OpenClassrooms

Projet P5:

Segmentez des clients d'un site e-commerce

N. Alves

Segmentez des clients d'un site e-commerce

Sommaire

Segmentez des clients d'un site e-commerce

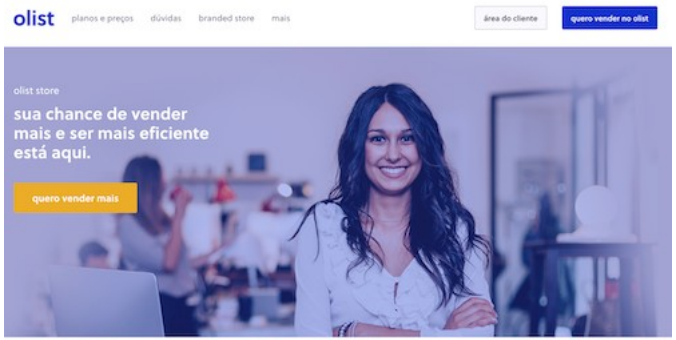
- **Contexte**
- Présentation des informations disponibles
- Étude du jeu de données, analyse
- Modélisation
- Conclusion et perspectives

Segmentez des clients d'un site e-commerce

Contexte

OLIST, solution de vente sur les places marchandes en ligne, a besoin d´une segmentation de sa clientèle pour ses campagnes de communication.

⇒ On souhaite définir une segmentation clientèle, sa logique sous-jacente éventuelle et une proposition de contrat de maintenance.



Site marchand OLIST

Segmentez des clients d'un site e-commerce

- Contexte
- **Présentation des informations disponibles**
- Étude du jeu de données, analyse
- Modélisation
- Conclusion et perspectives

Segmentez des clients d'un site e-commerce

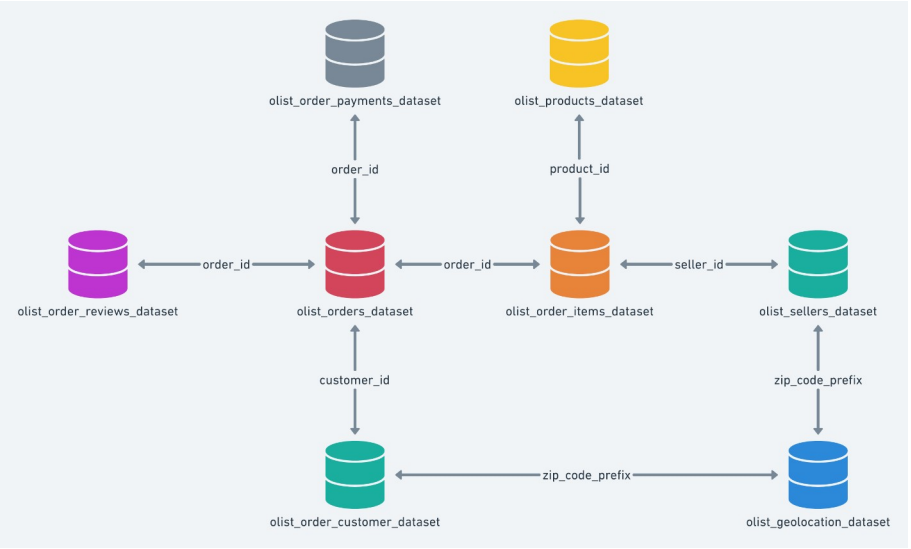
Présentation des informations disponibles

Huit fichiers au format ‘csv’ disponibles et joignables par une clé.

Ligne : article commandé

Colonne : caractéristique de la commande.

- lieu de livraison,
- date de commande et livraison,
- prix,
- catégories de produits,
- satisfaction client etc.



Les informations disponibles OLIST

Segmentez des clients d'un site e-commerce

- Contexte
- Présentation des informations disponibles
- **Étude du jeu de données, analyse**
- Analyse
- Modélisation
- Conclusion et perspectives

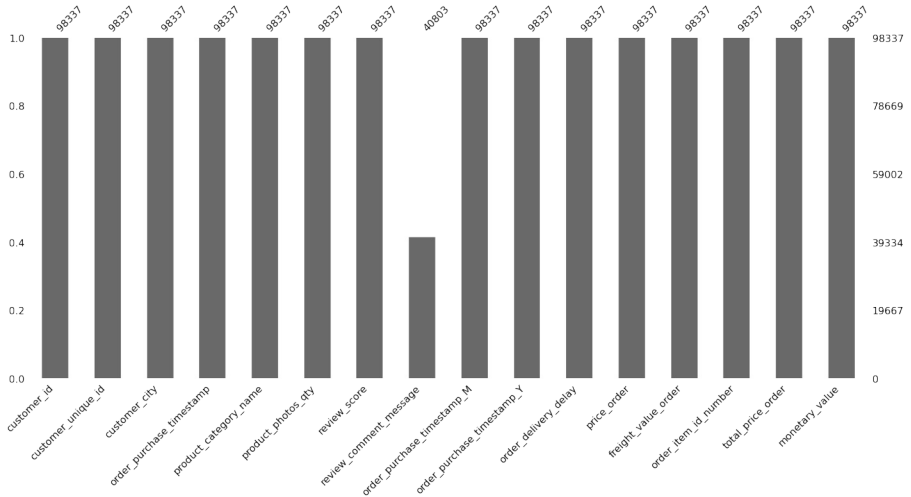
Segmentez des clients d'un site e-commerce

Étude du jeu de données

Création du jeu de donnée :

Fusion des fichiers en un seul jeu de donnée.

⇒ Après une première étude, on effectuera une sélection plus précise des données.



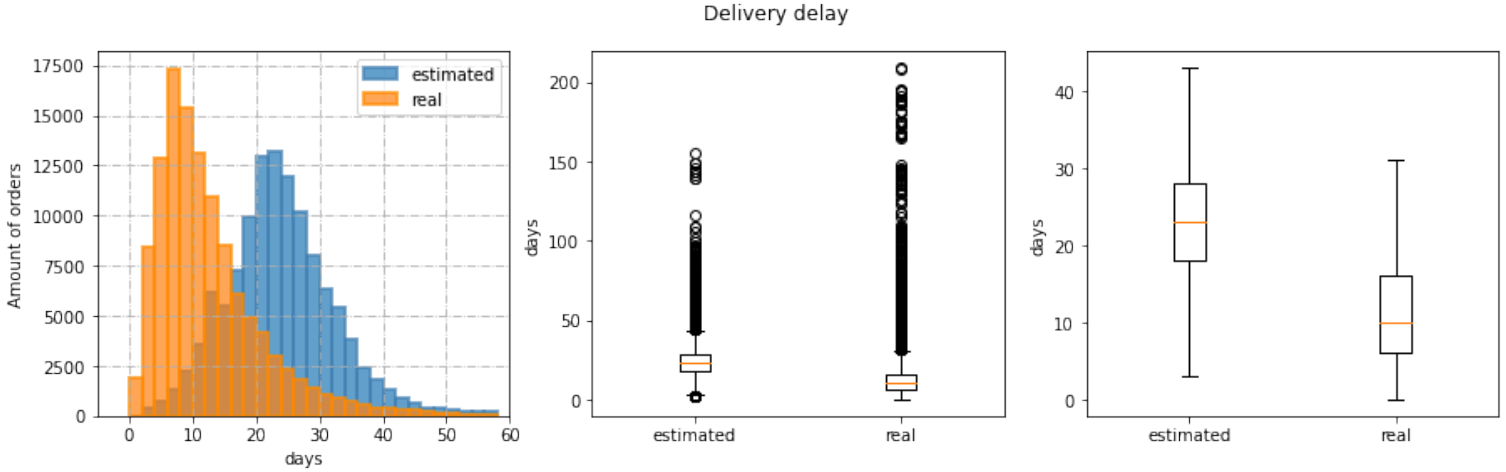
Les informations fusionnées, valeurs manquantes

Segmentez des clients d'un site e-commerce

Étude du jeu de données

Le délai de livraison : Cette variable est crée à partir des dates de commandes et de livraisons (réelles et estimées).

- 10 jours de livraison en moyenne
- Moins de 16 jours de délai pour 75 % des commandes et moins de 60 jours pour 99,7 % des commandes
- Les délais de livraisons réels sont inférieurs à ceux annoncés.

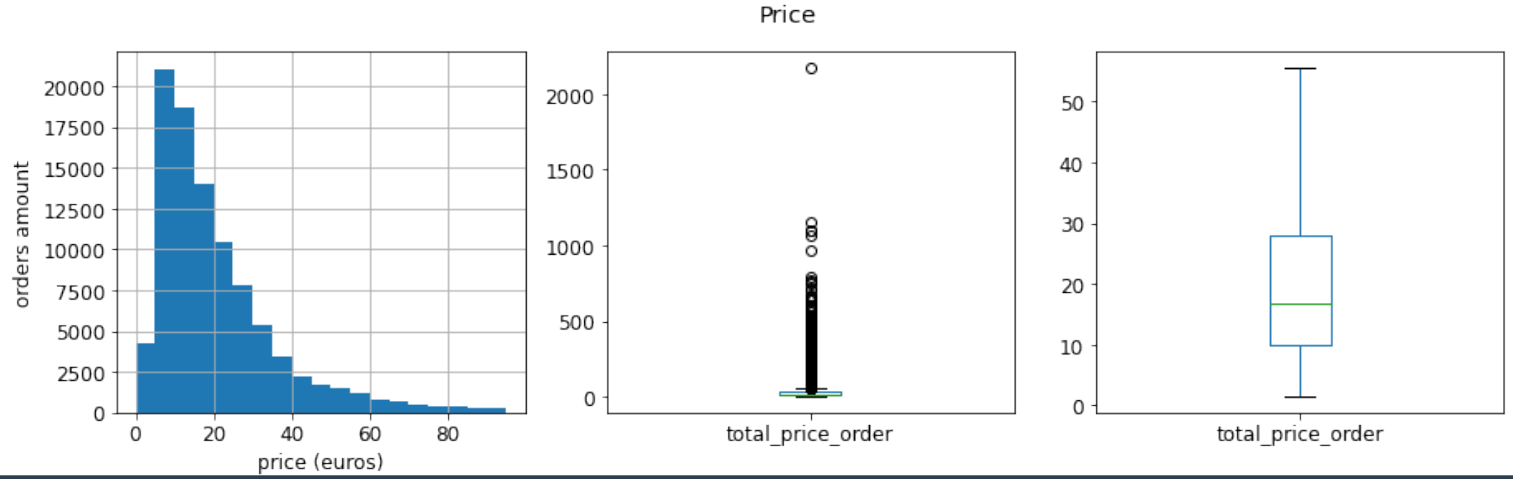


Segmentez des clients d'un site e-commerce

Étude du jeu de données

Le prix par commande :

- La valeur la plus fréquente se situe autour de ~10 euros
- 80 % des commandes ont une valeur de moins de 100 euros, 99 % des commandes ont une valeur de moins de 600 euros.



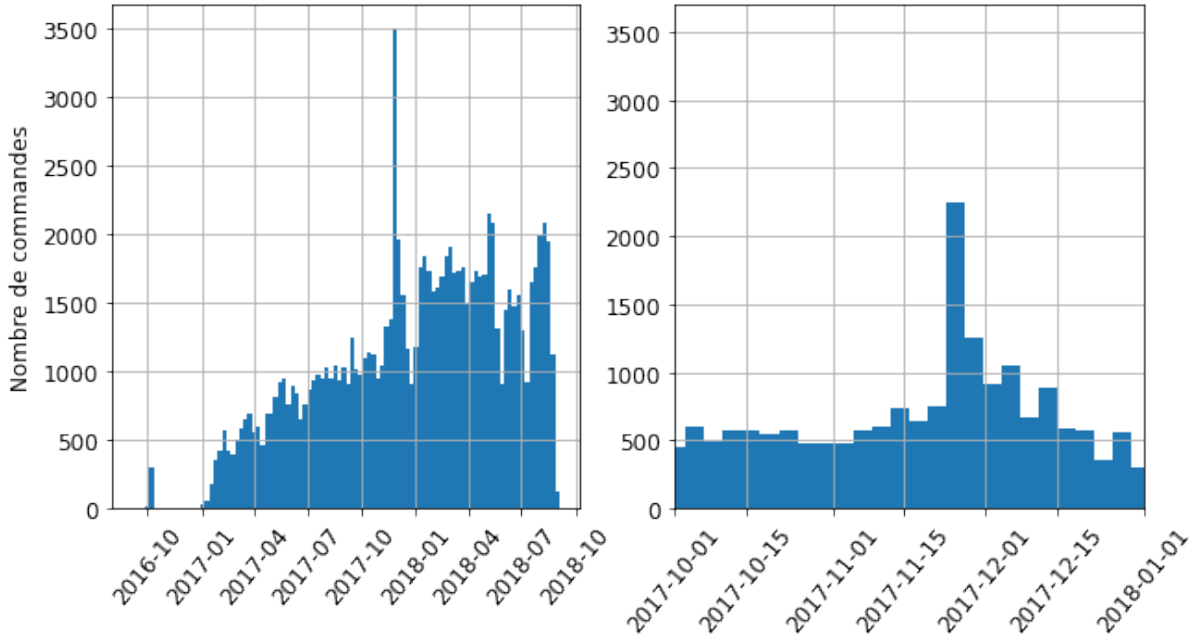
Segmentez des clients d'un site e-commerce

Étude du jeu de données

Les dates de commandes:

Le jeu de donnée est distribué sur une durée variant de début 2017 à mi-2018.

Un pic de vente est présent mi-novembre 2017.



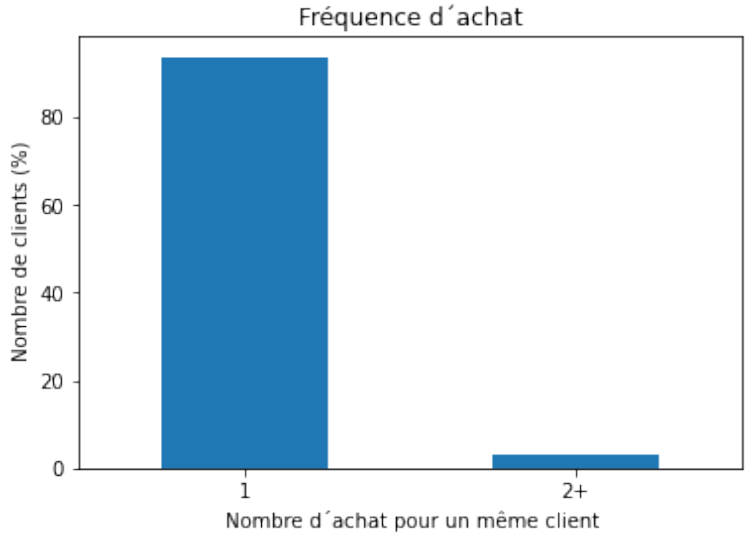
Segmentez des clients d'un site e-commerce

Étude du jeu de données

Caractéristiques complémentaires:

La fréquence client : 3 % des clients sont enregistrés comme ayant effectué plus de deux achats.

Le nombre d´article : 90% des commandes sont composées d´un seul article



Segmentez des clients d'un site e-commerce

Étude du jeu de données

Imputation :

Les dates de livraison non renseignées ont été remplacés par les dates de livraison estimées.

Valeurs extrêmes :

- Les commandes d´une valeur de plus de 600 euros ont été retiré.
- Les commandes avec un délai de livraison de plus de 60 jours ont été retiré.

Les variables sélectionnées:

Satisfaction client, délai de livraison, frais de transport, nombre d´article par commande, nombre de photos du produit, panier moyen, catégorie de produit.

⇒ **Après traitement, le jeu de donnée comporte comporte 96 500 lignes et 7 colonnes.**

Segmentez des clients d'un site e-commerce

- Contexte
- Présentation des informations disponibles
- Étude du jeu de données, analyse
- **Modélisation**
- Conclusion et perspectives

Segmentez des clients d'un site e-commerce

Modélisation

Méthodes : réduction dimensionnelle et algorithme de partition des données.

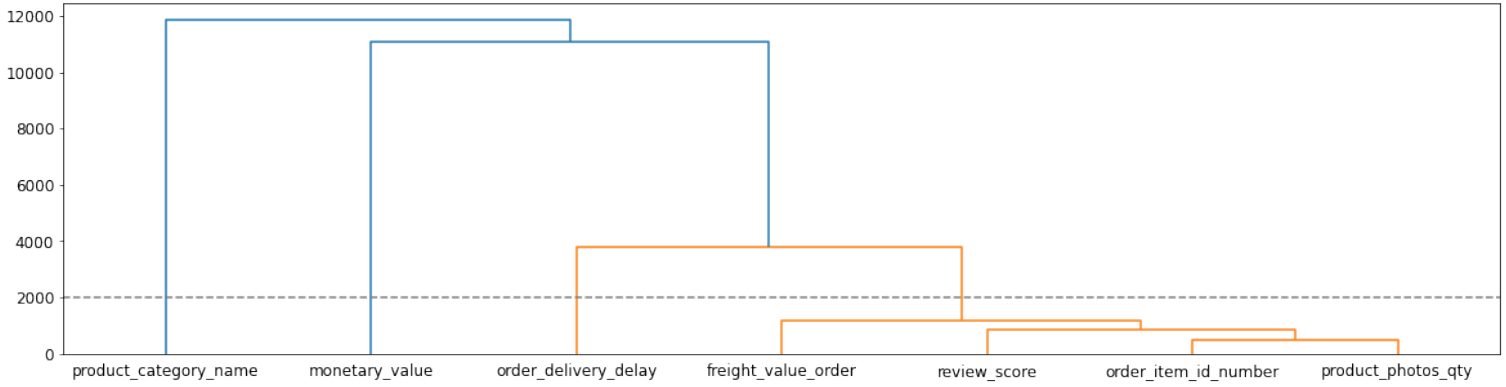
- Réduction dimensionnelle : analyse en composante principale et factorisation de matrice non-négative
- Partition des données : Kmeans et clustering hiérarchique
- Indicateur statistique : silhouette et indice de Rand ajusté.

Segmentez des clients d'un site e-commerce

Modélisation

Clustering hiérarchique :

Le clustering hiérarchique permet de réaliser un regroupement par récurrence, les résultats obtenus sont moins intéressant ici que dans le cas du kmeans.



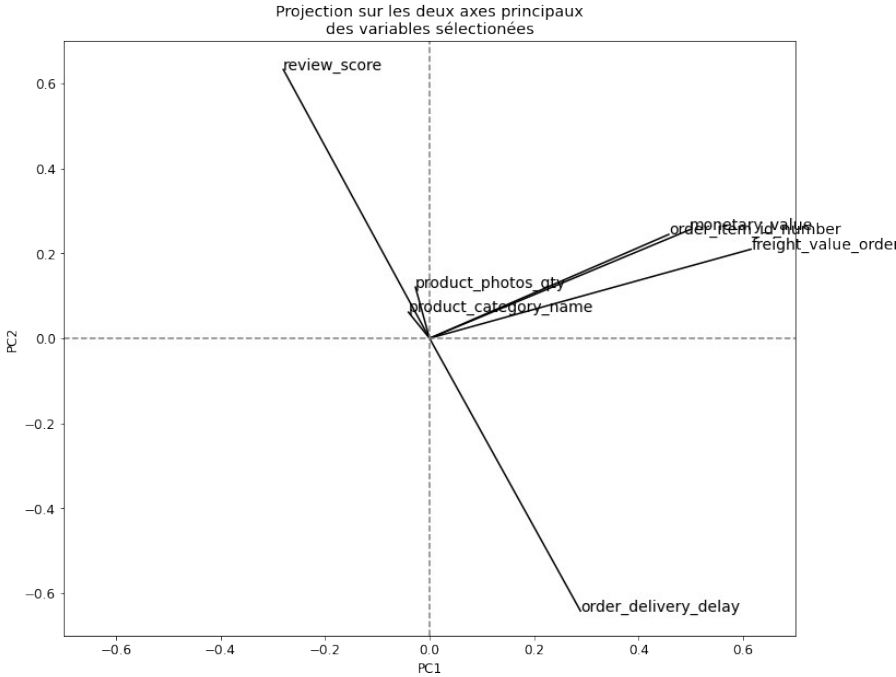
Segmentez des clients d'un site e-commerce

Modélisation

Réduction dimensionnelle

- ACP : nombre de 3, 4 et 5 composantes
  - ✔ Augmentation de variance expliquée de plus de 10%.
  - ✔ Variance totale entre 60 et 85%
- Projection sur PC1 : la satisfaction client et le délai de livraison en sens contraire, influence importante des autres variables aussi.

⇒ **Réalisation d´une étude croisée avec le Kmeans pour choisir les paramètres**



Segmentez des clients d'un site e-commerce

Modélisation

Réduction dimensionnelle

La décomposition NMF permet de définir les variables latentes au modèle :

review_score	order_delivery_delay	freight_value_order	order_item_id_number	monetary_value	product_category_name	product_photos_qty
0.4	0.0	1.2	0.2	1.2	0.1	0.1
0.0	0.0	0.0	0.0	3.6	0.0	0.0
2.7	0.8	0.0	0.5	0.0	53.2	1.2
0.4	13.8	0.0	0.2	1.1	0.0	0.4

Variables latentes:

- prix et satisfaction, sans le délai de livraison
- prix
- produit, satisfaction, nombre de photos
- le délai de livraison

Segmentez des clients d'un site e-commerce

Modélisation

Partition Kmeans :

Étude paramétrique en croisant :

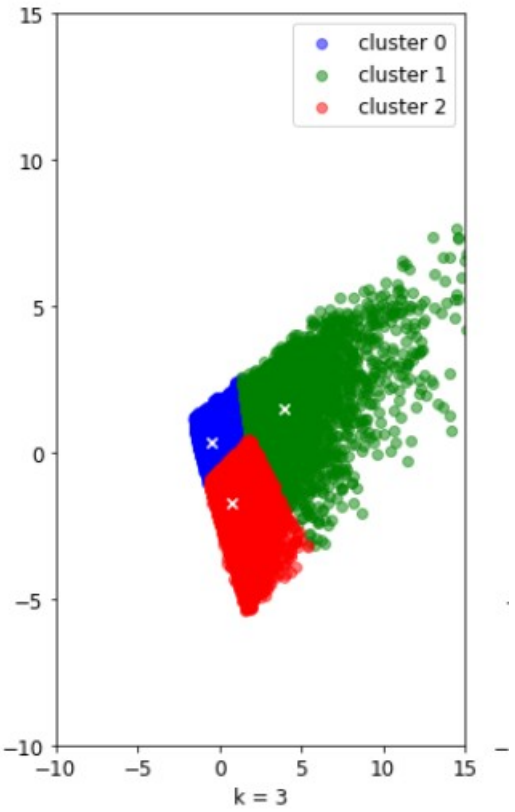
- le nombre de composantes de la réduction ACP (3,4 et 5 composantes)
- le nombre de partitions de l´algorithme Kmeans. (2 à 5 partitions)

Silhouette :

	k = 2	3	4	5
Silhouette	0.42	0.39	0.32	0.33

(3 composantes)

⇒ Le choix final s´est porté sur trois composantes et trois segments



Segmentez des clients d'un site e-commerce

Modélisation

La segmentation obtenue:

label	review_score	order_delivery_delay	freight_value_order	order_item_id_number	monetary_value	product_category_name	product_photos_qty	proportion (%)
0	5.0	9.0	3.0	1.0	20.0	health_beauty	1	75.0
1	4.0	13.0	12.0	2.0	103.0	furniture_decor	1	5.0
2	2.0	22.0	4.0	1.0	24.0	bed_bath_table	1	20.0

Segment 0 : 75 % de la population, panier de ~20 euros, satisfaction élevée, délai de livraison court.

Segment 1 : 5% de la population, panier de ~100 euros, satisfaction raisonnable, valeur de produit élevée.

Segment 2 : 20% de la population, panier de ~20 euros, satisfaction à améliorer, délai de livraison long.

Segmentez des clients d'un site e-commerce

Modélisation

Maintenabilité :

On s´intéresse à une évolution éventuelle du nombre de segments sur un semestre, ici le premier semestre 2018.

Évaluation de la concordance des segmentations, la mesure de performance ARI :

- Le passage de 2 à 3 segments est plus stable que les autres.
- Les passages de 2 à 4 segments ou de 3 à 4 segments sont équivalents.

	ARI
Clusters evolution 2 -> 3	0.78
3 -> 4	0.38
2 -> 4	0.34

Segmentez des clients d'un site e-commerce

Modélisation

- Utilisation de trois segments pour diversifier au mieux la clientèle, le modèle est entraîné sur le premier semestre 2018 à l´heure actuelle.
- Une maintenance de moins de 6 mois est conseillée pour actualiser les données.
- Entre ces périodes, il est aussi intéressant de prendre en compte les variables sous-jacentes pour renforcer la robustesse de la segmentation.

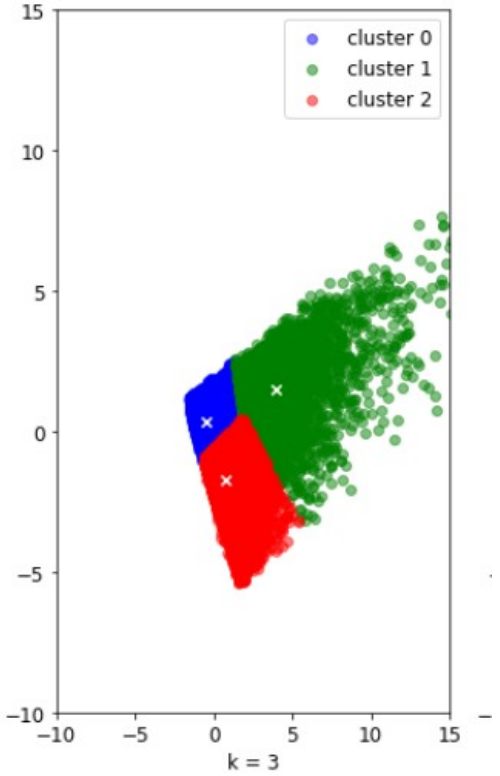
Segmentez des clients d'un site e-commerce

- Contexte
- Présentation des informations disponibles
- Étude du jeu de données
- Analyse
- Modélisation
- **Conclusion et perspectives**

Segmentez des clients d'un site e-commerce

Conclusion et perspectives

- Algorithme Kmeans avec une réduction dimensionnelle ACP
- Trois segments et une liste de variables latentes
- Période de maintenance de moins de six mois afin d´actualiser les données.



Merci de votre attention !