Pokemon - Overview Analysis

2020-11-23

Contents

1 Introduction

In this paper I will do some analysis of the Pokemon database available on Kaggle.

The database is structured with the following features:

Feature	Description
name	The English name of the Pokemon
japanese_name	The Original Japanese name of the Pokemon
pokedex_number	The entry number of the Pokemon in the National Pokedex
percentage_male	The percentage of the species that are male. Blank if the Pokemon is genderless.
type1	The Primary Type of the Pokemon
type2	The Secondary Type of the Pokemon
classification	The Classification of the Pokemon as described by the Sun and Moon Pokedex
height_m	Height of the Pokemon in metres
weight_kg	The Weight of the Pokemon in kilograms
capture_rate	Capture Rate of the Pokemon
base_egg_steps	The number of steps required to hatch an egg of the Pokemon
abilities	A stringified list of abilities that the Pokemon is capable of having
experience_growth	The Experience Growth of the Pokemon to reach level 100

Feature	Description
base_happiness	Base Happiness of the Pokemon
base_total	Sum of main statistics (hp, attack, defense, sp_attack, sp_defense, speed)
against_?	Eighteen features that denote the amount of damage taken against an attack of a particular type
hp	The Base HP of the Pokemon
attack	The Base Attack of the Pokemon
defense	The Base Defense of the Pokemon
sp_attack	The Base Special Attack of the Pokemon
sp_defense	The Base Special Defense of the Pokemon
speed	The Base Speed of the Pokemon
generation	The numbered generation which the Pokemon was first introduced
is_legendary	Denotes if the Pokemon is legendary.

Let's see a summary of our data :

```
##   abilities          against_bug      against_dark   against_dragon
##  Length:801         Min.   :0.2500   Min.   :0.250   Min.   :0.0000
##  Class :character   1st Qu.:0.5000   1st Qu.:1.000   1st Qu.:1.0000
##  Mode  :character   Median :1.0000   Median :1.000   Median :1.0000
##                     Mean   :0.9963   Mean   :1.057   Mean   :0.9688
##                     3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##                     Max.   :4.0000   Max.   :4.000   Max.   :2.0000
##
##  against_electric against_fairy   against_fight    against_fire
##  Min.   :0.000    Min.   :0.250   Min.   :0.000   Min.   :0.250
##  1st Qu.:0.500    1st Qu.:1.000   1st Qu.:0.500   1st Qu.:0.500
##  Median :1.000    Median :1.000   Median :1.000   Median :1.000
##  Mean   :1.074    Mean   :1.069   Mean   :1.066   Mean   :1.135
##  3rd Qu.:1.000    3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:2.000
##  Max.   :4.000    Max.   :4.000   Max.   :4.000   Max.   :4.000
##
##  against_flying  against_ghost   against_grass   against_ground
##  Min.   :0.250   Min.   :0.000   Min.   :0.250   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.500   1st Qu.:1.000
##  Median :1.000   Median :1.000   Median :1.000   Median :1.000
##  Mean   :1.193   Mean   :0.985   Mean   :1.034   Mean   :1.098
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000
##
##   against_ice    against_normal  against_poison   against_psychic
##  Min.   :0.250   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.500   1st Qu.:1.000   1st Qu.:0.5000   1st Qu.:1.000
##  Median :1.000   Median :1.000   Median :1.0000   Median :1.000
##  Mean   :1.208   Mean   :0.887   Mean   :0.9753   Mean   :1.005
##  3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :4.000   Max.   :1.000   Max.   :4.0000   Max.   :4.000
##
##   against_rock  against_steel    against_water       attack
##  Min.   :0.25   Min.   :0.2500   Min.   :0.250   Min.   :  5.00
##  1st Qu.:1.00   1st Qu.:0.5000   1st Qu.:0.500   1st Qu.: 55.00
```

```
##  Median :1.00   Median :1.0000   Median :1.000   Median : 75.00
##  Mean   :1.25   Mean   :0.9835   Mean   :1.058   Mean   : 77.86
##  3rd Qu.:2.00   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:100.00
##  Max.   :4.00   Max.   :4.0000   Max.   :4.000   Max.   :185.00
##
##  base_egg_steps  base_happiness     base_total    capture_rate
##  Min.   : 1280   Min.   :  0.00   Min.   :180.0   Length:801
##  1st Qu.: 5120   1st Qu.: 70.00   1st Qu.:320.0   Class :character
##  Median : 5120   Median : 70.00   Median :435.0   Mode  :character
##  Mean   : 7191   Mean   : 65.36   Mean   :428.4
##  3rd Qu.: 6400   3rd Qu.: 70.00   3rd Qu.:505.0
##  Max.   :30720   Max.   :140.00   Max.   :780.0
##
##  classfication         defense       experience_growth    height_m
##  Length:801         Min.   :  5.00   Min.   : 600000   Min.   : 0.100
##  Class :character   1st Qu.: 50.00   1st Qu.:1000000   1st Qu.: 0.600
##  Mode  :character   Median : 70.00   Median :1000000   Median : 1.000
##                     Mean   : 73.01   Mean   :1054996   Mean   : 1.164
##                     3rd Qu.: 90.00   3rd Qu.:1059860   3rd Qu.: 1.500
##                     Max.   :230.00   Max.   :1640000   Max.   :14.500
##                                                        NA's   :20
##        hp         japanese_name          name           percentage_male
##  Min.   :  1.00   Length:801         Length:801         Min.   :  0.00
##  1st Qu.: 50.00   Class :character   Class :character   1st Qu.: 50.00
##  Median : 65.00   Mode  :character   Mode  :character   Median : 50.00
##  Mean   : 68.96                                         Mean   : 55.16
##  3rd Qu.: 80.00                                         3rd Qu.: 50.00
##  Max.   :255.00                                         Max.   :100.00
##                                                         NA's   :98
##  pokedex_number   sp_attack        sp_defense         speed
##  Min.   :  1    Min.   : 10.00   Min.   : 20.00   Min.   :  5.00
##  1st Qu.:201    1st Qu.: 45.00   1st Qu.: 50.00   1st Qu.: 45.00
##  Median :401    Median : 65.00   Median : 66.00   Median : 65.00
##  Mean   :401    Mean   : 71.31   Mean   : 70.91   Mean   : 66.33
##  3rd Qu.:601    3rd Qu.: 91.00   3rd Qu.: 90.00   3rd Qu.: 85.00
##  Max.   :801    Max.   :194.00   Max.   :230.00   Max.   :180.00
##
##     type1              type2             weight_kg        generation
##  Length:801         Length:801         Min.   :  0.10   Min.   :1.00
##  Class :character   Class :character   1st Qu.:  9.00   1st Qu.:2.00
##  Mode  :character   Mode  :character   Median : 27.30   Median :4.00
##                                        Mean   : 61.38   Mean   :3.69
##                                        3rd Qu.: 64.80   3rd Qu.:5.00
##                                        Max.   :999.90   Max.   :7.00
##                                        NA's   :20
##   is_legendary
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.08739
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
```

As we can see, there are some missing values in our database. The columns *"height_m",* "percentage_male"_ and _"weight_kg". For my point of view, this columns are not crucial for the analysis, I decide to remove it from the data. I will also drop for the moment the columns "base_egg_steps" and "base_hapiness".

2 Data Wrangling

Right now, we can see that there are only 70 legendary pokemons (around 8%). Also, I would just mention that after checking on internet, it appaears that the 7th generation has 86 new pokemons (based on wikipedia, 81 new for the sun and moon version and 5 for ulta sun and ultra moon. All other generations are the exact amount of new pokemons. Also, for a future task, I would like to add a new column to indicate if a pokemon is a fisrt, a second, or a final evolution. I think it will be a good feature to do a better analyze.

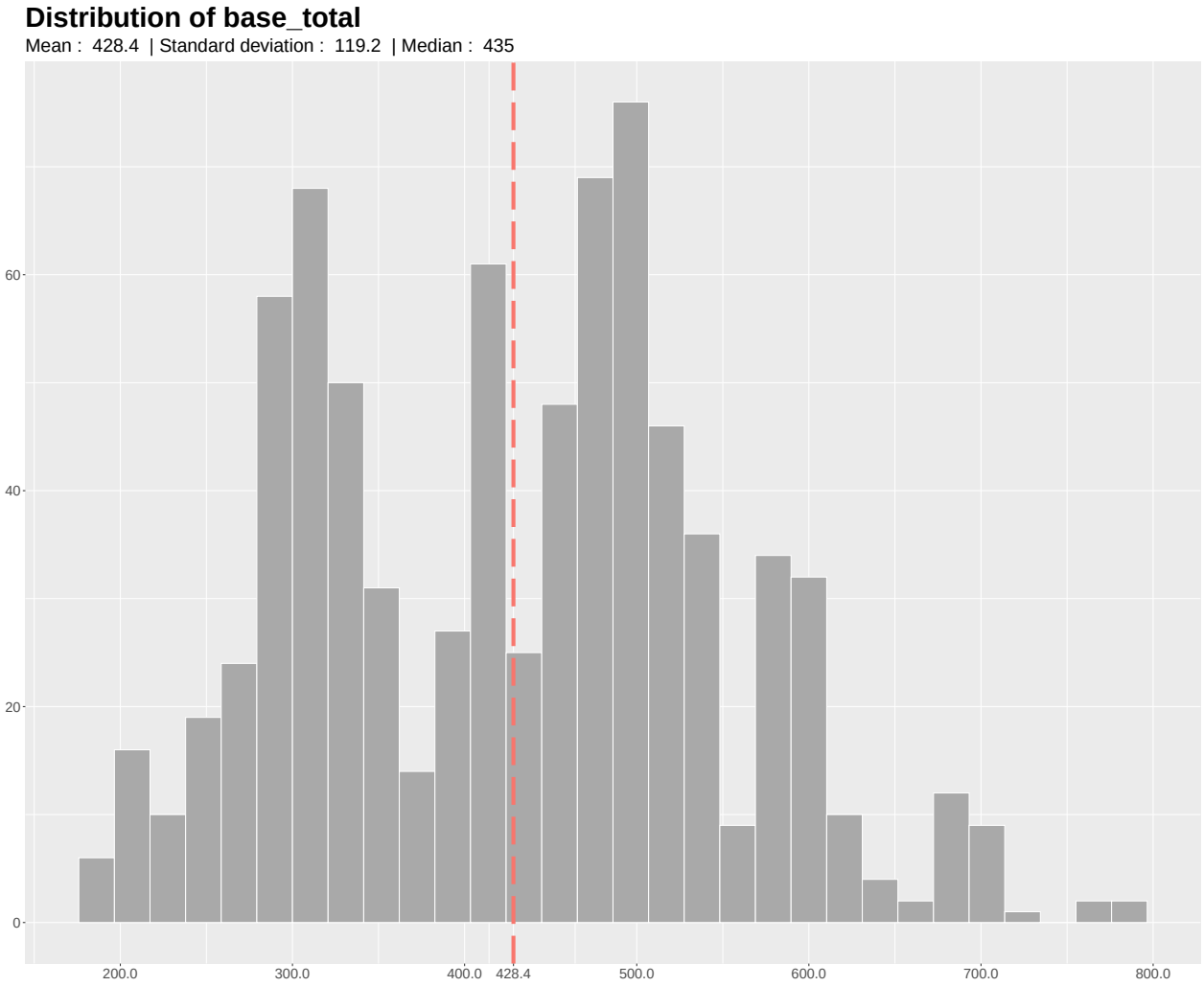
Also we can see an empty value for the type2 feature, this value means that the pokemon has only one type. But I want to give a more explicit value for this pokemon with not a second type. I choose to give the value "None".

```
##    Length     Class      Mode
##       801 character character
```

3 Analysis

3.0.1 Base_total Distribution

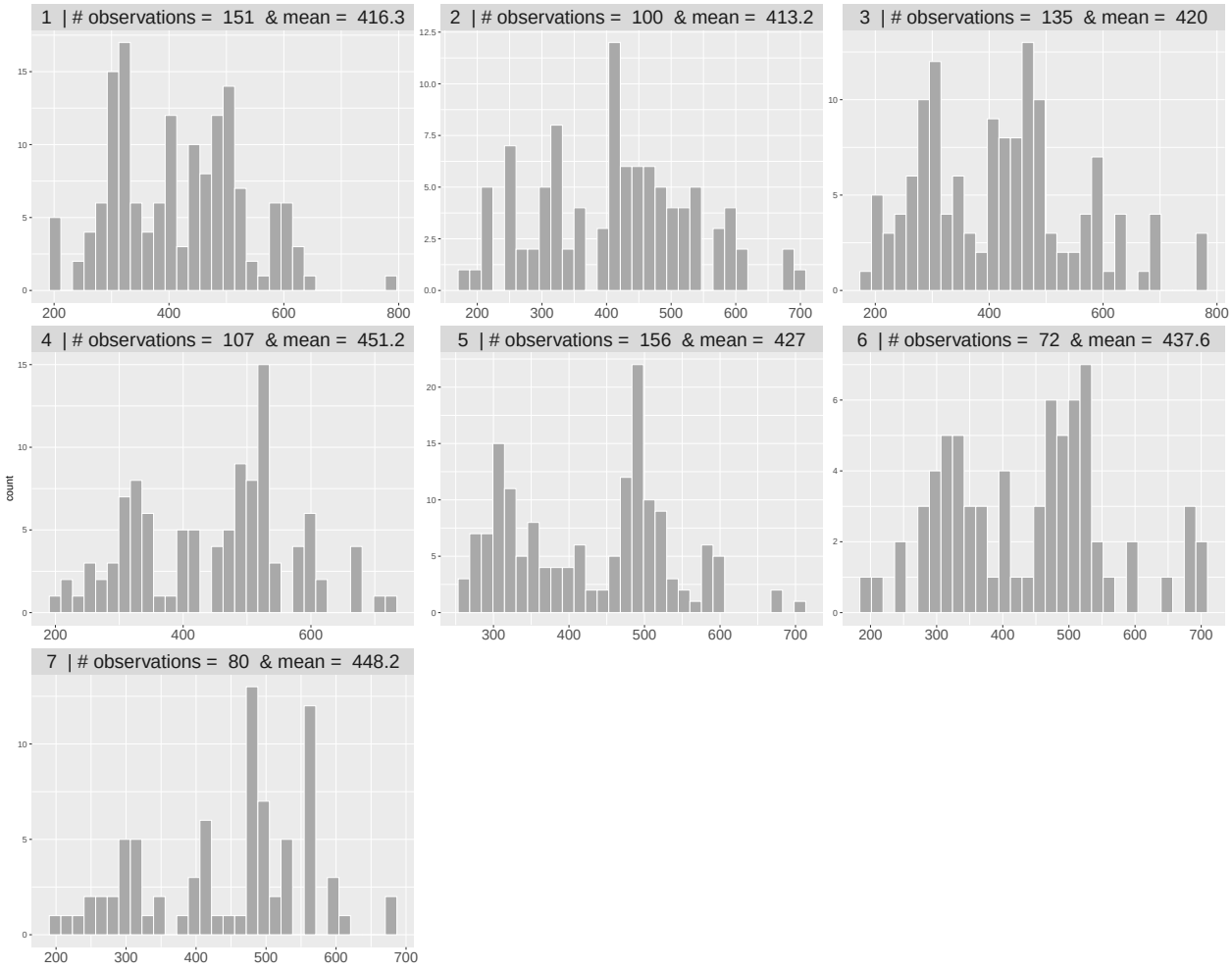
The "base_total" feature represents the sum of the general statistics for each pokemon, thus it is a really good value to determine the power of a pokemon So, It could be a good knowledge to see its repartition.



As wee can see below, the mean is equal to 428 and the median is equal to 435, there is not a big difference between this two estimators. So, despite the 4 outliers with a base_total superior to 700, the repartition is equal around 430. Furthermore the repartition is not distributed normally, there are two peak around 300 and 500, which with more information about if a pokemon is an evolution or not and of which pokemon, we could conclude and have some interesting insight of the repartition of pokemons

3.0.2 Base_total per generation

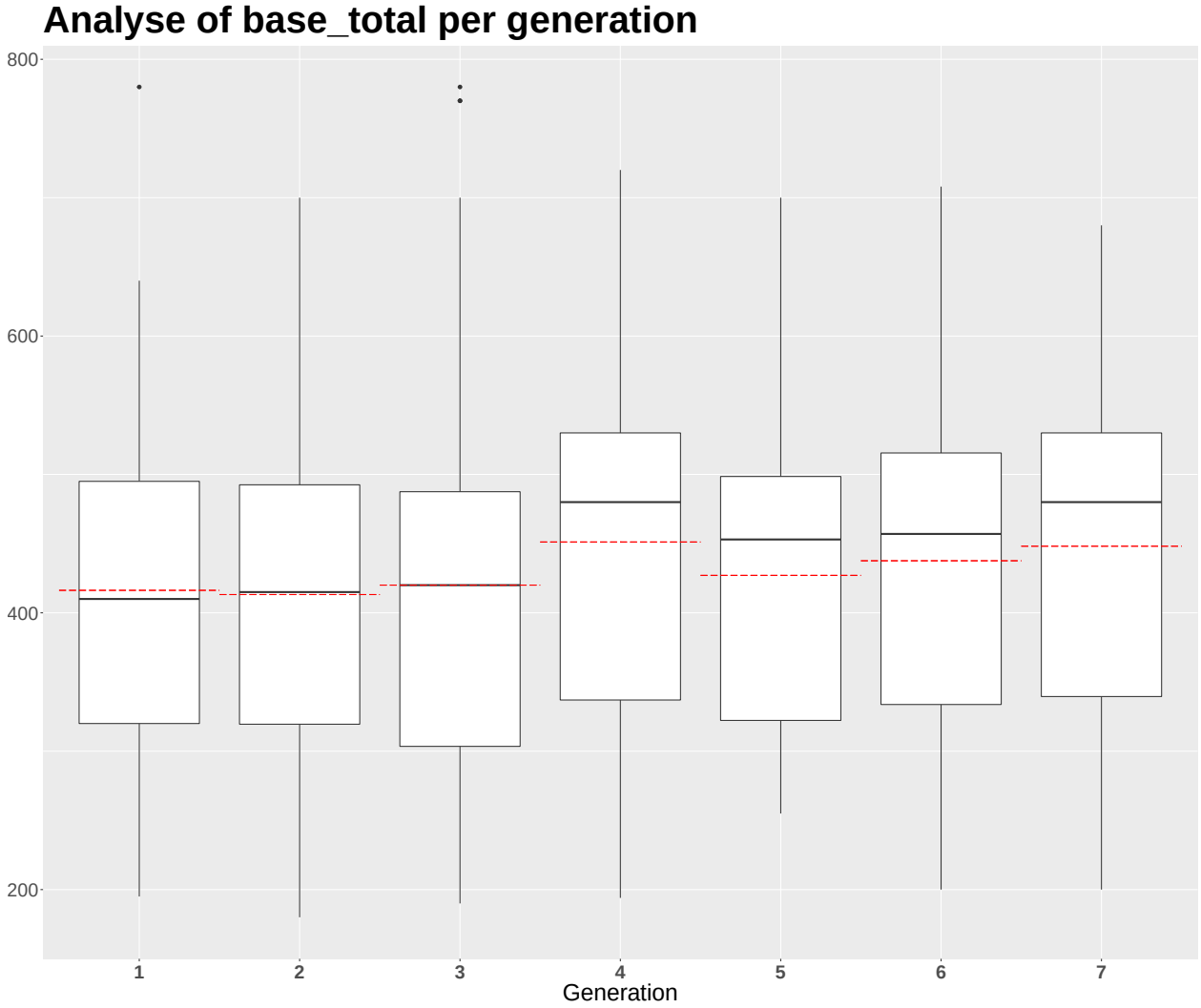
Let's see how differ the base_total according to the generations.



Here some observations :

- The first and secont generation are almost similar related to the mean. But they have the lowest mean.
- The fourth generation has the biggest mean.
- Only the fisrt and third generation have a pokemon around 800. The others the strongest pokemon are around 700.

Maybe we can visualize better with a box plot graph.

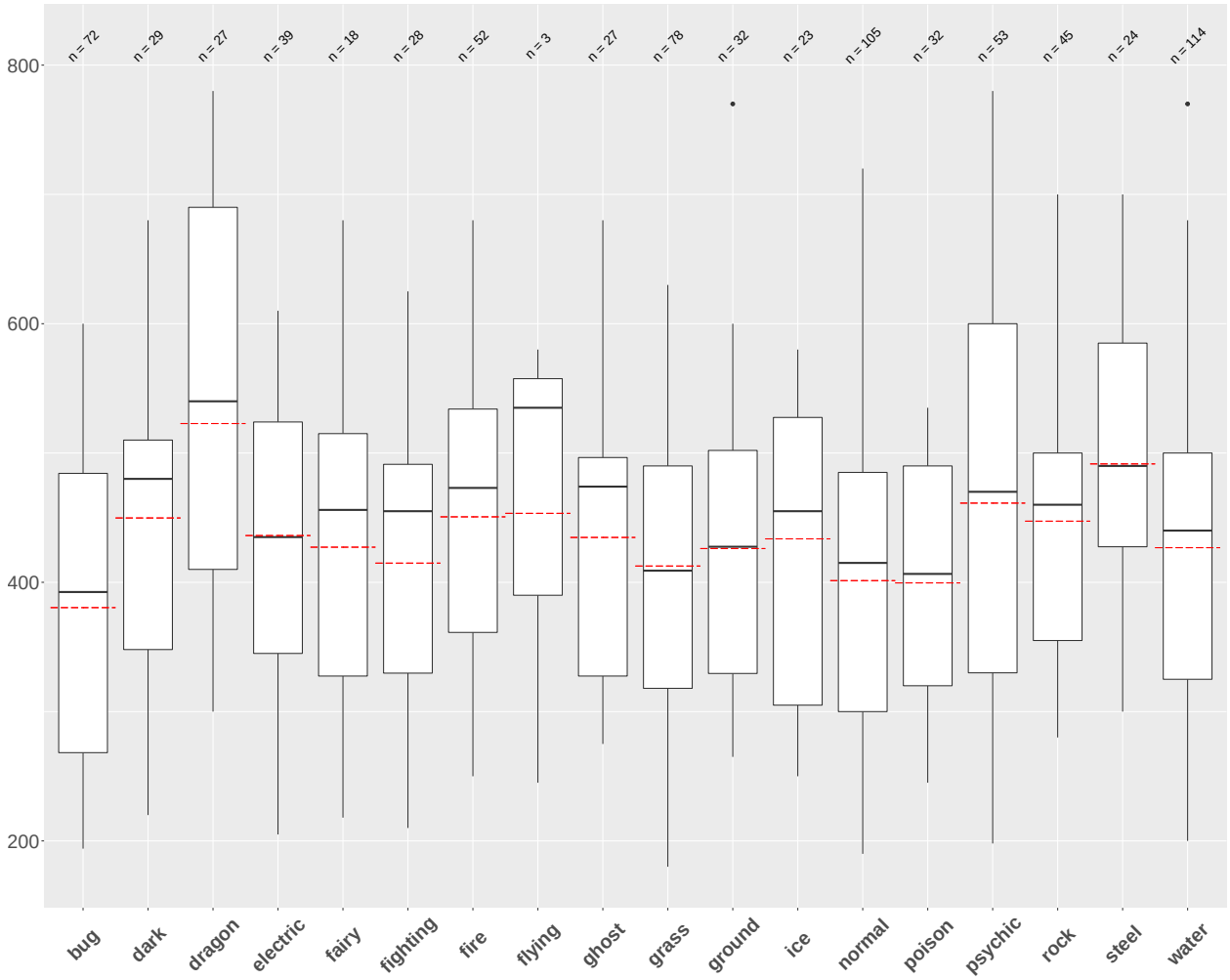


We can notice that the last 4 generation have a stronger pokemon than the first 3 generation. Also, the pokemons of the fifth generation start with a better base_total.

!!!!! Peut-être faire un test sur les moyennes !!!!! =)

3.0.3 Base_total per type

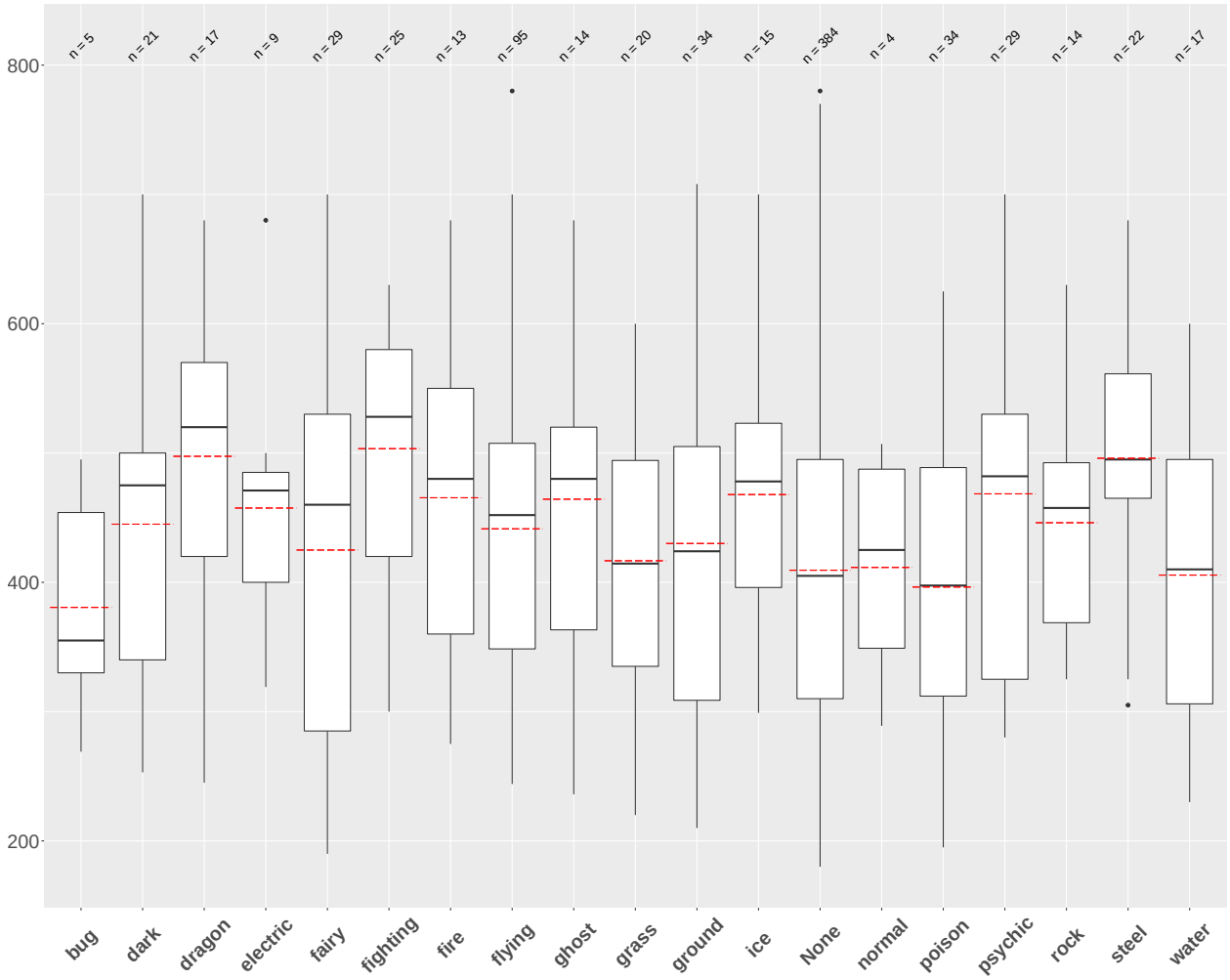
Insted of looking per generation, we can also see the difference are the types related the base_total value.



Some of this categories appears as extremely rare like flying, fairy with less than 20 pokemons with this type. But the few pokemons with flying in principal type are really good pokemon related on base_total caracteristic with two pokemons superior than 500 so superior than the mean of other type.

Furthermore, the top 3 is composed by dragon, steel and psychic type.

Then, I want to compare this results with the type2 variable, because some pokemon has 2 types.



There are 48 % of Pokemons who have not a second type, and if a pokemon has a second type, 20% of the time he will have a flying type.

Here, the top 3 is formed by fighting, dragon and steel.

In comparaison with the top of the type1 (dragon, steel and psychic), The pokemons with a dragon and steel type are on average strongest (also consider fire and psychic) .

Type	n_WO_Type2	prop_population_WO_Type2	mean_base_WO_Type2	n_W_Type2	mean_base_W_Type2
fairy	16	0.89	421.2	2	475.0
fighting	22	0.79	396.7	6	481.3
electric	26	0.67	418.7	13	471.1
psychic	35	0.66	441.3	18	500.0
normal	61	0.58	418.0	44	378.1
water	61	0.54	409.5	53	446.7
fire	27	0.52	407.3	25	497.4
ice	12	0.52	446.9	11	419.1
grass	37	0.47	393.5	41	429.8
dragon	12	0.44	403.9	15	617.9
poison	13	0.41	374.9	19	416.4
flying	1	0.33	580.0	2	390.0

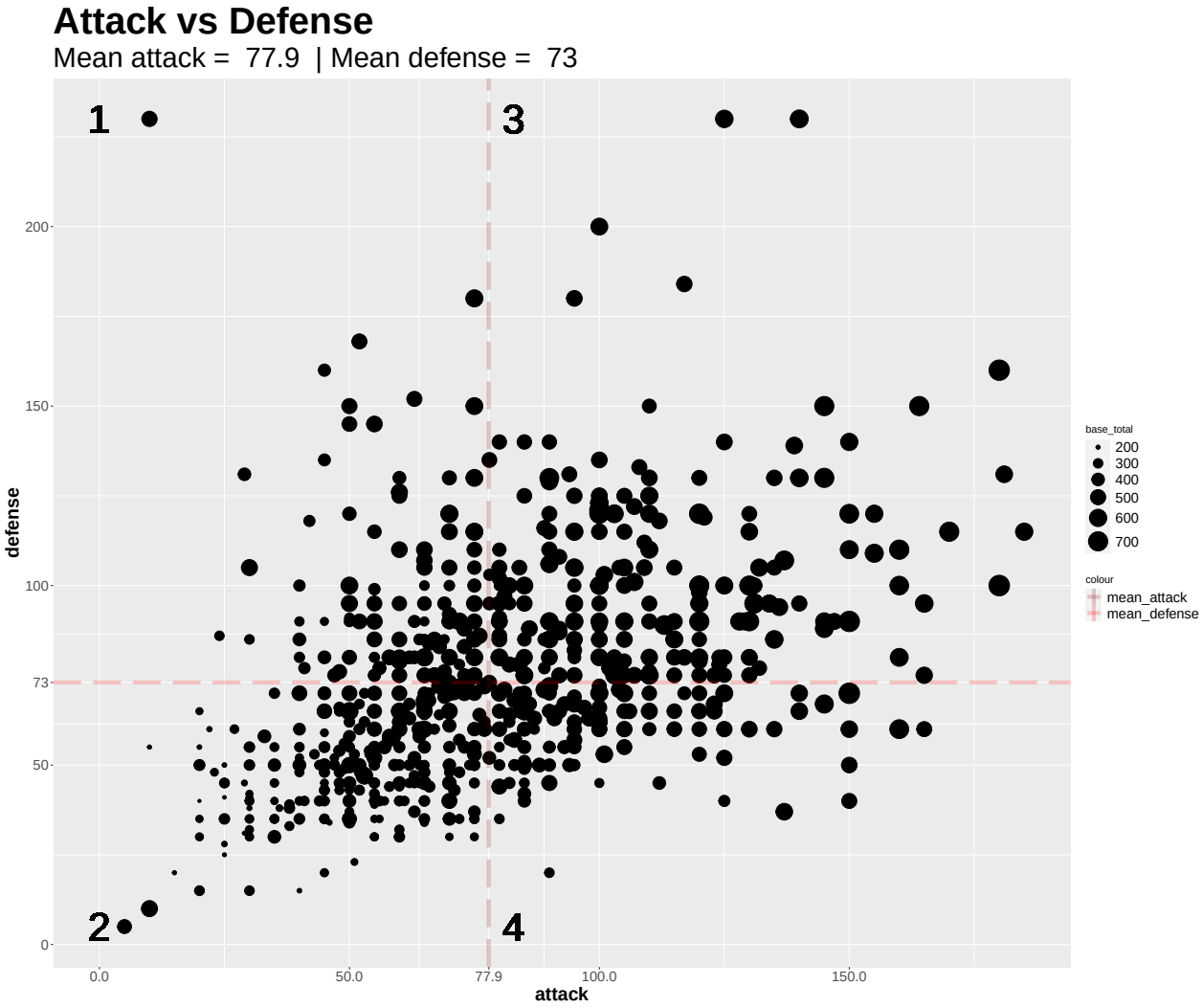
Type	n_WO_Type2	prop_population_WO_Type2	mean_base_WO_Type2	n_W_Type2	mean_base_W_Type2
ghost	9	0.33	426.8	18	438.7
dark	9	0.31	433.0	20	457.3
ground	10	0.31	427.8	22	425.3
bug	18	0.25	299.2	54	407.5
rock	11	0.24	404.7	34	461.0
steel	4	0.17	460.0	20	497.9

Regarding the proportion of pokemons with only one type, the fairy, fighting, electric and psychic are more common with only one type. Also, pokemons with two types are stronger than those with only one type.

3.1 Base_total per classification

3.2 Attack vs Defense

Another good way to compare pokemon between each other, it is to analyze the attack versus the defense.



As we can se on the graph below, we can separate the pokemons in 4 zones through the mean values of attack and defense.

- Zone 1 : This zone has the more defensive pokemon with also a average base_total score. If a player has a defensive strategic this pokemon could be a good fit.
- Zone 2 : This zone covers pokemons who are not really good, they are neither attacker pokemons nor defenser and they have a low base_total score.
- Zone 3 : This zone contains the more versatile pokemon with a good base_total score.
- Zone 4 : This zone has the more attacker pokemon with also a average base_total score. If a player has an offensive strategic this pokemon could be a good fit. They are the opposite of the defensive pokemon.

4 Classification

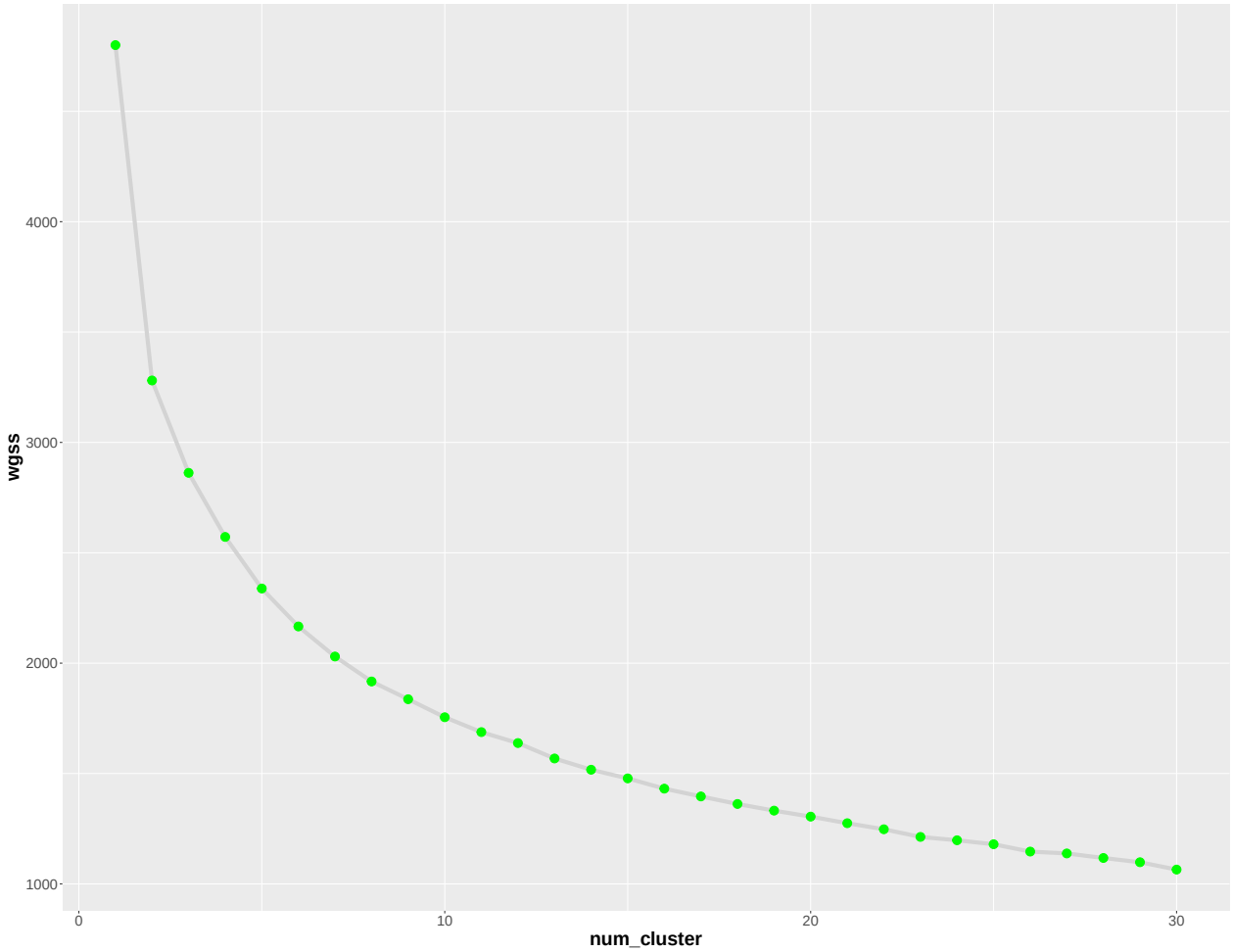
Now we have done some basic analysis, I want to see if we can group the pokemons with same caractetistic. To do this I will use some classification algorithm.

4.1 Classification with only the main statistic

For the first analysis I will use only the main statistic, attack, defense, hp, speed, sp_attack, sp_defense.

4.1.1 K-means Clustering

Here I will implement the K-means clustering algorithm. The first step is to determine the number of cluster we want.



Here, seven or eight cluster seem to be a good choices because it represents the “elbow bends”, so I will choose eight cluster.

The K-means analysis has split the data into 8 groups that we can see above. Let’s see some statistic for each cluster.

```
##
##      1  2  3  4  5  6  7
##   1 12 15 15 13 16 11 14
##   2 17 13 11 14 28  7 15
##   3  3  3  3  2  1  1  1
##   4 32 17 21 18 24 15  9
##   5 14  7 18 18 15  7  9
##   6 33 25 33 21 30 13 13
##   7  4  5  9  6  4  3  6
##   8 36 15 25 15 38 15 13
```

Here are some observations : - The Cluster 1 correponds to versatiles pokemon, those which are near or higher than the average of each meain caracteristics. - The Cluster 2 regroups the pokemon with a good attack stat - The Cluster 3 has the pokemon with a high HP. - The Cluster 4 regroups the medium pokemon. - The Cluster 5 contains the most of the legendary pokemon and the strongest pokemon. - The Cluster 6 regroups the weakest pokemon. - The Cluster 7 has the best defensive pokemon. - The Cluster 8 regroups

also the weakest pokemon.

4.1.2 SVM

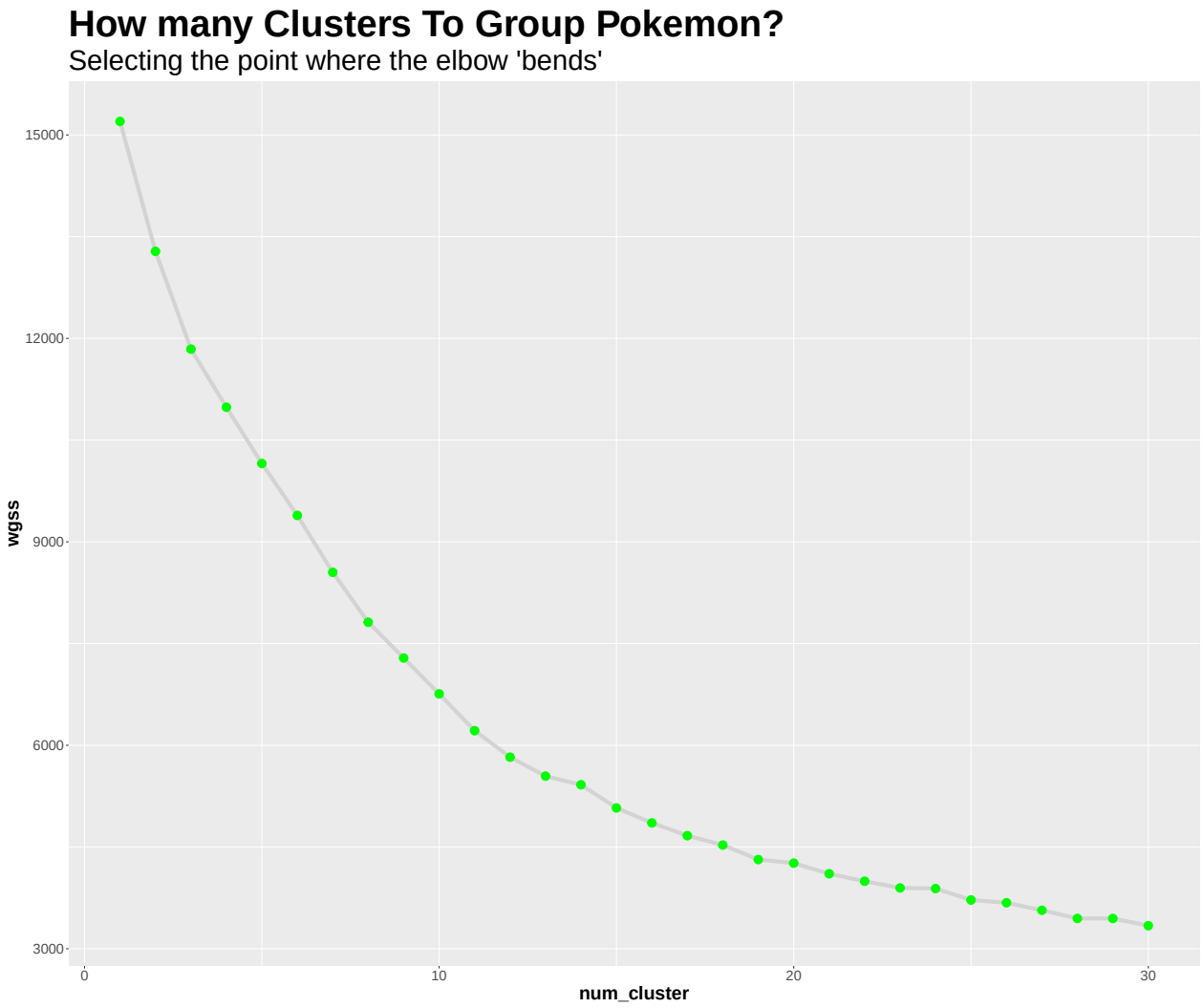
I will use this algorithm soon

4.2 All Stat Caracteristics

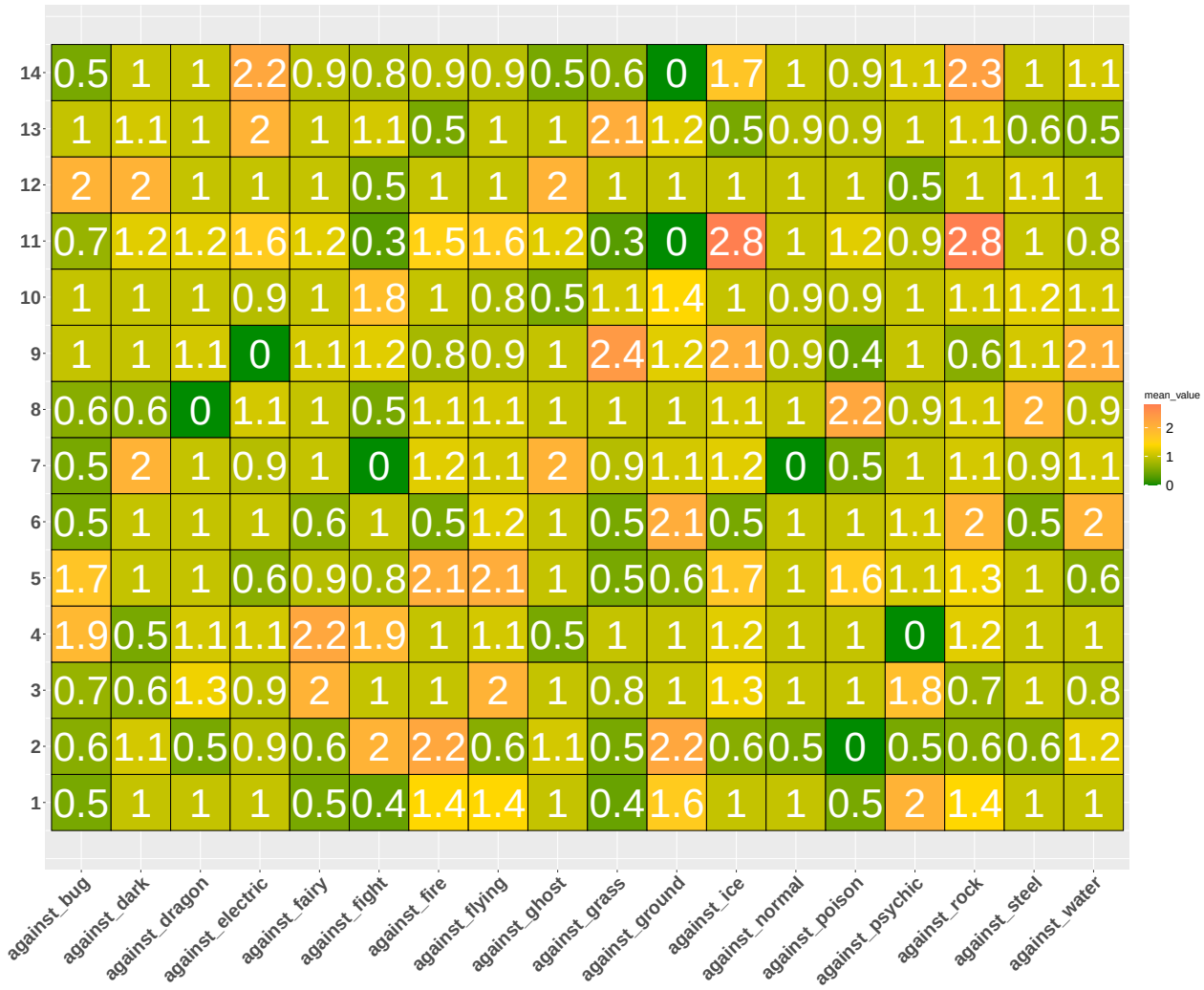
Now I will use all the numerical caracteristic.

4.2.1 K-means Clustering

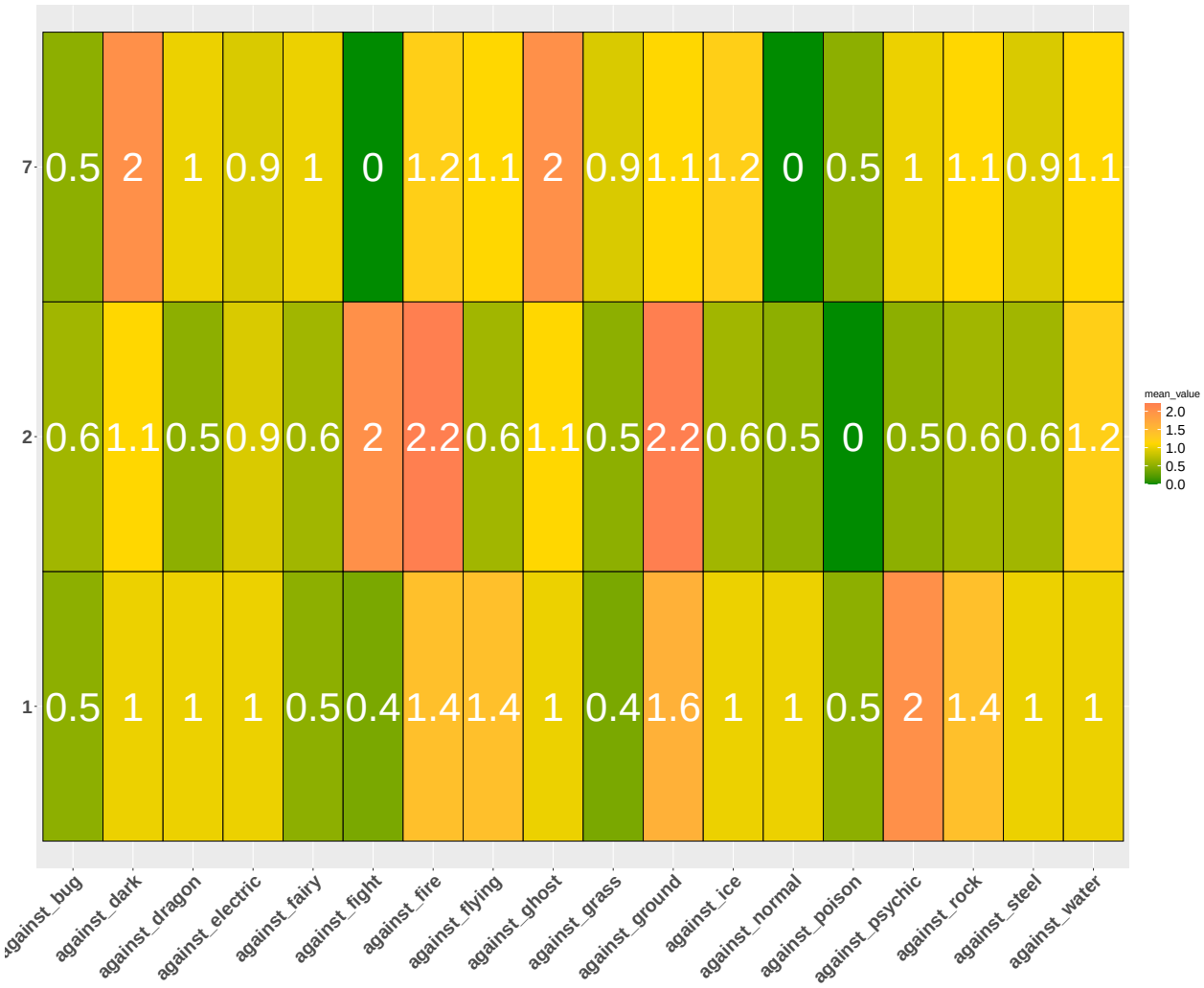
```
##  [1] "pokedex_number"    "name"              "against_bug"
##  [4] "against_dark"      "against_dragon"    "against_electric"
##  [7] "against_fairy"     "against_fight"     "against_fire"
## [10] "against_flying"    "against_ghost"     "against_grass"
## [13] "against_ground"    "against_ice"       "against_normal"
## [16] "against_poison"    "against_psychic"   "against_rock"
## [19] "against_steel"     "against_water"     "attack"
## [22] "base_total"        "capture_rate"      "classfication"
## [25] "defense"           "experience_growth" "hp"
## [28] "sp_attack"         "sp_defense"        "speed"
## [31] "type1"             "type2"             "generation"
## [34] "is_legendary"      "mean_base_stat"    "concattype"
```



The K-means analysis has split the data into 14 groups that we can see above. Let's plot the average score for each cluster for each against_type.



I would like to see the three top cluster related the average of all against group (so with the lowest average domage taken).



This 3 clusters don't seem to have one against group with a big domage taken which is a good point and some types against the value is zero. So, we can create a team with good stat for most of the battle.

To conclude I would like to see the top 3 pokemon related the base_total stat of each of this cluster.

name	base_total	type1	type2	generation	is_legendary	Cluster
Metagross	700	steel	psychic	3	0	2
Dialga	680	steel	dragon	4	1	2
Giratina	680	ghost	dragon	4	1	7
Hoopa	680	psychic	ghost	6	1	7
Solgaleo	680	psychic	steel	7	1	2
Lunala	680	psychic	ghost	7	1	7
Muk	500	poison	poison	1	0	1
Beedrill	495	bug	poison	1	0	1
Dragalge	494	poison	dragon	6	0	1

Only three of this pokemon are not legendary. Maybe it will be good to do a classification without the legendary pokemon (nest step).