# Analyse des Résiliations de Cartes de Crédit

Boujamaa Atrmouh - Christophe WANG

3 janvier 2024

## Contents

# Introduction

Les services de cartes de crédit, moteurs de l'économie moderne, font face au défi majeur de la résiliation client. Comprendre les tenants et aboutissants de cette décision devient impératif. Cette étude explore les facteurs et profils clients liés à la résiliation des cartes de crédit, avec pour objectif la prédiction de ce phénomène. Des analyses approfondies, allant de la corrélation à la modélisation prédictive, offriront des insights précieux pour les institutions financières.

## Source des données :

Les données utilisées dans cette analyse proviennent du jeu de données "Credit Card Customers" disponible sur Kaggle à l'adresse suivante : Credit Card Customers Dataset.

Les données ont été préalablement traitées, et vous pouvez vous référer au document de pré-analyse pour obtenir des détails sur ce processus. Ensuite, une phase de nettoyage a été effectuée pour préparer les données, lesquelles ont été ensuite sauvegardées dans un fichier CSV comprenant les informations traitées.

## Outils Utilisés :

- Langage de programmation : R
- Bibliothèques : ggplot2, skimr, VIM, outliers, psych, corrplot, cluster

## Problématique :

**Problématique :** Quels facteurs et profils de clients sont associés à la résiliation des services de cartes de crédit, et comment peut-on les prédire ?

## Objectifs :

1. **Compréhension des caractéristiques démographiques :**
   - **Question :** Quels sont les profils démographiques des clients résiliant leurs services de cartes de crédit?
   - **Méthodologie :** Analyses descriptives univariées pour chaque variable démographique.
2. **Étude des relations entre les variables :**
   - **Question :** Existe-t-il des relations significatives entre les différentes variables démographiques et transactionnelles?
   - **Méthodologie :** Analyses bivariées avec des tests de corrélation pour évaluer les relations.
3. **Facteurs sous-jacents à la résiliation :**
   - **Question :** Quels sont les facteurs sous-jacents qui contribuent le plus à la décision de résilier?
   - **Méthodologie :** Analyses factorielles pour réduire la dimensionnalité des données.
4. **Modélisation prédictive :**
   - **Question :** Peut-on développer un modèle de régression prédictive pour estimer la probabilité de résiliation en fonction des variables disponibles?
   - **Méthodologie :** Utilisation de méthodes de régression.
5. **Classification des clients :**
   - **Question :** Peut-on classer les clients en groupes distincts en fonction de leurs comportements et caractéristiques, en particulier en ce qui concerne la résiliation?
   - **Méthodologie :** Utilisation de méthodes de classification.

# Objectif 1: Compréhension des caractéristiques démographiques

L'objectif de cette première phase d'analyse est de comprendre les profils démographiques des clients résiliant leurs services de cartes de crédit à travers plusieurs variables clés. Les choix des variables à explorer ont été stratégiquement guidés par l'objectif de dévoiler des tendances significatives susceptibles d'informer les motifs de résiliation. Chacun de ces choix est étayé par une justification basée sur notre compréhension des comportements financiers et des facteurs socio-économiques.

## Justification des choix de variables :

1. **Customer_Age (Age du client) :** L'âge est une dimension cruciale pour appréhender les dynamiques financières. Notre décision d'inclure cette variable repose sur l'hypothèse qu'elle pourrait révéler des variations générationnelles dans les habitudes de résiliation.
2. **Gender (Genre) :** La diversité de genre peut influencer les préférences financières. Son inclusion dans notre analyse vise à identifier d'éventuelles disparités dans la résiliation en fonction du genre des clients.
3. **Dependent_count (Nombre de personnes à charge) :** Les responsabilités familiales sont des facteurs déterminants. L'inclusion de cette variable cherche à établir des liens entre le nombre de personnes à charge et la propension à résilier les services.
4. **Education_Level (Niveau d'éducation) :** Le niveau d'éducation est souvent corrélé à la stabilité financière. Nous avons choisi cette variable pour son potentiel à dévoiler des tendances entre le niveau d'éducation des clients et la résiliation.
5. **Marital_Status (Statut matrimonial) :** Le statut matrimonial peut refléter des engagements financiers et familiaux. Son inclusion dans notre analyse permet d'explorer les relations entre le statut matrimonial et la résiliation des cartes de crédit.
6. **Income_Category (Catégorie de revenu) :** Les disparités de revenus peuvent façonner les décisions financières. Nous avons intégré cette variable pour son rôle présumé dans la compréhension des variations de résiliation selon les catégories de revenus.

## Analyse :

Dans l'annexe (voir la partie Annexe : ***Profils démographiques***), vous trouverez des résultats détaillés comprenant des barplots illustrant visuellement les profils démographiques des clients résiliant leurs services de cartes de crédit pour chaque variable démographique. Cette section fournira une analyse détaillée de ces barplots, présentant ainsi une représentation graphique des distributions et des tendances démographiques associées à la résiliation des services de cartes de crédit.

1. **Customer_Age (Age du client) :** La distribution par âge révèle que la majorité des clients résiliants ont entre 40 et 50 ans, avec une concentration significative également dans la tranche de 50 à 60 ans. En revanche, les clients existants sont répartis sur une gamme plus large d'âges, avec une concentration relativement élevée dans la tranche de 40 à 50 ans. Il est essentiel de prendre en compte cette répartition lors de l'analyse des profils démographiques.
2. **Gender (Genre) :** En termes de genre, la répartition entre les clients résiliants et existants est relativement équilibrée. Les clients résiliants se composent à 51,75% de femmes et 48,25% d'hommes, tandis que les clients existants se répartissent à 46,92% de femmes et 53,08% d'hommes.
3. **Dependent_count (Nombre de personnes à charge) :** L'analyse du nombre de personnes à charge indique que les clients résiliants ont tendance à avoir une distribution plus homogène, tandis que les clients existants montrent une répartition plus variée.

4. **Education_Level (Niveau d'éducation) :** Le niveau d'éducation des clients résiliants et existants est assez similaire, avec une majorité ayant complété un diplôme de niveau collégial ou universitaire. Cette variable peut ne pas être un indicateur significatif de la résiliation.

5. **Marital_Status (Statut matrimonial) :** Le statut matrimonial montre que parmi les clients résiliants, une proportion importante est composée de personnes mariées (47,53%), suivies de près par les célibataires (44,20%). Pour les clients existants, la majorité est également mariée (50,87%), suivie de célibataires (41,13%).

6. **Income_Category (Catégorie de revenu) :** En termes de catégorie de revenu, les clients résiliants ont une distribution plus importante dans la catégorie "Moins de 40 000 $" (42,50%), tandis que les clients existants montrent une distribution plus équilibrée entre les catégories de revenus.

L'analyse des caractéristiques démographiques suggère que l'âge, le statut matrimonial et la catégorie de revenu peuvent être des facteurs importants à considérer lors de l'identification des profils démographiques des clients résiliant leurs services de cartes de crédit. Les clients résiliant leurs services de cartes de crédit sont souvent des individus dans la tranche d'âge de 40 à 50 ans, majoritairement mariés, avec un niveau d'éducation collégial ou universitaire, et une concentration plus élevée dans la catégorie de revenu "Moins de 40 000 $".

# Objectif 2: Étude des relations entre les variables

## Justification des choix de variables :

1. **Représentativité des domaines clés :**
   - **Customer_Age :** L'âge du client est souvent un facteur important dans de nombreuses analyses démographiques.
   - **Dependent_count :** Le nombre de personnes à charge peut influencer les habitudes de dépenses et la gestion financière.
   - **Credit_Limit :** La limite de crédit est un indicateur financier clé qui peut être lié à d'autres comportements financiers.
   - **Total_Trans_Amt :** Le montant total des transactions peut indiquer l'activité financière globale du client.
   - **Total_Trans_Ct :** Le nombre total de transactions peut également fournir des informations sur l'utilisation des services.
2. **Variables pertinentes pour la résiliation de carte de crédit :**
   - Ces variables sont susceptibles d'influencer la décision de résiliation des services de carte de crédit. Par exemple, l'âge, la situation familiale, les habitudes de dépenses (indiquées par la limite de crédit et le montant total des transactions), et l'activité de transaction (nombre total de transactions) pourraient tous être liés à la résiliation.
3. **Équilibre entre démographie et comportement financier :**
   - L'inclusion de variables démographiques (comme l'âge et le nombre de personnes à charge) en combinaison avec des variables transactionnelles permet d'obtenir une perspective équilibrée entre les caractéristiques personnelles et le comportement financier du client.

## Matrice de corrélation :

La matrice de corrélation, présentée dans la partie Annexe : ***Matrice corrélation***, expose les relations linéaires entre les variables démographiques et transactionnelles.

**Interprétation des résultats :**

- **Customer_Age et Dependent_count :**
  - Corrélation = -0.13
  - Interprétation : Une corrélation négative faible suggère que les clients plus âgés ont tendance à avoir moins de personnes à charge.
- **Customer_Age et Credit_Limit :**
  - Corrélation = 0.02
  - Interprétation : La corrélation est très faible, indiquant une relation presque nulle entre l'âge du client et la limite de crédit.
- **Customer_Age et Total_Trans_Amt :**
  - Corrélation = -0.05
  - Interprétation : La corrélation est faible et négative, suggérant une tendance à une légère diminution du montant total des transactions avec l'âge.

- **Customer_Age et Total_Trans_Ct :**
  - Corrélation = -0.07
  - Interprétation : Une corrélation négative faible suggère que les clients plus âgés ont tendance à effectuer moins de transactions.
- **Dependent_count et Credit_Limit :**
  - Corrélation = 0.08
  - Interprétation : Une corrélation positive faible indique que les clients avec plus de personnes à charge ont tendance à avoir une limite de crédit légèrement plus élevée.
- **Dependent_count et Total_Trans_Amt :**
  - Corrélation = 0.02
  - Interprétation : La corrélation est très faible, indiquant une relation presque nulle entre le nombre de personnes à charge et le montant total des transactions.
- **Dependent_count et Total_Trans_Ct :**
  - Corrélation = 0.06
  - Interprétation : Une corrélation positive faible suggère qu'il existe une tendance à une légère augmentation du nombre total de transactions avec le nombre de personnes à charge.
- **Credit_Limit et Total_Trans_Amt :**
  - Corrélation = 0.17
  - Interprétation : Il y a une corrélation positive modérée entre la limite de crédit et le montant total des transactions. Les clients avec une limite de crédit plus élevée ont tendance à effectuer des transactions de montant plus élevé.
- **Credit_Limit et Total_Trans_Ct :**
  - Corrélation = 0.08
  - Interprétation : La corrélation est positive faible, suggérant une relation modérée entre la limite de crédit et le nombre total de transactions.
- **Total_Trans_Amt et Total_Trans_Ct :**
  - Corrélation = 0.81
  - Interprétation : Il y a une forte corrélation positive entre le montant total des transactions et le nombre total de transactions. Cela indique que les clients effectuant un plus grand nombre de transactions ont tendance à avoir un montant total de transactions plus élevé.

## Test de corrélation :

Les tests utilisent la statistique de corrélation de Pearson ou corrélation linéaire.

(voir la partie Annexe : ***Test corrélation***)

**Interprétation des résultats :**

- **Customer_Age et Dependent_count :**
  - Corrélation significative (p-value = 2.98963e-27).
  - Interprétation : Il y a une corrélation statistiquement significative entre l'âge du client et le nombre de personnes à charge.

- **Customer_Age et Credit_Limit :**
  - Corrélation significative (p-value = 0.03722771).
  - Interprétation : Il existe une corrélation statistiquement significative entre l'âge du client et la limite de crédit, bien que la corrélation soit faible.
- **Customer_Age et Total_Trans_Amt :**
  - Corrélation significative (p-value = 0.0001127818).
  - Interprétation : Une corrélation négative significative existe entre l'âge du client et le montant total des transactions. Cela peut indiquer que les clients plus jeunes ont tendance à effectuer des transactions plus importantes.
- **Customer_Age et Total_Trans_Ct :**
  - Corrélation significative (p-value = 3.981838e-09).
  - Interprétation : Il y a une corrélation statistiquement significative entre l'âge du client et le nombre total de transactions. Les clients plus jeunes ont tendance à effectuer un nombre plus élevé de transactions.
- **Dependent_count et Credit_Limit :**
  - Corrélation significative (p-value = 7.178391e-12).
  - Interprétation : Il existe une corrélation significative entre le nombre de personnes à charge et la limite de crédit.
- **Dependent_count et Total_Trans_Amt :**
  - Corrélation non significative (p-value = 0.05111694).
  - Interprétation : Aucune corrélation statistiquement significative entre le nombre de personnes à charge et le montant total des transactions.
- **Dependent_count et Total_Trans_Ct :**
  - Corrélation significative (p-value = 1.649984e-06).
  - Interprétation : Une corrélation significative existe entre le nombre de personnes à charge et le nombre total de transactions.
- **Credit_Limit et Total_Trans_Amt :**
  - Corrélation significative (p-value = 4.594966e-48).
  - Interprétation : Une corrélation positive significative entre la limite de crédit et le montant total des transactions. Les clients avec une limite de crédit plus élevée ont tendance à effectuer des transactions plus importantes.
- **Credit_Limit et Total_Trans_Ct :**
  - Corrélation significative (p-value = 7.134977e-12).
  - Interprétation : Corrélation significative entre la limite de crédit et le nombre total de transactions. Les clients avec une limite de crédit plus élevée ont tendance à effectuer un plus grand nombre de transactions.

Les résultats de notre analyse démontrent qu'il existe des relations statistiquement significatives entre les caractéristiques démographiques, telles que l'âge du client, le nombre de personnes à charge, et les transactions financières, notamment la limite de crédit, le montant total des transactions, et le nombre total de transactions. Ces liens mettent en lumière l'importance de prendre en considération ces facteurs lors de l'évaluation des motifs de résiliation des services de cartes de crédit.

Il convient de noter que dans le contexte de notre étude, malgré les biais observés dans les résultats de corrélation, ces éléments n'impactent pas de manière significative les conclusions générales. Étant donné la nature exploratoire de notre analyse et l'objectif préliminaire de dégager des tendances, ces écarts ne remettent pas en question les grandes lignes de nos constatations.

# Objectif 3: Facteurs sous-jacents à la résiliation

## Justification des choix de variables :

Le choix des variables pour l'analyse factorielle a été guidé par l'objectif de réduire la dimensionnalité des données tout en capturant les aspects les plus significatifs liés à la résiliation des services de cartes de crédit. Les variables sélectionnées sont les suivantes :

- **Customer_Age :** L'âge peut être un indicateur important du comportement financier et de la stabilité dans la relation avec la banque. Les différences générationnelles peuvent également jouer un rôle dans la résiliation.
- **Credit_Limit :** La capacité de crédit d'un client est un aspect crucial de sa relation avec la banque. Des limites de crédit plus élevées pourraient être associées à une stabilité financière.
- **Total_Trans_Amt :** Les comportements de transaction, en particulier les montants dépensés, peuvent être des indicateurs importants du niveau d'engagement du client avec ses services financiers.
- **Attrition_Flag :** Il est essentiel d'inclure la variable cible dans l'analyse pour comprendre comment les autres variables sont liées à la résiliation.

## Analyse factorielle - Interprétation des facteurs (MR1 et MR2) :

Lors d'une analyse factorielle, les termes MR1 et MR2 font référence aux deux facteurs extraits. Ces termes sont souvent utilisés pour désigner les scores factoriels attribués à chaque observation (ici, chaque client) sur les deux facteurs respectifs. Chaque score factoriel est une combinaison linéaire des variables d'origine, pondérées par les charges factorielles. En d'autres termes, MR1 et MR2 représentent les scores attribués à chaque client sur les facteurs 1 et 2, respectivement.

### Nombre de facteurs :

Le nombre de facteurs est déterminé avec le critère de Kaiser-Guttman, qui retient les facteurs dont les valeurs propres sont supérieures à 1. (Voir la partie Annexe : ***Nombre de facteurs***)

## Interprétation des facteurs :

L'analyse factorielle avec la variable cible "Attrition_Flag" a été réalisée, et les résultats sont présentés ci-dessous. Pour une compréhension approfondie des charges factorielles, veuillez vous référer à la partie Annexe : ***Facteurs***.

- **MR1 (Facteur 1) :**
  - **Charge de Variable :** La variable ayant la charge la plus élevée sur MR1 est Total_Trans_Amt (0.95).
  - **Interprétation :** Les clients avec des montants de transactions élevés auront des scores plus élevés sur MR1. Cela suggère que des transactions financières importantes contribuent à ce facteur.
- **MR2 (Facteur 2) :**
  - **Charge de Variable :** La variable avec la charge la plus élevée sur MR2 est Credit_Limit (0.81).
  - **Interprétation :** Les clients avec des limites de crédit plus élevées auront des scores plus élevés sur MR2. Ainsi, la limite de crédit est un facteur distinct qui contribue à ce score.

### Facteurs sous-jacents à la décision de résilier :

Les résultats de l'analyse factorielle indiquent que deux facteurs principaux influent sur la décision de résilier les services de cartes de crédit :

- **Facteur 1 (MR1) - Transactions financières élevées :** Les clients avec des montants de transactions élevés sont associés à un score plus élevé sur MR1. Cela pourrait suggérer que des activités financières importantes ou fréquentes contribuent à la décision de résilier.
- **Facteur 2 (MR2) - Limite de crédit élevée :** Les clients ayant une limite de crédit plus élevée ont des scores plus élevés sur MR2. Cela indique que la limite de crédit est un autre facteur significatif lié à la résiliation.

Les facteurs sous-jacents à la résiliation des services de cartes de crédit semblent être liés aux comportements de transactions financières `Total_Trans_Amt` et aux conditions de crédit `Credit_Limit`.

# Objectif 4: Modélisation prédictive

L'objectif était de développer un modèle de régression logistique pour estimer la probabilité de résiliation des services de cartes de crédit en fonction des variables explicatives sélectionnées.

## Justification des choix de variables :

- **Customer_Age :** L'âge peut être un facteur clé dans la décision de résilier. Les habitudes bancaires et les besoins financiers peuvent varier considérablement entre les différentes tranches d'âge.
- **Dependent_count :** Le nombre de personnes à charge peut influencer la stabilité financière d'un ménage. Des responsabilités familiales plus importantes peuvent être liées à des besoins financiers différents.
- **Credit_Limit :** La limite de crédit peut refléter la stabilité financière du client. Des limites de crédit plus élevées pourraient indiquer une meilleure solvabilité et potentiellement une moindre propension à résilier.
- **Total_Trans_Amt et Total_Trans_Ct :** Ces variables représentent le montant total des transactions et le nombre total de transactions effectuées par le client. Des niveaux élevés dans ces deux domaines pourraient indiquer une utilisation fréquente et substantielle des services, ce qui pourrait être lié à la satisfaction et à la fidélité.

## Modèle de régression logistique :

### Transformation de la variable cible :

La variable cible `Attrition_Flag` a été transformée en une variable binaire où 1 représente les clients ayant résilié et 0 les clients non résiliés.

Une vérification des valeurs uniques de la variable transformée confirme qu'il y a deux catégories, 0 et 1, ce qui est conforme à l'attente. (Voir la partie Annexe : ***Transformation de la variable cible***)

### Sortie statistique :

Les détails complets de la sortie statistique du modèle de régression logistique, incluant les coefficients, les p-values, ainsi que d'autres métriques pertinentes, sont disponibles en annexe (voir la partie Annexe : ***Modèle de régression logistique***). Cette annexe offre une analyse détaillée de la performance du modèle et de la contribution individuelle des variables explicatives à la prédiction de la probabilité de résiliation des services de cartes de crédit.

### Interception (constante) significative :

L'interception du modèle est significative avec un coefficient estimé de 2.360 et une erreur standard de 0.2592. Cela indique que la probabilité de résiliation est significativement différente de zéro lorsque toutes les variables explicatives sont égales à zéro.

**Variables explicatives :**

- **Customer_Age :** La p-value de 0.0415 suggère une signification marginale de l'âge du client dans la probabilité de résiliation. Chaque année supplémentaire diminue la probabilité de résiliation.
- **Dependent_count :** Le coefficient positif de 0.1108 indique que la probabilité de résiliation augmente avec le nombre de personnes à charge.
- **Credit_Limit :** La p-value de 0.00679 suggère une signification, et le coefficient négatif indique que des limites de crédit plus élevées sont associées à une diminution de la probabilité de résiliation.
- **Total_Trans_Amt et Total_Trans_Ct :** Les coefficients négatifs des deux variables suggèrent que des montants de transaction totaux plus élevés et un nombre total de transactions plus élevé sont associés à une diminution de la probabilité de résiliation.

**Performance du modèle :**

- La p-value globale du modèle est très basse, indiquant qu'au moins une des variables explicatives est utile pour prédire la résiliation.
- La réduction significative de la deviance résiduelle par rapport à la deviance nulle suggère un bon ajustement du modèle.

Le modèle suggère que l'âge, le nombre de personnes à charge, le crédit limite, le montant total des transactions, et le nombre total de transactions sont des facteurs significatifs pour prédire la probabilité de résiliation des services de cartes de crédit. Les résultats indiquent des tendances cohérentes avec l'intuition, par exemple, des clients plus jeunes ou avec des limites de crédit plus élevées étant moins susceptibles de résilier.

# Objectif 5: Classification des clients

L'objectif était de classer les clients en groupes distincts en fonction de leurs comportements et caractéristiques, en particulier en ce qui concerne la résiliation des services de cartes de crédit. Pour ce faire, nous avons utilisé la méthode de classification k-means.

## Nombre Optimal de Clusters:

### Méthode du Coude:

Après avoir utilisé la méthode du coude pour déterminer le nombre optimal de clusters, nous avons identifié que le coude potentiel se situe autour de 2 clusters. Le graphique du coude montre que la somme totale des carrés intra-clusters diminue considérablement jusqu'à 2 clusters, après quoi la diminution devient plus lent. (Voir la partie Annexe : ***Méthode du Coude***)

### Méthode de la Silhouette:

La méthode de la silhouette suggère également que le nombre optimal de clusters est 2. La silhouette moyenne est maximale lorsque le nombre de clusters est fixé à 2. (Voir la partie Annexe : ***Méthode de la Silhouette***)

## Résultats de la Classification :

Référez-vous à la page annexe : ***Résultat de la classification*** pour une visualisation détaillée des résultats de la classification, y compris les graphiques avec (Voir la partie Annexe : ***Avec les points de données***) et sans points de données (Voir la partie Annexe : ***Sans les points de données***). Le modèle de k-means a généré deux clusters distincts, numérotés de 1 à 2. Voici un résumé des caractéristiques moyennes de chaque cluster :

- **Cluster 1 :**
  - Age moyen : 46.26 ans
  - Limite de crédit moyenne : 4,873.74
  - Montant total des transactions moyen : 4,169.55
  - Nombre total de transactions moyen : 63.65
- **Cluster 2 :**
  - Age moyen : 46.80 ans
  - Limite de crédit moyenne : 26,406.12
  - Montant total des transactions moyen : 5,512.59
  - Nombre total de transactions moyen : 68.80

### Interprétation :

Cette classification permet de distinguer deux profils distincts de clients en fonction de leurs caractéristiques.

- **Cluster 1 :** Il semble y avoir un groupe de clients plus jeunes que l'autre avec des limites de crédit plus basses et un nombre de transactions moins élevé. Cela pourrait représenter un segment de clients moins engagés ou débutants.
- **Cluster 2 :** Ce cluster semble être composé de clients plus âgés avec une limite de crédit moyenne élevée et un nombre élevé de transactions, ce qui peut indiquer des clients fidèles et actifs.

**Évaluation de la Classification :**

La qualité de la classification peut être évaluée en examinant la variabilité totale expliquée par le modèle. Cela peut être calculé à l'aide de la somme des carrés intra-cluster (within-cluster sum of squares, WSS) et la somme des carrés inter-cluster (between-cluster sum of squares, BSS). Dans notre cas :

$$\text{Variabilité totale expliquée (\%)} = \frac{\text{BSS}}{\text{WSS} + \text{BSS}} \times 100$$

Les composants du modèle k-means nous fournissent les valeurs nécessaires pour ce calcul :

$$\text{BSS} = \text{sum(kmeans\_model\$betweenss)}$$

$$\text{WSS} = \text{sum(kmeans\_model\$withinss)}$$

Substituons ces valeurs dans la formule pour obtenir la variabilité totale expliquée. Dans ce cas, environ 68.3 % de la variabilité totale des données est expliquée par la classification en clusters. Cela suggère une certaine robustesse du modèle dans la capture des tendances au sein des données.

**Utilisation des Résultats :**

Les résultats de la classification peuvent être utilisés pour mieux comprendre les clients et adapter les stratégies commerciales en conséquence. Par exemple, les entreprises peuvent cibler les clients du cluster 1 avec des offres de fidélisation, tandis que les entreprises peuvent cibler les clients du cluster 2 avec des offres de bienvenue.

# Conclusion

Cette étude approfondie visait à répondre à la problématique centrale : **"Quels facteurs et profils de clients sont associés à la résiliation des services de cartes de crédit, et comment peut-on les prédire ?"** Les résultats obtenus ont permis d'apporter des éclairages significatifs sur cette question complexe.

**Récapitulation des Découvertes :** L'analyse détaillée des variables, qu'elles soient démographiques ou transactionnelles, a révélé des relations subtiles mais significatives. L'âge, la situation familiale, la limite de crédit, et les comportements transactionnels ont émergé comme des facteurs clés influençant la décision de résiliation. Ces résultats offrent une perspective approfondie sur les facteurs et les profils de clients liés à la résiliation des services de cartes de crédit, fournissant ainsi une compréhension approfondie des éléments en jeu.

**Impact sur la Prise de Décision Stratégique :** Ces résultats ne sont pas simplement des observations, mais des outils puissants pour les institutions financières. Elles peuvent ajuster leurs stratégies de rétention en tenant compte des facteurs critiques. Par exemple, des offres personnalisées pour les segments démographiques spécifiques ou des ajustements de limites de crédit pourraient être envisagés pour renforcer la fidélité des clients.

**Ouverture vers des Études Futures :** Cette étude constitue une base pour approfondir la compréhension des motifs de résiliation des services de cartes de crédit. Des investigations futures pourraient se concentrer sur des aspects spécifiques, notamment l'impact des programmes de fidélisation, l'influence des canaux de communication sur la rétention, ou encore l'exploration des facteurs psychologiques liés à la décision de résilier. En explorant ces domaines, les chercheurs pourraient apporter des éclairages supplémentaires, contribuant ainsi à affiner davantage les stratégies de rétention des institutions financières.

Cette investigation vise à guider les décisions futures des institutions financières en fournissant des perspectives sur les dynamiques de résiliation des services de cartes de crédit. Les "insights" ou résultats obtenus peuvent servir de fondement pour des stratégies de rétention plus efficaces et une adaptation continue aux besoins changeants des clients.

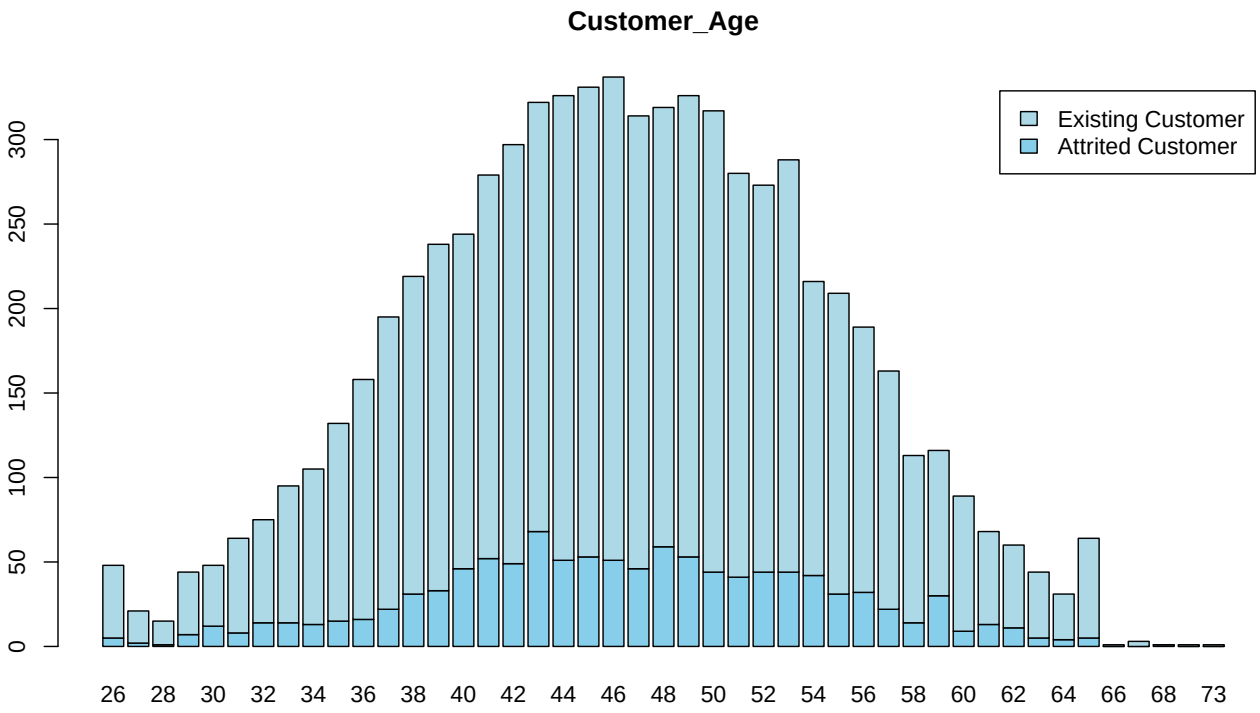
# Partie Annexe

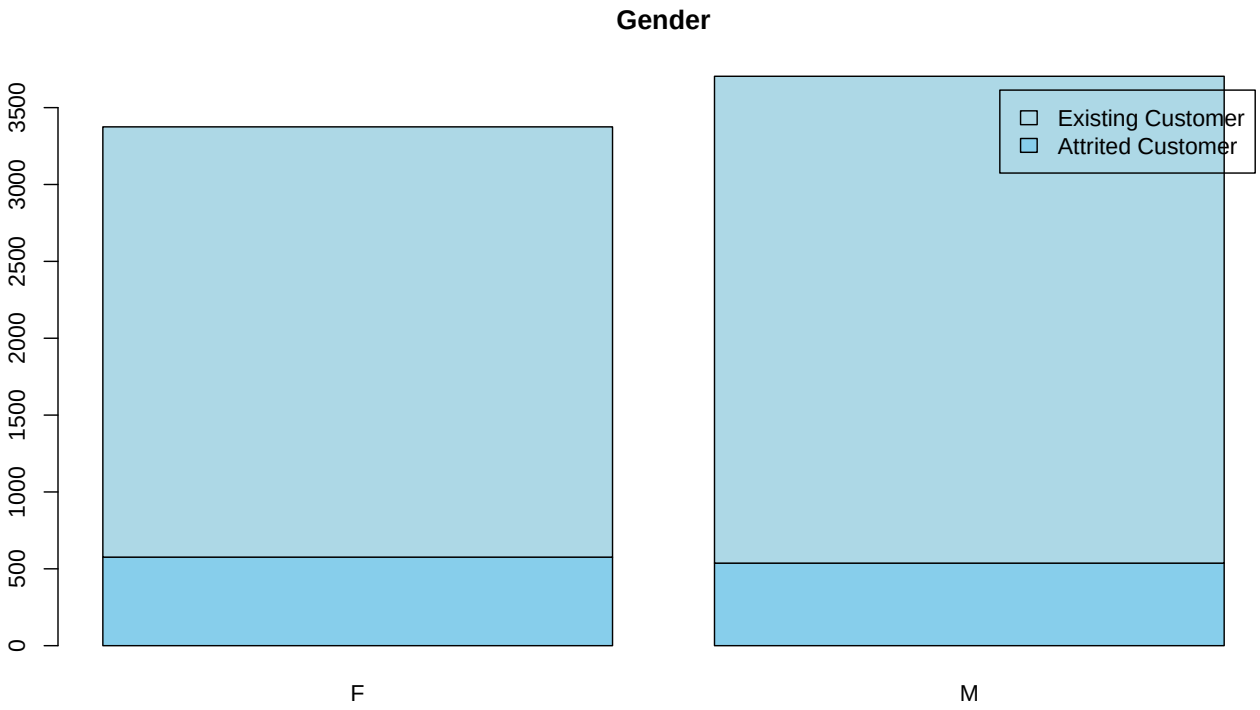
## 1 - Compréhension des caractéristiques démographiques

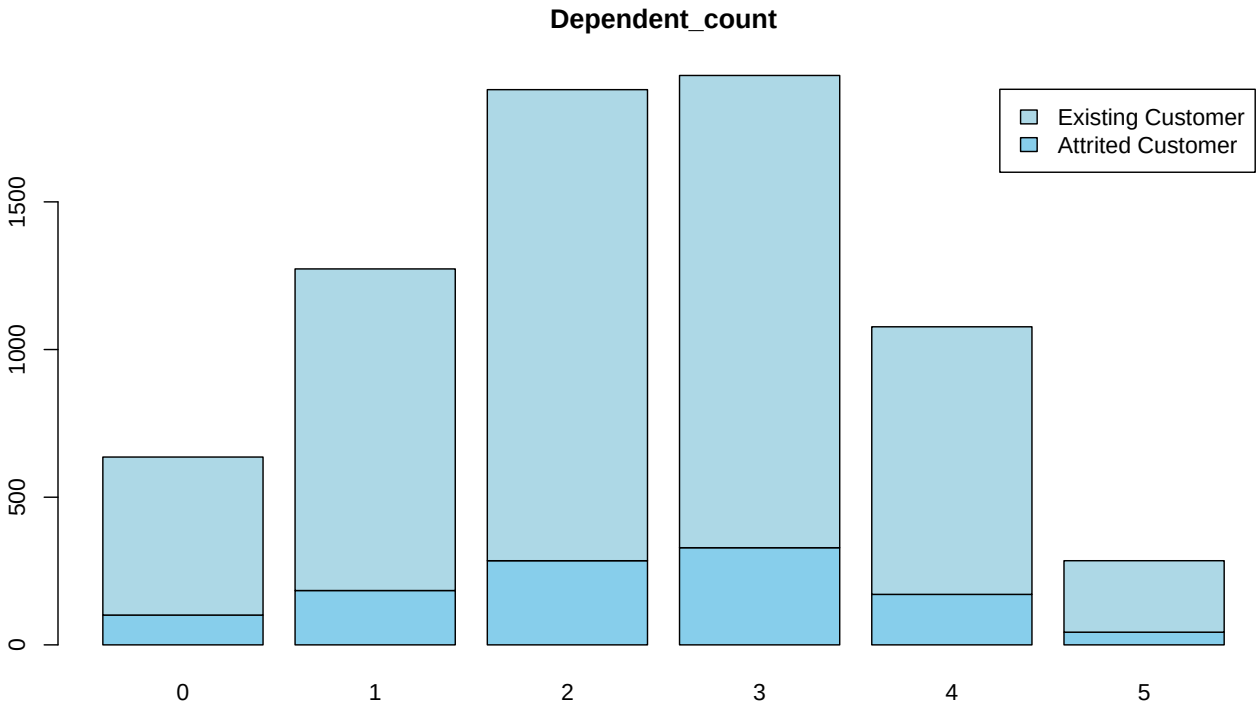
### Profils démographiques

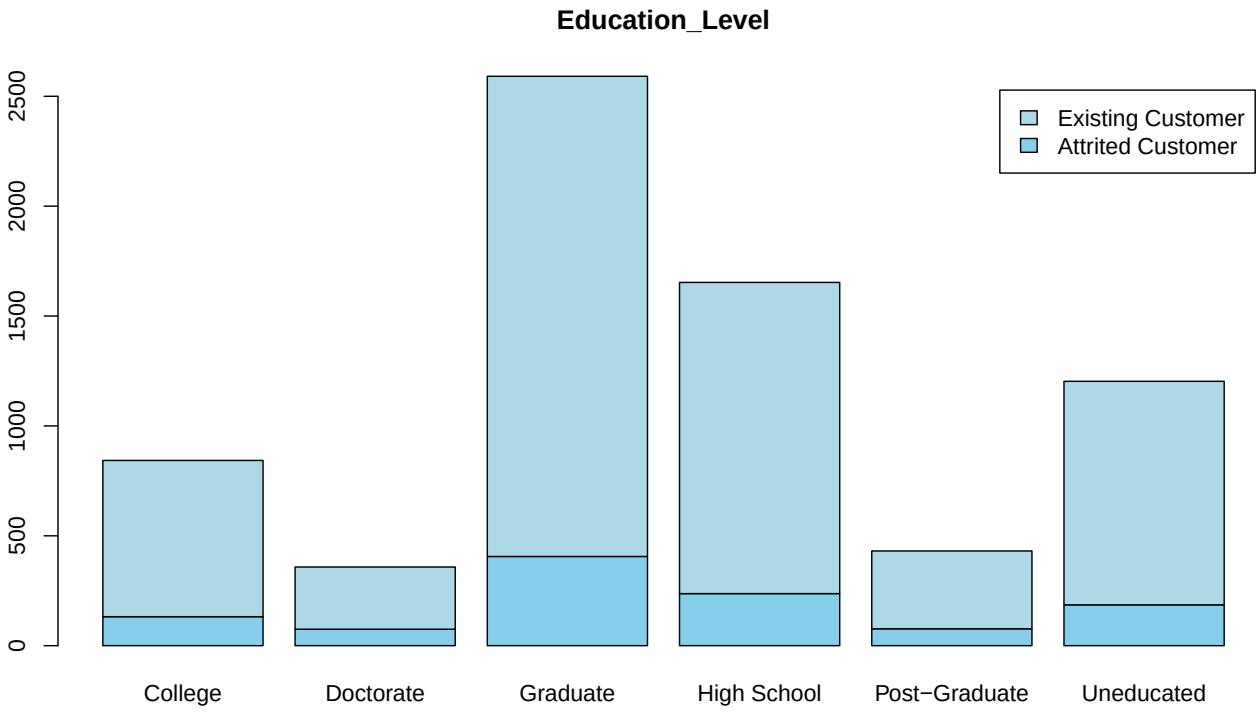
Cette annexe détaille les résultats de l'analyse des caractéristiques démographiques des clients ayant résilié leurs services de cartes de crédit. Les graphiques ci-dessous illustrent visuellement les profils démographiques en relation avec la résiliation des services.

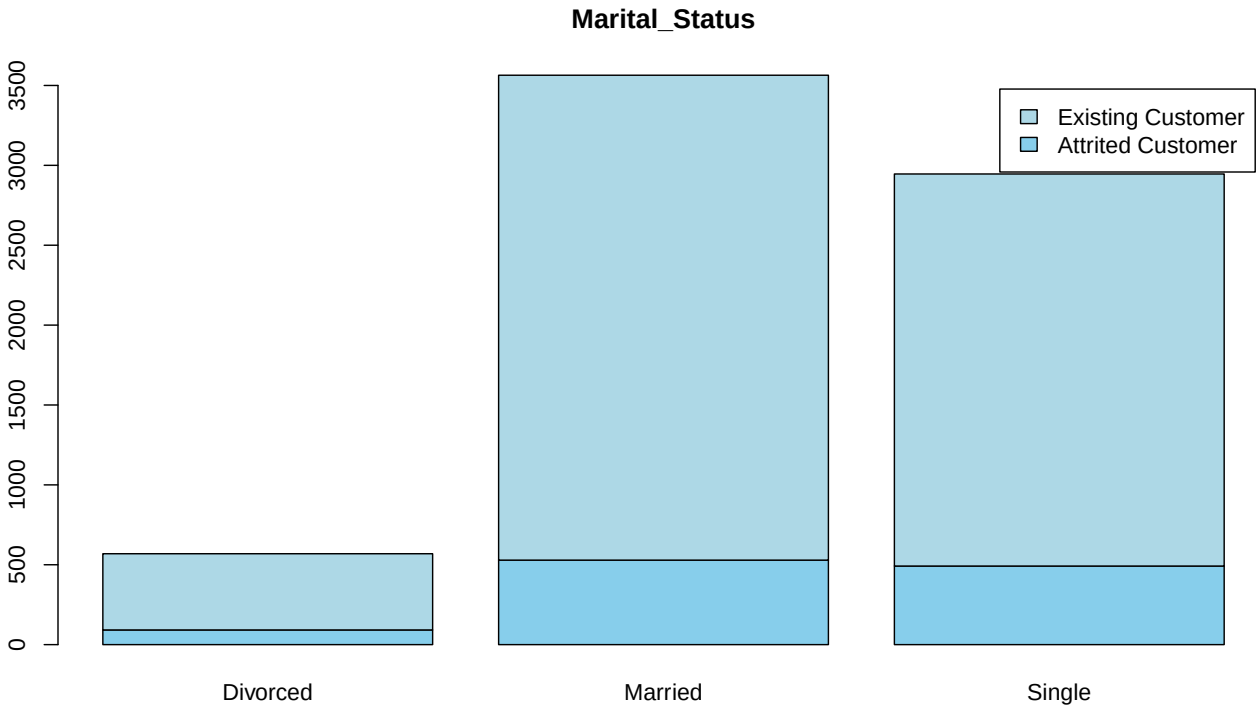
### En valeurs brutes

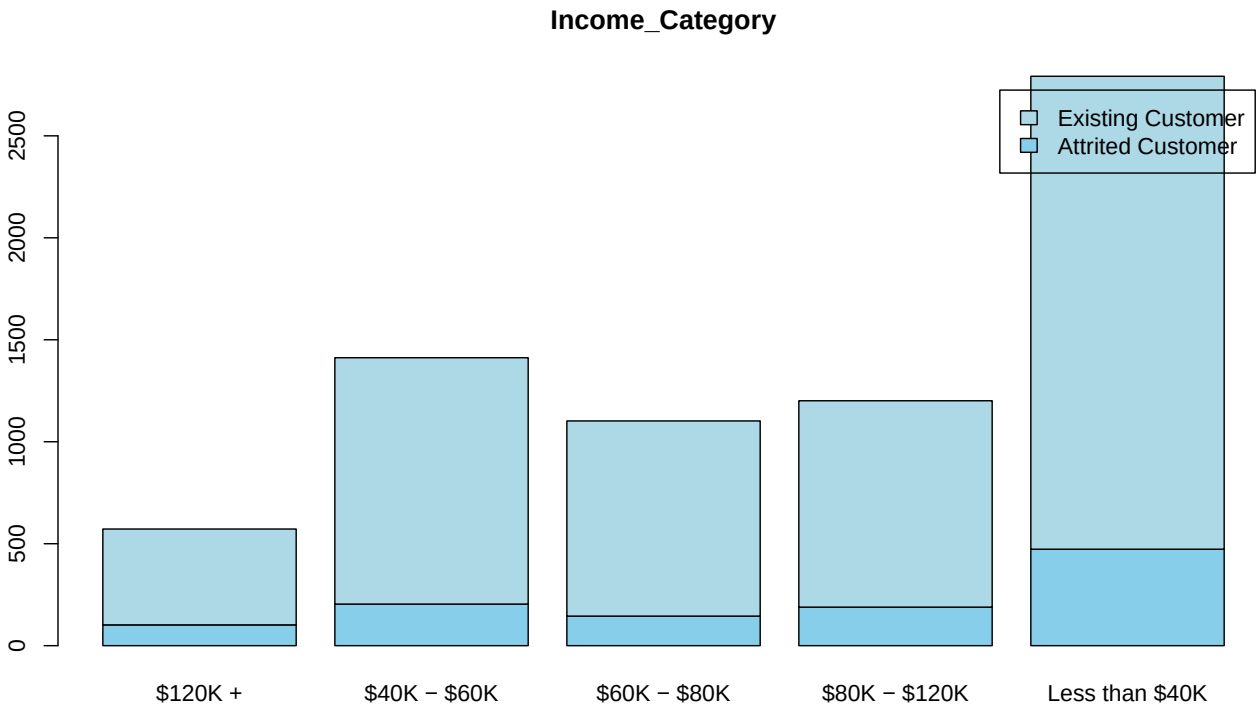



Gender
Existing Customer
Attrited Customer
0
500
1000
1500
2000
2500
3000
3500
F
M


Dependent_count
Existing Customer
Attrited Customer
0
500
1000
1500
0
1
2
3
4
5

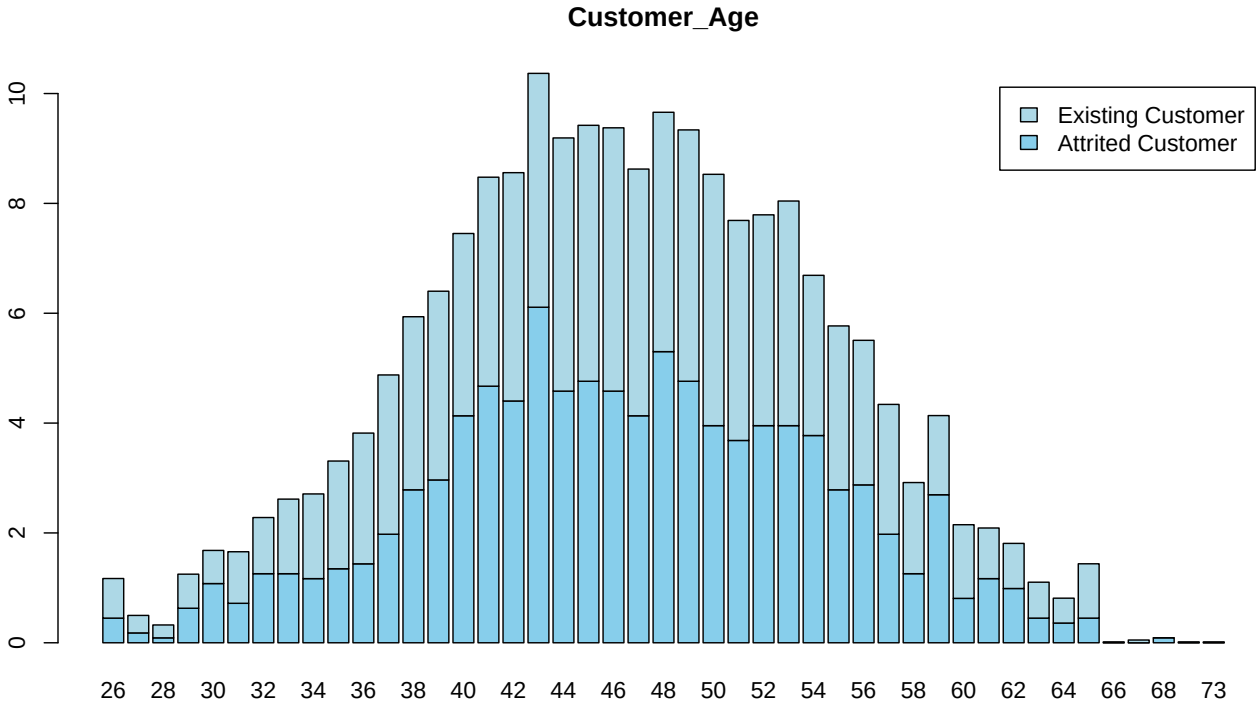

Education_Level
Existing Customer
Attrited Customer
0
500
1000
1500
2000
2500
College
Doctorate
Graduate
High School
Post−Graduate
Uneducated

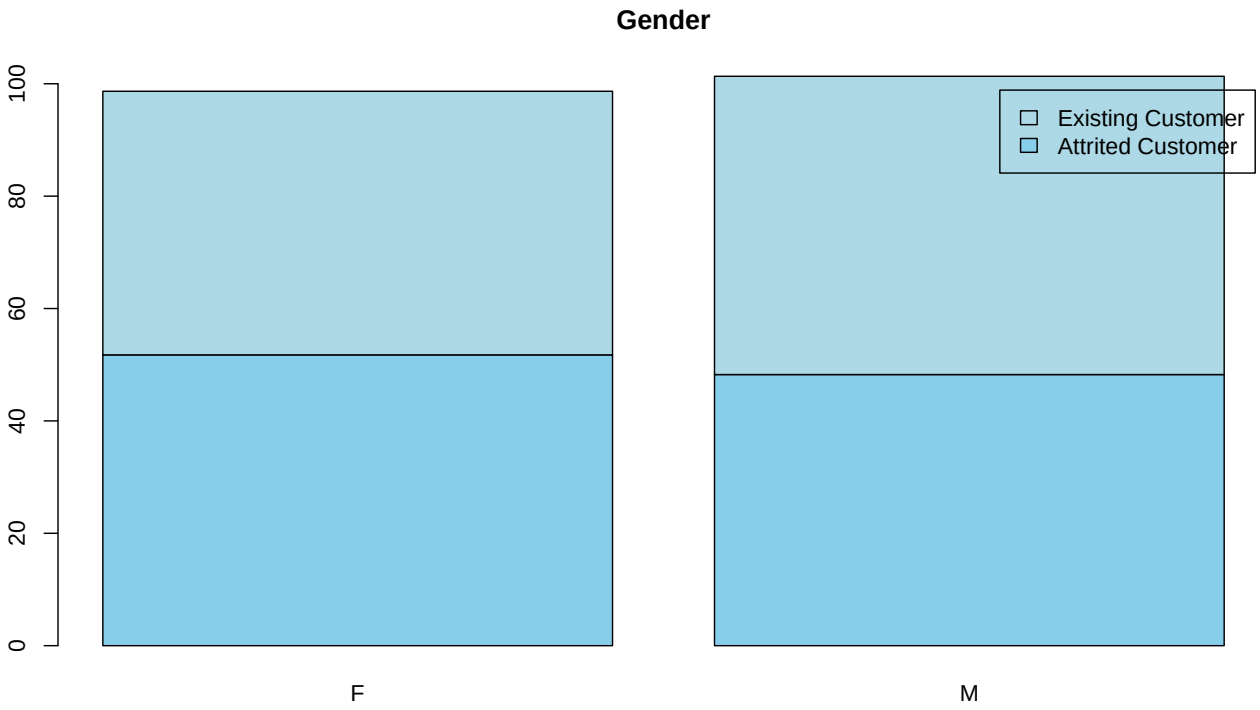

Marital_Status
Existing Customer
Attrited Customer
0
500
1000
1500
2000
2500
3000
3500
Divorced
Married
Single

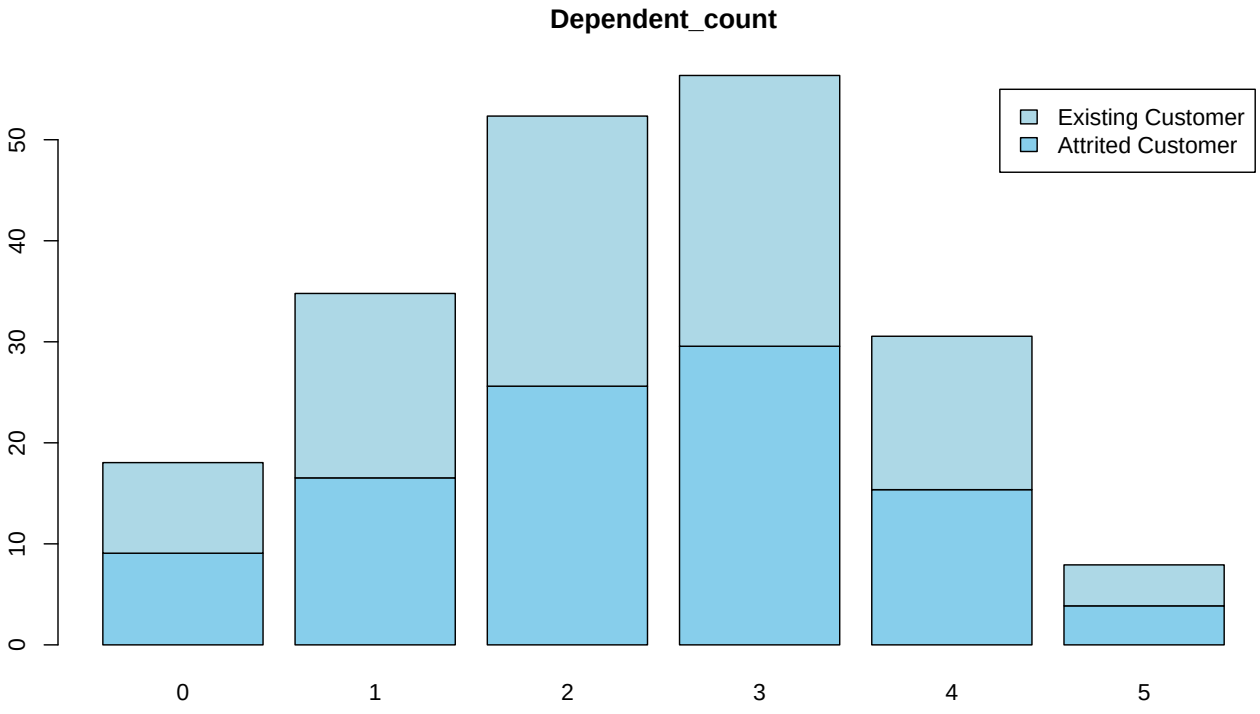


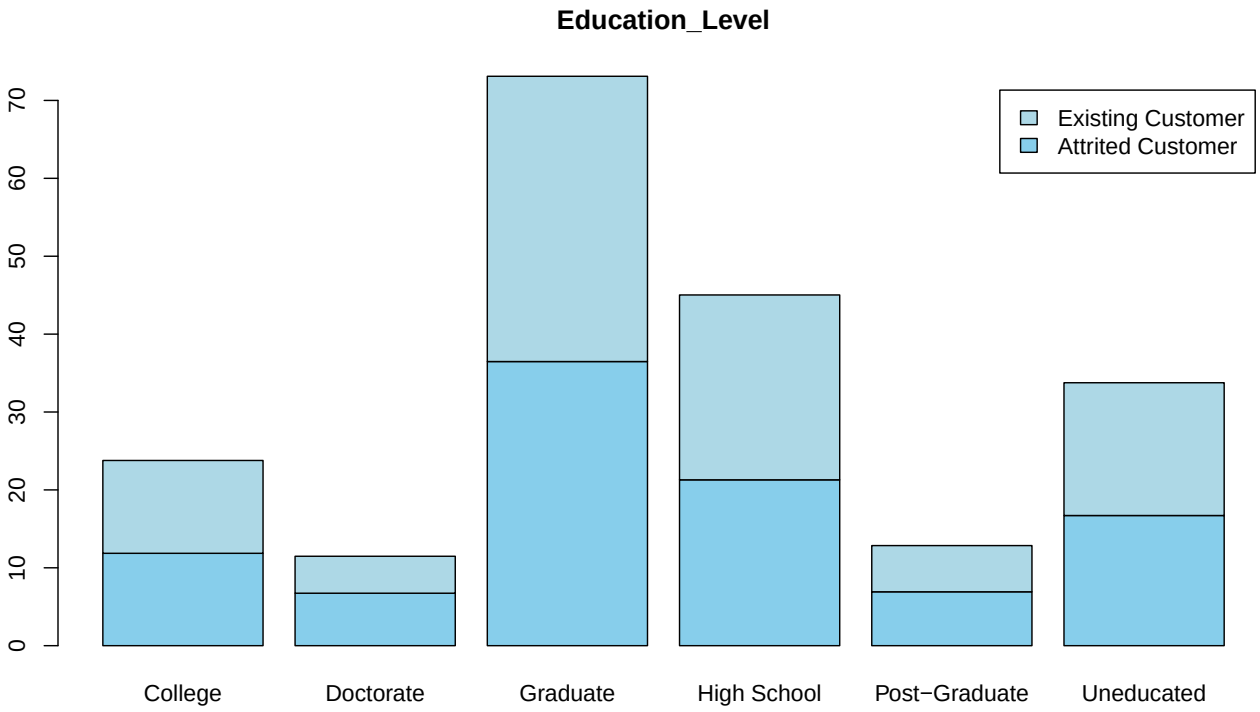

Les graphiques ci-dessus présentent la distribution des clients résiliant leurs services de cartes de crédit en fonction de différentes variables démographiques.

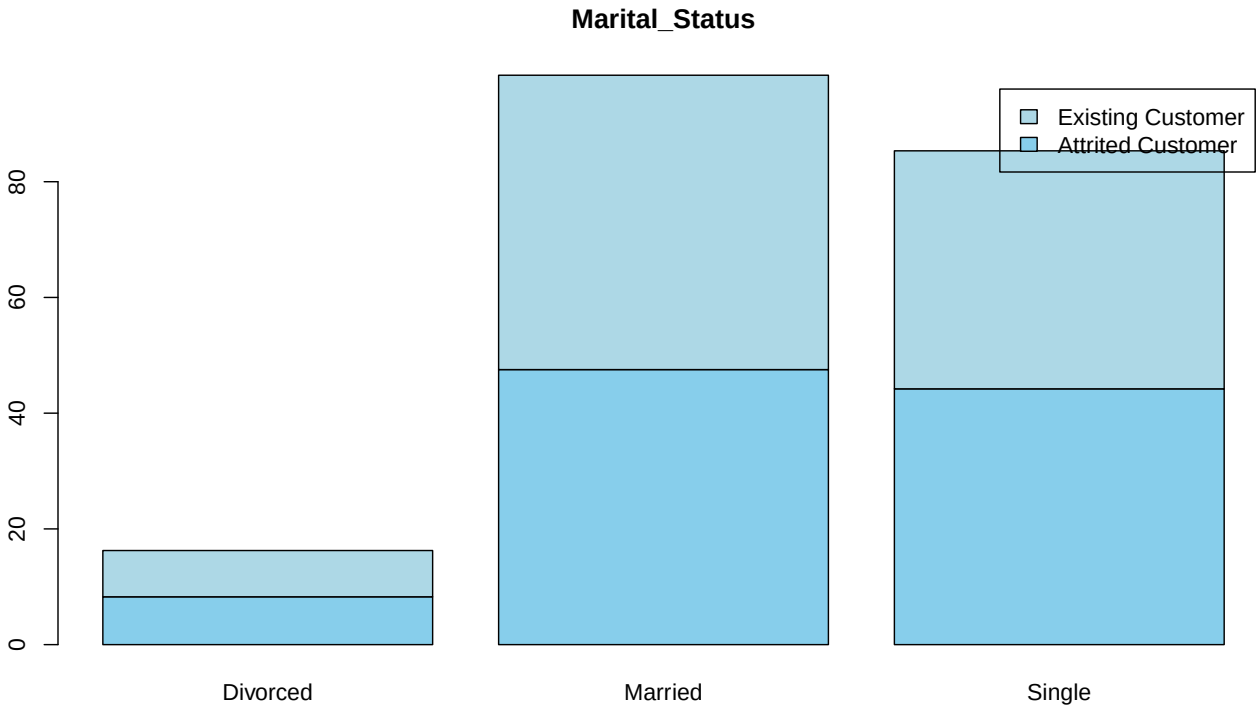
**En pourcentage**

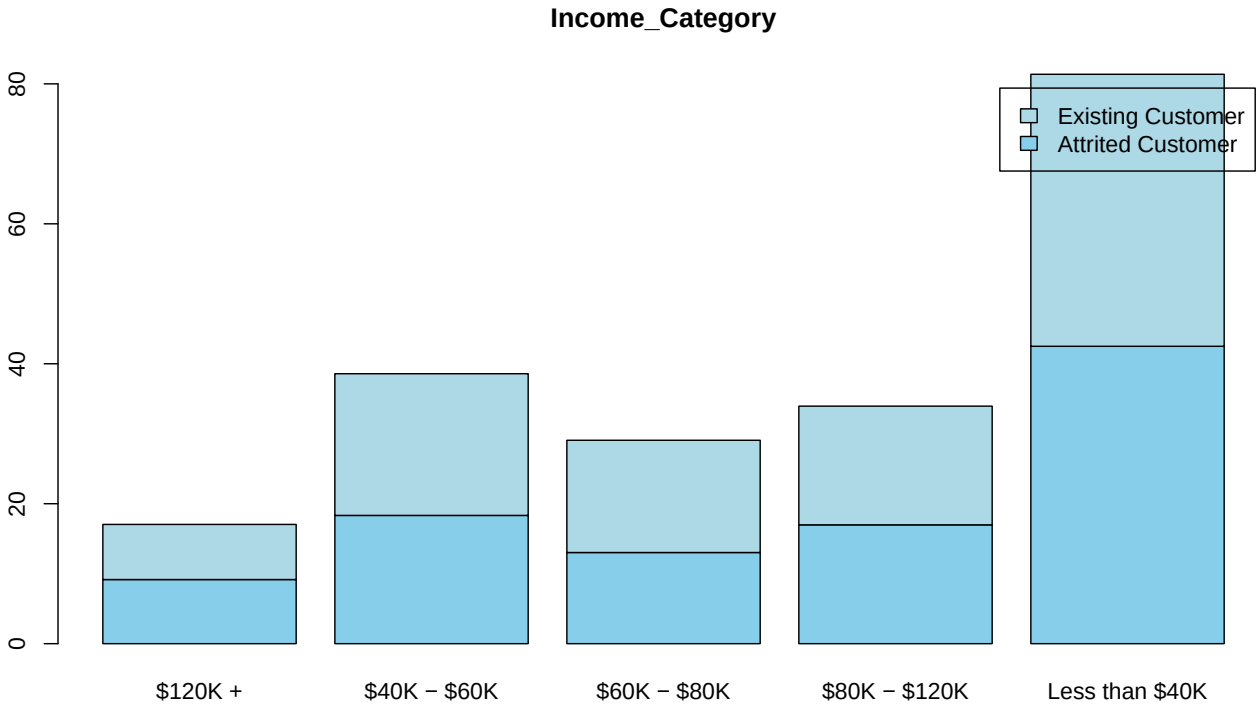



Gender
Existing Customer
Attrited Customer
0
20
40
60
80
100
F
M


Dependent_count
Existing Customer
Attrited Customer
0
10
20
30
40
50
0
1
2
3
4
5


Education_Level
Existing Customer
Attrited Customer
0
10
20
30
40
50
60
70
College
Doctorate
Graduate
High School
Post−Graduate
Uneducated


Marital_Status
Existing Customer
Attrited Customer
0
20
40
60
80
Divorced
Married
Single

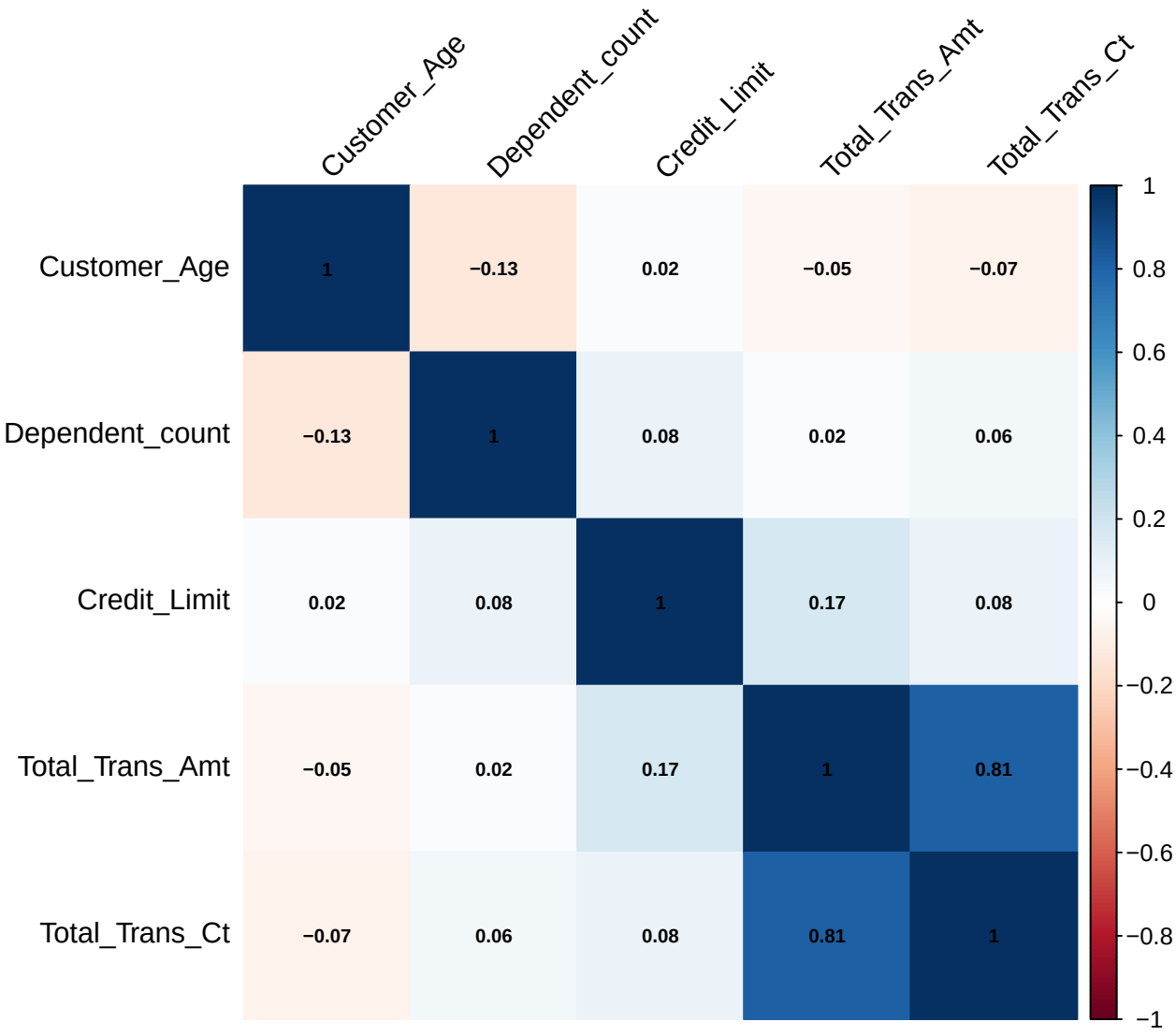



Les graphiques en pourcentage ci-dessus offrent une perspective plus détaillée, mettant en évidence la répartition relative des clients résiliant leurs services de cartes de crédit selon chaque variable démographique.

## 2 - Étude des relations entre les variables

### Matrice corrélation

La matrice de corrélation, générée à partir des variables démographiques et transactionnelles sélectionnées (Customer_Age, Dependent_count, Credit_Limit, Total_Trans_Amt, et Total_Trans_Ct), est présentée ci-dessous. Cette matrice met en évidence les relations linéaires entre ces variables, offrant ainsi un aperçu visuel des liens potentiels dans l'ensemble de données.



**Test corrélation**

Les tests de corrélation de Pearson ont été effectués pour évaluer la significativité des relations entre les paires de variables. Les résultats de ces tests, y compris les valeurs de corrélation et les p-values, sont résumés ci-dessous.

```
## Corrélation entre Customer_Age et Dependent_count :
## P-value = 2.98963e-27
##
## Corrélation entre Customer_Age et Credit_Limit :
## P-value = 0.03722771
##
## Corrélation entre Customer_Age et Total_Trans_Amt :
## P-value = 0.0001127818
##
## Corrélation entre Customer_Age et Total_Trans_Ct :
## P-value = 3.981838e-09
##
## Corrélation entre Dependent_count et Credit_Limit :
## P-value = 7.178391e-12
##
## Corrélation entre Dependent_count et Total_Trans_Amt :
## P-value = 0.05111694
##
## Corrélation entre Dependent_count et Total_Trans_Ct :
## P-value = 1.649984e-06
##
## Corrélation entre Credit_Limit et Total_Trans_Amt :
## P-value = 4.594966e-48
##
## Corrélation entre Credit_Limit et Total_Trans_Ct :
## P-value = 7.134977e-12
```

## 3 - Facteurs sous-jacents à la résiliation

### Nombre de facteurs

```
## [1] 1.2453792 1.0196309 0.9698897 0.7651002
```

Les valeurs propres des facteurs ont été évaluées pour déterminer le nombre de facteurs à retenir lors de l'analyse factorielle. Les facteurs retenus, MR1 et MR2, sont basés sur le critère de Kaiser-Guttman, qui conserve les facteurs dont les valeurs propres sont supérieures à 1.

### Facteurs

```
## Factor Analysis using method =  minres
## Call: fa(r = selected_data, nfactors = nb_facteurs, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                   MR1  MR2     h2    u2 com
## Customer_Age    -0.05 0.03 0.0041 0.996 1.7
## Credit_Limit     0.05 0.81 0.6519 0.348 1.0
## Total_Trans_Amt  0.95 0.15 0.9229 0.077 1.0
## Attrition_Flag   0.17 0.01 0.0282 0.972 1.0
##
##                        MR1  MR2
## SS loadings           0.93 0.67
## Proportion Var        0.23 0.17
## Cumulative Var        0.23 0.40
## Proportion Explained  0.58 0.42
## Cumulative Proportion 0.58 1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 2 factors are sufficient.
##
## df null model =  6  with the objective function =  0.06 with Chi Square =  420.59
## df of  the model are -1  and the objective function was  0
##
## The root mean square of the residuals (RMSR) is  0
## The df corrected root mean square of the residuals is  NA
##
## The harmonic n.obs is  7079 with the empirical chi square  0  with prob <  NA
## The total n.obs was  7079  with Likelihood Chi Square =  0  with prob <  NA
##
## Tucker Lewis Index of factoring reliability =  1.014
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                MR1  MR2
## Correlation of (regression) scores with factors   0.96 0.81
## Multiple R square of scores with factors          0.91 0.65
## Minimum correlation of possible factor scores     0.83 0.30
```

Les résultats de l'analyse factorielle, y compris les charges factorielles, sont présentés dans la table ci-dessus. Chaque score factoriel représente la contribution de chaque observation (client) aux deux facteurs extraits (MR1 et MR2).

## 4 - Modélisation prédictive

### Transformation de la variable cible

La variable cible `Attrition_Flag` a été transformée en une variable binaire dans le cadre du processus de modélisation prédictive. Cette transformation attribue la valeur 1 aux clients ayant résilié leurs services de cartes de crédit et la valeur 0 aux clients non résiliés. Une vérification des valeurs uniques de cette variable transformée est présentée dans le résultat suivant :

```
## [1] "Existing Customer" "Attrited Customer"
```

### Modèle de régression logistique

Le modèle de régression logistique a été élaboré pour estimer la probabilité de résiliation des services de cartes de crédit en fonction des variables explicatives sélectionnées. La variable `Attrition_Flag` a été transformée en une variable binaire, comme expliqué précédemment.

Le résultat ci-dessous présente la transformation de la variable cible et la construction du modèle de régression logistique :

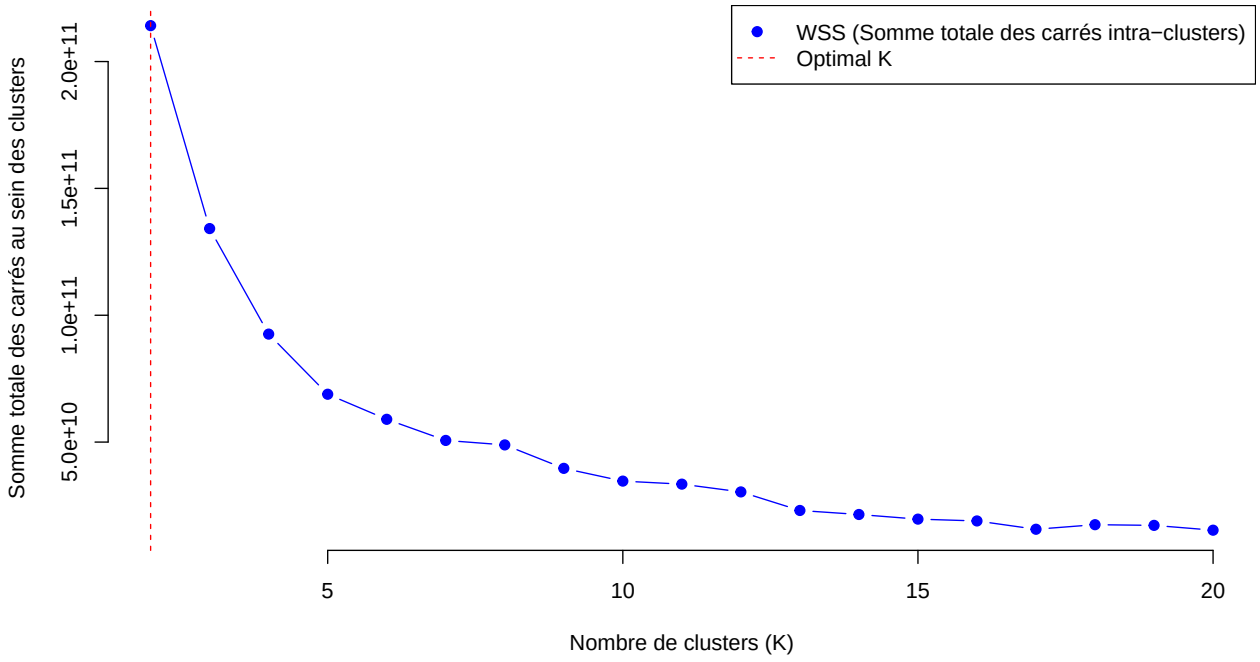
```
##
## Call:
## glm(formula = Attrition_Flag ~ Customer_Age + Dependent_count +
##     Credit_Limit + Total_Trans_Amt + Total_Trans_Ct, family = "binomial",
##     data = tab)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.360e+00  2.592e-01   9.104  < 2e-16 ***
## Customer_Age    -9.097e-03  4.462e-03  -2.039 0.041458 *
## Dependent_count  1.108e-01  2.889e-02   3.835 0.000125 ***
## Credit_Limit    -1.164e-05  4.300e-06  -2.707 0.006790 **
## Total_Trans_Amt  4.004e-04  2.112e-05  18.958  < 2e-16 ***
## Total_Trans_Ct  -9.493e-02  3.295e-03 -28.814  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6159.3  on 7078  degrees of freedom
## Residual deviance: 4804.0  on 7073  degrees of freedom
## AIC: 4816
##
## Number of Fisher Scoring iterations: 6
```

## 5 - Classification des clients

### Nombre optimal de clusters

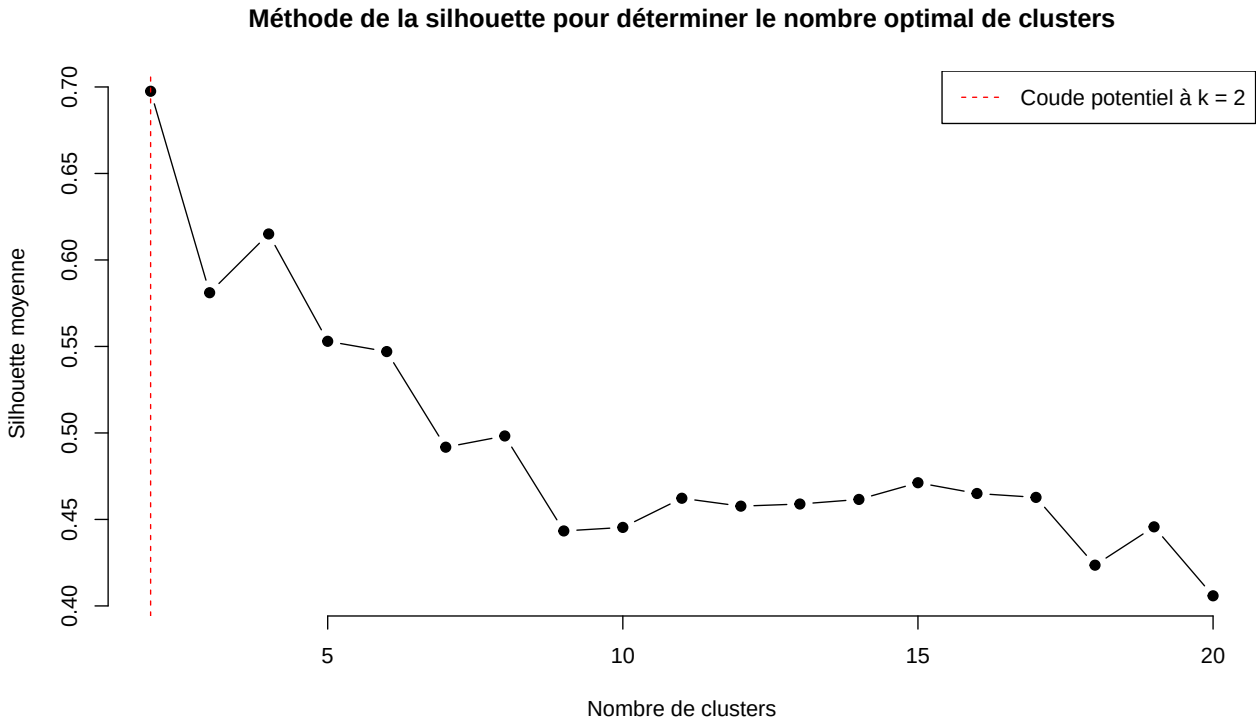
**Méthode du Coude** Le nombre optimal de clusters dans la classification des clients a été déterminé à l'aide de deux méthodes principales. La première, la Méthode du Coude, est illustrée dans le graphique ci-dessous. Cette méthode consiste à examiner la somme totale des carrés au sein des clusters (WSS) pour différents nombres de clusters (K). Le nombre optimal de clusters est souvent identifié au point où l'ajout d'un cluster supplémentaire ne contribue pas significativement à la réduction de la WSS. Dans ce cas, le meilleur nombre de clusters est indiqué par la ligne verticale rouge sur le graphique.

```
## Le meilleur nombre de clusters est: 2
```



**Méthode de la Silhouette** La deuxième méthode, la Méthode de la Silhouette, est une mesure de la qualité de la classification. Elle évalue à quel point chaque objet est similaire à son propre cluster (cohésion) par rapport aux autres clusters (séparation). Le graphique ci-dessous illustre la silhouette moyenne pour différentes valeurs de clusters. Le nombre optimal de clusters est identifié par la ligne verticale rouge à son point culminant.

```
## Le meilleur nombre de clusters est: 2
```



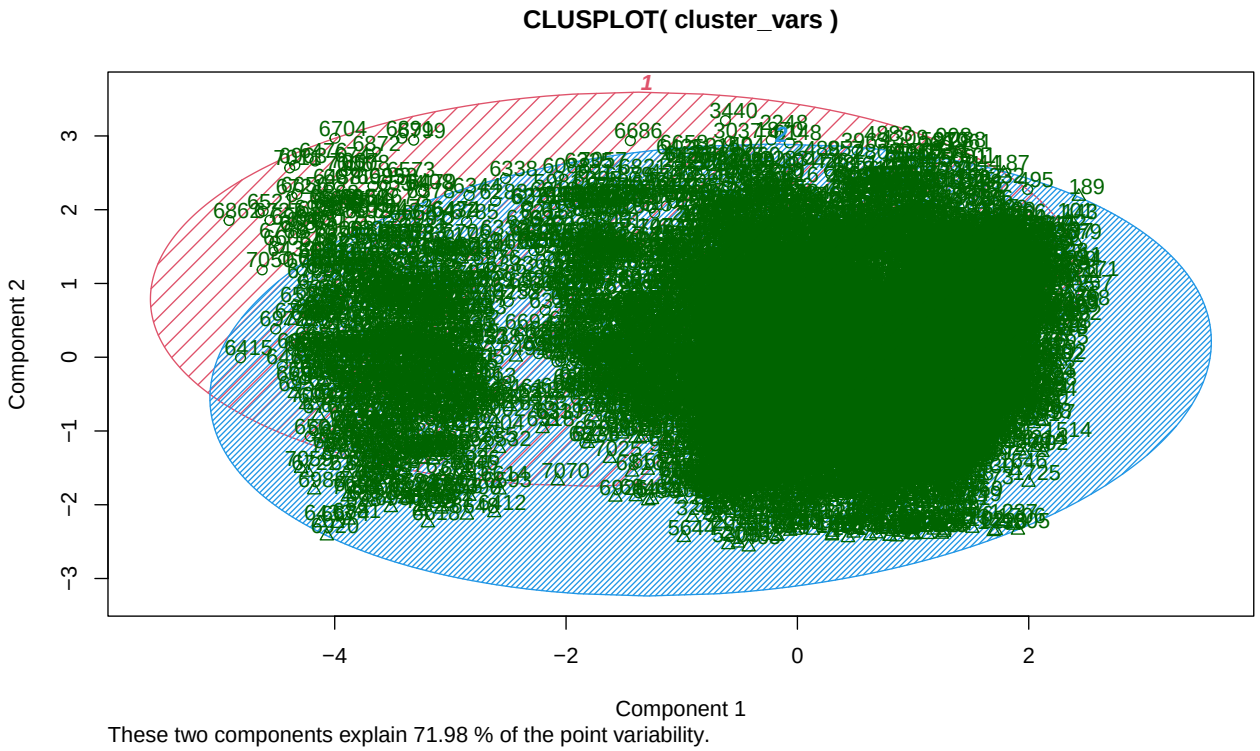
**Résultat de la classification**

Les résultats de la classification, incluant les centres des clusters, la somme totale des carrés (TOTSS = WSS + BSS), la somme totale des carrés intra-clusters (WSS), la somme totale des carrés inter-clusters (BSS), la taille des clusters, le nombre d'itérations, et les éventuelles erreurs, sont présentés dans le tableau ci-dessous.

```
## $centers
##   Customer_Age Credit_Limit Total_Trans_Amt Total_Trans_Ct
## 1     46.80488    26406.119        5512.592       68.80404
## 2     46.25535     4873.744        4169.554       63.65008
##
## $totss
## [1] 674617519143
##
## $withinss
## [1]  81811012784 132342761999
##
## $tot.withinss
## [1] 214153774783
##
## $betweenss
## [1] 460463744360
##
## $size
## [1] 1189 5890
##
## $iter
## [1] 1
##
## $ifault
## [1] 0
```

### Graphiques k-means

**Avec les points de données** Le graphique ci-dessous présente le résultat de la classification en utilisant la méthode k-means avec les points de données. Chaque point est coloré en fonction de son cluster d'appartenance.



**Sans les points de données** Le graphique ci-dessous illustre le résultat de la classification en utilisant la méthode k-means, mais cette fois-ci sans afficher les points de données individuels. Il met en évidence la répartition des clusters.