Analyse prédictive et segmentation de la clientèle e-commerce

Projet Olist

- Segmentation de la clientèle
- Analyse prédictive
- Exploitation des données transactionnelles

Réalisé par :
Darryl Momo
Demanou Levana
Laeticia Joyce Chuiedjui

Formation :
Data / IA – Analyse sectorielle
Année : 2024–2025

Contexte & Enjeux Business

Contexte

- **Olist** : plateforme e-commerce B2C
- **Forte volumétrie** de données transactionnelles
- **Marché** très concurrentiel

Message clé

Sans segmentation et analyse data, les décisions marketing sont génériques et peu rentables.

Enjeux Business

Mieux connaître les clients
Comprendre les comportements d'achat et les préférences

Identifier des segments à forte valeur
Distinguer des profils clients qui apportent la plus grande valeur

Optimiser les actions marketing et commerciales
Cibler efficacement les bonnes personnes avec les bons messages

Objectifs du Projet

Notre projet vise à exploiter les données e-commerce d'Olist afin d'obtenir des insights clients exploitables pour la prise de décision stratégique.

Comprendre les comportements clients

Analyser les patterns d'achat, préférences et tendances pour mieux anticiper les besoins

Segmenter la clientèle

Créer des segments data-driven pour ciblage personnalisé et optimisation des actions marketing

Relier les segments à des indicateurs métier

Connecter les segments à des métriques business pour mesurer l'impact stratégique

Préparer le socle pour des modèles prédictifs

Structurer les données et les insights pour des applications futures d'analyse prédictive

Message clé

L'objectif n'est pas juste de faire du clustering, mais de produire des segments compréhensibles et actionnables par les équipes business.

Données Utilisées

Sources de données Olist

Clients
Informations sur les clients

Commandes
Informations sur les commandes livrées

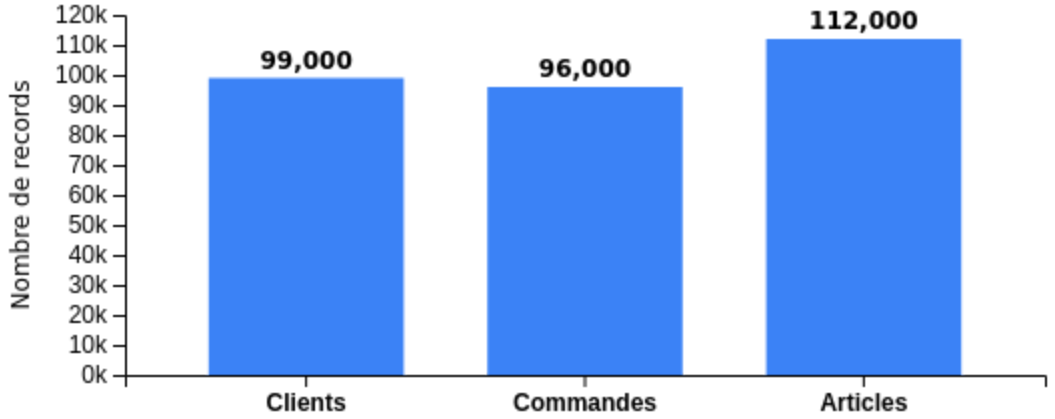
Articles commandés
Détails des articles dans chaque commande

Paiements
Informations sur les modes et statuts de paiement

Avis clients
Notes et commentaires des clients

Produits
Informations sur les produits vendus

Volumétrie des données



Message clé

Nous travaillons sur des données réelles, complexes et représentatives d'un e-commerce à grande échelle.

Architecture du Projet

Organisation professionnelle

data/raw
données brutes

data/interim
données nettoyées

notebooks
pipeline analytique structuré séparation claire des étapes

Approche

- Reproductible
- Modulaire
- Orientée métier

Message clé

Cette structure est proche des standards utilisés en entreprise data.

Nettoyage & Qualité des Données

Processus de Nettoyage

Suppression des doublons
Identification et removal des enregistrements dupliqués pour garantir la qualité des données

Gestion des valeurs manquantes
Analyse et traitement des champs vides ou null pour compléter les informations essentielles

Filtrage des commandes livrées
Restriction aux commandes ayant atteint leur statut "livrée" pour analyser uniquement les transactions completes

Vérification des clés critiques
Validation de l'intégrité des clés primaires et étrangères pour garantir la cohérence des relations

Résultat

- **Données fiables** pour l'analyse
- Aucun **biais majeur** introduit dans les résultats
- Base de données **prête pour l'analyse** approfondie

Message clé

Une segmentation n'a de valeur que si les données sont propres et cohérentes. Le nettoyage est la foundation de toute analyse data de qualité.

Analyse Exploratoire (EDA)

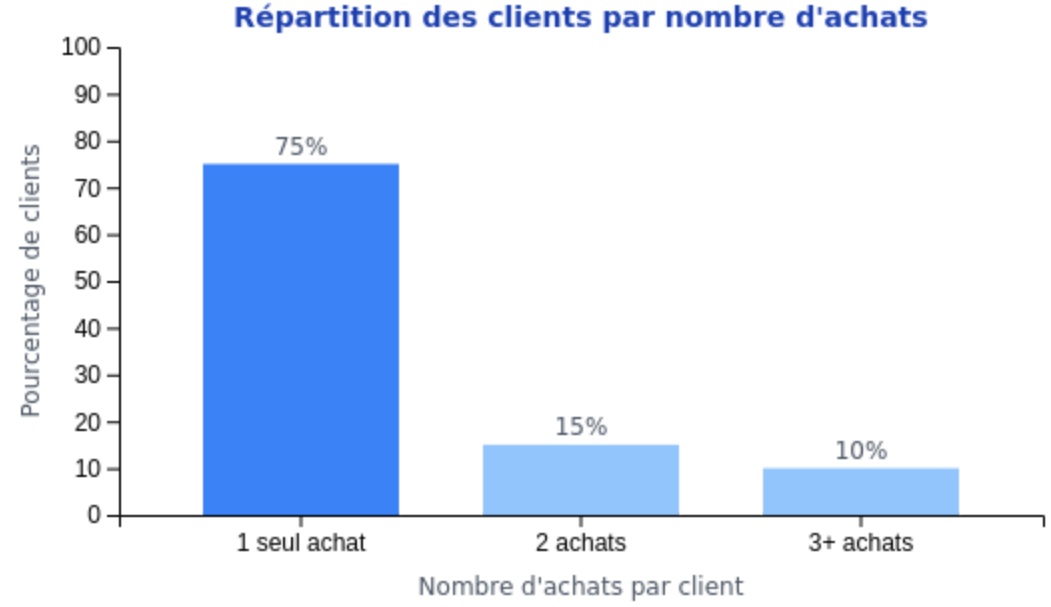
Observations clés

Panier moyen très hétérogène
Les tailles de paniers varient fortement selon les clients et les commandes

Majorité de clients à achat unique
La plupart des clients ne reviennent pas après leur première commande

Distribution asymétrique des dépenses
Une petite partie des clients génère une grande partie des ventes

Avis clients globalement positifs
Les commentaires laissés par les clients sont majoritairement favorables



Message clé

L'EDA confirme la nécessité d'une segmentation pour distinguer des profils clients très différents.

Construction des Features Clients

Dimension finale : 93 358 clients × 9 variables métiers

Analyses RFM

Recency

Temps écoulé depuis le dernier achat

Frequency

Nombre de fois que le client a acheté

Monetary

Somme totale dépensée par le client

Variables Complémentaires

Panier moyen

Valeur moyenne des commandes

Nombre d'articles

Nombre total d'articles achetés

Délai moyen de livraison

Temps moyen entre commande et livraison

Note moyenne des avis

Note moyenne des produits achetés

Nombre de catégories achetées

Nombre de catégories différentes achetées

Standardisation des Variables

Pourquoi scaler ?

- Variables sur des échelles différentes

K-means sensible aux distances

Méthode

StandardScaler

Centrer réduire les variables pour obtenir une moyenne = 0 et écart-type = 1

Message clé

Le scaling garantit une segmentation équitable entre les variables.

Illustration de la standardisation

Avant standardisation

Variable 1 (Panier)	Échelle: 0-1000
Variable 2 (Fréquence)	Échelle: 0-10
Variable 3 (Monetary)	Échelle: 0-5000

Après standardisation

Variable 1 (Panier)	Échelle: -2 à +2
Variable 2 (Fréquence)	Échelle: -2 à +2
Variable 3 (Monetary)	Échelle: -2 à +2

Les variables sont maintenant sur la même échelle, ce qui garantit une segmentation équitable.

Choix du Nombre de Clusters

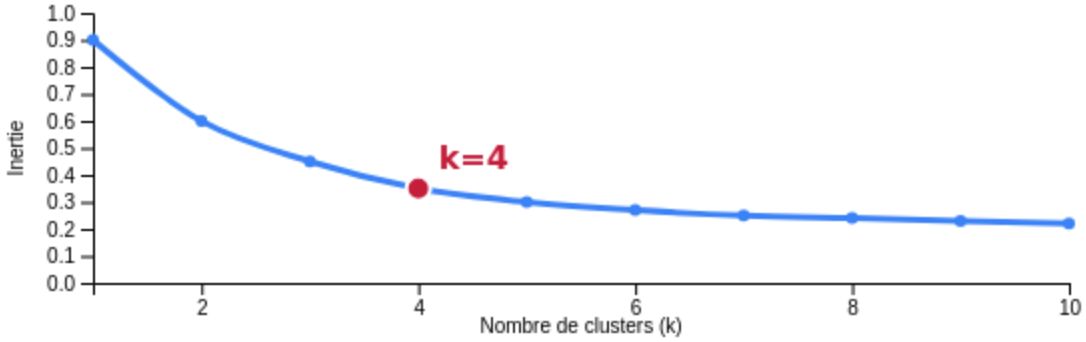
Méthode du coude (inertie)

- Recherche du "coude" dans la courbe d'inertie
- Point d'inflexion où l'inertie diminue moins rapidement

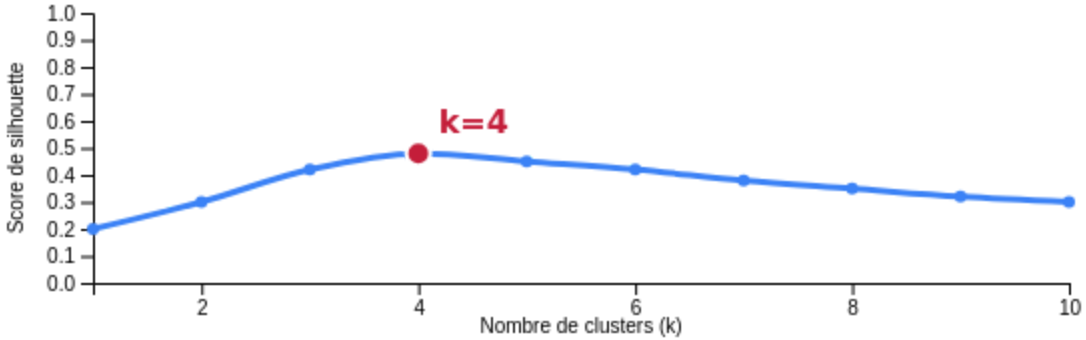
Score de silhouette

- Mesure la qualité de la séparation des clusters
- Score compris entre -1 et 1, plus proche de 1 = meilleur

Méthode du coude



Score de silhouette



Résultat : k = 4 clusters retenus

Bon compromis lisibilité / performance

Segmentation par K-Means

Algorithme

K-means
Algorithme de clustering non supervisé

- Minimise la variance intra-cluster
- Itérations jusqu'à convergence

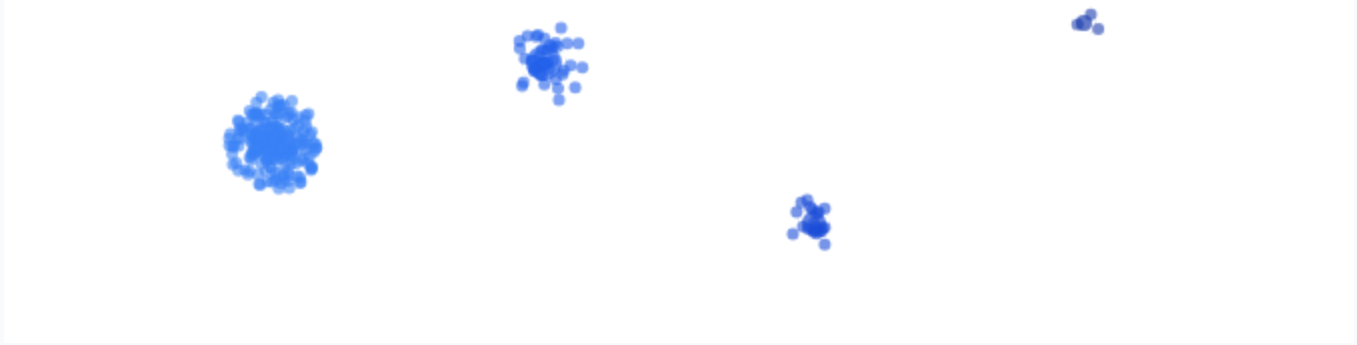
Résultat de la segmentation

Cluster 1 ~75%

Cluster 2 ~16%

Cluster 3 ~7%

Cluster 4 <1%



Résultats

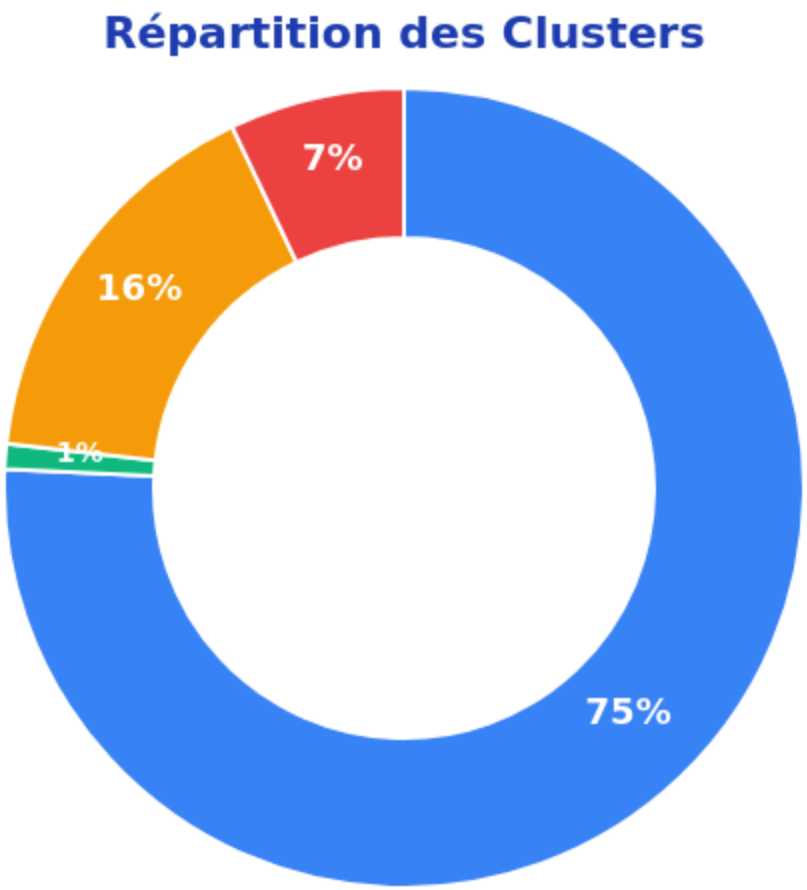
4 segments clients distincts
Chaque cluster représente un profil client différent

Répartition inégale mais réaliste
Reflects the natural distribution of customer value

Message clé

Les segments reflètent la réalité business : tous les clients n'ont pas la même valeur.

Profil des Clusters (Vue Globale)



Légende

Cluster 0: Clients occasionnels

Cluster 1: Clients premium

Cluster 2: Clients réguliers

Cluster 3: Bons clients

Message clé

Une petite partie des clients (Cluster 1) concentre une grande partie de la valeur.

Profil global

- La majorité des clients (75%) sont des clients occasionnels
- Seulement 16% des clients sont des clients réguliers
- Seulement 7% des clients font partie du segment des "bons clients"
- Moins de 1% des clients sont des clients premium

Interprétation Métier des Segments

Clients occasionnels

- Panier moyen faible
- Achat unique
- Cœur de la base

Clients réguliers

- Panier et dépenses modérés
- Bon potentiel de fidélisation

Bons clients

- Dépenses élevées
- Avis clients positifs
- Sensibles à la qualité de service

Clients premium

- Très forte valeur
- Très rares
- À traiter individuellement

Message clé

Chaque segment appelle une stratégie marketing différente.

Visualisation PCA

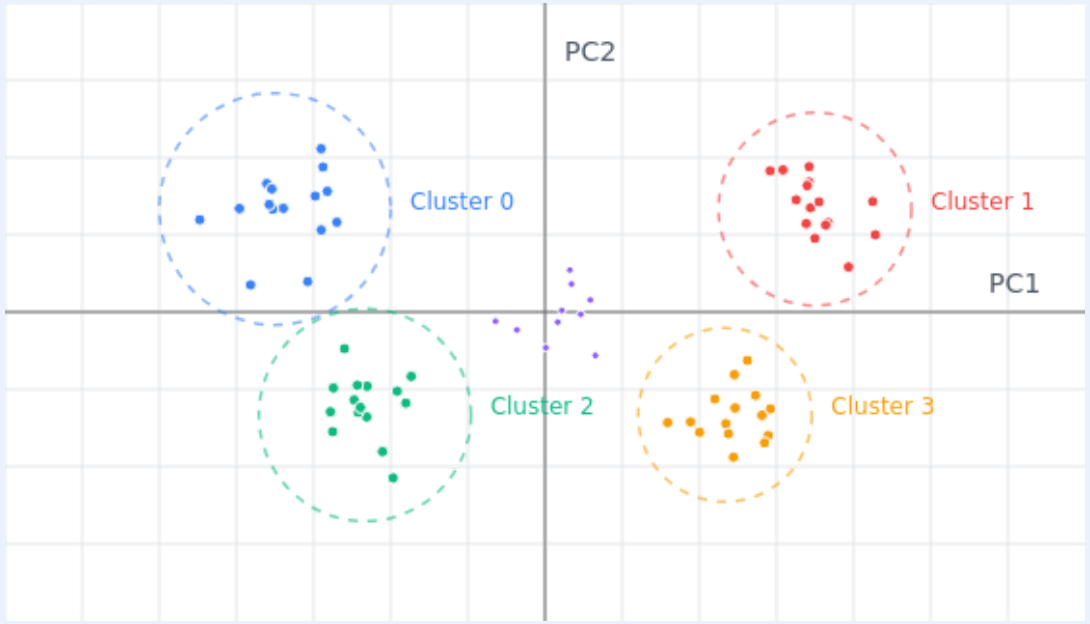
Objectif

Vérifier la séparation des clusters obtenus après la segmentation par k-means. Le PCA permet de visualiser les données dans un espace réduit tout en préservant les relations de distance entre les points.

Message clé

Le PCA confirme la cohérence de la segmentation. Les clusters sont bien séparés dans l'espace des composantes principales, ce qui valide la robustesse de notre approche de segmentation.

Résultat



- **Bonne séparation globale** des clusters
- **Chevauchements limités** entre clusters
- Validité **statistique** de la segmentation

Modélisation Supervisée (Appui à la Segmentation)

Objectif

- Prédire l'appartenance à un segment
- Comprendre les variables discriminantes

Modèle

- **Algorithme** : Random Forest
- **Validation** croisée

Message clé

Le modèle confirme la robustesse des segments.

Résultats

Performance

F1-score ≈ 0.996

Stabilité

- Modèle stable et robuste

Validité des segments

- Les segments sont **discriminants** pour le modèle
- Les segments sont **homogènes** et **hétérogènes** entre eux

Importance des Variables

Variables les plus discriminantes

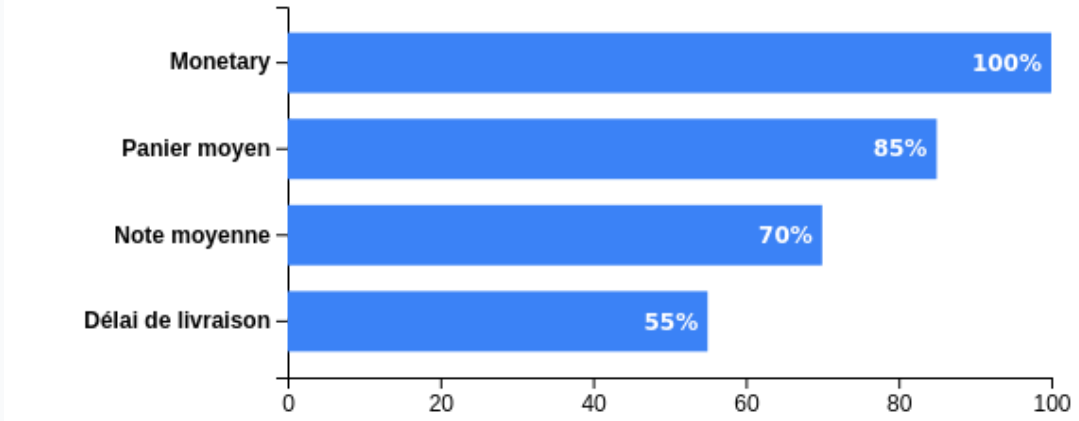
Monetary
Dépenses totales des clients

Panier moyen
Valeur moyenne des commandes

Note moyenne des avis
Satisfaction client

Délai de livraison
Temps d'attente moyen

Importance relative des variables



Message clé

La valeur client et l'expérience client (panier, notes, livraison) sont les facteurs clés de segmentation.

Pourquoi le modèle n'overfit pas (explication clé)

Point fondamental

Les clusters ont été construits à partir des variables suivantes :

- Recency
- Frequency
- Monetary
- Panier moyen
- Délai de livraison
- Note moyenne des avis
- Nombre de catégories

Interprétation correcte

Le modèle n'a pas pour objectif de prédire un comportement futur.

Il apprend à reproduire la frontière de décision du clustering.

Il s'agit d'un cas connu en data science : tautologie de segmentation

Modèle supervisé

Cluster 1 Cluster 2 Cluster 3 Cluster 4

La prédiction des clusters n'est pas un apprentissage prédictif, mais une dérivation logique des caractéristiques déjà utilisées pour la segmentation.

Objectif réel du modèle (vision entreprise)

Objectif du modèle supervisé

Pas prédire un comportement inconnu

Industrialiser la segmentation client

Utilité business

- Assigner rapidement de nouveaux clients à un segment
- Éviter de relancer un clustering complet
- Faciliter l'intégration CRM
- Outils marketing connectés

Déploiement en entreprise

Segmentation

Modèle supervisé

Déploiement

Pratique courante en entreprise

Cette approche est largement utilisée dans le monde des affaires pour automatiser et industrialiser les processus de segmentation tout en garantissant la cohérence des résultats.

Performance stable avec F1-score élevé

(0.996 sur le jeu de test)