Lapage

Analysez des ventes

Période : mars 2021 – fev 2023

PLAN

Nos données
- Clients
- Produits
- Transactions

Rapprochement des données

Nettoyage et préparation des données
- Valeurs manquantes
- Valeurs aberrantes

Analyse des données
- Prix
- Chiffre d'affaires
- Saisonnalité

Relations entre données

Nos données

Données clients

Identifiant client

Genre

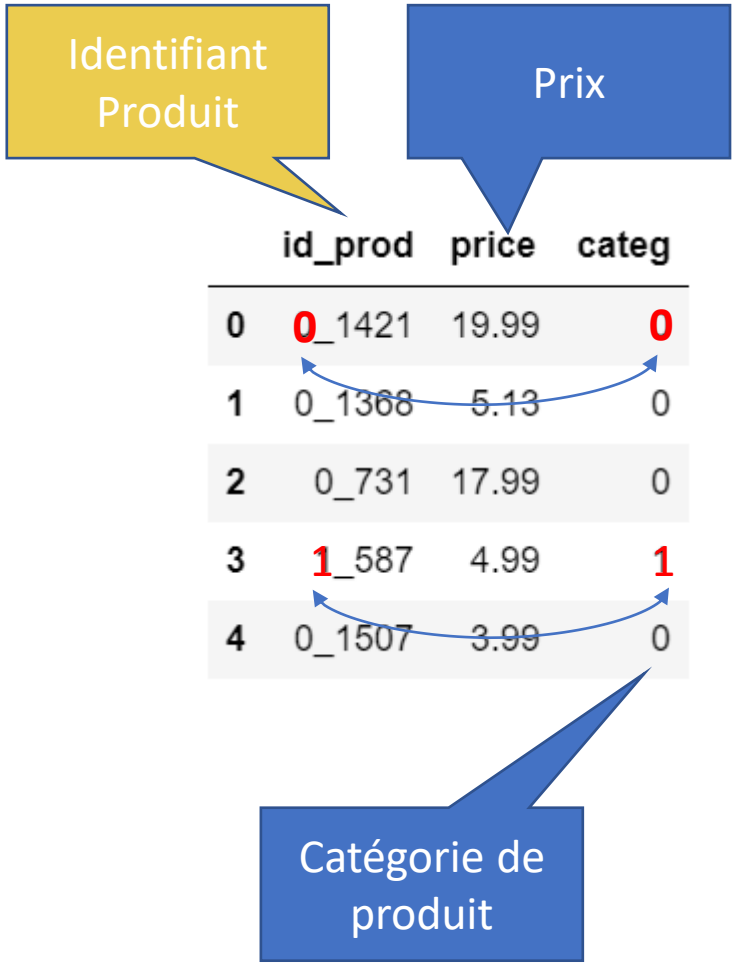
	client_id	sex	birth
0	c_4410	f	1967
1	c_7839	f	1975
2	c_1699	f	1984
3	c_5961	f	1962
4	c_5320	m	1943

Année de naissance

- 8623 enregistrements.
- 8623 identifiants uniques de clients de sexe masculin et féminin.
- 76 dates de naissance comprises entre 1929 et 2004.
- 21 clients n'ont pas effectué de transactions au cours des deux années.

```
('c_1223', 'c_2706', 'c_3017', 'c_3443', 'c_3526', 'c_3789',
'c_4086', 'c_4358', 'c_4406', 'c_4447', 'c_5223', 'c_5245', 'c_587',
'c_6735', 'c_6862', 'c_6930', 'c_7584', 'c_8253', 'c_8381', 'c_862',
'c_90' )
```

Données 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

- 3287 enregistrements.
- 3287 identifiants unique de produits appartenant à trois catégories distincts (0; 1; 2).
- Prix compris entre – 1 € et 300 €.
- 21 produits référencées qui n'ont pas fait l'objet d'une transaction sur les deux années.

```
('0_1014', '0_1016', '0_1025', '0_1062', '0_1119',
'0_1318', '0_1620', '0_1624', '0_1645', '0_1780', '0_1800',
'0_2308', '0_299', '0_310', '0_322', '0_510', '1_0',
'1_394', '2_72', '2_86', '2_87').
```

Données transactions

Identifiant produit

Date de la transaction

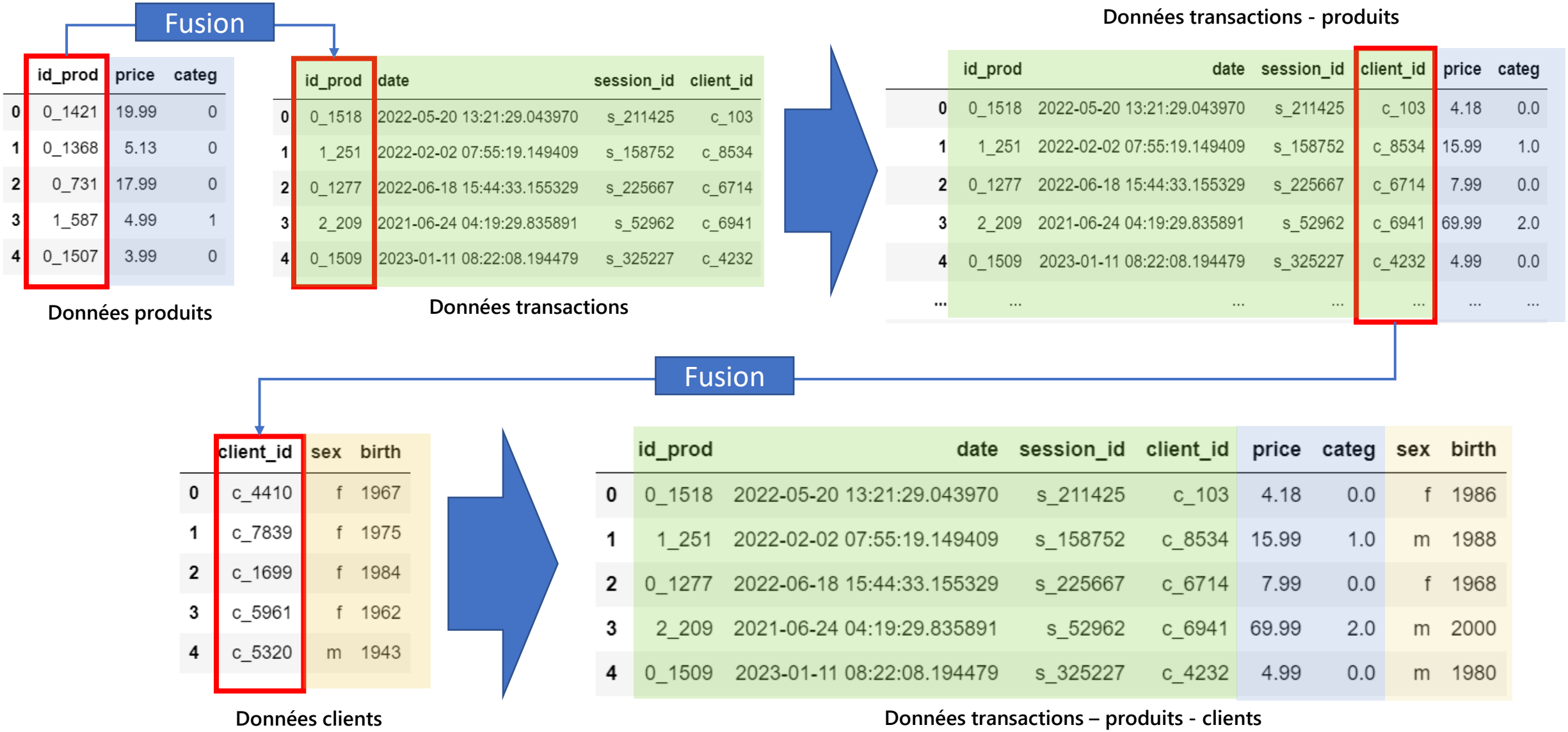
Identifiant client

	id_prod	date	session_id	client_id
0	0_1518	2022-05-20 13:21:29.043970	s_211425	c_103
1	1_251	2022-02-02 07:55:19.149409	s_158752	c_8534
2	0_1277	2022-06-18 15:44:33.155329	s_225667	c_6714
3	2_209	2021-06-24 04:19:29.835891	s_52962	c_6941
4	0_1509	2023-01-11 08:22:08.194479	s_325227	c_4232

Identifiant de la transaction

- 679532 enregistrements
- 3267 identifiants uniques de produits.
- 8602 identifiants unique de clients.
- Date de transaction pour une période allant du 01/03/2021 au 28/02/2023.
- 126 enregistrements en doubles.
- Le produit référencé "0_2245" de catégorie 0, vendu 221 fois sans posséder de données dans la table "produit".

Rapprochement des données


Fusion
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
Données produits
id_prod date session_id client_id
0 0_1518 2022-05-20 13:21:29.043970 s_211425 c_103
1 1_251 2022-02-02 07:55:19.149409 s_158752 c_8534
2 0_1277 2022-06-18 15:44:33.155329 s_225667 c_6714
3 2_209 2021-06-24 04:19:29.835891 s_52962 c_6941
4 0_1509 2023-01-11 08:22:08.194479 s_325227 c_4232
Données transactions
Données transactions - produits
id_prod date session_id client_id price categ
0 0_1518 2022-05-20 13:21:29.043970 s_211425 c_103 4.18 0.0
1 1_251 2022-02-02 07:55:19.149409 s_158752 c_8534 15.99 1.0
2 0_1277 2022-06-18 15:44:33.155329 s_225667 c_6714 7.99 0.0
3 2_209 2021-06-24 04:19:29.835891 s_52962 c_6941 69.99 2.0
4 0_1509 2023-01-11 08:22:08.194479 s_325227 c_4232 4.99 0.0
... ...
Fusion
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
Données clients
id_prod date session_id client_id price categ sex birth
0 0_1518 2022-05-20 13:21:29.043970 s_211425 c_103 4.18 0.0 f 1986
1 1_251 2022-02-02 07:55:19.149409 s_158752 c_8534 15.99 1.0 m 1988
2 0_1277 2022-06-18 15:44:33.155329 s_225667 c_6714 7.99 0.0 f 1968
3 2_209 2021-06-24 04:19:29.835891 s_52962 c_6941 69.99 2.0 m 2000
4 0_1509 2023-01-11 08:22:08.194479 s_325227 c_4232 4.99 0.0 m 1980
Données transactions – produits - clients

Nettoyage et préparation des données

	id_prod	date	session_id	client_id	price	categ	sex	birth
2633	0_2245	2022-09-23 07:22:38.636773	s_272266	c_4746	NaN	NaN	m	1940
10106	0_2245	2022-07-23 09:24:14.133889	s_242482	c_6713	NaN	NaN	f	1963
11727	0_2245	2022-12-03 03:26:35.696673	s_306338	c_5108	NaN	NaN	m	1978
15675	0_2245	2021-08-16 11:33:25.481411	s_76493	c_1391	NaN	NaN	m	1991
16377	0_2245	2022-07-16 05:53:01.627491	s_239078	c_7954	NaN	NaN	m	1973
...	...	...	...	...	...	...	...	...
669606	0_2245	2021-08-25 09:06:03.504061	s_80395	c_131	NaN	NaN		
670558	0_2245	2022-03-06 19:59:19.462288	s_175311	c_4167	NaN	NaN		
671162	0_2245	2022-05-16 11:35:20.319501	s_209381	c_4453	NaN	NaN		
675554	0_2245	2022-02-11 09:05:43.952857	s_163405	c_1098	NaN	NaN		
677871	0_2245	2021-12-14 22:34:54.589921	s_134446	c_4854	NaN	NaN		

221 rows × 8 columns

Catégorie 0

221 enregistrements avec des valeurs manquantes concernant un seul produit référencé "0_2245" de catégorie "0".

	id_prod	date	session_id	client_id	price	categ	sex	birth
2633	0_2245	2022-09-23 07:22:38.636773	s_272266	c_4746	10.64	0.0	m	1940
10106	0_2245	2022-07-23 09:24:14.133889	s_242482	c_6713	10.64	0.0	f	1963
11727	0_2245	2022-12-03 03:26:35.696673	s_306338	c_5108	10.64	0.0	m	1978
15675	0_2245	2021-08-16 11:33:25.481411	s_76493	c_1391	10.64	0.0	m	1991
16377	0_2245	2022-07-16 05:53:01.627491	s_239078	c_7954	10.64	0.0	m	1973
...	...	...	...	...	...	...	...	...
669606	0_2245	2021-08-25 09:06:03.504061	s_80395	c_131	10.64	0.0	m	1981
670558	0_2245	2022-03-06 19:59:19.462288	s_175311	c_4167	10.64	0.0	f	1979
671162	0_2245	2022-05-16 11:35:20.319501	s_209381	c_4453	10.64	0.0	m	1981
675554	0_2245	2022-02-11 09:05:43.952857	s_163405	c_1098	10.64	0.0	m	1986
677871	0_2245	2021-12-14 22:34:54.589921	s_134446	c_4854	10.64	0.0	m	1968

221 rows × 8 columns

Imputation par la moyenne des prix de la catégorie 0

Imputation par la catégorie 0

	id_prod	date	session_id	client_id	price	categ	sex	birth
3019	T_0	test_2021-03-01 02:30:02.237419	s_0	ct_0	-1.0	0.0	f	2001
5138	T_0	test_2021-03-01 02:30:02.237425	s_0	ct_0	-1.0	0.0	f	2001
9668	T_0	test_2021-03-01 02:30:02.237437	s_0	ct_1	-1.0	0.0	m	2001
10728	T_0	test_2021-03-01 02:30:02.237436	s_0	ct_0	-1.0	0.0	f	2001
15292	T_0	test_2021-03-01 02:30:02.237430	s_0	ct_0	-1.0	0.0	f	2001
...	...	...	...	...	...	...	...	...
577222	T_0	test_2021-03-01 02:30:02.237424	s_0	ct_0	-1.0	0.0	f	2001
592959	T_0	test_2021-03-01 02:30:02.237422	s_0	ct_1	-1.0	0.0	m	2001
607783	T_0	test_2021-03-01 02:30:02.237412	s_0	ct_0	-1.0	0.0	f	2001
625936	T_0	test_2021-03-01 02:30:02.237422	s_0	ct_0	-1.0	0.0	f	2001
670556	T_0	test_2021-03-01 02:30:02.237449	s_0	ct_1	-1.0	0.0	m	2001

74 rows × 8 columns

Suppression du préfixe « test_ »

Catégorie 0

Imputation par la moyenne des prix de la catégorie 0

74 enregistrements représentants des valeurs incohérentes

	id_prod	date	session_id	client_id	price	categ	sex	birth
3019	T_0	2021-03-01 02:30:02.237419	s_0	ct_0	10.64	0.0	f	2001
5138	T_0	2021-03-01 02:30:02.237425	s_0	ct_0	10.64	0.0	f	2001
9668	T_0	2021-03-01 02:30:02.237437	s_0	ct_1	10.64	0.0	m	2001
10728	T_0	2021-03-01 02:30:02.237436	s_0	ct_0	10.64	0.0	f	2001
15292	T_0	2021-03-01 02:30:02.237430	s_0	ct_0	10.64	0.0	f	2001
...	...	...	...	...	...	...	...	...
577222	T_0	2021-03-01 02:30:02.237424	s_0	ct_0	10.64	0.0	f	2001
592959	T_0	2021-03-01 02:30:02.237422	s_0	ct_1	10.64	0.0	m	2001
607783	T_0	2021-03-01 02:30:02.237412	s_0	ct_0	10.64	0.0	f	2001
625936	T_0	2021-03-01 02:30:02.237422	s_0	ct_0	10.64	0.0	f	2001
670556	T_0	2021-03-01 02:30:02.237449	s_0	ct_1	10.64	0.0	m	2001

74 rows × 8 columns

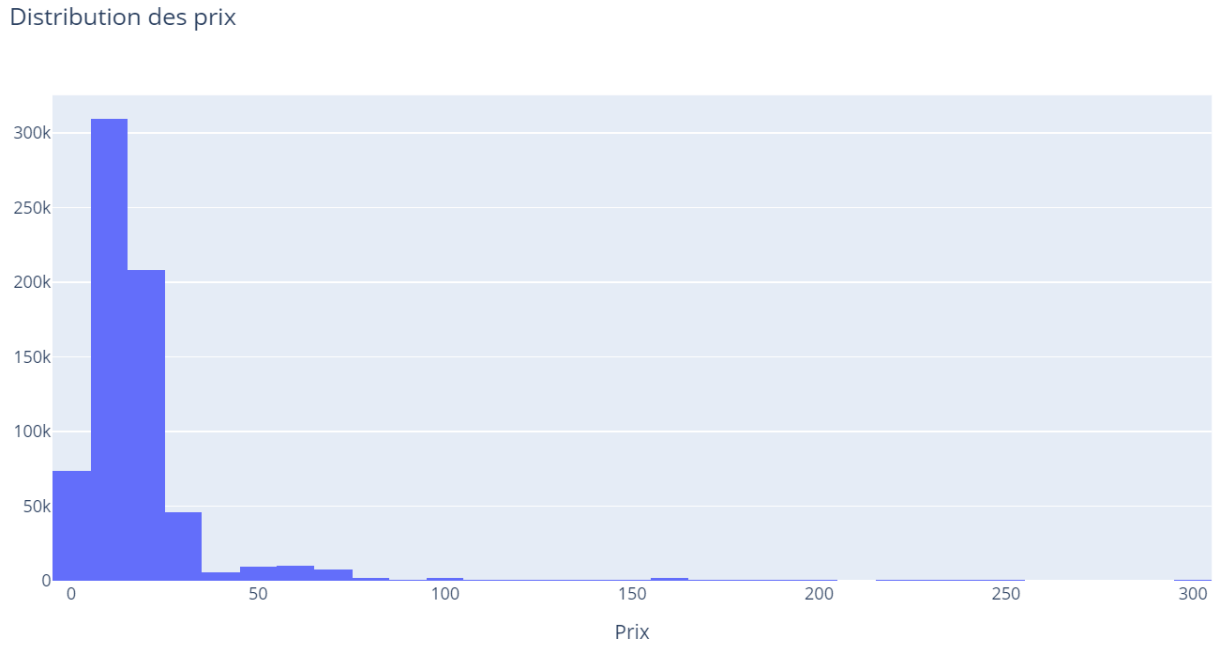
	id_prod	date	session_id	client_id	price	categ	sex	birth
0	0_1518	2022-05-20 13:21:29.043970	s_211425	c_103	4.18	0.0	f	1986
1	1_251	2022-02-02 07:55:19.149409	s_158752	c_8534	15.99	1.0	m	1988
2	0_1277	2022-06-18 15:44:33.155329	s_225667	c_6714	7.99	0.0	f	1968
3	2_209	2021-06-24 04:19:29.835891	s_52962	c_6941	69.99	2.0	m	2000
4	0_1509	2023-01-11 08:22:08.194479	s_325227	c_4232	4.99	0.0	m	1980

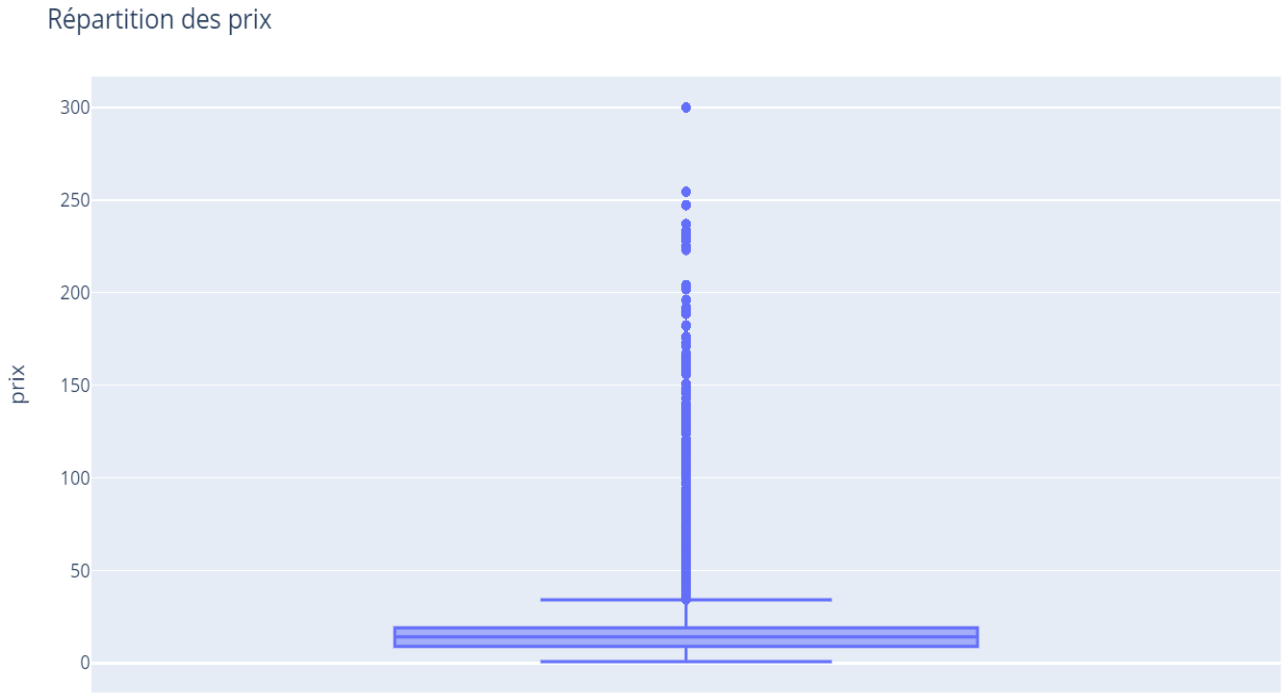
	id_prod	date	session_id	client_id	price	categ	sex	birth	jour	mois	trimestre	annee_exo	jour_semaine	week	age	cat_age
0	0_1518	2022-05-20 13:21:29.043970	s_211425	c_103	4.18	Cat 0	femme	1986	2022-05-20	2022-05	5	Année 2	Friday	semaine	36	30 - 50 ans
1	1_251	2022-02-02 07:55:19.149409	s_158752	c_8534	15.99	Cat 1	homme	1988	2022-02-02	2022-02	4	Année 1	Wednesday	semaine	34	30 - 50 ans
2	0_1277	2022-06-18 15:44:33.155329	s_225667	c_6714	7.99	Cat 0	femme	1968	2022-06-18	2022-06	6	Année 2	Saturday	weekend	54	> 50 ans
3	2_209	2021-06-24 04:19:29.835891	s_52962	c_6941	69.99	Cat 2	homme	2000	2021-06-24	2021-06	2	Année 1	Thursday	semaine	21	< 30 ans
4	0_1509	2023-01-11 08:22:08.194479	s_325227	c_4232	4.99	Cat 0	homme	1980	2023-01-11	2023-01	8	Année 2	Wednesday	semaine	43	30 - 50 ans

Analyse des données

Etude du prix de vente

- le prix modale est égal à : 15.99 €
- le prix moyen est égal à : 17.45 €
- le prix médian est égal à : 13.99 €
- La majorité des prix sont en dessous de 34 €

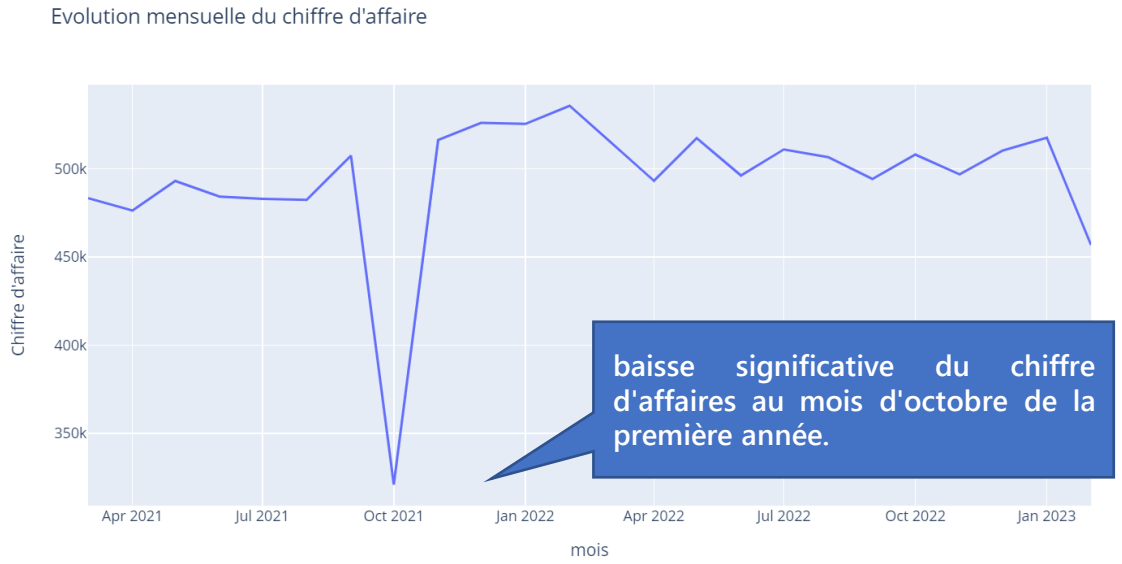




- Le prix maximal est de 300 €
- Le prix minimal est de 0.62 €
- 25% des prix sont inférieurs à : 8.87€
- La moitié des prix sont en dessous 14€
- 75% des prix sont inférieurs à 19€

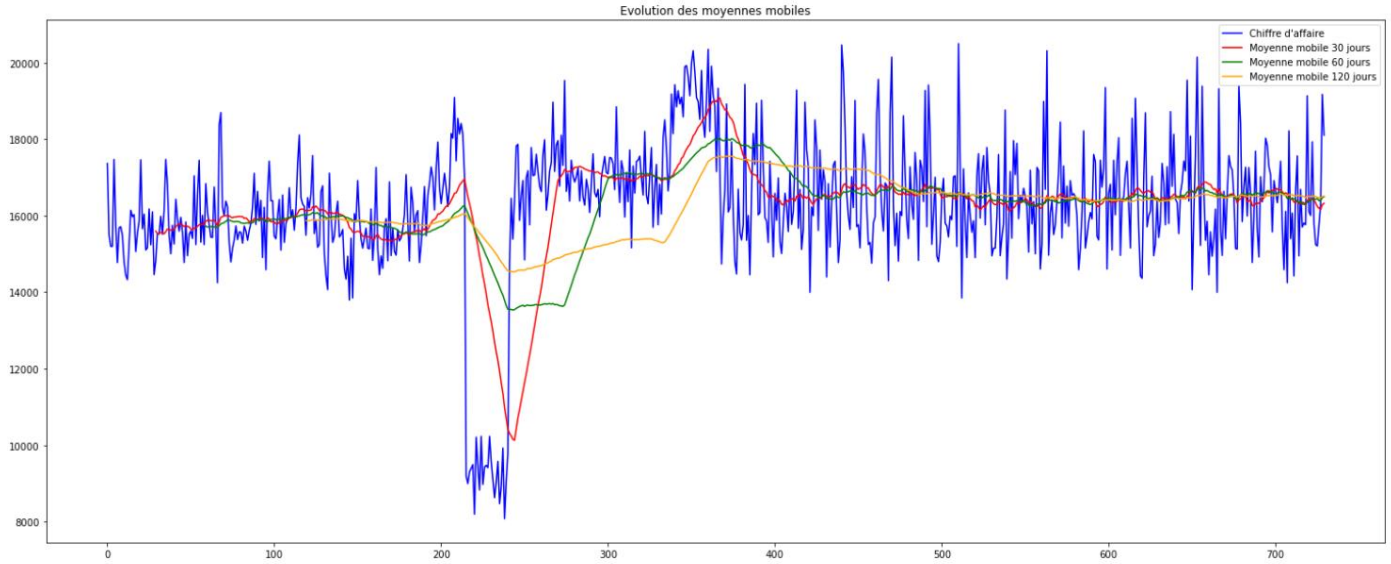
Evolution du chiffre d’affaire

Année	Chiffre d’affaire (en €)
1	5 833 620.33
2	6 023 247.15
Total	11 856 867.48



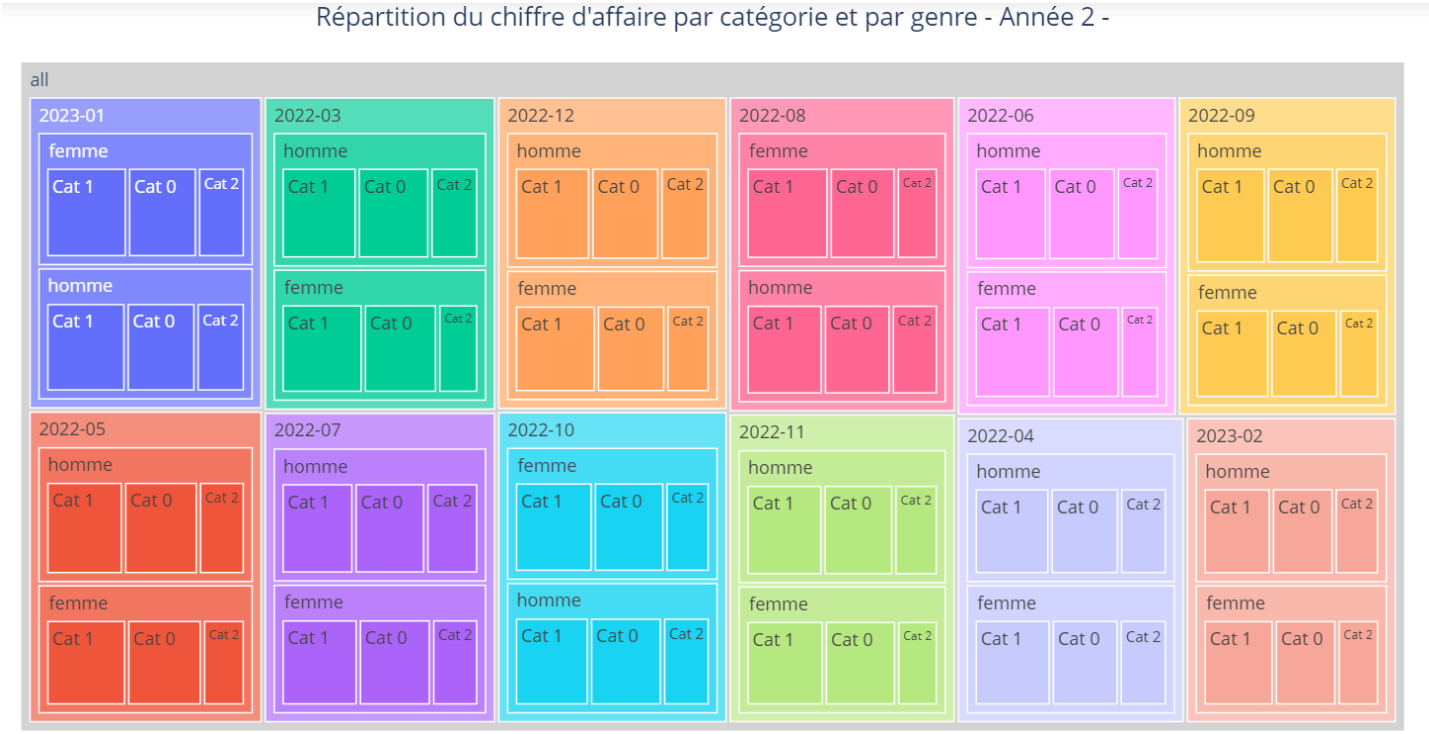
Tendance globale du chiffre d’affaire

L’étude de l’évolution de la moyenne mobile calculée sur des intervalles de 30, 60 et 120 jours montre que la tendance du chiffre d'affaire est bien stationnaire.



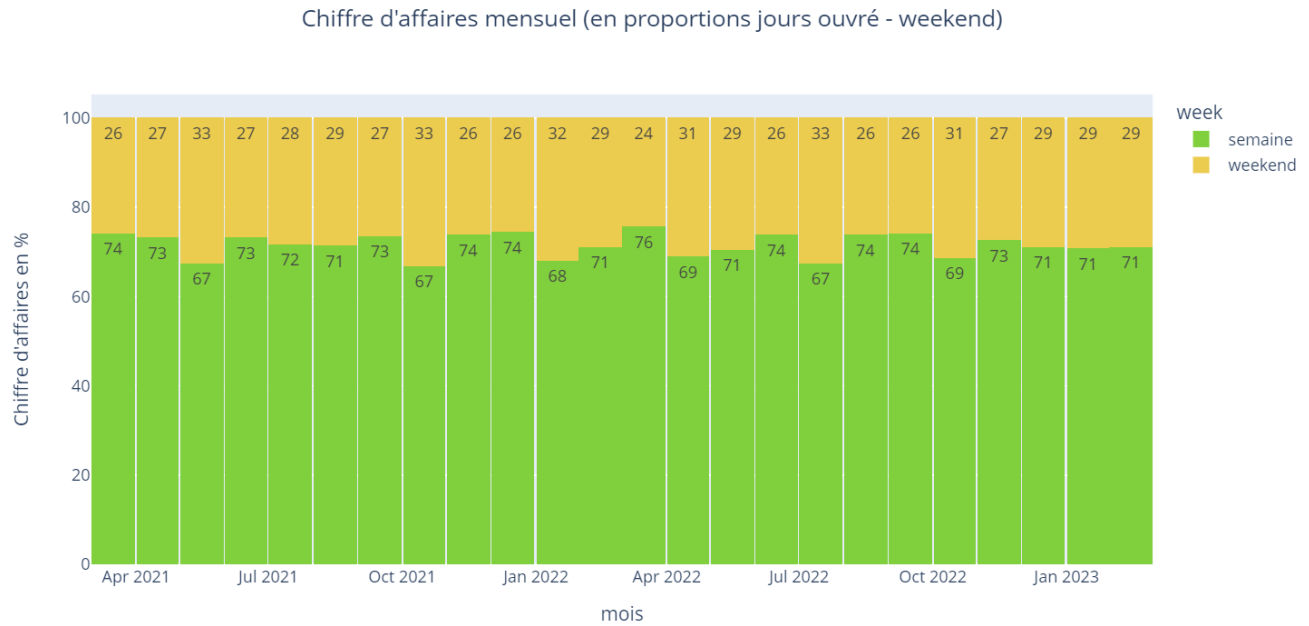
Saisonnalité annuelle

Aucune saisonnalité à l'année. Le chiffre d'affaire annuel par catégorie et par genre est réparti similairement sur tous les mois de l'année.

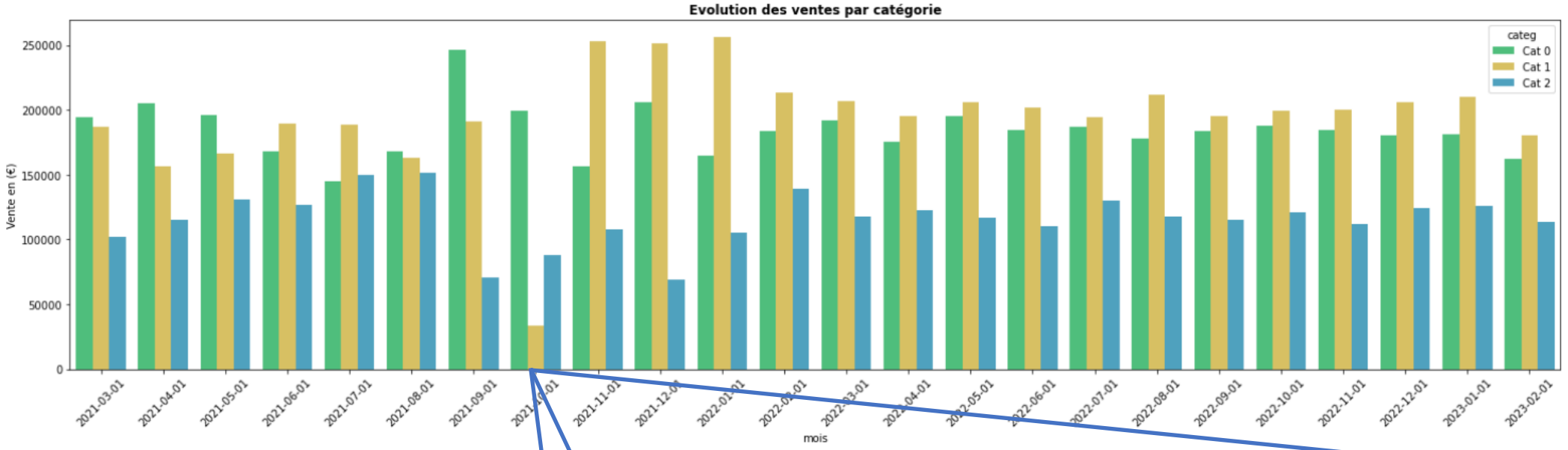


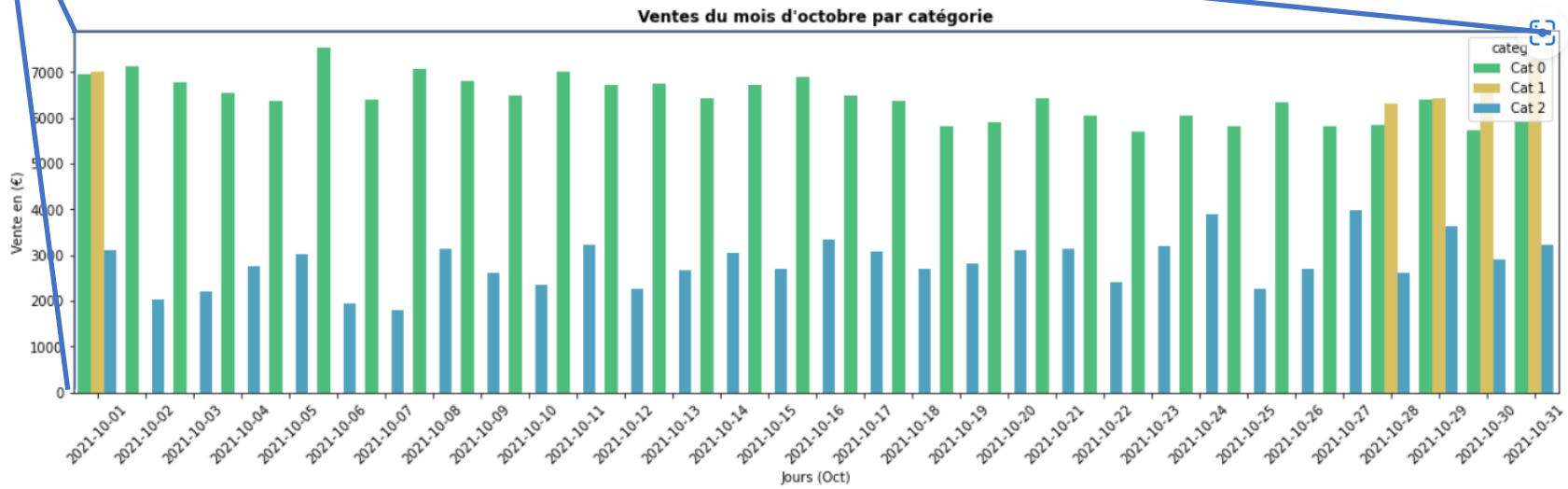
Chiffre d'affaire Weekend / jours ouvrés

Les ventes des weekend sont similaire à celle des jours ouvrés et cela sur toute une année.



Analyse des ventes du mois d'octobre 2021





Absence des ventes de la "catégorie 1" du 02 au 27 octobre 2021.

On peut supposer qu'il y a eu une rupture de stocks ou un problème d'approvisionnement ...

Chiffre d'affaires et références produits

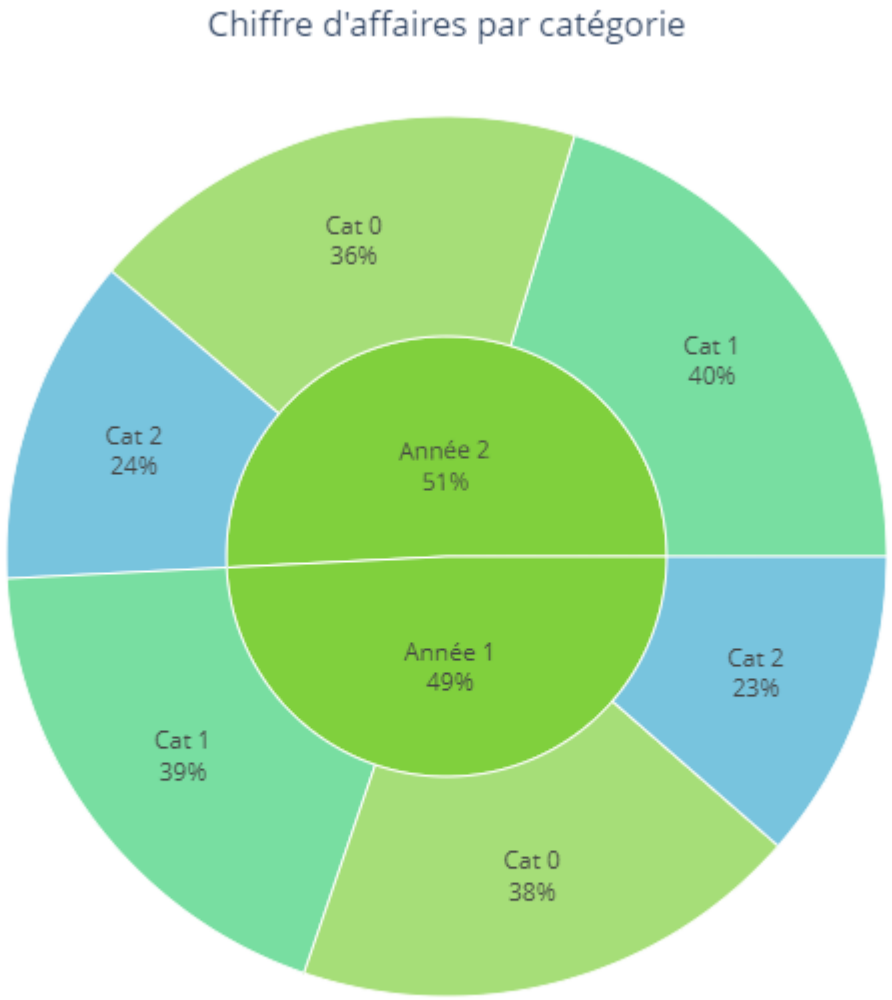
Les 10 premières références en terme de chiffre d'affaire. Toutes de catégorie 2 et achetées par des clients de moins de 30 ans.

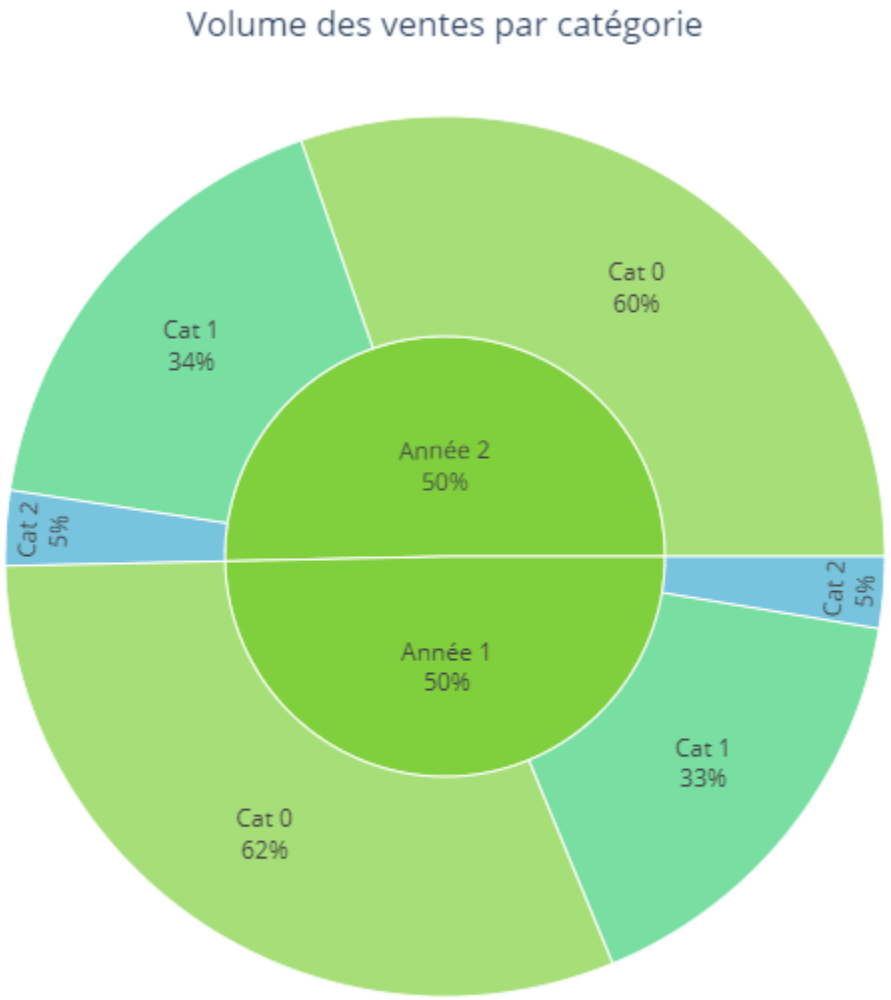
	id_prod	categ	cat_age	price
7854	2_135	Cat 2	< 30 ans	59538.37
7800	2_112	Cat 2	< 30 ans	57299.36
7774	2_102	Cat 2	< 30 ans	51747.50
7925	2_159	Cat 2	30 - 50 ans	48906.65
8071	2_209	Cat 2	< 30 ans	48783.03
7795	2_110	Cat 2	< 30 ans	47808.00
8193	2_39	Cat 2	< 30 ans	46739.94
7948	2_166	Cat 2	< 30 ans	45547.92
8206	2_43	Cat 2	< 30 ans	43953.72
7951	2_167	Cat 2	< 30 ans	41710.24

	id_prod	categ	cat_age	price
2921	0_202	Cat 0	> 50 ans	0.62
4517	0_528	Cat 0	< 30 ans	0.62
564	0_120	Cat 0	< 30 ans	0.66
475	0_1171	Cat 0	30 - 50 ans	0.99
577	0_1203	Cat 0	> 50 ans	0.99
805	0_1283	Cat 0	< 30 ans	0.99
1480	0_1511	Cat 0	> 50 ans	0.99
1562	0_1539	Cat 0	30 - 50 ans	0.99
1883	0_1651	Cat 0	> 50 ans	0.99
2182	0_1759	Cat 0	30 - 50 ans	0.99

Les 10 dernières références en terme de chiffre d'affaire. Toutes de catégorie 0 et achetées par des clients de tout âge.

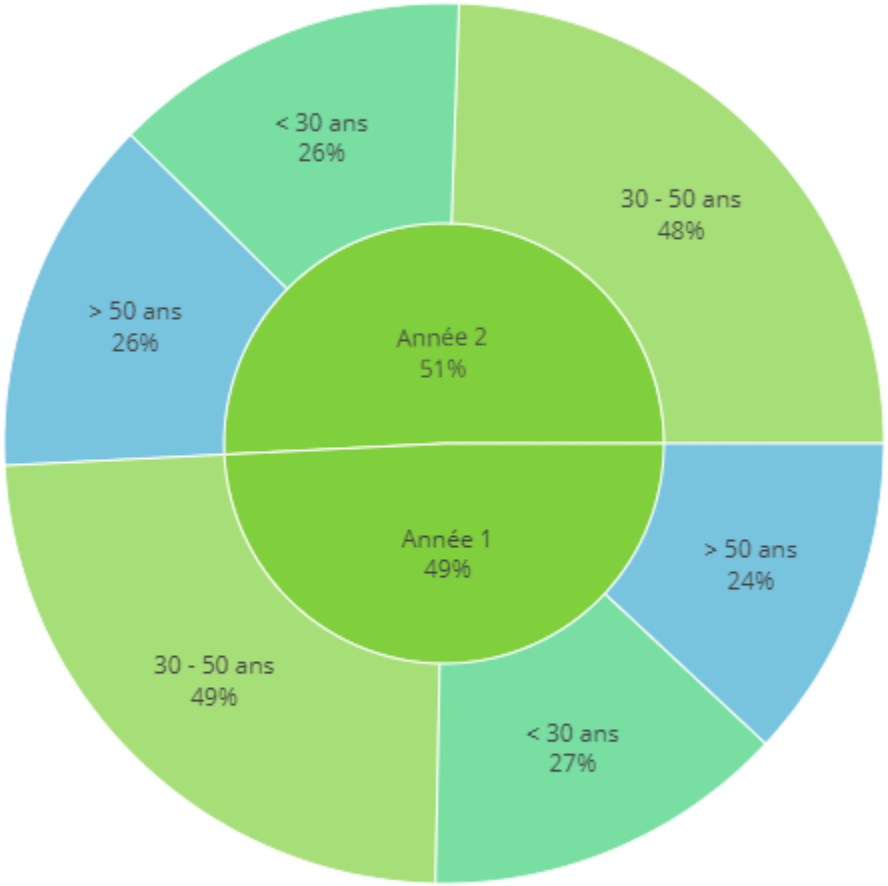
Chiffre d’affaire et catégorie de produit



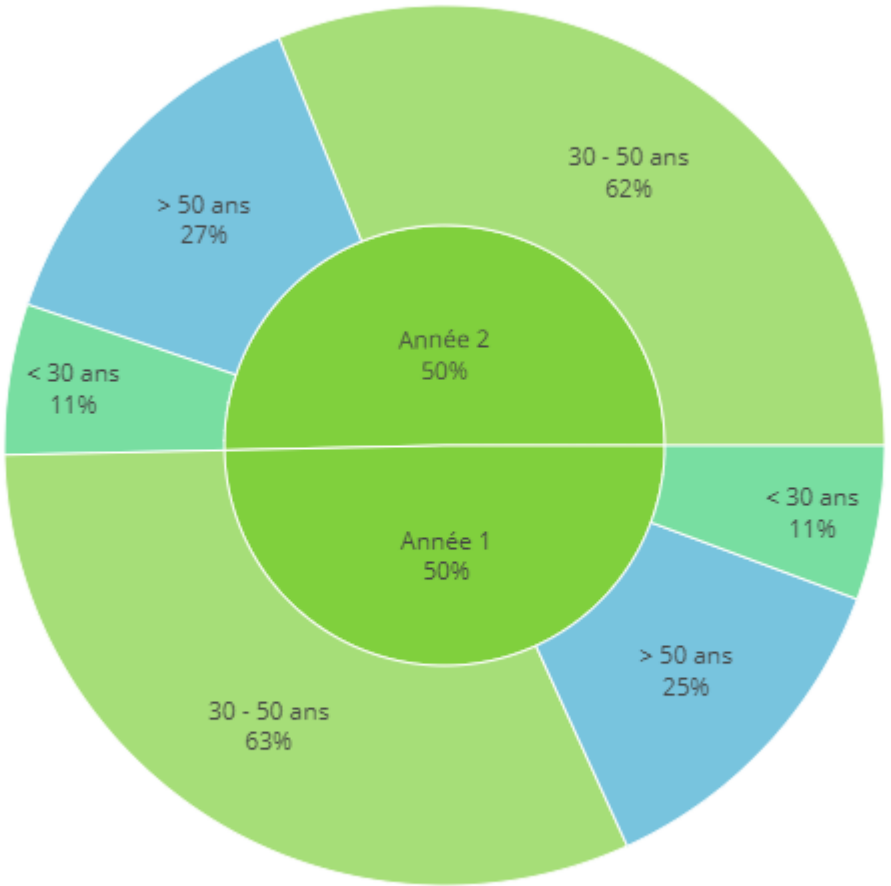


Chiffre d'affaires et catégories d'âge

Répartition du chiffre d'affaires par tranche d'age (Année 1-2)



Répartition par tranche d'age (Année 1-2)



Chiffre d'affaires et genre

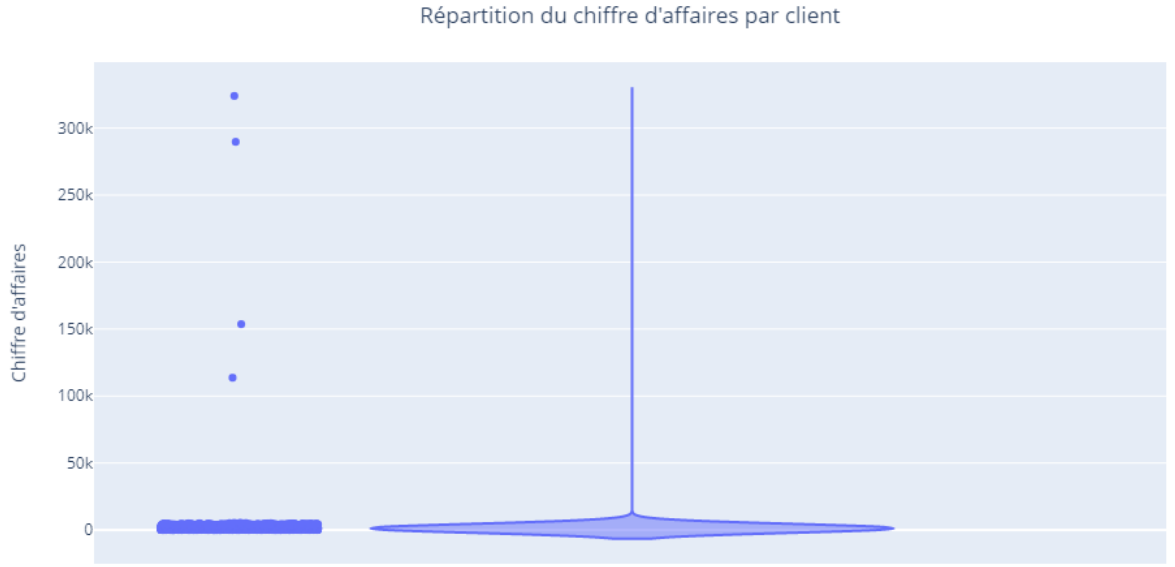
Répartition du chiffre d'affaire par genre (Année 1-2)

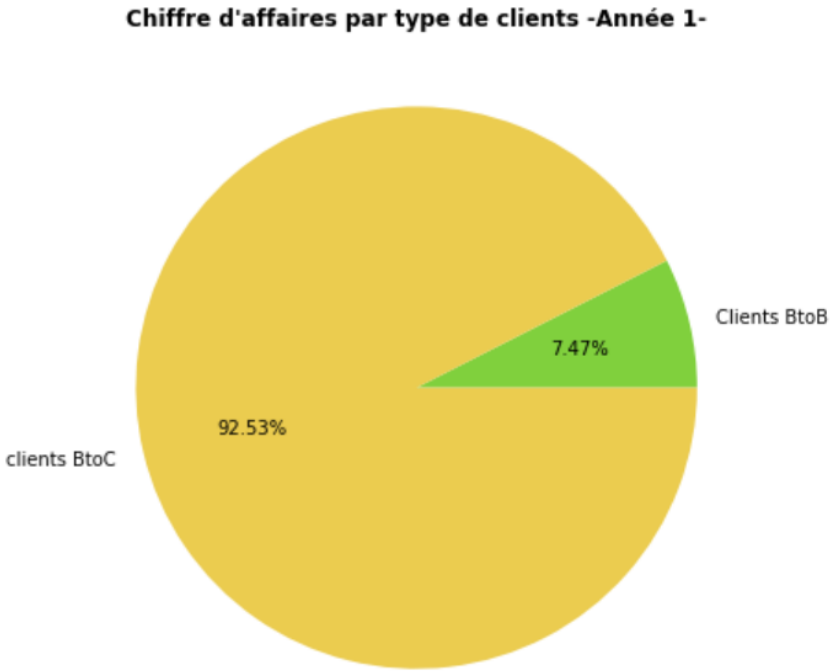
Répartition par genre (Année 1-2)

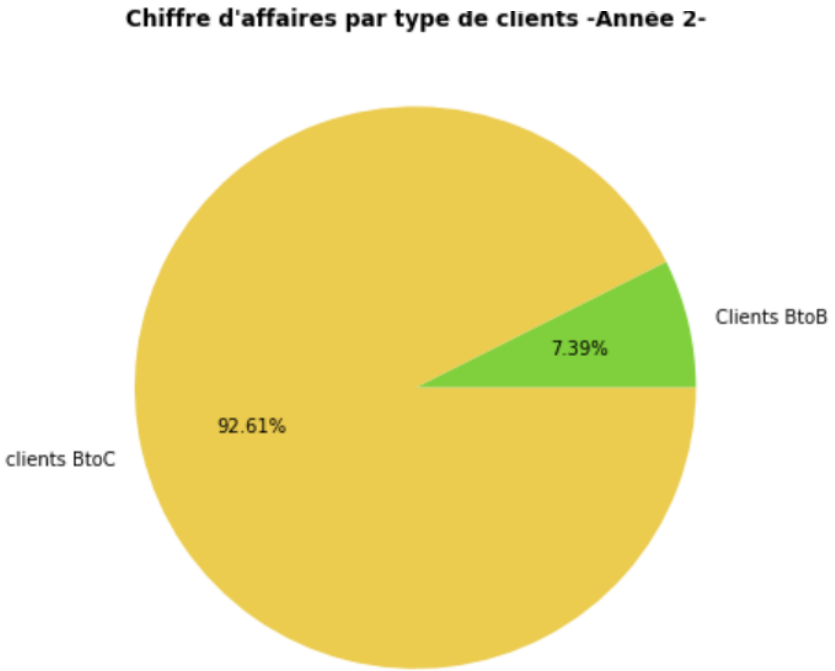
Profil des clients

Quatre clients identifiés (c_1609, c_4958, c_6714, c_3454) se distinguent par leur grande contribution au chiffre d'affaires généré.

On peut supposer qu'ils correspondent à des revendeurs (clients BtoB).

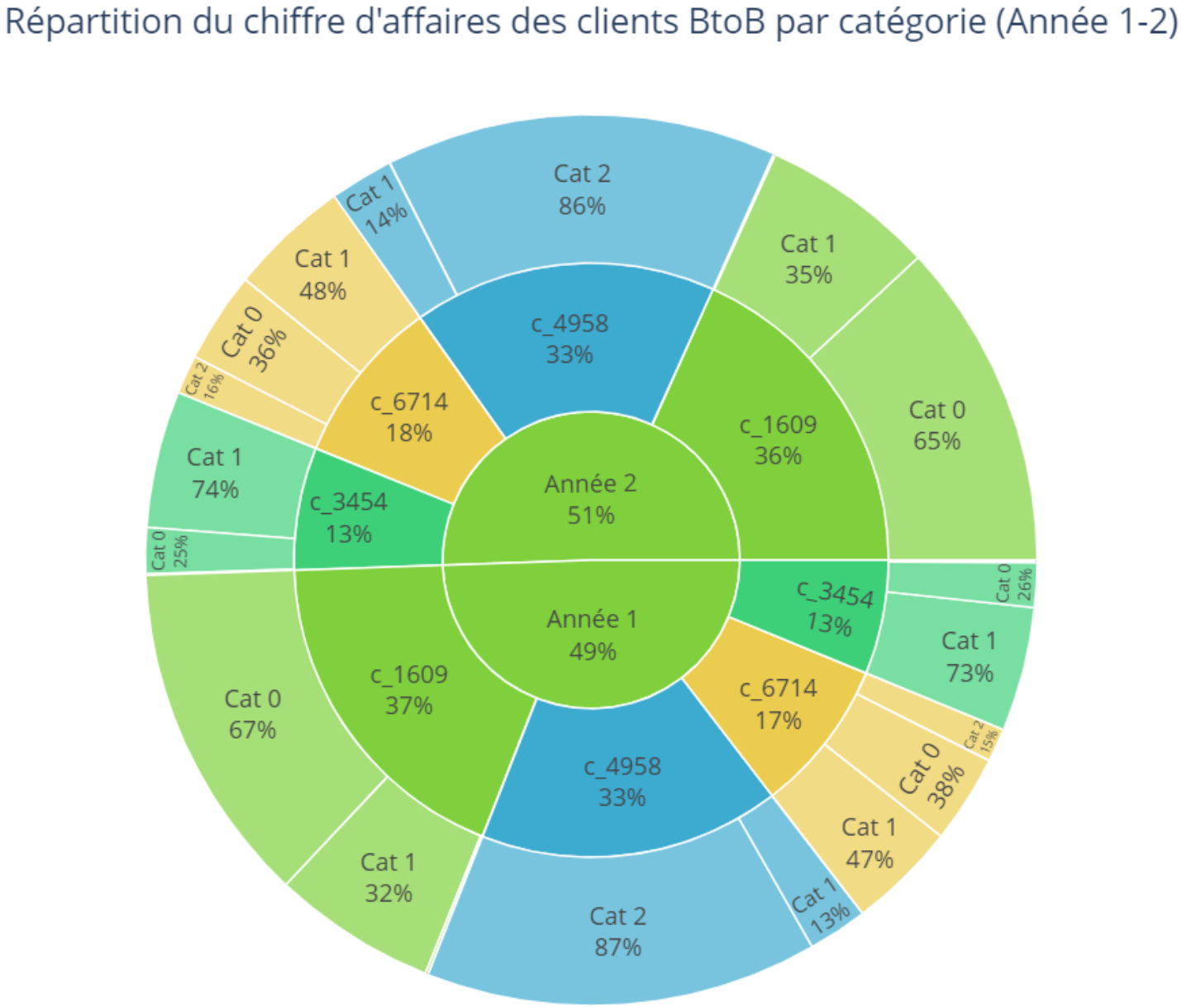




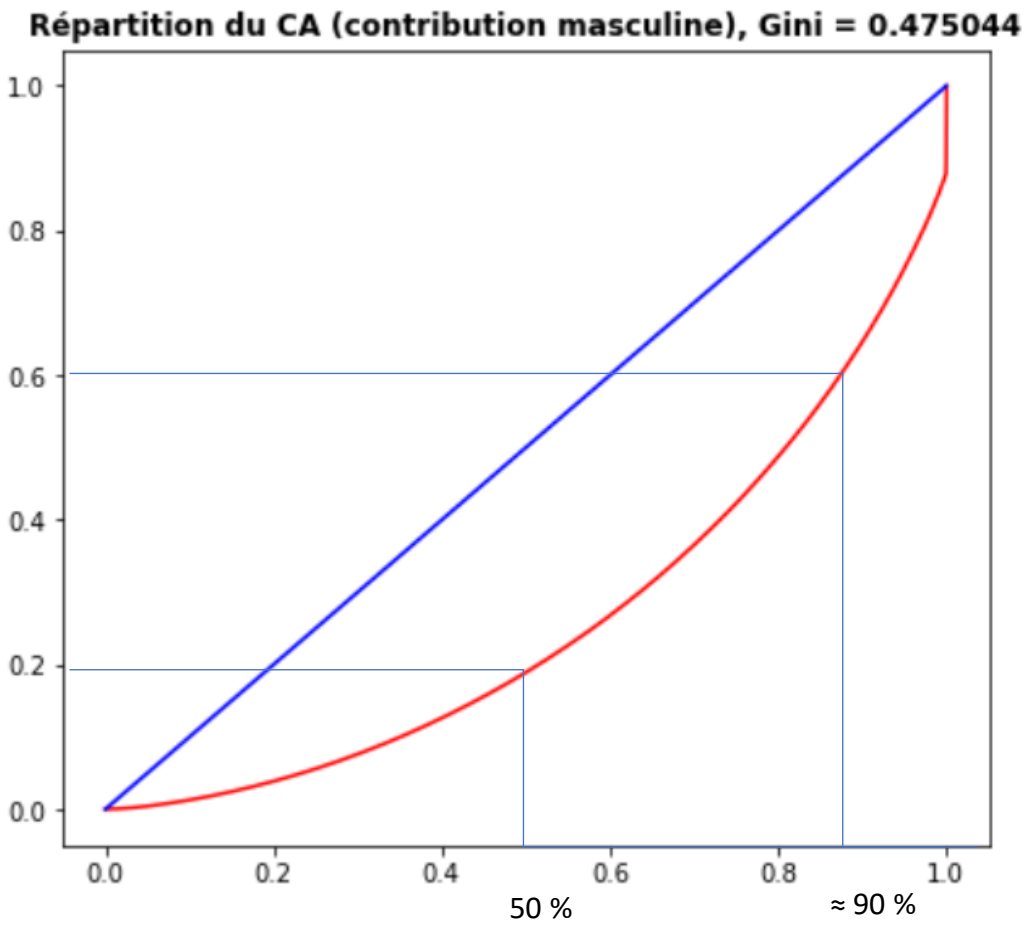


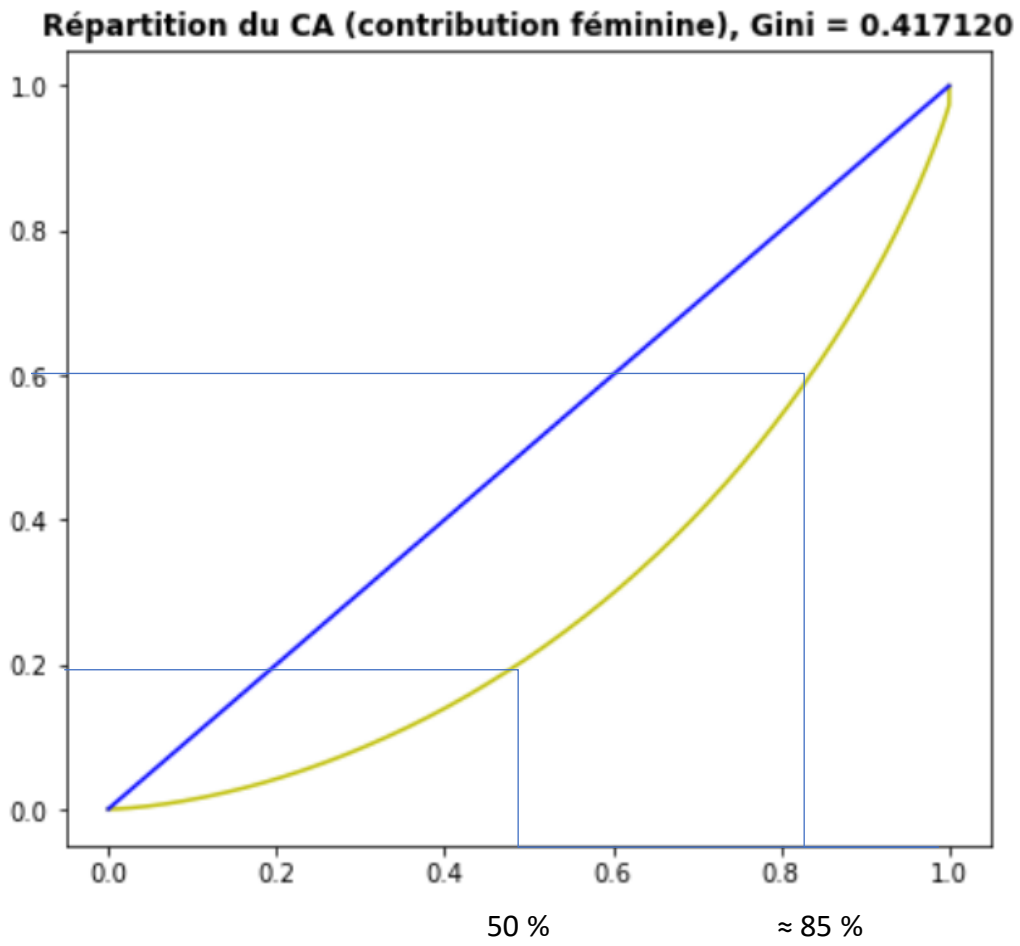
Les clients BtoB contribuent annuellement à 7% du chiffre d'affaires total.

Répartition du chiffre d'affaires entre clients BtoB



Contribution du genre dans la création du chiffre d'affaires



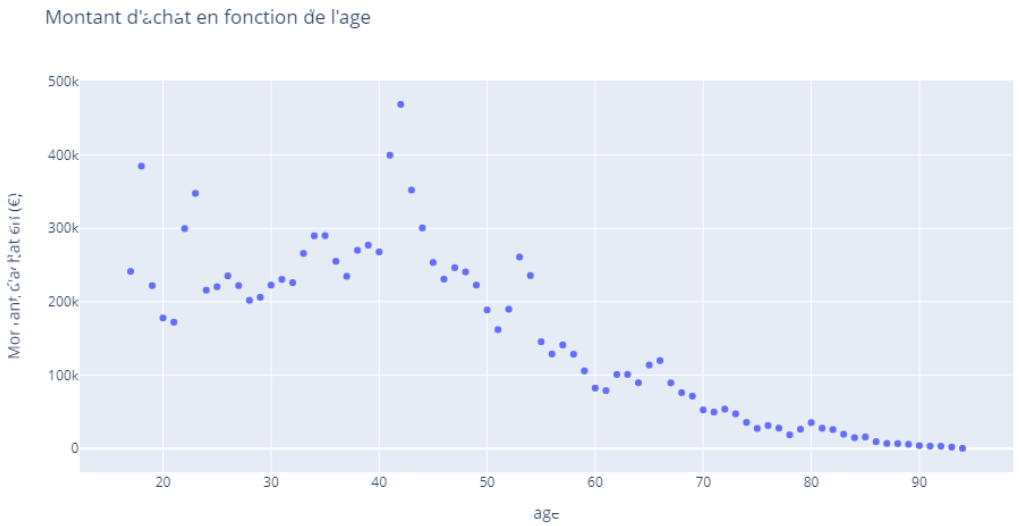


Les femmes et les hommes contribuent d'une manière moyennement équitable dans la création du chiffre d'affaires total.

Relations entre variables

Le lien entre, respectivement, le chiffre d'affaires et l'âge et le nombre d'achats et l'âge est partiellement linéaire. Ces données sont quantitatives, nous allons utiliser ici les tests de corrélation de Pearson et de Spearman.

Lien entre le chiffre d'affaire et l'âge des clients :



Posons les hypothèses de départ :
H0 : Variables indépendantes si p-value > 5%
H1 : Variables non indépendantes si p-value < 5%
Résultats des tests :

Pearson

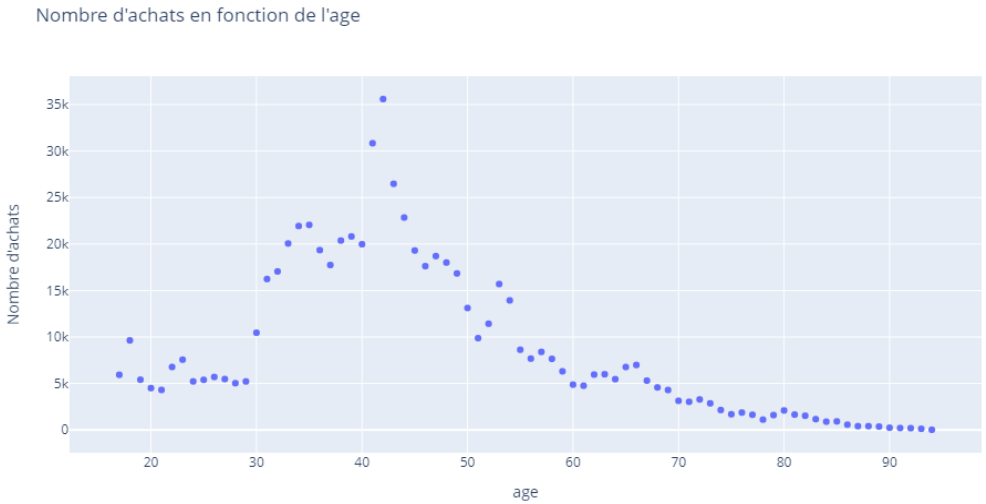
```
Corrélation = -0.856799660216099
P-value =  1.459407662056002 e-23 << 0.05
```

Spearman

```
Corrélation =-0.8720267074697454
p-value = 2.730425001740127 e-25 << 0.05
```

Le chiffre d'affaire est fortement corrélé à l'âge des clients, cependant cette corrélation est négative.

Lien entre le nombre d'achats et l'âge des clients :



Posons les hypothèses de départ :
H0 : Variables indépendantes si p-value > 5%
H1 : Variables non indépendantes si p-value < 5%
Résultats des tests :

Pearson

```
Corrélation = -0.5555773708371885
P-value = 1.2935855004108499 e-07 << 0.05
```

Spearman

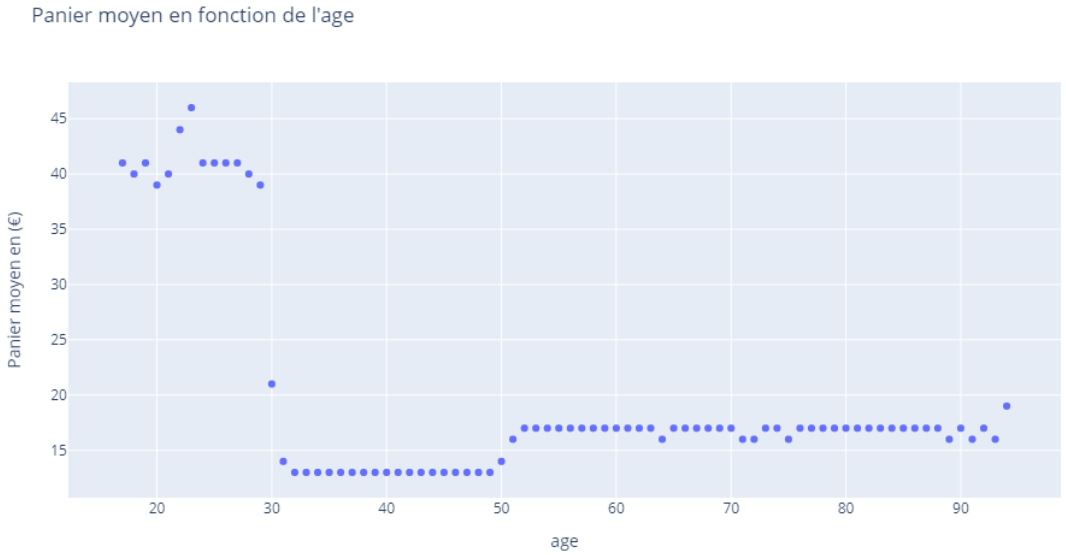
```
Corrélation = - 0.7059143388257313
p-value = 5.275863065515789 e-13 << 0.05
```

Le nombre de vente est moyennement corrélé à l'âge des clients, cependant cette corrélation est négative.

Lien entre le panier moyen et l'âge des clients :

Il est évident que le panier moyen est en lien avec des tranches d'âge bien définies. Il est de :

- Autour de 40 € pour les moins de 30 ans
- Autour de 13 € pour les plus de 30 ans et les moins de 50 ans
- Autour de 17 € pour les plus de 50 ans



Lien entre l'âge des clients et la catégorie de livres achetés :

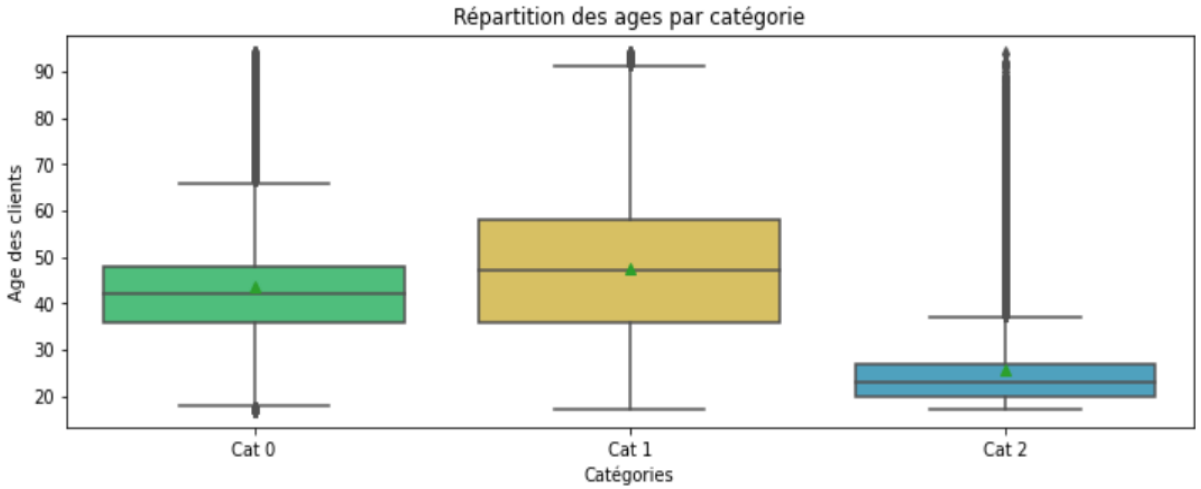
On va utiliser le test d'analyse de variance type ANOVA entre une variable qualitative "catégorie de livres" et quantitative « âge des clients » mais avant nous devons vérifier les hypothèses paramétriques de ce test :

- ✓ L'indépendance des données
- ✓ L'homoscédasticité ou l'égalité des variances
- ✓ La normalité des données

Test d'homoscédasticité :

Le test de Bartlett permet de tester si les variances sont significativement différentes ou non. Posons les hypothèses de départ :

H0 : Les variances de chaque groupe sont égales si p-value > 5%
H1 : Les variances de chaque groupe ne sont pas toutes égales si p-value < 5%




```
La statistique de bartlett est égale à 36527.06257330086
La p-value de bartlett est égale à 0.0 < 0.05
```

On rejette donc l'hypothèse nulle car la p-value est inférieure à 0.05. On peut dire que les variances de chaque groupe ne sont pas toutes égales.

Test de normalité :

Le test de Smirnov-Kermogolov permet de la normalité des données. Posons les hypothèses de départ :

H0 : La distribution suit une loi normale si p-value > 5%
H1 : La distribution ne suit pas une loi normale si p-value < 5%

```
La statistique de Smirnov-Kermogolov  est égale à 1.0
la p-value de Smirnov-Kermogolov est égale à 0
```

On rejette l'hypothèse nulle car la p-value est inférieure à 0.05. On peut dire que la distribution ne suit pas une loi normale.

Les hypothèses du test ANOVA ne sont pas réunies. Nous allons donc utiliser le test non paramétrique de Kruskal-Wallis.

Test non paramétrique de Kruskal-Wallis

Posons les hypothèses de départ :

H0 : Les trois catégories ne différent pas pour l'age si p-value > 5%
H1 : Au moins une catégorie différe des autres pour l'age si p-value < 5%

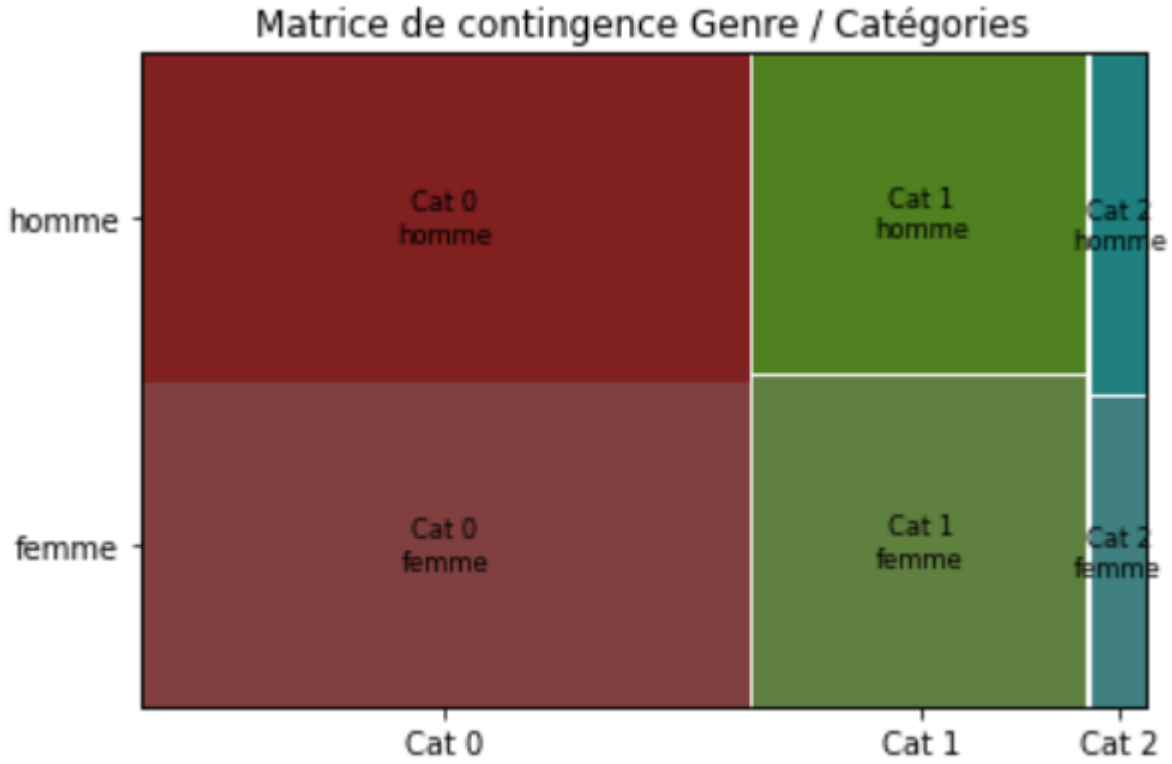
```
La statistique de Kruskal-Wallis est égale à 79491.9784
La p-value de Kruskal-Wallis est égale à 0.0
```

On rejette l'hypothèse nulle car la p-value est inférieure à 0.05. On peut donc supposer qu'au moins une des 3 catégories diffère des autres pour l'âge des clients.

Lien entre le genre d'un client et les catégories des livres achetés

Il s'agit ici de deux variables qualitatives "sexe du client" et "catégorie de livre". On utilisera le test d'indépendance du khi-deux. Posons les hypothèses de départ :

H0 : Le sexe des clients et la catégorie de livre sont indépendants si p-value > 5%
H1 : Le sexe des clients et la catégorie de livre sont non indépendants si p-value < 5%




```
La statistique Khi-2 est égale à 146.99906487909777
La p-value de Khi-2 est égale à 1.2010432285664067e-32 << 0.05
```

On rejette donc l'hypothèse nulle car p-value est inférieure au seuil fixé de 5% . On peut dire que le sexe des clients et la catégorie de livre sont non indépendants.