

Rapport d'une étude statistique sur les Pokémon

Auteurs : Nicolas BUSQUET et Manon ROULEAU

Vous trouverez dans ce document l'ensemble des graphiques construits en python avec un jeu de données sur les 801 premiers Pokémon du Pokédex.

Le notebook Jupyter de cette étude peut être trouver [ici](#) (Google Colab).

Présentation du jeu de données :

Nous avons un jeu de données de 801 entrées et 41 variables. Chaque entrée représente un Pokémon et elles sont classées dans l'ordre du Pokédex.

Il contient quelques valeurs invalides (null, NaN...), mais elles sont dans des proportions négligeables ou concernent des variables qui ne seront pas significatives dans nos analyses ultérieures.

I. Analyse univariée

1. *Visualisation des Pokémon par type*
2. *Visualisation de la proportion de Pokémon légendaires*
3. *Top 10 des talents les plus répandus*
4. *Statistiques sur le temps d'éclosion de Pokémon*

II. Résumés statistiques de quelques variables

1. *Variable base_egg_steps*
2. *Variable height_m*
3. *Variable hp*

III. Analyse multivariée

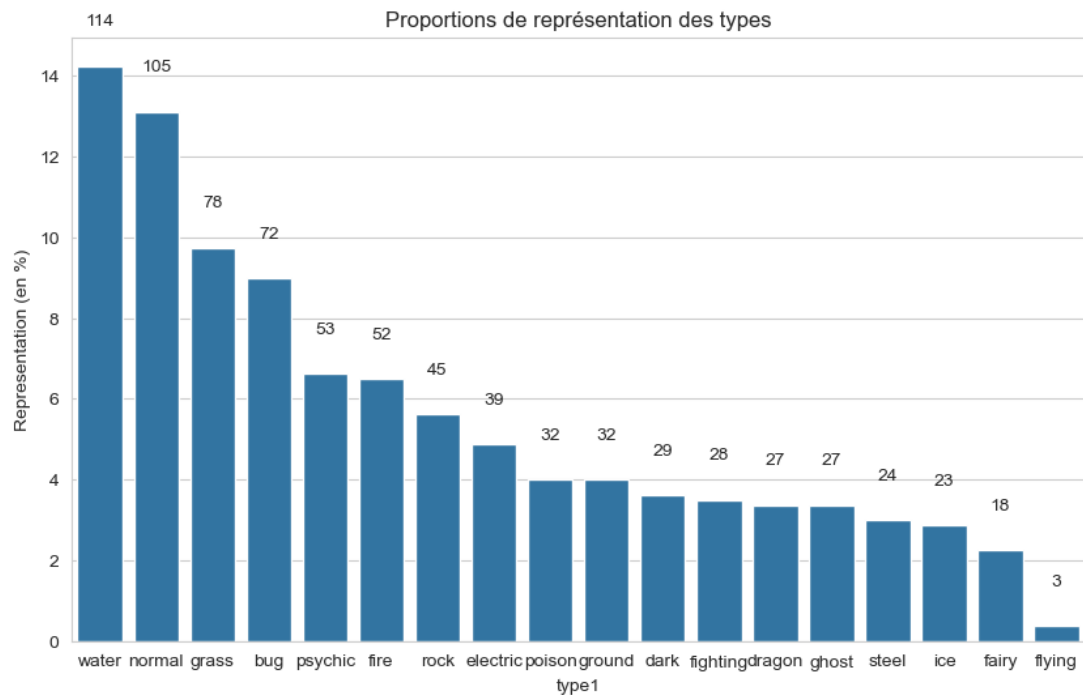
1. *Mise en relation des variables base_egg_steps et is_legendary*
2. *Relation entre les talents des Pokémon et leur type ou rareté*
3. *Distribution des Pokémon légendaires dans les différentes générations*
4. *Lien entre la statistique de Défense et le poids d'un Pokémon*
5. *Relation entre la défense et l'attaque d'un Pokémon*

IV. Analyse de corrélation

1. *Matrice de corrélation (avec heatmap)*
2. *Visualisation des corrélations*
 - a. *Relation entre la génération et le numéro de Pokédex d'un Pokémon*
 - b. *Relation entre le caractère légendaire d'un Pokémon et sa statistique d'attaque*
3. *Comparaison de la distribution de l'attaque spéciale par type (Combat, Psy, Normal)*

I. Analyse univariée

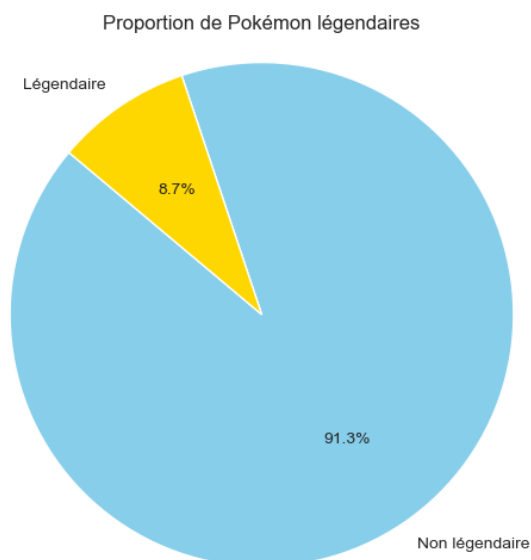
1. Visualisation des Pokémon par type



Moyenne de pokémon par type : 44.5

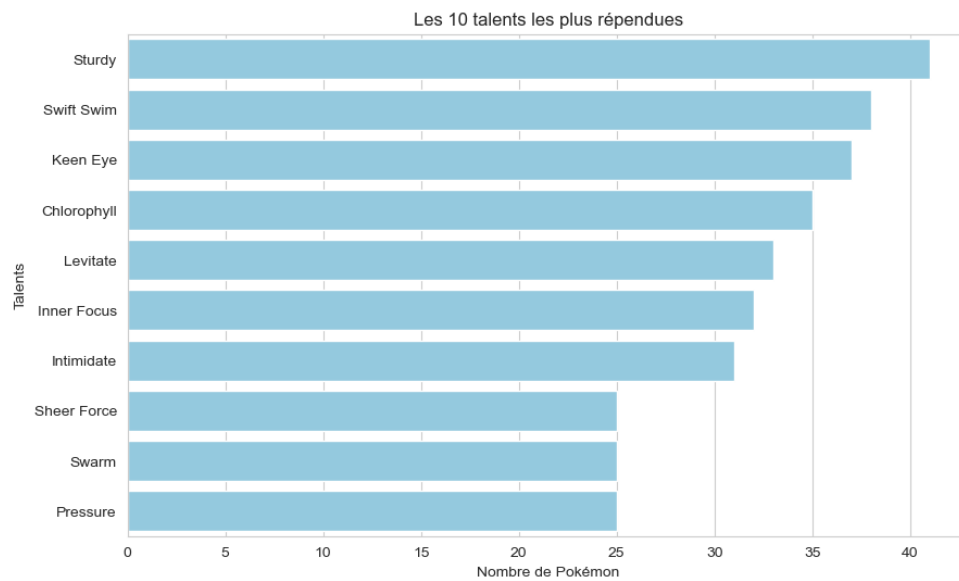
- Le type eau est le plus représenté parmi les Pokémon du jeu de données avec 114 Pokémon de ce type.
- En comparaison, le type Vol est le moins représenté avec environ 3 individus, ce qui nous fait une étendue de 111 Pokémon.
- On a environ 44.5 Pokémon par type.

2. Visualisation de la proportion de Pokémon légendaires



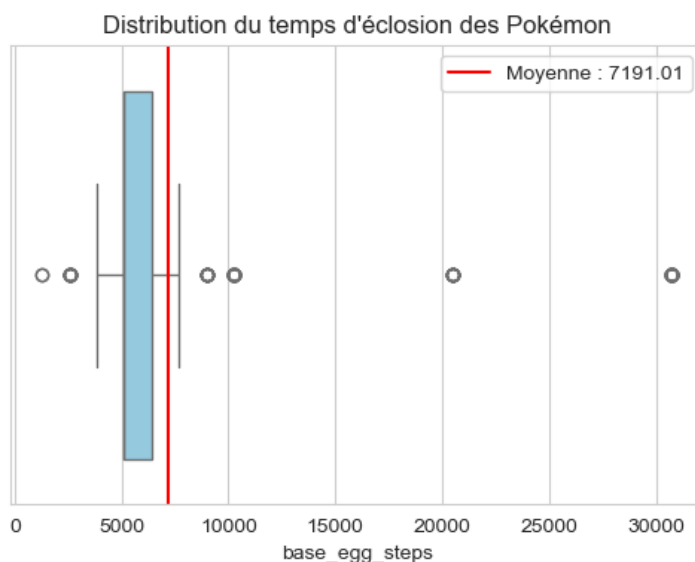
- On a 8.7% de Pokémon légendaires dans l'ensemble de notre jeu de données.
- Conséquemment, on trouve 91.3% de Pokémon normaux.

3. Top 10 des talents les plus répandus



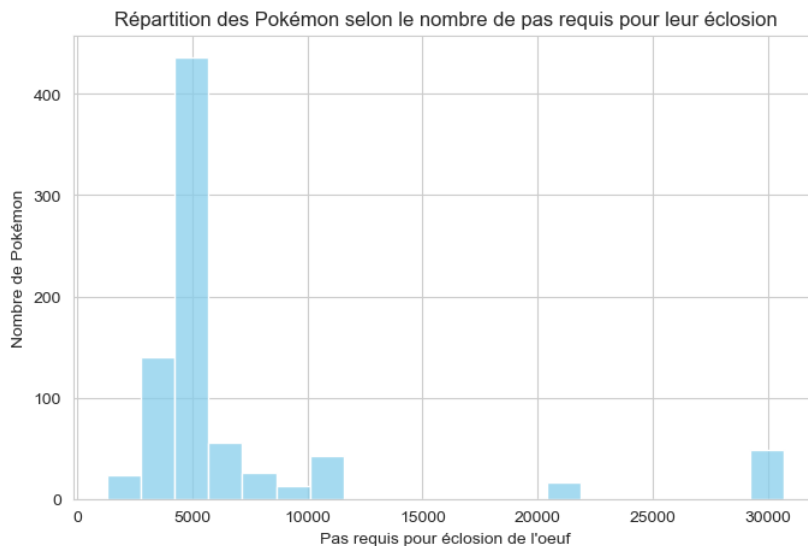
- On a récupéré les 10 talents les plus répandues parmi les Pokémon du jeu de données.
- Sturdy est le talent le plus utilisé, on peut présumer qu'il est répandu parmi les Pokémon eau ou alors parmi les Pokémon les plus communs.

4. Statistiques sur le temps d'éclosion de Pokémon



Avec ce visuel, on obtient plusieurs informations.

- Il faut marcher en moyenne 7191 pas pour faire un éclore un œuf pokémon.
 - La médiane est aux alentours de 5500 pas.
 - Le gros des valeurs est compris entre environ 4000 et 7500 pas.
 - On observe quelques "anomalies", des valeurs qui sortent du lot à 30000, 20000 et 10000 pas et quelques-unes encore plus près des autres valeurs.
 - On observe que la médiane et la moyenne sont plutôt éloignées l'une de l'autre.
- Essayons d'expliquer pourquoi.



- On obtient ici plus d'informations sur les "anomalies" visibles dans le précédent visuel. Il y a un nombre assez important de Pokémon nécessitant un très grand nombre de pas pour éclore, ce qui tire la moyenne vers le haut et l'éloigne de la médiane.
- Il serait intéressant de se demander si les Pokémon nécessitant beaucoup de pas pour éclore appartiennent à une catégorie particulière.

II. Résumés statistiques de quelques variables

1. Variable `base_egg_steps`

	base_egg_steps
count	801.000000
mean	7191.011236
std	6558.220422
min	1280.000000
25%	5120.000000
50%	5120.000000
75%	6400.000000
max	30720.000000

Ces statistiques descriptives nous confirment ce que nous disions précédemment avec des chiffres plus précis. On a en plus l'écart-type (std) qui nous confirme que les valeurs de cette variable sont plutôt dispersées dans notre échantillon.

2. Variable `height_m`

	height_m
count	781.000000
mean	1.163892
std	1.080326
min	0.100000
25%	0.600000
50%	1.000000
75%	1.500000
max	14.500000

- La variable `height_m` est très peu dispersé, même si son étendue est plutôt large.
- On a une étendue de 14.4 mètres, pour un écart-type de 1.08 mètre
- La plupart des Pokémon font autour de 1.16 mètre (moyenne). C'est également confirmé par la proximité des 1er et 3e quartiles.

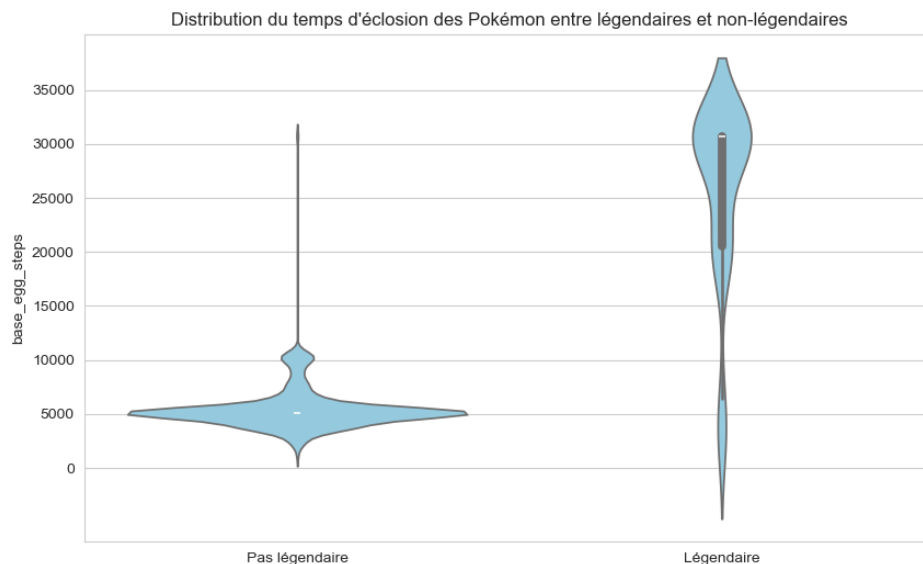
3. Variable hp

	hp
count	801.000000
mean	68.958801
std	26.576015
min	1.000000
25%	50.000000
50%	65.000000
75%	80.000000
max	255.000000

En nous basant sur l'écart-type et la différence entre les 1^{ers} et 3^e quartiles, on constate que les Pokémon ont tous un nombre de points de vie relativement proche, ~69 en moyenne.

III. Analyse multivariée

1. Mise en relation des variables `base_egg_steps` et `is_legendary`



On peut confirmer la conjecture faite plus tôt : les Pokémon légendaires sont ceux qui constituent les anomalies dans la variable `base_egg_step`. Seuls des Pokémon légendaires ont besoin de plus de 15 000 pas pour éclore.

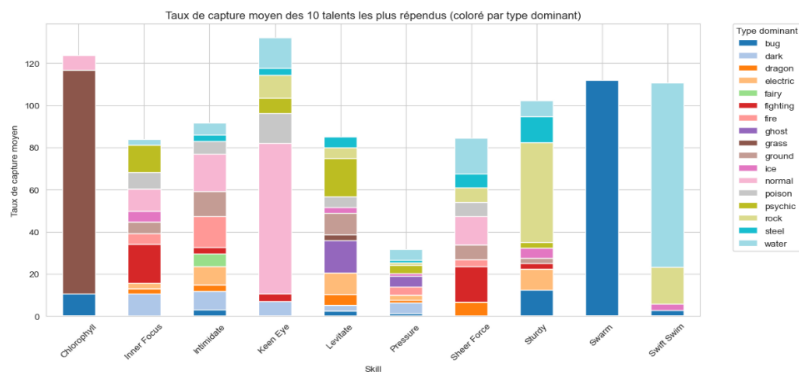
Seule une ligne évolutive de Pokémon contrevient à cette règle : "Type: Null" et "Silvally" les Pokémon Synthétiques. Ces Pokémon ont un statut particulier, il est cohérent qu'ils fassent exception.

2. Relation entre les talents des Pokémon et leur type ou rareté

Plus tôt, nous avons émis l'hypothèse que les talents les plus utilisés pourraient être plus présents chez les Pokémon de type eau, ces derniers étant plus nombreux. Nous allons donc vérifier ça maintenant.

Nous allons en même temps vérifier si cela peut avoir un rapport avec la rareté du Pokémon en prenant en compte la valeur du "capture rate" dans notre visuel.

	capture_rate_n
count	800.000000
mean	98.761250
std	76.257825
min	3.000000
25%	45.000000
50%	60.000000
75%	170.000000
max	255.000000



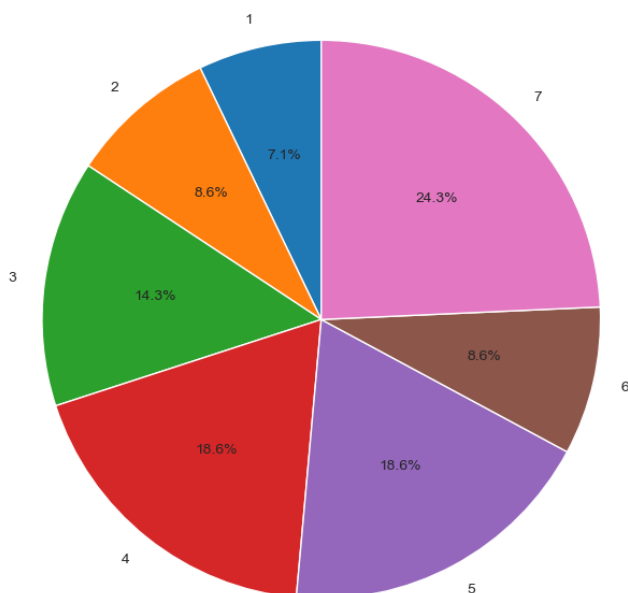
Si on constate que certains types sont dans plus de talents que d'autre, rien ne nous permet d'affirmer que la fréquence d'un type est liée à un ou deux types en particulier.

Aussi, en regardant les statistiques descriptives du *capture_rate* et le *capture_rate* moyen pour les Pokémon de chaque ability, il ne semble pas que la rareté ou l'absence de rareté des Pokémon parmi les utilisateurs d'une ability influence significativement la fréquence de cette ability.

La moyenne de *capture_rate* pour une ability est supérieure à la moyenne globale de *capture_rate* pour 5 abilities seulement.

3. Distribution des Pokémon légendaires dans les différentes générations

Répartition des Pokémon légendaires par génération

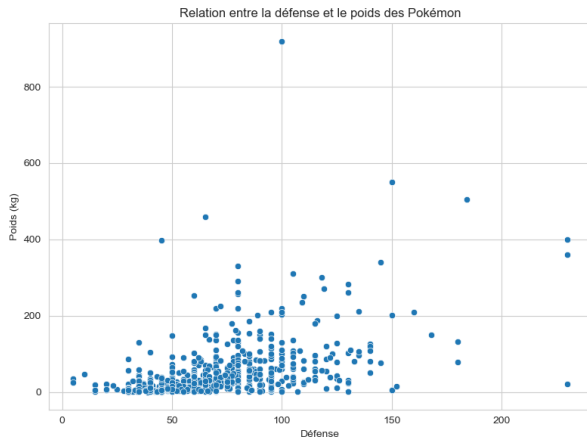


Ce graphique nous donne quelques informations sur les Pokémon légendaires au fil des générations :

- Exception faite de la 6^e génération, une génération a toujours au moins autant de légendaires que celle qui la précède.
- La 7^{ème} génération apporte presque le quart de l'ensemble des légendaires.

4. Lien entre la statistique de Défense et le poids d'un Pokémon

Nous allons construire un nuage de points prenant en abscisse la défense des Pokémon et en ordonnée leur poids. Ainsi, nous pourrions voir s'il y a une relation entre les deux. Il pourrait être logique qu'un Pokémon plus lourd ait plus de Défense.

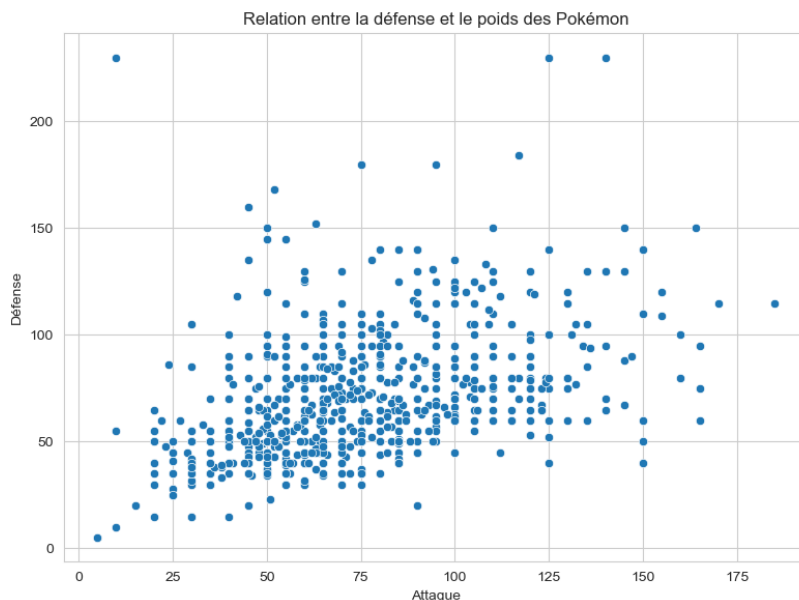


On observe une faible corrélation jusqu'à 200 kilos environ, dans la mesure où les Pokémon lourds ont pour la plupart plus de 75 de défense.

Cette corrélation n'est cependant pas marquée et ne semble pas permettre d'établir une règle claire.

5. Relation entre la défense et l'attaque d'un Pokémon

Également dans un nuage de point, nous allons visualiser l'attaque et la défense. Notre hypothèse est la suivante : un Pokémon ayant beaucoup d'attaque ne devrait pas avoir une défense très élevée. Vérifions ça.

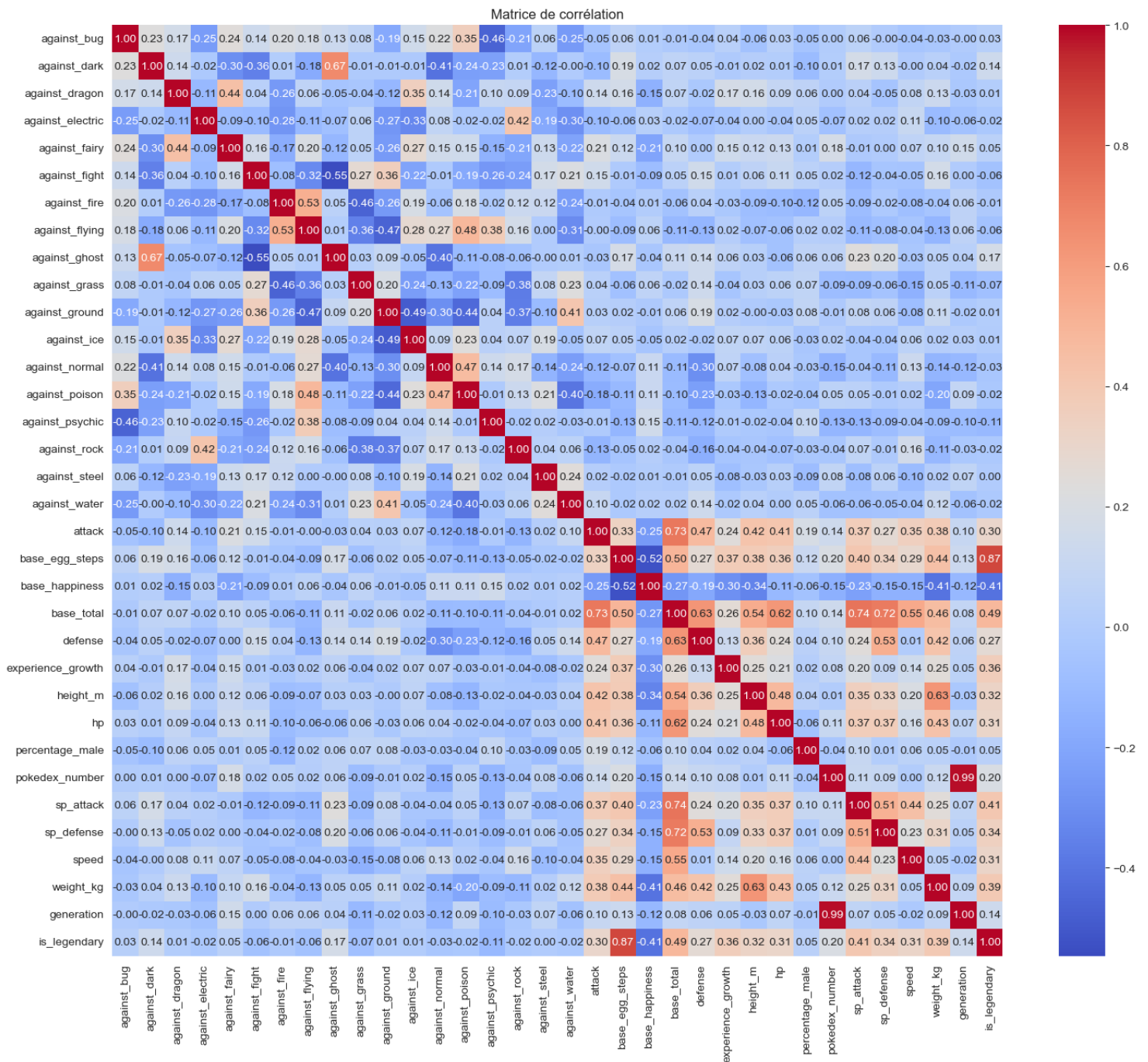


Bon, notre hypothèse est donc fautive, et il semble même que son contraire soit vrai. Nos points se rapprochent (très relativement) d'une droite d'équation $y=x$. En d'autres termes, les Pokémon ont une statistique défensive qui va plus ou moins suivre leur statistique offensive.

La plupart semblent se grouper autour du point en (50 ; 50).

IV. Analyse de corrélation

1. Matrice de corrélation (avec heatmap)



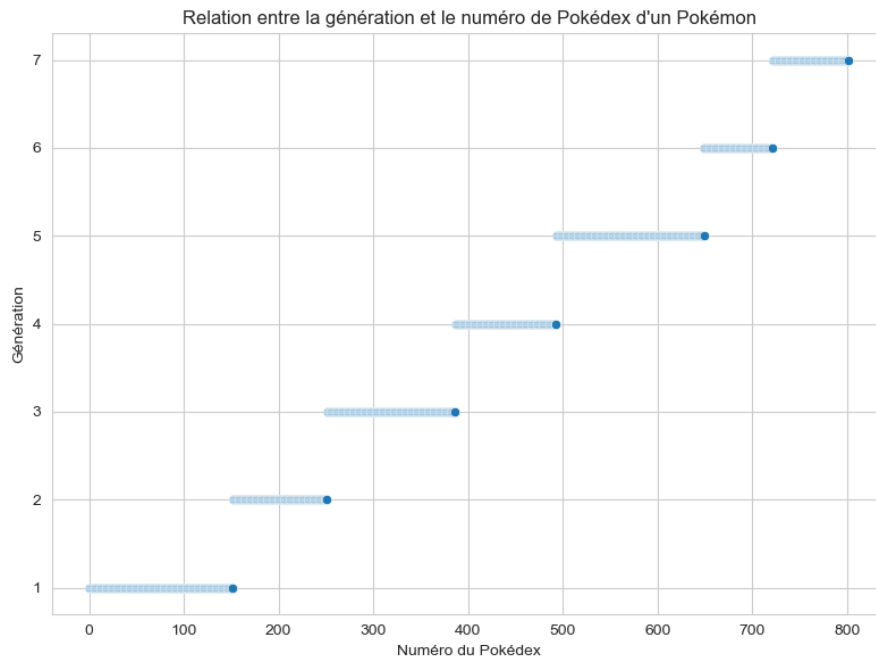
On observe quelques corrélations intéressantes, et d'autres plutôt évidentes. Notamment :

- Le numéro de Pokédex est évidemment proportionnel à la génération d'un pokémon.
- Le poids est corrélé à la taille d'un Pokémon.
- Plutôt étonnant : l'attaque est très impactée par le caractère légendaire d'un pokémon, mais les autres stats pas spécialement !
- Un pokémon légendaire mettra beaucoup de temps à éclore.
- Plus un pokémon met du temps à éclore, moins il est heureux "de base".

2. Visualisation des corrélations

a. Relation entre la génération et le numéro de Pokédex d'un Pokémon

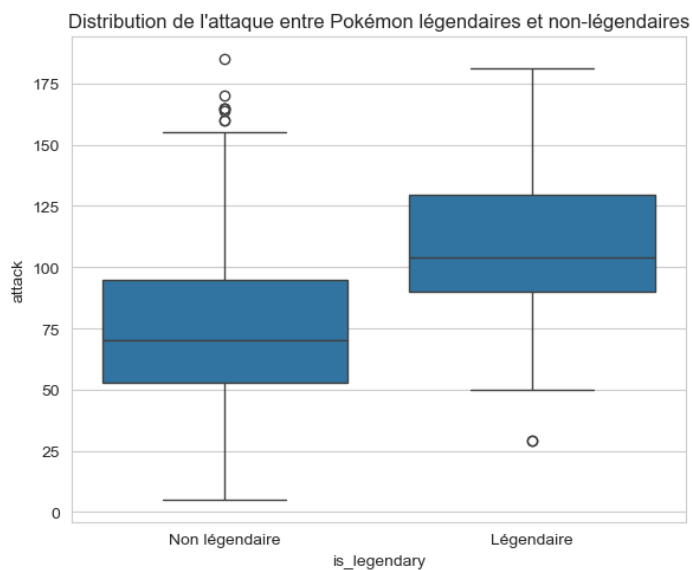
Nous allons visualiser le lien entre numéro du Pokédex et numéro de génération sous forme de nuage de point. On s'attend à voir les points former des paliers, chaque palier représentant une génération.



Ce graphe confirme nos attentes. Traduisons tout de même ce que cela nous montre : plus le numéro de Pokédex d'un Pokémon est petit, plus la génération dont il est issu est ancienne.

b. Relation entre le caractère légendaire d'un Pokémon et sa statistique d'attaque

La deuxième corrélation la plus importante de notre matrice est celle indiquant que les Pokémon légendaires ont généralement plus d'attaque que les autres. Nous allons la visualiser dans un boxplot.

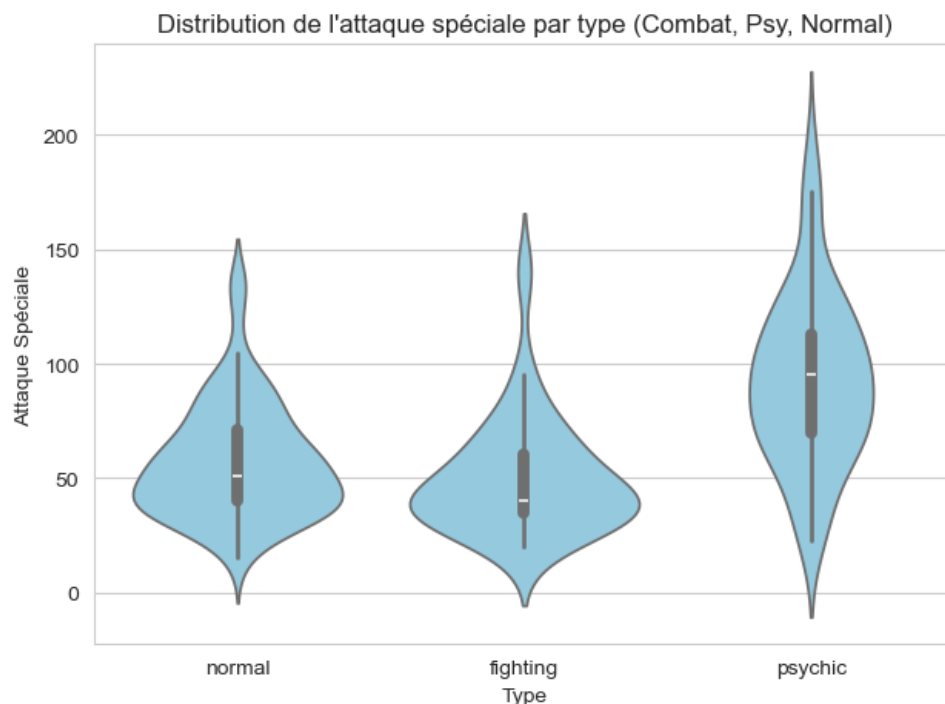


Même s'il y a bien sûr quelques exceptions, ce graphe semble nous confirmer la corrélation :

- La médiane pour les Pokémon légendaire (~102) est plus haute que pour les autres (~70).
- Cela vaut
 - o Premier quartile : ~51 (non-légendaires) contre ~85 (légendaires).
 - o Troisième quartile : ~90 (non-légendaires) contre ~130 (légendaires).
 - o Minimum : ~10 (non-légendaires) contre 50 (légendaires).
 - o Maximum : ~155 (non-légendaires) contre ~180 (légendaires).

3. Comparaison de la distribution de l'attaque spéciale par type (Combat, Psy, Normal)

Nous allons maintenant comparer la distribution de la statistique d'attaque spéciale entre les Pokémon de type Combat, de type Psy et de type Normal. Ce dernier type pourra nous donner une idée de la "normalité" pour cette statistique.



Il est clair et net que les Pokémon de type "Psy" ont une meilleure attaque spéciale. Finalement, les types "Normal" n'ont pas une attaque spéciale particulièrement supérieure à ceux de type "Combat".

La médiane pour le type "Normal" est à 50 et celle pour le type "Combat" est à environ 45. Pour les types "Psy" elle est à approximativement 95.