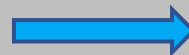
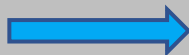






Lena Verboom

Analyse des ventes de Rester livres

Projet 4 : Analysez les ventes de votre entreprise



Contexte

- ❑ Analyse des données de ventes d'une grande chaîne de librairie : **Rester livres** 
- ❑ Entreprise française avec plusieurs magasins et une **boutique en ligne** 
- ❑ Analyse des caractéristiques **clients** (âge, sexe), des **produits** (prix, catégories) et des **transactions**

Est-ce qu'il y a une corrélation entre les différentes caractéristiques des données ?

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Données à ma disposition :

3 fichiers
CSV

Liste des clients

8 623 clients

Customers.csv

Liste des produits

3 287 produits

products.csv

Ventes

337 016 transactions

Transaction.csv

	client_id	sex	birth
0	c_4410	f	1967
1	c_7839	f	1975

	id_prod	price	categ
0	0_1421	19.99	0
1	0_1368	5.13	0

	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

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Suppression des valeurs négatives dans df_produits :

df_produits

	price	categ
count	3287.000000	3287.000000
mean	21.856641	0.370246
std	29.847908	0.615387
min	-1.000000	0.000000
25%	6.990000	0.000000
50%	13.060000	0.000000
75%	22.990000	1.000000
max	300.000000	2.000000

Prix négatif : impossible



Suppression des prix
inférieur à 0

df_produits

	price	categ
count	3286.000000	3286.000000
mean	21.863597	0.370359
std	29.849786	0.615446
min	0.620000	0.000000
25%	6.990000	0.000000
50%	13.075000	0.000000
75%	22.990000	1.000000
max	300.000000	2.000000

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Suppression des sessions de test dans df_ventes :

df_ventes

	id_prod	date	session_id	client_id
count	337016	337016	337016	337016
unique	3266	336855	169195	8602
top	1_369	test_2021-03-01 02:30:02.237413	s_0	c_1609
freq	1081	13	200	12855

Sélection des 200 sessions tests avec « s_0 »

	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

df_ventes_sans_test

	id_prod	date	session_id	client_id
count	336816	336816	336816	336816
unique	3265	336816	169194	8600
top	1_369	2021-06-12 00:15:34.384065	s_118668	c_1609
freq	1081	1	14	12855

Suppression des tests avec id_prod = T_0

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Merge des dataframes df_ventes_sans_test et df_clients :

df_ventes_sans_test

	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

+

df_clients

	client_id	sex	birth
0	c_4410	f	1967
1	c_7839	f	1975
2	c_1699	f	1984

Jointure de type inner sur client_id

df_clients_ventes

	id_prod	date	session_id	client_id	sex	birth
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450	f	1977
1	1_596	2021-09-12 02:11:24.774608	s_88567	c_4450	f	1977
2	1_278	2021-09-10 15:09:01.555889	s_87835	c_4450	f	1977

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Merge des dataframes df_clients_ventes et df_produits :

df_client_ventes

	id_prod	date	session_id	client_id	sex	birth
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450	f	1977
1	1_596	2021-09-12 02:11:24.774608	s_88567	c_4450	f	1977
2	1_278	2021-09-10 15:09:01.555889	s_87835	c_4450	f	1977



df_produits

	id_prod	price	categ
0	0_1421	19.99	0
1	0_1368	5.13	0
2	0_731	17.99	0

Jointure de type left sur id_prod



df_client_ventes
_produits

	id_prod	date	session_id	client_id	sex	birth	price	categ
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450	f	1977	4.99	0.0
1	1_596	2021-09-12 02:11:24.774608	s_88567	c_4450	f	1977	11.12	1.0
2	1_278	2021-09-10 15:09:01.555889	s_87835	c_4450	f	1977	19.18	1.0

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Recherche d'anomalies après la jointure left :

df_client_ventes_produits

```
id_prod      0
date          0
session_id    0
client_id     0
sex           0
birth         0
price        103
categ        103
dtype: int64
```

Sélection
des produits
concernés

	id_prod	date	session_id	client_id	sex	birth	price	categ
107	0_2245	2021-04-10 06:15:32.619826	s_18510	c_277	f	2000	NaN	NaN
3572	0_2245	2021-12-10 21:31:18.303110	s_132471	c_3519	m	1974	NaN	NaN
19047	0_2245	2021-04-10 09:22:57.768041	s_18566	c_8240	f	1978	NaN	NaN

Produit 0_2245 : absence des
caractéristiques prix et
catégorie

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Remplacement des valeurs NaN :

df_client_ventes_produits

id_prod							date	session_id	client_id	sex	birth	price	categ
107	0_2245	2021-04-10 06:15:32.619826		s_18510	c_277	f	2000	NaN	NaN				
3572	0_2245	2021-12-10 21:31:18.303110		s_132471	c_3519	m	1974	NaN	NaN				
19047	0_2245	2021-04-10 09:22:57.768041		s_18566	c_8240	f	1978	NaN	NaN				

Remplacement NaN par la moyenne
des prix des produits de la catégorie 0

Remplacement
NaN par la
catégorie 0

	id_prod	date	session_id	client_id	sex	birth	price	categ
107	0_2245	2021-04-10 06:15:32.619826	s_18510	c_277	f	2000	10.646828	0.0
3572	0_2245	2021-12-10 21:31:18.303110	s_132471	c_3519	m	1974	10.646828	0.0
19047	0_2245	2021-04-10 09:22:57.768041	s_18566	c_8240	f	1978	10.646828	0.0

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Création des colonnes âge et tranche d'âge :

df_client_ventes_produits

	id_prod	date	session_id	client_id	sex	birth	price	categ	age	range
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450	f	1977	4.99	0.0	45	25-45_y
1	1_596	2021-09-12 02:11:24.774608	s_88567	c_4450	f	1977	11.12	1.0	45	25-45_y

Détermination
de l'âge : 2022 –
date de
naissance des
clients

Création d'une tranche d'âge des
clients selon les modalités suivantes :

Range	15-25_y	25-45_y	45-65_y	65_y_et_+
age	$15 < \text{age} \leq 25$	$25 < \text{age} \leq 45$	$45 < \text{age} \leq 65$	$\text{age} > 65$

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Transformation de la date en datetime :

df_client_ventes_produits

#	Column	Non-Null Count	Dtype
0	id_prod	336816 non-null	object
1	date	336816 non-null	object
2	session_id	336816 non-null	object
3	client_id	336816 non-null	object
4	sex	336816 non-null	object
5	birth	336816 non-null	int64
6	price	336816 non-null	float64
7	categ	336816 non-null	float64
8	age	336816 non-null	int64
9	range	336816 non-null	object



df_client_ventes_produits

#	Column	Non-Null Count	Dtype
0	id_prod	336816 non-null	object
1	date	336816 non-null	datetime64[ns]
2	session_id	336816 non-null	object
3	client_id	336816 non-null	object
4	sex	336816 non-null	object
5	birth	336816 non-null	int64
6	price	336816 non-null	float64
7	categ	336816 non-null	float64
8	age	336816 non-null	int64
9	range	336816 non-null	object

Transformation en datetime : Année - Mois - Jour
- Heure - Minutes - Seconde

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Suppression de l'heure dans la colonne date :

df_client_ventes_produits

	id_prod	date	session_id	client_id	sex	birth	price	categ	age	range
0	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450	f	1977	4.99	0.0	45	25-45_y
1	1_596	2021-09-12 02:11:24.774608	s_88567	c_4450	f	1977	11.12	1.0	45	25-45_y



Suppression heure, minute et seconde

	id_prod	date	session_id	client_id	sex	birth	price	categ	age	range
0	0_1483	2021-04-10	s_18746	c_4450	f	1977	4.99	0.0	45	25-45_y
1	1_596	2021-09-12	s_88567	c_4450	f	1977	11.12	1.0	45	25-45_y

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Création de la colonne month et year_month :

df_client_ventes_produits

	id_prod	date	session_id	client_id	sex	birth	price	categ	age	range	month
0	0_1483	2021-04-10	s_18746	c_4450	f	1977	4.99	0.0	45	25-45_y	4
1	1_596	2021-09-12	s_88567	c_4450	f	1977	11.12	1.0	45	25-45_y	9

Prise en compte du
mois uniquement

Prise en compte du
mois et de l'année

	id_prod	date	session_id	client_id	sex	birth	price	categ	age	range	month	year_month
0	0_1483	2021-04-10	s_18746	c_4450	f	1977	4.99	0.0	45	25-45_y	4	2021-04
1	1_596	2021-09-12	s_88567	c_4450	f	1977	11.12	1.0	45	25-45_y	9	2021-09

Nettoyage

Analyse
univariée

Analyse
bivariée

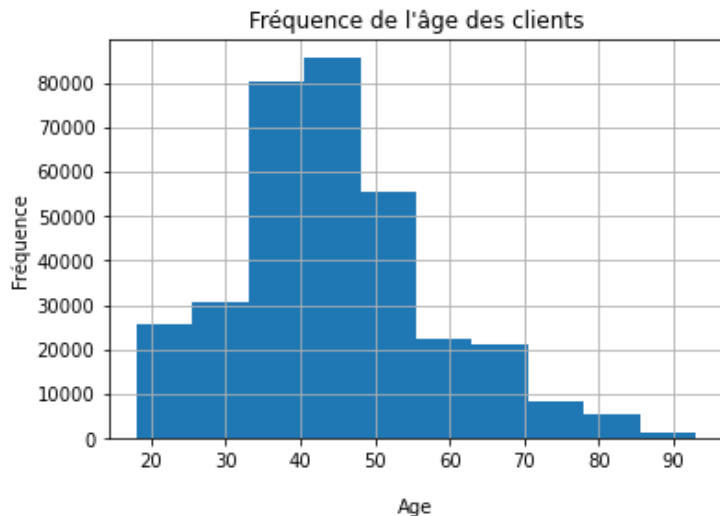
Analyse
statistique

Conclusion

Analyse des caractéristiques clients :

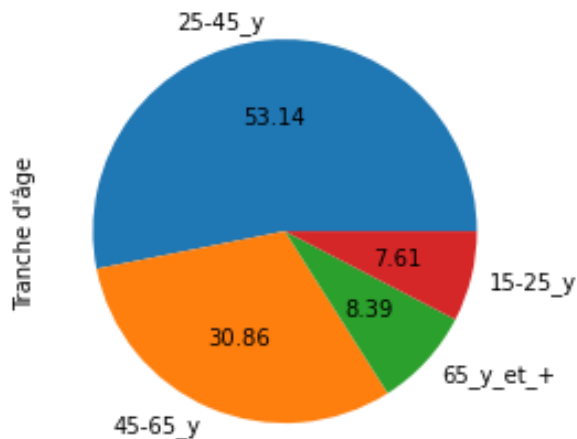
8600 clients

➔ Âge des clients



Minimum	Maximum	Moyenne	Médiane	Écart-type
18	93	44	42	13.5

Répartition des clients par tranche d'âge



La majorité des clients sont dans la tranche d'âge 25-45 (53%) et 45-65 (31%)

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

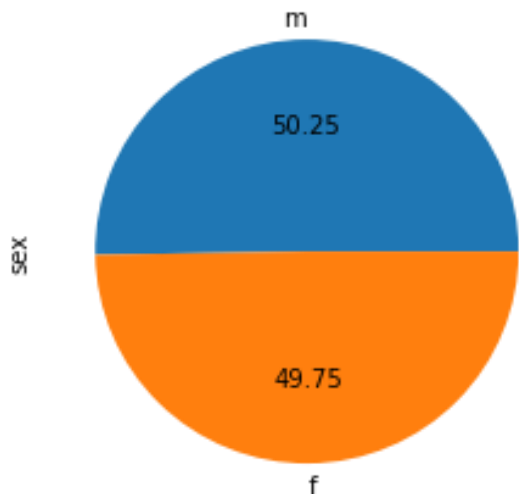
Analyse des caractéristiques clients :

8600 clients



Sexe des clients

Pourcentage de femme et d'homme parmi les clients



Répartition égalitaire entre les hommes et les femmes (environ 50/50)

Nettoyage

Analyse
univariéeAnalyse
bivariéeAnalyse
statistique

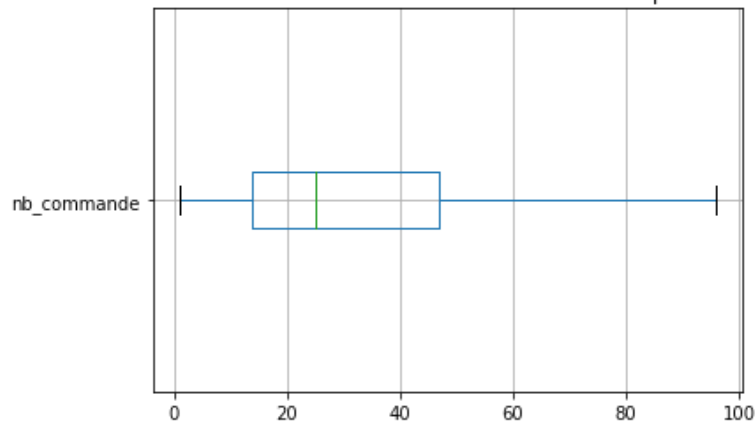
Conclusion

Analyse des caractéristiques clients :

8600 clients

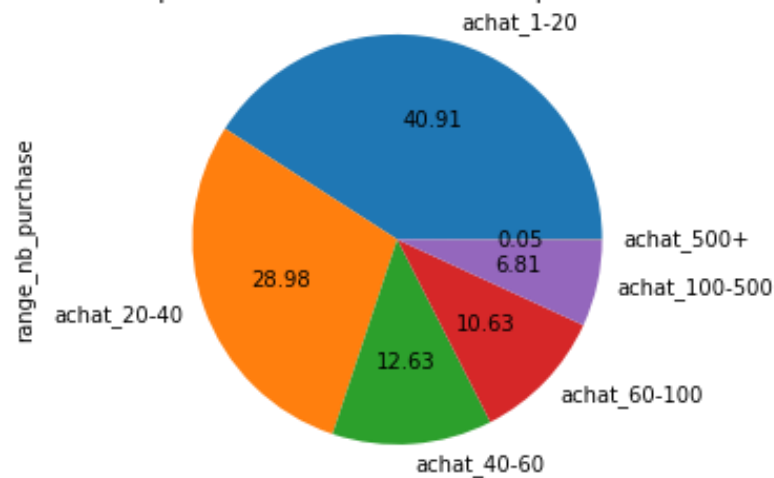
➔ Nombre de commande clients

Boîte à moustache des nombres de commande par client



Minimum	Maximum	Moyenne	Médiane	Écart-type
1	12855	39	25	156,4

Répartition du nombre d'achats par clients



La majorité des clients sont situés dans la tranche 1 à 20 achats (41%)

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

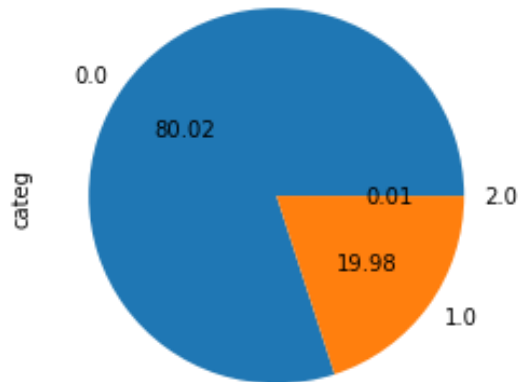
Analyse des caractéristiques clients : 8600 clients



Clients avec le plus d'achats

Répartition des catégories de produits achetés par le client 1609

Client_id	Nb_commande	Range_nb_purchase
C_1609	12855	Achat_500+
C_3454	3275	Achat_500+
C_4958	2562	Achat_500+
C_6714	4473	Achat_500+



**C_1609 est probablement une entreprise
et achètent majoritairement des produits
de la catégorie 0**

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

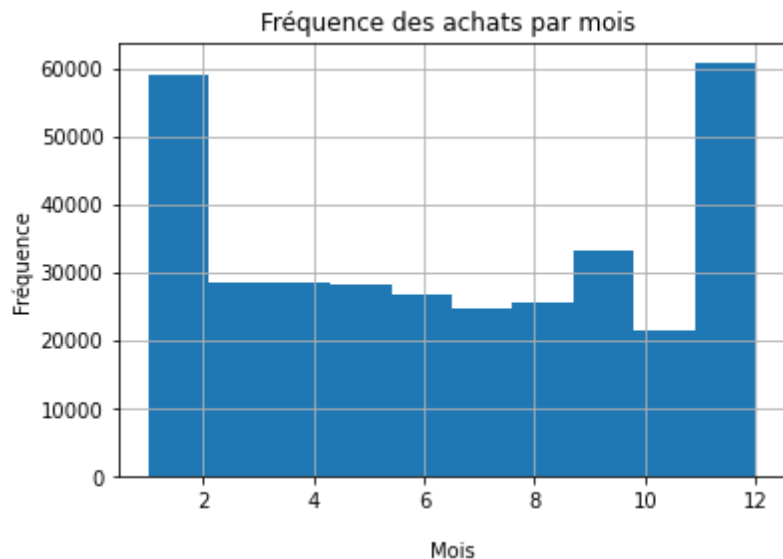
Conclusion

Analyse des caractéristiques clients :

8600 clients



Période d'achat



La majorité des achats clients sont effectués en janvier et en décembre

Nettoyage

Analyse
univariéeAnalyse
bivariéeAnalyse
statistique

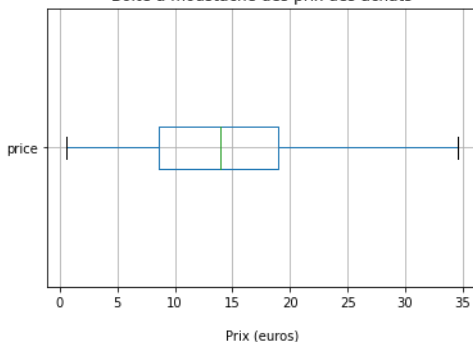
Conclusion

Analyse des caractéristiques produits : 3 265 produits

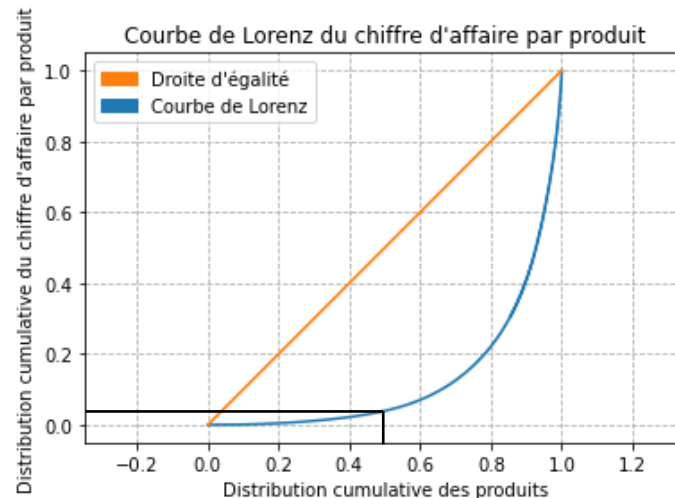
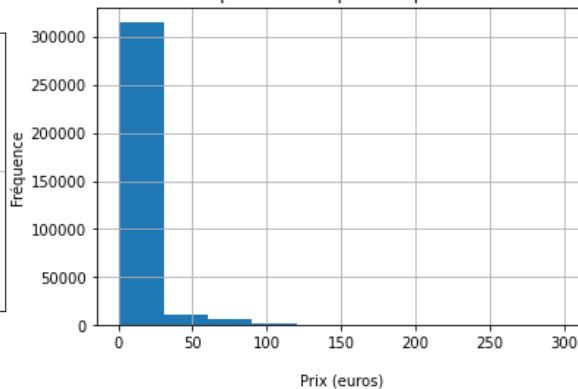


Prix des produits

Boîte à moustache des prix des achats



Répartition des prix des produits



Minimum	Maximum	Moyenne	Médiane	Écart-type	Coefficient de Gini
0.62	300	17.2	13.9	17.8	0.74

Forte inégalité entre le prix de chaque produit

Nettoyage

Analyse
univariée

Analyse
bivariée

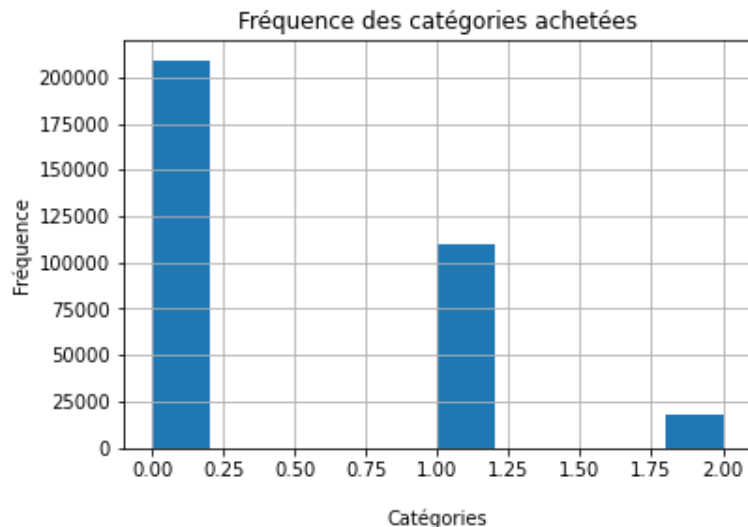
Analyse
statistique

Conclusion

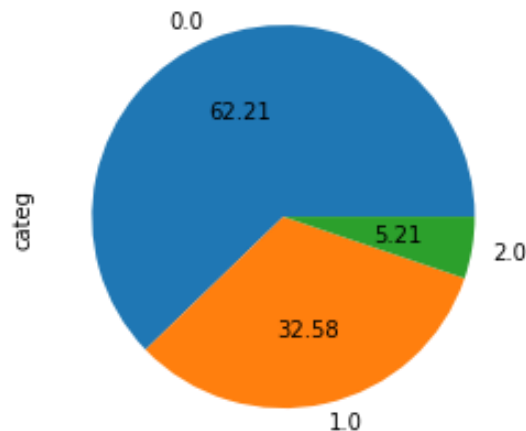
Analyse des caractéristiques produits : 3 265 produits



Catégories des produits



Répartition des catégories de produits achetés par les clients



Les produits de catégorie 0 sont achetés
plus fréquemment

Nettoyage

Analyse
univariée

Analyse
bivariée

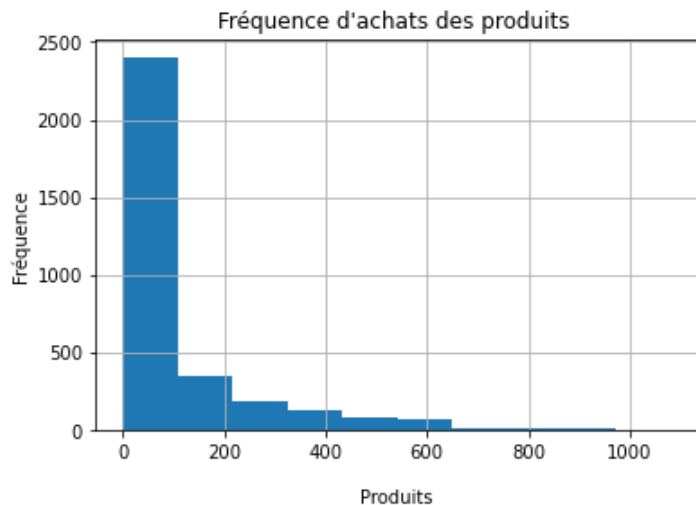
Analyse
statistique

Conclusion

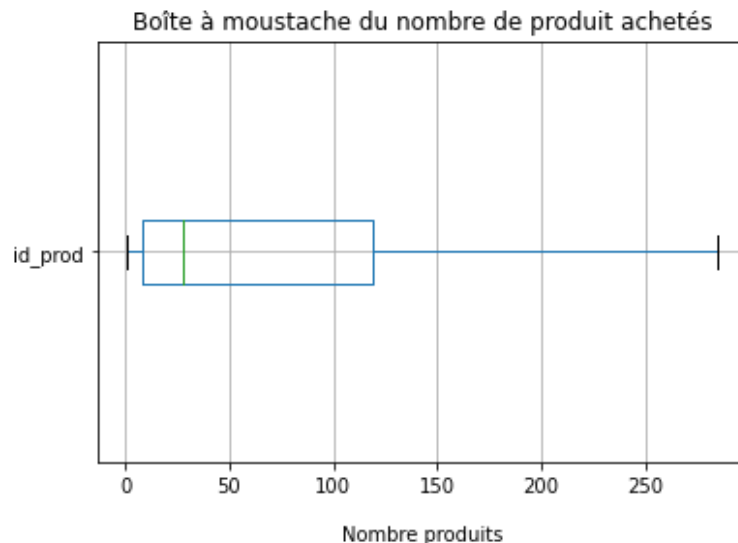
Analyse des caractéristiques produits : 3 265 produits



Nombre de produits achetés



Minimum	Maximum	Moyenne	Médiane	Écart-type
1	1081	103	28	163.2



La majorité des produits sont achetés
entre 1 et 120 fois

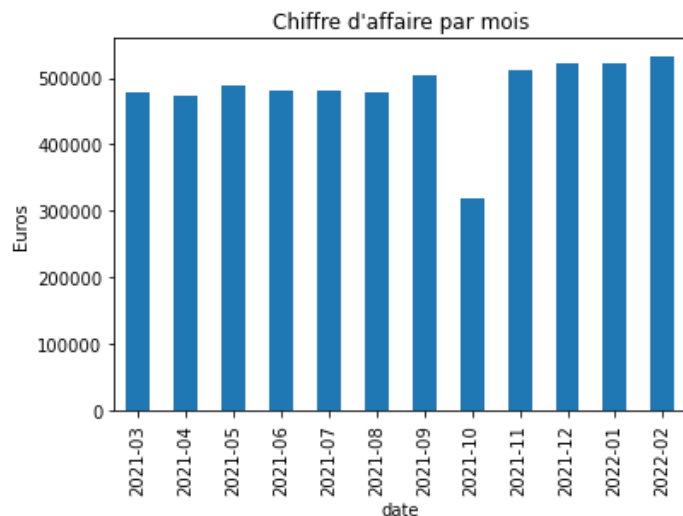
Nettoyage

Analyse
univariéeAnalyse
bivariéeAnalyse
statistique

Conclusion

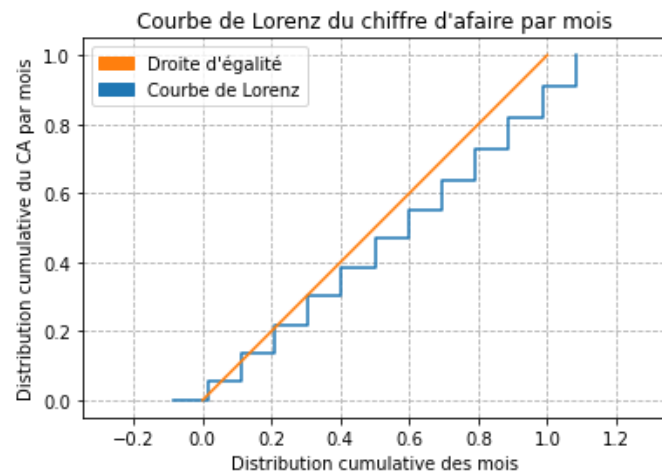
Analyse des dates et des prix :

➔ Chiffre d'affaire entre mars 2021 et février 2022



Moyenne	Médiane	Coefficient de Gini
483 139	485 392	0,048

5.8 millions euros



Le coefficient de Gini est proche de 0 donc
il y a une égalité entre les chiffres
d'affaires mensuels

Nettoyage

Analyse
univariée

Analyse
bivariée

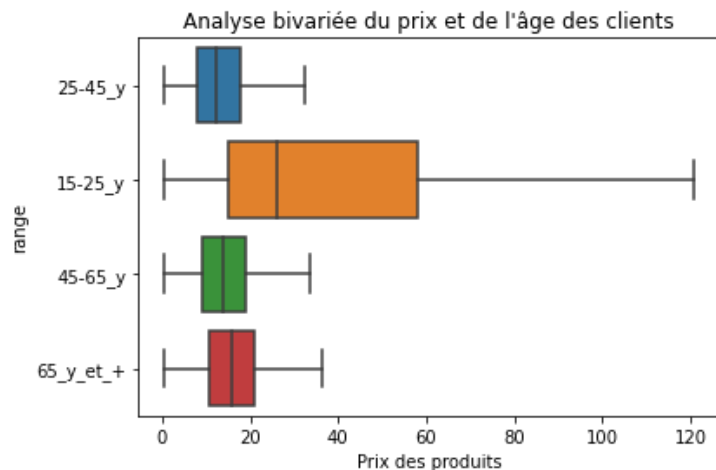
Analyse
statistique

Conclusion

Analyse des clients et des prix :

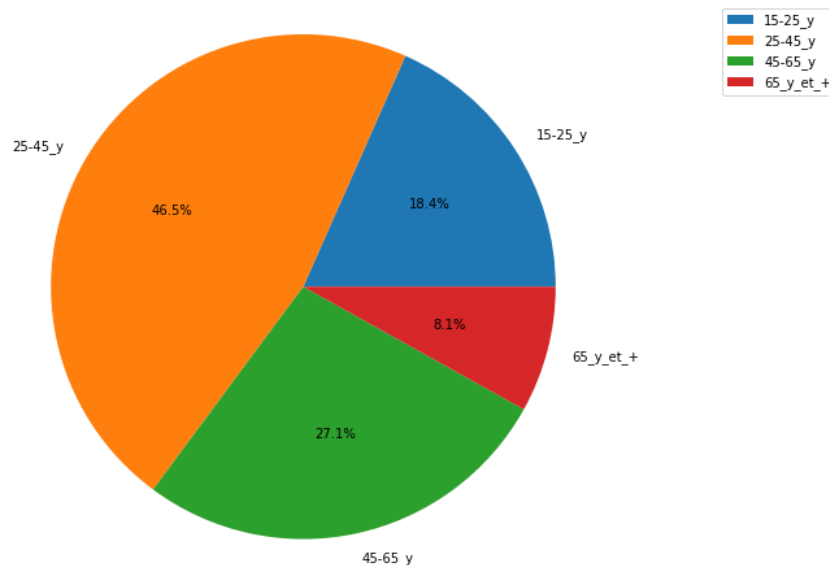


Âge des clients



La tranche d'âge 15-25 ans achète les produits les plus coûteux

Représentation du chiffre d'affaire par tranche d'âge



46 % du chiffre d'affaire est réalisé par des achats de clients dans la tranche d'âge 25-45 ans

Nettoyage

Analyse
univariée

Analyse
bivariée

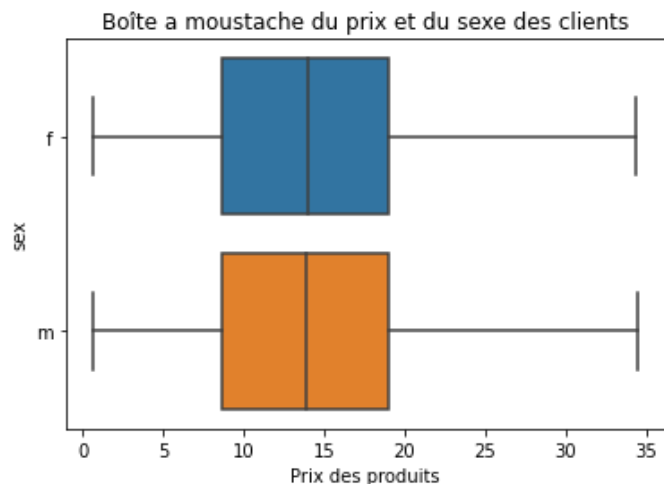
Analyse
statistique

Conclusion

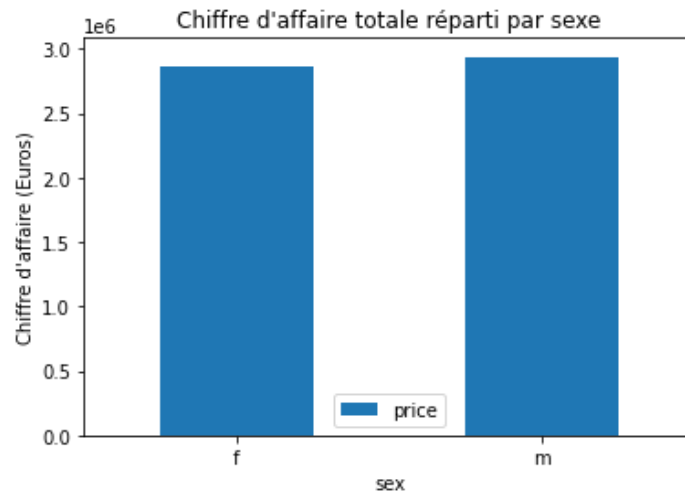
Analyse des clients et des prix :



Sexe des clients



Les prix d'achats des hommes et des femmes sont équivalents



Le chiffre d'affaire total pour les hommes et les femmes sont équivalents

Nettoyage

Analyse
univariée

Analyse
bivariée

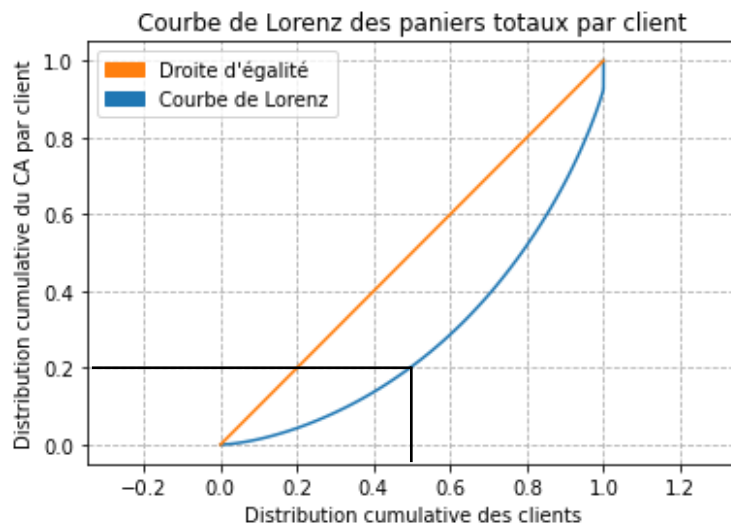
Analyse
statistique

Conclusion

Analyse des clients et des prix :



Chiffre d'affaire



Coefficient de Gini

0,44

Inégalité de distribution peu forte : chaque client ne contribue pas de la même manière au chiffre d'affaire

Nettoyage

Analyse
univariée

Analyse
bivariée

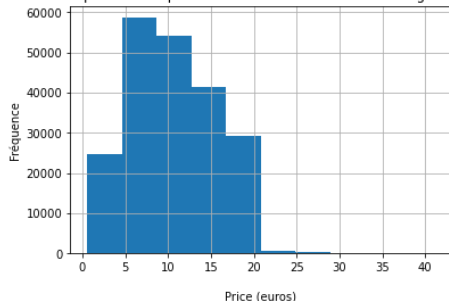
Analyse
statistique

Conclusion

Analyse des catégories produits et des prix :

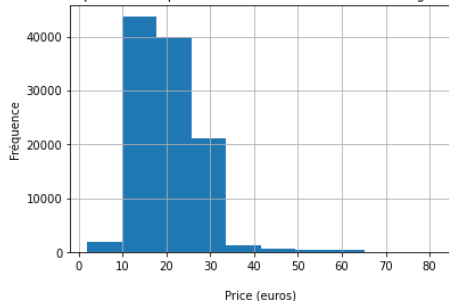
Catégorie 0

Dispersion des prix en fonction des différentes catégories



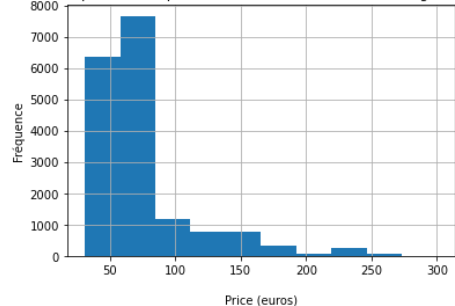
Catégorie 1

Dispersion des prix en fonction des différentes catégories

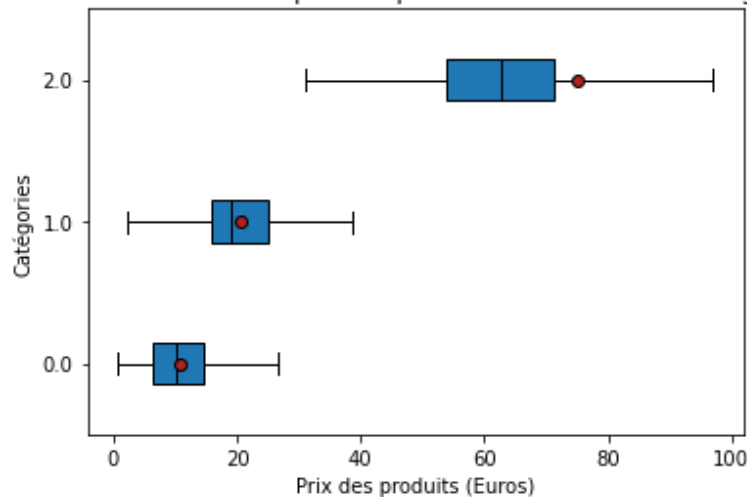


Catégorie 2

Dispersion des prix en fonction des différentes catégories



Boite à moustache des prix des produits en fonction des catégories



**Le prix des produits augmente en fonction
de la catégorie concerné
(catégorie 2 > catégorie 1 > catégorie 0)**

Nettoyage

Analyse
univariée

Analyse
bivariée

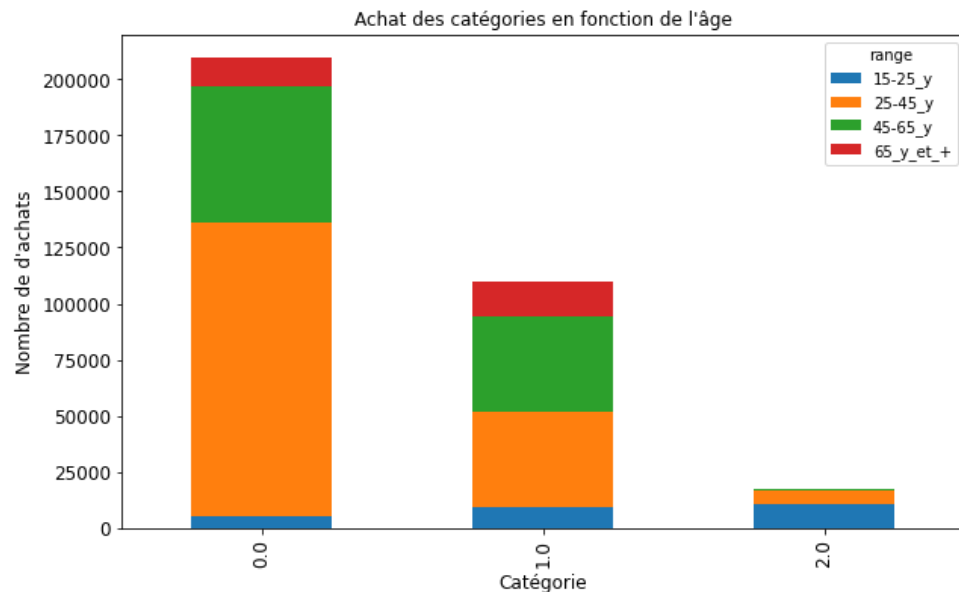
Analyse
statistique

Conclusion

Analyse des catégories produits et des clients:



Âge des clients



- La catégorie 0 est la plus achetée
- Categ 0 : 62% des achats 25-45 ans
- Categ 1 : 40% des achats 45-65 ans
- Categ 2 : 60% des achats 15-25 ans

Nettoyage

Analyse
univariée

Analyse
bivariée

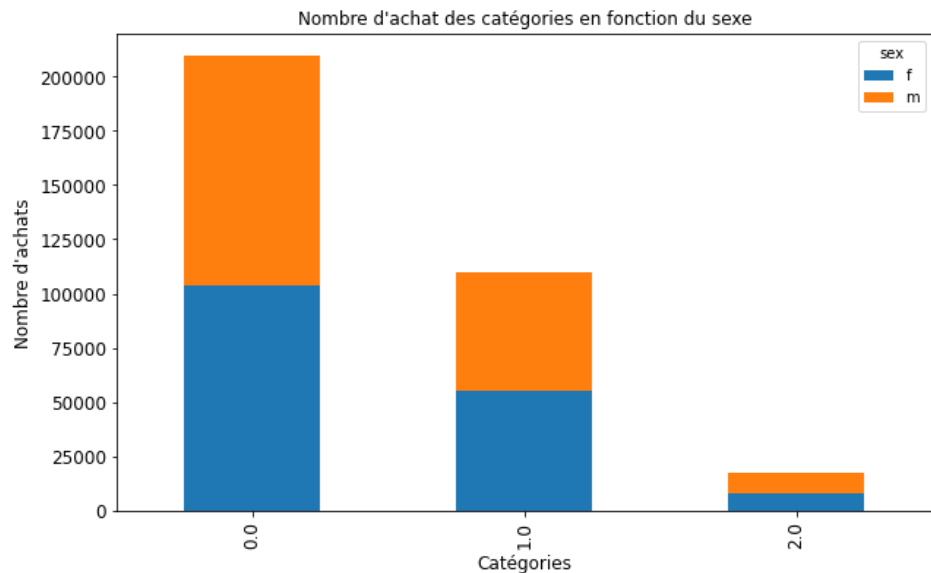
Analyse
statistique

Conclusion

Analyse des catégories produits et des clients:



Sexe des clients



Le nombre d'achat des catégories est équivalent entre les hommes et les femmes

Nettoyage

Analyse
univariée

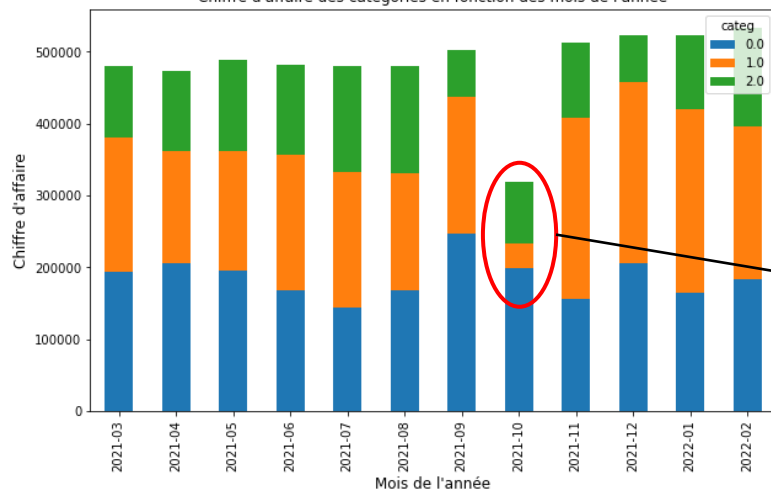
Analyse
bivariée

Analyse
statistique

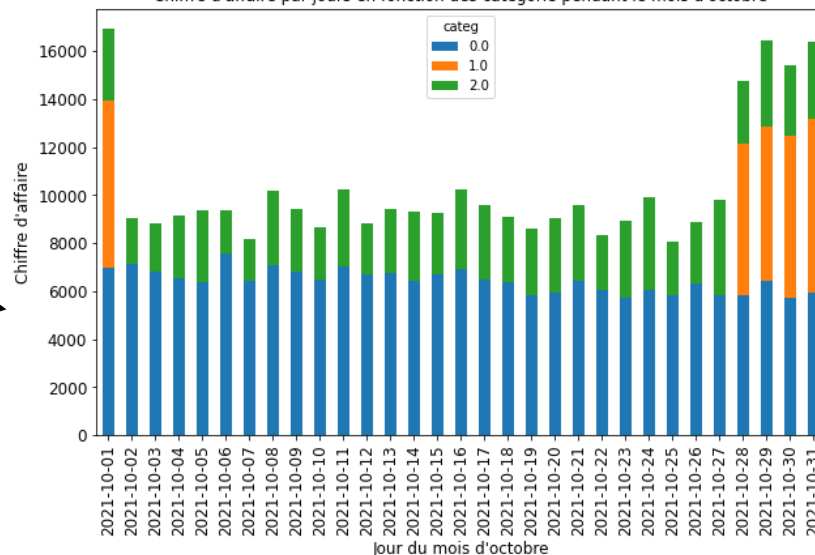
Conclusion

Analyse des catégories produits et du chiffre d'affaire:

Chiffre d'affaire des catégories en fonction des mois de l'année



Chiffre d'affaire par jours en fonction des catégorie pendant le mois d'octobre



Diminution CA catégorie 1 pour le mois
d'octobre

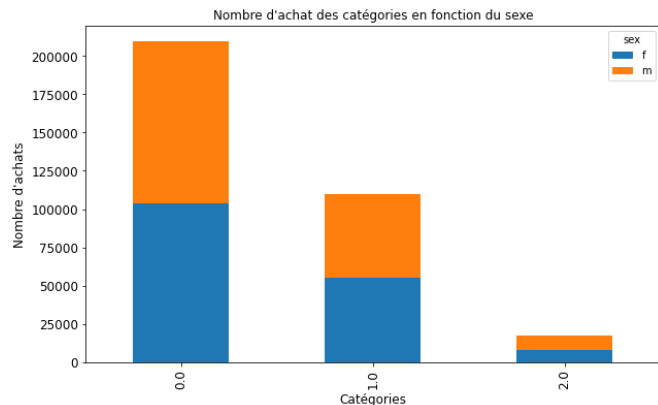
Plus de ventes de la catégorie 1 du 2 au 27
octobre

Nettoyage

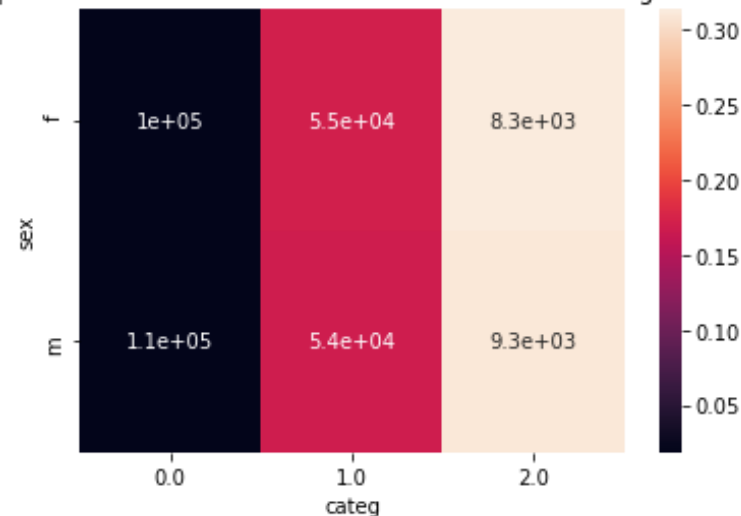
Analyse
univariéeAnalyse
bivariéeAnalyse
statistique

Conclusion

Analyse de corrélation des catégories produits et sexe des clients :



Heat map représentant le sexe des clients en fonction de la catégories de produits



Observé

Catégories	0	1	2
F	103846	55469	8260
M	105683	54266	9292

Attendu

Catégories	0	1	2
F	104246	54596	8732
M	105282	55138	8819

Test d'indépendance de Chi-2 : P-value : 1.78 e-18
Le sexe et la catégorie sont deux paramètres dépendants

Nettoyage

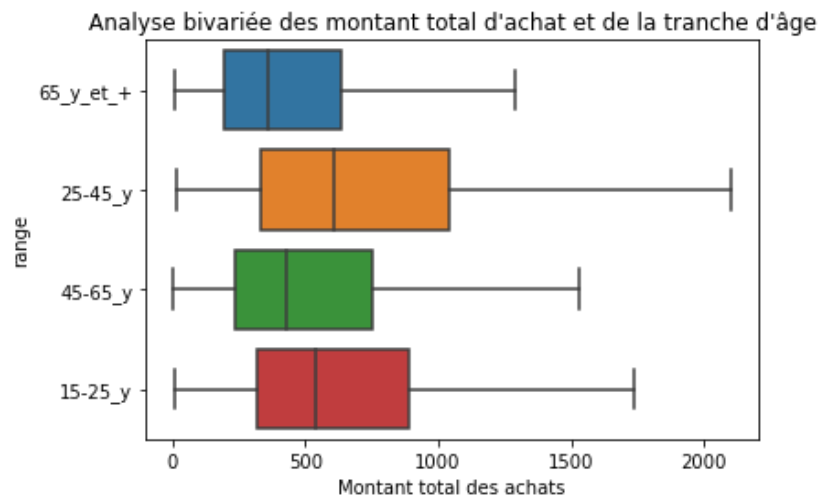
Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Analyse de corrélation entre l'âge clients et le montant total des achats :



Test d'ANOVA Welch :
P-value : 1.77 e-15
Dépendance entre l'âge
des clients et le
montant total des
achats

Nettoyage

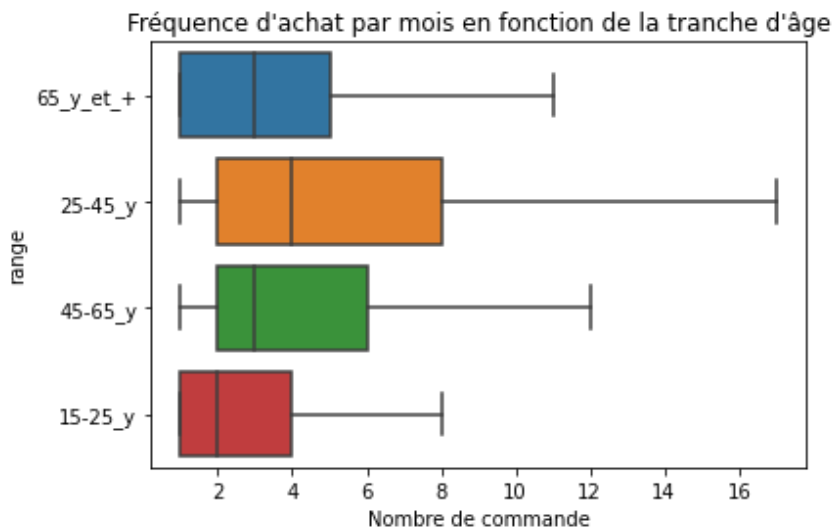
Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Analyse de corrélation entre l'âge clients et la fréquence d'achat :



Test d'ANOVA Welch :
P-value : 1.16e-156
Dépendance entre l'âge
des clients et la
fréquence d'achat

Nettoyage

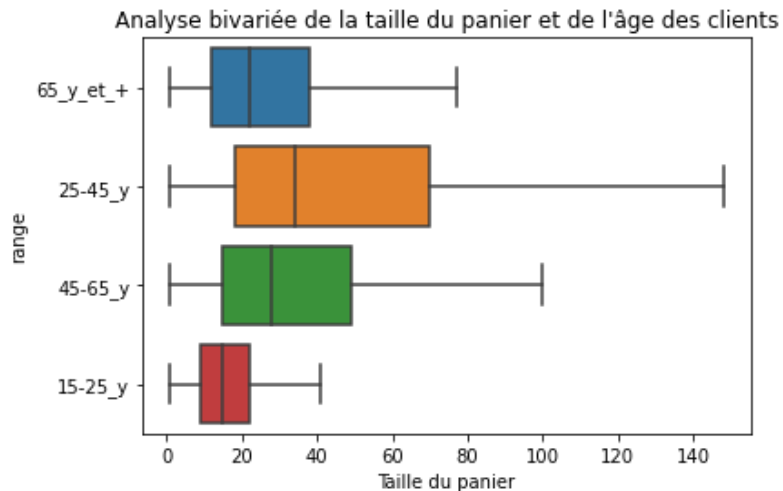
Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

Analyse de corrélation entre l'âge clients et la taille du panier moyen :



Test d'ANOVA Welch :
P-value : 1.10e-21
Dépendance entre
l'âge des clients et la
taille du panier moyen

Nettoyage

Analyse
univariéeAnalyse
bivariéeAnalyse
statistique

Conclusion

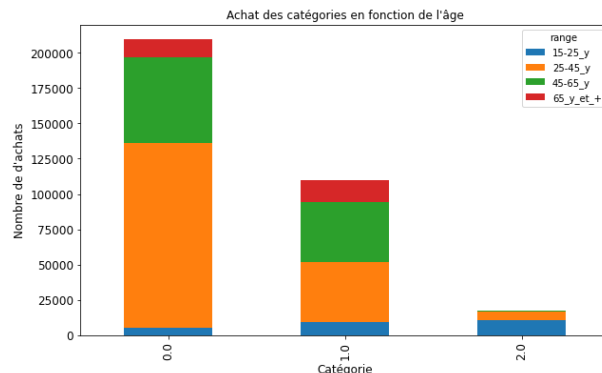
Analyse de corrélation entre l'âge clients et la catégorie de produits :

Observé

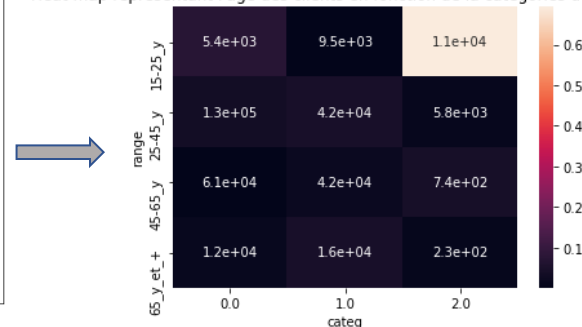
Catégories	0	1	2
15-25_y	5380	9491	10751
25-45_y	130758	42391	5832
45-65_y	60909	42295	739
65_y_et_+	12482	15558	230

Attendu

Catégories	0	1	2
15-25_y	15939	8347	1335
25-45_y	111341	58312	9326
45-65_y	64661	33864	5416
65_y_et_+	17586	9210	1473



Heat map représentant l'âge des clients en fonction de la catégories de produits



Test d'indépendance de Chi-2 : P-value : 0
L'âge et la catégorie sont deux paramètres dépendants

Nettoyage

Analyse
univariée

Analyse
bivariée

Analyse
statistique

Conclusion

- ❑ La plupart des clients sont situés dans la tranche d'âge **25 – 45 ans**
- ❑ La catégorie de produit **0** est celle la plus achetée (~10 euros)
- ❑ Les clients **25 – 45 ans** possèdent une fréquence d'achat la plus **élevée** (46% CA)
- ❑ **Diminution du CA** d'octobre : entre le 2 et le 27 disparition des ventes de la **catégorie 1**
- ❑ **Cibler** les caractéristiques des produits les **plus vendus** (type de livre, nombre de pages..)
- ❑ **Cibler** les clients âgés entre **25 et 45 ans**

Merci pour votre attention