Rapport du projet de classification des Pokémon

Elisa Floch, Amélie Goutard et Violette Marmion

1. Prise en main du jeu de données

Dans un premier temps, nous allons présenter la base de données ainsi que le sujet de notre étude de classification.

Le jeu de données concerne l'univers des Pokémon. La franchise a été créée en 1996. Elle est exploitée sous forme de jeux vidéos, mangas, dessins animés, jeux de cartes ou jeux en réalité augmentée. Le principe du jeu repose sur la capture, le dressage et les combats de créatures "animales" imaginaires appelées Pokémon et possédant chacun leurs caractéristiques propres.

Le jeu de données proposé contient différentes informations et statistiques concernant chacun des Pokémon de la première à la huitième génération.

La base de données initiale contenait 1045 lignes et 51 colonnes. Une ligne correspond à un Pokémon identifié, uniquement, par son nom et par son numéro dans la base des données des Pokémon, le Pokédex (non unique). Certains individus apparaissent plusieurs fois car ils ont plusieurs formes ou une évolution supplémentaire dite "Méga". Les variables représentent les caractéristiques nominales telles que le nom (en anglais, en allemand et en japonais), le numéro d'identification, le statut (normal, légendaire, etc), la génération et l'espèce. Mais aussi le type de la créature. Les Pokémon ont 1 ou 2 types parmi les 18 types existants : Grass, Fire, Water, Bug, Normal, Poison, Electric, Ground, Fairy, Fighting, Psychic, Rock, Ghost, Ice, Dragon, Dark, Steel et Flying.

Des caractéristiques quantitaves sont aussi présentes. Nous retrouvons la taille, le poids, les points de vie, d'attaque, de défense, de vitesse, des caractéristiques de dressage et les dommages reçus contre un certain type. Les valeurs des dommages sont égales à 0, 0.5, 1, 2 ou 4.

Dans un second temps, nous allons expliquer notre compréhension du jeu de données et montrer comment nous l'avons nettoyé.

Après avoir importé les données, nous regardons s'il y a des données manquantes. Par exemple, dans la colonne *type_2*, nous constatons qu'il y a 492 données manquantes. Ce qui semble normal puisque tous les Pokémon n'ont pas 2 types. Autre exemple, la colonne *percentage_male* comporte 173 données manquantes. Celles-ci correspondent aux Pokémon asexués. Au total, nous obtenons 2448 données manquantes.

Ensuite, nous avons regardé le nombre de Pokémon qui étaient en doublons, en fonction du numéro de Pokédex. Nous avons obtenu 147 doublons. Nous avons conservé la forme de base et supprimé les formes Méga de chaque Pokémon. Nous obtenons donc une base contenant 898 individus.

Puis, nous avons décidé de supprimer ces colonnes, selon nous, inutiles à notre analyse : *german_name*, *japanese_name*, *species*, *catch_rate*, *base_friendship*, *base_experience*, *growth_rate*, *egg_type_number*, *egg_type_1*, *egg_type_2*, *egg_cycles*, *abilities_number*, *ability_1*, *ability_2*, *ability_hidden*, *total_points*.

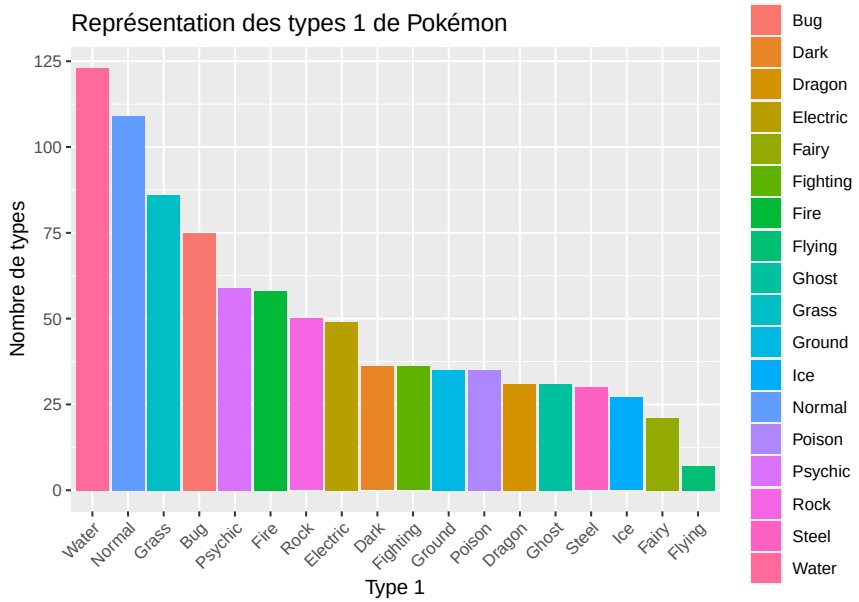
Une espèce est constituée d'un Pokémon et de ses évolutions. Ayant supprimé les évolutions de chaque Pokémon, nous avons jugé que cette variable *species* n'était pas interéssante. La variable *catch_rate* donne le niveau de rareté de capture d'un Pokémon. Afin de ne pas être redondant, nous avons décidé de garder uniquement comme caractéristique de rareté la variable *status*. Il n'était pas interéssant de garder l'amabilité d'un Pokémon donc nous avons supprimé la colonne *base_friendship*. Les variables *base_experience* et *growth_rate* déterminent le niveau et les points d'expérience des Pokémon. Ces variables ne sont comparables

car les niveaux sont propres à chaque Pokémon. Les variables abilities donnent le talent de chaque Pokémon, toutefois deux Pokémon identiques peuvent avoir des talents différents donc ce n'est pas une caractéristique d'un type de Pokémon. L'oeuf n'étant pas un Pokémon, nous avons décidé de supprimer les colonnes *egg*. La colonne *total_points* a été supprimée car étant la somme des variables *hp*, *attack*, *defense*, *sp_attack*, *sp_defense* et *speed* cette variable est redondante. Nous avons donc conservé 34 colonnes.

Ensuite, nous avons transformé la variable qualitative *status* en variable quantitative en lui attribuant les valeurs 1 pour Normal, 2 pour Sub Legendary, 3 pour Legendary et 4 pour Mythical.

Enfin, nous avons fait une exploration des données à l'aide de quelques statistiques descriptives.

Le graphique ci-dessous représente la variable *type_1* de Pokémon :



Le type *Water* est le plus repésenté puisque quasiment 125 créatures incarnent ce type.

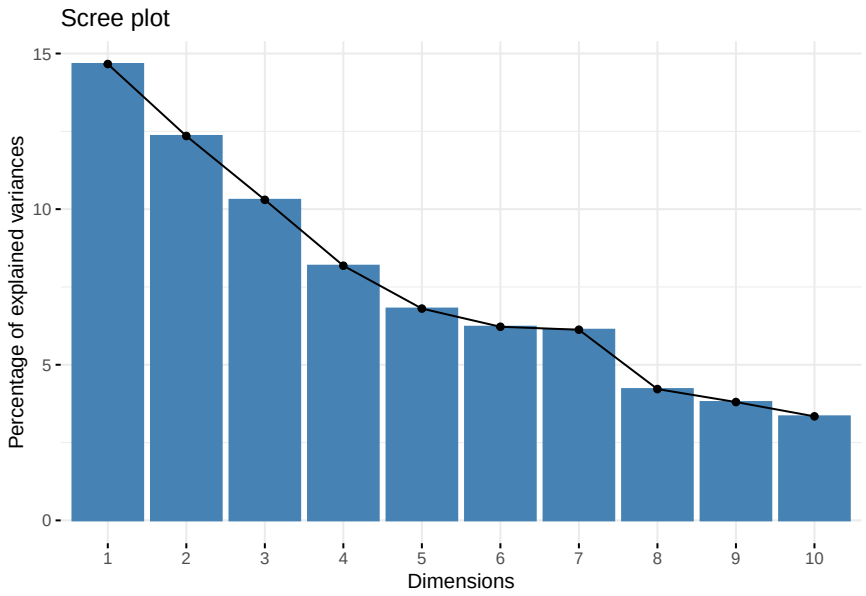
2. Classification des individus

2.1. ACP

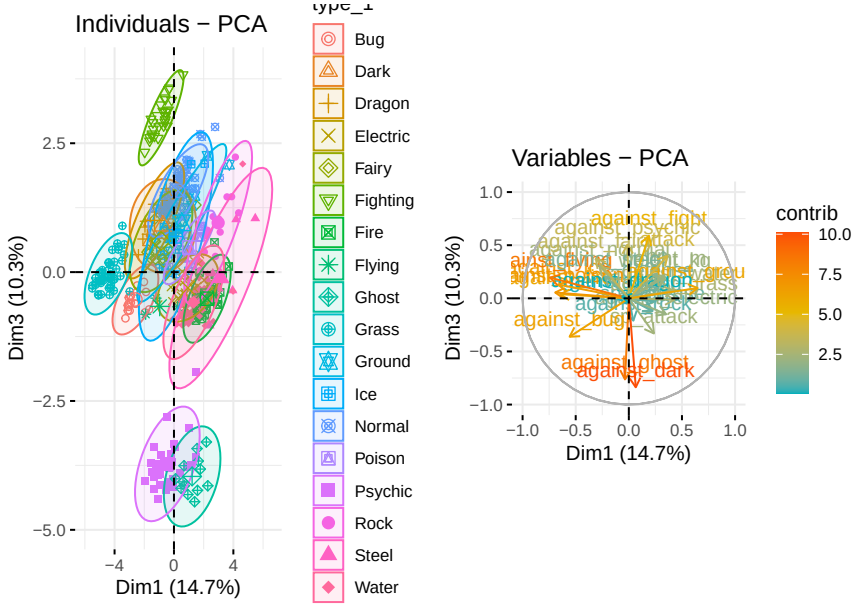
Avant de commencer la classification, nous allons réaliser une Analyse en Composantes Principales. L'analyse des données nous apporte une première interprétation et intuition. A partir des statistiques, nous cherchons à voir si nous pouvons regrouper les Pokémon par leurs types. Nous cherchons donc à représenter les individus, les Pokémon, et à voir graphiquement par une ACP si des groupes se forment, soit si certains individus ont des caractéristiques communes.

Ensuite, nous chercherons à vérifier notre intuition par une classification. L'objectif de notre classification est de prédire le type d'un Pokémon lorsqu'on en rencontre un nouveau.

Tout d'abord, nous nous focalisons sur les Pokémon ayant un seul type pour l'ACP. Nous standardisons les données car elles n'ont pas toutes la même unité.



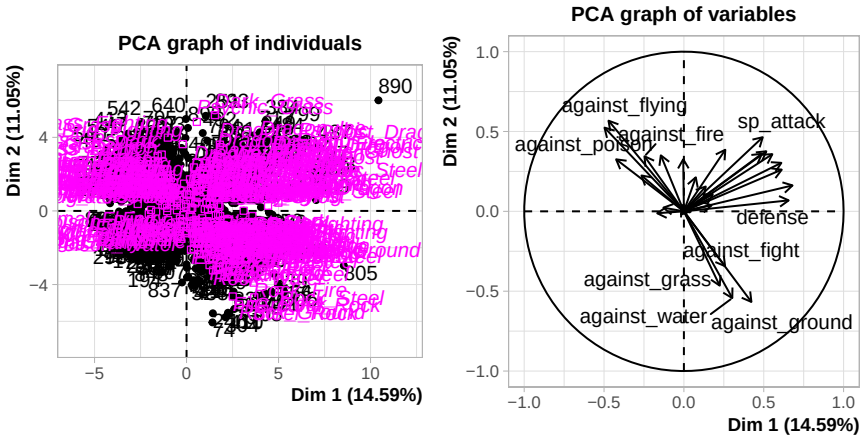
Nous représentons le diagramme en barre des pourcentages d'inertie de chaque axe. Nous représentons ensuite les individus et les variables sur les axes (1,2), (1,3), (1,4). La représentation qui permet de distinguer au mieux les individus est celle des axes 1 et 3.



Nous remarquons que 3 groupes semblent se distinguer. Le premier groupe contient les Pokémon de type Fighting, le second groupe contient les Pokémon de types Psychic et Ghost et enfin le dernier regroupe tous les types restants.

Nous réalisons désormais l'ACP avec tous les Pokémon. Pour cela, nous regroupons les colonnes *type_1* et *type_2* dans la colonne *type*. Cette colonne recense alors 177 types.

Nous calculons l'ACP sur les données standardisées.Nous choisissons de représenter les individus et les variables sur les axes 1 et 2 car c'est le plan qui distingue au mieux les individus.



Il est difficile de lire clairement les types étant donné qu'il y en a 177 mais 4 voire 5 groupes semblent se distinguer.

Conclusion de l'ACP : on aurait tendance à préférer le regroupement de type_1 et type_2 car les groupes semblent plus distincts.

2.2. CAH

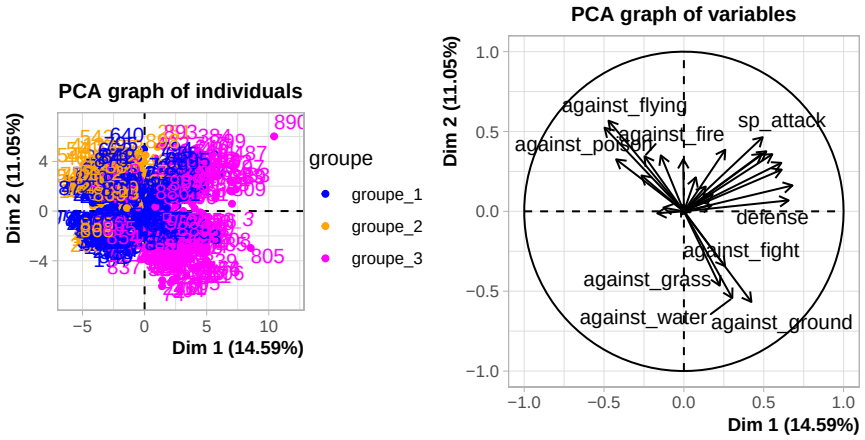
Le but est de créer une suite de partitions emboîtées : en n classes, puis en n-1 classes, etc jusqu'à obtenir une unique classe. Nous allons donc choisir le nombre de groupes le plus pertinent à la fin (par comparaison des partitions créées).

Les caractéristiques de la Classification Ascendante Hiérarchique sont :
- **Ascendant** : on regroupe des individus/classes à chaque étape
- **Hiérarchique** : les classes formées à chaque étape ne sont pas remises en cause
Le principe de cette méthode est de partir de la classification la plus fine : chaque individu est dans sa propre classe.

Pour appliquer cet algorithme, nous voyons qu'il nous faut :
- Définir une distance ou dissimilarité entre individus
- Définir une distance entre classes : pour le recalcul des distances à l'issue de chaque étape

Après avoir standardisé les données, nous avons calculé la distance euclidienne. Pour effectuer la CAH, nous avons utilisé la distance de Ward. Cette distance nécessite de disposer des distances euclidiennes (pour que la notion de barycentre comme centre du nuage de points ait un sens). Elle tend à créer des groupes sphériques et de tailles similaires. Elle est sensible à la présence de points outliers. C'est le critère le plus utilisé. À chaque étape, nous créons de nouvelles classes en agrégeant les classes (pouvant être réduites à un seul individu) les plus proches. À posteriori, on choisit le nombre de groupes le plus adapté. En regardant le tracé de la perte d'inertie, nous hésitons entre 3, 5 et 6 groupes. Le choix n'étant pas très naturel, nous décidons de tester les 3 cas et ainsi de choisir ce qui nous semble le plus convenable.

Commençons par faire une partition en 3 groupes. Pour cela, nous avons fait une représentation des groupes issus de la CAH sur le plan factoriel.



Nous constatons que les classes 2 et 3 sont bien distinctes. La classe 1 est globalement indépendante mais un peu plus mélangée avec les deux autres classes.

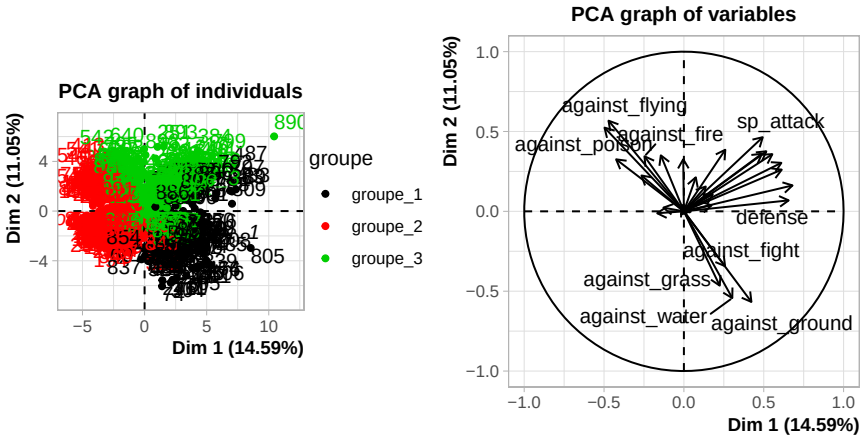
Enfin, nous avons testé notre modèle pour différentes valeurs de K.
- Pour 4 groupes : Nous retrouvons 2 classes distinctes (4 et 2) et 2 classes plus mélangées (1 et 3).
- Pour 5 groupes : Nous observons 2 classes distinctes (5 et 3) tandis que les autres classes sont moins évidentes. Si nous comparons les classes 2 et 4 à la classe 1, nous pouvons dire que cette dernière se distingue un peu plus de la classe 5.
- Pour 6 groupes : Nous remarquons que les classes 3, 5 et 6 sont distinctes. Elles correspondent bien au graphique des variables (3 groupes de flèches allant dans 3 directions différentes).

Ainsi, avec une CAH, nous décidons de conserver 3 classes. Ce choix n'a pas été très évident mais c'est celui qui nous semble le plus raisonnable. Regardons dans la partie suivante si notre choix se confirme avec les K-means.

2.3. K-means

Nous allons consolider les résultats de la CAH en utilisant l'algorithme de K-means. Cela va permettre la réallocation d'individus qui auraient été "mal" classés durant la CAH.

Nous avons commencé par une partition en 3 groupes. Pour cela, nous avons fait une représentation des groupes issus des K-means sur le premier plan factoriel.

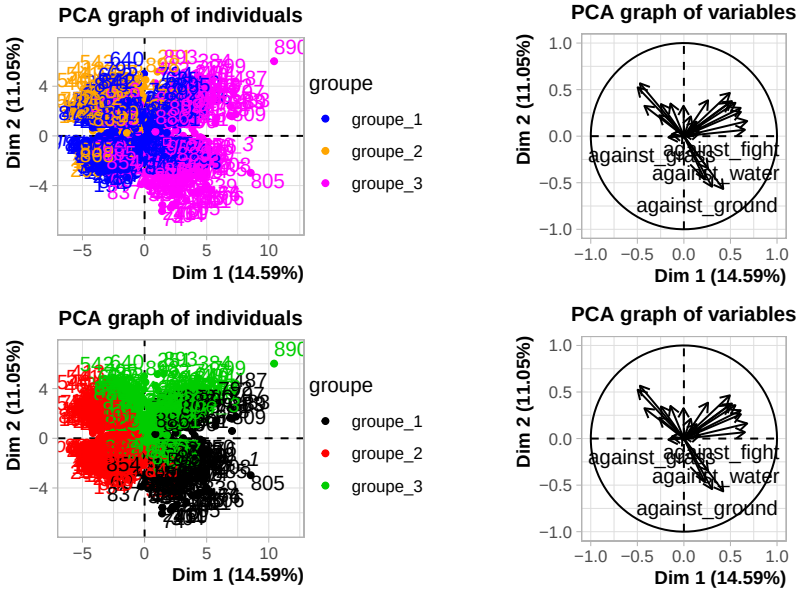


Nous constatons que les 3 classes sont distinctes, peut être même plus qu'avec la CAH. Puis, nous avons testé notre modèle pour différentes valeurs de K.

Ainsi, avec la méthode des K-means, nous décidons de conserver 3 classes. Notre choix est rassurant puisque le nombre de classes que nous avons gardé est le même avec les deux méthodes. Cependant, les groupes sont différents selon la méthode utilisée. Nous allons donc les comparer afin d'en retenir une seule.

2.4. Choix de la classification

Dans cette dernière partie, nous allons décider d'une classification finale. Pour cela, nous allons comparer les résultats de CAH et de K-means Pour rappel, voici la représentation des groupes sur le plan des axes 1 et 2 de la CAH et des K-means, respectivement.



Pour chaque méthode, nous avons compté, pour chaque classe, le nombre de Pokémon par type, en prenant en compte la variable *type_1*. Nous sommes conscientes que nous perdons en précision en regardant uniquement le premier type. Mais cela nous donne déjà une idée de la "justesse" des méthodes.

Ensuite, nous avons regardé les types de Pokémon les plus présents par classe. D'après nos connaissances sur le sujet, les classes issues de la K-means nous semblent plus cohérentes que celles issues de la CAH. En effet, le groupe 1 correspond, à priori, aux Pokémon ayant une résistance élevée ou pouvant annuler des attaques alors que le groupe 2 représente les Pokémon ayant des statistiques fortes en attaques spéciales. Enfin, le groupe 3 contient les Pokémon ayant de grosses faiblesses.

Pour comparer, nous remarquons que les groupes issus de la CAH ne sont pas très cohérents. Le groupe 1 n'est pas très ordonné, il semble contenir tout type de créatures tandis que le groupe 2 est composé, presque uniquement des Pokémon de types *Fairy* et *Grass*. Enfin, le groupe 3 rassemble les types *Ghost* et *Steel* alors qu'ils ont des types opposés.

Graphiquement, nous choisissons les groupes issus de la méthode des K-means De plus, vis-à-vis de nos connaissances et de la documentation sur les caractéristiques des types (plus de détails ici) nous trouvons que les résultats sont plus cohérents pour cette méthode.

3. Étude des groupes obtenus

	V1	V2	V3
groupe	1	2	3
nb_pokemon	168	372	358
poids_moy	145.61607	14.14651	79.71788
taille_moy	1.4714286	0.6155914	1.6446927
hp_moy	71.29167	52.50000	85.14804
attack_moy	89.60714	55.73387	92.43017
defense_moy	99.85119	51.67473	79.59777
sp_attack_moy	71.28571	52.50269	87.05866
sp_defense_moy	78.26786	52.42473	83.90503
speed_moy	57.12500	55.50806	80.94134
ag_normal_moy	0.4092262	0.9784946	0.9972067
ag_fire_moy	1.208333	1.200269	1.014665
ag_water_moy	1.5327381	0.9643817	0.9015363
ag_electric_moy	0.8690476	1.0893817	1.1026536
ag_grass_moy	1.3809524	0.9038978	0.9601955
ag_ice_moy	1.002976	1.585349	2.270950
ag_fight_moy	1.3422619	0.9247312	1.0914804
ag_poison_moy	0.3422619	1.1283602	1.1040503
ag_ground_moy	1.6607143	0.9146505	0.9546089
ag_flying_moy	0.733631	1.336022	1.240223
ag_psychic_moy	0.8258929	1.0645161	1.0055866
ag_bug_moy	0.7366071	1.0215054	1.0949721
ag_rock_moy	0.8794643	1.3790323	1.2765363
ag_ghost_moy	1.3363095	0.9247312	0.9050279
ag_dragon_moy	0.8392857	0.9543011	1.0502793
ag_dark_moy	1.2946429	1.0087366	0.9958101
ag_steel_moy	1.1369048	0.9549731	0.9357542
ag_fairy_moy	0.8735119	1.0275538	1.2283520

Suite au choix de la méthode des K-means, nous pouvons étudier les groupes obtenus. Le résumé statistique nous permet d'étudier les caractéristiques de chaque groupe.

- Groupe 1 : Ce groupe contient 168 Pokémon. Ce sont principalement des Pokémon de types Rock, Ghost et Steel. Ces Pokémon sont lourds et ont beaucoup de points de vie (HP). De plus, ce sont des Pokémon avec une attaque et une défense élevées mais ils ne sont pas très rapides pour attaquer. Ils resistent très bien aux types Normal et Poison mais ils sont assez faibles face aux types Water et Flying.
- Groupe 2 : Le deuxième groupe comprend 372 Pokémon de types Psychic, Dragon et Water principalement. Ces Pokémon sont grands et ont beaucoup de points de vie (HP). Leurs attaques classique et spéciale sont élevées et ils sont rapides pour attaquer. Les Pokémon du second groupe sont faibles face au type Ice.
- Groupe 3 : Le troisième et dernier groupe comporte 358 Pokémon. C'est le plus grand groupe. Il est principalement constitué des types Grass, Normal et Bug. Ces Pokémon sont très petits et légers. Ils n'ont pas beaucoup de points de vie (HP). Leurs attaques et leurs défenses sont faibles et ces Pokémon ne sont pas très rapides. Ils sont très faibles face au type Ice et plutôt faible face aux types Flying et Rock.

4. Conclusions, perspectives et critique

Nous pouvons conclure que les trois groupes obtenus avec les K-means semblent cohérents. En effet, les statistiques par classes rejoignent bien nos intuitions et connaissances sur les Pokémon après s'être documenté.

Cependant, il ne faut pas oublier que nos résultats découlent de nos choix antérieurs. Tout d'abord, nous avons fait le choix de ne pas traiter les variables concernant les groupes d'oeufs, l'espèce, l'habilité, le taux de capture, le lien affectif, l'expérience de base du Pokémon et le pourcentage de mâle.

Le pourcentage de mâle est une variable qui aurait pu être intéréssante afin de voir si certains types de Pokémon sont plutôt féminin, masculin ou asexué. Le problème de cette variable est qu'elle prend la valeur NA lorsque le Pokémon est asexué. Pour pouvoir l'intégrer à notre analyse, il aurait fallu prendre en compte les NA comme une catégorie à part entière. Une possibilité aurait été de transformer la variable en variable qualitative. Cependant en procédant de la sorte, nous perdons de l'information. Pour aller plus loin dans notre analyse, nous aurions pu intégrer la variable *percentage_male* dans notre classification. Nous aurions peut-être eu des resultats différents.