



Lapage Book Store

Analyse des ventes de mars 2021 à février 2023

Arnaud Golliot
Data Consultant
03 mars 2025

Sommaire

1. Préparation des données

- ✓ Données brutes collectées
- ✓ Processus de nettoyage, transformation et analyses des données
- ✓ Modèle de données
- ✓ Dataset global

2. Analyses descriptives

3. Analyses univariées

4. Analyses bivariées

5. Conclusions de l'étude

- ✓ Synthèse de l'étude et orientations proposées
- ✓ Limites de l'étude

6. Annexes

1 - PRÉPARATION DES DONNÉES

Données brutes collectées (3 fichiers CSV)

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

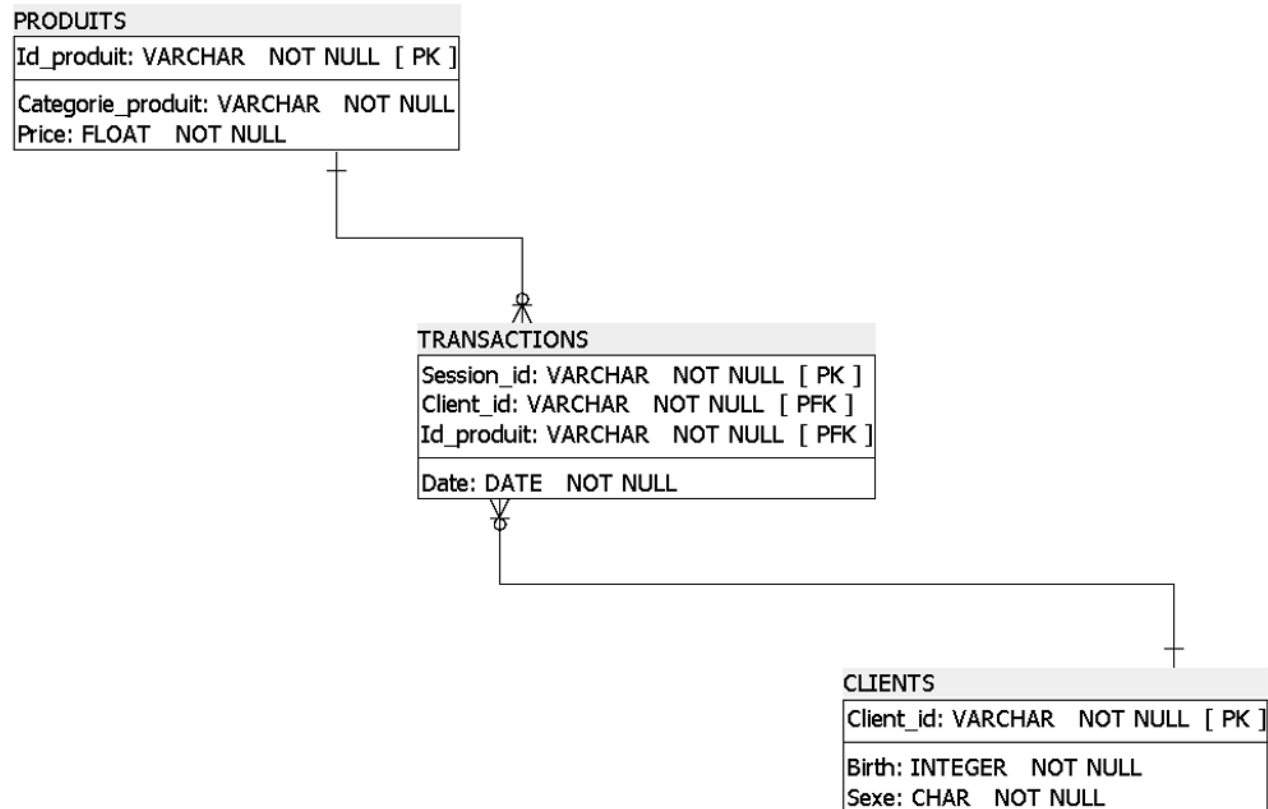
produits.csv

	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

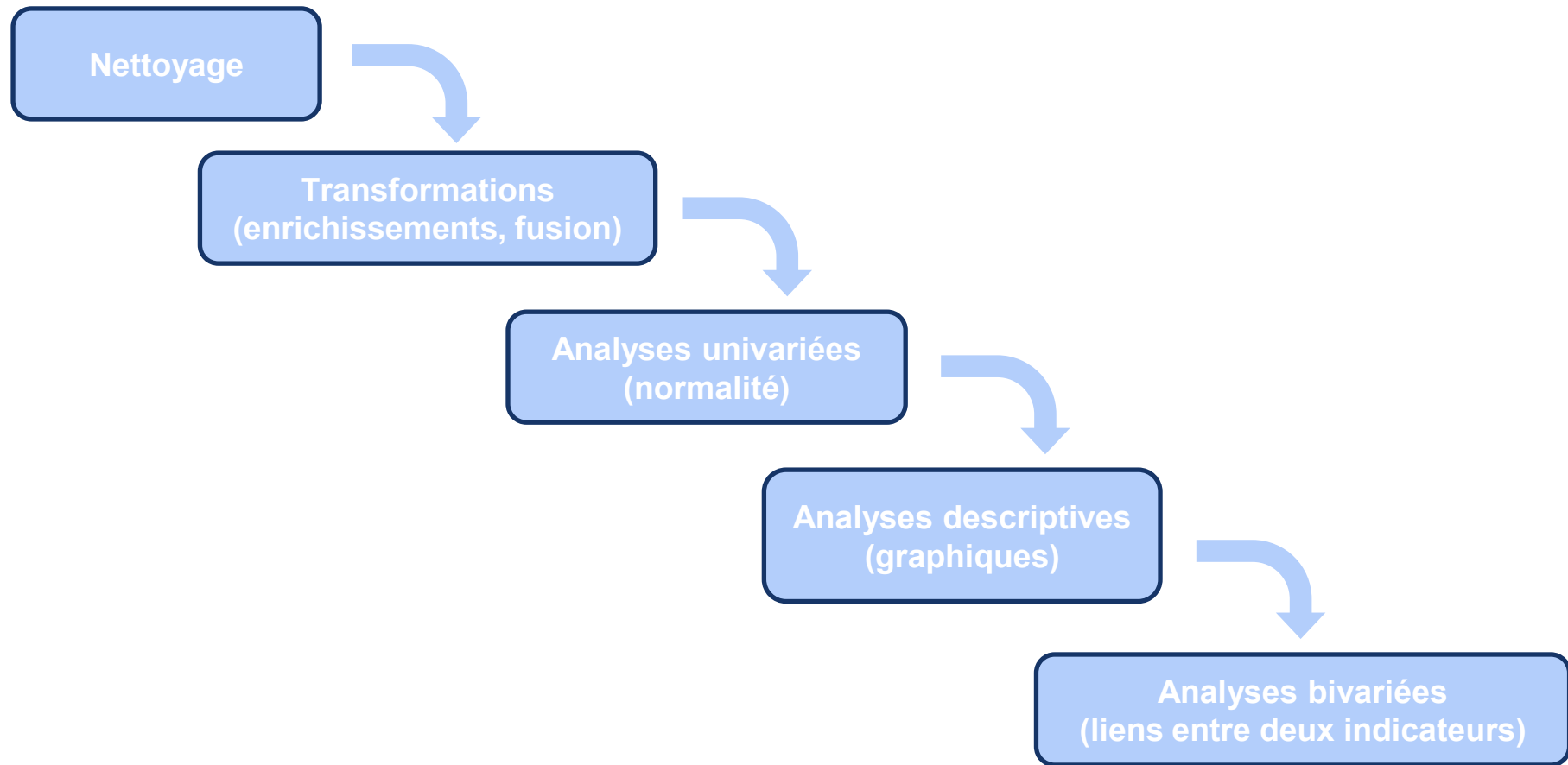
transactions.csv

	id_prod	date	session_id	client_id
0	0_1259	2021-03-01 00:01:07.843138	s_1	c_329
1	0_1390	2021-03-01 00:02:26.047414	s_2	c_664
2	0_1352	2021-03-01 00:02:38.311413	s_3	c_580
3	0_1458	2021-03-01 00:04:54.559692	s_4	c_7912
4	0_1358	2021-03-01 00:05:18.801198	s_5	c_2033

Modèle des données



Processus de nettoyage, transformation et analyses des données



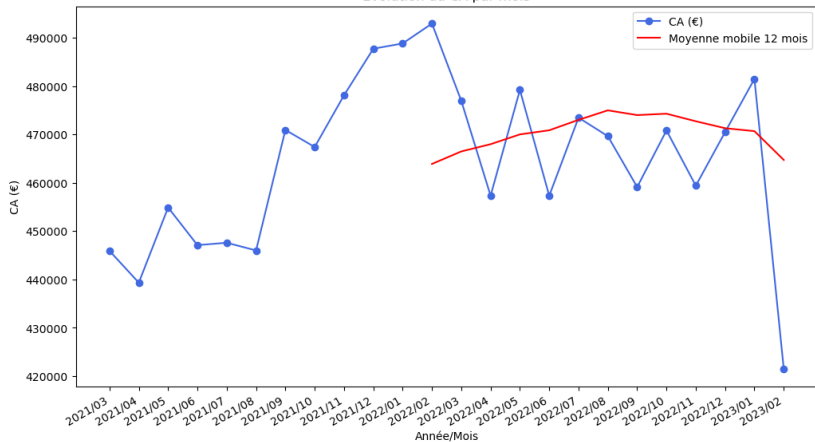
Dataset résultant

	id_prod	date	session_id	client_id	Converted_date	Année	Mois	Année_string	Mois_string	Année-Mois	price	categ	Catégorie	sex	birth	Âge	Tranche d'âge
0	0_1259	2021-03-01 00:01:07.843138	s_1	c_329	2021-03-01 00:01:07.843138	2021	3	2021	03	2021/03	11.99	0	Cat_0	f	1967	58	[51-60a]
1	0_1390	2021-03-01 00:02:26.047414	s_2	c_664	2021-03-01 00:02:26.047414	2021	3	2021	03	2021/03	19.37	0	Cat_0	m	1960	65	[61-70a]
2	0_1352	2021-03-01 00:02:38.311413	s_3	c_580	2021-03-01 00:02:38.311413	2021	3	2021	03	2021/03	4.50	0	Cat_0	m	1988	37	[31-40a]
3	0_1458	2021-03-01 00:04:54.559692	s_4	c_7912	2021-03-01 00:04:54.559692	2021	3	2021	03	2021/03	6.55	0	Cat_0	f	1989	36	[31-40a]
4	0_1358	2021-03-01 00:05:18.801198	s_5	c_2033	2021-03-01 00:05:18.801198	2021	3	2021	03	2021/03	16.49	0	Cat_0	f	1956	69	[61-70a]

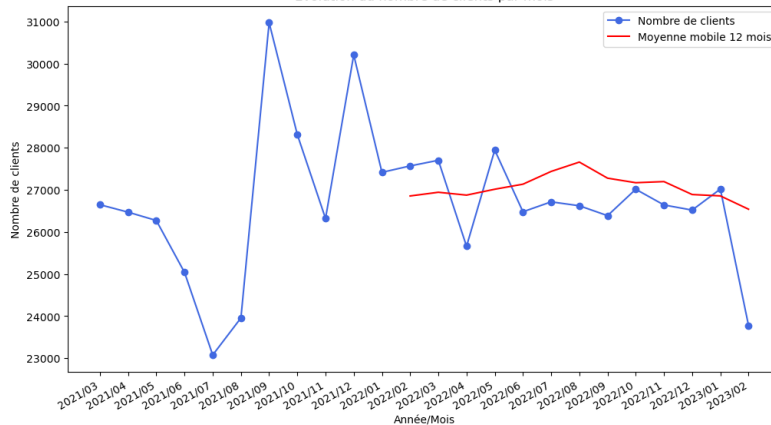
2 - ANALYSES DESCRIPTIVES

Courbes d'évolution mensuelle

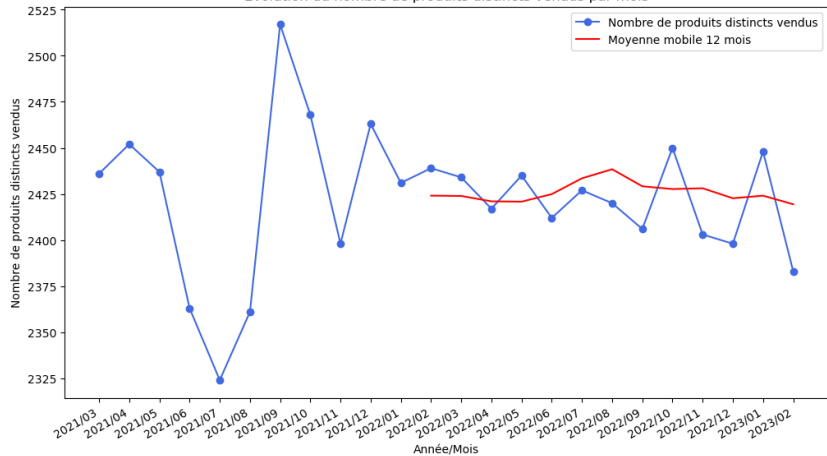
Évolution du CA par mois



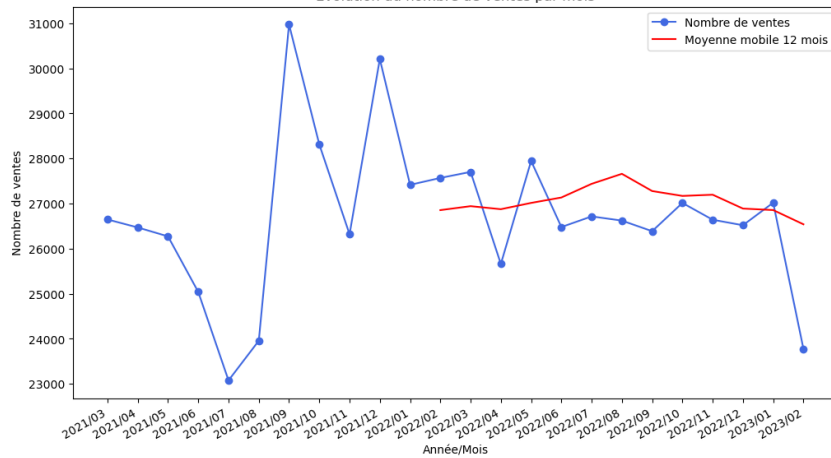
Évolution du nombre de clients par mois



Évolution du nombre de produits distincts vendus par mois

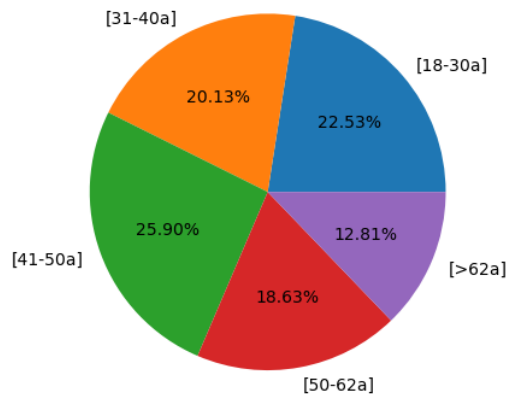


Évolution du nombre de ventes par mois

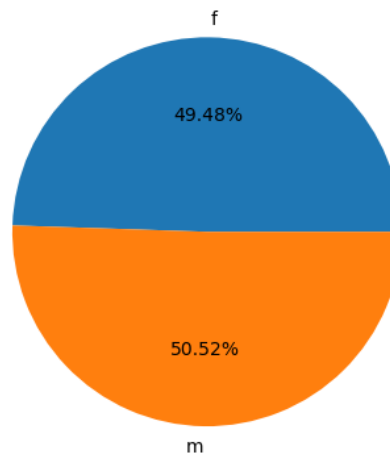


Répartition du CA (€)

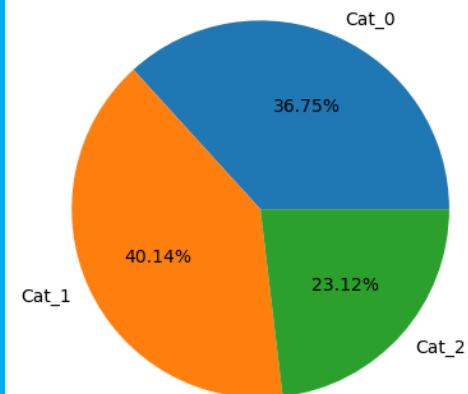
Répartition du CA(€) par tranche d'âge



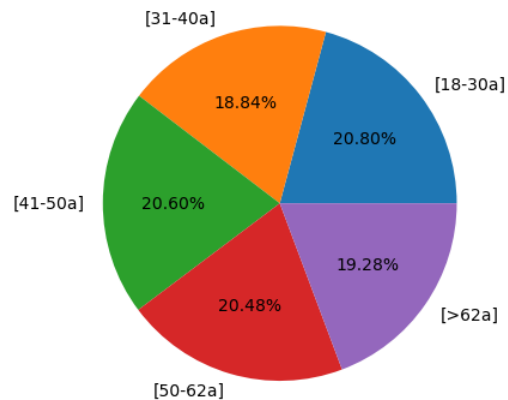
Répartition du CA(€) par sexe



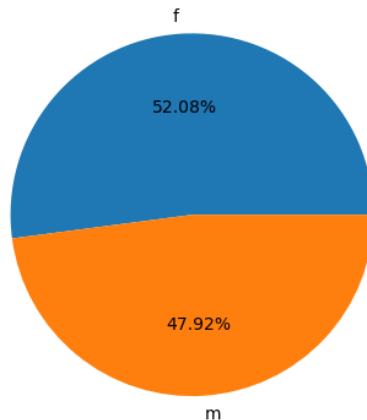
Répartition du CA(€) par catégorie de produit



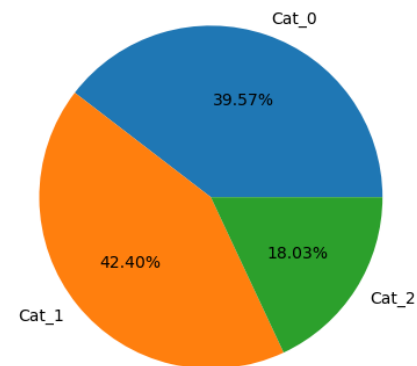
Répartition du nombre de clients



Répartition du nombre de clients par genre (M, F)

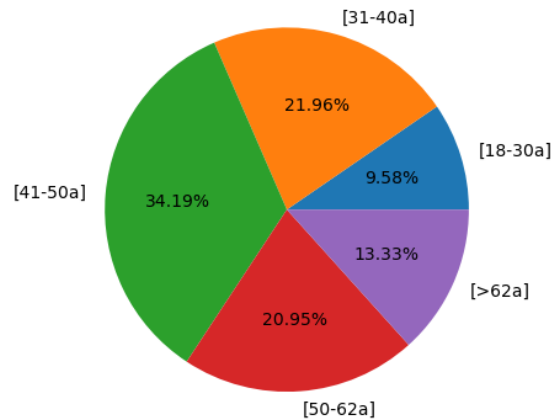


Répartition du nombre de clients par catégories de produit

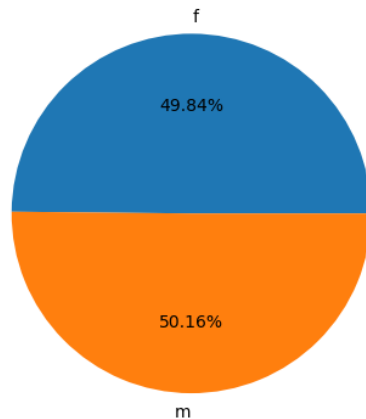


Répartition du nombre d'achats

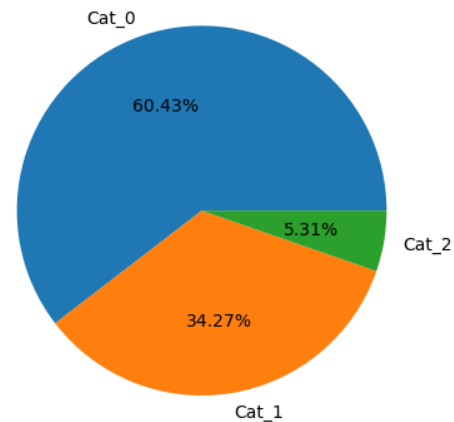
Répartition du nombre de transactions par tranche d'âge



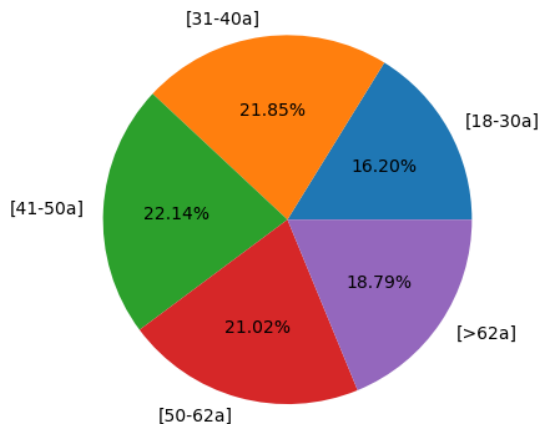
épartition du nombre de transactions par genre (M, F)



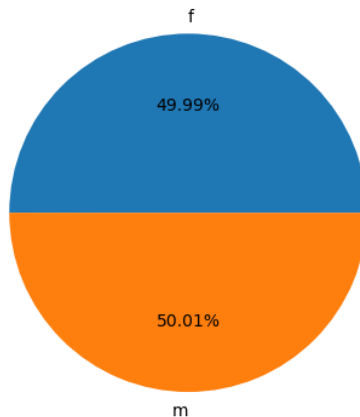
Répartition du nombre de transactions par catégories de produit



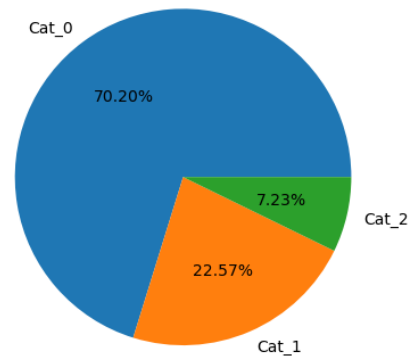
Répartition du nombre de produits



Répartition du nombre de produits vendus par genre (M, F)

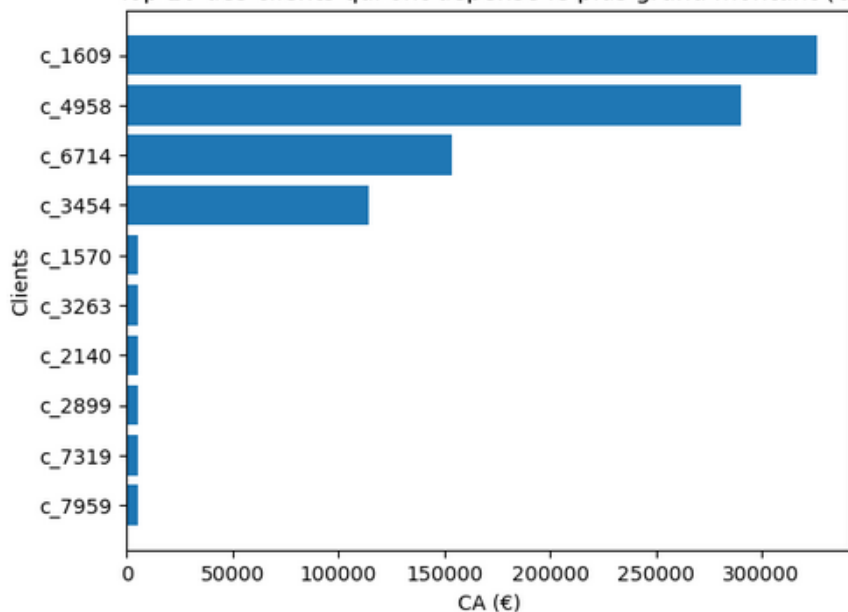


Répartition du nombre de produits vendus par catégorie de produit

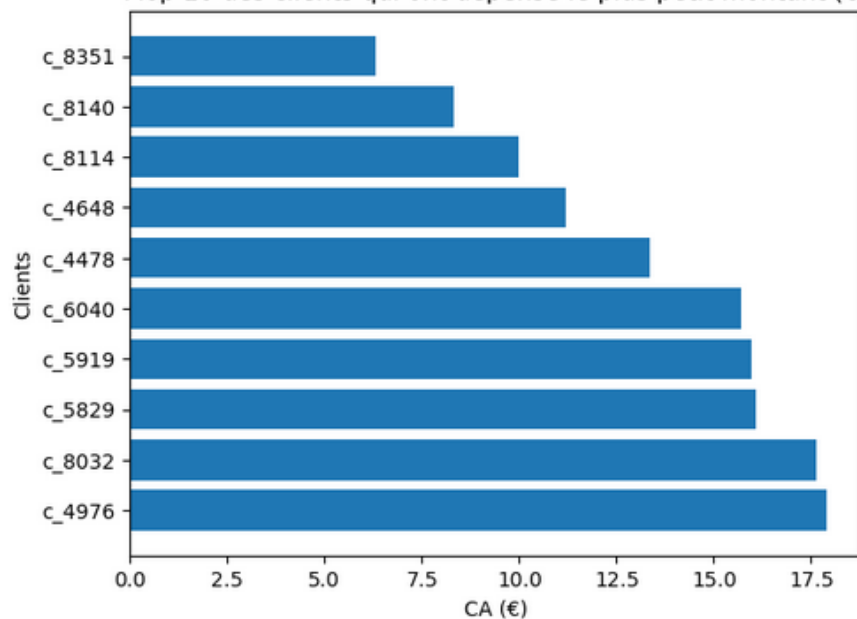


Top 10 et Flop 10 des clients qui ont acheté (€)

Top 10 des clients qui ont dépensé le plus grand montant (€)

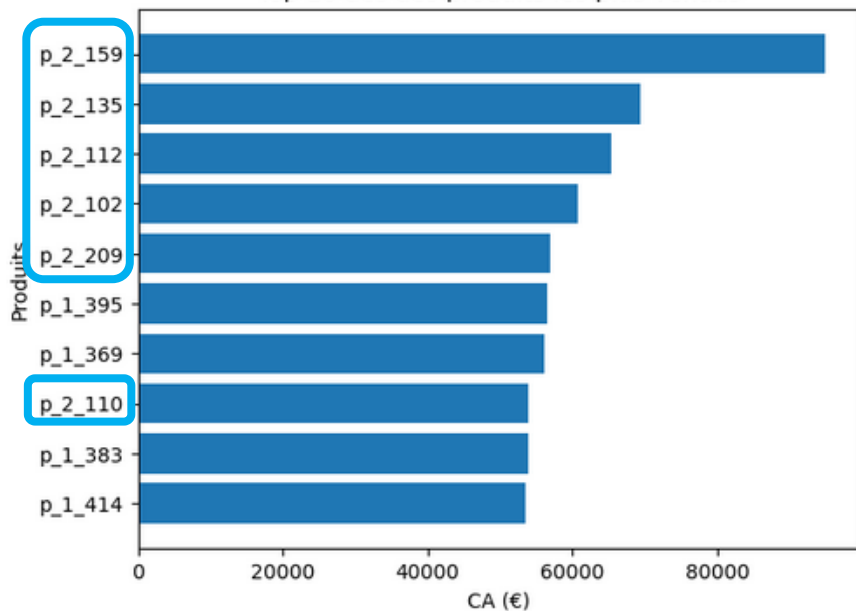


Flop 10 des clients qui ont dépensé le plus petit montant (€)

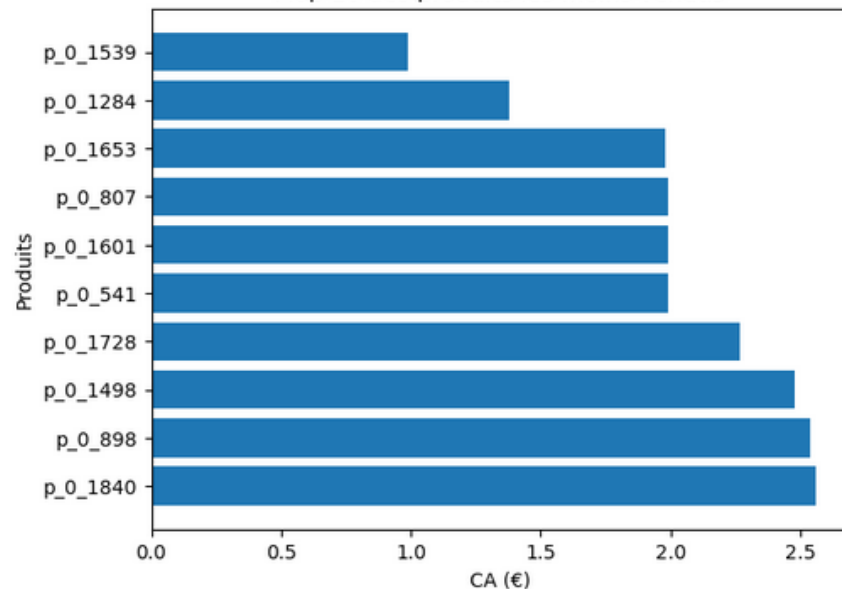


Top 10 et Flop 10 des produits vendus (€)

Top 10 des des produits les plus vendus



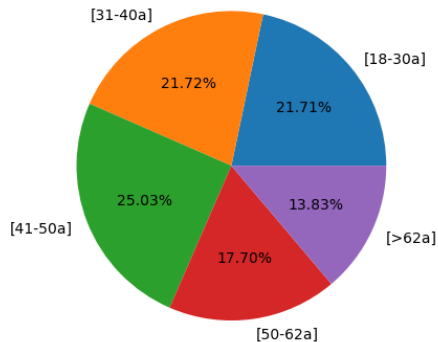
Flop 10 des produits les moins vendus



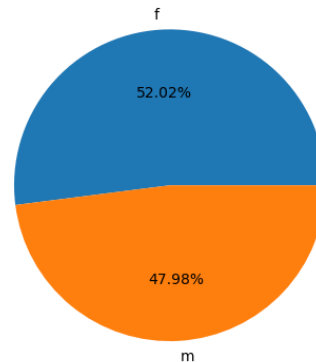
Répartition du CA (€) des clients particuliers versus professionnels

Clients particuliers

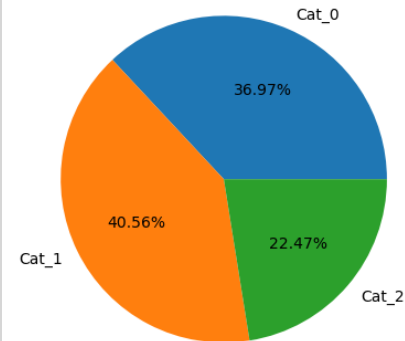
Répartition du CA(€) par tranche d'âge



Répartition du CA(€) par sexe

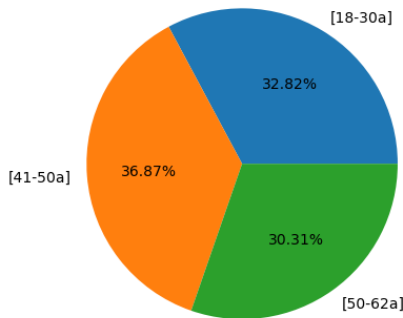


Répartition du CA(€) par catégorie de produit

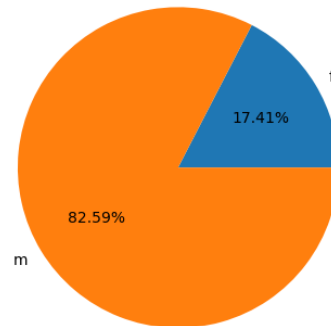


Clients professionnels

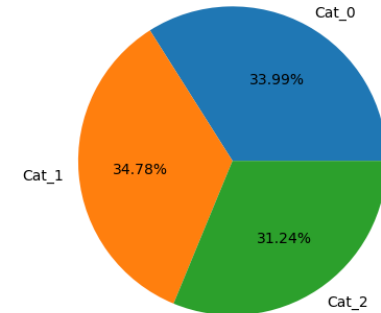
Répartition du CA(€) par tranche d'âge



Répartition du CA(€) par sexe

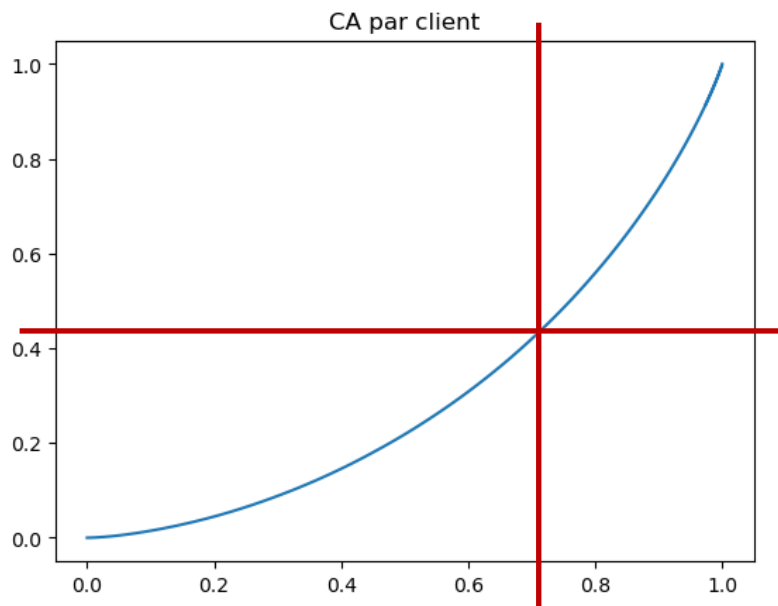


Répartition du CA(€) par catégorie de produit

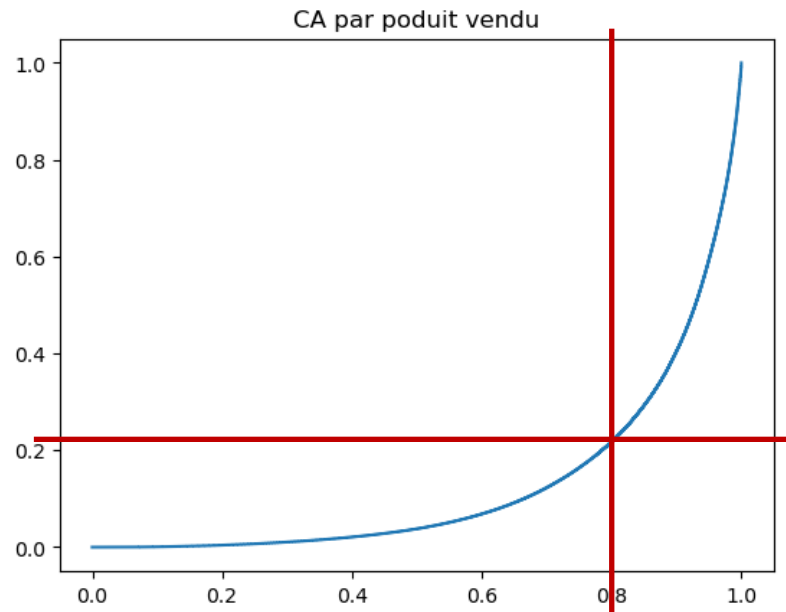


Courbes de lorenz

Sans les clients B to B (valeurs atypiques retirées)



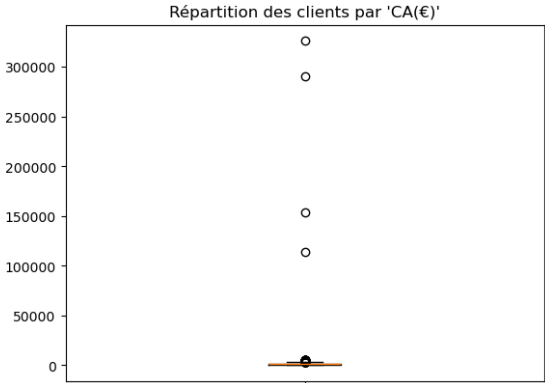
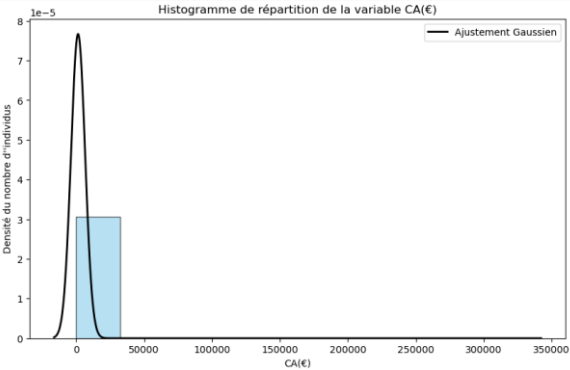
Près de 60% du CA est généré par 30% des clients



Près de 80% du CA est généré par 20% des produits vendus

3 - ANALYSES UNIVARIÉES

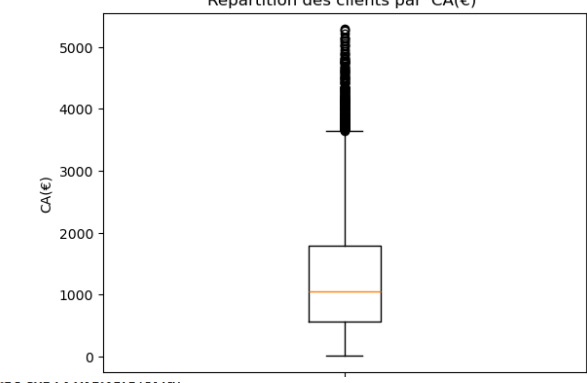
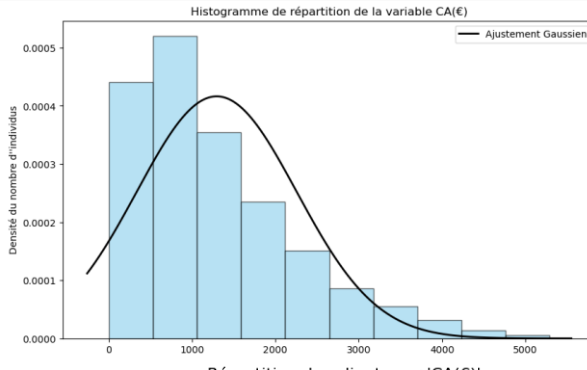
5 – Tests de normalité : répartition des clients par CA total (€)



Avec les clients B to B

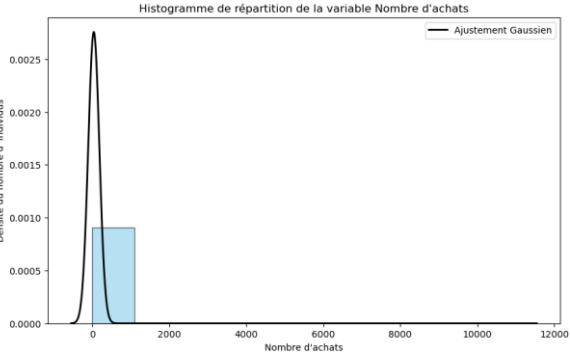
TEST DE SHAPIRO SUR LA VARIABLE 'CA(€)'	
H0 :	La répartition de la variable 'CA(€)' SUIE une loi normale
H1 :	La répartition de la variable 'CA(€)' NE SUIE PAS une loi normale
W :	0.06099
P value :	0.0
Analyse du Résultat :	W <= 0.95 ET p_value <= 0.05 (on rejette H0)
Conclusion :	La répartition de la variable 'CA(€)' NE SUIE PAS une loi normale

Sans les clients B to B

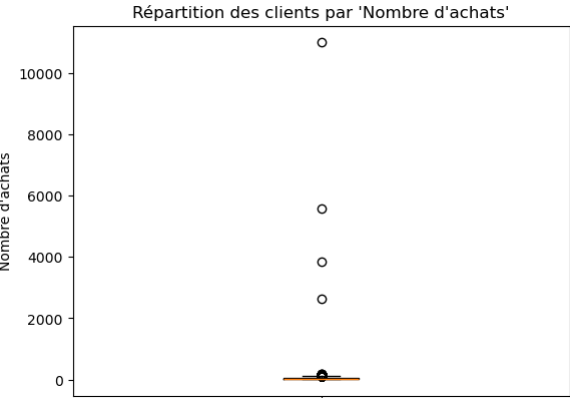


TEST DE SHAPIRO SUR LA VARIABLE 'CA(€)'	
H0 :	La répartition de la variable 'CA(€)' SUIE une loi normale
H1 :	La répartition de la variable 'CA(€)' NE SUIE PAS une loi normale
W :	0.904287
P value :	0.0
Analyse du Résultat :	W <= 0.95 ET p_value <= 0.05 (on rejette H0)
Conclusion :	La répartition de la variable 'CA(€)' NE SUIE PAS une loi normale

5 – Tests de normalité : répartition des clients par nombre d'achats

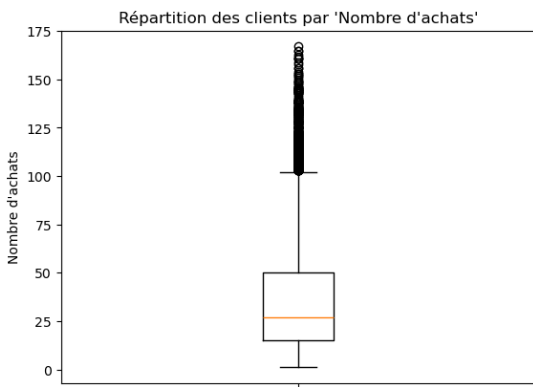
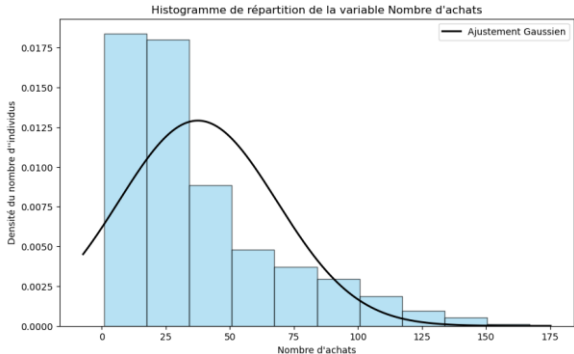


Avec les clients B to B



TEST DE SHAPIRO SUR LA VARIABLE 'Nombre d'achats'	
H0 :	La répartition de la variable 'Nombre d'achats' S'UIT une loi normale
H1 :	La répartition de la variable 'Nombre d'achats' NE S'UIT PAS une loi normale
W :	0.070596
P value :	0.0
Analyse du Résultat :	W <= 0.95 ET p_value <= 0.05 (on rejette H0)
Conclusion :	La répartition de la variable 'Nombre d'achats' NE S'UIT PAS une loi normale

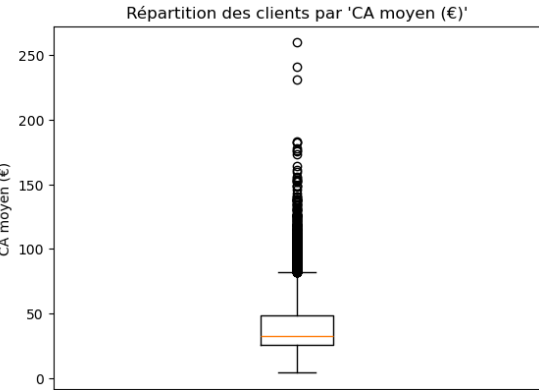
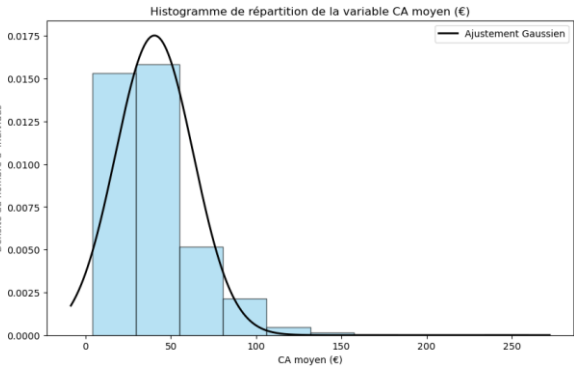
Sans les clients B to B



TEST DE SHAPIRO SUR LA VARIABLE 'Nombre d'achats'	
H0 :	La répartition de la variable 'Nombre d'achats' S'UIT une loi normale
H1 :	La répartition de la variable 'Nombre d'achats' NE S'UIT PAS une loi normale
W :	0.857498
P value :	0.0
Analyse du Résultat :	W <= 0.95 ET p_value <= 0.05 (on rejette H0)
Conclusion :	La répartition de la variable 'Nombre d'achats' NE S'UIT PAS une loi normale

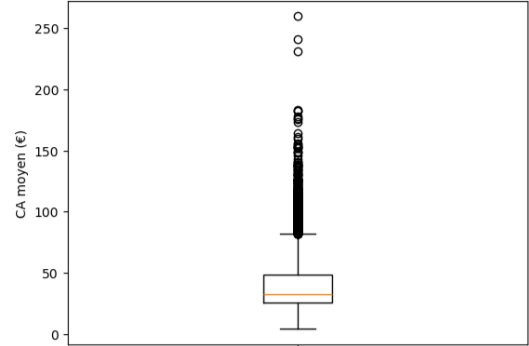
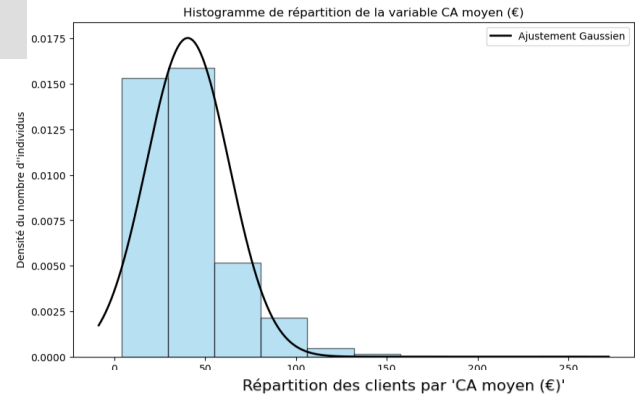
5 – Tests de normalité : répartition des clients par CA moyen (€)

Avec les clients B to B



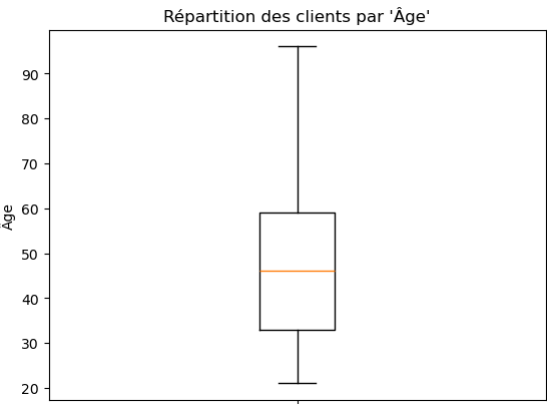
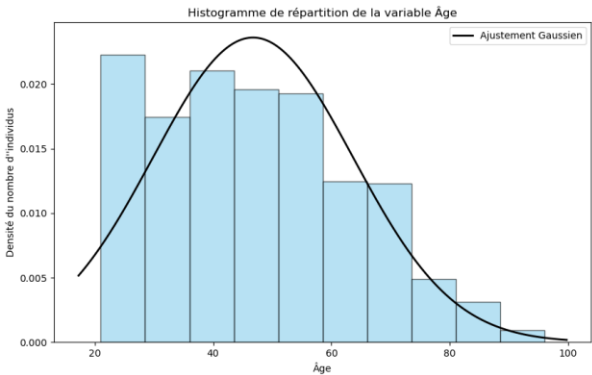
TEST DE SHAPIRO SUR LA VARIABLE 'CA moyen (€)'	
H0 :	La répartition de la variable 'CA moyen (€)' SUIT une loi normale
H1 :	La répartition de la variable 'CA moyen (€)' NE SUIT PAS une loi normale
W :	0.817257
P value :	0.0
Analyse du Résultat :	W <= 0.95 ET p_value <= 0.05 (on rejette H0)
Conclusion :	La répartition de la variable 'CA moyen (€)' NE SUIT PAS une loi normale

Sans les clients B to B



TEST DE SHAPIRO SUR LA VARIABLE 'CA moyen (€)'	
H0 :	La répartition de la variable 'CA moyen (€)' SUIT une loi normale
H1 :	La répartition de la variable 'CA moyen (€)' NE SUIT PAS une loi normale
W :	0.817162
P value :	0.0
Analyse du Résultat :	W <= 0.95 ET p_value <= 0.05 (on rejette H0)
Conclusion :	La répartition de la variable 'CA moyen (€)' NE SUIT PAS une loi normale

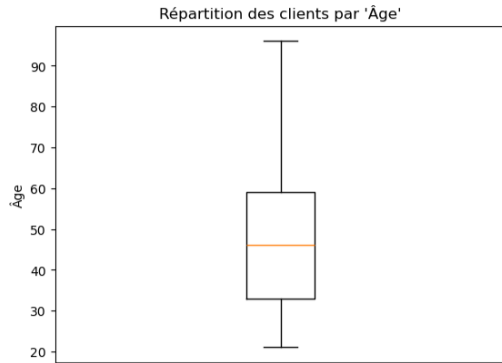
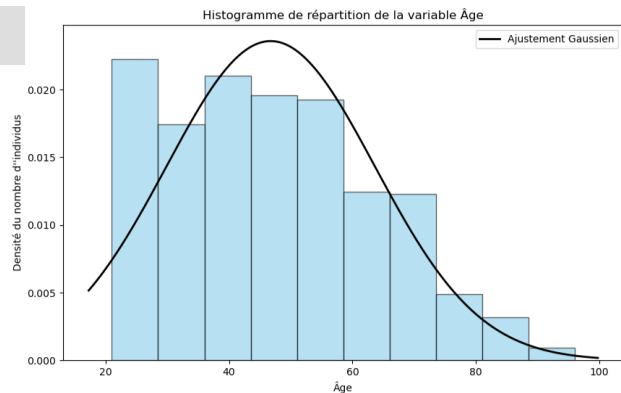
5 – Tests de normalité : répartition des clients par âge



TEST DE SHAPIRO SUR LA VARIABLE 'Âge'	
H0 :	La répartition de la variable 'Âge' SUIE une loi normale
H1 :	La répartition de la variable 'Âge' NE SUIE PAS une loi normale
W :	0.969747
P value :	0.0
Analyse du Résultat :	W > 0.95 ET p_value <= 0.05 (on rejette H0)
Conclusion :	La répartition de la variable 'Âge' NE SUIE PAS une loi normale

Avec les clients B to B

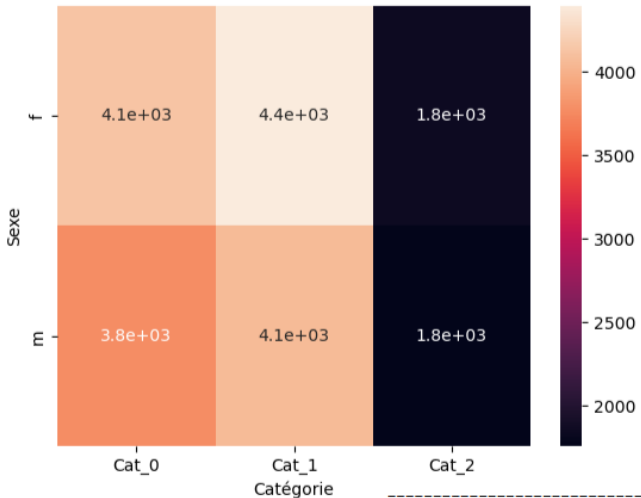
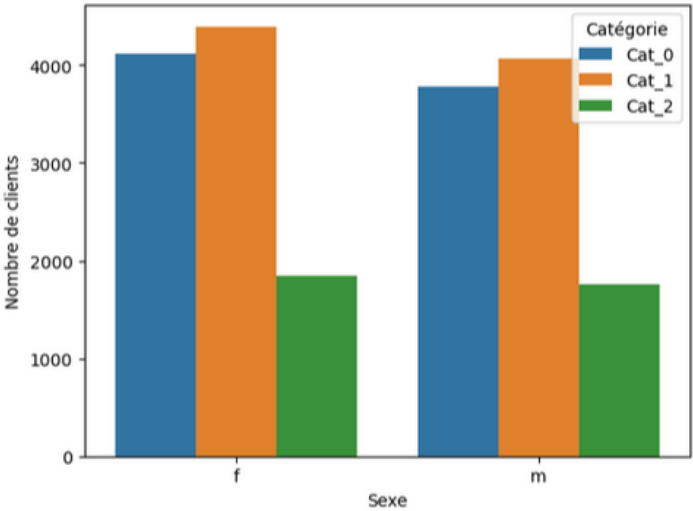
Sans les clients B to B



TEST DE SHAPIRO SUR LA VARIABLE 'Âge'	
H0 :	La répartition de la variable 'Âge' SUIE une loi normale
H1 :	La répartition de la variable 'Âge' NE SUIE PAS une loi normale
W :	0.969731
P value :	0.0
Analyse du Résultat :	W > 0.95 ET p_value <= 0.05 (on rejette H0)
Conclusion :	La répartition de la variable 'Âge' NE SUIE PAS une loi normale

4 - ANALYSES BIVARIÉES

6 – Analyses bivariées : Lien entre le genre et la catégorie de livres achetés



=====

Fréquences observées entre les variables 'Catégorie' et 'Genre'

=====

Catégorie	Cat_0	Cat_1	Cat_2
Sexe			
f	4119	4394	1839
m	3773	4061	1754

=====

Fréquences attendues entre les variables 'Catégorie' et 'Genre'

=====

[[4097.19077232 4389.47642929 1865.3327984]

[3794.80922768 4065.52357071 1727.6672016]]

=====

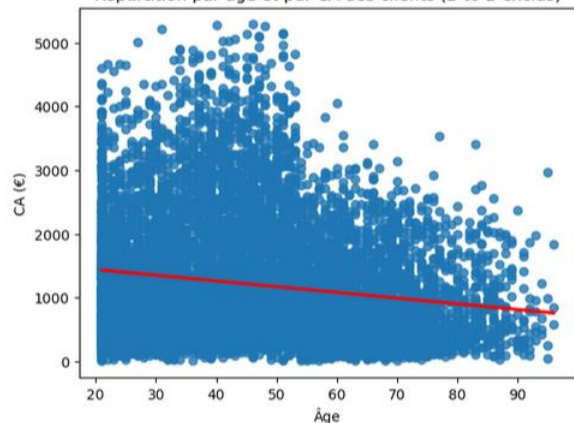
Résultat du test du Chi2 entre les variables 'Catégorie' et 'Genre'

=====

TEST DU CHI-2 SUR LES VARIABLES 'Sexe' et 'Catégorie'	
H0 :	Les variables 'Sexe' et 'Catégorie' NE SONT PAS corrélées
H1 :	Les variables 'Sexe' et 'Catégorie' SONT corrélées
Chi2 :	1.024224
Valeur P :	0.599229
Degré de liberté :	2
Analyse du Résultat :	p_value > 0.05 (on ne rejette pas H0)
Conclusion :	NON, les variables 'Sexe' et 'Catégorie' NE SONT PAS corrélées

6 – Analyses bivariées : Lien entre l'âge et le montant des achats des clients [1/2]

Répartition par âge et par CA des clients (B to B exclus)



1

TEST DE SPEARMAN SUR LES VARIABLES 'Âge' et 'CA (€)'

H0 : Les variables 'Âge' et 'CA (€)' NE SONT PAS corrélées

H1 : Les variables 'Âge' et 'CA (€)' SONT corrélées

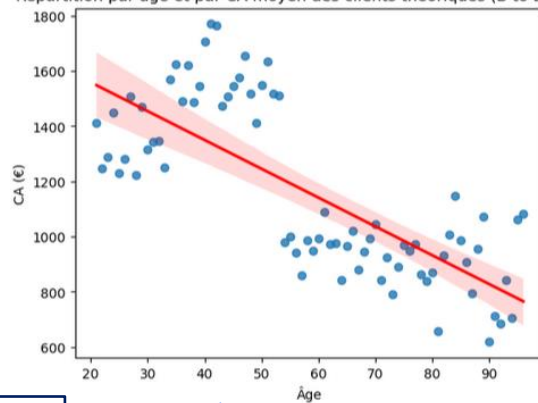
Coefficient de corrélation : -0.184538

Valeur P : 0.0

Analyse du Résultat : $p_value \leq 0.05$ (on rejette H0)

Conclusion : OUI, les variables 'Âge' et 'CA (€)' SONT corrélées

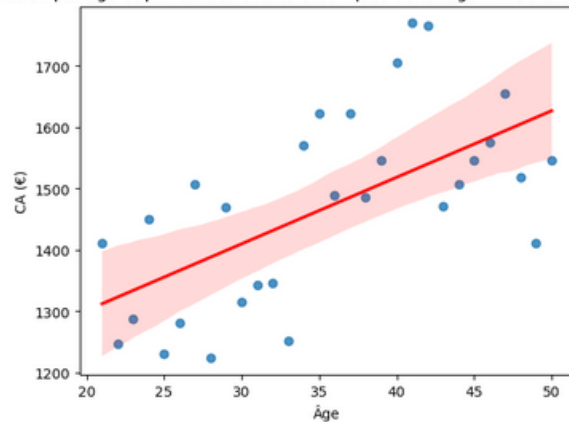
Répartition par âge et par CA moyen des clients théoriques (B to B exclus)



2

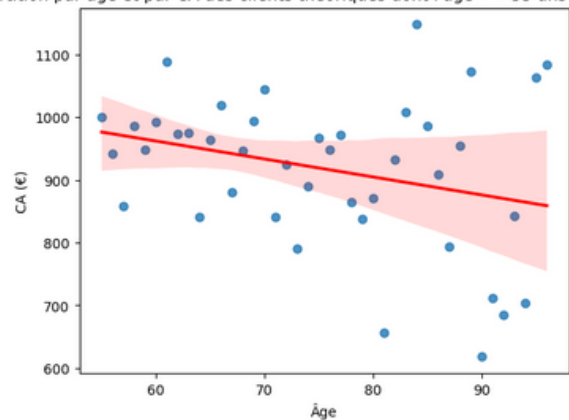
3

Répartition par âge et par CA des clients théoriques dont l'âge ≤ 50 ans (B to B exclus)



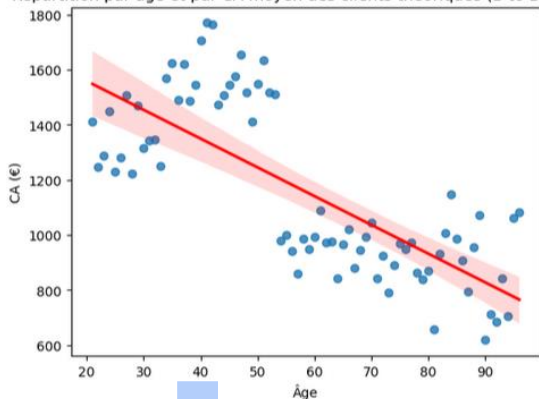
4

Répartition par âge et par CA des clients théoriques dont l'âge ≥ 55 ans (B to B exclus)



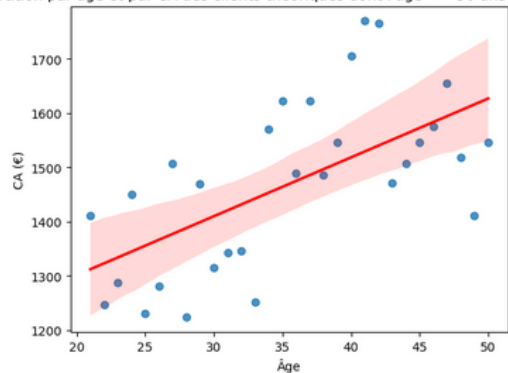
6 – Analyses bivariées : Lien entre l'âge et le montant des achats des clients [2/2]

Répartition par âge et par CA moyen des clients théoriques (B to B exclus)



Clients <= 50 ans

Répartition par âge et par CA des clients théoriques dont l'âge <= 50 ans (B to B exclus)



OLS Regression Results						
Dep. Variable:	CA (€)		R-squared:	0.388		
Model:	OLS		Adj. R-squared:	0.366		
Method:	Least Squares		F-statistic:	17.73		
Date:	Thu, 27 Feb 2025		Prob (F-statistic):	0.000239		
Time:	17:05:22		Log-Likelihood:	-186.18		
No. Observations:	30		AIC:	376.4		
Df Residuals:	28		BIC:	379.2		
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Âge	11.0281	2.619	4.210	0.000	5.663	16.394
Intercept	1080.8619	95.711	11.293	0.000	884.807	1276.917
Omnibus:	1.986		Durbin-Watson:	1.548		
Prob(Omnibus):	0.371		Jarque-Bera (JB):	1.351		
Skew:	0.270		Prob(JB):	0.509		
Kurtosis:	2.112		Cond. No.	154.		

Résidus Standardisés	
0	-85.462920
1	-210.751244
2	-91.117051
3	56.105712
4	-12.723761
5	-31.509471
6	-59.109593
7	-83.399673
8	221.232314
9	236.839576
10	183.267690
11	35.321688
12	-14.418285
13	133.220012
14	11.867148
15	155.866012
16	114.358543
17	-193.587841
18	-87.377127
19	-79.888305
20	-96.007350
21	68.405536
22	-166.292118
23	128.910709
24	-86.740172
25	-125.768552
26	104.333354
27	-46.287636
28	-77.403997
29	98.116802

4

TEST DE SHAPIRO SUR LA VARIABLE 'Résidus Standardisés'

H0 : La répartition de la variable 'Résidus Standardisés' S'UIT une loi normale

H1 : La répartition de la variable 'Résidus Standardisés' NE S'UIT PAS une loi normale

W : 0.957693

P value : 0.270122

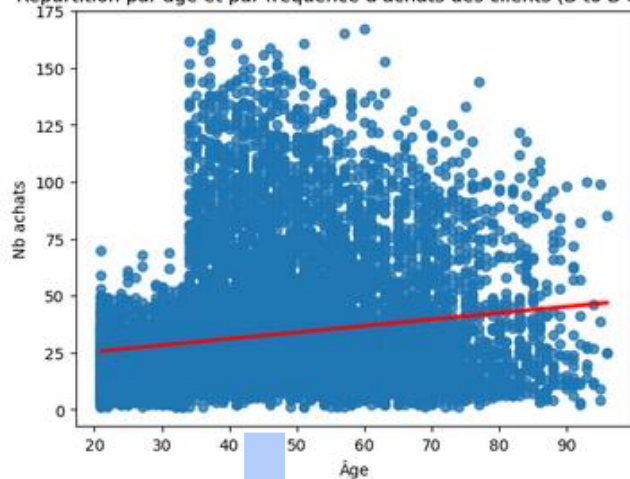
Analyse du Résultat : W > 0.95 ET p. value > 0.05 (on ne rejette pas H0)

Conclusion : OUI, La répartition de la variable 'Résidus Standardisés' S'UIT une loi normale

Âge	10	15	18	20	25	30	35	40	45	50
Prévision de CA moyen (€)	1191 €	1246 €	1279 €	1301 €	1357 €	1412 €	1467 €	1522 €	1577 €	1632 €
Intervalle de confiance (€)	+/- 71 €	+/- 58 €	+/- 51 €	+/- 47 €	+/- 36 €	+/- 27 €	+/- 23 €	+/- 26 €	+/- 34 €	+/- 44 €

6 – Analyses bivariées : Lien entre l'âge et la fréquence d'achat des clients [1/2]

Répartition par âge et par fréquence d'achats des clients (B to B exclus)



1

TEST DE SPEARMAN SUR LES VARIABLES 'Âge' et 'CA (€)'

H0 : Les variables 'Âge' et 'CA (€)' NE SONT PAS corrélées

H1 : Les variables 'Âge' et 'CA (€)' SONT corrélées

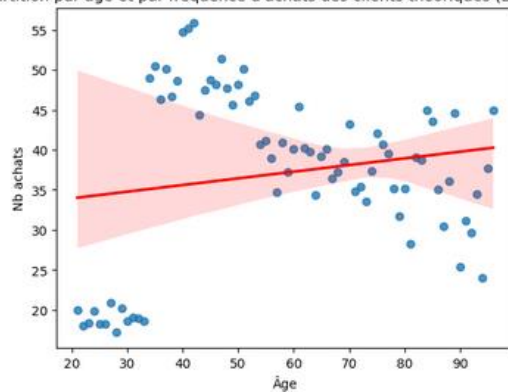
Coefficient de corrélation : -0.184538

Valeur P : 0.0

Analyse du Résultat : p_value <= 0.05 (on rejette H0)

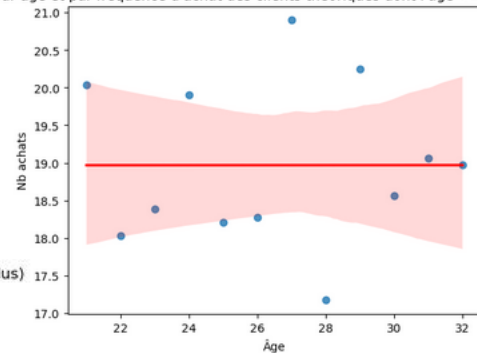
Conclusion : OUI, les variables 'Âge' et 'CA (€)' SONT corrélées

Répartition par âge et par fréquence d'achats des clients théoriques (B to B exclus)



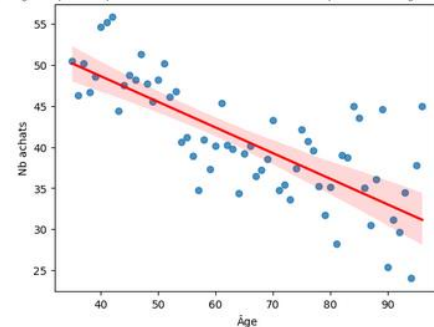
2

Répartition par âge et par fréquence d'achat des clients théoriques dont l'âge <= 32 ans (B to B exclus)



3

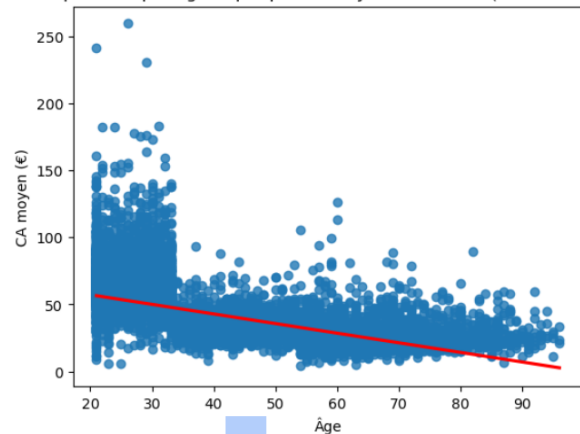
Répartition par âge et par fréquence d'achats des clients théoriques dont l'âge > 34 ans (B to B exclus)



4

6 – Analyses bivariées : Lien entre l'âge et le panier moyen des clients [1/2]

Répartition par âge et par panier moyen des clients (B to B exclus)



TEST DE SPEARMAN SUR LES VARIABLES 'Âge' et 'CA (€)'

H0 : Les variables 'Âge' et 'CA (€)' NE SONT PAS corrélées

H1 : Les variables 'Âge' et 'CA (€)' SONT corrélées

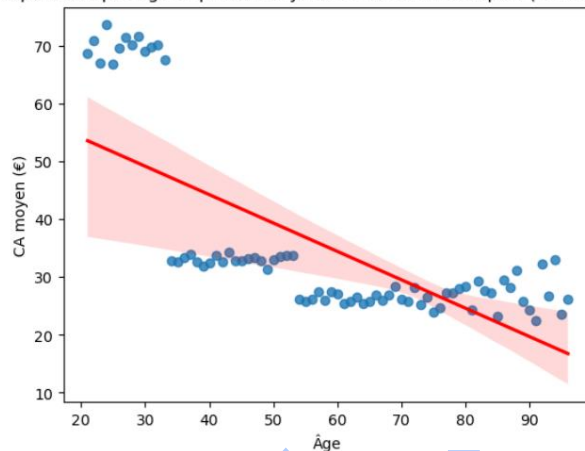
Coefficient de corrélation : -0.184538

Valeur P : 0.0

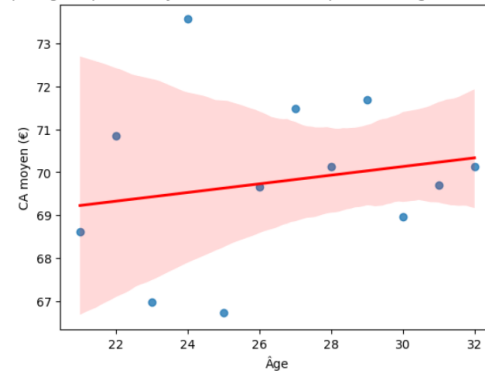
Analyse du Résultat : p_value <= 0.05 (on rejette H0)

Conclusion : OUI, les variables 'Âge' et 'CA (€)' SONT corrélées

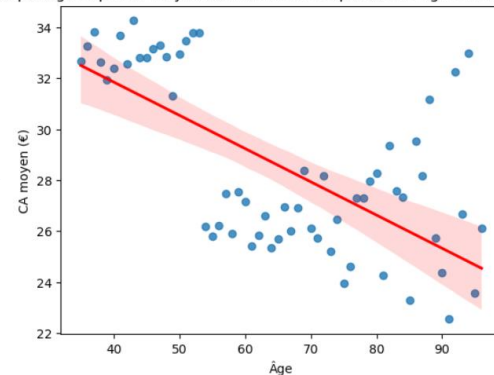
Répartition par âge et panier moyen des clients théoriques (B to B exclus)



Répartition par âge et panier moyen des clients théoriques dont l'âge <= 32 ans (B to B exclus)

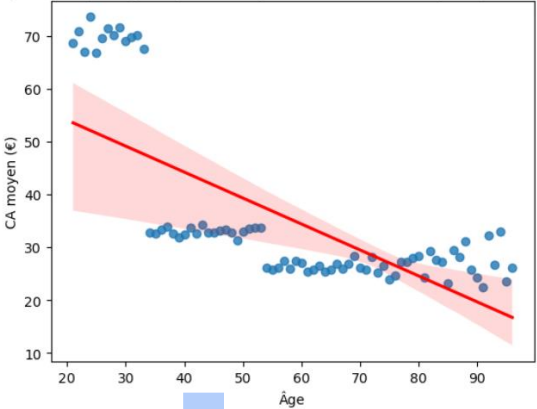


Répartition par âge et panier moyen des clients théoriques dont l'âge > 34 ans (B to B exclus)



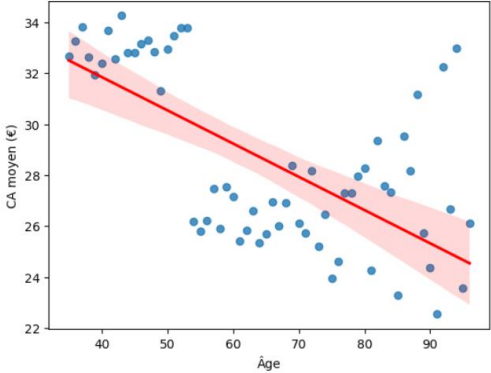
6 – Analyses bivariées : Lien entre l'âge et le panier moyen des clients [2/2]

Répartition par âge et panier moyen des clients théoriques (B to B exclus)



Clients <= 50 ans

Répartition par âge et panier moyen des clients théoriques dont l'âge > 34 ans (B to B exclus)



OLS Regression Results					
Dep. Variable:	CA moyen (€)		R-squared:	0.387	
Model:	OLS		Adj. R-squared:	0.377	
Method:	Least Squares		F-statistic:	37.94	
Date:	Fri, 28 Feb 2025		Prob (F-statistic):	6.64e-08	
Time:	15:46:17		Log-Likelihood:	-149.09	
No. Observations:	62		AIC:	302.2	
Df Residuals:	60		BIC:	306.4	
Df Model:	1				
Covariance Type:	nonrobust				
	coef.	std. err.	t	P> t	[0.025 0.975]
Âge	-0.1191	0.019	-6.159	0.000	-0.158 -0.080
Intercept	36.4604	1.313	27.777	0.000	33.835 39.086
Omnibus:	2.718		Durbin-Watson:	1.202	
Prob(Omnibus):	0.257		Jarque-Bera (JB):	2.418	
Skew:	0.482		Prob(JB):	0.299	
Kurtosis:	2.911		Cond. No.	258.	

Résidus Standardisés	
0	1.078599
1	-1.573674
2	7.735062
3	1.275099
4	6.742837
5	-3.085563
6	-1.372132
7	-0.120981
8	5.199746
9	2.084011
10	3.296229
11	-3.059216
12	0.879432
13	0.1017250
14	2.666257
15	-2.560855
16	1.358541
17	0.900148
18	0.129932
19	0.005255
20	-2.810388
21	-3.573997
22	-1.192512
23	-2.569159
24	0.278555
25	-2.269505
26	-2.004004
27	0.122723
28	-1.452483

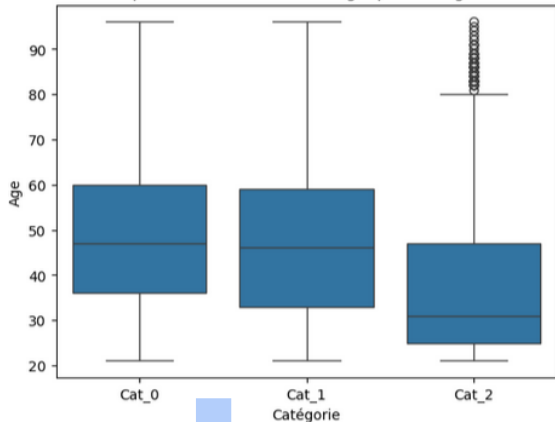
TEST DE SHAPIRO SUR LA VARIABLE 'Résidus Standardisés'	
H0 :	La répartition de la variable 'Résidus Standardisés' SUIT une loi normale
H1 :	La répartition de la variable 'Résidus Standardisés' NE SUIT PAS une loi normale
W :	0.951766
P value :	0.016243
Analyse du Résultat :	W > 0.95 ET p_value <= 0.05 (on rejette H0)
Conclusion :	NON, La répartition de la variable 'Résidus Standardisés' NE SUIT PAS une loi normale

Âge	35	40	45	50	55	60	65	70	75	80	85	90	95	100
Prévision de CA moyen (€)	32 €	32 €	31 €	31 €	30 €	29 €	29 €	28 €	28 €	27 €	26 €	26 €	25 €	25 €
Intervalle de confiance (€)	+/- 1 €	+/- 1 €	+/- 1 €	+/- 0 €	+/- 0 €	+/- 0 €	+/- 0 €	+/- 0 €	+/- 0 €	+/- 0 €	+/- 1 €	+/- 1 €	+/- 1 €	+/- 1 €

6 – Analyses bivariées : Lien entre la catégorie des produits achetés et l'âge[1/2]

Vérification des pré requis à l'applicabilité du test ANOVA

Répartition de la variable 'Âge' par 'Catégorie'



1

Catégorie "cat_0"

TEST DE SHAPIRO SUR LA VARIABLE 'Âge'

H0 : La répartition de la variable 'Âge' SUIIT une loi normale

H1 : La répartition de la variable 'Âge' NE SUIIT PAS une loi normale

W : 0.979668

P value : 0.0

Analyse du Résultat : W > 0.95 ET p_value <= 0.05 (on rejette H0)

Conclusion : NON, La répartition de la variable 'Âge' NE SUIIT PAS une loi normale

Catégorie "cat_1"

TEST DE SHAPIRO SUR LA VARIABLE 'Âge'

H0 : La répartition de la variable 'Âge' SUIIT une loi normale

H1 : La répartition de la variable 'Âge' NE SUIIT PAS une loi normale

W : 0.859389

P value : 0.0

Analyse du Résultat : W <= 0.95 ET p_value <= 0.05 (on rejette H0)

Conclusion : NON, La répartition de la variable 'Âge' NE SUIIT PAS une loi normale

Catégorie "cat_2"

TEST DE SHAPIRO SUR LA VARIABLE 'Âge'

H0 : La répartition de la variable 'Âge' SUIIT une loi normale

H1 : La répartition de la variable 'Âge' NE SUIIT PAS une loi normale

W : 0.970706

P value : 0.0

Analyse du Résultat : W > 0.95 ET p_value <= 0.05 (on rejette H0)

Conclusion : NON, La répartition de la variable 'Âge' NE SUIIT PAS une loi normale

2

TEST DE LEVENE SUR LA VARIABLE 'Âge' entre les 3 sous groupes 'Cat_0', 'Cat_1', et 'Cat_2'

H0 : Les variances de la variable 'Âge' sont équivalentes entre les 3 sous-groupes 'Cat_0', 'Cat_1', et 'Cat_2'

H1 : Au moins une des variances de la variable 'Âge' s'écarte des deux autres entre les 3 sous groupes 'Cat_0', 'Cat_1', et 'Cat_2'

Statistique de Levene : 41.220209

Valeur P : 0.0

Analyse du Résultat : p_value <= 0.05 (on rejette H0)

Conclusion : NON, au moins une des variances de la variable 'Âge' s'écarte des deux autres parmi les trois sous groupes 'Cat_0', 'Cat_1' et 'Cat_2'

3

Le test ANOVA n'est pas applicable

2

Effectuer un test non paramétrique de Kruskal-Wallis

4

6 – Analyses bivariées : Lien entre la catégorie des produits achetés et l'âge[2/2]

TEST DE KRUSKAL-WALLIS SUR LA VARIABLE 'Âge' entre les 3 sous groupes 'Cat_0', 'Cat_1', et 'Cat_2'

H0 : Les variances de la variable 'Âge' sont équivalentes entre les 3 sous-groupes 'Cat_0', 'Cat_1', et 'Cat_2'

H1 : Au moins une des variances de la variable 'Âge' s'écarte des deux autres entre les 3 sous groupes 'Cat_0', 'Cat_1', et 'Cat_2'

H de kruskal : 1251.55381

Valeur P : 0.0

Analyse du Résultat : p_value <= 0.05 (on rejette H0)

Conclusion : OUI, il y a une corrélation de la variable 'Âge' entre au moins deux sous-groupes parmi les trois 'Cat_0', 'Cat_1', et 'Cat_2' (au moins une des médianes s'écarte des deux autres)

2

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Cat_0	Cat_1	-1.4778	0.0	-2.0868	-0.8687	True
Cat_0	Cat_2	-10.9895	0.0	-11.7726	-10.2063	True
Cat_1	Cat_2	-9.5117	0.0	-10.2866	-8.7368	True

1

TEST DE MANN-WHITNEY SUR LA VARIABLE 'Âge' entre les 2 sous groupes 'Cat_0' ET 'Cat_2'

H0 : Aucune corrélation de la variable 'Âge' entre les 2 sous-groupes 'Cat_0' ET 'Cat_2' (les 2 médianes sont équivalentes)

H1 : Il y a une corrélation de la variable 'Âge' entre les deux sous-groupes 'Cat_0' ET 'Cat_2' (les deux médianes s'écartent de manière significative)

U de mann_whitney : 19834448.0

Valeur P : 0.0

Analyse du Résultat : p_value <= 0.05 (on rejette H0)

Conclusion : OUI, il y a une corrélation de la variable 'Âge' entre les deux sous-groupes 'Cat_0' ET 'Cat_2' (les deux médianes s'écartent de manière significative)

TEST DE MANN-WHITNEY SUR LA VARIABLE 'Âge' entre les 2 sous groupes 'Cat_0' ET 'Cat_1'

H0 : Aucune corrélation de la variable 'Âge' entre les 2 sous-groupes 'Cat_0' ET 'Cat_1' (les 2 médianes sont équivalentes)

H1 : Il y a une corrélation de la variable 'Âge' entre les deux sous-groupes 'Cat_0' ET 'Cat_1' (les deux médianes s'écartent de manière significative)

U de mann_whitney : 35158590.0

Valeur P : 0.0

Analyse du Résultat : p_value <= 0.05 (on rejette H0)

Conclusion : OUI, il y a une corrélation de la variable 'Âge' entre les deux sous-groupes 'Cat_0' ET 'Cat_1' (les deux médianes s'écartent de manière significative)

TEST DE MANN-WHITNEY SUR LA VARIABLE 'Âge' entre les 2 sous groupes 'Cat_0' ET 'Cat_2'

H0 : Aucune corrélation de la variable 'Âge' entre les 2 sous-groupes 'Cat_0' ET 'Cat_2' (les 2 médianes sont équivalentes)

H1 : Il y a une corrélation de la variable 'Âge' entre les deux sous-groupes 'Cat_0' ET 'Cat_2' (les deux médianes s'écartent de manière significative)

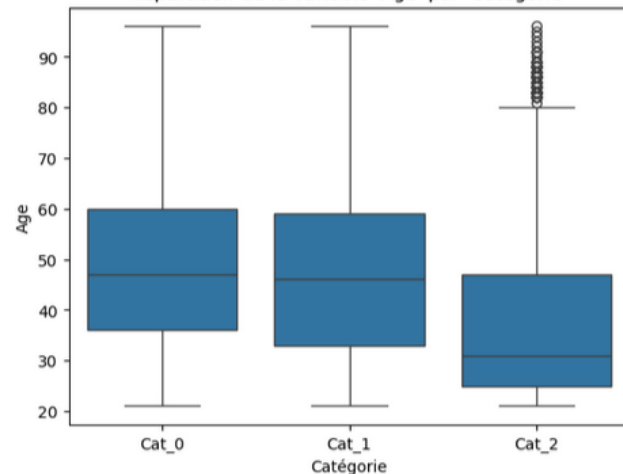
U de mann_whitney : 19834448.0

Valeur P : 0.0

Analyse du Résultat : p_value <= 0.05 (on rejette H0)

Conclusion : OUI, il y a une corrélation de la variable 'Âge' entre les deux sous-groupes 'Cat_0' ET 'Cat_2' (les deux médianes s'écartent de manière significative)

Répartition de la variable 'Âge' par 'Catégorie'



5 - CONCLUSIONS DE L'ÉTUDE

Conclusions et limites de l'étude

Conclusions

- ✓ Les ventes des 12 derniers mois (mars 2022 à février 2023) ont stagné - voire légèrement diminué - par rapport aux 12 mois précédents (mars 2021 à février 2022)
 - ✓ Parmi les livres qui se sont le mieux vendus, une majorité de livres de catégorie 2
 - ✓ Il y a pour les clients particuliers uniquement une sous-représentation du niveau de vente des livres de catégorie 2
 - ✓ Il y a un lien entre l'âge et la catégorie des livres achetés. Les clients jeunes (≤ 30 ans) se concentrent sur la catégorie 2, les clients senior (≥ 50 ans) sur les catégories 0 et 1. Par déduction les clients intermédiaires (30 à 50 ans) achètent indifféremment des livres dans les 3 catégories
- ➔ Un des moyens de juguler la tendance à la baisse des ventes, voire de l'augmenter, serait de définir un parcours client dédié pour la cible ≤ 30 ans (un parcours « jeune ») tout en étoffant et concentrant pour cette cible l'offre de livres de catégorie 2

Limites de l'étude

- ✓ L'abstraction des données pénalise la lisibilité des interprétations, notamment sur les orientations stratégiques à donner. Quand on parle de catégorie 2, celle qui est achetée par la cible des clients de moins de 30 ans, de quoi parle-t-on exactement ? Mangas, romans, culture ?
- ✓ Par manque de temps, certains renoncements ont été appliqués à l'étude, mais sans impact significatif sur les conclusions. Notamment le V de Cramer pour affirmer l'absence de lien entre le genre et la catégorie des livres achetés


ANNEXES

Annexe 1 : protocole d'analyses bivariées

Protocole des tests statistiques d'analyses bivariées

	Variable quantitative	Variable qualitative
Variable quantitative	<p>Effectuer un test de shapiro-wilk sur chacune des deux variables</p> <p>Si les deux variables suivent une loi normale</p> <p>Alors effectuer un test de Pearson (paramétrique)</p> <p>Sinon effectuer un test de Stearman (non paramétrique)</p> <p>En amont, établir le nuage de points entre les deux variables le heatmap associé</p>	<p>Effectuer un test de shapiro-wilk pour chacune des modalités de la variable qualitative</p> <p>Si Variable_Quali.Nb(modalités) <= 2</p> <p>Alors</p> <p>Si répartition normale pour chacune des modalités</p> <p>Alors effectuer un test T (paramétrique)</p> <p>Sinon effectuer un test de Mann-Whitney (non paramétrique)</p> <p>Si Variable_Quali.Nb(modalités) > 2</p> <p>Alors</p> <p>Si répartition normale pour chacune des modalités</p> <p>Alors effectuer un test ANOVA (paramétrique)</p> <p>Sinon effectuer un test de Kruskal-Wallis (non paramétrique)</p>
Variable qualitative	<p>Effectuer un test de shapiro-wilk pour chacune des modalités de la variable qualitative</p> <p>Si Variable_Quali.Nb(modalités) <= 2</p> <p>Alors</p> <p>Si répartition normale pour chacune des modalités</p> <p>Alors effectuer un test T (paramétrique)</p> <p>Sinon effectuer un test de Mann-Whitney (non paramétrique)</p> <p>Si Variable_Quali.Nb(modalités) > 2</p> <p>Alors</p> <p>Si répartition normale pour chacune des modalités</p> <p>Alors effectuer un test ANOVA (paramétrique)</p> <p>Sinon effectuer un test de Kruskal-Wallis (non paramétrique)</p>	<p>Test de shapiro-wilk non applicable !!!</p> <p>Tests de vérification de dépendance : Fischer ou Chi-2</p> <p>Si fréquence théorique des deux variables <= 5 (si matrice TCD <= 2x2)</p> <p>Alors effectuer un test de Fischer</p> <p>Sinon effectuer un test du Chi-2 -- matrice TCD > 2x2</p> <p>A compléter avec un histogramme groupé et un heatmap pour vulgariser le résultat</p>

6 – Analyses bivariées : lien quali x quali et V de Cramer

$0,50 \leq \text{Cramer's V} = 0,534 < 0,70$ La relation entre nos deux variables est donc forte. 

$0,70 \leq \text{Cramer's V ou Phi} \leq 1$	Relation très forte
$0,50 \leq \text{Cramer's V ou Phi} < 0,70$	Relation forte
$0,30 \leq \text{Cramer's V ou Phi} < 0,50$	Relation intéressante
$0,10 \leq \text{Cramer's V ou Phi} < 0,30$	Relation faible
$0,00 \leq \text{Cramer's V ou Phi} < 0,10$	Relation négligeable

FIG. 4.5 – Schéma de Davis : interprétation du Cramer's V et du Phi