Lapage

Contexte

- La Page : Librairie physique & site web depuis 2 ans

- Entrée dans l'équipe Marketing

- **Objectifs** : analyse des points forts, des points faibles, comportements clients, etc.

Sommaire

- Missions :

1) Demandes d'Antoine :
   Analyse des indicateurs

2) Demandes de Julie :
   Analyse des corrélations

- Ressources


taset clients


taset produits


taset transactions

- Outils

Langage : Pytho 

Logiciel : Jupyter 

Ressources

- Dataset clients 
- Dataset produits 
- Dataset transactions 

Outils

- Langage de programmation : Python

- Logiciel : Jupyter

Etapes de l'étude

1. Préparation des données
   - Exploration des données
   - Nettoyage des données
   - Merge
   - Enrichissement des données
2. Analyse des indicateurs
   - Demandes d'Antoine
3. Analyse des corrélation
   - Demande de Julie

Préparation des données

Préparation des données

Exploration du fichier clients

	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
...	...	...	...
8618	c_7920	m	1956
8619	c_7403	f	1970
8620	c_5119	m	1974
8621	c_5643	f	1968
8622	c_84	f	1982

✔ 8 623 clients tous différents

✔ Pas de doublons

✔ Pas de valeurs manquantes ou nulles

- Deux clients test

Préparation des données

Exploration du fichier produits

	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
...	...	...	...
3282	2_23	115.99	2
3283	0_146	17.14	0
3284	0_802	11.22	0
3285	1_140	38.56	1
3286	0_1920	25.16	0

✔ 2 281 produits tous différents

✔ 3 catégories

✔ Pas de doublon

✔ Pas de valeurs manquantes ou nulle

- Une valeur aberrante : « -1 » qui correspond à un produit test

Préparation des données

Exploration du fichier transactions

	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
...	...	...	...	...
679527	0_1551	2022-01-15 13:05:06.246925	s_150195	c_8489
679528	1_639	2022-03-19 16:03:23.429229	s_181434	c_4370
679529	0_1425	2022-12-20 04:33:37.584749	s_314704	c_304
679530	0_1994	2021-07-16 20:36:35.350579	s_63204	c_2227
679531	1_523	2022-09-28 01:12:01.973763	s_274568	c_3873

- ✔ 679 532 transactions
- ✔ Pas de valeurs manquantes ou nulles
- 183 lignes doublon (tests)
- Valeurs aberrantes (test_ , T_ , ct_)

Préparation des données

Nettoyage des données

	client_id	sex	birth
2735	ct_0	f	2001
8494	ct_1	m	2001

	id_prod	price	categ
731	T_0	-1.0	0

	id_prod	date	session_id	client_id
3019	T_0	test_2021-03-01 02:30:02.237419	s_0	ct_0
5138	T_0	test_2021-03-01 02:30:02.237425	s_0	ct_0
9668	T_0	test_2021-03-01 02:30:02.237437	s_0	ct_1
10728	T_0	test_2021-03-01 02:30:02.237436	s_0	ct_0
15292	T_0	test_2021-03-01 02:30:02.237430	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
670680	T_0	test_2021-03-01 02:30:02.237449	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

Suppression de tout les tests

- clients test (-2 lignes)
- produits test (-1 ligne)
- transaction test(-200 lignes)

Préparation des données

Merge

```
dataV1 = pd.merge(dataproducts, datatransactions, how="inner", on=["id_prod"])
dataV1
```

	id_prod	price	categ	date	session_id	client_id
0	0_1421	19.99	0	2022-02-20	s_168213	c_6389
1	0_1421	19.99	0	2022-11-19	s_299590	c_8364
2	0_1421	19.99	0	2021-09-19	s_92304	c_3544
3	0_1421	19.99	0	2023-01-11	s_325369	c_1025
4	0_1421	19.99	0	2021-08-01	s_70071	c_2298
...	...	...	...	...	...	...
679106	1_140	38.56	1	2022-06-30	s_231391	c_974
679107	0_1920	25.16	0	2023-01-30	s_334324	c_7748
679108	0_1920	25.16	0	2021-04-13	s_20115	c_7088
679109	0_1920	25.16	0	2021-05-30	s_41465	c_7748
679110	0_1920	25.16	0	2022-12-30	s_319303	c_7748

```
data = pd.merge(dataV1, datacustomers, how="inner", on=["client_id"])
data
```

	id_prod	price	categ	date	session_id	client_id	sex	birth
0	0_1421	19.99	0	2022-02-20	s_168213	c_6389	f	1991
1	0_1421	19.99	0	2022-10-20	s_285450	c_6389	f	1991
2	0_2131	8.99	0	2021-10-09	s_102458	c_6389	f	1991
3	0_1635	16.99	0	2021-04-28	s_26841	c_6389	f	1991
4	0_166	1.83	0	2021-07-15	s_62585	c_6389	f	1991
...	...	...	...	...	...	...	...	...
679106	2_163	68.99	2	2022-01-28	s_156517	c_7739	m	1997
679107	2_101	63.99	2	2021-07-21	s_65192	c_7089	m	2002
679108	2_101	63.99	2	2022-12-21	s_315267	c_7089	m	2002
679109	2_101	63.99	2	2022-10-21	s_285788	c_7089	m	2002
679110	2_101	63.99	2	2022-03-21	s_182240	c_7089	m	2002

Jointures finales : inner

- Ne pas avoir de valeurs nulles
- Etudes réalistes

Tests de jointure

- 21 produits non vendus
- 21 clients inactifs
- Ajout du produit 0_2245 non répertorier dans le fichier produit

Préparation des données

Enrichissement des données

	id_prod	price	categ	date	session_id	client_id	sex	birth	age	Groupe
679110	2_101	63.99	2	2022-03-21	s_182240	c_7089	m	2002	21	- 30 ans
546625	1_267	27.99	1	2021-08-28	s_81773	c_2831	m	2003	20	- 30 ans
546626	1_267	27.99	1	2023-01-28	s_333573	c_2831	m	2003	20	- 30 ans
546627	1_267	27.99	1	2022-09-28	s_274723	c_2831	m	2003	20	- 30 ans
546628	0_1479	15.99	0	2021-12-21	s_137668	c_2831	m	2003	20	- 30 ans
...	...	...	...	...	...	...	...	...	...	...
319972	1_376	17.49	1	2021-05-17	s_35555	c_5366	f	1941	82	+ 70 ans
319971	1_621	17.99	1	2021-12-09	s_131910	c_5366	f	1941	82	+ 70 ans
319970	1_621	17.99	1	2022-12-09	s_309595	c_5366	f	1941	82	+ 70 ans
319978	1_529	17.99	1	2021-11-12	s_118477	c_5366	f	1941	82	+ 70 ans
552662	0_1097	11.99	0	2022-06-05	s_219128	c_799	f	1945	78	+ 70 ans

Ajout je 2 variables

- Ajout de l'âge
- Ajout du groupe d'âge

Analyse des indicateurs : Chiffre d'affaire

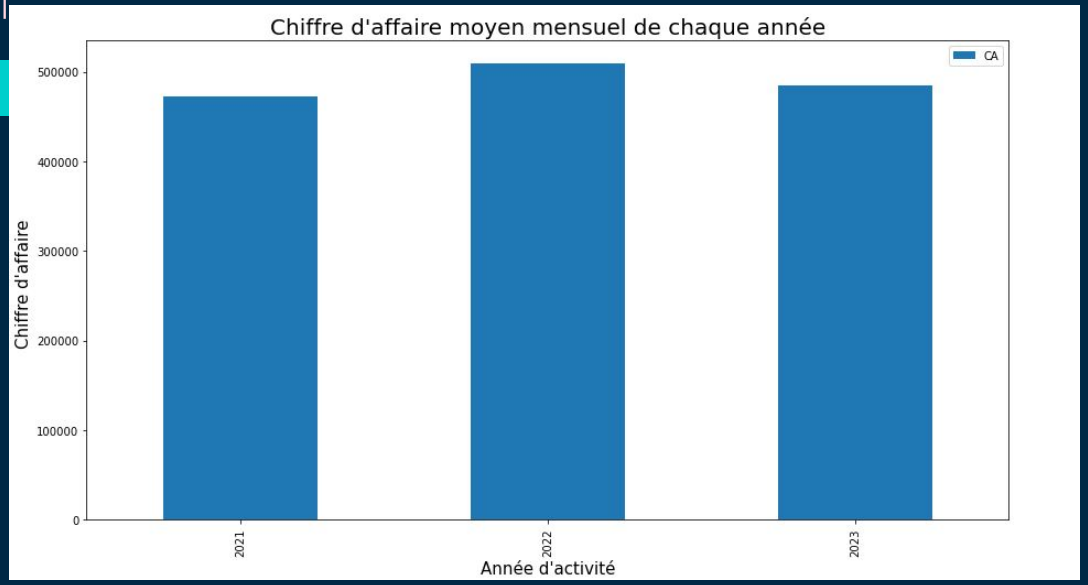
Analyse de données : Demandes d'Antoine

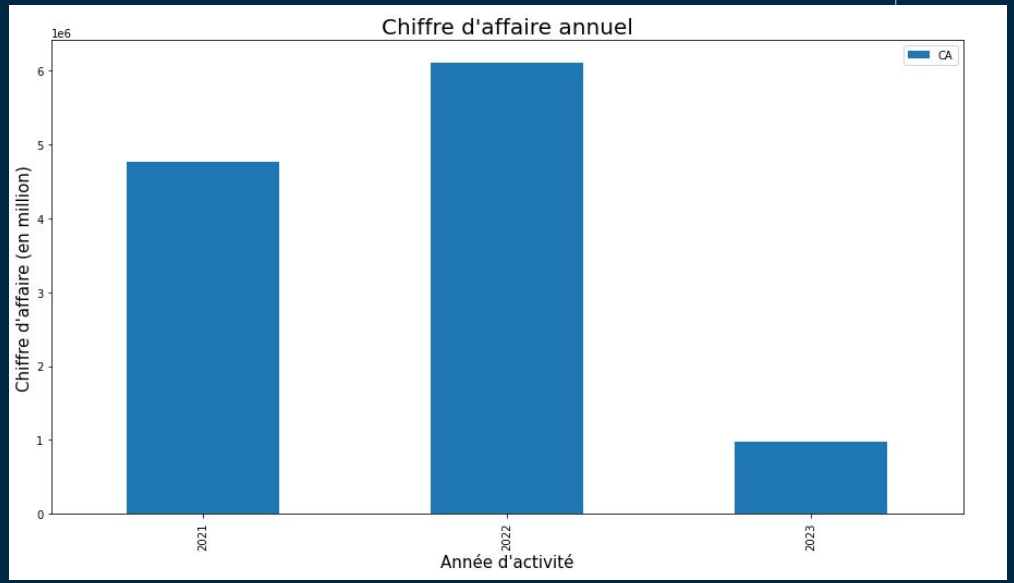
Etude du chiffre d'affaire

2022

- Meilleur chiffre d'affaires moyen mensuel

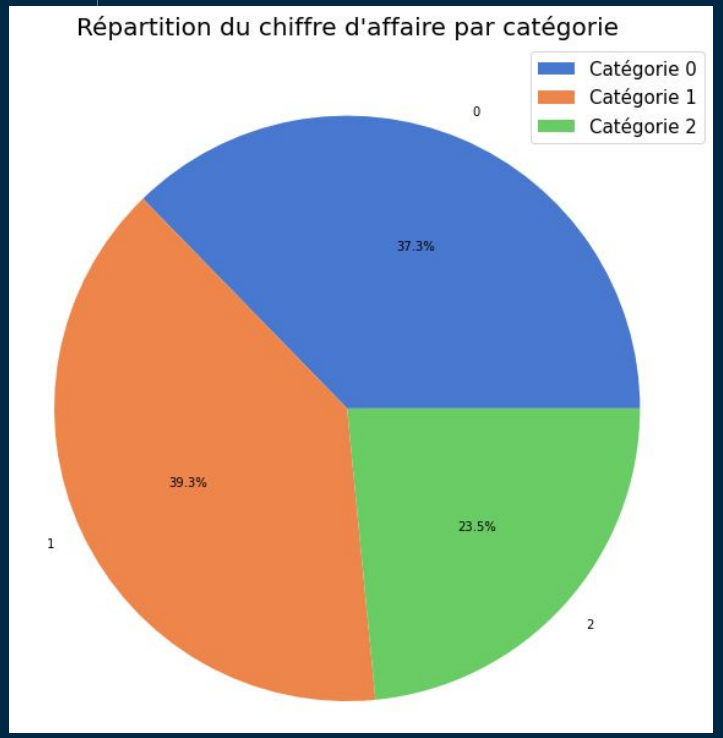
- Meilleur chiffre d'affaires annuel



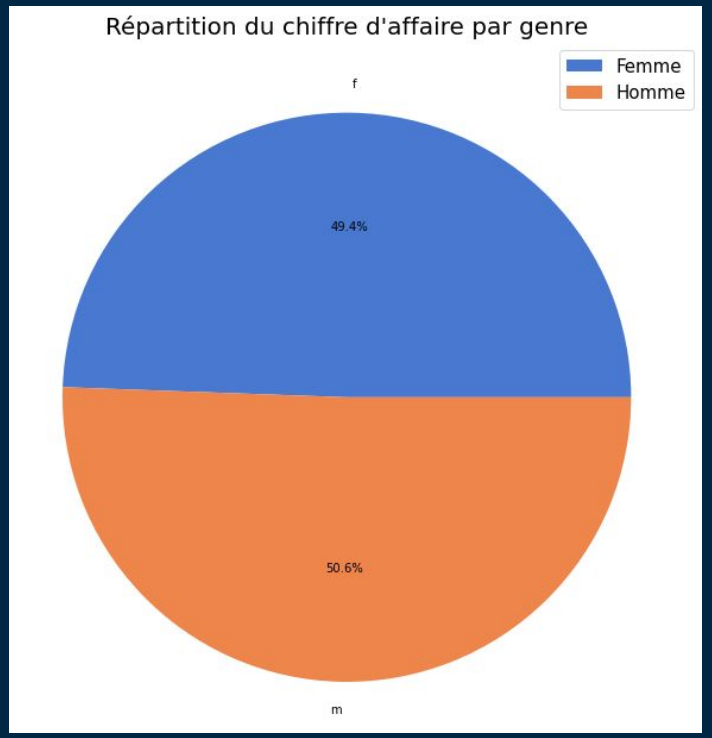


Analyse de données : Demandes d'Antoine

Etude du chiffre d'affaire



Les catégories 0 et 1 sont les plus achetés par les clients !



Les femmes et les hommes génèrent quasiment le même chiffre d'affaires !

Top 1 : Catégorie 1

Top 2 : Catégorie 0

Top 3 : Catégorie 2

Femme : 49,4 %

Homme : 50,6 %

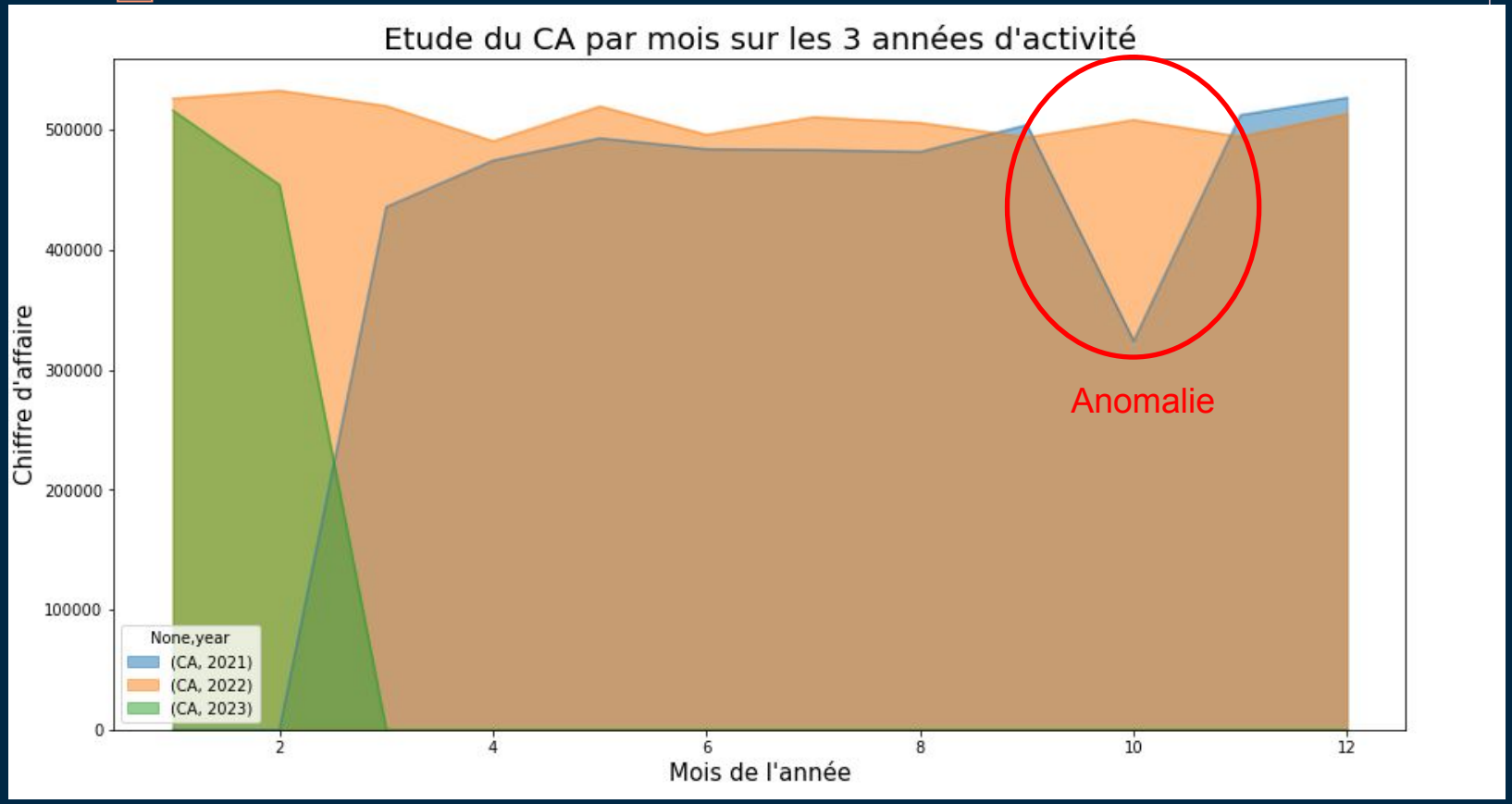
Analyse de données : Demandes d'Antoine

Etude du chiffre d'affaire

- 2021 : Augmentation continue du chiffre d'affaires (sans compter l'anomalie)

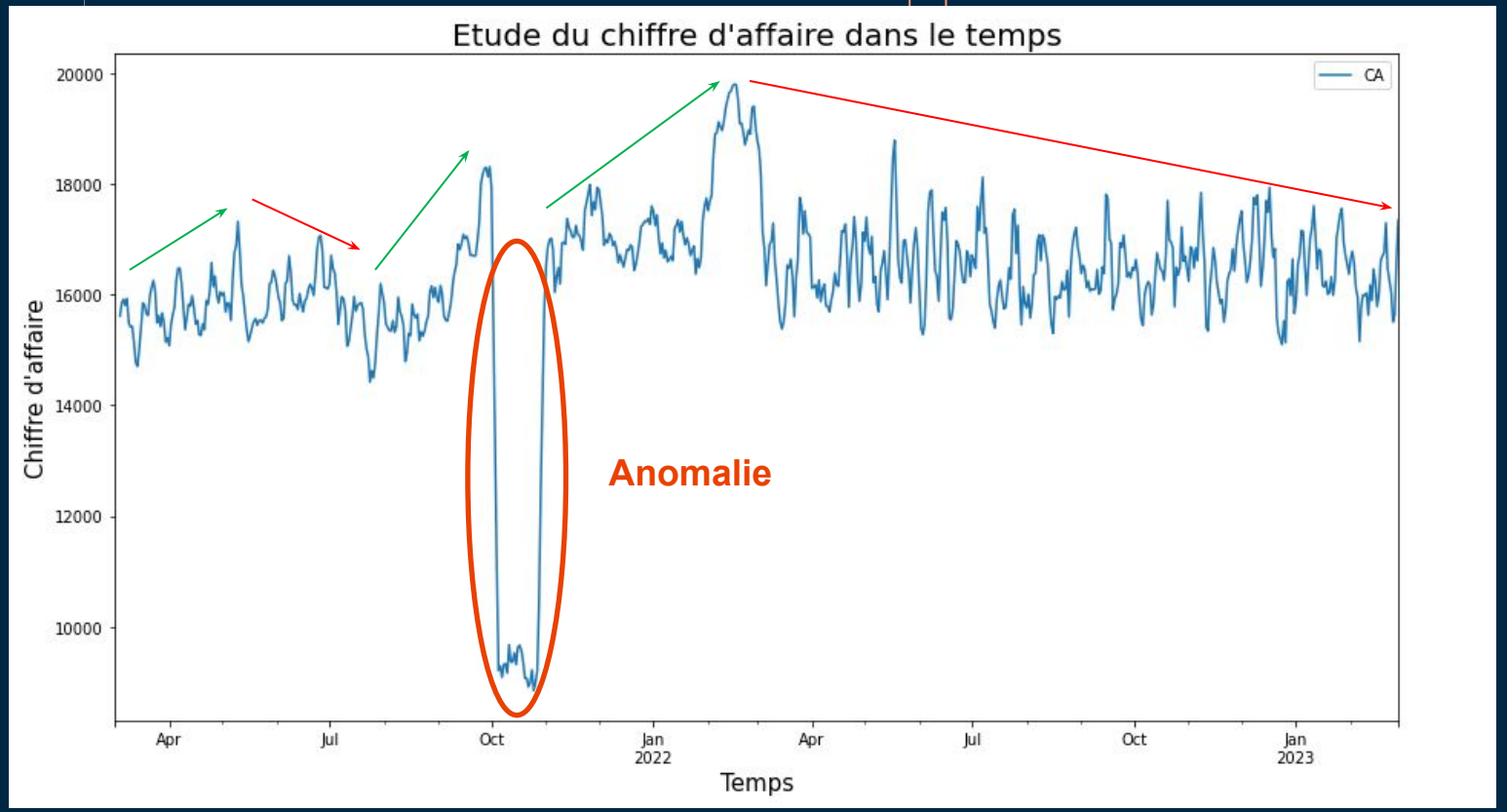
- 2022 : Stagnation du chiffre d'affaires

- 2023 : Grosse baisse significative au mois de février



Analyse de données : Demandes d'Antoine

Etude du chiffre d'affaire



Chiffre d'affaires total sur toute la période étudier

11 853 728 €

✔ Tendance générale à la hausse du début à la moitié de la période
(Sans compté l'exception)

- Anomalie : manque de données du mois d'octobre 2021

- Légère tendance à la baisse

Analyse de données : Demandes d'Antoine

Etude du chiffre d'affaire

	client_id	CA
677	c_1609	324033.35
4388	c_4958	289760.34
6337	c_6714	153670.92
2724	c_3454	113673.93

4 clients professionnels

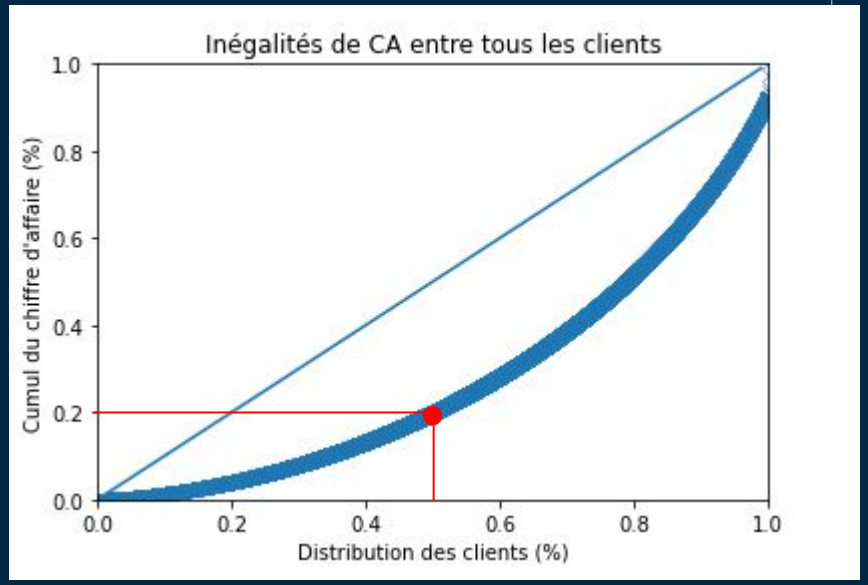
+ 800 000 € de CA

7,43 % du CA

Indice de Gini : 0,45

Indice de Gini assez fort

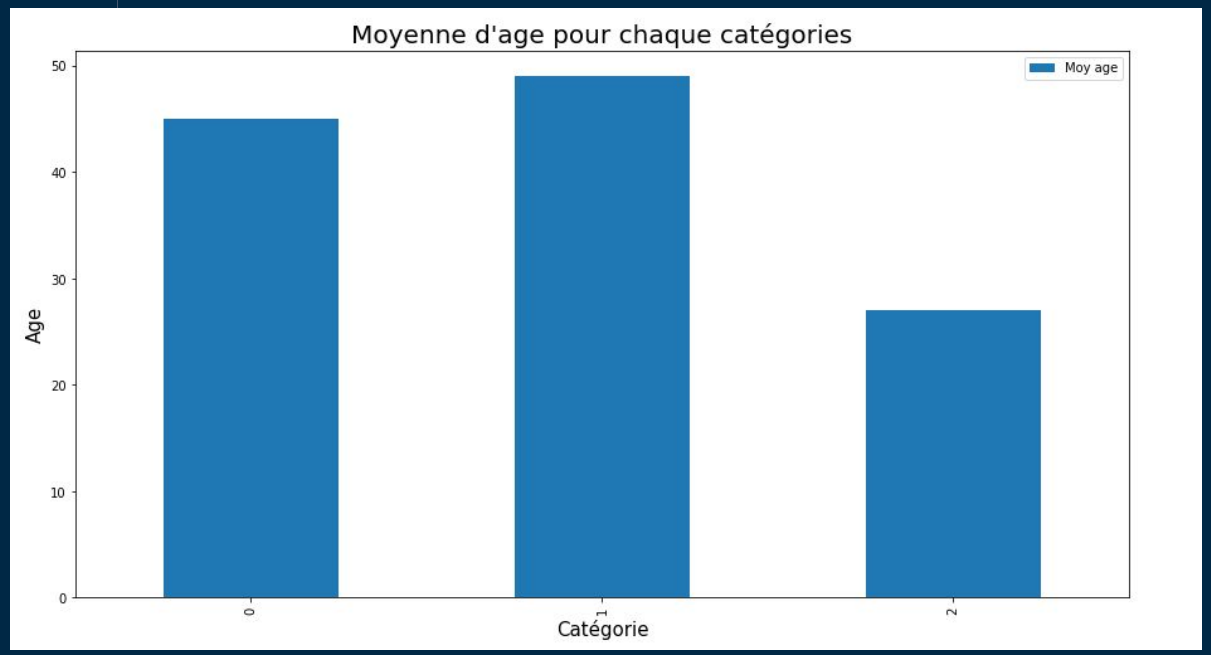
50% des clients génère 20% du chiffre d'affaires



Analyse des indicateurs : Profils clients

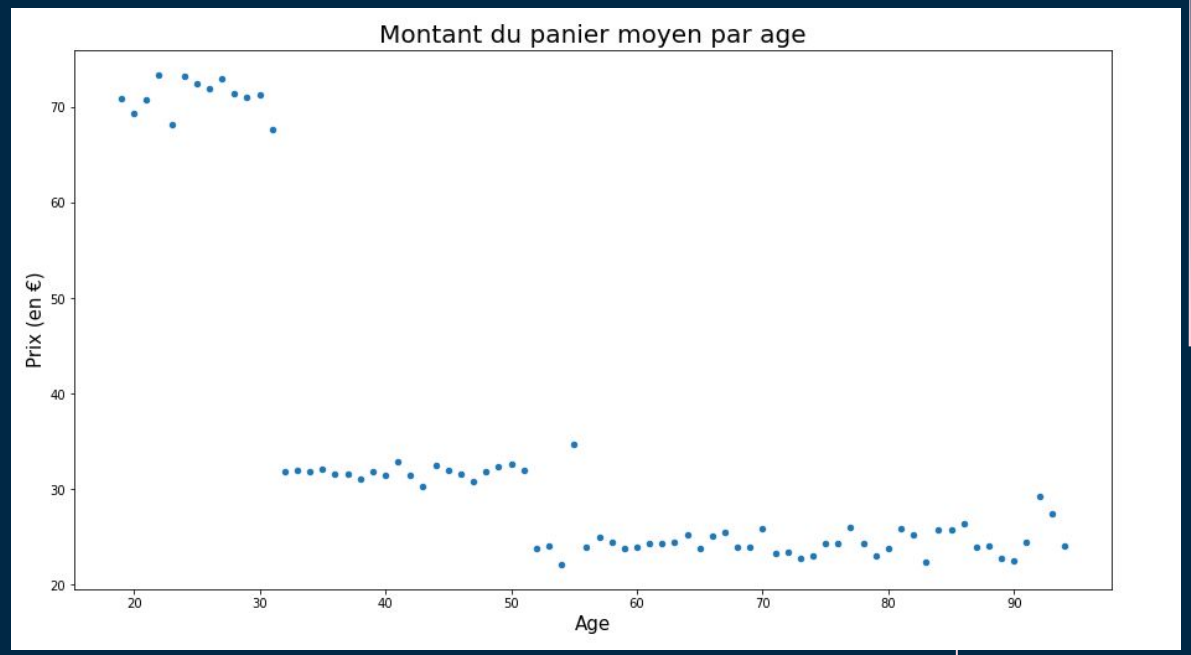
Analyse de données : Demandes d'Antoine

Etude des profils client



- Catégorie 0 : 45 ans
- Catégorie 1 : 49 ans
- Catégorie 2 : 27 ans

La catégorie 2 intéresse les plus jeunes !

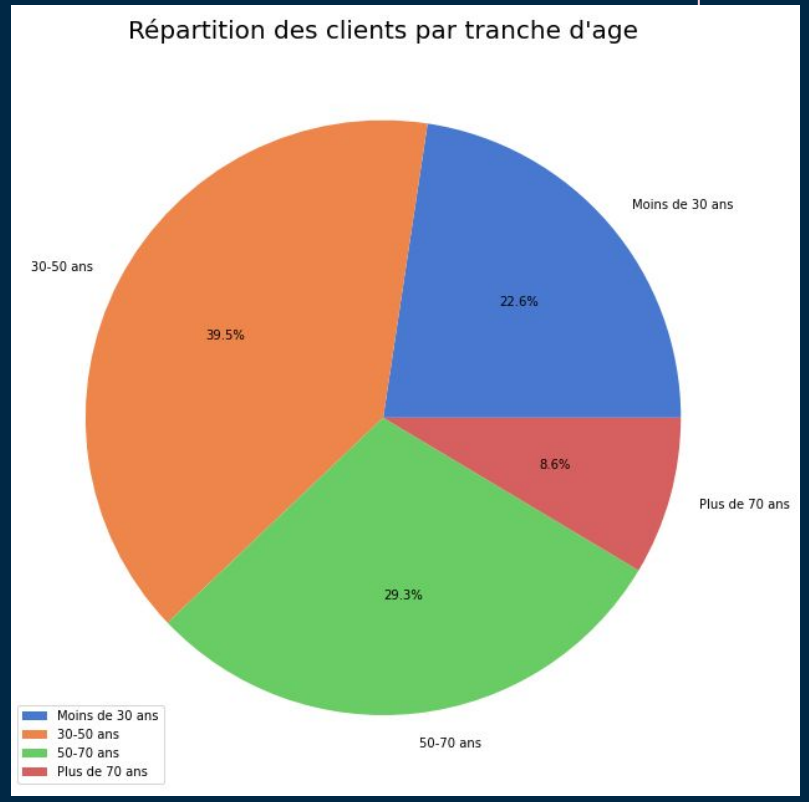


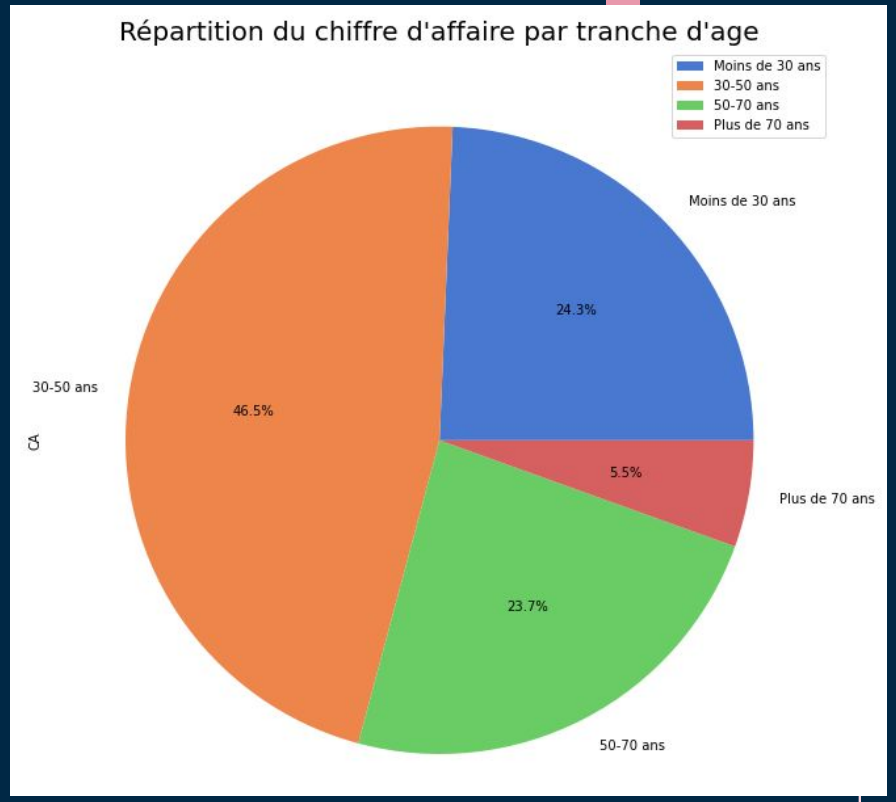
- Moins de 30 ans panier élevé (~70 €)
- 30/50 ans panier d'environ 30 €
- Plus de 50 ans panier d'environ

On remarque 3 groupes !

Analyse de données : Demandes d'Antoine

Etude des profils client



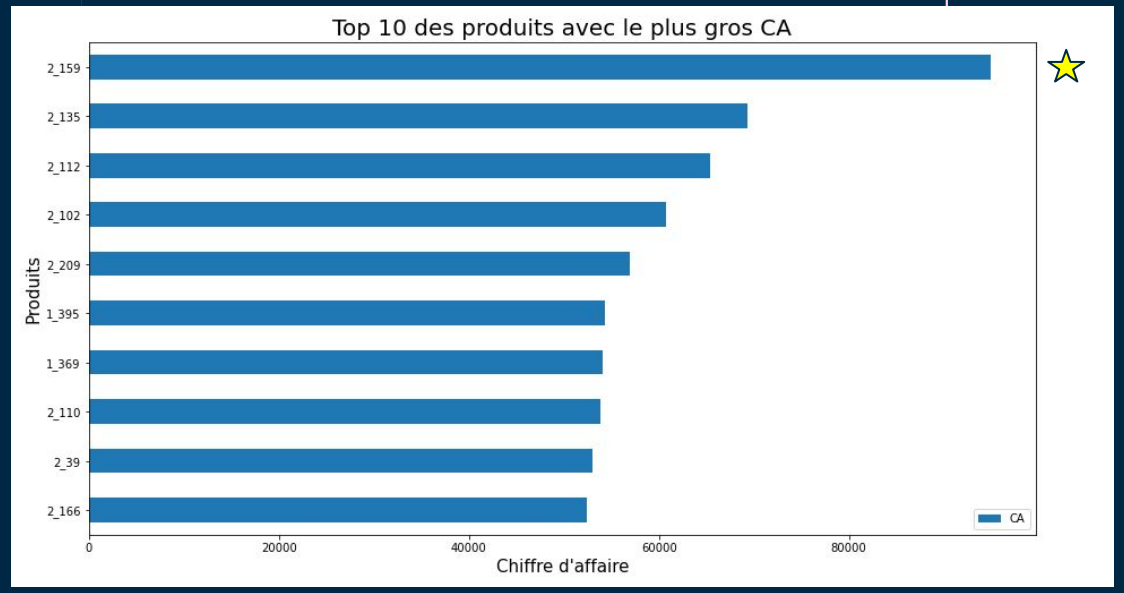


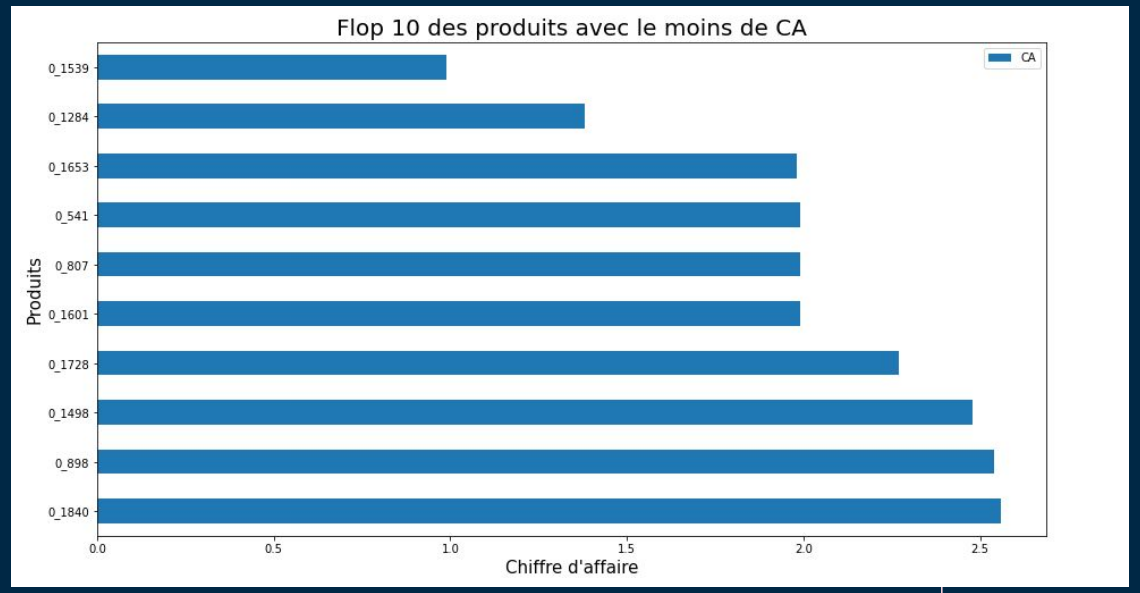
- 30-50 ans : N°1 en nombre + N°1 du CA
- Moins de 30 ans : N°3 en nombre mais N°2 du CA
- 50-70 ans : N°2 en nombre mais N°3 du CA
- Plus de 70 ans : peu en nombre + peu du CA

Analyse des indicateurs : Produits

Analyse de données : Demandes d'Antoine

Etude des produits : CA



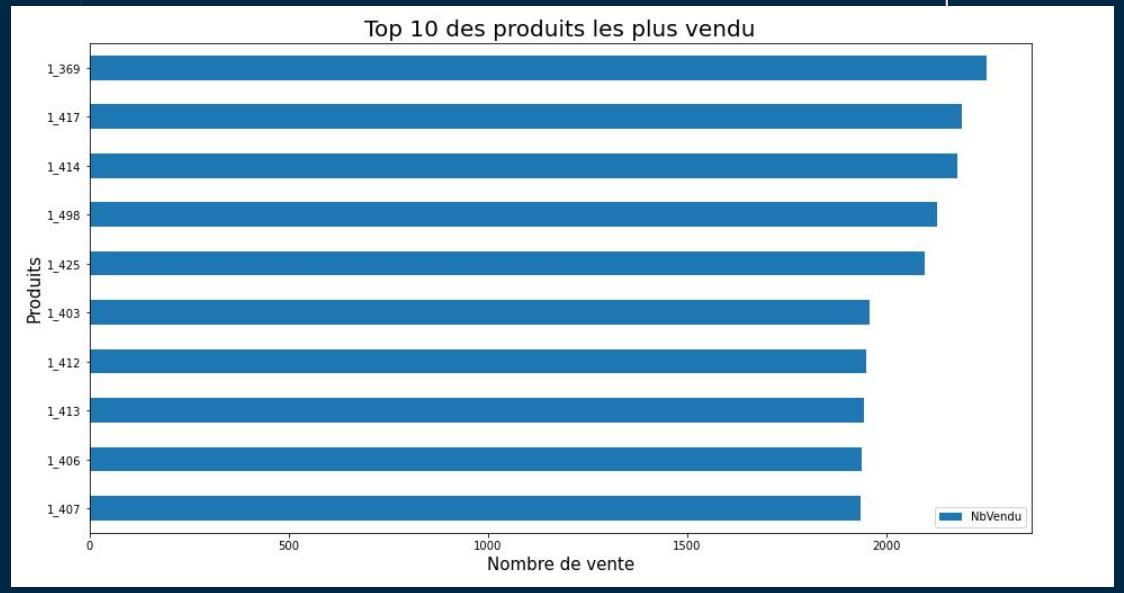


- Produit 2_159 best-seller : +80 000€
- Aucun produit de la catégorie 0

- Seulement des produits de la catégorie 0
- Beaucoup de produits avec un CA < 10€

Analyse de données : Demandes d'Antoine

Etude des produits : Nombre de ventes



	id_prod	NbVendu
3	0_100	3
9	0_1005	5
10	0_1006	8
17	0_1012	3
22	0_1019	8
...	...	...
3258	2_93	2
3259	2_94	8
3260	2_95	4
3263	2_98	1
3264	2_99	7

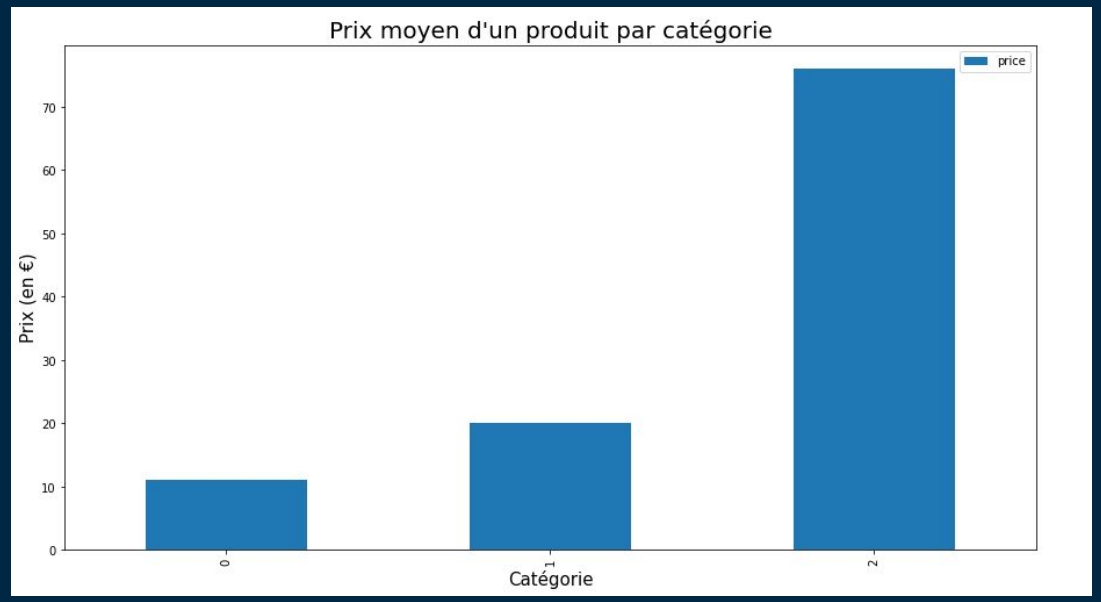
466 rows × 2 columns

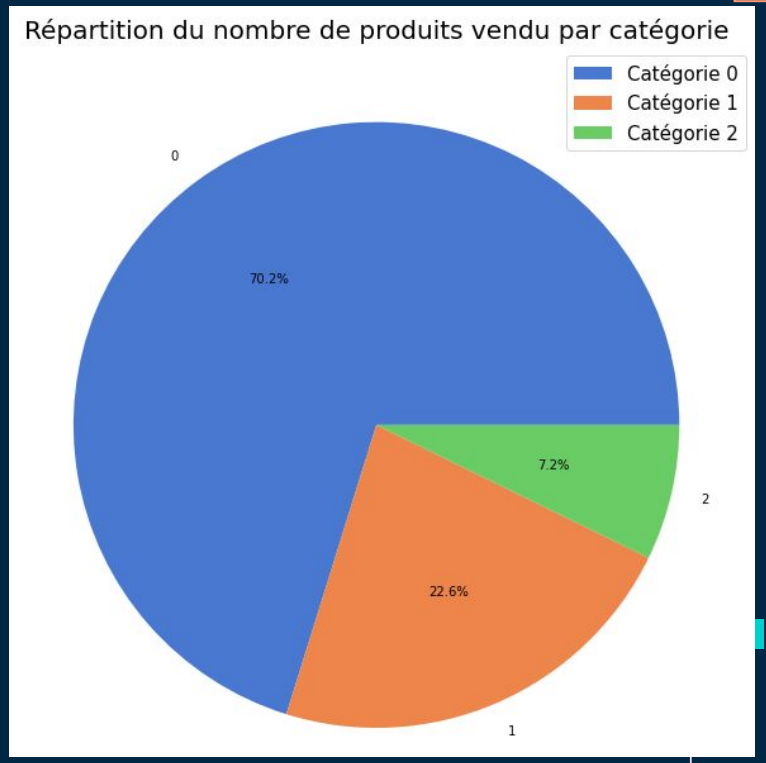
- Seulement des produits de la catégorie 1

- 466 produits avec moins de 10 ventes

Analyse de données : Demandes d'Antoine

Etude des produits : Catégories





- Catégorie 0 très vendu : prix moyen très bas
- Catégorie 2 peu vendu : prix moyen très élevé

Etudes des corrélations

Tests statistiques

Choix du test :

- Choix des variables
- Nombre de variables (1, 2 ou >2)
- Type des variables (quantitative/qualitative)

P-value :

- Entre 0 et 1
- Probabilité que H0 soit vraie
- Seuil de rejet de H0 = 5%

Rédaction des hypothèses :

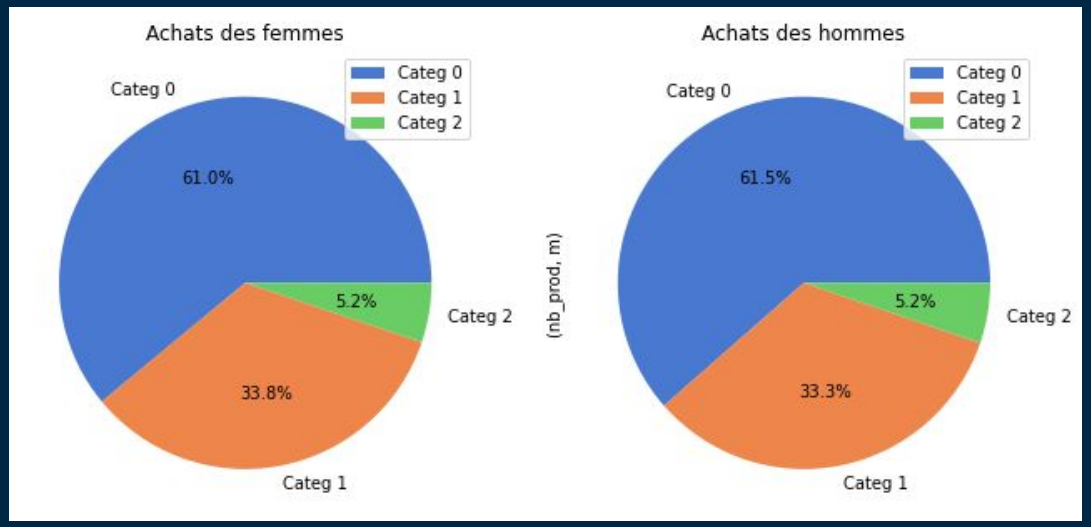
- H0 = Hypothèse nulle
(les variables sont indépendantes)
- H1 = Hypothèse alternative
(les variables ne sont pas indépendantes)

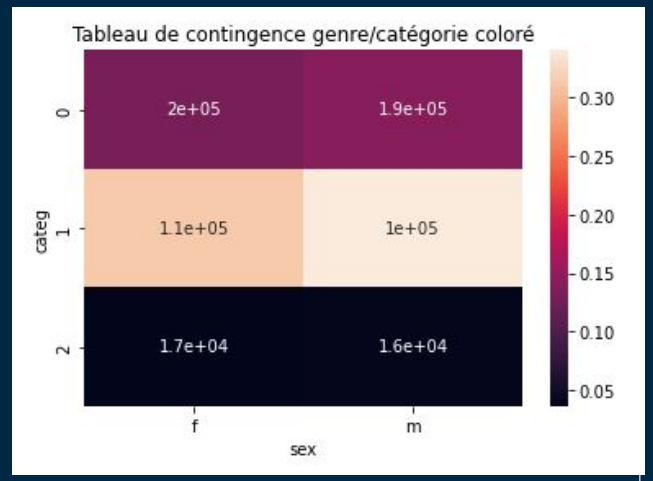
Gestion des outliers :

- Suppression des clients professionnels

Demandes de Julie

Analyse de corrélations entre le genre et la catégorie acheté





- Pour chaque genre répartition quasi identique des achats par catégorie

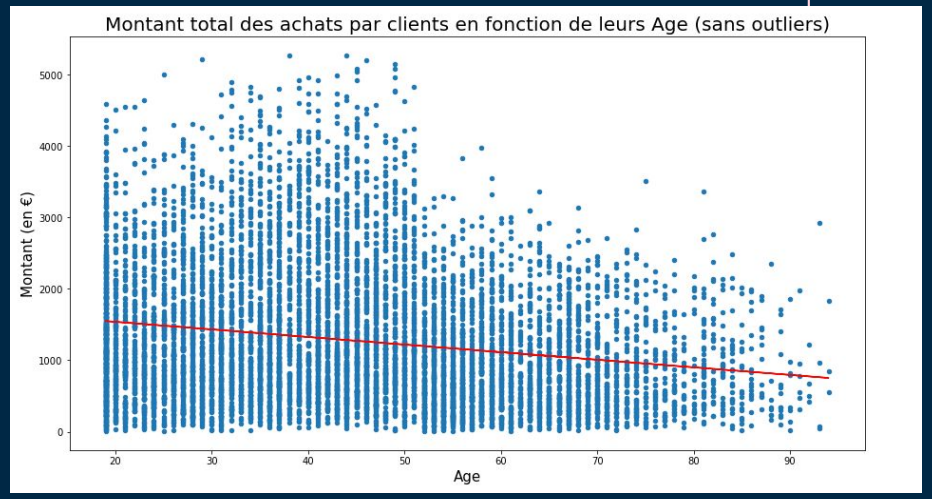
<u>Test Chi-2 :</u>

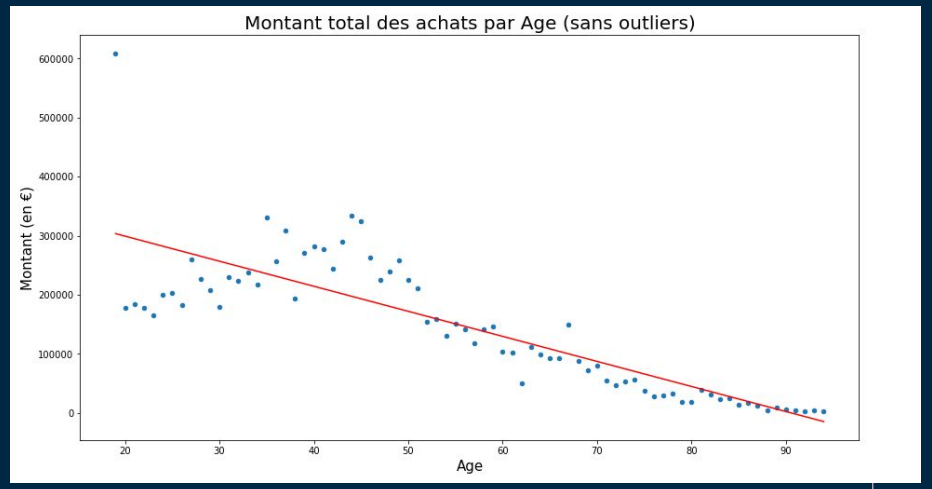
Chi2 stat : 20 + p-value < 5%

= Corrélation existante !

Demandes de Julie

Analyse de corrélations entre l'âge et le montant total des achats





- Tendance à la baisse en fonction de l'âge qui augmente

<u>Résultats du test de Pearson :</u>

Coef corrélation : -0,831 + p-value < 5%

= Corrélation existante !

Demandes de Julie

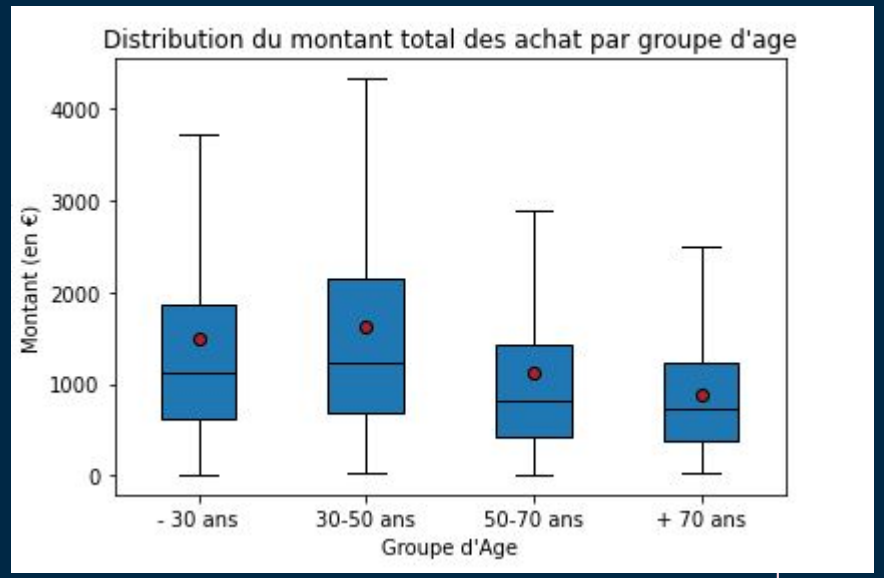
Analyse de corrélations entre le groupe d'âge et le montant total des achats

- Moyennes différentes entre les groupes

<u>Résultats du test de Welch's ANOVA :</u>

Coef corrélation : 0,066 **MAIS** p-value < 5%

= corrélation existante !

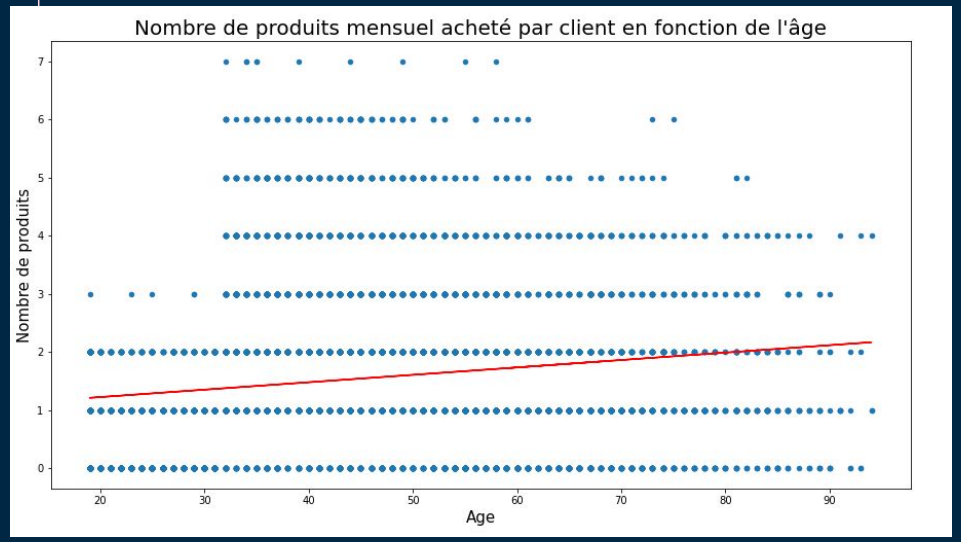


Demandes de Julie

Analyse de corrélations entre la fréquence d'achat et l'âge

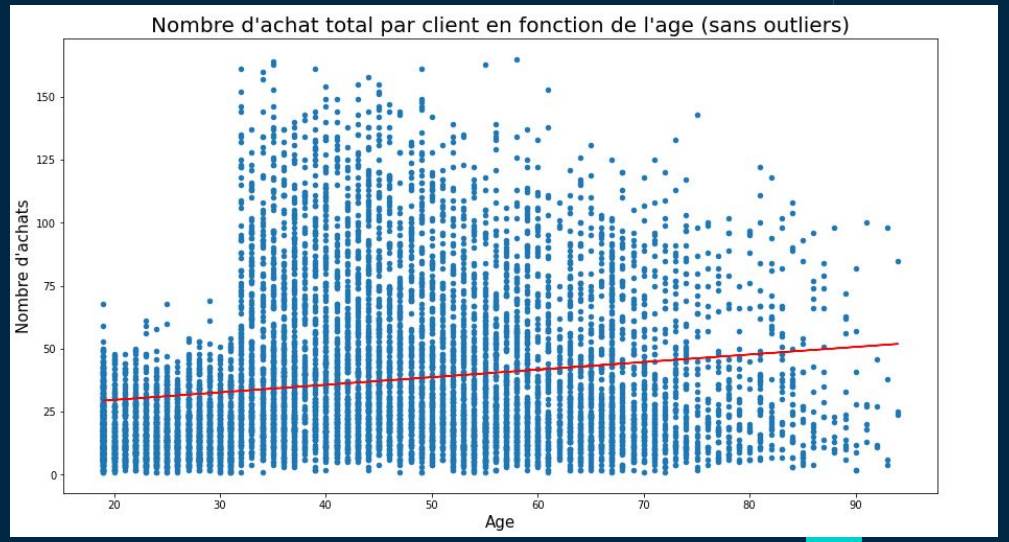
- Tendance positive en fonction de l'âge qui augmente

Méthode 1



Résultats du test de Pearson :
Coef corrélation : 0,160 **MAIS** p-value < 5%

Méthode 2

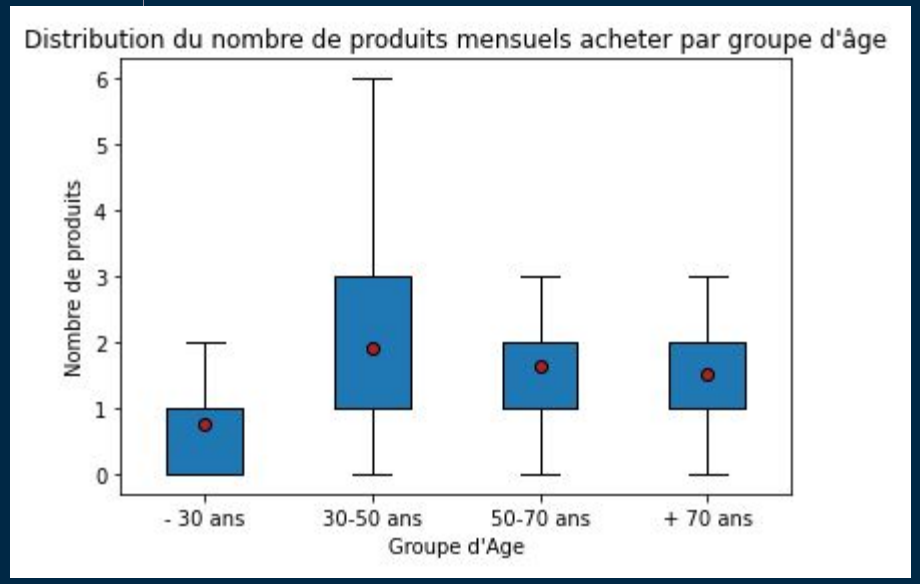


Résultats du test de Pearson :
Coef corrélation : 0,165 **MAIS** p-value < 5%

Demandes de Julie

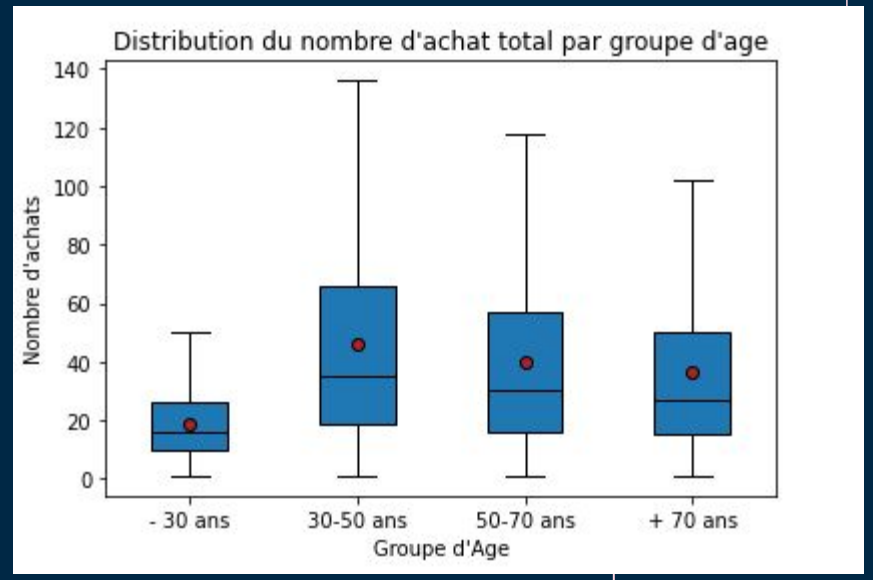
Analyse de corrélations entre la fréquence d'achat et le groupe d'âge

Méthode 1



- Moyennes différentes entre les groupes

Méthode 2



<u>Résultats du test de Welch's ANOVA :</u>

Rapport corrélation : 0,110

MAIS p-value < 5%

= corrélation existante !

<u>Résultats du test de Welch's ANOVA :</u>

Rapport corrélation : 0,114

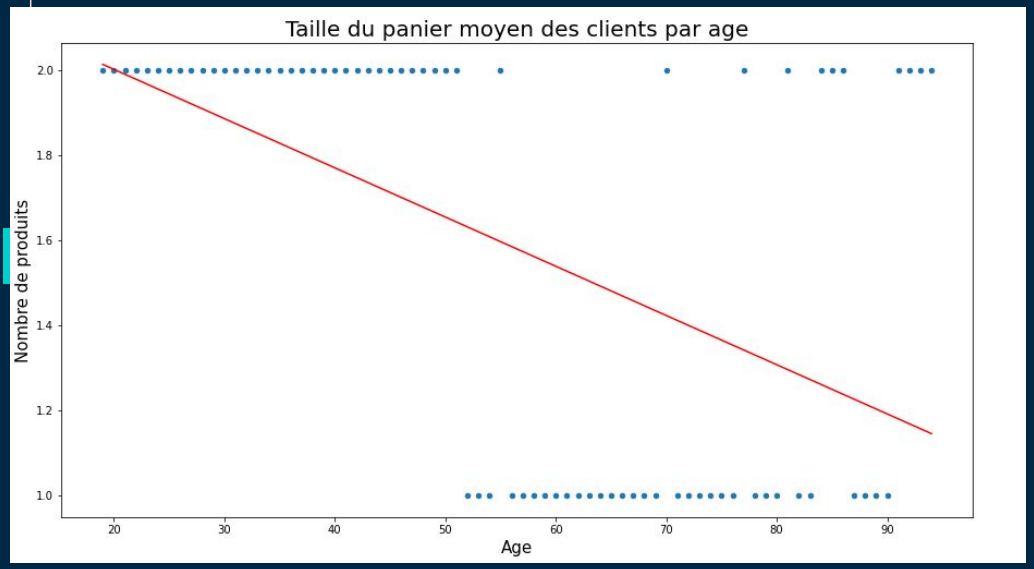
MAIS p-value < 5%

= corrélation existante !

Demandes de Julie

Analyse de corrélations entre la taille du panier et l'âge / groupe d'âge

- Tendance négative en fonction de l'âge qui augmente

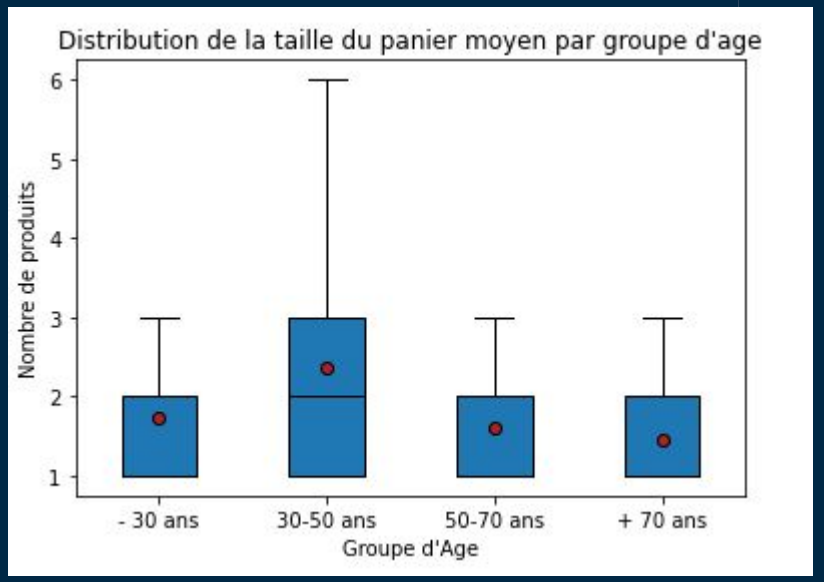


Résultats du test de Pearson :

Coef corrélation : -0,510 + p-value < 5%

= Corrélation existante !

- Moyennes différentes entre les groupes



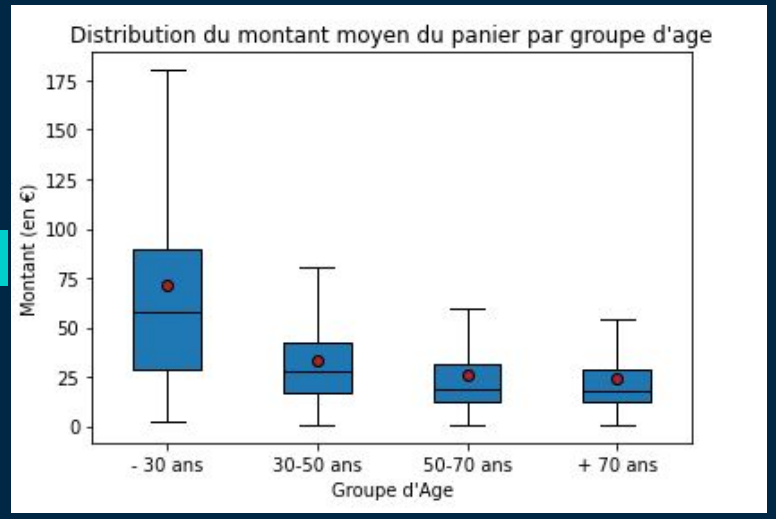
Résultats du test de Welch's ANOVA :

Rapport corrélation : 0,096 **MAIS** p-value < 5%

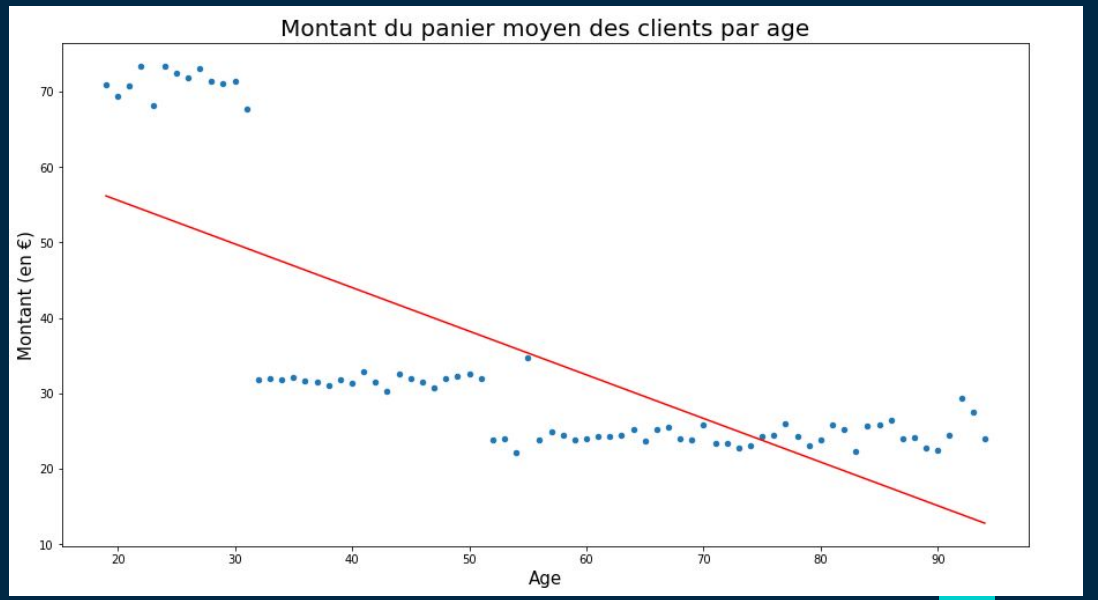
Demandes de Julie

Analyse de corrélations entre le montant moyen du panier et l'âge

- Moyennes différentes entre les groupes



- Tendance négative en fonction de l'âge qui augmente



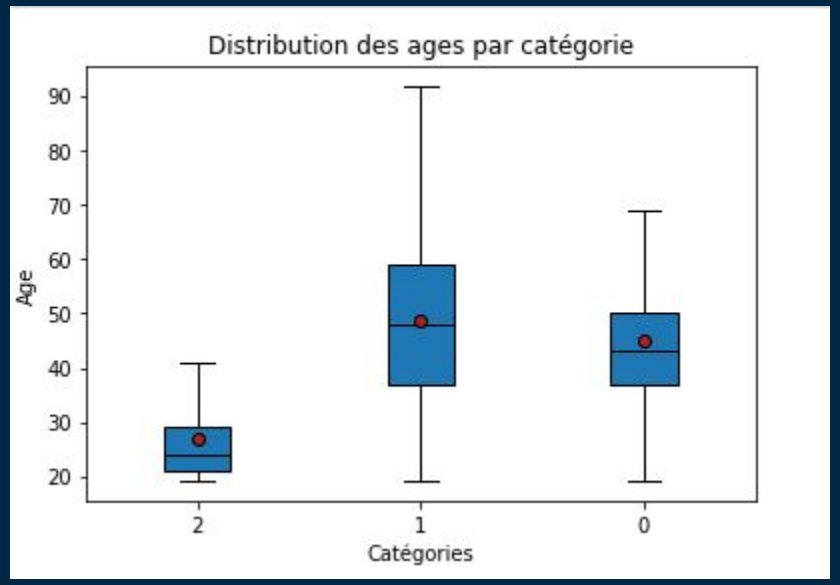
<u>Résultats du test de Welch's ANOVA :</u>
Rapport corrélation : 0,182 **MAIS** p-value < 5%

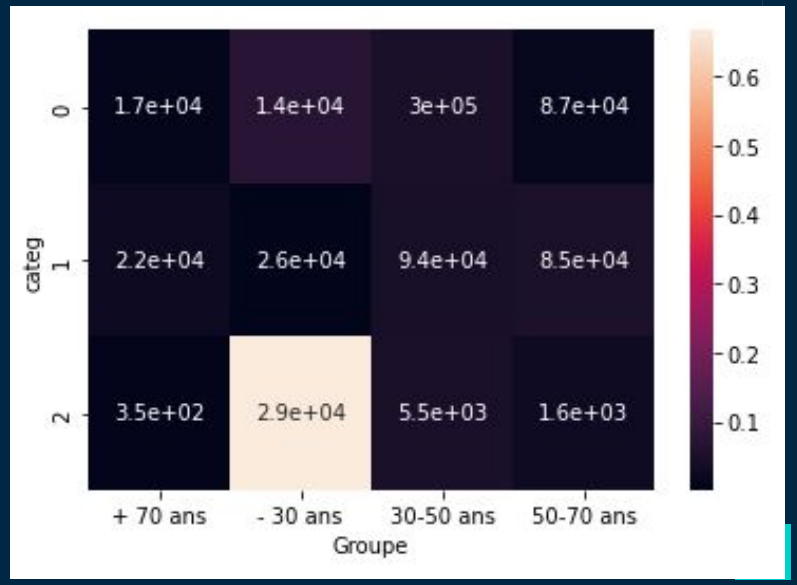
<u>Résultats du test de Pearson :</u>
Coef corrélation : -0,746 + p-value < 5%
= Corrélation existante !

Demandes de Julie

Analyse de corrélations entre la catégorie et l'âge / groupe d'âge

- Moyennes différentes entre les groupes
- Pour chaque groupe répartitions différentes des catégories





Résultats du test de Welch's ANOVA :
Rapport corrélation : 0,113 **MAIS** p-value < 5%

Résultats du test de Chi-2:
Chi2 stat : 257616 + p-value < 5%
= corrélation existante !

Synthèse

Analyse des indicateurs

- Données manquantes Octobre 2021
- Mauvais début de l'année 2023
- Catégorie 2 fait le moins de CA
- 4 clients professionnels

- Top 10 produits CA et ventes : catégorie 1 & 2
- Flop 10 produits CA et ventes : Catégorie 0

- 3 groupes distincts de clients
- Les plus jeunes achètent les produits les plus chers

Analyse des corrélations

- Fortes corrélations :
  Âge et montant total d'achat
  Âge et taille du panier
  Âge et montant moyen du panier

- Faibles corrélations :
  Genre et catégorie
  Âge et fréquence d'achat
  Âge et catégorie