Projet LSTAT2110 - Analyse des données

Ledune Lucien - 39301400 - DATS2M

Date de soumission :

Introduction

Pokémon est une franchise commercialisée par "The pokémon company", une société gérée par Nintendo, Game Freaks et Creatures.

Même si la franchise à d'abord commercialisé des jeux vidéos, celle-ci s'est étendue à d'autres marchés suite à la popularité montante de ses jeux.

L'ensemble des jeux pokémon à totalisé plus de 290 millions de ventes dans le monde. Les deux premiers jeux pokémon sont sortis en même temps au Japon en 1996. Il s'agissait de Pokémon version Rouge/Bleue. C'est sur cette première génération de 151 spécimens que portera notre analyse.

Le principe du jeu est celui-ci : On incarne un personnage qui parcours le monde à la recherche de pokémon, le but étant de tous les attraper. Ils servent majoritairement dans des combats, ils possèdent un certains nombre de statistiques (décrites dans la partie suivante) déterminant comment ils se comportent dans les combats.

Un autre point important est l'évolution. Àprès avoir combattu, certains pokémons peuvent évoluer, c'est à dire changer de forme et de statistiques. (Il pourra être intéressant de voir comment changent les pokémons évolués).

Il existe aussi un type particulier de pokémons, les légendaires, ceux-ci sont généralement plus rare et plus puissants que les autres, aussi nous pourrons essayer de les identifier grâce à l'analyse linéaire discriminante.

Présentation des données, analyse descriptive

Présentation des données

Pour réaliser ce projet d'analyses de données, nous devions trouver un jeu de données disposant d'au moins 30 individus, 6 variables continues, et une variable discrète pour l'analyse discriminante linéaire.

Le set de données retenu a été trouvé sur www.kaggle.com, et représente les 800 différents pokémons existants dans les 7 générations du jeu, représentés par 13 variables.

Cependant afin de faciliter l'analyse, certaines lignes et variables seront omises.

En effet pour l'analyse en composantes principales nous ne conserveront que les variables continues, car celle-ci ne s'applique pas aux variables discrètes (classement).

Ensuite afin de faciliter l'analyse discriminante linéaire ainsi que la représentation des données nous effectuerons les analyses sur la première génération uniquement, soit les 151 premiers pokémons.

Les différentes variables sont :

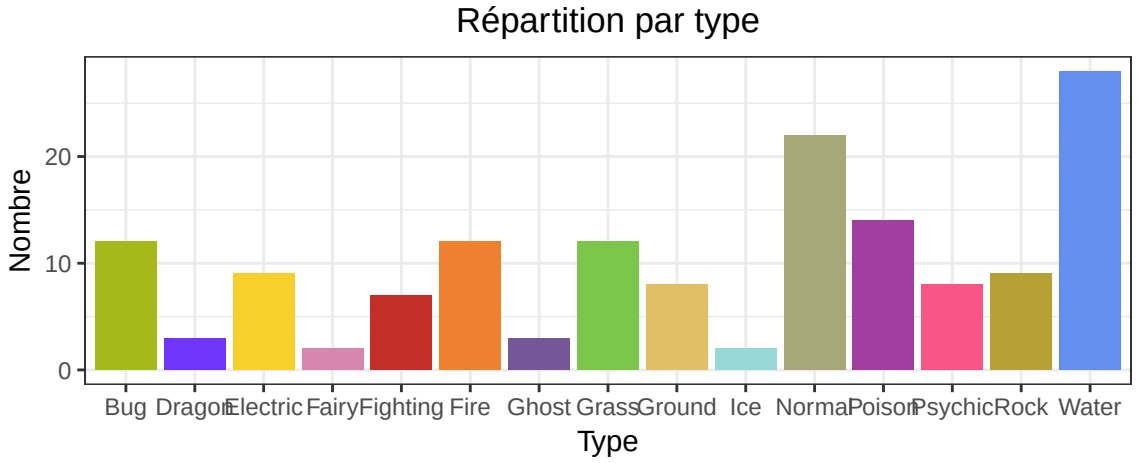
- X. : Un nombre désignant le numéro du pokémon, cette variable sera supprimée car peut pertinante dans notre analyse.

- Name : Le nom du pokémon, bien qu'ils ne seront pas utilisés à proprement parler dans l'analyse nous les conserverons afin de mieux représenter les pokémons sur certains graphiques.

- Type 1 : Le type principal du pokémon.
- Type 2 : L'éventuel type secondaire du pokémon.
- Total : Une variable représentant la somme des 6 prochaines variables continues. Celle-ce sera omise pour l'analyse en composantes principales car il est évident qu'elle sera corrélée avec les autres variables discrètes comme il s'agit de leur somme.
- HP : Les points de vie du pokémon, c'est le nombre de dégats qu'il peut prendre.
- Attack et Sp Attack : Représente la capacité offensive.
- Defense et Sp Defense : Représente la capacité défensive.
- Speed : C'est la vitesse du pokémon, l'individu ayant la plus grande valeur ici attaquera en premier.
- Generation : Représente la génération du pokémon, cetta variable nous permettra de sélectionner les pokémons issus de la première génération.
- Legendary : Boolean (True/False) indiquant si un pokémon est légendaire ou non.

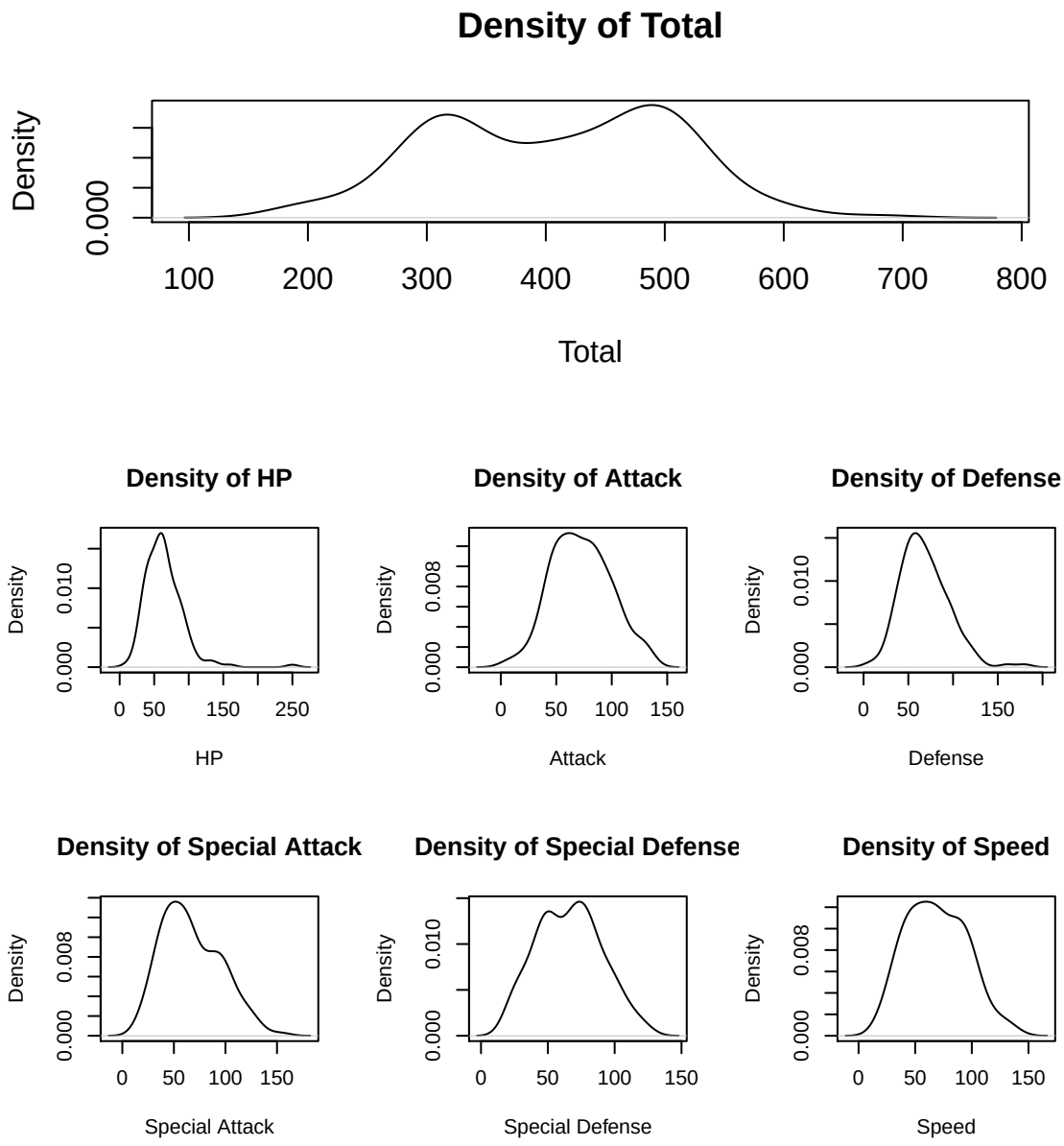
Analyse descriptive

Pour commencer nous pouvons observer la répartition des pokémons en fonction de certaines variables. Premièrement la distribution des pokémons en fonctions de leur type.



Nous pouvons observer sur le premier graphique que la répartition des spécimens par type est très inégale, certains type (water, normal,...) disposent d'un grand nombre d'individus contrairement à d'autres (dragon, fairy,...)

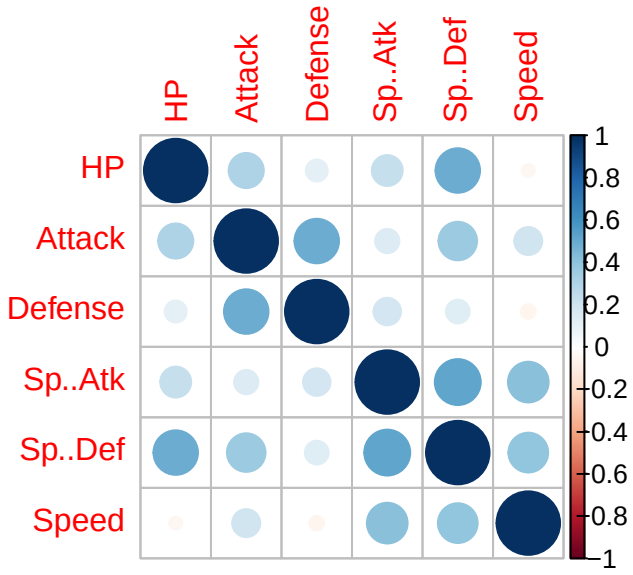
Nous allons maintenant nous intéresser à la distribution des différentes statistiques.



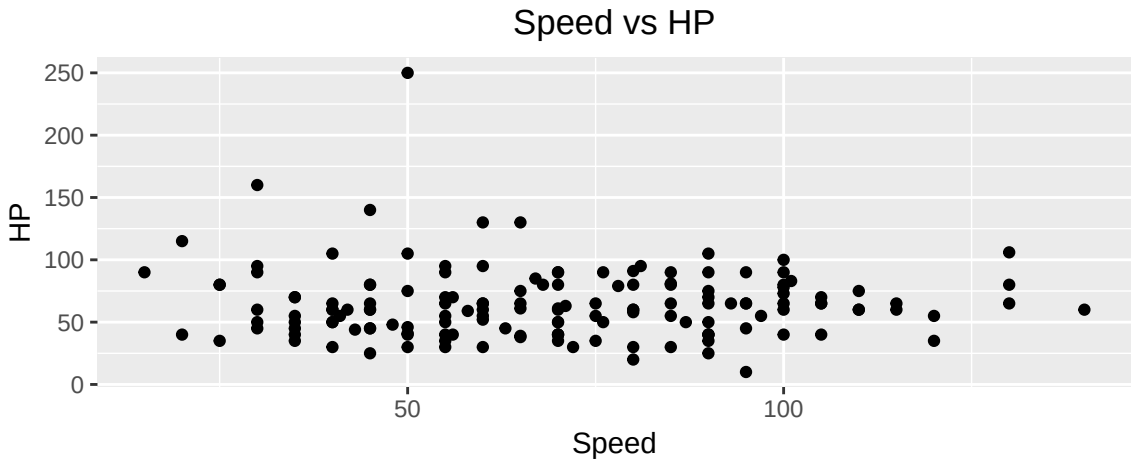
Nous pouvons voir que dans l'ensemble, les statistiques semblent suivre une loi gaussienne. Avec deux pics autours de 300 et 500 pour le total. (Sûrement expliqué par la différence de moyenne statistique entre un pokémon non-évolué et son évolution).

Regardons maintenant la matrice de corrélations des données.

```
corrplot 0.84 loaded
```



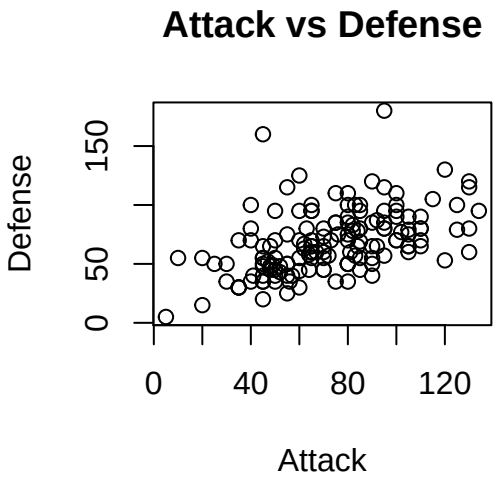
Sur ce graphique, nous pouvons voir que la majorité des variables sont positivement corrélées, ce qui peut vouloir dire que lorsqu'un individu gagne des statistiques, il en gagne de presque tous les types.
Cependant il y a des exceptions, ainsi la vitesse est légèrement négativement corrélée avec les HP et la Défense. Ce résultat est assez intuitif et signifie globalement que plus un individu est rapide, moins il serait résistant.

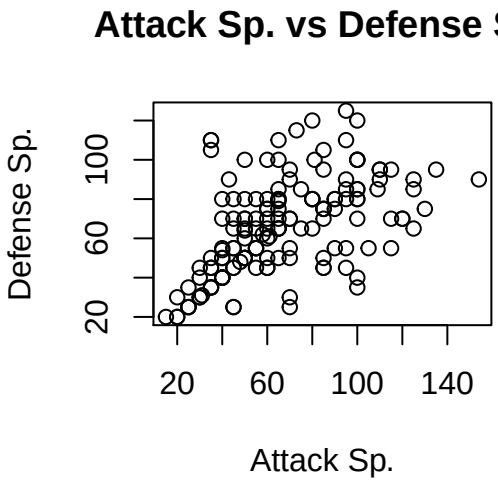


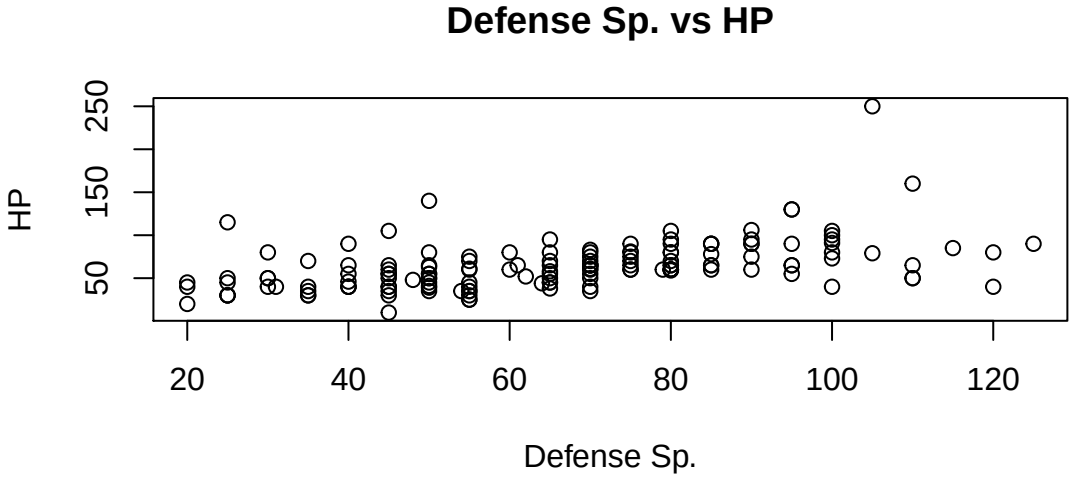
Sur le graphique nous voyons que cette corrélation est pratiquement négligeable.

Voici les autres relations intéressantes :

- L'attaque et la défense sont positivement corrélées.
- L'attaque spéciale et la défense spéciale sont également positivement corrélées
- La défense spéciale et les HP le sont aussi.

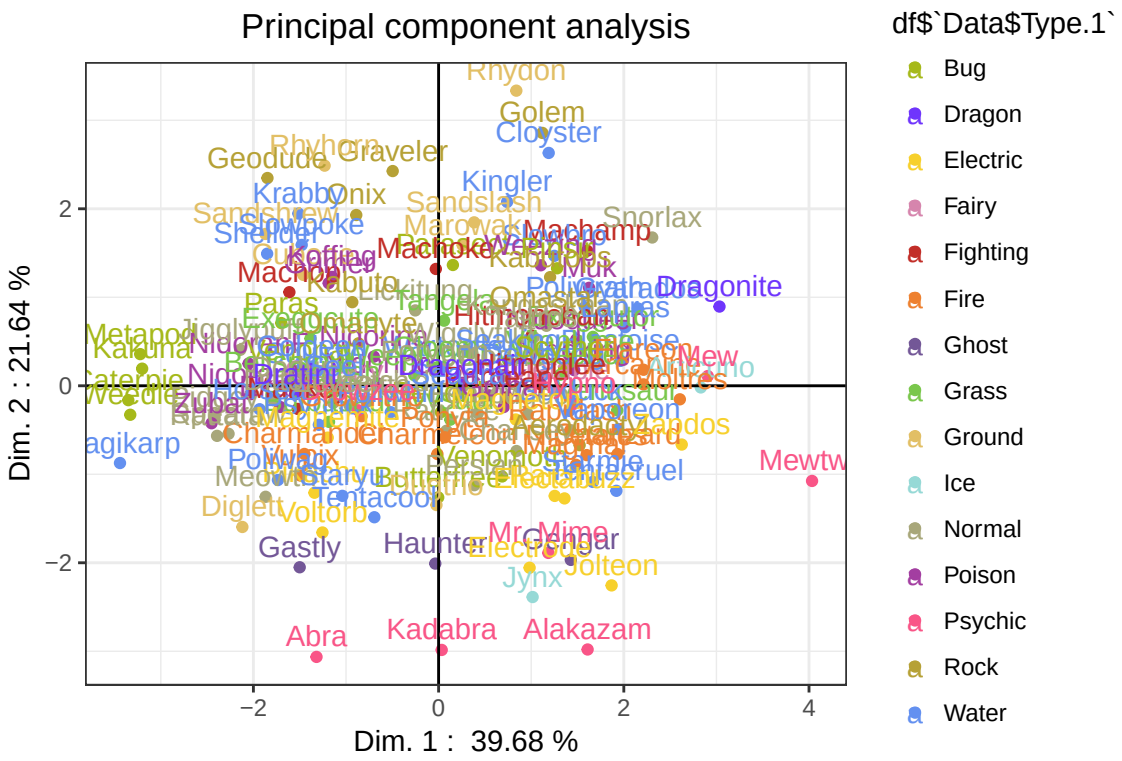

Attack vs Defense
Defense
0
50
150
0
40
80
120
Attack


Attack Sp. vs Defense Sp.
Defense Sp.
20
60
100
20
60
100
140
Attack Sp.


Defense Sp. vs HP
HP
50
150
250
20
40
60
80
100
120
Defense Sp.

Analyse en Composantes Principales

Nous allons maintenant passer à l'analyse en composantes principales de notre jeu de données.



eigenvalu	e percentag	e of variance cumulativ	e percentage of variance
comp 1	2.3810774	39.684623	39.68462
comp 2	1.2983823	21.639704	61.32433
comp 3	0.9841344	16.402240	77.72657
comp 4	0.6752113	11.253521	88.98009
comp 5	0.3418269	5.697115	94.67720
comp 6	0.3193677	5.322796	100.00000

Ce tableau nous indique que les trois premières dimensions conservent environ 75% de la variance.

Corrélation des axes et des variables

	Dim.1	Dim.2	Dim.3
HP	0.5825333	0.2196265	-0.7129478
Attack	0.6527662	0.4860255	0.2438546
Defense	0.4419946	0.6675834	0.4214754
Sp..Atk	0.6926334	-0.3779357	0.0735561
Sp..Def	0.8307983	-0.2026010	-0.2322905
Speed	0.5003026	-0.6199802	0.4235138

Sur le tableau ci-dessus, nous observons que la première dimension est assez bien caractérisée par toutes les variables (particulièrement l'attaque spéciale et la défense spéciale), seules la vitesse et la défense dont

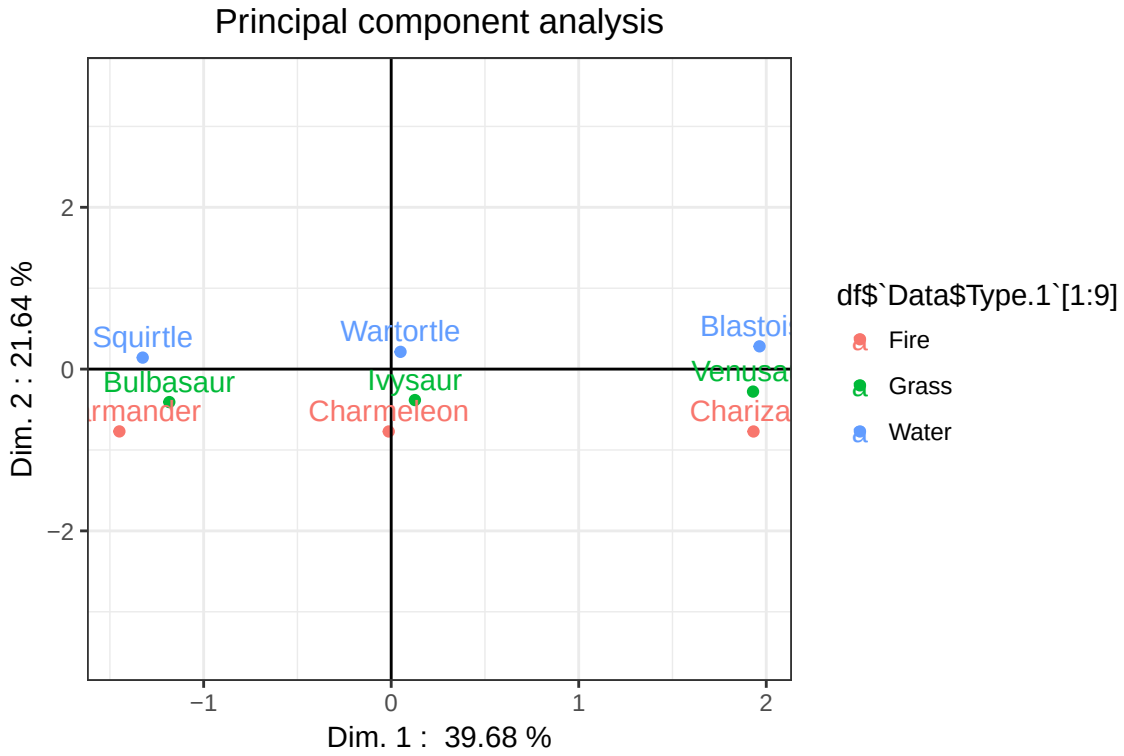
légèrement moins bien représentées.

Celles-ci sont en revanche les deux variables caractérisant le plus la deuxième dimension.
Nous pouvons donc dire que le premier axe va représenter les caractéristiques générales d'un pokémon, tandis que le deuxième va plutôt nous indiquer la défense ou la rapidité d'un pokémon.

Comme la défense est positivement corrélée avec le deuxième axe (au contraire de la vitesse), on peut dire que plus un pokémon sera haut sur la deuxième dimension, plus il sera résistant (et à fortiori plus il sera lent). Mais ces résultats restent influencés dans une mesure non négligeable par les autres variables et il faut en tenir compte.

Pour vérifier que le premier axe représente les caractéristiques générales d'un pokémon, nous pouvons vérifier que les évolutions d'un pokémon vont être representées de manière presque horizontale. (En évoluant un pokémon gagnera de toutes les statistiques).

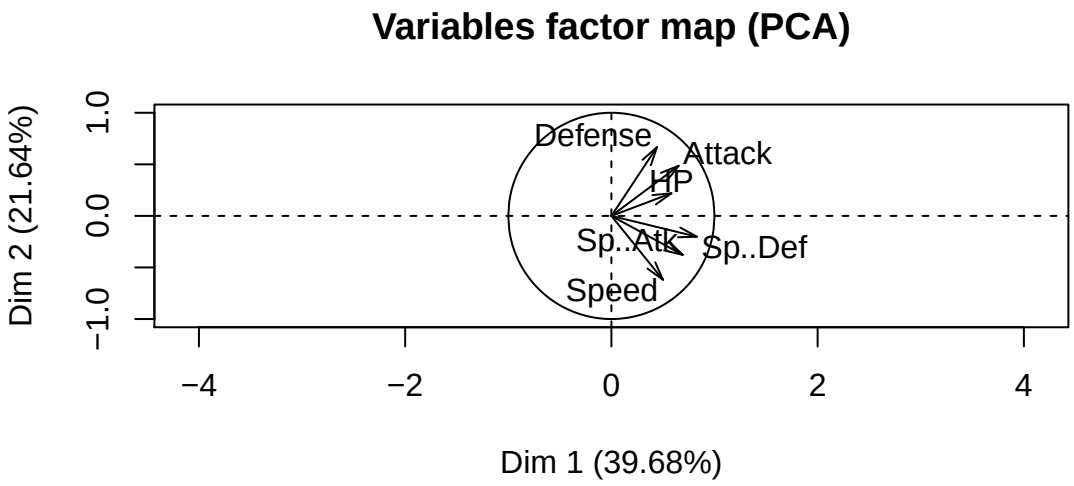
Voici les 9 premières lignes du jeu de données, représentant trois pokémons et leurs évolutions :



X. N	ame T	ype.1 T	ype.2	Total	HP	Attack	Defense	Sp..Atk	Sp..Def	Speed
1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45
2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60
3	Venusaur	Grass	Poison	525	80	82	83	100	100	80
4	Charmander	Fire		309	39	52	43	60	50	65
5	Charmeleon	Fire		405	58	64	58	80	65	80
6	Charizard	Fire	Flying	534	78	84	78	109	85	100
7	Squirtle	Water		314	44	48	65	50	64	43
8	Wartortle	Water		405	59	63	80	65	80	58
9	Blastoise	Water		530	79	83	100	85	105	78

Nous pouvons en effet constater dans le tableau que les statistiques évoluent toutes de manière relativement uniforme et que les évolutions des pokémons se déplacent bien de manière similaire sur la première dimension.

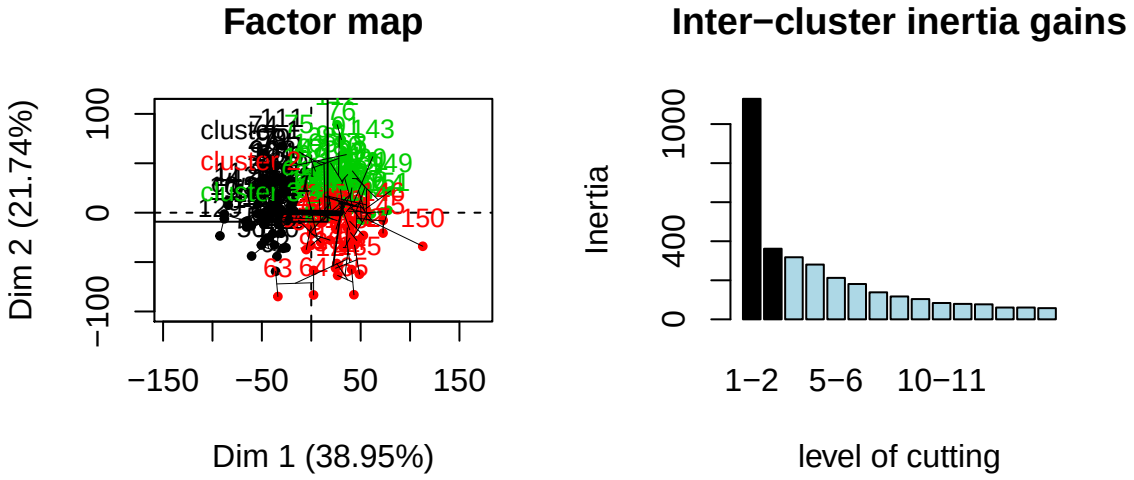
Nous avons aussi prédit précédemment que le deuxième axe indiquait la défense/vitesse d'un pokémon (respectivement vers le haut et vers le bas). Ces résultats se confirment aussi puisque l'individu bleu est en effet le plus résistant et le rouge est le plus rapide.



La sphère des corrélations quant à elle nous indique la même chose que la matrice des corrélation étudiée dans la section sur l'analyse descriptive et l'analyse des variables représentants les dimensions de la PCA.

Clustering

Passons maintenant au clustering de notre jeu de données. Nous utiliserons une méthode non supervisée : la classification hiérarchique. Voici les résultats :



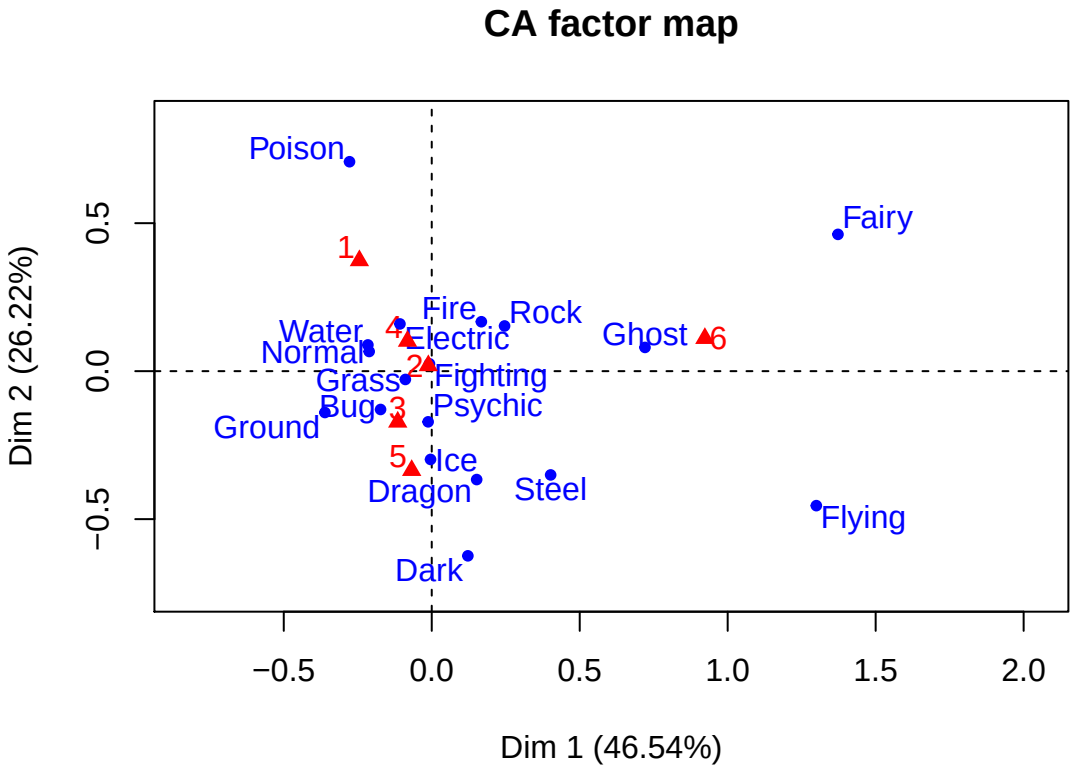
Le graphique en barres montre pourquoi nous divisons la population en 3 groupes différents. (Inter inertia => montre comment les groupes "s'éloignent")

- Le premier groupe semble représenter les individus faibles.

- Le deuxième rassemble les individus puissants et rapides.
- Le troisième représente les individus puissants et résistants.

Analyse des correspondances

Pour réaliser celle-ci, nous regarderons quelles sont les correspondances entre le type d'un pokémon et sa génération. (On reprend le jeu de donnée comprenant toutes les générations)



Sur le graphique nous pouvons observer les relations entre les différents types et les générations.

Nous pouvons en déduire :
* Le type poison semble être plus présent dans la première génération que dans les autres (La liste des générations des pokémons poisons sera affichée en dessous pour vérifier, ce sera aussi le cas pour les autres conclusions).

```
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 3 3 3 4 4 4 4 4 4 5 5 6 6
```

- Les types fairy, ghost et flying sont plus présents dans la 6ème génération que dans les autres.

```
 [1] 1 1 2 2 2 2 2 4 6 6 6 6 6 6 6 6 6
```

```
 [1] 1 1 1 2 3 3 3 3 4 4 4 4 4 4 4 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6
```

```
[1] 5 5 6 6
```

- Le type dark est proche de la 5ème génération.

```
 [1] 2 2 2 2 2 3 3 3 3 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6
```

Les listes semblent bien confirmer nos dires.

Les autres types sont dans l'ensemble répartis de façon plus homogène.

Conclusion

Dans ce travail, nous avons analysés un jeu de donnée représenté les statistiques des différents pokémons. Nous avons globalement utilisé uniquement la première génération (151 premiers) sauf dans le cas de l'analyse des correspondance, où une analyse des types de pokémons par génétion pouvait être intéressante.

Premièrement l'analyse descriptive du jeu de donnée nous a indiqué que les statistiques (continues) des pokémons semblent toutes suivre une loi normale tandis que les types étaient distribués de façon assez inégale. Nous avons aussi pu mettre en évidence certains liens entre les statistiques grâce à la matrice de corrélation des variables.

Ensuite, l'analyse en composantes principales (ACP) nous a permis de réduire le nombre de dimensions de notre jeu de données tout en conservant une grande partie de l'information, afin de pouvoir représenter les pokémons sur un graphique où la signification des deux axes a été interprétée. Le sphère des corrélations utilisée dans cette méthode a pu confirmer nos résultats sur les liens entre les variables.

De plus, le clustering (hiérarchique) nous a permis de classer de manière non supervisée les pokémons en différents groupes selon leur similarités au niveau des statistiques.

Enfin, l'analyse des correspondances (réalisée sur le set complet des 6 générations), nous a permis de découvrir l'existence de liens entre le type principal d'un pokémon et les différentes génération, et d'identifier quelles générations étaient plus propices à être composée d'un certains type de pokémon.

Dans l'ensemble, ce travail m'a permis d'appliquer les différentes méthodes vues au cours dans un cadre concret, et de réaliser une analyse complète d'un jeu de donnée dans le but d'en retirer des informations qui auraient étées bien plus difficiles a identifier sans la dite analyse.