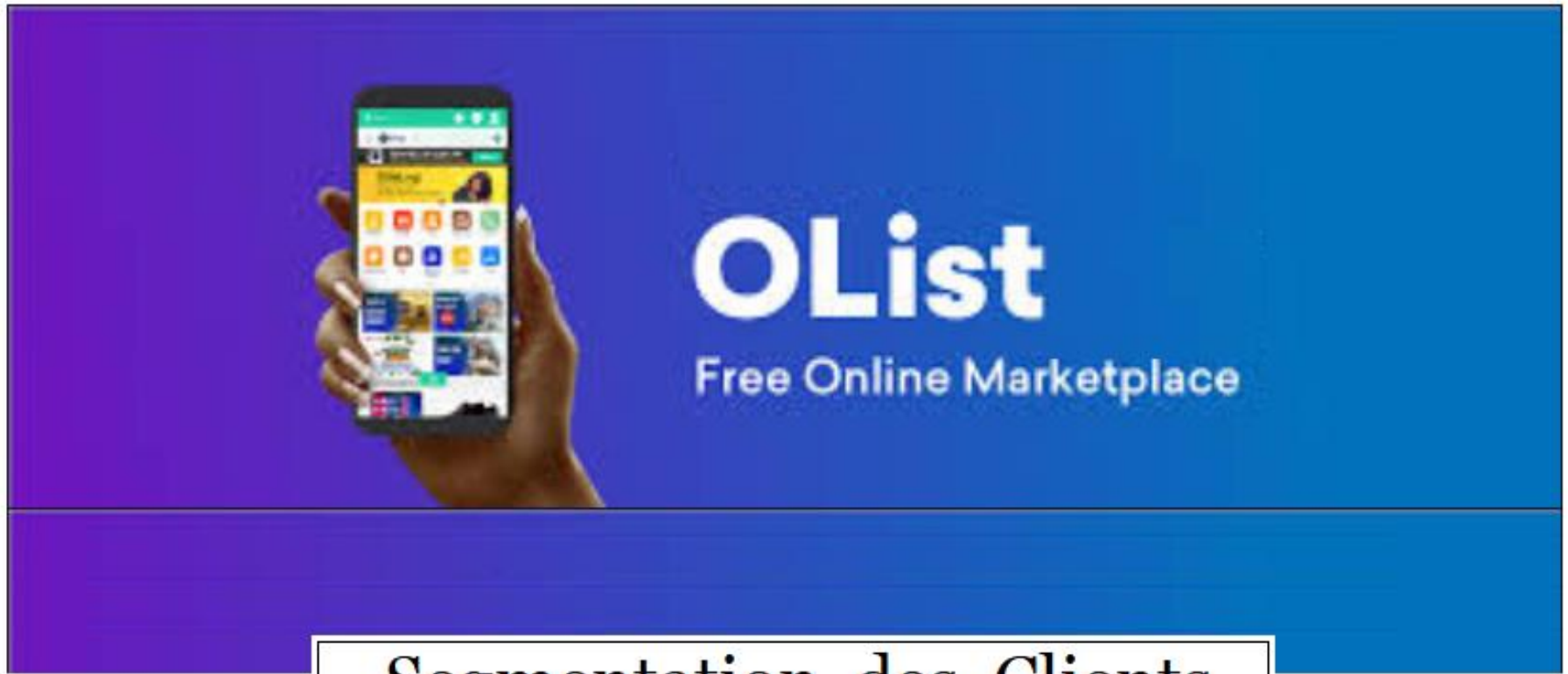
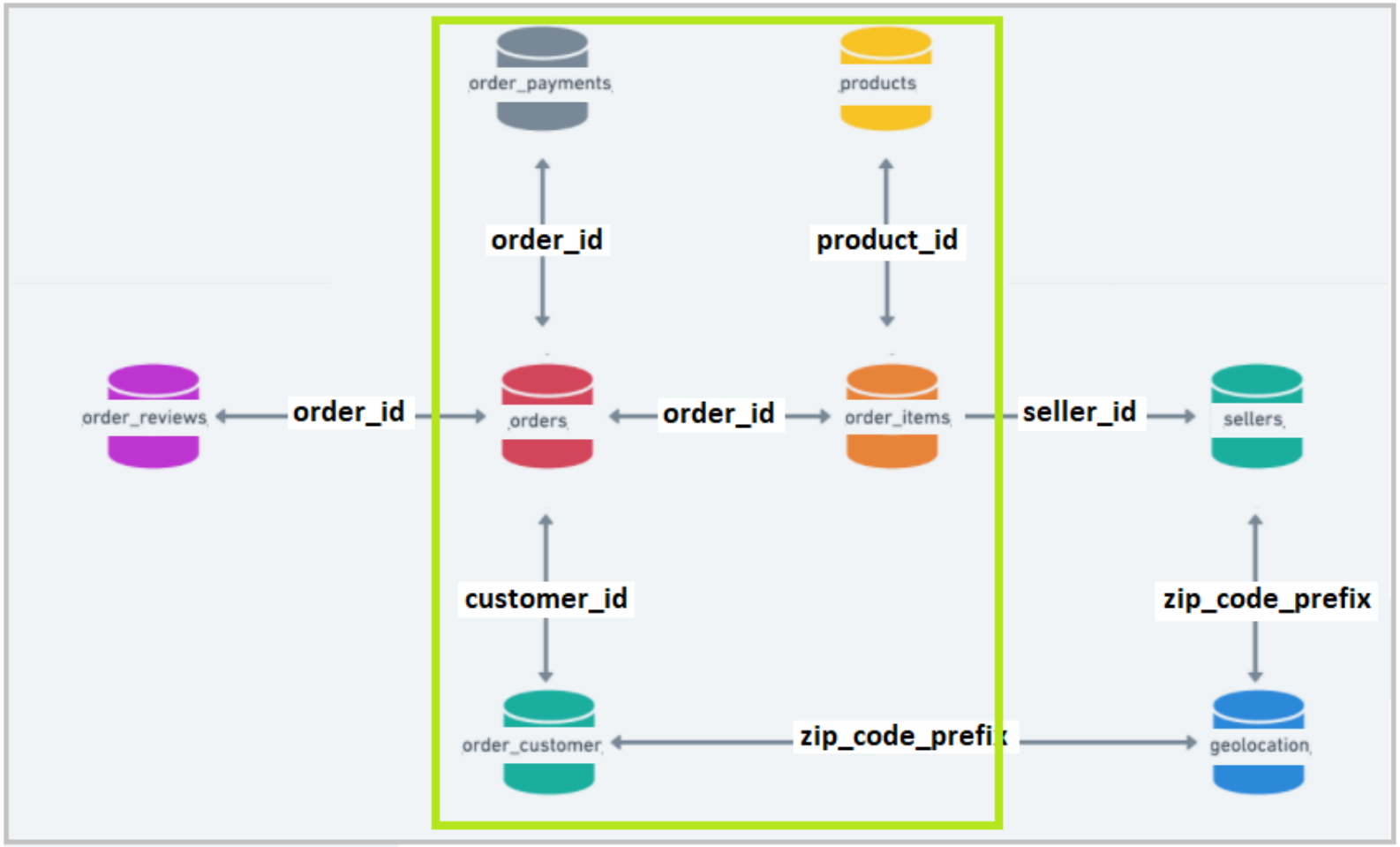


Segmentation des Clients

Sommaire

Source des Données : OLIST



: **tables utilisées**

Clients

customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
4a31da2ba757502ad7fcc9dc6768cfa5	b8b3c435a58aebd788a477bed8342910	95585	arroio do sal	RS
299f7b5125c8fbe1761a1b320c34fc7d	b8b3c435a58aebd788a477bed8342910	95585	arroio do sal	RS
5eaea72cbd8dda40b6b7ac932be5c393	b8b3c435a58aebd788a477bed8342910	95585	arroio do sal	RS
0c155574f4f4d27594dbdd37731a6ecb	b8b3c435a58aebd788a477bed8342910	95585	arroio do sal	RS

Données retenues

doublons

Cardinalité : **96096**

Nbre clients multi-commandes : **2886**

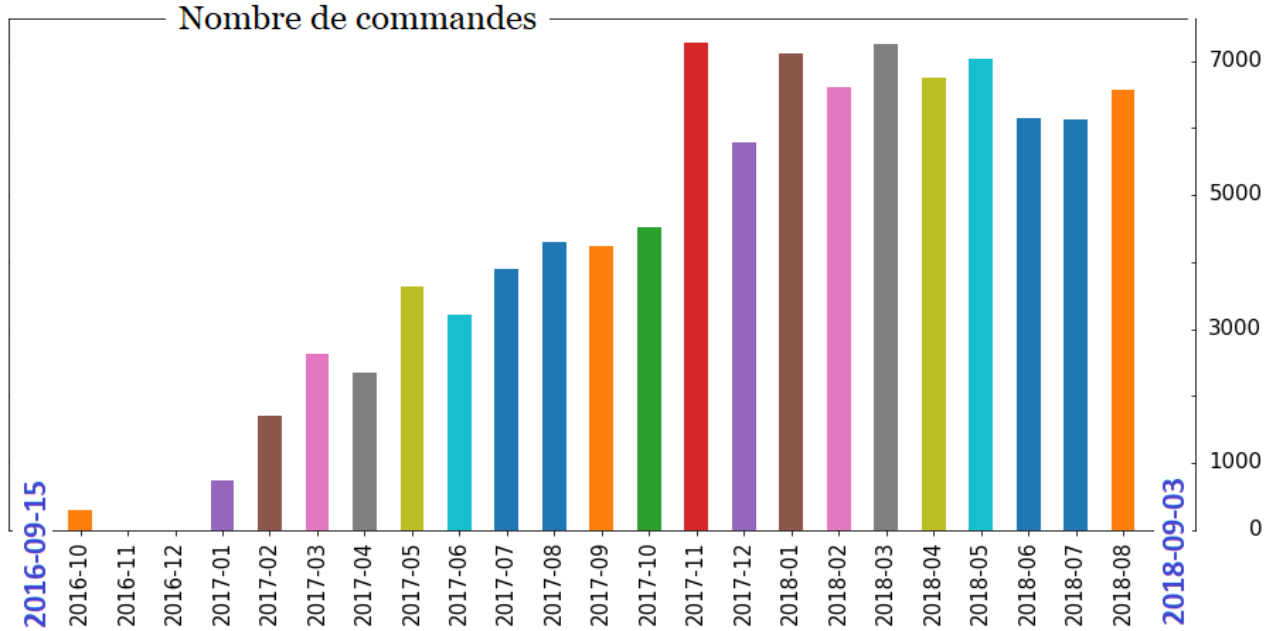
Commandes

order_id	customer_id	order_status	approved_at	delivered_carrier	delivered_customer	estimated_delivery
53cdb2fc8bc7dce0b6741e2150273451	af07308b275d755c9edb36a90c618231	delivered	26/07/2018 03:24	26/07/2018 14:31	07/08/2018 15:27	13/08/2018 00:00
47770eb9100c2d0c44946d9cf07ec65d	3a653a41f6f9fc3d2a113cf8398680e8	invoiced	null	08/08/2018 13:50	17/08/2018 18:06	04/09/2018 00:00
747996a66f5aa711deb8ae58f5ae46a0	12f5d6e1cbf93dafd9dcc19095df0b3d	shipped	null	11/01/2017 15:59	16/01/2017 15:18	13/02/2017 00:00
9b91ddcbd6cbceb83d4fd2462ca1f95e	12f5d6e1cbf93dafd9dcc19095df0b3d	processing	07/01/2017 03:44		16/01/2017 16:32	01/02/2017 00:00
464de32dc84484c1d26df3e8e38e708b	12f5d6e1cbf93dafd9dcc19095df0b3d	unavailable	07/01/2017 03:35	11/01/2017 15:59	17/01/2017 16:09	13/02/2017 00:00
17fed53ba6dfef9b594ee2268642e2aa	12f5d6e1cbf93dafd9dcc19095df0b3d	canceled	07/01/2017 03:35		16/01/2017 15:24	13/02/2017 00:00
ca5a215980675471f0cf8199c041909a	12f5d6e1cbf93dafd9dcc19095df0b3d	created	07/01/2017 03:44	11/01/2017 15:37		01/02/2017 00:00
8a784d47854e4cbc5562362393d504db	12f5d6e1cbf93dafd9dcc19095df0b3d	approved	07/01/2017 03:44	11/01/2017 16:08		13/02/2017 00:00

Données retenues

Cardinalité : **99441**

Nbre maximal de commandes / client : **16**



Lignes de commande

order_id	order_item_id	product_id	shipping_limit_date	price	freight_value
8272b63d03f5f79c56e9e4120aec44ef	1	270516a3f41dc035aa87d220228f844c	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	2	05b515fdc76e888aada3c6d66c201dff	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	3	05b515fdc76e888aada3c6d66c201dff	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	4	05b515fdc76e888aada3c6d66c201dff	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	5	05b515fdc76e888aada3c6d66c201dff	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	6	05b515fdc76e888aada3c6d66c201dff	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	7	05b515fdc76e888aada3c6d66c201dff	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	8	05b515fdc76e888aada3c6d66c201dff	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	9	05b515fdc76e888aada3c6d66c201dff	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	10	05b515fdc76e888aada3c6d66c201dff	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	11	05b515fdc76e888aada3c6d66c201dff	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	12	270516a3f41dc035aa87d220228f844c	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	13	270516a3f41dc035aa87d220228f844c	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	14	270516a3f41dc035aa87d220228f844c	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	15	270516a3f41dc035aa87d220228f844c	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	16	270516a3f41dc035aa87d220228f844c	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	17	270516a3f41dc035aa87d220228f844c	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	18	270516a3f41dc035aa87d220228f844c	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	19	270516a3f41dc035aa87d220228f844c	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	20	270516a3f41dc035aa87d220228f844c	21/07/2017 18:25	1.2	7.89
8272b63d03f5f79c56e9e4120aec44ef	21	79ce45dbc2ea29b22b5a261bbb7b7ee7	21/07/2017 18:25	7.8	6.57

Données retenues

Cardinalité : **112650**

Nombre max de
Lignes de commande / commande : **21**

Produits

product_id	category_name	weight_g	length_cm	height_cm	width_cm	volume
1e9e8ef04dbcff4541ed26657ea517e5	perfumaria	225.0	16.0	10.0	14.0	
3aa071139cb16b67ca9e5dea641aaa2f	artes	1000.0	30.0	18.0	20.0	
96bd76ec8810374ed1b65e291975717f	esporte_lazer	154.0	18.0	9.0	15.0	
cef67bcfe19066a932b7673e239eb23d	bebes	371.0	26.0	4.0	26.0	
9dc1a7de274444849c219cff195d0b71	utilidades_domesticas	625.0	20.0	17.0	13.0	

Cardinalité : **32951**

Données retenues

\+ length_cm * height_cm * width_cm = volume

Nombre de catégories de produit : **74**

Paiements

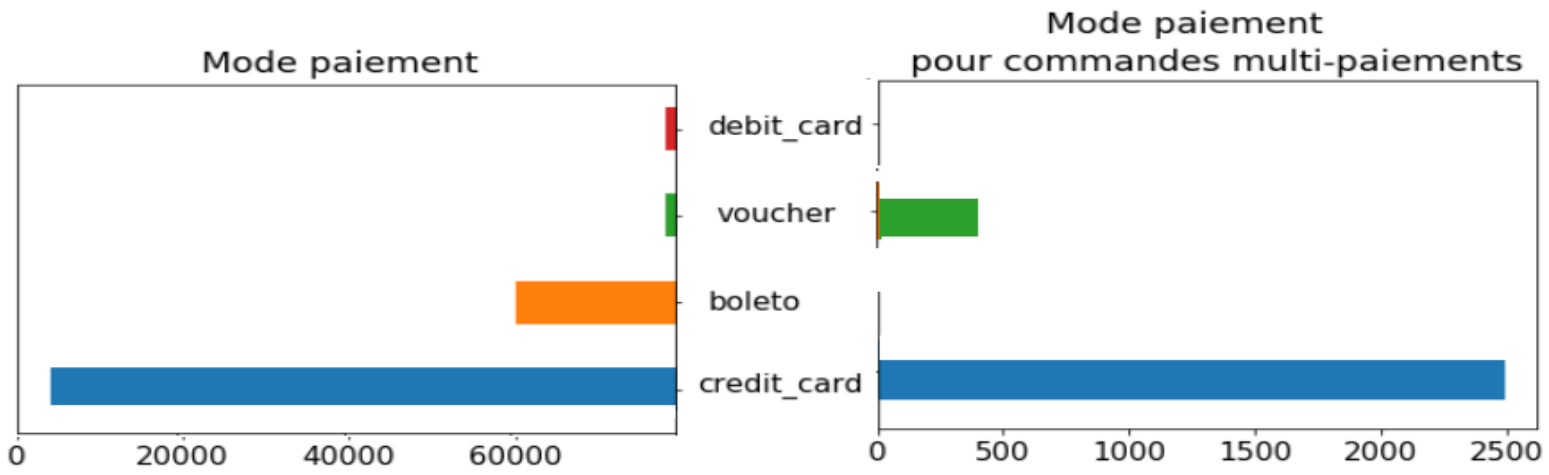
order_id	payment_sequential	payment_type	payment_installments	payment_value
b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	8	99.33
a9810da82917af2d9aefd1278f1dcfa0	1	debit_card	1	24.39
25e8ea4e93396b6fa0d3dd708e76c1bd	1	boleto	1	65.71
ba78997921bbcdc1373bb41e913ab953	1	voucher	8	107.78
42fdf880ba16b47b59251dd489d4441a	1	not_defined	2	128.45

Données retenues

Cardinalité : **103886**

Nbre commandes : **2920 / 98187** multi-paiements

Nbre paiements Max : **29**



Encodage du type de paiement

order_id	nb_payments	payment_type_boleto	payment_type_credit_card	payment_type_debit_card	payment_type_voucher
00010242fe8c5a6d1ba2dd792cb16214	1	0.0	1.0	0.0	0.0
00018f77f2f0320c557190d7a144bdd3	1	1.0	0.0	0.0	0.0
000229ec398224ef6ca0657da4fc703e	1	0.0	0	1.0	0.0
00024acbcdf0a6daa1e931b038114c75	1	0.0	0	0.0	1.0
00042b26cf59d7ce69dfabb4e55b4fd9	1	0.0	1.0	0.0	0.0

Construction du Dataset principal

order_id	price	freight	weight	volume
0008288aa423d2a3f00fcb17cd7d8719	49.9	13.37	1650.0	19800.0
0008288aa423d2a3f00fcb17cd7d8719	49.9	13.37	1650.0	19800.0

lignes de commande

Regroupement par sommation

order_id	price	freight	weight	volume
0008288aa423d2a3f00fcb17cd7d8719	99.8	26.74	3300.0	39600.0

commande

+ caractéristiques paiements

nb_payments	payment_type_boleto	payment_type_credit_card	payment_type_debit_card	payment_type_voucher
1	1.0	0.0	0.0	0.0

+ Préparation calcul fréquence

frequence	customer_id
1	9e4159995424971423b98c4a8bc11529

Construction du Dataset principal (suite)

customer_id	price	freight	weight	volume	nb_payments	frequence	type_boleto	type_credit_card	type_debit_card	type_voucher
b5b46581465f851df3529a30d2503f59	359.98	52.2	9200.0	114000.0	1	1	0.0	1.0	0.0	0.0
b5b46581465f851df3529a30d2503f59	51.98	32.22	1000.0	10240.0	1	1	0.0	1.0	0.0	0.0

Regroupement des commandes par Client

moy, min, max, moy, moy, moy, moy, sum, moy, moy, moy, moy

customer_id	price	price_min	price_max	freight	weight	volume	nb_payments	frequence	boleto	credit_card	debit_card	voucher
b5b46581465f851df3529a30d2503f59	205.98	51.98	359.98	42.21	5100.0	62120.0	1	2	0.0	1.0	0.0	0.0

+

recence
336

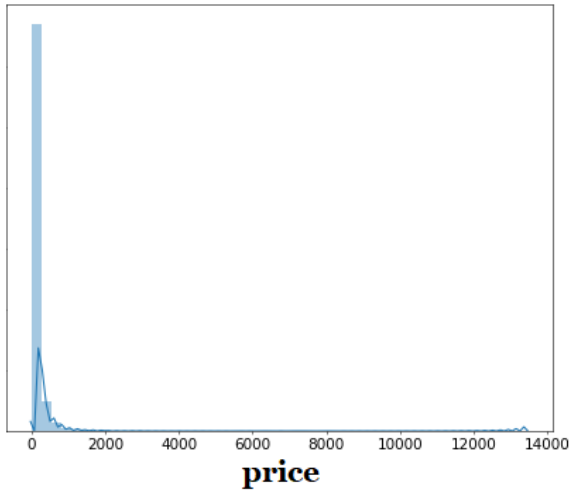
Caractéristiques retenues

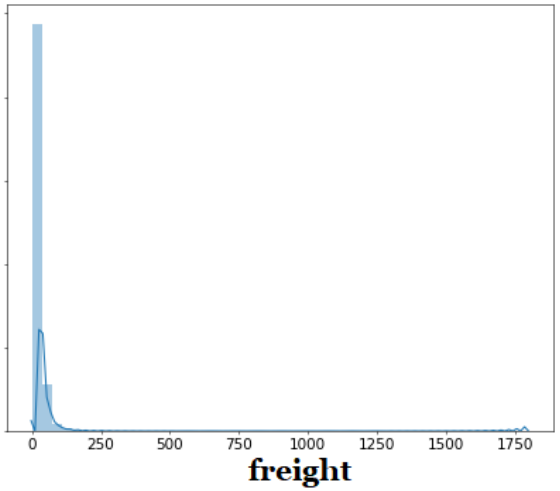
price	price_min	price_max	freight	weight	volume	nb_payments	frequence	recence	boleto	credit_card	debit_card	voucher
129.9	129.9	129.9	12.0	1500.0	7616.0	1.0	1	118	0.0	1.0	0.0	0.0
18.9	18.9	18.9	8.29	375.0	5148.0	1.0	1	121	0.0	1.0	0.0	0.0
69.0	69.0	69.0	17.22	1500.0	43750.0	1.0	1	544	0.0	1.0	0.0	0.0
25.99	25.99	25.99	17.63	150.0	1045.0	1.0	1	328	0.0	1.0	0.0	0.0
180.0	180.0	180.0	16.89	6050.0	528.0	1.0	1	295	0.0	1.0	0.0	0.0
154.0	154.0	154.0	12.98	3000.0	8241.0	1.0	1	153	0.0	1.0	0.0	0.0

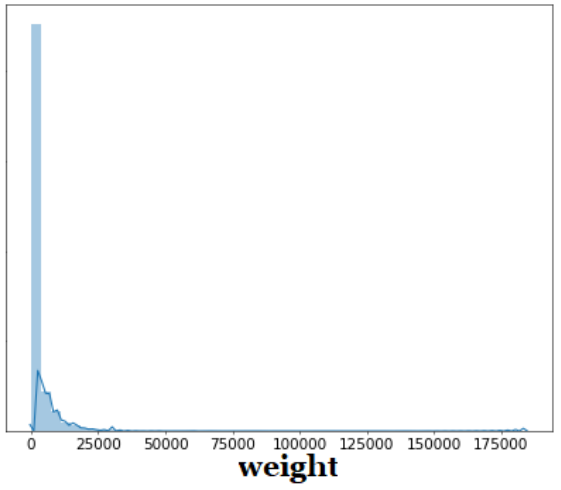
Statistiques Produits

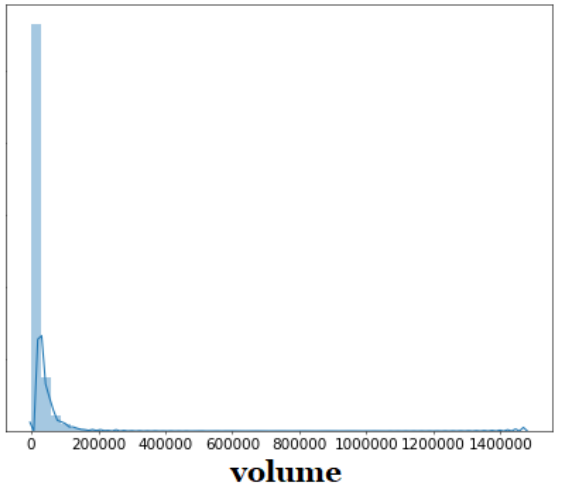
	price	price_min	price_max	freight	weight	volume	nb_payments	frequence	recence	boleto	credit_card	debit_card	voucher
count	94956.0	94956.0	94956.0	94956.0	94956.0	94956.0	94956.0	94956.0	94956.0	94956.0	94956.0	94956.0	94956.0
mean	137.91	136.58	139.28	22.82	2392.22	17397.24	1.04	1.03	244.74	0.2	0.77	0.02	0.02
median	87,2	85	89	17,25	750	7427	1	1	225	0	1	0	0
min	0.85	0.85	0.85	0.0	0.0	168.0	1.0	1.0	2.0	0.0	0.0	0.0	0.0
max	13440.0	13440.0	13440.0	1794.96	184400.0	1476000.0	29.0	16.0	701.0	1.0	1.0	1.0	1.0

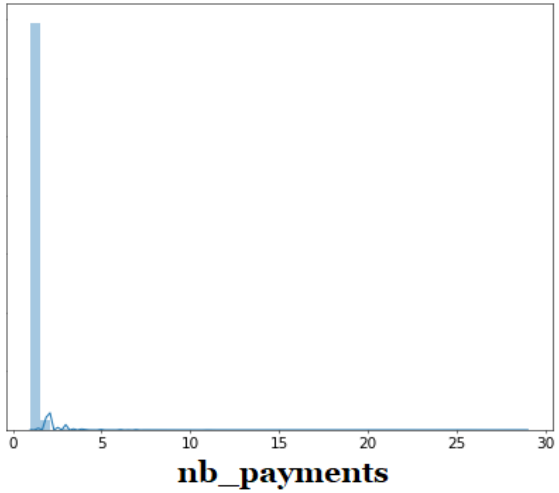
Analyse univariée

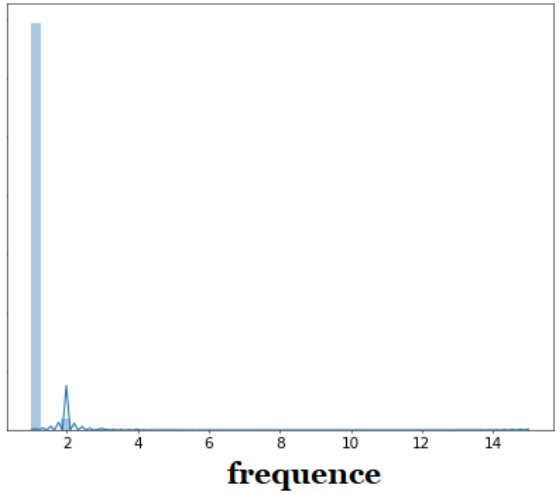


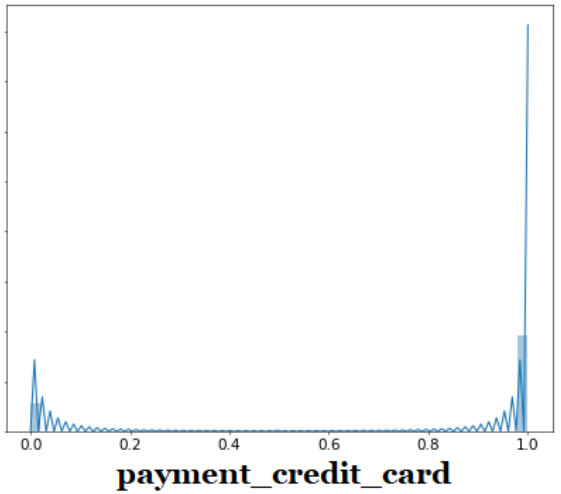


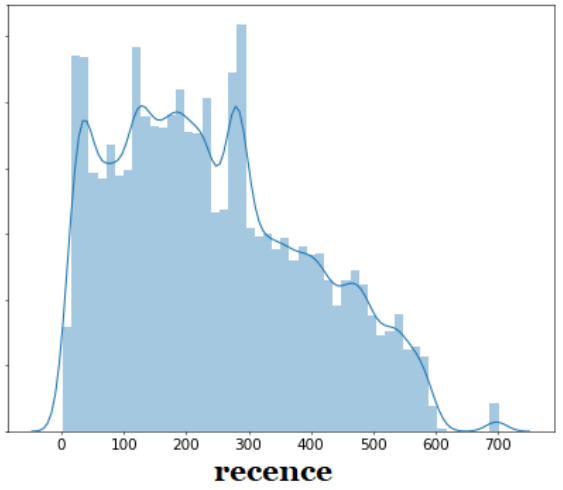








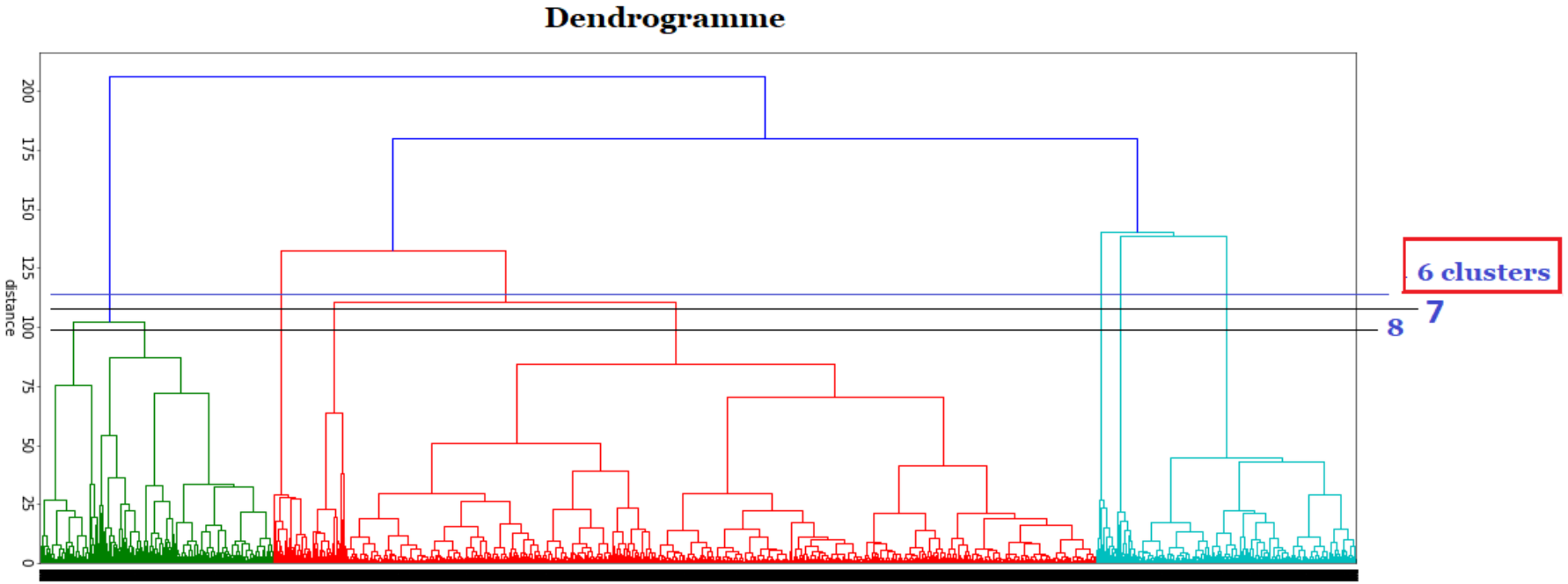


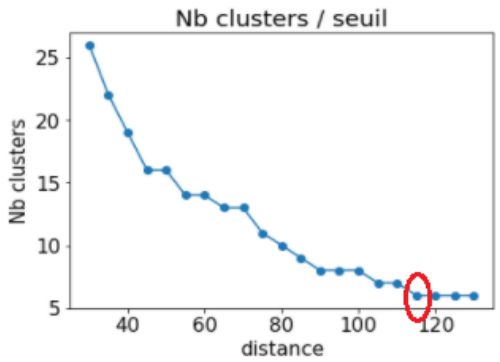


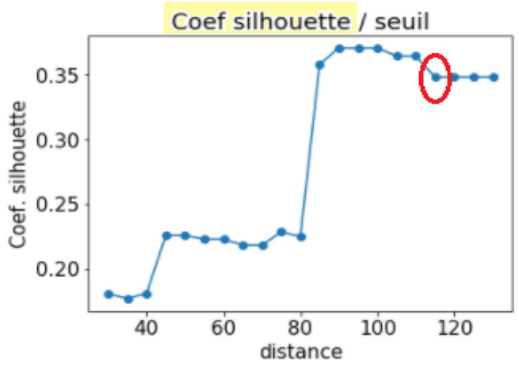
Clustering

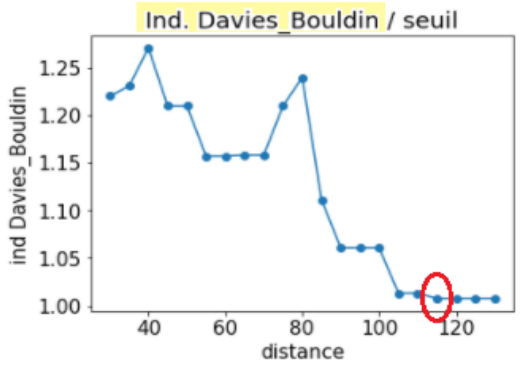
- Classification hiérarchique ascendante (CAH)
- K-means
- DBSCAN

Clustering avec CAH

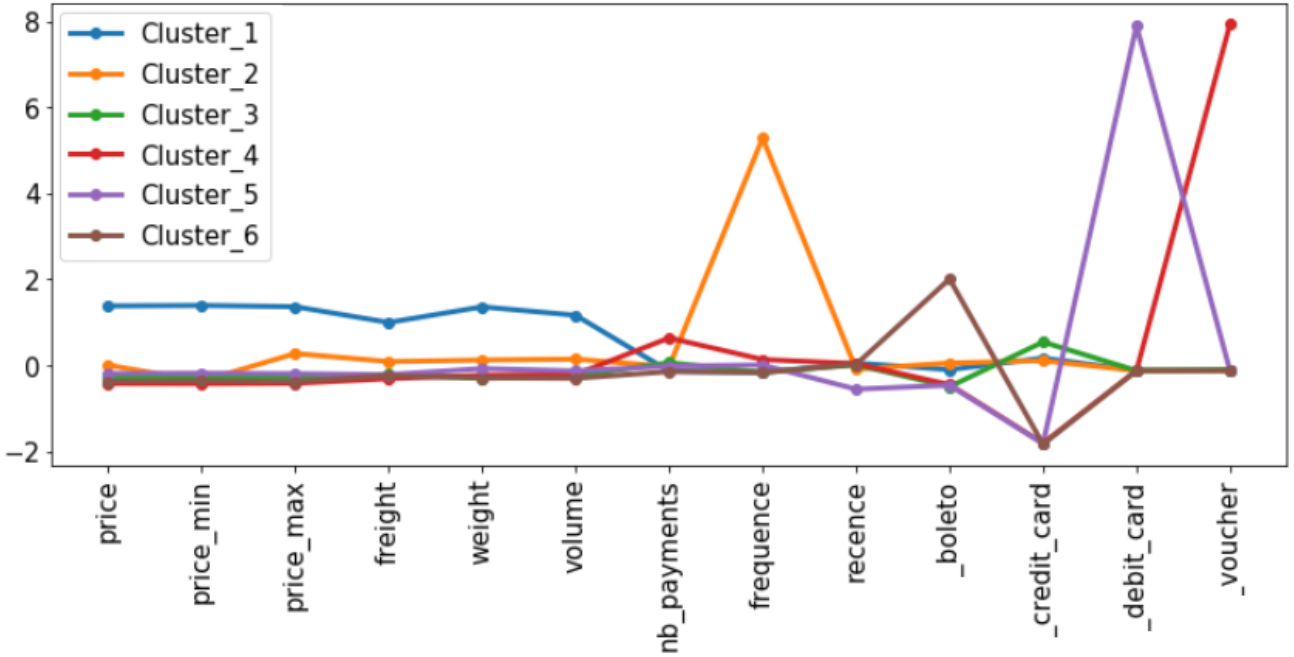


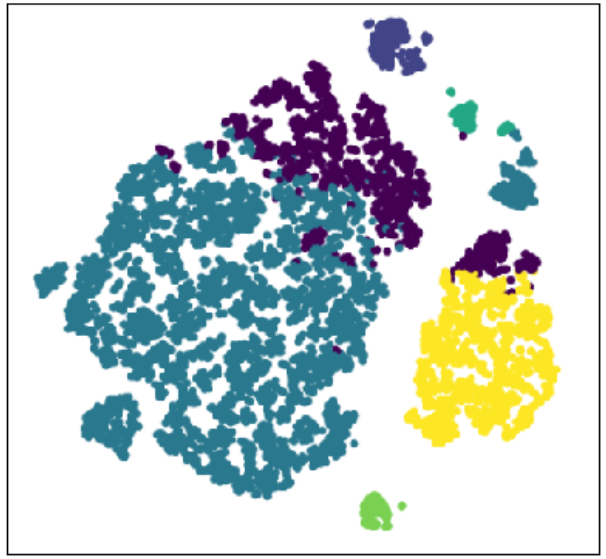






Clustering - CAH(2)

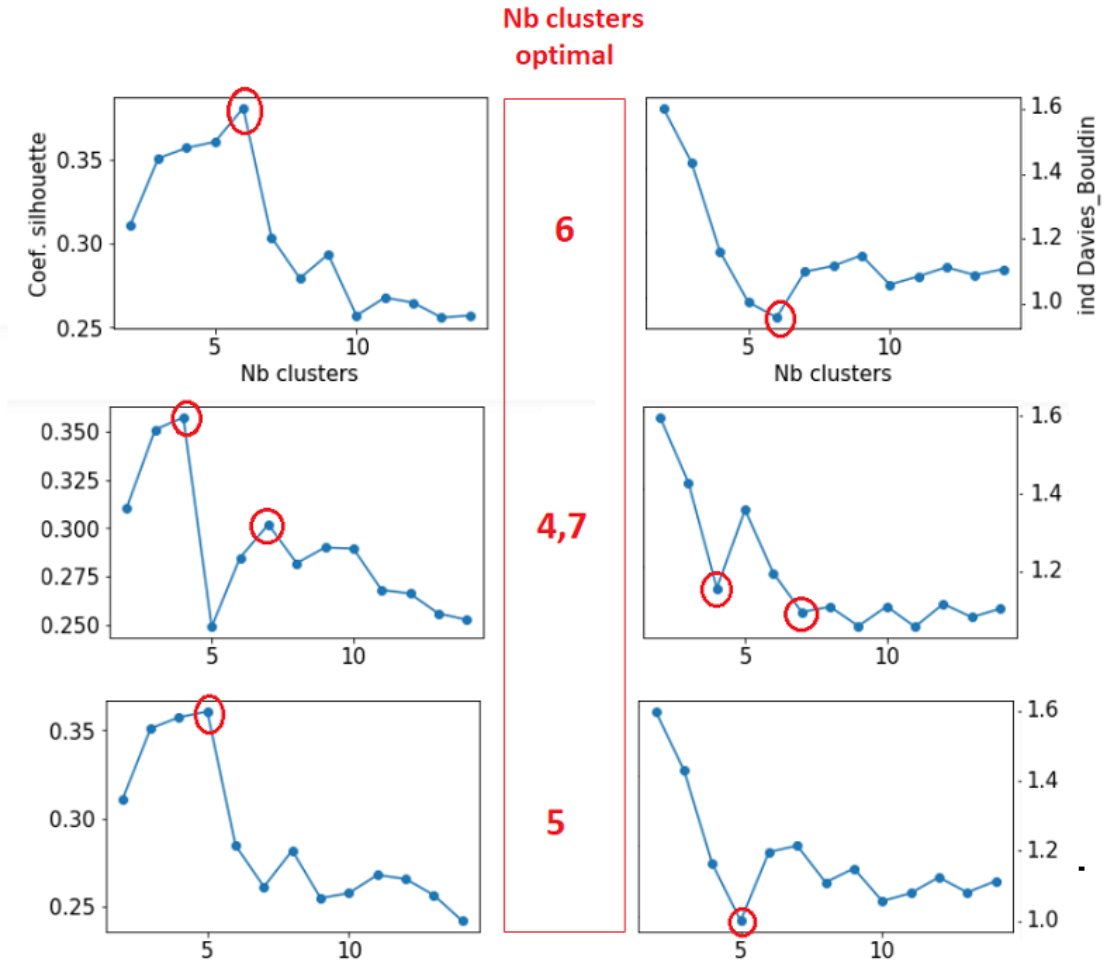


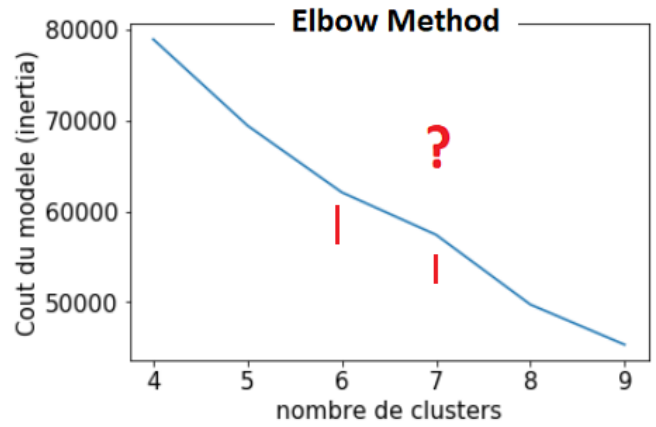


Représentation t-SNE des clients

cluster	price	price_min	price_max	freight	weight	volume	nb_payments	frequence	recence	boleto	credit_card	debit_card	voucher	nb_clients
Cluster_1	389.84	389.84	389.84	42.31	8188.4	49793.02	1.0	1.0	252.77	0.16	0.84	0.0	0.0	**1772**
Cluster_2	124.14	79.64	169.09	23.5	2605.35	18844.45	1.03	2.08	226.89	0.2	0.8	0.0	0.0	**297**
Cluster_3	87.46	87.45	87.46	18.83	1088.99	10197.11	1.06	1.0	246.24	0.0	1.0	0.0	0.0	**5962**
Cluster_4	75.25	74.28	76.29	17.38	1472.58	12061.47	1.21	1.06	249.82	0.01	0.01	0.0	0.98	**150**
Cluster_5	110.5	109.94	111.06	18.97	2119.35	14789.04	1.03	1.03	169.09	0.01	0.0	0.99	0.0	**148**
Cluster_6	77.73	77.73	77.73	18.26	1146.0	9396.01	1.0	1.0	248.26	1.0	0.0	0.0	0.0	**1671**

Clustering avec Kmeans

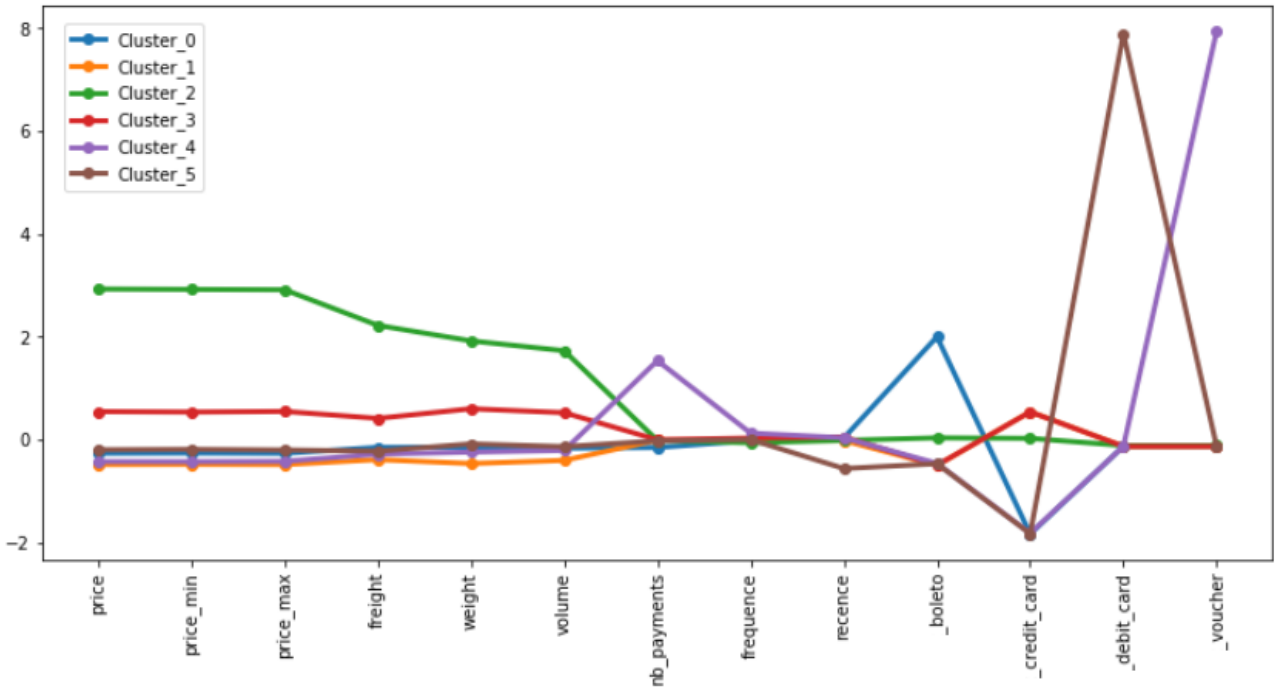


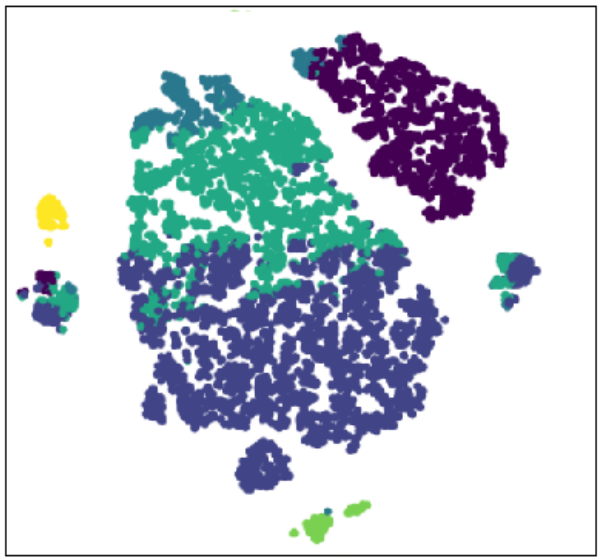


20 mesures du silhouette score pour un nombre de clusters compris dans [5, 6, 7, 8]

```
[0.2373407319449513, 0.2780404960669733, 0.27779395947915264, 0.2780404960669733]
[0.2811112131399136, 0.2811112131399136, 0.27994158708927286, 0.2780404960669733]
[0.2811112131399136, 0.2811112131399136, 0.2780404960669733, 0.31040477038813424]
[0.2780404960669733, 0.27826601718405314, 0.27832469471454085, 0.2811112131399136]
[0.2783607085856162, 0.27831364448468887, 0.2780404960669733, 0.2781979308478891]
[0.2809269463183091, 0.2811112131399136, 0.2801714141154879, 0.27779395947915264]
[0.2780404960669733, 0.2801714141154879, 0.2780404960669733, 0.2780404960669733]
[0.2780404960669733, 0.2780404960669733, 0.2780404960669733, 0.2780404960669733]
[0.2780404960669733, 0.2780404960669733, 0.2780404960669733, 0.2780404960669733]
[0.2780404960669733, 0.27779395947915264, 0.2780404960669733, 0.2780404960669733]
[0.2780404960669733, 0.2780404960669733, 0.28110432350191217, 0.280573409337362]
[0.2780404960669733, 0.2780404960669733, 0.2780404960669733, 0.28049156090504523]
[0.28002573001855624, 0.27984360682953363, 0.3121013974628259, 0.2780404960669733]
[0.2780404960669733, 0.2780404960669733, 0.2780404960669733, 0.2780404960669733]
[0.2791321916870242, 0.2780404960669733, 0.2780404960669733, 0.2811112131399136]
[0.2780404960669733, 0.2780404960669733, 0.28107769001216726, 0.27872799356163036]
[0.2780404960669733, 0.2785800985431838, 0.2780404960669733, 0.2811112131399136]
[0.28107769001216726, 0.23770816035240458, 0.28021006804363263, 0.2780404960669733]
[0.2811112131399136                                                  91434590009306]
[0.2809653317515452   Meilleur nombre de clusters : 6                811111213139136]
```

Clustering avec Kmeans(2)



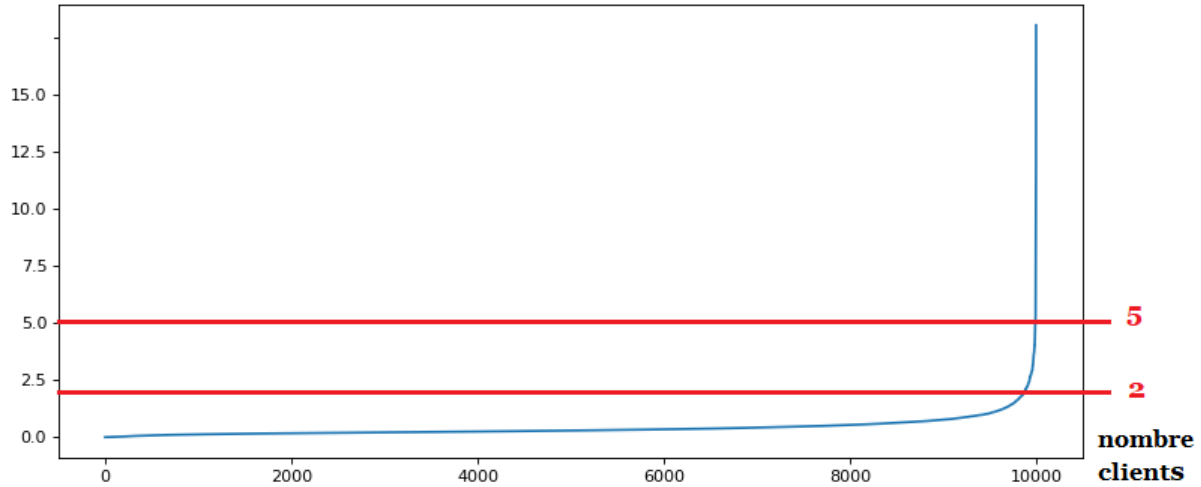


Représentation t-SNE des clients

cluster	price	price_min	price_max	freight	weight	volume	nb_payments	frequence	recence	boleto	credit_card	debit_card	voucher	nb_clients
Cluster_0	92.91	92.07	93.77	19.77	1747.11	12777.73	1.0	1.03	251.28	1.0	0.0	0.0	0.0	**1895**
Cluster_1	64.43	63.84	65.02	16.15	703.0	7674.54	1.04	1.03	242.63	0.0	1.0	0.0	0.0	**4681**
Cluster_2	794.6	790.71	798.49	72.9	12464.68	77650.53	1.05	1.02	244.61	0.21	0.78	0.0	0.0	**509**
Cluster_3	190.48	187.82	193.16	28.53	4126.57	27099.34	1.04	1.04	252.9	0.0	1.0	0.0	0.0	**2599**
Cluster_4	77.99	77.08	78.97	17.93	1434.76	12152.07	1.53	1.06	249.69	0.01	0.01	0.0	0.98	**168**
Cluster_5	110.5	109.94	111.06	18.97	2119.35	14789.04	1.03	1.03	169.09	0.01	0.0	0.99	0.0	**148**

Clustering avec DBSCAN

ε optimal

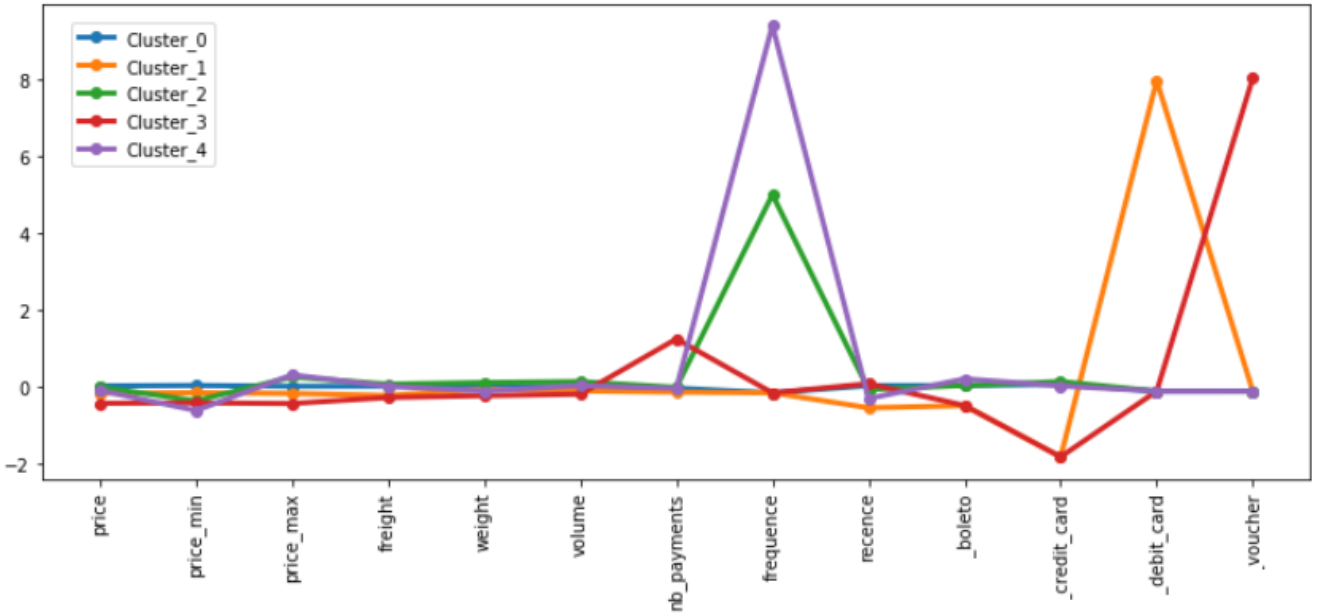


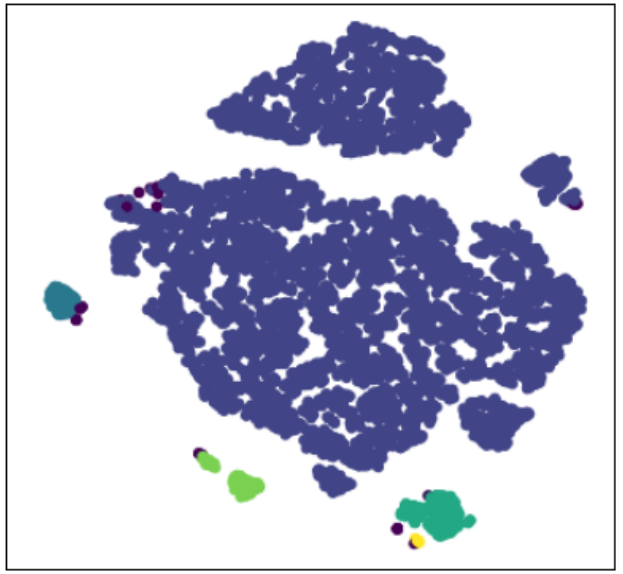
eps	nb clusters	Coef. Silhouette moyen
2	13	0.3559
3	7	0.3577
3.1	7	0.3602
3.2	7	0.3608
3.3	7	0.3611
3.4	7	0.3611
3.5	6	0.3272
3.6	6	0.3268
3.7	7	0.3242
3.8	5	0.3921
3.9	5	0.3937
4	4	0.4130
5	4	0.4148

'min_samples' optimal

min_samples	Nb clusters
5	7
6	6
7	6
8	5
9	5
10	5
15	4
20	4

Clustering avec DBSCAN (2)



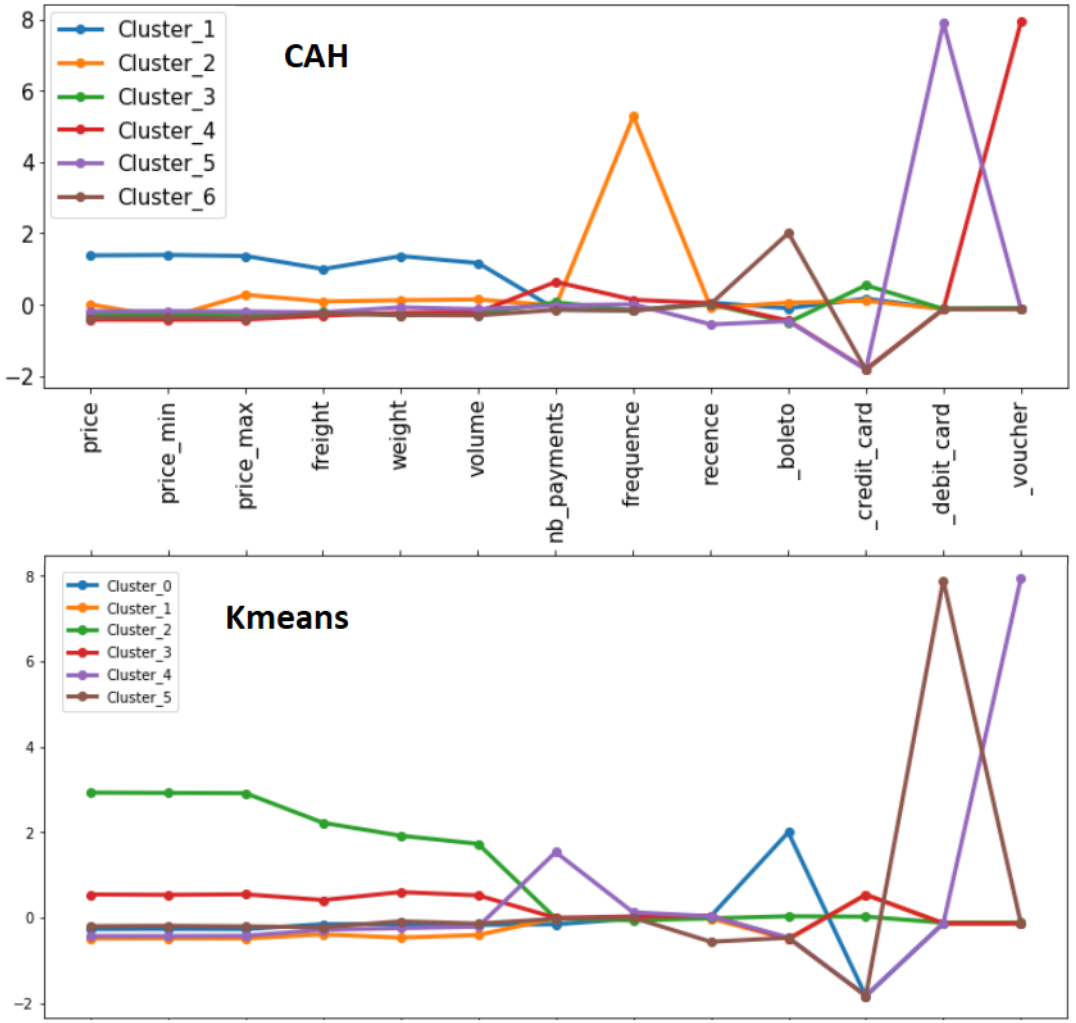


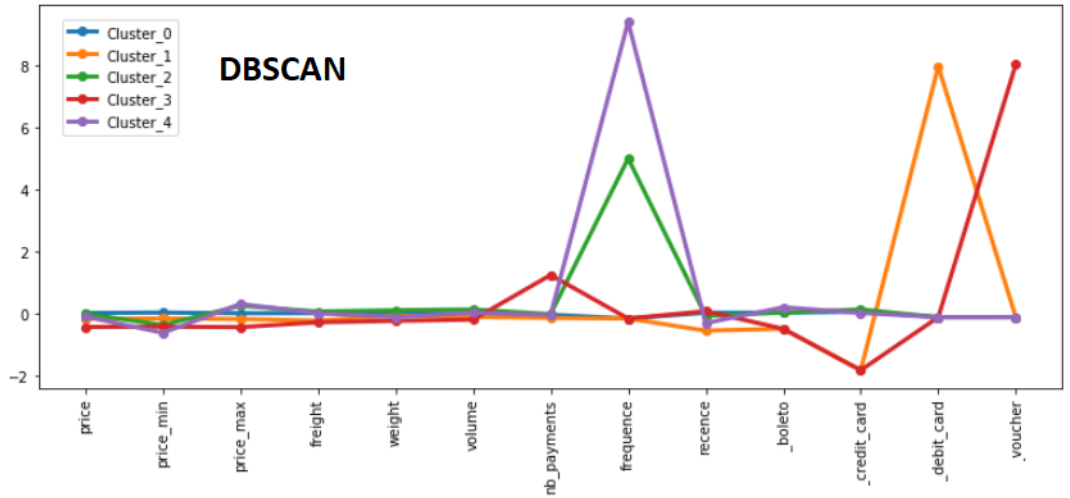
Représentation t-SNE des clients

cluster	price	price_min	price_max	freight	weight	volume	nb_payments	frequence	recence	boleto	credit_card	debit_card	voucher	nb_clients
Cluster_0	139.84	139.84	139.84	22.83	2389.24	17132.03	1.03	1.0	247.81	0.21	0.79	0.0	0.0	**9367**
Cluster_1	112.46	112.46	112.46	18.69	2157.92	15203.4	1.0	1.0	170.01	0.0	0.0	1.0	0.0	**138**
Cluster_2	120.16	77.21	163.11	22.87	2422.34	18037.22	1.03	2.0	228.31	0.19	0.81	0.0	0.0	**277**
Cluster_3	74.1	74.1	74.1	17.66	1420.94	12209.02	1.4	1.0	254.79	0.0	0.0	0.0	1.0	**154**
Cluster_4	105.22	50.13	165.12	21.17	1462.38	14338.95	1.02	3.13	193.27	0.24	0.76	0.0	0.0	**15**

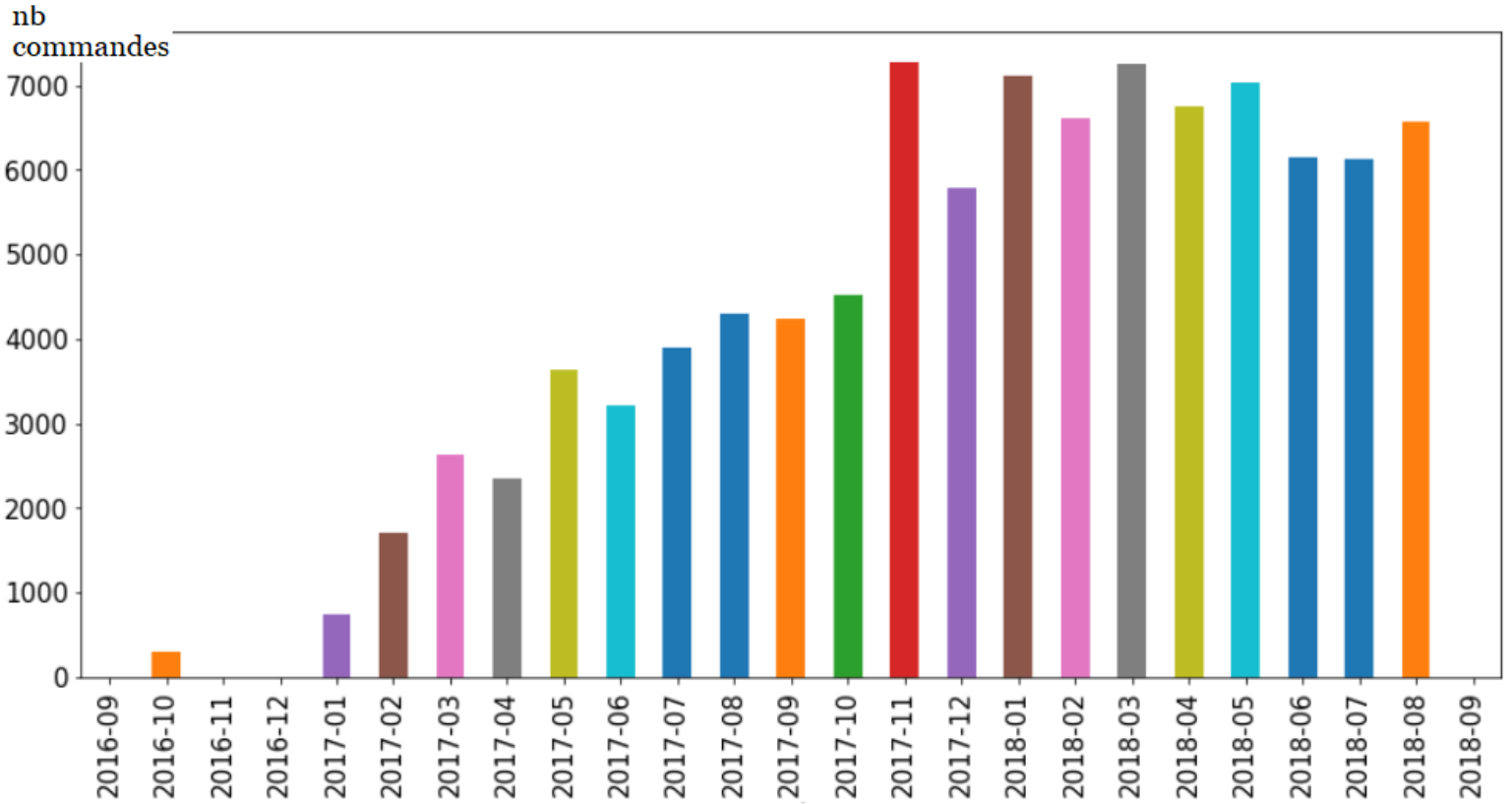
Estimated number of noise points : 49
Silhouette Coefficient : 0.444

Clustering comparatif





Evaluation comparative des 3 méthodes de Clustering: Principes



période 1 →

période 2 →

période 3 →

. . . . . .

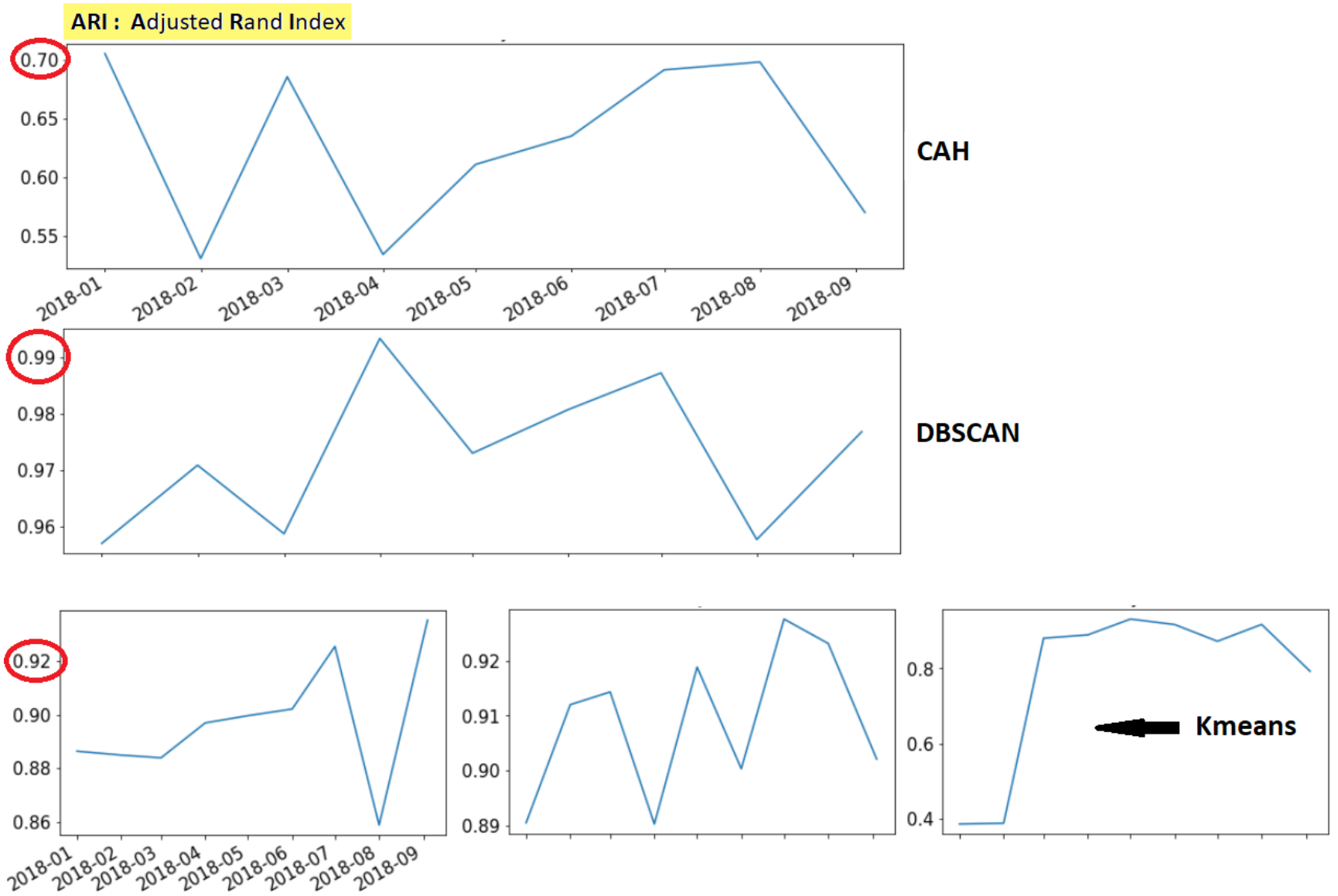
période 10 →

Principe

- **Clustering de clients multi-commandes sur 10 périodes.**
- **Clients identiques (1668 clients) sur les 10 périodes.**

période(n+1) = période(n) + 1 mois

Stabilité des partitions de Mois en Mois



Période de validité des segments