Análisis Exploratorio de Datos y gráficos depurados

Archibald Emmanuel Carrion Claeys

A. Primera parte

A.1. Información general

Empezamos realizando un lectura de los datos. En este caso, el dataset es un archivo CSV que contiene información sobre Pokémon.

```
df <- (read.csv(file.choose(), header = TRUE, encoding = "UTF-8"))
attach(df)

# Resumen informativo de los datos - tendencias
summary(df)
```

Podemos conseguir información general sobre el dataset usando str() y glimpse().

```
# Información básica
str(df)
# glimpse() es una función del paquete dplyr que proporciona una
# vista rápida de los datos
library(dplyr)
glimpse(df)

# adicionalmente tambien existe summary() que nos da un resumen de las
# variables, como  cuartiles  y  datos  máximos  y mínimos
summary(df)
```

No se agregaron las salidas de los 2 chunks anteriores, ya que son muy extensas, y pueden facilmente ser consultadas en el archivo csv adjunto. Algunos de los datos mas valiosos que se agregara al reporte son los siguientes - attack - defense - hp - weight_kg - height_m

```
summary(df$attack)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   55.00   75.00   77.86  100.00  185.00
```

```
summary(df$defense)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   50.00   70.00   73.01   90.00  230.00
```

```r
summary(df$hp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   50.00   65.00   68.96   80.00  255.00
```

```r
summary(df$weight_kg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.10    9.00   27.30   61.38   64.80  999.90      20
```

```r
summary(df$height_m)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.100   0.600   1.000   1.164   1.500  14.500      20
```

Las variables categoricas se pueden obtener usando la función table() o count(). En R, una variable categórica es aquella que puede tomar un número limitado de valores distintos, representando categorías o grupos.

```r
# Variables categoricas
table(df$type1)
```

```
##
##      bug     dark   dragon electric    fairy fighting     fire   flying
##       72       29       27       39       18       28       52        3
##    ghost    grass   ground      ice   normal   poison  psychic     rock
##       27       78       32       23      105       32       53       45
##    steel    water
##       24      114
```

```r
table(df$type2)
```

```
##
##               bug     dark   dragon electric    fairy fighting     fire
##      384        5       21       17        9       29       25       13
##   flying    ghost    grass   ground      ice   normal   poison  psychic
##       95       14       20       34       15        4       34       29
##     rock    steel    water
##       14       22       17
```

```r
count(df, type1)
```

```
##       type1   n
## 1       bug  72
## 2      dark  29
## 3    dragon  27
## 4  electric  39
## 5     fairy  18
## 6  fighting  28
## 7      fire  52
```

```
## 8    flying   3
## 9     ghost  27
## 10    grass  78
## 11   ground  32
## 12      ice  23
## 13   normal 105
## 14   poison  32
## 15  psychic  53
## 16     rock  45
## 17    steel  24
## 18    water 114
```

```r
count(df, type2)
```

```
##       type2   n
## 1           384
## 2       bug   5
## 3      dark  21
## 4    dragon  17
## 5  electric   9
## 6     fairy  29
## 7  fighting  25
## 8      fire  13
## 9    flying  95
## 10    ghost  14
## 11    grass  20
## 12   ground  34
## 13      ice  15
## 14   normal   4
## 15   poison  34
## 16  psychic  29
## 17     rock  14
## 18    steel  22
## 19    water  17
```

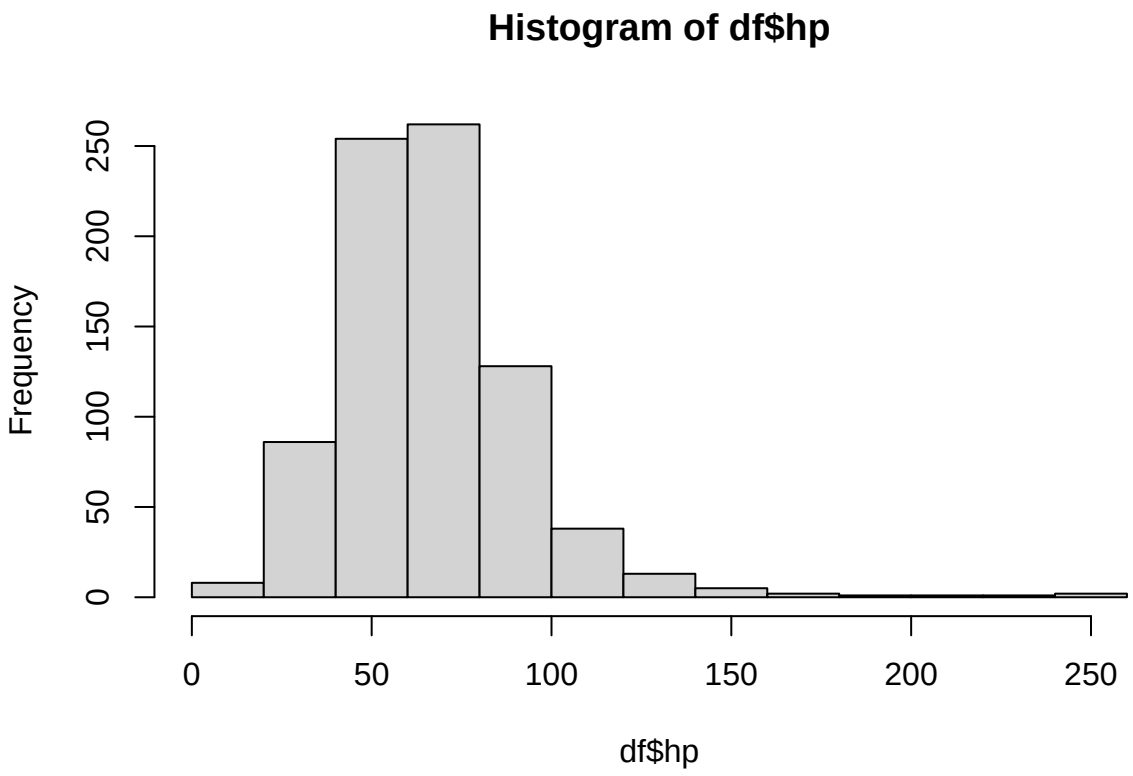
A.2. Histogramas

Un histograma es una representación gráfica de la distribución de un conjunto de datos que muestra la frecuencia de los valores en intervalos o "bins". Se usa para analizar la distribución de una variable continua, como la altura o el peso de los Pokémon.

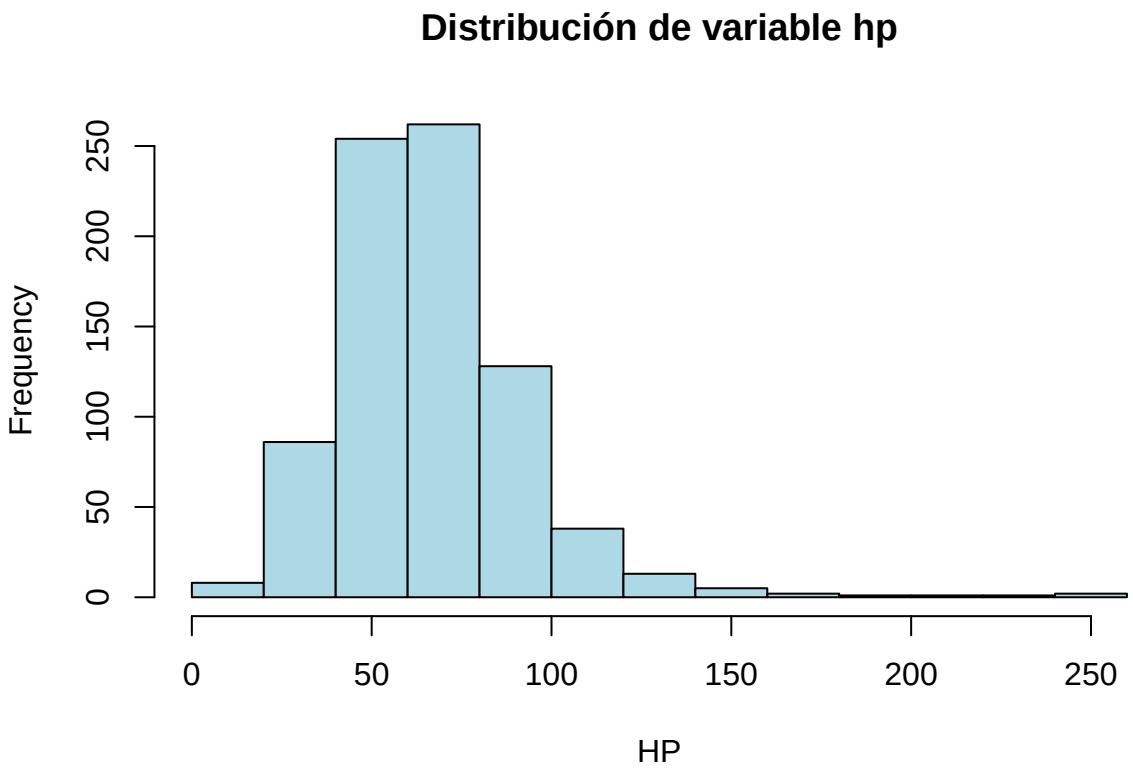
Histograma de la variable hp

```r
# Por ejemplo, el siguiente código construye un histograma para la variable hp

hist(df$hp)
```

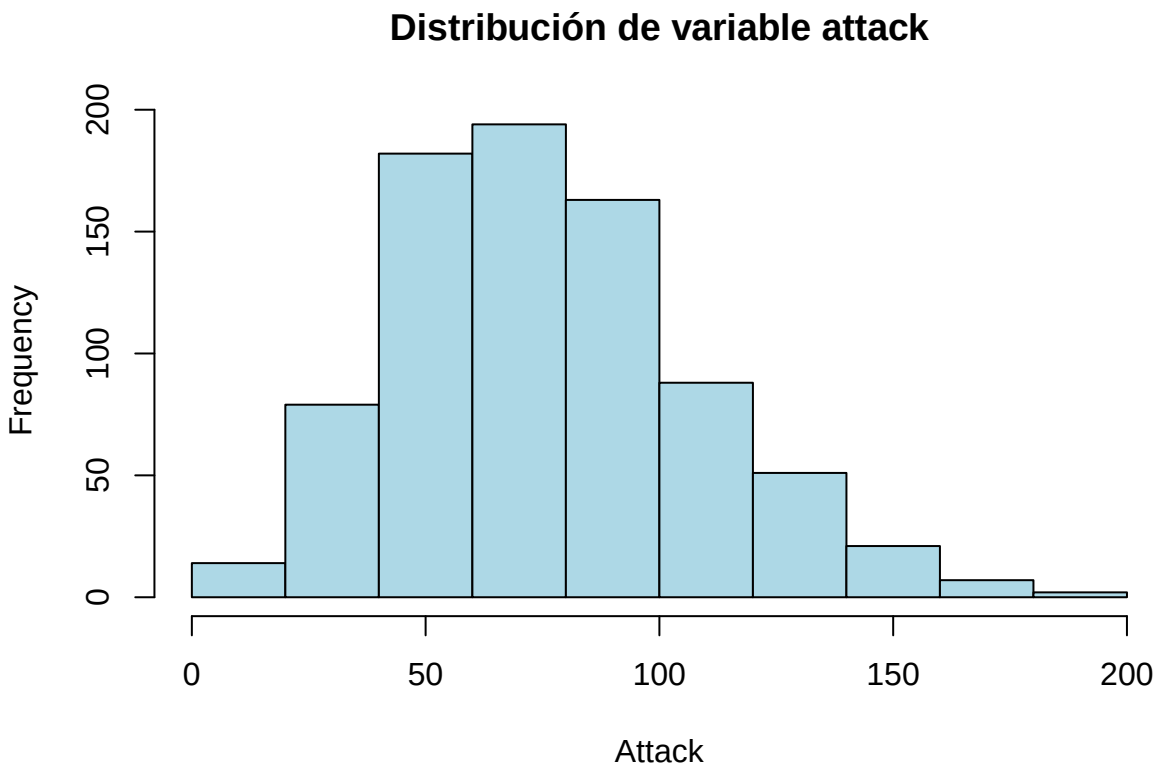



```r
# Se puede hacer un poco más claro agregando título y etiquetas:
hist(df$hp,
     main = "Distribución de variable hp",
     xlab = "HP",
     col = "lightblue",
     border = "black")
```



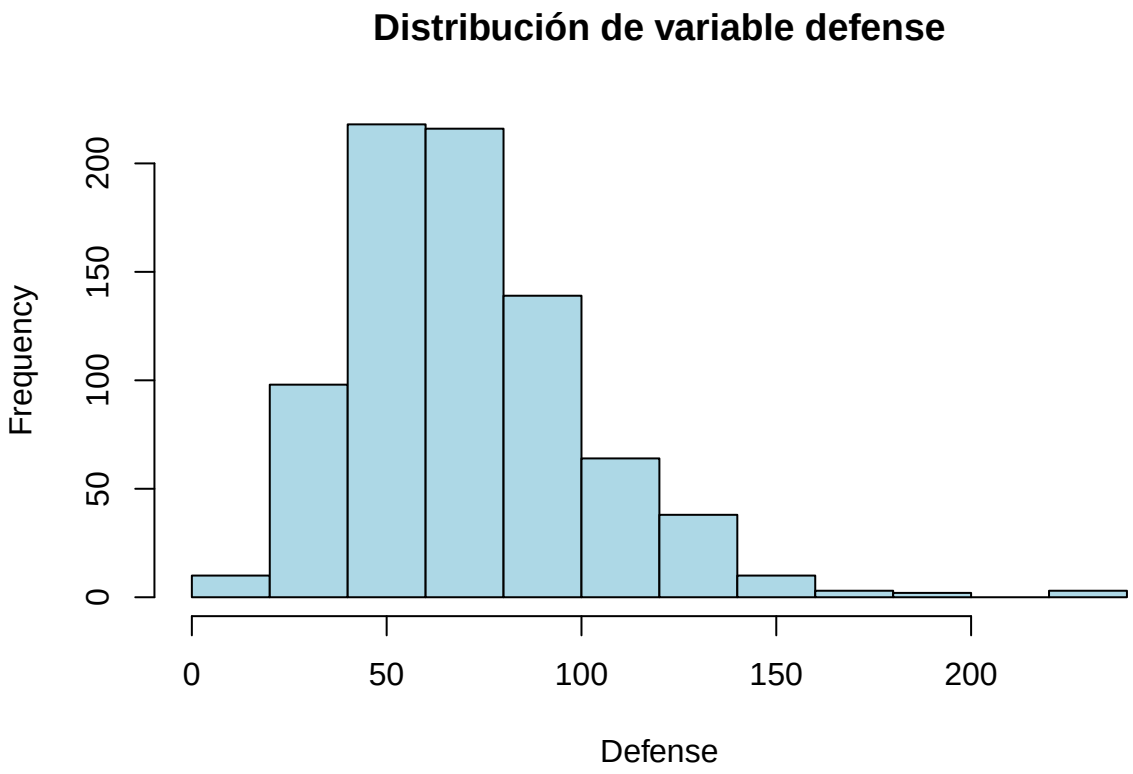
Histograma de la variable attack

```
# Histograma de la variable attack
hist(df$attack,
     main = "Distribución de variable attack",
     xlab = "Attack",
     col = "lightblue",
     border = "black")
```



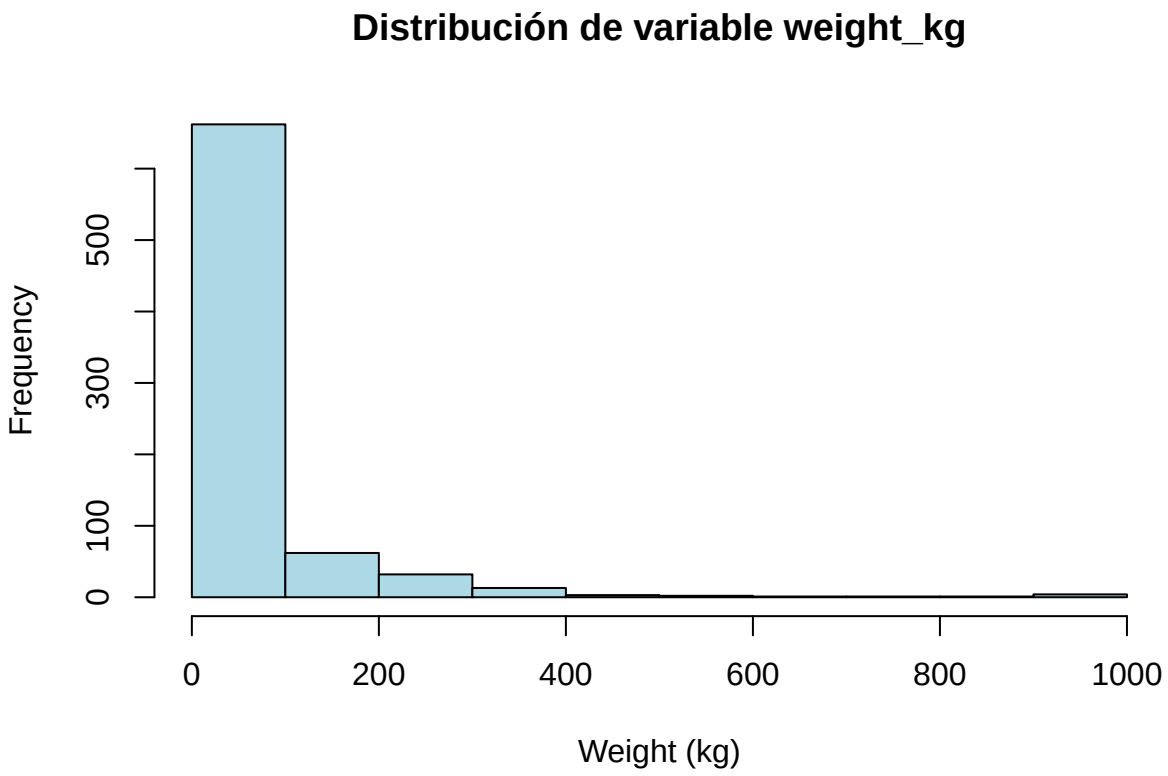
Histograma de la variable defense

```
# Histograma de la variable defense
hist(df$defense,
     main = "Distribución de variable defense",
     xlab = "Defense",
     col = "lightblue",
     border = "black")
```



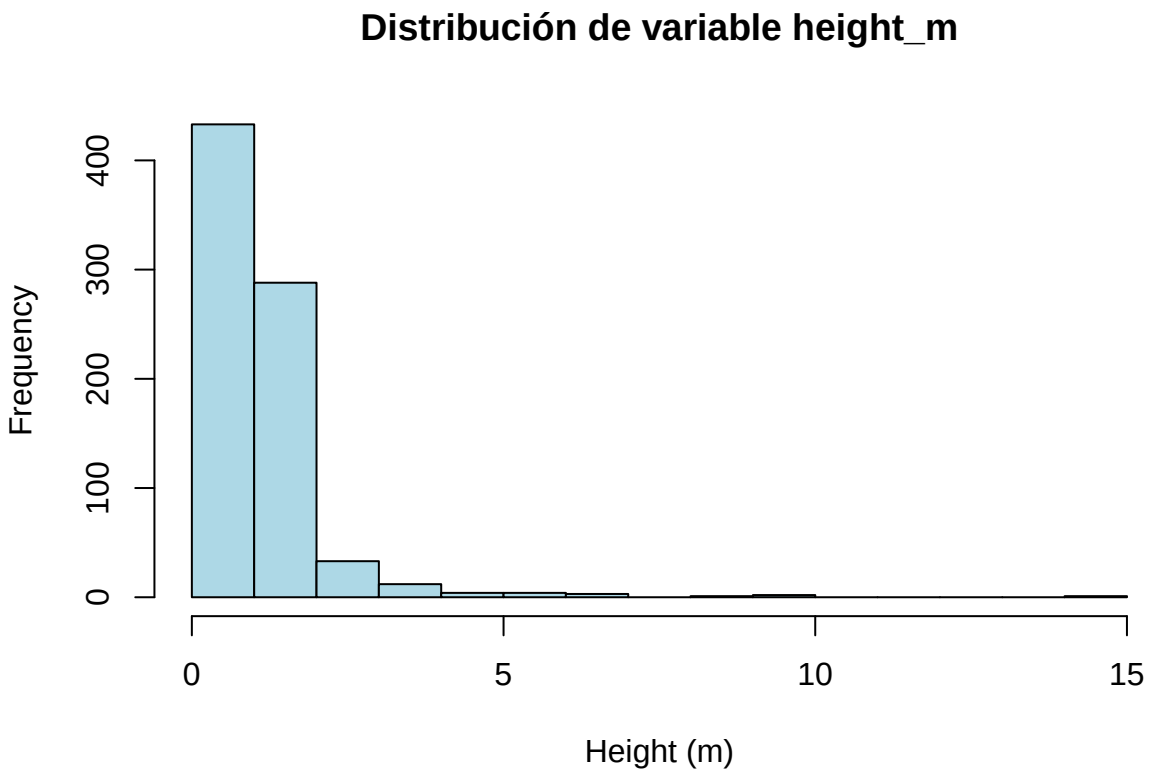
Histograma de la variable weight_kg

```
# Histograma de la variable weight_kg
hist(df$weight_kg,
     main = "Distribución de variable weight_kg",
     xlab = "Weight (kg)",
     col = "lightblue",
     border = "black")
```



Histograma de la variable height_m

```
# Histograma de la variable height_m
hist(df$height_m,
     main = "Distribución de variable height_m",
     xlab = "Height (m)",
     col = "lightblue",
     border = "black")
```


Distribución de variable height_m
Frequency
0
100
200
300
400
0
5
10
15
Height (m)

A.3. Boxplots