Rester livres

Analyse des ventes

ETUDE D'INTEGRATION

Frédéric Gainza - Mars 2021

SOMMAIRE

PARTIE 1. Nettoyage des données

1.1 Analyse des tables

1.2 Traitement des données

PARTIE 2. Analyse des données

2.1 Analyse de l'offre

2.2 Analyse de la demande

PARTIE 3. Etude de différentes corrélations

3.1 Entre sexe des clients et catégories achetées

3.2 Entre âge des clients et diverses variables

PARTIE 1

NETTOYAGE DES DONNEES

Rester livres

Analyse des tables à disposition

1. Table products

	id_prod	price	categ
1779	2_35	139.99	2
1621	0_882	24.75	0

```
Entrée [6]: produits.isna().sum()
            produits.duplicated().sum()

Out[6]: id_prod    0
        price      0
        categ      0

Out[6]: 0
```

Pas de doublons

Pas de valeurs nulles

	id_prod	price	categ
count	3287	3287.000000	3287.000000
unique	3287	NaN	NaN
mean	NaN	21.856641	0.370246
std	NaN	29.847908	0.615387
min	NaN	-1.000000	0.000000
max	NaN	300.000000	2.000000

3287 produits référencés

21,86€ comme prix moyen

Prix négatif

```
Entrée [9]: produits[produits['price'] <= 0]

Out[9]:
```

	id_prod	price	categ
731	T_0	-1.0	0

Prix < 0 concerne 1 article : id_prod = T_0

Analyse des tables à disposition

2. Table customers

	client_id	sex	birth
8549	c_720	f	1960
3639	c_8454	m	1959

```
Entrée [12]: clients.isna().sum()
             clients.duplicated().sum()

Out[12]: client_id    0
         sex          0
         birth        0

Out[12]: 0
```

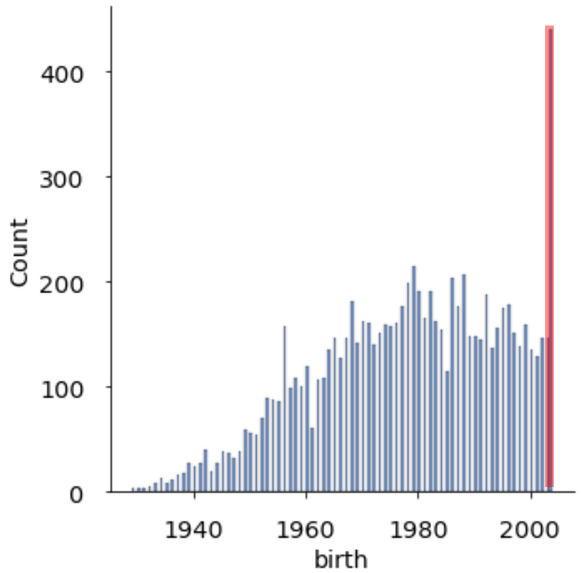
Pas de doublons

Pas de valeurs nulles

	client_id	sex	birth
count	8623	8623	8623.000000
unique	8623	2	NaN
top	c_5792	f	NaN
freq	1	4491	NaN

8623 clients enregistrés

2 valeurs pour la variable 'sex'



Année de naissance
2004 sur-représentée

Analyse des tables à disposition

3. Table transactions

	id_prod	date	session_id	client_id
311813	0_2061	2021-07-14 22:22:03.701623	s_62362	c_4491
19022	1_468	2022-02-04 13:41:57.028810	s_159873	c_2414

```
Entrée [20]: ventes.isna().sum()
             ventes.duplicated().sum()

Out[20]: id_prod       0
         date          0
         session_id    0
         client_id     0
         dtype: int64
```

→ Pas de valeurs nulles

```
Out[20]: 126
```

→ Présence de doublons

```
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   id_prod      337016 non-null  object
 1   date         337016 non-null  object
 2   session_id   337016 non-null  object
 3   client_id    337016 non-null  object
```

→ 'date' à mettre au format date

```
ventes[ventes.duplicated()].head(3)
```

	id_prod	date	session_id	client_id
34387	T_0	test_2021-03-01 02:30:02.237443	s_0	ct_0
54813	T_0	test_2021-03-01 02:30:02.237412	s_0	ct_1
57261	T_0	test_2021-03-01 02:30:02.237439	s_0	ct_1

Eléments de test :

- id_prod : T_0
- date avec préfixe 'test'
- session_id : s_0
- client avec préfixe 'ct'

	id_prod	date	session_id	client_id
count	337016	337016	337016	337016
unique	3266	336855	169195	8602
top	1_369	test_2021-03-01 02:30:02.237413	s_0	c_1609
freq	1081	13	200	12855

→ 337 016 transactions (1 transaction pour 1 produit vendu)

→ 8602 clients actifs (vs 8623 inscrits)
3266 produits différents vendus (vs 3287 référencés)

Traitement des données

1. Suppression des éléments de test

Les éléments à supprimer sont donc les suivants :

- df produits — le produit de test T_0
- df clients — les utilisateurs test ct_0 et ct_1
- df ventes — les dates qui commencent par 'test', c'est-à-dire celles qui correspondent aux sessions s_0

	client_id	sex	birth
2735	ct_0	f	2001
8494	ct_1	m	2001

	id_prod	price	categ
731	T_0	-1.0	0

	id_prod	date	session_id	client_id
1431	T_0	test_2021-03-01 02:30:02.237420	s_0	ct_1
...	...	...	...	...
332594	T_0	test_2021-03-01 02:30:02.237445	s_0	ct_0
332705	T_0	test_2021-03-01 02:30:02.237423	s_0	ct_1

200 rows × 4 columns

2. Typage des dates

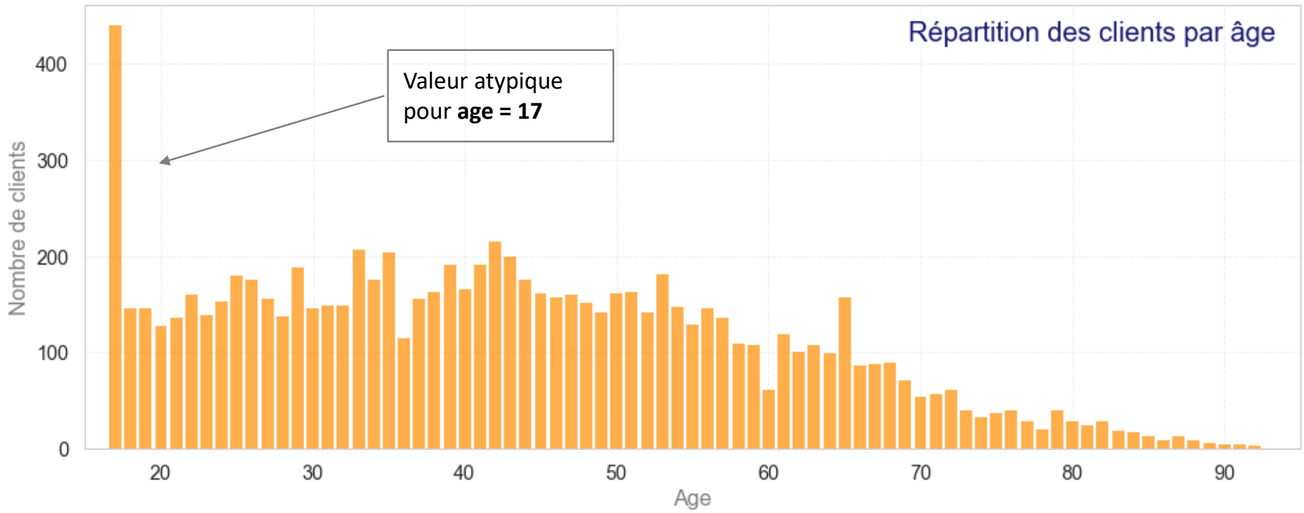
```
Entrée [246]: ventes['date'] = pd.to_datetime(ventes['date'], format='%Y-%m-%d %H:%M:%S.%f', errors='coerce')
              ventes.dtypes

Out[246]: id_prod                  object
          date            datetime64[ns]
```

	id_prod	date	session_id	client_id	s_year	s_month	s_week	s_day_name	s_day	s_hour	s_minute	s_date
329479	1_630	2021-05-05 21:18:00.766411	s_30426	c_5989	2021	5	18	Mercredi	5	21	18	2021-05-05
242149	0_1011	2021-08-25 01:01:02.229790	s_80264	c_4922	2021	8	34	Mercredi	25	1	1	2021-08-25
68278	0_1490	2021-12-24 11:18:28.803699	s_139247	c_1609	2021	12	51	Vendredi	24	11	18	2021-12-24

Traitement des données

3. Ajout de l'âge dans df 'clients'

	client_id	sex	birth	age
7274	c_628	m	1991	30
2571	c_3203	m	1965	56
5927	c_1582	f	1947	74



- Remarque sur âge < 18 ans
- Hypothèse pour expliquer ce pic :
  -> paramètres par défaut

4. Merge des 3 df

	id_prod	date	s_id	client_id	s_year	s_month	s_week	s_day_name	s_day	s_hour	s_minute	s_date
180400	0_1416	2021-11-17 11:39:25.825676	s_120961	c_7526	2021	11	46	Mercredi	17	11	39	2021-11-17
179638	1_308	2021-11-05 05:34:08.120657	s_114964	c_2208	2021	11	44	Vendredi	5	5	34	2021-11-05
108332	0_1048	2021-12-02 15:57:38.436661	s_128392	c_7471	2021	12	48	Jeudi	2	15	57	2021-12-02

	id_prod	price	categ
1118	0_1609	18.99	0
3048	0_1711	2.99	0
2177	0_542	18.18	0

	client_id	sex	birth	age
7402	c_5651	m	1964	57
2588	c_4353	f	1997	24
201	c_4694	f	1965	56

Traitement des données

```
Entrée [235]: general = pd.merge(ventes, produits, how='outer')
              general = pd.merge(general, clients, how='outer')
              general.sort_values('s_id')
```

Out[235]:

	id_prod	date	s_id	client_id	s_year	s_month	s_week	s_day_name	s_day	s_hour	s_minute	s_date	price	categ	sex	birth	age
272113	0_1259	2021-03-01 00:01:07.843138	s_1	c_329	2021.0	3.0	9.0	Lundi	1.0	0.0	1.0	2021-03-01	11.99	0.0	f	1967.0	54.0
212254	1_635	2021-03-01 00:10:33.163037	s_10	c_2218	2021.0	3.0	9.0	Lundi	1.0	0.0	10.0	2021-03-01	26.99	1.0	f	1970.0	51.0
24502	0_1451	2021-03-01 04:43:58.025677	s_100	c_3854	2021.0	3.0	9.0	Lundi	1.0	4.0	43.0	2021-03-01	19.99	0.0	f	1978.0	43.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
336856	NaN	NaT	NaN	c_90	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaT	NaN	NaN	m	2001.0	20.0
336857	NaN	NaT	NaN	c_587	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaT	NaN	NaN	m	1993.0	28.0
336858	NaN	NaT	NaN	c_3526	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaT	NaN	NaN	m	1956.0	65.0

336859 rows × 17 columns

5. Traitement des valeurs nulles

```
Entrée []: general.isna().sum()
```

```
Out[]: id_prod       21      s_hour       43
       date          43      s_minute     43
       s_id          43      s_date       43
       client_id     22      price       124
       s_year        43      categ       124
       s_month       43      sex          22
       s_week        43      birth        22
       s_day_name    43      age          22
       s_day         43
```

✓ <u>Un produit non référencé dans la table produits</u>

```
Entrée []: general[general['price'].isna()]['id_prod'].value_counts()
```

```
Out[]: 0_2245    103
       Name: id_prod, dtype: int64
```

le produit d'Id '0_2245', commandé 103 fois, n'apparait pas dans la table 'produits'

↳ On va imputer son prix par le prix moyen de sa catégorie (catégorie 0)

```
general.loc[general['id_prod'] == '0_2245', 'price'] = mean_cat_0
general.loc[general['id_prod'] == '0_2245', 'categ'] = 0.0
```

Traitement des données

✓ Des produits référencés jamais vendus

```
Entrée [281]: general[general['client_id'].isna()].sample(3)
              prod_sans_commande_ls = list(general[general['client_id'].isna()]['id_prod'])
              len(prod_sans_commande_ls)
```

Out[281]:

	id_prod	date	s_id	client_id	s_year	s_month	s_week	s_day_name	s_day	s_hour	s_minute	s_date	price	categ	sex	birth	age
336827	0_1025	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaT	24.99	0.0	NaN	NaN	NaN
336829	1_394	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaT	39.73	1.0	NaN	NaN	NaN
336830	2_72	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaT	141.32	2.0	NaN	NaN	NaN

Out[281]: 22

22 produits référencés sans aucune vente
→ soit 0,67% des produits

- On supprime ces produits de notre df général
- On crée un df pour ces produits

```
prod_sans_commande = produits[produits['id_prod'].isin(prod_sans_commande_ls)]
general = general.dropna(subset=['client_id'])
```

✓ Des clients inscrits sans aucun achat

```
Entrée [283]: general[general['id_prod'].isna()].sample(3)
              clients_sans_commande_ls = general[general['id_prod'].isna()]['client_id']
              len(clients_sans_commande_ls)
```

Out[283]:

	id_prod	date	s_id	client_id	s_year	s_month	s_week	s_day_name	s_day	s_hour	s_minute	s_date	price	categ	sex	birth	age
336858	NaN	NaT	NaN	c_3526	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaT	NaN	NaN	m	1956.0	65.0
336852	NaN	NaT	NaN	c_5223	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaT	NaN	NaN	m	2003.0	18.0
336850	NaN	NaT	NaN	c_6862	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaT	NaN	NaN	f	2002.0	19.0

Out[283]: 21

21 clients inscrits sans aucune commande
→ soit 0,24% des clients

- On supprime ces clients de notre df général
- On crée un df pour ces clients

```
clients_sans_commande = clients[clients['client_id'].isin(clients_sans_commande_ls)]
general = general.dropna(subset=['id_prod'])
```

Traitement des données

6. Organisation du df global

On renomme les colonnes en les préfixant avec :

- p pour produit
- c pour client
- s pour session

	c_id	c_sex	c_age	c_birth	p_id	p_categ	p_prix	s_panier_id	s_id_date	s_year	s_month	s_week	s_day_n	s_day	s_hour	s_minute	s_date
71849	c_2939	f	34.0	1987.0	0_1527	0.0	7.99	s_86755	2021-09-08 09:17:38.503748	2021.0	9.0	36.0	Mercredi	8.0	9.0	17.0	2021-09-08
110125	c_632	m	39.0	1982.0	1_376	1.0	17.49	s_29904	2021-05-04 18:45:45.163431	2021.0	5.0	18.0	Mardi	4.0	18.0	45.0	2021-05-04
79801	c_5062	m	33.0	1988.0	0_1583	0.0	15.99	s_155561	2022-01-26 17:14:50.233159	2022.0	1.0	4.0	Mercredi	26.0	17.0	14.0	2022-01-26
65021	c_3845	m	40.0	1981.0	0_1142	0.0	3.42	s_102583	2021-10-10 03:57:43.061085	2021.0	10.0	40.0	Dimanche	10.0	3.0	57.0	2021-10-10
128148	c_8548	m	43.0	1978.0	0_1303	0.0	7.99	s_3265	2021-03-08 02:46:32.924273	2021.0	3.0	10.0	Lundi	8.0	2.0	46.0	2021-03-08

PARTIE 2

ANALYSE DES DONNEES

Rester livres

Etude de l'offre

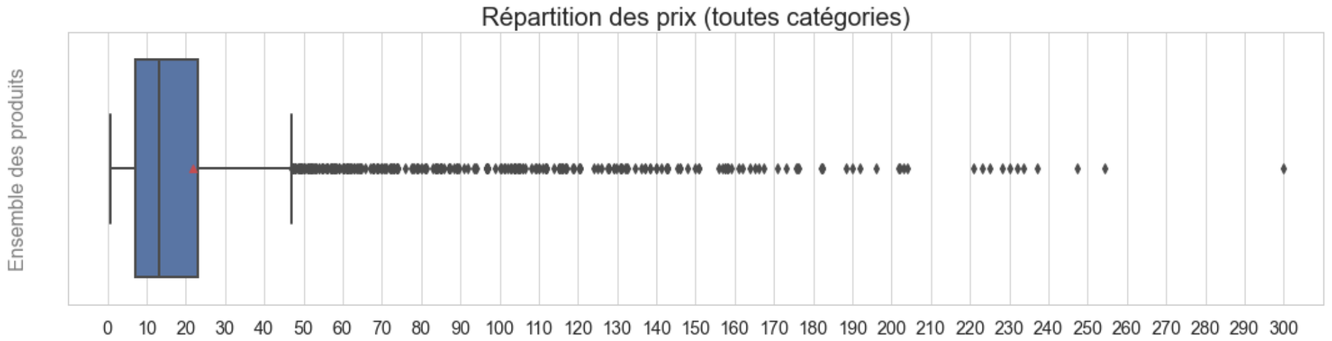
1. Analyse des prix des produits

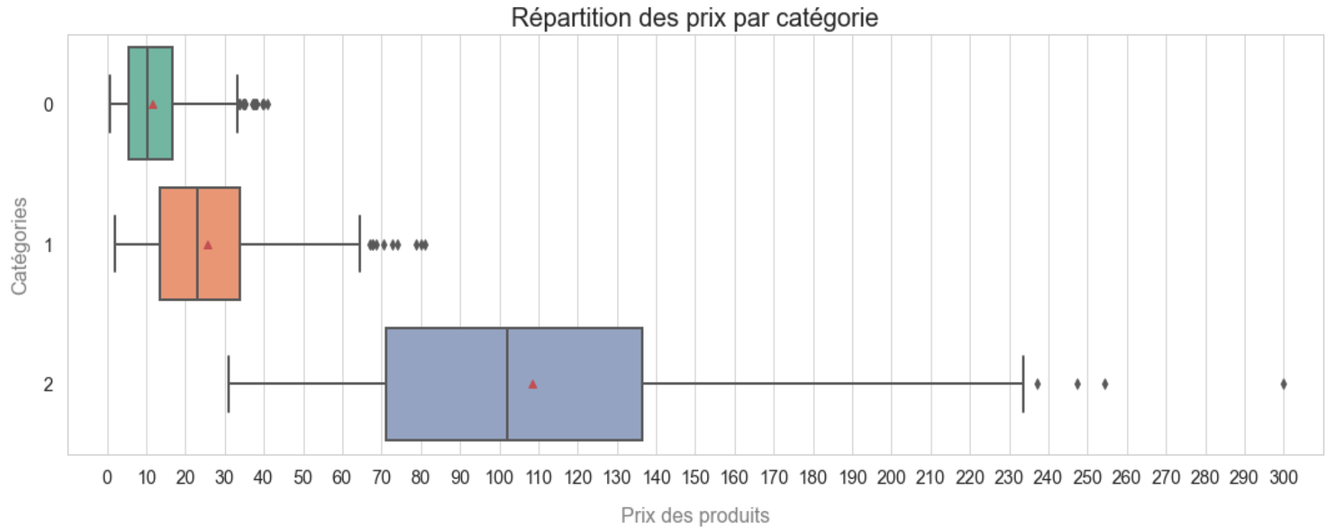
3287 produits proposés

	Nombre	Part	Prix Moyen
Categorie 0	2309	70.25%	11.73€
Categorie 1	739	22.5%	25.53€
Categorie 2	239	7.25%	108.35€
Total	3287	100%	21.86€

Répartition des prix

categ	count	mean	std	min	25%	50%	75%	max
0	2309.0	11.732795	7.564116	0.62	5.590	10.32	16.65	40.99
1	739.0	25.531421	15.425162	2.00	13.390	22.99	33.99	80.99
2	239.0	108.354686	49.561431	30.99	71.065	101.99	136.53	300.00
total	3287.0	21.860515	29.845766	0.62	6.99	13.06	22.99	300.0





Toutes catégories confondues :

- la moitié des produits ont un prix inférieur à 13€ environ
- le prix des 3/4 des produits est inférieur à 23€ environ

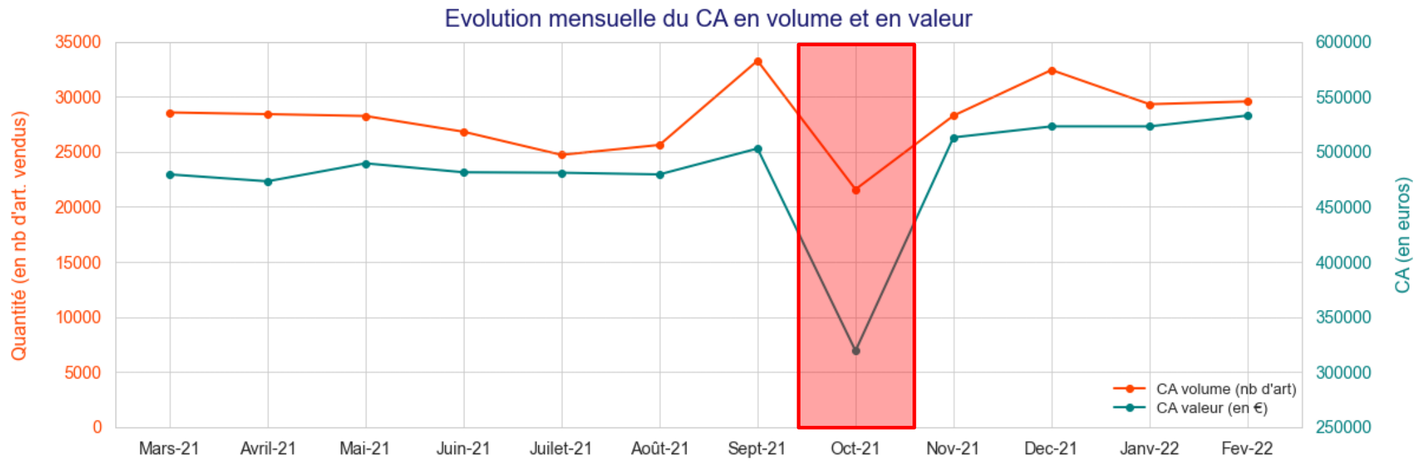
Par catégorie :

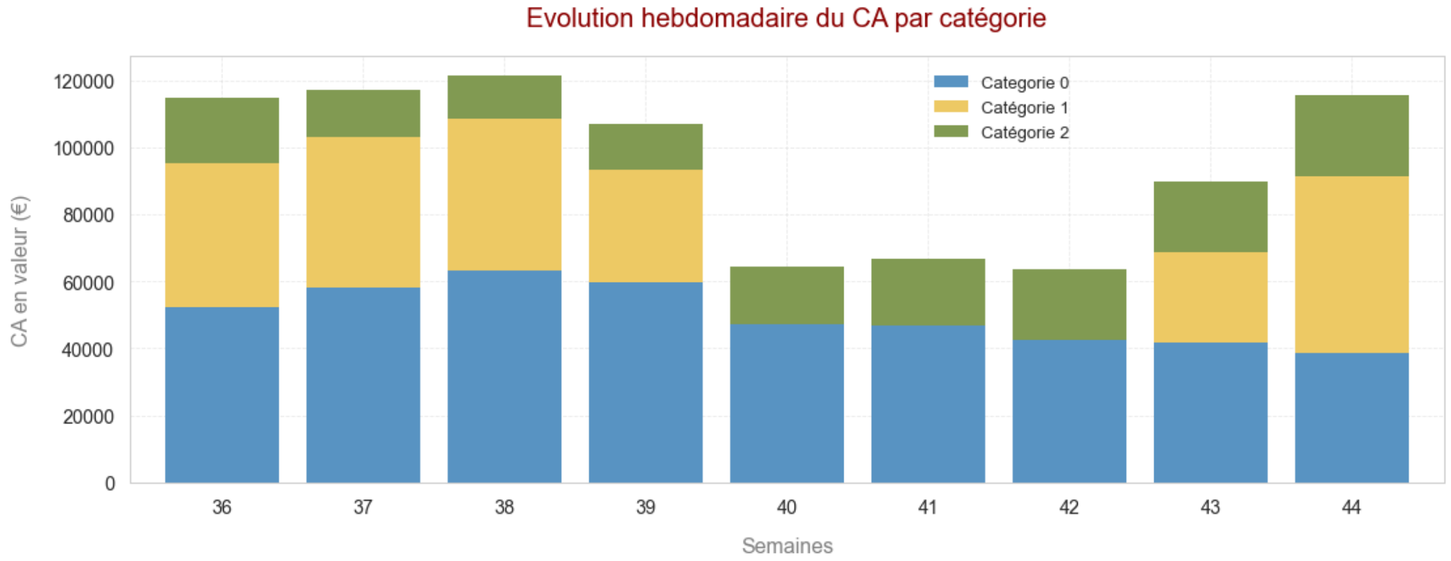
Plus la catégorie est élevée :

- plus le prix est important,
- plus la dispersion est importante

Etude de l'offre

2. Evolution du CA de manière globale





Le CA sur un an :

- en valeur **5 800 000€** (483 000€ / mois)
- en volume **336 800 produits** (28 000 / mois)

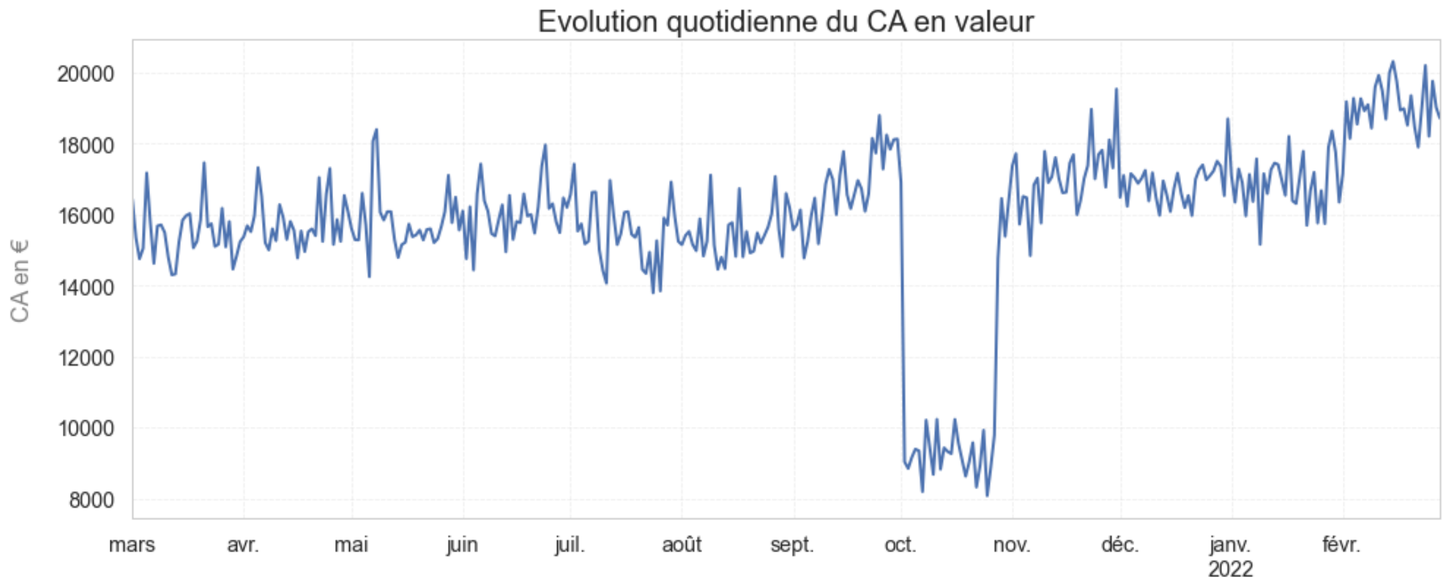
Evolution du CA relativement stable sauf pour le mois d'octobre :

↳ chute du CA : ↘ 35% en valeur

Pendant au moins 3 semaines :

↳ aucun produit de catégorie 1 n'est vendu
↳ évolution dans le trend pour les 2 autres catégories

Etude de l'offre



Evolution du CA journalier

-> absence de ventes de produits de catégorie 1 du 02 octobre au 27 octobre 2021

Hypothèse la plus probable :

-> problème d'enregistrement dans la bdd

-> ou problème d'accès à ces données

Traitement des données manquantes en 2 temps :

- estimer les ventes de produits de catégorie 1 → méthode 'interpolate'
- supprimer les données du mois d'octobre de notre df final (car l'estimation ne portera pas sur la répartition du CA entre les clients)

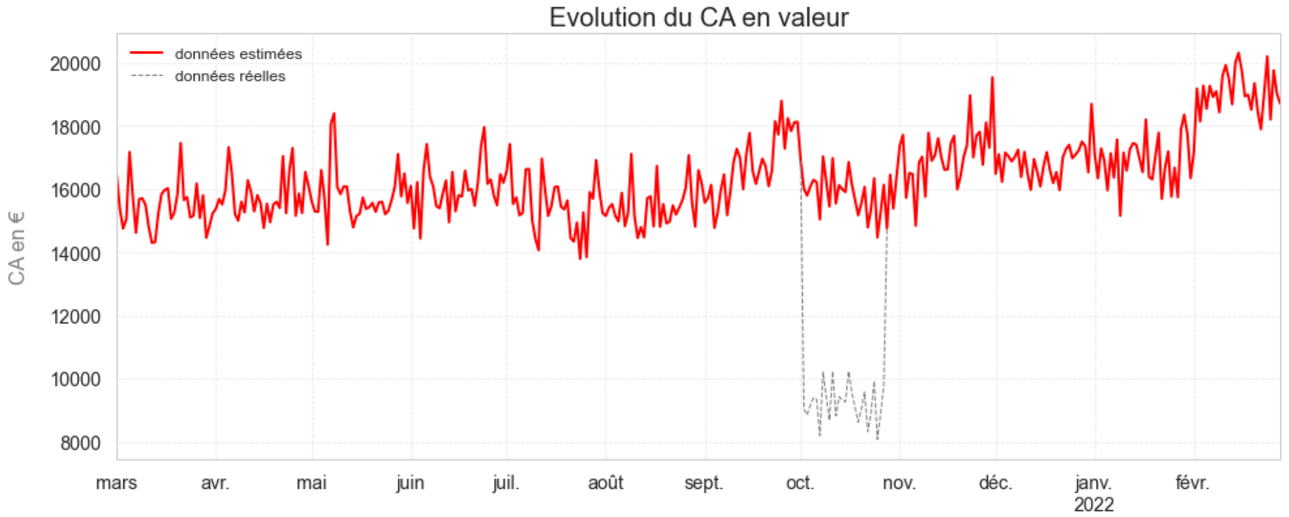
Etude de l'offre

3. Estimation des données manquantes

Utilisation de la méthode 'interpolate' pour estimer les données de vente de produit de catégorie 1

CA évolue de façon linéaire → données estimées par régression linéaire

	date	nb_p0	nb_p1	nb_p2	ca_p0	ca_p1	ca_p2	nb_total	ca_total
220	2021-10-07 00:00:00	597.00	337.78	26.00	6404.01	6851.39	1787.07	960.78	15042.47
221	2021-10-08 00:00:00	669.00	336.74	44.00	7069.53	6825.99	3137.82	1049.74	17033.34
222	2021-10-09 00:00:00	640.00	335.70	35.00	6808.69	6800.59	2616.67	1010.70	16225.95



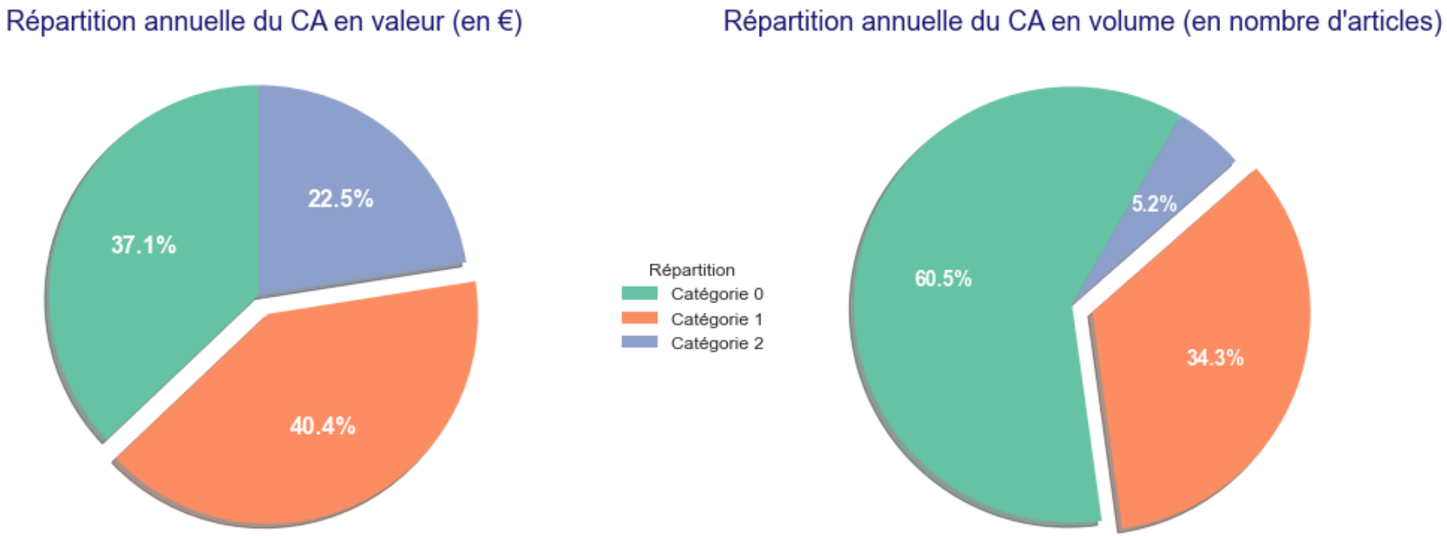
4. Suppression du mois d'octobre dans le df final

```
res_estimated = res.copy()
res_mask = res_estimated['s_month'] != 10.0
resultat = res_estimated[res_mask]
resultat.sample(3)
```

→ Le CA corrigé sur un an s'élève à **5 478 544€**, soit **498 049€ par mois**

Etude de l'offre

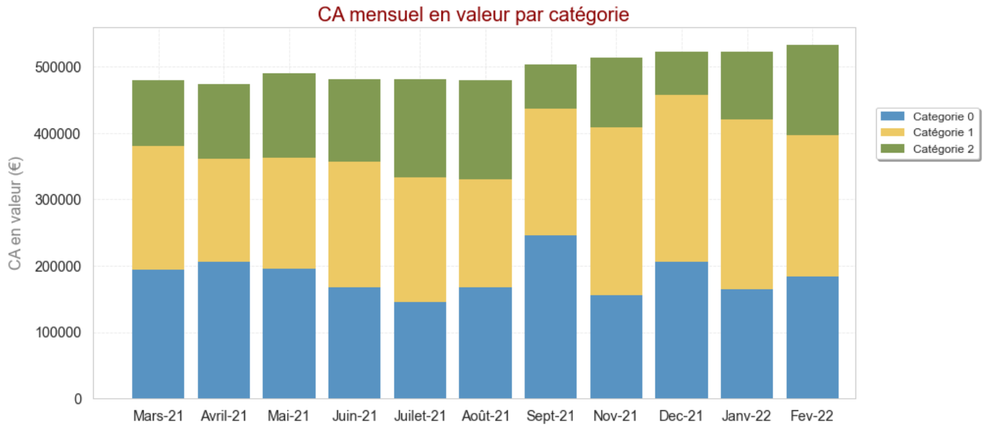
5. Répartition du CA par catégorie

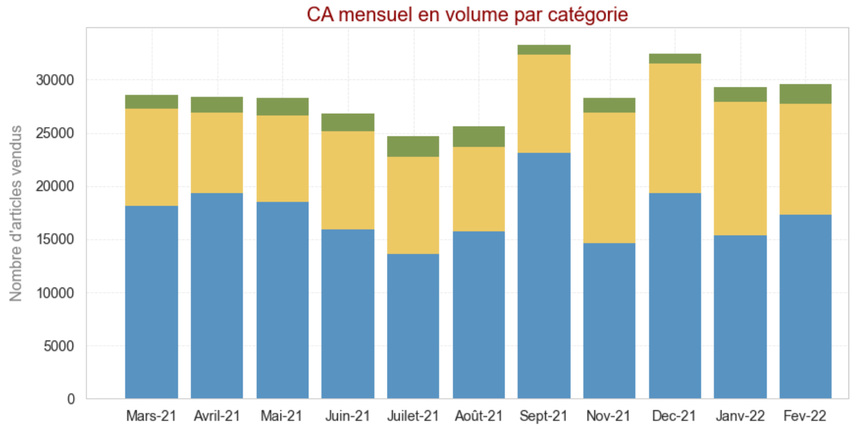


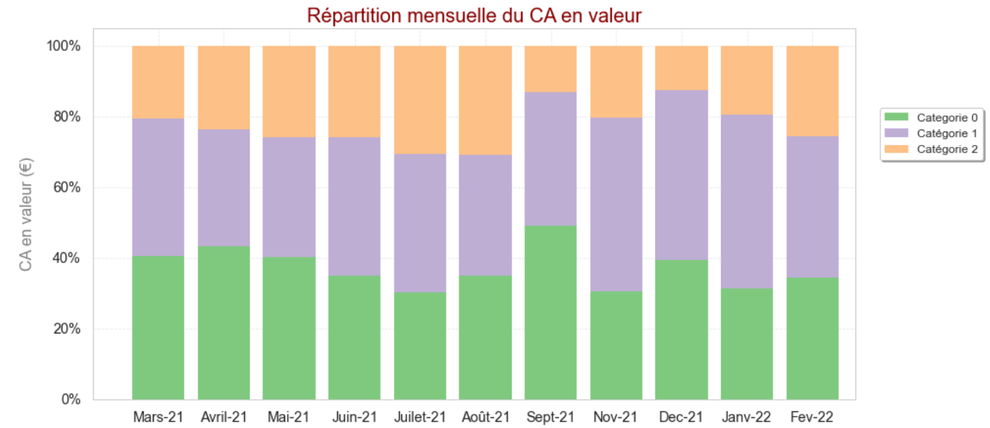
Les produits de **catégories 2** ne représentent que **5% du CA en terme de volume mais 22.5% en terme de valeur** (les produits de cette catégorie ont un prix de vente plus élevé)

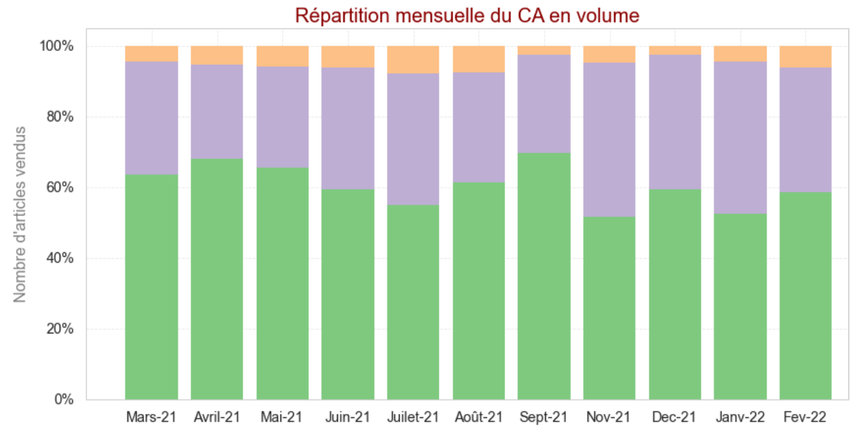
Etude de l'offre

6. Evolution du CA par catégorie









✓ en terme de valeur :

- les produits de catégorie 0 et 1 représentent entre 30 et 50% du CA
- la catégorie 2 entre 10 et 30%

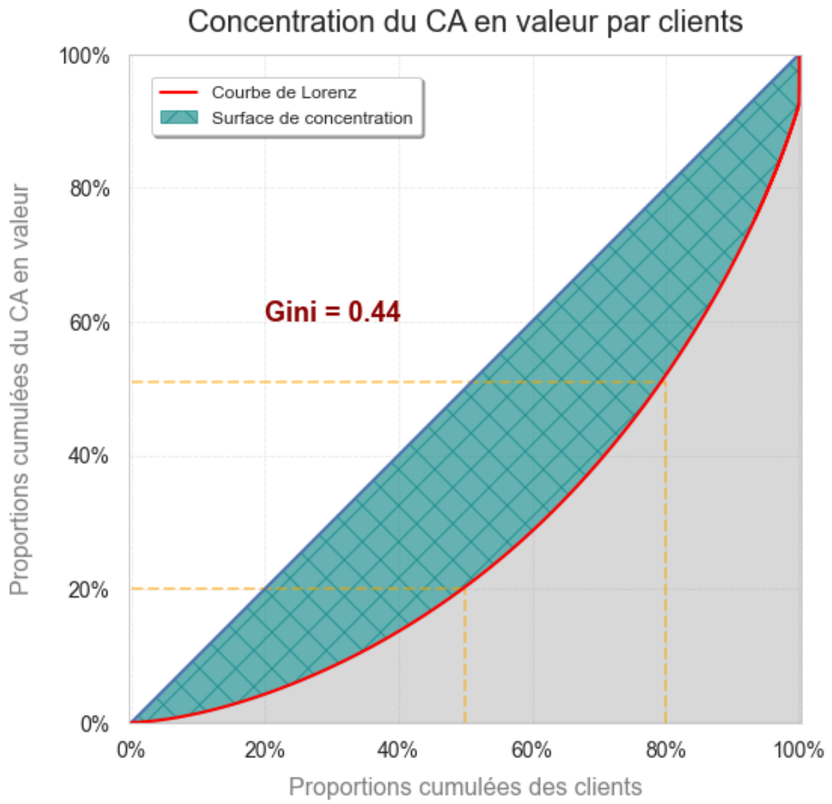
✓ en terme de volume :

- environ 60% pour la catégorie 0,
- 35% pour la catégorie 1,
- et 5% pour la catégorie 2

Etude de l'offre

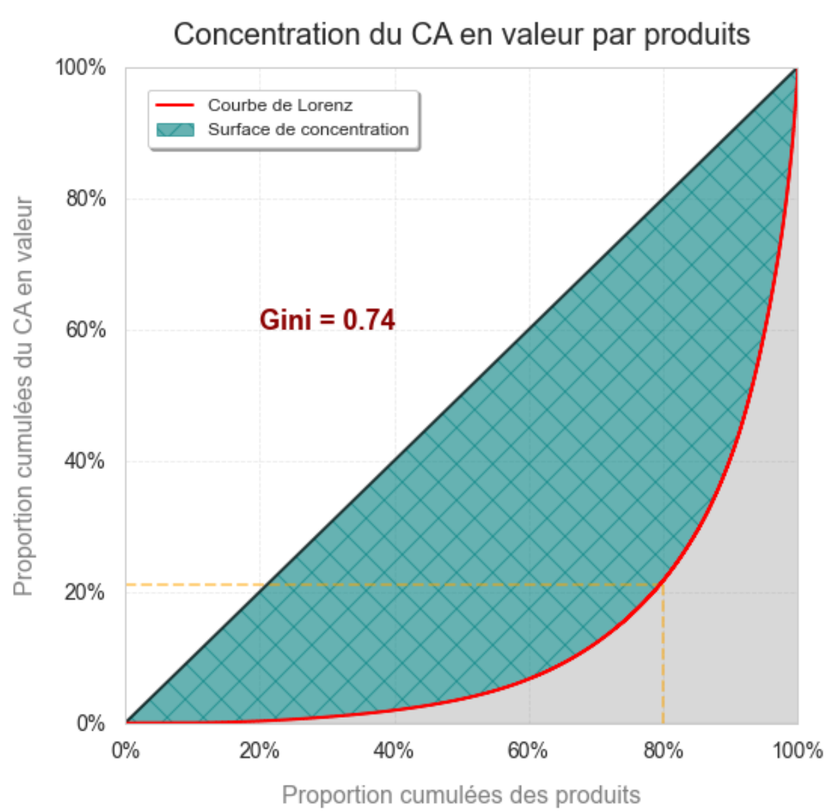
7. Concentration du CA

CA et clients



- la moitié des clients génèrent 20% du CA
- les 20% plus gros clients génèrent 50% du CA

CA et produits

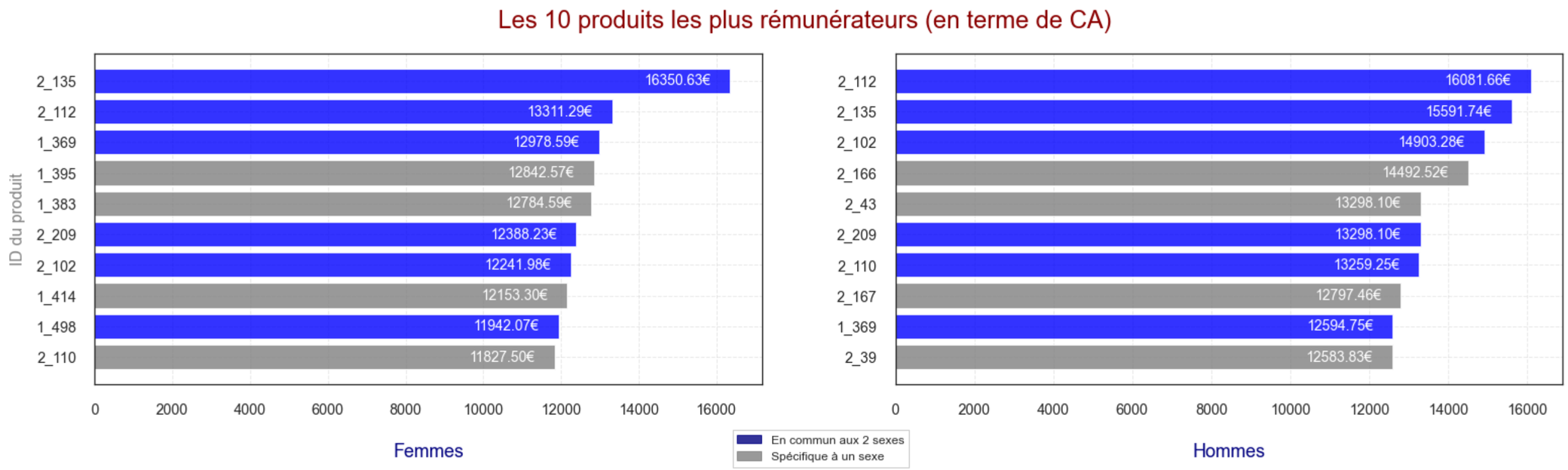


20% des produits génèrent près de 80% du CA

Etude de l'offre

8. Les meilleurs produits

Intérêt de **différencier hommes et femmes** pour les 10 produits les plus rémunérateurs



La proportion **de produits de catégorie 2** parmi les 10 produits les plus rémunérateurs :
- 90% chez les hommes
- 50% chez les femmes

Intéressant dans l'optique d'une **segmentation de la base client**

Etude de la demande

1. Données générales

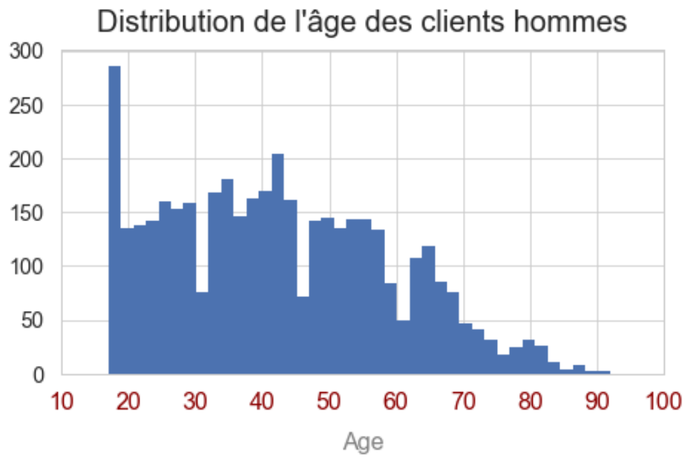
✓ 8623 clients enregistrés mais :

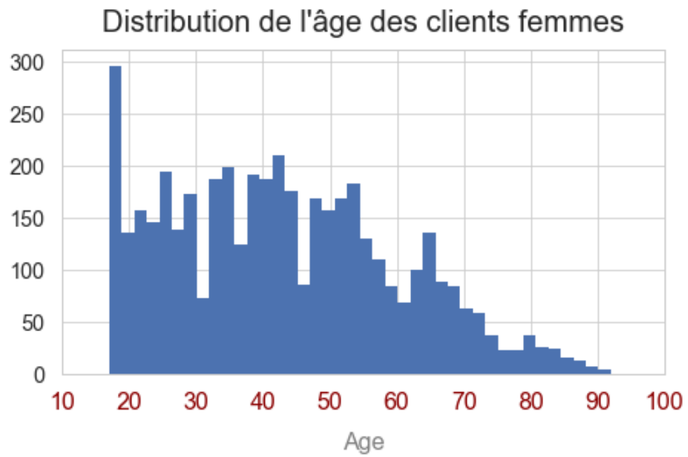
- 21 clients n'ont jamais commandés
- 2 clients étaient des clients test
- 2 clients n'ont commandé qu'au mois d'octobre 2011 (or on a supprimé cette période)

-> on compte donc **8598 clients actifs**

✓ Répartition hommes / femmes des clients :

- 4121 hommes, soit **47,93%**
- 4477 femmes, soit **52,07%**





↳pas de différence notable entre les sexes

Etude de la demande

2. Analyse des achats

Création de 2 df des commandes des clients :

- un df des **commandes** (*clé unique :* *s_panier_id*) :
  * infos sur le client (âge, genre)
  * détails de chaque commande (date, montant et nombre d'articles achetés)

Df des_commandes ('res_clients_all_orders') :

	c_nature	c_id	c_sex	c_age	s_panier_id	s_id_date	nb_art_panier	montant_panier	nb_art_cat_0	nb_art_cat_1	nb_art_cat_2	montant_cat_0	montant_cat_1	montant_cat_2
82522	professionnel	c_4958	m	22.0	s_63667	2021-07-17 22:02:12.871000	2.0	123.78	0.0	0.0	2.0	0.0	0.00	123.78
6687	particulier	c_1368	m	39.0	s_16610	2021-04-06 02:28:47.707052	1.0	19.87	0.0	1.0	0.0	0.0	19.87	0.00

- un df récapitulatif des **clients** (clé unique : *c_id*) :
  * infos sur le client (âge, genre)
  * nb de commandes réalisées et nb d'articles achetés au cours de l'année

Df récapitulatif des_clients ('res_clients_year') :

	c_nature	c_id	c_sex	c_age	nb_commandes	nb_articles	montant_total	nb_art_cat_0	nb_art_cat_1	nb_art_cat_2	montant_cat_0	montant_cat_1	montant_cat_2
3050	particulier	c_375	m	21.0	3	12.0	247.85	9.0	2.0	1.0	106.51	77.98	63.36
4112	particulier	c_471	f	24.0	6	13.0	523.72	6.0	4.0	3.0	57.92	98.28	367.52

Etude de la demande

2.1 Total des achats par client

```
Entrée [381]: res_clients_year.describe().T
```

Out[381]:

	count	mean	std	min	25%	50%	75%	max
c_age	8598.0	42.739591	16.909801	17.00	29.0000	42.000	55.0000	92.00
nb_commandes	8598.0	18.335427	66.536835	1.00	7.0000	12.000	23.0000	5042.00
nb_articles	8598.0	36.663410	144.875160	1.00	13.0000	24.000	44.0000	11839.00
montant_total	8598.0	637.188202	2417.248311	4.15	260.9825	475.675	822.5025	150729.07
nb_art_cat_0	8598.0	22.187834	107.144057	0.00	4.0000	11.000	26.0000	9303.00
nb_art_cat_1	8598.0	12.569086	40.788855	0.00	5.0000	9.000	16.0000	2535.00
nb_art_cat_2	8598.0	1.906490	17.196863	0.00	0.0000	0.000	1.0000	1558.00
montant_cat_0	8598.0	236.291121	1141.198347	0.00	41.9250	120.815	276.2775	99015.95
montant_cat_1	8598.0	257.457791	835.691078	0.00	93.7200	182.340	334.5750	51671.81
montant_cat_2	8598.0	143.439289	1306.427120	0.00	0.0000	0.000	110.0550	118353.71

CA annuel moyen par client est de **637€**

- mais de **fortes disparités**
  -> le plus *gros client* génère à lui tout seul un CA de plus de **15 000€**.

On trie de façon décroissante le CA par client

```
Entrée [382]: res_clients_year['montant_total'].nlargest(7)
```

```
Out[382]: 677     150729.070000
          4387    137151.480000
          6336     69405.635589
          2723     52744.145589
          7790      2436.232795
          7119      2406.170000
          7005      2366.200000
```

4 clients 'premiums'

→ 7.48% du CA total

→ A priori des professionnels

On ajoute une colonne *c_nature* aux 2 df clients pour spécifier la nature du client (particulier ou professionnel)

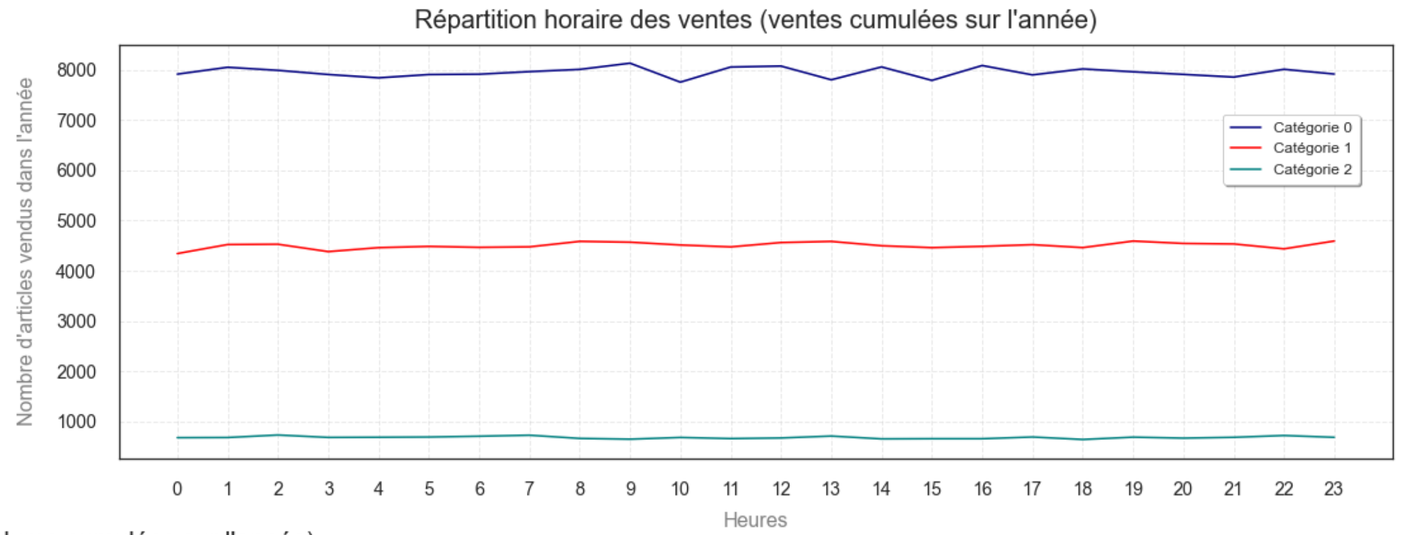
c_nature	c_id	c_sex	c_age	s_panier_id	s_id_date	nb_art_panier	montant_panier	nb_art_cat_0	nb_art_cat_1	nb_art_cat_2	montant_cat_0	montant_cat_1	montant_cat_2
particulier	c_1	m	66.0	s_114737	2021-11-04 17:28:13.934070	5.0	92.62	4.0	0.0	1.0	37.75	0.00	54.87
				s_120172	2021-11-15 20:40:00.586010	2.0	44.29	0.0	2.0	0.0	0.00	44.29	0.00
				s_134971	2021-12-15 23:32:41.632729	1.0	10.30	0.0	1.0	0.0	0.00	10.30	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...
professionnel	c_6714	f	53.0	s_97533	2021-09-29 18:05:03.844078	7.0	93.37	6.0	1.0	0.0	74.38	18.99	0.00
				s_97726	2021-09-30 03:27:58.500427	6.0	72.52	5.0	1.0	0.0	50.53	21.99	0.00
				s_97791	2021-09-30 06:06:51.038156	8.0	121.43	4.0	4.0	0.0	39.42	82.01	0.00

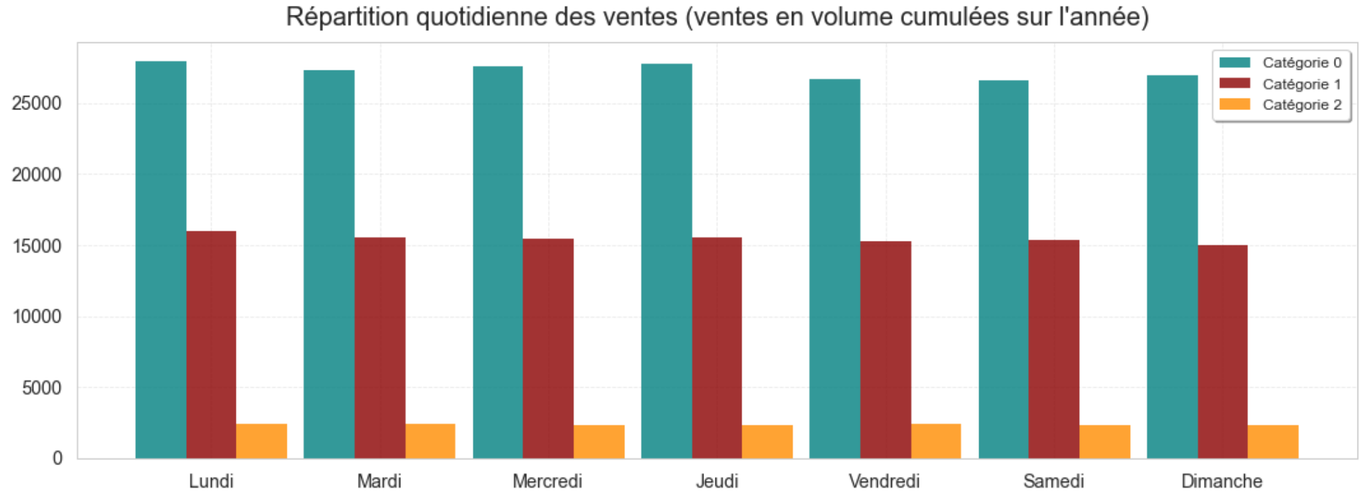
157648 rows × 9 columns

Etude de la demande

2.2 Répartition annuelle des achats

Les ventes sont réparties de manière homogène sur l'ensemble de la journée…





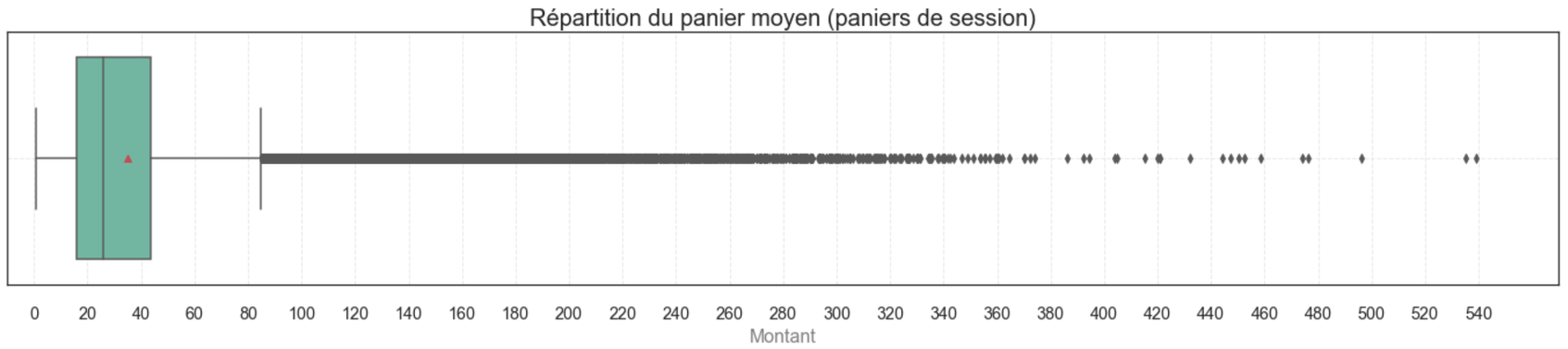
… ainsi que sur l'ensemble de la semaine

Etude de la demande

3. Le panier moyen

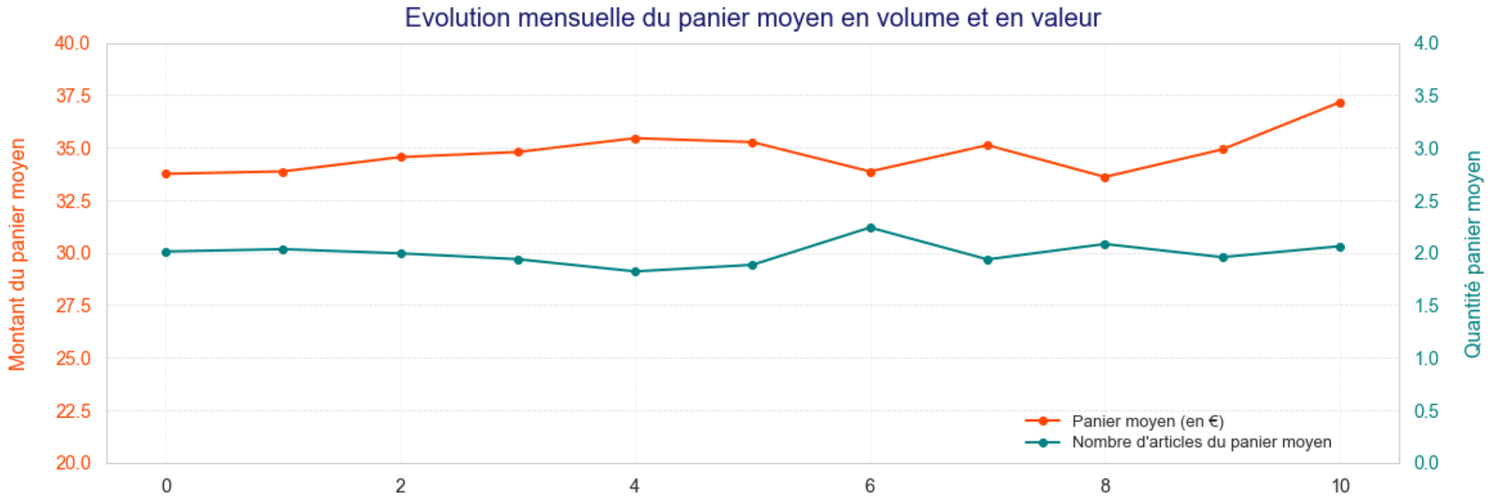
Remarque : 2 types de paniers moyens

- panier moyen de session (cf. la moyenne de toutes les commandes réalisées)
- panier moyen des clients (cf. la moyenne des paniers moyens de chaque client)



Pour les clients particuliers

	c_age	nb_art_panier	montant_panier
count	147046.000000	147046.000000	147046.000000
mean	45.180440	1.996709	34.468900
std	15.201614	1.280324	31.672006
min	17.000000	1.000000	0.620000
25%	34.000000	1.000000	15.810000
50%	43.000000	2.000000	25.990000
75%	55.000000	3.000000	43.250000
max	92.000000	14.000000	539.230000



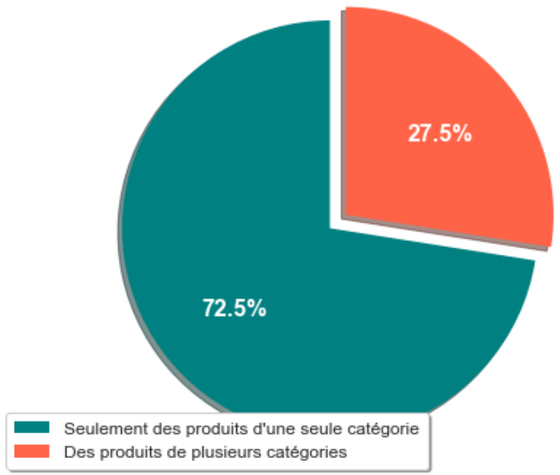
Caractéristiques du panier moyen :

- Montant : **34,5€**
- Composition : **2 articles**
- **Stable** sur l'année étudiée

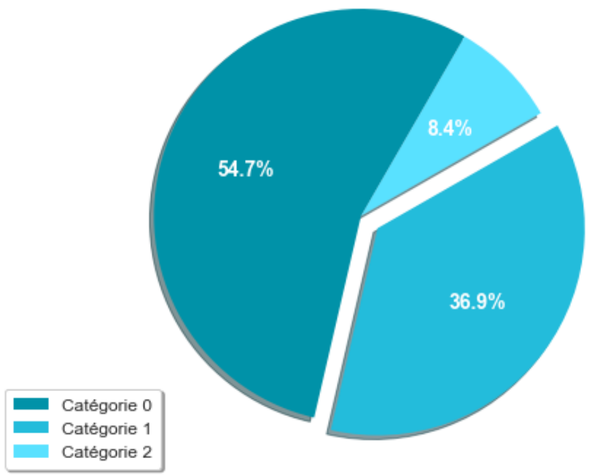
Etude de la demande

3.1 Répartition annuelle des achats

Composition du panier : 1 ou plusieurs catégories



Répartition des paniers d'une seule catégorie



✓ **Près de 75% des commandes** comportent uniquement des produits d'**une seule catégorie**

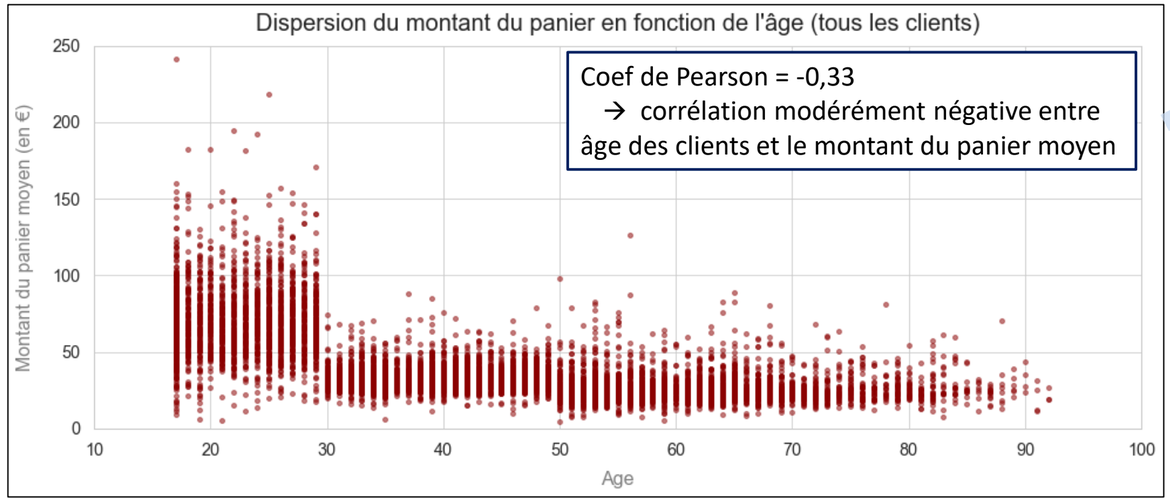
(en raisonnant en terme de client, 65% des produits qu'ils commandent sont de la même catégorie)

Intérêt de **s'interroger sur la raison de ce comportement**

- volonté des clients ?
- ergonomie du site qui n'incite pas à commander des produits de différentes catégories ?

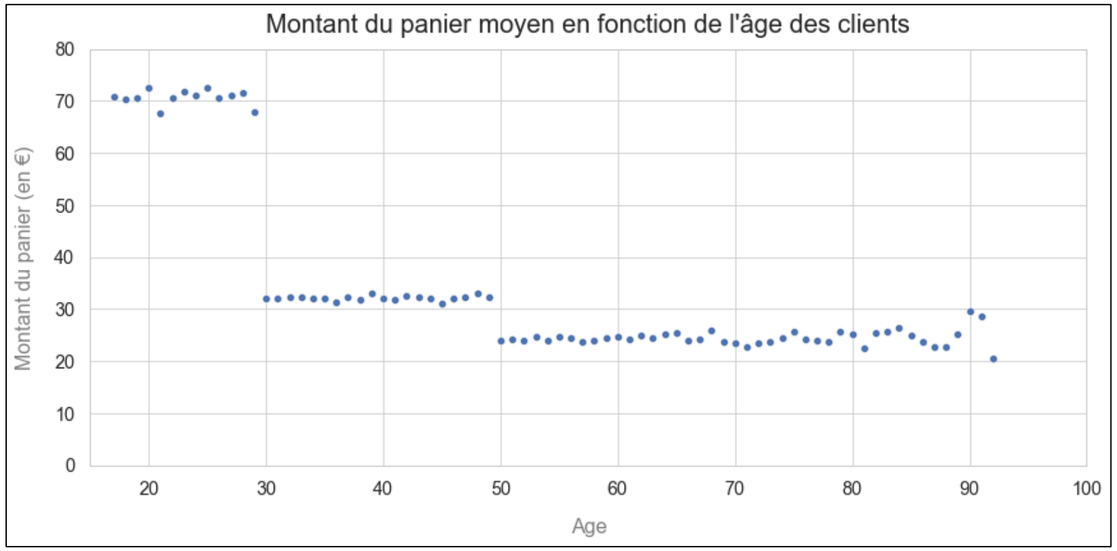
Etude de la demande

4. Relation entre âge des clients et le montant du panier moyen (base clients)



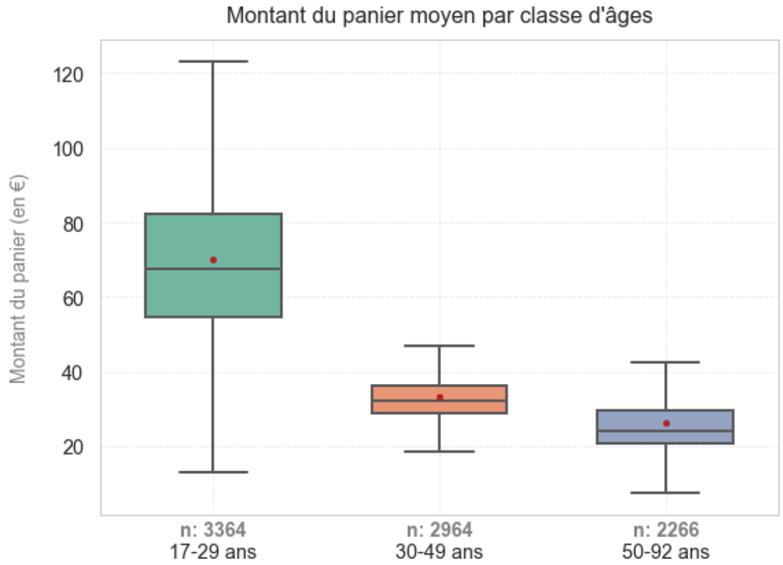
Coef de Pearson = -0,33
→ corrélation modérément négative entre âge des clients et le montant du panier moyen

Graphiquement, il semble qu'il existe **différents groupes d'âges**
→ On va agréger le montant des paniers par âge en calculant la moyenne
Obj : dégager des tendances graphiques



On peut regrouper les clients en 3 différentes classes d'âge :

- moins de 30 ans
- entre 30 et 49 ans
- 50 ans et plus



Etude de la demande

Analyse de la variance : ANOVA

Objectif : comparer les moyennes de trois groupes ou plus, créés par une variable catégorielle

Hypothèse nulle H0 :	La moyenne des différents groupes est égale
Hypothèse alternative H1 :	Au moins un groupe possède une moyenne différente
Seuil de signification choisi :	5%

Test F pour connaître le seuil de signification

$$\mathcal{F} = \frac{\frac{SCE}{p-1}}{\frac{SCR}{n-p}}$$

p : le nombre de groupes
n : le nombre d'observations

Analyse de la Variance : ANOVA

SCT = SCE + SCR

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$$

SCT : Variation totale
somme des carrés totaux
Total Sum of Squares

$$SCE = \sum_{i=1}^{k} n_i(\overline{y_i} - \overline{y})^2$$

SCE : Variation interclasse
somme des carrés expliqués
Sum of Squares of the Model
Variation entre les groupes

$$SCR = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{k} n_i s_i^2$$

SCR : Variation intraclasse
somme des carrés résiduels
Sum of Squares of the Error
Variation à l'intérieur des groupes

Calcul de eta² pour mesurer la corrélation

$$\text{eta}^2\ (\eta^2) = \frac{SCE}{SCT}$$

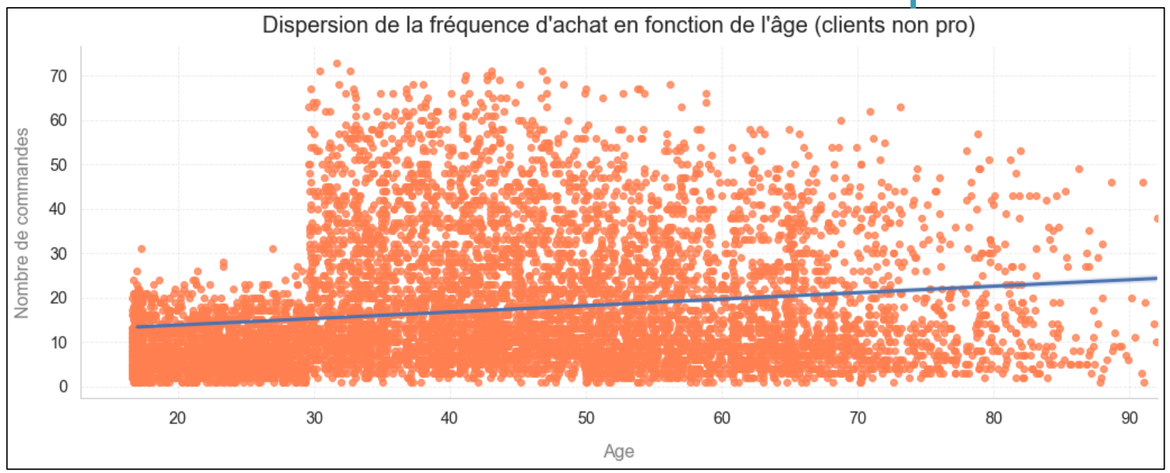
Ainsi, en appliquant ces calculs pour la relation entre âge et montant du panier moyen :
- la **p_value** obtenue (cf le seuil de signification) est inférieure à 0,05
  → donc on rejette H0 au profit de H1 : **au moins une moyenne d'un groupe d'âge est différente**

- **eta²** = 0,61538 → la corrélation entre nos variables est donc forte

Etude de la demande

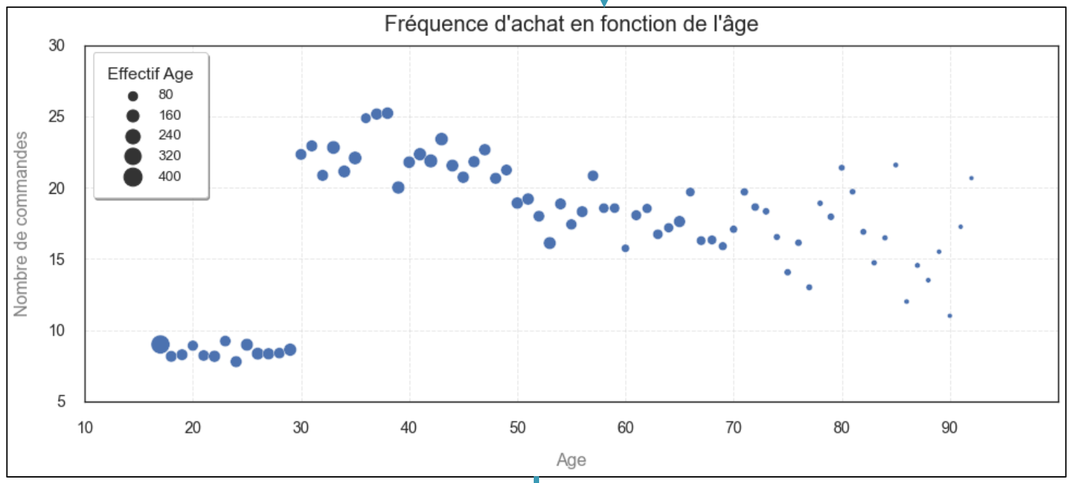
5. Relation entre âge des clients et la fréquence d'achat (nombre de commandes)

- 2 variables quantitatives
- On limite les données aux clients non professionnels



Pearson = 0,1747
→ les variables "âge" et "fréquence d'achat" sont **indépendantes**

→ on agrège les données (en faisant la moyenne des nombre de commandes par âge)



2 groupes d'âges se dégagent :

- les moins de 30 ans
- les 30 ans et plus

Etude de la corrélation pour chacune des 2 classes d'âges

Calcul du coef de Pearson pour mesurer la corrélation

- moins de 30 ans : - 0.0191 → pas de corrélation
- 30 ans et plus : - 0.1363 → pas de corrélation

PARTIE 3

ETUDE DE DIFFÉRENTES CORRELATIONS

Rester livres

Corrélation entre le sexe des clients et les catégories de produits achetés

2 variables qualitatives → test du Khi²

Hypothèse nulle H0 : Les deux variables qualitatives sont indépendantes
Hypothèse alternative H1 : Les deux variables qualitatives sont dépendantes
Seuil de signification choisi : 5%

Tableau de contingence
→ les données observées

	nb_art_cat_0	nb_art_cat_1	nb_art_cat_2	total_sex
f	92331.0	52993.0	7575.0	152899.0
m	85525.0	48061.0	7123.0	140709.0
total_cat	177856.0	101054.0	14698.0	293608.0

1 (177856 * 152899) / 293608 = 92620,107

Tableau des effectifs théoriques
→ Tableau d'indépendance

	nb_art_cat_0	nb_art_cat_1	nb_art_cat_2	total_sex
f	92620.107572	52924.77707	7054.115358	152899.0
m	85235.892428	48429.22293	7043.884642	140709.0
total_cat	177856.000000	101054.00000	14698.000000	293608.0

2 (92331 – 92620,108)² = 83583,188

	nb_art_cat_0	nb_art_cat_1	nb_art_cat_2
f	83583.188188	135588.126099	6259.239854
m	83583.188188	135588.126099	6259.239854

3 83583,188 / 92620,108 = 0,90243

Tableau des écarts à l'indépendance

	nb_art_cat_0	nb_art_cat_1	nb_art_cat_2	total_sex
f	0.90243	2.576507	0.817761	4.296699
m	0.98061	2.799717	0.888606	4.668934
total_cat	1.88304	5.376225	1.706368	8.965633

Valeur du Khi²

Seuil de signification
→ Table du Khi²

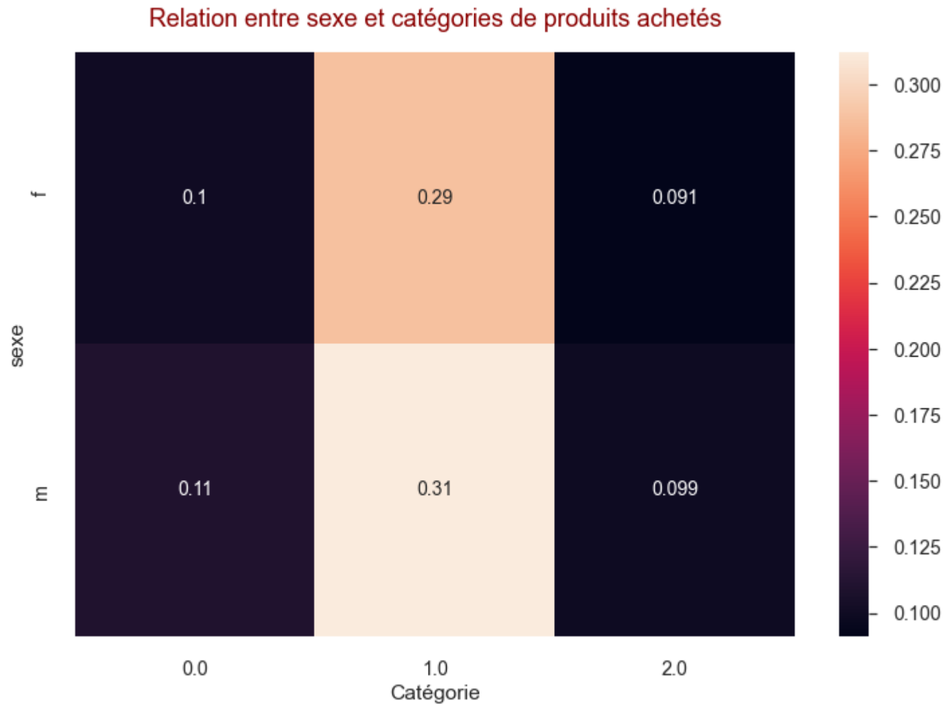
Résultat

Khi² = 8,9656
p_value < 0,025

⇒ On rejette H0
⇒ Les deux variables qualitatives sont dépendantes

Corrélation entre le sexe des clients et les catégories de produits achetés

- Représentation sous forme de heatmap



Liaison plus forte entre les variables 'sexe des clients' et 'catégories achetées' pour les produits de catégorie 1

- Calcul du V de Cramer

Permet de **mesurer le degré de dépendance** des 2 variables qualitatives

$$V = \sqrt{\frac{\chi^2}{\text{Effectif total} \times \min(\text{nombre de lignes} - 1, \text{nombre de colonnes} - 1)}}$$

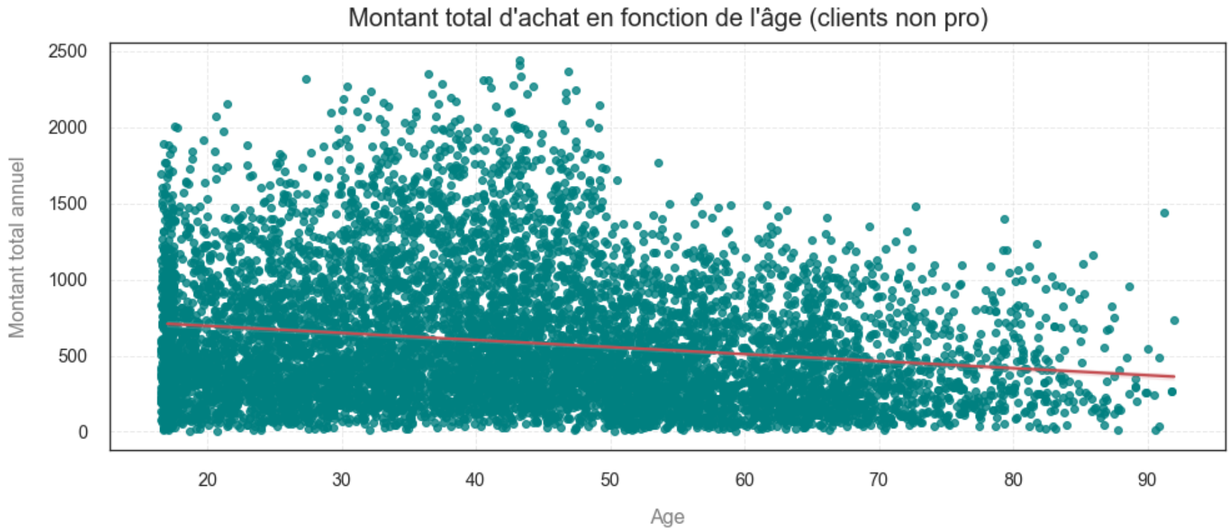
→ plus V est élevé, plus la dépendance entre les variables est importante

Ici, V = 0,01585

↳ V est proche de 0
→ **les variable 'sexe' et 'catégories' ont donc un degré de liaison très faible**

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

- âge des clients : variable **quantitative**
- montant total des achats : variable **quantitative**



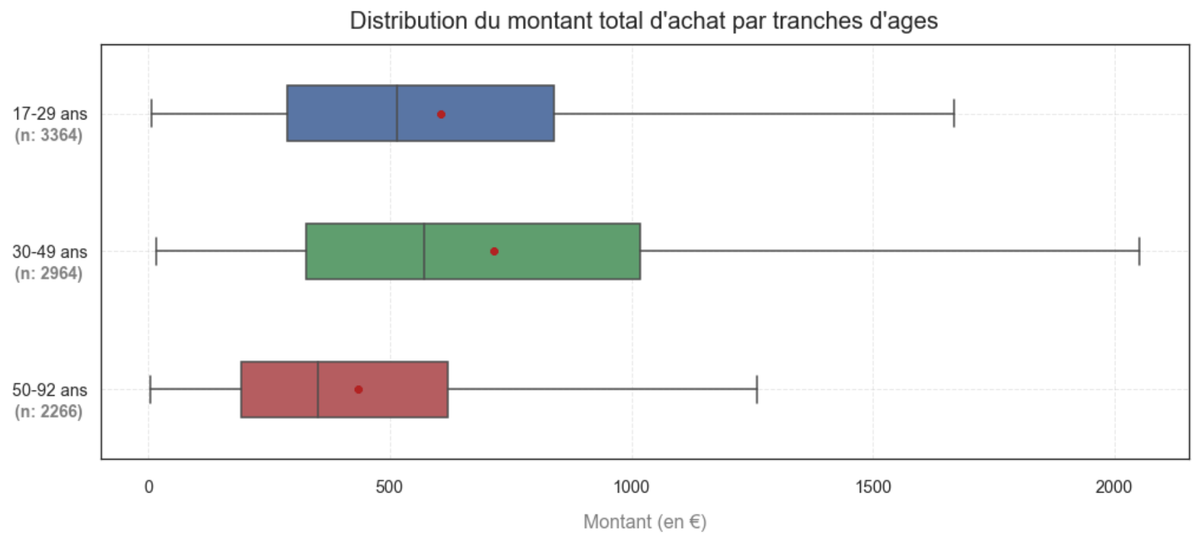
Pearson = - 0,1810 => faible corrélation entre l'âge des clients et le montant total des achats

⇒ Si on répartit les clients en 3 classes d'âges comme précédemment et que l'on calcule le coef de corrélation pour chaque classe

Classe "- de 30 ans"	Pearson = **- 0.0201**
Classe "30-49 ans"	Pearson = **- 0.0187**
Classe "50 ans et +"	Pearson = **- 0.0337**

} Pas de corrélation

- classe d'âge des clients : variable **qualitative**
- montant total des achats : variable **quantitative**



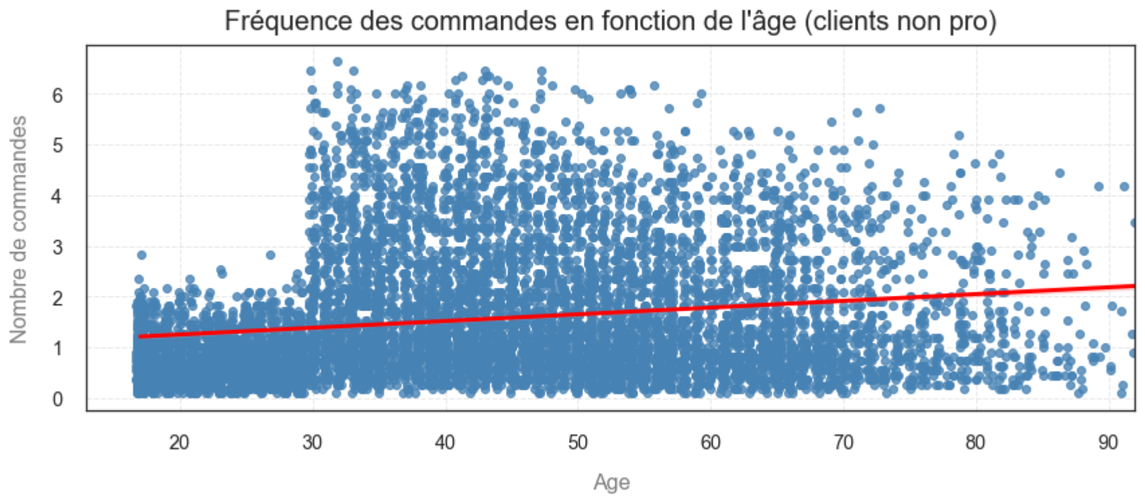
	df	sum_sq	mean_sq	F	PR(>F)	EtaSq
classe	2.0	1.247733e+08	6.238665e+07	361.896422	1.262731e-151	0.077704
Residual	8591.0	1.480986e+09	1.723881e+05	NaN	NaN	NaN

eta² = 0.0777 → Pas de corrélation

p_value ≃ 0 → on rejette H0 au profit de H1 : au moins une classe d'âge a une moyenne significativement différente

Corrélation entre l'âge des clients et la fréquence d'achat

- âge des clients : variable **quantitative**
- fréquence d'achat : variable **quantitative**
  (nombre de commandes annuelles mensualisées)



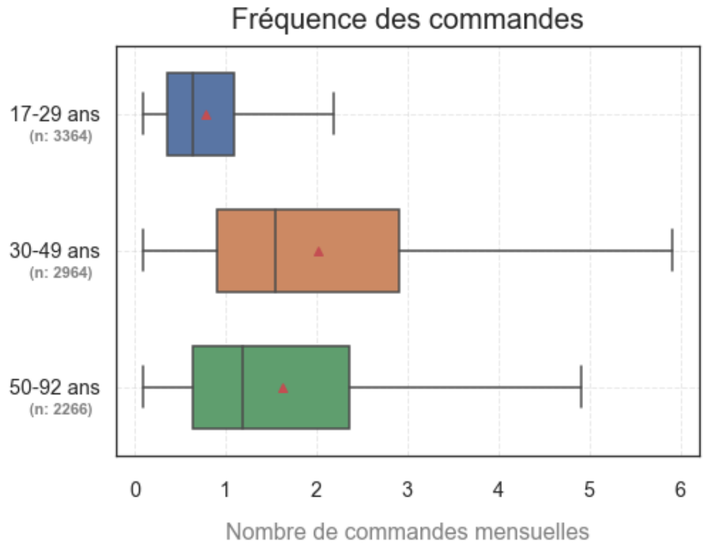
Rappel *(voir diapo 29)*

Pearson = 0,1747 => très faible corrélation entre l'âge des clients et la fréquence d'achat

⇒ Si on répartit les clients en 2 classes d'âges

Classe "- de 30 ans" Pearson = **- 0.0191**
Classe "30 ans et +" Pearson = **- 0.1364**
} Pas de corrélation

- classe d'âge des clients : variable **qualitative**
- fréquence d'achat : variable **quantitative**



	df	sum_sq	mean_sq	F	PR(>F)	EtaSq
classe	2.0	19508.301353	9754.150676	1687.936227	0.0	0.282101
Residual	8591.0	49645.186312	5.778744	NaN	NaN	NaN

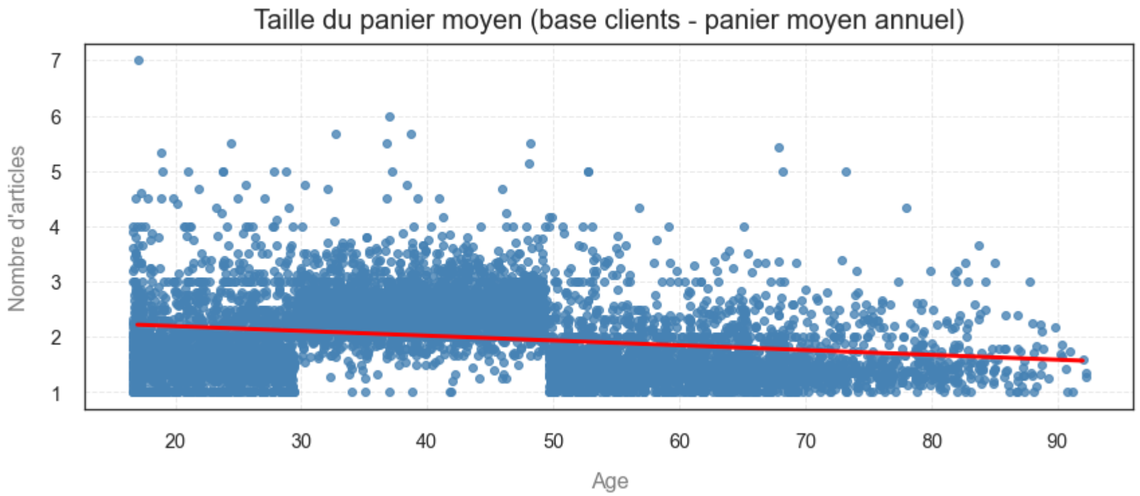
eta² = 0.2821 → faible corrélation entre les classes d'âges et le nombre de commandes

p_value ≃ 0 → on rejette H0 au profit de H1 : au moins une classe d'âge a une moyenne significativement différente

Corrélation entre l'âge des clients et la taille du panier moyen (en nb d'art.)

'panier moyen' :
soit panier annuel moyen pour chaque client
soit moyenne de l'ensemble des paniers

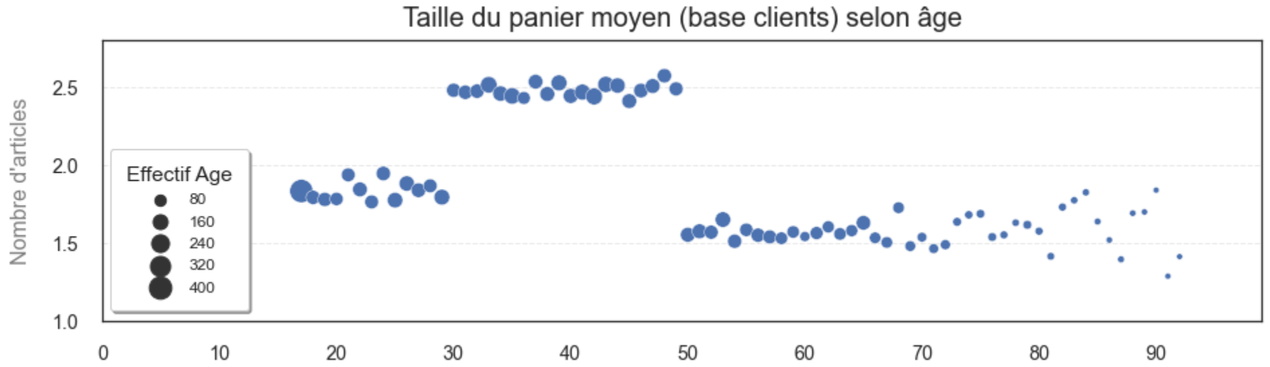
- âge des clients : variable **quantitative**
- taille du panier moyen : variable **quantitative**



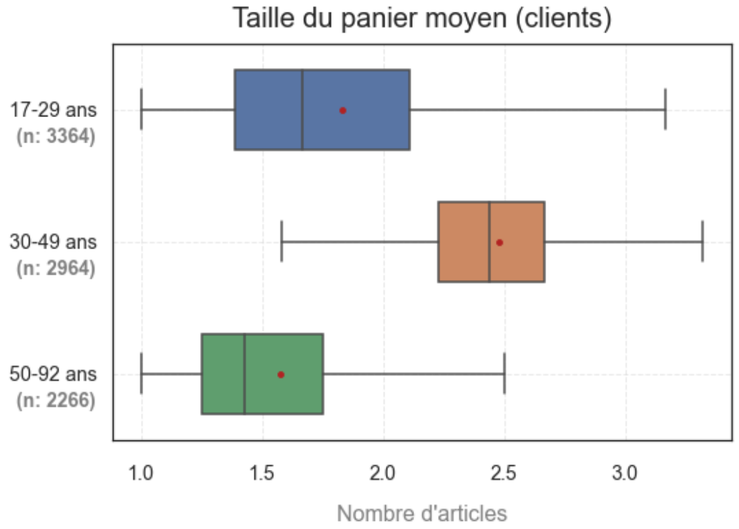
Droite de régression y = 2.3679 - 0.009β
Coef. de corrélation **R = - 0.2213**

=> **Faible corrélation négative** entre l'âge des clients et la taille du panier moyen

→ on agrège les données par âge



On retrouve les mêmes 3 groupes d'âges



- classes d'âges des clients :
  → variable **qualitative**
- taille du panier moyen :
  → variable **quantitative**

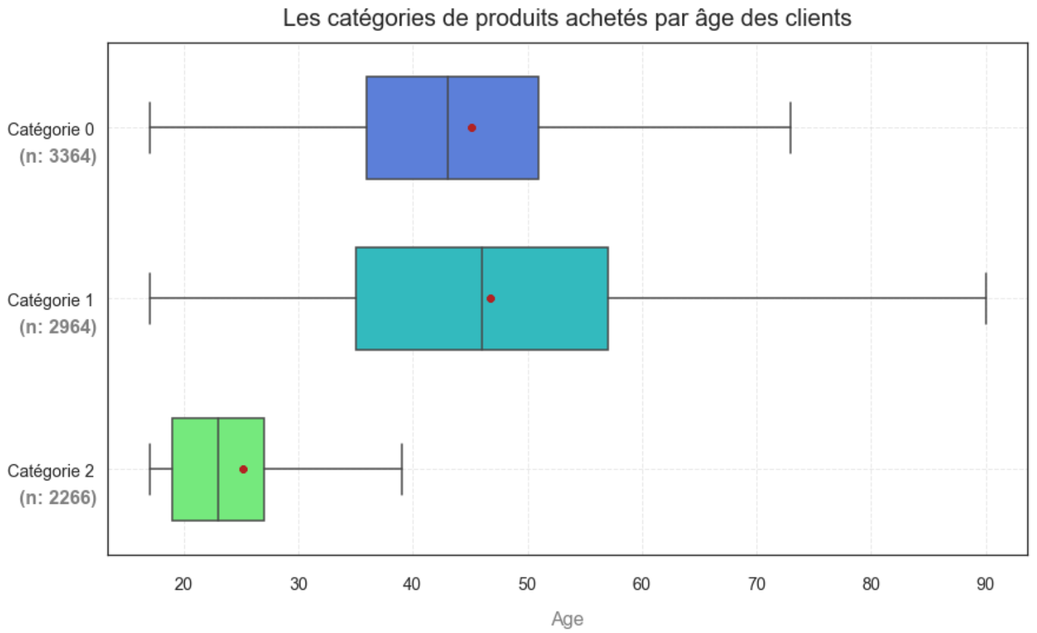
	df	sum_sq	mean_sq	F	PR(>F)	EtaSq
classe	2.0	1373.547156	686.773578	2480.155076	0.0	0.366039
Residual	8591.0	2378.912459	0.276908	NaN	NaN	NaN

eta² = 0.3660 → corrélation

p_value ≃ 0 → on rejette H0 au profit de H1 : au moins une classe d'âge a une moyenne significativement différente

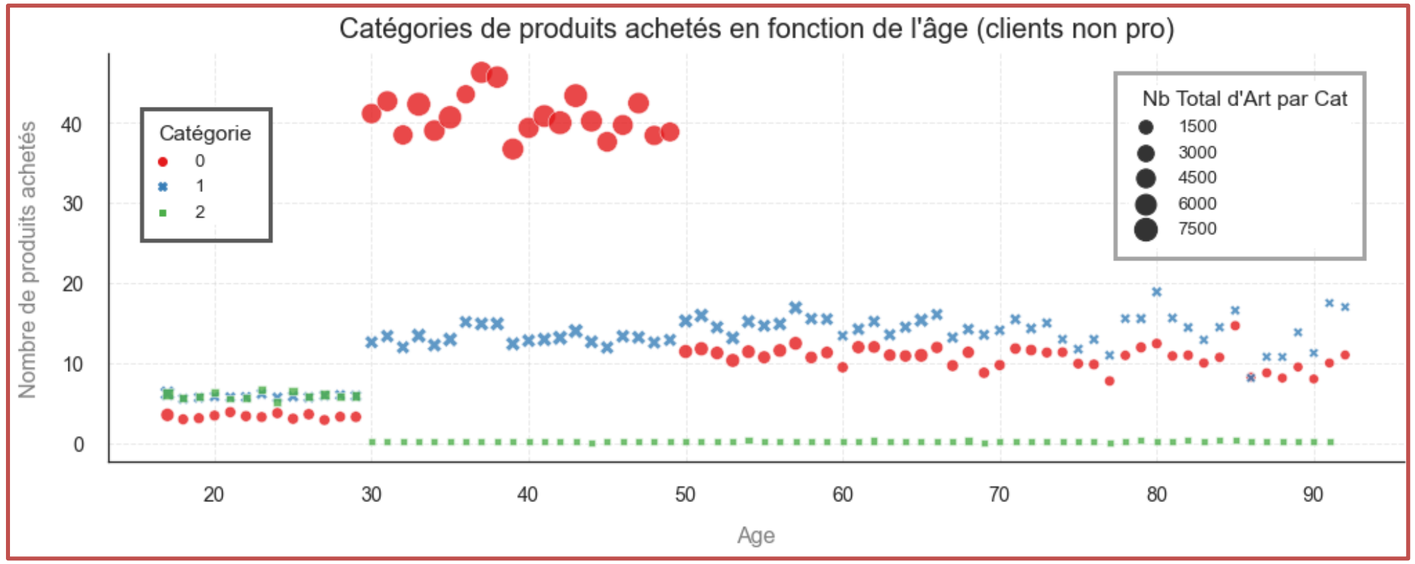
Corrélation entre l'âge des clients et les catégories de produits achetés

- âge des clients : variable **quantitative**
- catégories achetées : variable **qualitative**



→ on agrège les données par âge, en calculant le nombre d'articles commandés par catégorie en moyenne

	categ	age	effectif_age	nb_art_cat_total	nb_art_cat_moyen
0	0	17.0	437	1534.0	3.510297
1	1	17.0	437	2711.0	6.203661
2	2	17.0	437	2725.0	6.235698
3	0	18.0	145	429.0	2.958621



eta² = 0.1216 → pas de corrélation

p_value ≃ 0 → on rejette H0 au profit de H1 :
↳au moins une catégorie de produits a une moyenne significativement différente

Répartition des clients en 3 groupes d'âges, aux comportements spécifiques :

- les moins de 30 ans : cat 2 ++
- les 30-49 ans : cat 0 ++ / cat 1 +
- les 50 ans et plus : cat 0 -

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

- classe d'âge des clients :
  → variable **qualitative**
- catégories achetées :
  → variable **qualitative**

```
st.chi2_contingency(cont_test)
```

```
(80058.50215280065,
 0.0,
 4,
 array([[12523.92667708, 10350.07740468,  1601.99591824],
        [52357.34079531, 43269.37899805,  6697.28020664],
        [31393.73252761, 25944.54359727,  4015.72387512]]))
```

Khi² = 80058,50

p_value ≃ 0

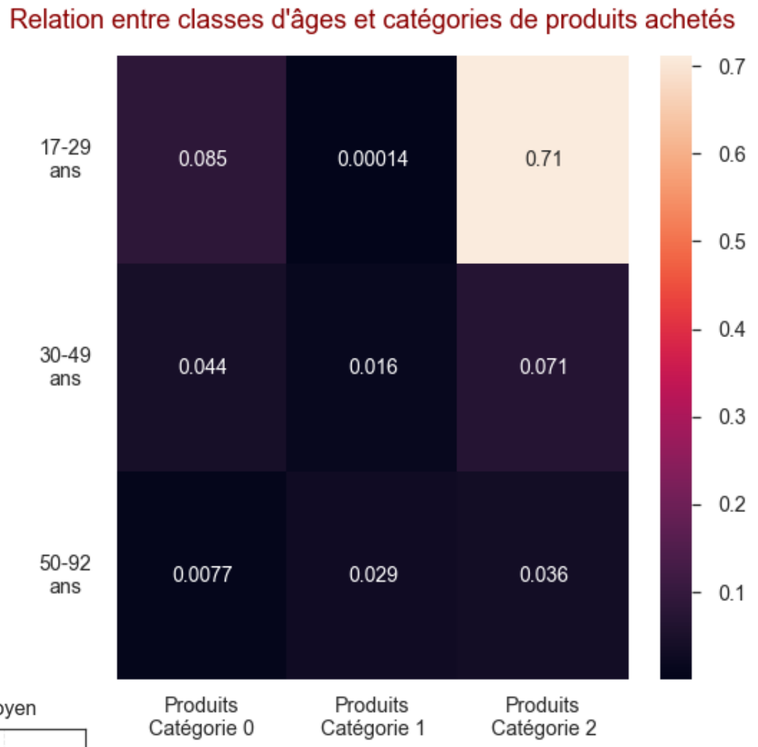
p_value ≃ 0 → on rejette H0 au profit de H1 :

↳Les variables 'âge' et 'catégories de produits' sont dépendantes

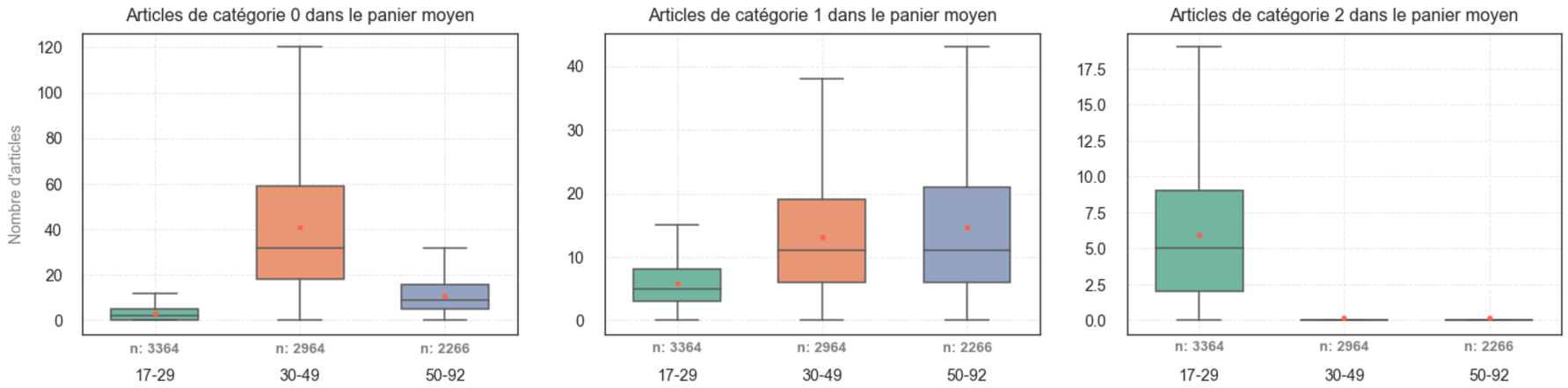
Khi-2 = 80058.50215280065
Effectif total : 188154
la plus petite dimension : 3

V = 0,4612

Important degré de dépendance entre les classes d'âges et les catégories achetées



Répartition du panier moyen par catégorie (base clients)



↳ **Liaison forte** entre clients de moins de 30 ans et produits de catégorie 2

Conclusion

Corrélations obtenues

Variable	Variable	Mesure de la relation	Relation
Âge	Montant total	R = - 0,18	Corrélation négative
Âge	Montant panier	R = - 0,33	Corrélation négative
Âge	Taille panier	R = - 0,22	Corrélation négative
Âge	Nombre de commandes	R = 0,17	Corrélation positive
Âge	Catégories	V = 0,4612	Liaison forte
Sexe	Catégories	V = 0,1585	Liaison faible

Ainsi, selon les corrélations obtenues, plus les clients sont âgés,
- moins leur montant annuel d'achat est important ;
- plus leur fréquence d'achat est importante
- moins leur panier moyen, en volume et en valeur, est important

La répartition des achats entre les catégories de produits achetés s'explique davantage par l'âge que par le sexe

Mais il semble plus pertinent de classer nos clients en 4 catégories, aux comportements spécifiques :

- **Les professionnels : 4 clients → 7,5% du CA**
  - Fréquence d'achat et montant total d'achat très importants

- **Les 30-49 ans : 35% des clients → 44% du CA**
  - Gros acheteurs de produits de catégorie 1
  - Taille du panier importante + Fréquence d'achat élevée => montant d'achat total important

- **Les moins de 30 ans : 40% des clients → 25% du CA**
  - Acheteurs quasi exclusifs des produits de catégorie 2
  - Faible fréquence d'achat mais montant du panier moyen élevé

- **Les 50 ans et plus : 25% des clients → 23,5% du CA**
  - Achat de produits de catégories 0 et 1 (en quantités équivalentes)
  - Taille et montant du panier moyen faibles => montant total d'achat faible

MERCI POUR VOTRE ATTENTION

QUESTIONS ?