

# P5 SEGMENTEZ DES CLIENTS D'UN SITE DE E-COMMERCE

Formation Data Scientist  
OpenClassrooms

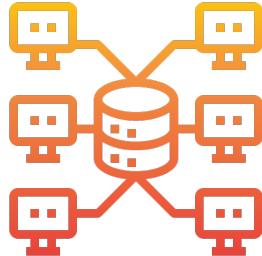
# CONTEXTE

oist

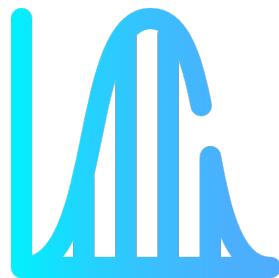
# OBJECTIFS

## Problématiques :

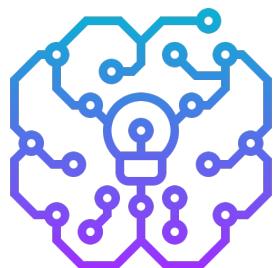
- **segmenter la clientèle** d'un site brésilien de e-commerce, afin d'améliorer leur communication
- proposer un **plan de maintenance** du modèle de segmentation



Nettoyer le jeu de données.



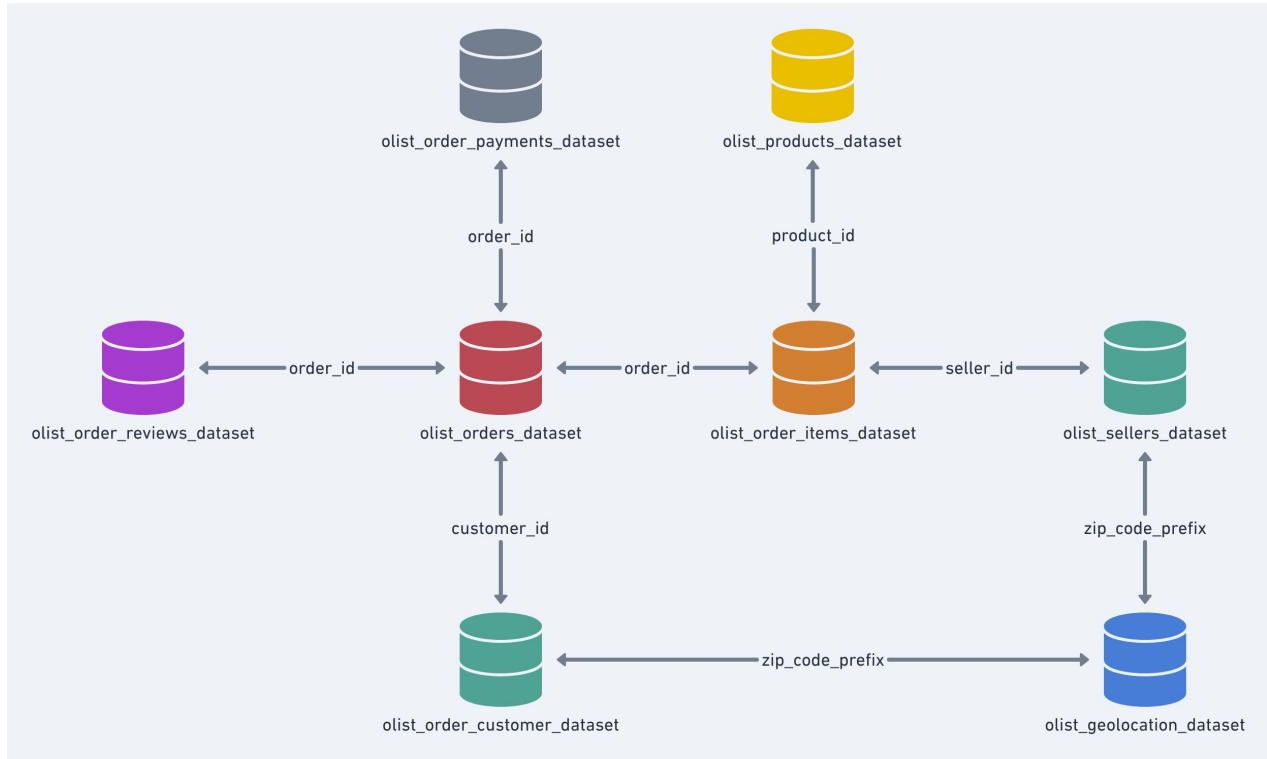
Effectuer une analyse exploratoire pour évaluer la pertinences des features.



Modélisation par Machine Learning : Segmentation avec K-means, Elbow curves, stabilité, plan de maintenance, interprétation des clusters

# DESCRIPTION DU JEU DE DONNÉES

- **Kaggle**  
(<https://www.kaggle.com/olistbr/brazilian-e-commerce>)
- 9 datasets à réunir (merge)
- Format : ~ 120k lignes  
~ 40 colonnes
- Lignes = produits commandés
- Colonnes = features :
  - Customer unique ID
  - Géolocalisation
  - Temporelles
  - Description produits
  - Prix et paiements
  - Reviews



Le merge doit être méthodique !  
Peu de valeurs manquantes (4%) surtout dans reviews  
Dtype → nombres 45% / chaînes de caractères 55%  
Utilisation de seulement 3% de la base de donnée

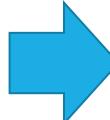
# FEATURE ENGINEERING POUR UNE ANALYSE DE TYPE RFM

Segmentation **RFM** :

**Recency** : nombre de jours depuis la dernière commande

**Frequency** : nombre de commandes

**Monetary value** : paiement moyen par commande



Proposition d'une segmentation **RFM+** :

**Recency**

**Frequency**

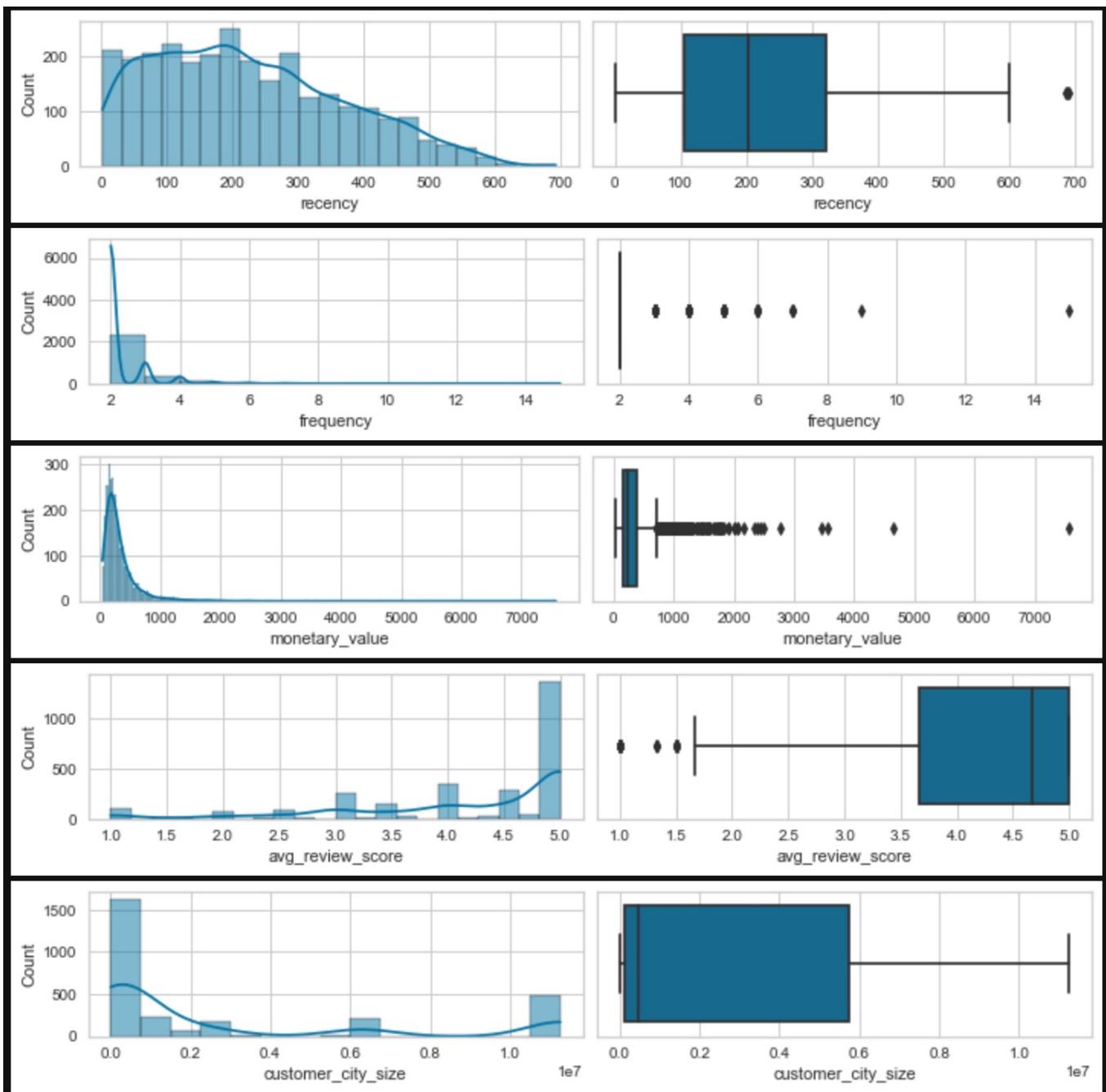
**Monetary value**

+

**Indicateur géographique** : taille de la ville du client  
→ merge avec un dataset trouvé sur kaggle

**Satisfaction** : score de review

# ANALYSE UNIVARIÉE DES FEATURES



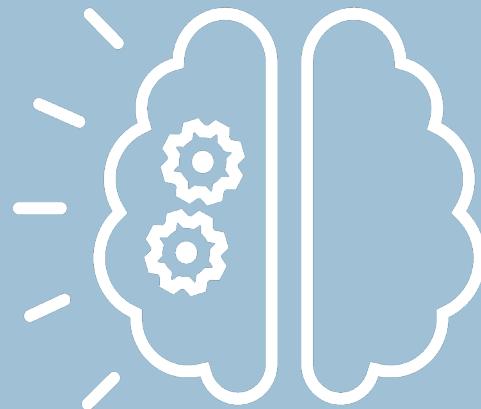
# ANALYSE DE CORRÉLATION DES FEATURES

	recency	frequency	monetary_value	avg_review_score	customer_city_size
recency	1.000000	0.098313	-0.033590	-0.034780	-0.042738
frequency	0.098313	1.000000	0.271220	0.006127	-0.002396
monetary_value	-0.033590	0.271220	1.000000	-0.047875	-0.044934
avg_review_score	-0.034780	0.006127	-0.047875	1.000000	-0.035040
customer_city_size	-0.042738	-0.002396	-0.044934	-0.035040	1.000000

-> Les features ne sont pas corrélées entre elles

# MACHINE LEARNING : K-MEANS CLUSTERING

Clustering using K-means, Elbow curves, stabilité, plan de maintenance, interprétation des clusters



# SÉLECTION DU NOMBRE DE CLUSTER

Calcul de la fonction de coût vs. nb clusters

Trouver une plage de k optimale  
→ « coude » de la courbe (ici 5 à 7)

Selectionner un k

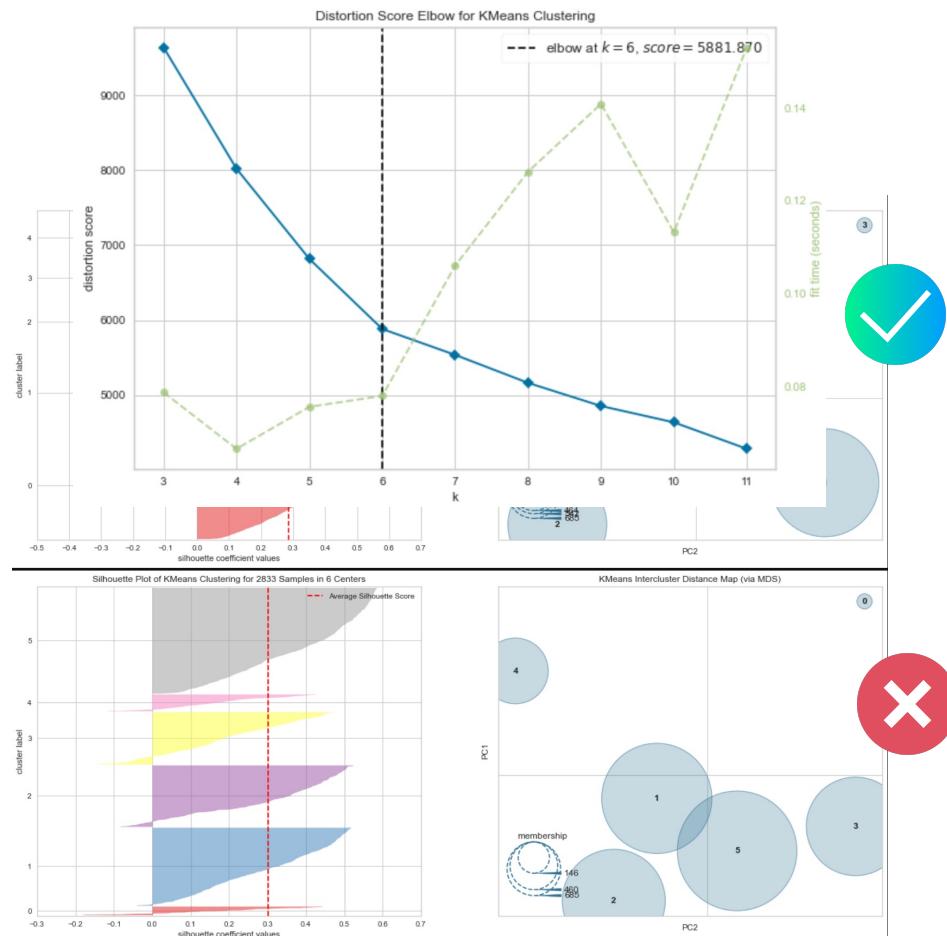
Fit et calcul de labels

Evaluer la pertinence des clusters

Réitérer en changeant k si besoin

## Hyperparamètres retenus

- K = 5
- Algo d'initialisation = Kmean++
- Nombre d'initialisation = 10
- 1000 itérations max.

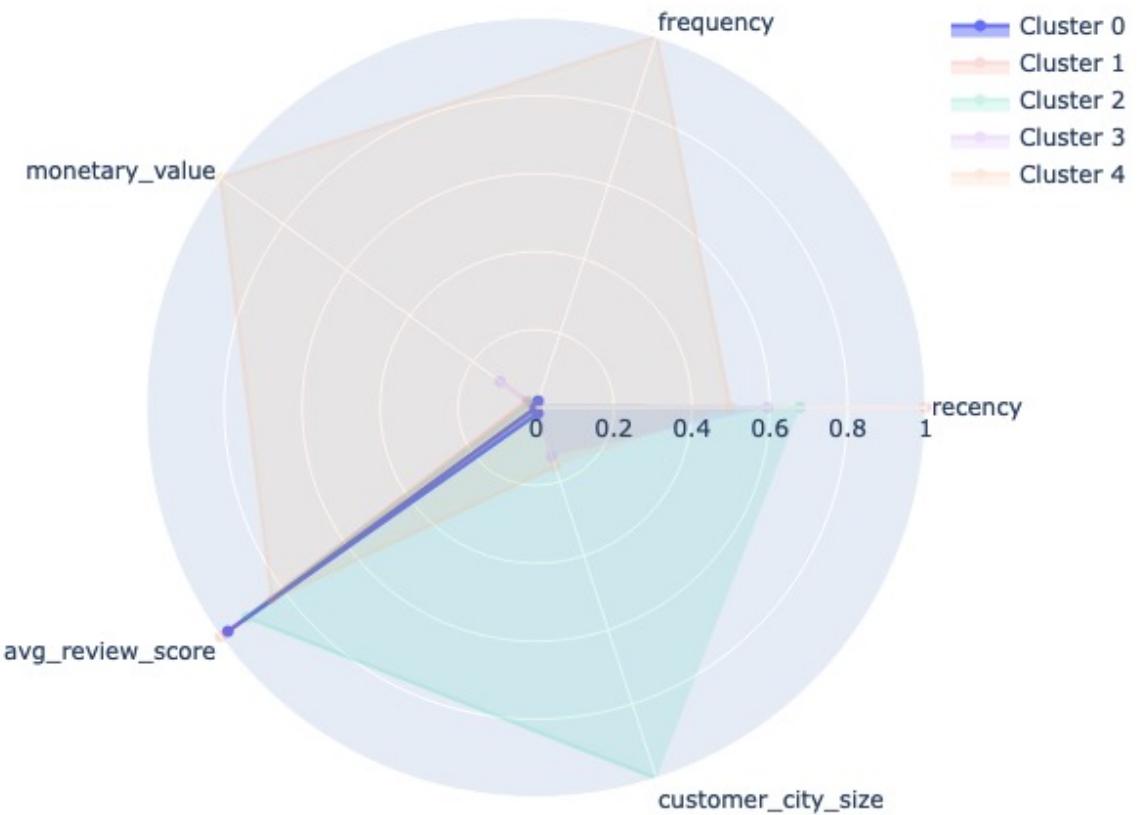


# FORMER AVERAGE CUSTOMER

Caractéristiques:

- Recency minimum de 246 jours
- Frequency moyenne de 2
- Monetary value entre 80 et 420 R\$
- Satisfaction : 90 % score > 4/5
- Villes < 1 millions d'habitants

Cluster's profile comparison

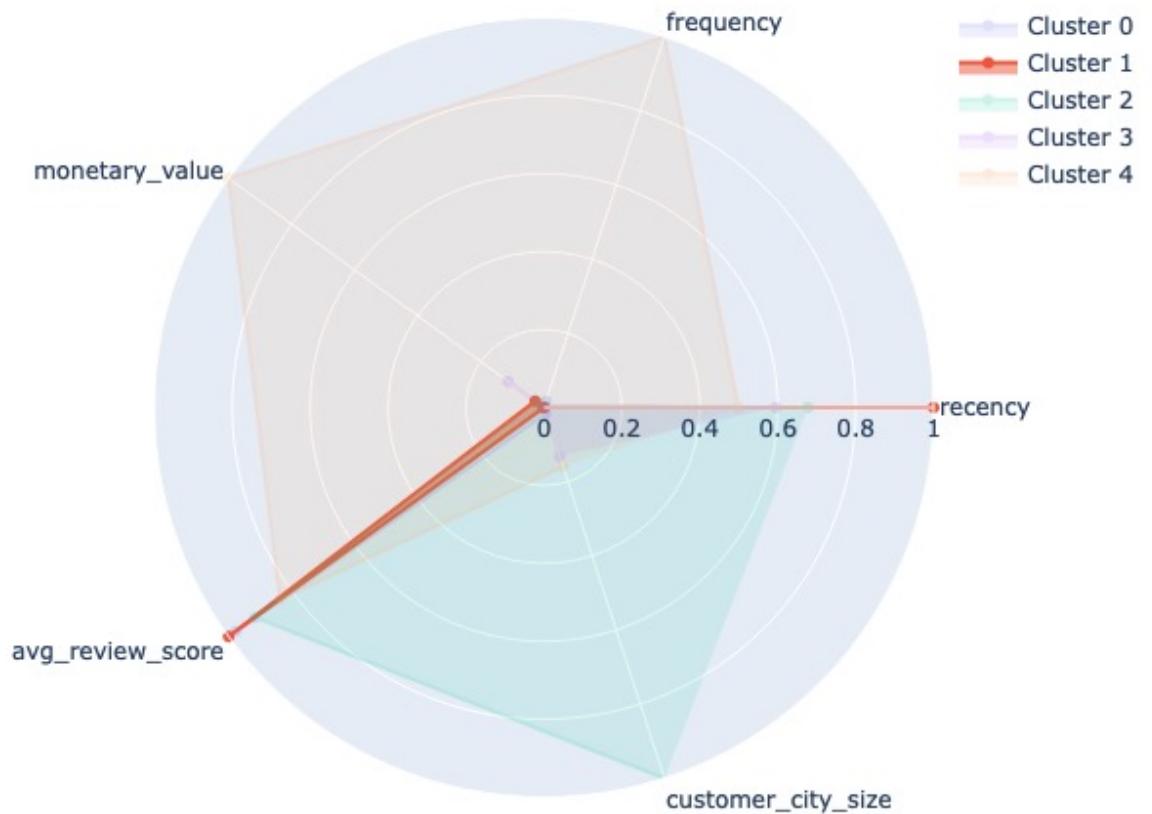


# CURRENT AVERAGE CUSTOMER

## Caractéristiques:

- Recency maximum de 250 jours
- Frequency moyenne de 2
- Monetary value entre 90 et 450 R\$
- Satisfaction 94 % score > 4/5
- Villes < 1 millions d'habitants

Cluster's profile comparison

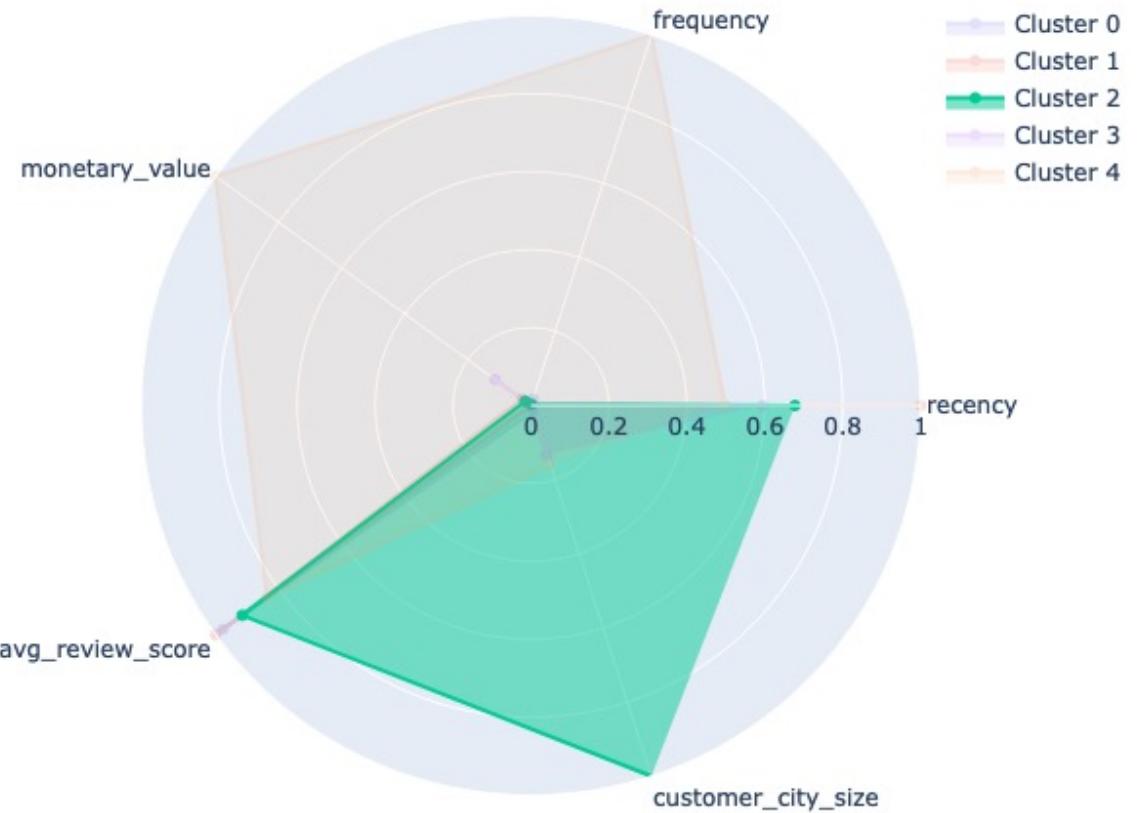


# BIG CITY CUSTOMER

## Caractéristiques:

- Recency de 1 à 592 jours
- Frequency moyenne de 2
- Monetary value entre 70 et 470 R\$
- Satisfaction 81 % score > 4/5
- **Mégapoles > 10 millions d'habitants**

Cluster's profile comparison

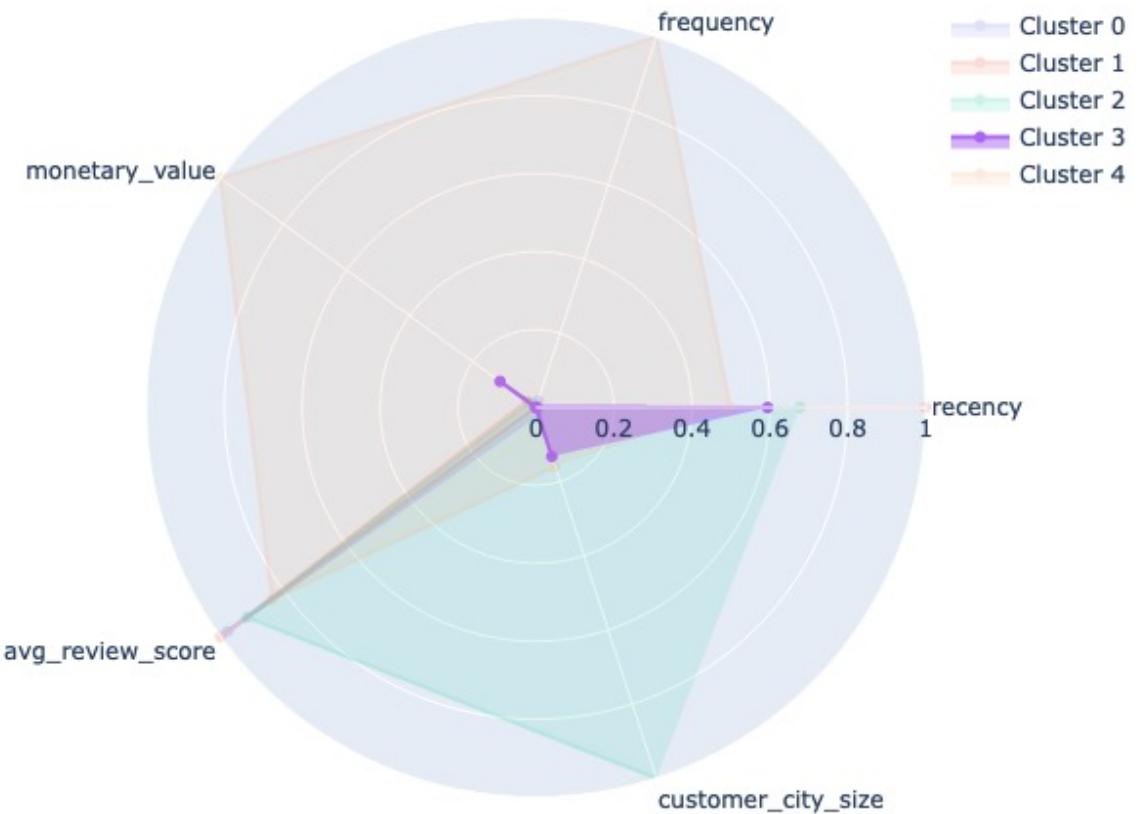


# UNSATISFIED CUSTOMER

## Caractéristiques:

- Recency de 1 à 585 jours
- Frequency moyenne de 2
- Monetary value entre 80 et 580 R\$
- Satisfaction tous score < 3.5/5 dont 40% < 2.5/5
- Villes de 2 millions d'habitant

Cluster's profile comparison

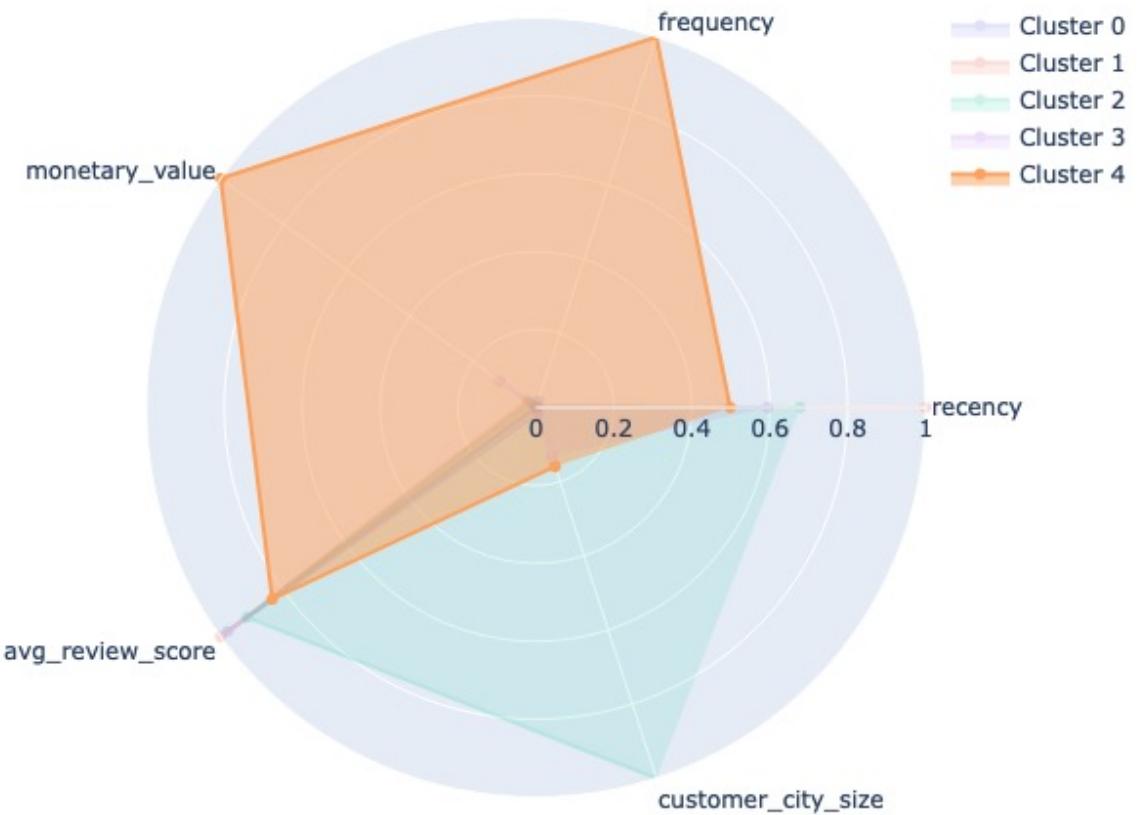


# CHAMPIONS CLIENT

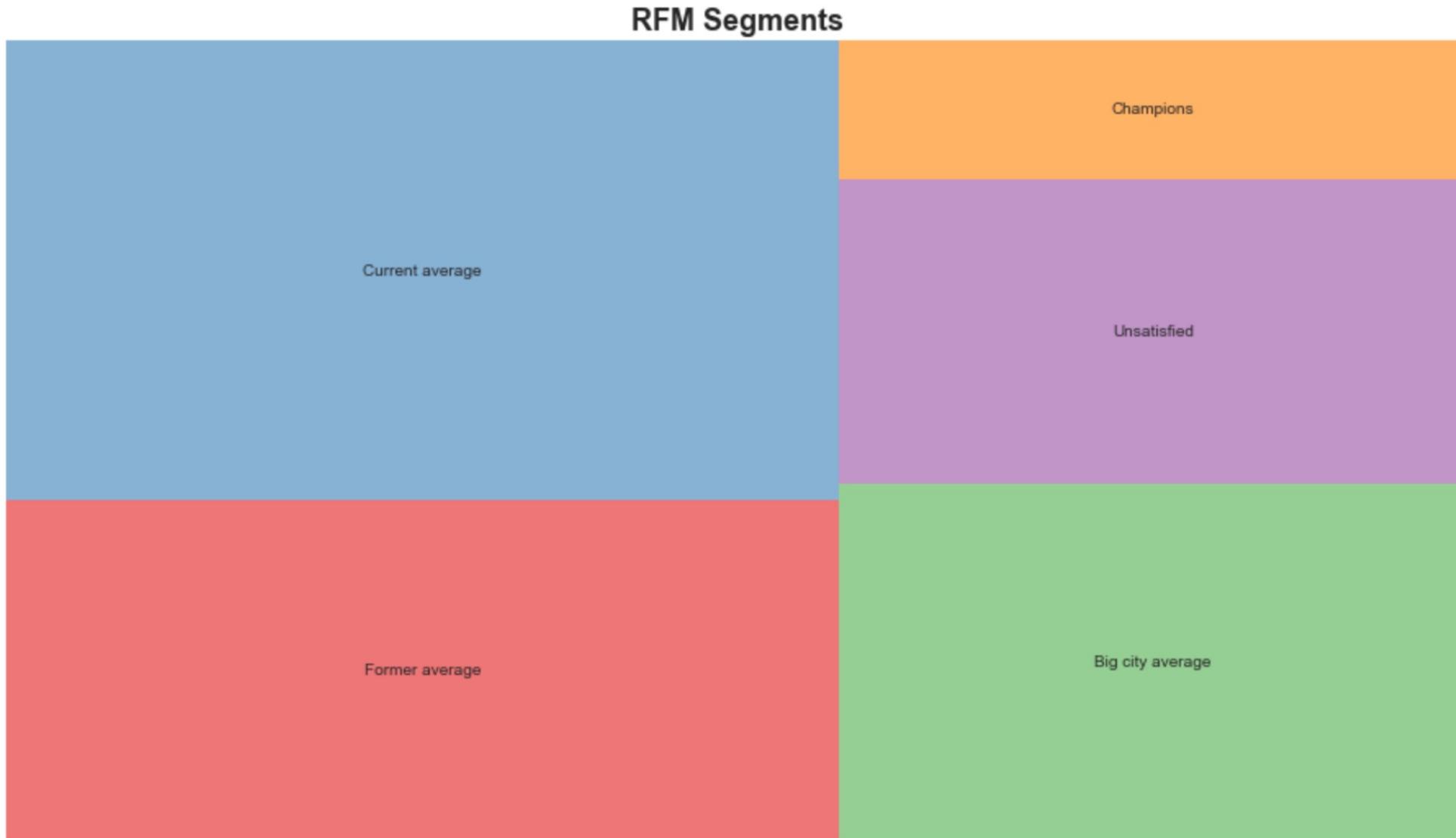
## Caractéristiques:

- Recency de 7 à 601 jours
- Frequency de 4 commandes
- Monetary value entre 682 et 1182 R\$
- Satisfaction 74 % score > 4/5
- Villes 2 millions d'habitant

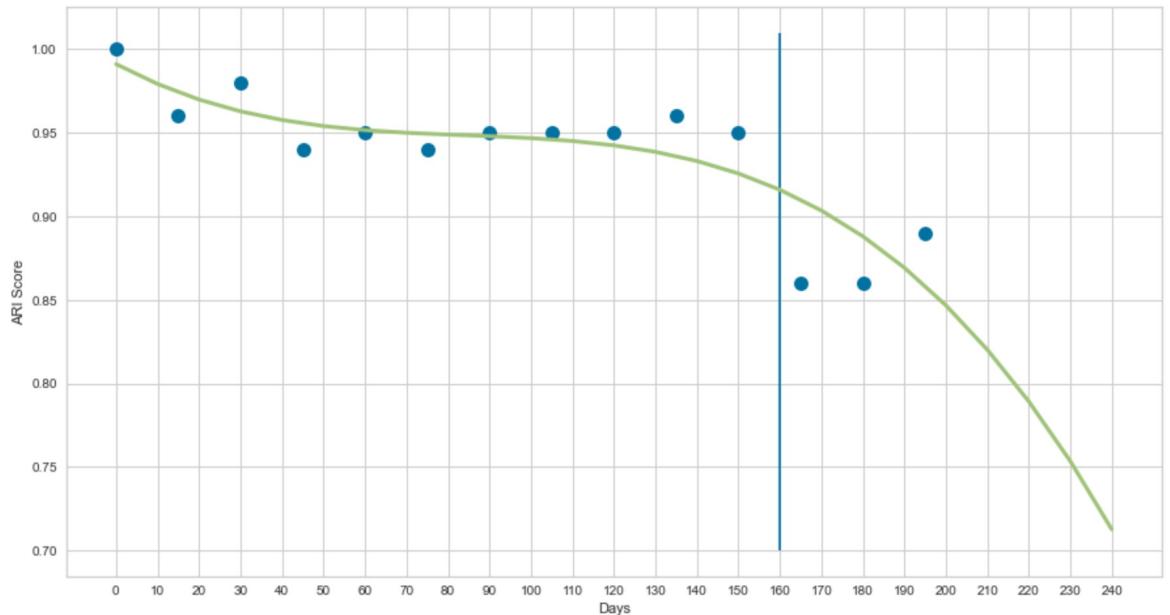
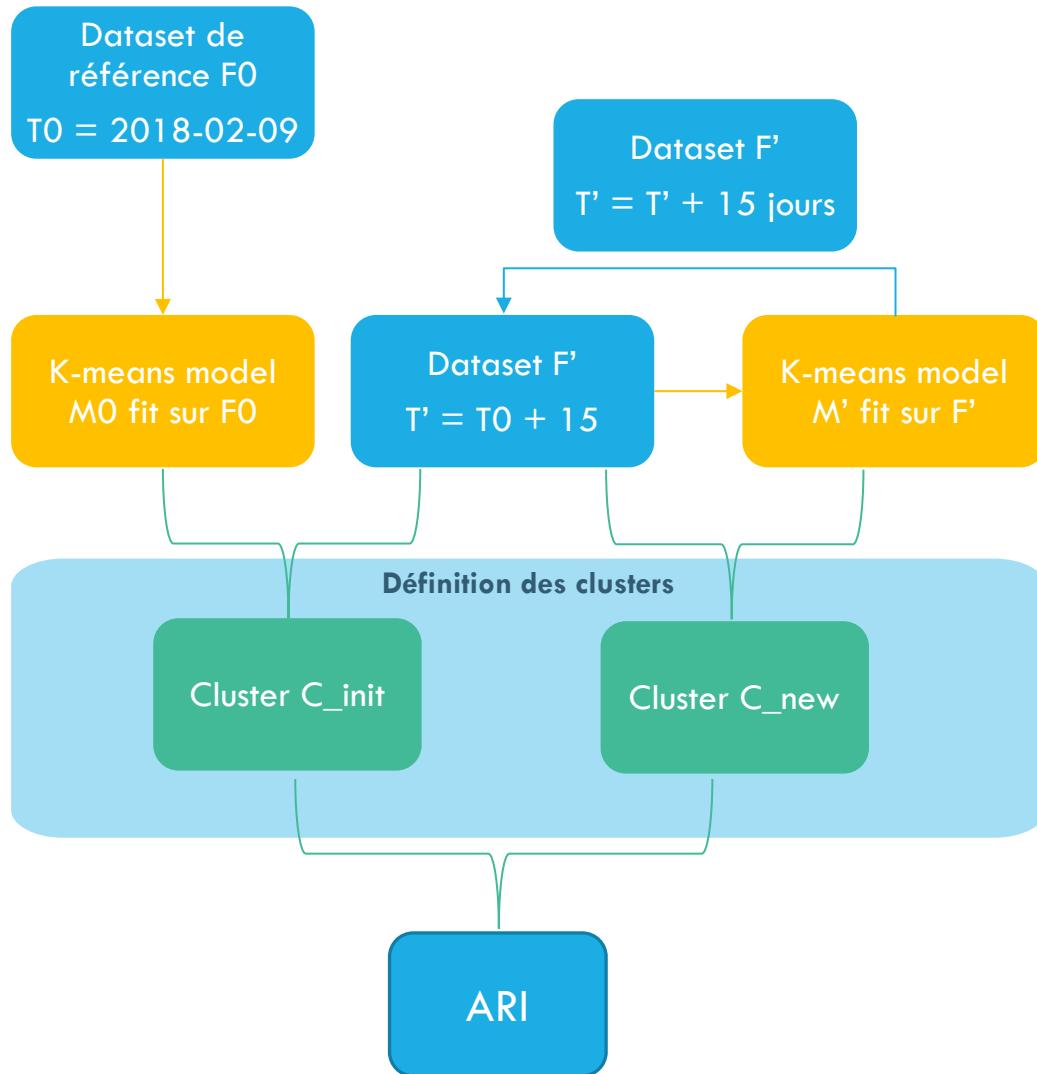
Cluster's profile comparison



# RÉPARTITIONS DES CLUSTERS



# PROPOSITION D'UN PLAN DE MAINTENANCE



- Le modèle est très stable
- plan de maintenance:
  - ARI > 90% → 5 mois

# BONUS

Déploiement d'une WebApp via Streamlit : [ici](#)

Création d'une API avec fastAPI et déploiement dans un container Docker

**Merci pour votre attention.**