



# Proposition de segmentation de la clientèle Olist

---

LO Ousmane, Machine Learning Engineer



01

## Introduction

Définition du problème,  
présentation des données

03

## Les segmentations

RFM, K-Means, CAH et  
DBSCAN

02

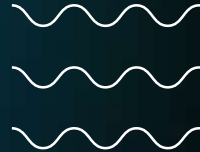
## Analyse descriptive

Analyses des comportements  
clients

04

## Conclusion

Personae, maintenance,  
avantages et inconvénients






# 01

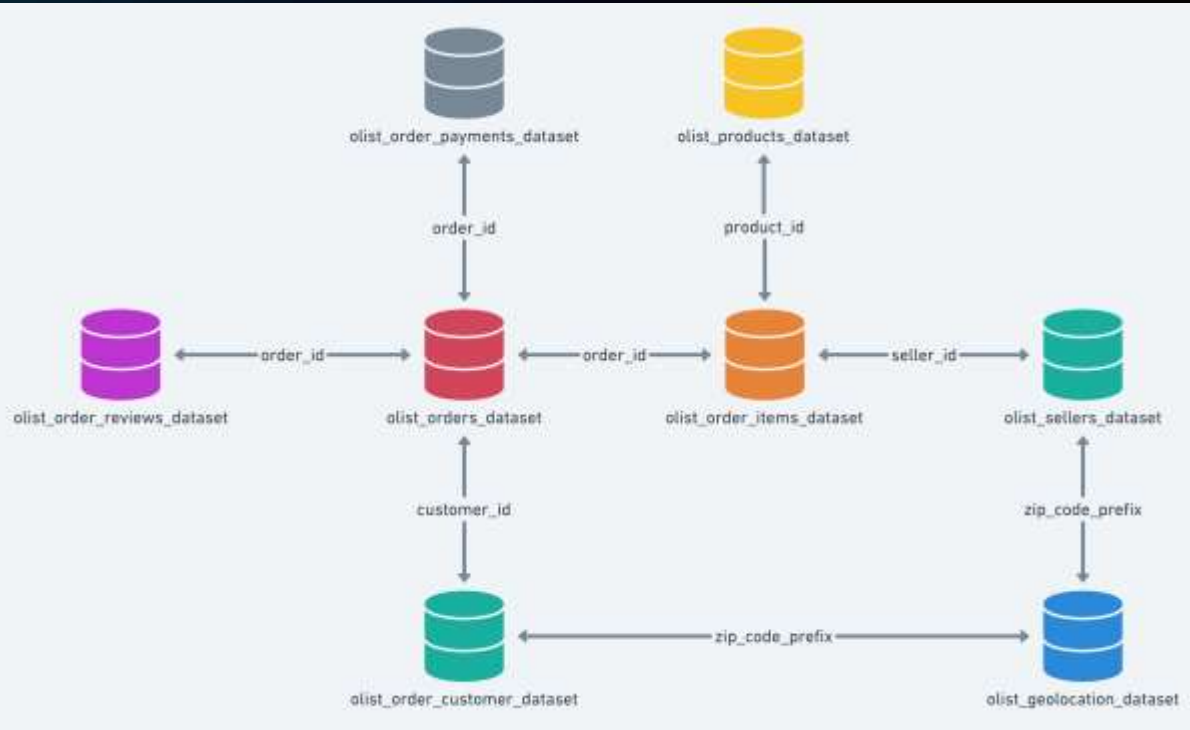
# Introduction

---

Passer d'une base de données non organisés à  
des segments actionnables

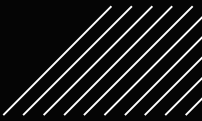


# Des sources de données multiples



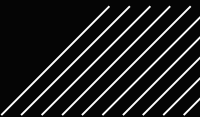
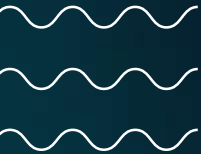
Exploration des données  
(Notebook « Etude  
Preliminaire »)

Analyse statistiques sur  
certaines données utiles pour la  
segmentation  
(Notebook « Notebook  
Analyse »)



# Objectifs

- Décrire les habitudes des clients avec des indicateurs simples
- Utiliser ces habitudes pour établir des groupes de clients





# 02



## Analyse descriptive



---

Mieux comprendre les comportements clients  
pour mieux les regrouper



# Analyses des clients

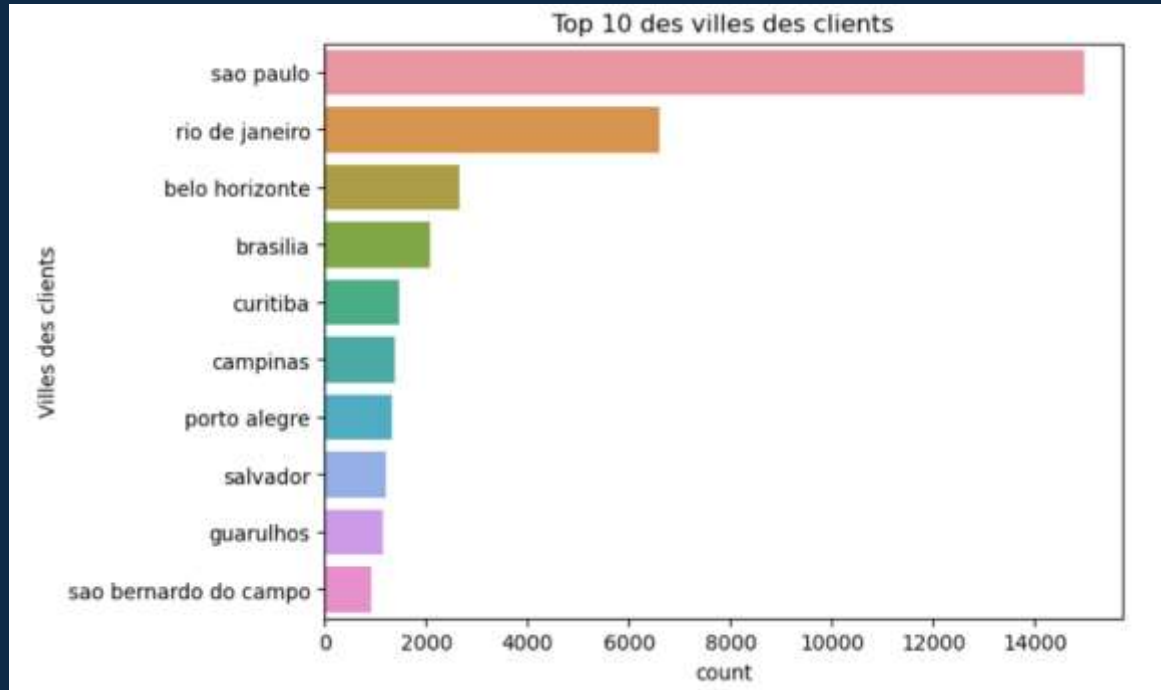
96 096 clients selon unique id

99 441 clients selon id

Conservation des id uniques

Affichage uniquement du  
Top 10 des villes

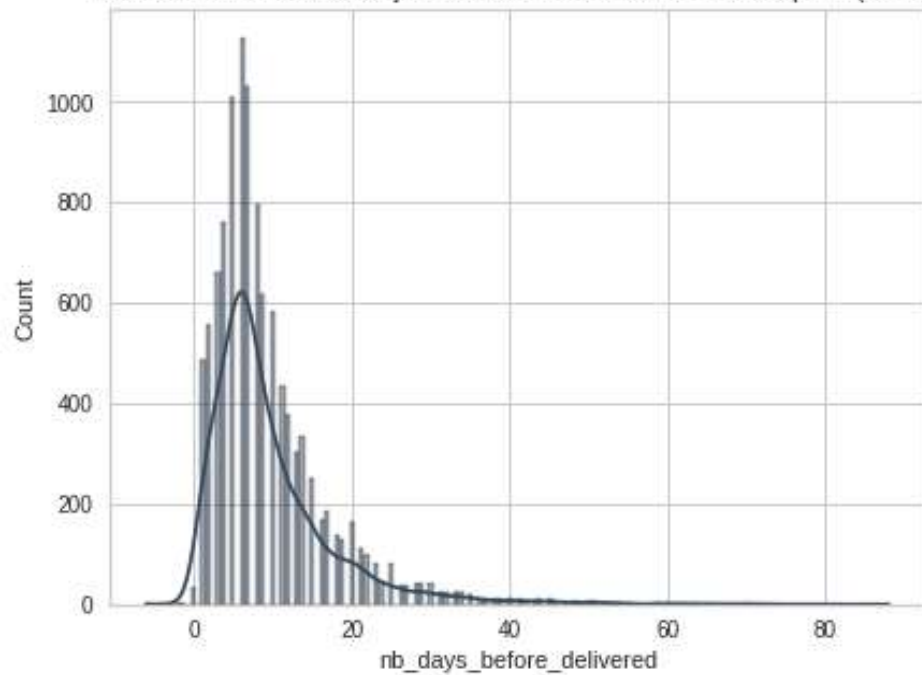
Analyse de la fréquence :  
Un achat dans 90%  
Conservation des clients  
avec minimum 2 achats



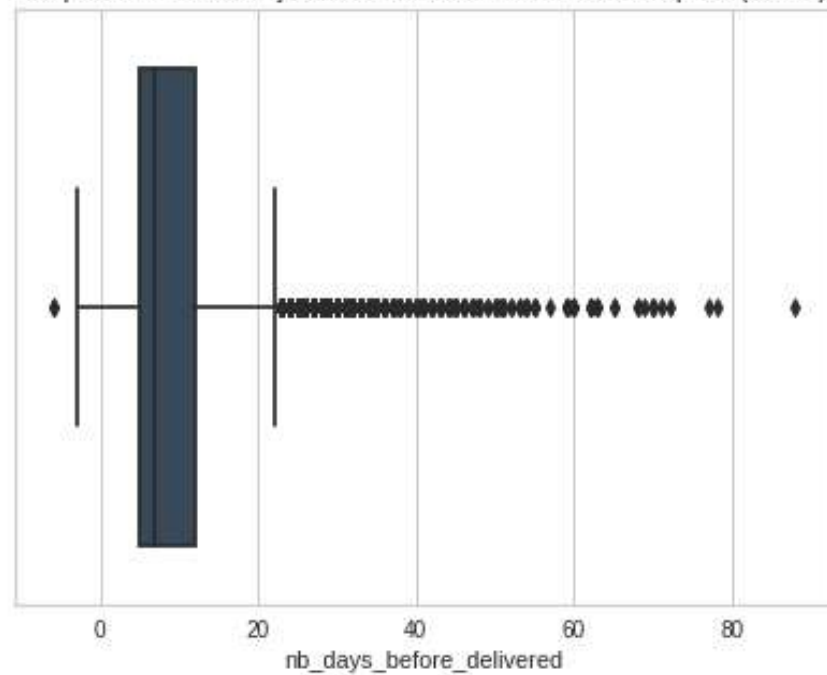
# Délai de livraison

Description du nombre de jours entre la commande et la réception (réalité)

Distribution du nombre de jours entre la commande et la réception (réalité)



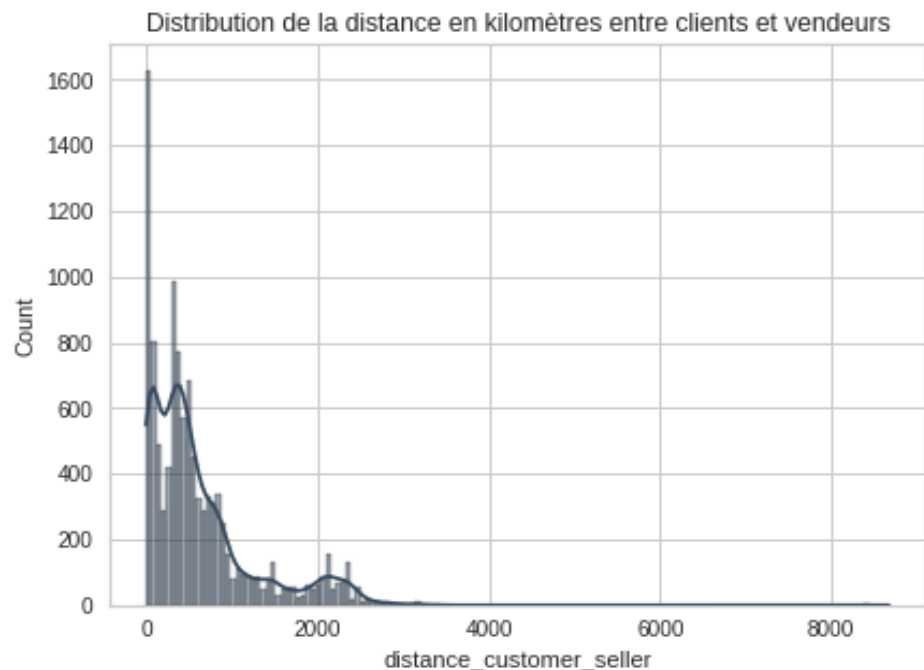
Boxplot du nombre de jours entre la commande et la réception (réalité)



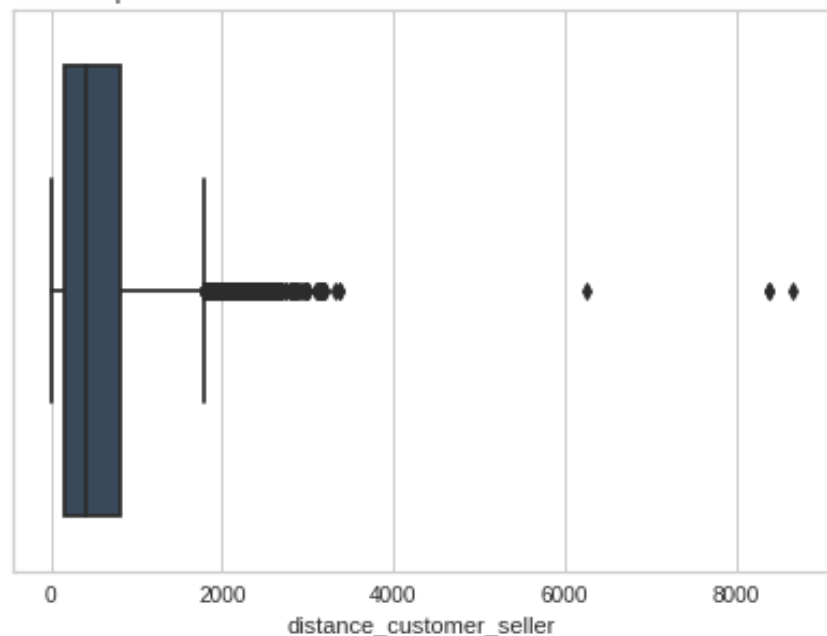


# Distance clients/vendeurs

Description de la distance en kilomètres entre clients et vendeurs

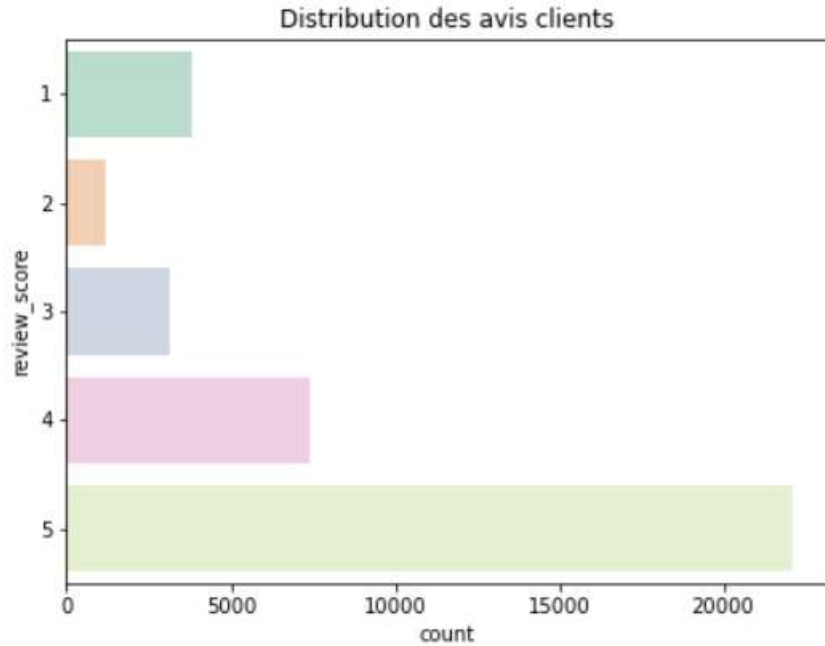


Boxplot de la distance en kilomètres entre clients et vendeurs

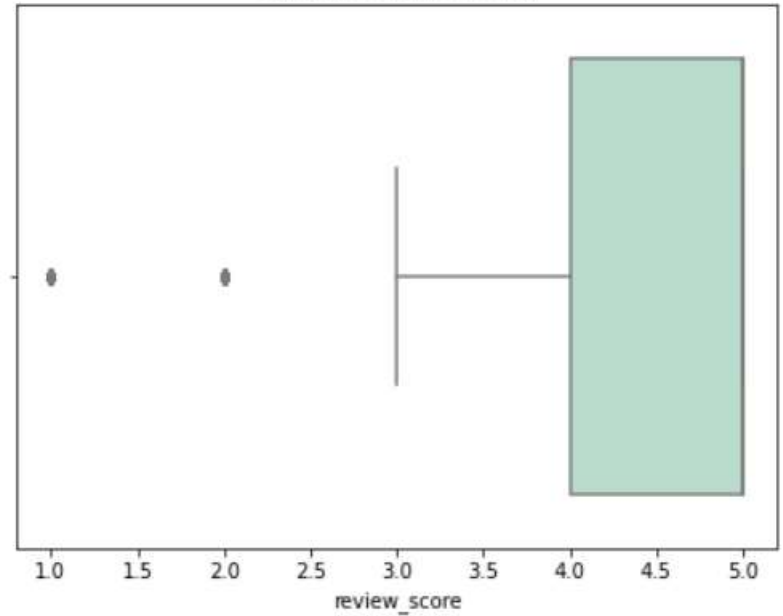


# Avis des clients

Description des avis clients



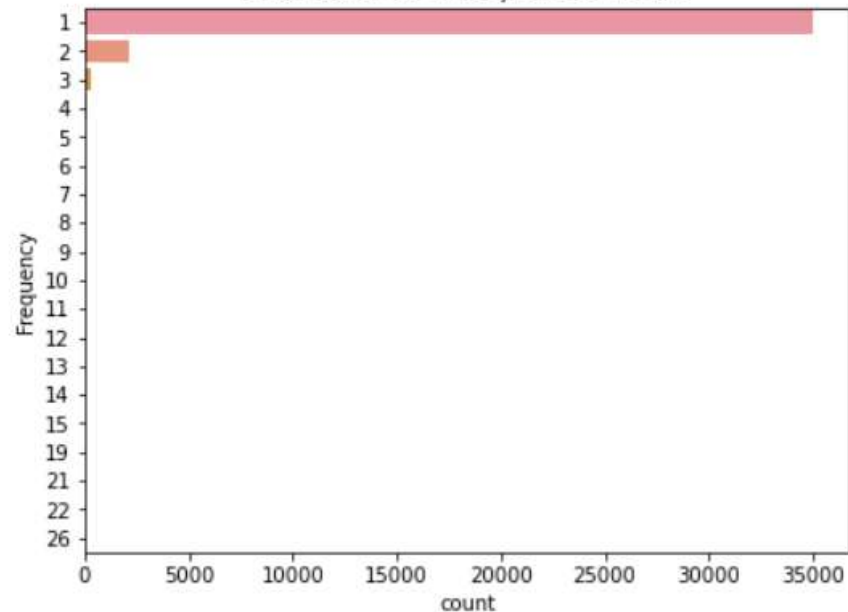
Boxplot des avis clients



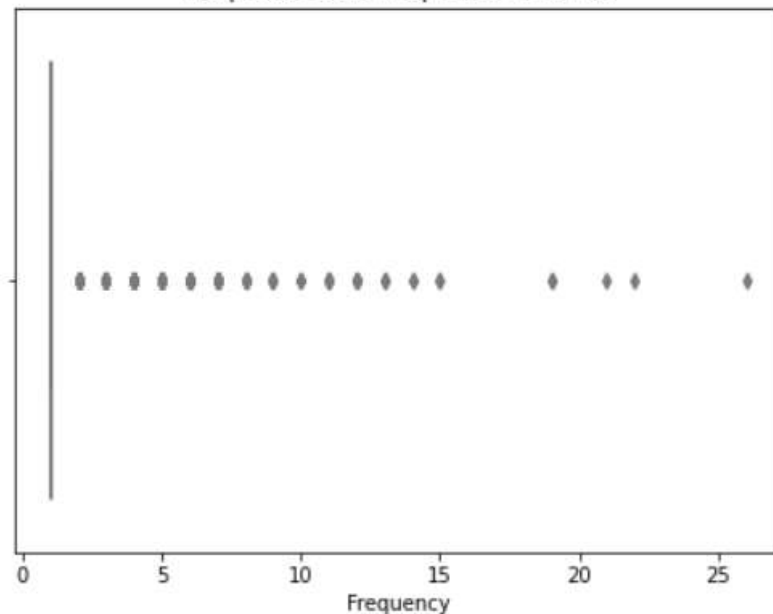
# Fréquence des achats

Description de la fréquence d'achats

Distribution de la fréquence d'achats



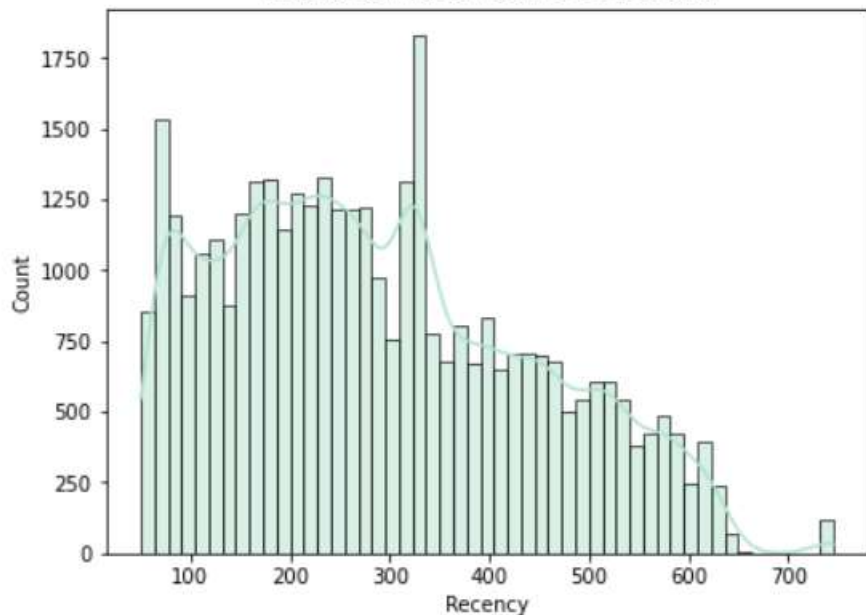
Boxplot de la la fréquence d'achats



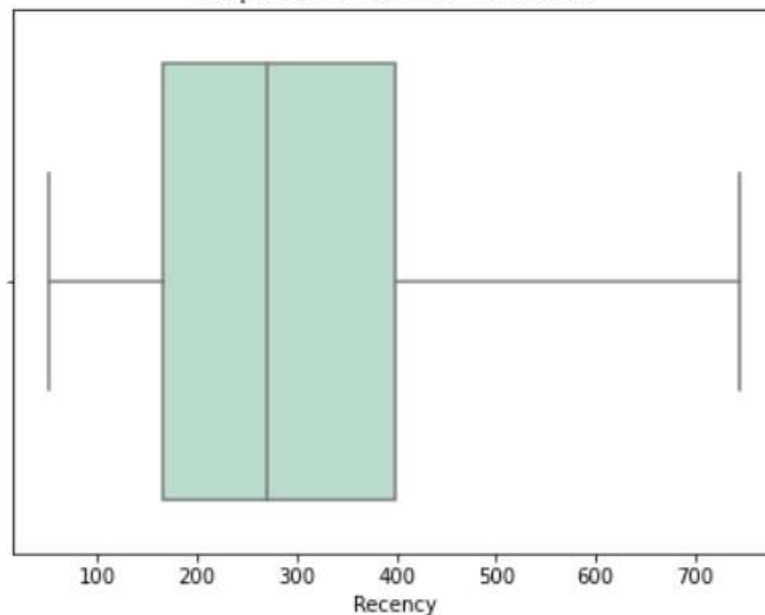
# Récence des achats

Description de la récence des achats

Distribution de la récence des achats

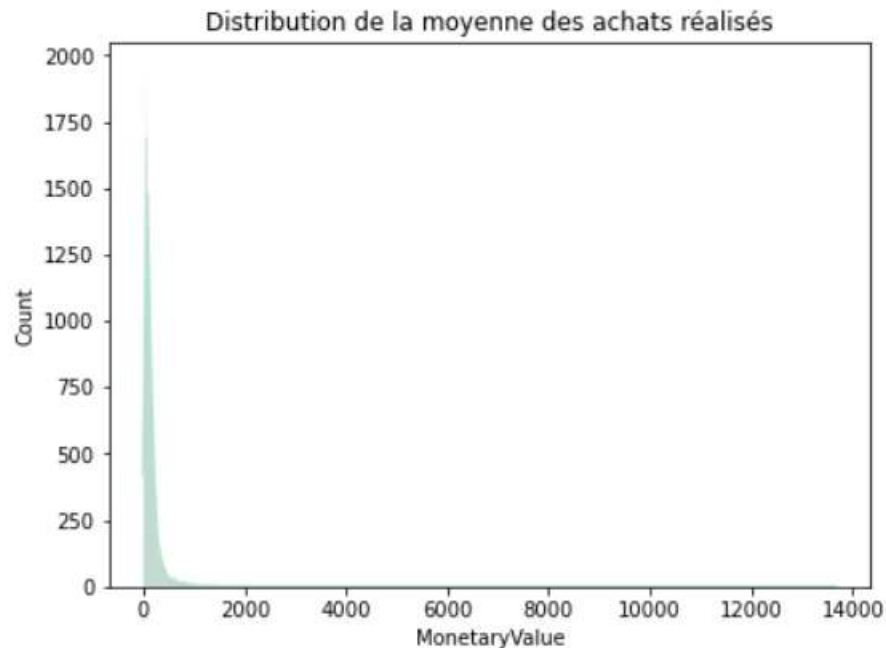


Boxplot de la récence des achats

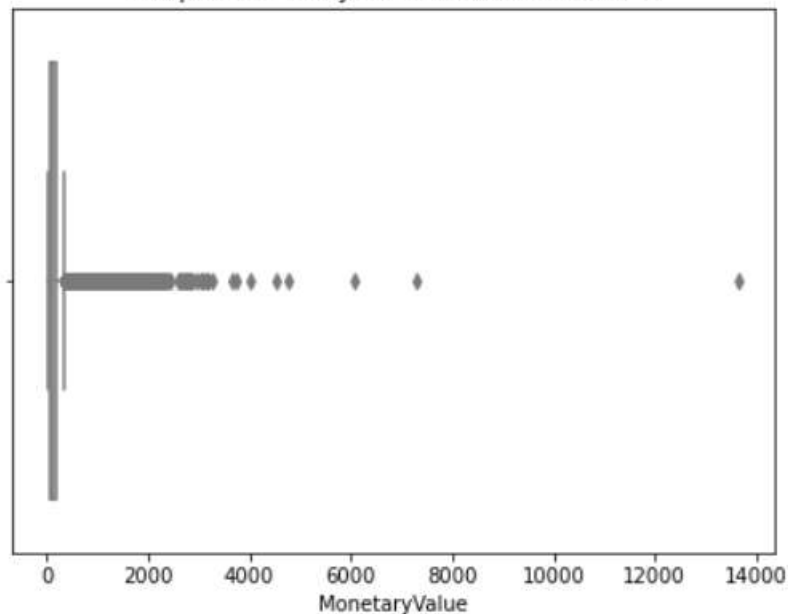


# Valeurs moyennes des achats

Description de la moyenne des achats réalisés



Boxplot de la moyenne des achats réalisés





# 03

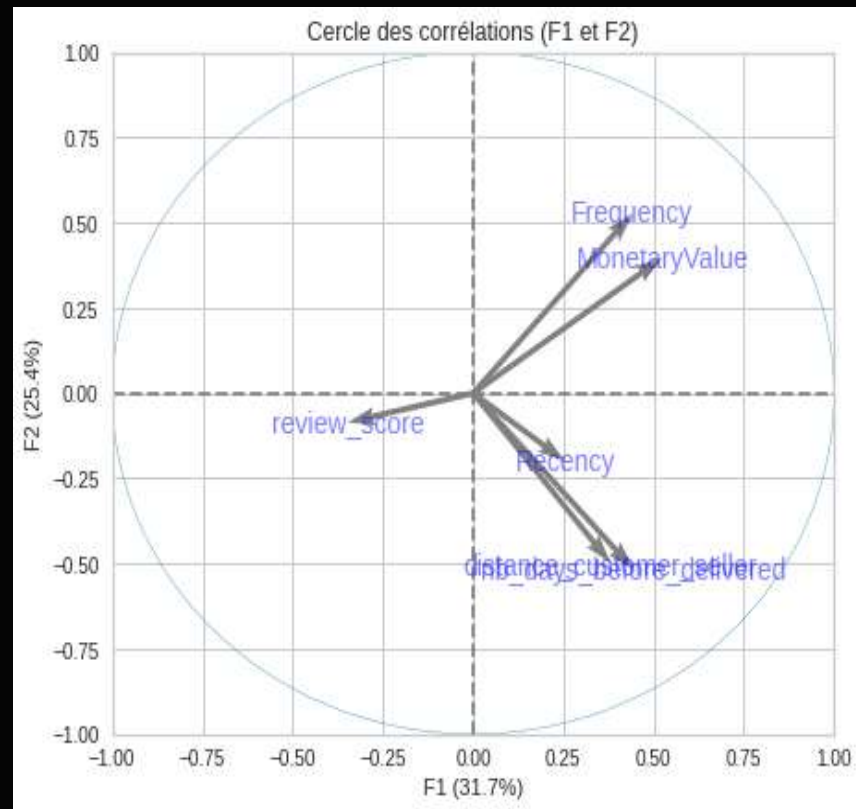
## Les segmentations

---

Segmentation RFM, K-Means, CAH, Dbscan

# Segmentation RFM

- Variables :  
Recency, Frequency, MonetaryValue, Reception Delay, Review Score, Distance between customers and sellers
- Transformation des variables en loglp
- Normalisation des variables
- Réduction de dimensions via une ACP
- Visualisations des clusters avec ACP

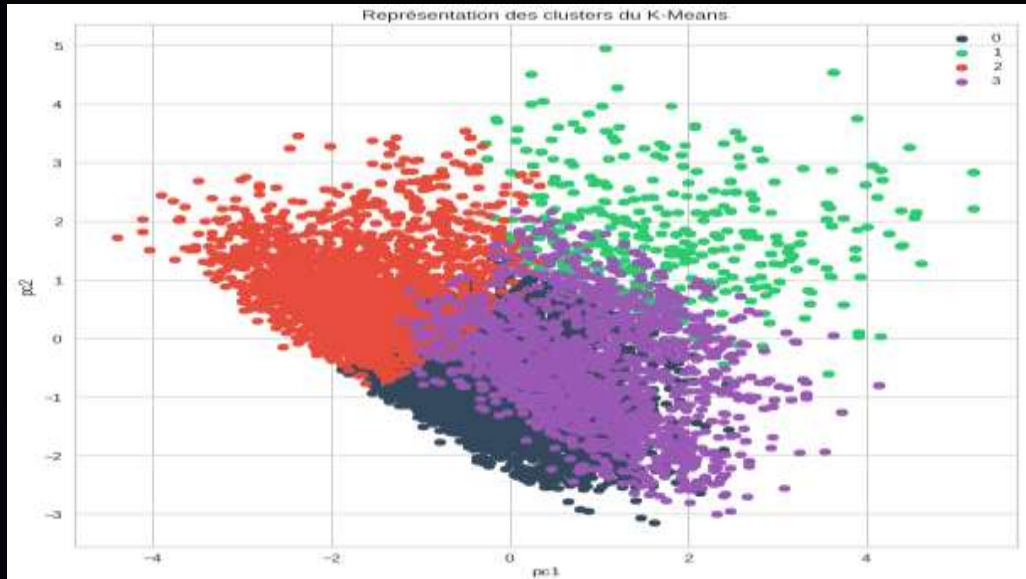
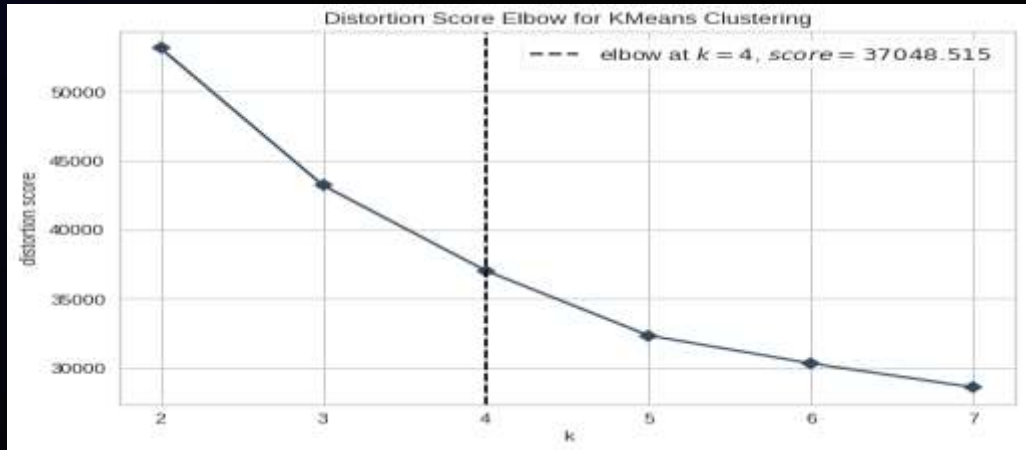


# K-Means (avec $K = 4$ )

- Détermination d'un nombre de groupe ( $K = 4$ )
- Visualisation en 2D via ACP
- Performance:

score de silhouette: 0.24

score de davies: 1.39



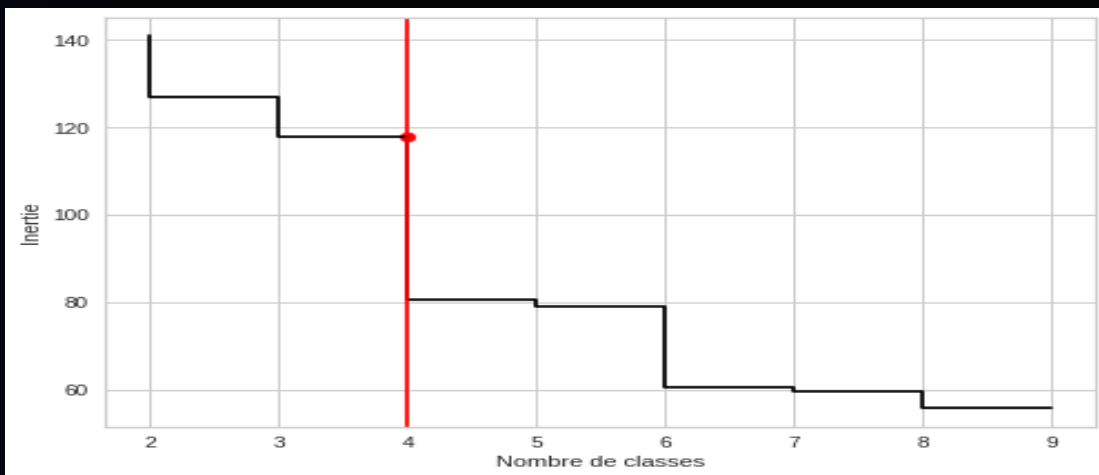
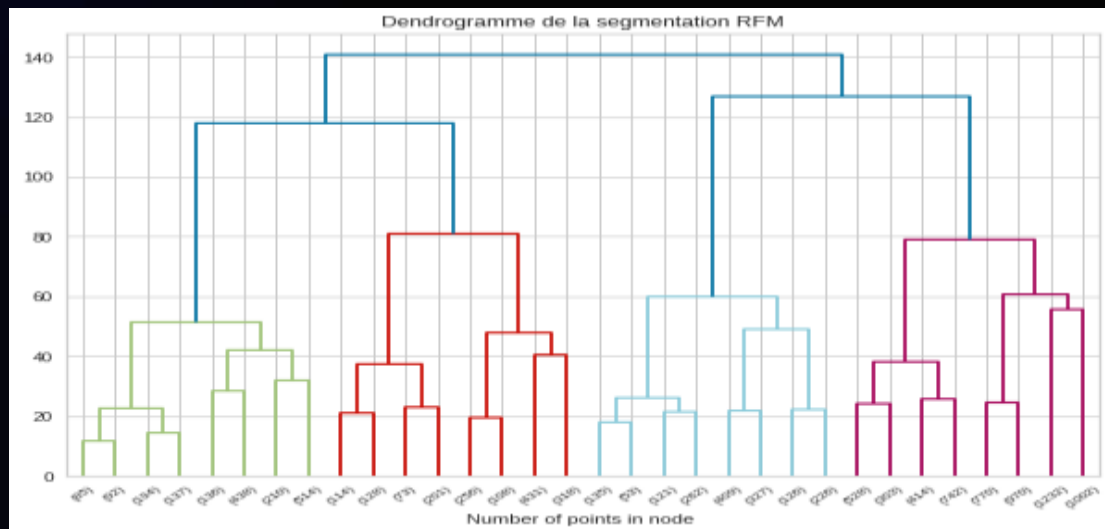


# CAH (avec $K = 4$ )

- Détermination d'un nombre de groupe ( $K = 4$ )
- Visualisation des clusters
- Performance:

score de silhouette: 0.22

score de davies: 1.49



# DBSCAN

(avec  $K = 4$ )

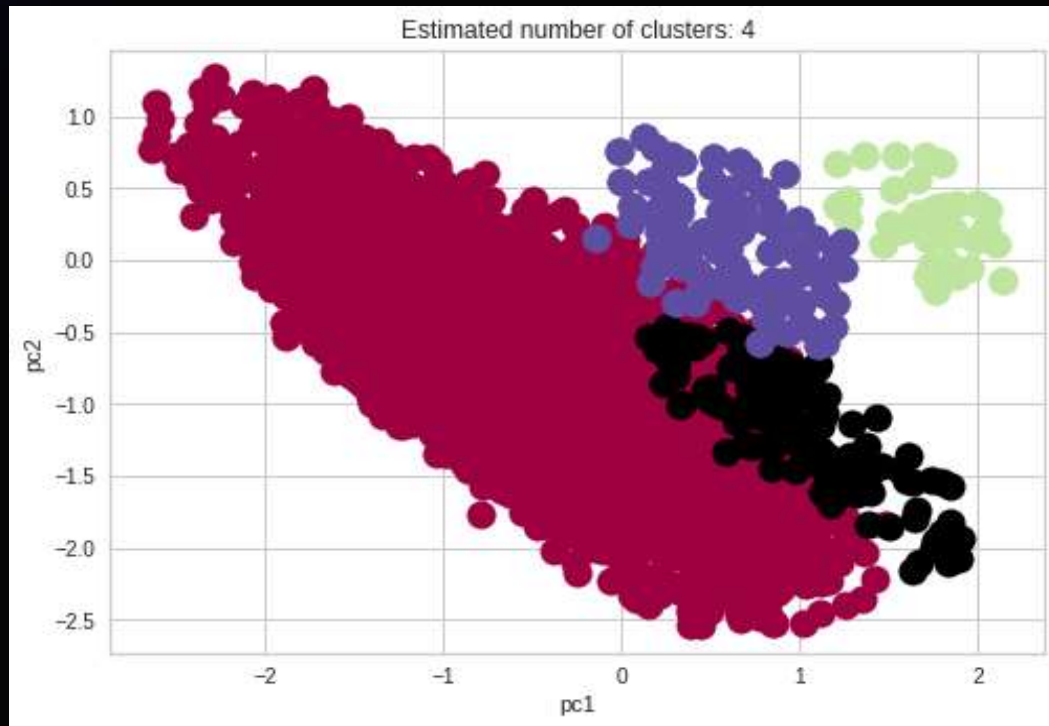
- Détermination d'un nombre de groupe ( $K = 4$ )

- Visualisation en 2D via ACP

- Performance:

score de silhouette: -0.007

score de davies: 2.5



# Comparaison des 3 modèles

	Model	silhouette_score	davies_bouldin_score
0	kmeans	0.248553	1.389337
1	CAH	0.216383	1.486746
2	dbscan	-0.007434	2.511110



- K-means plus performant sur les données

Modèle choisit : K-means

# Description des clients

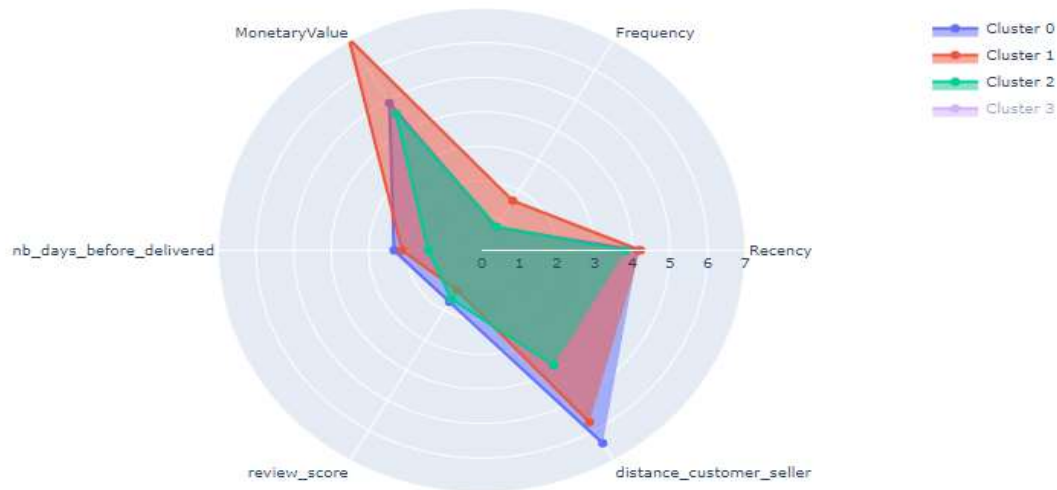
## Le client actif

- Plus de 6 achats
- Achat de fortes valeurs (> 1300 Réaux)
- 4.5/5

## Le client satisfait

- Plus de 2 achats
- Achats de valeurs moyennes
- 5/5

Analyse de la segmentation K-Means



## Le client moyen

- Plus de 2 achats
- Achat de valeurs moyennes
- 4.8/5
- 5 jours de délai

## Le client mécontent

- Plus de 2 achats
- Achat de valeurs fortes (>1097 Réaux)
- 2/5
- 13 jours de délai

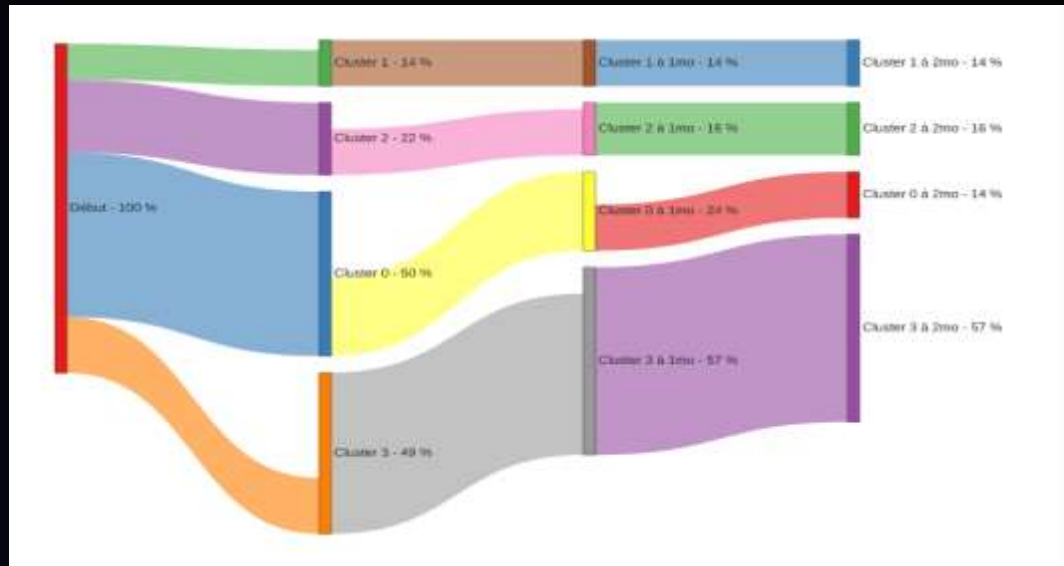
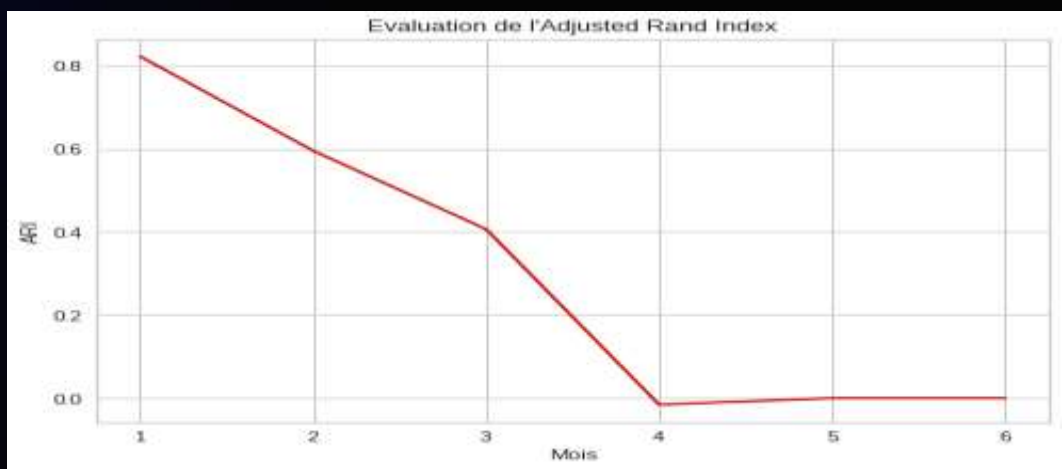
# Cartographie des clients par cluster

- Clients répartis majoritairement sur la cote
- Certains clients actifs (g1) sur les terres
- Plus part des clients mécontents (g3) sur SAO PAULO



# Maintenance

- Evaluation de la stabilité :
  - Score ARI sur une base de clients pendant 6 mois
  - Description des flux via diagramme de Sankey





# 04



## Conclusion



---

Quel segmentation choisir et perspectives



# Conclusion

- K-means :
  - Modèle avec meilleures performances, donc Plus adapté sur les données
  - Modèle stable sur 3 mois → Nécessite pas beaucoup de maintenance
- 4 profils de clients : Client actif, satisfait, moyen et mécontent
- Clusters 0 (clients satisfaits) et 2 (moyens) se vident aux profits du cluster 3 ( clients mécontents)
- Possibilité de mettre en place des stratégies différentes
  - Trouver une méthode pour « traquer » les clients pour avoir une meilleure estimation de la fréquence d'achat et de la récurrence
  - Utilisation de cookies
  - Création de compte
  - Récompenser les clients qui donnent un avis (pour augmenter le nombre d'avis)





CentraleSupélec

# THANKS



**CREDITS:** This presentation template was created by **Slidesgo**, and includes icons by **Flaticon** and infographics & images by **Freepik**

