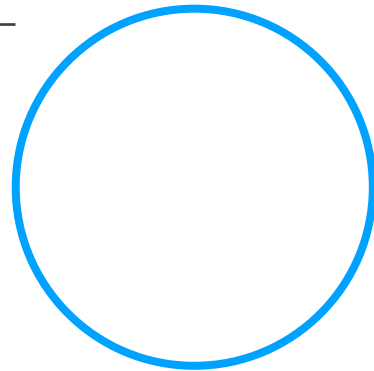


Étude de marché internationale

MAXIMEBCH – DATA ANALYST



INTRODUCTION

- Entreprise française de volaille souhaitant exporter ses produits à l'international
- Objectif : Réalisation d'une étude de marché pour identifier les pays à cibler





Sommaire

- 1 – NETTOYAGE DES DONNÉES
- 2 – ANALYSE DES DONNÉES
- 4 – CONCLUSIONS ET RECOMMANDATIONS



1 - NETTOYAGE DES DONNÉES



1 – NETTOYAGE DES DONNÉES

FICHIERS SOURCES

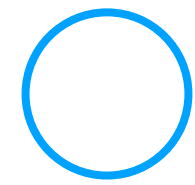


4 fichiers .csv extraits des données de la FAO :

- Bilan alimentaire : disponibilités alimentaires en calories et en protéines, d'origine animale ou non
- Marché du poulet : production, importation et exportation de poulet
- Population : chiffres de 2008 et 2018
- Sécurité et stabilité : PIB par habitant, indice de stabilité politique et d'absence de violence/terrorisme

Les fichiers ne présentent pas de doublons ou de valeurs manquantes.

On supprime la zone « Chine » qui est un agrégat de la Chine continentale, Taïwan, Hong Kong et Macao.



1 – NETTOYAGE DES DONNÉES

CRÉATION D'UN SEUL DATAFRAME

Après concaténation et pivot, les données sont rassemblées par pays dans un seul dataframe.

De nouvelles variables ont été créées :

- Ratio_protéines_animales = protéines animales / total protéines
- Population_croissance = population 2018 / population 2008
- Ppa (PIB par habitant) = PIB / population
- PIB_croissance = PIB 2018 / PIB 2008
- Poulet_import-export = importations 2018 / exportations 2018

pays	dispo_calories	dispo_proteines	ratio_proteines_animales	population	population_croissance	pib	ppa	pib_croissance	poulet_import-export
Afghanistan	2040.0	55.52	0.194344	37.171921	1.340868	2190.2	0.000059	1.379393	155.279221
Albania	3360.0	115.74	0.533523	2.882740	0.960056	13601.3	0.004718	1.344122	NaN
Algeria	3322.0	91.83	0.269302	42.228408	1.215885	11479.5	0.000272	1.080424	NaN
American Samoa	NaN	NaN	NaN	0.055465	0.964743	NaN	NaN	NaN	NaN
Andorra	NaN	NaN	NaN	0.077006	0.918247	NaN	NaN	NaN	NaN

1 – NETTOYAGE DES DONNÉES

CRÉATION DU DATAFRAME

Après suppression des valeurs manquantes et nulles :

- 106 pays
- 83% de la population mondiale


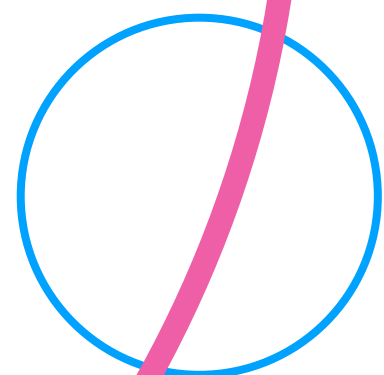

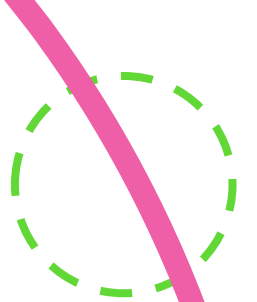
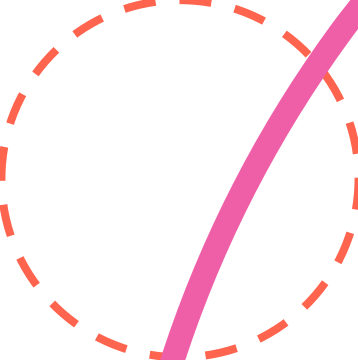

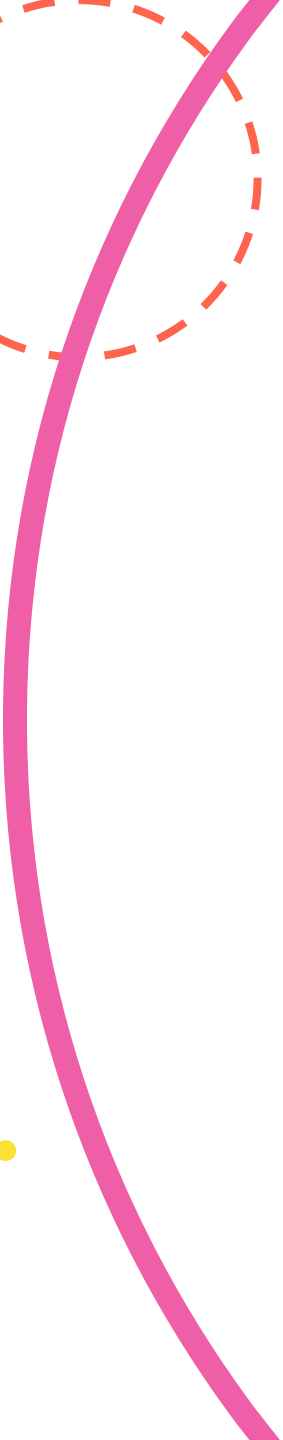
pays	dispo_calories	dispo_proteines	ratio_proteines_animales	population	population_croissance	pib	ppa	pib_croissance	poulet_import-export
Afghanistan	2040.000000	55.520000	0.194344	37.171921	1.340868	2190.200000	0.000059	1.379393	155.279221
Antigua and Barbuda	2445.000000	80.190000	0.645218	0.096286	1.127510	21115.800000	0.219303	0.914029	849.142857
Argentina	3307.000000	106.770000	0.646717	44.361150	1.106811	22745.900000	0.000513	0.984513	0.040777
Armenia	2997.000000	94.350000	0.480551	2.951745	1.015176	12715.000000	0.004308	1.214712	435.297297
Australia	3391.000000	105.940000	0.667359	24.898152	1.167158	49576.000000	0.001991	1.096894	0.145991
Austria	3695.000000	109.120000	0.603281	8.891388	1.065918	55687.200000	0.006263	1.041688	1.090448
Azerbaijan	3149.000000	94.420000	0.344524	9.949537	1.127826	14209.600000	0.001428	1.101255	54.939689
Bahamas	2655.000000	80.690000	0.649399	0.385637	1.122078	35500.500000	0.092057	0.902964	19536.000000
Bangladesh	2563.000000	60.730000	0.206817	161.376708	1.118309	4441.400000	0.000028	1.670327	110.500000
Barbados	2956.000000	88.690000	0.575262	0.286641	1.023915	15674.900000	0.054685	0.924953	14.460317

1 – NETTOYAGE DES DONNÉES

GEOPLOT

- GEOPLOT est une librairie Python permettant le traçage géospatial
- Jointure avec un DF contenant le continent et la géolocalisation de chaque pays

	continent	pays	iso_a3	geometry	dispo_calories	dispo_proteines	ratio_proteines_animaux	population	population_croissance	pib
0	Oceania	Fiji	FJI	MULTIPOLYGON (((180.00000 -16.06713, 180.00000...	2781.0	71.14	0.399635	0.883483	1.045096	13808.1
1	Africa	United Republic of Tanzania	TZA	POLYGON ((33.90371 -0.95000, 34.07262 -1.05982...	2373.0	58.93	0.206007	56.313438	1.345475	2590.0
2	North America	Canada	CAN	MULTIPOLYGON (((122.84000 49.00000, -122.9742...	3566.0	104.12	0.545044	37.074562	1.112093	48924.4



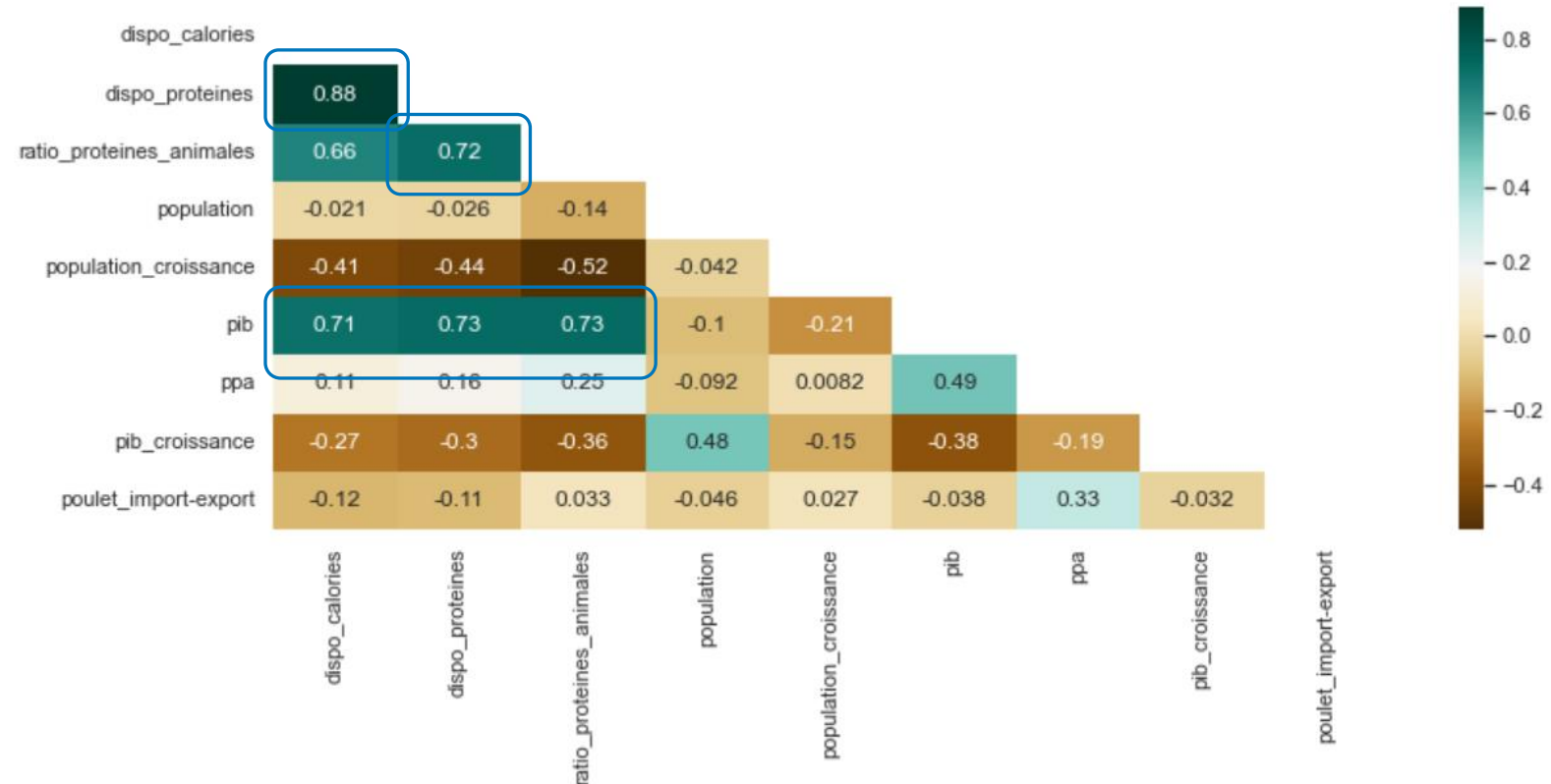
2 – CLASSIFICATION HIÉRARCHIQUE

1 – CLASSIFICATION HIÉRARCHIQUE

CORRÉLATIONS

La matrice ne révèle pas de fortes corrélations excepté :

- Entre la disponibilité en protéines et en calories
- Entre le PIB (niveau de richesse) et disponibilité alimentaire et ratio de protéines animales





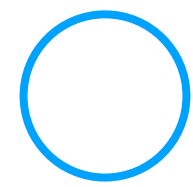
1 – CLASSIFICATION HIÉRARCHIQUE

NORMALISATION



Normalisation de variables (entre 0 et 1)

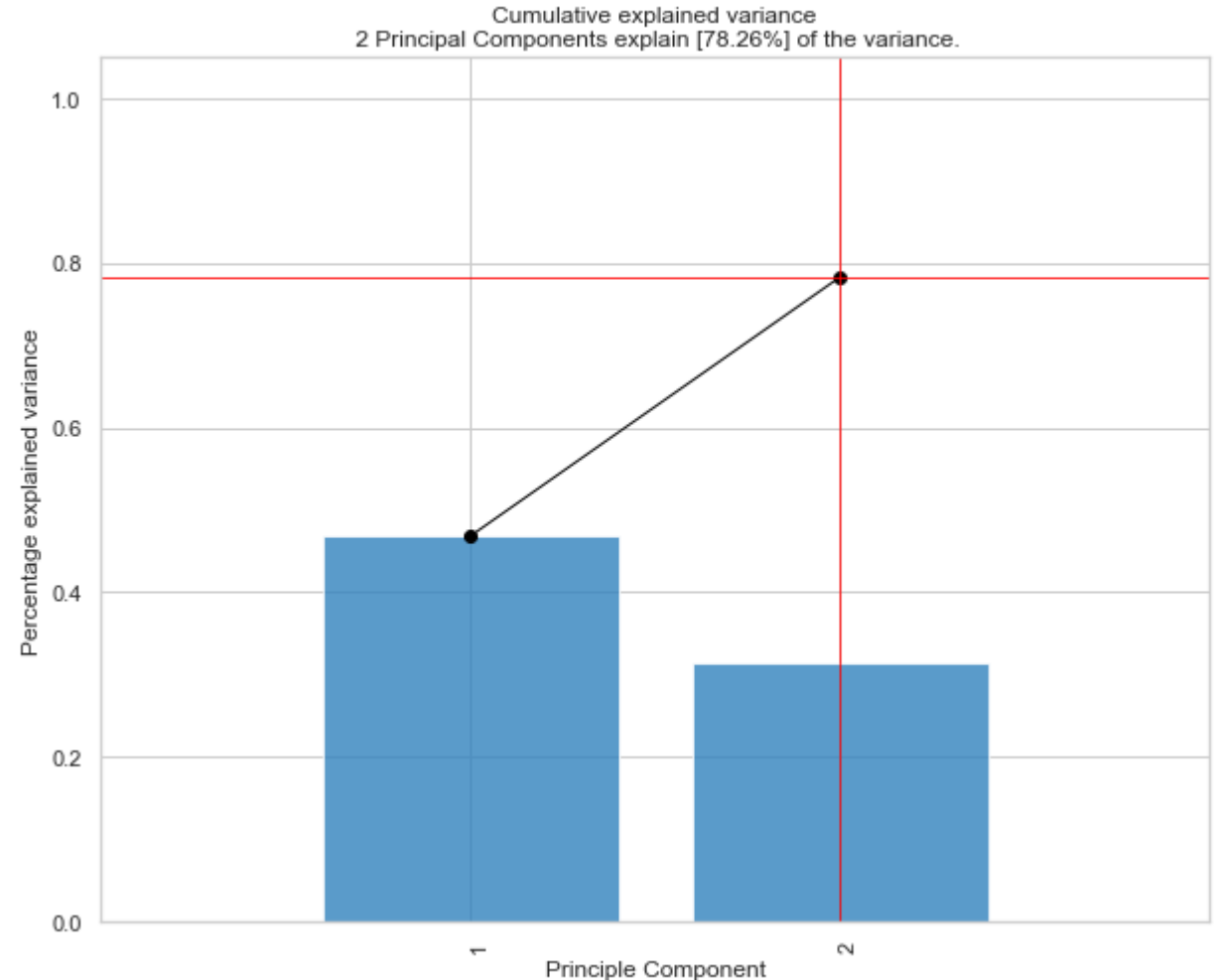
	ratio_proteines_animaux	population_croissance	pib_croissance	dispo_calories	dispo_proteines	population	pib	ppa	poulet_impo exp
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	1.000000e+02	100.0000
mean	0.454784	1.132678	1.197033	0.227394	0.006225	0.005891	0.941274	1.907658e-07	0.0309
std	0.143988	0.152900	0.218673	0.192240	0.005115	0.021517	0.105064	3.599276e-07	0.1157
min	0.141283	0.871890	0.691271	0.030351	0.000952	0.000005	0.421799	6.761622e-10	0.0000
25%	0.347529	1.029482	1.051039	0.082764	0.002498	0.000214	0.953808	1.930648e-08	0.0000
50%	0.487820	1.112964	1.185242	0.158154	0.004274	0.000761	0.987013	7.515395e-08	0.0001
75%	0.568652	1.198523	1.302479	0.280879	0.007577	0.003388	0.996512	1.837584e-07	0.0026
max	0.680987	1.755557	2.056280	0.905994	0.029119	0.190422	0.999539	2.266971e-06	0.7789



1 – CLASSIFICATION HIÉRARCHIQUE

ANALYSE EN COMPOSANTE PRINCIPALE

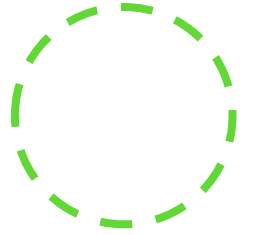
- Les variables catégorielles et la disponibilité en calories sont écartées
- Règle du coude : 2 composantes expliquent 78% de la variance





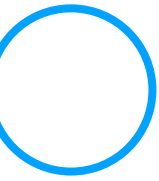
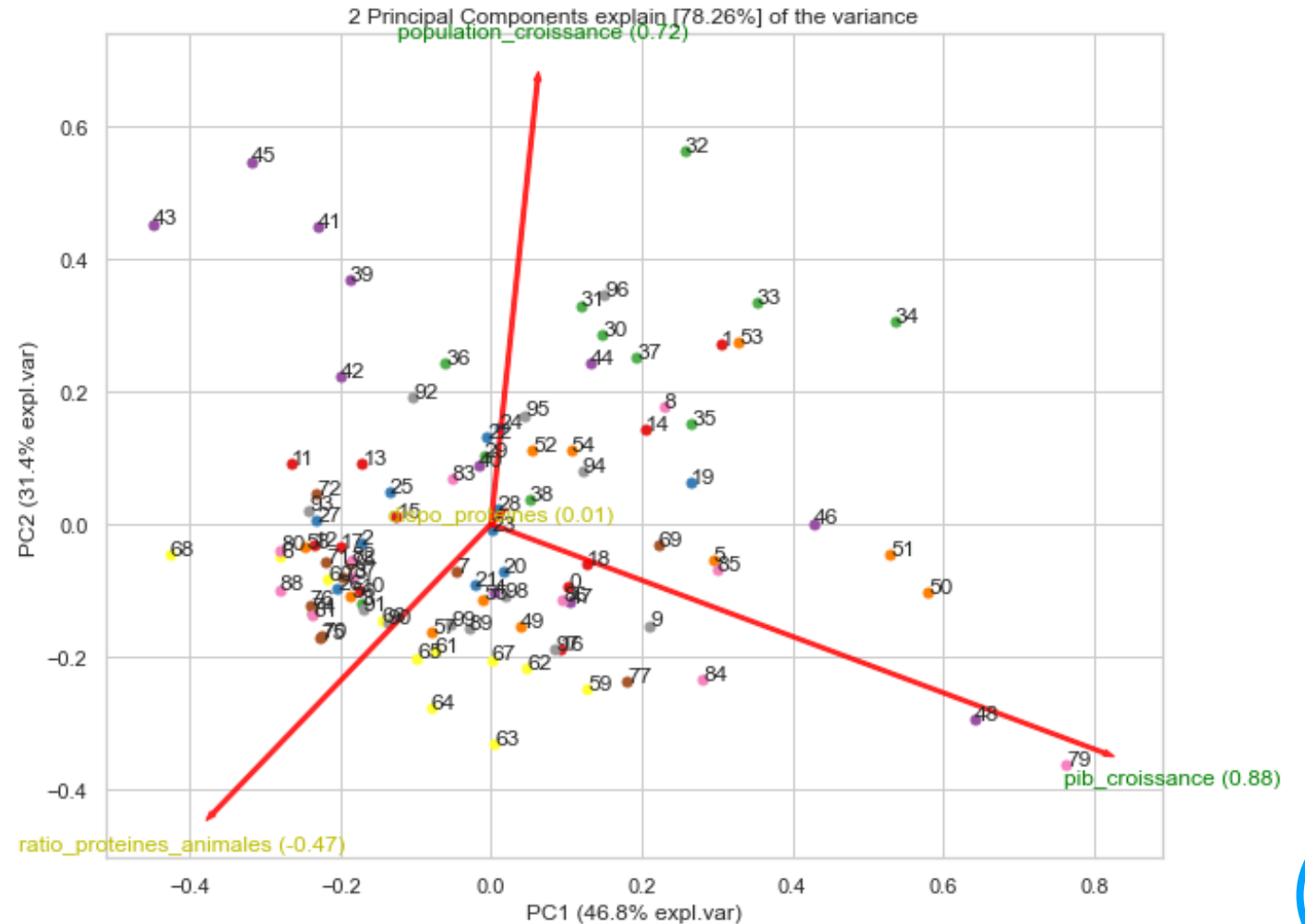
1 – CLASSIFICATION HIÉRARCHIQUE

REPRÉSENTATION EN 2 DIMENSIONS



La :

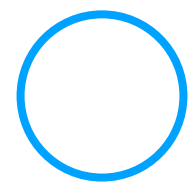
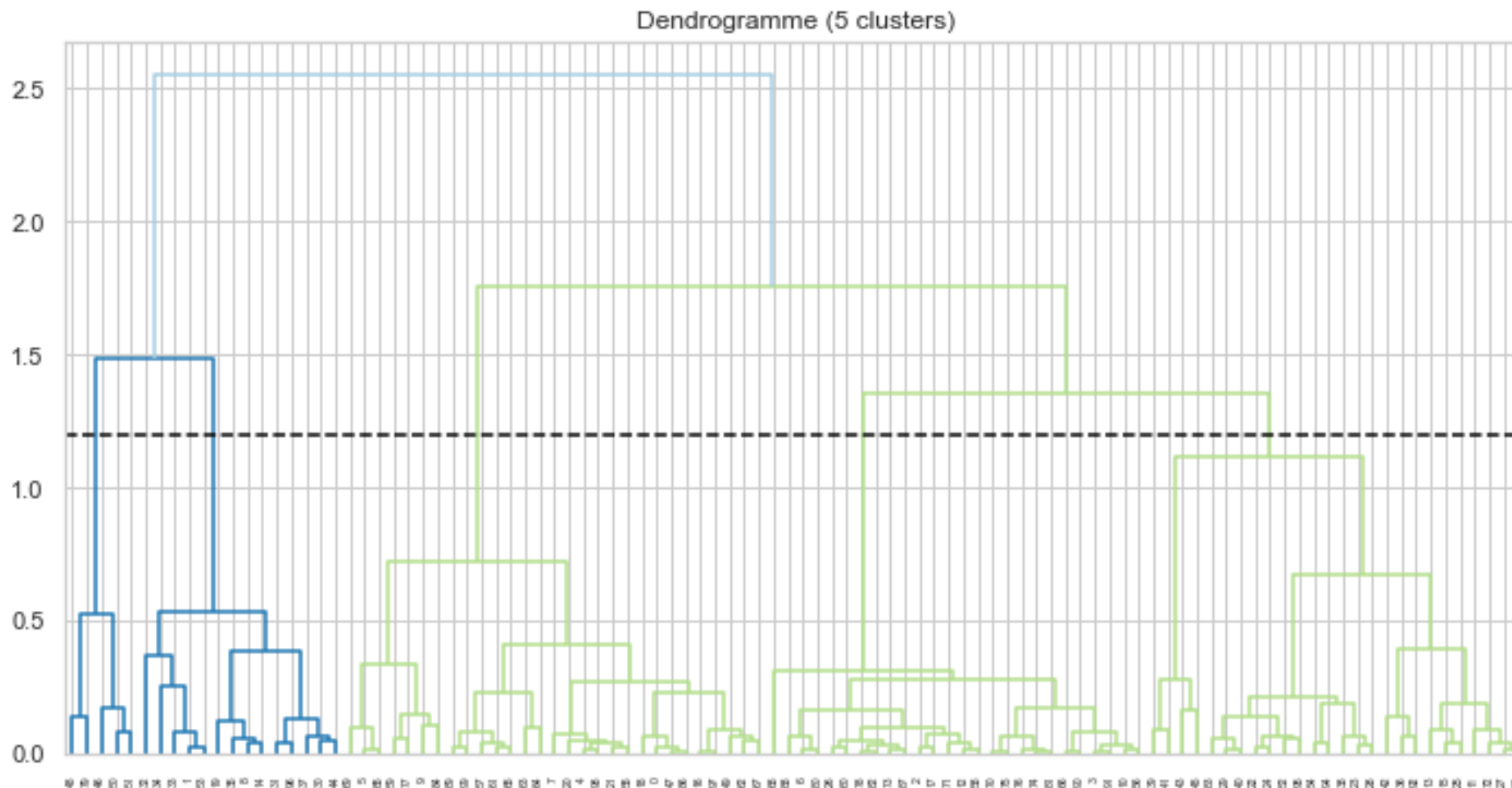
- Ent





1 – CLASSIFICATION HIÉRARCHIQUE

CLUSTERS



1 – NETTOYAGE DES DONNÉES

VALEURS ABERRANTES – OUTLIERS

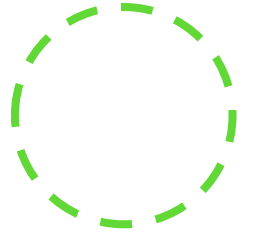
```
df = pd.merge(transactions, customers, on='client_id')
df = pd.merge(df, products, on = 'id_prod')
df['count'] = 1
df = df.groupby('client_id').sum().reset_index()
df = df.sort_values('count', ascending=False)
df = df[['client_id', 'count']]
df = pd.merge(df, customers, on='client_id')
df = df.sort_values('count', ascending=False)
top_10 = df.iloc[0:10]
print(top_10)
```

	client_id	count	sex	birth
0	c_1609	12855	m	1980
1	c_6714	4473	f	1968
2	c_3454	3275	m	1969
3	c_4958	2562	m	1999
4	c_2140	195	f	1977
5	c_7959	195	f	1974

```
mask = transactions.loc[(transactions['client_id'] == 'c_1609') | (transa
top_clients = mask.index.tolist()
transactions = transactions.drop(top_clients)
```

1 – NETTOYAGE DES DONNÉES

DONNÉES MANQUANTES : PRIX DE 0_2245



```
id_prod_false = transactions[transactions['id_prod_prod'] == False]
id_prod_false = id_prod_false.groupby('id_prod').mean()
id_prod_false
```

	id_prod_prod	client_id_custom
id_prod		
0_2245	False	True

```
products_0_2245 = products.loc[products['id_prod'] == '0_2245']
print(products_0_2245)
```

```
Empty DataFrame
Columns: [id_prod, price, categ]
Index: []
```

Le produit « 0_2245 » n'est pas dans nos données « products » mais dans celles « transactions ».

On peut lui attribuer la moyenne des prix de sa catégorie

```
transactions_m = pd.merge(transactions, products, on=['id_prod'])
transactions_m = pd.pivot_table(index='id_prod', columns='categ',
moy_cat0 = transactions_m[0].mean(skipna=True)
moy_cat0
```

11.718568310781567

```
products = products.sort_index()
print(products.tail())
```

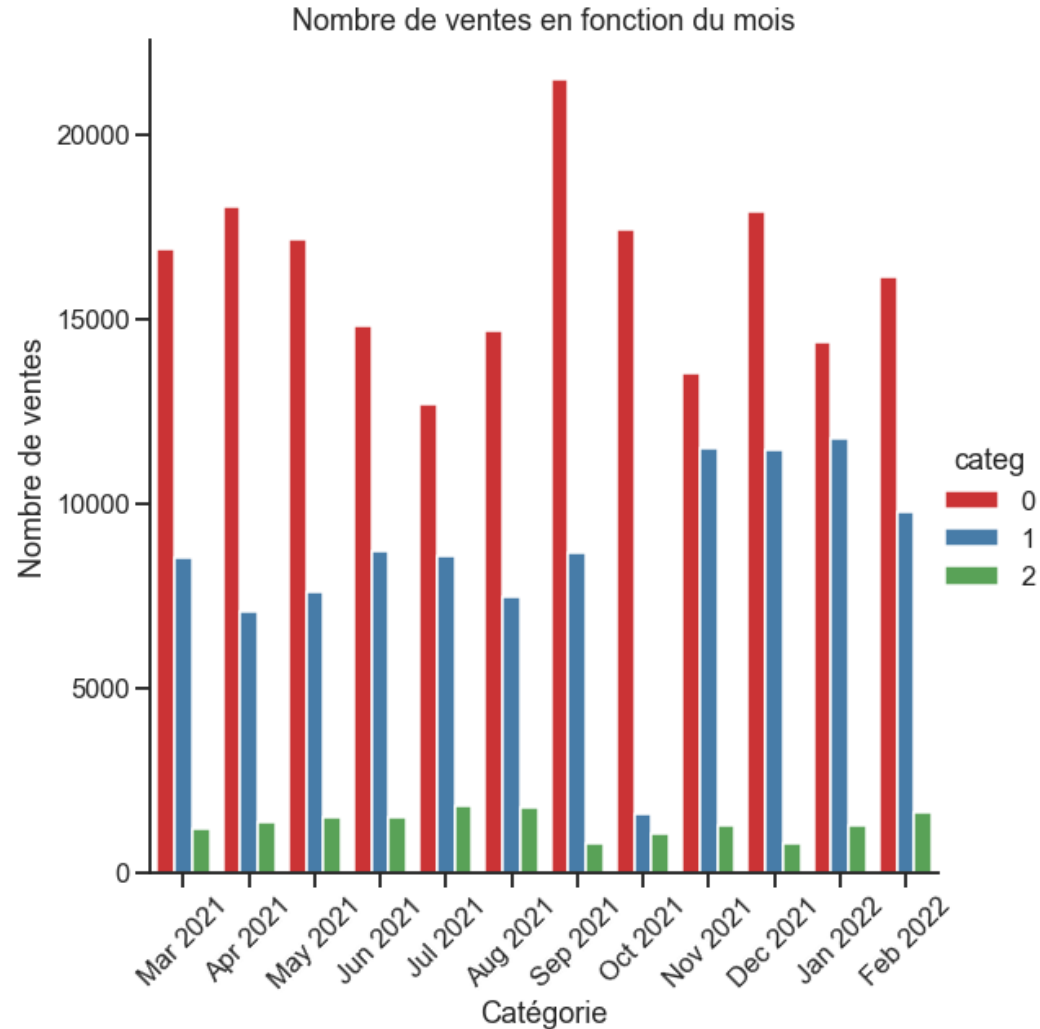
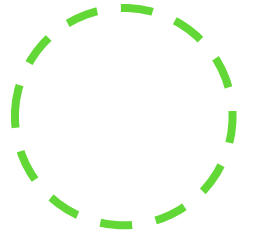
	id_prod	price	categ
3282	2_23	115.99	2
3283	0_146	17.14	0
3284	0_802	11.22	0
3285	1_140	38.56	1
3286	0_1920	25.16	0

```
products.loc[3287] = {'id_prod' : '0_2245', 'price' : 11.72, 'categ' : 0}
print(products.tail())
```

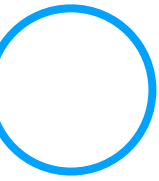
	id_prod	price	categ
3283	0_146	17.14	0
3284	0_802	11.22	0
3285	1_140	38.56	1
3286	0_1920	25.16	0
3287	0_2245	11.72	0

1 – NETTOYAGE DES DONNÉES

DONNÉES MANQUANTES : TRANSACTIONS D'OCTOBRE



- Perte de données : il manque des données des transactions d'octobre dans la catégorie 1



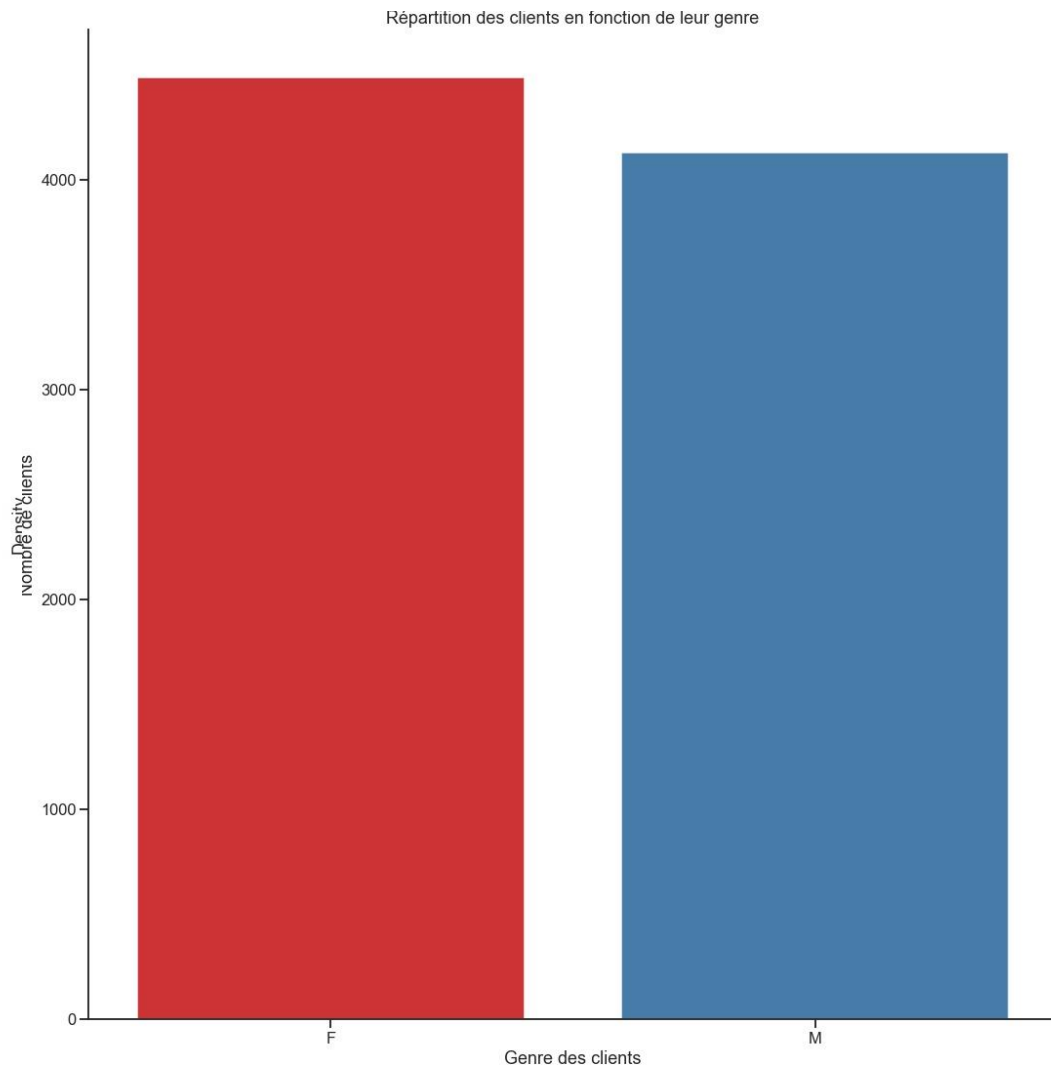


2 – ANALYSE DES DONNÉES

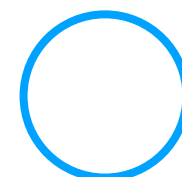


2 – ANALYSE DES DONNÉES

CLIENTS : DISTRIBUTION PAR GENRE

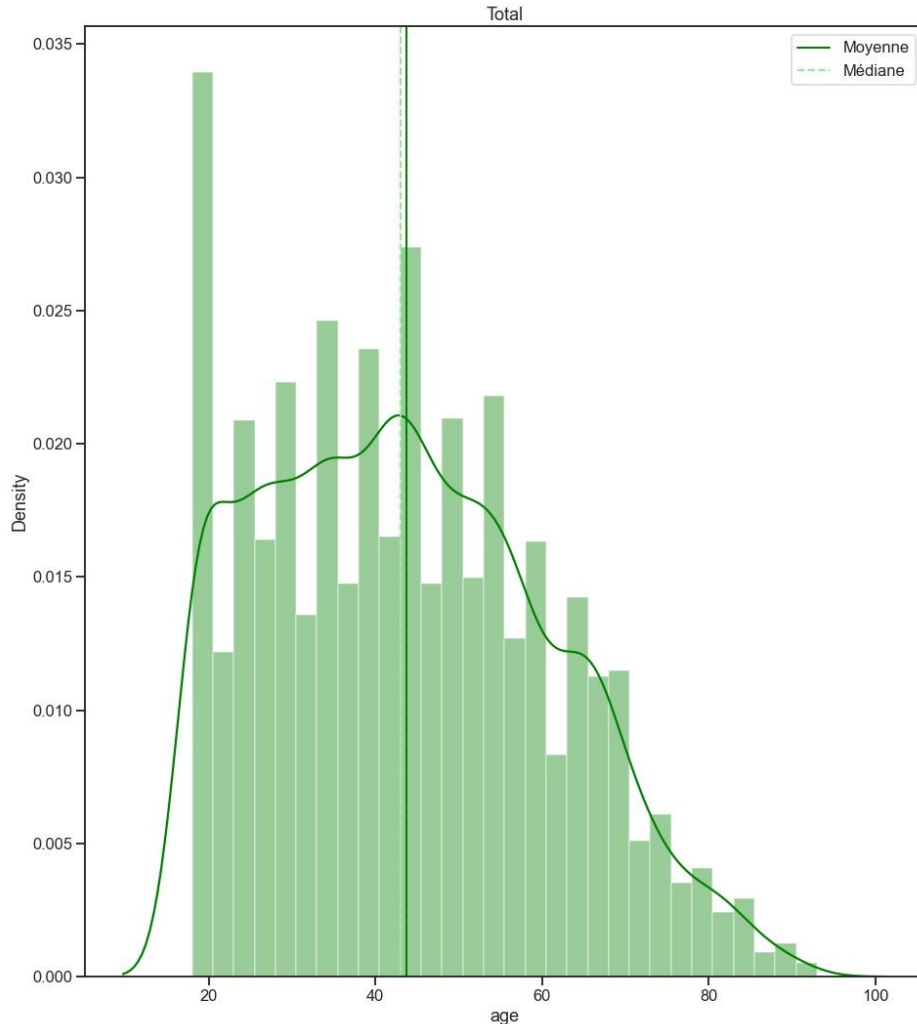


- Environ 8600 clients.
- Le nombre d'hommes et de femmes est pratiquement similaire.



2 – ANALYSE DES DONNÉES

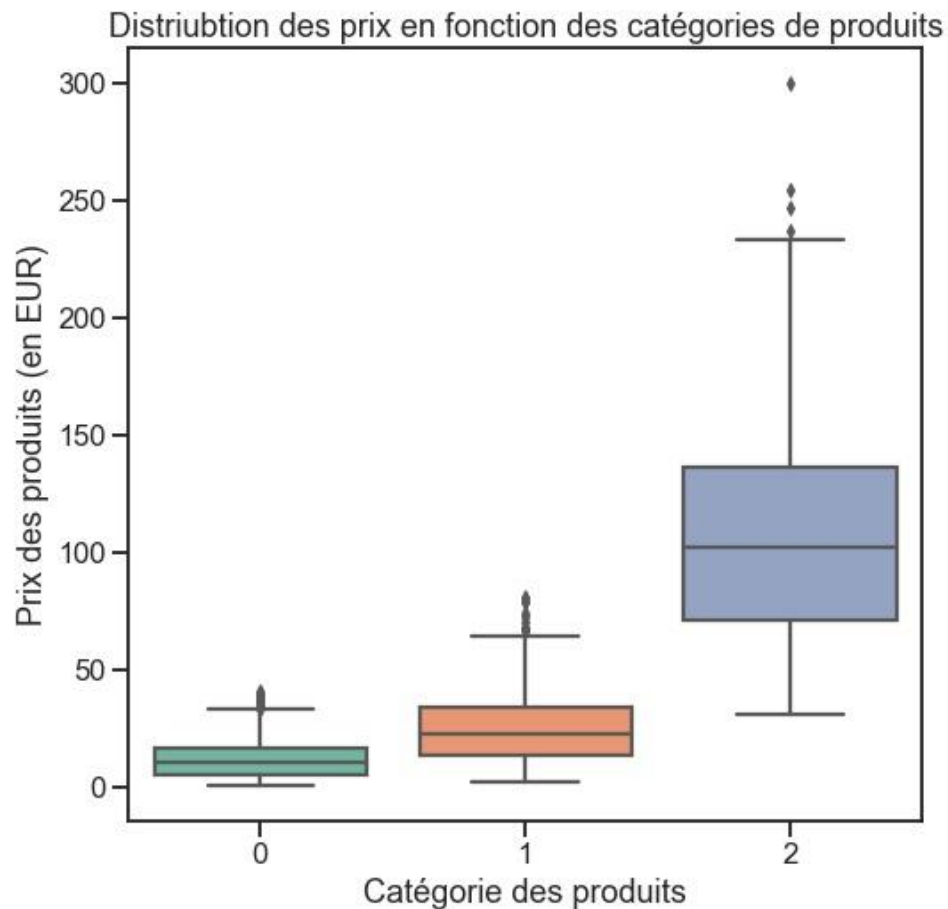
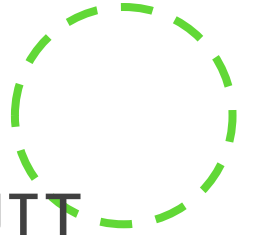
CLIENTS : DISTRIBUTION DES ÂGES



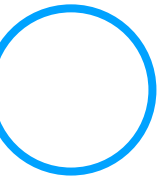
- La moyenne d'âge est de 43 ans.
- La distribution des âges est très symétriques chez les hommes et les femmes.
- Surreprésentation des acheteurs de 18 ans (conséquence de l'accès au site réservé aux majeurs).

2 – ANALYSE DES DONNÉES

PRIX : DISTRIBUTION SELON LES CATÉGORIES DE PRODUIT

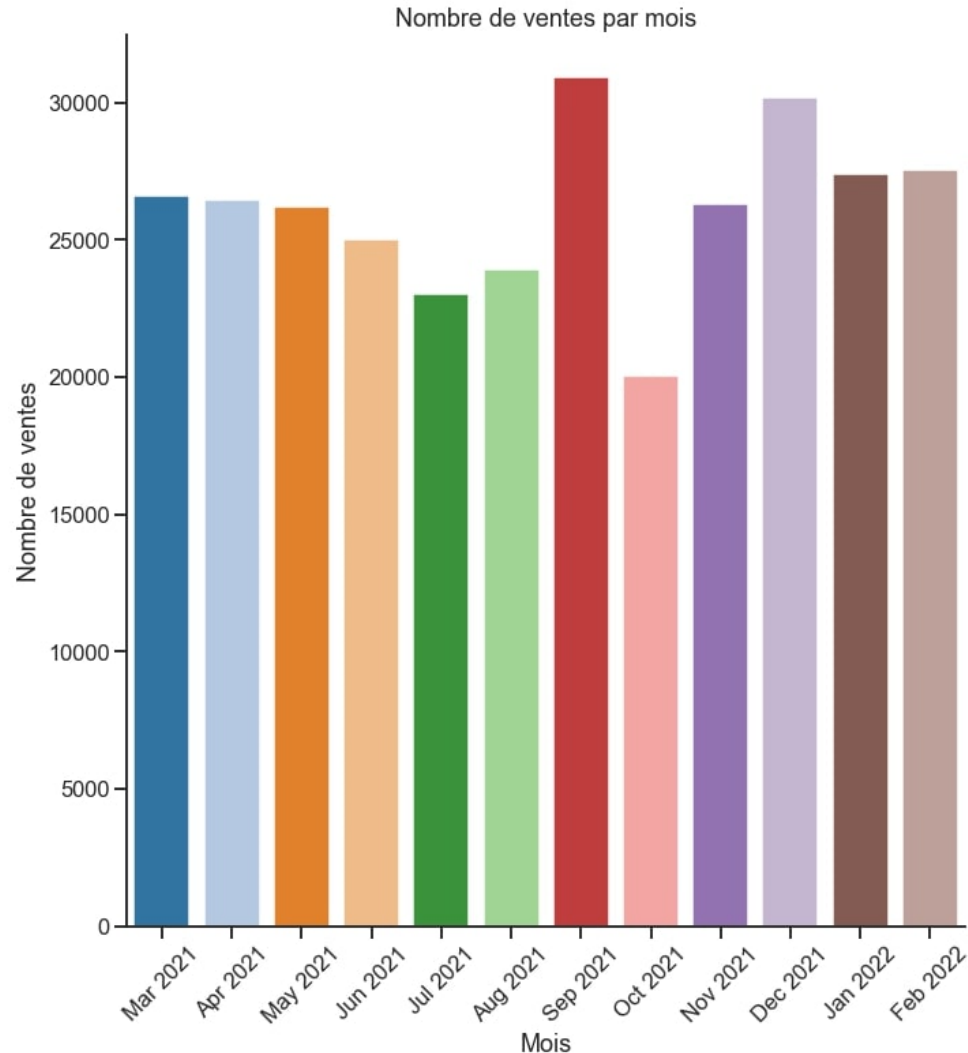


- Chaque catégorie correspond à un ordre de grandeur de prix croissant : la catégorie 0 a les prix les moins élevés et la catégorie 2 a les prix les plus élevés.

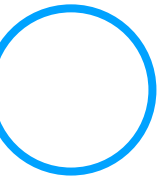


2 – ANALYSE DES DONNÉES

TRANSACTIONS : NOMBRE DE VENTES PAR MOIS

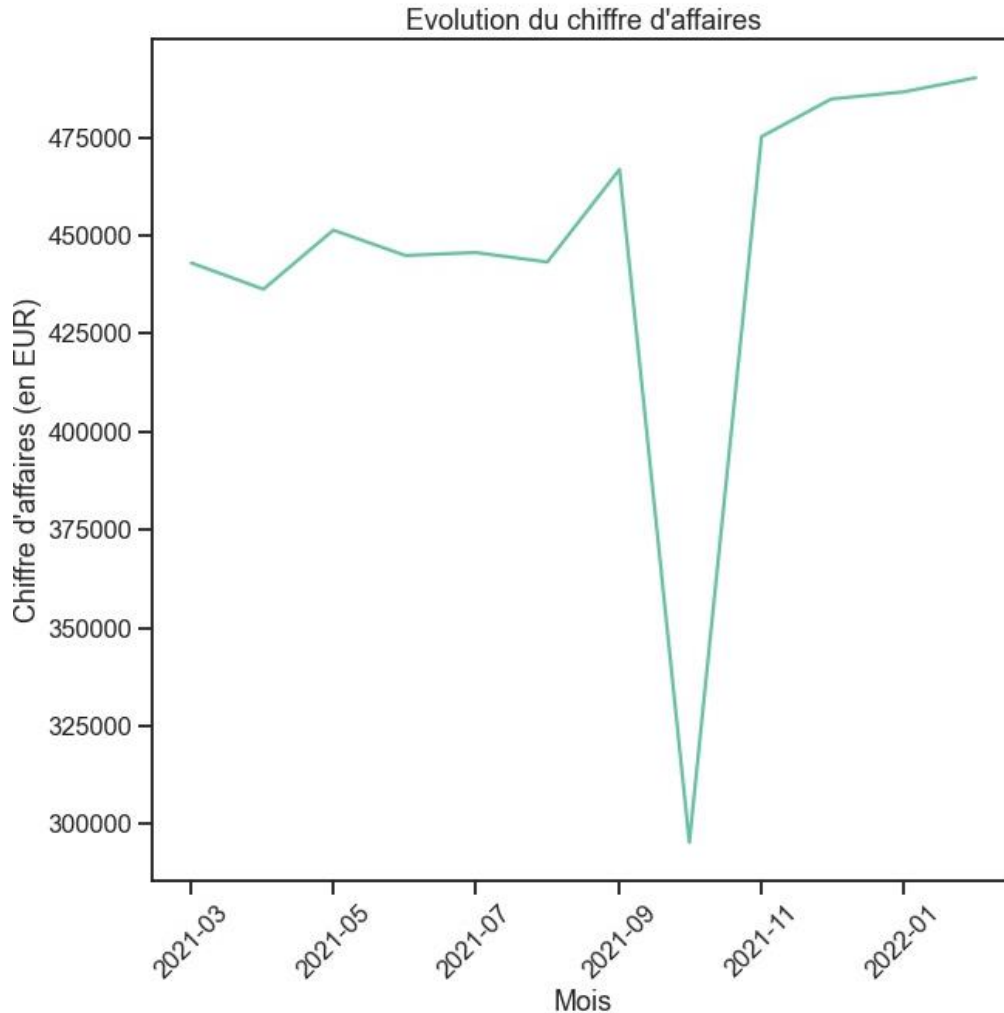


- Le nombre de vente est plutôt stable tout au long de l'année.
- Il y a cependant une baisse notable en octobre.
- Septembre (rentrée scolaire et littéraire) et décembre (fêtes de fin d'années) sont des moments importants.



2 – ANALYSE DES DONNÉES

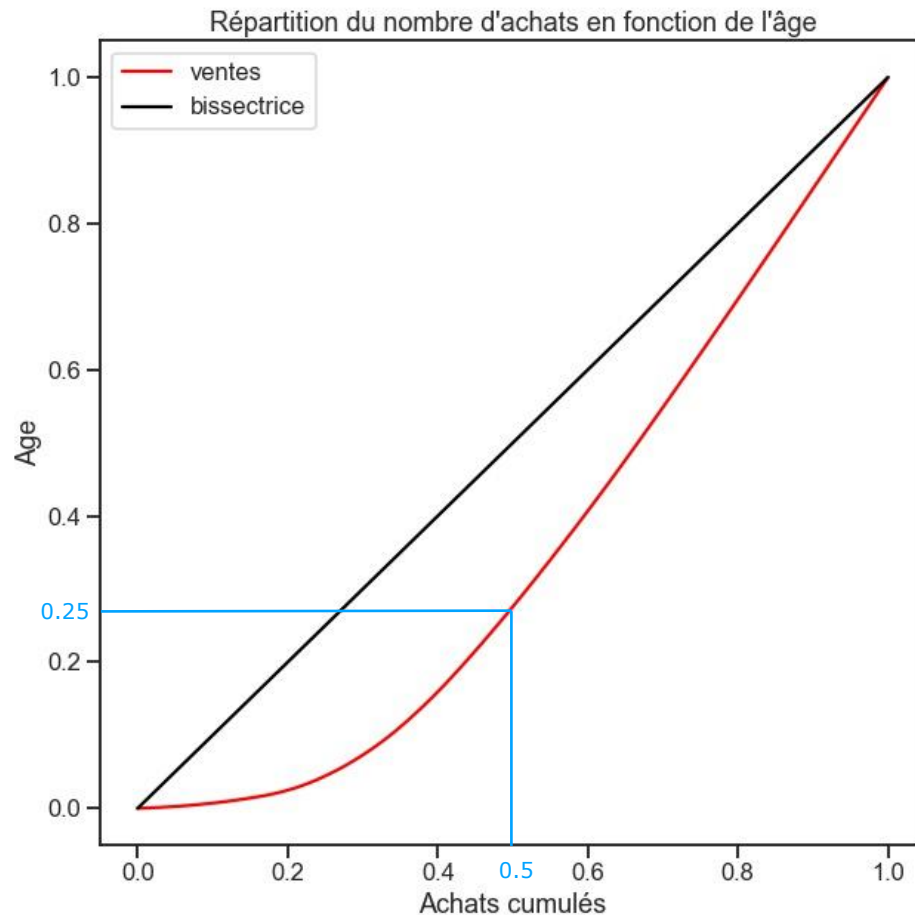
TRANSACTIONS : CHIFFRE D’AFFAIRES PAR MOIS



- Le chiffre d'affaires est en augmentation constante tout au long de l'année
- Il y a cependant une baisse notable en octobre, comme pour les ventes, que l'on peut expliquer par une perte de données.

2 – ANALYSE DES DONNÉES

TRANSACTIONS : CONCENTRATION DES ACHATS PAR AGE



- Le premier quartile des âges réalise 50% des achats.
- Avec un indice de Gini de 0,29, la répartition semble plutôt égalitaire.

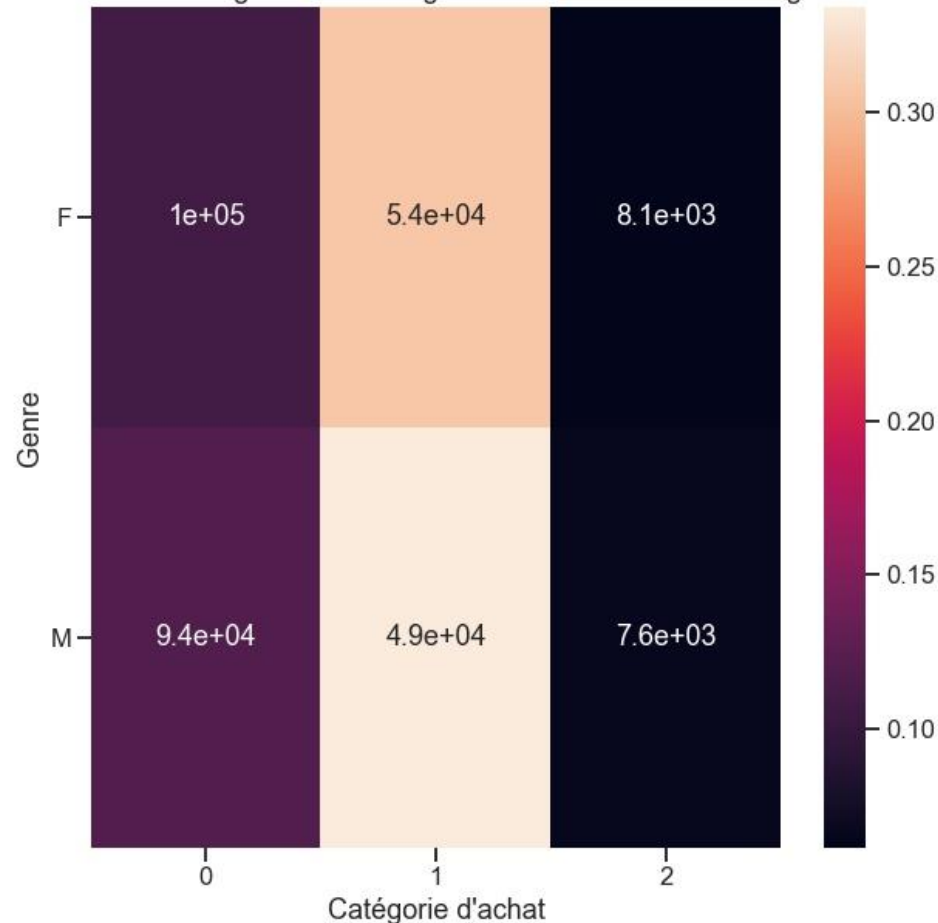


3 – ANALYSES BIVARIÉES

3 – ANALYSES BIVARIÉES

CORRÉLATION ENTRE GENRE ET CATÉGORIES D'ACHATS

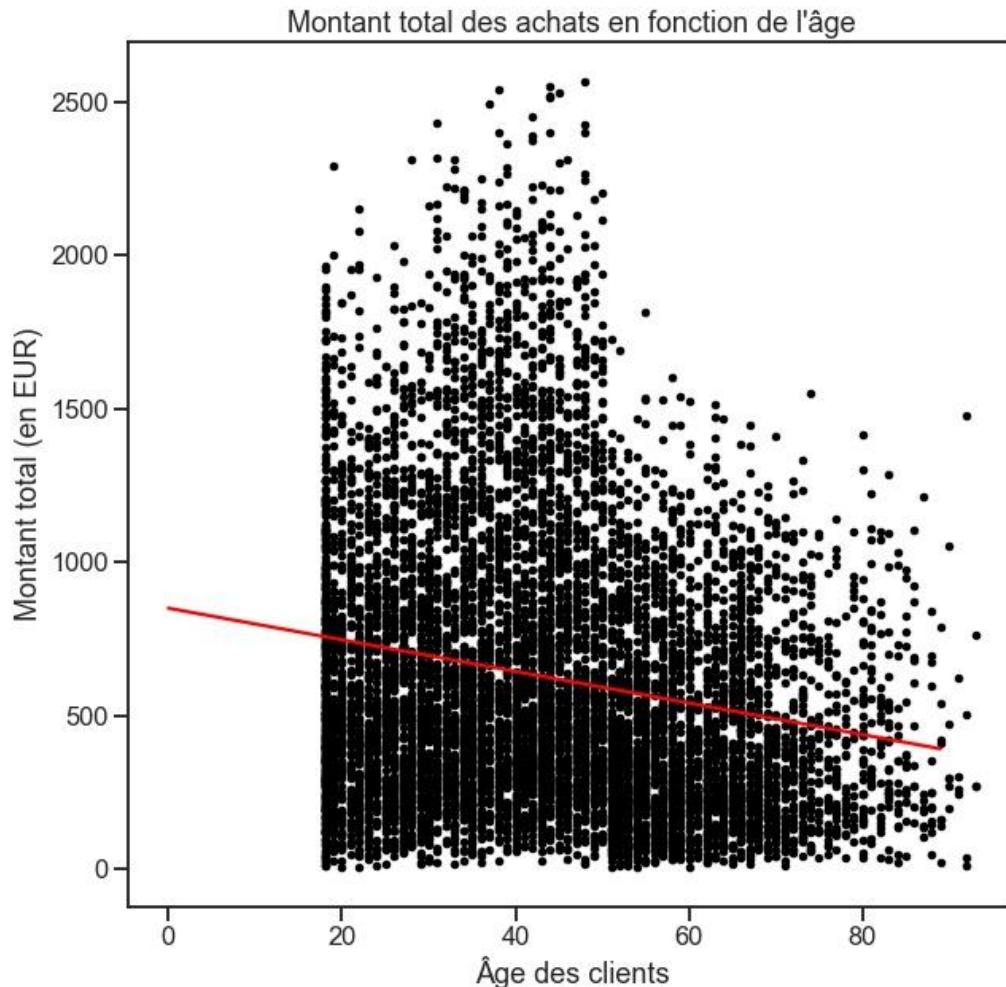
Tableau de contingence des catégories d'achat en fonction du genre



- Les femmes achètent majoritairement dans les catégories 1 et 2, et les hommes dans la catégorie 0.
- La relation n'est cependant pas significative statistiquement (selon le p-valeur).

3 – ANALYSES BIVARIÉES

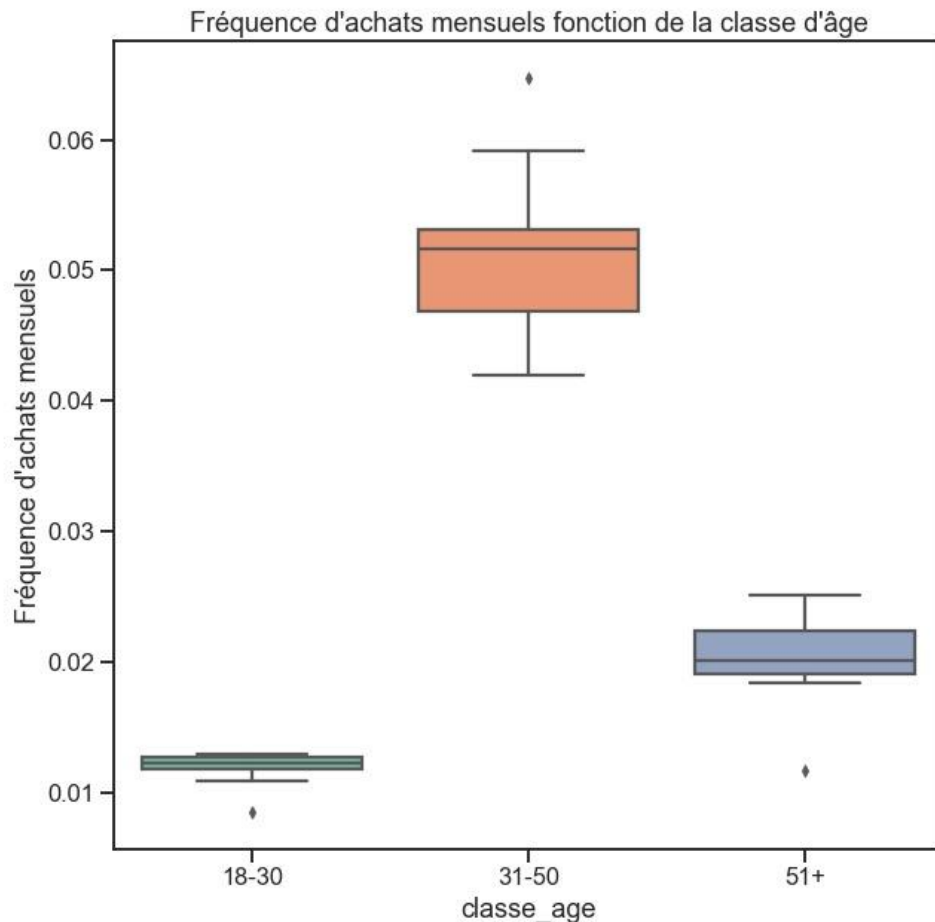
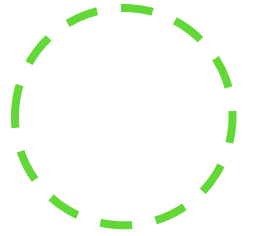
CORRÉLATION ENTRE ÂGE ET MONTANT TOTAL DES ACHATS



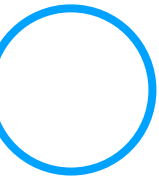
- Coefficient de corrélation de Pearson : -0.19
- Coefficient de détermination linéaire de Pearson (R^2) : 0,036
- Ces deux coefficients indiquent une absence de corrélation linéaire entre l'âge des clients et le montant total des achats.

3 – ANALYSES BIVARIÉES

CORRÉLATION ENTRE ÂGE ET FRÉQUENCE D'ACHAT

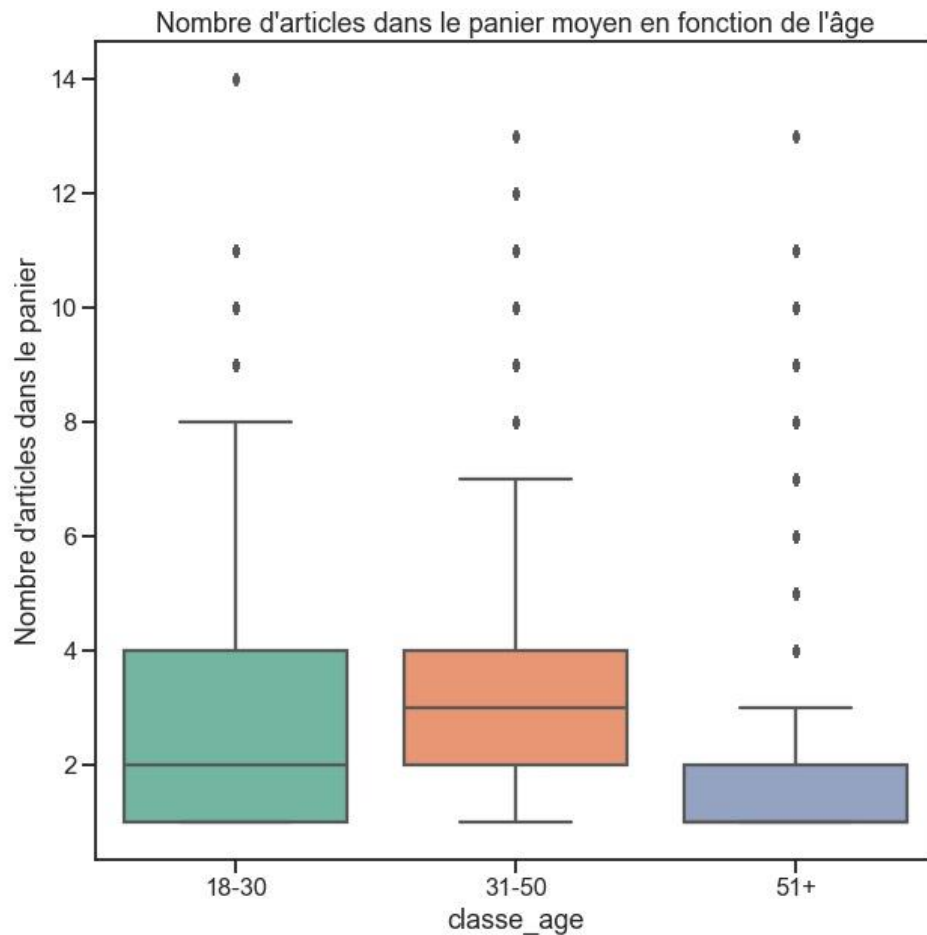
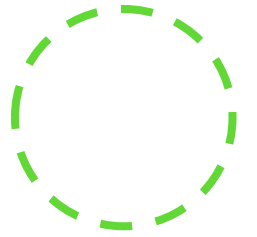


- $\text{Eta}^2 = 0,95$
- Cet eta^2 montre une importante corrélation entre la classe d'âge et la fréquence d'achat, et l'on constate sur le graphique que ce sont les 31-50 qui achètent le plus fréquemment.

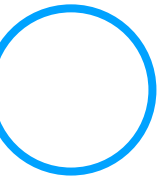


3 – ANALYSES BIVARIÉES

CORRÉLATION ENTRE ÂGE ET TAILLE DU PANIER MOYEN

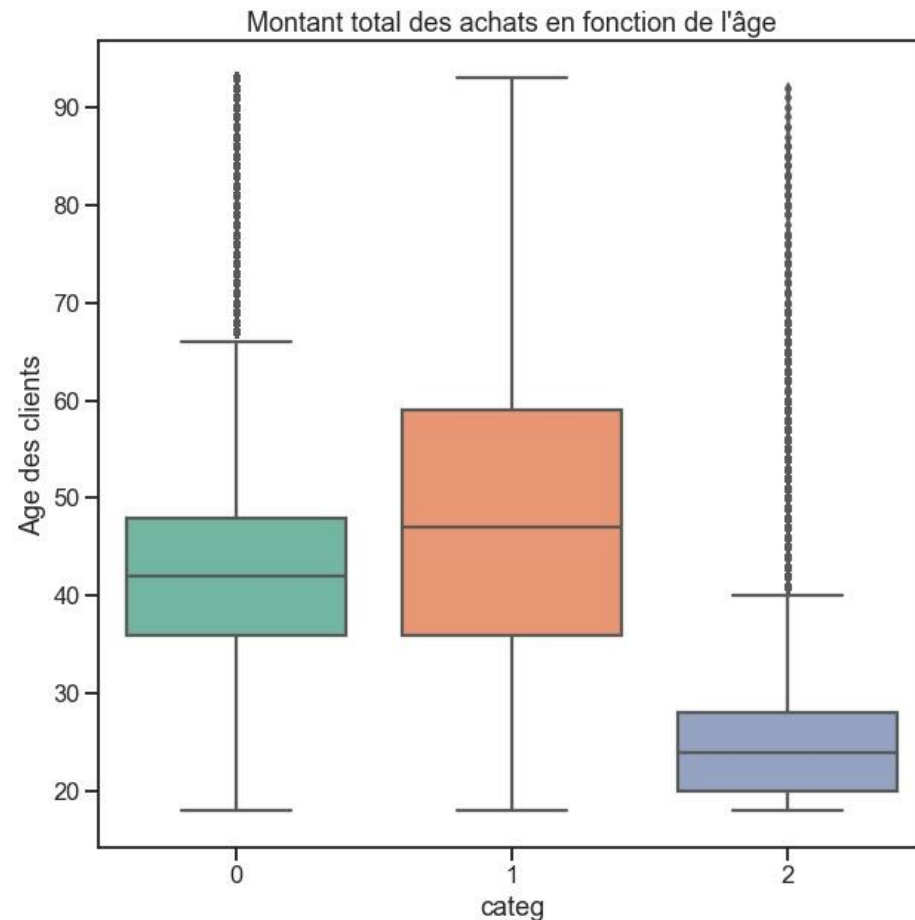


- $\text{Eta}^2 = 0,06$
- Le coefficient eta^2 montre une moyenne corrélation entre la classe d'âge et la taille du panier moyen.



3 – ANALYSES BIVARIÉES

CORRÉLATION ENTRE ÂGE ET CATÉGORIES DE PRODUITS ACHETÉS



- $\text{Eta}^2 = 0,11$
- Le coefficient eta^2 montre une forte corrélation entre l'âge des clients et la catégorie d'achat.
- Les clients les plus jeunes achètent principalement des produits de la catégorie 2, tandis que les autres deux autres catégories de produits sont achetées indistinctement par les clients.



4 – CONCLUSIONS ET RECOMMANDATIONS



- Deux aspects à corriger et prendre en compte dans les prochaines analyses :
 - Surreprésentation des clients de 18 ans.
 - Perte de données du mois d'octobre à récupérer.
 - Le client type de Rester Livres est un homme ou une femme de 35-50 ans achetant fréquemment des produits des catégories 1 et 2.
 - En revanche les 18-30 ans passent moins souvent à l'achat, même s'il s'agit principalement des produits plus chers de la catégorie 2. Il faut les inciter à acheter davantage dans les autres catégories et/ou plus régulièrement.
 - Les outliers pourraient être des entreprises, je recommande donc de séparer le B2B et B2C dans les prochaines analyses si c'est le cas.
- 