

SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE

KONAN KOFFI

PLAN DE PRESENTATION

I Présentation de la problématique, du cleaning effectué, du feature engineering et de l'exploration

II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

III Présentation de la simulation pour définir le délai de maintenance du modèle (contrat de maintenance)

IV Conclusion

I Présentation de la problématique, du cleaning effectué, du feature engineering et de l'exploration



olist

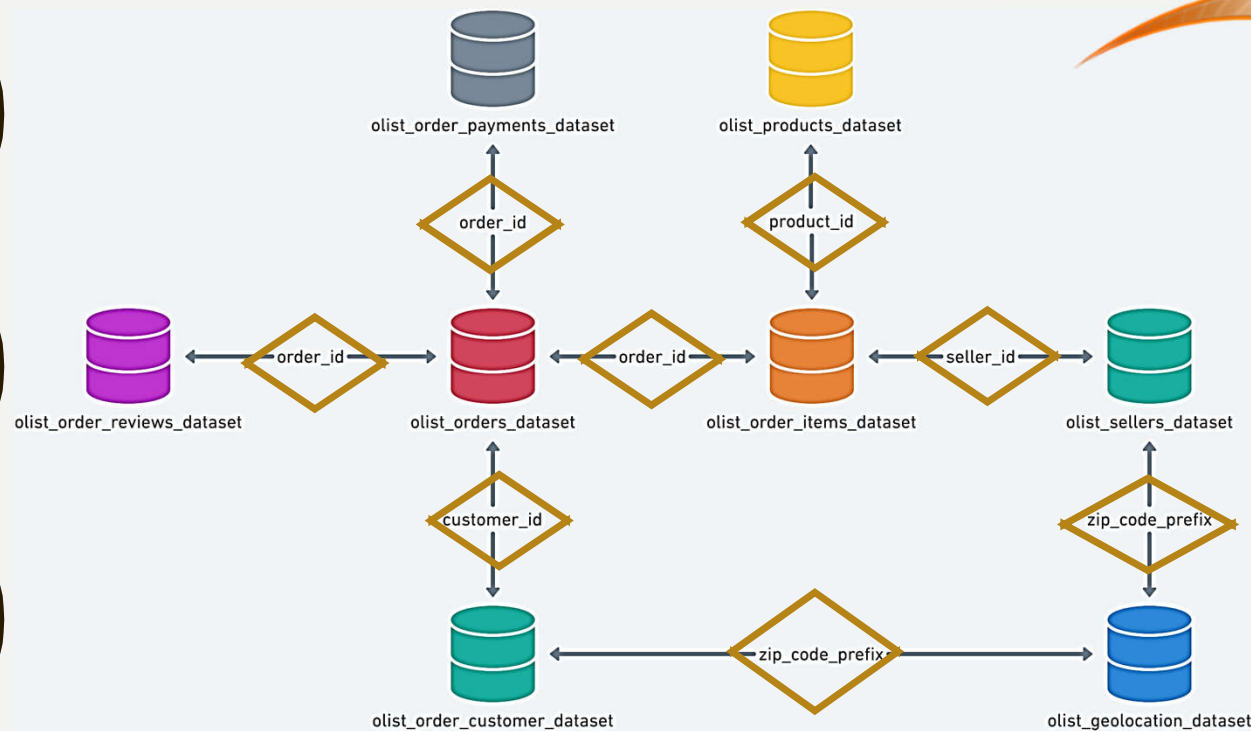
- ❖ Startup brésilienne qui opère dans le segment du e-commerce
- ❖ Propose une solution de vente sur les marketplaces en ligne.

Notre Mission

- Proposition de segmentation facilement exploitable par l'équipe Marketing d'Olist afin de les aider à mieux comprendre **les différents types d'utilisateurs**.
- Ressortir à minima les bons et moins bons clients en termes de **commandes et de satisfaction**.
- Code fourni selon la convention PEP8,

I Présentation de la problématique, du cleaning effectué, du feature engineering et de l'exploration

Données à notre disposition *



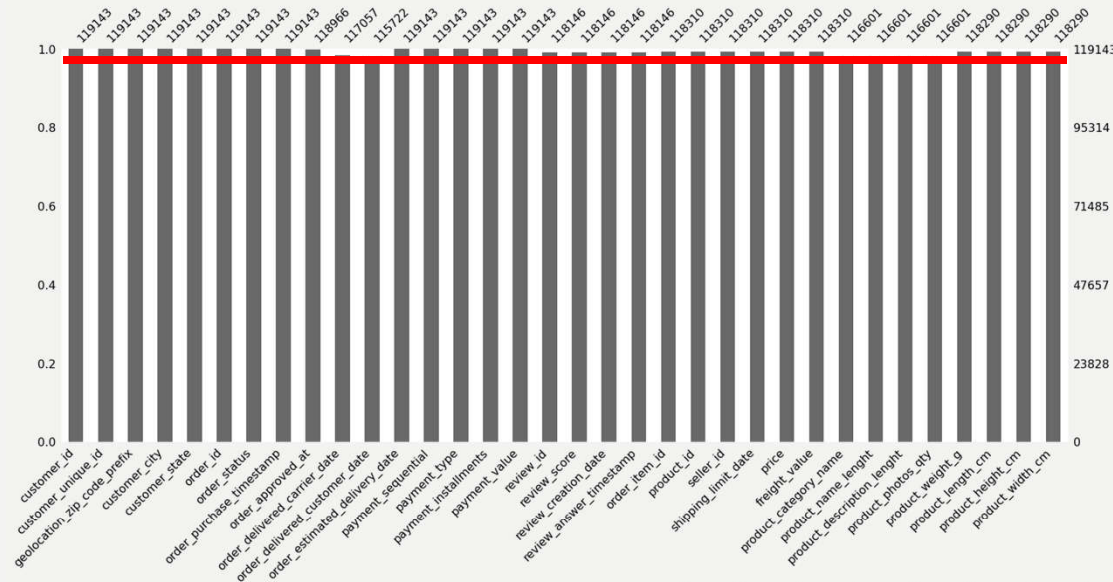
- Identifiants unique de Clients
- historique de commandes
- Articles achetés
- Vendeurs
- Satisfaction des clients
- Données de géolocalisation
- Date d'achat
- etc...



**3 % de clients:
Plus d'une commande**

(*): Source : https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?select=olist_orders_dataset.csv

I Présentation de la problématique, du cleaning effectué, du feature engineering et de l'exploration

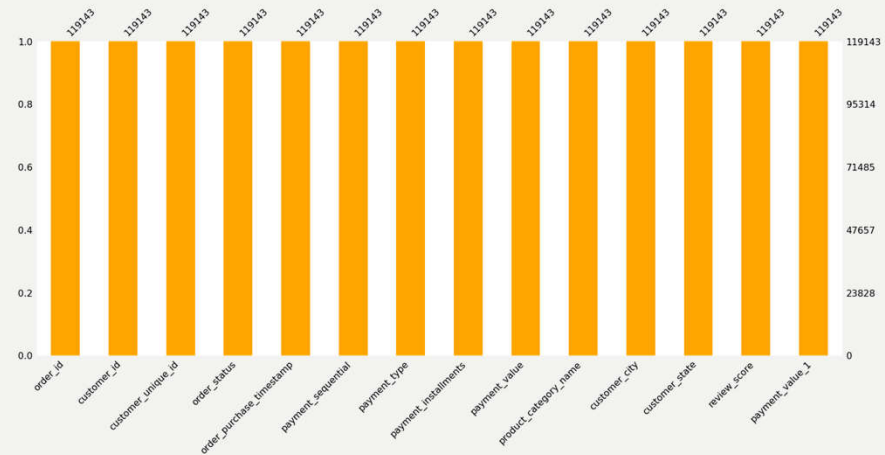


- Pas de valeurs négatives
- Très peu de données manquantes

- Beaucoup de duplicatas

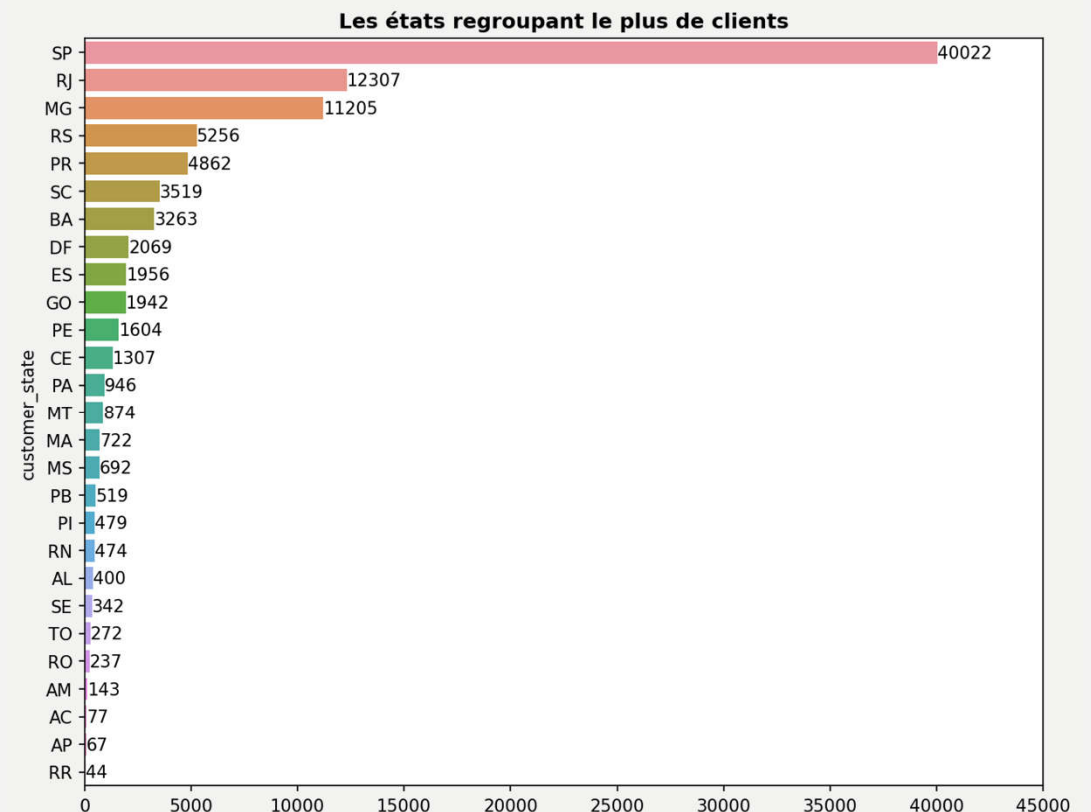
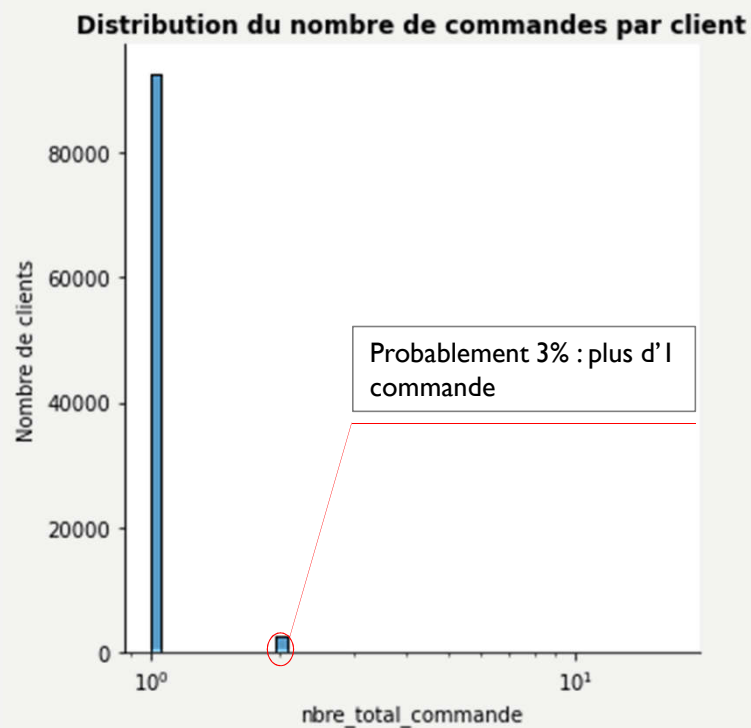


- Suppression de duplicatas dans la lignes de commandes

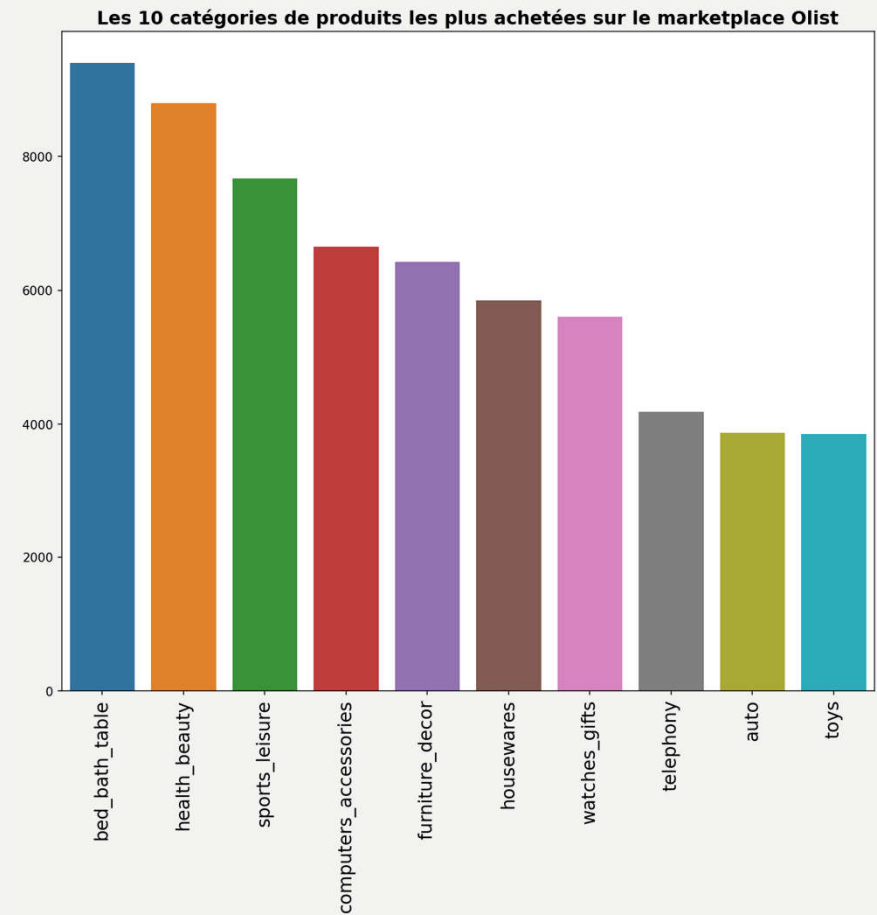
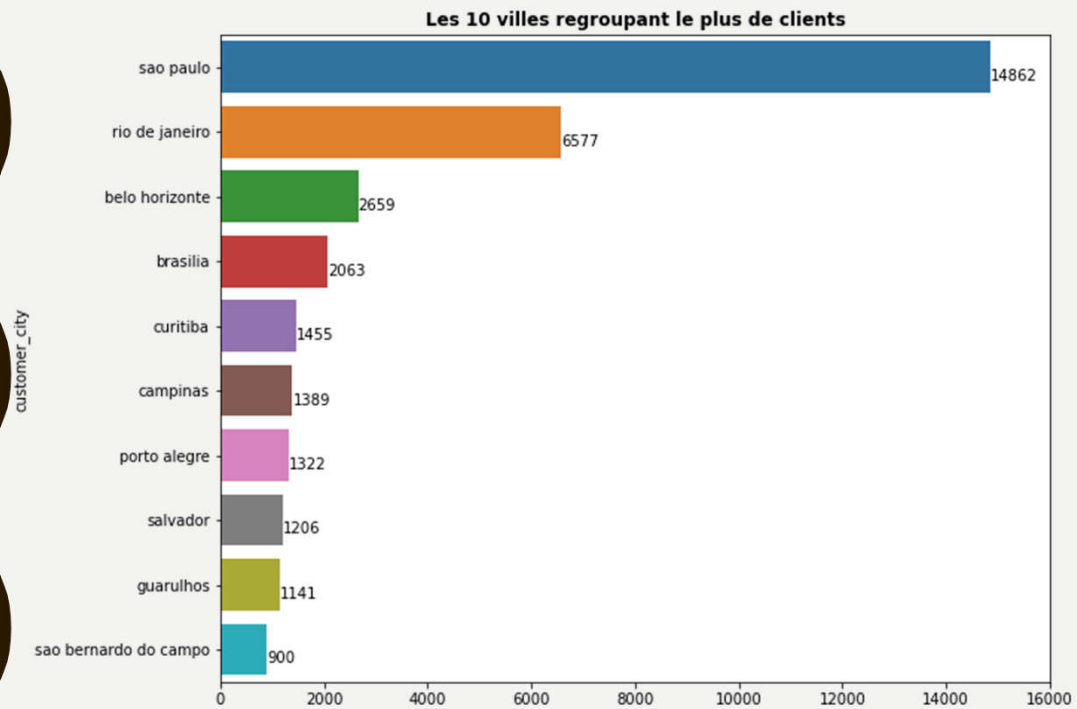


I Présentation de la problématique, du cleaning effectué, du feature engineering et de l'exploration

Exploration des données



I Présentation de la problématique, du cleaning effectué, du feature engineering et de l'exploration



I Présentation de la problématique, du cleaning effectué, du feature engineering et de l'exploration

Données RFM – Aspect Marketing

Feature engineering

	customer_unique_id	Recency_days	Frequency	Monetary	review_score
0	861eff4711a542e4b93843c6dd7febb0	474	1	146.87	4.0
1	9eae34bbd3a474ec5d07949ca7de67c0	298	1	275.79	1.0
2	3c799d181c34d51f6d44bbbc563024db	483	1	140.61	3.0
3	23397e992b09769faf5e66f9e171a241	211	1	137.58	4.0
4	567ab47ca4deb92d46dbf54dce07d0a7	528	1	142.05	4.0
...
95555	93d9e516a351a7747fc9830ae9525062	392	1	66.69	1.0
95556	f979a07fc18b2af3780a796ba14b96f4	329	1	54.09	5.0
95557	1b553902a5bbe6ee54a3aaa7cbfb6816	473	1	124.52	5.0
95558	d8bee9ec375c3a0f9ef8ed7456a51dcd	584	1	209.06	4.0
95559	141e824b8e0df709e3fcf6d982225a8e	350	1	115.45	1.0

- **Recency** : dernière commande du client
- **Frequency** : nombre de commandes
- **Monetary** : panier du client
- **Review_score** : satisfaction moyenne



Usage de fonctions d'agrégation

.mean() → review_score
.sum() → Monetary
.count() → Frequency

Différence de date

Type datetime → Recency_days

95560 clients au final / **96069**
Soit près de 509 clients fictifs

II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

Outils de segmentation de données / Apprentissage non supervisé

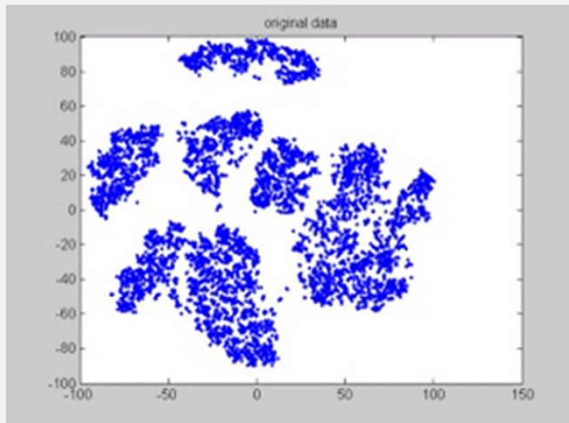
Fonction d'attribution $C : [1, N] \rightarrow [1, K]$ qui minimise une fonction coût.

→ Critères de proximité assurés par des mesures et classes de distance entre objets.

Exemple d'algorithme d'apprentissage non supervisé *

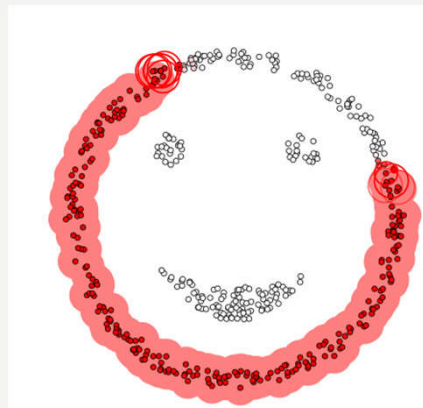
KMeans (K-moyennes)

Méthodes centroïdes



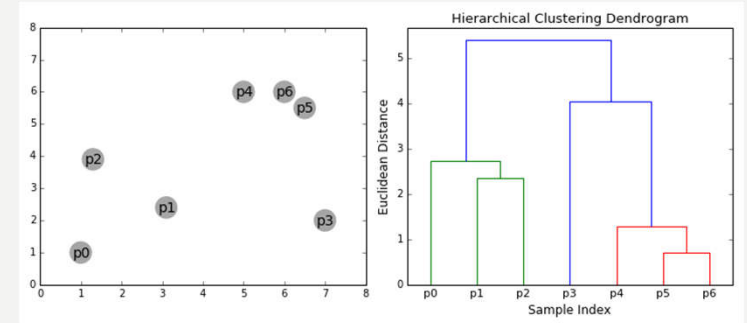
DBSCAN

Méthodes à densité



Clustering Agglomérative Hiérarchique

Méthodes hiérarchiques



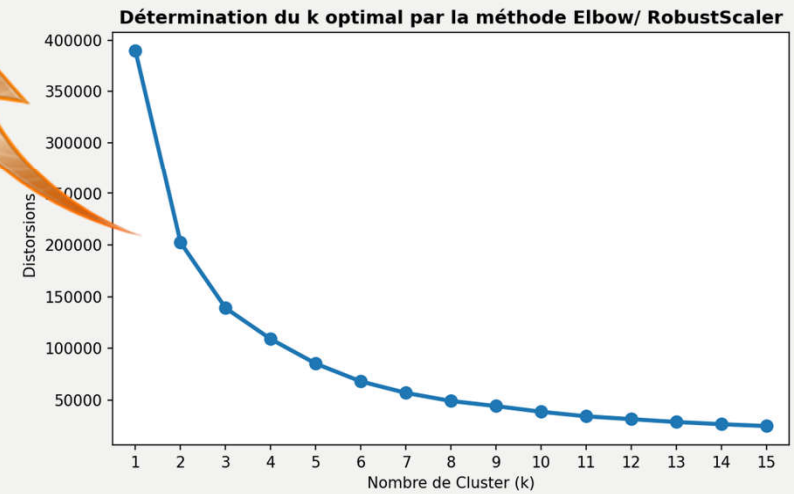
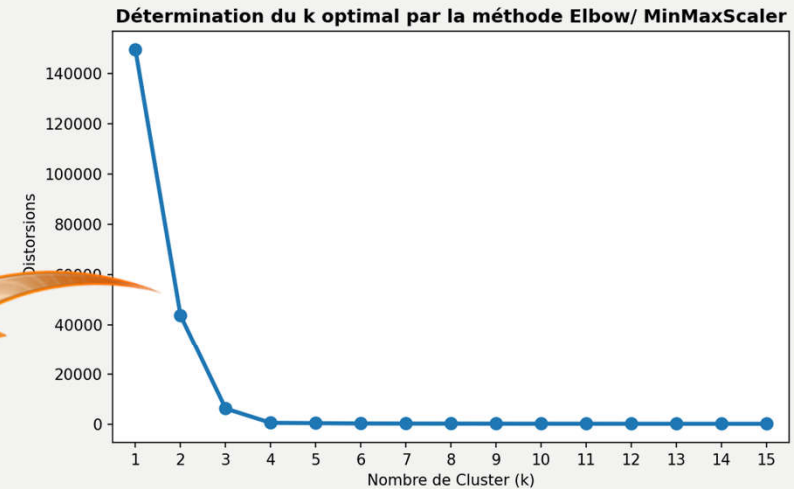
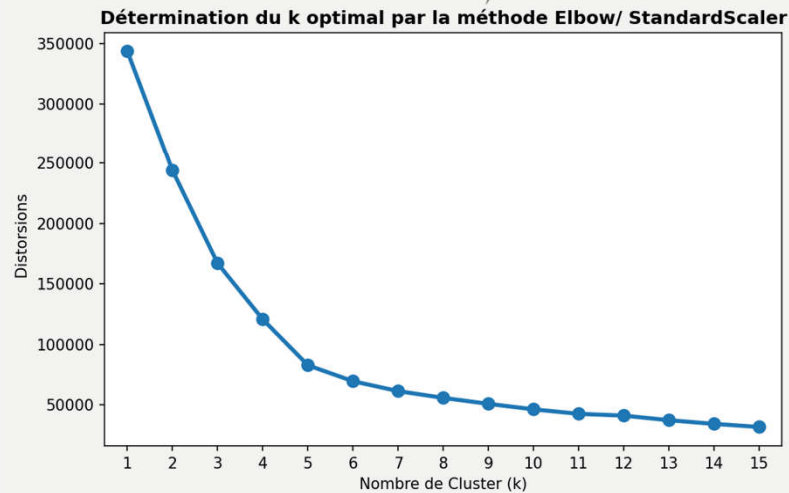
(*): Source: <https://larevueia.fr/clustering-les-3-methodes-a-connaître/>

II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

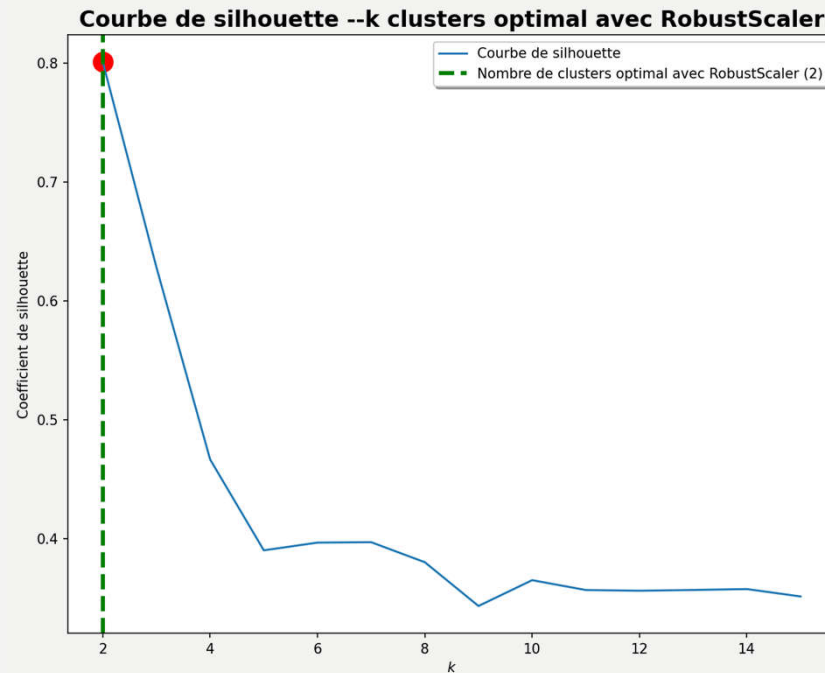
Méthode d'Elbow

Tracer la variance expliquée en fonction du nombre de clusters et à choisir le coude de la courbe comme nombre de clusters à utiliser.

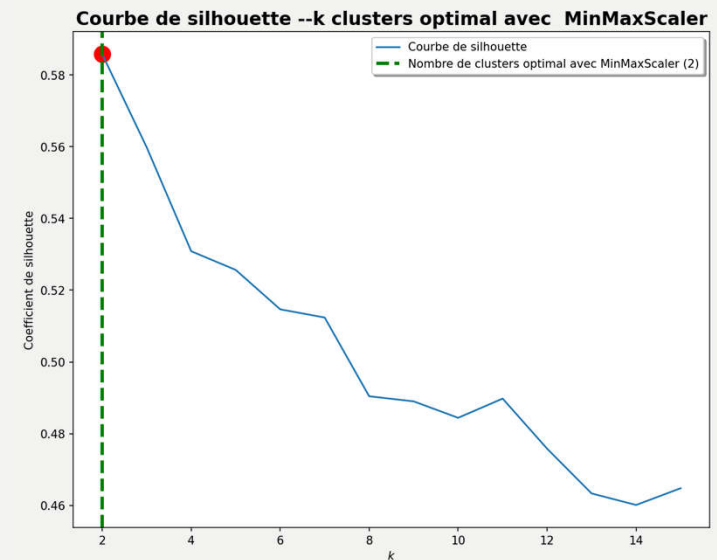
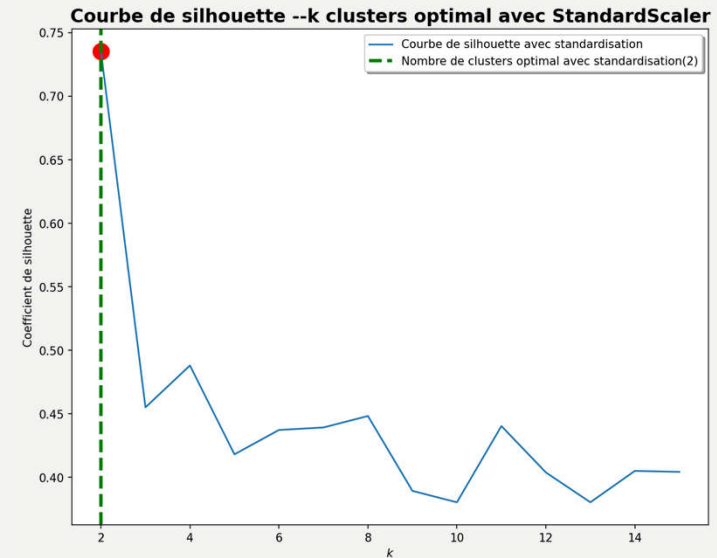
$k = 2$



II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

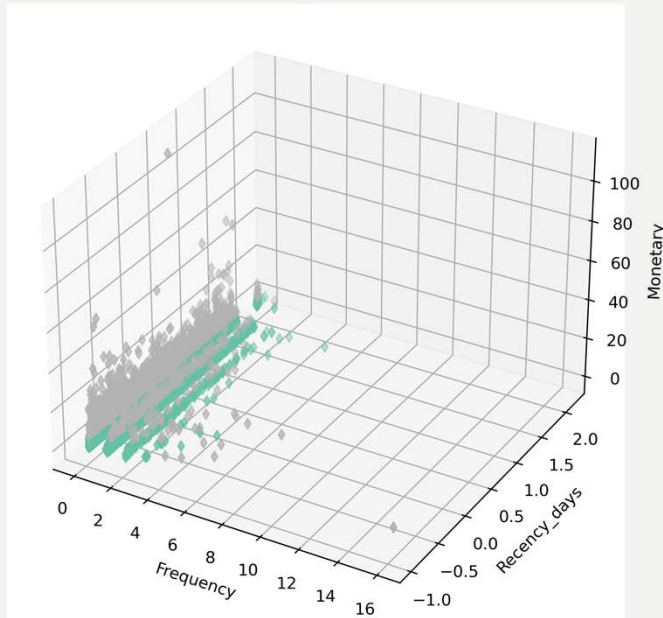


RobustScaler conduit à un k optimal avec pour coefficient de silhouette plus élevé.

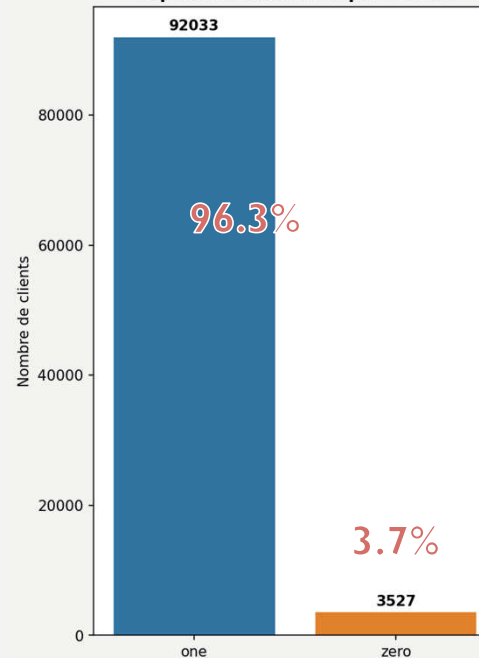


II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

Visualisation en 3D de la segmentation avec RobustScaler



Population de clients par cluster



Les tendances pour le cluster 0

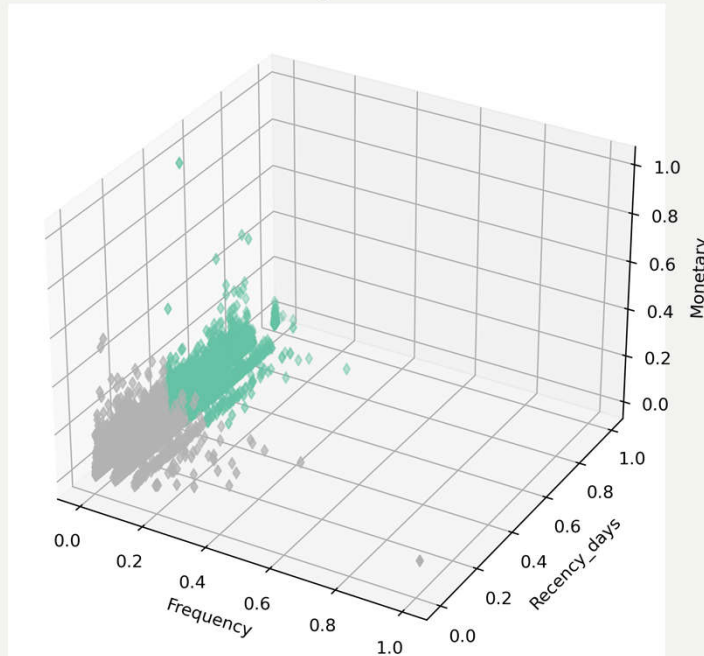
	count	mean	min	25%	50%	75%	max
Recency_days	92033.0	243.739159	0.00	120.00	225.00	353.00	729.00
Frequency	92033.0	1.030652	1.00	1.00	1.00	1.00	6.00
Monetary	92033.0	133.136323	9.59	61.80	104.28	171.03	578.78

Les tendances pour le cluster 1

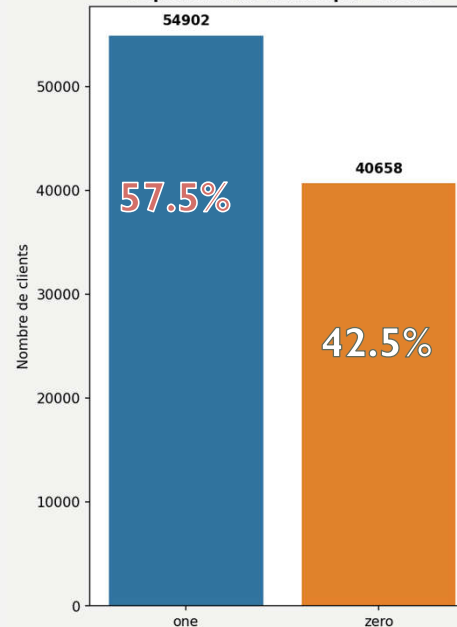
	count	mean	min	25%	50%	75%	max
Recency_days	3527.0	249.064361	6.00	119.000	231.00	361.000	699.00
Frequency	3527.0	1.123334	1.00	1.000	1.00	1.000	17.00
Monetary	3527.0	1024.333635	575.94	671.425	823.16	1154.325	13664.08

II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

Visualisation en 3D de la segmentation avec MinMaxScaler



Population de clients par cluster



Les tendances pour le cluster 0

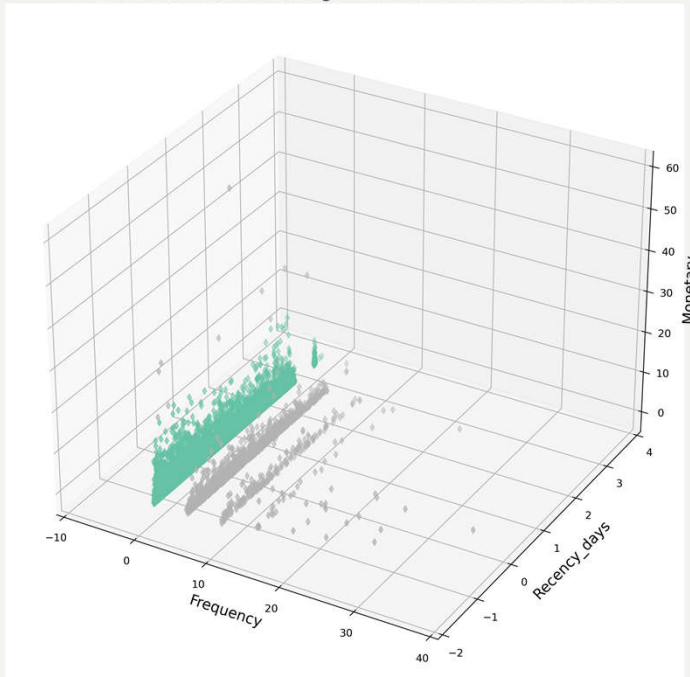
	count	mean	min	25%	50%	75%	max
Recency_days	40658.0	393.127626	264.00	305.00	380.00	467.000	729.00
Frequency	40658.0	1.029539	1.00	1.00	1.00	1.000	6.00
Monetary	40658.0	165.988695	10.07	62.99	106.17	181.635	13664.08

Les tendances pour le cluster 1

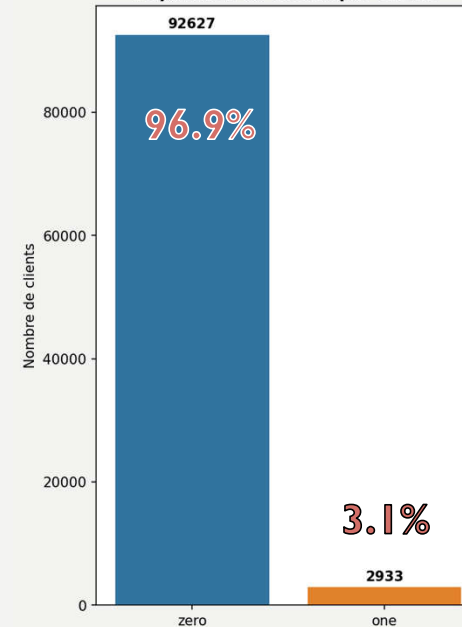
	count	mean	min	25%	50%	75%	max
Recency_days	54902.0	133.450749	0.00	70.00	135.0	195.00	263.00
Frequency	54902.0	1.037430	1.00	1.00	1.0	1.00	17.00
Monetary	54902.0	166.059372	9.59	63.27	109.5	184.03	7274.88

II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

Visualisation en 3D de la segmentation avec StandardScaler



Population de clients par cluster



Les tendances pour le cluster 0

	count	mean	min	25%	50%	75%	max
Recency_days	92627.0	244.480648	5.00	120.00	226.00	354.00	729.0
Frequency	92627.0	1.000000	1.00	1.00	1.00	1.00	1.0
Monetary	92627.0	160.912271	9.59	62.15	105.66	177.25	4445.5

Les tendances pour le cluster 1

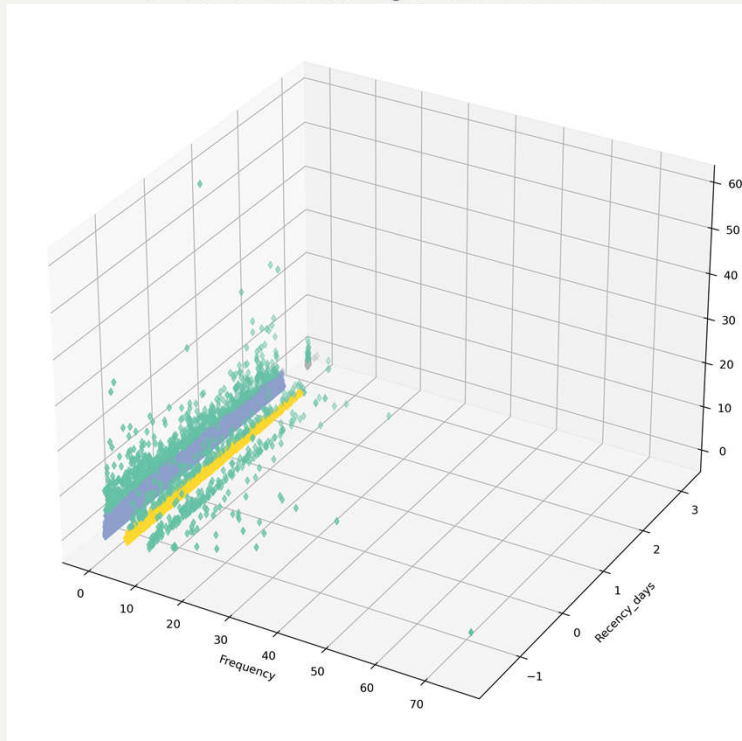
	count	mean	min	25%	50%	75%	max
Recency_days	2933.0	226.725878	0.00	111.00	206.00	325.00	696.00
Frequency	2933.0	2.110126	1.00	2.00	2.00	2.00	17.00
Monetary	2933.0	327.630075	35.94	145.75	225.63	361.02	13664.08

II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

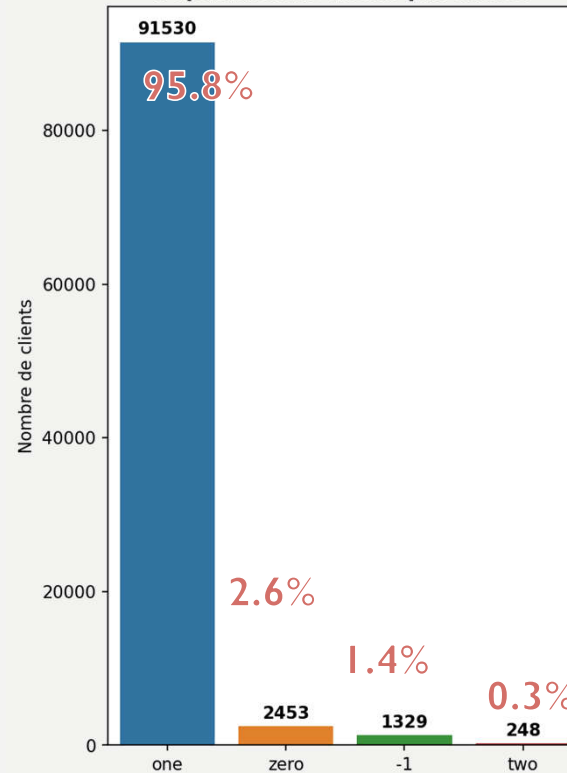
D'autres approches de segmentation

➤ DBSCAN (eps=0.5, min_samples=100)

Visualisation en 3D de la segmentation via DBSCAN



Population de clients par cluster



Les tendances pour le cluster 0

	count	mean
Recency_days	91530.0	243.640795
Frequency	91530.0	1.034186
Monetary	91530.0	166.469383

Les tendances pour le cluster 1

	count	mean
Recency_days	2453.0	252.720750
Frequency	2453.0	1.035874
Monetary	2453.0	156.830196

Les tendances pour le cluster -1

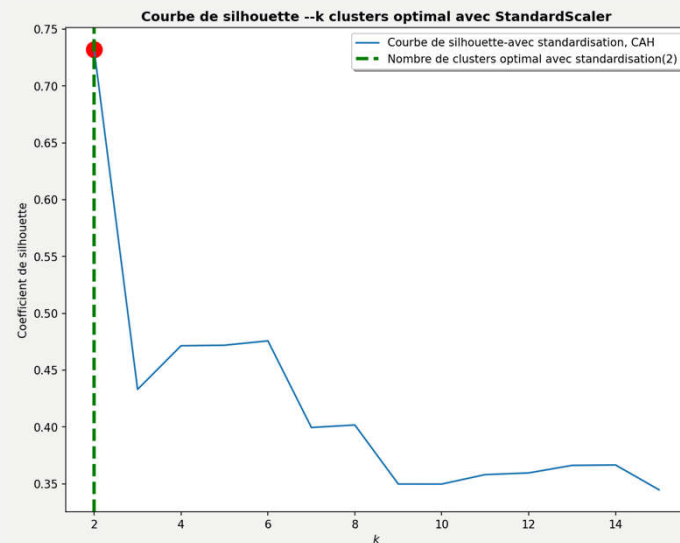
	count	mean
Recency_days	1329.0	247.833710
Frequency	1329.0	1.025583
Monetary	1329.0	153.980760

Les tendances pour le cluster 2

	count	mean
Recency_days	248.0	244.995968
Frequency	248.0	1.020161
Monetary	248.0	159.163266

II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

➤ Clustering Agglomérative Hiérarchique



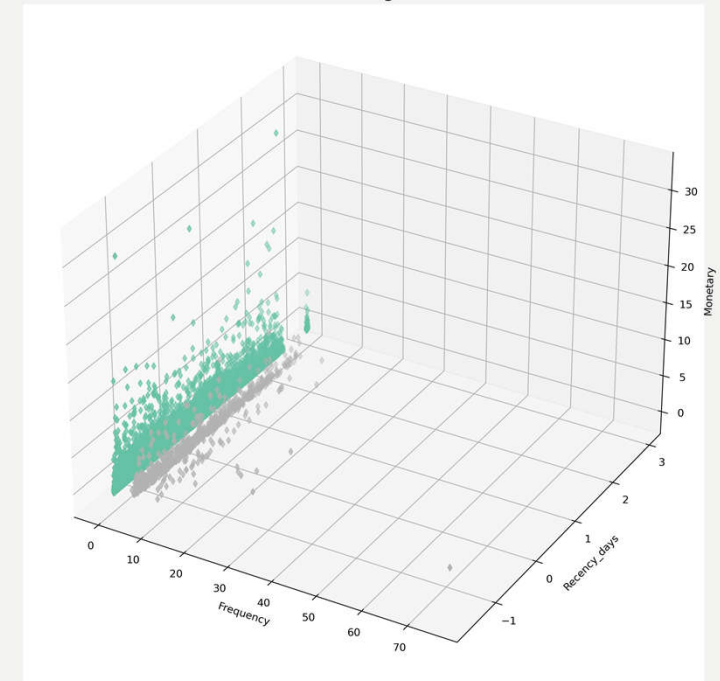
Les tendances pour le cluster 0

	count	mean	min	25%	50%	75%	max
Recency_days	27772.0	243.681982	5.00	120.00	225.00	354.000	700.00
Frequency	27772.0	1.032695	1.00	1.00	1.00	1.000	7.00
Monetary	27772.0	162.525060	10.07	62.69	107.87	182.495	7571.63

Les tendances pour le cluster 1

	count	mean	min	25%	50%	75%	max
Recency_days	896.0	247.632812	6.00	119.7500	229.00	355.0000	699.0
Frequency	896.0	1.034598	1.00	1.0000	1.00	1.0000	4.0
Monetary	896.0	160.204520	14.06	60.4375	99.33	174.9875	2566.9

Visualisation en 3D de la segmentation avec CAH



II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

Sans données de satisfaction

Les tendances pour le cluster 0

	count	mean	min	25%	50%	75%	max
Recency_days	92627.0	244.480648	5.00	120.00	226.00	354.00	729.0
Frequency	92627.0	1.000000	1.00	1.00	1.00	1.00	1.0
Monetary	92627.0	160.912271	9.59	62.15	105.66	177.25	4445.5

Les tendances pour le cluster 1

	count	mean	min	25%	50%	75%	max
Recency_days	2933.0	226.725878	0.00	111.00	206.00	325.00	696.00
Frequency	2933.0	2.110126	1.00	2.00	2.00	2.00	17.00
Monetary	2933.0	327.630075	35.94	145.75	225.63	361.02	13664.08

Ajout des données de satisfaction

Les tendances pour le cluster 0

	count	mean	min	25%	50%	75%	max
Recency_days	92626.0	243.580345	4.00	119.00	225.000	353.00	728.00
Frequency	92626.0	1.000000	1.00	1.00	1.000	1.00	1.00
Monetary	92626.0	160.866015	9.59	62.15	105.655	177.25	4194.76
review_score	92626.0	4.095913	1.00	4.00	5.000	5.00	5.00

Les tendances pour le cluster 1

	count	mean	min	25%	50%	75%	max
Recency_days	2934.0	225.802999	0.00	110.0000	205.000	324.0000	696.00
Frequency	2934.0	2.109748	1.00	2.0000	2.000	2.0000	17.00
Monetary	2934.0	329.033575	35.94	145.8125	225.655	361.0725	13664.08
review_score	2934.0	4.140811	1.00	3.5000	4.500	5.0000	5.00

Les clusters sont en moyenne satisfaits.

II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

Choix de $k = 4$ pour être moins généraliste comme avec $k=2$ sur la satisfaction



La stabilité des clusters et choix des paramètres de l'algorithme

Problème des optimums locaux

Configuration des clusters trouvés par K-Means peut ne pas être la plus **optimale**.

Solution

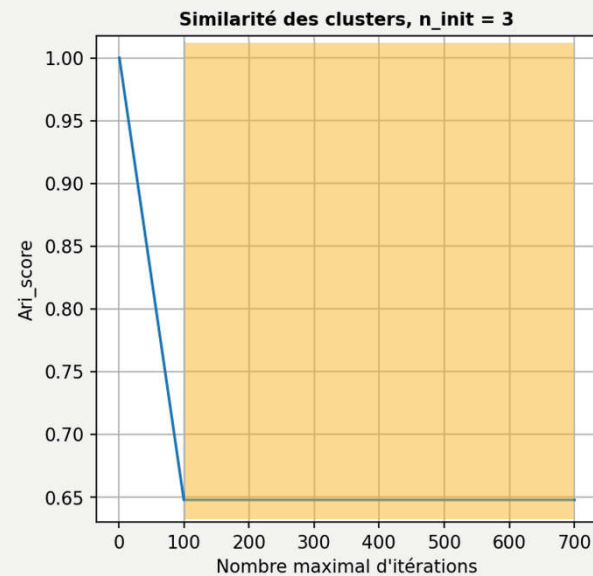
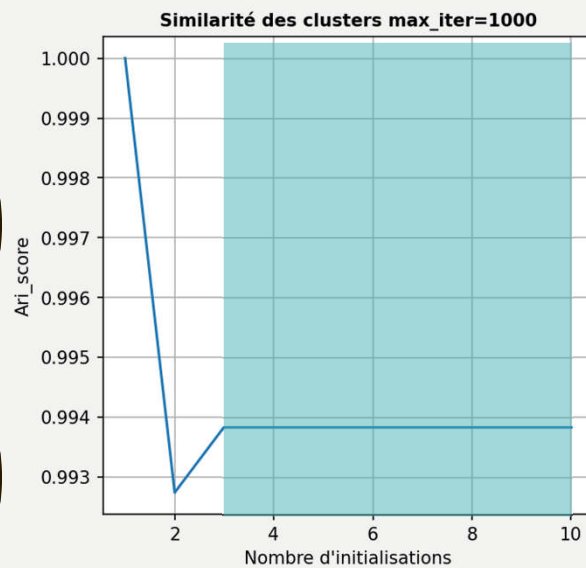
Lancer K-means **plusieurs fois** sur le jeu de données en fonction de :

- Nombre d'initialisations
- Nombre d'itérations

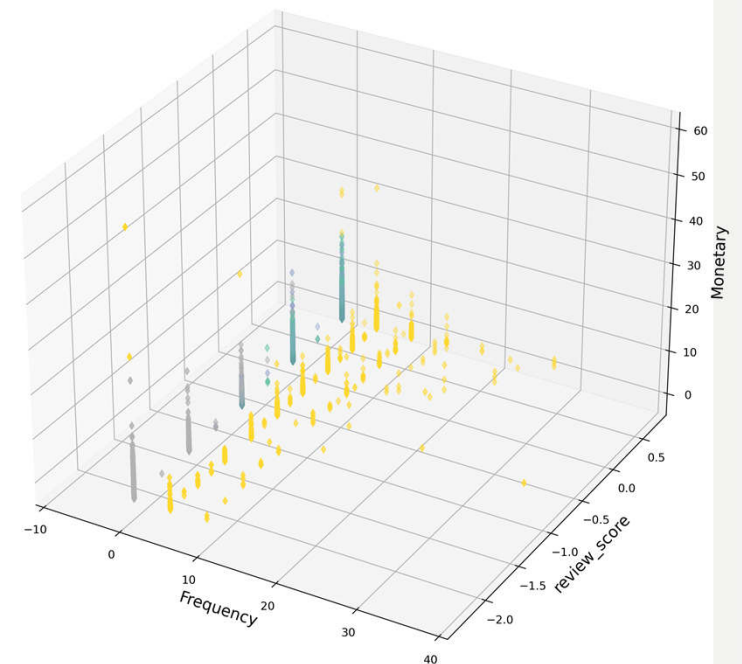
→ Faciliter le choix des hyperparamètres de l'algorithme de partitionnement

II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné

KMeans init = 'k-means ++'



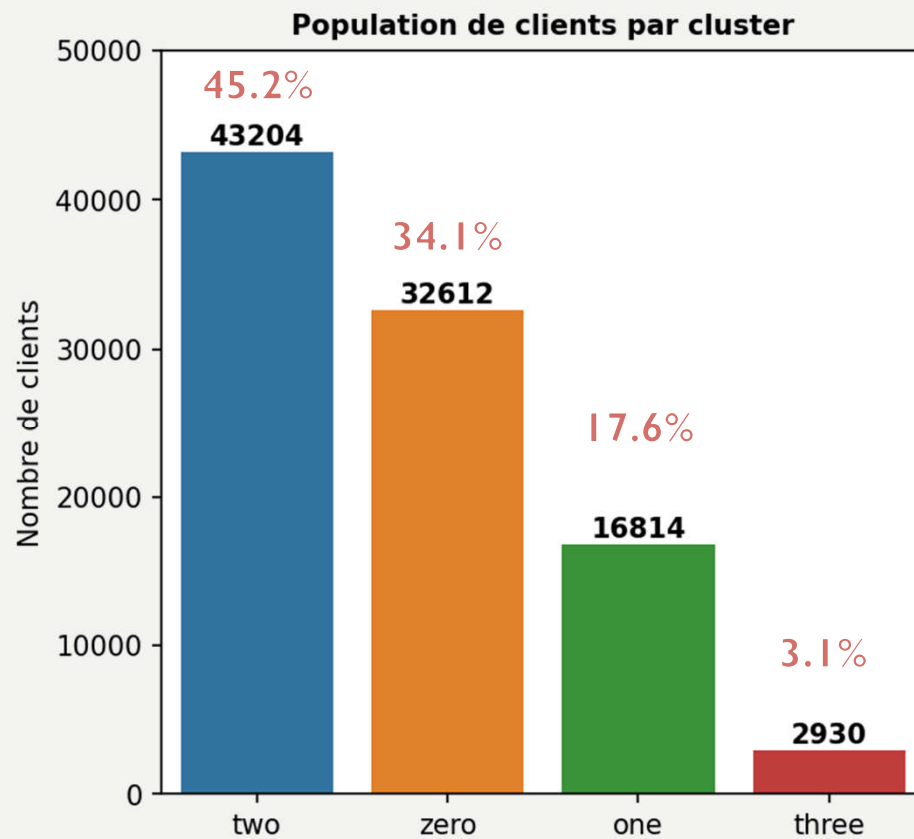
Visualisation en 3D de la segmentation avec StandardScaler



Modèle final stable :

KMeans(n_clusters=4, max_iter=1000, init='k-means++', n_init=3)

II Présentation des différentes pistes de modélisation effectuées et du modèle final sélectionné



Les tendances pour le cluster 0

	count	mean	min	25%	50%	75%	max
Recency_days	32612.0	397.438704	254.00	314.0000	387.00	470.0000	699.00
Frequency	32612.0	1.000000	1.00	1.0000	1.00	1.0000	1.00
Monetary	32612.0	157.345584	10.07	61.5875	102.69	172.5525	4764.34
review_score	32612.0	4.626089	2.00	4.0000	5.00	5.0000	5.00

Les tendances pour le cluster 1

	count	mean	min	25%	50%	75%	max
Recency_days	16814.0	245.970441	4.00	164.000	227.00	307.00	728.00
Frequency	16814.0	1.000000	1.00	1.000	1.00	1.00	1.00
Monetary	16814.0	187.514372	13.78	67.095	116.94	200.67	6081.54
review_score	16814.0	1.603426	1.00	1.000	1.00	2.00	4.00

Les tendances pour le cluster 2

	count	mean	min	25%	50%	75%	max
Recency_days	43204.0	126.510578	4.00	63.000	124.00	188.00	263.00
Frequency	43204.0	1.000000	1.00	1.000	1.00	1.00	1.00
Monetary	43204.0	153.595948	9.59	60.455	103.68	172.62	4513.32
review_score	43204.0	4.665702	3.00	4.000	5.00	5.00	5.00

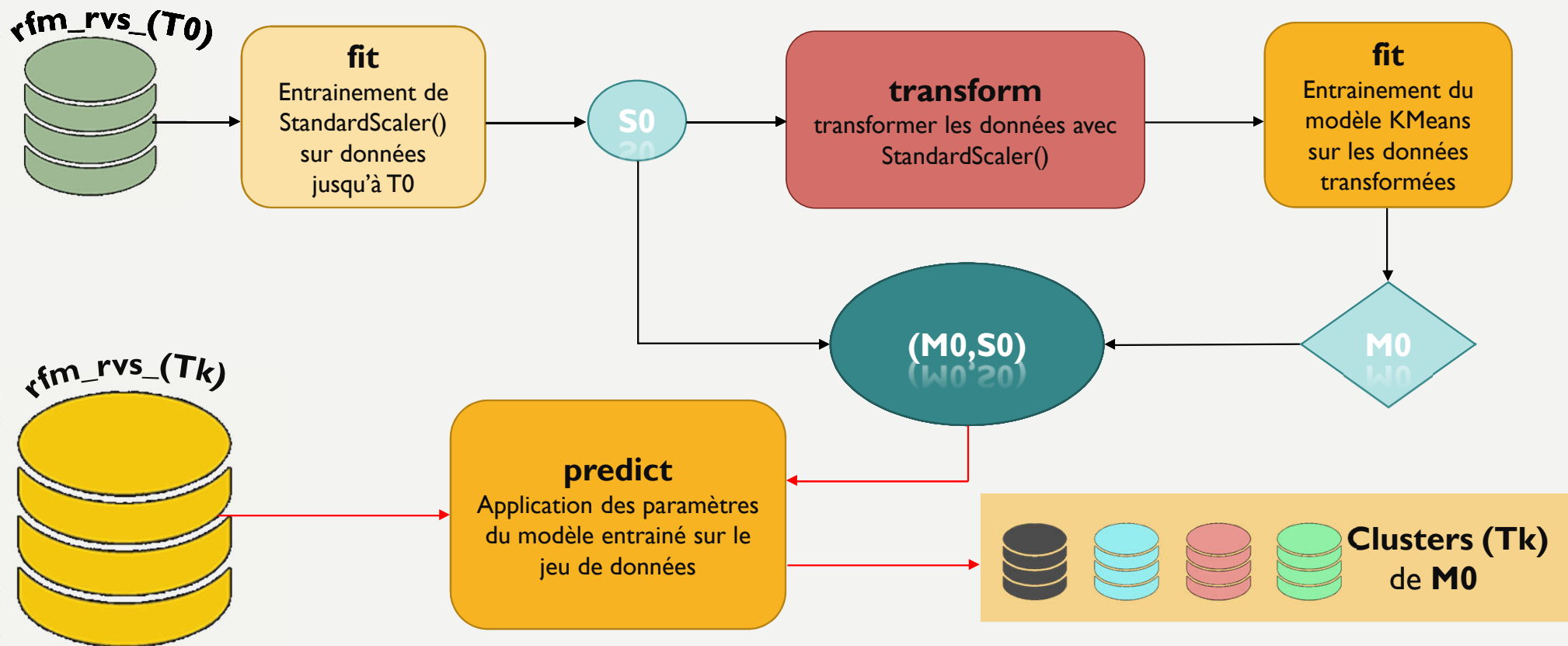
Les tendances pour le cluster 3

	count	mean	min	25%	50%	75%	max
Recency_days	2930.0	225.801365	0.00	110.0000	205.00	324.000	696.00
Frequency	2930.0	2.111263	1.00	2.0000	2.00	2.000	17.00
Monetary	2930.0	322.723485	35.94	145.7275	225.58	360.205	13664.08
review_score	2930.0	4.141345	1.00	3.5000	4.50	5.000	5.00

III Présentation de la simulation pour définir le délai de maintenance du modèle (contrat de maintenance)

```
S0 = StandardScaler.fit(rfm_rvs(T0))
```

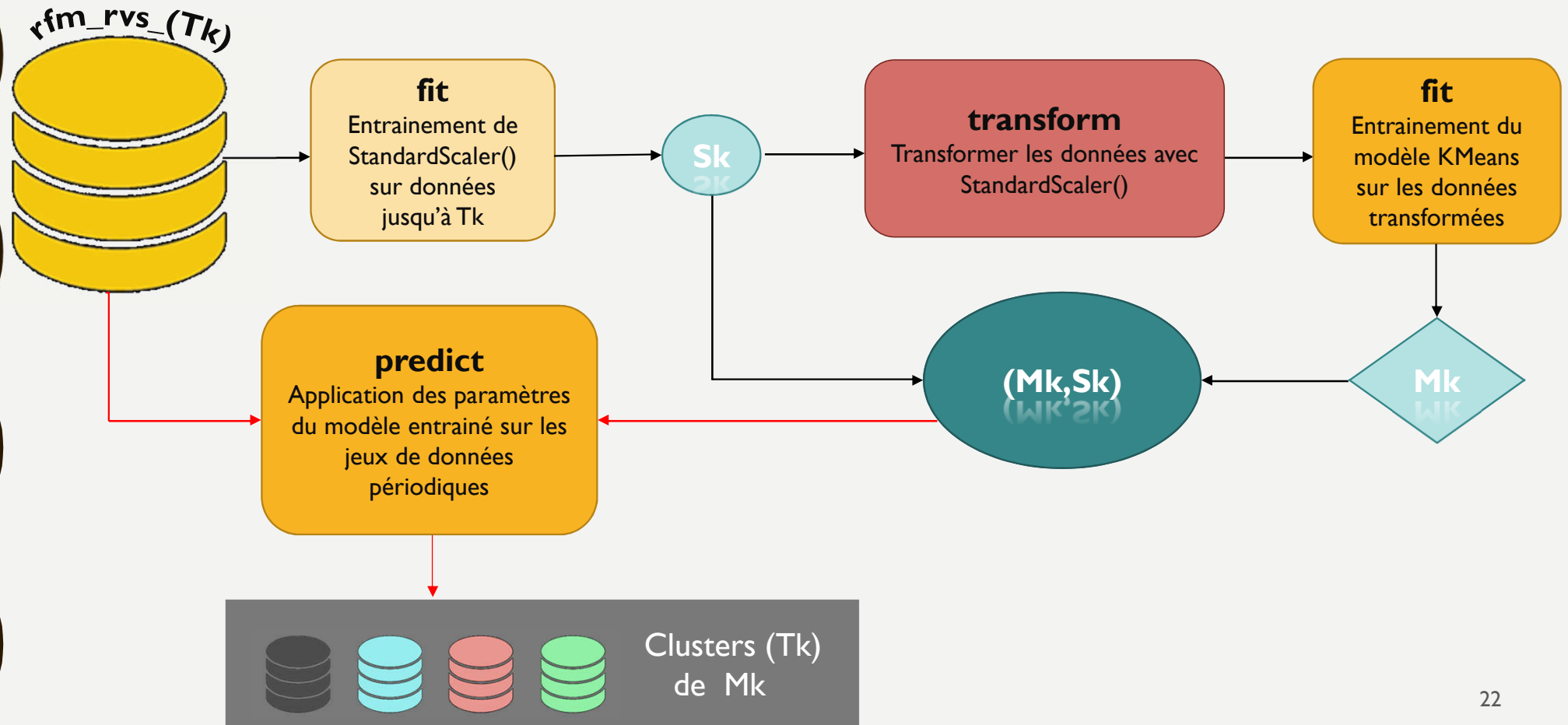
```
M0 = KMeans(n_clusters= 4, max_iter=1000, n_init=3, init='k-means++', random_state=0).fit(S0.transform(rfm_rvs(T0)))
```



III Présentation de la simulation pour définir le délai de maintenance du modèle (contrat de maintenance)

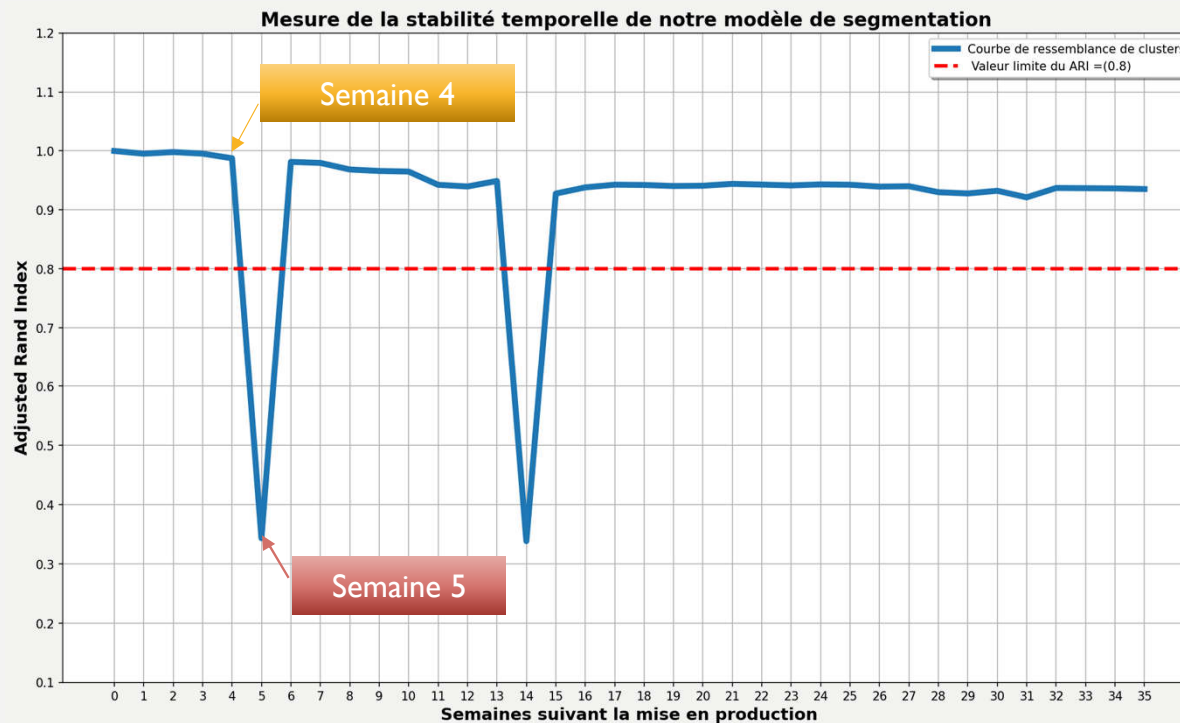
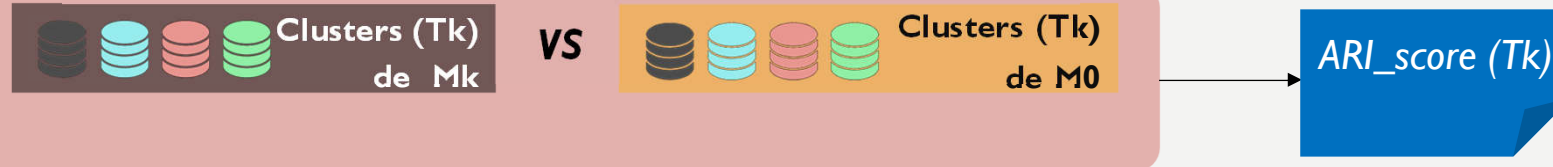
```
Sk = StandarScaler.fit(rfm_rvs(Tk))
```

```
Mk= KMeans(n_clusters= 4, max_iter=1000, n_init=3, init='k-means++', random_state=0).fit(Sk.transform (rfm_rvs(Tk)))
```



III Présentation de la simulation pour définir le délai de maintenance du modèle (contrat de maintenance)

Mesure de concordance avec **Adjusted_rand_score**







IV Conclusion

1

Choix de l'algorithme K-Means

2

Segmentation	Fréquence d'achat	Satisfaction
2 clusters		
4 clusters		

3

Simulation du modèle stable à 4 clusters à déployer sur les données



Contrat de Maintenance

4 semaines (1 mois) après le déploiement du modèle.