Analyse des ventes online

MAXIMEBCH- RESTER LIVRES

Sommaire

1 - NETTOYAGE DES DONNÉES

1 – NETTOYAGE DES DONNÉES

FICHIERS SOURCES

transactions.csv

	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
...	...	...	...	...
337011	1_671	2021-05-28 12:35:46.214839	s_40720	c_3454
337012	0_759	2021-06-19 00:19:23.917703	s_50568	c_6268
337013	0_1256	2021-03-16 17:31:59.442007	s_7219	c_4137
337014	2_227	2021-10-30 16:50:15.997750	s_112349	c_5
337015	0_1417	2021-06-26 14:38:19.732946	s_54117	c_6714

products.csv

	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
...	...	...	...
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

customers.csv

	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
...	...	...	...
8618	c_7920	m	1956
8619	c_7403	f	1970
8620	c_5119	m	1974
8621	c_5643	f	1968
8622	c_84	f	1982

1 – NETTOYAGE DES DONNÉES

VALEURS ABERRANTES – SESSIONS DE TEST « S_0 »

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

	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

```
transactions.loc[transactions['session_id'] == 's_0']
```

```
index_s_0 = transactions[transactions['session_id'] == 's_0'].index.values
transactions.drop(index_s_0, 0, inplace=True)
```

1 – NETTOYAGE DES DONNÉES

VALEURS ABERRANTES – PRODUIT TEST « T_0 »

```
products_test = products.sort_values('price', ascending=True)
products_test
```

	id_prod	price	categ
731	T_0	-1.00	0
2272	0_528	0.62	0
2355	0_202	0.62	0


```
products.drop(731, 0, inplace=True)
```

1 – NETTOYAGE DES DONNÉES

VALEURS ABERRANTES – CLIENTS DES SESSIONS DE TEST

```
ct_0 = customers.loc[customers['client_id'] == 'ct_0']
ct_1 = customers.loc[customers['client_id'] == 'ct_1']
print(ct_0)
print(ct_1)
```

```
     client_id sex  birth
2735      ct_0   f   2001
     client_id sex  birth
8494      ct_1   m   2001
```

```
customers.drop([2735, 8494], 0, inplace=True)
```

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	id_prod_prod	client_id_custom
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())
```

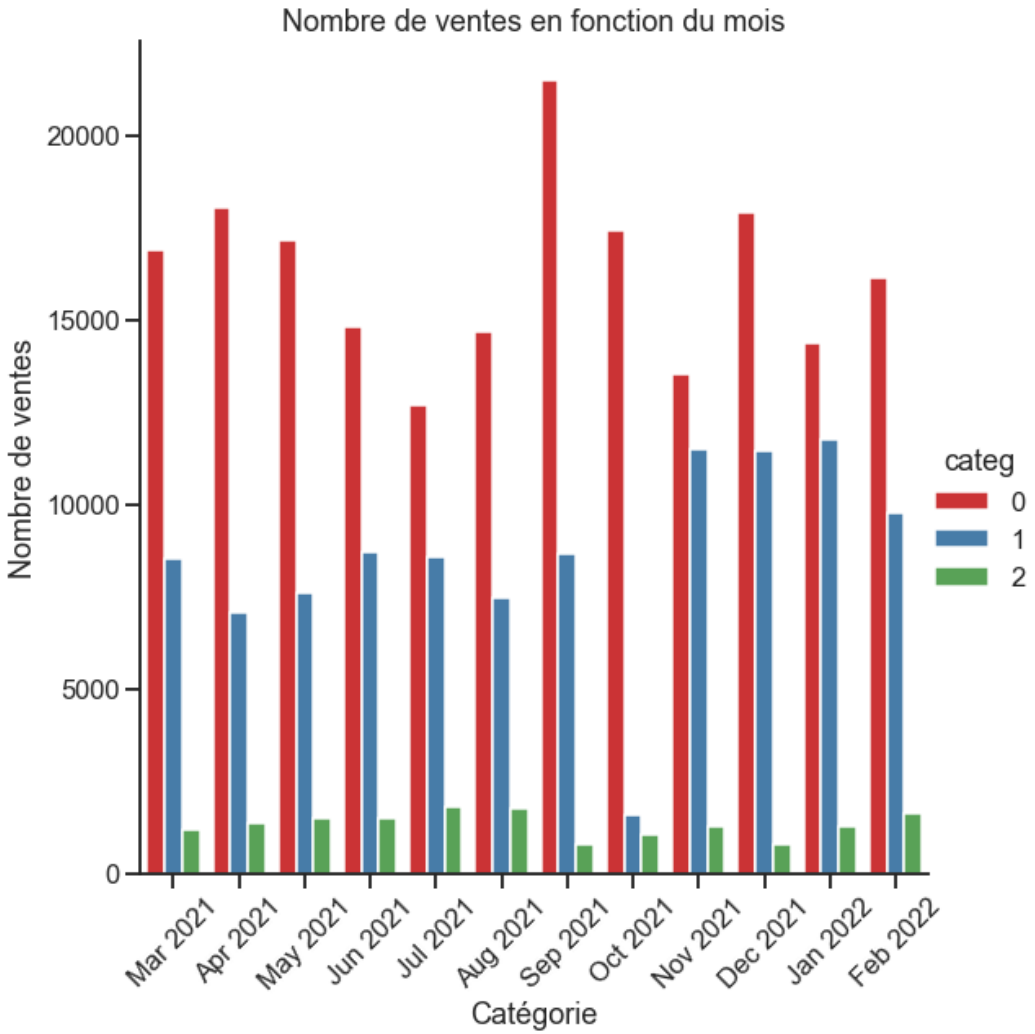
```
     id_prod   price categ
3282    2_23  115.99     2
3283   0_146   17.14     0
3284   0_802   11.22     0
3285   1_140   38.56     1
3286  0_1920   25.16     0
```

```
products.loc[3287] = {'id_prod' : '0_2245', 'price' : 11.72, 'categ'
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

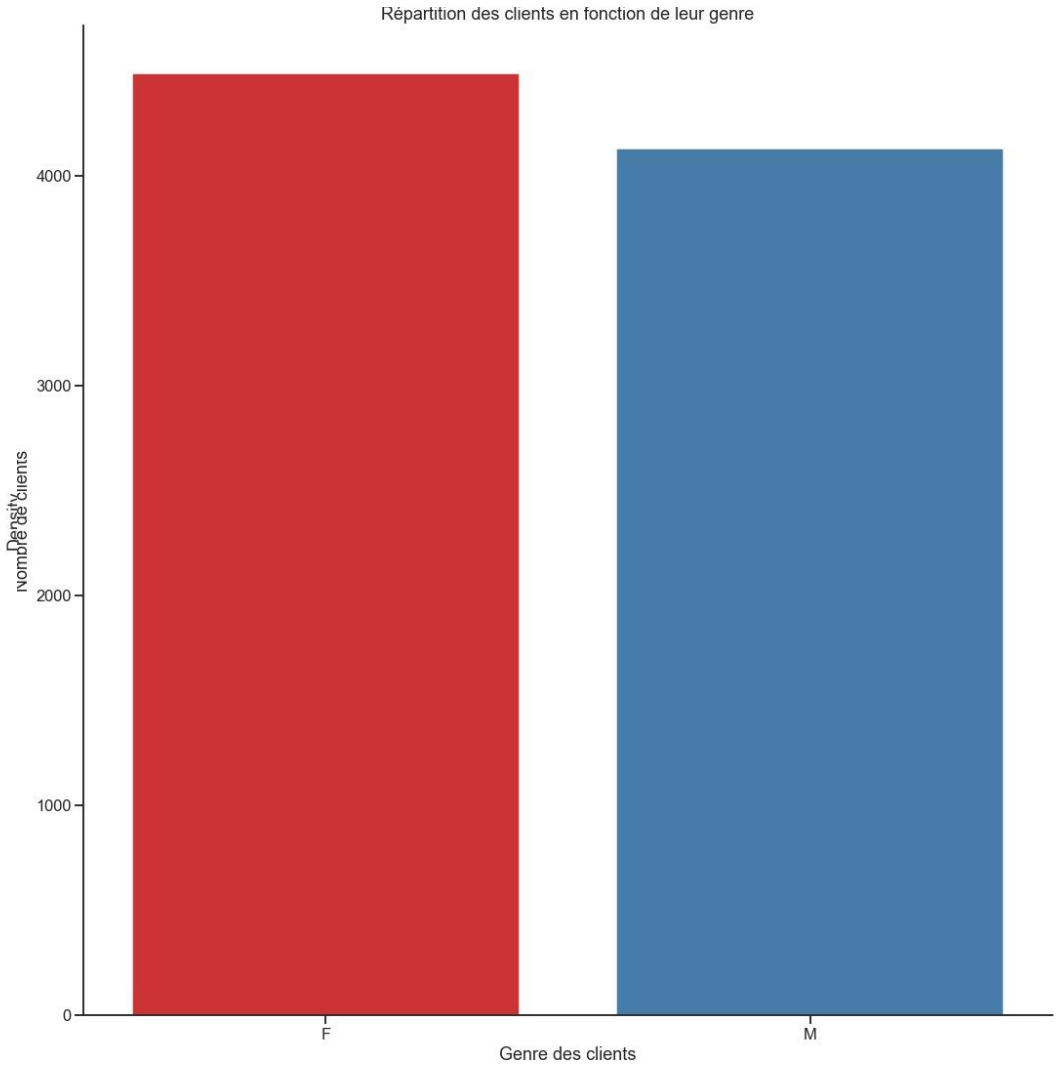


- Perte de données : il manque des données des transactions d'octobre dans la catégorie 1

2 – ANALYSE DES DONNÉES

2 – ANALYSE DES DONNÉES

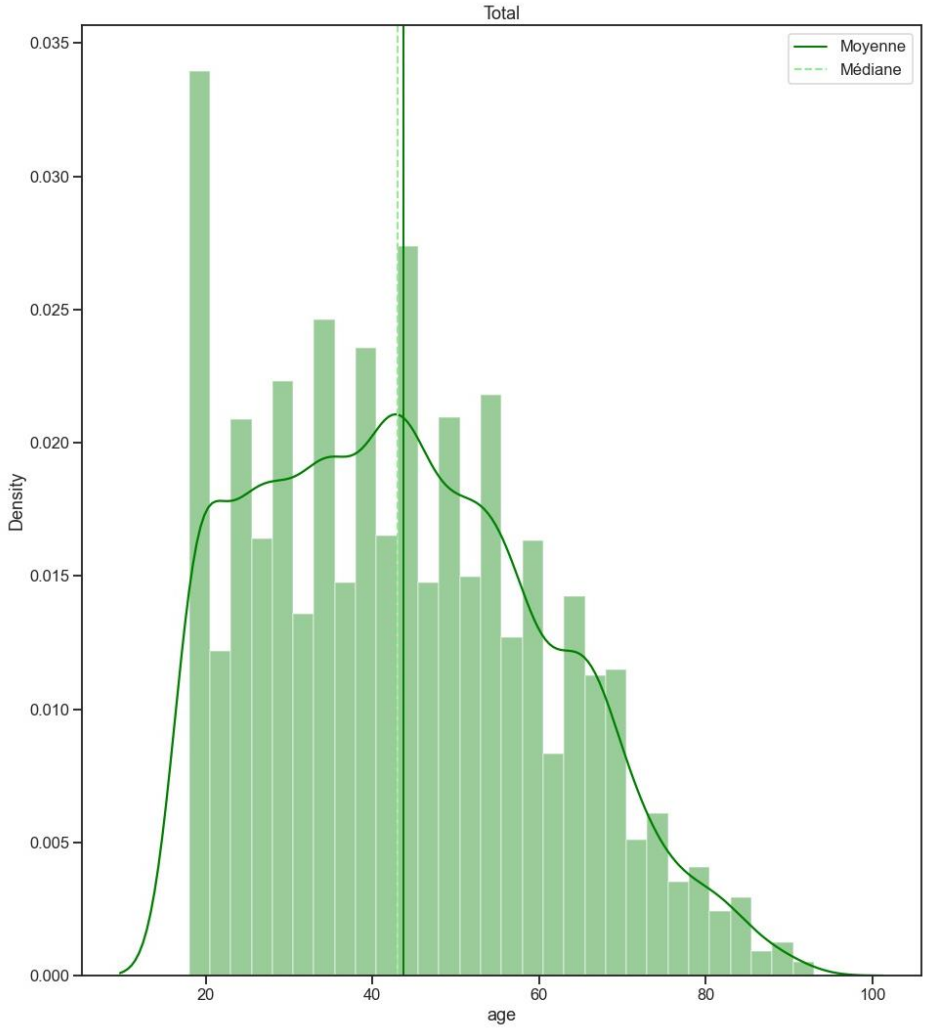
CLIENTS : DISTRIBUTION PAR GENRE



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

2 – ANALYSE DES DONNÉES

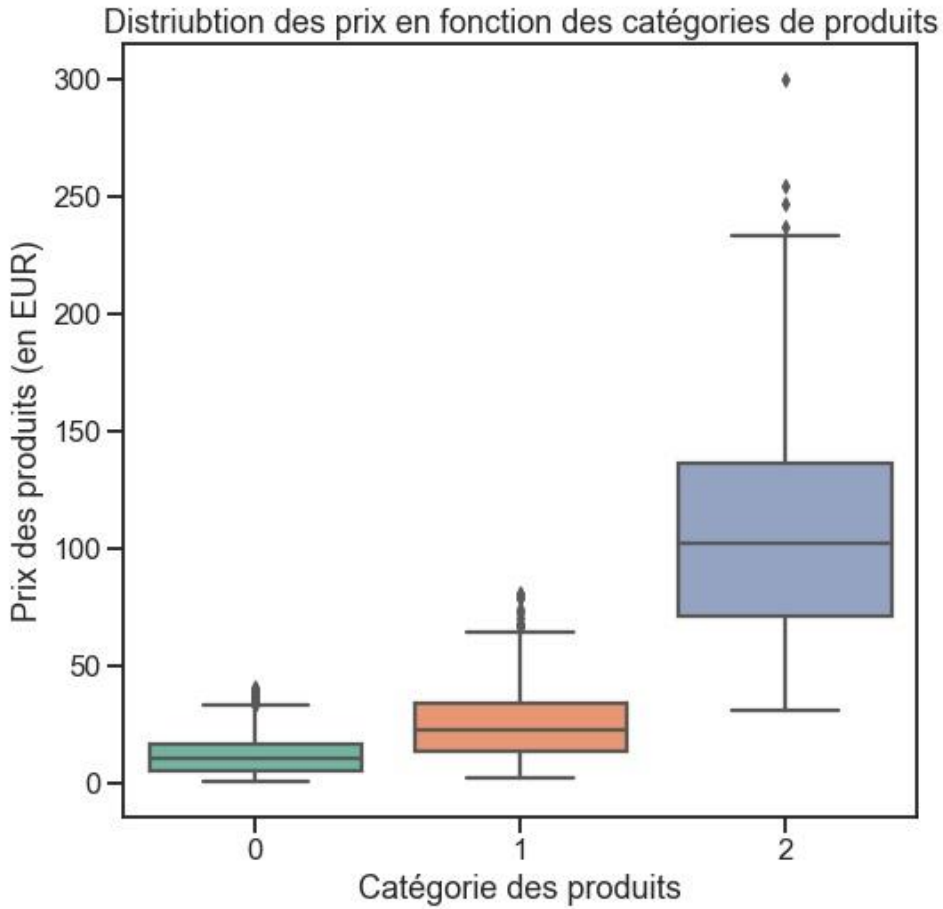
CLIENTS : DISTRIBUTION DES ÂGES



- 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

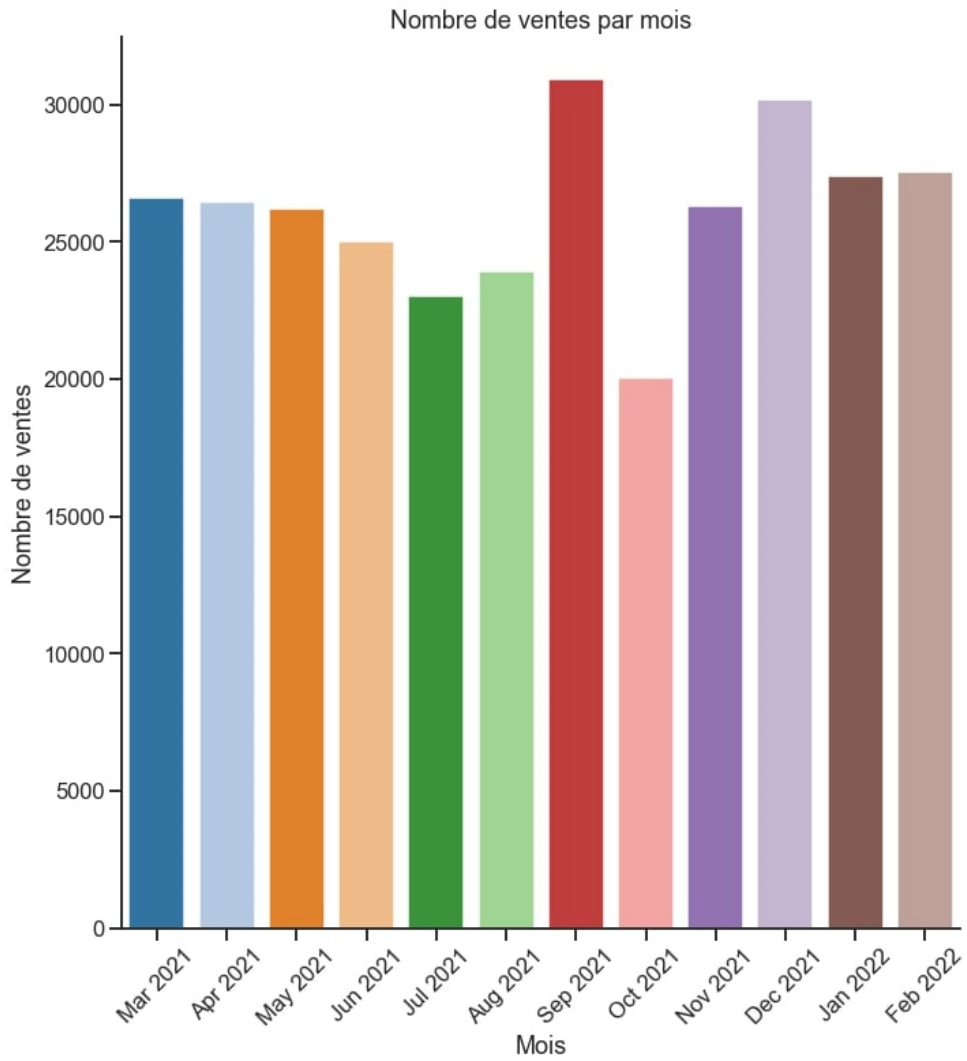
PRIX : DISTRIBUTION SELON LES CATÉGORIES DE PRODUIT



- 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

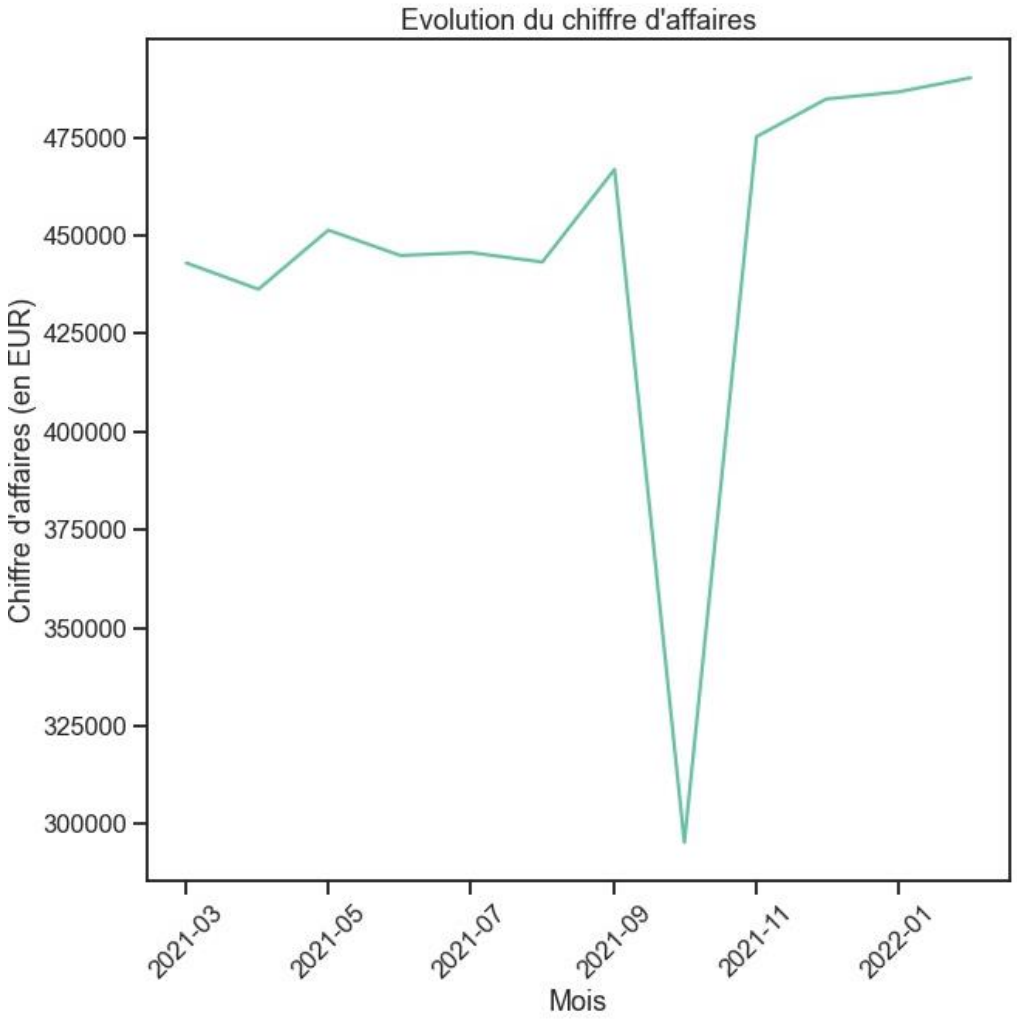
TRANSACTIONS : NOMBRE DE VENTES PAR MOIS



- 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

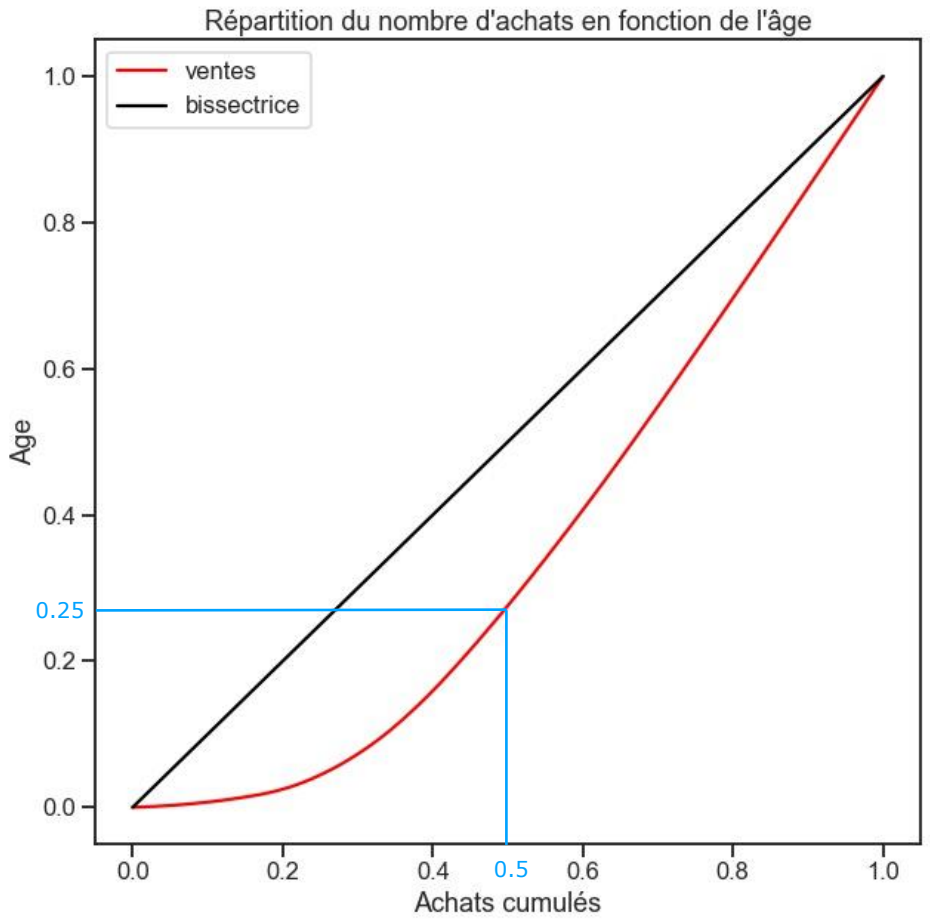
TRANSACTIONS : CHIFFRE D'AFFAIRES PAR MOIS



- 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

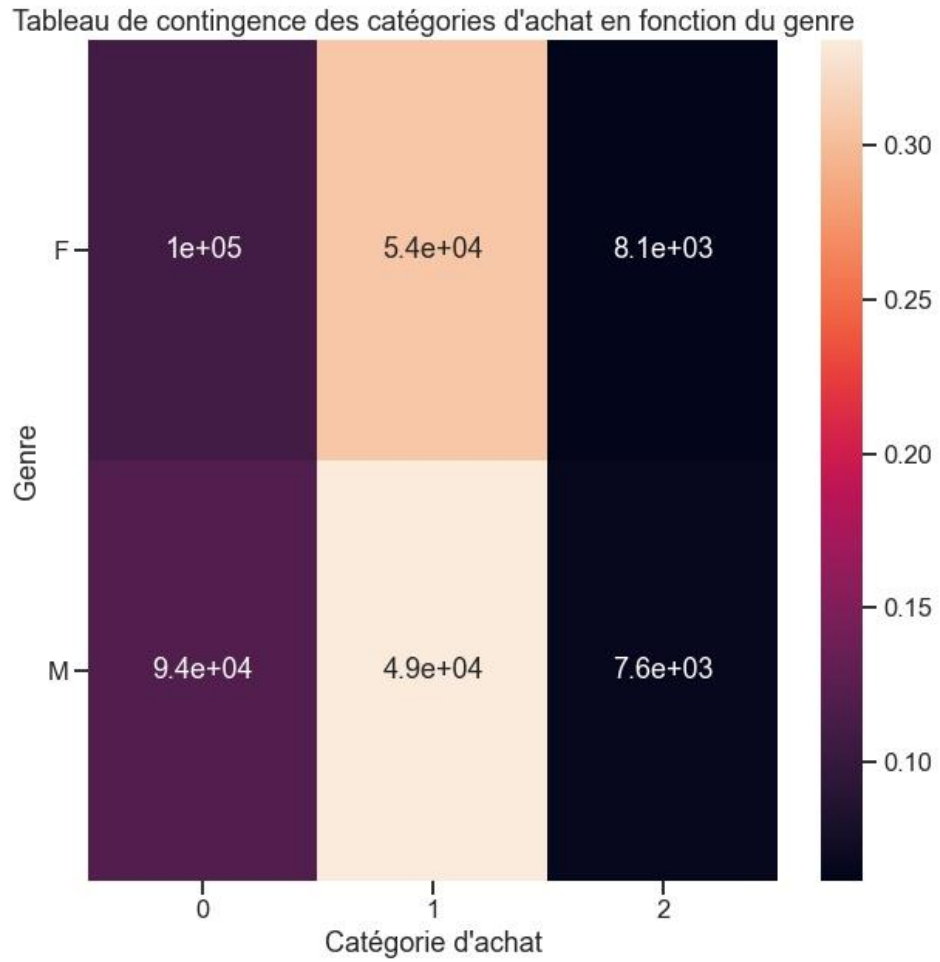


- Le premier quartile des âges réalise 50% des achats.
- Avec un indice de Gini de 0,29, la répartition semble plutôt égalitaire.

3 – ANALYSES BIVARIÉES

3 – ANALYSES BIVARIÉES

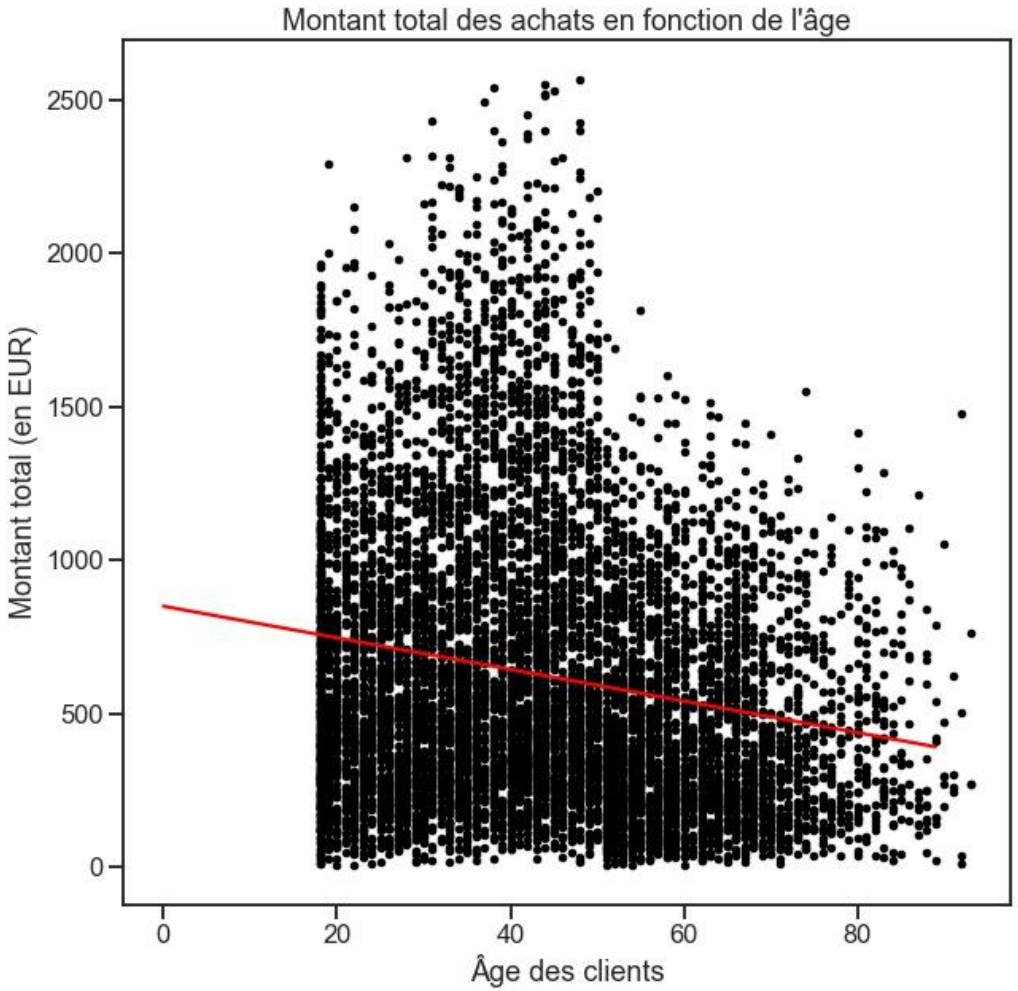
CORRÉLATION ENTRE GENRE ET CATÉGORIES D’ACHATS



- 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

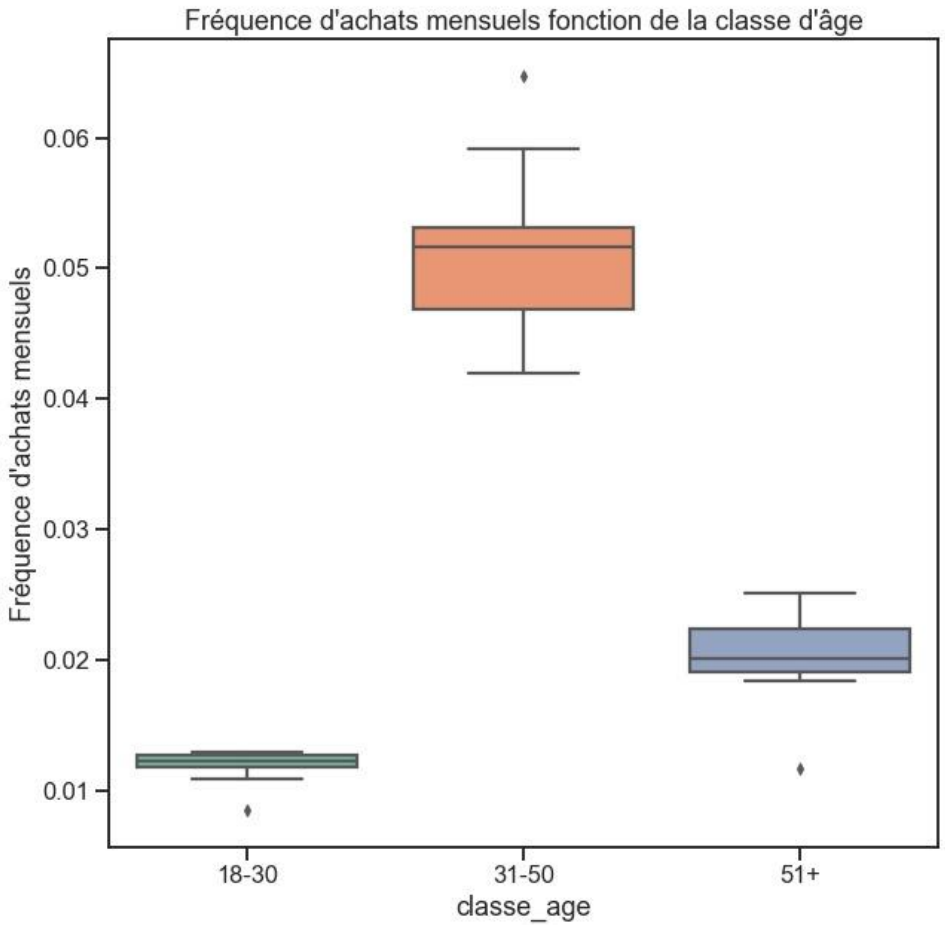
CORRÉLATION ENTRE ÂGE ET MONTANT TOTAL DES ACHATS



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

3 – ANALYSES BIVARIÉES

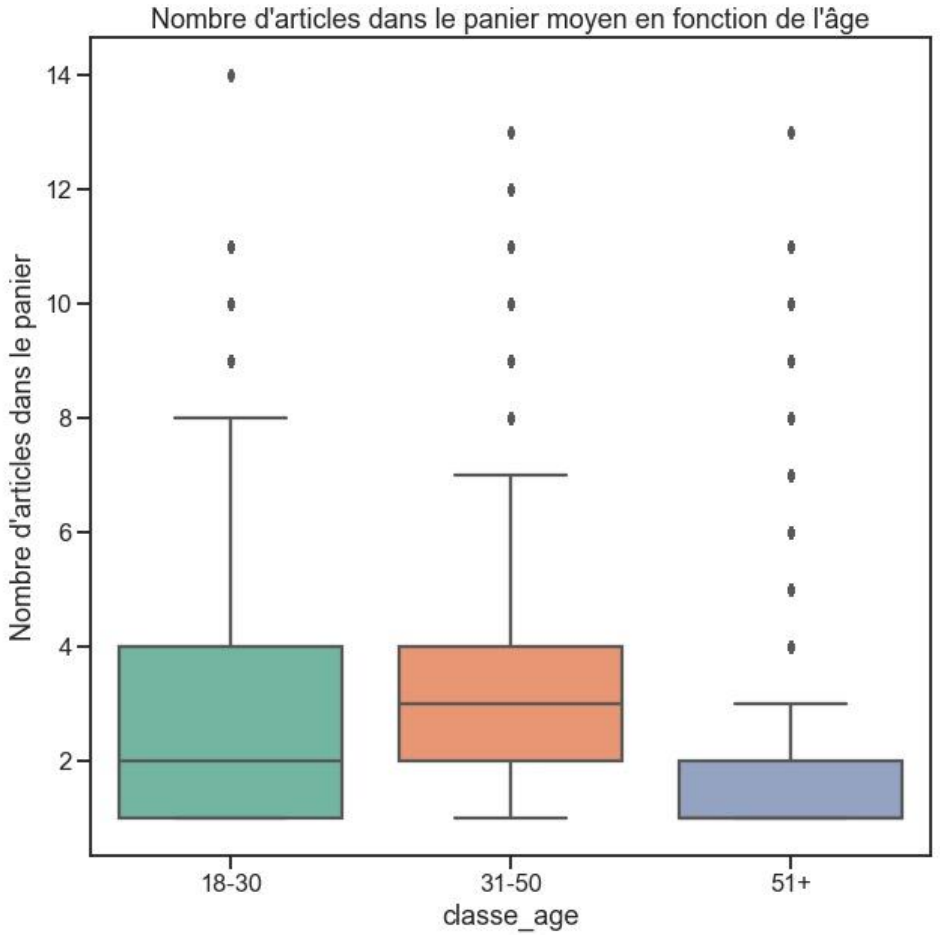
CORRÉLATION ENTRE ÂGE ET FRÉQUENCE D’ACHAT



- Eta² = 0,95
- Cet eta² 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

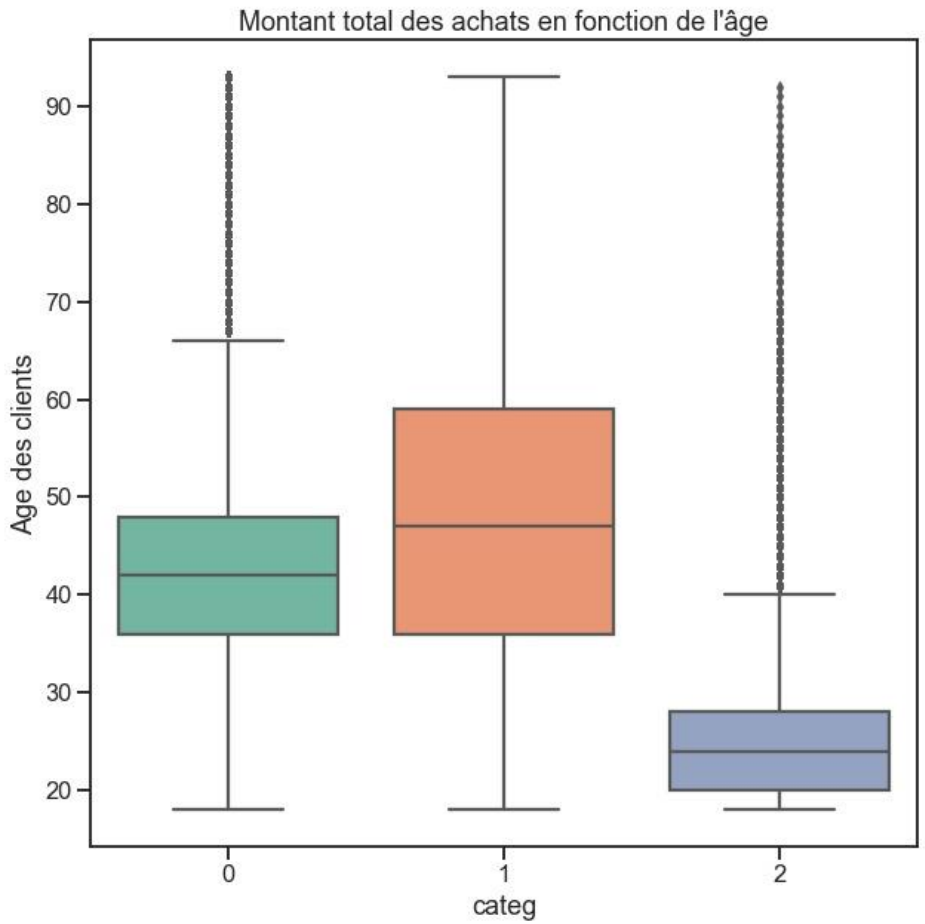
CORRÉLATION ENTRE ÂGE ET TAILLE DU PANIER MOYEN



- Eta² = 0,06
- Le coefficient eta² 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



- Eta² = 0,11
- Le coefficient eta² 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.