Datos Faltantes

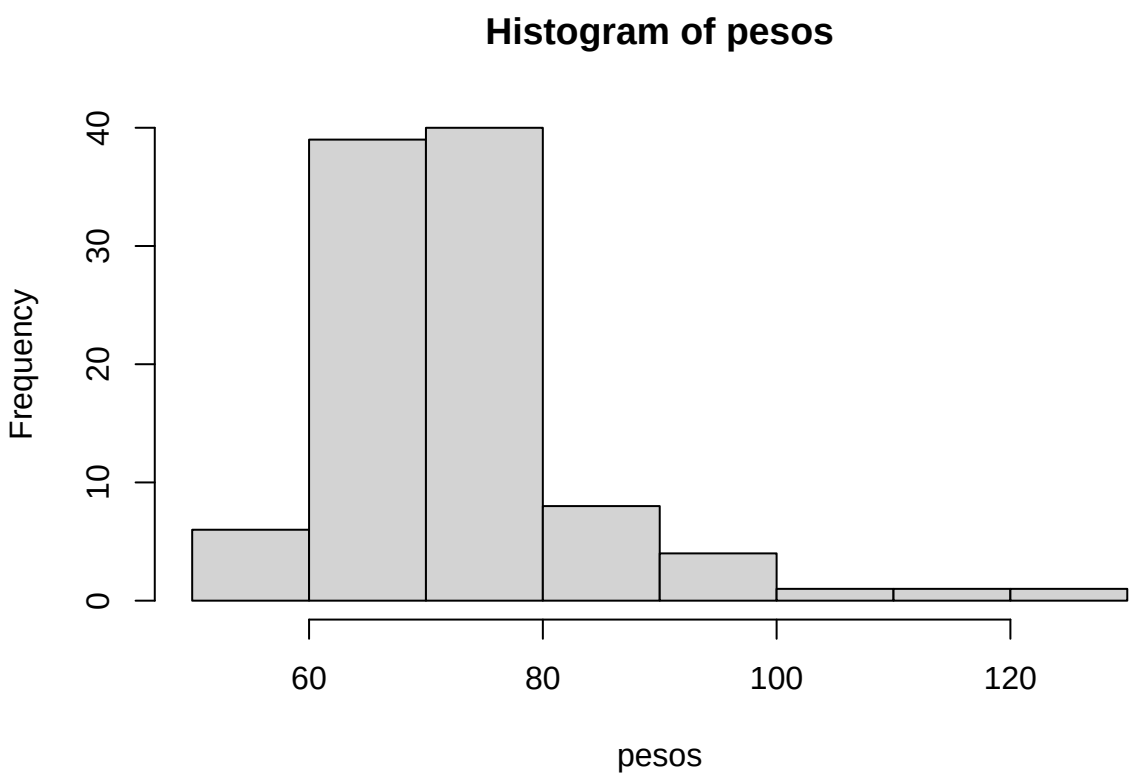
Paul Arevalo, Esteban Vizhñay

2024-07-06

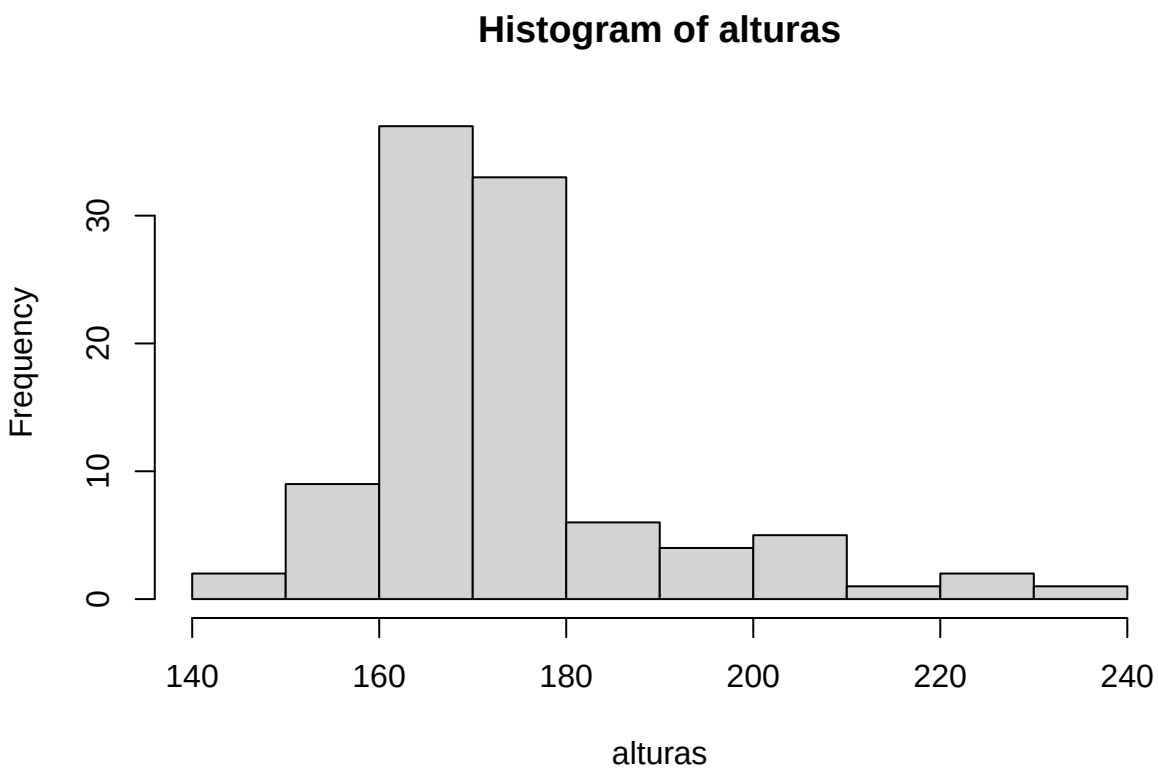
```r
# Generar distribución normal con media 50 y desviación estándar 10
pesos <- rnorm(100, mean = 68.5, sd = 7.5)
alturas <- rnorm(100, mean = 167.9, sd = 6.56)
# Introducir valores atípicos
for (i in 1:20) {
  index <- sample.int(length(pesos), 1)
  pesos[index] <- pesos[index] + runif(n = 1, min = 0, max = 50)
}
for (i in 1:20) {
  index <- sample.int(length(alturas), 1)
  alturas[index] <- alturas[index] + runif(n = 1, min = 0, max = 50)
}
```

Análisis univariable

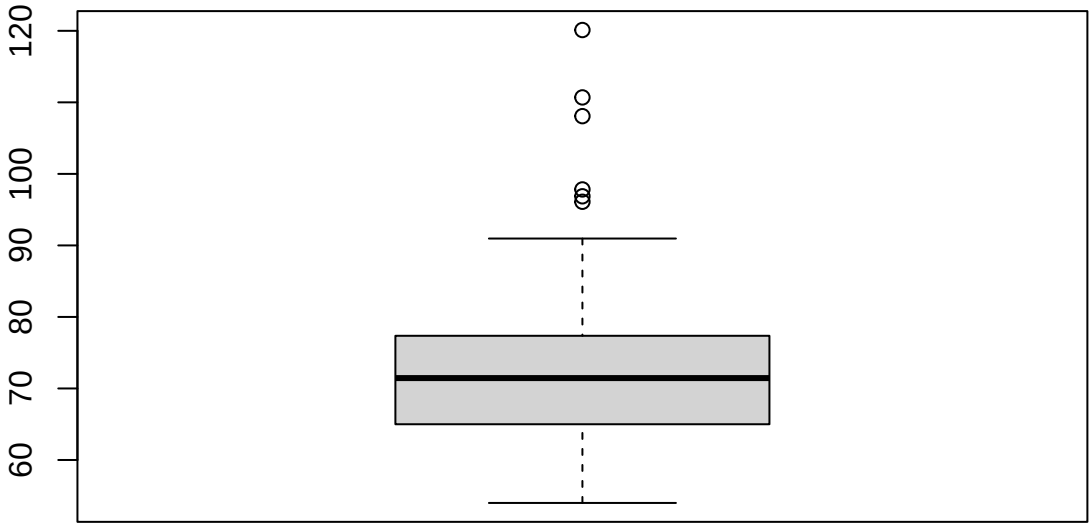
```r
hist(pesos)
```



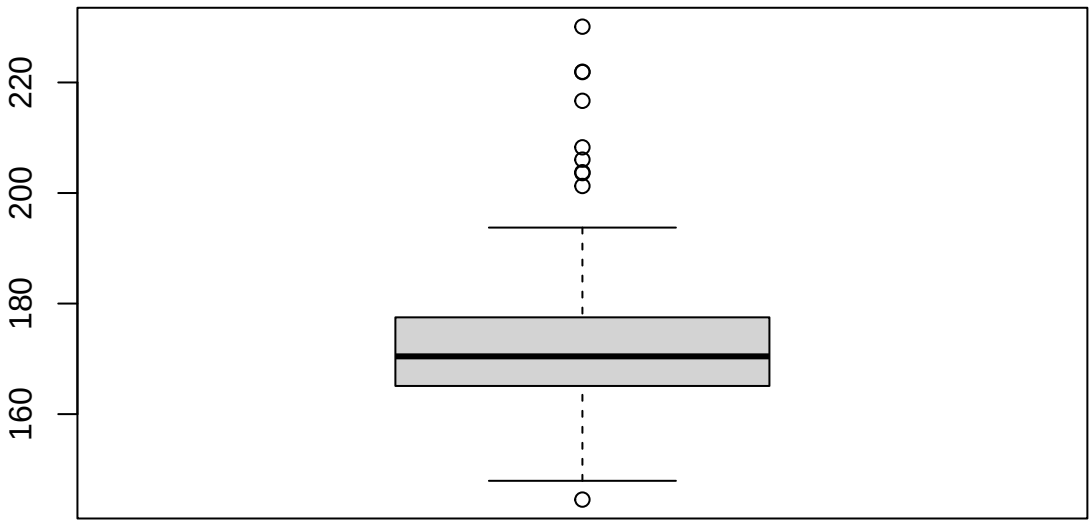
```r
hist(alturas)
```




```r
boxplot(pesos)
```




```r
boxplot(alturas)
```




```r
# Utilizar el método IQR (Rango intercuartílico)
iqr <- IQR(pesos)
limite_superior_p <- median(pesos) + 1.5 * iqr
limite_inferior_p <- median(pesos) - 1.5 * iqr
outliers_p <- pesos[pesos > limite_superior_p | pesos < limite_inferior_p]
```

```r
print(paste("[", limite_inferior_p, ",", limite_superior_p, "]"))
```

```
## [1] "[ 53.0186191393451 , 89.8823500839134 ]"
```

```r
outliers_p
```

```
## [1]  97.82401 120.11167 110.69214  96.86538 108.07195  90.96093  96.10296
```

```r
# Utilizar el método IQR (Rango intercuartílico)
iqr <- IQR(alturas)
limite_superior_a <- median(alturas) + 1.5 * iqr
limite_inferior_a <- median(alturas) - 1.5 * iqr
outliers_a <- alturas[alturas > limite_superior_a | alturas < limite_inferior_a]
```

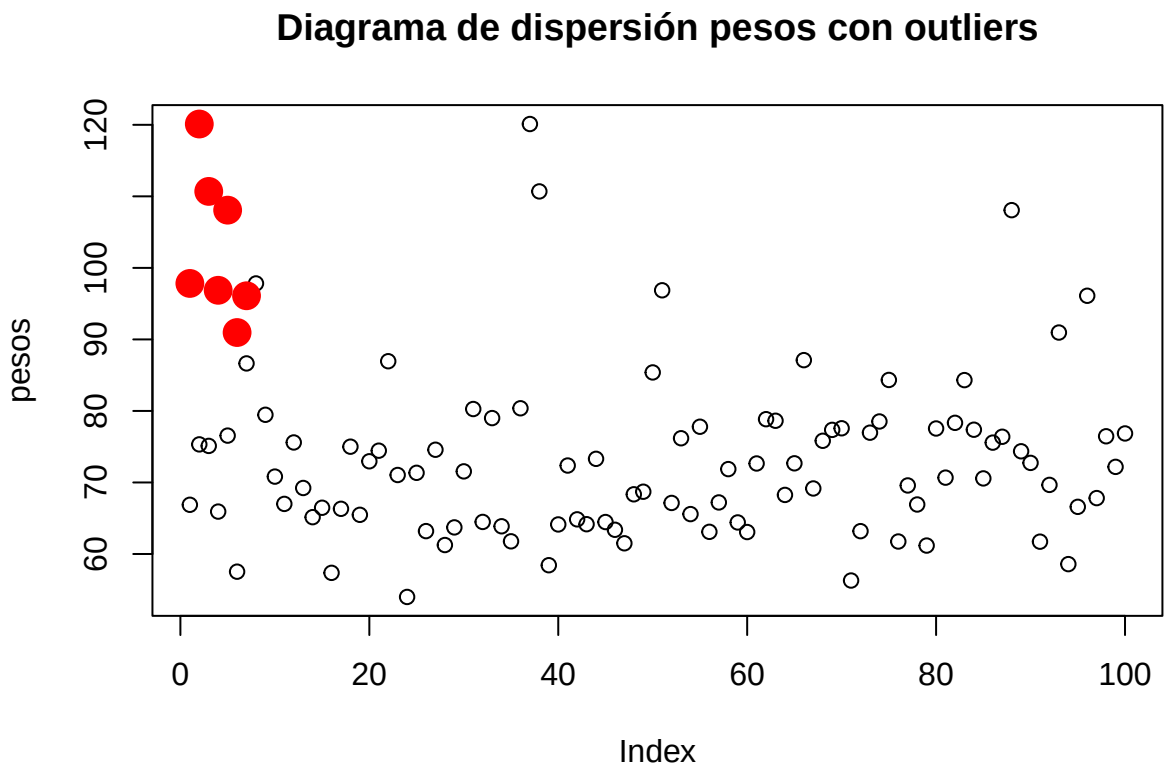
```r
print(paste("[", limite_inferior_a, ",", limite_superior_a, "]"))
```

```
## [1] "[ 151.907487489019 , 188.996232757087 ]"
```

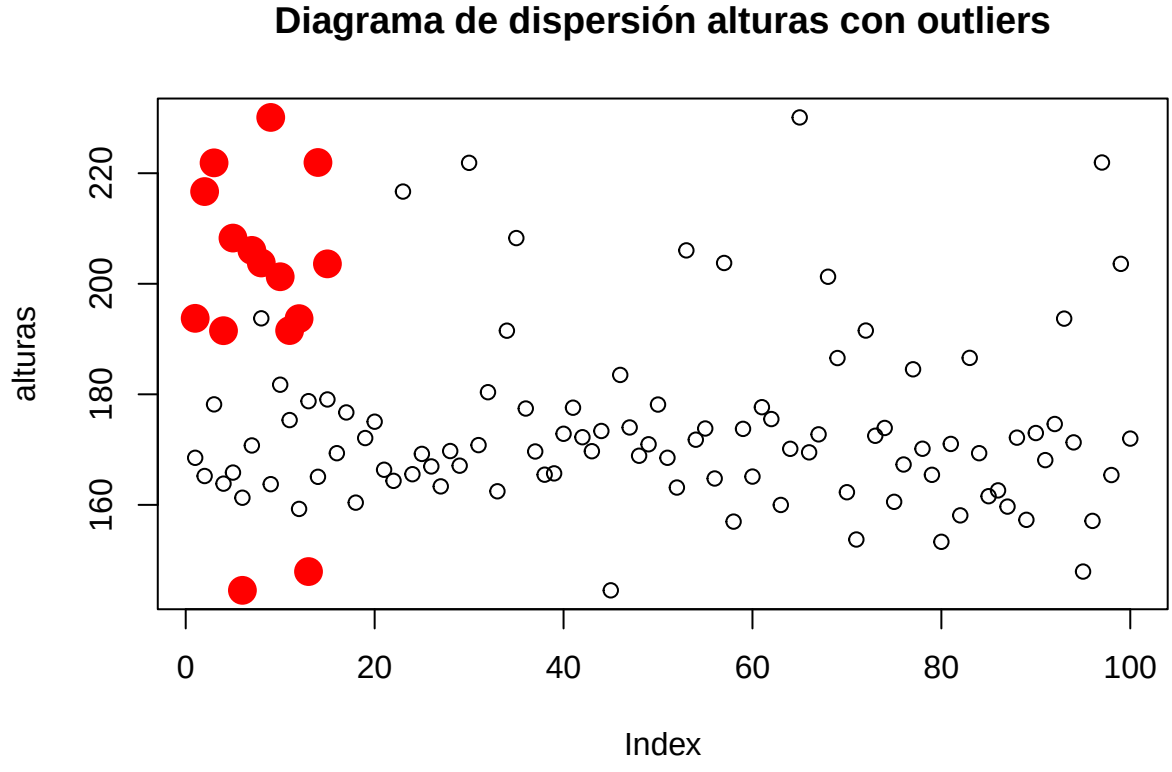
```r
outliers_a
```

```
##  [1] 193.7447 216.6826 221.8775 191.5125 208.2655 144.5408 206.0407 203.7597
##  [9] 230.0905 201.2710 191.5370 193.6932 147.9343 221.9387 203.6035
```

```r
plot(pesos, main = "Diagrama de dispersión pesos con outliers")
points(outliers_p, col = "red", pch = 16,cex = 2)
```




```
plot(alturas, main = "Diagrama de dispersión alturas con outliers")
points(outliers_a, col = "red", pch = 16,cex = 2)
```



Análisis Bivariable

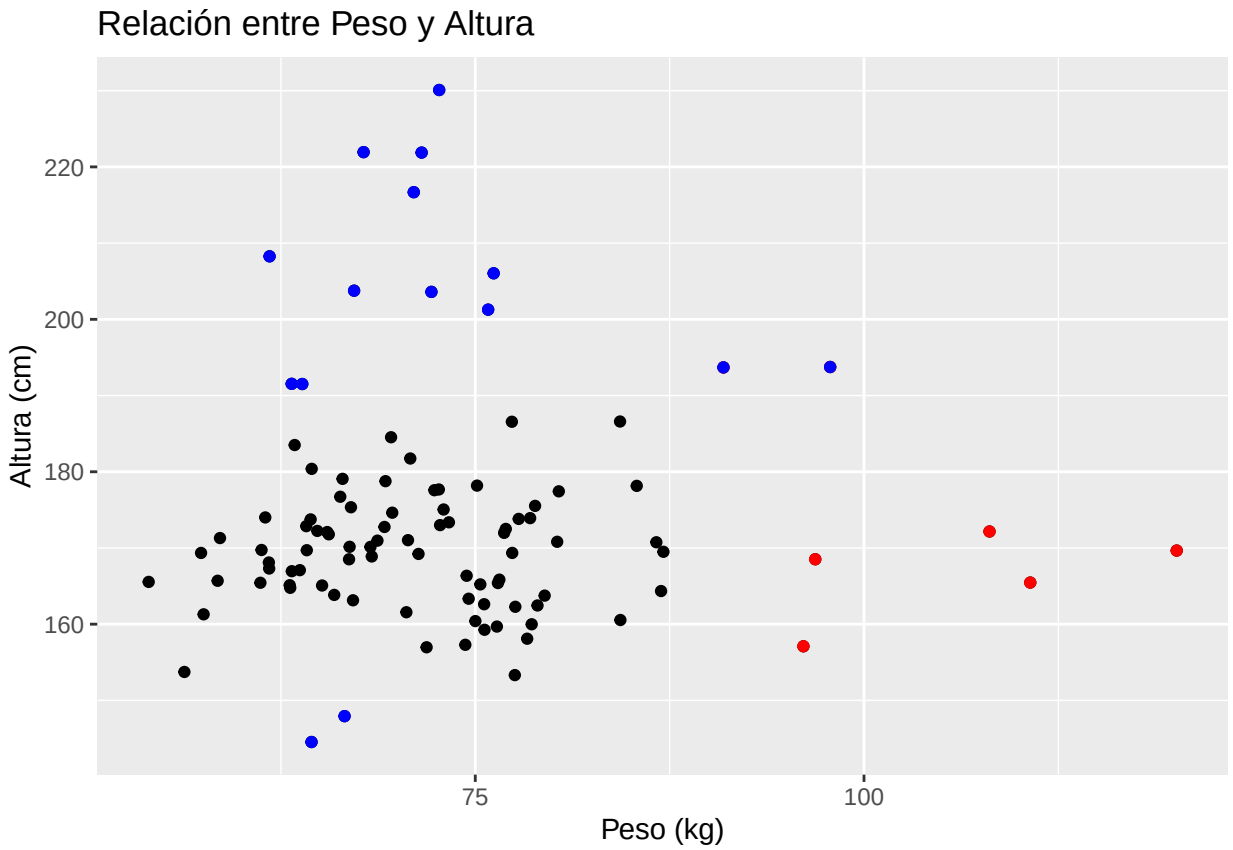
```
conjunto = cbind(pesos, alturas)
head(conjunto)
```

```
##         pesos  alturas
## [1,] 66.88133 168.5152
## [2,] 75.32565 165.2213
## [3,] 75.11222 178.1819
## [4,] 65.92864 163.8406
## [5,] 76.55041 165.8440
## [6,] 57.52898 161.2955
```

```
# Importar librerías
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
# Gráfico de dispersión
ggplot(conjunto, aes(x = pesos, y = alturas)) +
  geom_point() +
  geom_point(data = subset(conjunto, pesos %in% outliers_p),
             aes(x = pesos, y = alturas), color = "red") +  # Pesos
   geom_point(data = subset(conjunto, alturas %in% outliers_a),
             aes(x = pesos, y = alturas), color = "blue") +  # Pesos
labs(title = "Relación entre Peso y Altura",
x = "Peso (kg)", y = "Altura (cm)")
```


Relación entre Peso y Altura
Altura (cm)
220
200
180
160
75
100
Peso (kg)