RESTER LIVRES

Analyse des ventes d’une librairie en ligne

PARTIE 1
Nettoyage et préparation

Source : Base de données des ventes

Clients

	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

Produits

	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

Transactions

	id_prod	date	session_id	client_id
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450
1	2_226	2022-02-03 01:55:53.276402	s_159142	c_277
2	1_374	2021-09-23 15:13:46.938559	s_94290	c_4270
3	0_2186	2021-10-17 03:27:18.783634	s_105936	c_4597
4	0_1351	2021-07-17 20:34:25.800563	s_63642	c_1242

Source : Résumé

Clients

```
Année de naissance du client le plus vieux: 1929
Année de naissance du client le plus jeune: 2004

Clients sans genre défini:
```

client_id	sex	birth

```
Lignes avec des valeurs vides:
Aucune ligne n'a de valeur vide.
```

client_id	sex	birth

Produits

```
Nombre de ligne: 3287
Prix le plus bas: -1.0
Prix le plus haut: 300.0

Lignes avec des valeurs vides:
Aucune ligne n'a de valeur vide.
```

id_prod	price	categ

```
Prix inférieurs à 0 :
```

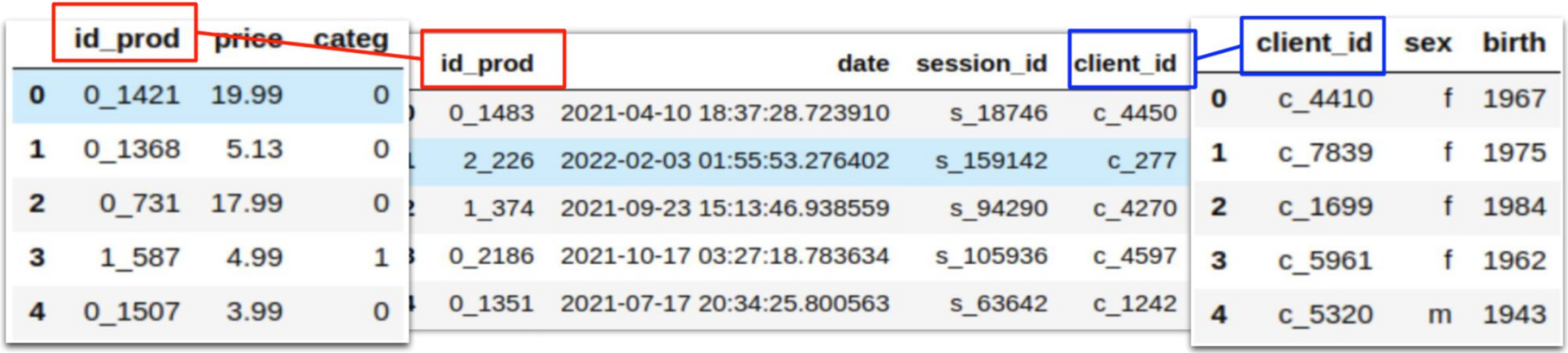
	id_prod	price	categ
731	T_0	-1.0	0

Transactions

```
Date la plus ancienne: 2021-03-01 00:02:26.047414
Date la plus récente: test_2021-03-01 02:30:02.237450


Dates érronées :
200 lignes

Sessions ID correspondant aux dates érronées:
['s_0']

Client ID correspondants aux dates érronées:
['ct_1' 'ct_0']

ID Produit correspondant aux dates érronées:
['T_0']

Lignes avec des valeurs vides:

Aucune ligne n'a de valeur vide.
```

id_prod	date	session_id	client_id

JOINTURE



	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

	id_prod	date	session_id	client_id
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450
1	2_226	2022-02-03 01:55:53.276402	s_159142	c_277
2	1_374	2021-09-23 15:13:46.938559	s_94290	c_4270
3	0_2186	2021-10-17 03:27:18.783634	s_105936	c_4597
4	0_1351	2021-07-17 20:34:25.800563	s_63642	c_1242

	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

Nettoyage : suppression et ajout

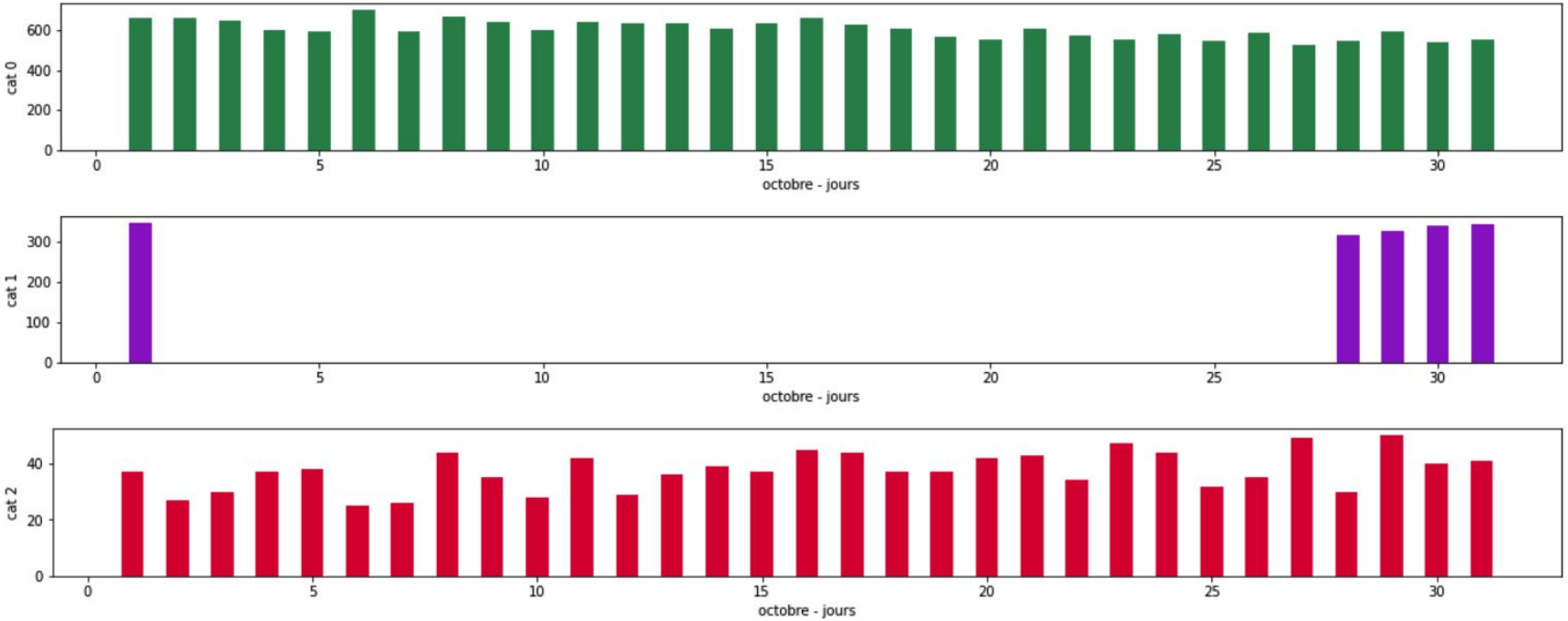
Suppression des tests

	c_id	c_sex	c_birth	p_id	p_price	p_cat	t_date	t_sess_id
108206	ct_0	f	2001.0	T_0	-1.0	0.0	test_2021-03-01 02:30:02.237446	s_0
108284	ct_0	f	2001.0	T_0	-1.0	0.0	test_2021-03-01 02:30:02.237419	s_0

Produit *0_2245* manquant

	c_id	c_sex	c_birth	p_id	p_price	p_cat	t_date	t_sess_id
2678	c_4505	m	1976.0	0_2245	**10.64**	**0**	2022-01-09 09:23:31.000720	s_147220

Nettoyage : suppression et ajout



Majorité des ventes catégorie 1 du mois d'Octobre manquantes
= suppression de toutes les ventes catégorie 1 du mois d'Octobre

Préparation : Discrétisation des âges

	client_id	client_sex	client_birth	product_id	product_price	product_cat	transaction_date	transaction_sess_id	client_age	client_tranche_age
0	c_4410	f	1967.0	1_385	25.99	1	2021-03-22 01:40:22.782925	s_9707	54.0	45-59
1	c_4410	f	1967.0	0_1110	4.71	0	2021-11-04 16:28:30.169021	s_114715	54.0	45-59
2	c_4410	f	1967.0	0_1111	19.99	0	2021-03-22 01:27:49.480137	s_9707	54.0	45-59
3	c_4410	f	1967.0	1_461	12.99	1	2021-08-11	s_74236	54.0	45-59

J'ai utilisé la méthode des amplitudes égales pour définir les 4 tranches d'âge suivantes:

-30
30 - 44
45 - 59
60+

Préparation : Discrétisation des prix

	client_id	client_sex	client_birth	product_id	product_price	product_cat	product_tranche_prix
0	c_4410	f	1967.0	1_385	25.99	1	+19€
1	c_4410	f	1967.0	0_1110	4.71	0	-9€
2	c_4410	f	1967.0	0_1111	19.99	0	+19€
3	c_4410	f	1967.0	1_461	12.99	1	9€-14€
4	c_4410	f	1967.0	1_536	11.21	1	9€-14€
5	c_4410	f	1967.0	1_190	14.53	1	14€-19€
6	c_4410	f	1967.0	0_1334	17.74	0	14€-19€
7	c_4410	f	1967.0	1_616	29.02	1	+19€
8	c_4410	f	1967.0	1_558	24.51	1	+19€

J'ai utilisé la méthode des quantiles pour définir les 4 tranches de prix suivantes:

-9€
9€ - 14€
14€ - 19€
+19€

Préparation : Éclatement des dates

	client_id	transaction_date	transaction_sess_id	transaction_year	transaction_month	transaction_month_part	transaction_weekday	transaction_hour	transaction_period_month
0	c_4410	2021-03-22 01:40:22.782925	s_9707	2021.0	3.0	2	0.0	1.0	3.0
1	c_4410	2021-11-04 16:28:30.169021	s_114715	2021.0	11.0	0	3.0	16.0	11.0
2	c_4410	2021-03-22 01:27:49.480137	s_9707	2021.0	3.0	2	0.0	1.0	3.0
3	c_4410	2021-08-11 08:40:47.495793	s_74236	2021.0	8.0	1	2.0	8.0	8.0
4	c_4410	2022-01-18 17:05:07.468131	s_151740	2022.0	1.0	2	1.0	17.0	13.0
5	c_4410	2021-11-12 18:11:43.280574	s_118628	2021.0	11.0	1	4.0	18.0	11.0
6	c_4410	2021-09-25 00:17:38.676453	s_94984	2021.0	9.0	2	5.0	0.0	9.0
7	c_4410	2021-12-01 07:31:51.359660	s_127714	2021.0	12.0	0	2.0	7.0	12.0
8	c_4410	2021-09-25	s_94984	2021.0	9.0	2	5.0	0.0	9.0

Exportation du dataframe

export_nettoyage.csv

```
client_id,client_sex,client_birth,product_id,product_price,product_cat,transaction_date,transaction_sess_id,
client_age,client_tranche_age,product_tranche_prix,transaction_year,transaction_month,transaction_month_part
,transaction_weekday,transaction_hour,transaction_period_month,Gros_client
c_4410,f,1967.0,1_385,25.99,1,2021-03-22 01:40:22.782925,s_9707,54.0,45-59,+19€,2021.0,3.0,2,0.0,1.0,3.0,No
c_4410,f,1967.0,0_1110,4.71,0,2021-11-04
16:28:30.169021,s_114715,54.0,45-59,-9€,2021.0,11.0,0,3.0,16.0,11.0,No
c_4410,f,1967.0,0_1111,19.99,0,2021-03-22 01:27:49.480137,s_9707,54.0,45-59,+19€,2021.0,3.0,2,0.0,1.0,3.0,No
c_4410,f,1967.0,1_461,12.99,1,2021-08-11
08:40:47.495793,s_74236,54.0,45-59,9€-14€,2021.0,8.0,1,2.0,8.0,8.0,No
c_4410,f,1967.0,1_536,11.21,1,2022-01-18
17:05:07.468131,s_151740,55.0,45-59,9€-14€,2022.0,1.0,2,1.0,17.0,13.0,No
c_4410,f,1967.0,1_190,14.53,1,2021-11-12
18:11:43.280574,s_118628,54.0,45-59,14€-19€,2021.0,11.0,1,4.0,18.0,11.0,No
c_4410,f,1967.0,0_1334,17.74,0,2021-09-25
00:17:38.676453,s_94984,54.0,45-59,14€-19€,2021.0,9.0,2,5.0,0.0,9.0,No
c_4410,f,1967.0,1_616,29.02,1,2021-12-01
07:31:51.359660,s_127714,54.0,45-59,+19€,2021.0,12.0,0,2.0,7.0,12.0,No
c_4410,f,1967.0,1_558,24.51,1,2021-09-25 00:11:19.292740,s_94984,54.0,45-59,+19€,2021.0,9.0,2,5.0,0.0,9.0,No
c_4410,f,1967.0,0_1376,16.24,0,2021-09-24
22:58:27.418343,s_94984,54.0,45-59,14€-19€,2021.0,9.0,2,4.0,22.0,9.0,No
c_4410,f,1967.0,0_1054,8.11,0,2021-10-01
13:16:27.958827,s_98432,54.0,45-59,-9€,2021.0,10.0,0,4.0,13.0,10.0,No
c_4410,f,1967.0,0_1455,8.99,0,2021-03-22 14:29:25.189266,s_9942,54.0,45-59,-9€,2021.0,3.0,2,0.0,14.0,3.0,No
c_4410,f,1967.0,1_653,25.99,1,2021-07-29
23:34:41.866951,s_68860,54.0,45-59,+19€,2021.0,7.0,3,3.0,23.0,7.0,No
c_4410,f,1967.0,1_91,20.99,1,2021-05-28 05:29:10.024293,s_40563,54.0,45-59,+19€,2021.0,5.0,3,4.0,5.0,5.0,No
c_4410,f,1967.0,0_521,23.99,0,2021-03-23
17:11:46.158290,s_10454,54.0,45-59,+19€,2021.0,3.0,2,1.0,17.0,3.0,No
c_4410,f,1967.0,1_407,15.99,1,2021-03-24
18:30:13.156928,s_10922,54.0,45-59,14€-19€,2021.0,3.0,2,2.0,18.0,3.0,No
c_4410,f,1967.0,0_1316,7.2,0,2021-12-29
09:11:18.860592,s_141762,54.0,45-59,-9€,2021.0,12.0,3,2.0,9.0,12.0,No
c_4410,f,1967.0,1_395,28.99,1,2021-09-24
23:57:35.138518,s_94984,54.0,45-59,+19€,2021.0,9.0,2,4.0,23.0,9.0,No
c_4410,f,1967.0,1_267,27.99,1,2021-09-24
23:15:59.919591,s_94984,54.0,45-59,+19€,2021.0,9.0,2,4.0,23.0,9.0,No
c_4410,f,1967.0,0_1977,5.99,0,2021-10-21
05:46:33.024343,s_107888,54.0,45-59,-9€,2021.0,10.0,2,3.0,5.0,10.0,No
c_4410,f,1967.0,0_1277,7.99,0,2021-09-25 00:03:39.156997,s_94984,54.0,45-59,-9€,2021.0,9.0,2,5.0,0.0,9.0,No
c_4410,f,1967.0,1_483,15.99,1,2021-03-13
21:35:55.949042,s_5913,54.0,45-59,14€-19€,2021.0,3.0,1,5.0,21.0,3.0,No
c_4410,f,1967.0,0_1121,19.99,0,2021-12-24
08:01:00.473579,s_139182,54.0,45-59,+19€,2021.0,12.0,2,4.0,8.0,12.0,No
c_4410,f,1967.0,1_312,24.56,1,2022-01-29
14:07:47.482092,s_156960,55.0,45-59,+19€,2022.0,1.0,3,5.0,14.0,13.0,No
c_4410,f,1967.0,0_1426,13.44,0,2021-08-04
10:02:26.066531,s_71224,54.0,45-59,9€-14€,2021.0,8.0,0,2.0,10.0,8.0,No
c_4410,f,1967.0,1_584,9.73,1,2021-05-28
05:19:16.367701,s_40563,54.0,45-59,9€-14€,2021.0,5.0,3,4.0,5.0,5.0,No
c_4410,f,1967.0,0_1612,13.54,0,2021-12-05
07:51:07.012251,s_129672,54.0,45-59,9€-14€,2021.0,12.0,0,6.0,7.0,12.0,No
c_4410,f,1967.0,1_621,17.99,1,2021-12-01
07:36:23.671688,s_127714,54.0,45-59,14€-19€,2021.0,12.0,0,2.0,7.0,12.0,No
c_4410,f,1967.0,0_1420,11.53,0,2021-03-22
22:31:25.825764,s_10092,54.0,45-59,9€-14€,2021.0,3.0,2,0.0,22.0,3.0,No
c_4410,f,1967.0,1_15,16.99,1,2021-08-22
02:23:56.577463,s_78992,54.0,45-59,14€-19€,2021.0,8.0,2,6.0,2.0,8.0,No
c_4410,f,1967.0,1_436,11.76,1,2022-01-04
```

PARTIE 2
Analyses

Sur la période

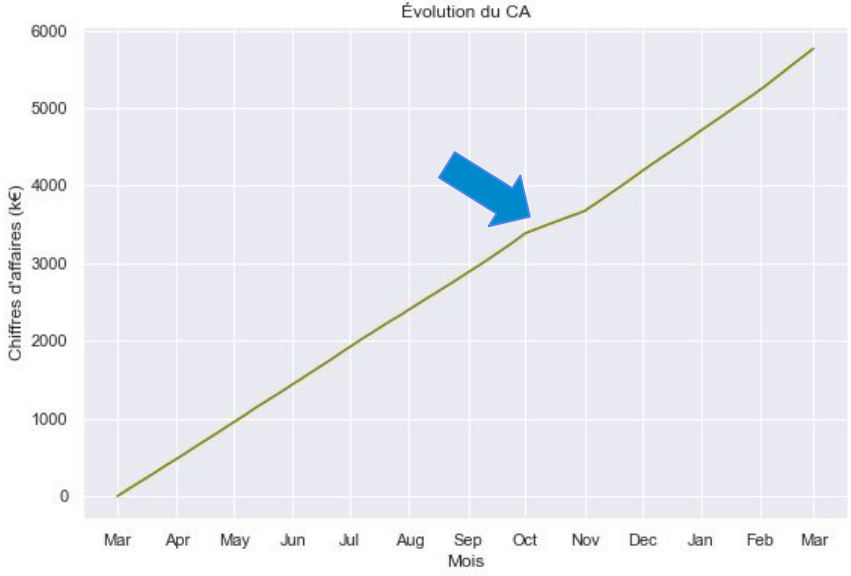
CA

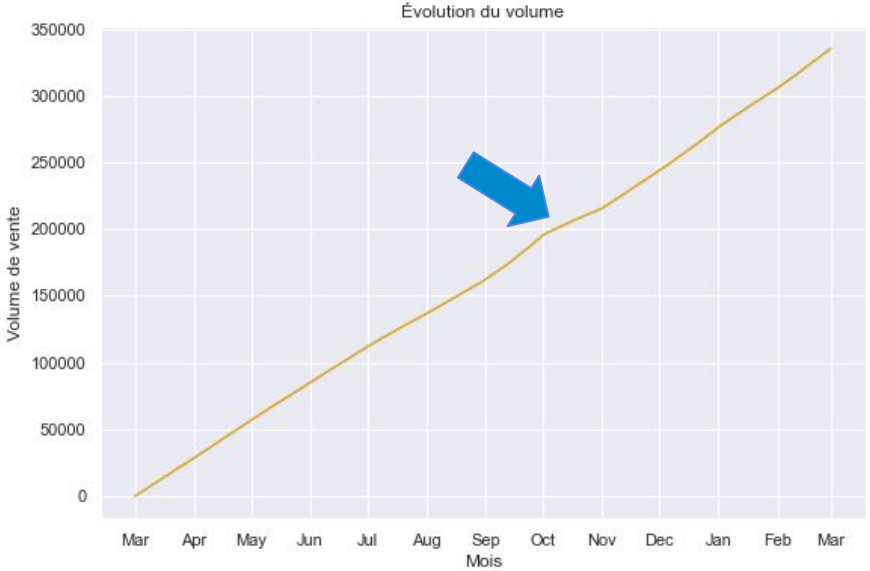
5,7 millions €

articles vendus

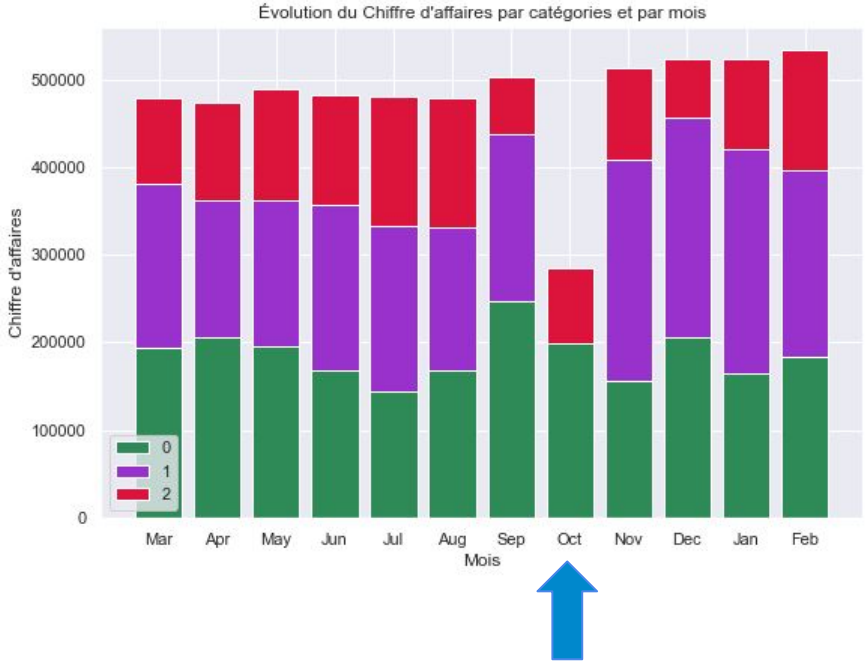
335 150

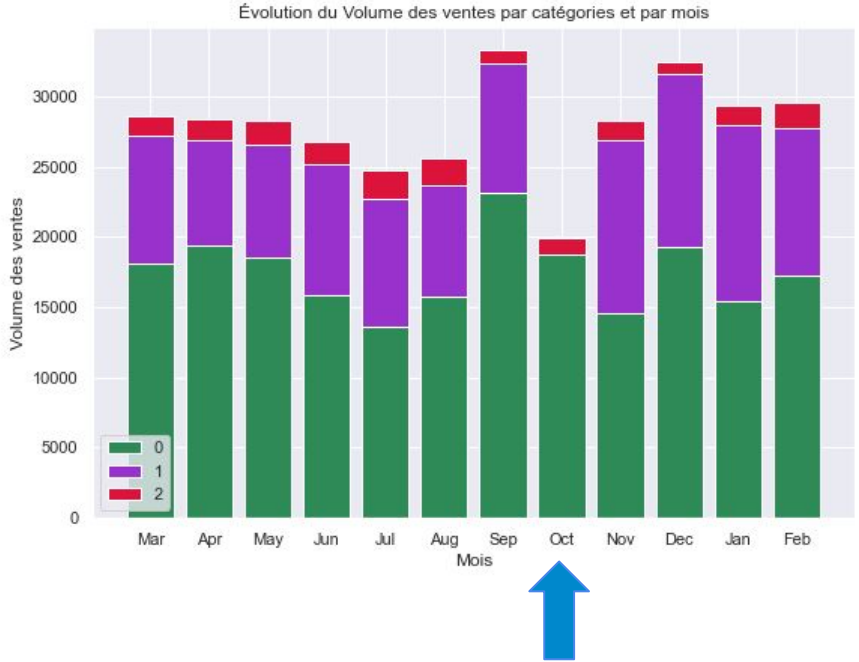
Analyse : CA et volume - vue globale



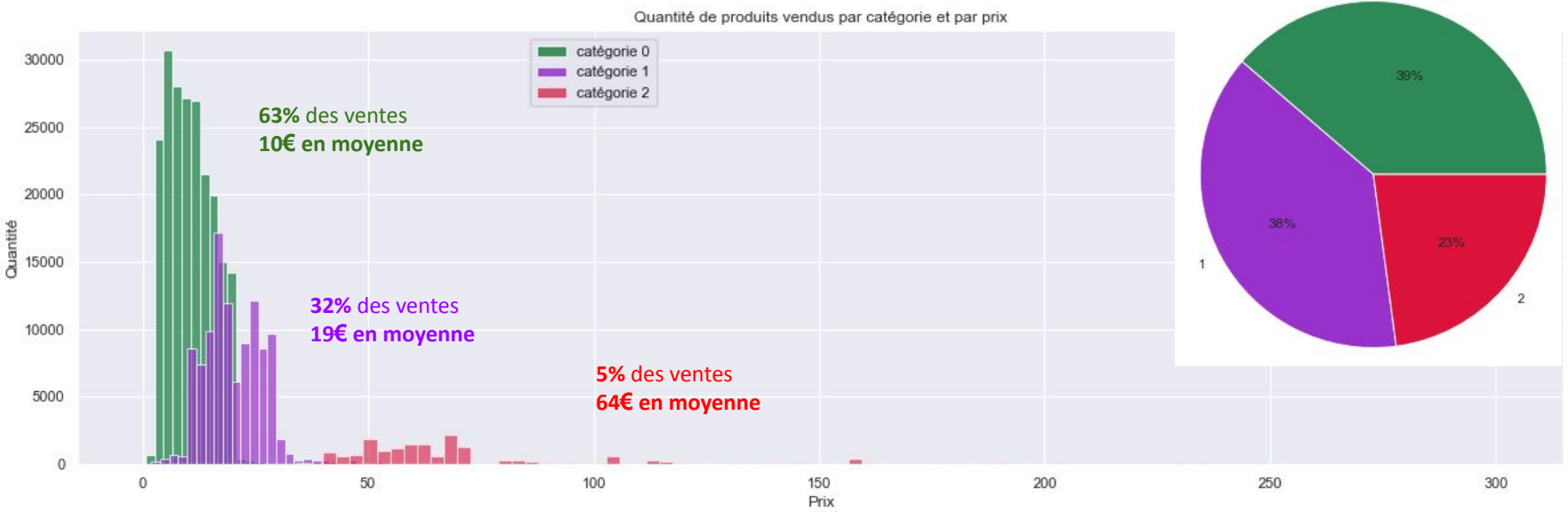


Analyse : CA et volume - par mois et catégories

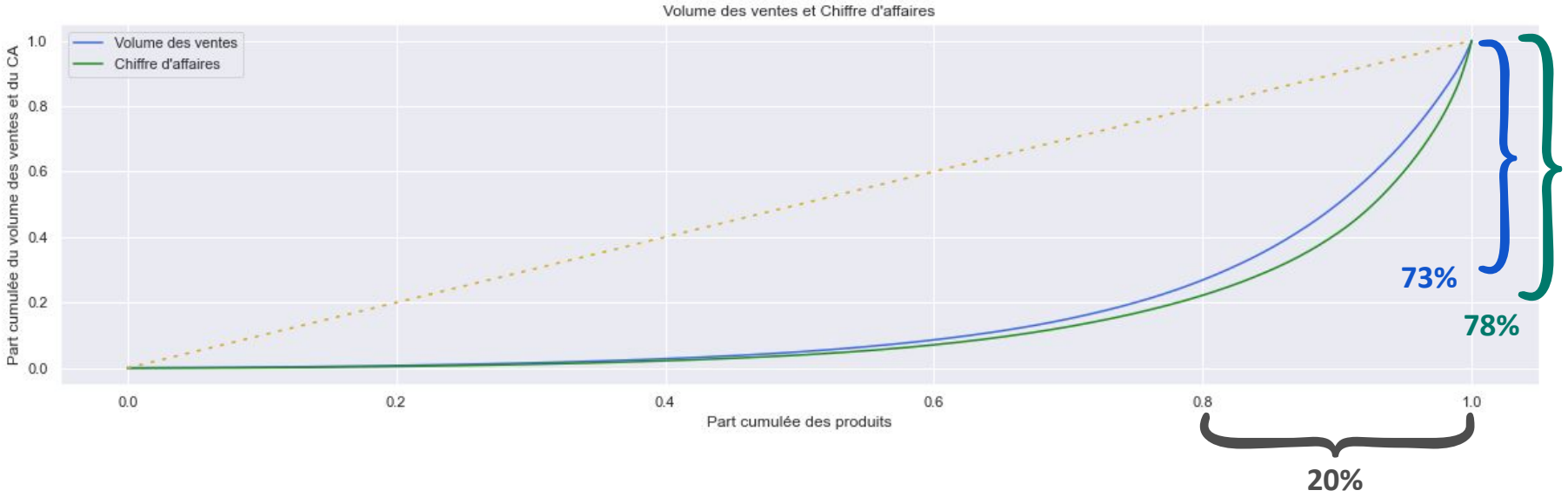




Analyse : distribution des prix par catégories



Analyse : concentration des ventes



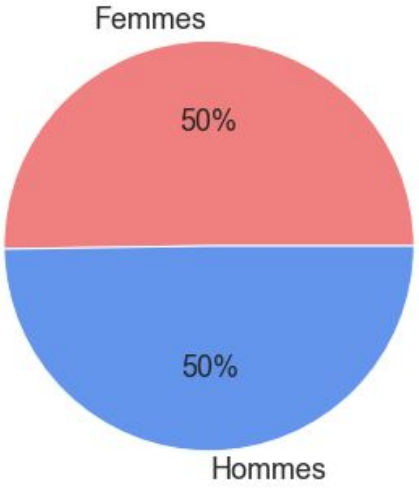
indice de gini sur chiffre d'affaires : **0.74**
indice de gini sur volume des ventes : **0.69**

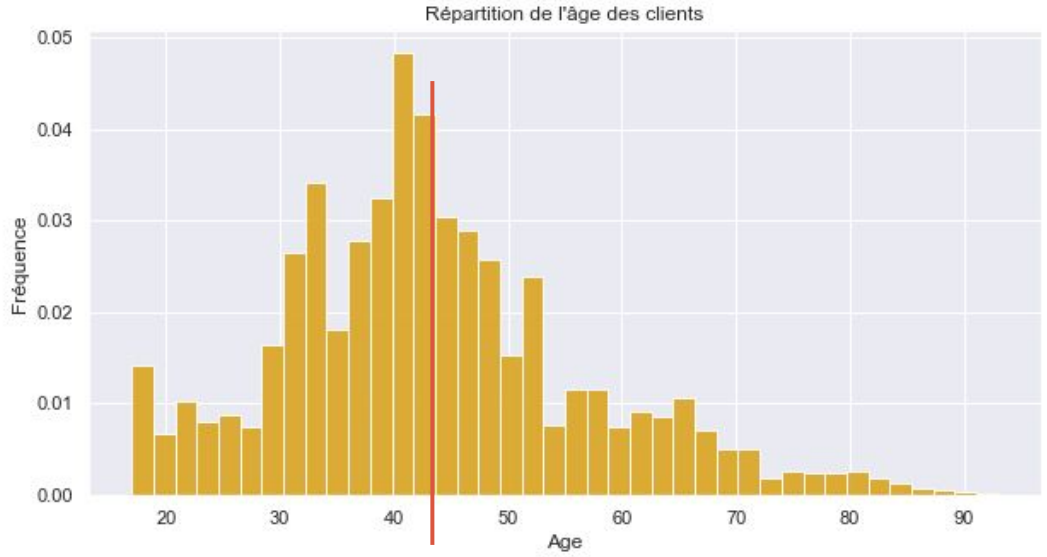
Analyse : Panier moyen

Panier moyen global : 34€

Entre 1 et 3 articles par panier

Analyse : Répartition des clients - genre et âge





âge moyen : 43,3 ans

Proche de la population française

PARTIE 3
Corrélations

Corrélation entre le sexe et la catégorie des produits achetés ?

TEST DU χ^2

H0 : Les deux variables sont indépendantes, elles ne sont pas corrélées.

H1 : Les deux variables ne sont pas indépendantes, elles sont corrélées.

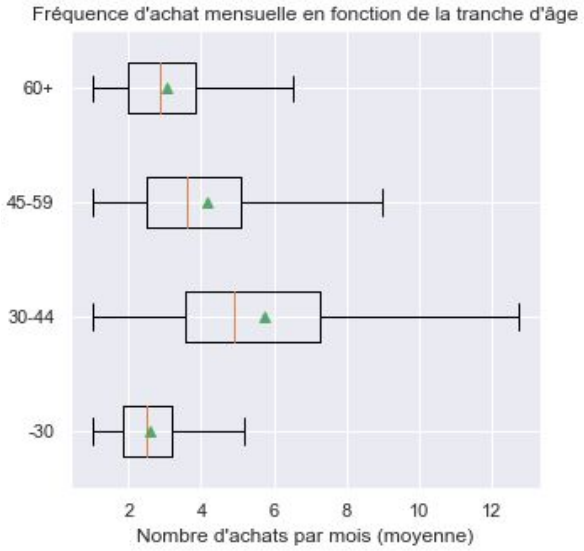
Le χ2 est de 11.41
La P-value est de 0.003
Le degré de liberté est de 2

La comparaison entre notre χ2 calculé et le χ2 théorique nous permet de trancher entre les deux hypothèses.

Pour α=5% : Le sexe des clients a une incidence sur la catégorie des produits achetés. C'est l'hypothèse alternative (H1) qui l'emporte.

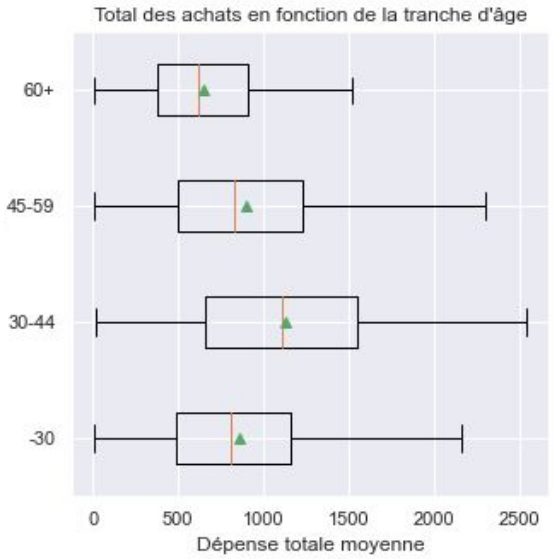
Corrélation entre l'âge du client et ...

la fréquence d'achat



$\eta^2 = 0.25$

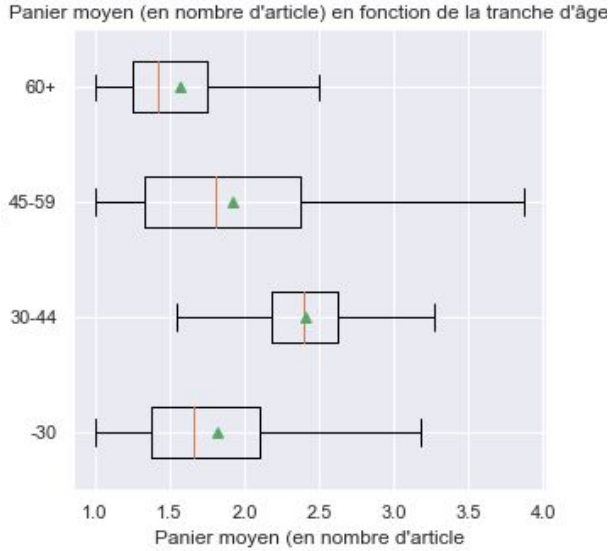
le montant total des achats



$\eta^2 = 0.10$

le panier moyen

(en nombre d'article)



$\eta^2 = 0.22$

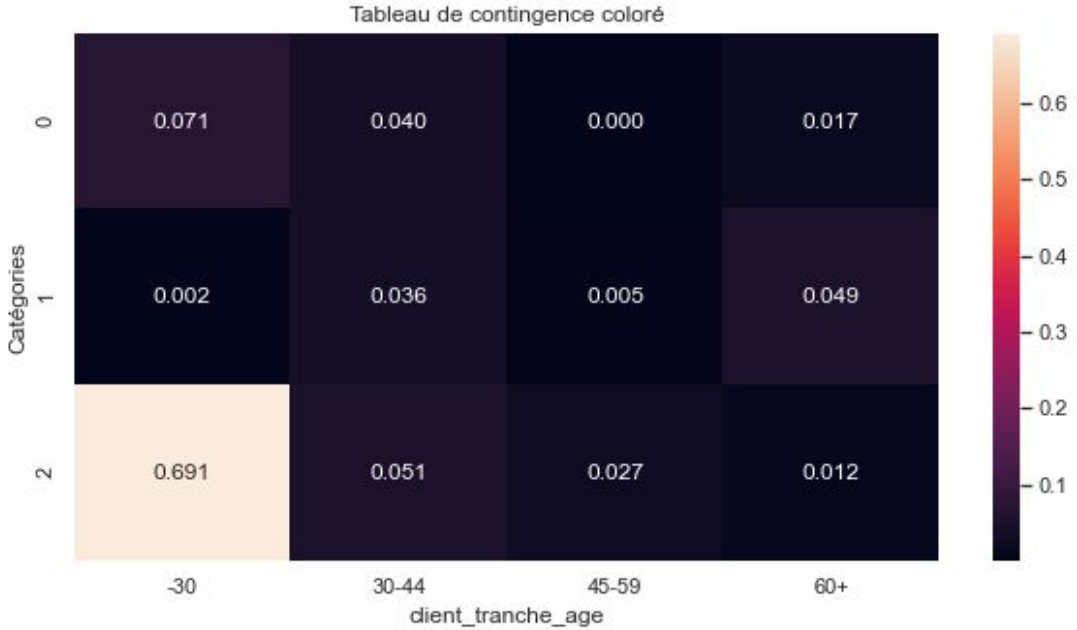
Dans les trois corrélations testées, η^2 est supérieur à 0.
Cela signifie que les moyennes des variables testées par tranche d'âge sont différentes: il existe donc à priori une relation entre les variables testées et l'âge des clients

Corrélation entre l’âge des clients et la catégorie des produits achetés?

TEST DU χ^2

H0 : Les deux variables sont indépendantes, elles ne sont pas corrélées.

H1 : Les deux variables ne sont pas indépendantes, elles sont corrélées.



Le χ2 est de 125228.98
La P-value est de 0.0
Le degré de liberté est de 6

La comparaison entre notre χ2 calculé et le χ2 théorique nous permet de trancher entre les deux hypothèses.

Pour α=5% : L’âge des clients a une incidence sur la catégorie des produits achetés. C'est l'hypothèse alternative H1 qui l'emporte.

Merci de votre attention

Avez-vous des questions ?