



P R E S E N T A T I O N

# Optimisation de l'expérience client par la segmentation des clients E-commerce

• • •

Comment identifier des segments de clients avec des comportements d'achat similaires afin de personnaliser les stratégies marketing et améliorer l'expérience client ?



# Agenda

• • •

## 1. Contexte & Préparation des données

Problématique métier  
Nettoyage des données  
Feature Engineering  
Exploration des données

## 2. Modélisation & Choix du modèle final

Pistes de modélisation explorées  
Comparaison des performances  
Sélection du modèle final

## 3. Simulation & Plan de maintenance

Simulation du comportement du modèle dans le temps  
Définition d'un délai optimal de maintenance  
Proposition de contrat de maintenance



# Contexte & Préparation des données



Search

Sign In Register

O LIST AND 3 COLLABORATORS · UPDATED 4 YEARS AGO

3530 Code Download :

## Brazilian E-Commerce Public Dataset by Olist

100,000 Orders with product, customer and reviews info

Data Card Code (584) Discussion (60) Suggestions (0)

### About Dataset

#### Brazilian E-Commerce Public Dataset by Olist

Welcome! This is a Brazilian ecommerce public dataset of orders made at [Olist Store](#). The dataset has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allows viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers. We also released a geolocation dataset that relates Brazilian zip codes to lat/lng coordinates.

This is real commercial data, it has been anonymised, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

Usability 10.00

License CC BY-NC-SA 4.0

Expected update frequency Never

Tags Business




**Contexte** : Structuration de l'équipe Data d'Olist

**Objectifs** : Construire une segmentation client exploitable pour le marketing, et définir sa fréquence de mise à jour optimale afin d'assurer sa pertinence dans le temps via un contrat de maintenance.

**Étape initiale** : Nettoyer et explorer les données pour comprendre la structure et préparer les modélisations.

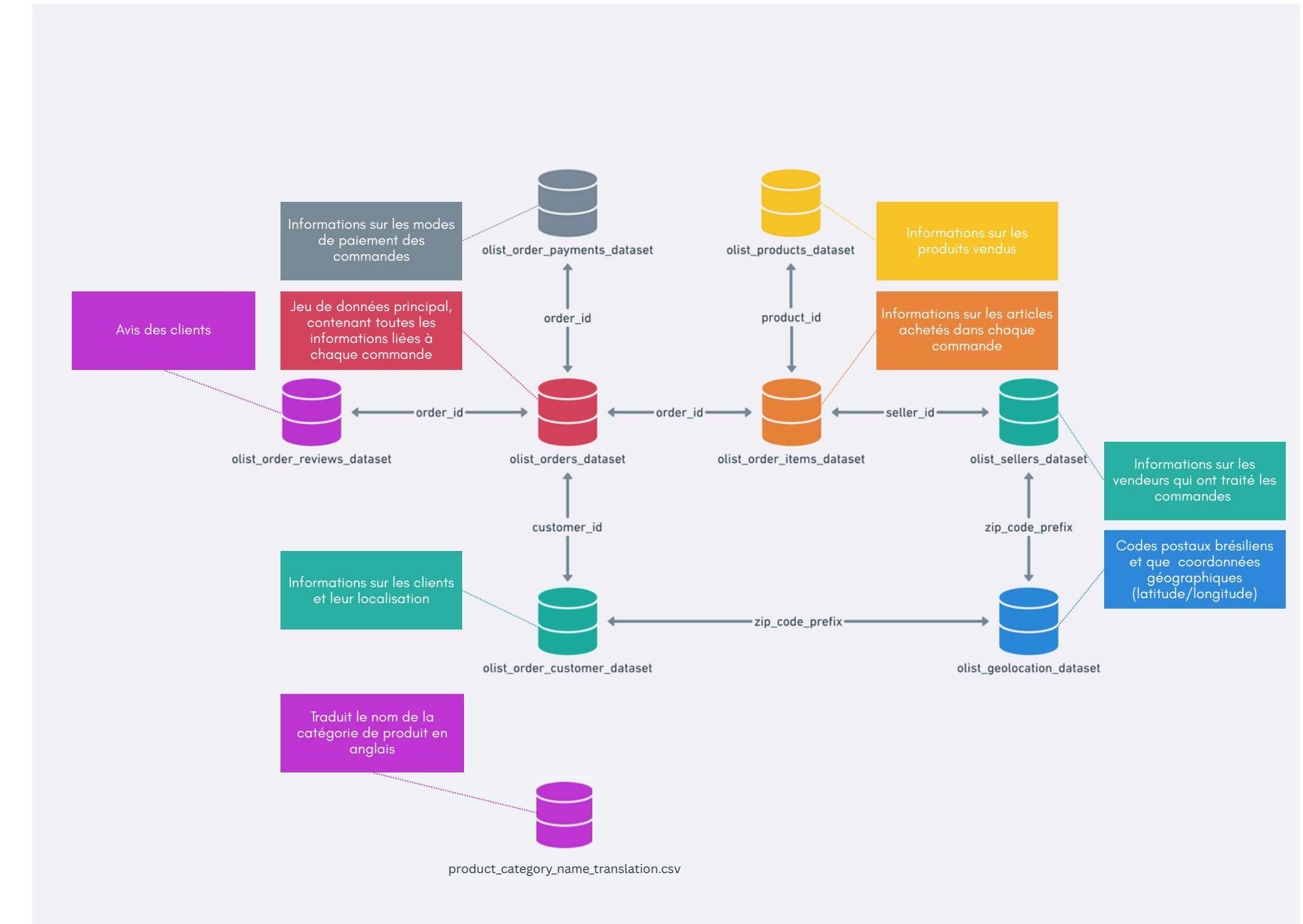
Contexte



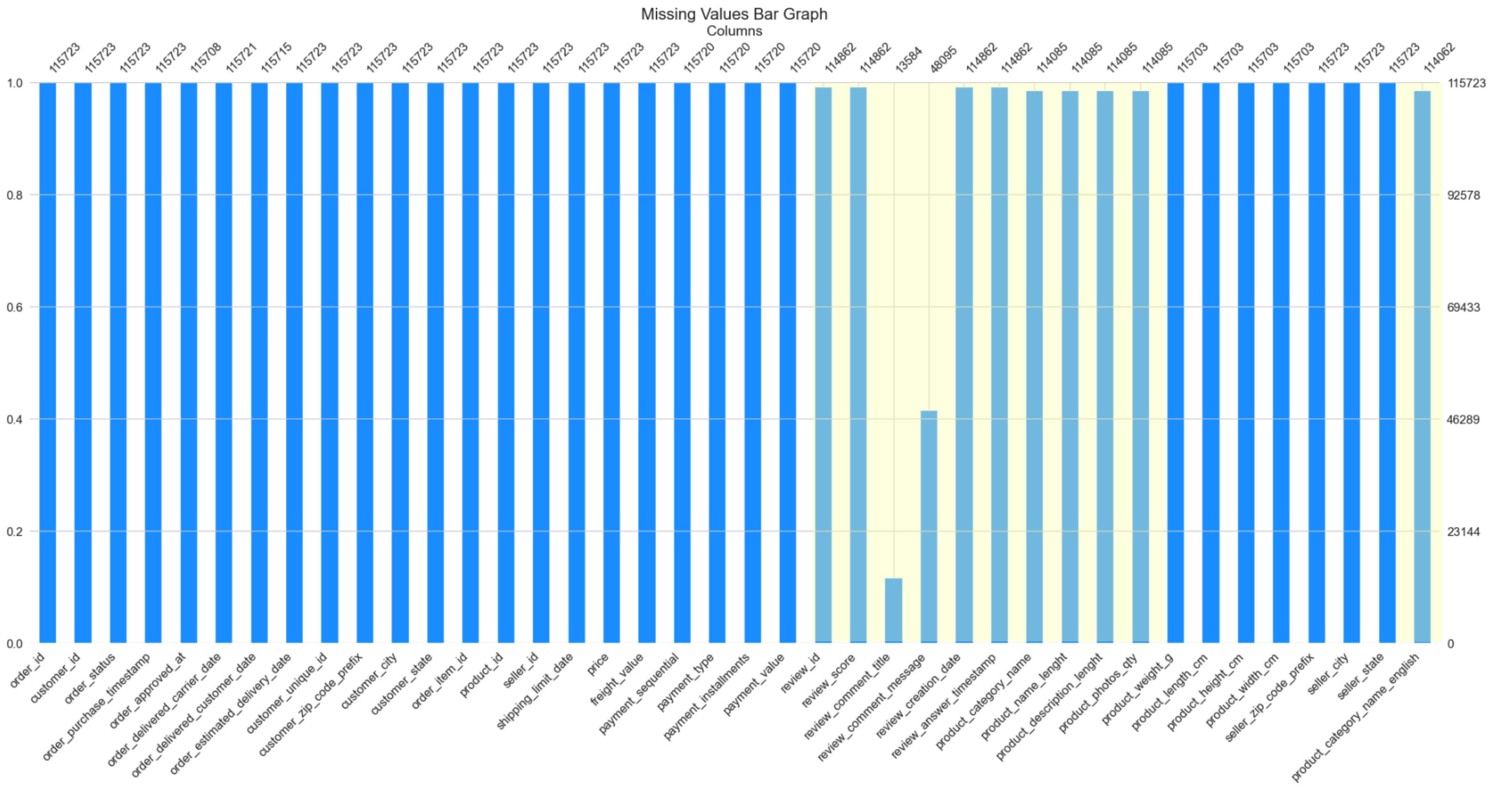
## Les données

Données anonymisées sur plus de **100 000 commandes** passées sur plusieurs marketplaces brésiliennes entre **2016 et 2018**

- CSV olist\_customers\_dataset.csv
- CSV olist\_geolocation\_dataset.csv
- CSV olist\_order\_items\_dataset.csv
- CSV olist\_order\_payments\_dataset.csv
- CSV olist\_order\_reviews\_dataset.csv
- CSV olist\_orders\_dataset.csv
- CSV olist\_products\_dataset.csv
- CSV olist\_sellers\_dataset.csv
- CSV product\_category\_name\_translation.csv



## Les données





## Nettoyage des valeurs manquantes et aberrantes

Filtrer le statut des commandes

[Livrées \(Delivered\)](#)

Remplissage des valeurs manquantes par la moyenne

[review\\_score](#)

Remplir toutes les valeurs manquantes dans les colonnes liées aux caractéristiques du produit par 0

Vérifier les commandes dupliquées

[Les lignes des commandes clients identiques](#)

Nettoyer les catégories des colonnes

[category, datetime](#)

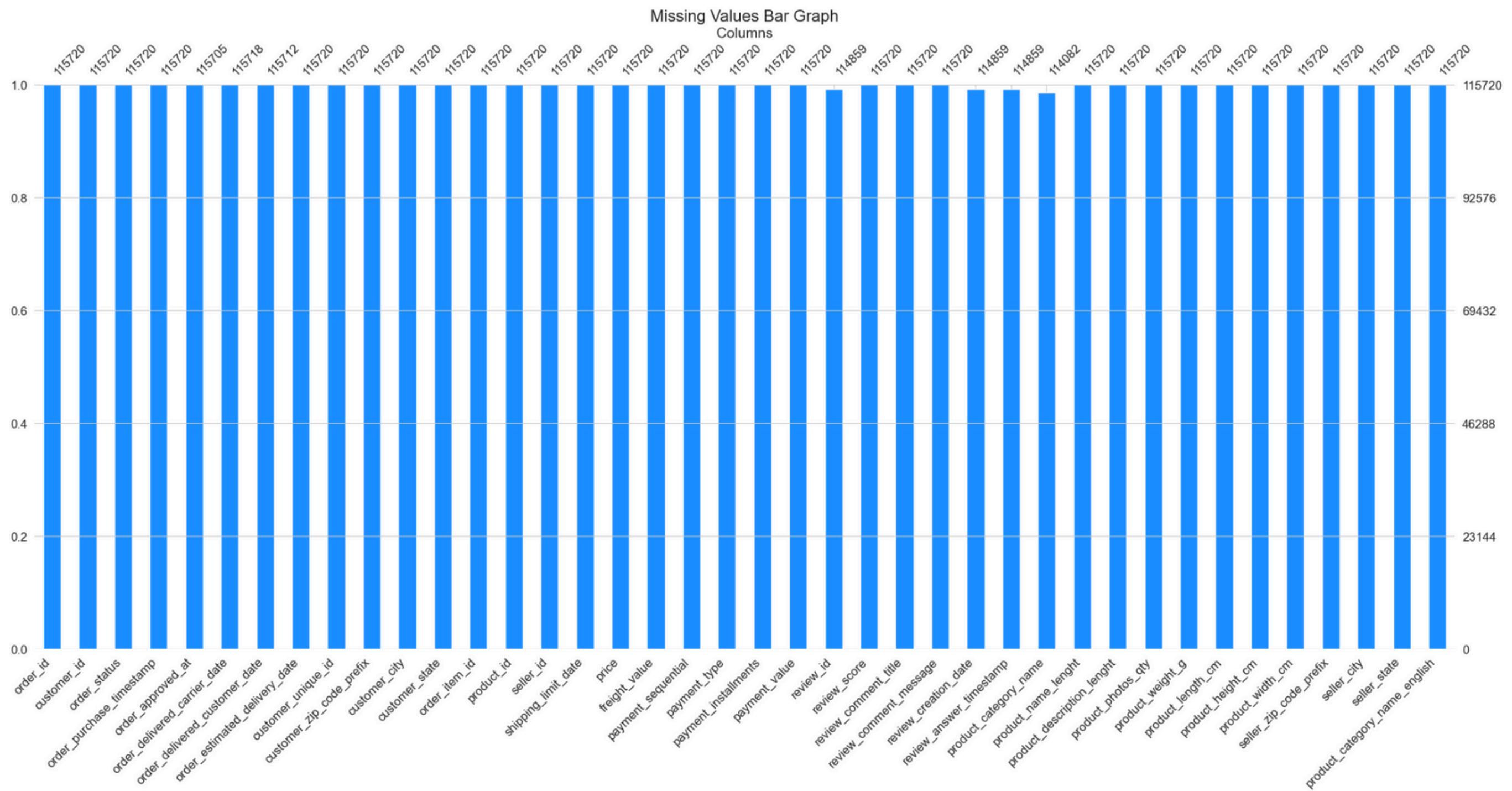
Compléter les catégories de produits en anglais en utilisant les traductions portugaises existantes

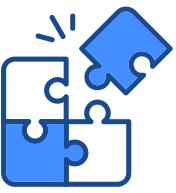
[product\\_category\\_name\\_english](#)



**Nettoyage des données**







Consolidation de plusieurs jeux de données



Création d'un dataset clients



## La récence d'achat

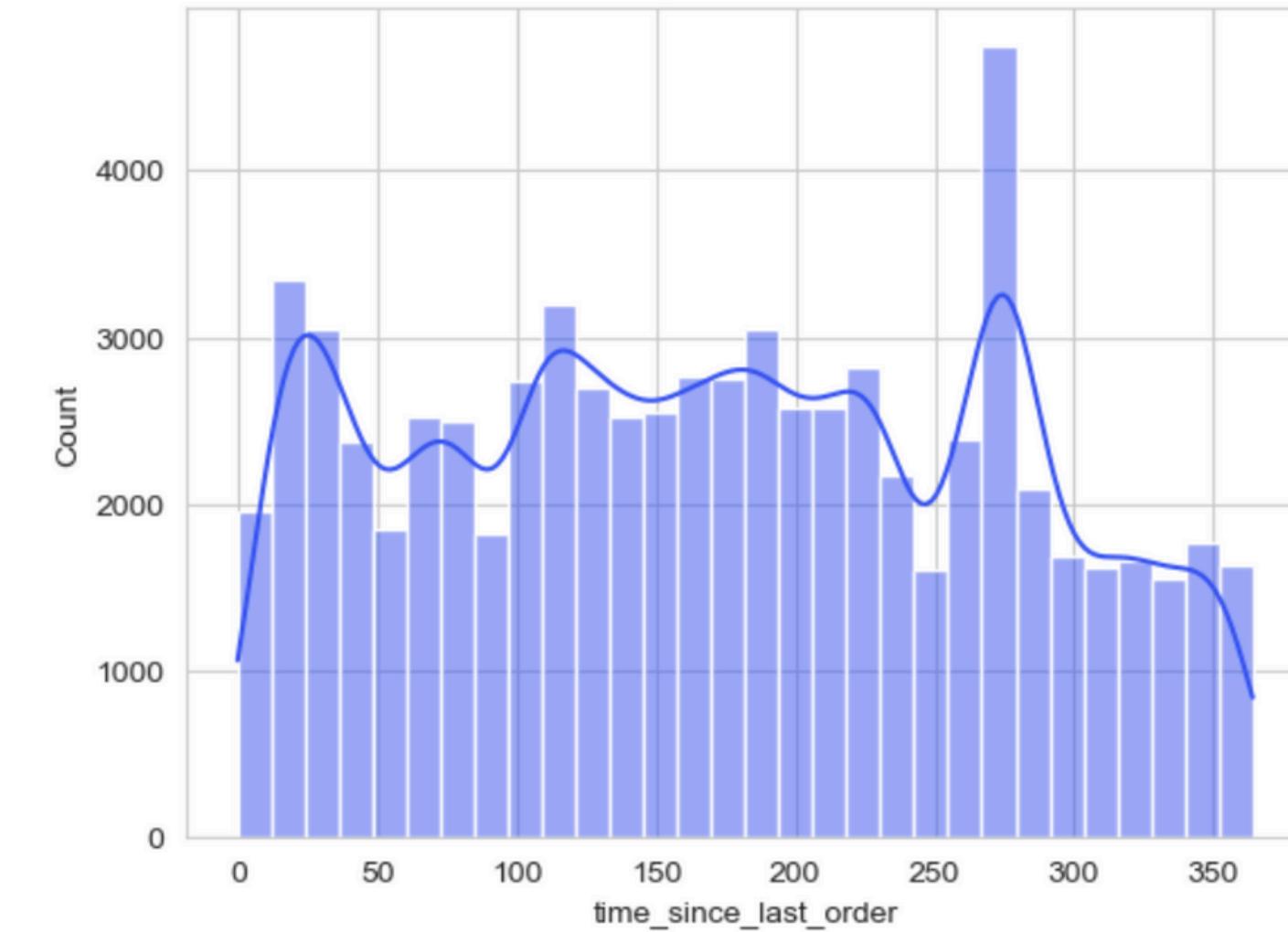
Temps depuis le dernier achat

**time\_since\_last\_order**



**Objectif :** Mesurer l'ancienneté et la récence de l'activité d'un client

Jours écoulés depuis la dernière commande



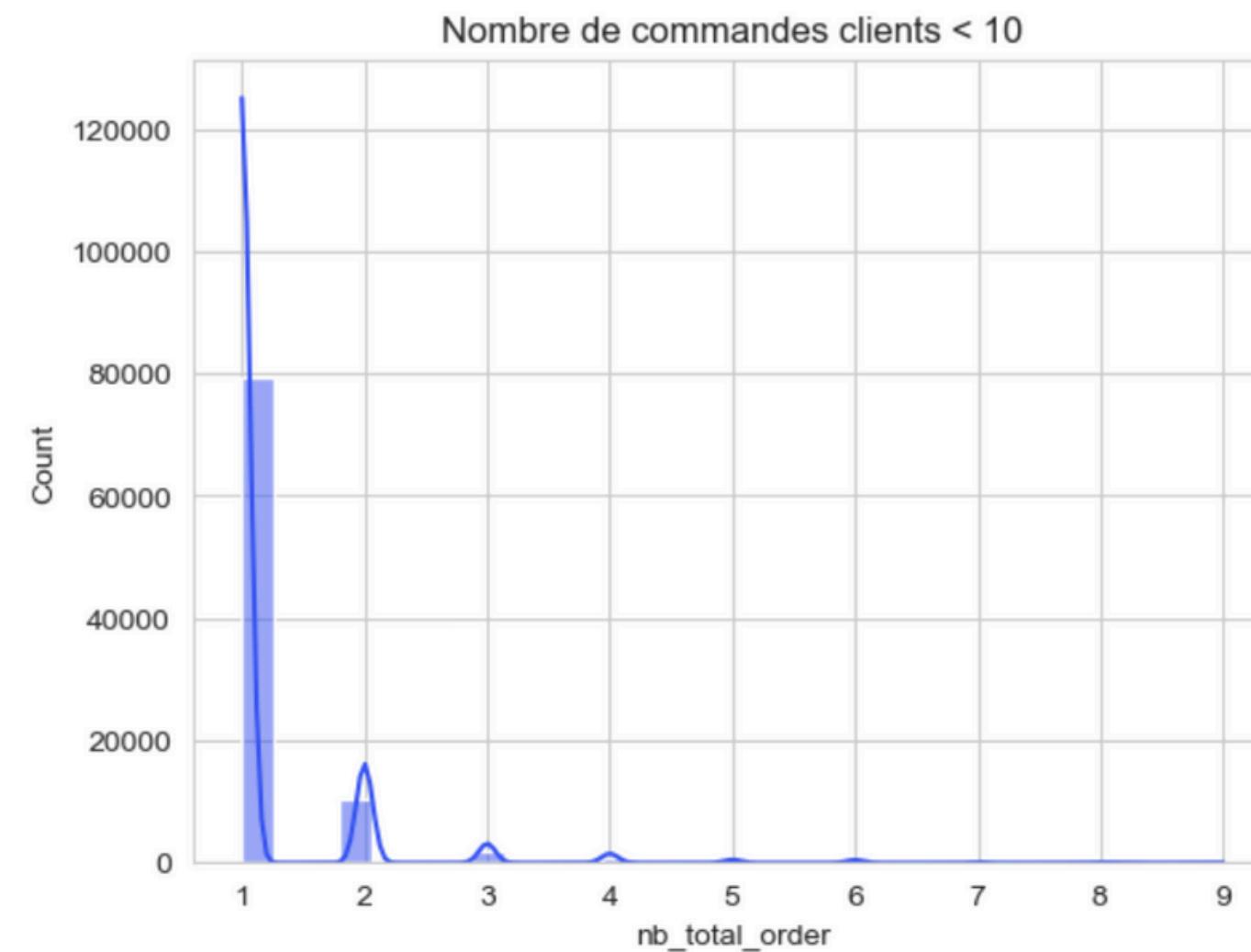
## La fréquence d'achat

Nombre total de commandes passées par chaque client

**nb\_total\_order**



**Objectif :** Mesurer la fréquence, l'engagement et la fidélité de chaque client

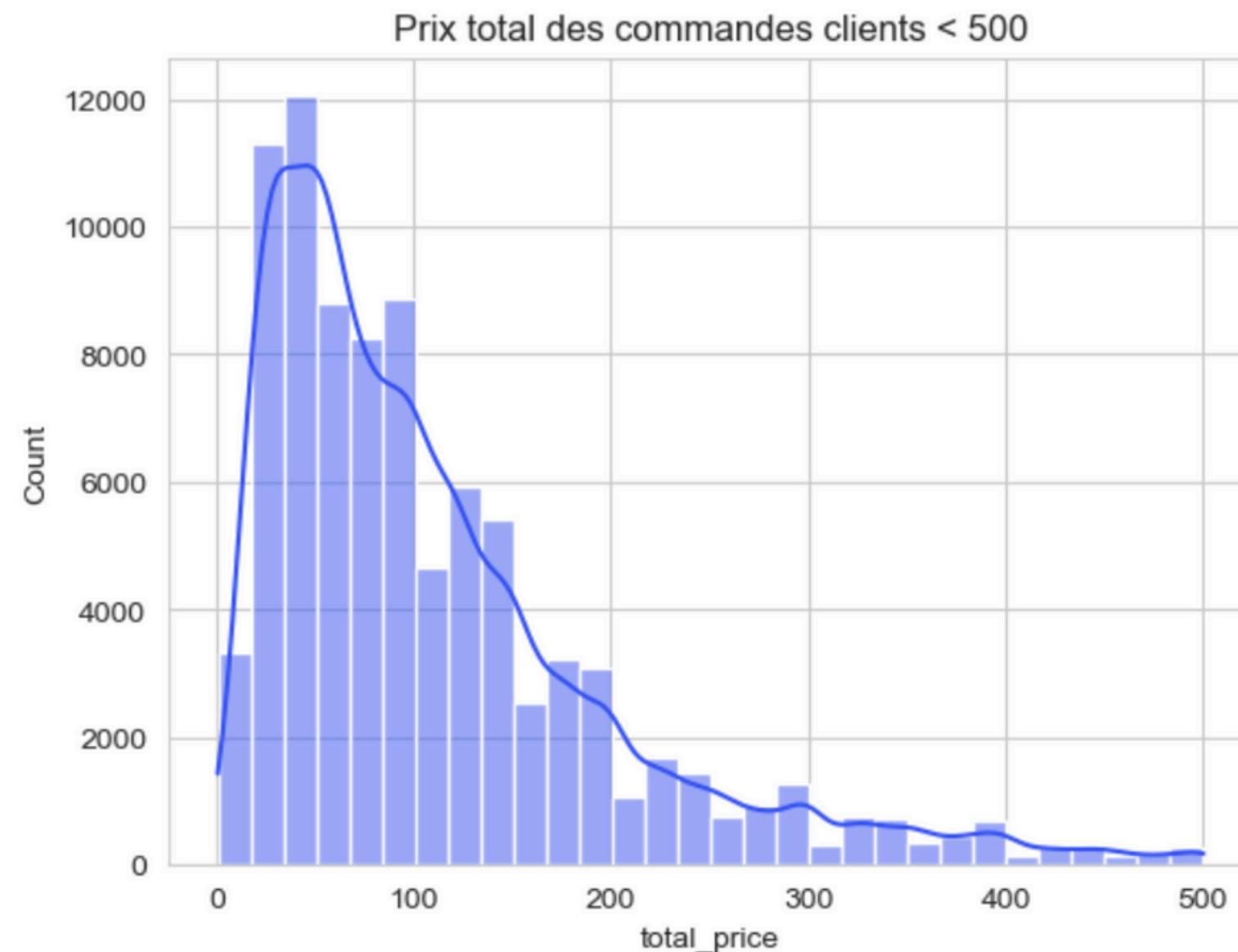


## Le montant d'achat

Prix total  
**total\_price**



**Objectif :** Evaluer le volume d'achats et le poids monétaire du client



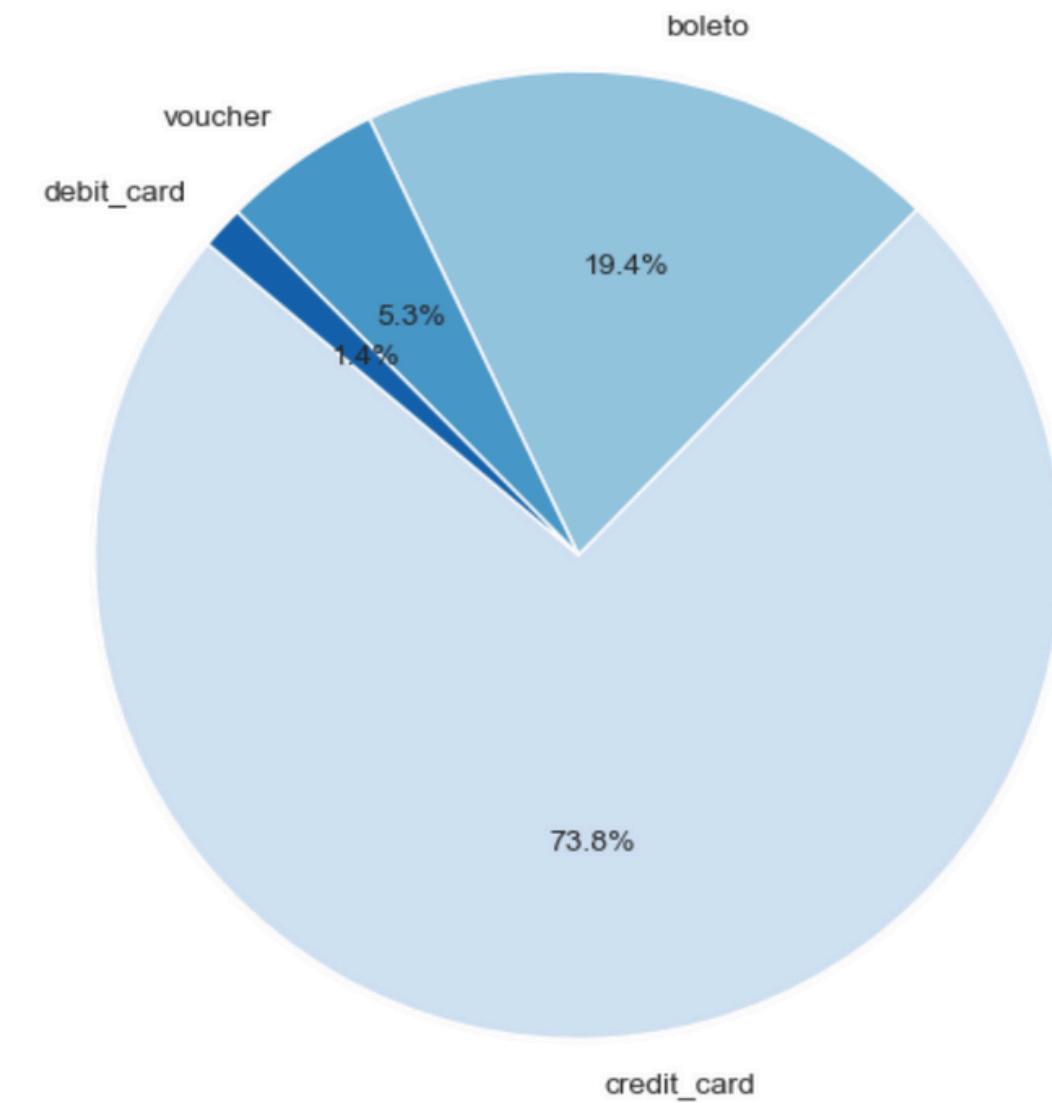
## Les moyens de paiement client

Nombre de paiements  
**payment\_type**



**Objectif :** Mesurer le type de paiement le plus utilisé par le client

Répartition des moyens de paiement



## La satisfaction client

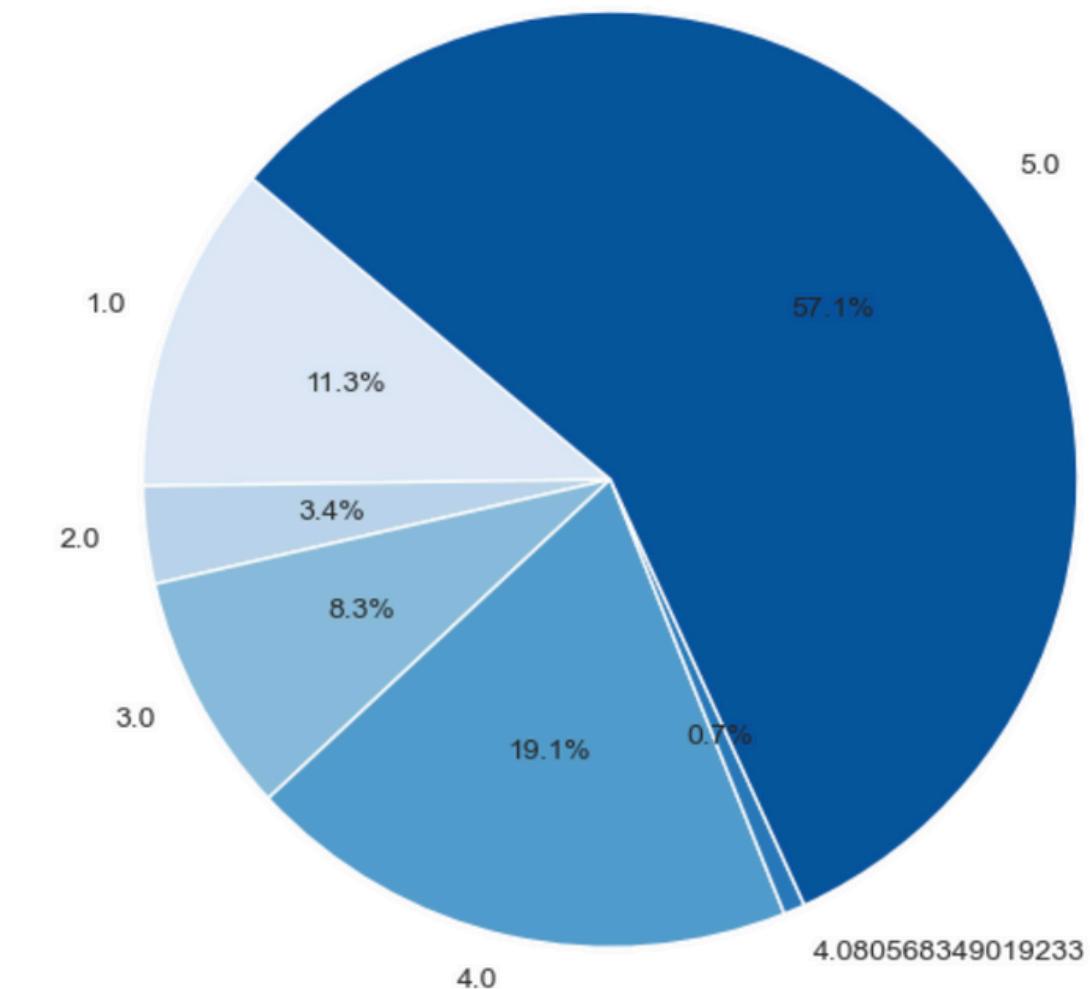
Note moyenne attribuée par le client

**mean\_review\_score**



**Objectif :** Mesurer le niveau de satisfaction moyen du client vis-à-vis de ses achats.  
Identifier les clients satisfaits et potentiellement insatisfaits

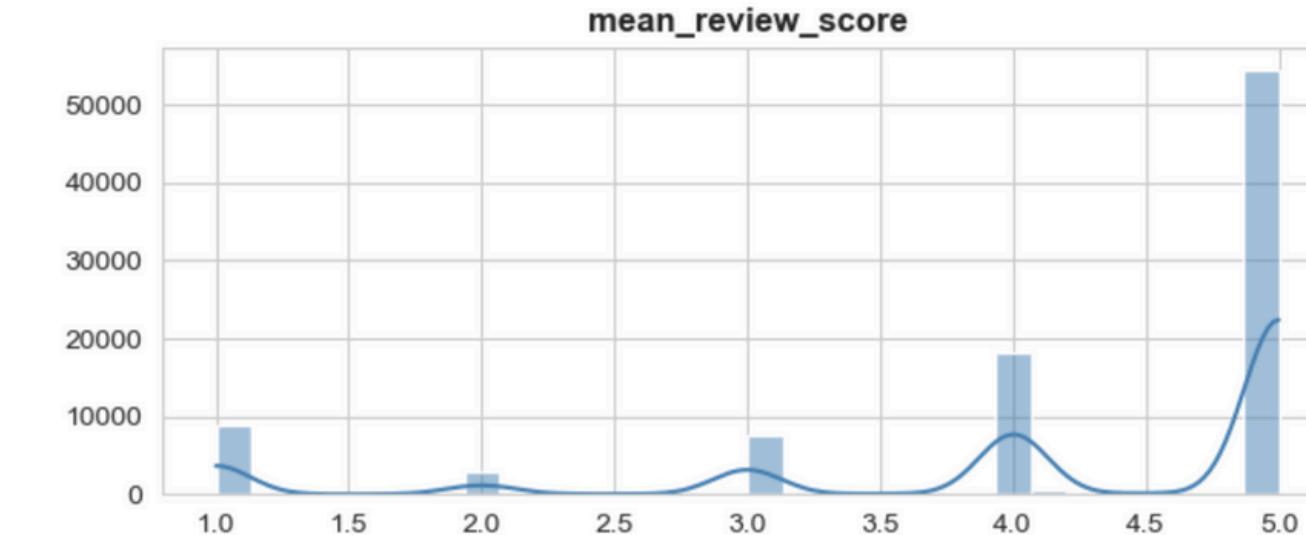
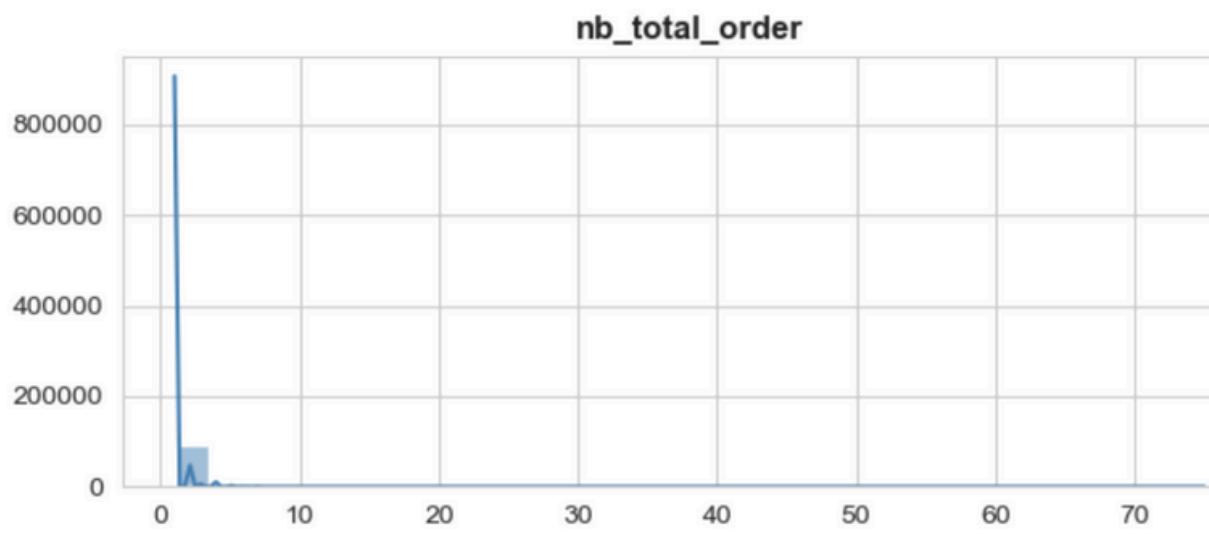
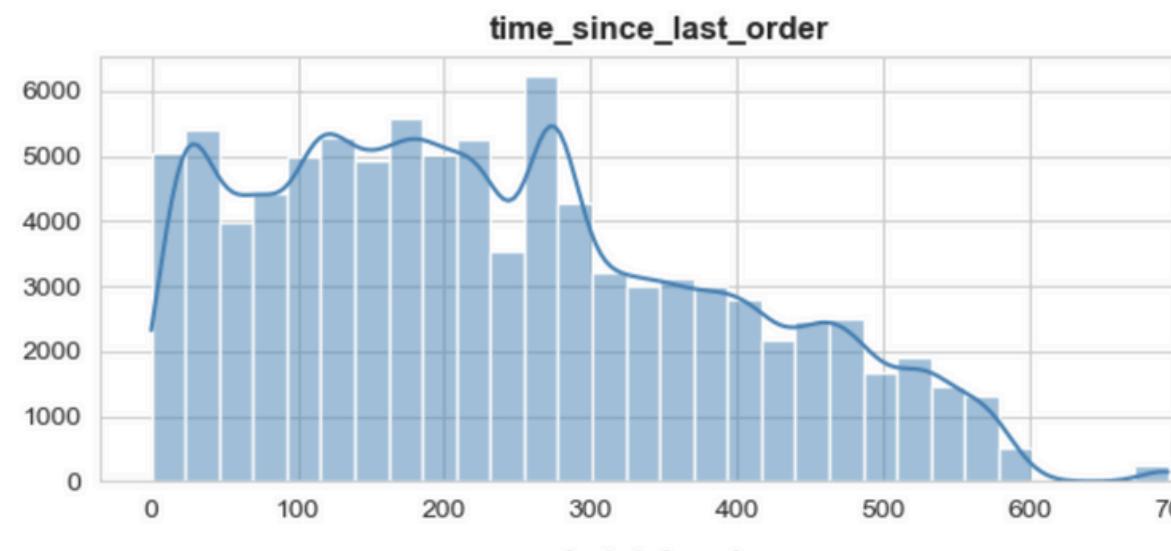
Répartition des avis clients



**Feature Engineering**



## Distribution des variables quantitatives



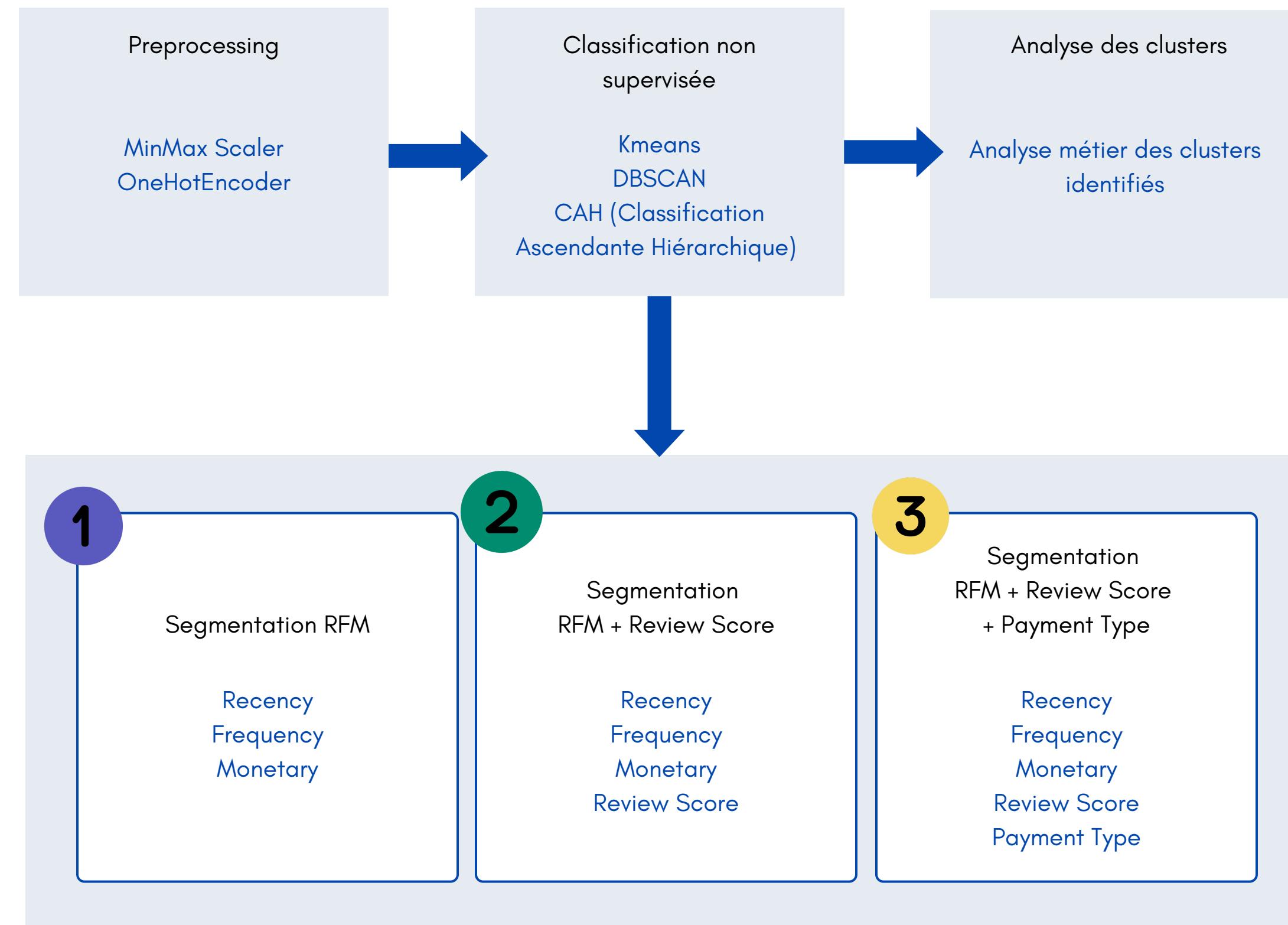
**Exploration des données**





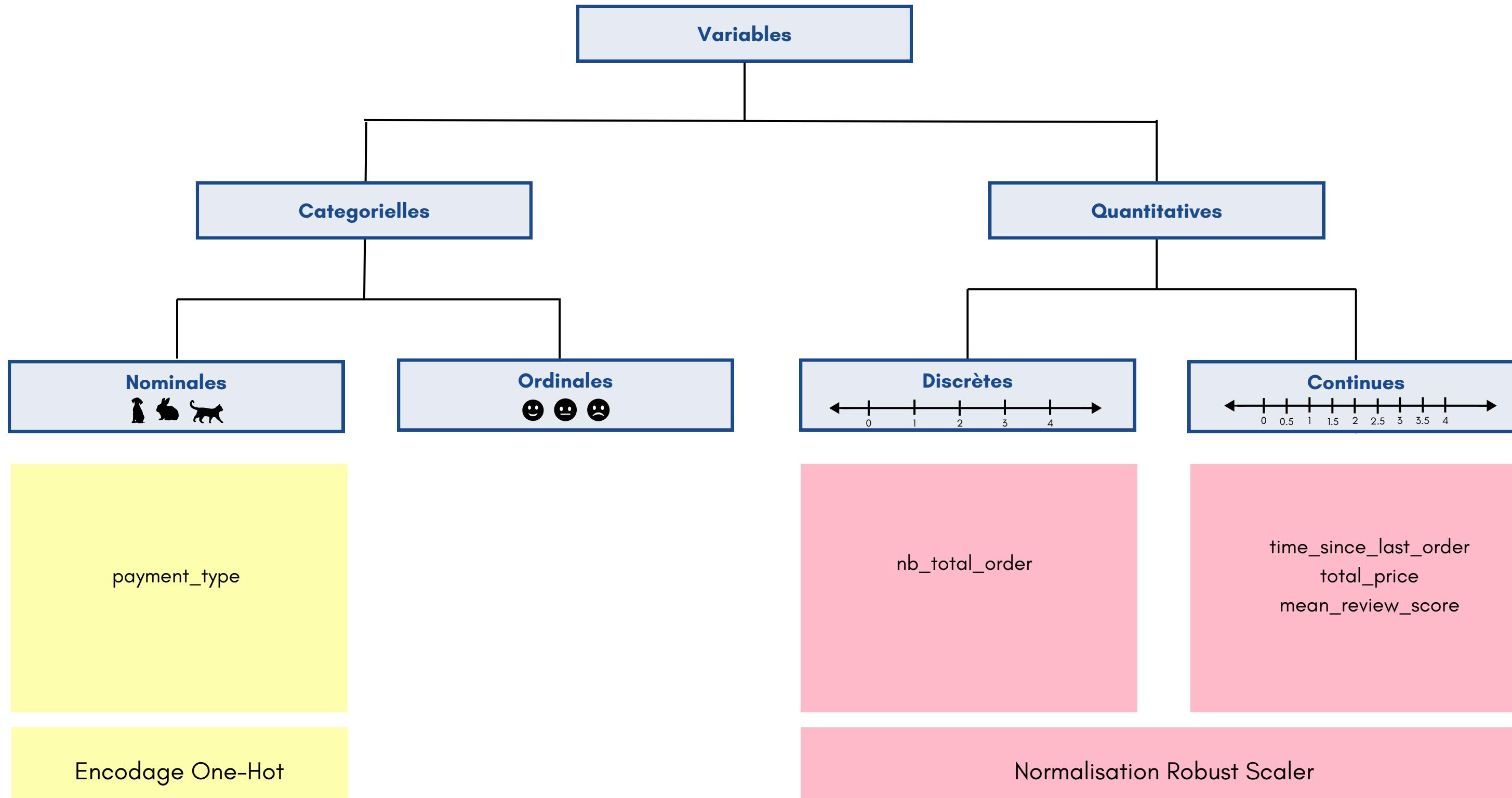
# Modélisation & Choix du modèle final





+

Pistes de modélisation

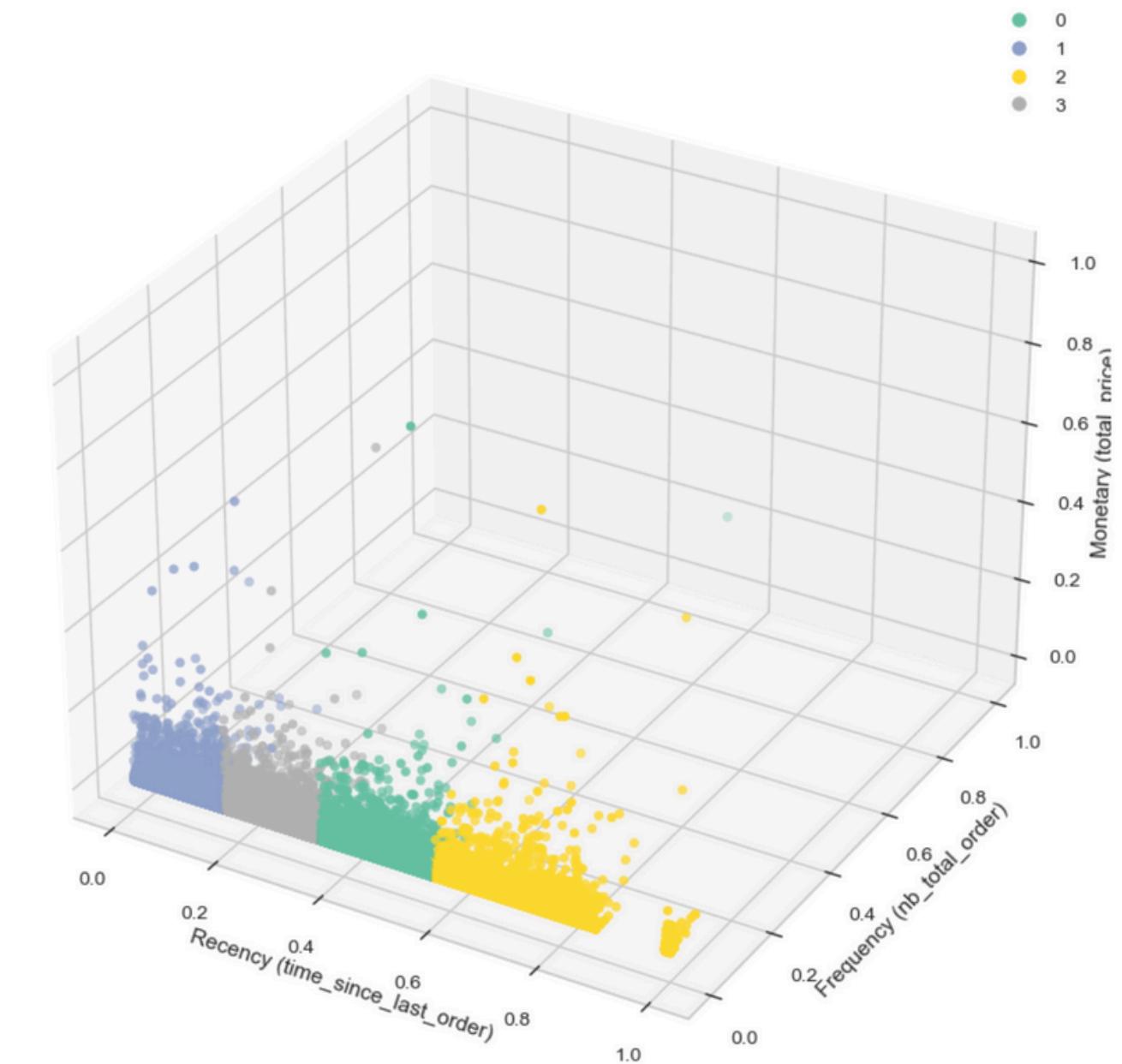


Preprocessing

## Kmeans



Clustering RFM avec Kmeans k=4

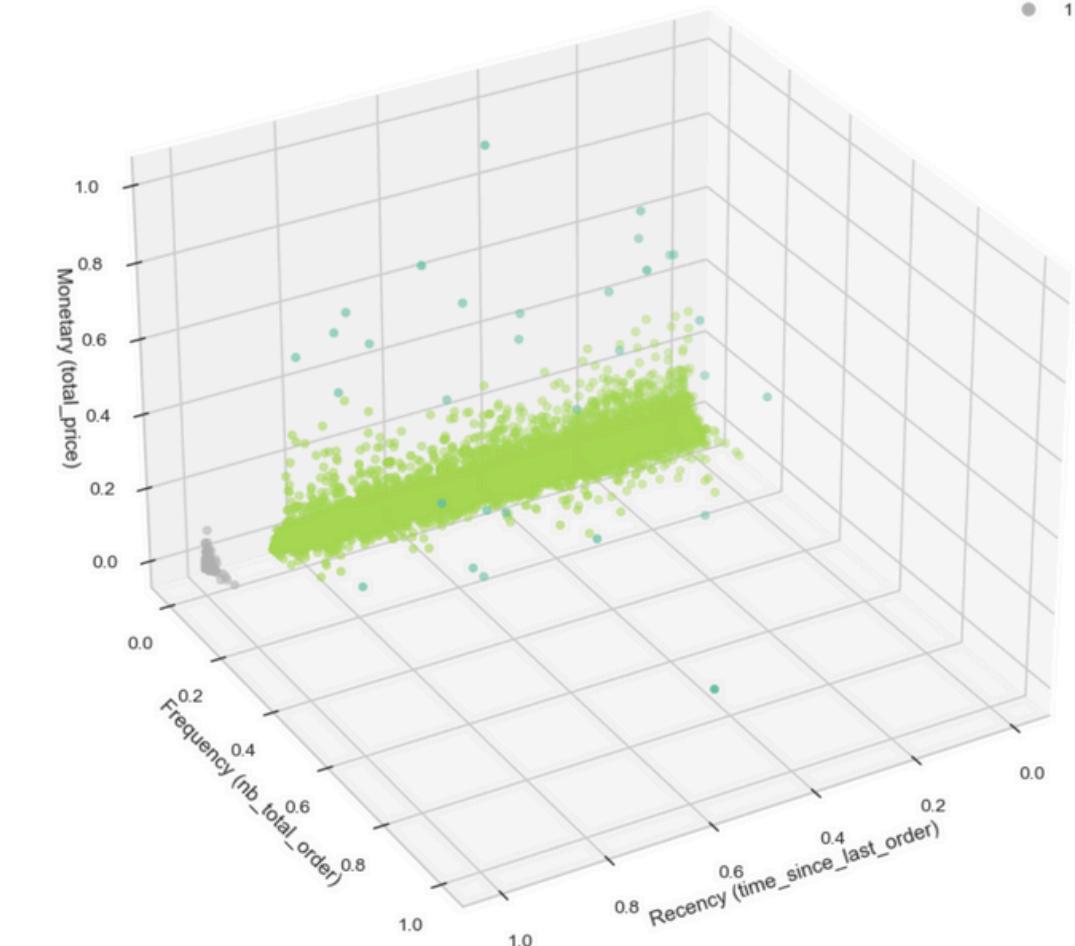


Segmentation  
RFM

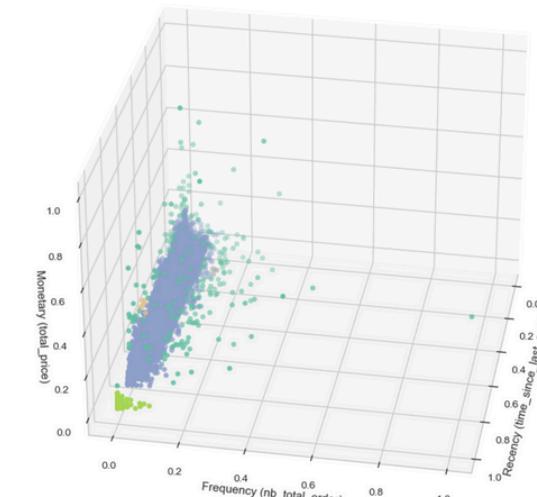
# DBSCAN

Paramètres du modèle	Clusters	Bruit	Silhouette	David Bouldin Score	Calinski Harabasz Score
epsilon = 0.1 min_samples = 6	2	31	0.48	1.00	1207
epsilon = 0.03 min_samples = 6	4	217	-0.17	1.55	623
epsilon = 0.05 min_samples = 6	2	82	0.42	1.33	1223

Clustering RFM avec DBSCAN (eps:0.1, min\_samples=6)



Clustering RFM avec DBSCAN (eps:0.03, min\_samples=6)

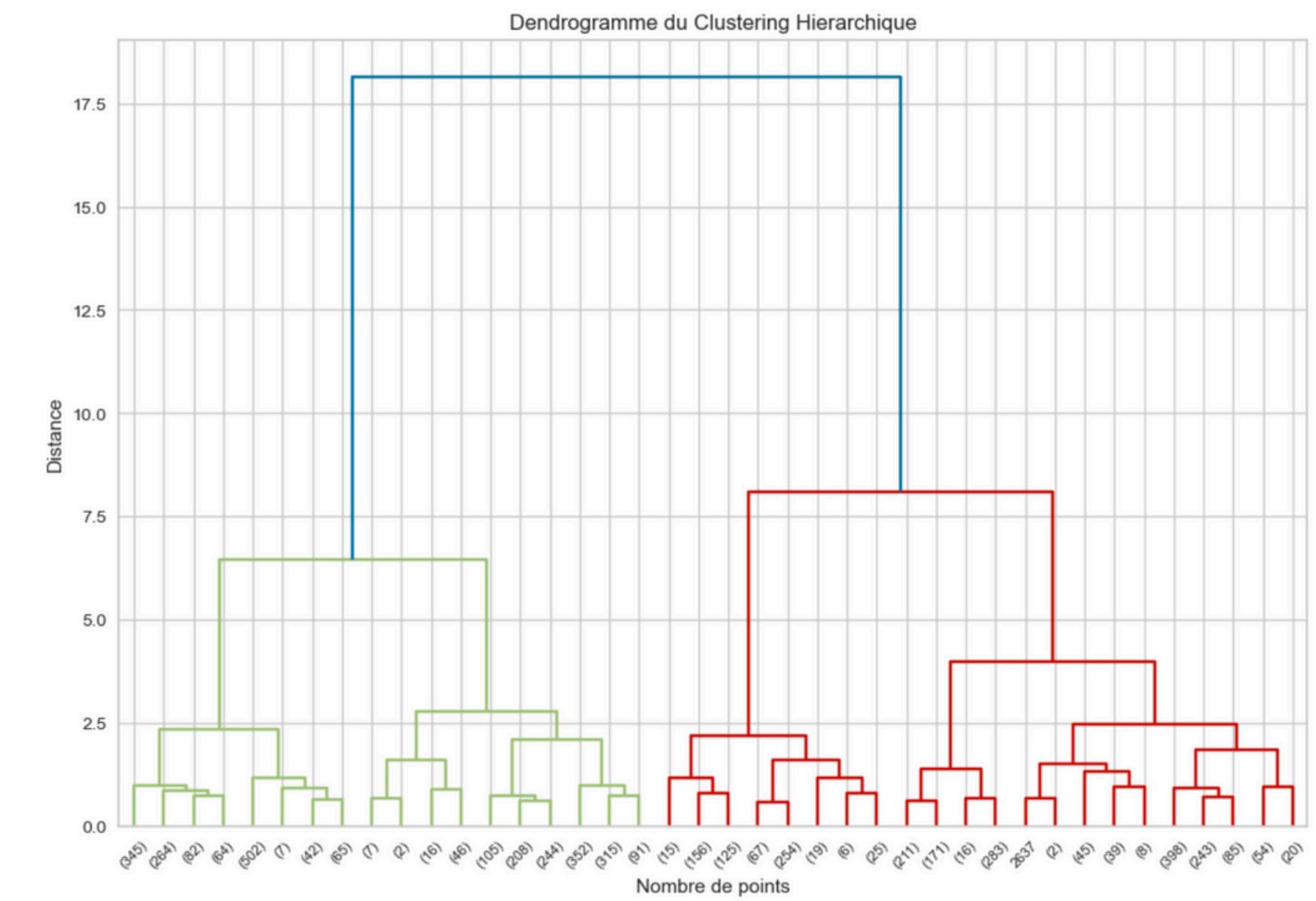


Segmentation  
RFM



## CAH (Classification Ascendante Hiérarchique)

Paramètres du modèle	Silhouette	David Bouldin Score	Calinski Harabasz Score
k=2	0.536	0.654	8430.3
k=3	0.475	0.666	7599.4
k=4	0.435	0.726	8224.5



## Autres ségmentations

Segmentation	Paramètres du modèle	Silhouette Score	Davies Bouldin Score	Calinski Harabasz Score
RFM + Score	Kmeans = 4	0.483	0.769	127 717
RFM + Score + Payment Type	Kmeans = 4	0.458	0.729	84 037



Autres  
segmentations

## RFM + Score

Kmeans = 4

Client inactif ou en perte de vitesse



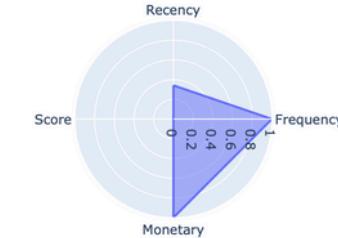
- Campagnes de réactivation
- Campagne email avec incitation
- Campagnes de retargeting

Nouveau client, potentiellement à fidélisé



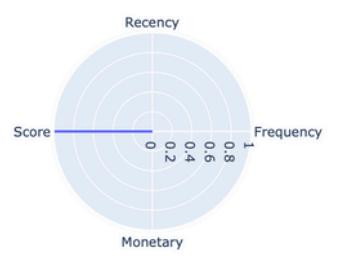
- Envoi d'un email de bienvenue personnalisé
- Proposer un avantage pour le 2e achat
- Campagnes de nurturing

Client précieux en danger



- Envoi d'un email personnalisé « Vous nous manquez » avec offre exclusive
- Mise en place d'un programme de fidélité VIP
- Campagnes de retargeting personnalisées

Prospect qui n'a jamais acheté



- Welcome series emails avec présentation de la marque + offre d'entrée
- Publicité retargeting avec mise en avant des best-sellers
- Proposer une offre d'essai ou produit découverte à petit prix



Modèle final





# Simulation & Plan de maintenance



# Méthode

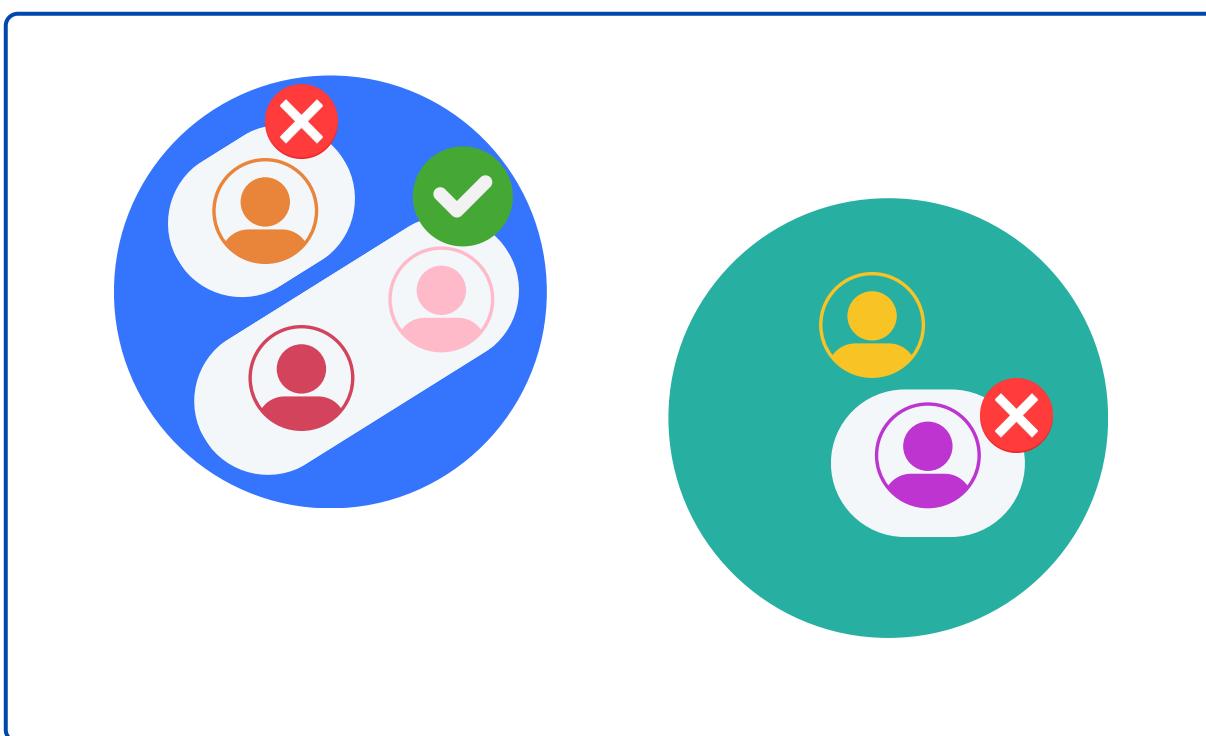


**Simulation du modèle dans le temps**

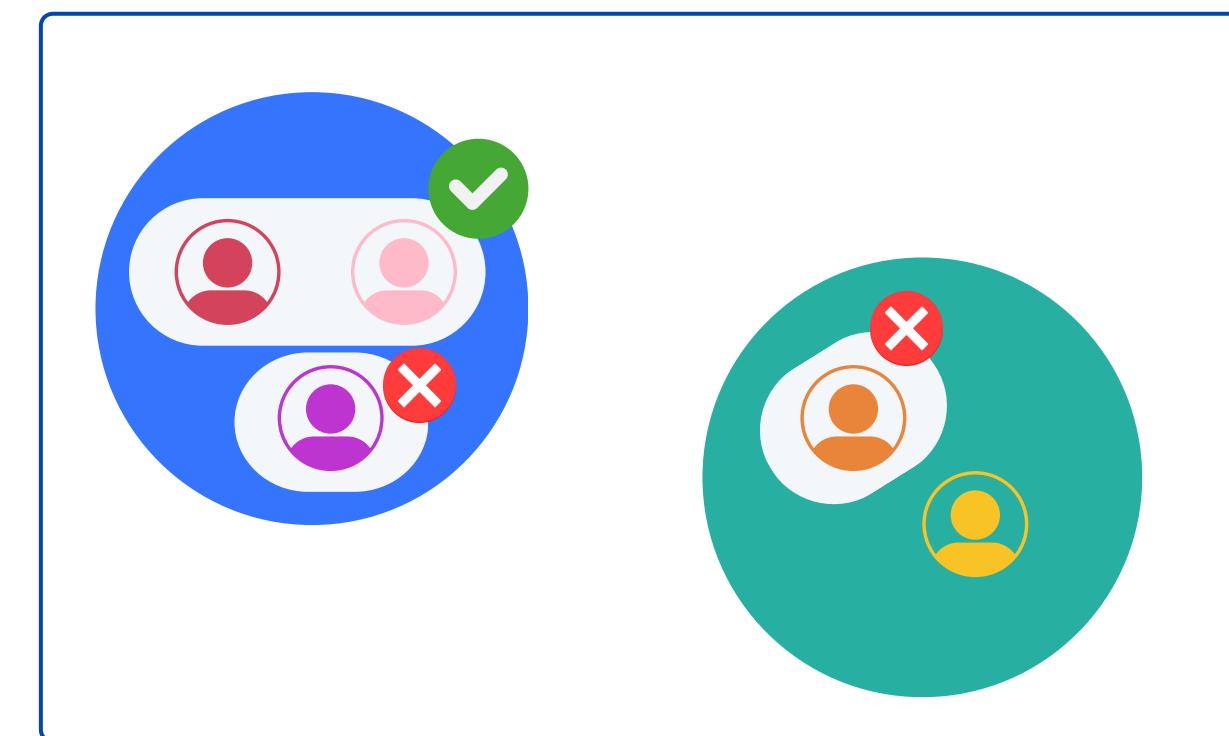


# Comment évaluer la qualité d'une segmentation client ?

Segmentation de référence



Segmentation automatique



ARI - Adjusted Rand Index



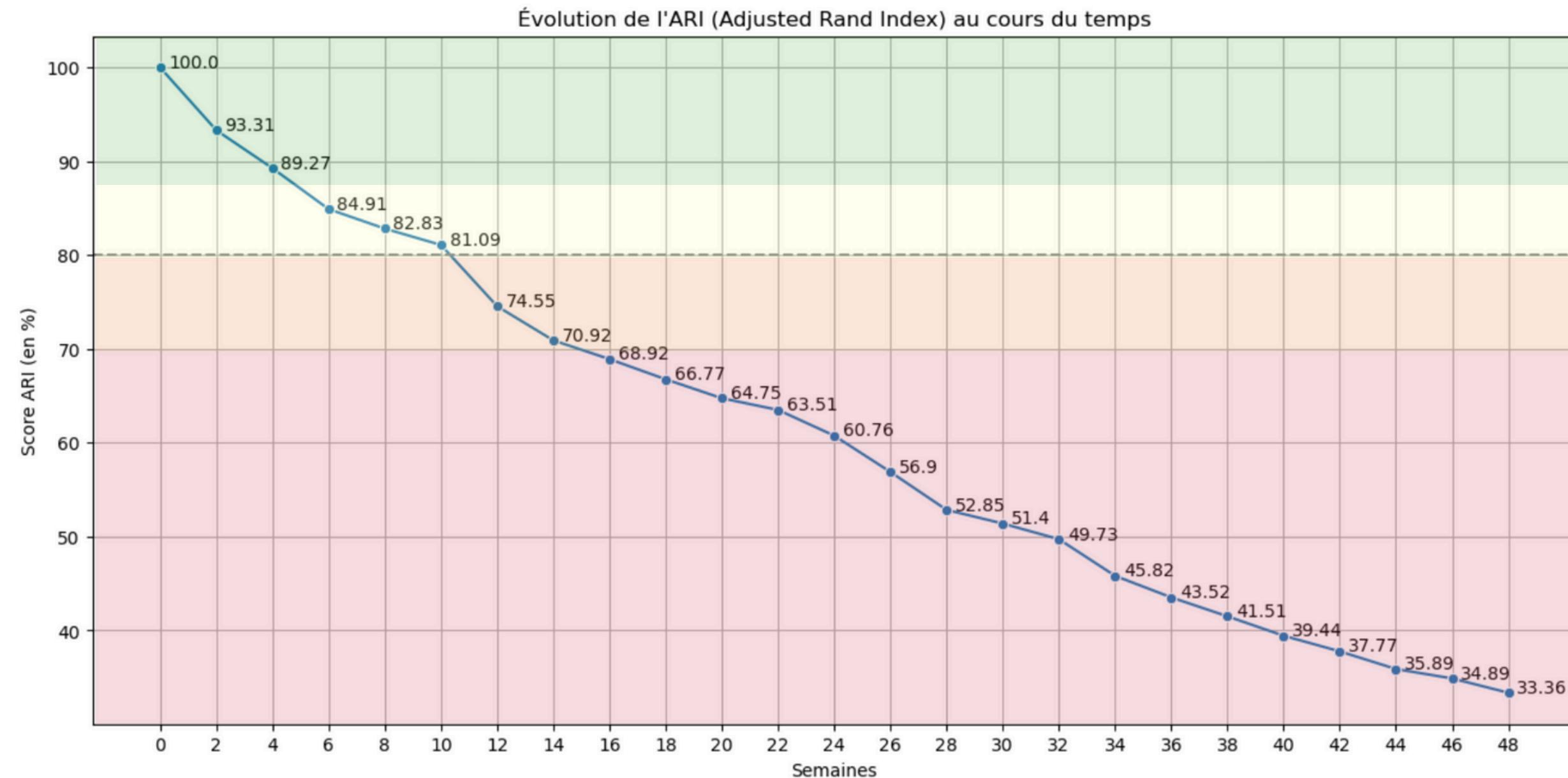
Contribue positivement au score



Contribue négativement au score

Simulation du  
modèle dans le  
temps

Réentraînement  
Tous les 2,5 mois :  
seuil de **80%**



Définition d'un  
délai optimal de  
maintenance



# Pourquoi définir ce seuil de score ARI ?

80% d'ARI comme seuil de ré-entraînement pour nos clusters clients



Réduire le gaspillage marketing



Protéger l'expérience client



Maintenir l'avantage compétitif



**Proposition de  
contrat de  
maintenance**

