Atelier Machine Learning

Analyse Comportementale Clientèle Retail

RAPPORT D'ANALYSE

Projet : E-commerce de Cadeaux

Exploration → Préparation → Modélisation

Préparé par : [TON NOM]

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

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

TABLE 1 – Pipeline du projet de Machine Learning

2 Description du Jeu de Données

- **Nombre de lignes :** 4 372 clients
- **Nombre de colonnes :** 52 features
- **Types de données :** 34 numériques, 18 catégorielles
- **Taille mémoire :** 1,7 MB

3 Analyse Exploratoire des Données (EDA)

3.1 Valeurs Manquantes

Alertes

— **Age :** 1 311 valeurs manquantes (**30,0%**)

— **AvgDaysBetweenPurchases :** 79 valeurs manquantes (1,8%)

3.2 Anomalies et Valeurs Suspectes

△ **SatisfactionScore :** 115 valeurs **-1**, 114 valeurs **99**

△ **SupportTicketsCount :** 43 valeurs **-1**, 87 valeurs **999**

△ **Gender :** 1 649 clients **"Unknown"** (37,7%)

△ **Quantités négatives :** 349 valeurs **-1** dans MinQuantity, 15 dans TotalQuantity, 16 dans AvgQuantityPerTransaction

△ **Codes 99 :** Recency (10), MonetaryMax (11), CustomerTenureDays (7), FirstPurchaseDaysAgo (7), UniqueProducts (13), UniqueDescriptions (10), TotalTransactions (11)

3.3 Visualisations

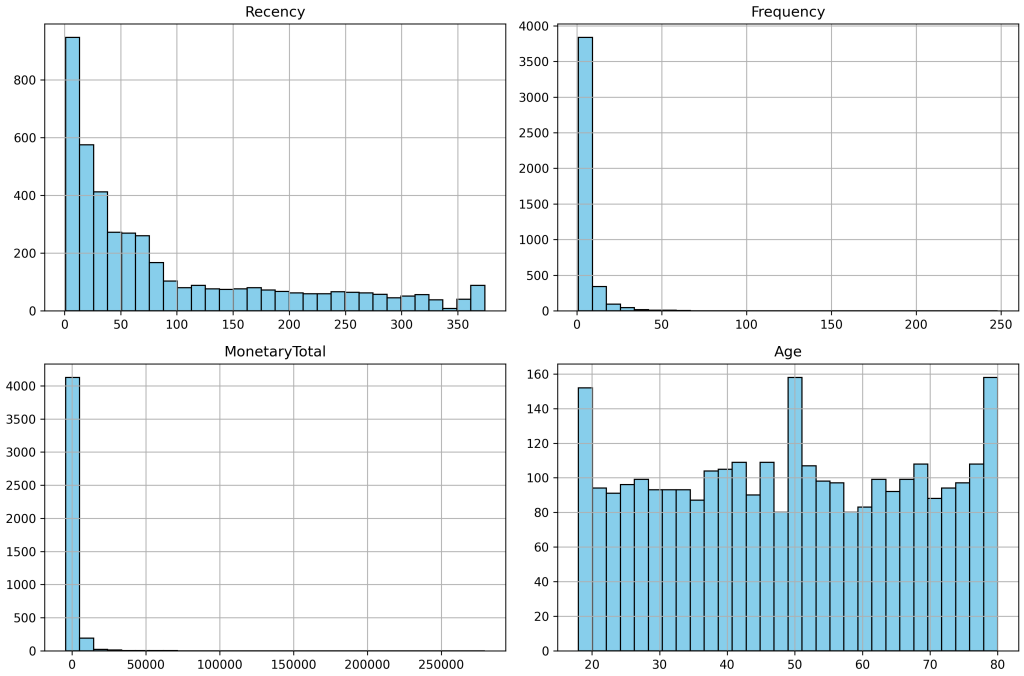

FIGURE 1 – Distribution des variables RFM et de l'âge

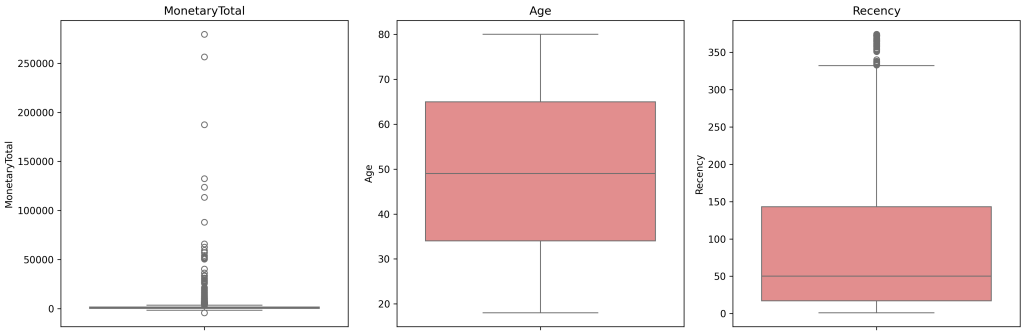

FIGURE 2 – Détection des outliers sur les variables monétaires

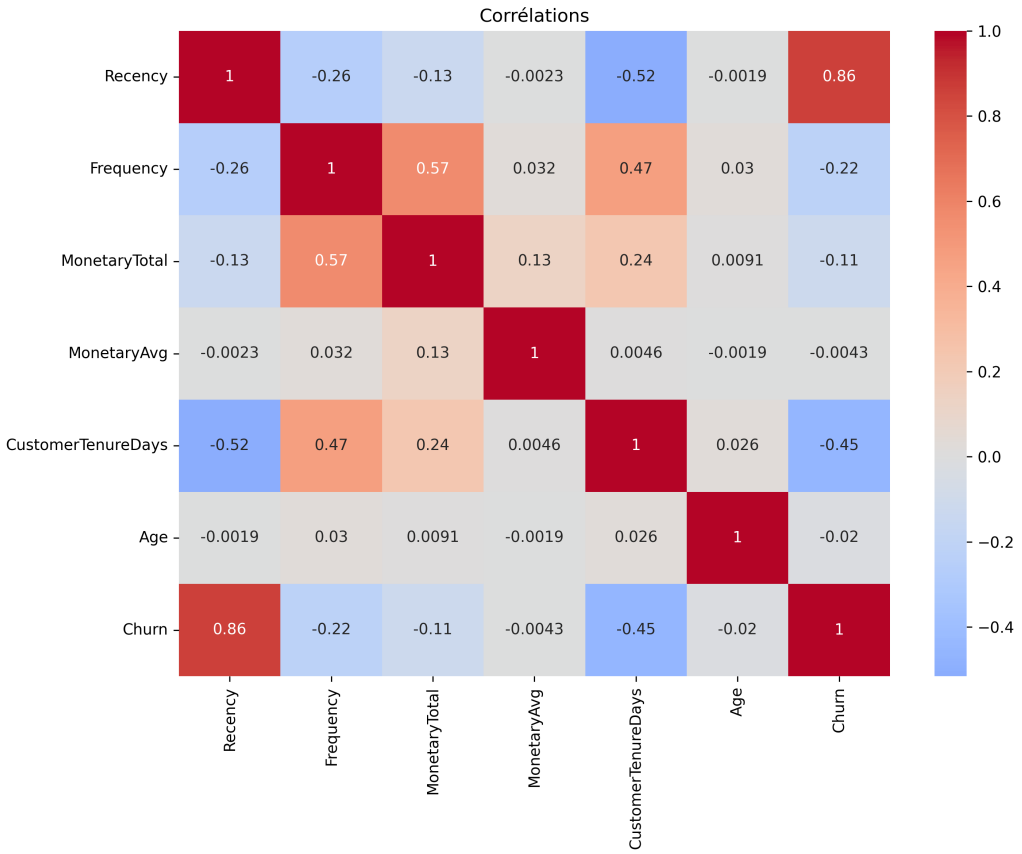

FIGURE 3 – Matrice de corrélation

Corrélations majeures :

— Recency ↔ Churn : $r = 0,86$ (forte)

— Frequency ↔ MonetaryTotal : $r = 0,57$ (modérée)

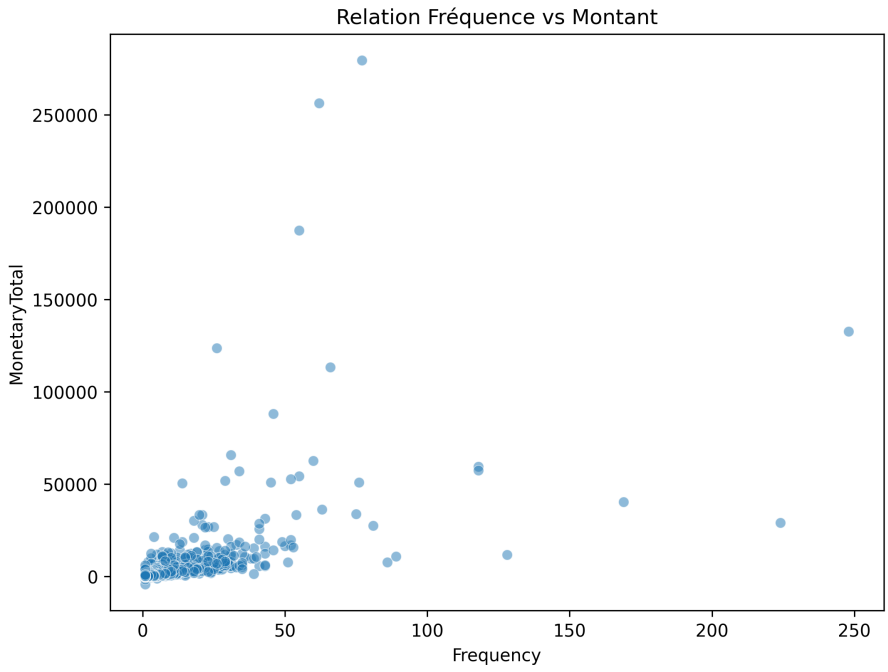

FIGURE 4 – Relation Fréquence vs Montant

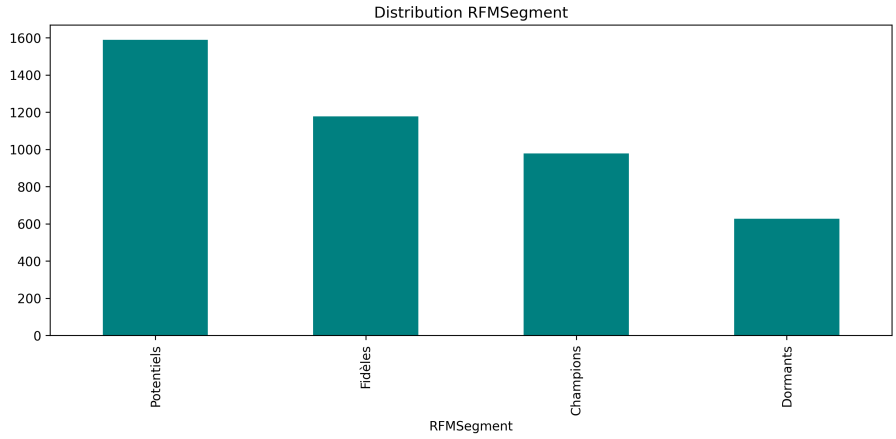

FIGURE 5 – Distribution des segments RFM

Segment	Effectif	%	Profil
Potentiels	1 589	36,3%	À activer
Fidèles	1 177	26,9%	Cœur de cible
Champions	979	22,4%	VIP
Dormants	627	14,3%	Risque départ

4 Préparation des Données (Préprocessing)

4.1 Étape 1 : Parsing des Dates (Guide §7)

Problème : Formats inconsistants (`17/07/10`, `2010-10-04`, `12/09/2009`)
Méthode : `pd.to_datetime(dayfirst=True, errors='coerce')`
Résultat : 4 nouvelles features : RegYear, RegMonth, RegDay, RegWeekday

4.2 Étape 2 : Imputation des Valeurs Manquantes (Guide §6)

Feature	Méthode	Valeur
Age (30% manquants)	Médiane	49,0 ans
AvgDaysBetweenPurchases	Médiane	25,4 jours
SupportTicketsCount (-1, 999)	Médiane après nettoyage	2,0
SatisfactionScore (-1, 99)	Médiane après nettoyage	3,0
Gender (Unknown)	Mode	M
Codes 99 (7 colonnes)	Médiane	par colonne
Valeurs négatives (-1)	Médiane positive	par colonne

TABLE 2 – Imputation des valeurs manquantes et aberrantes

4.3 Étape 3 : Suppression des Features Inutiles (Guide §5)

— **NewsletterSubscribed** : Supprimée (constante "Yes")
— **CustomerID** : Supprimée (identifiant non prédictif)

4.4 Étape 4 : Extraction LastLoginIP (Guide §5 + §7)

Méthode : Extraction du premier octet de l'IP
Exemple : "192.168.1.45" → "192" (feature IP_Prefix)

4.5 Étape 5 : Multicolinéarité (Guide §2)

Méthode : Matrice de corrélation avec seuil $|r| > 0,8$
Vérification : MonetaryTotal vs MonetaryAvg : $r = 0,73$
Décision : Conservation des deux variables (corrélation $< 0,8$)

4.6 Étape 6 : Feature Engineering (Guide §4)

3 nouvelles features créées :

$$MonetaryPerDay = \frac{MonetaryTotal}{Recency + 1}$$

$$AvgBasketValue = \frac{MonetaryTotal}{Frequency}$$

$$TenureRatio = \frac{Recency}{CustomerTenureDays + 1}$$

4.7 Étape 7 : Standardisation (Guide §1)

Méthode : StandardScaler (centrage-réduction)
Formule : $Z = \frac{X-\mu}{\sigma}$
Anti Data Leakage : Fit uniquement sur train, transform sur test

— Features standardisées : 48 colonnes

— Features binaires conservées : 6 colonnes

— Cible préservée : Churn (pas de standardisation)

4.8 Récapitulatif des Transformations

Étape	Entrée	Sortie	Guide
Chargement	-	52 cols	-
Parsing dates	52	54 cols	§7
Imputation	54	54 cols	§6
Suppression	54	52 cols	§5
Extraction IP	52	52 cols	§5+§7
Multicolinéarité	52	52 cols	§2
Feature Eng.	52	55 cols	§4
Standardisation	55	55 cols	§1

TABLE 3 – Évolution des dimensions du dataset

4.9 Vérification du Déséquilibre

— **Churn** : 67,2% fidèles / 32,8% partis (modéré)

— **AccountStatus** : 93,8% Active (quasi-constante)

5 Conclusion et Perspectives

Bilan

L'étape d'exploration et de préparation est terminée. Les données sont maintenant :

- Nettoyées (valeurs manquantes et aberrantes traitées)
- Enrichies (3 nouvelles features métier)
- Standardisées (même échelle pour toutes les variables)

Prochaine étape : Clustering avec K-Means pour identifier les segments clients actionnables.