

# Projet 5 : Segmentez des clients d'un site e-commerce

The logo for 'olist' is displayed within a solid blue square. The word 'olist' is written in a white, lowercase, sans-serif font, centered horizontally and vertically within the square.

**olist**

# Sommaire

- 1 – Présentation du Projet
- 2 – Préparation et analyse des données
- 3 – Segmentation de la clientèle
- 4 – Maintenance avec ari rand score

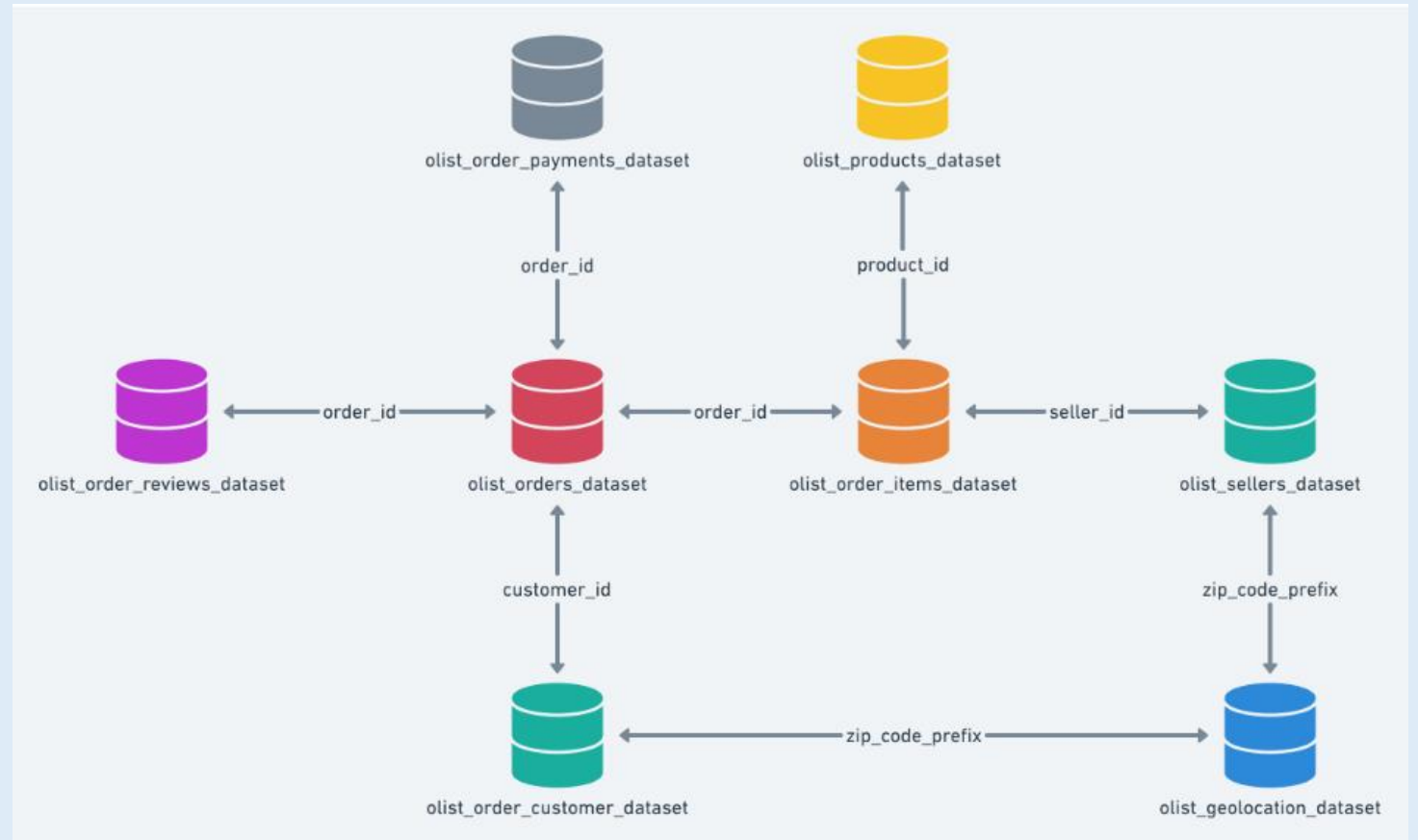
# 1 – Présentation du Projet

- Olist est une entreprise brésilienne qui souhaiterait une segmentation de sa clientèle ainsi qu'un contrat de maintenance de la segmentation.
- Dans les documents, seulement 3 % des clients ayant réalisé plusieurs commandes nous sont fournis.
- Le type de client doit être facilement repérable.

## 2 – Préparation et analyse des données

---

- 8 datasets
- Jointure des datasets effectuée en fonction de l'algèbre relationnelle fournie.



- Le dataframe contient 43 colonnes

Contient des informations sur les clients, les commande, le paiement, la satisfaction et la localisation, 28 020 689 données manquante

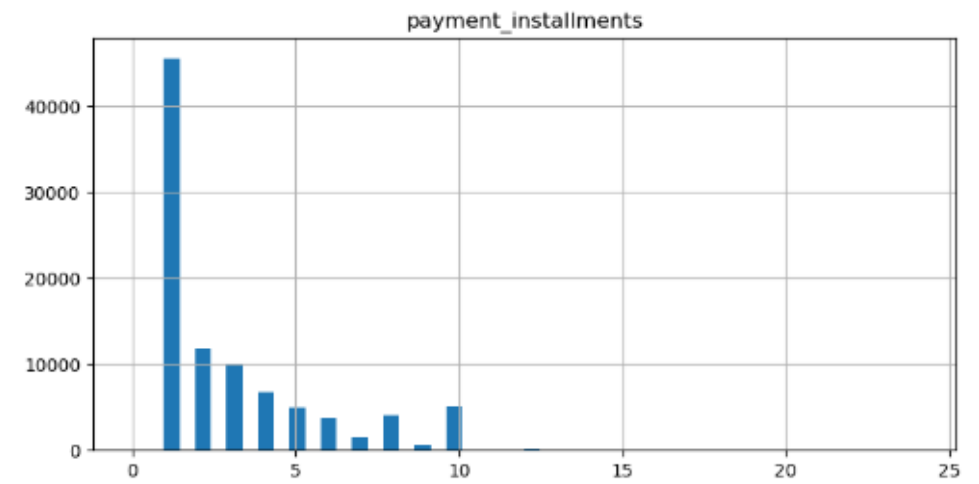
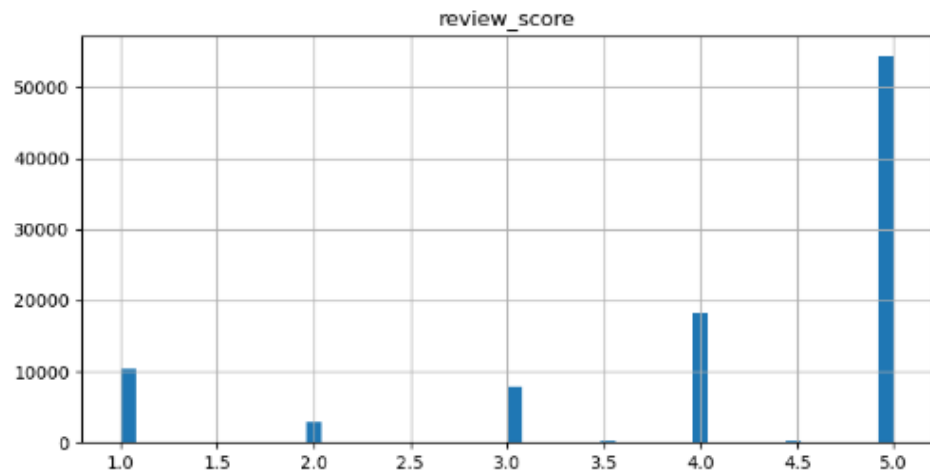
- Réduction du dataframe a 9 colonnes, 0 donnée manquante

```
customer_unique_id    0
customer_state         0
order_id              0
order_purchase_timestamp  0
payment_sequential     0
payment_type           0
payment_installments   0
payment_value          0
review_score           0
```

- Création d'une colonne pour connaître le nombre de produits total
- 102213 produits totaux commandé
- 17 818 736 de doublons

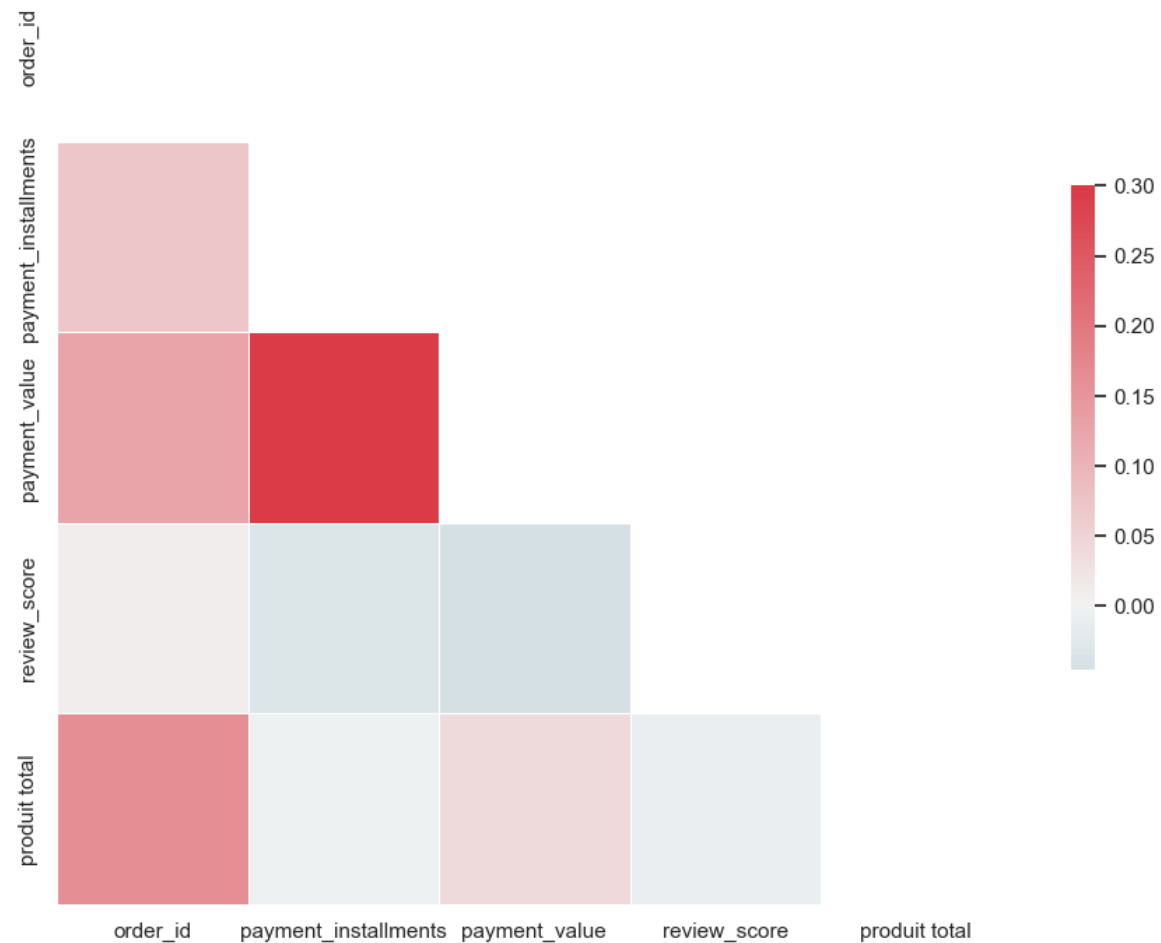
	customer_unique_id	produit total
0	0000366f3b9a7992bf8c76cfd3221e2	1
1	0000b849f77a49e4a4ce2b2a4ca5be3f	1
2	0000f46a3911fa3c0805444483337064	1
3	0000f6ccb0745a6a4b88665a16c9f078	1
4	0004aac84e0df4da2b147fca70cf8255	1
...	...	...
94452	fffcf5a5ff07b0908bd4e2dbc735a684	1
94453	fffea47cd6d3cc0a88bd621562a9d061	1
94454	ffff371b4d645b6ecea244b27531430a	1
94455	ffff5962728ec6157033ef9805bacc48	1
94456	ffffd2657e2aad2907e67c3e9daecbeb	1

	order_id	payment_installments	payment_value	review_score	produit total
count	94457.000000	94457.000000	94457.000000	94457.000000	94457.000000
mean	1.033730	2.939507	166.056596	4.102462	247.621690
std	0.210529	2.721002	227.673566	1.326289	1536.951535
min	1.000000	0.000000	9.590000	1.000000	1.000000
25%	1.000000	1.000000	63.100000	4.000000	58.000000
50%	1.000000	2.000000	107.900000	5.000000	117.000000
75%	1.000000	4.000000	183.120000	5.000000	240.000000
max	16.000000	24.000000	13664.080000	5.000000	314325.000000



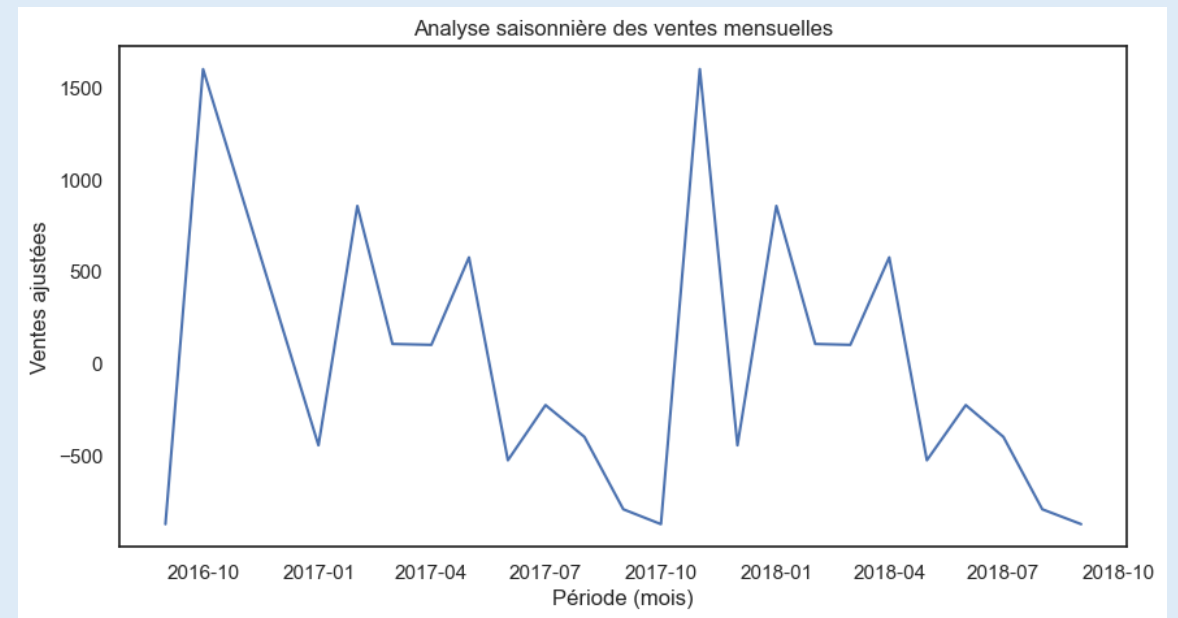
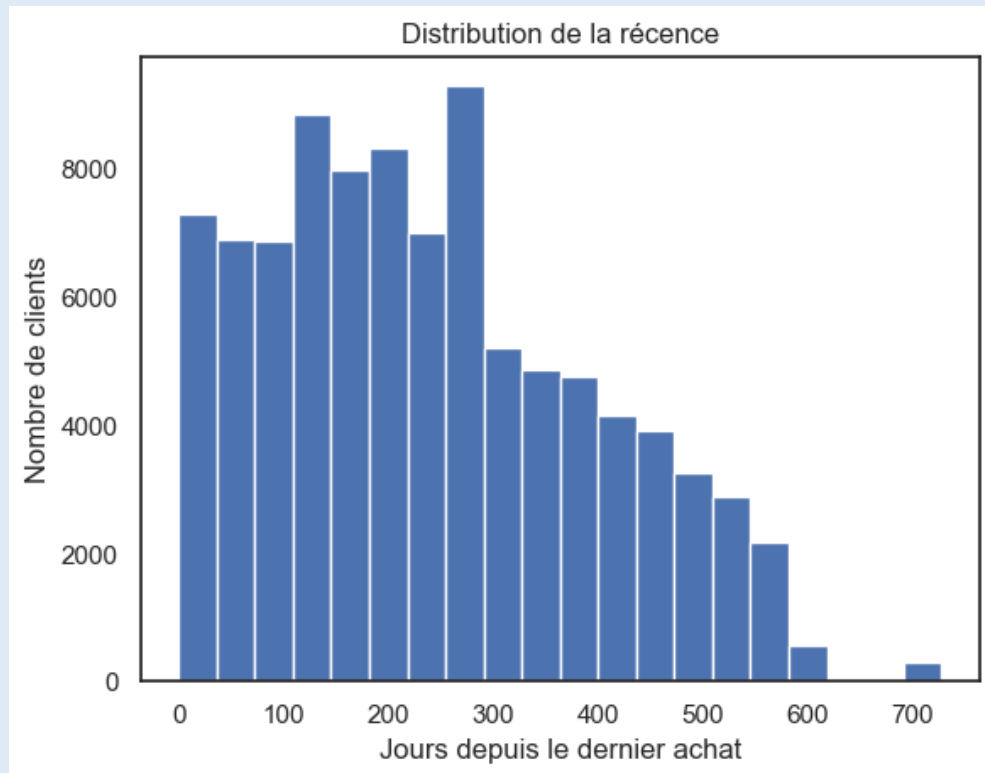


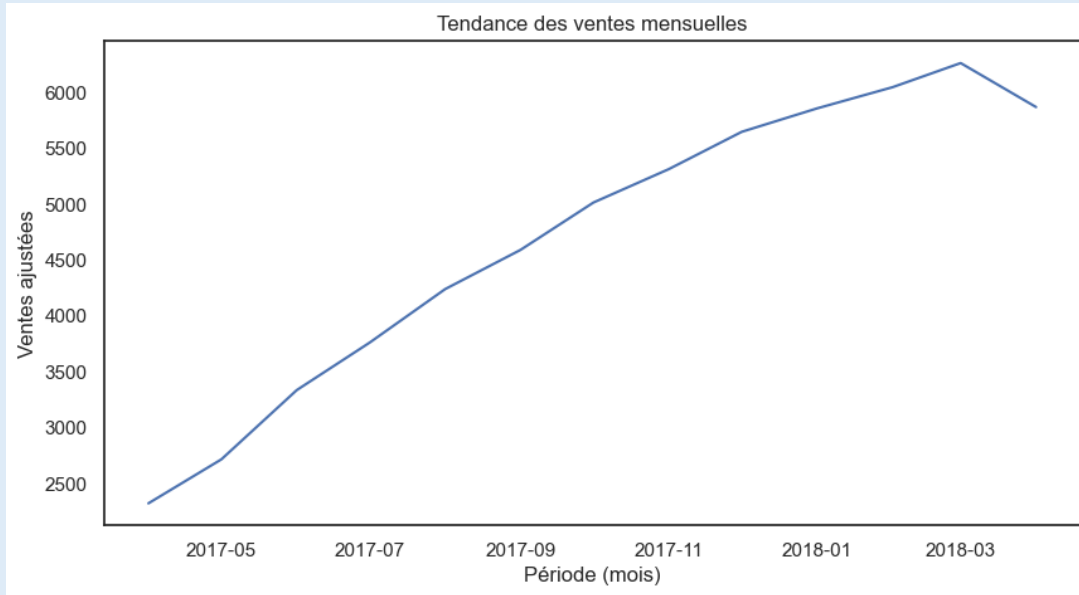
Forte corrélation  
entre paiement  
installment et  
paiement value



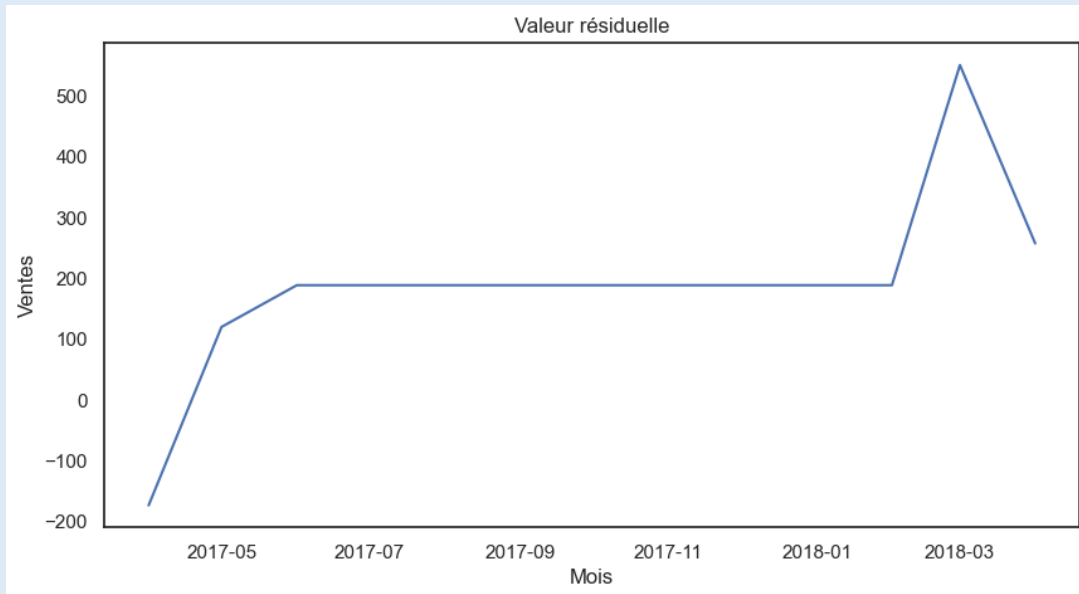
# RFM - Récence

- Seulement 23 mois disponibles, donc j'ai fait la saisonnalité sur 11 mois.
  - pic en octobre 2016 et novembre 2017
  - baisse de vente de juillet à septembre





Tendance croissante jusqu'au 03-2018



- On peut constater une diminution des ventes à partir de mai 2017, ce qui pourrait être dû à une baisse de la demande ou à une forte concurrence.
- En revanche, on observe une hausse significative des ventes à partir de mars 2018, ce qui peut être lié à une augmentation de la demande ou à une stratégie marketing réussie.

# frequence

97% des clients ont fait seulement 1 seule commande.

Une personne a fait 9 commandes et une autre personne a fait 16 commandes, ce sont les meilleurs clients.

1	0.969722
2	0.027812
3	0.001959
4	0.000307
5	0.000095
6	0.000053
7	0.000032
9	0.000011
16	0.000011

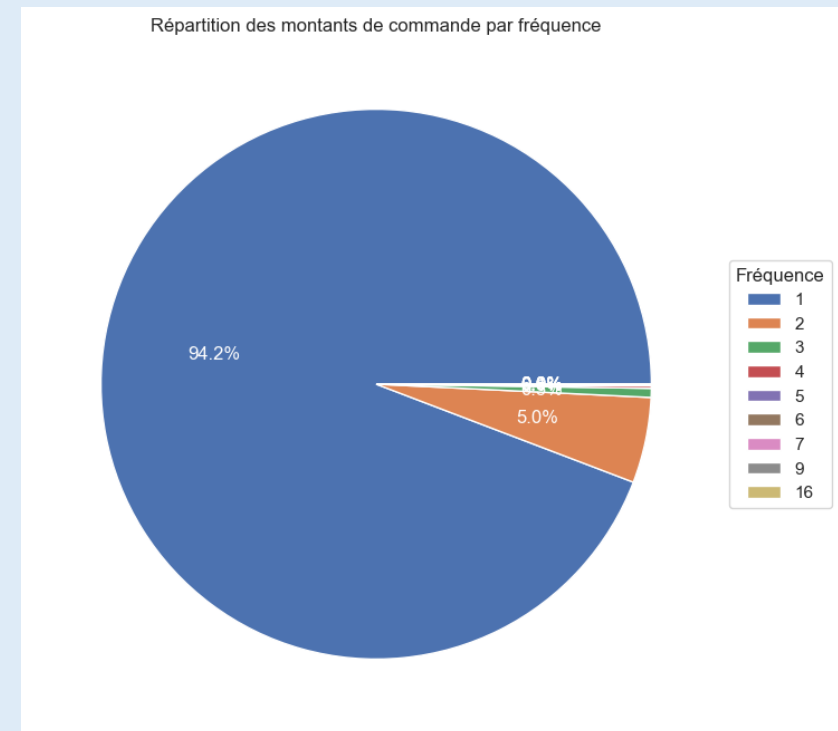
```
data2[data2['fréquence'] > 6]
```

	customer_unique_id	fréquence	o
<b>10185</b>	1b6c7548a2a1f9037c1fd3ddfed95f33	7	
<b>23077</b>	3e43e6105506432c953e165fb2acf44c	9	
<b>37172</b>	6469f99c1f9dfae7733b25662e7f1782	7	
<b>52090</b>	8d50f5eadf50201ccdcedfb9e2ac8455	16	
<b>74797</b>	ca77025e7201e3b30c44b472ff346268	7	

# Montant commande

Le montant total des commandes est proportionnel à la fréquence.

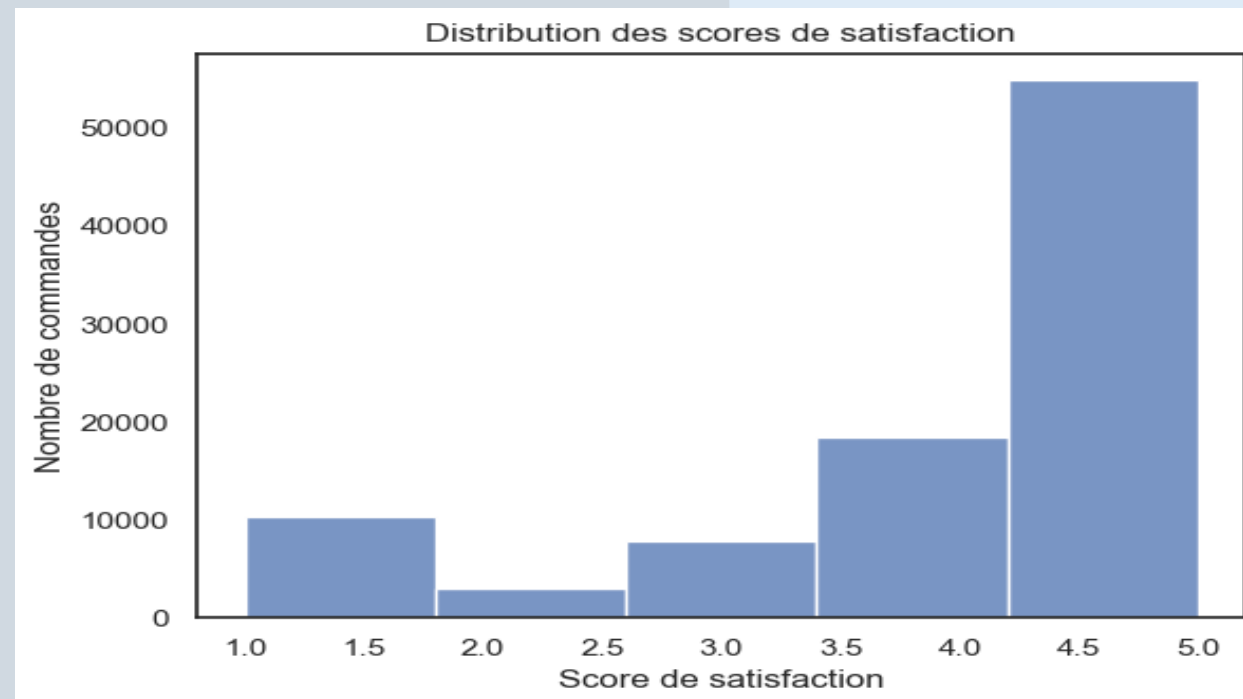
order_id	
1	14774909.96
2	789398.76
3	83051.28
4	22696.91
5	6719.17
6	3516.57
7	2840.56
9	1172.66
16	902.04



# Satisfaction des clients

- Les clients sont globalement satisfaits de leurs achats

count	94457
mean	4
std	1
min	1
25%	4
50%	5
75%	5
max	5



### 3- Segmentation de la clientèle : RFM et Satisfaction

- Les différentes méthodes de segmentations : Kmeans, Dbscan et clustering hiérarchique
- Les variables utilisées
- La démarche effectuée

# Sélections des variables et démarches

Variables pour la segmentation :

- Récence
- Fréquence
- Montant
- Satisfaction du client

Réduction de dimension à l'aide de l'ACP

Choix de l'algorithme :

- Kmeans
- Dbscan
- Clustering hiérarchique

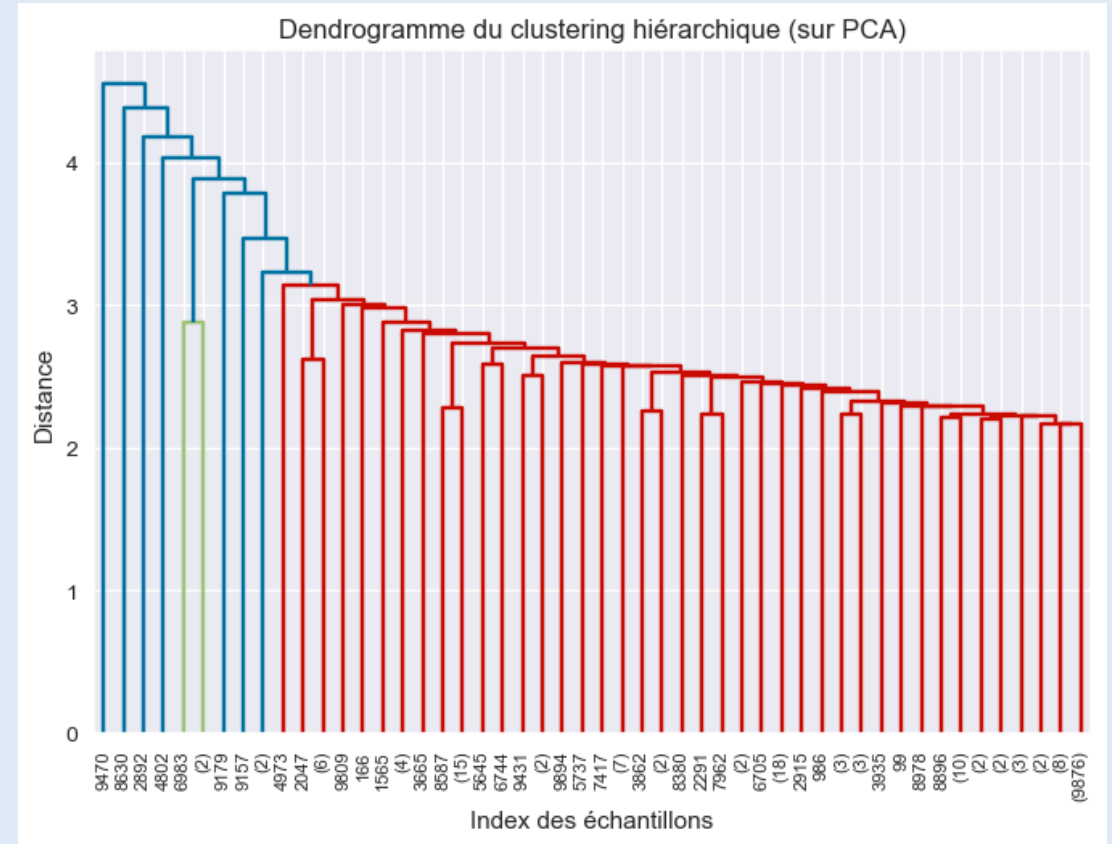
Visualisation et choix des Clusters





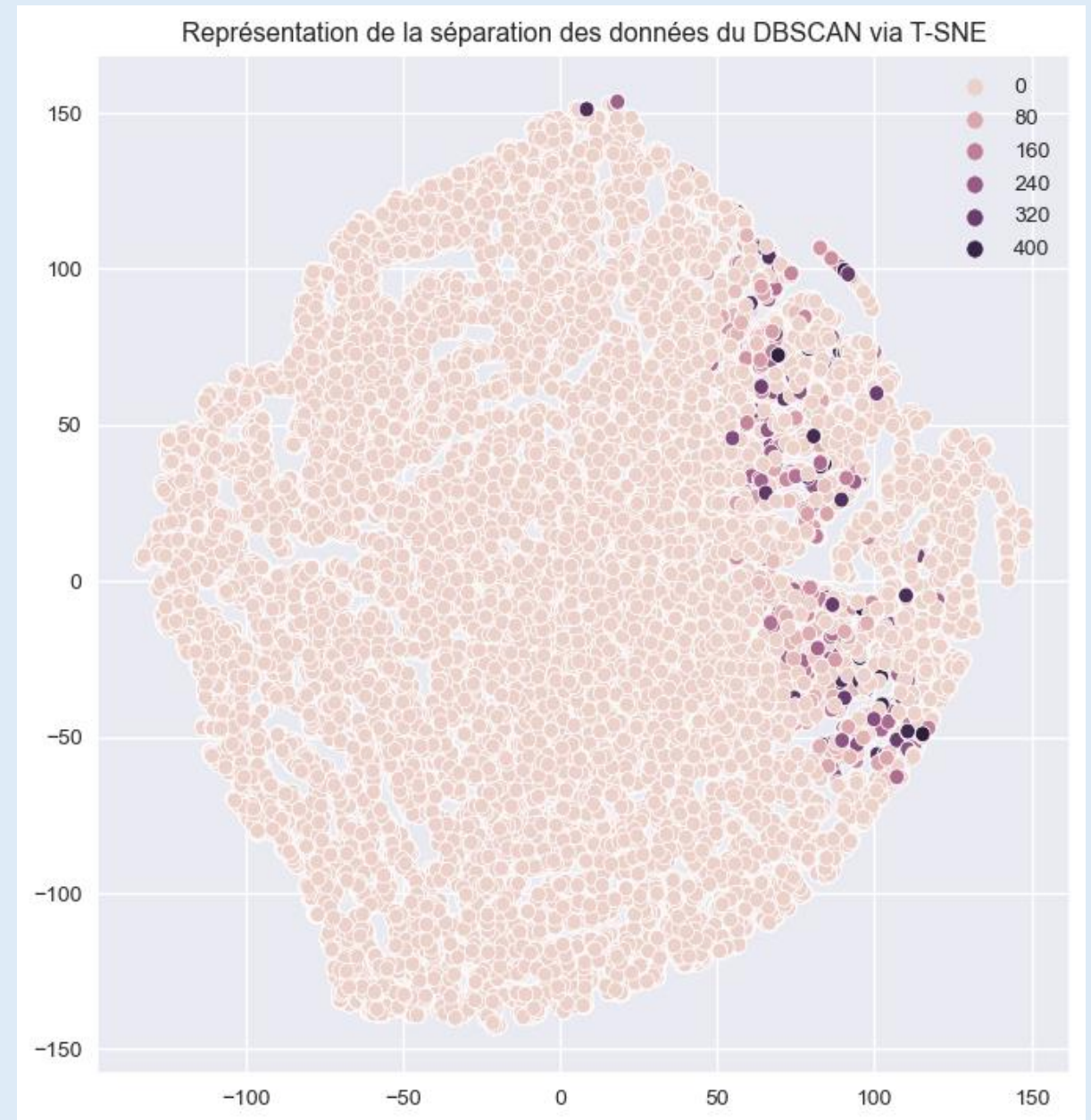
# Clustering hiérarchique

- Modèle qui nécessite un échantillon (ici j'ai pris 8000 données), ce qui fausse la segmentation. Clusters pas équilibrés.

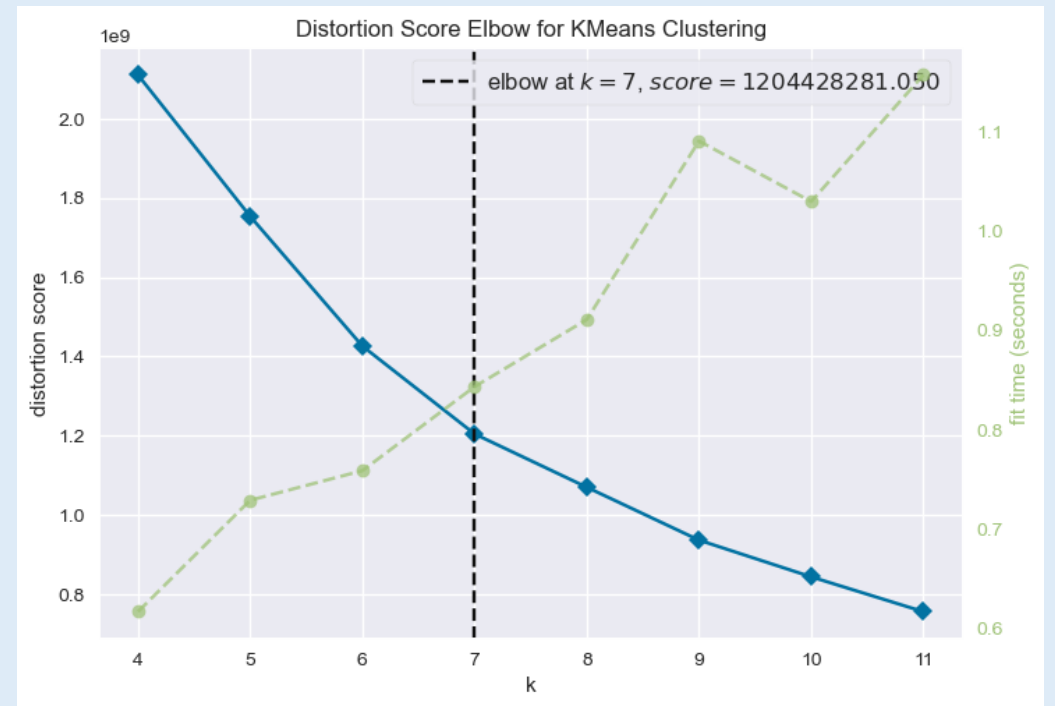
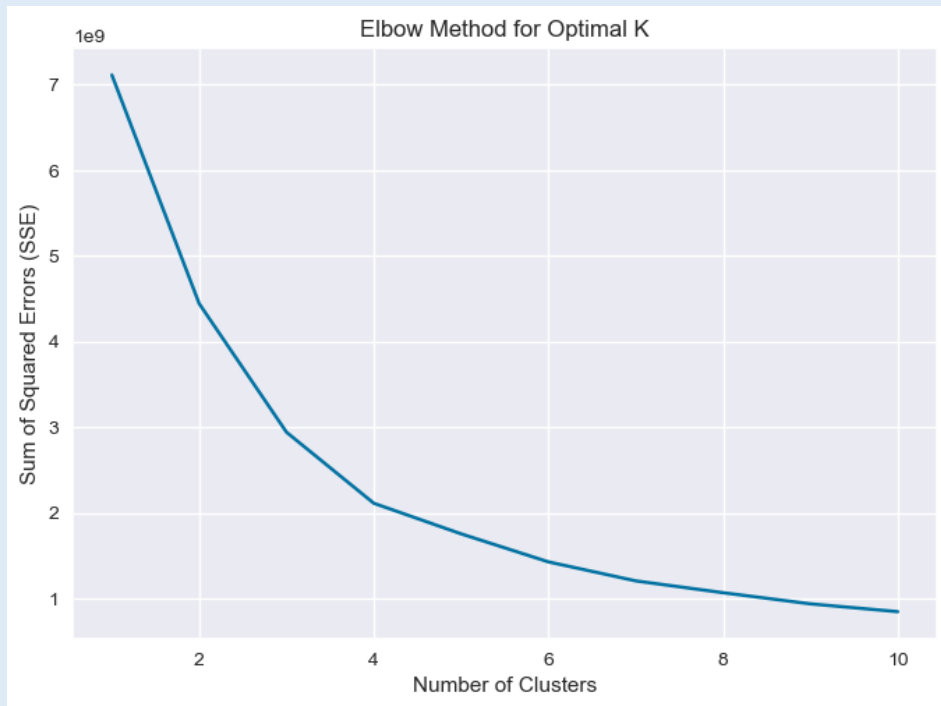


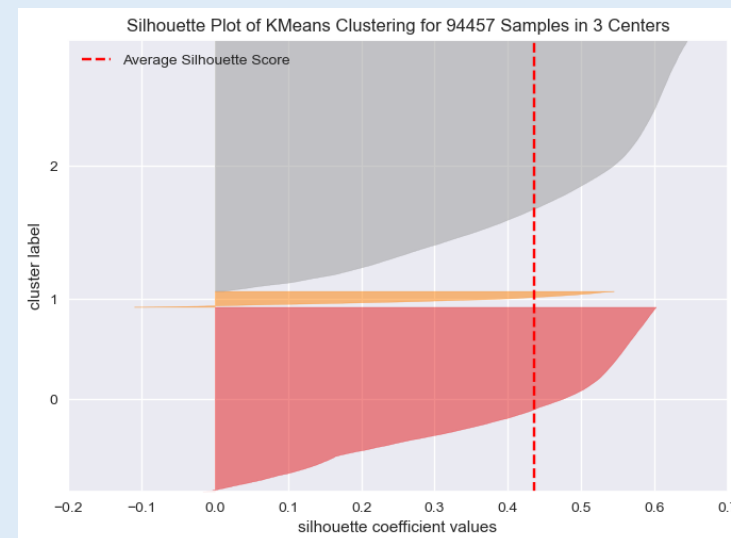
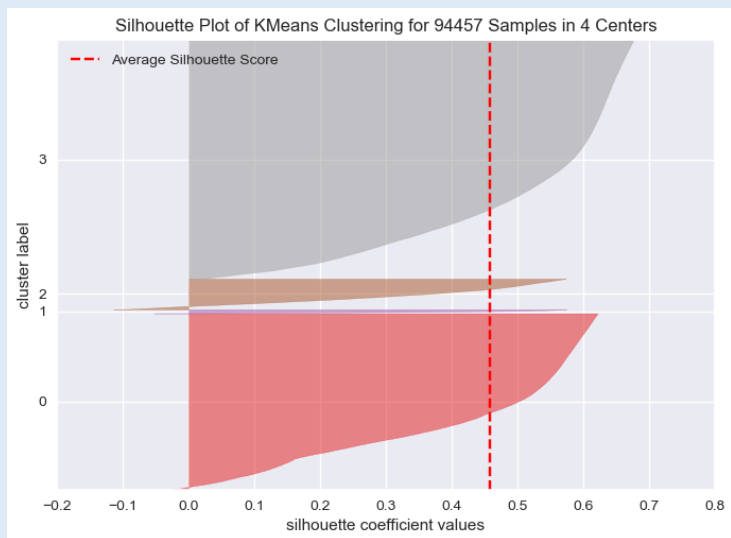
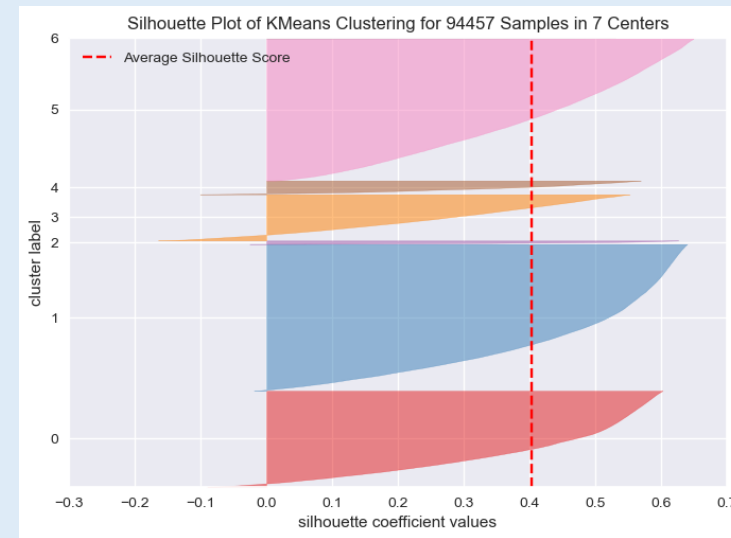
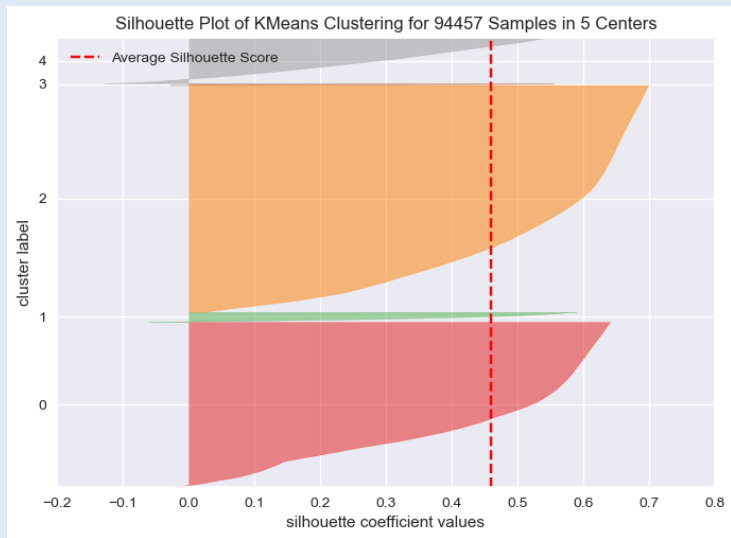
# DB SCAN

- Cluster difficilement identifiables
- Clusters peu équilibrés

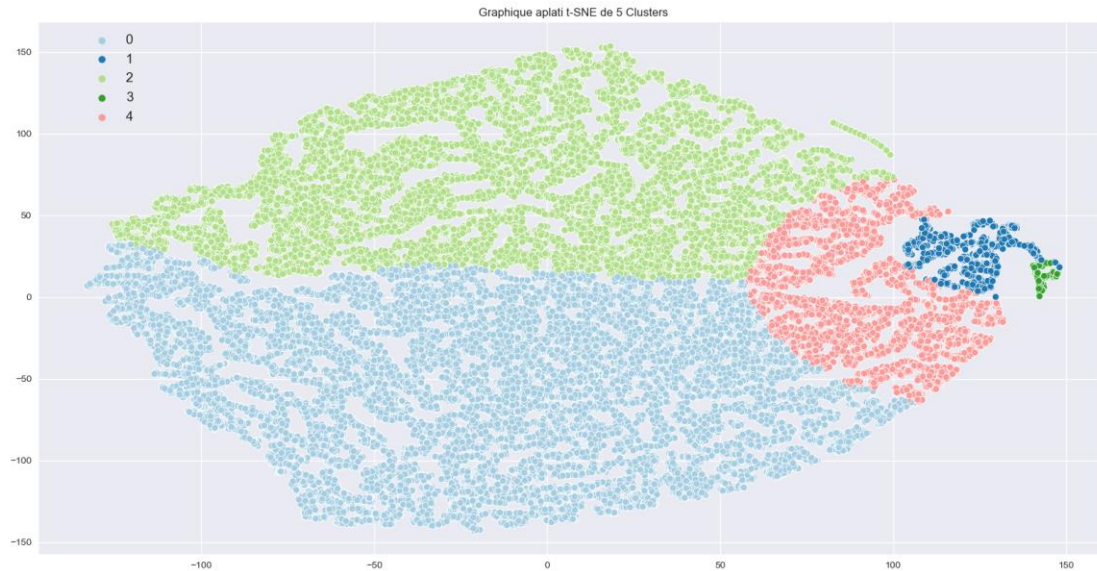


# Kmeans





# Visualisation des Clusters



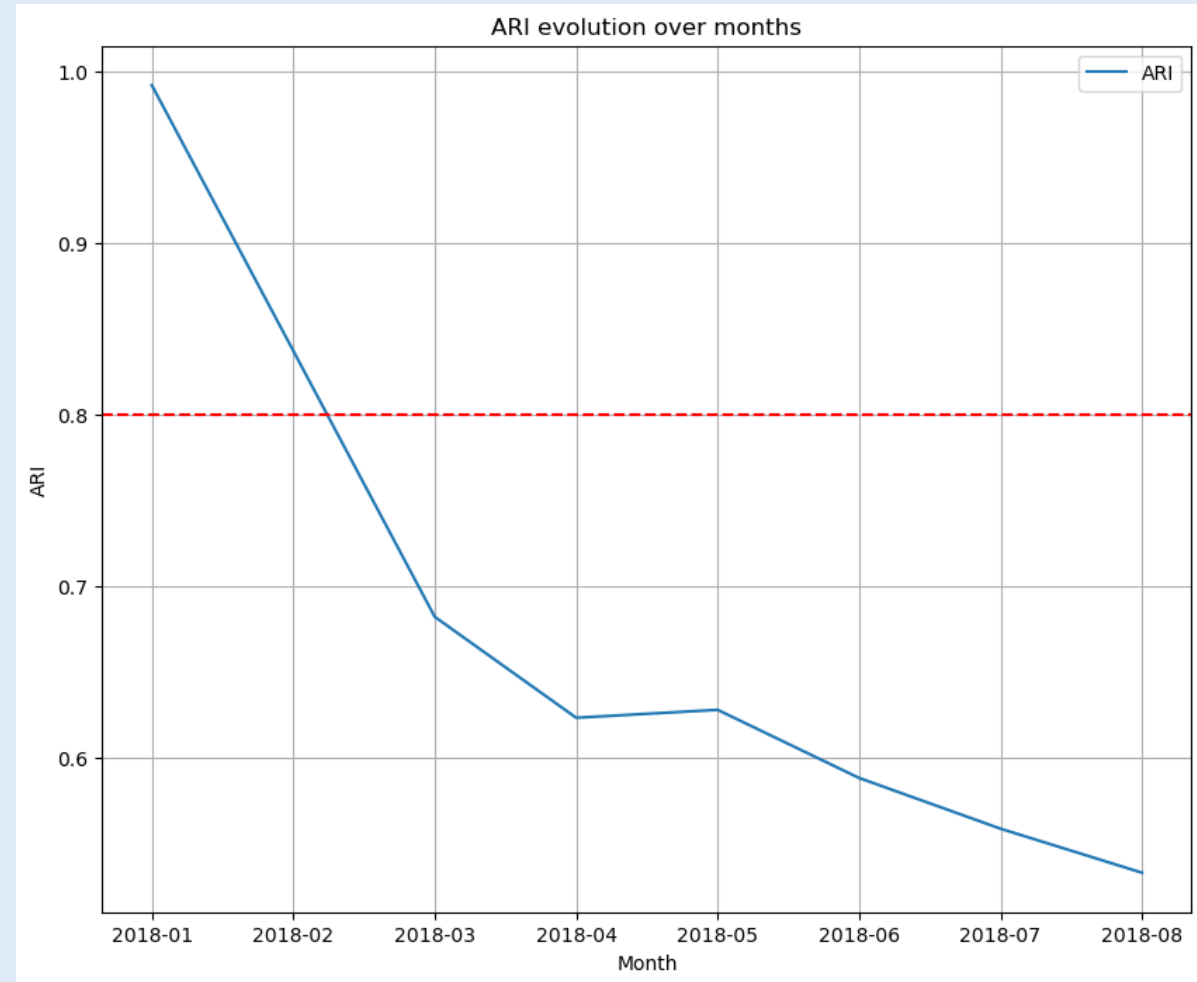
	recence	fréquence	payment_value		review_score
	mean	mean	mean	count	mean
cluster					
0	398.0	1.0	111.0	34702	4.0
1	133.0	1.0	108.0	47811	4.0
2	248.0	1.0	999.0	2089	4.0
3	224.0	1.0	400.0	9516	4.0
4	233.0	1.0	2402.0	339	4.0

- Cluster 0 : Clients anciens ayant passé énormément de commandes à faible coût.
- Cluster 1 : Clients très récents ayant passé énormément de commandes à faible coût.
- Cluster 2 : Clients récents ayant passé peu de commandes à prix élevé.
- Cluster 3 : Clients récents ayant passé beaucoup de commandes à prix moyen.
- Cluster 4 : Clients récents ayant passé très peu de commandes à prix très élevé.

# 4 – Maintenance avec ari rand score

Segmentations à effectuer au cours du deuxième mois (environ 40 jours)

```
{ '2018-01': 0.9923231295461841,  
  '2018-02': 0.8373117673283236,  
  '2018-03': 0.6820215410040537,  
  '2018-04': 0.6230848791908785,  
  '2018-05': 0.6277021130554848,  
  '2018-06': 0.587937217057098,  
  '2018-07': 0.5583239595357334,  
  '2018-08': 0.5326392544896735}
```



# CONCLUSION

## Client

- Création de 5 clusters à l'aide de K-means.
- Actualisation de la segmentation à faire tous les 40 jours.

## Marketing

- Bon de réduction pour les clients fidèles en fonction des produits qu'ils ont déjà achetés.
- Programme de fidélité pour les nouveaux clients, visant à les inciter à commander davantage.
- Livraison gratuite à partir d'un certain montant ou d'un certain nombre de produits, dans le but d'augmenter le prix ou le nombre de produits moyen dans le panier.