# Analysez les ventes d'une librairie avec Python

Lapage

# Contexte

L'entreprise Lapage était originellement une librairie physique avec plusieurs points de vente. Devant le succès de certains de ses produits et l'engouement de ses clients, elle a décidé depuis 2 ans d'ouvrir un site de vente en ligne.

La structure a besoin d'aide pour mieux comprendre ses données, c'est pourquoi j'ai été recruté.

J'interviens en tant que Data Analyst afin de faire le point sur l'activité.

# Plan

1. **Exploration des données et jointure**
    - Fichier : Comprendre la structure et les variables des fichiers de données.
    - Jointure

**2.Analyses du chiffre d'affaires**

- Calcul du CA : Déterminer le chiffre d'affaires total à partir des données disponibles.
- Analyse bivariée : Examiner les relations entre le CA et d’autres variables.
- Analyse temporelle : Étudier les tendances du chiffre d'affaires au fil du temps.

**3.Tests statistiques**

- Genre/catégorie
- Âge/montant total d'achat
- Âge/fréquence d'achat
- Âge/panier moyen
- Âge/catégorie des livres

# DataFrames

- Création de trois DataFrames : Transactions, Produits et Clients à partir des fichiers sources déjà nettoyés.
- Exploration de chaque DataFrame pour comprendre sa structure et identifier d'éventuelles anomalies.
- Ajustements mineurs effectués pour corriger les erreurs de format dans les dates et convertir les types de variables si nécessaire.
- Validation de l'unicité des clés primaires (session_id, id_prod, client_id).
- Préparation des DataFrames pour la jointure en vérifiant la cohérence des formats des variables.

```
Entrée [17]: # informations concernant le dataframe

transactions.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 687534 entries, 0 to 687533
Data columns (total 4 columns):
 #   Column      Non-Null Count   Dtype 
---  ------      --------------   ----- 
 0   id_prod     687534 non-null  object
 1   date        687534 non-null  object
 2   session_id  687534 non-null  object
 3   client_id   687534 non-null  object
dtypes: object(4)
memory usage: 21.0+ MB
```

```
Entrée [11]: # informations concernant le dataframe

produits.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3286 entries, 0 to 3285
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   id_prod  3286 non-null   object 
 1   price    3286 non-null   float64
 2   categ    3286 non-null   int64  
dtypes: float64(1), int64(1), object(1)
memory usage: 77.1+ KB
```

```
Entrée [4]: # informations concernant le dataframe

clients.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8621 entries, 0 to 8620
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   client_id  8621 non-null   object
 1   sex        8621 non-null   object
 2   birth      8621 non-null   int64 
dtypes: int64(1), object(2)
memory usage: 202.2+ KB
```

# Jointure

Entrée [29]:
```
#Création d'un dataframe global qui fusionne les tables transactions, produits et clients

df = pd.merge(transactions,produits, how='left', on='id_prod', indicator="Jointure")
df = df.merge(clients, how='right', on='client_id', indicator="merge")
df
```

Out[29]:

| | id_prod | date | session_id | client_id | Année | Mois | Jours | price | categ | Jointure | sex | birth | Âge | merge |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1_483 | 2021-03-13 21:35:55.949042 | s_5913 | c_4410 | 2021 | 03 | 13 | 15.99 | 1.0 | both | f | 1967 | 57 | both |
| **1** | 0_1111 | 2021-03-22 01:27:49.480137 | s_9707 | c_4410 | 2021 | 03 | 22 | 19.99 | 0.0 | both | f | 1967 | 57 | both |
| **2** | 1_385 | 2021-03-22 01:40:22.782925 | s_9707 | c_4410 | 2021 | 03 | 22 | 25.99 | 1.0 | both | f | 1967 | 57 | both |
| **3** | 0_1455 | 2021-03-22 14:29:25.189266 | s_9942 | c_4410 | 2021 | 03 | 22 | 8.99 | 0.0 | both | f | 1967 | 57 | both |
| **4** | 0_1420 | 2021-03-22 22:31:25.825764 | s_10092 | c_4410 | 2021 | 03 | 22 | 11.53 | 0.0 | both | f | 1967 | 57 | both |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **687550** | 0_1472 | 2022-05-14 00:24:49.391917 | s_208110 | c_84 | 2022 | 05 | 14 | 12.49 | 0.0 | both | f | 1982 | 42 | both |
| **687551** | 0_1438 | 2022-05-29 06:11:50.316631 | s_215697 | c_84 | 2022 | 05 | 29 | 9.31 | 0.0 | both | f | 1982 | 42 | both |
| **687552** | 1_459 | 2022-12-17 00:16:56.629536 | s_313173 | c_84 | 2022 | 12 | 17 | 15.99 | 1.0 | both | f | 1982 | 42 | both |
| **687553** | 0_1104 | 2022-12-17 00:24:14.357525 | s_313173 | c_84 | 2022 | 12 | 17 | 13.21 | 0.0 | both | f | 1982 | 42 | both |
| **687554** | 1_688 | 2022-12-17 00:36:24.148027 | s_313173 | c_84 | 2022 | 12 | 17 | 18.19 | 1.0 | both | f | 1982 | 42 | both |

687555 rows × 14 columns

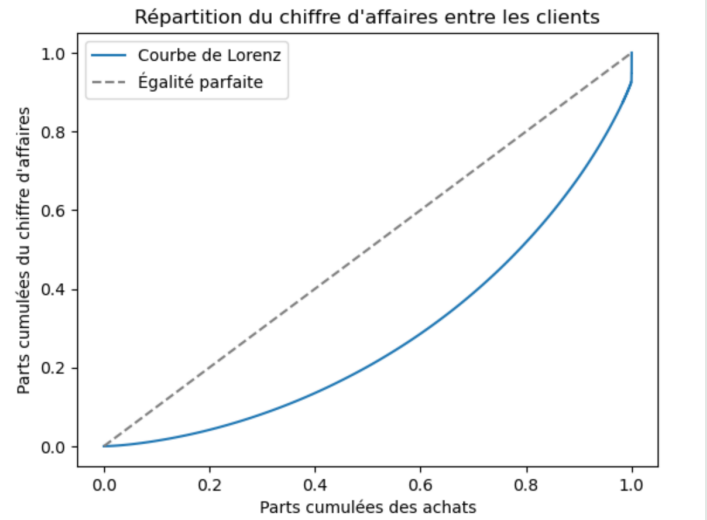
- Nous réalisons d'abord une jointure entre les DataFrames transactions et produits en utilisant id_prod, en effectuant une jointure par la gauche pour associer chaque transaction avec les détails des produits correspondants.
- Ensuite, nous fusionnons ce résultat avec le DataFrame clients en utilisant client_id, en réalisant une jointure par la droite pour inclure les informations sur les clients associées à chaque transaction et produit.
- Nous utilisons ensuite deux indicateurs, 'merge' et 'jointure', pour vérifier la réussite de la fusion des DataFrames.

# Calcul du Chiffre d'affaires

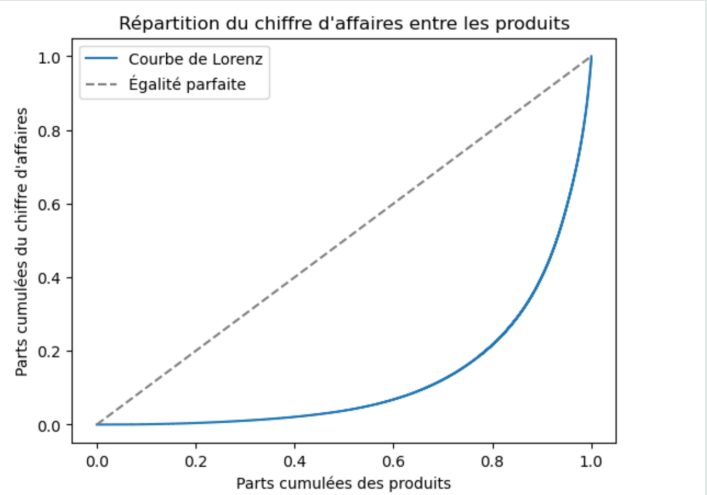
- Le chiffre d'affaires total s'élève donc à 12 027 663,1 euros.
- Grâce à une courbe de Lorenz, on peut observer la répartition du chiffre d'affaires entre les clients. L'indice de Gini de 0.442 indique une certaine inégalité modérée dans la répartition du chiffre d'affaires entre les clients. On remarque que 20% des clients génèrent 50% du chiffre d'affaires.
- Quant à la répartition du chiffre d'affaires entre les produits, la courbe de Lorenz et l'indice de Gini nous indiquent une inégalité encore plus prononcée. En effet, un peu près 20% des références produisent 80% du chiffre d'affaires total.

```
# Chiffre d'affaires total:

CA_tot = round(df['price'].sum(),1)

print ("Le chiffre d'affaires total est de", CA_tot)
```

```
Le chiffre d'affaire total est de 12027663.1
```




```
indice de Gini: 0.442
```



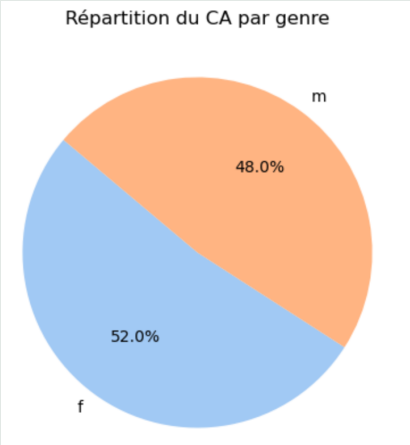

```
Indice de Gini pour la répartition du chiffre d'affaires par produit: 0.744
```

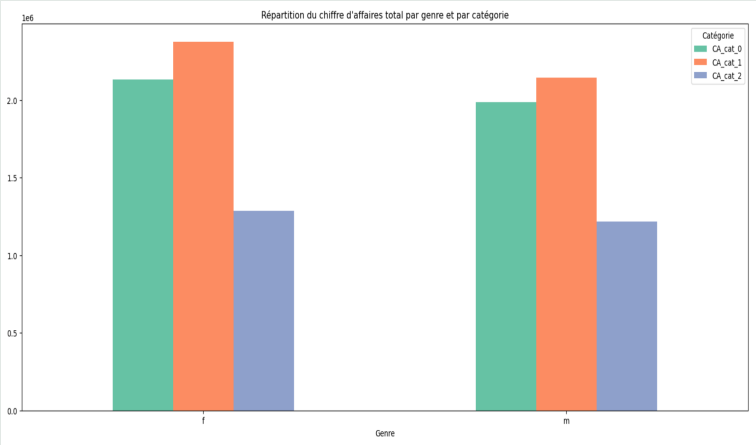
# Top/Flop

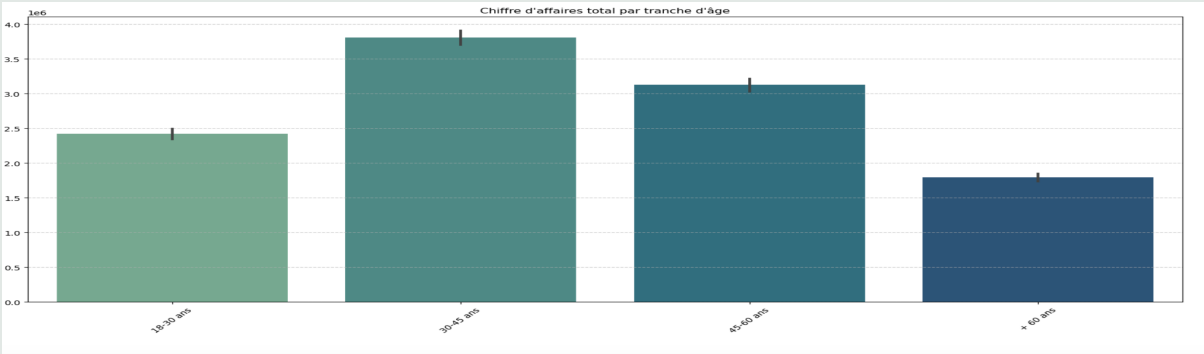
- Concernant les produits, nous constatons que le top 10 représente un pourcentage important du chiffre d'affaires, comme le démontre notre courbe de Lorenz. En revanche, les produits moins performants, les 'Flops', contribuent de manière négligeable au chiffre d'affaires total.
- Quant aux clients, nous notons que 4 d'entre eux se démarquent dans le top. En effet, ensemble, ils représentent un peu moins de 8% du chiffre d'affaires. Il est probable que ces clients soient des professionnels. Pour obtenir une vision plus cohérente de l'ensemble des données, nous exclurons ces clients de nos graphiques et de nos tests.

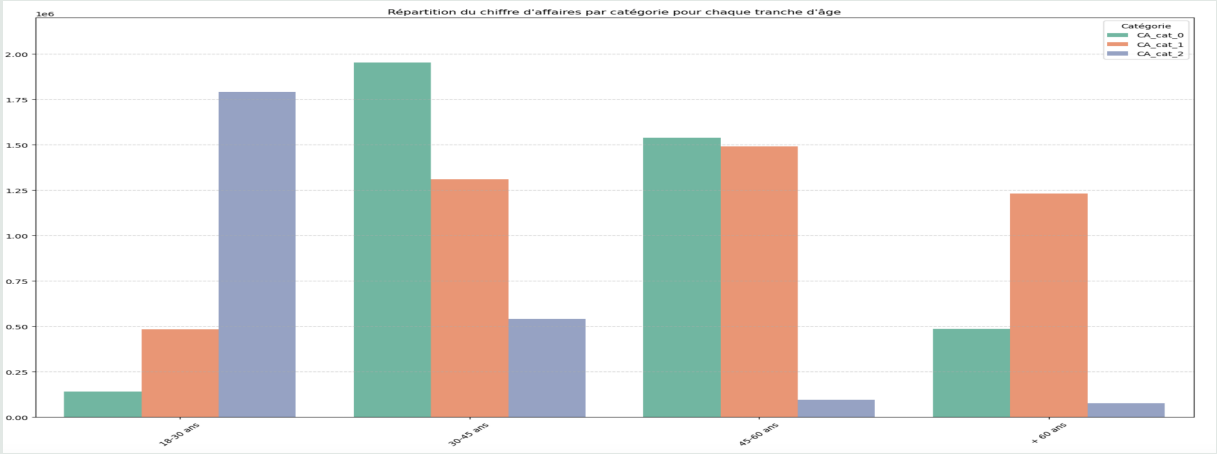
```
top 10 references
id_prod
2_159    94893.50
2_135    69334.95
2_112    65407.76
2_102    60736.78
2_209    56971.86
1_395    56617.47
1_369    56136.60
2_110    53846.25
1_383    53834.43
1_414    53522.18
Name: price, dtype: float64
top flop references
id_prod
0_1539    0.99
0_1284    1.38
0_1653    1.98
0_541     1.99
0_1601    1.99
0_807     1.99
0_1728    2.27
0_1498    2.48
0_898     2.54
0_1840    2.56
Name: price, dtype: float64
```

| client_id | sex | Âge | CA_total_client | CA_cat_0 | CA_cat_1 | CA_cat_2 | Tranche_age |
|---|---|---|---|---|---|---|---|
| **c_1609** | m | 44 | 326039.89 | 214447.24 | 110091.44 | 1501.21 | 30-45 ans |
| **c_4958** | m | 25 | 290227.03 | 48.76 | 39841.93 | 250336.34 | 18-30 ans |
| **c_6714** | f | 56 | 153918.60 | 57254.59 | 73566.22 | 23097.79 | 45-60 ans |
| **c_3454** | m | 55 | 114110.57 | 28779.69 | 84055.66 | 1275.22 | 45-60 ans |
| **c_1570** | f | 45 | 5285.82 | 2812.80 | 2327.03 | 145.99 | 45-60 ans |
| **c_3263** | f | 39 | 5276.87 | 3399.29 | 1877.58 | 0.00 | 30-45 ans |
| **c_2140** | f | 47 | 5260.18 | 3360.28 | 1753.91 | 145.99 | 45-60 ans |
| **c_2899** | f | 30 | 5214.05 | 25.38 | 779.01 | 4409.66 | 30-45 ans |
| **c_7319** | f | 50 | 5155.77 | 2936.17 | 2175.61 | 43.99 | 45-60 ans |
| **c_7959** | f | 50 | 5135.75 | 3467.40 | 1668.35 | 0.00 | 45-60 ans |



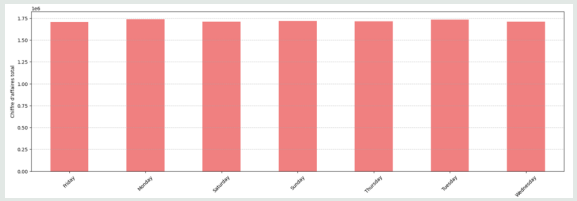



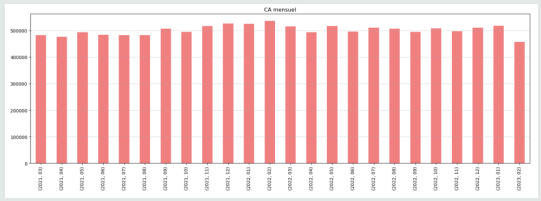



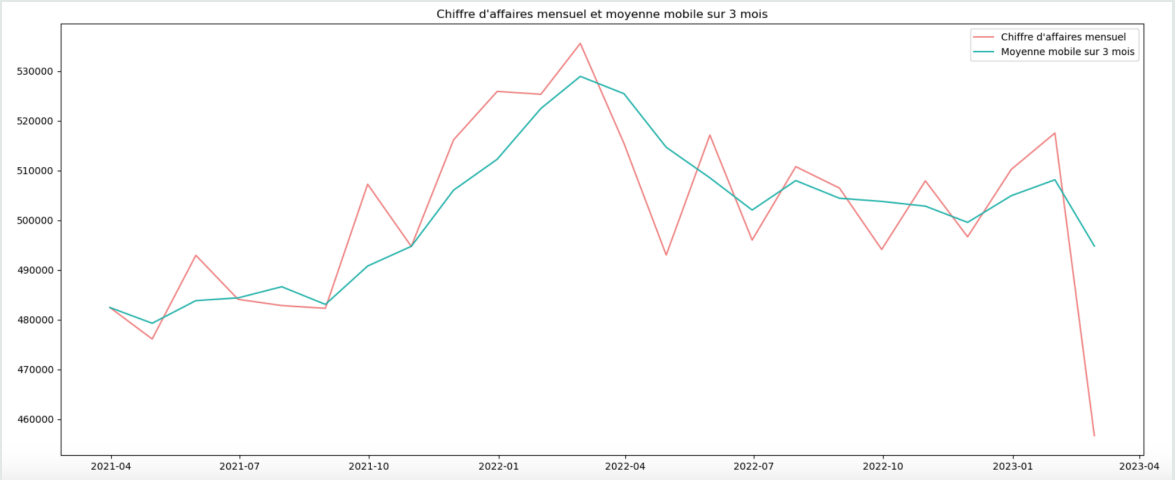

# Analyse du CA

- On observe que les différences sont minimes dans la répartition du chiffre d'affaires en fonction du genre et des catégories de produits.
- On peut noter cependant des variations certaines dans la répartition du chiffre d'affaires en fonction de l'âge, il faut noter que les tranches d'âge 18-30 ans et +60 ans soient sous-représentés dans l'échantillon.
- Il est important de noter qu'une corrélation est observée entre l'âge et la catégorie des produits.

# Analyse temporelle



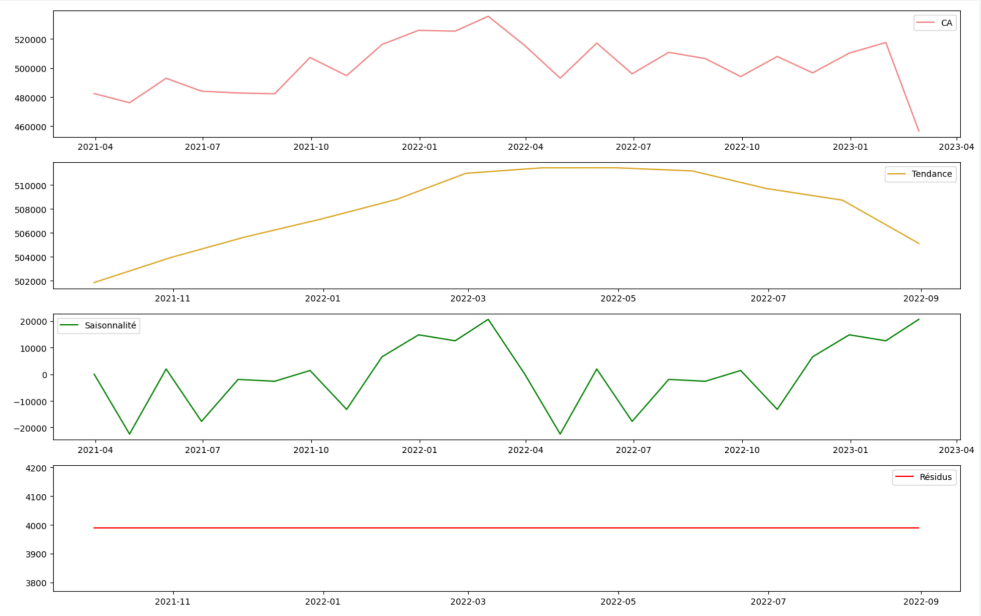




- Mensuellement et quotidiennement, nous ne notons pas de fluctuations importantes dans le chiffre d'affaires.
- Cependant, lorsque l'on examine le CA sur l'ensemble de l'échelle temporelle, on peut observer des fluctuations et des tendances qui se dégagent, notamment grâce à l'utilisation d'une moyenne mobile sur 3 mois.

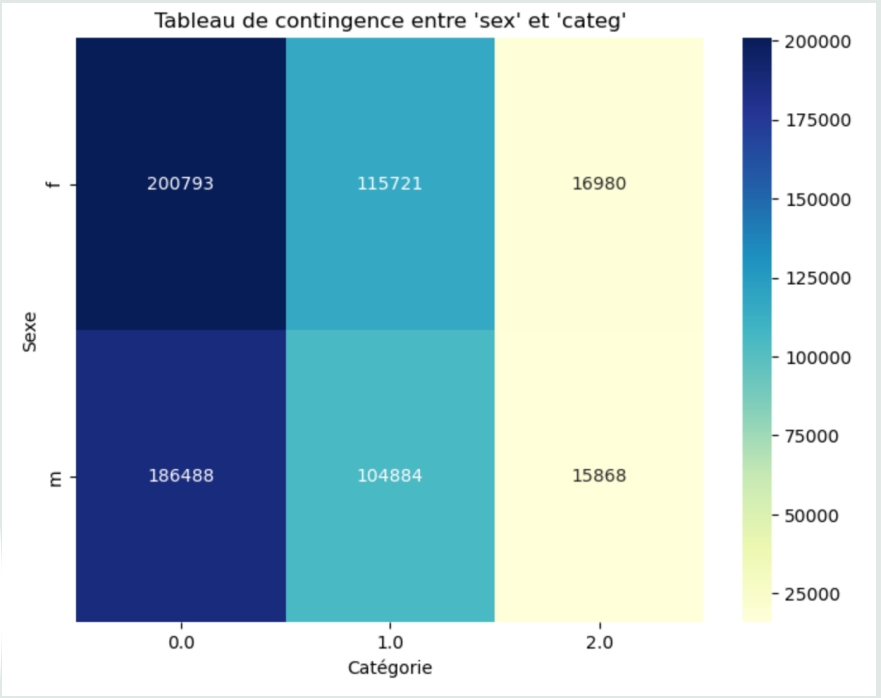
# Décomposition temporelle

- Nous avons observé une tendance haussière du chiffre d'affaires de avril 2021 jusqu'à mars 2022, suivie d'une période de consolidation jusqu'à juin 2022, puis d'une tendance baissière jusqu'à septembre 2022.
- Pour ce qui est de la saisonnalité, des variations régulières autour du niveau moyen sont perceptibles, avec une période de 12 mois.
- Par ailleurs, notre résidu plat suggère peu de fluctuations non expliquées qui persistent après avoir retiré les composantes saisonnières et de tendance. Cette observation renforce la validité du modèle et suggère que les variations observées dans le chiffre d'affaires peuvent être principalement attribuées à des tendances et des effets saisonniers plutôt qu'à des fluctuations aléatoires.



# Genre/Catégories

- H0 : Il n'y a pas de lien entre le genre du client et la catégorie de produit.
- H1 : Il existe un lien significatif entre le genre du client et la catégorie de produit.
- La p-valeur très faible (< 0,05) indique qu'il y a une association significative et que cette association n'est pas due au hasard. Par conséquent, nous rejetons l'hypothèse nulle selon laquelle le sexe des individus est indépendant de la catégorie des produits achetés.
- Cependant, malgré le lien significatif, le coefficient de Cramer est très bas, ce qui suggère une association faible entre le sexe des individus et la catégorie des produits achetés.

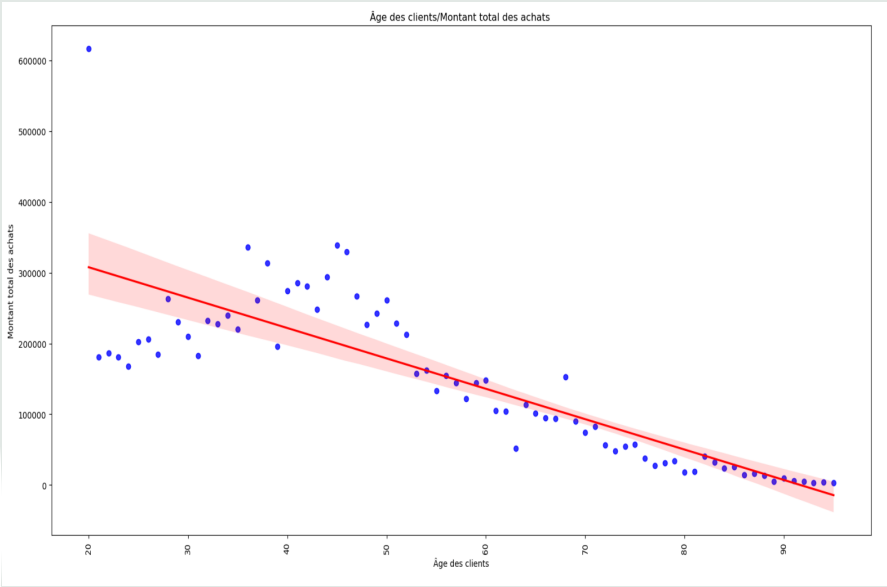



```
Résultats du test du chi2 :
Chi2 : 22.66856665178056
P-value : 1.1955928116587024e-05
Degrés de liberté : 2
Tableau attendu :
 [[201574.89662481 114822.13191434  17096.97146086]
 [185706.10337519 105782.86808566  15751.02853914]]

Coefficient de V de Cramer : 0.005742023164965101
```

# Âge/Montant Total d’achat

- Les tests de normalité pour nos deux variables indiquent que l'âge et le montant total des achats ne sont pas distribués selon une loi normale.

- Le coefficient de corrélation de Spearman entre l'âge et le montant total des achats est estimé à environ -0.185, ce qui indique une corrélation monotone négative faible entre les deux variables.
- La p-valeur associée est très proche de zéro(<0,05), ce qui renforce l'idée qu'il existe une corrélation statistiquement significative entre l'âge des clients et le montant total de leurs achats, bien que cette relation soit faible.
- En d'autres termes, plus l'âge des clients est élevé, moins leur montant total d'achats tend à être élevé, et vice versa.

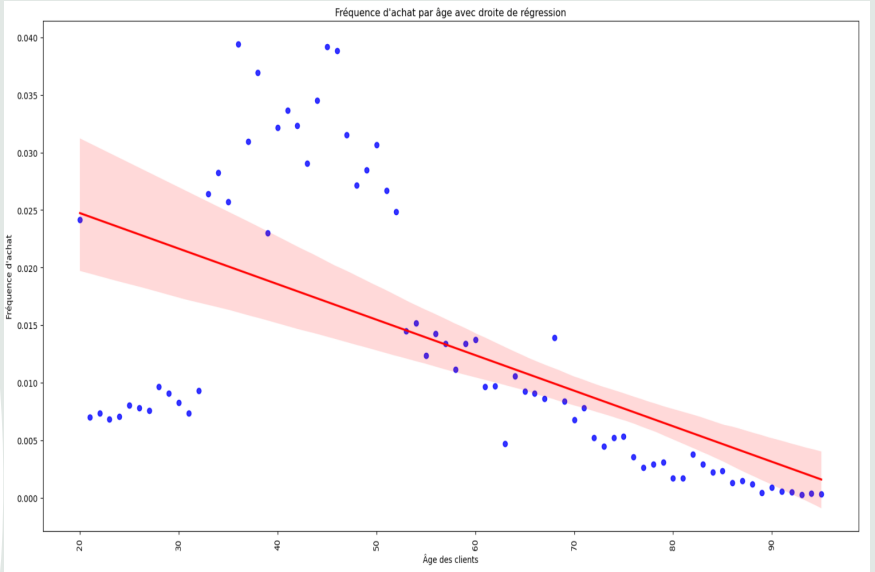



```
Test de normalité pour la variable Âge :
Statistique de test : 0.9697333574295044
P-valeur : 4.62984948847373e-39

Test de normalité pour la variable Montant total des achats :
Statistique de test : 0.9042858481407166
P-valeur : 0.0
```

```
Coefficient de corrélation de Spearman  -0.18453804793783096
P-valeur : 1.0212910436382683e-66
```

# Âge/Fréquence d’achats

- Le coefficient de corrélation de Spearman entre les deux variables est estimé à environ -0.676. Cette valeur indique une corrélation monotone négative modérée forte  entre l'âge des clients et le montant total de leurs achats.
- La p-valeur associée est très faible(<0,05), ce qui suggère une relation significative entre l'âge des clients et le montant total de leurs achats.
- En d'autres termes, plus l'âge des clients est élevé, moins leur fréquence d'achats tend à être élevé, et vice versa.

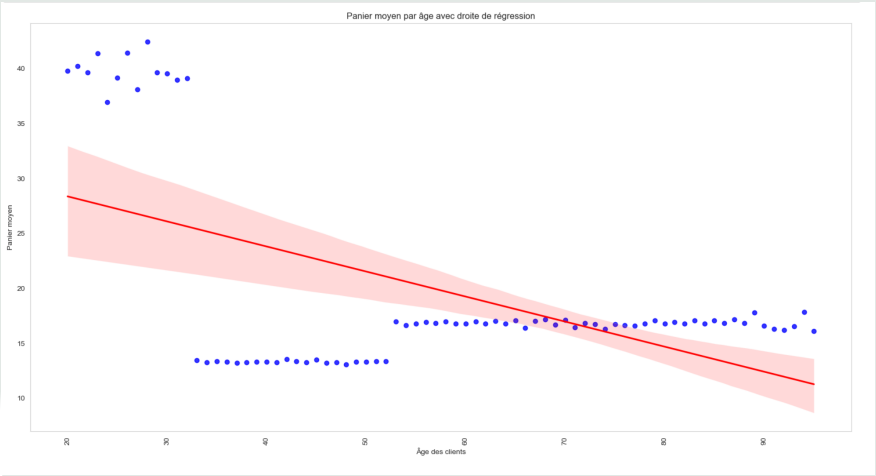



```
Test de normalité pour l'âge :
Statistique de test: 0.971943736076355
P-value: 0.0
La distribution de l'âge ne semble pas être normale (rejeter H0)

Test de normalité pour la fréquence d'achat :
Statistique de test: 0.8589646220207214
P-value: 5.637875801767223e-07
La distribution de la fréquence d'achat ne semble pas être normale (rejeter H0)
```

```
Coefficient de corrélation de Spearman: -0.6756527682843472
P-value: 2.146632000345534e-11
Il existe une corrélation négative significative entre l'âge et la fréquence d'achat.
```

# Âge/Panier moyen

- Le coefficient de corrélation de Spearman entre l'âge et le panier moyen est estimé à environ -0,077 indiquant une corrélation monotone négative très faible entre les deux variables.
- La p-valeur associée est supérieurs alpha(>0,05), ce qui signifie qu'il n'y a pas suffisamment de preuves pour rejeter l'hypothèse nulle selon laquelle il n'y a pas de corrélation significative entre les deux variables.
- Les résultats indiquent une tendance à une diminution du panier moyen avec l'augmentation de l’âge bien que cette tendance ne soit pas statistiquement significative.

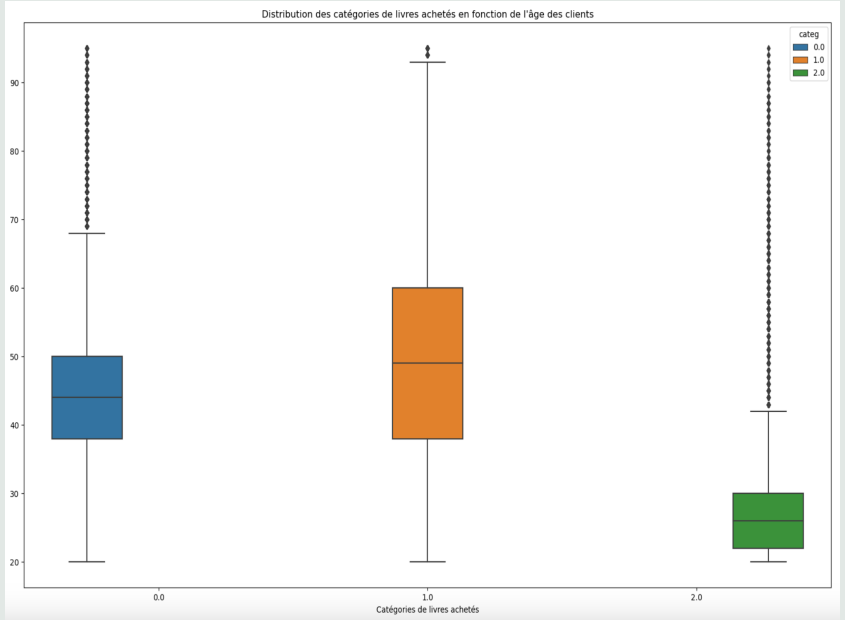



```
Test de normalité de Shapiro pour le panier moyen :
Statistique de test : 0.6062705516815186
P-value : 5.761508349404199e-13
La distribution du panier moyen ne semble pas être normale.
```

```
Coefficient de corrélation de Spearman entre l'âge et le panier moyen : -0.0777306903622693
P-valeur : 0.5045034885212738
Il n'y a pas de corrélation significative entre l'âge et le panier moyen
```

# Âge/Catégories

- La statistique de Kruskal-Wallis est de 71359.73 et la p-valeur est de 0.0.
- Ces résultats indiquent une différence significative entre les âges des clients pour les différentes catégories de livres
- En d'autres termes, les âges des clients varient de manière significative en fonction des catégories de livres achetés.




```
Statistique de Shapiro-Wilk sur les résidus : 0.9666518568992615
P-valeur sur les résidus : 0.0
Les résidus ne suivent pas une distribution normale
```

```
Statistique de Kruskal-Wallis : 71359.73412120914
P-valeur : 0.0
Il y a une différence significative entre les âges pour les différentes catégories de livres
```

# Conclusion

- Notre analyse nous a permis de déceler quatre clients professionnels qui contribuent grandement au chiffre d'affaires et qui nécessitent un traitement particulier pour optimiser leurs achats et leur satisfaction.
- Nous devrions envisager progressivement d'abandonner l'exploitation des produits qui ne parviennent pas à générer un chiffre d'affaires satisfaisant, afin de concentrer nos ressources sur nos produits phares et optimiser ainsi notre rentabilité à long terme.
- Nous avons remarqué que les clients plus jeunes ont tendance à dépenser davantage que leurs homologues plus âgés, et cette tendance est également observée dans leur fréquence d'achat. De plus, nous avons constaté que les préférences en termes de catégories de produits varient selon les différentes tranches d'âge, ce qui influe directement sur nos décisions en matière de stratégie marketing.
- Enfin, notre analyse temporelle a mis en évidence des tendances saisonnières dans le chiffre d'affaires, ce qui nous permettra de mieux planifier nos activités commerciales. En intégrant ces informations, nous sommes en mesure d'adapter des stratégies marketing pour maximiser les ventes et la satisfaction client.