OPENCLASSROOMS

Projet 4
Analysez les ventes de votre entreprise

Parcours Data Analyst

Jérôme Chalain – mai 2020

OPENCLASSROOMS

SOMMAIRE :

1. Contexte

2. Précisions concernant la préparation des données

3. Mon analyse des données

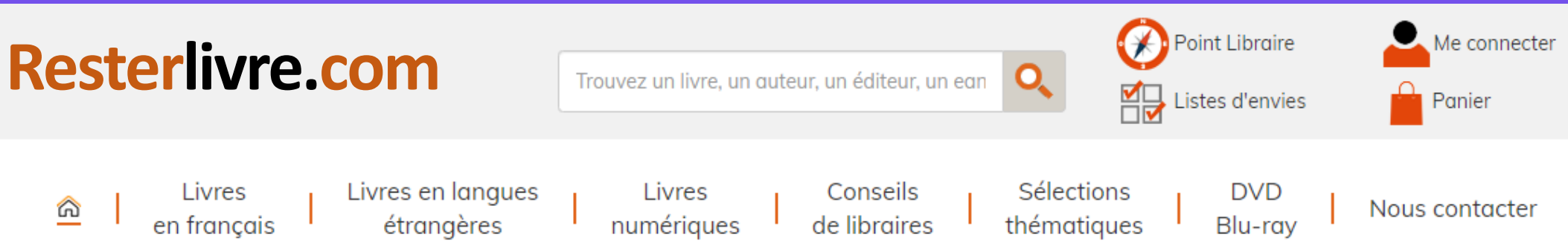
4. Analyse des corrélations

5. Conclusion

Jérôme Chalain – mai 2020

1. Contexte

Contexte



- La chaine de librairie « Rester livre » se développe en ligne !
- Afin de prendre les bonnes décisions il est important de faire un point précis sur les données recueillies

2. Détail de la préparation des données

Données de départ : 3 fichiers csv

1. Customers → clients

```
clients.shape
```

```
(8623, 3)
```

2. Products → produits

```
produits.shape
```

```
(3287, 3)
```

3. Transactions → ventes

```
ventes.shape
```

```
(337016, 4)
```

```
produits.head()
```

	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

```
clients.head()
```

	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

```
ventes.head()
```

	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

Jérôme Chalain – mai 2020

Détail de la préparation des données

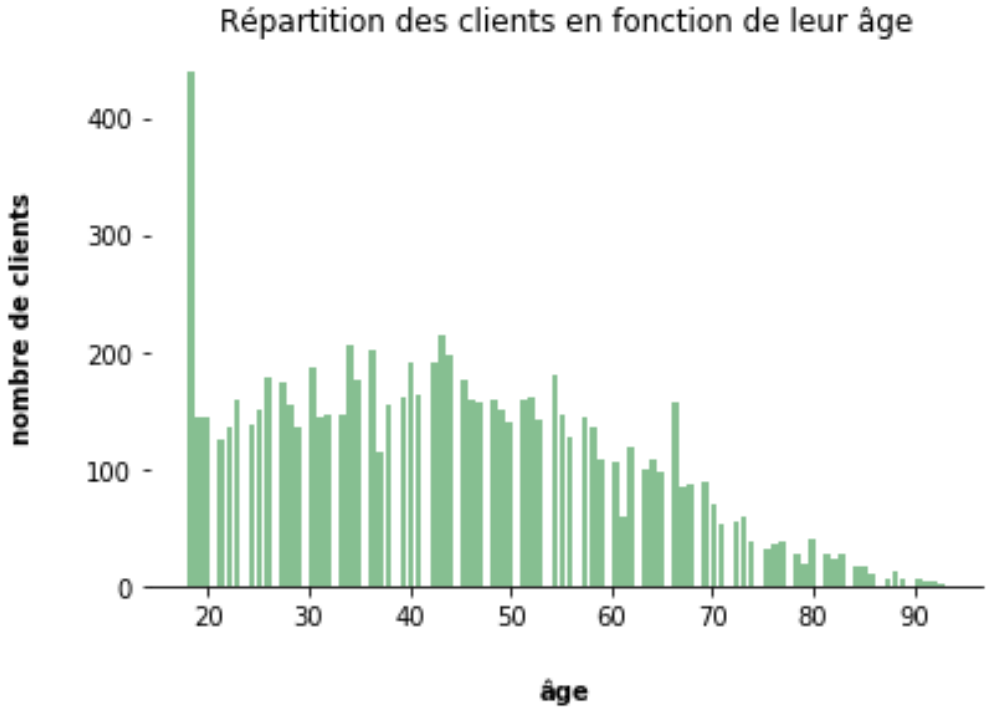
Table customers → clients

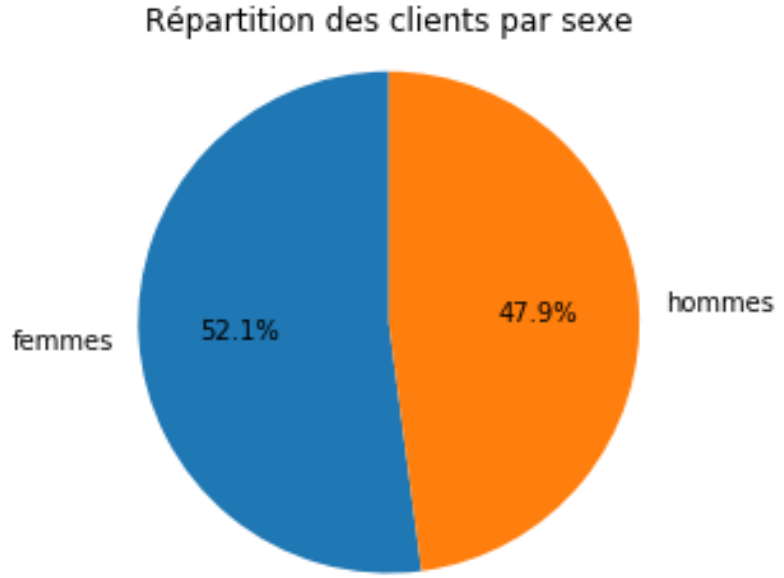
Ajout des colonnes « age » et « classe_age »

```
clients.head()
```

	client_id	sex	birth	age	classe_age
0	c_4410	f	1967	55	50+

- Plus de 8000 clients
- Nés entre 1929 et 2004.
- Age mini : 18 ans
- Age maxi : 93 ans
- Moyenne : 44 ans
- Médiane : 43 ans
- La moitié des clients a entre 30 et 56 ans.



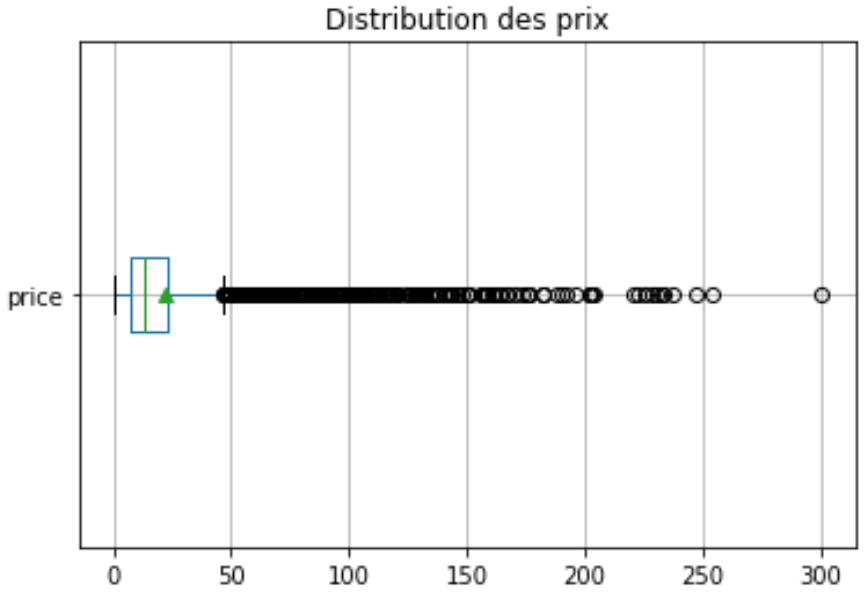


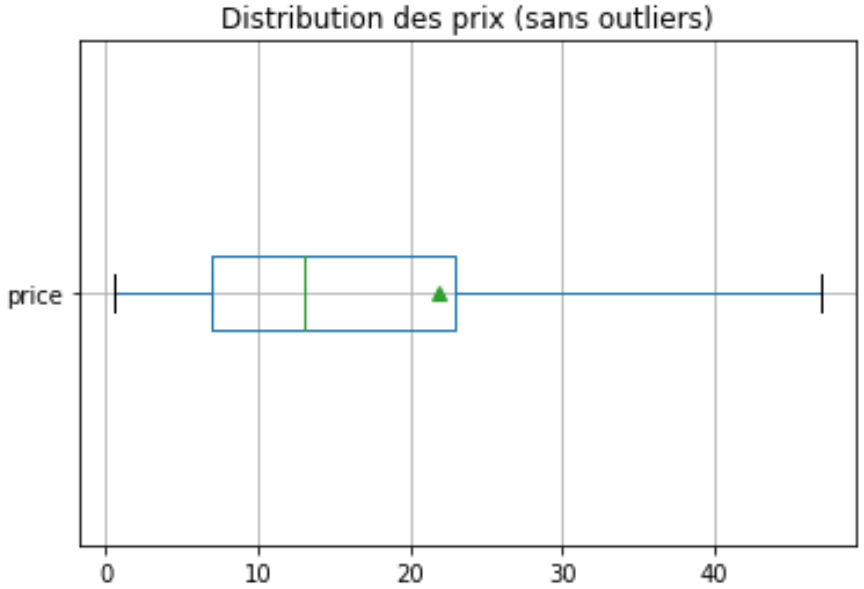
Détail de la préparation des données

Table products → produits

	id_prod	price	categ
731	T_0	-1.0	0

- La moyenne des prix est de 22 €.
- L'écart-type, plus élevé que la moyenne, montre une forte dispersion des prix.
- La moitié des prix est situe entre 7 et 23 €.
- Prix mini 0.62 €.
- Prix max 300 €.





Jérôme Chalain – mai 2020

Détail de la préparation des données

Table transactions → ventes

200 transactions correspondent à des tests

- 336 816 dates uniques.
- commence au 1/03/2021, fin au 28/02/2022 => 1 année complète.

```
# Modification du type "date" en datetime
ventes['date'] = pd.to_datetime(ventes['date'])
```

```
ventes.loc[(ventes["session_id"]=="s_0")].head()
```

	id_prod	date	session_id	client_id
1431	T_0	test_2021-03-01 02:30:02.237420	s_0	ct_1
2365	T_0	test_2021-03-01 02:30:02.237446	s_0	ct_1
2895	T_0	test_2021-03-01 02:30:02.237414	s_0	ct_1
5955	T_0	test_2021-03-01 02:30:02.237441	s_0	ct_0
7283	T_0	test_2021-03-01 02:30:02.237434	s_0	ct_1

```
ventes.describe(include='all')
```

	id_prod	date	session_id	client_id
count	336816	336816	336816	336816
unique	3265	336816	169194	8600
top	1_369	2021-06-30 11:35:55.387896	s_118668	c_1609
freq	1081	1	14	12855
first	NaN	2021-03-01 00:01:07.843138	NaN	NaN
last	NaN	2022-02-28 23:59:58.040472	NaN	NaN

Détail de la préparation des données

Jointure des tables

```
# Jointure 1 entre ventes et produits sur id_prod, outer pour garder toutes les données
cpv = ventes.merge(produits, how='outer')
```

```
# Jointure 2 entre le nouveau dataframe et clients sur client_id
cpv = cpv.merge(clients, how='outer')
cpv.head()
```

	id_prod	date	session_id	client_id	price	categ	sex	birth	age	classe_age
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450	4.99	0.0	f	1977.0	45.0	30-50
1	0_1085	2021-09-29 11:14:59.793823	s_97382	c_4450	3.99	0.0	f	1977.0	45.0	30-50
2	0_1453	2021-08-27 19:50:46.796939	s_81509	c_4450	7.99	0.0	f	1977.0	45.0	30-50
3	0_1405	2021-08-27 20:07:25.878440	s_81509	c_4450	4.99	0.0	f	1977.0	45.0	30-50
4	0_1392	2021-12-28 11:45:04.072281	s_141302	c_4450	6.30	0.0	f	1977.0	45.0	30-50

Détail de la préparation des données
Relancer les clients non acheteurs

Recherche des valeurs nulles

```
# Recherche des valeurs nulles
cpv_na = cpv.loc[cpv.isnull().any(axis=1)]
```

Ce que l'on constate :

- 22 produits n'ont pas été vendus (conservés)
- 21 clients n'ont rien acheté (conservés) → les relancer ?
- L'id_prod 0_2245 est enregistré dans ventes mais **pas dans produits**.

	id_prod	date	session_id	client_id	price	categ	sex	birth	age	classe_age
18202	0_2245	2021-06-05 17:04:43.982913	s_44481	c_6714	NaN	NaN	f	1968.0	54.0	50+

Détail de la préparation des données

Traitement du produit 0_2245 : j'ai pris le parti de supposer une anomalie à l'enregistrement de la vente. Je lui ai imputé la moyenne des prix de sa catégorie.

```
# Je commence par imputer la catégorie de ce produit
catprod = 0
cpv.loc[cpv.id_prod=='0_2245', 'categ'] = catprod
```

```
# Calcul de la moyenne des prix de la catégorie de produits 0:
moyprixcat0 = cpv.loc[cpv.categ == catprod, 'price'].mean()
moyprixcat0
```

```
10.647071995724621
```

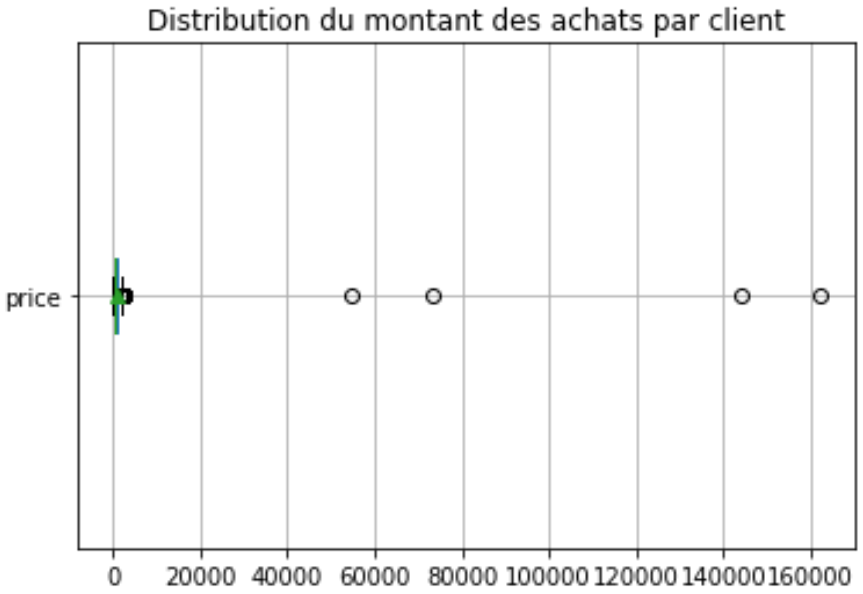
```
# Ajout de cette moyenne dans la colonne prix de l'id_prod 0_2245
cpv.loc[cpv.id_prod=='0_2245', 'price'] = moyprixcat0
```

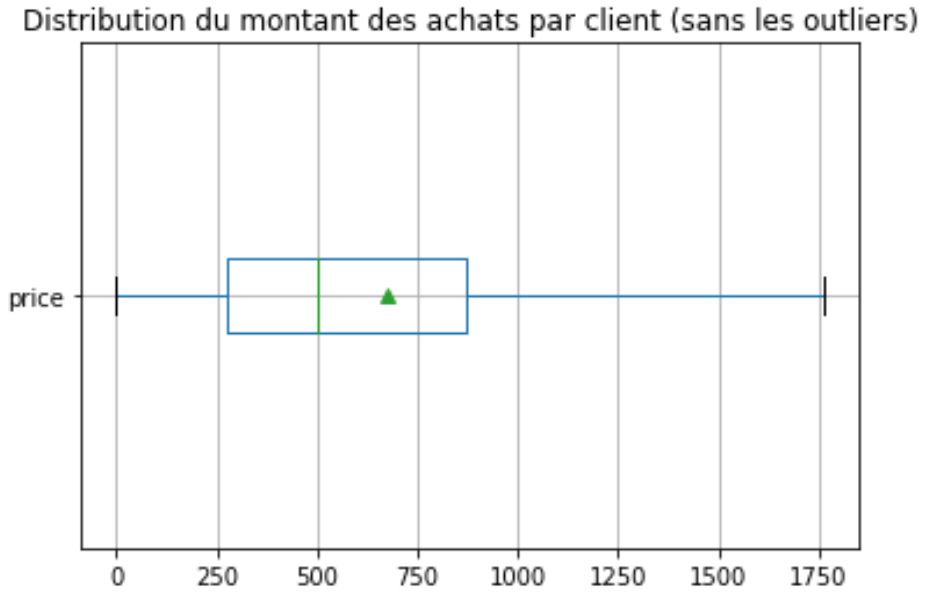
```
# Vérification
cpv[cpv.id_prod == '0_2245'].sample(1)
```

	id_prod	date	session_id	client_id	price	categ	sex	birth	age	classe_age
224547	0_2245	2021-09-06 14:03:40.171938	s_85877	c_1060	10.647072	0.0	f	1973.0	49.0	30-50

Détail de la préparation des données
Offrir des services particuliers à 4 clients exceptionnels

- 4 clients ont achetés chacun plus de 54 000 €, soit beaucoup plus que les autres clients.
- Ces clients ont dépensé plus de 400 000 €.
- Il représentent presque 7.5 % du CA.





	price
count	8621.000000
mean	672.506022
std	2563.075956
min	0.000000
25%	274.230000
50%	500.160000
75%	869.110000
max	162007.340000

Détail de la préparation des données

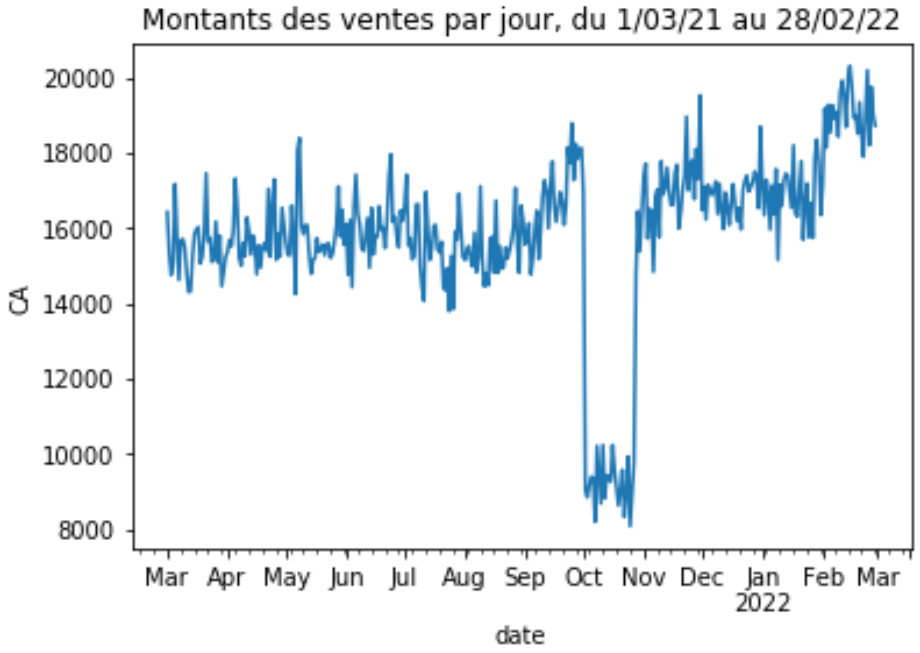
Finalement on obtient 3 tables faciles à exploiter :

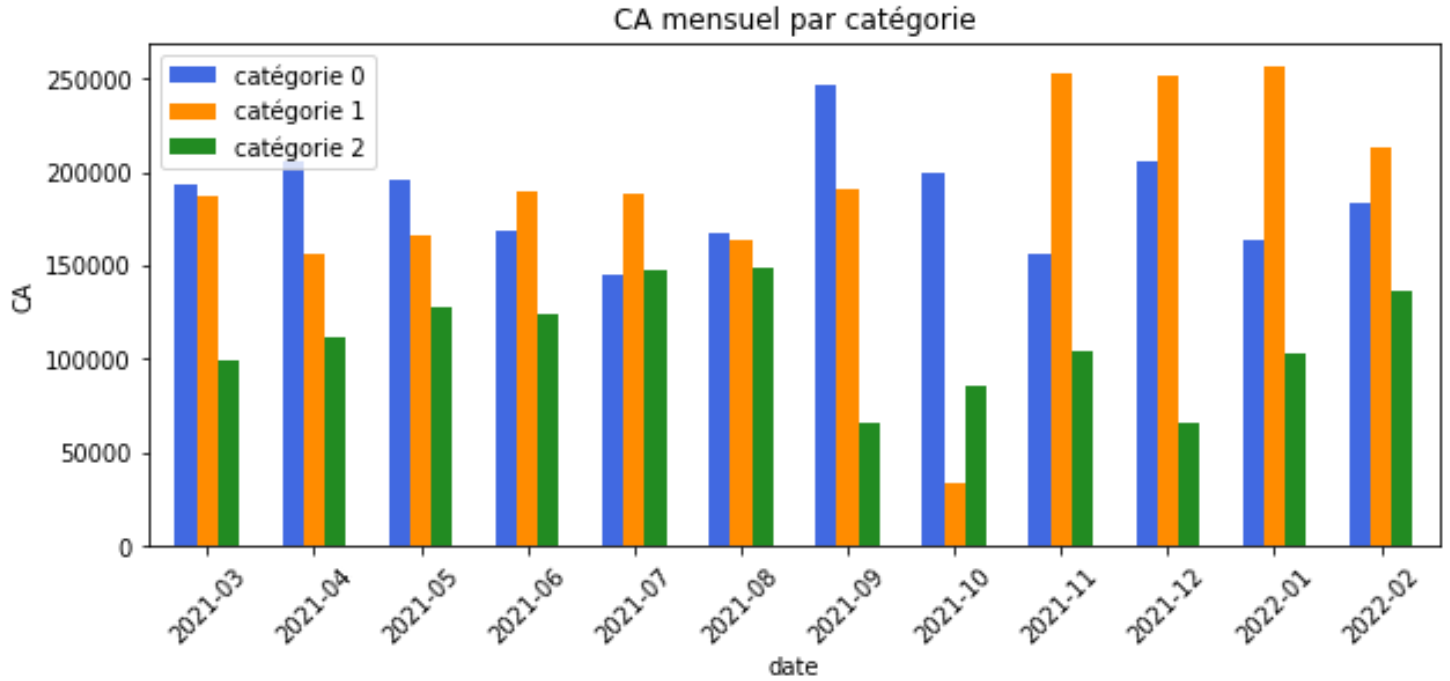
- 1 globale (table principale)
- 1 sans les clients « grands comptes »
- 1 sans les grands comptes et la classe d'âge 18 ans.

3. Analyse des données

Analyse des données
Que s'est-il passé en octobre ?...

1. **Attention** : données manquantes concernant les ventes ?

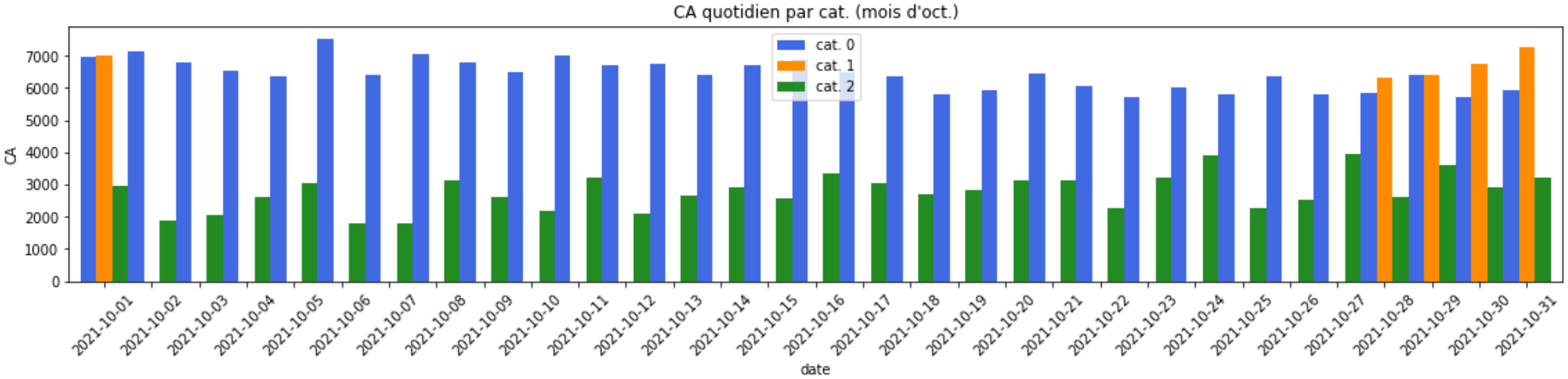




Jérôme Chalain – mai 2020

Analyse des données

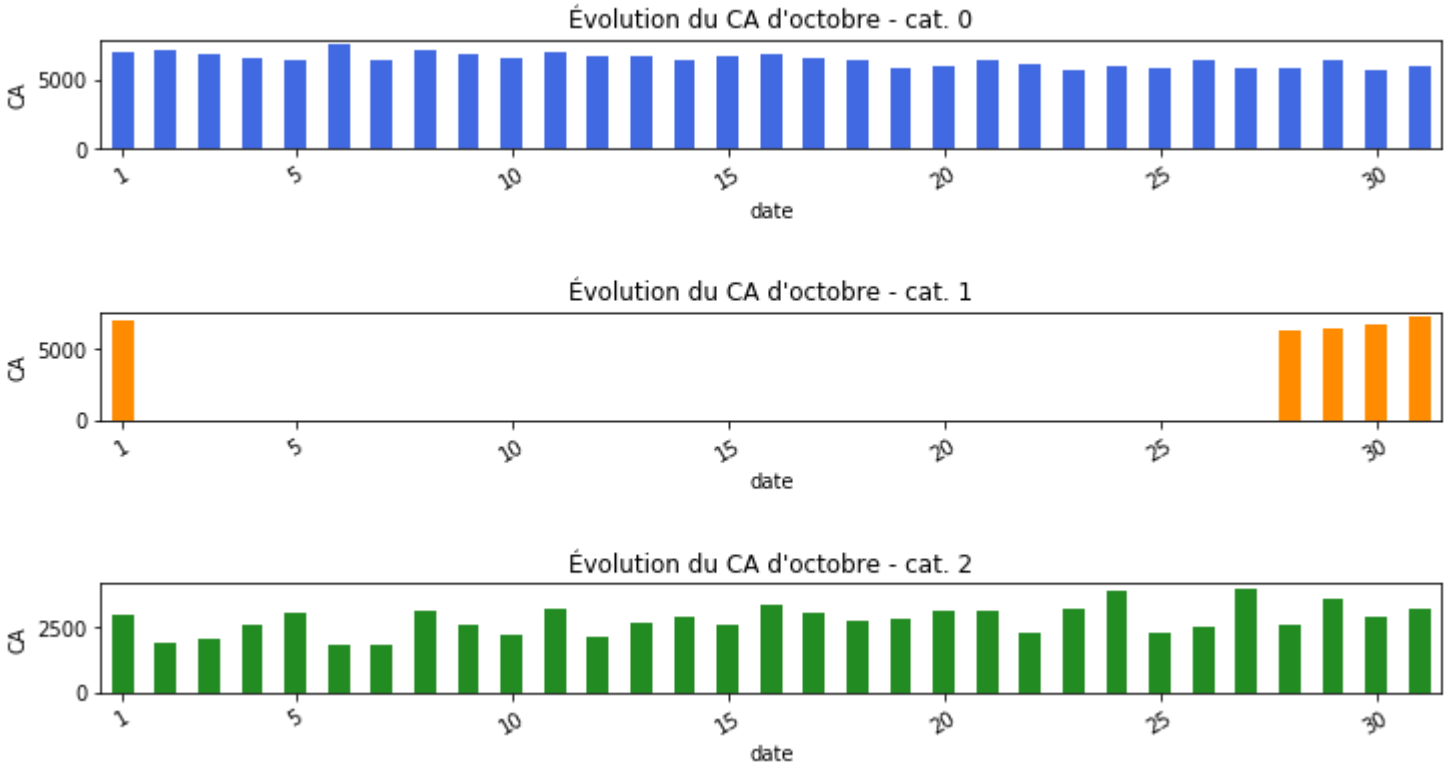
Que s’est-il passé en octobre...avec la catégorie 1 ?

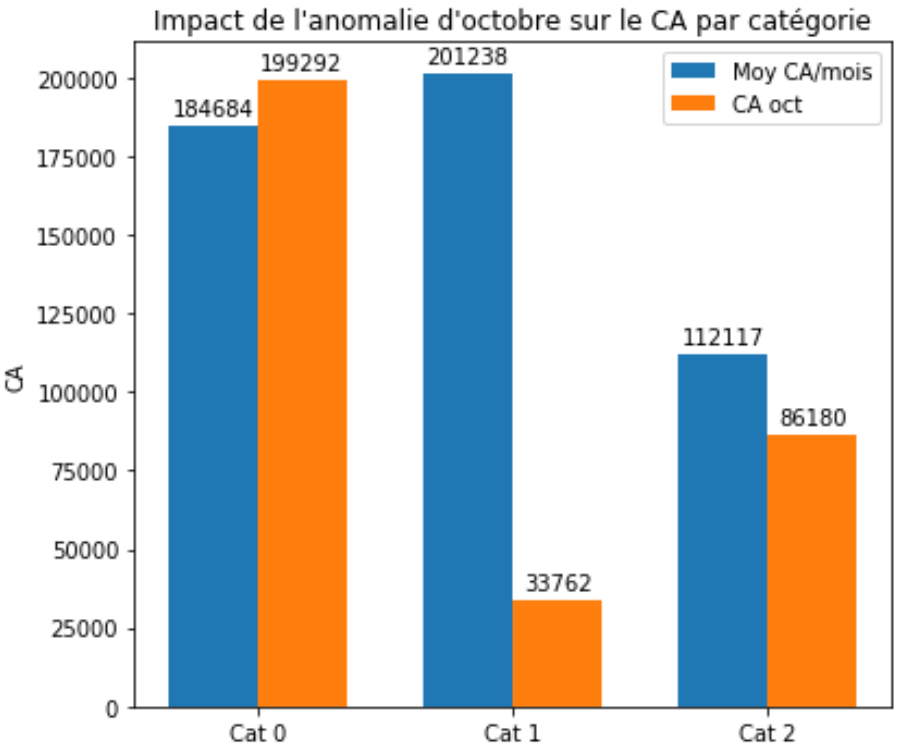


Problème d’enregistrement ?
Rupture de stock ?
A élucider.

Jérôme Chalain – mai 2020

Analyse des données
Impact de l’anomalie d’octobre



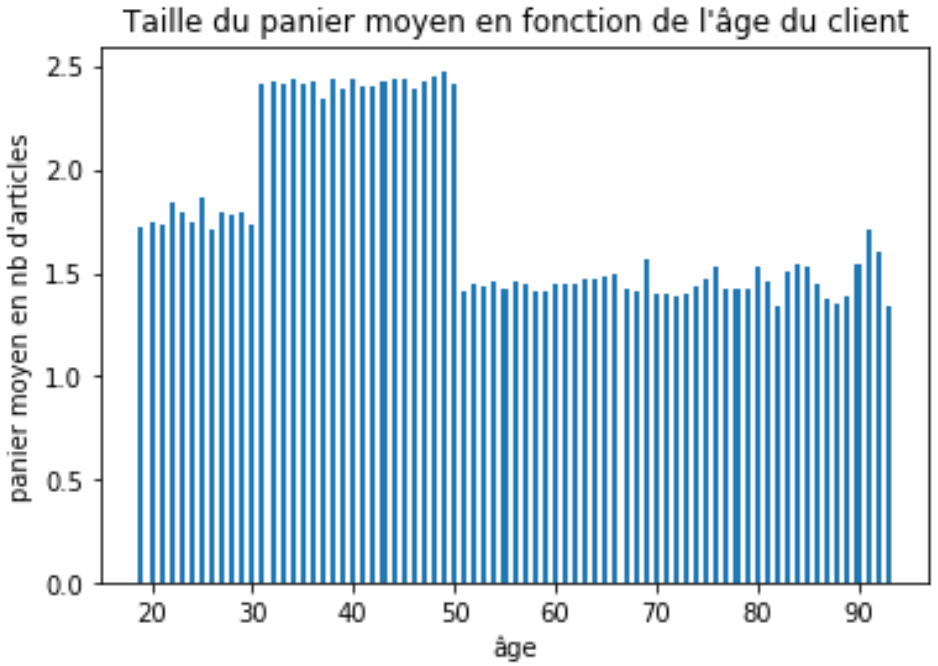


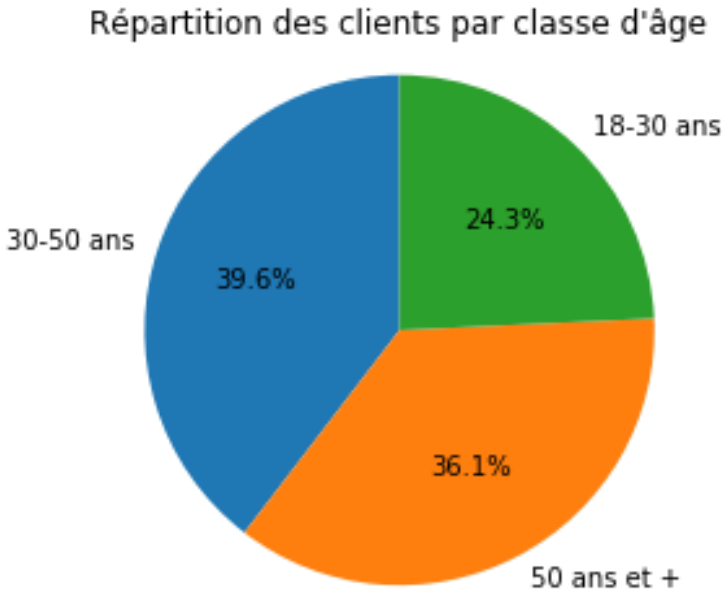
Possibilité d’évaluer les données manquantes.
J’ai choisi de laisser les informations telles quelles faute d’en savoir plus

Analyse des données
Prendre en compte les clients en fonction de leur âge et de leur classe d'âge

2. Concernant les clients

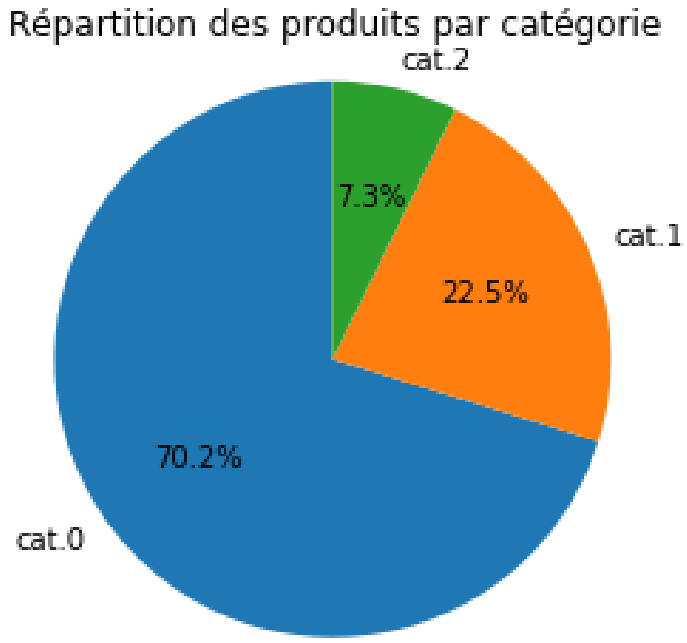
La répartition des achats incite à faire des analyses par classes d'âge

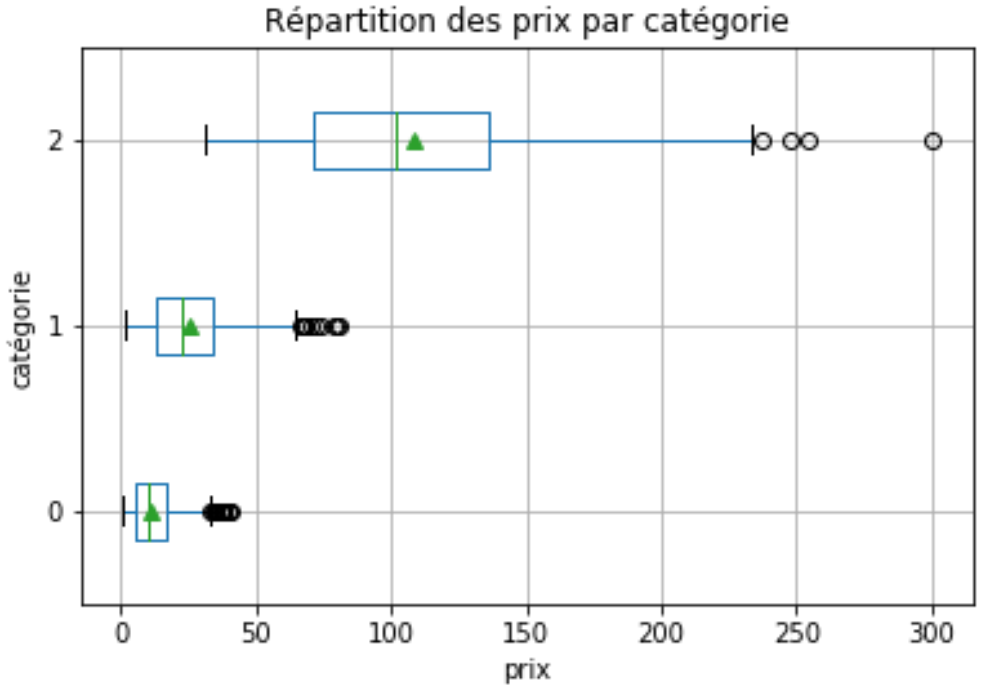




Analyse des données
Les produits, leurs catégories, leurs prix

3. Concernant les produits



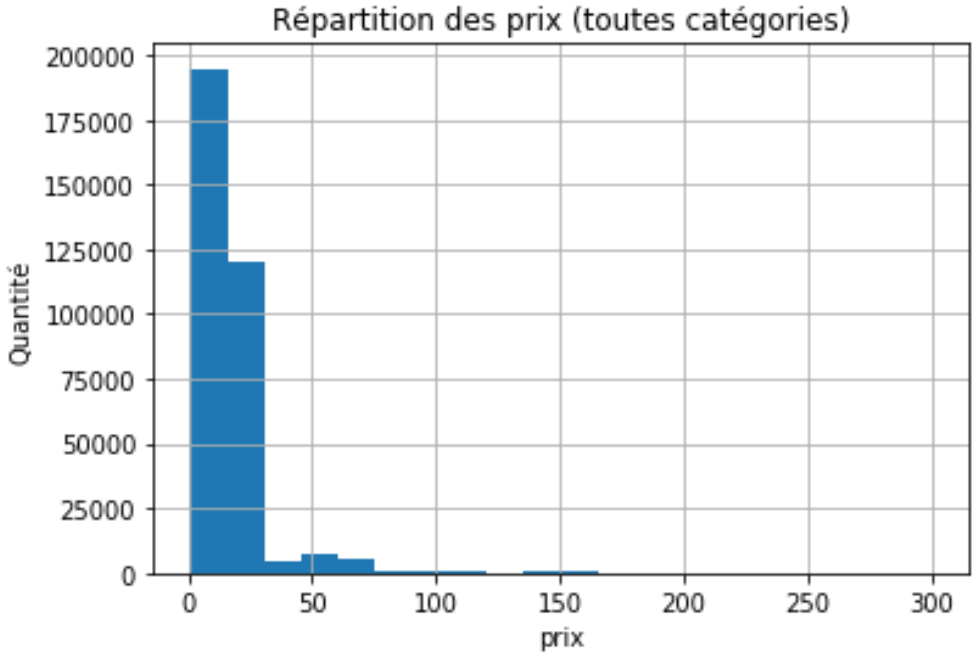


Jérôme Chalain – mai 2020

Analyse des données

3. Concernant les produits : focus sur les prix

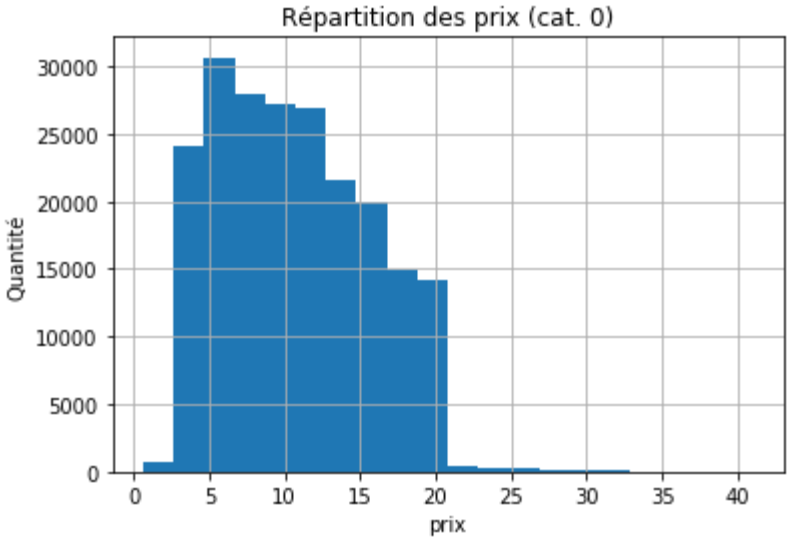
```
Toutes catégories confondues :
- Moyenne des prix: 17.21
- Médiane: 13.9
- Mode: 15.99
- Min: 0.62
- Max: 300.0
- Variance: 318.92935594812457
- Écart-type: 17.85859333621001
- Coefficient de variation générale: 1.0374186886958934
- Skewness empirique: 5.4805797800337315
- Kurtosis empirique: 45.43645969950359
Chiffre d'affaires (arrondi) :  6e+06 €
```



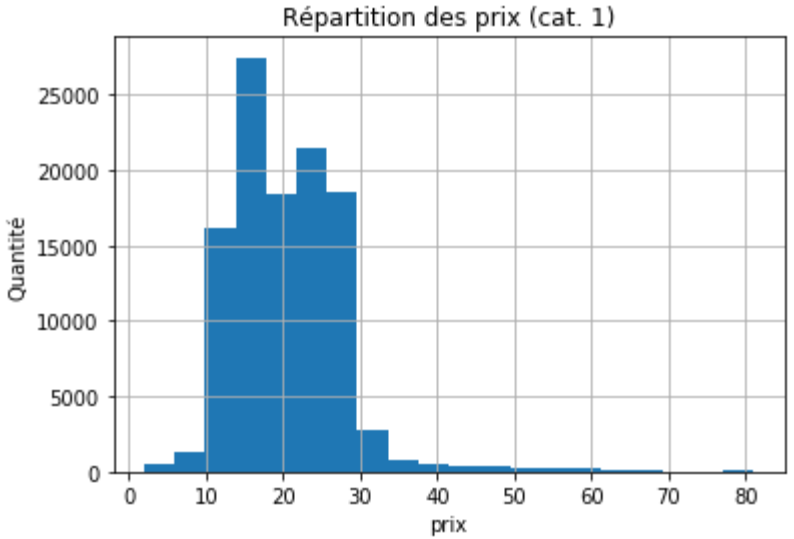
Analyse des données

3. Concernant les produits : focus sur les prix

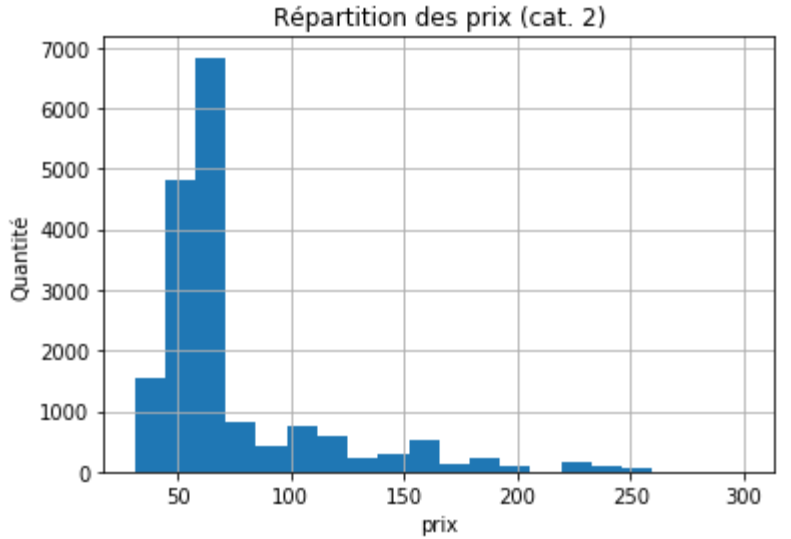
```
Catégorie 0 :
- Moyenne des prix: 10.65
- Médiane: 9.99
- Mode: 4.99
- Min: 0.62
- Max: 40.99
- Variance: 24.30230205875296
- Écart-type: 4.929736510073633
- Coefficient de variation générale: 0.46301224873288
- Skewness empirique: 0.4271396950359897
- Kurtosis empirique: -0.3700310009088872
Chiffre d'affaires (arrondi) :  2.23e+06 €
```




```
Catégorie 1 :
- Moyenne des prix: 20.48
- Médiane: 19.08
- Mode: 15.99
- Min: 2.0
- Max: 80.99
- Variance: 57.29260059571928
- Écart-type: 7.569187578315078
- Coefficient de variation générale: 0.36958061357470
- Skewness empirique: 1.727168284476662
- Kurtosis empirique: 8.152797423622884
Chiffre d'affaires (arrondi) :  2.25e+06 €
```



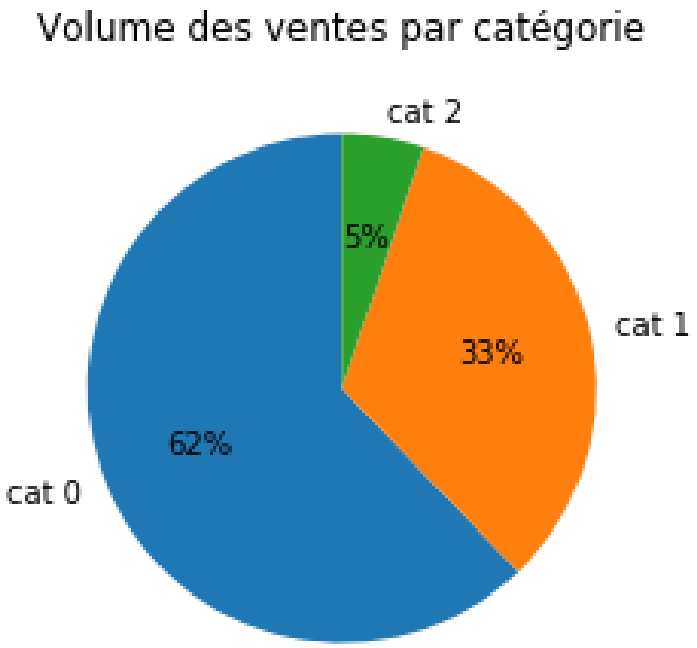

```
Catégorie 2 :
- Moyenne des prix: 75.19
- Médiane: 62.83
- Mode: 68.99
- Min: 30.99
- Max: 300.0
- Variance: 1528.6318301421995
- Écart-type: 39.09772154668606
- Coefficient de variation générale: 0.519968904497272
- Skewness empirique: 2.3143212706241654
- Kurtosis empirique: 5.528654721731269
Chiffre d'affaires (arrondi) :  1.32e+06 €
```

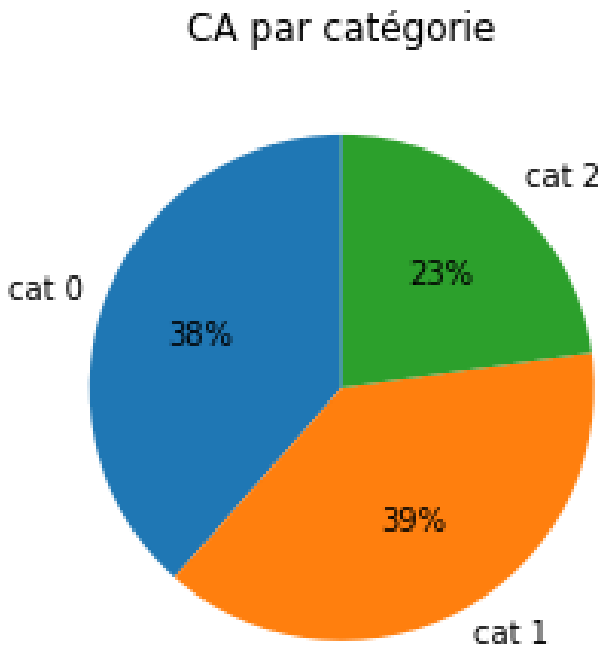


Analyse des données

Les produits : des volumes de ventes très différents, des contributions significatives au CA dans tous les cas !

3. Concernant les produits

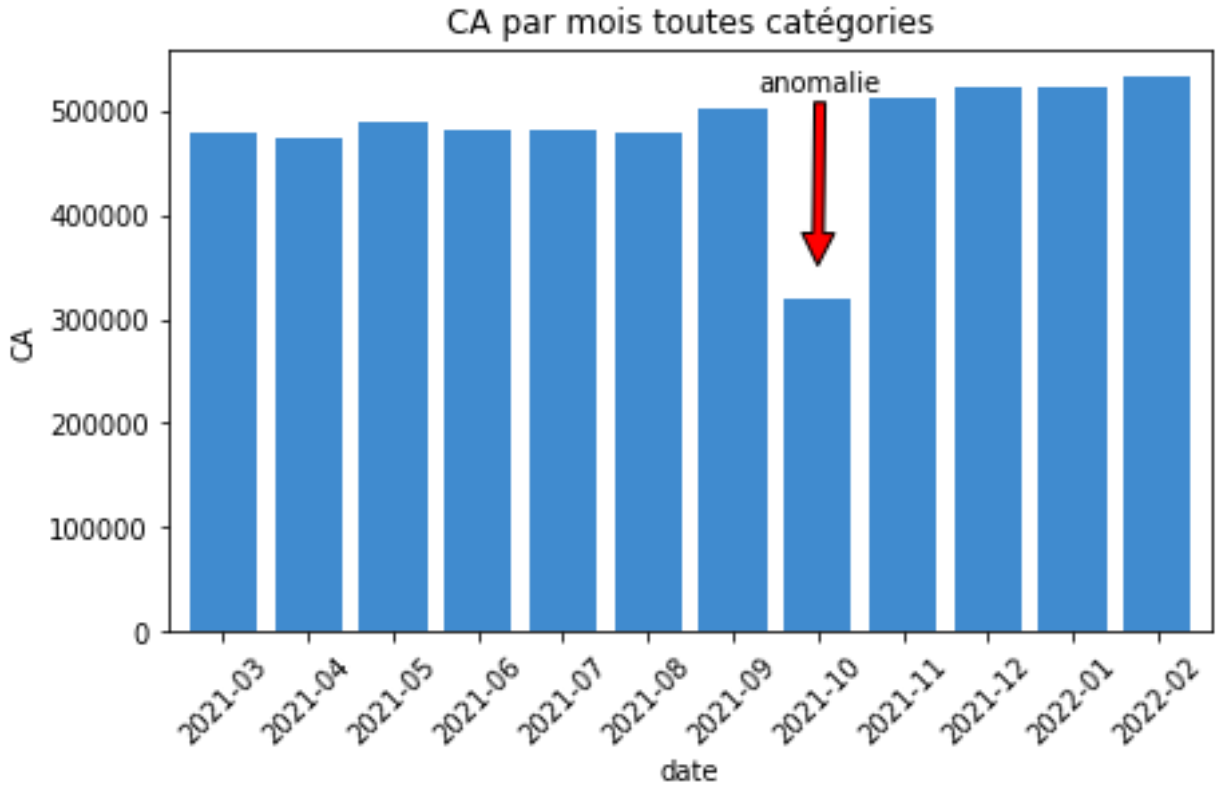





```
Sur la période, le CA global est de 5798472.68841556 € (environ 6e+06 €)
- cat 0 :  2.23e+06 €
- cat 1 :  2.25e+06 €
- cat 2 :  1.32e+06 €
```

Analyse des données
Un CA relativement constant sur l'année

4. Autres analyses : le CA

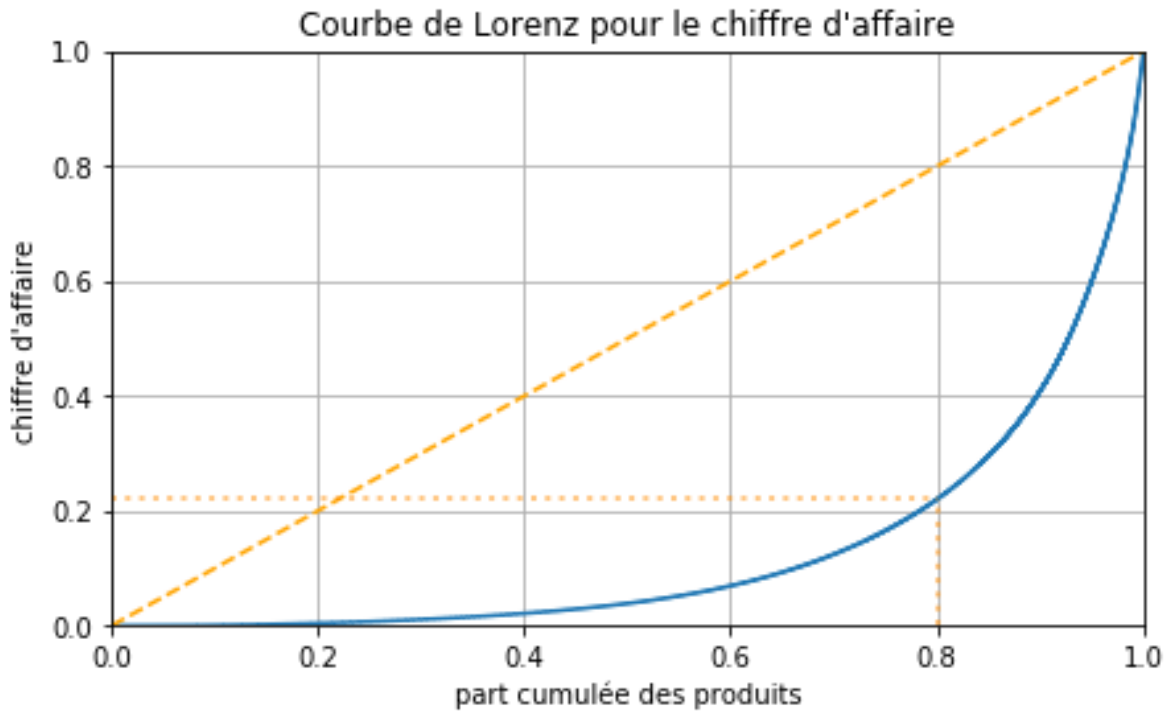


Jérôme Chalain – mai 2020

Analyse des données

4. Autres analyses : le CA

L'indice de Gini est égal à 0.74.
Rappel : l'inégalité est d'autant plus forte que l'indice de Gini est élevé.
Une petite partie des produits contribue à une grosse part du CA.

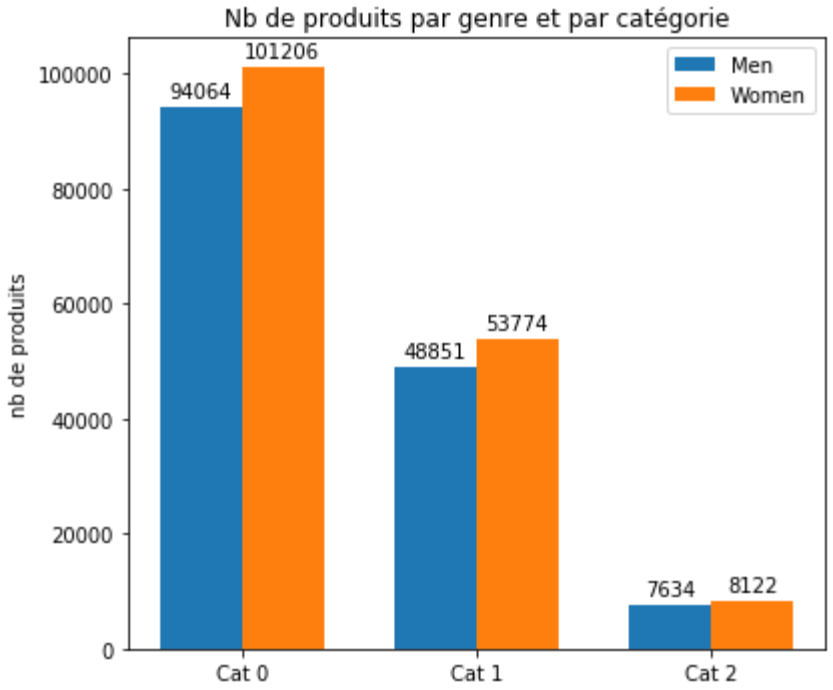


80 % du CA est fait par 20 % des produits

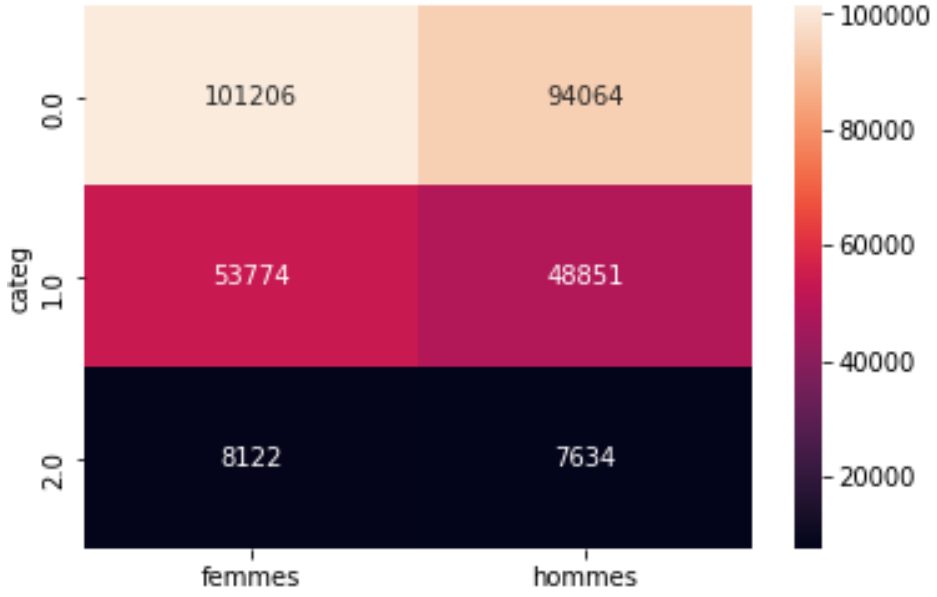
4. Analyse des corrélations

Analyse des corrélations

1. On observe une corrélation entre le sexe des clients et les catégories de produits achetés



Heatmap




```
p
```

```
0.006374607510181767
```

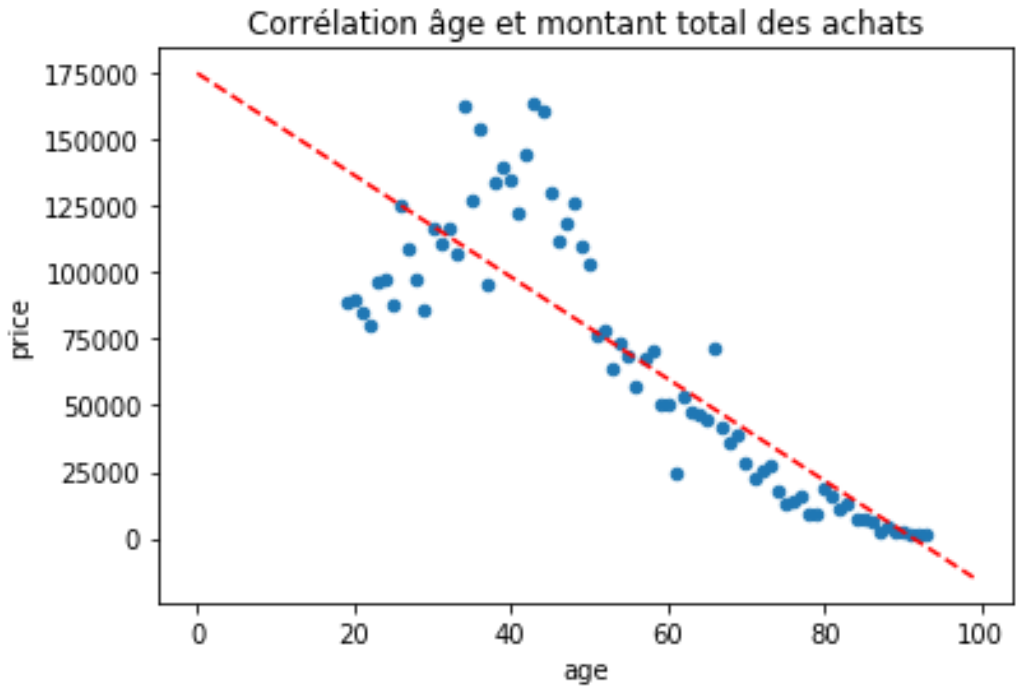
```
# Valeur du chi2 :
chi2
```

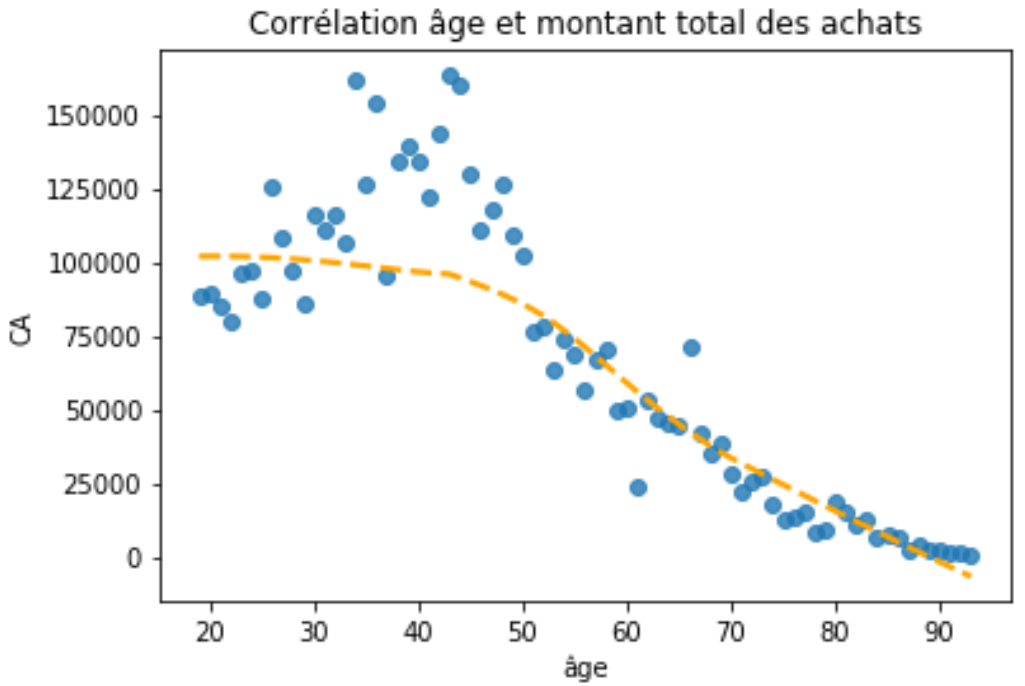
```
10.110865513726571
```

p-value < 0.05 et chi2 observé > chi2 théorique (5,99) => **rejet de l'hypothèse H0 d'indépendance entre les variables**

Analyse des corrélations

2a. On observe une corrélation entre l'âge des clients et le montant total des achats






```
Le coefficient de Pearson est de : -0.8502156672724805
Sa p-value est de : 4.956878115917968e-22
```

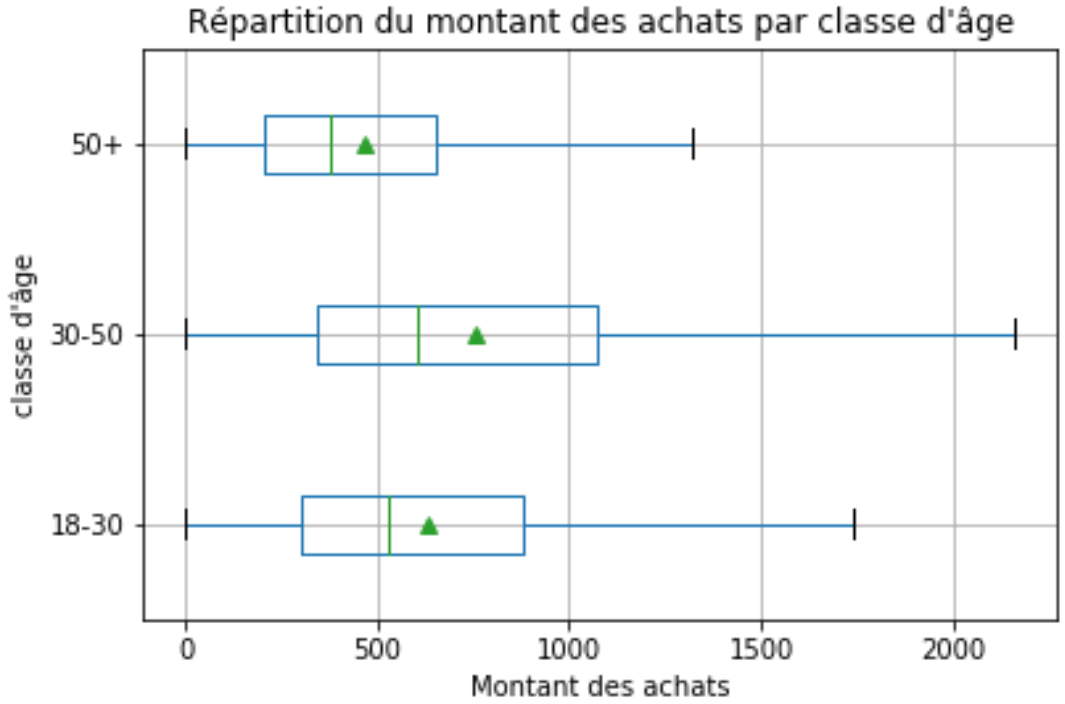
Recherche d'effets non-linéaires

Jérôme Chalain – mai 2020

Analyse des corrélations

2a. On observe une corrélation entre la classe d'âge des clients et le montant total des achats

ANOVA



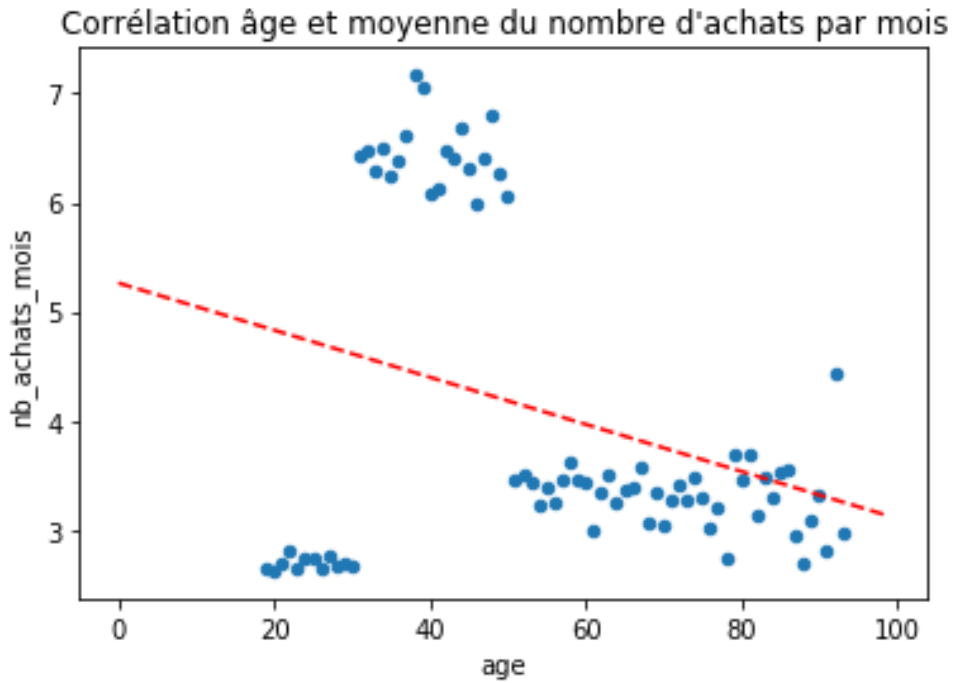
F ≈ 363 η2 ≈ 0.081

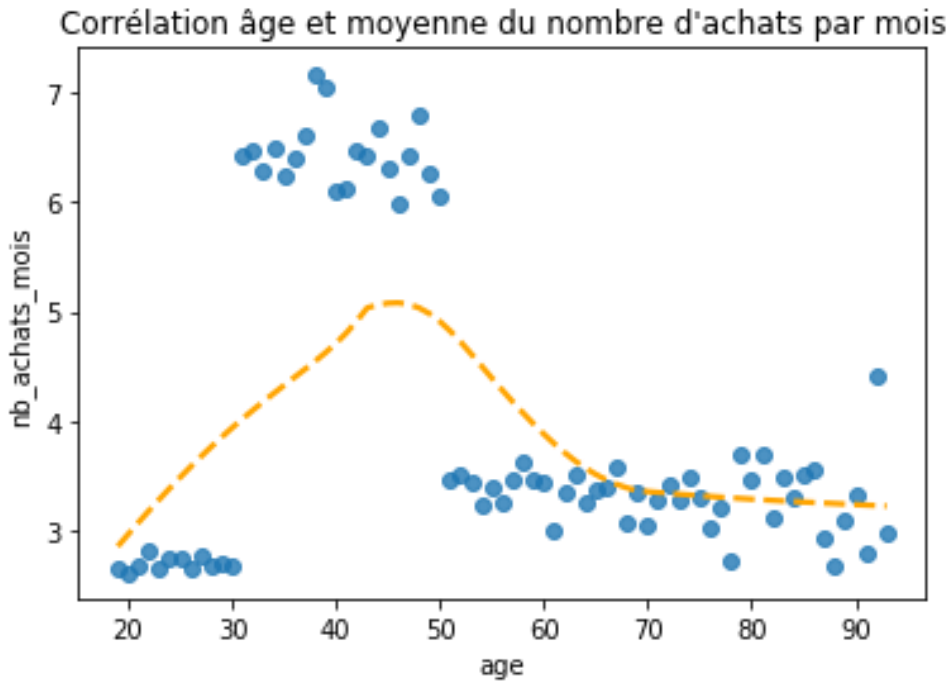
p ≈ 0

On rejette H0.

Analyse des corrélations

2b. Pas de corrélation entre l'âge des clients et la fréquence d’achat (nombre d'achats par mois par exemple)






```
Le coefficient de Pearson est de : -0.31517922107795426
Sa p-value est de : 0.005882460170220738
```

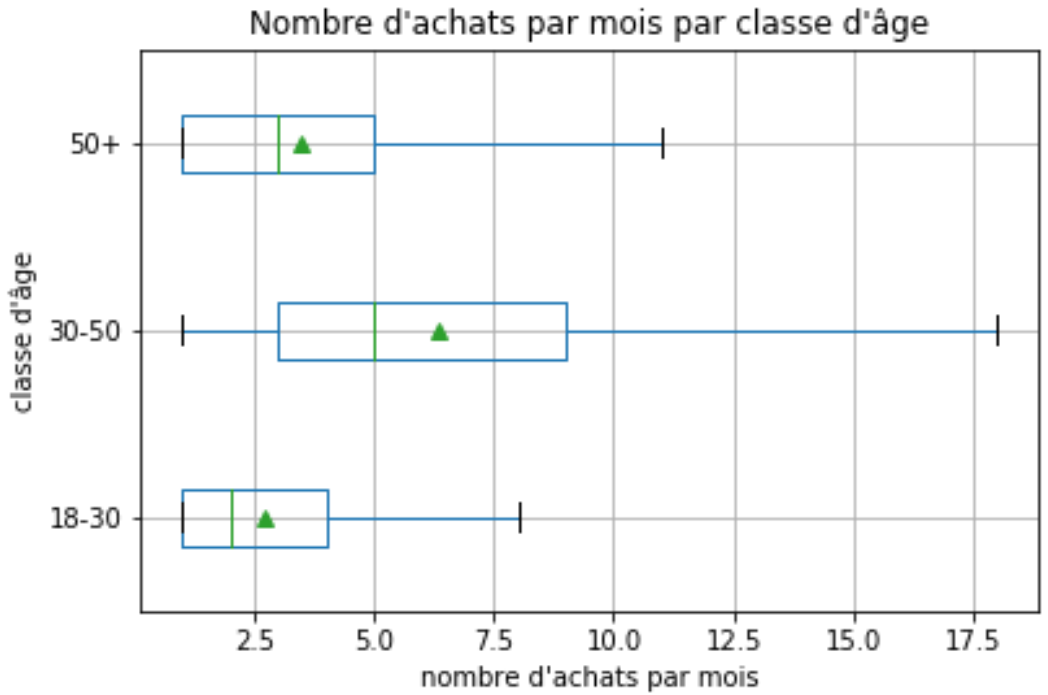
mais pas de relation linéaire

Jérôme Chalain – mai 2020

Analyse des corrélations

2b. Corrélation entre la classe d’âge des clients et la fréquence d’achat (nombre d'achats par mois par exemple)

ANOVA



F ≈ 5230

p ≈ 0

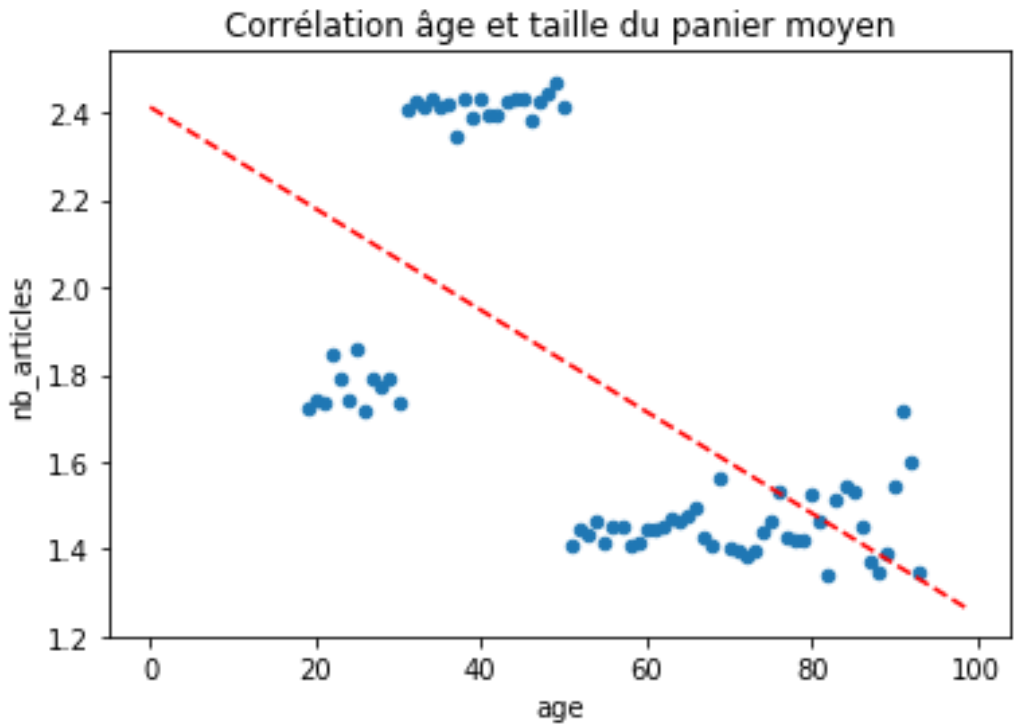
On rejette H0.

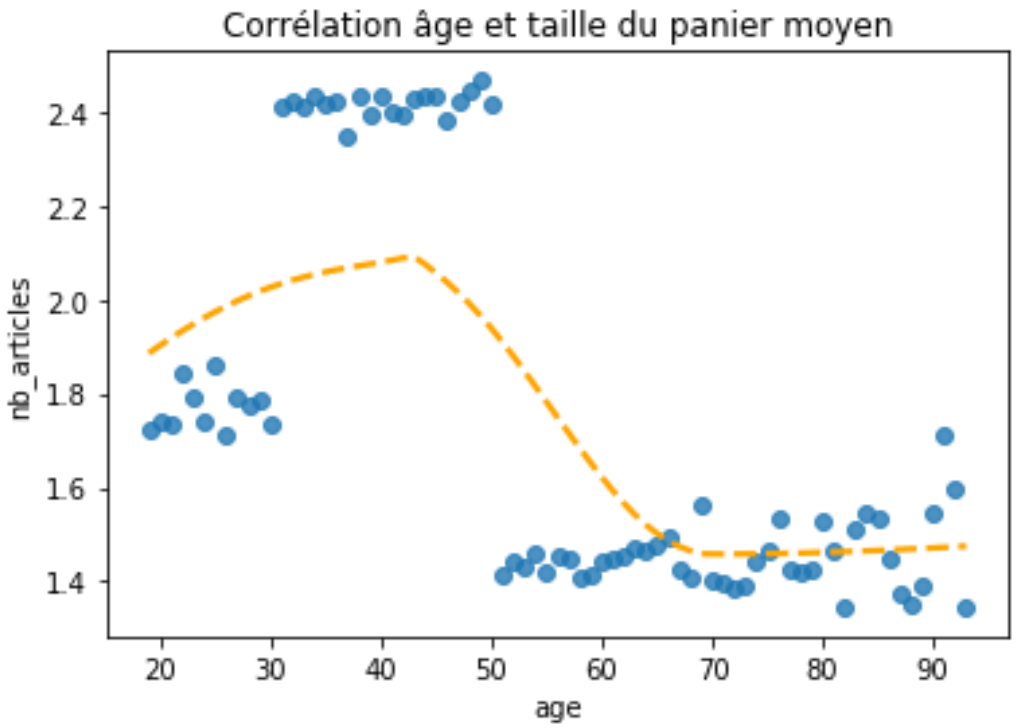
η2 ≈ 0.14

Jérôme Chalain – mai 2020

Analyse des corrélations

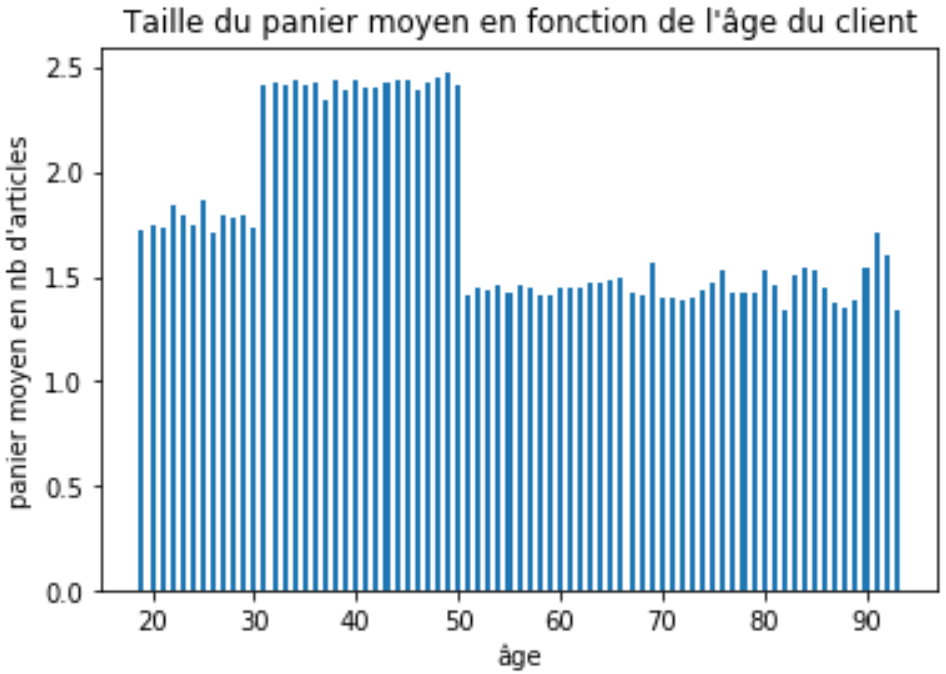
2c. Y a-t-il une corrélation entre l'âge des clients et la taille du panier moyen (en nombre d'articles)



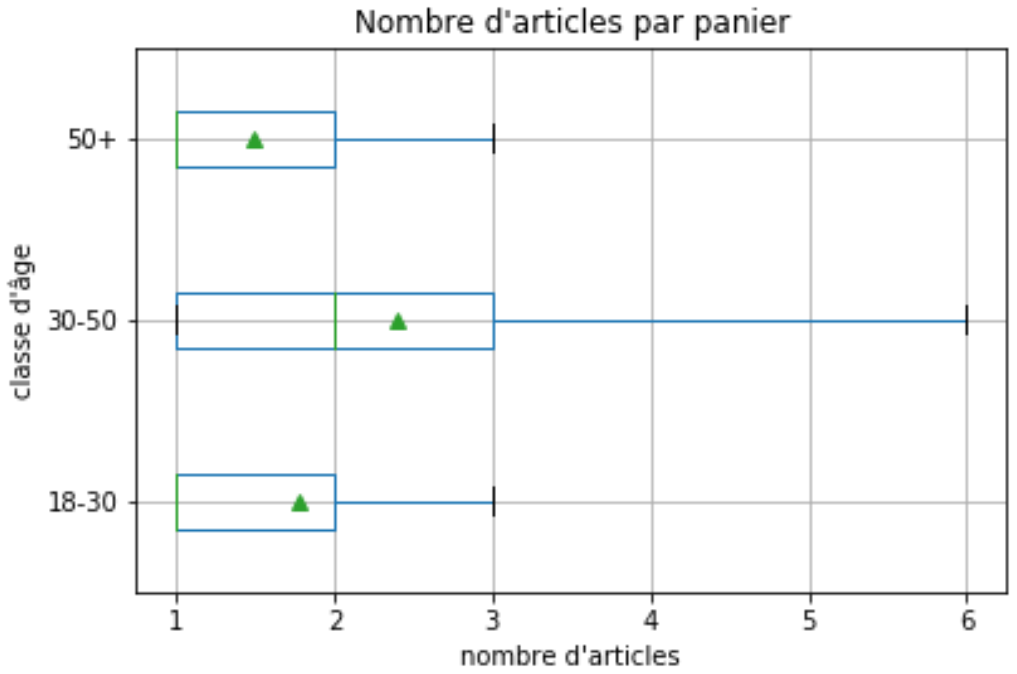


Analyse des corrélations

2c. Corrélation entre l'âge des clients et la taille du panier moyen (en nombre d'articles)



ANOVA



F ≈ 9993 η2 ≈ 0.11

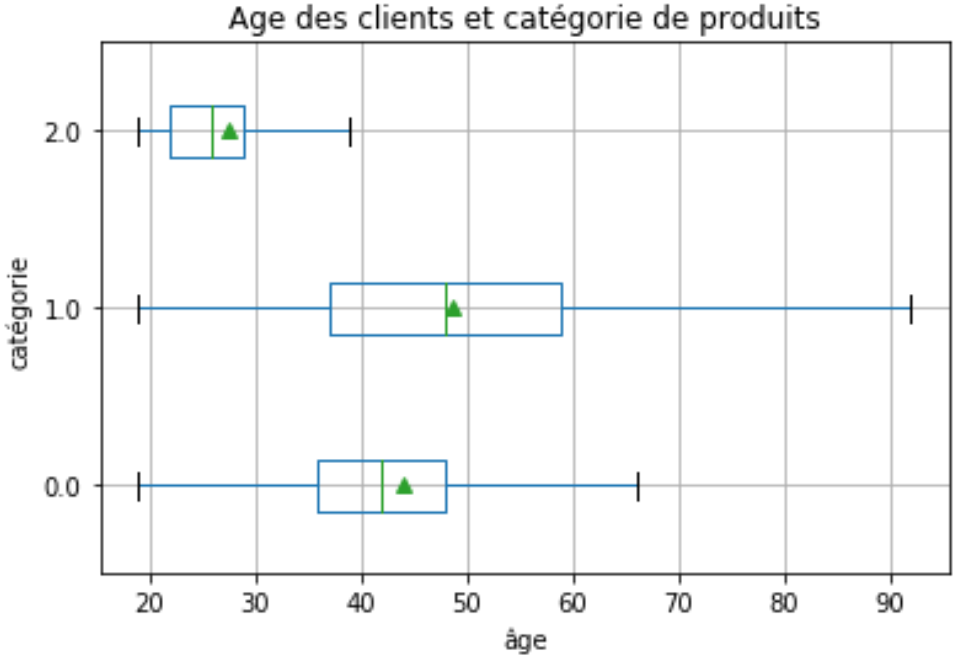
p ≈ 0

On rejette H0.

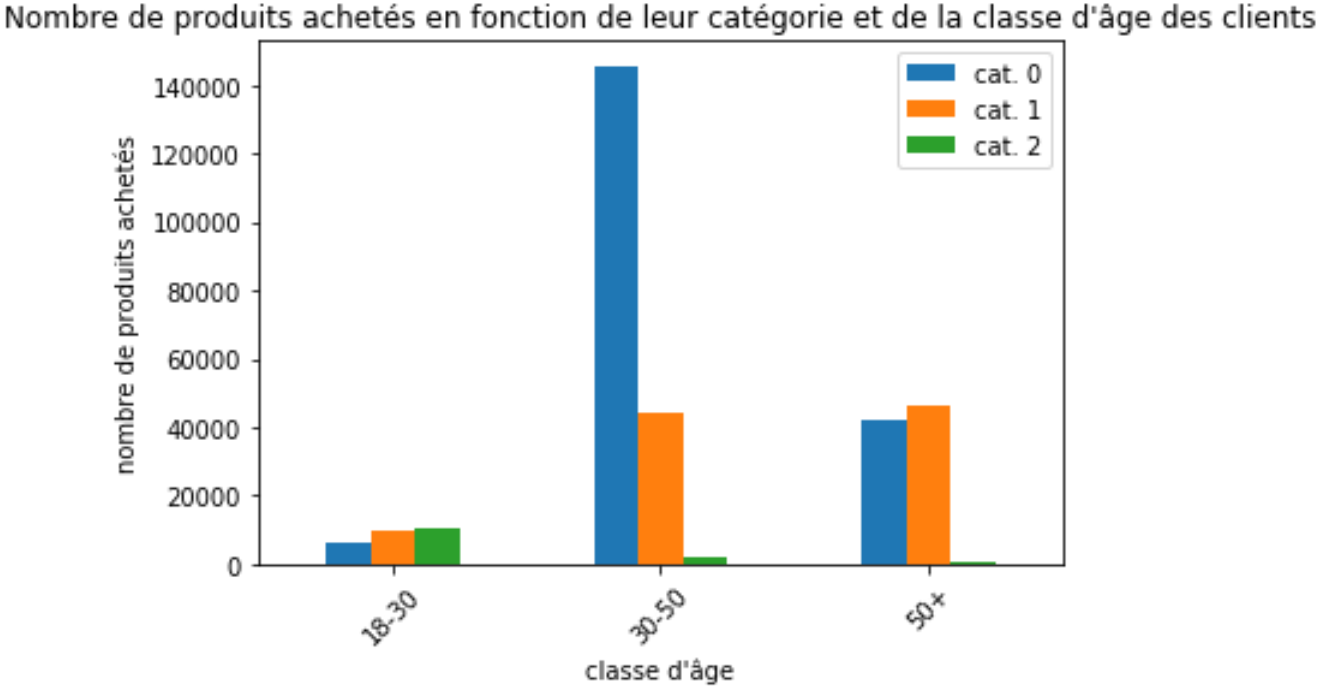
Analyse des corrélations

2d. On observe une corrélation entre l'âge des clients et les catégories de produits achetés

ANOVA



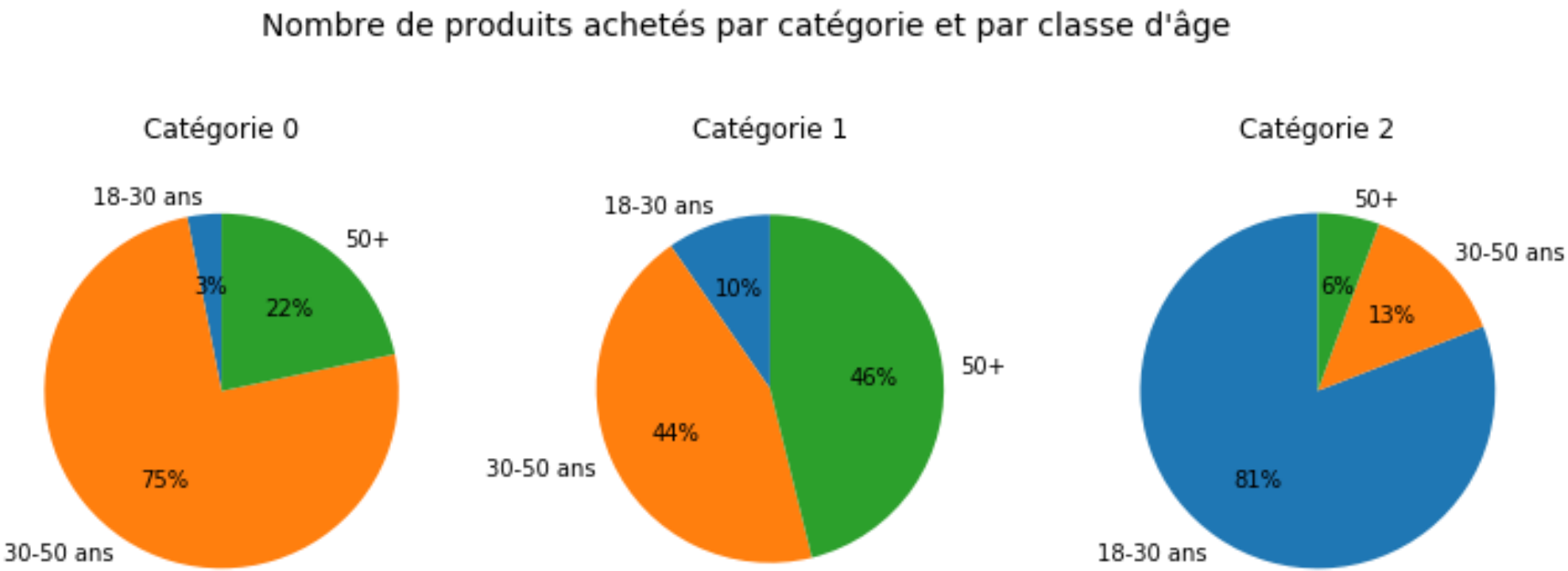
F ≈ 16906 et p ≈ 0.
On rejette H0.



p ≈ 0 et chi2 observé > chi2 théorique.
On rejette H0.

Analyse des corrélations

2d. On observe une corrélation entre l'âge des clients et les catégories de produits achetés



Jérôme Chalain – mai 2020

5. Conclusion

Conclusion
Nous connaissons mieux notre clientèle

- **Une classe d'âge à choyer** : **les 30-50 ans**

Ils représentent **40 %** des clients et plus de **47 % du CA**

Ils achètent **¾ des produits de la catégorie 0** et **44 % des produits de la catégorie 1**.

Conclusion
Nous connaissons mieux notre clientèle

- Les **18-30 ans** représentent **24 % des clients et 26 % du CA**.

Ils achètent plus de **80 % des produits de la catégorie 2**.

- Les **+ de 50 ans** représentent **36 % des clients et 27 % du CA**.

Ils achètent presque **la moitié des produits de la catégorie 1**.

Conclusion
Et maintenant ?

Propositions :

- **Développer** le potentiel de la classe d'âge 18-30 en priorité puis 50+.

Pour cela, pousser les ventes des produits de la catégorie 2 ?

- **Préciser** les données pour la classe d'âge 18 ans

- **Approfondir** les recherches sur les produits, à l'intérieur de chaque catégorie

- **Attention** aux données manquantes

- De nombreuses **autres analyses** sont envisageables !

Par exemple : ajuster les prix en fonction des ventes (yield management).

Merci de votre attention.