



Analysez les ventes d'une librairie avec R

Mohamad ALI



L'entreprise Lapage était originellement une librairie physique avec plusieurs points de vente. Mais devant le succès de certains de ses produits et l'engouement de ses clients, elle a décidé depuis 2 ans d'ouvrir un site de vente en ligne.

Notre mission; faire le point sur les différents indicateurs et chiffres clés de l'entreprise par

1. Une analyse des différents indicateurs de vente.
2. Une analyse plus ciblée sur les clients : l'objectif est de comprendre le comportement de nos clients en ligne, pour pouvoir ensuite comparer avec la connaissance acquise via nos librairies physiques.

Descriptions des données sources

- **Les Produits**

id_prod	price	categ
<chr>	<dbl>	<int>
0_1421	19.99	0
0_1368	5.13	0
0_731	17.99	0

- 3287 produits en 3 catégories :
- Categ 0 = 2039 (70%)
- Categ 1 = 739 (%)
- Categ 2 = 239 (%)
- prix plus élevé : 300€
- prix le plus bas : 0,62 €
- prix moyen : 17,45 €

- **Les Clients**

client_id	sex	birth
<chr>	<chr>	<int>
c_4410	f	1967
c_7839	f	1975
c_1699	f	1984

- 8623 Clients dont :
- 52% Femme
- 48% Homme
- Le plus âgé 94 ans
- Le plus jeune 19 ans

- **Les Transactions**

id_prod	date	session_id	client_id
<chr>	<chr>	<chr>	<chr>
0_1518	2022-05-20 13:21:29.043970	s_211425	c_103
1_251	2022-02-02 07:55:19.149409	s_158752	c_8534
0_1277	2022-06-18 15:44:33.155329	s_225667	c_6714

- 680 000 Transactions

Fusionner les données & Traiter les erreurs

1er : Fusionner Clients et Transactions par variable : client_id

client_id	sex	birth	id_prod		date	session_id
<chr>	<chr>	<int>	<chr>		<chr>	<chr>
c_1	m	1955	0_1470	2021-06-11 21:02:39.382765	s_47346	
c_1	m	1955	0_1429	2021-10-15 11:28:24.523566	s_105105	
c_1	m	1955	1_364	2021-12-15 23:32:41.632729	s_134971	

2eme: Fusionner Transactions_Clients et Produits par variable : id_prod

id_prod	price	categ	client_id	sex	birth		date	session_id
<chr>	<dbl>	<int>	<chr>	<chr>	<int>		<chr>	<chr>
0_0	3.75	0	c_1004	m	1973	2021-03-02 21:57:33.862118	s_908	
0_0	3.75	0	c_1011	f	1999	2022-03-18 16:40:10.068303	s_180968	
0_0	3.75	0	c_1011	f	1999	2022-02-18 16:40:10.068303	s_167174	

Valeurs manquantes NA

Remplacer 221 lignes de price par le prix moyen de categ 0

1. Supprimer 21 lignes de client_id (Clients inactifs)

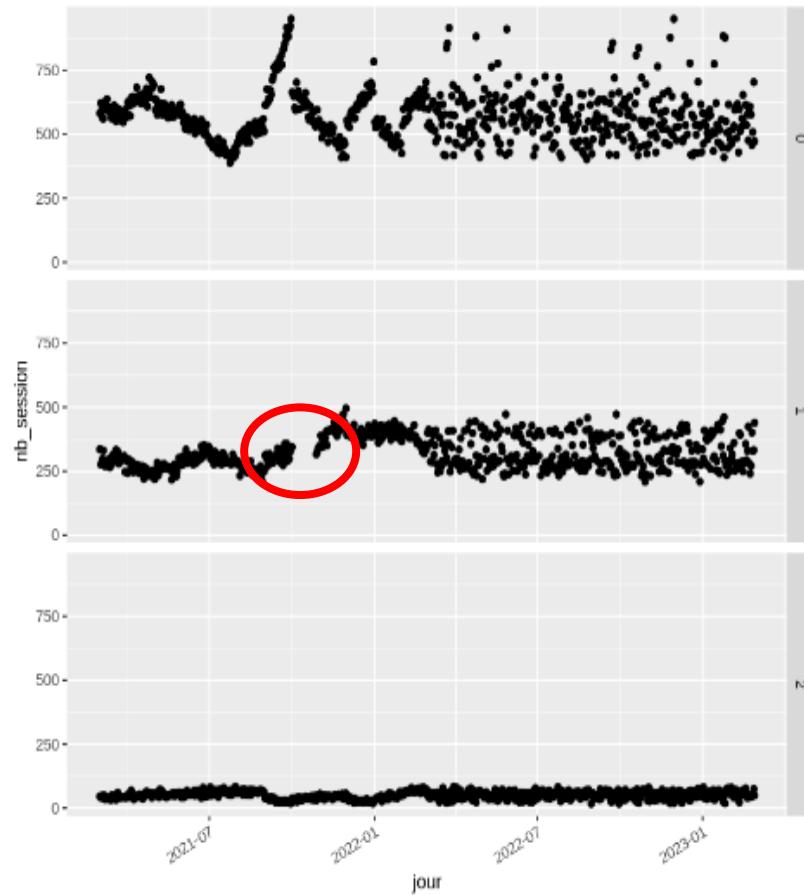
Valeurs <0

- Supprimer 221 lignes de price.

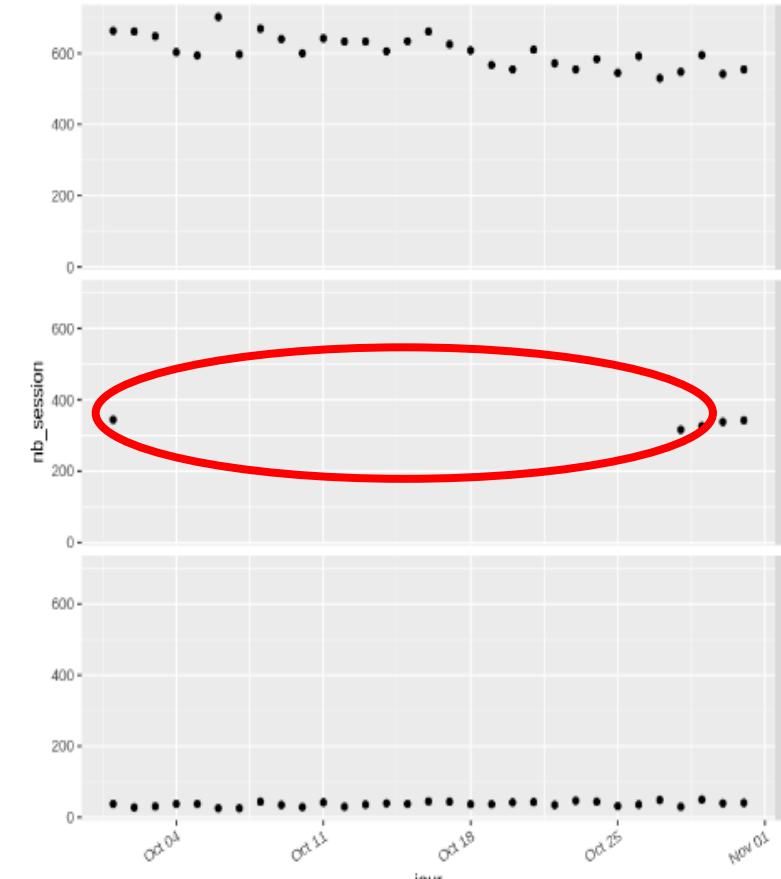
DéTECTER un trou dans les données

Un trou a été découvert en octobre (entre 2 et 27 octobre 2021) concernant les produits du catégories 1.

```
[1] "Le nombre total de jours de vente de catégorie 0 est : 730 ."  
[1] "Le nombre total de jours de vente de catégorie 1 est : 704 ."  
[1] "Le nombre total de jours de vente de catégorie 2 est : 730 ."  
session_j
```



```
[1] "Le nombre de jours de vente de catégorie 0 au mois d'octobre 2021 est : 31 ."  
[1] "Le nombre de jours de vente de catégorie 1 au mois d'octobre 2021 est : 5 ."  
[1] "Le nombre de jours de vente de catégorie 2 au mois d'octobre 2021 est : 31 ."  
categ
```



Analyse des données

1. Demandes d'Antoine

Zoom sur les références, les tops et les flops, la répartition par catégorie

Les tops 10 produits

<code>id_prod</code>	<code>categ</code>	<code>CA_prod</code>	<code>nbr_vente_tot</code>
<code><chr></code>	<code><int></code>	<code><dbl></code>	<code><int></code>
2_159	2	94893.50	649
2_135	2	69334.95	994
2_112	2	65407.76	960
2_102	2	60736.78	1025
2_209	2	56971.86	814
1_395	1	54356.25	1873
1_369	1	54025.48	2245
2_110	2	53846.25	865
2_39	2	53060.85	915
2_166	2	52449.12	228

Les flops 10 produits

<code>id_prod</code>	<code>categ</code>	<code>CA_prod</code>	<code>nbr_vente_tot</code>
<code><chr></code>	<code><int></code>	<code><dbl></code>	<code><int></code>
0_1539	0	0.99	1
0_1284	0	1.38	1
0_1653	0	1.98	2
0_1601	0	1.99	1
0_541	0	1.99	1
0_807	0	1.99	1
0_1728	0	2.27	1
0_1498	0	2.48	1
0_1539	0	0.99	1
0_1601	0	1.99	1
0_1633	0	24.99	1
0_1683	0	2.99	1
0_1728	0	2.27	1
0_1840	0	2.56	2
0_2201	0	20.99	1

Zoom sur les références, les tops et les flops, la répartition par catégorie

Les tops 10 clients du chiffre d'affaires

client_id	sex	sex_numeric	age	CA_client	frequence_achat	Taile_panier_moyenne	Per_CA_client
<chr>	<chr>	<dbl>	<int>	<dbl>	<int>	<dbl>	<dbl>
c_1609	m	1	43	324033.350	10997	29.46561	2.73305643
c_4958	m	1	24	289760.340	3851	75.24288	2.44398103
c_6714	f	0	55	153662.750	2620	58.64990	1.29606711
c_3454	m	1	54	113669.845	5573	20.39653	0.95874730
c_3263	f	0	38	5276.870	143	36.90119	0.04450771
c_1570	f	0	44	5271.620	158	33.36468	0.04446343
c_2899	f	0	29	5214.050	69	75.56594	0.04397786
c_2140	f	0	46	5208.820	147	35.43415	0.04393375
c_7319	f	0	49	5155.770	145	35.55703	0.04348630
c_8026	m	1	45	5093.218	146	34.88506	0.04295871

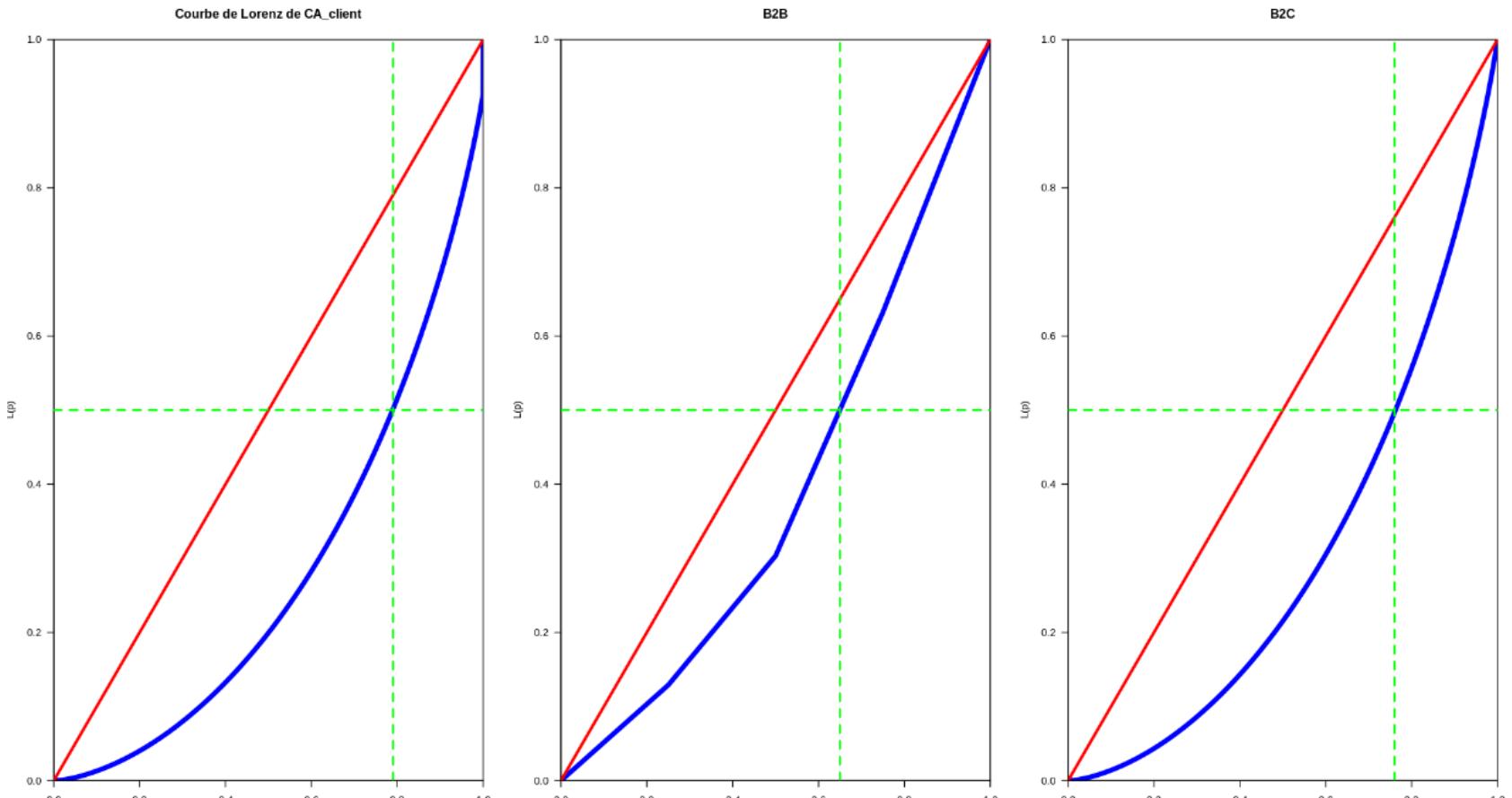
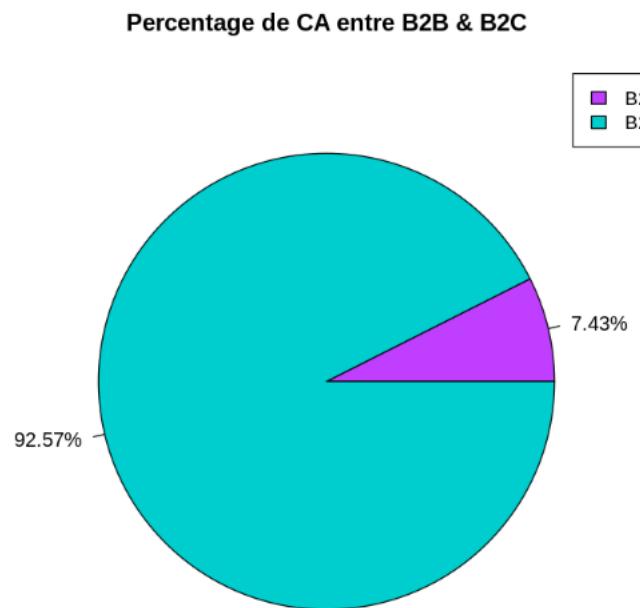
Les flops 10 clients du chiffre d'affaires

client_id	sex	sex_numeric	age	CA_client	frequence_achat	Taile_panier_moyenne	Per_CA_client
<chr>	<chr>	<dbl>	<int>	<dbl>	<int>	<dbl>	<dbl>
c_8351	f	0	55	6.31	1	6.31	5.322164e-05
c_8140	m	1	52	8.30	2	4.15	7.000628e-05
c_8114	m	1	61	9.98	2	4.99	8.417622e-05
c_240	m	1	25	11.06	1	11.06	9.328547e-05
c_4648	m	1	19	11.20	1	11.20	9.446630e-05
c_4478	f	0	53	13.36	1	13.36	1.126848e-04
c_5962	f	0	26	13.99	1	13.99	1.179985e-04
c_6040	f	0	49	15.72	1	15.72	1.325902e-04
c_5919	f	0	68	15.98	2	7.99	1.347832e-04
c_5829	f	0	34	16.07	1	16.07	1.355423e-04

Différents indicateurs et graphiques autour du chiffre d'affaires.

Le chiffre d'affaire totale est :

- 11 856 080 € dont :
- B2B est : 881 126 €
- B2C est : 10 974 953 €

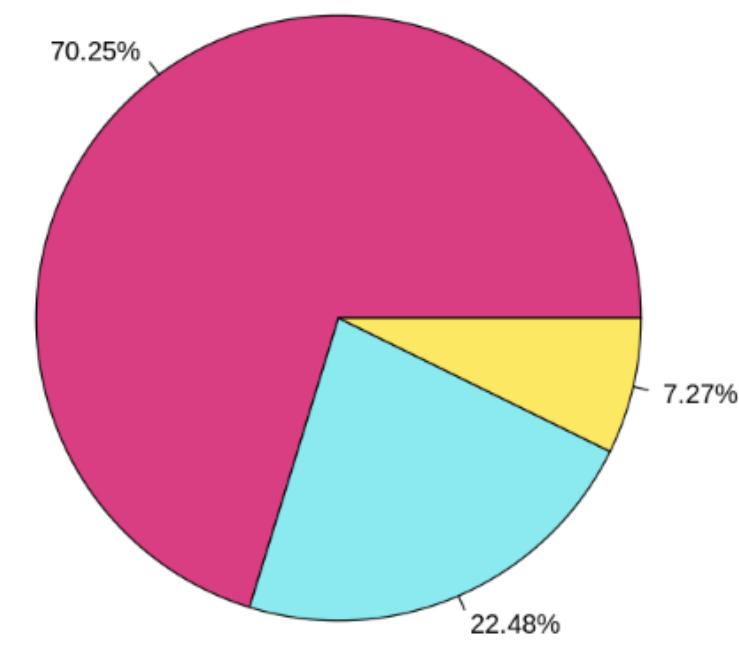


L'indice de Gini est :

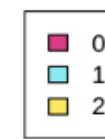
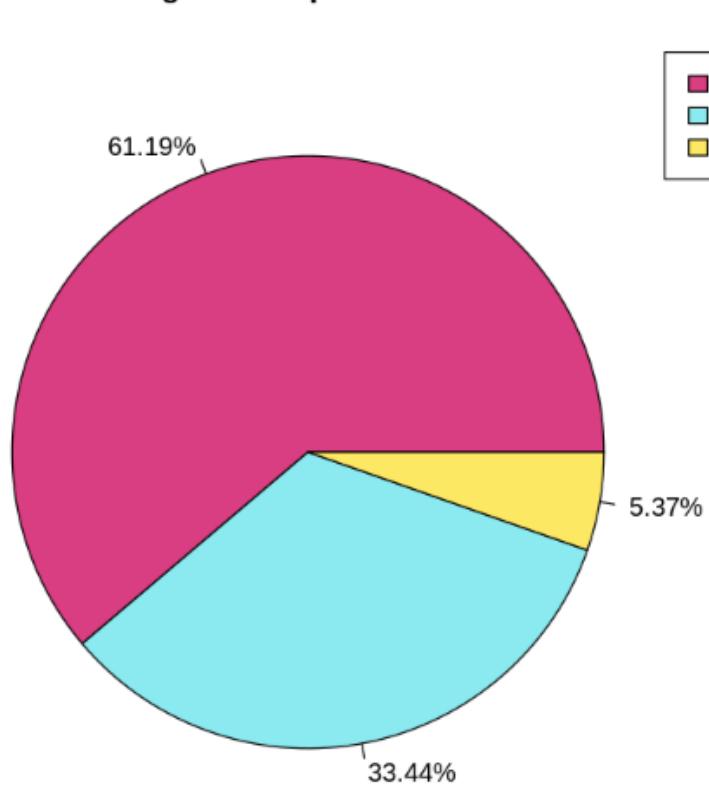
- Totale = 0.446 => la distribution des CA est relativement inégale.
- B2B = 0.218 => la distribution des CA est relativement équitable

Distribution des ventes par catégorie

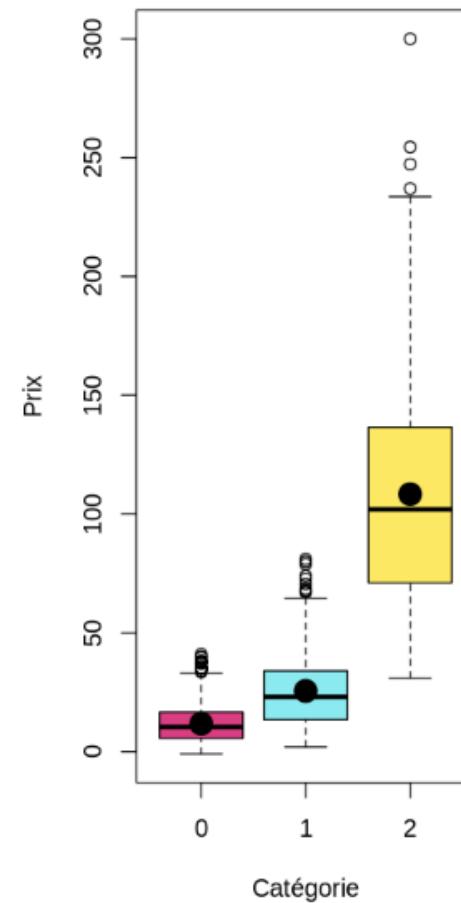
% catégories de produits disponibles



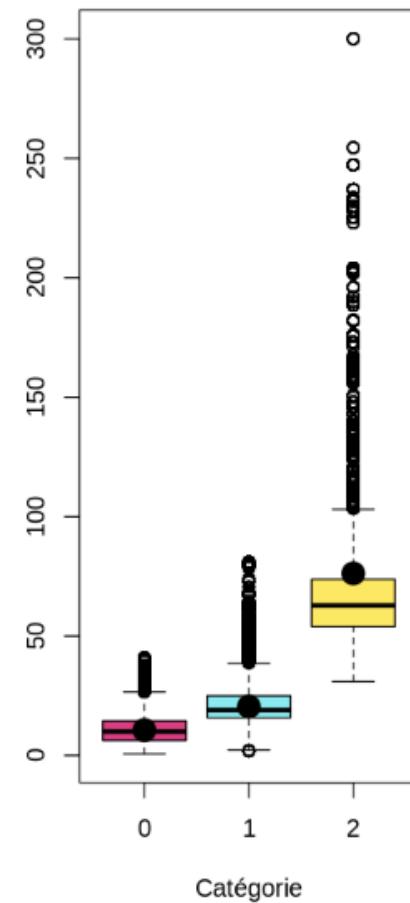
% catégories de produits vendus



prix prod disponibles & categ

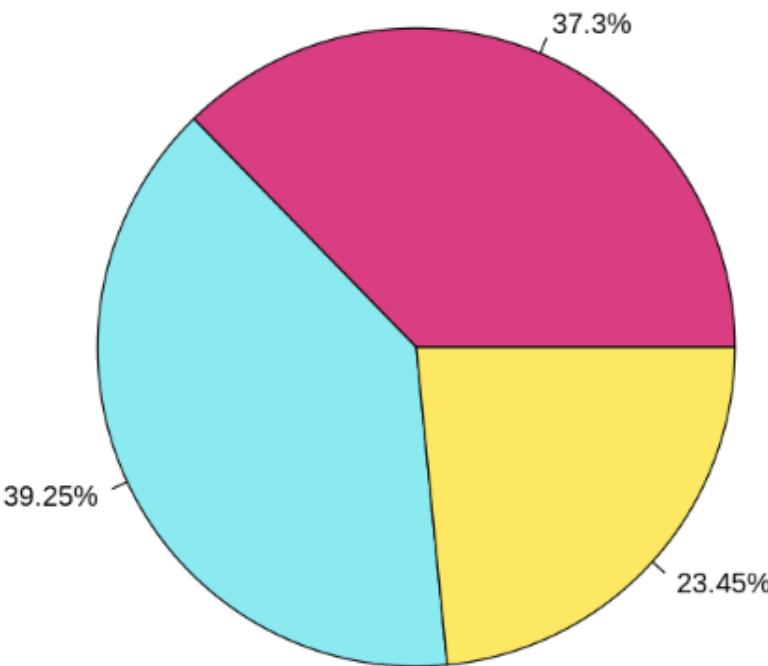


Prix prod vendus & categ

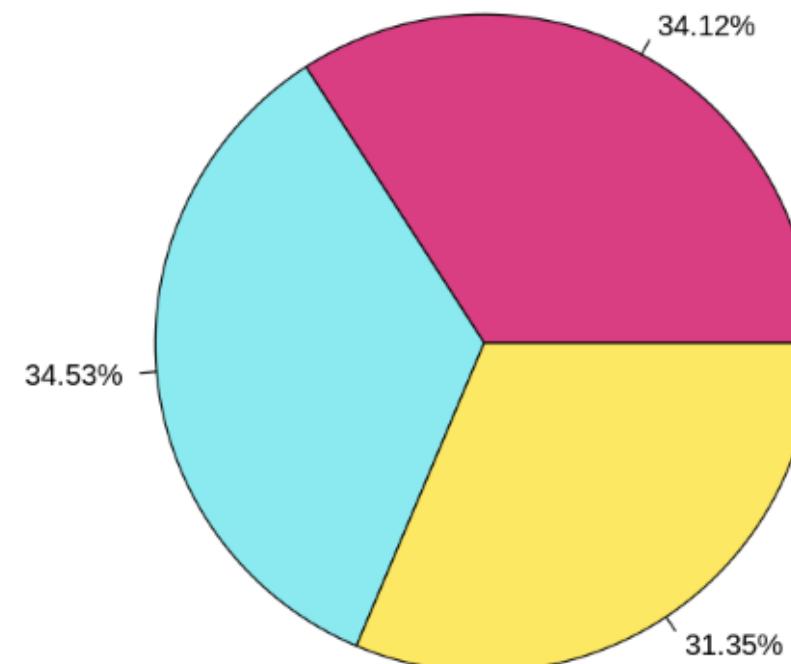


Différents indicateurs et graphiques autour du chiffre d'affaires.

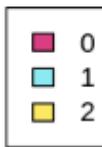
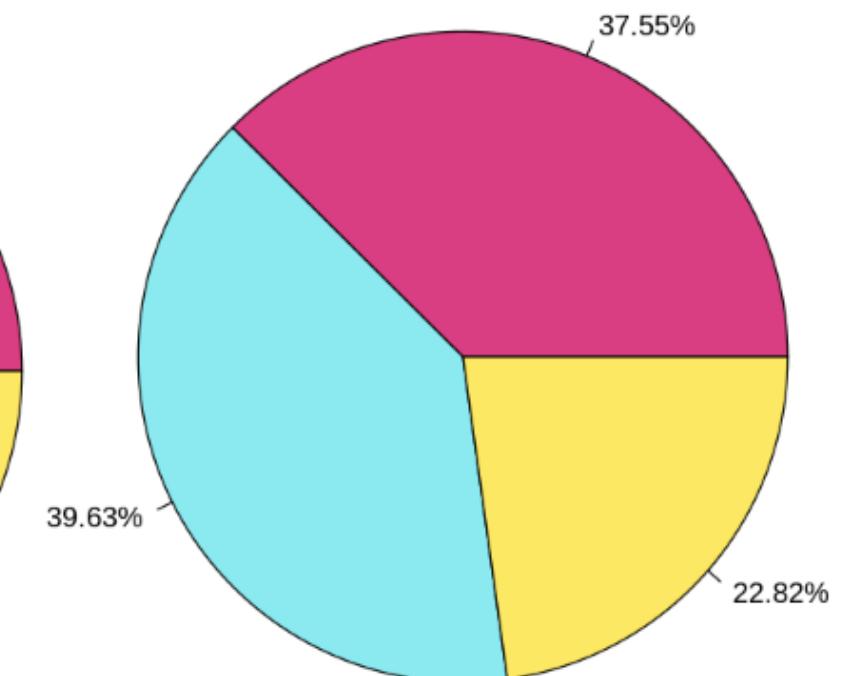
% CA par catégories



% CA par catégories pour B2B

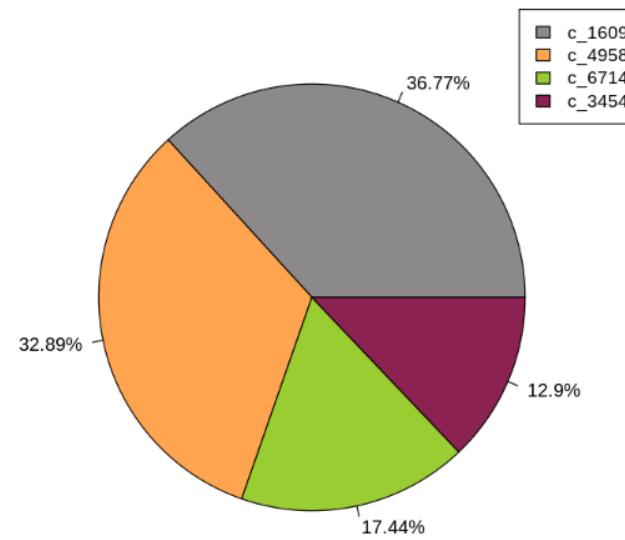


% CA par catégories B2C

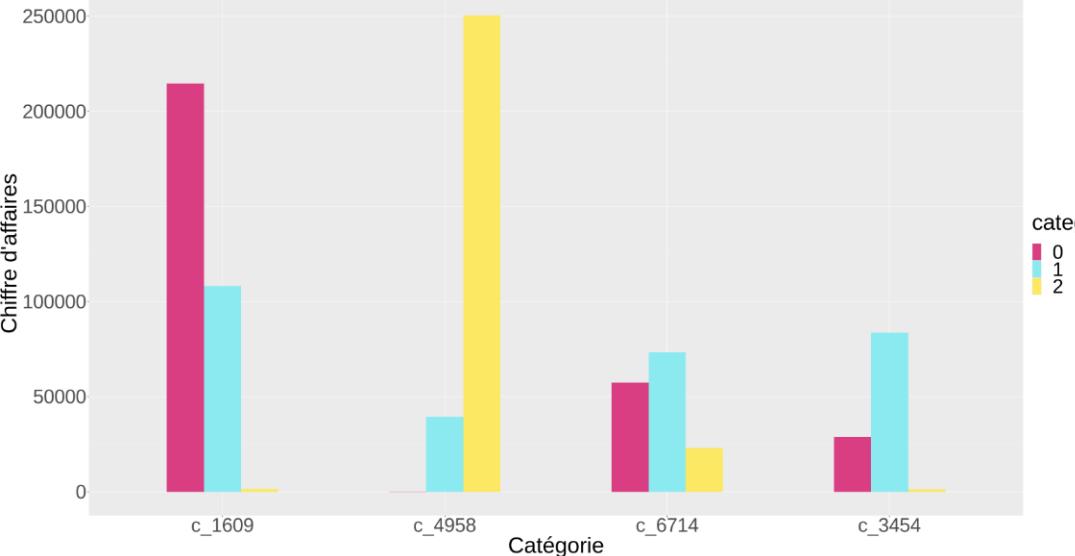


Différents indicateurs et graphiques autour du chiffre d'affaires_B2B

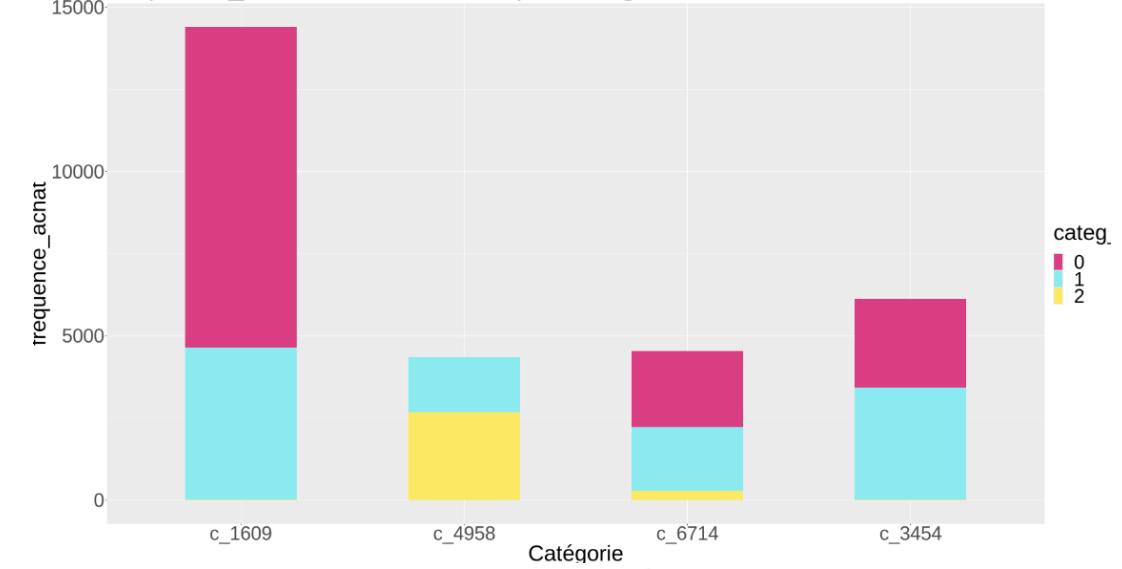
Percentage CA pour client B2B



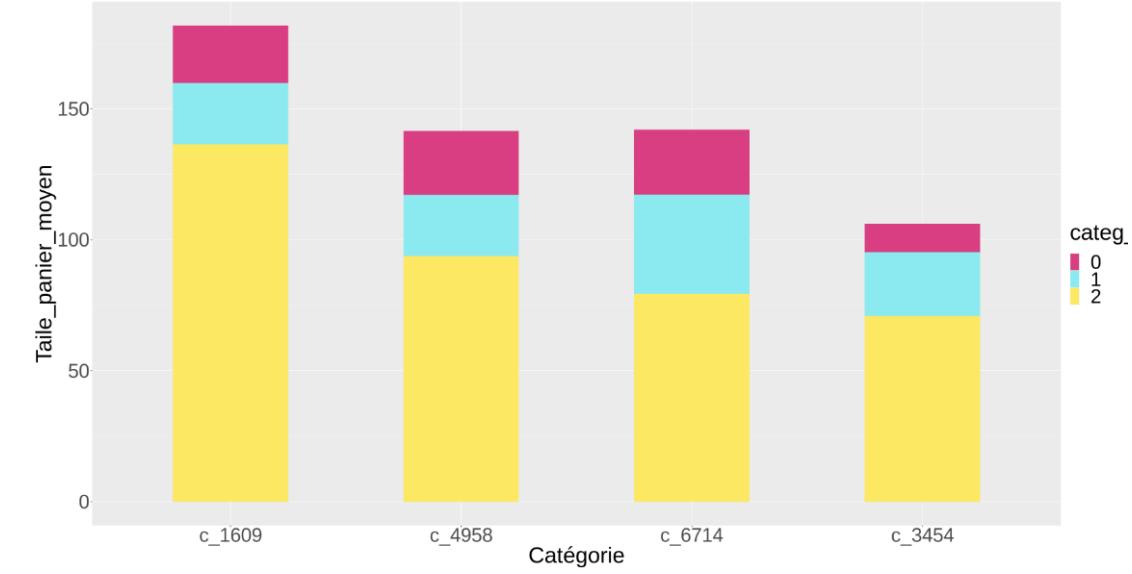
Chiffre d'affaires des clients B2B par Catégories



fréquence_achat des clients B2B par Catégories

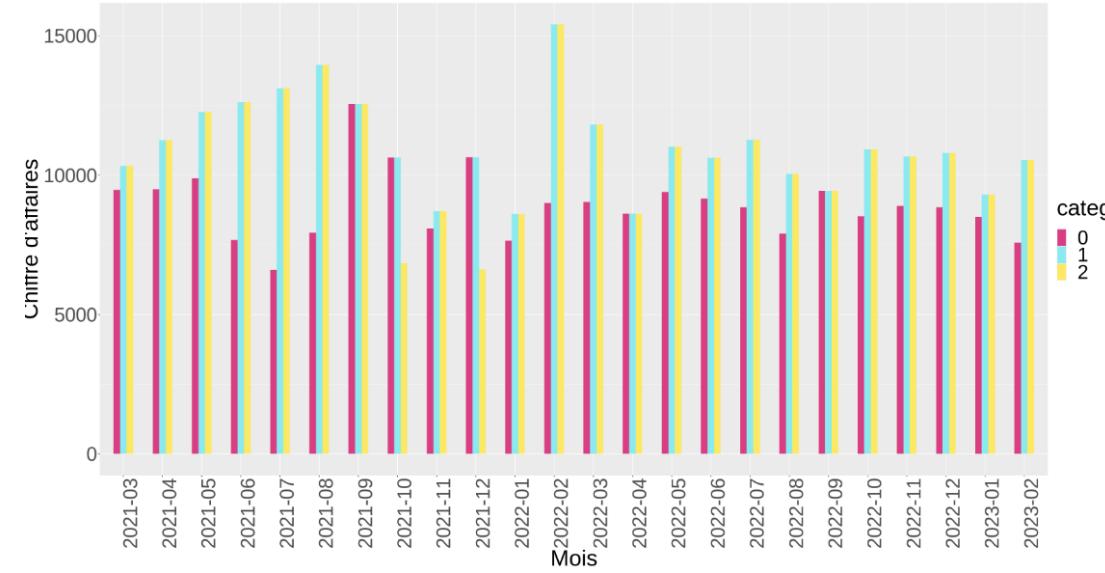


Taille_panier_moyen des clients B2B par Catégories

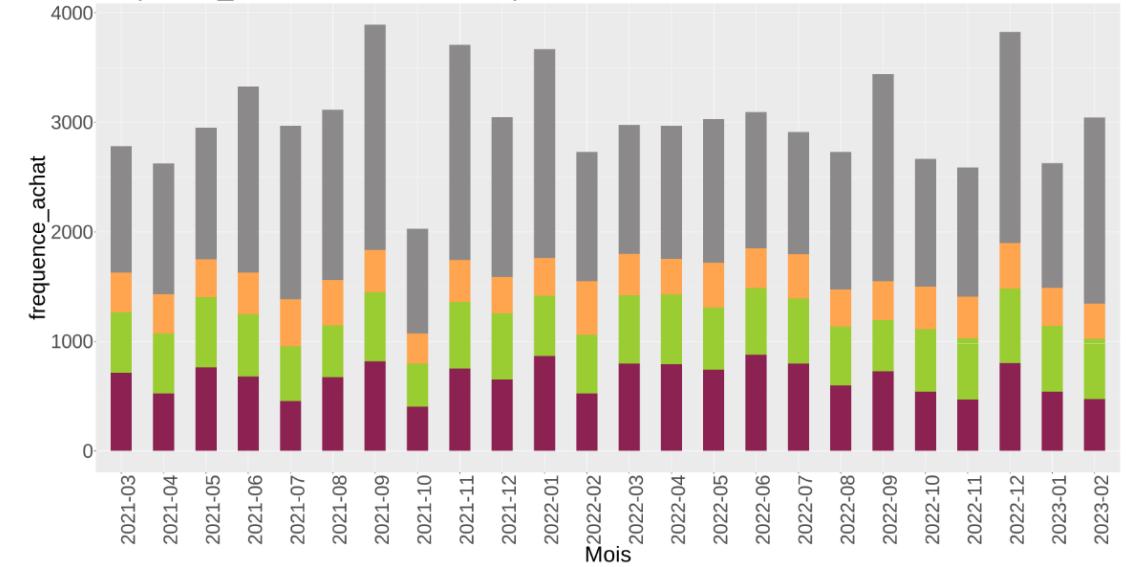


Différents indicateurs et graphiques autour du chiffre d'affaires_B2B

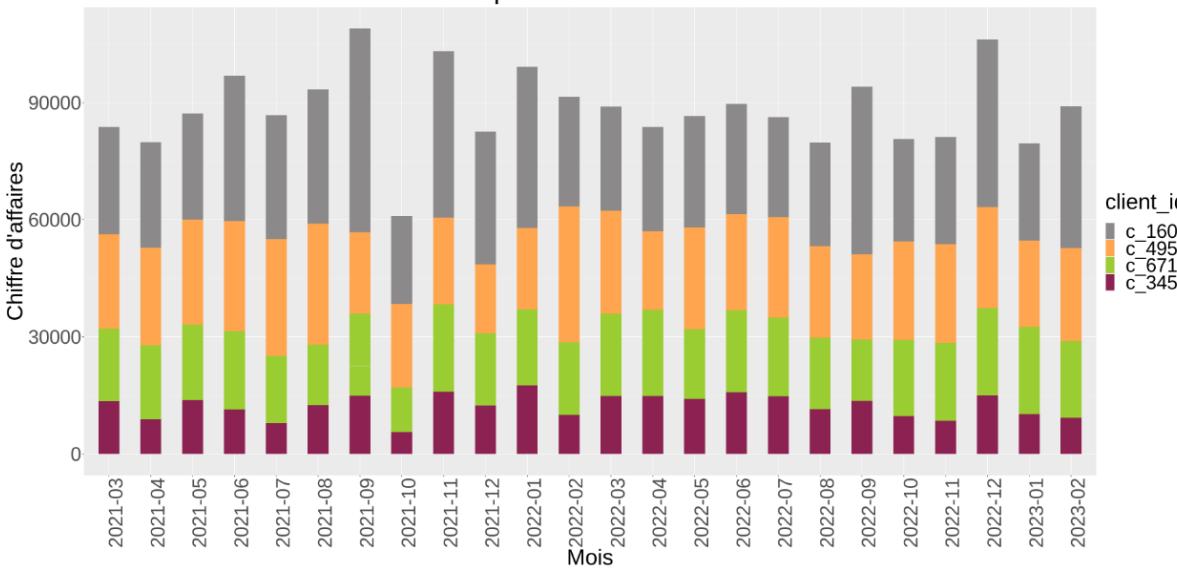
Chiffre d'affaires des Catégories B2B par mois



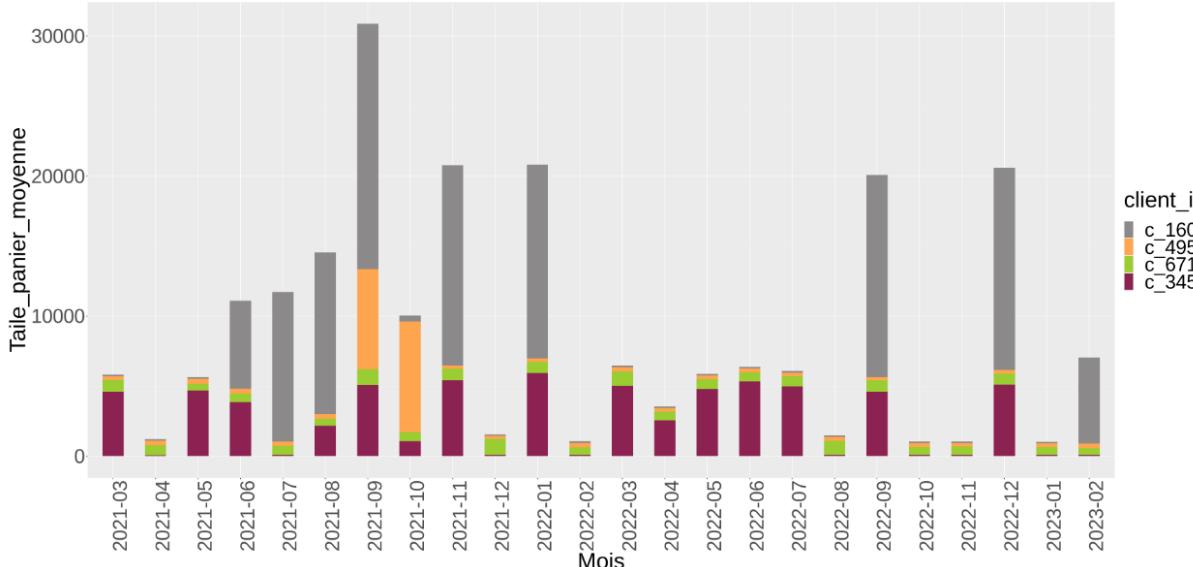
fréquence_achat des Clients B2B par mois



Chiffre d'affaires des Clients B2B par mois

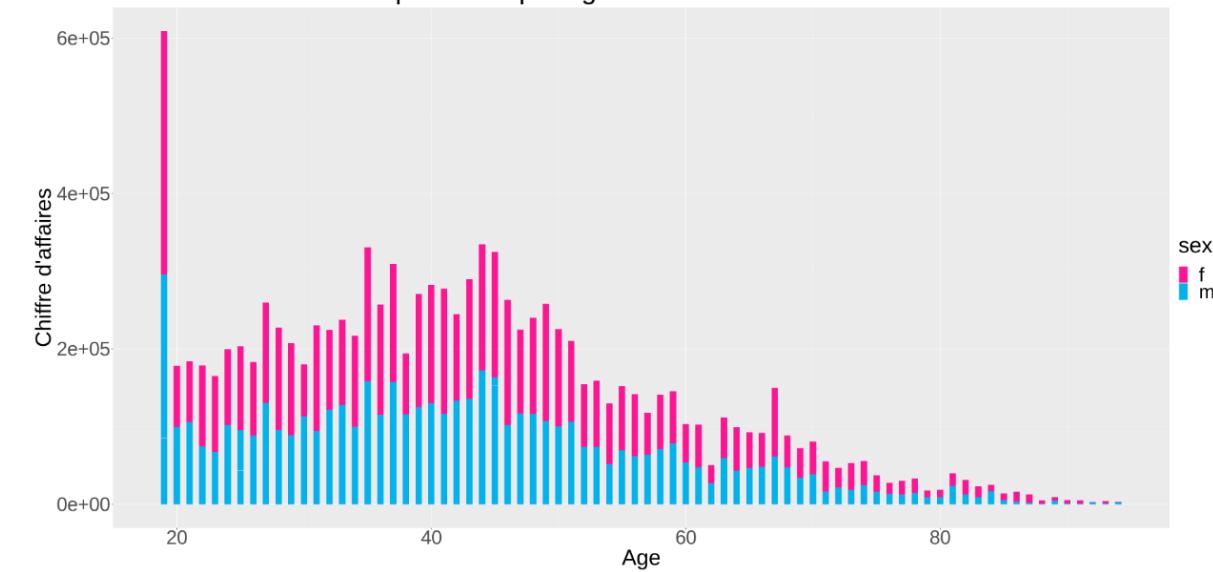


Taille_panier_moyenne des Clients B2B par mois

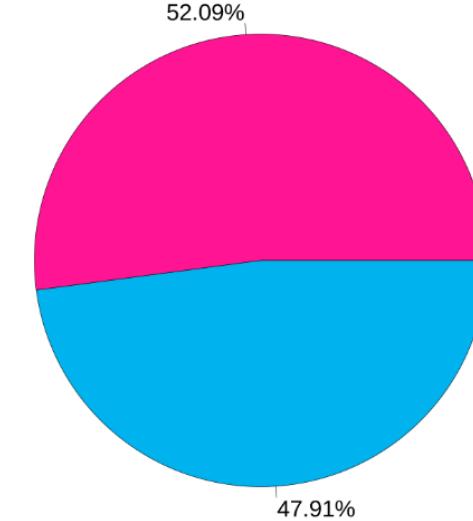


Différents indicateurs et graphiques autour du chiffre d'affaires _B2C

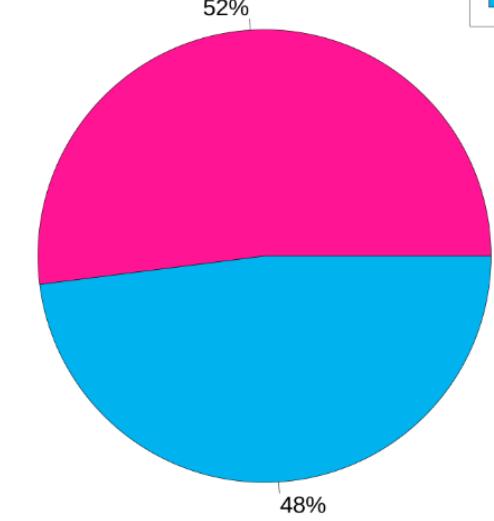
Chiffre d'affaires & sex pour B2C par age



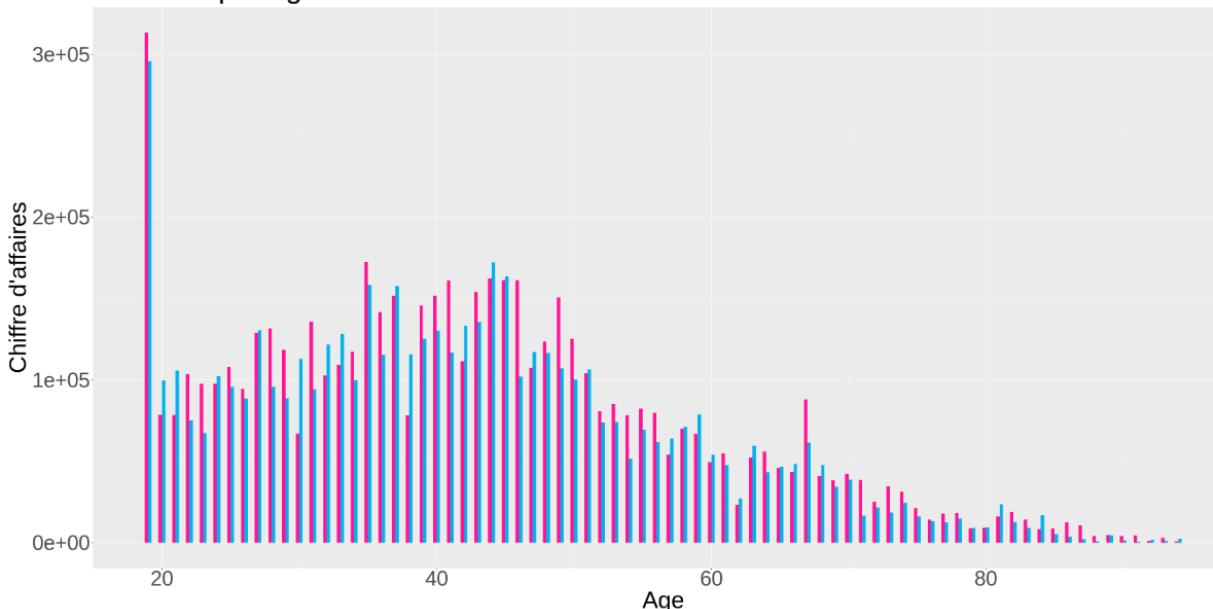
% Nombre de clients B2B



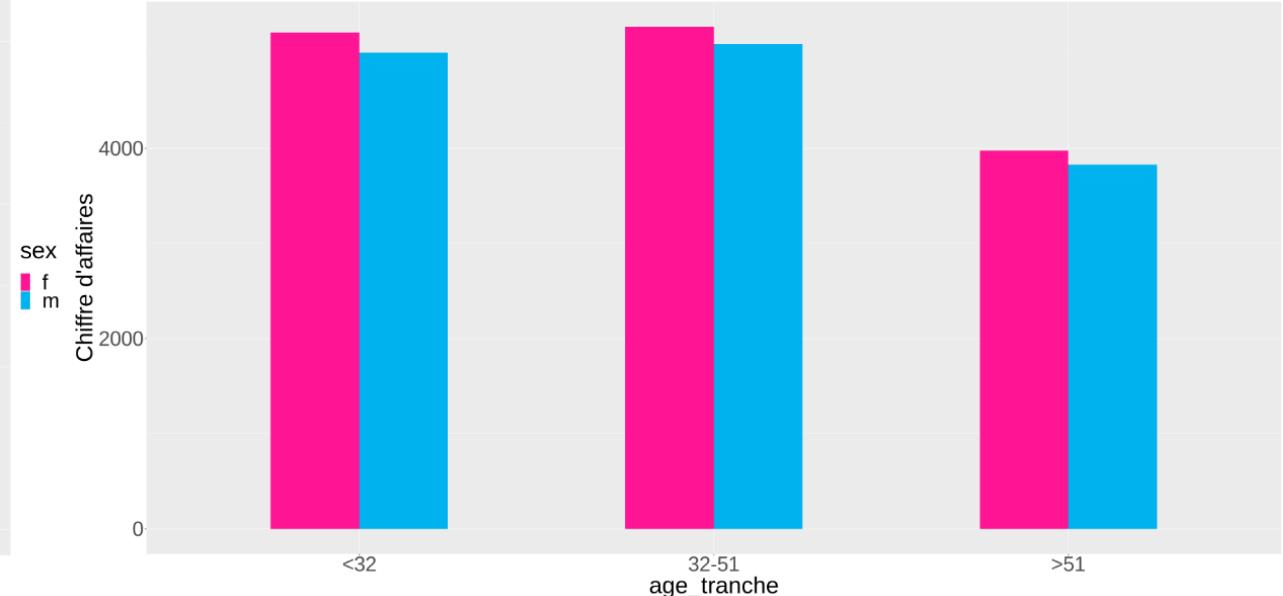
% CA selon le genre



Sex B2C par age

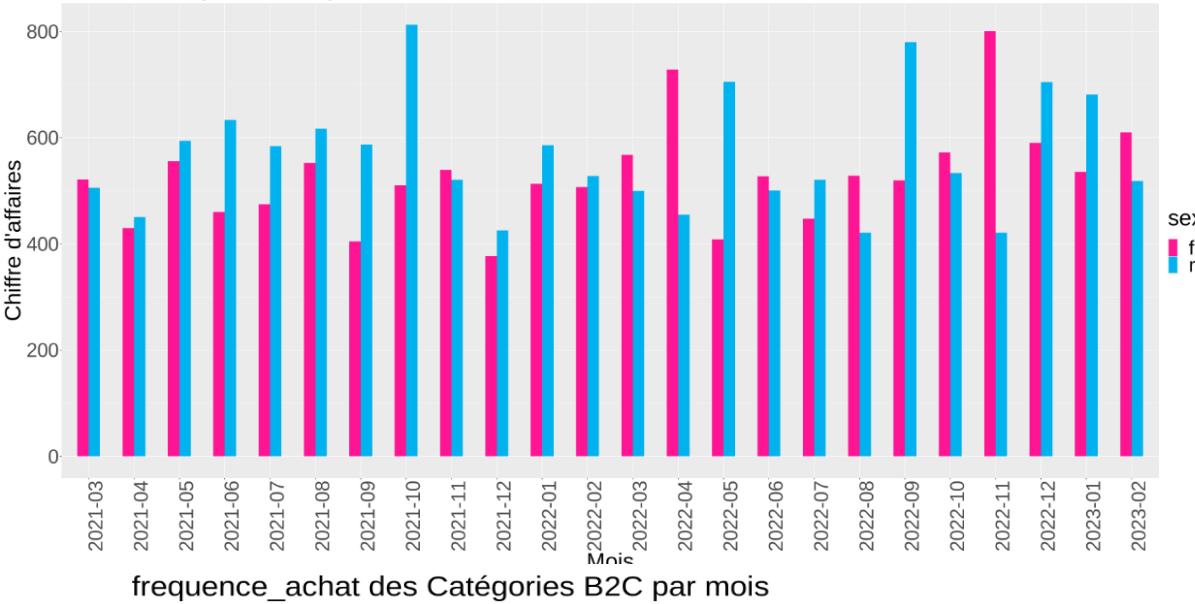


Chiffre d'affaires & sex pour B2C par age_tranche



Différents indicateurs et graphiques autour du chiffre d'affaires_B2C

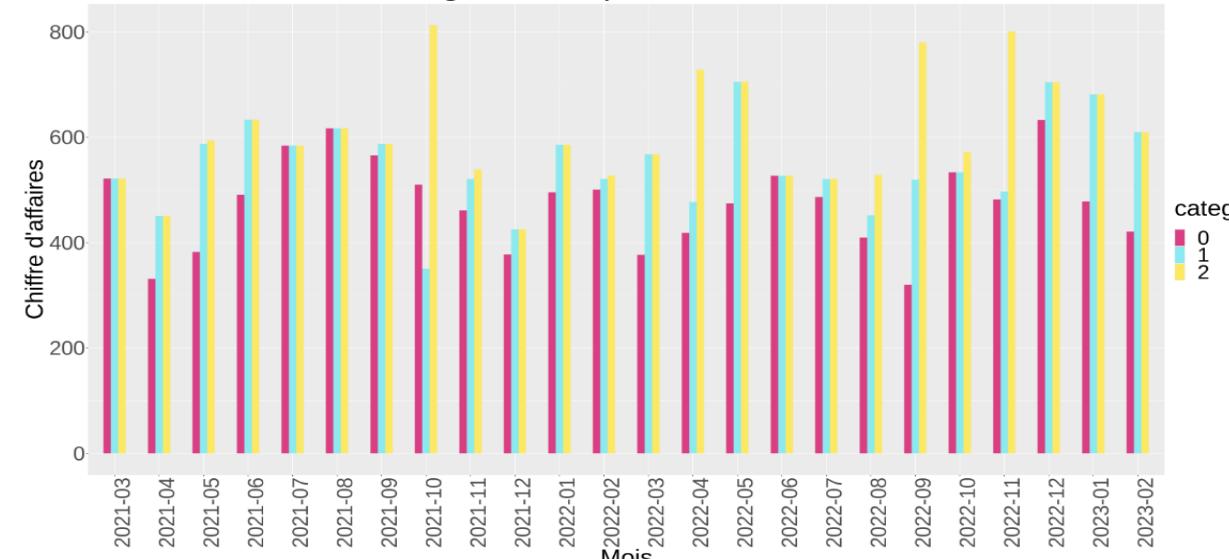
CA de Sex pour B2C par mois



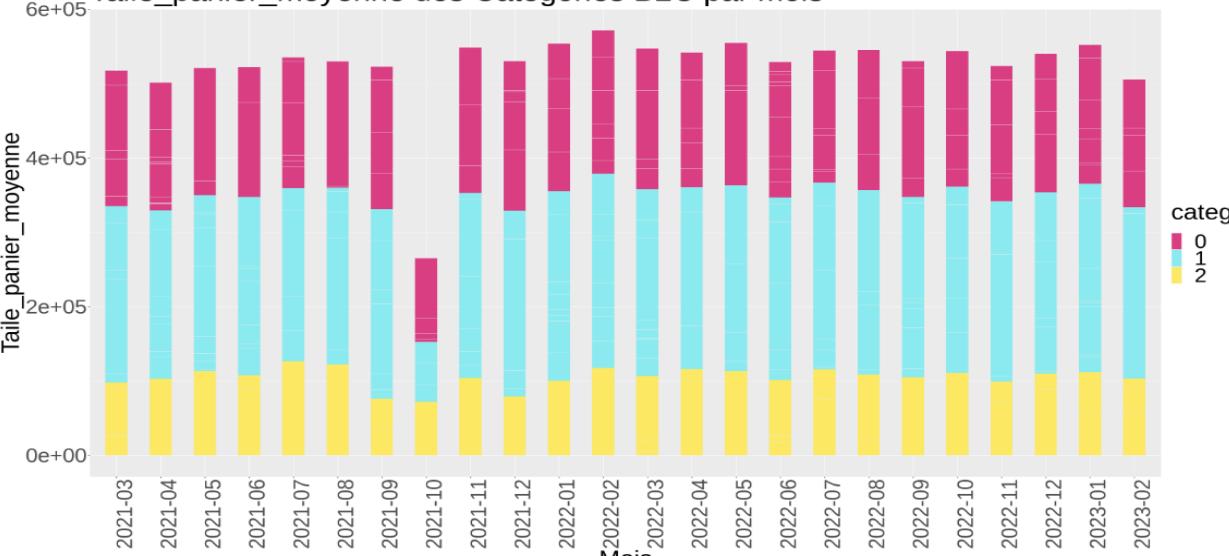
fréquence_achat des Catégories B2C par mois



Chiffre d'affaires des Catégories B2C par mois

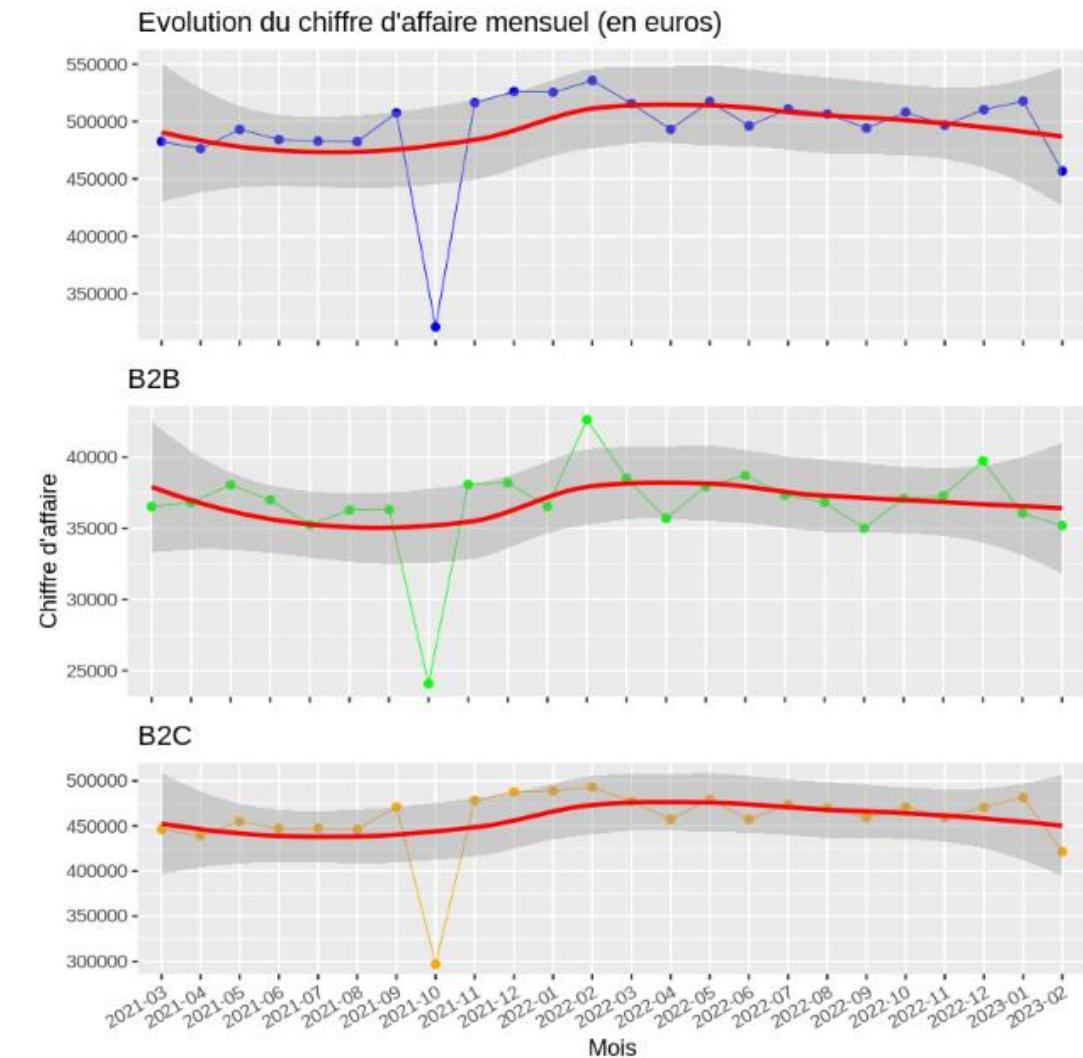


Taille_panier_moyenne des Catégories B2C par mois



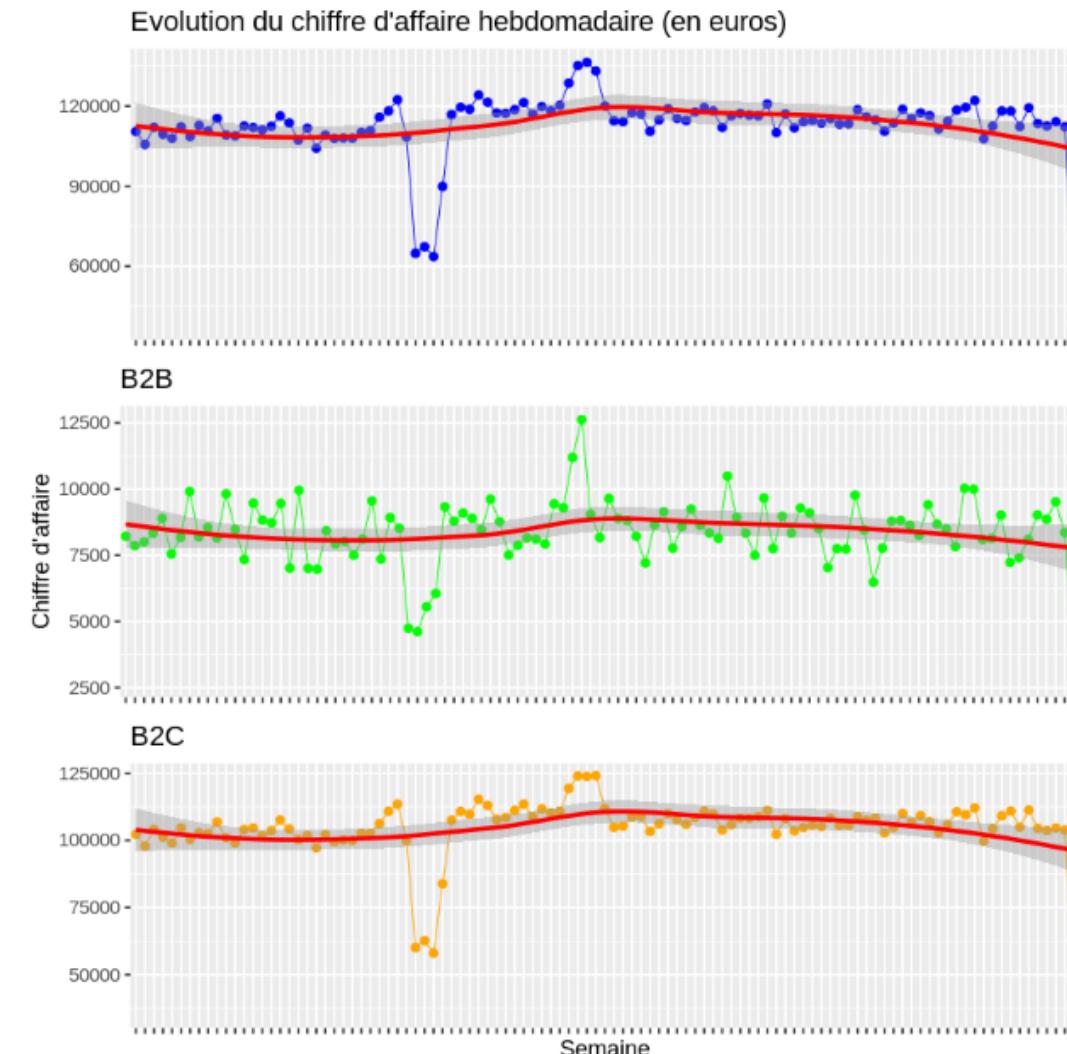
Evolution dans le temps et mettre en place une décomposition en moyenne mobile (par mois)

mois	CA_mois	mois	CA_mois_B2B	mois	CA_mois_B2C
<chr>	<dbl>	<chr>	<dbl>	<chr>	<dbl>
2022-02	535688.5	2022-02	42644.37	2022-02	493044.2
2021-12	525991.7	2022-12	39738.33	2022-01	488863.3
2022-01	525392.2	2022-06	38705.34	2021-12	487797.9
2023-01	517615.0	2022-03	38507.69	2023-01	481559.0
2022-05	517302.8	2021-12	38193.88	2022-05	479359.6
2021-11	516274.1	2021-11	38075.65	2021-11	478198.5
2022-03	515573.6	2021-05	38056.01	2022-03	477065.9
2022-07	510910.8	2022-05	37943.24	2022-07	473582.0
2022-12	510283.3	2022-07	37328.81	2021-09	471049.5
2022-10	508024.2	2022-11	37277.02	2022-10	470921.9



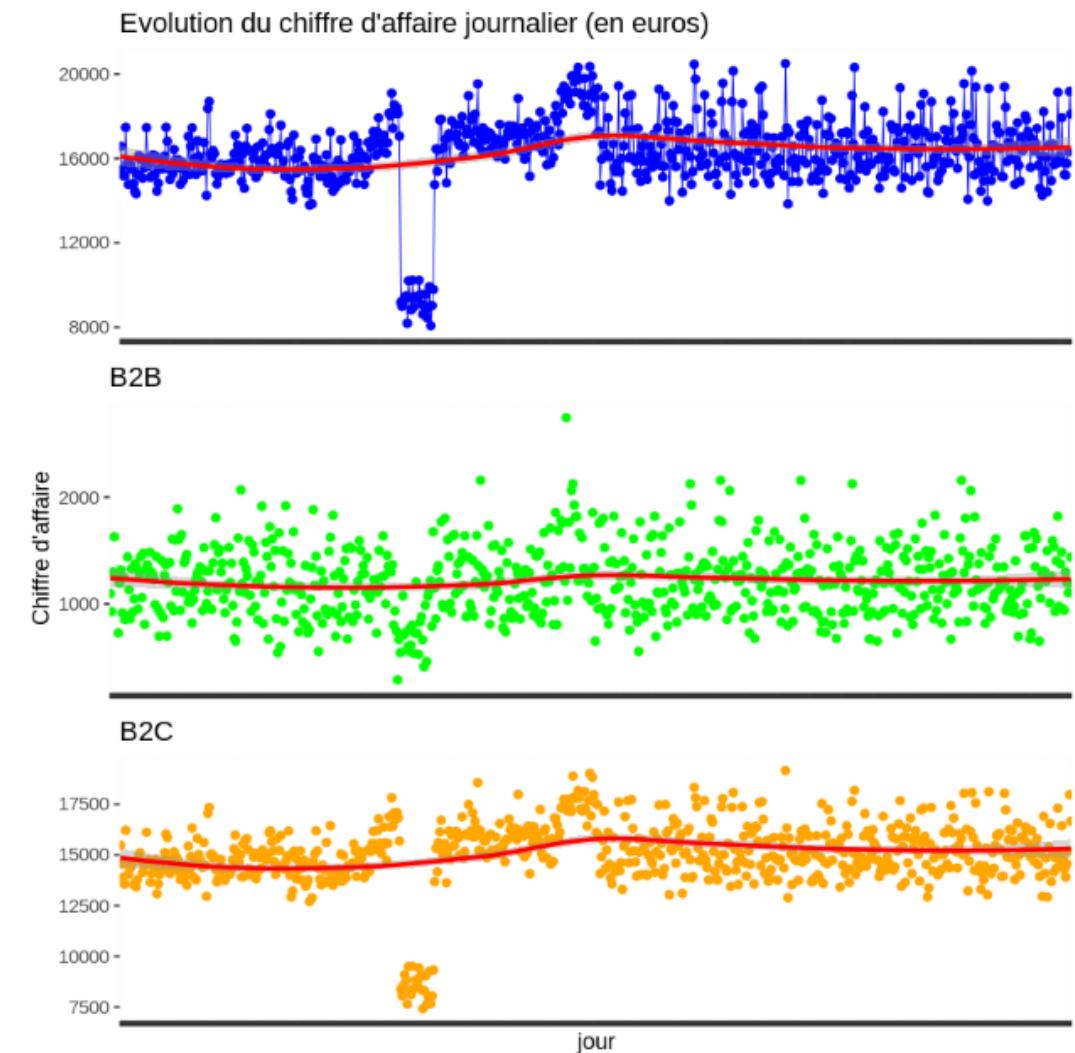
2. Evolution dans le temps et mettre en place une décomposition en moyenne mobile (par semaine)

date	CA_semaine	date	CA_semaine_B2B	date	CA_semaine_B2C
<chr>	<dbl>	<fct>	<dbl>	<fct>	<dbl>
2022-02-14	136457.8	2022-02-14	12629.100	2022-02-21	124109.8
2022-02-07	135187.9	2022-02-07	11196.700	2022-02-07	123991.2
2022-02-21	133157.7	2022-06-06	10495.660	2022-02-14	123828.7
2022-01-31	128661.9	2022-12-05	10031.558	2022-01-31	119368.2
2021-11-22	124184.1	2022-12-12	9998.760	2021-11-22	115287.2
2021-09-20	122401.7	2021-07-12	9957.470	2021-12-27	113492.0
2022-12-12	122115.0	2021-04-19	9913.038	2021-09-20	113476.6
2021-11-29	121455.9	2021-05-17	9821.390	2021-11-29	112988.8
2021-12-27	121370.7	2022-09-12	9770.580	2022-12-12	112116.2
2022-07-04	120788.7	2022-07-04	9662.450	2022-02-28	111830.7



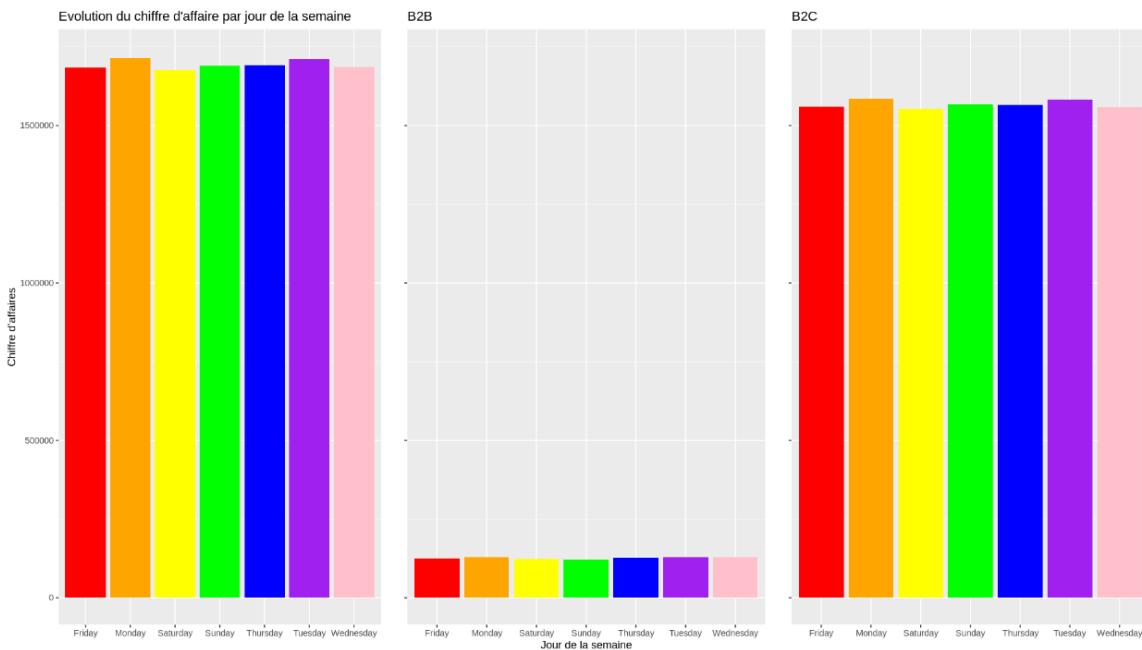
2. Evolution dans le temps et mettre en place une décomposition en moyenne mobile (par jour)

date	CA_jour	date	CA_jour_B2B	date	CA_jour_B2C
<chr>	<dbl>	<fct>	<dbl>	<fct>	<dbl>
2022-07-24	20498.32	2022-02-10	2745.50	2022-07-24	19166.54
2022-05-15	20462.51	2021-12-07	2157.74	2022-02-24	19017.56
2022-02-24	20349.34	2022-06-07	2157.74	2022-02-11	18890.98
2022-02-15	20313.53	2022-08-07	2157.74	2022-02-26	18836.10
2022-09-15	20313.53	2022-12-07	2157.74	2021-11-30	18573.62
2022-06-14	20146.21	2022-02-15	2125.37	2022-05-15	18337.14
2022-12-14	20146.21	2022-05-15	2125.37	2022-02-15	18188.16
2022-02-14	19997.23	2022-09-15	2125.37	2022-09-15	18188.16
2022-02-11	19924.33	2021-06-08	2068.64	2022-02-20	18134.23
2022-02-26	19908.56	2022-02-14	2061.73	2022-12-27	18125.05



2. Evolution dans le temps et mettre en place une décomposition en moyenne mobile (jour par semaine)

nom_jour	CA_jour	nom_jour	CA_jour_B2B	nom_jour	CA_jour_B2C
<chr>	<dbl>	<chr>	<dbl>	<chr>	<dbl>
Monday	1714470	Monday	129390.0	Monday	1585080
Tuesday	1711878	Tuesday	128216.1	Tuesday	1583661
Thursday	1692742	Wednesday	128063.6	Sunday	1567453
Sunday	1689445	Thursday	126268.7	Thursday	1566473
Wednesday	1687002	Friday	124558.5	Friday	1560165
Friday	1684724	Saturday	122637.3	Wednesday	1558938
Saturday	1675820	Sunday	121992.1	Saturday	1553183



Analyse des données

2.Demandes de Julie

Testes Statistiques

1. le lien entre le genre d'un client et les catégories des livres achetés (Test du Khi-Deux)

- Pour tester le lien entre deux variables catégorielles (qualitatives)

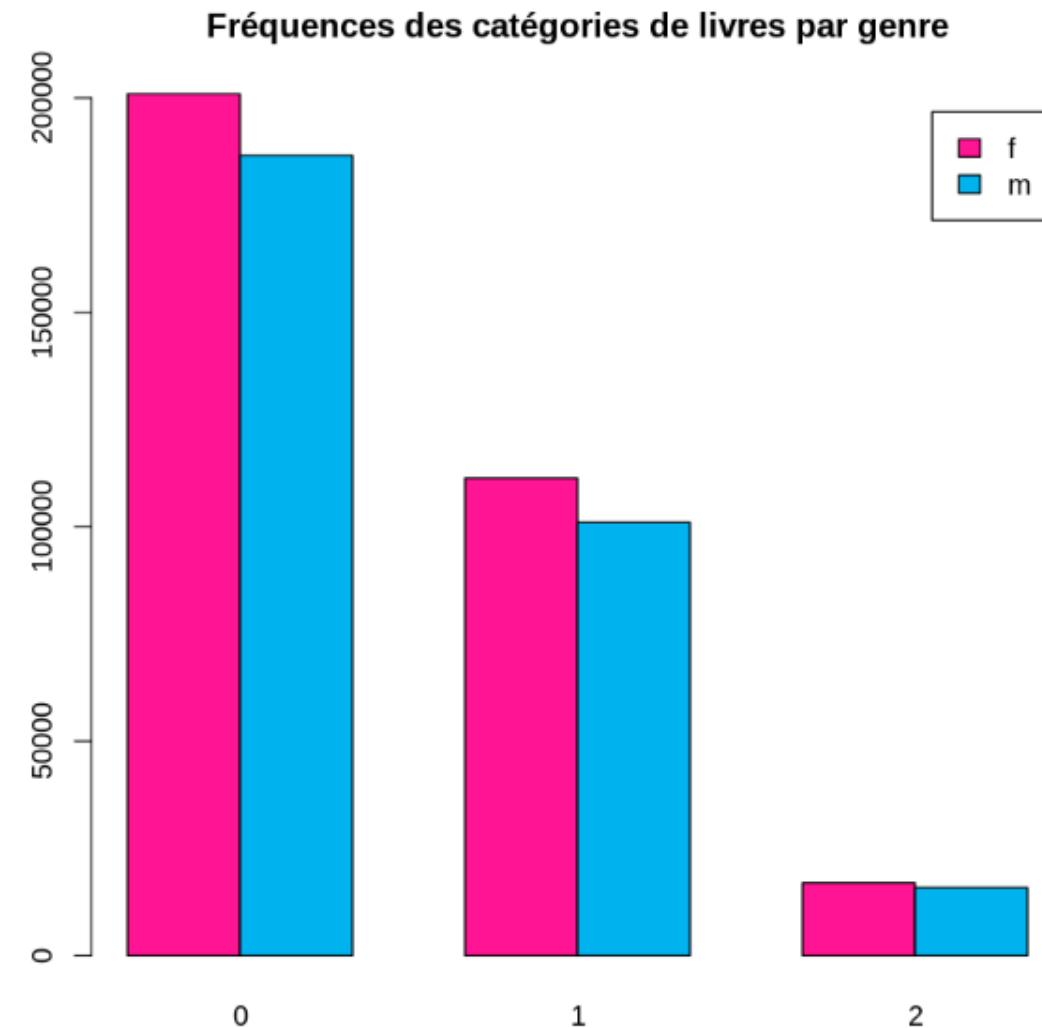
Hypothèses :

- H0:** Les variables sont indépendantes ($P>0,05$).
- H1:** Les variables ne sont pas indépendantes ($P<0,05$).

Dans notre résultat :

$p = 4,108e-05 < 0,05 \Rightarrow H1$ (rejeter l'hypothèse nulle d'indépendance)

- Il existe une association significative entre les variables categ et sex.**
- Les 2 variables ne sont pas indépendantes**



2. Le lien entre l'âge des clients et le montant total des achats

1. Le test de Shapiro-Wilk

Utilisé pour vérifier si une variable suit une distribution normale.

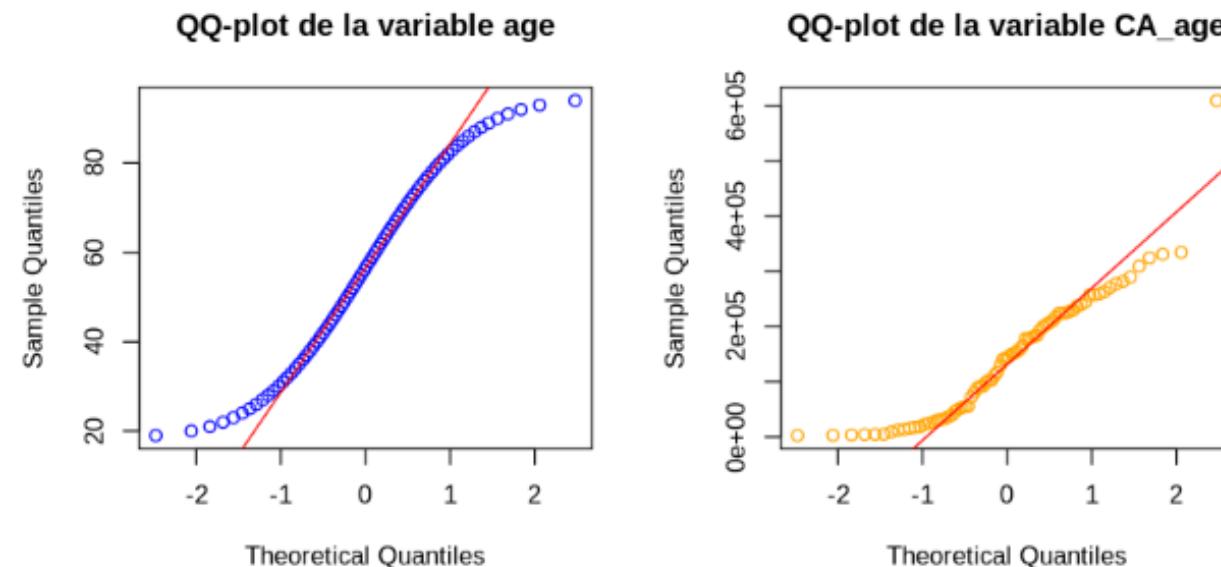
Hypothèses :

- H0 ($P>0,05$): Une variable suit une distribution normale.
- H1 ($P<0,05$): Une variable ne suit pas une distribution normale.

Dans notre résultat

- age : $0.008753 < 0,05 \Rightarrow$ ne suit pas une distribution normale.
- CA_age : $6.049e-05 < 0,05 \Rightarrow$ ne suit pas une distribution normale .

➤ **Cela signifie que les corrélations de Spearman ou de Kendall doivent être utilisées pour mesurer la relation monotone entre "age" et "CA_age".**



2.Le lien entre l'âge des clients et le montant total des achats

2.les corrélations de Spearman

Mesure de corrélation non paramétrique qui évalue la relation monotone entre 2 variables.

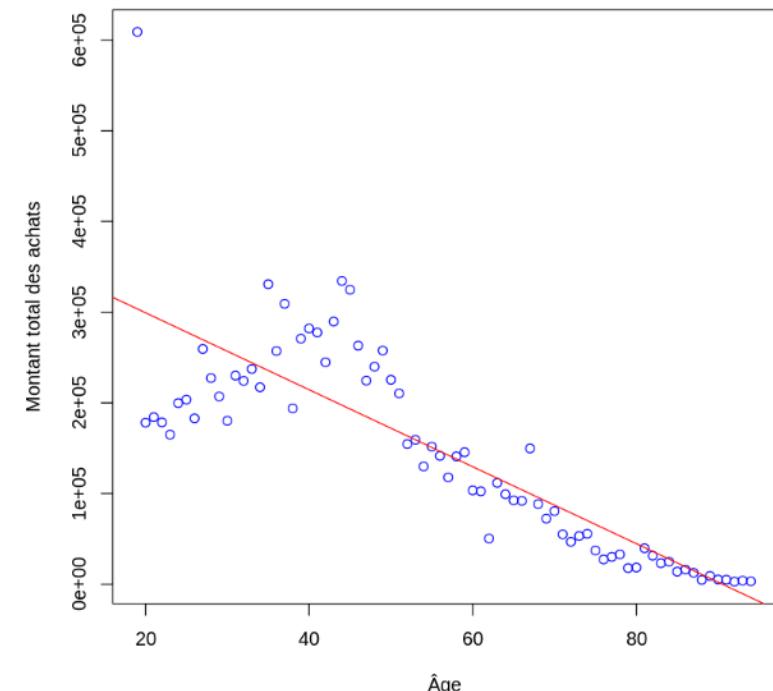
Hypothèses : Le coefficient varie entre -1 et 1, où

- une valeur 0 indique les deux variables sont indépendantes.
- une valeur différent de 0 indique les deux variables sont liées.

-0.8736842 => Une forte relation négative monotone entre les deux variables

➤ Plus l'âge est élevé, plus la valeur de CA age est faible et inversement.

Relation entre l'âge et CA_age



3.les corrélations de kendall

Mesure de corrélation non paramétrique qui évalue la relation entre 2 variables.

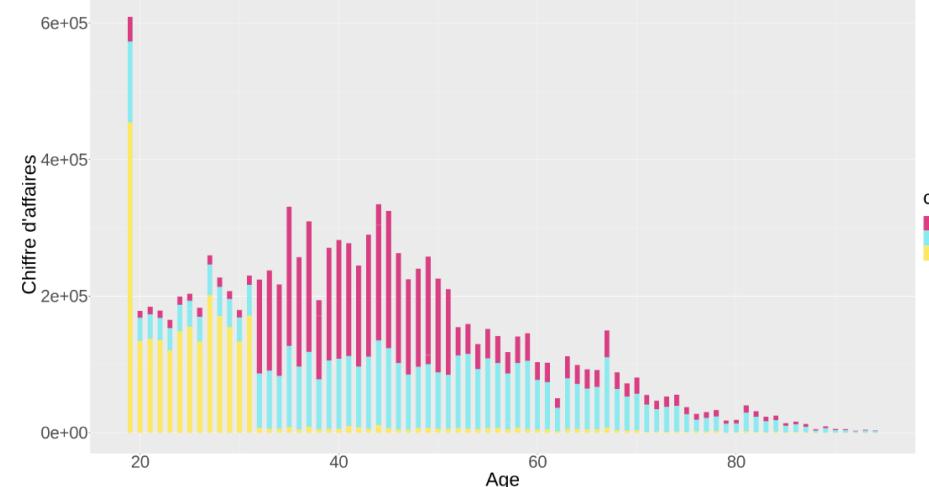
Hypothèses : Le coefficient varie entre -1 et 1, où

- une valeur 0 indique les deux variables sont indépendantes.
- une valeur différent de 0 indique les deux variables sont liées.

-0.7080702 => Une forte relation négative monotone entre les deux variables

➤ Plus l'âge est élevé, plus la valeur de CA age est faible et inversement.

Chiffre d'affaires & Catégories pour B2C par age



3. Le lien entre l'âge des clients et la fréquence d'achat

1. Le test de Shapiro-Wilk

Utilisé pour vérifier si une variable suit une distribution normale.

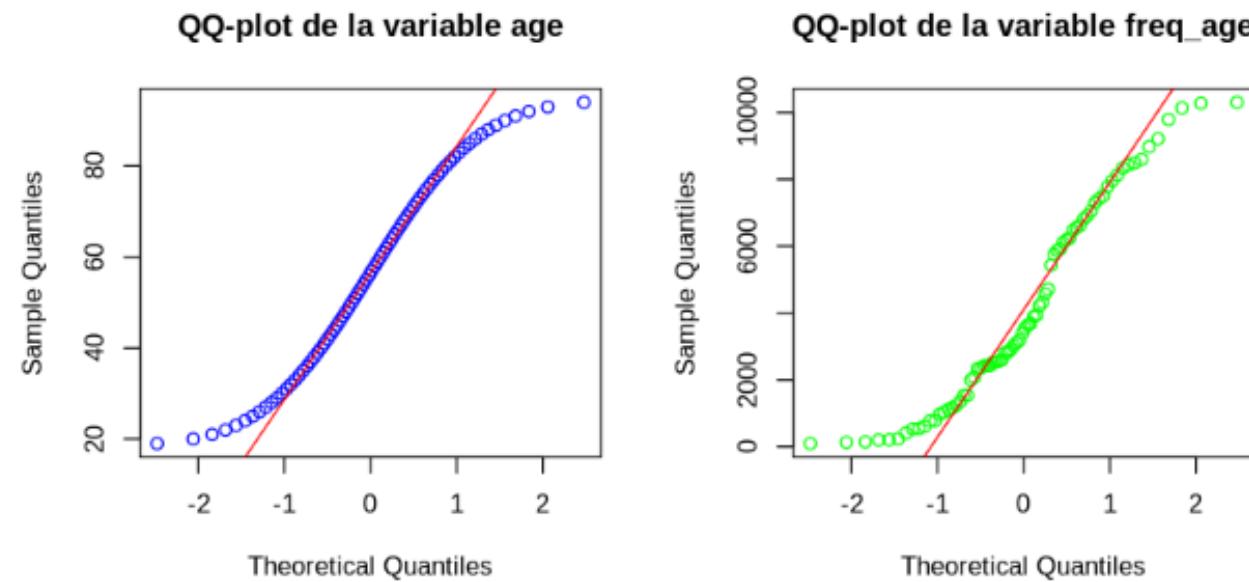
Hypothèses :

- $H_0 (P>0,05)$: Une variable suit une distribution normale.
- $H_1 (P<0,05)$: Une variable ne suit pas une distribution normale.

Dans notre résultat

- age : $0.008753 < 0,05 \Rightarrow$ ne suit pas une distribution normale.
- freq_age : $0.0005561 < 0,05 \Rightarrow$ ne suit pas une distribution normale .

➤ Cela signifie que les corrélations de Spearman ou de Kendall doivent être utilisées pour mesurer la relation monotone entre "age"et "freq_age".



3.Le lien entre l'âge des clients et la fréquence d'achat

2.les corrélations de Spearman

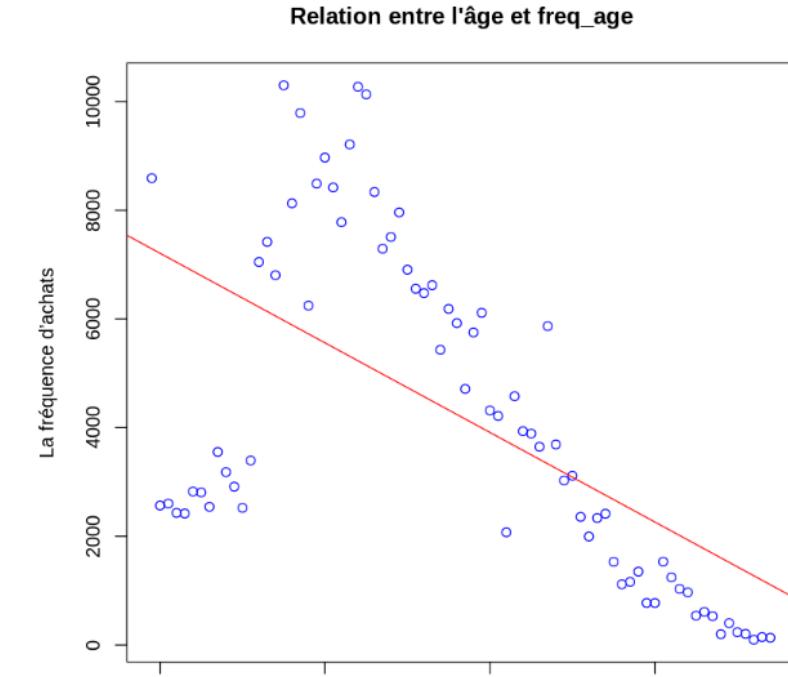
Mesure de corrélation non paramétrique qui évalue la relation monotone entre 2 variables.

Hypothèses : Le coefficient varie entre -1 et 1, où

- une valeur 0 indique les deux variables sont indépendantes.
- une valeur différent de 0 indique les deux variables sont liées.

-0.6593848=> Une forte relation négative monotone entre les deux variables

➤ Plus l'âge est élevé, plus la valeur de freq_age est faible et inversement.



3.les corrélations de kendall

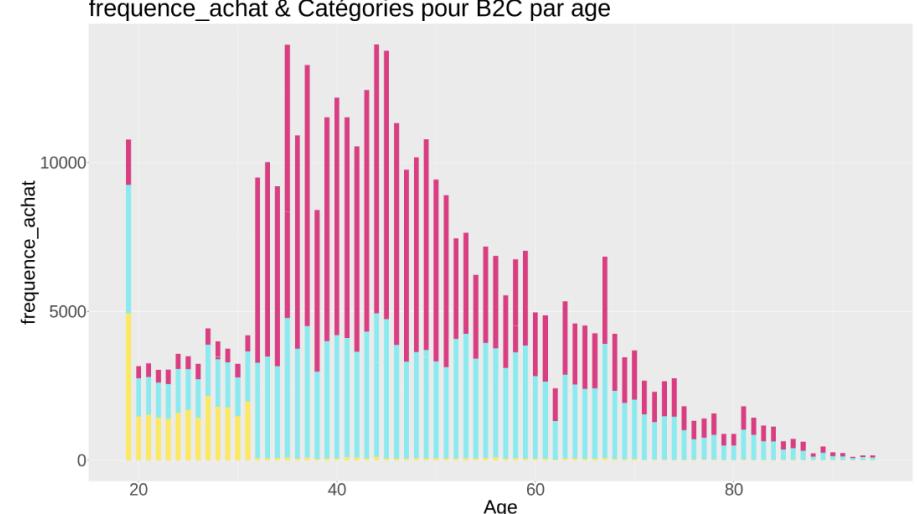
Mesure de corrélation non paramétrique qui évalue la relation entre 2 variables.

Hypothèses : Le coefficient varie entre -1 et 1, où

- une valeur 0 indique les deux variables sont indépendantes.
- une valeur différent de 0 indique les deux variables sont liées.

-0.5445614=> Une forte relation négative monotone entre les deux variables

➤ Plus l'âge est élevé, plus la valeur de freq_age est faible et inversement.



4. Le lien entre l'âge des clients et la taille du panier moyen

1. Le test de Shapiro-Wilk

Utilisé pour vérifier si une variable suit une distribution normale.

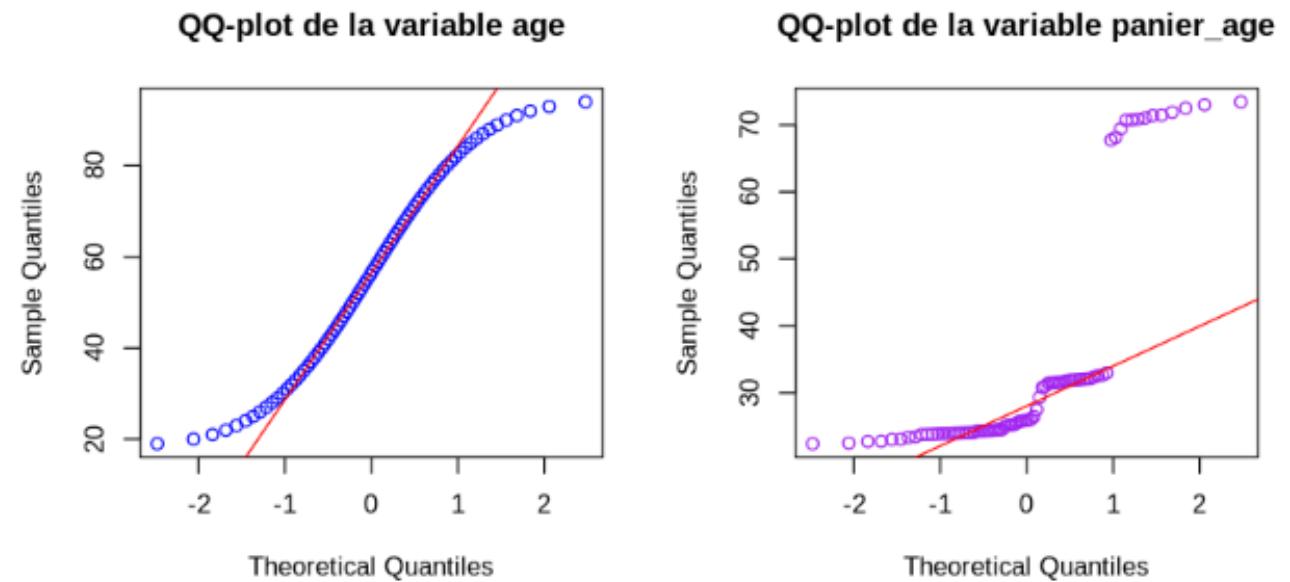
Hypothèses :

- H0 ($P>0,05$): Une variable suit une distribution normale.
- H1 ($P<0,05$): Une variable ne suit pas une distribution normale.

Dans notre résultat

- age : 0.008753 < 0,05 => ne suit pas une distribution normale.
- panier_age : 1.117e-12. < 0,05 => ne suit pas une distribution normale .

➤ **Cela signifie que les corrélations de Spearman ou de Kendall doivent être utilisées pour mesurer la relation monotone entre "age" et "panier_age".**



4. Le lien entre l'âge des clients et la taille du panier moyen

2. les corrélations de Spearman

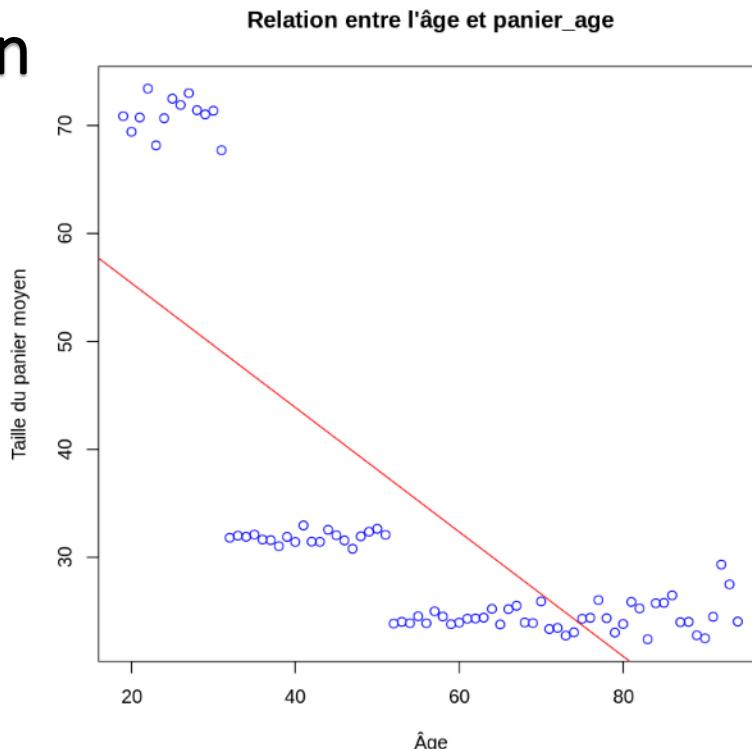
Mesure de corrélation non paramétrique qui évalue la relation monotone entre 2 variables.

Hypothèses : Le coefficient varie entre -1 et 1, où

- une valeur 0 indique les deux variables sont indépendantes.
- une valeur différent de 0 indique les deux variables sont liées.

- -0.7670267 => Une forte relation négative monotone entre les deux variables

➤ Plus l'âge est élevé, plus la valeur de panier_age est faible et inversement.



3. les corrélations de kendall

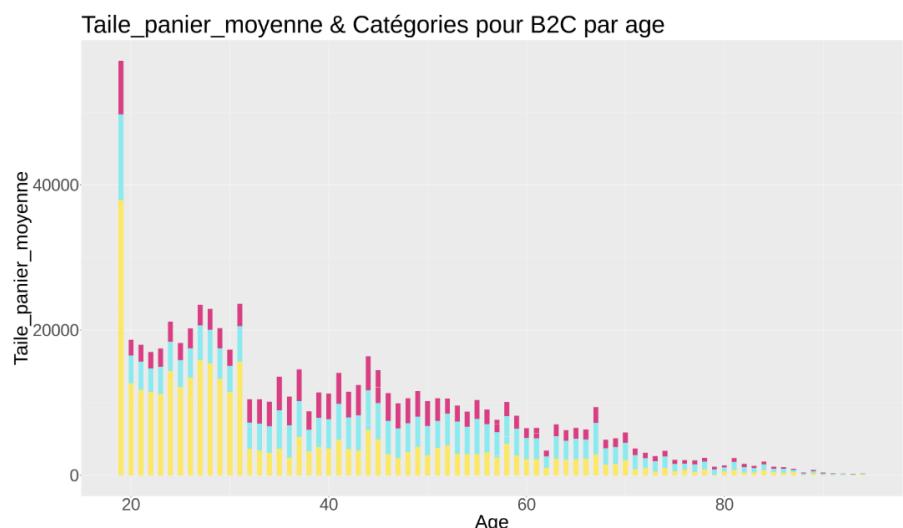
Mesure de corrélation non paramétrique qui évalue la relation entre 2 variables.

Hypothèses : Le coefficient varie entre -1 et 1, où

- une valeur 0 indique les deux variables sont indépendantes.
- une valeur différent de 0 indique les deux variables sont liées.

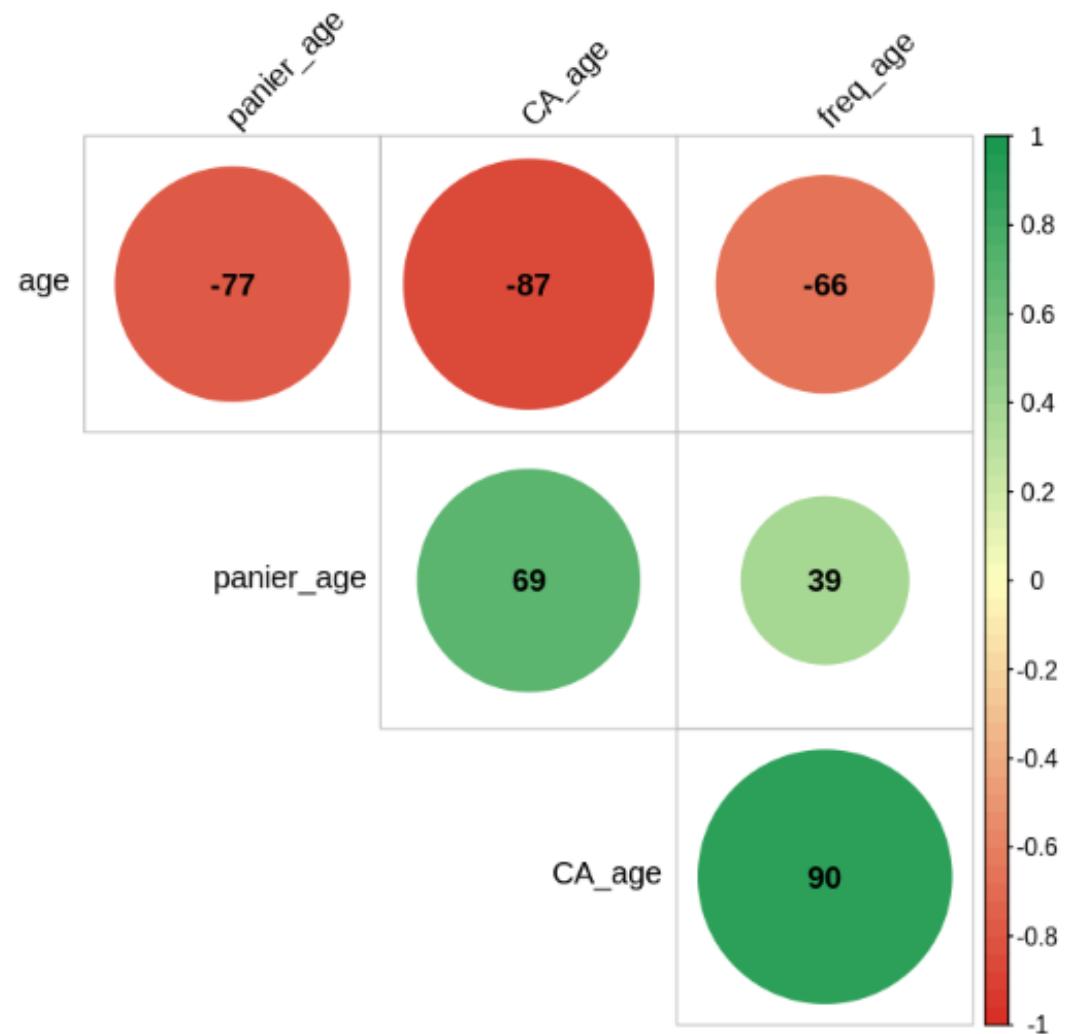
-0.5515789 => Une forte relation négative monotone entre les deux variables

➤ Plus l'âge est élevé, plus la valeur de panier_age est faible et inversement.

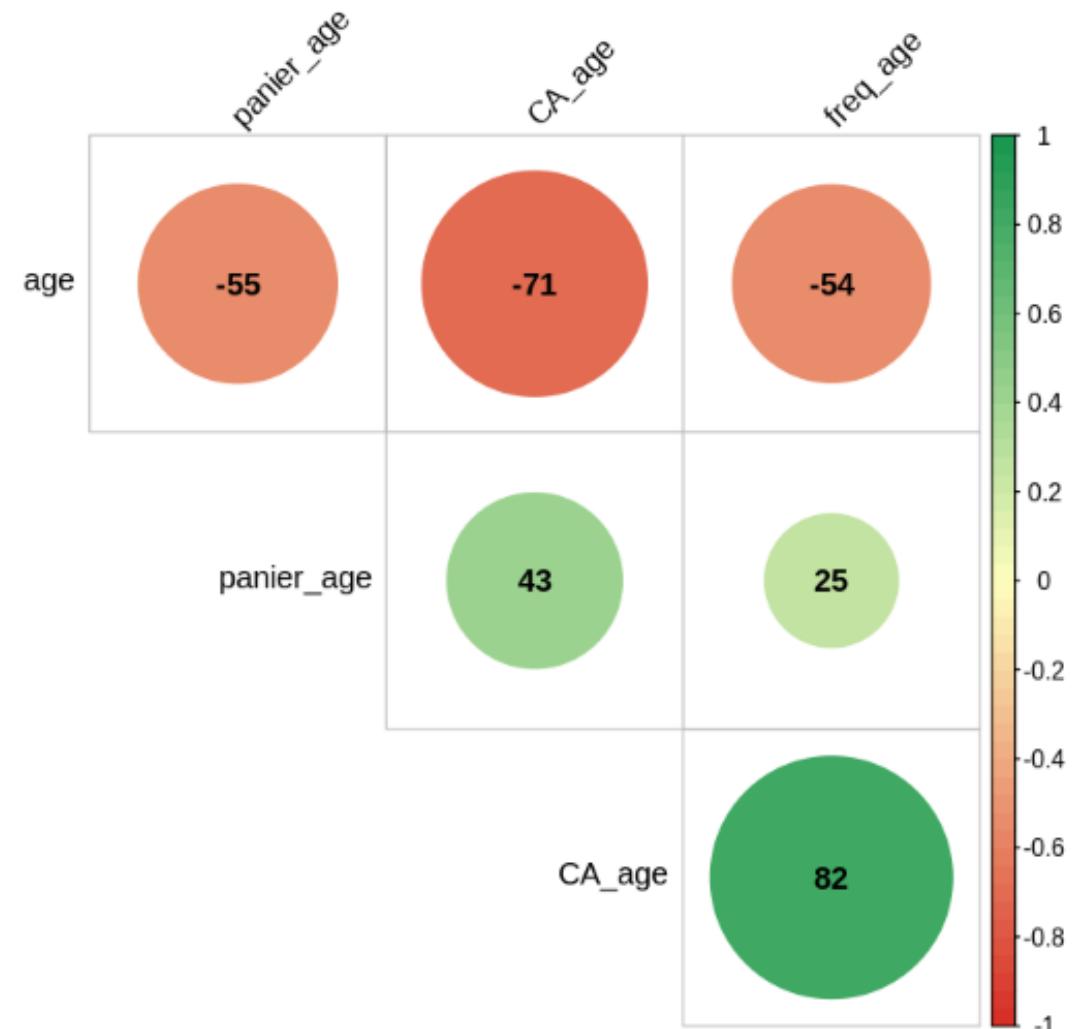


4. Matrix de Spearman & kendall pour panier_freq_CA_age_B2C

Matrice de corrélation de Spearman



Matrice de corrélation de Kendall



5. Le lien entre l'âge des clients et la catégorie des livres achetés

A

1. Le test de Shapiro-Wilk

Utilisé pour vérifier si une variable suit une distribution normale.

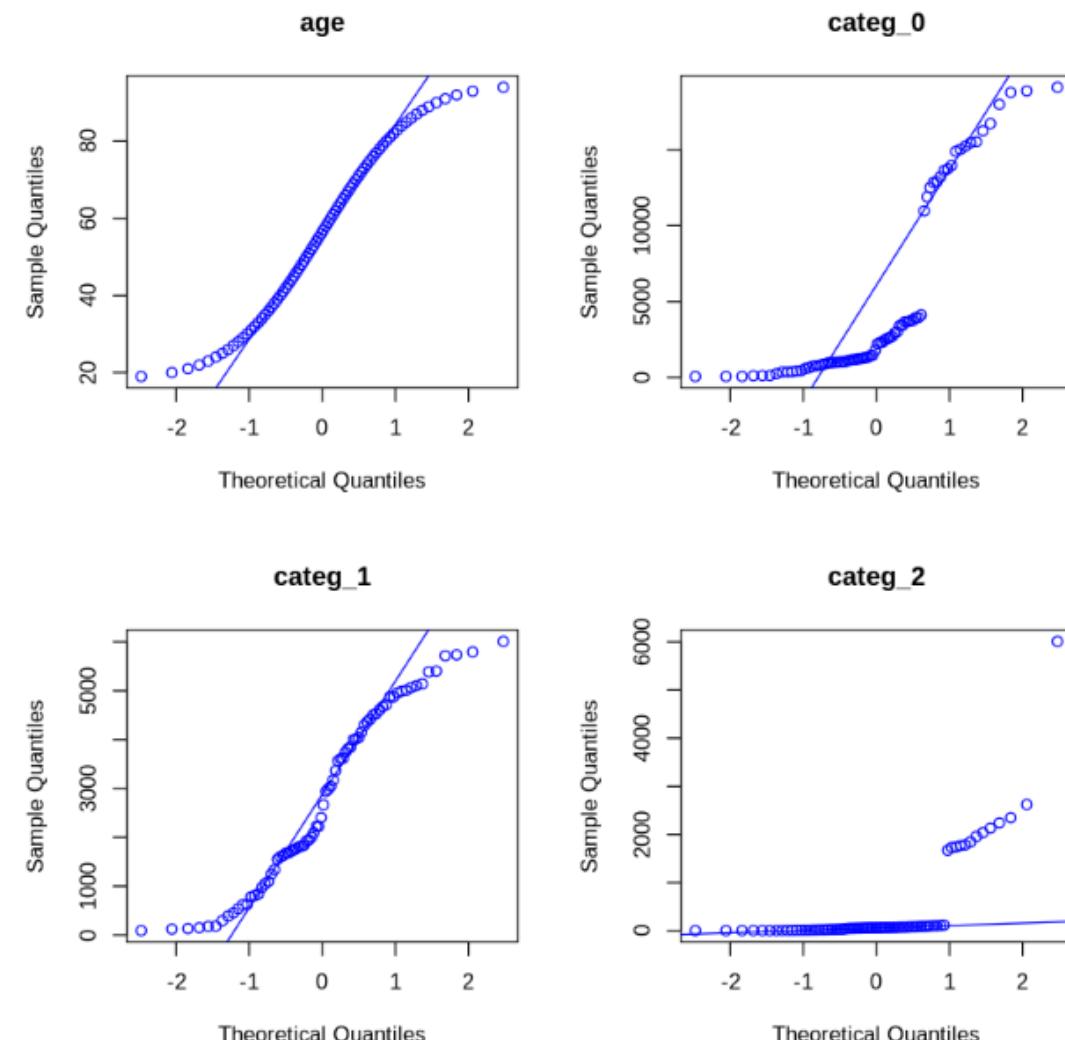
Hypothèses :

- H0 ($P>0,05$): Une variable suit une distribution normale.
- H1 ($P<0,05$): Une variable ne suit pas une distribution normale.

Dans notre résultat

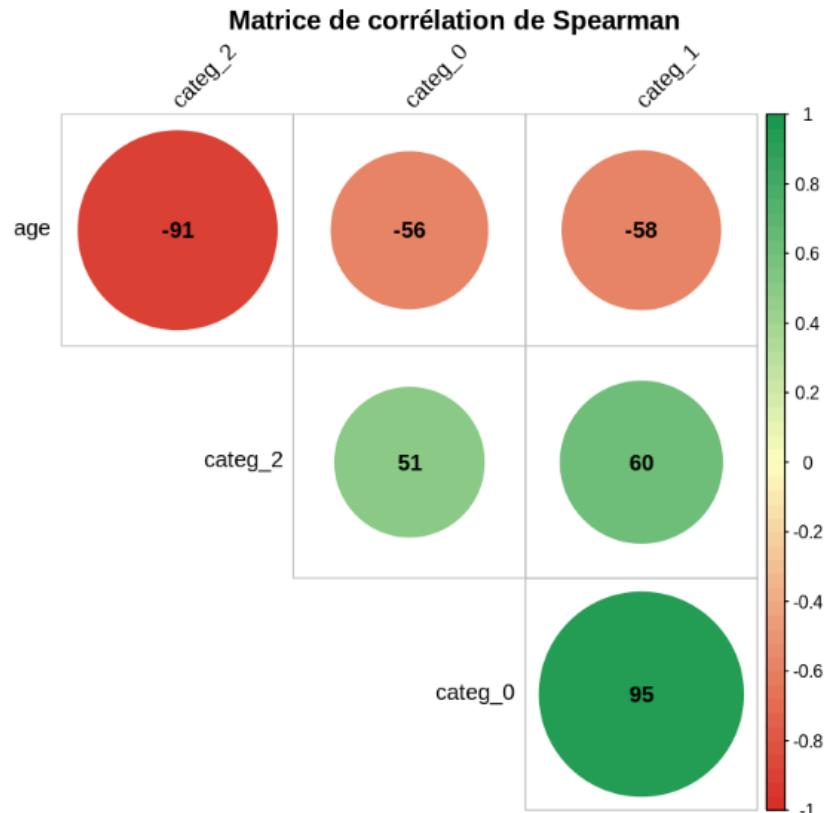
- Les données pour les variables :
- `age`, `categ_0`, `categ_1` et `categ_2`
- ne suivent pas une distribution normale

➤ Un test statistique non paramétrique couramment utilisé pour évaluer la relation entre deux variables quantitatives est soit le test de corrélation de Spearman ou Kendall.

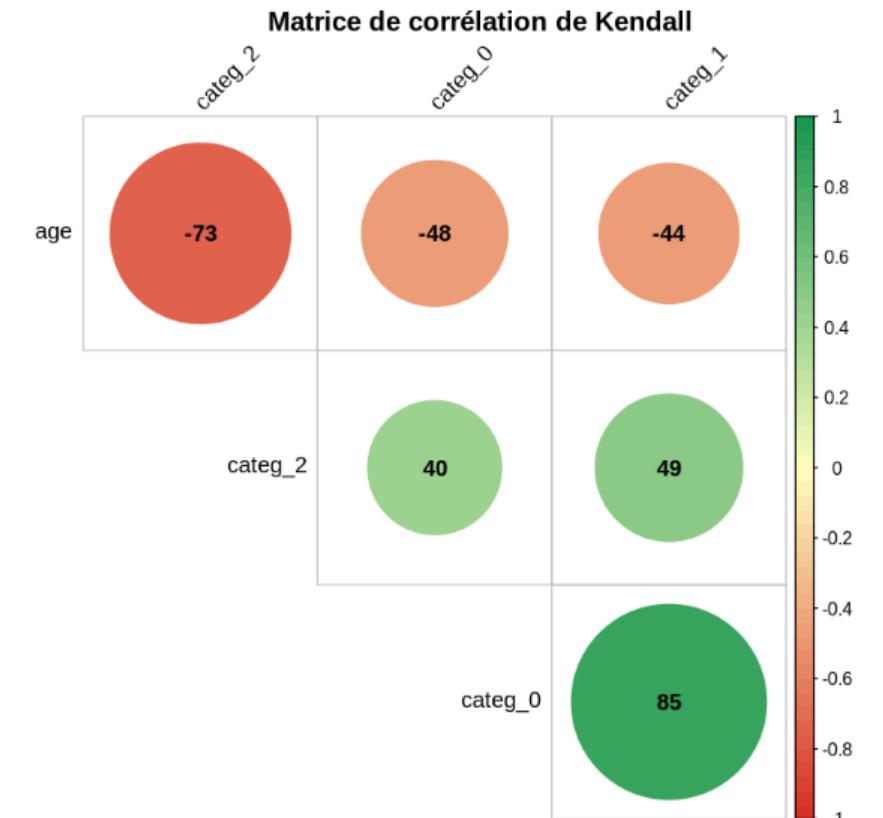


5. Le lien entre l'âge des clients et la catégorie des livres achetés

A



La variable **age** est négativement corrélée avec les variables **categ_0**, **categ_1** et **categ_2**. Cela signifie que plus l'âge est élevé, moins il y a de sessions pour les catégories 0, 1 et 2. La corrélation entre **age** et **categ_2** est particulièrement forte, avec une valeur de corrélation de -0,905661.



La variable **age** est négativement corrélée avec les variables **categ_0**, **categ_1** et **categ_2**. Cela signifie que plus l'âge est élevé, moins il y a de sessions pour les catégories 0, 1 et 2. La corrélation entre **age** et **categ_2** est particulièrement forte, avec une valeur de corrélation de -0,7324351.

5. Le lien entre l'âge des clients et la catégorie des livres achetés

B

1. Le test de Kolmogorov-Smirnov

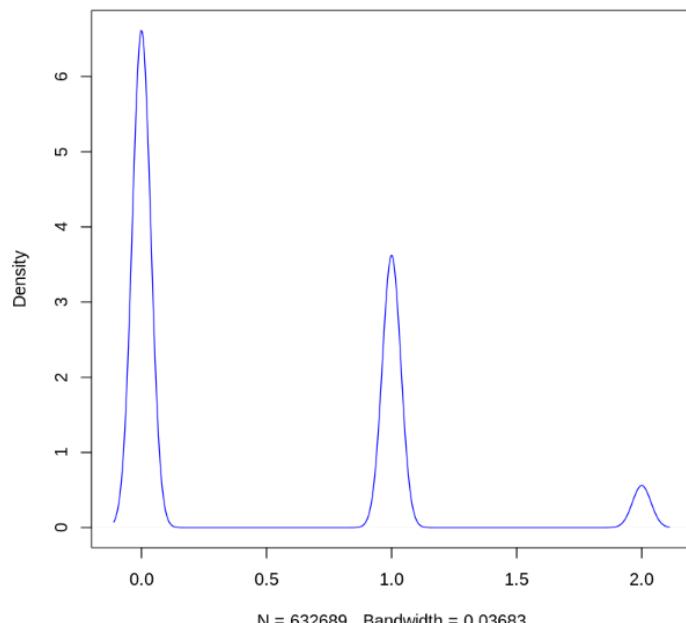
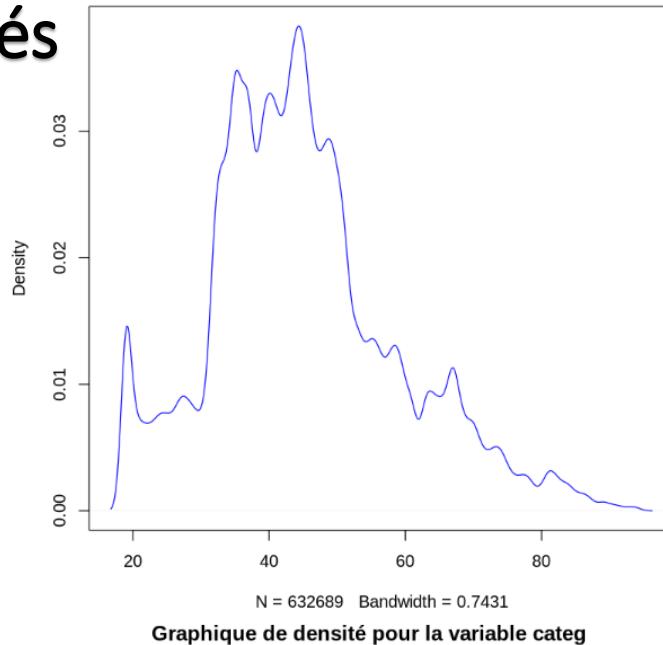
Ce test statistique couramment utilisé pour les échantillons de grande taille parce que le test Shapiro ne peut être utilisé que pour des échantillons de taille comprise entre 3 et 5000.

Hypothèses :

- H0 ($P > 0,05$): Une variable suit une distribution normale.
- H1 ($P < 0,05$): Une variable ne suit pas une distribution normale.

Dans notre résultat

- age et categ : $< 2.2e-16 < 0,05 \Rightarrow$ ne suit pas une distribution normale .
- Un test statistique non paramétrique couramment utilisé pour évaluer la relation entre une variable quantitative et une variable qualitative est le test de Kruskal-Wallis.



5. Le lien entre l'âge des clients et la catégorie des livres achetés

B

2. Le test de Kruskal-Wallis

Utilisé pour tester si les médianes de deux ou plusieurs groupes sont égales.

Hypothèses :

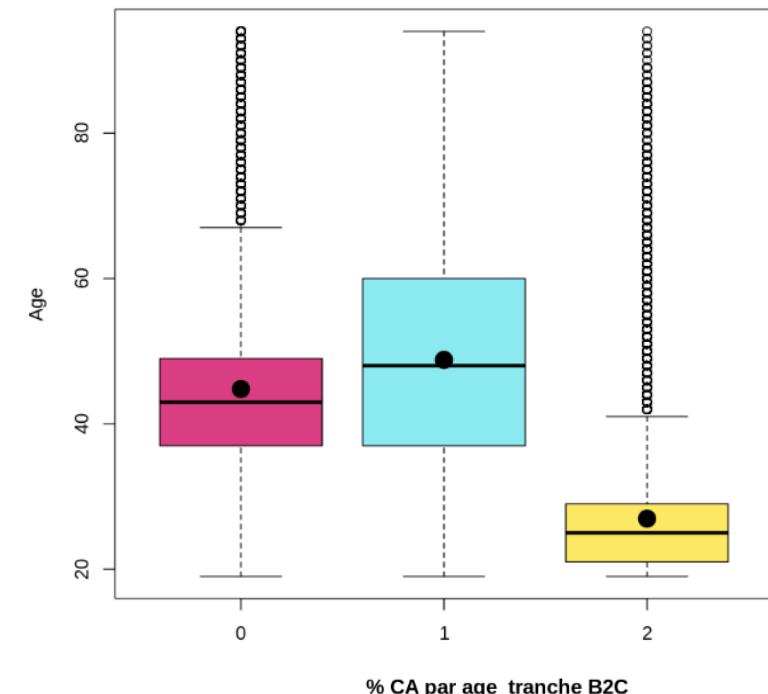
- H0 ($P>0,05$): les échantillons ayant la même distribution.
- H1 ($P<0,05$): les échantillons n'ayant pas la même distribution.

Dans notre résultat

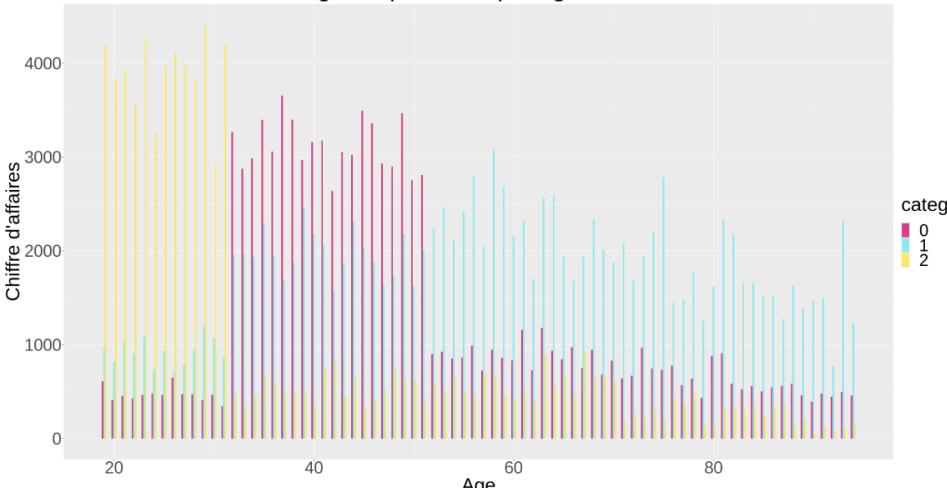
○ $< 2.2e-16 < 0,05 \Rightarrow$ ne suit pas la même distribution.

➤ Cela suggère que les groupes ne sont pas équivalents et que les médianes des groupes sont significativement différentes.

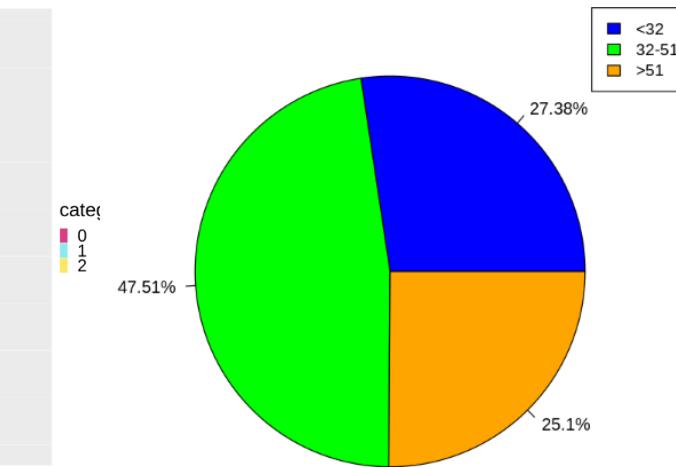
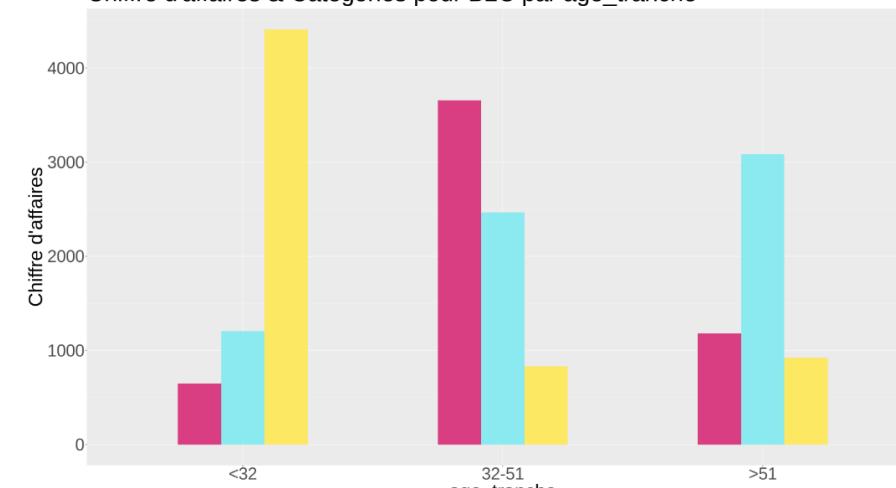
L'âge de Clients B2C en fonction des Catégorie



Chiffre d'affaires & Catégories pour B2C par age



Chiffre d'affaires & Catégories pour B2C par age_tranche



Conclusion et perspectives

1. Il n'y a pas d'évolution notable dans le temps des chiffres d'affaires
2. Il n'y a pas de mois, de semaine ou de jour préféré pour nos clients
3. Il y a un trou dans les ventes d'octobre 2021 en catégorie 1, peut-être dû à une erreur technique sur le site ou à des problèmes d'approvisionnement ou de stocks. Prévenir la cause de ce trou à l'avenir.
4. Nous avons deux types de clients : les clients importants (B2B =4 clients) qui n'ont jamais manqué un achat sur notre site même un seul jour avec un nombre de ventes important, et les clients réguliers (B2C =8 596 clients).
5. Il existe une association significative entre le genre d'un client et les catégories des livres achetés
6. Plus l'âge est élevé, plus la valeur de chiffres d'affaires, la fréquence d'achat et la taille du panier moyen sont faibles et inversement.
7. La catégorie 0 est privilégiée par les 32-51 ans, 2 par les <32 ans par contre 1 par tous nos clients
8. Les chiffres d'affaires et la fréquence d'achat sont plus élevés chez des 32-51 ans mais la taille panier moyen plus élevés chez 32<ans
9. Proposer des offres les jours fériés, les occasions, le rentrée et le week-end.
10. Conclure des accords commerciaux et proposer des offres spéciales à nos clients importants B2B.
11. Créer une offre par catégorie et par âge réservé à nos clients réguliers B2C.