2025-02-27

Antoine Laverdière-Allaire

HEC MONtreal

Devoir Segmentation

Analyse des Bases de Données en Marketing

Table des matières

[Objectif de la segmentation 2](#_Toc191455066)

[Exploration, nettoyage et traitement des données 2](#_Toc191455067)

[Members 2](#_Toc191455068)

[Reward 5](#_Toc191455069)

[Transaction 9](#_Toc191455070)

[Segmentation 14](#_Toc191455071)

[ACP 20](#_Toc191455072)

[Interprétation 21](#_Toc191455073)

[Annexe 25](#_Toc191455074)

## Objectif de la segmentation

L’objectif principal de cette segmentation est d’analyser l’engagement des membres du programme de fidélisation Air Miles afin de mieux comprendre leur relation avec le programme. Cette analyse permettra d’identifier les différents comportements d’achat, les interactions avec les récompenses et l’évolution de leur engagement au fil du temps. L’enjeu est de développer des stratégies marketing ciblées, notamment en personnalisant les offres, en réactivant les membres dont l’engagement a diminué, et en maximisant la valeur à long terme du programme pour l’entreprise.

Plutôt que d’adopter une segmentation trop générale, cette analyse se concentrera spécifiquement sur la manière dont les membres gagnent et utilisent leurs points Air Miles. En analysant comment et dans quelles circonstances les points sont accumulés et échangés, nous pourrons :

* Évaluer le niveau d’engagement global des membres envers le programme.
* Identifier des profils de comportements distincts, comme les membres qui épargnent leurs points sur le long terme versus ceux qui les utilisent rapidement.
* Comprendre leurs habitudes de dépenses et leurs préférences en matière de récompenses.

Cette segmentation fournira des insights stratégiques permettant à Loyalty d’optimiser la rentabilité du programme, en adaptant les récompenses et les offres aux préférences des membres, en développant des campagnes de fidélisation plus efficaces et en maximisant l'utilisation du programme pour augmenter l’engagement et la satisfaction des membres.

## Exploration, nettoyage et traitement des données

La première étape de cette analyse d’apprentissage non-supervisé consiste à approfondir l’exploration, le nettoyage et le traitement de la base de données afin d’acquérir une meilleure compréhension des variables et de leur structure. Cette étape est essentielle pour identifier d’éventuelles incohérences, valeurs manquantes ou extrêmes, et garantir la qualité des données avant d’entamer l’analyse de segmentation.

Étant donné qu’il y a trois bases de données distinctes à analyser, j’ai décidé de les explorer et de les traiter séparément afin d’assurer la cohérence et la qualité des données. Une fois cette étape complétée, je procéderai à leur fusion en une base de données finale, qui servira de fondation pour la segmentation. Cette approche me permet d’optimiser le nettoyage et la transformation des données avant d’entamer l’analyse. Débutons donc avec la base de données Members.

### Members

La base de données Members contient des informations essentielles sur les membres du programme Air Miles. Elle regroupe à la fois des variables socio-démographiques telles que l’âge, le genre et la province de résidence, ainsi que des données liées à l’historique d’adhésion et d’engagement dans le programme comme Tenure\_Months (Nombre de mois d’ancienneté).

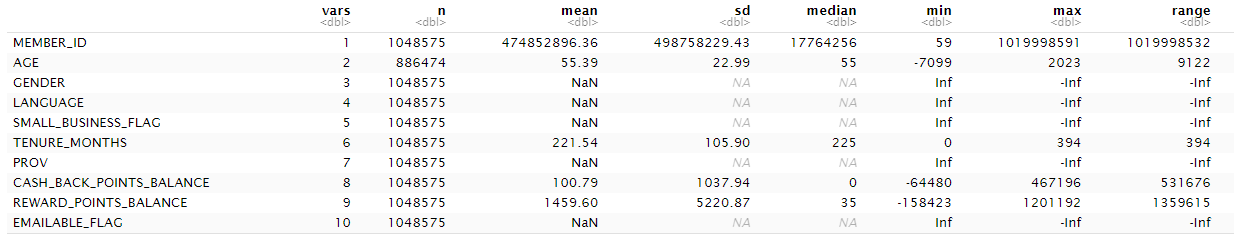
Avant d’entamer le traitement des variables, il est important de définir leur rôle dans l’analyse de segmentation. Étant donné que ces variables sont principalement descriptives et non comportementales, j’ai décidé de les utiliser pour caractériser et interpréter les segments obtenus, plutôt que de les inclure directement dans l’algorithme de segmentation. Ces informations permettront d’analyser le profil des segments, mais ne seront pas des critères déterminants dans la construction des clusters. Cette approche garantira une segmentation axée sur les comportements d’achat et d’utilisation des points, tout en permettant une analyse détaillée des profils des membres une fois la segmentation réalisée. Passons au nettoyage et traitement des variables.

En examinant la structure du jeu de données, j’ai constaté que plusieurs variables, telles que Gender, Prov, Emailable\_Flag et Mailable\_Flag, sont actuellement stockées sous forme de chaînes de caractères (character). Or, ces variables ne peuvent prendre que quelques valeurs distinctes, ce qui signifie qu’elles devraient plutôt être converties en facteurs (factor) pour mieux refléter leur nature catégorielle.

A screenshot of a computer

AI-generated content may be incorrect.

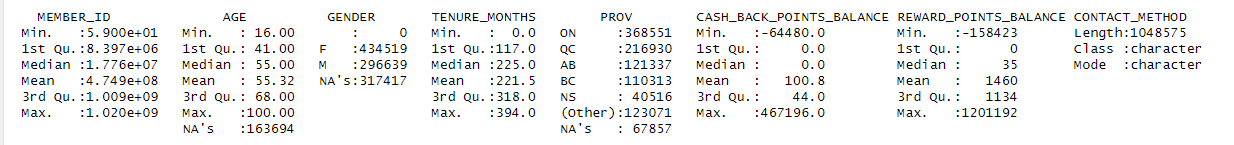
J’ai poursuivi l’exploration en regardant les statistiques descriptives.



Ma première observation est provenue de la variable Age. En analysant cette variable, j’ai identifié plusieurs valeurs aberrantes, notamment des âges négatifs ou extrêmement élevés, qui ne sont pas cohérents avec le contexte de notre étude. Pour garantir la qualité et la fiabilité des données, j’ai décidé de filtrer cette variable en conservant uniquement les membres âgés de 16 à 100 ans. 16 ans est l’âge minimum requis pour adhérer au programme Air Miles ce qui rend l’inclusion de membres plus jeunes incohérente. J’ai aussi fixé le seuil maximal à 100 ans afin d’exclure des valeurs extrêmes, car il est très rare que des individus dépassent cet âge, ce qui réduit le risque d'erreurs dans l'analyse.

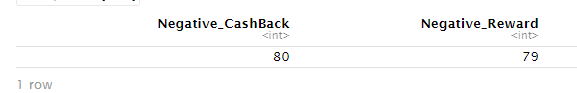
Ensuite, j’ai décidé de supprimer les variables Small\_Business\_Flag et Language, car elles n’apportent pas de grande valeur ajoutée à l’analyse. Small\_Business\_Flag n’est pas intéressante car elle contient uniquement la valeur **"**N" pour tous les membres, ce qui signifie qu’elle n’apporte aucune variation exploitable. Quant à elle, Language est fortement corrélée à la variable Prov, ce qui la rend redondante et peu pertinente pour expliquer les comportements des membres dans le programme.

Par la suite, je trouve que les variables Emailable Flag et Mailable Flag ne fournissent pas suffisamment d’informations pertinentes lorsqu’elles sont analysées séparément. Ainsi, afin d’optimiser leur utilisation, j’ai décidé de les fusionner en une seule variable appelée Contact\_Method. Cette nouvelle variable permettra de regrouper et identifier plus clairement les préférences de contact des membres, en affichant directement les méthodes disponibles sous forme textuelle, en fonction des réponses à ces deux indicateurs. Évidemment, après la création de Contact\_Method, j’ai retiré Emailable\_Flag et Mailable\_Flag de Members.



Voici maintenant à quoi ressemble le jeu de données. Je remarque qu’il y a plusieurs valeurs manquantes (NA) dans les colonnes Age et Gender. Il faut se rappeler que les variables de la table Members ne seront pas utilisées directement pour construire la segmentation. Leur rôle est plutôt d’aider à décrire et interpréter les segments obtenus une fois l’analyse terminée. Ainsi, dans ce contexte, la présence de valeurs manquantes dans Age et Gender n’aura pas d’impact significatif sur la qualité de la segmentation. Il n’est donc pas nécessaire de les imputer ou de les modifier à ce stade, et on conserve les valeurs manquantes telles quelles pour l’instant.

J’ai également constaté que les variables Cash\_Back\_Points\_Balance et Reward\_Points\_Balance contiennent quelques valeurs négatives. Après une analyse des valeurs minimales de ces colonnes, il apparaît que seulement 80 observations sont concernées par variable. Ces valeurs négatives pourraient être dues à des remboursements effectués par les membres ou à d’éventuelles erreurs de saisie.



Étant donné que l’objectif d’analyse ne porte pas sur les comportements de remboursement, et compte tenu du faible nombre de cas affectés, il est plus judicieux de supprimer ces lignes.

Finalement, j’ai observé que la variable Tenure\_Months présente une valeur minimale de zéro. Après vérification, il s’avère qu’il s’agit de membres récemment inscrits au programme Air Miles. Comme ces nouveaux membres représentent une proportion relativement faible de l’ensemble des données, leur présence ne risque pas d’altérer significativement les résultats de la segmentation. Par conséquent, j’ai décidé de les conserver dans le jeu de données, afin d’inclure toutes les typologies de membres et d’assurer une segmentation représentative de l’ensemble du programme.

Une fois le nettoyage terminé, voici l’état final de la table Members :

A white background with black text

AI-generated content may be incorrect.

On va la laisser de côté pour l’instant et y revenir plus tard lors de l’étape d’analyse de segmentation. Passons maintenant à l'exploration de la table Reward.

### Reward

Cette table contient les informations liées à l’utilisation des points Air Miles avec des variables précises comme Points\_Redeemed et Redemptions.

Je vais encore débuter l’étape de nettoyage en analysant la structure du dataframe.

A white background with black text

AI-generated content may be incorrect.

On constate que Rewards\_Category est enregistrée sous forme de caractère, alors qu’elle devrait être un facteur, étant donné qu’elle comporte un nombre limité de catégories distinctes. Je vais donc la convertir en facteur et afficher ses différentes modalités pour mieux comprendre sa distribution



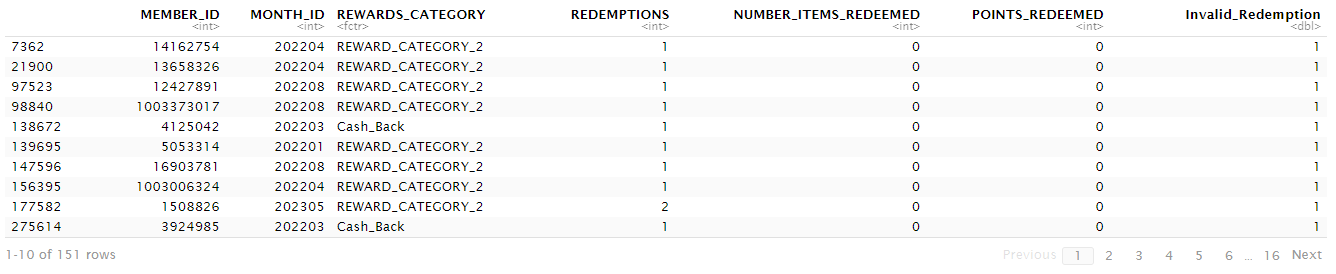
On voit que les catégories de récompenses se répartissent en cinq types distincts. Cependant, pour cette analyse, je vais principalement me concentrer sur la distinction entre les récompenses de type Cash Back et Reward, sans m’attarder sur les sous-catégories spécifiques des Rewards. Ainsi, lors de la création des nouvelles variables, l’analyse se fera exclusivement à travers ces deux grandes catégories (Cash Back vs. Reward) afin de simplifier l’interprétation des résultats et d’optimiser la pertinence de la segmentation.

On poursuit encore avec l’analyse des statistiques descriptives.

A screenshot of a computer

AI-generated content may be incorrect.

J’observe que la variable Number\_Items\_Redeemed contient des valeurs égales à zéro, alors que la variable Redemptions indique toujours qu’au moins une transaction a été effectuée (min = 1). Cette situation peut s’expliquer par soit des erreurs lors du processus de transaction ou des types d’échanges particuliers ne correspondant pas directement à l’obtention d’un article. De plus, aucun point n’a été gagné dans ces cas, ce qui soulève des questions quant à leur pertinence pour l’analyse. Avant de décider comment traiter ces observations, je vais d’abord évaluer leur fréquence dans le jeu de données afin de mesurer leur impact sur la segmentation.



Il y a 151 lignes qui sont concernées, soit un nombre relativement faible d’observations. Étant donné que l’objectif d’affaires est d’analyser l’accumulation et l’utilisation des points Air Miles, ces transactions ne sont pas pertinentes, car aucun point n’y est gagné. Ainsi, pour garantir la cohérence de l’analyse et éviter d’inclure des cas atypiques qui pourraient fausser les résultats, je vais supprimer les lignes du jeu de données lorsque Number\_Items\_Redeemed est égal à 0.

Ensuite, afin d’évaluer l’impact temporel des récompenses et de faciliter la création de nouvelles variables, je vais convertir la colonne Month\_ID en une nouvelle variable YEAR, représentant l’année de la transaction. YEAR sera ensuite converti en facteur.

Pourquoi choisir une granularité annuelle ?

1️. Une vision globale du programme

* Cette segmentation vise à offrir une analyse stratégique du programme Air Miles.
* Un niveau de détail trop précis (mensuel ou trimestriel) n'est pas nécessaire pour définir les segments et pourrait compliquer l’interprétation des résultats.

2️. Réduction de la complexité et optimisation des analyses

* Travailler avec des données mensuelles ou trimestrielles impliquerait la création d’un grand nombre de variables supplémentaires, rendant la segmentation plus complexe.
* Une approche annuelle permet de mieux identifier les tendances globales tout en garantissant une segmentation plus exploitable.

Je vais ainsi regrouper les transactions effectuées et les récompenses réclamées par année (2022, 2023 et 2024) afin de simplifier l’analyse et garantir une segmentation pertinente et efficace.

J’ai aussi identifié la présence de plusieurs valeurs extrêmes dans certaines variables du jeu de données. Plutôt que de les traiter immédiatement, je vais garder ce constat en tête et y revenir plus tard dans l’analyse, une fois que j’aurais créé le dataset final. Cette approche me permettra de prendre des décisions éclairées, en évaluant si ces valeurs influencent réellement la structure des segments ou si elles doivent être ajustées pour garantir une segmentation plus robuste et représentative.

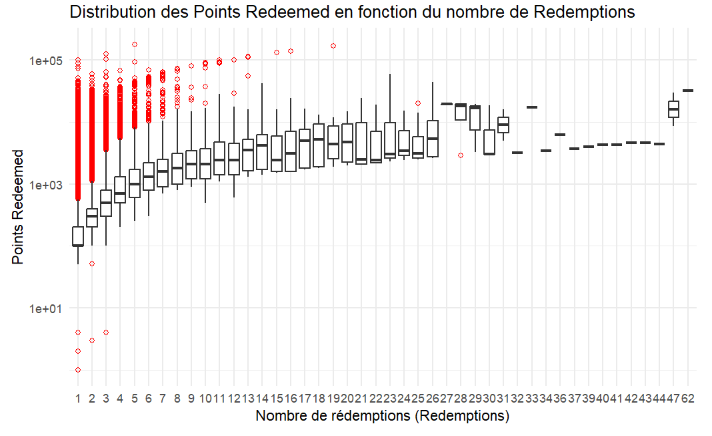
#### Création de variables

Je vais maintenant passer à la création de nouvelles variables.

Un indicateur clé de cette analyse est le nombre de points dépensés lors des réclamations de récompenses. Il est donc essentiel d’examiner le total des points réclamés par chaque membre, en différenciant :

* Les catégories de récompenses (CashBack vs Reward)
* L’année de la transaction (2022, 2023, 2024)

Par ailleurs, il serait également pertinent d’examiner le nombre de points dépensés par transaction (Points\_Per\_Redemption) afin d’identifier les petits consommateurs (utilisant leurs points de manière progressive) et les gros utilisateurs (dépensant beaucoup de points par réclamation).



Par exemple, en analysant ce boxplot représentant la distribution de Points\_Redeemed en fonction de Redemptions, plusieurs observations clés émergent :

1. Une relation croissante entre les deux variables : Plus le nombre de rédemptions (Redemptions) augmente, plus la quantité de points utilisés (Points\_Redeemed) a tendance à être élevée.
2. Une dispersion importante : La répartition des points réclamés est très large, ce qui indique une grande variabilité des comportements parmi les membres. Certains utilisent peu de points par transaction, tandis que d’autres en dépensent des montants considérables.
3. Présence de nombreux outliers : Plusieurs transactions affichent des valeurs extrêmement élevées, ce qui pourrait suggérer des cas particuliers nécessitant une analyse plus approfondie

Pour affiner l’analyse du nombre de points dépensés lors de réclamations de récompenses et obtenir une vision détaillé du comportement des membres, je vais donc créer les variables suivantes :

🔹 Total des points réclamés, répartis par catégorie de récompense (CashBack et Rewards) et par année (2022, 2023, 2024).

🔹 Nombre moyen de points utilisés par réclamation, afin d’identifier ceux qui dépensent beaucoup de points d’un coup vs ceux qui les utilisent progressivement. Cela servira à faire une distinction des profils de consommateurs.

Cependant, je veille également à ne pas générer un excès de variables dans la table Reward, afin d’éviter la multicollinéarité dans le jeu de données final. Un trop grand nombre de variables fortement corrélées pourrait fausser l’analyse et compliquer l’interprétation des segments.

#### Agrégation

Maintenant que les nouvelles variables ont été ajoutées à Reward, il est temps d’agréger la base de données afin de n’avoir qu’une seule ligne par membre (Member\_ID). Cette agrégation est essentielle pour obtenir une vue d’ensemble plus globale et détaillée du comportement de chaque membre vis-à-vis du programme de fidélité. Elle permet aussi de faciliter l’intégration des trois bases de données et de simplifier l’analyse et la segmentation.

L’agrégation des données aboutira à la création d’une nouvelle table nommée Reward\_agg. Lors de ce processus, j’ai appliqué différentes stratégies d’agrégation en fonction de la nature des variables pour garantit que l’ensemble des informations pertinentes est correctement consolidé

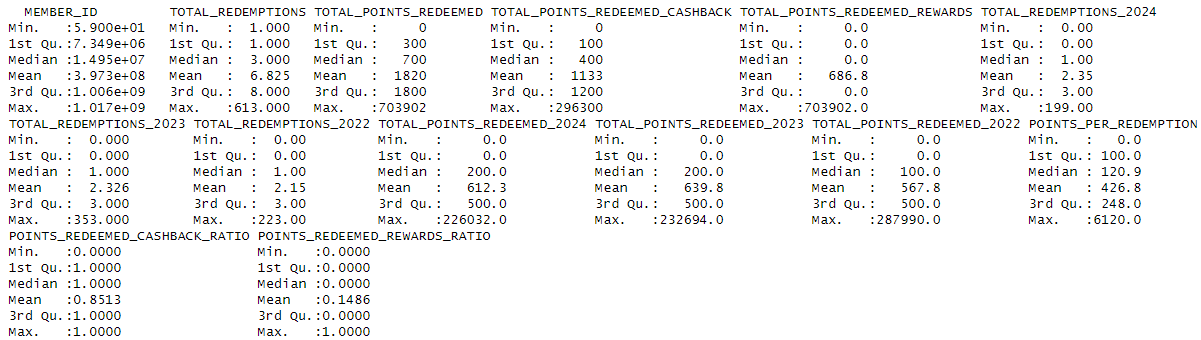
* Somme (sum()) pour les variables cumulatives, telles que Points\_Redeemed et Redemptions, afin d’obtenir le total de ces variables pour chaque membre.
* Maximum (max()) pour les nouvelles variables créées, qui conservent la même valeur pour chaque membre, afin de s’assurer que les bonnes données sont bien prises en compte.

Par la suite, je vais évaluer la corrélation entre les variables en créant une matrice de corrélation. (Voir Annexe 1)

L’analyse des corrélations révèle que Total\_Number\_Items\_Redeemed est fortement corrélé à Total\_Redemptions (> 0.9), ce qui indique une redondance d’information entre ces deux variables. J’ai donc décidé de supprimer Total\_Number\_Items\_Redeemed de Reward\_agg pour éviter la multicolinéarité et simplifier le modèle.

Finalement, en analysant le sommaire de Reward\_agg, j'ai constaté que la variable Points\_Per\_Redemption présente une valeur maximale anormalement élevée de 89 000 points. Une telle valeur est suspecte, car elle suggère que chaque utilisation de points est extraordinairement élevée, ce qui n'est pas réaliste dans un contexte typique. Pour traiter les valeurs extrêmes de la variable et éviter qu'elles ne faussent la segmentation, j'ai décidé d'appliquer une méthode de plafonnement. Cette technique consiste à remplacer les valeurs dépassant un certain seuil par la valeur de ce seuil, ici fixé au 99ᵉ percentile. Ainsi, les valeurs supérieures à ce percentile seront plafonnées, réduisant l'impact des anomalies tout en conservant l'intégrité des données.

Voici la structure de la table final Reward agg :



On aperçoit d’ailleurs que la valeur maximum de Points\_Per\_Redemption est désormais de 6120 ce qui va grandement aider à limiter l'influence des valeurs extrêmes sur la segmentation. L’approche du plafonnement sera probablement réutilisée plus tard pour traiter les valeurs extrêmes des autres variables du dataset final. Cependant, j’ai choisi de traiter Points\_Per\_Redemption maintenant, car il présentait un cas exceptionnel de valeurs extrêmes.

### Transaction

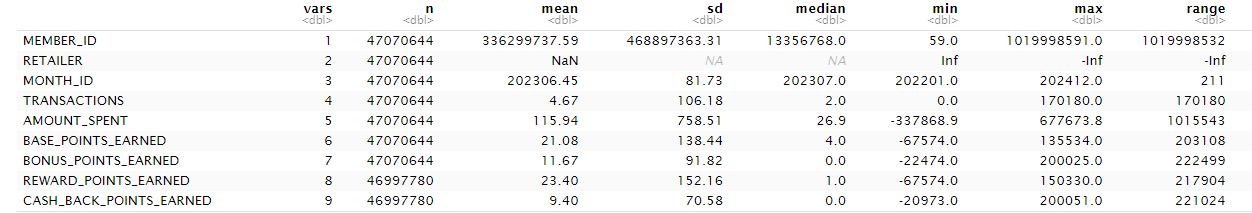
Cette table enregistre les achats et les points gagnés par les membres et permet d’identifier les comportements d’achats avec des variables comme Amount\_Spent et Base\_Points\_Earned.

Comme pour les autres variables, j’ai d’abord vérifié la structure de Transaction. Je me suis assuré que toutes les variables, à l’exception de Retailer et Amount\_Spent sont bien au format integer Retailer sera converti en facteur, puisqu’il s’agit d’une variable catégorielle et Amount\_Spent sera maintenu au format numérique (numeric), afin de permettre des calculs précis lors des analyses.

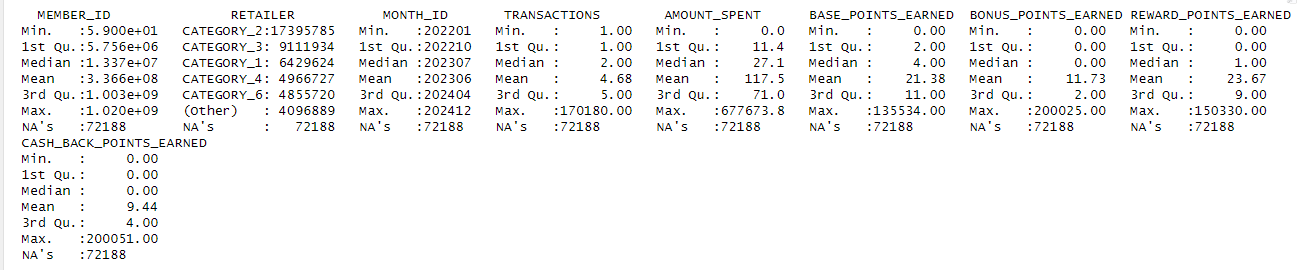
L’objectif étant d’analyser comment les membres ont accumulé leurs points Air Miles, il est essentiel de se concentrer sur le type d’achats effectués. On va donc analyser les différents types de catégories de Retailer.

J’estime que les transactions enregistrées sans catégorie spécifique (« ») ou classées sous Category\_8 ne fournissent pas d’informations exploitables sur les comportements des membres. Elles ne permettent donc pas d’identifier les tendances d’accumulation des points en fonction des types d’achats et on va les supprimer des catégories de Retailer.

Analysons les statistiques descriptives.



Je constate la présence de valeurs négatives dans plusieurs variables. Ces valeurs sont probablement dues à des remboursements effectués aux membres ou à des erreurs. Étant donné que l'objectif principal de cette étude est d'analyser les comportements d'achat positifs et non les retours ou annulations, il est judicieux d'exclure ces valeurs négatives de la segmentation. Cette approche permettra de se concentrer sur les transactions reflétant un engagement actif des membres envers le programme de fidélité.



L’analyse du summary de la base Transaction révèle la présence d’environ 112 000 valeurs manquantes (NA). Il est incertain si ces valeurs manquantes résultent d’erreurs de saisie ou si elles correspondent réellement à des données absentes.

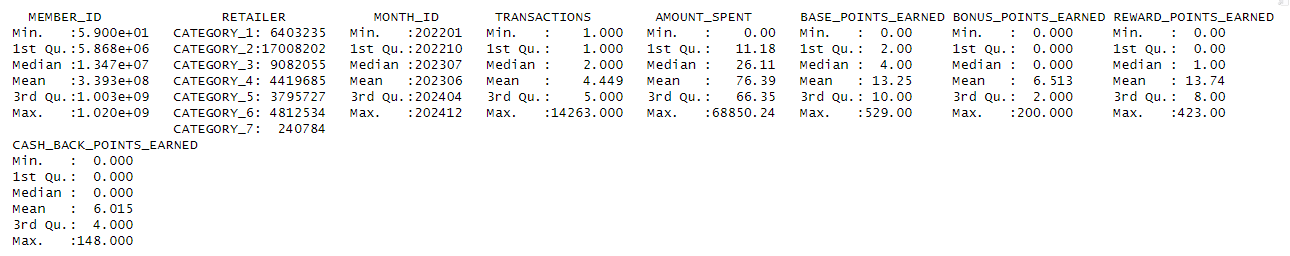
Je juge que ces observations ne sont pas essentielles pour atteindre l’objectif de cette segmentation. Afin de préserver la fiabilité des résultats et éviter d’introduire des biais liés aux données incomplètes, les transactions contenant des valeurs manquantes (NA) seront exclues.

En ce qui a trait aux valeurs extrêmes, l’analyse des données révèle la présence de valeurs anormalement élevées dans plusieurs variables clés comme Amount\_Spent, Bonus\_Points\_Earned, Rewards\_Points\_Earned et Cash\_Back\_Points\_Earned. Ces valeurs, bien que réelles, sont peu représentatives du comportement global des clients et risquent d’introduire des distorsions dans l’analyse. En effet, elles peuvent fausser la moyenne et l’écart-type. On se rappelle que l’objectif de cette segmentation est de cibler les comportements typiques, et non ceux d’un sous-ensemble restreint de clients ayant des habitudes d’accumulation et de dépenses atypiques.

Ainsi, comme effectué sur la table Reward\_agg plus tôt, j’ai appliqué un filtrage de plafonnement sur les valeurs extrêmes en éliminant le 1% des valeurs les plus élevées des variables suivantes :

* Amount\_Spent
* Bonus\_Points\_Earned
* Rewards\_Points\_Earned
* Cash\_Back\_Points\_Earned

Cette approche permet ainsi de garantir une segmentation plus robuste, exploitable et alignée avec les objectifs marketing.



On voit que l’état des valeurs extrêmes est désormais plus raisonnable. \*\*\* Mauvais screenshot

De plus, comme j’ai fait pour la variable Reward, j’ai transformé la colonne Month\_ID en YEAR.

Avant de poursuivre avec la création des nouvelles variables, il est important de bien comprendre le fonctionnement du système de points du programme Air Miles.

Lors de chaque transaction effectuée avec une carte Air Miles, les membres accumulent des Base Points, dont le montant dépend du montant d’argent dépensé et du type de programme auquel ils sont inscrits. Les Base Points sont attribués en fonction de paliers de dépenses chez les partenaires Air Miles et sont automatiquement enregistrés dans le système Loyalty, selon l’option de récompense sélectionnée par le membre.

Chaque membre peut personnaliser l’allocation de ses points Air Miles en choisissant l’une des trois options suivantes :

* Cash Back : Les points accumulés sont convertis en argent remis directement et peuvent être utilisés comme réduction immédiate sur des achats chez certains partenaires.
* Récompenses cadeaux (Rewards) : Les points sont stockés pour être échangés contre des voyages, articles, cartes-cadeaux ou expériences.
* Mix des deux : Les membres peuvent choisir un pourcentage précis pour allouer leurs points entre Cash Back et Dream Rewards, selon leurs préférences.

En plus des Base Points, les membres peuvent aussi gagner des points bonus avec des promotions spéciales chez les partenaires et des offres exclusives entre autres. Ces points bonus se rendent directement dans la balance de l’option de récompense sélectionnée par le membre.

#### Création de variables

Dans cette étape, je me concentre exclusivement sur les mécanismes d’accumulation et d’utilisation des points par les membres, car cela correspond à l’objectif principal du projet. Cette stratégie ne va donc pas prendre en compte la création de variable comme Total\_Transaction\_Category\_1 ou Total\_Transaction\_2024 puisqu’on va se concentrer seulement sur l’accumulation des points. Cela va réduire la corrélation excessive entre les variables, car des variables trop similaires risquent de brouiller les résultats de la segmentation et introduire du bruit statistique.

Dans cette section, j’ai grandement essayé, en vain, d’intégrer des variables de ratios et de moyennes afin d’enrichir la segmentation du dataset final. Cependant, je me suis souvent retrouvé face à deux défis :

🔹 Des valeurs très faibles et peu significatives, qui n’apportaient pas de réelle valeur ajoutée à l’analyse.  
🔹 Une forte corrélation avec d’autres variables, risquant d’introduire de la redondance et de fausser les résultats de la segmentation.

Ainsi, afin de préserver la cohérence avec l’objectif d’affaires, j’ai donc fait le choix de limiter le nombre de nouvelles variables, en me concentrant uniquement sur celles qui apportent une réelle différenciation entre les membres. Cette approche garantit une segmentation plus simple, efficace et directement exploitable.

Concernant les variables que j’ai effectivement créées, mon objectif était de capturer plusieurs dimensions clés du comportement des membres :

* Le volume total de points gagnés, en créant des variables qui reflètent l’accumulation annuelle des points.
* Une mesure de l’efficacité d’accumulation des points, permettant d’évaluer comment chaque membre optimise son programme de fidélité.
* L’analyse des différences de comportement selon les catégories d’achats, afin d’identifier quelles dépenses contribuent le plus à l’accumulation de points et repérer les préférences des membres.

Une relation qui m’a particulièrement intéressé est l’impact des points bonus sur l’accumulation totale de points par les membres. Plus précisément, je souhaite déterminer si les membres qui gagnent davantage de points bonus accumulent plus de points en général. Pour explorer cette relation, je vais créer une nouvelle variable appelée Bonus\_Points\_Earned\_Ratio, qui mesure la proportion de points bonus dans l’ensemble des points gagnés. Cette variable sera calculée comme suit : Bonus\_Points\_Earned / (Base\_Points\_Earned + Bonus\_Points\_Earned). Cette mesure permettra de comparer dans quelle mesure les bonus influencent l’accumulation totale de points, et d’identifier si certains membres optimisent davantage les offres promotionnelles que d’autres.

De la même manière que pour la table Reward, je souhaite analyser l’évolution de l’accumulation des points par les membres au fil des années. Cette analyse me permettra d’identifier les tendances d’engagement des membres dans le programme. Pour ce faire, je vais créer trois nouvelles variables : Points\_Earned\_2022, Points\_Earned\_2023 et Points\_Earned\_2024. Ces variables permettront de comparer l’évolution des points gagnés d’une année à l’autre, afin de repérer les membres dont l’accumulation de points augmente, diminue ou reste stable.

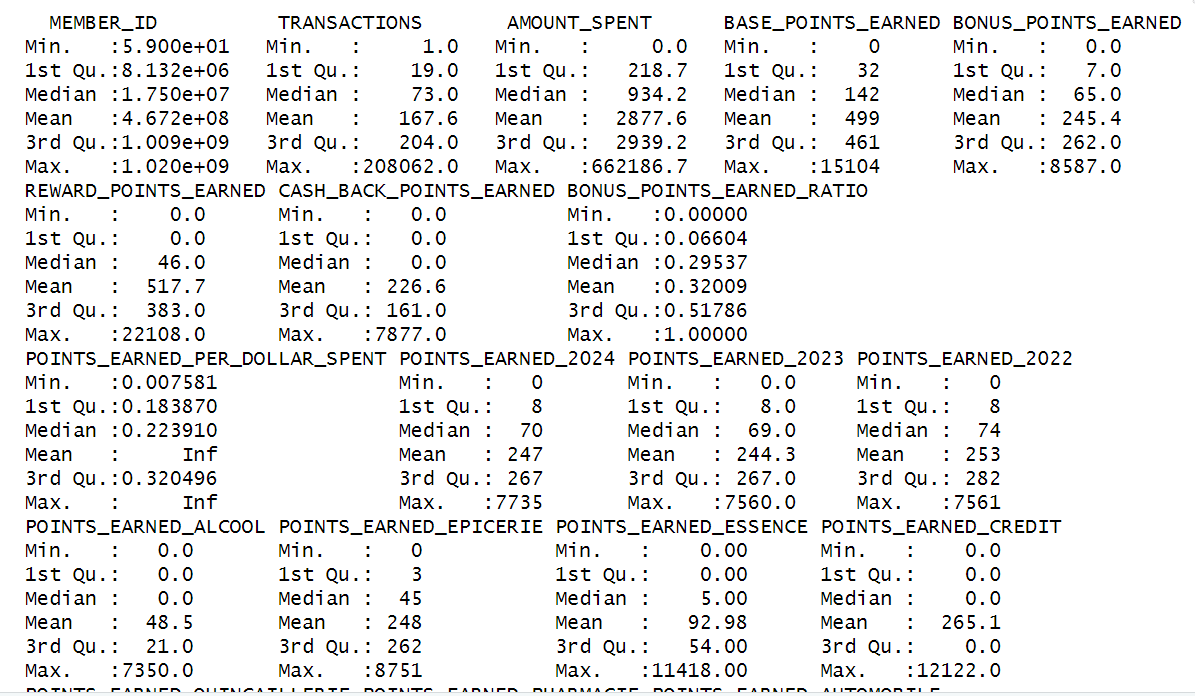
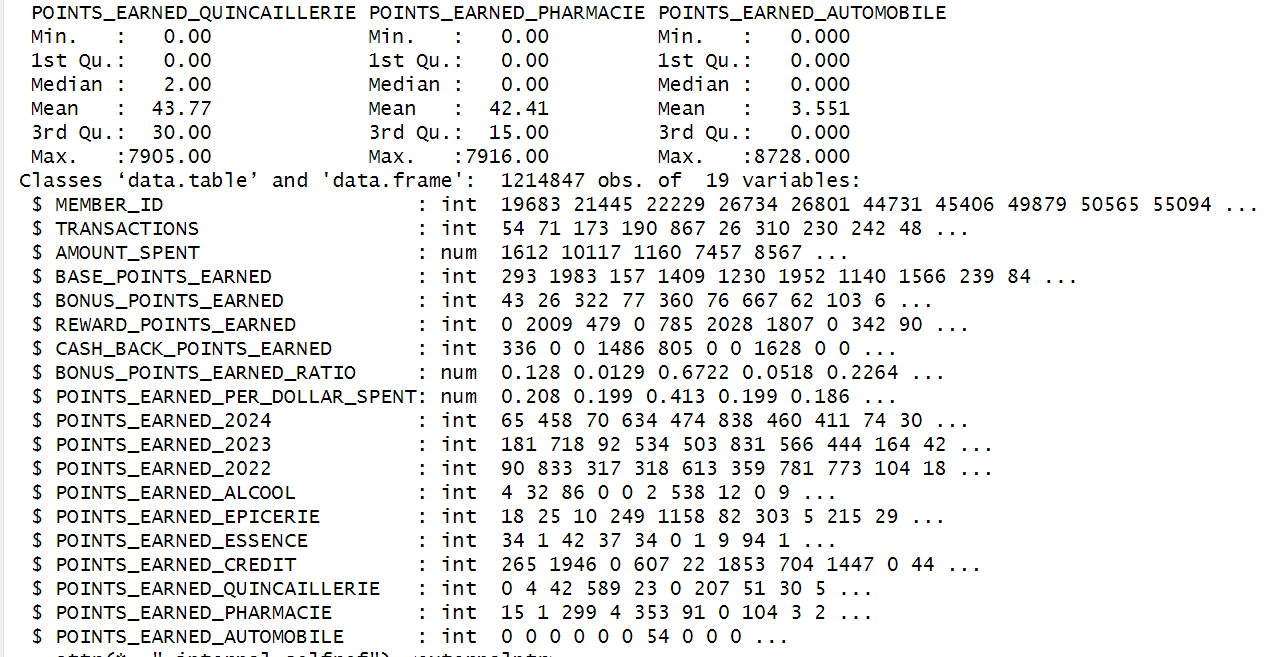
Enfin, dans la même logique, je souhaite analyser le comportement des membres en fonction du type d’achats qu’ils effectuent et l’impact de ces achats sur l’accumulation de points. Pour cela, je vais créer des variables représentant le total de points gagnés par catégorie de Retailer. Afin de rendre l’interprétation plus intuitive, ces variables seront renommées selon leur catégorie spécifique : Points\_Earned\_Alcool pour la Catégorie 1, Points\_Earned\_Epicerie pour la Catégorie 2, etc. Cette segmentation par catégorie permettra de mieux comprendre quels types d’achats contribuent le plus à l’accumulation de points et d’identifier les préférences des membres.

Une fois ces modifications effectuées, je vais supprimer les variables Retailer et Year du jeu de données et veillé à ce que toutes les nouvelles variables soient correctement formatées en integer.

#### Agrégation

On peut maintenant procéder à l’agrégation de la table Transaction, afin d’obtenir une seule ligne par membre (Member\_ID). Comme pour la table Reward, je vais effectuer la somme des variables de base (Amount\_Spent, Base\_Points\_Earned) pour calculer les totaux cumulés par membre, recalculer les ratios (afin d’obtenir des valeurs précises et représentatives après l’agrégation) et sélectionner la valeur maximale pour les variables constantes par membre.

Cette étape permettra d’avoir une vue consolidée du comportement d’accumulation des points des membres et facilitera l’intégration avec les autres tables pour la segmentation.

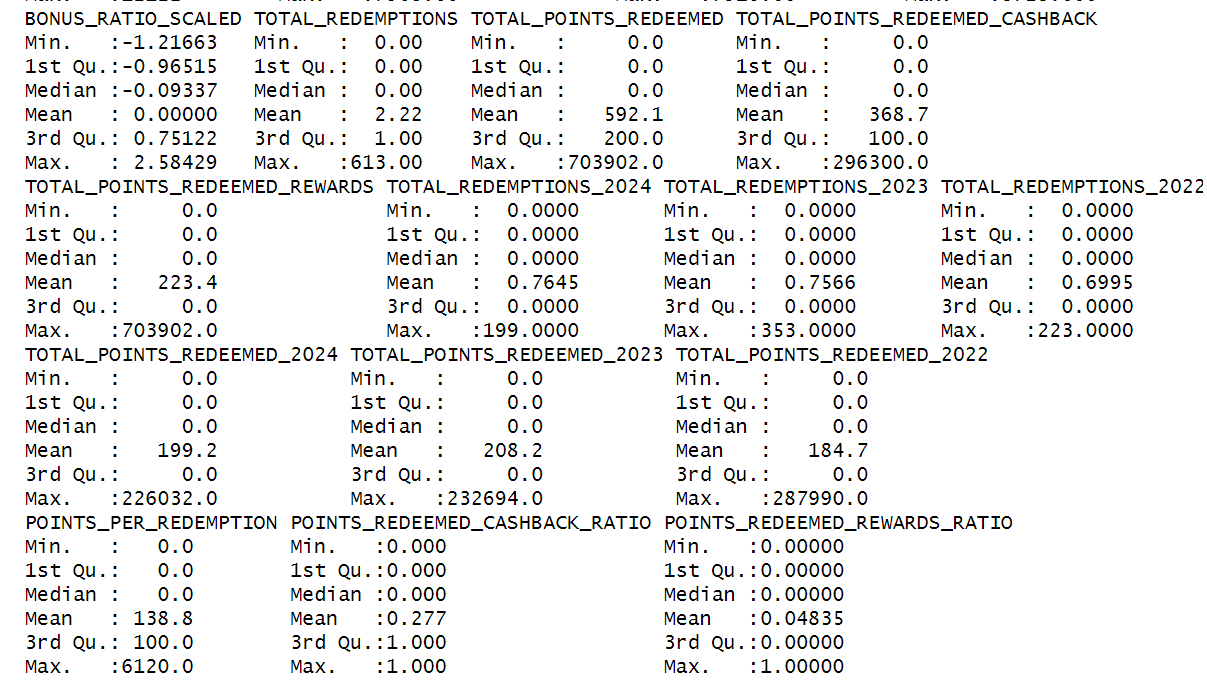
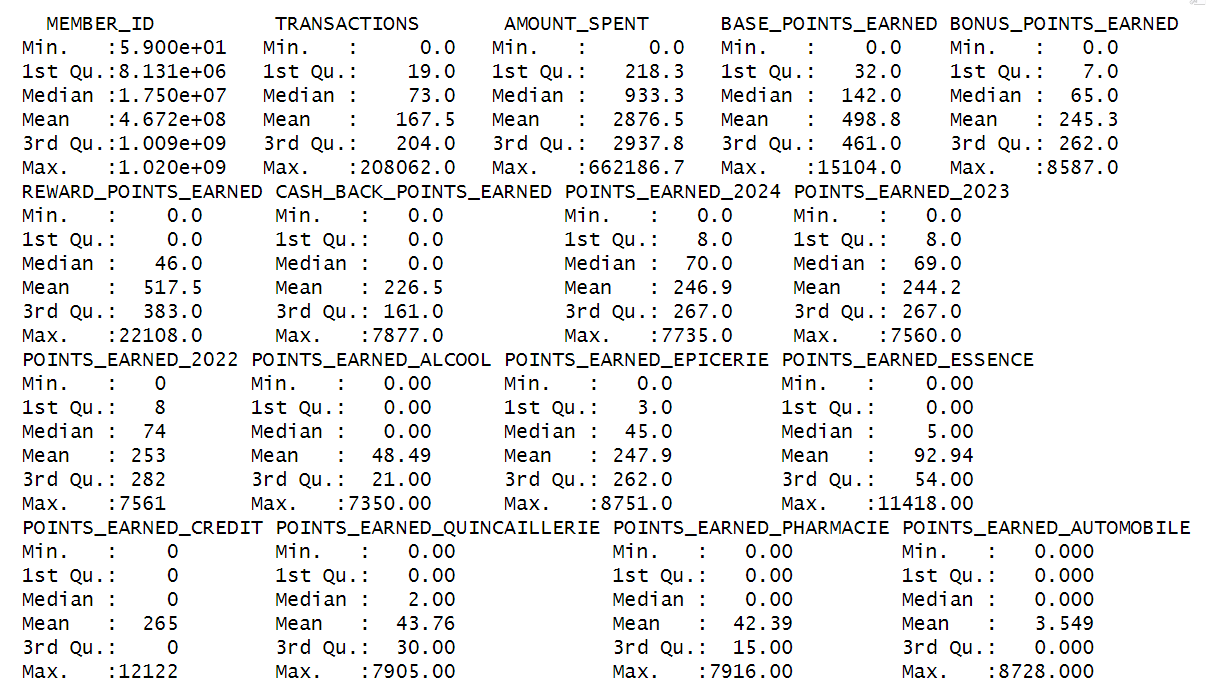
Je constate, dans le sommaire de Transaction\_agg, que Bonus\_Points\_Earned\_Ratio présente une faible variabilité et une dispersion limitée. Afin d’assurer une meilleure comparabilité entre les variables et d’éviter que cette variable soit sous-représentée dans l’algorithme de segmentation, je vais la normaliser.

J’ai ensuite analysé les corrélations entre les variables de la table agrégée à l’aide d’une matrice de corrélation. Les résultats démontrent qu’aucune variable ne présente de corrélations anormalement élevées, ce qui indique une bonne indépendance entre les variables. Comme attendu, les variables liées aux points accumulés et aux montants dépensés affichent une forte corrélation, ce qui est logique puisqu’elles reflètent directement les habitudes d’achat des membres. Cette relation avait d’ailleurs été anticipée et justifiée précédemment.

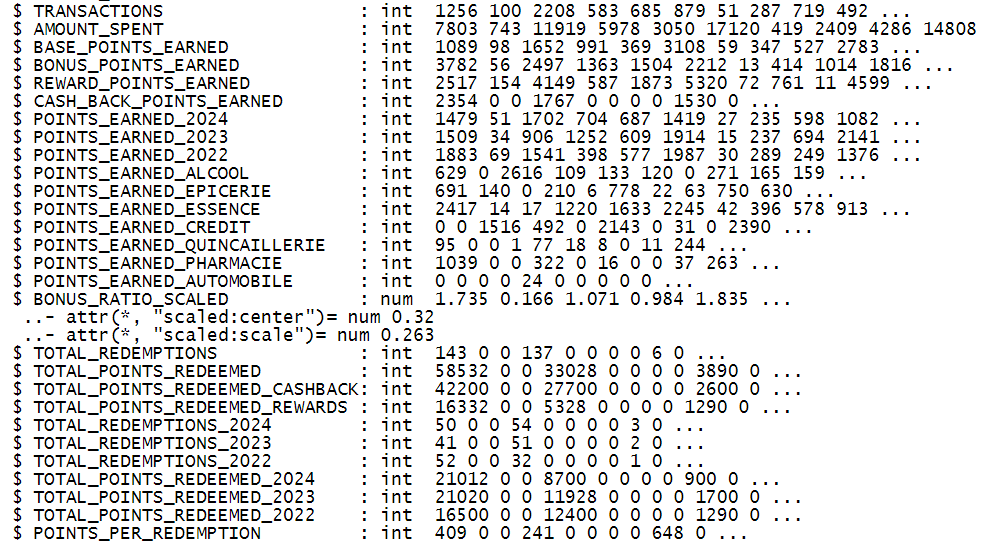
## Segmentation

Nous allons maintenant procéder à la jointure finale entre Transaction\_agg et Reward\_agg afin de créer le dataset final (Final\_agg), qui servira de base pour la segmentation. Cette fusion permettra de consolider les informations sur l’accumulation et l’utilisation des points par les membres, offrant ainsi une vue complète et optimisée pour l’analyse.

L’examen des corrélations entre les variables révèle que, à l’exception d’une forte corrélation entre Amount\_Spent et Total\_Redemptions, les autres variables ne présentent pas de relations significatives très fortes ou anormales entre elles. Bien que la corrélation ci-haut soit notable, ces deux variables restent essentielles pour l’analyse. Elles permettent de mieux comprendre les comportements d’achat et de rédemption des membres et ont également contribué à la création de plusieurs autres variables clés du dataset. Je ne ferais aucun autre changement concernant ces variables, mais il est important de prendre en compte cette corrélation dans les analyses ultérieures, car une multicolinéarité excessive pourrait potentiellement affecter la stabilité et l’interprétation des modèles statistiques utilisés dans la segmentation.



Le sommaire indique que toutes les variables de Final\_agg présentent des valeurs extrêmes dans leurs maximums, particulièrement en comparaison avec leurs troisièmes quartiles (Q3). Afin d’éviter que ces valeurs influencent de manière disproportionnée la segmentation, je vais encore appliquer une méthode de plafonnement au 99e percentile pour chaque variable. Cette approche permet de réduire l’impact des outliers tout en conservant une représentation fidèle des comportements des membres dans le dataset.



Voici la liste finale des variables de Final\_agg. On remarque que Member\_ID n’est plus présent. En effet, je l’ai supprimé pour éviter des erreurs futures de biais lors de la création des segments. Cela garantit que seules les variables pertinentes influencent la classification des membres. On va rajouter le Member\_ID après l’utilisation de l’algorithme pour lier les variables socio-démographiques afin de décrire les segments.

Bien que certaines de mes variables puissent introduire une certaine redondance, il est difficile de justifier leur exclusion dans le cadre de mon objectif d’affaires, d’autant plus qu’elles ne présentent pas de corrélations excessivement élevées entre elles (sauf pour les variables de points gagnés et Amount\_Spent).

Afin d’optimiser l’analyse tout en conservant l’essentiel de l’information, une approche envisageable serait de réaliser une Analyse en Composantes Principales (ACP). Cette méthode permettrait de réduire la dimensionnalité du dataset en regroupant les informations des variables existantes au sein de nouvelles variables synthétiques (PC1, PC2, ...). En appliquant l’ACP, je pourrais ainsi minimiser la redondance.

Par contre, avant de procéder à l’Analyse en Composantes Principales (ACP), je vais d’abord tenter de réaliser une segmentation manuelle afin d’obtenir une première analyse préliminaire rapide des profils de membres et peut-être obtenir un meilleur résultat de segmentation.

Pour débuter, il est essentiel d’identifier les variables clés qui seront utilisées pour la segmentation manuelle. En tenant compte de l’objectif d’affaires et après plusieurs essais-erreurs pour obtenir le meilleur ensemble, voici les variables que j’ai sélectionnées pour la segmentation:

|  |  |
| --- | --- |
| 1. Transactions | 1. Amount\_Spent |
| 1. Base\_Points\_Earned | 1. Bonus\_Points\_Earned |
| 1. Reward\_Points\_Earned | 1. Cash\_Back\_Points\_Earned |
| 1. Total\_Redemptions | 1. Total\_Points\_Redeemed |
| 1. Points\_Per\_Redemption |  |

J’ai choisi ces 9 variables car elles couvrent les deux dimensions essentielles de l’analyse : l’accumulation et l’utilisation des points. De plus, elles vont permettre une interprétation claire des segments, ce qui facilitera leur exploitation dans éventuelle stratégie marketing.

Avant d’appliquer la segmentation, j’ai réduit la dimensionnalité du dataset Final\_agg\_man et standardiser les variables afin de les ramener sur des échelles comparables.

J’ai ensuite consacré beaucoup d’efforts à la définition manuelle des règles de segmentation. Mon objectif est de créer des segments cohérents, basés sur des seuils pertinents en lien avec l’objectif d’affaires. Après plusieurs explorations préliminaires, j’ai identifié quatre segments clés, qui, selon moi, reflètent de manière fidèle la réalité de la base de données et permettent de différencier efficacement les comportements des membres.

1. **Grands acheteurs** (fort volume de transactions et montants élevés)

2. **Chasseurs de bonus** (volume élevé de points bonus ou cashback)

3. **Utilisateurs actifs** (rédemptions fréquentes et volume élevé de points utilisés)

4. **Membres inactifs** (faible engagement, peu de transactions et de points gagnés)

Je vais maintenant définir manuellement les règles pour faire partie de chaque segment.

Voici les règles que j'ai établies pour créer des segments clairs et bien différenciés :

**Segment 1 : Grands acheteurs :**

TRANSACTIONS ≥ 200 (> 3e quartile)

AMOUNT\_SPENT ≥ 2800 (> moyenne)

**Segment 2 : Chasseurs de bonus :**

BONUS\_POINTS\_EARNED ≥ 250 (> 3e quartile) ou

CASH\_BACK\_POINTS\_EARNED ≥ 160 (> 3e quartile)

TRANSACTIONS < 200 (pour éviter le chevauchement avec le segment 1)

**Segment 3 : Utilisateurs actifs :**

TOTAL\_REDEMPTIONS ≥ 1 (au moins 1 redemption, > 3e quartile)

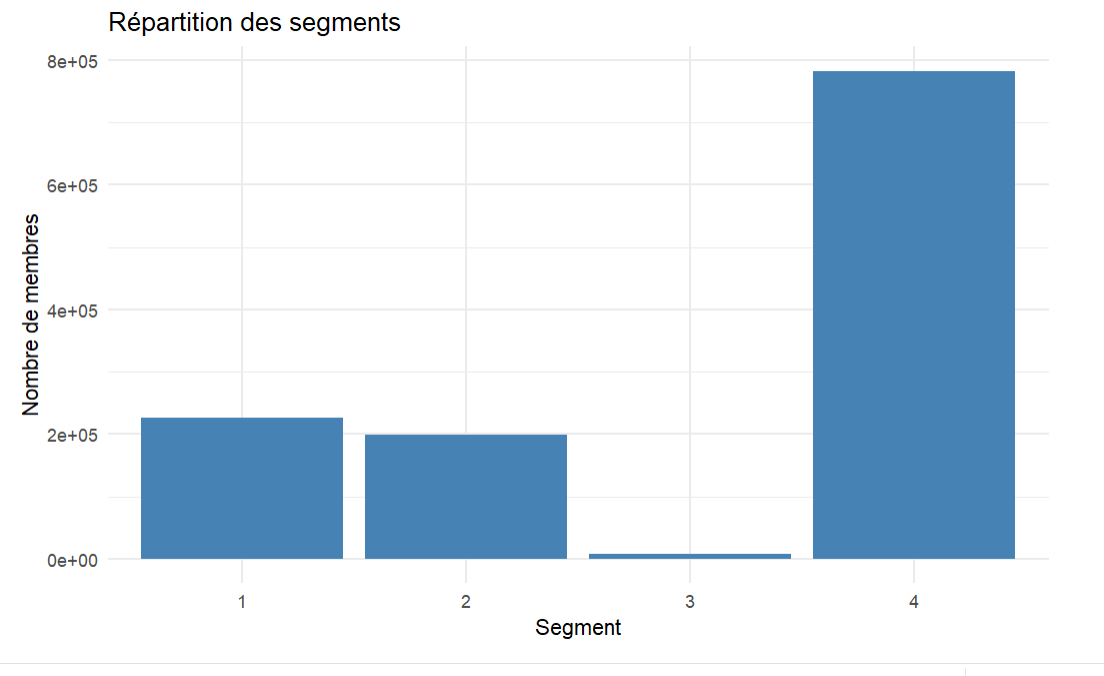
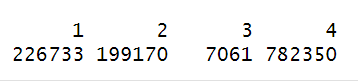
TOTAL\_POINTS\_REDEEMED ≥ 200 (> 3e quartile)

**Segment 4 : Membres à faible activité**

TRANSACTIONS < 20 (inférieur au premier quartile)

AMOUNT\_SPENT < 250 (inférieur au premier quartile)

TOTAL\_POINTS\_REDEEMED < 150



En évaluant la répartition des observations de la segmentation, on remarque que le segment 4 (65%) contient une majorité écrasante des données. Cela n’est vraiment pas optimale. De plus, le segment 3 ne contient que 7 000 données ce qui est très faible. Le choix de cette répartition compliquerait le ciblage marketing et limiterait la personnalisation des offres marketing.

Malgré plusieurs ajustements réalisés pour optimiser la répartition des membres, il s’agit de la segmentation manuelle qui offre le meilleur équilibre possible. Lors des différents essais, plusieurs configurations ont entraîné un nombre important de valeurs manquantes, ce qui a limité les possibilités d’affiner davantage la segmentation avec les variables sélectionnées.

Une segmentation plus équilibrée permettrait de mieux répartir les membres entre les segments et d’identifier des groupes plus exploitables pour des campagnes stratégiques. Ainsi, plutôt que d'affiner davantage la segmentation manuelle, ce qui pourrait entraîner une répartition déséquilibrée des observations, j’ai décidé d’explorer une approche plus robuste en utilisant la méthode CLARA sur R, une méthode de segmentation basée sur les k-medoids.

Avant d’aller de l’avant avec la segmentation avec CLARA, revenons au dataset Final\_agg.

Voici donc le tableau récapitulatif des variables qui seront utilisées dans la segmentation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Nom Variable | Description (pour chaque membre) | Origine |
| 1 | TRANSACTIONS | Nombre total de transactions effectuées | Originale (Somme) |
| 2 | AMOUNT\_SPENT | Montant total dépensé (en $) | Originale (Somme) |
| 3 | BASE\_POINTS\_EARNED | Nombre total de points de base accumulés. | Originale (Somme) |
| 4 | BONUS\_POINTS\_EARNED | Nombre total de points bonus gagnés. | Originale (Somme) |
| 5 | REWARD\_POINTS\_EARNED | Total des points accumulés pour des récompenses. | Originale (Somme) |
| 6 | CASH\_BACK\_POINTS\_EARNED | Total des points utilisés pour du cashback. | Originale (Somme) |
| 7 | POINTS\_EARNED\_2024 | Nombre de points gagnés en 2024. | Transformée |
| 8 | POINTS\_EARNED\_2023 | Nombre de points gagnés en 2023. | Transformée |
| 9 | POINTS\_EARNED\_2022 | Nombre de points gagnés en 2022. | Transformée |
| 10 | POINTS\_EARNED\_ALCOOL | Points gagnés sur des achats d’alcool. | Transformée |
| 11 | POINTS\_EARNED\_EPICERIE | Points gagnés sur des achats en épicerie. | Transformée |
| 12 | POINTS\_EARNED\_ESSENCE | Points gagnés sur des achats d’essence. | Transformée |
| 13 | POINTS\_EARNED\_CREDIT | Points gagnés sur des achats liées aux cartes de crédit. | Transformée |
| 14 | POINTS\_EARNED\_QUINCAILLERIE | Points gagnés sur des achats en quincaillerie. | Transformée |
| 15 | POINTS\_EARNED\_PHARMACIE | Points gagnés sur des achats en pharmacie | Transformée |
| 16 | POINTS\_EARNED\_AUTOMOBILE | Points gagnés sur des achats liés à l’automobile. | Transformée |
| 17 | BONUS\_RATIO\_SCALED | Ratio normalisé des points bonus gagnés sur le total des points. (Calcul: BONUS\_POINTS\_EARNED / (BASE\_POINTS\_EARNED + BONUS\_POINTS\_EARNED) | Transformée |
| 18 | TOTAL\_REDEMPTIONS | Nombre total de rédemptions effectuées. | Originale (Somme) |
| 19 | TOTAL\_POINTS\_REDEEMED | Nombre total de points dépensés en rédemptions. | Originale (Somme) |
| 20 | TOTAL\_POINTS\_REDEEMED\_  CASHBACK | Total des points dépensés en cashback. | Transformée |
| 21 | TOTAL\_POINTS\_REDEEMED\_  REWARDS | Nombre total de points utilisés pour des récompenses (cadeaux, voyages, etc.). | Transformée |
| 22 | TOTAL\_REDEMPTIONS\_2024 | Nombre total de rédemptions effectuées en 2024. | Transformée |
| 23 | TOTAL\_REDEMPTIONS\_2023 | Nombre total de rédemptions effectuées en 2023. | Transformée |
| 24 | TOTAL\_REDEMPTIONS\_2022 | Nombre total de rédemptions effectuées en 2022. | Transformée |
| 25 | TOTAL\_POINTS\_REDEEMED\_2024 | Nombre total de points dépensés en 2024. | Transformée |
| 26 | TOTAL\_POINTS\_REDEEMED\_2023 | Nombre total de points dépensés en 2023. | Transformée |
| 27 | TOTAL\_POINTS\_REDEEMED\_2022 | Nombre total de points dépensés en 2022. | Transformée |
| 28 | POINTS\_PER\_REDEMPTION | Nombre moyen de points dépensés par rédemption. | Transformée |

Les variables originales restent inchangées par rapport à leur définition initiale, mais elles ont été agrégées par membre. Elles représentent désormais la somme des valeurs de chaque variable pour un même Member\_ID. Par exemple, Transactions correspond à la somme totale des transactions effectuées par un membre et cumule toutes les lignes de ce dernier dans le jeu de données d’origine.

### ACP

Évidemment, 28 variables constituent un nombre trop élevé pour une segmentation efficace. Cela peut entraîner plusieurs problèmes, notamment :

* Un risque de sur-ajustement (overfitting), où le modèle capte des variations spécifiques aux données plutôt que des tendances générales.
* De la multicolinéarité, rendant certaines variables redondantes et pouvant fausser l’algorithme de clustering.
* Des difficultés d’interprétation, compliquant l’analyse des segments et la prise de décisions stratégiques.

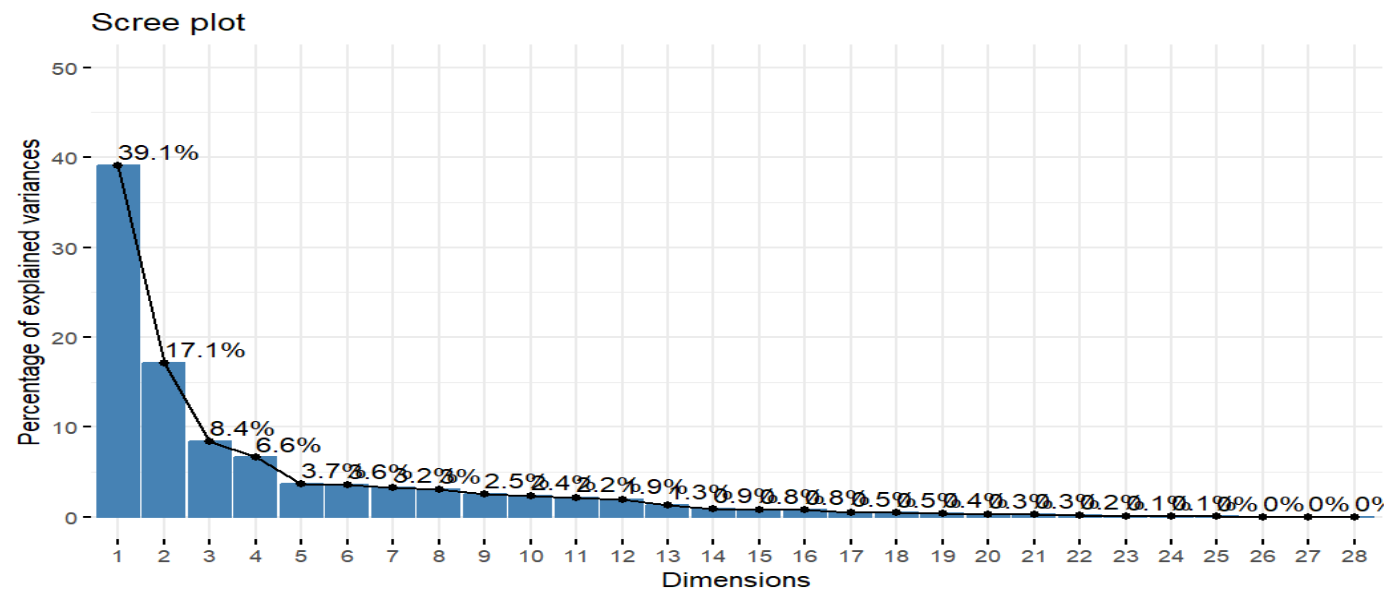
C’est pourquoi une Analyse en Composantes Principales (ACP) sera réalisée au préalable afin de réduire la dimensionnalité du jeu de données et préserver uniquement l’information essentielle pour la segmentation.

Après avoir remplacé les NA par des 0, j’ai appliqué l’ACP à Final\_agg. Voici les résultats :

A close-up of a number

AI-generated content may be incorrect.

Ces résultats montrent que les six premières composantes principales expliquent environ 78.5% de la variance du jeu de données. Pour mieux visualiser l’évolution de la variance expliquée par chaque composante, je vais générer un scree plot, qui aidera à identifier le nombre optimal de composantes à conserver.



Ce graphique valide le choix logique de retenir 6 composantes principales pour la segmentation. Avec ces 6 composantes, on capture environ 78.5% de la variance totale, ce qui est largement suffisant pour garantir une différenciation efficace des segments. Ce nombre de dimensions facilitera aussi la mise en œuvre de stratégies marketing ciblées. Chaque segment bénéficiera de caractéristiques distinctes, bien définies et exploitables d’un point de vue opérationnel.

On dispose désormais d’un nouveau jeu de données, nommé pca\_data, qui regroupe les 6 composantes principales issues de l’ACP. La prochaine étape consiste à déterminer le nombre optimal de clusters à utiliser pour la segmentation avec la méthode du coude. (Voir Annexe 2)

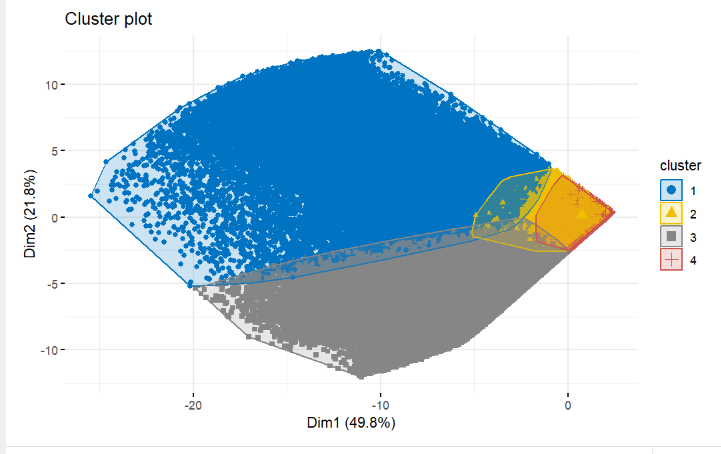
L’analyse du graphique de la méthode du coude indique que le nombre optimal de clusters se situe entre 3 et 4. J’ai donc choisi de retenir 4 clusters pour la segmentation.

Ce choix permet de garantir une bonne différenciation des segments, tout en évitant une segmentation trop complexe et un risque de redondance élevé.

On peut maintenant lancer la segmentation en fixant le seed pour garantir la reproductibilité des résultats. Les résultats de l’algorithme CLARA seront présentés sous forme d’un sommaire, affichant les moyennes des variables pour chaque cluster. Cela me permettra d’analyser les caractéristiques distinctives de chaque segment et d’identifier les principales tendances comportementales.



On observe déjà que la répartition des observations par segment est beaucoup plus équilibrée, rendant la segmentation plus représentative et exploitable pour de futures actions marketing. De plus, graphiquement, on voit que les 4 clusters sont bien visibles et assez distincts les uns des autres.



Pour aller plus loin, on va maintenant intégrer les clusters ainsi que les variables de la table Members à Final\_agg afin d’examiner l’influence des variables socio-démographiques sur le comportement des membres au sein de chaque segment. (Voir annexe 3 pour graphique)

J’ai poussé davantage l’analyse des variables socio-démographiques en créant des tables sommaires contenant les moyennes pertinentes pour l’interprétation des segments. (Voir annexe 4)

## Interprétation

Maintenant qu’on a toutes les informations nécessaires lié à l’impact des variables comportementales et socio-démographiques sur les segments, on peut effectuer l’interprétation complète des 4 segments obtenus. (Voir les annexes 4, 5 et 6 afin de consulter les tables des moyennes des variables pour chaque segment. Celles-ci ont servi à l’interprétation.)

**🚩 Segment 1 : « Seniors cashback »**

**Profil transactionnel :**

* **Transactions :** Modérées (416 transactions en moyenne).
* **Montants dépensés :** Modérés (5900 $ en moyenne).
* **Points gagnés :** Principalement en cashback (1360 pts moyens), utilisation modérée des points Reward.
* **Rédemptions :** Fréquentes avec une moyenne très élevée de points cashback (2300 pts).
* **Ancienneté :** Très élevée (plus de 20 ans majoritairement).

**Profil socio-démographique :**

* **Âge moyen :** Le plus élevé (60 ans en moyenne), forte proportion de membres âgés de 61-75 ans.
* **Genre :** Majoritairement féminin (64%).
* **Localisation :** Ontario et Québec principalement.
* **Méthode de contact :** Courrier traditionnel.

**📌 Interprétation & Recommandation :**

Ce groupe représente des clients seniors très fidèles utilisant principalement du cashback. Ils ont accumulé des points qu’ils utilisent régulièrement. L’âge avancé suggère une préférence pour la simplicité et les récompenses tangibles.

➡️ **Stratégie recommandée :**

* Maintenir une forte communication via le courrier traditionnel.
* Offrir des promotions cashback spécifiques, par exemple des bonus lors d'occasions spéciales (anniversaires, fêtes des grands-parents).
* Proposer des récompenses concrètes facilement utilisables dans leur quotidien (épicerie, pharmacie).

**🚩 Segment 2 : « Occasionnels »**

**Profil transactionnel :**

* **Transactions :** Faibles à modérées (150 transactions).
* **Montants dépensés :** Faibles (1500 $ en moyenne).
* **Points gagnés :** Faible en cashback (140 pts), Reward (300 pts).
* **Rédemptions :** Très peu de Redemptions, points principalement conservés.
* **Ancienneté :** Modérée (environ 19 ans en moyenne).

**Profil socio-démographique :**

* **Âge moyen :** Relativement jeune (55 ans), forte proportion de 30-45 ans.
* **Genre :** Majoritairement feminin (59%)
* **Localisation :** Ontario surtout.
* **Méthode de contact :** Principalement par courrier, très faible interaction digitale.

**📌 Interprétation & Recommandation :**

Ce segment est prudent, avec peu d'engagement dans les rédemptions. Ils accumulent lentement leurs points et les utilisent peu, ce qui suggère un faible intérêt ou un manque de compréhension du programme.

➡️ **Stratégie recommandée :**

* Communication pédagogique ciblée pour expliquer clairement les avantages du programme et encourager l'utilisation.
* Campagnes promotionnelles incitatives (offres limitées dans le temps pour encourager l'utilisation rapide des points accumulés).
* Tester une transition douce vers des canaux digitaux (courriel) pour accroître l’engagement régulier.

**🚩 Segment 3 : « Grands Dépensiers Rewards »**

**Profil transactionnel :**

* **Transactions :** Élevées (390 transactions).
* **Montants dépensés :** Très élevés (16 350 $ en moyenne, le plus élevé).
* **Points gagnés :** Très élevés en points Reward (3940 pts), faibles en cashback.
* **Rédemptions :** Modérées mais avec un solde Rewards très élevé (684 pts), ce sont clairement des accumulateurs stratégiques.
* **Ancienneté :** Très longue (plus de 20 ans).

**Profil socio-démographique :**

* **Âge moyen :** Élevé mais légèrement inférieur au segment 1 (58 ans).
* **Genre :** Assez équilibré.
* **Localisation :** Fortement représentés en Ontario, avec une légère diversité géographique (BC, AB).
* **Méthode de contact :** Préférence marquée pour Courriel + Mail (digitaux et traditionnels combinés).

**📌 Interprétation & Recommandation :**

Segment à haute valeur, membres très fidèles, dépensant beaucoup et accumulant stratégiquement des points Rewards. Ils montrent une excellente connaissance et utilisation stratégique du programme.

➡️ **Stratégie recommandée :**

* Programmes de récompenses VIP ou exclusifs (accès anticipé à des promotions, expériences haut de gamme).
* Incitations à utiliser davantage leurs points (réductions sur voyages, expériences exclusives).
* Renforcement de la personnalisation avec communication mixte digitale et physique (newsletter exclusive + courrier personnalisé).

**🚩 Segment 4 : « Faibles Dépensiers »**

**Profil transactionnel :**

* **Transactions :** Très faibles (33 transactions en moyenne).
* **Montants dépensés :** Très faibles (710 $ en moyenne).
* **Points gagnés :** Très faibles (principalement Rewards).
* **Rédemptions :** Très faibles, engagement minimal.
* **Ancienneté :** Faible (environ 16 ans, majorité moins de 10 ans).

**Profil socio-démographique :**

* **Âge moyen :** Jeune (55 ans), avec une forte proportion des moins de 30 ans et 30-45 ans.
* **Genre :** Majoritairement féminin (60%).
* **Localisation :** Ontario et Québec.
* **Méthode de contact :** Fortement orienté courrier traditionnel uniquement (61%).

**📌 Interprétation & Recommandation :**

Ce segment est peu actif, probablement des clients occasionnels, potentiellement nouveaux ou peu convaincus. Ils représentent une opportunité de croissance s’ils sont correctement activés.

➡️ **Stratégie recommandée :**

* Communications d'accueil ou de réactivation personnalisées via canaux digitaux (courriels, notifications mobiles) pour mieux rejoindre cette clientèle jeune.
* Offres incitatives fortes pour déclencher un premier ou nouveau comportement (bonus de bienvenue, prime de réactivation).
* Simplification des mécanismes de récompenses pour encourager leur engagement initial.

**🌟 Conclusion générale et recommandations stratégiques globales :**

La segmentation révèle des profils très distincts permettant une stratégie personnalisée et ciblée :

* **Segment 1 (Seniors Cashback)** et **Segment 3 (Grands Dépensiers Rewards)** représentent la majorité de la valeur du programme. Ils méritent une attention VIP et des récompenses adaptées à leur fidélité à long terme.
* **Segment 2 (Occasionnels)** et **Segment 4 (Faibles Dépensiers)** nécessitent des stratégies d'éducation et d'incitation à l'utilisation, afin de maximiser leur potentiel futur.

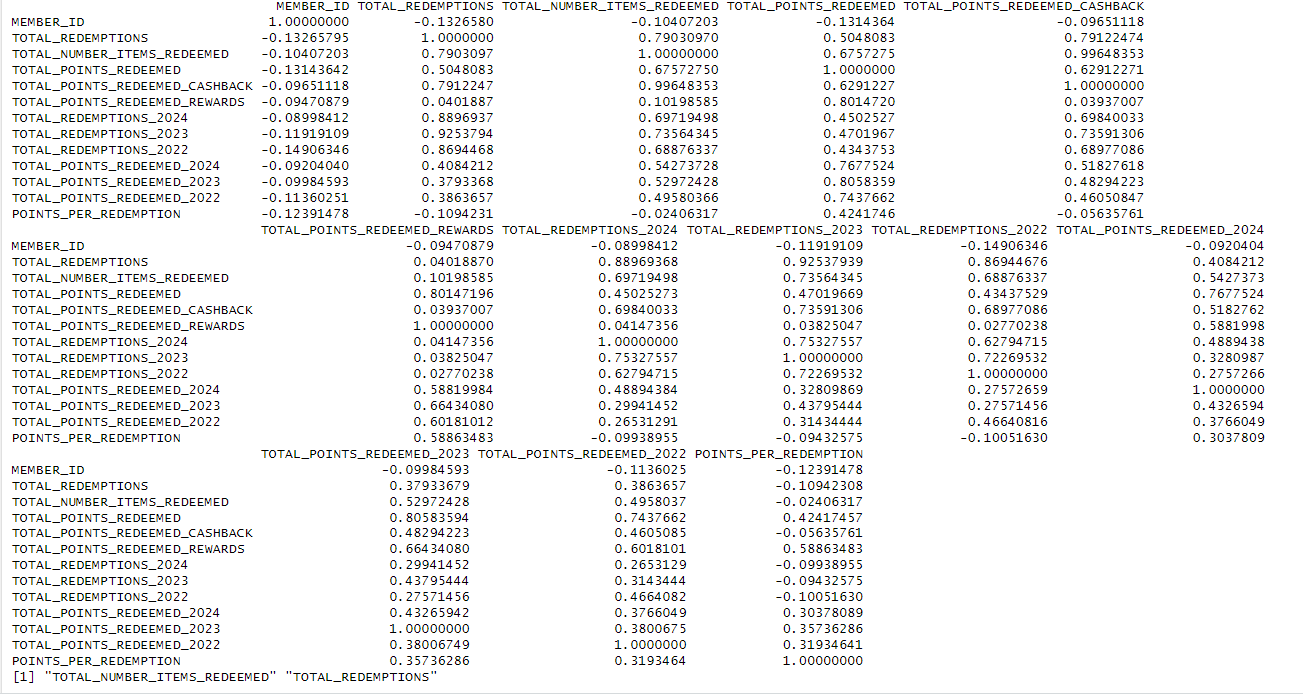
En appliquant ces recommandations stratégiques basées sur l'analyse fine des comportements transactionnels et du profil socio-démographique, Loyalty pourra optimiser la fidélisation, augmenter la satisfaction client, et maximiser la valeur à long terme des membres du programme Air Miles

\*\*\* Disclaimer \*\*\*

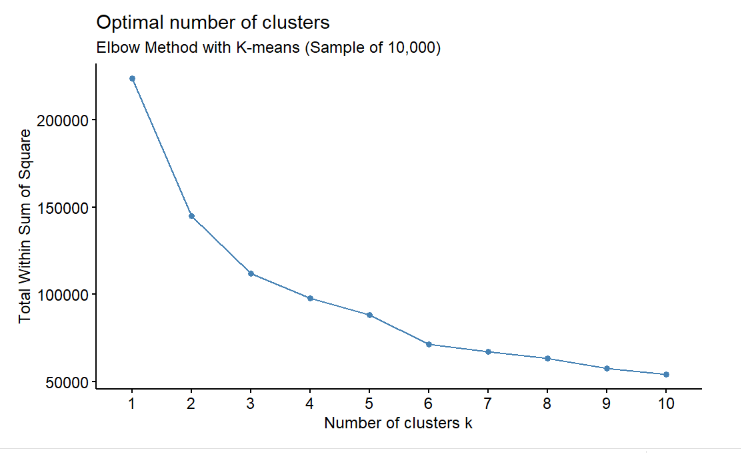
La mise en forme de l’interprétation des segments ainsi que le code R lié à la création de variables, l’ACP et la segmentation a été réalisée avec l’aide de Chat GPT.

## Annexe

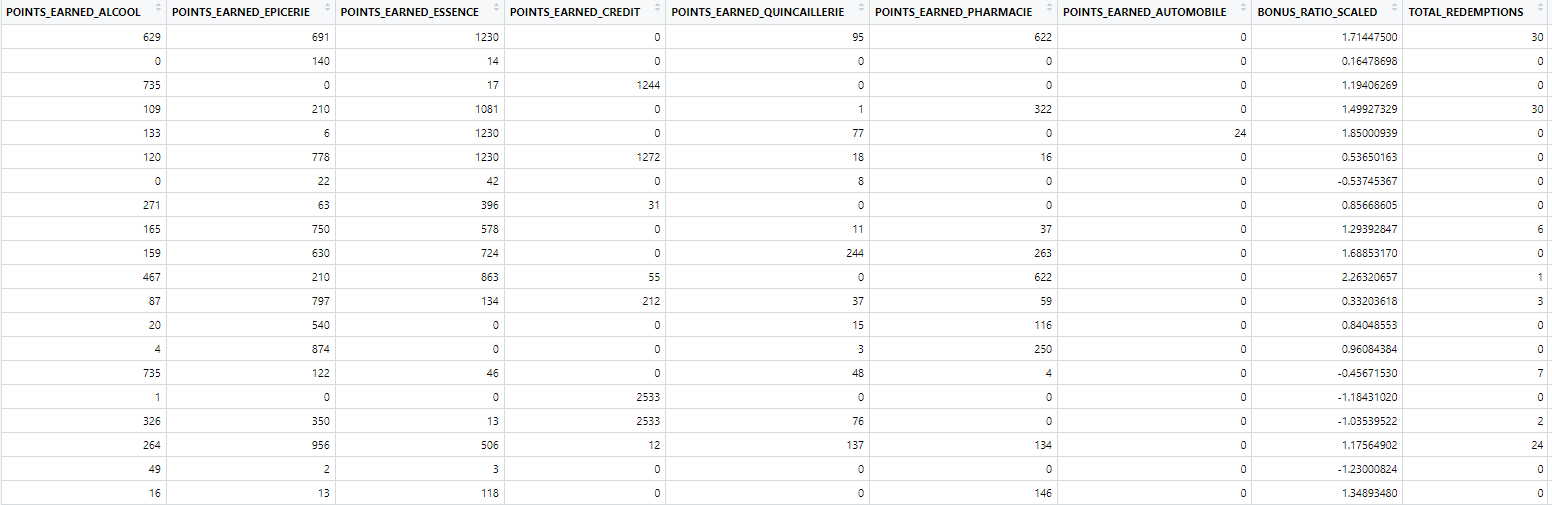
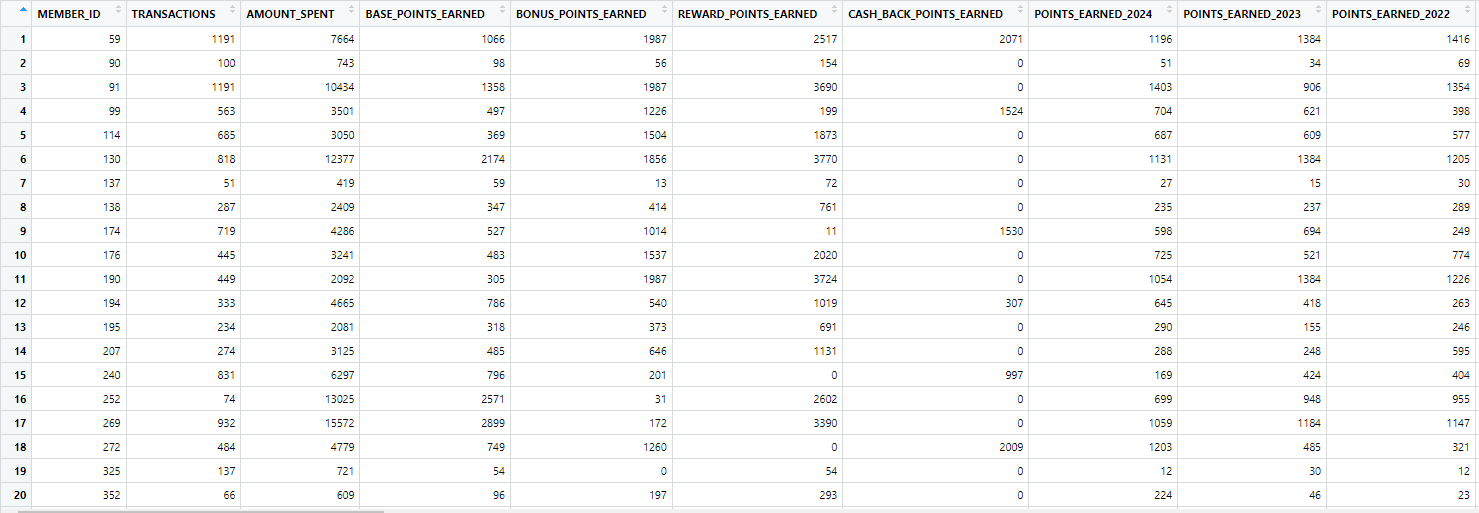
**Annexe 1 : Matrice de corrélation Reward\_agg**

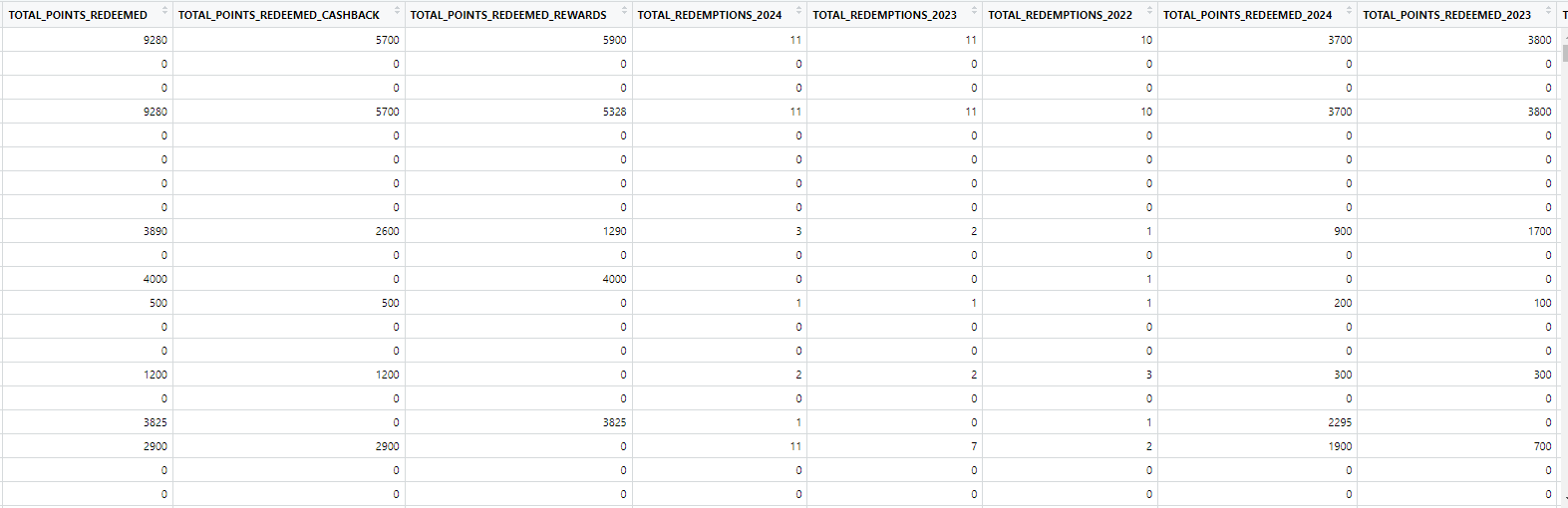


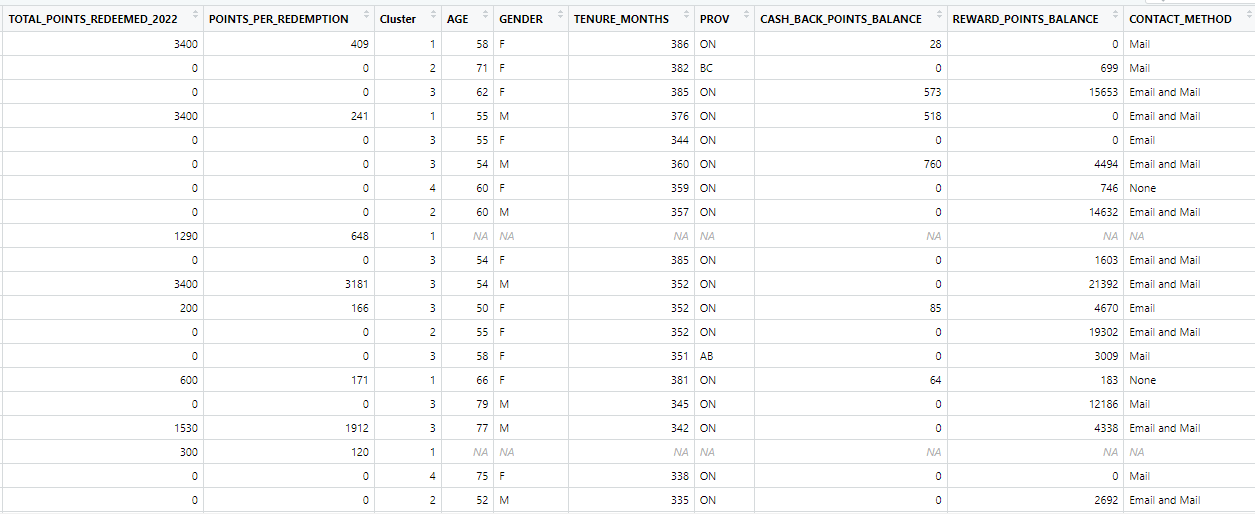
**Annexe 2 : Graphique méthode du coude pour segmentation**

****

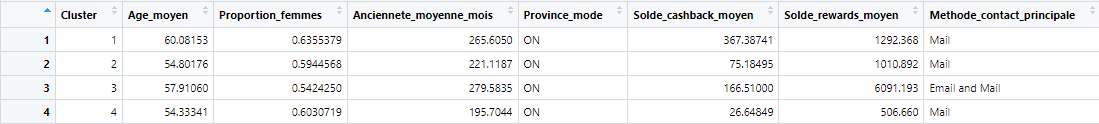
**Annexe 3 : Apercu de la Table finale Final\_agg avec les variables de Members**

****

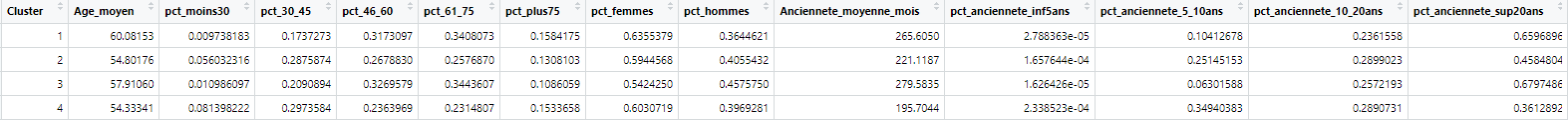
****

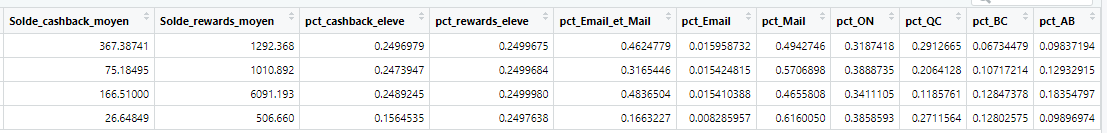
****

**Annexe 4 : Socio-Demo Summary**

****

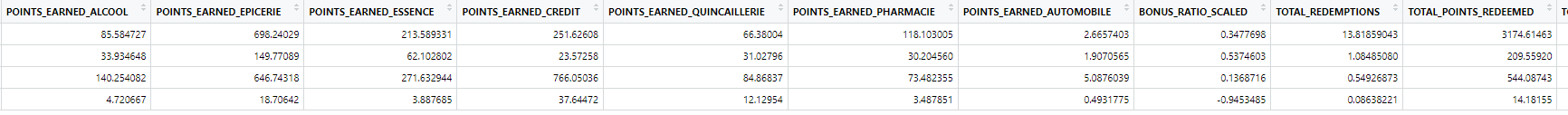
**Annexe 5 : Socio-Demo Detailed**

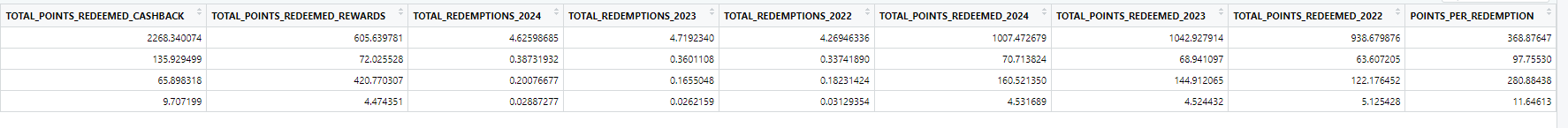
****

****

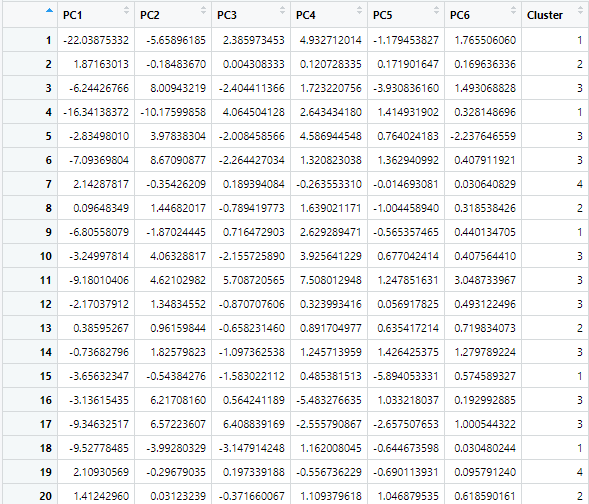
**Annexe 6 : Cluster Summary**

****

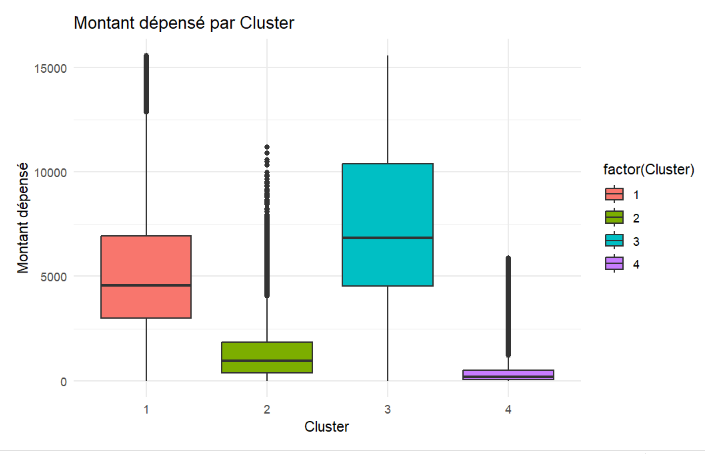
****

****

**Annexe 7 : Apercu de la table pca data**

****

**Annexe 8 : Graphique du Montant dépensé par Cluster**

****