Visualizzazione di dati in R

Corso di Bioinformatica

Nicola Vitacolonna
Corso di Laurea in Biotecnologie

Riepilogo: strutture dati in R

TIPO DI DATO	CLASSE DELL'OGGETTO IN R	ESEMPIO
stringa	character	"Io sono una stringa"
numero intero	integer	42L
numero reale	numeric	3.14
numero complesso	complex	17 + 24i
valore booleano	logical	TRUE
vettore	vector	c(1.0, 2.7, 3.1, 6.5)
vettore categoriale	factor	factor(c("a", "a", "b", "b", "b", "c"))
matrice	matrix	matrix(1:20, nrow = 5)
tabella di contingenza	table	table(c(4,4,5,5,5,6,7,7,7))
matrice dei dati	data.frame	data.frame(col.1 = c(3,2,4), col.2 = c("y","y","n"))
serie temporale	ts	ts(1:12, frequency = 12, start = c(2013, 3))

Riferimenti

- An Introduction to R (http://cran.r-project.org/doc/manuals/r-release/R-intro.html), Cap. 12
- Quick-R (http://www.statmethods.net)
- Producing Simple Graphs with R (http://www.harding.edu/fmccown/R/)

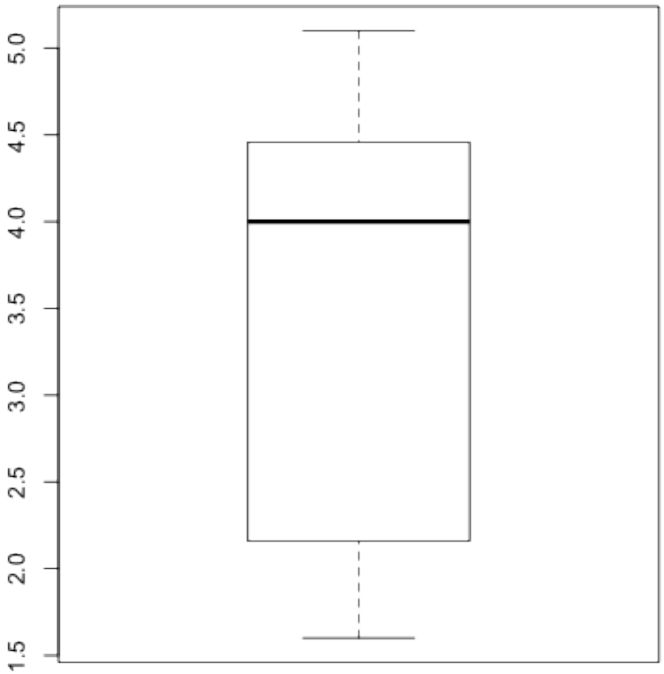
Alcune tipologie di grafici

- Diagramma a scatola e baffi (*boxplot*)
- Istogramma
- Sorted plot
- Boxplot multipli
- Strip chart
- Grafico di dispersione (scatterplot)
- Heat map
- Grafico a linee
- Grafico a barre (semplice, segmentato, multiplo)
- Grafico a bastoncino
- Dot chart
- Grafico a mosaico
- …

Diagramma a scatola e baffi (boxplot)

- Linea spessa: mediana
- Lato inferiore della scatola: primo quartile
- Lato superiore della scatola: terzo quartile
- Altezza della scatola: intervallo interquartile (o *IQR*, *Interquartile Range*)
- Estensione dei baffi: 1.5 volte l'IQR
- Eventuali valori al di fuori dei baffi sono considerati *valori anomali* (*outlier*)

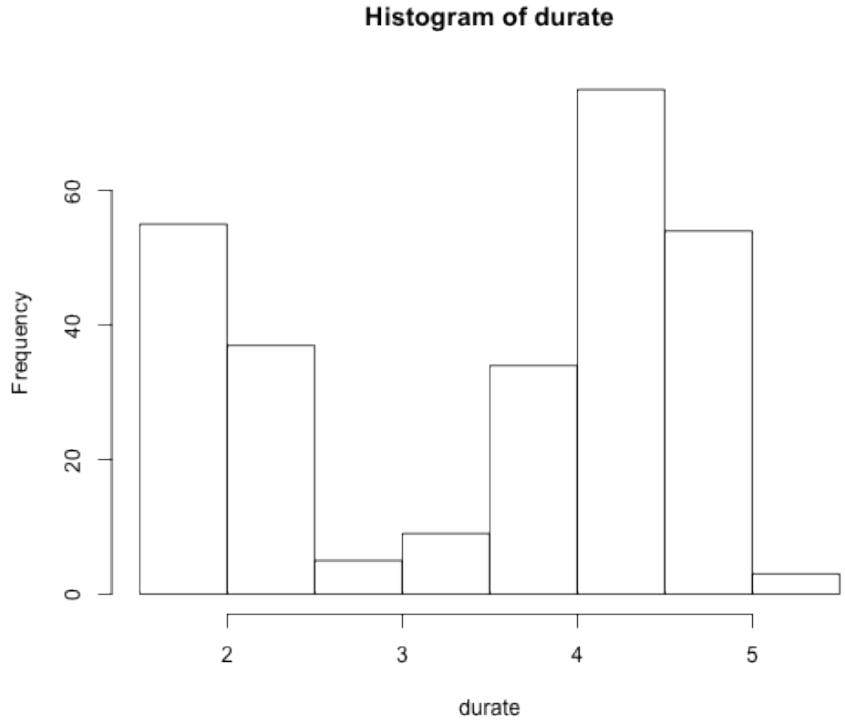
```
data(faithful)
durate <- faithful$eruptions
boxplot(durate)
```



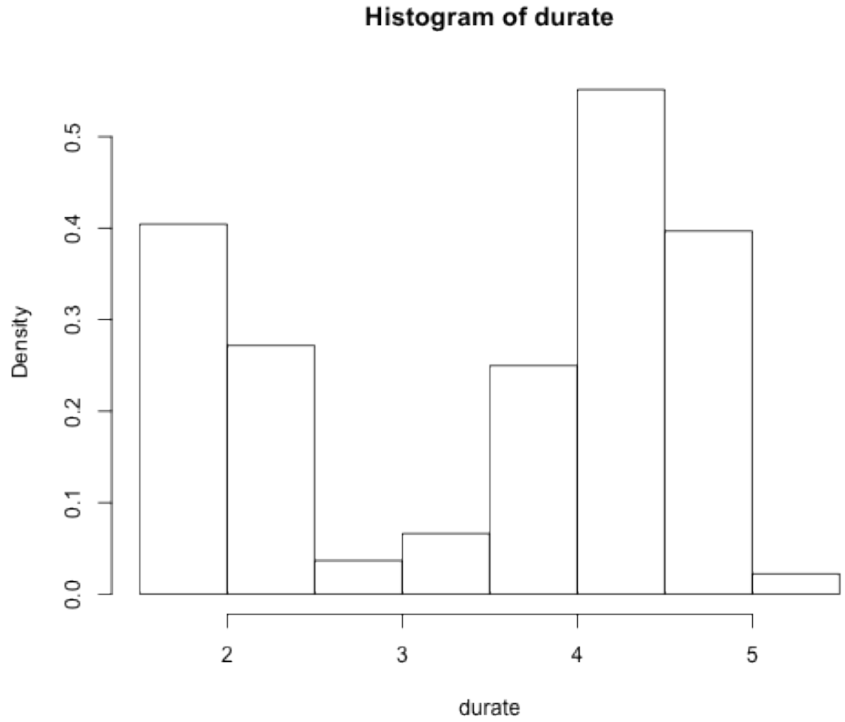
Istogramma

Visualizza una distribuzione di frequenze (assolute o relative)

```
hist(durate)
```



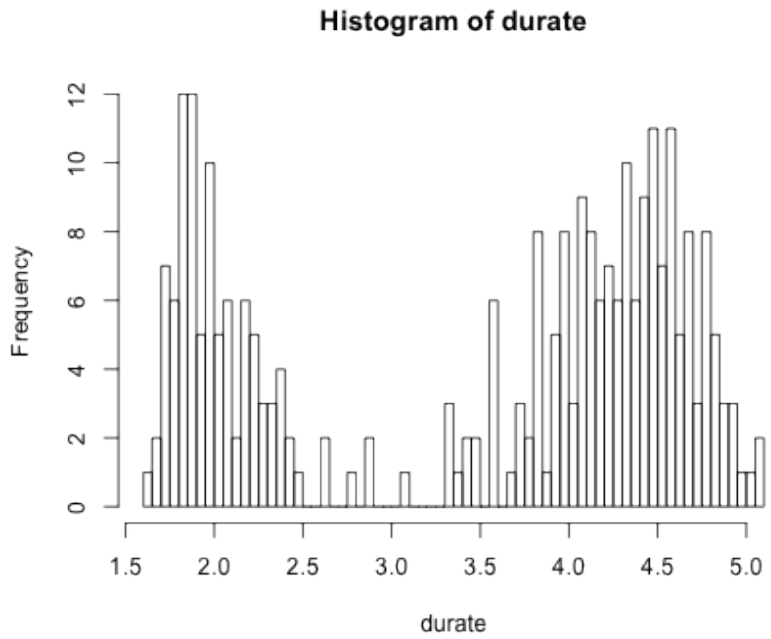

```
hist(durate, freq = FALSE)
```



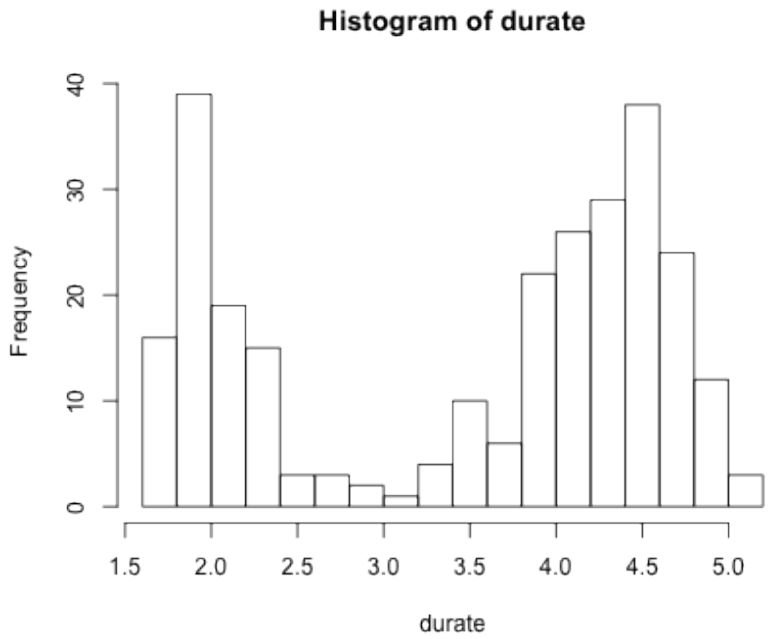
Istogrammi: specificare le classi

- Con il parametro `breaks` è possibile specificare il numero di classi
- Il numero di classi dev'essere appropriato al tipo di distribuzione

```
hist(durate, breaks = 50)
```



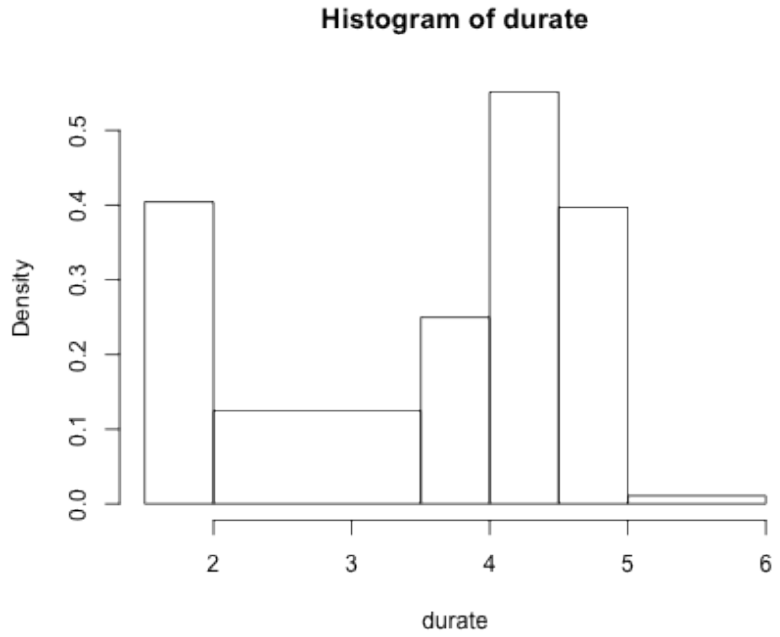

```
hist(durate, breaks = 20)
```



Istogrammi: specificare le classi

- È anche possibile specificare come raggruppare i valori mediante un vettore
- Gli intervalli non devono essere necessariamente equispaziati (in un istogramma, è l'*area* dei rettangoli che è proporzionale alla frequenza)

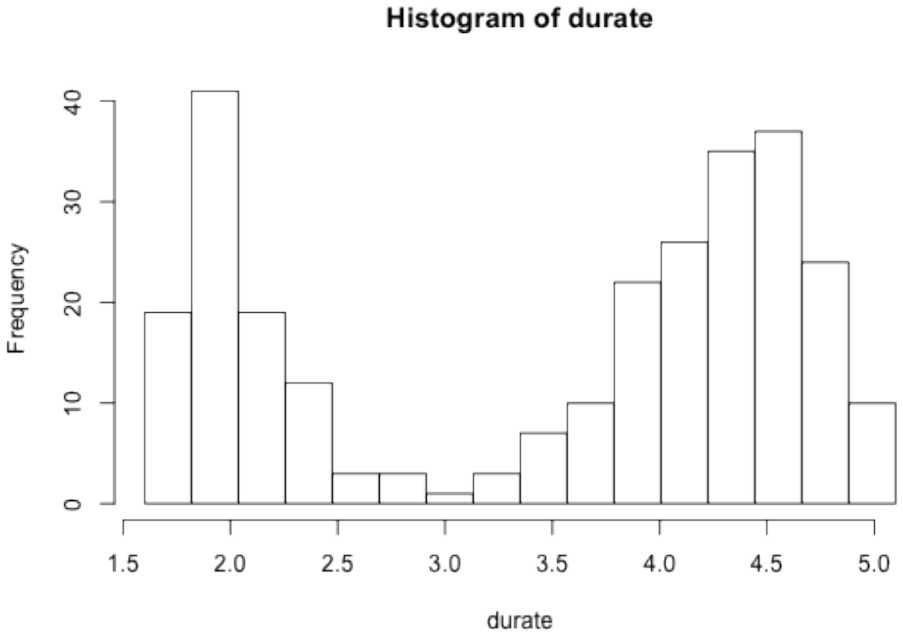
```
hist(durate, breaks = c(1.5, 2, 3.5, 4, 4.5, 5, 6))
```



Istogrammi: calcolo del numero di classi

Metodo della radice quadrata (http://en.wikipedia.org/wiki/Histogram): $n = \sqrt{|\bar{x}|}$

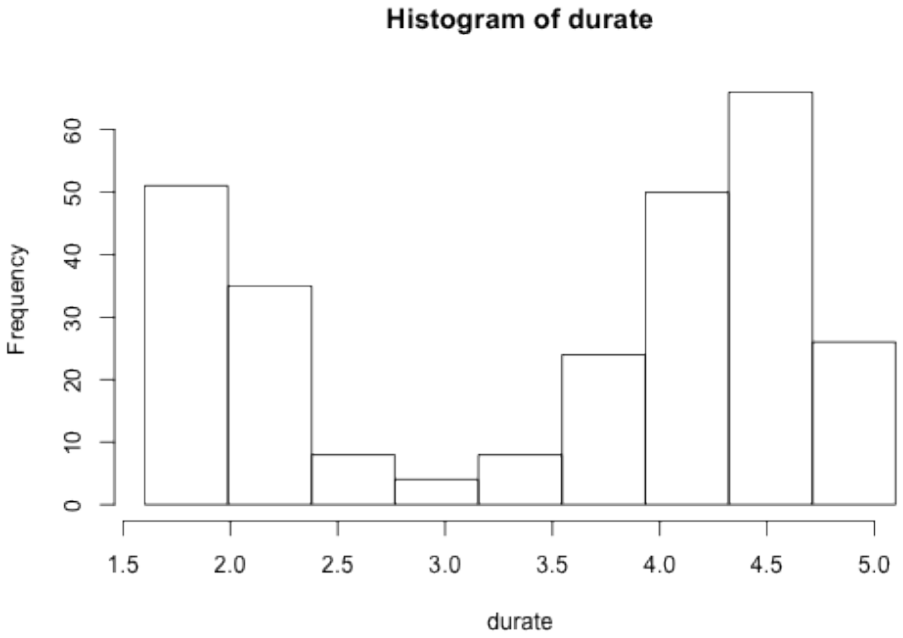
```
n <- round(sqrt(length(durate)))  # Radice quadrata (arrotondata) del numero di valori
classi <- seq(min(durate), max(durate), length = n + 1)
hist(durate, breaks = classi)
```



Istogrammi: calcolo del numero di classi

Metodo di Sturges (http://en.wikipedia.org/wiki/Histogram): $n = \lceil \log_2(\bar{x} + 1) \rceil$

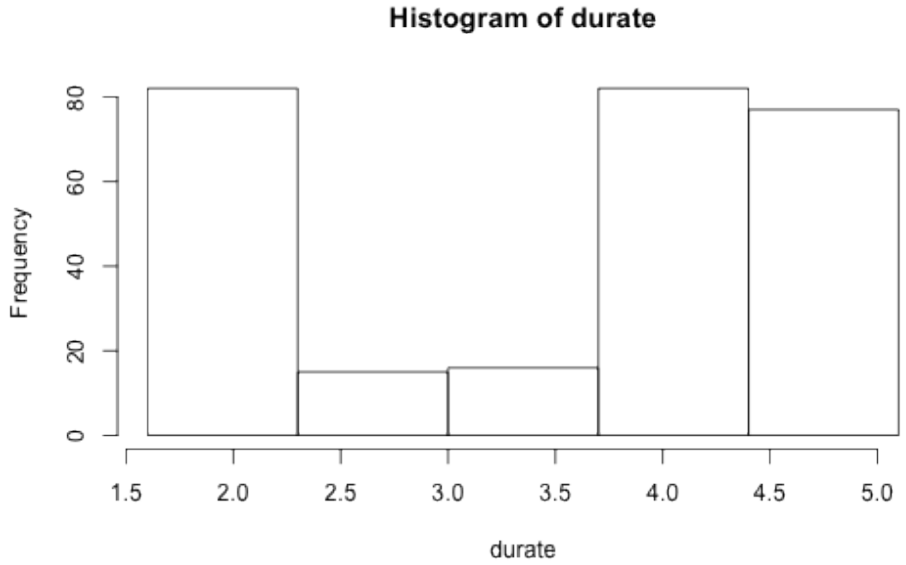
```
n <- ceiling(log2(length(durate) + 1))
classi <- seq(min(durate), max(durate), length = n + 1)
hist(durate, breaks = classi)
```



Istogrammi: calcolo del numero di classi

Metodo di Freedman–Diaconis (http://en.wikipedia.org/wiki/Histogram): $w = 2\,\frac{\mathrm{IQR}(\bar{x})}{|\bar{x}|^{1/3}}$

```
larghezza <- 2 * IQR(durate)/(length(durate)^(1/3))
n <- ceiling((max(durate) - min(durate))/larghezza)
classi <- seq(min(durate), max(durate), length = n + 1)
hist(durate, breaks = classi)
```



La scelta delle classi è importante!

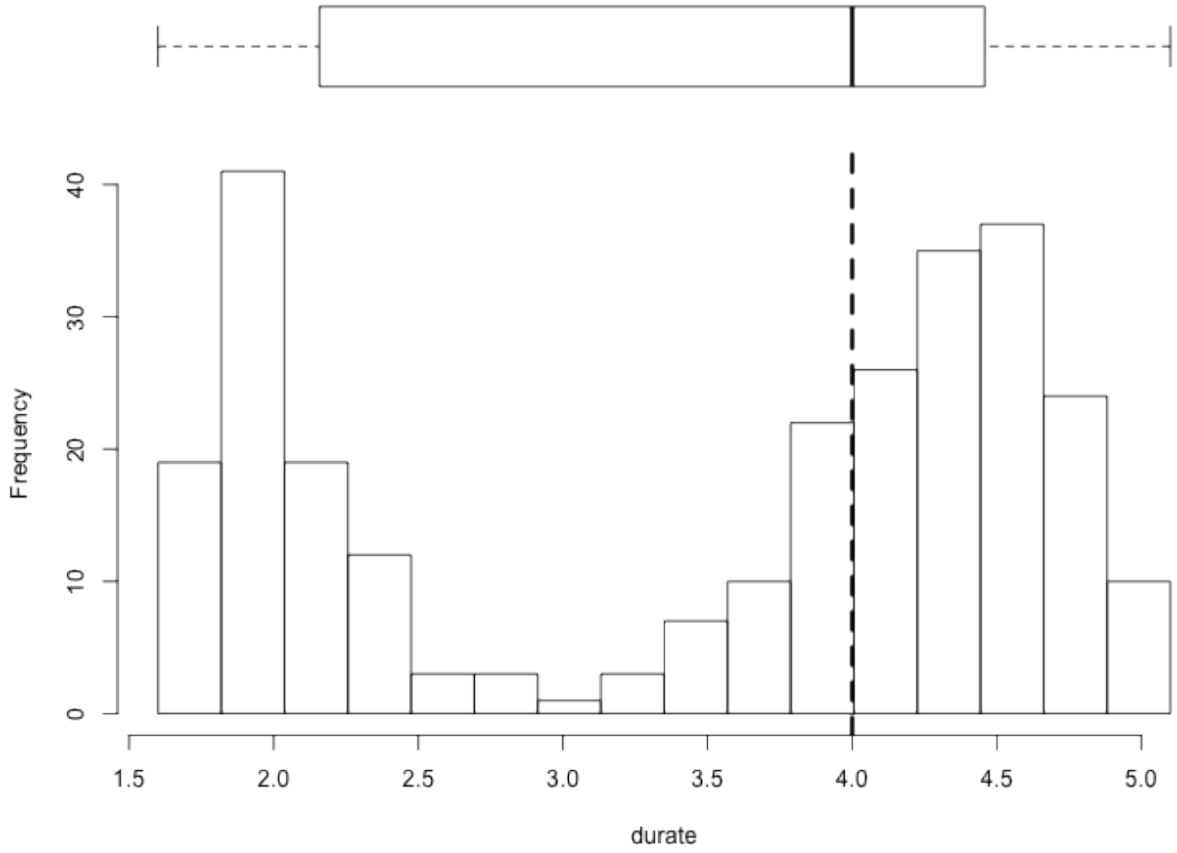
Assessing approximate distribution of data based on a histogram (http://stats.stackexchange.com/questions/51718/assessing-approximate-distribution-of-data-based-on-a-histogram/51753)

```
x <- c(1.03, 1.24, 1.47, 1.52, 1.92, 1.93, 1.94, 1.95, 1.96, 1.97, 1.98, 1.99,
    2.72, 2.75, 2.78, 2.81, 2.84, 2.87, 2.9, 2.93, 2.96, 2.99, 3.6, 3.64, 3.66,
    3.72, 3.77, 3.88, 3.91, 4.14, 4.54, 4.77, 4.81, 5.62)
```

```
hist(x, breaks = seq(0.3, 6.7, by = 0.8))
```

```
hist(x, breaks = 0:8)
```

Relazione tra istogramma e box-plot



Caricate in memoria il data set `iris` con il comando `data(iris)`. Disegnate un istogramma che consenta di valutare la frequenza di valori della variabile `Sepal.Width` compresi tra 2.9 e 3.0. Qual è tale frequenza?

- Circa 60.
- Poco piú di 35.
- Poco piú di 25.
- Minore di 25.

Submit Show Hint Show Answer Clear

Disegnate un istogramma della distribuzione delle larghezze dei petali della specie `virginica`. Il valore con frequenza maggiore è compreso tra

- 1.6 e 1.8.
- 1.8 e 2.0.
- 1.7 e 2.1.
- 2.2 e 2.5.

Submit Show Hint Show Answer Clear

Grafico dei valori ordinati (sorted plot)

La funzione `sort()` ordina gli elementi di un vettore

```
plot(sort(durate))
```

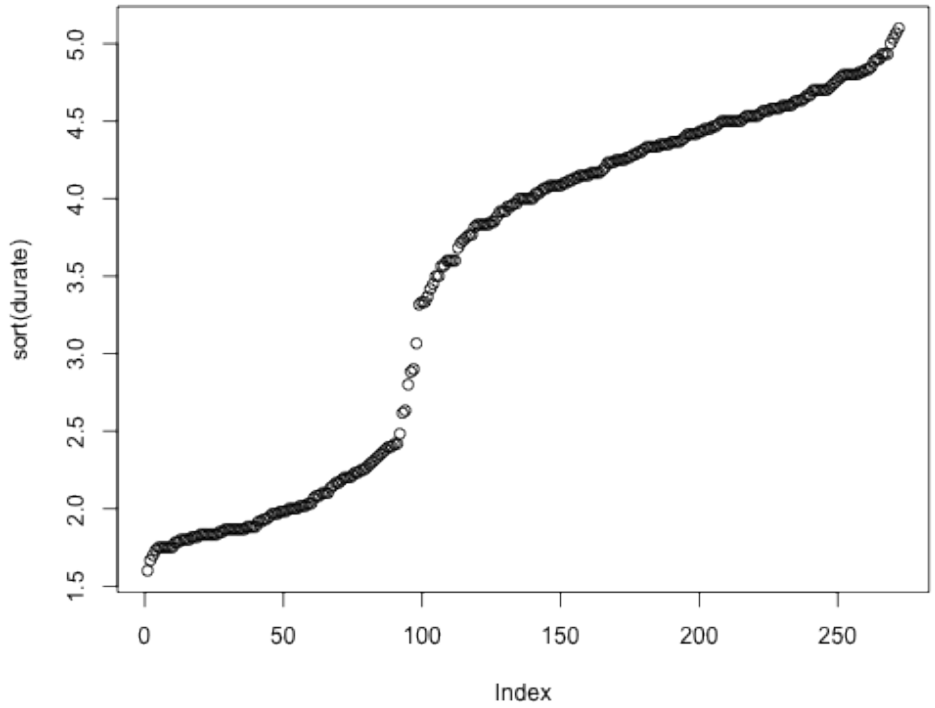
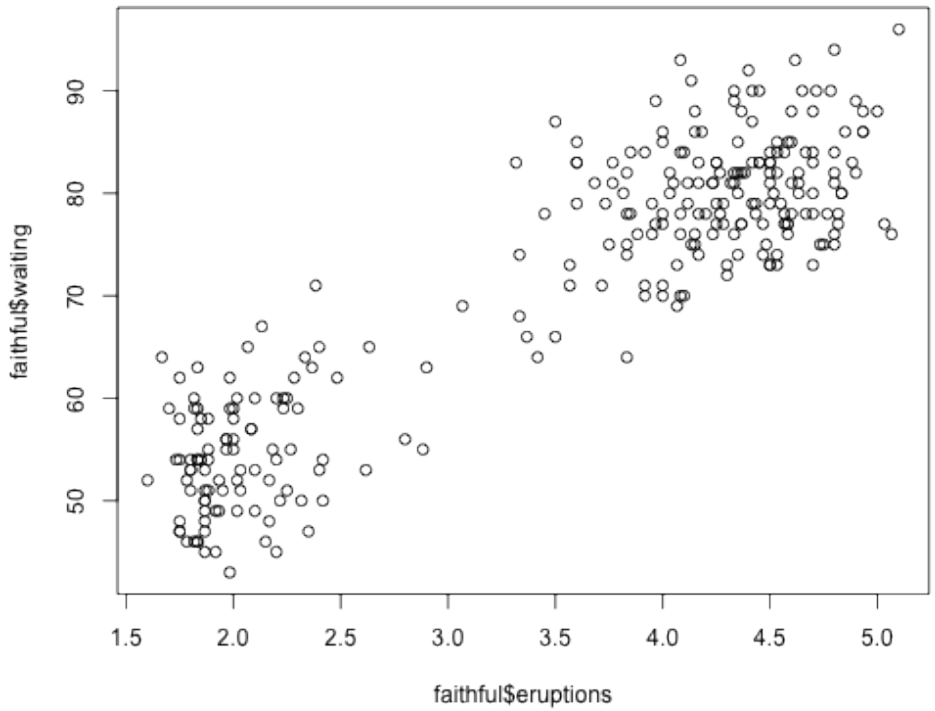


Grafico di dispersione (scatterplot)

```
plot(faithful$eruptions, faithful$waiting)  # plot(x,y)
```



Filtrate dal data frame `iris` le osservazioni relative alla specie `setosa` e memorizzate il risultato in una variabile chiamata `subset_setosa`. Disegnate un insieme di grafici di dispersione per confrontare tutte le coppie di variabili tra loro usando il comando `pairs(subset_setosa)`. Tra quali coppie di variabili ritenete sia piú ragionevole ipotizzare un'associazione?

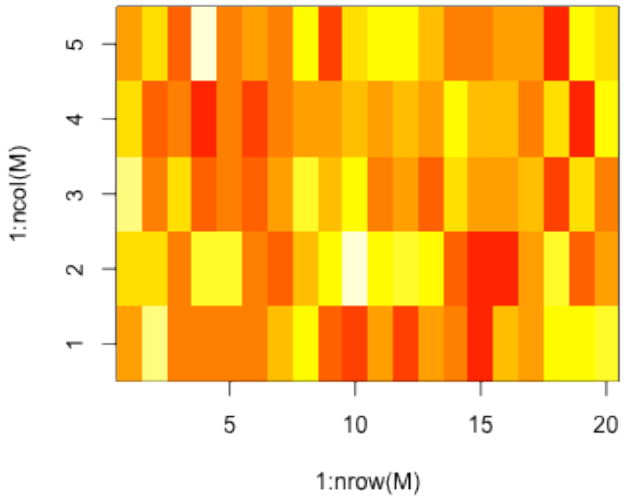
- Sepal.Width e Petal.Length.
- Sepal.Length e Petal.Width.
- Sepal.Length e Petal.Length.
- Sepal.Width e Sepal.Length.

Submit Show Hint Show Answer Clear

Heat map

- Si possono disegnare a partire da una matrice con la funzione `image()`
- Attenzione: nel grafico i dati sono trasposti (righe e colonne sono scambiate)!

```
dati <- rnorm(20 * 5)  # Genera 100 valori casuali con distribuzione normale
M <- matrix(dati, nrow = 20)  # Costruisce una matrice 20x5
image(1:nrow(M), 1:ncol(M), M)  # image(x, y, z)
```



Heat map

```
dati <- seq(1, 10, by = 1)
M <- matrix(dati, nrow = 5)
M
```

```
##      [,1] [,2]
## [1,]    1    6
## [2,]    2    7
## [3,]    3    8
## [4,]    4    9
## [5,]    5   10
```

Per disegnare correttamente una heat map bisogna:

1. riordinare le righe dall'ultima alla prima
2. trasporre la matrice con la funzione `t()`

```
M1 <- t(M[nrow(M):1, ])
image(1:nrow(M1), 1:ncol(M1), M1)
```

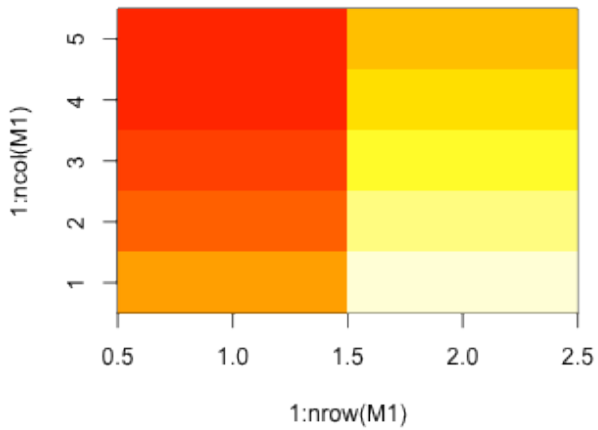


Grafico a linee

```
x <- seq(30, 50, length = 50)
y <- sample(0:100, 50)
plot(x, y, type = "l")
```

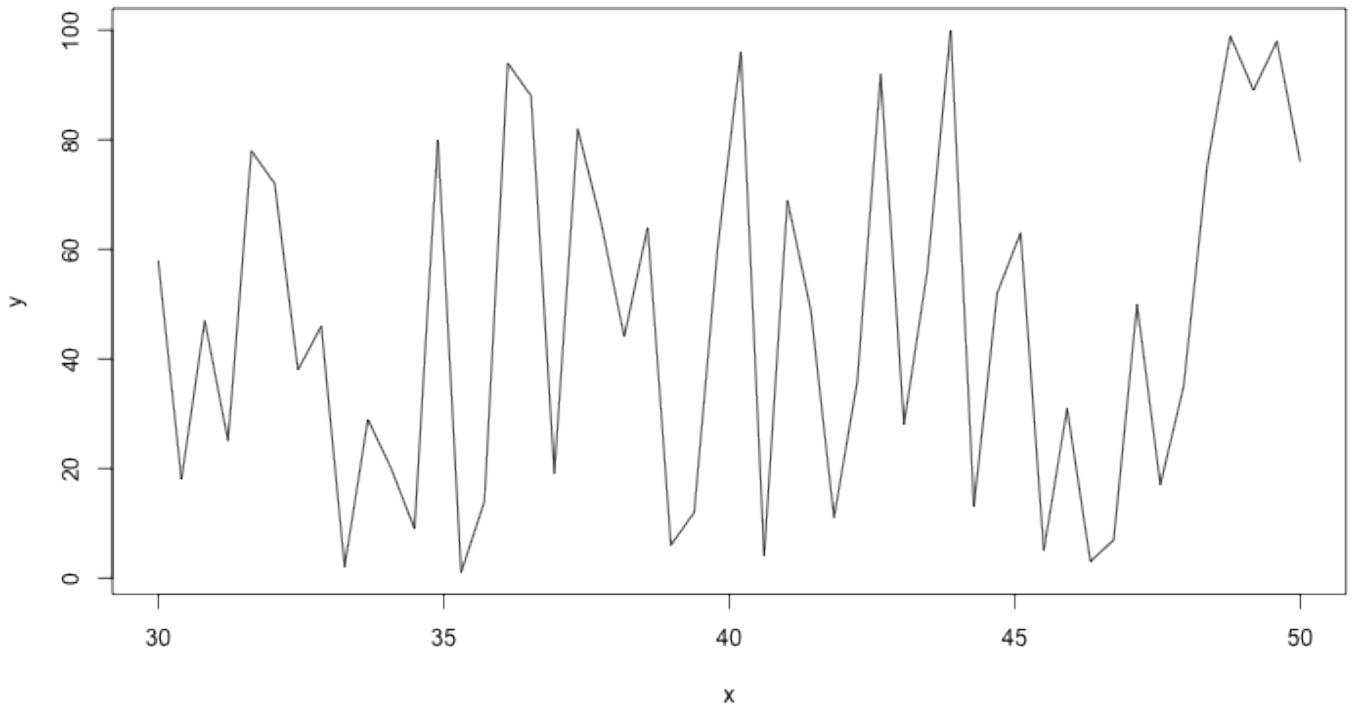
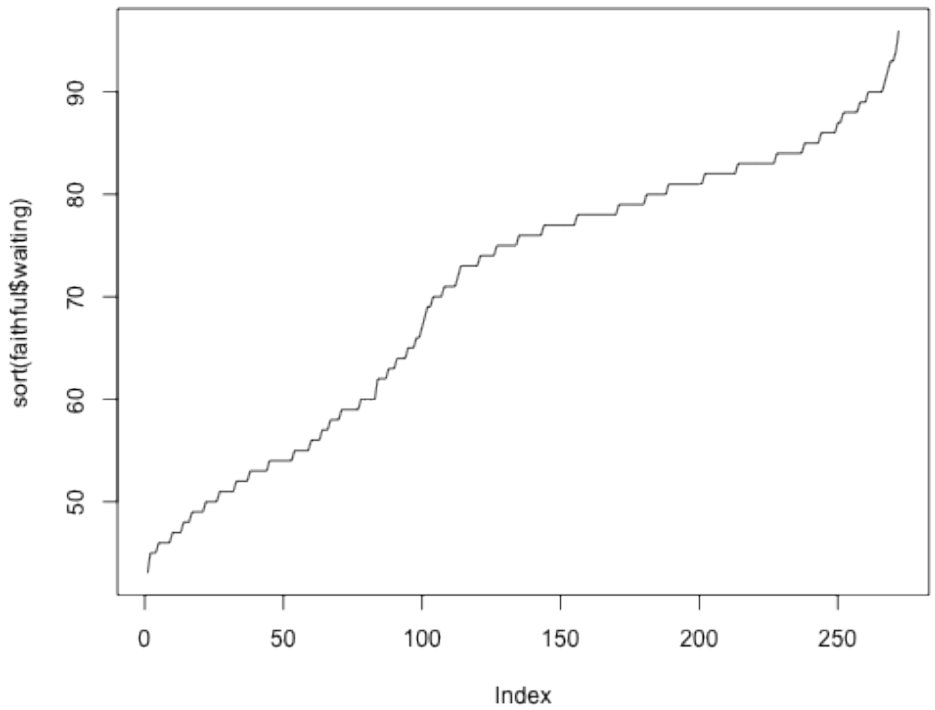


Grafico a linee dei valori ordinati

- Cf. sorted plot precedente

```
plot(sort(faithful$waiting), type = "l")
```



Serie temporale

- R ha una classe `ts` per manipolare serie temporali
- Se un oggetto ha classe `ts` allora `plot()` disegna un grafico a linee automaticamente
- Si può convertire esplicitamente un oggetto in una serie temporale con la funzione `as.ts()`

```
data(lynx)
class(lynx)
```

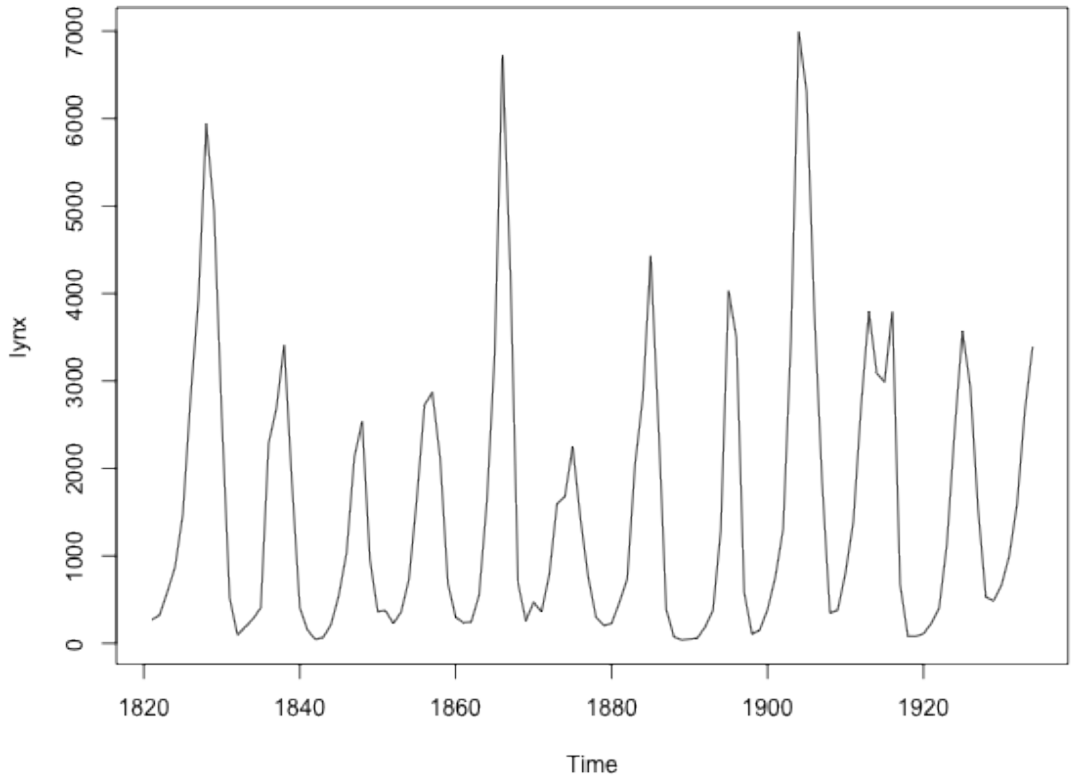
```
## [1] "ts"
```

```
str(lynx)
```

```
##  Time-Series [1:114] from 1821 to 1934: 269 321 585 871 1475 ...
```

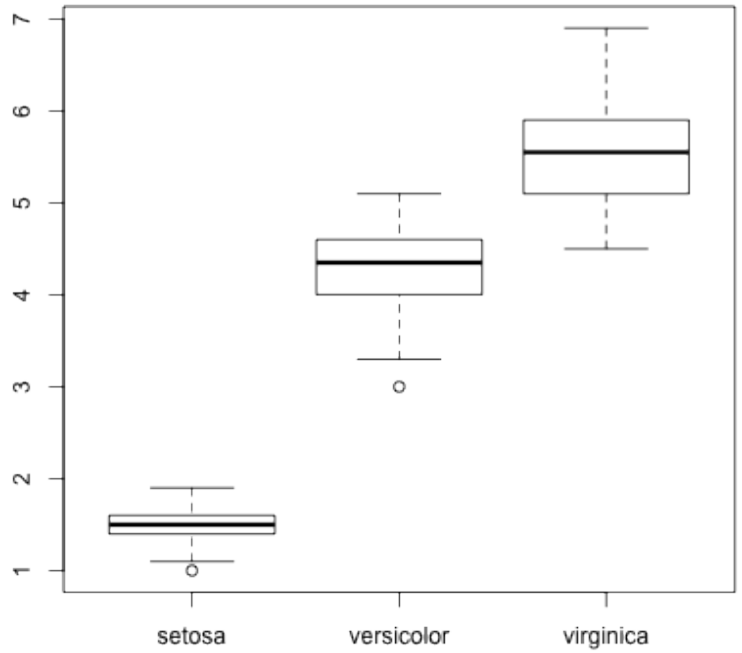
Serie temporale

```
plot(lynx)
```



Boxplot multipli

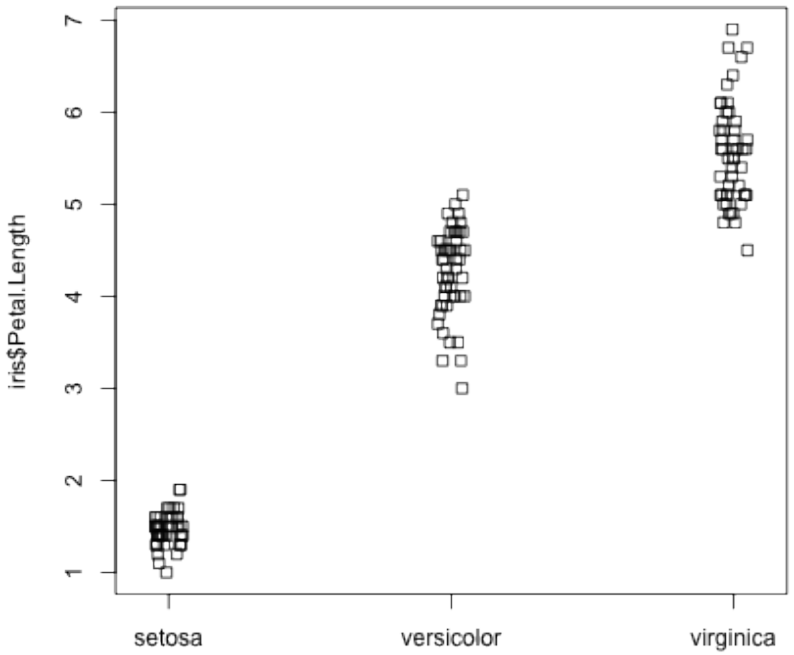
```
data(iris)
boxplot(iris$Petal.Length ~ iris$Species)  # Oppure, plot(iris$Species, iris$Petal.Length)
```



Strip chart

Alternativa al boxplot, utile specialmente per piccole quantità di dati

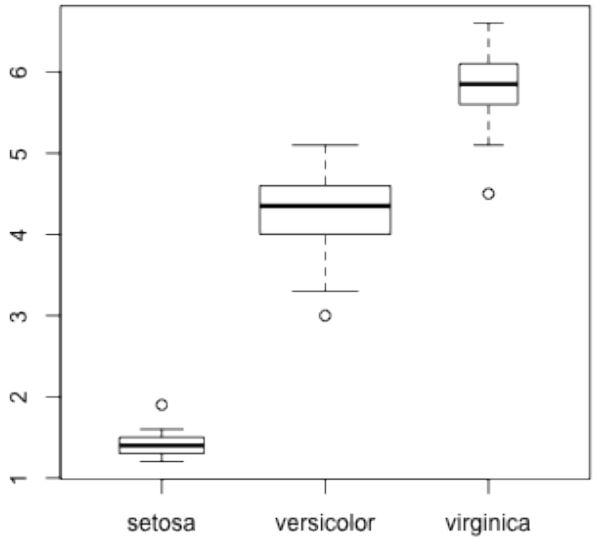
```
stripchart(iris$Petal.Length ~ iris$Species, vertical = T, method = "jitter",
    jitter = 0.05)
```



Boxplot: il parametro varwidth

Il parametro `varwidth` della funzione `boxplot()` consente di disegnare un boxplot la cui larghezza è proporzionale alla radice quadrata (perché?) del numero di osservazioni rappresentate dal boxplot:

```
iris_subset <- iris[30:110, ]
boxplot(iris_subset$Petal.Length ~ iris_subset$Species, varwidth = TRUE)
```



Aggiungete al data frame `iris` una colonna che consenta di distinguere tra fiori con petali “corti” e petali “lunghi”, scrivendo il seguente comando (usate `?cut` per l'aiuto sulla funzione `cut()`):

```
iris2 <- iris  # Crea una copia del data frame per non modificare l'originale
iris2$Petal.Type <- cut(iris2$Petal.Length, c(1, 4, 7))
```

Disegnate un grafico per confrontare le distribuzioni delle larghezze dei sepali per ciascun tipo di petalo (lungo o corto). Quale affermazione è corretta?

- La mediana della larghezza dei sepali è maggiore se il fiore ha petali lunghi.
- Sia la mediana sia la variabilità sono maggiori nei fiori con petali corti.
- La mediana e la variabilità sono all'incirca le stesse nei due casi.
- La mediana è all'incirca la stessa, ma la variabilità è maggiore nei fiori con petali corti.

Submit Show Hint Show Answer Clear

Grafico a barre semplice

- L'altezza di ciascuna barra è proporzionale alla frequenza del valore corrispondente
- Si possono usare anche per variabili quantitative discrete che assumono pochi valori distinti

```
barplot(table(iris_subset$Species))  # Oppure, plot(iris_subset$Species)
```

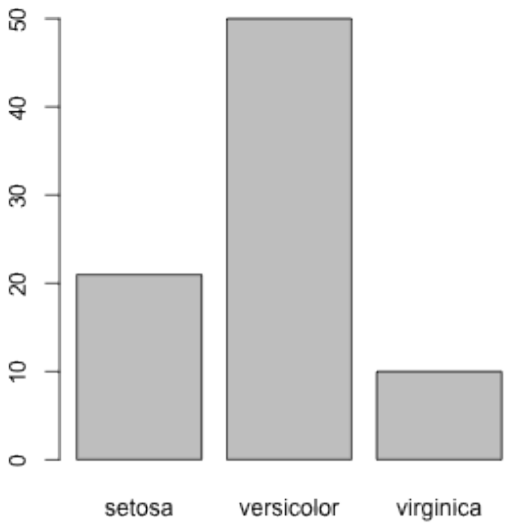
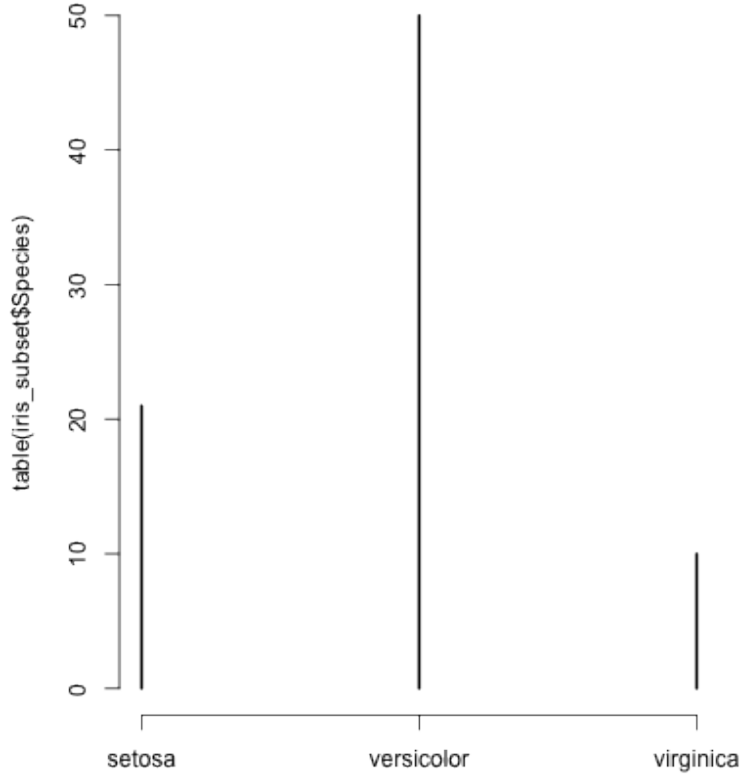


Grafico a bastoncino

```
plot(table(iris_subset$Species))
```



Grafici a barre da matrici

Esempio di matrice:

```
data(VADeaths)  # Morti per 1000 abitanti in Virginia nel 1940
class(VADeaths)
```

```
## [1] "matrix"
```

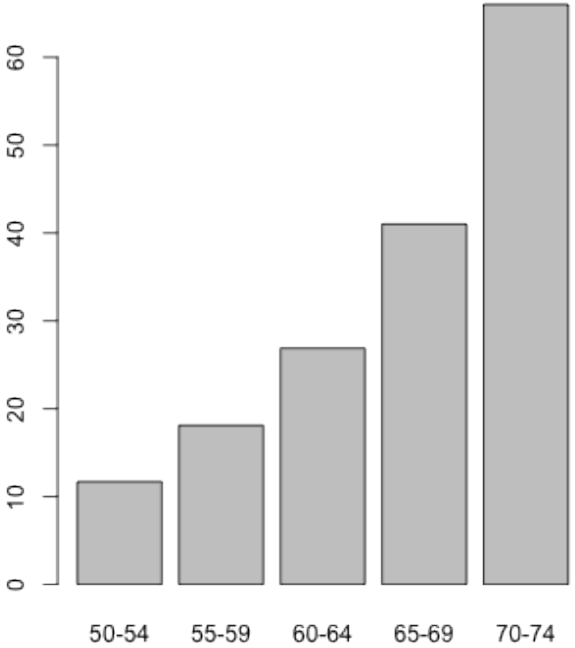
```
VADeaths
```

```
##       Rural Male Rural Female Urban Male Urban Female
## 50-54       11.7          8.7       15.4          8.4
## 55-59       18.1         11.7       24.3         13.6
## 60-64       26.9         20.3       37.0         19.3
## 65-69       41.0         30.9       54.6         35.1
## 70-74       66.0         54.3       71.1         50.0
```

Grafico a barre semplice da una matrice

Si ottiene selezionando una riga o una colonna della matrice

```
barplot(VADeaths[, "Rural Male"])
```




```
barplot(VADeaths["60-64", ])
```

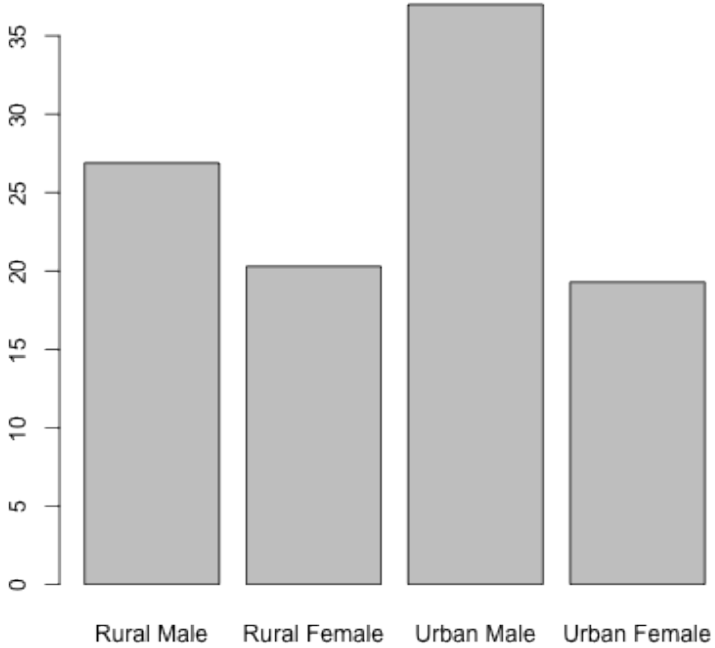
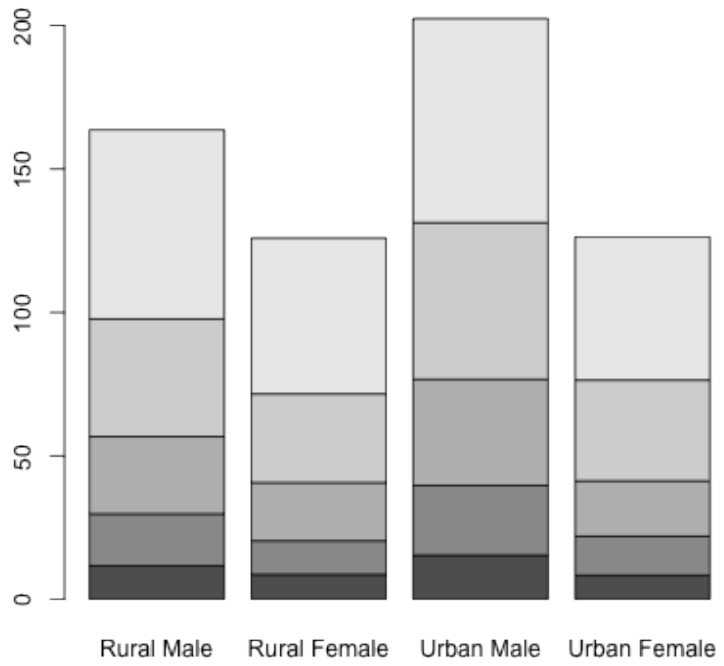


Grafico a barre segmentato

- Un grafico a barre per una matrice è automaticamente segmentato
- A volte è necessario trasporre la matrice per ottenere il risultato voluto

```
barplot(VADeaths)
```




```
barplot(t(VADeaths))
```

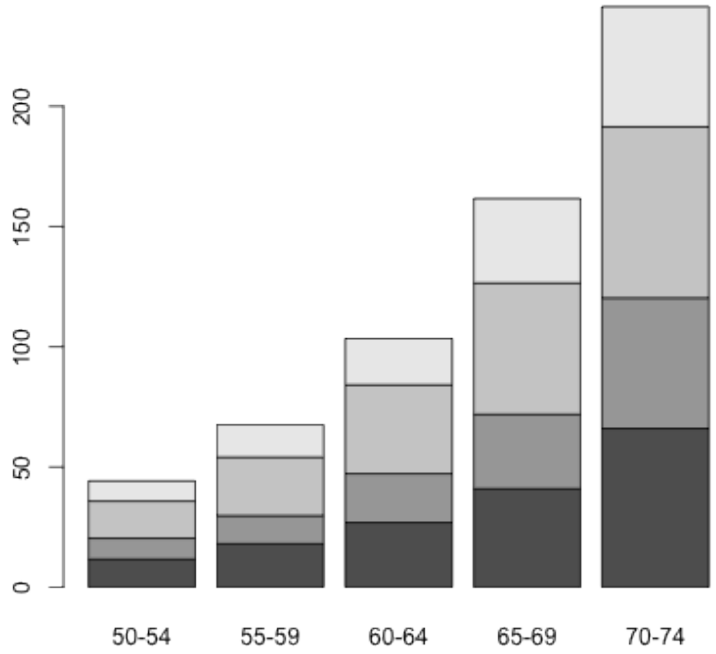
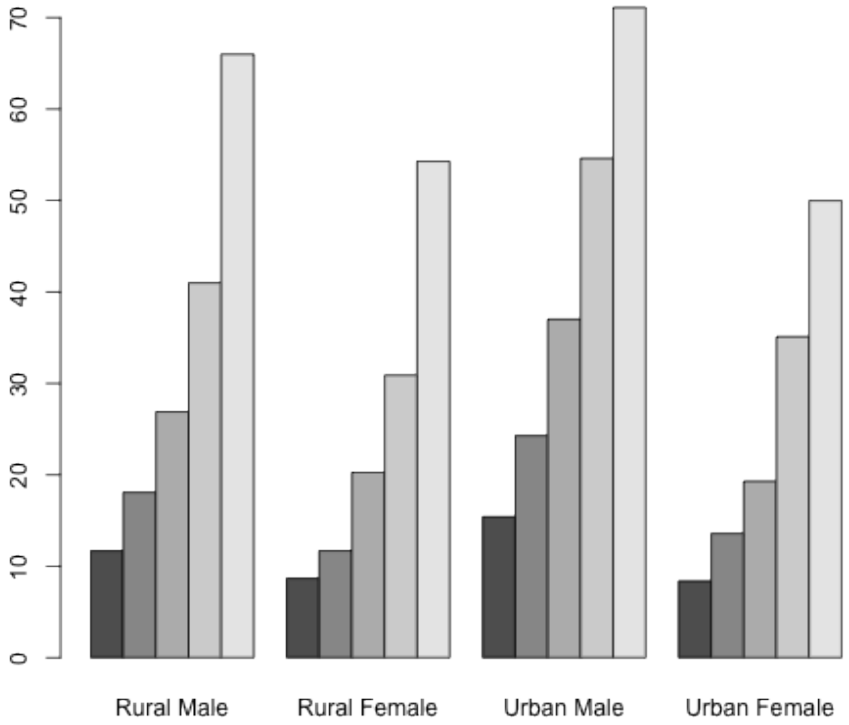


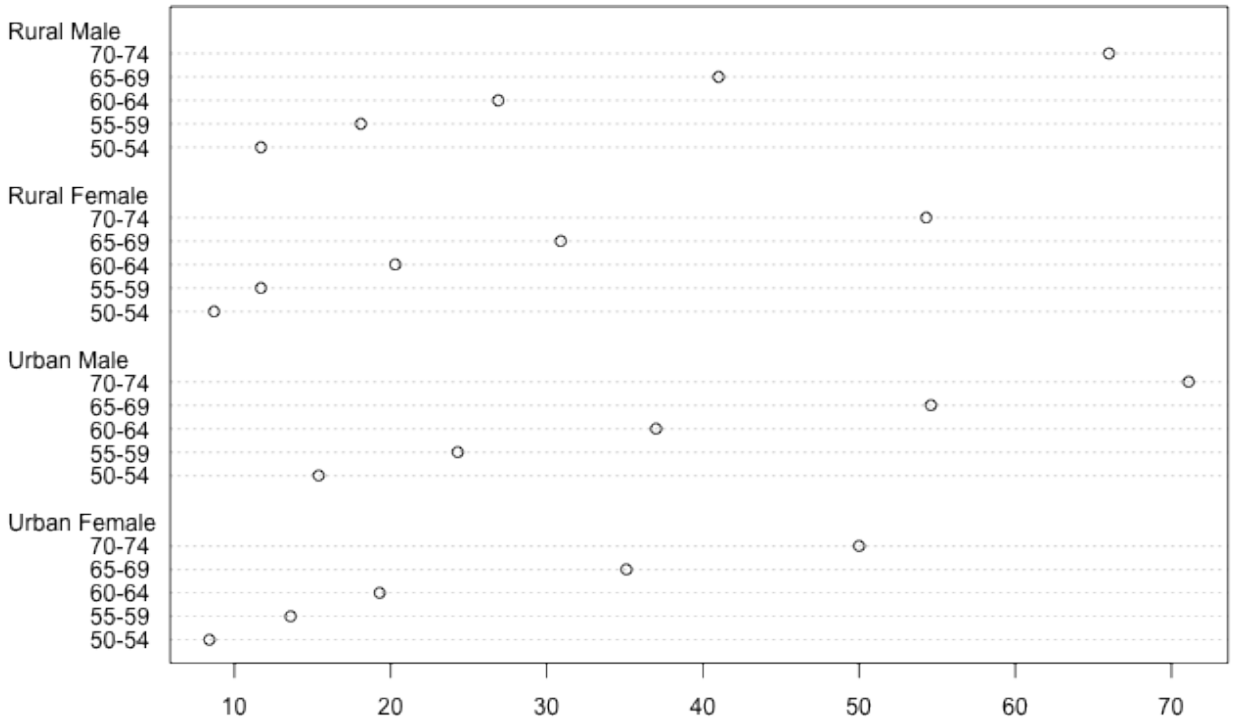
Grafico a barre multiplo

```
barplot(VADeaths, beside = TRUE)
```



Dot chart

```
dotchart(VADeaths)
```



Caricate in memoria il data frame `cars` scrivendo:

```
data(cars)
head(cars, 4)  # Esamina le prime 4 righe
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
```

Il comando `barplot(cars)` produce

- un grafico a barre semplice.
- tre grafici a barre semplici.
- un grafico a barre segmentato.
- un errore.

Submit Show Hint Show Answer Clear

Grafici per variabili categoriali

- Tipicamente, si vuole una rappresentazione grafica di una tabella di contingenza (classe `table`)
- Per una o due variabili, si possono usare i grafici a barre
- Per un numero arbitrario di variabili, un'alternativa è il **grafico a mosaico** (Hartigan e Kleiner, 1981)
- Un grafico a mosaico è ottenuto suddividendo una superficie rettangolare in un numero di rettangoli pari al numero di valori nella tabella di contingenza
- *L'area di ciascun rettangolo è proporzionale alla frequenza rappresentata*

```
##    A  B
## C  5 11
## D 21 15
```

Esempio: una tabella di contingenza a 4 vie

```
data(Titanic)
class(Titanic)
```

```
## [1] "table"
```

```
str(Titanic)
```

```
##  table [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
##  - attr(*, "dimnames")=List of 4
##   ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex     : chr [1:2] "Male" "Female"
##   ..$ Age     : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

Tabelle marginali

Tabella di contingenza per la seconda e quarta variabile:

```
margin.table(Titanic, c(2, 4))
```

```
##          Survived
## Sex         No  Yes
##   Male    1364  367
##   Female   126  344
```

Tabella di contingenza per la terza variabile:

```
margin.table(Titanic, 3)
```

```
## Age
## Child Adult
##   109  2092
```

Tabella di contingenza per la prima variabile:

```
margin.table(Titanic, 1)
```

```
## Class
##  1st  2nd  3rd Crew
##  325  285  706  885
```

Grafico a mosaico per due variabili categoriali

```
mosaicplot(margin.table(Titanic, c(1, 4)))
```

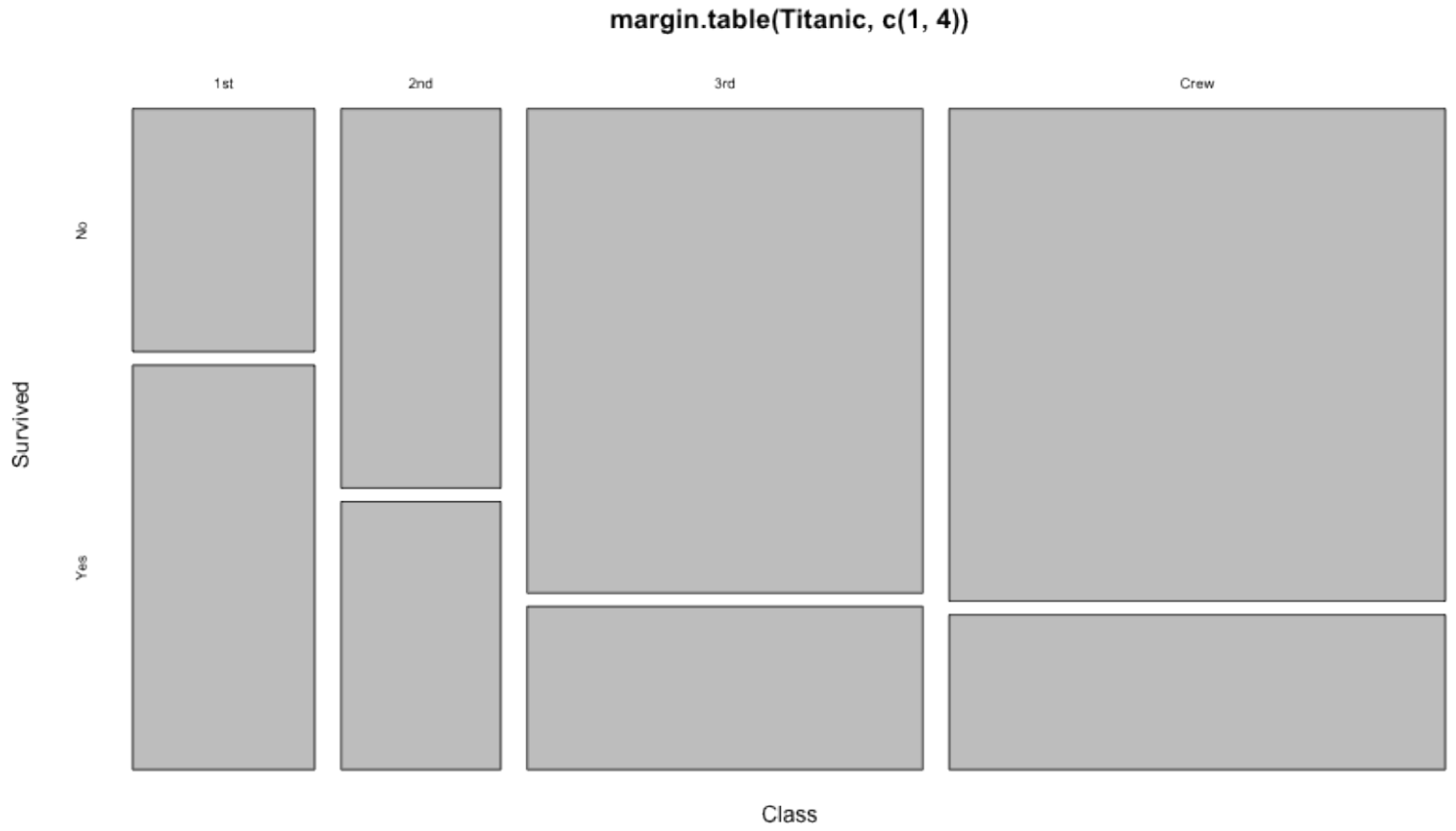
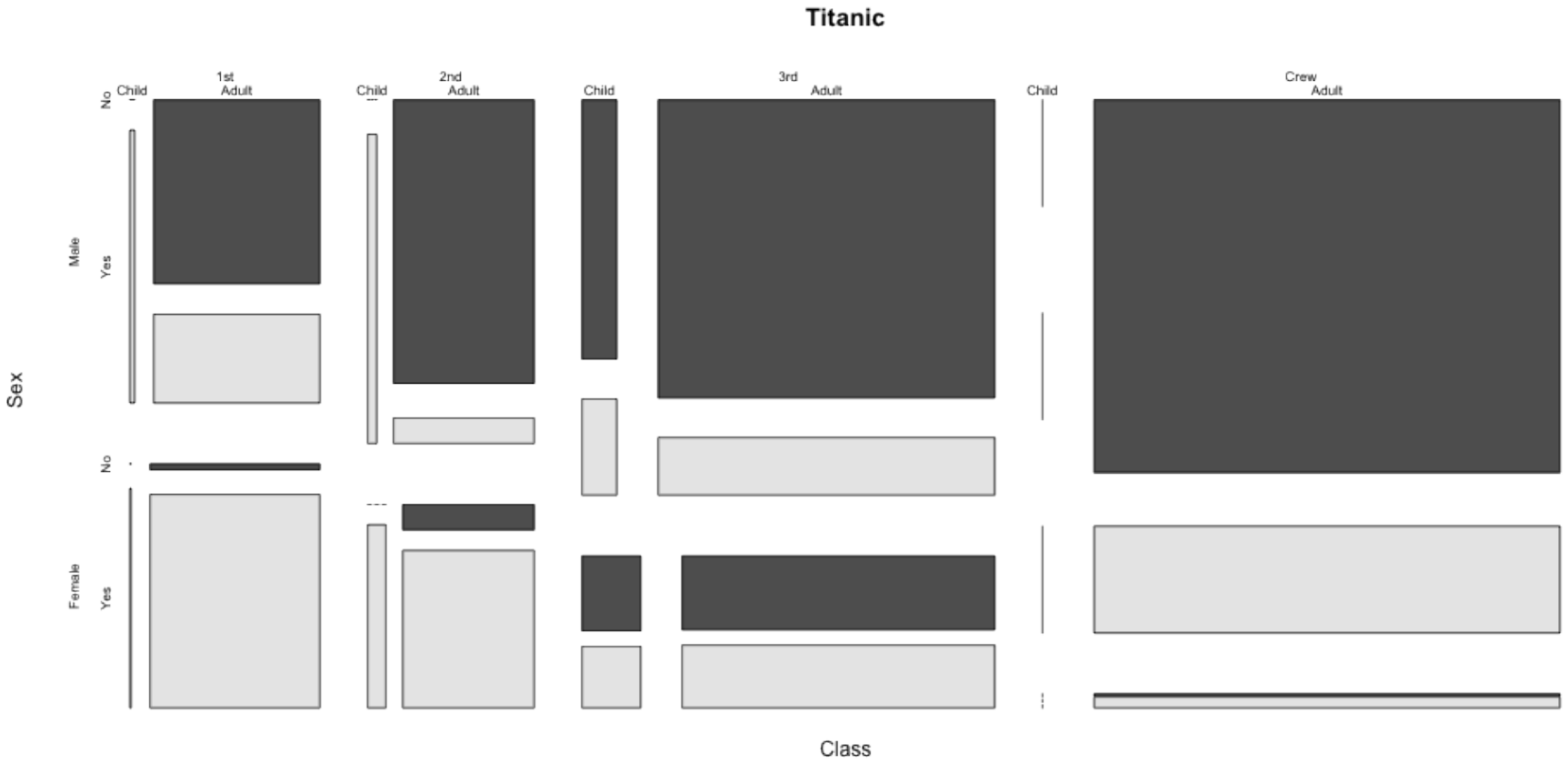


Grafico a mosaico multivariato

```
mosaicplot(Titanic, color = TRUE)
```



Costruzione manuale di un grafico a mosaico

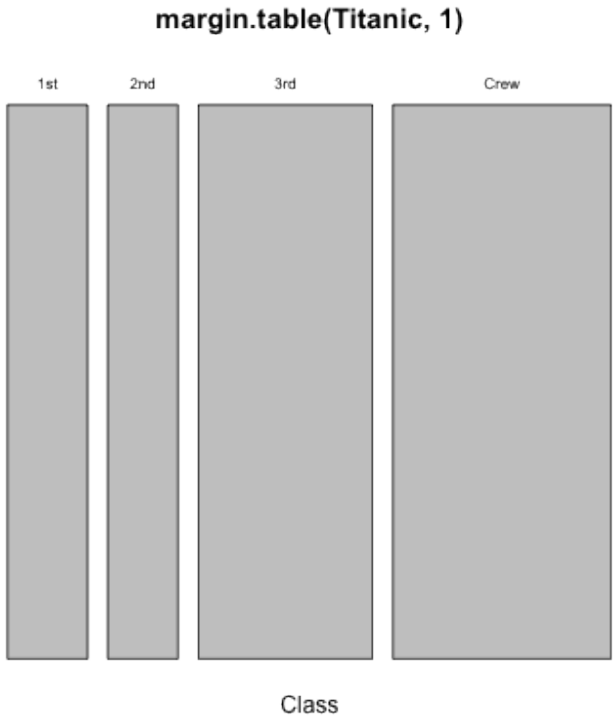
- Scegliere un ordinamento delle variabili (ad esempio, `Class`, `Sex`, `Age`, `Survived`)
- Dividere una superficie rettangolare in accordo alle frequenze marginali della prima variabile

```
T1 <- margin.table(Titanic, 1)
T1
```

```
## Class
##  1st  2nd  3rd Crew
##  325  285  706  885
```

```
T1/sum(T1)  # Freq. relative
```

```
## Class
##     1st     2nd     3rd    Crew
## 0.1477 0.1295 0.3208 0.4021
```



Costruzione manuale di un grafico a mosaico (cont.)

- Dividere ciascuno dei rettangoli ottenuti mediante l'operazione precedente lungo l'asse ortogonale a quello precedente in accordo alle frequenze marginali della variabile successiva nell'ordinamento scelto

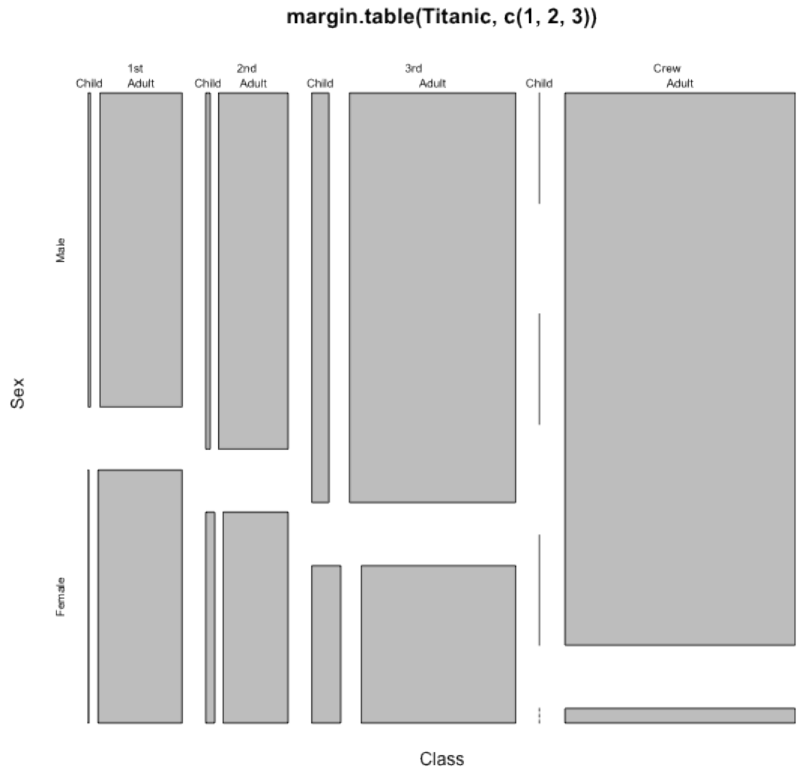
```
T2 <- margin.table(Titanic, c(1, 2))
  T2/rowSums(T2)  # Freq. relative
```

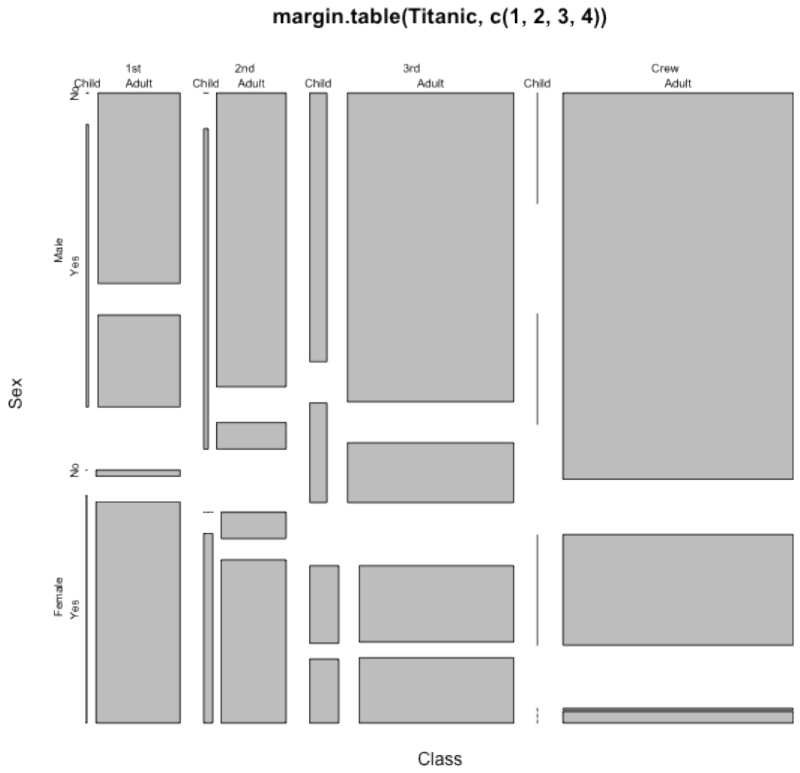
```
##       Sex
## Class     Male  Female
##   1st  0.55385 0.44615
##   2nd  0.62807 0.37193
##   3rd  0.72238 0.27762
##   Crew 0.97401 0.02599
```

```
mosaicplot(margin.table(Titanic, c(1, 2)))
```

Costruzione manuale di un grafico a mosaico (cont.)

- Ripetere l'operazione nella slide precedente finché non si sono esaurite tutte le variabili





Caricate il data set `Titanic` con il comando `data(Titanic)`. Disegnate un mosaico delle variabili `Sex` e `Survived`. Solamente osservando il grafico (senza stampare la tabella di contingenza), quale delle seguenti osservazioni si può fare?

- I sopravvissuti sono distribuiti in modo piú o meno uguale tra maschi e femmine.
- Il tasso di superstiti è piú alto tra le donne, ma le donne sono in valore assoluto meno degli uomini.
- Il tasso di superstiti è piú alto tra le donne, ma dal grafico non si può capire se le donne sono in valore assoluto di meno.
- Non è possibile confrontare il tasso di sopravvissuti tra maschi e femmine perché le aree sono diverse.

Submit Show Hint Show Answer Clear

Disegnate un mosaico delle variabili `Age`, `Class` e `Survived`. Quale delle seguenti affermazioni **non** è corretta?

- La seconda classe ha il minor numero di passeggeri.
- In ciascuna classe, il numero di bambini sopravvissuti è piú alto dei bambini non sopravvissuti.
- Il numero di bambini in prima classe è minore del numero di bambini in terza classe.
- Tra gli adulti, vi sono piú sopravvissuti tra i membri dell'equipaggio che non tra i passeggeri in terza classe.

Submit Show Hint Show Answer Clear