

# SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE

Présentation "Projet 5" chez "OPENCLASSROOM"  
Jaoid KRAIRI  
(Novembre 2021)

# SOMMAIRE

- ➡ Présentation de la problématique
- ➡ Présentation du cleaning effectué
- ➡ Présentation des différentes pistes de modélisation effectuées
- ➡ Présentation du modèle final
- ➡ Conclusion
- ➡ Remerciements

**Présentation de la problématique :**  
**1/ Rappel du contexte**

# Rappel du contexte

- ✓ Fournir aux équipes d'e-commerce une segmentation des clients pour les campagnes de communication
- ✓ Comprendre les différents types d'utilisateurs
- ✓ Fournir une description actionnable de la segmentation
- ✓ Faire une proposition de contrat de maintenance

## Présentation de la problématique : 2/ Interprétation

- Exploration des données et choix de caractéristiques adaptées,
- Problème de classification non supervisée,
- Les clusters devront être explicables et réutilisables pour des campagnes de communication.

# PRESENTATION DU CLEANING EFFECTUÈ :

## 1/ Cleaning

Données réparties en 9 jeu de données :

Clients / Géolocalisation / Commander des articles / Paiements de commande/ Avis de commande / Commande / Produits / Vendeurs / Traduction des catégories de produits.

Principales étapes du nettoyage:

- ✓ Définir le jeu de données à traiter,
- ✓ Créer un jeu de données utile,
- ✓ Séparation en 2 partie du jeu de données utile,
- ✓ Imputation des valeurs manquantes,
- ✓ Réunir les 2 jeux de données catégorique et numérique

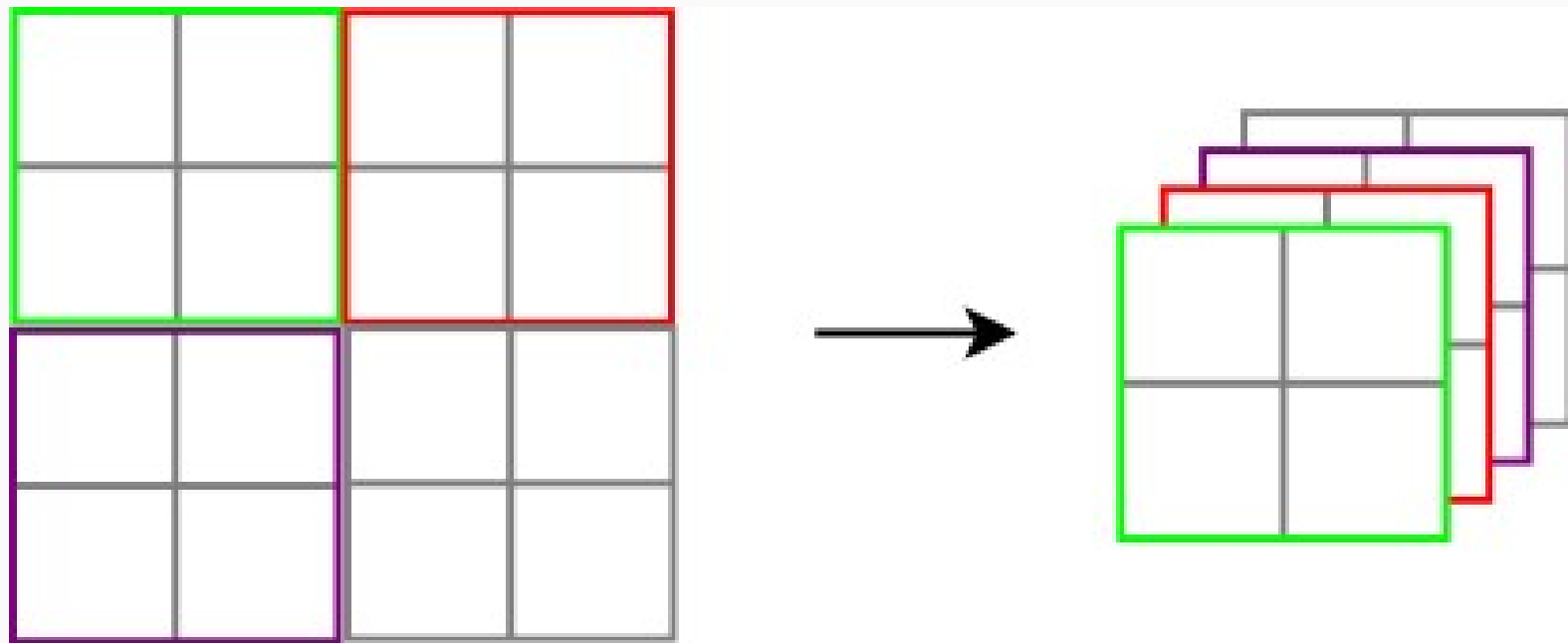
## PRESENTATION DU CLEANING EFFECTUÈ :

### 2/ Nombre de valeurs uniques par variable

customer_unique_id	96096
customer_id	99441
customer_city	4119
customer_state	27
order_id	99441
order_purchase_timestamp	98875
order_delivered_customer_date	95664
product_id	32951
payment_type	5
product_category_name_english	71
geolocation_lat	691481
geolocation_lng	691745
order_item_id	21
price	5968
freight_value	6999
payment_installments	24
payment_value	29077
review_score	5
dtype: int64	

## PRESENTATION DU CLEANING EFFECTUÈ :

### 3/ Créer des familles de produits



## PRESENTATION DU CLEANING EFFECTUÈ :

### 4/ Mettre au format les dates

#### Avant :

```
#   Column      Dtype
---  -
0   customer_unique_id  object
1   customer_id         object
2   customer_city       object
3   customer_state      object
4   order_id            object
5   order_purchase_timestamp  object
6   order_delivered_customer_date  object
7   product_id          object
8   payment_type        object
9   product_category_name_english  object
10  geolocation_lat     float64
11  geolocation_lng     float64
12  order_item_id       float64
13  price               float64
14  freight_value       float64
15  payment_installments float64
16  payment_value       float64
17  review_score        int64
dtypes: float64(7), int64(1), object(10)
memory usage: 2.4+ GB
```

#### Après :

```
#   Column      Dtype
---  -
0   customer_unique_id  object
1   customer_id         object
2   customer_city       object
3   customer_state      object
4   order_id            object
5   order_purchase_timestamp  datetime64[ns]
6   order_delivered_customer_date  datetime64[ns]
7   product_id          object
8   payment_type        object
9   product_category_name_english  object
10  geolocation_lat     float64
11  geolocation_lng     float64
12  order_item_id       float64
13  price               float64
14  freight_value       float64
15  payment_installments float64
16  payment_value       float64
17  review_score        int64
dtypes: datetime64[ns](2), float64(7), int64(1), object(8)
memory usage: 2.4+ GB
```



**PRESENTATION DU CLEANING EFFECTUÈ :**  
**5/ Création variables fonctionnels**

## Quatre variables fonctionnels:

- ✓ Délai dernier achat,
- ✓ Délai de livraison,
- ✓ Heure d'achat,
- ✓ Jour d'achat.

## PRESENTATION DU CLEANING EFFECTUÈ :

### 6/ Créer ma nouvelle base de données

Agrégation groupées sur customer\_unique\_id :

- ✓ Nombre de produits(Nb\_pdts),
- ✓ Nombre de commande(Nb\_com),
- ✓ Montant max d'achat(Mont\_max\_achats),
- ✓ Montant moyen d'achat(Mont\_moy\_achats),
- ✓ Note moyenne par commande(Note\_Moy\_Com),
- ✓ Date dernier achat(Date\_dernier\_achat),
- ✓ Villes,
- ✓ Catégorie de produits(Cat\_prod),
- ✓ Délai moyen dernier achat(Delai\_moy\_dernier\_achat),
- ✓ Délai moyen de livraison(Delai\_Moy\_Livraison),
- ✓ Heure d'achat moyen(Heure\_achat\_moy),
- ✓ Jour d'achat moyen(Jour\_achat\_moy),
- ✓ Type de paiement(Type\_paiement),
- ✓ Note facilité de paiement moyen(Note\_facil\_pai\_moy),
- ✓ Latitude,
- ✓ Longitude.

Jeu  
finale

96 096 clients

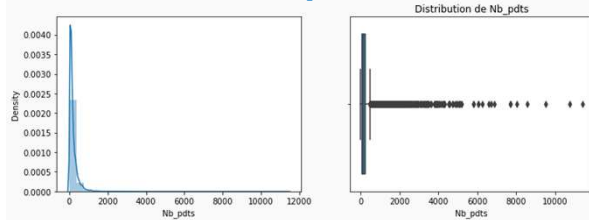
16 variables

# PRESENTATION DU CLEANING EFFECTUÈ :

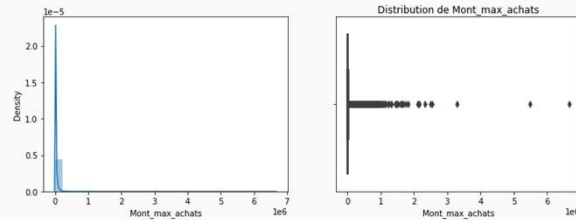
## 7/ Exploration du jeu de données

### 1/ Outliers atypiques

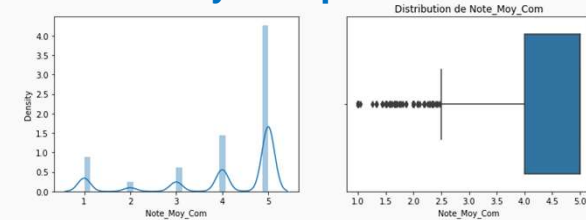
#### Nombre de produits acheté



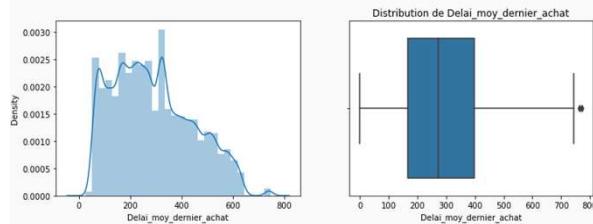
#### Montant max d'achat



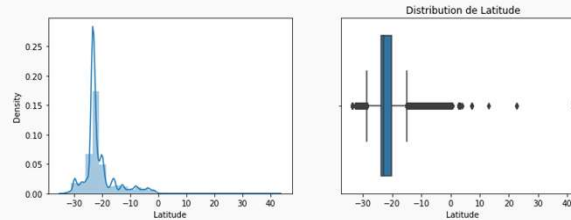
#### Note moyenne par commande



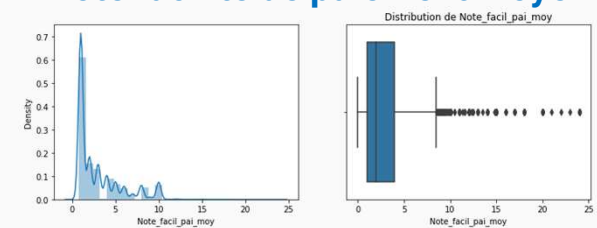
#### Délai moyen dernier achat



#### Latitude



#### Note facilité de paiement moyen

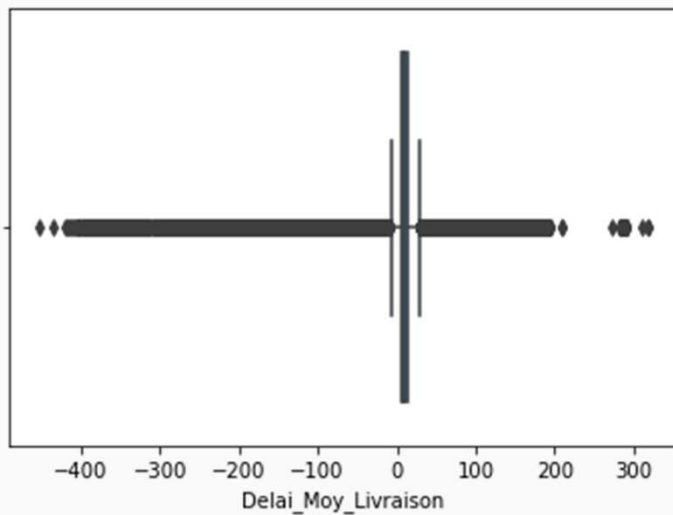


## PRESENTATION DU CLEANING EFFECTUÈ :

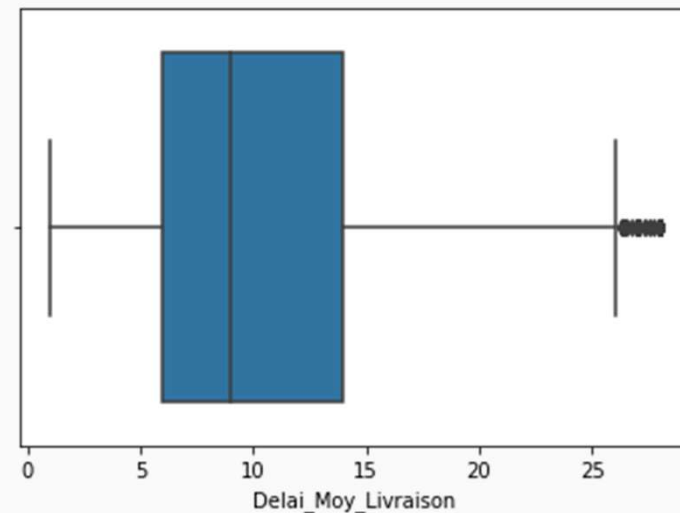
7/ Exploration du jeu de données

2/ Remplacer les valeurs aberrantes

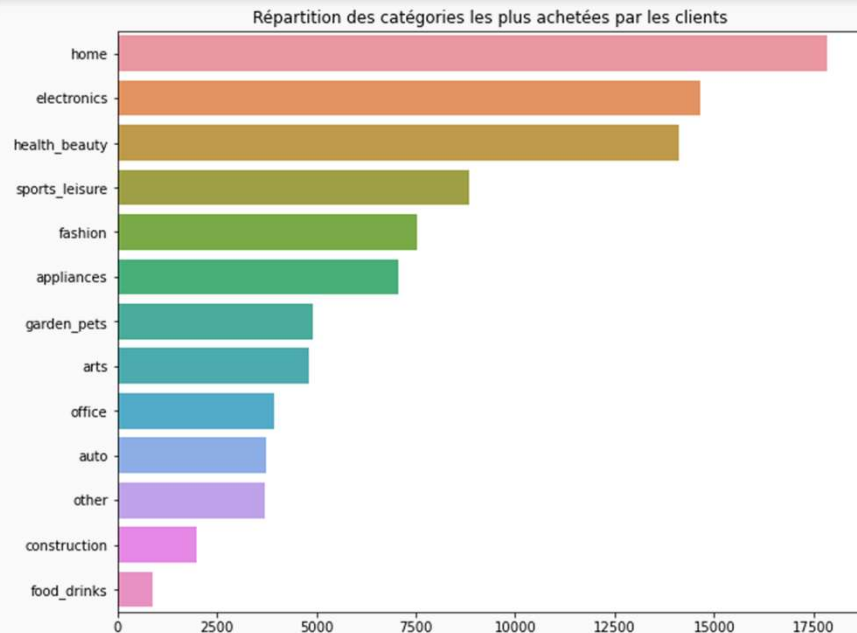
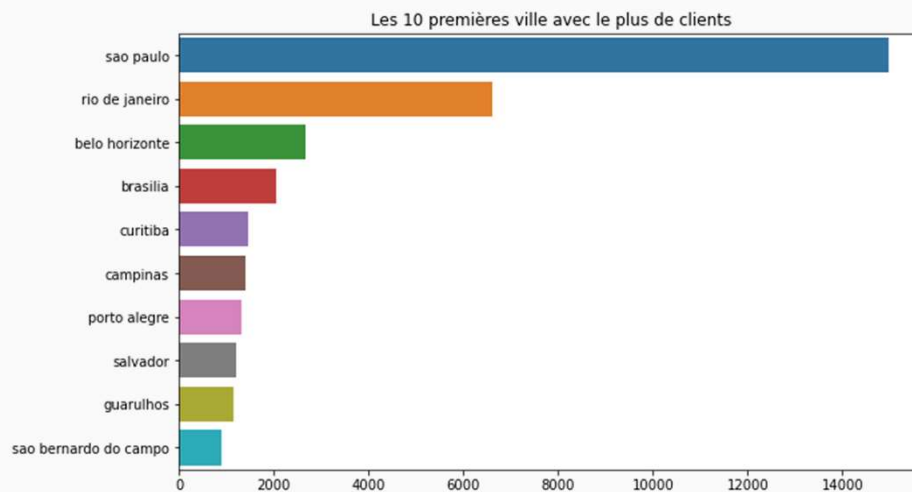
**Avant**



**Après**

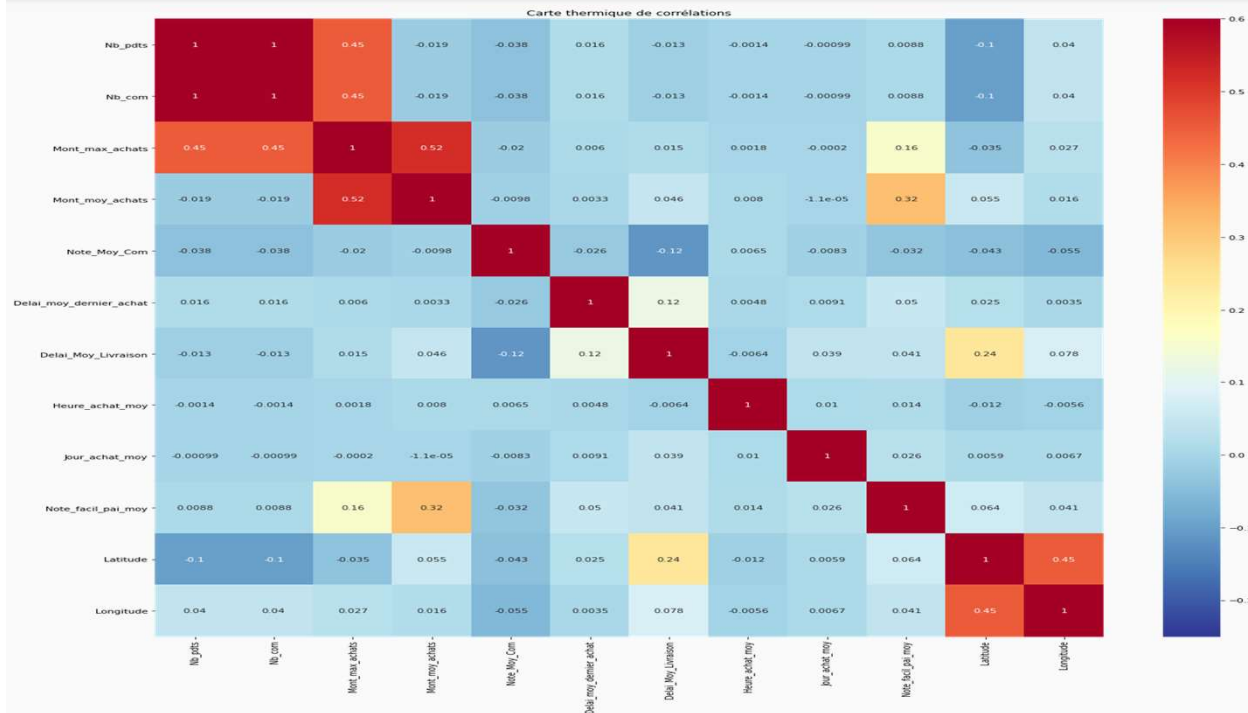


**PRESENTATION DU CLEANING EFFECTUÈ :**  
**7/ Exploration du jeu de données**  
**3/ Exemple de distribution 10 villes avec le plus de clients et la répartition des catégories les plus achetées par les clients**



## PRESENTATION DU CLEANING EFFECTUÈ :

### 8/ Carte thermique des corrélations



**Très forte corrélation:**

➤ Nb\_ppts = Nb\_com

1

**Légende en valeur absolu:**

- ✓ 0.00 à 0.19 "très faible",
- ✓ 0.20 à 0.39 "faible",
- ✓ 0.40 à 0.59 "modéré",
- ✓ 0.60 à 0.79 "fort",
- ✓ 0.80 à 1.0 "très fort"

# PRESENTATION DES DIFFERENTES PISTES DE MODELISATION EFFECTUEES :

## 1/ Segmentation de la clientèle à l'aide de l'analyse RFM

customer_unique_id	Nouveauté	Fréquence	Monétaire
0000366f3b9a7992bf8c76cfd3221e2	160	66	8573.40
0000b849f77a49e4a4ce2b2a4ca5be3f	163	144	2721.60
0000f46a3911fa3c0805444483337064	585	117	8073.00
0000f6ccb0745a6a4b88665a16c9f078	369	32	831.68
0004aac84e0df4da2b147fca70cf8255	336	133	23940.00

```
[17] quantiles = RFM_table.quantile(q=[0.2,0.4,0.6,0.8])  
quantiles
```

	Nouveauté	Fréquence	Monétaire
0.2	142.0	47.0	2729.09
0.4	227.0	89.0	6249.30
0.6	317.0	146.0	12448.34
0.8	433.0	276.0	27636.00

Conversion de quantiles en dictionnaire, plus facile à utiliser

```
[ ] quantiles = quantiles.to_dict()
```

## PRESENTATION DES DIFFERENTES PISTES DE MODELISATION EFFECTUEES :

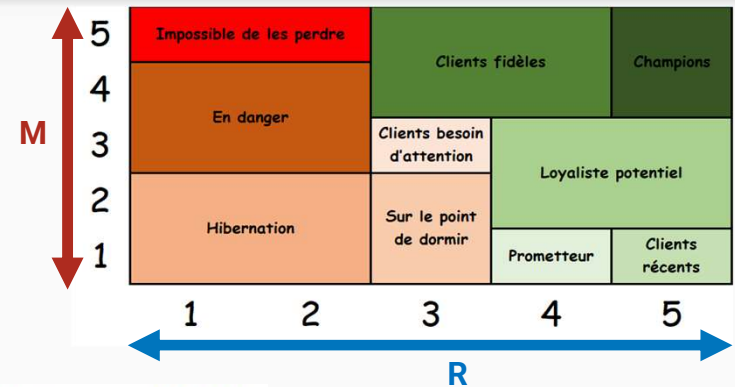
### 1/ Segmentation de la clientèle à l'aide de l'analyse RFM

	customer_unique_id	Nouveauté	Fréquence	Monétaire	R	F	M	RFMScore
0	0000366f3b9a7992bf8c76cfd3221e2	160	66	8573.40	4	2	3	423
1	0000b849f77a49e4a4ce2b2a4ca5be3f	163	144	2721.60	4	3	1	431
2	0000f46a3911fa3c0805444483337064	585	117	8073.00	1	3	3	133
3	0000f6ccb0745a6a4b88665a16c9f078	369	32	831.68	2	1	1	211
4	0004aac84e0df4da2b147fca70cf8255	336	133	23940.00	2	3	4	234



# PRESENTATION DES DIFFERENTES PISTES DE MODELISATION EFFECTUEES :

## 1/ Segmentation de la clientèle à l'aide de l'analyse RFM



	customer_unique_id	Nouveauté	Fréquence	Monétaire	R	F	M	RFMScore	Segment
0	0000366f3b9a7992bf8c76cfd3221e2	160	66	8573.40	4	2	3	423	loyaliste potentiel
1	0000b849f77a49e4a4ce2b2a4ca5be3f	163	144	2721.60	4	3	1	431	prometteur
2	0000f46a3911fa3c0805444483337064	585	117	8073.00	1	3	3	133	en danger
3	0000f6ccb0745a6a4b88665a16c9f078	369	32	831.68	2	1	1	211	hibernation
4	0004aac84e0df4da2b147fca70cf8255	336	133	23940.00	2	3	4	234	en danger

# PRESENTATION DES DIFFERENTES PISTES DE MODELISATION EFFECTUEES :

## 1/ Segmentation de la clientèle à l'aide de l'analyse RFM

Ségmentation client	Clients		Produits vendu		Catégorie de produits le plus acheté	Ville la plus représentée	Type de paiement le plus représenté	Montant moyen par achat	Fréquence d'achat moyen par client	Délais moyen de livraison	Jour d'achat en moyenne des clients	Heure d'achat en moyenne	Délais entre 2 achats en moyenne par client	Note satisfaction moyenne par commande		Note facilité de paiement moyen par commande		Chiffre d'affaire	
	Nombre de clients	% effectifs totale	Nombre de produit	% nbres Totale										Note	Ratio	Note	Ratio	Chiffre d'affaire par client	Pourcentage en fonction du chiffre d'affaire
Champions	7624	7,9%	8022	20,3%	health_beauty	rio de janeiro	credit_card	210.87€	1,052203568	8 jours	Mercredi	15h	99 jours	4,23	84,6%	3,5	14,6%	2 556 120 €	21,7%
Clients fidèles	15408	16,0%	10752	27,3%	home	rio de janeiro	credit_card	201.31€	0,697819315	11 jours	Mercredi	15h	229 jours	3,92	78,4%	3,42	14,3%	2 501 664 €	21,2%
Loyaliste potentiel	15586	16,2%	1928	4,9%	home	sao paulo	credit_card	84.14€	0,123700757	9 jours	Mercredi	15h	140 jours	4,15	83,0%	2,5	10,4%	12 448 €	0,1%
Clients récents	3869	4,0%	694	1,8%	electronics	sao paulo	credit_card	50.62€	0,179374515	8 jours	Mercredi	15h	93 jours	4,27	85,4%	1,99	8,3%	2 728 €	0,0%
Prometteur	3694	3,8%	518	1,3%	electronics	sao paulo	credit_card	50.38€	0,140227396	10 jours	Mercredi	15h	186 jours	4,02	80,4%	2,02	8,4%	2 729 €	0,0%
Clients ayant besoin d'attention	3880	4,0%	1460	3,7%	home	sao paulo	credit_card	96.83€	0,37628866	12 jours	Mercredi	15h	271 jours	3,92	78,4%	2,67	11,1%	12 448 €	0,1%
Sur le point de dormir	7605	7,9%	733	1,9%	electronics	sao paulo	credit_card	58.85€	0,096383958	12 jours	Mercredi	15h	270 jours	4,01	80,2%	2,13	8,9%	6 249 €	0,1%
En danger	15162	15,8%	2000	5,1%	home	sao paulo	credit_card	114.27€	0,131908719	11 jours	Mercredi	15h	445 jours	4,07	81,4%	3,15	13,1%	27 636 €	0,2%
Impossible de les perdre	7707	8,0%	11427	29,0%	home	rio de janeiro	credit_card	286.92€	1,482678085	11 jours	Mercredi	15h	445 jours	4,05	81,0%	4,27	17,8%	6 662 844 €	56,5%
Hibernation	15561	16,2%	1890	4,8%	home	sao paulo	credit_card	60.97€	0,12145749	11 jours	Mercredi	15h	446 jours	4,06	81,2%	2,42	10,1%	6 249 €	0,1%
								Nombre Totale de clients											Chiffre d'affaire Totale
								96096											11 791 115 €
									Nombre Totale de produit										
									39424										
									Fréquence d'achat totale moyen par client										
									0,41025641										

# PRESENTATION DES DIFFERENTES PISTES DE MODELISATION EFFECTUEES :

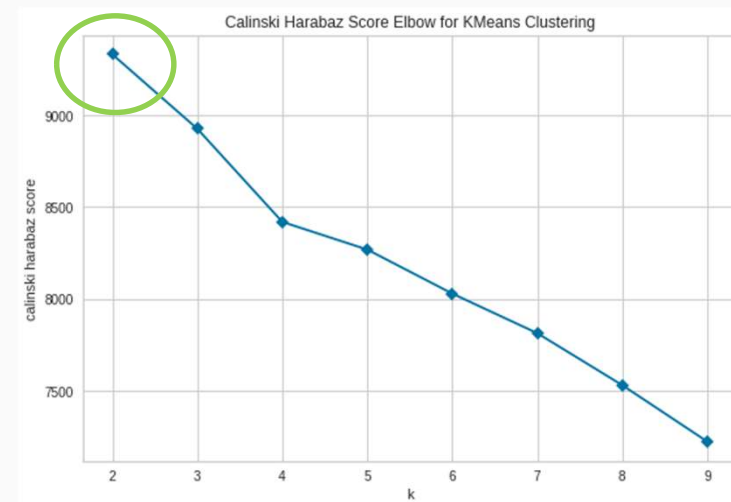
## 2/ Comparer 2 modèles

### 1/ KMeans

#### Préparation du jeu de données :

- ✓ Extraction des variables numériques,
- ✓ Extraction des variables catégoriques,
- ✓ Standardiser les données numériques,
- ✓ Joindre les données numérique au données catégorique
- ✓ Encodage à chaud des données catégoriques

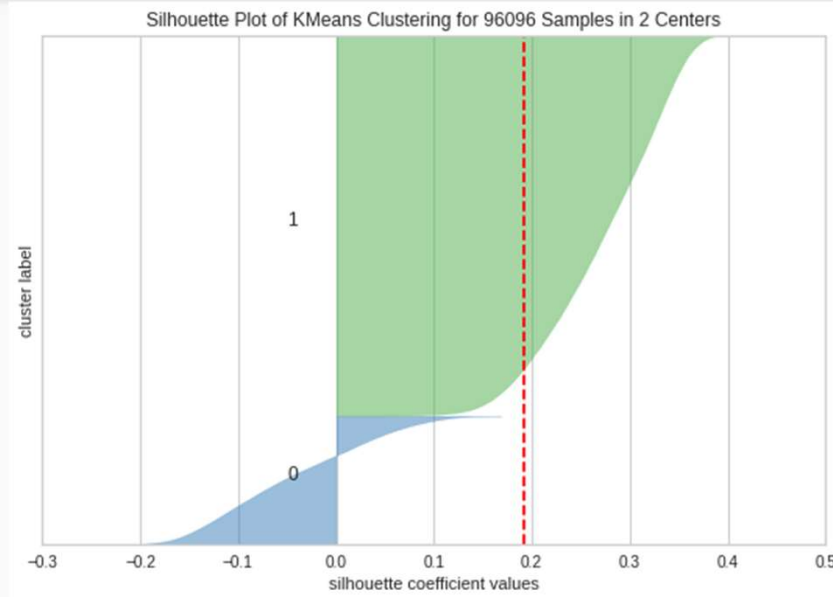
#### La technique du coude



# PRESENTATION DES DIFFERENTES PISTES DE MODELISATION EFFECTUEES :

## 2/ Comparer 2 modèles

### 1/ KMeans



**`silhouette_score(Kmeans) = 0.3395096597043642`**

# PRESENTATION DES DIFFERENTES PISTES DE MODELISATION EFFECTUEES :

## 2/ Comparer 2 modèles

### 2/ CAH(Classification ascendante hiérarchique)

#### Préparation du jeu de données :

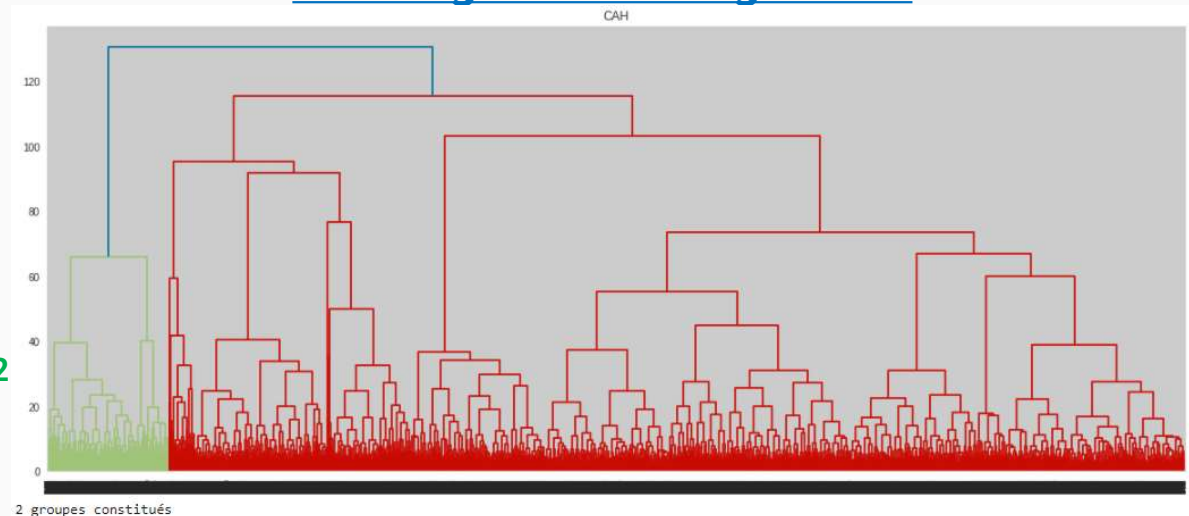
- ✓ Réduction du jeu de données trop volumineux
- ✓ Choix de 10 000 clients choisi de manière aléatoire

$\text{silhouette\_score}(\text{CAH}) = 0.31673381319636074$

<

$\text{silhouette\_score}(\text{KMeans}) = 0.3395096597043642$

#### Affichage du dendrogramme

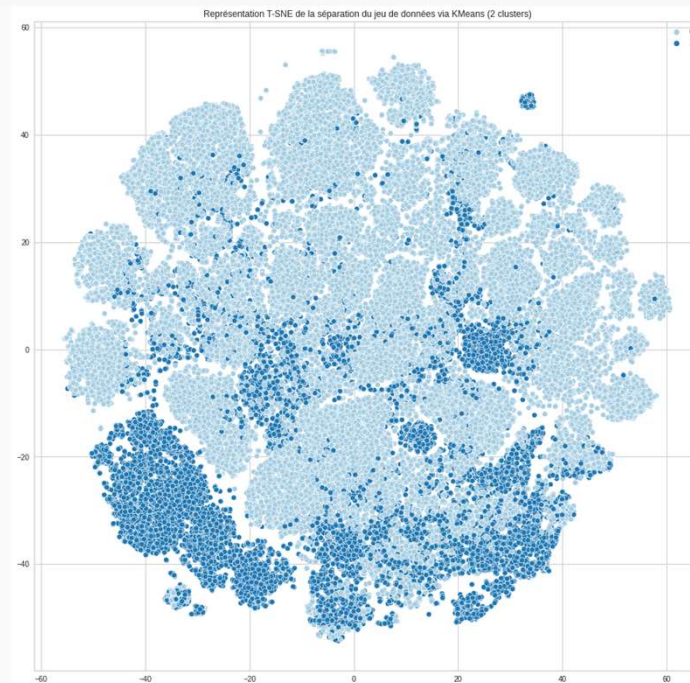


# PRESENTATION DU MODELE FINAL :

## 1/ Le modèle Kmeans

N° Clusters	Clients		Produits vendu		Comportement client le plus représenté	Catégorie de produits le plus acheté	Ville la plus représentée	Type de paiement le plus représenté	Montant moyen par achat	Fréquence d'achat moyen par client	Délais moyen de livraison	Jour d'achat en moyenne des clients	Heure d'achat en moyenne	Délais entre 2 achats en moyenne par client	Note satisfaction moyenne par commande		Note facilité de paiement moyen par commande		Chiffre d'affaire	
	Nombre de clients	% effectifs totale	Nombre de produit	% nbres Totale											Note	Ratio	Note	Ratio	Chiffre d'affaire par clusters	Pourcentage en fonction du chiffre d'affaire
0	71796	74,7%	5112	30,9%	loyaliste potentiel	home	sao paulo	credit_card	92,26 €	0,071201738	9,01 jours	Entre Mardi et Mercredi	14h47min24s	281,50 jours	4,21	84,2%	2,3	9,6%	382 064 €	5,4%
1	24300	25,3%	11427	69,1%	clients fidèles	home	rio de janeiro	credit_card	226,14 €	0,470246914	14,69 jours	Entre Mardi et Mercredi	14h45min36s	311,52 jours	3,66	73,2%	4,7	19,6%	6 662 844 €	94,6%
	Nombre Totale de clients		Nombre Totale de produit							Fréquence d'achat totale moyen par client									Chiffre d'affaire Totale	
	96096		16539							0,172109141									7 044 908 €	

## PRESENTATION DU MODELE FINAL : 2/ T-SNE



**PRESENTATION DU MODELE FINAL :**  
**2/ Déterminer la fréquence de mise à jours du jeu de données**  
**1/ Préparation de mes jeux de données**

**Préparation des jeux de données :**

- ✓ Déterminer les plages de date pour chaque fréquence de mise à jours,
- ✓ Extraction des variables numériques,
- ✓ Extraction des variables catégoriques,
- ✓ Standardiser les données numériques,
- ✓ Joindre les données numérique au données catégorique
- ✓ Encodage à chaud des données catégoriques

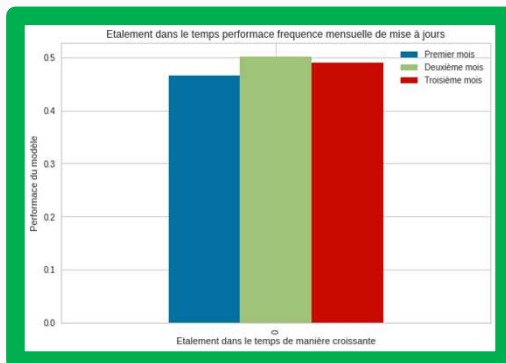


# PRESENTATION DU MODELE FINAL :

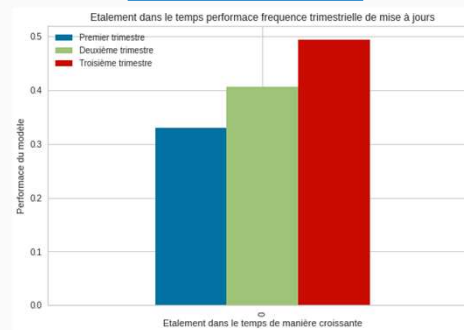
## 2/ Déterminer la fréquence de mise à jours du jeu de données

## 2/ Affichage des résultats des fréquences de mise à jours

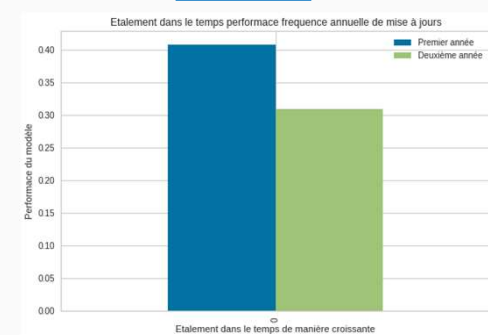
### MENSUELLE



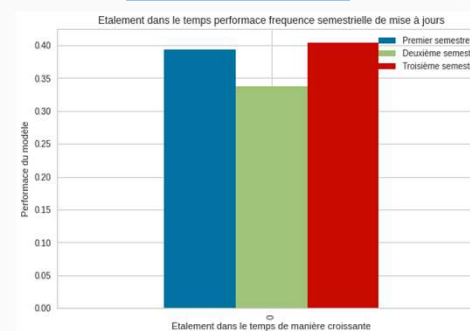
### TRIMESTRIELLE



### ANUELLE



### SEMESTRIELLE



## CONCLUSION

- ✓ 3 types de comportements champions, clients fidèles et impossibles de les perdre représentent à eux 3 près de 100% du chiffre d'affaire
- ✓ ¼ des clients réalisent à eux seuls pas loin de 100% du chiffre d'affaire.
- ✓ Devis de contrat maintenance (Mise à jours tous les mois)
- ✓ Piste d'amélioration

# REMERCIEMENT

Merci de m'avoir écouter

## REPONDRE AUX QUESTIONS