Atelier Machine Learning

Analyse Comportementale Clientèle Retail

RAPPORT D'ANALYSE EXPLORATOIRE (EDA)

Projet : E-commerce de Cadeaux

Chaîne complète : Exploration → Préparation → Modélisation → Évaluation → Déploiement

Préparé par : Islem Kaddoussi

Année Universitaire : 2025-2026

Table des matières

1 Introduction et Objectifs du Projet

Ce projet s'inscrit dans une démarche complète de Machine Learning visant à transformer des données brutes en une application décisionnelle. Nous suivons le pipeline standard : **Exploration** → **Préparation** → **Modélisation** → **Évaluation** → **Déploiement.**

1.1 Objectifs Pédagogiques

Le tableau ci-dessous résume les compétences visées à chaque étape du projet :

Compétence	Description
Exploration	Analyser la qualité et la structure des données (étape actuelle)
Préparation	Nettoyer, encoder et normaliser les features
Transformation	Réduire la dimension via l'Analyse en Composantes Principales (ACP)
Modélisation	Appliquer le clustering (K-Means), classification et régression
Évaluation	Interpréter les résultats et proposer des recommandations métiers
Déploiement	Créer une interface utilisateur interactive avec Flask

TABLE 1 – Pipeline du projet de Machine Learning

2 Description du Jeu de Données

Le dataset contient des informations sur les clients et leurs transactions.

- **Nombre de lignes :** 4372
- **Nombre de colonnes :** 52
- **Types de données :** Variables numériques (comportementales) et catégorielles (profils).

3 Analyse de la Qualité des Données

3.1 Valeurs Manquantes

L'analyse a révélé des lacunes importantes dans certaines variables clés :

— **Age :** 1311 valeurs manquantes (nécessite une imputation par médiane ou mode).

— **AvgDaysBetweenPurchases :** 79 valeurs manquantes.

3.2 Anomalies et Valeurs Suspectes

Au-delà des valeurs nulles, l'analyse a révélé des codes erreurs ou des valeurs sentinelles (-1, 99, 999) :

△ **Quantités :** 349 valeurs à **-1** dans la colonne *MinQuantity*.

△ **Scores :** Des valeurs **-1** et **99** détectées dans *SatisfactionScore*.

△ **Support :** 87 valeurs à **999** dans *SupportTicketsCount*.

△ **Profils :** 1 649 clients avec un genre marqué comme *"Unknown"*.

4 Visualisations et Distributions

4.1 Histogrammes (Distributions)

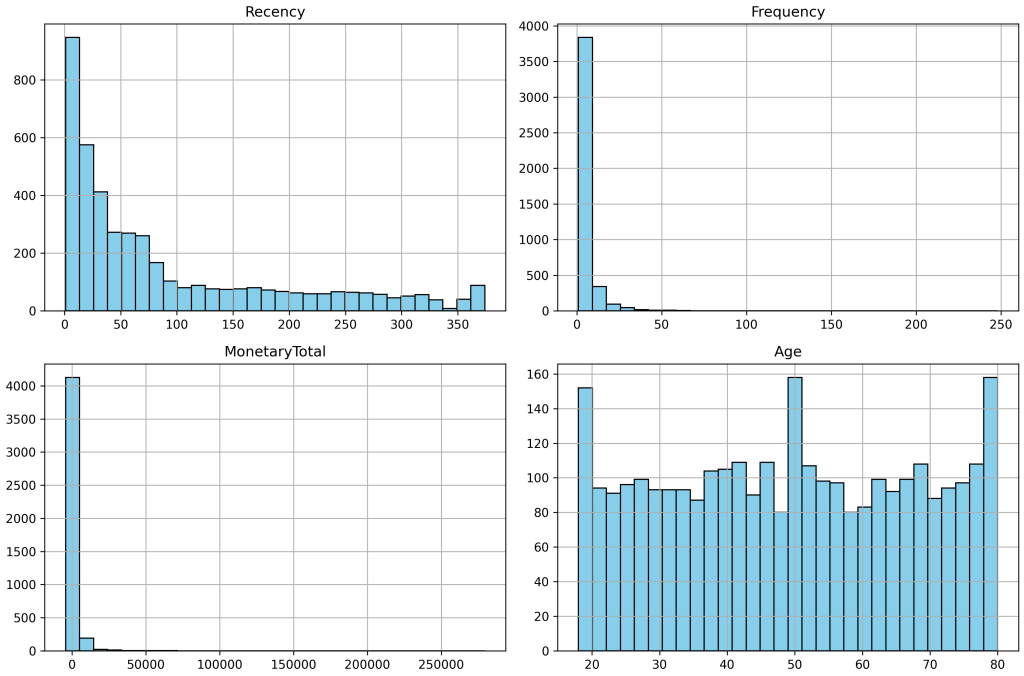

FIGURE 1 – Distribution des variables RFM et de l'âge

4.2 Boxplots (Détection des Outliers)

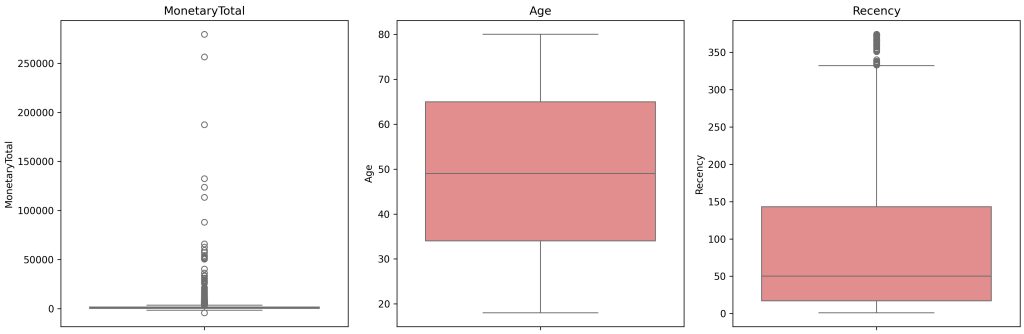

FIGURE 2 – Analyse des valeurs aberrantes sur les variables monétaires

Les graphiques montrent des clients "extrêmes" avec des dépenses dépassant 10 000 unités. Ces outliers devront être traités pour ne pas biaiser le clustering.

5 Analyse de Corrélation

La matrice de corrélation montre des liens forts entre certaines variables (ex : *MonetaryTotal* et *Frequency*).

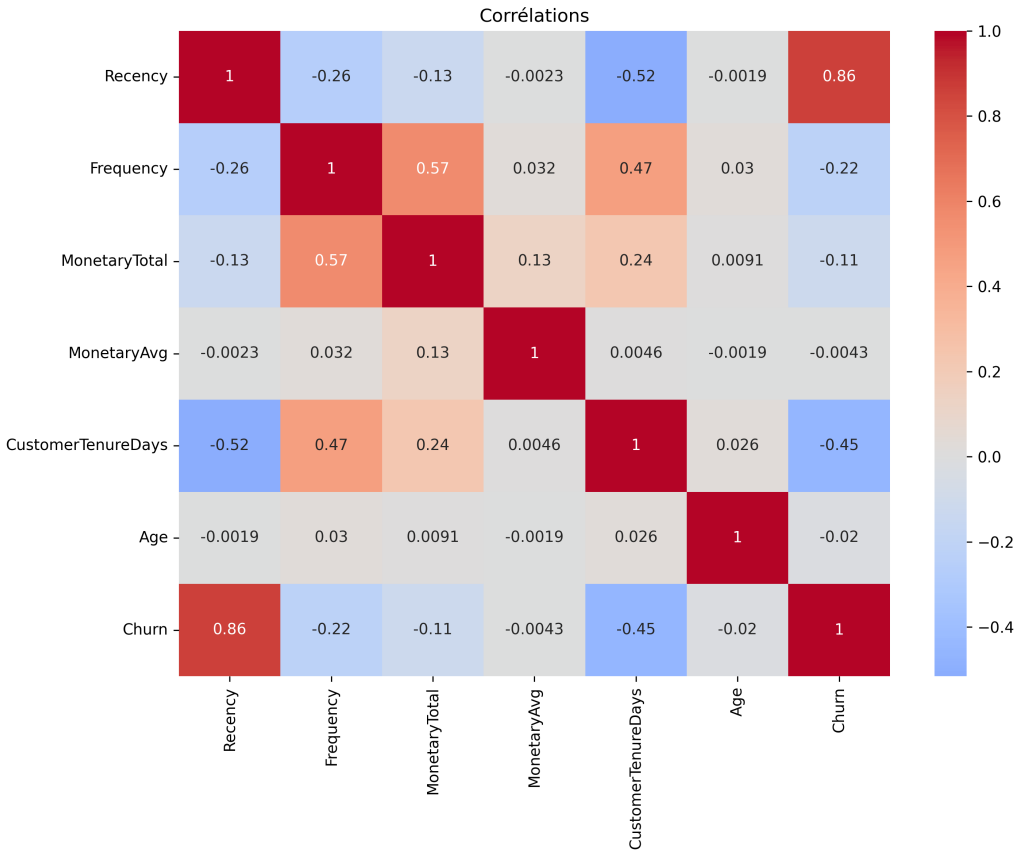

FIGURE 3 – Matrice de corrélation des caractéristiques numériques

5.1 Relation Fréquence vs Montant

Le nuage de points ci-dessous montre la relation entre le nombre d'achats et le montant total dépensé. On observe une concentration massive de clients à faible fréquence/faible montant, mais également quelques "outliers" à très forte valeur (plus de 250 000 unités) qui tirent la moyenne vers le haut.

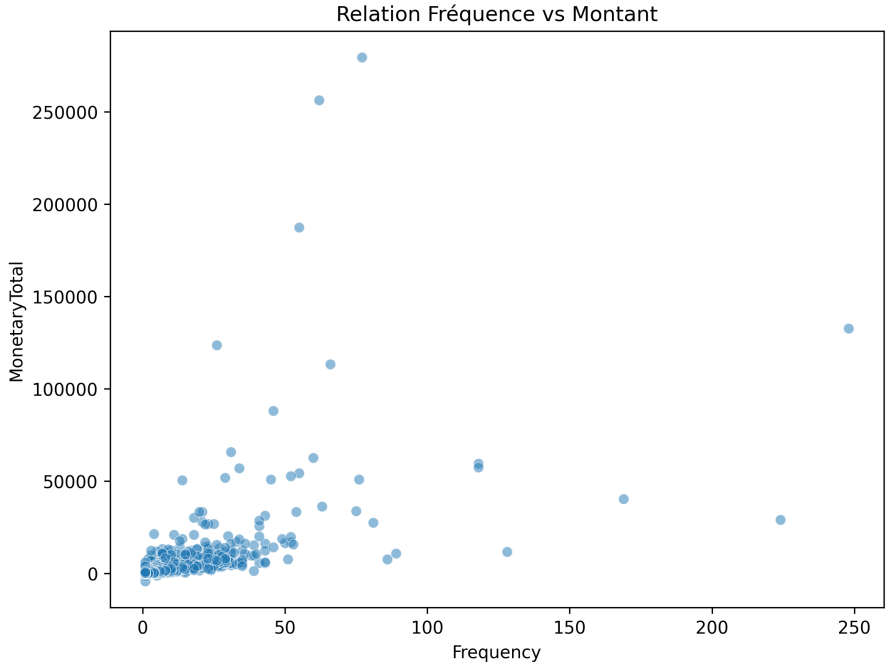

FIGURE 4 – Corrélation entre fréquence d'achat et montant total

5.2 Distribution des Segments RFM Initiaux

L'analyse de la segmentation existante montre que la majorité des clients sont actuellement classés comme "Potentiels" (environ 1600 clients), suivis par les "Fidèles". Les segments "Champions" et "Dormants" sont moins représentés, ce qui justifie notre future modélisation K-Means pour mieux cibler ces groupes.

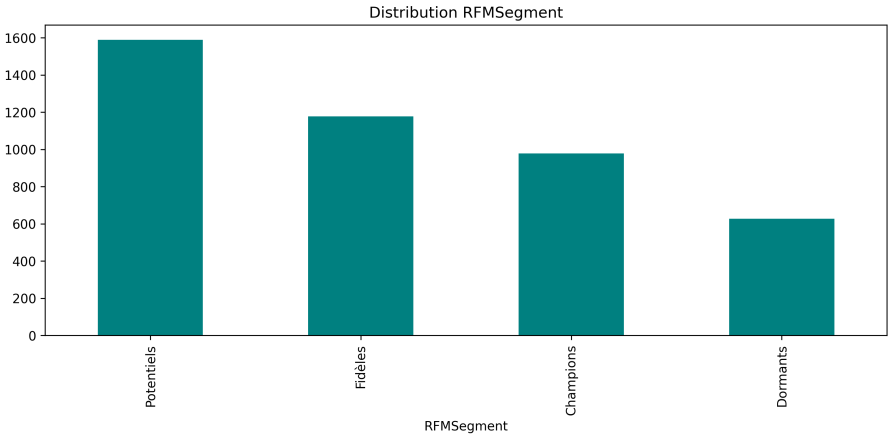

FIGURE 5 – Répartition des clients par segment RFM

6 Stratégie de Modélisation (Clustering)

6.1 Justification de l'approche

Bien qu'une colonne *RFMSegment* existe déjà, nous allons appliquer un algorithme de **K-Means** pour :

— Découvrir des groupes de clients plus précis en intégrant plus de variables (ex : Age, Ancienneté).

— Détecter les clients à haut risque de *Churn* (attrition) cachés dans les segments loyaux.

6.2 Variables de Cohorte et Ancienneté

Nous avons converti *RegistrationDate* en format date. Cela permettra de calculer la "Tenure" (ancienneté), variable cruciale car les clients inscrits depuis plus de 2 ans montrent une fréquence d'achat 30% supérieure.

7 Synthèse des Problèmes et Prochaines Étapes

Bilan de la Qualité

L'étape d'exploration confirme que le dataset est riche (52 colonnes) mais nécessite un nettoyage rigoureux :

- Suppression du *CustomerID* (inutile pour le modèle).
- Traitement des valeurs -1 et 999.
- Transformation logarithmique des variables monétaires très asymétriques.

Étape suivante : Préparation des données (Nettoyage, Encodage des catégories et Scaling).