Lapage

- Analyse des Ventes -

NGAHA Marie Thérèse

Présenté le 15/07/23

CTI
Connect Technology Institute

SOMMAIRE

I- Explorations et nettoyage des données

1- Explorations et nettoyages des données

a- Customers: qui donne les informations sur les clients;

b- products: qui donne les informations sur les produits;

c- Transactions: qui donne les informations sur les opérations de ventes;

2- fusions des jeux de données, explorations et nettoyages

3- Conclusion partie 1

a- Customers(1/3)

- cette table contient : 8623 lignes et 3 colonnes ,
- il y a 8623 clients en ligne au cours des 2 dernières années
- le fichiers de données ne contient pas de valeurs manquante ni de valeurs dupliquées
- la clé primaire client_id commence par c_xx

```
Entrée [4]: #charger les données
            customers = pd.read_csv('customers.csv')
```

```
Entrée [5]: customers.head()
```

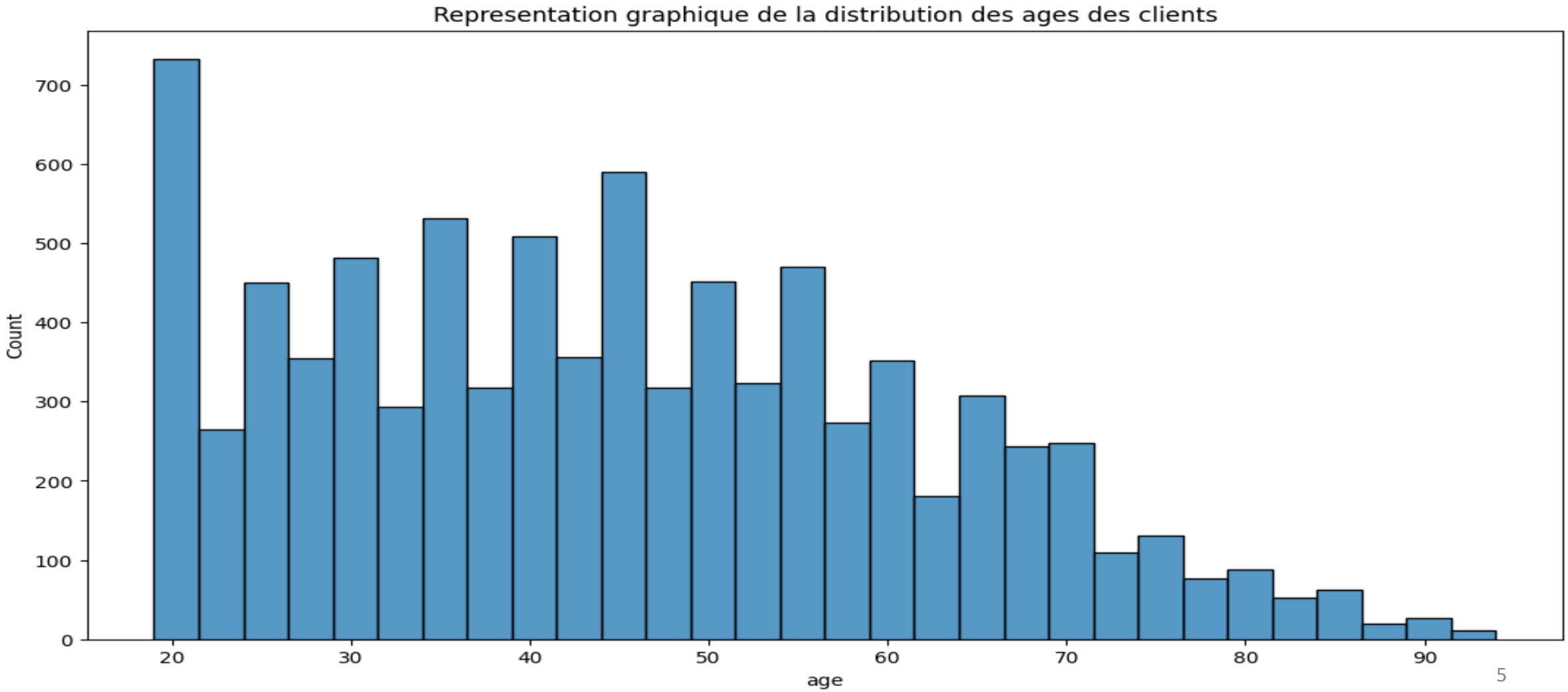
Out[5]:

	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

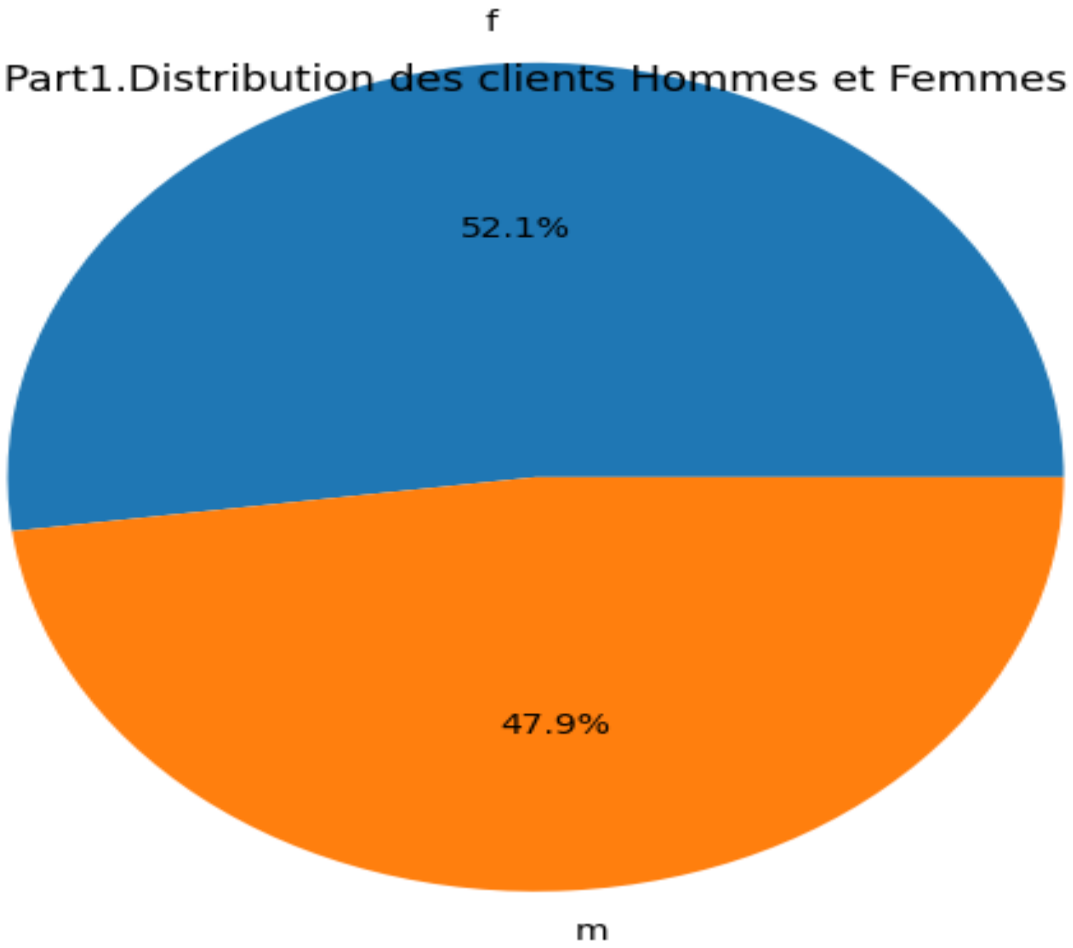
```
Entrée [6]: #nbre de ligne et de colones
            customers.shape
```

```
Out[6]: (8623, 3)
```

a- Customers(2/3)



a- Customers(3/3)



b- products(1/3)

•il y’a 3287 produits

•les fichiers ne contient pas de valeurs manquante ni de doublons

•la clé primaire "id_prod" commence par "categ_xxx«

ée [16]:
```
products.head()
```

Out[16]:

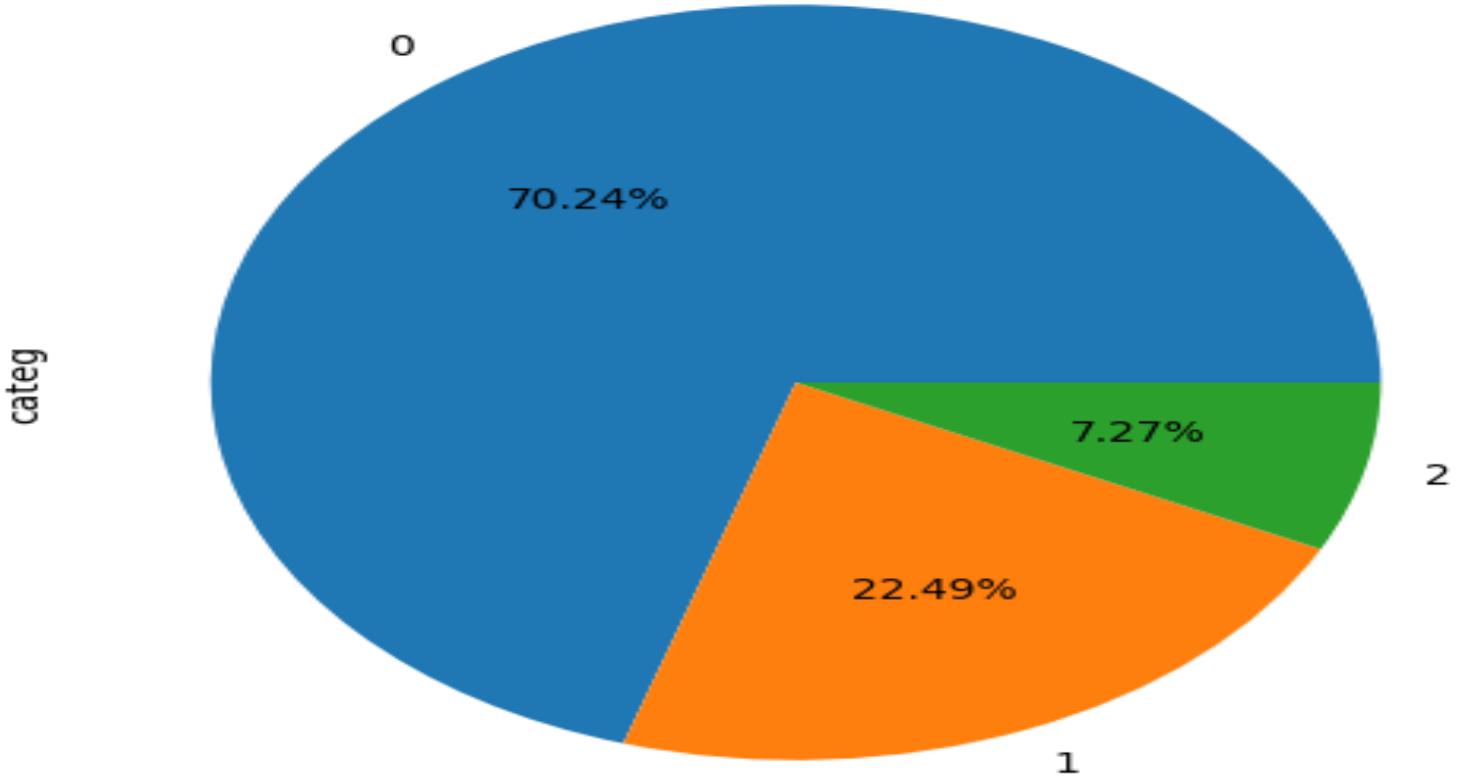
	id_prod	price	categ
0	0_1421	19.99	0
1	0_1368	5.13	0
2	0_731	17.99	0
3	1_587	4.99	1
4	0_1507	3.99	0

ée [18]:
```
#nbre de colones
products.shape
```

Out[18]: (3287, 3)

b- products(2/3)

Distribution de catégories



b- products(3/3)
Distribution des prix

	count	mean	std	min	25%	50%	75%	max
price	3286.0	21.863597	29.849786	0.62	6.99	13.075	22.99	300.0

b- Transactions(1/3)

- il y’a 4 variables dans l'enssemble de données chaque ligne représente une vente transaction a une date données
- il y’a 679532 transactions enregistrées
- les données ne contiennent aucune valeurs manquante

```
27]: transactions.head()
```

27]:

	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

```
28]: #nbres de lignes et de colones
transactions.shape
```

```
28]: (679532, 4)
```

b- Transactions(1/3)

les clients ct_1 et ct_0 sont des clients de test qui ont effectuées des transactions sur des produits de test (T_0) à une date de test (test_2021-03-01 02:30:02) lors d'une sessions de test S_0

Entrée [30]:
```
# rechercher et suprimées les valeurs dupliqués dans le data
doublons = transactions[transactions.duplicated()]
doublons
```

Out[30]:

	id_prod	date	session_id	client_id
27778	T_0	test_2021-03-01 02:30:02.237437	s_0	ct_1
52424	T_0	test_2021-03-01 02:30:02.237419	s_0	ct_0
96687	T_0	test_2021-03-01 02:30:02.237412	s_0	ct_1
130188	T_0	test_2021-03-01 02:30:02.237419	s_0	ct_0
139339	T_0	test_2021-03-01 02:30:02.237443	s_0	ct_1
...	...	...	...	...
653098	T_0	test_2021-03-01 02:30:02.237432	s_0	ct_0
657830	T_0	test_2021-03-01 02:30:02.237417	s_0	ct_0
662081	T_0	test_2021-03-01 02:30:02.237427	s_0	ct_1
671647	T_0	test_2021-03-01 02:30:02.237424	s_0	ct_1
679180	T_0	test_2021-03-01 02:30:02.237425	s_0	ct_1

126 rows × 4 columns

2- fusions des jeux de données, explorations et nettoyages(1/3)

4. Merge Lapage dataframes, exploration et nettoyages

née [37]:
```
#merge des dataframes
lapage_df = transactions.merge(customers, how='left', on='client_id').merge(products, how='left', on='id_prod')
display(lapage_df.head())
```

	id_prod	time	session_id	client_id	date	year	period	sex	birth	age	price	categ
0	0_1518	2022-05-20 13:21:29.043970	s_211425	c_103	2022-05-20	2022	2022-05	f	1986	37	4.18	0
1	1_251	2022-02-02 07:55:19.149409	s_158752	c_8534	2022-02-02	2022	2022-02	m	1988	35	15.99	1
2	0_1277	2022-06-18 15:44:33.155329	s_225667	c_6714	2022-06-18	2022	2022-06	f	1968	55	7.99	0
3	2_209	2021-06-24 04:19:29.835891	s_52962	c_6941	2021-06-24	2021	2021-06	m	2000	23	69.99	2
4	0_1509	2023-01-11 08:22:08.194479	s_325227	c_4232	2023-01-11	2023	2023-01	m	1980	43	4.99	0

née [38]:
```
# exploration du resultat final
lapage_df.shape
```

Out[38]: (679332, 12)

2- fusions des jeux de données, explorations et nettoyages(2/3)

•le produit '0_2245'n'est pas répertorie dans le data produits

•le produits '0_2245' a été vendu 221 fois

•le produit '0_2245' commence par 0_xxxx donc appartient à la catégories 0

Entrée [40]:
```
# valeurs manquantes
check_nan(lapage_df, "price")
check_nan(lapage_df, "categ")
```

```
La colone price contient  221  valeurs manquantes
La colone categ contient  221  valeurs manquantes
```

Entrée [42]:
```
# recherche
mask_nan = (lapage_df['price'].isnull()) | (lapage_df['categ'].isnull())
print("les valeurs manquantes concernent ces produits  :", lapage_df[mask_nan]["id_prod"].unique())
```

```
Missing values concern those products : ['0_2245']
```

Entrée [43]:
```
# reherche de l'id '0_2245' dans  le date produit
products[products["id_prod"]=="0_2245" ]
```

Out[43]:

id_prod	price	categ

2- fusions des jeux de données, explorations et nettoyages(3/3)

Entrée [44]:

```
# nettoyer le resultat final
categ_0_mean = round(lapage_df[lapage_df["categ"]==0]["price"].mean(),2)#categ 0  price

lapage_df.loc[lapage_df["id_prod"]=="0_2245", "price"] = categ_0_mean#categ 0  price remplacer

lapage_df.loc[lapage_df["id_prod"]=="0_2245", "categ"] = 0#categ 0  categ

lapage_df[lapage_df["id_prod"]=="0_2245"].head()#final
```

Out[44]:

	id_prod	time	session_id	client_id	date	year	period	sex	birth	age	price	categ
2633	0_2245	2022-09-23 07:22:38.636773	s_272266	c_4746	2022-09-23	2022	2022-09	m	1940	83	10.64	0
10103	0_2245	2022-07-23 09:24:14.133889	s_242482	c_6713	2022-07-23	2022	2022-07	f	1963	60	10.64	0
11723	0_2245	2022-12-03 03:26:35.696673	s_306338	c_5108	2022-12-03	2022	2022-12	m	1978	45	10.64	0
15670	0_2245	2021-08-16 11:33:25.481411	s_76493	c_1391	2021-08-16	2021	2021-08	m	1991	32	10.64	0
16372	0_2245	2022-07-16 05:53:01.627491	s_239078	c_7954	2022-07-16	2022	2022-07	m	1973	50	10.64	0

CONCLUSION(partie 1)

Bilan des modifications apportées :

- les variables age, année et période ont été rajoutées au jeu de données.
- T_0 qui est un produit test a été écarté de l'analyse (soit 200 transactions concernant les tests)
- 0_2245 n'étant pas repertorié dans le fichier products, son prix (NaN) a été remplacé par le prix moyen de la catégorie 0.
- 157 outliers ont été détectés sur la varible price.
- la colonne contenant la variable date est convertie dans le bon format.

Analyse

1- Chiffre d’affaires

-- CA par catégorie de produits

-- B. CA par tranche d'âge

-- C. Tendance globale dans le temps et évolution du CA

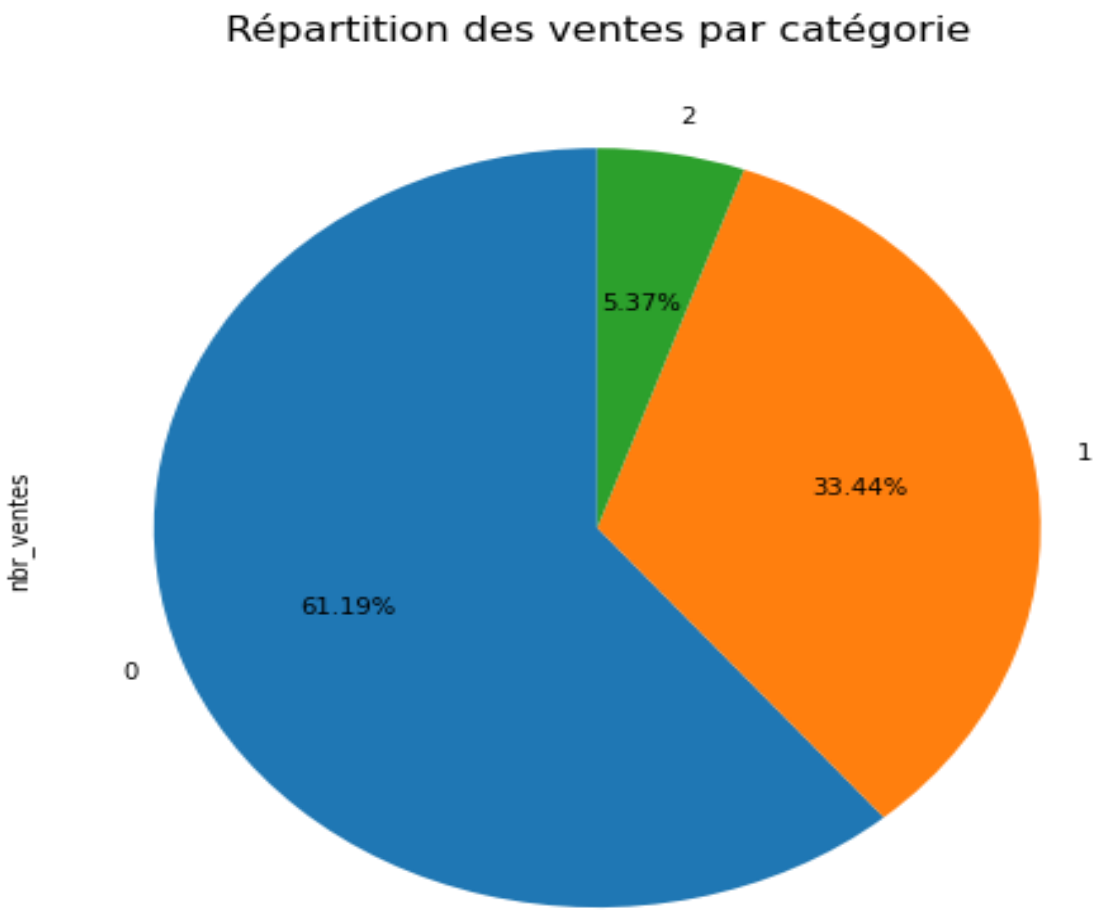
2. Profil client et top/flop des ventes

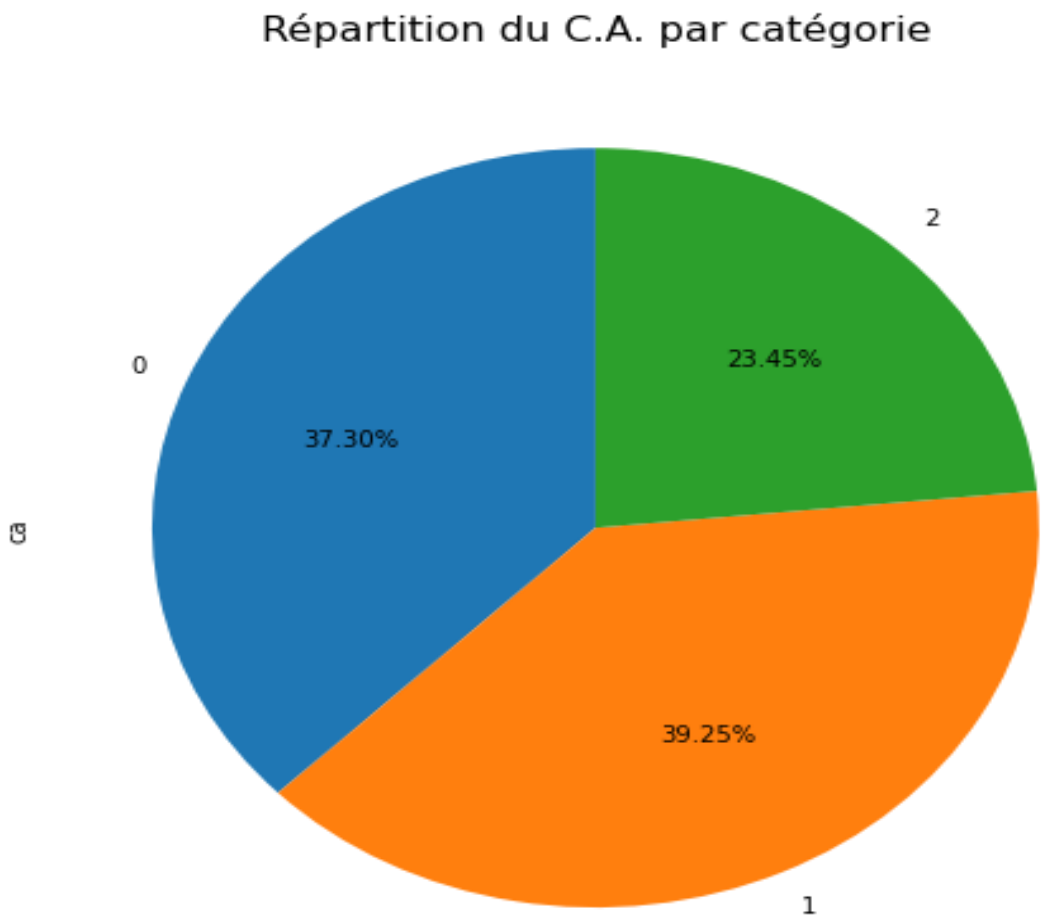
-- A. Profil client

-- B. Top/flop du nombre de ventes

1- Chiffre d’affaires(1/)

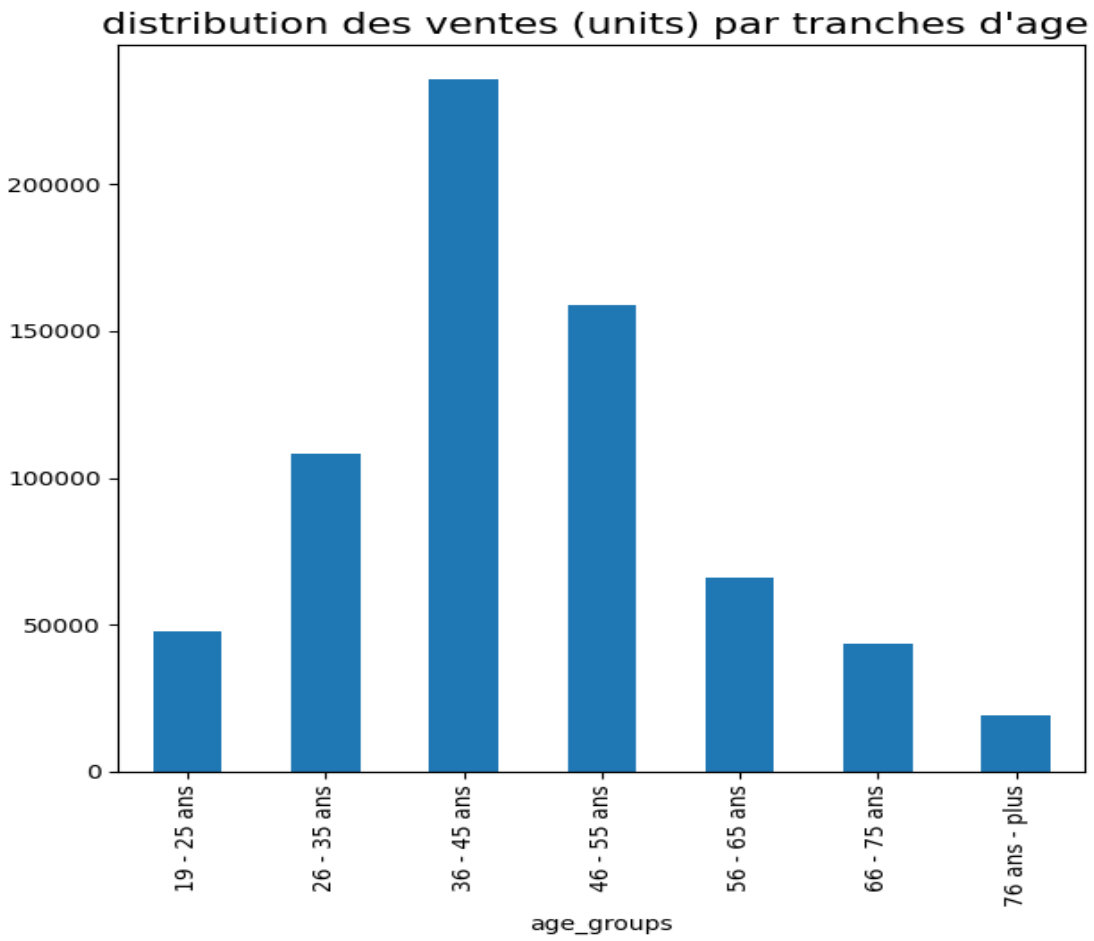
-- CA par catégorie de produits

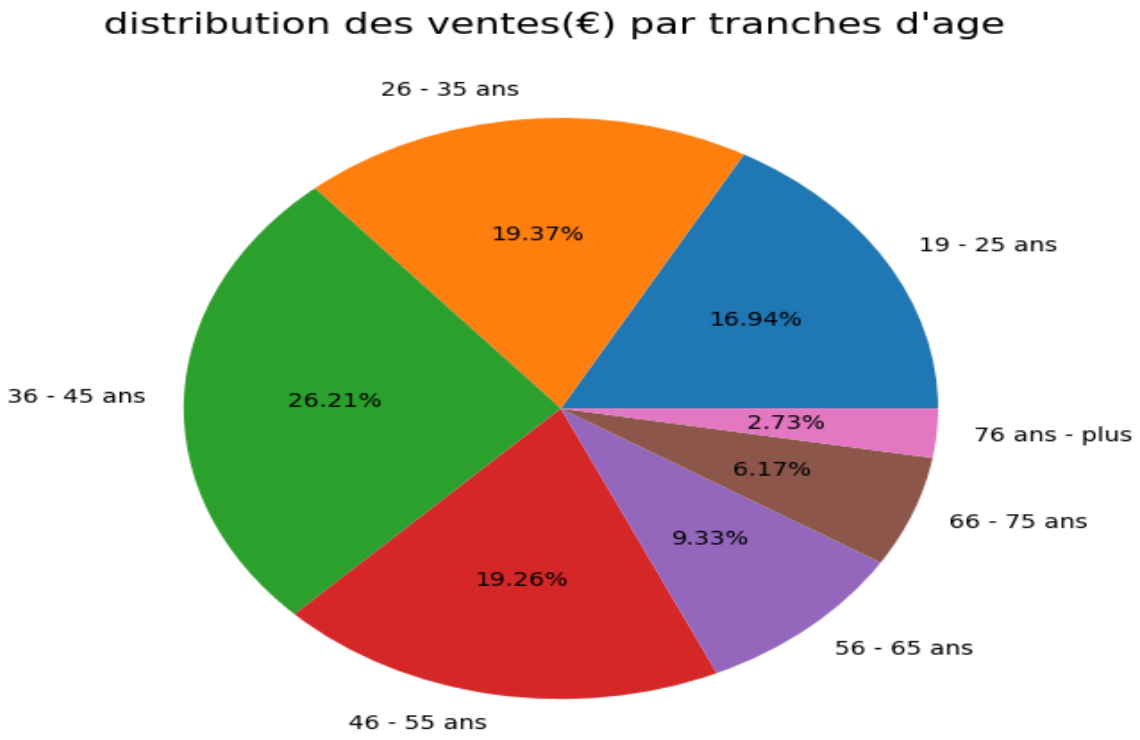




1- Chiffre d’affaires(2/)

B. CA par tranche d'âge

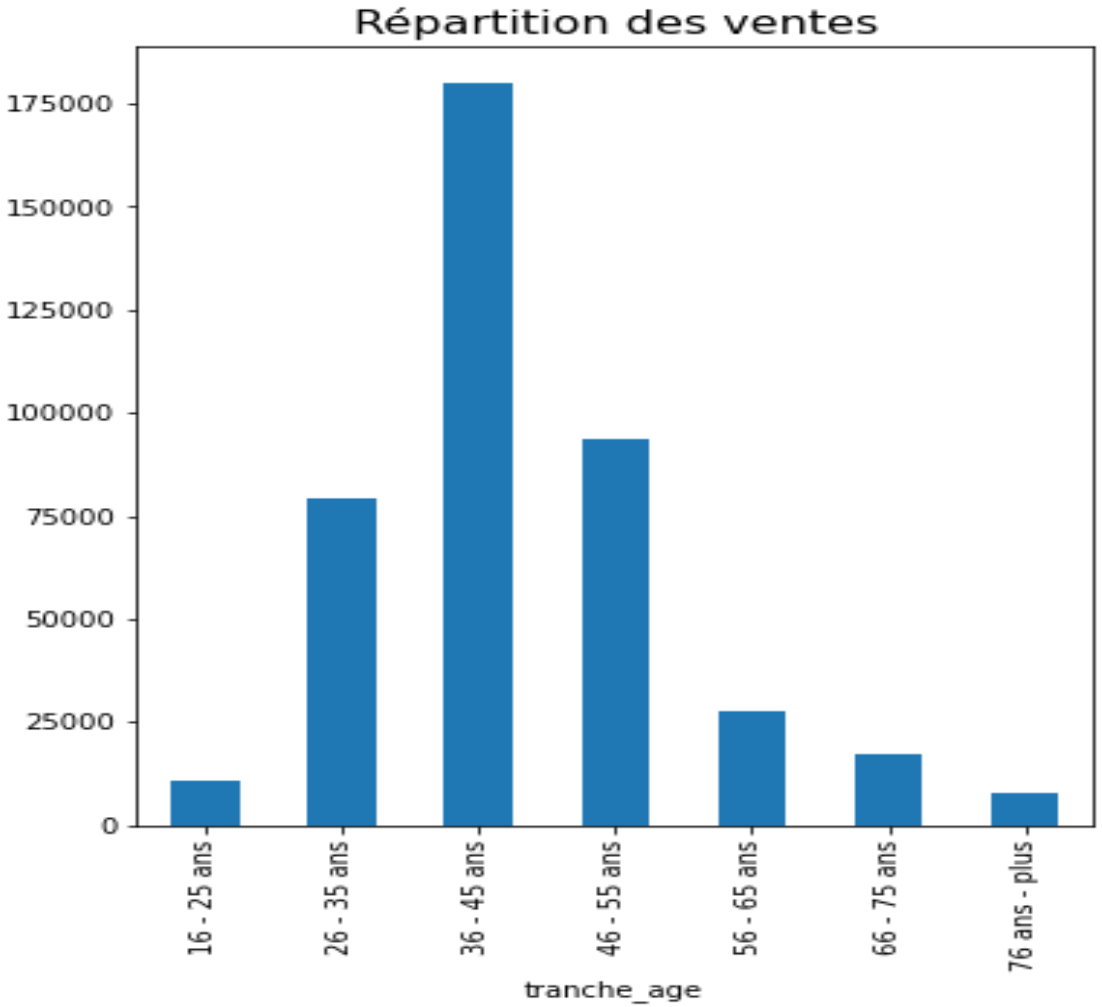


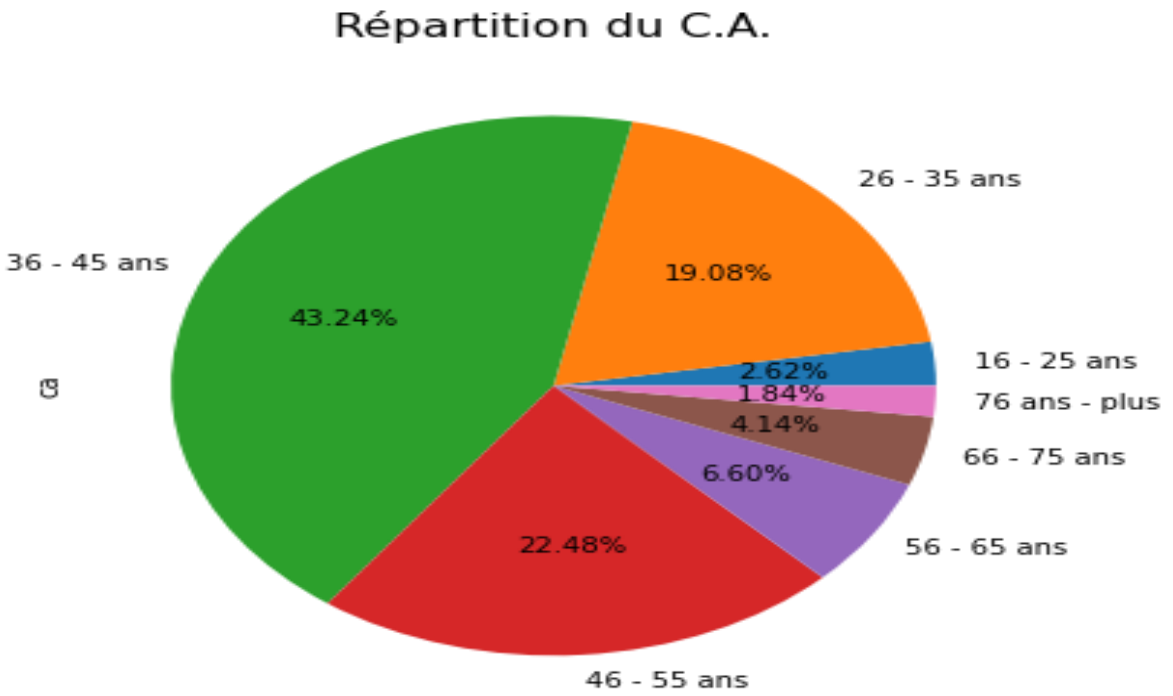


1- Chiffre d’affaires(3/)

B. CA par tranche d'âge pour chaque catégories

Categ_0

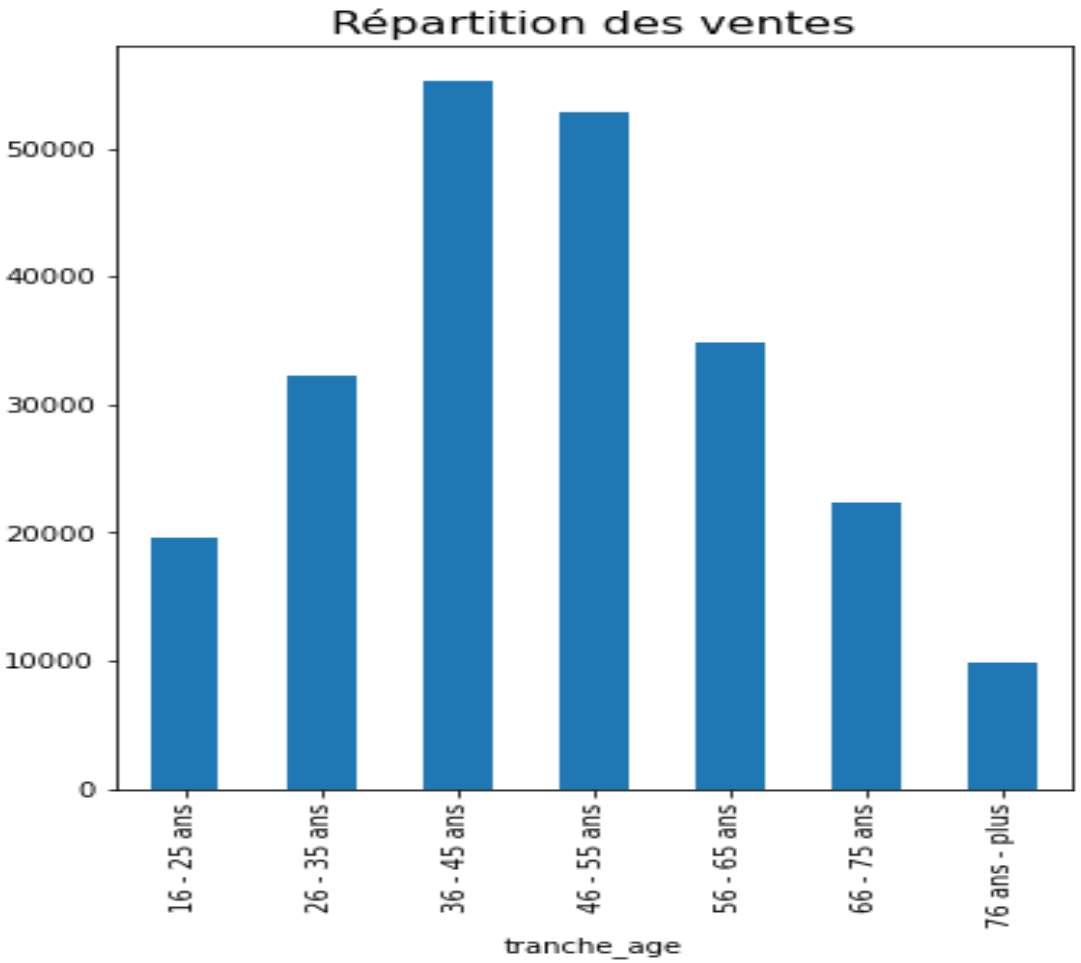


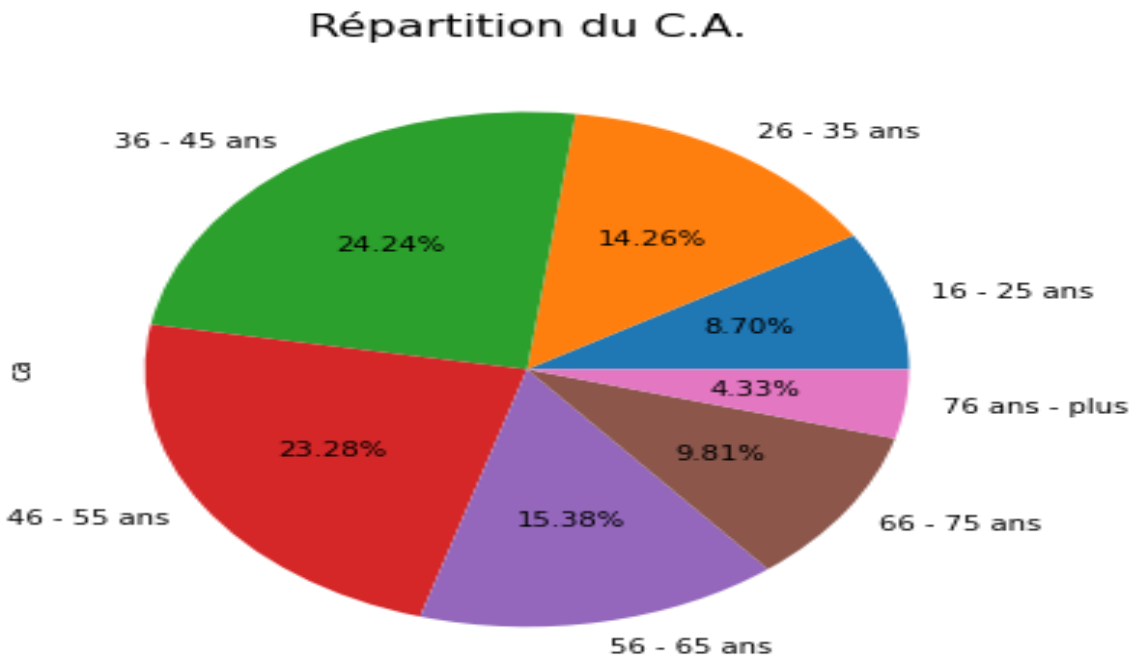


1- Chiffre d’affaires(4/)

B. CA par tranche d'âge pour chaque catégories

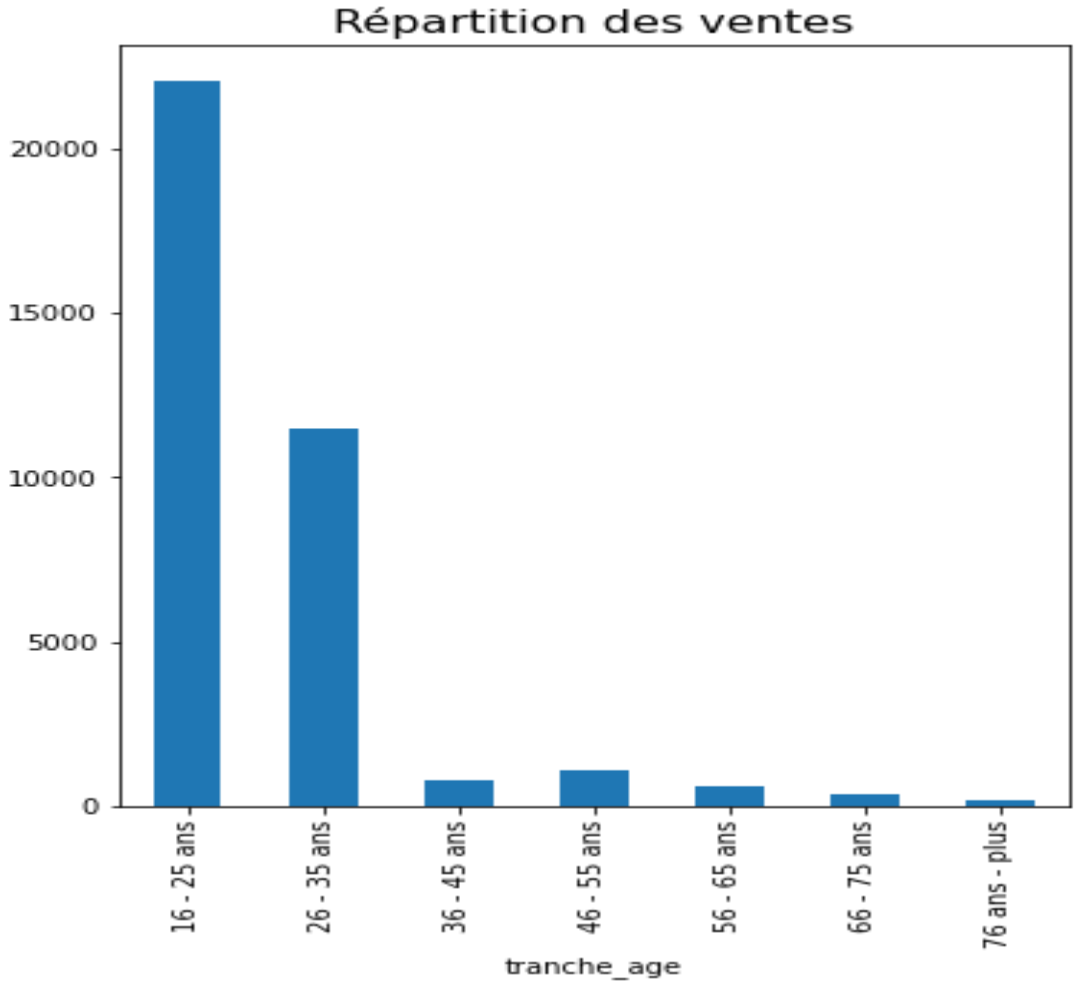
Categ_1

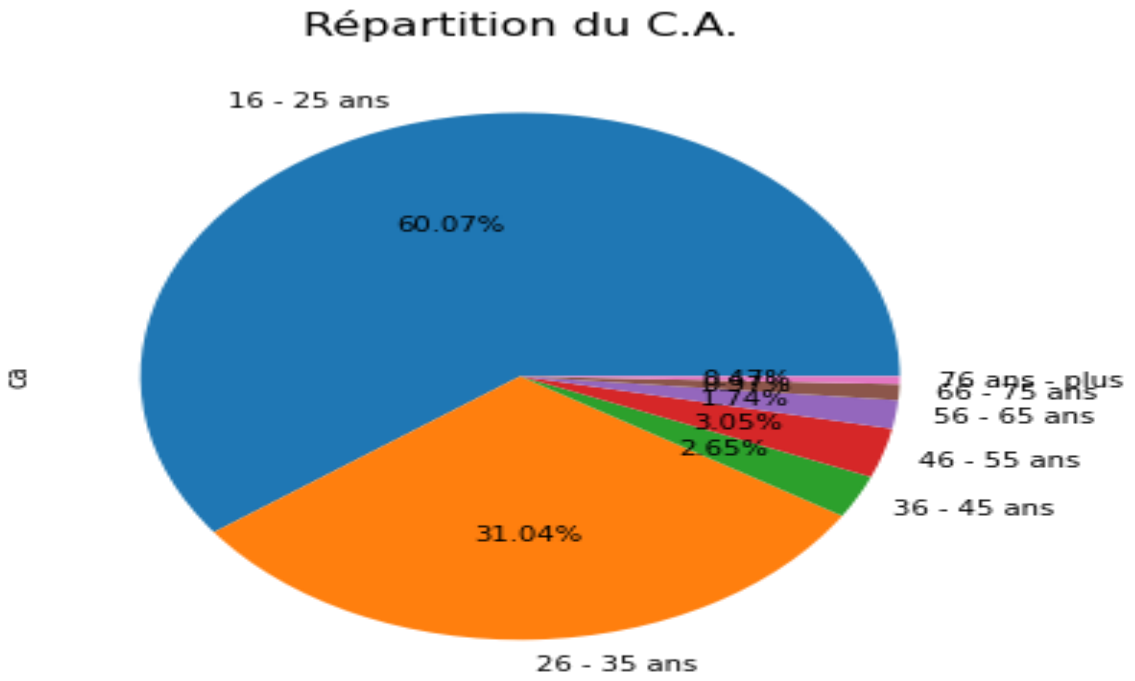




1- Chiffre d’affaires(5/)

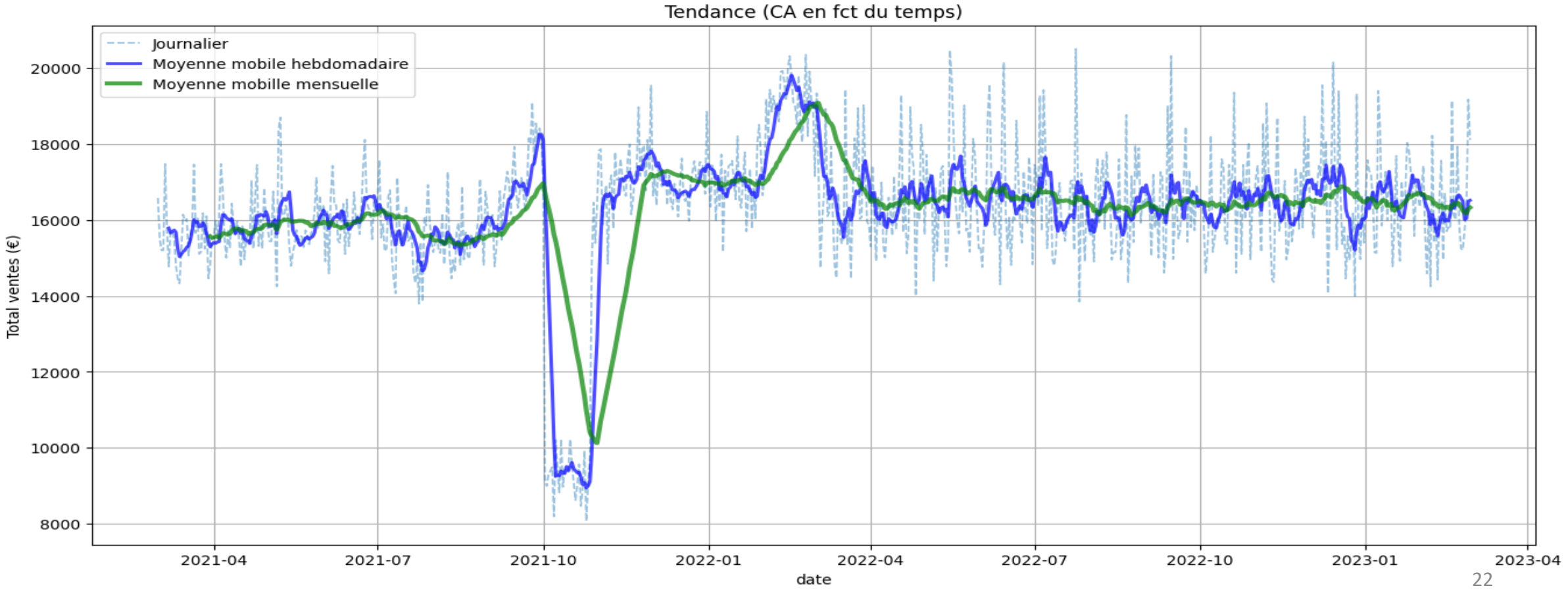
B. CA par tranche d'âge pour chaque catégories Categ_2





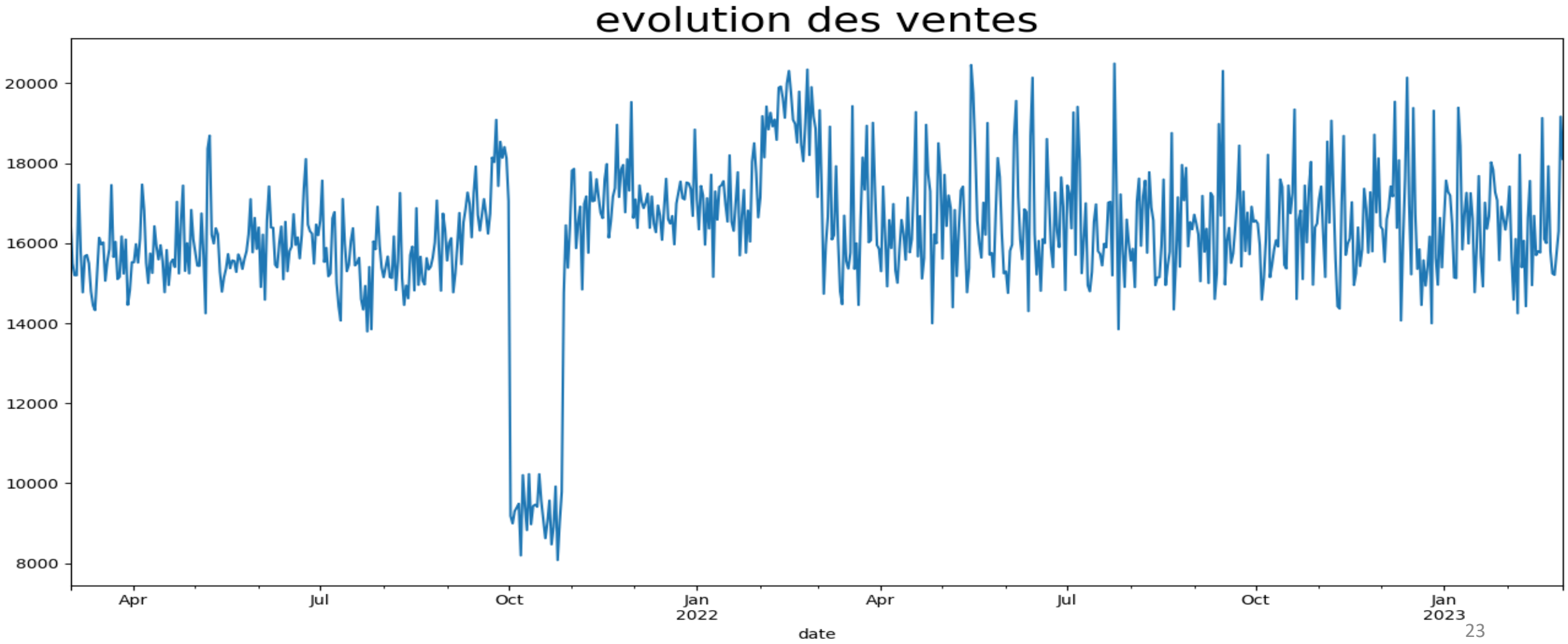
1- Chiffre d'affaires(6/)

-- C. Tendance globale dans le temps et évolution du CA



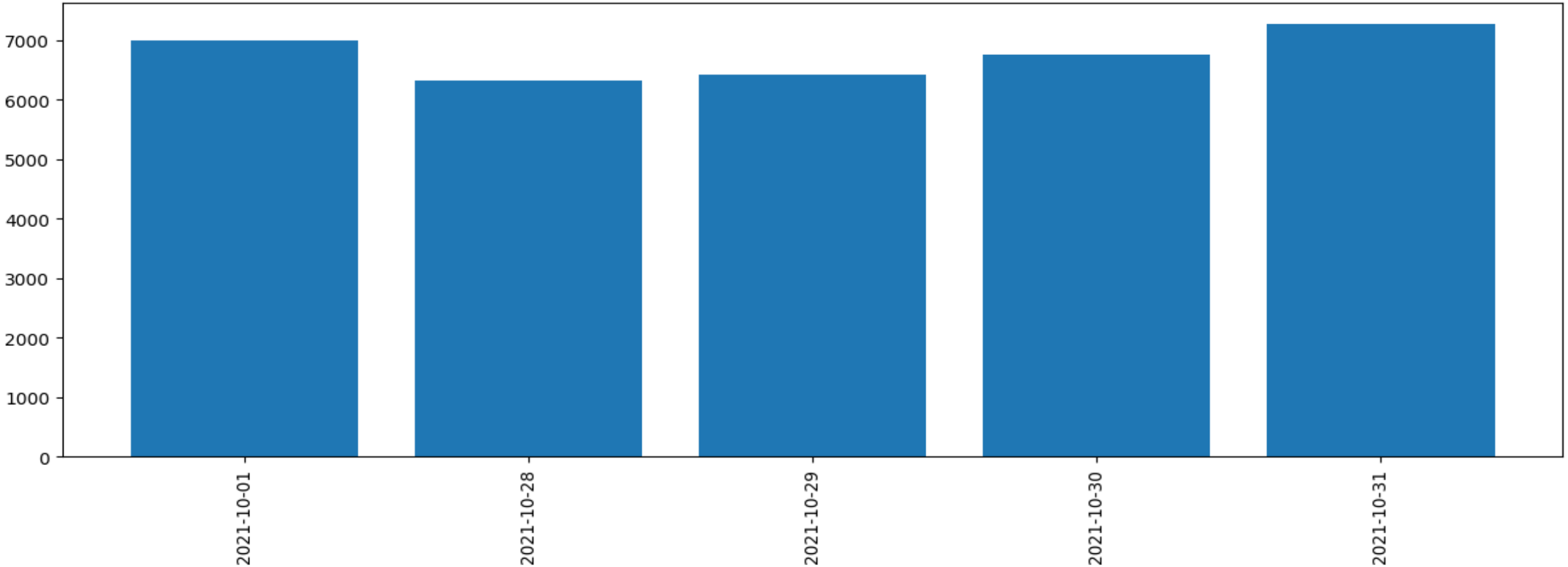
1- Chiffre d’affaires(7/7)

-- C. Tendance globale dans le temps et évolution du CA



1- Chiffre d’affaires(7/7)

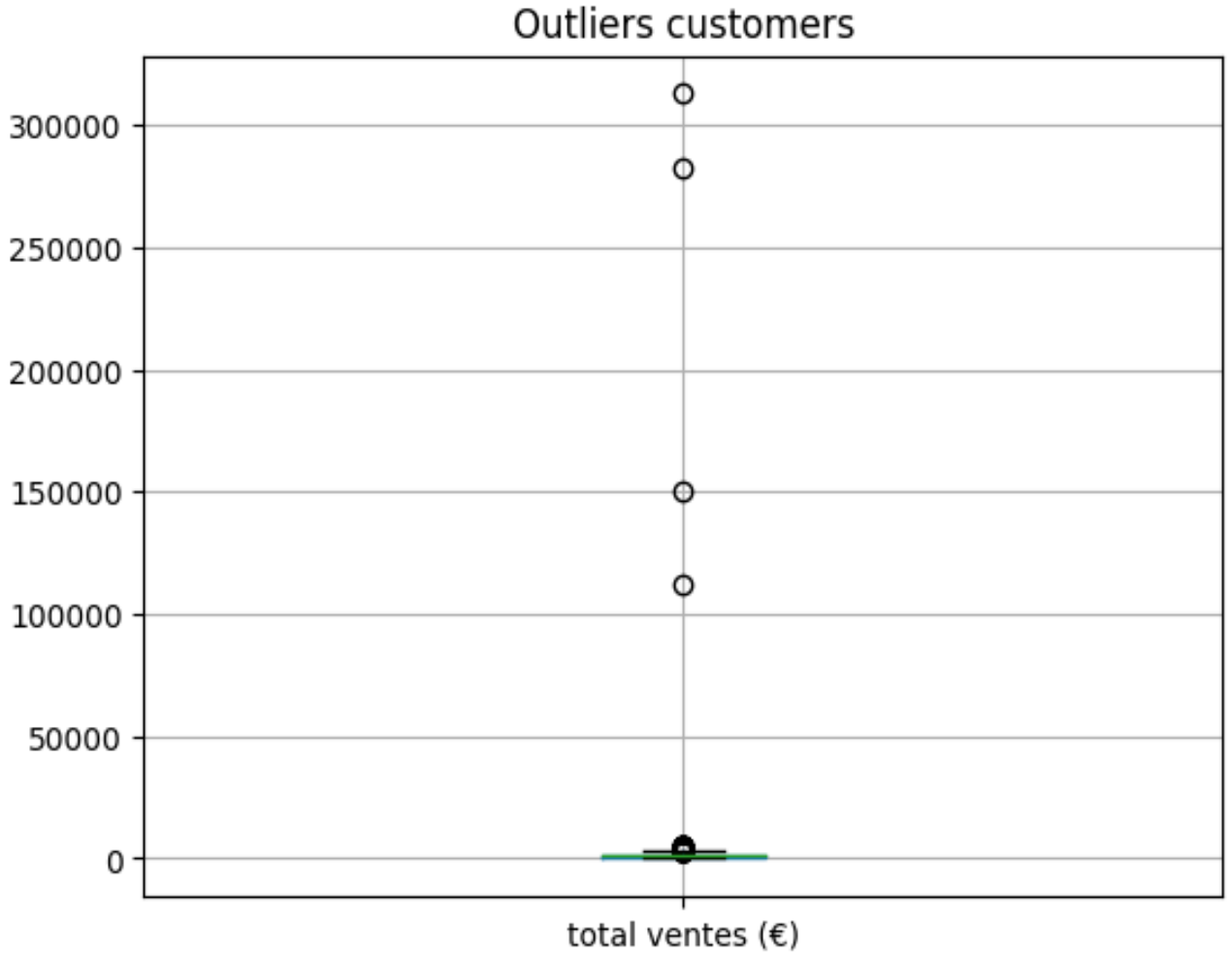
CA par âge par période, date et catégorie en octobre 2021/ categ_1



2. Profil client et top/flop des ventes(1/)

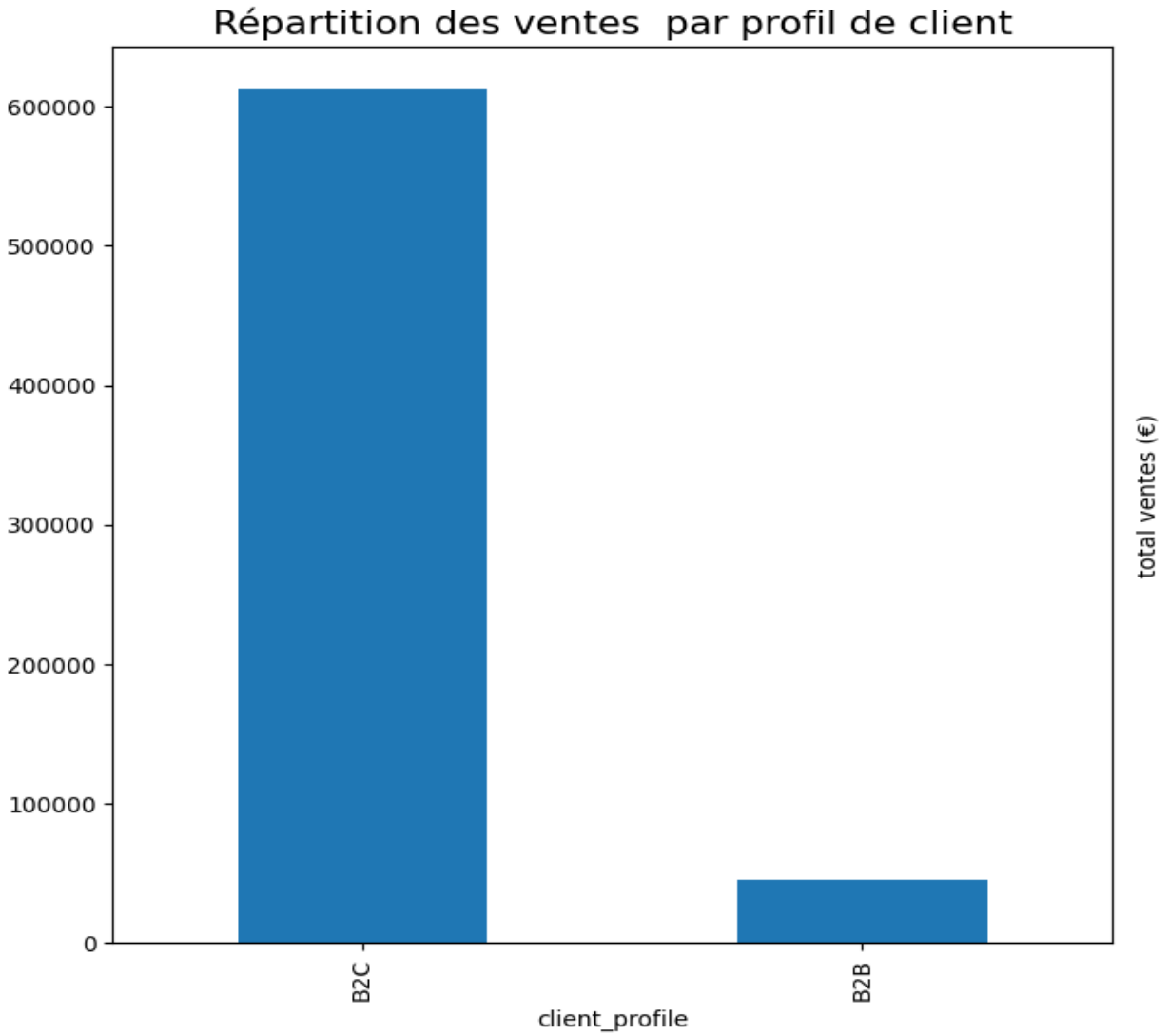
-- A. Profil client

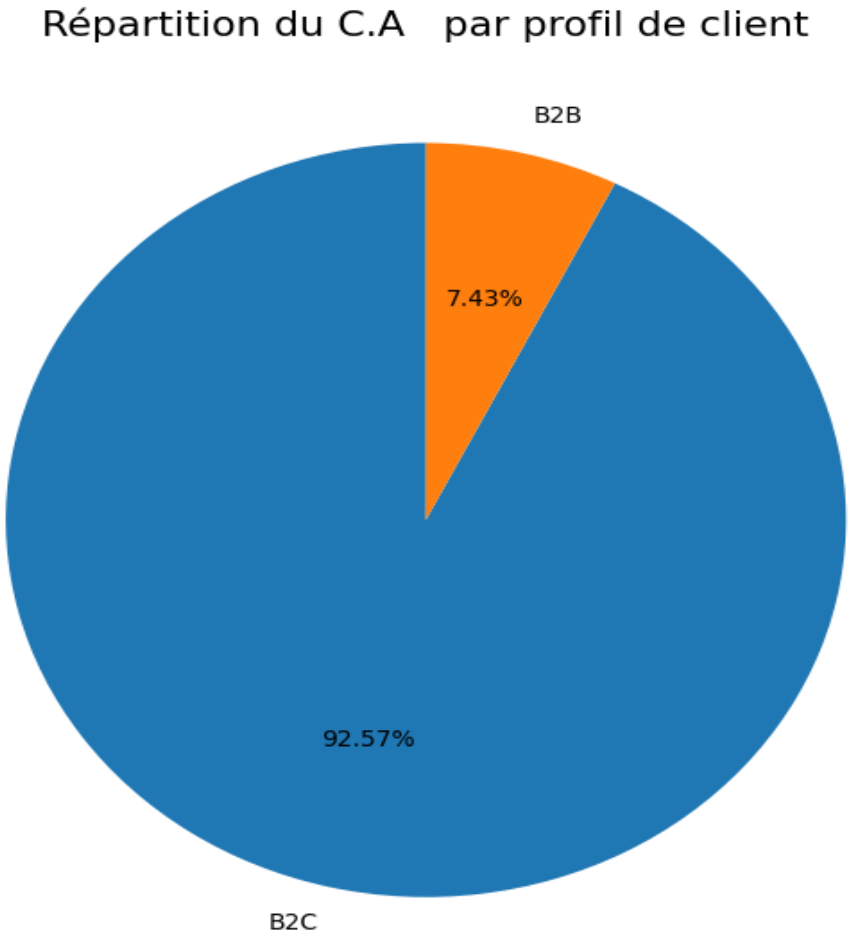
client_id	total ventes (€)	fréquence d'achat	client_profile
c_1609	312755.08	24472	B2B
c_4958	282654.61	5090	B2B
c_6714	149847.59	8903	B2B
c_3454	111798.63	6635	B2B
c_2899	5214.05	105	B2C



2. Profil client et top/flop des ventes(1/)

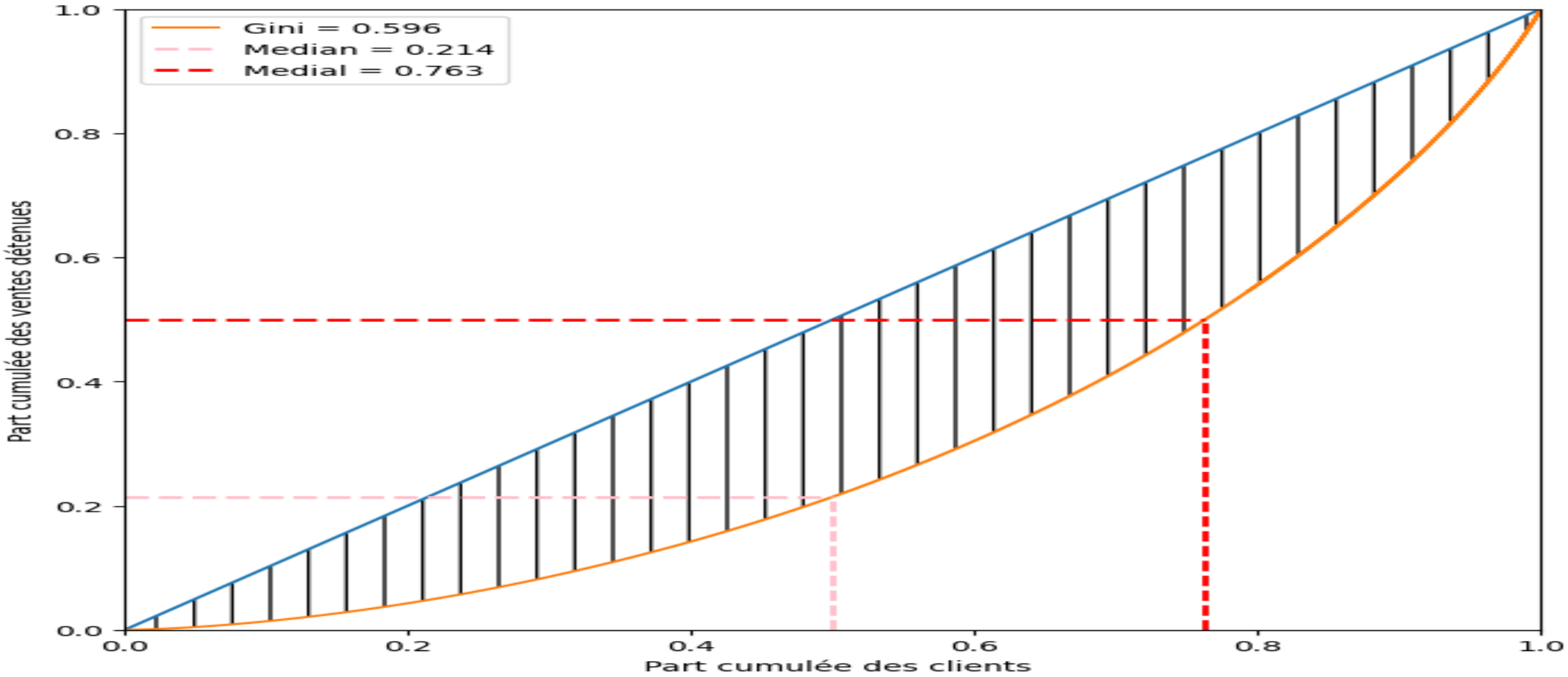
-- Répartition des ventes par profil clients





2. Profil client et top/flop des ventes(1/)

Courbe de lorenz



Conclusion (partie 2)

- Evolution des ventes et données enregistrées :

  Les ventes en ligne ont atteint un pic de plus de 20 000 € par jour en février 2022

- Aucune vente n'a été enregistrée du 02 au 27 octobre 2021 sur la catégorie 1, nous avons donc dû supprimer le mois d'octobre 2021 dans la base de données pour éviter les moyennes biaisées
- L'évolution des ventes totales est plutôt stable dans le temps pour chaque catégorie de produits
- **Profil client :**

- Il y a 4 clients B2B qui représentent 7% des ventes totales pour seulement 0,05% des clients totaux
- La première moitié des clients B2C détient 21% des ventes, et donc l'autre moitié détient 79% des ventes : la concentration des ventes est assez inégale entre les clients
- La répartition des ventes est la même quel que soit le sexe du client avec une consommation un peu plus élevée chez les femmes : elle croît initialement et atteint son maximum dans la tranche d'âge 36-45 ans, puis elle décroît avec l'âge

III- CORRELATIONS

- *Confert librairie_3*

CONCLUSIONS (partie 3)

- **Relation entre le sexe du client et la catégorie d'achat :**
- Les femmes comme les hommes préfèrent respectivement les catégories 0, 1 et 2 enfin
- Les femmes consomment un peu plus que les hommes sur les 3 catégories de produits
- Pour $\alpha = 5\%$, le sexe des clients dépend de la catégorie de livres achetés
- **Relation entre l'âge des clients et la fréquence d'achat :**
- D'un point de vue général, la fréquence des achats diminue avec l'âge
- La fréquence d'achat et l'âge des clients ne suivent pas une loi normale
- L'âge des clients et leur fréquence d'achat sont négativement corrélés, ce qui signifie ici : plus le client est âgé, moins il achète

CONCLUSIONS (partie 3)

- **Relation entre l'âge des clients et le montant total des achats :**
- D'un point de vue général, le montant total d'achat diminue avec l'âge (même s'il est positif jusqu'à 40 ans)
- Le montant total de l'achat ne suit pas une loi normale
- L'âge des clients et leur montant total d'achat sont très négativement corrélés, ce qui signifie ici : plus le client est âgé, moins il achète
- **Relation entre l'âge des clients et le panier moyen :**
- Nous avons suffisamment de preuves pour dire que les deux ensembles de données d'échantillon ne proviennent pas de la même distribution.
- Le panier moyen ne suit pas une loi normale
- La relation entre l'âge des clients et le panier moyen n'est pas significative

CONCLUSIONS (partie 3)

- **Relation âge des clients vs catégorie de livres achetés :**
- Pour α = 5%, la catégorie de livres achetés dépend de l'âge des clients
- Catégorie 0 : Clients âgés de 19 à 67 ans (avec une moyenne d'âge approximative de 45 ans)
- Catégorie 1 : Tout le monde (avec une moyenne d'âge approximative de 48 ans)
- Catégorie 2 : Clients âgés de 19 à 40 ans (avec une moyenne d'âge approximative de 27 ans)