olist

PRESENTATION

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 ?

JADEMAYALB

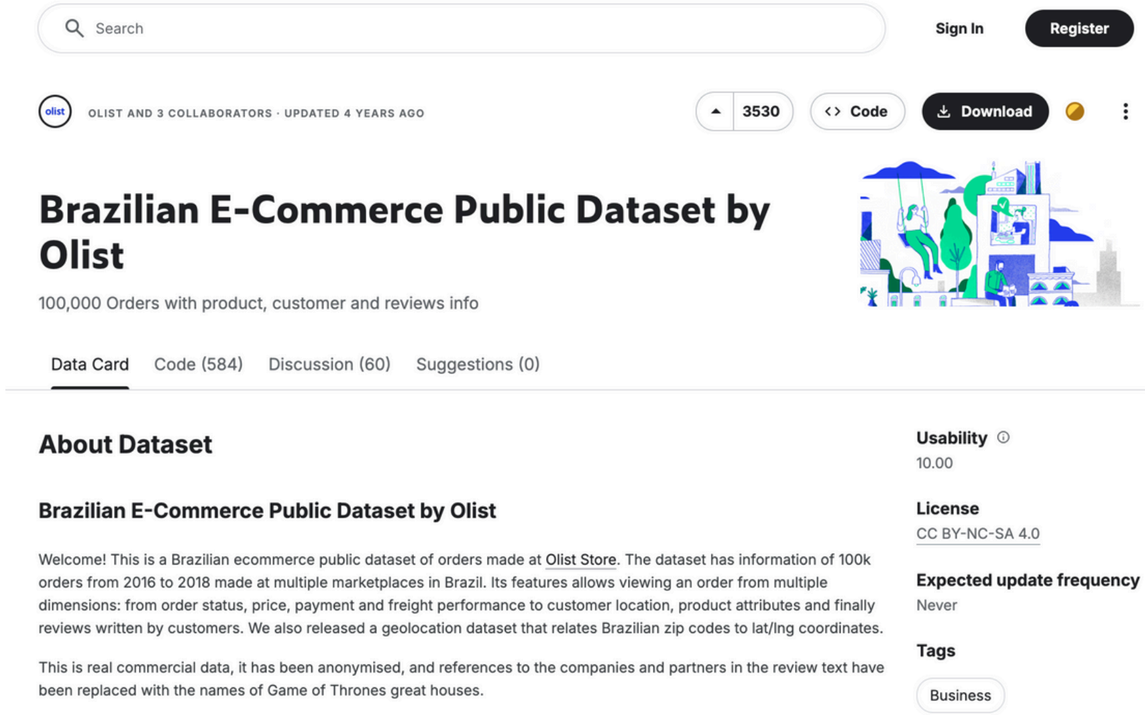
Agenda

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



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.

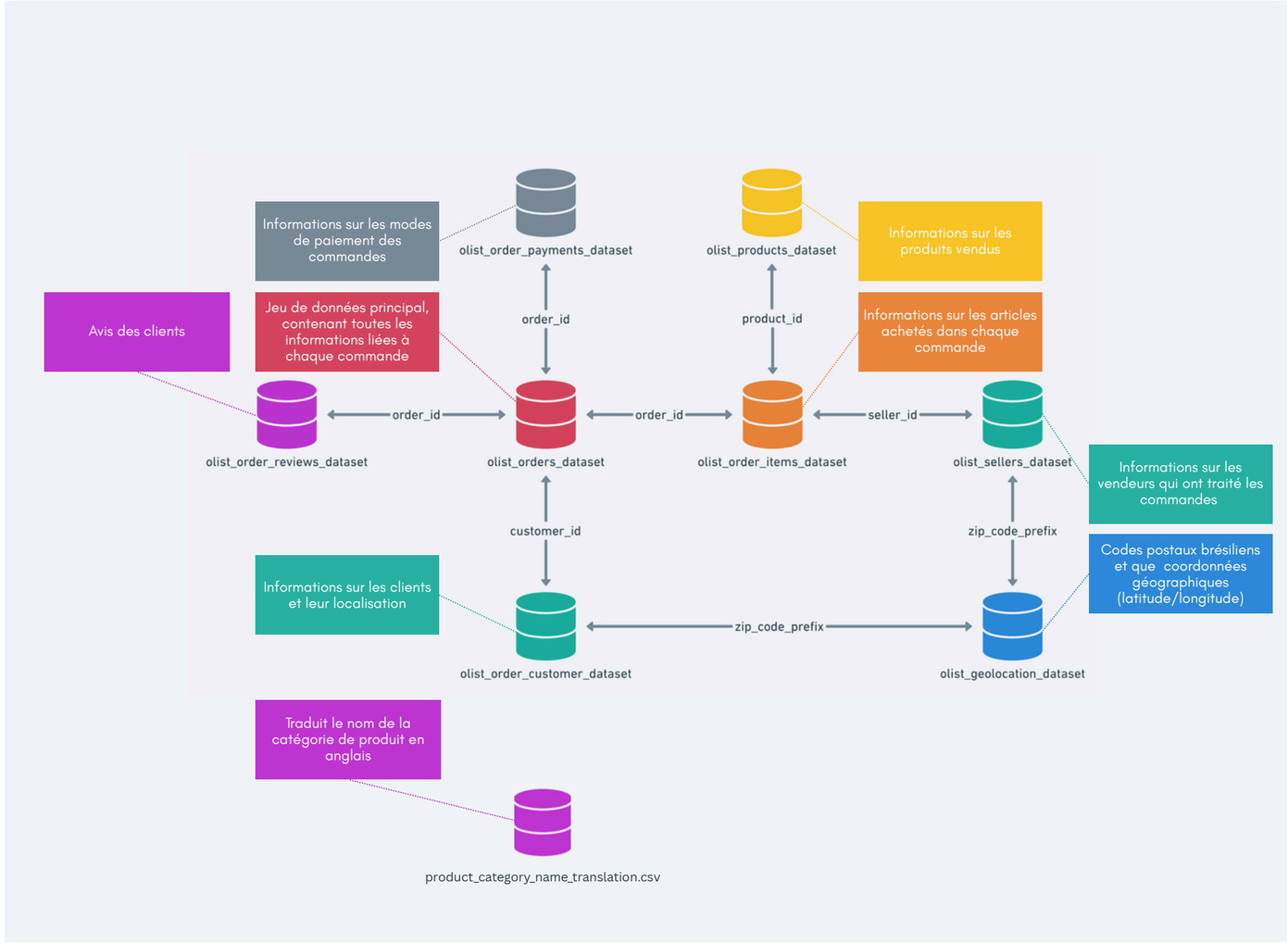
Étape initiale : Nettoyer et explorer les données pour comprendre la structure et préparer les modélisations.

Contexte

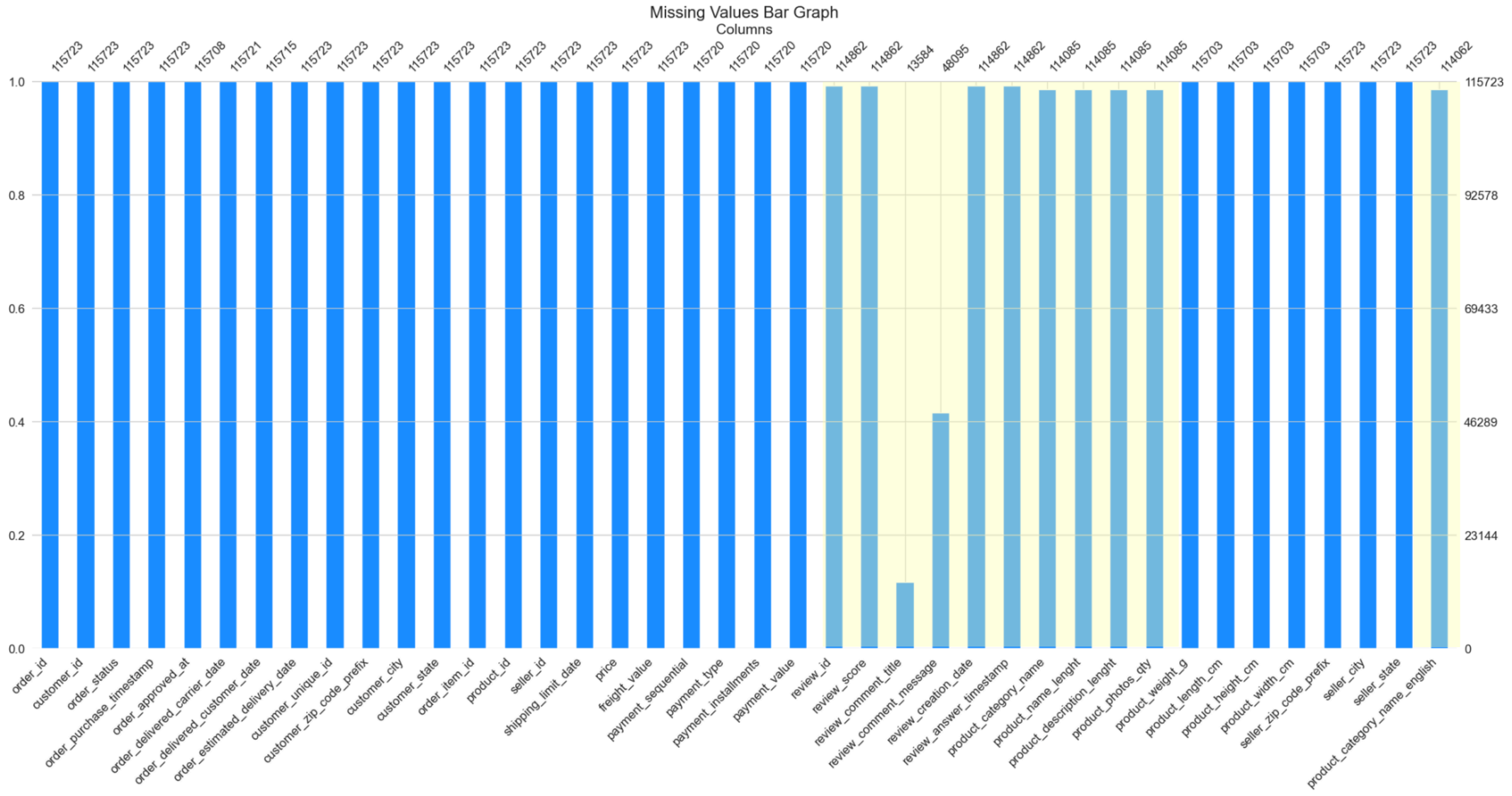
Les données

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

- olist_customers_dataset.csv
- olist_geolocation_dataset.csv
- olist_order_items_dataset.csv
- olist_order_payments_dataset.csv
- olist_order_reviews_dataset.csv
- olist_orders_dataset.csv
- olist_products_dataset.csv
- olist_sellers_dataset.csv
- product_category_name_translation.csv



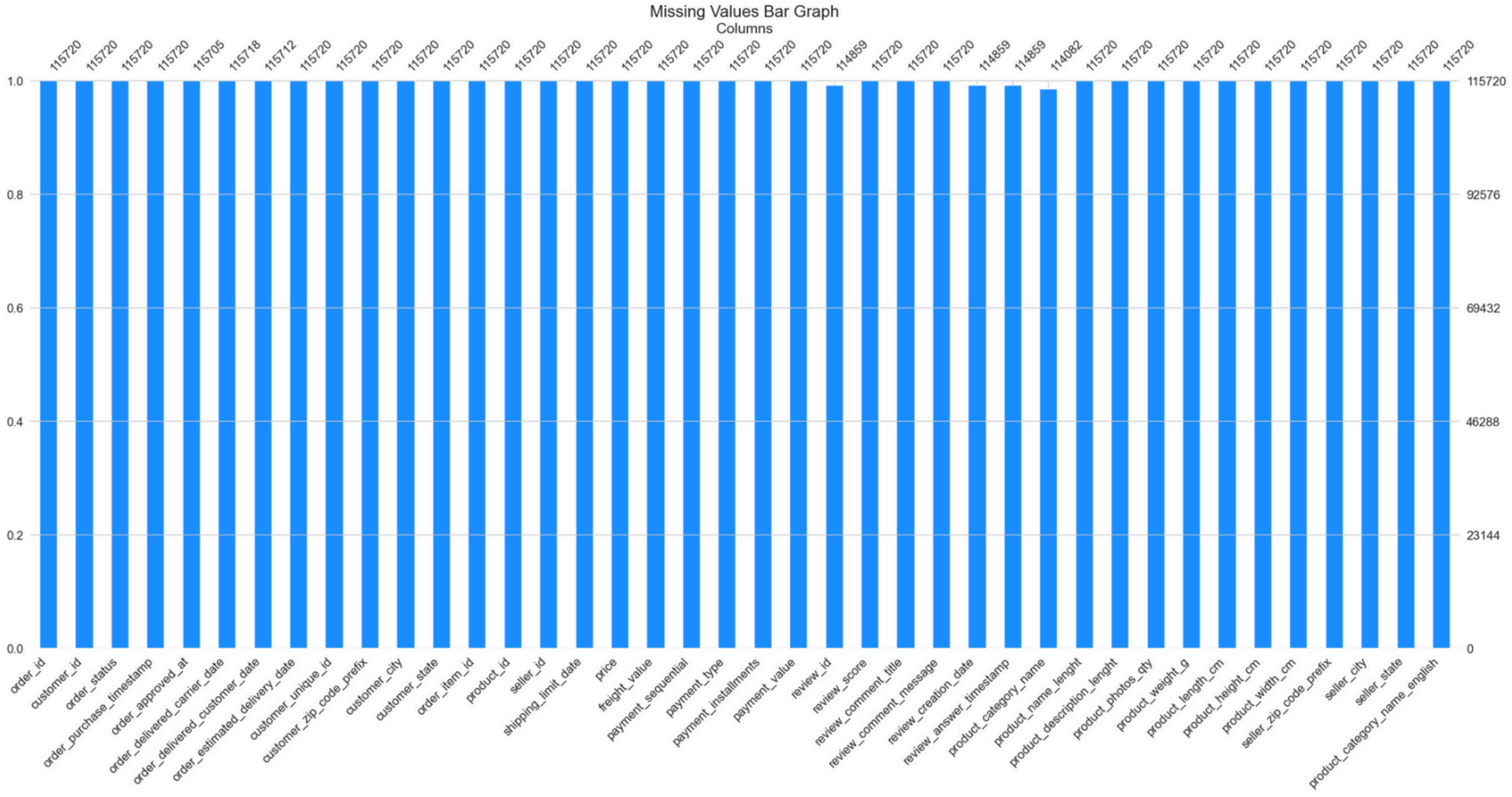
Les données


Missing Values Bar Graph
Columns
115723 115723 115723 115723 115708 115721 115715 115723 115723 115723 115723 115723 115723 115723 115723 115723 115723 115723 115723 115720 115720 115720 115720 114862 114862 13584 48095 114862 114862 114085 114085 114085 114085 115703 115703 115703 115703 115723 115723 115723 114062
1.0
0.8
0.6
0.4
0.2
0.0
115723
92578
69433
46289
23144
0
order_id
customer_id
order_status
order_purchase_timestamp
order_approved_at
order_delivered_carrier_date
order_delivered_customer_date
order_estimated_delivery_date
customer_unique_id
customer_zip_code_prefix
customer_city
customer_state
order_item_id
product_id
seller_id
shipping_limit_date
price
freight_value
payment_sequential
payment_type
payment_installments
payment_value
review_id
review_score
review_comment_title
review_comment_message
review_creation_date
review_answer_timestamp
product_category_name
product_name_lenght
product_description_lenght
product_photos_qty
product_weight_g
product_length_cm
product_height_cm
product_width_cm
seller_zip_code_prefix
seller_city
seller_state
product_category_name_english

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


Missing Values Bar Graph
Columns
1.0
0.8
0.6
0.4
0.2
0.0
115720
92576
69432
46288
23144
0
order_id
customer_id
order_status
order_purchase_timestamp
order_approved_at
order_delivered_carrier_date
order_delivered_customer_date
order_estimated_delivery_date
customer_unique_id
customer_zip_code_prefix
customer_city
customer_state
order_item_id
product_id
seller_id
shipping_limit_date
price
freight_value
payment_sequential
payment_type
payment_installments
payment_value
review_id
review_score
review_comment_title
review_comment_message
review_creation_date
review_answer_timestamp
product_category_name
product_name_lenght
product_description_lenght
product_photos_qty
product_weight_g
product_length_cm
product_height_cm
product_width_cm
seller_zip_code_prefix
seller_city
seller_state
product_category_name_english

Consolidation de plusieurs jeux de données

Données relatives à la commande du client

Un client unique

Clients
Commandes
Produits
Paiments
Vendeurs
Avis clients

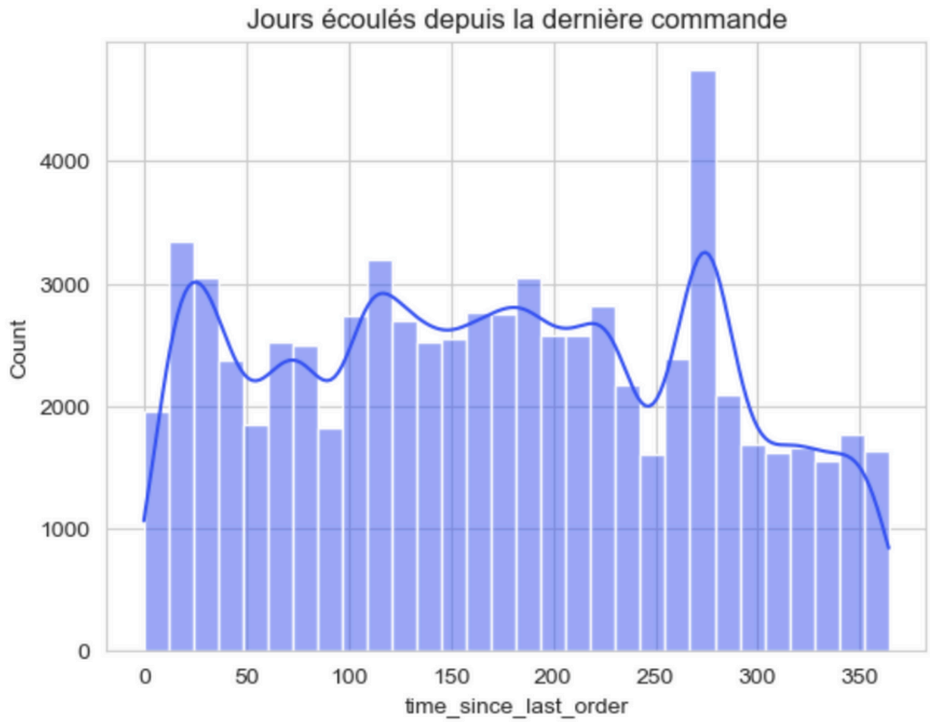
Jeu de données clients agrégé

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



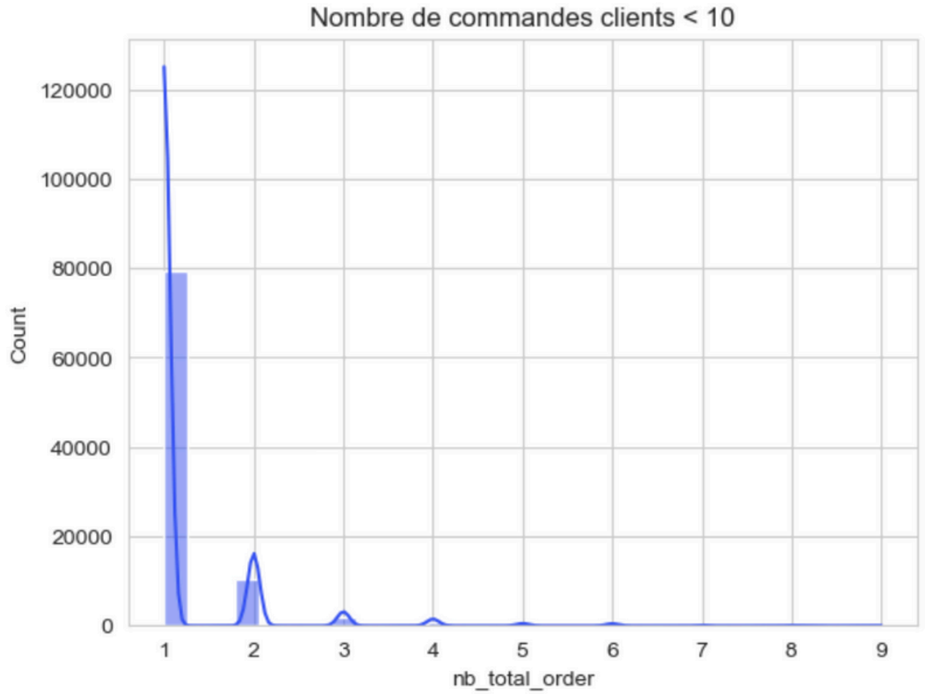
Feature Engineering

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

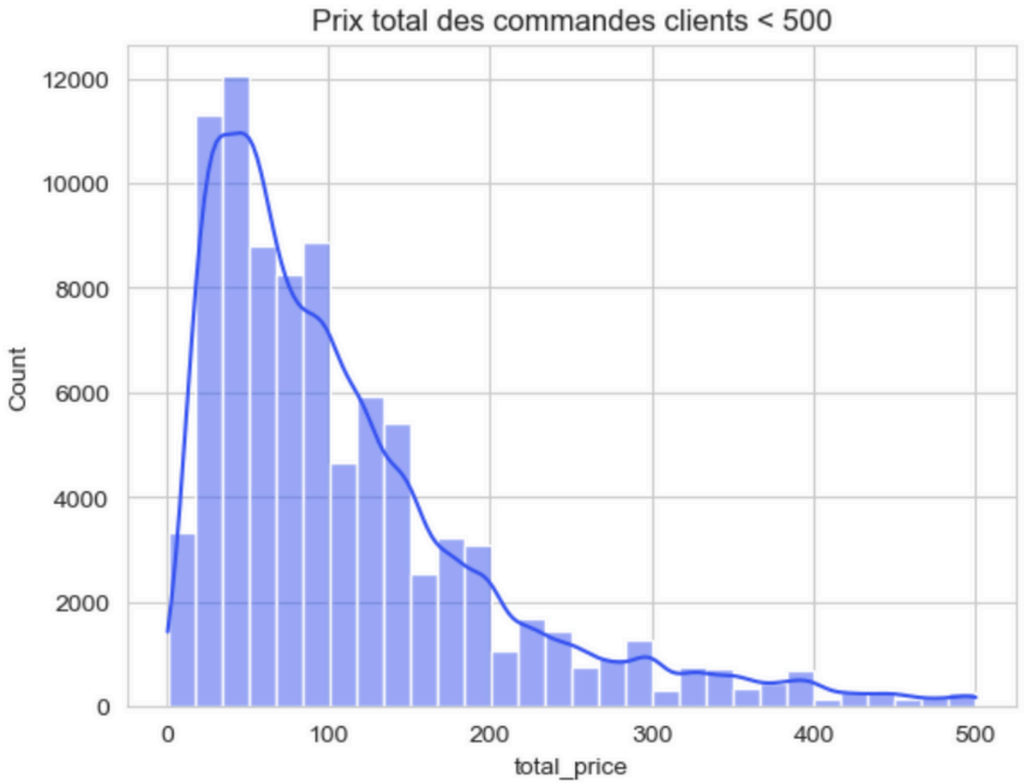


Feature Engineering

Le montant d'achat

Prix total
total_price

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

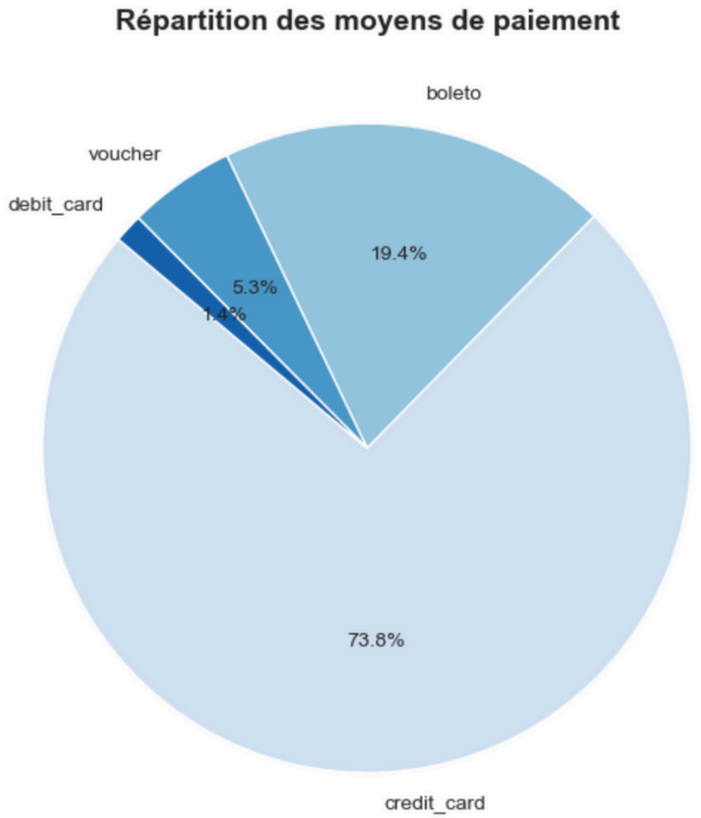


Feature Engineering

Les moyens de paiement client

Nombre de paiements
payment_type

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



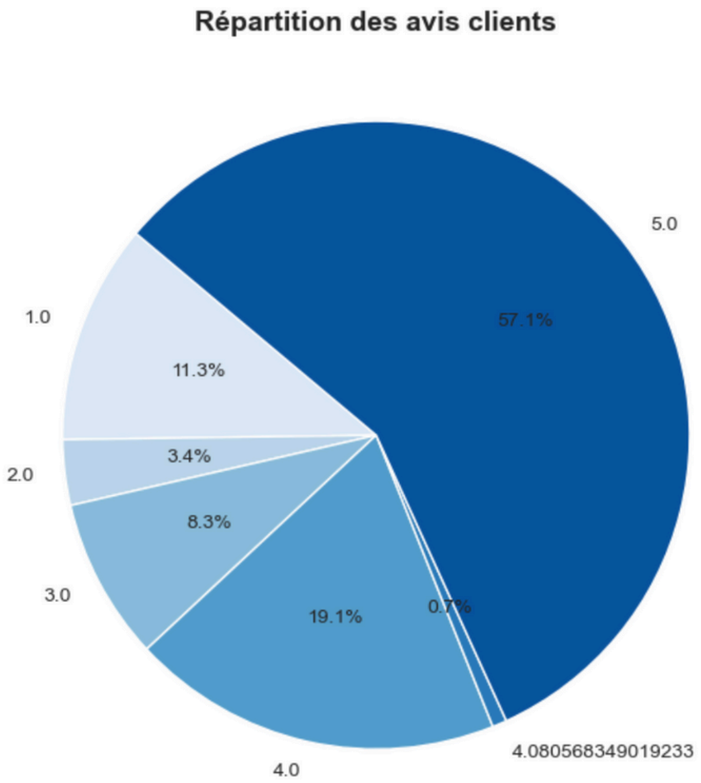
Feature Engineering

La satisfaction client

Note moyenne attribuée par le client

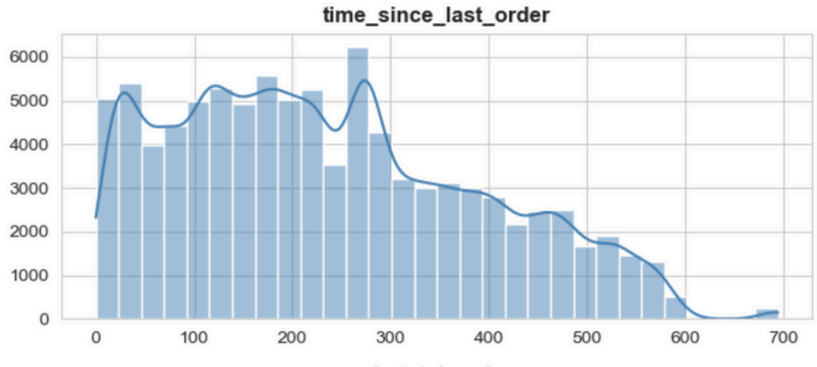
mean_review_score

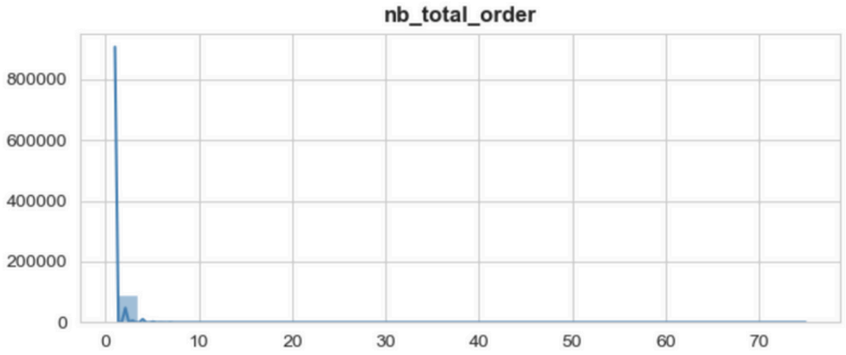
Objectif : Mesurer le niveau de satisfaction moyen du client vis-à-vis de ses achats.
Identifier les clients satisfaits et potentiellement insatisfaits

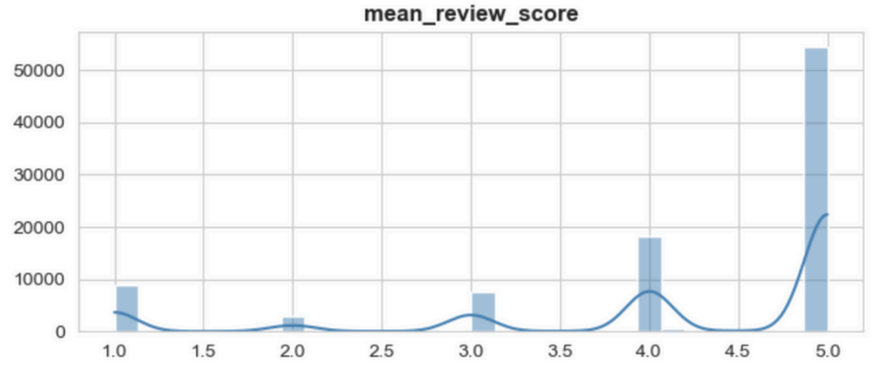


Feature Engineering

Distribution des variables quantitatives

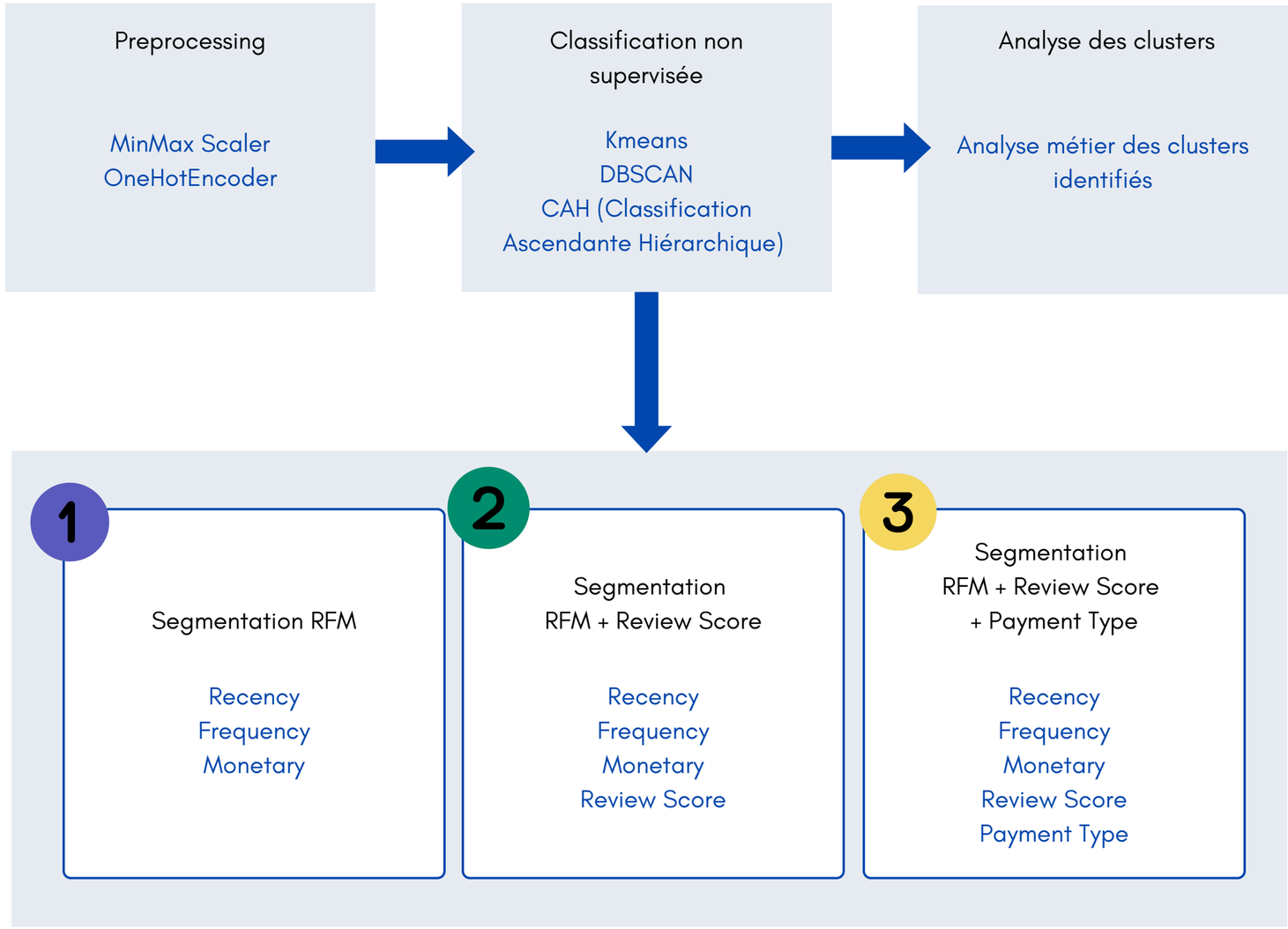






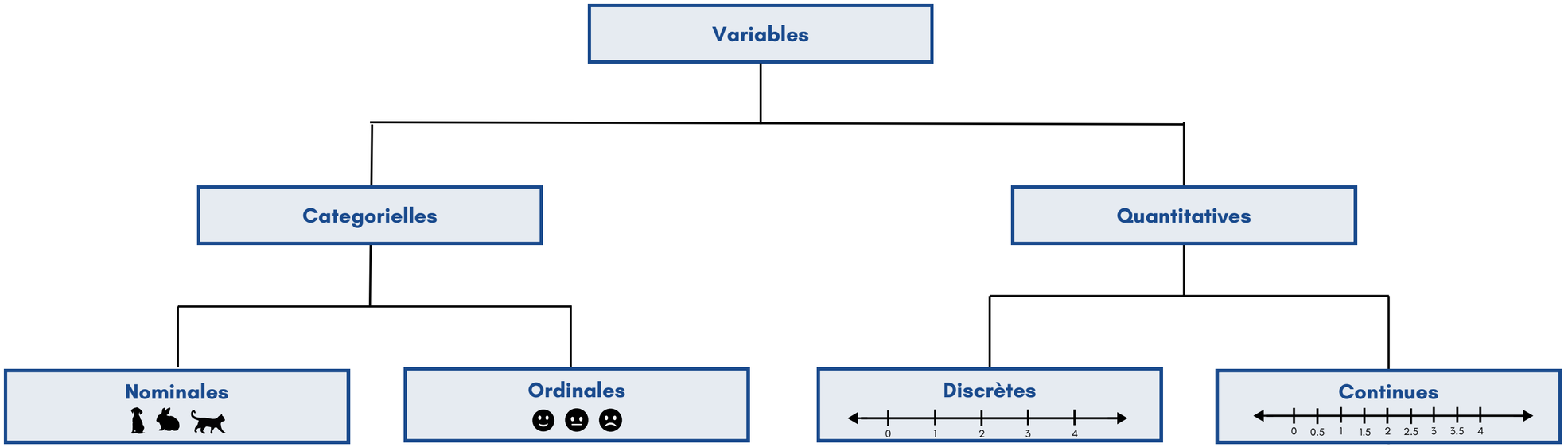
Exploration des données

Modélisation & Choix du modèle final



Pistes de modélisation

Preprocessing



payment_type

nb_total_order

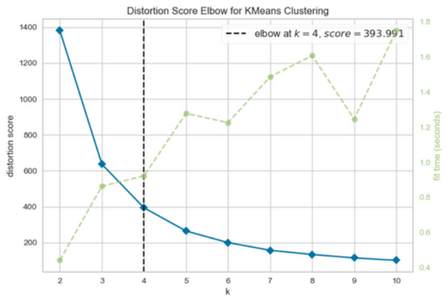
time_since_last_order
total_price
mean_review_score

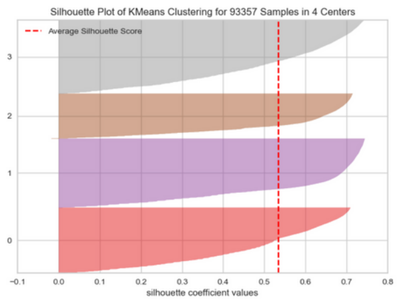
Encodage One-Hot

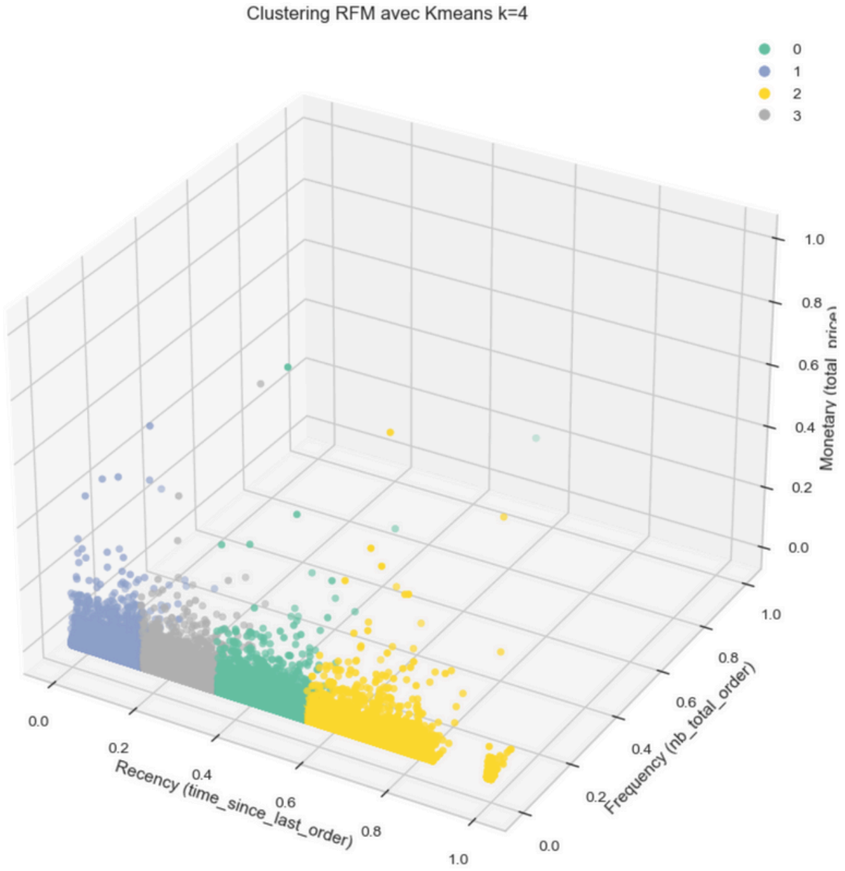
Normalisation Robust Scaler

Kmeans

Paramètres du modèle	Silhouette Score	Davies Bouldin Score	Calinski Harabasz Score
Kmeans = 4	0.53	0.55	328732



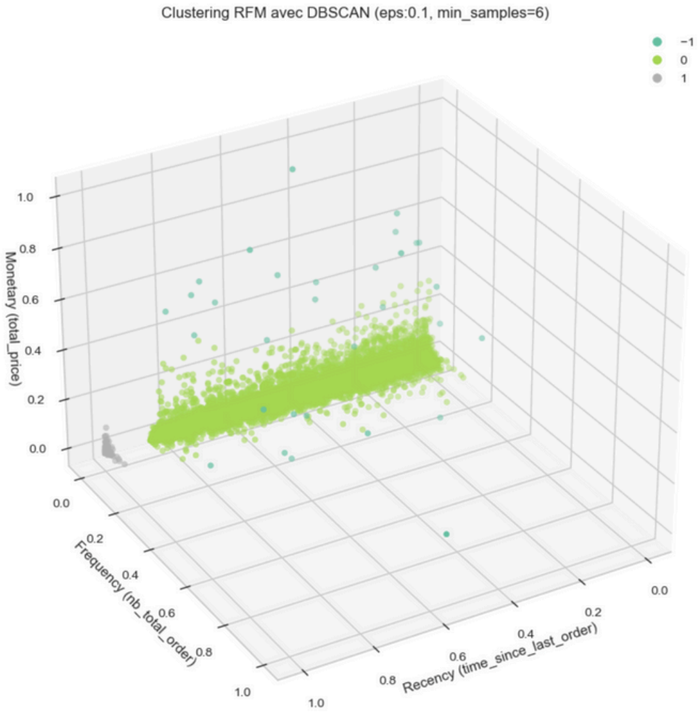


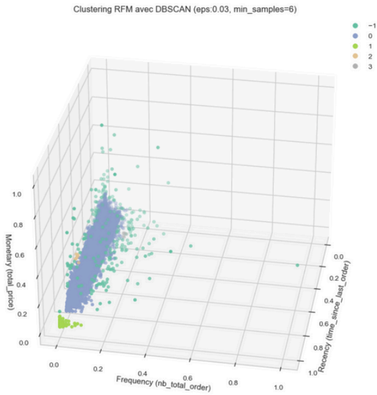


Segmentation RFM

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

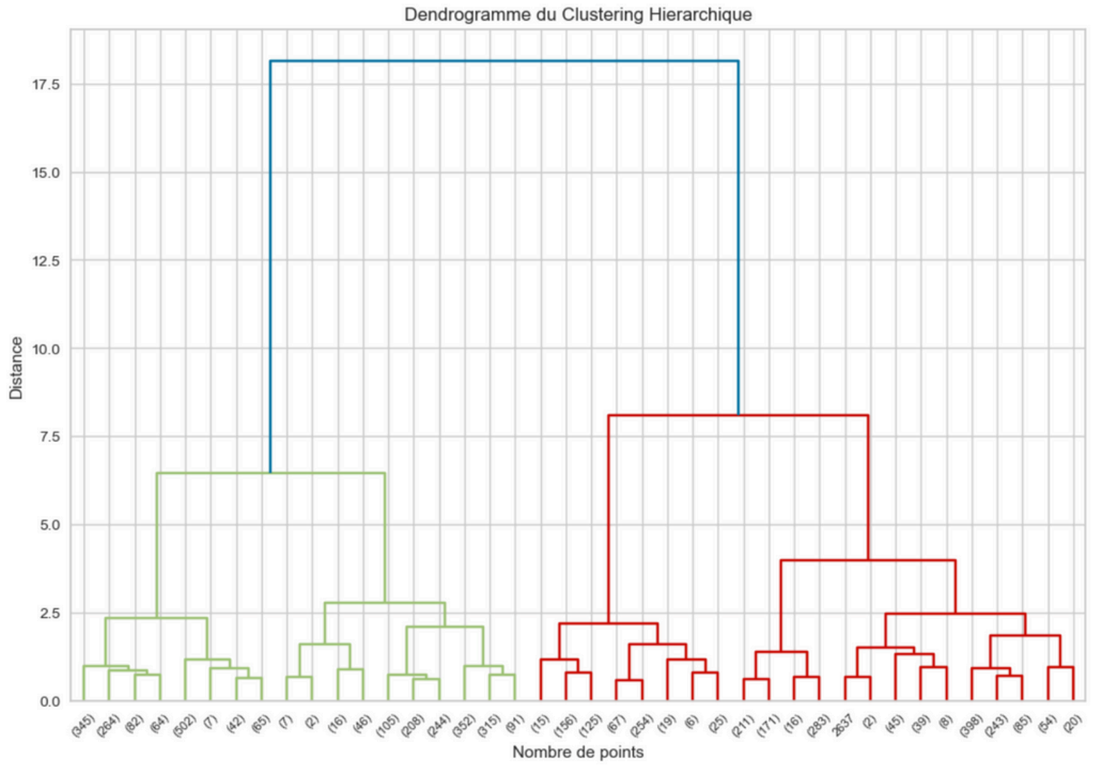




Segmentation RFM

CAH (Classification Ascendante Hiérarchique)

Paramètres du modèle	Silhouette	David Bouldin Score	Calinksi 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



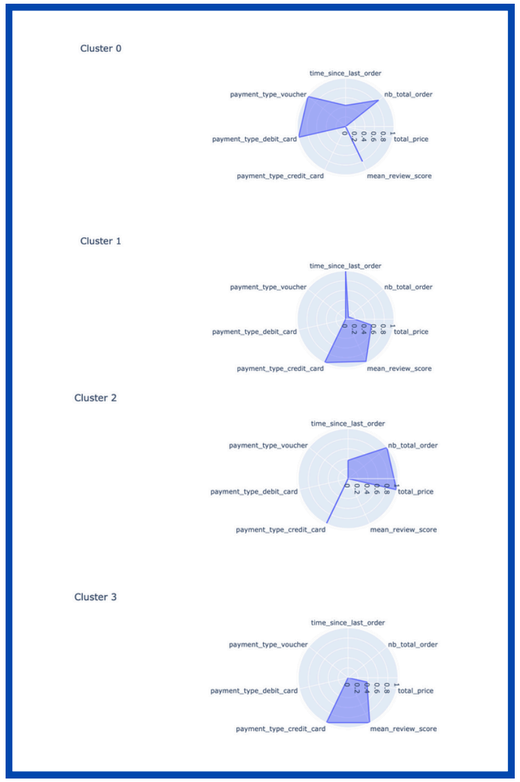
Segmentation RFM

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

RFM + Score

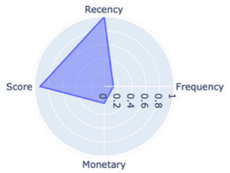
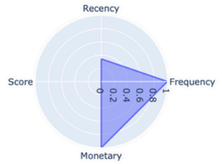
RFM + Score + Payment Type



Autres segmentations

RFM + Score

Kmeans = 4

Client inactif ou en perte de vitesse	Nouveau client, potentiellement à fidélisé	Client précieux en danger	Prospect qui n'a jamais acheté
			
• **Campagnes de réactivation** • **Campagne email avec incitation** • **Campagnes de retargeting**	• **Envoi d'un email de bienvenue personnalisé** • **Proposer un avantage pour le 2e achat** • **Campagnes de nurturing**	• **Envoi d'un email personnalisé « Vous nous manquez » avec offre exclusive** • **Mise en place d'un programme de fidélité VIP** • **Campagnes de retargeting personnalisées**	• **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

1

Découpage temporel pour suivre l'évolution des clients

Simuler la vie réelle des clients : ils changent avec le temps

2

Création du modèle de départ (Segmentation initiale)

Obtenir une segmentation de référence pour comparer les évolutions futures

3

Test de la stabilité dans le temps

Comparer les deux segmentations pour voir si les groupes restent stables ou évoluent.

4

Comparaison des groupes

Évaluer la stabilité d'une segmentation clients dans le temps

5

Fréquence de mise à jour de la segmentation

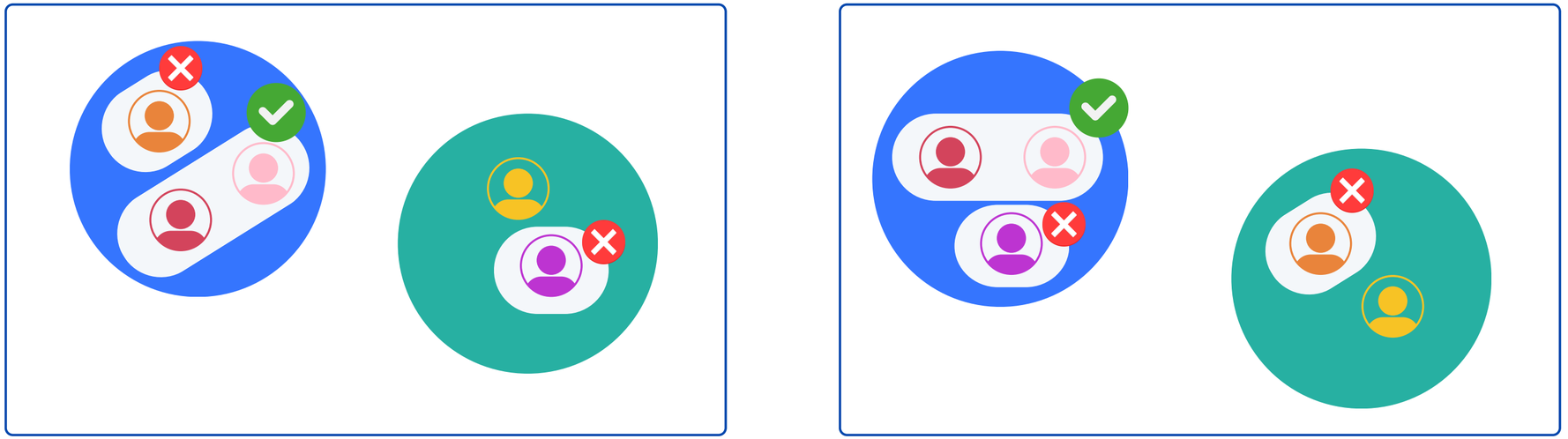
On observe au bout de combien de temps le score ARI devient trop faible

Simulation du modèle dans le temps

Comment évaluer la qualité d'une segmentation client ?

Segmentation de référence

Segmentation automatique



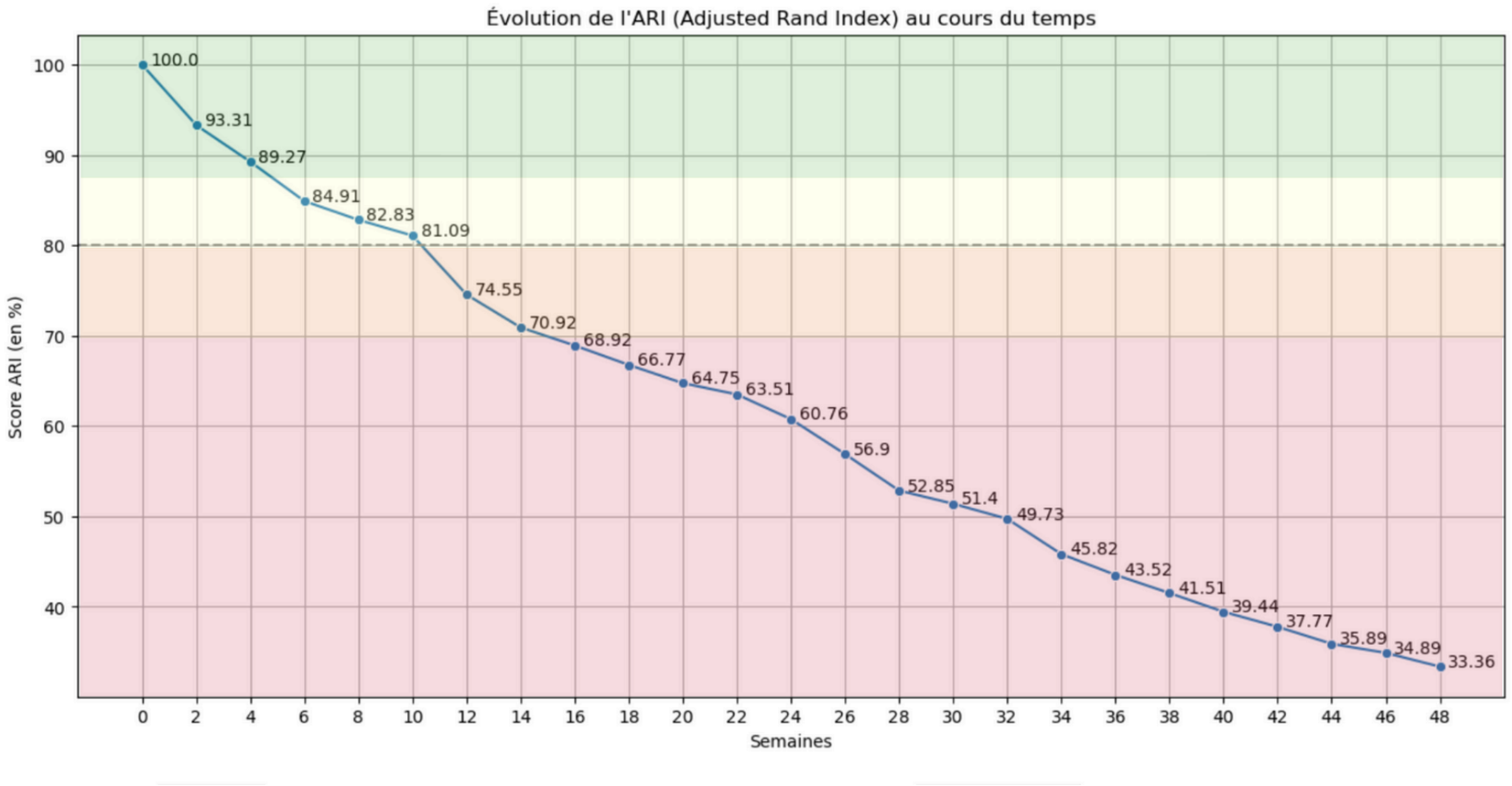
ARI - Adjusted Rand Index

Contibue 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