olist

Segmentez des clients d'un site e-commerce

...

Projet 5 Parcours Data Scientist
Python Jupyter Notebook

_____

Nalron Décembre 2020
OpenClassrooms - Centrale Supélec

Contexte général de la problématique du client

Donneur d'ordre

Olist solution de vente sur MarketPlace.

Besoin

Segmentation de la base clients pour les campagnes de communication.

Objectif

Comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.

Mission et méthodologie

Aider les équipes d'Olist à comprendre les différents types d'utilisateurs.

Proposer une segmentation exploitable et facile d'utilisation pour le marketing.

Evaluer la fréquence de mise à jour de la segmentation (*contrat de maintenance*).

Fournir un code qui respecte la convention PEP8, pour être utilisable par Olist.

Méthode envisagée

Méthodes non supervisées pour regrouper ensemble des profils similaires.
Création de features car peu de données fournies par Olist.

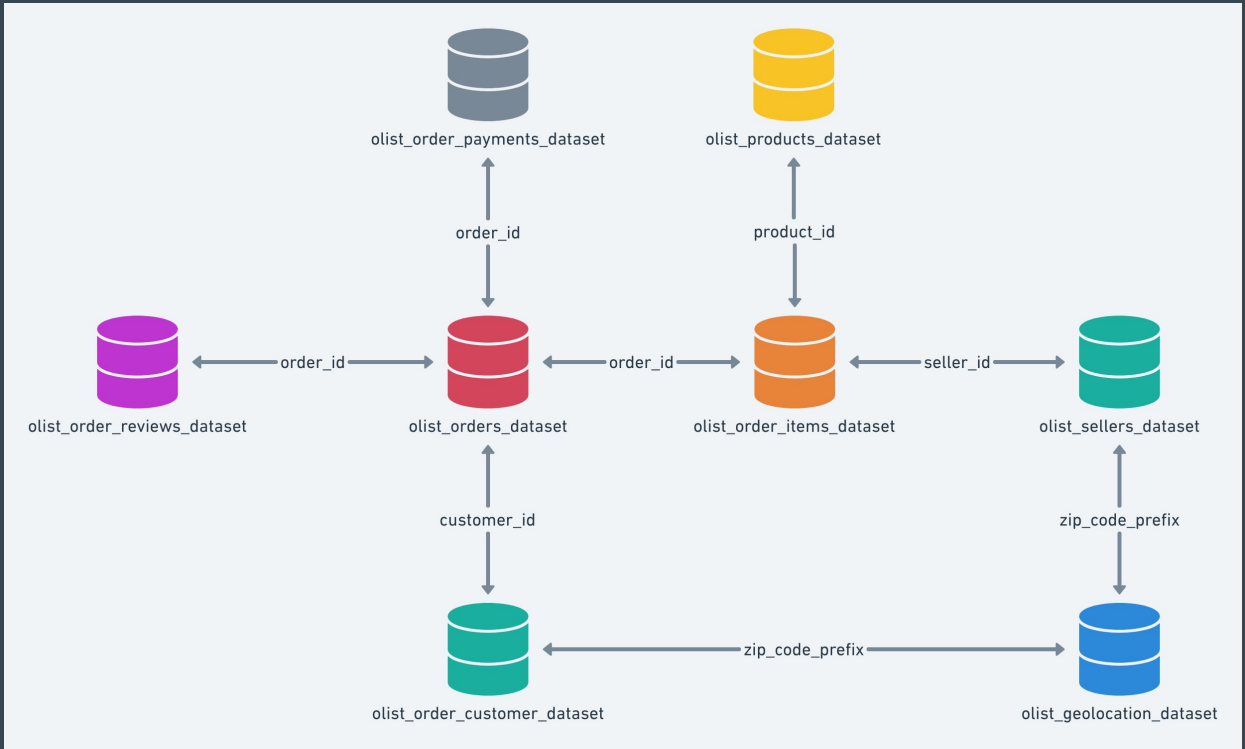
Présentation des données

Olist met à disposition sa base de données anonymisée sur l'historique des commandes (produits achetés, date d'achat, localisation, commentaires clients, ...) de septembre 2016 à octobre 2018.

lien des données :
https://www.kaggle.com/olistbr/brazilian-ecommerce

olist

Schéma synthétique de la base de données



Source : Kaggle

Vue synthétique des données : 9 fichiers

	Rows	Columns	%NaN	%Duplicate	object_dtype	float_dtype	int_dtype	bool_dtype	MB_Memory
customers	99441	5	0.00	0.00	4	0	1	0	3.793
geolocation	1000163	5	0.00	5.24	2	2	1	0	38.153
sellers	3095	4	0.00	0.00	3	0	1	0	0.095
products	32951	9	0.83	0.00	2	7	0	0	2.263
product_category	71	2	0.00	0.00	2	0	0	0	0.001
orders	99441	8	0.62	0.00	8	0	0	0	6.070
orders_items	112650	7	0.00	0.00	4	2	1	0	6.016
order_payments	103886	5	0.00	0.00	2	1	2	0	3.963
order_reviews	100000	7	20.93	0.00	6	0	1	0	5.341

Données réparties selon leurs fichiers issus de la database Olist.
Compréhension rapide et facilité des différents types d'informations.

Comparatif des principales features

Aucune donnée manquante sur les principales features explicatives.

Aucune duplication d'identifiant commande.

Analyse des outliers prix produits avec des valeurs élevées mais non absurdes dans le contexte business.

		features	dtype	nan	count	mean	std	min	max
customers	0	customer_zip_code_prefix	int64	0.0	99441.0	35137.0	29797.9	1003.0	99990.0
	0	geolocation_zip_code_prefix	int64	0.0	1000163.0	36574.0	30549.3	1001.0	99990.0
geolocation	1	geolocation_lat	float64	0.0	1000163.0	-21.0	5.7	-36.6	45.1
	2	geolocation_lng	float64	0.0	1000163.0	-46.0	4.3	-101.5	121.1
sellers	0	seller_zip_code_prefix	int64	0.0	3095.0	32291.0	32713.5	1001.0	99730.0
	0	product_name_lenght	float64	610.0	32341.0	48.0	10.2	5.0	76.0
	1	product_description_lenght	float64	610.0	32341.0	771.0	635.1	4.0	3992.0
	2	product_photos_qty	float64	610.0	32341.0	2.0	1.7	1.0	20.0
products	3	product_weight_g	float64	2.0	32949.0	2276.0	4282.0	0.0	40425.0
	4	product_length_cm	float64	2.0	32949.0	30.0	16.9	7.0	105.0
	5	product_height_cm	float64	2.0	32949.0	16.0	13.6	2.0	105.0
	6	product_width_cm	float64	2.0	32949.0	23.0	12.1	6.0	118.0
	0	order_item_id	int64	0.0	112650.0	1.0	0.7	1.0	21.0
orders_items	1	price	float64	0.0	112650.0	120.0	183.6	0.8	6735.0
	2	freight_value	float64	0.0	112650.0	19.0	15.8	0.0	409.7
	0	payment_sequential	int64	0.0	103886.0	1.0	0.7	1.0	29.0
order_payments	1	payment_installments	int64	0.0	103886.0	2.0	2.7	0.0	24.0
	2	payment_value	float64	0.0	103886.0	154.0	217.5	0.0	13664.1
order_reviews	0	review_score	int64	0.0	100000.0	4.0	1.4	1.0	5.0

Extrait du comparatif des features catégorielles

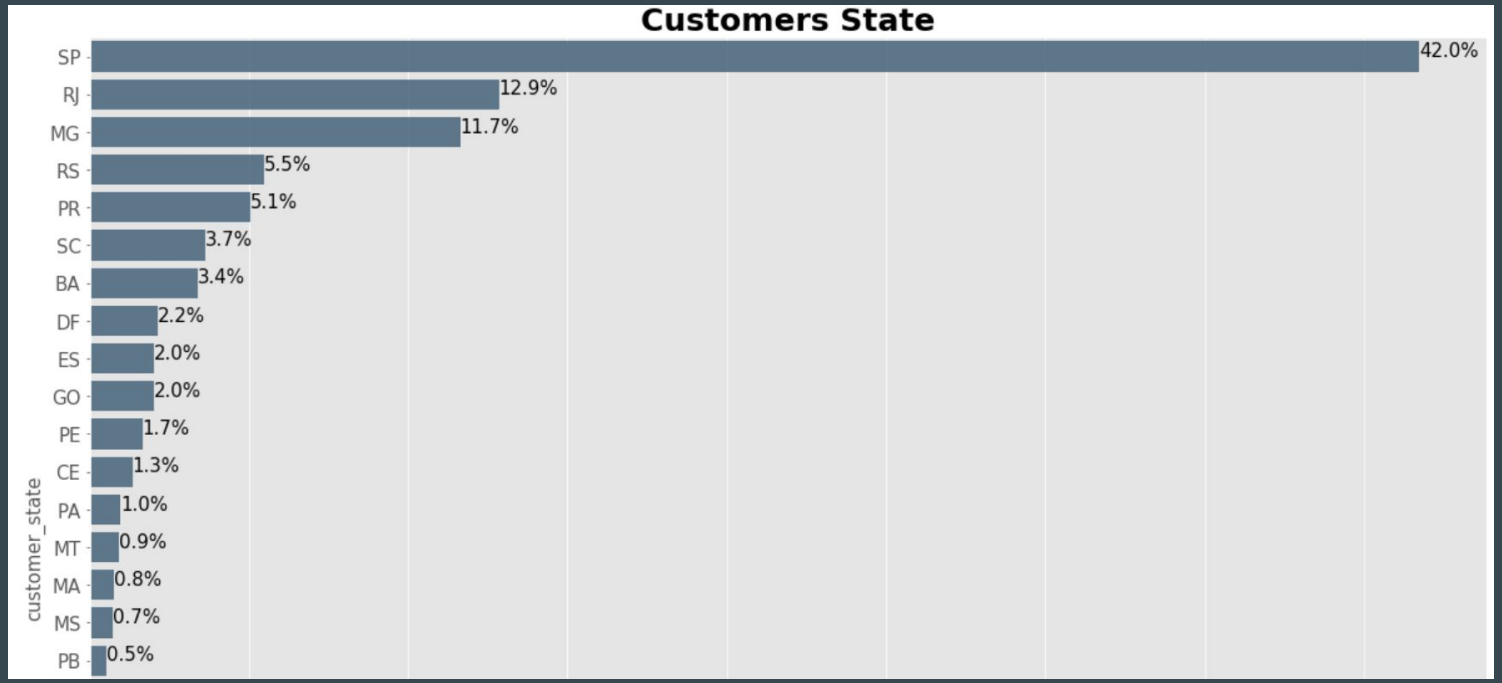
Clés principales :

- customer_unique_id
- order_id
- product_id

Agrégation des données nécessaires selon les clés de liaison.

		features	dtype	nan	count
customers	0	customer_id	object	0	99441
	1	customer_unique_id	object	0	99441
	2	customer_city	object	0	99441
	3	customer_state	object	0	99441
geolocation	0	geolocation_city	object	0	1000163
	1	geolocation_state	object	0	1000163
sellers	0	seller_id	object	0	3095
	1	seller_city	object	0	3095
	2	seller_state	object	0	3095
products	0	product_id	object	0	32951
	1	product_category_name	object	610	32341
product_category	0	product_category_name	object	0	71
	1	product_category_name_english	object	0	71
orders	0	order_id	object	0	99441
	1	customer_id	object	0	99441
	2	order_status	object	0	99441
	3	order_purchase_timestamp	object	0	99441
	4	order_approved_at	object	160	99281
	5	order_delivered_carrier_date	object	1783	97658
	6	order_delivered_customer_date	object	2965	96476
	7	order_estimated_delivery_date	object	0	99441

Localisation des clients



e-commerce exploité au Brésil.

Des zones géographiques plus significatives : São Paulo, Rio de Janeiro, Minas Gerais

Typologie des clients

97% des clients ont acheté qu'une seule fois en deux ans d'activité.
Clients très volatiles, aucune fidélisation ni adhésion à la plateforme...

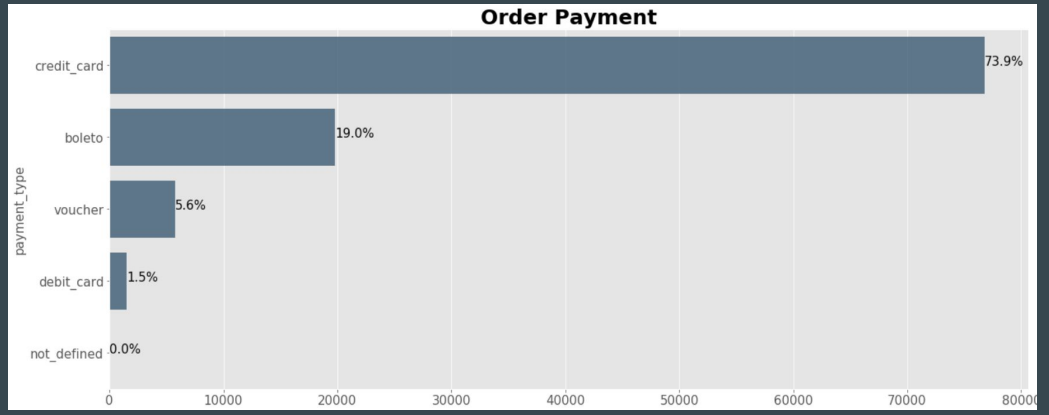
Achat One Shot

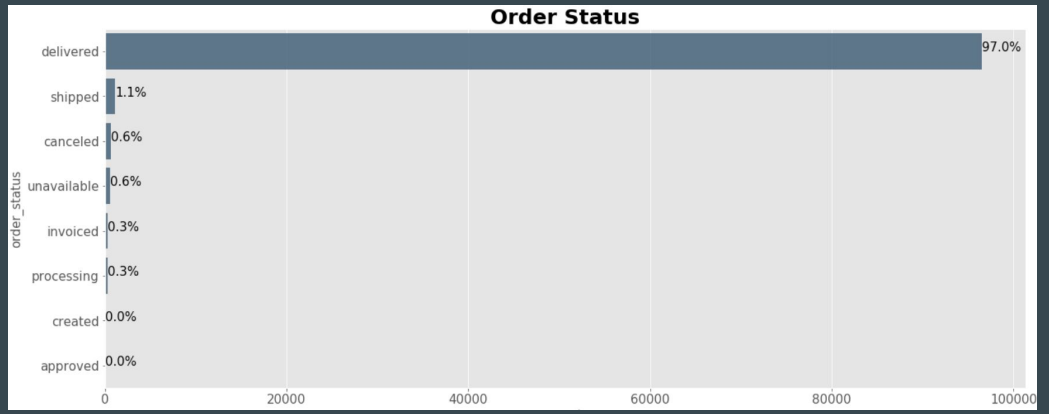
Exploration des features liées aux commandes

"Boleto" est un moyen de paiement en espèces le plus populaire du Brésil.

"Voucher" est une méthode de paiement prépayée simple.

Restriction sur les commandes valides.





Evolution du nombre de commandes dans le temps

2 années d'activité en évolution avec :
Un rebond entre 12/2016 et 03/2017.
Une stagnation depuis fin 2017.

count	99441
unique	98875
top	2018-03-31 15:08:21
freq	3
first	2016-09-04 21:15:19
last	2018-10-17 17:30:18
Name:	order_purchase_timestamp

Feature engineering

Olist met à disposition très peu de données pour cause de confidentialité.
Le clustering, quelque soit la méthode, demande des variables explicatives, une étape d'ajout de features est donc indispensable.

olist

Intégration de features KPI

KPI Conversion

- Panier moyen
- Montant total
- Montant Quartile < 25%
- Montant Quartile 25 à 50%
- Montant Quartile 50 à 75%
- Montant Quartile > 75%
- Montant Max. commande
- Montant Min. commande
- Montant d'achat catégorie
- Nombre de produits achetés
- Mode de paiement

KPI Rétention/Fidélité

- Nb de jour dernier achat
- Fréquence d'achat
- Note moyenne commentaires
- Nombre de commentaires

KPI Géolocalisation

- État d'origine du client

Jeu de données obtenu

Echantillon de travail

Indexé sur l'identifiant unique client

- 93 694 lignes
- 35 colonnes
- 0 doublon
- 0 NaN

Preprocessing à effectuer pour les différences d' échelle, l'encodage puis réduction de dimension PCA

```
Data columns (total 35 columns):
average_basket                    93694 non-null float64
total_spent                       93694 non-null float64
max_order_amount                  93694 non-null float64
min_order_amount                  93694 non-null float64
number_of_products_purchased      93694 non-null float64
number_orders                     93694 non-null float64
bucket_quartile_lower25           93694 non-null float64
bucket_quartile_25_50             93694 non-null float64
bucket_quartile_50_75             93694 non-null float64
bucket_quartile_upper75           93694 non-null float64
payment_boleto                    93694 non-null float64
payment_credit_card               93694 non-null float64
payment_debit_card                93694 non-null float64
payment_voucher                   93694 non-null float64
appliances                        93694 non-null float64
auto                              93694 non-null float64
construction                      93694 non-null float64
culture                           93694 non-null float64
electronics                       93694 non-null float64
fashion                           93694 non-null float64
food                              93694 non-null float64
garden                            93694 non-null float64
health_beauty                     93694 non-null float64
hobbies                           93694 non-null float64
home                              93694 non-null float64
office                            93694 non-null float64
others                            93694 non-null float64
pets                              93694 non-null float64
sports_leisure                    93694 non-null float64
toys                              93694 non-null float64
last_time_order                   93694 non-null float64
frequency_purchase                93694 non-null float64
average_review                    93694 non-null float64
number_review_comment             93694 non-null float64
customer_state                    93694 non-null object
dtypes: float64(34), object(1)
```

Standardisation - Encodage

Différence d'échelle entre les features :

Traitement par méthode *StandardScaler()*

Features catégorielles liées à la géolocalisation client :

Traitement par méthode *OneHotEncoding()*

Regroupement des états les moins représentatifs sous un intitulé unique 'other'.

Élaboration d'un échantillon réduit :

Optimisation *des temps de calcul pour la phase de test des modèles de clustering.*

```
X_sample = X.copy()
X_sample = X_sample.sample(n=10000)
X_sample.name = "X_sample"
X_sample.shape
```

```
(10000, 38)
```

Modélisation

3 types de modélisation : KMeans, Hiérarchique, densité

Algorithmes non supervisés : KMeans, AgglomerativeClustering, DBSCAN

Recherche du modèle d'apprentissage non supervisé adapté au problème métier.
Amélioration des hyperparamètres et évaluation des performances...
Test de stabilité sur la fréquence de mise à jour.

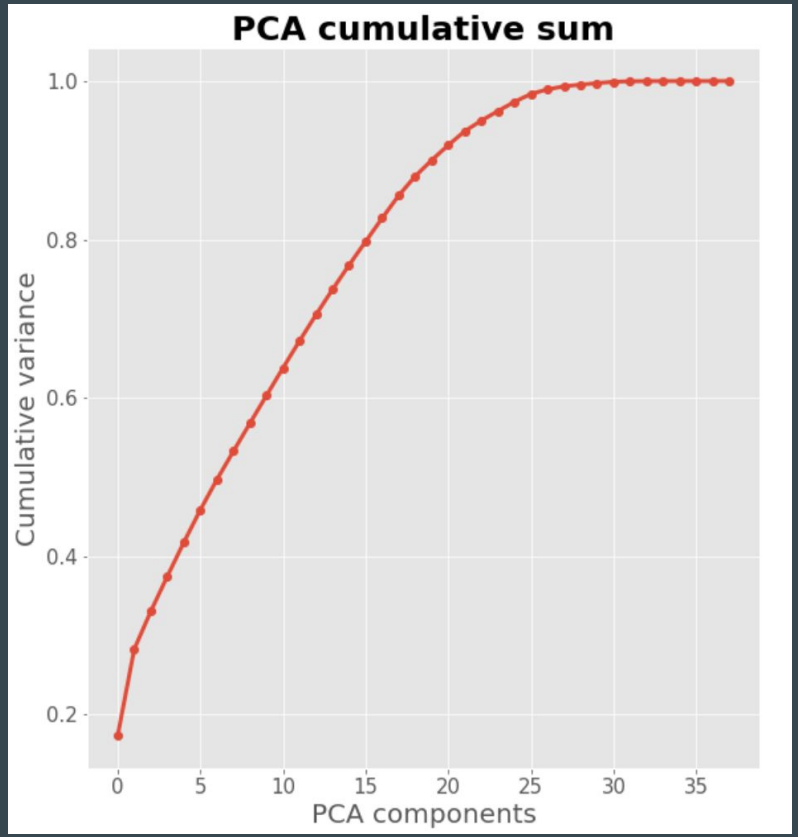
olist

Analyse en Composantes Principales

Prise de connaissance de la variance expliquée.

1er plan factoriel ≃ 30%

Réduction de dimension pour faciliter la visualisation des clusters.

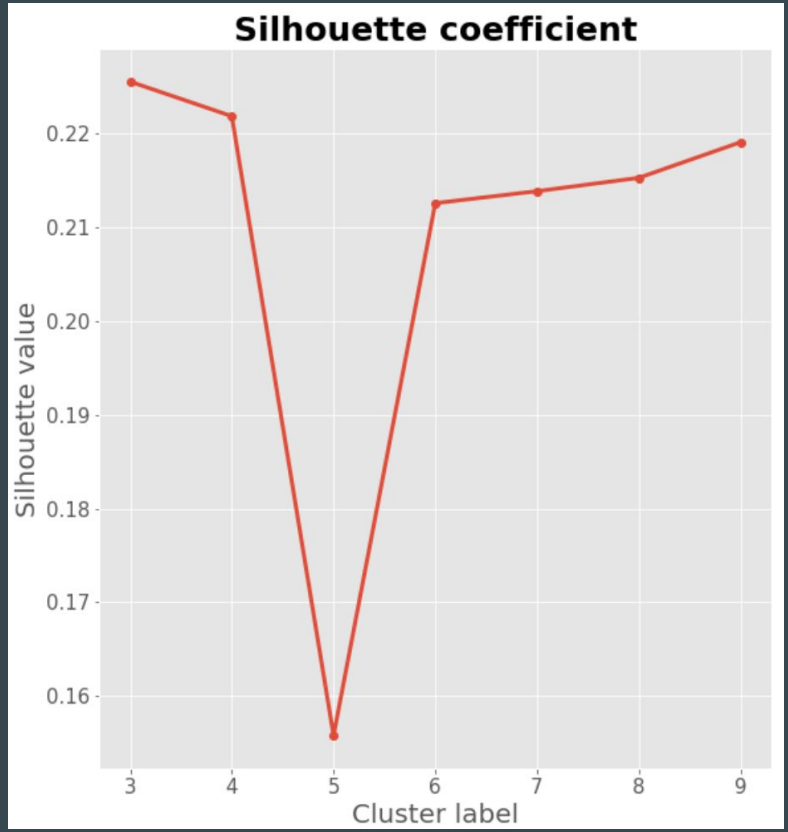


Réflexion sur le nombre k

Silhouette possible k=4

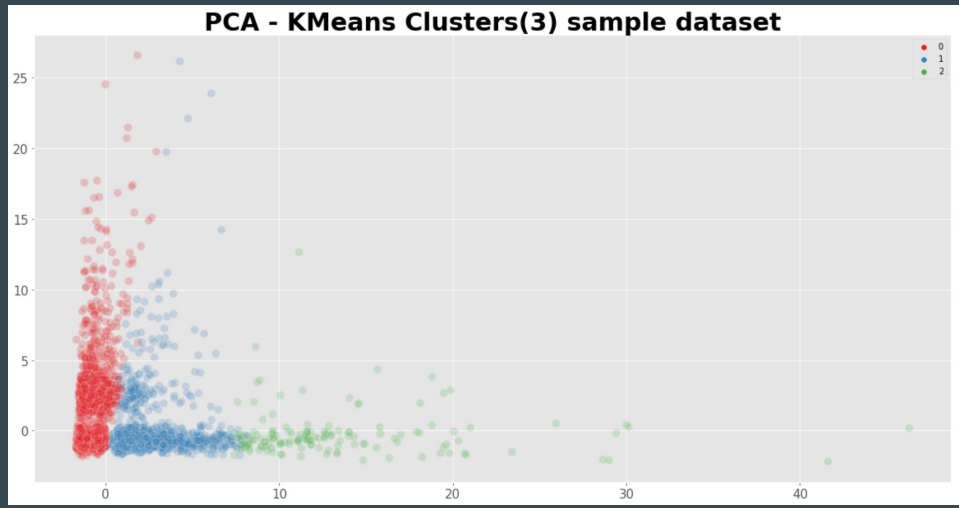
L'interprétabilité métier est également à prendre en compte.

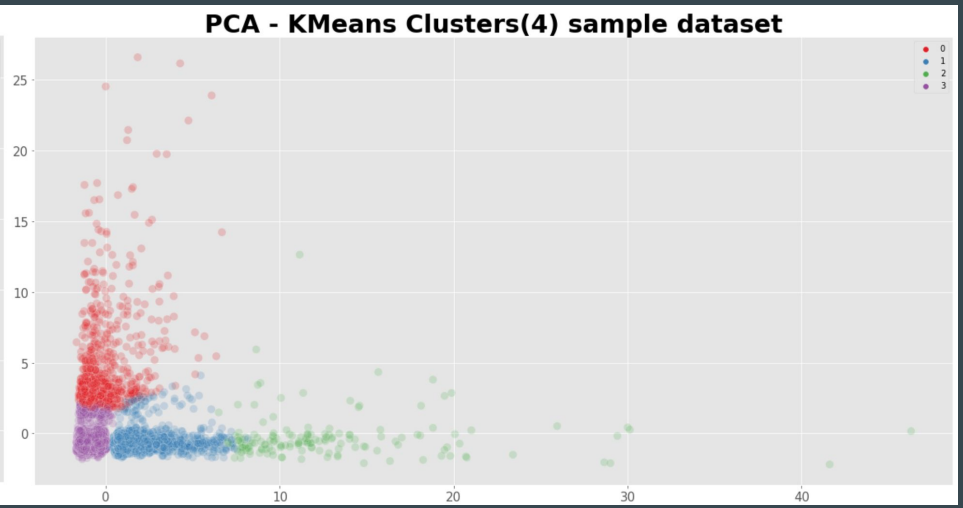
Un profilage des clusters clients peut devenir inadapté avec un nombre k trop élevé...



Clustering KMeans k= 3 et 4

Modélisation testée sur une population aléatoire de 10 000 individus.

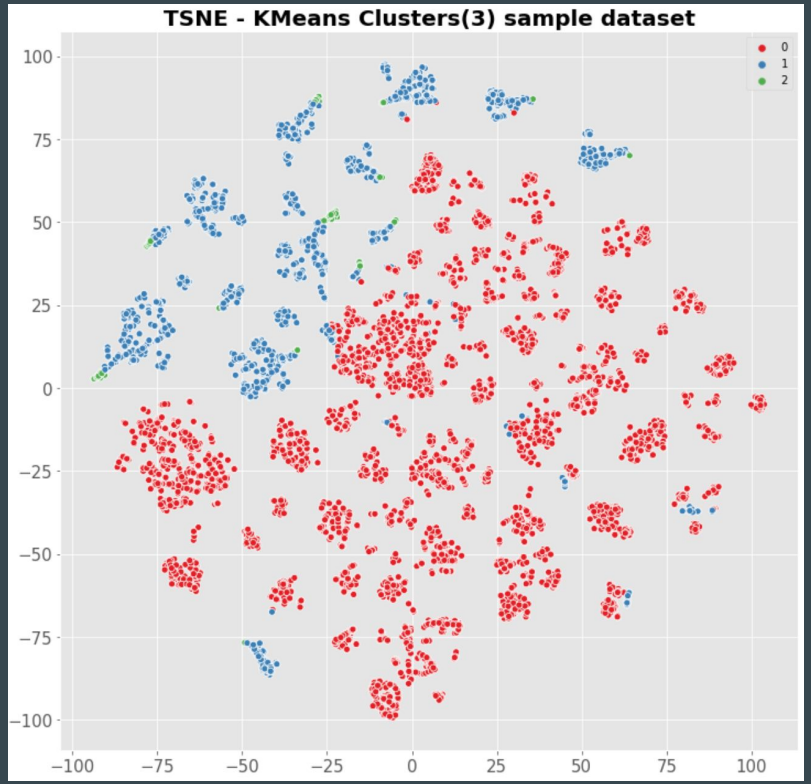


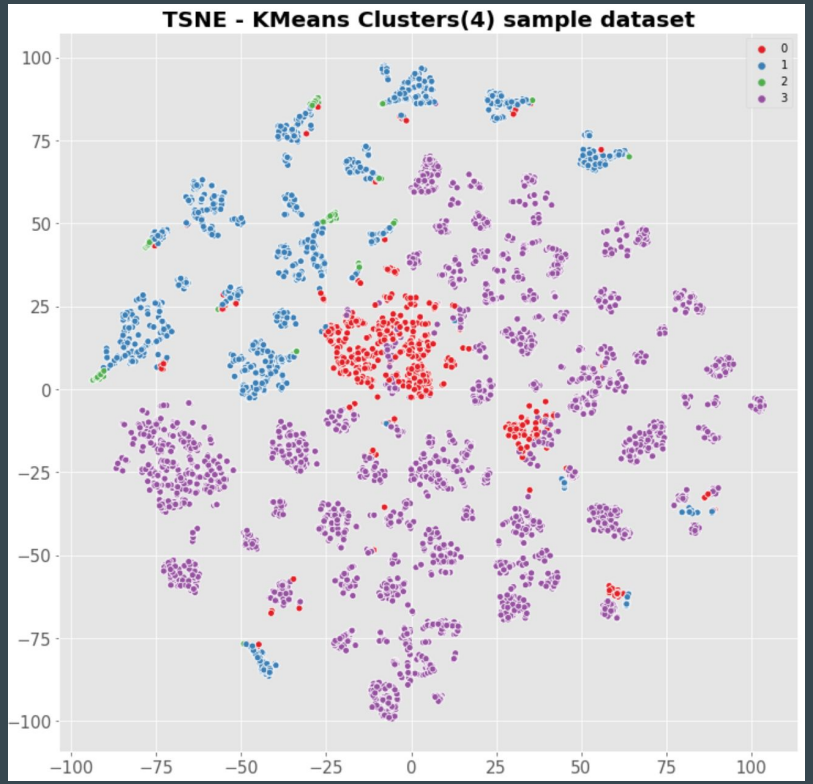


Silhouette : 0.227 - Time : 0.292 s

Silhouette : 0.222 - Time : 0.304 s

Visualisation facilité par T-SNE



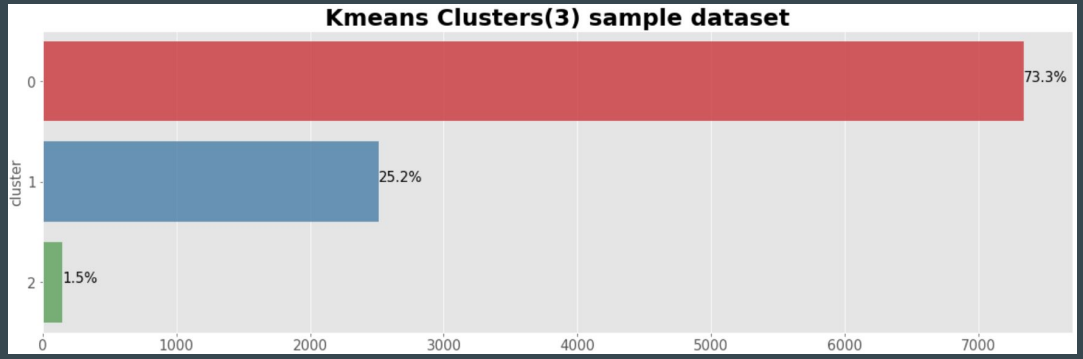


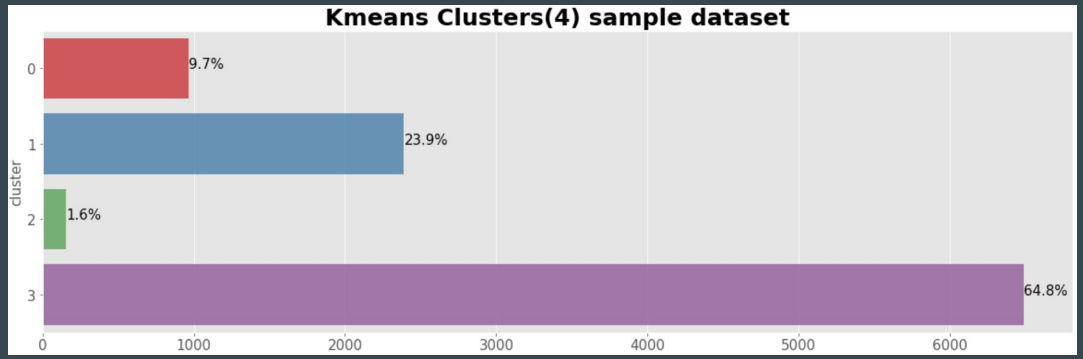
Algorithme de réduction de dimensionnalité plus lourd en temps de calcul.

Exploitation du CA selon les clusters

Le clustering en 4 partitions est plutôt intéressant, seulement l'interprétation métier du cluster 3 peut rester floue car 64.8% du CA.

Voyons si le clustering hiérarchique peut être utile dans la réflexion sur le nombre optimal de clusters.





Dendrogramme de regroupement hiérarchique

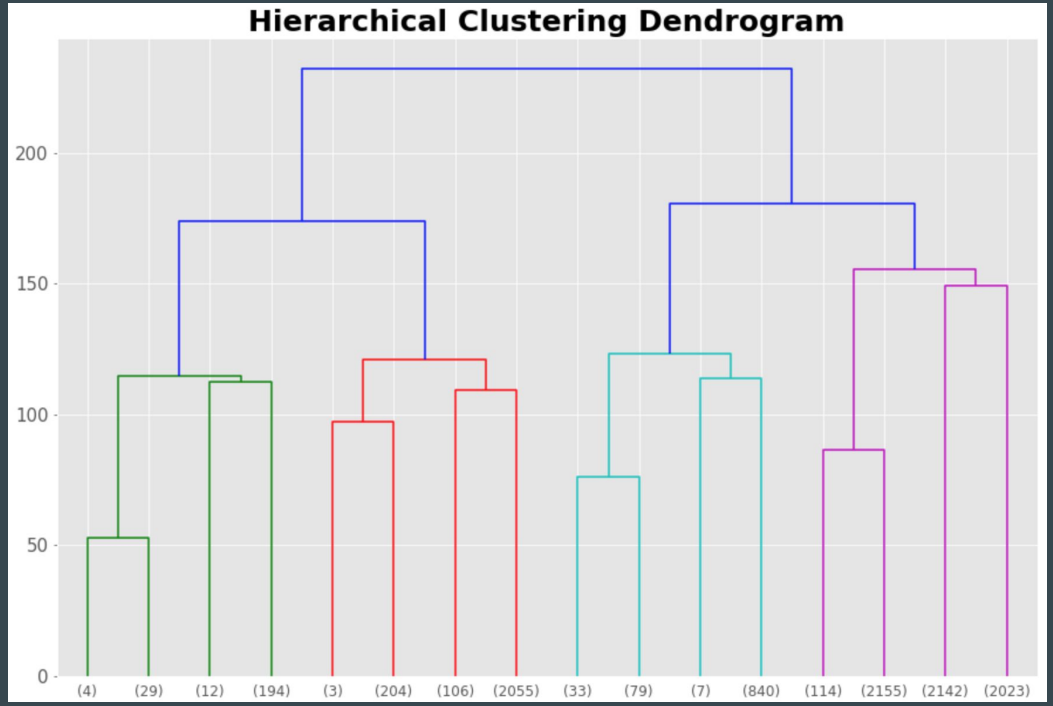
4 groupes différentiables

Test de modélisation Time : 4.728 s

```
AgglomerativeClustering(affinity='euclidean', compute_full_tree='auto',
                        connectivity=None, distance_threshold=None,
                        linkage='ward', memory=None, n_clusters=4)
```

Correspondance des groupes avec les 4 clusters du KMeans

kmeans / cah	0	1	2	3
1	21	52	2	164
2	232	561	38	1537
3	75	215	17	652
4	640	1562	100	4132

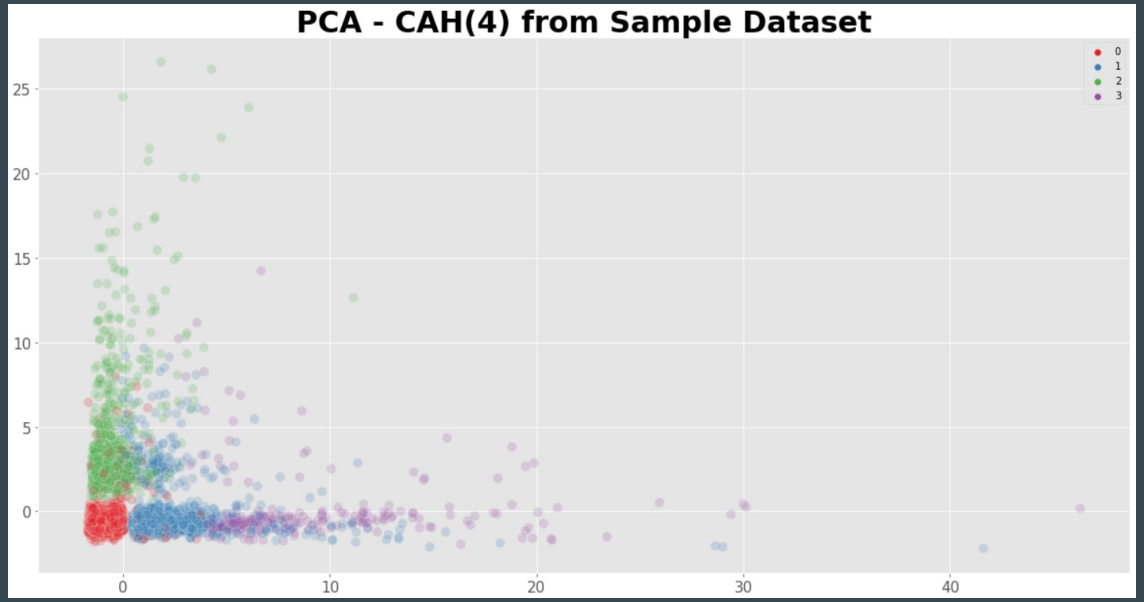


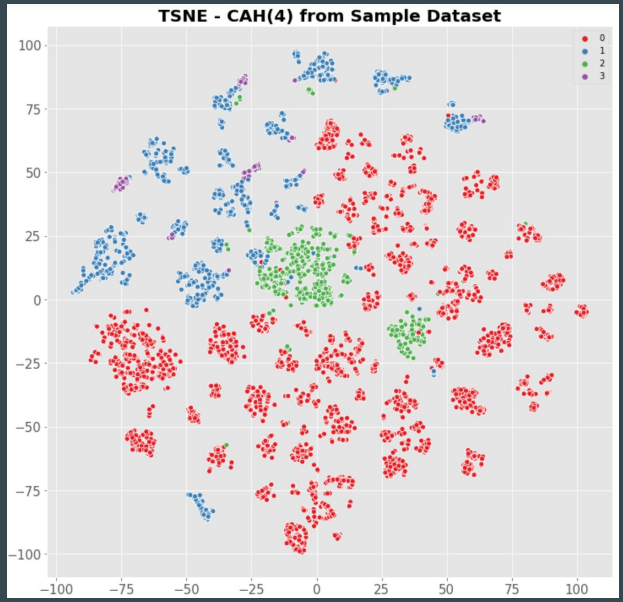
À chaque étape, on agrège les deux clusters les plus proches :

calculs lourds !

Clustering Hiérarchique

Modélisation testée sur une population aléatoire de 10 000 individus

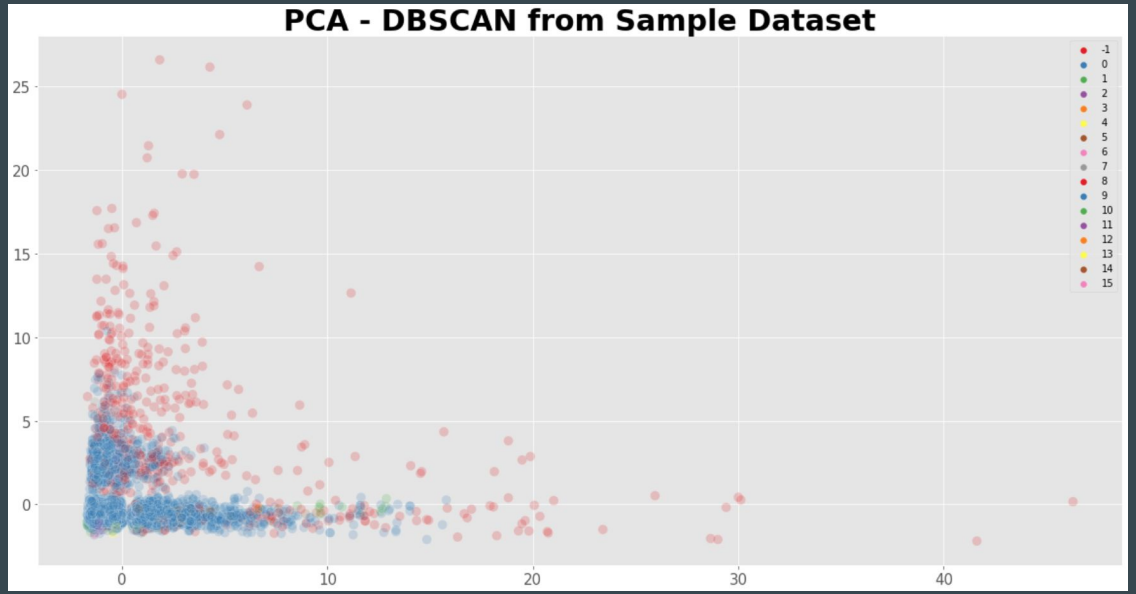


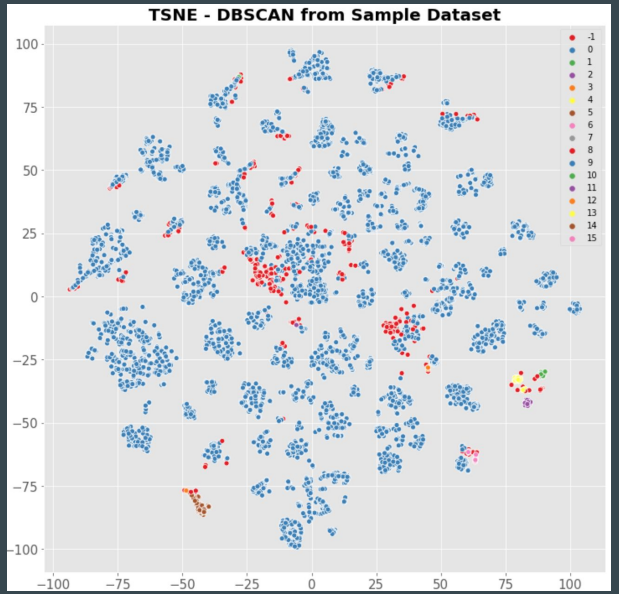


Silhouette : 0.211 - Algorithme AgglomerativeClustering

Clustering de densité

Modélisation testée sur une population aléatoire de 10 000 individus





Silhouette : - 0.249 Algorithme DBSCAN

Modélisation non adaptée à la structure de nos données

Modélisation sur l'échantillon complet

Les résultats des tests de clustering KMeans attestent d'une performance adaptée au type de population explorée dans le contexte métier.

Le modèle est-il stable pour l'ensemble de la population?
Le modèle est-il stable dans le temps?

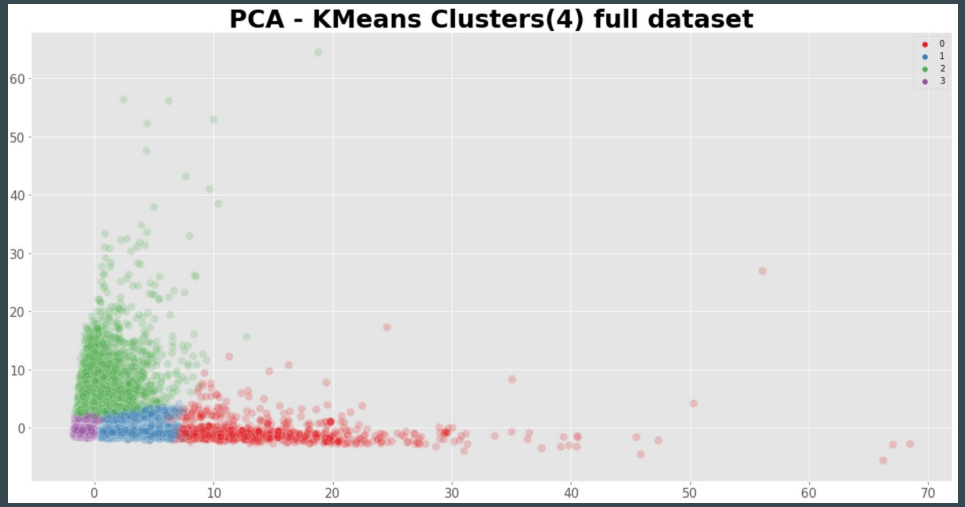
olist

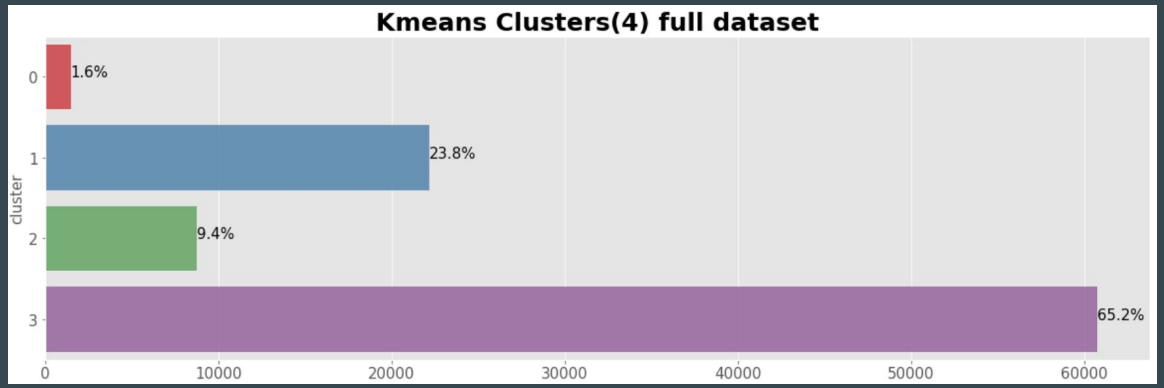
Clustering KMeans k=4

Modélisation effectuée sur l'ensemble de l'échantillon.

Silhouette : 0.222 - Time : 2.841 s

Stabilité avec une population 9 fois plus grande que celle des tests.





Clusters exprimés en %CA

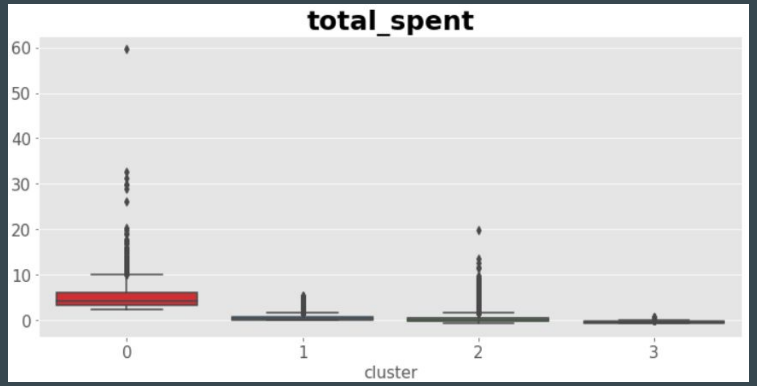
Cluster 3 représente 65% du CA

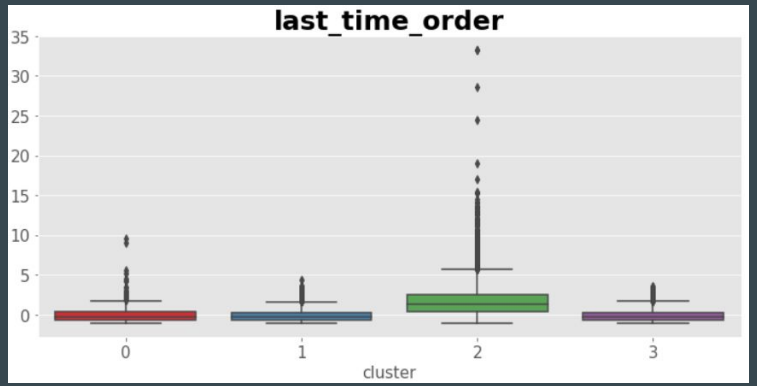
Centroïdes - Boxplot de comparaison

Approche comparative des clusters / features selon leurs centroïdes.

Visuellement des distinctions entre clusters est facilement identifiables par BoxPlot...

	0	1	2	3
average_basket	5.765688	0.630775	-0.263392	-0.332278
total_spent	5.297291	0.528437	0.343587	-0.370214
max_order_amount	5.759186	0.619737	-0.195626	-0.337781
min_order_amount	5.729365	0.635283	-0.320148	-0.324931
number_of_products_purchased	-0.135383	-0.229481	2.273738	-0.238349
bucket_quartile_lower25	-0.555014	-0.557619	0.237652	0.184985
bucket_quartile_25_50	-0.571902	-0.577405	0.118584	0.204996
bucket_quartile_50_75	-0.589235	-0.585504	-0.037311	0.231710
bucket_quartile_upper75	1.666572	1.670745	-0.305659	-0.604234
payment_boleto	-0.136085	-0.115241	0.421727	-0.013408
payment_credit_card	0.026951	-0.086799	1.618031	-0.202147
payment_debit_card	-0.010271	-0.025851	0.012269	0.008906
payment_voucher	-0.016606	-0.039816	0.130671	-0.002820
appliances	1.551207	0.066376	0.059568	-0.070210
auto	1.123695	0.118634	-0.010431	-0.068712
construction	0.890106	0.089639	0.061477	-0.062948
culture	0.206617	0.010704	-0.009362	-0.007490
electronics	2.328797	0.094065	0.049612	-0.098301
fashion	2.140404	0.198470	-0.052618	-0.117814
food	-0.050571	-0.006620	0.087213	-0.008934
garden	1.174623	0.033676	0.091222	-0.055097
health_beauty	1.855972	0.238432	-0.030110	-0.127642
hobbies	1.442459	0.019179	-0.026489	-0.038173
home	0.598219	0.147156	0.462829	-0.133901





Comparaison des clusters / CA global

L'approche Business permet de comparer des volumes de CA entre eux.

Des thématiques très peu représentatives de l'activité : Animaux, Alimentation, Culture...

Une segmentation plus détaillée du Cluster 3 est recommandée.

Voyons également les différences des clusters sur les aspects de fidélisation, de fréquence d'achat et de localisation client.

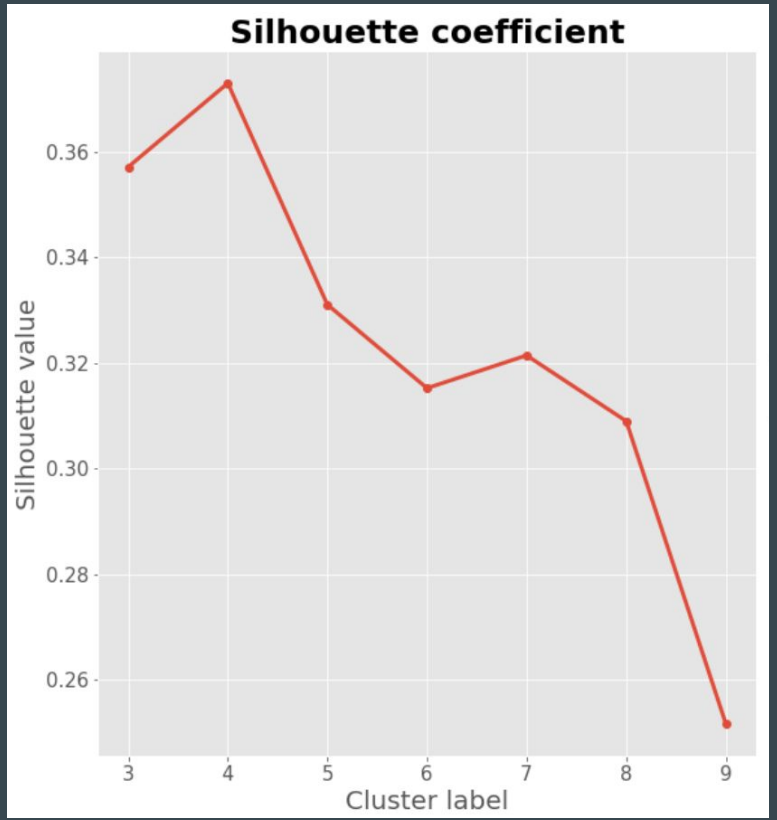
Cluster	0	1	2	3
CA(%)	1.60	23.80	9.40	65.20
appliances	0.12	1.31	0.53	3.57
auto	0.13	0.96	0.34	2.31
construction	0.03	0.61	0.25	1.69
culture	0.00	0.17	0.05	0.45
electronics	2.24	2.94	1.13	7.74
fashion	0.28	2.08	0.84	5.75
food	0.00	0.08	0.03	0.29
garden	0.05	0.69	0.26	2.09
health_beauty	0.33	3.23	1.27	8.47
hobbies	0.04	0.27	0.13	0.79
home	0.22	3.46	1.35	8.89
office	0.06	1.02	0.41	2.62
pets	0.01	0.29	0.14	0.92
sports_leisure	0.15	1.74	0.70	4.60
toys	0.03	0.77	0.29	2.00

Réflexion sur le nombre k

Silhouette optimale k=4

Analyse effectuée sur le cluster 3 (65% CA) du modèle précédent.

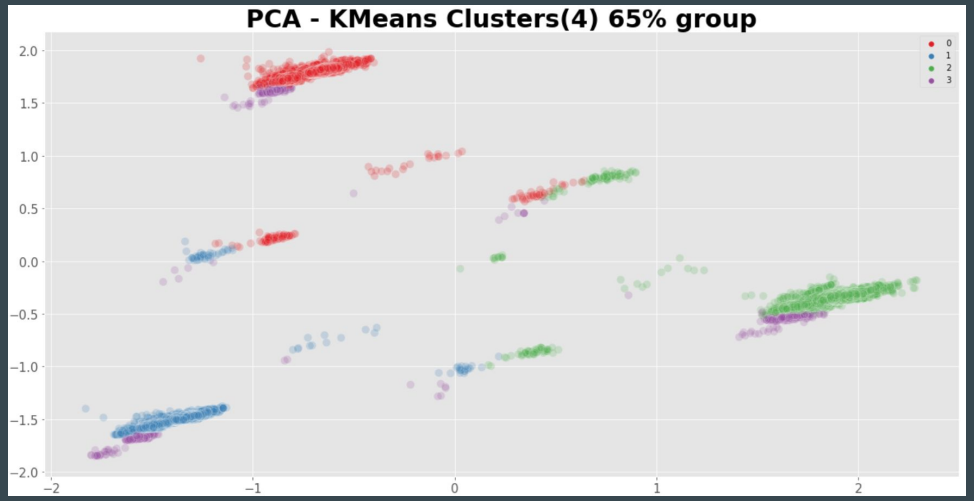
Voyons si un nouveau clustering KMeans peut être utile pour mieux comprendre le volume important du cluster 3.

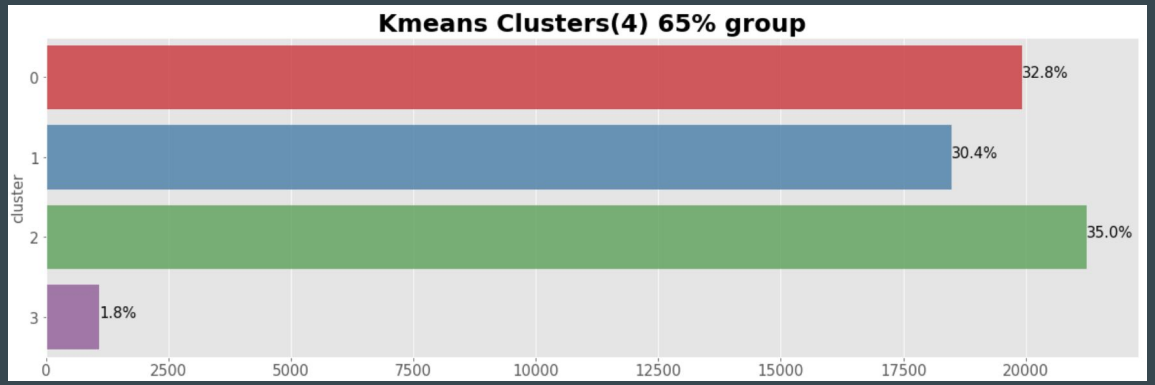


Clustering KMeans du groupe à 65%

Modélisation effectuée sur le cluster 3 (65% CA) du modèle précédent.

Silhouette : 0.373 - Time : 0.835 s





Comparaison des clusters / CA global

Cluster	0	1	2	3
CA(%)	1.60	23.80	9.40	65.20
appliances	0.12	1.31	0.53	3.57
auto	0.13	0.96	0.34	2.31
construction	0.03	0.61	0.25	1.69
culture	0.00	0.17	0.05	0.45
electronics	2.24	2.94	1.13	7.74
fashion	0.28	2.08	0.84	5.75
food	0.00	0.08	0.03	0.29
garden	0.05	0.69	0.26	2.09
health_beauty	0.33	3.23	1.27	8.47
hobbies	0.04	0.27	0.13	0.79
home	0.22	3.46	1.35	8.89
office	0.06	1.02	0.41	2.62
pets	0.01	0.29	0.14	0.92
sports_leisure	0.15	1.74	0.70	4.60
toys	0.03	0.77	0.29	2.00

	3.0	3.1	3.2	3.3
CA(%)	30.40	32.80	35.00	1.80
appliances	7.84	5.91	4.91	0.27
auto	7.42	6.65	4.17	0.32
construction	6.50	6.16	4.06	0.27
culture	2.07	1.54	1.56	0.02
electronics	8.38	10.02	7.80	0.35
fashion	12.65	10.75	7.76	0.67
food	1.62	0.97	1.96	0.19
garden	7.38	3.24	4.00	0.25
health_beauty	16.82	12.38	4.58	0.71
hobbies	3.47	3.25	3.40	0.17
home	20.38	12.15	2.84	0.77
office	9.46	8.65	0.79	0.43
pets	5.78	4.13	1.67	0.05
sports_leisure	12.50	8.79	1.26	0.46
toys	9.04	5.44	0.50	0.14

Cluster	0	1	2	3
payment_boleto	7.55	1.09	9.07	3.09
payment_credit_card	30.99	24.45	20.07	9.47
payment_debit_card	14.81	15.98	17.02	52.82
payment_voucher	1.40	1.20	2.92	0.89

Achat One Shot > Peu de fidélisation > Faible fréquence d'achat et d'avis produits.

Clusters peu différenciables selon les critères de géolocalisation.
Tous les clusters sont mieux représentés dans l'état de São Paulo suivi par Rio de Janeiro.

Identification Marketing des clusters

Cluster 0 ***"Gros consommateur"***	• Les plus dépensiers • Fort intérêt pour les technologies et la mode
Cluster 1 ***"Gros consommateur"***	• Les dépensiers (*qui prennent soins d'eux*) • Soins du corps et tenue vestimentaire
Cluster 2 ***"Petit consommateur"***	• Les économes • L'univers de l'indoor avant tout...
Cluster(s) 3 ***"Consommateur généraliste"***	• Multiple visage • Multi-produits

Cluster 3.0 ***"Petit consommateur"***	• Les économes généralistes • Multi-familles de produits
Cluster 3.1 ***"Moyen consommateur"***	• Les semi-dépensiers généralistes • Multi-familles de produits
Cluster 3.2 ***"Gros consommateur"***	• Les plus dépensiers généralistes • Multi-familles de produits
Cluster 3.3 **"Petit/Moyen consommateur"**	• Les adeptes du débit immédiat • Multi-familles de produits

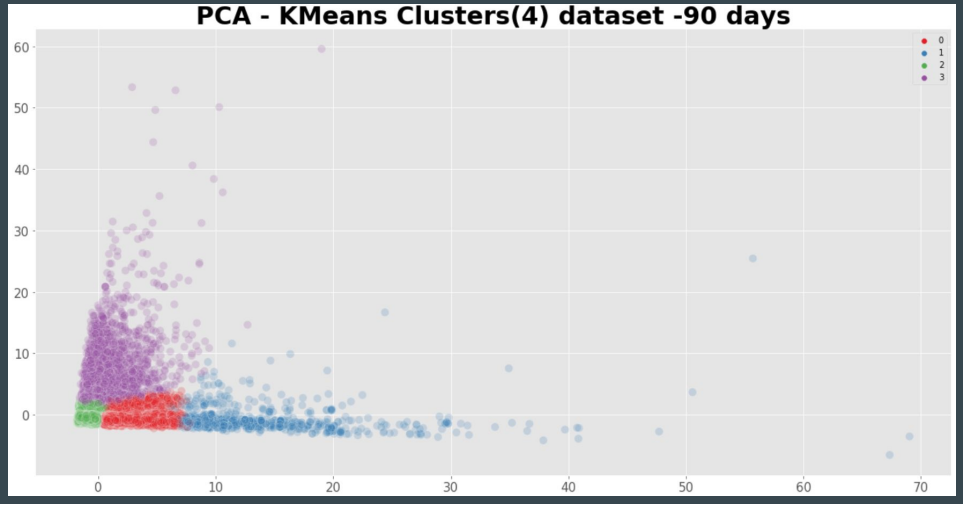
Test de stabilité des clusters

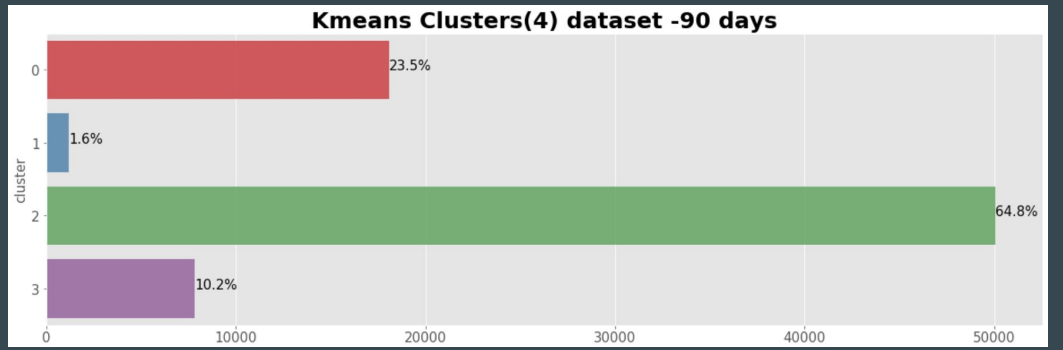
Modélisation effectuée sur l'échantillon complet -90 jours d'historique

Silhouette : 0.224 - Time : 2.311 s

Quelque soit l'échantillon notons une excellente stabilité du partitionnement.

`ARI score : 0.836 permet de valider les similitudes du modèle.`





Contrat de maintenance

La proposition de contrat de maintenance est basée sur une analyse de la stabilité des segments au cours du temps.

olist

Méthodes utilisées

Analyse exploratoire des données temporelles :

Fréquence identifiée 3 Mois.

Possibilité d'implémenter une méthode *.predict()* du modèle pour établir un update hebdomadaire ou mensuel.

Analyse de la stabilité du Clustering :

Entraînement sur les données à -90 Jours.

Stabilité du nombre de clusters.

Stabilité du coef. de silhouette.

ARI score proche de 1.

Conclusion

Olist peut améliorer ses campagnes de communication en exploitant périodiquement la segmentation proposée...

olist

Segmentation en 4 partitions stables au cours du temps via KMeans.
Amélioration de l'analyse par un clustering supplémentaire sur le groupe majoritaire (65% CA).

Similarités communes identifiées sur :
la faible fréquence d'achat, le peu de récence ou encore la localisation.

Discussion possible sur l'idée de combiner une méthode RFM complémentaire au modèle KMeans.

Amélioration possible du manque d'adhésion au site :
Vente One Shot > Communauté à développer.

Code fourni réutilisable (convention PEP8) par Olist.