Harmadik gyak

Monos Attila

2022-11-07

Leíró statisztika

A való életben szinte sose ismerjük a valódi paramétereit egy eloszlásnak – ez főleg abszolút folytonos eloszlásoknál fordulhat elő, pl. normális, exponenciális, Gamma eloszlások.

Emiatt sokszor mért adatok vizsgálatára szorítkozunk. A háttérben levő folyamatot, ami a minta elemeit adta, egy véletlen folyamatnak fogjuk fel, így van eloszlása – ám ezt nem ismerjük. Ennek első eleme a leíró statisztika, mely a mintának (vagyis az adathalmaznak) az eloszlását nem próbálja megtalálni, csak leírja egyes tulajdonságait, pl.:

- Mintaátlag:
$$\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$
- Tapasztalati szórás:
$$S_n = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2}$$
- Korrigált tapasztalati szórás:
$$S_n^* = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2}$$
- Szórási együttható (százalékként is szokták írni):
$$V = \frac{S_n}{\overline{X}}$$
- Korrigált szórási együttható (százalékként is szokták írni):
$$V = \frac{S_n^*}{\overline{X}}$$
- k. tapasztalati momentun:
$$m_k = \sum_{i=1}^{n} X_i^k$$

1. Feladat

Egy szabályos dobókockával négyszer dobtunk, és a következő eredményeket kaptuk: $1, 3, 6, 1$. Számoljuk ki a mintaátlagot, a tapasztalati szórást, a korrigált tapasztalati szórást, a korrigált szórási együtthatót és a második tapasztalati momentumot!

Mintaátlag:

```
x <- c(1, 3, 6, 1);
mean(x)
```

```
## [1] 2.75
```

Tapasztalati szórásnégyzet:

```
sqrt(mean((x - mean(x))^2))
```

```
## [1] 2.046338
```

Korrigált tapasztalati szórásnégyzet:

```
sqrt(1/3 * sum((x - mean(x))^2))
```

```
## [1] 2.362908
```

```
sd(x)
```

```
## [1] 2.362908
```

Szórási együttható:

```
sd(x)/mean(x)
```

```
## [1] 0.8592392
```

```
round(sd(x)/mean(x), 4)*100
```

```
## [1] 85.92
```

Második tapasztalati momentum:

```
mean(x^2)
```

```
## [1] 11.75
```

Mindez összefoglalva:

```
cat("Átlag:", mean(x),
    "\nSzórás:", sqrt(mean((x - mean(x))^2)),
    "\nKorrigált szórás:", sd(x),
    "\nSzórási együttható:", sd(x)/mean(x),
    "\nTapasztalati második momentum:", mean(x^2), '\n')
```

```
## Átlag: 2.75
## Szórás: 2.046338
## Korrigált szórás: 2.362908
## Szórási együttható: 0.8592392
## Tapasztalati második momentum: 11.75
```

Toljuk el 100-al az előző adatokat! Hogyan változik a mintaátlag és a korrigált tapasztalati szórás?

```r
x_new <- x + 100
cat("Átlag:", mean(x_new),
    "\nRégi átlag:", mean(x),
    "\nKorrigált szórás:", sd(x_new),
    "\nRégi korrigált szórás:", sd(x_new)/mean(x_new), '\n')
```

```
## Átlag: 102.75
## Régi átlag: 2.75
## Korrigált szórás: 2.362908
## Régi korrigált szórás: 0.02299667
```

Most szorozzuk meg -3-al az eredeti adatokat! Ekkor hogyan változik a mintaátlag és a korrigált tapasztalati szórás?

```r
x_new <- -3*x
cat("Átlag:", mean(x_new),
    "\nRégi átlag:", mean(x),
    "\nKorrigált szórás:", sd(x_new),
    "\nRégi korrigált szórás:", sd(x_new), '\n')
```

```
## Átlag: -8.25
## Régi átlag: 2.75
## Korrigált szórás: 7.088723
## Régi korrigált szórás: 7.088723
```

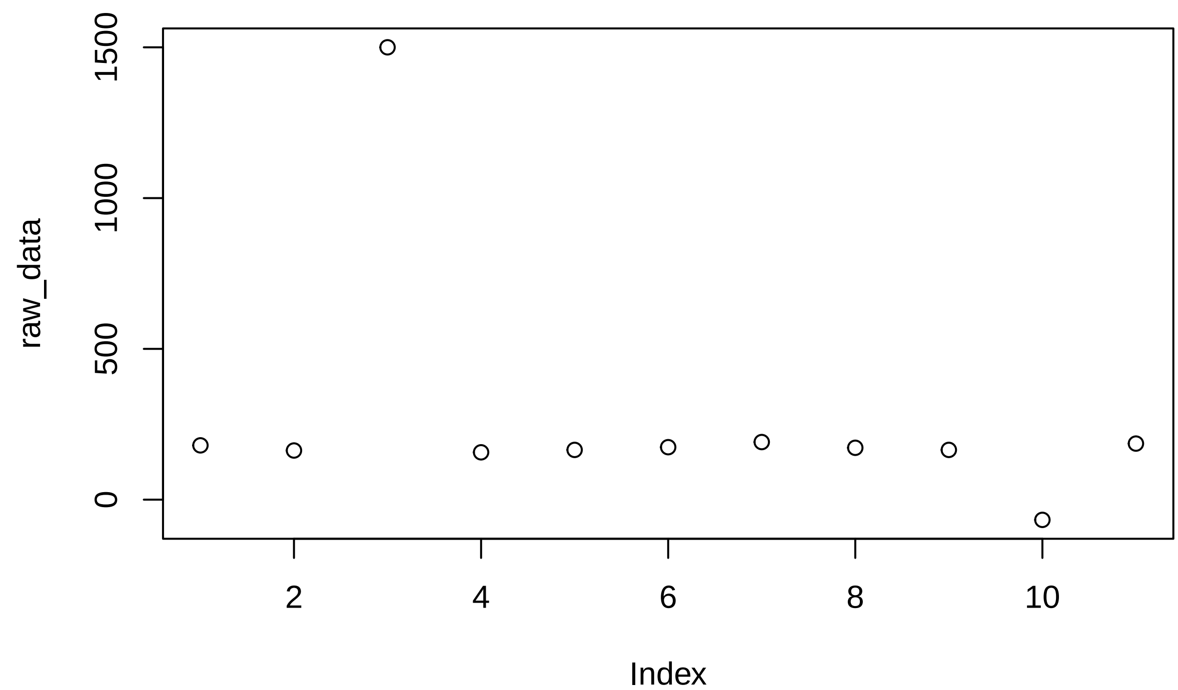
Extra Feladat Magyarázzuk meg a fenti változásokat! A többi adat változik-e?

2. Feladat Egy csoportban a hallgatók magassága cm-ben:

$$180, 163, 1500, 157, 165, 174, 191, 172, 165, 1-68, 186$$

Ezek reális adatok? Az esetleges adathibákat javítsuk! Ezt nem lehet csak úgy random tenni, attól, hogy egy adat "csúnya", még meg kell nézni, hogy tényleg hibás-e.

```r
raw_data <- c(180, 163, 1500, 157, 165, 174, 191, 172, 165, 1-68, 186)
n <- length(raw_data)
plot(raw_data)
```



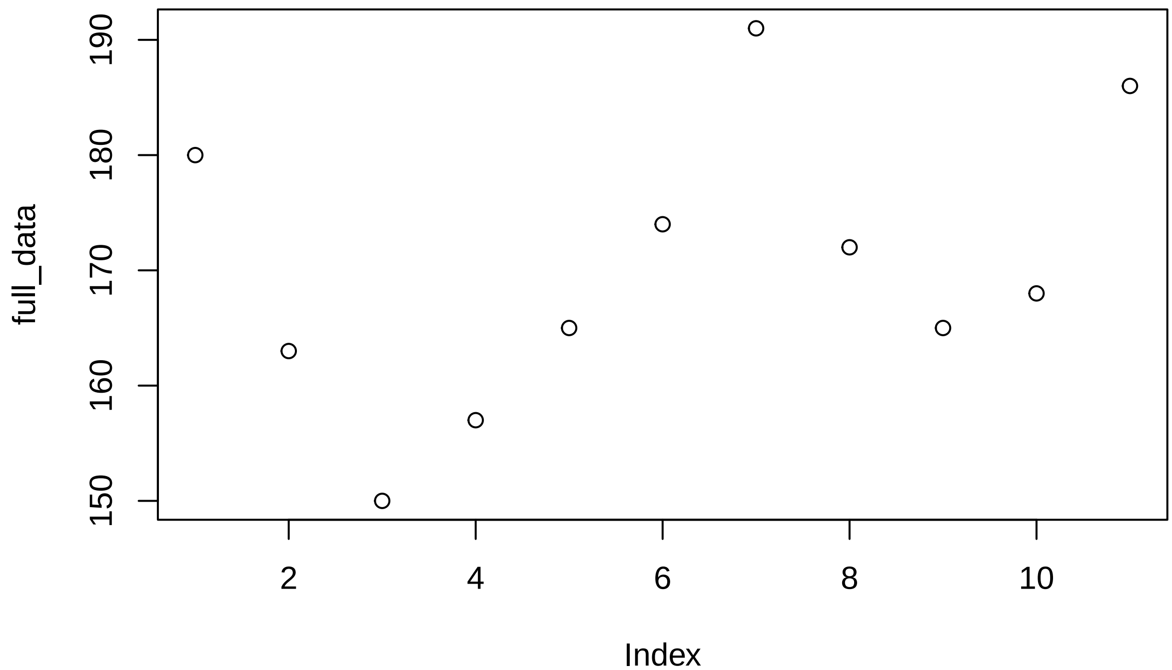

```
summary(raw_data)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -67.0   164.0   172.0   271.5   183.0  1500.0
```

```
data.entry(raw_data)
```

Látjuk, hogy két adat van, ami hibás: 1500 és $1-68$, ezeket javítjuk kézzel:

```
full_data <- raw_data
full_data[3] <- 150 #corrected from 1500, extra zero
full_data[10] <- 168 #corrected from 1-68, extra -
plot(full_data)
```




```r
summary(full_data)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   150.0   164.0   168.0   170.1   177.0   191.0
```

Most elemezzük az adatokat az alapstatisztikákkal:

- Átlag,
- korrigált tapasztalati szórás,
- szórási együttható (korrigált tapasztalati szórásból),
- kvartilisek,
- terjeledem,
- interkvartilis terjedelem!

```r
mean(full_data)
```

```
## [1] 170.0909
```

```r
sd(full_data)
```

```
## [1] 12.20209
```

```
sd(full_data)/mean(full_data)
```

```
## [1] 0.07173862
```

```
min(full_data)
```

```
## [1] 150
```

```
max(full_data)
```

```
## [1] 191
```

```
max(full_data) - min(full_data)
```

```
## [1] 41
```

```
quarts = quantile(full_data, probs = c(1/4, 1/2, 3/4), type = 6)
quarts[3] - quarts[1]
```

```
## 75%
##  17
```

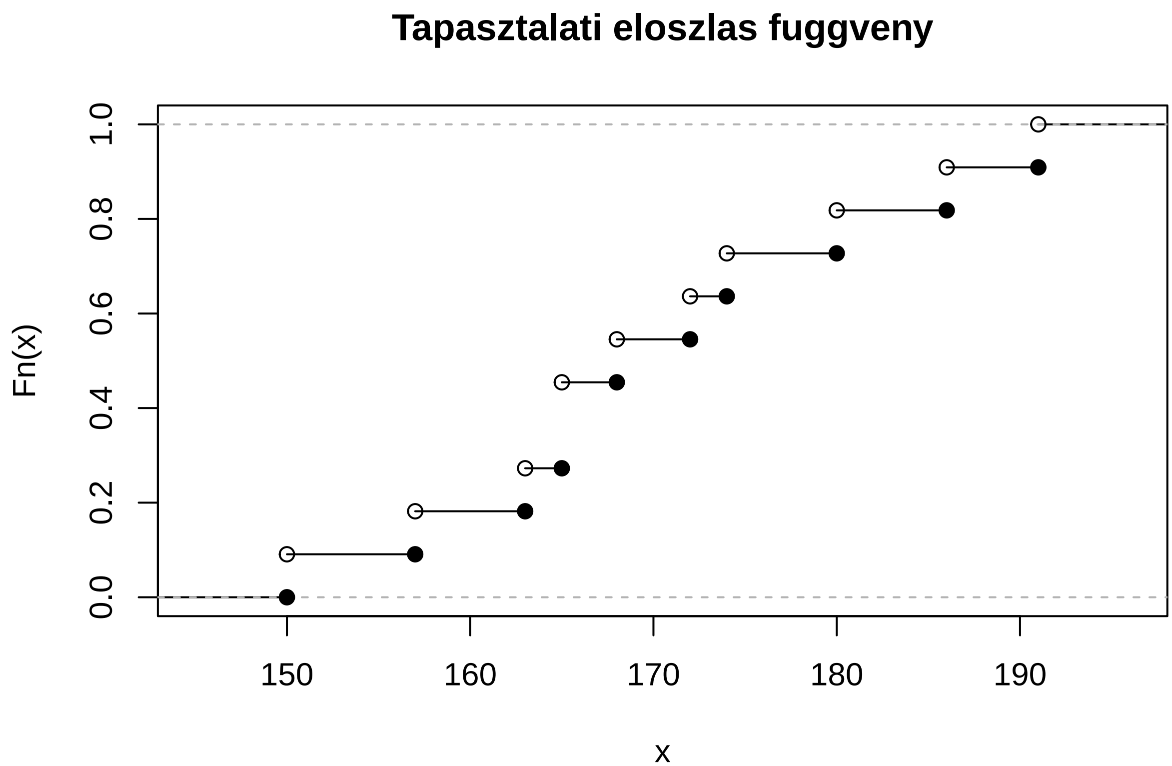
Adjuk meg a rendezett mintát!

```
x <- sort(full_data); x
```

```
##  [1] 150 157 163 165 165 168 172 174 180 186 191
```

Rajzoljuk fel a tapasztalati eloszlásfüggvényt, és olvassuk le az értékét a 180 helyen! Szövegesen mit jelent ez?

```
Fn <- ecdf(full_data)
plot(Fn, do.points = FALSE, ylab = 'Fn(x)', main = "Tapasztalati eloszlas fuggveny")
points(unique(x), unique(c(0, Fn(x)))[1:length(unique(x))], pch = 19)
points(unique(x), unique(Fn(x)), pch = 21)
```




```
Fn(180)
```

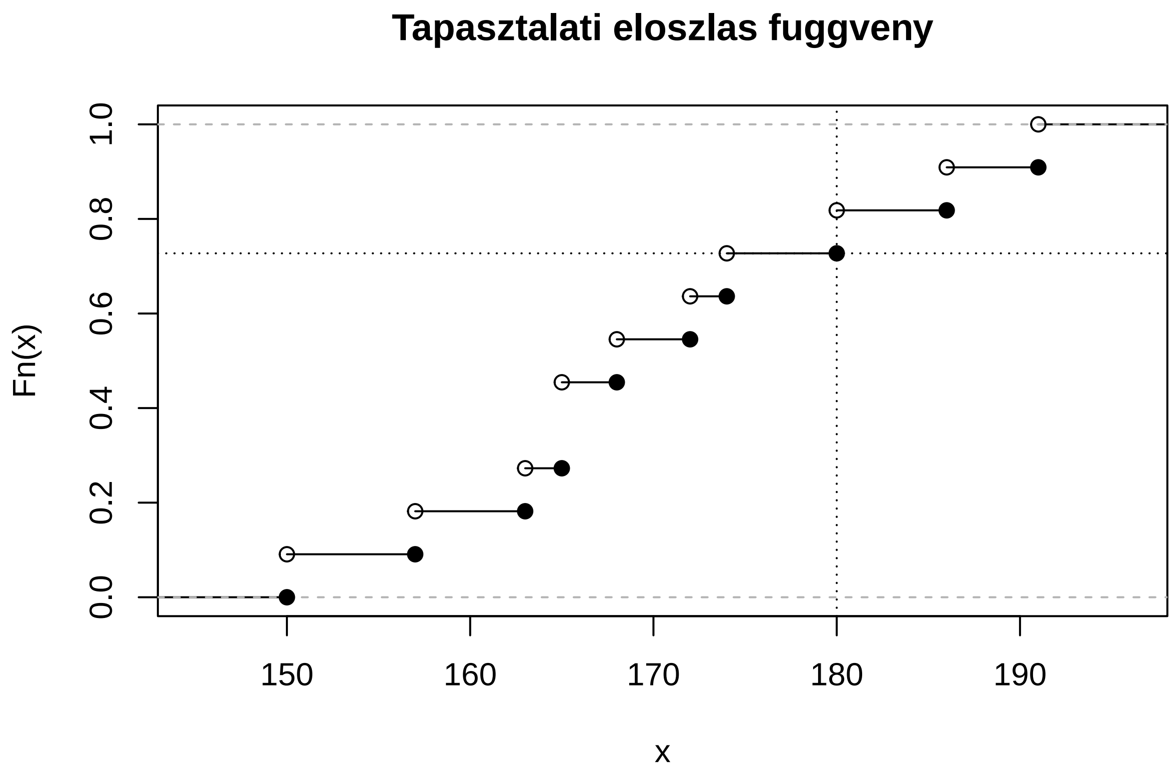
```
## [1] 0.8181818
```

A tapasztalati eloszlásfüggvény jobbról folytonos. Átalakítható úgy, hogy balról folytonos legyen:

```
Fn_lc <- function(x) {mean(full_data<x)}
Fn_lc(180)
```

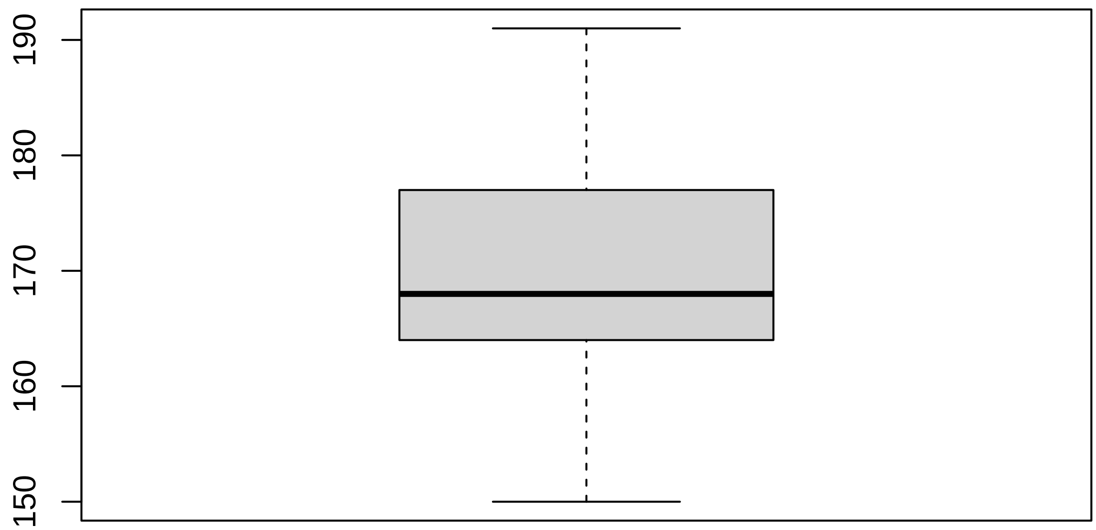
```
## [1] 0.7272727
```

```
plot(Fn, do.points = FALSE, ylab = "Fn(x)", main = "Tapasztalati eloszlas fuggveny")
points(unique(x), unique(c(0, Fn(x)))[1:length(unique(x))], pch = 19)
points(unique(x), unique(Fn(x)), pch = 21)
abline(v = 180, lty = 3)
abline(h = Fn_lc(180), lty = 3)
```

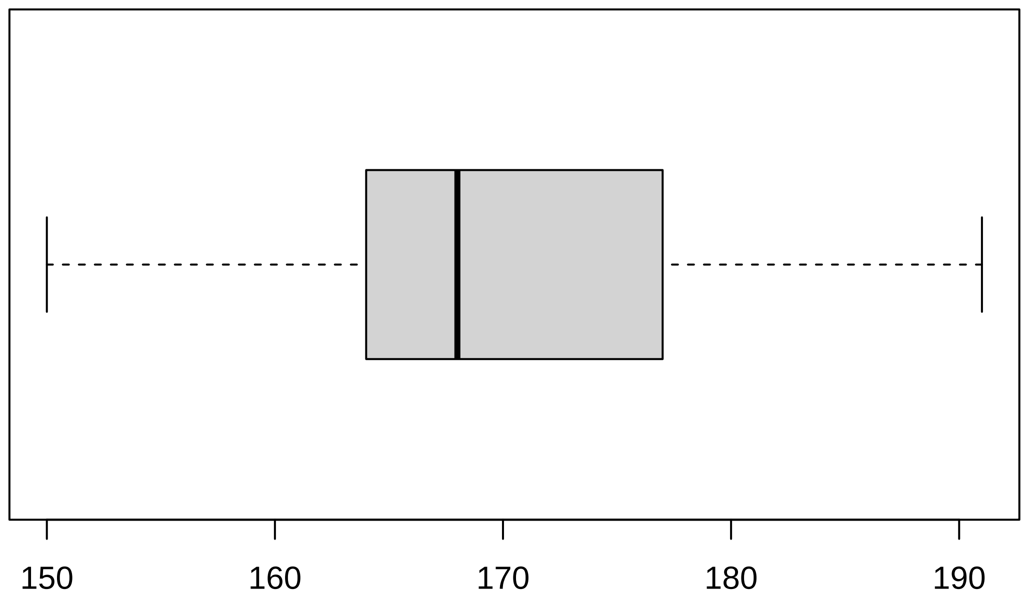


Rajzoljuk le az adatok boxplot ábráját!

```
boxplot(full_data)
```



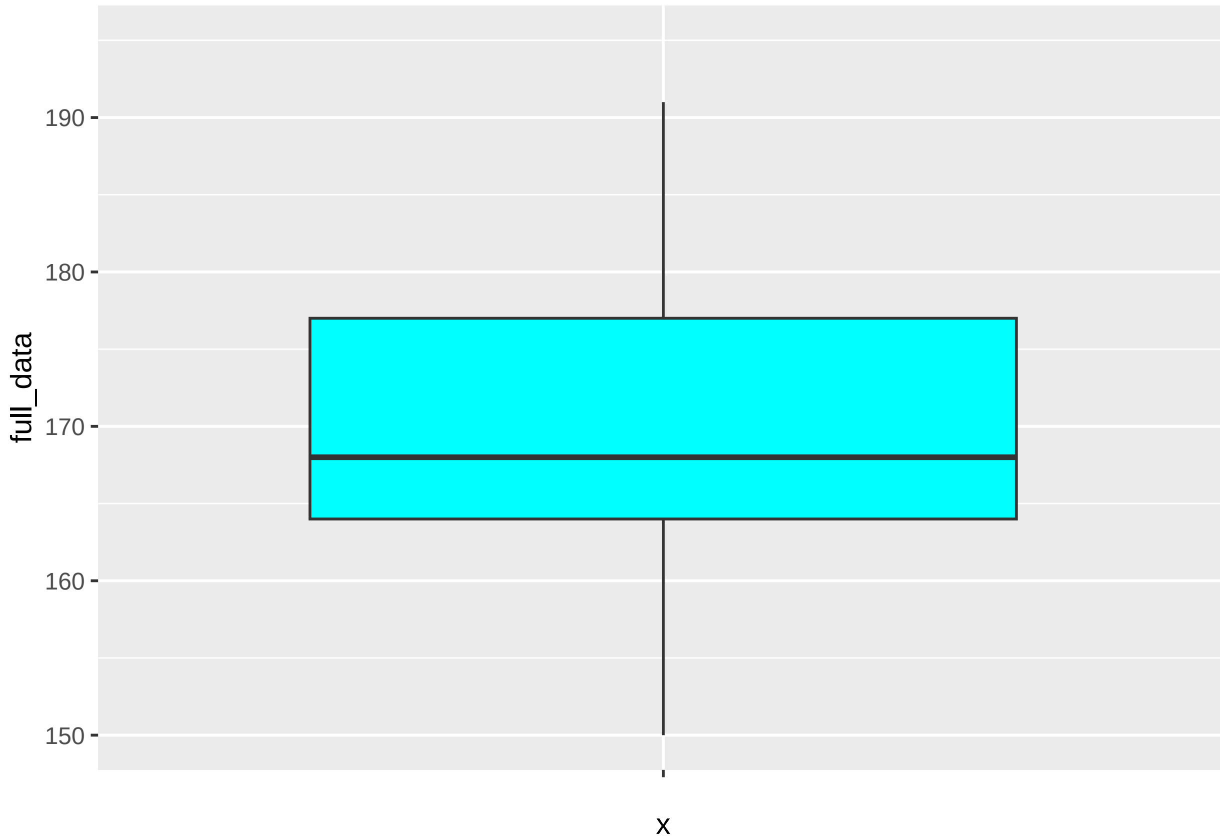

```
boxplot(full_data, horizontal = TRUE)
```




```r
#install.packages("ggplot2")
library(ggplot2) #Needs rlang 1.0.6. Close down all .rmd's, restart RStudio, then              #
```

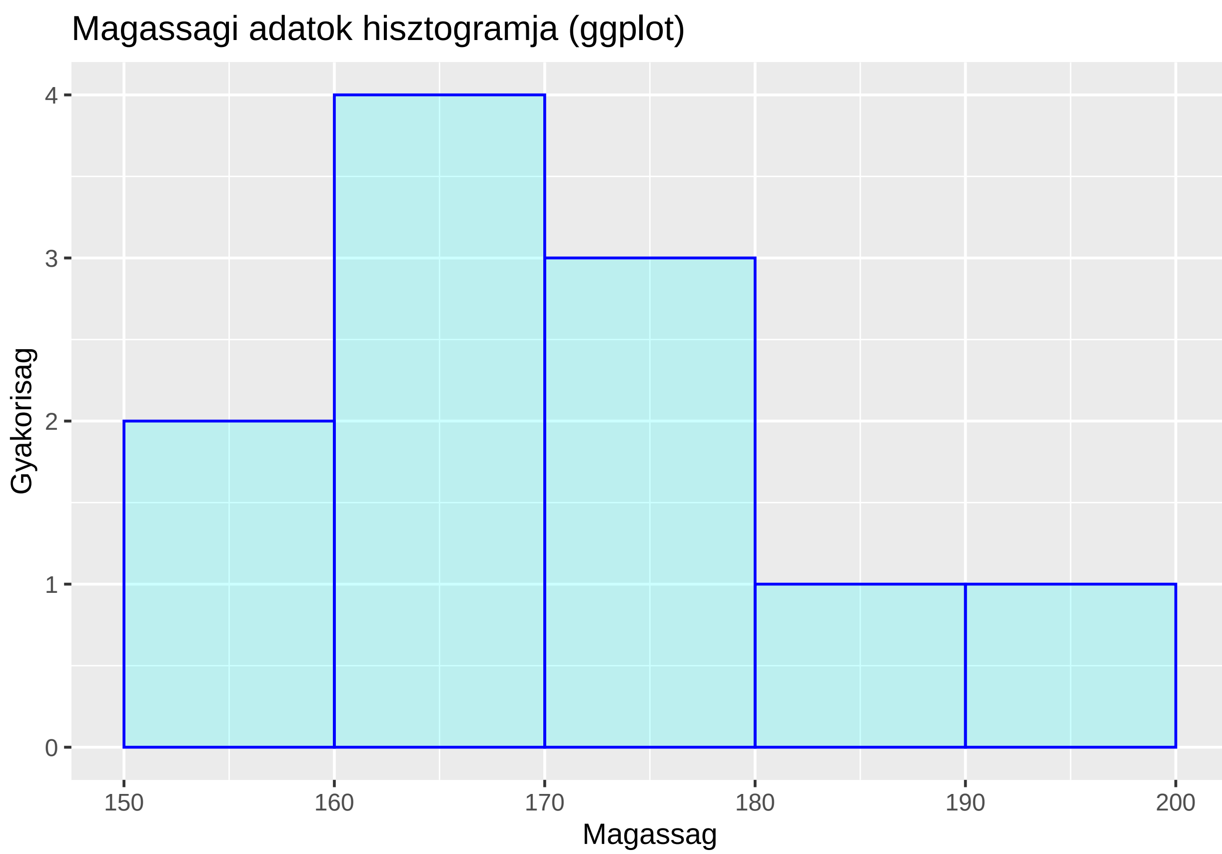
```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```r
df <- data.frame(full_data)
ggplot(data = df, aes(x = "", y = full_data)) +
  geom_boxplot(fill="cyan") + coord_cartesian(ylim = c(150, 195))
```

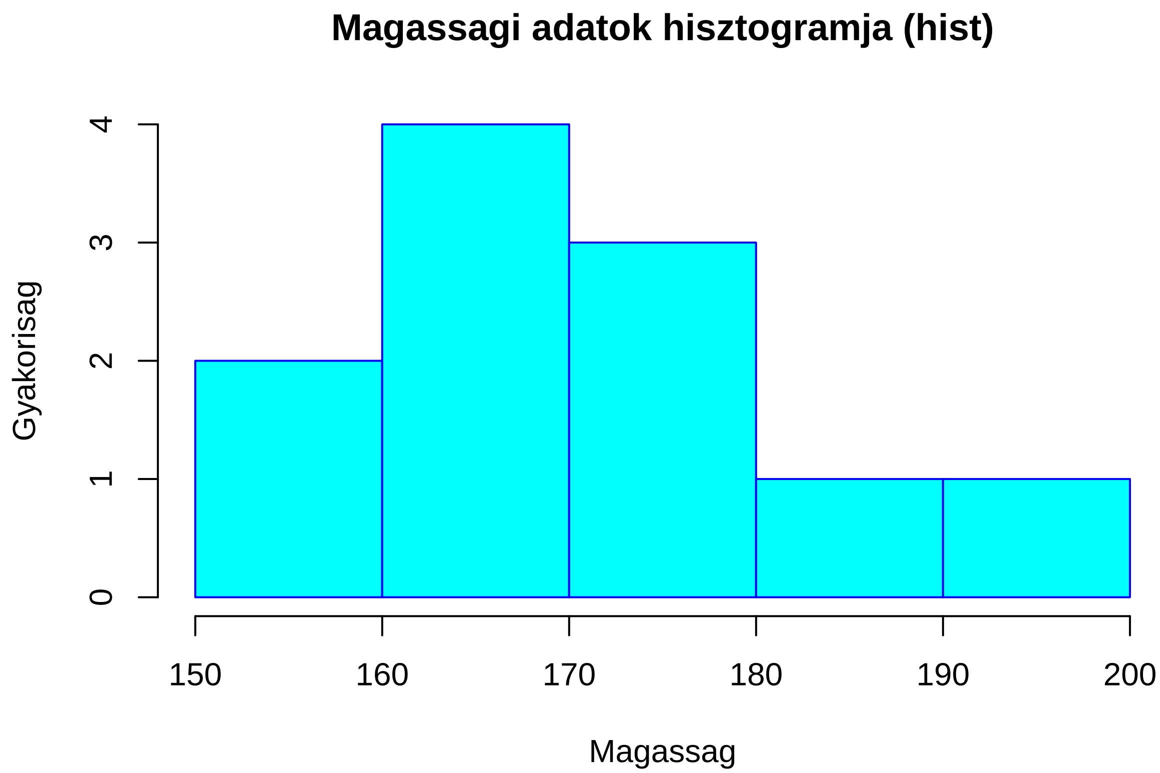


Készítsünk hisztogramot az adatokról!

```
ggplot(data = df, aes(x = full_data)) +
       geom_histogram(breaks = seq(150, 200, by = 10),
                       col="blue",
                       fill="cyan",
                       alpha = .2) +
       labs(title = "Magassagi adatok hisztogramja (ggplot)", x = "Magassag",
            y = "Gyakorisag")
```



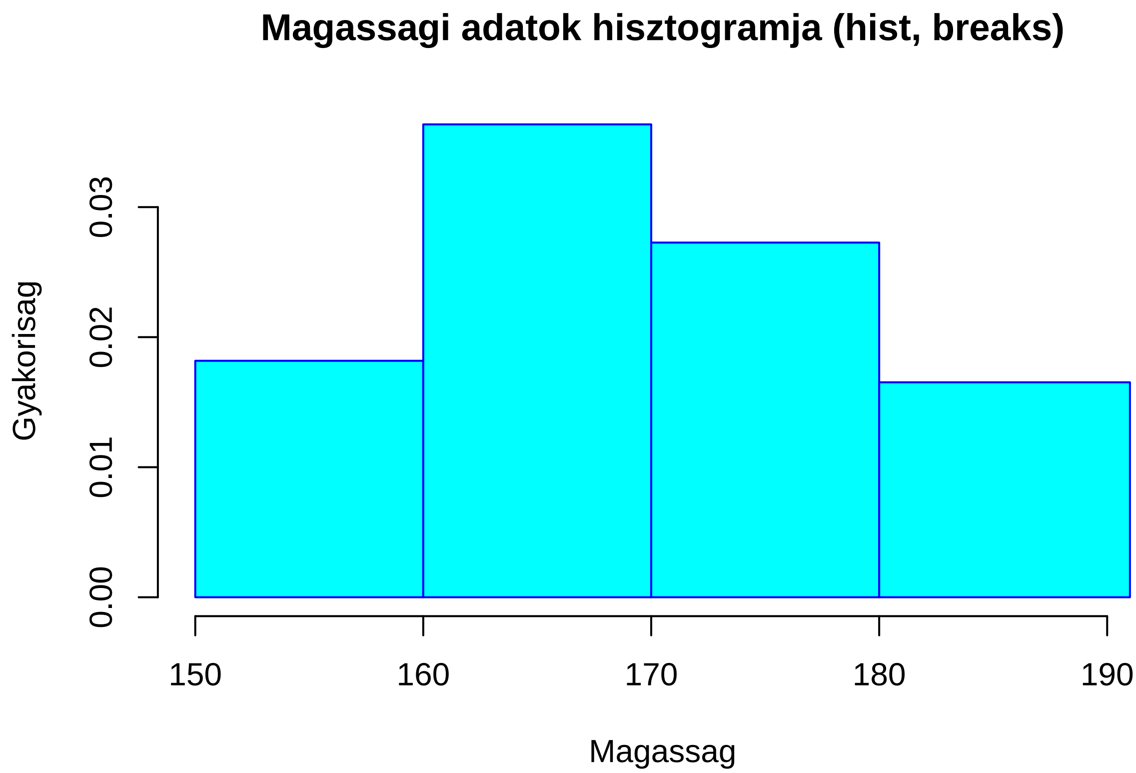

```r
histo <- hist(full_data,
              breaks = 5,
              xlab = "Magassag",
              ylab = "Gyakorisag",
              main = "Magassagi adatok hisztogramja (hist)",
              col = "cyan",
              border = "blue")
```




```
histo$counts
```

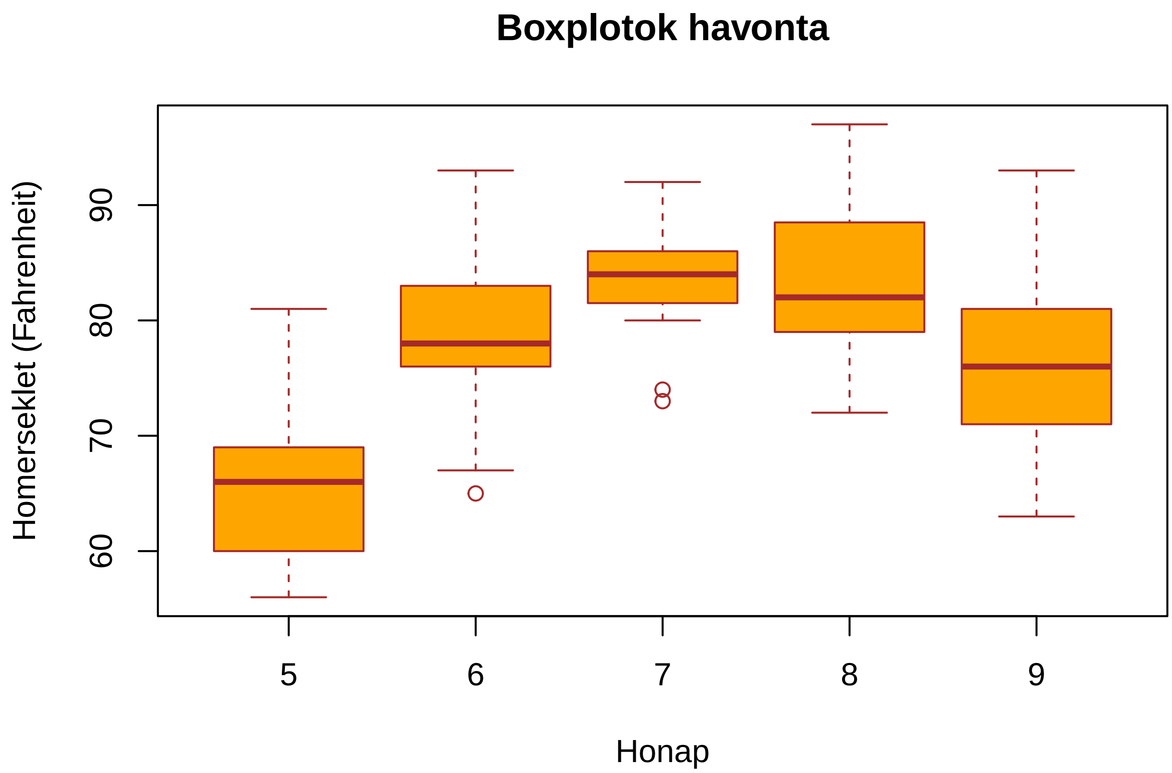
```
## [1] 2 4 3 1 1
```

```
hist(full_data, breaks = c(min(full_data), 160, 170, 180, max(full_data)),
     xlab = "Magassag",
     ylab = "Gyakorisag",
     main = "Magassagi adatok hisztogramja (hist, breaks)",
     col = "cyan",
     border = "blue")
```



Az R-nek vannak beépített adathalmazai itt.

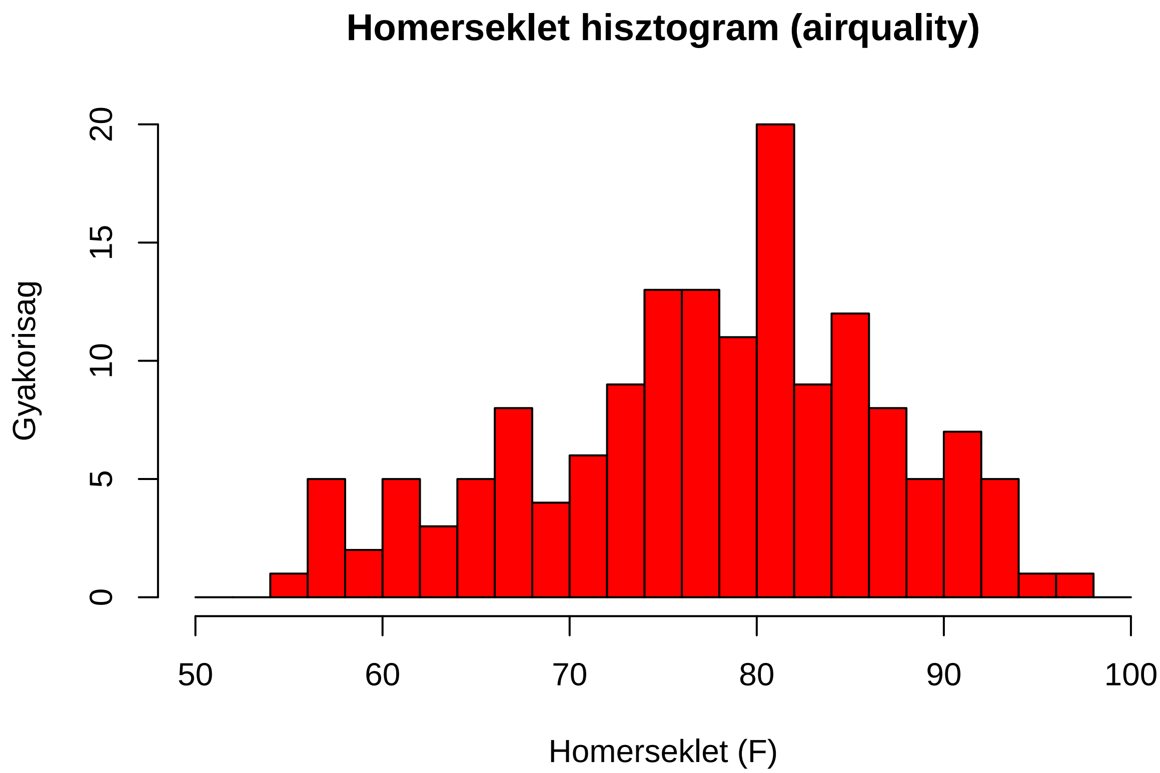
```
boxplot(Temp ~ Month,
        data = airquality,
        main = "Boxplotok havonta",
        xlab = "Honap",
        ylab = "Homerseklet (Fahrenheit)",
        col = "orange",
        border = "brown")
```




```
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

```
h <- hist(airquality$Temp,
          breaks = seq(50, 100, 2),
          #breaks = seq(50, 100, 10),
          col = "red",
          xlab = "Homerseklet (F)",
          ylab = "Gyakorisag",
          main = "Homerseklet hisztogram (airquality)")
```



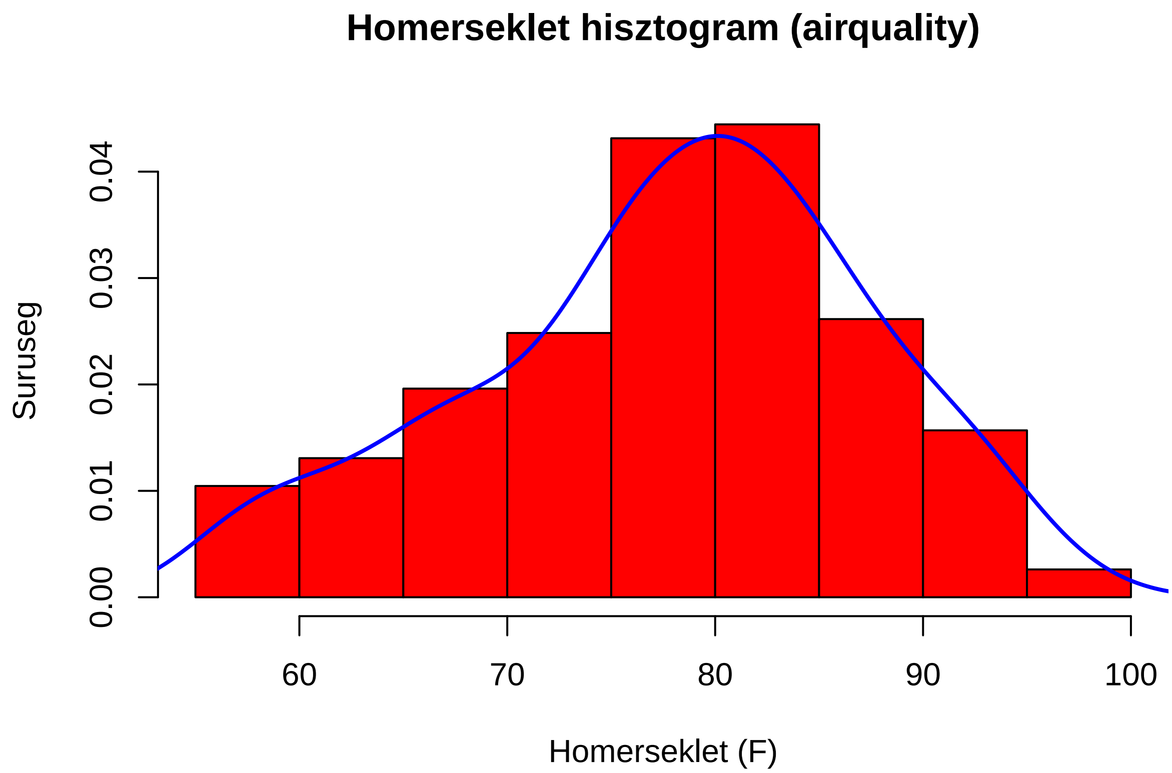

```
h$breaks
```

```
##  [1]  50  52  54  56  58  60  62  64  66  68  70  72  74  76  78  80  82  84  86
## [20]  88  90  92  94  96  98 100
```

```
h$counts
```

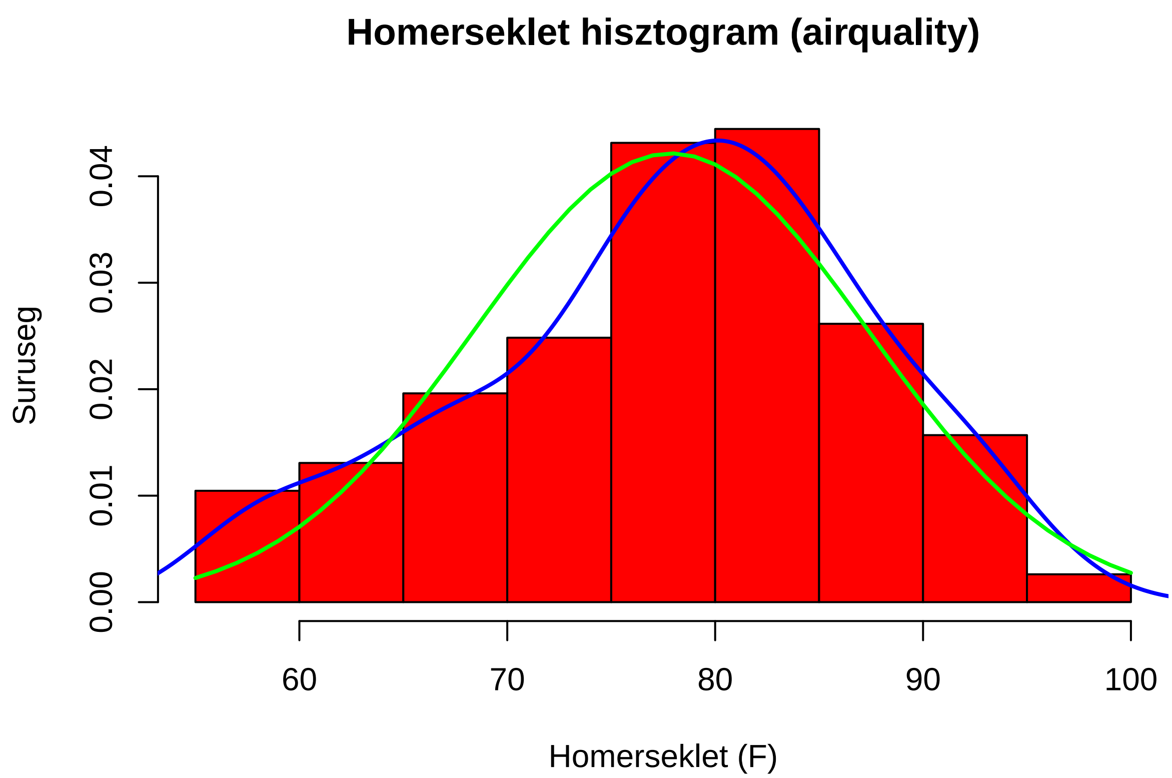
```
##  [1]  0  0  1  5  2  5  3  5  8  4  6  9 13 13 11 20  9 12  8  5  7  5  1  1  0
```

```
hist(airquality$Temp, freq = FALSE,
     col = "red",
     xlab = "Homerseklet (F)",
     ylab = "Suruseg",
     main = "Homerseklet hisztogram (airquality)")
lines(density(airquality$Temp), lwd = 2, col = "blue")
```



Ez már közelíthető normális eloszlással (ennek megsejtésére van a statisztika)

```
hist(airquality$Temp, freq = FALSE,
     col = "red",
     xlab = "Homerseklet (F)",
     ylab = "Suruseg",
     main = "Homerseklet hisztogram (airquality)")
lines(density(airquality$Temp), lwd = 2, col = "blue")
x <- seq(55, 100, 1)
dn <- dnorm(x, mean = mean(airquality$Temp), sd = sd(airquality$Temp))
lines(x, dn, type = "l", lwd = 2, col = "green")
```



Extra feladatok

Ezeket érdemes megcsinálni, mert sokat fog mesélni arról, hogyan működik az R. Ha bármelyikkel kapcsolatban van kérdésetek, keressetek!

```
adat = c(2,0,1,0,8,3,5,7,8,2,3,5,1,7,8,3,5,3,2,8)
```

Mit számolnak az alábbi R programok?

1.

```
sum(adat < 3)
```

```
## [1] 7
```

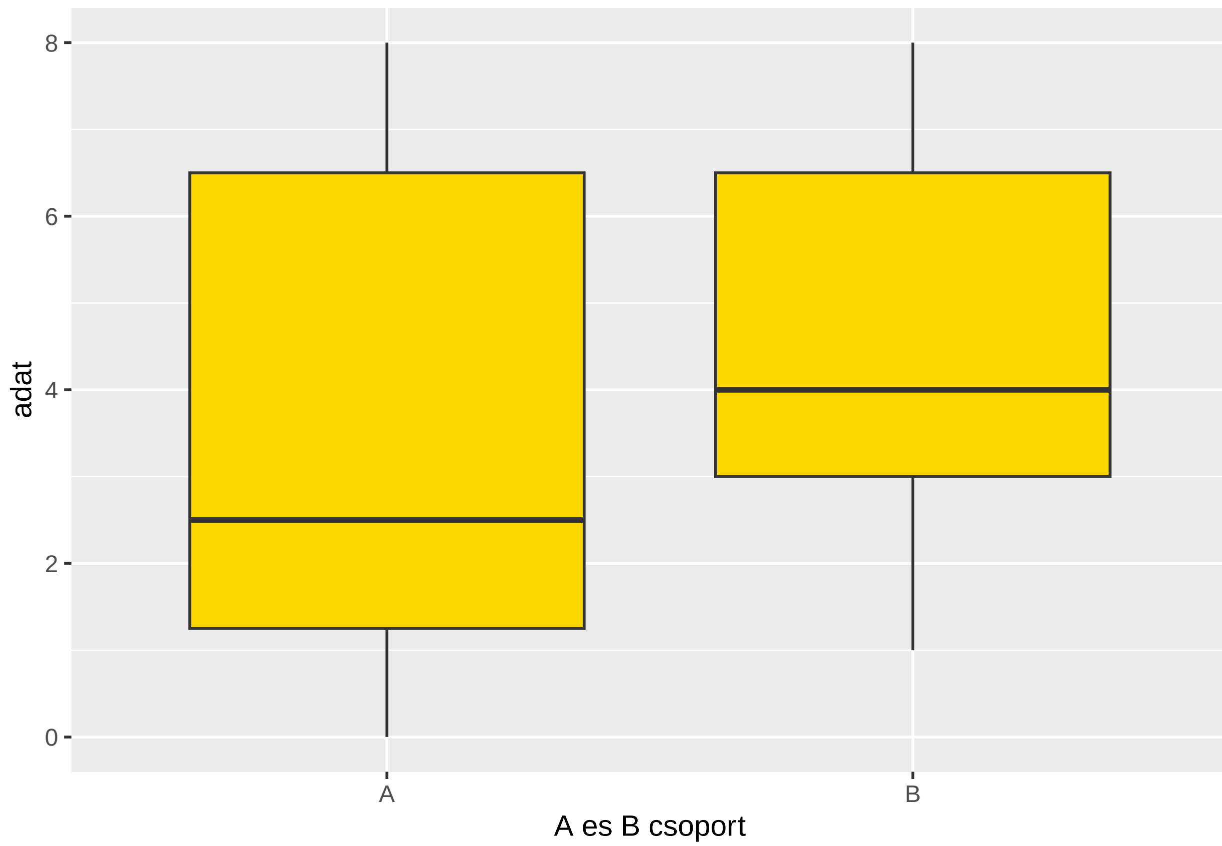
2.

```
t_adat <- table(adat)
names(t_adat)[t_adat == max(t_adat)]
```

```
## [1] "3" "8"
```

3.

```r
rep <- rep(c("A", "B"), c(10, 10))
df <- data.frame(adat, rep = rep)
ggplot(df, aes(x = rep, y = adat)) +
  geom_boxplot(fill = "gold") +
  scale_x_discrete(name = "A es B csoport")
```



Az alábbi érték TRUE, vagy FALSE? Ha FALSE, akkor hogyan javítható?

```r
sd(adat) == sqrt(sum((adat-mean(adat))^2)/length(adat))
```

Tapasztalati eloszlásfüggvény

Adott egy mintánk, azaz egy $X_1, \ldots, X_n$ valószínűségi változósorozat. Feltesszük, hogy ezek függetlenek, és azonos eloszlásúak. Ezek egy realizációja az $x_1, \ldots, x_n$ sorozat.

Statisztikának nevezzük a minta bármely függvényét, ilyen pl. az átlag, a tapasztalati szórásnégyzet, vagy a tapasztalati eloszlásfüggvény:

$$F_n(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{I}(X_i < x),$$

ahol $\mathbb{I}(X_i < x)$ értéke 1, ha $X_i > x$, különben 0 – ez az indikátorfüggvény (emlékezzünk vissza az indikátoreloszlásra!)

Glivenko-Cantelli tétel:

$|F_n(x) - F(x)| \to 0$ egyenletesen, 1 valószínűséggel. Vagyis ha a minta elemszáma elég nagy, akkor (szinte) minden x-re $F_n(x)$ értéke közel van $F(x)$ értékéhez – azaz a tapasztalati eloszlásfüggvény tekinthető a valós eloszlásfüggvénynek.

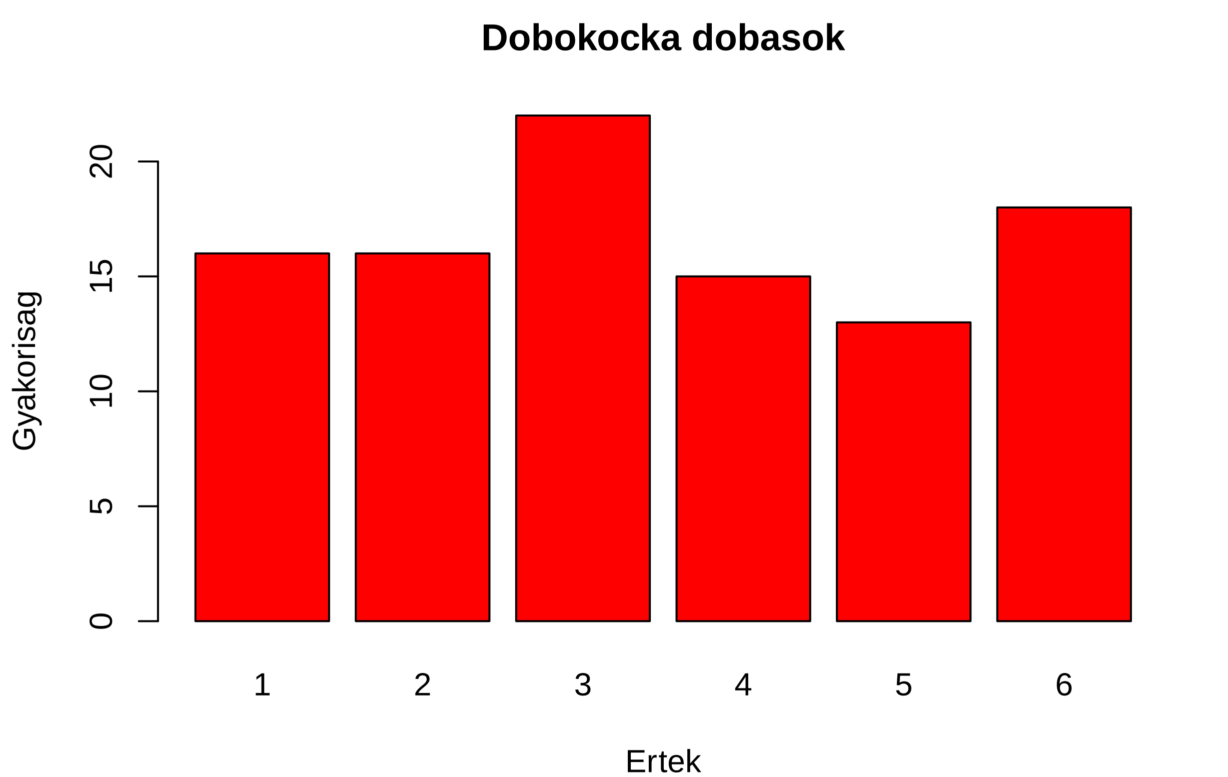
3. Feladat

Generáljunk 100 kockadobást, és ábrázoljuk annak tapasztalati eloszlásfüggvényét!

```
num_values <- 100
x_sample <- sample(1:6, size = num_values, replace = TRUE)
table(x_sample)
```

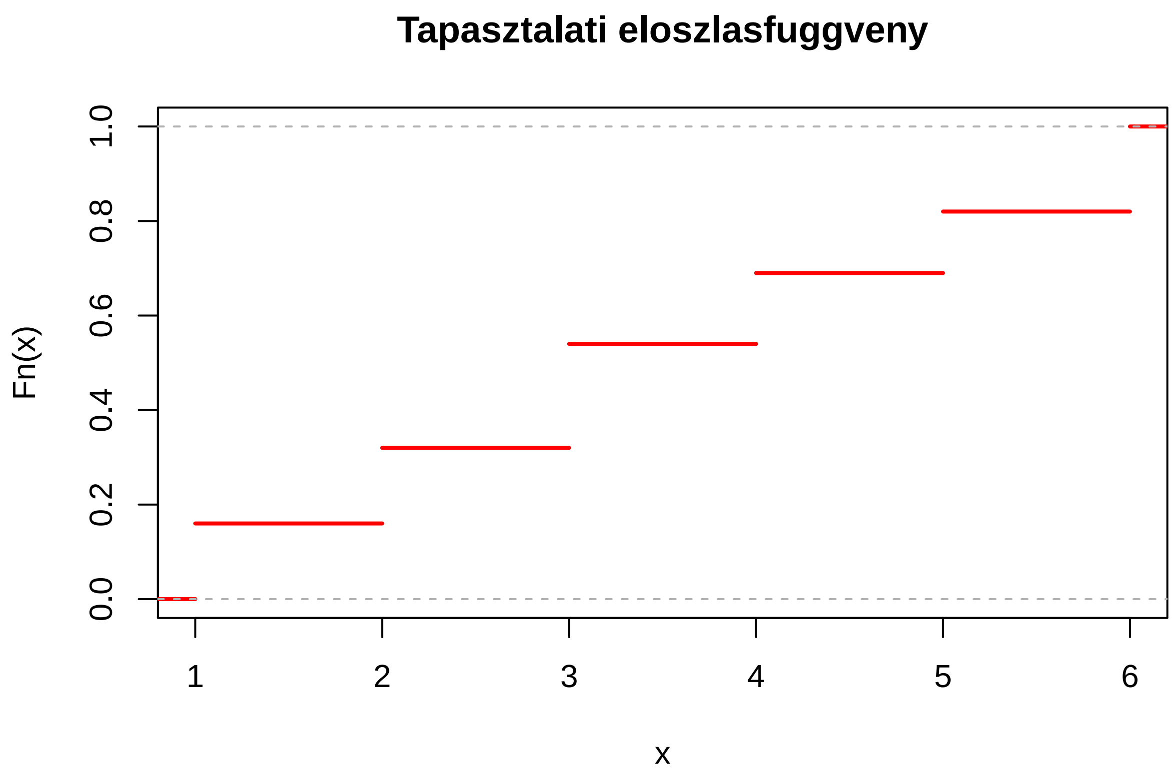
```
## x_sample
##  1  2  3  4  5  6
## 16 16 22 15 13 18
```

```
barplot(table(x_sample),
        col = "red",
        xlab = "Ertek",
        ylab = "Gyakorisag",
        main = "Dobokocka dobasok")
```




```
plot(ecdf(x_sample),
     do.points = FALSE,
     col = "red",
     lwd = 2,
```

```
     xlim = c(1, 6),
     main = "Tapasztalati eloszlasfuggveny")
```



Szimuláljunk egy kicsit, és nézzük meg, hogy egy eloszlás elméleti és tapasztalati eloszlásfüggvénye hogyan viszonyul egymáshoz!

```
x <- 1:6
trueF <- ecdf(x)
n_values <- 100
x_sample <- sample(1:6, size = n_values, replace = TRUE)
```

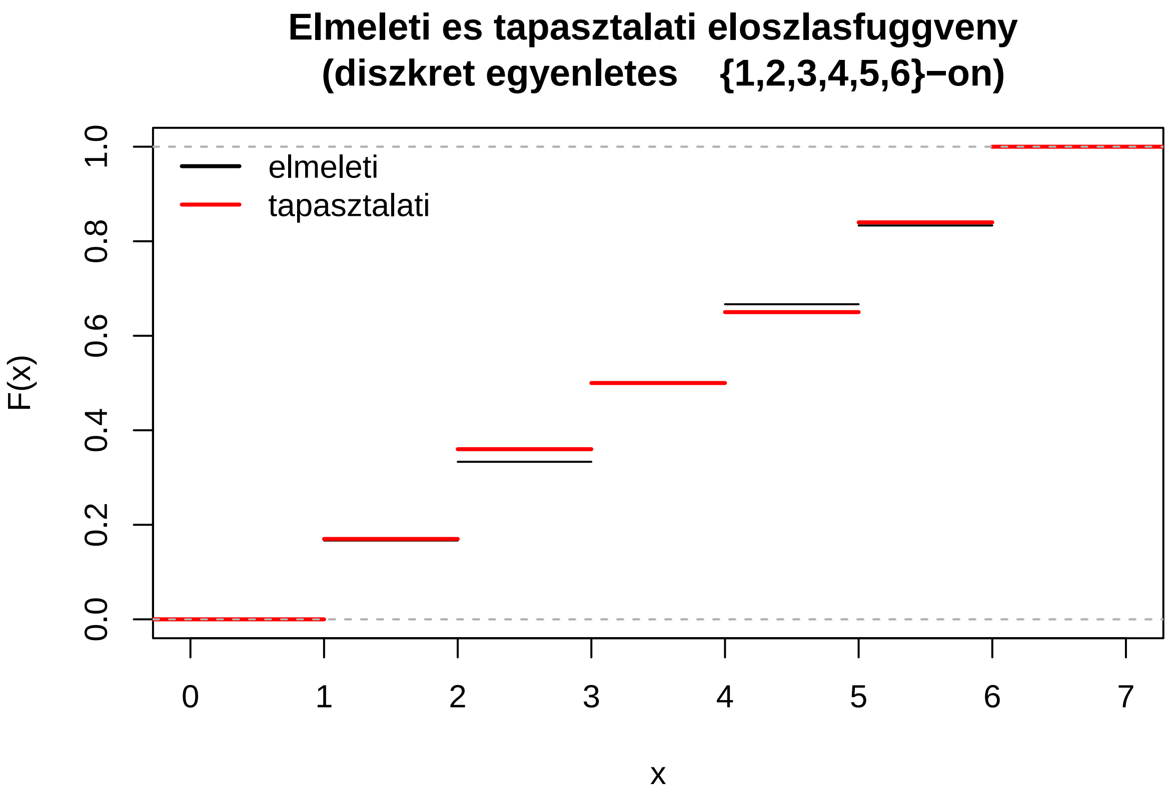
```
floor(runif(n_values, min = 1, max = 7))
```

```
##   [1] 2 4 5 1 2 1 6 4 3 4 1 4 5 5 5 3 3 5 4 6 5 4 3 2 3 2 5 3 4 6 6 4 6 5 5 2 3
##  [38] 6 1 4 4 2 4 3 6 2 1 3 2 3 5 6 3 1 1 2 5 5 6 5 2 5 4 3 6 3 5 6 3 5 5 4 4 3
##  [75] 3 4 3 5 3 3 2 2 5 3 3 5 2 2 6 3 2 5 4 6 3 3 6 4 6 2
```

Hasonlítsuk most ezeket össze:

```
plot(trueF,
     do.points = FALSE,
     ylab = "F(x)",
     main = "Elmeleti es tapasztalati eloszlasfuggveny \n (diszkret egyenletes    {1,2,3,4,5,6}-on)")
plot(ecdf(x_sample), add = TRUE,
```

```
      do.points = FALSE,
      xlim = c(1, 6),
      col = "red",
      lwd = 2)
legend(x = 'topleft',
       bty = 'n',
       legend = c("elmeleti", "tapasztalati"),
       col = c("black", "red"),
       lwd = 2)
```



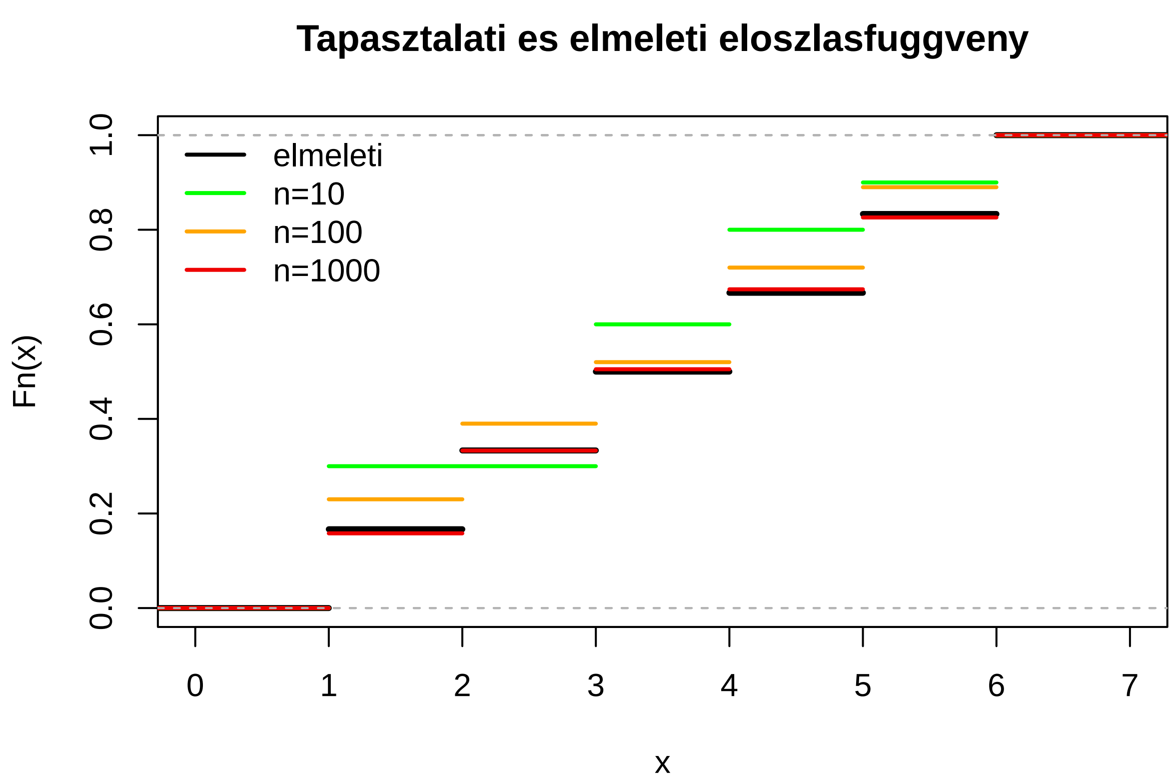

```
#points(unique(x), unique(c(0, trueF(x)))[1:length(unique(x))], pch = 19)
#points(unique(x), unique(trueF(x), pch = 21))
```

Több szimulációt is meg tudunk nézni egy ábrán:

```
n_values <- c(10, 100, 1000)
cols <- c("black", "green", "orange", "red2")
#colorspace::diverge_hsv(length(n_values) + 1)

i <- 0
plot(trueF,
      do.points = FALSE,
      col = cols[i <- i + 1],
      lwd = 3,
      main = "Tapasztalati es elmeleti eloszlasfuggveny")
```

```
for (n in n_values)
{
  x_sample <- sample(1:6, size = n, replace = TRUE)
  plot(ecdf(x_sample),
       add = TRUE,
       do.points = FALSE,
       col = cols[i <- i + 1],
       lwd = 2)
}
legend(x = 'topleft',
       bty = 'n',
       col = cols,
       lwd = 2,
       legend = c('elmeleti', paste('n', n_values, sep="=")))
```



4. Feladat

Szimuláljuk ugyanezt a (standard) normális eloszlásra is!

```
n <- 12
x <- rnorm(n)

plot(ecdf(x),
     do.points = FALSE,
     xlim = c(-3.2, 3.2),
     col = "red",
```

```
      lwd = 2,
      main = "Elmeleti es tapasztalati eloszlasfuggveny \n (abszolut folytonos val. valt.: standard norma
      ylab = "")

x_sample <- seq(-3.2, 3.2, 0.01)
lines(x_sample, pnorm(x_sample), lwd = 2) #elmeleti
legend(x = "topleft",
       bty = "n",
       col = c("black", "red"),
       lwd = 2,
       legend = c("elmeleti", "tapasztalati"))
```