Pràctica 2

Pere Moles i Cristina Vilageliu

Descripció del dataset

El joc de dades seleccionat és PokemonData. Hem obtingut aquest dataset del banc de dades Kaggle (disponible a https://www.kaggle.com/datasets/abcsds/pokemon).

El món dels videojocs guanya cada dia més adeptes i fins i tot hi ha gent que viu de jugar-hi, ja sigui participant a tornejos, fent videos a Youtube o jugant en directe en plataformes d'streaming.

Un dels grans clàssics dels videojocs, i també dels dibuixos animats i còmics japonesos, és sense dubte **Pokemon**.

El videojoc ens posa en la pell d'un entrenador d'aquests particulars animals fantàstics anomenats Pokemon. Els Pokemon tenen diferents característiques, evolucionen i gràcies a ells els seus entrenadors poden guanyar tornejos. Els entrenadors, a més, han de reunir tants Pokemon com puguin i aspiren a tenir els millors per a poder guanyar en les lligues que s'organitzen. Però, quins son els millors?

Gràcies a aquest dataset podrem analitzar les diferents característiques dels Pokemon per a poder determinar quins son els millors tipus de Pokemon i competir per a guanyar tots els tornejos!

Integració i selecció.

El fitxer de dades conté 800 observacions i 12 variables.

Tenim 3 variables categòriques, 8 variables numèriques i 1 variable dicotòmica.

Les variables que formen aquest dataset son:

'#' (identificador), 'Name'(nom), 'Type1' (grup al que pertany un Pokemon depenent de les seves característiques), 'Type2'(alguns Pokemons tenen dos tipus, 'HP' (Salut), 'Attack' (potència d'atac), 'Defense' (capacitat de defensa), 'SpAtk' (atac especial), 'SpDef' (defensa especial), 'Speed' (velocitat), 'Generation' (generació) i 'Legendary' (si son llegendaris o no ho son).

Les variables d'interès seran 'HP', 'Attack', 'Defense', 'SpAtk', 'SpDef', 'Speed', 'Generations' i 'Type 1'.

Neteja de les dades

Zeros i elements buits

Anem a donar un cop d'ull a les dades per a veure si aquestes contenen valors nuls o elements buits.

```
colSums(is.na(pokemon))
```

```
##        Num       Name      Type1      Type2         HP     Attack    Defense
##          0          0          0          0          0          0          0
##      SpAtk      SpDef      Speed Generation  Legendary
##          0          0          0          0          0
```

Veiem que no hi ha cap columna amb elements buits.

```
colSums(pokemon == "")
```

```
##        Num       Name      Type1      Type2         HP     Attack    Defense
##          0          0          0        386          0          0          0
##      SpAtk      SpDef      Speed Generation  Legendary
##          0          0          0          0          0
```

També veiem que no hi ha cap observació en blanc.

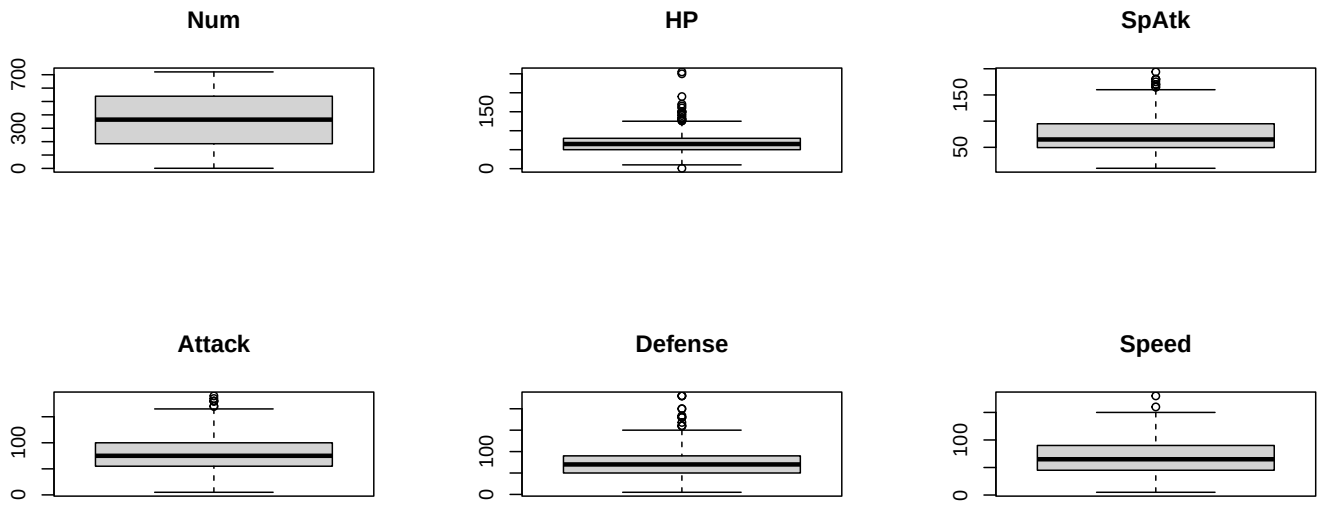
```
colSums(pokemon == 0)
```

```
##        Num       Name      Type1      Type2         HP     Attack    Defense
##          0          0          0          0          0          0          0
##      SpAtk      SpDef      Speed Generation  Legendary
##          0          0          0          0        735
```

I, per últim, veiem que no hi ha cap variable amb zeros o valors nuls. La única columna que en té és la de 'Legendary', però aquesta és una variable dicotòmica on 0 és un valor amb significat (el Pokemon en qüestió NO és legendari).

Valors extrems

Anem a analitzar ara els valors extrems i determinar si son errors o no.



Veiem que les variables numèriques amb més outliers són 'HP' (o salut), 'Defense' i 'SpDef'. No els podem considerar outliers ja que determinen valors de salut i defensa per sobre de la mitja, és a dir, descriuen a Pokemon amb molta resistència. Aquesta dada podria ser interessant a l'hora d'escollir quin és el millor Pokemon així que no elimanerem els outliers ja que no són errors i ens poden donar informació valiosa.

Anàlisi de les dades

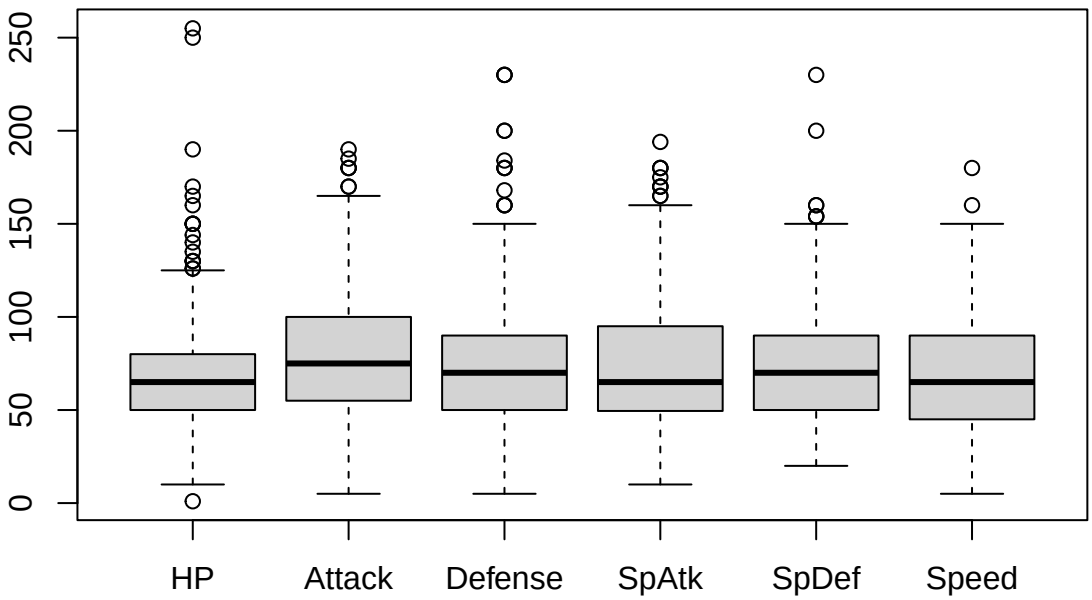
Anàlisi descriptiva

Estadistiques Pokemon

```
summary(pokemon_stats)
```

```
##        HP             Attack       Defense           SpAtk
##  Min.   :  1.00   Min.   :  5   Min.   :  5.00   Min.   : 10.00
##  1st Qu.: 50.00   1st Qu.: 55   1st Qu.: 50.00   1st Qu.: 49.75
##  Median : 65.00   Median : 75   Median : 70.00   Median : 65.00
##  Mean   : 69.26   Mean   : 79   Mean   : 73.84   Mean   : 72.82
##  3rd Qu.: 80.00   3rd Qu.:100   3rd Qu.: 90.00   3rd Qu.: 95.00
##  Max.   :255.00   Max.   :190   Max.   :230.00   Max.   :194.00
##      SpDef           Speed
##  Min.   : 20.0   Min.   :  5.00
##  1st Qu.: 50.0   1st Qu.: 45.00
##  Median : 70.0   Median : 65.00
##  Mean   : 71.9   Mean   : 68.28
##  3rd Qu.: 90.0   3rd Qu.: 90.00
##  Max.   :230.0   Max.   :180.00
```

```
boxplot(pokemon_stats)
```



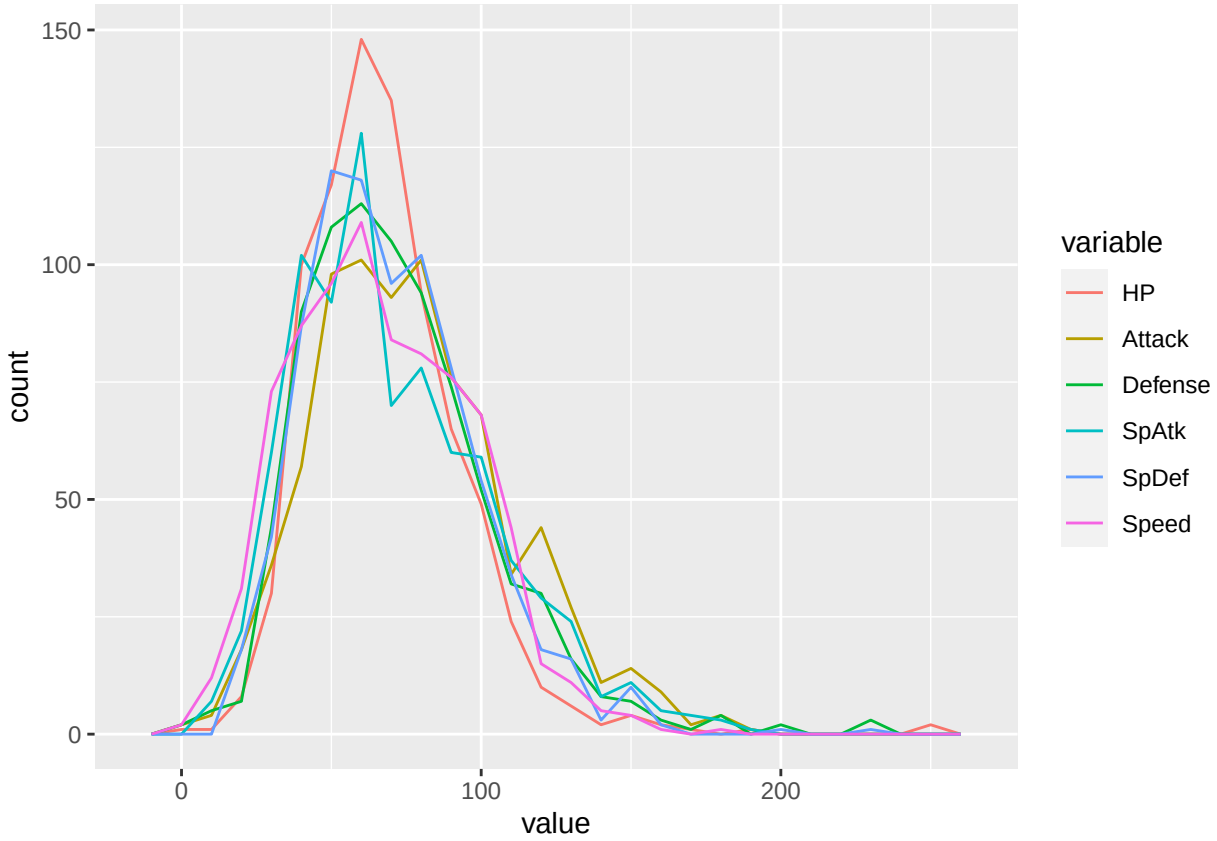
La funció summary ens permet obtenir un resum de les mesures de tendència central i dispersió mes importants per tal de realitzar l'anàlisi descriptiu de les dades.

Hem de tenir en compte que els stats de Pokemon son en base a 255. Nomes s'assoleix aquesta xifra en algun Pokemon per l'estadistica de HP. Alguns stats com velocitat (180), atac (190) o atac especial (194) queden lluny d'aquesta xifra.

Podem observar diferències en la mitjana dels stats, i que la mitjana mes alta es la d'Atac (amb 79) i la mes baixa la Velocitat (amb 68.28).

Observem gràficament com està distribuida cada estadística de Pokemon. Ho agrupem en intervals de 10.

```
library(ggplot2)
library(tidyverse)
```



Observem com la distribucio de les diferents estadistiques es bastant similar.

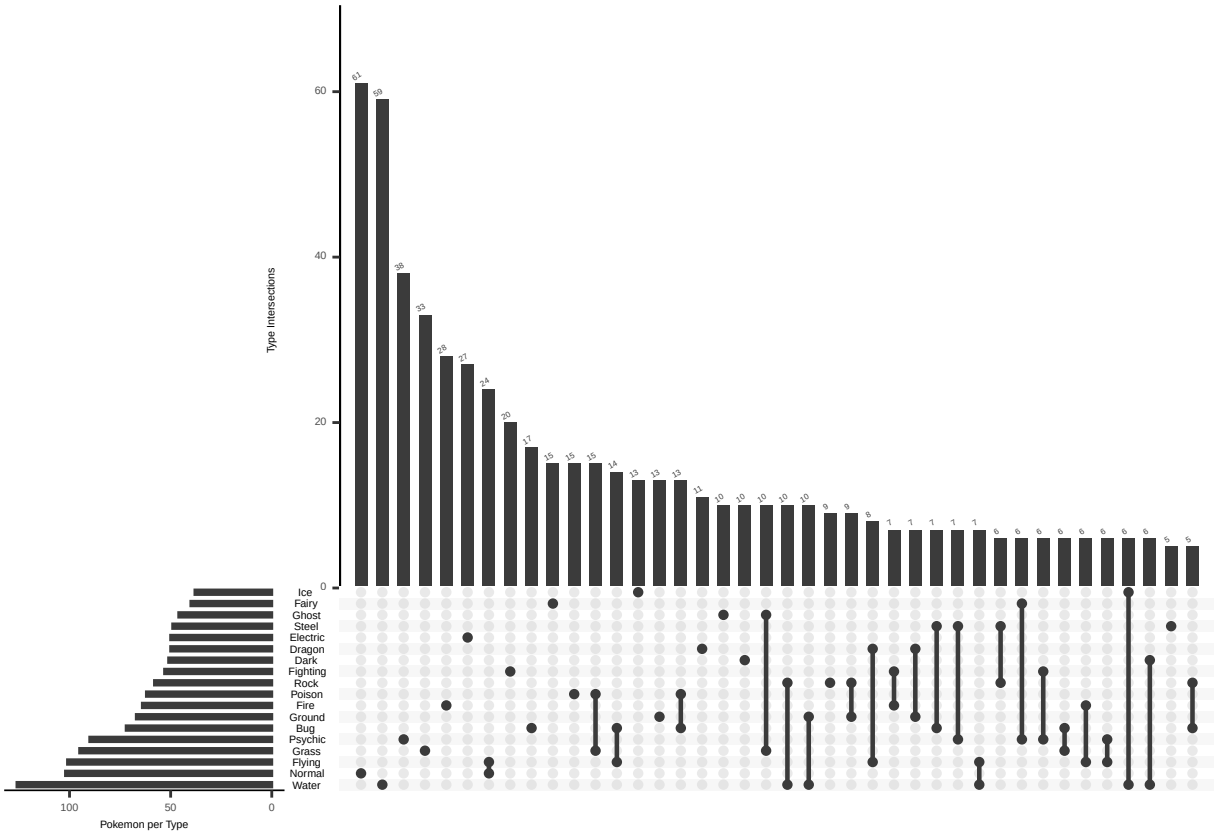
Freqüència dels tipus de Pokemon

Realitzarem una gràfica utilitzant la llibreria Upset que ens permet visualitzar millor la freqüència d'interseccions. Així doncs, podrem visualitzar els parells de tipus més comuns, i també els tipus individualment.

Primer de tot crearem les següents columnes a partir de les variables Type 1 i Type 2:

```
type_list <- unique(pokemon$Type1)

pokemon[type_list]<-0


for(type in type_list){
  pokemon[type] <- ifelse(pokemon$Type1 == type | pokemon$Type2 == type,1,0)
}

pokemon$type<-NULL
```

A continuació creem la gràfica.



Observem que els tipus de Pokemon mes populars son Aigua i Normal, i les combinacions mes populars son Normal-Volador i Planta-Veri.

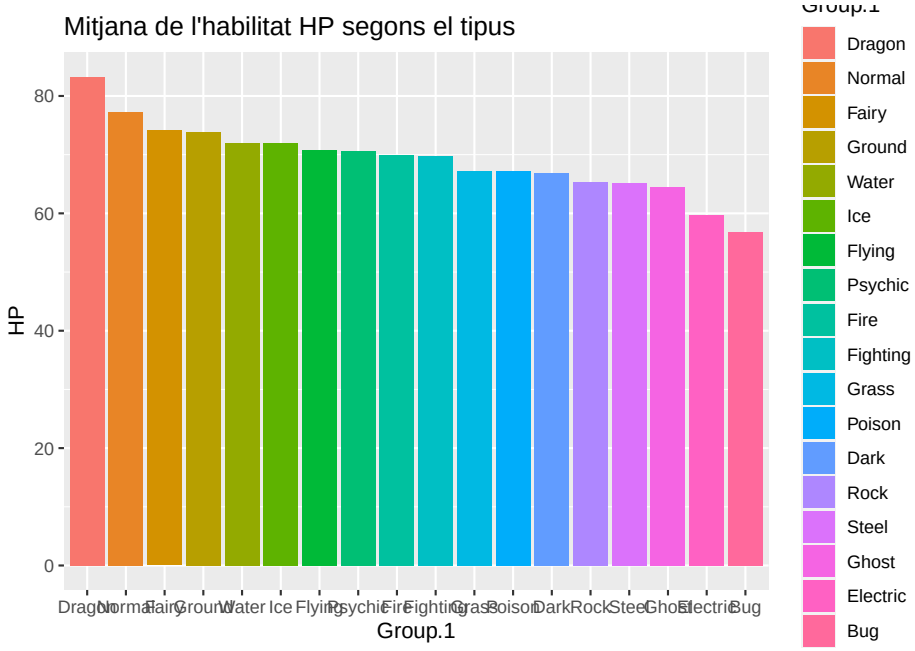
Mitjana de les estadístiques per tipus de Pokemon

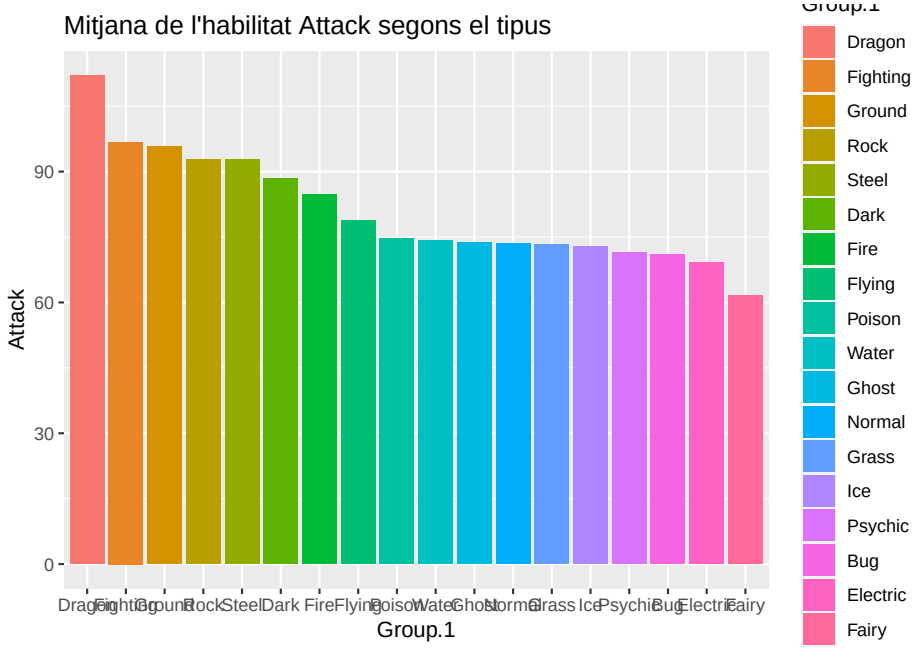
A continuació volem veure la mitjana de cada estadística depenent del tipus de Pokemon.

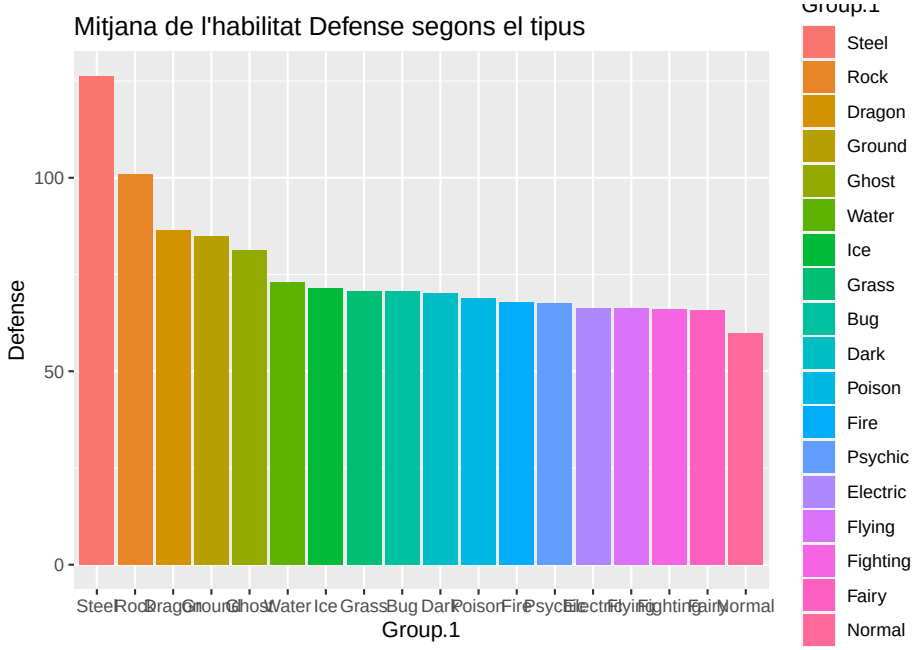
```
df <- aggregate(pokemon$HP, list(pokemon$Type1), FUN=mean)
colnames(df)[2] <- 'HP'


for(i in 6:11){
  df<- merge(df,aggregate(pokemon[i], list(pokemon$Type1), FUN=mean),by="Group.1")
}
df
```

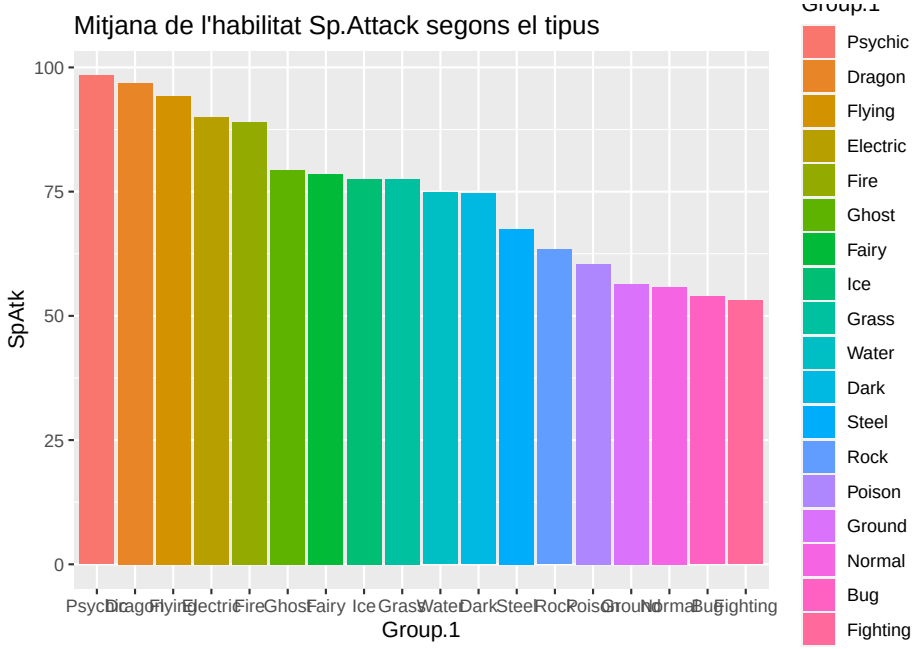
```
##     Group.1       HP    Attack   Defense    SpAtk    SpDef     Speed Generation
## 1       Bug 56.88406  70.97101  70.72464 53.86957 64.79710  61.68116   3.217391
## 2      Dark 66.80645  88.38710  70.22581 74.64516 69.51613  76.16129   4.032258
## 3    Dragon 83.31250 112.12500  86.37500 96.84375 88.84375  83.03125   3.875000
## 4  Electric 59.79545  69.09091  66.29545 90.02273 73.70455  84.50000   3.272727
## 5     Fairy 74.11765  61.52941  65.70588 78.52941 84.70588  48.58824   4.117647
## 6  Fighting 69.85185  96.77778  65.92593 53.11111 64.70370  66.07407   3.370370
## 7      Fire 69.90385  84.76923  67.76923 88.98077 72.21154  74.44231   3.211538
## 8    Flying 70.75000  78.75000  66.25000 94.25000 72.50000 102.50000   5.500000
## 9     Ghost 64.43750  73.78125  81.18750 79.34375 76.46875  64.34375   4.187500
```

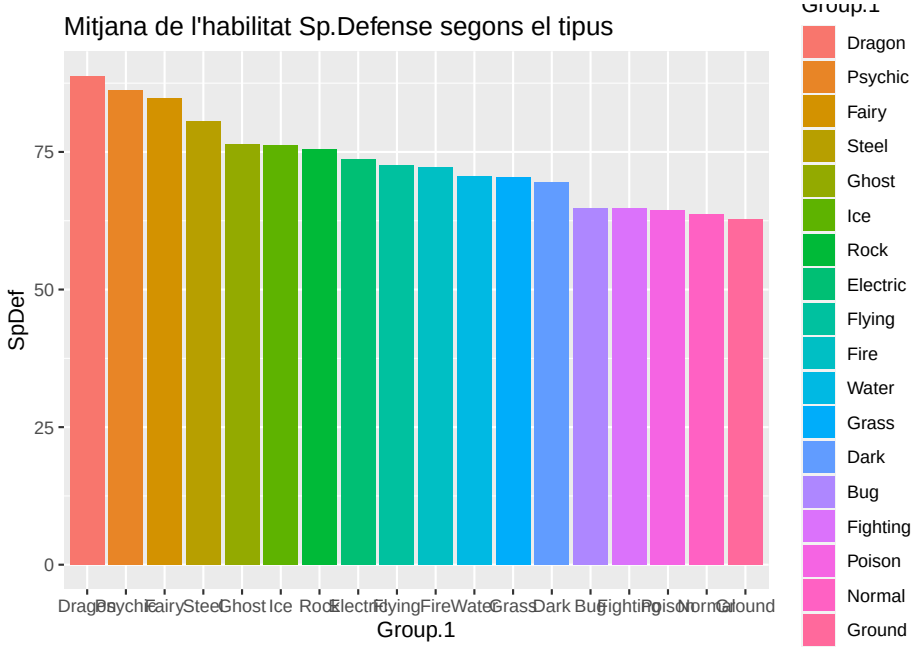
```
## 10    Grass 67.27143  73.21429  70.80000 77.50000 70.42857  61.92857   3.357143
## 11   Ground 73.78125  95.75000  84.84375 56.46875 62.75000  63.90625   3.156250
## 12      Ice 72.00000  72.75000  71.41667 77.54167 76.29167  63.45833   3.541667
## 13   Normal 77.27551  73.46939  59.84694 55.81633 63.72449  71.55102   3.051020
## 14   Poison 67.25000  74.67857  68.82143 60.42857 64.39286  63.57143   2.535714
## 15  Psychic 70.63158  71.45614  67.68421 98.40351 86.28070  81.49123   3.385965
## 16     Rock 65.36364  92.86364 100.79545 63.34091 75.47727  55.90909   3.454545
## 17    Steel 65.22222  92.70370 126.37037 67.51852 80.62963  55.25926   3.851852
## 18    Water 72.06250  74.15179  72.94643 74.81250 70.51786  65.96429   2.857143
```

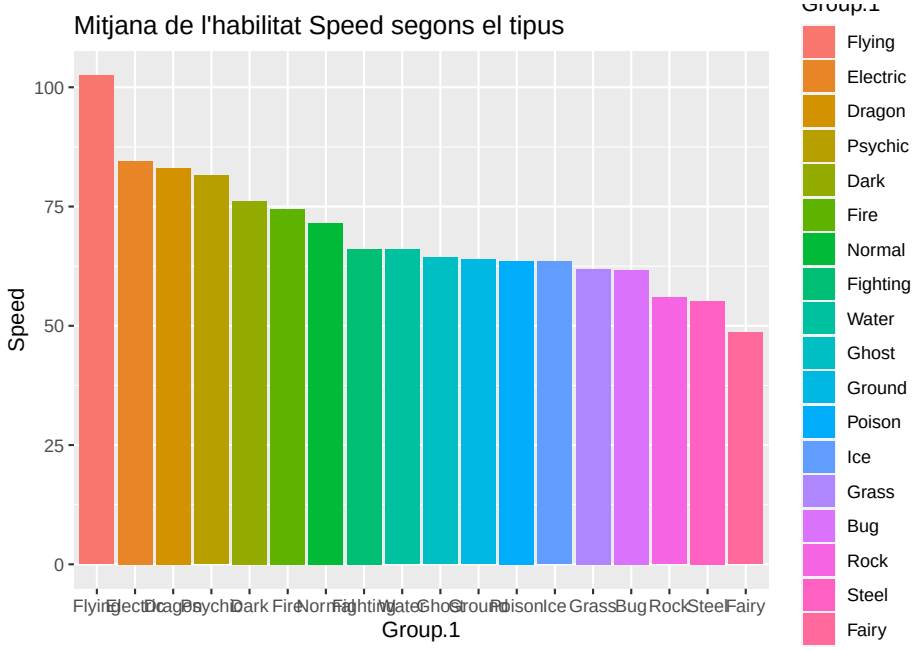





Mitjana de l'habilitat Defense segons el tipus
Defense
100
50
0
Steel
Rock
Dragon
Ground
Ghost
Water
Ice
Grass
Bug
Dark
Poison
Fire
Psychic
Electric
Flying
Fighting
Fairy
Normal
Group.1
Group.1
Steel
Rock
Dragon
Ground
Ghost
Water
Ice
Grass
Bug
Dark
Poison
Fire
Psychic
Electric
Flying
Fighting
Fairy
Normal


Mitjana de l'habilitat Sp.Attack segons el tipus
SpAtk
100
75
50
25
0
Psychic
Dragon
Flying
Electric
Fire
Ghost
Fairy
Ice
Grass
Water
Dark
Steel
Rock
Poison
Ground
Normal
Bug
Fighting
Group.1
Group.1
Psychic
Dragon
Flying
Electric
Fire
Ghost
Fairy
Ice
Grass
Water
Dark
Steel
Rock
Poison
Ground
Normal
Bug
Fighting





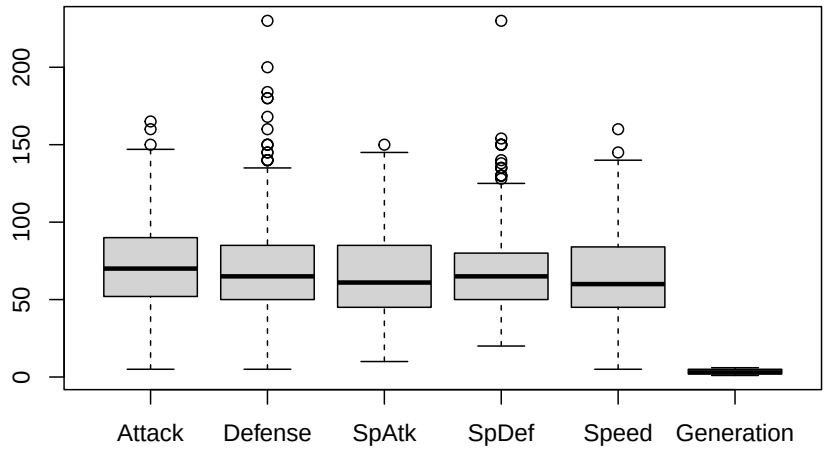
Cas particular: descartem Pokemon Llegendaris i Mega-Evolucions

A continuacio analitzem els canvis que es produeixen si descartem Pokemon Llegendaris i Mega-evolucions, que son els Pokemon mes poderosos del joc.

```
summary(no_legends_and_mega_stats)
```

```
##      Attack          Defense           SpAtk            SpDef
##  Min.   :  5.00   Min.   :  5.00   Min.   : 10.00   Min.   : 20.00
##  1st Qu.: 52.00   1st Qu.: 50.00   1st Qu.: 45.00   1st Qu.: 50.00
```

```
##  Median : 70.00   Median : 65.00   Median : 61.00   Median : 65.00
##  Mean   : 72.83   Mean   : 69.41   Mean   : 66.17   Mean   : 67.19
##  3rd Qu.: 90.00   3rd Qu.: 85.00   3rd Qu.: 85.00   3rd Qu.: 80.00
##  Max.   :165.00   Max.   :230.00   Max.   :150.00   Max.   :230.00
##      Speed          Generation
##  Min.   :  5.00   Min.   :1.000
##  1st Qu.: 45.00   1st Qu.:2.000
##  Median : 60.00   Median :3.000
##  Mean   : 63.93   Mean   :3.338
##  3rd Qu.: 84.00   3rd Qu.:5.000
##  Max.   :160.00   Max.   :6.000
```



Observem una clara disminució en les diferents variables respecte al resum anterior on estaven inclosos els Pokemon Llegendaris i Mega-Evolucions.

Anàlisi de la normalitat i homoscedasticitat

Comprovem si les dades tenen una distribució normal amb el test de Shapiro-Wilk.

```
shapiro.test(pokemon$HP)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pokemon$HP
## W = 0.91583, p-value < 2.2e-16
```

```
shapiro.test(pokemon$Attack)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  pokemon$Attack
## W = 0.97893, p-value = 2.472e-09
```

```
shapiro.test(pokemon$Defense)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pokemon$Defense
## W = 0.93806, p-value < 2.2e-16
```

```
shapiro.test(pokemon$SpAtk)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pokemon$SpAtk
## W = 0.95954, p-value = 4.665e-14
```

```
shapiro.test(pokemon$SpDef)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pokemon$SpDef
## W = 0.96077, p-value = 8.252e-14
```

```
shapiro.test(pokemon$Speed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pokemon$Speed
## W = 0.98416, p-value = 1.31e-07
```

Com veiem, totes les variables d'interès tenen un p-valor més petit que el nivell de significança així que podem dir que cap de les nostres variables té una distribució normal.

Anem ara a comprobar la homoscedasticitat amb el test de Fligner-Killeen.

```
fligner.test(Attack ~ SpAtk, data = pokemon)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Attack by SpAtk
## Fligner-Killeen:med chi-squared = 145.49, df = 104, p-value = 0.004538
```

```
fligner.test(Defense ~ SpDef, data = pokemon)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Defense by SpDef
## Fligner-Killeen:med chi-squared = 117.68, df = 91, p-value = 0.03141
```

```
fligner.test(HP ~ Speed, data = pokemon)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  HP by Speed
## Fligner-Killeen:med chi-squared = 192.91, df = 107, p-value = 7.087e-07
```

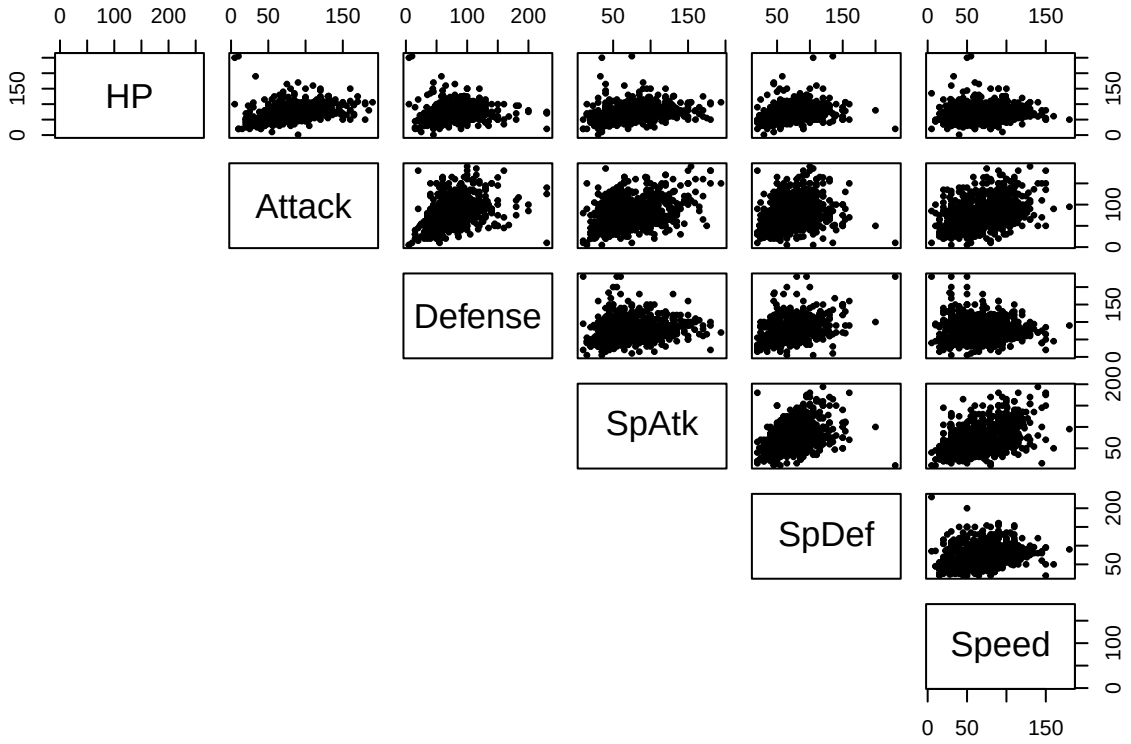
Aquestes son algunes de les comparacions que hem fet per analitzar la igualtat de variàncies. Podem veure, en la mostra més rellevant, que el p-valor està per sota del nivell de significança en tots els casos així que tampoc podem garantir la homoscedasticitat de els dades.

Anàlisi de correlacions

Volem veure si existeix alguna correlació entre les diferents estadístiques dels Pokemon.

Primer de tot, dibuixem els diferents diagrames de dispersió de totes les variables.

```
pairs(pokemon_stats, pch = 19,  cex = 0.5,  lower.panel=NULL)
```



A primera vista, sembla que podem observar alguna lleugera correlació entre algunes variables.

Per a comprovar-ho, observarem la matriu de correlacions.

```
corr <- cor(pokemon_stats)
corr
```

```
##                HP    Attack   Defense     SpAtk     SpDef     Speed
## HP      1.0000000 0.4223860 0.2396223 0.3623799 0.3787181 0.1759521
## Attack  0.4223860 1.0000000 0.4386871 0.3963618 0.2639896 0.3812397
## Defense 0.2396223 0.4386871 1.0000000 0.2235486 0.5107466 0.0152266
## SpAtk   0.3623799 0.3963618 0.2235486 1.0000000 0.5061214 0.4730179
## SpDef   0.3787181 0.2639896 0.5107466 0.5061214 1.0000000 0.2591331
## Speed   0.1759521 0.3812397 0.0152266 0.4730179 0.2591331 1.0000000
```

Podem observar que les parelles de stats que estan més correlacionades son:

- La defensa i la defensa especial
- L'atac especial i la defensa especial
- Velocitat i atac especial

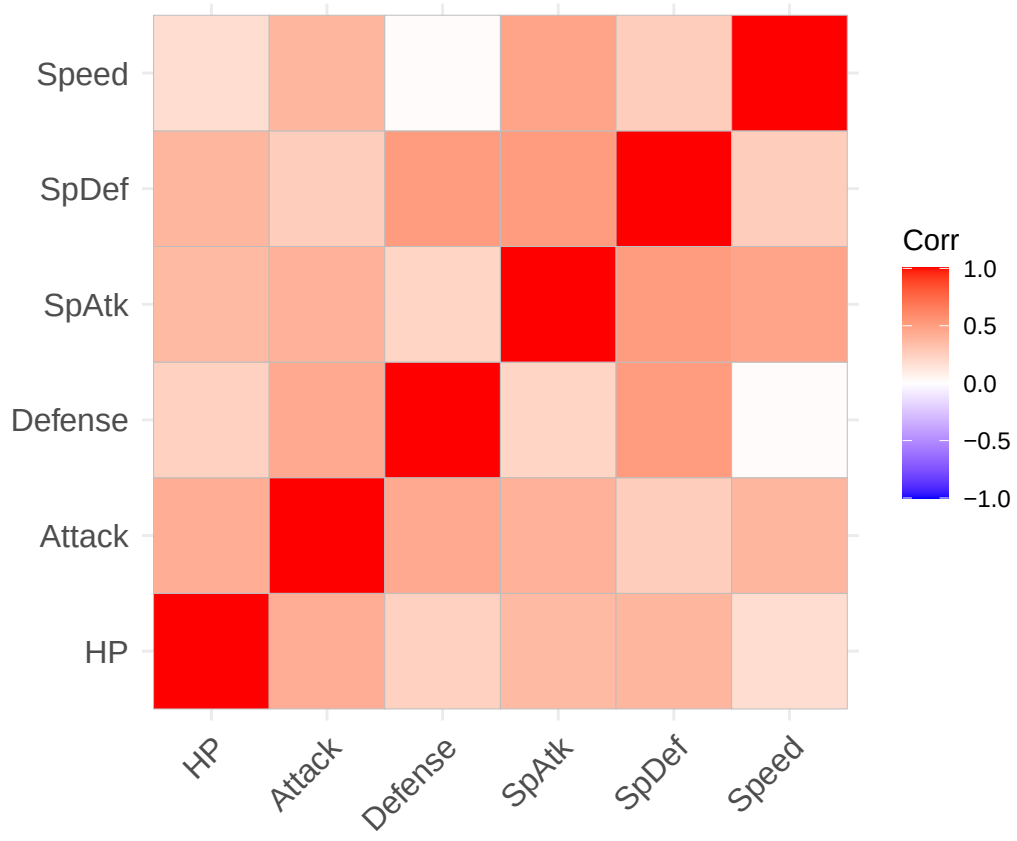
En canvi, les variables menys correlacionades son:

- Velocitat i Defensa
- Velocitat i HP

- Defensa i atac especial.

Podem visualitzar-ho amb la seguent gràfica

```
library(ggcorrplot)
ggcorrplot(corr)
```



Comparació entre més de dos grups

Per a seguir amb el nostre anàlisi, farem una comparació entre la variable 'Type1' i la resta de variables numèriques. EL què pretenem és veure si tots els possibles tipus de Pokemon tenen característiques similars o si, per el contrari, difereixen.

Per a fer-ho, farem servir l'alternativa no paramètrica als contrastos d'hipòtesis de més de 2 grups: el test de Kruskal-Wallis. Utilitzem aquest mètode ja que les dades no segueixen una distribució normal.

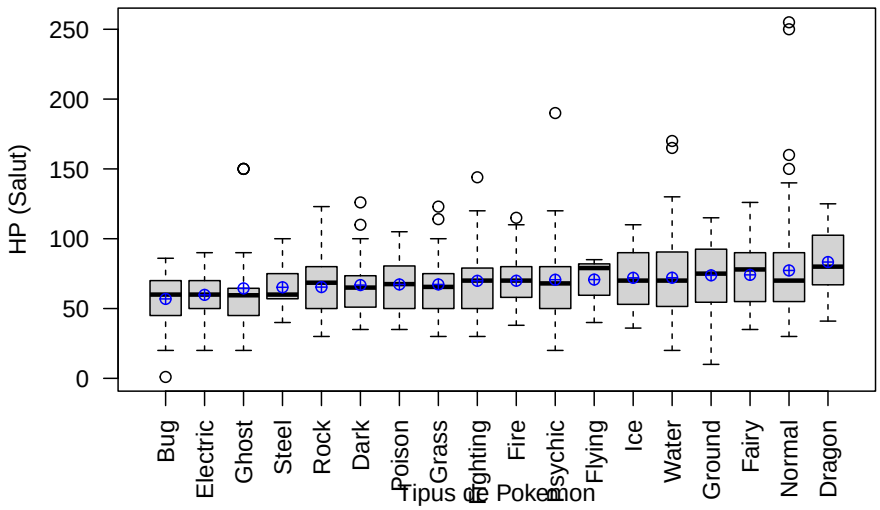
El nivell de confiança que utilitzarem és del 95%.

```
kruskal.test(HP ~ Type1, data = pokemon)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  HP by Type1
## Kruskal-Wallis chi-squared = 49.375, df = 17, p-value = 5.28e-05
```

El p-valor és més petit que el nivell de significança (0.05) així que podem concloure que els diferents tipus de Pokemon ('Type1') tenen diferents nivells de 'HP' (Salut).

Observem en un gràfic les diferents mitjanes de cada tipus.



Veiem que els Pokemon de tipus Bug, Electric i Ghost son els que menys 'HP' tenen.

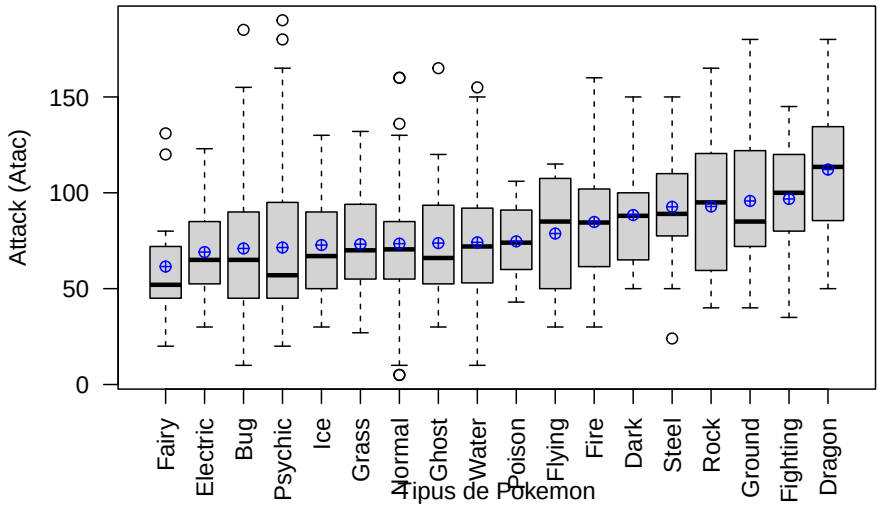
Els tipus Dragon, Normal i Fairy són els que més 'HP' tenen.

```
kruskal.test(Attack ~ Type1, data = pokemon)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Attack by Type1
## Kruskal-Wallis chi-squared = 92.874, df = 17, p-value = 1.826e-12
```

Per a la variable 'Attack' (Atac) el p-valor és més petit que el nivell de significança (0.05) així que podem concloure que els diferents tipus de Pokemon ('Type1') tenen diferents nivells de 'Attack'.

Observem en un gràfic les diferents mitjanes de cada tipus.



Els tipus de Pokemon amb un atac més fluix són Fairy, Electric i Bug.

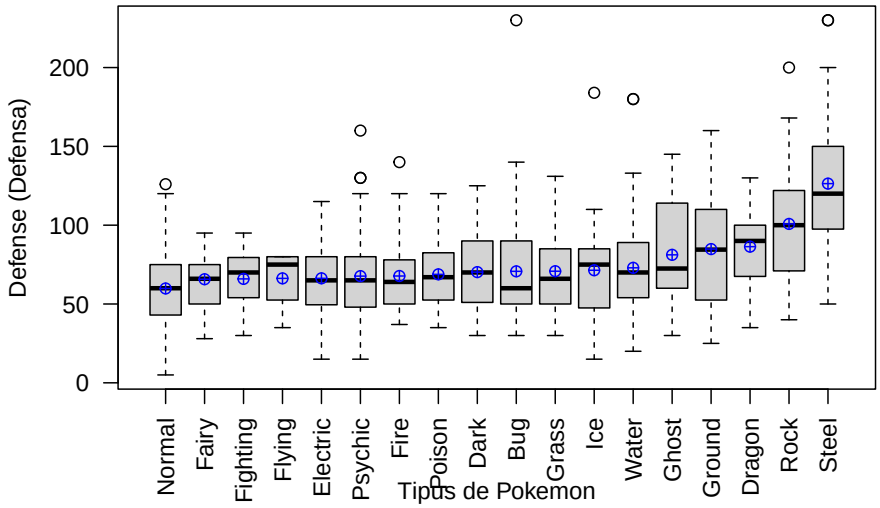
Per altre banda, els tipus de Pokemon amb millor atac són: Dragon, Fighting i Ground.

```
kruskal.test(Defense ~ Type1, data = pokemon)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Defense by Type1
## Kruskal-Wallis chi-squared = 110.4, df = 17, p-value = 1.016e-15
```

Per a la variable 'Defense' (Defensa) el p-valor és més petit que el nivell de significança (0.05) així que podem concloure que els diferents tipus de Pokemon ('Type1') tenen diferents nivells de 'Defense'.

Observem en un gràfic les diferents mitjanes de cada tipus.



En el gràfic boxplot podem veure que els Pokemon amb pitjor capacitat defensiva son els de tipus Normal, Fairy i Fighting.

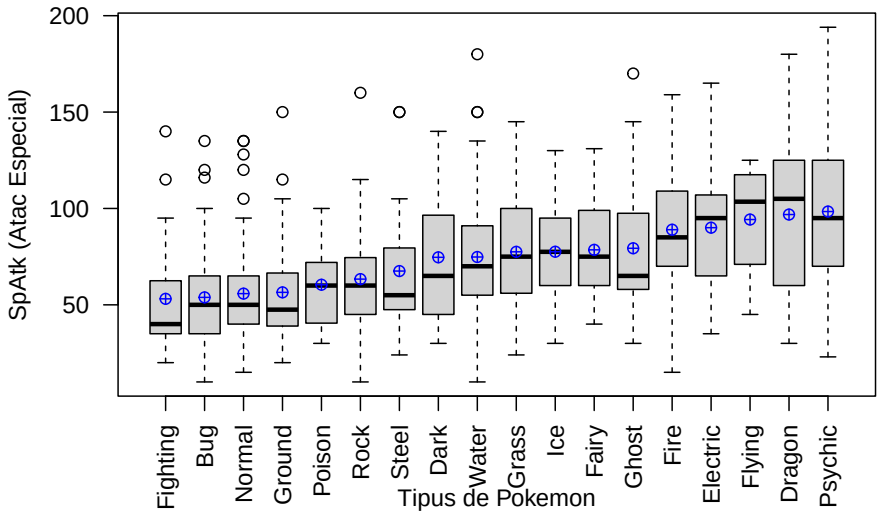
També veiem que els Pokemon amb millor defensa son els de tipus Steel, Rock i Dragon.

```
kruskal.test(SpAtk ~ Type1, data = pokemon)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SpAtk by Type1
## Kruskal-Wallis chi-squared = 166.71, df = 17, p-value < 2.2e-16
```

Per a la variable 'SpAtk' (Atac Especial) el p-valor és més petit que el nivell de significança (0.05) així que podem concloure que els diferents tipus de Pokemon ('Type1') tenen diferents nivells de 'SPAtk'.

Observem en un gràfic les diferents mitjanes de cada tipus.



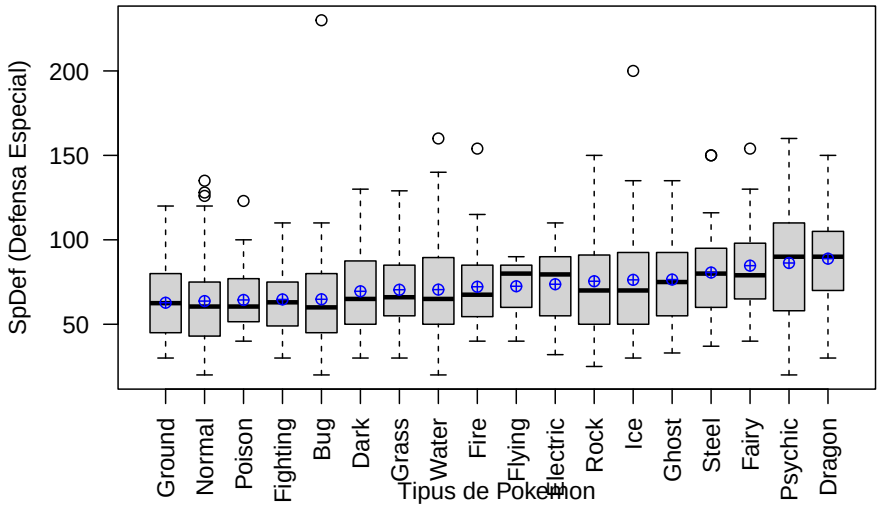
Per als atacs especials és millor no comptar amb Pokemon de tipus Fighting, Bug o Normal i apostar millor per als de tipus Flying, Dragon o Psychic.

```
kruskal.test(SpDef ~ Type1, data = pokemon)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SpDef by Type1
## Kruskal-Wallis chi-squared = 54.96, df = 17, p-value = 6.955e-06
```

Per a la variable 'SpDef' (Defensa Especial) el p-valor és més petit que el nivell de significança (0.05) així que podem concloure que els diferents tipus de Pokemon ('Type1') tenen diferents nivells de 'SPDef'.

Observem en un gràfic les diferents mitjanes de cada tipus.



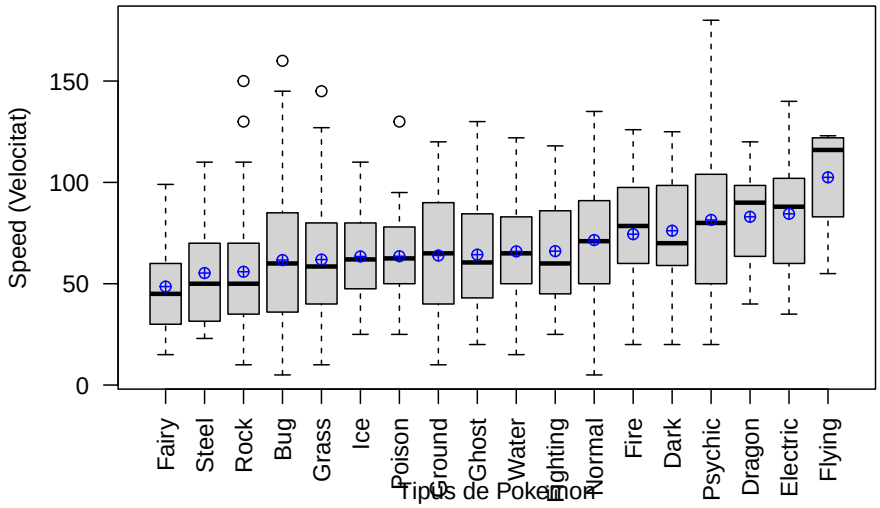
Els Pokemon amb pitjor 'SpDef' de mitjana son Ground, Normal i Poison i els que tenen millor defensa especial son Fairy, Psychic i Dragon.

```
kruskal.test(Speed ~ Type1, data = pokemon)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Speed by Type1
## Kruskal-Wallis chi-squared = 78.827, df = 17, p-value = 6.191e-10
```

Per últim, veiem que la variable 'Speed' (Velocitat) també té un p-valor per sota del nivell de significança (0.05) així que també determinem que les mitjanes de velocitat per als diferents tipus de Pokemon son diferents.

Mirem en un gràfic com es distribueixen els diferents tipus segons la seva velocitat.



Veiem que els Pokemon més lents de mitjana son els de tipus Fairy, Steel i Rock. En canvi, els més ràpids son els de tipus Dragon, Electric i Flying.

Com a conclusió, podem dir que les característiques analitzades per a cada tipus de Pokemon són molt diferents, veiem clarament diferències entre els tipus ('Type1') de Pokemon.

Com a jugadors, ens interessarà reunir la millor quantitat de Pokemon de tipus Dragon, com per exemple Dragonite o Salamence.

Per altre banda, segons el nostre estudi de les característiques principals dels Pokemon, els Pokemon de tipus Bug (Caterpie o Volbeat), Fairy (Clefairy o Swirlix) o Normal (Pidgey o Rattata) no ens interessarien tant ja que son els que pitjors estadístiques tenen.

Conclusions

El procés de neteja ha sigut senzill donat que les dades no contenen valors nuls ni zeros. Hem trobat una variable amb valors en blanc però no ha afectat en el nostre estudi així que no ens ha fet falta tractar-la.

L'estadística de Pokemon amb la mitjana més alta és Attack. En canvi, l'estadística amb la mitjana més baixa es Speed.

Si excloem els Pokemon llegendaris i les Mega-Evolucions, observem una disminució notable en les mitjanes de les diferents estadístiques dels Pokemon.

Hem comprovat que les dades no segueixen una distribució normal ni tenen variàncies iguals (homoscedasticitat).

Els tipus de Pokemon més freqüents son Aigua i Normal, i les combinacions més populars son normal-Volador i Planta-Veri.

Observem una notable correlació entre la defensa i la defensa especial, l'atac especial i la defensa especial i per ultim, la velocitat i l'atac especial. En canvi, observem una correlació baixa entre Velocitat i Defensa, Velocitat i HP, Defensa i atac especial.

Hem comparat els diferents tipus de Pokemon en funció de cada característica i hem vist que hi ha diferències d' 'stats' en cada grup. Els Pokemon més forts són els de tipus Dragon i els més dèbils són els tipus Bug, Fairy i Normals, segons el nostre estudi.

Contribucions	Signatura
Investigació prèvia	PM/CV
Redacció de les respostes	PM/CV
Desenvolupament del codi	PM/CV
Participació al vídeo	PM/CV