

Formation Data Analyst


Projet 4 : Analysez les ventes de votre entreprise

26/02/2021

Aïssa MOUACHA



SOMMAIRE

- ❑ Introduction
 - ❑ Data cleaning
 - ❑ Analyse des ventes
 - ❑ Corrélations
 - ❑ Conclusions
- 

- Grande chaîne de librairie « Rester livres »
- Développement activité par vente en ligne
- Analyser les données disponibles (transactions, clients, produits)



1. Vue d'ensemble

Téléchargement
*csv

Jointures (3 df)

Aperçu global

Jointure des 3 dataset

```
# on agrège toutes nos données
table = pd.merge(df1, df2, how="left", on="client_id")
table = pd.merge(table, df3, how="left", on="id_prod")
table.head()
```

```
# on jette un oeil sur l'aspect général de notre dataset
table.describe()
```

	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

nombre différent de variables

valeurs négatives et marginales

présence de 3 catégories

2. Traitement valeurs aberrantes

	id_prod		date	session_id	client_id	sex	birth	price	categ
1431	T_0	test_2021-03-01 02:30:02.237420		s_0	ct_1	m	2001	-1.0	0.0
2365	T_0	test_2021-03-01 02:30:02.237446		s_0	ct_1	m	2001	-1.0	0.0
2895	T_0	test_2021-03-01 02:30:02.237414		s_0	ct_1	m	2001	-1.0	0.0

Affichage de la table où valeur achat (= -1)
Correspondance avec libellé («test, T_, ct_»)

NB of price negative values : 200

NB of id_prod starting with T_ : 200

Matching between both families : True is unique value



Suppression des 200 lignes représentant
0.06% du dataframe global

Valeur achats ≥ 300 : nb = 8



Conservation de ces achats

3. Traitement valeurs manquantes

Identification du (des) Id_produit(s) concerné(s):

`['0_2245']`

Identification de la catégorie associée:

	id_prod	date	session_id	client_id	sex	birth	price	categ
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450	f	1977	4.990000	0.0
1	2_226	2022-02-03 01:55:53.276402	s_159142	c_277	f	2000	65.750000	2.0
2	1_374	2021-09-23 15:13:46.938559	s_94290	c_4270	f	1979	10.710000	1.0
3	0_2186	2021-10-17 03:27:18.783634	s_105936	c_4597	m	1963	4.200000	0.0
4	0_1351	2021-07-17 20:34:25.800563	s_63642	c_1242	f	1980	8.990000	0.0



Imputation (VAR = price) par moyenne

```
# Imputation par la moyenne [price]
average = table_alt2['price'].mean()
table_alt2['price'].fillna(round(average,1), inplace=True)
# Imputation par la valeur connue pour le produit identifié [categ]
cat0 = 0.0
table_alt2['categ'].fillna(cat0, inplace=True)
```

	id_prod	session_id	client_id	date	heure	sex	birth	categ	price	année	mois	age
6235	0_2245	s_49705	c_1533	2021-06-17	03:03:12	m	1972	0.0	17.2	2021	6	50
10802	0_2245	s_49323	c_7954	2021-06-16	05:53:01	m	1973	0.0	17.2	2021	6	49
14051	0_2245	s_124474	c_5120	2021-11-24	17:35:59	f	1975	0.0	17.2	2021	11	47
17486	0_2245	s_172304	c_4964	2022-02-28	18:08:49	f	1982	0.0	17.2	2022	2	40
21078	0_2245	s_3	c_580	2021-03-01	00:09:29	m	1988	0.0	17.2	2021	3	34

↓
Id_produit

↓ ↓
c0 moy

4. Format Date

Par partition , conversion et ajout de nouvelles colonnes



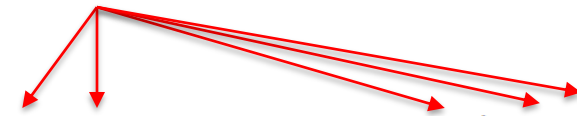
Formatage colonnes date/heure (split)

```
# on split [date] -> [date] + [heure]
table_alt1[['date', 'heure']] = table_alt1['date'].str.split(' ', n=1, expand=True)

# on convertit & reformate
table_alt1['date'] = pd.to_datetime(table_alt1['date']) # conversion D en datetime
table_alt1['heure'] = pd.to_timedelta(table_alt1['heure']) # conversion H en timedelta
table_alt1['heure'] = table_alt1['heure'].dt.floor('s') # format secondes tronqué

# on conserve seulement les colonnes suivantes
table_alt1 = table_alt1[['id_prod', 'session_id', 'client_id', 'date', 'heure', 'sex', 'birth', 'categ', 'price']]

# on crée des colonnes supplémentaires à partir des existantes
table_alt1['année'] = table_alt1['date'].dt.year
table_alt1['mois'] = table_alt1['date'].dt.month
table_alt1['age'] = 2022 - table_alt1['birth']
```



	id_prod	session_id	client_id	date	heure	sex	birth	categ	price	année	mois	age
0	0_1483	s_18746	c_4450	2021-04-10	18:37:28	f	1977	0.0	4.99	2021	4	45
1	2_226	s_159142	c_277	2022-02-03	01:55:53	f	2000	2.0	65.75	2022	2	22
2	1_374	s_94290	c_4270	2021-09-23	15:13:46	f	1979	1.0	10.71	2021	9	43
3	0_2186	s_105936	c_4597	2021-10-17	03:27:18	m	1963	0.0	4.20	2021	10	59
4	0_1351	s_63642	c_1242	2021-07-17	20:34:25	f	1980	0.0	8.99	2021	7	42

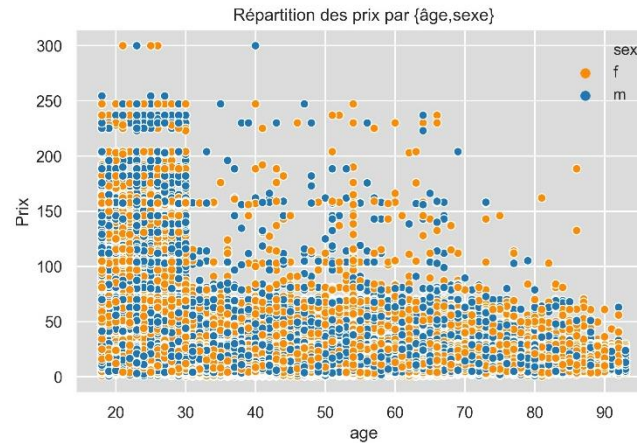
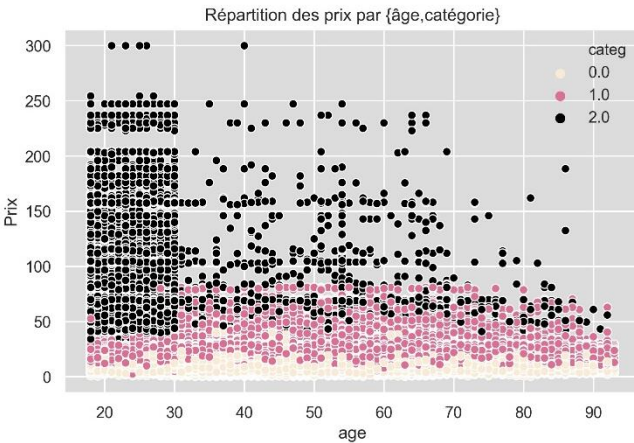
5. Résumé

	Cleaning
Size	337016 rows
Valeurs aberrantes	Suppression (200 rows)
Valeurs manquantes	Imputation par moyenne ou médiane (103 rows)
Format date	DD + HH
Size post-cleaning	336816 rows
Exports	« <i>table_imputations_moyenne</i> » ou « <i>table_imputations_mediane</i> »

Le double exercice d'imputation (par moy/med) conduit à des résultats extrêmement proches (rappel 0,03% du dataset global).

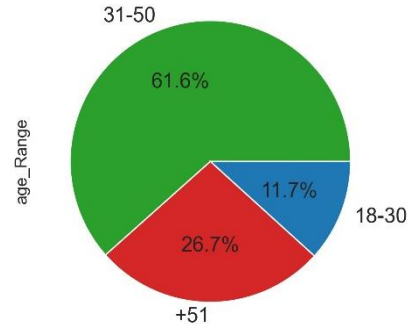
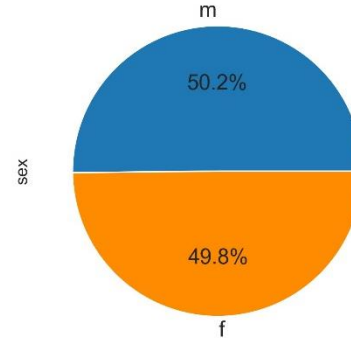
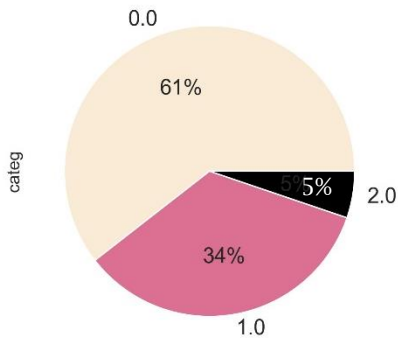
On poursuivra donc l'analyse avec la version avec imputation par la moyenne.

1. Présentation des graphs



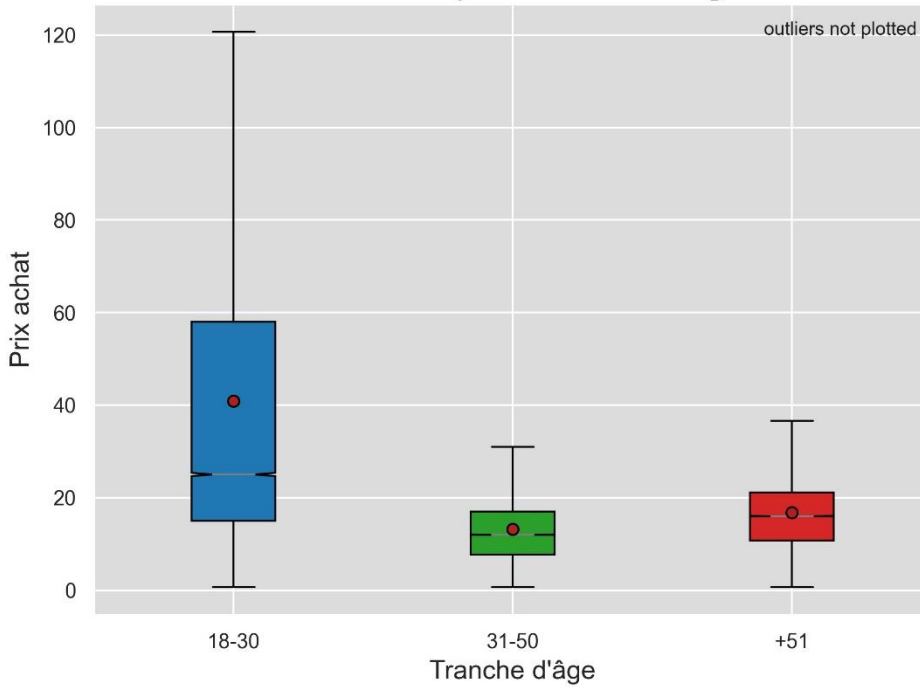
Etat des lieux du dataset :

- Produit c2
 - dispersion >
 - valeur achat >>
 - ~5% des ventes
 - plébiscité par les 18-30 ans
- Produit c0
 - ~2/3 des ventes
- Profil client (genre)
 - équi-répartition (h/f)
- Profil client (âge)
 - ~2/3 de « mid-age »
 - ~1/4 de « seniors »
 - ~1/10 de « jeunes »

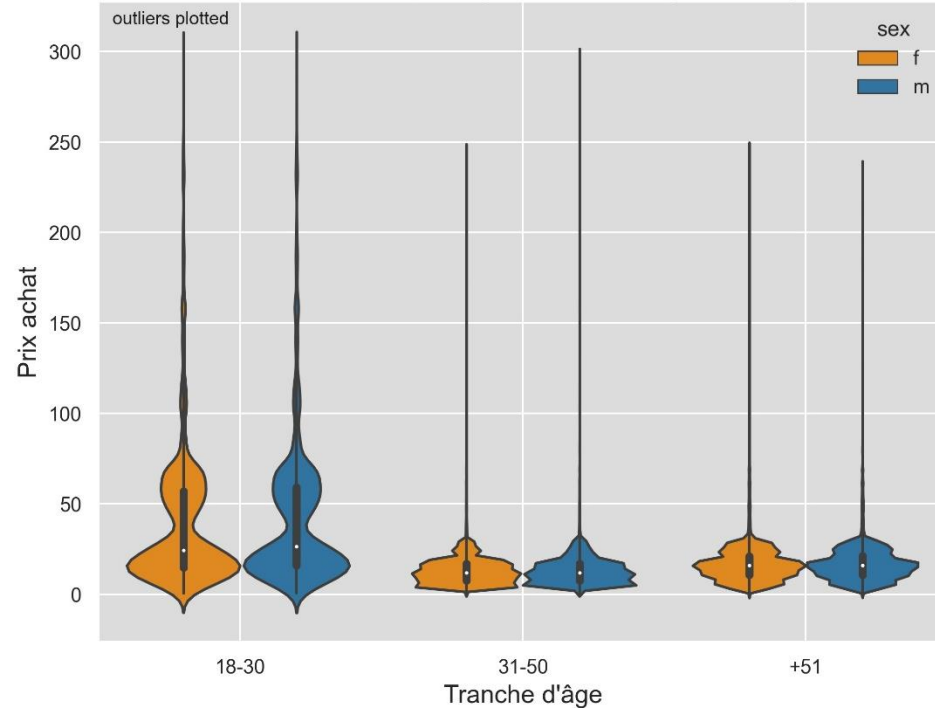


1. Présentation des graphs

Prix achat par tranche d'âge



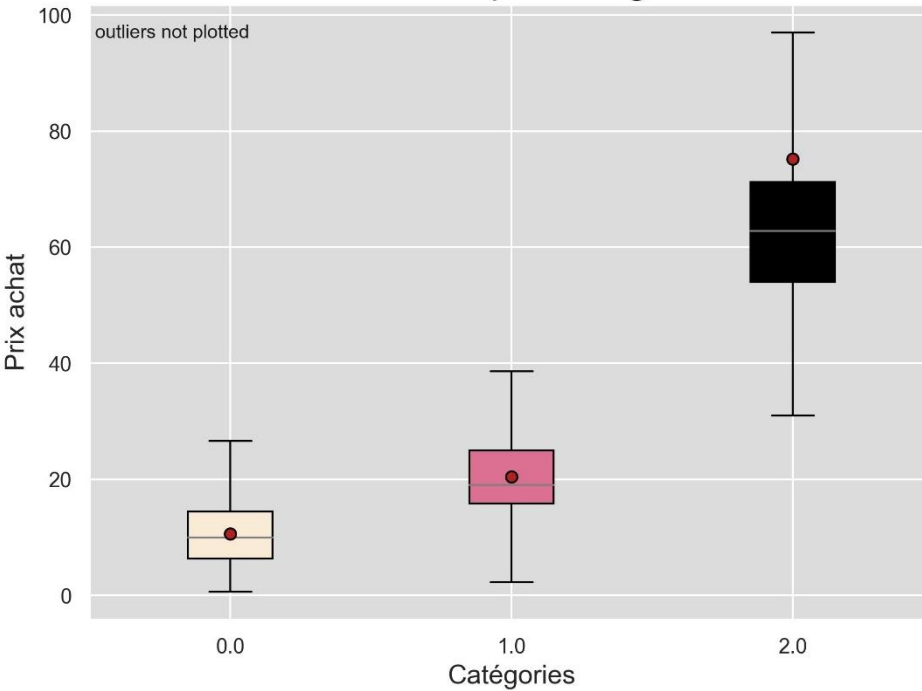
Prix achat par {tranche d'âge, sexe}



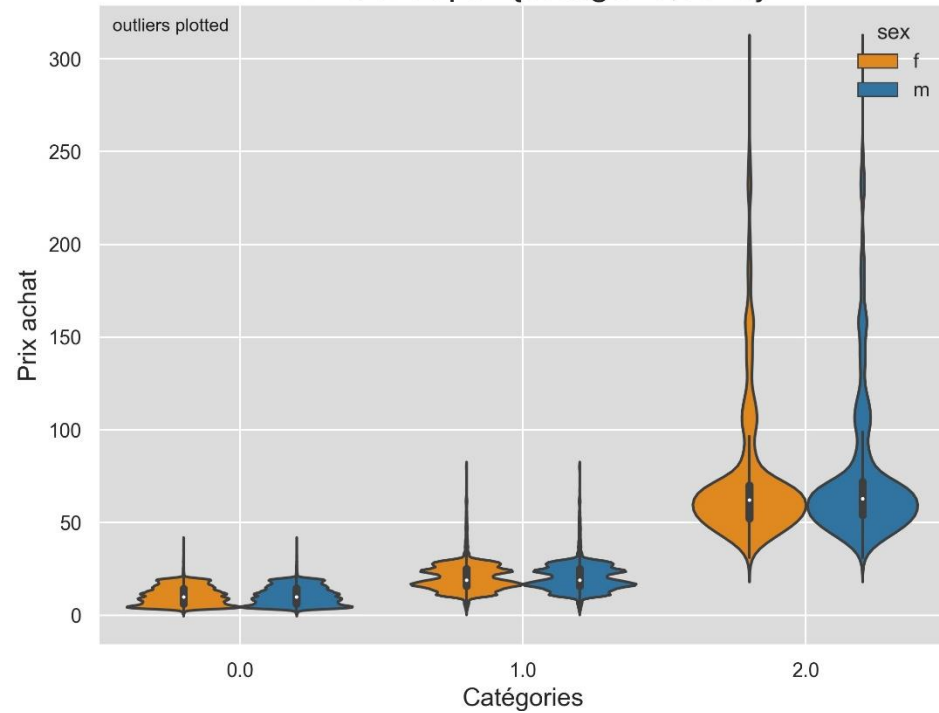
- Valeur achat moyen la plus élevée et dispersée chez les « jeunes »
- Confirmation aucune distinction h/f

1. Présentation des graphs

Prix achat par catégorie

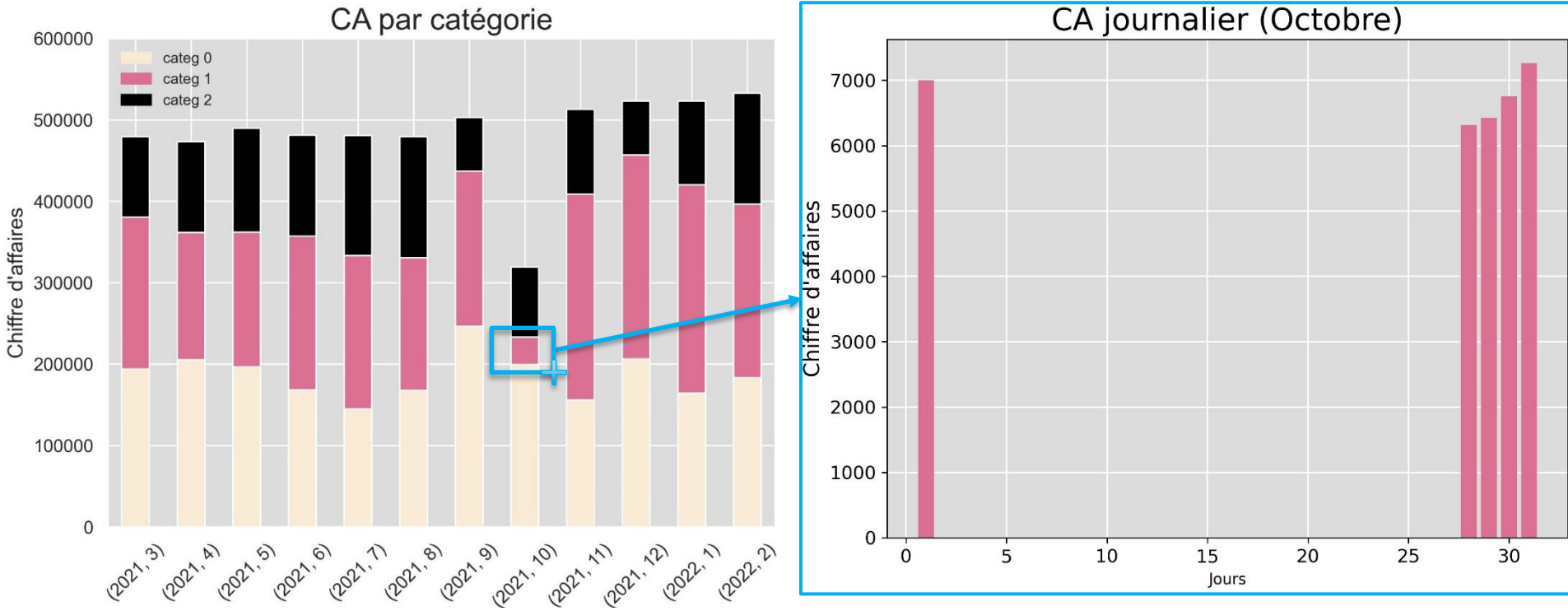


Prix achat par {catégorie,sexe}



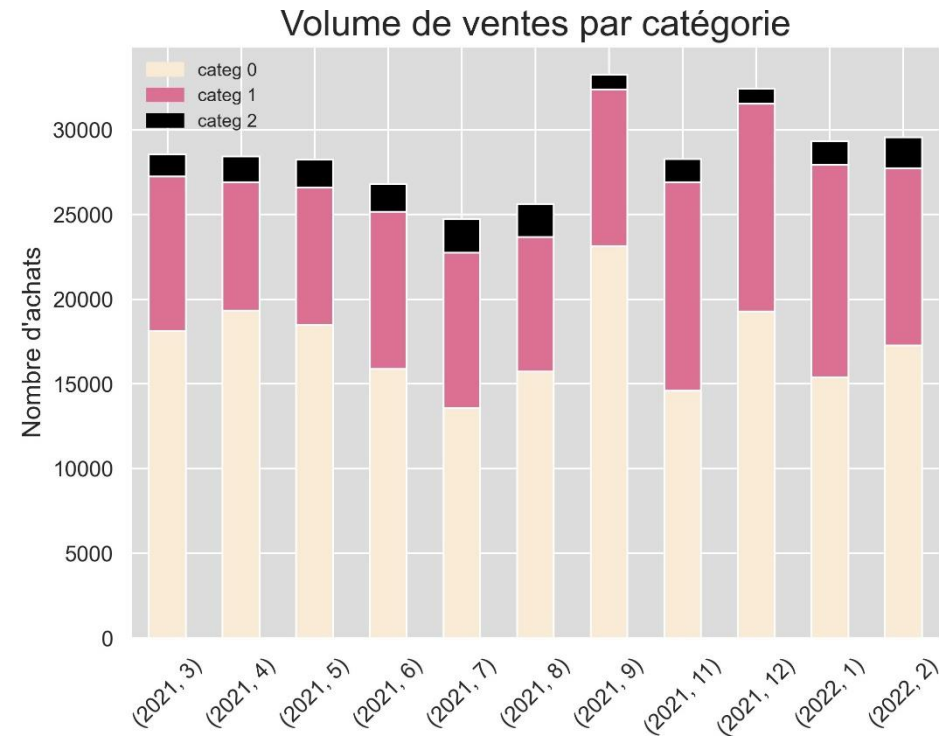
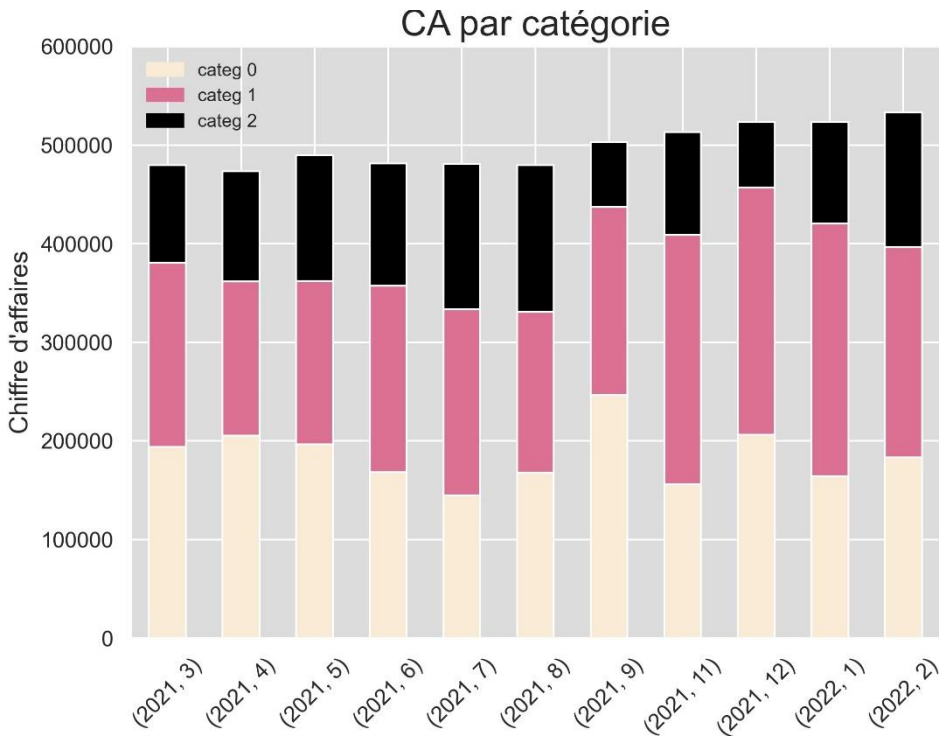
- Valeur achat moyen la plus élevée pour les produits de catégorie 2
- Confirmation aucune distinction h/f

1. Présentation des graphes



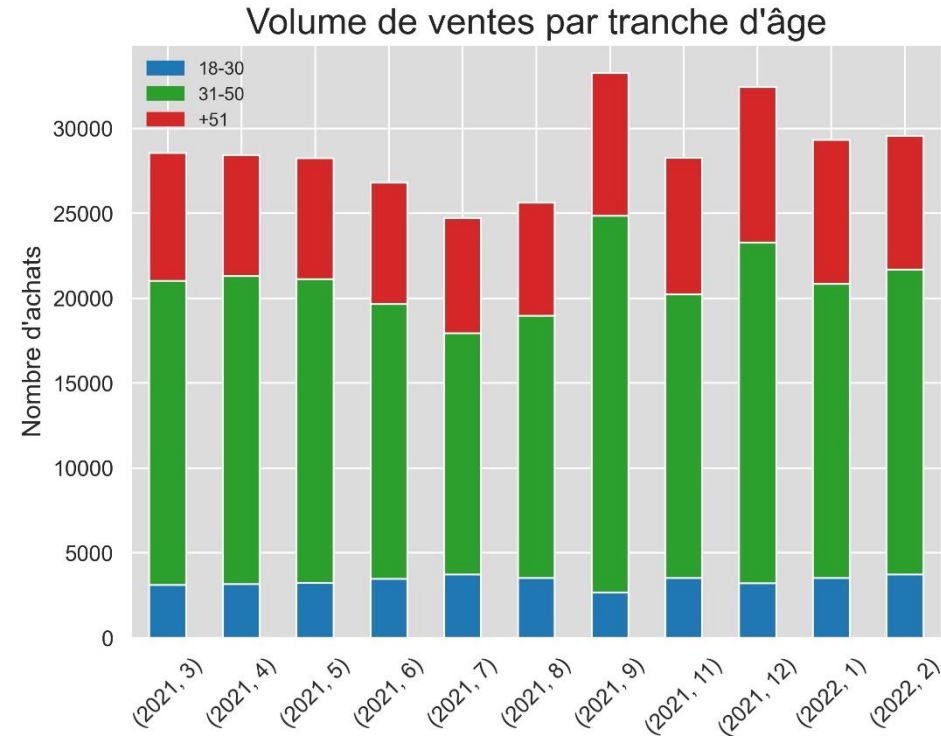
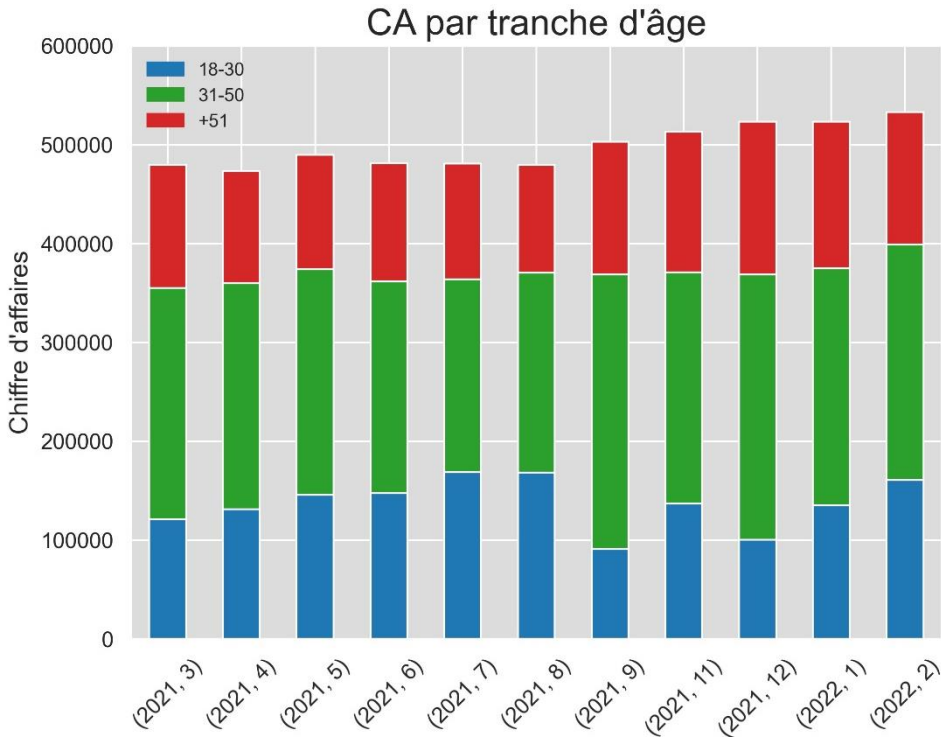
- Ventes du mois d'octobre ont subi une brutale chute / incident dans les produits de catégorie 1
- Pour ne pas biaiser notre analyse, les données du mois d'octobre ne seront pas prises en compte.

1. Présentation des graphes



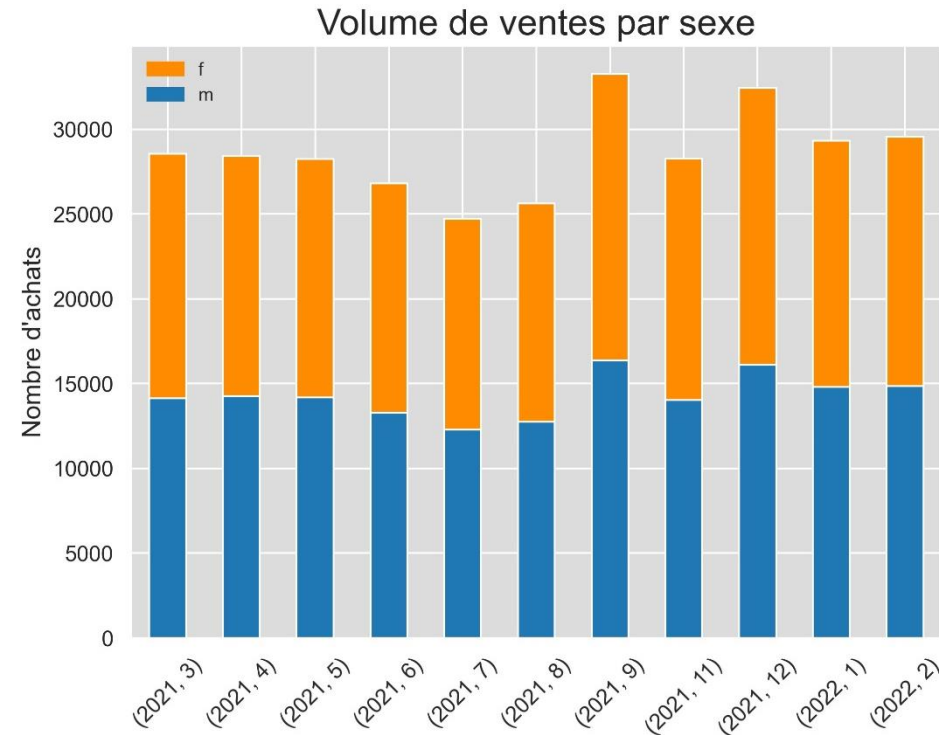
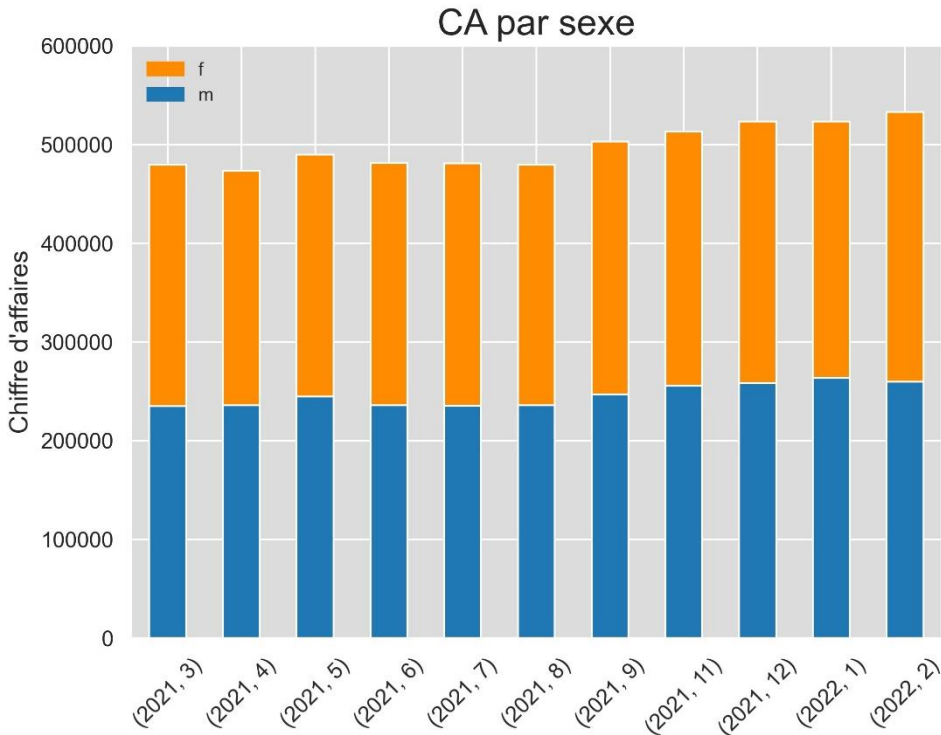
- CA stable sur les 6 premiers mois, puis augmentation
- Impact visible des achats produits c2 (peu en nombre mais élevés en valeur)

1. Présentation des graphes



- Croissance du CA 2d semestre visible sur tranches « seniors » puis « jeunes »
- CA des 18-30 ans important car moins de clients mais portés sur les produits c2 (+ chers)

1. Présentation des graphes



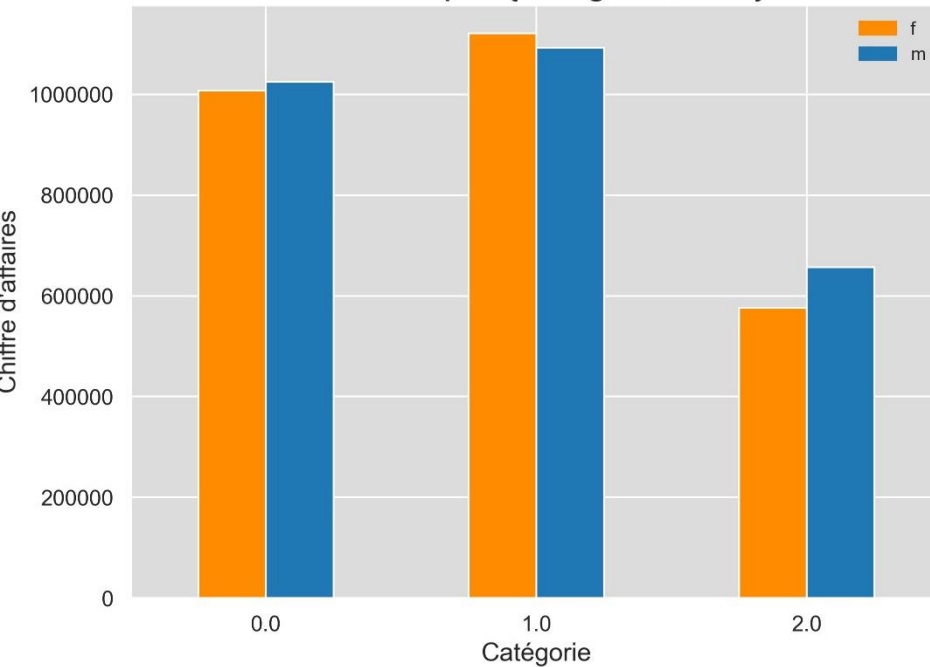
- H/F contribuent de façon équitable au CA
- Profil des clients (pour les mois 9 et 12)



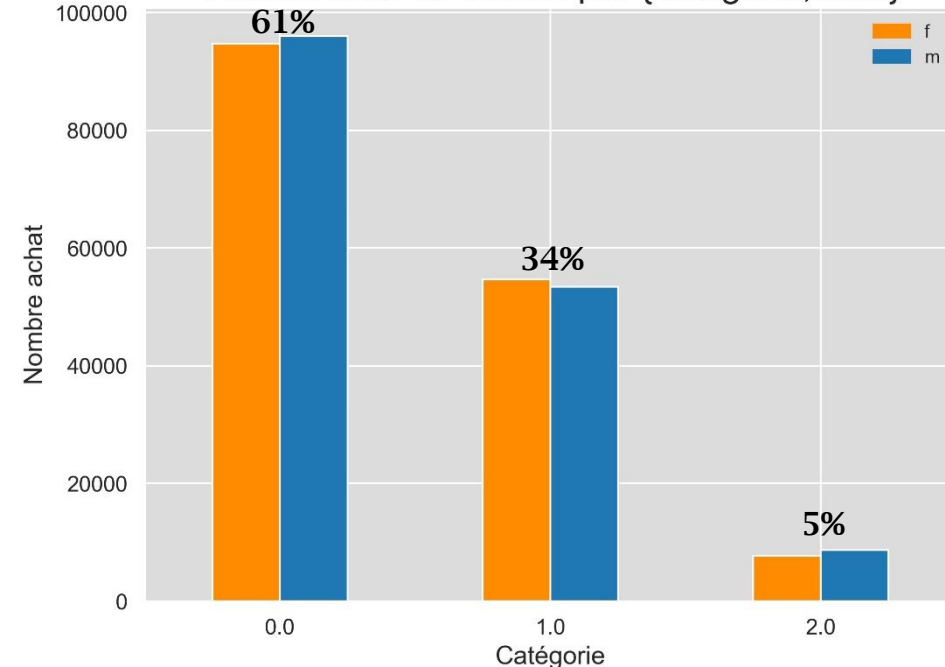
+ 31-50 ans + Produits c0

1. Présentation des graphs

CA total par {catégorie,sexe}

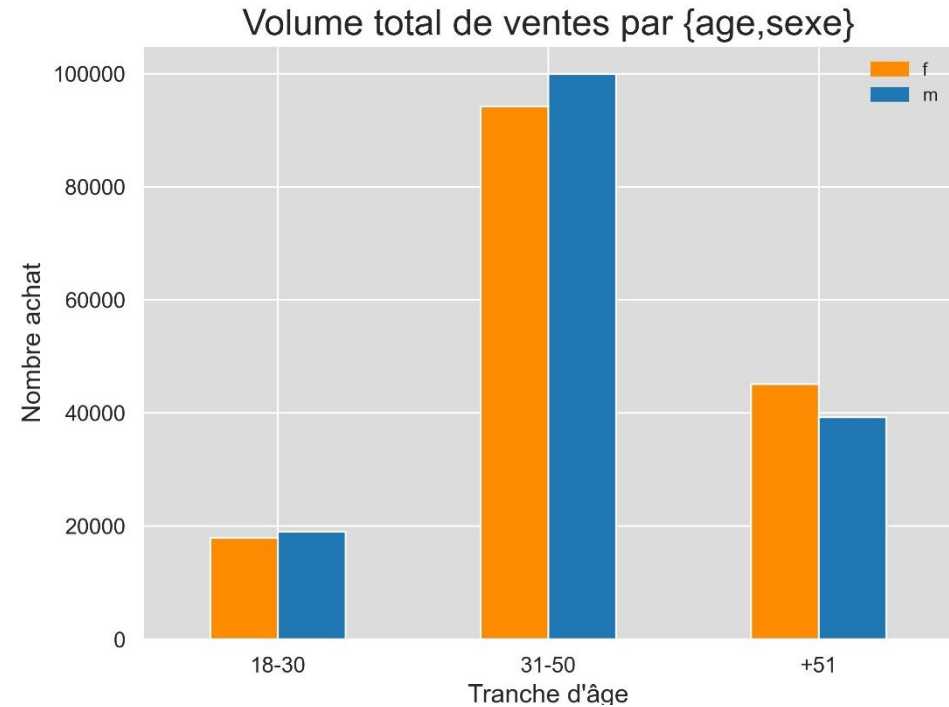
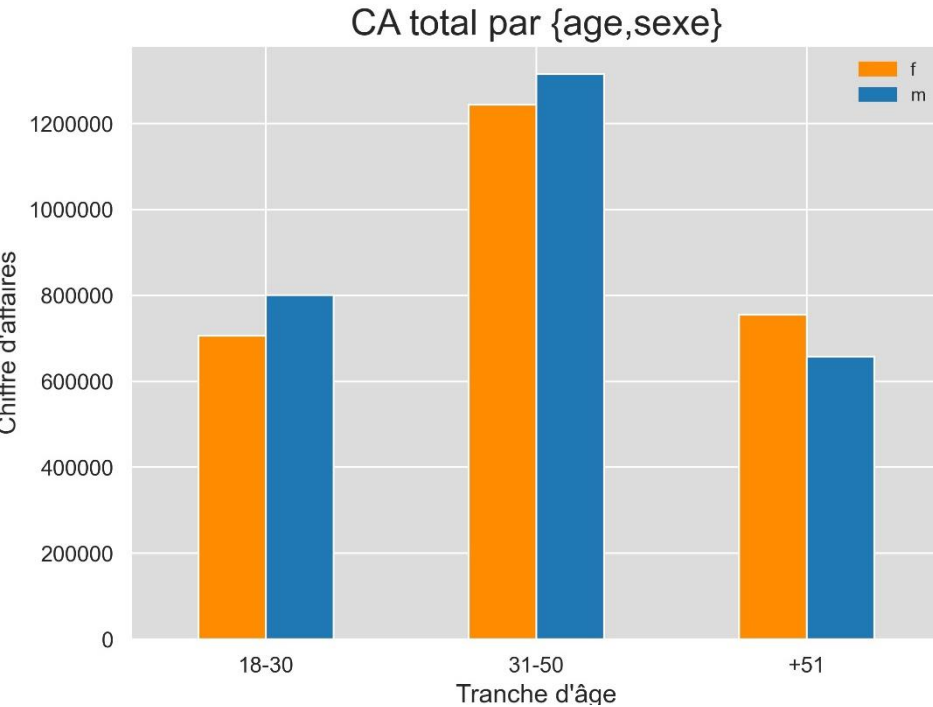


Volume total de ventes par {catégorie,sexe}



- Rapports respectifs (volume/CA):
 - c0 : ~1 pour 10
 - c1 : ~1 pour 20
 - c2 : ~1 pour 60

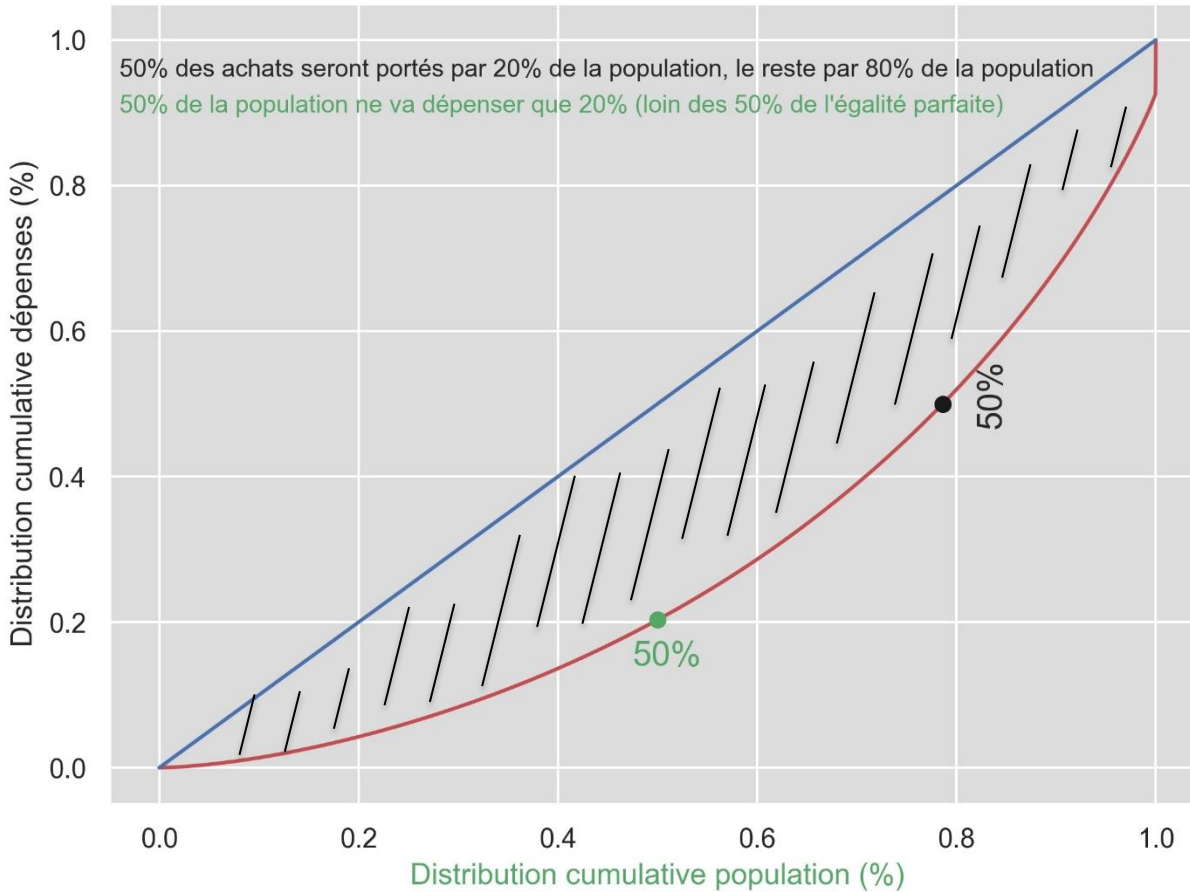
1. Présentation des graphs



- Socle du CA = clientèle « mid-âge »
- Importance de la clientèle « jeune »

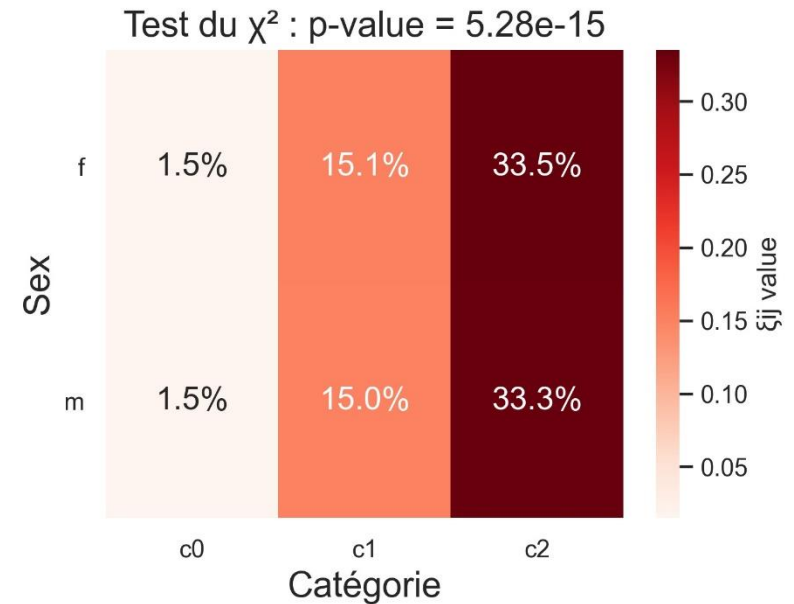
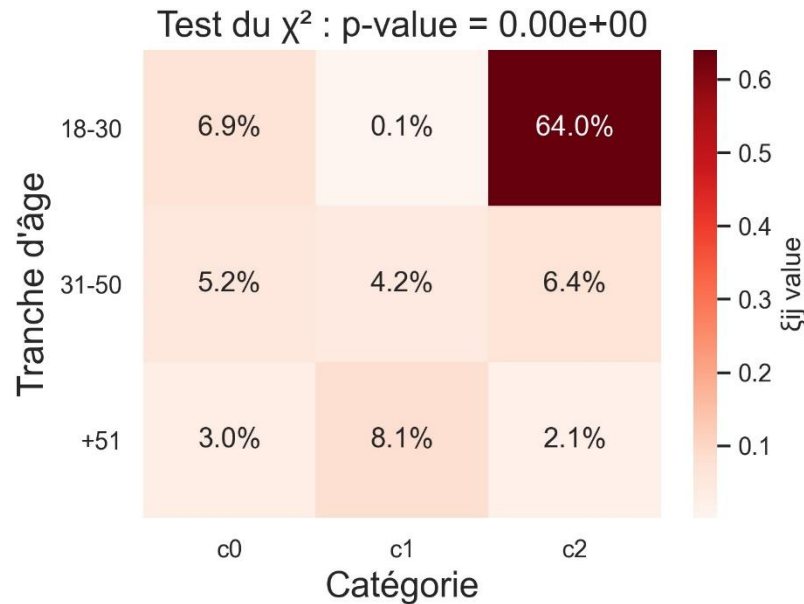
1. Présentation des graphs

GINI=0.4405



- Inégalité visible au travers de l'aire représentée ci-contre (indice de Gini $\gg 0$)
- Variabilité du montant du panier d'achat

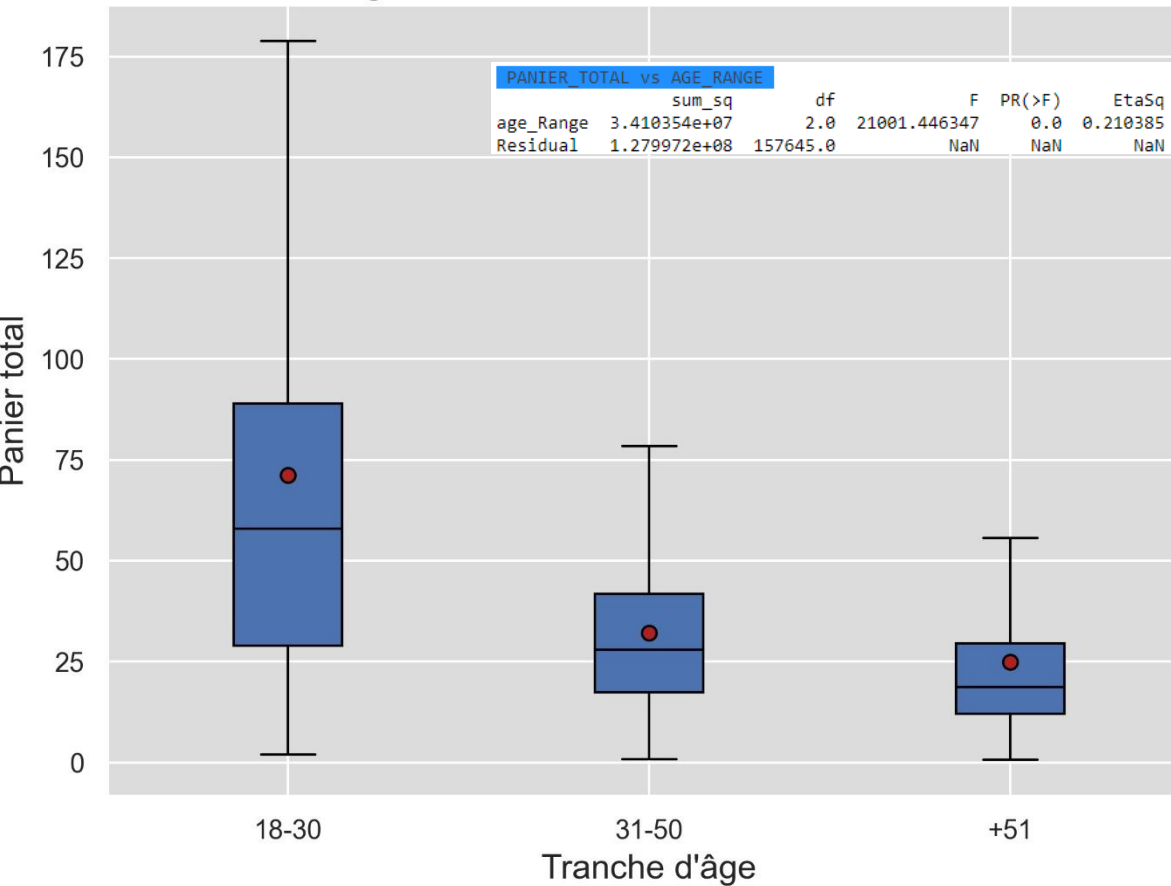
1. Test du Khi-2



- Avec $p_value < 5\%$ (degré de significativité fixé), on peut rejeter l'hypothèse d'indépendance (H_0).
 - Les variables « Age des clients » et « Catégorie des produits » sont corrélées.
 - Les variables « Sexe des clients » et « Catégorie des produits » sont corrélées.
- Le fort attrait des « jeunes » pour les produits de « catégorie 2 » est confirmée ici.

2. ANOVA

Corrélation Age des clients vs Panier total



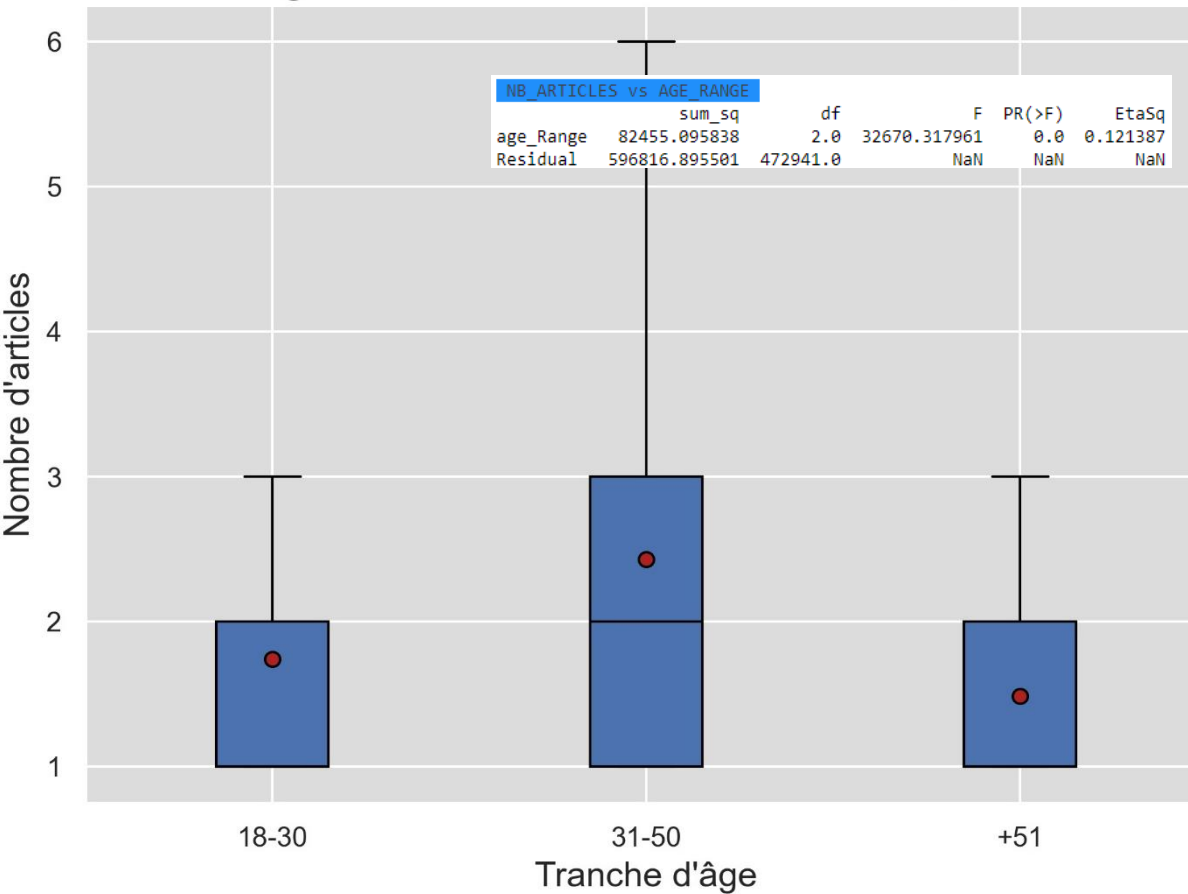
Parametric test assumptions

- Population distributions are normal
- Samples have equal variances
- Independence

- Inégalité des moyennes visible (en rouge)
- $F > 1$ et $p_value < 5\% \Rightarrow$ rejet H_0
- $\eta^2 = 0.21$ (forte corrélation)
- + le client est « jeune » + le montant du panier augmente

2. ANOVA

Corrélation Age des clients vs Nombre d'articles



Parametric test assumptions

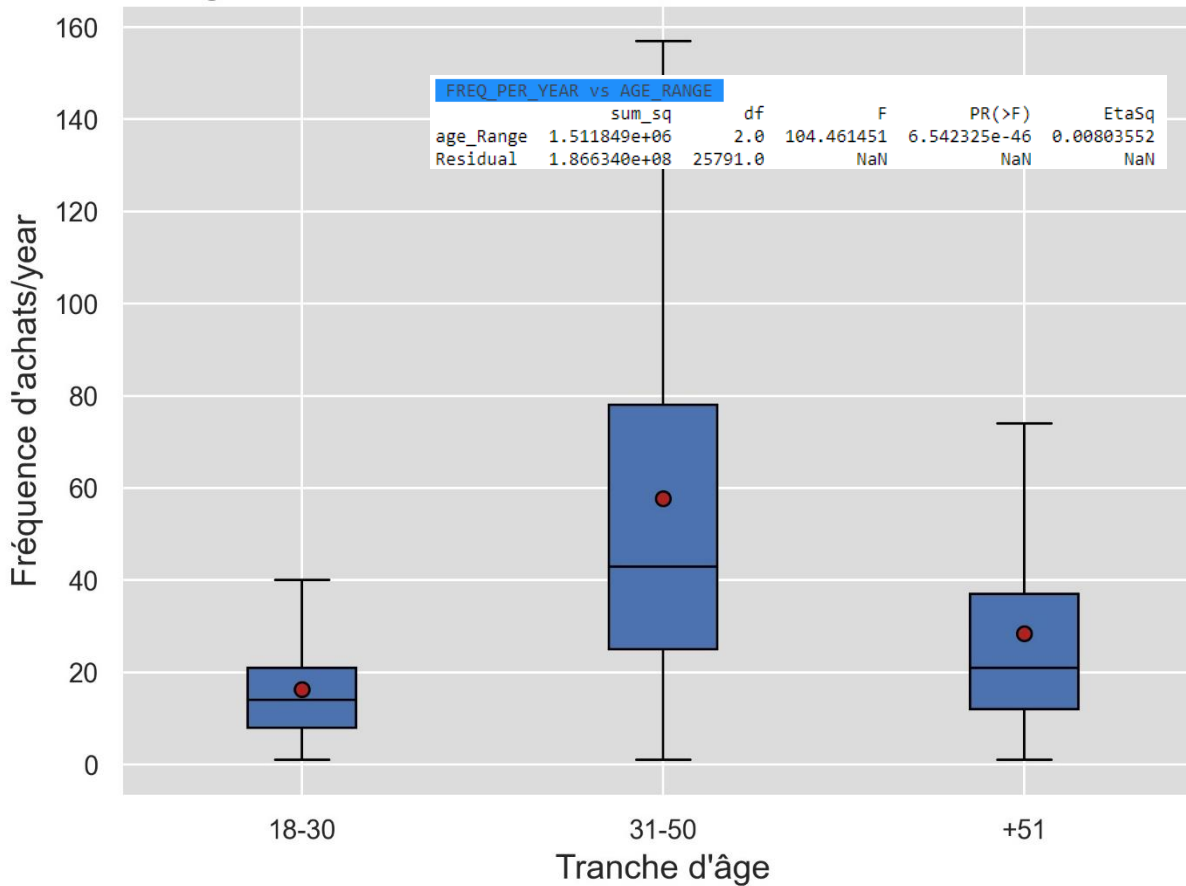
- Population distributions are normal
- Samples have equal variances
- Independence

- Inégalité des moyennes visible (en rouge)
- $F > 1$ et $p_value < 5\% \Rightarrow$ rejet H_0
- $\eta^2 = 0.12$ (corrélation moyenne)
- Les clients « mid-âge » achète plus d'articles (de façon assez équilibrée sur les 3 catégories cf. test Khi-2)

2. ANOVA

Corrélation

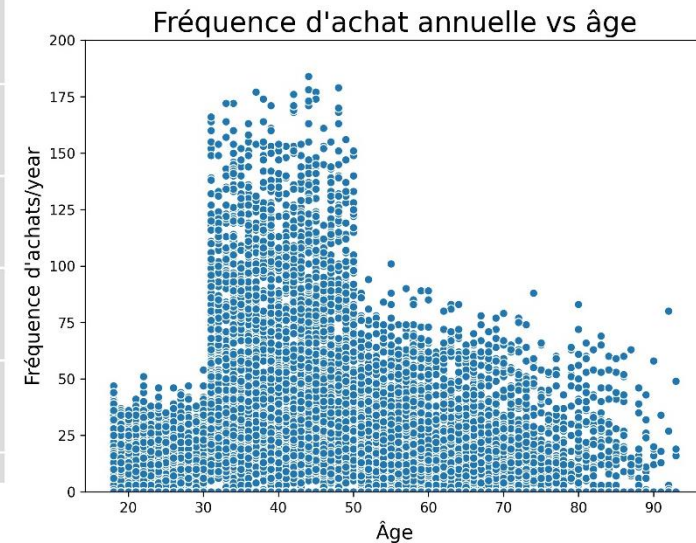
Age des clients vs Fréquence d'achat annuelle



Parametric test assumptions

- Population distributions are normal
- Samples have equal variances
- Independence

- Inégalité des moyennes visible (en rouge)
- $F > 1$ et $p_value < 5\% \Rightarrow$ rejet H_0
- $\eta^2 < 0.01$ (faible corrélation)

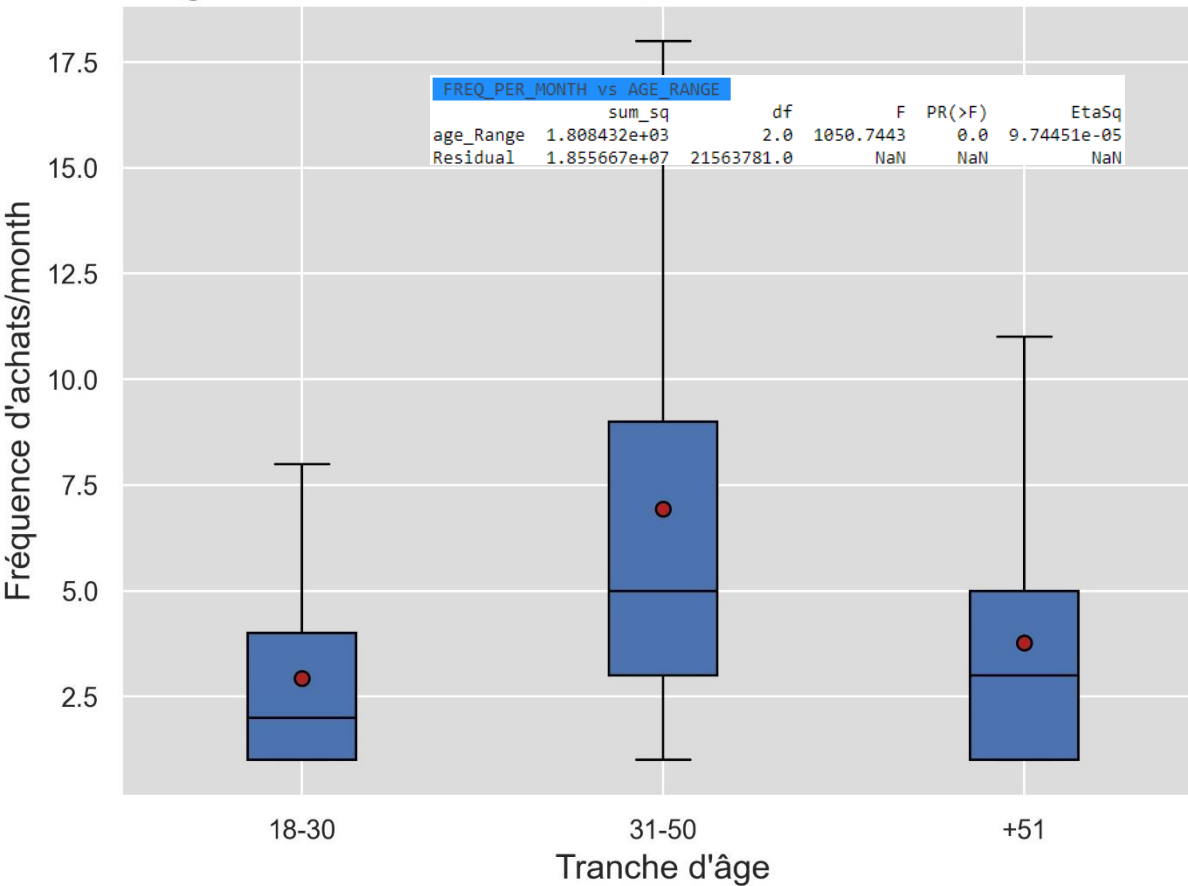


- sur une année, les clients « mid-âge » achètent jusqu'à 3 fois plus que les « jeunes » et 2 fois plus que les « seniors »

2. ANOVA

Corrélation

Age des clients vs Fréquence d'achat mensuelle

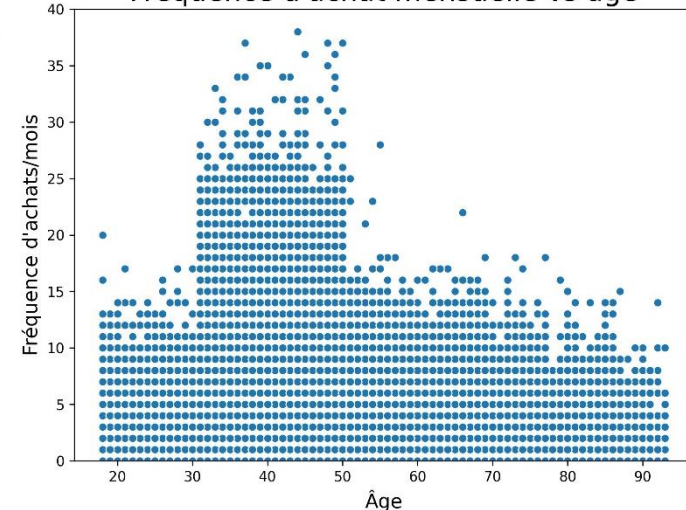


Parametric test assumptions

- Population distributions are normal
- Samples have equal variances
- Independence

- Inégalité des moyennes visible (en rouge)
- $F > 1$ et $p_value < 5\% \Rightarrow$ rejet H_0
- $\eta^2 < 0.01$ (faible corrélation)

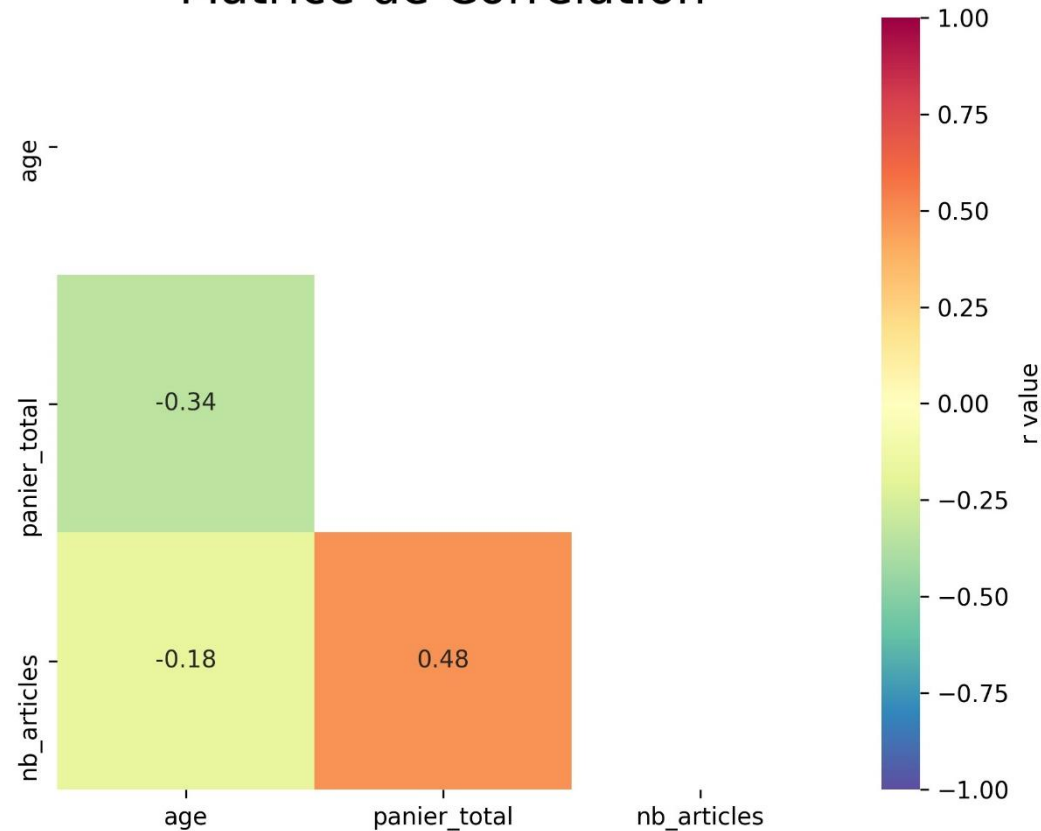
Fréquence d'achat mensuelle vs âge



- mêmes tendances que pour la fréquence annuelle d'achats

3. Coefficient de corrélation de Pearson

Matrice de Corrélation



# Corrélation	$r < 0$	$r > 0$
# Faible	-0,5 à 0,0	0,0 à 0,5
# Forte	-1,0 à -0,5	0,5 à 1,0

- Corrélation négative :

Panier ↗ avec l'âge ↘
(cf. 1^{ère} ANOVA)

- En toute logique :

Panier ↗ avec nb articles ↗

- CA en croissance
- Les « jeunes » dépensent peu mais s'orientent vers les produits les plus onéreux
- Les « mid-âge » dépensent en grand nombre et plus souvent que les autres clients (ciblent les produits les moins chers)
- Les « seniors » couvrent le reste du panel avec un attrait pour les produits médians

Merci

