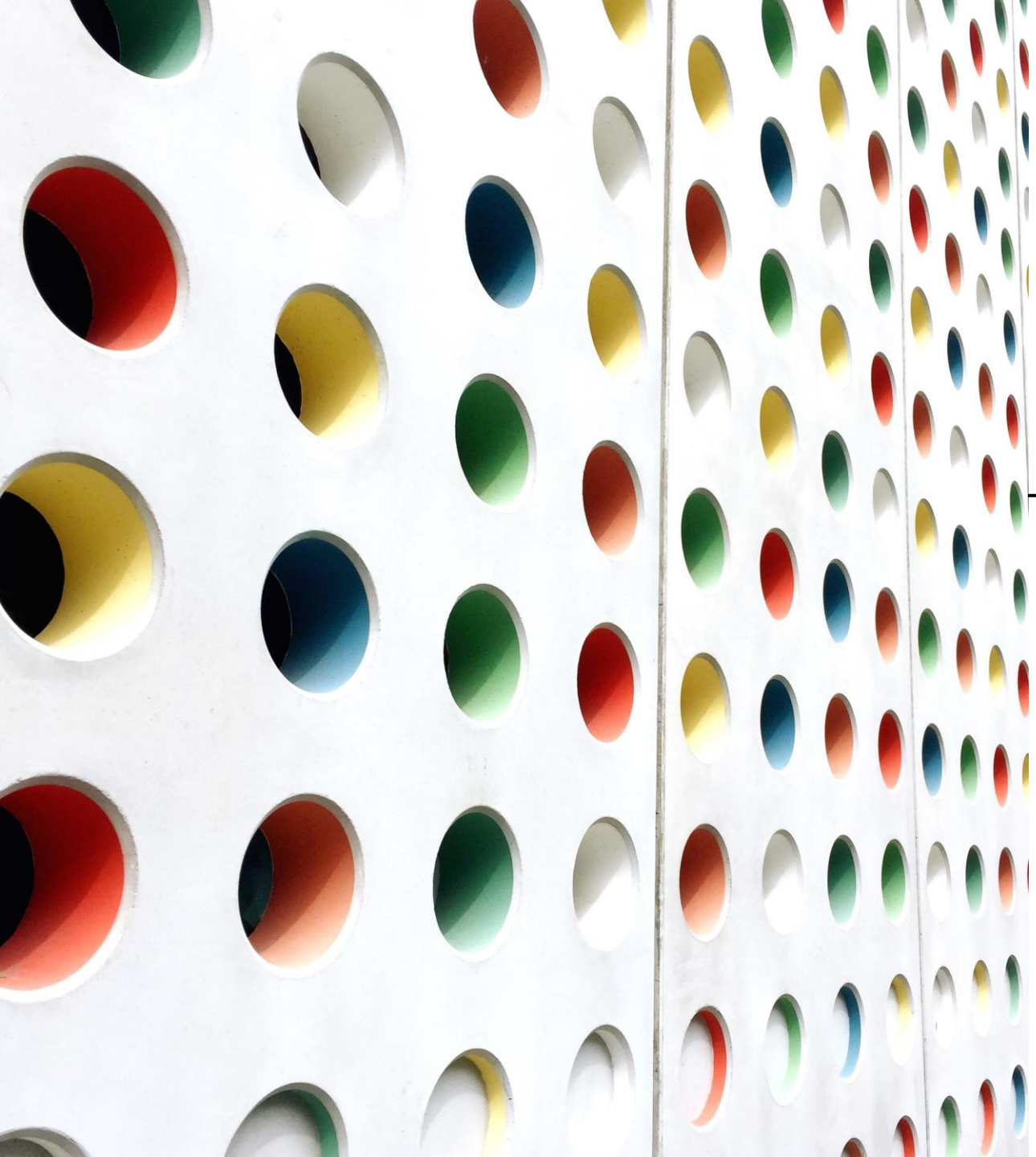


Segmentez des clients d'un site e-commerce

olist

PROJET: #5

Formation: Data Scientist

Mentor: Medina Hadjem

Agenda

olist

Introduction
&
Problématique

Modélisation

Simulation
&
Démonstration

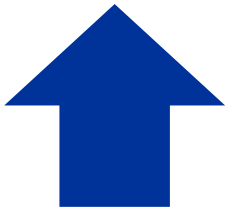
Agenda

olist

Introduction
&
Problématique

Modélisation

Simulation
&
Démonstration

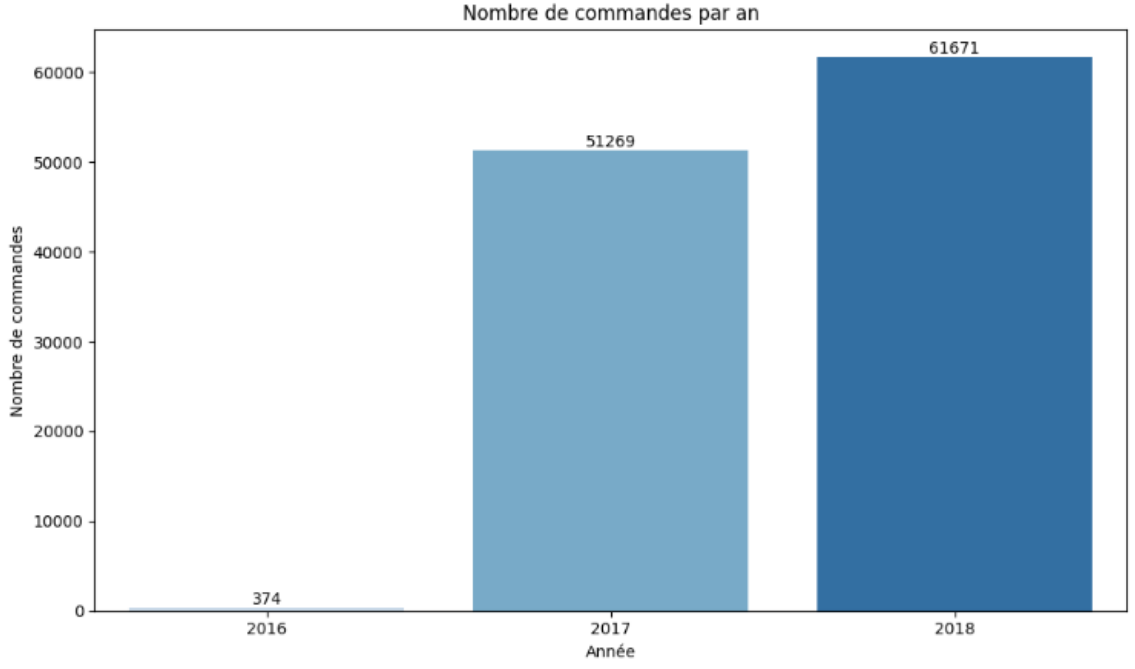


Présentation générale d'Olist

OLIST :

Site e-commerce brésilien qui propose une solution de vente sur les marketplaces.

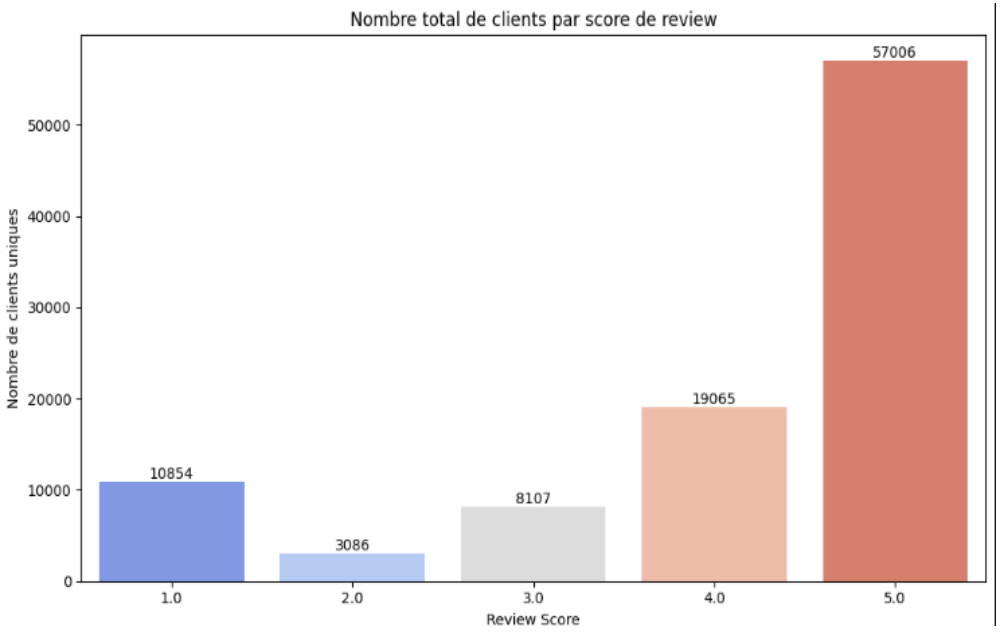
- Nombre de commandes en croissance de 2016 à 2018
- La connaissance des clients est cruciale pour le développement stratégique d'Olist



CLIENTS :

Internautes qui commandent des produits via la plateforme e-commerce.

- La livraison des commandes est globalement bien notée

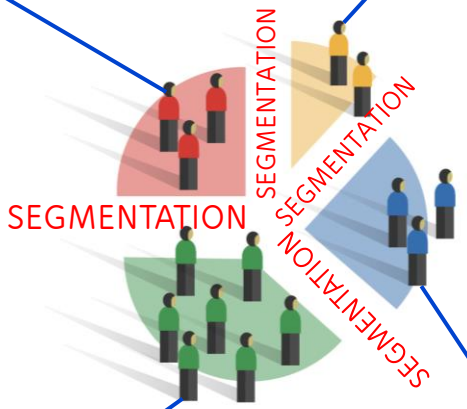


Importance et enjeux de la segmentation des clients

olist

Connaître les comportements d'achat de ses clients

Optimiser la relation client



Identifier les clients en fonction de leurs segments

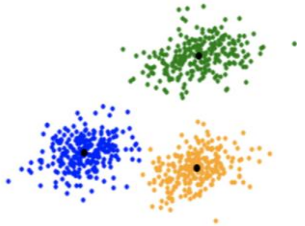
Adapter les campagnes de communication à destination des clients

PROBLEMATIQUE

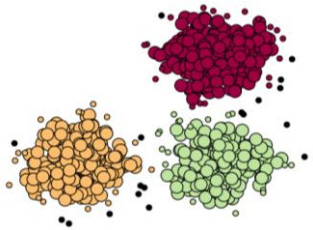
Comment mieux connaître les clients d'OLIST ?

Segmenter les clients à l'aide de modèles de Machine Learning non supervisés

K - MEANS



DB SCAN



Etapes du traitement de la problématique

Dataset OLIST

113. 314 lignes X 27 colonnes

Sélection et fusion des tables:

- Orders_items : Information produits et livraisons
- Orders : Statut des commandes
- Products : Caractéristiques des produits
- Customers : Identifiants et localisations des clients
- Order_reviews: Avis sur les commandes

EDA

Analyse et sélection des features
RFM :

- *Recensement*: Nombre de jours total depuis la dernière commande par client.
[« order_purchase_timestamp », «customer_unique_id »]
- *Fréquence* : Nombre total de commandes par client.
[«customer_unique_id », « order_id »]
- *Montant* : Montant total des commandes par client.
[«customer_unique_id », « Price »]

Feature Engineering

- Suppression des lignes avec des prix négatifs ou égaux à 0
- Feature Review_Score:
Traitement des valeurs manquantes = Imputation de la moyenne des scores de revus (942 valeurs, 0,83 % du dataset OLIST)
- Feature « Review_Score » : One-Hot Encoding = *Review _score_1, Review_score_2, Review_score_3, Review_Score_4, Review_Score_5.*
Réponse : (1 Oui 0: Non)
- Standardisation des données :
Mise à l'échelle des données RFM avec StandardScaler
- Conversion de la colonne :
« order_purchase_timestamp » en type *« datetime ».*

Modeling et évaluation

ETAPES :

- Application et analyse des résultats de clustering : Algorithme K-Means
- Application et analyse des résultats de clustering : Algorithme DBSCAN
- Simulation pour estimer la durée d'un contrat de maintenance et l'évaluation du modèle K-means avec l'Adjusted Rand Index

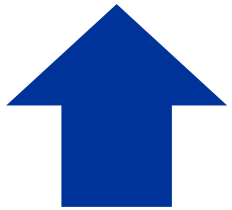
Agenda

olist

Introduction
&
Problématique

Modélisation

Simulation
&
Démonstration



Modélisation de modèles non supervisés en Machine Learning

olist

K - Means

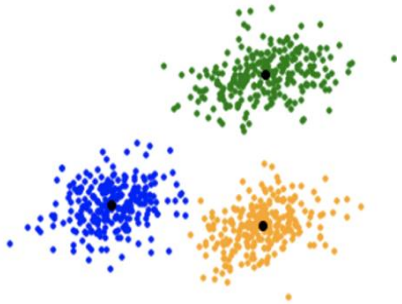
Algorithme qui regroupe les **points en K clusters autour de centroïdes**, en minimisant **la distance euclidienne** entre les points et leurs centroïdes respectifs.

Partitionne les données en **K clusters** prédéfinis.

Chaque cluster est représenté par un **centroïde** (moyenne des points).

Distance **euclidienne** pour assigner les points au cluster proche.

Convient à des données bien séparées, mais reste **sensible aux valeurs aberrantes**.



K-Means

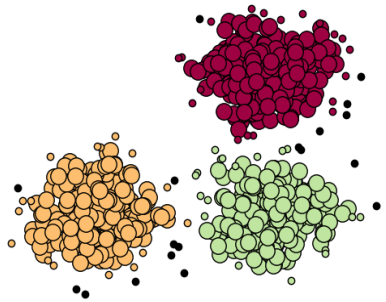
SEGMENTATION

DB SCAN

Algorithme basé sur la **densité** qui identifie les clusters et les points bruités. Il peut identifier des clusters de forme variée.

Regroupe les points en fonction de leur **densité locale**.

Identifie les outliers (points isolés) comme du **bruit (-1)**.



DB SCAN

Pas besoin de spécifier le nombre de clusters

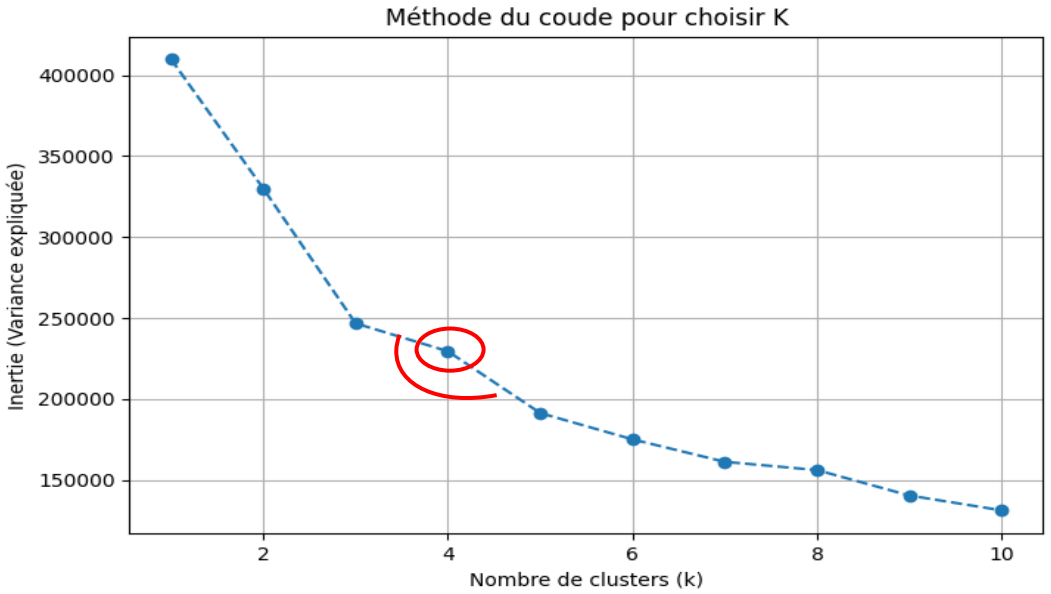
Performant pour **les données de formes complexes**, mais sensible aux paramètres **epsilon** (Eps).

Modélisation de modèles non supervisés en Machine Learning

olist

K – Means sur RFM et Review scores

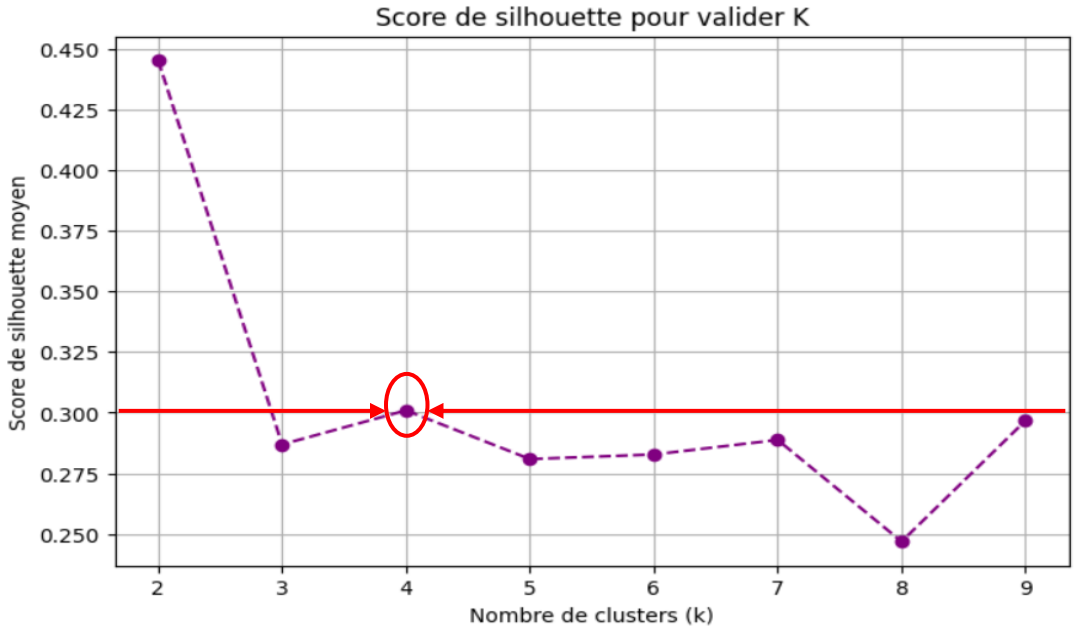
1. On détermine le nombre de clusters optimal



Le « coude » marque le ralentissement de la diminution de **l'inertie** ici, il est visible pour

$k = 4$

2. On mesure la qualité des clusters avec le score de silhouette



Le score de silhouette montre que les clusters sont cohérents et bien séparés, confirmant

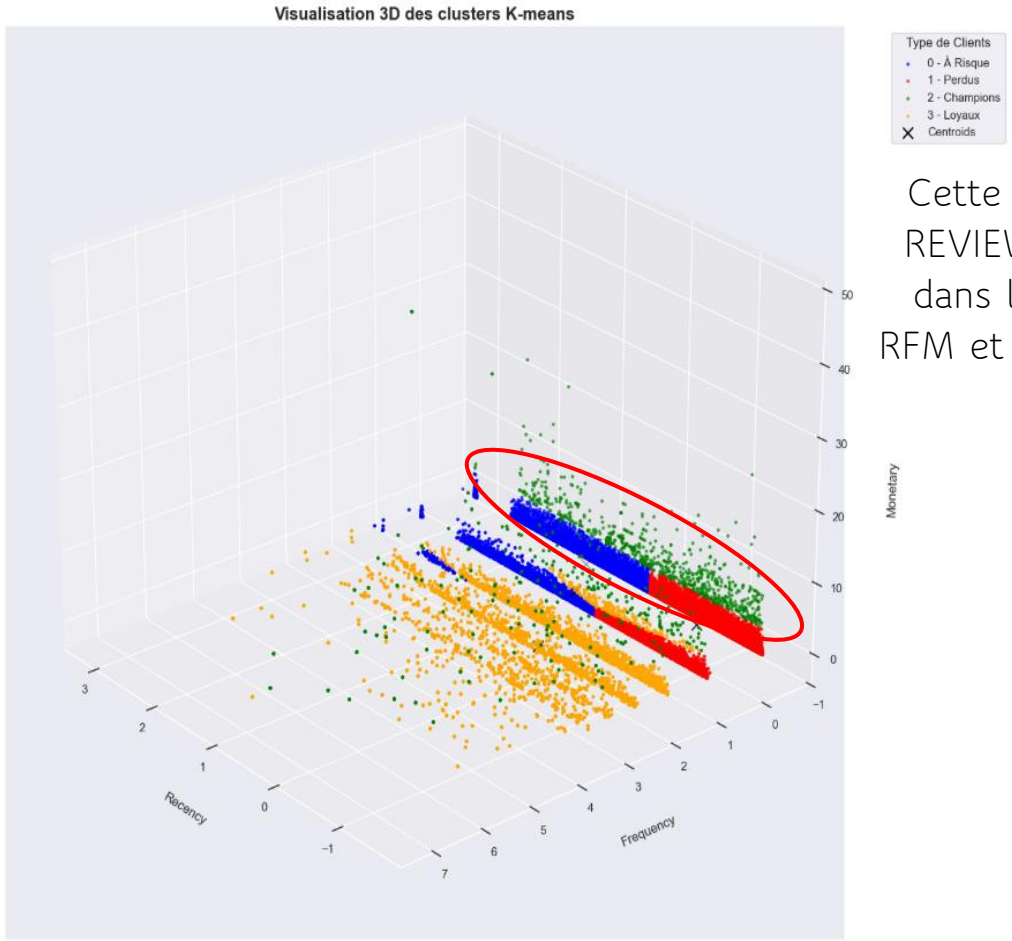
k 4

→ Avec 4 clusters le score de silhouette de 0.30 est suffisant pour distinguer les clients d'OLIST en 4 groupes de clients hétérogènes

Modélisation de modèles non supervisés en Machine Learning

K – Means / Application de l'algorithme non supervisé sur 2 jeux de données

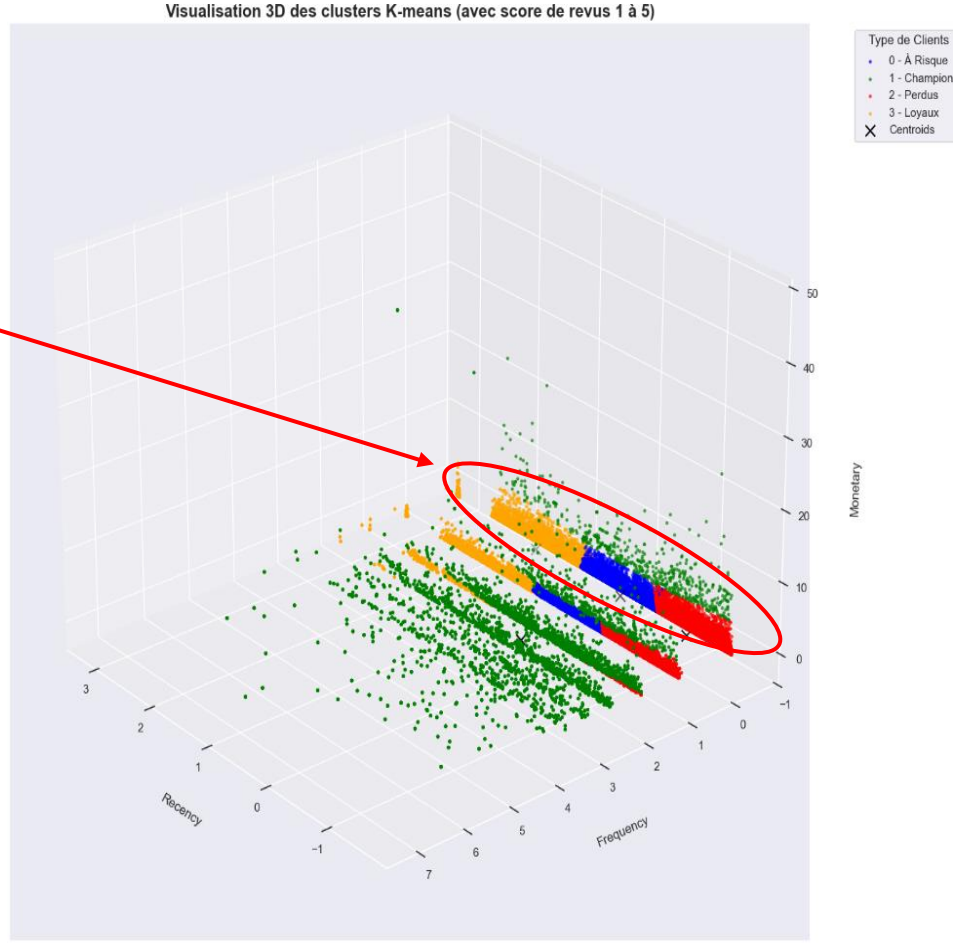
RFM : Clusters peu distincts, plus **difficile à interpréter**.



RFM montre des clusters moins distincts

Cette segmentation enrichie avec les REVIEW SCORES de (1 à 5), s'inscrit dans la continuité des travaux sur la RFM et offre une vision plus précise des clients.*

RFM + Review Score : Meilleure précision et **séparation claire des clusters**



RFM + Review Score = améliore la précision et offre une séparation plus claire.

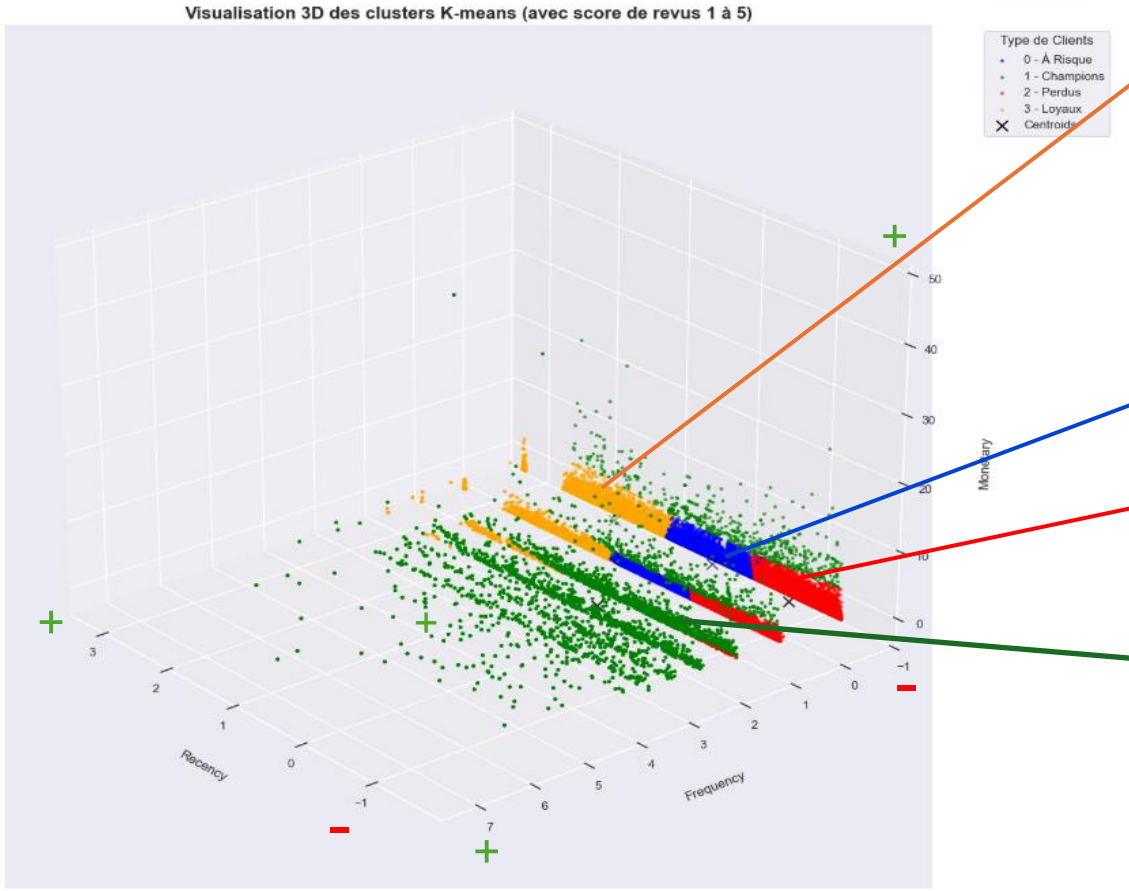
* *La segmentation précédente était basée uniquement sur RFM (Récence, Fréquence, Montant), sans inclure les scores de satisfaction.*

Modélisation de modèles non supervisés en Machine Learning

olist

K – Means / Visualisation des Clusters & Mapping des clients

3. Résultat appliqué de K-Means (Algorithme Non Supervisé) RFM + Review Score



4. Le mapping des clients

Loyaux : fréquence moyenne, récense récente et un montant modéré indiquant une relation stable.

A risque : fréquence et montant faible. l'engagement est réduit.

Perdus : faible fréquence et récence éloignée, signe de désengagement.

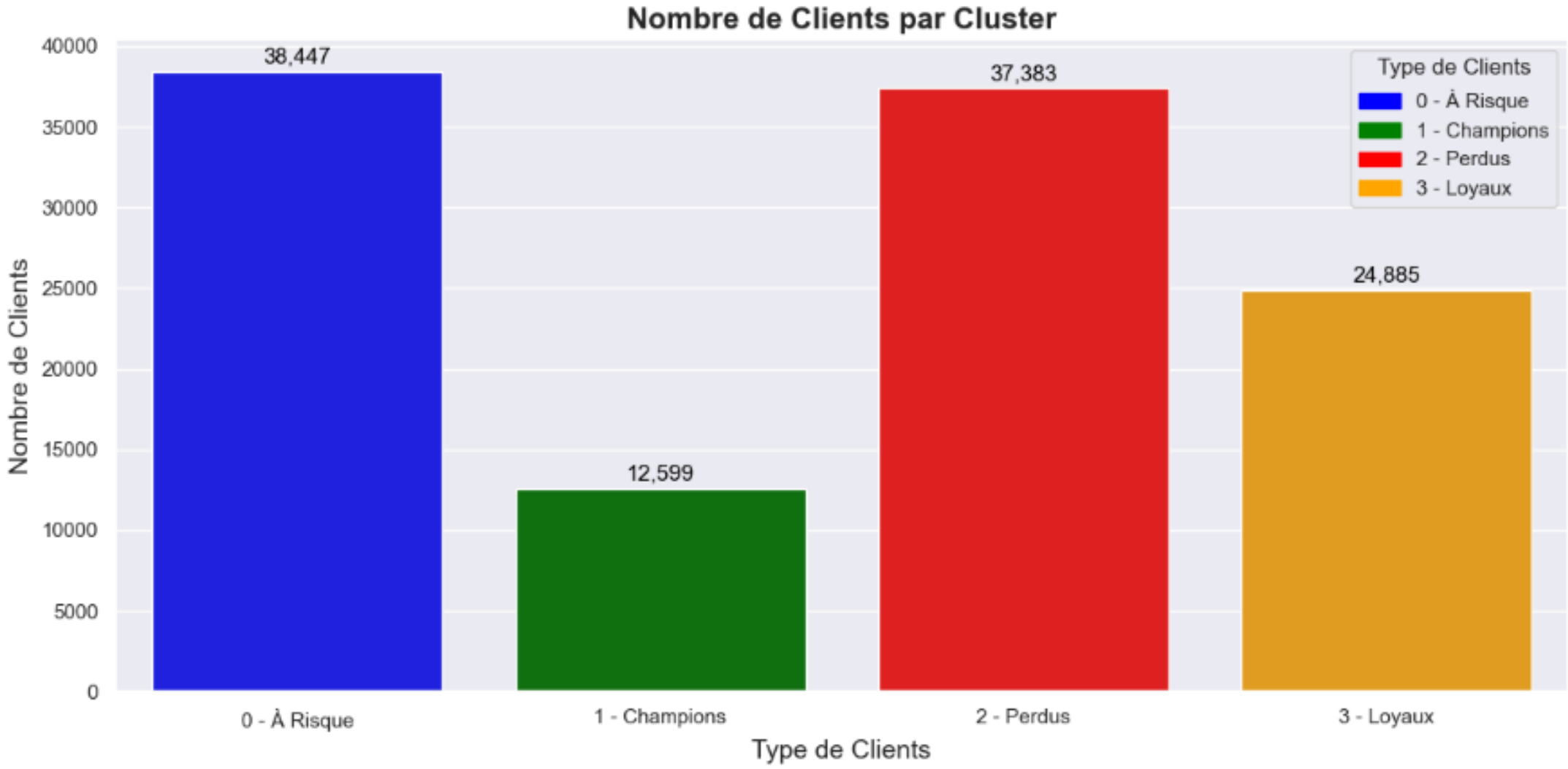
Champions : fréquence d'achat élevée, générant un montant significatif.

Ce mapping nous permet de segmenter les clients d'OLIST selon leur comportement sur la plateforme.

Modélisation de modèles non supervisés en Machine Learning

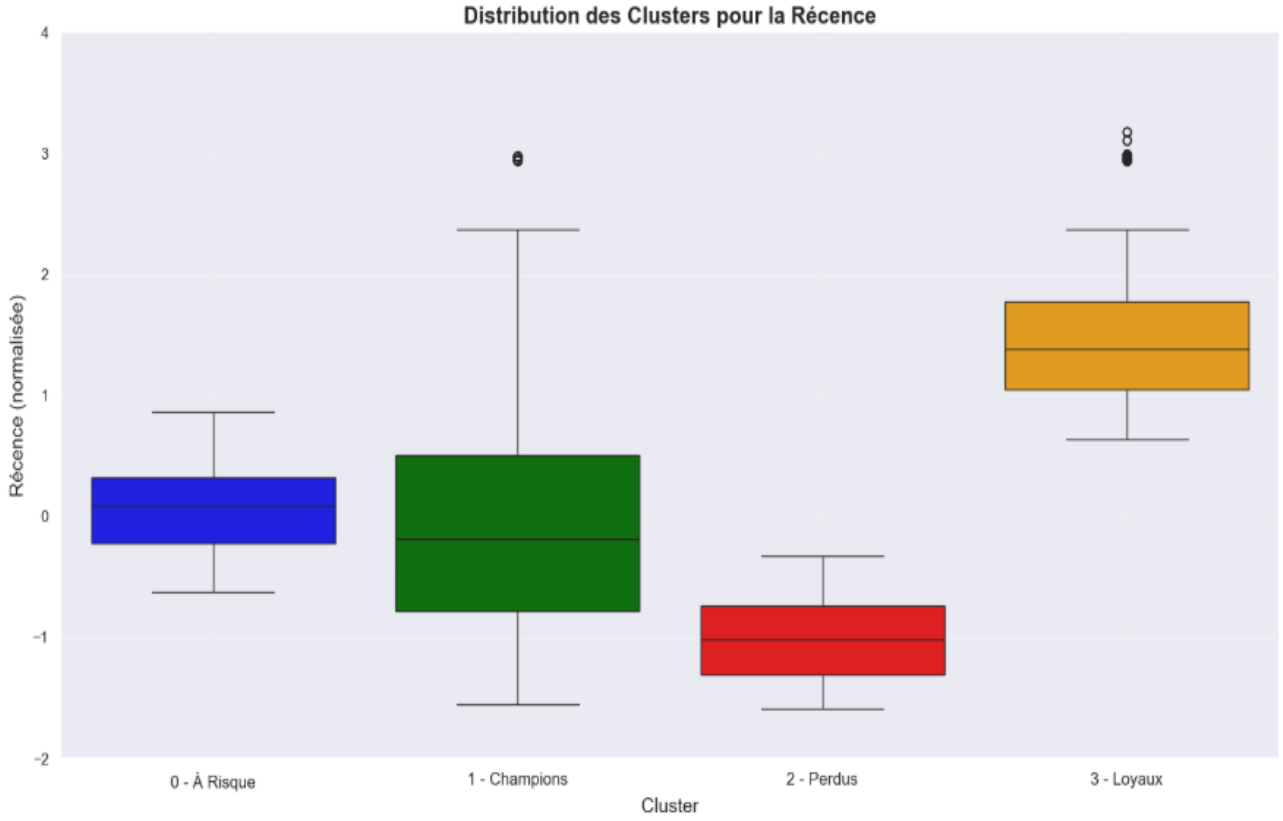
olist

K – Means

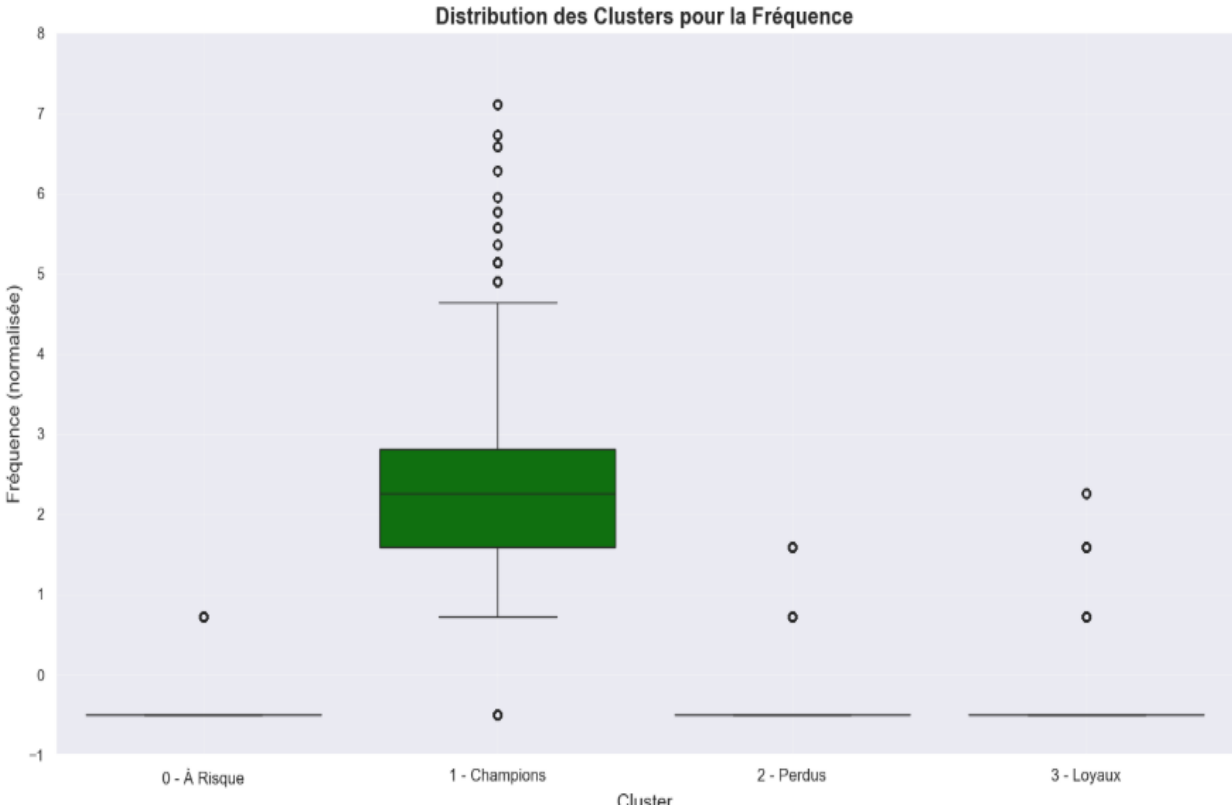


Modélisation de modèles non supervisés en Machine Learning

K – Means



- Loyaux : Ils ont commandé récemment.
- Champions : Une récence modérée indique une activité régulière.
- Perdus : Ils ont une récence très faible, ce qui confirme qu'ils n'ont pas commandé depuis longtemps.
- A Risque : Une récence faible, suggère qu'ils commencent à s'éloigner.

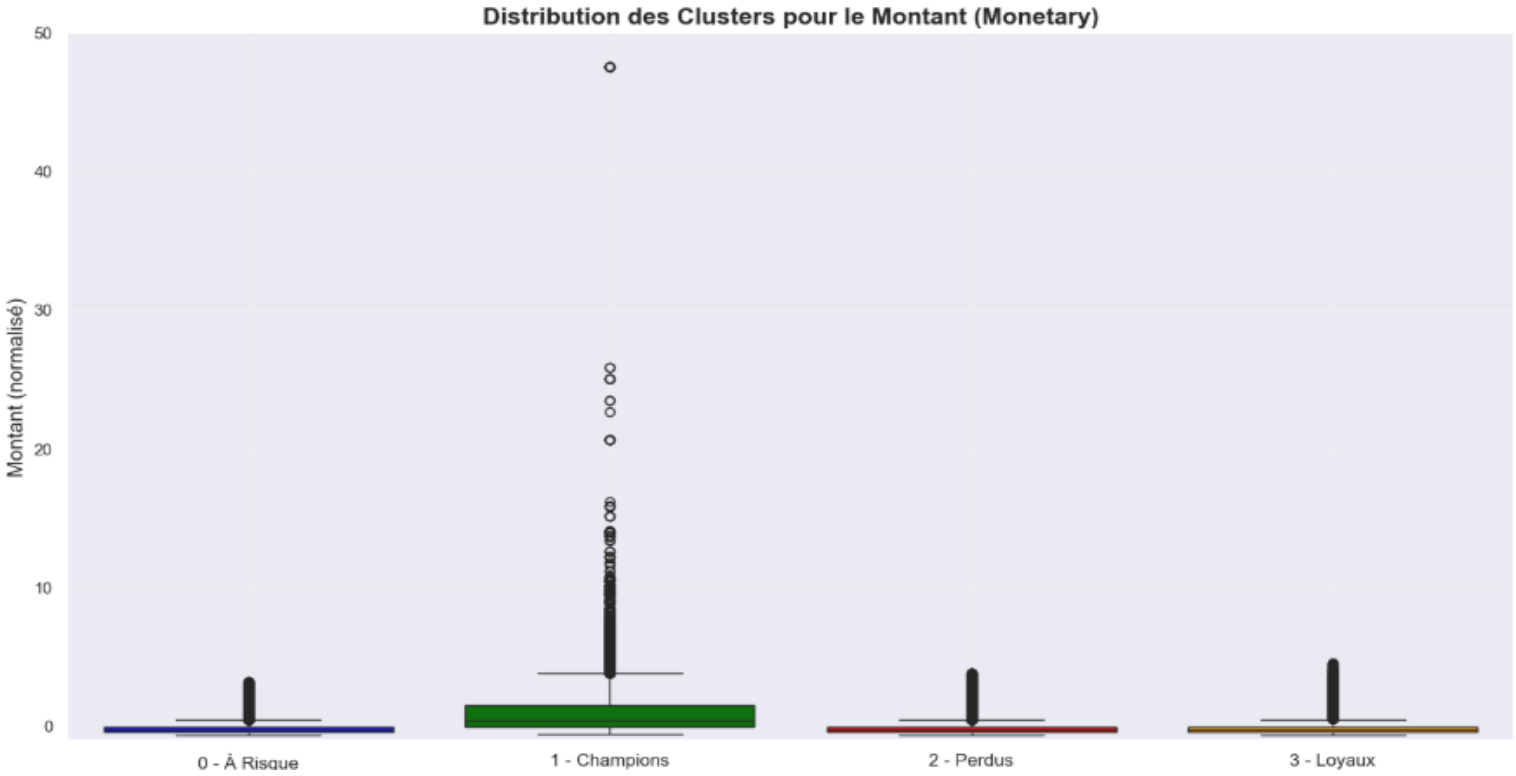


- Champions : La Fréquence d'achat élevée, ils contribuent fortement aux achats sur la plateforme d'OLIST.
- Loyaux : La fréquence est modérée le nombre total de commandes est stable.
- Perdus, A risque : Ils ont une fréquence faible, ce qui indique un désengagement.

Modélisation de modèles non supervisés en Machine Learning

olist

K – Means / Visualisation de la R F M



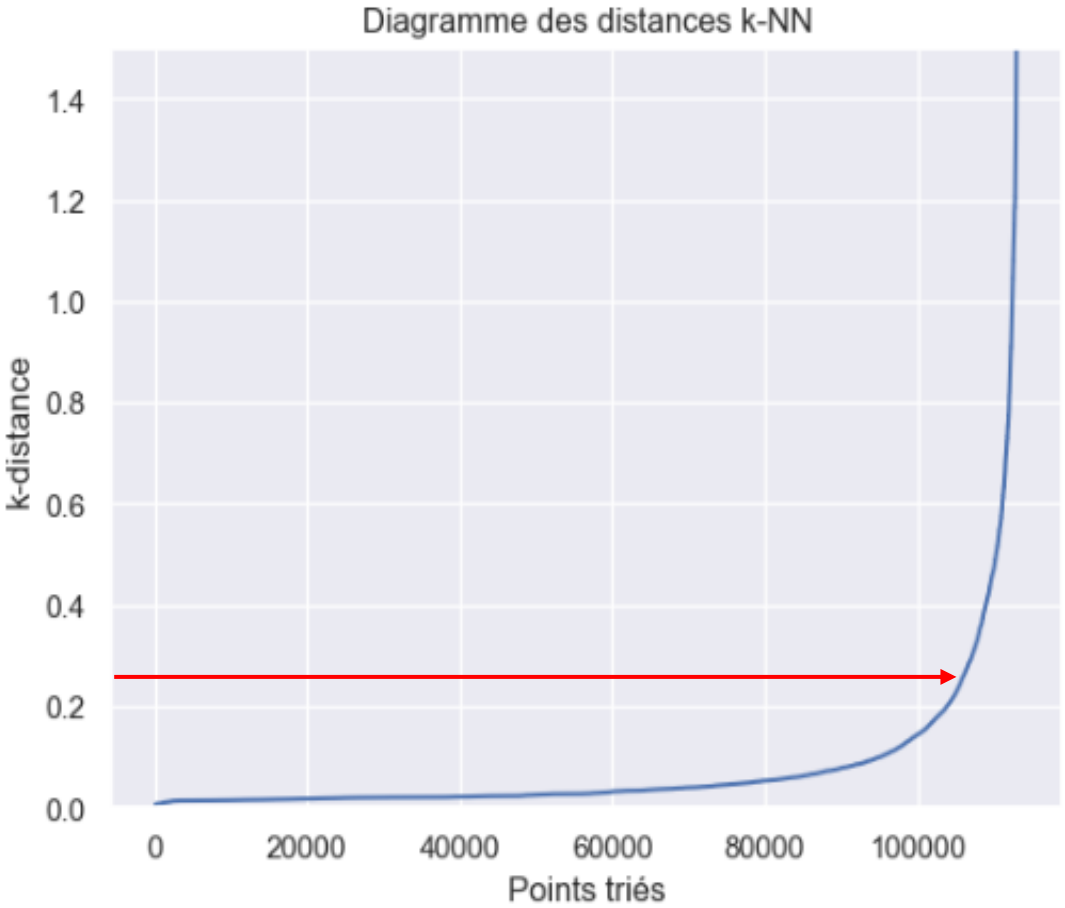
- Champions : Montant d'achat très élevé, avec des valeurs extrêmes. Ils sont les clients les plus précieux (le panier d'achat d'un client champion est élevé en moyenne).
- Loyaux : Montant modéré, la relation est stable.
- A Risque : Montants très faibles, ce qui confirme un faible engagement.
- Perdus : Montants très faibles également.

Modélisation de modèles non supervisés en Machine Learning

olist

Application de DBSCAN : Détection de clusters sur RFM

1. On détermine la valeur optimale de l' "ε" (epsilon)



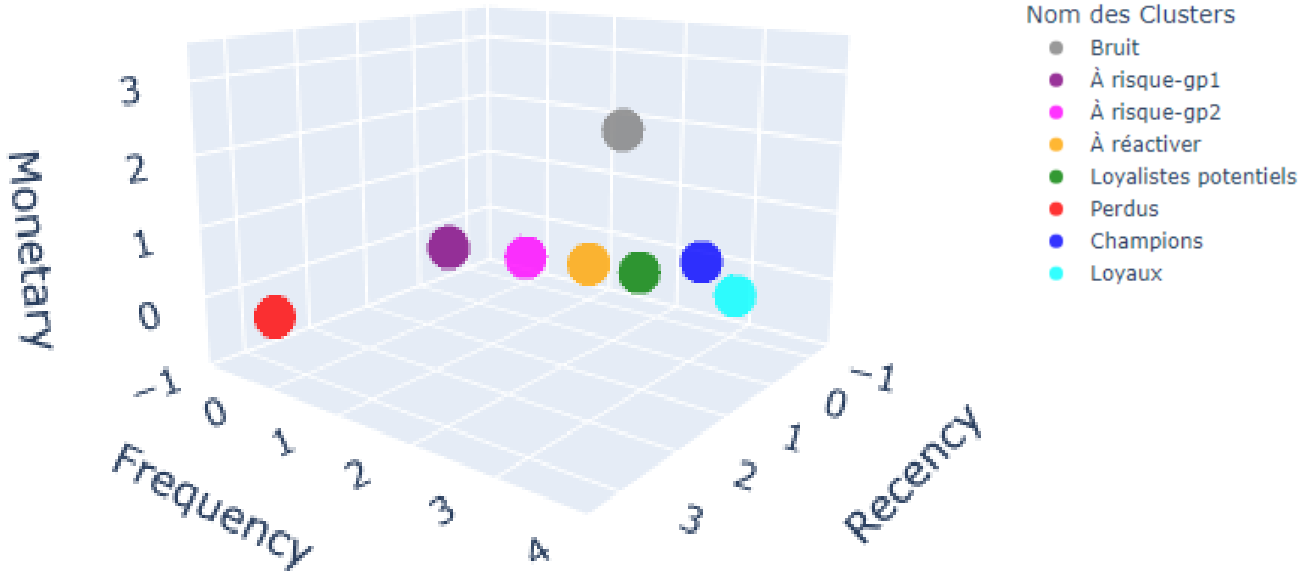
La courbure ("genou") marque la transition entre clusters denses (basses distances) et points isolés (distances élevées).

2. On applique l'algorithme DB SCAN avec les paramètres suivants

- On définit l' ε" (**epsilon**) à **0.3** pour qu'un point soit considéré comme voisin grâce au **genou** observé sur le diagramme des distances k-NN *(« k-Nearest Neighbors »,« Plus proches voisins »).*

- **Nb. Min de points pour former un cluster = 100** car le dataset contient (plus de 100. 000 pts.)

Observation de la Moyenne des clusters

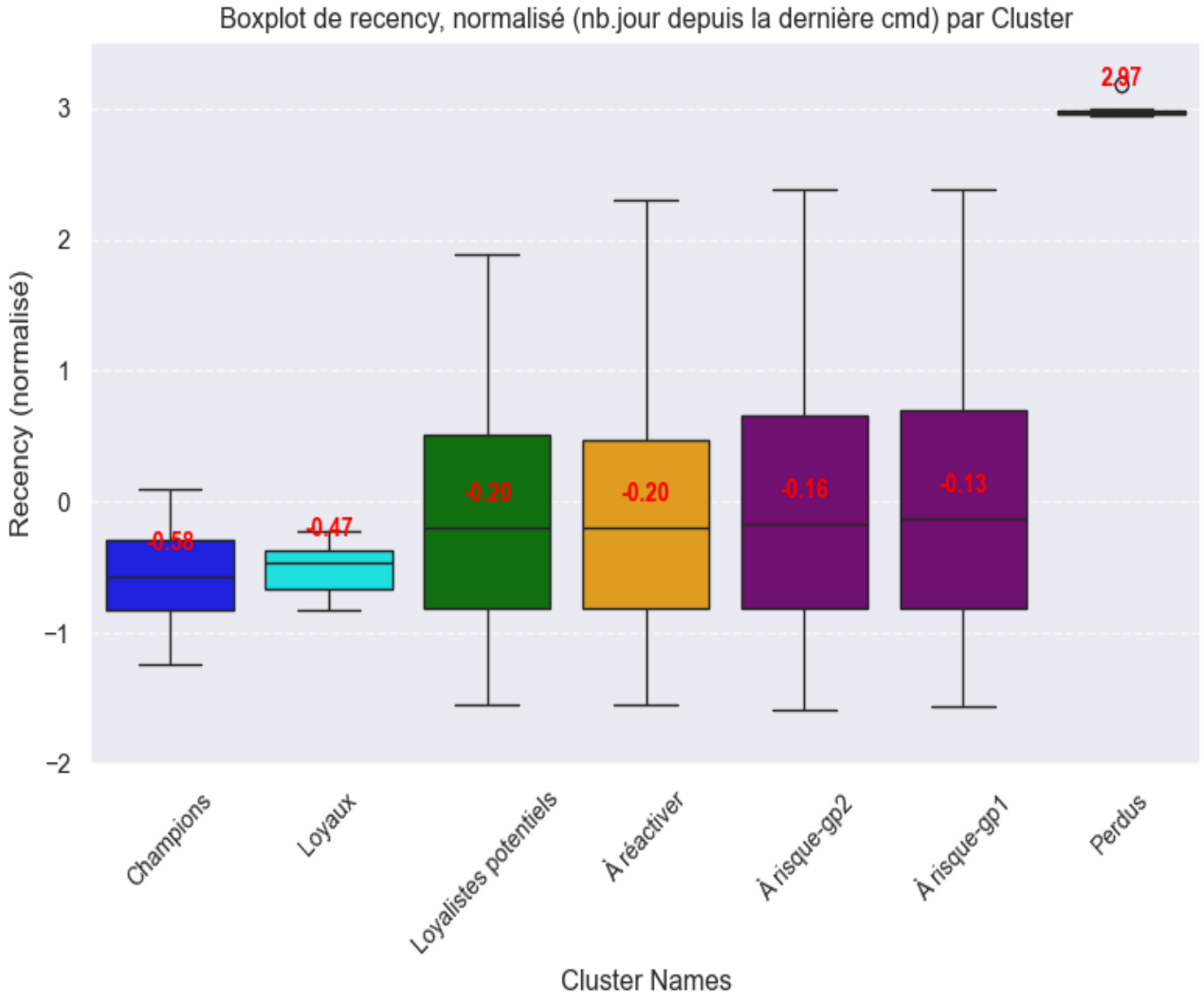


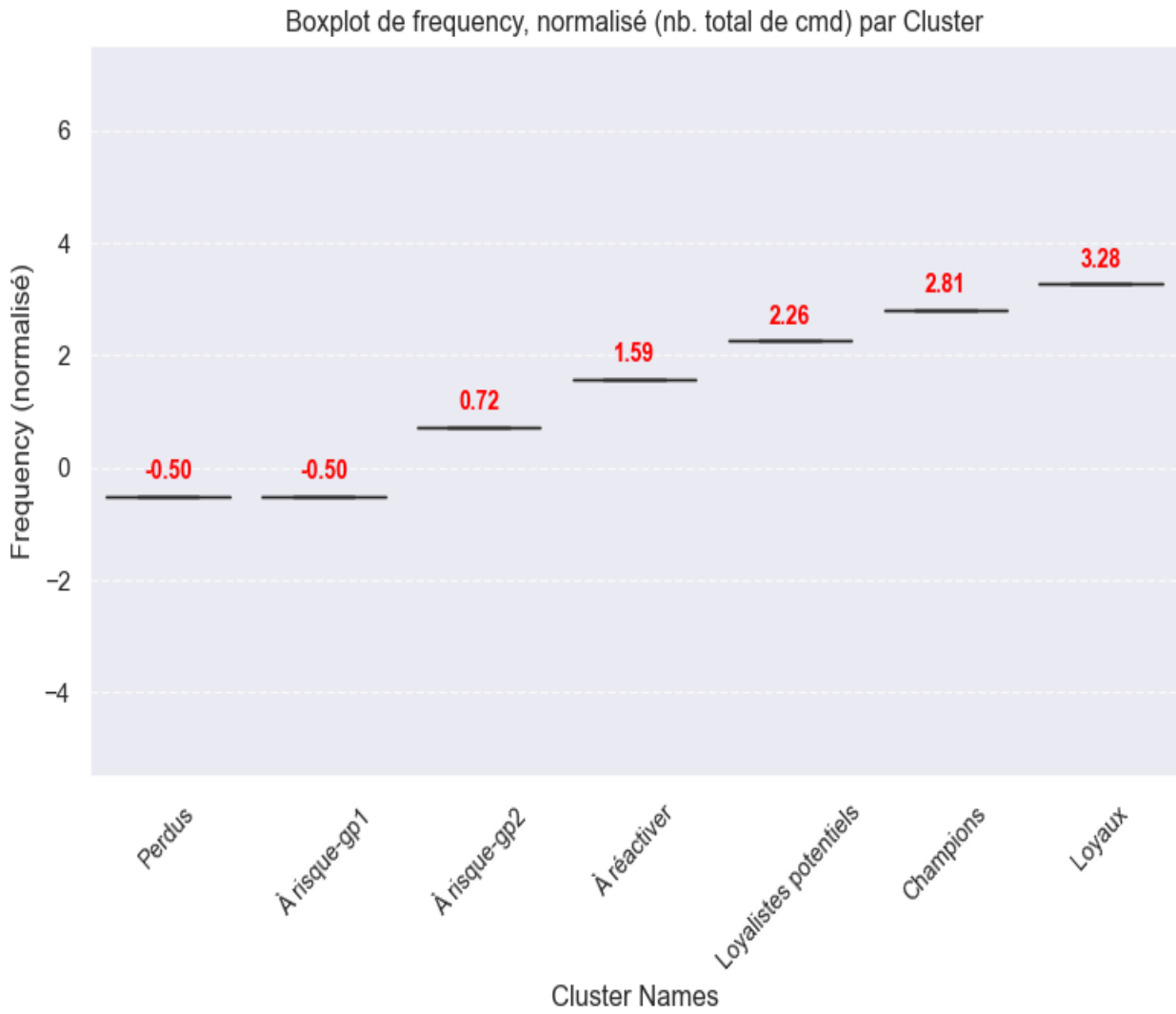
Modélisation de modèles non supervisés en Machine Learning

olist

Application de DBSCAN : Détection de clusters sur RFM

4. Confirmation des labels attribués par la visualisation des boxplots des variables RFM



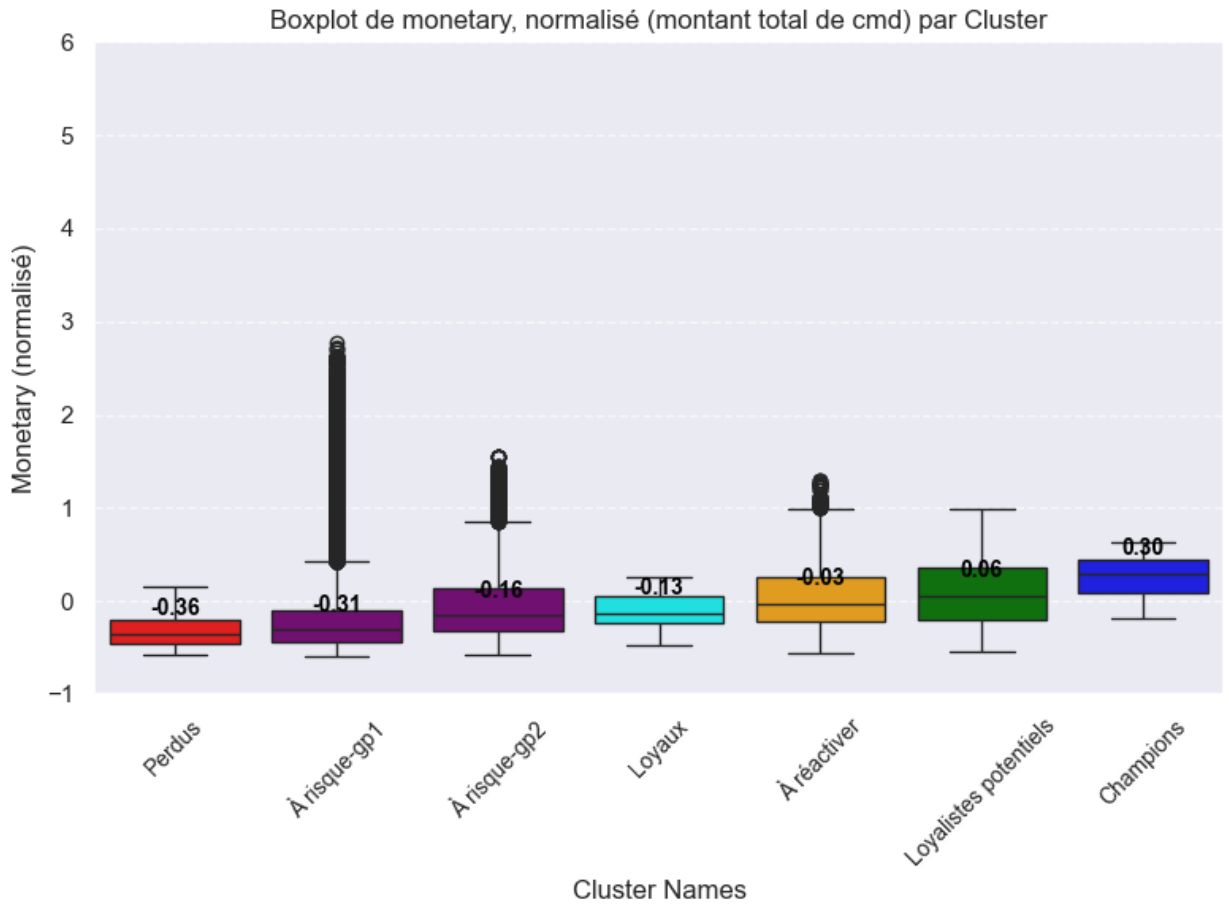


Modélisation de modèles non supervisés en Machine Learning

olist

Application de DBSCAN : Détection de clusters sur RFM

4. Confirmation des labels attribués par la visualisation de box plot des variables RFM



Pour chaque variable RFM, on observe des variations claires entre les clusters (en visualisant les médianes) , cela nous confirme la pertinence des labels obtenus grâce à l'application de l'algorithme non supervisé Density-Based Spatial Clustering of Applications of Noise.

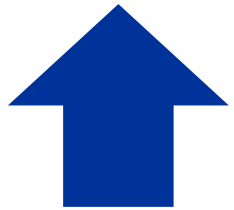
Agenda

olist

Introduction
&
Problématique

Modélisation

Simulation
&
Démonstration



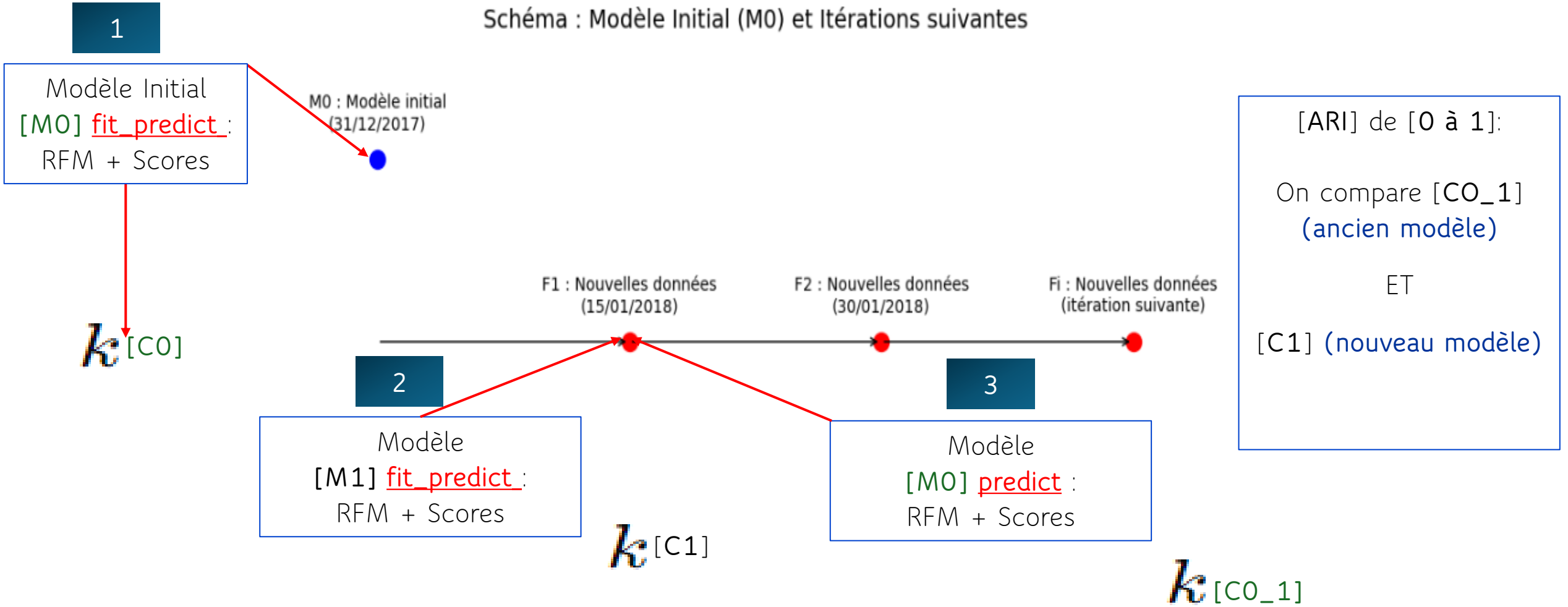
Simulation d'un contrat de maintenance

olist

Utilisation de la métrique « ARI » - Adjusted Rand Index

L'ARI (Adjusted Rand Index) est une métrique utilisée pour évaluer la similarité entre deux partitions d'un jeu de données.
Ici on utilise l'ARI pour évaluer la qualité du clustering K-Means RFM + Review Scores.

Schéma : Modèle Initial (M0) et Itérations suivantes



Simulation d'un contrat de maintenance

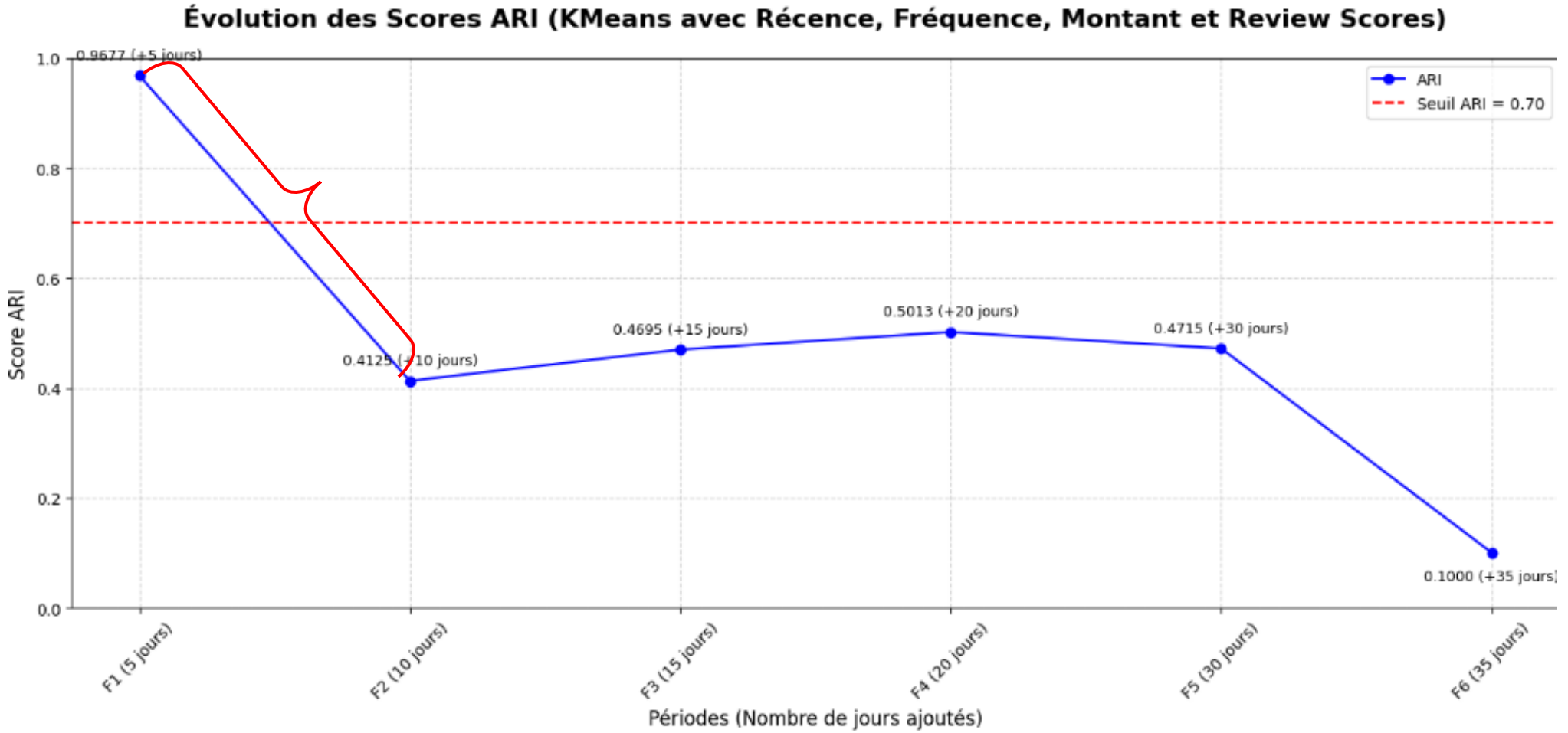
olist

Utilisation de la métrique « ARI » - Adjusted Rand Index

Un ARI (Adjusted Rand Index) proche de 1 indique une performance optimale

ARI = > 0

ARI optimal = 1



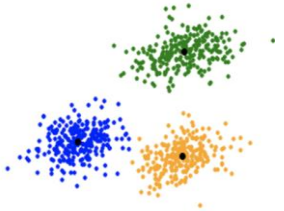
Au vu du seuil modéré de l'ARI 0.70 Le modèle [M0] doit être maintenu tous les 15 jours pour garantir une **segmentation optimale des clients**

Conclusion

La segmentation K – Means nous a permis d'identifier 4 segments de clients:

1. Les Champions: ils représentent les clients les plus précieux, leur nombre total de commandes ainsi que le montant total de leurs achats est élevé (fréquence et montant élevés).
2. Les Loyaux: ils représentent les clients qui sont réguliers. Ils interagissent souvent avec la plateforme d'Olist.
3. Les à Risque: ils représentent les clients en désengagement progressif.
4. Les Perdus : Ils sont faiblement engagés.

K – MEANS



- L'ajout des scores de satisfaction (Review Scores) aux analyses de la RFM a considérablement amélioré la segmentation client d'OLIST.
- La simulation de maintenance du modèle grâce à la métrique de l'ARI nous a montré qu'une mise à jour mensuelle est optimale pour maintenir une segmentation fiable.

Perspective :

Déployer des campagnes marketing personnalisées pour maximiser l'engagement client. Ex :

- Champions : Offrir des avantages exclusifs pour maintenir leur fidélité.
- Loyaux : Stimuler les achats en proposant des offres complémentaires.
- À Risque et Perdus : Lancer des campagnes de réactivation, comme des réductions ou des rappels personnalisés.

Mettre en place un suivi de performance commerciale par segment client.

QUESTIONS & REPONSES

MERCI