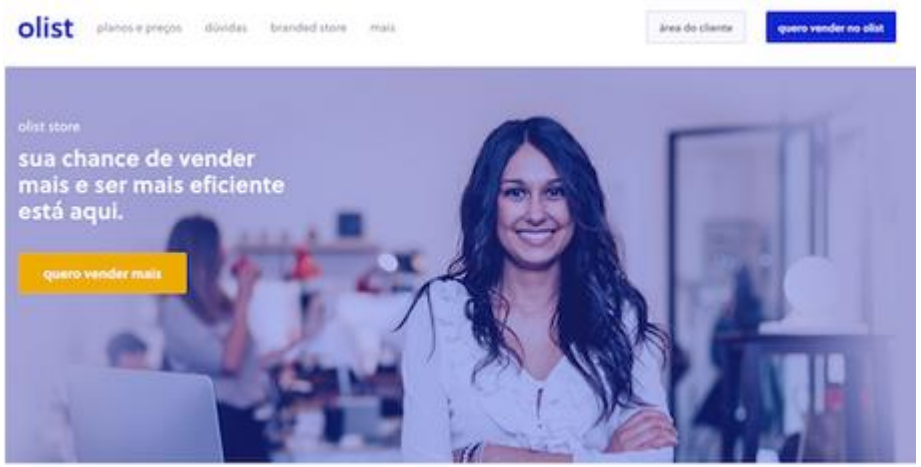


#5 Segmentez des clients d'un site e-commerce

Soutenance Emilie Groschêne le 20/01/2023
Evaluateur: Zied Jemai
Mentor: Lea Naccache

Sommaire

I. PRESENTATION DE LA PROBLEMATIQUE

I. Présentation de la problématique

olist est une entreprise brésilienne permettant de vendre sur les marketplaces en ligne. Elle souhaite **segmenter ses clients** et rendre ainsi plus **efficace** ses **campagnes de communication**.

Objectifs:

- Comprendre les différents **types d'utilisateurs** grâce à leur comportement et à leurs données personnelles

Mission:

- Extraire les données permettant de **caractériser les clients**
- Mettre en place un modèle **d'apprentissage non supervisé** pour segmenter les clients
- Rendre interprétables les segments d'un **point de vue métier**
- Proposer un **contrat de maintenance** basé sur une analyse de la stabilité des segments au cours du temps

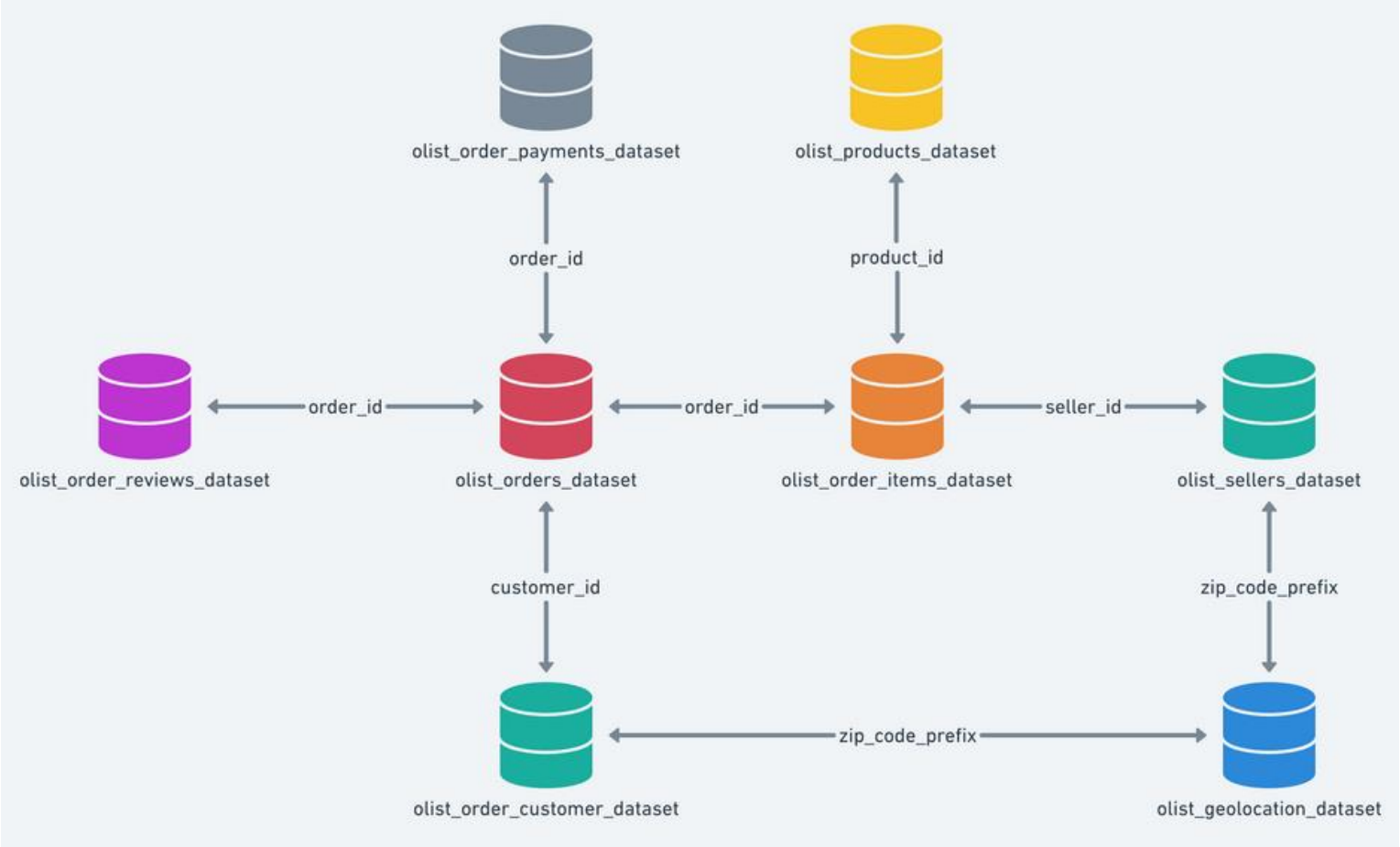
Cahier des charges:

- Le code doit respecter le format **PEP8**

II. PREPARATION DU JEU DE DONNEES ET EXPLORATION

II. Préparation du jeu de données – Architecture

- Base de données **anonymisée**
- **8 datasets** reliés entre eux par différentes **clés** (order_id, product_id etc)
- Contenant des informations sur les **clients**, la **localisation**, les **commandes** et **produits** achetés, les **options de paiement**, les **avis** clients et les **vendeurs**
- Un dataset supplémentaire permettant de **traduire** les catégories de produits en anglais



II. Préparation du jeu de données – Les données

olist_products_dataset.csv			
Variables	**Nb_lignes**	**% distinct**	**% NaN**
product_id	32 951	100	0
product_category_name		0,22	1,85
product_name_lenght		0,2	1,85
product_description_lenght		8,98	1,85
product_photos_qty		0,06	1,85
product_weight_g		6,69	0,01
product_length_cm		0,3	0,01
product_height_cm		0,31	0,01
product_width_cm		0,29	0,01

olist_order_payments_dataset.csv			
Variables	**Nb_lignes**	**% distinct**	**% NaN**
order_id	103 886	95,72	0
payment_sequential		0,03	0
payment_type		0	0
payment_installments		0,02	0
payment_value		27,99	0

olist_order_reviews_dataset.csv			
Variables	**Nb_lignes**	**% distinct**	**% NaN**
review_id	99 224	99,18	0
order_id		99,44	0
review_score		0,01	0
review_comment_title		4,56	88,34
review_comment_message		36,44	58,70
review_creation_date		0,64	0
review_answer_timestamp		99,02	0

olist_orders_dataset.csv			
Variables	**Nb_lignes**	**% distinct**	**% NaN**
order_id	99 441	100	0
customer_id		100	0
order_status		0,01	0
order_purchase_timestamp		99,43	0
order_approved_at		91,24	0,16
order_delivered_carrier_date		81,47	1,79
order_delivered_customer_date		96,2	2,98
order_estimated_delivery_date		0,46	0

olist_order_items_dataset.csv			
Variables	**Nb_lignes**	**% distinct**	**% NaN**
order_id	112 650	87,59	0
order_item_id		0,02	0
product_id		29,25	0
seller_id		2,75	0
shipping_limit_date		82,84	0
price		5,3	0
freight_value		6,21	0

olist_sellers_dataset.csv			
Variables	**Nb_lignes**	**% distinct**	**% NaN**
seller_id	3 095	100	0
seller_zip_code_prefix		72,57	0
seller_city		19,74	0
seller_state		0,74	0

olist_customers_dataset.csv			
Variables	**Nb_lignes**	**% distinct**	**% NaN**
customer_id	99 441	100	0
customer_unique_id		97	0
customer_zip_code_prefix		15	0
customer_city		4	0
customer_state		0	0

olist_geolocation_dataset.csv			
Variables	**Nb_lignes**	**% distinct**	**% NaN**
geolocation_zip_code_prefix	1 000 163	2	0
geolocation_lat		72	0
geolocation_lng		72	0
geolocation_city		1	0
geolocation_state		0	0

II. Préparation du jeu de données – Etapes suivies

Consolidation des données

- Jointure gauche

par order_id

- Agrégation au niveau de la **commande**
- Suppression des données manquantes
- Création d'indicateurs par commande

par customer_unique_id

- Agrégation de la table des commandes au niveau du **client unique**
- Création d'indicateurs

- **Montant total dépensé (Monetary)**
- **Montant moyen par commande**
- **Prix moyen des articles commandés**
- **Prix article le moins cher**
- **Prix article le plus cher**
- **Prix moyen du transport**

- **Nombre de commandes (Frequency)**
- **Nombre d'articles moyens / commande**
- **Nombre d'article achetés**
- **Nombre d'articles différents achetés**

- **Délai passé depuis la dernière commande (Recency)**
- **Durée moyenne de livraison**
- **Jours de retard moyens de livraison**
- **Nombre de jours moyens pour rédiger un commentaire**

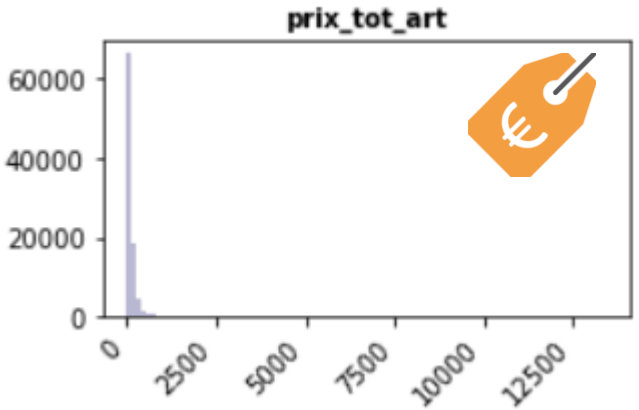
- **Moyen de paiement préféré**

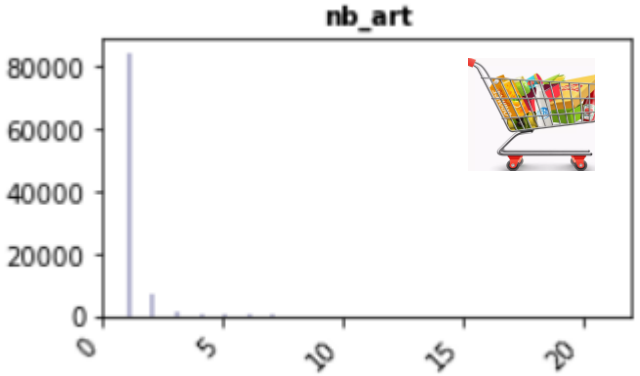
- **Distance moyenne vendeur / client**

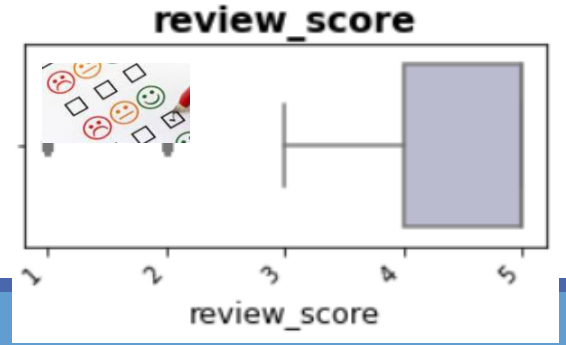
- **Note moyenne des commandes**

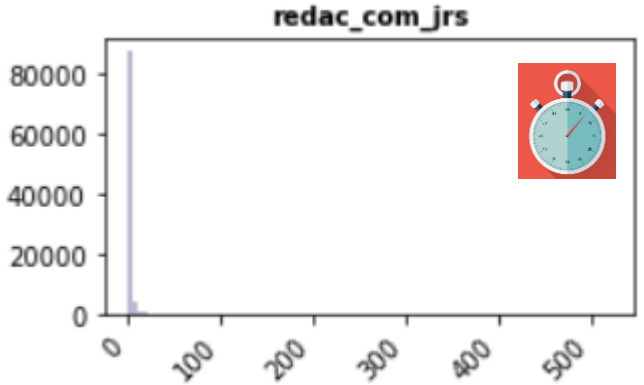
- **Catégorie du produit le plus acheté**

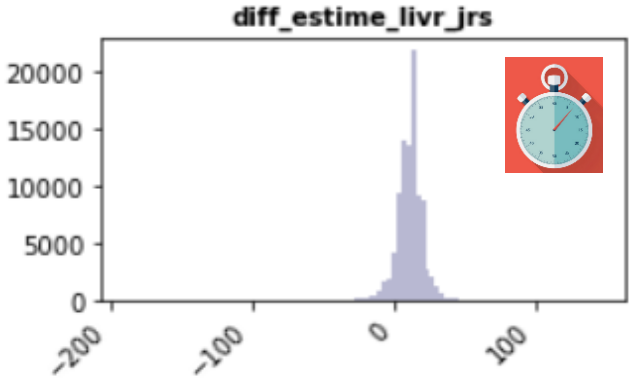
II. Préparation du jeu de données – Exploration (1/4)

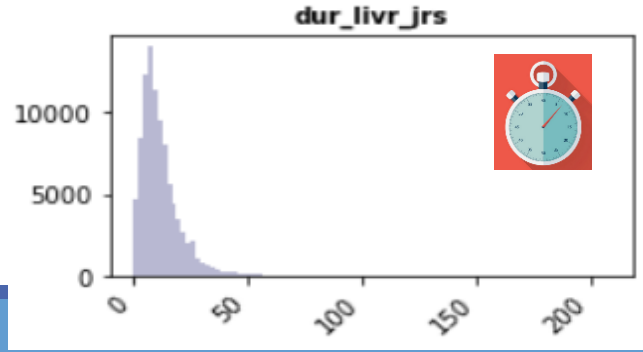


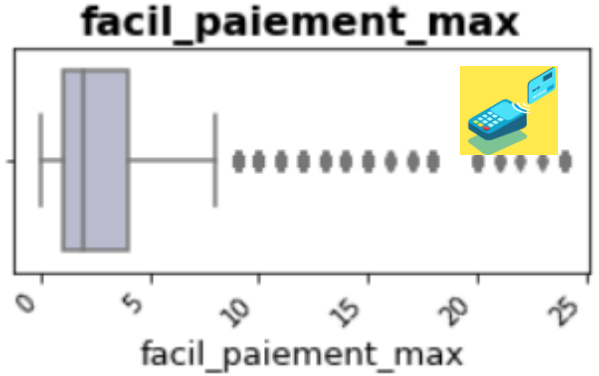






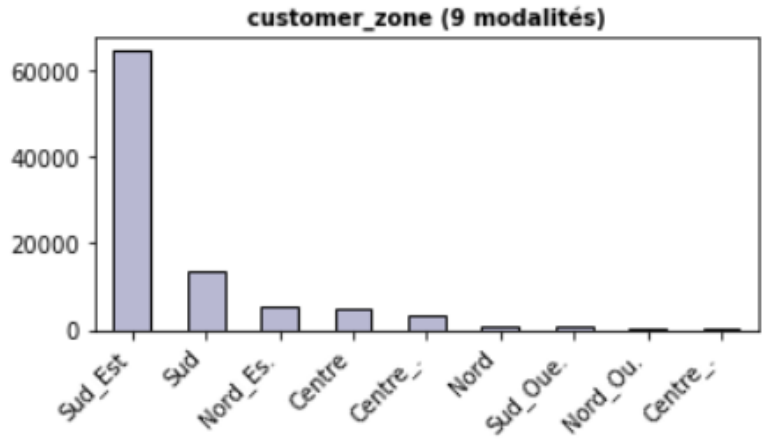


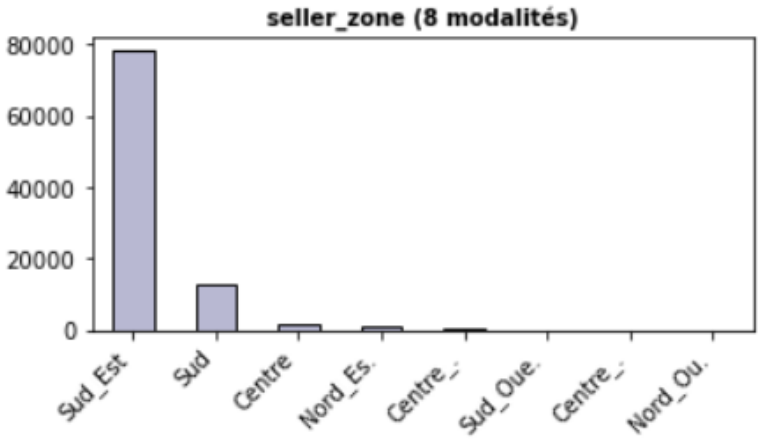


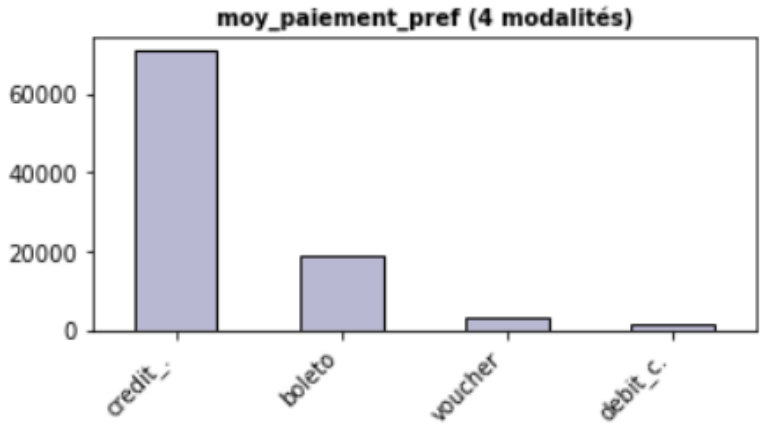


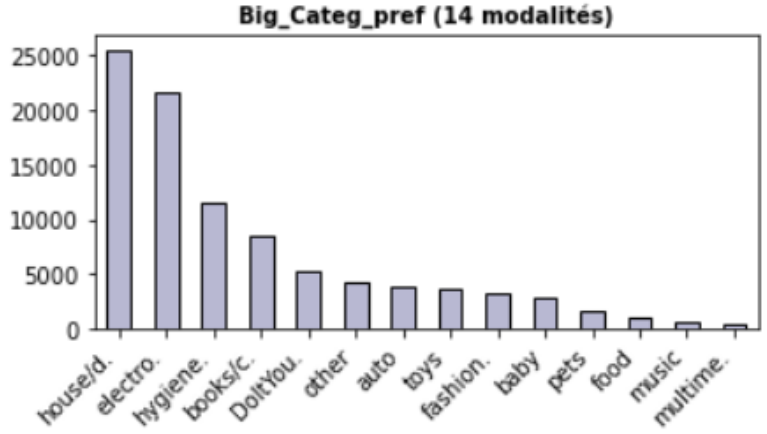
- **Montant moyen** par commande peu élevé de **137 Réals** + 27 Réals de livraison
- **1 article** par commande en moyenne
- Majorité de **clients satisfaits**
- Le client met **2,5 jours** en moyenne par commande pour rédiger un **commentaire**
- **12 jours** de **livraison** en moyenne
- **11 jours de retard** moyen dans les **livraisons** (parfois livraison en avance)
- Le nombre de **facilités de paiement** tourne autour de **3** en moyenne mais peuvent être beaucoup plus important

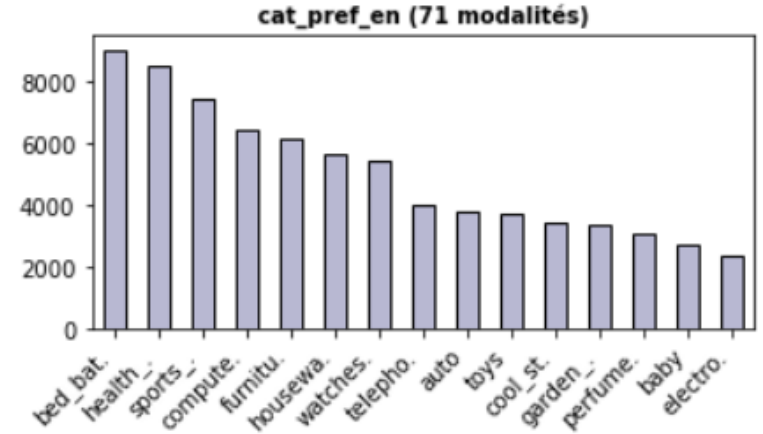
II. Préparation du jeu de données – Exploration (2/4)







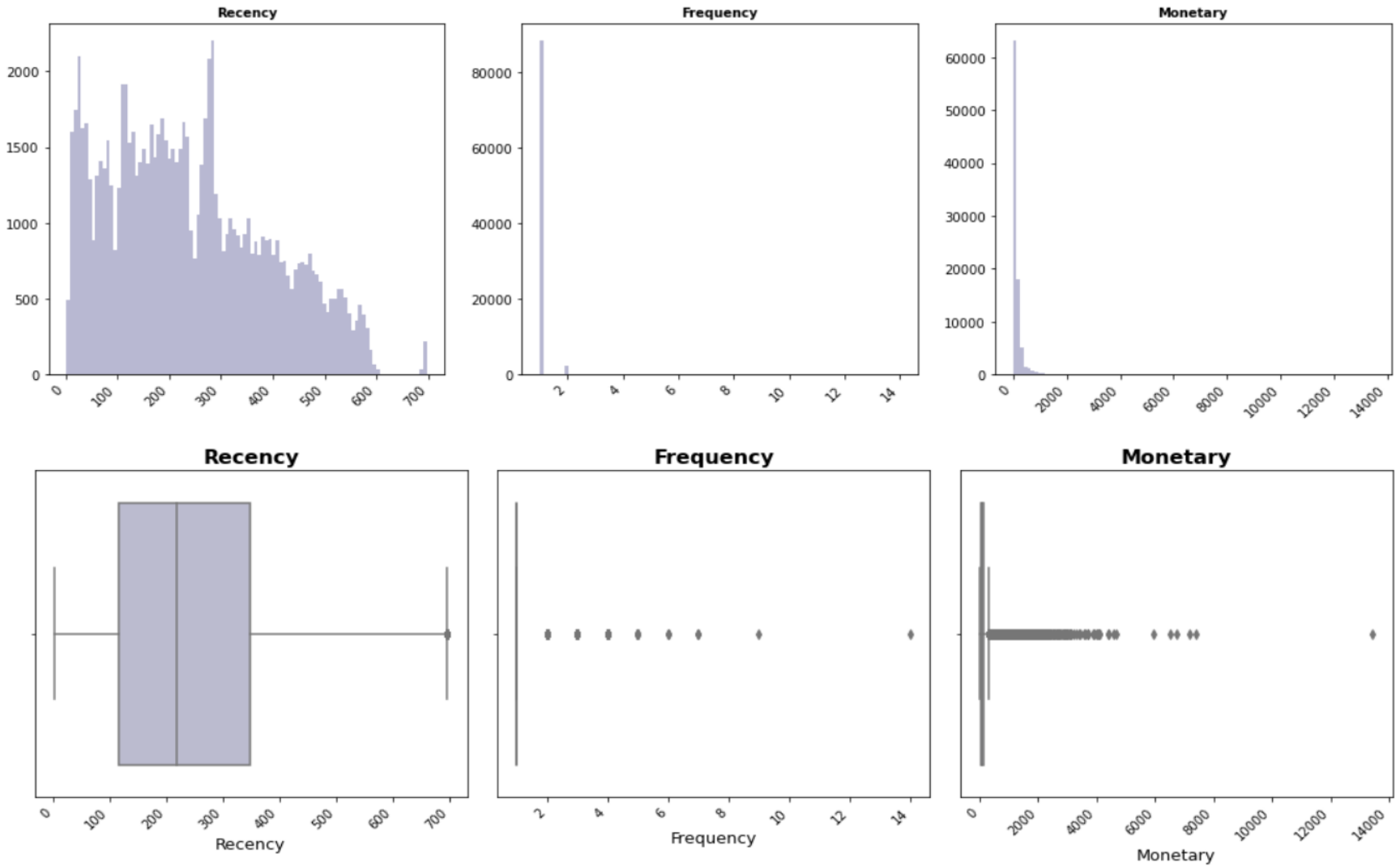




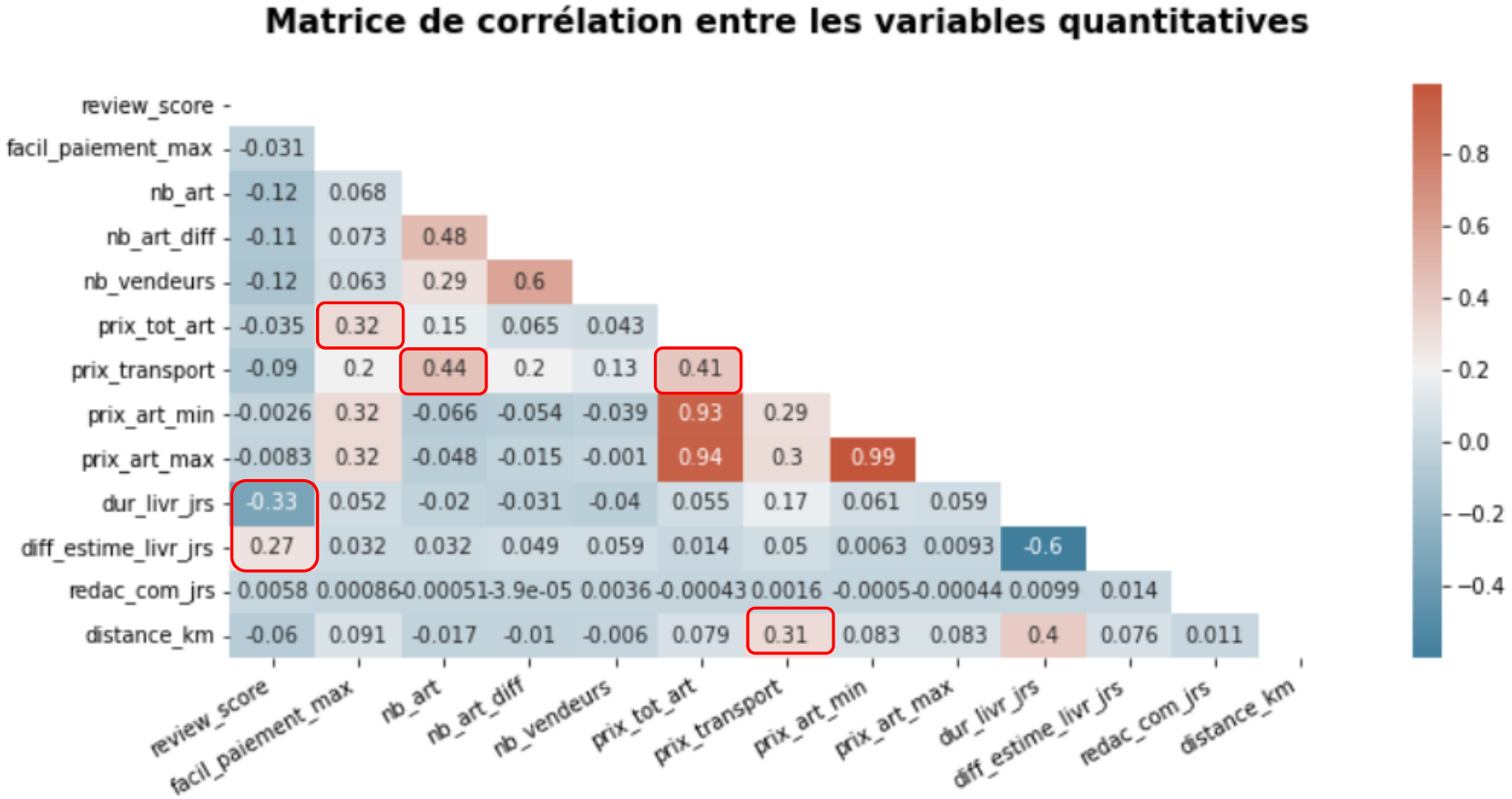
Période analysée:
3/10/2016 au
29/08/2018

Seuls 3% des clients ont passé plusieurs commandes

II. Préparation du jeu de données – Exploration (3/4)



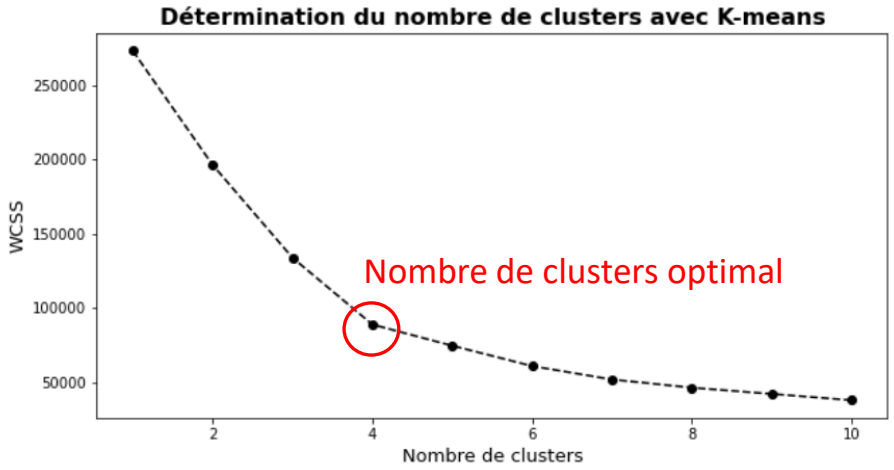
II. Préparation du jeu de données – Exploration (4/4)

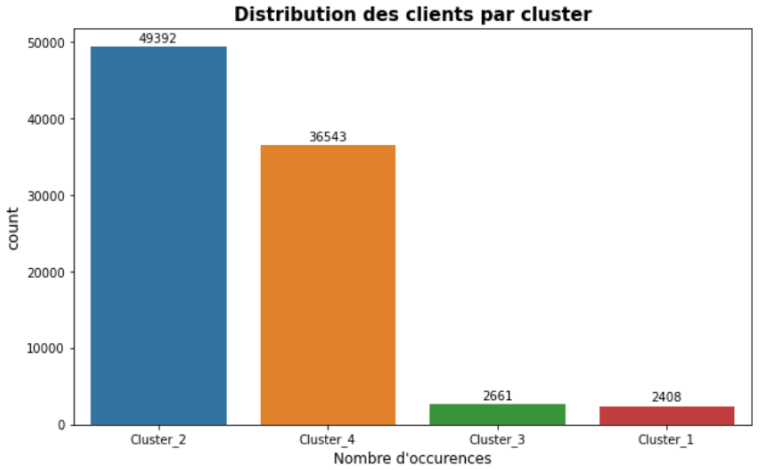


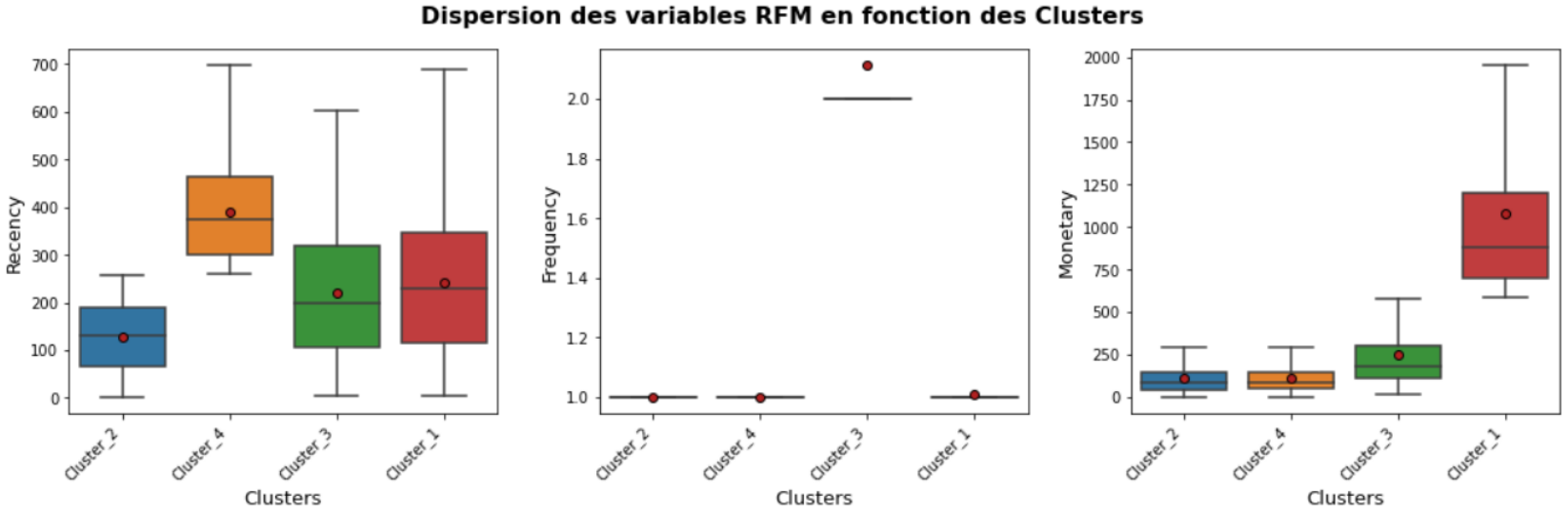
- La **note** associée à la commande semble être surtout liée aux variables en rapport avec la **livraison** de la commande. Les clients attribuent les meilleures notes lorsque les délais de livraison sont plus courts et la date de livraison respectée ou en avance.
- Le **montant** total de la **commande** est corrélé au nombre de **facilités de paiement**
- Le montant du **transport** est lié au **nombre** et **prix** des articles et à la **distance** entre le client et le vendeur

III. PISTES DE MODELISATION

III. Pistes de modélisation – K-means sur indicateurs RFM

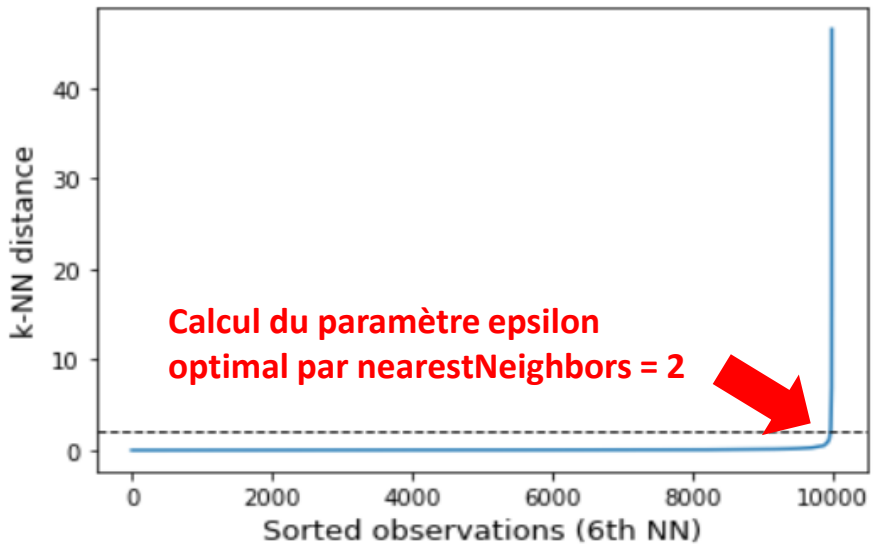


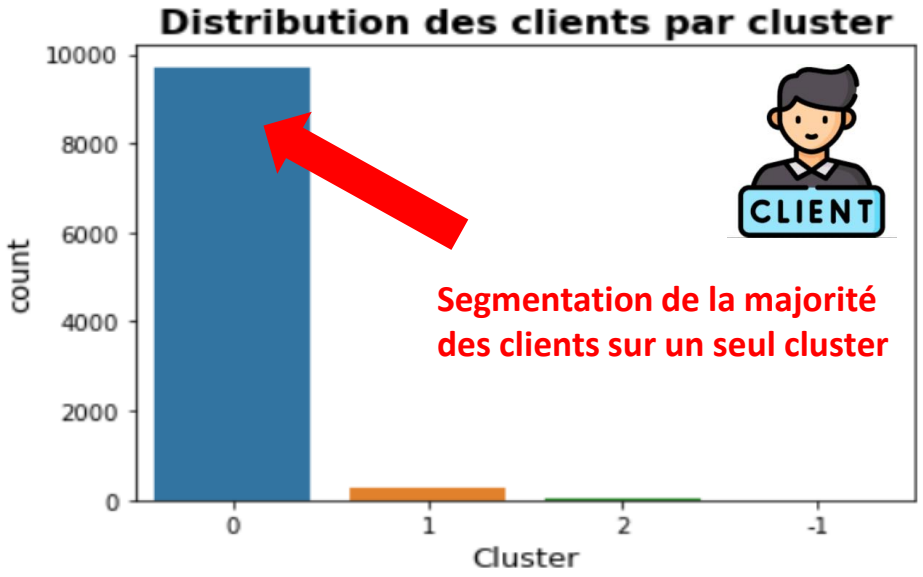


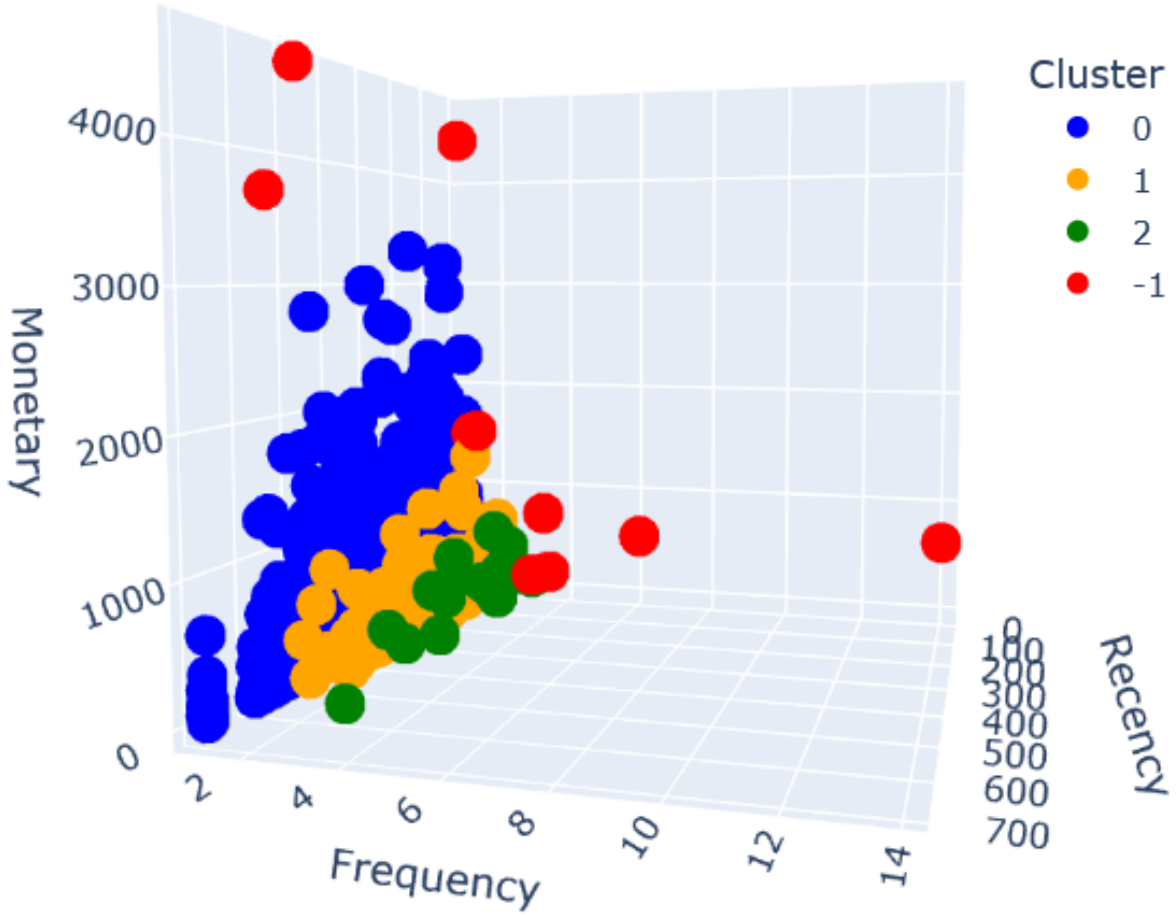


- Le cluster 1 représente les clients qui ont passé **une seule commande avec un montant élevé**
- Le cluster 2 représente les clients qui ont passé **une seule commande de faible montant mais plus récemment**
- Le cluster 3 représente les clients qui ont passé **plusieurs commandes**
- Le cluster 4 représente les clients qui ont passé **une seule commande, de faible montant, il y a longtemps**

III. Pistes de modélisation – DBSCAN sur indicateurs RFM

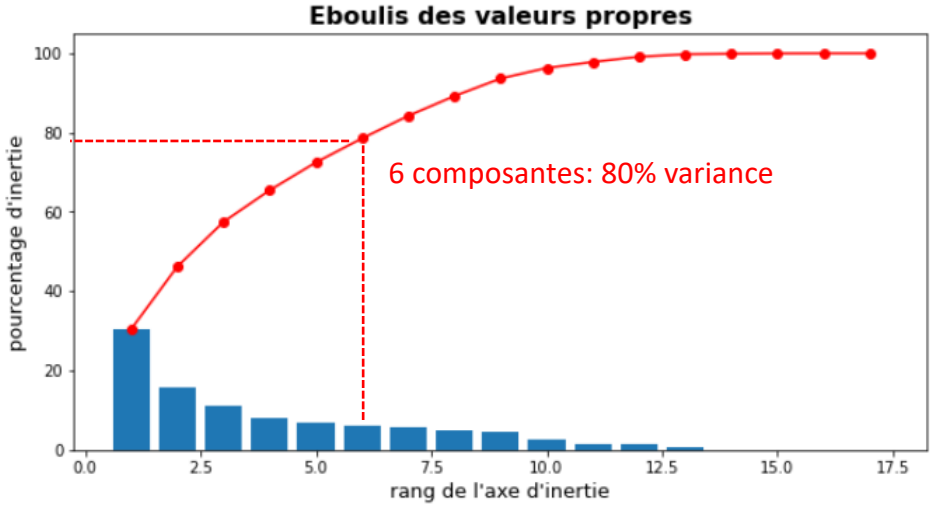


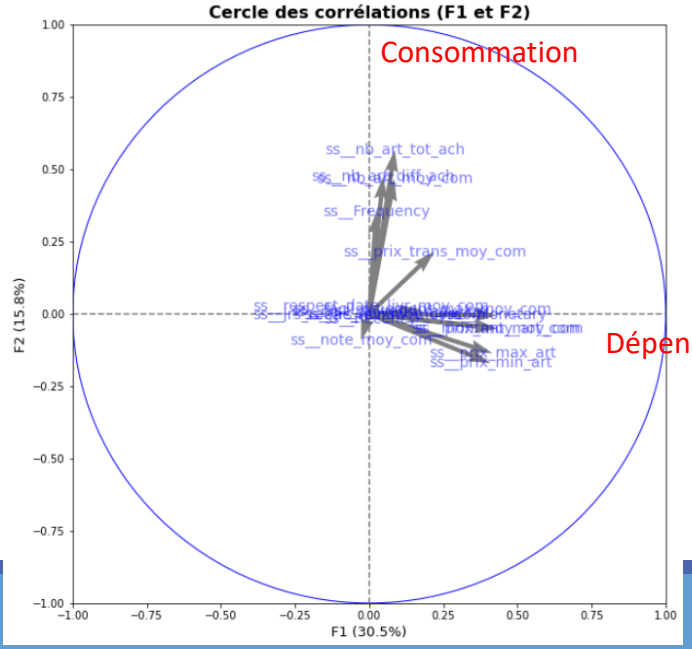


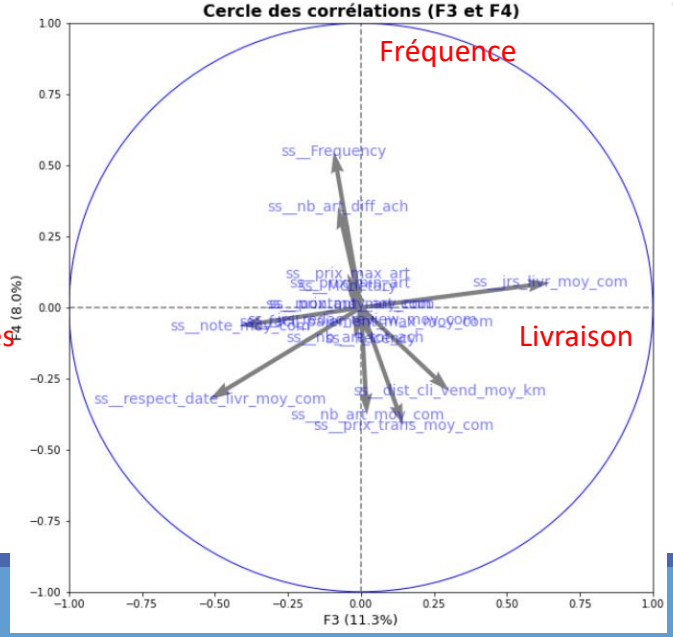


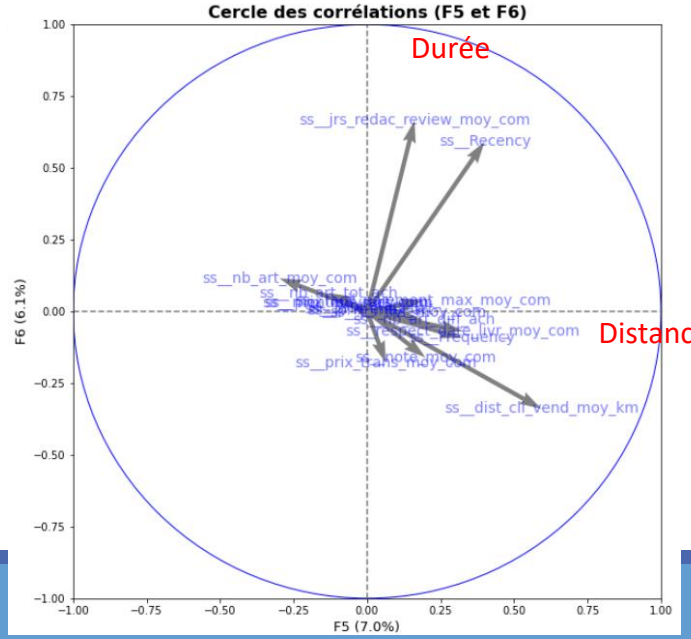

```
Nombre estimé de clusters: 3
Nombre estimé d'outliers: 10
```

III. Pistes de modélisation – K-means sur ACP (1/2)

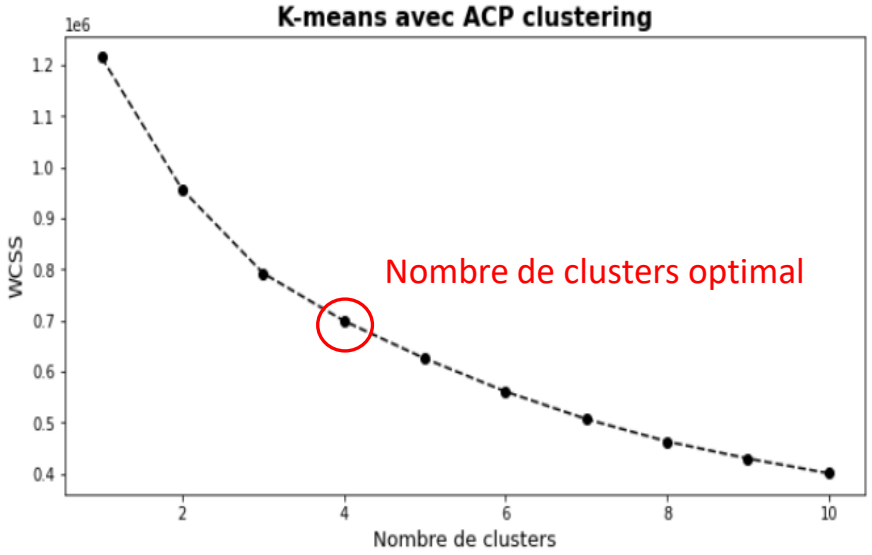


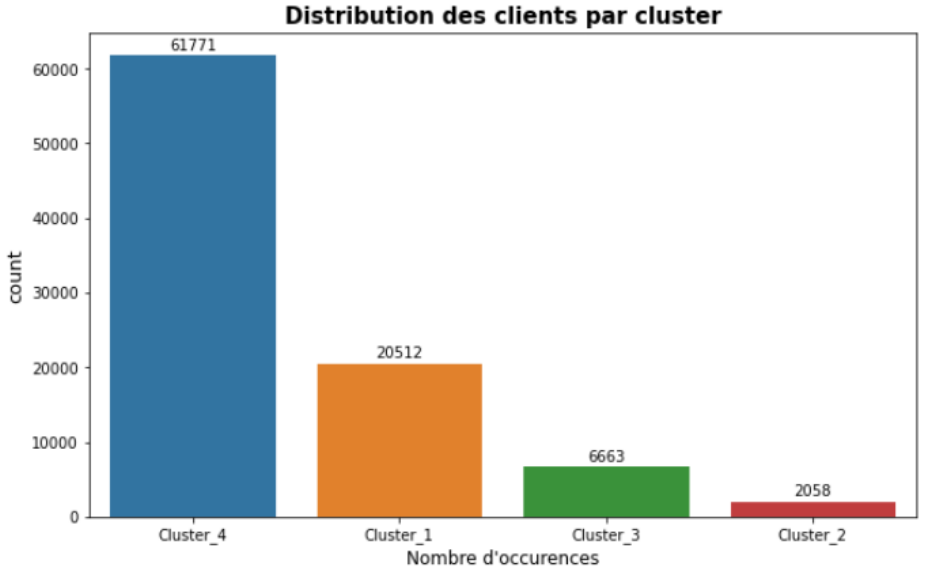


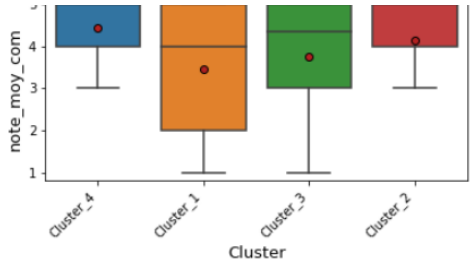


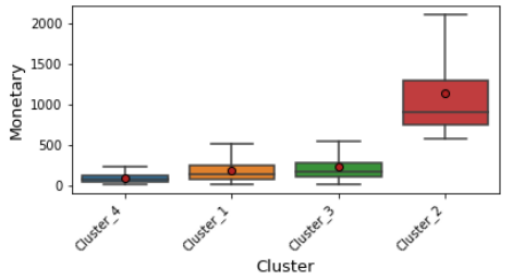


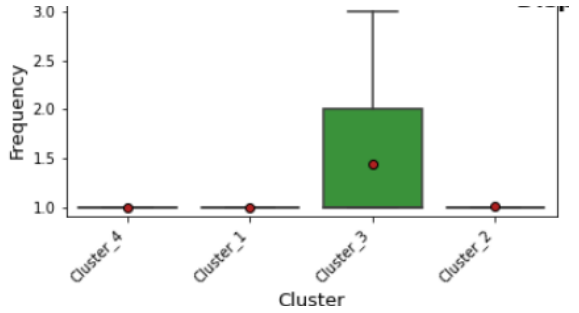
III. Pistes de modélisation – K-means sur ACP (2/2)







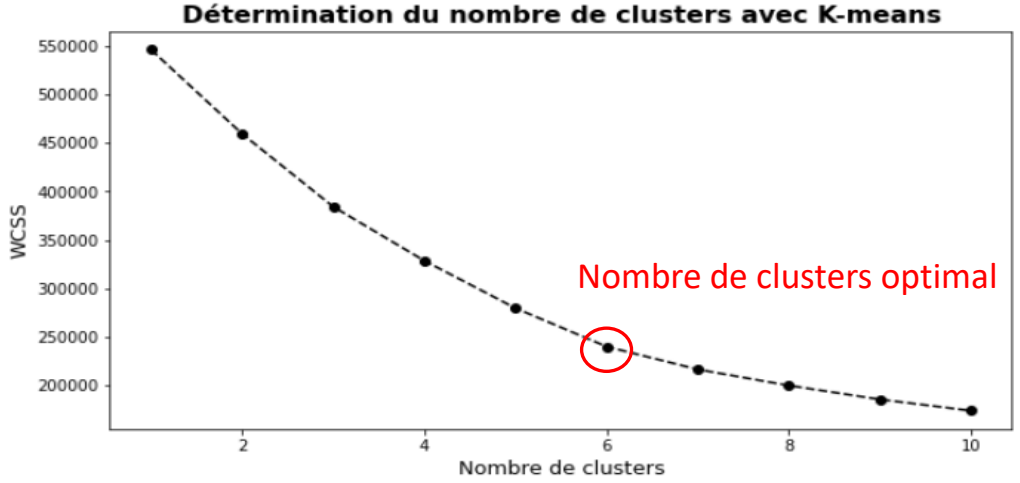


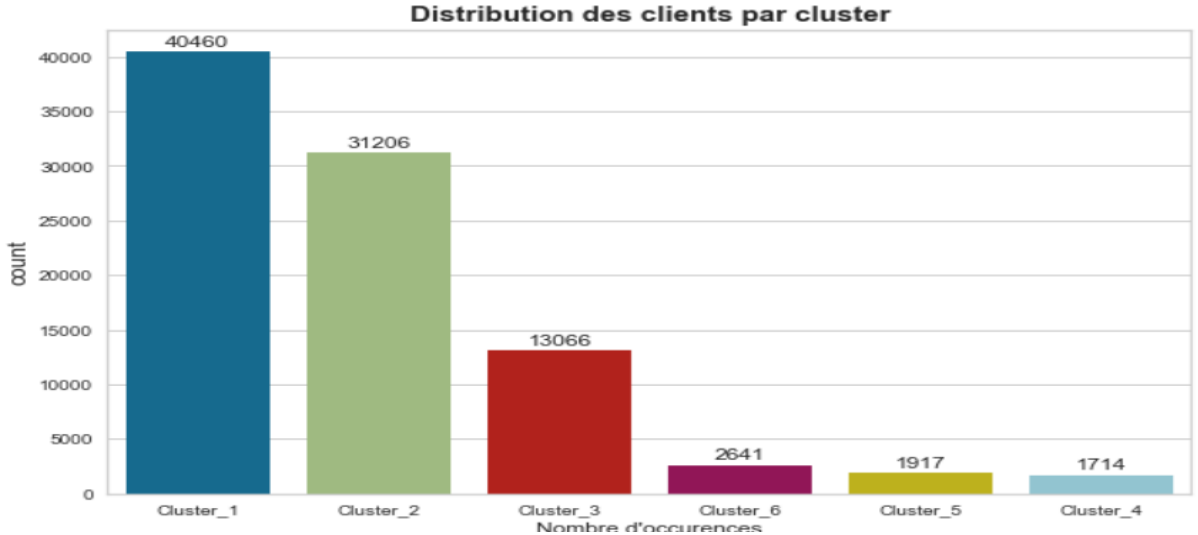


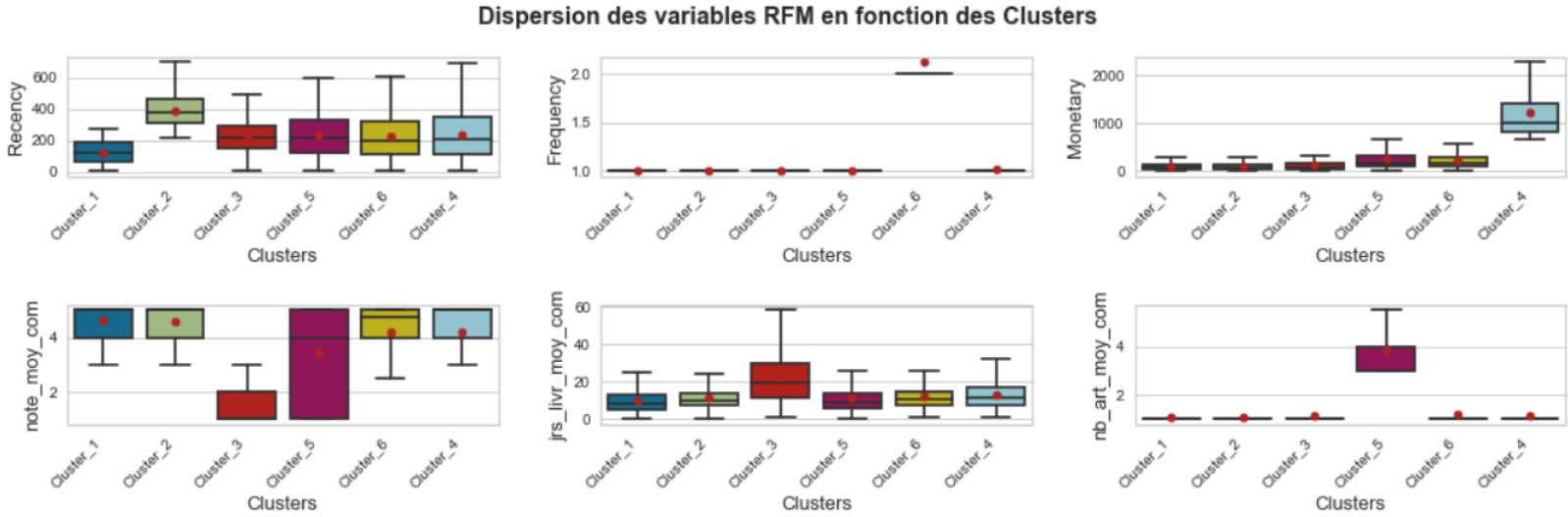
- Le cluster 1 représente les clients les **moins satisfaits** avec un nombre de jours de livraison supérieur aux autres clients et ayant effectué une seule commande
- Le cluster 2 représente les clients qui ont **dépensé le plus** et qui ont effectué une seule commande
- Le cluster 3 représente les clients qui ont passé **plus d'une commande**
- Le cluster 4 représente les clients les plus satisfaits mais ayant effectué **une seule commande**

IV. MODELE FINAL

IV. Modèle final – K-means sur variables les plus différenciantes (1/2)





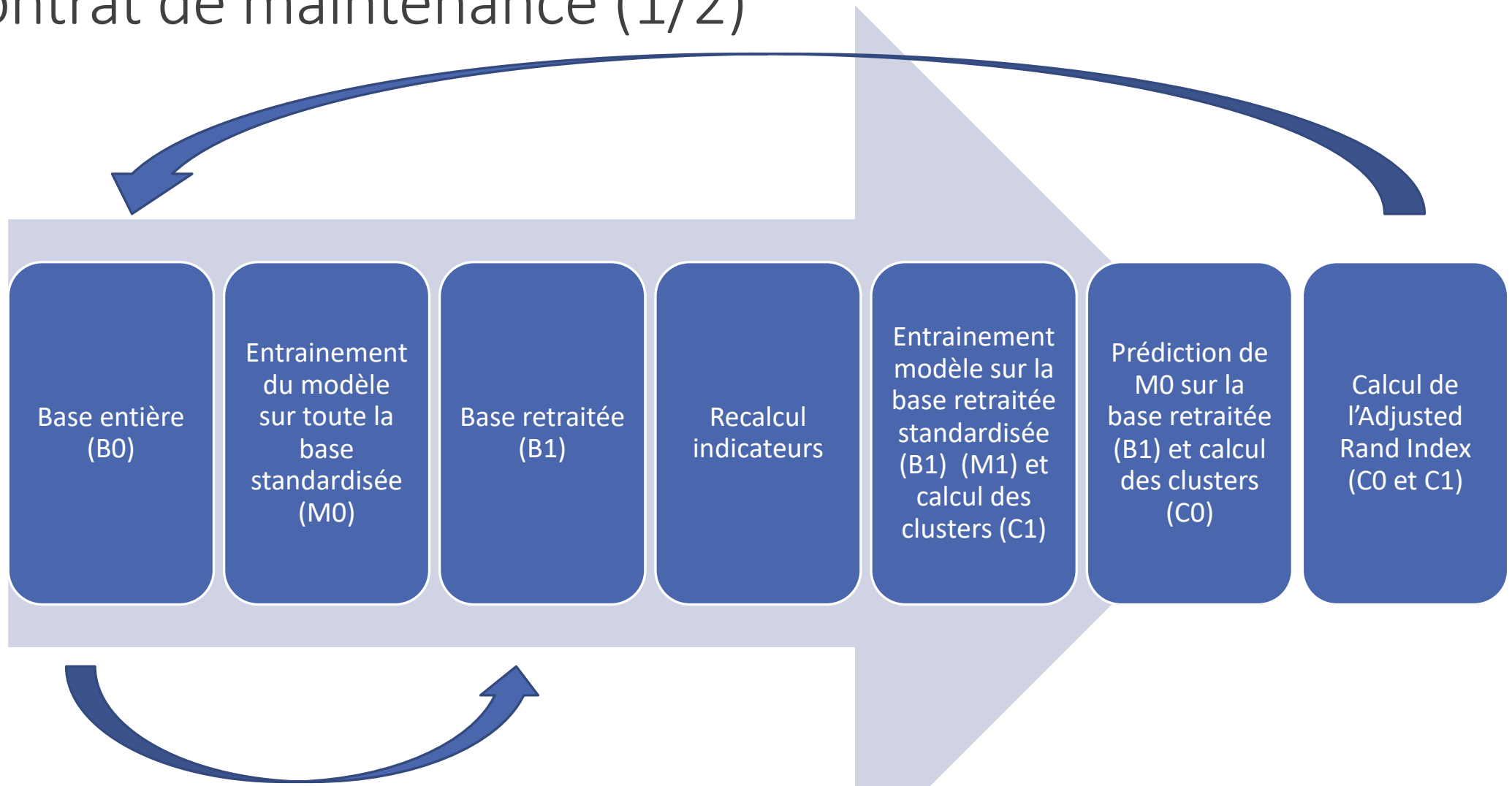


IV. Modèle final – K-means sur variables les plus différenciantes (2/2)

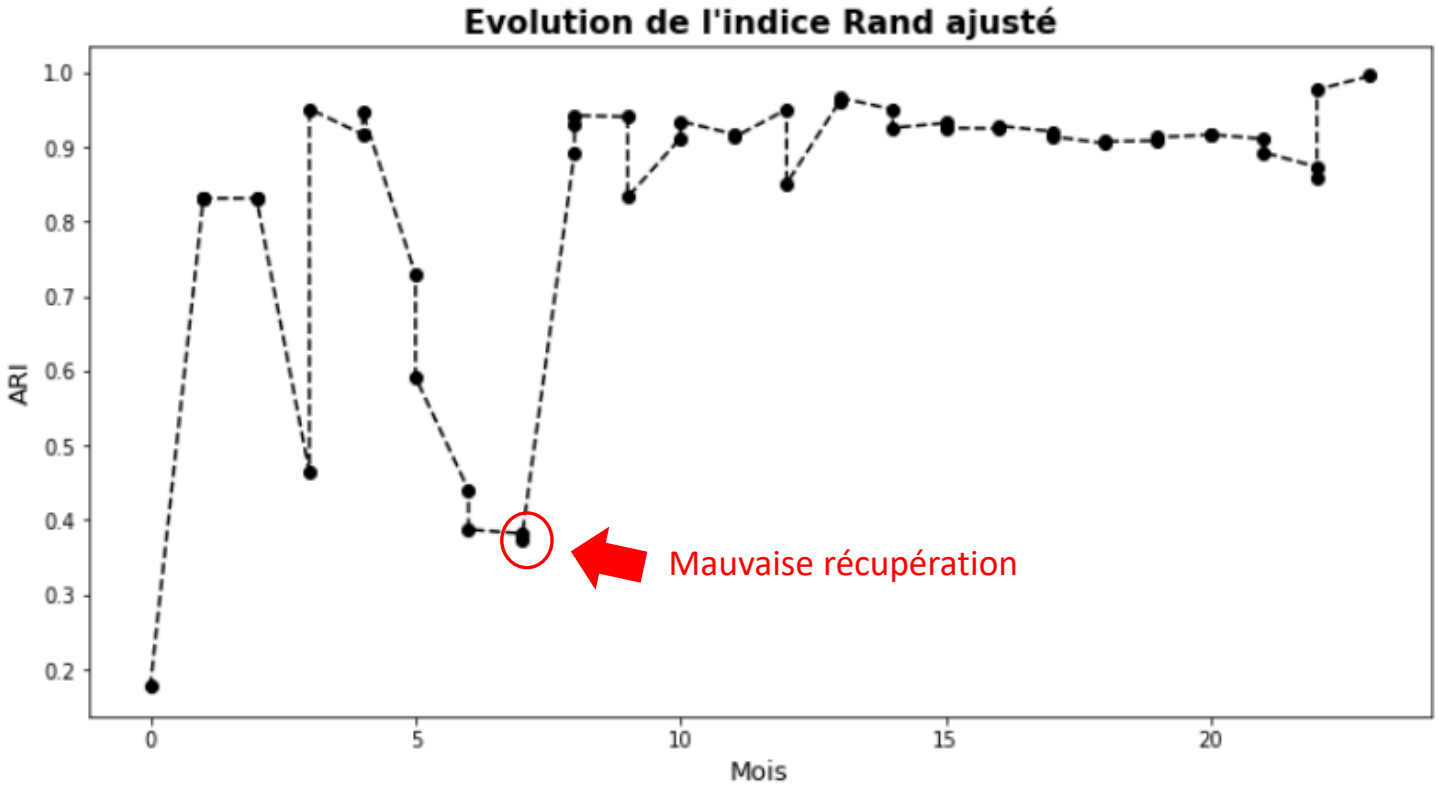
- **Le cluster 1** représente les clients qui ont passé une seule commande récemment avec un montant faible et qui sont satisfaits
- **Le cluster 2** représente les clients qui ont passé une seule commande avec un montant faible il y a un certain temps et qui sont satisfaits
- **Le cluster 3** représente les clients qui ont passé une seule commande avec un montant faible et qui ne sont pas satisfaits (livraison longue)
- **Le cluster 4** représente les clients qui ont passé une seule commande, de montant élevé et qui sont plutôt satisfaits
- **Le cluster 5** représente les clients qui ont passé une seule commande de plusieurs articles, de montant faible
- **Le cluster 6** représente les clients qui ont passé plusieurs commandes de montant faible et qui sont relativement satisfaits

V. CONTRAT DE MAINTENANCE

V. Contrat de maintenance (1/2)



V. Contrat de maintenance (2/2)



L'ARI doit être interprété comme suit :

- ARI >= 0,90 **excellente récupération**
- 0,80 =< ARI < 0,90 **bonne récupération**
- 0,65 =< ARI < 0,80 **récupération modérée**
- ARI < 0,65 **mauvaise récupération**

	Jours retranchés	ARI	days_update	month_update
0	7	0.995060	688	23.0
1	21	0.976340	674	22.0
2	35	0.859557	660	22.0
3	49	0.873206	646	22.0
4	63	0.892216	632	21.0
5	77	0.910530	618	21.0
6	91	0.916372	604	20.0
7	105	0.916049	590	20.0
8	119	0.912720	576	19.0
9	133	0.908181	562	19.0
10	147	0.907051	548	18.0
11	161	0.904575	534	18.0
12	175	0.913960	520	17.0
30	427	0.940121	268	9.0
31	441	0.942073	254	8.0
32	455	0.930041	240	8.0
33	469	0.892741	226	8.0
34	483	0.372862	212	7.0
35	497	0.381971	198	7.0
36	511	0.387485	184	6.0

VI. CONCLUSIONS

V. Conclusions

- Un premier modèle de clustering issu des données est possible (k-means sur variables différenciantes)
- Les 6 clusters sont facilement interprétables par le métier : axes **montant**, **satisfaction**, **fréquence** de commande, **récence**
- Une **maintenance tous les 7 mois** est à prévoir
- La segmentation pourrait être améliorée en introduisant de **nouvelles features sur les clients** par exemple ou en se focalisant sur les clients ayant passé plusieurs commandes

MERCI