EA Stephen
Data scientist

Segmentez des clients d'un site e-commerce

olist

SOMMAIRE

- Présentation de la problématique, du cleaning effectué, du feature engineering et de l'exploration
- Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné
- Présentation de la simulation pour définir le délai de maintenance du modèle (contrat de maintenance)

La mission

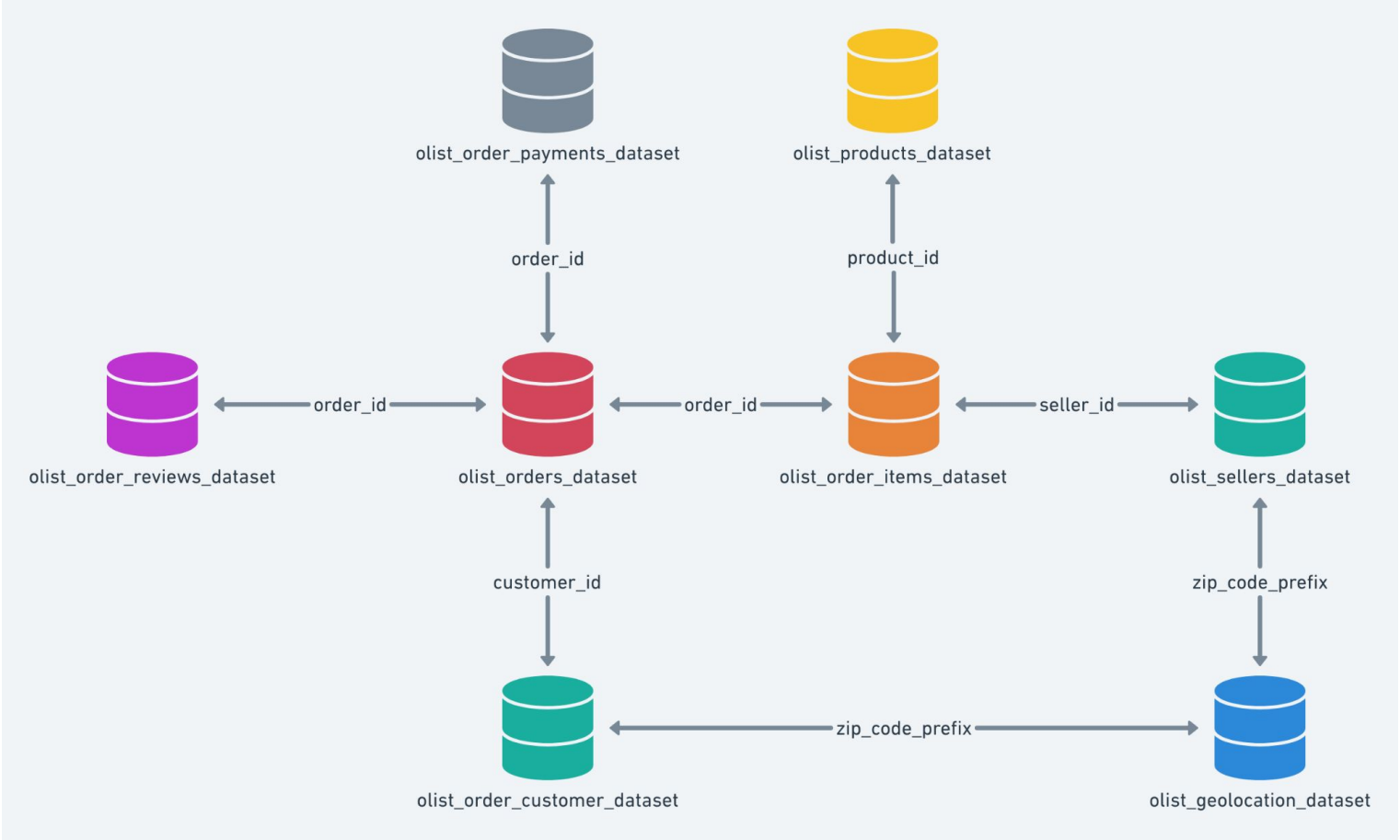
Olist (solution de vente sur les marketplaces en ligne souhaite fournir à ses équipes d'e-commerce)

Une segmentation des clients pour leurs campagnes de communication.
La segmentation proposée doit être exploitable et facile d'utilisation pour l'équipe marketing

Une recommandation de fréquence à laquelle la segmentation doit être mise à jour pour rester pertinente, afin de pouvoir effectuer un devis de contrat de maintenance.

olist

Le schéma de donnée



olist

Les produits

Le dataset : DF_PRODUCT_CATEGORY
Les dimensions du df_product_category : (71, 2)
Les valeurs manquantes dans chaque colonne :

	Count_NaN	%_NaN_col	Types	Total_NaN_in_dataset	%_NaN_in_dataset
product_category_name	0	0.0	object	0	0.0
product_category_name_english	0	0.0	object	0	0.0

	product_category_name	product_category_name_english
0	beleza_saude	health_beauty

Jointure left

	product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm
0	1e9e8ef04dbcff4541ed26657ea517e5	perfumaria	40.0	287.0	1.0	225.0	16.0	10.0	14.0

df_product_final

32951 rows × 10 columns

Le dataset : DF_PRODUCTS
Les dimensions du df_products : (32951, 9)
Les valeurs manquantes dans chaque colonne :

	Count_NaN	%_NaN_col	Types	Total_NaN_in_dataset	%_NaN_in_dataset
product_category_name	610	1.85	object	2448	0.83
product_name_lenght	610	1.85	float64	2448	0.83
product_description_lenght	610	1.85	float64	2448	0.83
product_photos_qty	610	1.85	float64	2448	0.83
product_weight_g	2	0.01	float64	2448	0.83
product_length_cm	2	0.01	float64	2448	0.83
product_height_cm	2	0.01	float64	2448	0.83
product_width_cm	2	0.01	float64	2448	0.83
product_id	0	0.0	object	2448	0.83

olist

Fusion entre customers et orders

df_customers

Shape:(99441,5)

df_orders

Shape: 96478,8
Orders_status: delivered

Jointure inner

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state	order_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date
0	06b8999e2fba1a1fbc88172c00ba8bc7	1	14409	franca	SP	00e7ee1b050b8499577073aeb2a297a1	delivered	2017-05-16 15:05:35	2017-05-16 15:22:12	2017-05-23 10:47:57	2017-05-25 10:35:35	2017-06-05 00:00:00
1	18955e83d337fd6b2def6b18a428ac77	2	9790	sao bernardo do campo	SP	29150127e6685892b6eab3eec79f59c7	delivered	2018-01-12 20:48:24	2018-01-12 20:58:32	2018-01-15 17:14:59	2018-01-29 12:41:19	2018-02-06 00:00:00
2	4e7b3e00288586ebd08712fdd0374a03	3	1151	sao paulo	SP	b2059ed67ce144a36e2aa97d2c9e9ad2	delivered	2018-05-19 16:07:45	2018-05-20 16:19:10	2018-06-11 14:31:00	2018-06-14 17:58:51	2018-06-13 00:00:00
3	b2b6027bc5c5109e529d4dc6358b12c3	4	8775	mogi das cruzes	SP	951670f92359f4fe4a63112aa7306eba	delivered	2018-03-13 16:06:38	2018-03-13 17:29:19	2018-03-27 23:22:42	2018-03-28 16:04:25	2018-04-10 00:00:00
4	4f2d8ab171c80ec8364f7c12e35b23ad	5	13056	campinas	SP	6b7d50bd145f6fc7f33cebabd7e49d0f	delivered	2018-07-29 09:51:30	2018-07-29 10:10:09	2018-07-30 15:16:00	2018-08-09 20:55:48	2018-08-15 00:00:00
...	...	...	...	...	...	...	...	...	...	...	...	...
96473	17ddf5dd5d51696bb3d7c6291687be6f	96092	3937	sao paulo	SP	6760e20addcf0121e9d58f2f1ff14298	delivered	2018-04-07 15:48:17	2018-04-07 16:08:45	2018-04-11 02:08:36	2018-04-13 20:06:37	2018-04-25 00:00:00
96474	e7b71a9017aa05c9a7fd292d714858e8	96093	6764	taboao da serra	SP	9ec0c8947d973db4f4e8dcf1fbfa8f1b	delivered	2018-04-04 08:20:22	2018-04-04 08:35:12	2018-04-05 18:42:35	2018-04-11 18:54:45	2018-04-20 00:00:00
96475	5e28dfe12db7fb50a4b2f691faecea5e	96094	60115	fortaleza	CE	fed4434add09a6f332ea398efd656a5c	delivered	2018-04-08 20:11:50	2018-04-08 20:30:03	2018-04-09 17:52:17	2018-05-09 19:03:15	2018-05-02 00:00:00
96476	56b18e2166679b8a959d72dd06da27f9	96095	92120	canoas	RS	e31ec91cea1ecf97797787471f98a8c2	delivered	2017-11-03 21:08:33	2017-11-03 21:31:20	2017-11-06 18:24:41	2017-11-16 19:58:39	2017-12-05 00:00:00
96477	274fa6071e5e17fe303b9748641082c8	96096	6703	cotia	SP	28db69209a75e59f20ccbb5c36a20b90	delivered	2017-12-19 14:27:23	2017-12-19 18:50:39	2017-12-21 19:17:21	2017-12-26 18:42:36	2018-01-08 00:00:00

96478 rows × 12 columns

df_customers_orders

Le nombres de clients uniques: 93358
Le nombres de références clients: 96478
Le status des commandes: ['delivered']

olist

Fusion entre les commandes clients et le moyen de paiement

df_customers_orders
(96478, 12)

df_order_payments
(103886, 5)

Jointure left

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
0	06b8999e2fba1a1fbc88172c00ba8bc7	1	14409	franca	SP
1	18955e83d337fd6b2def6b18a428ac77	2	9790	sao bernardo do campo	SP
2	4e7b3e00288586ebd08712fdd0374a03	3	1151	sao paulo	SP
3	b2b6027bc5c5109e529d4dc6358b12c3	4	8775	mogi das cruzes	SP
4	4f2d8ab171c80ec8364f7c12e35b23ad	5	13056	campinas	SP
...	...	...	...	...	...
100752	17ddf5dd5d51696bb3d7c6291687be6f	96092	3937	sao paulo	SP
100753	e7b71a9017aa05c9a7fd292d714858e8	96093	6764	taboao da serra	SP
100754	5e28dfe12db7fb50a4b2f6911aecea5e	96094	60115	fortaleza	CE
100755	56b18e2166679b8a959d72dd06da27f9	96095	92120	canoas	RS
100756	274fa6071e5e17fe303b9748641082c8	96096	6703	cotia	SP

100757 rows × 16 columns

order_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date	payment_sequential	payment_type	payment_installments	payment_value
00e7ee1b050b8499577073aeb2a297a1	delivered	2017-05-16 15:05:35	2017-05-16 15:22:12	2017-05-23 10:47:57	2017-05-25 10:35:35	2017-06-05 00:00:00	1.0	credit_card	2.0	146.87
29150127e6685892b6eab3eec79f59c7	delivered	2018-01-12 20:48:24	2018-01-12 20:58:32	2018-01-15 17:14:59	2018-01-29 12:41:19	2018-02-06 00:00:00	1.0	credit_card	8.0	335.48
b2059ed67ce144a36e2aa97d2c9e9ad2	delivered	2018-05-19 16:07:45	2018-05-20 16:19:10	2018-06-11 14:31:00	2018-06-14 17:58:51	2018-06-13 00:00:00	1.0	credit_card	7.0	157.73
951670f92359f4fe4a63112aa7306eba	delivered	2018-03-13 16:06:38	2018-03-13 17:29:19	2018-03-27 23:22:42	2018-03-28 16:04:25	2018-04-10 00:00:00	1.0	credit_card	1.0	173.30
6b7d50bd145f6fc7f33cebabd7e49d0f	delivered	2018-07-29 09:51:30	2018-07-29 10:10:09	2018-07-30 15:16:00	2018-08-09 20:55:48	2018-08-15 00:00:00	1.0	credit_card	8.0	252.25
...	...	...	...	...	...	...	...	...	...	...
6760e20addcf0121e9d58f2f1ff14298	delivered	2018-04-07 15:48:17	2018-04-07 16:08:45	2018-04-11 02:08:36	2018-04-13 20:06:37	2018-04-25 00:00:00	1.0	credit_card	6.0	88.78
9ec0c8947d973db4f4e8dcf11bfa8f1b	delivered	2018-04-04 08:20:22	2018-04-04 08:35:12	2018-04-05 18:42:35	2018-04-11 18:54:45	2018-04-20 00:00:00	1.0	credit_card	3.0	129.06
fed4434add09a6f332ea398efd6556a5c	delivered	2018-04-08 20:11:50	2018-04-08 20:30:03	2018-04-09 17:52:17	2018-05-09 19:03:15	2018-05-02 00:00:00	1.0	credit_card	5.0	56.04
e31ec91cea1ecf97797787471f98a8c2	delivered	2017-11-03 21:08:33	2017-11-03 21:31:20	2017-11-06 18:24:41	2017-11-16 19:58:39	2017-12-05 00:00:00	1.0	credit_card	2.0	711.07
28db69209a75e59f20ccbb5c36a20b90	delivered	2017-12-19 14:27:23	2017-12-19 18:50:39	2017-12-21 19:17:21	2017-12-26 18:42:36	2018-01-08 00:00:00	1.0	credit_card	1.0	21.77

payment_sequential	payment_type	payment_installments	payment_value
...	...	...	...

df_customers_orders_payments

100757 rows × 16 columns

```
Le nombres de clients uniques:  93358
Le nombres de références clients:  96478
```

olist

Fusion entre les commandes et la satisfaction client

df_customers_orders_payments
(100757, 16)

df_order_reviews
(99224, 7)

Jointure inner

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state	order_id
0	06b8999e2fba1a1fbc88172c00ba8bc7	1	14409	franca	SP	00e7ee1b050b8499577073aeb2a297a1
1	18955e83d337fd6b2def6b18a428ac77	2	9790	sao bernardo do campo	SP	29150127e6685892b6eab3eec79f59c7
2	4e7b3e00288586ebd08712fdd0374a03	3	1151	sao paulo	SP	b2059ed67ce144a36e2aa97d2c9e9ad2
3	b2b6027bc5c5109e529d4dc6358b12c3	4	8775	mogi das cruzes	SP	951670f92359f4fe4a63112aa7306eba
4	4f2d8ab171c80ec8364f7c12e35b23ad	5	13056	campinas	SP	6b7d50bd145f6fc7f33cebabd7e49d0f
...	...	...	...	...	...	...
100645	17ddf5dd5d51696bb3d7c6291687be6f	96092	3937	sao paulo	SP	6760e20addcf0121e9d58f2f1ff14298
100646	e7b71a9017aa05c9a7fd292d714858e8	96093	6764	taboao da serra	SP	9ec0c8947d973db4f4e8dcf1fbfa8f1b
100647	5e28dfe12db7fb50a4b2f691faecea5e	96094	60115	fortaleza	CE	fed4434add09a6f332ea398efd656a5c
100648	56b18e2166679b8a959d72dd06da27f9	96095	92120	canoas	RS	e31ec91cea1ecf97797787471f98a8c2
100649	274fa6071e5e17fe303b9748641082c8	96096	6703	cotia	SP	28db69209a75e59f20ccbb5c36a20b90

	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date	payment_sequential	payment_type	payment_installments	payment_value
0	delivered	2017-05-16 15:05:35	2017-05-16 15:22:12	2017-05-23 10:47:57	2017-05-25 10:35:35	2017-06-05 00:00:00	1.0	credit_card	2.0	146.87
1	delivered	2018-01-12 20:48:24	2018-01-12 20:58:32	2018-01-15 17:14:59	2018-01-29 12:41:19	2018-02-06 00:00:00	1.0	credit_card	8.0	335.48
2	delivered	2018-05-19 16:07:45	2018-05-20 16:19:10	2018-06-11 14:31:00	2018-06-14 17:58:51	2018-06-13 00:00:00	1.0	credit_card	7.0	157.73
3	delivered	2018-03-13 16:06:38	2018-03-13 17:29:19	2018-03-27 23:22:42	2018-03-28 16:04:25	2018-04-10 00:00:00	1.0	credit_card	1.0	173.30
4	delivered	2018-07-29 09:51:30	2018-07-29 10:10:09	2018-07-30 15:16:00	2018-08-09 20:55:48	2018-08-15 00:00:00	1.0	credit_card	8.0	252.25
...	...	...	...	...	...	...	...	...	...	...
100645	delivered	2018-04-07 15:48:17	2018-04-07 16:08:45	2018-04-11 02:08:36	2018-04-13 20:06:37	2018-04-25 00:00:00	1.0	credit_card	6.0	88.78
100646	delivered	2018-04-04 08:20:22	2018-04-04 08:35:12	2018-04-05 18:42:35	2018-04-11 18:54:45	2018-04-20 00:00:00	1.0	credit_card	3.0	129.06
100647	delivered	2018-04-08 20:11:50	2018-04-08 20:30:03	2018-04-09 17:52:17	2018-05-09 19:03:15	2018-05-02 00:00:00	1.0	credit_card	5.0	56.04
100648	delivered	2017-11-03 21:08:33	2017-11-03 21:31:20	2017-11-06 18:24:41	2017-11-16 19:58:39	2017-12-05 00:00:00	1.0	credit_card	2.0	711.07
100649	delivered	2017-12-19 14:27:23	2017-12-19 18:50:39	2017-12-21 19:17:21	2017-12-26 18:42:36	2018-01-08 00:00:00	1.0	credit_card	1.0	21.77

	review_id	review_score	review_comment_title	review_comment_message	review_creation_date	review_answer_timestamp
0	8b8b52d46df026a9d1ad2136a59b30b	4	NaN	NaN	2017-05-26 00:00:00	2017-05-30 22:34:40
1	02fc48a9efa3e3d0f1a8ea26507eeec3	5	NaN	NaN	2018-01-30 00:00:00	2018-02-10 22:43:29
2	5ad6695d76ee186dc473c42706984d87	5	NaN	NaN	2018-06-15 00:00:00	2018-06-15 12:10:59
3	059a801bb31f6aab2266e672cab87bc5	5	NaN	NaN	2018-03-29 00:00:00	2018-04-02 18:36:47
4	8490879d58d6c5d7773f2739a03f089a	5	a melhor nota	O baratheon è esxelente Amo adoro o baratheon	2018-08-10 00:00:00	2018-08-17 01:59:52
...	...	...	...	...	...	...
100645	36e2cdbaa9f639b57c53b37ac798fee8	4	Recomendo	NaN	2018-04-14 00:00:00	2018-04-28 11:15:04
100646	b273b431c3aedb4eed18643309652940	5	NaN	NaN	2018-04-12 00:00:00	2018-04-16 10:36:05
100647	fa4f16891e6b2edd1354668d07f5648b	1	Chateada	Esperava qualidade no atendimento e estou tend...	2018-05-04 00:00:00	2018-05-08 01:41:07
100648	0bcdc9e450ea500811a8d39ee993cd47	5	NaN	NaN	2017-11-17 00:00:00	2017-11-19 23:34:18
100649	95d701d062d2bb88041768a63aa750d	5	NaN	Produto chegou bem antes do prazo , gostei mui...	2017-12-27 00:00:00	2017-12-27 23:21:06

100650 rows × 22 columns

df_customers_orders_payments_reviews

100650 rows × 22 columns

Le nombres de clients uniques: 92755
Le nombres de références clients: 95832

olist

Fusion entre les produits vendus par Olist et les catégories des produits

df_order_items

(112650, 7)

df_products

(32951, 10)

Jointure left

	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm	product_category_name_english
0	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	2017-09-19 09:45:35	58.90	13.29	cool_stuff	58.0	598.0	4.0	650.0	28.0	9.0	14.0	cool_stuff
1	00018f77f2f0320c557190d7a144bdd3	1	e5f2d52b802189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	2017-05-03 11:05:13	239.90	19.93	pet_shop	56.0	239.0	2.0	30000.0	50.0	30.0	40.0	pet_shop
2	000229ec398224ef6ca0657da4fc703e	1	c777355d18b72b67abbeef9df44fd0fd	5b51032eddd242adc84c38acab88f23d	2018-01-18 14:48:30	199.00	17.87	moveis_decoracao	59.0	695.0	2.0	3050.0	33.0	13.0	33.0	furniture_decor
3	00024acbcdf0a6daa1e931b038114c75	1	7634da152a4610f1595efa32f14722fc	9d7a1d34a5052409006425275ba1c2b4	2018-08-15 10:10:18	12.99	12.79	perfumaria	42.0	480.0	1.0	200.0	16.0	10.0	15.0	perfumery
4	00042b26cf59d7ce69dfabb4e55b4fd9	1	ac6c3623068f30de03045865e4e10089	df560393f3a51e74553ab94004ba5c87	2017-02-13 13:57:51	199.90	18.14	ferramentas_jardim	59.0	409.0	1.0	3750.0	35.0	40.0	30.0	garden_tools
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
112645	fffc94f6ce00a00581880bf54a75a037	1	4aa6014eceb68207f9dc4bffebc05b0	b8bc237ba3788b23da09c0f1f3a3288c	2018-05-02 04:11:01	299.99	43.41	utilidades_domesticas	43.0	1002.0	3.0	10150.0	89.0	15.0	40.0	housewares
112646	fffcd46ef2263f404302a634eb57f7eb	1	32e07fd915822b0765e448c4dd74c828	f3c38ab652836d21de61fb8314b69182	2018-07-20 04:31:48	350.00	36.53	informatica_acessorios	31.0	232.0	1.0	8950.0	45.0	26.0	38.0	computers_accessories
112647	fffce4705a9662cd70adb13d4a31832d	1	72a30483855e2eafc67aee5dc2560482	c3cfdc648177fdbbbb35635a37472c53	2017-10-30 17:14:25	99.90	16.95	esporte_lazer	43.0	869.0	1.0	967.0	21.0	24.0	19.0	sports_leisure
112648	fffe18544ffabc95dfada21779c9644f	1	9c422a519119dcad7575db5af1ba540e	2b3e4a2a3ea8e01938cabda2a3e5cc79	2017-08-21 00:04:32	55.99	8.72	informatica_acessorios	56.0	1306.0	1.0	100.0	20.0	20.0	20.0	computers_accessories
112649	fffe41c64501cc87c801fd61db3f6244	1	350688d9dc1e75ff97be326363655e01	f7ccf836d21b2fb1de37564105216cc1	2018-06-12 17:10:13	43.00	12.79	cama_mesa_banho	47.0	511.0	1.0	600.0	30.0	3.0	19.0	bed_bath_table

112650 rows × 16 columns

df_order_items_products

112650 rows × 16 columns

Fusion entre le client et les articles achetés

df_order_items_products

(112650, 16)

df_customers_orders_payments_reviews

(100650, 22)

Jointure left

	order_id	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm	product_category_name_english
0	00010242fe8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	2017-09-19 09:45:35	58.9	13.29	cool_stuff	58.0	598.0	4.0	650.0	28.0	9.0	14.0	cool_stuff

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state	order_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date	payment_sequential	payment_type	payment_installments	payment_value	review_id
0	06b8999e2fba1a1fbc88172c00ba8bc7	1	14409	franca	SP	00e7ee1b050b8499577073aeb2a297a1	delivered	2017-05-16 15:05:35	2017-05-16 15:22:12	2017-05-23 10:47:57	2017-05-25 10:35:35	2017-06-05 00:00:00	1.0	credit_card	2.0	146.87	88b8b52d46df026a9d1ad2136a59b30b

review_score	review_comment_title	review_comment_message	review_creation_date	review_answer_timestamp
4	NaN	NaN	2017-05-26 00:00:00	2017-05-30 22:34:40

df_customers_orders_payments_reviews_items_products

114862 rows × 37 columns

Le nombres de clients uniques: 92755
Le nombres de références clients: 95832

Fusion avec la géolocation

df_customers_orders_payments_reviews_items_products
(114862, 37)

df_olist_geolocation
(738332, 5)

Jointure left

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state	order_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date	payment_sequential	payment_type	payment_installments	payment_value	review_id
0	06b8999e2fba1a1fbc88172c00ba8bc7	1	14409	franca	SP	00e7ee1b050b8499577073aeb2a297a1	delivered	2017-05-16 15:05:35	2017-05-16 15:22:12	2017-05-23 10:47:57	2017-05-25 10:35:35	2017-06-05 00:00:00	1.0	credit_card	2.0	146.87	88b8b52d46df026a9d1ad2136a59b30b

review_score	review_comment_title	review_comment_message	review_creation_date	review_answer_timestamp	order_item_id	product_id	seller_id	shipping_limit_date	price	freight_value	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm
4	NaN	NaN	2017-05-26 00:00:00	2017-05-30 22:34:40	1	a9516a079e37a9c9c36b9b78b10169e8	7c67e1448b00f6e969d365cea6b010ab	2017-05-22 15:22:12	124.99	21.88	moveis_escritorio	41.0	1141.0	1.0	8683.0	54.0	64.0	31.0

product_category_name_english
office_furniture

	customer_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state
0	1037	-23.545621	-46.639292	sao paulo	SP

df_final_fusion

114862 rows × 41 columns

Le nombres de clients uniques: 92755
Le nombres de références clients: 95832

Data cleaning et features sélections

Les variables supprimés

customer id	0
customer zip code prefix	0
customer city	0
customer state	0
order status	0
order approved at	15
order delivered carrier date	2
order estimated delivery date	0
payment sequential	0
payment type	0
payment installments	0
review id	0
review comment title	101275
review comment message	0
review creation date	0
review answer timestamp	0
product id	0
seller id	0
shipping limit date	0
price	0
freight value	0
product category name	0
product name lenght	1626
product description lenght	1626
product photos qty	1626
product weight g	20
product length cm	20
product height cm	20
product width cm	20
product category name english	0
geolocation lat	302
geolocation lng	302
geolocation city	302
geolocation_state	302

Features sélections

customer unique id	0
order id	0
order item id	0
order_purchase_timestamp	0
order delivered customer date	8
payment value	3
review_score	0

Supprimer les valeurs manquantes

Feature engineering

weekday_order_purchase	0
month_order_purchase	0
hour_order_purchase	0
recence	0
Duration_delevery_order	0

olist

Dataframe pour la segmentation

Groupby

customer_unique_id

Aggrégation

order id	count
order item id	mean
payment value	sum
review_score	mean
recence	mean
Duration_delevery_order	mean

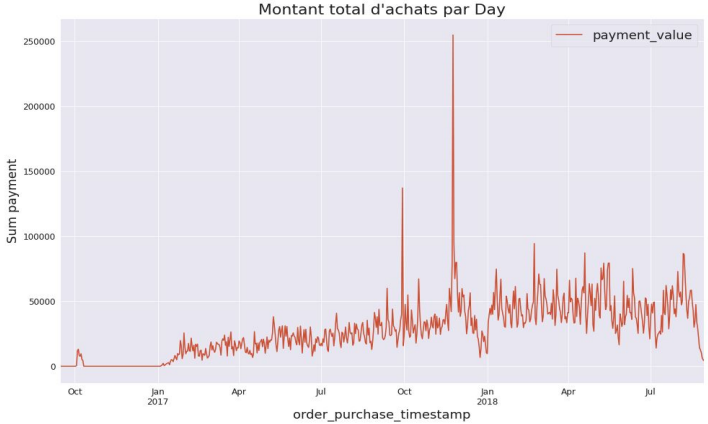
Drop outlier

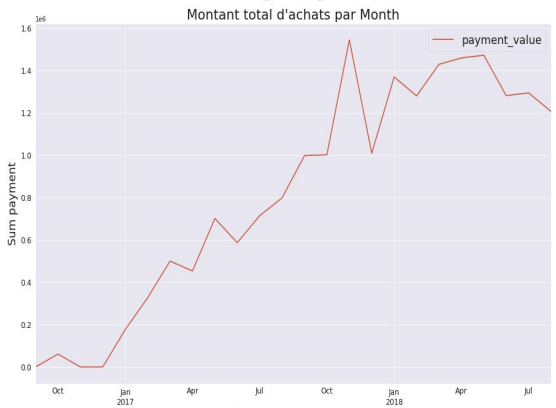
Client 6186 sensibilise le clustering et la stabilité

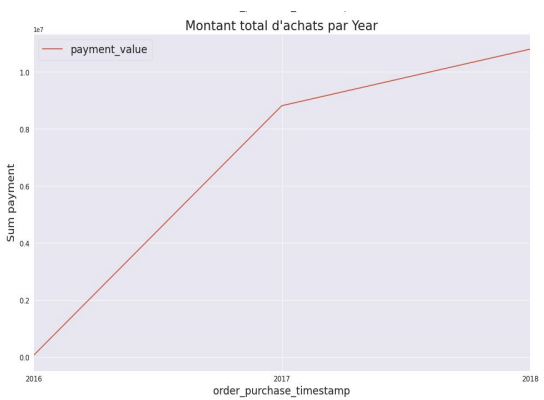
	customer_unique_id	recence	frequence	mean_order_item_id	Monnaie	review_score_mean	Duration_delevery_order_mean
0	1	470.0	1.0	1.0	146.87	4.0	9.0
1	2	229.0	1.0	1.0	335.48	5.0	17.0
2	3	102.0	1.0	1.0	157.73	5.0	26.0
3	4	169.0	1.0	1.0	173.30	5.0	15.0
4	5	31.0	1.0	1.0	252.25	5.0	11.0
...	...	...	...	...	...	...	...
92740	96092	144.0	1.0	1.0	88.78	4.0	6.0
92741	96093	147.0	1.0	1.0	129.06	5.0	7.0
92742	96094	143.0	1.0	1.0	56.04	1.0	31.0
92743	96095	299.0	1.0	1.0	711.07	5.0	13.0
92744	96096	253.0	1.0	1.0	21.77	5.0	7.0

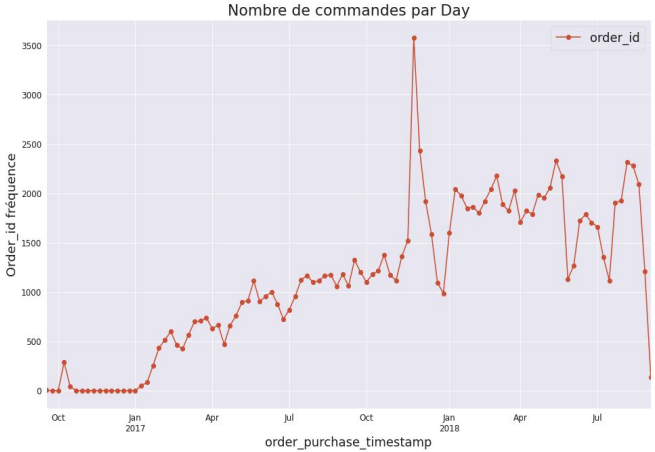
92745 rows × 7 columns

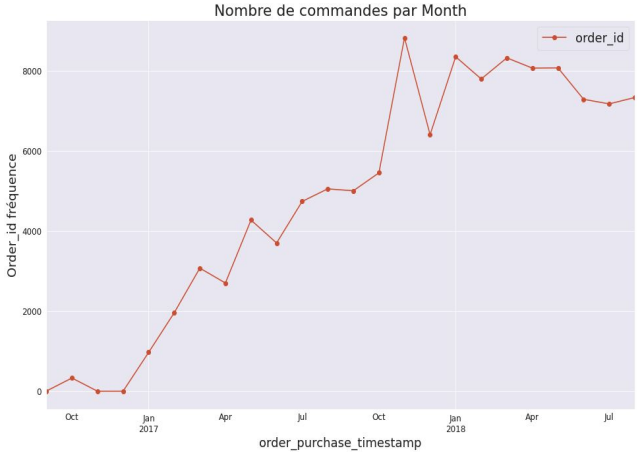
Analyse temporelle

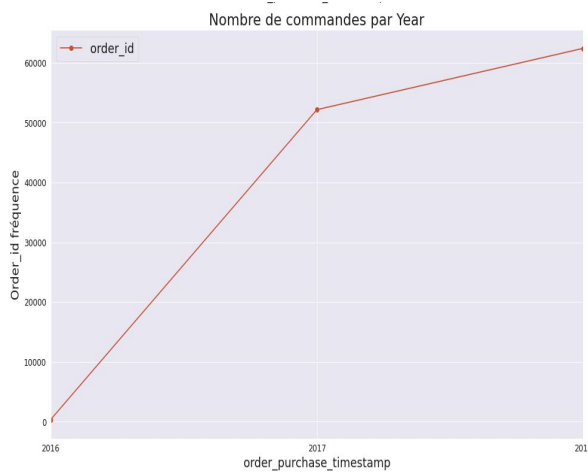






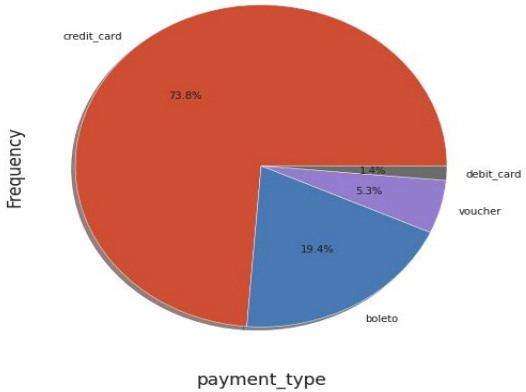




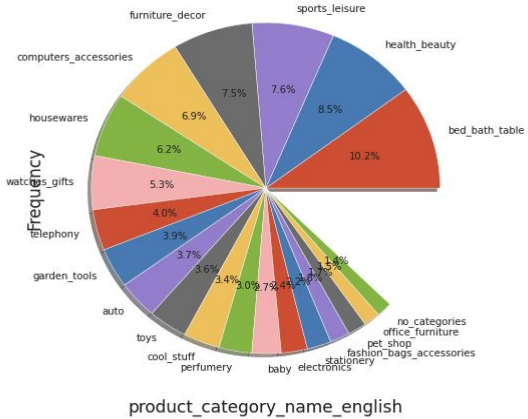


Analyse catégorielle

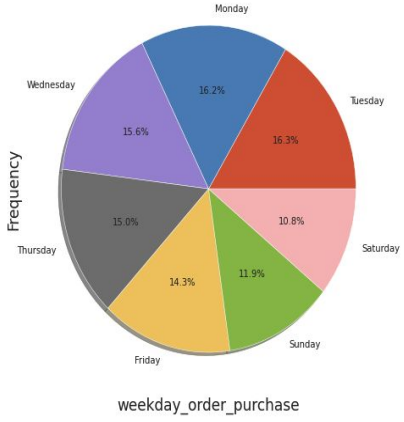
Répartition de la variable payment_type



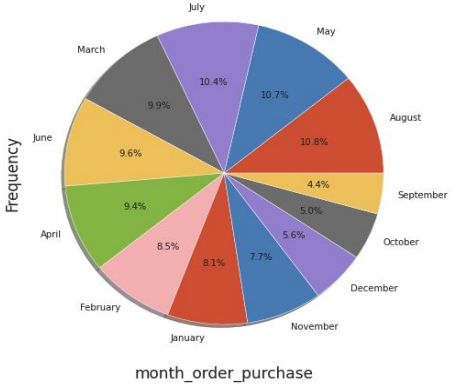
Répartition de la variable product_category_name_english



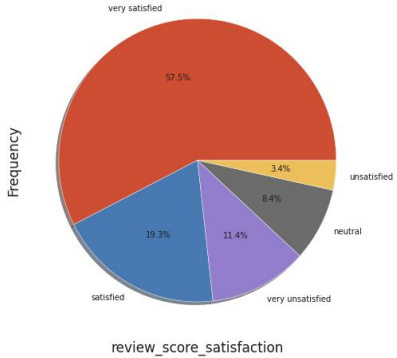
Répartition de la variable weekday_order_purchase



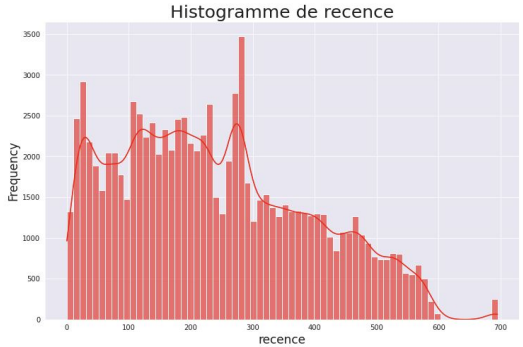
Répartition de la variable month_order_purchase

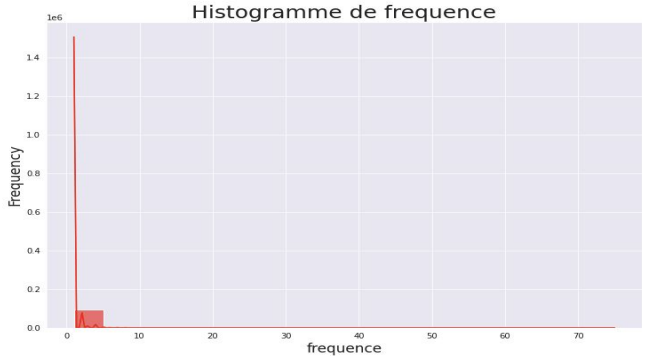


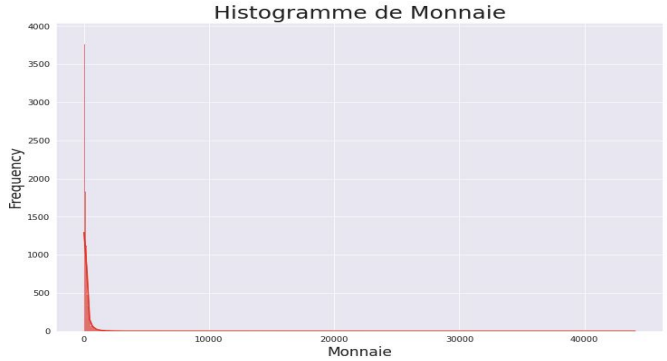
Répartition de la variable review_score_satisfaction

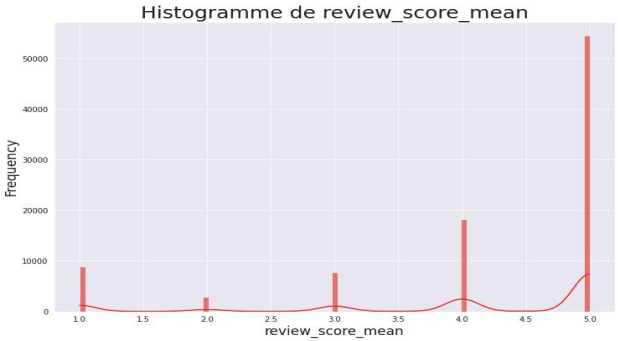


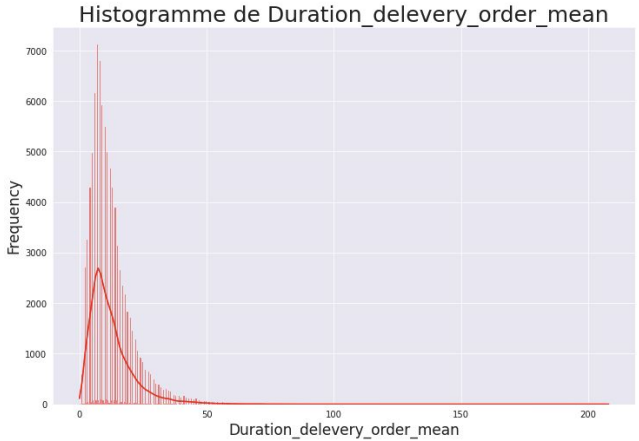
Data exploration











olist

Approche clustering (kmeans)

Kmeans RFM

Evaluation best K

n_clusters = 4

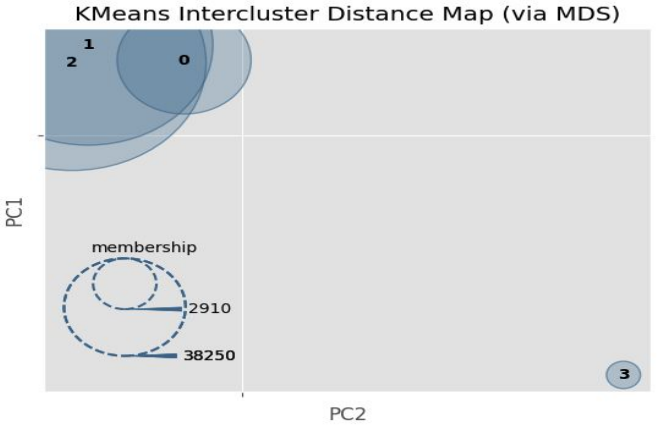
- méthode du coude , davies_bouldin_score(0.76)
- silhouette score(0.46) ,calinski_harabasz(41342)

Stabilité à l'initialisation

- Stabilité moyenne pour 10 itérations

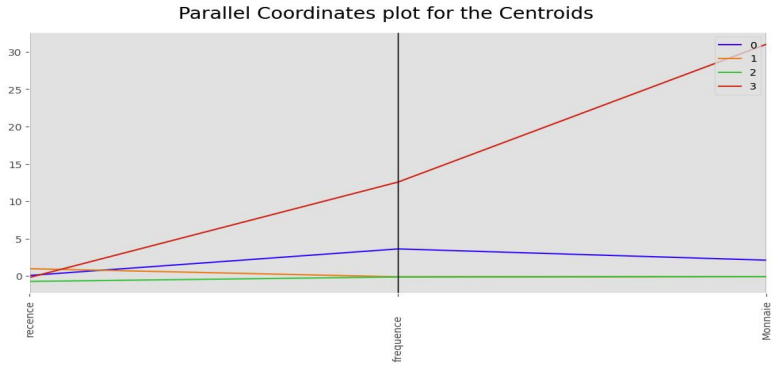
Distance

- Les clusters (1 et 2) trop similaire en terme F,M
-Cluster 3 suffisamment distant mais trop peu de clients



Effectifs des clusters

- pas homogène
- pas pertinent pour le besoin métier
- La fréquence à trop de poids dans la segmentation (97% clients ont passé 1 commande)

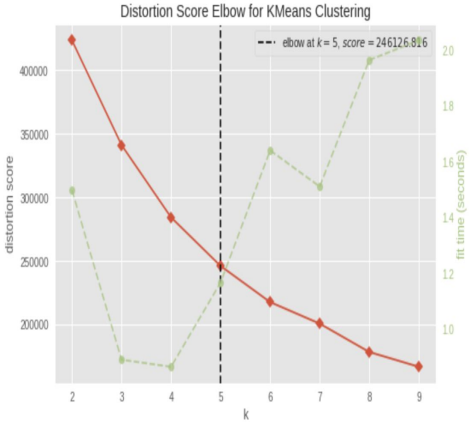


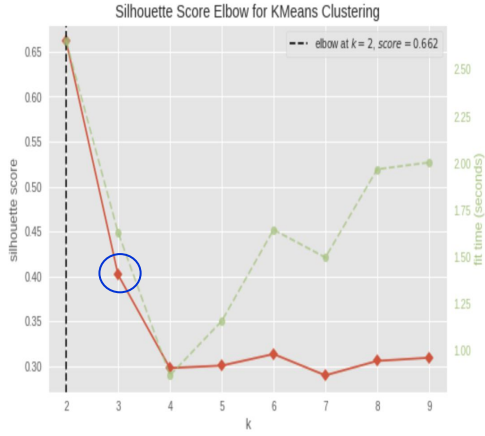
cluster 0 : 2910
cluster 1 : 38250
cluster 2 : 51547
cluster 3 : 38

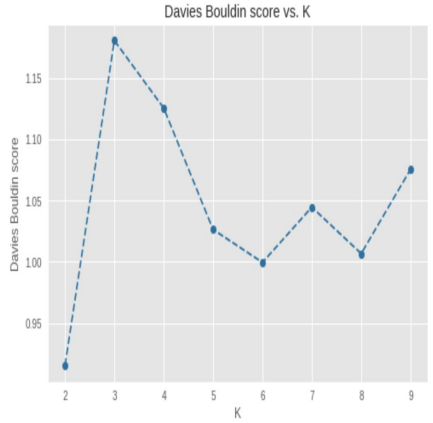
Conclusion

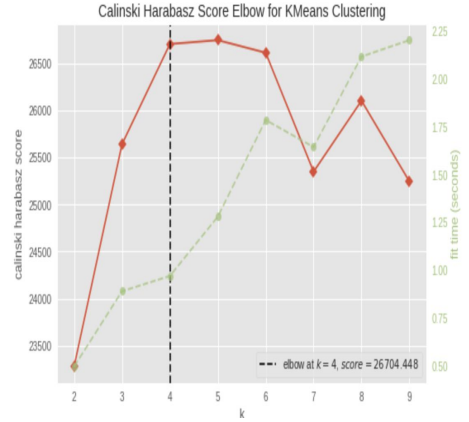
La segmentation RFM pas pertinente d'un point de vue métier.

KMeans évaluation du best K (RFM, mean satisfaction, mean items, mean durée livraison)









Méthode du coude (distortion)

La moyenne de la somme des carrés des distances au centroïde le proche (coude)

Silhouette score

Le coefficient de silhouette ou score de silhouette est une mesure utilisée pour calculer
la qualité d'une technique de regroupement.
Sa valeur est comprise entre -1 et 1 .
1 : Signifie que les grappes sont bien séparées les unes des autres et clairement distinguées.
a= distance intra-cluster moyenne,
c'est-à-dire la distance moyenne entre chaque point d'un cluster.

Davies Bouldin

Le score est défini comme la mesure de similarité moyenne de chaque cluster avec son cluster le plus similaire, où la similarité est le rapport des distances intra-cluster aux distances inter-cluster. Ainsi, les clusters plus éloignés et moins dispersés donneront un meilleur score.Le score minimum est de zéro, les valeurs inférieures indiquant un meilleur regroupement.

Calinski

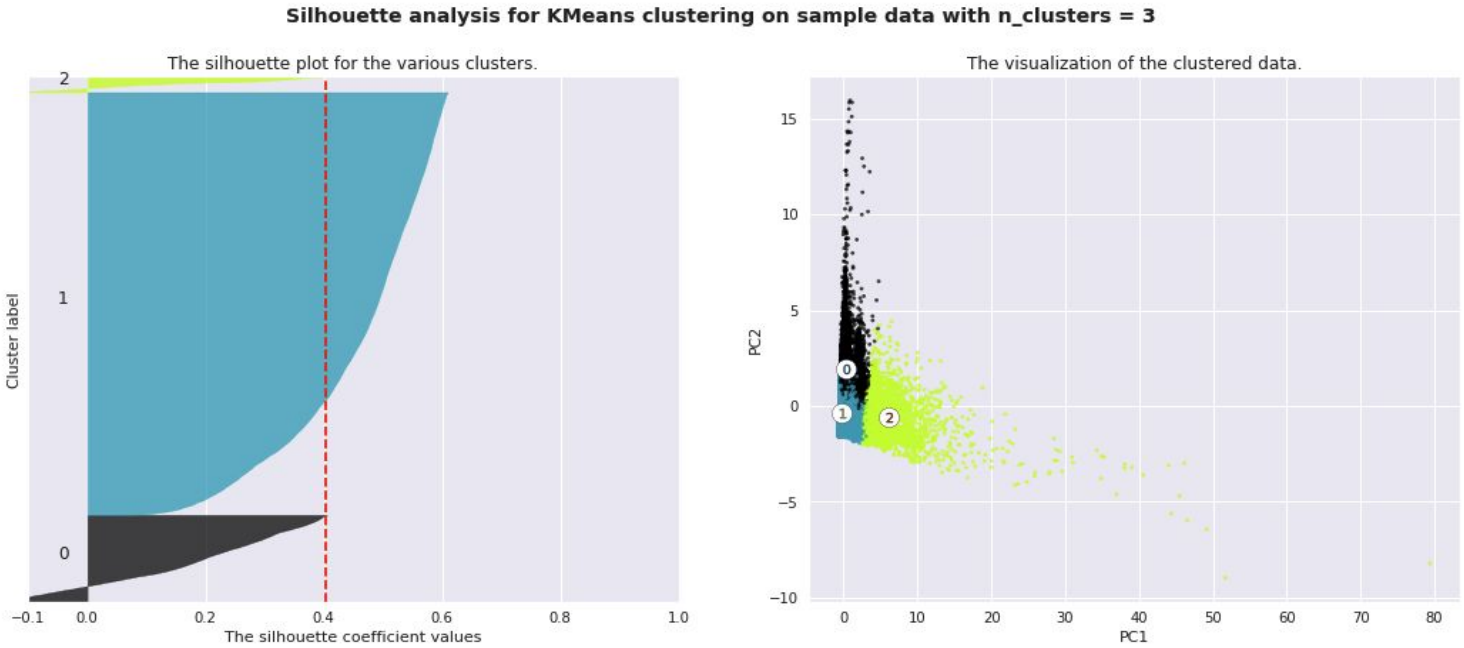
Le score est défini comme le rapport entre la dispersion intra-cluster et la dispersion inter-cluster.

Stabilité à l'initialisation avec Kmeans

```
-------
Iter_1(n_cluster 3):
-------
Fit_time:  2.384185791015625e-07
Le score ARI  0.9999521516389113
Le score HOMO  0.9997709474649992
Le score AMI 0.9997857656100236
-------
Iter_2(n_cluster 3):
-------
Fit_time:  7.152557373046875e-07
Le score ARI  0.9959405702565177
Le score HOMO  0.9901246797240195
Le score AMI 0.9888804399395965
-------
Iter_3(n_cluster 3):
-------
Fit_time:  9.5367431640625e-07
Le score ARI  0.9999521516389113
Le score HOMO  0.9997709474649992
Le score AMI 0.9997857656100236
-------
Iter_4(n_cluster 3):
-------
Fit_time:  4.76837158203125e-07
Le score ARI  0.999904303273828
Le score HOMO  0.9995678771782378
Le score AMI 0.9995975095952087
-------
Iter_5(n_cluster 3):
-------
Fit_time:  2.384185791015625e-07
Le score ARI  0.9964650441254504
Le score HOMO  0.991210947206712
Le score AMI 0.9901260596176928
-------
```

```
-------
Iter_6(n_cluster 3):
-------
Fit_time:  7.152557373046875e-07
Le score ARI  0.9999521516389113
Le score HOMO  0.9997709474649992
Le score AMI 0.9997857656100236
-------
Iter_7(n_cluster 3):
-------
Fit_time:  1.1920928955078125e-06
Le score ARI  0.9998086065314611
Le score HOMO  0.9991877183510793
Le score AMI 0.9992469672824071
-------
Iter_8(n_cluster 3):
-------
Fit_time:  2.384185791015625e-07
Le score ARI  1.0
Le score HOMO  1.0
Le score AMI 1.0
-------
Iter_9(n_cluster 3):
-------
Fit_time:  9.5367431640625e-07
Le score ARI  0.999904303273828
Le score HOMO  0.9995678771782378
Le score AMI 0.9995975095952087
-------
Iter_10(n_cluster 3):
-------
Fit_time:  2.384185791015625e-07
Le score ARI  0.999904303273828
Le score HOMO  0.9995678771782378
Le score AMI 0.9995975095952087
```

Forme des clusters

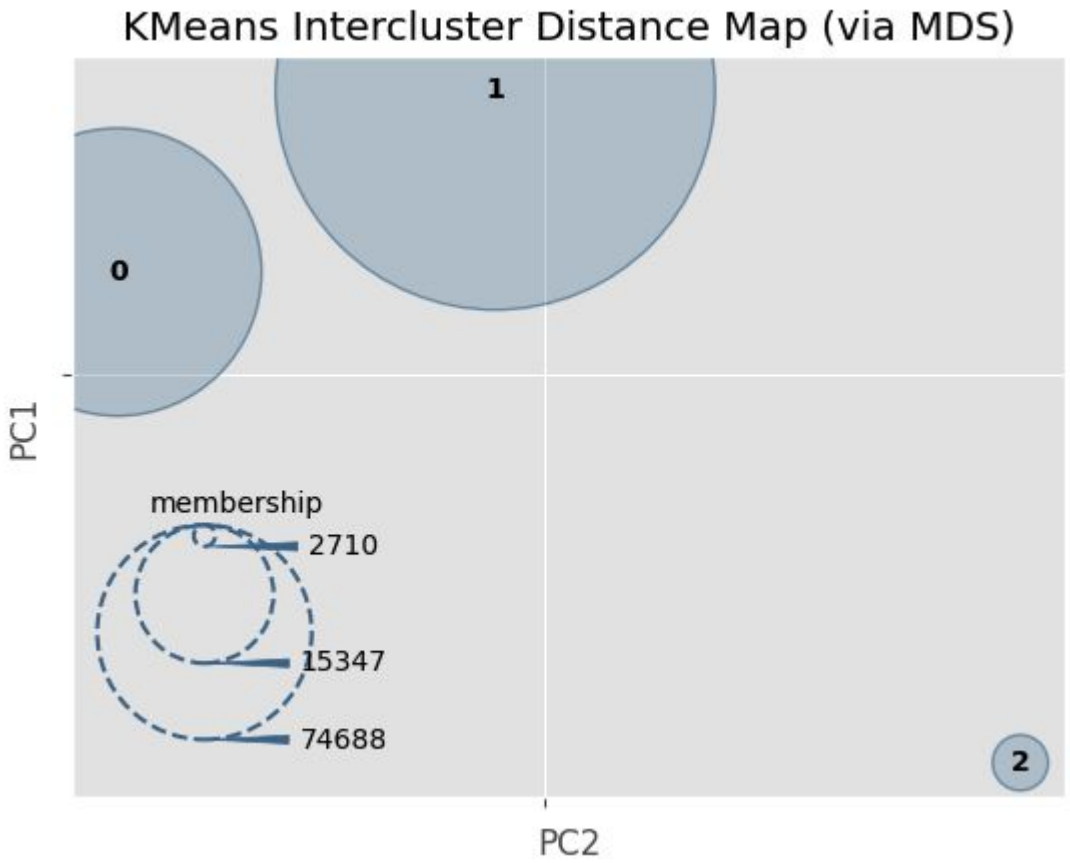


Cluster 0: 15347 clients
Cluster 1: 74688 clients
Cluster 2: 2710 clients

- Silhouette Score: 0.40
- calinski_harabasz Score: 25600
- davies_bouldin Score: 1.18

Distance entre les clusters

La distance intercluster permet de voir une dissimilarité entre les clusters



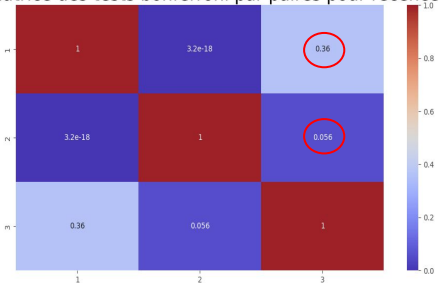
Déterminer si un cluster est différents des autres

Kruskal-wallis Test 5%

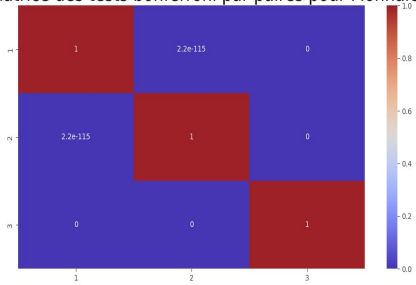
- stat=60105.071, p=0.000
  recence moyenne
- stat=19657.082, p=0.000
  fréquence
- stat=6645.377, p=0.000
  Monnaie
- stat=41808.725, p=0.000
  review_score_mean
- stat=12074.576, p=0.000
  Duration_delevery_order_mean
- stat=25552.940, p=0.000
  mean_order_item_id

Déterminer si les clusters par paires sont différents

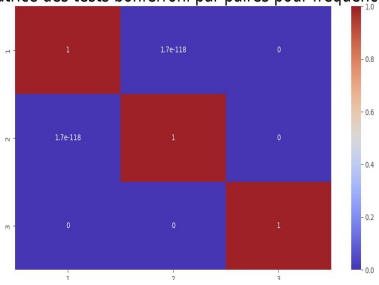
La matrice des tests bonferroni par paires pour recence



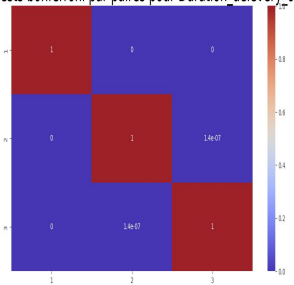
La matrice des tests bonferroni par paires pour Monnaie



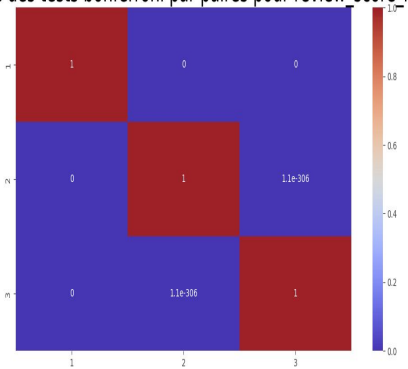
La matrice des tests bonferroni par paires pour frequence



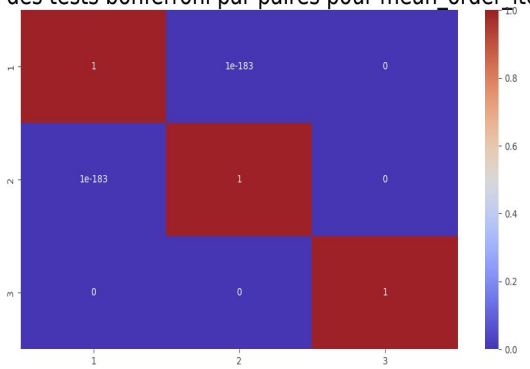
La matrice des tests bonferroni par paires pour Duration_delevery_order_mean



La matrice des tests bonferroni par paires pour review_score_mean



La matrice des tests bonferroni par paires pour mean_order_item_id



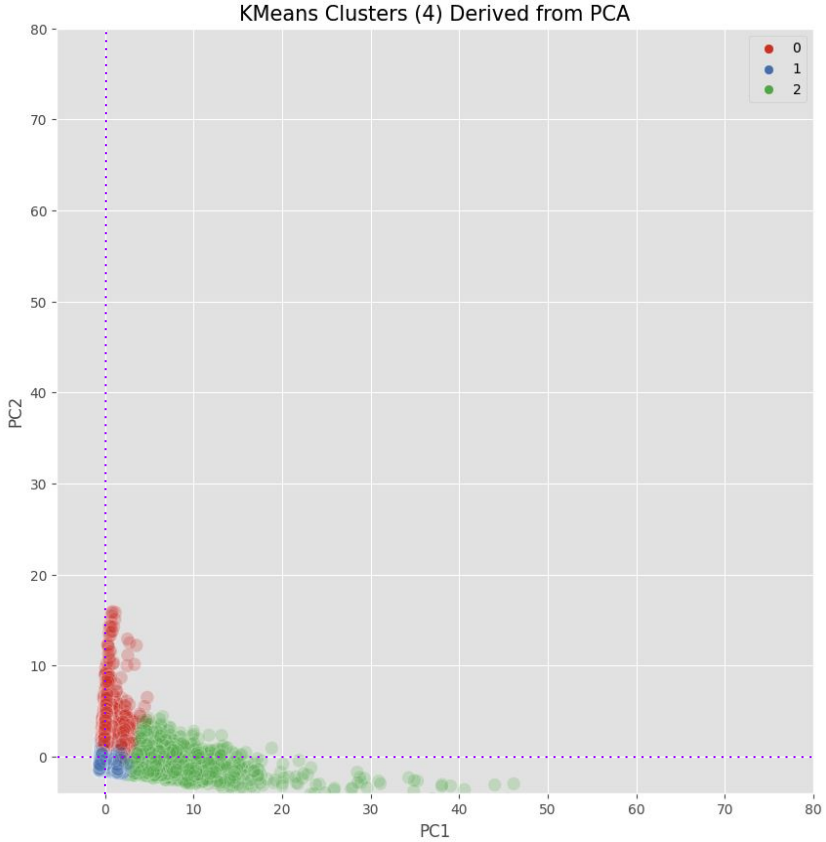
Test Dunn 5%

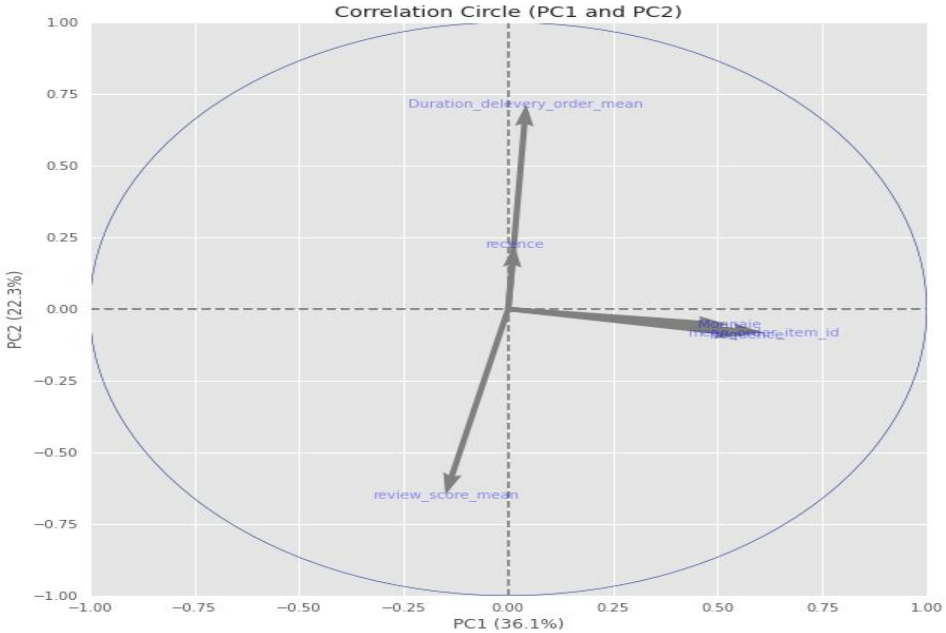
Hypothèses des test paramétriques non respectées (Variance et normalité)

Les clusters non différents sont (1 et 3) (2 et (2 et 3) pour la récence

Les autres paires sont différentes pour les autres features

ACP visualisation des clusters

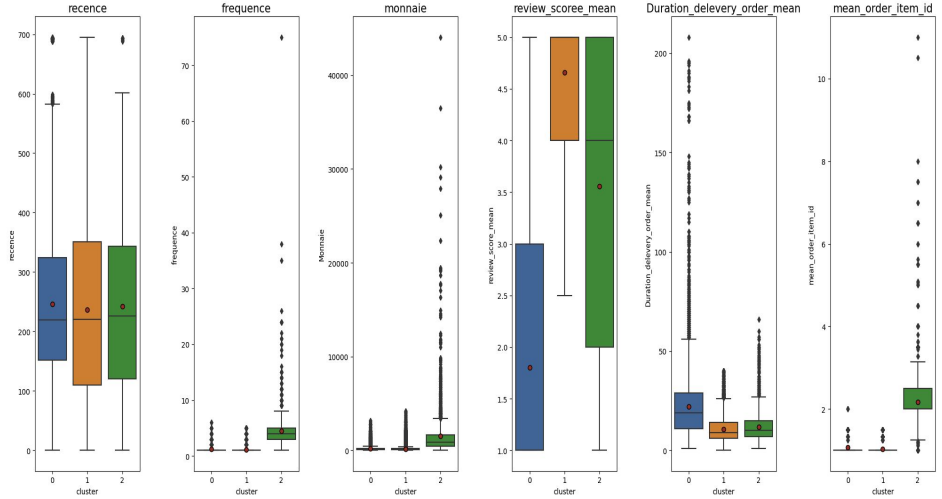


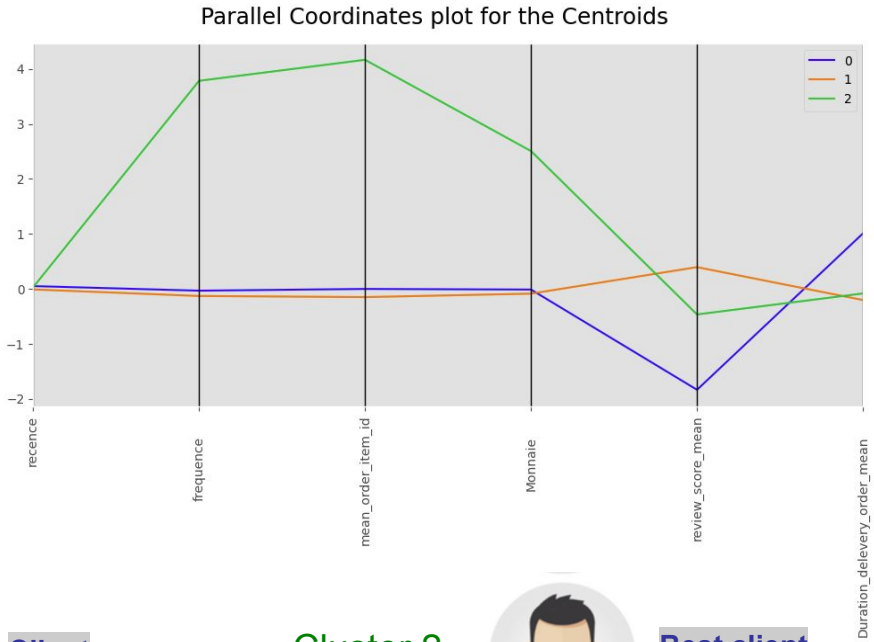


PC1 :Monnaie, mean_order_item_d, fréquence

PC2 : review_score_mean, Duration_delevery_order_mean

Analyse des clusters





cluster 0

- fréquence : 1
- Moyenne articles : 1
- Dépense moyenne : 203
- Satisfaction moyenne : 1.80
- Durée livraison moyenne : 22 jours

Client très insatisfait peu dépensier 16.55 %

cluster 1

- fréquence : 1
- Moyenne articles : 1
- Dépense moyenne : 165.26
- Satisfaction moyenne : 4.65
- Durée livraison moyenne : 10 jours

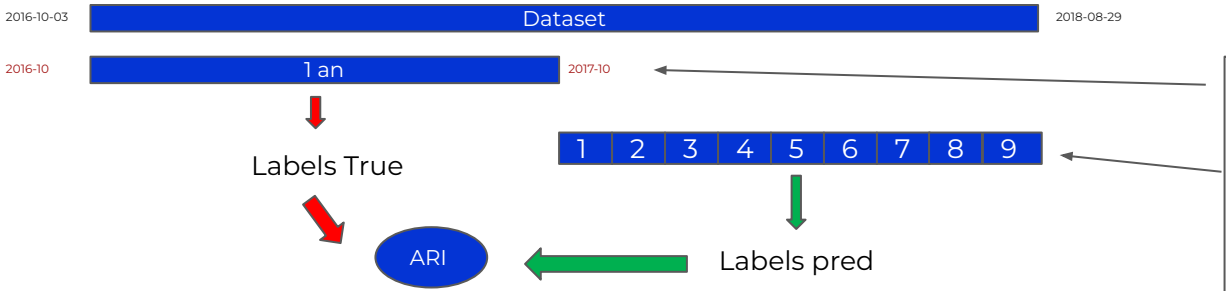
Client très satisfait peu dépensier 80.53 %

Cluster 2

- fréquence : 4
- Moyenne articles : 2
- Dépense moyenne : 1509.46
- Satisfaction moyenne : 3.55
- Durée livraison moyenne : 11 jours

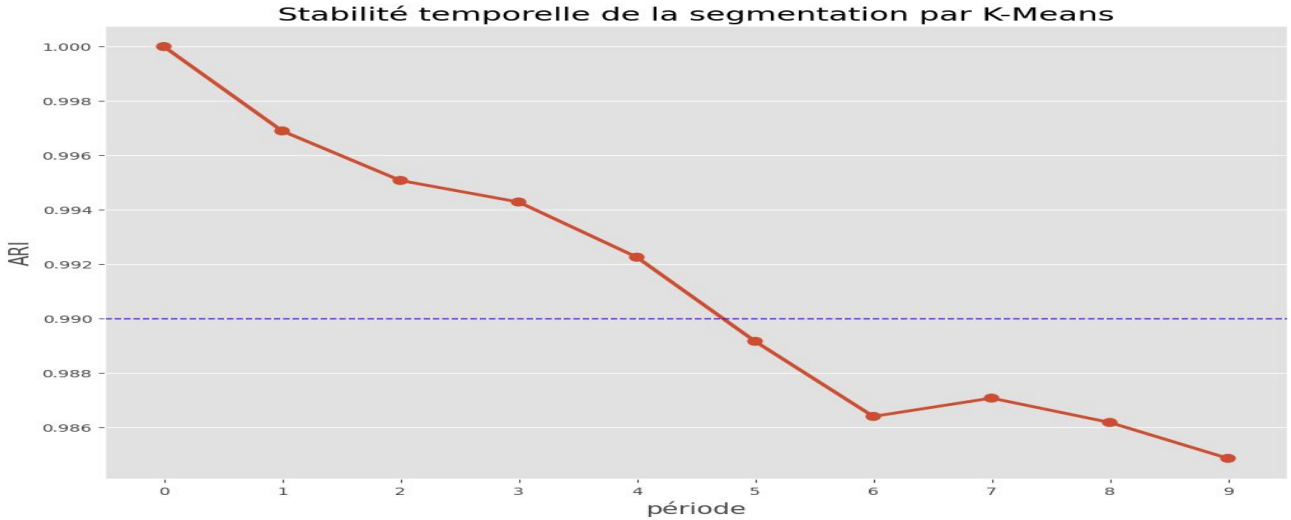
Best client satisfaction moyenne très dépensier 2.92 %

Stabilité temporelle

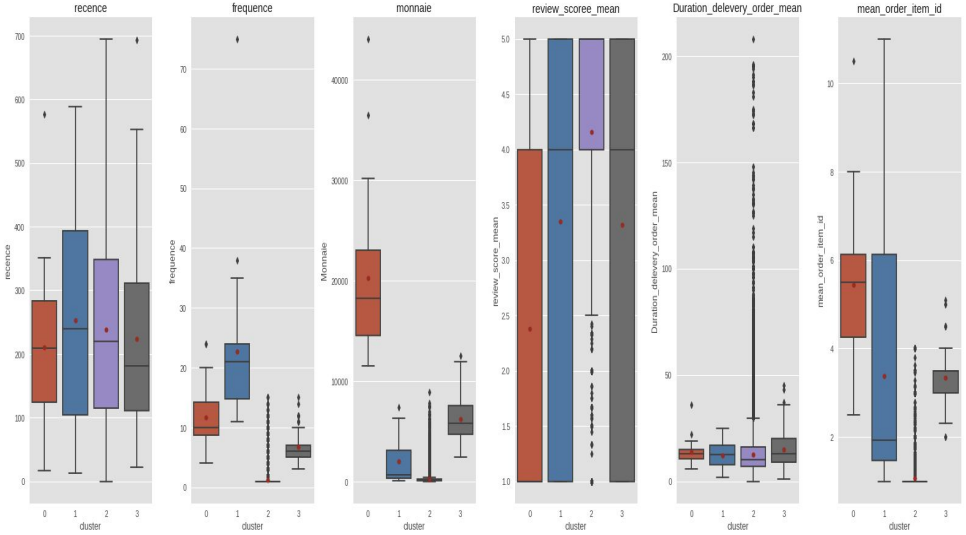


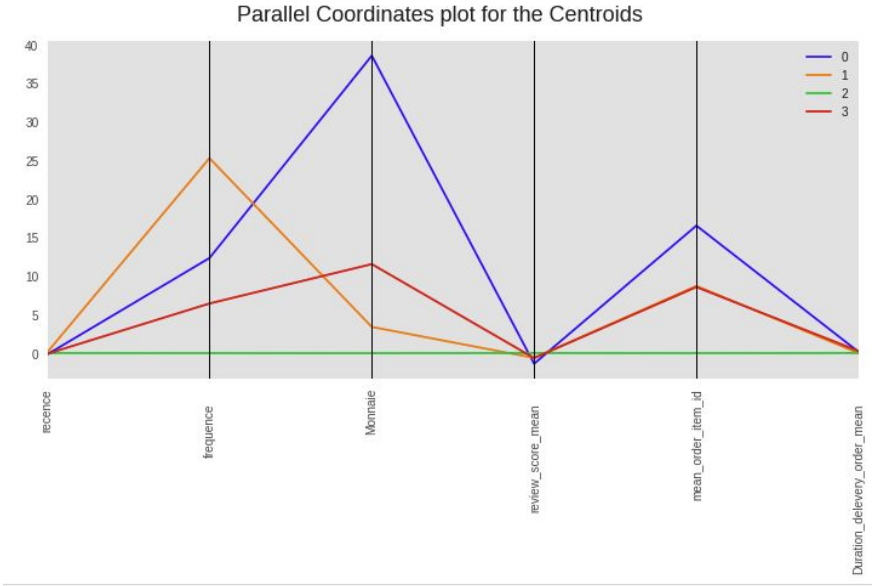
- Clients communs pour chaque périodes par rapport à la période de 1 an
- Agg des variables
- PCA
-Kmeans(n_clusters=3,init='k-means++', random_state=39)

	periode	ARI
0	0	1.000000
1	1	0.996899
2	2	0.995078
3	3	0.994286
4	4	0.992258
5	5	0.989166
6	6	0.986405
7	7	0.987079
8	8	0.986183
9	9	0.984864



Algorithme BIRCH





Effectifs

cluster 0 : 24
cluster 1 : 24
cluster 2 : 92589
cluster 3 : 108

Evaluation

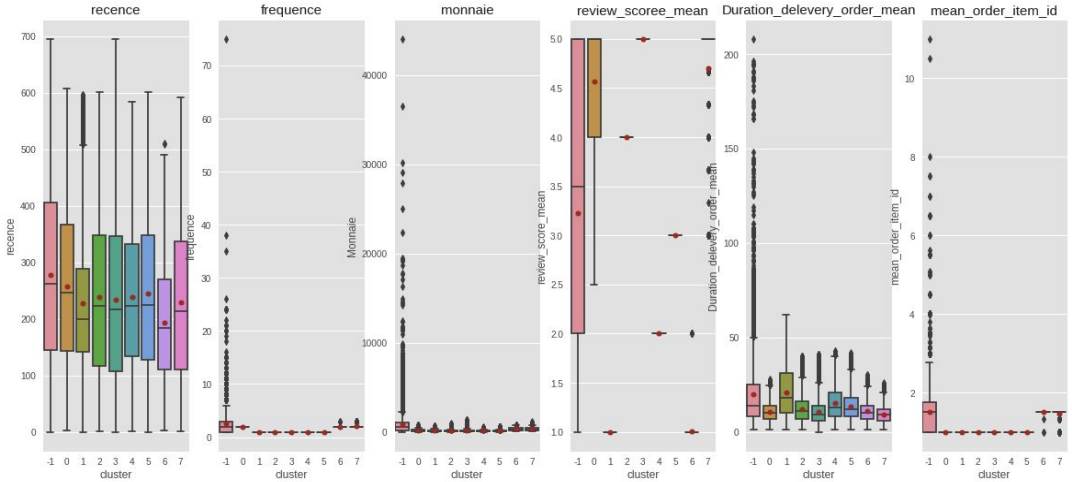
- Silhouette Score: 0.84
- calinski_harabasz Score: 6164.79
- davies_bouldin Score: 0.68

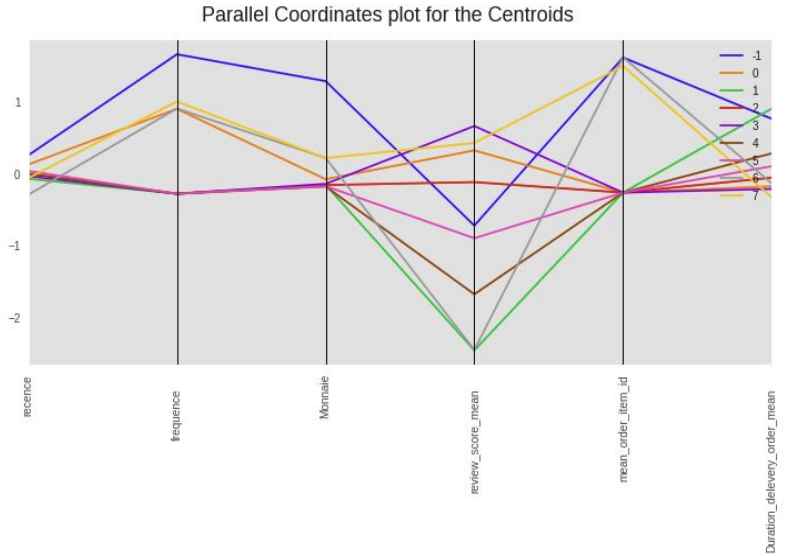
Conclusion

Le clustering ne permet pas de répondre à la problématique métier (Les effectifs des clusters sont trop hétérogène pour effectuer des stratégies marketing)

olist

Algorithme HDBSCAN





Effectifs

cluster -1 (bruit): 8350
cluster 0 : 3003
cluster 1 : 6092
cluster 2 : 46826
cluster 3 : 15506
cluster 4 : 1897
cluster 5 : 6055
cluster 6 : 867
cluster 7 : 4149

Evaluation

- Silhouette Score: 0.09
- calinski_harabasz Score: 5985.50
- davies_bouldin Score: 2.62

Conclusion

- Le clustering ne permet pas de répondre à la problématique métier (Les effectifs des clusters sont trop hétérogène pour effectuer des stratégies marketing)
- Les métriques indiquent que la segmentation n'est pas pertinente

Conclusion

La segmentation finale porte sur 6 variables :

- Recence
- Fréquence
- Monnaie
- Review_score_mean
- Mean_order_item_id
- Duration_delevery_order_mean

Le modèle final choisis est le Kmeans:

- 3 clusters
- Stable à l'initialisation avec 10 itérations
- Le clustering permet d'identifier des profils clients afin d'adapter une stratégie marketing pour chaque clusters
- Le clustering permet de segmenter les bons et moins bons clients en termes de commandes et de satisfaction.

contrat de maintenance

- Stabilité fortement dépendante des variables choisies
- La fréquence à laquelle la segmentation doit être mise à jour est 5 mois pour qu'elle reste pertinente