

Rester livres

Analyse des ventes

Mars 2021 → février 2022



1. Exploration

- 3 jeux de données
- Clients
- Produits
- Transactions
- Jointures

2. Nettoyage

- Lignes test
- Dates manquantes
- Valeurs manquantes
- Variables supplémentaires

3. Analyse

- Clients particuliers et professionnels
- B2B : typologie
- B2B : prix et catégories
- B2B : dates d'achat
- Sexe
- Âge et catégorie
- Âge et chiffre d'affaires
- Âge, fréquence et panier
- Catégories et dates d'achat
- Volumes et chiffres d'affaires
- Catégories et prix

4. Tests

- Hypothèses nulles et alternatives
- Valeur p
- ANOVA : Catégorie et âge
- ANOVA : Conditions
- Chi 2 : Catégorie et sexe

Typologie des clients

Typologie des produits

Questions

- Y a-t-il un lien entre le sexe et la catégorie ?
- Y a-t-il un lien entre l'âge et le CA ?
- Y a-t-il un lien entre l'âge et la fréquence ?

Exploration des données

1

clients

8623 individus, aucun doublon

	client_id	sex	birth
7876	c_596	m	2002
7885	c_4626	m	1991
1125	c_7744	m	1997

produits

3287 individus, aucun doublon

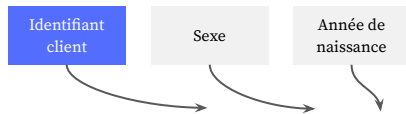
	id_prod	price	categ
2983	1_637	32.99	1
2307	0_1541	11.99	0
2586	0_458	6.75	0

transactions

337016 individus, 126 doublons

	id_prod	date	session_id	client_id
50882	0_1507	2021-04-14 19:46:12.306370	s_20586	c_7929
212527	0_1391	2021-11-28 21:42:09.512676	s_126513	c_159
123784	0_1173	2021-03-04 02:25:49.498064	s_1423	c_4831

Source : github.com/gllmfrnr/oc/tree/master/p4/sources



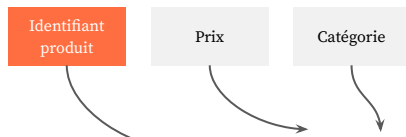
	client_id	sex	birth
0	c_4410	f	1967
1	c_7839	f	1975
2	c_1699	f	1984
3	c_5961	f	1962
4	c_5320	m	1943
...
8618	c_7920	m	1956
8619	c_7403	f	1970
8620	c_5119	m	1974
8621	c_5643	f	1968
8622	c_84	f	1982
8623 rows x 3 columns			
Dataset clients			
8653	c_2043	f	1888
8654	c_2043	f	1888
8655	c_2118	m	1814

→ Hommes et femmes de 17 à 92 ans

→ Aucune valeur manquante

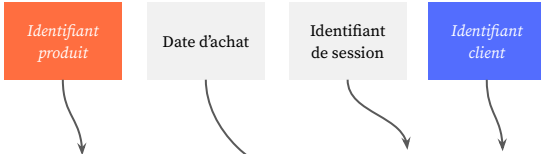
→ Clé primaire : '*client_id*'

Clients tous distincts



	id_prod	price	categ
0	0_1421	19.99	0
1	0_1368	5.13	0
2	0_131	17.99	0
3	1_587	4.99	1
4	0_1507	3.99	0
...
3282	2_23	115.99	2
3283	0_146	17.14	0
3284	0_802	11.22	0
3285	1_140	38.56	1
3286	0_1920	25.16	0
3287 rows x 3 columns			
Dataset produits			
3588	0_1878	252.00	0
3589	1_5730	528.00	1
3590	0_200	141.33	0

- 3 catégories
Identiques aux préfixes de 'id_prod'
- Prix de -1 à 300
Valeur aberrante : -1
Correspond au produit test T_0
- Aucune valeur manquante
- Clé primaire : 'id_prod'
Produits tous distincts



	in		id_prod		date	session_id	client_id
0	38	0	0_1483		2021-04-10 18:37:28.723910	s_18746	c_4450
1	203	1	2_226		2022-02-03 01:55:53.276402	s_159142	c_277
2	245	2	1_374		2021-09-23 15:13:46.938559	s_94290	c_4270
3	317	3	0_2186		2021-10-17 03:27:18.783634	s_105936	c_4597
4	26	4	0_1351		2021-07-17 20:34:25.800563	s_63642	c_1242
...
70	57	337011	1_671		2021-05-28 12:35:46.214839	s_40720	c_3454
71	59	337012	0_759		2021-06-19 00:19:23.917703	s_50568	c_6268
72	133	337013	0_1256		2021-03-16 17:31:59.442007	s_7219	c_4137
73	22	337014	2_227		2021-10-30 16:50:15.997750	s_112349	c_5
74	79	337015	0_1417		2021-06-26 14:38:19.732946	s_54117	c_6714
75 rows 337016 rows x 4 columns							

Dataset transactions

→ 2 clés étrangères

vers **clients** ('client_id')

vers **produits** ('id_prod')

→ Valeurs aberrantes (dates)

74 préfixes test_

Produit T_0

'session_id' s_0

→ Aucune valeur manquante

→ Clé primaire : 'date' + 'client_id'

→ Différences entre les clés étrangères

21 clients inactifs

22 livres invendus

1 livre vendu non référencé

→ Jointures des 3 tables

Ne préserve que les clés de **transactions**

Clé primaire : 'date' + 'client_id'

→ 96 valeurs manquantes

Dans prix et catégorie

Ne concernent que le produit 0_2245

produits				clients			
	id_prod	price	categ		client_id	sex	birth
	2983	1_637	32.99		7876	c_596	m 2002
	2307	0_1541	11.99		7885	c_4626	m 1991
	2586	0_458	6.75		1125	c_7744	m 1997

index	id_prod	date	session_id	client_id	sex	birth	price	categ
6231	0_2245	2021-06-17 03:03:12.668129	s_49705	c_1533	m	1972	NaN	NaN
10797	0_1507	2021-04-14 19:46:12.306370	s_20586	c_7954	m	1973	NaN	NaN
14045	0_1391	2021-11-28 21:42:09.512676	s_126513	c_159	f	1975	NaN	NaN
17480	0_1173	2021-03-04 02:25:49.498064	s_1423	c_4964	f	1982	NaN	NaN
21071	0_2245	2021-03-01 00:09:29.301897	s_3	c_580	m	1988	NaN	NaN
...
322523	0_2245	2021-04-06 19:50:19.462289	s_16036	c_4167	f	1970	NaN	NaN
329226	0_1483	2021-04-10 18:37:28.723910	s_18746	c_4450	f	1977	4.99	0.00
330297	2_226	2022-02-03 01:55:53.276402	s_159142	c_277	f	2000	65.75	2.00
335331	1_374	2021-09-23 15:13:46.938559	s_94290	c_4270	f	1979	10.71	1.00
336020	0_2186	2021-10-17 03:27:18.783634	s_105936	c_4597	m	1963	4.20	0.00
96 rows	4	0_1351	2021-07-17 20:34:25.800563	s_63642	f	1980	8.99	0.00
...
336885	1_671	2021-05-28 12:35:46.214839	s_40720	c_3454	m	1969	31.99	1.00
336886	0_759	2021-06-19 00:19:23.917703	s_50568	c_6268	m	1991	22.99	0.00
336887	0_1256	2021-03-16 17:31:59.442007	s_7219	c_4137	f	1968	11.03	0.00
336888	2_227	2021-10-30 16:50:15.997750	s_112349	c_5	f	1994	50.99	2.00
336889	0_1417	2021-06-26 14:38:19.732946	s_54117	c_6714	f	1968	17.99	0.00

336890 rows x 8 columns

Dataframe finale

Nettoyages

2

id_prod		Date	session_id	client_id	Sexe	Naissance	Âge	Prix	Catégorie
1431	T_0	test_2021-03-01 02:30:02.237420	s_0	ct_1	m	2001	19	-1.0	0.0
2365	T_0	test_2021-03-01 02:30:02.237446	s_0	ct_1	m	2001	19	-1.0	0.0
2895	T_0	test_2021-03-01 02:30:02.237414	s_0	ct_1	m	2001	19	-1.0	0.0
5955	T_0	test_2021-03-01 02:30:02.237441	s_0	ct_0	f	2001	19	-1.0	0.0
7283	T_0	test_2021-03-01 02:30:02.237434	s_0	ct_1	m	2001	19	-1.0	0.0
...
264229	T_0	test_2021-03-01 02:30:02.237416	s_0	ct_1	m	2001	19	-1.0	0.0
288815	T_0	test_2021-03-01 02:30:02.237415	s_0	ct_1	m	2001	19	-1.0	0.0
293003	T_0	test_2021-03-01 02:30:02.237421	s_0	ct_0	f	2001	19	-1.0	0.0
298292	T_0	test_2021-03-01 02:30:02.237423	s_0	ct_1	m	2001	19	-1.0	0.0
317233	T_0	test_2021-03-01 02:30:02.237448	s_0	ct_0	f	2001	19	-1.0	0.0
74 rows x 9 columns									
Les 74 transactions du produit test									
341522	T_0	test_2021-03-01 05:30:05.521418	s_0	ct_0	f	2001	18	-1.0	0.0
588385	T_0	test_2021-03-01 05:30:05.521453	s_0	ct_1	m	2001	18	-1.0	0.0
800005	T_0	test_2021-03-01 05:30:05.521461	s_0	ct_0	f	2001	18	-1.0	0.0

→ 74 lignes test

'id_prod' T_0

'date' préfixe test_

'prix' -1

'session_id' s_0

'client_id' ct_0, ct_1

→ Suppression des lignes

→ Nouvelle clé primaire : 'date'

→ Conversion des dates au format *datetime*

→ Données étalées sur 1 an

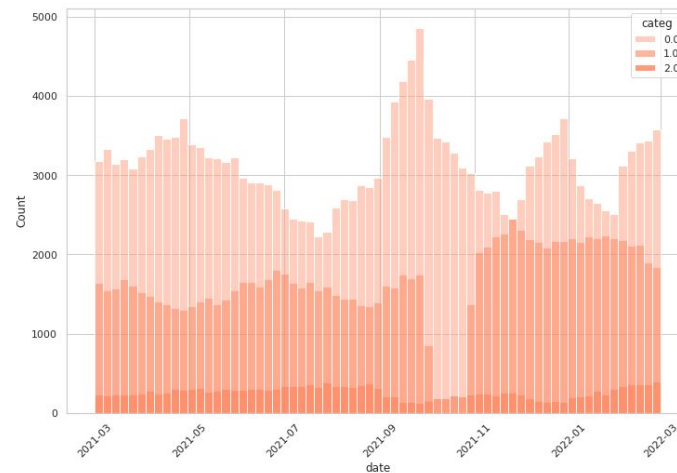
Catégorie 1 manquante

→ Anomalie au mois d'octobre

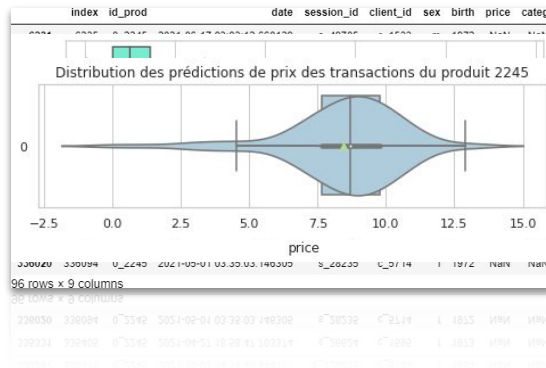
Catégorie 1 manquante

Suppression du mois entier

6% du dataset supprimé



Nombre de commandes par date et catégorie



→ Prix et catégorie manquantes pour le produit 0_2245
96 lignes
0.03% du dataset

→ 3 approches de remplacement :

1. Supprimer les lignes
2. Remplacer par une valeur fixe
3. Modéliser une valeur

→ Valeurs de remplacement

Catégorie : 0 (préfixe de l'identifiant produit)

Prix : 9.99 (prix moyen de la catégorie 0)

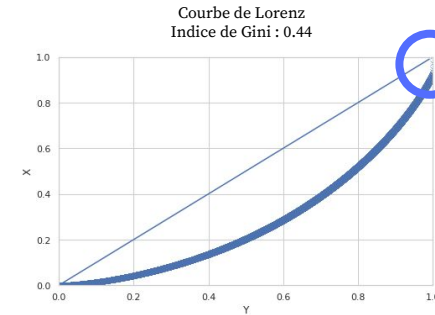
Valeurs confirmées avec des modélisations

	index	id_prod	date	session_id	client_id	sex	birth	price	categ	mois	age	classe_age	total_ventes	ventes_mensuelles	taille_panier_moyen	panier_moyen	total_achats
222582	237190	0_488	2021-03-07 12:31:13.741385	s_3029	c_5452	f	1981	4.39	0.00	3	40	40-50	42	4.00	2.62	12.73	529.09
230815	245968	0_1050	2021-08-22 13:38:34.543695	s_79199	c_2604	m	1956	12.51	0.00	8	65	60-70	66	6.00	1.74	15.88	1,114.80
97485	103950	1_673	2021-11-20 02:56:27.310335	s_122256	c_8252	f	1989	12.99	1.00	11	32	30-40	119	11.00	2.59	12.94	1,490.03
81469	103920	1_613	2021-11-20 05:29:51.310332	s_155520	c_8525	f	1989	15.88	1.00	11	35	30-40	118	11.00	5.20	15.84	1,880.03

Analyse clients et produits

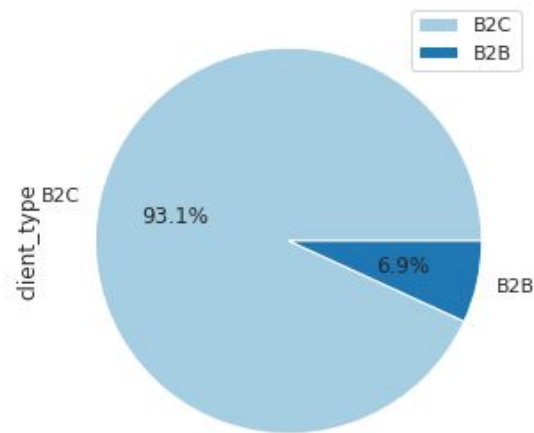
3

- Indice d'inégalité du CA : 0.44
- 4 clients professionnels
 - + de 50.000€ par an
 - + de 250 livres achetés par mois
- Analyse clients en 2 axes : B2B et B2C



Classement des clients par chiffre d'affaires annuel

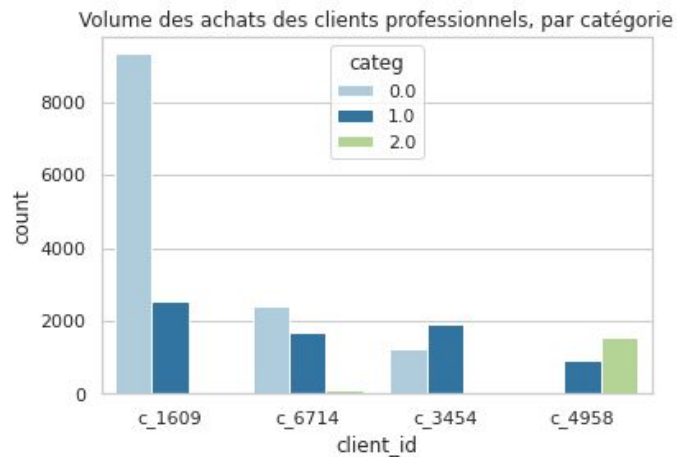
	client_id	panier_moyen	taille_panier_moyen	total_achats	total_ventes	ventes_mensuelles
0	c_1609	12.72	35.41	151,018.91	11861	1,078.00
1	c_4958	55.31	7.42	137,456.83	2463	224.00
2	c_6714	16.66	12.98	69,493.36	4193	381.00
3	c_3454	16.62	9.39	52,845.11	3145	286.00
4	c_8026	13.47	2.88	2,434.49	184	17.00
5	c_7421	13.58	2.83	2,406.17	178	16.00
6	c_7319	13.57	2.75	2,366.20	168	15.00
7	c_3263	13.82	3.05	2,346.34	177	16.00
8	c_8392	13.76	2.76	2,332.08	171	16.00
9	c_2899	53.69	1.62	2,313.54	47	4.00
0	c_5888	23.88	1.85	5,313.24	11	4.00
8	c_8385	13.18	5.18	5,335.08	111	18.00
1	c_3583	13.85	2.00	5,310.34	111	18.00



Proportion des transactions entre professionnels et particuliers

Client :	1609	6714	3454	4958
CA annuel	151,018	137,456	69,493	52,845
Panier moyen	\$	\$ \$	\$ \$	\$ \$ \$

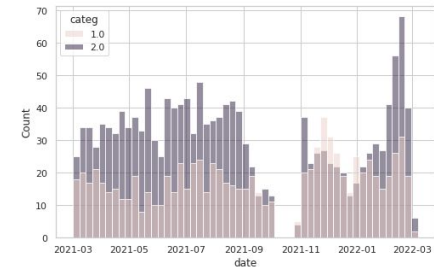
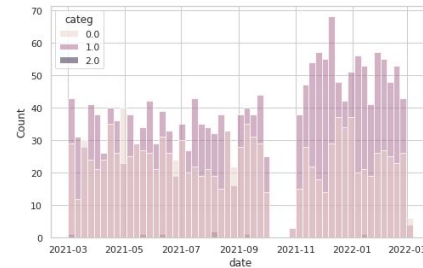
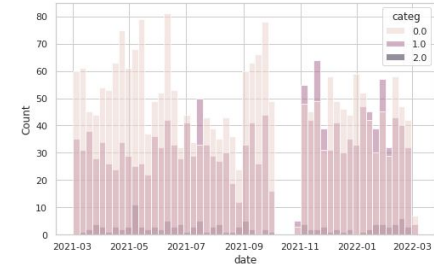
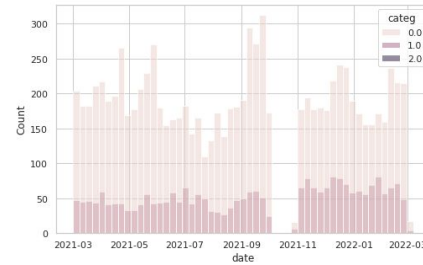
Typologie des clients professionnels



→ Catégorie 0 : prix et panier moyen faibles

→ Catégorie 2 : prix et panier moyen élevés

- Catégorie 0 : septembre
- Catégorie 1 : décembre
- Catégorie 2 : février

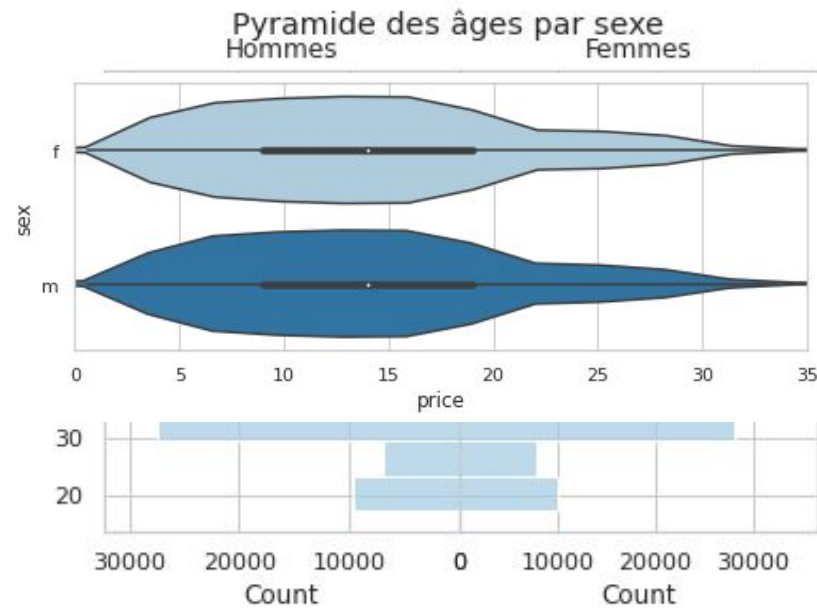


B2B

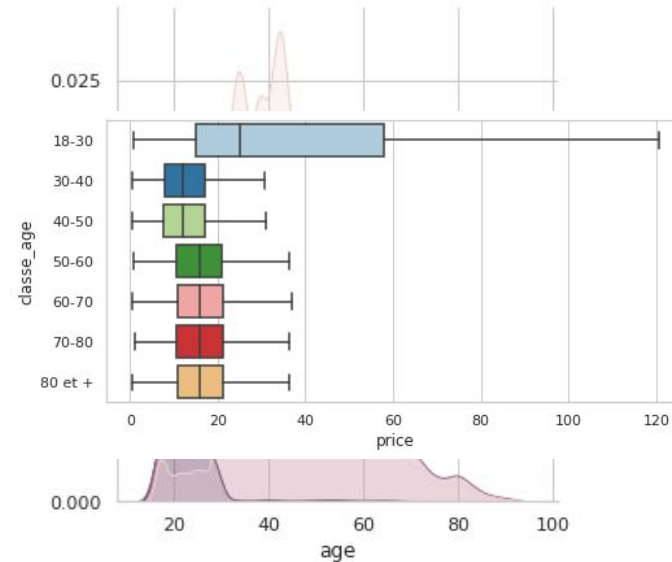
- 7% du chiffre d'affaires
- Prix d'achat liés aux catégories
- Dates d'achat liées aux catégories

Identifiant client	1609	6714	3454	4958
CA annuel	151,018	137,456	69,493	52,845
Panier moyen	\$	\$ \$	\$ \$	\$ \$ \$
Catégorie favorite	0	0, 1	1	2
Pic d'achat	Septembre	Septembre	Décembre	Février

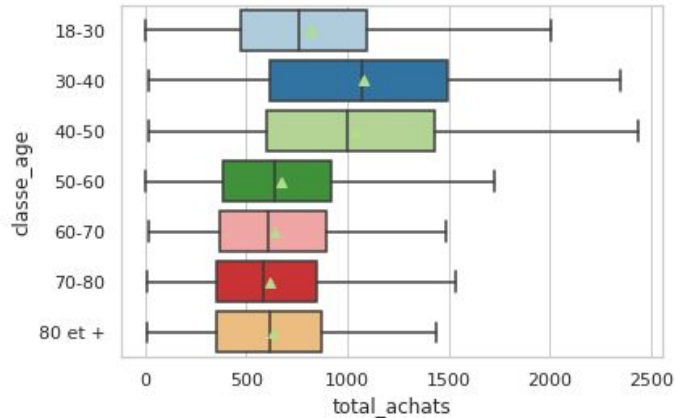
Aucune corrélation entre le sexe et les autres variables



- 3 groupes ordonnés par âge
- Moins de 30 ans
 - Consommateurs quasi-exclusifs de la catégorie 2
 - Prix beaucoup plus hauts
- 30-50 ans
 - Part significative des ventes
 - Consommateurs principaux de la catégorie 0
 - Prix plus bas que les autres clients



Volume des ventes par âge et catégorie



30-50 ans

Les plus gros chiffres d'affaires annuels

Le plus gros volume de ventes

Consommateurs principaux de la catégorie 0



Moins de 30 ans

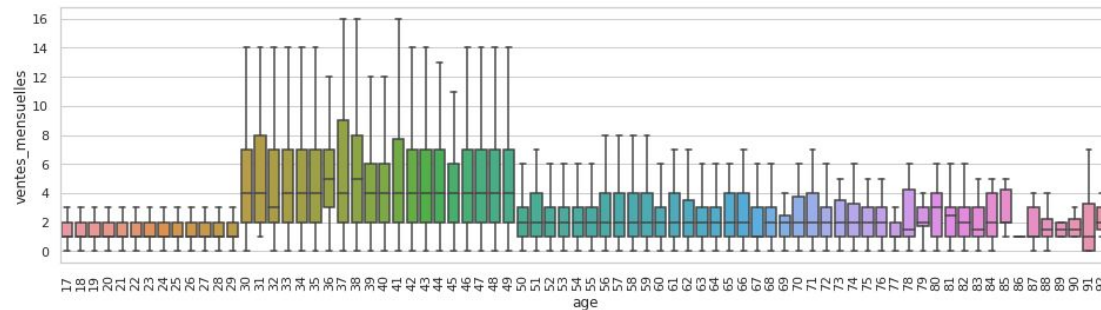
De plus gros chiffres d'affaires que les plus de 50 ans

Consommateurs principaux de la catégorie 2

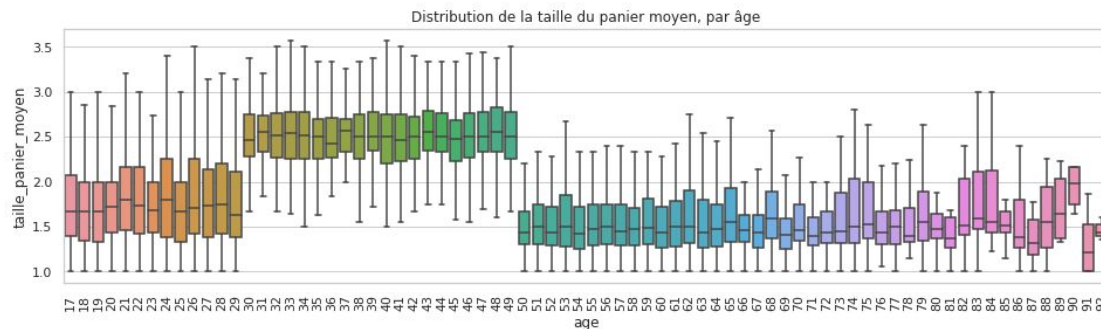
Distribution des chiffres d'affaires annuels, par classe d'âge

Confirmation des 3 groupes

Nombre moyen
d'achats par mois
Moins de 30 ans homogènes



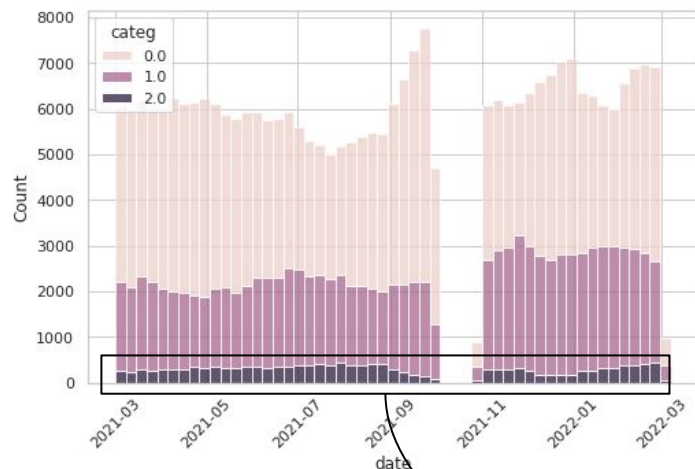
Nombre moyen
d'articles par panier



B2C

- 3 catégories de clients, par âge
- Prix d'achat liés aux catégories
- Pas de corrélation avec le sexe

Tranche d'âge	Moins de 30 ans	De 30 à 50 ans	Plus de 50 ans
Catégorie phare	2	0	0, 1
Panier moyen	\$ \$ \$	\$	\$ \$
Taille du panier	++	+++	+
Profil	Étudiant post-bac	Lecteur de livres de poche	Tous profils



→ Catégorie 0

Pic en septembre

Second pic en décembre

→ Catégorie 1

Pic aux fêtes de fin d'année

Cadeaux de fin d'année

Vacances estivales

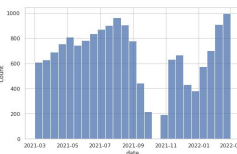
→ Catégorie 2

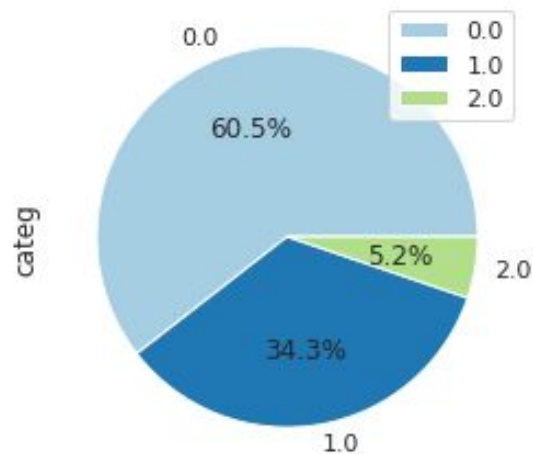
Pic en septembre et février

Constant toute l'année

Chute après septembre

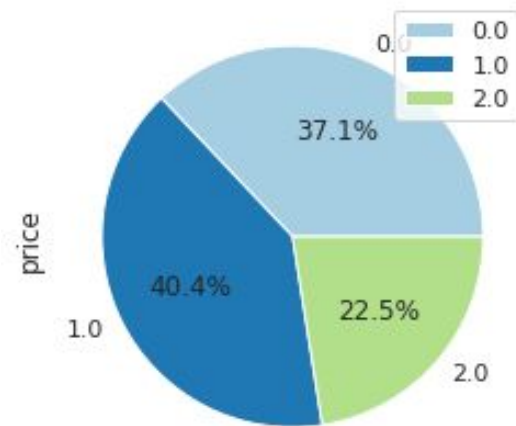
Creux à Noël





Volume des ventes

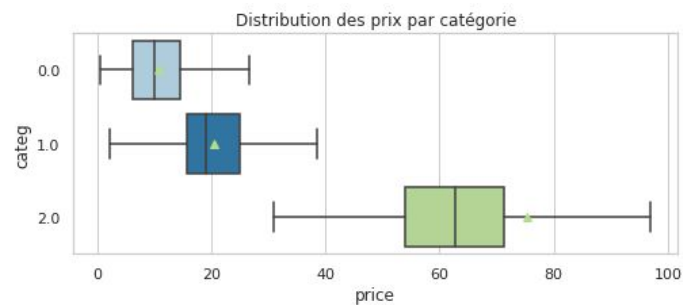
Catégorie 0 : 60%
Catégorie 1 : 35%
Catégorie 2 : 5%



Chiffre d'affaires

Catégorie 0 : 37%
Catégorie 1 : 40%
Catégorie 2 : 23%

Confirmation des 3 catégories de prix ordonnées

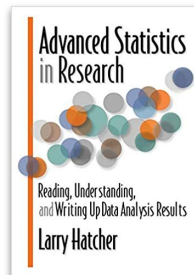
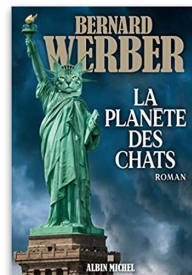


Catégorie :	0	1	2
Prix médian	9.99	19.08	62.83
Prix minimum	0.62	2	30.99
Prix maximum	40.99	80.99	300
Profil type	Livres de poche	Romans, nouvelles sorties	Livres scolaires, beaux livres

Produits

3 catégories de prix ordonnées

- Équilibre des chiffres d'affaires
- Volumes de transactions inégaux
- Périodicités différentes



Catégorie :	0	1	2
Pic d'achat	Septembre	Décembre	Février, août
Occasion	Rentrée scolaire	Vacances, cadeaux	Rentrées universitaires
Prix	\$	\$ \$	\$ \$ \$
Profil type	Livres de poche	Romans, nouvelles sorties	Livres scolaires, spécialisés, beaux livres

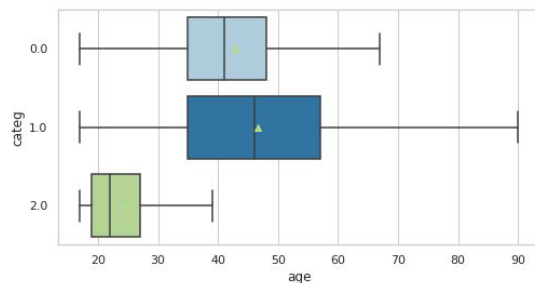
Tests statistiques

4

Un test statistique nécessite :

1. Des données
2. Une hypothèse nulle
3. Une hypothèse alternative

Catégorie et âge



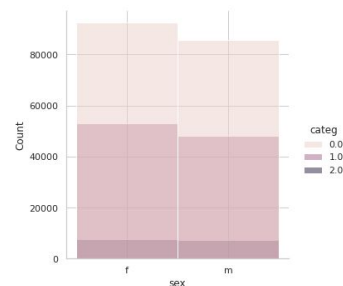
H0

L'âge n'influe pas sur la catégorie.

H1

L'âge influe sur la catégorie.

Catégorie et sexe



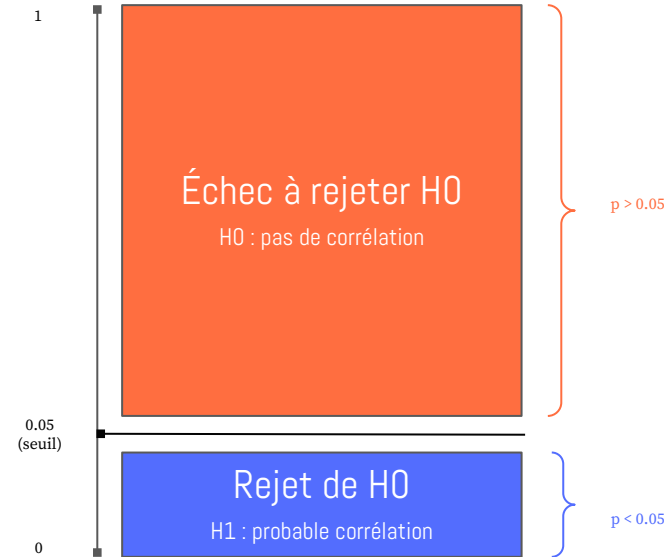
H0

Le sexe n'a pas d'impact sur la catégorie.

H1

Le sexe a un impact sur la catégorie.

- Nombre entre 0 et 1
- Probabilité que l'hypothèse nulle soit vraie
- Seuil généralement fixé à 0.05



→ Analyse de variance entre groupes

Variable qualitative : 'categ'

Variable quantitative : 'age'

→ Hypothèse nulle

Les moyennes des groupes sont égales

Probablement pas de corrélation

→ Hypothèse alternative

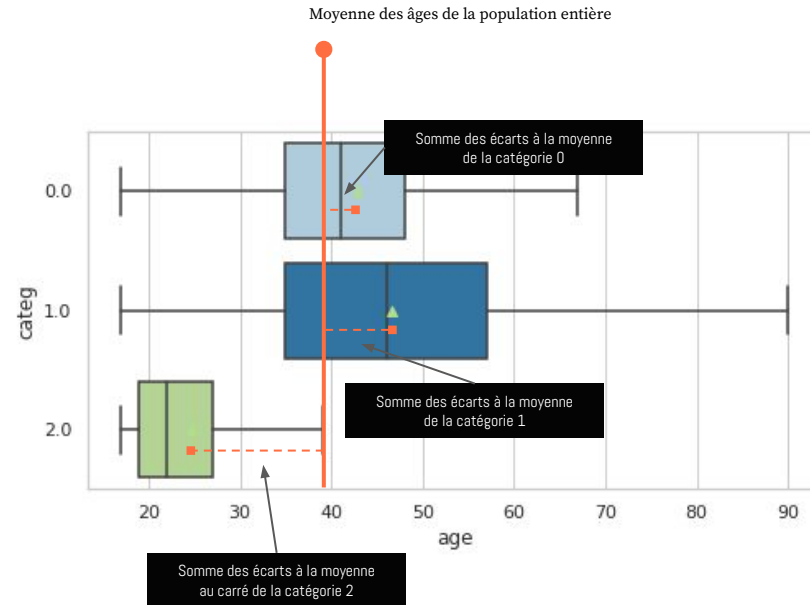
Les moyennes des groupe sont différentes

Probable corrélation des variables

Test d'ANOVA

p-value : 0.0

H0 rejetée : Probable corrélation



Indépendance des échantillons

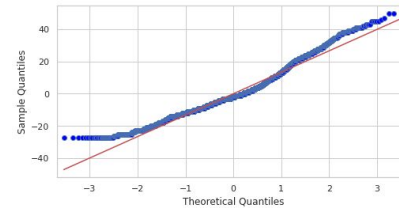
Normalité de la distribution

Test de Shapiro

Sans boxcox

P-value : 0.0

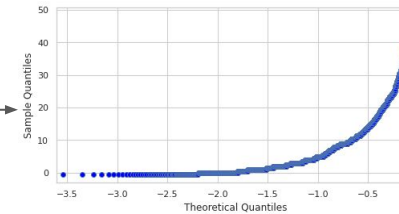
Rejet de l'hypothèse 0 : la distribution n'est probablement pas normale (condition requise pour l'ANOVA)



Après boxcox

P-value : 1.0

Hypothèse 0 acceptée : la distribution est probablement normale, ANOVA est validée



Homogénéité des variances

Test de Levene

p-value : 0.0

h1 : les variances ne sont pas égales

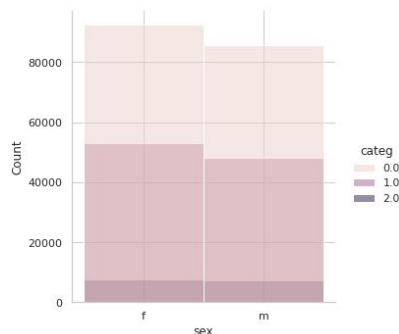


Test alternatif : Welch ANOVA

p-value : 0.0

h1 : les moyennes sont différentes





↓ Catégorie	Femme	Homme
0	92617	85774
1	53161	48221
2	7599	7136

→ Test d'indépendance entre 2 variables catégoriques

→ 2 degrés de liberté

(Nombre de modalités de 'sex' - 1) x (Nombre de modalités de 'categ' - 1)

→ Conditions

Pas besoin de normalité (uniquement pour quantitatives continues)

Au moins 1 valeur dans chaque cellule de la table de contingence

Au moins 80% des valeurs égales ou supérieures à 5

→ Résultats

P-value : 0.2

H0 acceptée : indépendance des échantillons

Résultats

- 3 catégories de livres ordonnées par prix
- 3 groupes cibles de clients
 - ◆ Grands lecteurs de 30 à 50 ans
 - ◆ Étudiants de moins de 30 ans
- Corrélations
 - Le sexe n'est pas corrélé à la catégorie
 - L'âge est corrélé au chiffre d'affaires

Recommandations

- Promouvoir chaque catégorie en fonction de la période de l'année
- Offrir un espace en ligne réservé aux professionnels