BILAN COMMERCIAL 2021-2023:

ANALYSE DES VENTES DE LA LIBRAIRIE LAPAGE

*Fatiha **AKDIM** – Service Marketing – Librairie en ligne **Lapage***

AU PROGRAMME

1. CONTEXTE COMMERCIAL, OBJECTIFS ET MOYENS
2. PRESENTATION ET NETTOYAGE DES FICHIERS
3. ANALYSE DES INDICATEURS DE VENTE
4. ANALYSE DES PRODUITS ET DES CATEGORIES
5. ANALYSE DU PROFIL DES CLIENTS
6. LES TESTS STATISTIQUES
7. SYNTHESE ET SUGGESTIONS

1.CONTEXTE COMMERCIAL OBJECTIFS ET MOYENS

1.1. Le Contexte commercial:

Face au succès et à l'engouement des clients pour les produits commercialisés par la librairie « **Lapage** » ,
la société a fait le choix depuis 2 ans de compléter ses points de vente physique par la commercialisation en ligne de ses produits depuis 2 ans .

1.2. Les objectifs de l'analyse:

Après 2 ans d'exercice l'entreprise souhaite faire un bilan de ses points forts et ses points faible.

Ma mission a donc consisté à :

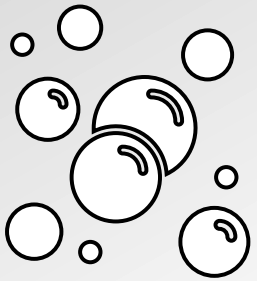
- Analyser l'évolution et la structure du Chiffre d'affaires
- Analyser la structure des produits
- Analyser le profil des clients
- Analyser les potentiels liens existantes entre les comportements des clients et les indicateurs de ventes/achats

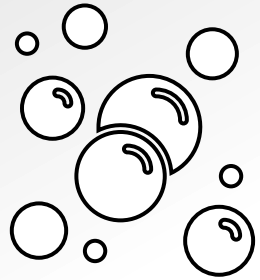
1.3. Les moyens:

Mise à disposition de 3 fichiers csv:

- Customers
- Products
- Transactions

2. OBSERVATIONS ET NETTOYAGE DES FICHIERS





2.1. Présentation des fichiers

- Fichier Customers :

	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

- Fichier Product:

	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

- Fichier Transaction :

	id_prod	date	session_id	client_id
0	0_1518	2022-05-20 13:21:29.043970	s_211425	c_103
1	1_251	2022-02-02 07:55:19.149409	s_158752	c_8534
2	0_1277	2022-06-18 15:44:33.155329	s_225667	c_6714
3	2_209	2021-06-24 04:19:29.835891	s_52962	c_6941
4	0_1509	2023-01-11 08:22:08.194479	s_325227	c_4232

2.2. Les anomalies détectées

- Présence d'une valeur négative dans la colonne prix
- Fichier concerné: **PRODUCT**
- Choix: Suppression de la vigne concernée

	id_prod	price	categ
731	T_0	-1.0	0

- Présence de clients tests ct_1 ct_0
- Fichier concerné: **CUSTOMERS**
- Choix: suppression des lignes

- Présence de 126 doublons
- Fichier concerné: **TRANSACTIONS**
- Choix: Suppression des lignes T_0 qui sont des clients tests et qui génèrent des doublons . 200 Lignes concernées par les T_0

<u>Suppression des lignes</u>:
- T_0
- CT_1, CT_0

- Format date exprimé en objet au lieu de Date
- Fichier concerné: **TRANSACTIONS**
- Choix: modification du format au profit du format AAAA-MM-JJ et suppression de l'heure

	id_prod	date	session_id	client_id
0	0_1518	2022-05-20 13:21:29.043970	s_211425	c_103

	id_prod	date	session_id	client_id
0	0_1518	2022-05-20	s_211425	c_103

- Présence d'une valeur dans le fichier transaction non référencée dans le fichier 'Product'.
- Fichier concerné: **PRODUCT**
- Choix: Ajout du produit dans le fichier 'Product' en lui attribuant une valeur et une catégorie

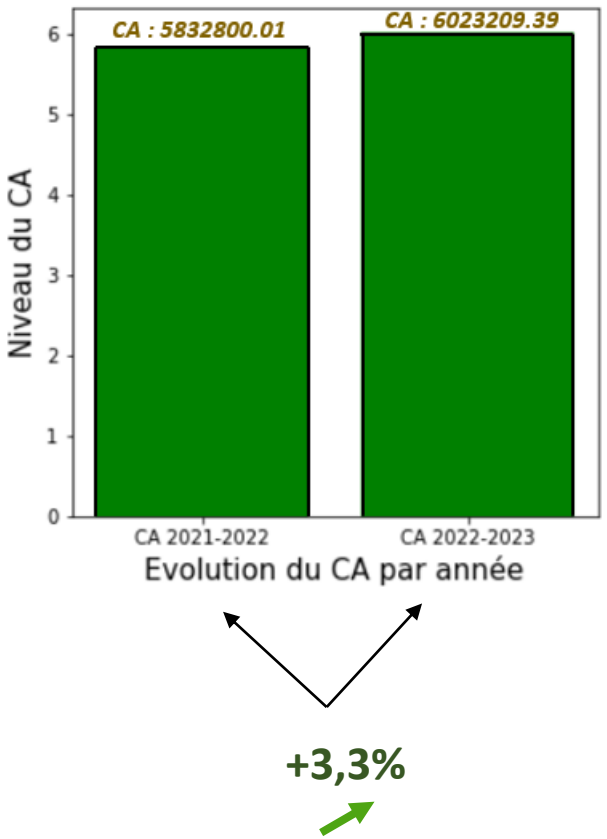
	id_prod	price	categ
0	0_1421	19.99	0
...	...	...	...
3287	0_2245	10.32	0

3287 rows × 3 columns

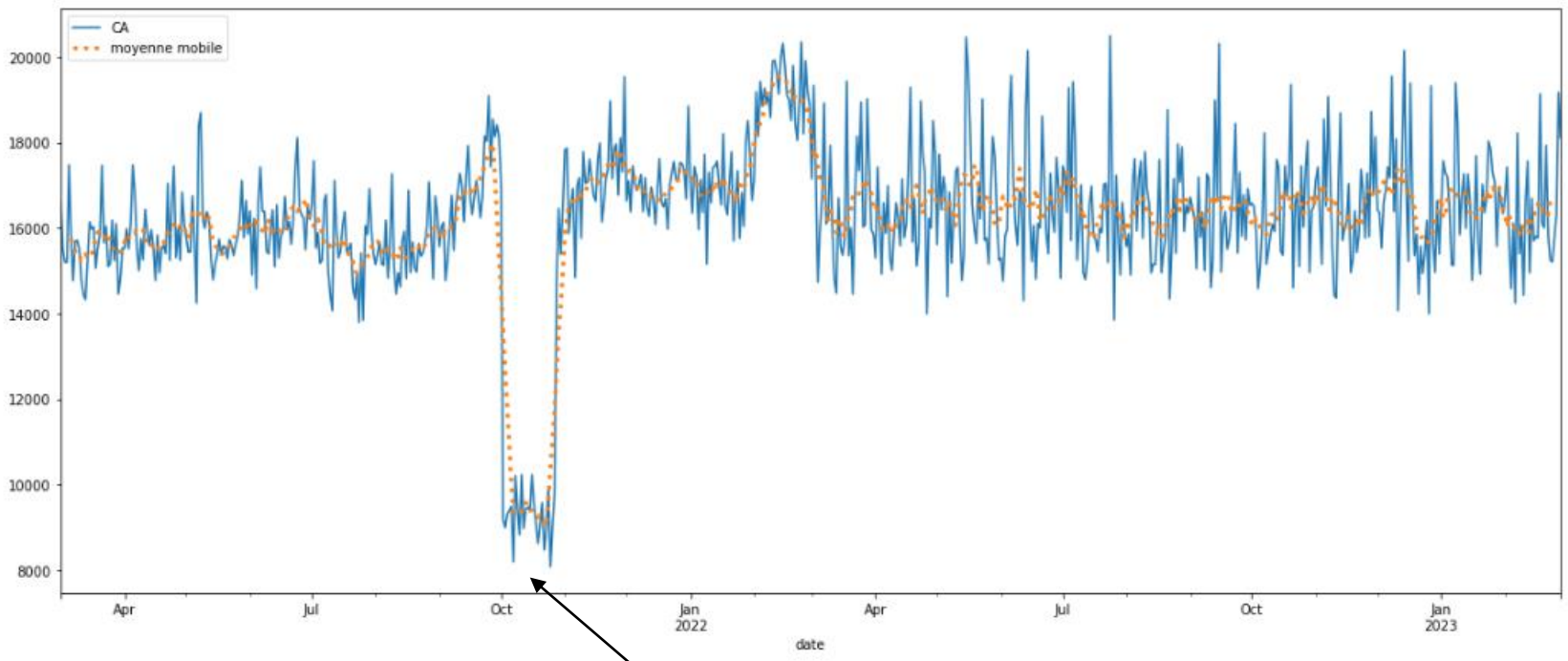
3 . ANALYSE DES INDICATEURS DE VENTE

3.1. Analyse de l'évolution du Chiffre d'affaires

- **Evolution du CA 2021-2023**



- **Evolution du CA 2021-2023 et moyenne mobile**



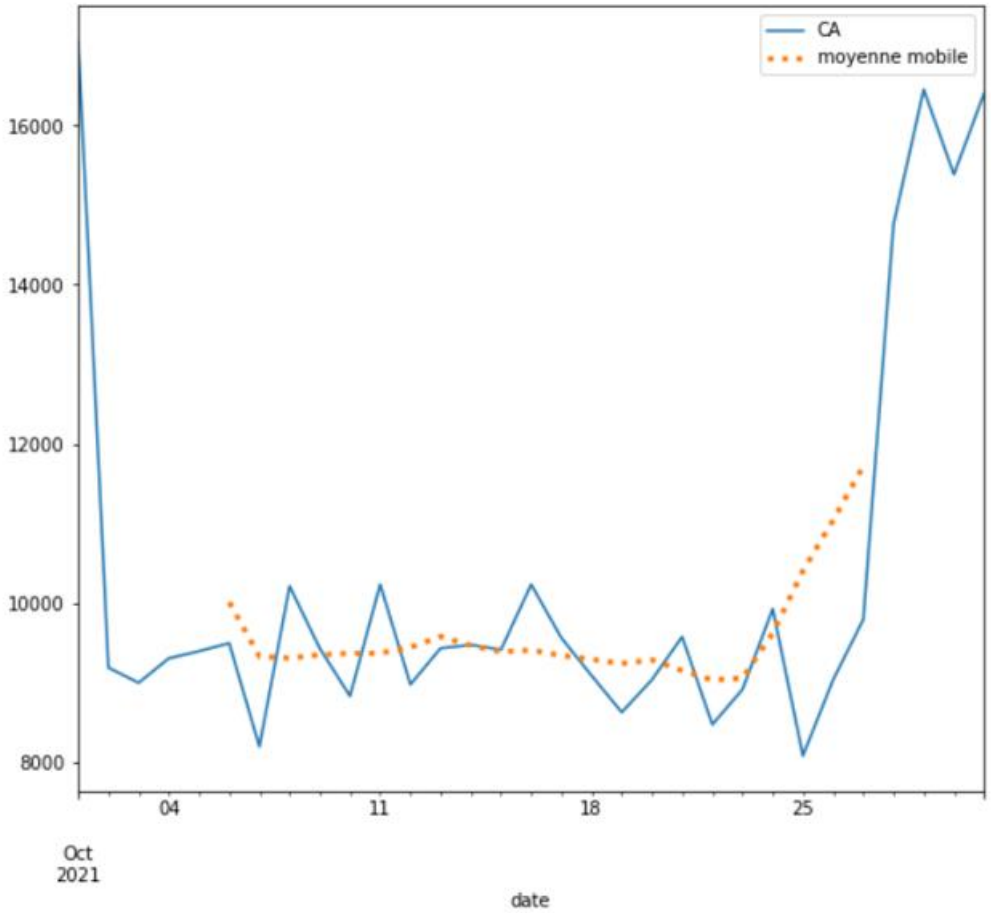
* `rolling(window=10, center=True)`

- Chiffre d'affaire total 2021-2023 : 11 856 009.4 euros

Observation d'une baisse anormal du chiffre d'affaire

3. Analyse de la baisse du Chiffre d'affaires d'octobre 2021

▪ GRAPHIQUE : CHIFFRE D'AFFAIRE OCT. 2021



- *Baisse du chiffre d'affaire du 02 au 25/10/2021*

▪ ETUDE DE LA REPARTITION DU CHIFFRE D'AFFAIRE PAR CATEGORIE

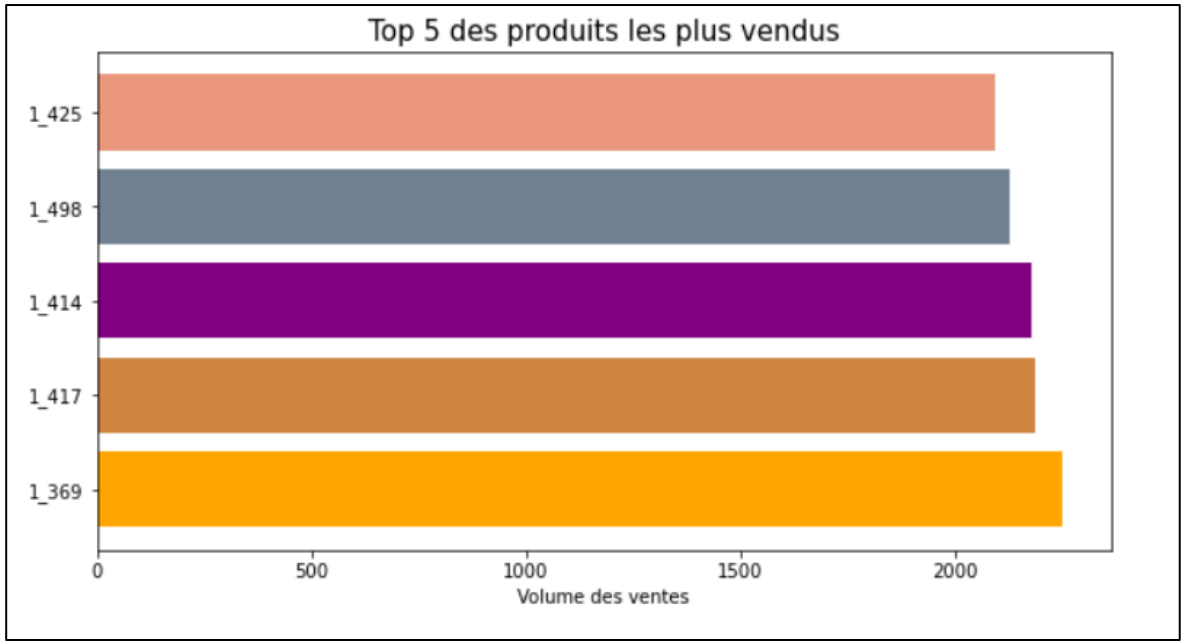
```
#répartition du CA pour le mois d'octobre
df_oct.rename(columns={'price': 'CA'}, inplace= True)
dfcateg=df_oct[['categ','CA']].groupby(['categ'], as_index=False).sum()
dfcateg
```

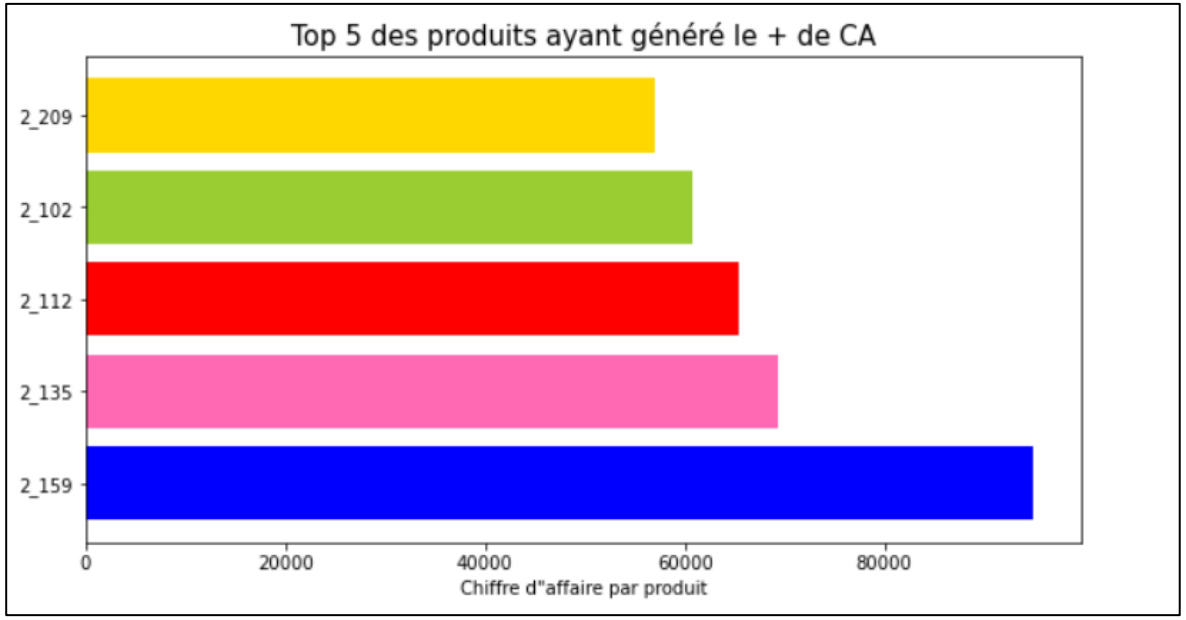
	categ	CA
0	0	156298.13
1	2	65689.92

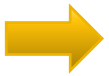
- **On constate une absence de CA sur la catégorie 1**
  (On peut supposer une rupture du stock sur cette catégorie ce qui pourrait expliquer la baisse de chiffre d'affaire pour cette période)

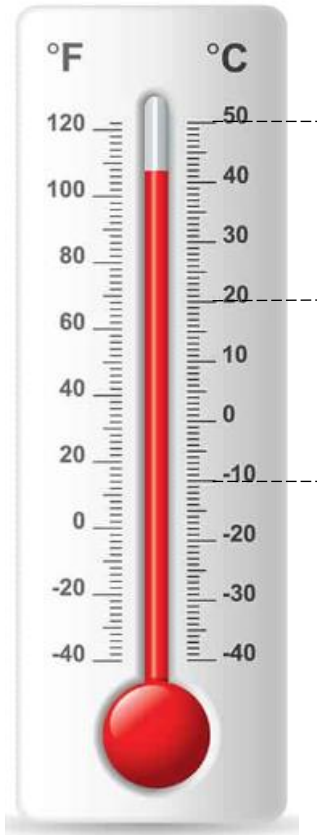
4. ANALYSE DES PRODUITS ET DES CATEGORIES

4.1. Analyse des produits







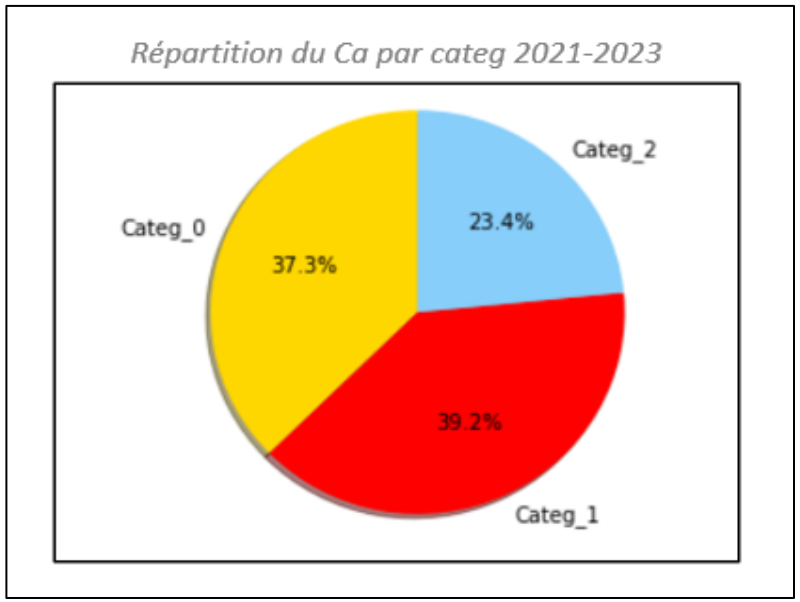


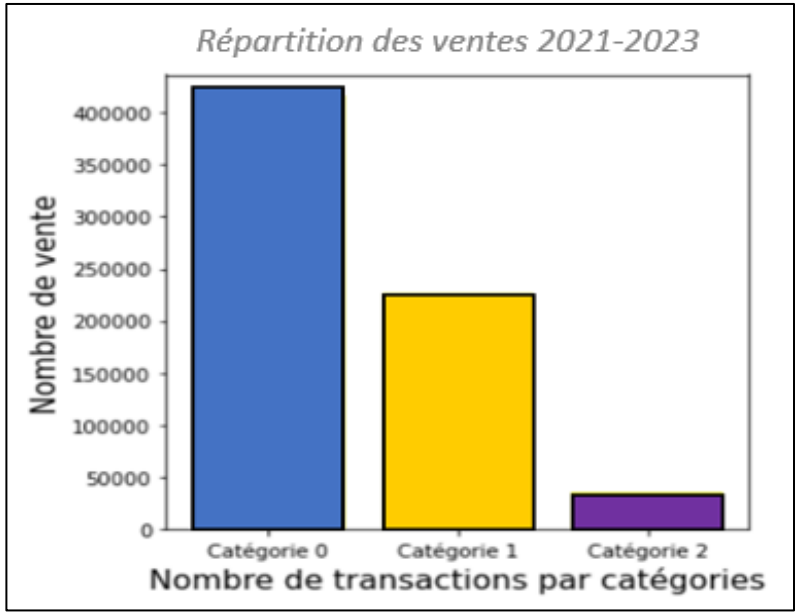
Le produit ayant généré le - de CA est le 0_1539

18 produits n'ont été vendus qu'une seule fois

20 produits invendus

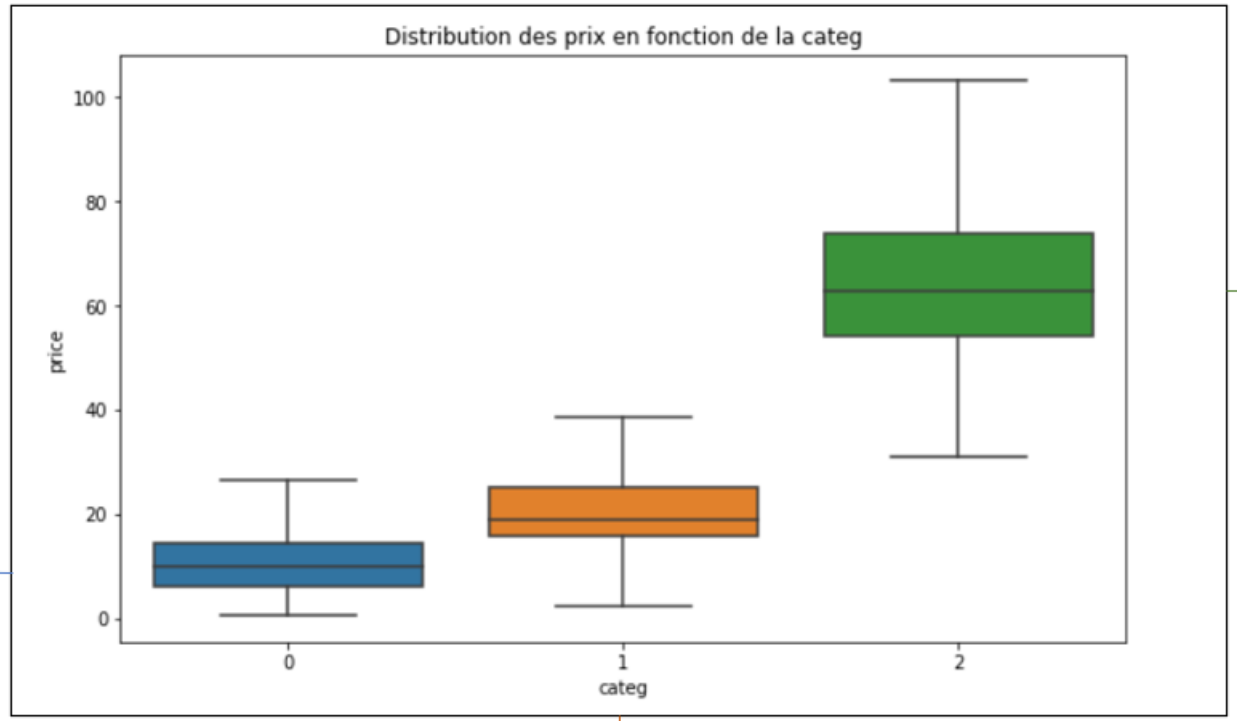
4.2. Analyse des Catégories





4.2. Top et Flop Catégories

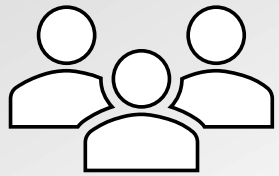
Catégorie 0 :
- Prix moyen : 10,64



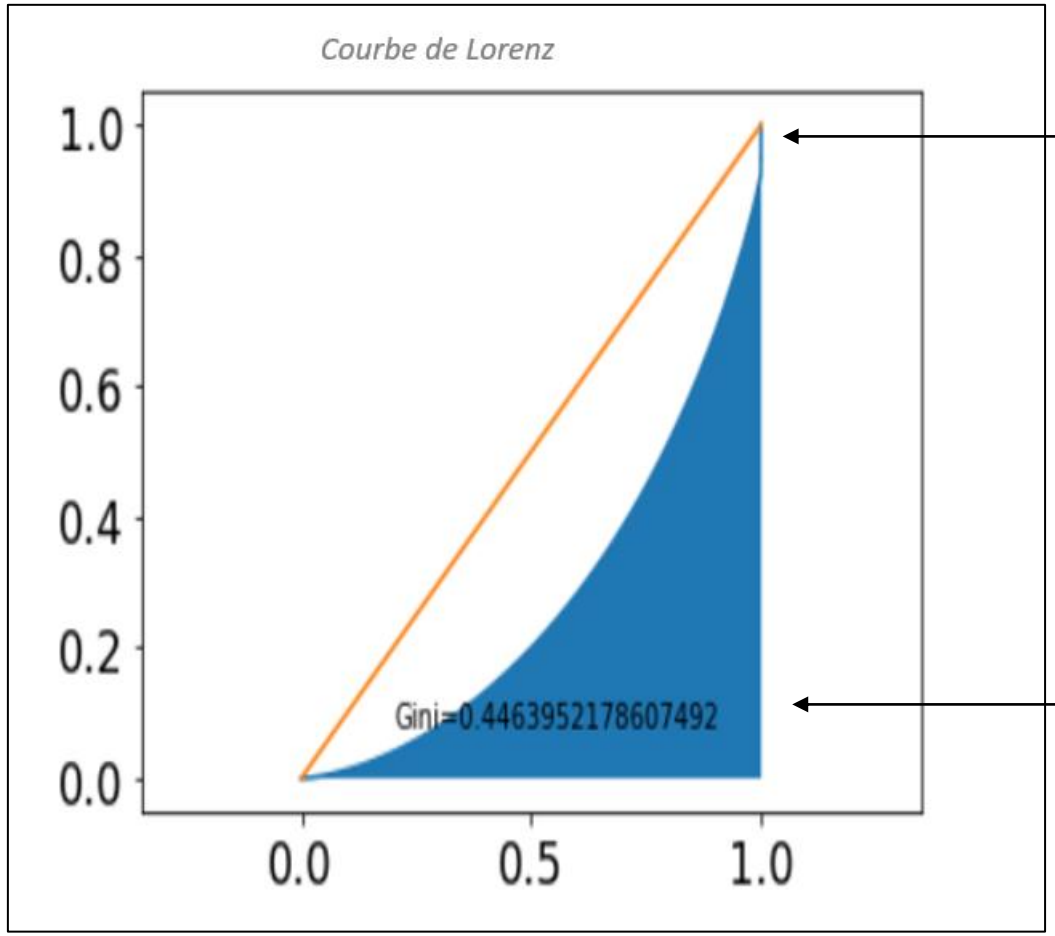
Catégorie 2 :
- Prix moyen : 76,21

Catégorie 1 :
- Prix moyen : 20,49

5. ANALYSE DU PROFIL DES CLIENTS



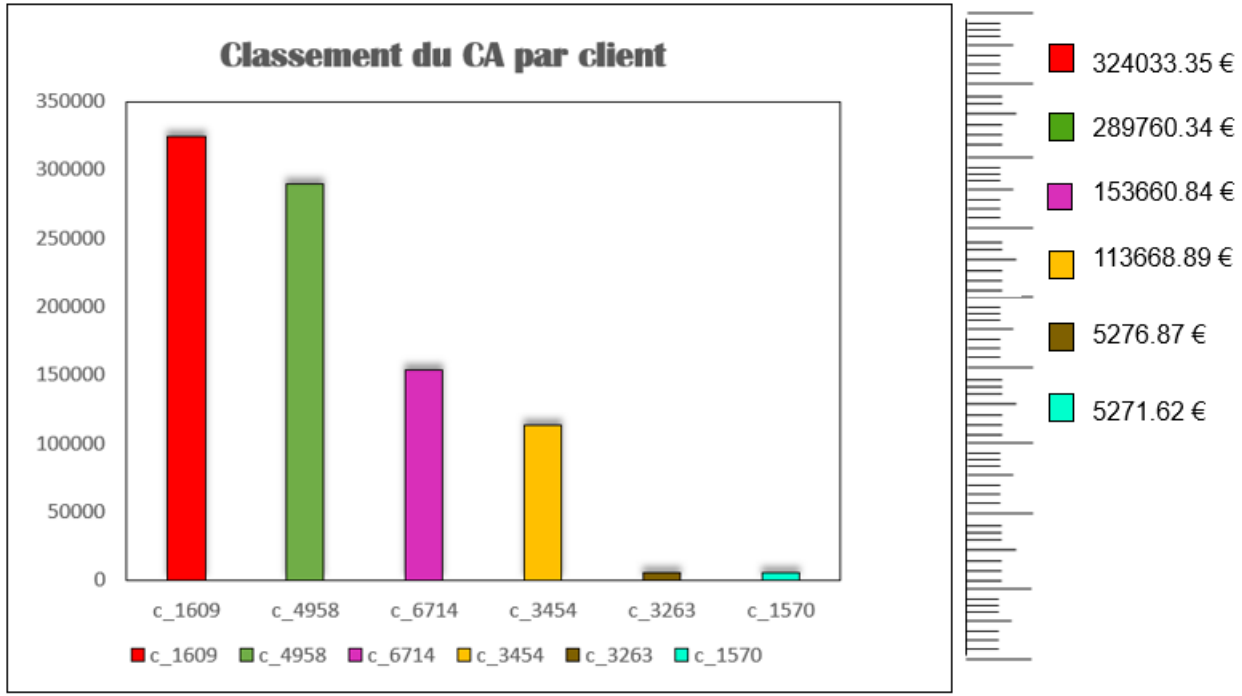
5.1. Inégalité de la distribution du Chiffre d'affaire

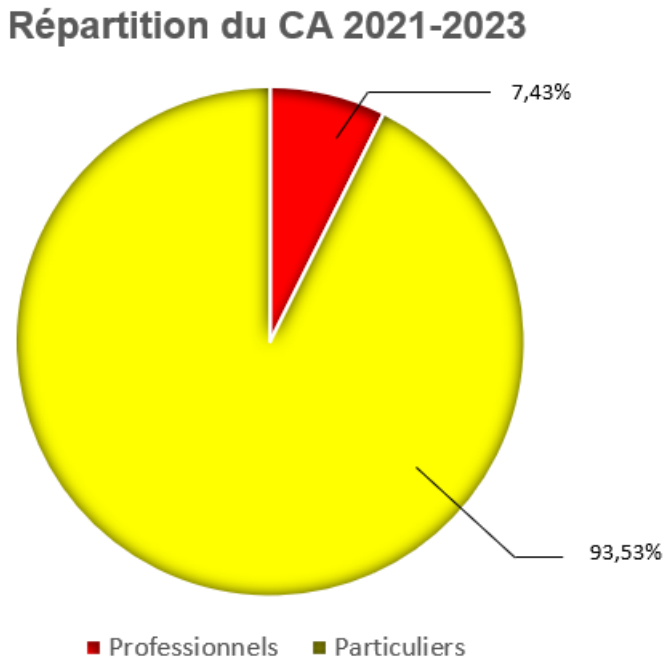


La droite d'Equidistribution
Représente une situation parfaitement égalitaire

Coef de Gini
Mesure l'inégalité de la distribution du CA

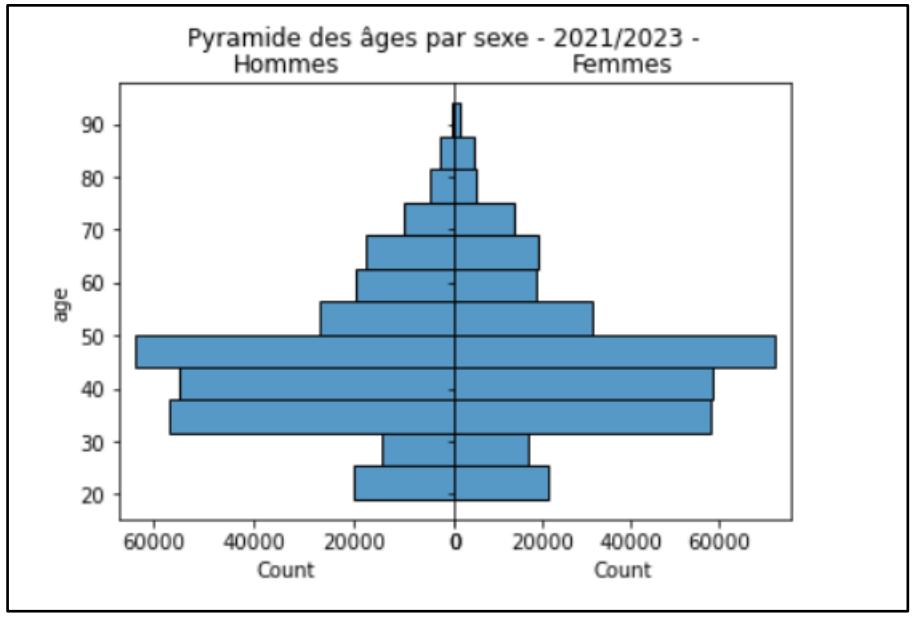
5.2. Analyse en fonction de la typologie des clients

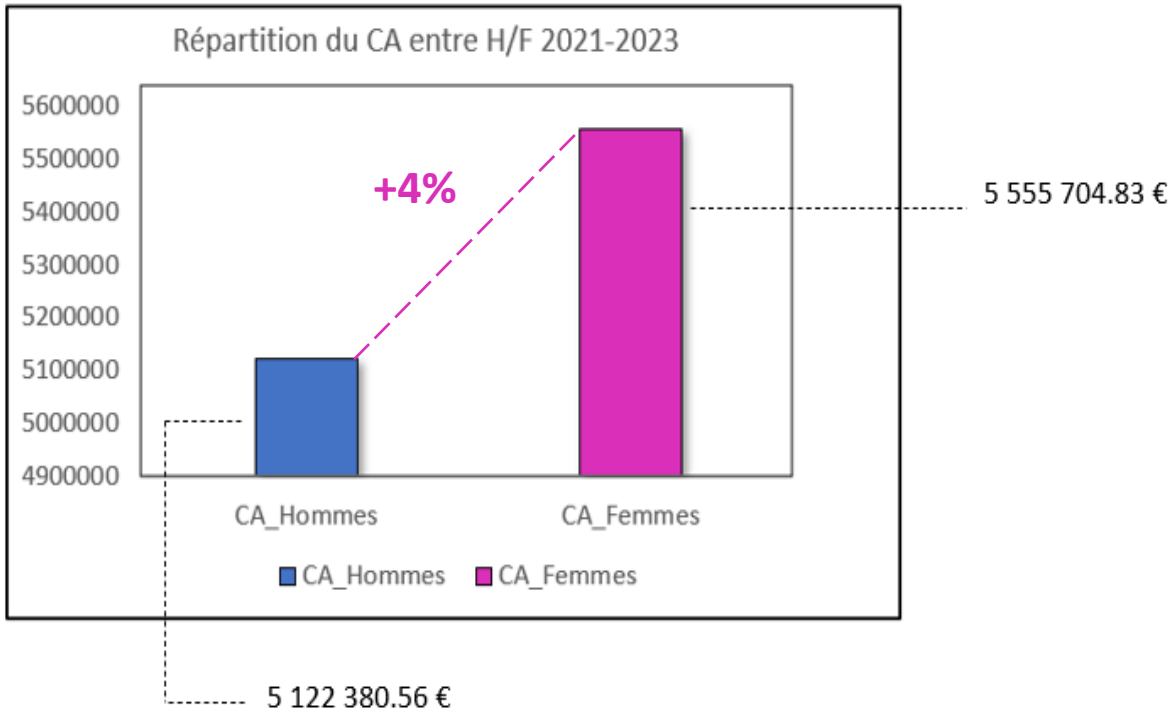




** Résultats commerciaux 2021-2023*

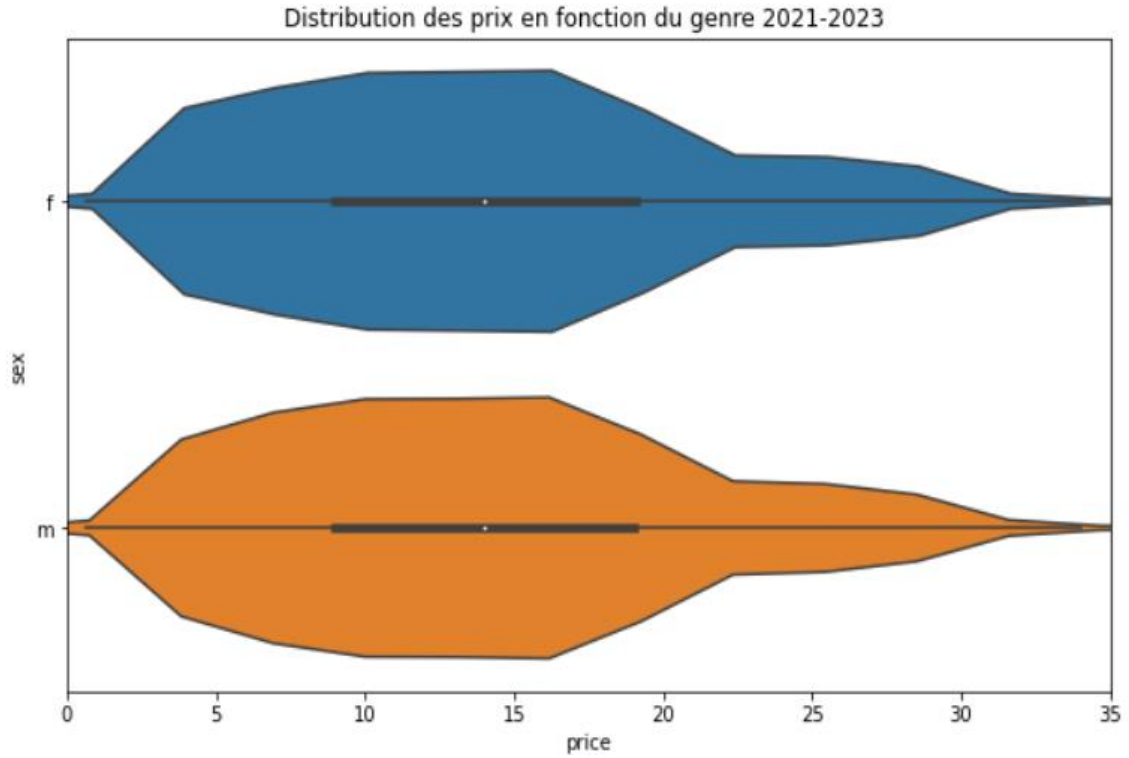
5.3. Analyse en fonction du genre



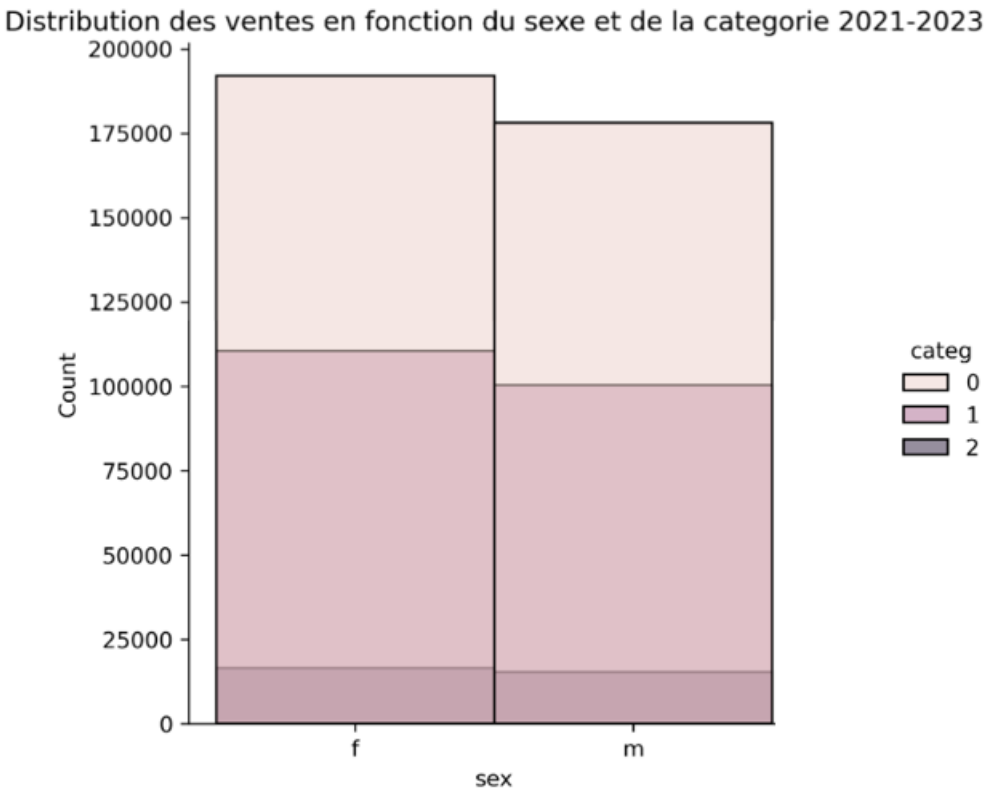


- *Equilibre H/F*
- *Tranche 30/50 ans dominante*
- *Tous les âges sont représentés*

- *Les femmes génèrent 4 % de Chiffre d'affaire en + par rapport aux hommes.*



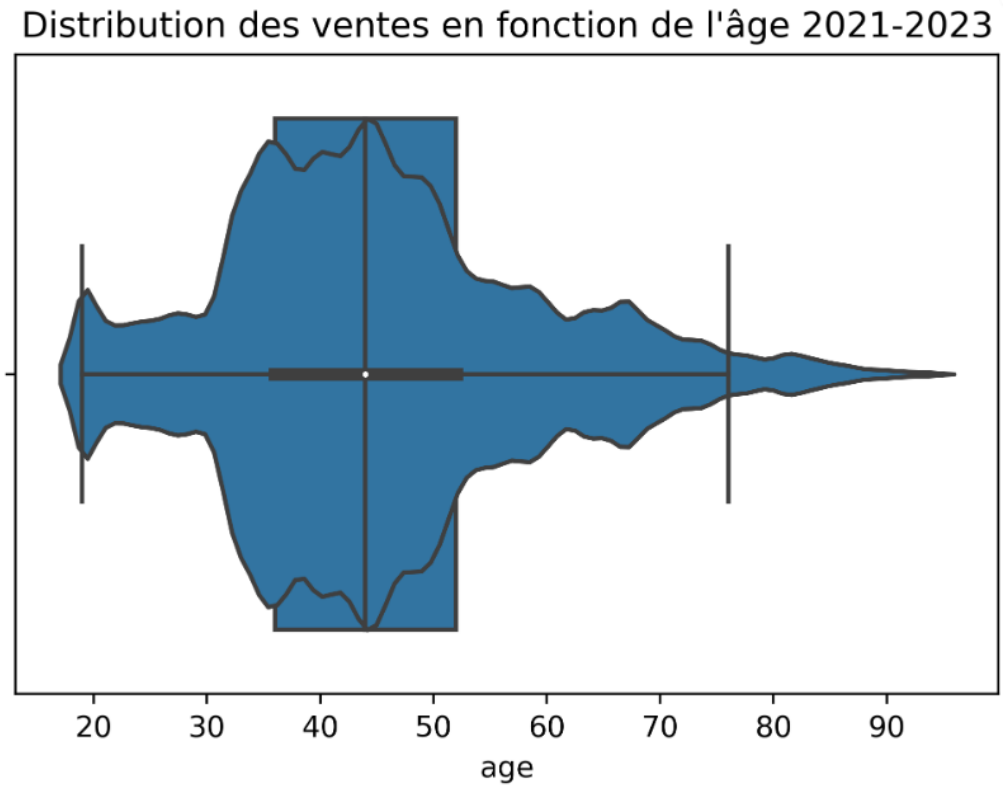
- *Distribution des prix en fonction du genre identique*

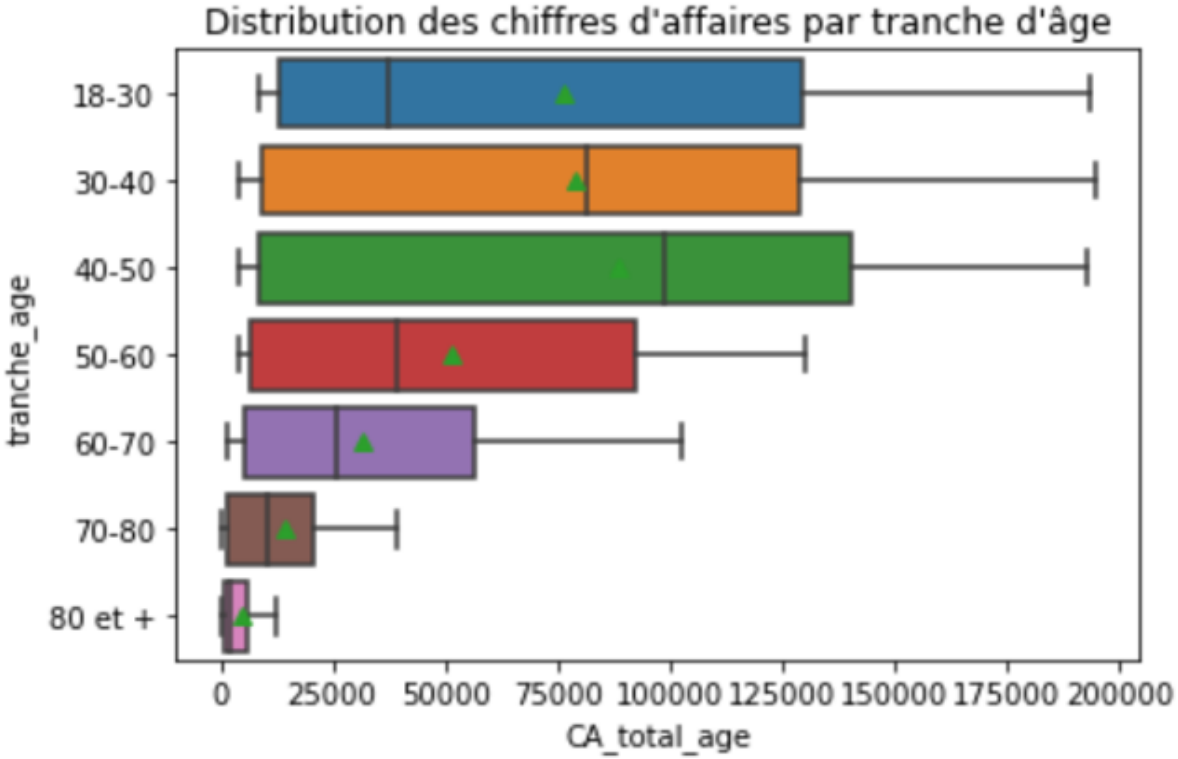


- *Categ n°1 : nombre d'achat réalisé par les femmes soit légèrement supérieur à celui des hommes*

- *Pour les categ 0 et 2 elles semblent a peu près identiques entre homme et femme.*

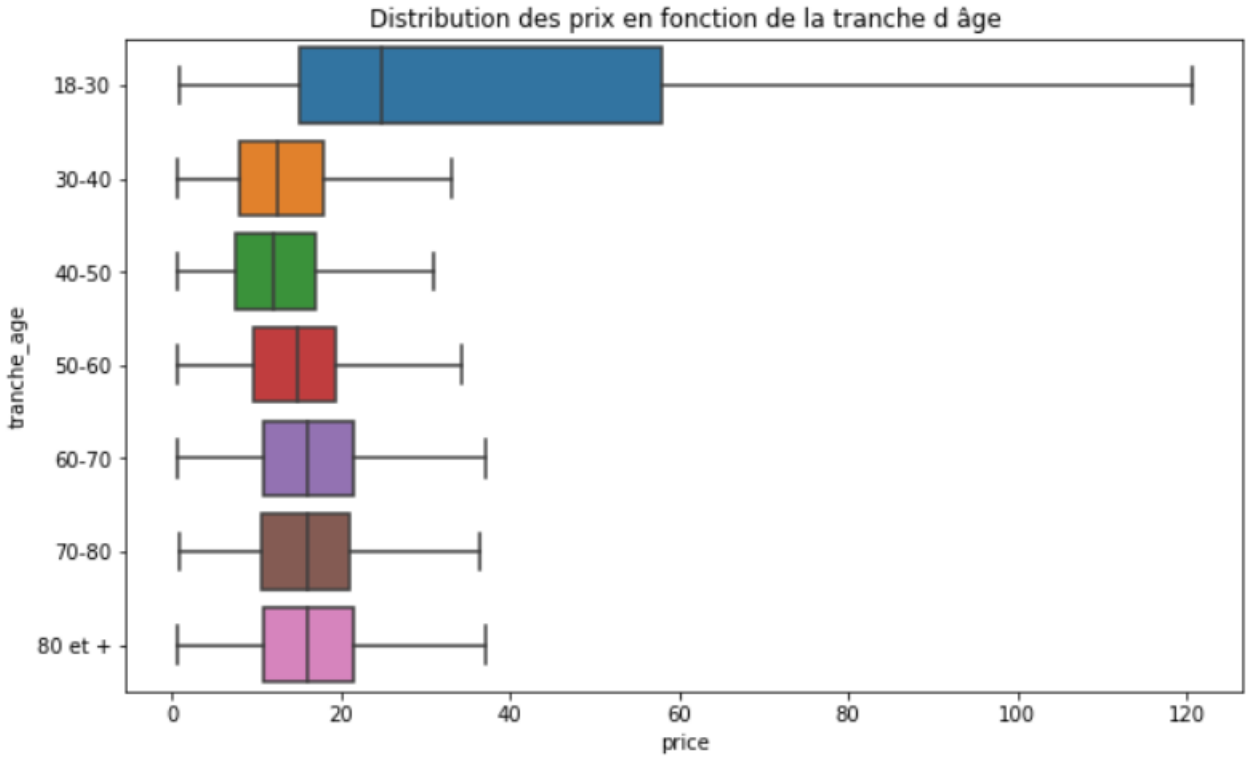
5.4. Analyse en fonction de l'âge

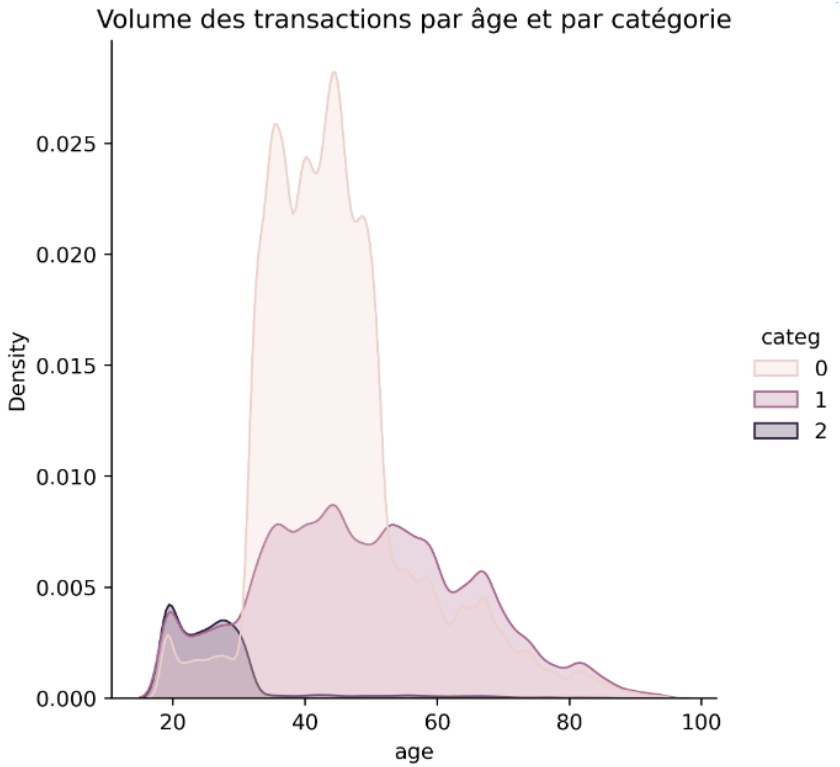




- *Les 35-55 ans réalisent plus de la moitié des achats*

- *Les 40-50 ans génèrent le + de CA*
- *Les 18-40 génèrent le même niveau de CA*





- *Les 18-30 ans achètent les produits ayant les prix les plus élevés.*
- *Equilibre des autres tranches d'âge*

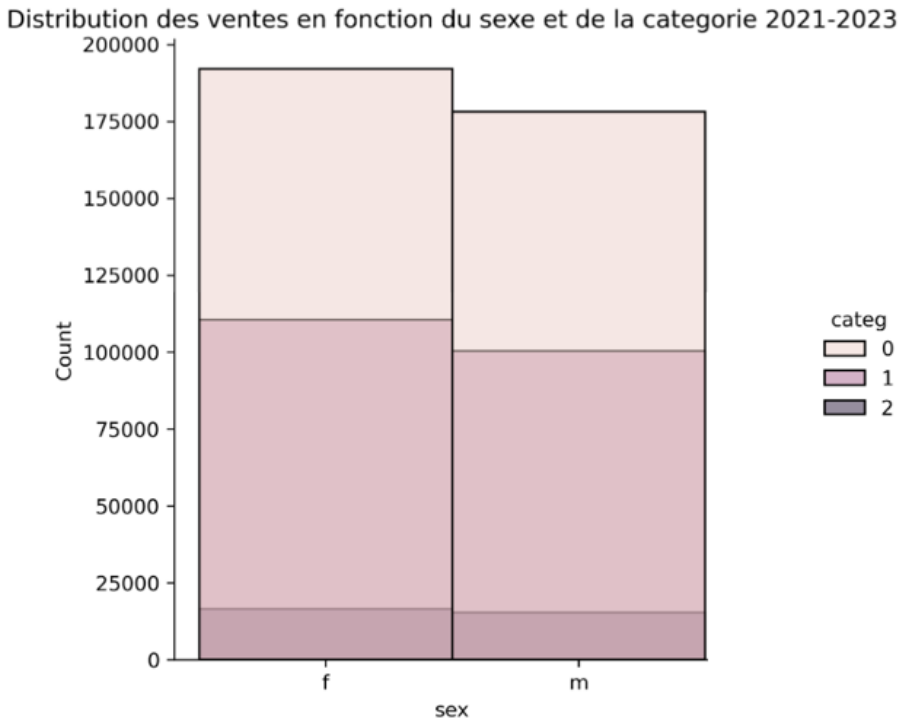
- *La categ 0 est davantage achetée par les 35-55 ans*
- *La categ 2 par les 18-35*
- *La categ 1 par tout le monde*
- *Les 60-80 achètent peu en volume*

6 . LES TESTS STATISTIQUES

6.1. Existe-t-il un lien entre le genre des clients et la catégorie de livre acheté?

6.1.1. Préparation du test

- **Observations:**



- *Categ n°1 : nombre d'achat réalisé par les femmes > à celui des hommes*
- *Categ 0 et 2 : identiques entre homme et femme*

- **Types de variables:** 2 variables qualitatives

- **Choix du test:** Test du khi-deux d'indépendance

- **Conditions du test:**

1) Avoir un échantillon aléatoire simple

2) Concerner 2 variables catégorielles

3) Avoir plus de cinq valeurs attendues pour chaque combinaison de variables

- **Hypothèses:**

H0 = les variables sont indépendantes
H1 = les variables sont dépendantes
On pose $\alpha = 0{,}05$ *(seuil de décision. Ici 5%)*

6.1.2. Test KHI2

- **La méthode:**

- Elle consiste à comparer 2 matrices , l'une reprenant l'ensemble des dénombrement de nos variables, l'autre reprenant les démembrements attendues si les variables étaient indépendantes l'une de l'autre.
- En cas de relation entre les variables, les 2 matrices seront différents.
- Le calcul du Khi 2 permet de quantifié l'écart entre les 2 matrices:

1 - Créer une tableau de contingence: Matrice des valeurs observées

```
# Table de contingence
#la table de contingence résume sous forme de tableau les données
X = "categ"
Y = "sex"
cont = dfstat[[X, Y]].pivot_table(index=X, columns=Y, aggfunc=len,
                                  margins=True,margins_name="Total").fillna(0).copy()
tx = dfstat[X].value_counts()
ty = dfstat[Y].value_counts()
cont = cont.astype(int)
cont
```

sex	f	m	Total
categ			
0	192025	178044	370069
1	110550	100227	210777
2	16429	15351	31780
Total	319004	293622	612626

2 - Crée la matrice 'Valeurs attendues'

```
#création de la matrice valeurs attendues
#dénombrements attendus des cellules comme vu juste avant manuellement
tx_df = pd.DataFrame(tx)
tx_df.columns = ["c"]
ty_df = pd.DataFrame(ty)
ty_df.columns = ["c"]
# Valeurs totales observées
n = len(dfstat)
# Produit matriciel. On utilise pd.T pour pivoter une des deux séries.
indep = (tx_df.dot(ty_df.T) / n)
indep
```

	f	m
0	192700.752622	177368.247378
1	109754.901209	101022.098791
2	16548.346169	15231.653831

Plus de cinq valeurs attendues pour chaque combinaison de variables

Matrice 2 : Valeurs attendues

Correspond aux fréquences attendues si les variables étaient indépendantes

3 - Crée la matrice 'écart au carré normalisé de la valeur attendue VS valeur observée'

```
# Matrice
freq = (cont-indep)**2/indep
freq
```

	Total	f	m
0	NaN	2.369693	2.574540
1	NaN	5.759944	6.257859
2	NaN	0.860721	0.935126
Total	NaN	NaN	NaN

Matrice 3 : Ecart au carré normalisé

Calcul la différence entre les 2 matrices dans chaque cellules élevée au carré divisées par la valeur attendue

**Le carré donne la même importance aux combinaisons avec moins de valeurs observées qu'attendues et aux combinaisons avec plus de valeurs observées qu'attendues*

4 - Calcul du khi 2

```
#Calcul du Chi2
#Tester l'hypothèse nulle consiste à comparer les variables observées (celles déjà dans le tableau)
#avec les variables attendues.
chi2 = freq.sum().sum()
chi2
```

```
18.757882515638208
```

Khi 2: L'écart entre les 2 matrices
Somme des cellules de la matrice n°3

5 - Calcul de la p_value

```
st_p
```

```
8.448460261159054e-05
```

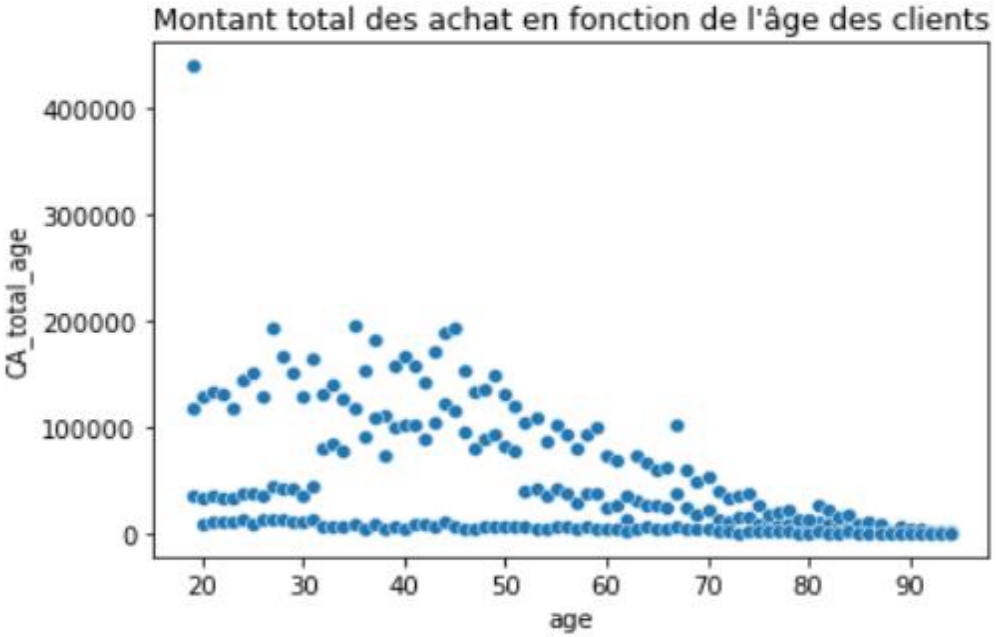
6 - Conclusion

P value ‹ α
Rejet de l'hypothèse nulle, les variables sont bien dépendantes

6.2. Existe-t-il un lien entre l'âge des clients et le montant total des achats?

6.2.1. Préparation du test

- **Observations:**



- **Types de variables:**

Quantitatives

- **Choix du test:**

Conditionné à la réalisation d 'un test de normalité :

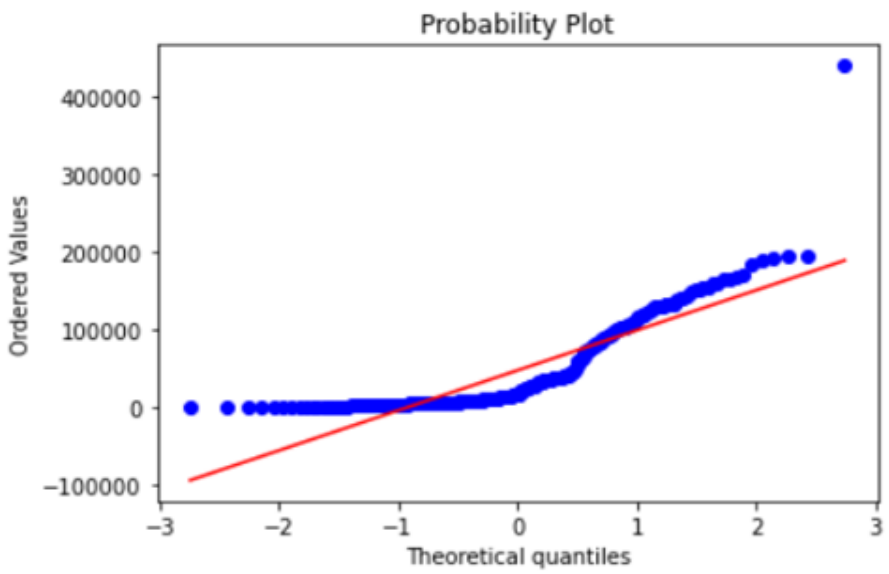
- Si positif : Test de Pearson (paramétrique)
- Si négatif: Test de Spearman (non paramétrique)

La représentation graphique des données semble indiquer que plus les clients sont âgés plus le montant des achats est diminue

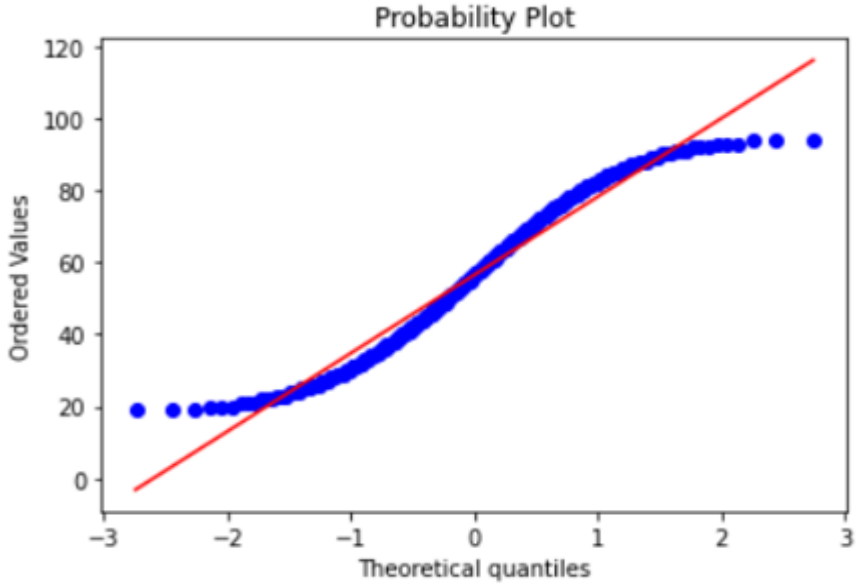
6.2.2. Tests de normalité

- **Tracé Q-Q : Diagramme Quantile-Quantile**

- Variable : 'Montant total des achats'



- Variable : 'Âge'



Distribution normale

Distribution de la variable concernée

Test Anderson Darling

Hypothèses:

H0 = Les données proviennent d'une distribution spécifiée
H1 = Les données proviennent d'une distribution non spécifiée
On pose $\alpha = 0{,}05$

Test:

- **Variable : Âge**

```
from scipy.stats import anderson

result = (anderson(newDf['age'], dist='norm'))

print(f"A-D statistic: {result[0]}")
print(f"Critical values: {result[1]}")
print(f"Significance levels: {result[2]}")
```

```
A-D statistic: 2.524546128434281
Critical values: [0.566 0.645 0.774 0.903 1.074]
Significance levels: [15.  10.   5.   2.5  1. ]
```

- **Variable : Montant total des achats**

```
result = (anderson(newDf['CA_total_age'], dist='norm'))

print(f"A-D statistic: {result[0]}")
print(f"Critical values: {result[1]}")
print(f"Significance levels: {result[2]}")
```

```
A-D statistic: 17.653045060771774
Critical values: [0.566 0.645 0.774 0.903 1.074]
Significance levels: [15.  10.   5.   2.5  1. ]
```

Résultats:

<u>Test de l'hypothèse risque 5</u> :

- Statistique de test A-D (2.52) est › à la valeur critique (0,77)
- Rejet de l'hypothèse nulle et concluons que les données ne proviennent pas d'une distrib normale:
- Orientation vers test non paramétrique Spearman au lieu de Pearson

6.2.3. Tests non paramétrique de Spearman

- **Explications et fonctionnement**

La corrélation de Spearman est l'équivalent non-paramétrique de la corrélation de Pearson. Elle mesure le lien entre deux variables. Si les variables sont ordinales, discrètes ou qu'elles ne suivent pas une loi normale, on utilise la corrélation de Spearman

- *Le coefficient de correlation varie entre -1 et +1,*
- *0 reflétant une relation nulle entre les deux variables,*
- *Une valeur négative (correlation négative) signifiant que lorsqu'une des variable augmente, l'autre diminue ;*
- *Une valeur positive (correlation positive) indique que les deux variables varient ensemble dans le même sens*

- **Hypothèses**

H0 = il n'y a pas de relation entre l'âge et le montant total des achats
H1 = Il y a bien une relation entre l'âge et le montant total des achats
On pose $\alpha = 0,05$

- **Tableau de corrélation de Spearman:**

```
#tableau : calcul du coef de correlation de spearman
newDf.corr(method="spearman")
```

	age	CA_total_age
age	1.000000	-0.620862
CA_total_age	-0.620862	1.000000

Coef négatif

Lorsqu'une variable augmente l'autre diminue ce qui confirme l'hypothèse selon laquelle + un client est âgé plus le montant des achats diminue

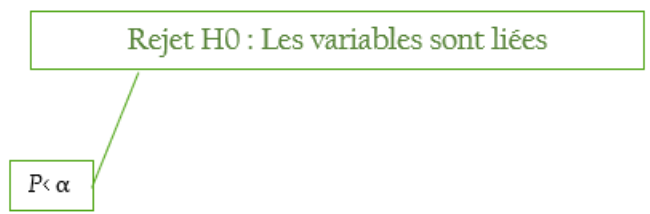
- **Niveau de correlation:**

```
#création boucle pour indiquer le niveau de corrélation
rs= spearmanr(newDf)[0]

if abs(rs)< .10:
    qual = 'negligeable'
elif abs(rs)< .20:
    qual = 'faible'
elif abs(rs)< .40:
    qual = 'modéré'
elif abs(rs)< .60:
    qual = 'relativement fort'
elif abs(rs)< .80:
    qual = 'fort'
else:
    qual= 'très fort'
qual
```

```
'fort'
```

- **P_Value:**




```
#Spearman et p_value
from scipy.stats import spearmanr
spearmanr(newDf)
```

```
SpearmanrResult(correlation=-0.6208620706292934, pvalue=1.0809773406014902e-25)
```

6.2.4. Comparaison avec le test paramétrique de Pearson

- **Test de Pearson** *Hypothèses identiques*

```
#test de pearson(test parametrique)
st.pearsonr(newDf['age'],newDf['CA_total_age'])
```

```
(-0.5098441854808893, 1.7329812860262953e-16)
```

P< α Hypothèse H0 rejetée

- **Explications**

Le coef de pearson est pertinent lorsqu'il existe une relation linéaire entre les variables et que les variables suivent une distribution normale ce qui n'est pas le cas ici, pour autant son fonctionnement est le suivant:

- Le coefficient de Pearson est un indice reflétant une relation linéaire entre deux variables continues.
- Le coefficient de corrélation varie entre -1 et +1,0 reflétant une relation nulle entre les deux variables,
- une valeur négative (corrélation négative) signifiant que lorsqu'une des variable augmente, l'autre diminue
- une valeur positive (corrélation positive) indique que les deux variables varient ensemble dans le même sens

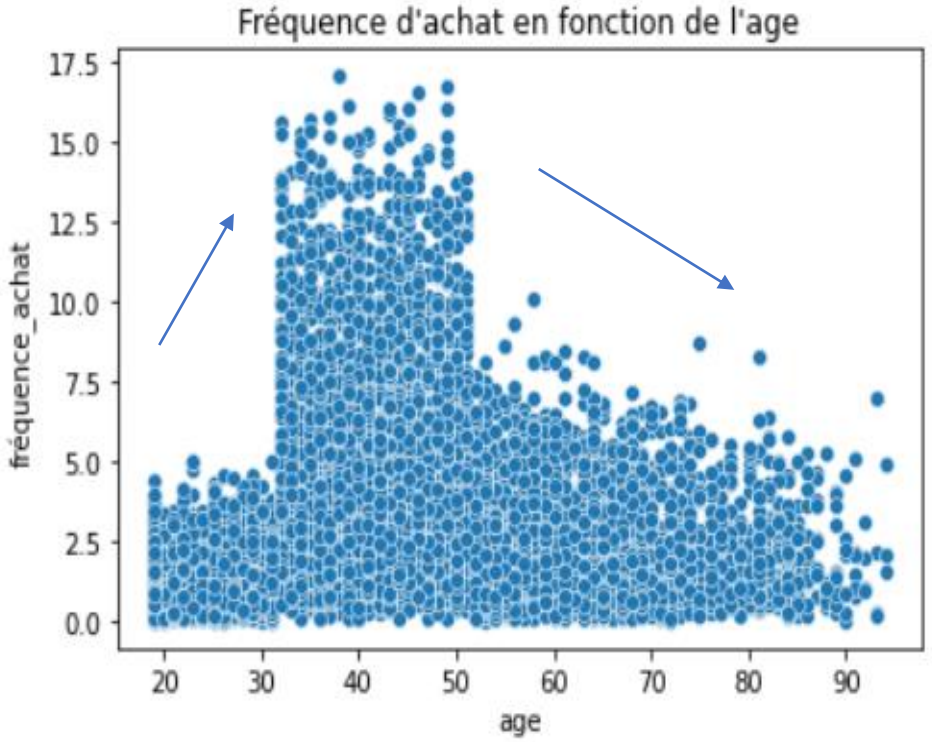
- **Resultats**

On remarque que: *La pvalue est proche de -1 on peut en déduire qu'il y a une forte corrélation entre les 2 variables ce qui confirme la déduction précédente concernant la diminution du montant des achats avec l'augmentation de l'âge ainsi que les résultats du test de Spearman

6.3. Existe-t-il un lien entre l'âge des clients et la Fréquence d'achat?

6.3.1. Préparation du test

- **Observations:**



Les graphique montre que la fréquence d'achat augmente entre 20 et 50 ans puis diminue entre 50 et 90 ans

- **Types de variables:**

2 variables quantitatives

- **Choix du test:**

Conditionné à la réalisation d 'un test de normalité

- **Résultat:**

Négatif les variables ne suivent pas une distribution normale, choix du test de Spearman

6.3.3. Tests non paramétrique de Spearman

- **Tableau de corrélation de Spearman:**

```
#tableau coef de correlation de spearman
newDf_freq.corr(method="spearman")
```

	age	fréquence_achat
age	1.000000	0.129604
fréquence_achat	0.129604	1.000000

Coef positif mais proche de 0

- **P_Value:**

```
spearmanr(newDf_freq)
```

```
SpearmanrResult(correlation=0.12960358760068205,
pvalue=1.6381331276761283e-33)
```

P< α on peut rejeter H0

- **Niveau de correlation:**

```
rs= spearmanr(newDf_freq)[0]

if abs(rs)< .10:
    qual = 'negligeable'
elif abs(rs)< .20:
    qual = 'faible'
elif abs(rs)< .40:
    qual = 'modéré'
elif abs(rs)< .60:
    qual = 'relativement fort'
elif abs(rs)< .80:
    qual = 'fort'
else:
    qual= 'très fort'
qual
```

```
'faible'
```

- *Le coefficient de correlation varie entre -1 et +1,*
- *0 reflétant une relation nulle entre les deux variables,*
- *Une valeur négative (correlation négative) signifiant que lorsqu'une des variable augmente, l'autre diminue ;*
- *Une valeur positive (correlation positive) indique que les deux variables varient ensemble dans le même sens*

On peut rejeter l'hypothèse nulle, pour autant le lien entre les 2 variable reste faible et partiellement vrai lorsque l'on regarde le graphique:

- Vrai de 19 à 30 ans
- Faux de 50 à 90 ans

6.3.4. Comparaison avec le test paramétrique de Pearson

```
#test de pearson(test parametrique)
st.pearsonr(newDf_freq['age'],newDf_freq['fréquence_achat'])
```

```
(0.03287075896343302, 0.002306572042379856)
```

$P < \alpha$ Hypothèse H0 rejetée

Le coef de pearson est pertinent lorsqu'il existe une relation linéaire entre les variables et que les variables suivent une distribution normale ce qui n'est pas le cas ici, pour autant son fonctionnement est le suivant:

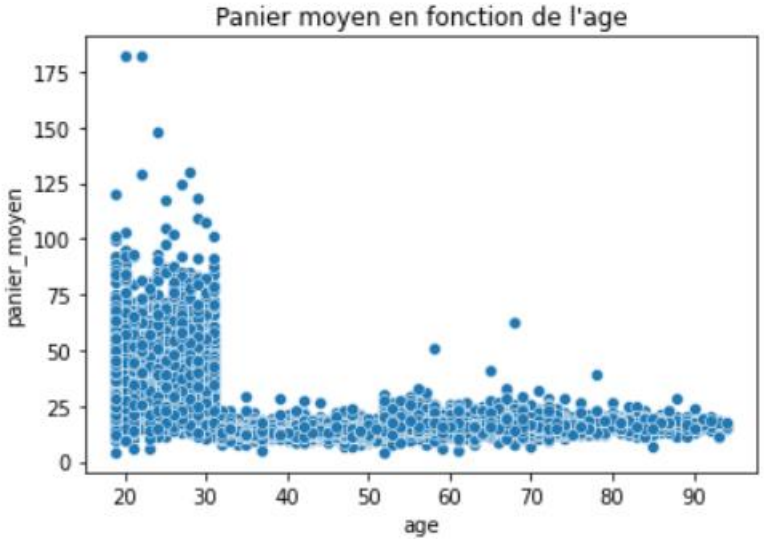
- Le coefficient de Pearson est un indice reflétant une relation linéaire entre deux variables continues.
- Le coefficient de corrélation varie entre -1 et +1,0 reflétant une relation nulle entre les deux variables,
- une valeur négative (corrélation négative) signifiant que lorsqu'une des variable augmente, l'autre diminue
- une valeur positive (corrélation positive) indique que les deux variables varient ensemble dans le même sens

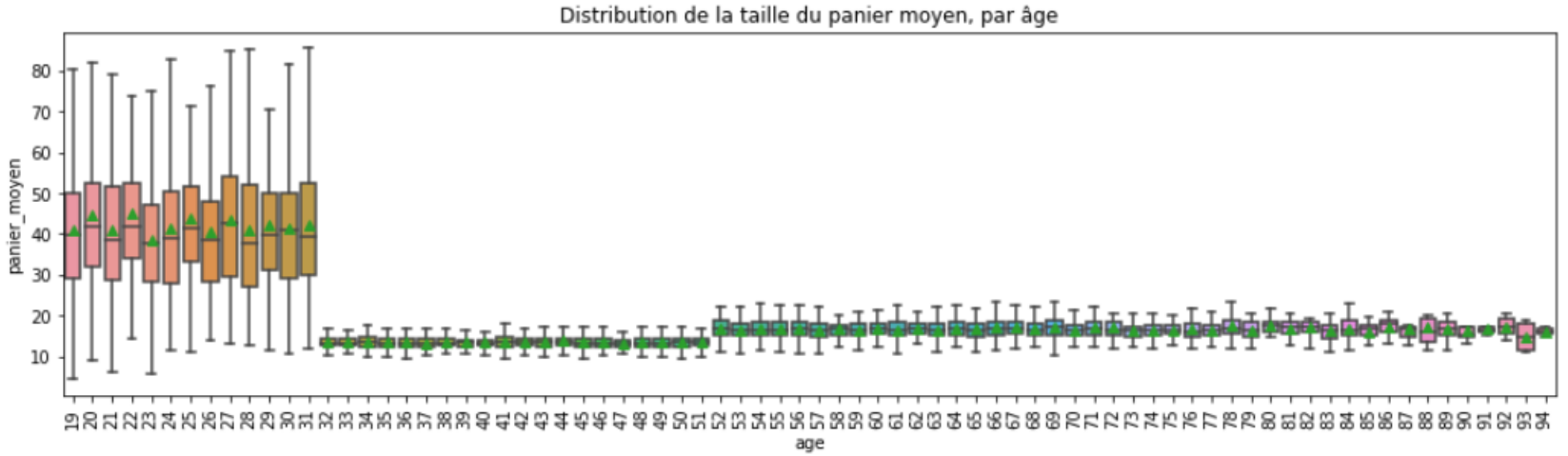
<u>Remarque</u>:
Les résultats de Pearson vont dans le même sens.
La p_value est pour autant moins significative même si < à α

6.4. Existe-t-il un lien entre l'âge des clients et le panier moyen des clients?

6.4.1. Préparation du test

- **Observations:**





- *Comme vu précédemment les moins de 30 ans ont un panier plus important malgré une fréquence d'achat moins importante*
- *Après 30 ans les paniers moyens sont homogènes*
- *Le panier moyen le plus élevé est celui des 19/31 ans ce qui confirme les résutats vu précédemment: les 32/51 achètent moins cher mais en volume plus important*

Vérifions qu'il existe bien un lien entre les variables âge et montant total des achats et qu'il ne s'agit pas d'un simple hasard

- **Types de variables**

2 variables quantitatives

- **Choix du test**

Conditionné à la réalisation d 'un test de normalité

- **Résultat:**

Négatif, orientation vers le test de Spearman

6.4.3. Tests non paramétrique de Spearman

- **Tableau de corrélation de Spearman:**

```
#tableau coef de correlation de spearman
newDfp.corr(method="spearman")
```

	age	panier_moyen
age	1.000000	-0.325921
panier_moyen	-0.325921	1.000000

- *Le coefficient de correlation varie entre -1 et +1,*
- *0 reflétant une relation nulle entre les deux variables,*
- *Une valeur négative (correlation négative) signifiant que lorsqu'une des variable augmente, l'autre diminue ;*
- *Une valeur positive (correlation positive) indique que les deux variables varient ensemble dans le même sens*

- **P_Value:**

```
spearmanr(newDfp)
```

```
SpearmanrResult(correlation=-0.3259206528690354,
pvalue=7.929082716253137e-212)
```

P< α on peut rejeter H0

- **Niveau de correlation:**

```
#création boucle pour indiquer le niveau de corrélation
rs= spearmanr(newDfp)[0]

if abs(rs)< .10:
    qual = 'negligeable'
elif abs(rs)< .20:
    qual = 'faible'
elif abs(rs)< .40:
    qual = 'modéré'
elif abs(rs)< .60:
    qual = 'relativement fort'
elif abs(rs)< .80:
    qual = 'fort'
else:
    qual= 'très fort'
qual
```

```
'modéré'
```

On peut rejeter l'hypothèse nulle pour autant le lien entre les 2 variable reste modéré et partiellement vrai lorsque l'on regarde le graphique:

- Faux de 19 à 30 ans
- Vrai de 30 à 90 ans

6.4.4. Comparaison avec le test paramétrique de Pearson

```
#test de pearson(test parametrique)
st.pearsonr(newDfp['age'],newDfp['panier_moyen'])
```

```
(-0.5089196147145202, 0.0)
```

$P < \alpha$ Hypothèse H0 rejetée

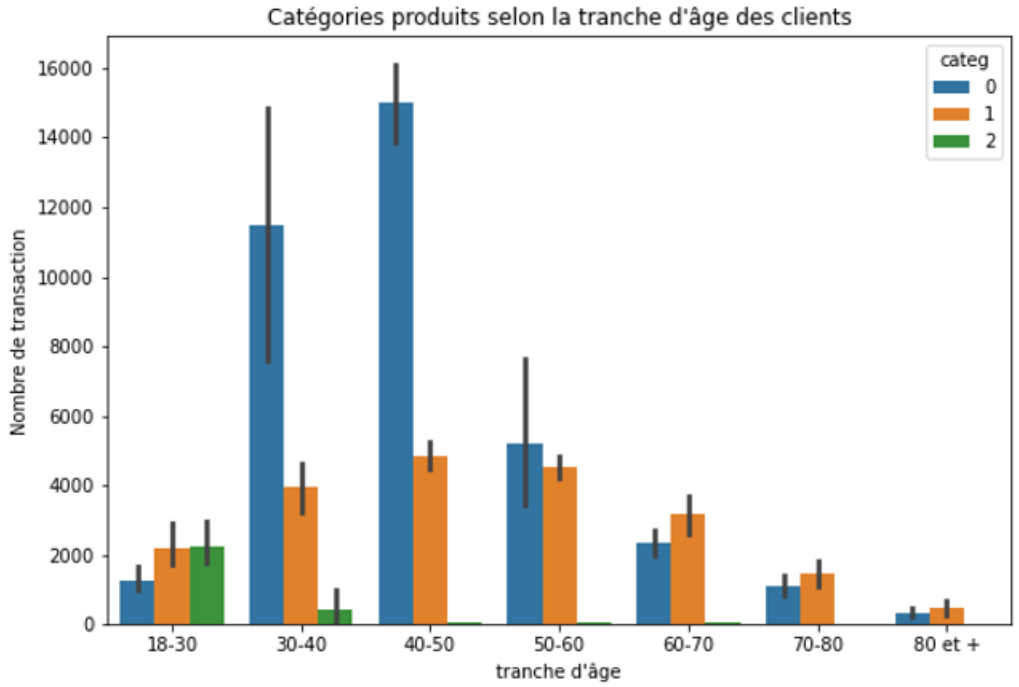
Le coef de pearson est pertinent lorsqu'il existe une relation linéaire entre les variables et que les variables suivent une distribution normale ce qui n'est pas le cas ici, pour autant son fonctionnement est le suivant:

- Le coefficient de Pearson est un indice reflétant une relation linéaire entre deux variables continues.
- Le coefficient de corrélation varie entre -1 et +1,0 reflétant une relation nulle entre les deux variables,
- une valeur négative (corrélation négative) signifiant que lorsqu'une des variable augmente, l'autre diminue
- une valeur positive (corrélation positive) indique que les deux variables varient ensemble dans le même sens

6.5. Existe-t-il un lien entre l'âge des clients et la catégorie de livre acheté

6.5.1. Préparation du test

- **Observations:**



Le graphique ainsi que les observations précédentes indiquent que :

- *La categ 0 serait d'avantage achetée par les 30-50 ans*
- *La categ 2 par les 18-40*
- *La categ 1 par tout le monde*

Vérifions qu'il existe bien un lien entre ces 2 variables et qu'il ne s'agit pas d'un hasard

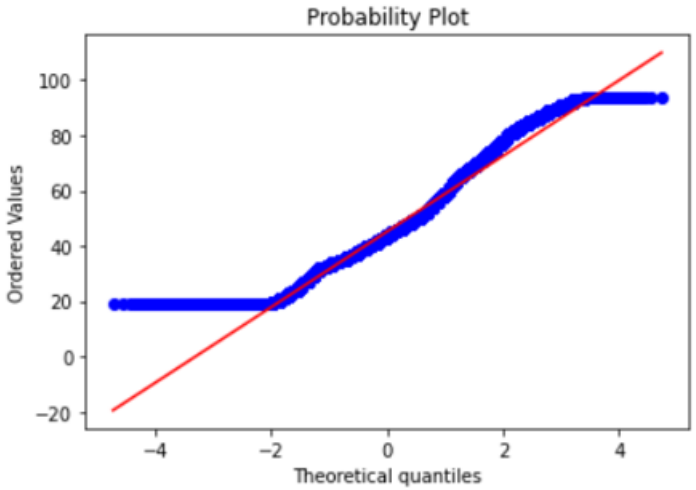
- **Types de variables:**

Nous étudions ici le potentiel lien entre 1 variable qualitative et 1 variable quantitative
Le choix du test est conditionné à la réalisation d 'un test de normalité

6.5.2. Tests de normalité

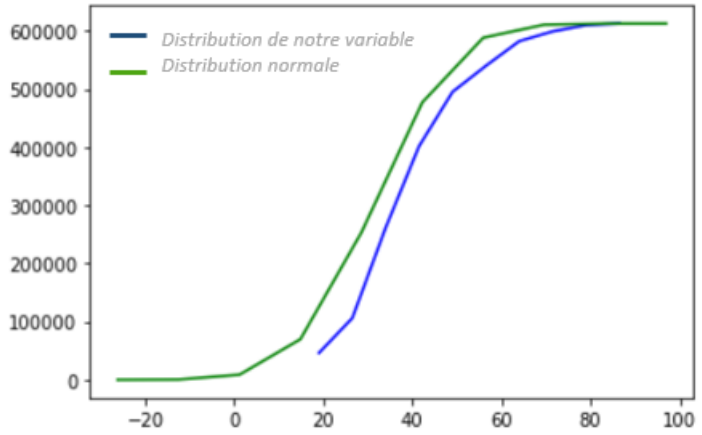
- **Tracé Q-Q**

Distribution de la variable âge vs distribution normale



- **Tracé Kolmogorov-Smirnov**

Distribution de la variable âge vs distribution normale



- **Test Anderson Darling**

```
result = (anderson(df2['age'], dist='norm'))

print(f"A-D statistic: {result[0]}")
print(f"Critical values: {result[1]}")
print(f"Significance levels: {result[2]}")
```

```
A-D statistic: 4996.998227147036
Critical values: [0.576 0.656 0.787 0.918 1.092]
Significance levels: [15.  10.   5.   2.5  1. ]
```

La variable ne suit pas une distribution normale. Choix du test non paramétrique de Kruskal et Wallis

6.5.3. Test non paramétrique de KRUSKAL ET WALLIS

- **Explications et fonctionnement**

  - Pour une variable qualitative avec plus de deux modalités, on fait une ANOVA (Analysis Of Variance) qui compare les moyennes des groupes. Ce test est un test paramétrique
  - L'alternative est l'utilisation du test de Kruskal et Wallis. Il va permettre de déterminer si l'âge influence la catégorie en comparant les médianes des 3 groupes et indiquer si les 3 groupes sont différents, auquel cas

- **Hypothèses**

  - H0: les médianes sont égales = il n'y a pas de difference entre les groupes
  - H1: les médianes sont différentes = les groupes sont different

- **Test**

1) Créer 1 groupe pour chaque catégorie:

```
#on crée 3 groupe
Groupe1=df2.query('categ==0')['age'].tolist()
Groupe2=df2.query('categ==1')['age'].tolist()
Groupe3=df2.query('categ==2')['age'].tolist()
```

2) Test de Kruskal:

```
#on lance le test Krustal wallis
st.kruskal(Groupe1, Groupe2, Groupe3)
```

```
KruskalResult(statistic=69911.21351199699, pvalue=0.0)
```

Rejet H0

- **Conclusion:**

Les groupes sont donc différents ce qui signifie que la variable âge a une influence sur le choix de la catégorie les variables sont donc liées

**Médiane = rang divisé par nombre de valeurs*

7.SYNTHESE ET PRECONISATIONS

POINTS FORTS

- Le chiffre d'affaire est en progression
- Fort succès des catégories 0 et 1

AXES D'AMELIORATION

- Présence de produits invendus et d'autres très peu vendus
- Vigileance sur les ruptures de stock (exemples categ 1 au mois d'octobre)
- Vigileance sur les plus de 30 ans (baisse du panier moyen)
- Vigileance sur les plus de 50 ans (baisse de la fréquence d'achat)

SUGGESTIONS

- Développer le Btb
- Développer une offre dédiée pour les 50-90 ans