SAP - Prva auditorna vježba

Case study *Iris data*: Deskriptivna statistika i vizualizacija podataka

Tessa Bauman, Stjepan Begušić, David Bojanić, Krunoslav Jurčić, Tomislav Kovačević, Andro Merćep

20.10.2021.

Uvod

Vježbe i projekt na predmetu "Statistička analiza podataka" izvode se u programskom jeziku R, radnoj okolini RStudio, u obliku R Markdown izvještaja koji kombiniraju pisanje teksta s programskim kodom i rezultatima izvođenja koda.

Pojedine auditorne vježbe bavit će se konkretnim case study-jem kroz koji će se demonstrirati praktična strana obrađenog gradiva.

Case study: *Iris data*

R uključuje razne ugrađene skupove podataka u sklopu paketa `datasets`.

```
library(help = "datasets")
```

Jedan od poznatijih skupova podataka su podatci Edgara Andersona o duljinama i širinama lapova i latica cvjetova irisa.

```
help(iris) #help ili ?
```

```
## starting httpd help server ... done
```

Dataset *iris* sastoji se od 3 vrste cvijeta iris - *Iris setosa*, *versicolor*, i *virginica*. 150 je primjera u datasetu; svaki primjer sastoji se od 5 varijabli.

Prije svega, bitno je znati kontekst podataka! Interpretacija podataka je značajan dio obrade podataka. Upoznajmo se s datasetom *iris*:

```
# Učitavanje built-in dataseta i pregled prvih nekoliko redaka
irisdata = iris
head(irisdata)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```r
#irisdata
```

Što predstavljaju pojedine varijable? Koja je koja? Što možemo iz njih naslutiti? Koja je svrha? Kakve analize možemo provesti? Kakve rezultate potencijalno možemo dobiti?

```r
knitr::include_graphics("iris-machinelearning.png")
```

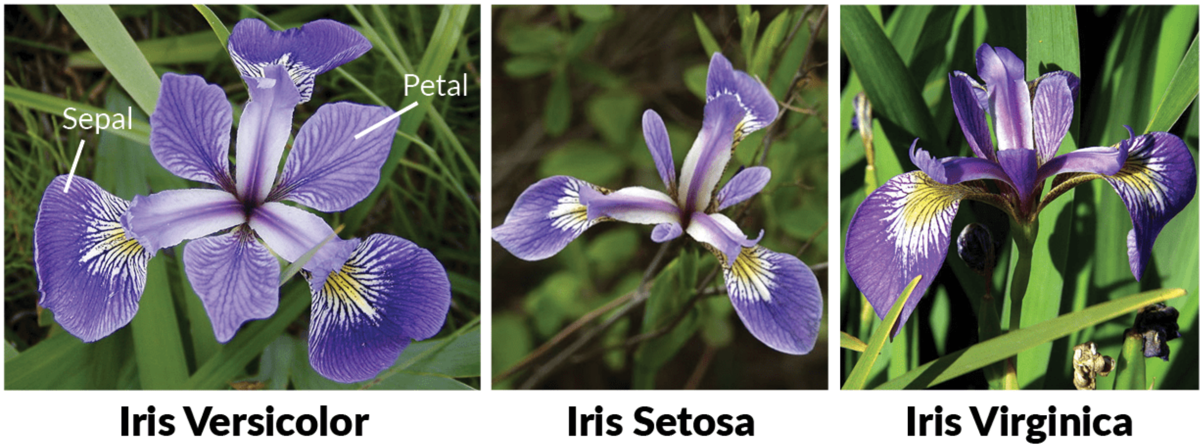


Figure 1: Iris species

```r
knitr::include_graphics("iris_petal-sepal-width-length.png")
```

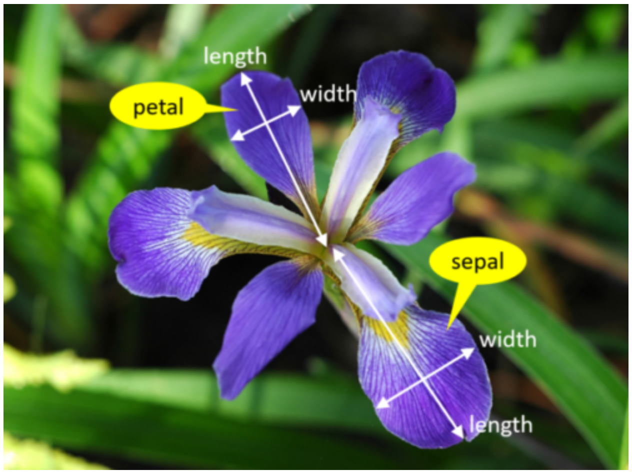


Figure 2: Sepal and petal width and length

Osnovne manipulacije nad datasetom:

```r
# Dimenzije dataseta:
dim(irisdata)  # broj redaka, broj stupaca (broj primjera, broj varijabli)
```

```
## [1] 150   5
```

```
nrow(irisdata) # broj redaka
```

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

```
ncol(irisdata) # broj stupaca -> što daje length?
```

```
## [1] 5
```

```
names(irisdata) # imena stupaca
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

```
class(irisdata)
```

```
## [1] "data.frame"
```

```
# Uvodna analiza, pristup stupcima data.frame objekta preko imena pomocu operatora $
irisdata$Sepal.Length
```

```
##   [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1
##  [19] 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.0
##  [37] 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6 5.3 5.0 7.0 6.4 6.9 5.5
##  [55] 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1
##  [73] 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5
##  [91] 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3
## [109] 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2
## [127] 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8
## [145] 6.7 6.7 6.3 6.5 6.2 5.9
```

```
class(irisdata$Petal.Width)
```

```
## [1] "numeric"
```

```
# klasa ove varijable je "numeric" -- varijabla na intervalnoj/racionalnoj skali - koja od njih u ovom

irisdata$Species
```

```
##   [1] setosa     setosa     setosa     setosa     setosa     setosa
##   [7] setosa     setosa     setosa     setosa     setosa     setosa
##  [13] setosa     setosa     setosa     setosa     setosa     setosa
##  [19] setosa     setosa     setosa     setosa     setosa     setosa
##  [25] setosa     setosa     setosa     setosa     setosa     setosa
##  [31] setosa     setosa     setosa     setosa     setosa     setosa
##  [37] setosa     setosa     setosa     setosa     setosa     setosa
##  [43] setosa     setosa     setosa     setosa     setosa     setosa
##  [49] setosa     setosa     versicolor versicolor versicolor versicolor
##  [55] versicolor versicolor versicolor versicolor versicolor versicolor
##  [61] versicolor versicolor versicolor versicolor versicolor versicolor
##  [67] versicolor versicolor versicolor versicolor versicolor versicolor
##  [73] versicolor versicolor versicolor versicolor versicolor versicolor
##  [79] versicolor versicolor versicolor versicolor versicolor versicolor
##  [85] versicolor versicolor versicolor versicolor versicolor versicolor
##  [91] versicolor versicolor versicolor versicolor versicolor versicolor
##  [97] versicolor versicolor versicolor versicolor virginica  virginica
## [103] virginica  virginica  virginica  virginica  virginica  virginica
## [109] virginica  virginica  virginica  virginica  virginica  virginica
## [115] virginica  virginica  virginica  virginica  virginica  virginica
## [121] virginica  virginica  virginica  virginica  virginica  virginica
## [127] virginica  virginica  virginica  virginica  virginica  virginica
## [133] virginica  virginica  virginica  virginica  virginica  virginica
## [139] virginica  virginica  virginica  virginica  virginica  virginica
## [145] virginica  virginica  virginica  virginica  virginica  virginica
## Levels: setosa versicolor virginica
```

```
class(irisdata$Species)
```

```
## [1] "factor"
```

```
# klasa ove varijable je "factor" -- varijabla na nominalnoj/ordinalnoj skali - koja od njih u ovom slu
```

```
# Indeksiranje

# Jedan indeks izdvaja stupac ili sve osim određenih stupaca
irisdata[2]
```

```
##     Sepal.Width
## 1           3.5
## 2           3.0
## 3           3.2
## 4           3.1
## 5           3.6
## 6           3.9
## 7           3.4
## 8           3.4
## 9           2.9
## 10          3.1
## 11          3.7
## 12          3.4
## 13          3.0
## 14          3.0
## 15          4.0
## 16          4.4
## 17          3.9
## 18          3.5
## 19          3.8
## 20          3.8
## 21          3.4
## 22          3.7
## 23          3.6
## 24          3.3
## 25          3.4
## 26          3.0
## 27          3.4
## 28          3.5
## 29          3.4
## 30          3.2
## 31          3.1
## 32          3.4
## 33          4.1
## 34          4.2
## 35          3.1
## 36          3.2
## 37          3.5
## 38          3.6
## 39          3.0
## 40          3.4
## 41          3.5
## 42          2.3
## 43          3.2
## 44          3.5
## 45          3.8
## 46          3.0
## 47          3.8
## 48          3.2
## 49          3.7
## 50          3.3
## 51          3.2
## 52          3.2
## 53          3.1
## 54          2.3
## 55          2.8
## 56          2.8
## 57          3.3
## 58          2.4
## 59          2.9
## 60          2.7
## 61          2.0
## 62          3.0
## 63          2.2
## 64          2.9
## 65          2.9
## 66          3.1
## 67          3.0
## 68          2.7
## 69          2.2
## 70          2.5
## 71          3.2
## 72          2.8
## 73          2.5
## 74          2.8
## 75          2.9
## 76          3.0
## 77          2.8
## 78          3.0
## 79          2.9
## 80          2.6
## 81          2.4
## 82          2.4
## 83          2.7
## 84          2.7
## 85          3.0
## 86          3.4
## 87          3.1
## 88          2.3
## 89          3.0
## 90          2.5
## 91          2.6
## 92          3.0
## 93          2.6
## 94          2.3
## 95          2.7
## 96          3.0
## 97          2.9
## 98          2.9
## 99          2.5
## 100         2.8
## 101         3.3
## 102         2.7
## 103         3.0
## 104         2.9
## 105         3.0
## 106         3.0
## 107         2.5
## 108         2.9
## 109         2.5
## 110         3.6
## 111         3.2
## 112         2.7
## 113         3.0
## 114         2.5
## 115         2.8
## 116         3.2
## 117         3.0
## 118         3.8
## 119         2.6
## 120         2.2
## 121         3.2
## 122         2.8
## 123         2.8
## 124         2.7
## 125         3.3
## 126         3.2
## 127         2.8
## 128         3.0
## 129         2.8
## 130         3.0
## 131         2.8
## 132         3.8
## 133         2.8
## 134         2.8
## 135         2.6
## 136         3.0
## 137         3.4
## 138         3.1
## 139         3.0
## 140         3.1
## 141         3.1
## 142         3.1
## 143         2.7
## 144         3.2
## 145         3.3
## 146         3.0
## 147         2.5
## 148         3.0
## 149         3.4
## 150         3.0
```

```
irisdata[c(2,4)]
```

```
##     Sepal.Width Petal.Width
## 1           3.5         0.2
## 2           3.0         0.2
## 3           3.2         0.2
## 4           3.1         0.2
## 5           3.6         0.2
## 6           3.9         0.4
## 7           3.4         0.3
## 8           3.4         0.2
## 9           2.9         0.2
## 10          3.1         0.1
## 11          3.7         0.2
## 12          3.4         0.2
## 13          3.0         0.1
## 14          3.0         0.1
## 15          4.0         0.2
## 16          4.4         0.4
## 17          3.9         0.4
## 18          3.5         0.3
## 19          3.8         0.3
## 20          3.8         0.3
## 21          3.4         0.2
## 22          3.7         0.4
## 23          3.6         0.2
## 24          3.3         0.5
## 25          3.4         0.2
## 26          3.0         0.2
## 27          3.4         0.4
## 28          3.5         0.2
## 29          3.4         0.2
## 30          3.2         0.2
## 31          3.1         0.2
## 32          3.4         0.4
## 33          4.1         0.1
## 34          4.2         0.2
## 35          3.1         0.2
## 36          3.2         0.2
## 37          3.5         0.2
## 38          3.6         0.1
## 39          3.0         0.2
## 40          3.4         0.2
## 41          3.5         0.3
## 42          2.3         0.3
## 43          3.2         0.2
## 44          3.5         0.6
## 45          3.8         0.4
## 46          3.0         0.3
## 47          3.8         0.2
## 48          3.2         0.2
## 49          3.7         0.2
## 50          3.3         0.2
## 51          3.2         1.4
## 52          3.2         1.5
## 53          3.1         1.5
## 54          2.3         1.3
## 55          2.8         1.5
## 56          2.8         1.3
## 57          3.3         1.6
## 58          2.4         1.0
## 59          2.9         1.3
## 60          2.7         1.4
## 61          2.0         1.0
## 62          3.0         1.5
## 63          2.2         1.0
## 64          2.9         1.4
## 65          2.9         1.3
## 66          3.1         1.4
## 67          3.0         1.5
## 68          2.7         1.0
## 69          2.2         1.5
## 70          2.5         1.1
## 71          3.2         1.8
## 72          2.8         1.3
## 73          2.5         1.5
## 74          2.8         1.2
## 75          2.9         1.3
## 76          3.0         1.4
## 77          2.8         1.4
## 78          3.0         1.7
## 79          2.9         1.5
## 80          2.6         1.0
## 81          2.4         1.1
## 82          2.4         1.0
## 83          2.7         1.2
## 84          2.7         1.6
## 85          3.0         1.5
## 86          3.4         1.6
## 87          3.1         1.5
## 88          2.3         1.3
## 89          3.0         1.3
## 90          2.5         1.3
## 91          2.6         1.2
## 92          3.0         1.4
## 93          2.6         1.2
## 94          2.3         1.0
## 95          2.7         1.3
## 96          3.0         1.2
## 97          2.9         1.3
## 98          2.9         1.3
## 99          2.5         1.1
## 100         2.8         1.3
## 101         3.3         2.5
## 102         2.7         1.9
## 103         3.0         2.1
## 104         2.9         1.8
## 105         3.0         2.2
## 106         3.0         2.1
## 107         2.5         1.7
## 108         2.9         1.8
## 109         2.5         1.8
## 110         3.6         2.5
## 111         3.2         2.0
## 112         2.7         1.9
## 113         3.0         2.1
## 114         2.5         2.0
## 115         2.8         2.4
## 116         3.2         2.3
## 117         3.0         1.8
## 118         3.8         2.2
## 119         2.6         2.3
## 120         2.2         1.5
## 121         3.2         2.3
## 122         2.8         2.0
## 123         2.8         2.0
## 124         2.7         1.8
## 125         3.3         2.1
## 126         3.2         1.8
## 127         2.8         1.8
## 128         3.0         1.8
## 129         2.8         2.1
## 130         3.0         1.6
## 131         2.8         1.9
## 132         3.8         2.0
## 133         2.8         2.2
## 134         2.8         1.5
## 135         2.6         1.4
## 136         3.0         2.3
## 137         3.4         2.4
## 138         3.1         1.8
## 139         3.0         1.8
## 140         3.1         2.1
## 141         3.1         2.4
## 142         3.1         2.3
## 143         2.7         1.9
## 144         3.2         2.3
## 145         3.3         2.5
## 146         3.0         2.3
## 147         2.5         1.9
## 148         3.0         2.0
## 149         3.4         2.3
## 150         3.0         1.8
```

```
irisdata[-c(2,4)]
```

```
##     Sepal.Length Petal.Length    Species
## 1            5.1          1.4     setosa
## 2            4.9          1.4     setosa
## 3            4.7          1.3     setosa
## 4            4.6          1.5     setosa
## 5            5.0          1.4     setosa
## 6            5.4          1.7     setosa
## 7            4.6          1.4     setosa
## 8            5.0          1.5     setosa
## 9            4.4          1.4     setosa
## 10           4.9          1.5     setosa
## 11           5.4          1.5     setosa
## 12           4.8          1.6     setosa
## 13           4.8          1.4     setosa
## 14           4.3          1.1     setosa
## 15           5.8          1.2     setosa
## 16           5.7          1.5     setosa
## 17           5.4          1.3     setosa
## 18           5.1          1.4     setosa
## 19           5.7          1.7     setosa
## 20           5.1          1.5     setosa
## 21           5.4          1.7     setosa
## 22           5.1          1.5     setosa
## 23           4.6          1.0     setosa
## 24           5.1          1.7     setosa
## 25           4.8          1.9     setosa
## 26           5.0          1.6     setosa
## 27           5.0          1.6     setosa
## 28           5.2          1.5     setosa
## 29           5.2          1.4     setosa
## 30           4.7          1.6     setosa
## 31           4.8          1.6     setosa
## 32           5.4          1.5     setosa
## 33           5.2          1.5     setosa
## 34           5.5          1.4     setosa
## 35           4.9          1.5     setosa
## 36           5.0          1.2     setosa
## 37           5.5          1.3     setosa
## 38           4.9          1.4     setosa
## 39           4.4          1.3     setosa
## 40           5.1          1.5     setosa
## 41           5.0          1.3     setosa
## 42           4.5          1.3     setosa
## 43           4.4          1.3     setosa
## 44           5.0          1.6     setosa
## 45           5.1          1.9     setosa
## 46           4.8          1.4     setosa
## 47           5.1          1.6     setosa
## 48           4.6          1.4     setosa
## 49           5.3          1.5     setosa
## 50           5.0          1.4     setosa
## 51           7.0          4.7 versicolor
## 52           6.4          4.5 versicolor
## 53           6.9          4.9 versicolor
## 54           5.5          4.0 versicolor
## 55           6.5          4.6 versicolor
## 56           5.7          4.5 versicolor
## 57           6.3          4.7 versicolor
## 58           4.9          3.3 versicolor
## 59           6.6          4.6 versicolor
## 60           5.2          3.9 versicolor
## 61           5.0          3.5 versicolor
## 62           5.9          4.2 versicolor
## 63           6.0          4.0 versicolor
## 64           6.1          4.7 versicolor
## 65           5.6          3.6 versicolor
## 66           6.7          4.4 versicolor
## 67           5.6          4.5 versicolor
## 68           5.8          4.1 versicolor
## 69           6.2          4.5 versicolor
## 70           5.6          3.9 versicolor
## 71           5.9          4.8 versicolor
## 72           6.1          4.0 versicolor
## 73           6.3          4.9 versicolor
## 74           6.1          4.7 versicolor
## 75           6.4          4.3 versicolor
## 76           6.6          4.4 versicolor
## 77           6.8          4.8 versicolor
## 78           6.7          5.0 versicolor
## 79           6.0          4.5 versicolor
## 80           5.7          3.5 versicolor
## 81           5.5          3.8 versicolor
## 82           5.5          3.7 versicolor
## 83           5.8          3.9 versicolor
## 84           6.0          5.1 versicolor
## 85           5.4          4.5 versicolor
## 86           6.0          4.5 versicolor
## 87           6.7          4.7 versicolor
## 88           6.3          4.4 versicolor
## 89           5.6          4.1 versicolor
## 90           5.5          4.0 versicolor
## 91           5.5          4.4 versicolor
## 92           6.1          4.6 versicolor
## 93           5.8          4.0 versicolor
## 94           5.0          3.3 versicolor
## 95           5.6          4.2 versicolor
## 96           5.7          4.2 versicolor
## 97           5.7          4.2 versicolor
## 98           6.2          4.3 versicolor
## 99           5.1          3.0 versicolor
## 100          5.7          4.1 versicolor
## 101          6.3          6.0  virginica
## 102          5.8          5.1  virginica
## 103          7.1          5.9  virginica
## 104          6.3          5.6  virginica
## 105          6.5          5.8  virginica
## 106          7.6          6.6  virginica
## 107          4.9          4.5  virginica
## 108          7.3          6.3  virginica
## 109          6.7          5.8  virginica
## 110          7.2          6.1  virginica
## 111          6.5          5.1  virginica
## 112          6.4          5.3  virginica
## 113          6.8          5.5  virginica
## 114          5.7          5.0  virginica
## 115          5.8          5.1  virginica
## 116          6.4          5.3  virginica
## 117          6.5          5.5  virginica
## 118          7.7          6.7  virginica
## 119          7.7          6.9  virginica
## 120          6.0          5.0  virginica
## 121          6.9          5.7  virginica
## 122          5.6          4.9  virginica
## 123          7.7          6.7  virginica
## 124          6.3          4.9  virginica
## 125          6.7          5.7  virginica
## 126          7.2          6.0  virginica
## 127          6.2          4.8  virginica
## 128          6.1          4.9  virginica
## 129          6.4          5.6  virginica
## 130          7.2          5.8  virginica
## 131          7.4          6.1  virginica
## 132          7.9          6.4  virginica
## 133          6.4          5.6  virginica
## 134          6.3          5.1  virginica
## 135          6.1          5.6  virginica
## 136          7.7          6.1  virginica
## 137          6.3          5.6  virginica
## 138          6.4          5.5  virginica
## 139          6.0          4.8  virginica
## 140          6.9          5.4  virginica
## 141          6.7          5.6  virginica
## 142          6.9          5.1  virginica
## 143          5.8          5.1  virginica
## 144          6.8          5.9  virginica
## 145          6.7          5.7  virginica
## 146          6.7          5.2  virginica
## 147          6.3          5.0  virginica
## 148          6.5          5.2  virginica
## 149          6.2          5.4  virginica
## 150          5.9          5.1  virginica
```

```
# Kod vektora od dva indeksa prvi predstavlja redak a drugi stupac
irisdata[c(2,5,6), 3:5]
```

```
##   Petal.Length Petal.Width Species
## 2          1.4         0.2  setosa
## 5          1.4         0.2  setosa
## 6          1.7         0.4  setosa
```

```
# primijetiti razliku:
# retci 3 i 4 i svi stupci
irisdata[c(3,4),]
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
```

```
# svi retci i stupci 3 i 4
irisdata[,c(3,4)]
```

```
##     Petal.Length Petal.Width
## 1            1.4         0.2
## 2            1.4         0.2
## 3            1.3         0.2
## 4            1.5         0.2
## 5            1.4         0.2
## 6            1.7         0.4
## 7            1.4         0.3
## 8            1.5         0.2
## 9            1.4         0.2
## 10           1.5         0.1
## 11           1.5         0.2
## 12           1.6         0.2
## 13           1.4         0.1
## 14           1.1         0.1
## 15           1.2         0.2
## 16           1.5         0.4
## 17           1.3         0.4
## 18           1.4         0.3
## 19           1.7         0.3
## 20           1.5         0.3
## 21           1.7         0.2
## 22           1.5         0.4
## 23           1.0         0.2
## 24           1.7         0.5
## 25           1.9         0.2
## 26           1.6         0.2
## 27           1.6         0.4
## 28           1.5         0.2
## 29           1.4         0.2
## 30          1.6         0.2
## 31          1.6         0.2
## 32          1.5         0.4
## 33          1.5         0.1
## 34          1.4         0.2
## 35          1.5         0.2
## 36          1.2         0.2
## 37          1.3         0.2
## 38          1.4         0.1
## 39          1.3         0.2
## 40          1.5         0.2
## 41          1.3         0.3
## 42          1.3         0.3
## 43          1.3         0.2
## 44          1.6         0.6
## 45          1.9         0.4
## 46          1.4         0.3
## 47          1.6         0.2
## 48          1.4         0.2
## 49          1.5         0.2
## 50          1.4         0.2
## 51          4.7         1.4
## 52          4.5         1.5
## 53          4.9         1.5
## 54          4.0         1.3
## 55          4.6         1.5
## 56          4.5         1.3
## 57          4.7         1.6
## 58          3.3         1.0
## 59          4.6         1.3
## 60          3.9         1.4
## 61          3.5         1.0
## 62          4.2         1.5
## 63          4.0         1.0
## 64          4.7         1.4
## 65          3.6         1.3
## 66          4.4         1.4
## 67          4.5         1.5
## 68          4.1         1.0
## 69          4.5         1.5
## 70          3.9         1.1
## 71          4.8         1.8
## 72          4.0         1.3
## 73          4.9         1.5
## 74          4.7         1.2
## 75          4.3         1.3
## 76          4.4         1.4
## 77          4.8         1.4
## 78          5.0         1.7
## 79          4.5         1.5
## 80          3.5         1.0
## 81          3.8         1.1
## 82          3.7         1.0
## 83          3.9         1.2
## 84           5.1         1.6
## 85           4.5         1.5
## 86           4.5         1.6
## 87           4.7         1.5
## 88           4.4         1.3
## 89           4.1         1.3
## 90           4.0         1.3
## 91           4.4         1.2
## 92           4.6         1.4
## 93           4.0         1.2
## 94           3.3         1.0
## 95           4.2         1.3
## 96           4.2         1.2
## 97           4.2         1.3
## 98           4.3         1.3
## 99           3.0         1.1
## 100          4.1         1.3
## 101          6.0         2.5
## 102          5.1         1.9
## 103          5.9         2.1
## 104          5.6         1.8
## 105          5.8         2.2
## 106          6.6         2.1
## 107          4.5         1.7
## 108          6.3         1.8
## 109          5.8         1.8
## 110          6.1         2.5
## 111          5.1         2.0
## 112          5.3         1.9
## 113          5.5         2.1
## 114          5.0         2.0
## 115          5.1         2.4
## 116          5.3         2.3
## 117          5.5         1.8
## 118          6.7         2.2
## 119          6.9         2.3
## 120          5.0         1.5
## 121          5.7         2.3
## 122          4.9         2.0
## 123          6.7         2.0
## 124          4.9         1.8
## 125          5.7         2.1
## 126          6.0         1.8
## 127          4.8         1.8
## 128          4.9         1.8
## 129          5.6         2.1
## 130          5.8         1.6
## 131          6.1         1.9
## 132          6.4         2.0
## 133          5.6         2.2
## 134          5.1         1.5
## 135          5.6         1.4
## 136          6.1         2.3
## 137          5.6         2.4
## 138          5.5         1.8
## 139          4.8         1.8
## 140          5.4         2.1
## 141          5.6         2.4
## 142          5.1         2.3
## 143          5.1         1.9
## 144          5.9         2.3
## 145          5.7         2.5
## 146          5.2         2.3
## 147          5.0         1.9
## 148          5.2         2.0
## 149          5.4         2.3
## 150          5.1         1.8
```

```
irisdata[c(3,4)]
```

```
##     Petal.Length Petal.Width
## 1            1.4         0.2
## 2            1.4         0.2
## 3            1.3         0.2
## 4            1.5         0.2
## 5            1.4         0.2
## 6            1.7         0.4
## 7            1.4         0.3
## 8            1.5         0.2
## 9            1.4         0.2
## 10           1.5         0.1
## 11           1.5         0.2
## 12           1.6         0.2
## 13           1.4         0.1
## 14           1.1         0.1
## 15           1.2         0.2
## 16           1.5         0.4
## 17           1.3         0.4
## 18           1.4         0.3
## 19           1.7         0.3
## 20           1.5         0.3
## 21           1.7         0.2
## 22           1.5         0.4
## 23           1.0         0.2
## 24           1.7         0.5
## 25           1.9         0.2
## 26           1.6         0.2
## 27           1.6         0.4
## 28           1.5         0.2
## 29           1.4         0.2
## 30           1.6         0.2
## 31           1.6         0.2
## 32           1.5         0.4
## 33           1.5         0.1
## 34           1.4         0.2
## 35           1.5         0.2
## 36           1.2         0.2
## 37           1.3         0.2
## 38           1.4         0.1
## 39           1.3         0.2
## 40           1.5         0.2
## 41           1.3         0.3
## 42           1.3         0.3
## 43           1.3         0.2
## 44           1.6         0.6
## 45           1.9         0.4
## 46           1.4         0.3
## 47           1.6         0.2
## 48           1.4         0.2
## 49           1.5         0.2
## 50           1.4         0.2
## 51           4.7         1.4
## 52           4.5         1.5
## 53           4.9         1.5
## 54           4.0         1.3
## 55           4.6         1.5
## 56           4.5         1.3
## 57           4.7         1.6
## 58           3.3         1.0
## 59           4.6         1.3
## 60           3.9         1.4
## 61           3.5         1.0
## 62           4.2         1.5
## 63           4.0         1.0
## 64           4.7         1.4
## 65           3.6         1.3
## 66           4.4         1.4
## 67           4.5         1.5
## 68           4.1         1.0
## 69           4.5         1.5
## 70           3.9         1.1
## 71           4.8         1.8
## 72           4.0         1.3
## 73           4.9         1.5
## 74           4.7         1.2
## 75           4.3         1.3
## 76           4.4         1.4
## 77           4.8         1.4
## 78           5.0         1.7
## 79           4.5         1.5
## 80           3.5         1.0
## 81           3.8         1.1
## 82           3.7         1.0
## 83           3.9         1.2
## 84           5.1         1.6
## 85           4.5         1.5
## 86           4.5         1.6
## 87           4.7         1.5
## 88           4.4         1.3
## 89           4.1         1.3
## 90           4.0         1.3
## 91           4.4         1.2
## 92           4.6         1.4
## 93           4.0         1.2
## 94           3.3         1.0
## 95           4.2         1.3
## 96           4.2         1.2
## 97           4.2         1.3
## 98           4.3         1.3
## 99           3.0         1.1
## 100          4.1         1.3
## 101          6.0         2.5
## 102          5.1         1.9
## 103          5.9         2.1
## 104          5.6         1.8
## 105          5.8         2.2
## 106          6.6         2.1
## 107          4.5         1.7
## 108          6.3         1.8
## 109          5.8         1.8
## 110          6.1         2.5
## 111          5.1         2.0
## 112          5.3         1.9
## 113          5.5         2.1
## 114          5.0         2.0
## 115          5.1         2.4
## 116          5.3         2.3
## 117          5.5         1.8
## 118          6.7         2.2
## 119          6.9         2.3
## 120          5.0         1.5
## 121          5.7         2.3
## 122          4.9         2.0
## 123          6.7         2.0
## 124          4.9         1.8
## 125          5.7         2.1
## 126          6.0         1.8
## 127          4.8         1.8
## 128          4.9         1.8
## 129          5.6         2.1
## 130          5.8         1.6
## 131          6.1         1.9
## 132          6.4         2.0
## 133          5.6         2.2
## 134          5.1         1.5
## 135          5.6         1.4
## 136          6.1         2.3
## 137          5.6         2.4
## 138          5.5         1.8
## 139          4.8         1.8
## 140          5.4         2.1
## 141          5.6         2.4
## 142          5.1         2.3
## 143          5.1         1.9
## 144          5.9         2.3
## 145          5.7         2.5
## 146          5.2         2.3
## 147          5.0         1.9
## 148          5.2         2.0
## 149          5.4         2.3
## 150          5.1         1.8
```

```
# Još osnovnih manipulacija stupcima:
irisdata[c(2,5,6), c("Sepal.Width","Species")]
```

```
##   Sepal.Width Species
## 2         3.0  setosa
## 5         3.6  setosa
## 6         3.9  setosa
```

```
irisdata[c(2,5,6), -c(5)]
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 2          4.9         3.0          1.4         0.2
## 5          5.0         3.6          1.4         0.2
## 6          5.4         3.9          1.7         0.4
```

```
# Izdvojiti sve redove gdje je Sepal Width veći od 3.3:
irisdata[irisdata$Sepal.Width > 3.3,]
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1            5.1         3.5          1.4         0.2     setosa
## 5            5.0         3.6          1.4         0.2     setosa
## 6            5.4         3.9          1.7         0.4     setosa
## 7            4.6         3.4          1.4         0.3     setosa
## 8            5.0         3.4          1.5         0.2     setosa
## 11           5.4         3.7          1.5         0.2     setosa
## 12           4.8         3.4          1.6         0.2     setosa
## 15           5.8         4.0          1.2         0.2     setosa
## 16           5.7         4.4          1.5         0.4     setosa
## 17           5.4         3.9          1.3         0.4     setosa
## 18           5.1         3.5          1.4         0.3     setosa
## 19           5.7         3.8          1.7         0.3     setosa
## 20           5.1         3.8          1.5         0.3     setosa
## 21           5.4         3.4          1.7         0.2     setosa
## 22           5.1         3.7          1.5         0.4     setosa
## 23           4.6         3.6          1.0         0.2     setosa
## 25           4.8         3.4          1.9         0.2     setosa
## 27           5.0         3.4          1.6         0.4     setosa
## 28           5.2         3.5          1.5         0.2     setosa
## 29           5.2         3.4          1.4         0.2     setosa
## 32           5.4         3.4          1.5         0.4     setosa
## 33           5.2         4.1          1.5         0.1     setosa
## 34           5.5         4.2          1.4         0.2     setosa
## 37           5.5         3.5          1.3         0.2     setosa
## 38           4.9         3.6          1.4         0.1     setosa
## 40           5.1         3.4          1.5         0.2     setosa
## 41           5.0         3.5          1.3         0.3     setosa
## 44           5.0         3.5          1.6         0.6     setosa
## 45           5.1         3.8          1.9         0.4     setosa
## 47           5.1         3.8          1.6         0.2     setosa
## 49           5.3         3.7          1.5         0.2     setosa
## 86           6.0         3.4          4.5         1.6 versicolor
## 110          7.2         3.6          6.1         2.5  virginica
## 118          7.7         3.8          6.7         2.2  virginica
## 132          7.9         3.8          6.4         2.0  virginica
## 137          6.3         3.4          5.6         2.4  virginica
## 149          6.2         3.4          5.4         2.3  virginica
```

```
# Izdvojiti sve stupce osim stupca Species:
irisdata[names(irisdata) != "Species"]
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1            5.1         3.5          1.4         0.2
## 2            4.9         3.0          1.4         0.2
## 3            4.7         3.2          1.3         0.2
## 4            4.6         3.1          1.5         0.2
## 5            5.0         3.6          1.4         0.2
## 6            5.4         3.9          1.7         0.4
## 7            4.6         3.4          1.4         0.3
## 8            5.0         3.4          1.5         0.2
## 9            4.4         2.9          1.4         0.2
## 10           4.9         3.1          1.5         0.1
## 11           5.4         3.7          1.5         0.2
## 12           4.8         3.4          1.6         0.2
## 13           4.8         3.0          1.4         0.1
## 14           4.3         3.0          1.1         0.1
## 15           5.8         4.0          1.2         0.2
## 16           5.7         4.4          1.5         0.4
## 17           5.4         3.9          1.3         0.4
## 18           5.1         3.5          1.4         0.3
## 19           5.7         3.8          1.7         0.3
## 20           5.1         3.8          1.5         0.3
## 21           5.4         3.4          1.7         0.2
## 22           5.1         3.7          1.5         0.4
## 23           4.6         3.6          1.0         0.2
## 24           5.1         3.3          1.7         0.5
## 25           4.8         3.4          1.9         0.2
## 26           5.0         3.0          1.6         0.2
## 27           5.0         3.4          1.6         0.4
## 28           5.2         3.5          1.5         0.2
## 29           5.2         3.4          1.4         0.2
## 30           4.7         3.2          1.6         0.2
## 31           4.8         3.1          1.6         0.2
## 32           5.4         3.4          1.5         0.4
## 33           5.2         4.1          1.5         0.1
## 34           5.5         4.2          1.4         0.2
## 35           4.9         3.1          1.5         0.2
## 36           5.0         3.2          1.2         0.2
## 37           5.5         3.5          1.3         0.2
## 38           4.9         3.6          1.4         0.1
## 39           4.4         3.0          1.3         0.2
## 40           5.1         3.4          1.5         0.2
## 41           5.0         3.5          1.3         0.3
## 42           4.5         2.3          1.3         0.3
## 43           4.4         3.2          1.3         0.2
## 44           5.0         3.5          1.6         0.6
## 45           5.1         3.8          1.9         0.4
## 46           4.8         3.0          1.4         0.3
## 47           5.1         3.8          1.6         0.2
## 48           4.6         3.2          1.4         0.2
## 49           5.3         3.7          1.5         0.2
## 50           5.0         3.3          1.4         0.2
## 51           7.0         3.2          4.7         1.4
## 52           6.4         3.2          4.5         1.5
## 53           6.9         3.1          4.9         1.5
## 54           5.5         2.3          4.0         1.3
## 55           6.5         2.8          4.6         1.5
## 56           5.7         2.8          4.5         1.3
## 57           6.3         3.3          4.7         1.6
## 58           4.9         2.4          3.3         1.0
## 59           6.6         2.9          4.6         1.3
## 60           5.2         2.7          3.9         1.4
## 61           5.0         2.0          3.5         1.0
## 62           5.9         3.0          4.2         1.5
## 63           6.0         2.2          4.0         1.0
## 64           6.1         2.9          4.7         1.4
## 65           5.6         2.9          3.6         1.3
## 66           6.7         3.1          4.4         1.4
## 67           5.6         3.0          4.5         1.5
## 68           5.8         2.7          4.1         1.0
## 69           6.2         2.2          4.5         1.5
## 70           5.6         2.5          3.9         1.1
## 71           5.9         3.2          4.8         1.8
## 72           6.1         2.8          4.0         1.3
## 73           6.3         2.5          4.9         1.5
## 74           6.1         2.8          4.7         1.2
## 75           6.4         2.9          4.3         1.3
## 76           6.6         3.0          4.4         1.4
## 77           6.8         2.8          4.8         1.4
## 78           6.7         3.0          5.0         1.7
## 79           6.0         2.9          4.5         1.5
## 80           5.7         2.6          3.5         1.0
## 81           5.5         2.4          3.8         1.1
## 82           5.5         2.4          3.7         1.0
## 83           5.8         2.7          3.9         1.2
## 84           6.0         2.7          5.1         1.6
## 85           5.4         3.0          4.5         1.5
## 86           6.0         3.4          4.5         1.6
## 87           6.7         3.1          4.7         1.5
## 88           6.3         2.3          4.4         1.3
## 89           5.6         3.0          4.1         1.3
## 90           5.5         2.5          4.0         1.3
## 91           5.5         2.6          4.4         1.2
## 92           6.1         3.0          4.6         1.4
## 93           5.8         2.6          4.0         1.2
## 94           5.0         2.3          3.3         1.0
## 95           5.6         2.7          4.2         1.3
## 96           5.7         3.0          4.2         1.2
## 97           5.7         2.9          4.2         1.3
## 98           6.2         2.9          4.3         1.3
## 99           5.1         2.5          3.0         1.1
## 100          5.7         2.8          4.1         1.3
## 101          6.3         3.3          6.0         2.5
## 102          5.8         2.7          5.1         1.9
## 103          7.1         3.0          5.9         2.1
## 104          6.3         2.9          5.6         1.8
## 105          6.5         3.0          5.8         2.2
## 106          7.6         3.0          6.6         2.1
## 107          4.9         2.5          4.5         1.7
## 108          7.3         2.9          6.3         1.8
## 109          6.7         2.5          5.8         1.8
## 110          7.2         3.6          6.1         2.5
## 111          6.5         3.2          5.1         2.0
## 112          6.4         2.7          5.3         1.9
## 113          6.8         3.0          5.5         2.1
## 114          5.7         2.5          5.0         2.0
## 115          5.8         2.8          5.1         2.4
## 116          6.4         3.2          5.3         2.3
## 117          6.5         3.0          5.5         1.8
## 118          7.7         3.8          6.7         2.2
## 119          7.7         2.6          6.9         2.3
## 120          6.0         2.2          5.0         1.5
## 121          6.9         3.2          5.7         2.3
## 122          5.6         2.8          4.9         2.0
## 123          7.7         2.8          6.7         2.0
## 124          6.3         2.7          4.9         1.8
## 125          6.7         3.3          5.7         2.1
## 126          7.2         3.2          6.0         1.8
## 127          6.2         2.8          4.8         1.8
## 128          6.1         3.0          4.9         1.8
## 129          6.4         2.8          5.6         2.1
## 130          7.2         3.0          5.8         1.6
## 131          7.4         2.8          6.1         1.9
## 132          7.9         3.8          6.4         2.0
## 133          6.4         2.8          5.6         2.2
## 134          6.3         2.8          5.1         1.5
## 135          6.1         2.6          5.6         1.4
## 136          7.7         3.0          6.1         2.3
## 137          6.3         3.4          5.6         2.4
## 138          6.4         3.1          5.5         1.8
## 139          6.0         3.0          4.8         1.8
## 140          6.9         3.1          5.4         2.1
## 141          6.7         3.1          5.6         2.4
## 142          6.9         3.1          5.1         2.3
## 143          5.8         2.7          5.1         1.9
## 144          6.8         3.2          5.9         2.3
## 145          6.7         3.3          5.7         2.5
## 146          6.7         3.0          5.2         2.3
## 147          6.3         2.5          5.0         1.9
## 148          6.5         3.0          5.2         2.0
## 149          6.2         3.4          5.4         2.3
## 150          5.9         3.0          5.1         1.8
```

```
# Izdvojiti sve stupce koji opisuju duljinu:
irisdata[names(irisdata) %in% c("Sepal.Length","Petal.Length")]
```

```
##     Sepal.Length Petal.Length
## 1            5.1          1.4
## 2            4.9          1.4
## 3            4.7          1.3
## 4            4.6          1.5
## 5            5.0          1.4
## 6            5.4          1.7
## 7            4.6          1.4
## 8            5.0          1.5
## 9            4.4          1.4
## 10           4.9          1.5
## 11           5.4          1.5
## 12           4.8          1.6
## 13           4.8          1.4
## 14           4.3          1.1
## 15           5.8          1.2
## 16           5.7          1.5
## 17           5.4          1.3
## 18           5.1          1.4
## 19           5.7          1.7
## 20           5.1          1.5
## 21           5.4          1.7
## 22           5.1          1.5
## 23           4.6          1.0
## 24           5.1          1.7
## 25           4.8          1.9
## 26           5.0          1.6
## 27           5.0          1.6
## 28           5.2          1.5
## 29           5.2          1.4
## 30           4.7          1.6
## 31           4.8          1.6
## 32           5.4          1.5
## 33           5.2          1.5
## 34           5.5          1.4
## 35           4.9          1.5
## 36           5.0          1.2
## 37           5.5          1.3
## 38           4.9          1.4
## 39           4.4          1.3
## 40           5.1          1.5
## 41           5.0          1.3
## 42           4.5          1.3
## 43           4.4          1.3
## 44           5.0          1.6
## 45           5.1          1.9
## 46           4.8          1.4
## 47           5.1          1.6
## 48           4.6          1.4
## 49           5.3          1.5
## 50           5.0          1.4
## 51           7.0          4.7
## 52           6.4          4.5
## 53           6.9          4.9
## 54           5.5          4.0
## 55           6.5          4.6
## 56           5.7          4.5
## 57           6.3          4.7
## 58           4.9          3.3
## 59           6.6          4.6
## 60           5.2          3.9
## 61           5.0          3.5
## 62           5.9          4.2
## 63           6.0          4.0
## 64           6.1          4.7
## 65           5.6          3.6
## 66           6.7          4.4
## 67           5.6          4.5
## 68           5.8          4.1
## 69           6.2          4.5
## 70           5.6          3.9
## 71           5.9          4.8
## 72           6.1          4.0
## 73           6.3          4.9
## 74           6.1          4.7
## 75           6.4          4.3
## 76           6.6          4.4
## 77           6.8          4.8
## 78           6.7          5.0
## 79           6.0          4.5
## 80           5.7          3.5
## 81           5.5          3.8
## 82           5.5          3.7
## 83           5.8          3.9
## 84           6.0          5.1
## 85           5.4          4.5
## 86           6.0          4.5
## 87           6.7          4.7
## 88           6.3          4.4
## 89           5.6          4.1
## 90           5.5          4.0
## 91           5.5          4.4
## 92           6.1          4.6
## 93           5.8          4.0
## 94           5.0          3.3
## 95           5.6          4.2
## 96           5.7          4.2
## 97           5.7          4.2
## 98           6.2          4.3
## 99           5.1          3.0
## 100          5.7          4.1
## 101          6.3          6.0
## 102          5.8          5.1
## 103          7.1          5.9
## 104          6.3          5.6
## 105          6.5          5.8
## 106          7.6          6.6
## 107          4.9          4.5
## 108          7.3          6.3
## 109          6.7          5.8
## 110          7.2          6.1
## 111          6.5          5.1
## 112          6.4          5.3
## 113          6.8          5.5
## 114          5.7          5.0
## 115          5.8          5.1
## 116          6.4          5.3
## 117          6.5          5.5
## 118          7.7          6.7
## 119          7.7          6.9
## 120          6.0          5.0
## 121          6.9          5.7
## 122          5.6          4.9
## 123          7.7          6.7
## 124          6.3          4.9
## 125          6.7          5.7
## 126          7.2          6.0
## 127          6.2          4.8
## 128          6.1          4.9
## 129          6.4          5.6
## 130          7.2          5.8
## 131          7.4          6.1
## 132          7.9          6.4
## 133          6.4          5.6
## 134          6.3          5.1
## 135          6.1          5.6
## 136          7.7          6.1
## 137          6.3          5.6
## 138          6.4          5.5
## 139          6.0          4.8
## 140          6.9          5.4
## 141          6.7          5.6
## 142          6.9          5.1
## 143          5.8          5.1
## 144          6.8          5.9
## 145          6.7          5.7
## 146          6.7          5.2
## 147          6.3          5.0
## 148          6.5          5.2
## 149          6.2          5.4
## 150          5.9          5.1
```

```
# Ispisati sve Sepal Width za koji je Petal Length veci od 1.4:
irisdata$Sepal.Width[irisdata$Petal.Length > 1.4]
```

```
##   [1] 3.1 3.9 3.4 3.1 3.7 3.4 4.4 3.8 3.8 3.4 3.7 3.3 3.4 3.0 3.4 3.5 3.2 3.1
##  [19] 3.4 4.1 3.1 3.4 3.5 3.8 3.8 3.7 3.2 3.2 3.1 2.3 2.8 2.8 3.3 2.4 2.9 2.7
##  [37] 2.0 3.0 2.2 2.9 2.9 3.1 3.0 2.7 2.2 2.5 3.2 2.8 2.5 2.8 2.9 3.0 2.8 3.0
##  [55] 2.9 2.6 2.4 2.4 2.7 2.7 3.0 3.4 3.1 2.3 3.0 2.5 2.6 3.0 2.6 2.3 2.7 3.0
##  [73] 2.9 2.9 2.5 2.8 3.3 2.7 3.0 2.9 3.0 3.0 2.5 2.9 2.5 3.6 3.2 2.7 3.0 2.5
##  [91] 2.8 3.2 3.0 3.8 2.6 2.2 3.2 2.8 2.8 2.7 3.3 3.2 2.8 3.0 2.8 3.0 2.8 3.8
## [109] 2.8 2.8 2.6 3.0 3.4 3.1 3.0 3.1 3.1 3.1 2.7 3.2 3.3 3.0 2.5 3.0 3.4 3.0
```

```
# ili:
irisdata[irisdata$Petal.Length > 1.4,]$Sepal.Width
```

```
##   [1] 3.1 3.9 3.4 3.1 3.7 3.4 4.4 3.8 3.8 3.4 3.7 3.3 3.4 3.0 3.4 3.5 3.2 3.1
##  [19] 3.4 4.1 3.1 3.4 3.5 3.8 3.8 3.7 3.2 3.2 3.1 2.3 2.8 2.8 3.3 2.4 2.9 2.7
##  [37] 2.0 3.0 2.2 2.9 2.9 3.1 3.0 2.7 2.2 2.5 3.2 2.8 2.5 2.8 2.9 3.0 2.8 3.0
##  [55] 2.9 2.6 2.4 2.4 2.7 2.7 3.0 3.4 3.1 2.3 3.0 2.5 2.6 3.0 2.6 2.3 2.7 3.0
##  [73] 2.9 2.9 2.5 2.8 3.3 2.7 3.0 2.9 3.0 3.0 2.5 2.9 2.5 3.6 3.2 2.7 3.0 2.5
##  [91] 2.8 3.2 3.0 3.8 2.6 2.2 3.2 2.8 2.8 2.7 3.3 3.2 2.8 3.0 2.8 3.0 2.8 3.8
## [109] 2.8 2.8 2.6 3.0 3.4 3.1 3.0 3.1 3.1 3.1 2.7 3.2 3.3 3.0 2.5 3.0 3.4 3.0
```

```
# Izdvojiti sve pozicije (indekse) za koje vrijedi uvjet Sepal.Width > 3.3 i ispisati sve pripadne prim
ind = which(irisdata$Sepal.Width > 3.3)
irisdata$Sepal.Width > 3.3
```

```
##   [1]  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
##  [13] FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
##  [25]  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE
##  [37]  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE
##  [49]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [85] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [109] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## [133] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [145] FALSE FALSE FALSE FALSE  TRUE FALSE
```

```
ind
```

```
##  [1]   1   5   6   7   8  11  12  15  16  17  18  19  20  21  22  23  25  27  28
## [20]  29  32  33  34  37  38  40  41  44  45  47  49  86 110 118 132 137 149
```

```
irisdata[ind,]
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1            5.1         3.5          1.4         0.2     setosa
## 5            5.0         3.6          1.4         0.2     setosa
## 6            5.4         3.9          1.7         0.4     setosa
## 7            4.6         3.4          1.4         0.3     setosa
## 8            5.0         3.4          1.5         0.2     setosa
## 11           5.4         3.7          1.5         0.2     setosa
## 12           4.8        3.4         1.6        0.2     setosa
## 15           5.8        4.0         1.2        0.2     setosa
## 16           5.7        4.4         1.5        0.4     setosa
## 17           5.4        3.9         1.3        0.4     setosa
## 18           5.1        3.5         1.4        0.3     setosa
## 19           5.7        3.8         1.7        0.3     setosa
## 20           5.1        3.8         1.5        0.3     setosa
## 21           5.4        3.4         1.7        0.2     setosa
## 22           5.1        3.7         1.5        0.4     setosa
## 23           4.6        3.6         1.0        0.2     setosa
## 25           4.8        3.4         1.9        0.2     setosa
## 27           5.0        3.4         1.6        0.4     setosa
## 28           5.2        3.5         1.5        0.2     setosa
## 29           5.2        3.4         1.4        0.2     setosa
## 32           5.4        3.4         1.5        0.4     setosa
## 33           5.2        4.1         1.5        0.1     setosa
## 34           5.5        4.2         1.4        0.2     setosa
## 37           5.5        3.5         1.3        0.2     setosa
## 38           4.9        3.6         1.4        0.1     setosa
## 40           5.1        3.4         1.5        0.2     setosa
## 41           5.0        3.5         1.3        0.3     setosa
## 44           5.0        3.5         1.6        0.6     setosa
## 45           5.1        3.8         1.9        0.4     setosa
## 47           5.1        3.8         1.6        0.2     setosa
## 49           5.3        3.7         1.5        0.2     setosa
## 86           6.0        3.4         4.5        1.6 versicolor
## 110          7.2        3.6         6.1        2.5  virginica
## 118          7.7        3.8         6.7        2.2  virginica
## 132          7.9        3.8         6.4        2.0  virginica
## 137          6.3        3.4         5.6        2.4  virginica
## 149          6.2        3.4         5.4        2.3  virginica
```

Mjere centralne tendencije

Mjere centralne tendencije (ili središnje mjere) opisuju skup podataka jednom vrijednošću oko koje se podatci grupiraju. Najčešće korištene mjere centralne tendencije su: aritmetička sredina, medijan, mod i podrezana aritmetička sredina.

```r
# Aritmeticka sredina - mean
mean(irisdata$Petal.Length)
```

```
## [1] 3.758
```

```r
# Podrezana aritmeticka sredina s uklanjanjem po 20% najmanjih i najvecih podataka
mean(irisdata$Petal.Length, trim=0.2)
```

```
## [1] 3.842222
```

```r
# Medijan - robusna mjera centralne tendencije(točno 50% podataka je manje i 50% podataka veće od te vr
median(irisdata$Petal.Length)
```

```
## [1] 4.35
```

```
# 1., 2. i 3. kvartil
quantile(irisdata$Petal.Length, probs = c(0.25,0.5,0.75)) # Koji kvartil je ujedno i medijan?
```

```
##  25%  50%  75%
## 1.60 4.35 5.10
```

```
# Mod (most frequent value) - vrijednost koja se najčešće pojavljuje u podatcima. Kada ova mjera ima sm
require(modeest)
```

```
## Loading required package: modeest
```

```
mfv(irisdata$Petal.Length)
```

```
## [1] 1.4 1.5
```

Mjere rasipanja

Mjere rasipanja opisuju varijabilnost podataka, koliko su podatci koncentrirani ili rašireni. Najčešće korištene mjere su: rang, interkvartilni rang, varijanca, standardna devijacija i koeficijent varijacije.

```
# Rang- razlika između najvećeg i najmanjeg iznosa u podatcima
max(irisdata$Petal.Length)-min(irisdata$Petal.Length)
```

```
## [1] 5.9
```

```
# Interkvartilni rang - razlika trećeg i prvog kvartila podataka --> Zašto je ovo robusnija mjera od pr
IQR(irisdata$Petal.Length)
```

```
## [1] 3.5
```

```
# Varijanca i standardna devijacija - najčešće korištene mjere rasipanja
var(irisdata$Petal.Length)
```

```
## [1] 3.116278
```

```
sd(irisdata$Petal.Length)
```

```
## [1] 1.765298
```

```
sqrt(var(irisdata$Petal.Length))
```

```
## [1] 1.765298
```

```
# Računa li var() nepristranu procjenu varijance?
help(var)

# Koeficijent varijacije -  relativna mjera rasipanja koja opisuje rasipanje podataka u odnosu na njiho
#suppressWarnings(require(raster,quietly = TRUE))
#cv(irisdata$Petal.Length)
sd(irisdata$Petal.Length)/mean(irisdata$Petal.Length)
```

```
## [1] 0.4697441
```

Osnovna deskriptivna statistika i (napredna) manipulacija podataka

```
# Osnovna deskriptivna statistika:
summary(irisdata)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
##        Species  
##  setosa    :50  
##  versicolor:50  
##  virginica :50  
##                 
##                 
## 
```

Izračunajmo srednje vrijednosti i medijane svih mjera irisa, zasebno za svaku vrstu irisa u podatcima. Koliko se razlikuju srednje vrijednosti i medijani za svaku vrstu i što to govori o obliku distribucije tih mjera?

Izračunajmo potom robusniju procjenu računajući podrezanu aritmetičku sredinu s uklanjanjem 10% najvećih i najmanjih vrijednosti.

```
# tidyverse - vrlo koristan skup biblioteka koji omogućuje jos elegantniju manipulaciju data frame-ovim
# https://www.tidyverse.org/packages/

library(tidyverse)
```

```
## Registered S3 method overwritten by 'httr':
##   method         from  
##   print.response rmutil

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5 
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0 
## v readr   2.1.3      v forcats 0.5.2 
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)

irisdata %>% group_by(Species) %>% summarise(
          Mean.Sep.Len = mean(Sepal.Length),
          Mean.Pet.Len = mean(Petal.Length),
          Mean.Sep.Wid = mean(Sepal.Width),
          Mean.Pet.Wid = mean(Petal.Width)
            ) -> summary.result1
summary.result1
```

```
## # A tibble: 3 x 5
##   Species    Mean.Sep.Len Mean.Pet.Len Mean.Sep.Wid Mean.Pet.Wid
##   <fct>             <dbl>        <dbl>        <dbl>        <dbl>
## 1 setosa             5.01         1.46         3.43        0.246
## 2 versicolor         5.94         4.26         2.77        1.33
## 3 virginica          6.59         5.55         2.97        2.03
```

```
irisdata %>% group_by(Species) %>% summarise(
          Med.Sep.Len = median(Sepal.Length),
          Med.Pet.Len = median(Petal.Length),
          Med.Sep.Wid = median(Sepal.Width),
          Med.Pet.Wid = median(Petal.Width)
            ) -> summary.result2
summary.result2
```

```
## # A tibble: 3 x 5
##   Species    Med.Sep.Len Med.Pet.Len Med.Sep.Wid Med.Pet.Wid
##   <fct>            <dbl>       <dbl>       <dbl>       <dbl>
## 1 setosa             5          1.5         3.4         0.2
## 2 versicolor         5.9        4.35        2.8         1.3
## 3 virginica          6.5        5.55        3           2
```

```
# Podrezana srednja vrijednost - zašto je ovo robusnija metoda u odnosu na običnu srednju vrijednost?
irisdata %>% group_by(Species) %>% summarise(
          MeanTr.Sep.Len = mean(Sepal.Length, trim = 0.1),
          MeanTr.Pet.Len = mean(Petal.Length, trim = 0.1),
          MeanTr.Sep.Wid = mean(Sepal.Width, trim = 0.1),
          MeanTr.Pet.Wid = mean(Petal.Width, trim = 0.1)
            ) -> summary.result3
summary.result3
```

```
## # A tibble: 3 x 5
##   Species    MeanTr.Sep.Len MeanTr.Pet.Len MeanTr.Sep.Wid MeanTr.Pet.Wid
##   <fct>               <dbl>          <dbl>          <dbl>          <dbl>
## 1 setosa               5.00           1.46           3.42          0.238
## 2 versicolor           5.94           4.29           2.78          1.32
## 3 virginica            6.57           5.51           2.96          2.03
```

```
# Do sada smo računali mjere centralne tendencije za sve vrste zajedno - ali vidimo da kad ih razdvojim
# Usporedite razliku između medijana i meana za petal length izračunatih za sve vrste zajedno, potom iz
```

```
# Pomoću summary-ja statistike za pojedinu vrstu:
summary(irisdata[irisdata["Species"] == c("setosa"),])
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100  
##  1st Qu.:4.800   1st Qu.:3.200   1st Qu.:1.400   1st Qu.:0.200  
##  Median :5.000   Median :3.400   Median :1.500   Median :0.200  
##  Mean   :5.006   Mean   :3.428   Mean   :1.462   Mean   :0.246  
##  3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300  
##  Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600  
##        Species  
##  setosa    :50  
##  versicolor: 0  
##  virginica : 0  
##                 
##                 
##                 
```

```r
summary(irisdata[irisdata["Species"] == c("versicolor"),])
```

```
##   Sepal.Length    Sepal.Width     Petal.Length   Petal.Width          Species  
##  Min.   :4.900   Min.   :2.000   Min.   :3.00   Min.   :1.000   setosa    : 0  
##  1st Qu.:5.600   1st Qu.:2.525   1st Qu.:4.00   1st Qu.:1.200   versicolor:50  
##  Median :5.900   Median :2.800   Median :4.35   Median :1.300   virginica : 0  
##  Mean   :5.936   Mean   :2.770   Mean   :4.26   Mean   :1.326                  
##  3rd Qu.:6.300   3rd Qu.:3.000   3rd Qu.:4.60   3rd Qu.:1.500                  
##  Max.   :7.000   Max.   :3.400   Max.   :5.10   Max.   :1.800                  
```

```r
summary(irisdata[irisdata["Species"] == c("virginica"),])
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.900   Min.   :2.200   Min.   :4.500   Min.   :1.400  
##  1st Qu.:6.225   1st Qu.:2.800   1st Qu.:5.100   1st Qu.:1.800  
##  Median :6.500   Median :3.000   Median :5.550   Median :2.000  
##  Mean   :6.588   Mean   :2.974   Mean   :5.552   Mean   :2.026  
##  3rd Qu.:6.900   3rd Qu.:3.175   3rd Qu.:5.875   3rd Qu.:2.300  
##  Max.   :7.900   Max.   :3.800   Max.   :6.900   Max.   :2.500  
##        Species  
##  setosa    : 0  
##  versicolor: 0  
##  virginica :50  
##                 
##                 
##                 
```

```r
# Još jedan način:
aggregate(irisdata[names(irisdata) != "Species"], list(irisdata$Species), mean)
```

```
##      Group.1 Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     setosa        5.006       3.428        1.462       0.246
## 2 versicolor        5.936       2.770        4.260       1.326
## 3  virginica        6.588       2.974        5.552       2.026
```

```r
aggregate(irisdata[names(irisdata) != "Species"], list(irisdata$Species), median)
```

```
##      Group.1 Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     setosa          5.0         3.4         1.50         0.2
## 2 versicolor          5.9         2.8         4.35         1.3
## 3  virginica          6.5         3.0         5.55         2.0
```

```
aggregate(irisdata[names(irisdata) != "Species"], list(irisdata$Species), mean,trim=0.1)
```

```
##      Group.1 Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     setosa       5.0025      3.4150       1.4600      0.2375
## 2 versicolor       5.9375      2.7800       4.2925      1.3250
## 3  virginica       6.5725      2.9625       5.5100      2.0325
```

```
# Prednost tidyverse-a?
```

Kada ima smisla (za kakve podatke) koristiti ovu deskriptivnu statistiku? Recimo da je dan rastući vremenski niz. Biste li primijenili mjere centralne tendencije na takav dataset?

Ponovno se vraćamo na značajnost konteksta, interpretacije podataka!

Izračunajmo sada interkvartilni rang (IQR) i standardnu devijaciju svih mjera za svaku od vrsta irisa zasebno.

```
irisdata %>% group_by(Species) %>% summarise(
          IQR.Sep.Len = IQR(Sepal.Length),
          IQR.Pet.Len = IQR(Petal.Length),
          IQR.Sep.Wid = IQR(Sepal.Width),
          IQR.Pet.Wid = IQR(Petal.Width)
            ) -> summary.result
summary.result
```

```
## # A tibble: 3 x 5
##   Species    IQR.Sep.Len IQR.Pet.Len IQR.Sep.Wid IQR.Pet.Wid
##   <fct>            <dbl>       <dbl>       <dbl>       <dbl>
## 1 setosa           0.400       0.175       0.475         0.1
## 2 versicolor       0.7         0.600       0.475         0.3
## 3 virginica        0.675       0.775       0.375         0.5
```

```
irisdata %>% group_by(Species) %>% summarise(
          sd.Sep.Len = sd(Sepal.Length),
          sd.Pet.Len = sd(Petal.Length),
          sd.Sep.Wid = sd(Sepal.Width),
          sd.Pet.Wid = sd(Petal.Width)
            ) -> summary.result
summary.result
```

```
## # A tibble: 3 x 5
##   Species    sd.Sep.Len sd.Pet.Len sd.Sep.Wid sd.Pet.Wid
##   <fct>           <dbl>      <dbl>      <dbl>      <dbl>
## 1 setosa          0.352      0.174      0.379      0.105
## 2 versicolor      0.516      0.470      0.314      0.198
## 3 virginica       0.636      0.552      0.322      0.275
```

```
# Usporedite opet razliku ovih mjera za petal length izračunatih za sve vrste zajedno, potom izračunati
# Možemo li iz ovih statistika zaključiti nešto o varijabilnosti i raspršenosti sepal length-a različit


# Kada je koja od ovih mjera rasipanja primjenjivija? Koja je primjenjivija za iris dataset?
```

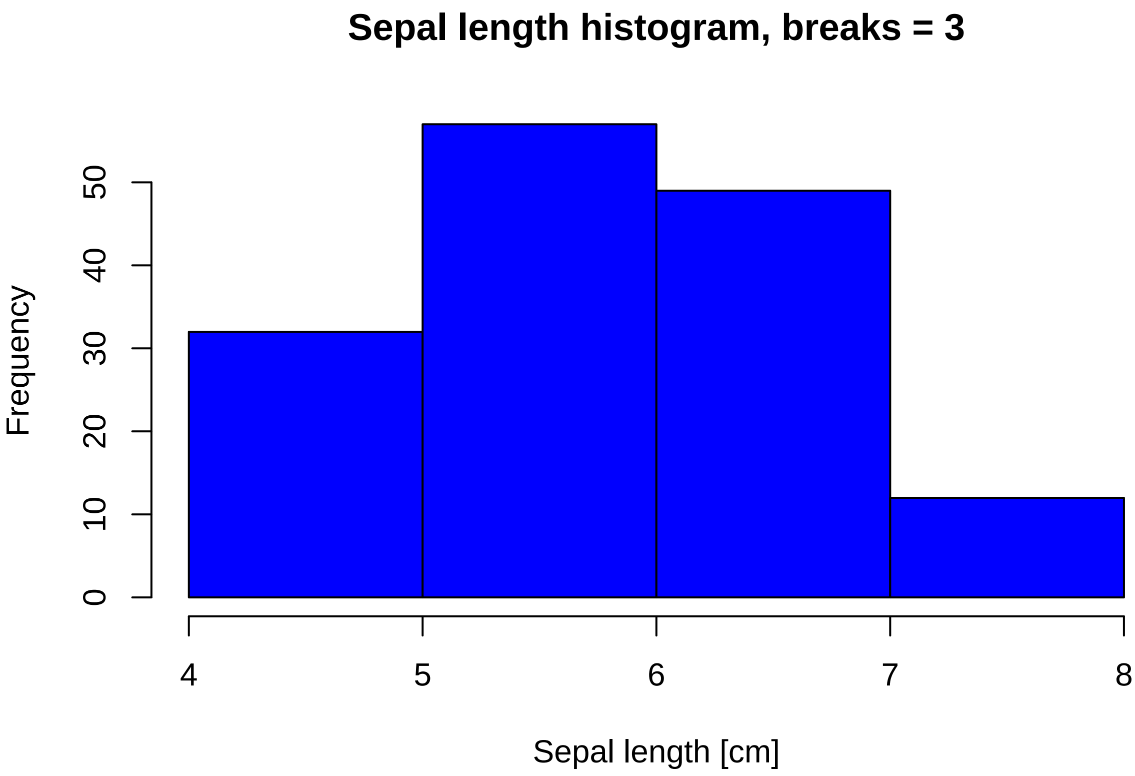
Vizualizacija podataka

Opet uvelike ovisi o kontekstu podataka, a neki od osnovnih načina vizualizacije podataka su:

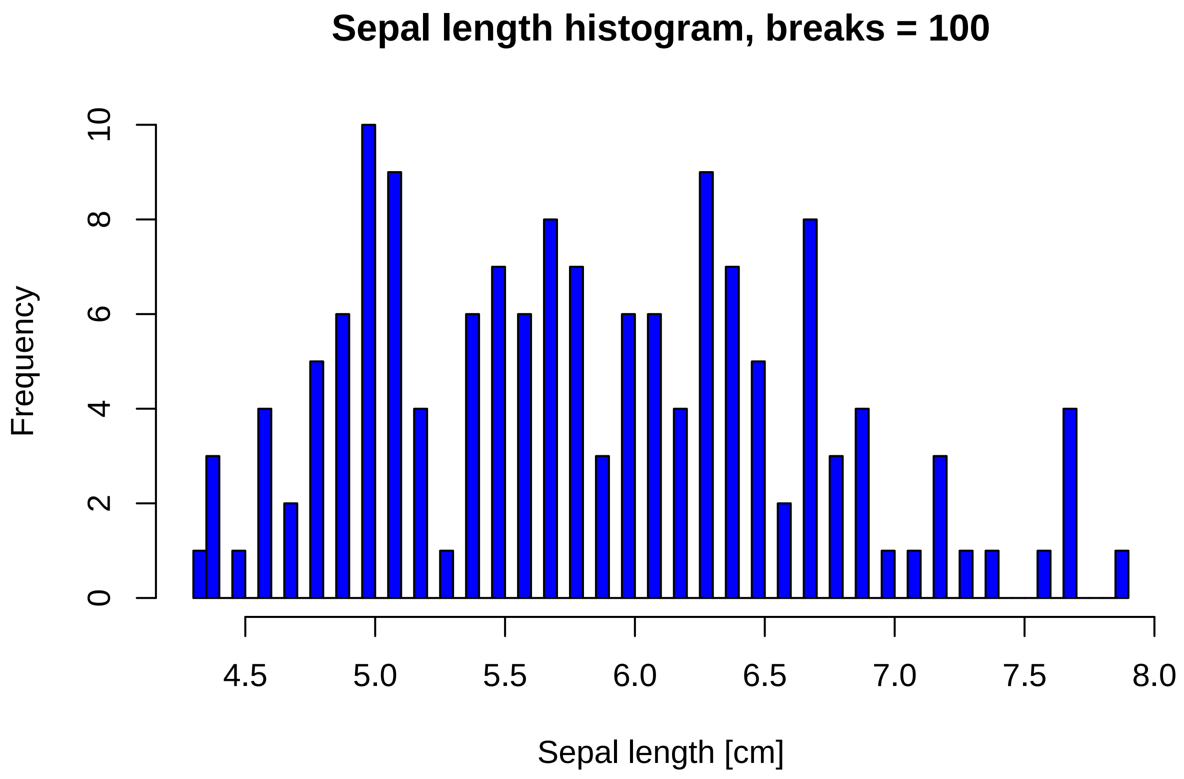
- Histogram - pokazuje oblik distribucije i gustoću podataka, a zasnovan je na grupiranju varijabli u razrede
- Pravokutni dijagram (box plot) - kombinira prikaz medijana, kvartila podataka, te najmanje i najveće vrijednosti. Pravokutni dijagram prikazuje i stršeće vrijednosti, koje se standardno definiraju kao podatci koji su iznad $Q_3 + 1.5 \cdot IQR$ ili ispod $Q_1 - 1.5 \cdot IQR$.
- Dijagram raspršenja (scatter plot) - jedan je od najvažnijih načina prikaza bivarijantnih podataka, te daje informaciju o povezanosti varijabli

Zanima nas kako je distribuirana duljina lapa - prikažimo histogramom. Kako izabrati broj razreda? Koje su granice razreda? Je li bolje prikazati podatke agregirano ili grupirano? Želimo li prikazati frekvencije ili relativne frekvencije (td. je površina histograma = 1)?

```
h = hist(irisdata$Sepal.Length,
         breaks=3,
         main="Sepal length histogram, breaks = 3",
         xlab="Sepal length [cm]",
         ylab='Frequency',
         col="blue"
         )
```

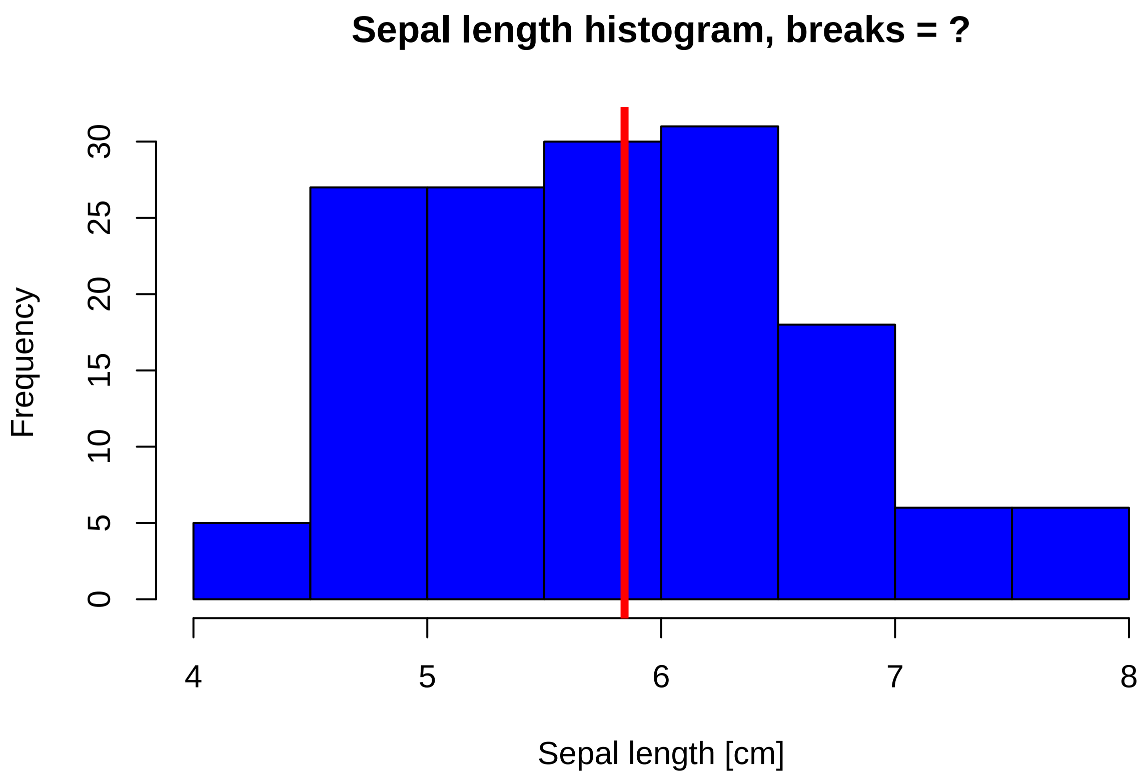


```
h = hist(irisdata$Sepal.Length,
         breaks=100,
         main="Sepal length histogram, breaks = 100",
         xlab="Sepal length [cm]",
         ylab='Frequency',
         col="blue"
         )
```



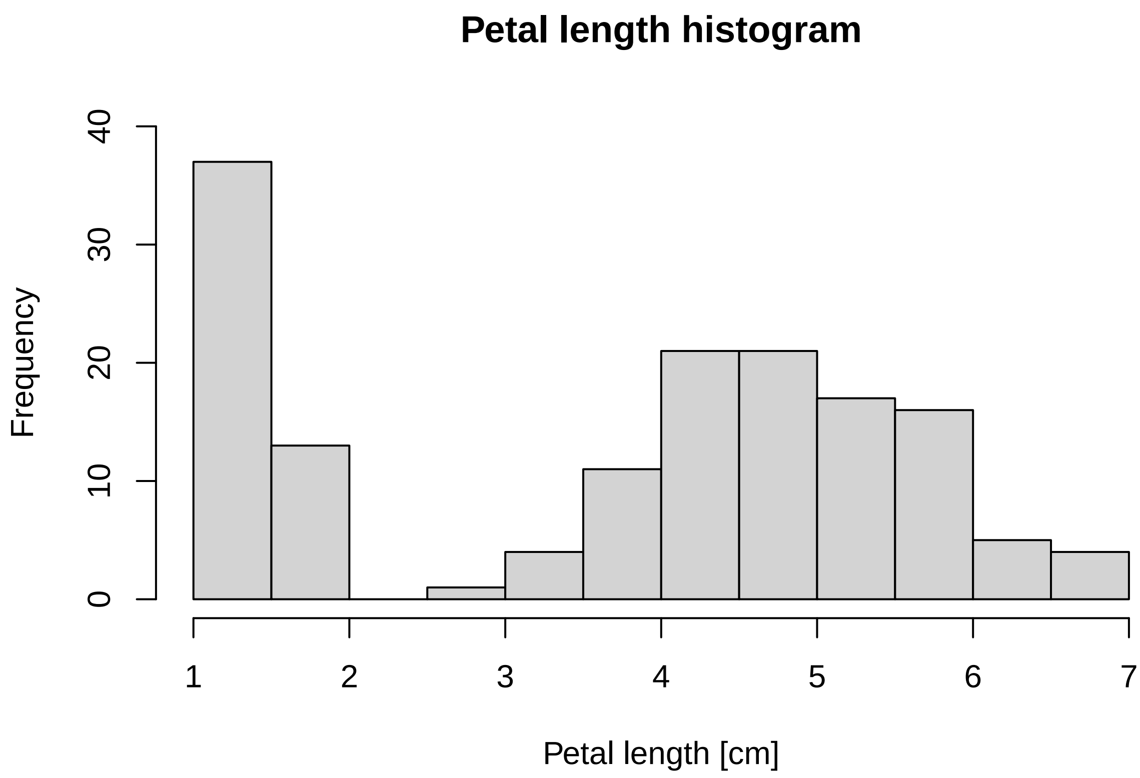
```
h = hist(irisdata$Sepal.Length,
         main="Sepal length histogram, breaks = ?",
         xlab="Sepal length [cm]",
         ylab='Frequency',
         col="blue"
         )

abline(v = mean(irisdata$Sepal.Length), col = "red", lwd = 4)
```



```
# Možemo li iz ovog histograma isčitati da se radi o multimodalnoj distribuciji?
```

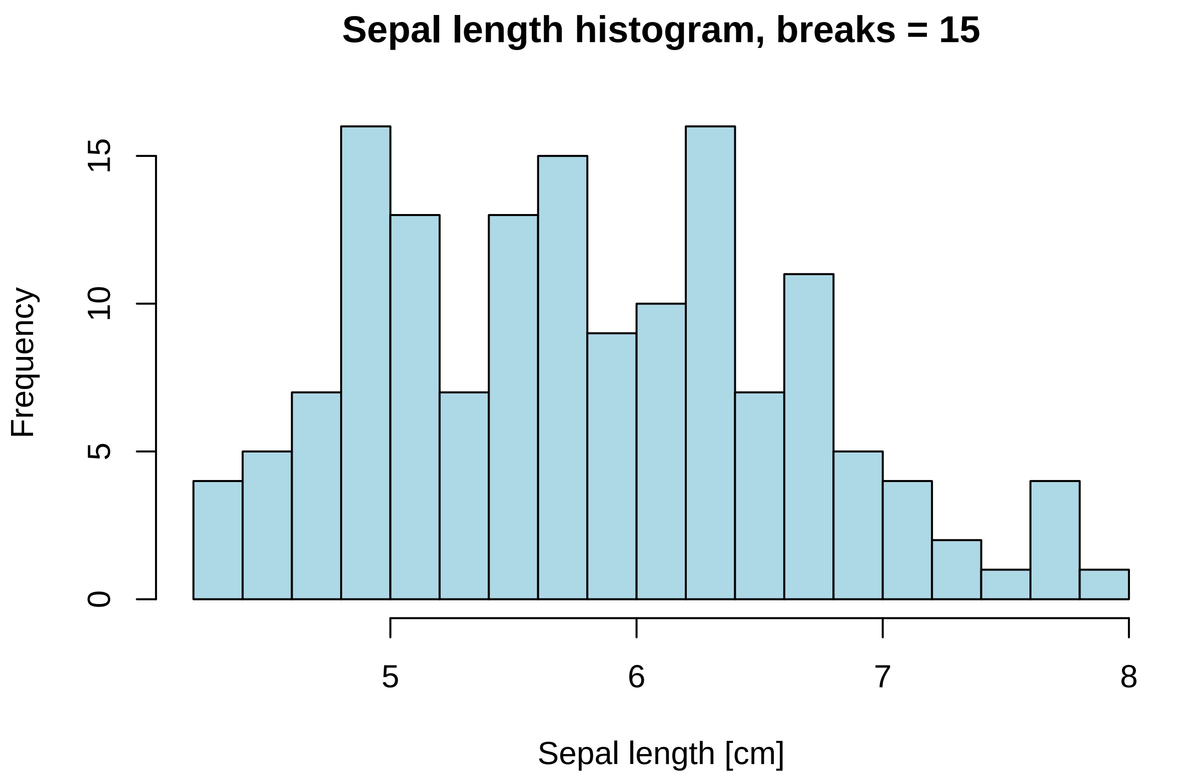
```
h = hist(irisdata$Petal.Length,
         main="Petal length histogram",
         xlab="Petal length [cm]",
         ylab='Frequency',
         ylim= c(0,40)
         )
```



```
mfv(irisdata$Petal.Length)
```

```
## [1] 1.4 1.5
```

```
h = hist(irisdata$Sepal.Length,
         breaks=15,
         main="Sepal length histogram, breaks = 15",
         xlab="Sepal length [cm]",
         ylab='Frequency',
         col="lightblue"
         )
```



```
# Histogram duljine lapa s cca. 15 razreda (broj razreda shvaca kao "sugestiju")
# Kako možemo doći do breakpoint-ova:
h$breaks
```

```
##  [1] 4.2 4.4 4.6 4.8 5.0 5.2 5.4 5.6 5.8 6.0 6.2 6.4 6.6 6.8 7.0 7.2 7.4 7.6 7.8
## [20] 8.0
```

```
length(h$breaks)
```

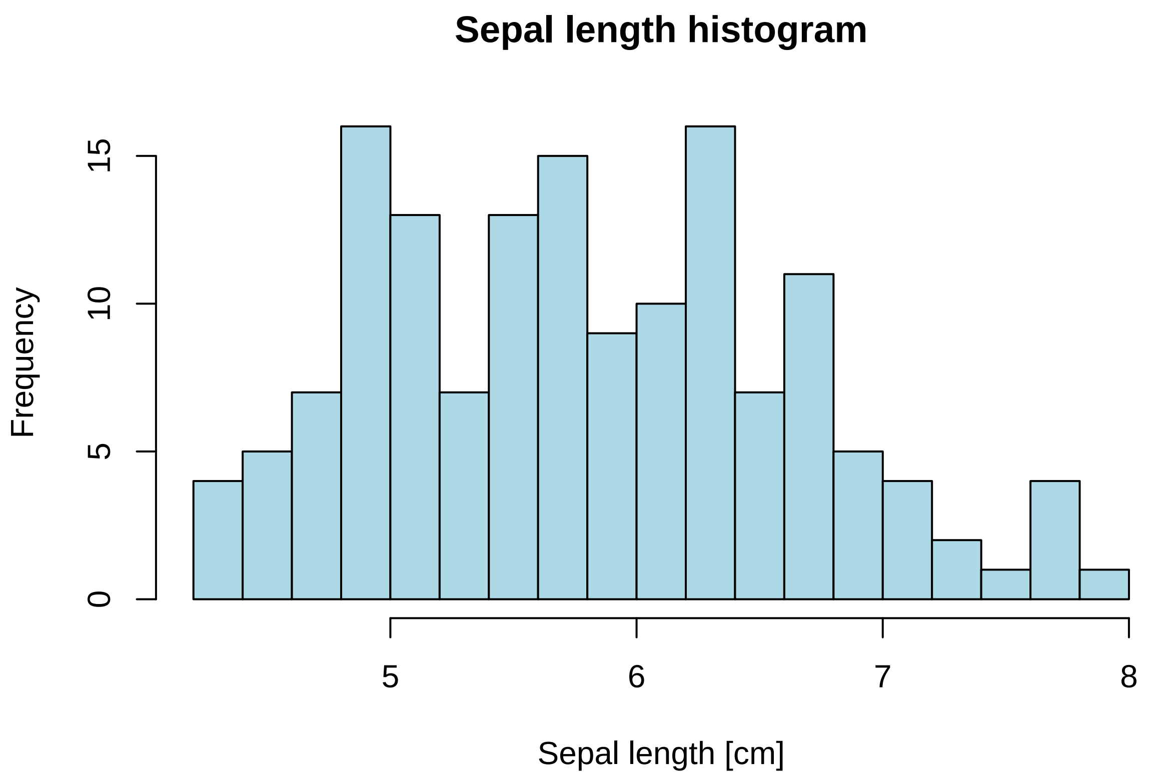
```
## [1] 20
```

```
# Ako želimo dati točan broj razreda, moramo definirati točke breakpoint-ova

b = seq(min(irisdata$Sepal.Length) - 0.1,max(irisdata$Sepal.Length) + 0.1,0.2)
length(b)
```

```
## [1] 20
```

```
h = hist(irisdata$Sepal.Length,
         breaks=b,
         main="Sepal length histogram",
         xlab="Sepal length [cm]",
         ylab='Frequency',
         col="lightblue"
         )
```



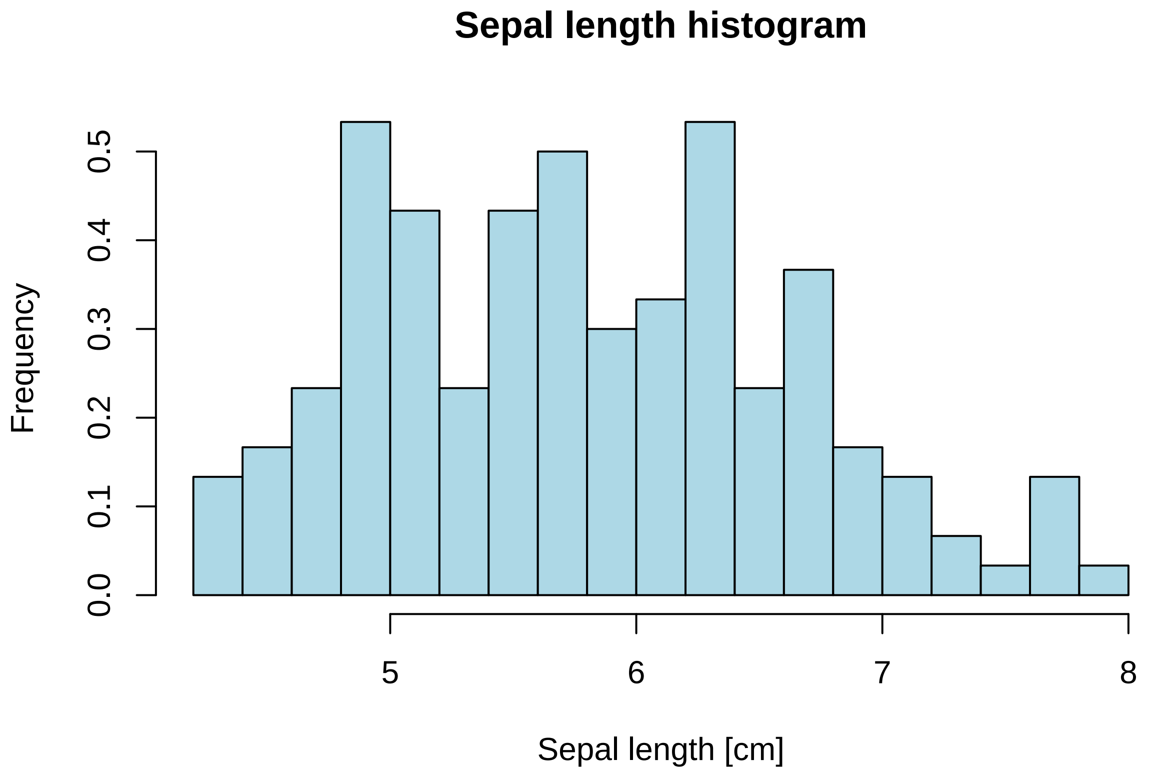
```
# Broj razreda i frekvencije:
length(h$breaks)
```

```
## [1] 20
```

```
h$counts
```

```
##  [1]  4  5  7 16 13  7 13 15  9 10 16  7 11  5  4  2  1  4  1
```

```
# Histogram s prikazom relativnih frekvencija:
h = hist(irisdata$Sepal.Length,
         prob=TRUE,
         breaks=15,
         main="Sepal length histogram",
         xlab="Sepal length [cm]",
         ylab='Frequency',
         col="lightblue"
         )
```



```
# Broj razreda i relativne frekvencije:
length(h$breaks)
```

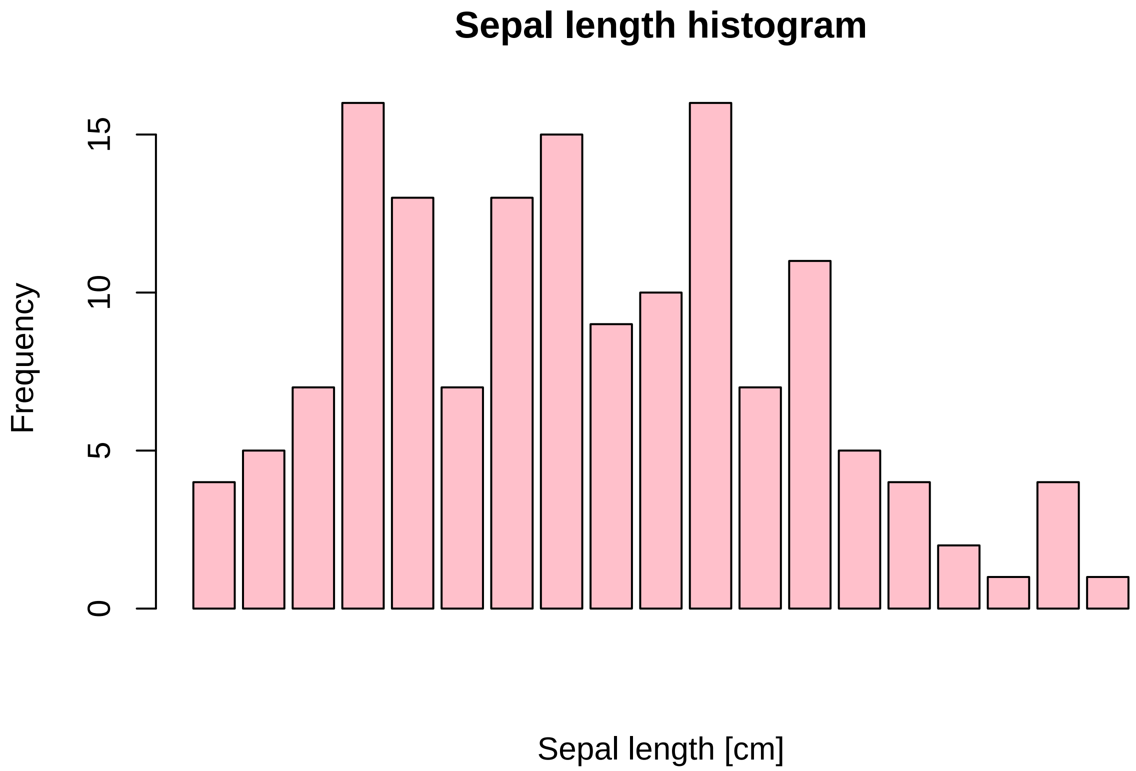
```
## [1] 20
```

```
h$density
```

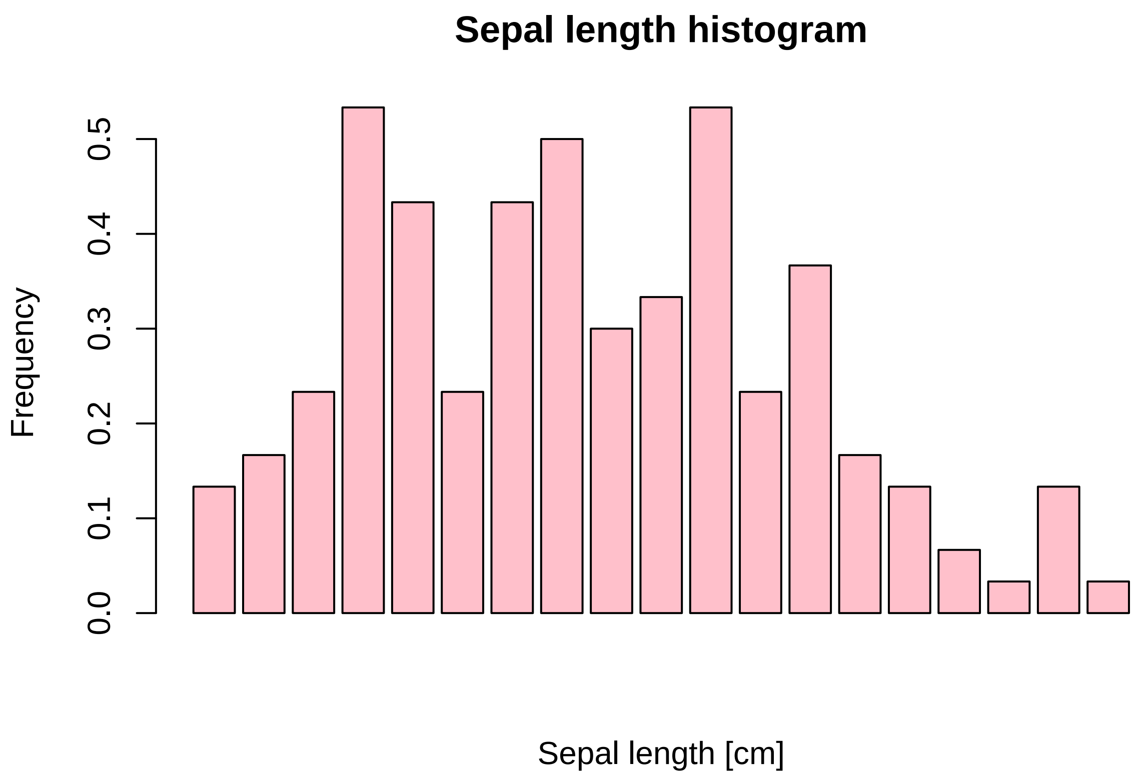
```
##  [1] 0.13333333 0.16666667 0.23333333 0.53333333 0.43333333 0.23333333
##  [7] 0.43333333 0.50000000 0.30000000 0.33333333 0.53333333 0.23333333
## [13] 0.36666667 0.16666667 0.13333333 0.06666667 0.03333333 0.13333333
## [19] 0.03333333
```

Stupčasti dijagram (barplot):

```
# Ako već imamo frekvencije:
data.counts = h$counts
barplot(data.counts,
        main="Sepal length histogram",
        xlab="Sepal length [cm]",
        ylab='Frequency',
        col="pink")
```



```
data.counts = h$density
barplot(data.counts,
        main="Sepal length histogram",
        xlab="Sepal length [cm]",
        ylab='Frequency',
        col="pink")
```



Usporedba grupiranih podataka:

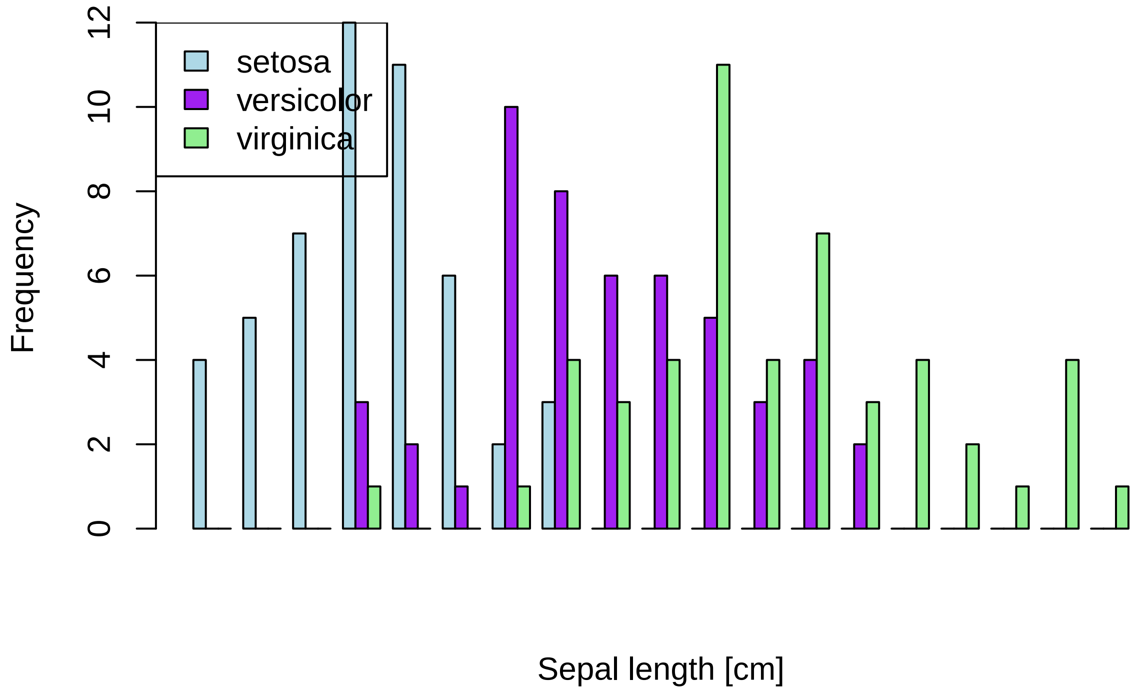
```
# Ako grupiramo podatke i onda radimo histogram:
b = seq(min(irisdata$Sepal.Length) - 0.1,max(irisdata$Sepal.Length) + 0.1,0.2)

h1 = hist(irisdata[irisdata["Species"] == c("setosa"),]$Sepal.Length,
         breaks=b,
         plot=FALSE)
h2 = hist(irisdata[irisdata["Species"] == c("versicolor"),]$Sepal.Length,
         breaks=b,
         plot=FALSE)
h3 = hist(irisdata[irisdata["Species"] == c("virginica"),]$Sepal.Length,
         breaks=b,
         plot=FALSE)

data <- t(cbind(h1$counts,h2$counts,h3$counts))
data
```

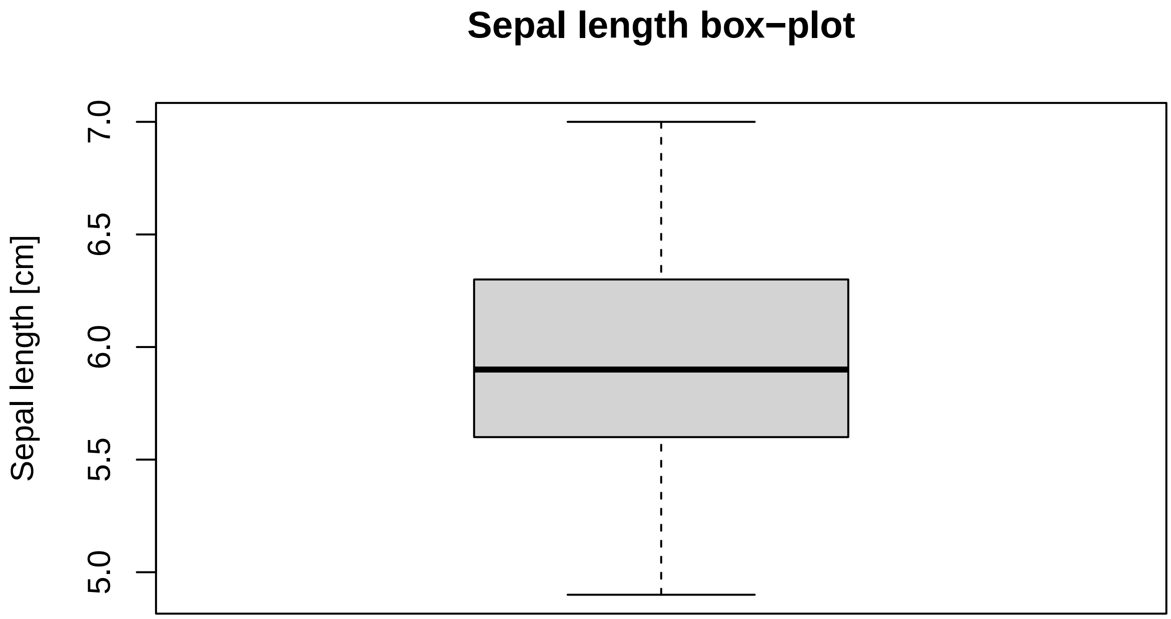
```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## [1,]    4    5    7   12   11    6    2    3    0     0     0     0     0     0
## [2,]    0    0    0    3    2    1   10    8    6     6     5     3     4     2
## [3,]    0    0    0    1    0    0    1    4    3     4    11     4     7     3
##      [,15] [,16] [,17] [,18] [,19]
## [1,]     0     0     0     0     0
## [2,]     0     0     0     0     0
## [3,]     4     2     1     4     1
```

```
barplot(data,beside=TRUE, col=c("lightblue", "purple", "lightgreen"), xlab="Sepal length [cm]", ylab='F
legend("topleft",c("setosa","versicolor","virginica"),fill = c("lightblue", "purple", "lightgreen"))
```



Usporedite pravokutne dijagrame različitih vrsta za pojedine varijable.

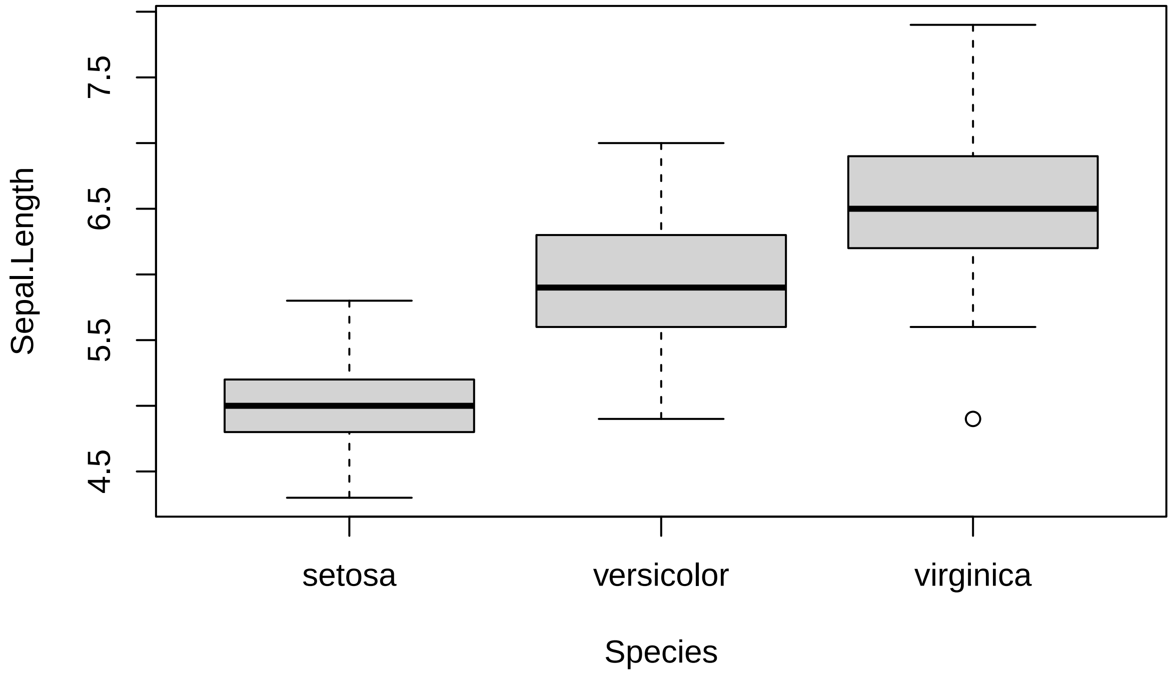
```
# Pravokutni dijagram versicolor vrste za duljinu lapa:
boxplot(irisdata[irisdata["Species"]=="versicolor",]$Sepal.Length,
        main='Sepal length box-plot',
        ylab='Sepal length [cm]')
```



```
# Vrijednosti pravokutnog dijagrama dolaze iz deskriptivne statistike:
summary(irisdata[irisdata["Species"]=="versicolor",]$Sepal.Length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.900   5.600   5.900   5.936   6.300   7.000
```

```
# Pravokutni dijagrami vrsta za duljinu lapa:
boxplot(Sepal.Length ~ Species,data=irisdata)
```



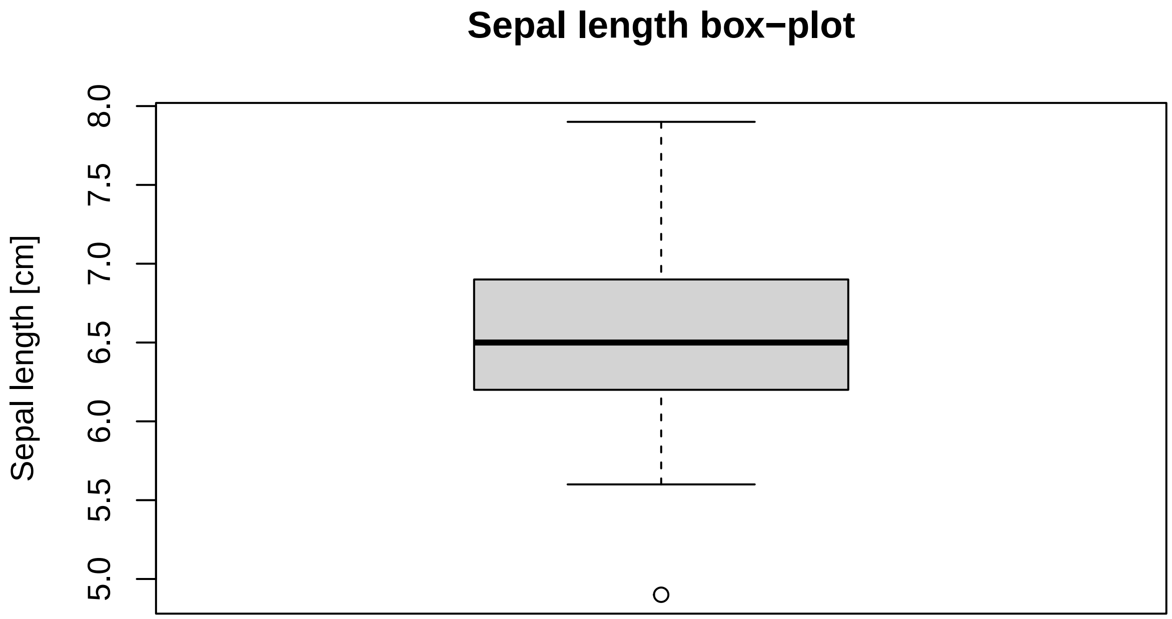
```
aggregate(irisdata[names(irisdata) != "Species"]$Sepal.Length, list(irisdata$Species), median)
```

```
##      Group.1   x
## 1     setosa 5.0
## 2 versicolor 5.9
## 3  virginica 6.5
```

```
summary(irisdata[irisdata["Species"]=="virginica",]$Sepal.Length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.900   6.225   6.500   6.588   6.900   7.900
```

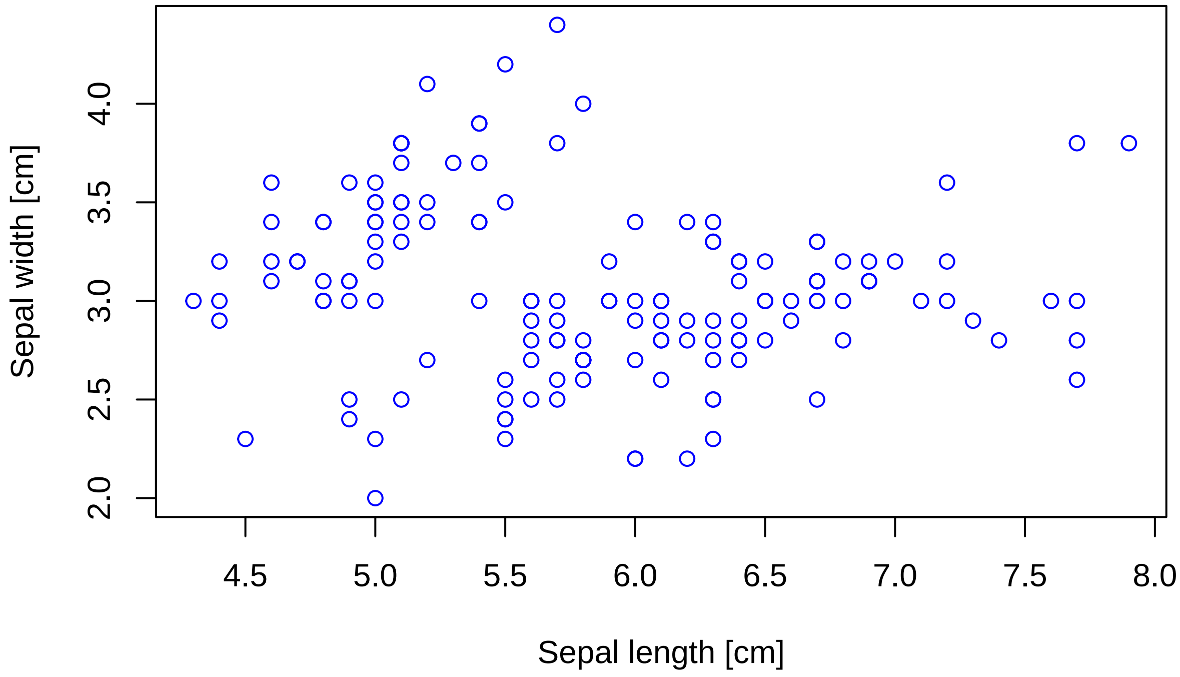
```
boxplot(irisdata[irisdata["Species"]=="virginica",]$Sepal.Length,
        main='Sepal length box-plot',
        ylab='Sepal length [cm]',
        range=1.5)
```



Možemo li iz dijagrama raspršenja naslutiti kakvu vezu između duljine i širine lapa?

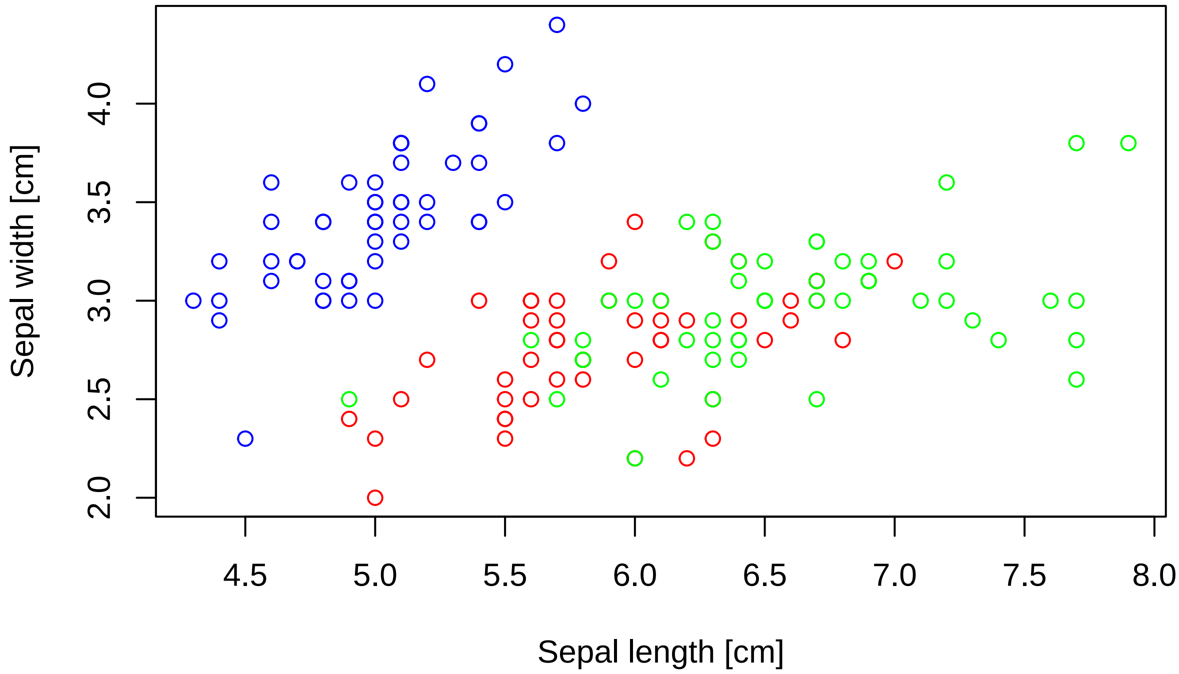
Neka iz grafa bude jasno koja točka zastupa koju vrstu irisa - možemo li što naslutiti iz tog prikaza?

```
# Ne razlikujemo vrste irisa:
plot(irisdata$Sepal.Length,irisdata$Sepal.Width,
     col="blue",
     xlab='Sepal length [cm]',
     ylab='Sepal width [cm]')
```



```
# Razlikujemo vrste irisa:
plot(irisdata$Sepal.Length[irisdata$Species=='setosa'],
     irisdata$Sepal.Width[irisdata$Species=='setosa'],
     col='blue',
     xlim=c(min(irisdata$Sepal.Length),max(irisdata$Sepal.Length)),
     ylim=c(min(irisdata$Sepal.Width),max(irisdata$Sepal.Width)),
     xlab='Sepal length [cm]',
     ylab='Sepal width [cm]')

points(irisdata$Sepal.Length[irisdata$Species=='versicolor'],
       irisdata$Sepal.Width[irisdata$Species=='versicolor'],col='red')
points(irisdata$Sepal.Length[irisdata$Species=='virginica'],
       irisdata$Sepal.Width[irisdata$Species=='virginica'],col='green')
```

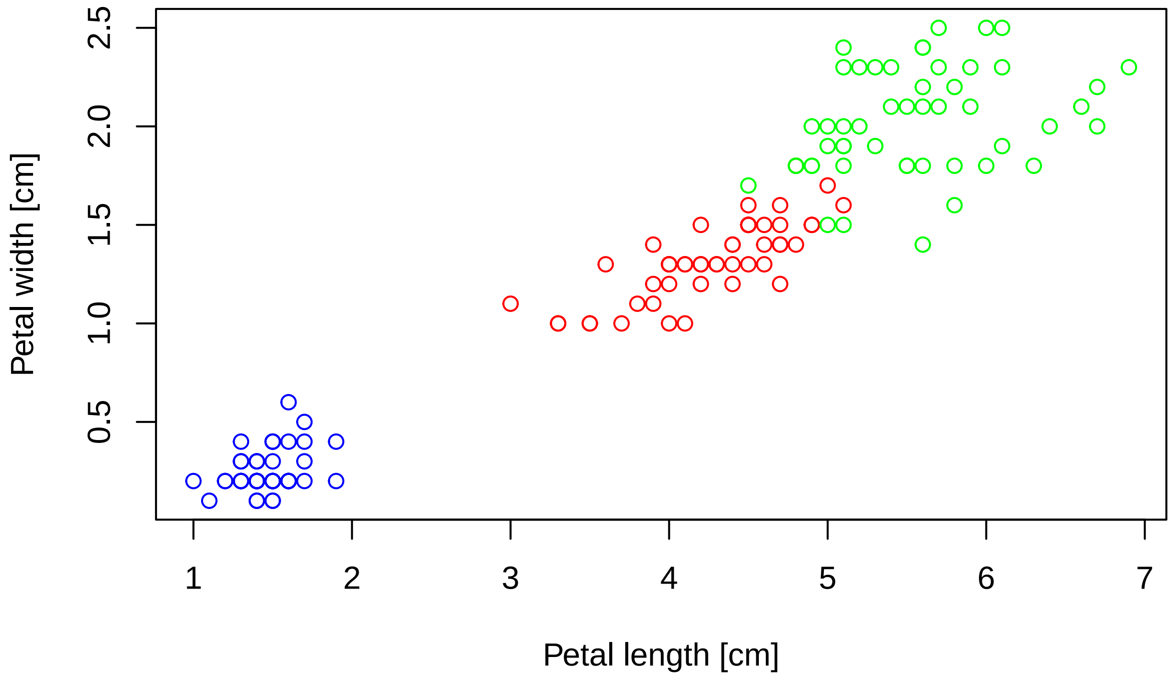


Što možemo naslutiti ako nacrtamo dijagram raspršenja za duljine i širine latica?

```
# Provjerimo kako izgleda scatterplot za latice:

plot(irisdata$Petal.Length[irisdata$Species=='setosa'],
     irisdata$Petal.Width[irisdata$Species=='setosa'],
     col='blue',
     xlim=c(min(irisdata$Petal.Length),max(irisdata$Petal.Length)),
     ylim=c(min(irisdata$Petal.Width),max(irisdata$Petal.Width)),
     xlab='Petal length [cm]',
     ylab='Petal width [cm]')

points(irisdata$Petal.Length[irisdata$Species=='versicolor'],
       irisdata$Petal.Width[irisdata$Species=='versicolor'],col='red')
points(irisdata$Petal.Length[irisdata$Species=='virginica'],
       irisdata$Petal.Width[irisdata$Species=='virginica'],col='green')
```



Prljavi podatci

Osim ugrađenih skupova podataka, u R možemo učitati i podatke iz datoteka različitih formata.

Što smo do sada zanemarili provjeriti, odnosno podrazumijevali?

Pri učitavanju podataka iz datoteka može se dogoditi da su tipovi nekih varijabli krivo prepoznati – u tom slučaju potrebno je provjeriti tipove i ručno ih ispraviti. Također, moguće je da u podatcima nedostaju neke vrijednosti, koje u učitanom `data.frame`-u poprimaju vrijednost `NA`. Moguće je i da su neke vrijednosti krivo unesene ili krivo učitane.

```
# Učitavanje podataka iz csv datoteke:
iris.modif = read.table("iris_mod.txt")
head(iris.modif)
```

```
##              V1                                                   V2
## 1 Sepal.Length ,"Sepal.Width","Petal.Length","Petal.Width","Species"
## 2             1                                 ,5.1,3.5,1.4,0.2,"1"
## 3             2                                   ,4.9,3,1.4,0.2,"1"
## 4             3                                 ,4.7,3.2,1.3,0.2,"1"
## 5             4                                 ,4.6,3.1,1.5,0.2,"1"
## 6             5                                   ,5,3.6,1.4,0.2,"1"
```

Vidimo da se sve krivo ucitalo jer nije dobar separator pa cemo popraviti separator:

```r
iris.modif = read.table("iris_mod.txt", sep = ",")
head(iris.modif)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2       1
## 2          4.9         3.0          1.4         0.2       1
## 3          4.7         3.2          1.3         0.2       1
## 4          4.6         3.1          1.5         0.2       1
## 5          5.0         3.6          1.4         0.2       1
## 6          5.4         3.9          1.7         0.4       1
```

```r
dim(iris.modif)
```

```
## [1] 150   5
```

Jesu li tipovi stupaca ispravni?

```r
class(iris.modif$Species)
```

```
## [1] "integer"
```

```r
# Klasa je integer - to ne želimo jer se radi o tipu irisa --> kategorijska varijabla!
iris.modif$Species = as.factor(iris.modif$Species)
class(iris.modif$Species)
```

```
## [1] "factor"
```

```r
iris.modif$Species
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [75] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3
## [112] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [149] 3 3
## Levels: 1 2 3
```

```r
iris.modif$Petal.Width
```

```
##   [1]   0.2   0.2   0.2   0.2   0.2   0.4   0.3   0.2   0.2   0.1   0.2   0.2
##  [13]   0.1   0.1   0.2   0.4   0.4   0.3   0.3   0.3   0.2   0.4   0.2   0.5
##  [25]   0.2   0.2    NA   0.2   0.2   0.2   0.2   0.4   0.1   0.2   0.2   0.2
##  [37]   0.2   0.1   0.2   0.2   0.3   0.3   0.2   0.6   0.4   0.3   0.2   0.2
##  [49]   0.2   0.2   1.4   1.5   1.5   1.3   1.5   1.3   1.6   1.0   1.3   1.4
##  [61]   1.0   1.5   1.0   1.4   1.3   1.4   1.5   1.0   1.5   1.1   1.8   1.3
##  [73]   1.5   1.2   1.3   1.4   1.4   1.7   1.5   1.0   1.1   1.0   1.2   1.6
##  [85]   1.5   1.6   1.5   1.3   1.3   1.3   1.2   1.4   1.2   1.0   1.3   1.2
##  [97]   1.3   1.3   1.1   1.3   2.5   1.9   2.1   1.8   2.2   2.1   1.7   1.8
## [109] 100.8   2.5   2.0   1.9   2.1   2.0   2.4   2.3   1.8   2.2   2.3   1.5
## [121]   2.3   2.0   2.0   1.8   2.1   1.8   1.8   1.8   2.1   1.6   1.9   2.0
## [133]   2.2   1.5   1.4   2.3   2.4   1.8   1.8   2.1   2.4   2.3   1.9   2.3
## [145]   2.5   2.3   1.9   2.0   2.3   1.8
```

Ima li nedostajućih vrijednosti?

```
# is.na ce nam vratiti logical vektor koji ima TRUE na mjestima gdje pod$Petal.Length ima NA:
sum(is.na(iris.modif$Petal.Length)) # Koliko?
```

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

```
sum(is.na(iris.modif$Sepal.Length)) # Koliko?
```

```
## [1] 4
```

```
sum(is.na(iris.modif$Petal.Width)) # Koliko?
```

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

```
sum(is.na(iris.modif$Sepal.Width)) # Koliko?
```

```
## [1] 2
```

```
sum(is.na(iris.modif$Species)) # Koliko?
```

```
## [1] 0
```

```
# complete.cases ce vratiti logical vrijednost za svaki redak;
# Vrijednost je FALSE --> barem jedan element retka NA
sum(!complete.cases(iris.modif))
```

```
## [1] 8
```

```
# Koji su to retci?
iris.modif[!complete.cases(iris.modif),]
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 11           5.4         3.7           NA         0.2       1
## 27           5.0         3.4          1.6          NA       1
## 51            NA         3.2          4.7         1.4       2
## 89            NA         3.0          4.1         1.3       2
## 110          7.2          NA          6.1         2.5       3
## 116           NA         3.2          5.3         2.3       3
## 119           NA         2.6          6.9         2.3       3
## 129          6.4          NA          5.6         2.1       3
```

```
# Izbacit ćemo nedostajuce vrijednosti
iris.modif.full = iris.modif[complete.cases(iris.modif),]

iris.modif.full %>% group_by(Species) %>% summarise(
          count = n())
```

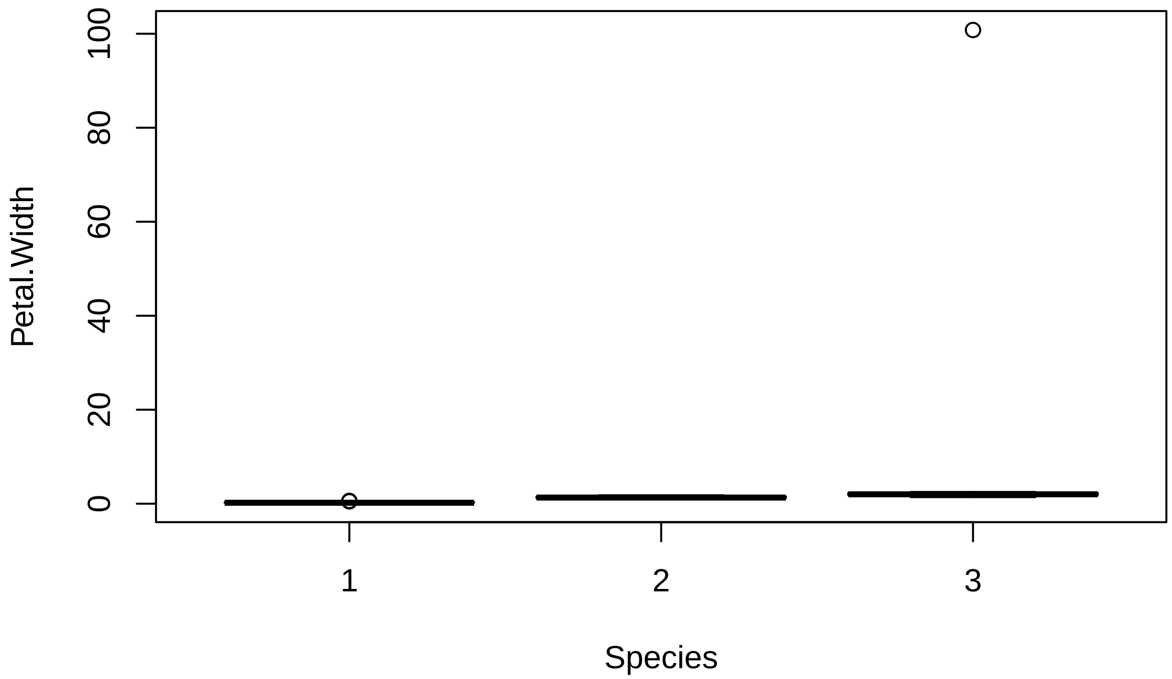
```
## # A tibble: 3 x 2
##   Species count
##   <fct>   <int>
## 1 1          48
## 2 2          48
## 3 3          46
```

Deskriptivna statistika:

```
summary(iris.modif.full)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width      Species
##  Min.   :4.300   Min.   :2.00   Min.   :1.000   Min.   :  0.100   1:48
##  1st Qu.:5.100   1st Qu.:2.80   1st Qu.:1.525   1st Qu.:  0.300   2:48
##  Median :5.800   Median :3.00   Median :4.300   Median :  1.300   3:46
##  Mean   :5.815   Mean   :3.05   Mean   :3.718   Mean   :  1.876
##  3rd Qu.:6.400   3rd Qu.:3.30   3rd Qu.:5.100   3rd Qu.:  1.800
##  Max.   :7.900   Max.   :4.40   Max.   :6.700   Max.   :100.800
```

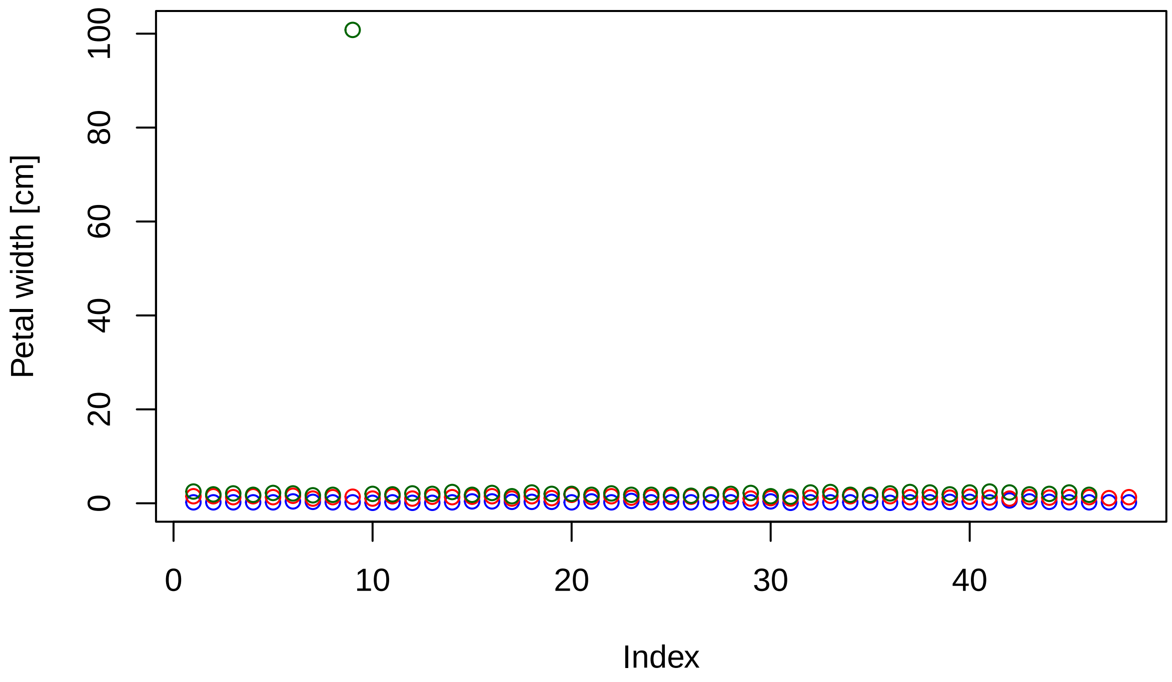
```
boxplot(Petal.Width ~ Species,data=iris.modif.full)
```



```
iris.modif.full %>% group_by(Species) %>% summarise(
          sd = sd(Petal.Width),
          IQR = IQR(Petal.Width),
          mean= mean(Petal.Width)
            ) -> summary.result
summary.result
```

```
## # A tibble: 3 x 4
##   Species     sd   IQR  mean
##   <fct>    <dbl> <dbl> <dbl>
## 1 1        0.105 0.1   0.244
## 2 2        0.202 0.3   1.32
## 3 3       14.6   0.475 4.15
```

```
plot(iris.modif.full$Petal.Width[iris.modif.full$Species==1],
     col='blue',
     ylim=c(min(iris.modif.full$Petal.Width),max(iris.modif.full$Petal.Width)),
     ylab='Petal width [cm]')
points(iris.modif.full$Petal.Width[iris.modif.full$Species==2],col='red')
points(iris.modif.full$Petal.Width[iris.modif.full$Species==3],col='dark green')
```



Koji je to outlier? Je li taj outlier točna vrijednost?

```
ind = which(iris.modif.full$Petal.Width >20)
iris.modif.full[ind,]
```

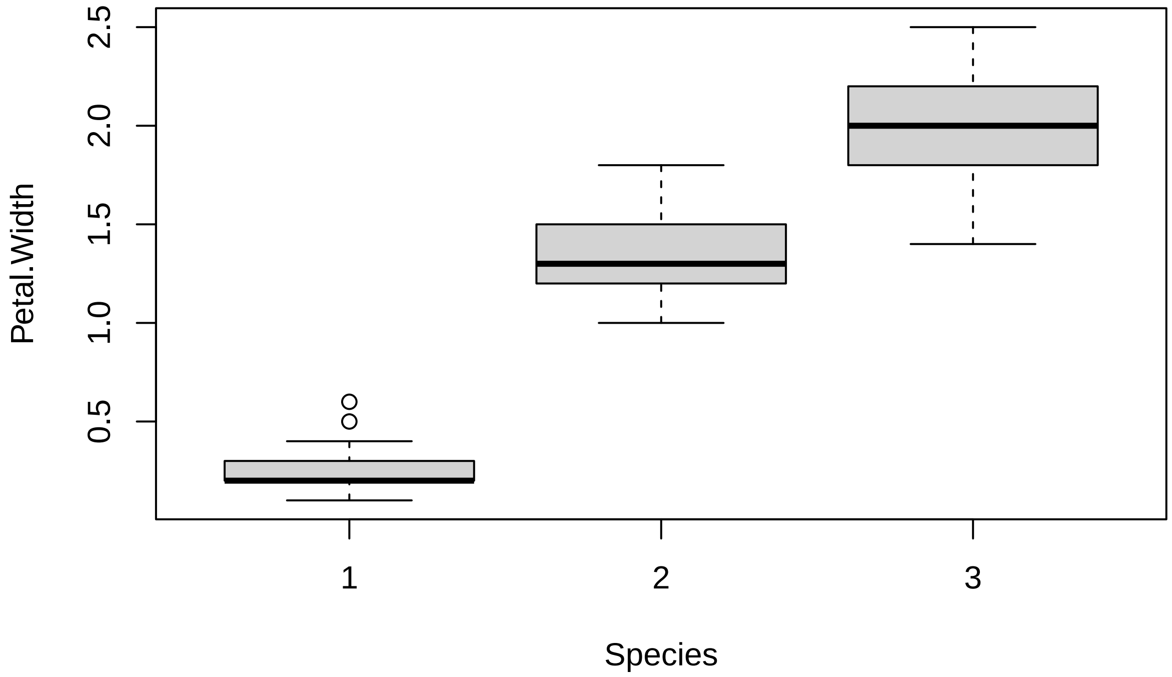
```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 109          6.7         2.5          5.8       100.8       3
```

```
iris.cleaned = iris.modif.full[-ind,]

summary(iris.cleaned)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width    Species
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   1:48
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.500   1st Qu.:0.300   2:48
##  Median :5.800   Median :3.000   Median :4.300   Median :1.300   3:45
##  Mean   :5.809   Mean   :3.054   Mean   :3.703   Mean   :1.174
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.700   Max.   :2.500
```

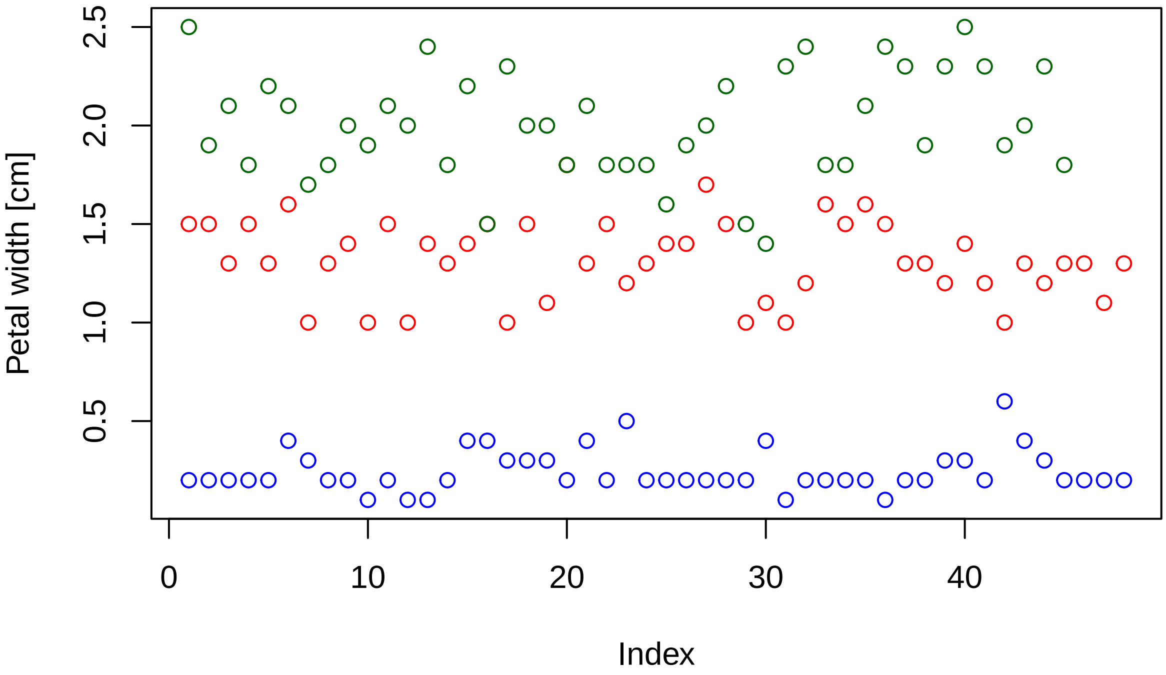
```
boxplot(Petal.Width ~ Species,data=iris.cleaned)
```



```
iris.cleaned %>% group_by(Species) %>% summarise(
          sd = sd(Petal.Width),
          IQR = IQR(Petal.Width),
          mean= mean(Petal.Width)
            ) -> summary.result

plot(iris.cleaned$Petal.Width[iris.cleaned$Species==1],
     col='blue',
     ylim=c(min(iris.cleaned$Petal.Width),max(iris.cleaned$Petal.Width)),
     ylab='Petal width [cm]')
```

```
points(iris.cleaned$Petal.Width[iris.cleaned$Species==2],col='red')
points(iris.cleaned$Petal.Width[iris.cleaned$Species==3],col='dark green')
```



Je li izbacivanjzue redaka s nedostajućim vrijednostima uvijek najbolje rješenje?

Npr., pretpostavimo da imamo 1000 podataka (redaka) opisanih sa 150 varijabli (stupaca), te u 90% njih varijabla pod rednim brojem 84. ima NA (ostale su prisutne). Ako bismo maknuli sve retke kod kojih funkcija `complete.cases()` poprima vrijednost `FALSE`, drastično bismo smanjili skup podataka (10 puta!). S druge strane, ako maknemo samo stupac 84, još uvijek imamo 1000 podataka (no jednu varijablu tj. stupac manje). Nekad ćemo moći ručno popuniti nedostajuće vrijednosti, a nekad će nedostajuća vrijednost nositi neku dodatnu informaciju.

Kako ćemo tretirati nedostajuće vrijednosti ovisit će prvenstveno o samom datasetu, odnosno kontekstu podataka. Za kraj ovih vježbi, još jednom, naglašavamo značajnost interpretacije podataka!