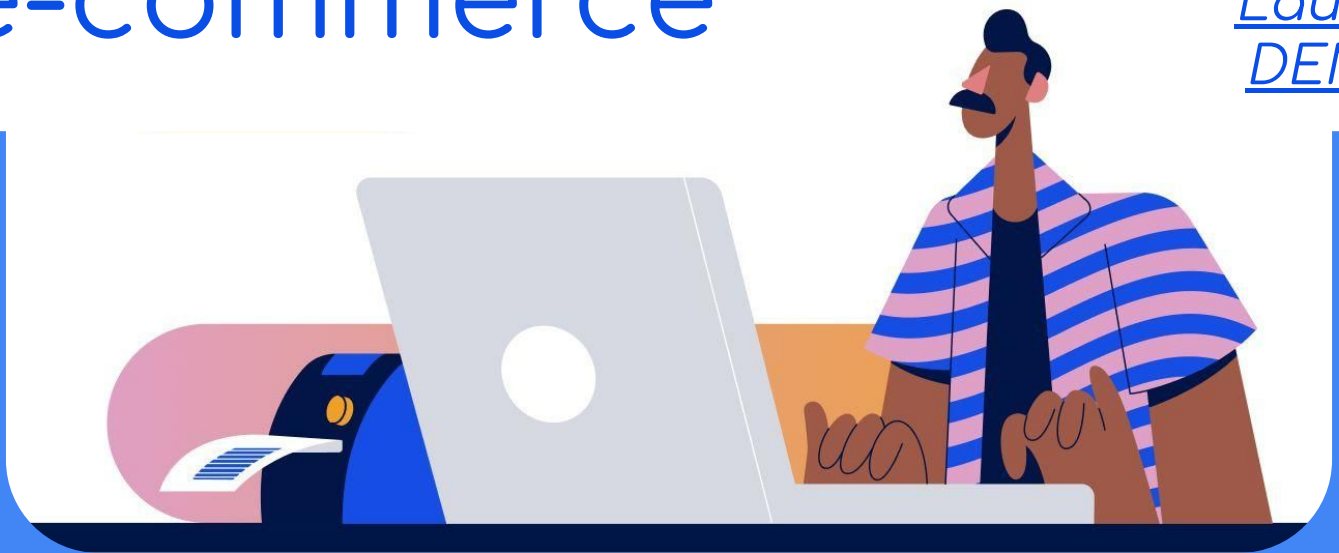
Segmenter des clients d'un site e-commerce

OPENCLASSROOMS

Projet n°5

Laureenda DEMEULE



olist

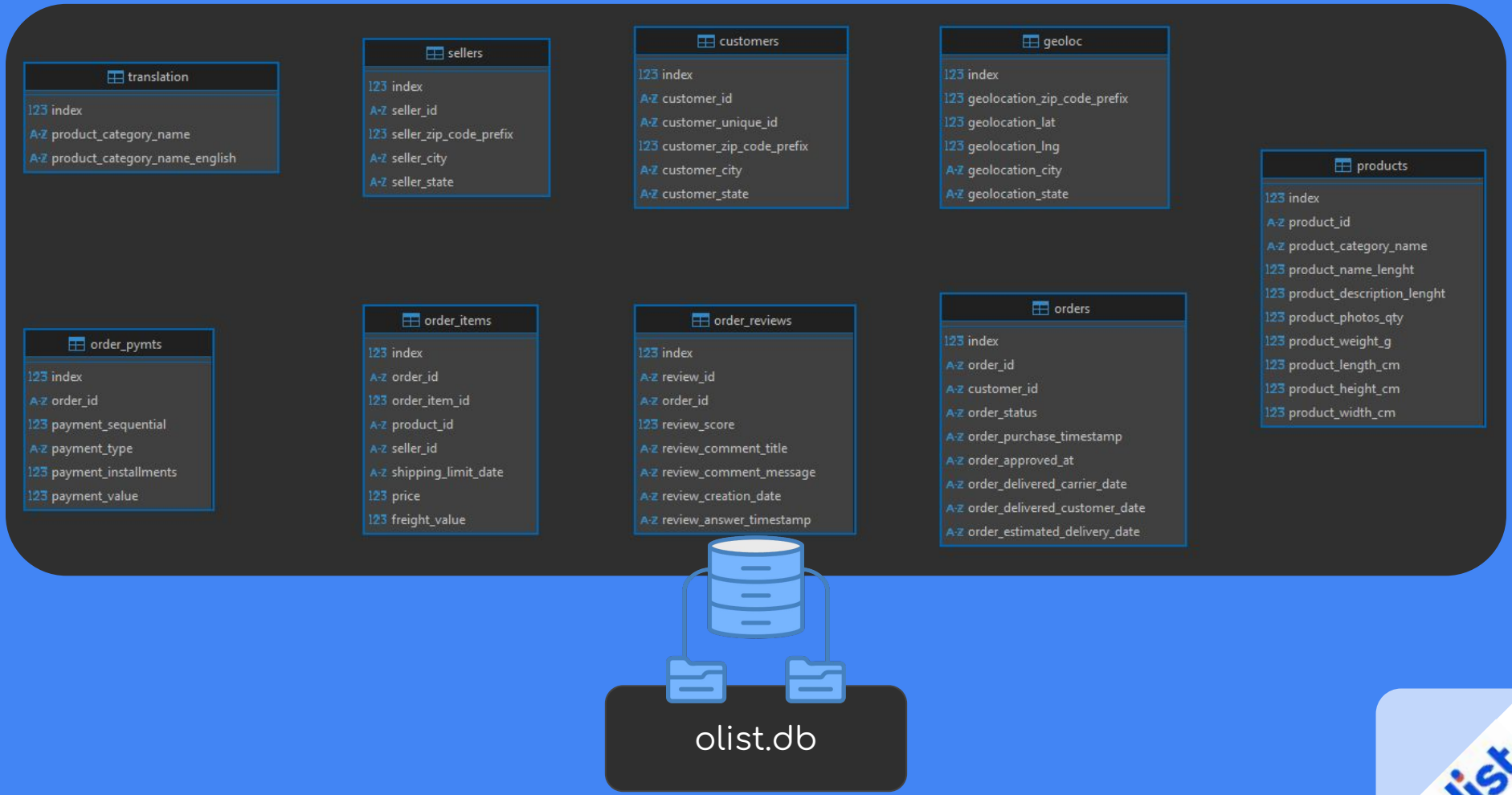
Sommaire

Contexte

- Marketplace brésilienne
  - expédition de produits
  - services logistiques
- Le jeu de données couvre une période de 24 mois
- Explorer les dynamiques du marché

Objectifs

- Répondre aux requêtes SQL
- Fournir une segmentation exploitable par l'équipe Marketing
- Différenciation des clients selon des critères de performance


translation
123 index
A-Z product_category_name
A-Z product_category_name_english
sellers
123 index
A-Z seller_id
123 seller_zip_code_prefix
A-Z seller_city
A-Z seller_state
customers
123 index
A-Z customer_id
A-Z customer_unique_id
123 customer_zip_code_prefix
A-Z customer_city
A-Z customer_state
geoloc
123 index
123 geolocation_zip_code_prefix
123 geolocation_lat
123 geolocation_lng
A-Z geolocation_city
A-Z geolocation_state
products
123 index
A-Z product_id
A-Z product_category_name
123 product_name_lenght
123 product_description_lenght
123 product_photos_qty
123 product_weight_g
123 product_length_cm
123 product_height_cm
123 product_width_cm
order_pymts
123 index
A-Z order_id
123 payment_sequential
A-Z payment_type
123 payment_installments
123 payment_value
order_items
123 index
A-Z order_id
123 order_item_id
A-Z product_id
A-Z seller_id
A-Z shipping_limit_date
123 price
123 freight_value
order_reviews
123 index
A-Z review_id
A-Z order_id
123 review_score
A-Z review_comment_title
A-Z review_comment_message
A-Z review_creation_date
A-Z review_answer_timestamp
orders
123 index
A-Z order_id
A-Z customer_id
A-Z order_status
A-Z order_purchase_timestamp
A-Z order_approved_at
A-Z order_delivered_carrier_date
A-Z order_delivered_customer_date
A-Z order_estimated_delivery_date
olist.db

```
# Création d'un objet Cursor pour interagir avec la base de données
cur = conn.cursor()

# Récupération de la date de la dernière commande par client
```

	CustomerId	Importance	Recency	Frequency	Monetary	Review
0	0000366f3b9a7992bf8c76cfdf3221e2	False	2018-05-10 10:56:27	1	141.90	5.0
1	0000b849f77a49e4a4ce2b2a4ca5be3f	False	2018-05-07 11:11:27	1	27.19	4.0
2	0000f46a3911fa3c0805444483337064	False	2017-03-10 21:05:03	1	86.22	3.0
3	0000f6ccb0745a6a4b88665a16c9f078	False	2017-10-12 20:29:41	1	43.62	4.0
4	0004aac84e0df4da2b147fca70cf8255	False	2017-11-14 19:45:42	1	196.89	5.0

```
cur.execute("""
SELECT customer_unique_
FROM customers
JOIN orders ON customer
JOIN order_pymts ON orders.order_id = order_pymts.order_id
GROUP BY customer_unique_id;
""")
Monetary = cur.fetchall()
```

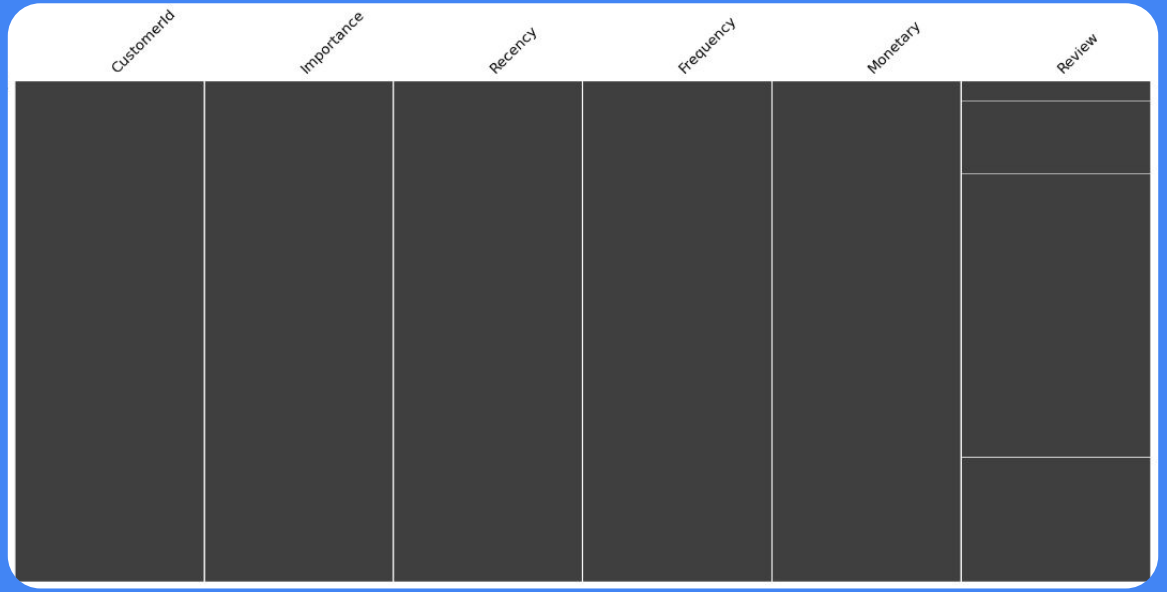
Conversion et Feature Engineering : Création de features par client

RFM

olist

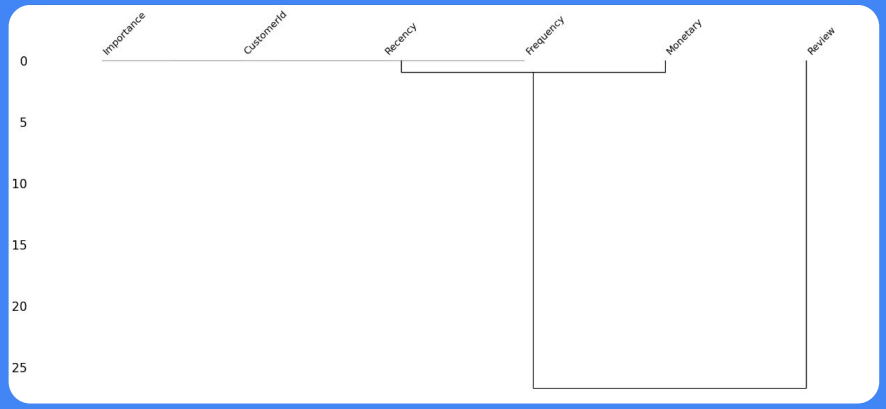
Exploration et nettoyage des données

- Identification et traitement des valeurs doublons
- Identification et traitement des valeurs manquantes

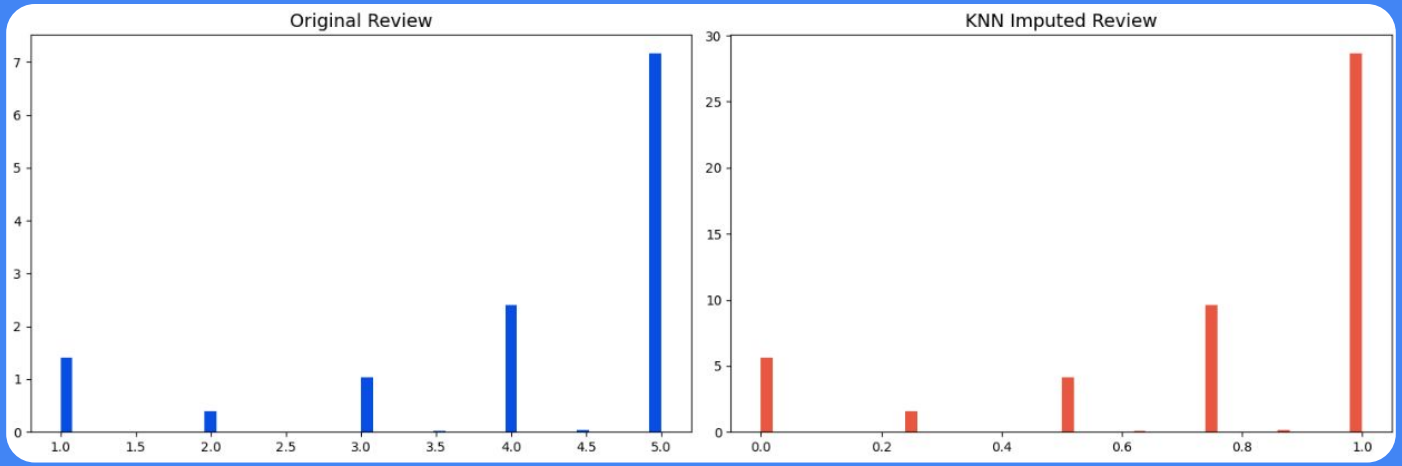



```
Nombre total de valeurs manquantes dans le dataset =  717 environ 0.12 % du dataset.
```

```
column_name : nbr_values -> nbr_of_Na

CustomerId : 96096 valeurs -> 0 Na
Importance : 96096 valeurs -> 0 Na
Recency : 96096 valeurs -> 0 Na
Frequency : 96096 valeurs -> 0 Na
Monetary : 96095 valeurs -> 1 Na
Review : 95380 valeurs -> 716 Na
```



	Nom du test	Statistique de test	p-valeur	Normalité
0	Shapiro-Wilk	0.691650	0.000	Non
1	Kolmogorov-Smirnov	0.863798	0.000	Non
2	Anderson-Darling	12118.263484	NaN	Non
3	D'Agostino-Pearson	18164.113444	0.000	Non
4	Lilliefors	0.324883	0.001	Non



Exploration et nettoyage des données

- Identification et traitement des valeurs doublons
- Identification et traitement des valeurs manquantes
- Identification et traitement des valeurs aberrantes

CustomerId Importance Recency Frequency Monetary Review

	Variable	Shapiro-Wilk	Kolmogorov-Smirnov	Anderson-Darling	D'Agostino-Pearson
0	Review	False	False	False	False
1	Monetary	False	False	False	False
2	Frequency	False	False	False	False
3	Recency	False	False	False	False

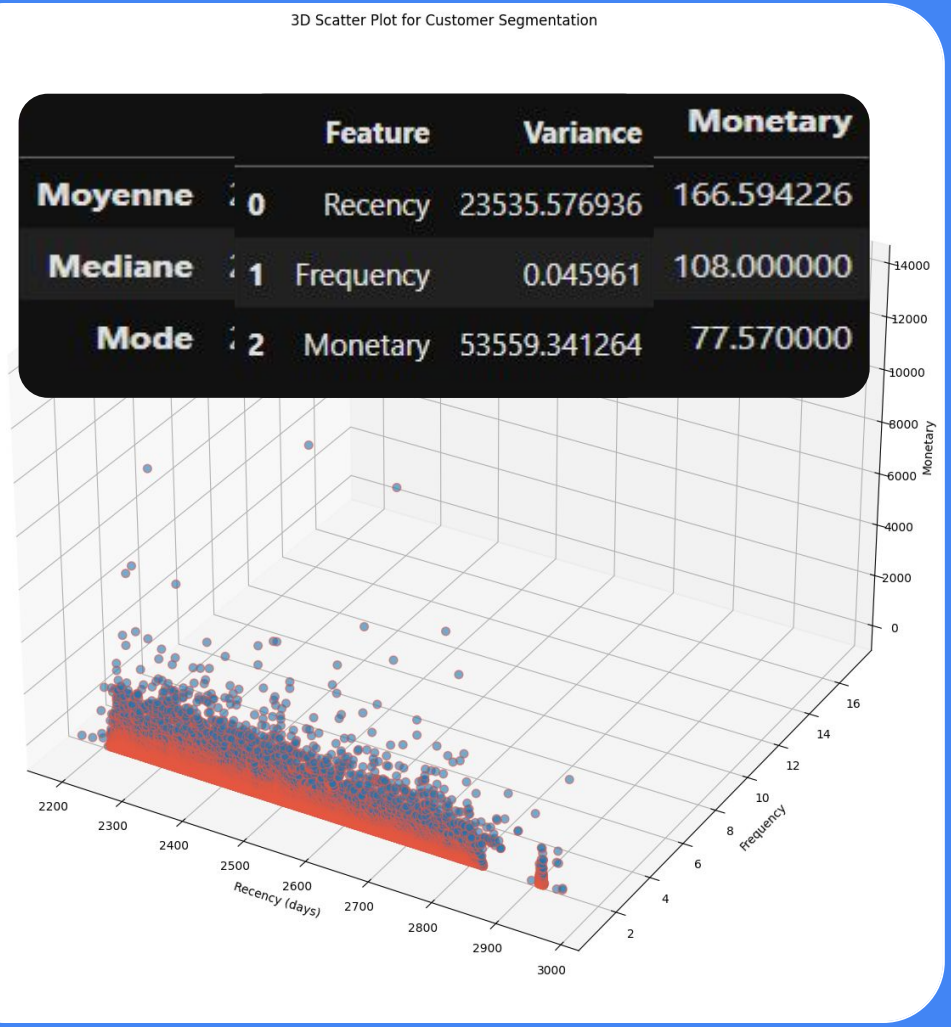
	Centile	Monetary	Frequency
0	25%	63.1200	1.0
1	50%	108.0000	1.0
2	75%	183.5300	1.0
3	90%	319.5700	1.0
4	95%	476.1520	1.0
5	99%	1122.4662	2.0

```
Nombre total de valeurs m                    environ 0.0 % du dataset.

column                                                    Na

Custom
Import
Recenc
Frequenc
Monetary : 96095 valeurs -> 0 Na
Review : 96095 valeurs -> 0 Na
```

Exploration et nettoyage des données

- Identification et traitement des valeurs doublons
- Identification et traitement des valeurs manquantes
- Identification et traitement des valeurs aberrantes
- Visualisation des données

		Feature	Variance	Monetary
Moyenne	0	Recency	23535.576936	166.594226
Mediane	1	Frequency	0.045961	108.000000
Mode	2	Monetary	53559.341264	77.570000

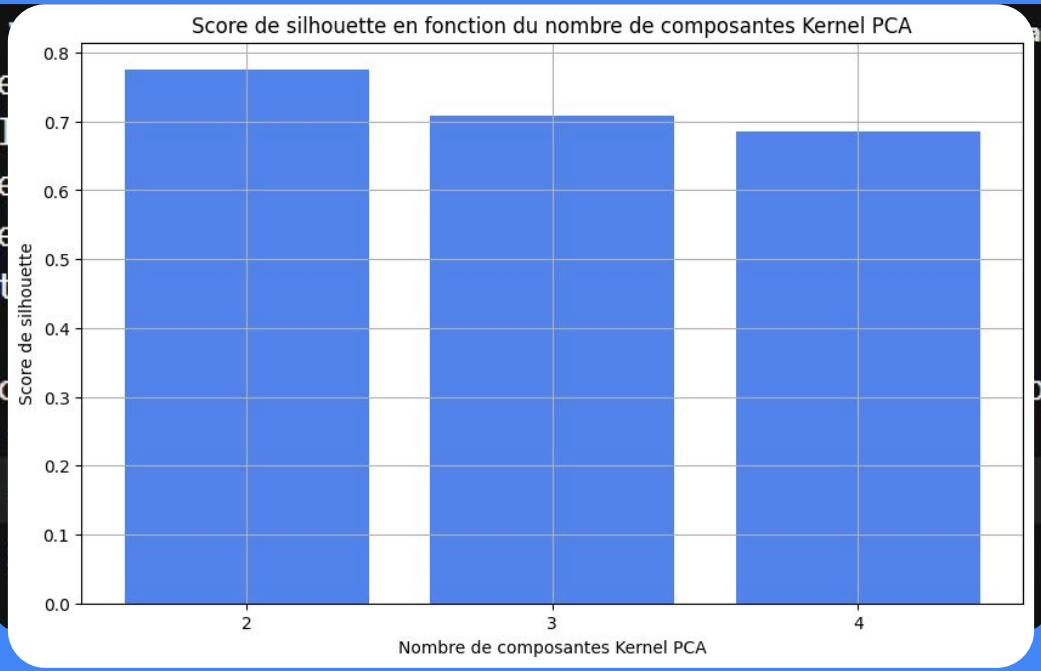


Choix des méthodes de segmentation des données

Etude de la normalisation

Etude de la réduction dimensionnelle

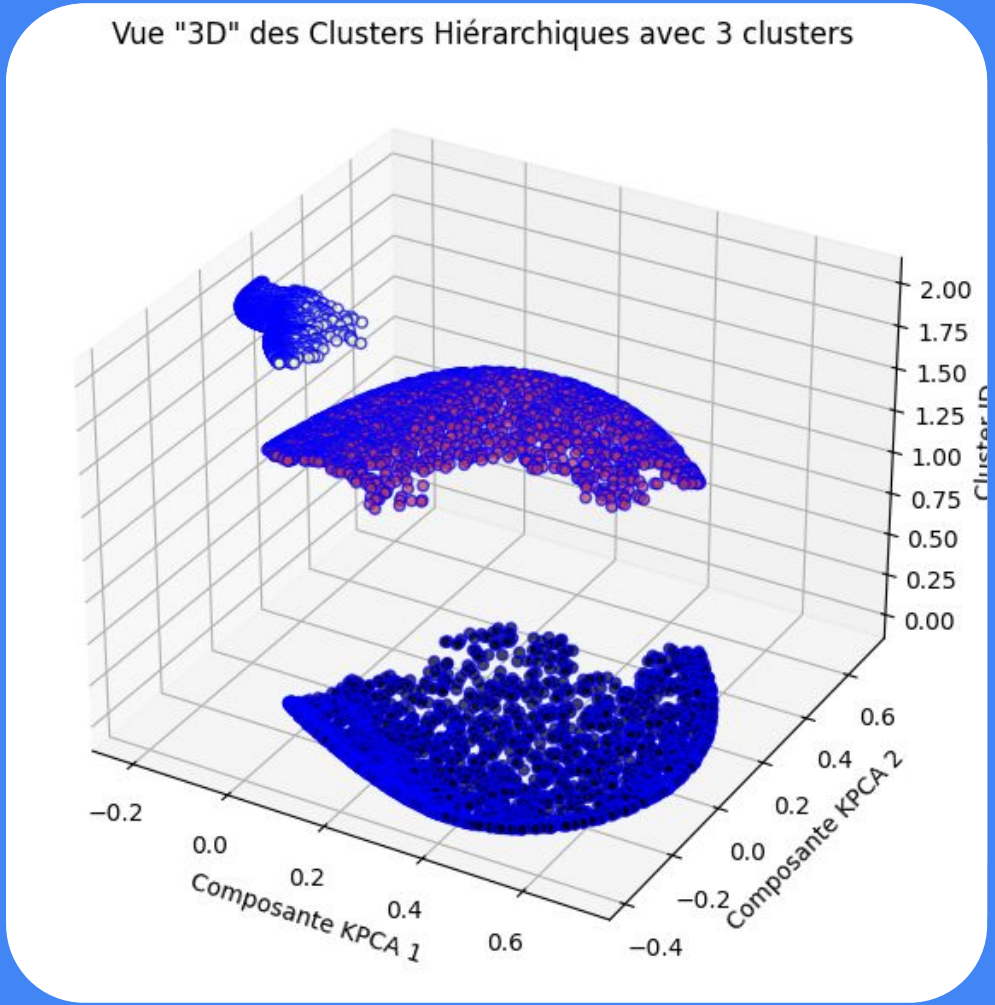
```
PCA Silhouette
Kernel PCA Sil
Isomap Silhoue
LLE Silhouette
t-SNE Silhouet

La méthode rec
```




```
ouette de 0.76
```

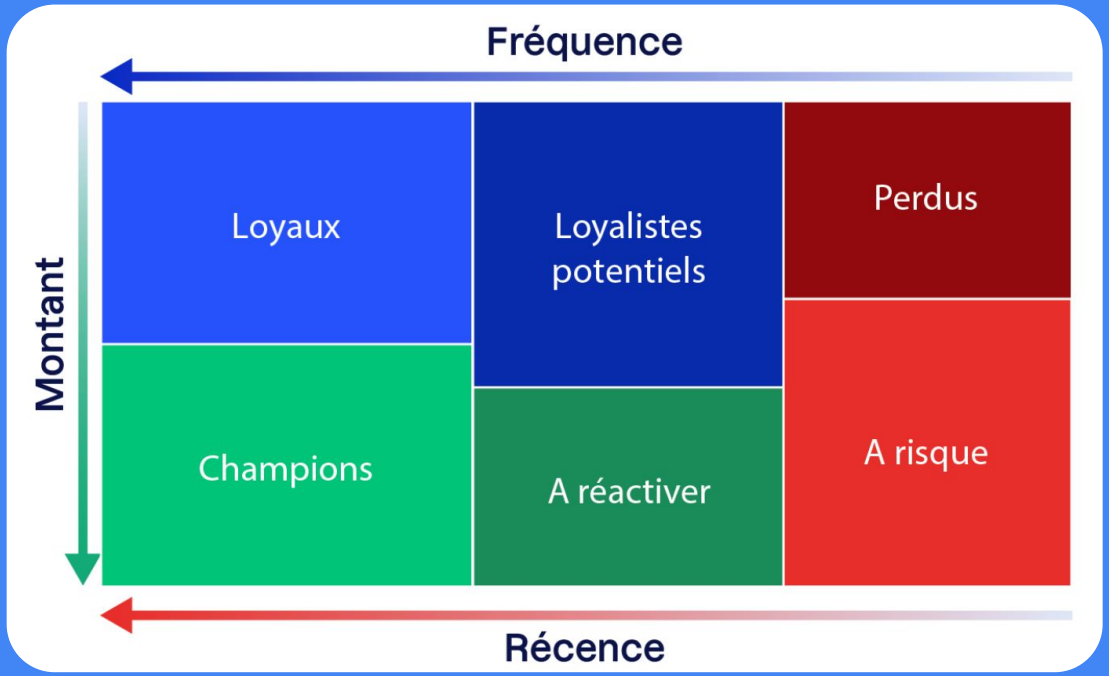
Choix des méthodes de segmentation des données

Modélisation

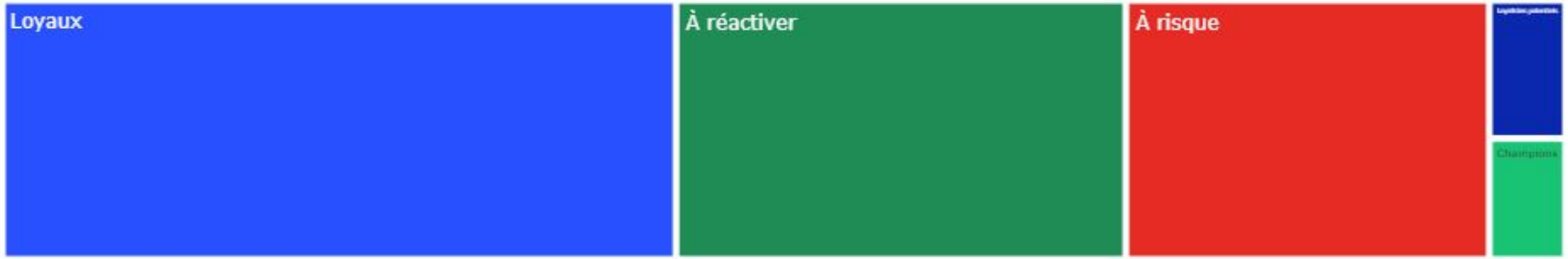


Choix des méthodes de segmentation des données

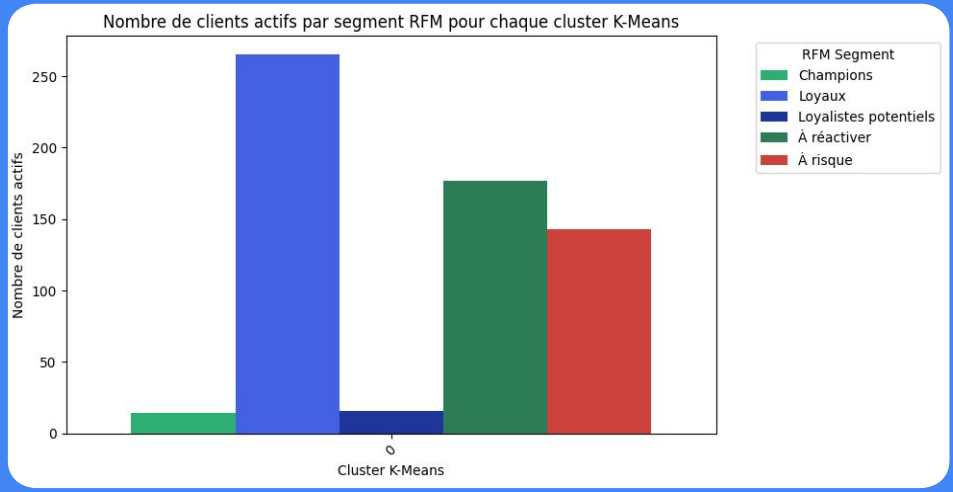
Segmentation RFM

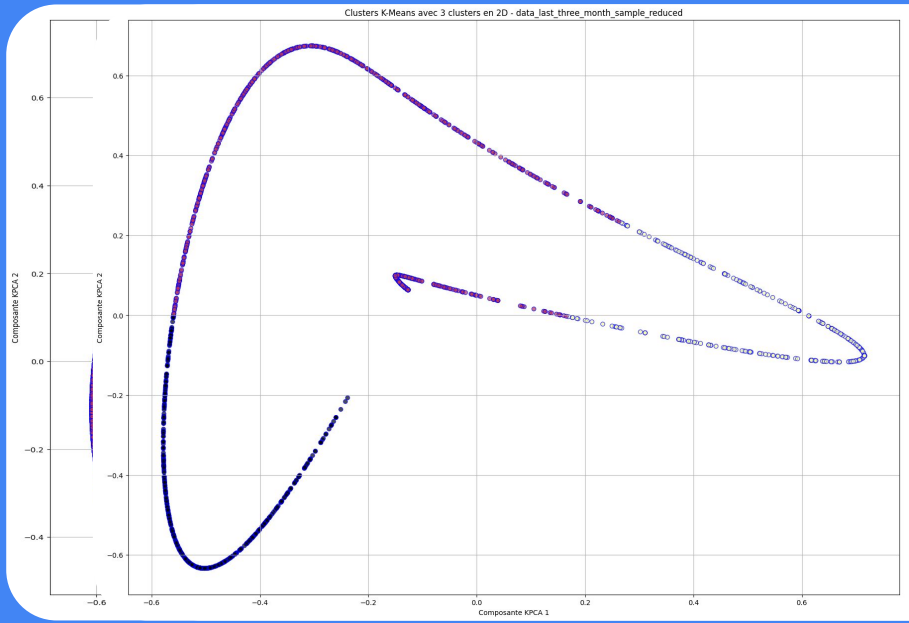


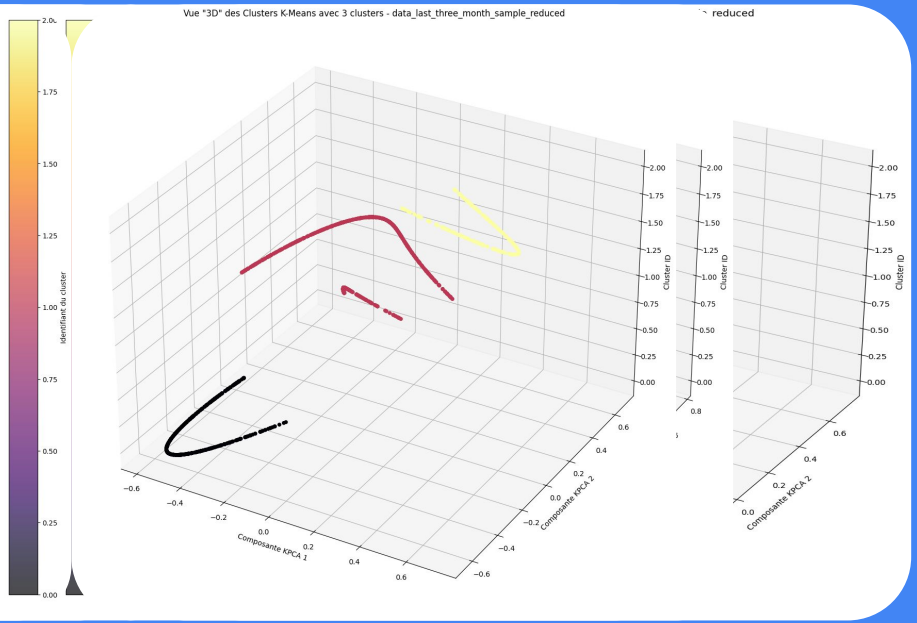
Treemap des segments RFM pour les clients actifs dans le Cluster 0



	Recency_Score	Frequency_Score	Monetary_Score	RFM_Segment
45311	2	1	3	À réactiver
15221	3	1	3	À risque
73285	2	1	4	Loyalistes potentiels
23140	4	1	3	Loyaux
48776	5	1	5	Champions

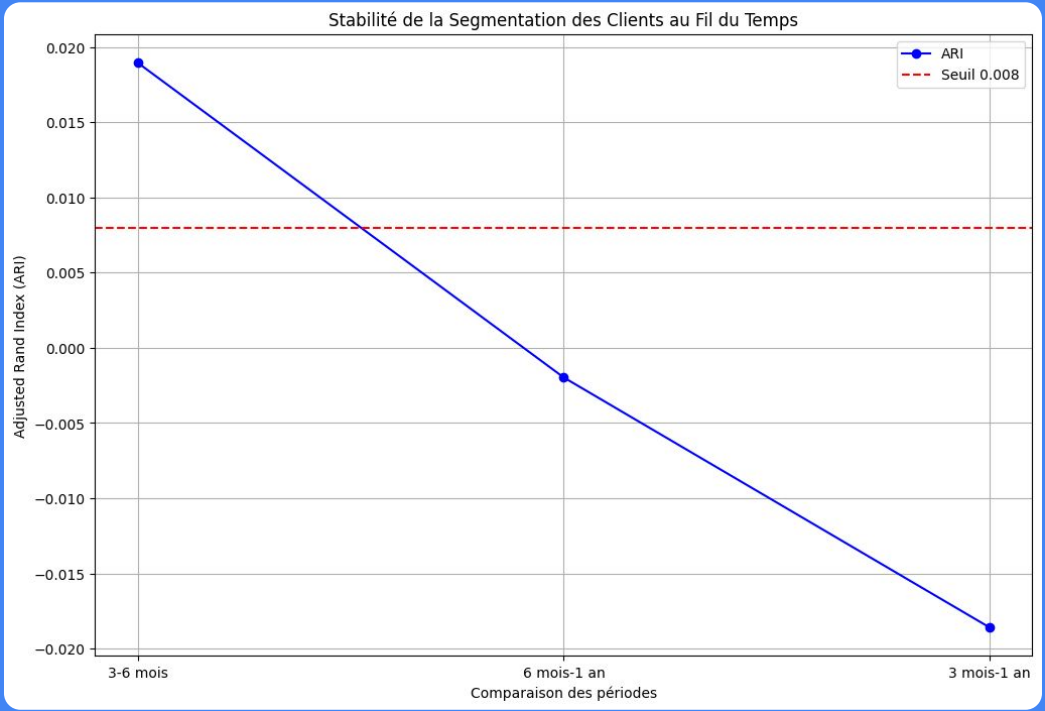






Analyse de stabilité

3 mois → 6 mois → 1 an



Fréquence de mise à jour → tous les 3 mois

Merci de votre attention !