- Analyse des Ventes -

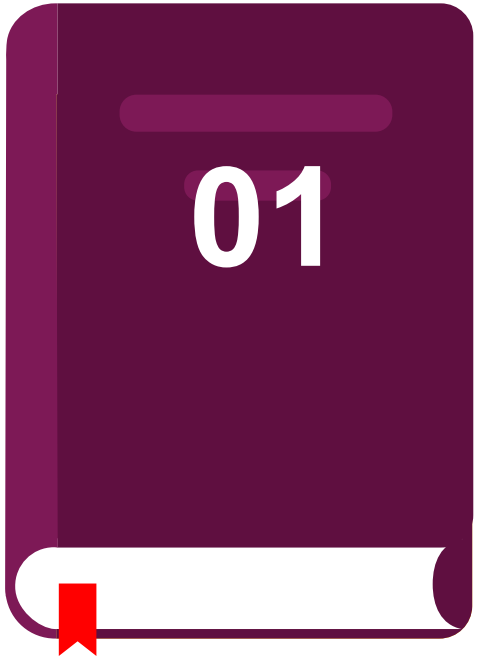
SOMMAIRE

Nettoyage des données

1- Gestion des tests

2- Gestion des NaN

3 - Autres opérations



1- Gestion des tests

1/4 - Identification du problème

```
Entrée [7]: products.describe(include='all')
```

Out[7]:

	id_prod	price	categ
count	3287	3287.000000	3287.000000
unique	3287	NaN	NaN
top	0_2060	NaN	NaN
freq	1	NaN	NaN
mean	NaN	21.856641	0.370246
std	NaN	29.847908	0.615387
min	NaN	-1.000000	0.000000
25%	NaN	6.990000	0.000000
50%	NaN	13.060000	0.000000
75%	NaN	22.990000	1.000000
max	NaN	300.000000	2.000000

2/4 - Recherche d'informations

Entrée [22]:
```
#On verifie le prix -1 que l'on à aperçu auparavant
df.loc[df.price== -1.0]
```

Out[22]:

	id_prod		date	session_id	client_id	sex	age	price	categ
309439	T_0	test_	2021-03-01 02:30:02.237419	s_0	ct_0	f	22	-1.0	0.0
309440	T_0	test_	2021-03-01 02:30:02.237425	s_0	ct_0	f	22	-1.0	0.0
309441	T_0	test_	2021-03-01 02:30:02.237436	s_0	ct_0	f	22	-1.0	0.0
309442	T_0	test_	2021-03-01 02:30:02.237430	s_0	ct_0	f	22	-1.0	0.0
309443	T_0	test_	2021-03-01 02:30:02.237449	s_0	ct_0	f	22	-1.0	0.0
...	...		...	...	...	...	...	...	...
517697	T_0	test_	2021-03-01 02:30:02.237420	s_0	ct_1	m	22	-1.0	0.0
517698	T_0	test_	2021-03-01 02:30:02.237427	s_0	ct_1	m	22	-1.0	0.0
517699	T_0	test_	2021-03-01 02:30:02.237449	s_0	ct_1	m	22	-1.0	0.0
517700	T_0	test_	2021-03-01 02:30:02.237424	s_0	ct_1	m	22	-1.0	0.0
517701	T_0	test_	2021-03-01 02:30:02.237425	s_0	ct_1	m	22	-1.0	0.0

200 rows × 8 columns

3/4 - On retire les lignes « test » et on les place dans un autre dataset

Entrée [23]:
```
#On les stock dans un df appelé test
test = df.loc[df.date.str.contains("test"),:]
```

Entrée [24]:
```
#On les suprimme du df principal
df=df[~df.date.str.contains("test")]
```

4/4 - Conclusion

Entrée [25]:
```
#On verifie si il reste des prix à -1.0 qui ne sont pas des test
df.loc[df.price== -1.0]
#ils n'en restent pas
```

Out[25]:

id_prod	date	session_id	client_id	sex	age	price	categ

Toutes les valeurs négatives étaient des tests.

2 - Gestion des NaN

221 NaN → **0_2245** → *price ?* → imputation par la **moyenne** : 10,63 €

↓

categ ?

↓

categ : 0

3 - Autres opérations

1/2 - Transformation de la colonne « date »

date
2022-05-20 13:21:29.043970
2022-06-18 05:55:31.816994
2023-02-08 17:31:06.898425

date	hour	day_week
2022-05-20	13	Ven
2022-06-18	6	Sam
2023-02-08	18	Merc

2/2 - Création d'un df « commandes »

session_id	client_id	age	price	number_items	date	hour	day_week
s_1	c_329	56	11.99	1	2021-03-01	0	Lun
s_10	c_2218	53	26.99	1	2021-03-01	0	Lun
s_100	c_3854	45	33.72	2	2021-03-01	4	Lun

Analyse univariée

1 - Etude du CA

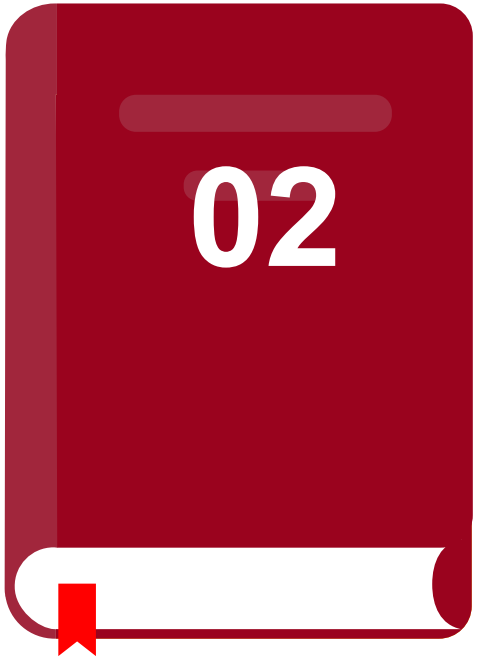
- CA annuel
- CA mensuel
- CA Journalier
- CA par catégorie
- CA mensuel selon la catégorie
- CA journalier selon la catégorie
- CA par heure et catégorie

2 - Etude des produits

- Prix selon la catégorie
- Répartition du CA
- Meilleurs et Pires produits

3 - Etude des clients

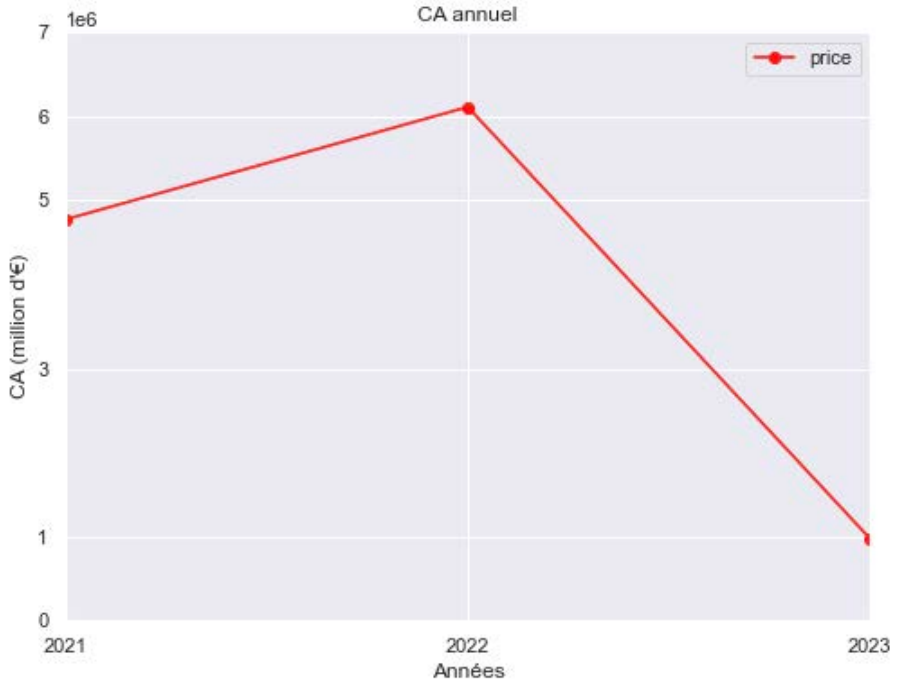
- Âge et sexe
- Visualisations en fonction du sexe
- Répartition du CA
- Meilleurs clients



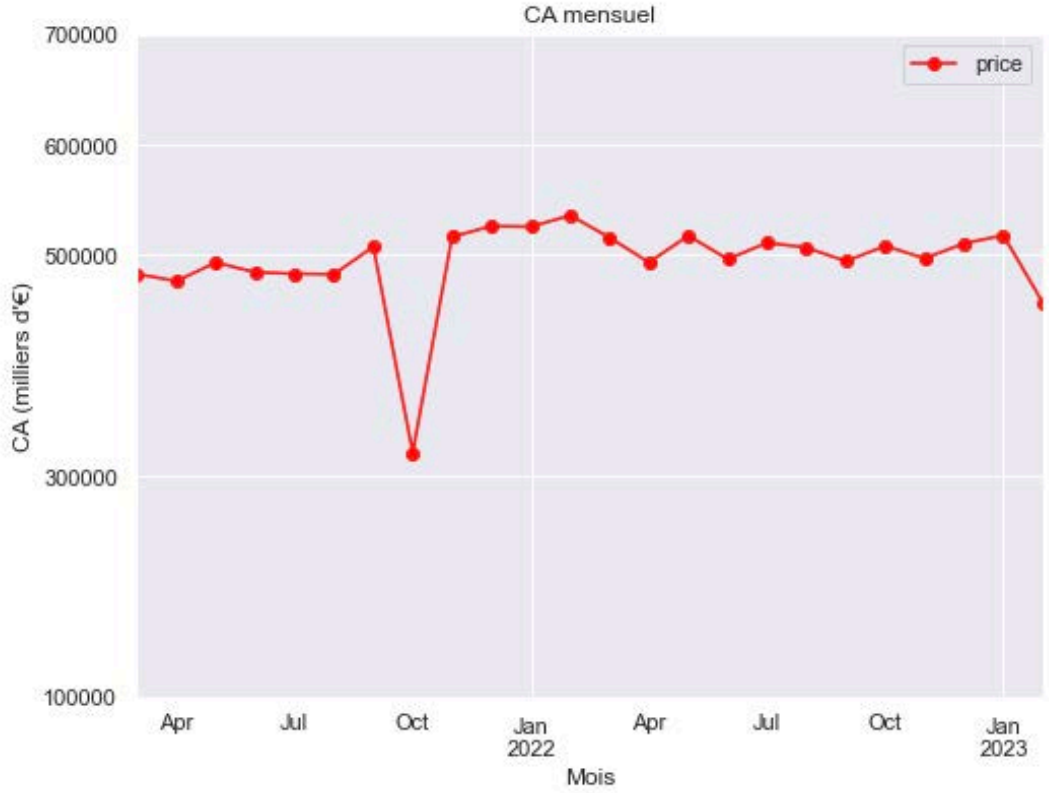
1 - Etude du CA

1 - Etude du CA

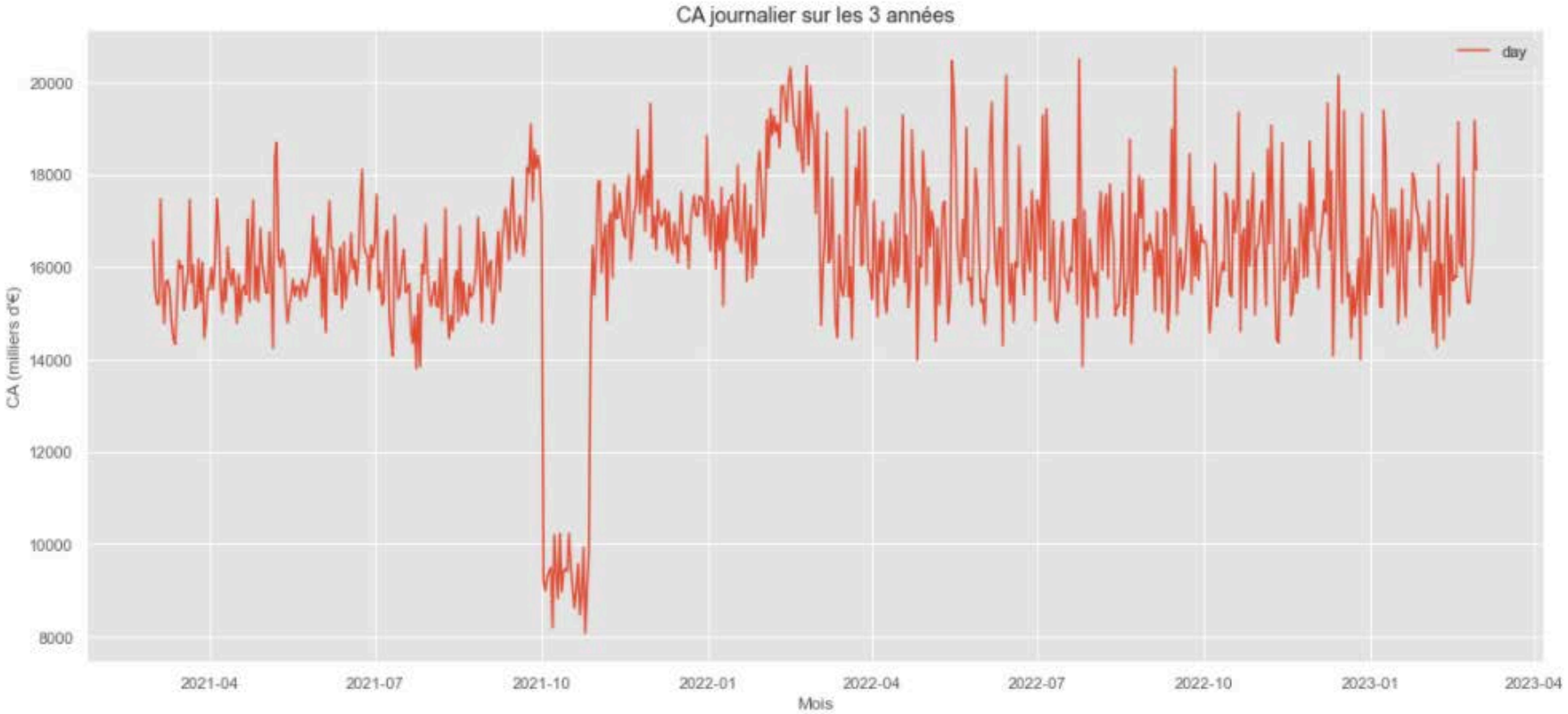
CA annuel 2022: 6,1 Millions €
CA de Mars 2021 à Mars 2023: 11,8 Millions €
CA annuel: 5,9 Millions €



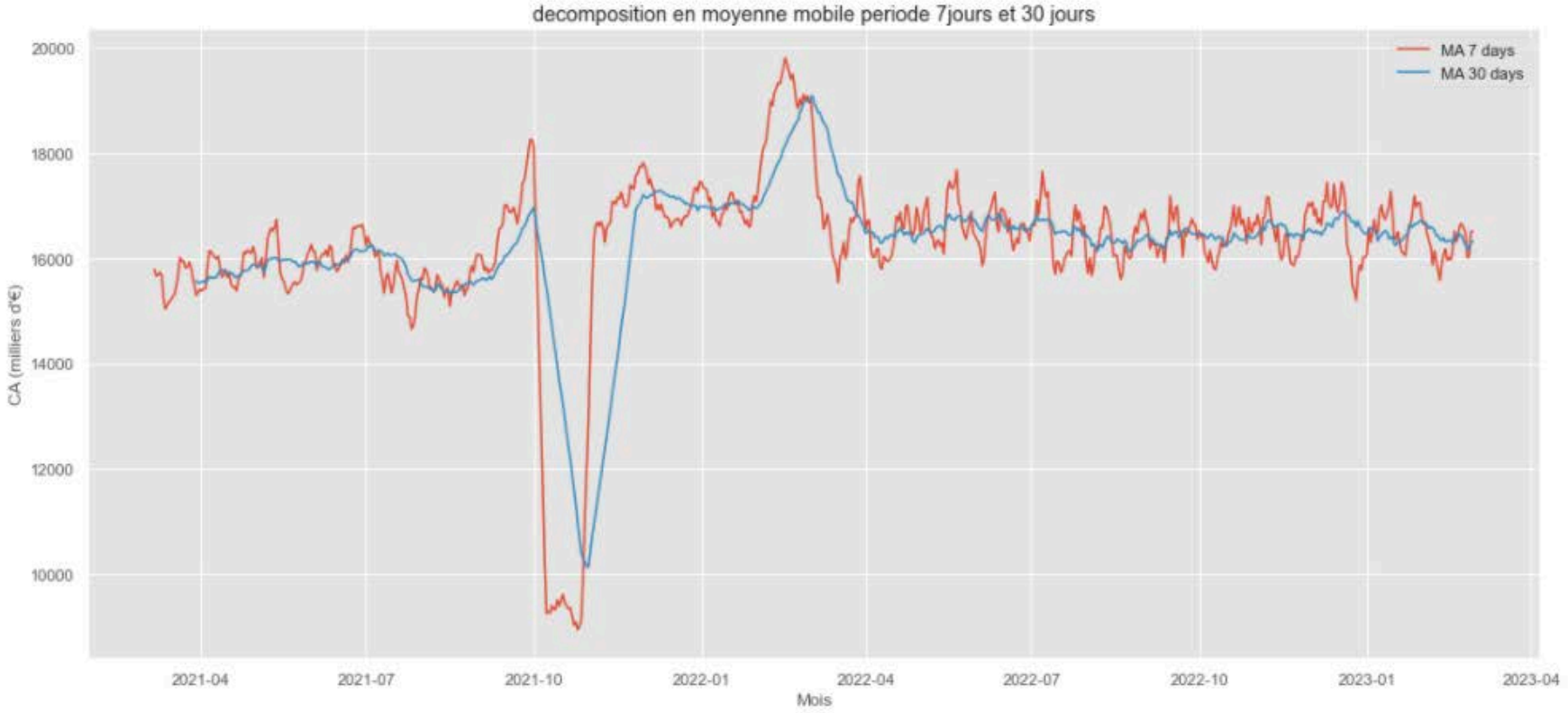
1 - Etude du CA


CA mensuel
price
700000
600000
500000
300000
100000
CA (milliers d'€)
Apr
Jul
Oct
Jan
2022
Apr
Jul
Oct
Jan
2023
Mois

1 - Etude du CA

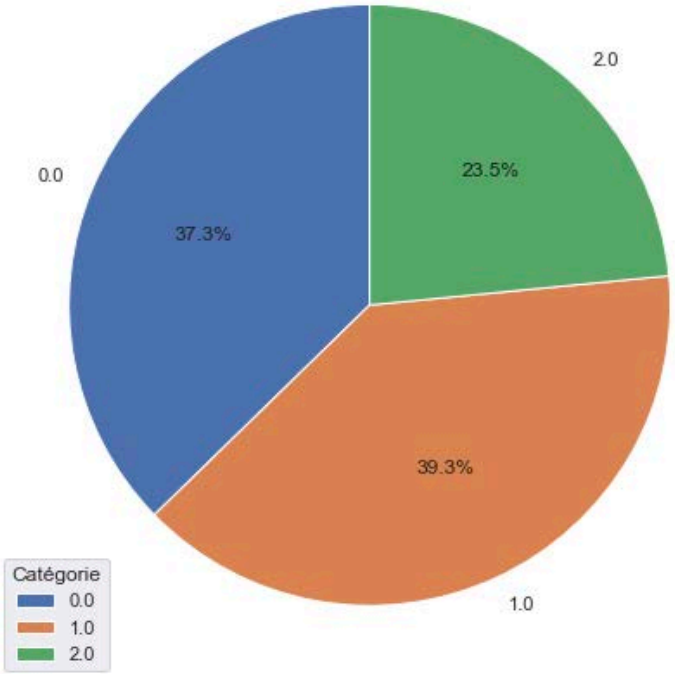

CA journalier sur les 3 années
day
CA (milliers d'€)
20000
18000
16000
14000
12000
10000
8000
2021-04
2021-07
2021-10
2022-01
2022-04
2022-07
2022-10
2023-01
2023-04
Mois

1 - Etude du CA

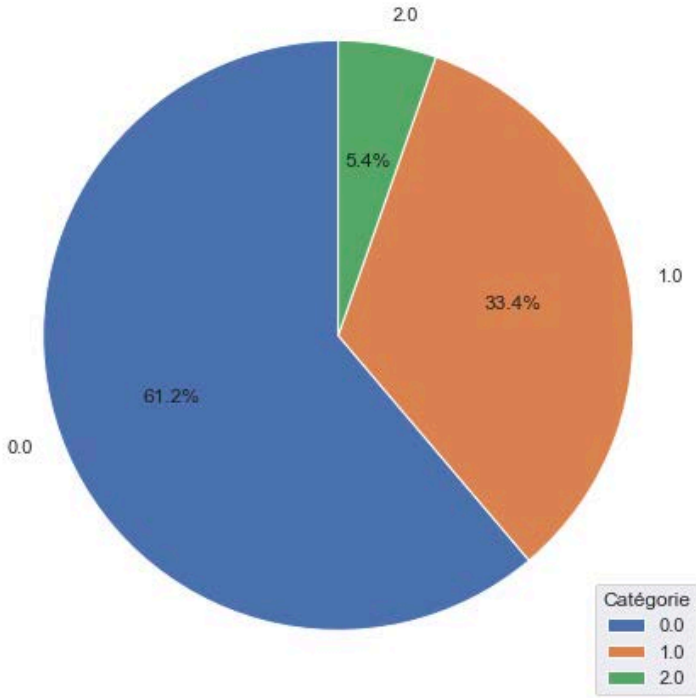

decomposition en moyenne mobile periode 7jours et 30 jours
MA 7 days
MA 30 days
CA (milliers d'€)
20000
18000
16000
14000
12000
10000
2021-04
2021-07
2021-10
2022-01
2022-04
2022-07
2022-10
2023-01
2023-04
Mois

1 - Etude du CA

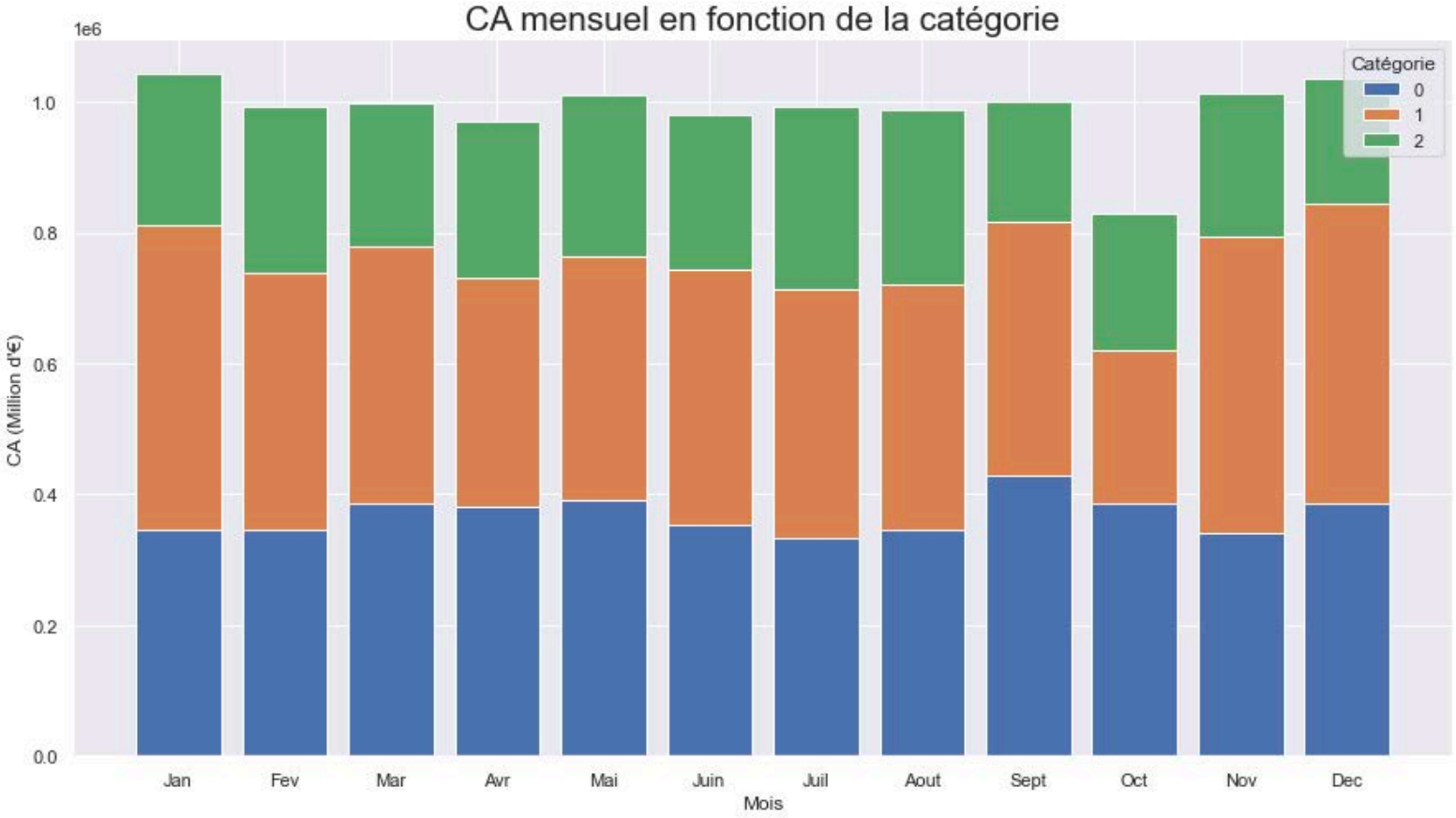
CA en fonction des catégories



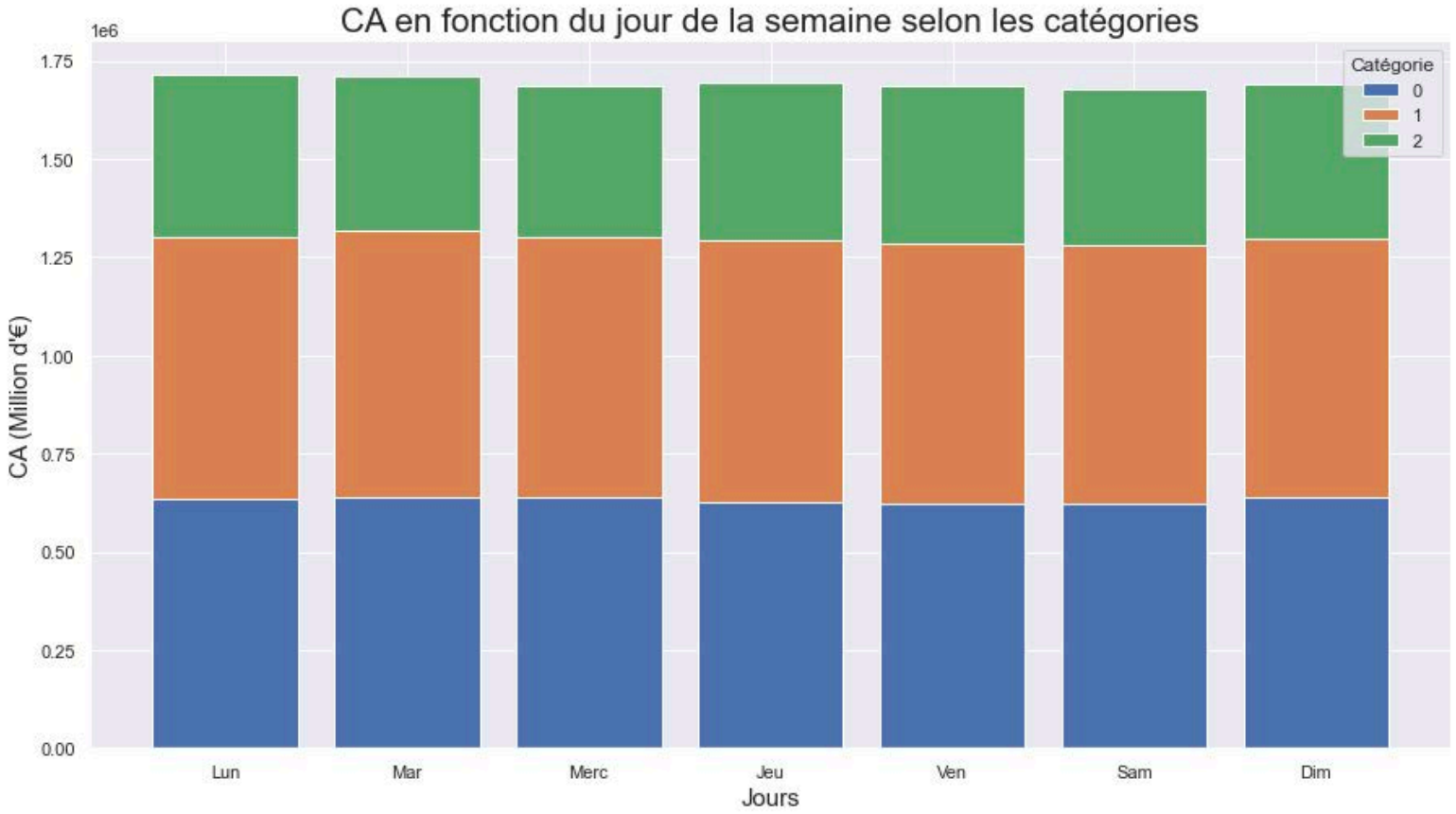
Fréquence d'achat selon les catégories



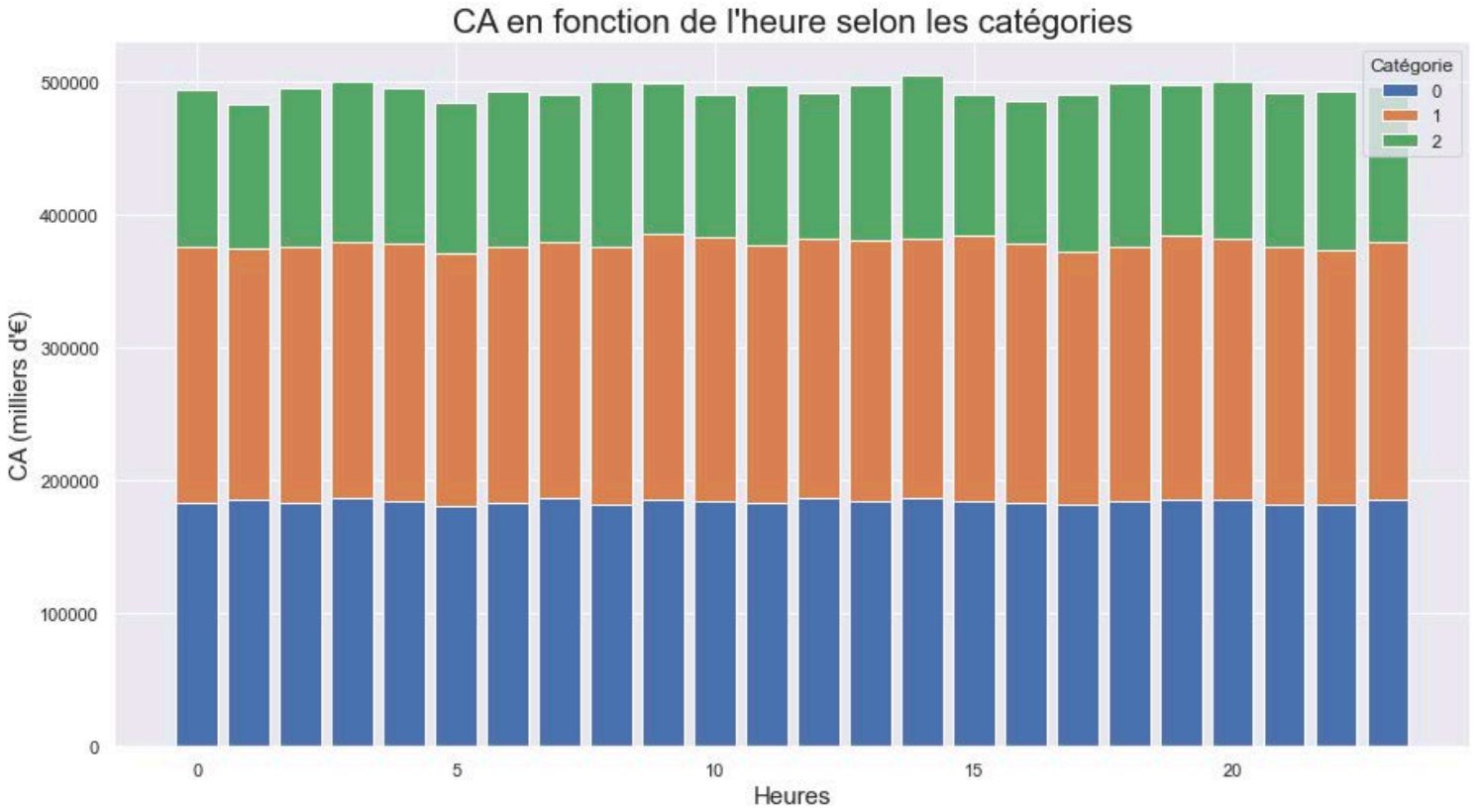
1 - Etude du CA


CA mensuel en fonction de la catégorie
1e6
1.0
0.8
0.6
0.4
0.2
0.0
CA (Million d'€)
Catégorie
0
1
2
Jan
Fev
Mar
Avr
Mai
Juin
Juil
Aout
Sept
Oct
Nov
Dec
Mois

1 - Etude du CA

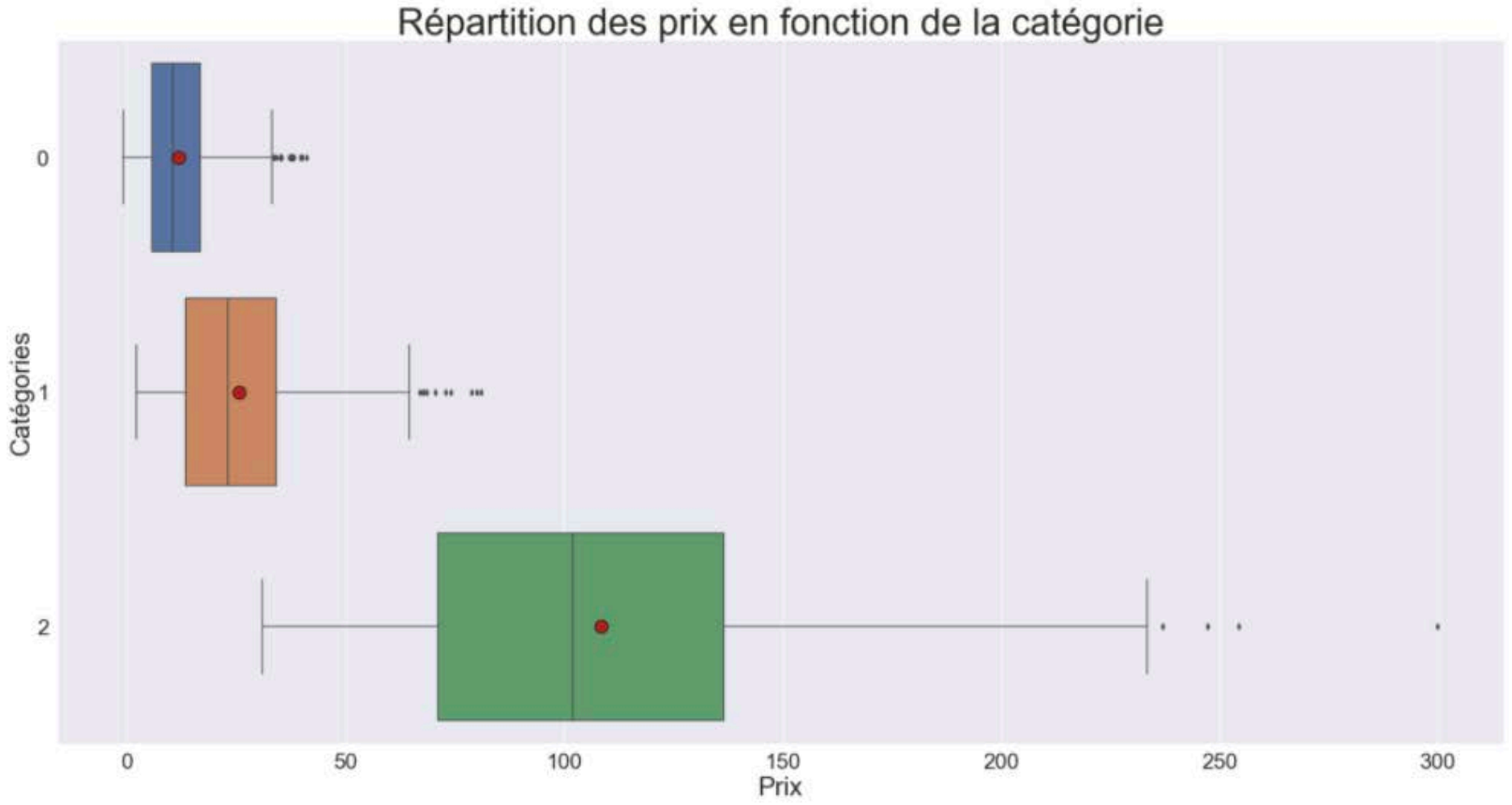

CA en fonction du jour de la semaine selon les catégories
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
CA (Million d'€)
Catégorie
0
1
2
Lun
Mar
Merc
Jeu
Ven
Sam
Dim
Jours

1 - Etude du CA

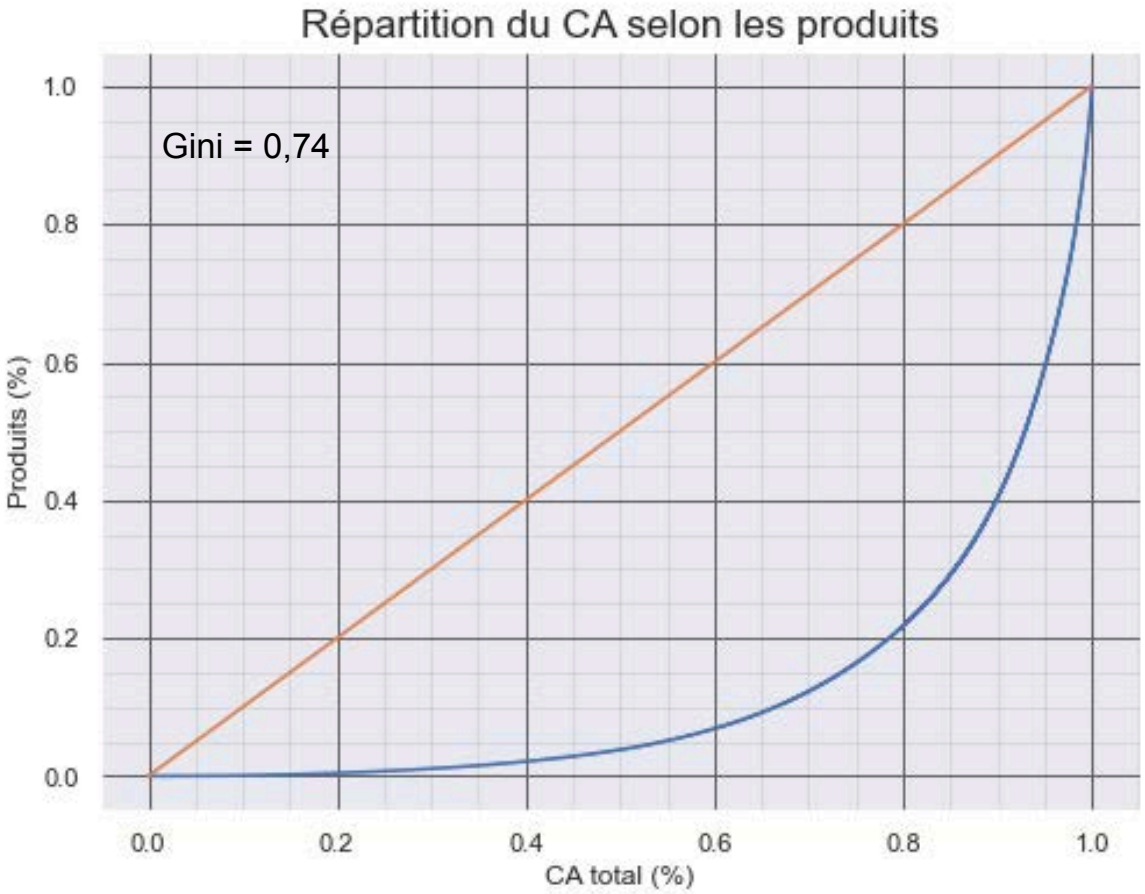

CA en fonction de l'heure selon les catégories
CA (milliers d'€)
500000
400000
300000
200000
100000
0
0
5
10
15
20
Heures
Catégorie
0
1
2

2 - Etude des produits

2 - Etude des produits



2 - Etude des produits



Les meilleurs produits…

selon le montant des achats.

id_prod	number_purchases	monetary_value
2_159	650	94893.50
2_135	1005	69334.95
2_112	968	65407.76
2_102	1027	60736.78
2_209	814	56971.86
1_395	1875	54356.25
1_369	2252	54025.48
2_110	865	53846.25
2_39	915	53060.85
2_166	228	52449.12

selon le nombre d'achats.

id_prod	number_purchases	monetary_value
1_369	2252	54025.48
1_417	2189	45947.11
1_414	2180	51949.40
1_498	2128	49731.36
1_425	2096	35611.04
1_403	1960	35260.40
1_412	1951	32484.15
1_413	1945	34990.55
1_406	1939	48106.59
1_407	1935	30940.65

Les pires produits…

selon le montant des achats.

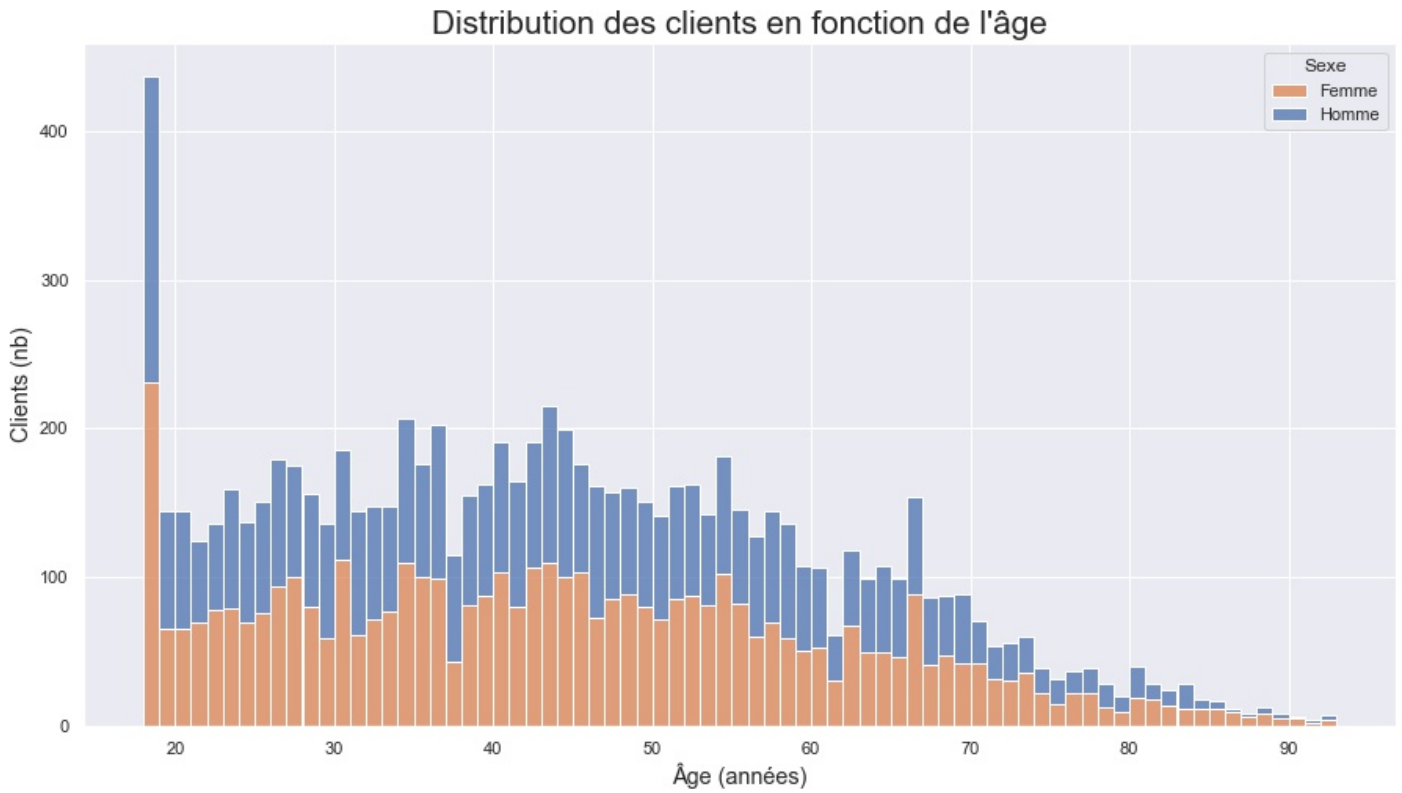
id_prod	number_purchases	monetary_value
0_1539	1	0.99
0_1284	1	1.38
0_1653	2	1.98
0_541	1	1.99
0_807	1	1.99
0_1601	1	1.99
0_1728	1	2.27
0_1498	1	2.48
0_898	2	2.54
0_1840	2	2.56

selon le nombre d'achats.

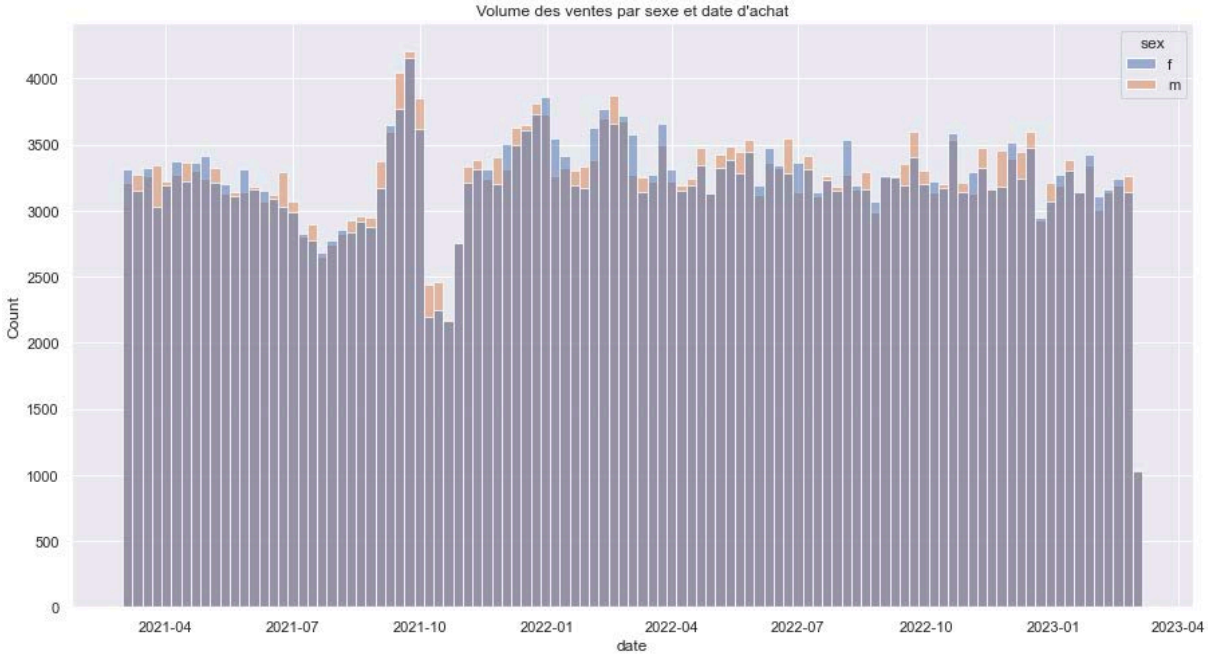
id_prod	number_purchases	monetary_value
0_549	1	2.99
0_2201	1	20.99
2_23	1	115.99
0_1284	1	1.38
0_1683	1	2.99
0_833	1	2.99
2_98	1	149.74
0_1633	1	24.99
0_1601	1	1.99
2_81	1	86.99

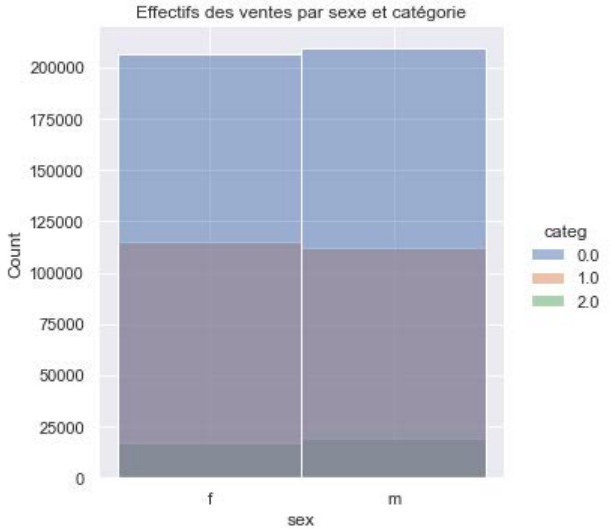
3 - Etude des clients

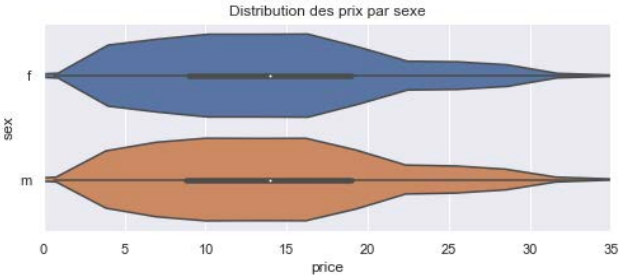
3 - Etude des clients



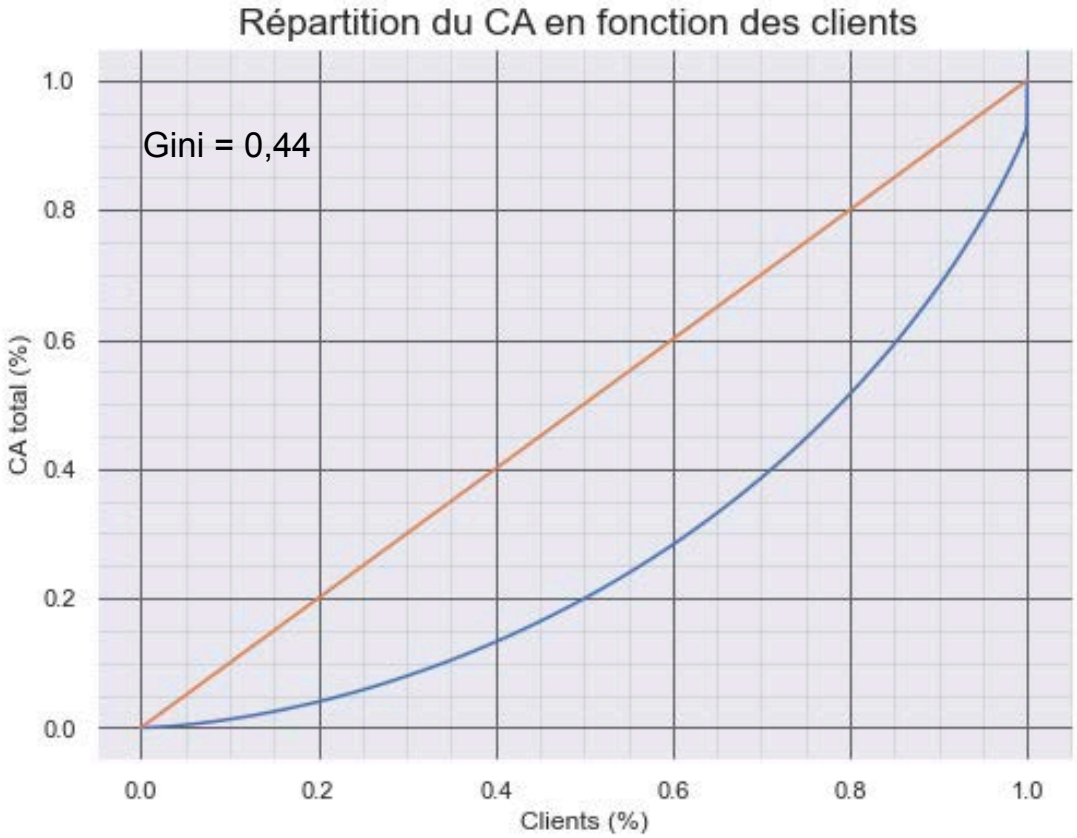
3 - Etude des clients







3 - Etude des clients



Les meilleurs clients…

selon le montant des achats.

	client_id	Recency	Frequency	MonetaryValue
0	c_1609	1	10997	324033.350000
1	c_4958	1	3851	289760.340000
2	c_6714	1	2620	153662.749128
3	c_3454	1	5573	113669.844564
4	c_3263	3	143	5276.870000
5	c_1570	8	158	5271.620000
6	c_2899	8	69	5214.050000
7	c_2140	1	147	5208.820000
8	c_7319	3	145	5155.770000
9	c_8026	3	146	5093.218188

selon la fréquence d'achat.

	client_id	Recency	Frequency	MonetaryValue
0	c_1609	1	10997	324033.350000
1	c_3454	1	5573	113669.844564
2	c_4958	1	3851	289760.340000
3	c_6714	1	2620	153662.749128
4	c_8526	8	165	3975.060000
5	c_1637	5	164	4698.870000
6	c_669	3	163	4499.360000
7	c_2265	1	163	3271.280000
8	c_682	1	161	4102.180000
9	c_8510	4	161	4798.630000

Analyse des corrélations

1 - Sexe et catégorie

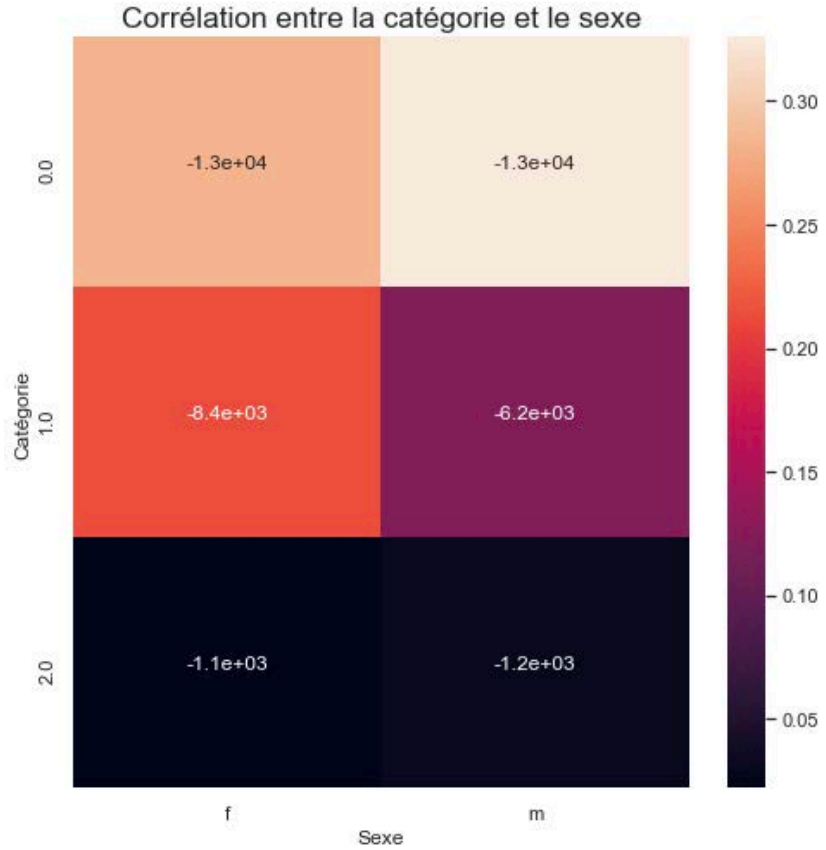
2 - Âge et taille du panier

3 - Âge et montant total

4- Âge et fréquence d’achat

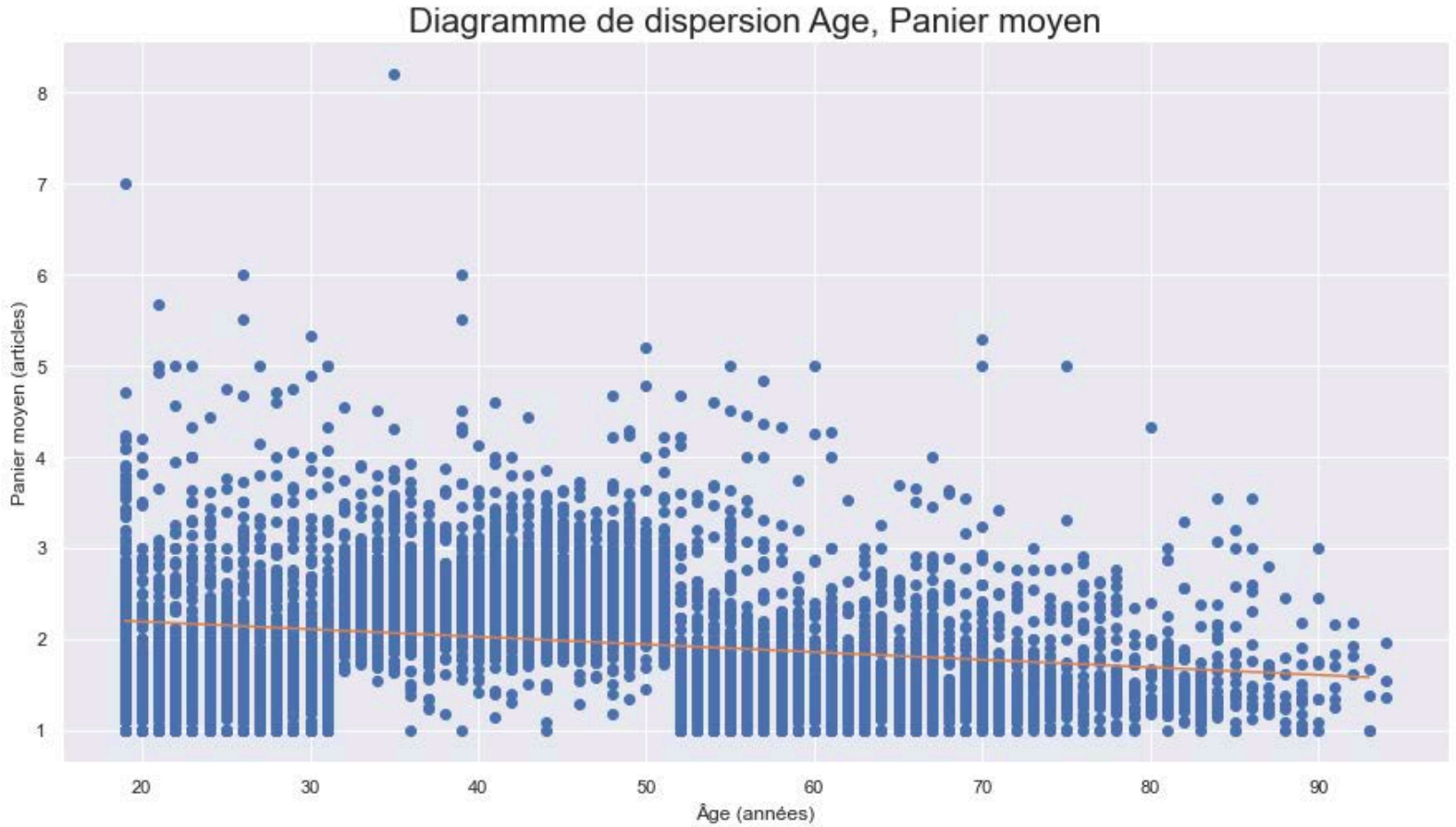
5 - Âge et catégorie

1 - Sexe et catégorie

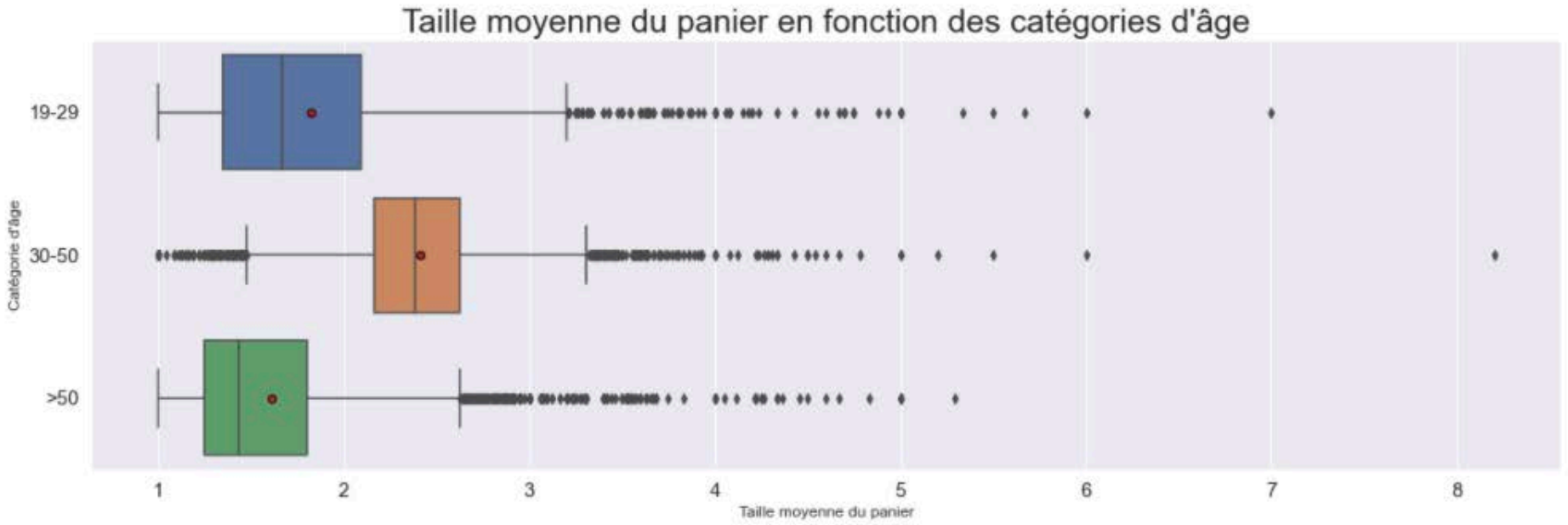


2 - Âge et taille du panier

2 - Âge et taille du panier



2 - Âge et taille du panier



2 - Âge et taille du panier

> : H0 vérifiée :
pas de corrélation

R < ? > seuils de Cohen

p-value < ? > seuil (0.05)

< : H1 vérifiée :
corrélation

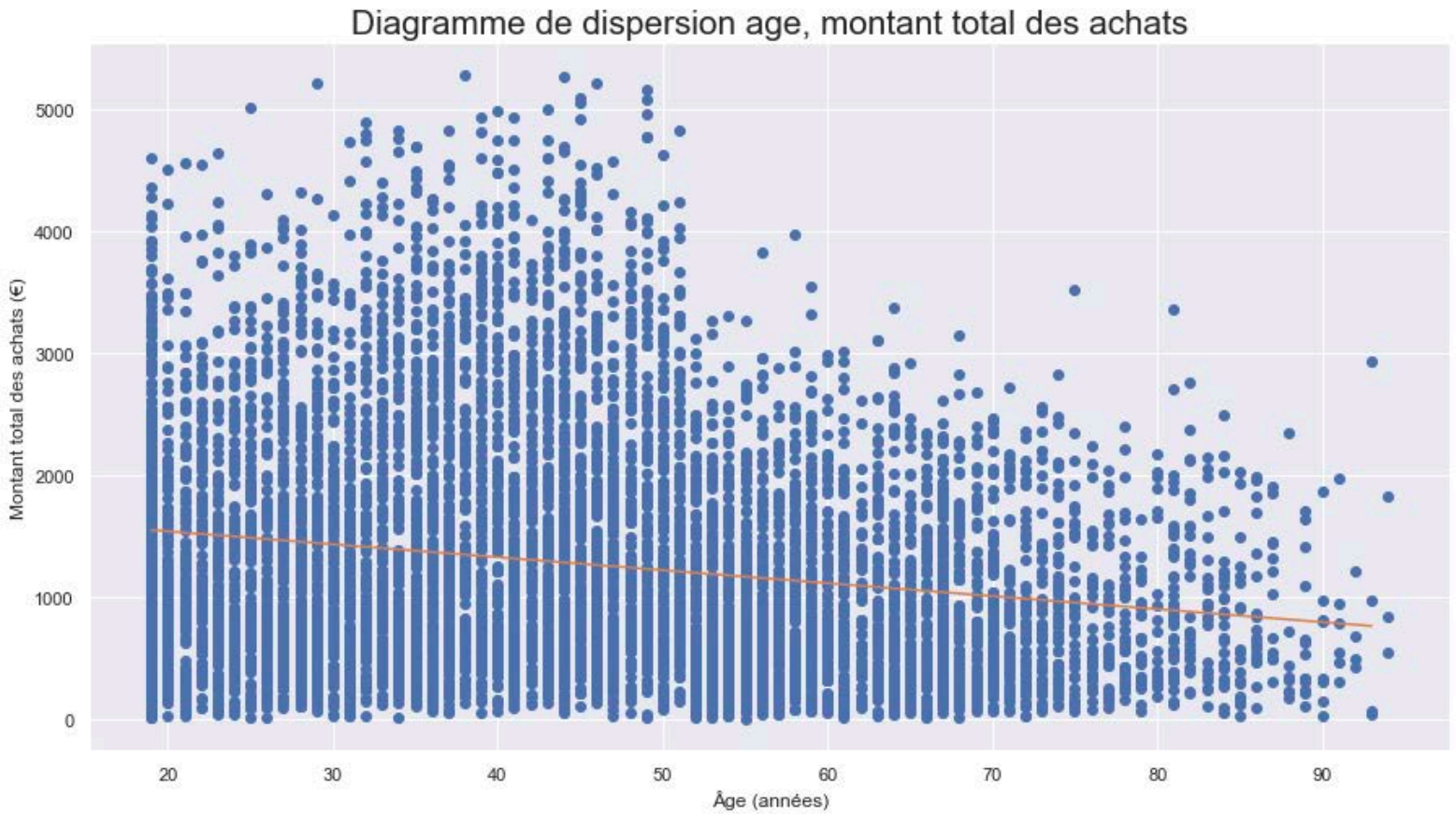
- faible
- moyenne
- forte

Ici :
p-value = 0
R = 0,54

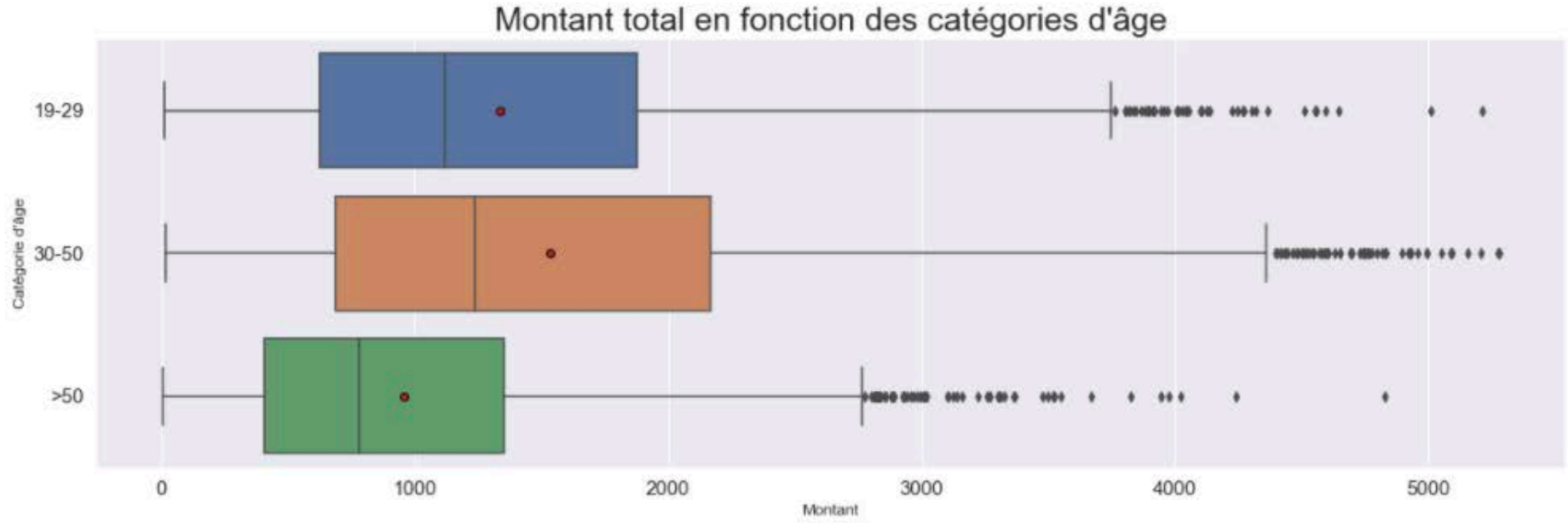
On rejette H0 : il y a une corrélation forte entre la catégorie d'âge et la taille du panier.

3 - Âge et montant total

3 - Âge et montant total



3 - Âge et montant total



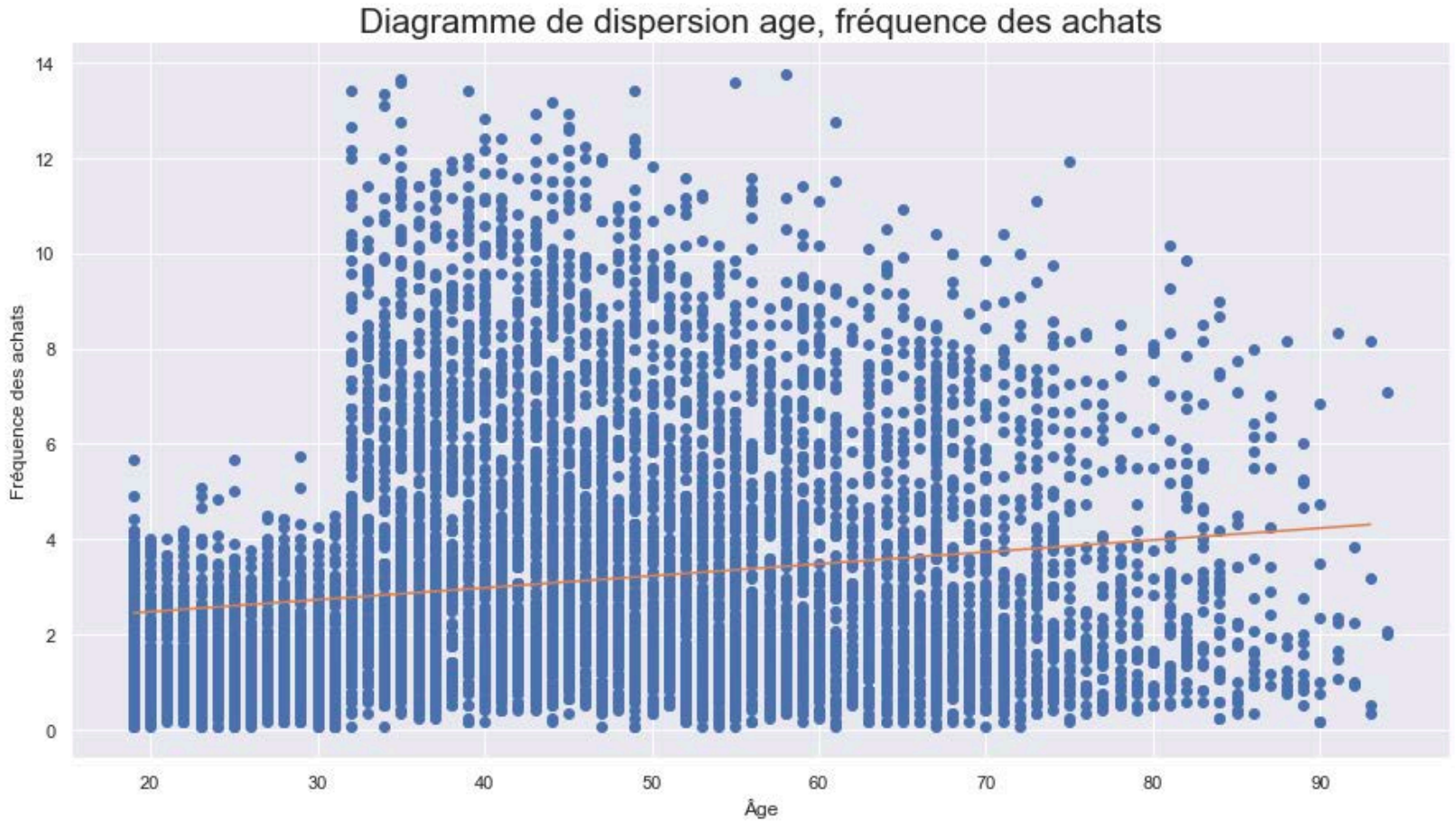
3 - Âge et montant total

Ici :
p-value = 3,12e-138
R = 0,26

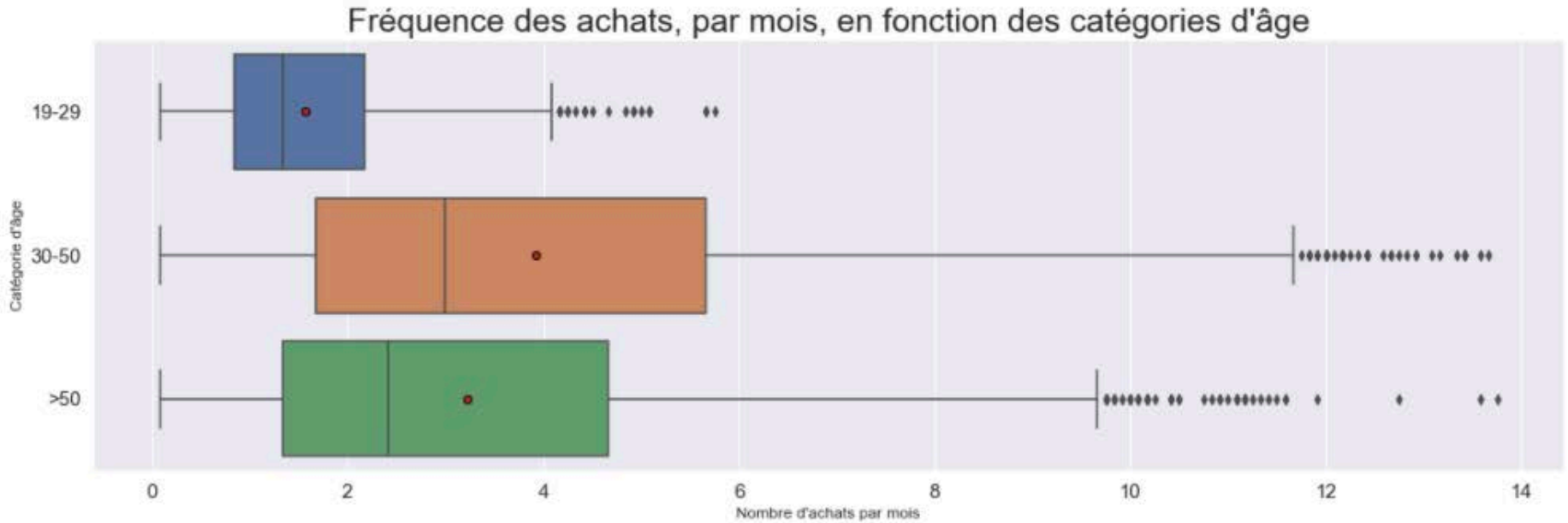
On rejette H0 : il y a une corrélation faible entre la catégorie d'âge et la taille du panier.

4- Âge et fréquence d’achat

4- Âge et fréquence d’achat



4- Âge et fréquence d’achat



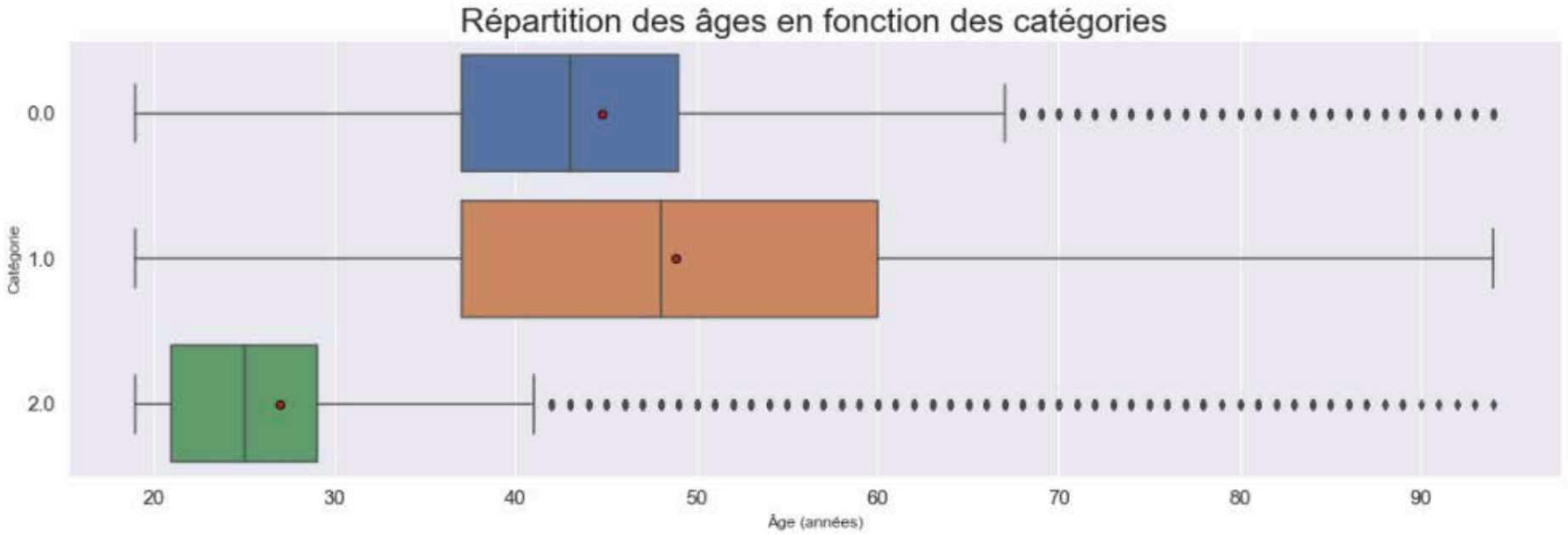
4- Âge et fréquence d’achat

Ici :
p-value = 2,48e-255
R = 0,35

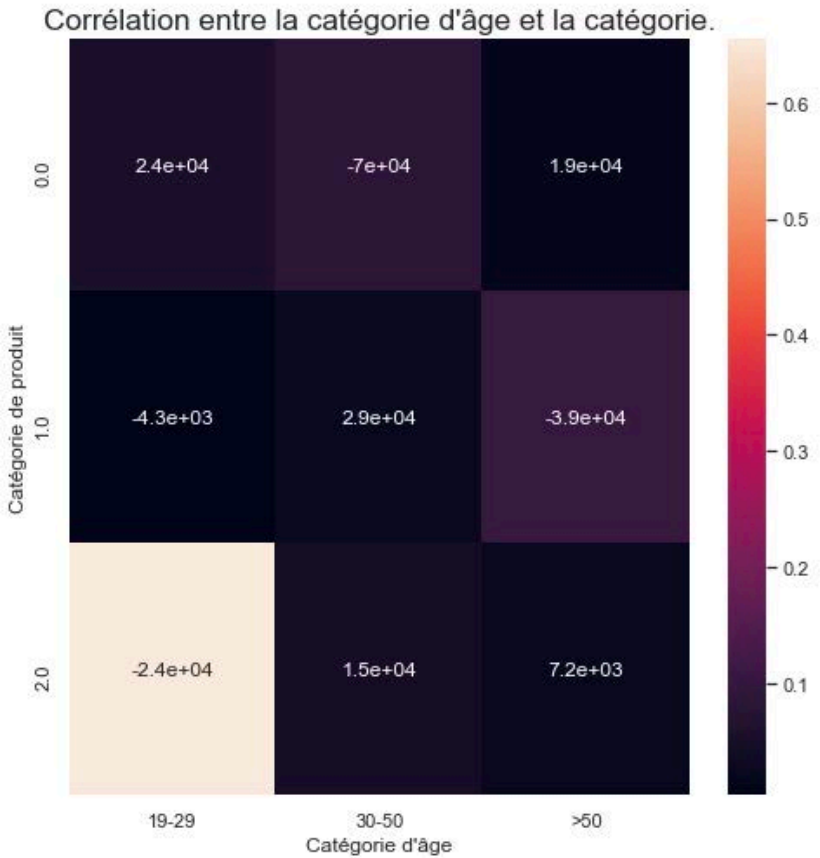
On rejette H0 : il y a une corrélation moyenne entre la catégorie d’âge et la taille du panier.

5 - Âge et catégorie

5 - Âge et catégorie



5 - Âge et catégorie



Conclusion

Conclusion générale

Conclusion des corrélations

Conclusion générale

- Gamme de prix différente selon les catégories

- Stabilité des ventes : selon les mois, les jours, les heures

- Equilibre (CA) des différentes catégories (pas de catégorie particulièrement faible)

Conclusion des corrélations

- Corrélation variable entre le sexe et la catégorie :
  - nulle concernant la catégorie 2
  - très faible pour la catégorie 1
  - faible pour la catégorie 0

- Pas de corrélation linéaire entre les ventes et l'âge, mais corrélation entre les ventes et la catégorie d'âge

- Corrélation forte entre la catégorie 2 et la plus jeune tranche d'âge

Probabilité (Question supplémentaire)

Quel est la probabilité qu'un client achète la référence 0_525 sachant qu'il a acheté la référence 2_159 ?

```
df_2_159 = df[df['id_prod'] == '2_159']
df_0_525 = df[df['id_prod'] == '0_525']
df_2_159 =df_2_159['client_id'].drop_duplicates()
df_0_525 =df_0_525['client_id'].drop_duplicates()
nb_client_commun = len(df_0_525.isin(df_2_159))
proba=(nb_client_commun/len(df_2_159))*100
```

```
#On selectionne toutes les références 2_159
#On selectionne toutes les références 0_525
#On selectionne les clients unique de la référence 2_159
#On selectionne les clients unique de la référence 0_525
#On selectionne les clients unique qui ont commandés les deux références
#On calcul la probabilité
```

```
#On affiche la probabilité:`
```

La Probabilité qu'un client achète la référence 0_525 sachant qu'il a acheté la référence 2_159 est de: 86.5 %