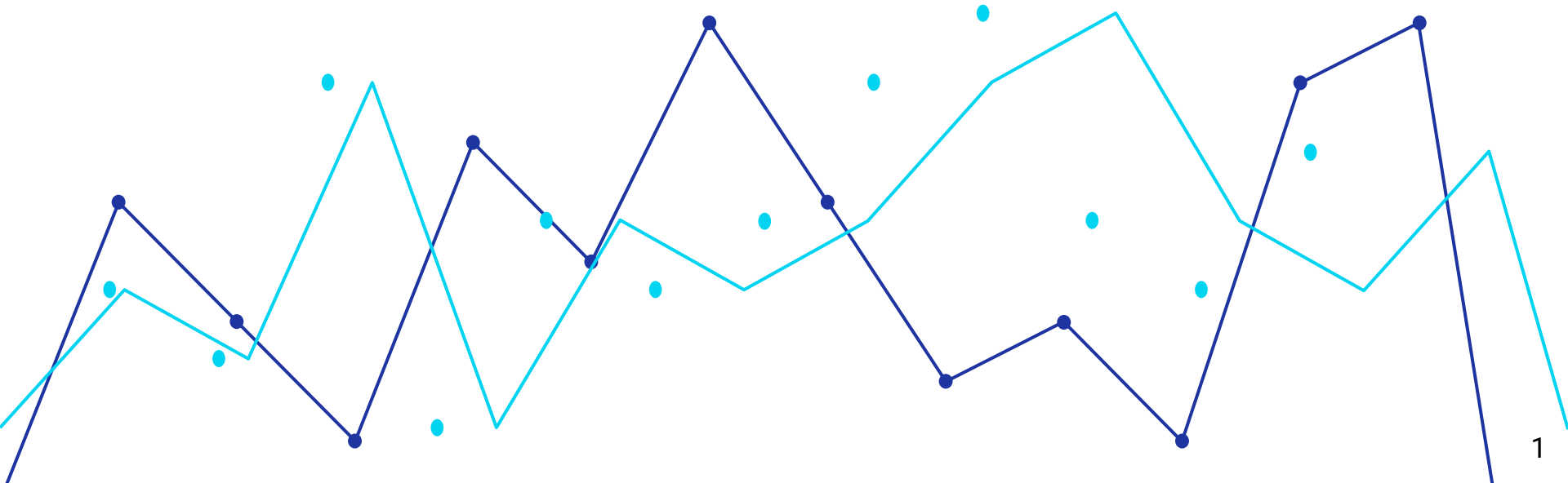
Data Analyst



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 qui se nomme jupyter tout c’est processus sont soumis à une gestion des containers 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

Orphan Data

Data

Afin de pouvoir interpréter les différentes valeurs, nous devons donc les fusionner en une seule dataframe. Cependant, selon la méthod de jointure on se retrouve avec une second difficulté. Nous obtenons des données orpheline

Donnée orpheline

	client_id	sex	user_birthday	id_product	sell_date	session_id	transaction_date	sell_year	month	price	category_id
266960	c_4505	m	1976.0	0_2245	2022-01-09 09:23:31.000720	s_147220	2022-01-09	2022.0	1.0	17.216434	NaN
266961	c_3468	f	1981.0	0_2245	2021-09-11 10:52:05.205583	s_88251	2021-09-11	2021.0	9.0	17.216434	NaN
266962	c_1403	f	1978.0	0_2245	2022-02-15 14:26:50.187952	s_165575	2022-02-15	2022.0	2.0	17.216434	NaN
266963	c_3065	f	1977.0	0_2245	2022-01-26 13:34:33.440366	s_155484	2022-01-26	2022.0	1.0	17.216434	NaN
266964	c_7102	m	1983.0	0_2245	2021-04-25 19:58:42.716401	s_25704	2021-04-25	2021.0	4.0	17.216434	NaN
...	...	...	...	...	...	...	...	...	...	...	...
336833	NaN	NaN	NaN	0_525	NaN	NaN	NaN	NaN	NaN	2.990000	0.0
336834	NaN	NaN	NaN	2_86	NaN	NaN	NaN	NaN	NaN	132.360000	2.0
336835	NaN	NaN	NaN	0_299	NaN	NaN	NaN	NaN	NaN	22.990000	0.0
336836	NaN	NaN	NaN	0_510	NaN	NaN	NaN	NaN	NaN	23.660000	0.0
336837	NaN	NaN	NaN	0_2308	NaN	NaN	NaN	NaN	NaN	20.280000	0.0

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

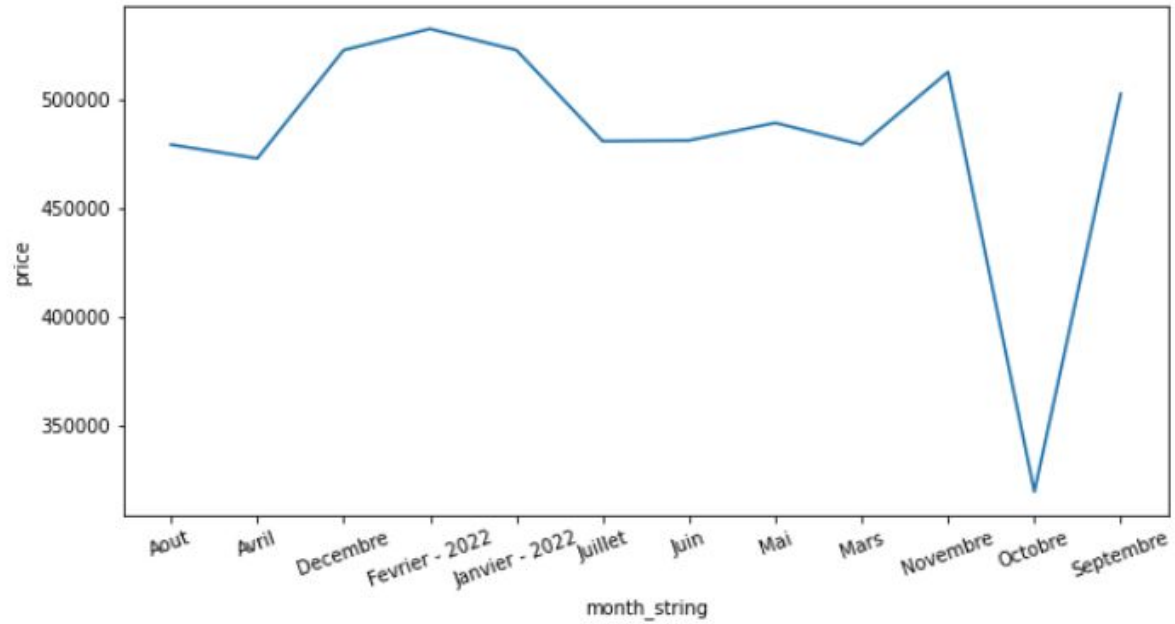
Première analyses

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.

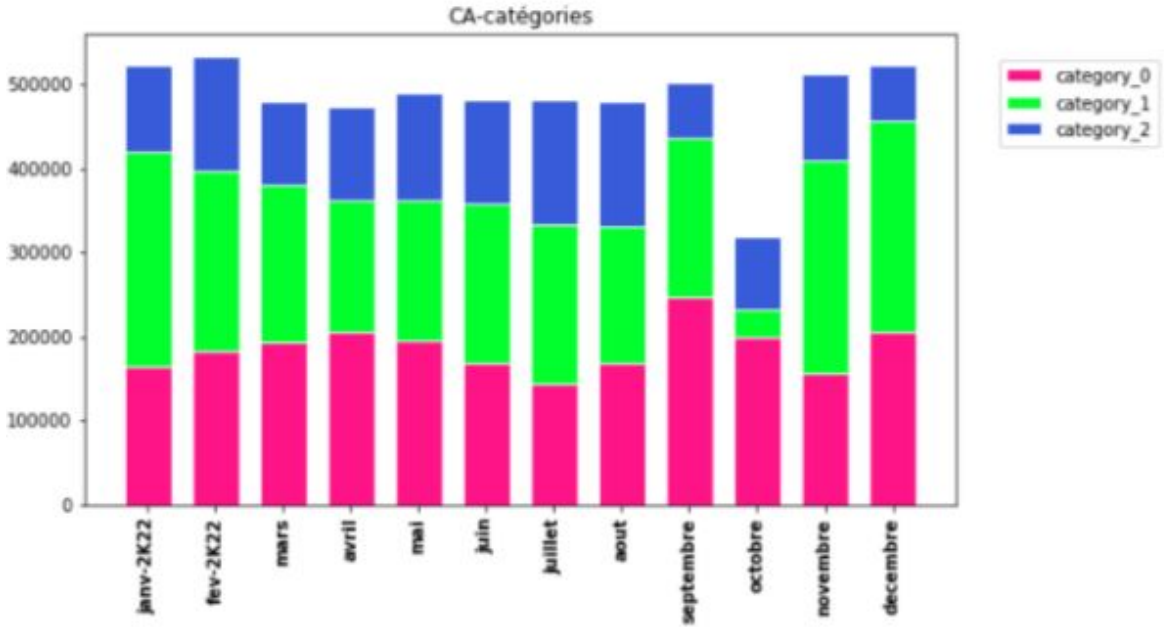
La société à générer plus de **250 000** transactions en 12 mois.

Chiffre d'affaires



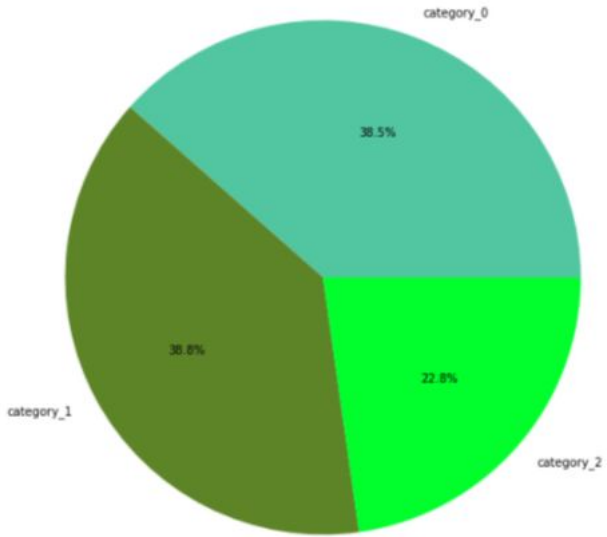
Ca sur 12 mois

Chiffre d'affaires / Catégorie



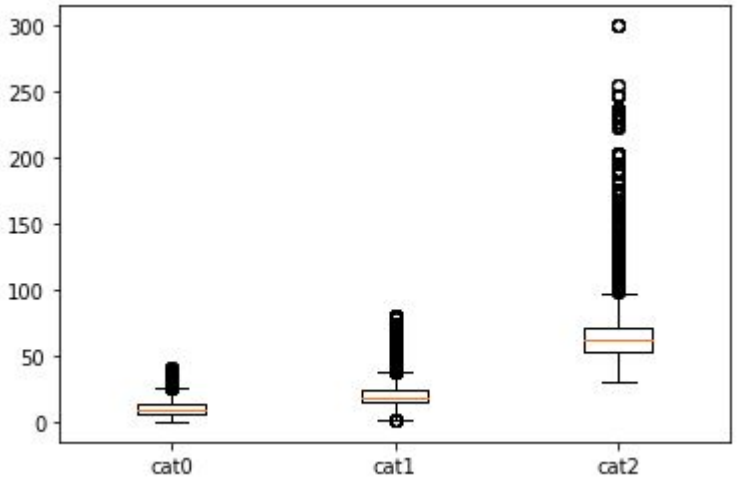
Ca / Catégories par Mois / sur 12 Mois

Par catégorie / transaction



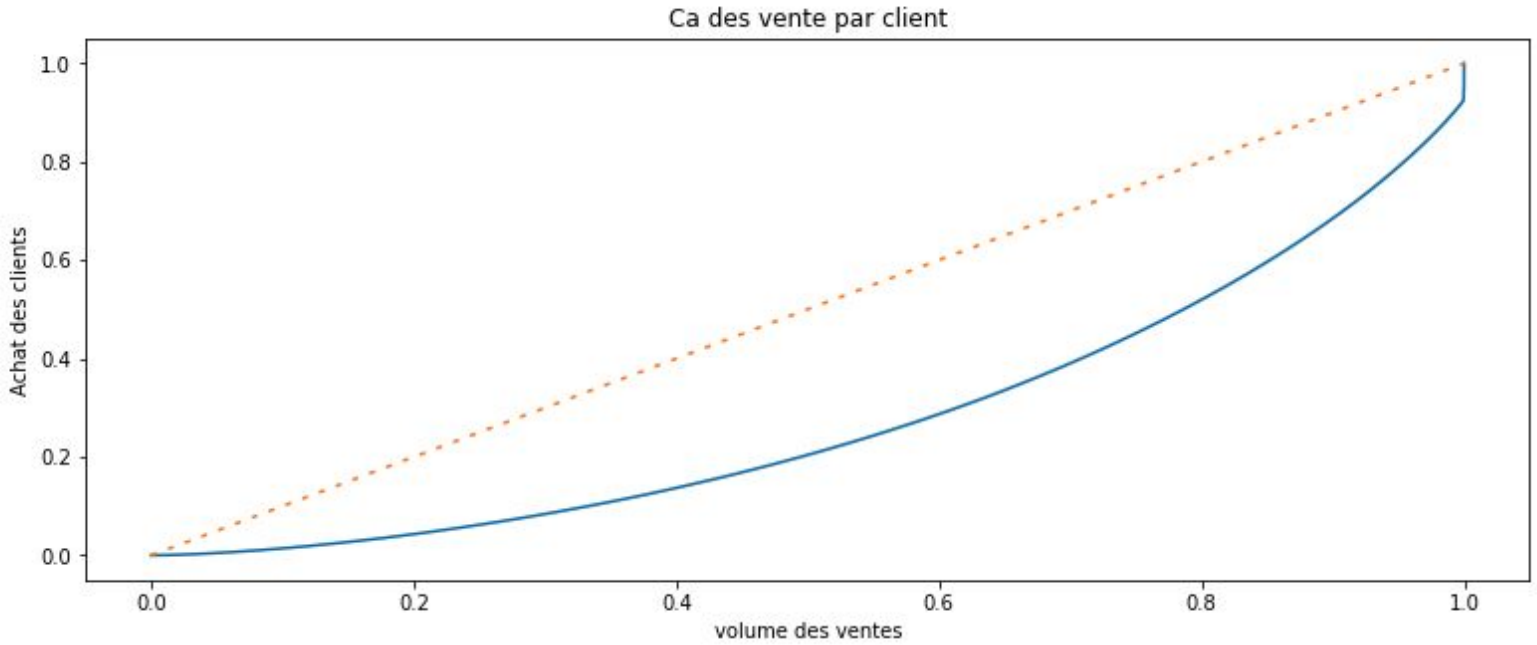
Répartition des transactions par catégories

Catégorie / achat



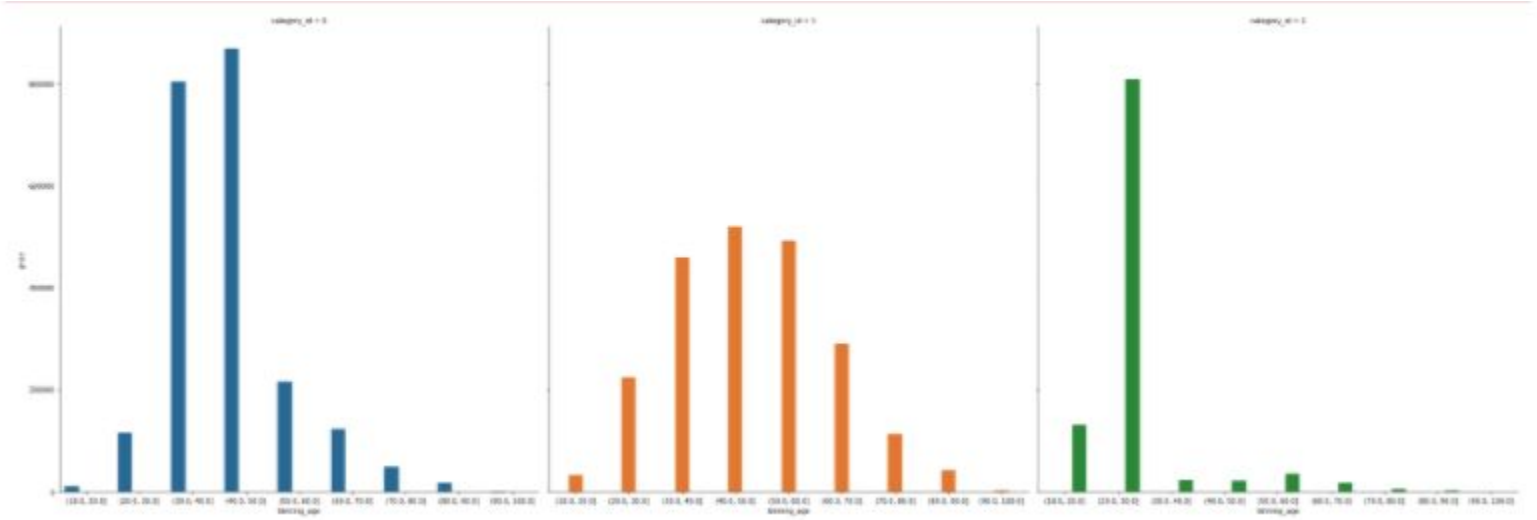
Les dépense via les catégories.

Volumes des ventes / clients



Indice de gini de 0.4 / lorenz curve (très peu de client représente beaucoup de CA)

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



Gamme de produit

Mean

En faisant une moyennes de prix par catégorie on peut déduire que le produit peut être catégorisé en tant que gamme

	category_id	price
0	0	10.646828
1	1	20.480106
2	2	75.174949

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
- 0n peut redéfinir les catégories en gamme de produit
- 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

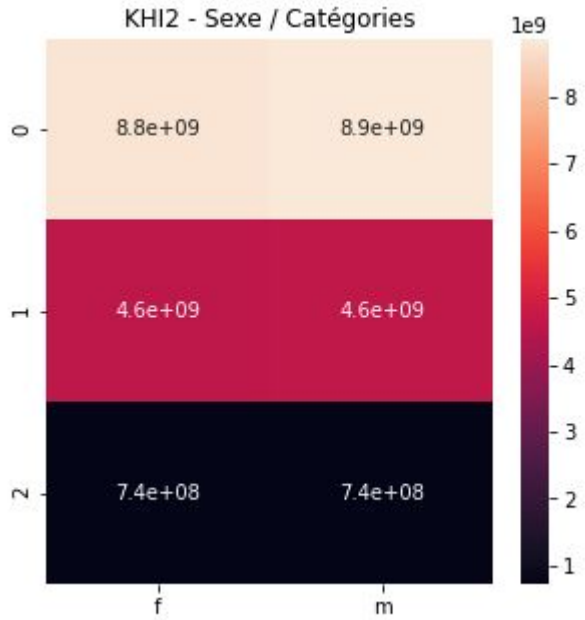
Y-a t-il une corrélation entre le sexe des client et les catégories achetés ?

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 .



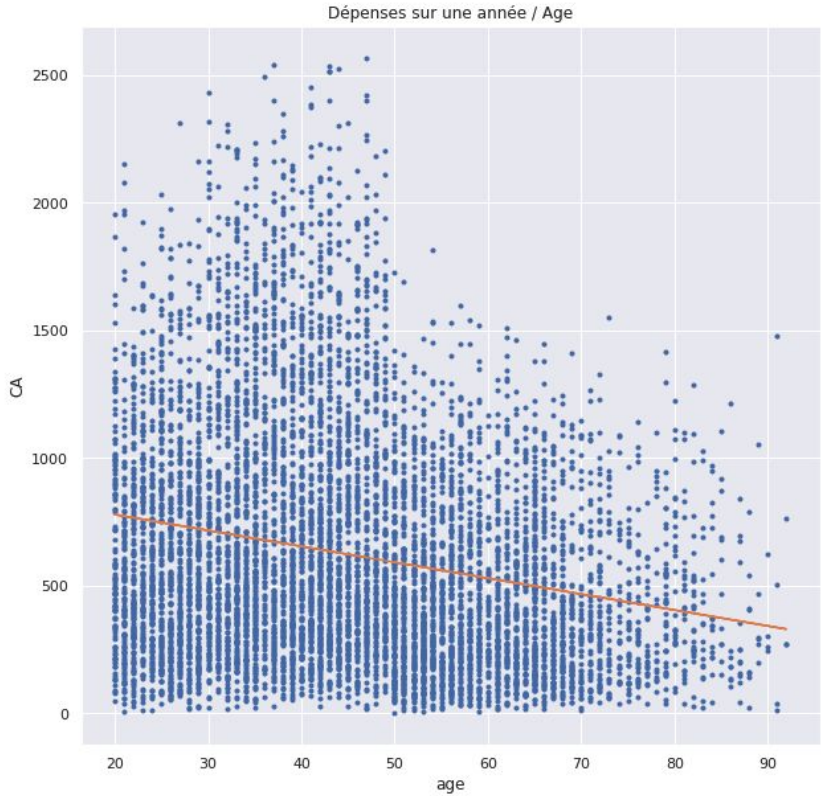
Y-a t-il une corrélation entre l'âge des clients et le montant des achats ?

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



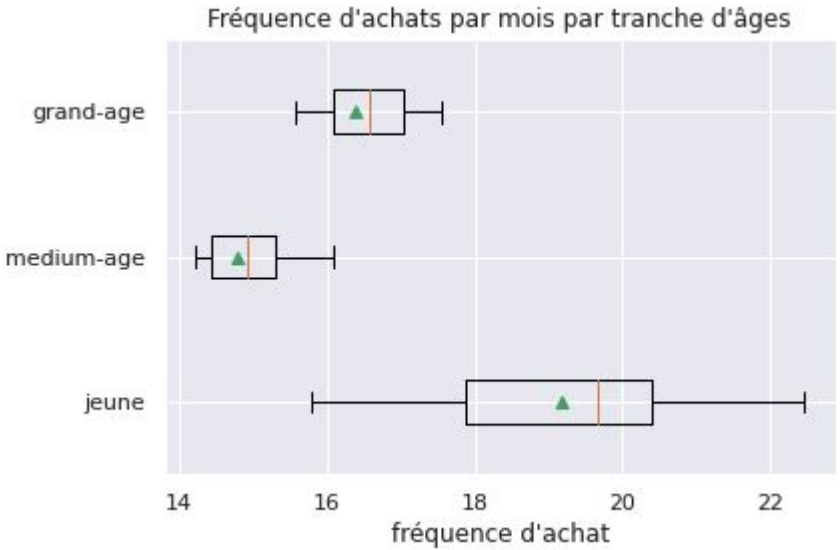
Y-a t-il une corrélation entre l'âge et la fréquence des achats ?

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



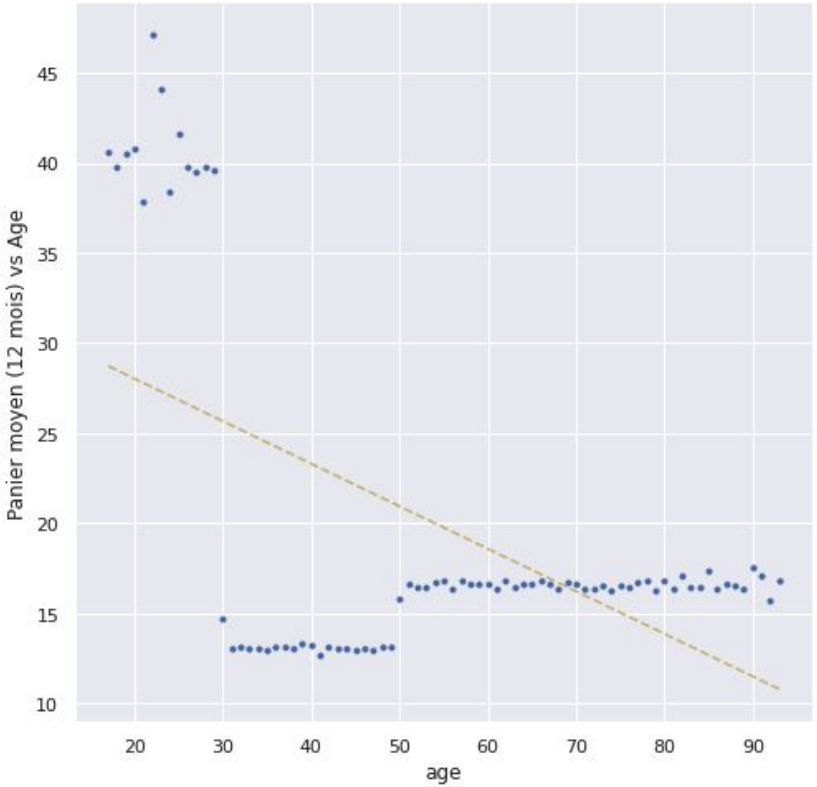
Y-a t-il une corrélation entre l'âge et le panier moyen ?

Regréssion linéaire

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

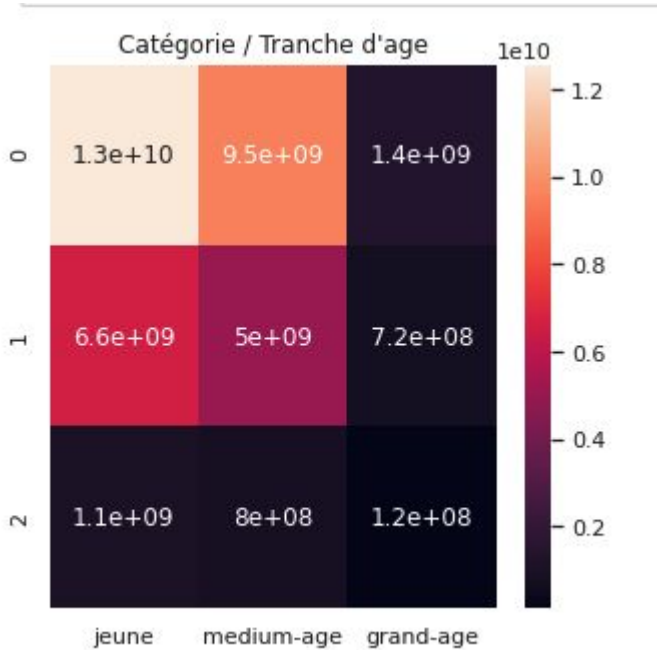


Y-a t-il une corrélation entre l'âge et les catégorie + produit achetées ?

Contingence

Ce tableau confirme qu'il y a bien une corrélation entre l'âge et les catégories

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 ?