NGUYEN Manh Hung | HANNAOUI Youness | PHAM Thi Minh Ngoc

PROJET CLASSIFICATION NON-SUPERVISÉE

2021/2022

Classification de l’univers des Pokémons

I. Prise en main du jeu de données

I.1. Compréhension du jeu de données

La première étape fondamentale pour mener à bien une étude est de comprendre parfaitement les données dont nous disposons.

Le nom Pokémon n'est pas étranger à l'enfance de beaucoup de gens. Les histoires de ces dessins animés ont toujours un attrait étrange pour les enfants et les adultes et deviennent l'une des fiertés du Japon. Avec n'importe quelle génération de jeu, les joueurs doivent également entraîner des Pokémon pour capturer et entraîner d'autres Pokémon afin qu'ils deviennent une armée puissante, puis se battre avec les Pokémon des autres joueurs pour gagner. Ce sont les 3 thèmes principaux du jeu Pokémon : la capture, l'entraînement et le combat incluant les jeux vidéo, les animés, les séries manga ou encore le jeu de cartes à collectionner Pokémon.

Il existe de nombreux types de Pokémon aux caractéristiques différentes, il sera donc intéressant d'analyser leur classement en fonction de leurs données.

Il faut savoir que cette base de données est constituée de 51 colonnes et 1045 lignes. Pour autant, ce nombre de lignes ne correspond pas au nombre d'individus analysés, certains individus apparaissent plusieurs fois, car ils ont plusieurs formes ou une évolution supplémentaire dite "méga". Comme on ne pourra garder qu'une unique forme pour chacun, on a choisi le premier individu pour chaque forme.

Nous remarquons qu'il existe à la fois des variables quantitatives et qualitatives basées sur les caractéristiques du pokémon. Parmi ces 50 variables, nous n'avons retenu que celles à des fins de classification, qui seront présentées plus en détail dans la section suivante.

Nous avons également remarqué que de nombreuses colonnes manquaient de données, nous devions donc les traiter avant de pouvoir procéder à l'analyse.

I.2. Nettoyage des données

Il est primordial de nettoyer les données. En effet, négliger cette étape pourrait biaiser nos résultats et notre classification finale pourrait se révéler fausse. Commençons par les valeurs manquantes, nous devons vérifier s'il y en a et si c'est le cas les gérer afin de ne pas biaiser notre étude. Il s'avère que nous n'avons aucune valeur manquante à cette étape de nettoyage. Ensuite, nous vérifions si nous avons des valeurs aberrantes, si tel est le cas, nous allons devoir traiter ces valeurs en les retirant par exemple.

On commence par repérer puis supprimer les colonnes qui possèdent le plus des valeurs manquantes (> 100):
"type_2","ability_2","ability_hidden","base_friendship","base_experience", "egg_type_2","percentage_male".

Les valeurs manquantes sur le jeu de données sont liées directement avec la définition des différents Pokémons. Les valeurs manquantes dans la variable "type_2" viennent du fait que les Pokémons possèdent parfois deux types et d'autres fois un seul types, par exemple le Pokémon Charmander est définie sur pokebip avec un seul type qui est Fire par contre le pokémon Bulbasaur est définie avec deux types qui sont Grass et Poison.

Alors, pour les colonnes, nous supprimons 7 variables, pour les lignes, nous gardons une seule instance de chaque type, et supprimons également les lignes qui contiennent des valeurs manquantes, nous nous retrouvons avec un tableau de données avec 898 individus et 44 variables.

I.3. Choix des variables dans le cadre d'une classification

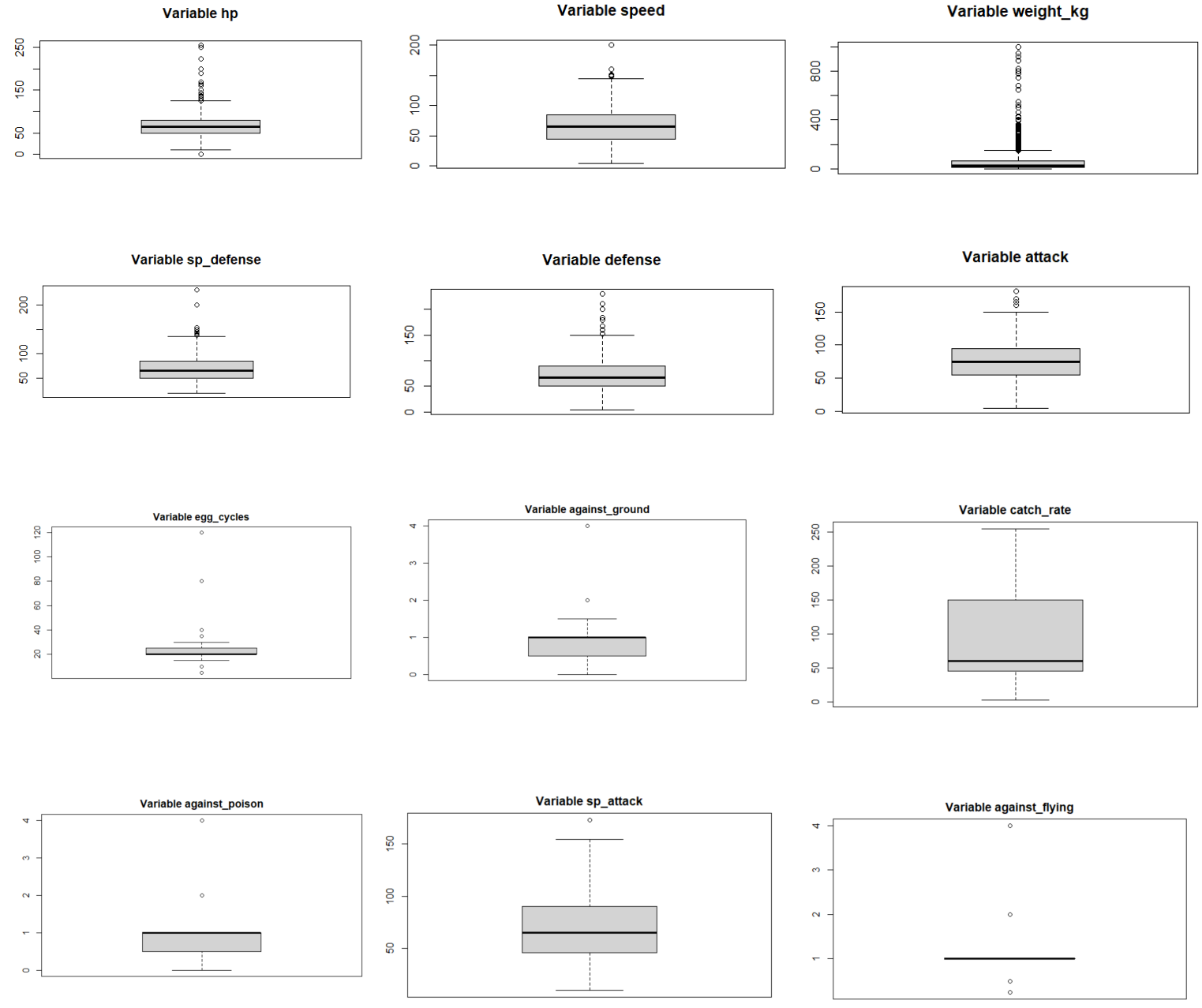
Afin de pouvoir réaliser la classification, on va sélectionner les variables quantitatives les moins corrélées entre eux et qui semblent apportées le plus d'informations concernant les compétences des Pokémons en combat, des

caractéristiques de leur poids et de leur taille, de plus, nous avons trouvé des informations telles que la capturabilité et des informations sur les œufs très intéressantes, alors nous allons donc faire des recherches là-dessus :
"hp","attack","defense","sp_attack","speed","weight_kg","sp_defense",
"catch_rate","egg_cycles","against_poison","against_ground","against_flying"

Sachant que les Pokémons peuvent effectivement être classés très soigneusement et dans de nombreux groupes, nous avons essayé d'expérimenter sur des variables pour pouvoir regrouper le plus de groupes possibles. Après de nombreuses expérimentations et sélections de variables, nous avons décidé que le groupe de variables ci-dessus donnerait les meilleurs résultats.

I.4. Standardisation des données

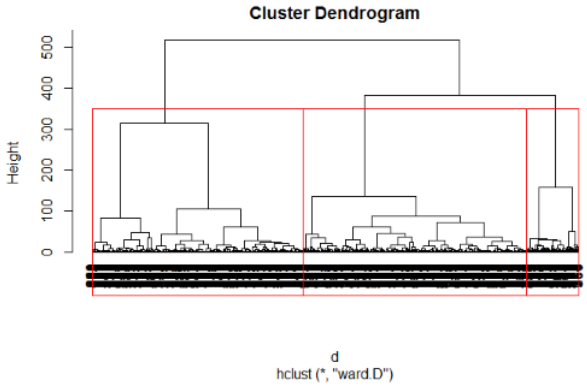
Les boxplots des 12 variables choisis montre en général l'existence de beaucoup de variables aberrantes surtout pour les variables "hp", "weight kg" et "defense". De plus, les unités de variables "against_poison", "against_ground", "against_flying" sont différentes des autres, ce qui nous obligent à standardiser les données en les centrant puis les réduisant. Quelques distributions avec des valeurs aberrantes:

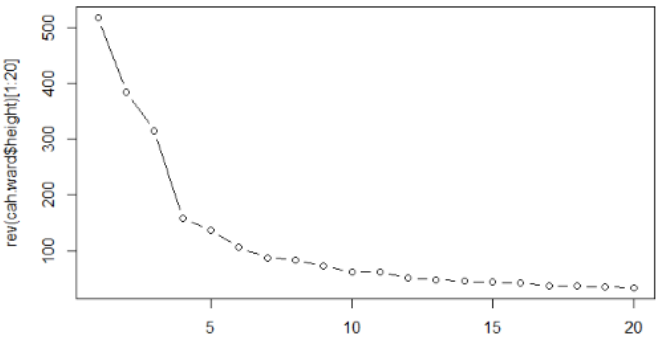


II. Classification des individus

II.1. Application des algorithmes de classification CAH et K-means

Après le nettoyage et la standardisation de nos données, on va mettre en place une classification des Pokémons sélectionnées en se basant premièrement sur la stratégie de Ward de la CAH pour avoir une idée sur le nombre de groupes qu'on peut choisir, puis on va appliquer l'algorithme de K-means.





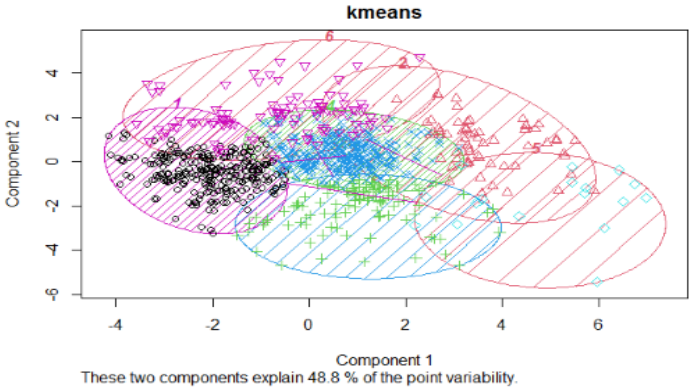
On remarque une perte d'inertie rapide entre le premier et le troisième saut, nous remarquons qu'une classification en trois groupes est possible.

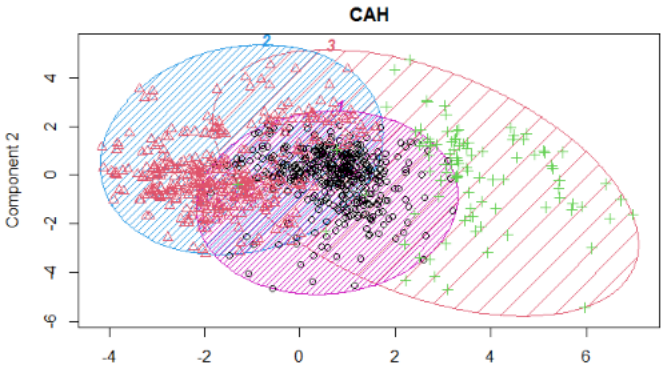
En remarque que le dendrogramme au-dessus, qui représente les partitions obtenues avec ces trois groupes ne sont pas de même proportion. Deux groupes, c'est proche au niveau de la taille, puis un autre plus petit.

Après avoir une idée générale sur le nombre de groupes à choisir avec la stratégie de Ward en se basant sur l'inertie intra-classe, on va appliquer l'algorithme de K-means pour requalifier les données qui auraient été mal placées.

II.2. Classification finale

Comparaison **K-means** et **CAH**:





Nous remarquons dans les représentations factorielles de **CAH** que l'ensemble des groupes se mélangent. Même le groupe noir s'est mêlé à tous les autres groupes. Ainsi, nous constatons que plus de la moitié des individus de chacun des trois autres groupes sont mélangés et aucun groupe n'est reconnaissable. Cette méthode nous semble donc inappropriée dans le cadre de notre étude.

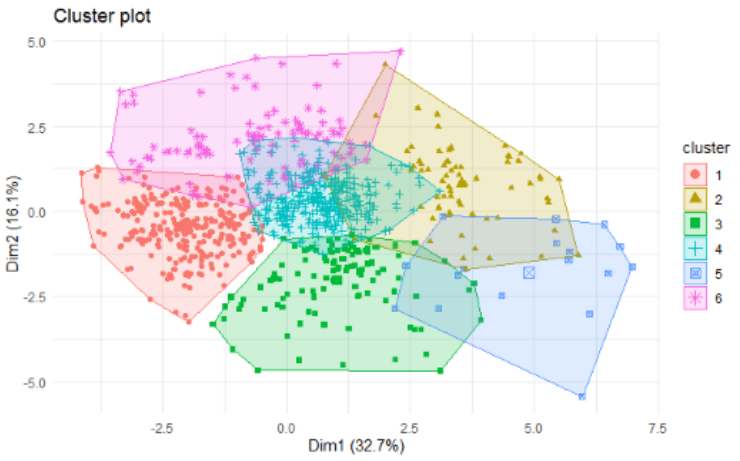
Nous avons classifié avec la méthode des **K-means**. En ce qui concerne les représentations factorielles des individus, nous observons que six groupes sont distincts, ceux-ci se différencient sur la combinaison de l'axe 1 et de l'axe 2.

Cette méthode produit une représentation factorielle de groupes mieux que **CAH**, ces résultats seront présentés par la suite.

III. Etude des groupes obtenus

III.1. Représentation des résultats

On observe sur ce clusplot obtenu via la méthode **K-means**, qu'on a bien six groupes distincts.



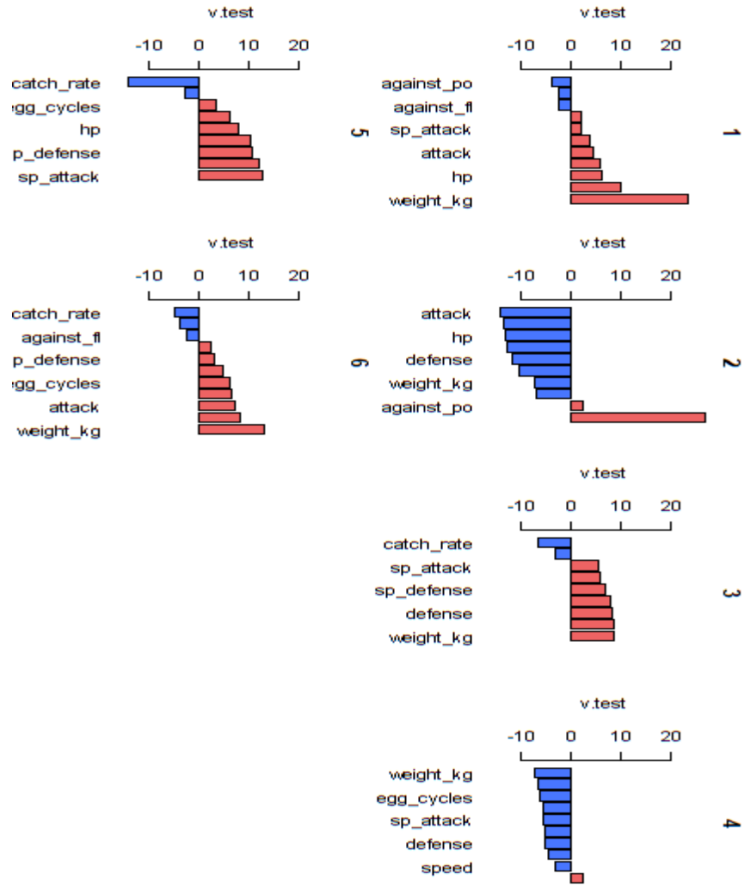
On notera qu'avec nos données, 37,2 % des variations sont portées par les deux premiers axes principaux. Non seulement les positions, ce clusplot nous montre également la densité individuelle de Pokémon de chaque groupe.
Nous observons que pour des groupes de même niveau sur l'axe de la composante 1, ils seront séparés sur l'axe 2, et inversement, c'est pourquoi nous observons que cette division est raisonnable:
Par exemple, sur l'axe 1, avec les trois groupes 1 (rouge), 3 (vert), 5 (bleu), leurs clusters sont séparés même s'ils ont le même degré sur la deuxième composante. L'axe 1 nous aide également à séparer le groupe 6 (violet) et le groupe 2 (jaune).
Sur l'axe 2, les groupes 6 (violet), 5 (turquoise) et 3 (vert) occupent des positions différentes, alors, les Pokémons de ces trois groupes ont trois niveaux différents en fonction de la caractéristique représentée sur l'axe 2. L'axe 2 nous aide également à séparer le groupe 3 (vert) et le groupe 2 (jaune) par exemple, même s'ils ont le même degré sur la première composante.

III.2. Caractérisation des partitions obtenues

On cherche à savoir ce qui caractérise les six groupes que l'on a obtenus afin de déterminer les critères qui les ont construits.

Nous mettrons en évidence les caractéristiques intéressantes de chaque groupe comme suit:

Cluster 1: On peut dire que ce sont de très bons guerriers et qu'ils ont un avantage en termes d'attaque. Pour un match où la défense n'est pas très importante, ces Pokémon conviennent. Si vous avez besoin de choisir un guerrier capable de prendre des coups contre ses adversaires et de vivre longtemps (variable *hp*), c'est un excellent choix.



Cluster 2: Le groupe a globalement les statistiques les plus faibles. Ils sont particulièrement faibles en attaque et en défense. Mais ils ont un très bon avantage de pouvoir frapper des Pokémon de type eau. Nous pensons donc que ce n'est pas le groupe le plus faible.

Cluster 3: Un groupe de Pokémon avec une puissance d'attaque assez uniforme et une variété de compétences, y compris l'attaque spéciale et la défense spéciale. Nous considérons qu'il s'agit du deuxième groupe le plus fort, car comparés aux groupes 1 et 6, ils ont des capacités de combat similaires, mais ce groupe 3 a une meilleure défense spéciale.

Cluster 4: Les pokémons de ce groupe ont tous des fonctions faibles, même si le groupe 2 a également des statistiques faibles, ils sont toujours les meilleurs pour vaincre les Pokémons de type eau, nous supposons donc que le groupe 4 est l'espèce qui combat le plus mal les Pokémons car ils n'ont pas de statistiques particulières meilleures que les autres clusters.

Cluster 5: Ils sont considérés comme des Pokémon le plus puissants et nous trouvons que ce groupe est le plus complet en termes de compétences de combat. Les capacités de défense et d'attaque des Pokémon appartenant à ce groupe sont supérieures aux autres groupes. Si vous avez besoin de guerriers avec des compétences spéciales pour vaincre vos adversaires, c'est le choix parfait.

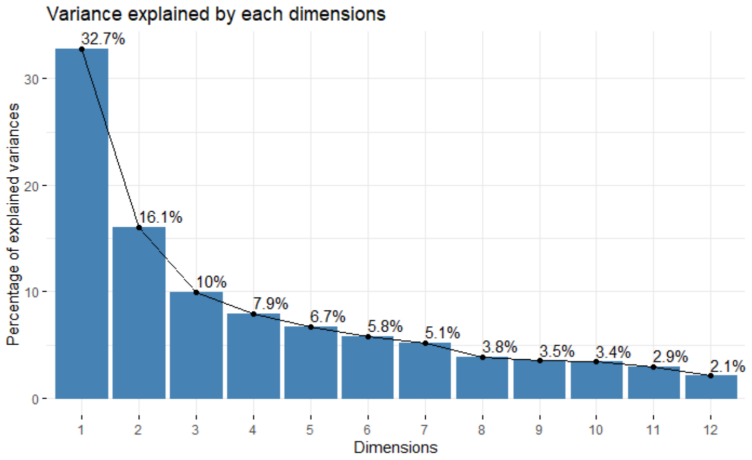
Cluster 6: Ce groupe de Pokémon est classé parmi les trois plus forts car ils ont de bonnes capacités de défense et d'attaque, c'est un groupe de Pokémon avec un avantage de poids et il a un meilleur avantage défensif que le groupe

1, c'est pourquoi nous l'avons mis dans le groupe des trois Pokémon les plus forts (les trois groupes les plus forts sont les groupes 3, 5, 6).

III.3. Représentation informative des résultats

On cherche maintenant à représenter les informations détenues par les données. Pour cela, on va utiliser une analyse à composantes principales.

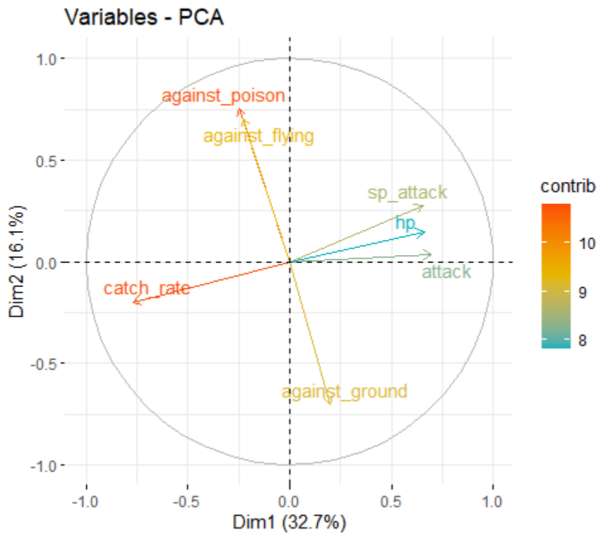
III.3.1 Représentation sur les axes principaux d'une ACP



Le graphique ci-dessus nous permet d'observer que 48,8% et 58.8% des variances sont expliquées et conservées par les deux et trois premiers axes principaux respectivement.ce qui représente un taux important , sachant qu'on travaille sur un espace de dix dimensions.

On peut donc utiliser ces axes afin de pouvoir représenter au mieux les données.

III.3.2 Analyse de corrélation des variables avec les axes principaux

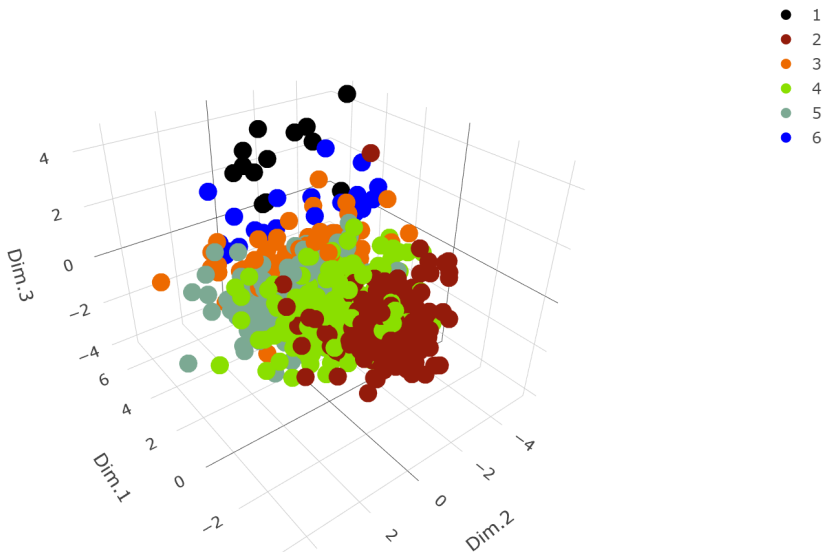


On remarque que :

- La variable hp et et non corrélée avec against_poison, against_flying et against_ground.
- Les deux variables catch_rate et hp sont négativement corrélés.
- La variable against_ground est négativement corrélée avec against_poison et against_flying.
- La variable catch_rate et non corrélée avec against_poison, against_flying et against_ground.

On peut observer aussi que les variables hp, sp_attack, attack et catch_rate contribuent fortement à la création du premier axe principal, puis les variables against_poison ,against_flying et against_ground contribut fortement à la création du deuxième axe principal.

II.3.3 Répartition des individus dans le plan factoriel



On représente ici les partitions des classes obtenues, sur trois dimensions.
Cependant, comme nous l'avons vu, la réduction de dimensionnalité ne nous suffit pas pour visualiser le clustering de nos données, représenté par le chevauchement des clusters si nous n'utilisons que les 2 premières dimensions.

IV. Conclusions, perspectives et critique

IV.1. Points qui peuvent être critiqués au niveau des choix faits

Le nombre des variables choisies peut être critiqué. Le but était de pouvoir faire une étude sur un maximum de caractéristiques des Pokémon qui semblaient pertinentes mais il est possible que les résultats aient été plus clairs mais moins précis, si on avait choisi de prendre 12 variables.

IV.2. Pistes qui pourraient être explorées pour aller plus loin et/ou mieux explorer ces données

Pour améliorer notre travail, on aurait pu développer notre étude sur d'autres aspect telles que les type des Pokémon,dont les 18 déja existants(eau,air,feu,acier,combat,etc,..) ou les dommages reçus contre un certain type X(variables against X).