# 01 Worksheet. Crash Course in Statistics (Summer 2025)

Neuroscience Center Zurich, University of Zurich

Zofia Baranczuk

2025-08-25

## 1. GDP.

Read the data set GDP. We will focus on year 2023. Which country has the highest and which one has the lowest GDP in 2023? What are GDPs of these countries? What is the GDP of Switzerland? Plot the histogram and the boxplot of GDP. How many columns and how many rows does this data set have? (Extra:) What is GDP for all the countries starting with letter "S"?

```
library(readr)
library(here)
```

```
## here() starts at C:/Users/zosia/OneDrive - Hochschule Luzern/Documents/math_uzh/HS25_Neuro_R
```

```
GDP <- read_csv(here("Data", "GDP.csv"), show_col_types = FALSE)
ind_max <- which.max(GDP$`2023`) #index of the (first, if many)
#maximal element in the column
GDP$`Country Name`[ind_max] #country name corresponding to the index above
```

```
## [1] "Monaco"
```

```
GDP$`2023`[ind_max] # GDP value corresponding to the index above
```

```
## [1] 256580.5
```

```
# - index -- the row number for which GDP 2023 was the highest

#if issues with NAs:
#1. na.omit on a smaller data set.
# Careful: if you have more columns than the one you are interested in, you can loose more data
#2. m <- max(GDP$`2023`, na.rm = TRUE)
#

ind_min <- which.min(GDP$`2023`)
GDP$`Country Name`[ind_min]
```

```
## [1] "Burundi"
```

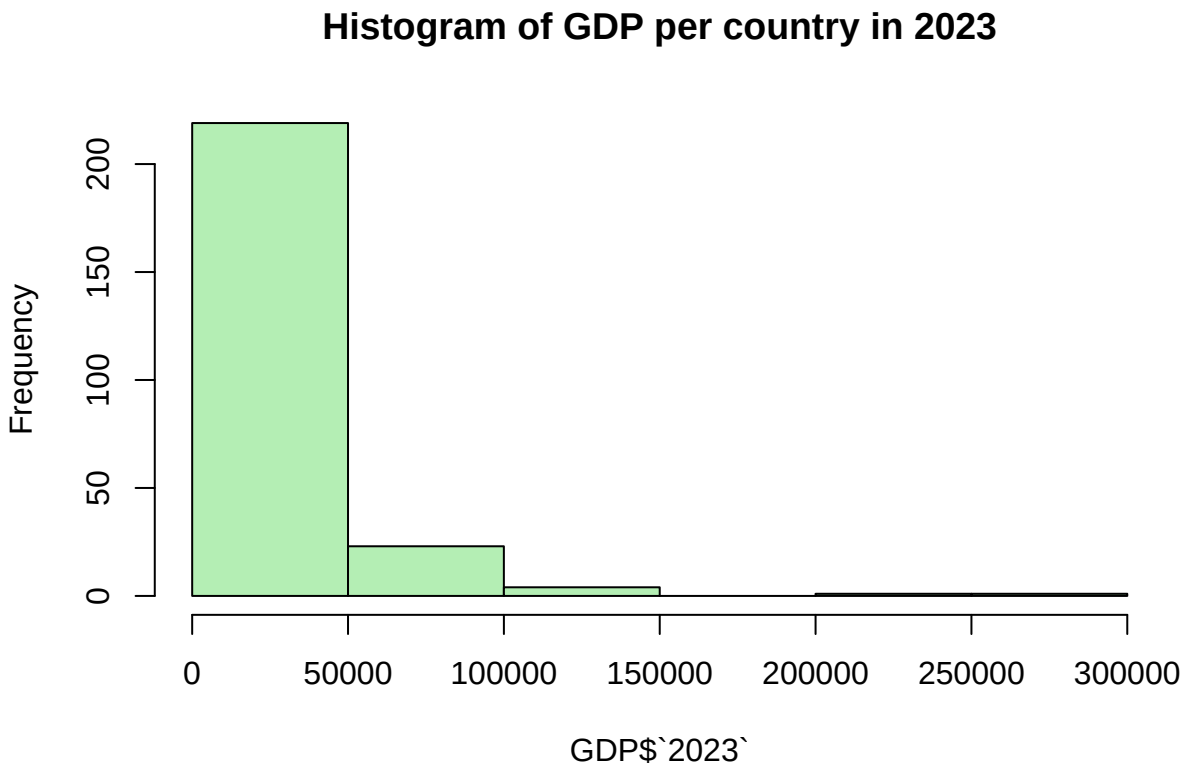
```r
GDP$`2023`[ind_min]
```

```
## [1] 192.0743
```

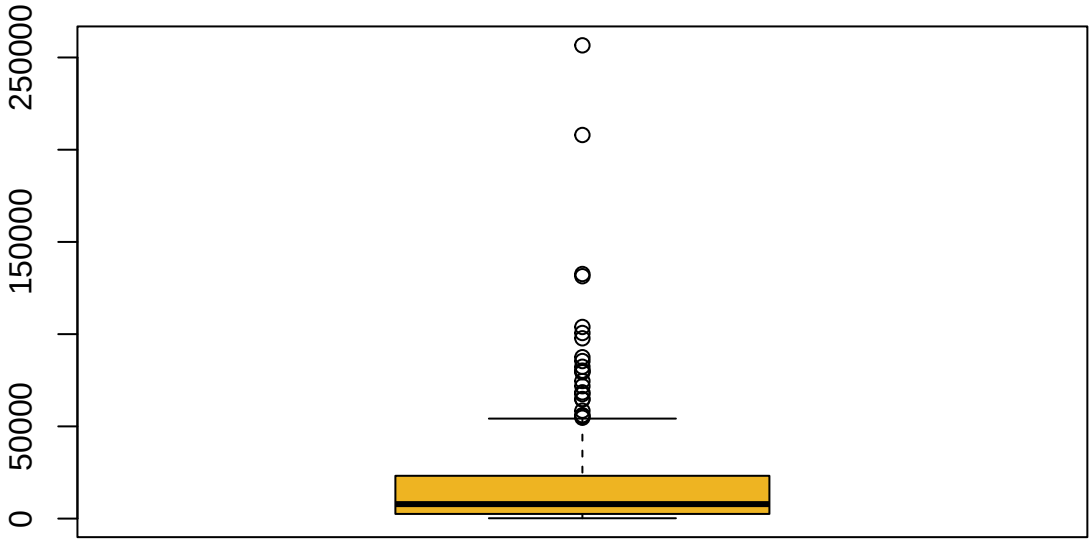
```r
GDP$`2023`[GDP$`Country Name`== "Switzerland"]
```

```
## [1] 100631.8
```

```r
hist(GDP$`2023`, col = "darkseagreen2", main= "Histogram of GDP per country in 2023")
```




```r
boxplot(GDP$`2023`, col = "goldenrod2",main = "GDP per country, 2023")
```




```
print("nrow:")
```

```
## [1] "nrow:"
```

```
nrow(GDP)
```

```
## [1] 262
```

```
print("ncol:")
```

```
## [1] "ncol:"
```

```
ncol(GDP)
```

```
## [1] 26
```

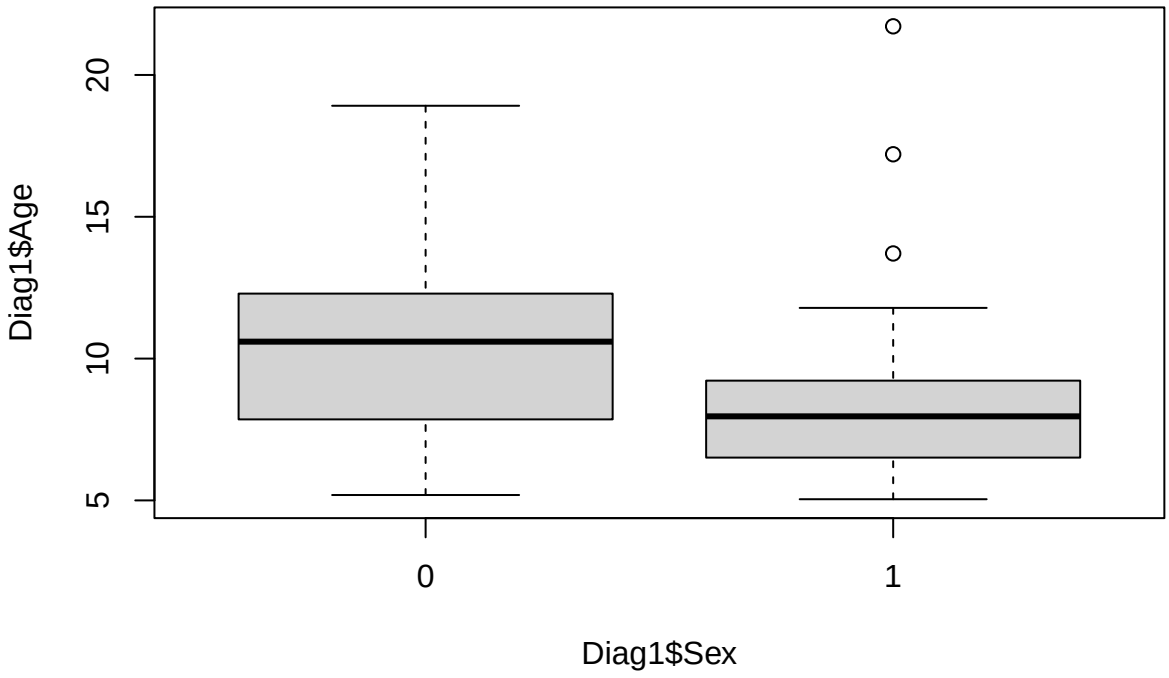
```
idx <-startsWith(GDP$`Country Name`, "S")
S_GDP<- GDP[idx, c("Country Name", "2023")]
S_GDP
```

```
## # A tibble: 32 x 2
##    `Country Name`       `2023`
##    <chr>                 <dbl>
##  1 Switzerland         100632.
##  2 Spain                33509.
##  3 St. Kitts and Nevis  22600.
##  4 St. Lucia            13555.
##  5 Sri Lanka             3799.
##  6 South Asia            2532.
##  7 Saudi Arabia         36157.
##  8 Sudan                  797.
##  9 Senegal               1698.
## 10 Singapore            85412.
## # i 22 more rows
```

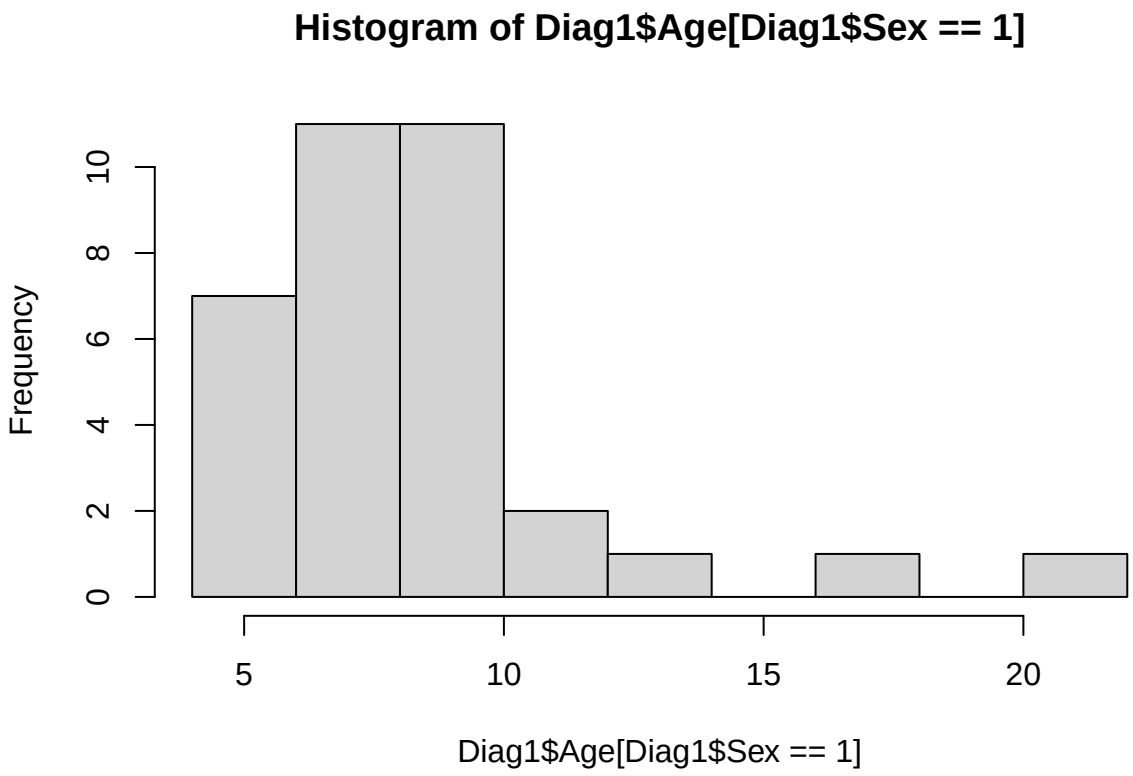
## 2. Choose one data set of interest from 02DataSets (or use your won data set).

Create 1–2 simple plots (e.g., histogram, boxplot, scatter). For each plot, add a one-sentence rationale: What do you learn from this view?

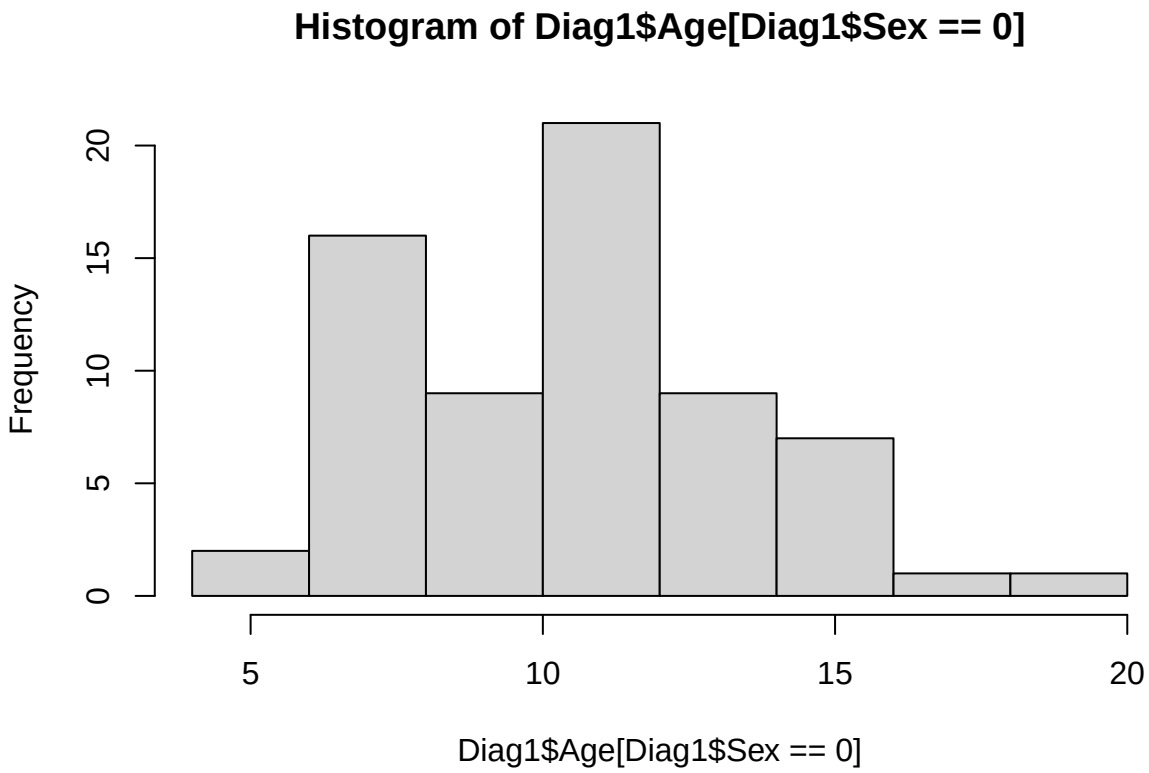
```
load(here("Data", "Diag1.RData"))
boxplot(Diag1$Age~Diag1$Sex)
```




```
hist(Diag1$Age[Diag1$Sex==1])
```




```
hist(Diag1$Age[Diag1$Sex==0])
```



```r
#Extra: or, to make the histogram a bit more useful, but quite tedious.
x0 <- Diag1$Age[Diag1$Sex == 0]
x1 <- Diag1$Age[Diag1$Sex == 1]

# Common breaks (adjust n if you want more/fewer bins)
rng <- range(c(x0, x1), na.rm = TRUE)
breaks <- pretty(rng, n = 20)

# Precompute on the same bins
h0 <- hist(x0, breaks = breaks, plot = FALSE)
h1 <- hist(x1, breaks = breaks, plot = FALSE)

# Y limit to fit both, so that we have the same range for boys and girls
ylim <- c(0, max(h0$density, h1$density, na.rm = TRUE))

# Draw
plot(h0, freq = FALSE,
     col = rgb(0.2, 0.6, 0.9, 0.4),
     xlab = "Age", ylab = "Density",
     main = "Age distribution by Sex",
     ylim = ylim)

plot(h1, freq = FALSE, add = TRUE,
     col = rgb(0.9, 0.3, 0.3, 0.4))

legend("topright",
       fill = c(rgb(0.2,0.6,0.9,0.4), rgb(0.9,0.3,0.3,0.4)),
       border = "grey30",
       legend = c("Sex = 0", "Sex = 1"))
```