HEC MONTRÉAL

Analyse de Bases de Données en Marketing
Devoir INDIVIDUEL (Leroy Tiojip - 11348337)

SEGMENTATION

Problématique d'affaires

Comment Loyalty Air Miles peut-elle optimiser son programme de fidélisation en segmentant efficacement ses membres afin d'améliorer l'engagement, la rétention et la valeur à long terme des clients tout en adaptant ses actions marketing aux comportements d'achat et d'utilisation des récompenses ?

Justification de la problématique :

Le programme de fidélité Air Miles repose sur l'accumulation et l'utilisation de points par ses membres. Cependant, avec **plus de 1,25 million de membres**, il est essentiel de **comprendre les différences comportementales** entre eux pour proposer des offres adaptées. L'objectif de cette segmentation est d'aider **Loyalty Air Miles** à :

1. **Identifier les segments de clients les plus engageants et rentables**
2. **Différencier les comportements d'accumulation et d'utilisation des points**
3. **Cibler les segments à risque pour limiter l'attrition et stimuler l'engagement**
4. **Optimiser les offres promotionnelles et les campagnes marketing**

Orientation du devoir

1. **Sélection des variables clés** qui influencent l'engagement des clients dans le programme.
2. **Segmentation des clients en groupes homogènes** en fonction de leur comportement d'achat et d'utilisation des récompenses.
3. **Analyse des segments** pour comprendre leurs caractéristiques et leurs différences. Et **recommandations stratégiques** pour **ajuster les actions marketing** selon les segments identifiés.

Partie I : Sélection des variables clés

Justification des variables sélectionnées

a. Variables transactionnelles et financières

Variable	Constituée à partir de	Rôle dans la segmentation
TOTAL_TRANSACTIONS	TRANSACTIONS (somme)	Indique la fréquence d'utilisation du programme.
TOTAL_AMOUNT_SPENT	AMOUNT_SPENT (somme)	Mesure la valeur du client en fonction de ses dépenses totales.
Avg_Spending_Per_Transaction	TOTAL_AMOUNT_SPENT / TOTAL_TRANSACTIONS	Segmente les membres en petits acheteurs vs gros dépensiers.
Transaction_Frequency	TOTAL_TRANSACTIONS / TENURE_MONTHS	Mesure la fréquence mensuelle d'achat.

Approche stratégique

Elles permettent de **catégoriser les membres** selon leur engagement transactionnel :

- **Achats fréquents vs occasionnels**
- **Gros dépensiers vs petits consommateurs**

b. Variables liées à l'accumulation des points

Variable	Constituée à partir de	Rôle dans la segmentation
TOTAL_BASE_POINTS_EARNED	BASE_POINTS_EARNED (somme)	Points accumulés via des achats normaux.
TOTAL_BONUS_POINTS_EARNED	BONUS_POINTS_EARNED (somme)	Points gagnés via des promotions.
TOTAL_REWARD_POINTS_EARNED	BASE_POINTS_EARNED + BONUS_POINTS_EARNED	Total des points accumulés.

TOTAL_CASH_BACK_POINTS_EARNED	CASH_BACK_POINTS_EARNED (somme)	Points convertis en cashback.
CASH_BACK_POINTS_BALANCE	CASH_BACK_POINTS_BALANCE (valeur récente)	Solde final des points de cashback.
REWARD_POINTS_BALANCE	REWARD_POINTS_BALANCE (valeur récente)	Solde final des points de récompenses.

Approche stratégique

Elles permettent de **différencier les clients** en fonction de leur manière d'accumuler des points :

- **Ceux qui accumulent régulièrement vs ceux qui ne gagnent presque rien**
- **Ceux qui réagissent aux promotions vs ceux qui ne maximisent pas leurs gains**

c. Variables liées à l'utilisation des points

Variable	**Constituée à partir de**	**Rôle dans la segmentation**
TOTAL_POINTS_REDEEMED	POINTS_REDEEMED (somme)	Nombre total de points utilisés par le membre.
TOTAL_REDEMPTIONS	REDEMPTIONS (somme)	Nombre total de transactions d'échange de points.
TOTAL_ITEMS_REDEEMED	NUMBER_ITEMS_REDEEMED (somme)	Nombre total d'articles obtenus via les rédemptions.
Points_Utilization_Score	TOTAL_POINTS_REDEEMED / (TOTAL_BASE_POINTS_EARNED + TOTAL_BONUS_POINTS_EARNED + 1) (+1, cest pour éviter une forme indéterminée)	Score d'utilisation des points (clients actifs vs accumulateurs passifs).

Cash_Back_Usage_Ratio	TOTAL_CASH_BACK_POINTS_EARNED / (TOTAL_CASH_BACK_POINTS_EARNED + TOTAL_REWARD_POINTS_EARNED + 1) (+1, cest pour éviter une forme indéterminée)	Identifie les clients préférant le cashback aux récompenses classiques.
Redemption_Frequency	TOTAL_REDEMPTIONS / TOTAL_TRANSACTIONS	Mesure la fréquence d'échange des points.

Approche stratégique

Elles permettent de **classifier les clients** en fonction de **leur comportement d'utilisation des points** :

- **Accumulateurs passifs** (beaucoup de points non utilisés)
- **Utilisateurs réguliers** (points échangés fréquemment)
- **Préférences entre cashback et récompenses traditionnelles**

d. Variables liées aux préférences d'achat et de récompenses

Variable	**Constituée à partir de**	**Rôle dans la segmentation**
UNIQUE_RETAILERS	RETAILER (nombres distincts)	Nombre de catégories de détaillants fréquentés.
Top_Retailer_Category_Share	UNIQUE_RETAILERS / TOTAL_TRANSACTIONS	Identifie la diversité des achats du client.
UNIQUE_REWARD_CATEGORIES	REWARDS_CATEGORY (nombres distincts)	Nombre de types de récompenses différentes utilisées.
Top_Reward_Category_Share	UNIQUE_REWARD_CATEGORIES / TOTAL_REDEMPTIONS	Identifie la répartition des choix de récompenses.

Approche stratégique

Elles permettent de **comprendre les préférences des membres** et de **les segmenter en fonction de leurs habitudes d'achat et d'échange** :

- **Clients fidèles à une catégorie vs clients explorateurs**
- **Utilisateurs de récompenses diversifiées vs spécialisés dans un type de récompense**

e. Variables globales d'engagement et de fidélité

Variable	**Constituée à partir de**	**Rôle dans la segmentation**
TENURE_MONTHS	TENURE_MONTHS (valeur unique)	Ancienneté du membre dans le programme.
Loyalty_Score	(TOTAL_TRANSACTIONS + TOTAL_REDEMPTIONS + REWARD_POINTS_BALANCE) / (TENURE_MONTHS + 1)	Score global de fidélité basé sur plusieurs facteurs.

Approche stratégique

Elles aident à **comparer les membres selon leur engagement global** dans le programme :

- **Anciens membres fidèles vs nouveaux entrants actifs**
- **Clients ayant une valeur élevée pour le programme vs utilisateurs dormants**

Justification de l'exclusion de certaines variables

Certaines variables ont été exclues car elles ne reflètent pas directement le comportement d'achat ou d'utilisation du programme (ex. **GENDER, LANGUAGE**), concernent des préférences de communication sans impact sur la fidélité (ex. **MAAILABLE_FLAG, EMAILABLE_FLAG**), sont des informations géographiques ou démographiques non essentielles à la segmentation comportementale (ex. **AGE, PROV**), ou sont des variables techniques utilisées uniquement pour l'agrégation des données (ex. **MEMBER_ID, MONTH_ID**), évitant ainsi d'introduire du bruit dans l'analyse.

Partie II : Segmentation des clients en groupes homogènes

a. Le Nettoyage

Chargement des bases de données

Chez **Loyalty Air Miles**, chaque client accumule et dépense ses points à travers un programme de fidélisation stratégique. Pour comprendre **comment les membres interagissent avec le programme**, il est essentiel de disposer de **données propres et exploitables**. Nous avons donc entrepris **une phase rigoureuse de nettoyage et de transformation des données**, afin de garantir une **analyse fiable et pertinente**.

Dès le départ, nous récupérons les trois principales sources d'informations nécessaires à l'étude :

- **MEMBERS_DIM** : Contient des informations sur l'ancienneté des membres et leur solde de points.
- **TRANSACTION_FACT** : Recense l'ensemble des transactions effectuées par les clients.
- **REWARD_FACT** : Enregistre les rédemptions et le type de récompenses choisies par les membres.

Ces bases sont **volumineuses** et doivent être nettoyées avant toute exploitation.

Identification et traitement des valeurs manquantes

Avant toute analyse, nous devons **évaluer l'exhaustivité des données** pour éviter toute distorsion dans l'interprétation des comportements des clients. Cette première analyse nous permet de détecter des colonnes avec des valeurs manquantes et d'évaluer leur impact potentiel sur l'analyse.

Correction des valeurs manquantes dans les transactions

Certains enregistrements ne précisent pas le détaillant (RETAILER), ce qui pourrait poser problème lors de l'analyse des habitudes d'achat. Nous attribuons "No Retailer" aux transactions sans détaillant, permettant ainsi de conserver ces transactions dans l'analyse sans distorsion.

D'autres colonnes présentent des valeurs manquantes pour les points accumulés (REWARD_POINTS_EARNED) et les cashbacks (CASH_BACK_POINTS_EARNED). Plutôt que de supprimer ces transactions, nous utilisons **l'imputation par la médiane**, en nous basant sur la catégorie de détaillant (RETAILER). Cette approche garantit que les valeurs imputées reflètent les tendances des autres transactions effectuées chez le même détaillant.

Nous appliquons **la même correction** pour les cashbacks, Ainsi, nous minimisons la perte de données tout en conservant des valeurs cohérentes avec les tendances du programme.

Sélection des variables essentielles

Nous filtrons les variables les plus pertinentes pour notre **analyse comportementale et notre segmentation**.

- Extraction des données transactionnelles : ***MEMBER_ID, MONTH_ID, RETAILER, TRANSACTIONS, AMOUNT_SPENT, BASE_POINTS_EARNED, BONUS_POINTS_EARNED, REWARD_POINTS_EARNED, CASH_BACK_POINTS_EARNED***. Cette sélection nous permet de conserver uniquement les informations clés sur la fréquence des transactions, les dépenses et les points accumulés.
- Extraction des données sur les membres : ***MEMBER_ID, TENURE_MONTHS, CASH_BACK_POINTS_BALANCE, REWARD_POINTS_BALANCE.*** Nous nous concentrons sur l'ancienneté des membres et leur solde de points disponibles, deux indicateurs essentiels de fidélité.

- Extraction des données sur les récompenses : ***MEMBER_ID,MONTH_ID, REWARDS_CATEGORY, REDEMPTIONS, NUMBER_ITEMS_REDEEMED, POINTS_REDEEMED.*** Nous gardons uniquement les données relatives aux échanges de points pour mieux comprendre comment les membres utilisent leurs récompenses.

Dans le cadre de notre analyse des données du programme de fidélité Air Miles, il est essentiel de conserver les valeurs négatives résultant de remboursements ou d'annulations. Ces transactions reflètent des interactions réelles des membres avec le programme et leur exclusion pourrait fausser la compréhension de leur comportement. Par exemple, lors de l'annulation d'une réservation de voyage payée en milles Rêves, les milles sont reversés sur le compte du membre, entraînant une transaction négative. Selon les conditions générales d'Air Miles, il est recommandé de souscrire une assurance annulation/interruption de voyage afin de garantir le reversement des milles dans les comptes d'adhérent en cas d'annulation. Ainsi, intégrer ces valeurs négatives dans notre analyse nous permet de capturer fidèlement l'ensemble des activités des membres et d'adapter nos stratégies en conséquence.

Quelques exemples de variables :

```
 AMOUNT_SPENT         BASE_POINTS_EARNED
 Min.    :-150675.0   Min.    :-30136.00
```

Vérification : Assurer la qualité des données extraites

Avant de passer à l'analyse, nous faisons **une dernière vérification** des valeurs manquantes. Cette étape permet de s'assurer que toutes les imputations ont été correctement appliquées et que nos datasets sont prêts pour l'analyse.

b. Échantillonnage

Face à la **taille importante de la base de données Air Miles**, contenant des millions de transactions et d'interactions des membres, il était essentiel d'**optimiser le traitement des données** sans compromettre la qualité des analyses. Pour cela, nous avons opté pour un **échantillonnage stratégique**, visant à obtenir un **sous-ensemble représentatif** de la population initiale. Cette approche permet de **réduire la charge computationnelle**, tout en conservant la diversité des comportements des membres, afin de garantir une segmentation fiable et pertinente.

```
summary(dt_merged_2)
```

```
   MEMBER_ID          MONTH_ID.x      RETAILER          TRANSACTIONS       AMOUNT_SPENT       BASE_POINTS_EARNED
 Min.   :5.900e+01   Min.   :202201   Length:181860035   Min.   :     0.00   Min.   :-150675.0   Min.   :-30136.00
 1st Qu.:4.997e+06   1st Qu.:202210   Class :character   1st Qu.:     1.00   1st Qu.:     13.6   1st Qu.:     2.00
 Median :1.203e+07   Median :202307   Mode  :character   Median :     3.00   Median :     36.0   Median :     5.00
 Mean   :2.583e+08   Mean   :202306                      Mean   :     7.36   Mean   :    144.3   Mean   :    25.49
 3rd Qu.:1.000e+09   3rd Qu.:202404                      3rd Qu.:     6.00   3rd Qu.:     97.9   3rd Qu.:    15.00
 Max.   :1.017e+09   Max.   :202412                      Max.   :170180.00   Max.   : 677673.8   Max.   :135534.00
 BONUS_POINTS_EARNED REWARD_POINTS_EARNED CASH_BACK_POINTS_EARNED TENURE_MONTHS CASH_BACK_POINTS_BALANCE REWARD_POINTS_BALANCE
 Min.   :-31500.0    Min.   :-29052.00    Min.   :-20973.0        Min.   :  8   Min.   :-64480.0        Min.   : -4280
 1st Qu.:     0.0    1st Qu.:     0.00    1st Qu.:     2.0        1st Qu.:213   1st Qu.:    35.0        1st Qu.:    13
 Median :     0.0    Median :     0.00    Median :     7.0        Median :285   Median :    76.0        Median :   135
 Mean   :    19.8    Mean   :    10.71    Mean   :    34.6        Mean   :272   Mean   :   414.7        Mean   :  1350
 3rd Qu.:     8.0    3rd Qu.:     0.00    3rd Qu.:    26.0        3rd Qu.:342   3rd Qu.:   166.0        3rd Qu.:   855
 Max.   :500000.0    Max.   :285049.00    Max.   :385000.0        Max.   :394   Max.   :467196.0        Max.   :767478
   MONTH_ID.y     REWARDS_CATEGORY    REDEMPTIONS     NUMBER_ITEMS_REDEEMED POINTS_REDEEMED
 Min.   :202201   Length:181860035   Min.   : 1.000   Min.   :  0.000       Min.   :     0.0
 1st Qu.:202210   Class :character   1st Qu.: 1.000   1st Qu.:  1.000       1st Qu.:   100.0
 Median :202307   Mode  :character   Median : 1.000   Median :  1.000       Median :   100.0
 Mean   :202309                      Mean   : 1.272   Mean   :  2.364       Mean   :   370.1
 3rd Qu.:202404                      3rd Qu.: 1.000   3rd Qu.:  2.000       3rd Qu.:   300.0
 Max.   :202412                      Max.   :62.000   Max.   :318.000       Max.   :178440.0
```

```{r}
# Vérifier les colonnes restantes
colnames(dt_sample_merged_1)
```

```
 [1] "MEMBER_ID"              "RETAILER"               "TRANSACTIONS"           "AMOUNT_SPENT"
 [5] "BASE_POINTS_EARNED"     "BONUS_POINTS_EARNED"    "REWARD_POINTS_EARNED"   "CASH_BACK_POINTS_EARNED"
 [9] "TENURE_MONTHS"          "CASH_BACK_POINTS_BALANCE" "REWARD_POINTS_BALANCE" "REWARDS_CATEGORY"
[13] "REDEMPTIONS"            "NUMBER_ITEMS_REDEEMED"  "POINTS_REDEEMED"
```

```{r}
summary(dt_sample_merged_1)
```

```
   MEMBER_ID          RETAILER          TRANSACTIONS       AMOUNT_SPENT       BASE_POINTS_EARNED  BONUS_POINTS_EARNED
 Min.   :5.900e+01   Length:16683395   Min.   :     0.00   Min.   :-150675.0   Min.   :-30136.00   Min.   :-16500.0
 1st Qu.:5.308e+06   Class :character   1st Qu.:     1.00   1st Qu.:     13.7   1st Qu.:     2.00   1st Qu.:     0.0
 Median :1.237e+07   Mode  :character   Median :     3.00   Median :     36.3   Median :     5.00   Median :     0.0
 Mean   :2.738e+08                      Mean   :     7.58   Mean   :    146.2   Mean   :    25.73   Mean   :    20.1
 3rd Qu.:1.001e+09                      3rd Qu.:     6.00   3rd Qu.:     98.7   3rd Qu.:    16.00   3rd Qu.:     9.0
 Max.   :1.017e+09                      Max.   :170180.00   Max.   : 655730.8   Max.   :111761.00   Max.   :318917.0
 REWARD_POINTS_EARNED CASH_BACK_POINTS_EARNED TENURE_MONTHS   CASH_BACK_POINTS_BALANCE REWARD_POINTS_BALANCE REWARDS_CATEGORY
 Min.   :-10463.00    Min.   :-20973.0        Min.   :  8.0   Min.   :-64480.0         Min.   : -4280       Length:16683395
 1st Qu.:     0.00    1st Qu.:     2.0        1st Qu.:206.0   1st Qu.:    36.0         1st Qu.:    12       Class :character
 Median :     0.00    Median :     8.0        Median :268.3   Median :    76.0         Median :   133       Mode  :character
 Mean   :    10.13    Mean   :    35.7        Mean   :268.3   Mean   :   422.1         Mean   :  1320
 3rd Qu.:     0.00    3rd Qu.:    26.0        3rd Qu.:339.0   3rd Qu.:   169.0         3rd Qu.:   839
 Max.   :111761.00    Max.   :318917.0        Max.   :394.0   Max.   :467196.0         Max.   :767478
   REDEMPTIONS     NUMBER_ITEMS_REDEEMED POINTS_REDEEMED
 Min.   : 1.00    Min.   :  0.000       Min.   :     0.0
 1st Qu.: 1.00    1st Qu.:  1.000       1st Qu.:   100.0
 Median : 1.00    Median :  1.000       Median :   100.0
 Mean   : 1.28    Mean   :  2.392       Mean   :   373.7
 3rd Qu.: 1.00    3rd Qu.:  2.000       3rd Qu.:   300.0
 Max.   :62.00    Max.   :318.000       Max.   :178440.0
```

Les moyennes et les ecart-type des variables du dataset initial et de l'échantilllon issu du dataset initial sont sensiblement égaux, Ce qui confirme la répresentativité de notre évhantillon.

Fusion des bases de données : Construire une vue client unique

Le premier défi est de **rassembler toutes les données essentielles** pour **avoir une vision globale de chaque membre**. Ici, nous associons les transactions des membres (TRANSACTION_FACT) à leurs informations de profil (MEMBERS_DIM) pour mieux comprendre leurs habitudes d'achat et leur ancienneté dans le programme. *Nous enrichissons ensuite cette base en intégrant les informations sur les récompenses (REWARD_FACT), ce qui nous permet d'avoir une vision complète des membres : combien ils dépensent, combien de points ils accumulent et comment ils les utilisent.*

Une première **analyse descriptive** est effectuée pour **évaluer la distribution des données** et détecter d'éventuelles anomalies. *Cette étape nous permet d'identifier la répartition des valeurs, les éventuelles incohérences et les données manquantes avant de procéder à la segmentation.*

Sélection d'une période pertinente : Exclure les anomalies post-COVID

L'un des enjeux majeurs de l'analyse est de **capturer des comportements clients représentatifs**. Or, la période COVID-19 a profondément modifié les habitudes d'achat et de fidélité. Nous excluons toutes les transactions antérieures à janvier 2023 pour éviter les biais liés aux restrictions sanitaires et aux perturbations du marché.

Création d'un échantillon représentatif (20%)

Afin d'optimiser l'analyse et réduire la charge computationnelle, nous sélectionnons un **échantillon aléatoire de 20% des membres**, tout en garantissant **la reproductibilité des résultats**. L'utilisation d'un seed garantit que le même échantillon sera sélectionné à chaque exécution du script, assurant ainsi la cohérence des analyses.

Ensuite, nous supprimons les colonnes MONTH_ID.x et MONTH_ID.y car elles ne sont plus nécessaires après la sélection temporelle. Nous réduisons ainsi le bruit dans les données et nous concentrons uniquement sur les variables pertinentes.

Une **dernière vérification** est effectuée pour s'assurer que toutes les colonnes nécessaires sont bien présentes.

Agrégation des données : Construire les indicateurs de segmentation

Une fois l'échantillon prêt, nous consolidons les informations **au niveau du membre** (MEMBER_ID). L'objectif est de créer **des indicateurs comportementaux robustes** pour différencier les profils clients. Cette étape nous permet de résumer les comportements de chaque client à travers des variables clés telles que la fréquence des achats, l'utilisation des points et la diversité des détaillants fréquentés.

Création de nouvelles variables d'analyse comportementale

Pour affiner la segmentation, nous introduisons **des indicateurs avancés**, calculés à partir des données agrégées. Ces variables nous aident à identifier des tendances spécifiques parmi les clients : qui sont les plus engagés, qui utilisent activement leurs points et qui sont plus passifs.

c. **Optimisation des données**

Après avoir soigneusement **nettoyé et échantillonné** les données du programme de fidélité Air Miles, nous entrons dans une phase clé : **l'optimisation et la transformation des variables**. L'objectif est d'**améliorer la qualité des données** avant l'application des modèles de segmentation, en **réduisant l'impact des valeurs extrêmes**, en **rendant les distributions plus homogènes** et en **standardisant les valeurs pour éviter les biais**.

Identifier et classifier les variables : Une approche stratégique

Dans une base de données aussi riche et variée, il est essentiel de **catégoriser les variables selon leur nature et leur rôle dans l'analyse**. Nous excluons MEMBER_ID car il est uniquement utilisé pour identifier les membres et n'a pas de valeur analytique.

Nous classons ensuite les variables en trois grandes catégories :

- **Variables financières (Montants cumulés)** : Ces variables représentent des montants dépensés ou gagnés et sont souvent sujettes à de grandes variations. Ex : "TOTAL_AMOUNT_SPENT", "TOTAL_BASE_POINTS_EARNED", "TOTAL_BONUS_POINTS_EARNED", "TOTAL_REWARD_POINTS_EARNED", "TOTAL_CASH_BACK_POINTS_EARNED", "CASH_BACK_POINTS_BALANCE", "REWARD_POINTS_BALANCE"

- **Variables de ratio et de comportement** : Ces variables mesurent des comportements et des proportions (ex : combien de points sont réellement utilisés par rapport à ceux gagnés ?). Ex : "Points_Utilization_Score", "Cash_Back_Usage_Ratio", "Avg_Spending_Per_Transaction"
- **Variables de fréquence et d'activité :** Elles reflètent l'intensité d'utilisation du programme et la fidélité des membres. Ex : "TOTAL_TRANSACTIONS", "TOTAL_REDEMPTIONS", "TOTAL_ITEMS_REDEEMED", "TOTAL_POINTS_REDEEMED", "Transaction_Frequency", "Redemption_Frequency", "TENURE_MONTHS", "Loyalty_Score"

Réduction de l'impact des valeurs extrêmes : Gérer les outliers

- **Transformation logarithmique des variables financières :** Les **montants et les points accumulés** présentent souvent **des distributions très asymétriques** (quelques membres ont des valeurs très élevées). Pour lisser ces écarts, nous utilisons une **transformation logarithmique modifiée**. Cette transformation permet de **réduire l'effet des valeurs extrêmes** tout en conservant l'interprétabilité des données.
- **Winsorization des ratios et scores :** Les **ratios et scores comportementaux** peuvent être fortement influencés par des valeurs aberrantes. Nous appliquons une **Winsorization**, une méthode qui **ramène les valeurs extrêmes à un seuil prédéfini**. Nous nous assurons que **les ratios restent dans une plage raisonnable**, sans être faussés par quelques cas extrêmes.
- **Standardisation robuste des fréquences et volumes :** Les variables comme **le nombre total de transactions et de rédemptions** peuvent varier énormément d'un membre à l'autre. Pour harmoniser les échelles, nous utilisons **une standardisation robuste basée sur la médiane et l'IQR** (Intervalle Interquartile), qui est **moins sensible aux valeurs extrêmes**. *Cette méthode nous permet de comparer équitablement **les membres ayant des volumes d'activités différents**, sans être influencés par quelques cas extrêmes.*

Avant de passer à la segmentation, nous effectuons **une dernière vérification** pour s'assurer que toutes les transformations ont été correctement appliquées. Nous nous assurons que **les valeurs aberrantes ont été maîtrisées** et que **les données sont prêtes pour l'analyse**.

Standardisation finale des variables numériques

Pour garantir que **toutes les variables ont une importance égale** dans le modèle de clustering, nous appliquons une **standardisation Z-score** sur les variables numériques. Cette transformation assure que **chaque variable a une contribution équivalente dans la segmentation** et empêche que certaines dominent le modèle.

d. Réduction de Dimensionnalité pour la Segmentation

Après avoir **optimisé et standardisé** notre base de données Air Miles, nous faisons face à un **nouveau défi** :

- **Éliminer les corrélations redondantes entre les variables** pour éviter que certaines aient une influence excessive sur la segmentation.
- **Réduire la complexité du modèle** en conservant uniquement les informations les plus pertinentes.

L'**Analyse en Composantes Principales (ACP)** est l'outil idéal pour atteindre ces objectifs. Elle permet d'**identifier les axes les plus significatifs** dans nos données tout en **réduisant le nombre de variables**, sans perdre trop d'informations.

Sélection des variables numériques

Avant d'appliquer l'ACP, nous devons nous assurer que **seules les variables numériques sont utilisées**, car l'ACP ne fonctionne pas avec des variables catégorielles. Nous filtrons uniquement les variables numériques afin de garantir que l'ACP fonctionne correctement.

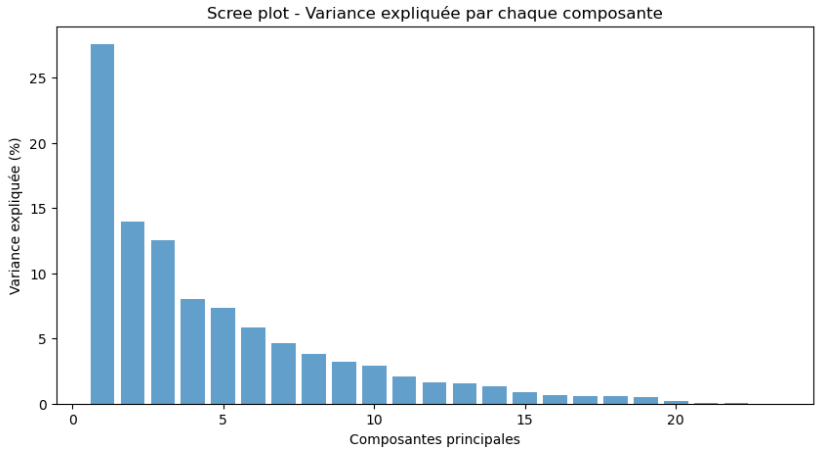
Application de l'ACP : Révéler les axes majeurs des comportements clients

L'ACP consiste à **transformer** nos variables d'origine en **nouvelles dimensions appelées "composantes principales"**, qui capturent **l'essence des comportements des membres**. Nous déterminons le **nombre maximum de composantes** pouvant être extraites, puis nous appliquons l'ACP sur notre dataset standardisé.

Analyse de la variance expliquée : Combien d'informations garder ?

Chaque composante principale représente **un certain pourcentage de la variance totale** des données.
Nous voulons **conserver un maximum d'informations tout en réduisant le nombre de variables**. Nous mesurons combien d'informations chaque nouvelle dimension capture.

Nous affichons **un graphique de la variance expliquée**, aussi appelé **Scree Plot**, pour visualiser le point où l'ajout de nouvelles composantes n'apporte plus de gains significatifs.



Ce graphique nous permet d'identifier le "point de coude" où l'ajout de composantes supplémentaires devient inutile.

Nous résumons les résultats dans un tableau :

	Composante	Variance expliquée (%)	Variance accumulative (%)
0	PC1	27.554260	27.554260
1	PC2	13.993017	41.547278
2	PC3	12.528772	54.076050
3	PC4	8.042491	62.118540
4	PC5	7.364998	69.483538
5	PC6	5.870467	75.354005
6	PC7	4.624004	79.978009
7	PC8	3.822187	83.800196
8	PC9	3.183374	86.983570
9	PC10	2.893204	89.876775

Cette analyse nous montre combien de dimensions sont nécessaires pour capturer **au moins 70% de l'information totale**.

Sélection du nombre optimal de composantes

Nous avons choisi **6 composantes**, car elles expliquent **75.35% de la variance totale**, bien au-dessus du seuil recommandé de 70%. Nous réduisons ainsi nos nombreuses variables en **6 dimensions clés**, tout en conservant l'essentiel des informations.

Création du dataset final avec les nouvelles composantes

Nous créons un **nouveau dataset basé sur ces 6 composantes principales**, qui serviront à la segmentation.

	PC1	PC2	PC3	PC4	PC5	PC6
0	17.761694	9.717008	-4.545914	18.450895	-1.431848	-7.471902
1	14.347248	5.903542	-2.858465	13.678610	-1.828108	-4.044271
2	2.452436	-0.631725	0.372185	-0.144273	0.328590	0.712669
3	1.423233	1.045267	-1.798198	-1.611410	1.317838	0.893853
4	1.152674	-0.801404	0.439103	-0.832140	0.255271	0.957992

Nous obtenons ainsi une base de données optimisée, où chaque membre est décrit par 6 dimensions représentatives de son comportement.

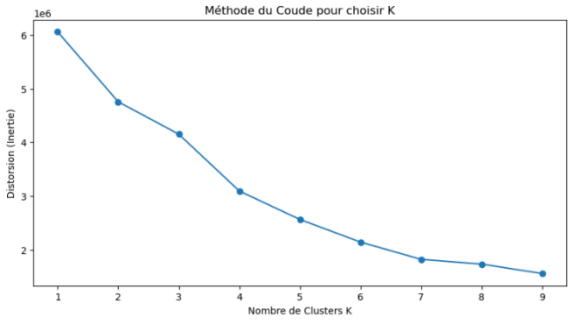
e. Storytelling du Clustering Avancé pour la Segmentation

Après avoir **optimisé et réduit la dimensionnalité des données**, nous entrons dans **la phase clé de la segmentation** : **le clustering**. L'objectif est d'identifier des **profils homogènes** parmi les membres du programme Air Miles, afin de permettre à l'entreprise de **personnaliser ses offres et stratégies marketing**. Cependant, avec un **grand volume de données**, nous avons choisi **une approche hybride et avancée**, combinant :

- **K-Means** pour un premier regroupement rapide des données.
- **Clustering hiérarchique** pour affiner la structure des clusters.
- **K-Means final** pour stabiliser les groupes finaux.

Détermination du nombre optimal de clusters : La Méthode du Coude

Avant d'appliquer un modèle de clustering, il est essentiel de **déterminer le nombre de groupes optimal**. Nous utilisons **la méthode du coude**, qui analyse la **variance intra-cluster** (inertie).



Ce graphe nous permet d'identifier le "point de coude", où l'ajout de clusters supplémentaires n'améliore plus significativement la segmentation.

Application de K-Means pour structurer les données

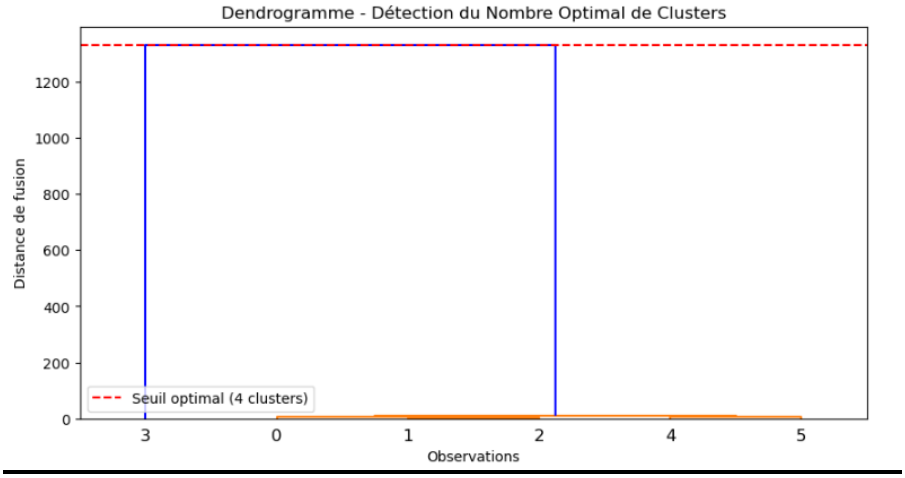
Une fois le nombre optimal de clusters estimé (n_clusters = 6), nous appliquons **K-Means** pour obtenir une première segmentation. *K-Means nous aide à **rapide structurer** les observations en groupes distincts.*

Nous **moyennisons** ensuite les clusters pour obtenir les centroïdes, qui serviront d'entrée au **clustering hiérarchique**. *Nous condensons les caractéristiques des groupes pour mieux comprendre leurs différences.*

Clustering hiérarchique pour affiner les groupes

Le **clustering hiérarchique** nous permet de **visualiser la relation entre les groupes** et d'**ajuster le nombre de clusters**. Nous utilisons la **méthode de Ward**, qui minimise la variance intra-cluster pour obtenir des groupes plus homogènes. Nous analysons ensuite les **distances de fusion** pour déterminer où **casser l'arbre hiérarchique**. Nous détectons

le **nombre optimal de groupes** en identifiant le plus grand saut entre deux fusions. Nous affichons ensuite le **dendrogramme**, qui nous aide à visualiser comment les clusters se forment.



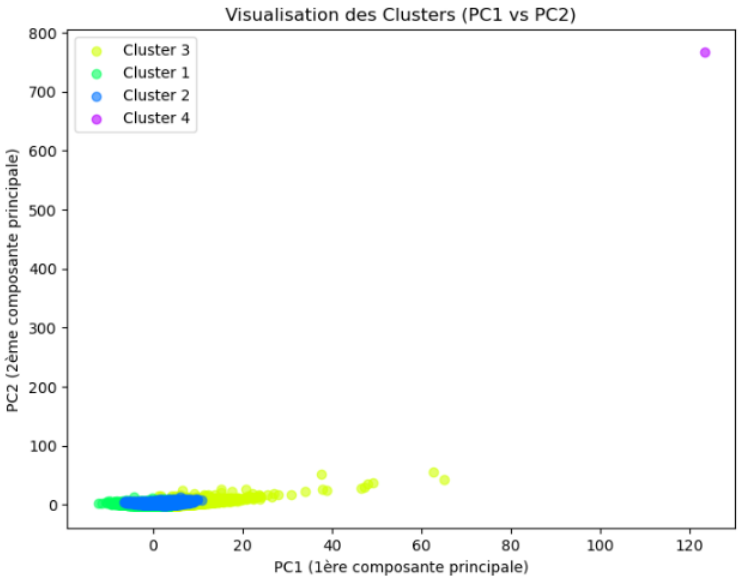
Nous observons **les regroupements naturels des clusters** et confirmons le nombre optimal de groupes **(4)** à utiliser.

Associer les résultats hiérarchiques aux clusters initiaux

Nous assignons chaque observation à son cluster **final** basé sur le clustering hiérarchique. Nous combinons les avantages des **deux méthodes de clustering** pour stabiliser les groupes. Nous vérifions ensuite la **répartition des clusters** :

```
📌 Nombre d'observations dans chaque cluster après Hiérarchique :
Final_Cluster
1    312930
3     20755
2     15959
4         1
Name: count, dtype: int64
```

Nous nous assurons que les clusters sont équilibrés et représentatifs.



Application finale de K-Means pour affiner les segments

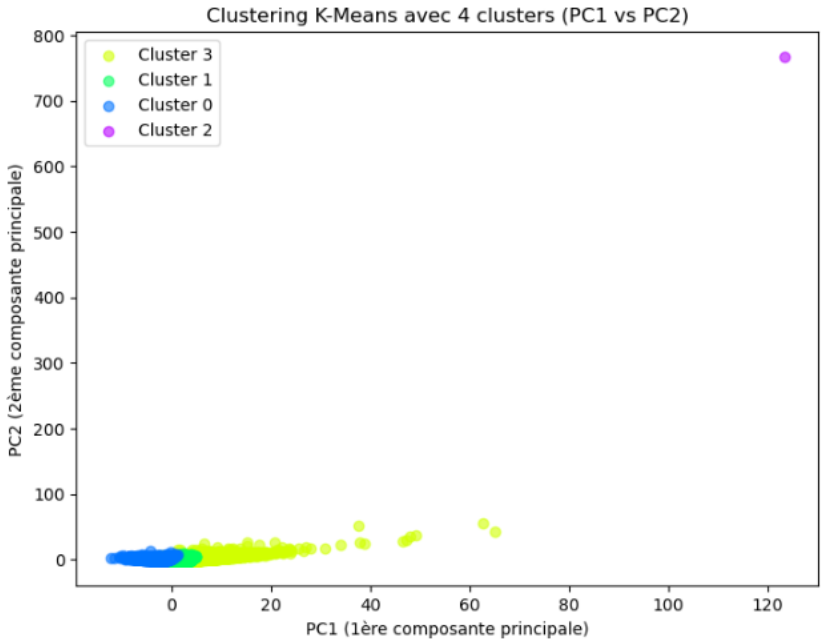
Nous appliquons une **dernière phase de K-Means** avec le nombre optimal de clusters trouvé (k_optimal = 4). Cette étape permet de **fixer les segments finaux** avec une meilleure stabilité. Nous vérifions la répartition finale des clusters :

```
📌 Nombre d'observations dans chaque cluster final après K-Means :
Cluster
1    210803
0    102448
3     36393
2         1
Name: count, dtype: int64
```

Les groupes finaux sont bien formés et prêts à être analysés.

Visualisation des clusters : Comprendre la segmentation

Nous **visualisons les clusters** en utilisant **les deux premières composantes principales** (PC1 et PC2).



*Cette visualisation nous aide à **comprendre comment les clients sont répartis** en fonction de leurs comportements.*

Partie III : Analyse des segments et recommandations stratégiques

Après avoir appliqué **une segmentation avancée basée sur K-Means et le clustering hiérarchique**, nous devons maintenant **interpréter les segments** et comprendre leurs caractéristiques clés.

- L'objectif est de répondre aux questions suivantes : **Quels sont les principaux segments identifiés ?**
- **Quelles sont leurs caractéristiques en termes d'âge, dépenses et fidélité ?**
- **Les variables sociodémographiques influencent-elles significativement cette segmentation ?**

Analyse descriptive des segments : Qui sont les clients Air Miles ?

Nous avons identifié **quatre segments distincts**, avec des caractéristiques bien définies en termes d'âge, de dépenses et d'utilisation des points.

Cluster	Âge Moyen	Interprétation
Cluster 0	56.2 ans	Clients actifs avec une utilisation régulière des points

Cluster 1	59.0 ans	Clients fidèles, mais dépensant de façon modérée
Cluster 2	76.0 ans	Membres très âgés, utilisant peu leurs points
Cluster 3	59.7 ans	Utilisateurs occasionnels du programme

Le segment 2 est significativement plus âgé que les autres, ce qui pourrait suggérer un comportement d'accumulation de points sans rédemption fréquente.

Genre (GENDER)

Le test du Chi-2 révèle que le genre est significativement différent entre les segments (p-value < 2.2e-16).
Implication : Certains segments pourraient contenir plus d'hommes ou de femmes, influençant les choix de récompenses et la fidélité.

Langue (LANGUAGE)

La langue parlée a une influence sur l'appartenance à un segment (p-value < 2.2e-16).
Implication : Les préférences en matière de communication et d'offres pourraient varier entre les francophones et anglophones.

Province (PROV)

Les clients ne sont pas répartis uniformément entre les provinces (p-value < 2.2e-16).
Implication : Certains segments pourraient être plus représentés dans certaines provinces, indiquant des tendances régionales dans l'utilisation du programme.

Préférences de communication : Recevabilité postale et email

Les préférences pour recevoir des offres par e-mail ou par courrier varient selon les segments (p-value < 2.2e-16).
Implication : Certains segments pourraient être plus réceptifs aux campagnes numériques, tandis que d'autres préfèrent des communications traditionnelles.

Statut d'entreprise (SMALL_BUSINESS_FLAG)

Le test montre une forte association entre le statut d'entreprise et l'appartenance à un segment (p-value < 2.2e-16).
Implication : Les petites entreprises utilisent le programme Air Miles différemment des particuliers, avec une potentielle accumulation de points pour des achats professionnels.

Cette analyse, nous avons découvert que **les variables sociodémographiques influencent significativement la segmentation**.

- **L'âge est un facteur clé**, avec un segment (Cluster 2) nettement plus âgé.
- **Le genre, la langue et la province impactent la segmentation**, suggérant des différences culturelles et régionales.
- **Les préférences de communication varient entre les segments**, ce qui permet d'adapter les campagnes marketing.
- **Les petites entreprises ont des comportements distincts**, pouvant nécessiter des offres spécifiques.

offres

Grâce à cette segmentation avancée, Air Miles peut **optimiser son programme de fidélité** en adaptant ses offres, sa communication et ses stratégies d'engagement aux comportements et profils de ses membres. Les **utilisateurs actifs et réguliers** bénéficieront de **bonus sur les transactions fréquentes**, tandis que les **accumulateurs passifs** recevront des **incitations spécifiques pour échanger leurs points**, comme des offres limitées ou des alertes d'expiration. Les **utilisateurs occasionnels** seront ciblés avec des **récompenses instantanées** pour stimuler leur engagement. La segmentation révèle également des **différences significatives selon les variables sociodémographiques** : la province et la langue influencent les préférences d'utilisation du programme, ce qui permet d'adapter les campagnes marketing à chaque région et groupe linguistique. De plus, les **préférences de communication (courrier vs e-mail)** permettront un ciblage plus précis pour maximiser l'engagement. Un **programme dédié aux PME**, incluant des avantages exclusifs et des bonus pour les achats professionnels, aidera à fidéliser cette clientèle spécifique. Enfin, des **stratégies interactives** comme la gamification (défis et récompenses), les **notifications personnalisées**, et une **intégration omnicanal avec les partenaires** viendront renforcer la fidélité des membres et optimiser l'efficacité des campagnes marketing d'Air Miles.