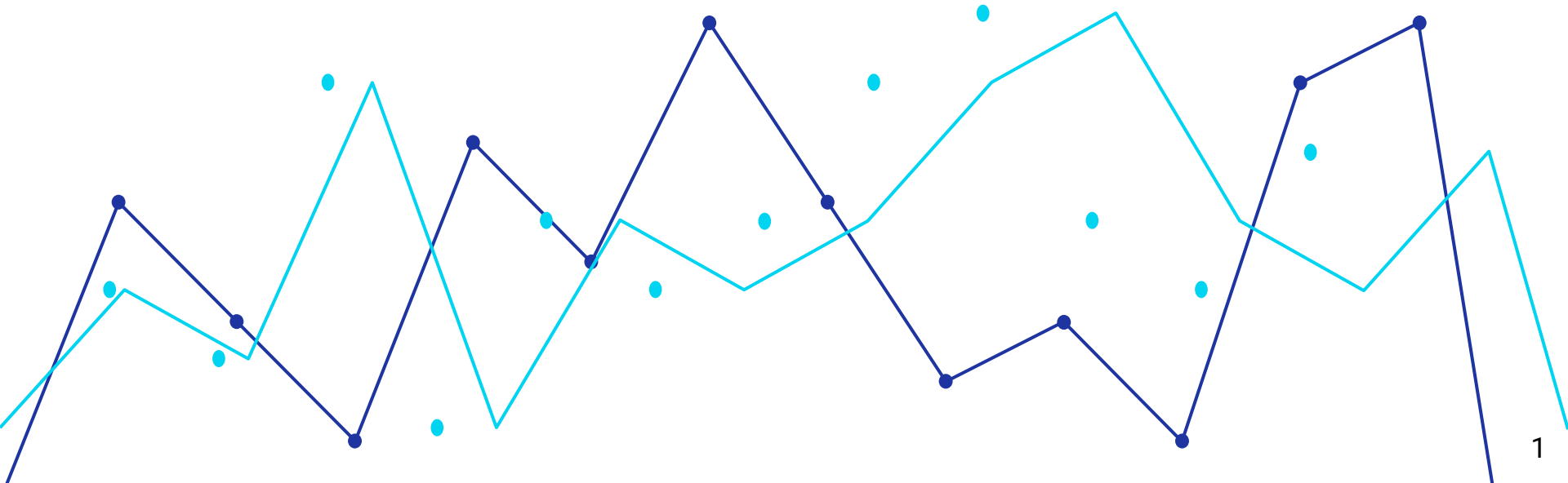
Data Analyst



Problèmatique

Comment établir un état des lieux via les chiffres d'affaires d'une entreprise ?

Index

Stack technique

Nous avons mis en place tout un environnement afin de pouvoir manipuler les donnée de manière sécurisée. Le projet utilise le langage python via l’IDE Jupyter et un gestionnaire de container qui est docker

Traitement des données

Data's

Data

On interagit avec 3 dataframes différents:

1. Les produits
2. Les transactions
3. Les clients

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

Transactions

	id_prod	date	session_id	client_id
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450
1	2_226	2022-02-03 01:55:53.276402	s_159142	c_277
2	1_374	2021-09-23 15:13:46.938559	s_94290	c_4270
3	0_2186	2021-10-17 03:27:18.783634	s_105936	c_4597
4	0_1351	2021-07-17 20:34:25.800563	s_63642	c_1242

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

Traitement des données

Pour chacune de ces dataframes évoqués nous allons appliquer un traitement de données. Ceci implique plusieurs faits :

1. Correction des valeurs aberrantes
2. Enlever les valeurs négatives
3. Identifier les valeurs nulles

Produit

Data

Pour chaque items, elle correspondent à une catégorie unique et un prix fixe.

Problématique ?
Oui, car elle à des valeurs négatives mais aussi de test.

Test Values

```
detectTestValue = products[products['id_prod'].str.contains('T', na=True) == True]
detectTestValue
```

	id_prod	price	categ
731	T_0	-1.0	0

Negative values

	id_prod	price	categ
731	T_0	-1.0	0

Transaction

Data

Pour chaque colonne, elle correspondent à un id de produit, une date, un client et une session

Problématique ?
Oui, nous obtenons 200 lignes de test.

Afin de mieux traités les données. On applique une fonction de datatime

Test Values

	id_prod	date	session_id	client_id
1431	T_0	test_2021-03-01 02:30:02.237420	s_0	ct_1
2365	T_0	test_2021-03-01 02:30:02.237446	s_0	ct_1
2895	T_0	test_2021-03-01 02:30:02.237414	s_0	ct_1
5955	T_0	test_2021-03-01 02:30:02.237441	s_0	ct_0
7283	T_0	test_2021-03-01 02:30:02.237434	s_0	ct_1
...	...	...	...	...
332594	T_0	test_2021-03-01 02:30:02.237445	s_0	ct_0
332705	T_0	test_2021-03-01 02:30:02.237423	s_0	ct_1
332730	T_0	test_2021-03-01 02:30:02.237421	s_0	ct_1
333442	T_0	test_2021-03-01 02:30:02.237431	s_0	ct_1
335279	T_0	test_2021-03-01 02:30:02.237430	s_0	ct_0

200 rows × 4 columns

Client

Data

Pour chaque items, elle correspondent à un client unique, un sexe , age

Problématique ?
Oui, car elle à des valeurs de test

Test Values

	client_id	sex	birth
2735	ct_0	f	2001
8494	ct_1	m	2001

Données manquantes

Data

Suite aux jointures des dataframes on se retrouve avec des données qui n'ont pas de prix et de catégorie.
Indice : code_produit liée à la catégorie
Et application d'une moyenne

CA avant = 5797674
CA après = 5797674

Donnée orpheline

	id_product	sell_date	session_id	client_id	transaction_date	sell_year	month	sex	user_birthday	price	category_id
6231	0_2245	2021-06-17 03:03:12.668129	s_49705	c_1533	2021-06-17	2021	6	m	1972	NaN	NaN
10797	0_2245	2021-06-16 05:53:01.627491	s_49323	c_7954	2021-06-16	2021	6	m	1973	NaN	NaN
14045	0_2245	2021-11-24 17:35:59.911427	s_124474	c_5120	2021-11-24	2021	11	f	1975	NaN	NaN
17480	0_2245	2022-02-28 18:08:49.875709	s_172304	c_4964	2022-02-28	2022	2	f	1982	NaN	NaN
21071	0_2245	2021-03-01 00:09:29.301897	s_3	c_580	2021-03-01	2021	3	m	1988	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
322523	0_2245	2021-04-06 19:59:19.462288	s_16936	c_4167	2021-04-06	2021	4	f	1979	NaN	NaN
329226	0_2245	2021-03-30 23:29:02.347672	s_13738	c_7790	2021-03-30	2021	3	f	1983	NaN	NaN
330297	0_2245	2021-12-03 14:14:40.444177	s_128815	c_6189	2021-12-03	2021	12	f	1984	NaN	NaN
335331	0_2245	2021-04-27 18:58:47.703374	s_26624	c_1595	2021-04-27	2021	4	f	1973	NaN	NaN
336020	0_2245	2021-05-01 03:35:03.146305	s_28235	c_5714	2021-05-01	2021	5	f	1972	NaN	NaN

103 rows × 11 columns

Enregistrement des données

Une fois les données approuvées et conforme à nos attentes nous allons les enregistrer dans un dossier ‘data_clear_values’. Ce dossier contient nos 4 dataframes différents pour rappel :

1. Clients
2. Produits
3. Transactions
4. Merge (fusion des de ces dataframes)

Récapitulatif : traitement des données

Sur l’ensembles des dataframes, suites aux divers traitement effectuez une pertes de plus de 400 lignes sur l’ensembles des données. Il faut donc informer le clients de cette perte et comment la palier

Analyse de données

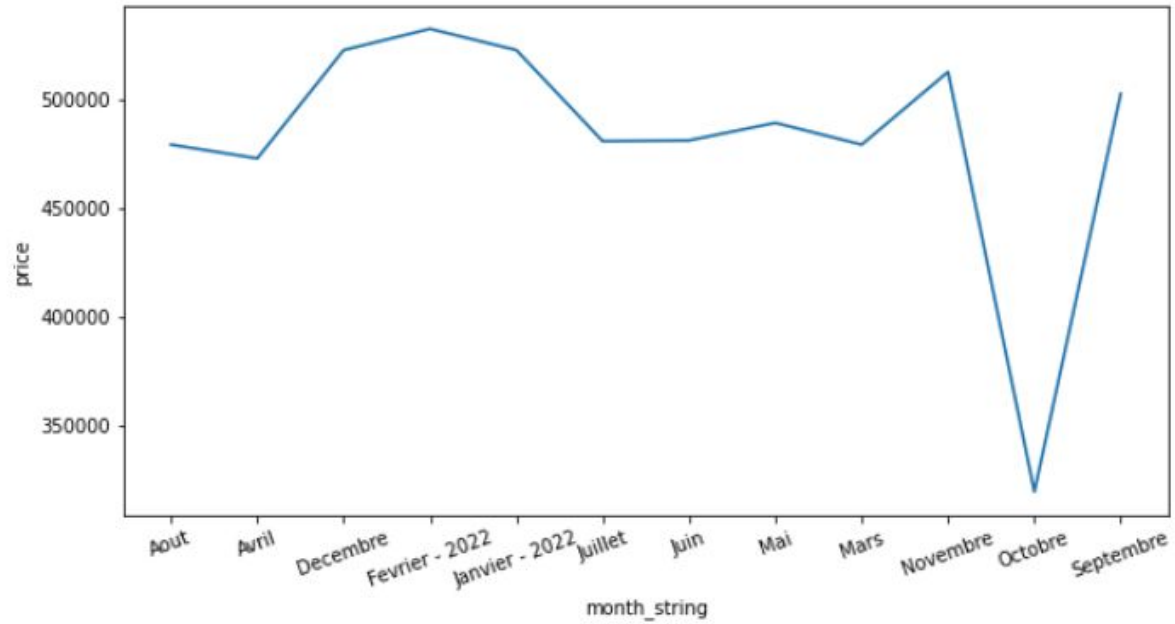
Première analyses

En 2021 : cette librairie réalise un chiffre d'affaire de de plus de **4 millions d'euros en CA.**

Nous pouvons décomposer les ventes en 3 grande catégorie.
La catégorie 0 représente **39% des transactions.**
On remarque que ces catégories se traduit aussi en gamme de prix.

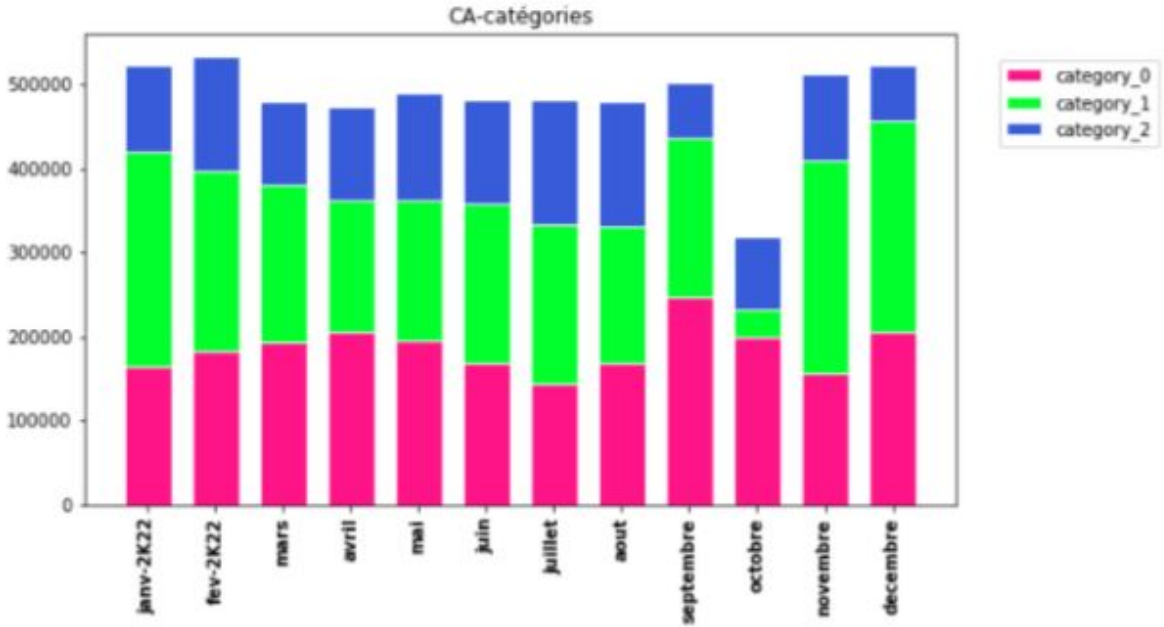
La société à générer plus de **250 000** transactions en 12 mois.

Chiffre d'affaires



Ca sur 12 mois

Chiffre d'affaires / Catégorie

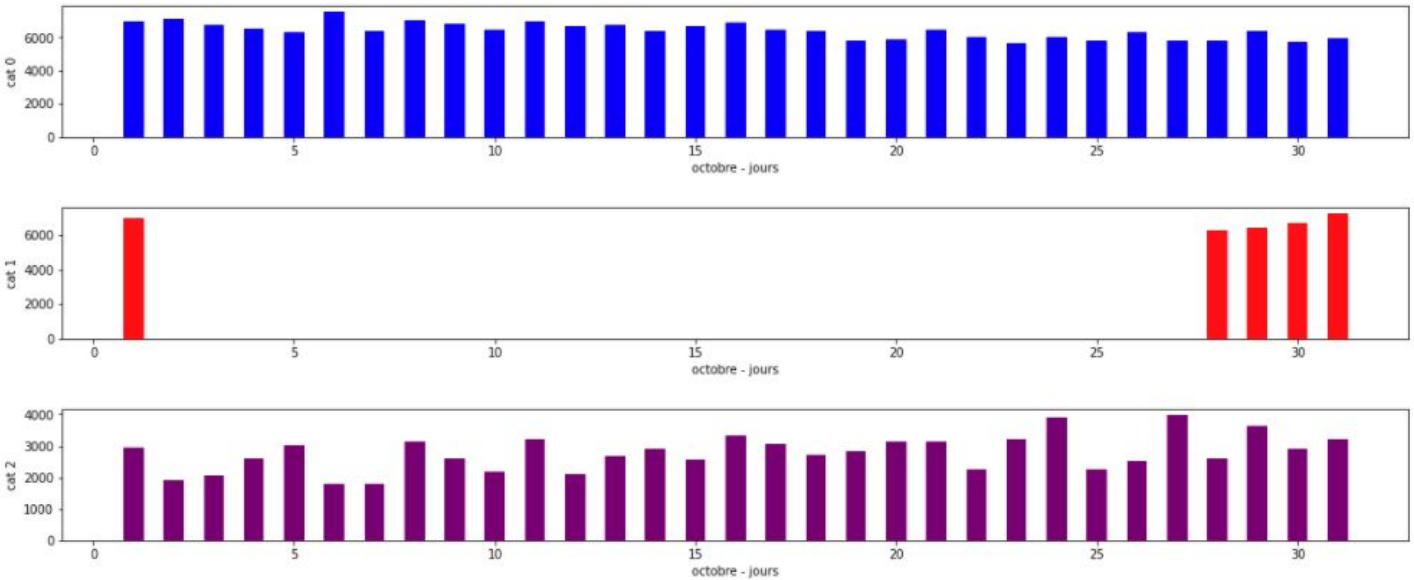


Ca / Catégories par Mois / sur 12 Mois

Chiffre d’affaires octobre

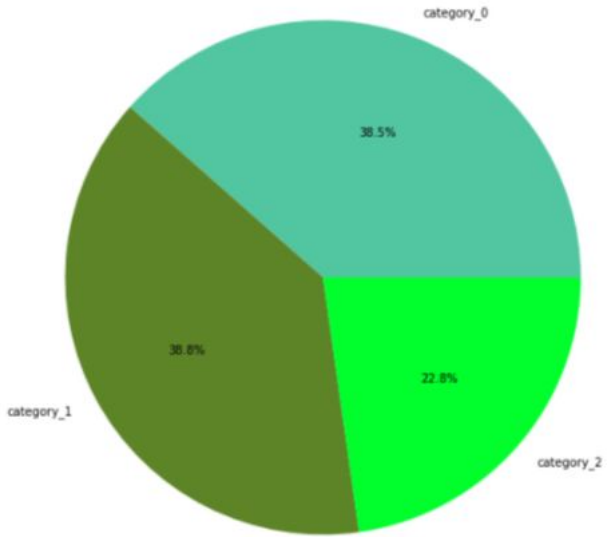
categories

On constate un manque de data. Il peut s’expliquer de deux manière soit un manque de donnée. Soit il y a eu un bug des serveurs. On enregistre que 5 jour pour la catégorie 1



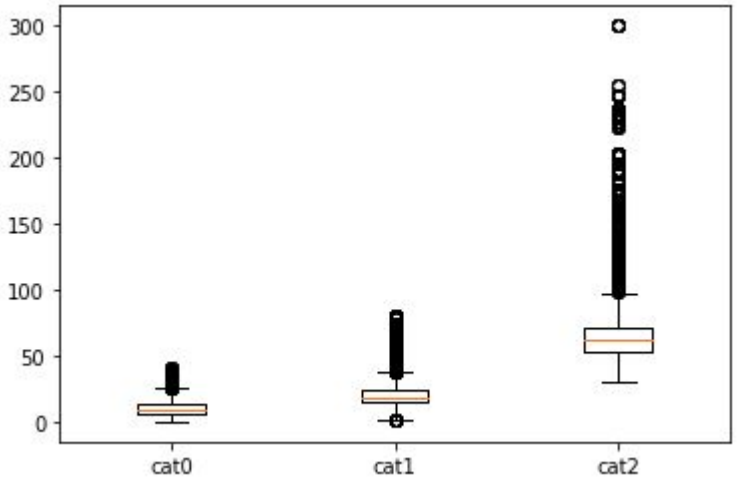
Ca / catégorie

Par catégorie / transaction



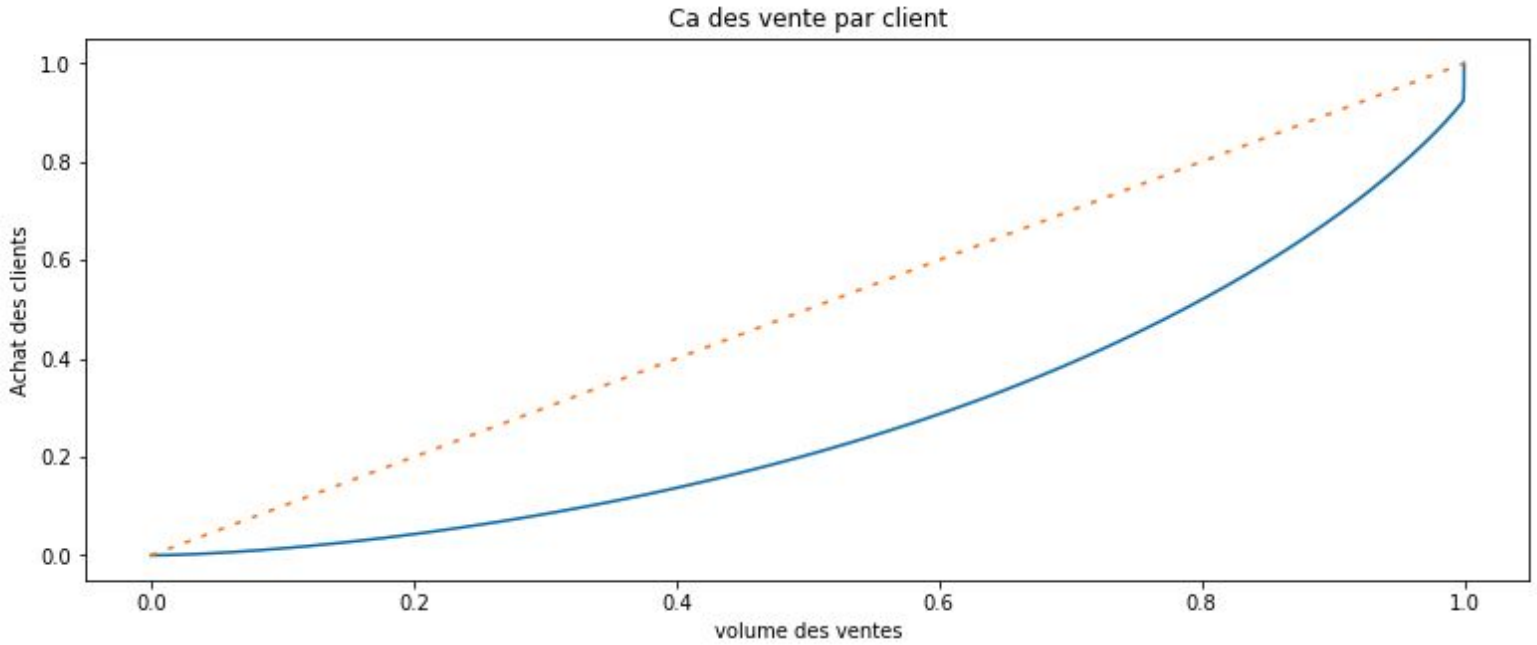
Répartition des transactions par catégories

Catégorie / achat



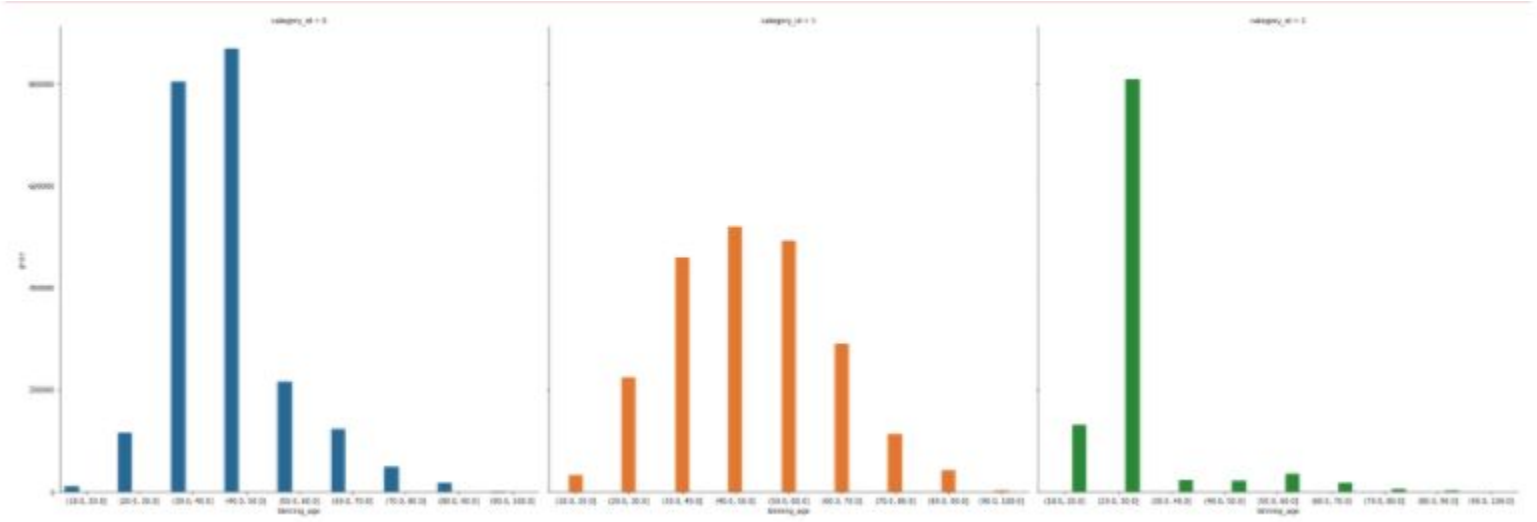
Les dépense via les catégories.

Volumes des ventes / clients



Indice de gini de 0.4 : corréspond a deux fois l'air entre la courbe bleu et les pointillés.(plus elle est basse plus elle est inégalitaire)
lorenz curve (très peu de client représente beaucoup de CA)

Analyse bi-varié CA = Catégorie / tranche d'âge



Identification des plus gros clients

4 plus gros clients

On constates que c'est 4 clients représente plus de 7% du chiffre d'affaire global sur ces 12 mois. (hypothèse c'est sans doute une école ou une université)

CA par Client

	client_id	price	client_id	sex	user_age
677	c_1609	162007.34	c_8233	m	58
4388	c_4958	144257.21	c_1123	f	42
6337	c_6714	73197.34	c_1503	m	35
2724	c_3454	54442.92	c_1476	m	50

Synthèse de l'analyse de donnée

Suite à cette première analyse on constate :

- Plus de 50% des clients sont de la même catégorie
- Une baisse significative du CA en Octobre
- Les 3 plus gros clients représentent plus de 4% du Chiffre d'affaire global (représenté aussi par la courbe de lorenz
- Le panier moyen toute catégorie est de 33 euros

Corrélation des variables

Corrélation des variables

Le but d’avoir des grosses données l’optimisation du CA dans ce cas précis en extrayant différentes variable on peut voir le comportement des consommateurs

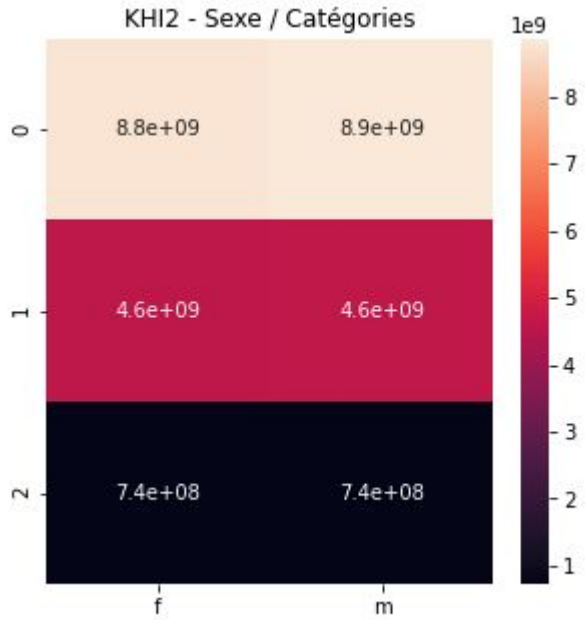
Y-a t-il une corrélation entre le sexe des client et les catégories achetés ?

CHI-2

Une utilise cette technique afin de voir si deux variable qualitative sont corréler.
Visuellement on utilise un tableau de contingence

La p-value : 1
Chi-2 : 1.92

Légère corrélation entre la catégorie 1 et 2 . La troisième n'impacte pas .



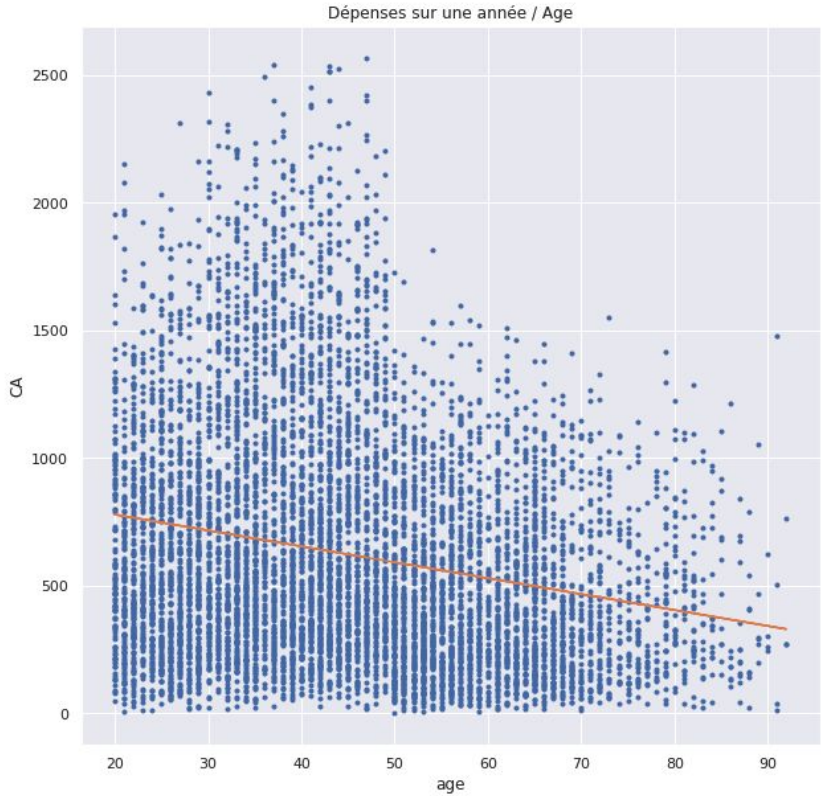
Y-a t-il une corrélation entre l'âge des clients et le montant des achats ?

Régression linéaire

Une utilise cette technique de corrélation car les variables sont quantitatives.

Corrélation de pearson = 0.2
Cette corrélation négative faible entre l'âge des clients et leur dépenses.

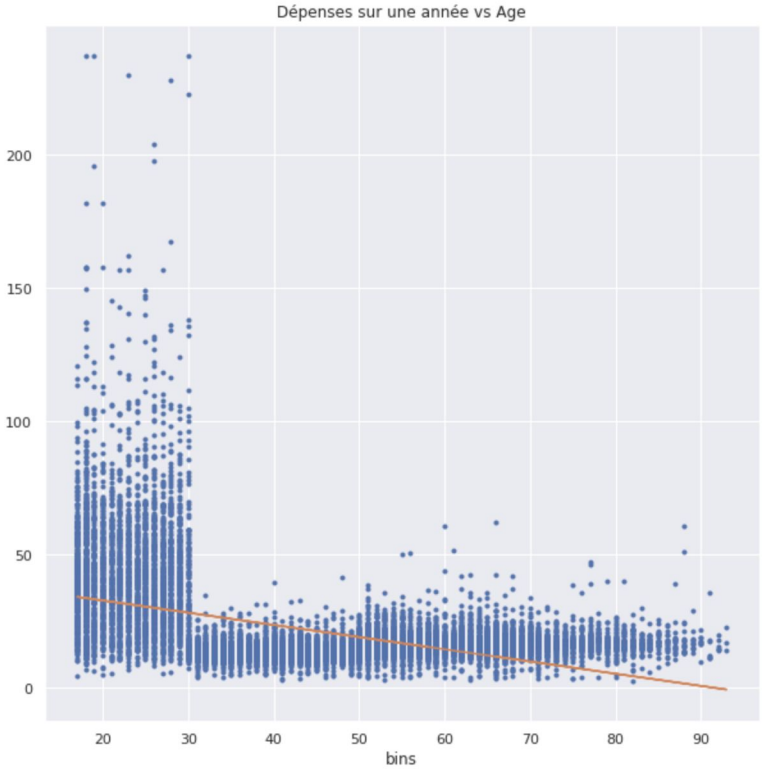
P-val : 3.8



Y-a t-il une corrélation entre l'âge des clients et le montant des achats ?

Dispersion

Une utilise cette technique de corrélation car les variables sont quantitatives.



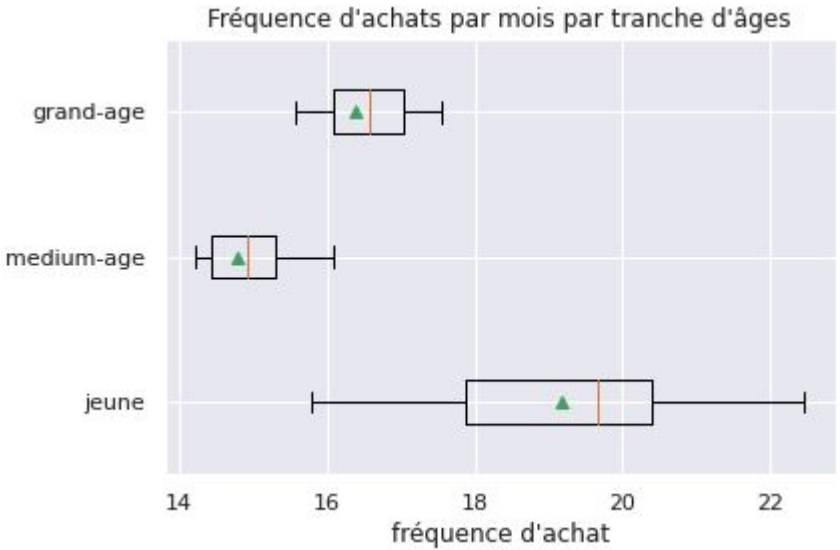
Y-a t-il une corrélation entre l'âge et la fréquence des achats ?

Box Plot

On peut faire le rapprochement avec la régression linéaire.

L'écart type des jeunes et ceux qui achète le plus souvent

Etat**2 = 0.37



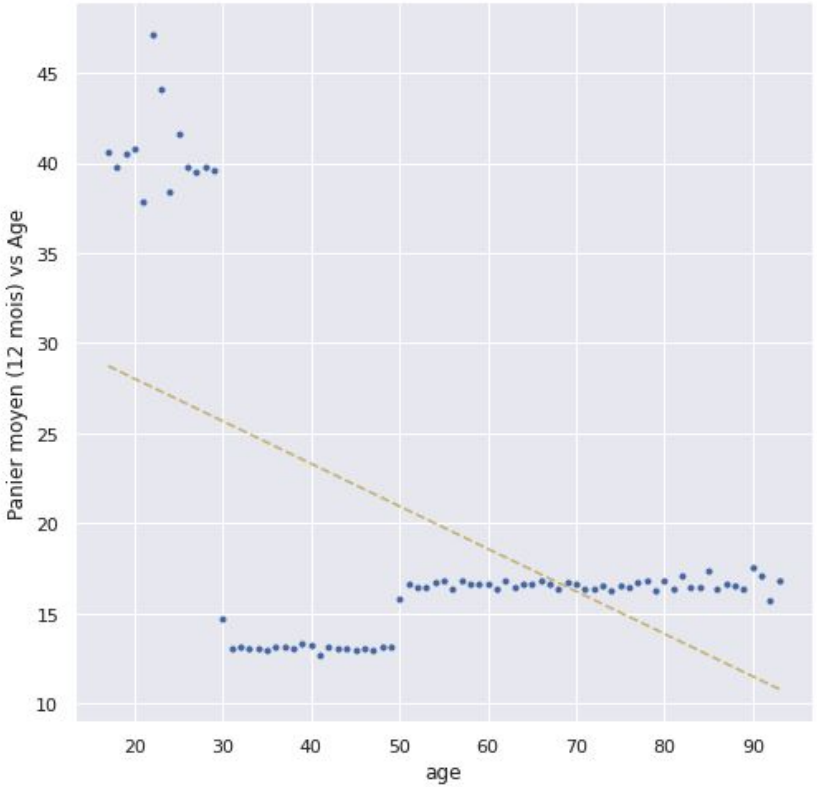
Y-a t-il une corrélation entre l'âge et le panier moyen ?

Regréssion

Tranche d'âge et panier moyen

Cette régression n'est pas explicative. On référence que 30%. On constate visuellement qu'il n'y a pas de corrélation entre elle.

R**2 = 0.30

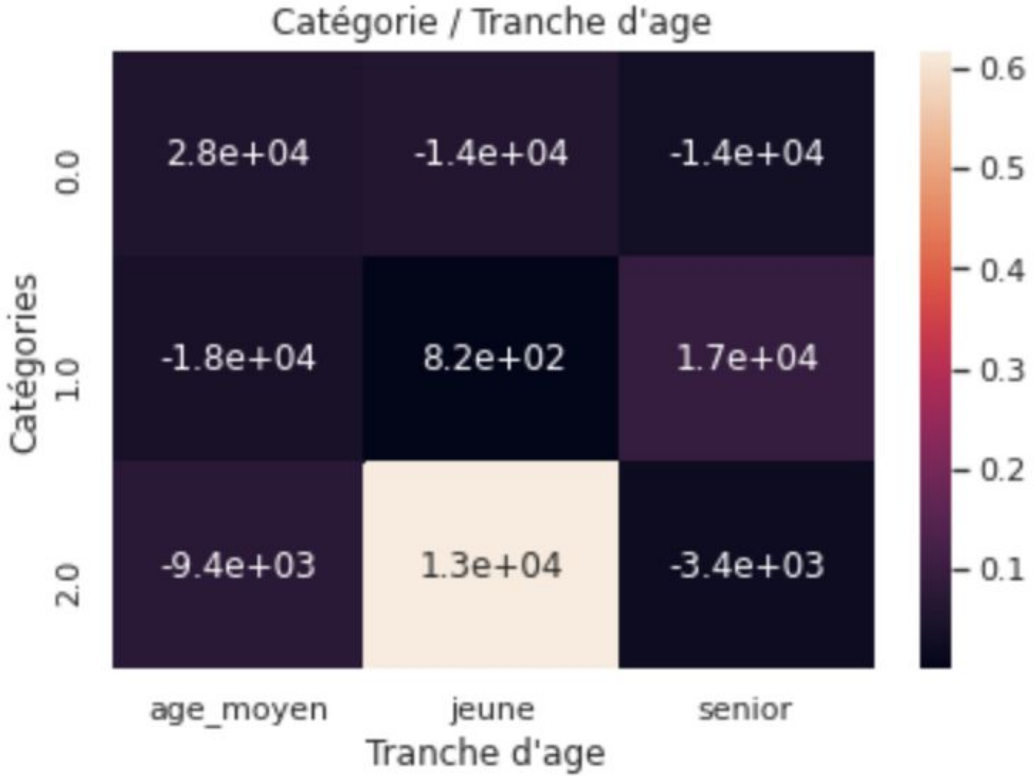


Y-a t-il une corrélation entre l'âge et les catégorie + produit achetées ?

Contingence

Intérêt des jeune pour la catégorie 2.
Pour le autres tranches d'âges.
Il n'y a pas d'impact significatif entre les autres catégories et les autres tranches d'âges.

xi_n = 0
P-value = 0



Conclusion

1. Une augmentation du chiffre d'affaire à part au mois d'octobre grosse baisse
2. Le sexe n'interfaire pas avec le CA
3. Les 30-40 sont ce qui dépensent le plus
4. Il y a une niche commercial pour les plus de 40 ans car ce groupe achète les livre les plus chères
5. Cela nécessite énormément de traitement avant d'obtenir des datas clear

Questions ?