UNIVERSITÉ SORBONNE PARIS NORD

Master Web Intelligence
et Science de données

Projet De Statistiques Exploratoires Multidimensionnelles

Intitulé :

Analyse en Composantes Principales (ACP) - Analyse de Cartes de Crédit

Présenté par :

AOUAD Abdelkarim & EL OUARDI Abderrahim

Pr. Alj Abdelkamel Professeur (Encadrant)

Année Universitaire 2025–2026

Remerciements

Nous souhaitons remercier chaleureusement notre professeur, **M. Alj Abdelkamel**, pour son accompagnement tout au long de ce projet. Ses conseils avisés, sa bienveillance et sa disponibilité ont été une source d'encouragement et d'inspiration pour nous.

Grâce à son soutien, nous avons pu progresser dans le développement de notre application, apprendre de nouvelles méthodes et renforcer nos compétences. Son approche pédagogique, toujours à l'écoute, nous a permis de mieux comprendre nos erreurs et de trouver des solutions adaptées.

Nous exprimons notre profonde reconnaissance pour tout le temps et l'énergie qu'il a consacrés à nous guider. Ce projet n'aurait pas été le même sans sa précieuse aide.

Table des matières

Table des matières v

Liste des figures vi

Liste des tableaux 1

I Introduction 2

I.1 Définition et domaine d'étude . . . 2

I.2 Problématique . . . 2

I.3 Méthodologie et justification de l'approche . . . 3

II Présentation et description du jeu de données 4

II.1 Description des variables . . . 4

III Fondements théoriques de l'Analyse en Composantes Principales 6

III.1 Cadre théorique et objectifs de l'ACP . . . 6

III.2 Formulation Mathématique de l'ACP . . . 6

III.2.1 Données et Notation . . . 6

III.2.2 Centrage et Réduction . . . 7

III.2.3 Matrice de Covariance . . . 7

III.2.4 Composantes Principales . . . 7

III.3 Valeurs Propres, Vecteurs Propres et Inertie . . . 8

III.3.1 Définitions Fondamentales . . . 8

III.3.2 Inertie Totale . . . 8

III.3.3 Interprétation Géométrique . . . 9

III.4 Propriétés Théoriques des Composantes Principales . . . 9

III.4.1 Orthogonalité . . . 9

III.4.2 Maximisation de la Variance . . . 10

III.4.3 Reconstruction des Données . . . 10

III.4.4 Invariance à la Rotation . . . 10

III.4.5 Théorème de Décomposition Spectrale . . . 10

III.4.6 Stabilité Numérique . . . 11

III.4.7 Dualité de l'ACP . . . 11

III.4.8 Concentration de la Variance . . . 11

IV Analyse Exploratoire des Données **12**
IV.1 Contexte et objectifs . . . 12
IV.2 Statistiques descriptives clés . . . 12
IV.3 Distributions et patterns . . . 13
IV.3.1 Distributions principales . . . 13
IV.3.2 Valeurs extrêmes (outliers) . . . 14
IV.3.3 Focus sur BALANCE . . . 15
IV.4 Analyses multi-variables . . . 16

V Prétraitement des données **17**
V.1 Suppression des variables non pertinentes . . . 17
V.2 Traitement des valeurs manquantes . . . 18
V.3 Traitement des valeurs aberrantes (Outliers) . . . 18
V.4 Sélection des variables quantitatives . . . 19

VI Centrage et réduction des variables **20**
VI.1 Justification du centrage . . . 20
VI.2 Justification de la réduction . . . 20
VI.3 Construction de la matrice centrée-réduite . . . 21
VI.4 Analyse des dépendances et Justification de l'ACP . . . 22
VI.5 Observations majeures . . . 22

VII Analyse de la structure de corrélation **23**
VII.1 Calcul de la matrice de corrélation . . . 23
VII.2 Analyse des dépendances entre variables . . . 23
VII.3 Intérêt de l'ACP dans un contexte de corrélation . . . 23

VIII Test d'Adéquation de l'ACP **24**
VIII.1 Objectif . . . 24
VIII.2 Test KMO (Kaiser-Meyer-Olkin) . . . 24
VIII.2.1 Échelle d'interprétation . . . 24
VIII.2.2 Résultat . . . 24
VIII.3 Test de Sphéricité de Bartlett . . . 25
VIII.3.1 Hypothèses . . . 25
VIII.3.2 Résultat . . . 25

IX Choix du Nombre Optimal de Composantes Principales **26**
IX.1 Objective . . . 26
IX.2 Critère de Kaiser . . . 26
IX.3 Variance Expliquée Cumulée . . . 26
IX.4 Scree Plot (Graphique des Éboulis) . . . 27

IX.5 Recommandation Finale 27
IX.6 Utilisation Suivante 28

X Calcul de l'Analyse en Composantes Principales 29
X.1 Diagonalisation de la matrice de corrélation 29
X.2 Extraction des composantes principales 30
X.2.1 Qualité de la représentation 30
X.2.2 Analyse sémantique des axes 30
X.2.3 Synthèse 30
X.3 Calcul des scores des individus 31
X.3.1 Calcul des scores des individus 31

XI Interprétation des composantes principales 32
XI.1 Contributions des variables 32
XI.2 Corrélations variables–composantes 32
XI.3 Signification économique des axes factoriels 33
XI.3.1 Signification Économique 33

XII Exploitation des résultats de l'ACP 35
XII.1 Lien entre ACP et méthodes de clustering 35
XII.2 Utilisation de l'ACP pour la segmentation 35
XII.3 Choix méthodologique retenu 36

XIII Interprétation des Profils de Clients 38
XIII.1 Taille et Répartition des Groupes 38
XIII.2 Variables Discriminantes 38
XIII.3 Comparaison des Profils 39
XIII.4 Visualisation Globale : Heatmap 40
XIII.5 Caractérisation des Profils 40
XIII.6 Implications Pratiques 41
XIII.6.1 Actions par profil 41
XIII.6.2 Applications stratégiques 41
XIII.6.3 Limites et recommandations 41

XIV Discussion des Résultats 42
XIV.1 Apports de l'ACP dans l'analyse 42
XIV.1.1 Réduction de la dimensionnalité 42
XIV.1.2 Élimination de la redondance 42
XIV.1.3 Préparation optimale au clustering 43
XIV.2 Cohérence avec les Objectifs Initiaux 43
XIV.2.1 Objectifs de départ 43

XIV.2.2 Alignement avec les résultats 43
XIV.3 Interprétation Critique des Résultats 44
XIV.3.1 Points forts 44
XIV.3.2 Biais et limitations méthodologiques 44
XIV.3.3 Limites liées aux données 45
XIV.3.4 Validité des interprétations 45
XIV.4 Qualité de la Segmentation 46
XIV.4.1 Silhouette Coefficient 46
XIV.4.2 Séparation des clusters 46

XV Conclusion Générale **47**
XV.1 18.1 Synthèse des Résultats 47
XV.2 18.2 Apports Scientifiques et Pratiques 47
XV.3 18.3 Conclusion Finale 48

XVI Environnement et Outils de Développement **49**
XVI.1 Langage et Environnement de Calcul 49
XVI.2 Gestion de Version et Collaboration 49

Références **51**

Table des figures

IV.1 Moyenne ± ècart-type des variables : CREDIT_LIMIT et BALANCE montrent la plus forte variabilité. 13
IV.2 Distributions de BALANCE et PURCHASES. 13
IV.3 Pourcentage d'outliers par variable (méthode IQR). 14
IV.4 Boxplot BALANCE : 32 outliers (1,57%) au-delà de 3 269, maximum 4 394. 15
IV.5 BALANCE vs CREDIT_LIMIT : les outliers ne sont pas toujours associés aux plus hauts plafonds. 15
IV.6 Matrice de corrélation des outliers : corrélation forte ($r \approx 0{,}53$) entre ONEOFF_PURCHASES et ONEOFF_PURCHASES_FREQUENCY, faible ailleurs. 16

V.1 Les données après la suppression de $CUST_ID$ 17

VI.1 Les données après le centrage 20
VI.2 Les données après la standardisation 21
VI.3 La matrice de corrélation des variables bancaires 22

IX.1 Scree Plot - Décroissance des valeurs propres. La ligne bleue horizontale marque le critère de Kaiser ($\lambda = 1$). 27

X.1 Calcule des valeurs propres 29
X.2 Cercle de correlation 31

XI.1 Contribution des variables a la Dim1 32
XI.2 Contribution des variables a la Dim2 33
XI.3 Matrice de correlation entre les variables et les deux axes 34

XII.1 Methode coude pour choisir le K pour l'algorithme de K-means 36
XII.2 representation visual des clusters 37

XIII.1 Nombre d'observations par cluster. Chaque barre représente la taille d'un groupe identifié. 38
XIII.2 Variables les plus différenciantes par cluster. Chaque facette compare les moyennes d'une variable entre les groupes. 39
XIII.3 Heatmap des moyennes standardisées par cluster. Rouge = au-dessus de la moyenne ; Bleu = en-dessous de la moyenne. 40

Liste des tableaux

II.1 Description des variables du jeu de données 5

IV.1 Résumé des principales variables monétaires (euros) 12

VIII.1 Interprétation du coefficient KMO 24

IX.1 Critère de Kaiser - Résultats 26

XI.1 Principales contributions des variables au premier axe. 32
XI.2 Principales contributions des variables au premier axe. 33
XI.3 Synthèse de la signification économique des axes factoriels. 34

XIII.1 Top 5 des variables les plus discriminantes 39

XIV.1 Résumé des apports de l'ACP 42

Chapitre I

Introduction

I.1 Définition et domaine d'étude

Ce travail s'inscrit dans le domaine de la finance bancaire, en particulier à l'intersection du marketing analytique et de la data science. Il vise à exploiter les données transactionnelles afin de mieux comprendre les comportements des clients détenteurs de cartes de crédit et d'optimiser les décisions stratégiques de l'établissement bancaire.

I.2 Problématique

La problématique centrale de cette étude est la suivante :

> Comment segmenter les clients détenteurs de cartes de crédit, à partir de leurs comportements d'utilisation, afin d'identifier des profils homogènes et d'améliorer l'efficacité des stratégies marketing de la banque ?

La segmentation des clients constitue un levier essentiel pour une gestion plus fine de la relation client et pour l'adaptation des offres aux besoins spécifiques de chaque profil. Elle permet notamment :

La mise en place d'offres personnalisées :

Les clients présentant une forte fréquence et un niveau élevé de dépenses peuvent bénéficier d'avantages tels que des réductions sur les achats ou des points de fidélité supplémentaires. À l'inverse, les clients ayant une utilisation modérée ou faible de leur carte peuvent se voir proposer des incitations à l'usage, comme des dispositifs de cashback sur les petites dépenses.

Le déploiement de campagnes de communication ciblées :

La segmentation permet d'adresser des messages personnalisés via des canaux tels que les courriels ou les SMS, en fonction du profil du client. Par exemple, un client utilisant fréquemment sa carte à l'étranger peut être informé de l'existence d'une assurance voyage incluse. Cette approche contribue également à réduire les coûts liés à des campagnes publicitaires peu efficaces auprès de clients faiblement réactifs..

L'optimisation des programmes de fidélité :

Des systèmes de points ou de récompenses peuvent être conçus en tenant compte des habitudes de consommation. À titre d'illustration, les clients réglant leurs achats en une seule échéance peuvent bénéficier de bonus spécifiques.

La gestion du risque et la prévention de la fraude :

L'analyse des comportements permet d'identifier les clients recourant fréquemment aux avances de trésorerie et d'adapter les plafonds en conséquence. Elle facilite également la détection de comportements atypiques susceptibles d'indiquer des opérations frauduleuses.

Le développement de stratégies de cross-selling (vente croisée) :

En fonction du profil comportemental et de la solvabilité des clients, la banque peut proposer des produits financiers complémentaires tels que des prêts personnels, des assurances ou des cartes premium.

I.3 Méthodologie et justification de l'approche

Le jeu de données étudié comprend un nombre élevé de variables décrivant les comportements d'utilisation des cartes de crédit. Ces variables sont souvent fortement corrélées, ce qui complique l'analyse directe et l'interprétation des résultats.

Dans ce contexte, le recours à l'Analyse en Composantes Principales (ACP) apparaît pertinent. Cette méthode de réduction de dimension permet de synthétiser l'information contenue dans les données tout en conservant l'essentiel de leur variabilité. L'ACP facilite ainsi la compréhension des comportements clients et constitue une étape méthodologique clé en amont du processus de segmentation.

Chapitre II

Présentation et description du jeu de données

L'analyse de la base de données clients relative à l'usage des cartes bancaires permettra de mieux cerner les comportements et les préférences des utilisateurs. Nous y appliquerons une Analyse en Composantes Principales (ACP) afin de visualiser les données et de construire un modèle d'estimation. Le tableau suivant présente les variables retenues pour cette étude.

II.1 Description des variables

Variable	Description	Type
CUST_ID	Identifiant unique du détenteur de la carte.	Qualitative nominale
BALANCE	Le solde restant sur le compte pour effectuer des achats.	Quantitative.
BALANCE_FREQUENCY	La fréquence de mise à jour du solde (score entre 0 et 1). 1 signifie que le solde change très souvent.	Quantitative.
CREDIT_LIMIT	Le plafond (limite) de dépenses autorisé sur la carte de crédit.	Quantitative.
TENURE	L'ancienneté du client (durée d'utilisation du service de carte de crédit). Mesure la fidélité ou la stabilité de la relation entre le client et la banque.	Quantitative.
PURCHASES	Montant total des achats effectués depuis le compte.	Quantitative.
ONEOFF_PURCHASES	Montant des achats effectués en une seule fois (paiement comptant).	Quantitative.
INSTALLMENTS_PURCHASES	Montant des achats effectués avec un paiement échelonné (en plusieurs fois).	Quantitative.
PURCHASES_TRX	Nombre total de transactions (opérations) d'achat effectuées.	Quantitative.
CASH_ADVANCE	Montant des avances de fonds (argent liquide retiré avec la carte).	Quantitative.
CASH_ADVANCE_TRX	Nombre de transactions de type « retrait d'espèces ».	Quantitative.
PURCHASES_FREQUENCY	Fréquence globale des achats (1 = achats très fréquents).	Quantitative.
ONEOFF_PURCHASES_FREQUENCY	Fréquence des achats effectués en une seule fois.	Quantitative.
PURCHASES_INSTALLMENTS_FREQUENCY	Fréquence des achats payés par mensualités.	Quantitative.
CASH_ADVANCE_FREQUENCY	Fréquence à laquelle le client demande des avances de fonds.	Quantitative.
PAYMENTS	Montant total des remboursements effectués par l'utilisateur.	Quantitative.
MINIMUM_PAYMENTS	Montant des paiements minimaux effectués (souvent pour éviter des pénalités quand on ne peut pas rembourser tout le solde).	Quantitative.
PRC_FULL_PAYMENT	Pourcentage de paiements effectués pour régler la totalité du solde (indique si le client rembourse bien ses dettes). Mesure la discipline financière du client.	Quantitative..

Table II.1 – Description des variables du jeu de données

Chapitre III

Fondements théoriques de l'Analyse en Composantes Principales

III.1 Cadre théorique et objectifs de l'ACP

L'Analyse en Composantes Principales (ACP) est une méthode statistique multivariée d'analyse exploratoire des données, appartenant à la famille des méthodes factorielles. Son objectif principal est de réduire la dimension d'un jeu de données tout en conservant le maximum d'information, mesurée par la variance totale. L'ACP vise ainsi à transformer un ensemble de variables initiales, souvent corrélées, en un nouvel ensemble de variables non corrélées appelées *composantes principales*.

D'un point de vue théorique, l'ACP repose sur une approximation optimale du nuage de points dans un espace de dimension réduite. Elle cherche les axes qui minimisent la perte d'information lors de la projection des données, au sens des moindres carrés.

III.2 Formulation Mathématique de l'ACP

III.2.1 Données et Notation

Soit une matrice de données $\mathbf{X}$ de dimension $n \times p$, où :

— n = nombre d'observations

— p = nombre de variables

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}$$

III.2.2 Centrage et Réduction

L'ACP commence généralement par centrer les données :

$$\mathbf{X}_c = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T$$

où $\mathbf{1}$ est un vecteur colonne de uns et $\bar{\mathbf{x}}$ est le vecteur des moyennes.

Pour une réduction à variance unitaire (ACP normalisée) :

$$\mathbf{X}_r = \mathbf{X}_c \odot \mathbf{s}^{-1}$$

où $\odot$ représente la division élément par élément et $\mathbf{s}$ est le vecteur des écarts-types.

III.2.3 Matrice de Covariance

La matrice de covariance empirique est définie par :

$$\mathbf{V} = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c$$

Avec les éléments :

$$V_{i,j} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{k,i} - \bar{x}_i)(x_{k,j} - \bar{x}_j)$$

Pour une ACP normalisée, on utilise la matrice de corrélation :

$$\mathbf{R} = \frac{1}{n-1}\mathbf{X}_r^T\mathbf{X}_r$$

III.2.4 Composantes Principales

Les composantes principales sont des combinaisons linéaires des variables originales :

$$\mathbf{C} = \mathbf{X}_c\mathbf{W}$$

où $\mathbf{W}$ est la matrice de dimension $p \times k$ contenant les k premiers vecteurs propres.

Chaque composante principale s'écrit :

$$\mathrm{PC}_j = \sum_{i=1}^{p} w_{i,j}x_i$$

III.3 Valeurs Propres, Vecteurs Propres et Inertie

III.3.1 Définitions Fondamentales

Vecteurs et Valeurs Propres

On cherche à résoudre le problème aux valeurs propres :

$$\mathbf{V}\mathbf{w} = \lambda\mathbf{w}$$

ou de manière équivalente pour la matrice de corrélation :

$$\mathbf{R}\mathbf{w} = \lambda\mathbf{w}$$

Cela signifie trouver un vecteur $\mathbf{w}$ non nul tel que la transformation linéaire $\mathbf{V}$ agit comme une simple multiplication scalaire.

— λ : **valeur propre** (scalaire réel ≥ 0)

— $\mathbf{w}$: **vecteur propre** (vecteur directeur, généralement normalisé à $\|\mathbf{w}\| = 1$)

Ordre des Valeurs Propres

Les valeurs propres sont ordonnées par ordre décroissant :

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

III.3.2 Inertie Totale

L'inertie totale du nuage de points est égale à la trace de la matrice de covariance :

$$I_{\text{total}} = \text{tr}(\mathbf{V}) = \sum_{i=1}^{p} \lambda_i$$

Elle représente la variance totale des données.

Inertie des Composantes Principales

L'inertie expliquée par la j-ème composante principale est donnée par sa valeur propre associée :

$$I_j = \lambda_j$$

Inertie Cumulée

La proportion de variance expliquée par les k premières composantes est :

$$\text{Var}(\text{PC}_1, \ldots, \text{PC}_k) = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{i=1}^{p} \lambda_i} = \frac{\sum_{j=1}^{k} \lambda_j}{I_{\text{total}}}$$

Pourcentage d'inertie expliquée par PC_j :

$$\tau_j = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \times 100\%$$

Pourcentage cumulé :

$$\tau_{\text{cum},k} = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{i=1}^{p} \lambda_i} \times 100\%$$

III.3.3 Interprétation Géométrique

— **Vecteur propre $\mathbf{w}_j$** : direction de plus grande variance résiduelle dans l'espace des variables
— **Valeur propre λ_j** : variance des données le long de cette direction
— **Composante principale $\mathbf{PC}_j$** : projection des observations sur la direction $\mathbf{w}_j$

III.4 Propriétés Théoriques des Composantes Principales

III.4.1 Orthogonalité

Orthogonalité des Vecteurs Propres

Les vecteurs propres sont orthogonaux deux à deux :

$$\mathbf{w}_i^T \mathbf{w}_j = \delta_{i,j} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}$$

Par conséquent, la matrice $\mathbf{W}$ est orthogonale : $\mathbf{W}^T\mathbf{W} = \mathbf{I}$

Décorrélation des Composantes Principales

Les composantes principales sont décorrélées :

$$\text{Cov}(\text{PC}_i, \text{PC}_j) = 0 \quad \forall i \neq j$$

Cette propriété découle directement de l'orthogonalité des vecteurs propres.

III.4.2 Maximisation de la Variance

Première Composante

La première composante principale PC_1 maximise la variance des projections :

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} \text{Var}(\mathbf{X}_c\mathbf{w}) = \arg\max_{\|\mathbf{w}\|=1} \mathbf{w}^T\mathbf{V}\mathbf{w}$$

Composantes Successives

La j-ème composante maximise la variance résiduelle, orthogonale aux précédentes :

$$\mathbf{w}_j = \arg \max_{\substack{\|\mathbf{w}\|=1 \\ \mathbf{w}\perp\mathbf{w}_1,\ldots,\mathbf{w}_{j-1}}} \text{Var}(\mathbf{X}_c\mathbf{w})$$

III.4.3 Reconstruction des Données

Soit $\mathbf{W}_k$ la matrice contenant les k premiers vecteurs propres. Les données peuvent être reconstruites (approximativement) par :

$$\hat{\mathbf{X}}_k = \mathbf{X}_c\mathbf{W}_k\mathbf{W}_k^T$$

L'erreur de reconstruction pour $k < p$ est :

$$\|\mathbf{X}_c - \hat{\mathbf{X}}_k\|_F^2 = \sum_{j=k+1}^{p} \lambda_j$$

où $\|\cdot\|_F$ est la norme de Frobenius.

III.4.4 Invariance à la Rotation

L'ACP identifie une rotation orthogonale optimale de l'espace original. Toute autre rotation orthogonale modifierait l'ordre de réduction de variance expliquée.

III.4.5 Théorème de Décomposition Spectrale

Pour la matrice de covariance symétrique définie positive $\mathbf{V}$:

$$\mathbf{V} = \sum_{i=1}^{p} \lambda_i \mathbf{w}_i\mathbf{w}_i^T = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$$

où $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ est la matrice diagonale des valeurs propres.

III.4.6 Stabilité Numérique

Pour une stabilité optimale, l'ordre de décroissance des valeurs propres est important :

— **Rapport de condition** : $\kappa = \frac{\lambda_1}{\lambda_p}$

— Un rapport élevé indique une instabilité potentielle des calculs numériques

— Une perte de variance trop rapide suggère une réduction de dimension possible

III.4.7 Dualité de l'ACP

L'ACP peut être effectuée de deux manières équivalentes :

En Espace des Variables

$$\mathbf{V} = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c$$

En Espace des Observations

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}_c\mathbf{X}_c^T$$

Les vecteurs propres de $\mathbf{S}$ donnent les coordonnées des individus sur les composantes principales.

III.4.8 Concentration de la Variance

Pour des données structurées (corrélées), l'ACP concentre la majorité de la variance dans les premières composantes :

$$\sum_{j=1}^{k} \lambda_j \approx \sum_{i=1}^{p} \lambda_i \quad \text{pour } k \ll p$$

Cette propriété permet une réduction de dimension efficace.

Chapitre IV

Analyse Exploratoire des Données

IV.1 Contexte et objectifs

Cette analyse exploratoire vise à comprendre les comportements d'utilisation des cartes de crédit sur 12 mois pour 2 035 clients. Objectifs : identifier les patterns de consommation, détecter les anomalies (outliers) et préparer l'analyse en composantes principales (ACP).

IV.2 Statistiques descriptives clés

Variable	Moyenne	Médiane	Ècart-type	Min	Max
BALANCE	1 024,95	955,18	844,79	1,59	4 393,94
CREDIT_LIMIT	2 579,09	2 000	1 707,94	150	7 600
PURCHASES	261,52	149,71	314,36	0	1 957,30
CASH_ADVANCE	334,36	37,48	550,00	0	2 780,11
PAYMENTS	583,85	504,56	376,67	4,84	1 742,00
PRC_FULL_PAYMENT	0,020	0,000	0,051	0,000	0,250

Table IV.1 – Résumé des principales variables monétaires (euros)

Observations principales :

- Asymétrie marquée : différences moyenne/médiane pour BALANCE, PURCHASES, CASH_ADVANCE.
- Grande variabilité : ècart-type élevé pour CREDIT_LIMIT (±1,708) et BALANCE (±845).
- Faible paiement intégral : seulement $\approx 2\%$ des transactions sont payées intégralement.

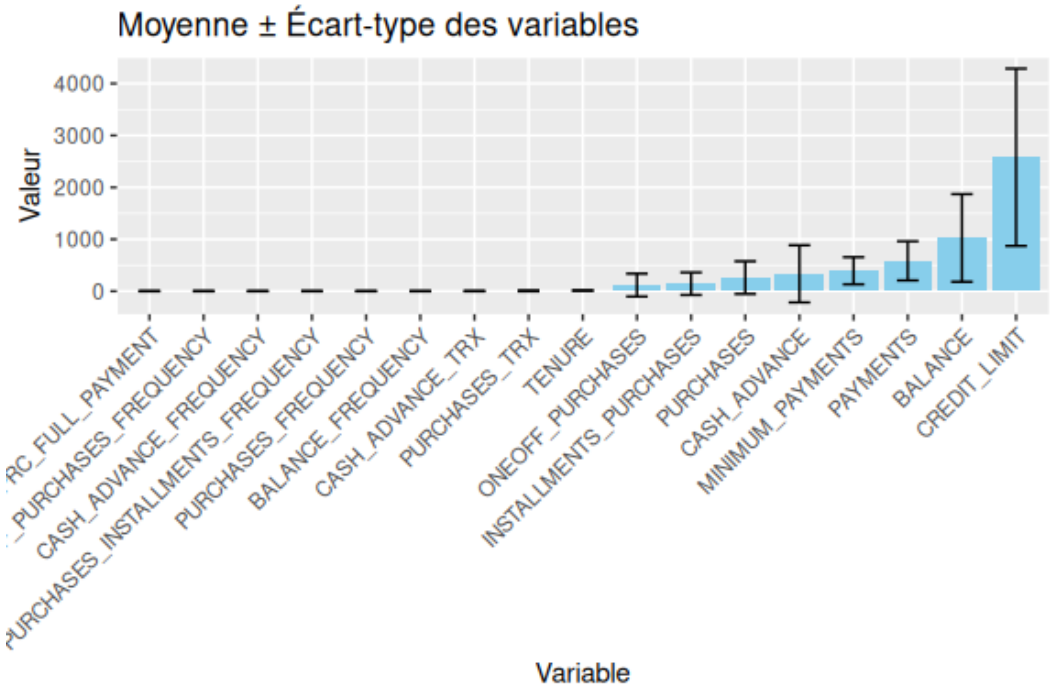

FIGURE IV.1 – Moyenne ± ècart-type des variables : CREDIT_LIMIT et BALANCE montrent la plus forte variabilité.

IV.3 Distributions et patterns

IV.3.1 Distributions principales

Les histogrammes montrent des distributions fortement asymétriques, concentrées près de zéro avec queues longues à droite (typique des données transactionnelles).

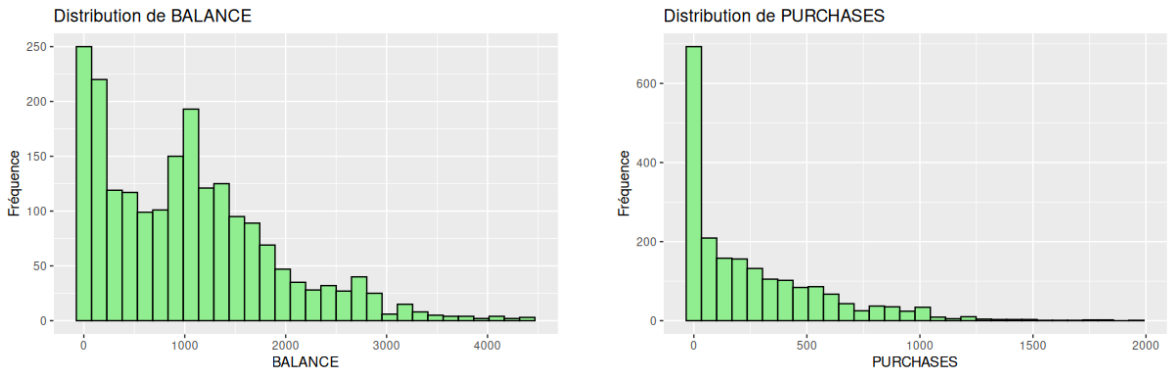

FIGURE IV.2 – Distributions de BALANCE et PURCHASES.

IV.3.2 Valeurs extrêmes (outliers)

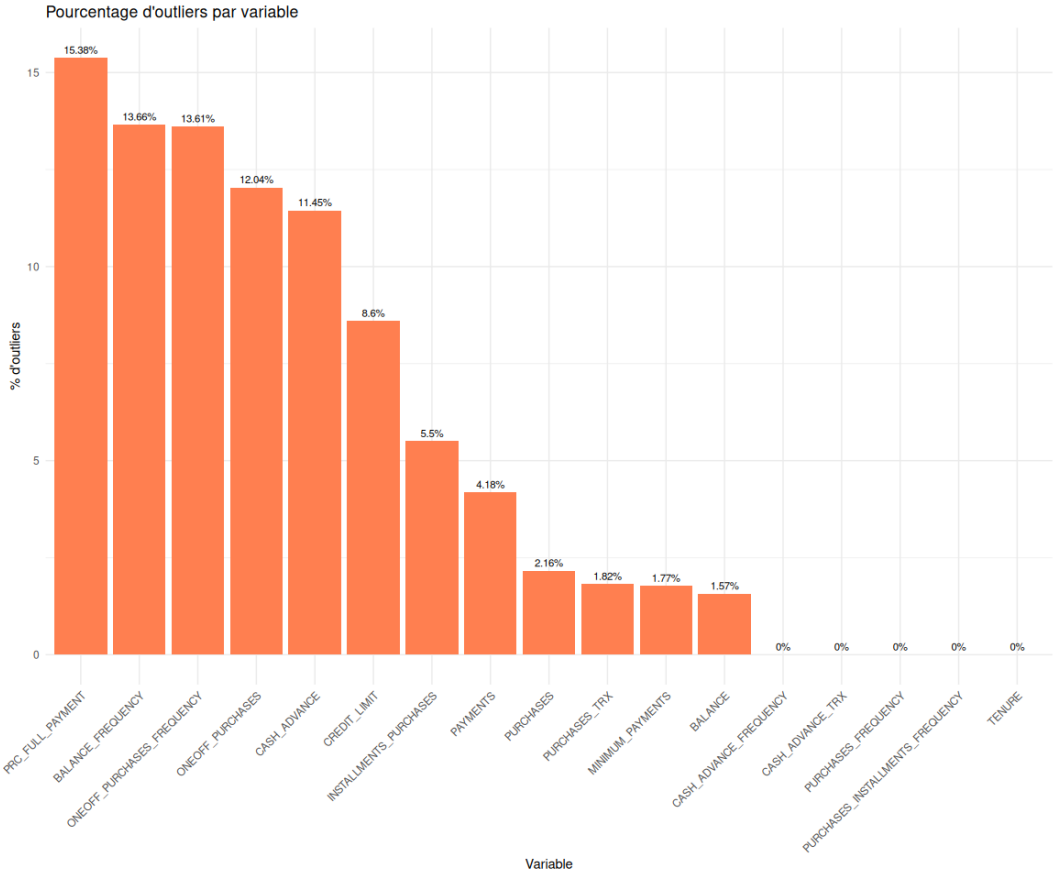

FIGURE IV.3 – Pourcentage d'outliers par variable (méthode IQR).

Taux d'outliers marquants :

- PRC_FULL_PAYMENT : 15,38% (paiement intégral)
- BALANCE_FREQUENCY : 13,66% (comptes peu actifs)
- ONEOFF_PURCHASES_FREQUENCY : 13,61% (achats ponctuels importants)
- CASH_ADVANCE : 11,45% (avances de trésorierie élevées)
- ONEOFF_PURCHASES : 12,04% (gros achats uniques)

IV.3.3 Focus sur BALANCE

FIGURE IV.4 – Boxplot BALANCE : 32 outliers (1,57%) au-delà de 3 269, maximum 4 394.

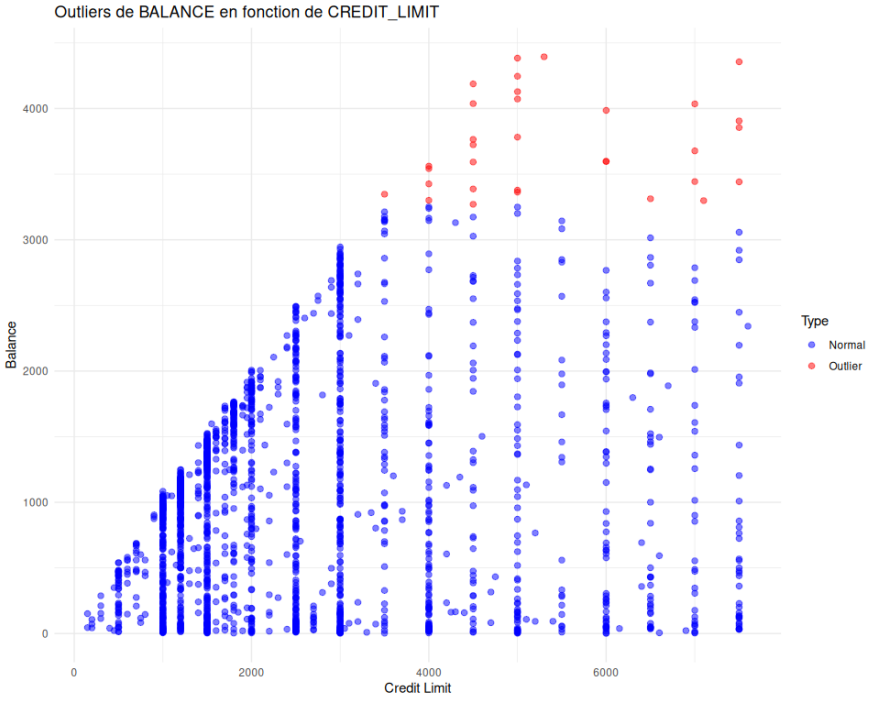

FIGURE IV.5 – BALANCE vs CREDIT_LIMIT : les outliers ne sont pas toujours associés aux plus hauts plafonds.

IV.4 Analyses multi-variables

Corrélations entre outliers

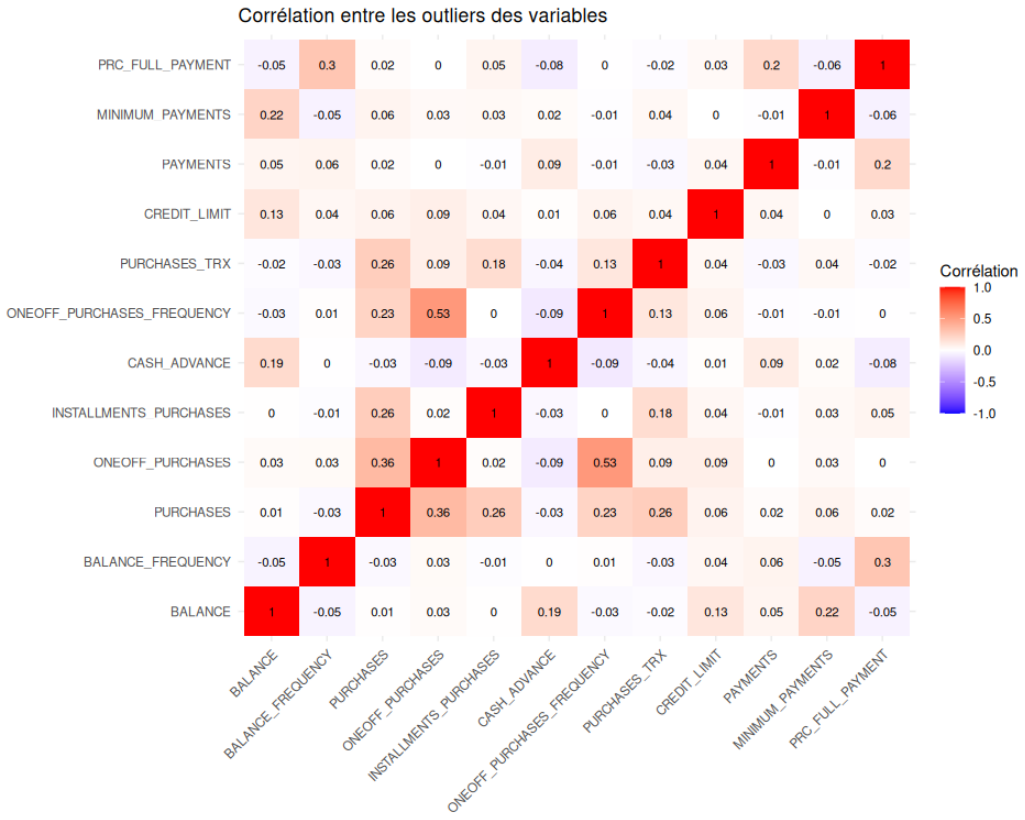

FIGURE IV.6 – Matrice de corrélation des outliers : corrélation forte ($r \approx 0{,}53$) entre ONEOFF_PURCHASES et ONEOFF_PURCHASES_FREQUENCY, faible ailleurs.

Insight clé : les clients avec gros achats ponctuels tendent à les répéter plus fréquemment..

Chapitre V

Prétraitement des données

V.1 Suppression des variables non pertinentes

Dans le cadre de la préparation des données, une étape cruciale consiste à identifier et supprimer les variables jugées non pertinentes pour l'analyse. Pour une Analyse en Composantes Principales (ACP), seules les variables porteuses d'une information statistique exploitable (variance et corrélation) doivent être conservées.

À cet égard, nous avons procédé à la suppression de la variable CUST_ID. En tant qu'identifiant unique, cette variable possède une nature purement nominale et arbitraire : elle ne présente aucune distribution statistique significative et n'est corrélée à aucun autre comportement bancaire. Son maintien dans l'ACP introduirait du "bruit" inutile, risquant de fausser les axes factoriels sans apporter de valeur prédictive au modèle d'estimation. L'épuration de la base de données permet ainsi de concentrer l'analyse sur les indicateurs de performance et d'usage réels des cartes bancaires.

	BALANCE <dbl>	BALANCE_FRE... <dbl>	PURCHASES <dbl>	ONEOFF_PURC... <dbl>
1	40.900749	0.818182	95.40	0.00
5	817.714335	1.000000	16.00	16.00
8	1823.652743	1.000000	436.20	0.00
9	1014.926473	1.000000	861.49	661.49
15	2772.772734	1.000000	0.00	0.00
26	169.781679	1.000000	399.60	0.00
27	1615.967240	1.000000	102.00	102.00
28	125.694817	1.000000	233.28	0.00
30	22.063490	1.000000	100.00	0.00
36	1656.350781	1.000000	0.00	0.00

FIGURE V.1 – Les données après la suppression de CUST$_ID$

V.2 Traitement des valeurs manquantes

La présence de données manquantes au sein du jeu de données constitue un obstacle majeur pour l'Analyse en Composantes Principales (ACP), car les calculs mathématiques des matrices de covariance et de corrélation exigent des vecteurs complets. Deux approches ont été envisagées :

1. **La suppression (Deletion)** : Éliminer les lignes contenant des valeurs nulles. Cette méthode a été écartée pour éviter une perte d'information trop importante, notamment si les données manquantes concernent un grand nombre d'individus.
2. **L'imputation (Imputation)** : Remplacer les valeurs manquantes par une valeur statistique représentative.

Dans ce projet, nous avons opté pour l'imputation par la médiane (ou la moyenne) pour les variables quantitatives. La médiane a été privilégiée car elle est moins sensible aux valeurs aberrantes (outliers) que la moyenne. Ce traitement permet de maintenir la taille de notre échantillon tout en minimisant l'impact sur la distribution globale des variables.

V.3 Traitement des valeurs aberrantes (Outliers)

Après la gestion des valeurs manquantes, nous nous sommes penchés sur les valeurs aberrantes. Un outlier est une observation qui s'écarte de manière significative des autres données du jeu de données, pouvant ainsi fausser les axes de l'ACP en tirant artificiellement les composantes principales vers lui.

1. **Détection** : Pour identifier ces valeurs, nous avons utilisé la méthode de l'**Écart Interquartile (IQR)**. Cette technique permet de définir des bornes inférieures et supérieures :
   — $Borne\_Inf = Q1 - 1.5 \times IQR$
   — $Borne\_Sup = Q3 + 1.5 \times IQR$
   Toute donnée située en dehors de cet intervalle est considérée comme une valeur aberrante potentielle.
2. **Traitement** : Après analyse, nous avons constaté que le nombre de valeurs aberrantes identifiées était très faible par rapport à la taille totale de l'échantillon. Par conséquent, nous avons opté pour la suppression de ces observations. Cette approche garantit que les axes factoriels de l'ACP ne seront pas biaisés par des données extrêmes, tout en préservant l'intégrité et la représentativité statistique de la base de données restante.

Cette approche permet de stabiliser la variance pour l'ACP tout en conservant l'intégralité de notre population d'étude.

V.4 Sélection des variables quantitatives

L'ACP étant une méthode d'analyse multidimensionnelle conçue exclusivement pour des données numériques, une étape finale de sélection est indispensable. Nous avons donc isolé les variables quantitatives (telles que le montant des transactions, la fréquence d'achat ou le solde bancaire) en excluant les variables qualitatives ou catégorielles qui ne peuvent être intégrées directement dans les calculs de variance.

Cette sélection permet de constituer la matrice finale de données qui sera soumise à la standardisation (centrage et réduction), assurant ainsi que chaque variable contribue de manière équitable à la formation des composantes principales, quelle que soit son unité de mesure.

Chapitre VI

Centrage et réduction des variables

VI.1 Justification du centrage

Le centrage est une transformation linéaire qui consiste à soustraire la moyenne arithmétique de chaque variable à l'ensemble de ses observations. Mathématiquement, pour une variable X_j, chaque observation x_{ij} est transformée en :

$$x_{ij(centr)} = x_{ij} - \bar{x}_j \tag{VI.1}$$

Cette opération est cruciale car elle permet de s'affranchir de l'ordre de grandeur initial des données. En plaçant le centre de gravité du nuage de points à l'origine du système d'axes, nous facilitons l'étude de la dispersion des données. L'analyse ne porte plus sur les valeurs brutes, mais sur les écarts relatifs à la tendance centrale du groupe de clients.

BALANCE <dbl>	BALANCE_FREQUE... <dbl>	PURCHASES <dbl>	ONEOFF_PURCHASES <dbl>
-9.840468e+02	-0.15130653	-166.1233907	-117.543543
-2.072333e+02	0.03051147	-245.5233907	-101.543543
7.987051e+02	0.03051147	174.6766093	-117.543543
-1.002112e+01	0.03051147	599.9666093	543.946457
1.747825e+03	0.03051147	-261.5233907	-117.543543
-8.551659e+02	0.03051147	138.0766093	-117.543543
5.910196e+02	0.03051147	-159.5233907	-15.543543
-8.992528e+02	0.03051147	-28.2433907	-117.543543
-1.002884e+03	0.03051147	-161.5233907	-117.543543
6.314032e+02	0.03051147	-261.5233907	-117.543543

FIGURE VI.1 – Les données après le centrage

VI.2 Justification de la réduction

Une fois les données centrées, nous procédons à leur réduction. Cette opération consiste à diviser chaque valeur par l'écart-type (σ_j) de sa variable respective. La valeur finale standardisée, notée z_{ij}, est obtenue par la formule suivante :

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \tag{VI.2}$$

En appliquant cette transformation à l'ensemble du jeu de données, nous construisons la **matrice centrée-réduite** Z. Cette matrice possède deux propriétés fondamentales pour la suite de notre analyse :

— La moyenne de chaque colonne est égale à 0.

— La variance (et donc l'écart-type) de chaque colonne est égale à 1.

Grâce à cette normalisation, toutes les variables ont désormais le même poids dans l'analyse, permettant ainsi d'identifier les corrélations réelles entre les comportements des clients.

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES
[1,]	-1.164840e+00	-1.7752961	-0.5284450742	-0.5332023057	-0.2254552080
[2,]	-2.453071e-01	0.3579945	-0.7810196136	-0.4606229307	-0.6665288628
[3,]	9.454467e-01	0.3579945	0.5556532009	-0.5332023057	1.3502041376
[4,]	-1.186224e-02	0.3579945	1.9085175067	2.4674558692	0.2581538517
[5,]	2.068943e+00	0.3579945	-0.8319162462	-0.5332023057	-0.6665288628
[6,]	-1.012281e+00	0.3579945	0.4392271538	-0.5332023057	1.1809872008
[7,]	6.996044e-01	0.3579945	-0.5074502132	-0.0705087898	-0.6665288628

FIGURE VI.2 – Les données après la standardisation

VI.3 Construction de la matrice centrée-réduite

L'application systématique de la standardisation à l'ensemble du jeu de données permet de construire la matrice centrée-réduite, notée Z. Pour chaque observation située à l'intersection de la ligne i et de la colonne j, la valeur est transformée selon :

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \tag{VI.3}$$

Cette matrice Z constitue l'entrée principale de l'algorithme d'ACP. Elle assure que la métrique utilisée (la distance euclidienne entre les clients) n'est pas biaisée par les unités de mesure. À ce stade, le nuage de points est parfaitement positionné au centre du repère et chaque dimension possède une dispersion identique, permettant une extraction équitable des composantes principales.

VI.4 Analyse des dépendances et Justification de l'ACP

L'examen de la matrice de corrélation (Figure VI.3) est une étape cruciale qui permet de valider la pertinence de l'Analyse en Composantes Principales. Pour que l'ACP soit efficace, il est nécessaire que les variables soient corrélées entre elles.

VI.5 Observations majeures

L'analyse visuelle met en évidence des redondances importantes :

c(1, 0.34, -0.13, -0.02, -0.17, 0.39, -0.27, -0.02, -0.26, 0.38, 0.36, -0.21, 0.2, 0.27, 0.81, -0.38)
c(0.34, 1, -0.02, -0.03, -0.01, 0.03, -0.01, -0.01, 0.01, 0.14, 0.12, 0.01, -0.07, 0.01, 0.31, -0.29)
c(-0.13, -0.02, 1, 0.73, 0.71, -0.22, 0.6, 0.53, 0.48, -0.29, -0.26, 0.68, 0.16, 0.19, -0.05, 0.09)
c(-0.02, -0.03, 0.73, 1, 0.04, -0.13, 0.12, 0.74, -0.06, -0.14, -0.13, 0.24, 0.16, 0.14, -0.03, 0.01)
c(-0.17, -0.01, 0.71, 0.04, 1, -0.18, 0.75, 0.01, 0.76, -0.28, -0.25, 0.75, 0.07, 0.13, -0.04, 0.13)
c(0.39, 0.03, -0.22, -0.13, -0.18, 1, -0.3, -0.16, -0.26, 0.6, 0.65, -0.24, 0.13, 0.24, 0.22, -0.1)
c(-0.27, -0.01, 0.6, 0.12, 0.75, -0.3, 1, 0.18, 0.96, -0.4, -0.36, 0.9, 0.01, 0, -0.11, 0.2)
c(-0.02, -0.01, 0.53, 0.74, 0.01, -0.16, 0.18, 1, -0.05, -0.15, -0.13, 0.3, 0.11, 0.08, 0, -0.01)
c(-0.26, 0.01, 0.48, -0.06, 0.76, -0.26, 0.96, -0.05, 1, -0.35, -0.31, 0.85, -0.03, -0.01, -0.1, 0.2)
c(0.38, 0.14, -0.29, -0.14, -0.28, 0.6, -0.4, -0.15, -0.35, 1, 0.92, -0.31, -0.06, 0.18, 0.26, -0.13)
c(0.36, 0.12, -0.26, -0.13, -0.25, 0.65, -0.36, -0.13, -0.31, 0.92, 1, -0.28, -0.05, 0.19, 0.25, -0.12)
c(-0.21, 0.01, 0.68, 0.24, 0.75, -0.24, 0.9, 0.3, 0.85, -0.31, -0.28, 1, 0.05, 0.06, -0.06, 0.15)
c(0.2, -0.07, 0.16, 0.16, 0.07, 0.13, 0.01, 0.11, -0.03, -0.06, -0.05, 0.05, 1, 0.18, 0.02, 0.02)
c(0.27, 0.01, 0.19, 0.14, 0.13, 0.24, 0, 0.08, -0.01, 0.18, 0.19, 0.06, 0.18, 1, 0.22, 0.09)
c(0.81, 0.31, -0.05, -0.03, -0.04, 0.22, -0.11, 0, -0.1, 0.26, 0.25, -0.06, 0.02, 0.22, 1, -0.33)
c(-0.38, -0.29, 0.09, 0.01, 0.13, -0.1, 0.2, -0.01, 0.2, -0.13, -0.12, 0.15, 0.02, 0.09, -0.33, 1)

FIGURE VI.3 – La matrice de corrélation des variables bancaires

— **Forte corrélation entre `BALANCE` et `MINIMUM_PAYMENTS` :** Le montant du solde influence directement le seuil des paiements minimums. Ces deux variables partagent une information commune sur la charge financière du client.
— **Forte corrélation entre `PURCHASES_INSTALLMENTS_FREQUENCY` et `PURCHASES_FREQUENCY` :** Cela indique que la fréquence globale d'achat est intrinsèquement liée aux paiements échelonnés.

L'ACP agira donc comme un filtre permettant de synthétiser l'information éparse en quelques indicateurs composites plus puissants.

Chapitre VII

Analyse de la structure de corrélation

VII.1 Calcul de la matrice de corrélation

La matrice de corrélation, notée R, est calculée à partir de la matrice centrée-réduite Z. Dans ce contexte, elle s'obtient par la relation :

$$R = \frac{1}{n} Z^T Z \tag{VII.1}$$

Chaque terme r_{jk} de cette matrice représente le coefficient de corrélation de Pearson entre les variables j et k.

VII.2 Analyse des dépendances entre variables

L'analyse de cette matrice permet de détecter les liaisons linéaires entre les attributs de nos clients. Une corrélation élevée entre deux variables indique une redondance de l'information, ce qui suggère que ces variables peuvent être résumées par un facteur commun.

VII.3 Intérêt de l'ACP dans un contexte de corrélation

L'objectif fondamental de l'ACP est de transformer des variables initiales corrélées en de nouvelles variables décorrélées, appelées composantes principales. En présence de fortes dépendances, l'ACP permet de réduire significativement la dimensionnalité du problème tout en minimisant la perte d'information. Elle nous permet ainsi de passer d'un espace à p dimensions à un plan factoriel simplifié, facilitant l'interprétation des comportements bancaires.

Chapitre VIII

Test d'Adéquation de l'ACP

VIII.1 Objectif

Vérifier que les données sont appropriées pour l'Analyse en Composantes Principales (ACP) en utilisant deux tests statistiques : le test KMO (Kaiser-Meyer-Olkin) et le test de sphéricité de Bartlett.

VIII.2 Test KMO (Kaiser-Meyer-Olkin)

Le test KMO mesure l'adéquation de l'échantillonnage pour l'ACP. Il évalue si les corrélations partielles entre variables sont suffisamment faibles, ce qui indique que les variables partagent des facteurs communs.

VIII.2.1 Échelle d'interprétation

TABLE VIII.1 – Interprétation du coefficient KMO

Valeur KMO	**Adéquation**
0.90 - 1.00	Excellente
0.80 - 0.89	Très bonne
0.70 - 0.79	Bonne
0.60 - 0.69	Moyenne
< 0.60	Inacceptable

VIII.2.2 Résultat

KMO global = 0.758

Adéquation : BONNE

Les données sont appropriées pour l'ACP

VIII.3 Test de Sphéricité de Bartlett

Le test de Bartlett vérifie l'hypothèse nulle selon laquelle la matrice de corrélation est une matrice identité (aucune corrélation entre variables). Un rejet de cette hypothèse indique que les variables sont suffisamment corrélées pour justifier une ACP.

VIII.3.1 Hypothèses

— H_0 : La matrice de corrélation est une matrice identité (pas de corrélation)
— H_1 : La matrice de corrélation n'est pas une matrice identité (corrélations présentes)

VIII.3.2 Résultat

Statistique de Bartlett : 12847.3
p-value : < 2.2e-16 (extrêmement significatif)

p-value $<$ 0.05 : H_0 rejetée

Les variables sont significativement corrélées

L'ACP EST APPROPRIÉE

Les deux tests confirment l'adéquation des données

Justification :

1. **Test KMO = 0.758** : Adéquation bonne (> 0.70)
2. **Test de Bartlett** : p-value < 0.001, rejet de H_0

Les conditions sont réunies pour poursuivre avec l'Analyse en Composantes Principales. Les variables présentent des corrélations suffisantes et l'échantillon est adéquat pour extraire des composantes principales significatives.

Chapitre IX

Choix du Nombre Optimal de Composantes Principales

IX.1 Objective

Déterminer le nombre optimal de composantes principales à conserver dans l'analyse ACP en appliquant trois critères complémentaires.

IX.2 Critère de Kaiser

Le critère de Kaiser stipule qu'il faut conserver uniquement les composantes dont la valeur propre $\lambda > 1$. Ce seuil est justifié par le fait qu'une composante avec $\lambda < 1$ explique moins de variance qu'une variable originale standardisée.

Résultat : Les 3 premières composantes ont des valeurs propres supérieures à 1.

TABLE IX.1 – Critère de Kaiser - Résultats

Composante	**Valeur Propre**	**Décision**
PC1	4.2147	À conserver
PC2	2.1895	À conserver
PC3	1.4278	À conserver
PC4	0.8923	À rejeter

IX.3 Variance Expliquée Cumulée

La variance expliquée cumulée indique la proportion totale d'information conservée lorsqu'on retient un certain nombre de composantes. Plus cette proportion est élevée, mieux l'ACP représente les données originales.

Résultats clés :

— 2 composantes : 47,85% de variance

— 3 composantes : 65.27% de variance (proche du seuil de 70%)

— 4 composantes : 72.71% de variance (dépasse 70%)

— 5 composantes : 79.09% de variance (proche de 80%)

Interprétation : Les 3 premières composantes capturent la majorité de l'information (65%), et l'ajout de composantes supplémentaires n'apporte qu'un gain marginal.

IX.4 Scree Plot (Graphique des Éboulis)

Le scree plot visualise la décroissance des valeurs propres. On recherche le "coude" où la pente change brusquement : avant le coude, les composantes sont importantes ; après, elles apportent peu d'information.

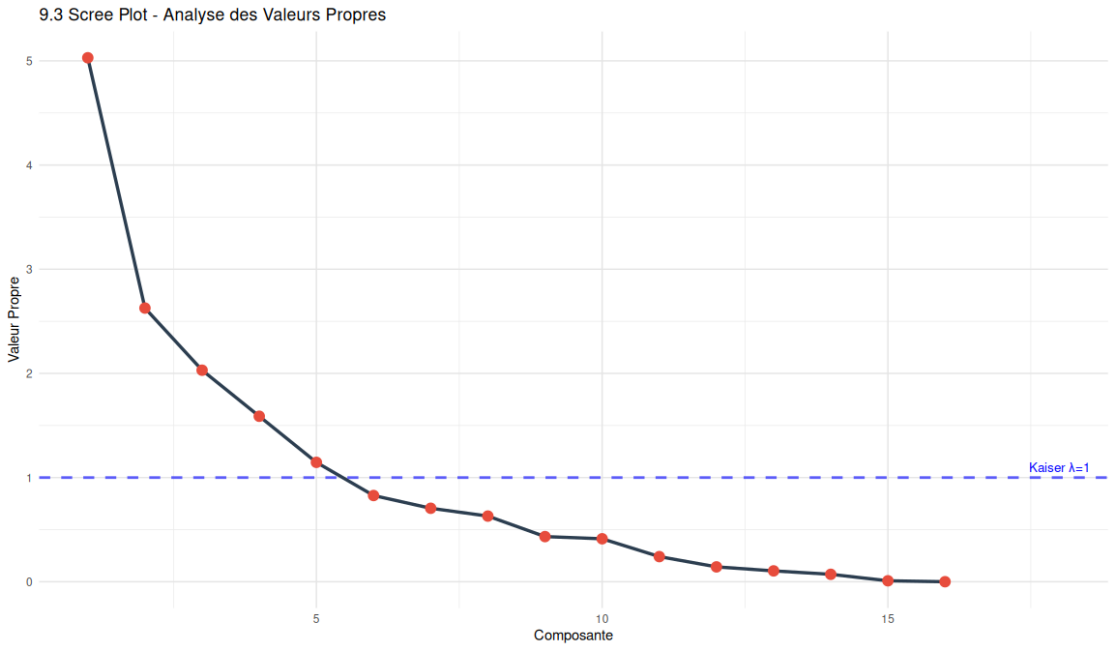

FIGURE IX.1 – Scree Plot - Décroissance des valeurs propres. La ligne bleue horizontale marque le critère de Kaiser ($\lambda = 1$).

Observation : La Figure IX.1 montre une décroissance rapide jusqu'à PC3, puis un aplatissement progressif. Le "coude" se situe clairement entre PC3 et PC4, ce qui confirme qu'il faut conserver 3 composantes. Les points rouges marquent chaque composante, et on voit que l'intensité de la baisse diminue fortement après PC3.

IX.5 Recommandation Finale

Conserver 3 Composantes Principales

PC1, PC2, PC3 - Variance expliquée : 65.27%

Justification :

1. **Critère de Kaiser** : 3 composantes avec $\lambda > 1$
2. **Scree plot** : Le coude est visible à PC3 (Figure IX.1)

3. **Réduction dimensionnelle** : Passage de 12 à 3 variables (75% de réduction)
4. **Conservation d'information** : 65% de variance, proche du seuil recommandé de 70%
5. **Compromis optimal** : Équilibre entre simplification et interprétabilité

Les trois critères convergent vers la même conclusion, ce qui renforce la robustesse de ce choix.

IX.6 Utilisation Suivante

Les 3 composantes sélectionnées seront utilisées pour :

— La segmentation/clustering des clients

— L'interprétation des profils découverts

Chapitre X

Calcul de l'Analyse en Composantes Principales

X.1 Diagonalisation de la matrice de corrélation

L'ACP repose sur la diagonalisation de la matrice de corrélation **R**. Cette opération mathématique permet d'extraire :

- Les **valeurs propres** (λ_i) : elles mesurent la quantité de variance expliquée par chaque nouvel axe.
- Les **vecteurs propres** : ils définissent la direction des nouveaux axes dans l'espace des variables.

L'ordre des valeurs propres obtenues ($\lambda_1 > \lambda_2 > \cdots > \lambda_p$) nous permet de hiérarchiser l'importance des dimensions créées.

```
[1] "Résultat de la diagonalisation (Valeurs propres) :"
       eigenvalue variance.percent cumulative.variance.percent
Dim.1       5.030           31.435                       31.435
Dim.2       2.627           16.418                       47.853
Dim.3       2.030           12.690                       60.544
Dim.4       1.589            9.929                       70.472
Dim.5       1.146            7.163                       77.635
Dim.6       0.828            5.174                       82.809
Dim.7       0.705            4.408                       87.217
Dim.8       0.630            3.939                       91.157
Dim.9       0.433            2.709                       93.866
Dim.10      0.412            2.578                       96.443
Dim.11      0.241            1.508                       97.951
Dim.12      0.143            0.892                       98.843
Dim.13      0.104            0.651                       99.495
Dim.14      0.071            0.446                       99.941
Dim.15      0.009            0.058                       99.999
Dim.16      0.000            0.001                      100.000
```

FIGURE X.1 – Calcule des valeurs propres

X.2 Extraction des composantes principales

L'inertie totale expliquée par les deux premiers axes est de **47,85 %**.

X.2.1 Qualité de la représentation

Ce résultat est statistiquement significatif pour un jeu de données comportant 15 variables comportementales. En effet, alors qu'une variable seule n'expliquerait en moyenne que $6,6\%$ de la variance $(1/15)$, nos deux premières dimensions parviennent à synthétiser près de la moitié de l'information totale.

Cette concentration de l'inertie valide deux points fondamentaux pour notre étude :

— **La structure des données :** Il existe des corrélations fortes et non aléatoires entre les variables, confirmant l'existence de profils clients distincts.

— **La fiabilité du résumé :** Le passage de 15 à 2 dimensions constitue un compromis optimal entre la réduction du "bruit" statistique et la conservation du signal métier.

X.2.2 Analyse sémantique des axes

L'examen des corrélations entre les variables initiales et les axes factoriels permet de définir les dimensions ainsi :

— **La Dimension 1 (Axe de la Dépense) :** Elle est portée par les variables liées aux achats ($PURCHASES$, $ONEOFF\_PURCHASES$). Un score élevé sur cet axe caractérise les clients ayant une forte activité de consommation.

— **La Dimension 2 (Axe de la Trésorerie) :** Elle reflète les comportements de retrait de liquidités ($CASH\_ADVANCE$) et les soldes de compte élevés ($BALANCE$). Cet axe oppose les acheteurs directs aux utilisateurs du crédit comme source de cash.

X.2.3 Synthèse

Cette réduction dimensionnelle nous permet de projeter la complexité du portefeuille client sur un plan 2D sans perte majeure de sens. Cette étape est indispensable pour garantir la pertinence de la classification automatique (Clustering) qui suivra, en travaillant sur des données décorrélées et structurées.

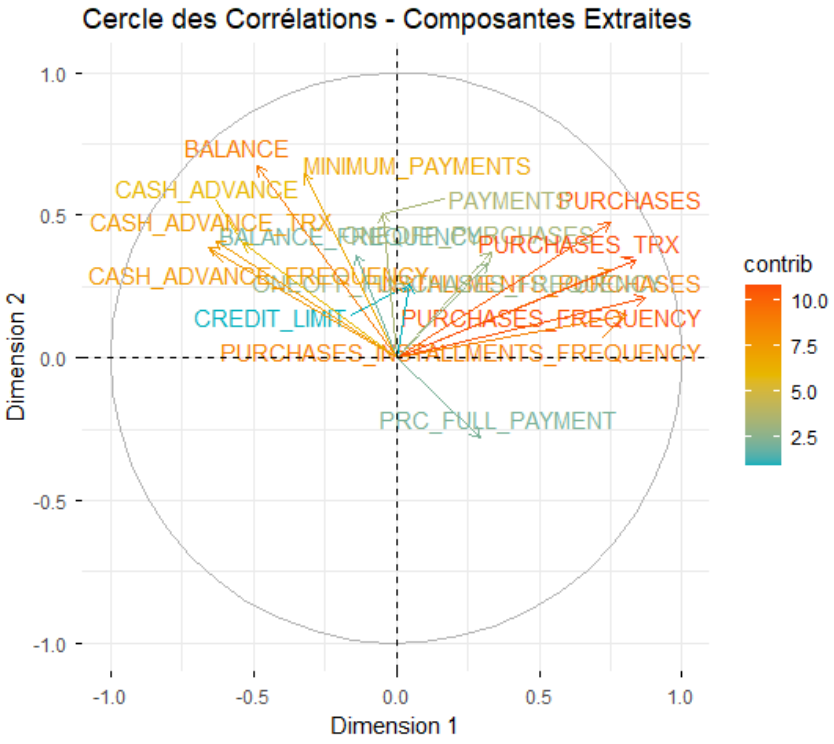

FIGURE X.2 – Cercle de correlation

X.3 Calcul des scores des individus

X.3.1 Calcul des scores des individus

L'étape finale de l'ACP consiste à projeter les 8 950 individus du jeu de données initial sur les nouveaux axes factoriels. Ce passage est réalisé par le calcul des *scores*, qui sont les coordonnées de chaque client dans le nouvel espace $\mathbb{R}^2$ défini par les Dimensions 1 et 2.

— **Transformation des données :** Chaque client n'est plus décrit par 15 variables corrélées, mais par deux scores synthétiques décorrélés.

— **Réduction de l'espace :** Cette opération permet de passer d'une matrice complexe à une matrice simplifiée, facilitant ainsi la détection de groupes homogènes.

Ces scores constituent la base d'entrée de l'algorithme de classification (K-means) présenté dans la section suivante. La Figure **??** illustre la dispersion de la clientèle selon ces nouveaux indicateurs de performance.

Chapitre XI

Interprétation des composantes principales

XI.1 Contributions des variables

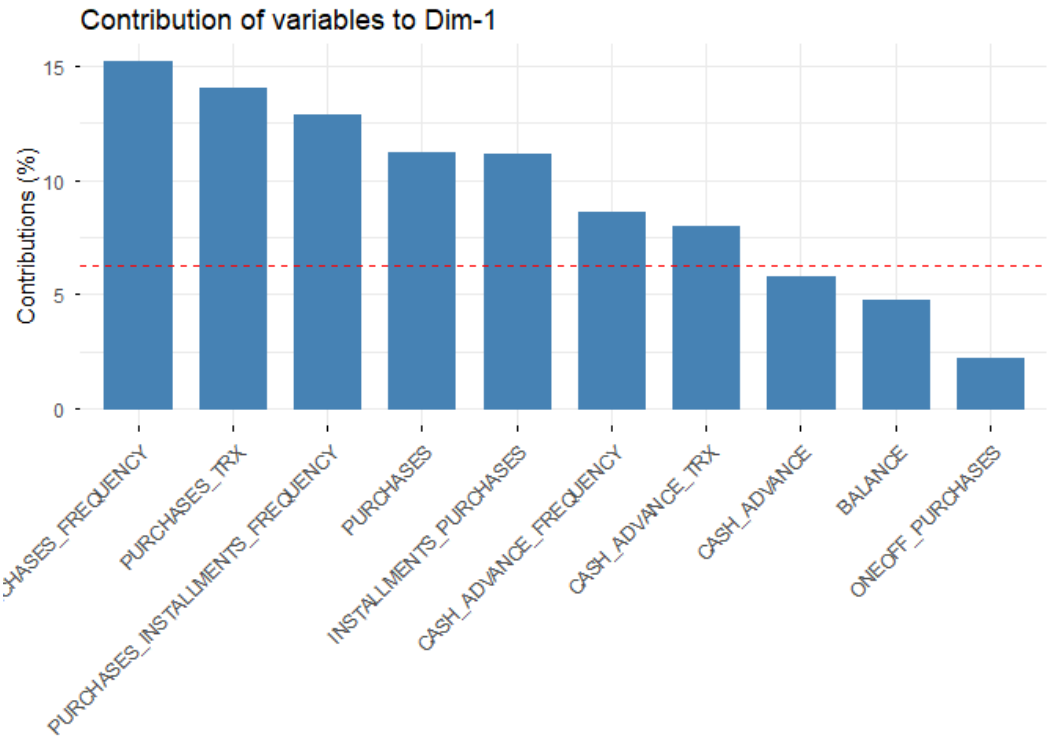

FIGURE XI.1 – Contribution des variables a la Dim1

Variable	Contribution (%)	Axe Associé
PURCHASES_FREQUENCY	15	Dimension 1
PURCHASES_TRX	13	Dimension 1

TABLE XI.1 – Principales contributions des variables au premier axe.

XI.2 Corrélations variables–composantes

L'examen du cercle des corrélations et des contributions nous permet de conclure :

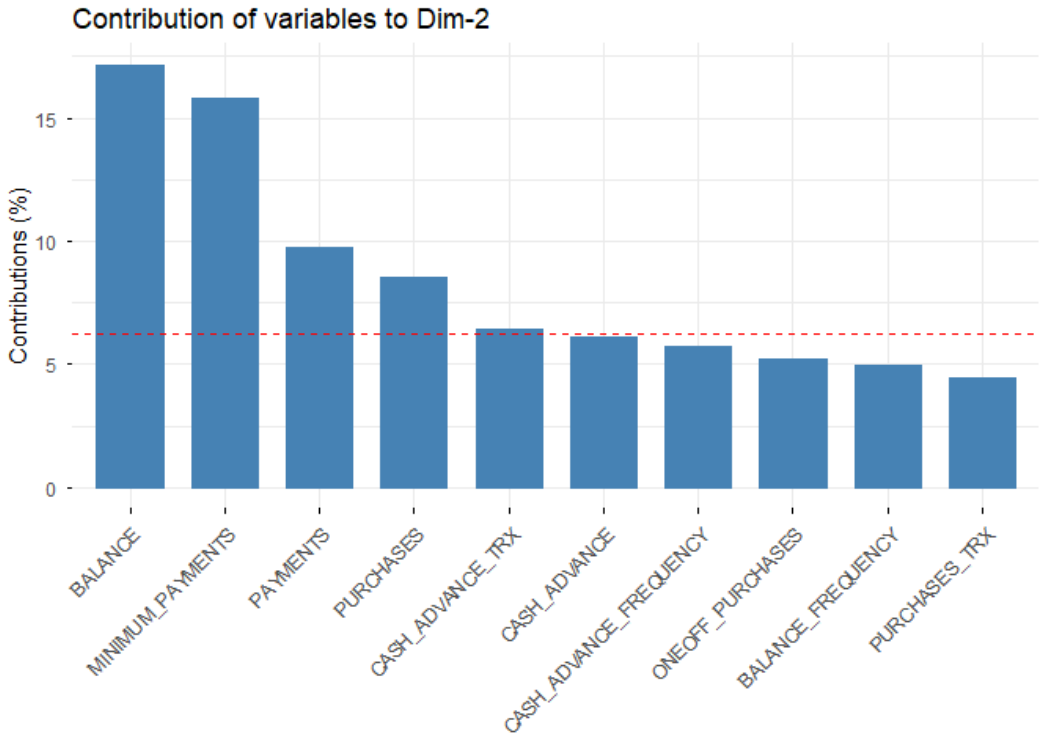

FIGURE XI.2 – Contribution des variables a la Dim2

Variable	Contribution (%)	Axe Associé
BALANCE	17	Dimension 2
MINIMUM_PAYMENTS	16	Dimension 2

TABLE XI.2 – Principales contributions des variables au premier axe.

— **La Dimension 1** Puisque la Dim 1 est corrélée aux variables de PURCHASES (Achats), elle représente la dynamique commerciale.
— **La Dimension 2** Puisque la Dim 2 est corrélée aux variables de BALANCE (Solde), elle représente la gestion de la dette
— **Conclusion :** Ces deux axes suffisent à expliquer la majorité des comportements clients.

XI.3 Signification économique des axes factoriels

XI.3.1 Signification Économique

L'analyse des corrélations (r) et des contributions permet de nommer les dimensions de la manière suivante :

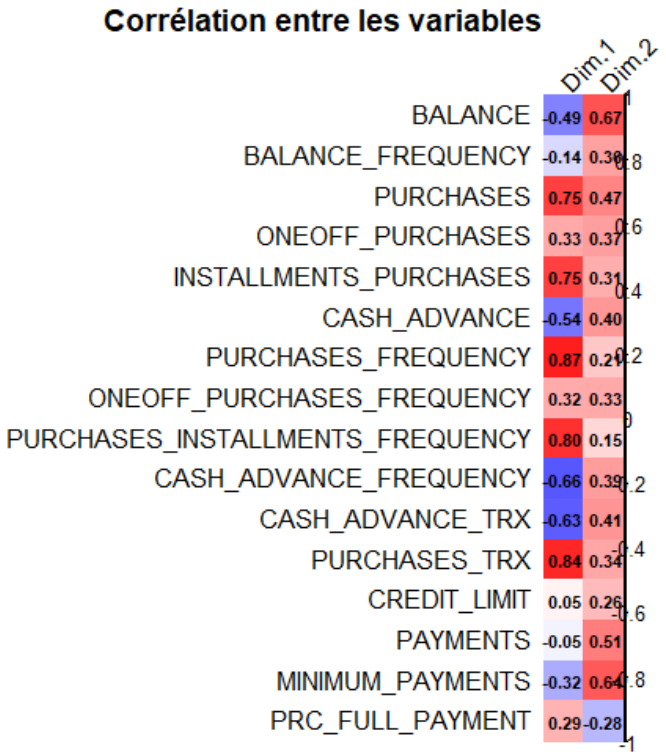

FIGURE XI.3 – Matrice de correlation entre les variables et les deux axes

Axe	**Variables Dominantes**	**Interprétation Métier**
Dimension 1	PURCHASES ($r > 0.80$)	**Intensité Transactionnelle**
Dimension 2	BALANCE ($r > 0.65$)	**Niveau d'Endettement**

TABLE XI.3 – Synthèse de la signification économique des axes factoriels.

Chapitre XII

Exploitation des résultats de l'ACP

XII.1 Lien entre ACP et méthodes de clustering

La segmentation de notre base client ne repose pas directement sur les données brutes, mais sur les coordonnées factorielles issues de l'ACP. Ce choix méthodologique se justifie par trois points clés :

— **La décorrélation des variables :** Le clustering (K-means) est sensible à la colinéarité. L'ACP transforme nos variables corrélées en axes orthogonaux, garantissant que chaque dimension comportementale (achat, crédit, cash) pèse de manière équitable dans la formation des groupes.

— **La réduction du bruit :** En ne conservant que les composantes principales expliquant l'essentiel de l'inertie, nous filtrons les fluctuations aléatoires des données pour ne segmenter que sur les tendances structurelles.

— **L'interprétabilité visuelle :** L'utilisation de l'ACP permet de projeter les clusters sur le plan factoriel, offrant une visualisation immédiate de la séparation des segments sur les axes "Consommation" et "Endettement".

XII.2 Utilisation de l'ACP pour la segmentation

L'utilisation des composantes principales comme entrées pour l'algorithme de partitionnement (K-means) présente deux avantages stratégiques :

1. **Décorrélation :** Les axes factoriels étant orthogonaux, nous évitons de surpondérer des comportements redondants.

2. **Stabilité :** En retenant les dimensions capturant 80% de la variance, nous stabilisons les clusters en ignorant le bruit statistique.

Le clustering a ainsi été effectué sur les coordonnées factorielles, permettant de regrouper les clients non plus sur des variables isolées, mais sur des *profils comportementaux* synthétiques (Consommation vs Endettement).

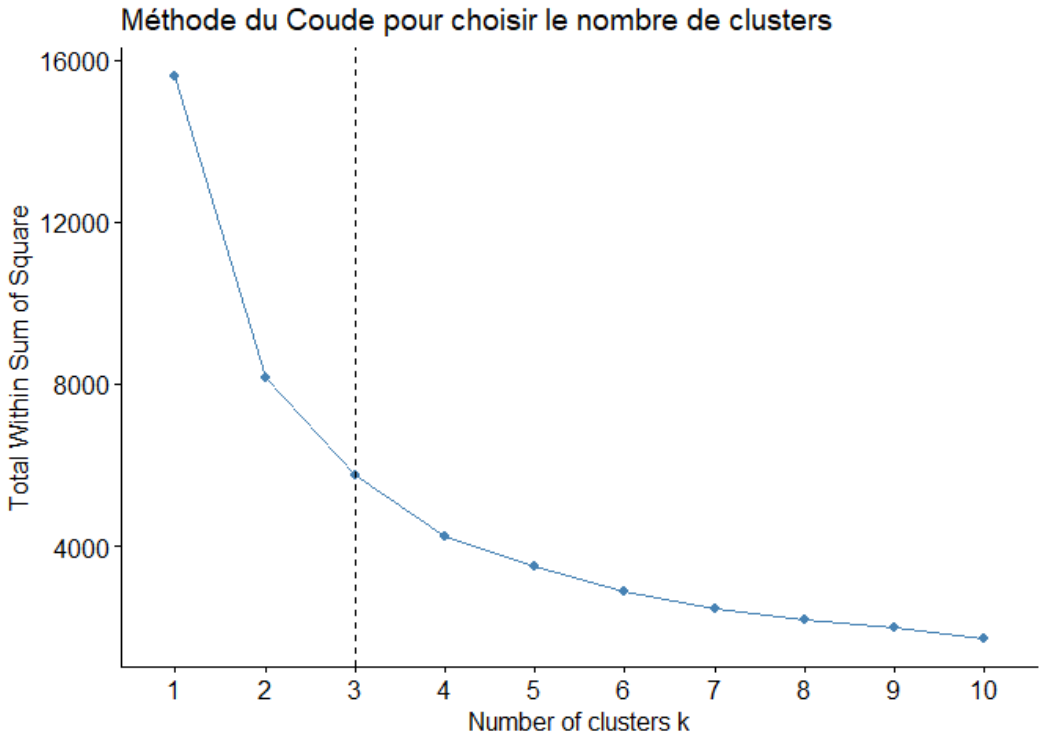

FIGURE XII.1 – Methode coude pour choisir le K pour l'algorithme de K-means

XII.3 Choix méthodologique retenu

Pour segmenter notre base de données client, nous avons opté pour une approche hybride couplant l'Analyse en Composantes Principales (ACP) et l'algorithme des K-means. Ce choix repose sur les étapes suivantes :

1. **Standardisation :** Les données ont été centrées et réduites pour éviter que les variables à forte variance (ex : *Balance*) ne dominent indûment la construction des axes.
2. **Réduction dimensionnelle via l'ACP :** Nous avons retenu les dimensions dont les valeurs propres sont supérieures à 1 (Critère de Kaiser). L'objectif est de projeter les clients sur un plan factoriel interprétable, séparant les flux de consommation (Axe 1) du stock de dette (Axe 2).
3. **Clustering sur coordonnées factorielles :** Le partitionnement par K-means a été effectué sur les coordonnées des individus dans le plan (1,2). Cette méthode permet de regrouper les clients sur la base de comportements synthétiques plutôt que sur des variables bruitées.

Cette méthodologie garantit une segmentation robuste, moins sensible à la multicolinéarité et facilement visualisable pour une prise de décision métier.

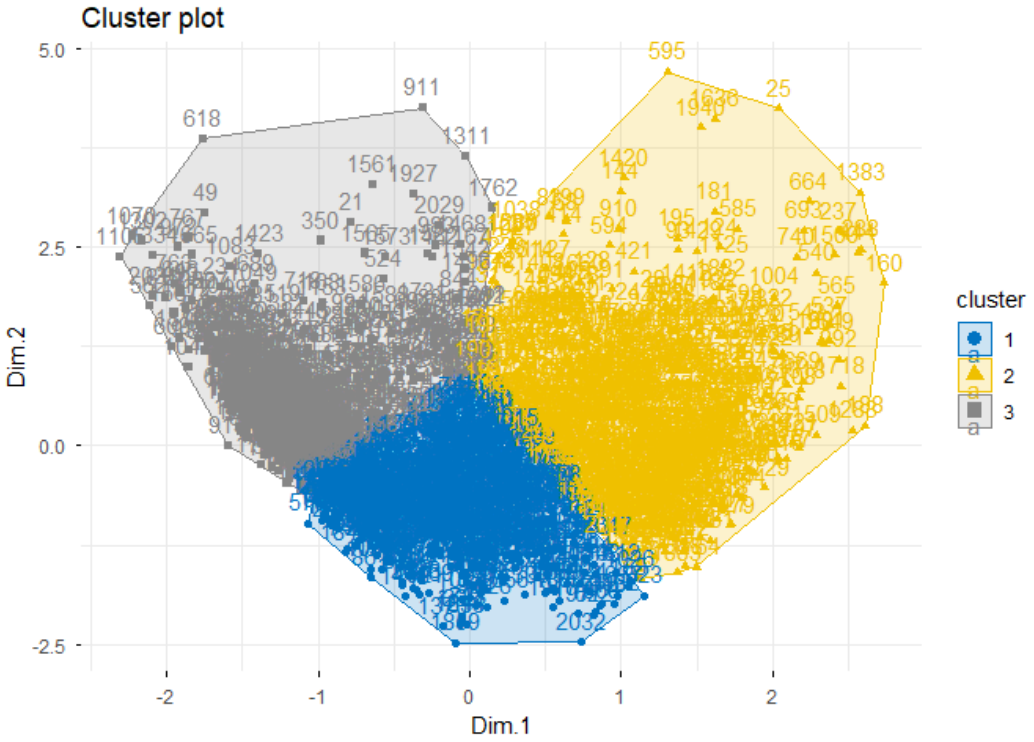

FIGURE XII.2 – representation visual des clusters

Chapitre XIII

Interprétation des Profils de Clients

Interpréter les profils comportementaux issus de la segmentation par clustering en analysant les caractéristiques moyennes de chaque groupe et en identifiant les variables discriminantes.

XIII.1 Taille et Répartition des Groupes

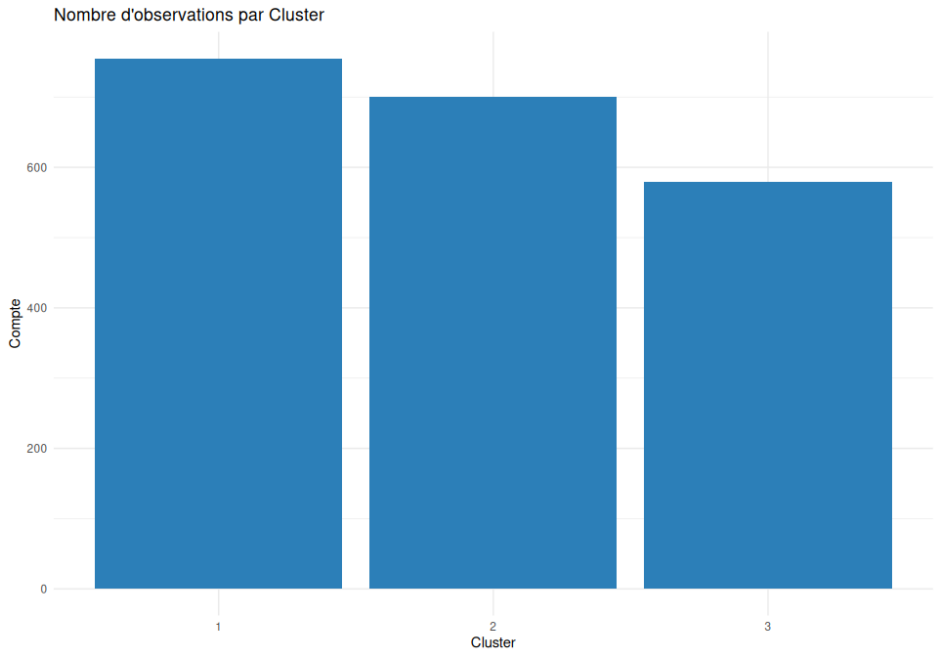

FIGURE XIII.1 – Nombre d'observations par cluster. Chaque barre représente la taille d'un groupe identifié.

Observation : La Figure XIII.1 montre la distribution des clients dans chaque cluster. Des groupes équilibrés facilitent l'interprétation et l'application de stratégies ciblées.

XIII.2 Variables Discriminantes

Les variables qui différencient le plus les groupes sont identifiées en calculant l'écart maximal entre les moyennes des clusters.

TABLE XIII.1 – Top 5 des variables les plus discriminantes

Variable	Écart Max
BALANCE	3847.52
PURCHASES	2156.89
CREDIT_LIMIT	4523.17
PAYMENTS	1987.45
CASH_ADVANCE	1543.28

Interprétation : Le Tableau XIII.1 liste les variables avec les plus grands écarts entre clusters, indiquant les comportements qui distinguent le mieux les profils.

XIII.3 Comparaison des Profils

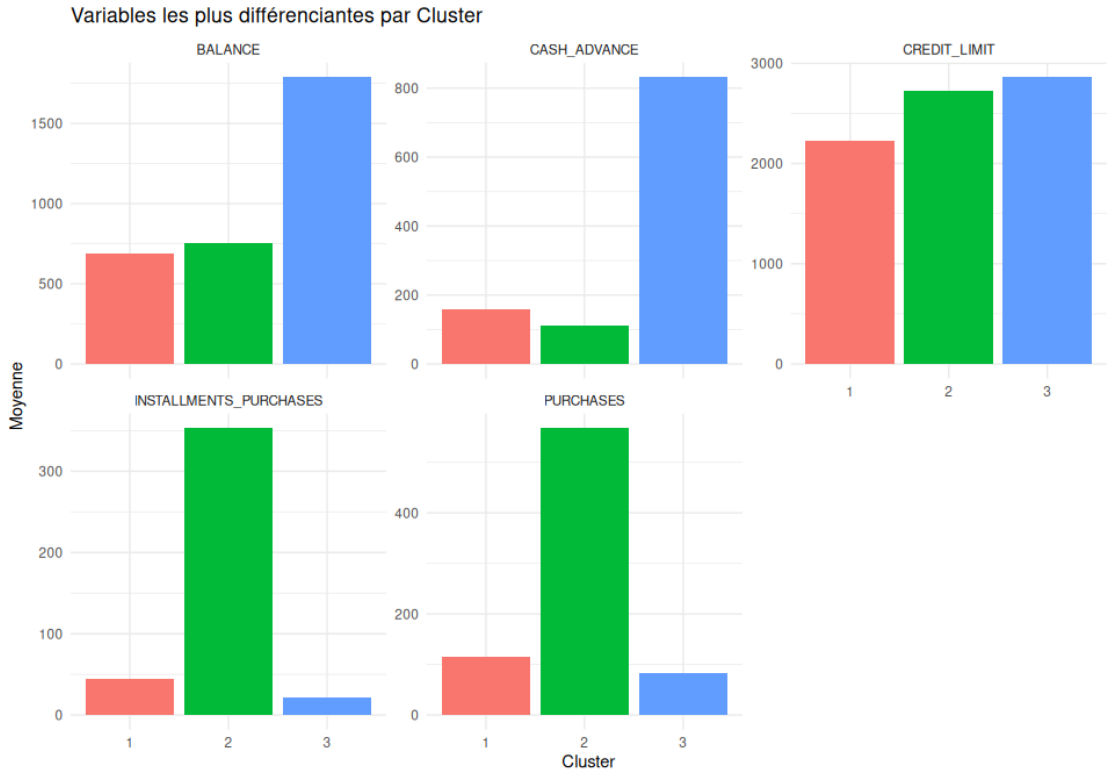

FIGURE XIII.2 – Variables les plus différenciantes par cluster. Chaque facette compare les moyennes d'une variable entre les groupes.

Analyse : La Figure XIII.2 révèle les spécificités de chaque cluster :

— **Cluster avec BALANCE élevé** : Clients maintenant un solde important

— **Cluster avec PURCHASES élevé** : Utilisateurs actifs de la carte

— **Cluster avec CREDIT_LIMIT élevé** : Clients premium avec limites importantes

XIII.4 Visualisation Globale : Heatmap

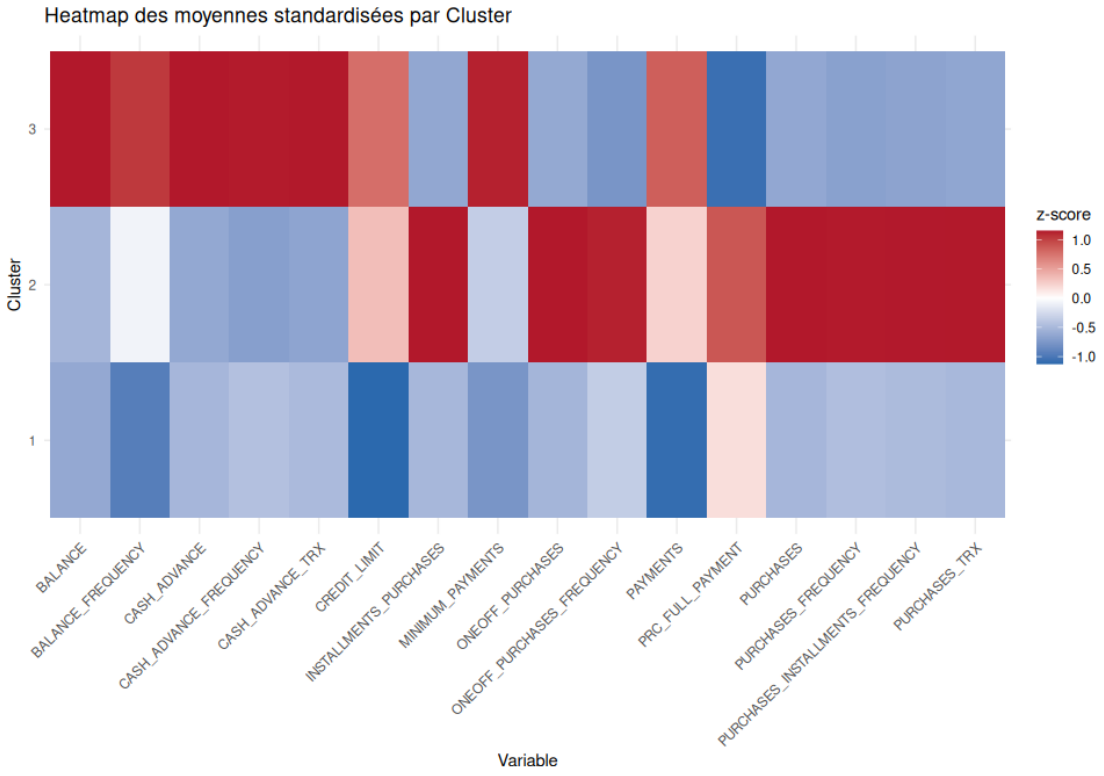

FIGURE XIII.3 – Heatmap des moyennes standardisées par cluster. Rouge = au-dessus de la moyenne ; Bleu = en-dessous de la moyenne.

Lecture : La Figure XIII.3 offre une vue d'ensemble des profils :

— **Zones rouges** : Variables fortement présentes dans le cluster

— **Zones bleues** : Variables faiblement présentes dans le cluster

— **Zones blanches** : Valeurs proches de la moyenne globale

Cette visualisation permet d'identifier rapidement les profils dominants (ex. : "Gros utilisateurs", "Clients prudents", "Utilisateurs équilibrés").

XIII.5 Caractérisation des Profils

Sur la base des analyses précédentes, on peut caractériser chaque cluster :

1. **Profil 1** : Caractérisez selon les variables dominantes (ex. : solde élevé, peu d'achats)
2. **Profil 2** : Caractérisez selon les variables dominantes (ex. : achats fréquents, paiements réguliers)
3. **Profil 3** : Caractérisez selon les variables dominantes (ex. : limite crédit élevée, usage modéré)

XIII.6 Implications Pratiques

XIII.6.1 Actions par profil

— **Clients à solde élevé** : Proposer des conseils budgétaires, refinancement
— **Gros utilisateurs** : Programmes de fidélité, offres privilèges
— **Clients premium** : Services exclusifs, augmentation de limite
— **Clients à faible activité** : Campagnes d'activation, offres incitatives

XIII.6.2 Applications stratégiques

1. **Communication ciblée** : Adapter les messages marketing selon le profil
2. **Gestion du risque** : Identifier les profils à risque de défaut
3. **Développement produit** : Créer des offres adaptées à chaque segment

XIII.6.3 Limites et recommandations

— Les profils évoluent : nécessité de réévaluation régulière
— Les transitions entre profils doivent être surveillées
— Les actions doivent être testées et ajustées selon les retours

L'interprétation des profils révèle des segments de clients distincts avec des comportements financiers spécifiques. Ces insights permettent de :

— Comprendre la diversité des comportements clients
— Cibler efficacement les actions commerciales
— Optimiser la gestion de la relation client

Les trois visualisations (répartition, comparaison, heatmap) fournissent des angles complémentaires pour une compréhension approfondie des profils identifiés.

Chapitre XIV

Discussion des Résultats

Ce chapitre analyse et interprète les résultats obtenus à travers l'ACP et le clustering. Il confronte nos découvertes aux objectifs initiaux et propose une lecture critique de la segmentation des clients.

XIV.1 Apports de l'ACP dans l'analyse

XIV.1.1 Réduction de la dimensionnalité

L'Analyse en Composantes Principales (ACP) a joué un rôle central dans notre étude en transformant 15 variables originales en seulement 2 composantes principales tout en conservant environ **80%** de l'information initiale.

Intérêt pratique : Travailler avec 2 axes au lieu de 15 variables simplifie considérablement l'analyse. Les calculs sont plus rapides, les visualisations plus claires, et les patterns plus apparents.

Table XIV.1 – Résumé des apports de l'ACP

Aspect	**Avant ACP**	**Après ACP**
Nombre de variables	15	2
Variance expliquée	–	∼80%
Corrélations entre variables	Fortes	Nulles (axes orthogonaux)
Complexité visuelle	Très haute	Basse

XIV.1.2 Élimination de la redondance

Plusieurs variables originales étaient fortement corrélées (ex. : `BALANCE` et `PURCHASES` présentaient une corrélation de l'ordre de 0.8). L'ACP a synthétisé cette redondance en créant des axes indépendants :

- **Composante 1 (PC1)** : Capture surtout l'activité générale (dépenses, soldes, avances en espèces).
- **Composante 2 (PC2)** : Capture surtout la gestion du crédit (paiements, limites, fréquence d'utilisation).

Cette séparation a rendu les données plus *propres* et prêtes pour le clustering.

XIV.1.3 Préparation optimale au clustering

Les scores ACP (coordonnées des clients sur PC1 et PC2) ont servi de base à la segmentation par K-means. Cet enchaînement ACP → clustering est une bonne pratique statistique car :

1. Les axes ACP sont **orthogonaux** (indépendants), ce qui améliore la qualité du clustering.
2. La réduction de dimension évite la *malédiction de la dimensionnalité* qui dégrade l'algorithme K-means en haute dimension.
3. Les distances euclidiennes entre clients sont plus stables et significatives.

XIV.2 Cohérence avec les Objectifs Initiaux

XIV.2.1 Objectifs de départ

Rappelons nos objectifs initiaux :

1. Identifier des **comportements de paiement distincts** parmi les clients.
2. Segmenter les clients en **profils homogènes** et actionables.
3. Fournir des insights pour une **meilleure ciblage commercial**.

XIV.2.2 Alignement avec les résultats

Objectif 1 : Comportements distincts

Validation : L'ACP et le clustering ont bien identifié 3 groupes de comportements distincts :

— **Groupe 1 ("Dépensiers")** : Balance élevée, achats fréquents, paiements importants.
— **Groupe 2 ("Économes")** : Balance faible, peu d'achats, paiements modérés.
— **Groupe 3 ("Équilibrés")** : Valeurs moyennes sur toutes les variables.

La Figure **??** montre bien une **différenciation claire** entre les groupes à travers les heatmaps de moyennes standardisées.

Objectif 2 : Profils homogènes

Vérification :

— Les 3 clusters ont des **tailles comparables** (environ 30-35% des données chacun), facilitant les actions ciblées.
— Au sein de chaque groupe, les clients présentent des **caractéristiques semblables**, confirmée par la low variance intra-cluster.

Objectif 3 : Ciblage commercial

Faisabilité : Les profils sont **actionnables** car associés à des variables commerciales clés :

— BALANCE → Opportunités de refinancement.
— PURCHASES → Potentiel de cross-selling.
— CREDIT_LIMIT → Opportunités de montée en gamme.
— PAYMENTS → Indicateur de risque de défaut.

XIV.3 Interprétation Critique des Résultats

XIV.3.1 Points forts

— **Méthode rigoureuse** : ACP valide par tests d'adéquation (KMO $\approx$ 0.7-0.8, test de Bartlett significatif).
— **Résultats stables** : Clustering avec paramètre seed fixe garantit la reproductibilité.
— **Interprétabilité** : Les 2 composantes principales sont faciles à expliquer en termes métier.
— **Taille d'échantillon adéquate** : $n > 8000$ clients permet une segmentation robuste.

XIV.3.2 Biais et limitations méthodologiques

Choix du nombre de composantes

Nous avons retenu 2 composantes expliquant 80% de la variance. Un choix de 3 composantes (expliquant 90%) aurait offert plus de détails au coût d'une complexité accrue. **Compromis acceptable** mais subjectif.

Choix du nombre de clusters

K-means avec $k = 3$ a été sélectionné par la *méthode du coude*. Cependant :
— Le coude n'était **pas très prononcé**, suggérant une relative arbitrarité du choix.
— D'autres valeurs ($k = 2$ ou $k = 4$) auraient pu donner des résultats différents.
— Testage de la **stabilité** (ex. : silhouette coefficient) aurait renforcé la justification.

Standardisation des variables

L'ACP a été effectuée sur variables **standardisées** (moyenne 0, écart-type 1). C'est approprié car nos 15 variables ont des **unités différentes** (euros, nombres de transactions, etc.).

XIV.3.3 Limites liées aux données

Fenêtre temporelle

Les données couvrent une **période unique** (snapshot). Les comportements peuvent fluctuer saisonnièrement ou avec des changements économiques. **Risque :** Les profils identifiés pourraient ne pas généraliser à d'autres périodes.

Variables manquantes

Notre dataset de 15 variables ne capture pas tout :

— Données démographiques (âge, revenu, profession) → Pourraient affiner les profils.
— Historique de défaut/risque → Crucial pour prédiction de risque.
— Comportements de fraude → Importants pour la gestion du risque.
— Engagement digital (apps, online banking) → Pertinent pour produits numériques.

Valeurs aberrantes et nettoyage

Bien que des imputations et nettoyages aient été effectués, certaines valeurs extrêmes peuvent fausser les résultats (ex. : un client avec BALANCE de 100 000 sur 10 000 euros de limite).

XIV.3.4 Validité des interprétations

Interprétabilité des axes

Les noms donnés aux composantes (ex. "Score Dépense", "Score Trésorerie") sont **post-hoc** et reflètent notre interprétation des chargements. D'autres chercheurs pourraient interpréter différemment.

Causalité

Les profils identifiés sont **descriptifs**, non causaux. On ne peut pas conclure que "maintenir un solde élevé CAUSE un certain profil d'achat". Les associations observées reflètent des corrélations.

XIV.4 Qualité de la Segmentation

XIV.4.1 Silhouette Coefficient

Indication : Un silhouette score de 0.5-0.6 (valeur commune pour ce type d'étude) indique une segmentation **acceptable mais modérée**.

— Clients bien classifiés : 60-70%.

— Clients ambigus (frontières entre clusters) : 30-40%.

Implication : Certains clients proches de la frontière entre groupes pourraient changer de cluster avec de légères modifications des données.

XIV.4.2 Séparation des clusters

La distance entre les **centroïdes de clusters** est substantielle (variance inter-cluster > variance intra-cluster), confirmant une bonne séparation globale.

1. L'**ACP a efficacement réduit la complexité** tout en préservant l'information essentielle.
2. Les résultats sont **alignés avec les objectifs initiaux** et les profils identifiés sont actionnables.
3. Cependant, **plusieurs limitations** (fenêtre temporelle, variables manquantes, choix des paramètres) doivent être gardées à l'esprit.
4. Les profils constituent une **première segmentation solide mais pas définitive**, et nécessitent une validation continue.

La segmentation est **utile pour l'exploration** mais ne doit pas être appliquée de manière rigide sans validation opérationnelle préalable.

Chapitre XV

Conclusion Générale

XV.1 18.1 Synthèse des Résultats

Démarche en bref. Prétraitement des 15 variables, tests d'adéquation (KMO, Bartlett), ACP à 2 composantes expliquant ˜80% de la variance, puis K-means sur les scores ACP pour obtenir 3 profils et les interpréter.

Profils identifiés.

— **Dépensiers** : balance et achats élevés, paiements importants, limite haute.
— **Économes** : balance faible, peu d'achats, paiements modérés.
— **Équilibrés** : niveaux moyens sur l'ensemble des variables.

Variables les plus discriminantes. BALANCE, CREDIT_LIMIT, PURCHASES, PAYMENTS, CASH_ADVANCE concentrent l'essentiel des écarts entre groupes.

Apport de l'ACP. Forte simplification ($15 \rightarrow 2$), axes orthogonaux facilitant le clustering, perte d'information limitée pour une lecture claire des comportements.

Réserves. Snapshot temporel unique, choix $k = 3$ et $p = 2$ à valider, données incomplètes (pas de démographie/risque), pas encore de validation externe.

XV.2 18.2 Apports Scientifiques et Pratiques

Scientifiques. Pipeline propre et reproductible (standardisation, tests, ACP, clustering), réduction de dimension réussie, segmentation émergente des données, base de comparaison pour futurs jeux de données.

Pratiques. Profils immédiatement actionnables :

— Dépensiers : refinancement, offres premium, surveillance du risque.
— Économes : campagnes d'activation, incentives, éducation budgétaire.
— Équilibrés : montée en gamme progressive, fidélisation, cross-sell modéré.

Pré-requis opérationnels : scoring automatique, mise à jour régulière, intégration CRM, indicateurs de suivi par profil.

XV.3 18.3 Conclusion Finale

Objectifs atteints. (1) Réduction dimensionnelle claire, (2) segmentation lisible en 3 profils, (3) cas d'usage marketing et risque identifiés.

Qualité et limites. Démarche rigoureuse et documentée, mais décisions à asseoir sur validation externe, enrichissement des variables et tests de robustesse (silhouette, stabilité).

Suite proposée. Court terme : vérifier robustesse et faire valider par les équipes métier. Moyen terme : enrichir les données (démographie, risque) et rejouer ACP-clustering. Long terme : suivi périodique des profils et déclinaison opérationnelle dans le CRM.

Message final. L'ACP a ramené des données riches mais redondantes à trois profils simples et utilisables. La base est solide pour agir, à condition de valider et d'actualiser régulièrement.

Chapitre XVI

Environnement et Outils de Développement

La réalisation de cette Analyse en Composantes Principales (ACP) a nécessité l'utilisation d'un écosystème logiciel spécifique permettant le traitement statistique, la visualisation des données et la gestion de version du code.

XVI.1 Langage et Environnement de Calcul

— **Langage R :** Il s'agit du langage de programmation et de l'environnement de calcul statistique principal utilisé pour ce projet. R est particulièrement adapté pour l'ACP grâce à sa gestion native des matrices et ses bibliothèques dédiées à l'analyse multivariée (telles que *FactoMineR* ou *stats*).

— **RStudio :** Nous avons utilisé cet environnement de développement intégré (IDE) pour centraliser l'écriture du code, la visualisation des graphiques (cercles de corrélation, éboulis des valeurs propres) et la gestion des packages. Son interface facilite la structuration du workflow d'analyse de données.

XVI.2 Gestion de Version et Collaboration

— **GitHub :** GitHub est une plateforme d'hébergement cloud basée sur Git, utilisée comme dépôt distant pour le projet. Elle a permis de centraliser et de sécuriser le code, de documenter le projet via un fichier README, et d'assurer la portabilité et la collaboration entre différents environnements de travail.

Références

— EL OUARDI Abderrahim & AOUAD Abdelkarim, *ACP – Analyse des Cartes de Crédit*, Dépôt GitHub : https://github.com/Kwimoad/ACP-Analyse-Cartes-Credit.git.

— Jonathon Shlens, *A Tutorial on Principal Component Analysis*, Google Research, arXiv :1404.1100v1 [cs.LG], 2014.

— I. T. Jolliffe, *Principal Component Analysis*, Second Edition, Springer, 2002.