Analyse des ventes du site web

Lapage

Introduction

Faire le point après deux ans d’exercice du site, et pouvoir analyser ses points forts, ses points faibles, les comportements clients, etc.

Il y a 3 fichiers csv qui sont mis à disposition :

- customers.csv : un fichier qui contient l'identifiant du client, son sexe et l'année de naissance.
- products.csv : un fichier qui contient la référence produit, son prix et la catégorie.
- transactions.csv : un fichier qui concerne les transactions du site. Il fait le lien entre le fichier customers et products. Il contient l'identifiant du client et la référence du produit ainsi que la date de la transaction et son identifiant.

Point abordé pendant la présentation :

- Observation des différents dataframes.
- Rapprochement des dataframes.
- Analyse des ventes.
  - Chiffre d'affaires.
    - Répartition par catégorie.
    - Répartition par tranche d'âge.
    - Répartition par catégorie dans le temps.
    - Moyenne mobile sur 3 et 5 mois.
  - Zoom sur les références.
    - Meilleures et moins bonnes ventes globales.
    - Meilleures et moins bonnes ventes par catégorie.
  - Courbe de Lorenz.
- Analyse de la clientèle.
  - Sexe des clients par rapport aux catégories de livres.
  - Age des clients et le montant total des achats.
  - Age et la fréquence d'achat.
  - Age et la taille du panier moyen.
  - Age et les catégories de livres achetés.

Observation du dataframe clients

	client_id	sex	birth
8494	ct_1	m	2001
2735	ct_0	f	2001
7358	c_999	m	1964
2145	c_998	m	2001
94	c_997	f	1994
...	...	...	...
3426	c_1001	m	1982
8472	c_1000	f	1966
2137	c_100	m	1992
6894	c_10	m	1956
4299	c_1	m	1955

	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

La clé primaire est ici la colonne client_id, elle est composée du préfixe “c_” suivi de chiffre.

La colonne sex contient les 2 valeurs f et m.

L’année de naissance va de 1929 à 2004, cette colonne va changer, on va afficher plutôt l’âge des clients.

2 lignes vont être supprimées, “ct_0” et “ct_1” car ce sont des valeurs de test.

Observation du dataframe produits

	id_prod	price	categ
731	T_0	-1.00	0
3188	2_99	84.99	2
3088	2_98	149.74	2
2698	2_97	160.99	2
2576	2_96	47.91	2
...	...	...	...
922	0_1000	6.84	0
663	0_100	20.60	0
2691	0_10	17.95	0
803	0_1	10.99	0
1001	0_0	3.75	0

La clé primaire est la colonne “id_prod”, elle est composée du préfixe qui correspond à la catégorie (0_ pour 0, 1_ pour 1, etc).

La colonne catégorie a 3 valeurs : 0, 1 et 2.

Il y a une ligne qui concerne des valeurs de tests a supprimé avec un prix négatif et un id produit à “T_0”.

Observation du dataframe 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

C’est un dataframe qui fait le lien entre clients et produits par l'intermédiaire de id_prod et client_id.

Chaque transaction est désignée par l’identifiant de la session.

Il y a aussi la présence de valeurs de tests a supprimé, dans la colonne date, il y a en préfixe “test_” suivi de la date et “s_0” dans la colonne session_id.

Rapprochement des dataframe

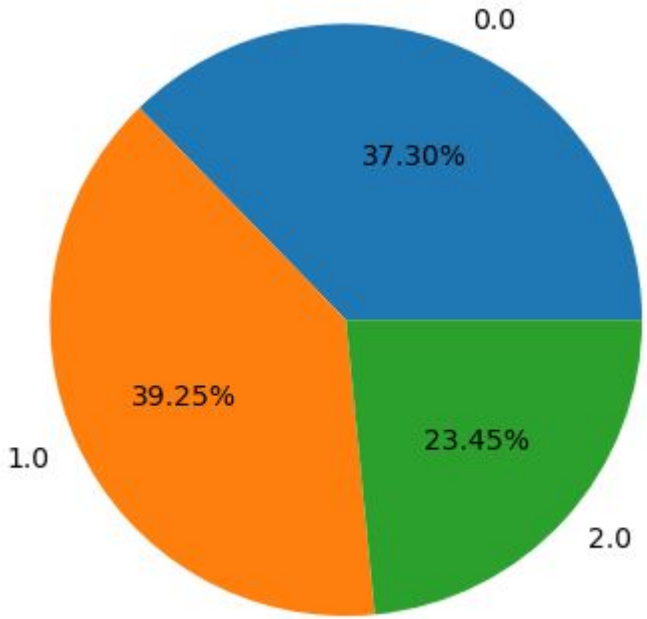
	client_id	sex	age	id_prod	date	session_id	price	categ
2633	c_4746	m	83	0_2245	2022-09-23 07:22:38.636773	s_272266	10.64	0.0
10106	c_6713	f	60	0_2245	2022-07-23 09:24:14.133889	s_242482	10.64	0.0
11727	c_5108	m	45	0_2245	2022-12-03 03:26:35.696673	s_306338	10.64	0.0
15675	c_1391	m	32	0_2245	2021-08-16 11:33:25.481411	s_76493	10.64	0.0
16377	c_7954	m	50	0_2245	2022-07-16 05:53:01.627491	s_239078	10.64	0.0
...	...	...	...	...	...	...	...	...
669730	c_131	m	42	0_2245	2021-08-25 09:06:03.504061	s_80395	10.64	0.0
670682	c_4167	f	44	0_2245	2022-03-06 19:59:19.462288	s_175311	10.64	0.0
671286	c_4453	m	42	0_2245	2022-05-16 11:35:20.319501	s_209381	10.64	0.0
675679	c_1098	m	37	0_2245	2022-02-11 09:05:43.952857	s_163405	10.64	0.0
677996	c_4854	m	55	0_2245	2021-12-14 22:34:54.589921	s_134446	10.64	0.0

Rassemblement des différentes dataframes en un seul appelé ventes pour pourvoir faire les analyses par la suite.

Il y a des valeurs manquantes du au rapprochement pour le produit “0_2245” sur les colonnes prix et catégorie.

Comme le préfixe est “0_”, la catégorie va être 0 et pour le prix, c’est la moyenne des prix pour la catégorie 0 qui va être utilisée pour remplir les valeurs manquantes.

Répartition du chiffre d'affaires par catégorie

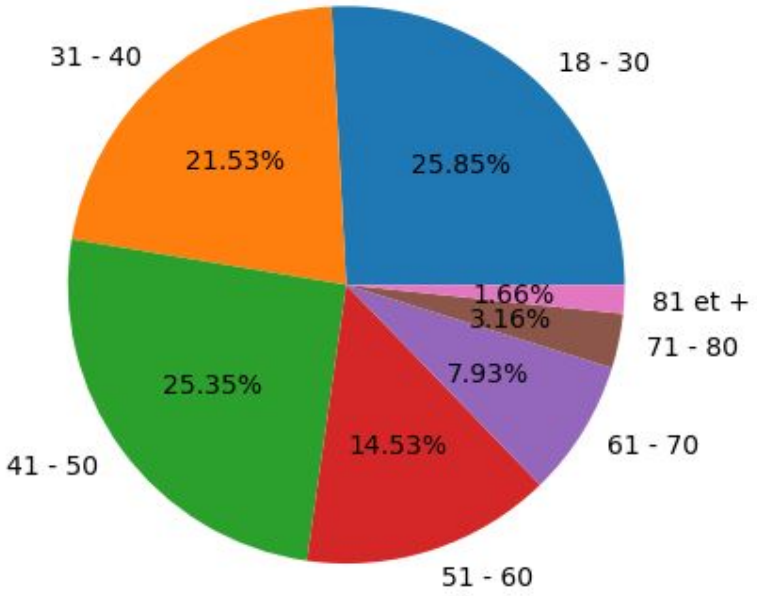


Les catégories 0 et 1 sont celles qui génèrent le plus de chiffre d’affaires.

La catégorie 1 représente 39,25 % et la catégorie 0 avec 37,30 %.

La catégorie 2 génère le moins de chiffre d’affaires avec 23,45 %.

Répartition du chiffre d'affaires par tranche d’âge

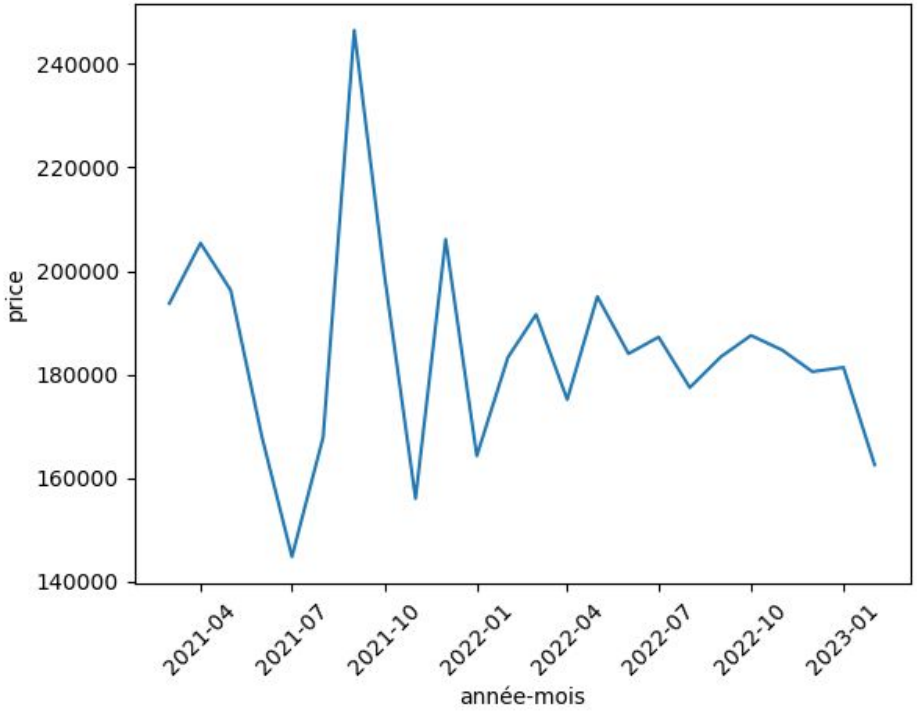


Il y a 3 parts qui sont le plus représentée :

- Les 18-30 avec 25,85 %.
- Les 41-50 ans avec 25,35 %.
- Les 31-40 ans avec 21,53 %.

La part qui est le moins représentée est les 81 et + avec 1,66 %.

Répartition du chiffre d'affaires pour la catégorie 0

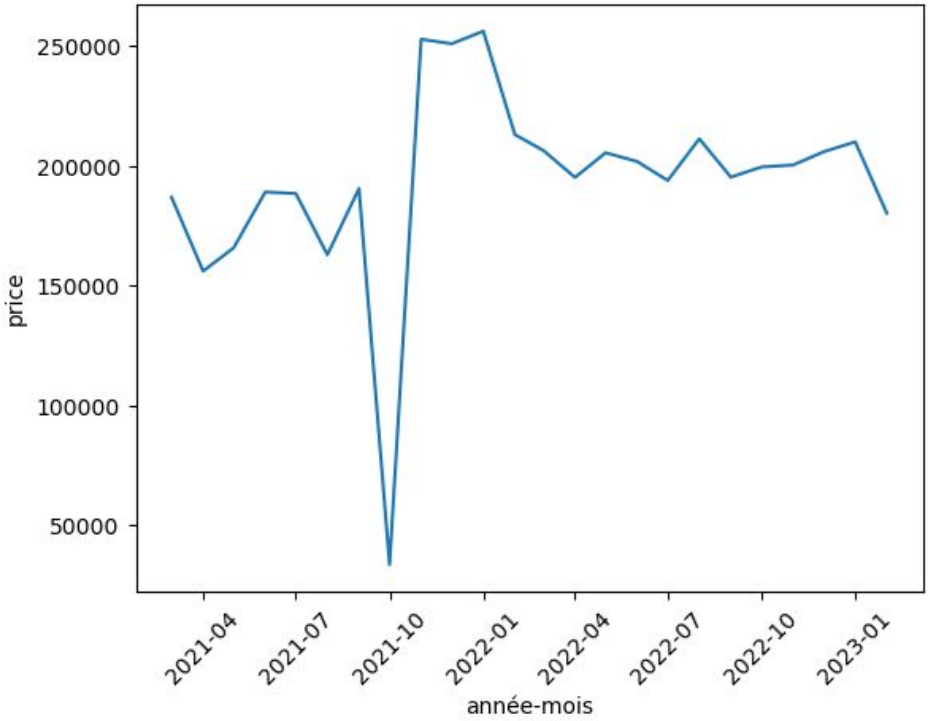


Globalement, le chiffre d'affaires est entre 160 000 € et 210 000 € pendant cette période.

Plus fortes ventes en septembre 2021 avec environ 240 000 € du chiffre d'affaires.

La période qui a le moins rapporté est en juillet 2021 avec environ 145 000 €.

Répartition du chiffre d'affaires pour la catégorie 1

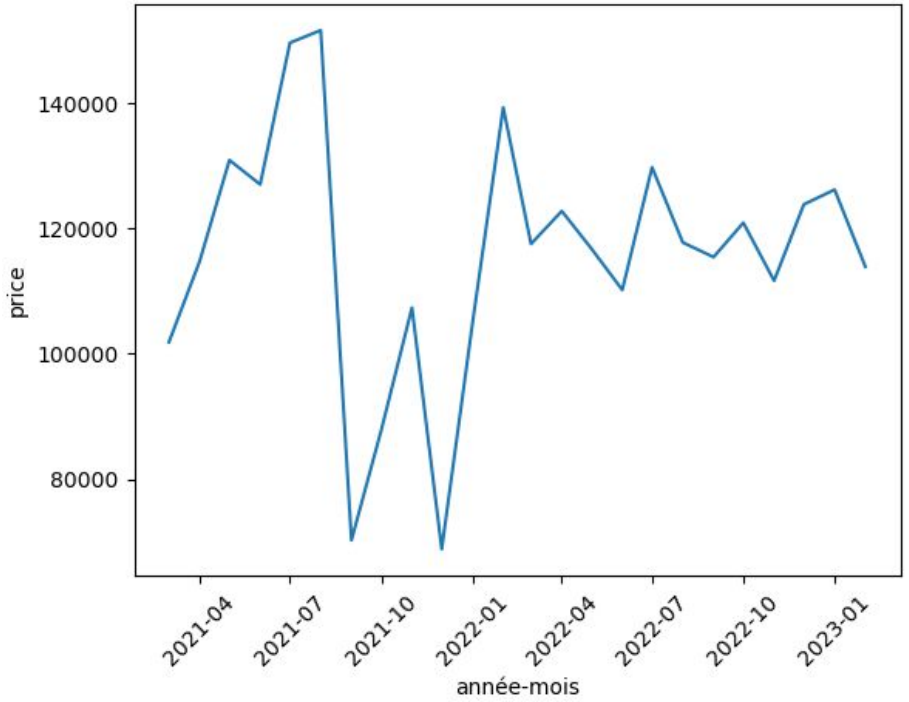


Les périodes de novembre et décembre 2021 ont généré le plus de chiffre d'affaires avec 250 000 €.

Globalement sur la période, le chiffre d'affaires est entre 150 000 € et 210 000 €.

Il y a une anomalie pour le mois d'octobre, il n'y a pas d'information pour cette période.

Répartition du chiffre d'affaires pour la catégorie 2

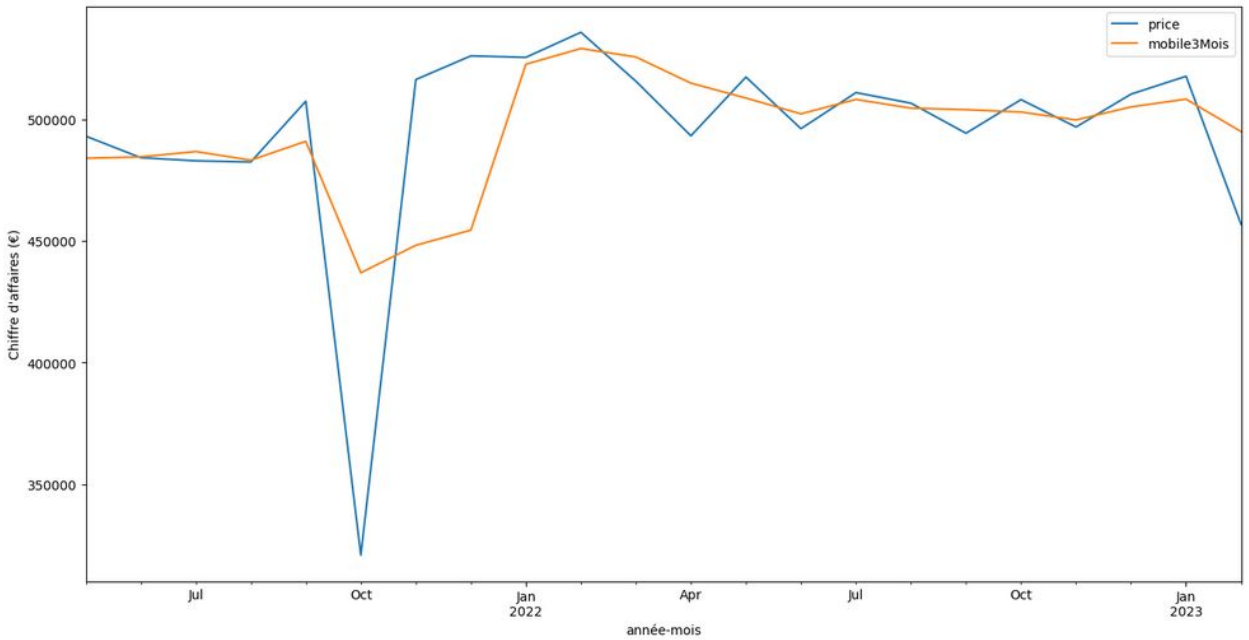


Globalement, le chiffre d'affaires est entre 100 000 € et 140 000 €.

2 pics positifs et négatifs dans la courbe :

- En juillet/août 2021 avec 150 000 € pour le positif.
- Une baisse à 70 000 € pour septembre et décembre 2021.

Moyenne mobile sur 3 mois

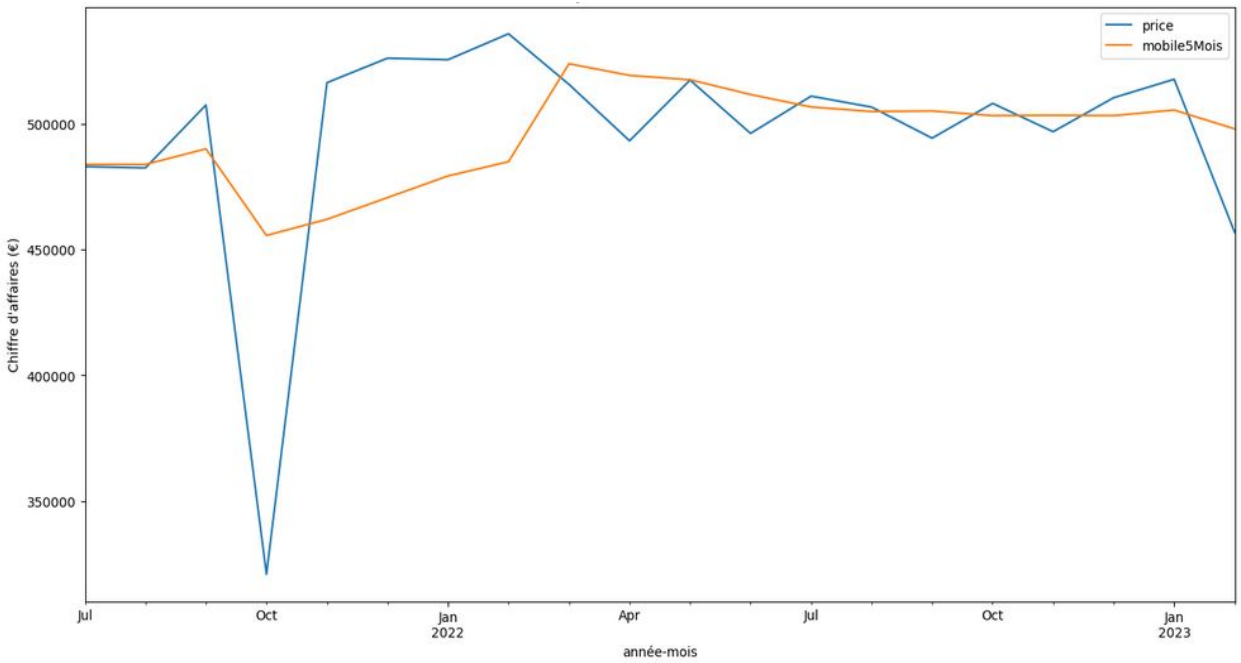


La moyenne mobile est généralement autour des 500 000 €.

Période la plus forte entre janvier et mars 2022.

Le mois d'octobre est la période la plus basse dû à l'absence de données pour la catégorie 1.

Moyenne mobile sur 5 mois



La moyenne mobile est généralement autour des 500 000 € comme précédemment.

Même période forte entre janvier et mars 2022.

Même problème pour le mois d'octobre 2021.

Plus la période choisie est importante et plus la courbe tant à ce lisser.

Meilleures et moins bonnes ventes globales

Meilleures ventes

	id_prod	categ	Nombre de vente
2592	1_369	1.0	2252
2645	1_417	1.0	2189
2642	1_414	1.0	2180
2734	1_498	1.0	2128
2654	1_425	1.0	2096
2630	1_403	1.0	1960
2640	1_412	1.0	1951
2641	1_413	1.0	1945
2633	1_406	1.0	1939
2634	1_407	1.0	1935

Moins bonnes ventes

	id_prod	categ	Nombre de vente
1327	0_2201	0.0	1
166	0_1151	0.0	1
802	0_1728	0.0	1
3248	2_81	2.0	1
595	0_1539	0.0	1
313	0_1284	0.0	1
1793	0_549	0.0	1
549	0_1498	0.0	1
1785	0_541	0.0	1
2167	0_886	0.0	1

Meilleures et moins bonnes ventes pour la catégorie 0

Meilleures ventes

	id_prod	categ	Nombre de vente
466	0_1422	0.0	1292
476	0_1431	0.0	1282
469	0_1425	0.0	1266
477	0_1432	0.0	1254
454	0_1411	0.0	1246
472	0_1428	0.0	1245
0	0_0	0.0	1242
468	0_1424	0.0	1238
487	0_1441	0.0	1235
479	0_1434	0.0	1235

Moins bonnes ventes

	id_prod	categ	Nombre de vente
2080	0_807	0.0	1
1327	0_2201	0.0	1
166	0_1151	0.0	1
802	0_1728	0.0	1
595	0_1539	0.0	1
313	0_1284	0.0	1
1793	0_549	0.0	1
549	0_1498	0.0	1
1785	0_541	0.0	1
2167	0_886	0.0	1

Meilleures et moins bonnes ventes pour la catégorie 1

Meilleures ventes

	id_prod	categ	Nombre de vente
2592	1_369	1.0	2252
2645	1_417	1.0	2189
2642	1_414	1.0	2180
2734	1_498	1.0	2128
2654	1_425	1.0	2096
2630	1_403	1.0	1960
2640	1_412	1.0	1951
2641	1_413	1.0	1945
2633	1_406	1.0	1939
2634	1_407	1.0	1935

Moins bonnes ventes

	id_prod	categ	Nombre de vente
2313	1_117	1.0	4
2453	1_243	1.0	4
2432	1_224	1.0	4
2635	1_408	1.0	3
2631	1_404	1.0	3
2345	1_146	1.0	3
2636	1_409	1.0	3
2753	1_514	1.0	2
2649	1_420	1.0	2
2629	1_402	1.0	2

Meilleures et moins bonnes ventes pour la catégorie 2

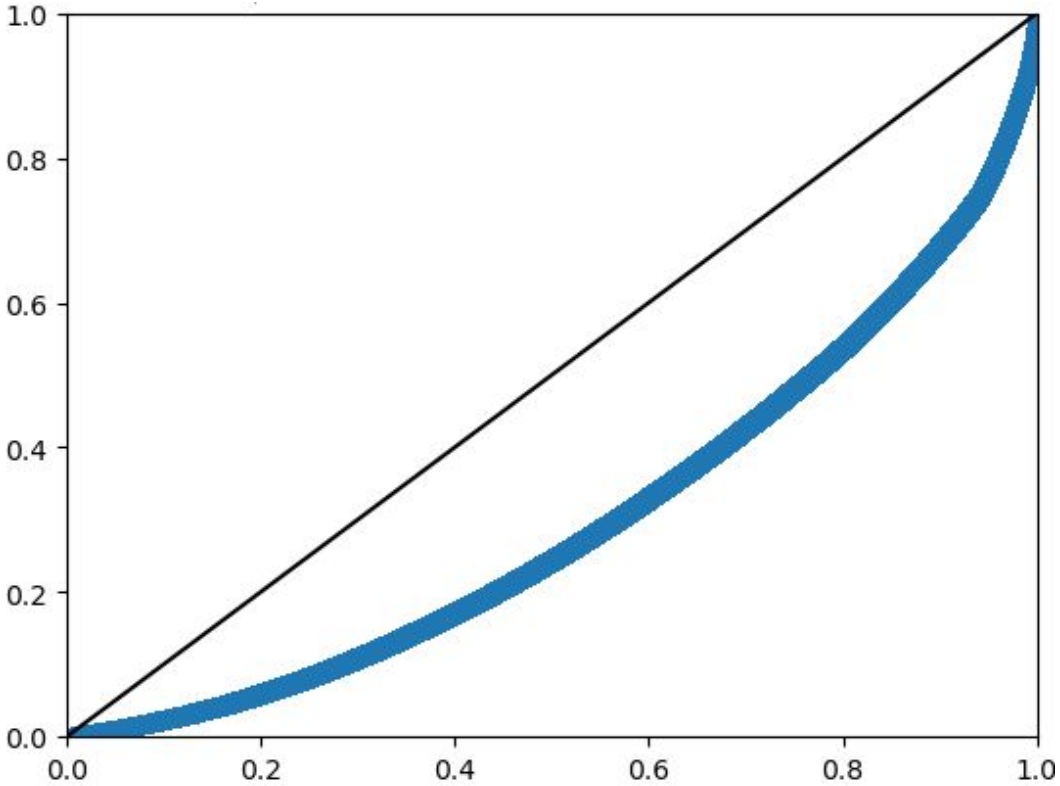
Meilleures ventes

	id_prod	categ	Nombre de vente
3035	2_102	2.0	1027
3071	2_135	2.0	1005
3046	2_112	2.0	968
3202	2_39	2.0	915
3200	2_37	2.0	882
3044	2_110	2.0	865
3152	2_208	2.0	831
3153	2_209	2.0	814
3151	2_207	2.0	786
3042	2_109	2.0	744

Moins bonnes ventes

	id_prod	categ	Nombre de vente
3232	2_66	2.0	3
3244	2_78	2.0	3
3190	2_28	2.0	3
3066	2_130	2.0	3
3067	2_131	2.0	3
3243	2_77	2.0	2
3259	2_93	2.0	2
3264	2_98	2.0	1
3176	2_23	2.0	1
3248	2_81	2.0	1

Courbe de Lorenz

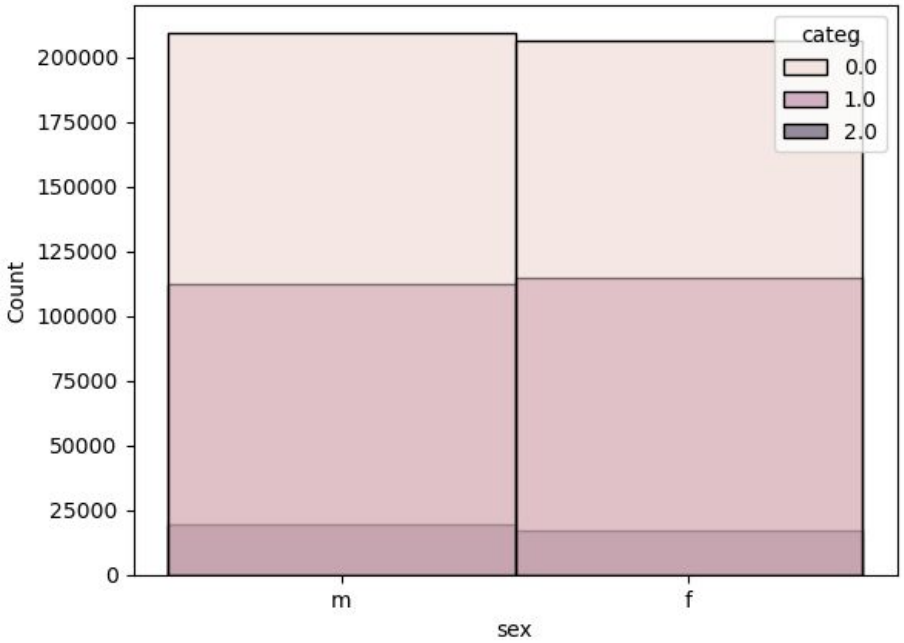


La courbe va représenter la répartition du chiffre d’affaires rapporté par les clients.

L'indice de Gini est de 0.39.

Elle montre qu'il y a une inégalité du chiffre d'affaires entre les clients.

Sexe des clients par rapport aux catégories de livres

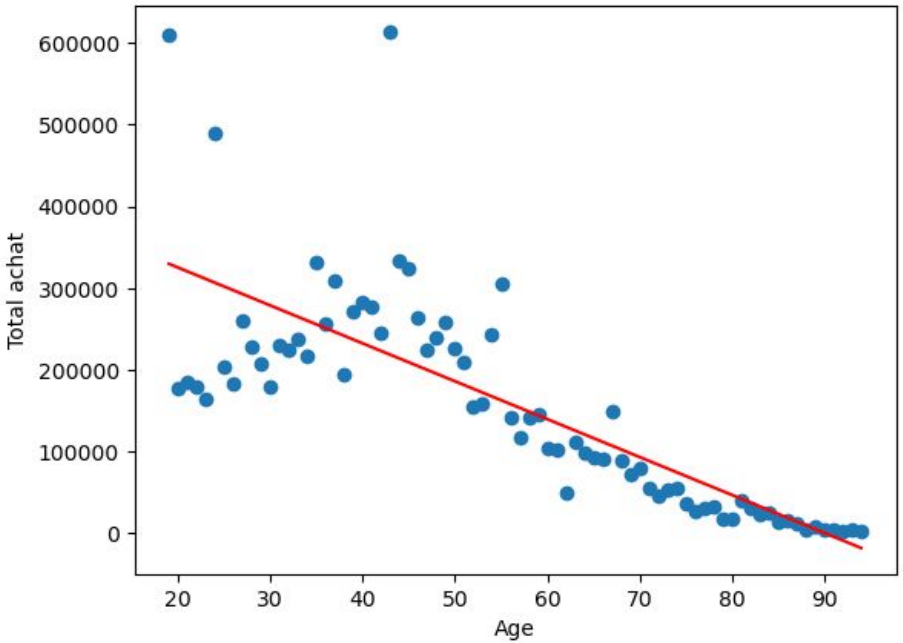


On utilise le test de Chi-2 pour voir un lien entre ces 2 éléments.

Répartition du nombre de ventes est égale entre les genres, on peut émettre comme hypothèse que le sexe des clients n'a pas d'impact sur les catégories de livres achetés.

La p-value est de 1,19, elle est donc faible, ce qui confirme l’hypothèse qu’il n’y a pas de lien entre le sexe des clients et la catégorie de livres achetés.

Age des clients et le montant total des achats

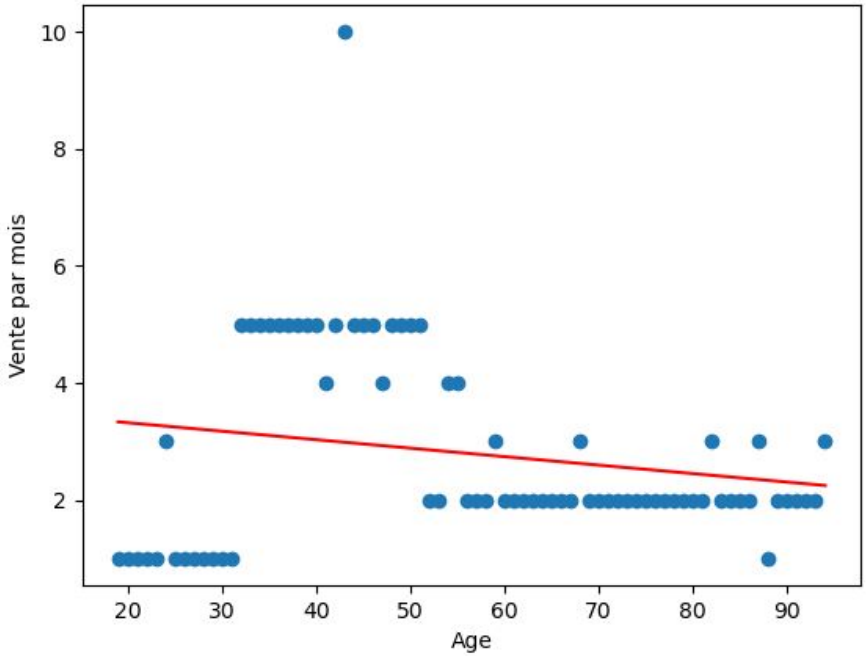


Utilisation d’une régression linéaire, pour voir le lien entre l'âge des clients et le montant total des achats.

Montant de la courbe commence à environ à 320 000 €.

Plus les clients sont âgés et plus le montant total des achats diminue.

Age des clients et la fréquence d’achat



Utilisation d’une régression linéaire.

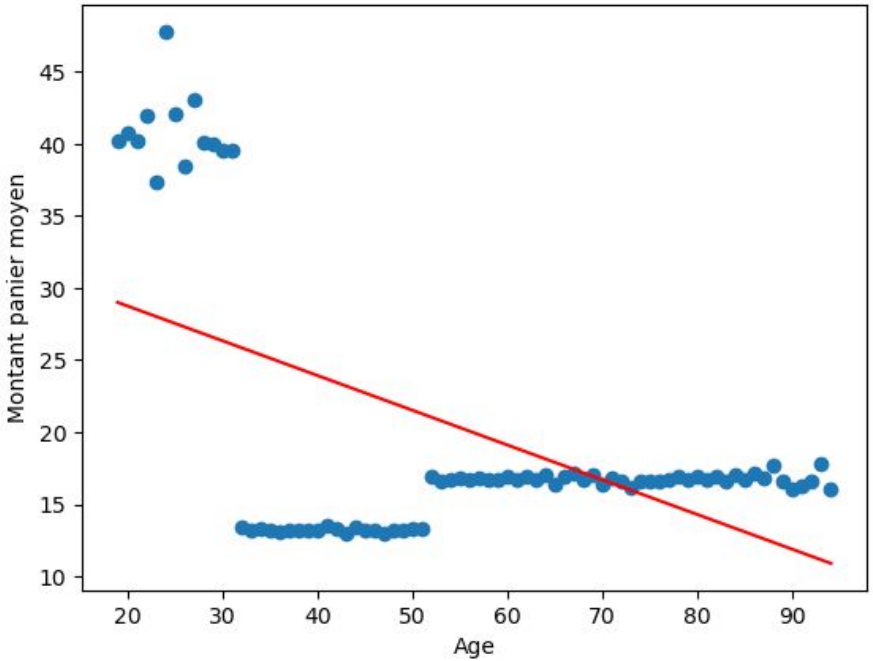
La droite montre que la fréquence est de 3 livres par mois pour n'importe quel âge.

Entre 18 et 31 ans, la fréquence est la moins élevée avec 1 livre par mois.

La fréquence la plus importante est entre 32 - 51 ans avec 5 livres par mois.

La majorité des clients achète 2 livres par mois en moyenne.

Age des clients et la taille du panier moyen



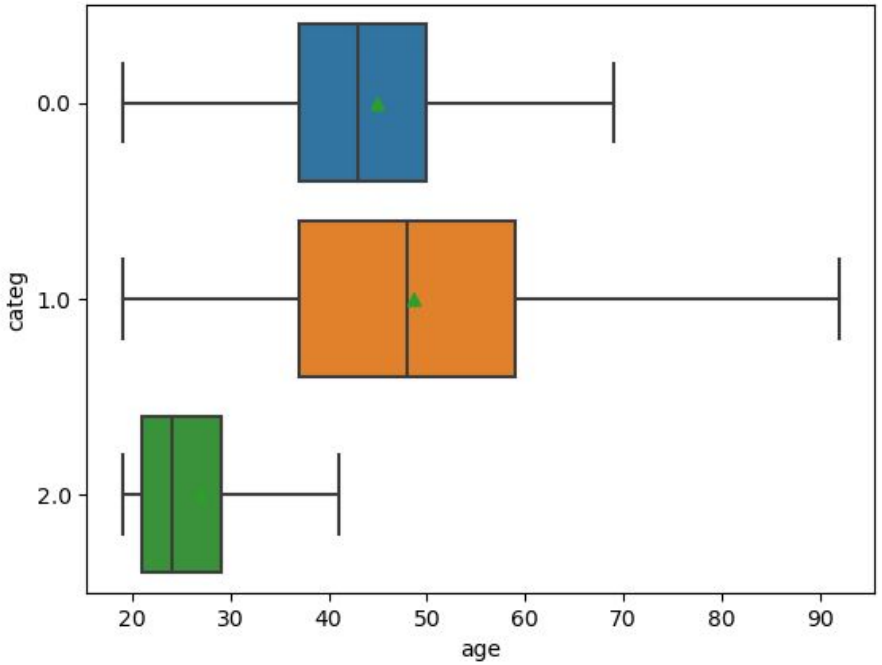
Utilisation aussi d’une régression linéaire.

La régression linéaire commence à 30 € du panier moyen pour finir à environ 10 €.

Le panier moyen le plus important est pour la tranche d'âge de 18 - 30 ans qui va avec un panier allant de 37 € à 47 €.

Les 30 - 50 ans représente ceux où le panier moyen est le moins élevé avec une valeur de 15 €.

Age des clients et les catégories de livres achetés



Analyse par ANOVA.

Les triangles représentent l'âge moyen.

La catégorie 2 est achetée par les 18 - 40 ans avec une forte concentration entre les 20- 30 ans.

La catégorie 1 est achetée par tous les âges avec les 35 - 60 ans qui représentent 50 % des achats.

La catégorie 0 est achetée par une bonne partie des clients, l'âge va de 18 à 70 ans, avec une concentration de 35 à 50 ans.

La corrélation entre l’âge et la catégorie est de 0,1, ce qui montre une relation faible, cette relation faible est due à la catégorie 2.

Conclusion

- Les catégories qui rapportent le plus sont la 0 et 1.
- La catégorie 2 est exclusivement achetée par les 18 - 40 ans.
- Plus les clients sont âgés et plus le chiffre d’affaires diminue donc forte clientèle jeune.
- Panier moyen fort pour les 18 - 30 ans mais une fréquence faible.
- Effet inverse pour les 30 - 50 ans.