Working with R data

Michael Friendly

2022-12-13

Data in R packages

```
data()               # list data in the datasets package
data(package="vcd")  # in the vcd package
```

typically, load data from a package using data()

```
data(UCBAdmissions)
str(UCBAdmissions)
```

```
##  'table' num [1:2, 1:2, 1:6] 512 313 89 19 353 207 17 8 120 205 ...
##  - attr(*, "dimnames")=List of 3
##   ..$ Admit : chr [1:2] "Admitted" "Rejected"
##   ..$ Gender: chr [1:2] "Male" "Female"
##   ..$ Dept  : chr [1:6] "A" "B" "C" "D" ...
```

```
sum(UCBAdmissions)
```

```
## [1] 4526
```

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

```
## Admit
## Admitted Rejected
##     1755     2771
```

```
margin.table(UCBAdmissions, 2:3)
```

```
##         Dept
## Gender     A   B   C   D   E   F
##   Male   825 560 325 417 191 373
##   Female 108  25 593 375 393 341
```

data frames

```
library(vcd)         # load the vcd package & make its datasets available
```

```
## Loading required package: grid
```

```
data(Arthritis)
str(Arthritis)
```

```
## 'data.frame':    84 obs. of  5 variables:
##  $ ID       : int  57 46 77 17 36 23 75 39 33 55 ...
##  $ Treatment: Factor w/ 2 levels "Placebo","Treated": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Sex      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
```

```
##  $ Age      : int  27 29 30 32 46 58 59 59 63 63 ...
##  $ Improved : Ord.factor w/ 3 levels "None"<"Some"<..: 2 1 1 3 3 3 1 3 1 1 ...
```

```
head(Arthritis)      # see the first few lines
```

```
##   ID Treatment  Sex Age Improved
## 1 57   Treated Male  27     Some
## 2 46   Treated Male  29     None
## 3 77   Treated Male  30     None
## 4 17   Treated Male  32   Marked
## 5 36   Treated Male  46   Marked
## 6 23   Treated Male  58   Marked
```

making tables from data frame variables

```
table(Arthritis$Improved)
```

```
##
##   None   Some Marked
##     42     14     28
```

```
table(Arthritis$Treatment, Arthritis$Sex)
```

```
##
##           Female Male
##   Placebo     32   11
##   Treated     27   14
```

```
with(Arthritis, table(Treatment, Sex))
```

```
##          Sex
## Treatment Female Male
##   Placebo     32   11
##   Treated     27   14
```

xtabs() is often easier

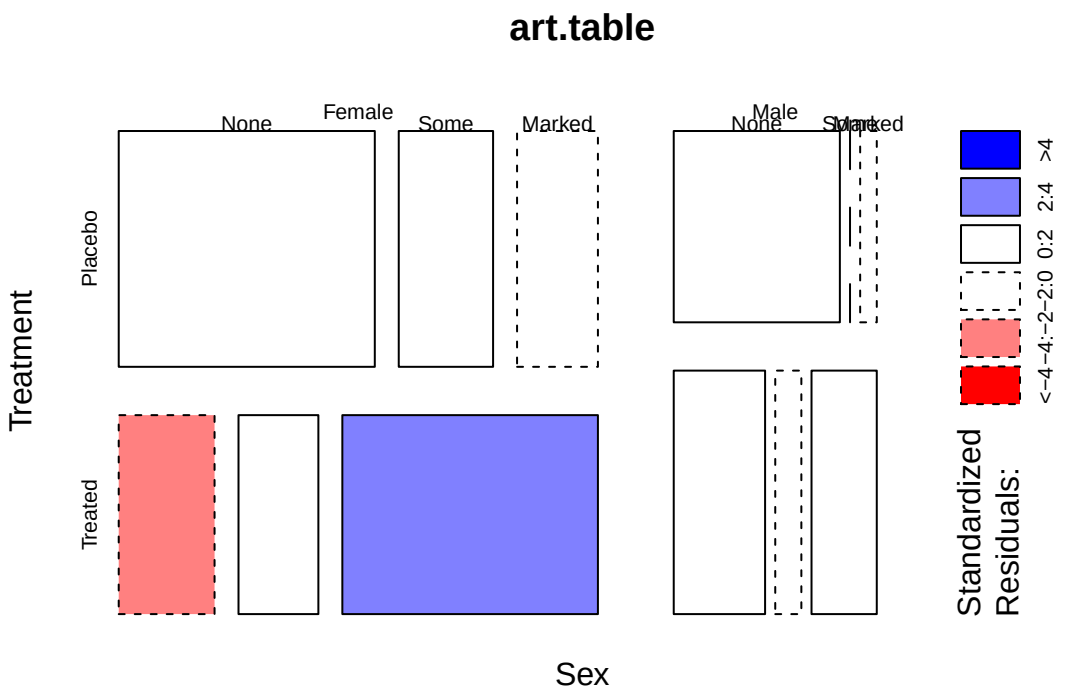
```
art.table <- xtabs(~ Sex + Treatment + Improved, data=Arthritis)
ftable(art.table)    # display as flattened table
```

```
##                 Improved None Some Marked
## Sex    Treatment
## Female Placebo              19    7      6
##        Treated               6    5     16
## Male   Placebo              10    0      1
##        Treated               7    2      5
```

```
summary(art.table)   # chi-square test for mutual independence
```

```
## Call: xtabs(formula = ~Sex + Treatment + Improved, data = Arthritis)
## Number of cases in table: 84
## Number of factors: 3
## Test for independence of all factors:
##  Chisq = 19.6, df = 7, p-value = 0.006501
##  Chi-squared approximation may be incorrect
```

```
plot(art.table, shade=TRUE)
```



Reading data from external files

read a data table from a local file (NB: '/' not '\' for all systems)

```
# arthritis <- read.csv("N:/psy6136/data/arthritis.csv")
# arthritis <- read.csv(file.choose())
# or, read the same data from a web URL ...
arthritis <- read.csv("https://raw.githubusercontent.com/friendly/psy6136/master/data/Arthritis.csv")

levels(arthritis$Improved)
```

```
## NULL
```

make an ordered factor

```
arthritis$Improved <- ordered(arthritis$Improved,
                              levels=c("None", "Some", "Marked"))
```

make a variable discrete: cut and arithmetic

```
table(10*floor(arthritis$Age/10))
```

```
##
## 20 30 40 50 60 70
##  4 11 11 29 26  3
```

```
table(cut(arthritis$Age, breaks=6))
```

```
##
```

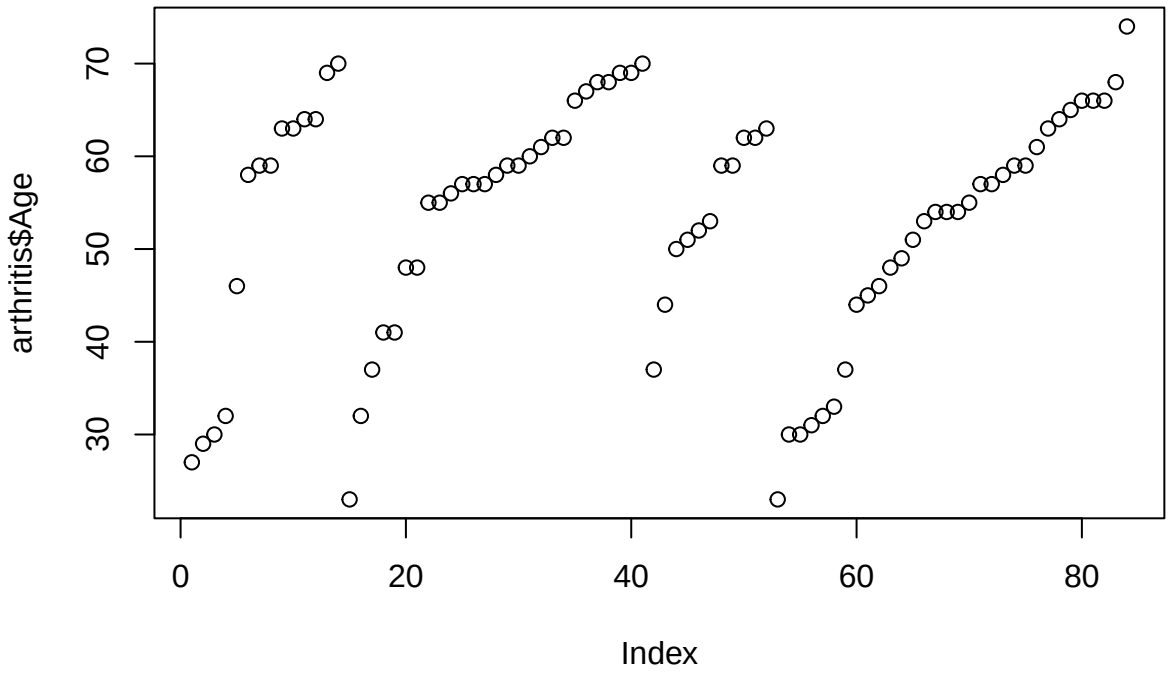
```
## (22.9,31.5]   (31.5,40]   (40,48.5]   (48.5,57]   (57,65.5] (65.5,74.1]
##           8           7          10          19          26          14
```

assign new variable in data frame

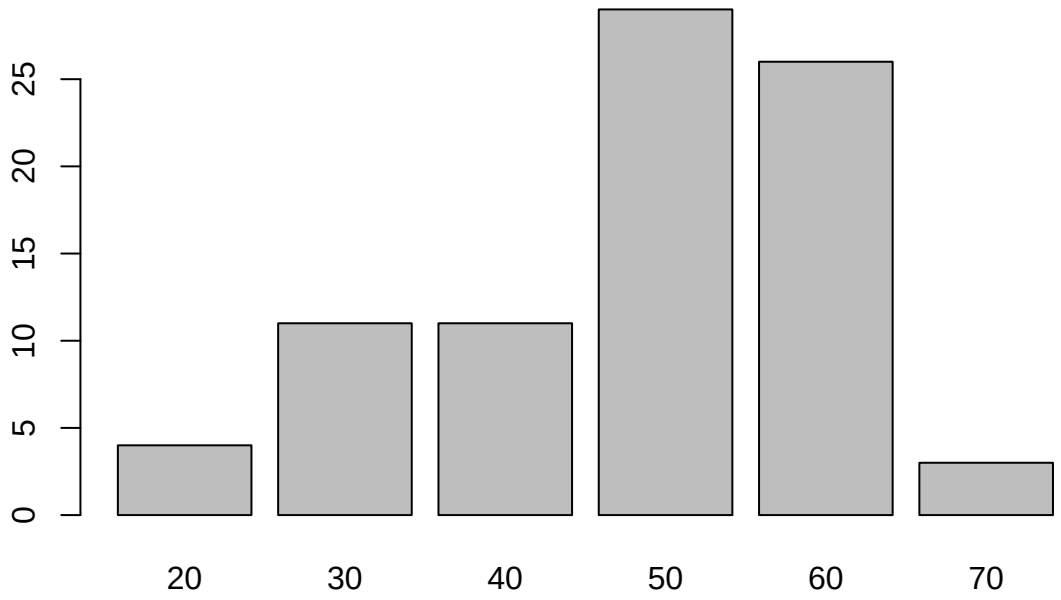
```
arthritis$AgeGroup <- factor(10*floor(arthritis$Age/10))
```

simple plots

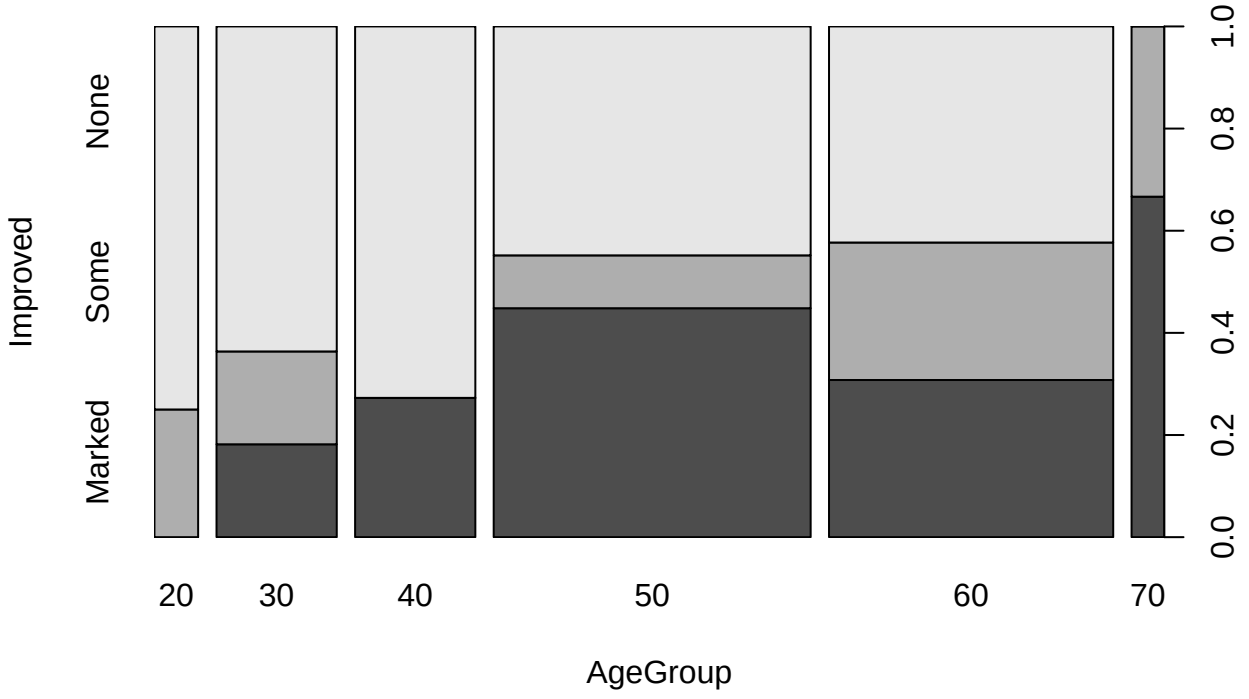
```
plot(arthritis$Age)       # index plot
```



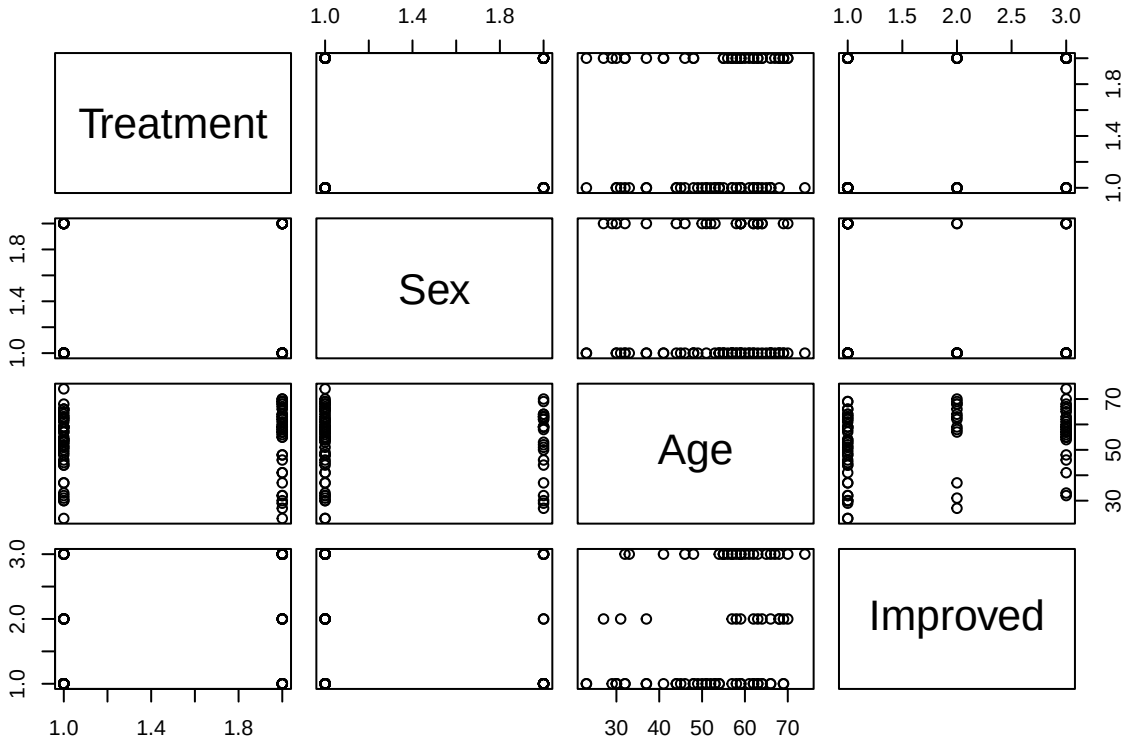

```
plot(arthritis$AgeGroup) # barplot for a factor
```




```
plot(Improved ~ AgeGroup, data=arthritis) # spineplot for two factors
```




```
plot(arthritis[,2:5])     # scatterplot matrix; not too useful for factors
```


Treatment
Sex
Age
Improved
1.0
1.4
1.8
1.0
1.5
2.0
2.5
3.0
1.0
1.4
1.8
1.0
1.4
1.8
30
50
70
1.0
2.0
3.0
1.0
1.4
1.8
30
40
50
60
70