur 'id_prod' qui doit être de type : categ + '_' + 999  
products.id_prod.str.contains('[0-2]_[0-9]+', regex=True, na=False).sum()  
wrongformat=products[~products.id_prod.str.contains('[0-2]_[0-9]+', regex=True, na=False)]  
wrongformat
```

```
3286
```

	id_prod	price	categ
731	T_0	-1.0	0

Fichier “transactions.csv”

→ pas de problème de formatage incohérent de donnée sauf pour la colonne « date » dont il faut modifier le type de donnée

→ pas de valeur nulle dans le jeu de données

→ 200 lignes concernent le produit T_0 : elles correspondent à des tests (date, session_id et client_id au format test)

	id_prod	date	session_id	client_id
0	0_1518	2022-05-20 13:21:29.043970	s_211425	c_103
1	1_251	2022-02-02 07:55:19.149409	s_158752	c_8534
2	0_1277	2022-06-18 15:44:33.155329	s_225667	c_6714
3	2_209	2021-06-24 04:19:29.835891	s_52962	c_6941
4	0_1509	2023-01-11 08:22:08.194479	s_325227	c_4232

(679532, 4)

```
id_prod      object
date         object
session_id   object
client_id    object
dtype: object
```

```
# Vérification de la présence du produit "T_0" dans les transactions
transactions.loc[transactions['id_prod']=="T_0"].count()
transactions.loc[transactions['id_prod']=="T_0"].head(5)
```

```
id_prod      200
date         200
session_id   200
client_id    200
dtype: int64
```

	id_prod	date	session_id	client_id
3019	T_0	test_2021-03-01 02:30:02.237419	s_0	ct_0
5138	T_0	test_2021-03-01 02:30:02.237425	s_0	ct_0
9668	T_0	test_2021-03-01 02:30:02.237437	s_0	ct_1
10728	T_0	test_2021-03-01 02:30:02.237436	s_0	ct_0
15292	T_0	test_2021-03-01 02:30:02.237430	s_0	ct_0

Fichier "transactions.csv" (suite)

→ les codes clients « ct_0 » et « ct_1 » ne correspondent qu'aux lignes de tests (200 transactions concernées) → suppression de ces données dans les différents dataframes pour la suite de notre analyse

→ conversion colonne « date » au format datetime

→ pas de doublon dans le dataframe

→ Il s'agit de l'ensemble des transactions passées entre le 01/03/2021 et le 28/02/2023 soit 24 mois

```
# Vérification que les clients "ct_0" et "ct_1" ne correspondent qu'aux transactions de tests
transactions.loc[(transactions['client_id']=="ct_0") | (transactions['client_id']=="ct_1") ].count()
```

11

```
id_prod      200
date          200
session_id    200
client_id     200
dtype: int64
```

```
#Suppression du code produit "T_0" dans le dataframe "products"
products=products[products['id_prod']!="T_0"]
products.shape
```

(3286, 3)

```
#Suppression des codes client "ct_0" et "ct_1" dans le dataframe "customers"
customers=customers[(customers['client_id']!="ct_0") & (customers['client_id']!="ct_1") ]
customers.shape
```

(8621, 3)

```
#Suppression des dates au format "test_xxxx" dans le dataframe "transactions" ce qui revient à supprimer les
#transactions effectuées par le client_id "T_0"
transactions=transactions[transactions['id_prod']!="T_0"]
transactions.shape
```

(679332, 4)

```
#Conversion de la colonne 'date' au format datetime
transactions['date']= pd.to_datetime(transactions['date'],errors = 'coerce')
transactions['date'].isna().sum()
transactions.dtypes
```

0

```
id_prod      object
date         datetime64[ns]
session_id    object
client_id     object
dtype: object
```

```
#Vérification de la présence de doublons dans le dataframe
transactions.duplicated().sum()
```

0

Création dataframe "join_final" (jointure des 3 tables)

→ jointure en 2 étapes entre les 3 tables

Constat :

- 21 clients n'ont jamais passé de commandes → suppression des lignes
- aucune commande passée concernant 21 produits → suppression des lignes
- 1 produit non référencé dans le catalogue (id_prod='0_2245') : 221 transactions concernées (environ 0.03% des données)
→ Choix de suppression des lignes (autre choix possible : remplacement du prix manquant en réalisant une imputation par la moyenne du prix des produits de la catégorie 0)

→ conversion colonne « date » au format datetime

→ conversion colonnes « categ » et « birth » en entier

→ ajout colonnes « Age » et « mois »

→ pas de doublon dans le dataframe

Etape 1 : Jointure entre les dataframes "transactions" et "customers"

```
#Réalisation d'une jointure entre les dataframes "transactions" et "customers"
#de type 'outer'et analyse de la colonne "indicator"

join_transactions_customers=transactions.merge(customers, left_on='client_id', right_on='client_id', how='outer', indicator=True)
join_transactions_customers['_merge'].unique()

join_transactions_customers[join_transactions_customers['_merge']=='both']['_merge'].count()
join_transactions_customers[join_transactions_customers['_merge']=='right_only']['_merge'].count()

['both', 'right_only']
Categories (2, object): ['both', 'right_only']

679332

21
```

Etape 2 : jointure entre les dataframes "join_transactions_customers" et "products"

```
#Réalisation d'une jointure entre les dataframes "join_transactions_customers" et "products"
#de type 'outer'et vérification de la colonne "indicator"

join_final=join_transactions_customers.merge(products, left_on='id_prod', right_on='id_prod', how='outer', indicator=True)
join_final['_merge'].unique()

#Analyse des différents types de jointure
join_final[join_final['_merge']=='both']['_merge'].count()
join_final[join_final['_merge']=='right_only']['_merge'].count()
join_final[join_final['_merge']=='left_only']['_merge'].count()
join_final[join_final['_merge']=='left_only']['id_prod'].unique()

['both', 'left_only', 'right_only']
Categories (3, object): ['both', 'left_only', 'right_only']

679111

21

221

array(['0_2245'], dtype=object)

#Ajout des colonnes 'Age' et 'mois'
join_final['Age']=2022-join_final['birth']
join_final['mois']=join_final['date'].dt.strftime("%b %Y")
```

	id_prod	date	session_id	client_id	sex	birth	price	categ	Age	mois
0	0_1518	2022-05-20 13:21:29.043970	s_211425	c_103	f	1986	4.18	0	36	May 2022
1	0_1518	2021-07-20 13:21:29.043970	s_64849	c_103	f	1986	4.18	0	36	Jul 2021
2	0_1518	2022-08-20 13:21:29.043970	s_255965	c_103	f	1986	4.18	0	36	Aug 2022
3	0_1518	2021-05-09 11:52:55.100386	s_32104	c_6714	f	1968	4.18	0	54	May 2021
4	0_1518	2022-05-30 01:17:07.487046	s_216118	c_6714	f	1968	4.18	0	54	May 2022

PARTIE 2

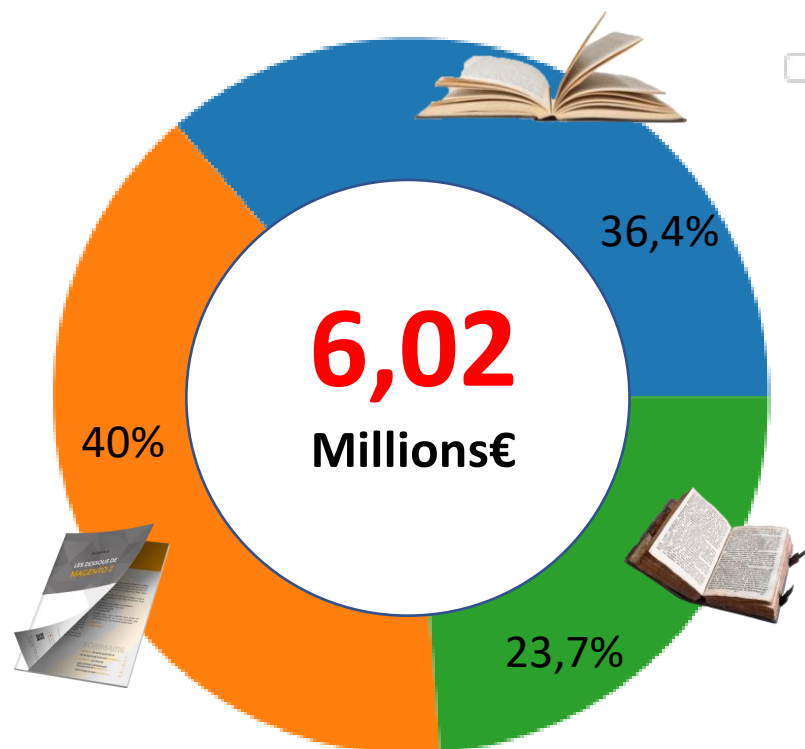
Analyse des données

-

Analyses univariées

CHIFFRES D'AFFAIRES du 01/03/2022 au 28/02/2023

(période 2) – Analyse globale



Catégorie 0

2 189 302€



Catégorie 1

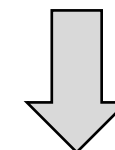
2 406 338€



Catégorie 2

1 426 350€

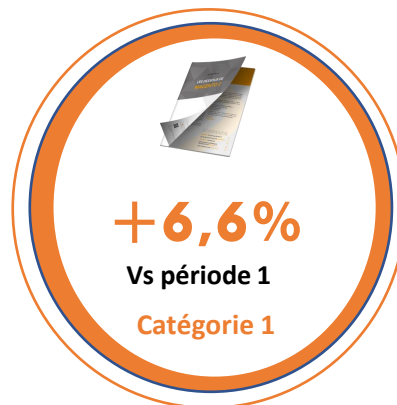
8 621 clients



173 102 achats soit
341 926 livres achetés

Evolution CA par rapport période 1

(du 01/03/2021 au 28/02/2022)



Chiffre d'affaires global

+ 3,15%

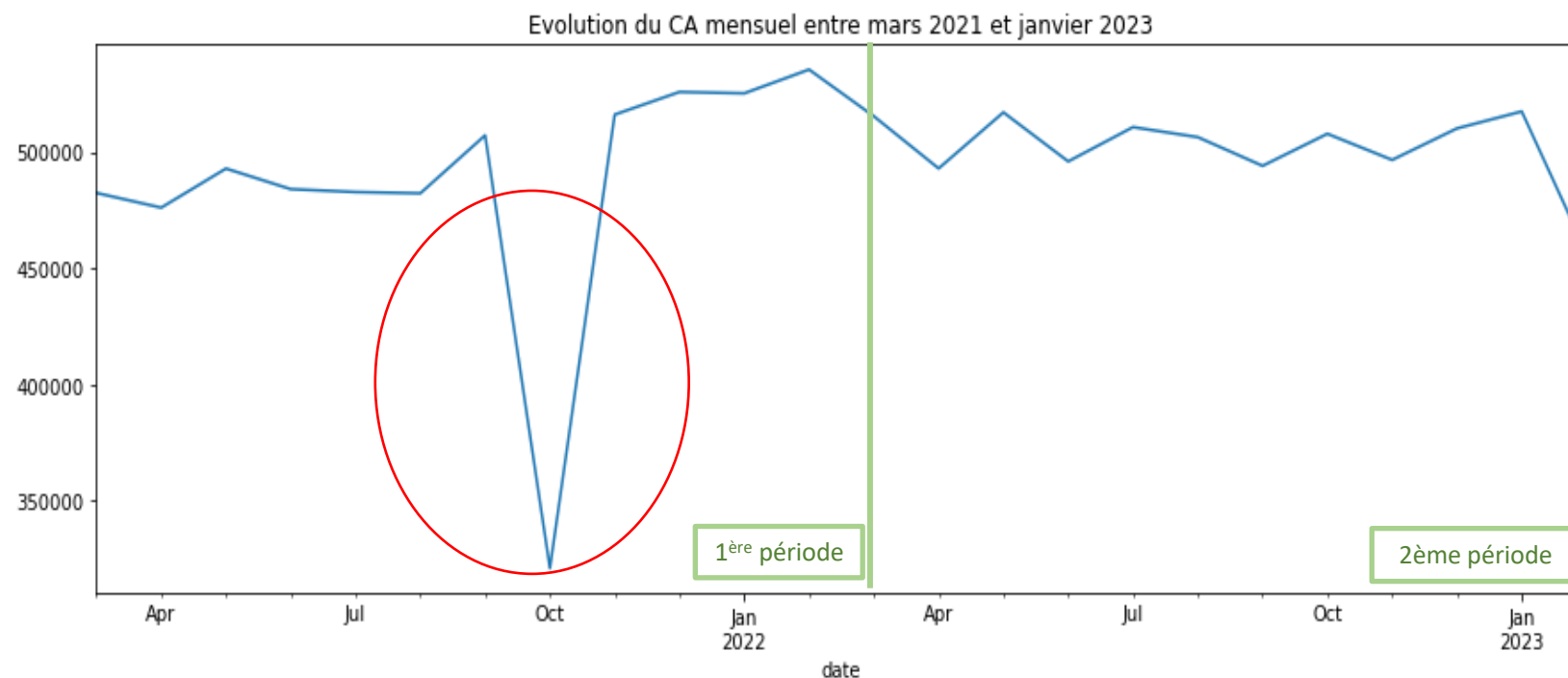
Vs période 1



+2,27% achats
+1,4% livres achetés

Détail évolution CHIFFRE D'AFFAIRES mensuels sur les 2 périodes

(mars 2021 => février 2023)



CA du 01/03/2021 au 28/02/2022 (période 1)
5 831 737 €

CA du 01/03/2022 au 28/02/2023 (période 2)
6 021 992 €



anomalie de données en octobre 2021

➔ Focus anomalie octobre 2021

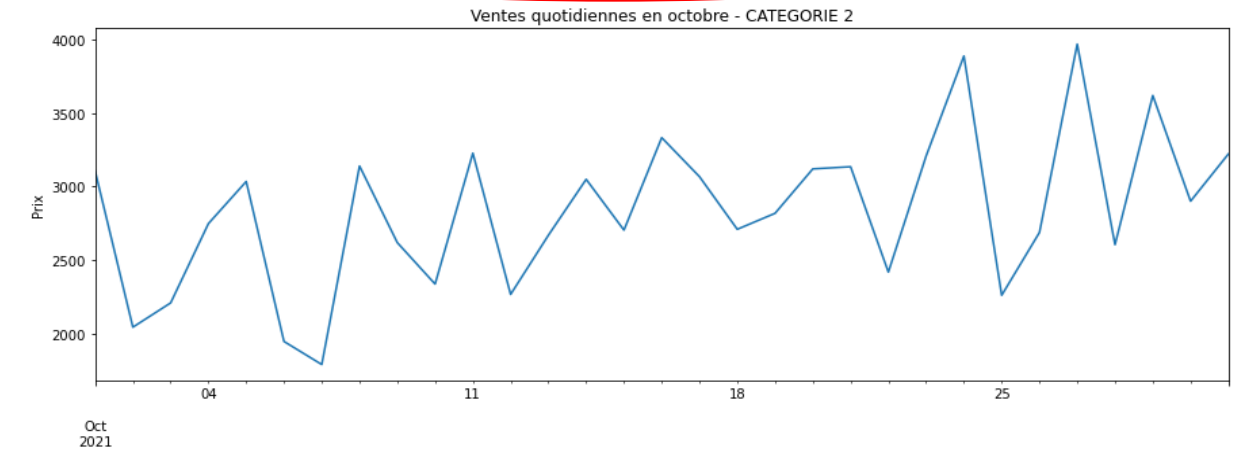
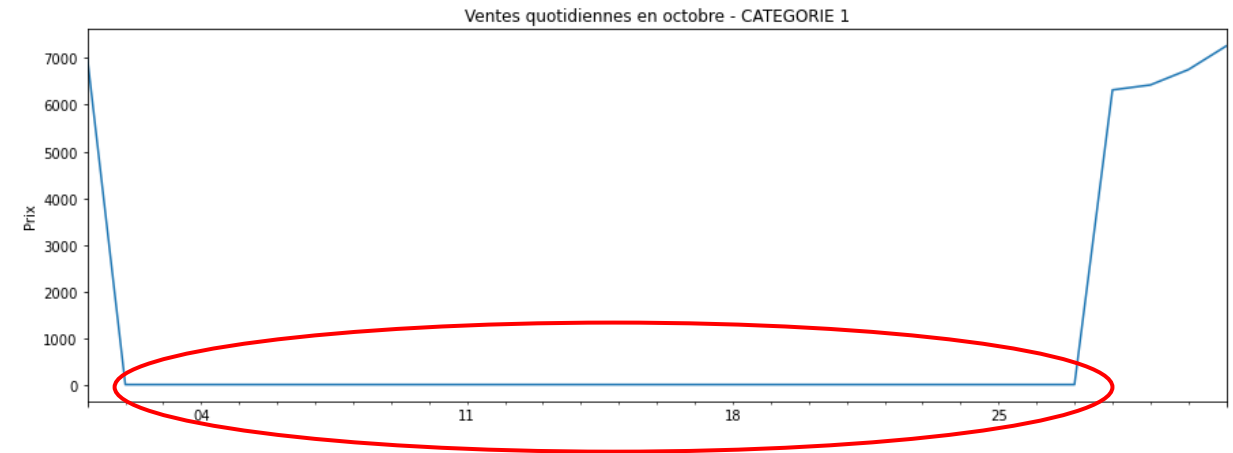
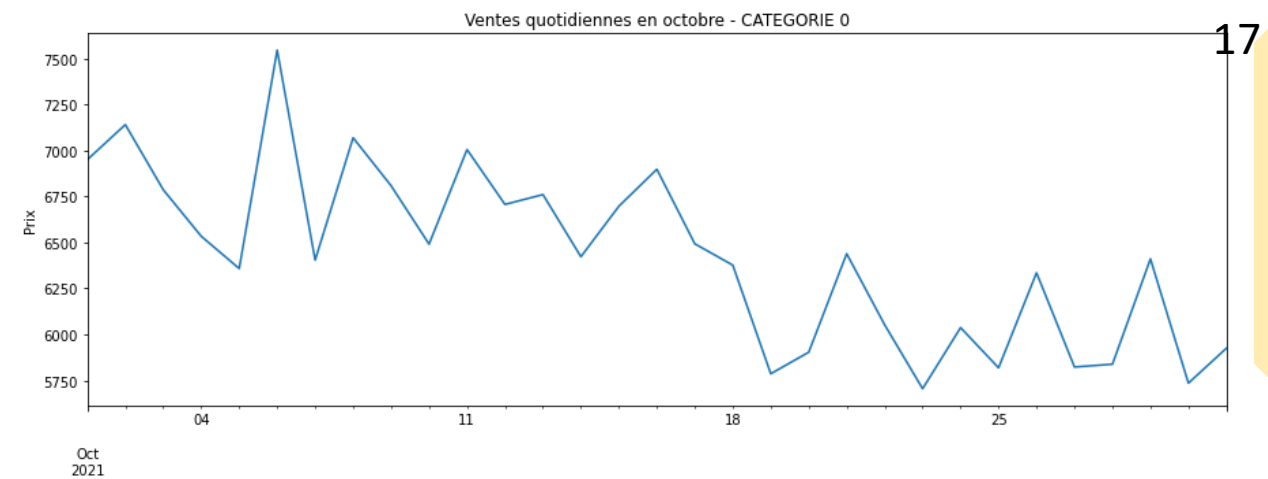


Entre le 2 et le 27 octobre : pas de transactions pour la catégorie 1.

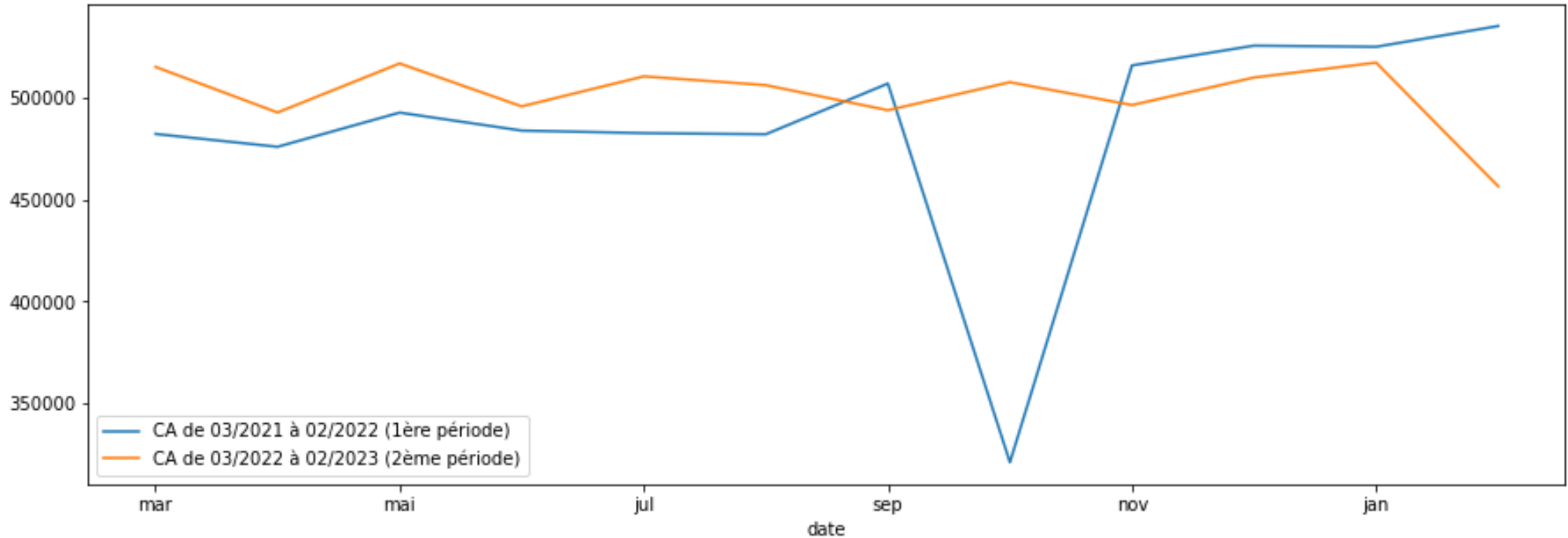
➔ Causes possibles :

- pas de vente dans la catégorie 1
- problème technique de remontée de données
- produits qui ne sont plus en stock

A voir avec les services informatiques et marketing



Evolution du CA sur 2 périodes



- Mars ➔ Août : montant ventes 2ème période > montant ventes 1ère période
- à partir de septembre : montant ventes 1ère période > montant ventes 2ème période (non confirmé pour octobre)

CHIFFRE D'AFFAIRES

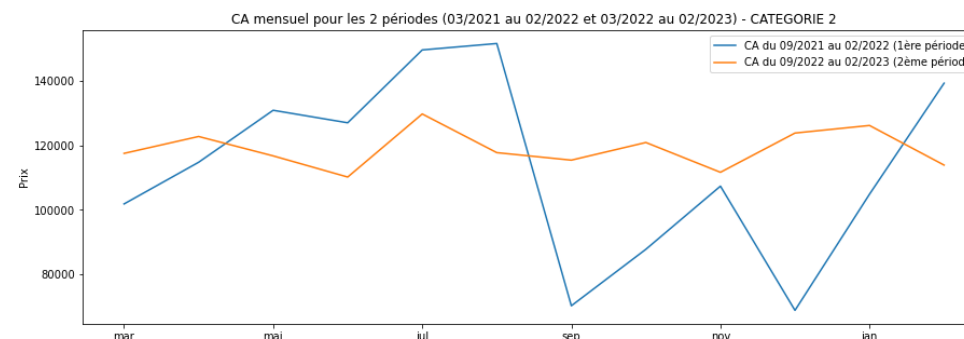
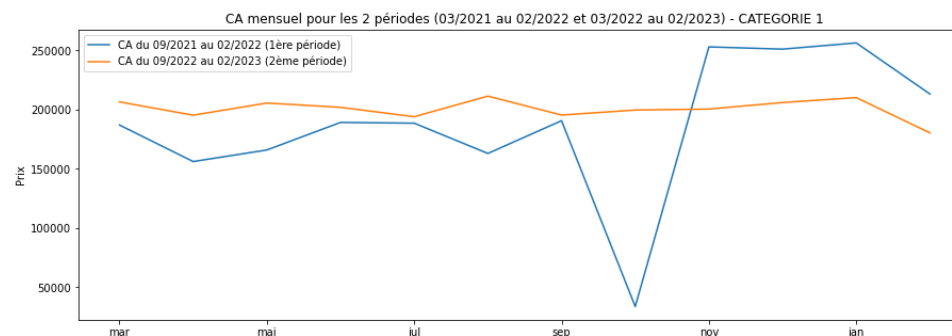
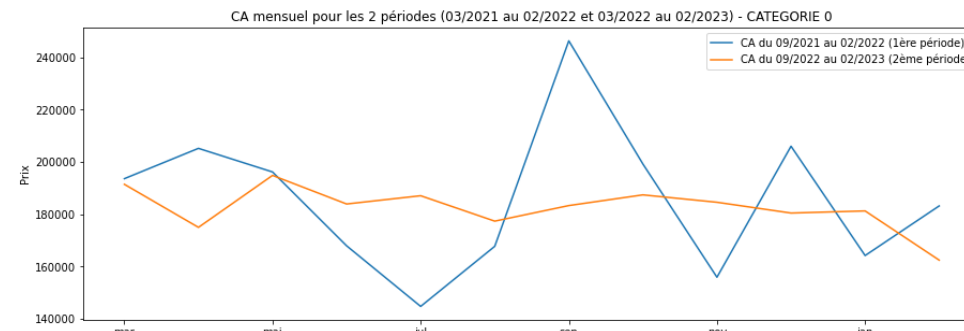
Comparaison entre les 2 périodes / catégorie (1/2)

	Cat 0	Cat 1	Cat 2
CA du 01/03/2021 au 28/02/2022	2 230 428.29€	2 247 384.41€	1 353 924.35€
CA du 01/03/2022 au 28/02/2023	2 189 302.68€	2 406 338.28€	1 426 350.672€
Evolution entre les 2 périodes	-1.87%	+6.6% (*)	+5.08%

(*) à confirmer suite incident octobre 2021


Pour les 3 catégories :

- 1^{ère} période : fluctuations des ventes très marquées
- 2^{ème} période : ventes beaucoup plus stables mais baisse marquée à partir de janvier 2023



CHIFFRE D'AFFAIRES

Comparaison entre les 2 périodes / catégorie (2/2)

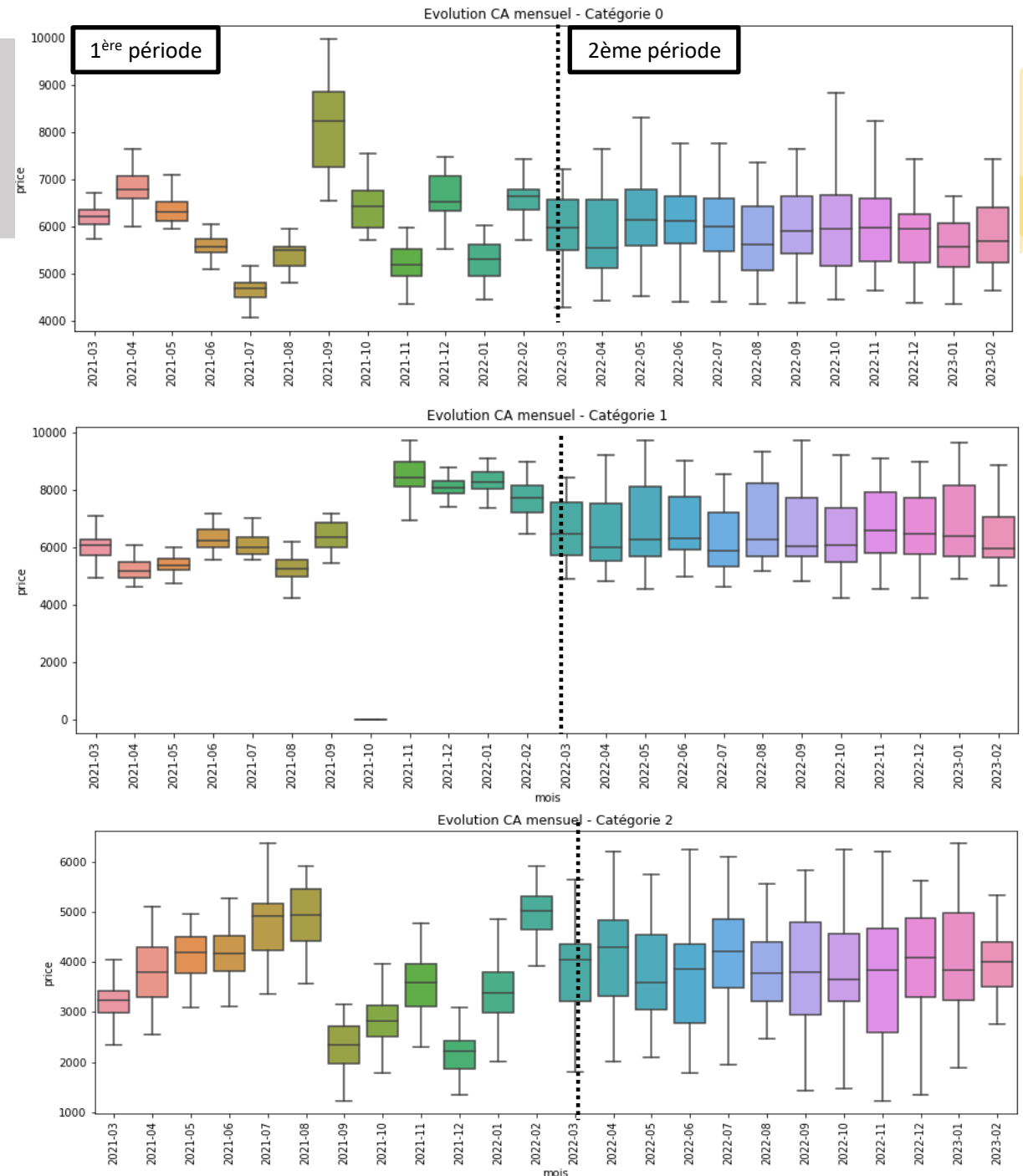
- **1ère période** : fluctuations de ventes
-  1ère année d'existence du site : demande non anticipée → fluctuation importante des stocks
- **2ème période** : ventes beaucoup plus stables mais montant des achats plus variés

Saisonnalité :

Catégorie 0 : plus de ventes en septembre (rentrée scolaire)

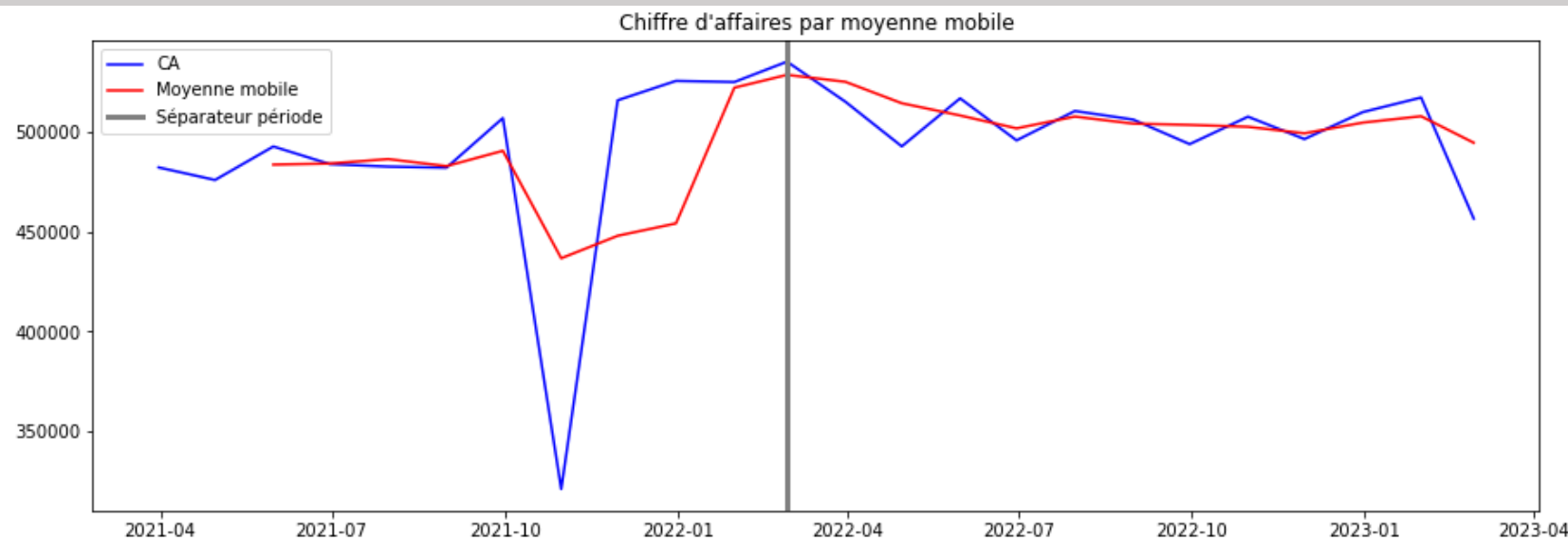
Catégorie 1 : plus de ventes en novembre, décembre et janvier (fêtes de fin d'année, semestre universitaire)

Catégorie 2 : plus de ventes en juillet, août et février (rentrée scolaire, semestre universitaire)



CHIFFRE D'AFFAIRES

Décomposition en Moyenne mobile d'ordre 3 entre mars 2021 et février 2023



Les tendances :

- mars 2021 → février 2022 :  du CA (incident en octobre 2021)
- à partir de mars 2022 :  avec une baisse plus nette en février 2023.

Nécessité pour compléter l'analyse de coupler ces données avec celles de la base de gestion des stocks et de l'historique de l'évolution du prix des articles.

CHIFFRE D'AFFAIRES

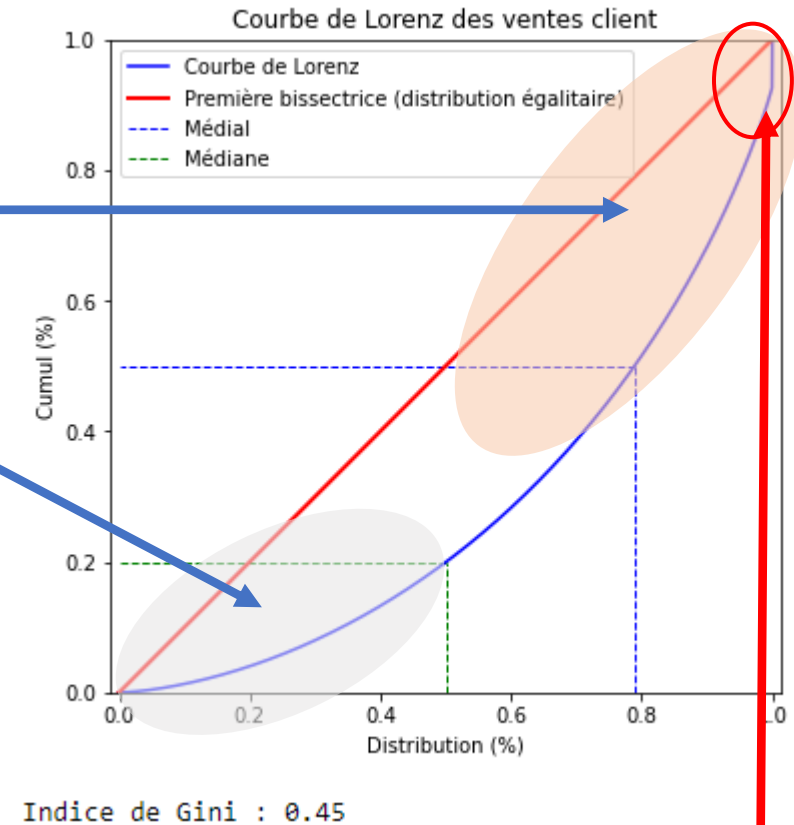
Répartition entre clients - analyse de concentration

→ Répartition du CA par client inégalitaire

- 21% des clients réalisent 50% du CA
- 50% des clients réalisent 20% du CA

Inégalité confirmée par l'indice de Gini à 0.45

On note une anomalie de la courbe qui correspond à une distribution totalement inégalitaire.



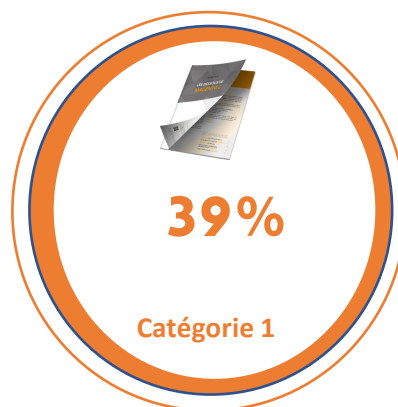
CHIFFRE D'AFFAIRES

Répartition par catégorie sur les 2 périodes

Part du chiffre d'affaires par catégorie



Le plus de livres proposés
Le plus de livres vendus



3 fois - de livres proposés que cat 0
1,8 fois - de livres vendus que cat 0



10 fois - de livres proposés que cat 0
12 fois - de livres vendus que cat 0



→ Réflexion nécessaire sur l'élargissement de l'offre des produits des catégories 1 et 2

Distribution des prix

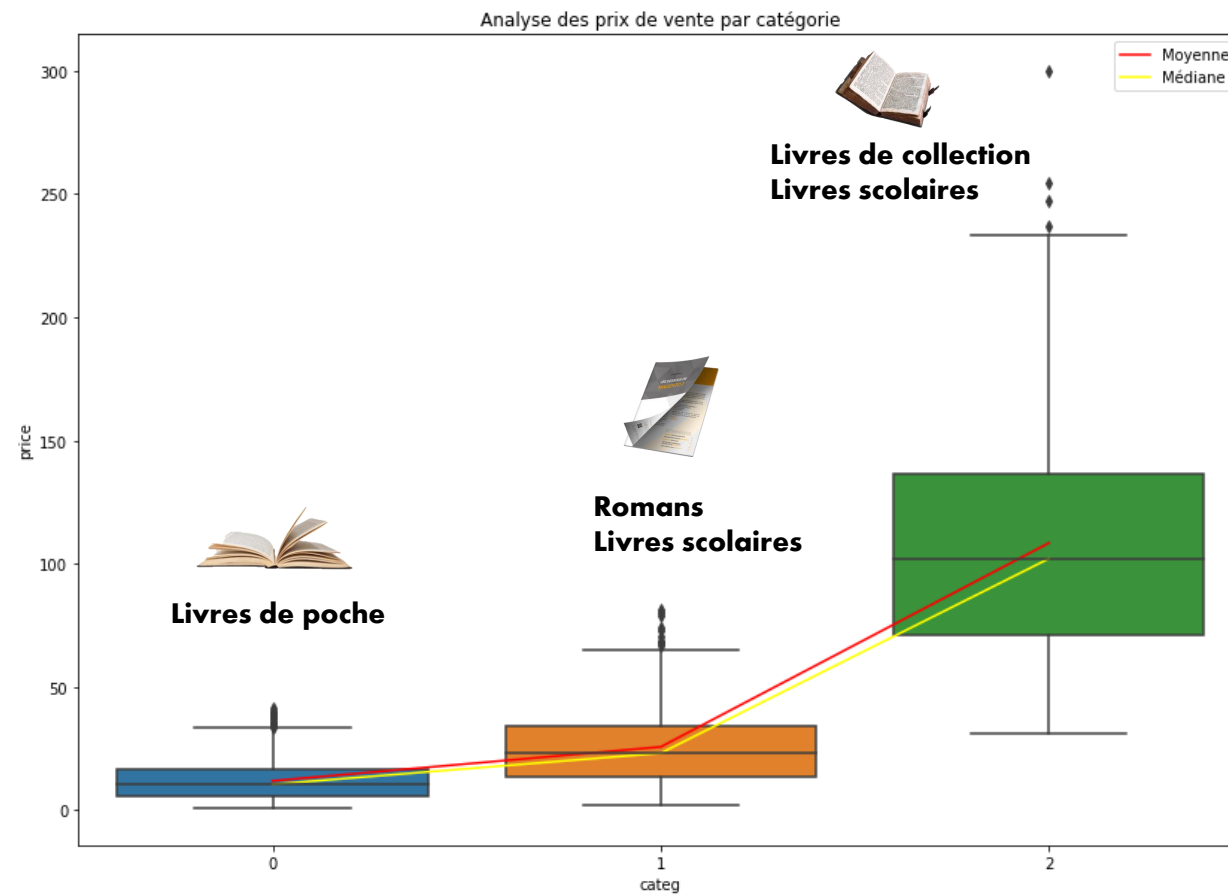
- **3 286** livres proposés répartis en 3 catégories
- Plage des prix : **de 0,62€ à 300€**
- 75% des prix < 13,07€

	Cat 0	Cat 1	Cat 2
Prix minimum	0,62€	2€	30,99€
Prix maximum	40,99€	80,99€	300€
Prix médian	10,32€	22,99€	101,99€

Les - chers

Prix
intermédiaires

Les + chers



ZOOM sur les références littéraires



Référence produit	Cat	Prix	Nombre livres vendus
1_369	1	23.99	2252
1_417	1	20.99	2189
1_414	1	23.83	2180
1_498	1	23.37	2128
1_425	1	16.99	2096
1_403	1	17.99	1960
1_412	1	16.65	1951
1_413	1	17.99	1945
1_406	1	24.81	1939
1_407	1	15.99	1935



VS



Référence produit	Cat	Prix	Nombre livres vendus
0_2201	0	20.99	1
0_1601	0	1.99	1
0_549	0	2.99	1
2_81	2	86.99	1
0_807	0	1.99	1
0_1683	0	2.99	1
0_1151	0	2.99	1
0_1633	0	24.99	1
0_833	0	2.99	1
0_886	0	21.82	1

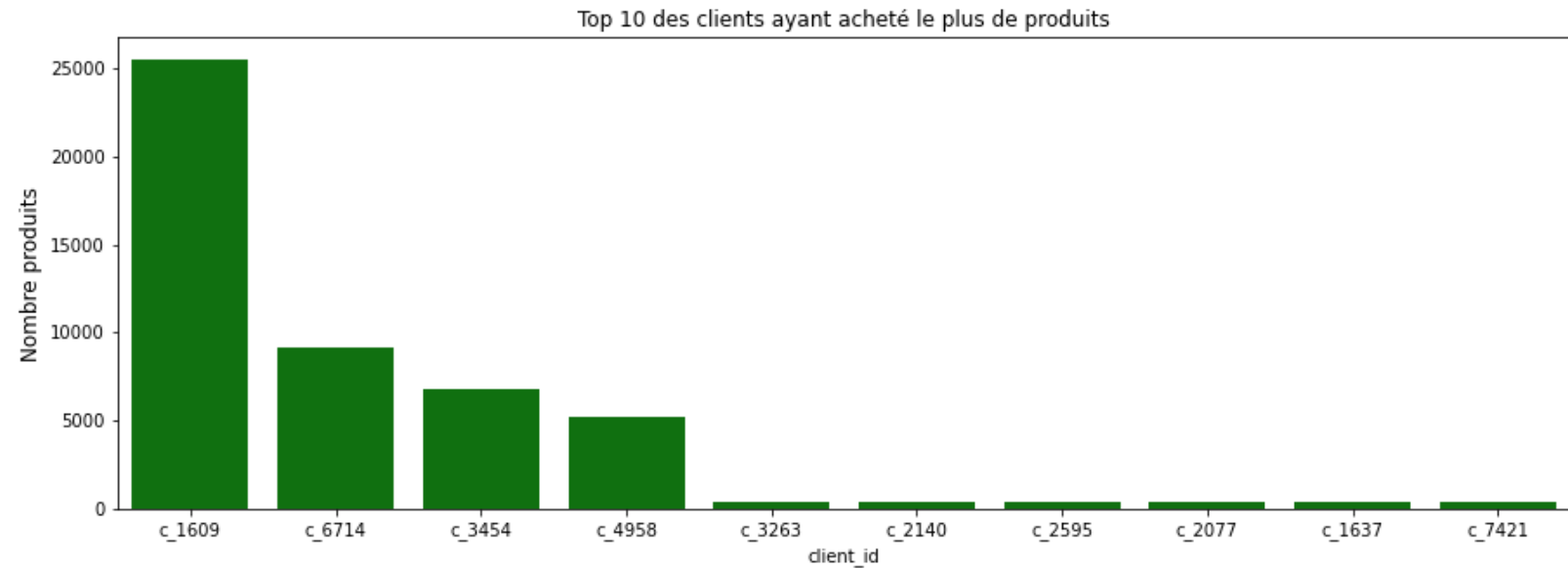


Essentiellement Catégories 0

- Revoir le catalogue des livres de la catégorie 0
- Intensifier l'offre des livres de la catégorie 1

Profil des clients

- **8621** clients avec nombre de femmes > nombre hommes
- Présence de 4 clients professionnels



- Inégalité très marquée du nombre de produits achetés
→ Analyse distincte entre clients professionnels et particuliers

Clients professionnels : analyse globale

4 clients :

Total achats des professionnels : 881 030,54 €

% du CA total : 7,43 %



	C_1609	C_3454	C_4958	C_6714
Age	42	53	23	54
Sexe	M	M	M	F
Cat 0	\$\$\$	\$		\$\$
Cat 1	\$\$\$	\$\$	\$	\$\$
Cat 2	\$	\$	\$\$\$	\$\$
Montant CA	324 033,35€	113 637,93€	289 760,34€	153 598,92€

Analyse clients non professionnels par genre

- **% de produits achetés** : %F > %H
- **% montant des ventes par genre** : %F > %H
- En proportion :
 - les hommes et les femmes achètent quasiment autant de produits
 - les hommes et les femmes génèrent quasiment le même chiffre d'affaires

	Homme	Femme
Fichier client	47,92%	52,08%
% de produits achetés	47,97%	52,03%
%montant des ventes	48%	52%
Nombre transactions par genre /Nombre personnes par genre	73,4	73,3
CA par genre /Nombre personnes par genre	145 266,90€	144 945,80€

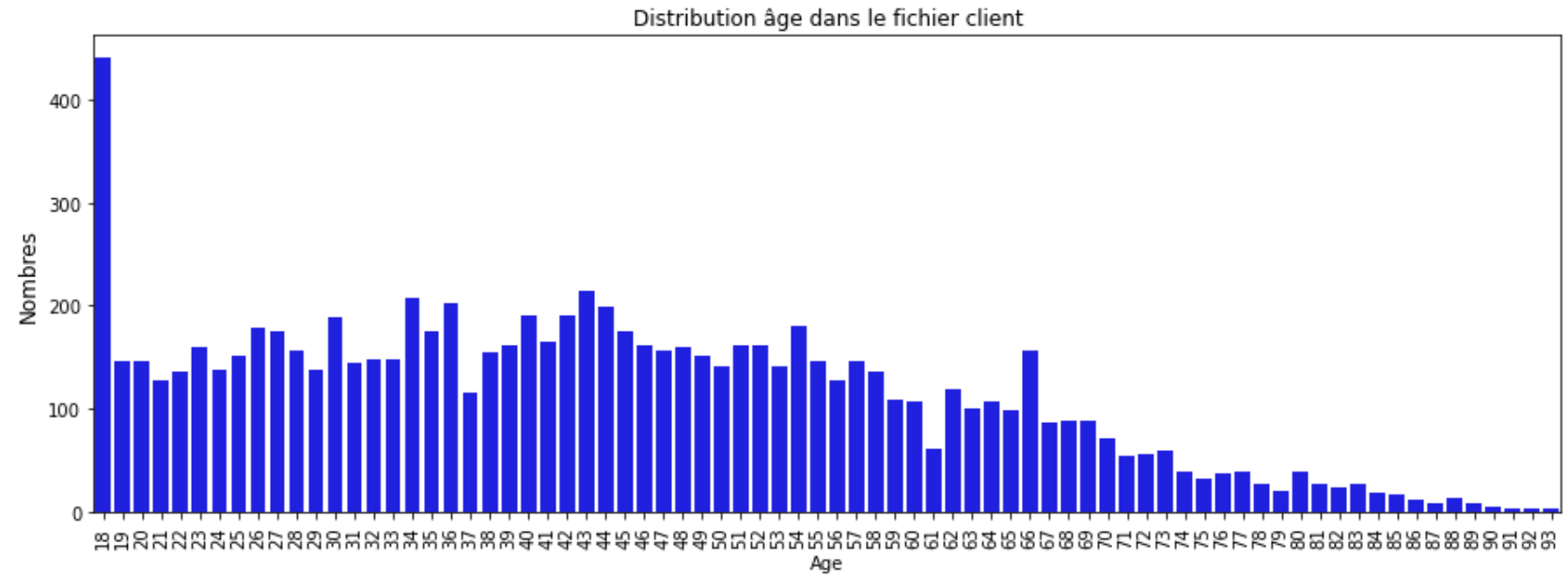


Analyse clients non professionnels par âge

Dans le fichier « client » :

- des clients de 18 à 93 ans
- Age moyen : 43,7 ans
- Age médian : 43 ans

Les clients les + nombreux : les 18 ans

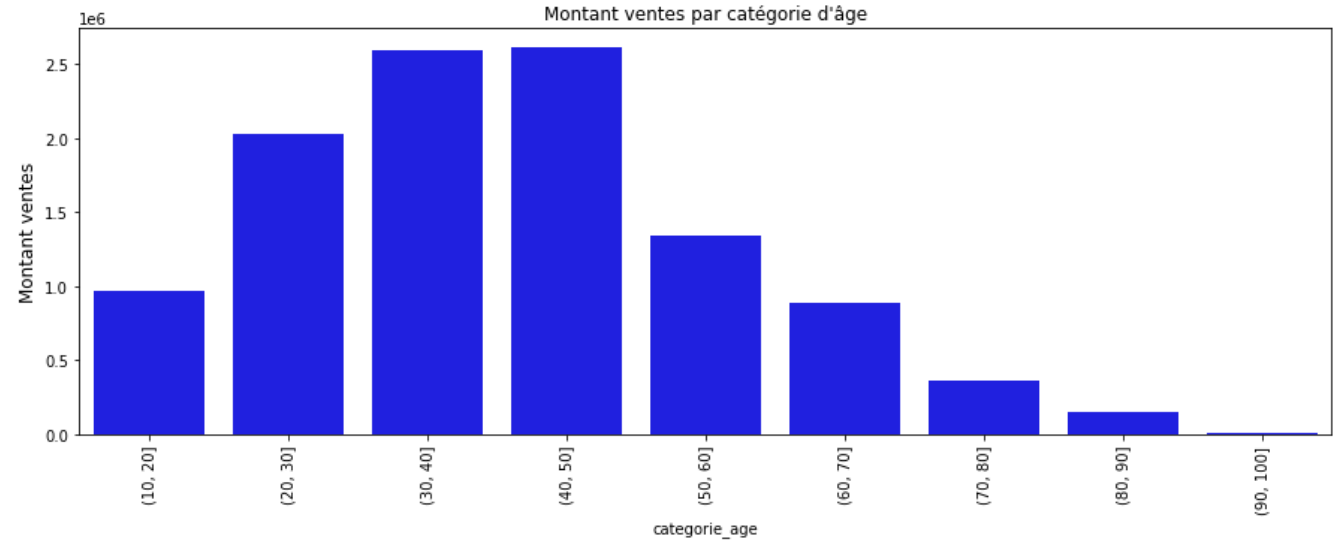


75% des clients ont moins de 56 ans

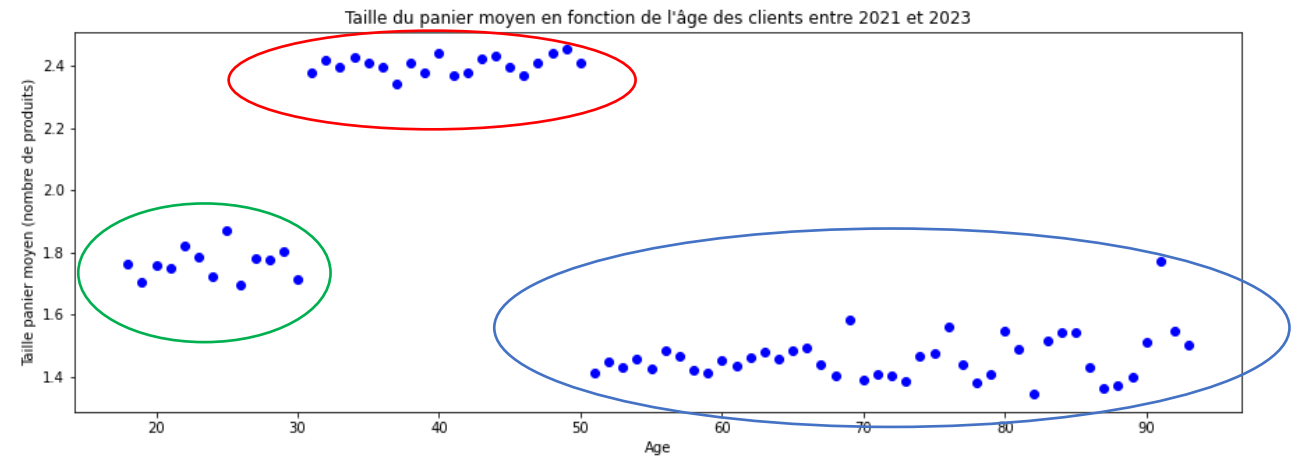
50% des clients ont entre 30 ans et 56 ans

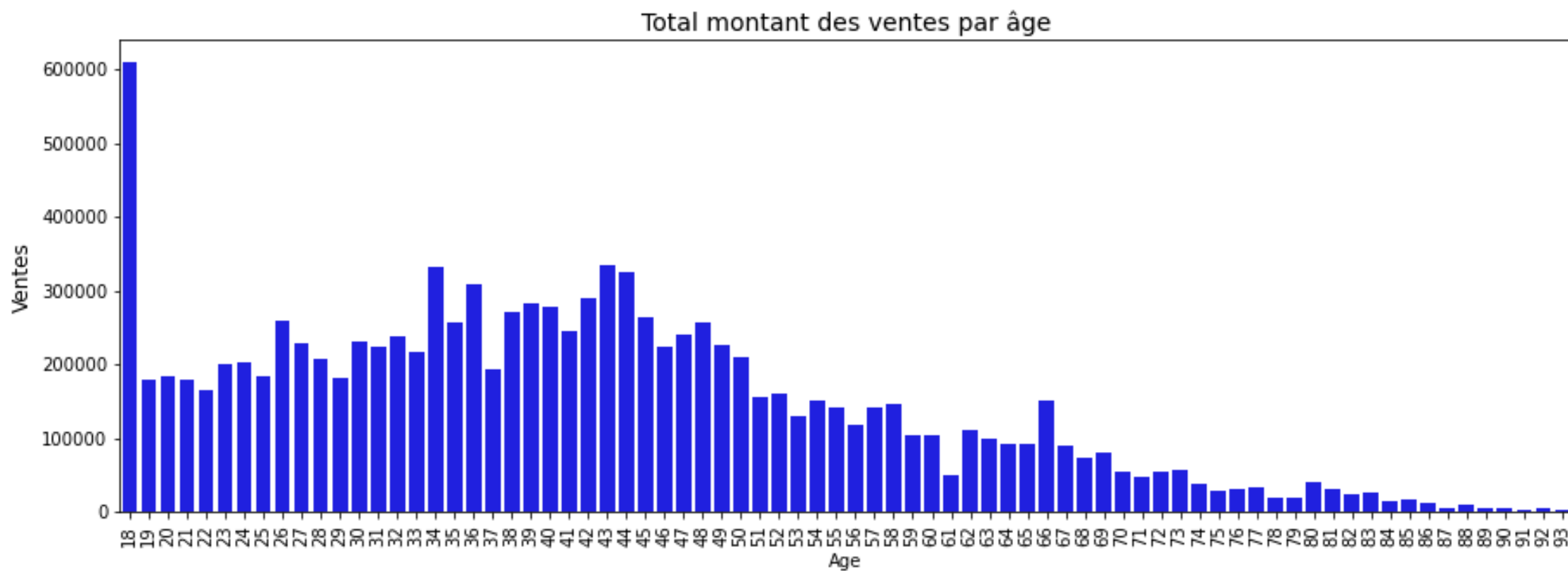
3 groupes de distinguent:

- **les – de 30 ans** : achètent – mais + chers
- **les 30-50 ans** : achètent souvent et beaucoup
- **les + de 50 ans** : achètent mais - chers



	- de 30 ans	de 30 à 50 ans	+ 50 ans
Fréquence d'achats			
Montant des achats	\$\$	\$\$\$	\$
Taille panier moyen			
Les plus gros acheteurs par catégorie	2	0 et 1	





- Les + gros acheteurs : les 18 ans
- Les – gros acheteurs : les 91 et 93 ans

PARTIE 3

Analyse des données

-

Test statistiques



Hypothèses à vérifier

Dépendance entre le genre des clients et la catégorie de livres achetée

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

Dépendance entre l'âge des clients et la fréquence des achats

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

Dépendance entre l'âge des clients et la catégorie d'achat

Corrélation genre client – catégories des livres achetés

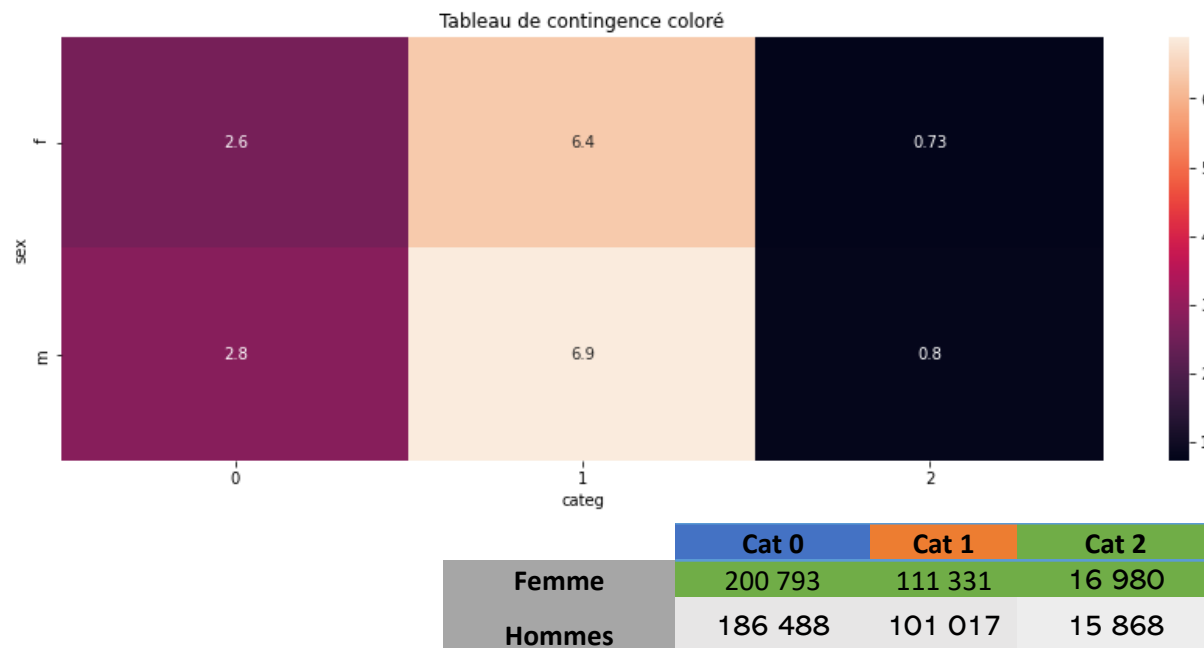
- Test d'indépendance entre 2 **variables qualitatives**
→ Test du Khi2
- **2 degrés de liberté**
(nombre de genres - 1) x (nombre de catégories - 1) = 2
- **Pré-requis du test**
Taille minimale de 20 unités statistiques (ici 679 111)
Tous les effectifs théoriques doivent être ≥ 5 (l'effectif le plus petit est 18173)
- **Risque α : 5%**
- **2 hypothèses :**

H0 (hypothèse nulle) : les variables 'sex' et 'categ' sont indépendantes

H1 (hypothèse alternative) : les variables 'sex' et 'categ' sont dépendantes

- **Résultat du test** : $\chi^2 = 20.22 > 5.99 = \text{seuil critique}$ et $p\text{-value} = 4.08 \times 10^{-5} < 0.01$

Test statistique significatif : Rejet hypothèse H0 : les variables catégorie d'achat et genre sont probablement dépendantes l'une de l'autre.



Femmes : celles qui achètent le + en nombre dans toutes les catégories même si la différence n'est pas très marquée

Corrélation âge client – montant total des achats

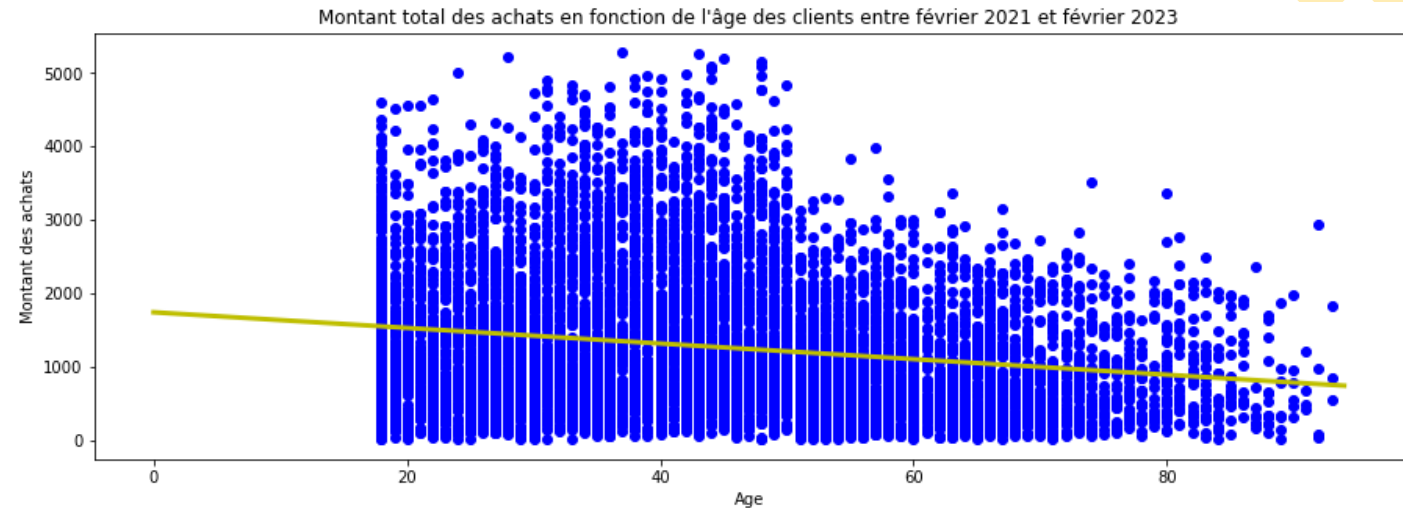
- Test d'indépendance entre **2 variables quantitatives**
 → **Test de régression linéaire** : détermination du **coefficient de corrélation linéaire de Pearson** pour analyse de l'intensité du lien
- Risque α : 5%
- 2 hypothèses :
H0 (hypothèse nulle) : Variables indépendantes si $p\text{-value} > 5\%$
H1 (hypothèse alternative) : Variables non indépendantes si $p\text{-value} < 5\%$

→ Coefficient de corrélation linéaire de Pearson = -0.188
 $R^2(\text{coefficient de détermination}) = 0.035$
 $p\text{-value} = 1.7101053363389715e-69$

Corrélation linéaire négative entre l'âge du client et le montant des achats et $p\text{-value} < 5\%$



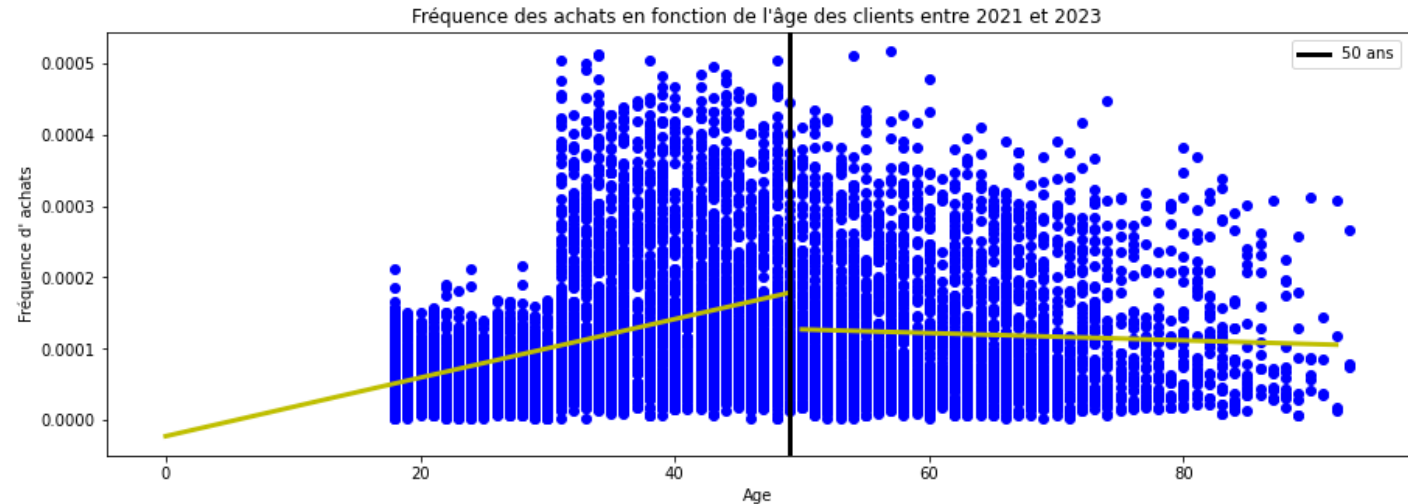
Rejet H0 : les 2 variables sont probablement dépendantes l'une de l'autre



Au + le client est âgé, au plus le montant de ses achats est faible
Au - le client est âgé, au plus le montant de ses achats est important

Corrélation âge client – fréquence des achats

- Test d'indépendance entre **2 variables quantitatives**
 → **Test de régression linéaire** : détermination du **coefficient de corrélation linéaire de Pearson** pour analyse de l'intensité du lien
- Risque α : 5%
- 2 hypothèses :
H0 (hypothèse nulle) : Variables indépendantes si $p\text{-value} > 5\%$
H1 (hypothèse alternative) : Variables non indépendantes si $p\text{-value} < 5\%$
- Cas âge < 50 ans
 → Coefficient de corrélation linéaire de Pearson = 0.396
 $R^2(\text{coefficient de détermination}) = 0.157$
 $p\text{-value} = 2.787281947606434e-205$
- Cas âge > 50 ans
 → Coefficient de corrélation linéaire de Pearson = -0.052
 $R^2(\text{coefficient de détermination}) = 0.003$
 $p\text{-value} = 0.0035491062359061065$



Corrélation linéaire positive entre l'âge du client et le montant des achats et $p\text{-value} < 5\%$: **les 2 variables sont probablement dépendantes l'une de l'autre**
 Au + l'âge augmente, au + la fréquence d'achat augmente

Corrélation linéaire négative entre l'âge du client et le montant des achats et $p\text{-value} < 5\%$
les 2 variables sont probablement dépendantes l'une de l'autre
 Au + l'âge augmente, au + la fréquence d'achat diminue

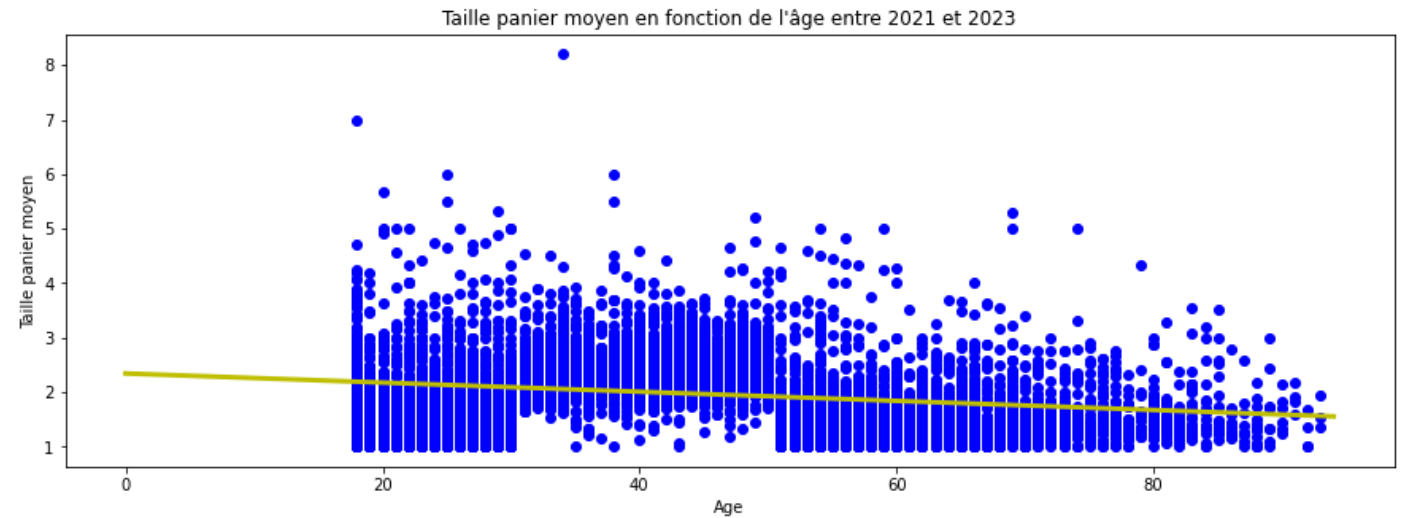
Corrélation âge client – taille panier moyen

- Test d'indépendance entre **2 variables quantitatives**
 → **Test de régression linéaire** : détermination du **coefficient de corrélation linéaire de Pearson** pour analyse de l'intensité du lien
- 2 hypothèses :
H0 (hypothèse nulle) : Variables indépendantes si $p\text{-value} > 5\%$
H1 (hypothèse alternative) : Variables non indépendantes si $p\text{-value} < 5\%$
- **Coefficient de corrélation linéaire de Pearson** = -0.213
 $R^2(\text{coefficient de détermination}) = 0.045$
 $p\text{-value} = 1.2372377168355913e-88$

Corrélation linéaire négative entre l'âge du client et le montant des achats et $p\text{-value} < 5\%$



Rejet de H0 : les 2 variables sont probablement dépendantes l'une de l'autre.
Au + l'âge augmente, au + la taille du panier moyen diminue



Corrélation âge client – catégorie d'achat

- Test d'indépendance entre 1 variable quantitative (âge) et 1 variable qualitative (catégorie)
→ ANOVA à 1 facteur (analyse de la variance) basée sur le test de Fisher
- Risque α : 5%
- 2 hypothèses :
H0 (hypothèse nulle) :
Les 3 moyennes entre les groupes sont égales et probablement pas de

H1(hypothèse alternative) :

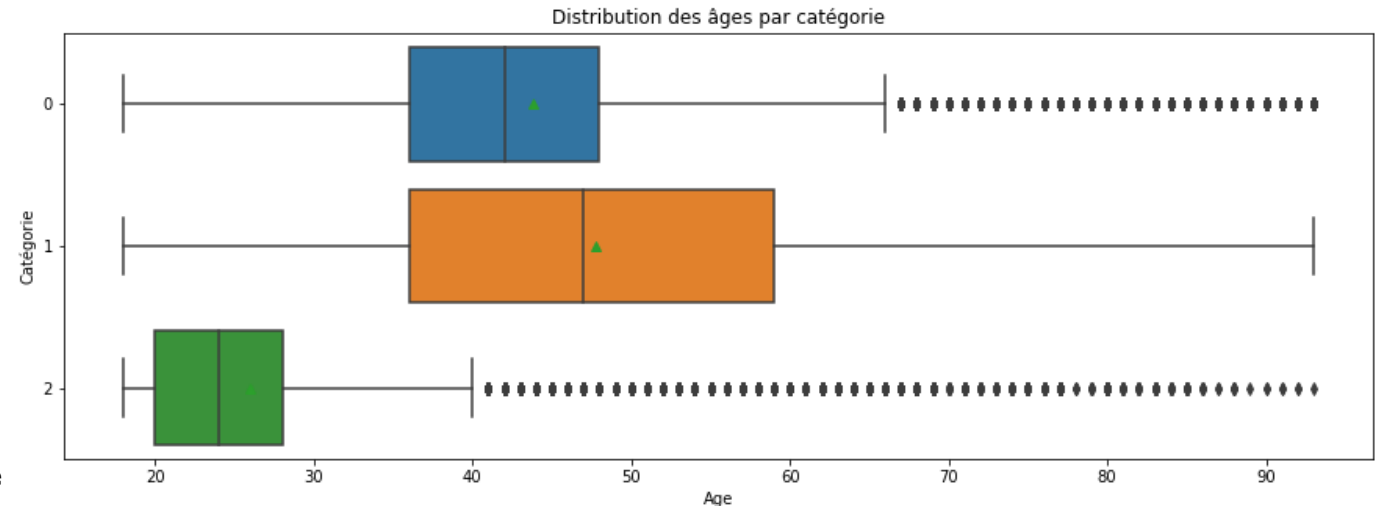
Les 3 moyennes entre les groupes sont différentes et probablement corrélation

- → Résultat ANOVA / test de Fisher :

p_value=0 et p_value < 0,05



Rejet hypothèse H0 : le choix de la catégorie est probablement influencé par l'âge



Vérification des conditions de l'application de l'ANOVA

- **Indépendance entre les échantillons** : donnée par le contexte de l'étude, il y a 3 catégories distinctes

1ère condition validée

- **Egalité des variances** avec le test de Bartlett (étude des variances entre les 3 groupes)
avec comme hypothèses de départ :

- H_0 : Les variances de chaque groupe sont égales si $p\text{-value} > 5\%$
- H_1 : Les variances de chaque groupe ne sont pas toutes égales $< 5\%$

→ la $p\text{-value} < 5\%$: rejet de H_0 , les variances ne sont pas toutes égales

2ème condition non validée

- **Test de la normalité des résidus** : Test de Shapiro-Wilk inefficace car effectif échantillon > 5000 donc test Kolmogorov-Smirnov avec

- H_0 : Les résidus suivent une loi normale si $p\text{-value} > 5\%$
- H_1 : Les résidus ne suivent pas une loi normale si $p\text{-value} < 5\%$

→ la $p\text{-value} < 5\%$: rejet de H_0 , les résidus ne suivent pas une loi normale, 3ème condition non validée.



Utilisation test non paramétrique : « Kruskal-Wallis »

Utilisation test non paramétrique : « Kruskal-Wallis »

Ce test travaille sur les rangs des valeurs.

Hypothèses de départ :

- H_0 : même rang moyen pour chaque groupe si $p\text{-value} > 5\%$
- H_1 : au moins 2 groupes ont des rangs moyens différents si $p\text{-value} < 5\%$

Après calcul, on obtient $p\text{-value}=0 < 5\%$



Rejet de H_0 : au moins une des 3 catégories diffère des autres, sa distribution n'est pas égale.

On en déduit donc que le choix de la catégorie est probablement influencé par l'âge.



Recommandations

- **Anomalie de données en octobre 2021** : vérifier l'origine de l'anomalie (absence de ventes, de stocks, problèmes techniques) afin de confirmer les % de progression du CA (global et par catégorie)
- **Fluctuation des ventes (dont baisse à partir de février 2023)** : analyse à coupler avec la base des stocks et de l'historique des prix pratiqués
- **Données manquantes dans la base client** : date d'inscription des clients et type profil client (b2b ou b2c) pour une analyse des données plus performante
- **Modification du catalogue des produits pour les catégories 1 et 2** : ils représentent un fort taux du chiffre d'affaires en dépit de leur faible représentation en catalogue. A coupler avec les tops des ventes en magasin physique ? Questionnaire sur les habitudes de lecture ?
- **3 groupes de clients distincts** : programme de fidélisation spécifique à développer ?
- **Promotions des catégories de produits** en fonction de la période de l'année