RESTER LIVRES

- Analyse des Ventes -

Alexandre Monod pour OpenClassrooms le 27/11/2020

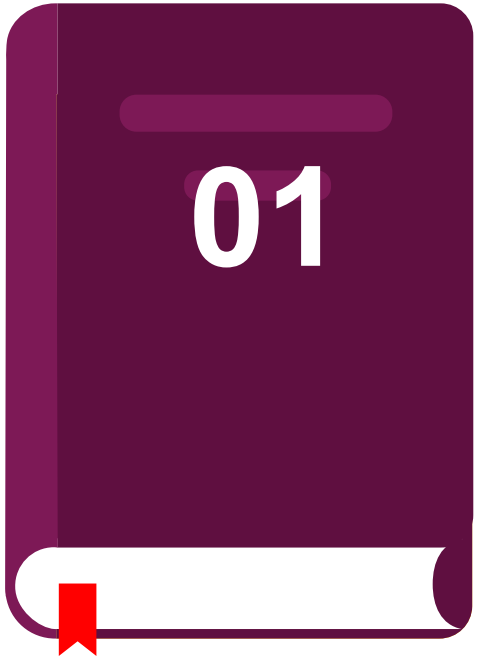
SOMMAIRE

Nettoyage des données

1- Gestion des tests

2- Gestion des NaN

3 - Autres opérations



1- Gestion des tests

1/4 - Identification du problème

```
Entrée [17]: data = pd.merge(data_temp, products, how="left")
             data.describe()
```

Out[17]:

	birth	price	categ
count	337016.000000	336913.000000	336913.000000
mean	1977.837150	17.204376	0.429900
std	13.531686	17.855658	0.590999
min	1929.000000	-1.000000	0.000000
25%	1971.000000	8.580000	0.000000
50%	1980.000000	13.900000	0.000000
75%	1987.000000	18.990000	1.000000
max	2004.000000	300.000000	2.000000

2/4 - Recherche d'informations

```
Entrée [19]: data.loc[data.price== -1.0]
```

Out[19]:

	id_prod	date	session_id	client_id	sex	age	price	categ
98695	T_0	test_2021-03-01 02:30:02.237420	s_0	ct_1	m	21	-1.0	0.0
98696	T_0	test_2021-03-01 02:30:02.237446	s_0	ct_1	m	21	-1.0	0.0
98697	T_0	test_2021-03-01 02:30:02.237414	s_0	ct_1	m	21	-1.0	0.0
98698	T_0	test_2021-03-01 02:30:02.237434	s_0	ct_1	m	21	-1.0	0.0
98699	T_0	test_2021-03-01 02:30:02.237412	s_0	ct_1	m	21	-1.0	0.0
...	...	...	...	...	...	...	...	...
217612	T_0	test_2021-03-01 02:30:02.237437	s_0	ct_0	f	21	-1.0	0.0
217613	T_0	test_2021-03-01 02:30:02.237438	s_0	ct_0	f	21	-1.0	0.0
217614	T_0	test_2021-03-01 02:30:02.237436	s_0	ct_0	f	21	-1.0	0.0
217615	T_0	test_2021-03-01 02:30:02.237445	s_0	ct_0	f	21	-1.0	0.0
217616	T_0	test_2021-03-01 02:30:02.237430	s_0	ct_0	f	21	-1.0	0.0

3/4 - Suppression des lignes « test »

```
# Sélection des lignes dans lesquelles le mot "test" n'apparaît pas (suppression des lignes contenant "test")
data=data[~data.date.str.contains("test")]
```

4/4 - Conclusion

```
# Vérification de la présence de valeurs négatives qui ne soient pas des tests
data.loc[data.price== -1.0]
```

id_prod	date	session_id	client_id	sex	age	price	categ

Toutes les valeurs négatives étaient des tests.

2 - Gestion des NaN

103 NaN → **0_2245** → *price ?* → imputation par la **médiane** : 9,99 €

↓

categ ?

↓

categ : 0

3 - Autres opérations

1/2 - Transformation de la colonne « date »

date
2021-10-27 04:56:38.293970
2021-12-27 11:11:12.123067
2021-04-10 18:37:28.723910

date	hour	day_week
2021-10-27	5	Sam
2021-12-27	11	Dim
2021-04-10	19	Mar

2/2 - Création d'un df « commandes »

session_id	client_id	age	price	number_items	date	hour	day_week
s_1	c_329	55	11.99	1	2021-10-27	5	Sam
s_10	c_2218	52	26.99	1	2021-12-27	11	Dim
s_100	c_3854	44	33.72	2	2021-04-10	19	Mar

Analyse univariée

1 - Etude des clients

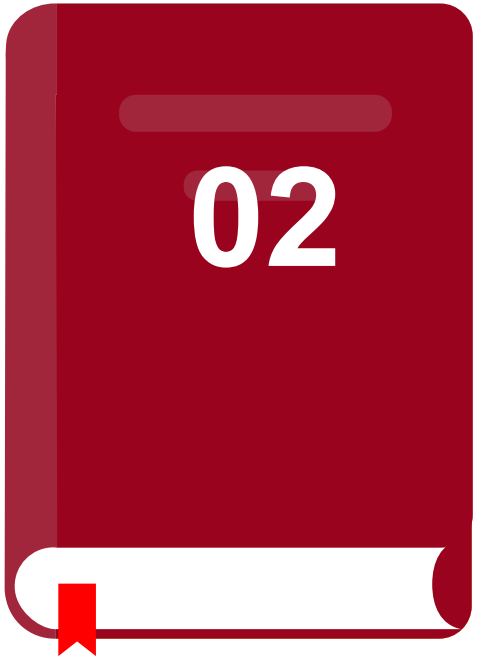
- Âge et sexe
- Répartition du CA
- Meilleurs clients
- Taux d'attrition

2 - Etude des produits

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

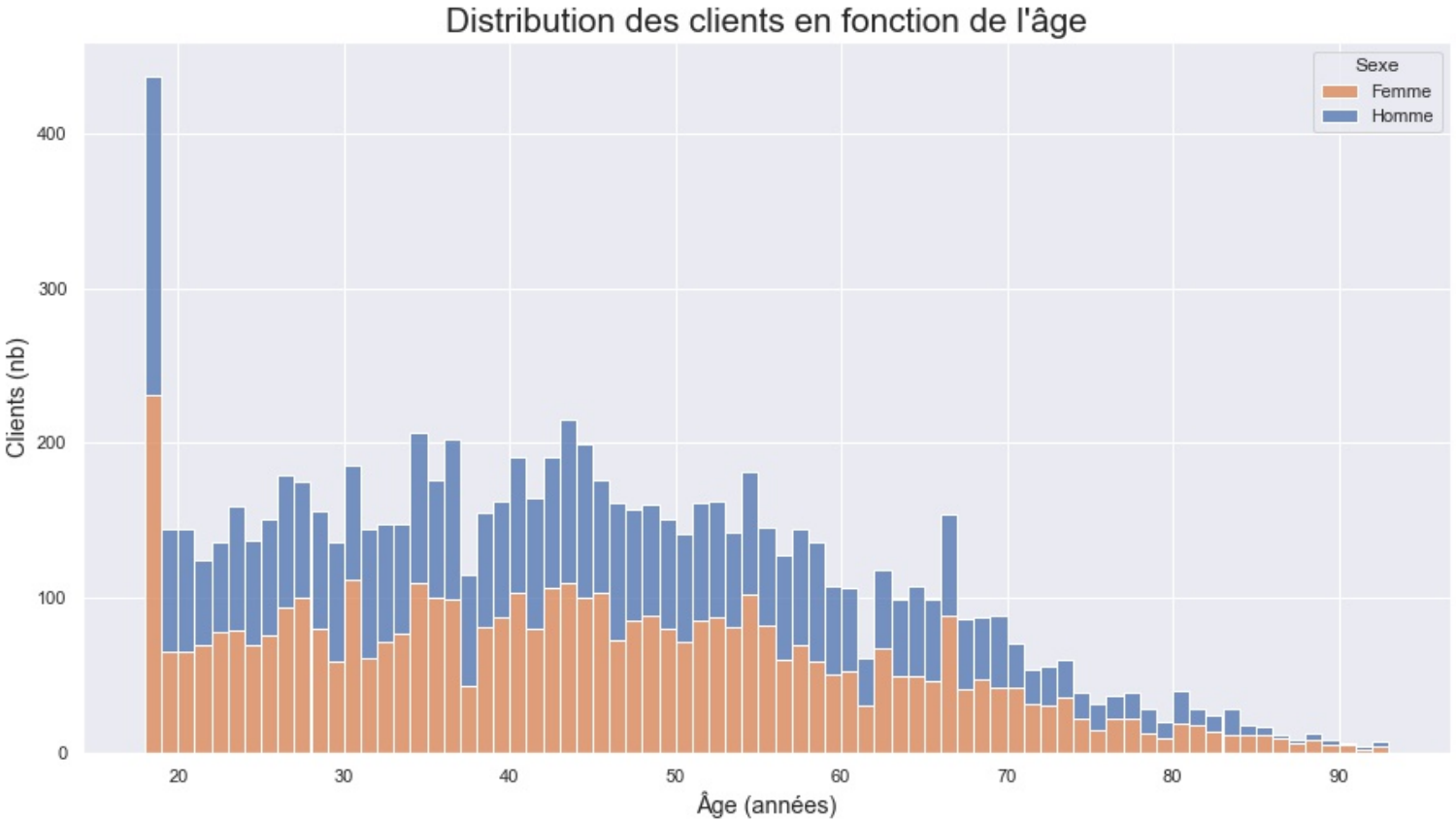
3 - Etude des transactions

- CA annuel et mensuel
- CA selon l'heure du jour
- CA selon le jour de la semaine
- CA et fréquence selon la catégorie

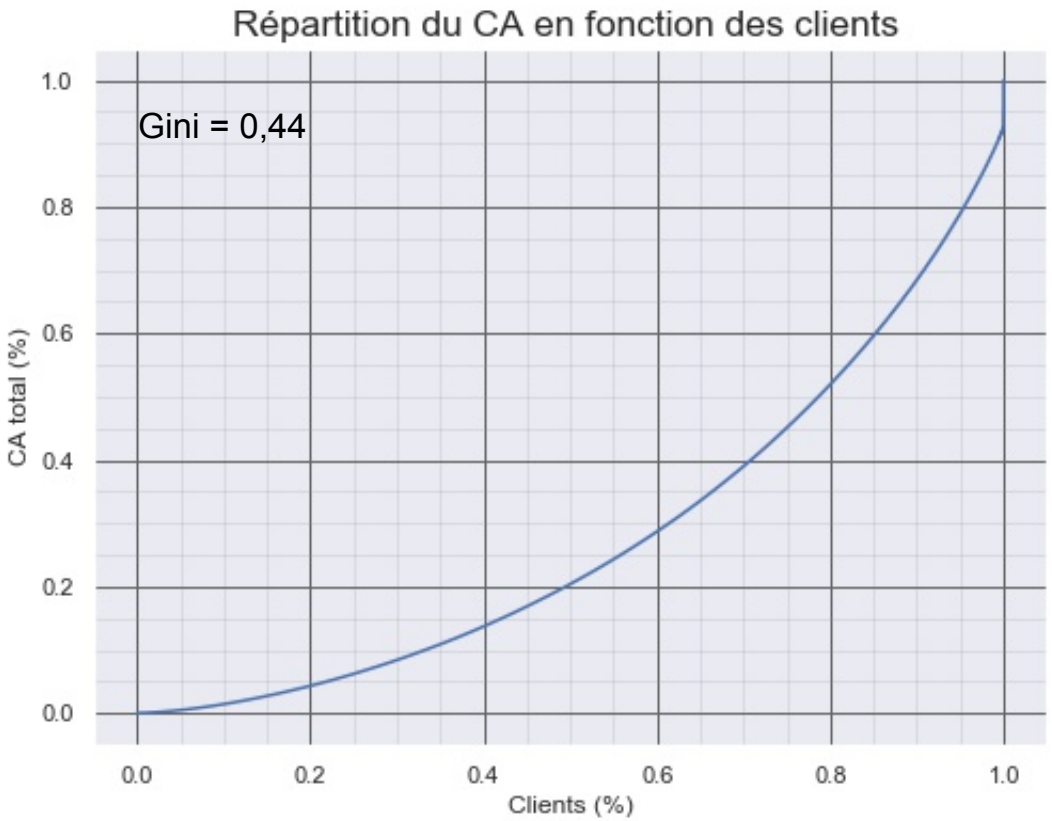


1 - Etude des clients

1 - Etude des clients



1 - Etude des clients



Les meilleurs clients…

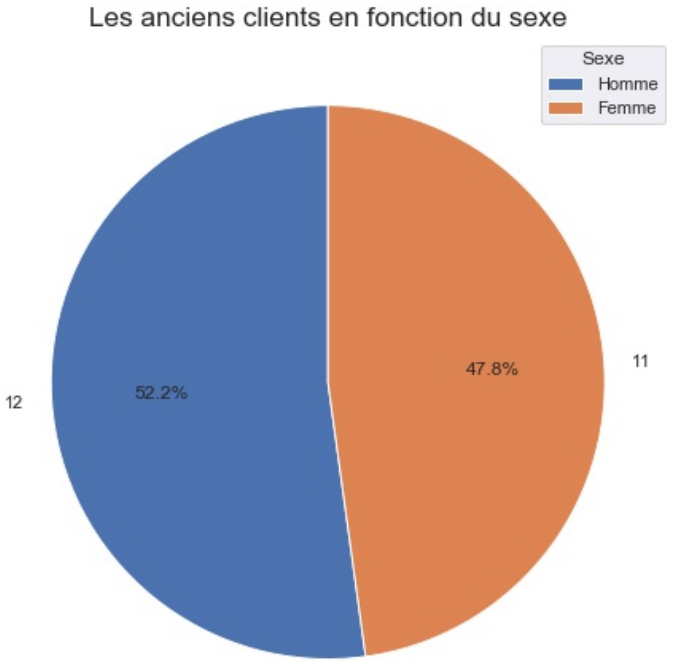
selon le montant des achats.

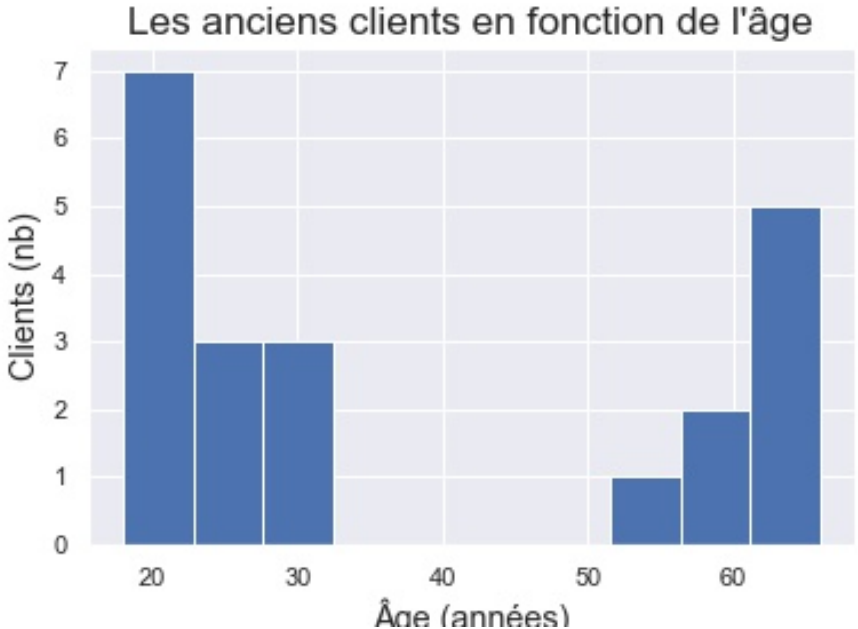
	client_id	Recency	Frequency	MonetaryValue	R	F	M	RFM_Score	sex	age
0	c_1609	1	5501	162007.34	4	4	4	12	m	42
1	c_3454	1	2711	54462.90	4	4	4	12	m	53
2	c_4958	1	1888	144257.21	4	4	4	12	m	23
3	c_6714	1	1286	73217.32	4	4	4	12	f	54
4	c_682	5	84	2264.49	4	4	4	12	f	48
5	c_8392	1	79	2515.98	4	4	4	12	f	44
6	c_8510	6	77	2318.67	4	4	4	12	m	31
7	c_8026	4	77	2547.66	4	4	4	12	m	44
8	c_5602	1	77	2309.84	4	4	4	12	m	33
9	c_7421	4	77	2511.98	4	4	4	12	m	44

selon la fréquence d'achat.

	client_id	Recency	Frequency	MonetaryValue	R	F	M	RFM_Score	sex	age
0	c_1609	1	5501	162007.34	4	4	4	12	m	42
1	c_4958	1	1888	144257.21	4	4	4	12	m	23
2	c_6714	1	1286	73217.32	4	4	4	12	f	54
3	c_3454	1	2711	54462.90	4	4	4	12	m	53
4	c_7959	1	76	2564.25	4	4	4	12	f	48
5	c_8026	4	77	2547.66	4	4	4	12	m	44
6	c_4491	1	73	2540.53	4	4	4	12	f	38
7	c_2140	7	74	2527.01	3	4	4	11	f	45
8	c_8392	1	79	2515.98	4	4	4	12	f	44
9	c_7421	4	77	2511.98	4	4	4	12	m	44

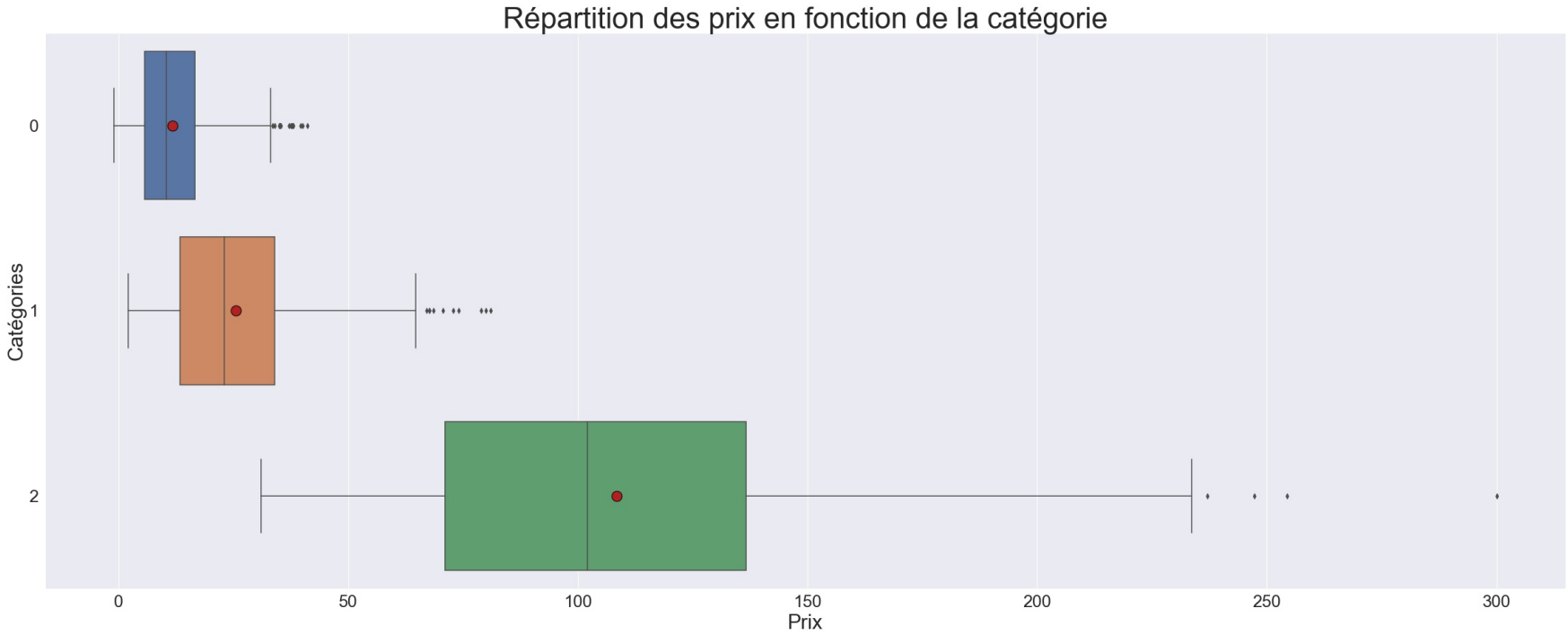
Taux d'attrition : 0,27 %



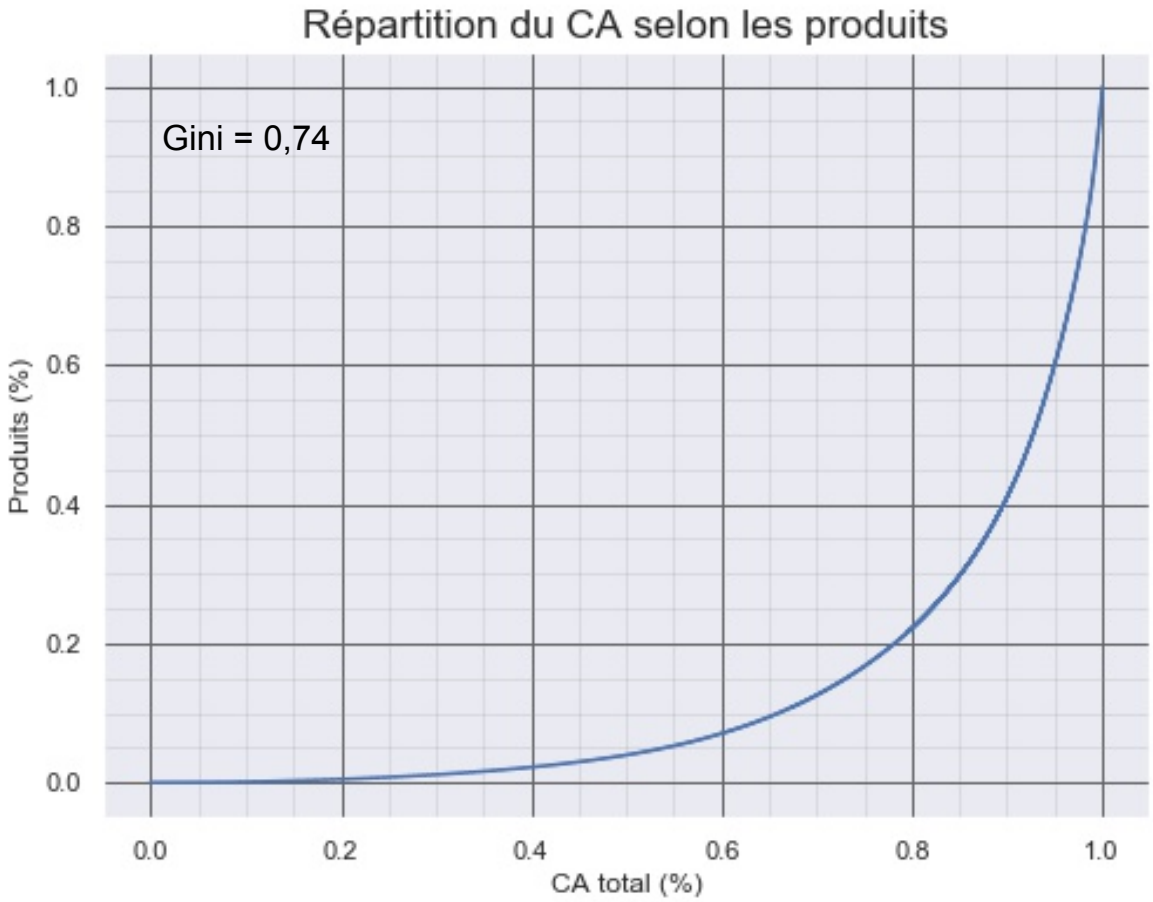


2 - Etude des produits

2 - Etude des produits



2 - Etude des produits



Les meilleurs produits…

selon le montant des achats.

id_prod	number_purchases	monetary_value
2_135	491	33874.09
2_112	473	31960.61
2_102	489	28919.46
2_209	390	27296.10
2_110	434	27016.50
1_369	1081	25933.19
1_395	891	25830.09
2_166	111	25534.44
2_43	361	25266.39
2_39	435	25225.65

selon le nombre d'achats.

id_prod	number_purchases	monetary_value
1_369	1081	25933.19
1_417	1062	22291.38
1_498	1036	24211.32
1_414	1027	24473.41
1_425	1013	17210.87
1_398	952	9681.84
1_406	946	23470.26
1_413	944	16982.56
1_403	939	16892.61
1_407	933	14918.67

Les pires produits…

selon le montant des achats.

id_prod	number_purchases	monetary_value
0_1653	1	0.99
0_1539	1	0.99
0_1840	1	1.28
0_1284	1	1.38
0_1858	1	1.83
0_1912	1	1.89
0_643	2	1.98
0_1191	2	1.98
0_1601	1	1.99
0_541	1	1.99

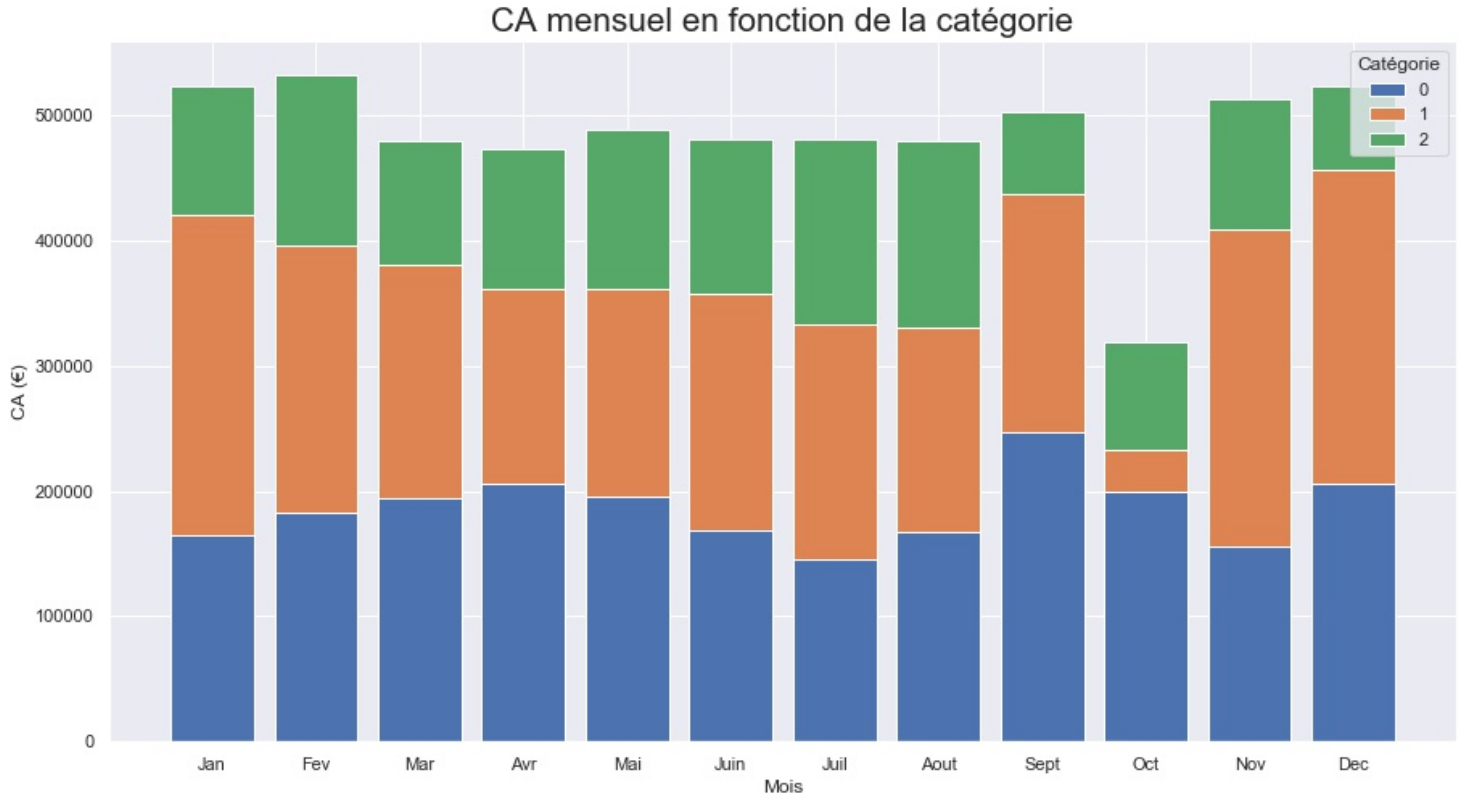
selon le nombre d'achats.

id_prod	number_purchases	monetary_value
0_1683	1	2.99
0_1379	1	2.99
0_1912	1	1.89
0_1165	1	2.99
2_28	1	103.50
2_23	1	115.99
0_568	1	29.76
1_402	1	34.52
0_1151	1	2.99
0_549	1	2.99

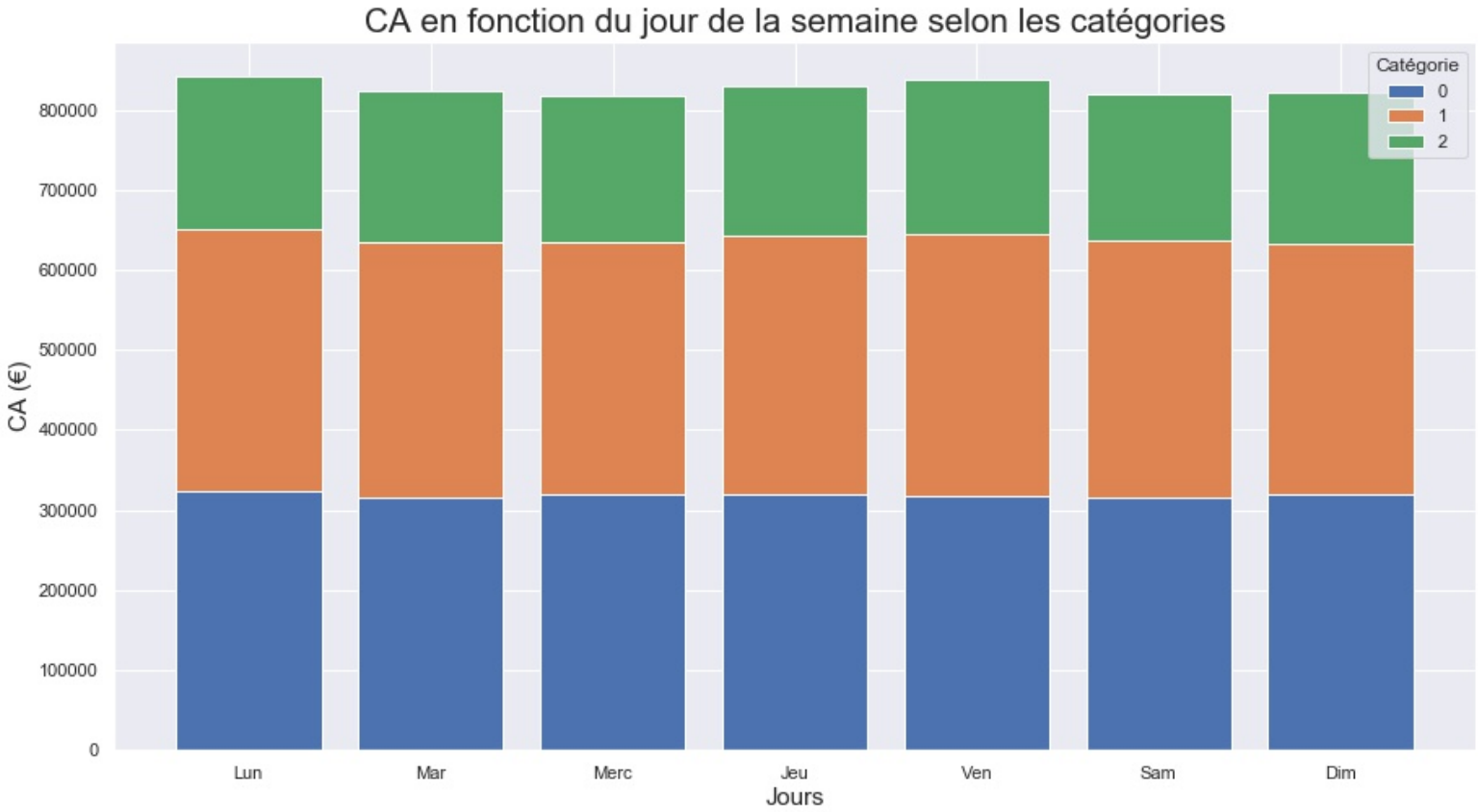
3 - Etude des transactions

3 - Etude des transactions

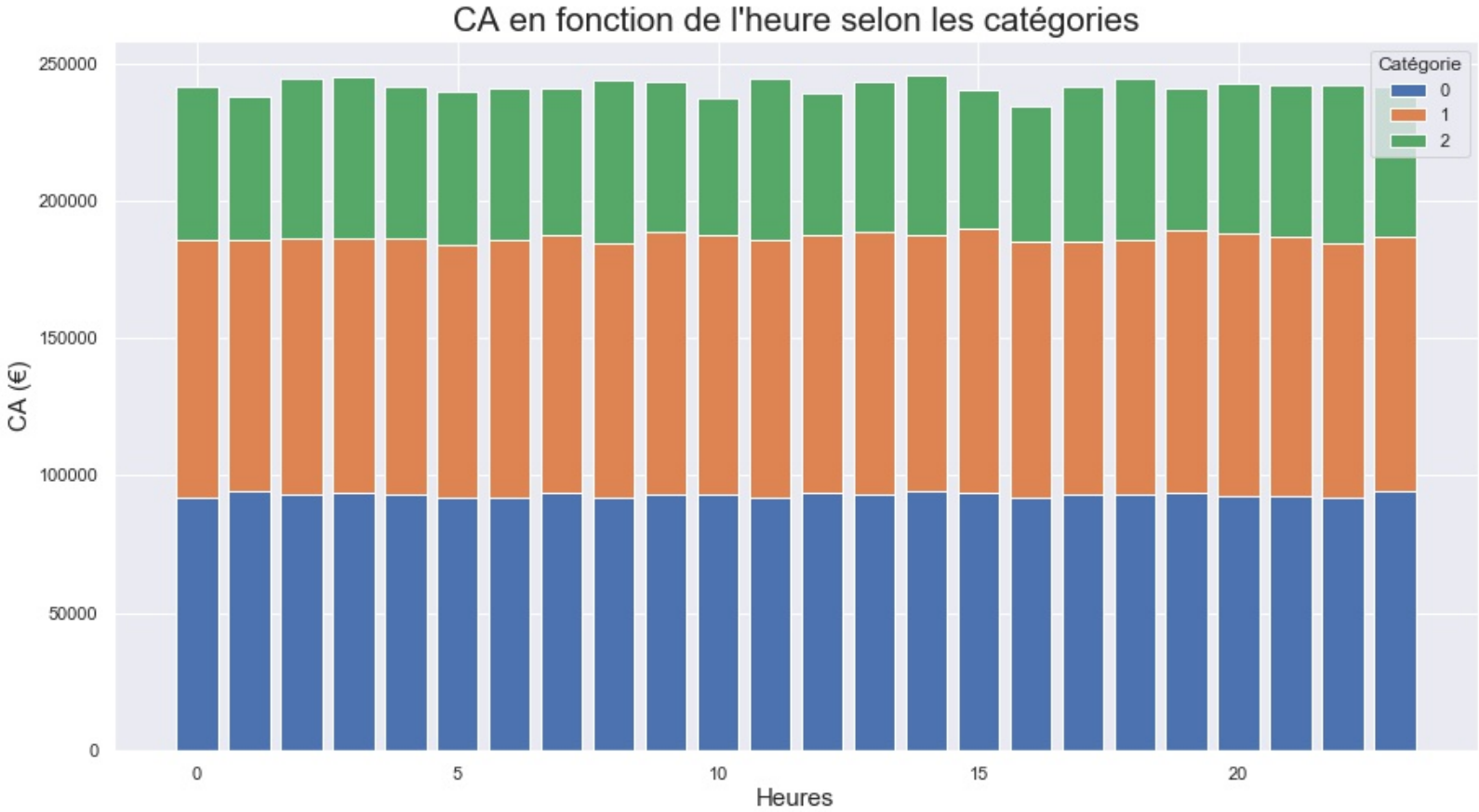
CA annuel : 5,8 Millions €



3 - Etude des transactions

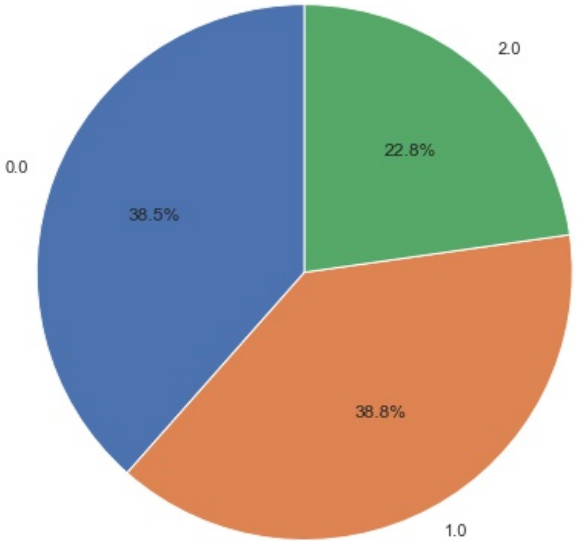


3 - Etude des transactions

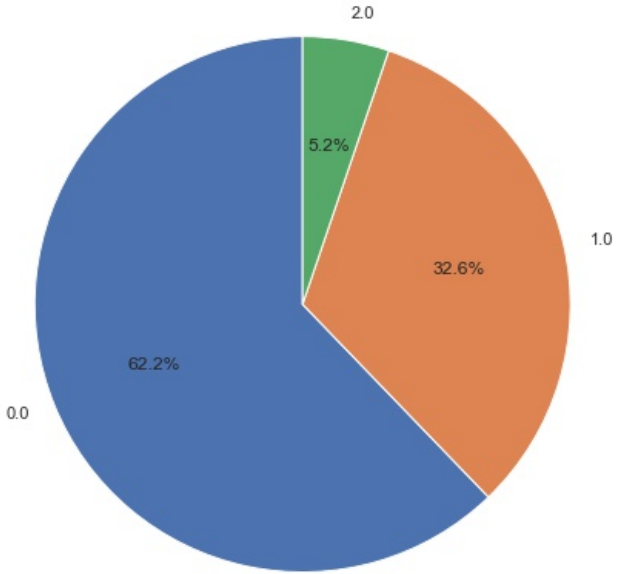


3 - Etude des transactions

CA en fonction des catégories



Fréquence d'achat selon les catégories



Analyse des corrélations

1 - Sexe et catégorie

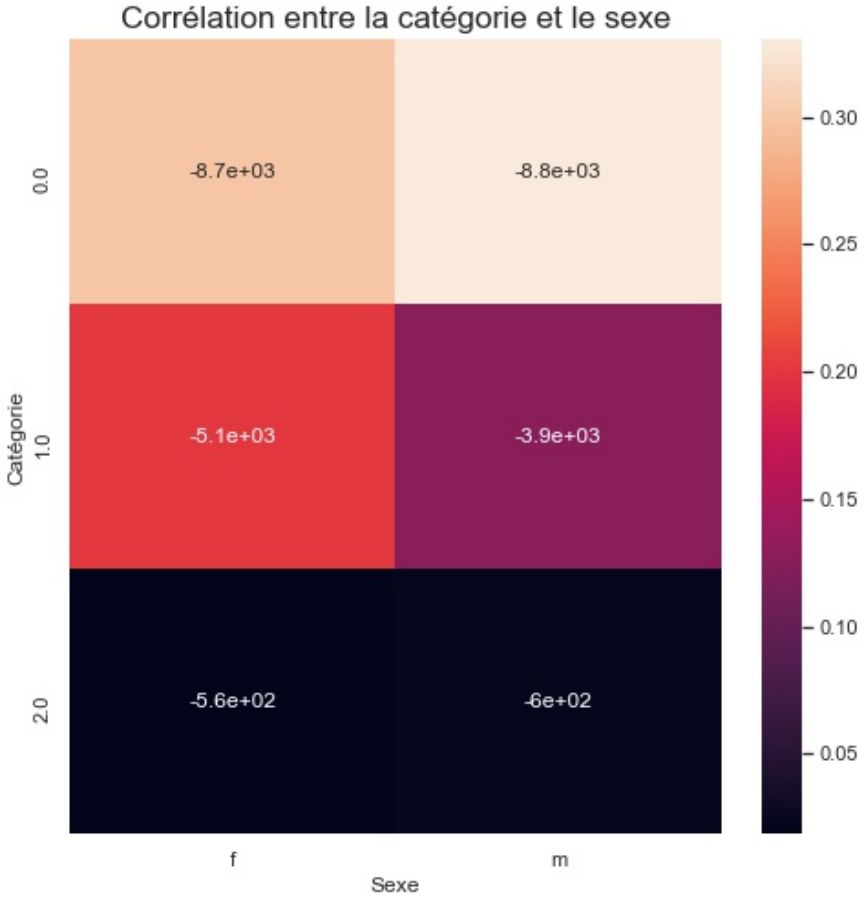
2 - Âge et taille du panier

3 - Âge et montant total

4- Âge et fréquence d’achat

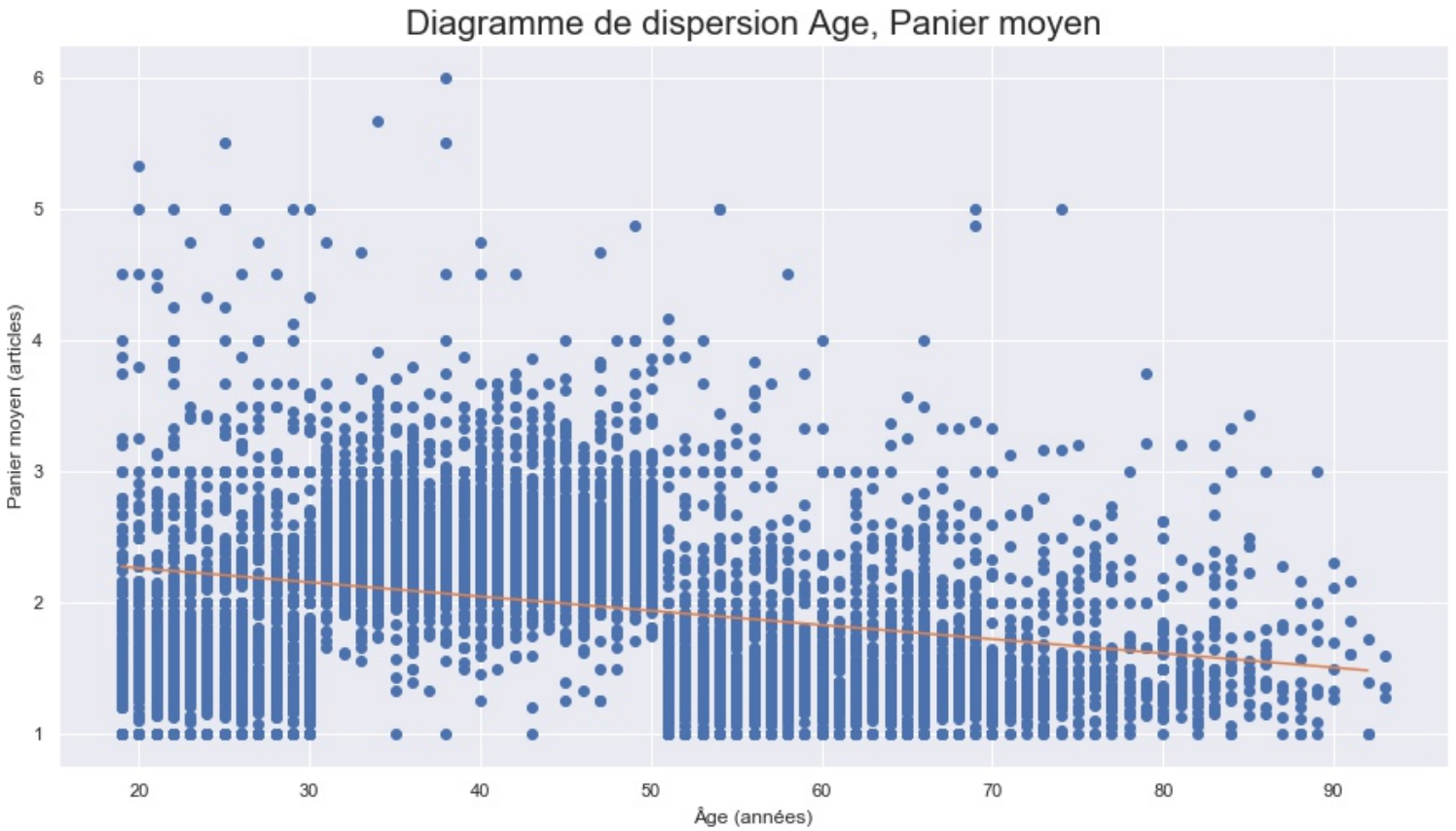
5 - Âge et catégorie

1 - Sexe et catégorie

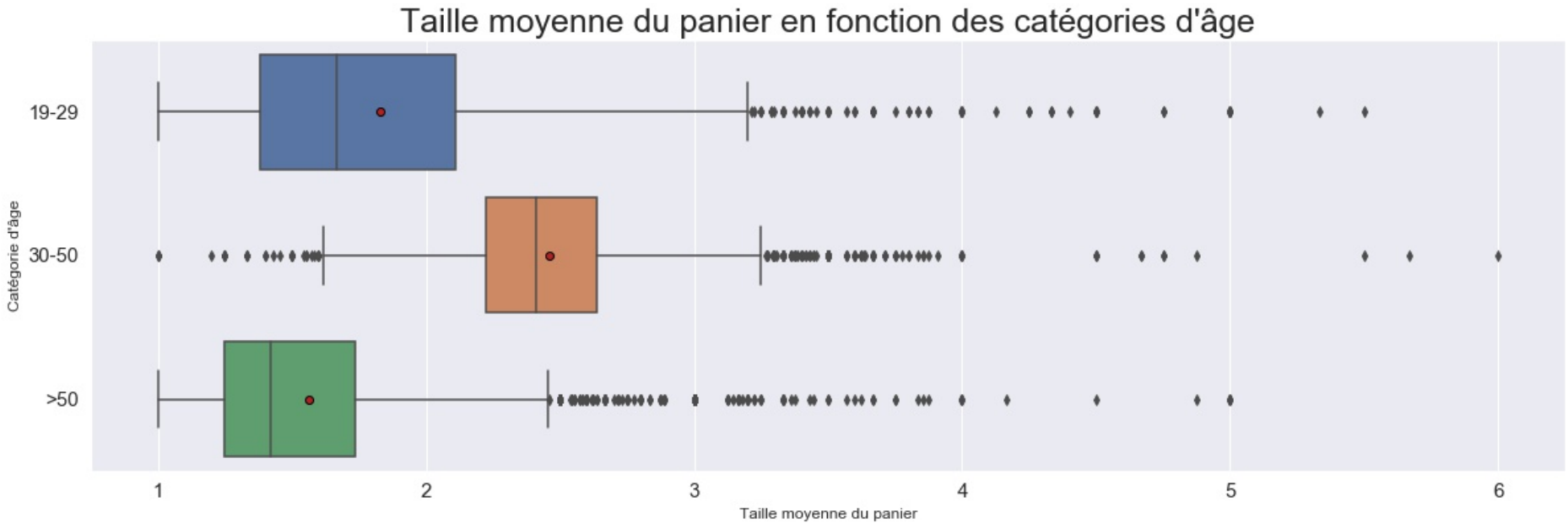


2 - Âge et taille du panier

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

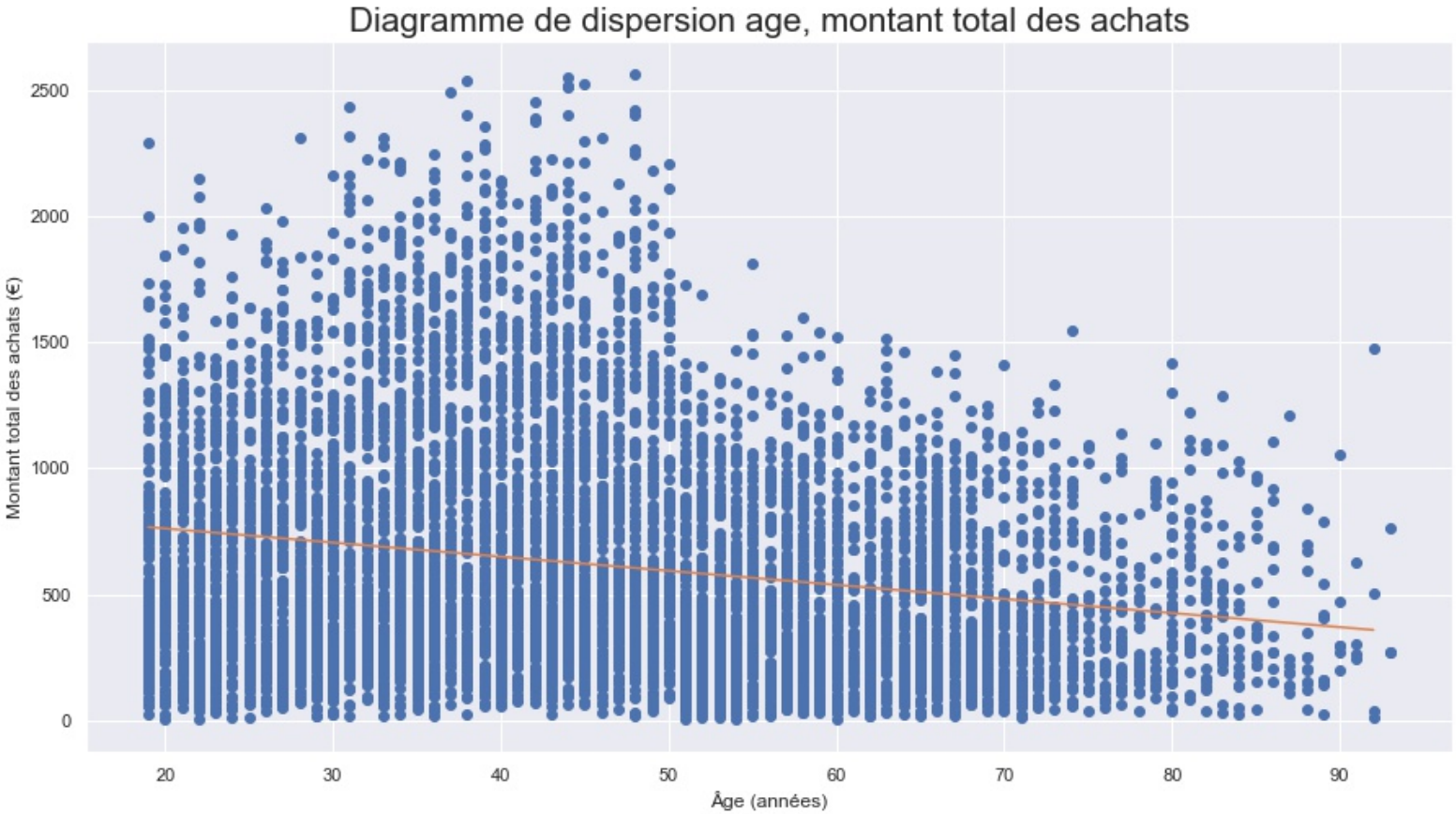
- faible
- moyenne
- forte

Ici :
p-value = 0
R = 0,63

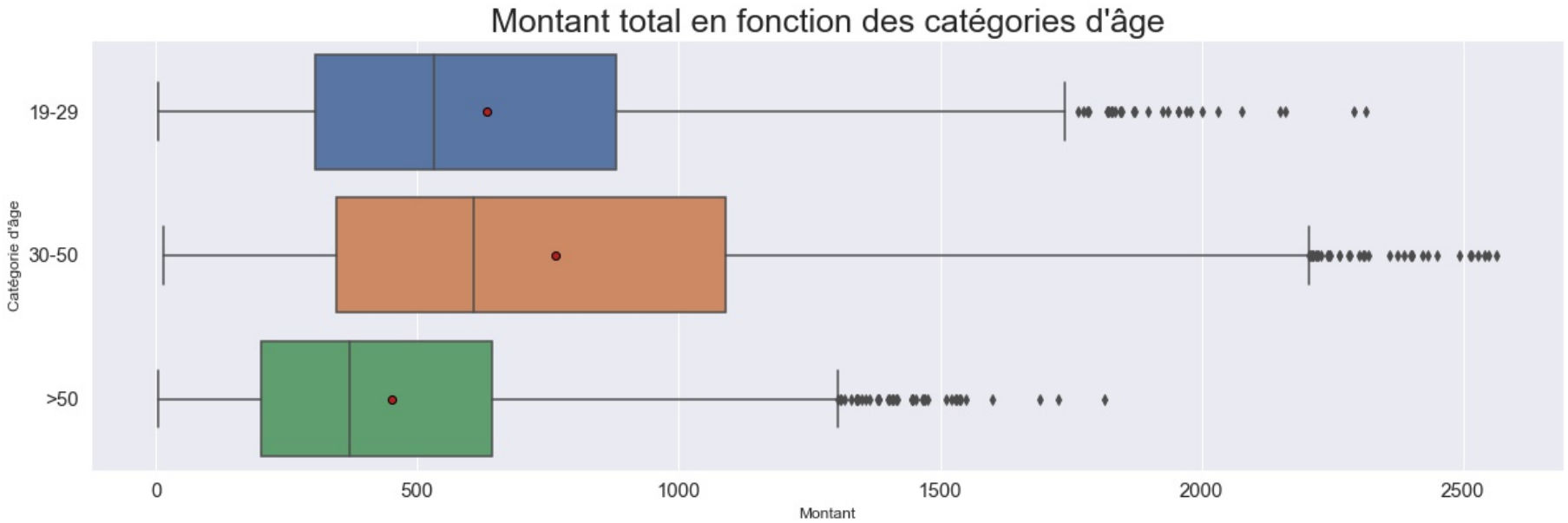
On rejette H0 : il y a une corrélation forte entre la catégorie d'âge et la taille du panier.

3 - Âge et montant total

3 - Âge et montant total



3 - Âge et montant total



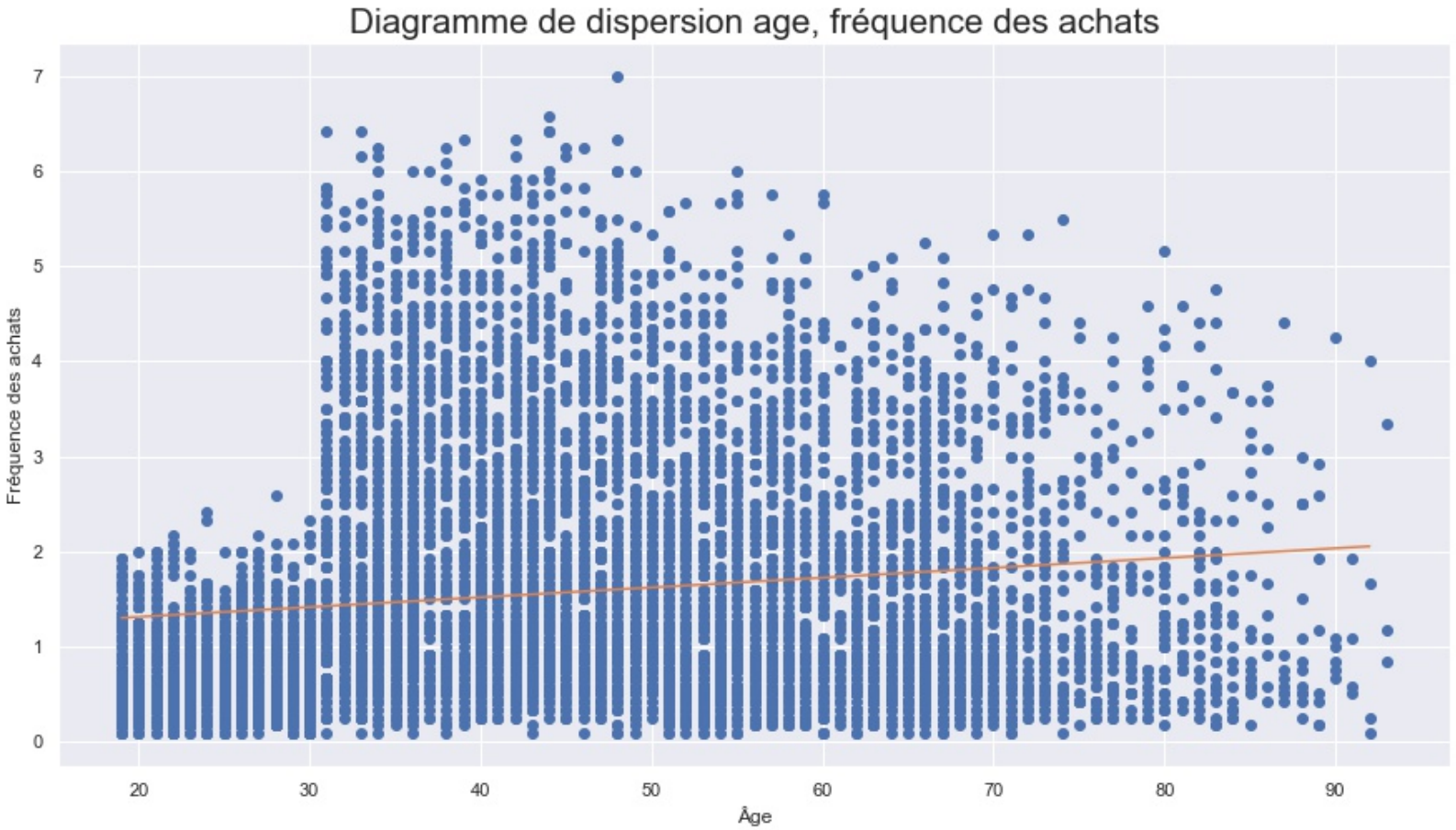
3 - Âge et montant total

Ici :
p-value = $1,9*10^{-166}$
R = 0,3

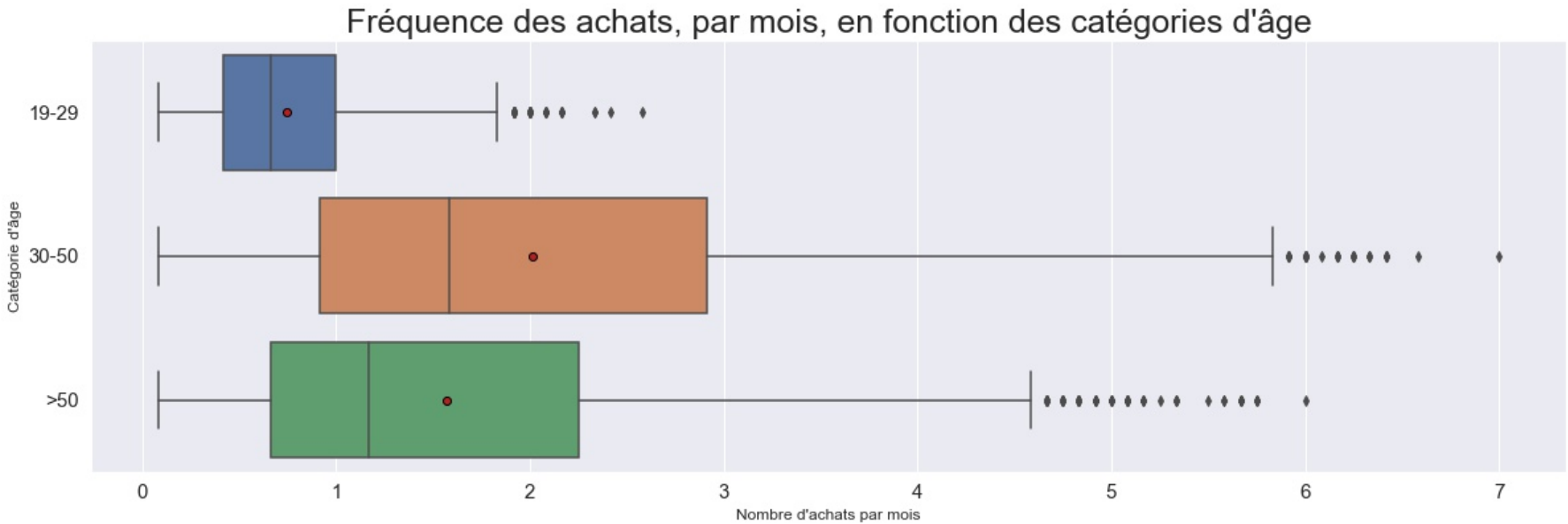
On rejette H0 : il y a une corrélation moyenne entre la catégorie d'âge et la taille du panier.

4- Âge et fréquence d’achat

4- Âge et fréquence d’achat



4- Âge et fréquence d’achat



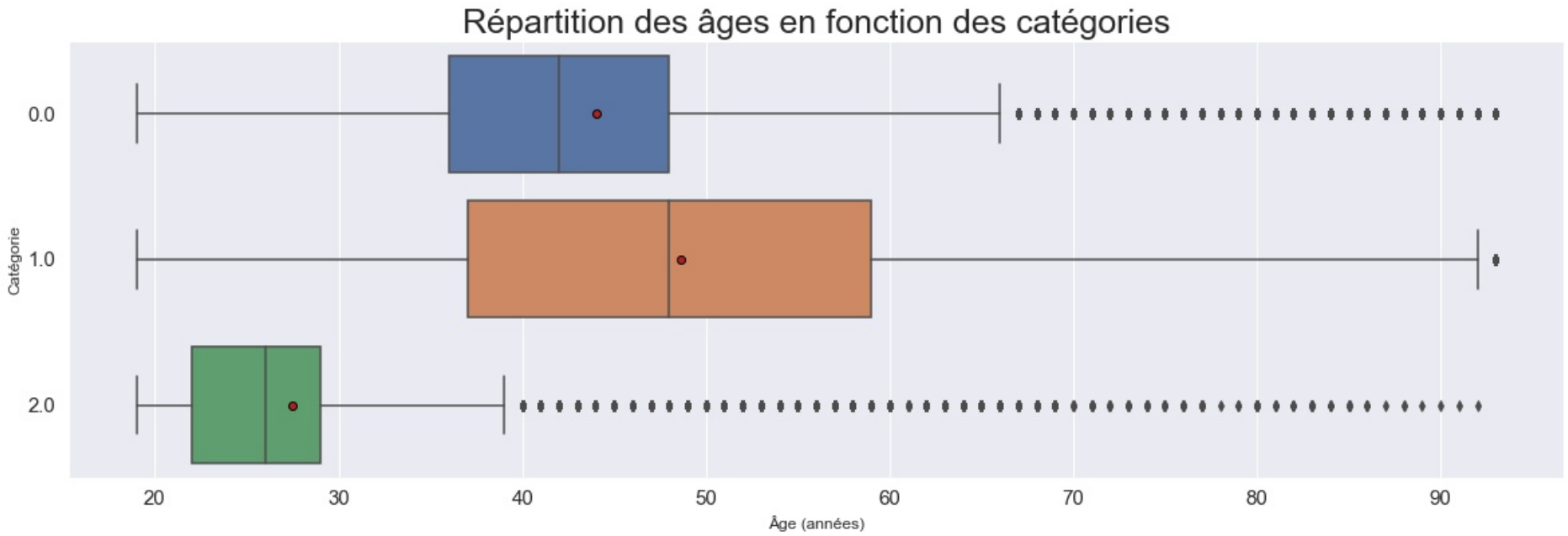
4- Âge et fréquence d’achat

Ici :
p-value = $1,7*10^{-269}$
R = 0,38

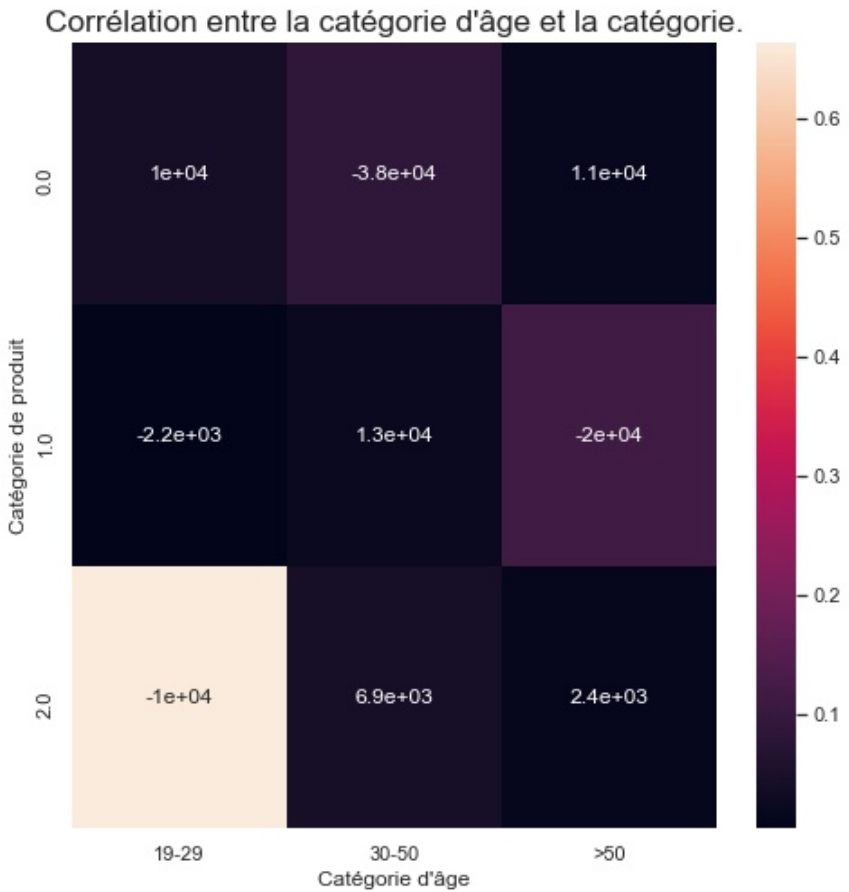
On rejette H0 : il y a une corrélation moyenne entre la catégorie d’âge et la taille du panier.

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

- Taux d'attrition faible

- 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