PROJET 5

SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE

Plan de la présentation

1. Sujet
2. Analyse Exploratoire
3. Clustering
4. Contrat de mainteancne
5. Conclusion

Sujet

Je suis consultant pour Olist, une entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne.

Olist souhaite que je fournisse aux équipes d'e-commerce une segmentation des clients qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.
Mon objectif est de comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.
Je dois fournir à l'équipe marketing une description actionable de votre segmentation et de sa logique sous-jacente pour une utilisation optimale, ainsi qu'une proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps.

Les données

Plusieurs jeux de données concernant :

- les clients
- les commandes
- les avis des commandes
- les paiements

- les articles
- les vendeurs
- les localisations
- les traductions des catégories de produit

Objectif : réunir les variables pertinentes dans un seul et unique jeu de données où une ligne = un client unique

Analyse exploratoire
Les clients

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
count	99441	99441	99441.000000	99441	99441
unique	99441	96096	NaN	4119	27
top	06b8999e2fba1a1fbc88172c00ba8bc7	8d50f5eadf50201ccdcedfb9e2ac8455	NaN	sao paulo	SP
freq	1	17	NaN	15540	41746
mean	NaN	NaN	35137.474583	NaN	NaN
std	NaN	NaN	29797.938996	NaN	NaN
min	NaN	NaN	1003.000000	NaN	NaN
25%	NaN	NaN	11347.000000	NaN	NaN
50%	NaN	NaN	24416.000000	NaN	NaN
75%	NaN	NaN	58900.000000	NaN	NaN
max	NaN	NaN	99990.000000	NaN	NaN

Commandes

	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at
count	99441	99441	99441	99441	99281
unique	99441	99441	8	98875	90733
top	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2018-04-11 10:48:14	2018-02-27 04:31:10
freq	1	1	96478	3	9

Dans un premier temps, on souhaite attribuer une commande à un customer_id non unique.
On souhaite conserver order_purchase_timestamp qui est la date qui nous intéresse le plus parmi les variables proposées.

Articles achetés

	order_id	order_item_id	product_id	seller_id	shipping_limit_date
count	112650	112650.000000	112650	112650	112650
unique	98666	NaN	32951	3095	93318
top	8272b63d03f5f79c56e9e4120aec44ef	NaN	aca2eb7d00ea1a7b8ebd4e68314663af	6560211a19b47992c3666cc44a7e94c0	2017-07-21 18:25:23
freq	21	NaN	527	2033	21
mean	NaN	1.197834	NaN	NaN	NaN
std	NaN	0.705124	NaN	NaN	NaN
min	NaN	1.000000	NaN	NaN	NaN
25%	NaN	1.000000	NaN	NaN	NaN
50%	NaN	1.000000	NaN	NaN	NaN
75%	NaN	1.000000	NaN	NaN	NaN
max	NaN	21.000000	NaN	NaN	NaN

Ici nous souhaitons récupérer order_item_id qui est le numéro attribué à l'article dans une commande.
Lorsque nous regrouperons par commande, cela nous servira pour savoir le nombre de produits achetés lors d'une commande.

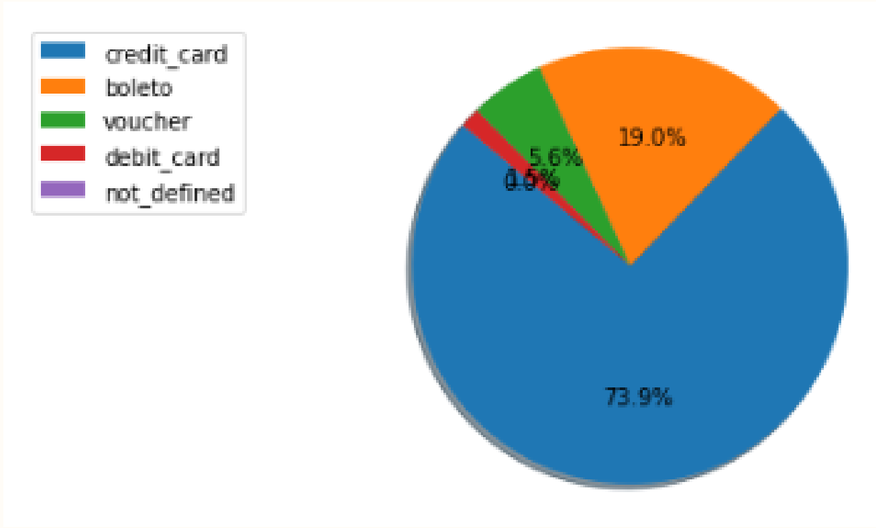
Avis des commandes

	review_id	order_id	review_score	review_comment_title	review_comment_message
count	99224	99224	99224.000000	11568	40977
unique	98410	98673	NaN	4527	36159
top	7b606b0d57b078384f0b58eac1d41d78	c88b1d1b157a9999ce368f218a407141	NaN	Recomendo	Muito bom
freq	3	3	NaN	423	230
mean	NaN	NaN	4.086421	NaN	NaN
std	NaN	NaN	1.347579	NaN	NaN
min	NaN	NaN	1.000000	NaN	NaN
25%	NaN	NaN	4.000000	NaN	NaN
50%	NaN	NaN	5.000000	NaN	NaN
75%	NaN	NaN	5.000000	NaN	NaN
max	NaN	NaN	5.000000	NaN	NaN

On souhaite conserver uniquement la note, le review_score, attribué à la commande.

Les paiements

	order_id	payment_sequential	payment_type	payment_installments	payment_value
count	103886	103886.000000	103886	103886.000000	103886.000000
unique	99440	NaN	5	NaN	NaN
top	fa65dad1b0e818e3ccc5cb0e39231352	NaN	credit_card	NaN	NaN
freq	29	NaN	76795	NaN	NaN
mean	NaN	1.092679	NaN	2.853349	154.100380
std	NaN	0.706584	NaN	2.687051	217.494064
min	NaN	1.000000	NaN	0.000000	0.000000
25%	NaN	1.000000	NaN	1.000000	56.790000
50%	NaN	1.000000	NaN	1.000000	100.000000
75%	NaN	1.000000	NaN	4.000000	171.837500
max	NaN	29.000000	NaN	24.000000	13664.080000



On souhaite conserver uniquement la note, le review_score, attribué à la commande.

Dataframe final

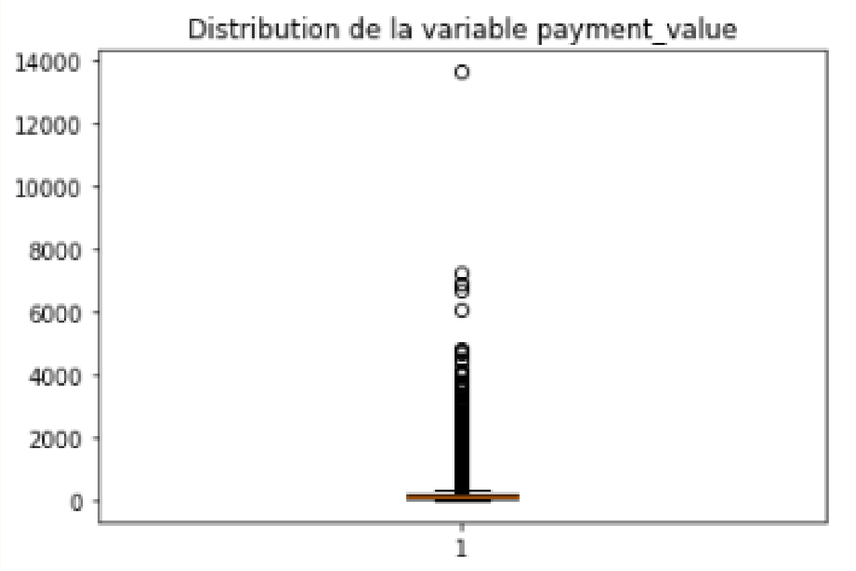
	days_since_last_purchase	christmas_purchase	number_of_products_bought	review_score	payment_installments	payment_value	number_of_purchase
count	96093.000000	96093.000000	96093.000000	96093.000000	96093.000000	96093.000000	96093.000000
mean	289.102337	0.123076	1.139099	4.085067	2.901509	158.717482	1.087228
std	153.128232	0.327104	0.525246	1.336532	2.677539	219.499964	0.493251
min	0.000000	0.000000	1.000000	1.000000	0.000000	1.856818	1.000000
25%	165.000000	0.000000	1.000000	4.000000	1.000000	60.850000	1.000000
50%	271.000000	0.000000	1.000000	5.000000	2.000000	103.750000	1.000000
75%	398.000000	0.000000	1.000000	5.000000	4.000000	175.090000	1.000000
max	772.000000	1.000000	21.000000	5.000000	24.000000	13664.080000	33.000000

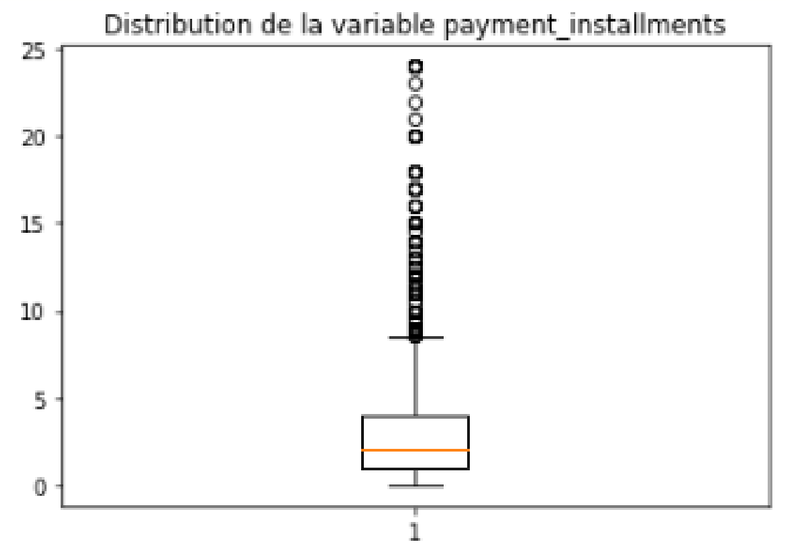
Transformer mes données

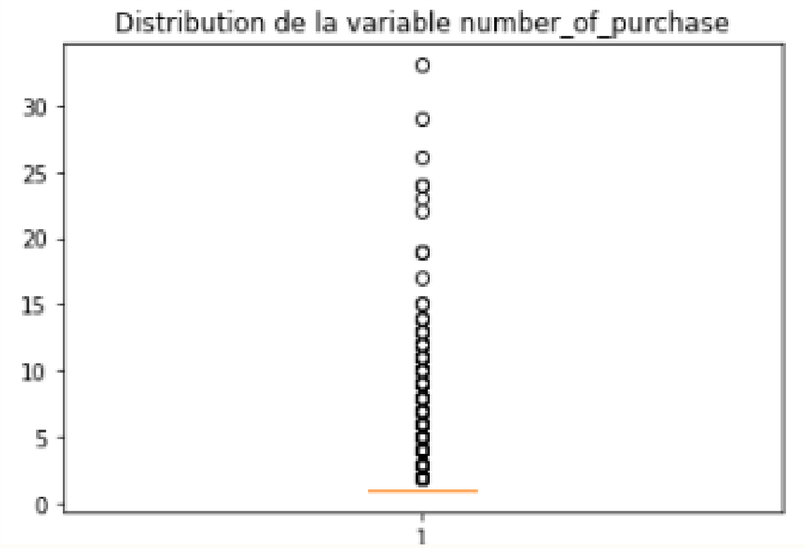
Plusieurs trasnformations :

- Création de nouvelles variables :
  - days_since_last_purchase
  - christmas_purchase
  - number_of_purchase
- Standard Scaler
- MinMax Scaler
- Passage au log10
- Création de plusieurs dataframes

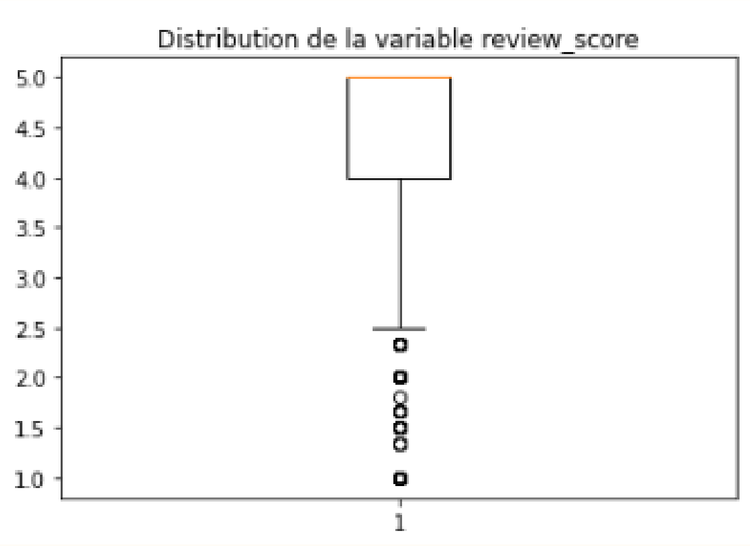
Quelques graphiques





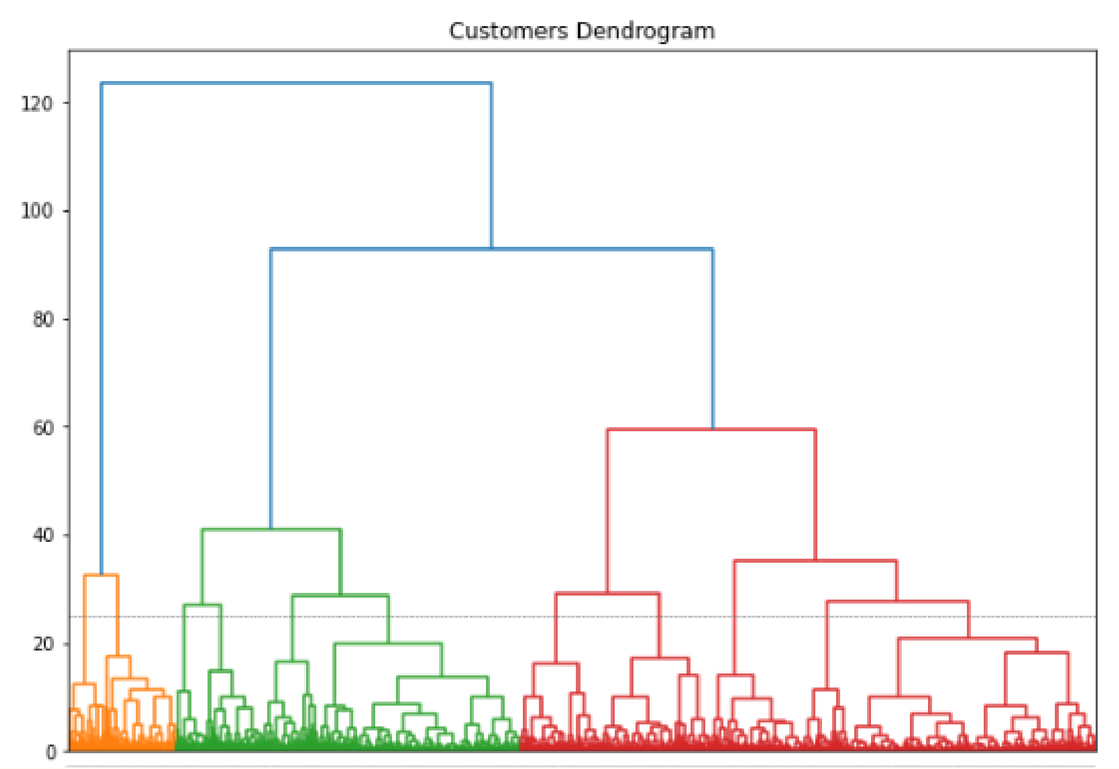


1	90285
2	4569
3	740
4	256
5	82
6	63
7	31
8	13
9	12
10	8
11	7
12	6
13	4
14	3
15	2
24	2
19	2
22	1
26	1
29	1
17	1
33	1
23	1

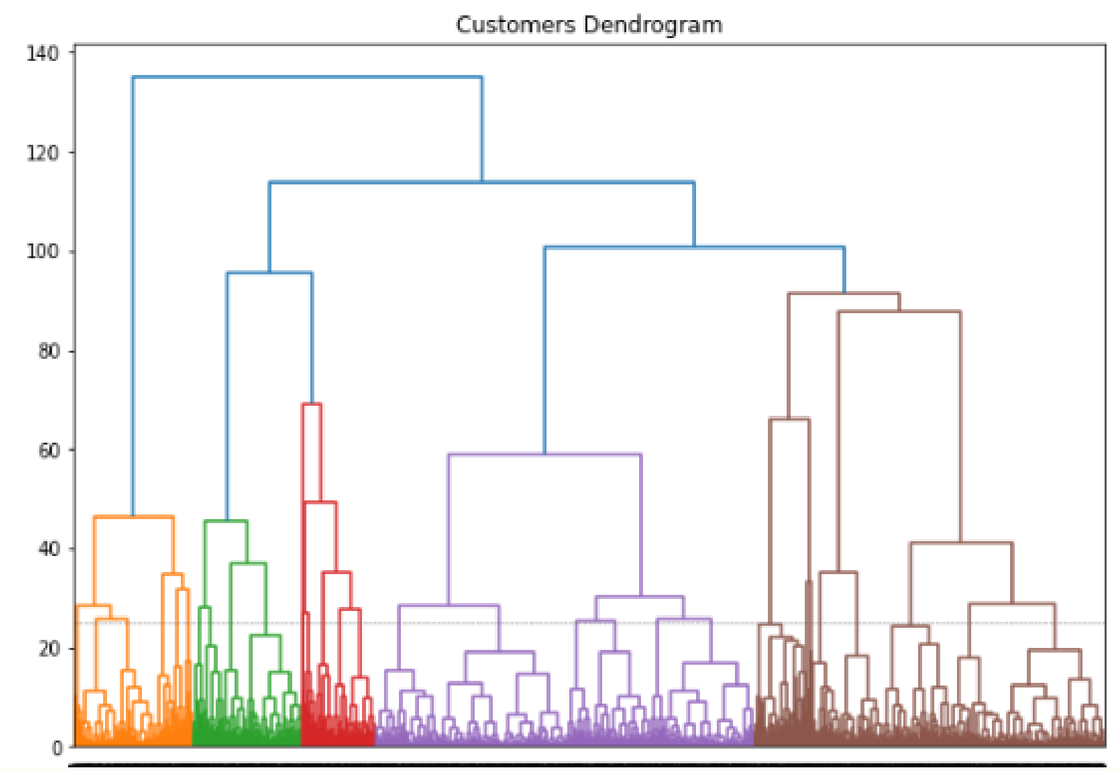


Clustering hiérarchique

Données min max scalées

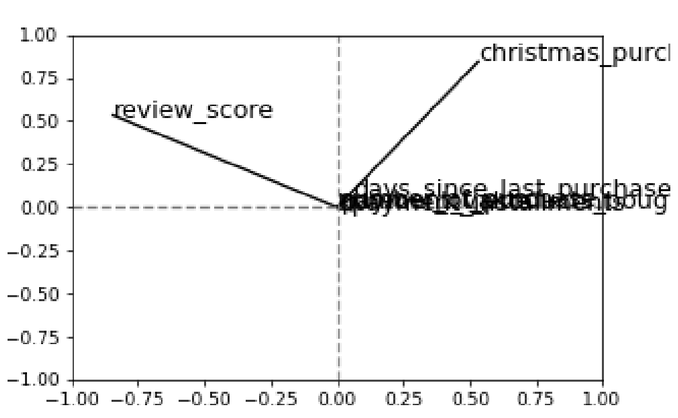


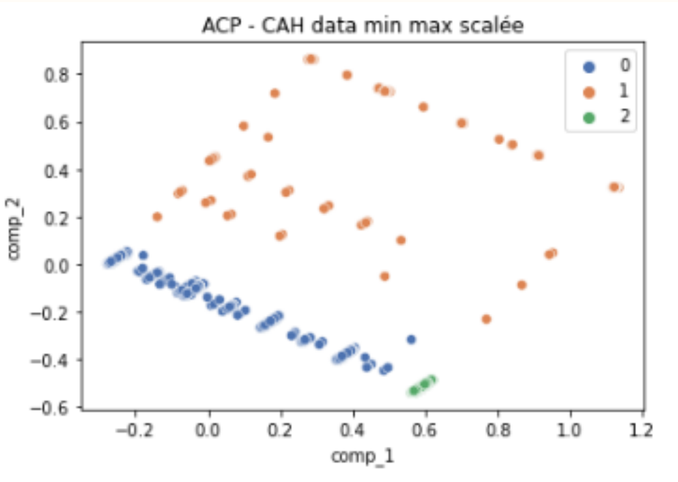
Données standardisées



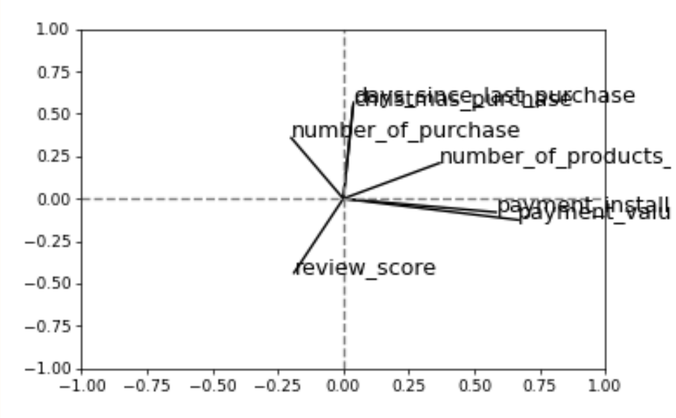
ACP

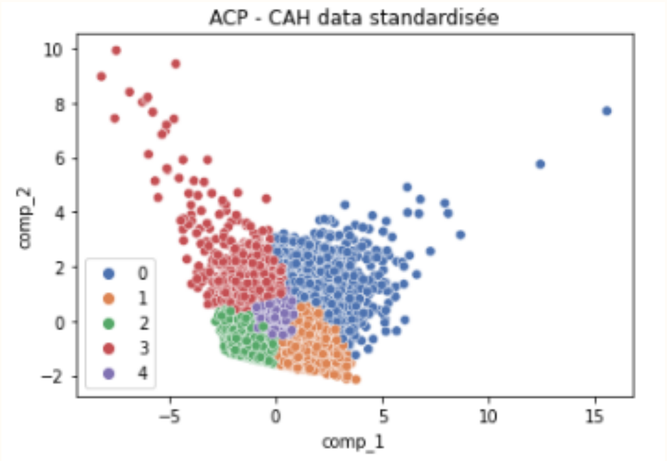
Données min max scalées



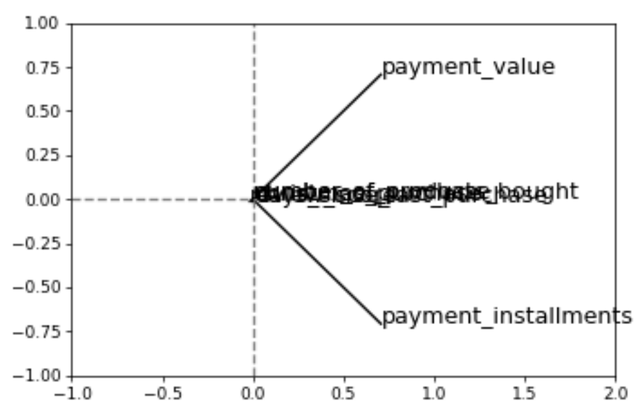


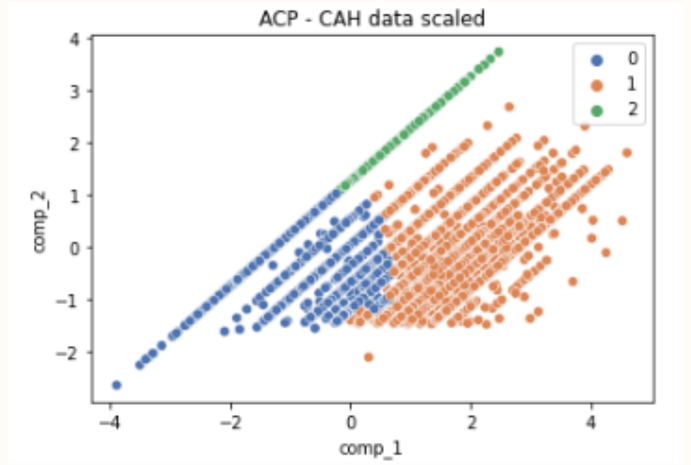
Données standardisées





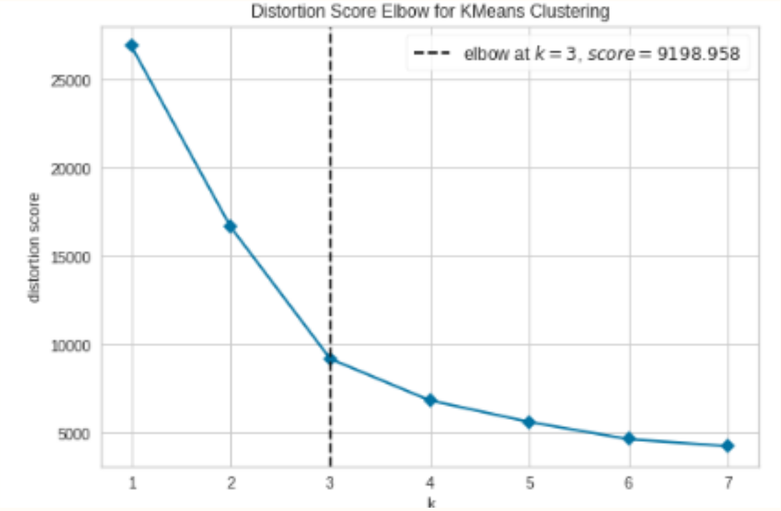
Données mélange de scalers

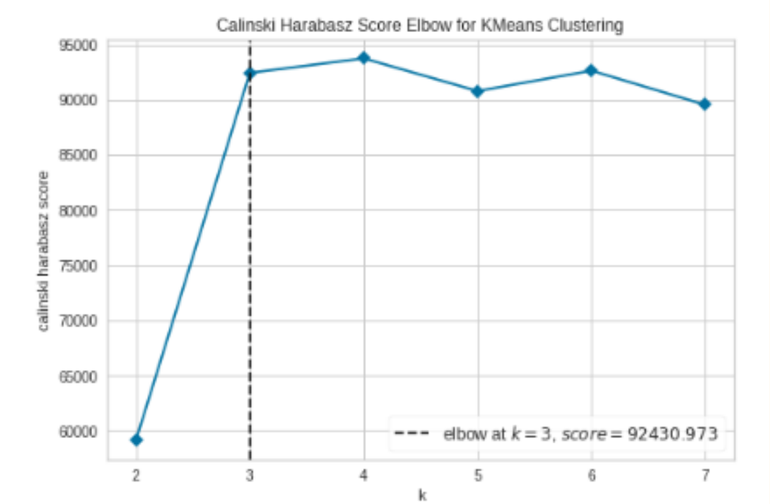




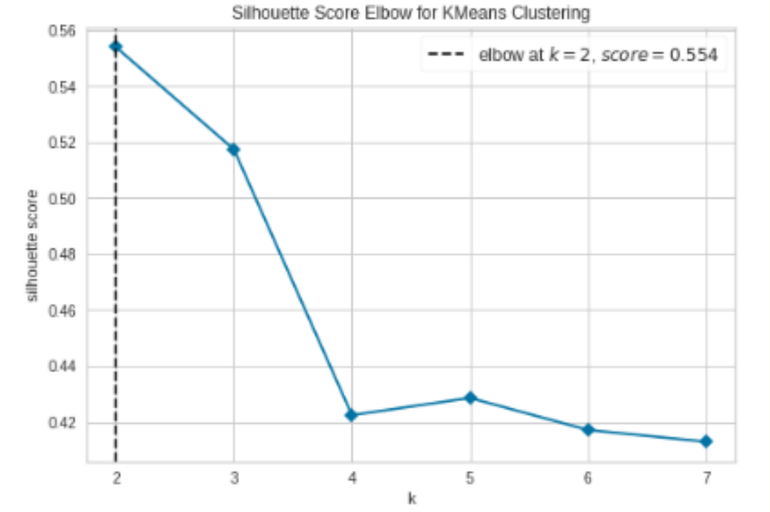
Clustering Kmeans

Difficulté : trouver le bon "compromis" entre des metrics intéressantes
et un clustering interprétable d'un point de vue métier

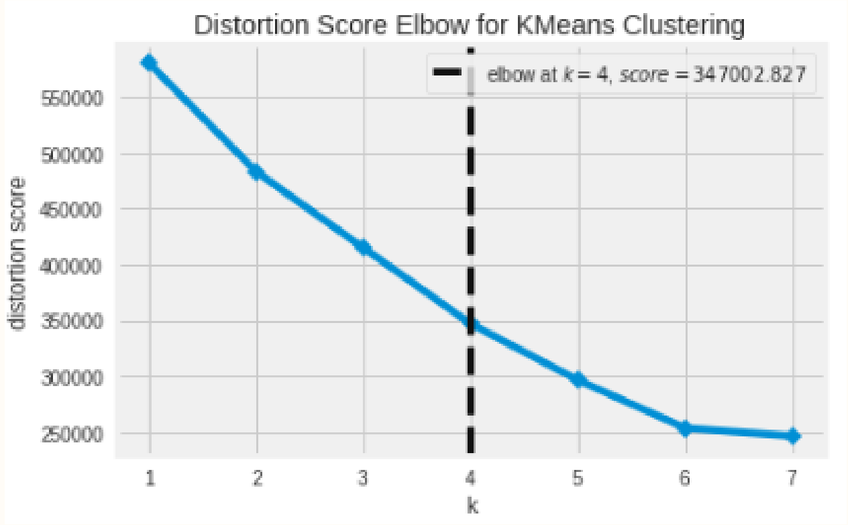


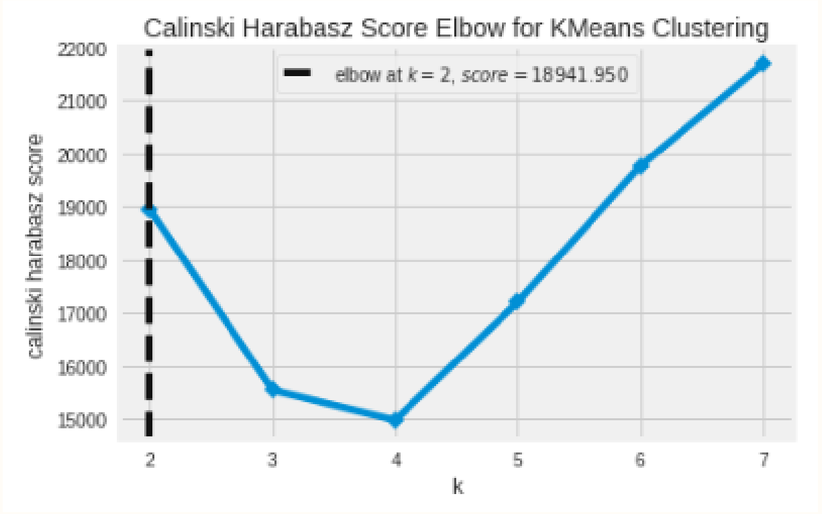


Data min max scalées

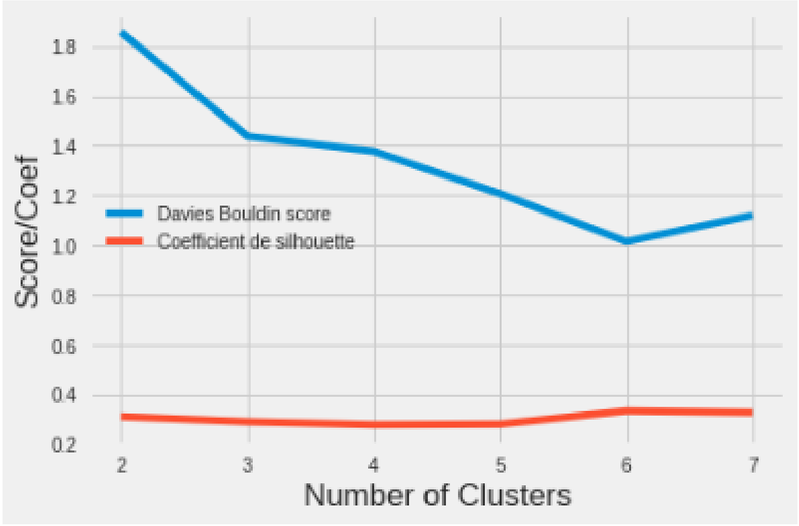


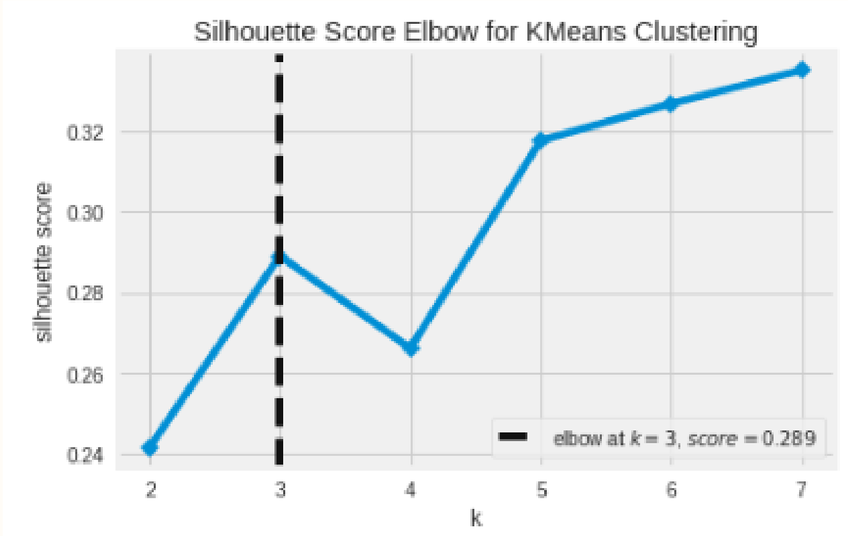
Clustering Kmeans





Data standardisées





Clustering Kmeans

Clustering sur données min max scalées

label	days_since_last_purchase	christmas_purchase	number_of_products_bought	review_score	payment_installments	payment_value	number_of_purchase
0	315.822441	0.005257	1.168386	2.887843	2.935604	151.038329	1.140864
1	170.487552	0.000502	1.105958	4.746711	2.714502	153.664245	1.070139
2	322.786790	0.992506	1.113251	4.537874	2.930370	149.301370	1.079303
3	272.950636	0.000262	1.288647	1.153402	3.093553	190.679203	1.071878
4	471.796368	0.001030	1.115853	4.737539	3.072781	157.027882	1.107806
5	323.171504	0.996421	1.292313	1.215825	3.328850	185.759742	1.092496

Clustering sur données mélange de scalers

label	days_since_last_purchase	christmas_purchase	number_of_products_bought	review_score	payment_installments	payment_value	number_of_purchase
0	295.126351	0.127057	1.169919	4.038313	5.921799	166.991818	1.062976
1	282.541631	0.122328	1.097548	4.124591	1.004280	90.893549	1.068058
2	298.481194	0.121264	1.061768	4.127832	3.105582	75.981788	1.072841
3	287.273941	0.118469	1.020338	4.186168	1.072095	37.627720	1.211096
4	293.499138	0.120974	1.336054	3.903945	7.476134	605.589701	1.037198
5	279.436262	0.127456	1.292701	4.010081	1.307953	266.641440	1.049742

Clustering Kmeans

Clustering sur données standardisées

label	days_since_last_purchase	christmas_purchase	number_of_products_bought	review_score	payment_installments	payment_value	number_of_purchase
0	322.619786	0.995183	1.078630	3.958909	2.815991	134.921435	1.074374
1	285.096479	0.138225	3.829310	3.400479	3.746511	372.774070	1.071702
2	168.005245	0.000668	1.058264	4.649303	1.944422	114.618087	1.063125
3	282.928119	0.001191	1.143777	1.500095	2.520850	135.162720	1.080516
4	473.759252	0.002676	1.065256	4.582109	2.202950	112.927885	1.156145
5	291.855887	0.025945	1.110768	4.280131	8.378752	434.783158	1.046697

Interprétation

- 0 : Commande pour les fêtes de fin d'année, satisfaits, payent en 3 fois et panier moyen de 135$
- 1 : Commande de plusieurs articles, plutôt satisfaits, payent en 4 fois, panier moyen de 370$
- 2 : Commande de moins de 6 mois, très satisfaits, payent en 2 fois, panier moyen de 115$
- 3: Commande il y a moins d'un an, plutôt mécontents, payent en 2-3 fois, panier moyen de 135$
- 4 : Commande il y a plus d'un an, pourtant très satisfaits, payent en 2 fois, panier moyen de 115$
- 5 : Commande il y a moins d'un an, très satisfaits, payent en 8 fois, panier moyen de 435$

Proposition marketing

- Relancer les anciens clients, remercier les nouveaux
- Proposer le paiement en plusieurs fois, à ceux qui payent en une fois comme à ceux qui ont déjà cette habitude
- Stimuler les clients de Noel pour les faire commander à d'autres périodes de l'année
- Proposer une promotion aux clients mécontents

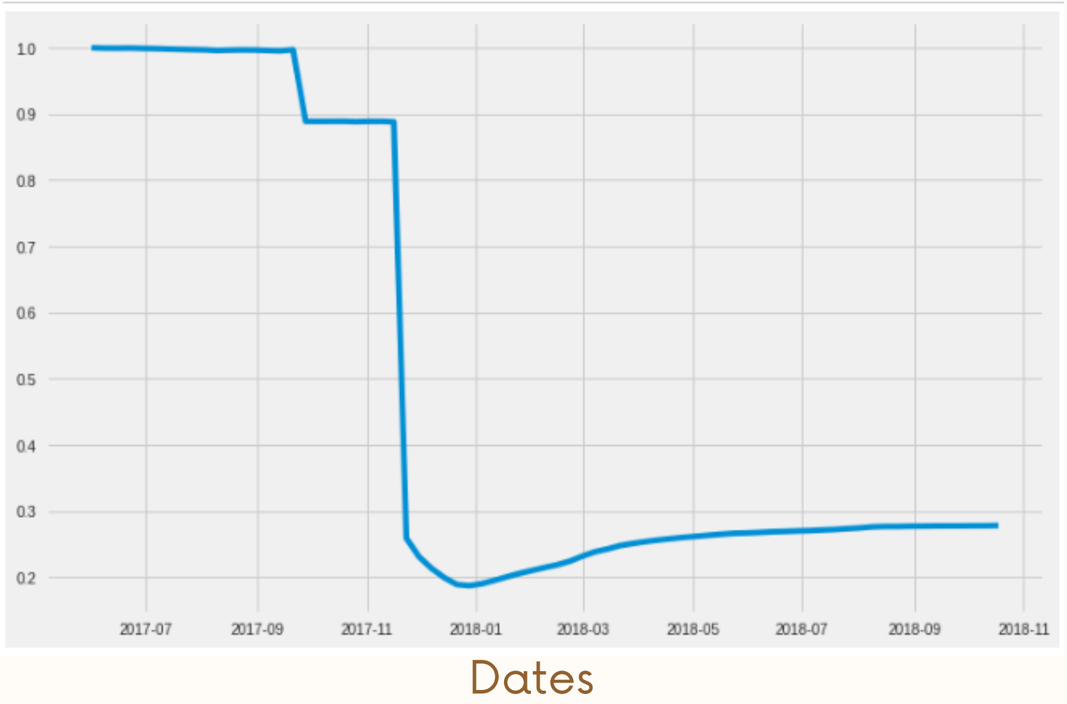
Contrat de maintenance

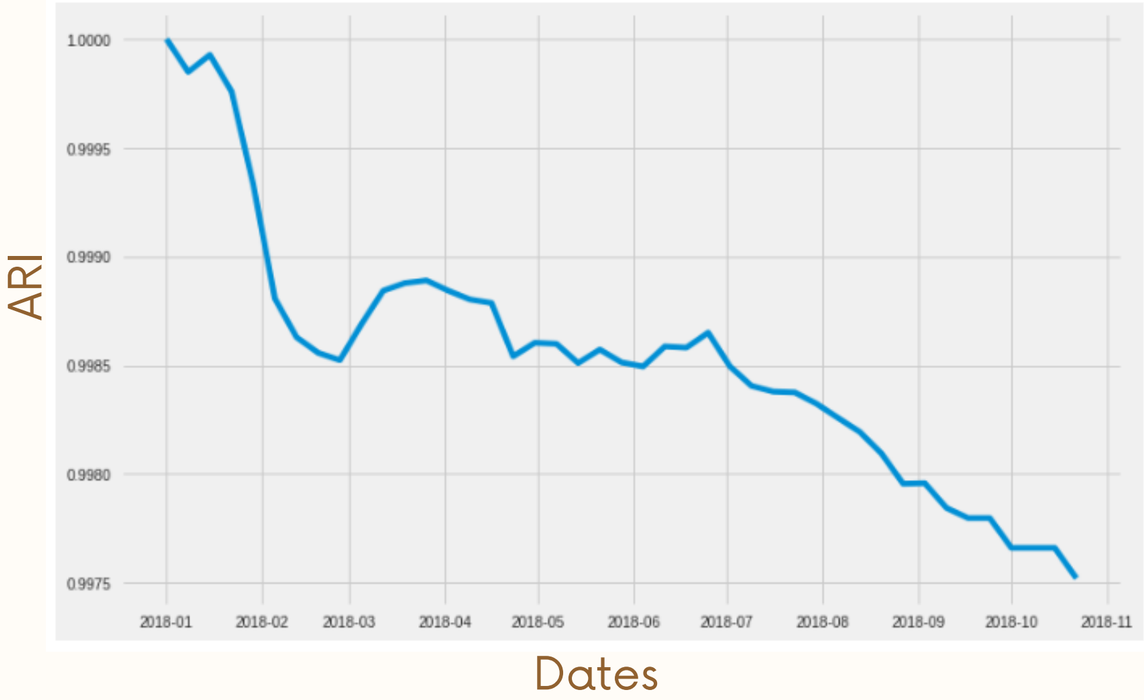
Démarche:

1. Proposer un modèle de clustering M0 entraîné sur un jeu de données de clients C0 à un temps T0
2. Partir du principe qu'a un temps T1, le meilleur modèle de clustering M1 est entraîné sur un jeu de données clients C1
3. Faire une prédiction des clusters du jeu C1 avec le modèle M0
4. Mesurer la performance du modèle M0 par rapport à M1 sur le jeu C1
5. Itérer de l'étape 2 à 4 en augmentant l'intervalle entre T0 et T1

Contrat de maintenance

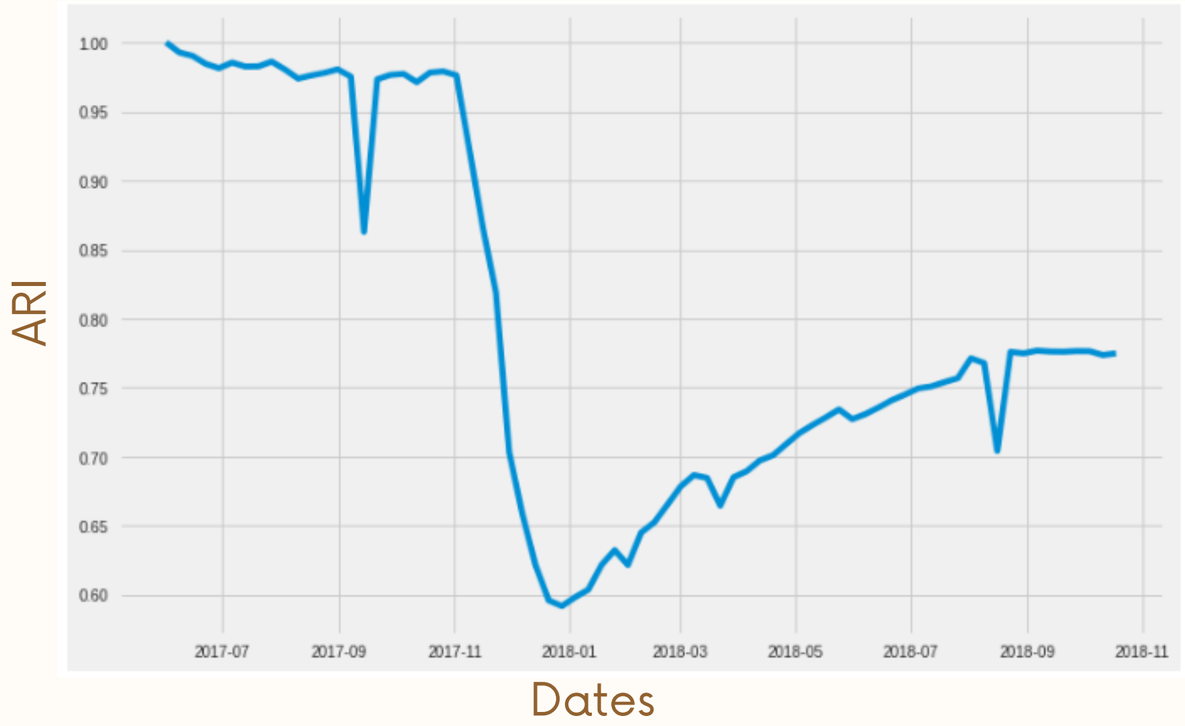
Avec les données min max scalées

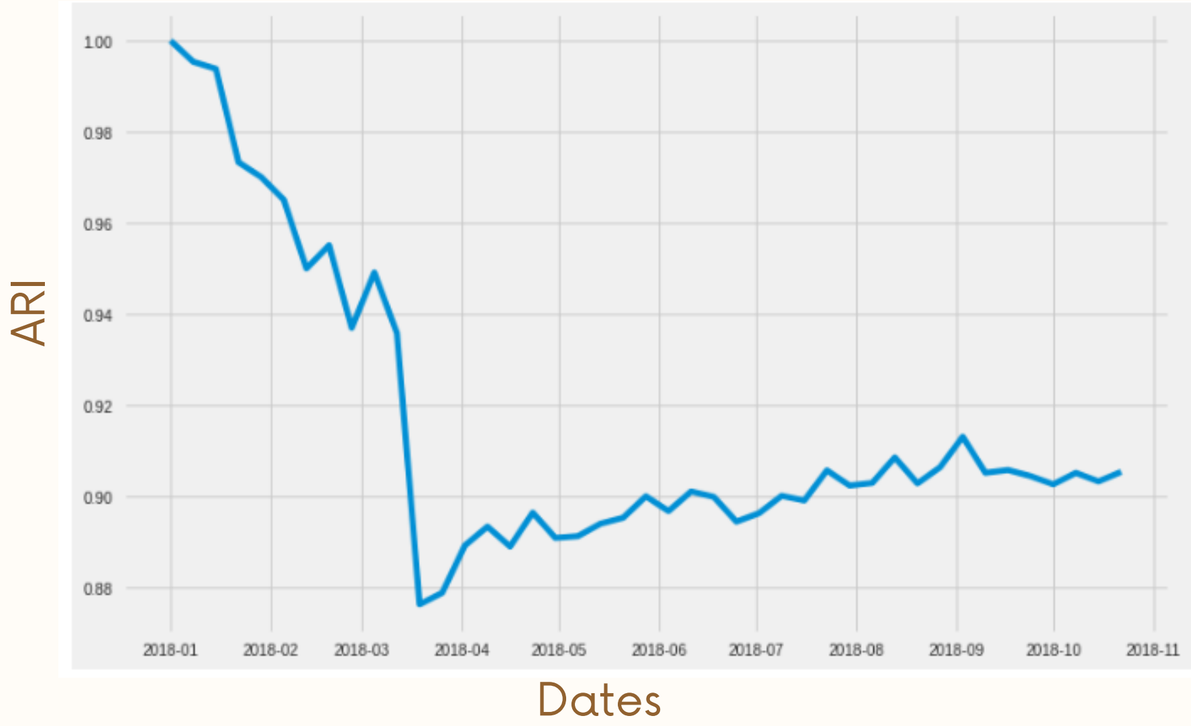




Contrat de maintenance

Avec les données standardisées





Merci de votre attention