

A thick orange vertical bar on the left side of the slide, which tapers into a triangular shape at the bottom.

ANALYSE DE VENTES

RESTER LIVRES



RESTER LIVRE, CETTE LIBRAIRIE CRÉÉE EN 2001, A COMME POINT DE
VENTE DES BOUTIQUES ET SON SITE WEB.
SES PRINCIPAUX CLIENTS SONT DES PARTICULIERS DE TOUS ÂGES.



LES DONNÉES

	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

	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

	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

FICHER
TRANSACTIONS

FICHER
CUSTOMERS

FICHER
PRODUCTS

MISSION 1

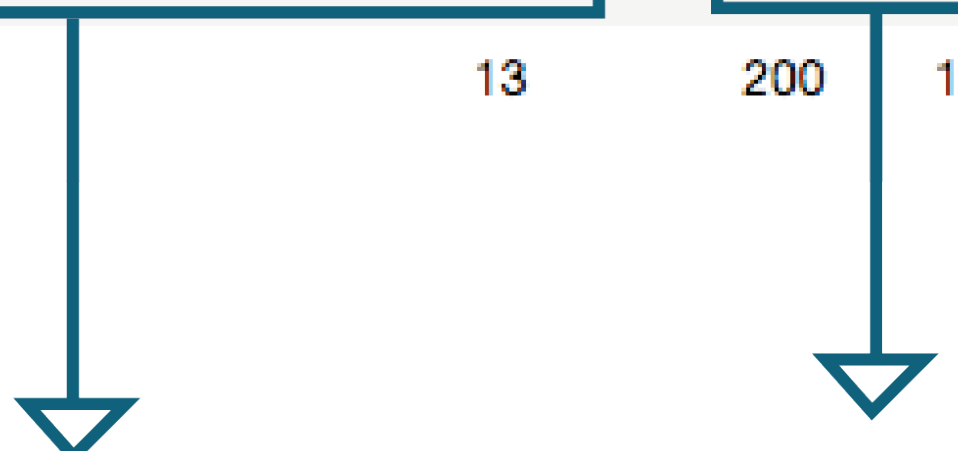


TRAITEMENT
DES
DONNÉES



ON A SUPPRIMÉ LES VALEURS ABERRANTES

	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



	id_prod	date	session_id	client_id
57755	T_0	test_2021-03-01 02:30:02.237413	s_0	ct_1
59043	T_0	test_2021-03-01 02:30:02.237413	s_0	ct_0
95537	T_0	test_2021-03-01 02:30:02.237413	s_0	ct_0
100544	T_0	test_2021-03-01 02:30:02.237413	s_0	ct_0

ÉLIMINATIONS DE CLIENT TEST

```
ind_drop = df_transactions[df_transactions['id_prod'].apply(lambda x: x.startswith('T_0'))].index
```

```
# on supprime le produit avec l'id_prod 'T_0'
df_transactions = df_transactions.drop(ind_drop)
```

VÉRIFICATION DES DONNÉES ABERRANTES (PRIX)

EN PARALLÈLE, ON A VÉRIFIÉ QU'IL N'Y
AVAIT PAS DE VALEURS MANQUANTES
VIA LA FONCTION `.ISNULL()`

price	
count	3287.000000
mean	21.856641
std	29.847908
min	-1.000000



price	
count	330312.000000
mean	15.716898
std	11.351166
min	0.620000

ON SUPPRIME LES VALEURS NÉGATIVES

PRODUIT QUI APPARAÎT DANS LA TABLE TRANSACTIONS MAIS
NE PAS DANS LA TABLE PRODUCTS

IMPUTATIONS PAR LA MOYENNE DU PRODUITS o_2245

id_prod	price	categ
---------	-------	-------



	id_prod	price	categ	date	session_id	client_id
336713	0_2245	11.0	0.0	2021-06-17 03:03:12.668129	s_49705	c_1533
336714	0_2245	11.0	0.0	2021-06-16 05:53:01.627491	s_49323	c_7954
336715	0_2245	11.0	0.0	2021-11-24 17:35:59.911427	s_124474	c_5120
336716	0_2245	11.0	0.0	2022-02-28 18:08:49.875709	s_172304	c_4964

CRÉATION DE VARIABLE ÂGE

```
# on creee la colonne age  
df['age'] =(2022 - df['birth'])
```

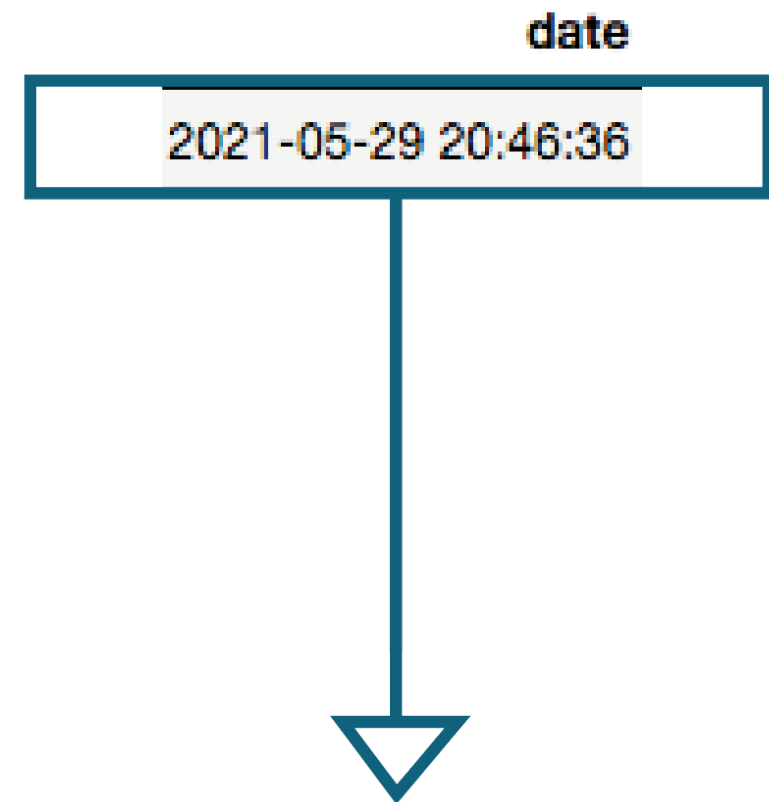
age

count	330312.000000
mean	44.178340
std	13.034259
min	18.000000
25%	36.000000
50%	42.000000
75%	51.000000
max	84.000000

VÉRIFICATION DES
DONNÉES
ABERRANTES

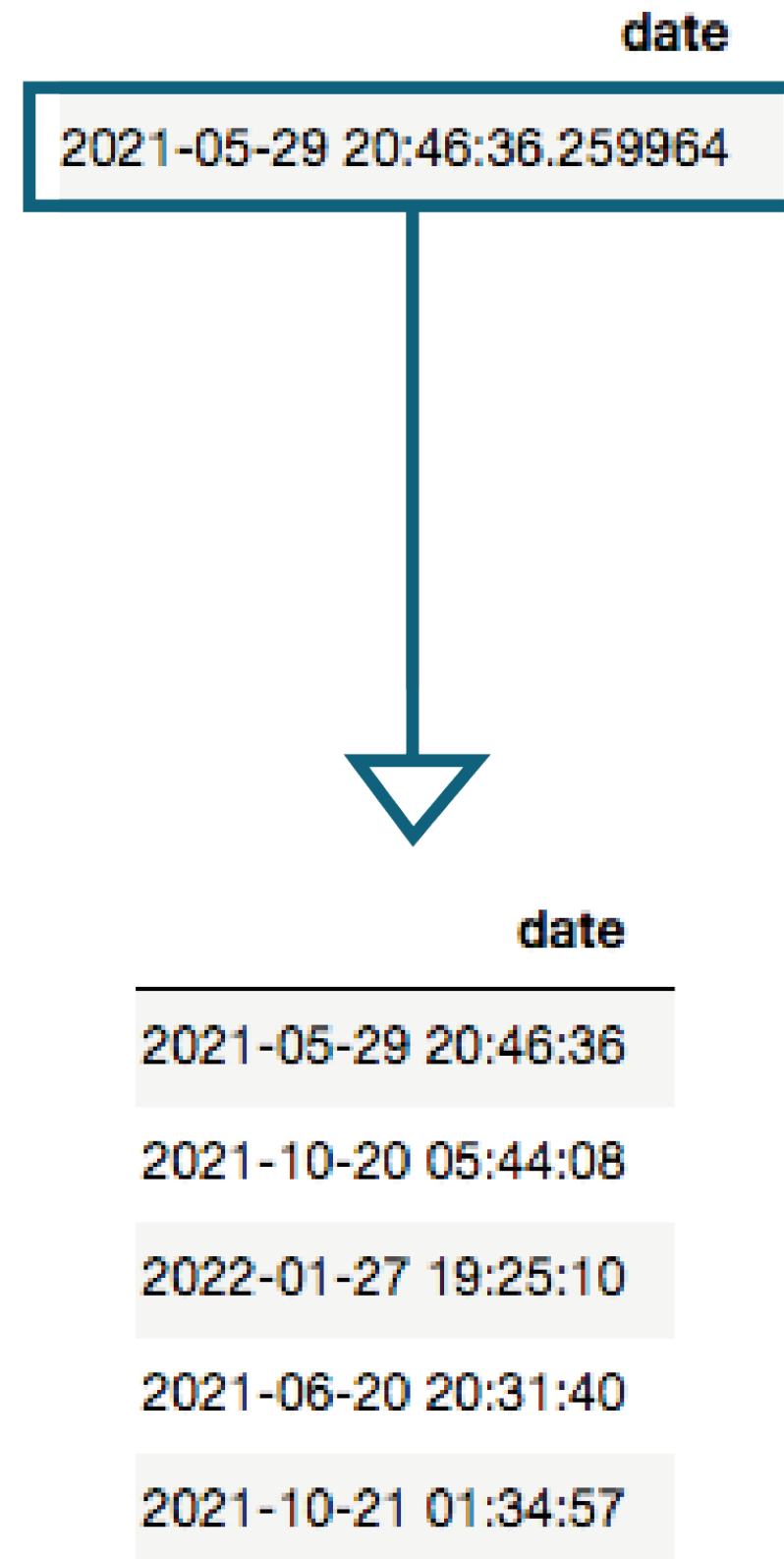


ON A COUPÉ LA
COLONNE 'DATE' EN
'DATE' ET 'HEURE'



heure	date
20:46:36	2021-05-29
05:44:08	2021-10-20
19:25:10	2022-01-27
20:31:40	2021-06-20
01:34:57	2021-10-21

NETTOYAGE DE LA NOMENCLATURE



ON A SUPPRIMÉ LES MICROSECONDES

NETTOYAGE DE LA NOMENCLATURE



client_id

c_3757



client_id s

3757

5606

7453

7075

2276

ON A ENLEVÉ LE
C_DE LA COLONNE
CLIENT_ID

DATAFRAME FINAL

	id_prod	price	categ	date	session_id	client_id	sex	birth	age	heure
169775	0_202	0.62	0	2021-03-05	s_2176	5277	m	1976	46	17:06:24
9757	0_528	0.62	0	2021-05-28	s_40833	1609	m	1980	42	18:10:03
68901	0_528	0.62	0	2021-06-24	s_53240	5636	f	1989	33	18:55:00
239135	0_202	0.62	0	2022-01-06	s_145956	2069	m	1986	36	18:57:12
80717	0_528	0.62	0	2021-10-14	s_104606	4951	m	1984	38	11:22:30

ON A FUSIONNÉ LES TROIS DF

VARIABLE ANNEXE

	session_id	panier
0	s_41352	0.62
1	s_107417	0.62
2	s_156085	40.16
3	s_51439	30.61
4	s_107804	15.55

DF PANIER

MISSION 2



ANALYSE DES DONNÉES

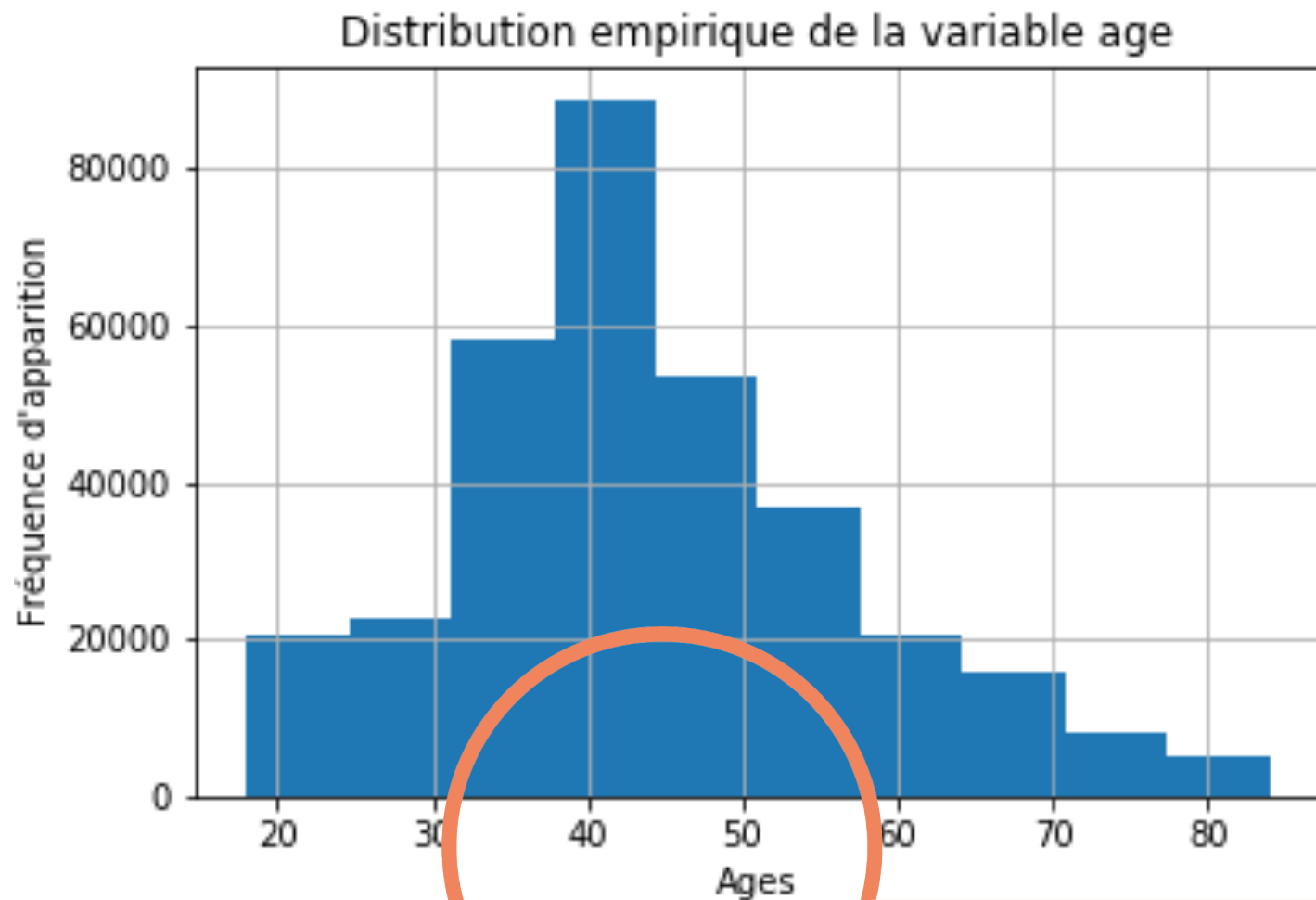


L'ÂGE

LES INDICATEURS DE TENDANCE CENTRALE POUR L'ÂGE DES
CLIENTS : MODE, MOYENNE, MÉDIANE

TOP TEN ÂGE CLIENTS

age
42.0
43.0
34.0
44.0
36.0
39.0
38.0
40.0
45.0
35.0



L'ÂGE MOYENNE DE CLIENTS EST 44 ANS

L'ÂGE MÉDIANE DE CLIENTS EST 42 ANS

L'ÂGE PLUS FRÉQUENT DE CLIENTS EST 42 ANS

LE CLIENT PLUS ÂGÉ A 84 ANS
ET LE PLUS JEUNE 18 ANS

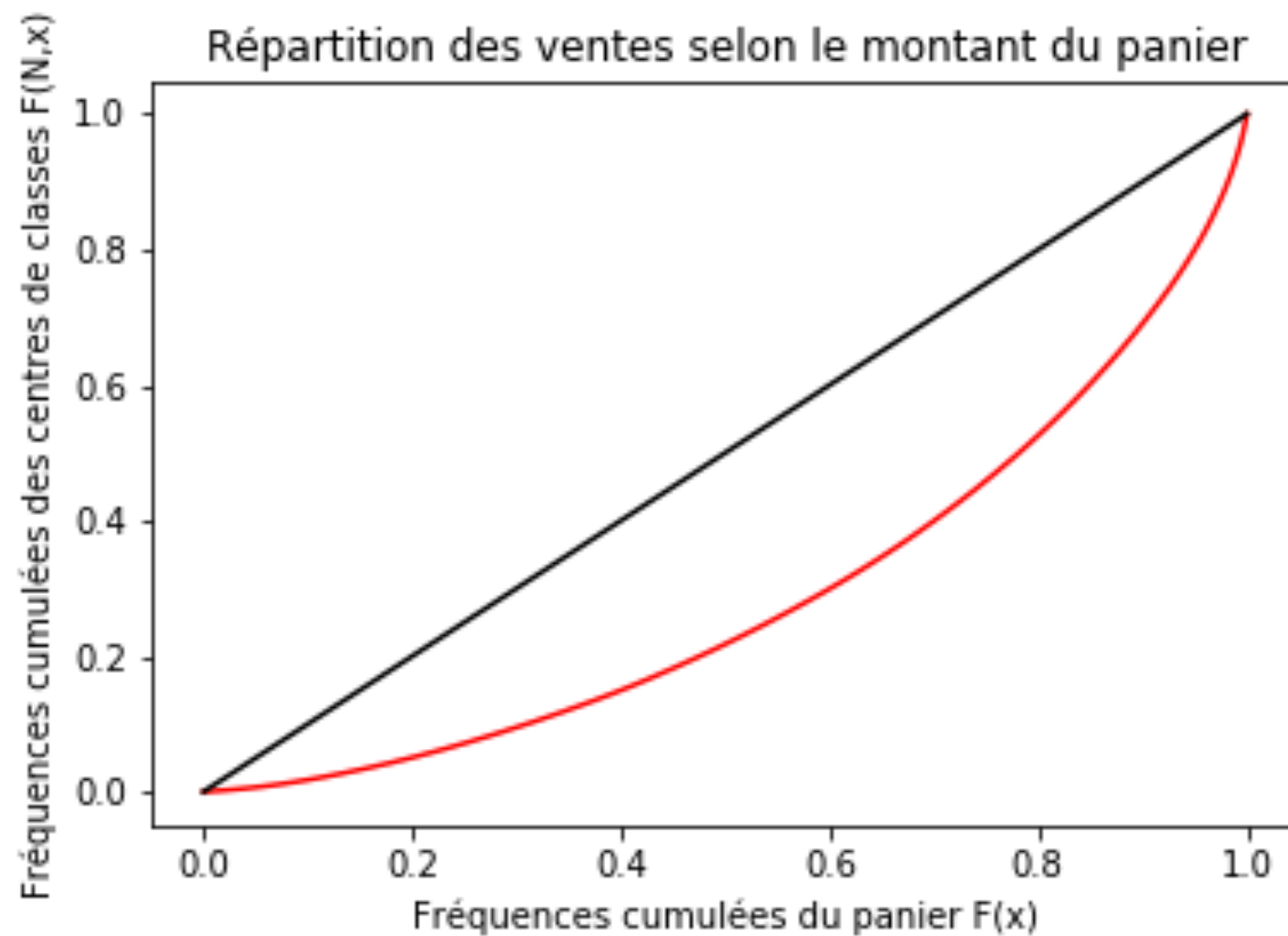
- 42 ANS

+ 42 ANS



LE PANIER

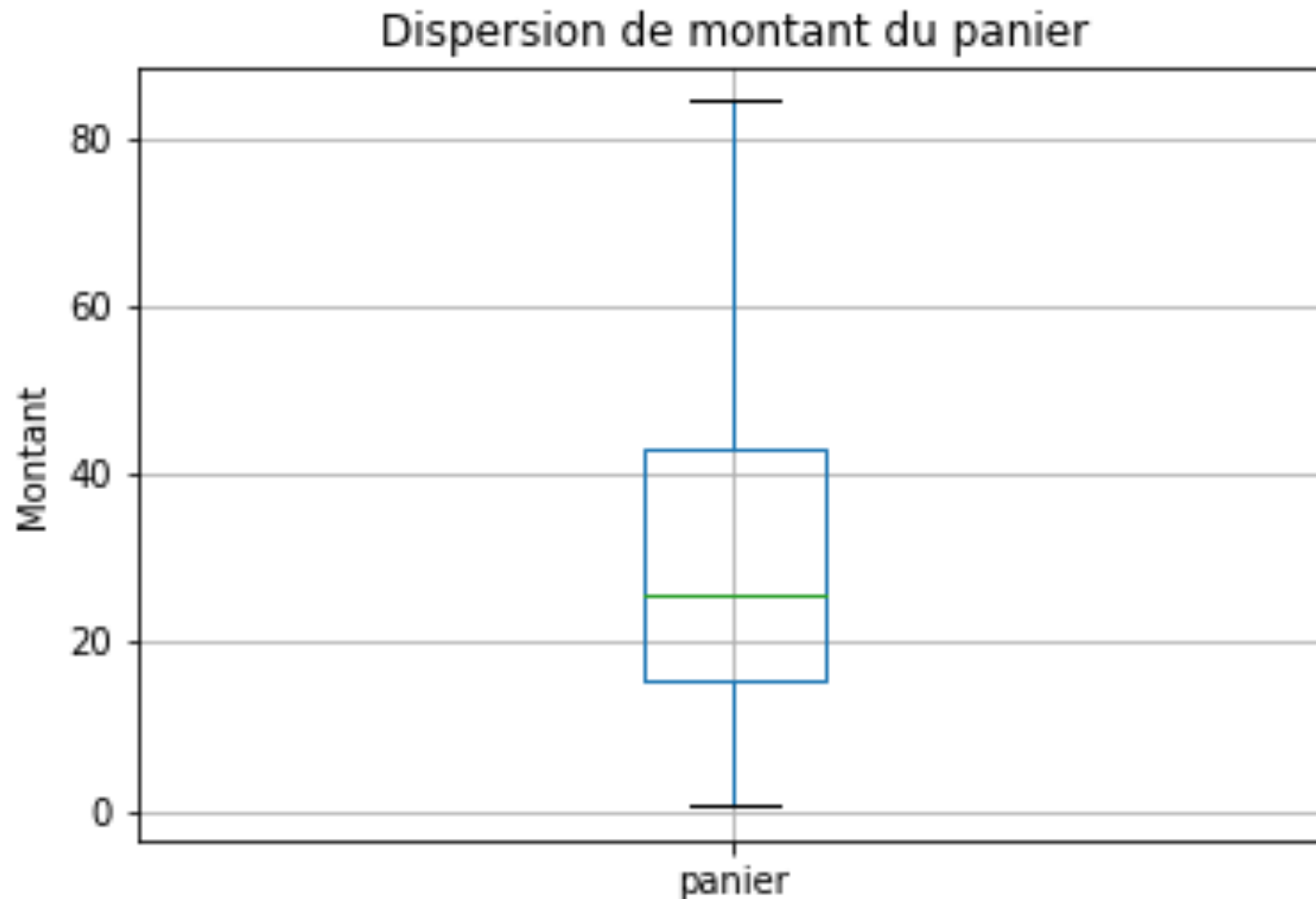
UNE ANALYSE DE CONCENTRATION, VIA UNE COURBE DE LORENZ ET UN INDICE DE GINI



L'INDICE DE GINI
EST ÉGAL À 0.42

CONCENTRATION RELATIVEMENT MODÉRÉE DE
MONTANT DU PANIER

INDICATEURS DE DISPERSION



MAXIMUM DE DÉPENSÉ 539€

75% DES CLIENTS DÉPENSÉ MOINS DE 43€

50% DES CLIENTS DÉPENSÉ MOINS DE 25€

25% DES CLIENTS DÉPENSÉ MOINS DE 15€

MINIMUM DE DÉPENSÉ 0,62€

LA VARIANCE DES DÉPENSES EST ÉGAL À 1020 EUROS

LA MOYENNE DES DÉPENSES EST ÉGAL À 35 EUROS

L'ÉCART TYPE DES DÉPENSES EST ÉGAL À 32 EUROS

L'ÉCART INTER-QUARTILE DES DÉPENSES EST ÉGAL À 28 EUROS

- 25 EUROS

+ 25 EUROS

- 43 EUROS

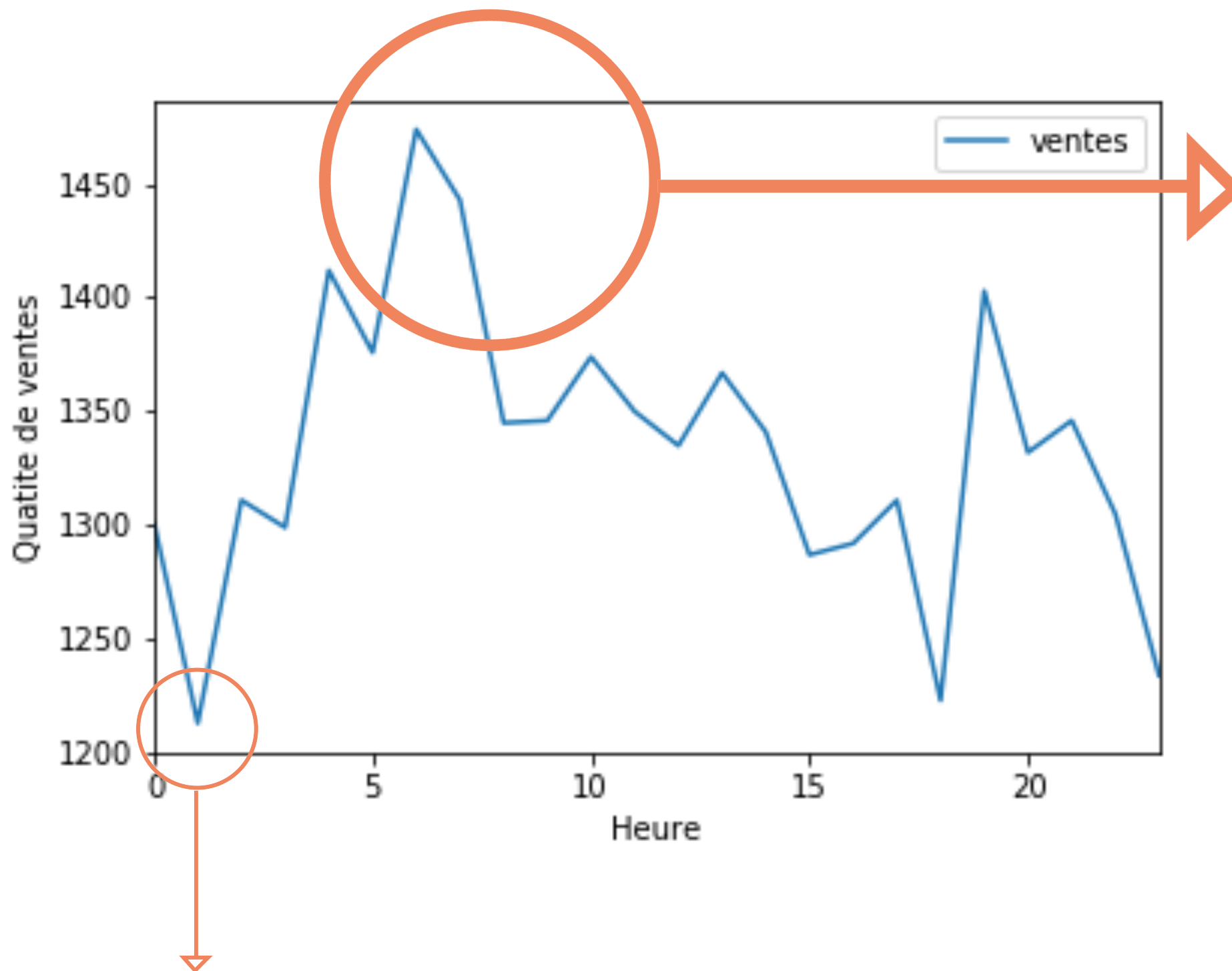
+ 43 EUROS



CALCULE DES VENTES EN FONCTION DE L'HEURE

DES ANALYSES BIVARIÉES

VENTES PAR HEURE MOIS DE DÉCEMBRE



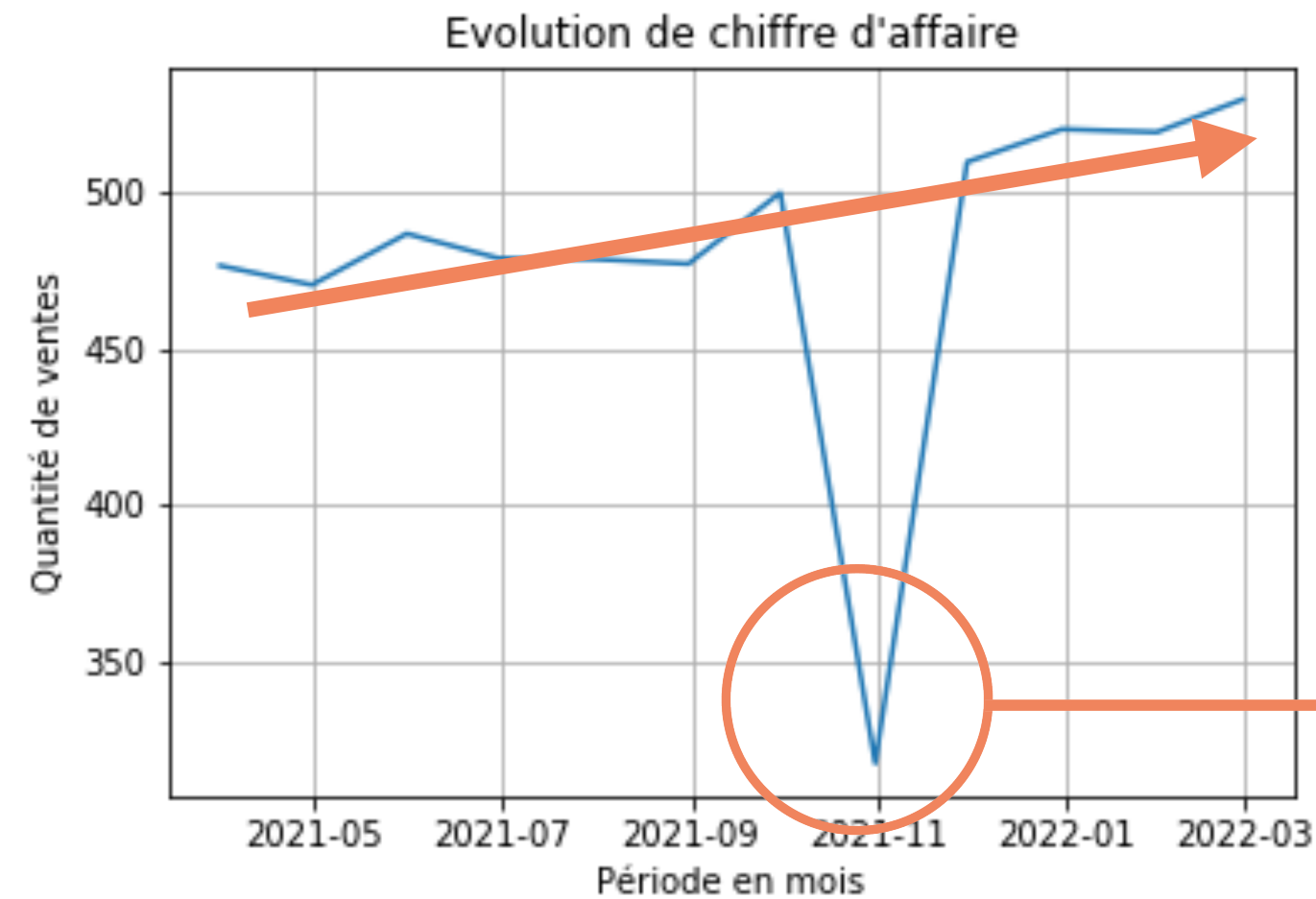
HEURE DE VENTE
PLUS ÉLEVÉE.

HEURE DE VENTE PLUS FAIBLE.



ANALYSE DE LA STRUCTURE DES VENTES

LE CHIFFRE D'AFFAIRES ET RÉPARTITION DES VENTES SELON LES
CATÉGORIES DE PRODUIT : REPRÉSENTATION D'UNE SÉRIE TEMPORELLE



LA TENDANCE GÉNÉRALE EST
À LA HAUSSE

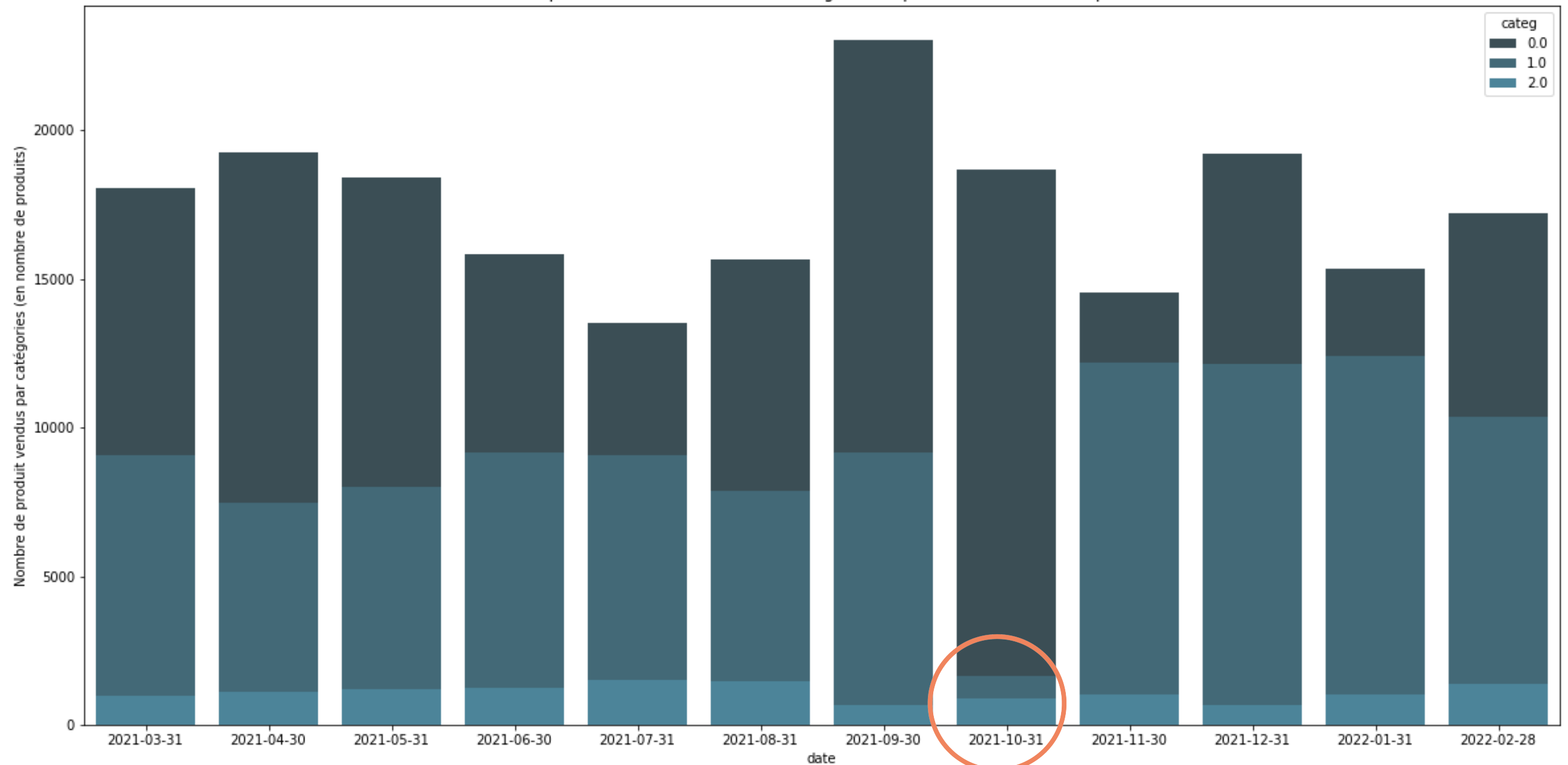
VARIATION SAISONNIÈRE OU
OUBLIE D'INDIVIDUS ?

Nombre de ventes par jour des produits categorie 1



ABSENCE DE VENTE
DE PRODUITS DE
LA CATÉGORIE 1

Répartition des ventes selon les catégories de produit au cours du temps



INFORMATION DE VENTES DES PRODUITS DE CATÉGORIE 1

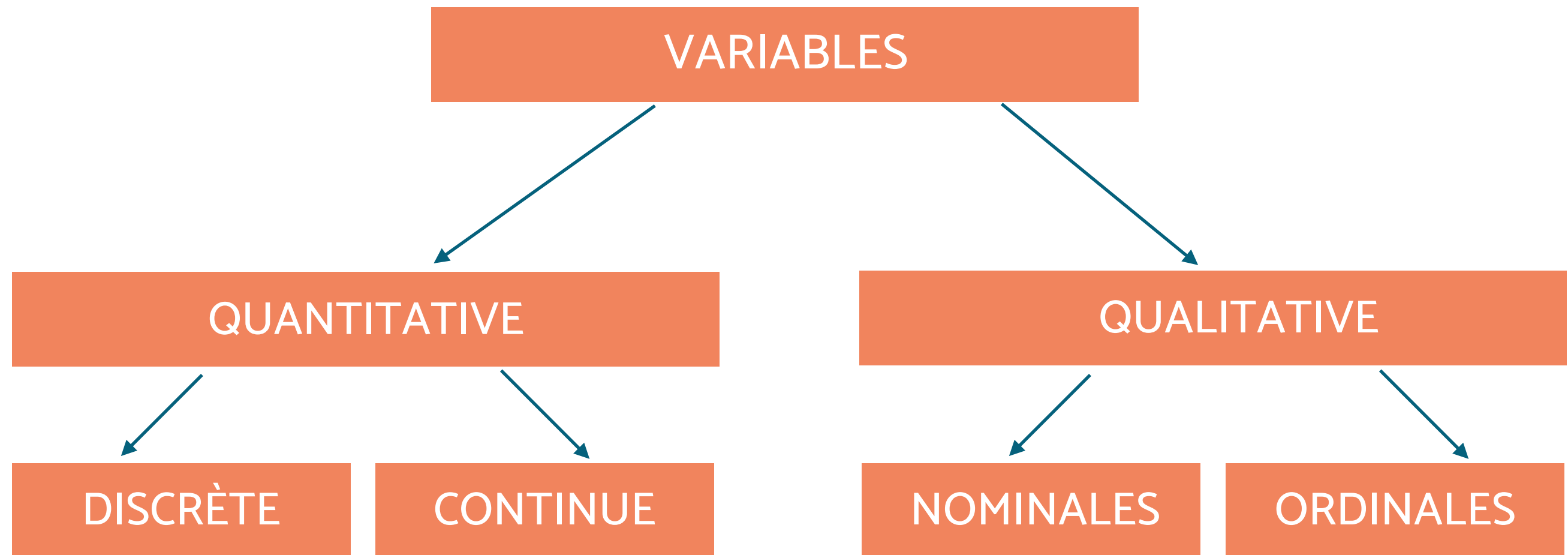
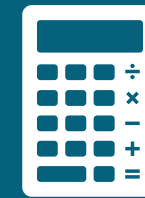
date	catég	id_prod	price	session_id	client_id	sex
2021-10-31	1.0	1645	1645	1645	1645	1645

LES PRIX DE PRODUITS CATÉGORIE 0 VONT DE 0€ À 41€
 LES PRIX DE PRODUITS CATÉGORIE 1 VONT DE 2€ À 81€
 LES PRIX DE PRODUITS CATÉGORIE 2 VONT DE 31€ À 300€

MISSION 3

ANALYSE DES CORRÉLATIONS

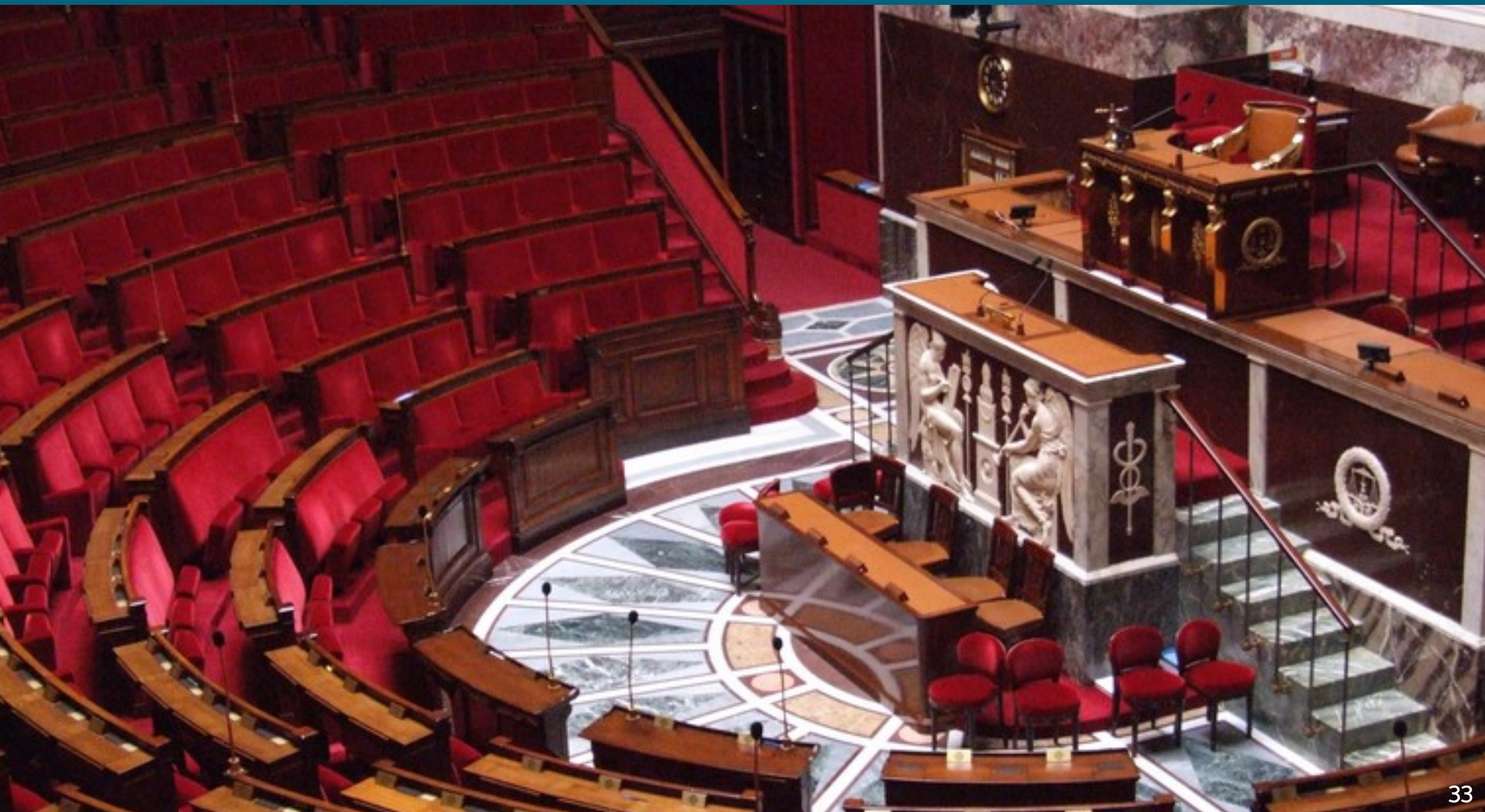




LA DISTINCTION ENTRE LES TYPES DE VARIABLES DÉCIDE DES ANALYSES STATISTIQUES UTILISABLES, DES REPRÉSENTATIONS GRAPHIQUES UTILISABLES OU ENCORE DES TESTS STATISTIQUES

LES TEST STATISTIQUES

TROIS SCÉNARIOS POSSIBLES



DEUX VARIABLES QUANTITATIVES

COEFFICIENT DE CORRÉLATION LINÉAIRE

LE COEFFICIENT DE CORRÉLATION DE DEUX VARIABLES X ET Y CORRESPOND À UNE NORMALISATION DE LEUR COVARIANCE PAR LE PRODUIT DES ÉCART-TYPES DE X ET Y . IL MESURE LE DEGRÉ DE DÉPENDANCE LINÉAIRE DE X ET Y .

$$r(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

LE RÉSULTAT DU COEFFICIENT EST UNE VALEUR ENTRE -1 ET 1.

SON INTERPRÉTATION EST LA SUIVANTE :

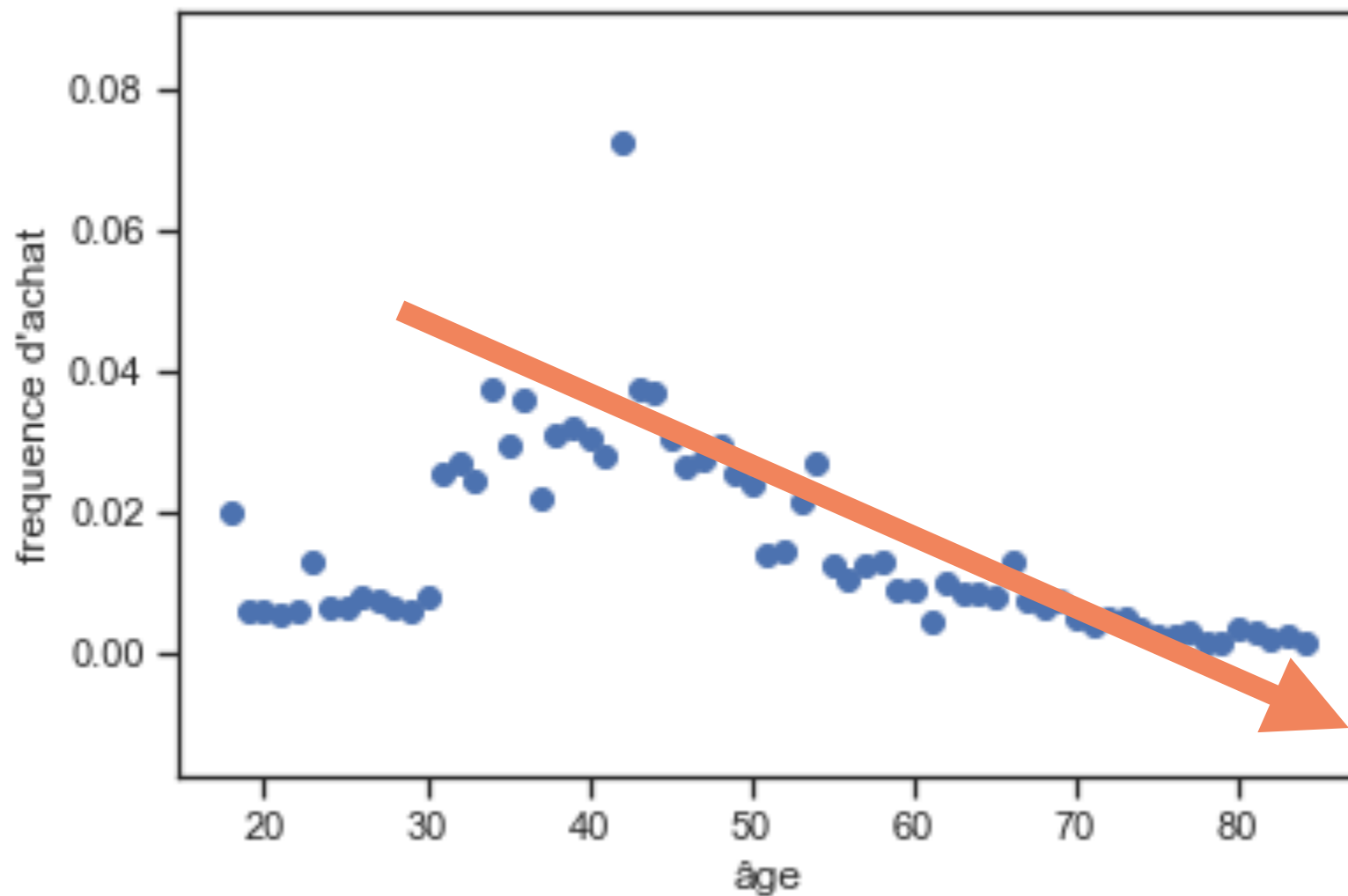
1 : IL EXISTE UNE FORTE RELATION LINÉAIRE POSITIVE ENTRE X ET Y.

0 : IL N'Y A PAS DE RELATION LINÉAIRE ENTRE X ET Y.

-1 : IL EXISTE UNE FORTE RELATION LINÉAIRE NÉGATIVE ENTRE X ET Y.



Y A-T-IL UNE CORRÉLATION ENTRE L'ÂGE DES CLIENTS
ET
LA FRÉQUENCE D'ACHAT (IE. NOMBRE D'ACHATS PAR
MOIS PAR EXEMPLE)?



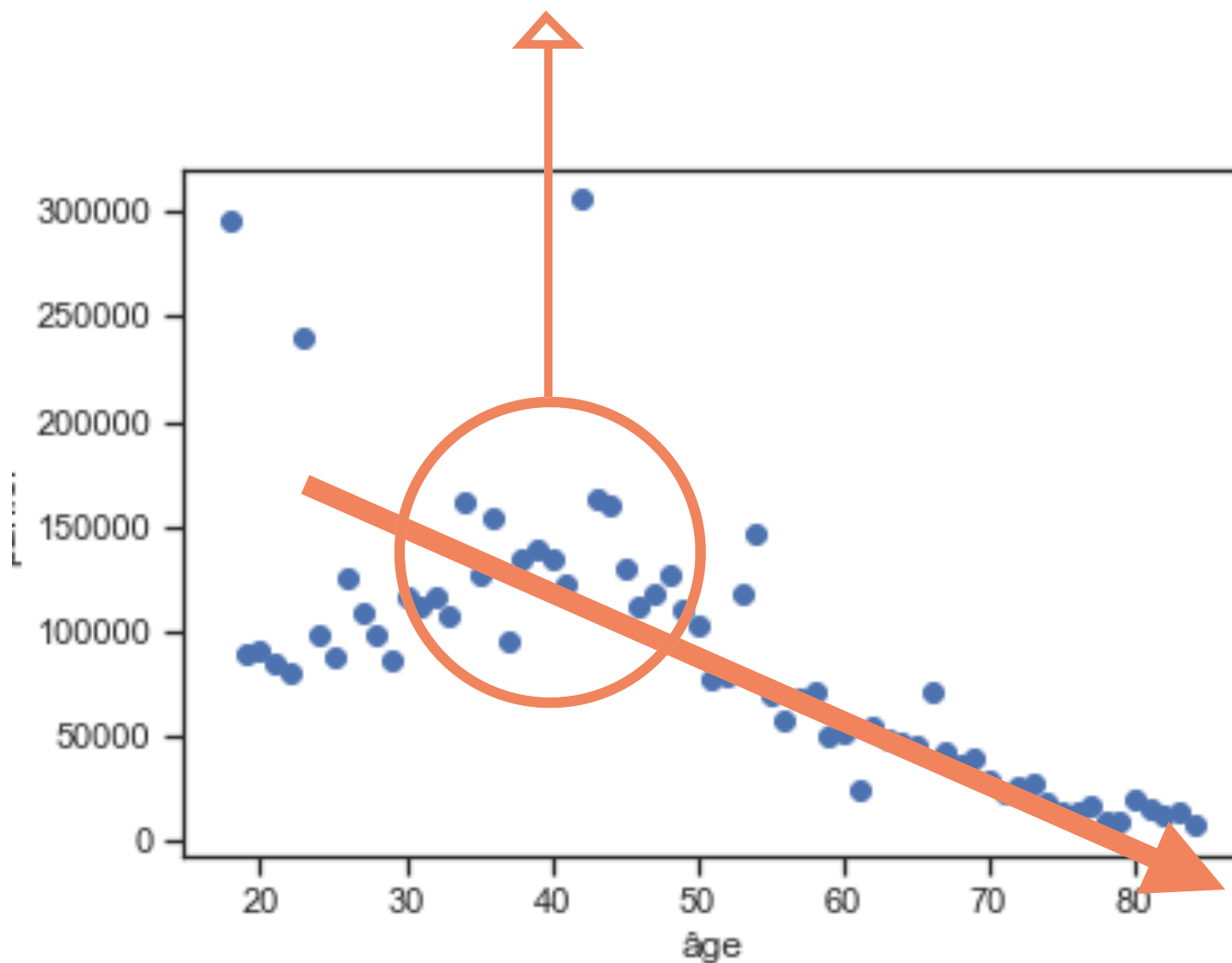
ON OBSERVE QUE PLUS LES CONSOMMATEURS SONT ÂGÉS PLUS LA FRÉQUENCE DE LEURS ACHATS EST FAIBLE AVEC UNE HAUSSE DE LA FRÉQUENCE DANS LA TRANCHE D'ÂGE DE 30 À 50 ANS

	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	67	-0.425	[-0.6, -0.21]	0.181	0.155	0.000337	82.091	0.955



Y A-T-IL UNE CORRÉLATION ENTRE L'ÂGE DES CLIENTS
ET
LE MONTANT TOTAL DES ACHATS?

LES CLIENTS ENTRE 30 ET 50 ANS
DÉPENSENT PLUS QUE LES AUTRES.



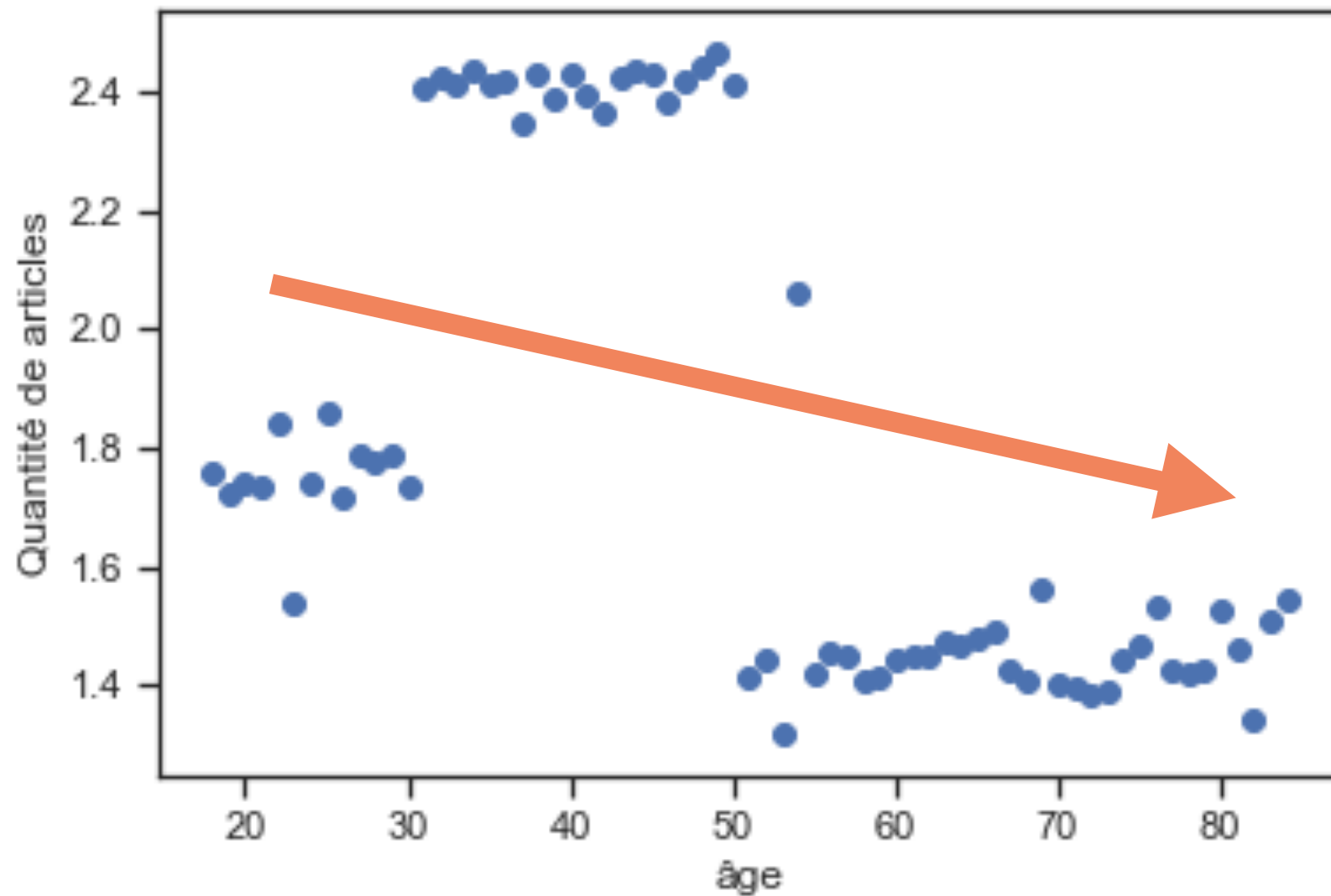
A FUR ET MESURE QUE
L'ÂGE DES CLIENTS
MONTENT LES
MONTANT DU PANIER
DESCEND.

	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	67	-0.719	[-0.82, -0.58]	0.517	0.502	7.000234e-12	1.472e+09	1.0



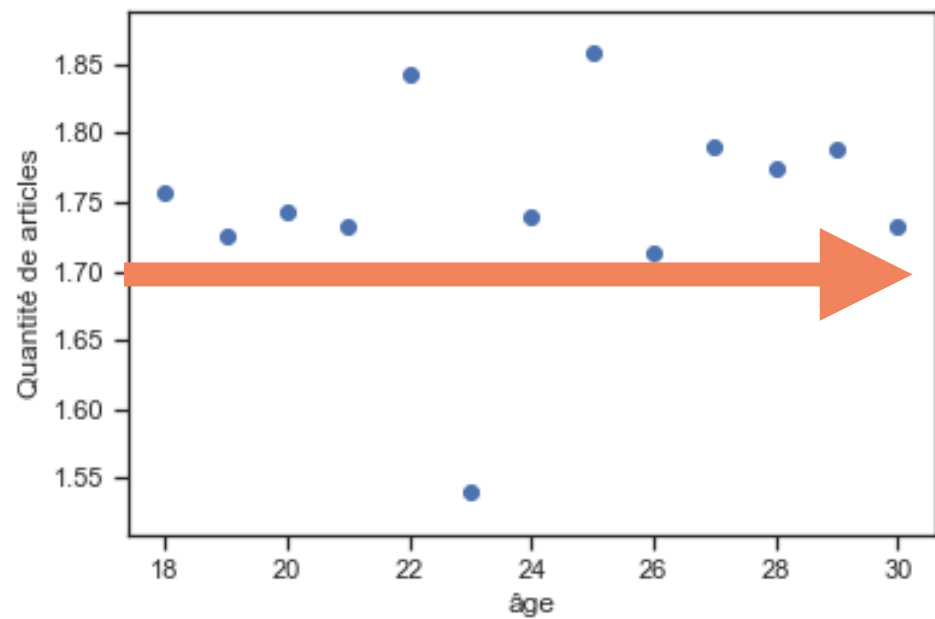
Y A-T-IL UNE CORRÉLATION ENTRE L'ÂGE DES CLIENTS
ET
LA TAILLE DU PANIER MOYEN (EN NOMBRE D'ARTICLES)?

TOP TEN ÂGE/ QUANTITÉ DE ARTICLES



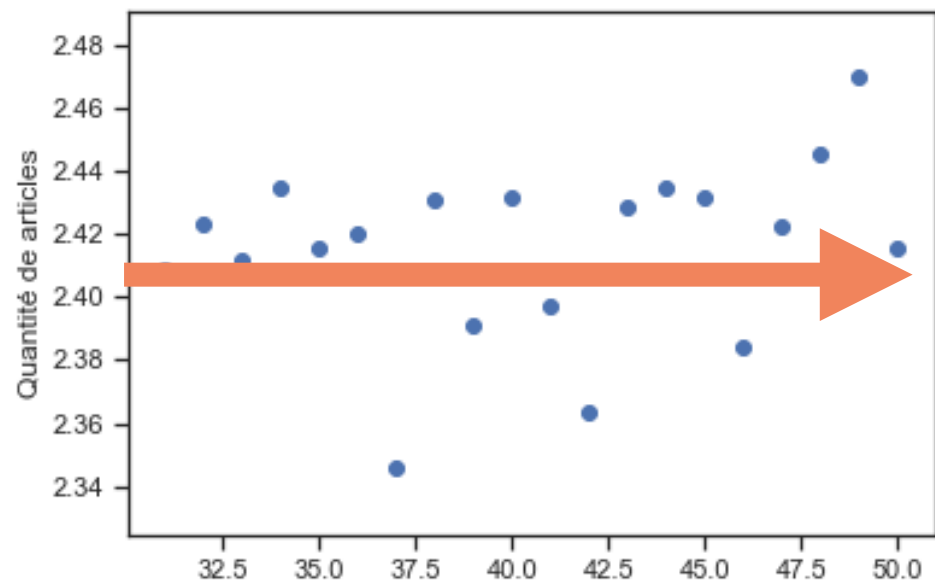
	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	67	-0.552	[-0.7, -0.36]	0.305	0.283	0.000001	1.466e+04	0.999

âge
49
48
44
34
40
45
38
43
32
47



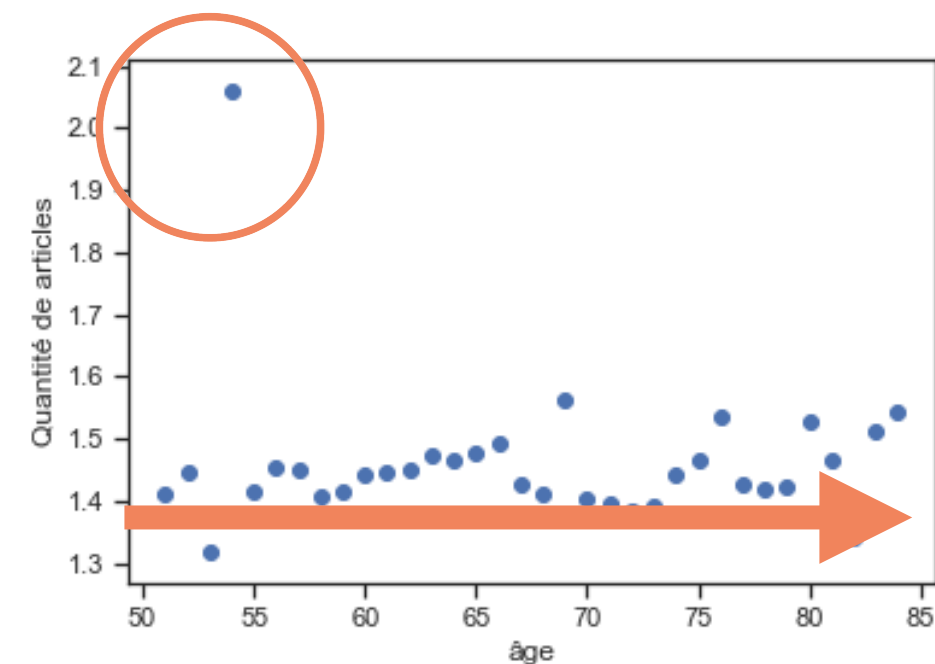
CLIENTS DE MOINS DE 30 ANS

	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	13	0.148	[-0.44, 0.65]	0.022	-0.174	0.629614	0.379	0.077



CLIENTS DE 30 ANS À 50 ANS

	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	20	0.241	[-0.23, 0.62]	0.058	-0.053	0.306742	0.451	0.178



CLIENTS DE PLUS DE 50 ANS

	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	34	-0.091	[-0.42, 0.26]	0.008	-0.056	0.609509	0.242	0.08

DEUX VARIABLES QUALITATIVE

LE TEST D'INDÉPENDANCE DU KHI-CARRÉ

LE TEST D'INDÉPENDANCE DU KHI-CARRÉ QUI PERMET DE CONTRÔLER
L'INDÉPENDANCE DE DEUX CARACTÈRES DANS UNE POPULATION
DONNÉE

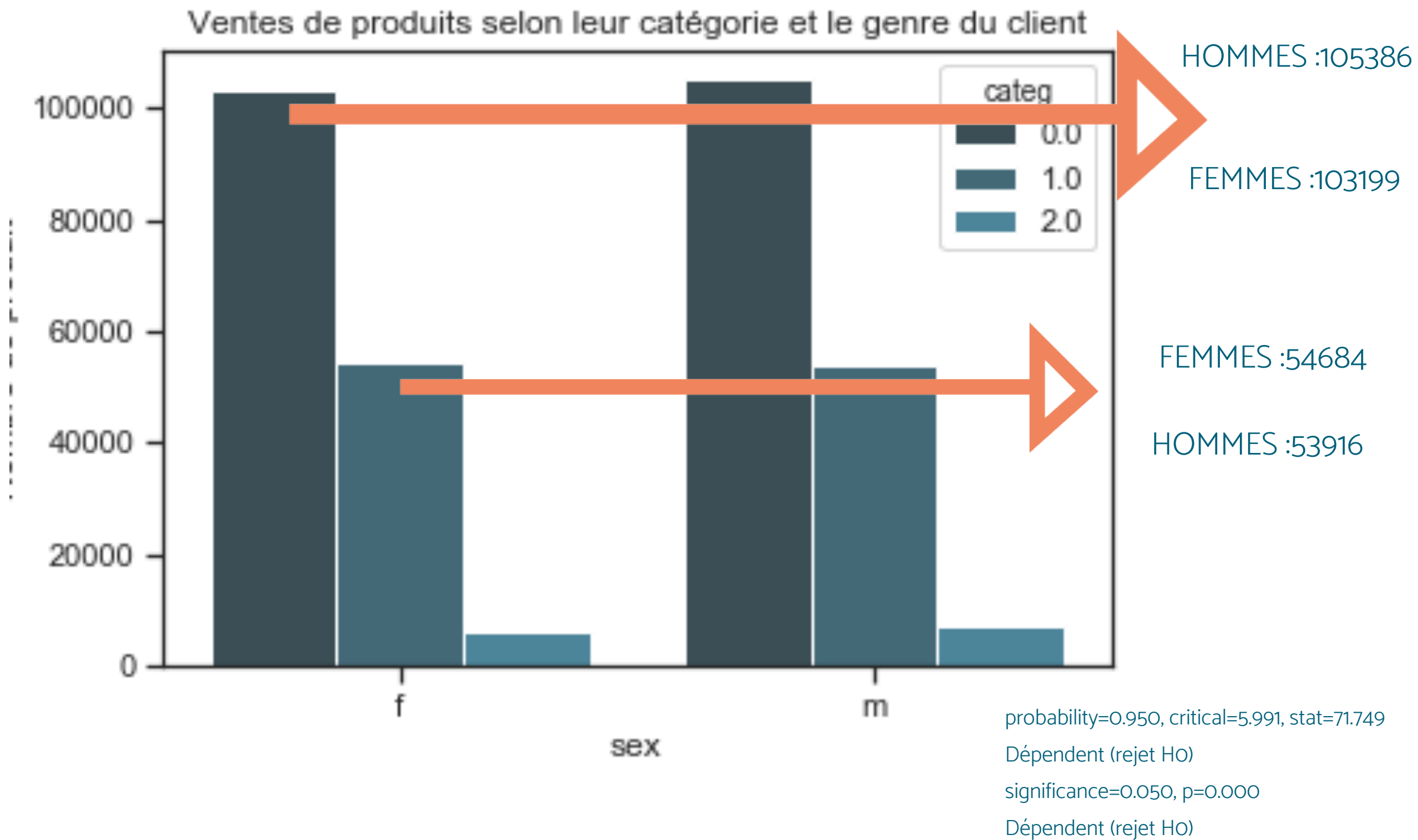
LA RÈGLE EST LA SUIVANTE :

SI LE KHI-CARRÉ CALCULÉ EST INFÉRIEUR AU KHI-CARRÉ THÉORIQUE : INDÉPENDANCE

SI LE KHI-CARRÉ CALCULÉ EST SUPÉRIEUR AU KHI-CARRÉ THÉORIQUE : DÉPENDANCE



Y A-T-IL UNE CORRÉLATION ENTRE LE SEXE DES CLIENTS
ET LES CATÉGORIES DE PRODUITS ACHETÉS ?



EXISTE UNE DÉPENDANCE ENTRE LA CATÉGORIE DE PRODUITS ACHETÉS ET LE GENRE DE CLIENTS

UNE VARIABLES QUALITATIVE ET UNE VARIABLE QUANTITATIVE

ANOVA : ANALYSE DE VARIANCE UNIVARIÉE

L'ANOVA CORRESPOND À UN MODÈLE LINÉAIRE GAUSSIEN DANS LEQUEL TOUTES LES VARIABLES EXPLICATIVES (LES X_j) SONT QUALITATIVES.

POUR DÉCIDER L'ACCEPTATION OU LE REJET DE L'HYPOTHÈSE NULLE, IL RESTE À COMPARER LA P-VALUE, PAR CONVENTION, LORSQUE :

LA VALEUR P EST 0.001 : NOUS DISONS QU'IL Y A DES PREUVES SOLIDES QUE LA CORRÉLATION SOIT SIGNIFICATIVE.

LA VALEUR P EST DE 0,05 : IL Y A DES PREUVES MODÉRÉES QUE LA CORRÉLATION SOIT SIGNIFICATIVE.

LA VALEUR DE P EST 0.1 : IL Y A PEU DE PREUVES QUE LA CORRÉLATION SOIT SIGNIFICATIVE.



Y A-T-IL UNE CORRÉLATION ENTRE L'ÂGE DES CLIENTS
ET
LES CATÉGORIES DE PRODUITS ACHETÉS?

=====

ANOVA SUMMARY

=====

Source	ddof1	ddof2	F	p-unc	np2
-----	-----	-----	-----	-----	-----
categ	2	334946	23283.327	0.000	0.122

DEGRÉ DE LIBERTÉ (SCT)

$M.N-1 = 334948$

DDL(SCINTRA) = 334946

DDL (SCINTER) = 2

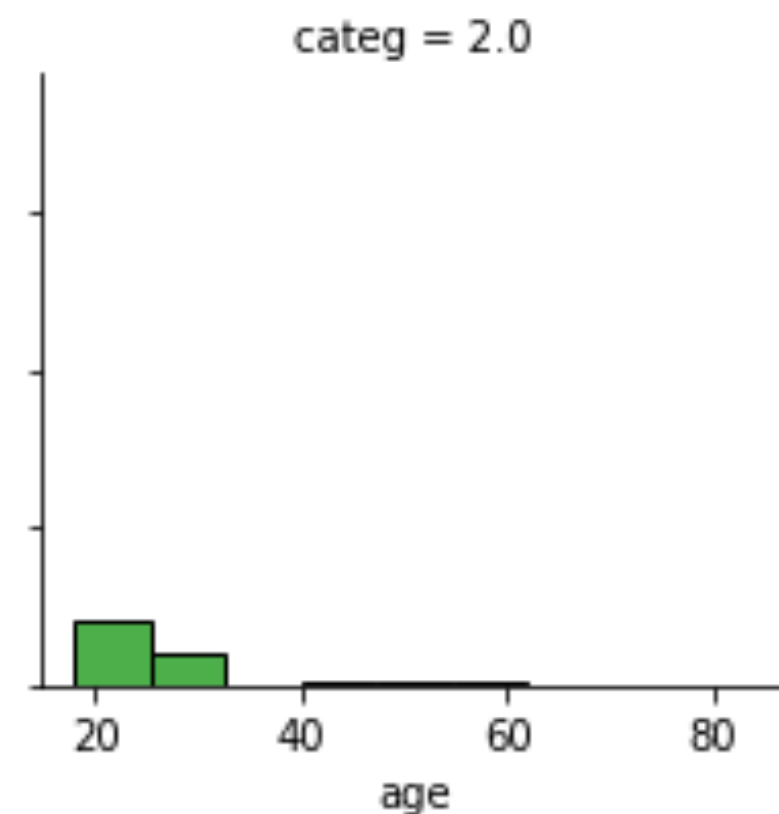
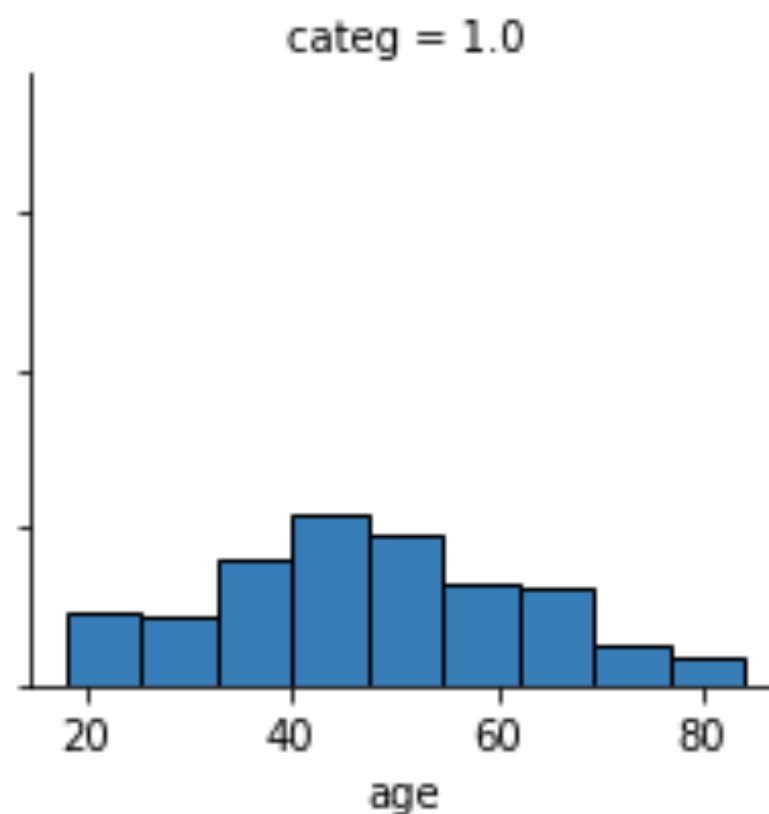
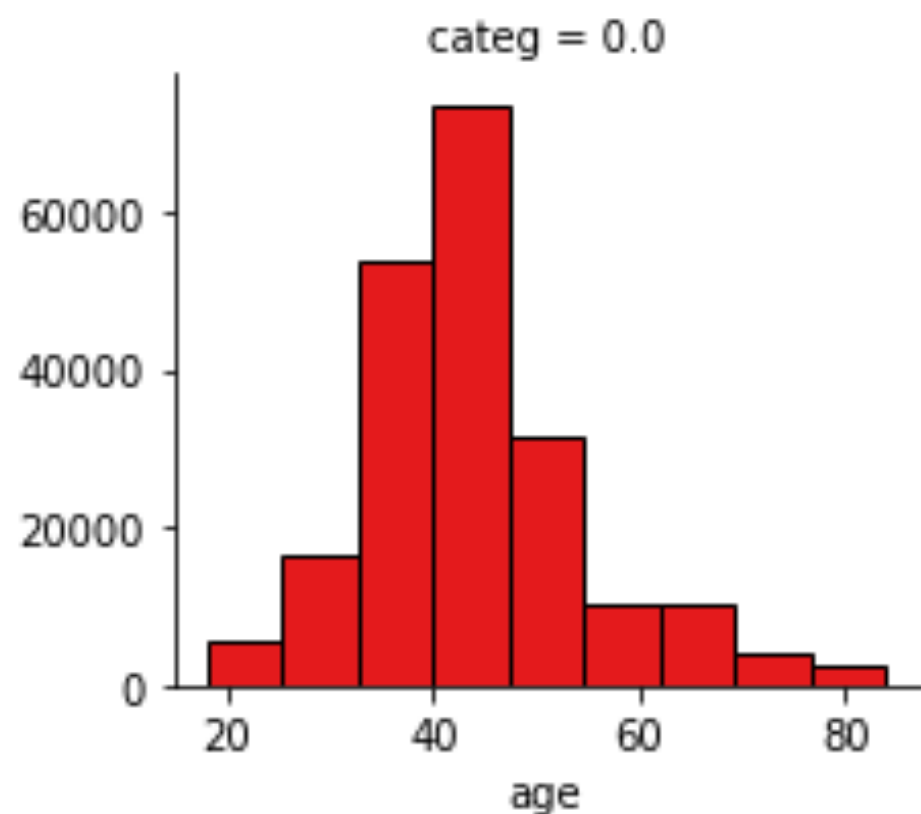
STATISTIQUE F

LOI DE FISHER-SNEDECOR

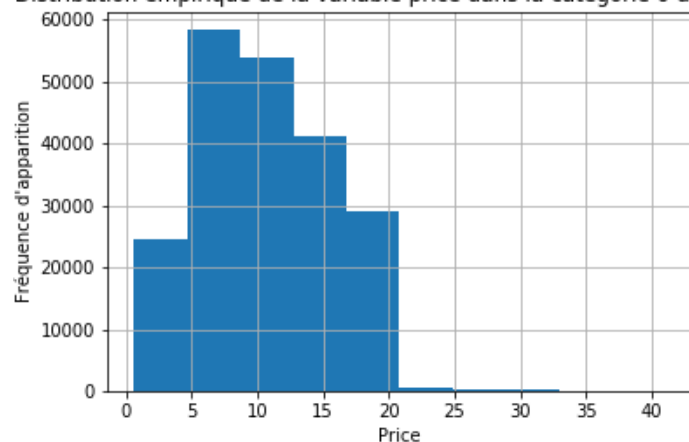
F TRÈS ELEVÉ = DÉPENDANCE

F TRÈS FAIBLE = INDEPENDENCE

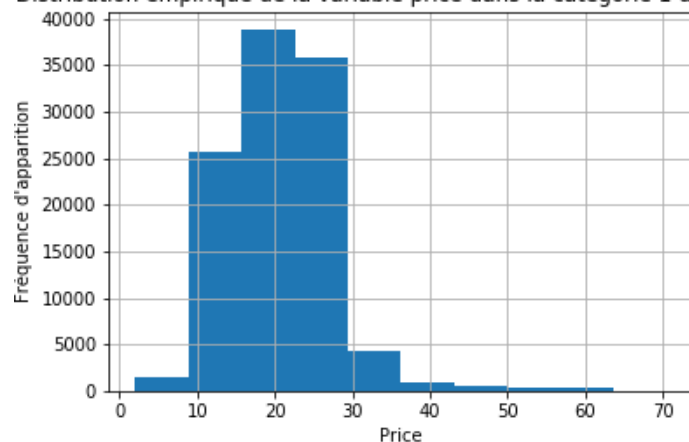
P-VALEUR



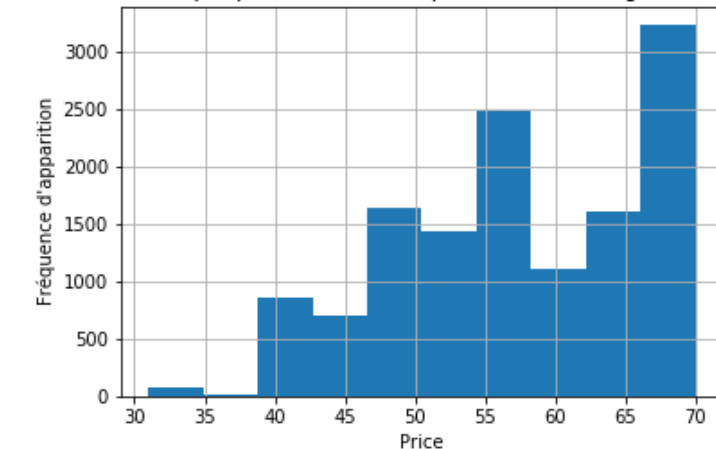
Distribution empirique de la variable price dans la categorie 0 de produ



Distribution empirique de la variable price dans la categorie 1 de produ



Distribution empirique de la variable price dans la categorie 2 de produ



LES PRODUITS CATÉGORIE 0
(PRIX FAIBLE) SONT ACHETÉS
PRINCIPALEMENT POUR LES
CLIENTS ENTRE 30 ET 50

LES PRODUITS CATÉGORIE 1
(PRIX MODÉRÉ) SONT ACHETÉS
DE FORME HOMOGÈNE POUR
LES CLIENTS DE TOUTES ÂGES

LES PRODUITS CATÉGORIE 2
(PRIX PUS ELEVÉ) SONT
ACHETÉS QUE POUR LES
CLIENTS ENTRE 18 ET 30



CONCLUSIONS ET RECOMMANDATIONS

CONCLUSIONS

L'ÂGE DES CLIENTS EST DÉTERMINANT PAR RAPPORT AU MONTANT DU PANIER, LEUR FRÉQUENCE D'ACHATS ET LA QUANTITÉ D'ARTICLES.



L'ÂGE DES CLIENTS EST DÉTERMINANT PAR RAPPORT À LA CATÉGORIE DE PRODUITS QU'ACHÈTENT.



STRATEGIE MARKETING

CIBLER LES CLIENTS DE 30 À 50 ANS EN LEUR
PROPOSANT LES PRODUITS DE CATÉGORIE 1 ET 2



CIBLER LES CLIENTS DE MOINS DE 30 ANS EN LEUR
PROPOSANT LES PRODUITS DE CATÉGORIE 0 ET 1



CIBLER LES CLIENTS DE PLUS DE 40 ANS EN LEUR
PROPOSANT LES PRODUITS DE CATÉGORIE 1 ET 2



...**FIN**