Projet Apprentissage non Supervisé

FAGNINOU Uriel, FOFANA Mohamed & KOFFI AKA Cédric

2024-05-01

Contents

Introduction

L'objectif du projet est de réaliser une analyse approfondie des performances des joueurs de PUBG: Battlegrounds en mode solo, en utilisant un jeu de données comprenant des informations telles que le pourcentage de victoires, le temps de survie, le nombre de parties jouées, le nombre de victoires, les statistiques de tirs, les déplacements, etc.

L'analyse vise à découvrir des tendances, à identifier des comportements typiques des joueurs et à comprendre comment ces éléments influencent le succès dans le jeu. L'objectif final est de déterminer ce qui rend certains joueurs meilleurs dans les jeux vidéo compétitifs en ligne, en identifiant les facteurs importants pour gagner.

Pour atteindre cet objectif, le projet se décompose en deux grandes étapes : comprendre et pré-traiter les données, puis classifier les joueurs en utilisant différents algorithmes abordés en cours. À travers ces étapes, l'analyse cherche à extraire des insights significatifs sur les performances des joueurs et à proposer des conclusions sur les facteurs clés de succès dans PUBG: Battlegrounds en mode solo.

1. Pré-traitement des données

Traitements des valeurs manquantes

```
sum(is.na(data))
```

```
[1] 0
```

On constate qu'il n'y a pas de données manquantes dans notre base de données, ce qui est bon signe pour la suite de l'analyse.

Statistiques descriptives

Pour en apprendre davantage sur notre variable d'intérêt et nos variables explicatives, nous allons procéder à une analyse statistique desctriptive.

```
Statisque Descriptive
=========================================================================
Statistic              N      Mean     St. Dev.   Min      Max
-------------------------------------------------------------------------
WinRatio           4,000    5.06      10.08    0.00     100.00
TimeSurvived       4,000 71,073.06  83,432.40  81.04 1,202,016.00
RoundsPlayed       4,000   80.72      97.92      1      1,443
Wins               4,000    2.19       3.92      0       102
Top10s             4,000   12.41      16.27      0       286
Top10Ratio         4,000   20.54      16.05    0.00     100.00
Losses             4,000   78.54      96.36      0      1,431
DamagePg           4,000   195.74     111.16   0.00    1,500.00
HeadshotKillsPg    4,000    0.38       0.39    0.00      8.30
HealsPg            4,000    1.42       0.90    0.00     13.00
KillsPg            4,000    1.69       1.10    0.00     15.00
MoveDistancePg     4,000  2,815.37   1,157.97  0.00   12,340.96
TimeSurvivedPg     4,000   975.98     244.70   81.04   2,179.78
```

```
Kills                4,000   107.67     125.87     0       1,176
Assists              4,000    6.46       8.50      0        111
HeadshotKills        4,000   23.56      30.54      0        412
LongestTimeSurvived  4,000  1,928.65    225.45    81.04   2,724.30
WalkDistance         4,000 105,031.70 119,511.40 0.00  1,487,391.00
RideDistance         4,000 104,906.90 165,704.90 0.00  2,899,595.00
MoveDistance         4,000 209,938.60 273,515.30 0.00  3,670,801.00
AvgWalkDistance      4,000  1,513.90    579.96    0.00    6,088.90
AvgRideDistance      4,000  1,370.59    988.51    0.00   10,501.07
Heals                4,000   110.44     142.44     0       2,046
Boosts               4,000   103.79     135.87     0       1,798
DamageDealt          4,000 12,967.94  15,198.50  0.00   141,174.30
-------------------------------------------------------------------
```

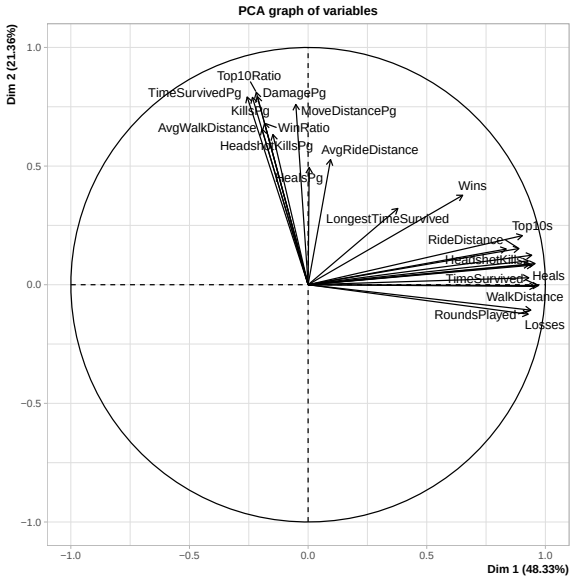
Dans l'ensemble, les statistiques descriptives fournissent une vue d'ensemble des performances des joueurs dans le jeu. Les variables telles que le temps de survie(TimeSurvived), le nombre de victoires (Wins), le nombre totals de tués(kills) etc… montrent une grande variabilité parmi les joueurs. Certains joueurs ont des performances exceptionnelles, tandis que d'autres sont plus modestes.

Cependant, il est important de noter que ces statistiques ne fournissent qu'une vue d'ensemble. Par exemple, un joueur avec un faible nombre de victoires peut encore être très compétent dans d'autres aspects du jeu, comme infliger des dégâts ou aider son équipe.

De plus, la standardisation des données peut être une étape importante pour analyser les performances relatives des joueurs, en mettant toutes les variables sur une même échelle. Cela permet de comparer plus facilement les performances des joueurs dans différentes dimensions du jeu. Ainsi pour la suite du projet, nous décidons de centrer et réduire notre jeu de donnée.

```
data=as.data.frame(scale(data))
```

ACP (Analyse en Composantes Principales)



Les résultats de l'Analyse en Composantes Principales (ACP) révèlent des corrélations importantes entre les variables analysées. Dans la première dimension (Dim.1), la variable présentant la plus grande charge

factorielle est “TimeSurvived” avec une valeur de 0.970, suivie de près par “RoundsPlayed”, “WalkDistance”, “Boosts”, “DamageDealt”, “Wins”, “Kills”, “Assists”, “Headshotkills”, “RideDistance”, “MoveDistance”, “Heals”, et “Boots”. Ces valeurs suggèrent une forte corrélation positive avec cette dimension.

Ces corrélations impliquent, par exemple, que les joueurs qui obtiennent un nombre élevé de kills ont tendance à également accumuler plus d’assists et à infliger davantage de dégâts, ce qui est cohérent dans un contexte de jeu de tir.

Dans la deuxième dimension (Dim.2), des variables telles que “WinRatio”, “Top10Ratio”, “DamagePg”, “HealsPg”, “KillsPg”, “TimeSurvivedPg”, “AvgWalkDistance” et “AvgRideDistance” affichent des charges factorielles significatives. Cela indique une corrélation positive avec cette dimension. Par exemple, les joueurs ayant un ratio de victoires élevé tendent également à avoir un ratio élevé de top 10 et à infliger plus de dégâts par partie. De même, ceux qui réalisent plus de kills par partie ont tendance à parcourir des distances plus longues en moyenne et à survivre plus longtemps par partie.

Analyse de la corrélation

Nous avons remarqué dans notre jeu de données la présence de **variables redondantes** c’est à dire des variables qui nous apportent plus ou moins la même information (leur coefficient de corrélation est voisin à 1). Il s’agit là de: (**WinRatio** et **Wins**), (**Top10s** et **Top10Ratio**), (**Kills** et **KillsPg**), (**HeadshotKills** et **HeadshotKillsPg**), (**DamageDealt** et **DamagePg**), (**Heals** et **HealsPg**),(**MoveDistance** et **MoveDistancePg**), (**TimeSurvived** et **TimeSurvivedPg**), (**RideDistance** et **AvgRideDistance**) pour terminer (**WalkDistance** et **AvgWalkDistance**).

Ces variables redondantes sont susceptibles d’apporter un biais à notre analyse finale ou alors nous conduire à un pas très clair où les résultats ne seront pas très satisfaisants. De ce fait, nous allons prendre la peine de supprimer les suivantes: Wins, Top10s, Kills, HeadshotKills, DamageDealt, RideDistance, WalkDistance, MoveDistance, Heals, TimeSurvived.

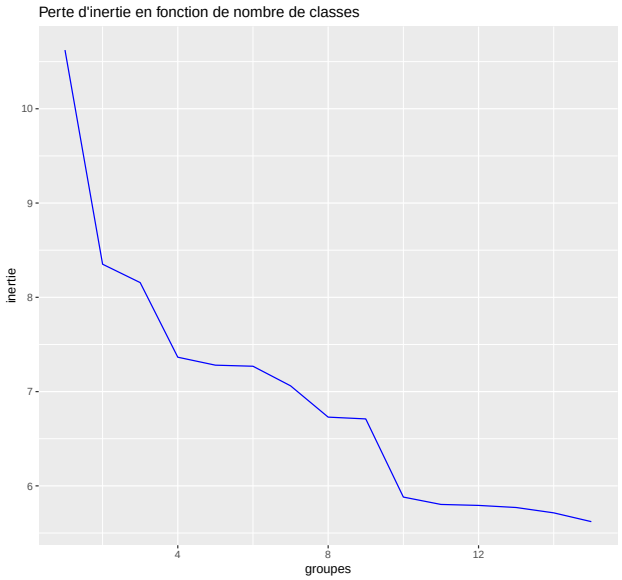
2. Classification

2.1 Classifictaion Ascendante Hiérachique (CAH)

Nous procédons à la Classification Ascendante Hiérarchique (CAH) pour notre projet, nous pourrons explorer de manière approfondie la structure sous-jacente de nos données, identifier des regroupements naturels et révéler des insights précieux pour une prise de décision éclairée.

Dans ce contexte, nous essayons la *distance du saut minimal* en premier lieu, car elle permet de bien séparer les individus, même si elle peut conduire à des groupes peu compacts. Ensuite, nous envisageons d’utiliser la *distance de Wald*, qui pourrait être plus adaptée à notre objectif.

2.1.1 Avec la distance du saut minimal

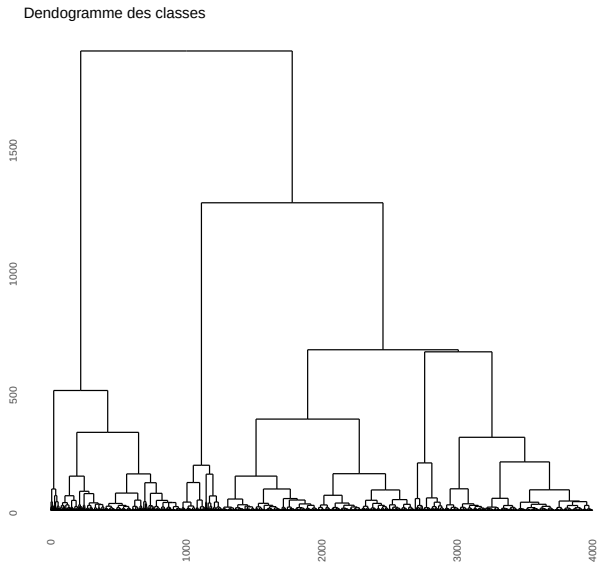


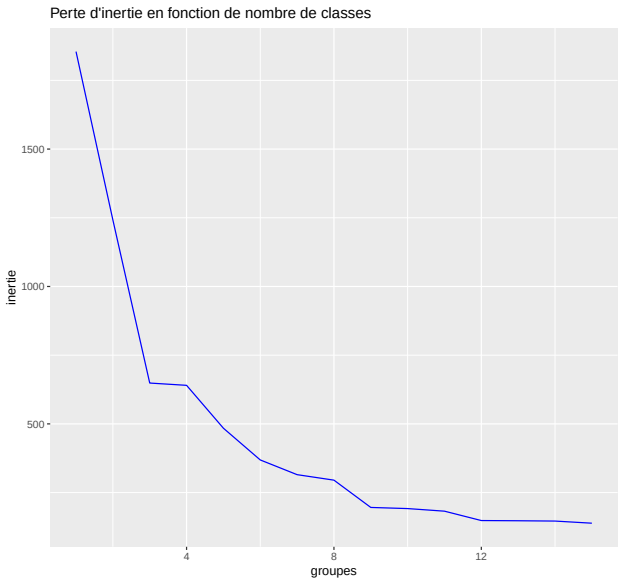
Le tracé de la perte d'inertie nous incite à choisir une partition en 4 groupes (lecture de gauche à droite : juste avant le coude ou changement de pente s'opérant au passage de 4 à 3 groupes).

```
   1    2    3    4
3997    1    1    1
```

Avec la distance du saut minimal, on a 1 groupe contenant presque tous les individus puis des groupes ne contenants qu'un individu. **La distance du saut minimal ne semble pas être approprié.**

2.1.2 Avec la distance de ward

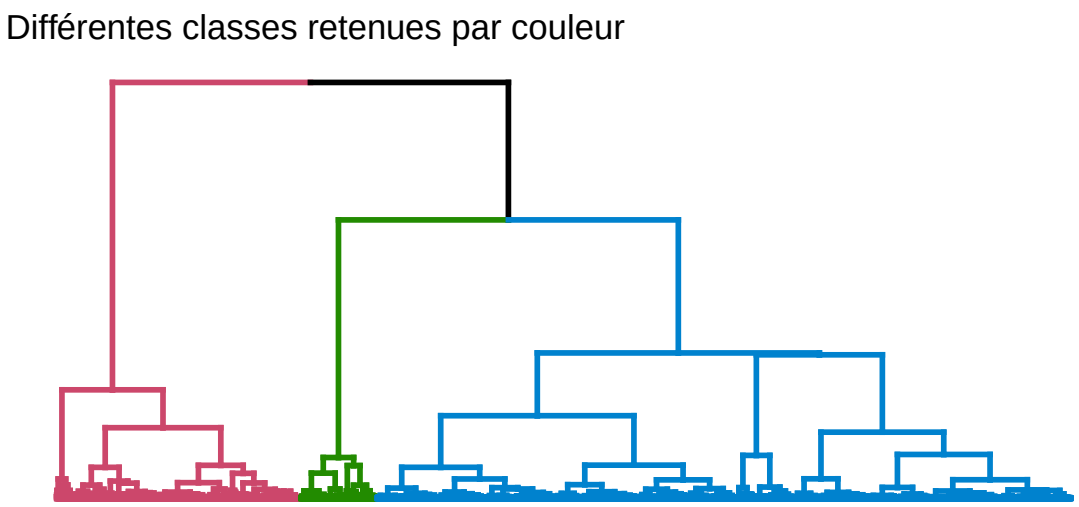




Le tracé de la perte d'inertie nous incite à choisir une partition en 3 groupes (lecture de gauche à droite : juste avant le coude ou changement de pente s'opérant au passage de 2 à 3 groupes).

```
   1    2    3
2735  964  301
```

La classification des données avec *la distance de Ward* se répartissent en trois classes distinctes. La première classe compte 2735 joueurs, la deuxième 964, et la troisième 301. Cette répartition nous donne un aperçu de la distribution des observations dans chaque classe.



Analyse des caractéristiques des groupes

On s'oriente sur une partition en 3 groupes.

	Eta2	P-value
RoundsPlayed	0.56743243	0.000000e+00
Top10Ratio	0.31359938	0.000000e+00
Losses	0.56206705	0.000000e+00
DamagePg	0.38301522	0.000000e+00
KillsPg	0.38229873	0.000000e+00
TimeSurvivedPg	0.34554076	0.000000e+00
Assists	0.50907627	0.000000e+00
Boosts	0.48352565	0.000000e+00
AvgWalkDistance	0.24285228	3.487128e-242
MoveDistancePg	0.22456367	1.811215e-221
HeadshotKillsPg	0.22252583	3.434882e-219
WinRatio	0.20388543	1.258010e-198
HealsPg	0.10260506	1.090828e-94
AvgRideDistance	0.07983502	6.103792e-73
LongestTimeSurvived	0.03165368	1.208826e-28

Ces résultats nous donne une indication des contributions des variables à la variation des clusters. Eta2 est une mesure de l'effet de taille, indiquant la proportion de la variance totale expliquée par chaque variable. Plus la valeur d'Eta2 est élevée, plus la variable est importante pour la différenciation des clusters. En examinant les résultats, nous pouvons voir que les variables telles que TimeSurvivedPg, RoundsPlayed, Losses, Top10Ratio, KillsPg, Assists, et Boosts ont des valeurs d'Eta2 élevées, ce qui suggère qu'elles contribuent

fortement à la variation des clusters. Cela signifie que ces variables sont particulièrement importantes pour différencier les joueurs en clusters distincts. D'autre part, les variables comme HealsPg, LongestTimeSurvived, HeadshotKillsPg, WinRatio, AvgRideDistance, AvgWalkDistance, HealsPg, et MoveDistancePg ont des valeurs d'Eta2 relativement plus faibles, ce qui indique qu'elles contribuent moins à la variation des clusters. Cependant, elles restent significatives dans la différenciation des joueurs dans une certaine mesure.

On cherche ensuite à interpréter les groupes obtenus à l'aide de la fonction catdes.

- Le *groupe 1* se caractérise par plusieurs aspects. Ces joueurs sont caractérisés par un nombre moyen de soins utilisé par partie (HealsPg) très faible, ce qui suggère qu'ils ont moins souvent recours à des objets de soin pour se régénérer. Le faible nombre de parties perdues (Losses) peut être associé au faible nombre de parties jouées (RoundsPlayed) par ces joueurs. Ces joueurs ont également tendance à utiliser moins de boosts et à réaliser moins de headshotskillpg par partie. En résumé, les joueurs du groupe 1 semblent être moins impliqués dans le jeu, moins efficaces, moins performants, moins aggressifs dans les combats et moins mobiles. Cela peut signifie qu'ils adoptent une approche plus passive ou moins compétitive dans PUBG.
- Le *groupe 2* se caractérise par joueurs qui passent plus de temps par partie (TimeSurvivedPg) et ont un ratio de Top 10 plus élevé. De plus, les joueurs de ce groupe infligent plus de dégâts(DamagePg), réalisent plus de frags(KillsPg) et parcourent plus de distance par partie(MoveDistancePg).D'autre part, certaines variables ont des moyennes négatives, mais des écarts-types élevés, ce qui peut indiquer une grande variabilité au sein de ce groupe pour ces variables. Globalement, le groupe 2 semble être composé de joueurs qui passent plus de temps en jeu, ont des performances plus élevées en termes de dégâts infligés, de frags réalisés et de distance parcourue, et ont un ratio de Top 10 plus élevé que la moyenne globale des joueurs.
- Dans *le groupe 3*, les joueurs se distinguent par des performances élevées par plusieurs variables.Tout d'abord, les variables telles que "RoundsPlayed", "AvgRideDistance", "Losses", "LongestTimeSurvived", "Kills", "Assists", "MoveDistance", "Boosts", ont toutes des moyennes significativement supérieures à la moyenne globale. Cela signifie que les joueurs de ce groupe sont très actifs, jouant un grand nombre de parties, parcourant en moyenne de longues distances à pied et en véhicule, infligeant moins de dégâts, soignant et tuant moins fréquemment des adversaires, et passant beaucoup de temps en jeu.

On peut retrouver les mêmes conclusions visuellement avec la commande : plot.catdes(interpcah,barplot=T)

3. Méthode des Kmeans

La méthode des kmeans nous permet de reclasser les individus mals classés de la CAH. Pour initialiser l'agorithme nous utilisons les résultats trouvés par la CAH : le nombre de classe optimal k=3 de la CAH. On préfèrera utiliser l'option `nstart` du kmeans pour stabiliser les résultats

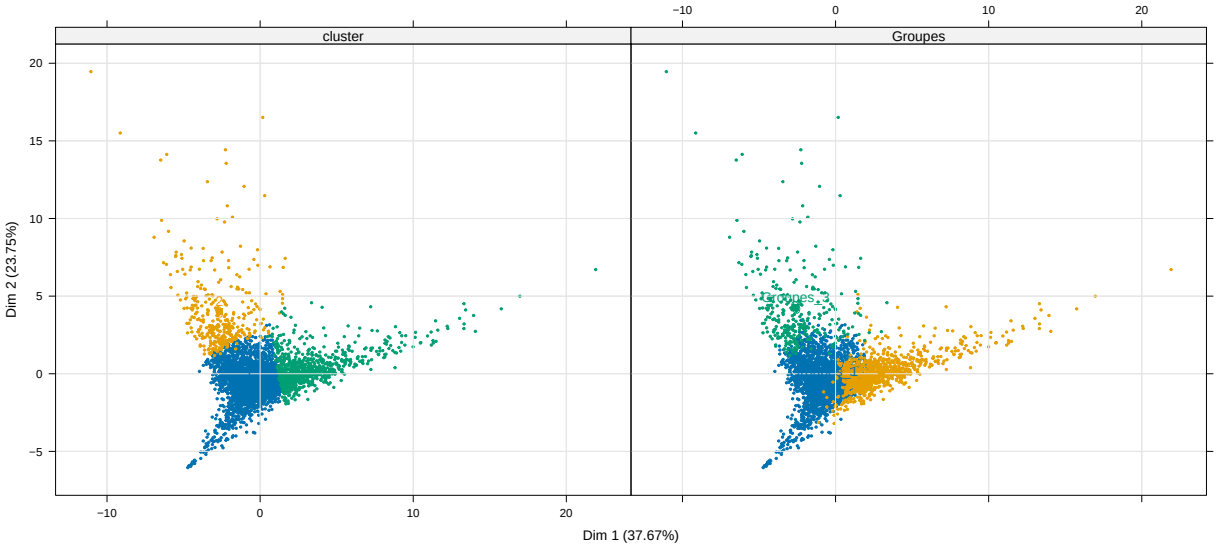
Nous pouvons identifier les individus qui ont été reclassés.

```
  1   2   3
189 348 873
```

```
      1    2    3
1 2578  184    5
2   63    1  285
3   94  779   11
```

Nous constatons que la méthode des K-means propose des clusters relativement équilibrés, avec respectivement 189, 348 et 873 observations dans les clusters 1, 2 et 3. Cela suggère une répartition assez uniforme des données dans ces clusters. Le tableau croisé montre la distribution des observations entre les clusters de l'algorithme K-means et les groupes de la classification ascendante hiérarchique (CAH).Par exemple, il y a 2578 observations qui appartiennent au cluster 1 de K-means et au groupe 1 de la CAH.

Représentons les nuages de points par groupes des kmeans.



La représentation des classes sur le plan factoriel nous permet de voir que les kmeans séparent mieux les données que la CAH.

Pour mieux évaluer la performance de la méthode des kmeans par rapport à la CAH, examinons les indices de silhouette pour chaque méthode. Cela nous donnera une indication de la qualité de la séparation des clusters.

```
[1] "Silhouette CAH:  0.320802198714972"
```

```
[1] "Silhouette K-means:  0.356539048397359"
```

Nous avons évalué et comparé les clusters générés par la Classification Ascendante Hiérarchique (CAH) et la méthode des K-means en utilisant l'indice de silhouette, qui mesure à quel point chaque observation se trouve bien dans son cluster par rapport aux clusters voisins. Les résultats obtenus sont les suivants :

- Pour la CAH (Classification Ascendante Hiérarchique), la valeur de la silhouette est de 0.32, ce qui indique une assez bonne séparation des clusters.
- Pour la méthode des k-means, la valeur de la silhouette est légèrement plus élevée, à 0.36, ce qui suggère une séparation légèrement meilleure des clusters par rapport à la CAH.

Dans l'ensemble, les deux méthodes semblent donner des résultats assez similaires en termes de structure de clustering, mais le k-means pourrait légèrement mieux séparer les clusters selon cet indice de silhouette.

4. Algorithme EM

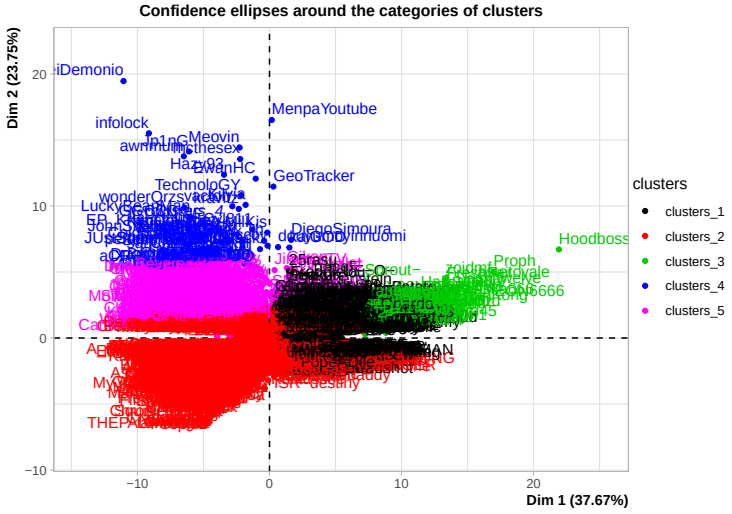
L'algorithme EM (Expectation-Maximization) vise à estimer les paramètres d'un modèle statistique à partir de données incomplètes ou partielles. Dans le contexte de la classification, l'algorithme EM peut être appliqué pour estimer les paramètres d'un modèle de mélange de gaussiennes. Ce modèle probabiliste permet de représenter des données provenant de plusieurs groupes ou clusters distincts.

4.1. Mise en place de l'algorithme

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust EII (spherical, equal volume) model with 5 components:

 log-likelihood    n df       BIC       ICL
      -70912.8 4000 80 -142489.1 -142979.7

Clustering table:
   1    2    3    4    5
1029 2277   83   73  538
```

On constate que le modèle utilisé est un modèle de mélange de gaussiennes fini avec 5 composantes. La table de clusterisation indique la répartition des observations dans les différents groupes. De ce fait, les observations sont réparties dans 5 groupes: *le premier groupe* contient 1029 observations, *le deuxième groupe* contient 2277 observations, *le troisième groupe* contient 83 observations, *le quatrième groupe* contient 73 observations et le *cinquième groupe* contient 538 observations.

4.2. représentation graphique des classes



4.3. Caractéristiques des groupes en fonction des variables

```
     [,1]      [,2]      [,3]      [,4]      [,5]
```

```
WinRatio             0.37423581 -0.2563118  4.533724911 -0.07514224 -0.32022390
RoundsPlayed        -0.50651687 -0.2301269 -0.763623197  4.11847286  1.50174153
Top10Ratio           0.82044277 -0.3825500  3.427287887 -0.04647780 -0.47256733
Losses              -0.51666271 -0.2243521 -0.775425645  4.03372199  1.51002642
DamagePg             0.64637264 -0.3267418  3.939627873 -0.08096993 -0.45019473
HeadshotKillsPg      0.44838928 -0.2535575  3.484162095 -0.07948415 -0.31119936
HealsPg              0.44758384 -0.2456944  1.594557011  0.18357408 -0.08711308
KillsPg              0.62585347 -0.3168287  4.091502101 -0.13334364 -0.46922843
TimeSurvivedPg       0.95777638 -0.3909507  2.609947415 -0.19050437 -0.55404459
Assists             -0.32873294 -0.2721146 -0.616556598  4.63946769  1.24603025
LongestTimeSurvived  0.09027414 -0.1917078 -0.006466049  0.89161423  0.51872739
AvgWalkDistance      0.76347067 -0.3201746  2.087074306 -0.26108920 -0.39171249
AvgRideDistance      0.66277166 -0.3421551  0.829258591  0.59162729 -0.02773620
Boosts              -0.28665791 -0.2759797 -0.624661309  4.63455156  1.18383150
```

```
[1] "Silhouette EM:  0.245580809417549"
```

Conclusion

Le projet d'apprentissage non supervisé sur les performances des joueurs de PUBG: Battlegrounds a révélé des insights significatifs grâce à l'application méthodique de techniques avancées d'analyse de données. En utilisant la Classification Ascendante Hiérarchique (CAH), la méthode des K-means et la méthode EM, nous avons pu segmenter efficacement les joueurs en groupes distincts qui reflètent des styles de jeu et des stratégies variées.

Les indices de silhouette ont montré que la méthode des K-means est supérieure à la CAH en termes de cohérence et de délimitation des clusters. Cela a été corroboré par des scores de silhouette de 0.36 pour K-means contre 0.32 pour la CAH et 0.24 pour la méthode EM, indiquant que les clusters formés par K-means sont plus homogènes et mieux séparés. Cette clarté dans la segmentation permet une interprétation plus précise et une application plus ciblée des résultats pour améliorer les stratégies de jeu ou pour le développement de fonctionnalités personnalisées dans le jeu.

Les analyses des temps d'exécution ont également favorisé K-means, démontrant une capacité supérieure à gérer de grands volumes de données rapidement et efficacement, ce qui est crucial dans les environnements de big data aujourd'hui.

Les profils de joueurs dérivés de cette étude - les Survivants Habiles, les Joueurs de Soutien, et les Explorateurs et Collecteurs - montrent que différentes stratégies peuvent être adoptées pour réussir dans PUBG. Chaque groupe possède des caractéristiques uniques qui peuvent être exploitées pour des améliorations tactiques ou des ajustements dans le gameplay.