Package ‘openintro’

Type Package

Title Data Sets and Supplemental Functions from ‘OpenIntro’ Textbooks

Version 1.7.1

Date 2017-09-05

Author David M Diez, Christopher D Barr, and Mine Cetinkaya-Rundel

Maintainer David M Diez <david@openintro.org>

Description Supplemental functions and data for ‘OpenIntro’ resources, which includes open-source textbooks and resources for introductory statistics at <http://www.openintro.org>. The package contains data sets used in our open-source textbooks along with custom plotting functions for reproducing book figures. Note that many functions and examples include color transparency; some plotting elements may not show up properly (or at all) when run in some versions of Windows operating system.

License GPL-2 | GPL-3

LazyLoad yes

LazyData yes

URL http://www.openintro.org/

Depends graphics, grDevices, stats, utils, R (>= 2.10)

NeedsCompilation no

Repository CRAN

Date/Publication 2017-09-08 03:55:44 UTC

R topics documented:

openintro-package ......................................................... 5
abbr2state ....................................................................... 7
acs12 ............................................................................... 8
ageAtMar .......................................................................... 10
ami.occurrences ............................................................... 10
ArrowLines ........................................................................ 11
association.1.3 ................................................................. 12
association.4.6 ................................................................. 13
R topics documented:

association.7.12 ........................................ 13
assortive.mating ....................................... 14
avandia .................................................. 14
babies .................................................... 15
babies.crawl ............................................ 16
bac ....................................................... 17
ballBearing ............................................. 17
bdims ..................................................... 18
births ..................................................... 20
books ..................................................... 21
boxPlot .................................................. 21
Braces ................................................... 23
buildAxis .............................................. 24
burger .................................................... 26
cancer.in.dogs ......................................... 27
cards ..................................................... 28
cars ..................................................... 28
cats ...................................................... 30
ccHousing ................................................ 30
CCP ...................................................... 31
census ................................................... 32
cherry ................................................... 33
chick.wts ............................................... 34
china .................................................... 34
ChiSquareTail ......................................... 35
classData ............................................... 36
COL ..................................................... 36
contTable ............................................... 37
corr.match ............................................ 38
county ................................................... 39
county.w.sm.ban ...................................... 40
countyComplete ....................................... 42
cpr ....................................................... 44
credits .................................................. 44
CT2DF ................................................... 45
densityPlot ............................................. 46
diamonds ............................................... 48
dlsegments ............................................. 49
dotPlot .................................................. 50
dotPlotStack .......................................... 52
dream .................................................... 53
drug.use ............................................... 53
ebola.survey .......................................... 54
edaPlot .................................................. 55
elmhurst ................................................ 55
email ..................................................... 56
email50 .................................................. 58
epa2012 .................................................. 60
### R topics documented:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>exams</td>
<td>61</td>
</tr>
<tr>
<td>fadeColor</td>
<td>62</td>
</tr>
<tr>
<td>fcid</td>
<td>63</td>
</tr>
<tr>
<td>fheights</td>
<td>64</td>
</tr>
<tr>
<td>friday</td>
<td>64</td>
</tr>
<tr>
<td>full.body.scan</td>
<td>65</td>
</tr>
<tr>
<td>gear.company</td>
<td>66</td>
</tr>
<tr>
<td>gender.discrimination</td>
<td>66</td>
</tr>
<tr>
<td>gifted</td>
<td>67</td>
</tr>
<tr>
<td>global.warming.pew</td>
<td>68</td>
</tr>
<tr>
<td>goog</td>
<td>69</td>
</tr>
<tr>
<td>gov.poll</td>
<td>69</td>
</tr>
<tr>
<td>govRace10</td>
<td>70</td>
</tr>
<tr>
<td>gpa</td>
<td>71</td>
</tr>
<tr>
<td>gpa_study_hours</td>
<td>72</td>
</tr>
<tr>
<td>gradesTV</td>
<td>73</td>
</tr>
<tr>
<td>gsearch</td>
<td>74</td>
</tr>
<tr>
<td>healthcare.law.survey</td>
<td>74</td>
</tr>
<tr>
<td>healthcare.survey</td>
<td>75</td>
</tr>
<tr>
<td>health_coverage</td>
<td>76</td>
</tr>
<tr>
<td>heartTr</td>
<td>76</td>
</tr>
<tr>
<td>helium</td>
<td>77</td>
</tr>
<tr>
<td>histPlot</td>
<td>78</td>
</tr>
<tr>
<td>house</td>
<td>80</td>
</tr>
<tr>
<td>houseRace10</td>
<td>81</td>
</tr>
<tr>
<td>housing</td>
<td>83</td>
</tr>
<tr>
<td>hsb2</td>
<td>83</td>
</tr>
<tr>
<td>husbands.wives</td>
<td>84</td>
</tr>
<tr>
<td>immigration</td>
<td>85</td>
</tr>
<tr>
<td>infMortRate</td>
<td>86</td>
</tr>
<tr>
<td>ipod</td>
<td>87</td>
</tr>
<tr>
<td>iris</td>
<td>87</td>
</tr>
<tr>
<td>jury</td>
<td>88</td>
</tr>
<tr>
<td>leg_mari</td>
<td>88</td>
</tr>
<tr>
<td>linResPlot</td>
<td>89</td>
</tr>
<tr>
<td>lmPlot</td>
<td>90</td>
</tr>
<tr>
<td>loop</td>
<td>92</td>
</tr>
<tr>
<td>lsegments</td>
<td>92</td>
</tr>
<tr>
<td>makeTube</td>
<td>93</td>
</tr>
<tr>
<td>male.heights</td>
<td>95</td>
</tr>
<tr>
<td>mammals</td>
<td>96</td>
</tr>
<tr>
<td>mammogram</td>
<td>97</td>
</tr>
<tr>
<td>marathon</td>
<td>98</td>
</tr>
<tr>
<td>marioKart</td>
<td>98</td>
</tr>
<tr>
<td>migraine</td>
<td>100</td>
</tr>
<tr>
<td>MLB</td>
<td>101</td>
</tr>
<tr>
<td>mlbBat10</td>
<td>102</td>
</tr>
<tr>
<td>MosaicPlot</td>
<td>104</td>
</tr>
</tbody>
</table>
myPDF  ......................................................... 105
nba.heights .............................................. 106
ncbirths .................................................. 107
normTail .................................................. 108
nuclear.survey .......................................... 109
offshore.drilling ...................................... 110
orings ..................................................... 110
oscars ..................................................... 111
outliers .................................................... 112
PlotWLine ................................................ 113
pm25.2011.durham ..................................... 114
poker ....................................................... 115
possum ..................................................... 115
president .................................................. 116
prof.evaltns.beauty.public ......................... 117
prRace08 .................................................. 120
res.demo.1 ............................................... 121
res.demo.2 ............................................... 122
run10 ....................................................... 122
sat.improve .............................................. 124
satGPA ..................................................... 124
scotus_healthcare ..................................... 125
senateRace10 ............................................. 126
sinusitis ................................................... 127
sleep.deprivation ...................................... 128
smoking .................................................... 128
sp500.seq ............................................... 130
speed.gender.height ................................. 130
stanford ................................................... 131
starbucks .................................................. 132
stats.scores .............................................. 132
stem.cell .................................................. 133
stem.cells ............................................... 134
stent30 .................................................... 135
stent365 .................................................. 135
student.housing ....................................... 136
student.sleep .......................................... 137
sulphinpyrazone ...................................... 137
supreme.court .......................................... 138
textbooks ............................................... 139
tgSpending .............................................. 140
tips ......................................................... 140
toohey ...................................................... 142
tourism ..................................................... 142
toy.anova .................................................. 143
transplant ............................................... 143
treeDiag ................................................... 144
trees ....................................................... 145
Description

This package is a supplement to OpenIntro Statistics, which is a free textbook available at openintro.org (at-cost paperbacks are also available for under $10 on Amazon). The package contains data sets used in the textbook along with custom plotting functions for reproducing book figures. Note that many functions and examples include color transparency. Some plotting elements may not show up properly (or at all) in some Windows versions.

Details

Package: openintro
Type: Package
Version: 1.5.0
Date: 2013-10-28
License: GPL-2 | GPL-3
LazyLoad: yes

boxPlot, buildAxis, densityPlot, dotPlot, edaPlot, histPlot, normTail, cars, marioKart, possum, runQP, satGPA, textbooks

Some colors include transparency, which means they will not be plotted in some operating systems (e.g. Windows). However, the plots may be viewed if they are written to a PDF or PNG file first. For a discussion of this topic, please see http://yihui.name/en/2007/09/semi-transparent-colors-in-r-color-image-as-an-example/

Two new functions, myPDF and myPNG, were created in this package and may also be used to set up nice plotting files that allow for transparency.

Author(s)

David M Diez, Christopher D Barr, Mine Cetinkaya-Rundel
Maintainer: DM Diez <david.m.diez@gmail.com>
Examples

```r
### boxPlot example 1
data(run10)
par(mfrow=c(2,2))
boxplot(run10$time)
boxplot(run10$time)

### histPlot example 1
histPlot(run10$time[run10$gender == 'M'], probability=TRUE, xlim=c(30, 180), ylim=c(0, 0.025), hollow=TRUE)
histPlot(run10$time[run10$gender == 'F'], probability=TRUE, add=TRUE, hollow=TRUE, lty=1)
legend('topleft', col=c('black', 'red'), lty=2:3, legend=c('M', 'F'))
histPlot(run10$time, col=fadeColor('yellow', '33'), border='darkblue', probability=TRUE, breaks=seq(30, 180)
histPlot(run10$time, probability=TRUE, breaks=brks, col=fadeColor('darkgoldenrod4', '33'))

### histPlot example 2
data(cars)
histPlot(cars$price[cars$type == 'small'], probability=TRUE, hollow=TRUE, xlim=c(0, 50))
histPlot(cars$price[cars$type == 'midsize'], probability=TRUE, hollow=TRUE, add=TRUE, border='red', lty=3)
histPlot(cars$price[cars$type == 'large'], probability=TRUE, hollow=TRUE, add=TRUE, border='blue', lty=4)
legend('topright', lty=2:4, col=c('black', 'red', 'blue'), legend=c('small', 'midsize', 'large'))

### densityPlot example
data(tips)
densityPlot(tips$tip, tips$day)
legend('topright', col=c('black', 'red'), lty=1:2, legend=c('Tuesday', 'Friday'))

### identifying reasons for outliers
data(marioKart)
par(mfrow=c(1,1))
boxPlot(marioKart$totalPr, marioKart$cond, horiz=TRUE)
these <- which(marioKart$totalPr > 80)
# see the data collection criteria for
# why these observations do not belong.
lines(rep(marioKart$totalPr[these[1]], 2), c(2.4, 2))
text(marioKart$totalPr[these[1]], 2.4, marioKart$title[these[1]],
```
pos=3, cex=0.5)  
lines(rep(marioKart$totalPr[these[2]], 2), c(1.6, 2))  
text(marioKart$totalPr[these[2]], 1.6, marioKart$title[these[2]],  
pos=1, cex=0.5)  

### compare plotting methods <===#  
data(cars)  
par(mfrow=c(1,1))  
histPlot(cars$price, ylim=c(0, 0.1), axes=FALSE, ylab='',  
probability=TRUE, xlab='price')  
axis(1)  
boxPlot(cars$price, width=0.03, horiz=TRUE, add=0.067, axes=FALSE)  
dotPlot(cars$price, at=0.095, add=TRUE)  
densityPlot(cars$price, add=TRUE)  

### controlling the number of axis labels <===#  
# specify the number of labels  
data(textbooks)  
x <- textbooks$diff  
par(mfrow=c(3,1))  
histPlot(x, axes=FALSE)  
buildAxis(x, n=4, nMin=4, nMax=4)  
histPlot(x, axes=FALSE)  
buildAxis(x, n=5, nMin=5, nMax=5)  
histPlot(x, axes=FALSE)  
# no decent axis is found for this data with exactly six labels  
# no min or max specified, only a target number of labels:  
buildAxis(x, n=6)  

### creating normal plots with tails <===#  
par(mfrow=c(2,3), mar=c(3,3,1,1), mgp=c(1.7, 0.7, 0))  
normTail(L=-2)  
normTail(U=1, xLab='symbol', cex.axis=0.7)  
normTail(M=c(-2, -0.3), col='#22558833')  
normTail(5, 13, L=-5, M=c(0,3), U=12, xAxisIncr=2)  
normTail(102, 4, xLim=c(97,110), M=c(100,103))  
normTail(-10.0, 5.192, M=c(-5,2), digits=1, xAxisIncr=2)  

### Exploratory Data Analysis Plot <===#  
data(mlbBat10)  
#edaPlot(mlbBat10)

---

**Description**

Two utility functions. One converts state names to the state abbreviations, and the second does the opposite.
Usage

abbr2state(abbr)

state2abbr(state)

Arguments

state A vector of state name, where there is a little fuzzy matching.
abbr A vector of state abbreviation.

Value

Returns a vector of the same length with the corresponding state names or abbreviations.

Author(s)

David Diez

See Also

county, countyComplete

Examples

state2abbr("Minnesota")
abbr2state("MN")

#_____ Some Spelling/Capitalization Errors Okay _____#
state2abbr("mInnesta")

acs12 American Community Survey, 2012

Description

Results from the US Census American Community Survey, 2012.

Usage

data("acs12")
Format

A data frame with 2000 observations on the following 13 variables.

income  Annual income.
employment Employment status.
hrs_work Hours worked per week.
race Race.
age Age, in years.
gender Gender.
citizen Whether the person is a U.S. citizen.
time_to_work Travel time to work, in minutes.
lang Language spoken at home.
marrried Whether the person is married.
edu Education level.
disability Whether the person is disabled.
birth_qrtr The quarter of the year that the person was born, e.g. Jan thru Mar.

Source

http://www.census.gov/acs

Examples

data(acs12)
d <- subset(acs12, 30 <= age & age <= 60)
d <- subset(d, employment == "employed" & income > 0)
d <- na.omit(d[c("age", "income")])

par(mfrow = c(2, 1), mar = c(5, 5, 1, 1))
plot(d$age, d$income)
m <- lm(income ~ age, data = d)
abline(m)
summary(m)

plot(d$age, log(d$income))
ml <- lm(log(income) ~ age, data = d)
abline(ml)
summary(ml)
ageAtMar: Age at first marriage of 5,534 US women.

Description
Age at first marriage of 5,534 US women who responded to the National Survey of Family Growth (NSFG) conducted by the CDC in the 2006 and 2010 cycle.

Usage
data(ageAtMar)

Format
A data frame with 5,534 observations and 1 variable.

Source

Examples
data(ageAtMar)
histPlot(ageAtMar$age)

ami.occurrences: Acute Myocardial Infarction (Heart Attack) Events

Description
This data set is simulated but contains realistic occurrences of AMI in NY City.

Usage
data("ami.occurrences")

Format
A data frame with 365 observations on the following variable.

ami a numeric vector
**Description**

Similar to `lines`, this function will include endpoints that are solid points, open points, or arrows (mix-and-match ready).

**Usage**

```r
ArrowLines(x, y, lty = 1, lwd = 2.5, col = 1, length = 0.1, af = 3,
          cex.pch = 1.2, ends = c("a", "a"), ...)
```

**Arguments**

- `x` A vector of the x-coordinates of the line to be drawn.
- `y` A vector of the y-coordinates of the line to be drawn. This vector should have the same length as that of `x`.
- `lty` The line type.
- `lwd` The line width.
- `col` The line and endpoint color.
- `length` If an end point is an arrow, then this specifies the sizing of the arrow. See the `length` argument in the `arrows` help file for additional details.
- `af` A tuning parameter for creating the arrow. Usually the default (3) will work. If no arrow is shown, make this value larger. If the arrow appears to extend off of the line, then specify a smaller value.
- `cex.pch` Plotting character size (if open or closed point at the end).
- `ends` A character vector of length 2, where the first value corresponds to the start of the line and the second to the end of the line. A value of "a" corresponds to an arrow being shown, "o" to an open circle, and "c" for a closed point.
- `...` All additional arguments are passed to the `lines` function.

**Author(s)**

David Diez

**References**

OpenIntro, openintro.org
Simulated data for association plots, set 1

Description

Simulated data set.

Usage

data("association.1.3")

Format

A data frame with 121 observations on the following 4 variables.

- x1 a numeric vector
- y1 a numeric vector
- y2 a numeric vector
- y3 a numeric vector

Examples

data(association.1.3)
plot(association.1.3$x1, association.1.3$y1)
Simulated data for association plots, set 2

Description
Simulated data set.

Usage
data("association.4.6")

Format
A data frame with 121 observations on the following 4 variables.
ex2  a numeric vector
y4  a numeric vector
y5  a numeric vector
y6  a numeric vector

Examples
data(association.4.6)
plot(association.4.6$x2, association.4.6$y4)

Simulated data for association plots, set 3

Description
Simulated data set.

Usage
data("association.7.12")

Format
A data frame with 141 observations on the following 7 variables.
ex3  a numeric vector
ey7  a numeric vector
ey8  a numeric vector
ey9  a numeric vector
ey10 a numeric vector
ey11 a numeric vector
ey12 a numeric vector
Examples

data(association.7.12)
plot(association.7.12$x3, association.7.12$y7)

assortive.mating    Eye color of couples

Description

Colors of the eye colors of male and female partners.

Usage

data("assortive.mating")

Format

A data frame with 204 observations on the following 2 variables.

self_male  a factor with levels blue brown green
partner_female  a factor with levels blue brown green

Source


Examples

data(assortive.mating)
table(assortive.mating)

avandia    Cardiovascular problems for two types of Diabetes medicines

Description

A comparison of cardiovascular problems for Rosiglitazone and Pioglitazone.

Usage

data("avandia")
babies

Format
A data frame with 227571 observations on the following 2 variables.
treatment a factor with levels Pioglitazone Rosiglitazone
cardiovascular_problems a factor with levels no yes

Source

Examples
data(havandia)
table(havandia)

babies The Child Health and Development Studies

Description
The Child Health and Development Studies investigate a range of topics. One study, in particular, considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. The goal is to model the weight of the infants (bwt, in ounces) using variables including length of pregnancy in days (gestation), mother’s age in years (age), mother’s height in inches (height), whether the child was the first born (parity), mother’s pregnancy weight in pounds (weight), and whether the mother was a smoker (smoke).

Usage
babies

Format
A data frame with 1236 rows and 8 variables:
case  id number
bwt  birthweight, in ounces
gestation  length of gestation, in days
parity  binary indicator for a first pregnancy (0=first pregnancy)
age  mother’s age in years
height  mother’s height in inches
weight  mother’s weight in pounds
smoke  binary indicator for whether the mother smokes
Source

These data come from Child Health and Development Studies

See Also

Gestation

babies.crawl   Crawling age

Description

Crawling age of babies along with the average outdoor temperature at 6 months of age.

Usage

data("babies.crawl")

Format

A data frame with 12 observations on the following 5 variables.

birth_month   a factor with levels April August December February January July June March May November October September
avg_crawling_age   a numeric vector
SD   a numeric vector
n   a numeric vector
temperature   a numeric vector

Source


Examples

data(babies.crawl)
plot(babies.crawl$temperature, babies.crawl$avg_crawling_age)
**bac**  

*Beer and blood alcohol content*

**Description**

Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer.

**Usage**

```r
data("bac")
```

**Format**

A data frame with 16 observations on the following 3 variables.

- `student` a numeric vector
- `beers` a numeric vector
- `bac` a numeric vector

**Source**


**Examples**

```r
data(bac)
plot(bac$beers, bac$bac)
```

---

**ballBearing**  

*Lifespan of ball bearings*

**Description**

A simulated data set on lifespan of ball bearings.

**Usage**

```r
data(ballBearing)
```

**Format**

A data frame with 75 observations on the following variable.

- `lifespan` Lifespan of ball bearings (in hours).
Source

Simulated data.

Examples

data(ballBearing)
par(mfrow=c(1,2))
histPlot(ballBearing$lifeSpan, col='#22558833')
qqnorm(ballBearing$lifeSpan)

bdims

Body measurements of 507 physically active individuals.

Description

Body girth measurements and skeletal diameter measurements, as well as age, weight, height and
gender, are given for 507 physically active individuals - 247 men and 260 women. These data can
be used to provide statistics students practice in the art of data analysis. Such analyses range from
simple descriptive displays to more complicated multivariate analyses such as multiple regression
and discriminant analysis.

Usage

data(bdims)

Format

A data frame with 507 observations on the following 25 variables.

bia.di A numerical vector, respondent’s biacromial diameter in centimeters.

bii.di A numerical vector, respondent’s biiliac diameter (pelvic breadth) in centimeters.

bit.di A numerical vector, respondent’s bitrochanteric diameter in centimeters.

che.de A numerical vector, respondent’s chest depth in centimeters, measured between spine and
sternum at nipple level, mid-expiration.

che.di A numerical vector, respondent’s chest diameter in centimeters, measured at nipple level,
mid-expiration.

elb.di A numerical vector, respondent’s elbow diameter in centimeters, measured as sum of two
elbows.

wri.di A numerical vector, respondent’s wrist diameter in centimeters, measured as sum of two
wrists.

kne.di A numerical vector, respondent’s knee diameter in centimeters, measured as sum of two
knees.

ank.di A numerical vector, respondent’s ankle diameter in centimeters, measured as sum of two
ankles.
sho. gi  A numerical vector, respondent’s shoulder girth in centimeters, measured over deltoid muscles.
che. gi  A numerical vector, respondent’s chest girth in centimeters, measured at nipple line in males and just above breast tissue in females, mid-expiration.
wai. gi  A numerical vector, respondent’s waist girth in centimeters, measured at the narrowest part of torso below the rib cage as average of contracted and relaxed position.
nav. gi  A numerical vector, respondent’s navel (abdominal) girth in centimeters, measured at umbilicus and iliac crest using iliac crest as a landmark.
hip. gi  A numerical vector, respondent’s hip girth in centimeters, measured at at level of bi-trochanteric diameter.
thi. gi  A numerical vector, respondent’s thigh girth in centimeters, measured below gluteal fold as the average of right and left girths.
bic. gi  A numerical vector, respondent’s bicep girth in centimeters, measured when flexed as the average of right and left girths.
for. gi  A numerical vector, respondent’s forearm girth in centimeters, measured when extended, palm up as the average of right and left girths.
kne. gi  A numerical vector, respondent’s knee diameter in centimeters, measured as sum of two knees.
cal. gi  A numerical vector, respondent’s calf maximum girth in centimeters, measured as average of right and left girths.
ank. gi  A numerical vector, respondent’s ankle minimum girth in centimeters, measured as average of right and left girths.
wri. gi  A numerical vector, respondent’s wrist minimum girth in centimeters, measured as average of right and left girths.
age  A numerical vector, respondent’s age in years.
wgt  A numerical vector, respondent’s weight in kilograms.
hgt  A numerical vector, respondent’s height in centimeters.
sex  A categorical vector, 1 if the respondent is male, 0 if female.

Source


Examples

data(bdims)
histPlot(bdims$hgt)
boxPlot(bdims$hgt)
plot(bdims$wgt ~ bdims$hgt)
plot(bdims$hgt ~ bdims$sho.gi)
plot(bdims$wgt ~ bdims$hip.gi)
Description

Data on a random sample of 100 births for babies in North Carolina where the mother was not a smoker and another 50 where the mother was a smoker.

Usage

data(births)

Format

A data frame with 150 observations on the following 14 variables.

- fAge Father’s age.
- mAge Mother’s age.
- weeks Weeks at which the mother gave birth.
- premature Indicates whether the baby was premature or not.
- visits Number of hospital visits.
- gained Weight gained by mother.
- weight Birth weight of the baby.
- sexBaby Gender of the baby.
- smoke Whether or not the mother was a smoker.

Source

These birth records were

References

Birth records released by North Carolina in 2004.

Examples

data(births)
boxPlot(births$weight, births$smoke)
**books**

<table>
<thead>
<tr>
<th>books</th>
<th>Sample of books on a shelf</th>
</tr>
</thead>
</table>

**Description**

Simulated data set.

**Usage**

```r
data("books")
```

**Format**

A data frame with 95 observations on the following 2 variables.

- `type` a factor with levels fiction nonfiction
- `format` a factor with levels hardcover paperback

**Examples**

```r
data(books)
table(books)
```

---

**boxPlot**

<table>
<thead>
<tr>
<th>Box plot</th>
</tr>
</thead>
</table>

**Description**

An alternative to boxplot. Equations are not accepted. Instead, the second argument, `fact`, is used to split the data.

**Usage**

```r
boxPlot(x, fact=NULL, horiz=FALSE, width=2/3, lwd=1,
        lcol='black', medianLwd=2, pch=20, pchCex=1.8,
        col=rgb(0,0,0,0.25), add=FALSE, key=NULL,
        axes=TRUE, xlab='', ylab='', xlim=NULL, ylim=NULL,
        na.rm=TRUE, ...)
```
Arguments

- **x**: A numerical vector.
- **fact**: A character or factor vector defining the grouping for side-by-side box plots.
- **horiz**: If TRUE, the box plot is oriented horizontally.
- **width**: The width of the boxes in the plot. Value between 0 and 1.
- **lwd**: Width of lines used in box and whiskers.
- **lcol**: Color of the box, median, and whiskers.
- **medianLwd**: Width of the line marking the median.
- **pch**: Plotting character of outliers.
- **pchCex**: Size of outlier character.
- **col**: Color of outliers.
- **add**: If FALSE, a new plot is created. Otherwise, the boxplots are added to the current plot for values of TRUE or a numerical vector specifying the locations of the boxes.
- **key**: The order in which to display the side-by-side boxplots. If locations are specified in add, then the elements of add will correspond to the elements of key.
- **axes**: Whether to plot the axes.
- **xlab**: Label for the x axis.
- **ylab**: Label for the y axis.
- **xlim**: Limits for the x axis.
- **ylim**: Limits for the y axis.
- **na.rm**: Indicate whether NA values should be removed.
- **...**: Additional arguments to plot.

Author(s)

David Diez

See Also

histPlot, dotPlot, densityPlot

Examples

data(run10)
par(mfrow=1:2)

###> comparison 1 <###
boxPlot(run10$time)
boxplot(run10$time)

###> comparison 2 <###
boxPlot(run10$time, run10$gender, col=fadeColor('black', '22'))
boxplot(run10$time ~ run10$gender)
### Braces

#### Plot a Braces Symbol

**Description**

This function is not yet very flexible.

**Usage**

```
Braces(x, y, face.radians = 0, long = 1, short = 0.2, ...)
```
Arguments

- `x`: x-coordinate of the center of the braces.
- `y`: y-coordinate of the center of the braces.
- `face.radians`: Radians of where the braces should face. For example, the default with `face.radians = 0` has the braces facing right. Setting to `pi / 2` would result in the braces facing up.
- `long`: The units for the long dimension of the braces.
- `short`: The units for the short dimension of the braces. This must be less than or equal to half of the long dimension.
- `...`: Arguments passed to `lines`.

Author(s)

- David Diez (david@openintro.org)

See Also

- `dlsegments`

Examples

```r
plot(0:1, 0:1, type = 'n')
braces(0.5, 0.5, face.radians = 3 * pi / 2)
```

Description

The function `buildAxis` is built to provide more control of the number of labels on the axis. This function is still under development.

Usage

```r
buildAxis(side, limits, n, nMin = 2, nMax = 10, extend = 2, eps = 10^-12, ...)
```

Arguments

- `side`: The side of the plot where to add the axis.
- `limits`: Either lower and upper limits on the axis or a data set.
- `n`: The preferred number of axis labels.
- `nMin`: The minimum number of axis labels.
- `nMax`: The maximum number of axis labels.
- `extend`: How far the axis may extend beyond `range(limits)`.
- `eps`: The smallest increment allowed.
- `...`: Arguments passed to `axis`
buildAxis

Details
The primary reason behind building this function was to allow a plot to be created with similar
features but with different data sets. For instance, if a set of code was written for one data set and
the function `axis` had been utilized with pre-specified values, the axis may not match the plot of a
new set of data. The function `buildAxis` addresses this problem by allowing the number of axis
labels to be specified and controlled.

The axis is built by assigning penalties to a variety of potential axis setups, ranking them based on
these penalties and then selecting the axis with the best score.

Value
A vector of the axis plotted.

Author(s)
David M Diez

See Also
`histPlot`, `dotPlot`, `boxPlot`, `densityPlot`

Examples
```r
### 0 ===#
limits <- rnorm(100, 605490, 10)
hist(limits, axes=FALSE)
buildAxis(1, limits, 2, nMax=4)

### 1 ===#
x <- seq(0, 500, 10)
y <- 8*x+rnorm(length(x), mean=6000, sd=200)
plot(x, y, axes=FALSE)
buildAxis(1, limits=x, n=5)
buildAxis(2, limits=y, n=3)

### 2 ===#
x <- 9528412 + seq(0, 200, 10)
y <- 8*x+rnorm(length(x), mean=6000, sd=200)
plot(x, y, axes=FALSE)
temp <- buildAxis(1, limits=x, n=4)
buildAxis(2, y, 3)

### 3 ===#
x <- seq(367, 1251, 10)
y <- 7.5*x+rnorm(length(x), mean=6000, sd=800)
plot(x, y, axes=FALSE)
buildAxis(1, limits=x, n=4, nMin=3, nMax=3)
buildAxis(2, limits=y, n=4, nMin=3, nMax=5)

### 4 ===#
x <- seq(367, 367.1, 0.001)
```
y <- 7.5*x+rnorm(length(x), mean=6000, sd=0.01)
plot(x, y, axes=FALSE)
buildAxis(1, limits=x, n=4, nMin=5, nMax=6)
buildAxis(2, limits=y, n=2, nMin=3, nMax=4)

### 5 <=#
x <- seq(-0.05, -0.003, 0.0001)
y <- 50 + 20*x + rnorm(length(x), sd=0.1)
plot(x, y, axes=FALSE)
buildAxis(1, limits=x, n=4, nMin=5, nMax=6)
buildAxis(2, limits=y, n=4, nMax=5)
abline(lm(y ~ x))

### 6 <=#
x <- seq(-0.0097, -0.008, 0.0001)
y <- 50 + 20*x + rnorm(length(x), sd=0.1)
plot(x, y, axes=FALSE)
buildAxis(1, limits=x, n=4, nMin=5, nMax=6)
buildAxis(2, limits=y, n=4, nMax=5)
abline(lm(y ~ x))

### 7 <=#
x <- seq(0.03, -0.003099, -0.000001)
y <- 50 + 20*x + rnorm(length(x), sd=0.1)
plot(x, y, axes=FALSE)
buildAxis(1, limits=x, n=4, nMin=2, nMax=5)
buildAxis(2, limits=y, n=4, nMax=5)
abline(lm(y ~ x))

### 8 - repeat <=#
m <- runif(1)/runif(1) +
rgamma(1, runif(1)/runif(1), runif(1)/runif(1))
s <- rgamma(1, runif(1)/runif(1), runif(1)/runif(1))
x <- rnorm(50, m, s)
hist(x, axes=FALSE)
buildAxis(1, limits=x, n=5, nMin=4, nMax=6, eps=10^-12)
if(diff(range(x)) < 10^-12){
cat("too small\n")
}

---

**burger**

**Burger preferences**

**Description**

Sample burger place preferences versus gender.

**Usage**

`data("burger")`
cancer.in.dogs

Format

A data frame with 500 observations on the following 2 variables.

best_burger_place  Burger place.
gender  a factor with levels Female Male

Source

SurveyUSA, Results of SurveyUSA News Poll #17718, data collected on December 2, 2010.

Examples

data(burger)
table(burger)

cancer.in.dogs  Cancer in dogs

Description

A study in 1994 examined 491 dogs that had developed cancer and 945 dogs as a control group to determine whether there is an increased risk of cancer in dogs that are exposed to the herbicide 2,4-dichlorophenoxyacetic acid (2,4-D).

Usage

data("cancer.in.dogs")

Format

A data frame with 1436 observations on the following 2 variables.

order  a factor with levels 2,4-D no  2,4-D
response a factor with levels cancer no cancer

Source


Examples

data(cancer.in.dogs)
table(cancer.in.dogs)
cards

Description

All the cards in a standard deck.

Usage

data("cards")

Format

A data frame with 52 observations on the following 4 variables.

value  a factor with levels 10 2 3 4 5 6 7 8 9 A J K Q
color a factor with levels black red
suit  a factor with levels Club Diamond Heart Spade
face  a logical vector

Examples

data(cars)
table(cars$value)
table(cars$color)
table(cars$suit)
table(cars$face)
table(cards$suit, cards$face)

cars

Description

A data frame with 54 rows and 6 columns. The columns represent the variables type, price, mpgCity, drivetrain, passengers, weight for a sample of 54 cars from 1993. This data is a subset of the Cars93 data set from the MASS package.

Usage

data(cars)
**Format**

A data frame with 54 observations on the following 6 variables.

- **type**  The vehicle type with levels large, midsize, and small.
- **price**  Vehicle price (USD).
- **mpgCity**  Vehicle mileage in city (miles per gallon).
- **driveTrain**  Vehicle drive train with levels 4WD, front, and rear.
- **passengers**  The vehicle passenger capacity.
- **weight**  Vehicle weight (lbs).

**Details**

These cars represent a random sample for 1993 models that were in both *Consumer Reports* and *PACE Buying Guide*. Only vehicles of type 'small', 'midsize', and 'large' were included.


**Source**


**References**


**Examples**

data(cars)

```r
### Vehicle price by type
par(mfrow=c(1,3))
histPlot(cars$price[cars$type=='small'], probability=TRUE, hollow=TRUE, xlim=c(0,50))
histPlot(cars$price[cars$type=='midsize'], probability=TRUE, hollow=TRUE, border='red', lty=3)
histPlot(cars$price[cars$type=='large'], probability=TRUE, hollow=TRUE, border='blue', lty=4)
legend('topleft', lty=2:4, col=c('black', 'red', 'blue'), legend=c('small', 'midsize', 'large'))
```

```r
### Vehicle price versus weight
plot(cars$weight, cars$price, col=fadeColor('magenta', '88'), pch=20, cex=2)

### Mileage versus weight
plot(cars$weight, cars$mpgCity, type="n")
temp <- c(seq(1000, 5000, 100), rev(seq(1000, 5000, 100)))*1000
hold <- 87.71 - 0.03508*temp + 0.000004432*temp^2 + 7*c(rep(-1, 41), rep(1, 41), 1)
polygon(temp, hold, col="#E2E2E2",
```
The `cats` dataset contains a summary of 144 cats. It includes three variables:

- **Sex**: a factor with levels `F` and `M`.
- **Bwt**: a numeric vector.
- **Hwt**: a numeric vector.

The dataset is used for various statistical analyses. For example, the `hist` function can be used to create a histogram of the body weight (`Bwt`).

The `ccHousing` dataset represents simulated data of housing prices at a community college. These data are intended to represent housing prices of students at a community college.
Format
A data frame with 75 observations on the following variable.

price Monthly housing price, simulated.

References
OpenIntro Statistics, openintro.org

Examples
data(ccHousing)
hist(ccHousing$price)

---

Plot a Cartesian Coordinate Plane

Description
Create a Cartesian Coordinate Plane.

Usage
CCP(xlim = c(-4, 4), ylim = c(-4, 4),
mar = rep(0, 4), length = 0.1, tcl = 0.007,
xlab = FALSE, ticks = 1, ticklabs = 1,
xpos = 1, ypos = 2,
cex.coord = 1, cex.xylab = 1.5, add = FALSE)

Arguments

- **xlim** The x-limits for the plane (vector of length 2).
- **ylim** The y-limits for the plane (vector of length 2).
- **mar** Plotting margins.
- **length** The length argument is passed to the arrows function and is used to control the size of the arrow.
- **tcl** Tick size.
- **xlab** Whether x and y should be shown next to the labels.
- **ticks** How frequently tick marks should be shown on the axes. If a vector of length 2, the first argument will correspond to the x-axis and the second to the y-axis.
- **ticklabs** How frequently tick labels should be shown on the axes. If a vector of length 2, the first argument will correspond to the x-axis and the second to the y-axis.
- **xpos** The position of the labels on the x-axis. See the pos argument in the text function for additional details.
 ypos

The position of the labels on the y-axis. See the pos argument in the text function for additional details.

cex.coord

Inflation factor for font size of the coordinates, where any value larger than zero is acceptable and 1 corresponds to the default.

cex.xylab

Inflation factor for font size of the x and y labels, where any value larger than zero is acceptable and 1 corresponds to the default.

add

Indicate whether a new plot should be created (FALSE, the default) or if the Cartesian Coordinate Plane should be added to the existing plot.

Author(s)

David Diez

References

OpenIntro, openintro.org

See Also

lsegments, dlsegments, ArrowLines

Examples

CCP()

CCP(xlab=TRUE, ylim=c(-3.5, 2), xpos=3, cex.coord=1)

CCP(xlim=c(-8, 8), ylim=c(-10, 6), ticklabs=c(2,2), cex.xylab=0.8)

---

| census            | Random sample of 2000 U.S. Census Data |

Description

A random sample of 500 observations from the 2000 U.S. Census Data.

Usage

data(census)

Format

A data frame with 500 observations on the following 8 variables.

censusYear  Census Year.
stateFIPScode Name of state.
totalFamilyIncome Total family income (in U.S. dollars).
age  Age.
sex  Sex with levels Female and Male.
raceGeneral  Race with levels American Indian or Alaska Native, Black, Chinese, Japanese, Other Asian or Pacific Islander, Two major races, White and Other.
maritalStatus  Marital status with levels Divorced, Married/spouse absent, Married/spouse present, Never married/single, Separated and Widowed.
totalPersonalIncome  Total personal income (in U.S. dollars).

Source

http://factfinder.census.gov/

Examples

data(census)
str(census)
these <- census[,3] > 0  # income greater than 0
histPlot(log(census$totalFamilyIncome[these]), xlab="log(total family income")

--------
cherry  Summary information for 31 cherry trees

Description

Researchers wanting to understand the relationship between these variables for black cherry trees collected data from 31 trees in the Allegheny National Forest, Pennsylvania.

Usage

data("cherry")

Format

A data frame with 31 observations on the following 3 variables.

diam  diameter in inches (at 54 inches above ground)
height  height is measured in feet
volume  volume in cubic feet

Source


Examples

data(cherry)
model <- lm(Volume ~ Diam + Height, cherry)
summary(model)
plot(model)
chick.wts  

*Chicken weights vs feed type*

**Description**

Experiment for chick weights and feed type.

**Usage**

```r
data("chick.wts")
```

**Format**

A data frame with 71 observations on the following 2 variables.

- `weight` a numeric vector
- `feed` a factor with levels `casein horsebean linseed meatmeal soybean sunflower`

**Source**


**Examples**

```r
data(chick.wts)
boxplot(chick.wts$weight ~ chick.wts$feed)
```

---

china  

*Child care hours*

**Description**

The China Health and Nutrition Survey aims to examine the effects of the health, nutrition, and family planning policies and programs implemented by national and local governments.

**Usage**

```r
data("china")
```

**Format**

A data frame with 9788 observations on the following 3 variables.

- `gender` a numeric vector
- `edu` a numeric vector
- `child_care` a numeric vector
ChiSquareTail

Source

Examples

```r
data(china)
summary(china)
```

---

ChiSquareTail  
Plot upper tail in chi-square distribution

Description
Plot a chi-square distribution and shade the upper tail.

Usage

```r
ChisquareTail(u, df, xlim = c(0, 10),
   col = fadeColor("black", "22"),
   axes = TRUE, ...)
```

Arguments

- `u` Cut off for the upper tail.
- `df` Degrees of freedom.
- `xlim` Limits for the plot.
- `col` Color of the shading.
- `axes` Whether to plot an x-axis.
- `...` Currently ignored.

Value
Nothing is returned from the function.

Author(s)
David Diez

See Also

`normTail`

Examples

```r
data(COL)
ChisquareTail(11.7, 7,
   c(0, 25),
   col = COL[1])
```
### classData

*Simulated class data*

**Description**

This data is simulated and is meant to represent students scores from three different lectures who were all given the same exam.

**Usage**

```r
data(classData)
```

**Format**

A data frame with 164 observations on the following 2 variables.

- `m1` Represents a first midterm score.
- `lecture` Three classes: a, b, and c.

**References**

OpenIntro Statistics, Chapter 8.

**Examples**

```r
data(classData)
anova(lm(m1 ~ lecture, classData))
```

---

### COL

*OpenIntro Statistics colors*

**Description**

These are the core colors used for the OpenIntro Statistics textbook. The blue, green, yellow, and red colors are also gray-scaled, meaning no changes are required when printing black and white copies.

**Usage**

```r
data(COL)
```

**Format**

A 7-by-4 matrix of 7 colors with four fading scales: blue, green, yellow, red, black, gray, and light gray.
Source
Colors selected by OpenIntro’s in-house graphic designer, Meenal Patel.

References

Examples
data(COL)
plot(1:7, 7:1, col=COL, pch=19, cex=6, xlab="", ylab="",
xlim=c(0.5,7.5), ylim=c(-2.5,8), axes=FALSE)
text(1:7, 7:1+0.7, paste("COL[", 1:7, "]", sep=""), cex=0.9)
points(1:7, 7:1+0.7, col=COL[,2], pch=19, cex=6)
points(1:7, 7:1-1.4, col=COL[,3], pch=19, cex=6)
points(1:7, 7:1-2.1, col=COL[,4], pch=19, cex=6)

contTable

Generate Contingency Tables for \LaTeX

Description
Input a data frame or a table, and the \LaTeX output will be returned. Options exist for row and column proportions as well as for showing work.

Usage
contTable(x, prop = c("none", "row", "col"),
show = FALSE, digits = 3)

Arguments
x A data frame (with two columns) or a table.
prop Indicate whether row ("r", "R", "row") or column ("c", "C", "col") proportions should be used. The default is to simply print the contingency table.
show If row or column proportions are specified, indicate whether work should be shown.
digits The number of digits after the decimal that should be shown for row or column proportions.

Details
The \texttt{contTable} function makes substantial use of the \texttt{cat} function.

Author(s)
David Diez
References
OpenIntro Statistics, openintro.org

See Also
e-mail, cars, possum, marioKart

Examples
data(email)
table(email[,c("spam", "sent_email")])
contTable(email[,c("spam", "sent_email")])

corr.match Sample data sets for correlation problems

Description
Simulated data.

Usage
data("corr.match")

Format
A data frame with 121 observations on the following 9 variables.

x  a numeric vector
y1 a numeric vector
y2 a numeric vector
y3 a numeric vector
y4 a numeric vector
y5 a numeric vector
y6 a numeric vector
y7 a numeric vector
y8 a numeric vector

Source
Simulated data set.

Examples
data(corr.match)
plot(corr.match$x, corr.match$y1)
cor(corr.match$x, corr.match$y1)
United States Counties

Description

Data for 3143 counties in the United States. See the countyComplete data set for additional variables.

Usage

data(county)

Format

A data frame with 3143 observations on the following 15 variables.

name  County names.
state  State names.
fed_spend  Federal spending per capita
poverty  Percent of population in poverty.
homeownership  Homeownership rate, 2006-2010.
multiunit  Percent of housing units in multi-unit structures, 2006-2010.
income  Income per capita income.
med_income  Median income.

Source

These data were collected from http://quickfacts.census.gov/qfd/states/ (no longer available) and its accompanying pages.

References

~~ OpenIntro Statistics, openintro.org ~~

See Also

email, email50, countyComplete

Examples

data(county)

p00 <- county$pop2000
p10 <- county$pop2010
hist((p10 - p00)/p00)
Description

County-level data, including information on county-level smoking bans.

Usage

data("county.w.sm.ban")

Format

A data frame with 14444 observations on the following 54 variables.

name  County name.
state  State name.
FIPS  a numeric vector
pop2010 a numeric vector
pop2000 a numeric vector
age_under_5 a numeric vector
age_under_18 a numeric vector
age_over_65 a numeric vector
female a numeric vector
white a numeric vector
black a numeric vector
native a numeric vector
asian a numeric vector
pac_isl a numeric vector
two_plus_races a numeric vector
hispanic a numeric vector
white_not_hispanic a numeric vector
no_move_in_one_plus_year a numeric vector
foreign_born a numeric vector
foreign_spoken_at_home a numeric vector
hs_grad a numeric vector
bachelors a numeric vector
veterans a numeric vector
mean_work_travel a numeric vector
housing_units a numeric vector
county.w.sm.ban

home_ownership a numeric vector
housing_multi_unit a numeric vector
median_val_owner_occupied a numeric vector
households a numeric vector
persons_per_household a numeric vector
per_capita_income a numeric vector
median_household_income a numeric vector
poverty a numeric vector
private_nonfarm_establishments a numeric vector
private_nonfarm_employment a numeric vector
percent_change_private_nonfarm_employment a numeric vector
nonemployment_establishments a numeric vector
firms a numeric vector
black_owned_firms a numeric vector
native_owned_firms a numeric vector
asian_owned_firms a numeric vector
pac_isl_owned_firms a numeric vector
hispanic_owned_firms a numeric vector
women_owned_firms a numeric vector
manufacturer_shipments_2007 a numeric vector
mercent_whole_sales_2007 a numeric vector
sales a numeric vector
sales_per_capita a numeric vector
accommodation_food_service a numeric vector
building_permits a numeric vector
fed_spending a numeric vector
area a numeric vector
density a numeric vector
smoking_ban a factor with levels comprehensive none partial

Source

These data were collected from http://quickfacts.census.gov/qfd/states/ (no longer available) and its accompanying pages. Smoking ban data were from a variety of sources.

Examples

data(county.w.sm.ban)
Description

Data for 3143 counties in the United States.

Usage

`data(countyComplete)`

Format

A data frame with 3143 observations on the following 53 variables.

- `state`: State.
- `name`: County name.
- `fips`: FIPS code.
- `age_under_5`: Percent of population under 5 (2010).
- `age_under_18`: Percent of population under 18 (2010).
- `age_over_65`: Percent of population over 65 (2010).
- `female`: Percent of population that is female (2010).
- `white`: Percent of population that is white (2010).
- `black`: Percent of population that is black (2010).
- `native`: Percent of population that is a Native American (2010).
- `asian`: Percent of population that is Asian (2010).
- `pac_isl`: Percent of population that is Hawaii or Pacific Islander (2010).
- `two_plus_races`: Percent of population that identifies as two or more races (2010).
- `hispanic`: Percent of population that is Hispanic (2010).
- `white_not_hispanic`: Percent of population that is white and not Hispanic (2010).
- `no_move_in_one_plus_year`: Percent of population that has not moved in at least one year (2006-2010).
- `foreign_born`: Percent of population that is foreign-born (2006-2010).
- `foreign_spoken_at_home`: Percent of population that speaks a foreign language at home (2006-2010).
- `hs_grad`: Percent of population that is a high school graduate (2006-2010).
- `bachelors`: Percent of population that earned a bachelor’s degree (2006-2010).
- `veterans`: Percent of population that are veterans (2006-2010).
mean_work_travel  Mean travel time to work (2006-2010).
housing_units  Number of housing units (2010).
home_ownership  Homeownership rate (2006-2010).
housing_multi_unit  Housing units in multi-unit structures (2006-2010).
median_val_owner_occupied  Median value of owner-occupied housing units (2006-2010).
households  Households (2006-2010).
persons_per_household  Persons per household (2006-2010).
per_capita_income  Per capita money income in past 12 months (2010 dollars, 2006-2010)
median_household_income  Median household income (2006-2010).
poverty  Percent below poverty level (2006-2010).
private_nonfarm_employment  Private nonfarm employment (2009).
percent_change_private_nonfarm_employment  Private nonfarm employment, percent change from 2000 to 2009.
nonemployment_establishments  Nonemployer establishments (2009).
firms  Total number of firms (2007).
black_owned_firms  Black-owned firms, percent (2007).
native_owned_firms  Native American-owned firms, percent (2007).
asiaindian_owned_firms  Asian-owned firms, percent (2007).
pac_isl_owned_firms  Native Hawaiian and other Pacific Islander-owned firms, percent (2007).
hispanic_owned_firms  Hispanic-owned firms, percent (2007).
women_owned_firms  Women-owned firms, percent (2007).
manufacturer_shipments_2007  Manufacturer shipments, 2007 ($1000).
mercent_whole_sales_2007  Merchance wholesaler sales, 2007 ($1000).
sales  Retail sales, 2007 ($1000).
sales_per_capita  Retail sales per capita, 2007.
accommodation_food_service  Accommodation and food services sales, 2007 ($1000).
building_permits  Building permits (2010).
fed_spending  Federal spending (2009).
area  Land area in square miles (2010).
density  Persons per square mile (2010).

**Source**

http://quickfacts.census.gov/qfd/states/ (no longer available)

**References**

OpenIntro Statistics, openintro.org/os

**Examples**

data(countyComplete)
Description

These patients were randomly divided into a treatment group where they received a blood thinner or the control group where they did not receive a blood thinner. The outcome variable of interest was whether the patients survived for at least 24 hours.

Usage

data("cpr")

Format

A data frame with 90 observations on the following 2 variables.
group a factor with levels control treatment
outcome a factor with levels died survived

Source


Examples

data(cpr)
table(cpr)

Description

A simulated data set of number of credits taken by college students each semester.

Usage

data(credits)

Format

A data frame with 100 observations on the following variable.

credits Number of credits.
Source

Simulated data.

Examples

data(credits)
histPlot(credits$credits)

CT2DF

Contingency Table to Data Frame

Description

Take a 2D contingency table and create a data frame representing the individual cases.

Usage

CT2DF(x, rn = row.names(x), cn = colnames(x), dfn = c("row.var", "col.var"))

Arguments

x 
Contingency table as a matrix.

rn 
Character vector of the row names.

cn 
Character vector of the column names.

dfn 
Character vector with 2 values for the variable representing the rows and columns.

Value

A data frame with two columns.

Author(s)

David Diez

See Also

MosaicPlot

Examples

a <- matrix( 459, 727, 854, 385, 99, 4198, 6245, 4821, 1634, 578), 2, byrow = TRUE)
b <-
CT2DF(  a, c("No", "Yes"),
densityPlot

Description

Compute kernel density plots, written in the same structure as boxPlot. Histograms can be automatically added for teaching purposes.

Usage

densityPlot(x, fact = NULL, bw = "nr0", histo = c("none", "faded", "hollow"), breaks = "Sturges", fading = "0E", fadingBorder = "25", lty = NULL, lwd = 1, col = c("black", "red", "blue"), key = NULL, add = FALSE, adjust = 1, kernel = c("gaussian", "epanechnikov", "rectangular", "triangular", "biweight", "cosine", "optcosine"), weights = NULL, n = 512, from, to, na.rm = FALSE, xlim = NULL, ylim = NULL, main = "", ...)  

Arguments

x A numerical vector.

fact A character or factor vector defining the grouping for data in x.

bw Bandwidth. See density.

histo Whether to plot a faded histogram (‘faded’) or hollow histogram (‘hollow’) in the background. By default, no histogram will be plotted.

breaks The breaks argument for histPlot if histo is ‘faded’ or ‘hollow’.

fading Character value of hexadecimal, e.g. ‘22’ or ‘5D’, describing the amount of fading inside the rectangles of the histogram if histo="faded".

fadingBorder Character value of hexadecimal, e.g. ‘22’ or ‘5D’, describing the amount of fading of the rectangle borders of the histogram if histo is ‘faded’ or ‘hollow’.

lty Numerical vector describing the line type for the density curve(s). Each element corresponds to a different level of the argument fact

lwd Numerical vector describing the line width for the density curve(s). Each element corresponds to a different level of the argument fact

col Numerical vector describing the line color for the density curve(s). Each element corresponds to a different level of the argument fact

key An argument to specify ordering of the factor levels.
add
If TRUE, the density curve is added to the plot.
adjust
Argument passed to density to adjust the bandwidth.
kernel
Argument passed to density to select the kernel used.
weights
Argument passed to density to weight observations.
n
Argument passed to density to specify the detail in the density estimate.
from
Argument passed to density specifying the lowest value to include in the density estimate.
to
Argument passed to density specifying the largest value to include in the density estimate.
n.a.rm
Argument passed to density specifying handling of NA values.
xlim
x-axis limits.
ylim
y-axis limits.
main
Title for the plot.
... If add=FALSE, then additional arguments to plot.

Author(s)
David Diez

See Also
histPlot, dotPlot, boxPlot

Examples

data(tips)
par(mfrow=c(2,2))
histPlot(tips$tip[tips$day == 'Tuesday'], hollow=TRUE, xlim=c(0, 30),
  lty=1, main='Tips by day')
histPlot(tips$tip[tips$day == 'Friday'], hollow=TRUE, border='red',
  add=TRUE, main='Tips by day')
legend('topright', col=c('black', 'red'), lty=1:2,
  legend=c('Tuesday', 'Friday'))
densityPlot(tips$tip, tips$day, col=c('black', 'red'), main='Tips by day')
legend('topright', col=c('black', 'red'), lty=1:2,
  legend=c('Tuesday', 'Friday'))
data(run10)
densityPlot(run10$time, histo='faded', breaks=15, main='Run time')
densityPlot(run10$time, histo='hollow', breaks=30, fadingBorder='66',
  lty=1, main='Run time')
Description

Summaries of nearly 54k diamonds.

Usage

data("diamonds")

Format

A data frame with 53940 observations on the following 10 variables.

carat  a numeric vector
cut    a factor with levels Fair Good Ideal Premium Very Good
color  a factor with levels D E F G H I J
clearance  a factor with levels I1 IF SI1 SI2 VS1 VS2 VVS1 VVS2
depth  a numeric vector
table  a numeric vector
price  a numeric vector
x      a numeric vector
y      a numeric vector
z      a numeric vector

Source


Examples

data(diamonds)
Create a Double Line Segment Plot

Description

Create a plot showing two line segments. The union or intersection of those line segments can also be generated by utilizing the type argument.

Usage

dlsegments(x1 = c(3, 7), x2 = c(5, 9), l = c("o", "o"), r = c("c", "c"),
           type = c("n", "u", "i"), COL = 2, lwd = 2.224, ylim = c(-0.35, 2),
           mar = rep(0, 4), hideOrig = FALSE)

Arguments

x1          The endpoints of the first interval. Values larger (smaller) than 999 (-999) will be interpreted as (negative) infinity.
x2          The endpoints of the second interval. Values larger (smaller) than 999 (-999) will be interpreted as (negative) infinity.
l          A vector of length 2, where the values correspond to the left end point of each interval. A value of "o" indicates the interval is open at the left and "c" indicates the interval is closed at this end.
r          A vector of length 2, where the values correspond to the right end point of each interval. A value of "o" indicates the interval is open at the right and "c" indicates the interval is closed at this end.
type        By default, no intersection or union of the two lines will be shown (value of "n"). To show the union of the line segments, specify "u". To indicate that the intersection be shown, specify "i".
COL          If the union or intersection is to be shown (see the type argument), then this parameter controls the color that will be shown.
lwd         If the union or intersection is to be shown (see the type argument), then this parameter controls the width of any corresponding lines or open points in the union or intersection.
ylim        A vector of length 2 specifying the vertical plotting limits, which may be useful for fine-tuning plots. The default is c(-0.35, 2).
mar          A vector of length 4 that represent the plotting margins.
hideOrig     An optional argument that to specify that the two line segments should be shown (hideOrig takes value FALSE, the default) or that they should be hidden (hideOrig takes value TRUE).

Author(s)

David Diez
References

OpenIntro, openintro.org

See Also

lsegments, CCP, ArrowLines

Examples

dlsegments(c(-3,3), c(1, 1000),
   r=c("o", "o"), l=c("c", "o"), COL=COL[4])
dlsegments(c(-3,3), c(1, 1000),
   r=c("o", "o"), l=c("c", "o"), type="un", COL=COL[4])
dlsegments(c(-3,3), c(1, 1000),
   r=c("o", "o"), l=c("c", "o"), type="in", COL=COL[4])

dotPlot

Description

Plot observations as dots.

Usage

dotPlot(x, fact = NULL, vertical = FALSE, at = 1, key = NULL,
   pch = 20, col = fadeColor("black", "66"), cex = 1.5,
   add = FALSE, axes = TRUE, xlim = NULL, ylim = NULL, ...)

Arguments

x A numerical vector.

fact A character or factor vector defining the grouping for data in x.

vertical If TRUE, the plot will be oriented vertically.

at The vertical coordinate of the points, or the horizontal coordinate if vertical=TRUE.

key The factor levels corresponding to at, pch, col, and cex.

pch Plotting character. If fact is given, then different plotting characters can be specified for each factor level. If key is specified, the elements of pch will correspond to the elements of key.

col Plotting character color. If fact is given, then different colors can be specified for each factor level. If key is specified, the elements of col will correspond to the elements of key.
dotPlot

Plotting character size. If fact is given, then different character sizes can be specified for each factor level. If key is specified, the elements of cex will correspond to the elements of key.

add
If TRUE, then the points are added to the plot.

axes
If FALSE, no axes are plotted.

xlim
Limits for the x axis.

ylim
Limits for the y axis.

... Additional arguments to be passed to plot if add=FALSE or points if add=TRUE.

Author(s)
David Diez

See Also
histPlot, densityPlot, boxPlot

Examples

### example 1
```r
data(cars)
dotPlot(cars$price, cars$type, key=c('large', 'midsize', 'small'), cex=1:3)
```

### example 2
```r
data(run10)
layout(matrix(1:2,2), heights=c(2.7,1.5))
par(las=1)
these <- run10$gender=='M'
dotPlot(run10$time[these], run10$div[these],
col=fadeColor('black', '11'))
# disorganized levels in the above plot, which we could
# organize with key. an example of organizing the levels...
dotPlot(run10$time[these], run10$div[these],
col=fadeColor('black', '11'),
key=c('20-24', '25-29', '30-34', '35-39'))
par(las=0, mfrow=c(1,1))
```

### example 3
```r
data(marioKart)
dotPlot(marioKart$totalPr, marioKart$cond, ylim=c(0.5,2.5),
xlim=c(25, 80), cex=1) # miss the outliers
boxPlot(marioKart$totalPr, marioKart$cond, add=1:2+0.1,
key=c('new', 'used'), horiz=TRUE, axes=FALSE)
```
dotPlotStack

Add a Stacked Dot Plot to an Existing Plot

Description

Add a stacked dot plot to an existing plot. The locations for the points in the dot plot are returned from the function in a list.

Usage

dotPlotStack(x, radius = 1, seed = 1, addDots = TRUE, ...)

Arguments

x A vector of numerical observations for the dot plot.
radius The approximate distance that should separate each point.
seed A random seed (integer). Different values will produce different variations.
addDots Indicate whether the points should be added to the plot.
... Additional arguments are passed to points.

Value

Returns a list with a height that can be used as the upper bound of ylim for a plot, then also the x- and y-coordinates of the points in the stacked dot plot.

Author(s)

David Diez

References

~~ OpenIntro Statistics, openintro.org ~~

See Also

dotPlot, histPlot

Examples

#
dream

Survey on views of the DREAM Act

Description
A SurveyUSA poll.

Usage
data("dream")

Format
A data frame with 910 observations on the following 2 variables.

ideology a factor with levels Conservative Liberal Moderate
stance a factor with levels No Not sure Yes

Source
SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

Examples
data(dream)
table(dream)

drug.use
Drug use of students and parents

Description
Summary of 445 student-parent pairs.

Usage
data("drug.use")

Format
A data frame with 445 observations on the following 2 variables.

student a factor with levels not uses
parents a factor with levels not used
Source


Examples

data(drug.use)
table(drug.use)

data(“ebola.survey”)
table(“ebola.survey”)

data(“ebola.survey”)
table(“ebola.survey”)

Description

In New York City on October 23rd, 2014, a doctor who had recently been treating Ebola patients in Guinea went to the hospital with a slight fever and was subsequently diagnosed with Ebola. Soon thereafter, an NBC 4 New York/The Wall Street Journal/Marist Poll asked New Yorkers whether they favored a “mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient”. This poll included responses of 1,042 New York adults between October 26th and 28th, 2014.

Usage

data(“ebola.survey”)”

Format

A data frame with 1042 observations on the following variable.

quarantine Indicates whether the respondent is in favor or against the mandatory quarantine.

Source

Poll ID NY141026 on maristpoll.marist.edu.

Examples

data(ebola.survey)
table(ebola.survey)
edaPlot

**Exploratory data analysis plot**

**Description**

Explore different plotting methods using a click interface.

**Usage**

```r
edaplot(dataFrame, Col=c("#888888", "#FF0000", "#222222",
                          "#FFFFFF", "#CCCCCC", "#3377AA"))
```

**Arguments**

- `dataFrame` A data frame.
- `Col` A vector containing six colors. The colors may be given in any form.

**Author(s)**

David Diez

**See Also**

`histPlot, densityPlot, boxPlot, dotPlot`

**Examples**

```r
data(mlbBat10)
bat <- mlbBat10[mlbBat10$AB > 200,]
#edaPlot(bat)

data(mariokart)
mk <- mariokart[mariokart$totalPr < 100,]
#edaPlot(mk)
```

---

**elmhurst**

*Elmhurst College gift aid*

**Description**

A random sample of 50 students gift aid for students at Elmhurst College.

**Usage**

```r
data("elmhurst")
```
**Format**

A data frame with 50 observations on the following 3 variables.

- *family_income*  Family income of the student.
- *gift_aid*  Gift aid, in $1000s.
- *price_paid*  Price paid by the student (tuition - gift aid).

**Source**

These data were sampled from a table of data for all freshman from the 2011 class at Elmhurst College that accompanied an article titled What Students Really Pay to Go to College published online by The Chronicle of Higher Education: [http://chronicle.com/article/What-Students-Really-Pay-to-Go/131435](http://chronicle.com/article/What-Students-Really-Pay-to-Go/131435).

**Examples**

```r
data(elmhurst)
plot(elmhurst$family_income, elmhurst$gift_aid)
m <- lm(gift_aid ~ family_income, data = elmhurst)
abline(m)
summary(m)
```

---

**email**

*Data frame representing information about a collection of emails*

**Description**

These data represent incoming emails for the first three months of 2012 for an email account (see Source).

**Usage**

```r
data(email)
data(email_test)
```

**Format**

A *email* (*email_sent*) data frame has 3921 (1252) observations on the following 21 variables.

- *spam*  Indicator for whether the email was spam.
- *to_multiple*  Indicator for whether the email was addressed to more than one recipient.
- *from*  Whether the message was listed as from anyone (this is usually set by default for regular outgoing email).
- *cc*  Indicator for whether anyone was CCed.
- *sent_email*  Indicator for whether the sender had been sent an email in the last 30 days.
- *time*  Time at which email was sent.
image  The number of images attached.
attach  The number of attached files.
dollar  The number of times a dollar sign or the word “dollar” appeared in the email.
winner  Indicates whether “winner” appeared in the email.
inherit  The number of times “inherit” (or an extension, such as “inheritance”) appeared in the email.
viegra  The number of times “viagra” appeared in the email.
password  The number of times “password” appeared in the email.
num_char  The number of characters in the email, in thousands.
line_breaks  The number of line breaks in the email (does not count text wrapping).
format  Indicates whether the email was written using HTML (e.g. may have included bolding or active links).
re_subj  Whether the subject started with “Re:”, “RE:”, “re:”, or “rE:”
exclaim_subj  Whether there was an exclamation point in the subject.
urgent_subj  Whether the word “urgent” was in the email subject.
exclaim_mess  The number of exclamation points in the email message.
number  Factor variable saying whether there was no number, a small number (under 1 million), or a big number.

Source

David Diez’s Gmail Account, early months of 2012. All personally identifiable information has been removed.

References

~~ OpenIntro Statistics, openintro.org ~~

See Also

email50, county

Examples

data(email)
e <- email

#______ Variables For Logistic Regression ______#
# Variables are modified to match
#  OpenIntro Statistics, Second Edition
#  As Is (7): spam, to_multiple, winner, format,
#    re_subj, exclaim_subj
#  Omitted (6): from, sent_email, time, image,
#    viagra, urgent_subj, number
#  Become Indicators (5): cc, attach, dollar,
#    inherit, password
e$cc <- ifelse(email$cc > 0, 1, 0)
e$attach <- ifelse(email$attach > 0, 1, 0)
e$dollar <- ifelse(email$dollar > 0, 1, 0)
e$inherit <- ifelse(email$inherit > 0, 1, 0)
e$password <- ifelse(email$password > 0, 1, 0)
# Transform (3): num_char, line_breaks, exclaim_mess
#e$num_char <- cut(email$num_char, c(0, 1, 5, 10, 20, 10000))
#e$line_breaks <- cut(email$line_breaks, c(0, 100, 1000, 100000))
#e$exclaim_mess <- cut(email$exclaim_mess, c(-1, 0, 1, 5, 100000))
g <- glm(spam ~ to_multiple + winner + format +
  re_subj + exclaim_subj +
  cc + attach + dollar +
  inherit + password, # +
  num_char + line_breaks + exclaim_mess,
  data=email, family=binomial)
summary(g)

#------ Variable Selection Via AIC ------#
g. <- step(g)
plot(predict(g., type="response"), email$spam)

#------ Splitting num_char by html ------#
x <- log(email$num_char)
bw <- 0.004
R <- range(x) + c(-1, 1)
wt <- sum(email$format)/nrow(email)
htmlAll <- density(x, bw=0.4, from=R[1], to=R[2])
htmlNo <- density(x[email$format != 1L], bw=0.4,
  from=R[1], to=R[2])
htmlYes <- density(x[email$format == 1L], bw=0.4,
  from=R[1], to=R[2])
htmlNo$y <- htmlNo$y # (1-wt)
htmlYes$y <- htmlYes$y #* wt + htmlNo$y
plot(htmlAll, xlim=c(-4, 6), ylim=c(0, 0.4))
  lines(htmlNo, col=4)
  lines(htmlYes, lwd=2, col=2)

---

email50

Sample of 50 emails

Description

This is a subsample of the email data set.

Usage

data(email50)
Format

A data frame with 50 observations on the following 21 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>spam</td>
<td>Indicator for whether the email was spam.</td>
</tr>
<tr>
<td>to_multiple</td>
<td>Indicator for whether the email was addressed to more than one recipient.</td>
</tr>
<tr>
<td>from</td>
<td>Whether the message was listed as from anyone (this is usually set by default for regular outgoing email).</td>
</tr>
<tr>
<td>cc</td>
<td>Indicator for whether anyone was CCed.</td>
</tr>
<tr>
<td>sent_email</td>
<td>Indicator for whether the sender had been sent an email in the last 30 days.</td>
</tr>
<tr>
<td>time</td>
<td>Time at which email was sent.</td>
</tr>
<tr>
<td>image</td>
<td>The number of images attached.</td>
</tr>
<tr>
<td>attach</td>
<td>The number of attached files.</td>
</tr>
<tr>
<td>dollar</td>
<td>The number of times a dollar sign or the word “dollar” appeared in the email.</td>
</tr>
<tr>
<td>winner</td>
<td>Indicates whether “winner” appeared in the email.</td>
</tr>
<tr>
<td>inherit</td>
<td>The number of times “inherit” (or an extension, such as “inheritance”) appeared in the email.</td>
</tr>
<tr>
<td>viagra</td>
<td>The number of times “viagra” appeared in the email.</td>
</tr>
<tr>
<td>password</td>
<td>The number of times “password” appeared in the email.</td>
</tr>
<tr>
<td>num_char</td>
<td>The number of characters in the email, in thousands.</td>
</tr>
<tr>
<td>line_breaks</td>
<td>The number of line breaks in the email (does not count text wrapping).</td>
</tr>
<tr>
<td>format</td>
<td>Indicates whether the email was written using HTML (e.g. may have included bolding or active links).</td>
</tr>
<tr>
<td>re_subj</td>
<td>Whether the subject started with “Re:”, “RE:”, “re:”, or “rE:”</td>
</tr>
<tr>
<td>exclaim_subj</td>
<td>Whether there was an exclamation point in the subject.</td>
</tr>
<tr>
<td>urgent_subj</td>
<td>Whether the word “urgent” was in the email subject.</td>
</tr>
<tr>
<td>exclaim_mess</td>
<td>The number of exclamation points in the email message.</td>
</tr>
<tr>
<td>number</td>
<td>Factor variable saying whether there was no number, a small number (under 1 million), or a big number.</td>
</tr>
</tbody>
</table>

Source

David Diez’s Gmail Account, early months of 2012. All personally identifiable information has been removed.

References

~~ OpenIntro Statistics, openintro.org ~~

See Also

email, county
Examples

```r
data(email50)
data(email)
set.seed(5)
d <- email[sample(nrow(email), 50), ][c(1:25, 27:50), ]
identical(d, email50)

# the "[c(1, 26, 2:25, 27:50),]" was added to reorder the cases
```

---

**epa2012**  
*Vehicle info from the EPA*

---

**Description**

Details from the EPA.

**Usage**

```r
data("epa2012")
```

**Format**

A data frame with 1129 observations on the following 28 variables.

- `model_yr` a numeric vector
- `mfr_name` Manufacturer name.
- `division` Vehicle division.
- `carline` Vehicle line.
- `mfr_code` a factor with levels `adx asx azd bex bgt bmx cda crx dsx fjx fmx gmx hnx hyx jcx kmx lrx ltx max mbx mtx nlx nsx prx rii rrg sax skx tkx tvp tyx vxv vxw`
- `model_type_index` a numeric vector
- `engine_displacement` a numeric vector
- `no_cylinders` a numeric vector
- `city_mpg` a numeric vector
- `hwy_mpg` a numeric vector
- `comb_mpg` a numeric vector
- `guzzler` a factor with levels `N Y`
- `air_aspir_method` a factor with levels `SC TC`
- `air_aspir_method_desc` a factor with levels `Naturally Aspirated Supercharged Turbocharged`
- `transmission` a factor with levels `A AM CVT M OT SA SCV`
transmission_desc a factor with levels Automated Manual Automatic Continuously Variable
Manual Other Selectable Continuously Variable (e.g. CVT with paddles) Semi-Automatic

no_gears a numeric vector

trans_lockup a factor with levels N Y

trans_creeper_gear a factor with levels N

drive_sys a factor with levels 4 A F P R

drive_desc a factor with levels 2-Wheel Drive, Front 2-Wheel Drive, Rear 4-Wheel Drive
All Wheel Drive Part-time 4-Wheel Drive

fuel_usage a factor with levels DU EL G GM GP GPR H

fuel_usage_desc a factor with levels Diesel Electricity Gasoline (Mid Grade Unleaded Recommended) Gasoline (Premium Unleaded Recommended) Gasoline (Premium Unleaded Required) Gasoline (Regular Unleaded Recommended) Hydrogen

class a factor with levels Compact Cars Large Cars Midsize Cars Midsize Station Wagons Minicompact Cars Small Pick-up Trucks 2WD Small Pick-up Trucks 4WD Small Station Wagons Special Purpose Vehicle 2WD Special Purpose Vehicle, minivan 2WD Special Purpose Vehicle, minivan 4WD Special Purpose Vehicle, SUV 2WD Special Purpose Vehicle, SUV 4WD Standard Pick-up Trucks 2WD Standard Pick-up Trucks 4WD Subcompact Cars Two Seaters Vans, Cargo Types Vans, Passenger Type

car_truck a factor with levels 1 2 car

release_date Date of vehicle release.

fuel_cell a factor with levels N Y

Source


Examples

data(epa2012)

<table>
<thead>
<tr>
<th>exams</th>
<th>Exam scores</th>
</tr>
</thead>
</table>

Description

Exam scores from a class of 19 students.

Usage

data("exams")

Format

A data frame with 19 observations on the following variable.

scores a numeric vector
Examples

data(exams)
hist(exams$scores)

fadeColor

Fade colors

Description
Fade colors so they are transparent.

Usage
fadeColor(col, fade = "FF")

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>col</td>
<td>An integer, color name, or RGB hexadecimal.</td>
</tr>
<tr>
<td>fade</td>
<td>The amount to fade <code>col</code>. This value should be a character in hexadecimal from '00' to 'FF'. The smaller the value, the greater the fading.</td>
</tr>
</tbody>
</table>

Author(s)
David Diez

See Also
dotPlot

Examples

data(marioKart)
new <- marioKart$cond == 'new'
used <- marioKart$cond == 'used'
par(mfrow=1:2)

### color numbers

dotPlot(marioKart$totalPr[new], ylim=c(0,3), xlim=c(25, 80), pch=20, col=2, cex=2, main='using regular colors')
dotPlot(marioKart$totalPr[used], at=2, add=TRUE, col=4, pch=20, cex=2)
dotPlot(marioKart$totalPr[new], ylim=c(0,3), xlim=c(25, 80), col=fadeColor(2, '22'), pch=20, cex=2, main='fading the colors first')
dotPlot(marioKart$totalPr[used], at=2, add=TRUE, col=fadeColor(4, '22'), pch=20, cex=2)

### color names

dotPlot(marioKart$totalPr[new], ylim=c(0,3), xlim=c(25, 80), pch=20,
fcid

Summary of male heights from USDA Food Commodity Intake Database

Description

Sample of heights based on the weighted sample in the survey.

Usage

data("fcid")

Format

A data frame with 100 observations on the following 2 variables.

height a numeric vector
num_of_adults a numeric vector
Examples
data(fcid)

| fheights       | Female college student heights, in inches |

Description
24 sample observations.

Usage
data("fheights")

Format
A data frame with 24 observations on the following variable.

| heights       | height, in inches |

Examples
data(fheights)
hist(fheights$heights)

friday            Friday the 13th

Description
This data set addresses issues of how superstitions regarding Friday the 13th affect human behavior, and whether Friday the 13th is an unlucky day. Scanlon, et al. collected data on traffic and shopping patterns and accident frequency for Fridays the 6th and 13th between October of 1989 and November of 1992.

There are three types of observations: traffic, shopping, and accident. For traffic, the researchers obtained information from the British Department of Transport regarding the traffic flows between junctions 7 to 8 and junctions 9 to 10 of the M25 motorway. For shopping, they collected the numbers of shoppers in nine different supermarkets in southeast England. For accidents, they collected numbers of emergency admissions to hospitals due to transport accidents.

Usage
data(friday)
full.body.scan

**Format**
A data frame with 61 observations and 6 variables.

- **type** Type of observation, traffic, shopping, or accident.
- **date** Year and month of observation.
- **sixth** Counts on the 6th of the month.
- **thirteenth** Counts on the 13th of the month.
- **diff** Difference between the sixth and the thirteenth.
- **location** Location where data is collected.

**Source**

[http://lib.stat.cmu.edu/DASL/DataFiles/Fridaythe13th.html](http://lib.stat.cmu.edu/DASL/DataFiles/Fridaythe13th.html)

**Examples**

```r
data(friday)
par(mfrow = c(1,2))
boxPlot(friday$sixth[friday$type == "traffic"], xlab = "sixth")
boxPlot(friday$thirteenth[friday$type == "traffic"], xlab = "thirteenth")
```

---

**full.body.scan**

Poll about use of full-body airport scanners

**Description**
Poll about use of full-body airport scanners, where about 4-in-5 people supported the use of the scanners.

**Usage**

data("full.body.scan")

**Format**
A data frame with 1137 observations on the following 2 variables.

- **answer** a factor with levels do not know / no answer should should not
- **party.affiliation** a factor with levels Democrat Independent Republican

**Source**
gender.discrimination

Examples

data(full.body.scan)
## maybe str(full.body.scan) ; plot(full.body.scan) ...

gear.company

Fake data for a gear company example

Description

Made-up data for whether a sample of two gear companies’ parts pass inspection.

Usage

data("gear.company")

Format

A data frame with 2000 observations on the following 2 variables.

company  a factor with levels current prospective
outcome  a factor with levels not pass

Examples

data(gear.company)
## maybe str(gear.company) ; plot(gear.company) ...

gender.discrimination

Bank manager recommendations based on gender

Description

Study from the 1970s about whether gender influences hiring recommendations.

Usage

data("gender.discrimination")

Format

A data frame with 48 observations on the following 2 variables.

gender  a factor with levels female male
decision a factor with levels not promoted
Source


Examples

data(gender.discrimination)
## maybe str(gender.discrimination) ; plot(gender.discrimination) ...

---

Analytical skills of young gifted children

description

An investigator is interested in understanding the relationship, if any, between the analytical skills of young gifted children and the following variables: father's IQ, mother's IQ, age in month when the child first said 'mummy' or 'daddy', age in month when the child first counted to 10 successfully, average number of hours per week the child's mother or father reads to the child, average number of hours per week the child watched an educational program on TV during the past three months, average number of hours per week the child watched cartoons on TV during the past three months. The analytical skills are evaluated using a standard testing procedure, and the score on this test is used as the response variable.

Data were collected from schools in a large city on a set of thirty-six children who were identified as gifted children soon after they reached the age of four.

Usage

data(gifted)

Format

A data frame with 36 observations and 8 variables.

- score  Score in test of analytical skills.
- fatheriq  Father's IQ.
- motheriq  Mother's IQ.
- speak  Age in months when the child first said 'mummy' or 'daddy'.
- count  Age in months when the child first counted to 10 successfully.
- read  Average number of hours per week the child's mother or father reads to the child.
- edutv  Average number of hours per week the child watched an educational program on TV during the past three months.
- cartoons  Average number of hours per week the child watched cartoons on TV during the past three months.
**Source**


**Examples**

```r
data(gifted)
histPlot(gifted$count)
histPlot(gifted$fatheriq)
histPlot(gifted$motheriq)
histPlot(gifted$motheriq - gifted$fatheriq)
plot(gifted$score ~ gifted$motheriq)
lm(gifted$score ~ gifted$motheriq + gifted$fatheriq + gifted$speak +
gifted$count + gifted$read +
gifted$edtv + gifted$cartoons)
```

---

**Pew survey on global warming**

**Description**

A 2010 Pew Research poll asked 1,306 Americans, “From what you’ve read and heard, is there solid evidence that the average temperature on earth has been getting warmer over the past few decades, or not?”

**Usage**

```r
data("global.warming.pew")
```

**Format**

A data frame with 2253 observations on the following 2 variables.

- `party_or_ideology` a factor with levels Conservative Republican Liberal Democrat Mod/Cons Democrat Mod/Lib Republican
- `response` Response

**Source**

Pew Research Center, Majority of Republicans No Longer See Evidence of Global Warming, data collected on October 27, 2010.

**Examples**

```r
data(global.warming.pew)
## maybe str(global.warming.pew) ; plot(global.warming.pew) ...```
goog  

*Google stock data*

**Description**

Google stock data from 2006 to early 2014, where data from the first day each month was collected.

**Usage**

`data("goog")`

**Format**

A data frame with 98 observations on the following 7 variables.

- `date` a factor with levels 2006-01-03, 2006-02-01, and so on
- `open` a numeric vector
- `high` a numeric vector
- `low` a numeric vector
- `close` a numeric vector
- `volume` a numeric vector
- `Adj.Close` a numeric vector

**Source**

Yahoo! Finance.

**Examples**

```r
data(goog)
## maybe str(goog); plot(goog) ...
```

---

gov.poll  

*Pew Research poll on government approval ratings*

**Description**

The poll’s focus is on Obama and then Democrats and Republicans in Congress.

**Usage**

`data("gov.poll")`
Format
A data frame with 4223 observations on the following 2 variables.

- poll: a factor with levels approve disapprove
- eval: a factor with levels Democrats Obama Republicans

Source
See the Pew Research website: www.people-press.org/2012/03/14/romney-leads-gop-contest-trails-in-matchup-with-obama. The counts in Table 6.19 are approximate.

Examples
data(gov.poll)
## maybe str(gov.poll) ; plot(gov.poll) ...

govRace10  

---

Election results for 2010 Governor races in the U.S.

Description
Election results for 2010 Governor races in the U.S.

Usage
data(govRace10)

Format
A data frame with 37 observations on the following 23 variables.

- id: Unique identifier for the race, which does not overlap with other 2010 races (see houseRace10 and senateRace10)
- state: State name
- abbr: State name abbreviation
- name1: Name of the winning candidate
- perc1: Percentage of vote for winning candidate (if more than one candidate)
- party1: Party of winning candidate
- votes1: Number of votes for winning candidate
- name2: Name of candidate with second most votes
- perc2: Percentage of vote for candidate who came in second
- party2: Party of candidate with second most votes
- votes2: Number of votes for candidate who came in second
- name3: Name of candidate with third most votes
Source

Data was collected from MSNBC.com on November 9th, 2010.

Examples

data(govRace10)
table(govRace10[,c("party1", "party2")])

---

gpa  
 Survey of Duke students on GPA, studying, and more

Description

A survey of 55 Duke University students asked about their GPA, number of hours they study at night, number of nights they go out, and their gender.

Usage

data("gpa")

Format

A data frame with 55 observations on the following 5 variables.

gpa  a numeric vector
studyweek  a numeric vector
sleepnight  a numeric vector
out  a numeric vector
gender  a factor with levels female male

Examples

data(gpa)
## maybe str(gpa) ; plot(gpa) ...
A data frame with 193 rows and 2 columns. The columns represent the variables `gpa` and `study_hours` for a sample of 193 undergraduate students who took an introductory statistics course in 2012 at a private US university.

Usage

```r
data(gpa_study_hours)
```

Format

A data frame with 193 observations on the following 2 variables.

- `gpa` Grade point average (GPA) of student.
- `study_hours` Number of hours students study per week.

Details

GPA ranges from 0 to 4 points, however one student reported a GPA > 4. This is a data error but this observation has been left in the dataset as it is used to illustrate issues with real survey data. Both variables are self reported, hence may not be accurate.

Source

Collected at a private US university as part of an anonymous survey in an introductory statistics course.

Examples

```r
data(gpa_study_hours)

#=>> gap vs. study hours <===
plot(gpa_study_hours$gpa ~ gpa_study_hours$study_hours,
     xlab = "Study hours/week", ylab = "GPA",
     pch = 20, col = COL[1:2])
```
Simulated data for analyzing the relationship between watching TV and grades

Description

This is a simulated data set to be used to estimate the relationship between number of hours per week students watch TV and the grade they got in a statistics class.

Usage

data(gradesTV)

Format

A data frame with 25 observations on the following 2 variables.

TV  Number of hours per week students watch TV.
Grades Grades students got in a statistics class (out of 100).

Details

There are a few potential outliers in this data set. When analyzing the data one should consider how (if at all) these outliers may affect the estimates of correlation coefficient and regression parameters.

Source

Simulated data

Examples

data(gradesTV)
str(gradesTV)

plot(gradesTV)
makeTube(gradesTV$TV, gradesTV$Grades, 1.5, type='robust', homosk=FALSE)

lmPlot(gradesTV$TV, gradesTV$Grades, xAxis=4, xlab='time watching TV', yR=0.2, highlight=c(1,15,20))
**gsearch**  
*Simulated Google search experiment*

**Description**

The data were simulated to look like sample results from a Google search experiment.

**Usage**

```r
data("gsearch")
```

**Format**

A data frame with 10000 observations on the following 2 variables.

- **type**: a factor with levels `new search` `no new search`
- **outcome**: a factor with levels `current test` `Q test` `R`  

**Examples**

```r
data(gsearch)
## maybe str(gsearch) ; plot(gsearch) ...
```

---

**healthcare.law.survey**  
*Pew Research Center poll on health care, including question variants*

**Description**

For example, Pew Research Center conducted a survey with the following question: “As you may know, by 2014 nearly all Americans will be required to have health insurance. [People who do not buy insurance will pay a penalty] while [People who cannot afford it will receive financial help from the government]. Do you approve or disapprove of this policy?” For each randomly sampled respondent, the statements in brackets were randomized: either they were kept in the order given above, or the two statements were reversed.

**Usage**

```r
data("healthcare.law.survey")
```

**Format**

A data frame with 1503 observations on the following 2 variables.

- **order**: a factor with levels `cannot_afford_second` `penalty_second`
- **response**: a factor with levels `approve` `disapprove` `other`
healthcare.survey

Source
www.people-press.org/2012/03/26/public-remains-split-on-health-care-bill-opposed-to-mandate/. Sample sizes for each polling group are approximate.

Examples
data(healthcare.law.survey)
## maybe str(healthcare.law.survey) ; plot(healthcare.law.survey) ...

---

healthcare.survey   Healthcare Survey with Two Different Wordings

Description
How does survey wording and order of options impact responses?

Usage
data("healthcare.survey")

Format
A data frame with 1503 observations on the following 2 variables.

  wording  Which phrasing came first: first: pay penalty or first: receive help.
  response  Response classification of approve, disapprove, or other.

Details
The survey question, where the content in the brackets was switched for each respondent. "As you may know, by 2014 nearly all Americans will be required to have health insurance. [People who do not buy insurance will pay a penalty] while [People who cannot afford it will receive financial help from the government]. Do you approve or disapprove of this policy?"

Source

Examples
data(healthcare.survey)
table(healthcare.survey)
**health_coverage**  
*Health Coverage and Health Status*

<table>
<thead>
<tr>
<th>health_coverage</th>
<th>Description</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Survey responses for 20,000 responses to the Behavioral Risk Factor Surveillance System.</td>
<td>data(&quot;health_coverage&quot;)</td>
</tr>
<tr>
<td>coverago</td>
<td>Whether the person had health coverage or not.</td>
<td></td>
</tr>
<tr>
<td>health_status</td>
<td>The person’s health status.</td>
<td></td>
</tr>
</tbody>
</table>

| heartTr        | Heart Transplant Data  |

**Description**  
The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated officially a heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Then the actual heart transplant occurs between a few weeks to several months depending on the availability of a donor. Very few candidates during this waiting period show improvement and get deselected as a heart transplant candidate, but for the purposes of this experiment those patients were kept in the data as continuing candidates.

**Usage**  
data(heartTr)
Format

A data frame with 103 observations on the following 8 variables.

- **id**: ID number of the patient.
- **acceptyear**: Year of acceptance as a heart transplant candidate.
- **age**: Age of the patient at the beginning of the study.
- **survived**: Survival status with levels alive and dead.
- **survtime**: Number of days patients were alive after the date they were determined to be a candidate for a heart transplant until the termination date of the study.
- **prior**: Whether or not the patient had prior surgery with levels yes and no.
- **transplant**: Transplant status with levels control (did not receive a transplant) and treatment (received a transplant).
- **wait**: Waiting Time for Transplant

Source

http://www.stat.ucla.edu/~jsanchez/data/stanford.txt

References


Examples

```r
data(heartTr)
str(heartTr)
boxplot(heartTr$survtime, heartTr$transplant, ylab = 'Survival Time (days)'
mosaicplot(~ transplant + survived, data = heartTr)
```

helium  

*Helium football*

Description

At the 1976 Pro Bowl, Ray Guy, a punter for the Oakland Raiders, punted a ball that hung mid-air long enough for officials to question whether the pigskin was filled with helium. The ball was found to be filled with air, but since then many have tossed around the idea that a helium-filled football would outdistance an air-filled one. Students at Ohio State University conducted an experiment to test this myth. They used two identical footballs, one air filled with air and one filled with helium. Each football was kicked 39 times and the two footballs were alternated with each kick.

Usage

```r
data(helium)
```
**Format**

A data frame with 39 observations on the following 3 variables.

- **trial**: Trial number.
- **air**: Distance in years for air-filled football.
- **helium**: Distance in years for helium-filled football.

**Details**


**Source**


**Examples**

```r
data(helium)
par(mfrow = c(1,2))
boxPlot(helium$air, xlab = "air")
boxPlot(helium$helium, xlab = "helium")
```

---

**histPlot**

*Histogram or hollow histogram*

**Description**

Create histograms and hollow histograms. This function permits easy color and appearance customization.

**Usage**

```r
histPlot(x,
  col = fadeColor("black", "22"), border = "black",
  breaks = "default", probability = FALSE,
  hollow = FALSE, add = FALSE, lty = 2, lwd = 1,
  freqTable=FALSE, right=TRUE, axes = TRUE,
  xlab = NULL, ylab = NULL, xlim = NULL, ylim = NULL,
  ...)"
```
**Arguments**

- **x**  
  Numerical vector or a frequency table (matrix) where the first column represents the observed values and the second column the frequencies. See also freqTable argument.

- **col**  
  Shading of the histogram bins.

- **border**  
  Color of histogram bin borders.

- **breaks**  
  A vector for the bin boundaries or an approximate number of bins.

- **probability**  
  If FALSE, the frequency is plotted. If TRUE, then a probability density.

- **hollow**  
  If TRUE, a hollow histogram will be created.

- **add**  
  If TRUE, the histogram is added to the plot.

- **lty**  
  Line type. Applies only if hollow=TRUE.

- **lwd**  
  Line width. Applies only if hollow=TRUE.

- **freqTable**  
  Set to TRUE if x is a frequency table.

- **right**  
  Set to FALSE to assign values of x that fall on a bin margin to the left bin. Otherwise the ties default to the right bin.

- **axes**  
  If FALSE, the axes are not plotted.

- **xlab**  
  Label for the x axis.

- **ylab**  
  Label for the y axis.

- **xlim**  
  Limits for the x axis.

- **ylim**  
  Limits for the y axis.

- ...  
  Additional arguments to plot. If add is TRUE, these arguments are ignored.

**Author(s)**

David Diez

**See Also**

boxPlot, dotPlot, densityPlot

**Examples**

data(run10)
par(mfrow=c(2,2))
histPlot(run10$gender=='M', probability=TRUE, xlab=c(30, 180),
          ylab=c(0, 0.025), hollow=TRUE)
histPlot(run10$gender=='F', probability=TRUE, add=TRUE,
          hollow=TRUE, lty=3, border='red')
legend('topleft', col=c('black', 'red'), lty=2:3, legend=c('M', 'F'))
histPlot(run10$gender=='F', probability=TRUE, border='darkblue',
          breaks=c(0, 100, 200), col=fadeColor('yellow', '33'))
legend('topleft', col=c('black', 'red'), lty=2:3, legend=c('M', 'F'))
histPlot(run10$gender=='M', probability=TRUE, breaks=brks,
          col=fadeColor('darkgoldenrod4', '33'))
Description

The make-up of the United States House of Representatives every two years since 1789. The last Congress included is the 112th Congress, which completes its term in 2013.

Usage

data(house)

Format

A data frame with 112 observations on the following 12 variables.

congress  The number of that year’s Congress
yearStart  Starting year
yearEnd  Ending year
seats  Total number of seats
p1  Name of the first political party
np1  Number of seats held by the first political party
p2  Name of the second political party
np2  Number of seats held by the second political party
other  Other
vac  Vacancy
del  Delegate
res  Resident commissioner

Source

Office of the Clerk of the U.S. House of Representatives Party Divisions:

http://clerk.house.gov/art_history/house_history/partyDiv.html

Data for Congresses 1-111 was recorded from the website above on November 1st, 2010. It appears this page was later moved to

http://artandhistory.house.gov/house_history/partyDiv.aspx (no longer available)

where data for Congress 112 was recorded on April 21, 2011.
Examples

data(house)

# Examine two-party relationship since 1855
th <- 34:112
COL <- c("#EEDDDB", "#DDEEBB", "#DDDDDD",
        "#BBDEEE", "#EEEEEE", "#EECCCC")
party <- c("#2222FF", "#FF2222")
par(las=1)
plot(house$yearStart[these], 100*house$np1[these]/house$seats[these],
     type="n", xlab="Year", ylab="Percent of House seats", ylim=c(11, 93))
rect(1861.3, -1000, 1865.3, 1000, col=COL[1], border="FFFFFF")
rect(1914.5, -1000, 1918.9, 1000, col=COL[2], border="FFFFFF")
rect(1929, -1000, 1939, 1000, col=COL[3], border="FFFFFF")
rect(1939.7, -1000, 1945.6, 1000, col=COL[4], border="FFFFFF")
rect(1955.8, -1000, 1965.3, 1000, col=COL[5], border="E2E2E2")
rect(1965.3, -1000, 1975.4, 1000, col=COL[6], border="E2E2E2")
lines(house$yearStart[these], 100*house$np1[these]/house$seats[these],
      col=party[1])
lines(house$yearStart[these], 100*house$np2[these]/house$seats[these],
      col=party[2])
legend("topleft", lty=c(1,1), col=party,
      c("Democrats", "Republicans"), bg="FFFFFF")
legend("topright", fill=COL,
      c("Civil War", "World War I", "Great Depression", "World War II",
        "Vietnam War Start", "Vietnam War Escalated"),
      bg="FFFFFF", border="FFFFFF")

---

houseRace10  

**Election results for the 2010 U.S. House of Representatives races**

**Description**

Election results for the 2010 U.S. House of Representatives races

**Usage**

data(houseRace10)

**Format**

A data frame with 435 observations on the following 24 variables.

- **id** Unique identifier for the race, which does not overlap with other 2010 races (see `govRace10` and `senateRace10`)
- **state** State name
- **abbr** State name abbreviation
- **num** District number for the state
name1 Name of the winning candidate
perc1 Percentage of vote for winning candidate (if more than one candidate)
party1 Party of winning candidate
votes1 Number of votes for winning candidate
name2 Name of candidate with second most votes
perc2 Percentage of vote for candidate who came in second
party2 Party of candidate with second most votes
votes2 Number of votes for candidate who came in second
name3 Name of candidate with third most votes
perc3 Percentage of vote for candidate who came in third
party3 Party of candidate with third most votes
votes3 Number of votes for candidate who came in third
name4 Name of candidate with fourth most votes
perc4 Percentage of vote for candidate who came in fourth
party4 Party of candidate with fourth most votes
votes4 Number of votes for candidate who came in fourth
name5 Name of candidate with fifth most votes
perc5 Percentage of vote for candidate who came in fifth
party5 Party of candidate with fifth most votes
votes5 Number of votes for candidate who came in fifth

Details

This analysis in the Examples section was inspired by and is similar to that of Nate Silver’s district-level analysis on the FiveThirtyEight blog in the New York Times:


Source

Data was collected from MSNBC.com on November 9th, 2010.

Examples

data(houseRace10)
hr <- table(houseRace10[,c("abbr", "party")])
nr <- apply(hr, 1, sum)

data(prRace08)
pr <- prRace08[prRace08$state != "DC",c("state", "pObama")]
hr <- hr[as.character(pr$state),]
(fit <- glm(hr ~ pr$pObama, family=binomial))

x1 <- pr$pObama[match(houseRace10$abbr, pr$state)]
y1 <- (houseRace10$party == "Democrat")+0
housing

Simulated data set on student housing

Description
Each observation represents a simulated rent price for a student.

Usage
data("housing")

Format
A data frame with 75 observations on the following variable.

cost a numeric vector

Examples
data(housing)
## maybe str(housing) ; plot(housing) ...

hsb2
High School and Beyond survey

Description
Two hundred observations were randomly sampled from the High School and Beyond survey, a
survey conducted on high school seniors by the National Center of Education Statistics.
Usage

```
data(hsb2)
```

Format

A data frame with 200 observations and 11 variables.

- `id`: Student ID.
- `gender`: Student’s gender, with levels `female` and `male`.
- `race`: Student’s race, with levels `african american`, `asian`, `hispanic`, and `white`.
- `ses`: Socio economic status of student’s family, with levels `low`, `middle`, and `high`.
- `schtyp`: Type of school, with levels `public` and `private`.
- `prog`: Type of program, with levels `general`, `academic`, and `vocational`.
- `read`: Standardized reading score.
- `write`: Standardized writing score.
- `math`: Standardized math score.
- `science`: Standardized science score.
- `socst`: Standardized social studies score.

Source


Examples

```
data(hsb2)
boxPlot(hsb2$read ~ hsb2$write, fact = hsb2$gender,
       ylab = "diff. bet. reading and writing scores")
```

---

**husbands.wives**

*Married couples in Great Britain*

Description

The Great Britain Office of Population Census and Surveys collected a random sample of 170 married couples in Britain.

Usage

```
data("husbands.wives")
```
**immigration**

**Format**

A data frame with 199 observations on the following 8 variables.

- Age_Husband: a numeric vector
- Ht_Husband: a numeric vector
- Age_Wife: a numeric vector
- Ht_Wife: a numeric vector
- Age_Husb_at_Marriage: a numeric vector
- Years_Married: a numeric vector
- Age_Wife_At_Marriage: a numeric vector
- Duration: a factor with levels <= 20 >20

**Source**


**Examples**

```r
data(husbands.wives)
## maybe str(husbands.wives); plot(husbands.wives) ...
```

---

**immigration**

*Poll on illegal workers in the US*

**Description**

910 randomly sampled registered voters in Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country as well as their political ideology.

**Usage**

```r
data("immigration")
```

**Format**

A data frame with 910 observations on the following 2 variables.

- response: a factor with levels Apply for citizenship Guest worker Leave the country Not sure
- political: a factor with levels conservative liberal moderate

**Source**

SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.
Examples

data(immigration)
## maybe str(immigration); plot(immigration) ...

infMortRate  

Infant Mortality Rates, 2012

Description

This entry gives the number of deaths of infants under one year old in 2012 per 1,000 live births in the same year. This rate is often used as an indicator of the level of health in a country.

Usage

data(infMortRate)

Format

A data frame with 222 observations on the following 2 variables.

country Name of country.
infMortRate Infant mortality rate per 1,000 live births.

Details

The data is given in decreasing order of infant mortality rates. There are a few potential outliers.

Source


Examples

data(infMortRate)
histPlot(infMortRate$infMortRate)
boxPlot(infMortRate$infMortRate)
### ipod

**Length of songs on an iPod**

**Description**
A simulated data set on lengths of songs on an iPod.

**Usage**
data(ipod)

**Format**
A data frame with 3000 observations on the following variable.

- **songLength** Length of song (in minutes).

**Source**
Simulated data.

**Examples**
data(ipod)
histPlot(ipod$songLength)

---

### iris

*Iris data (flowers)*

**Description**
Classical data set from R.A. Fisher.

**Usage**
data("iris")

**Format**
A data frame with 150 observations on the following 5 variables.

- **Sepal.Length** a numeric vector
- **Sepal.Width** a numeric vector
- **Petal.Length** a numeric vector
- **Petal.Width** a numeric vector
- **Species** a factor with levels *setosa* *versicolor* *virginica*
**Source**


**Examples**

```r
data(iris)
## maybe str(iris); plot(iris) ...
```

---

**jury**  
**Simulated juror data set**

**Description**

Simulated data set of registered voters proportions and representation on juries.

**Usage**

```r
data("jury")
```

**Format**

A data frame with 275 observations on the following variable.

- **race** a factor with levels black hispanic other white

**Examples**

```r
data(jury)
## maybe str(jury); plot(jury) ...
```

---

**leg_mari**  
**Legalization of Marijuana Support in 2010 California Survey**

**Description**

In a 2010 Survey USA poll, 70

**Usage**

```r
data("leg_mari")
```

**Format**

A data frame with 119 observations on the following variable.

- **response** One of two values: oppose or support.
linResPlot

Source
Survey USA, Election Poll #16804, data collected July 8-11, 2010.

Examples

data(leg_mari)
table(leg_mari)

linResPlot Create simple regression plot with residual plot

Description
Create a simple regression plot with residual plot.

Usage
linResPlot(x, y, axes = FALSE, wBox = TRUE, wLine = TRUE,
lCol = "#00000088", lty = 1, lwd = 1,
main = "", xlab = "", ylab = "", marRes = NULL,
col = fadeColor(4, "88"), pch = 20, cex = 1.5,
yR = 0.1, ylim = NULL, subset = NULL, ...)

Arguments
x Predictor variable.
y Outcome variable.
axes Whether to plot axis labels.
wBox Whether to plot boxes around each plot.
wLine Add a regression line.
lCol Line color.
lty Line type.
lwd Line width.
main Title for the top plot.
xlab x-label.
ylab y-label.
marRes Margin for the residuals plot.
col Color of the points.
pch Plotting character of points.
cex Size of points.
yR An additional vertical stretch factor on the plot.
ylim y-limits.
subset Boolean vector, if wanting a subset of the data.
... Additional arguments passed to both plots.
See Also

`makeTube`

Examples

```r
# Currently seems broken for this example.
n <- 25
x <- runif(n)
y <- 5 * x + rnorm(n)
myMat <- rbind(matrix(1:2, 2))
myW <- 1
myH <- c(1, 0.45)
par(mar = c(0.35, 0.654, 0.35, 0.654))
layout(myMat, myW, myH)
linResPlot(x, y, col = COL[1, 2])
```

---

**lmPlot**

*Linear regression plot with residual plot*

**Description**

Plot data, the linear model, and a residual plot simultaneously.

**Usage**

```r
lmPlot(x, y, xAxis = 0, yAxis = 4, resAxis = 3, resSymm = TRUE,
       wBox = TRUE, wLine = TRUE, lCol = "#00000000", lty = 1,
       lwd = 1, xlab = "", ylab = "", marRes = NULL,
       col = "#22558888", pch = 20, cex = 1.5, xR = 0.02,
       yR = 0.1, xlim = NULL, ylim = NULL, subset = NULL,
       parCustom = FALSE, myHeight = c(1, 0.45),
       plots = c("both", "mainOnly", "resOnly"), highlight = NULL,
       hlCol = NULL, hlCex = 1.5, hlPch = 20, na.rm=TRUE, ...)
```

**Arguments**

- `x` The x coordinates of points in the plot.
- `y` The y coordinates of points in the plot.
- `xAxis` The maximum number of x axis labels.
- `yAxis` The maximum number of y axis labels.
- `resAxis` The maximum number of y axis labels in the residual plot.
- `resSymm` Boolean determining whether the range of the residual plot should be symmetric about zero.
- `wBox` Boolean determining whether a box should be added around each plot.
- `wLine` Boolean determining whether to add a regression line to the plot.
<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>lCol</code></td>
<td>The color of the regression line to be added.</td>
</tr>
<tr>
<td><code>lty</code></td>
<td>The line type of the regression line to be added.</td>
</tr>
<tr>
<td><code>lwd</code></td>
<td>The line width of the regression line to be added.</td>
</tr>
<tr>
<td><code>xlab</code></td>
<td>A label for the x axis.</td>
</tr>
<tr>
<td><code>ylab</code></td>
<td>A label for the y axis</td>
</tr>
<tr>
<td><code>marRes</code></td>
<td>Margin specified for the residuals.</td>
</tr>
<tr>
<td><code>col</code></td>
<td>Color of points.</td>
</tr>
<tr>
<td><code>pch</code></td>
<td>Plotting character.</td>
</tr>
<tr>
<td><code>cex</code></td>
<td>Plotting character size.</td>
</tr>
<tr>
<td><code>xR</code></td>
<td>Scaling the limits of the x axis. Ignored if <code>xlim</code> specified.</td>
</tr>
<tr>
<td><code>yR</code></td>
<td>Scaling the limits of the y axis. Ignored if <code>ylim</code> specified.</td>
</tr>
<tr>
<td><code>xlim</code></td>
<td>Limits for the x axis.</td>
</tr>
<tr>
<td><code>ylim</code></td>
<td>Limits for the y axis.</td>
</tr>
<tr>
<td><code>subset</code></td>
<td>A subset of the data to be used for the linear model.</td>
</tr>
<tr>
<td><code>parCustom</code></td>
<td>If TRUE, then the plotting margins are not modified automatically. This value should also be TRUE if the plots are being placed within a plot of multiple panels.</td>
</tr>
<tr>
<td><code>myHeight</code></td>
<td>A numerical vector of length 2 representing the ratio of the primary plot to the residual plot, in height.</td>
</tr>
<tr>
<td><code>plots</code></td>
<td>Not currently utilized.</td>
</tr>
<tr>
<td><code>highlight</code></td>
<td>Numerical vector specifying particular points to highlight.</td>
</tr>
<tr>
<td><code>hlCol</code></td>
<td>Color of highlighted points.</td>
</tr>
<tr>
<td><code>hlCex</code></td>
<td>Size of highlighted points.</td>
</tr>
<tr>
<td><code>h1Pch</code></td>
<td>Plotting characters of highlighted points.</td>
</tr>
<tr>
<td><code>na.rm</code></td>
<td>Remove cases with NA values.</td>
</tr>
<tr>
<td><code>...</code></td>
<td>Additional arguments to <code>plot</code>.</td>
</tr>
</tbody>
</table>

**Author(s)**

David M Diez <david.m.diez@gmail.com>

**See Also**

- `makeTube`

**Examples**

```r
data(satGPA)
lmPlot(satGPA$SATSum, satGPA$FYGPA)

data(gradesTV)
lmPlot(gradesTV$TV, gradesTV$Grades, xAxis=4, xlab='time watching TV', yR=0.2, highlight=c(1,15,20))
```
loop

Output a message while inside a loop

Description

(Note: `txtProgressBar` and `setTxtProgressBar` are better.) Output a message while inside a for loop to update the user on progress. This function is useful in tracking progress when the number of iterations is large or the procedures in each iteration take a long time.

Usage

```r
loop(i, n = NULL, every = 1, extra=NULL)
```

Arguments

- `i` The index value used in the loop.
- `n` The last entry in the loop.
- `every` The number of loops between messages.
- `extra` Additional information to print.

Author(s)

David M Diez

See Also

`myPDF`

Examples

```r
for(i in 1:60){
  loop(i, 160, 20, paste("iter", i))
}
```

lsegments

Create a Line Segment Plot

Description

Create a simple plot showing a line segment.

Usage

```r
lsegments(x = c(3, 7), l = "o", r = "c", ticks = TRUE, labs = 1,
          add = 0, ylim = c(-0.75, 0.25))
```
Arguments

x  The endpoints of the interval. Values larger (smaller) than 999 (-999) will be interpreted as (negative) infinity.

l  Indicate whether the left end point should be open ("o") or closed ("c").

r  Indicate whether the right end point should be open ("o") or closed ("c").

ticks  Indicate whether to show tick marks (TRUE) or not (FALSE).
labs  The position for the point labels. Set to 0 if no labels should be shown.

add  Indicate whether the line segment should be added to an existing plot (TRUE) or a new plot should be created (FALSE).

ylim  A vector of length 2 specifying the vertical plotting limits, which may be useful for fine-tuning plots. The default is c(-0.75, 0.25).

Author(s)

David Diez

References

OpenIntro, openintro.org

See Also

dlsegments, CCP, ArrowLines

Examples

lsegments(c(2,7), "o", "c", ylim=c(-0.3, 0.2))

lsegments(c(5,7), "c", "c", ylim=c(-0.3, 0.2))

lsegments(c(4,1000), "o", "o", ylim=c(-0.3, 0.2))

Description

Produce a linear, quadratic, or nonparametric tube for regression data.

Usage

makeTube(x, y, Z=2, R=1, col='#00000022', border='#00000000',
type=c('lin', 'quad', 'robust'), stDev=c('constant', 'linear', 'other'),
length.out=99, bw='default', plotTube=TRUE, addLine=TRUE, ...)

Arguments

- **x**: x coordinates.
- **y**: y coordinates.
- **Z**: Number of standard deviations out from the regression line to extend the tube.
- **R**: Control of how far the tube extends to the left and right.
- **col**: Fill color of the tube.
- **border**: Border color of the tube.
- **type**: The type of model fit to the data. Here ‘robust’ results in a nonparametric estimate.
- **stDev**: Choices are constant variance (‘constant’), the standard deviation of the errors changes linearly (‘linear’), or the standard deviation of the errors should be estimated using nonparametric methods (‘other’).
- **length.out**: The number of observations used to build the regression model. This argument may be increased to increase the smoothing of a quadratic or nonparametric curve.
- **bw**: Bandwidth used if type='robust' or homosk=FALSE.
- **plotTube**: Whether the tube should be plotted.
- **addLine**: Whether the linear model should be plotted.
- **...**: Additional arguments passed to the *lines* function if addLine=TRUE.

Value

- **X**: x coordinates for the regression model.
- **Y**: y coordinates for the regression model.
- **tubeX**: x coordinates for the boundary of the tube.
- **tubeY**: y coordinates for the boundary of the tube.

Author(s)

David M Diez

See Also

*lmPlot*

Examples

```r
### possum example
data(possum)
x <- possum$total
y <- possum$headL
plot(x,y)
maketube(x,y,1)
maketube(x,y,2)
maketube(x,y,3)
```
males.heights

Sample of 100 male heights

Description
Random sample based on Food Commodity Intake Database distribution

Usage
data("male.heights")

Format
A data frame with 100 observations on the following variable.

heights a numeric vector
References

http://fcid.foodrisk.org/

Examples

data(male.heights)
## maybe str(male.heights) ; plot(male.heights) ...

mammals

Sleep in Mammals

Description

This data set includes data for 39 species of mammals distributed over 13 orders. The data were used for analyzing the relationship between constitutional and ecological factors and sleeping in mammals. Two qualitatively different sleep variables (dreaming and non dreaming) were recorded. Constitutional variables such as life span, body weight, brain weight and gestation time were evaluated. Ecological variables such as severity of predation, safety of sleeping place and overall danger were inferred from field observations in the literature.

Usage

data(mammals)

Format

A data frame with 62 observations on the following 11 variables.

Species  Species of mammals
BodyWt   Total body weight of the mammal (in kg)
BrainWt  Brain weight of the mammal (in kg)
NonDreaming  Number of hours of non dreaming sleep
Dreaming  Number of hours of dreaming sleep
TotalSleep  Total number of hours of sleep
LifeSpan  Life span (in years)
Gestation  Gestation time (in days)
Predation  An index of how likely the mammal is to be preyed upon. 1 = least likely to be preyed upon. 5 = most likely to be preyed upon.
Exposure  An index of how exposed the mammal is during sleep. 1 = least exposed (e.g., sleeps in a well-protected den). 5 = most exposed.
Danger  An index of how much danger the mammal faces from other animals. This index is based upon Predation and Exposure. 1 = least danger from other animals. 5 = most danger from other animals.
mammogram

Source

http://www.statsci.org/data/general/sleep.txt

References


Examples

data(mammals)
lmPlot(log(mammals$BodyWt), log(mammals$BrainWt))

---

mammogram Experiment with Mammogram Randomized

Description

An experiment where 89,835 women were randomized to either get a mammogram or a non-mammogram breast screening. The response measured was whether they had died from breast cancer within 25 years.

Usage

data("mammogram")

Format

A data frame with 89835 observations on the following 2 variables.

treatment a factor with levels control mammogram
breast_cancer_death a factor with levels no yes

Source


Examples

data(mammogram)
table(mammogram)
chisq.test(table(mammogram))
**marioKart**

---

**mariokart**

**Wii Mario Kart auctions from Ebay**

---

**Description**

Auction data from Ebay for the game Mario Kart for the Nintendo Wii. This data was collected in early October, 2009.

**Usage**

data(marioKart)

---

**marathon**

*New York City Marathon Times*

---

**Description**


**Usage**

data(marathon)

**Format**

A data frame with 60 observations on the following 3 variables.

- **Year**  Year
- **Gender**  Gender
- **Time**  Running time (in hours)

**Source**

http://www.webcitation.org/5kx7i1F8p

**Examples**

data(marathon)
str(marathon)
histPlot(marathon$Time)
boxPlot(marathon$Time, horiz = TRUE, fact = marathon$Gender)
Format

A data frame with 143 observations on the following 12 variables. All prices are in US dollars.

- **id**: Auction ID assigned by Ebay.
- **duration**: Auction length, in days.
- **nBids**: Number of bids.
- **cond**: Game condition, either new or used.
- **startPr**: Start price of the auction.
- **shipPr**: Shipping price.
- **totalPr**: Total price, which equals the auction price plus the shipping price.
- **shipSp**: Shipping speed or method.
- **sellerRate**: The seller’s rating on Ebay. This is the number of positive ratings minus the number of negative ratings for the seller.
- **stockPhoto**: Whether the auction feature photo was a stock photo or not. If the picture was used in many auctions, then it was called a stock photo.
- **wheels**: Number of Wii wheels included in the auction. These are steering wheel attachments to make it seem as though you are actually driving in the game. When used with the controller, turning the wheel actually causes the character on screen to turn.
- **title**: The title of the auctions.

Details

There are several interesting features in the data. First off, note that there are two outliers in the data. These serve as a nice example of what one should do when encountering an outlier: examine the data point and remove it only if there is a good reason. In these two cases, we can see from the auction titles that they included other items in their auctions besides the game, which justifies removing them from the data set.

This data set includes all auctions for a full week in October, 2009. Auctions were included in the data set if they satisfied a number of conditions. (1) They were included in a search for “wii mario kart” on ebay.com, (2) items were in the Video Games > Games > Nintendo Wii section of Ebay, (3) the listing was an auction and not exclusively a “Buy it Now” listing (sellers sometimes offer an optional higher price for a buyer to end bidding and win the auction immediately, which is an optional Buy it Now auction), (4) the item listed was the actual game, (5) the item was being sold from the US, (6) the item had at least one bidder, (7) there were no other items included in the auction with the exception of racing wheels, either generic or brand-name being acceptable, and (8) the auction did not end with a Buy It Now option.

References

http://www.ebay.com/
http://www.openintro.org/
Examples

data(marioKart)

# Identify the outliers
boxPlot(marioKart$totalPr, marioKart$cond, horiz=TRUE)
toss <- which(marioKart$totalPr > 80)
lines(rep(marioKart$totalPr[toss[1]], 2), c(2.4, 2))
text(marioKart$totalPr[toss[1]]-55, 2.4, marioKart$title[toss[1]],
pos=3, cex=0.5)
lines(rep(marioKart$totalPr[toss[2]], 2), c(1.6, 2))
text(marioKart$totalPr[toss[2]], 1.6, marioKart$title[toss[2]],
pos=1, cex=0.5)
marioKart[toss, ]

# the other two points marked on the boxplot are legitimate auctions

# Replot without the outliers
boxPlot(marioKart$totalPr[-toss], marioKart$cond[-toss], horiz=TRUE)

# Fit a Multiple Regression Model
mk <- marioKart[-toss,]
summary(lm(totalPr ~ cond + stockPhoto + duration + wheels, mk))
summary(lm(totalPr ~ cond + stockPhoto + wheels, mk))
summary(fit <- lm(totalPr ~ cond + wheels, mk))

# Fit Diagnostics

e <- fit$res
f <- fit$fit
par(mfrow=c(2,3), mar=c(4, 4, 2, 1))
qqnorm(e, ylab="Residuals", main="")
plot(e, xlab="Order of collection", ylab="Residuals")
plot(f, e, xlab="Fitted values", ylab="Residuals")
plot(f, (abs(e)), xlab="Fitted values", ylab="Absolute value of residuals")
boxPlot(e, mk$cond, xlab="Condition", ylab="Residuals")
plot(mk$wheels, e, xlab="Number of wheels", ylab="Residuals",
main="Notice curvature")

migraine

Migraines and acupuncture

Description

Experiment involving acupuncture and sham acupuncture (as placebo) in the treatment of migraines.

Usage

data("migraine")
MLB

Format

A data frame with 89 observations on the following 2 variables.

- group: a factor with levels control treatment
- pain_free: a factor with levels no yes

Source


Examples

```r
data(migraine)
## maybe str(migraine); plot(migraine) . . .
```

MLB

Salary data for Major League Baseball (2010)

Description

Salary data for Major League Baseball players in the year 2010.

Usage

```r
data(MLB)
```

Format

A data frame with 828 observations on the following 4 variables.

- player: Player name
- team: Team
- position: Field position
- salary: Salary (in $1000s)

Source

Collected from the following page (and its linked pages) on February 23rd, 2011:

http://content.usatoday.com/sportsdata/baseball/mlb/salaries/team
Examples

data(MLB)

# _____ Basic Histogram _____ #
hist(MLB$salary / 1000, breaks = 15,
    main = "", xlab = "Salary (millions of dollars)", ylab = "",
    axes = FALSE,
    col = "#22558844")
axis(1, seq(0, 40, 10))
axis(2, c(0, 500))
axis(2, seq(100, 400, 100), rep("", 4), tcl = -0.2)

# _____ Histogram on Log Scale _____ #
hist(log(MLB$salary / 1000), breaks=15,
    main = "", xlab = "log(Salary)", ylab = "",
    axes = FALSE, col = "#22558844")
axis(1) #, seq(0, 40, 10))
axis(2, seq(0, 300, 100))

# _____ Box plot of log(salary) against position _____ #
par(las = 1, mar = c(4, 8, 1, 1))
boxPlot(log(MLB$salary / 1000), MLB$position, horiz = TRUE, ylab = "")

mlbBat10

Major League Baseball Player Hitting Statistics for 2010

Description


Usage

data(mlbBat10)

Format

A data frame with 1199 observations on the following 19 variables.

name Player name
team Team abbreviation
position Player position
G Number of games
AB Number of at bats
R Number of runs
H Number of hits
2B Number of doubles
3B  Number of triples
HR  Number of home runs
RBI  Number of runs batted in
TB  Total bases, computed as $3*HR + 2*3B + 1*2B + H$
BB  Number of walks
SO  Number of strikeouts
SB  Number of stolen bases
CS  Number of times caught stealing
OBP  On base percentage
SLG  Slugging percentage (TB / AB)
AVG  Batting average

Source

Data was collected from MLB.com on April 22nd, 2011.

Examples

data(mlbBat10)
d <- mlbBat10[mlbBat10$AB > 200,]
pos <- c("OF", "1B", "2B", "3B", "SS", "DH", "C")
POS <- c("OF", "1B", "2B", "3B")

# On-base Percentage Across Positions
out <- c()
gp <- c()
for(i in 1:length(pos)){
  these <- which(d$pos %in% pos[i])
  out <- c(out, d[these,"OBP"])
  gp <- c(gp, rep(POS[i], length(these)))
}
plot(out ~ as.factor(gp))
summary(lm(out ~ as.factor(gp)))
anova(lm(out ~ as.factor(gp)))

# Batting Average Across Positions
out <- c()
gp <- c()
for(i in 1:length(pos)){
  these <- which(d$pos %in% pos[i])
  out <- c(out, d[these,"AVG"])
  gp <- c(gp, rep(POS[i], length(these)))
}
plot(out ~ as.factor(gp))
summary(lm(out ~ as.factor(gp)))
anova(lm(out ~ as.factor(gp)))

# Home Runs Across Positions

MosaicPlot

Custom Mosaic Plot

Description

Plot a mosaic plot custom built for a particular figure.

Usage

MosaicPlot(formula, data, col = "#00000022", border = 1,
            dir = c("v", "h"), off = 0.01, cex.axis = 0.7,
            col.dir = "v", flip = c("v"), ...)

Arguments

- **formula**: Formula describing the variable relationship.
- **data**: Data frame for the variables, optional.
- **col**: Colors for plotting.
- **border**: Ignored.
- **dir**: Ignored.
- **off**: Fraction of white space between each box in the plot.
- **cex.axis**: Axis label size.
- **col.dir**: Direction to lay out colors.
Whether to flip the ordering of the vertical ("v") and/or horizontal ("h") ordering in the plot.

Ignored.

Author(s)
David Diez (david@openintro.org)

Examples
data(email)
data(COL)
email$spam <- ifelse(email$spam == 0, "not\ns\spam", "\spam")
par(las = 1)
MosaicPlot(number ~ spam, email, col = COL[1:3], off = 0.02)

Description
A similar function to pdf and png, except that different defaults are provided, including for the plotting parameters.

Usage
myPDF(fileName, width = 5, height = 3,
      mar = c(3.9, 3.9, 1, 1),
      mgp = c(2.8, 0.55, 0),
      las = 1, tcl=-0.3, ...)

myPNG(fileName, width = 600, height = 400,
       mar = c(3.9, 3.9, 1, 1),
       mgp = c(2.8, 0.55, 0),
       las = 1, tcl=-0.3, ...)

Arguments

fileName File name for the image to be output. The name should end in .pdf.
width The width of the image file (inches). Default: 5.
height The height of the image file (inches). Default: 3.
mar Plotting margins. To change, input a numerical vector of length 4.
mgp Margin graphing parameters. To change, input a numerical vector of length 3. The first argument specifies where x and y labels are placed; the second specifies the axis labels are placed; and the third specifies how far to pull the entire axis from the plot.
las  Orientation of axis labels. Input 0 for the default.
tcl  The tick mark length as a proportion of text height. The default is ~0.5.
...  Additional arguments to par.

Author(s)
David M Diez

See Also
edaPlot

Examples

data(marioKart)
#=>> Save a plot to a PDF <==>
# myPDF("myPlot.pdf")
data(run10)
histPlot(run10$time)
# dev.off()

#=>> Save a plot to a PNG <==>
# myPNG("myPlot.png")
data(run10)
histPlot(run10$time)
# dev.off()

nba.heights  *NBA Player heights from 2008-9*

Description
Heights of all NBA players from the 2008-9 season.

Usage
data("nba.heights")

Format
A data frame with 435 observations (players) on the following 4 variables.

last.name  Last name.
first.name  First name.
h.meters  Height, in meters.
h.in  Height, in inches.
ncbirths

Source

Collected from \url{http://www.nba.com}.

Examples

\begin{verbatim}
data(nba.heights)
qqnorm(nba.heights$h.meters)
\end{verbatim}

\begin{verbatim}
ncbirths
\end{verbatim}

\textit{North Carolina births}

Description

In 2004, the state of North Carolina released to the public a large data set containing information on births recorded in this state. This data set has been of interest to medical researchers who are studying the relation between habits and practices of expectant mothers and the birth of their children. This is a random sample of 1,000 cases from this data set.

Usage

\begin{verbatim}
data(ncbirths)
\end{verbatim}

Format

A data frame with 1000 observations on the following 13 variables.

\begin{verbatim}
fage  Father's age in years.
mage  Mother's age in years.
mature Maturity status of mother.
weeks Length of pregnancy in weeks.
premie Whether the birth was classified as premature (premie) or full-term.
visits Number of hospital visits during pregnancy.
gained Weight gained by mother during pregnancy in pounds.
weight Weight of the baby at birth in pounds.
lowbirthweight Whether baby was classified as low birthweight (low) or not (not low).
gender Gender of the baby, female or male.
habit Status of the mother as a nonsmoker or a smoker.
marital Whether mother is married or not married at birth.
whitemom Whether mom is white or not white.
\end{verbatim}

Examples

\begin{verbatim}
data(ncbirths)
boxPlot(ncbirths$weight, fact = ncbirths$habit)
boxPlot(ncbirths$visits, fact = ncbirths$whitemom)
boxPlot(ncbirths$gained, fact = ncbirths$mature)
\end{verbatim}
**normTail**

*Normal distribution tails*

**Description**

Produce a normal (or t) distribution and shaded tail.

**Usage**

```r
normTail(m = 0, s = 1, L = NULL, U = NULL, M = NULL, df = 1000,
         curveColor = 1, border = 1, col = "#CCCCCC", xlim = NULL,
         ylim = NULL, xlab = "", ylab = "", digits = 2, axes = 1,
         detail = 999, xLab = c("number", "symbol"), cex.axis = 1,
         xAxisIncr = 1, add = FALSE, ...)
```

**Arguments**

- **m**
  - Numerical value for the distribution mean.
- **s**
  - Numerical value for the distribution standard deviation.
- **L**
  - Numerical value representing the cutoff for a shaded lower tail.
- **U**
  - Numerical value representing the cutoff for a shaded upper tail.
- **M**
  - Numerical value representing the cutoff for a shaded central region.
- **df**
  - Numerical value describing the degrees of freedom. Default is 1000, which results in a nearly normal distribution. Small values may be useful to emphasize small tails.
- **curveColor**
  - The color for the distribution curve.
- **border**
  - The color for the border of the shaded area.
- **col**
  - The color for filling the shaded area.
- **xlim**
  - Limits for the x axis.
- **ylim**
  - Limits for the y axis.
- **xlab**
  - A title for the x axis.
- **ylab**
  - A title for the y axis.
- **digits**
  - The maximum number of digits past the decimal to use in axes values.
- **axes**
  - A numeric value denoting whether to draw both axes (3), only the vertical axes (2), only the horizontal axes (1, the default), or no axes (0).
- **detail**
  - A number describing the number of points to use in drawing the normal curve. Smaller values correspond to a less smooth curve but reduced memory usage in the final file.
- **xLab**
  - If "number", then the axis is drawn at the mean, and every standard deviation out until the third standard deviation. If "symbol", then Greek letters are used for standard deviations from three standard deviations from the mean.
- **cex.axis**
  - Numerical value controlling the size of the axis labels.
Description

A simple random sample of 1,028 US adults in March 2013 found that 56% support nuclear arms reduction.

Usage

data("nuclear.survey")

Format

A data frame with 1028 observations on the following variable.

arms_reduction Responses of favor or against.

Source


Examples

data(nuclear.survey)
table(nuclear.survey)
Description
A 2010 survey asking a randomly sample of registered voters in California for their position on drilling for oil and natural gas off the Coast of California.

Usage
data("offshore.drilling")

Format
A data frame with 827 observations on the following 2 variables.

position a factor with levels do not know oppose support
college_grad a factor with levels no yes

Source
Survey USA, Election Poll #16804, data collected July 8-11, 2010.

Examples
data(offshore.drilling)
## maybe str(offshore.drilling); plot(offshore.drilling) ...
**Format**

A data frame with 23 observations on the following 2 variables.

- **temp** Temperature, in Fahrenheit.
- **damage** Number of damaged O-rings (out of 6).

**Source**

https://archive.ics.uci.edu/ml/datasets/Challenger+USA+Space+Shuttle+O-Ring

**Examples**

```r
data(orings)
## maybe str(orings); plot(orings) ...
```

---

**oscars**  
*Oscar winners, 1929 to 2012*

---

**Description**

Best actor and actress Oscar winners from 1929 to 2012.

**Usage**

```r
data(oscars)
```

**Format**

A data frame with 170 observations on the following 10 variables.

- **gender** Gender of winner, female or male.
- **oscar_no** Denotes which Oscar ceremony.
- **oscar_yr** Denotes which Oscar year.
- **name** Name of winning actor or actress.
- **movie** Name of movie actor or actress got the Oscar for.
- **age** Age at which the actor or actress won the Oscar.
- **birth_pl** State where the actor or actress was born, country if foreign.
- **birth_mo** Birth month of actor or actress.
- **birth_d** Birth day of actor or actress.
- **birth_y** Birth year of actor or actress.

**Details**

Although there have been only 84 Oscar ceremonies until 2012, there are 85 male winners and 85 female winners because ties happened on two occasions (1933 for the best actor and 1969 for the best actress).
Source


Examples

data(oscars)
boxPlot(oscars$age, oscars$gender)
barplot(oscars$birth_mo)
barplot(table(oscars$birth_pl))

---

outliers              Simulated data sets for different types of outliers

Description

Data sets for showing different types of outliers

Usage

data("outliers")

Format

A data frame with 50 observations on the following 5 variables.

- x  a numeric vector
- y  a numeric vector
- xInf  a numeric vector
- yLev  a numeric vector
- yOut  a numeric vector

Examples

data(outliers)
## maybe str(outliers) ; plot(outliers) ...
PlotWLine

Plot data and add a regression line.

Description

Plot data and add a regression line.

Usage

\[
\text{PlotWLine}(x, y, xlab = "", ylab = "", \\
\quad \text{col} = \text{fadeColor}(4, "88"), \text{cex} = 1.2, \\
\quad \text{pch} = 20, n = 4, \text{nMax} = 4, \\
\quad yR = 0.1, \text{axes} = \text{TRUE}, \ldots)
\]

Arguments

- **x**: Predictor variable.
- **y**: Outcome variable.
- **xlab**: x-axis label.
- **ylab**: y-axis label.
- **col**: Color of points.
- **cex**: Size of points.
- **pch**: Plotting character.
- **n**: The preferred number of axis labels.
- **nMax**: The maximum number of axis labels.
- **yR**: y-limit buffer factor.
- **axes**: Boolean to indicate whether or not to include axes.
- **...**: Passed to `plot`.

See Also

- `maketube`

Examples

\[
\text{PlotWLine}(1:10, \text{seq}(-5, -2, \text{length.out} = 10) + \text{rnorm}(10))
\]
**Description**

Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency.

**Usage**

```r
data("pm2.2011.durham")
```

**Format**

A data frame with 449 observations on the following 20 variables.

- **Date** Date
- **AQS_SITE_ID** a factor with levels 37-063-0015
- **POC** a numeric vector
- **Daily.Mean.PM2.5.Concentration** a numeric vector
- **UNITS** a factor with levels ug/m3 LC
- **DAILY_AQI_VALUE** a factor with levels 12 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 43 44 45 46 48 49 52 54 55 56 57 60 62 65 9
- **DAILY_OBS_COUNT** a numeric vector
- **PERCENT_COMPLETE** a numeric vector
- **AQS_PARAMETER_CODE** a numeric vector
- **AQS_PARAMETER_DESC** a factor with levels Acceptable PM2.5 AQI & Speciation Mass PM2.5 – Local Conditions
- **CSA_CODE** a numeric vector
- **CSA_NAME** a factor with levels Raleigh–Durham–Cary, NC
- **CBSA_CODE** a numeric vector
- **CBSA_NAME** a factor with levels Durham, NC
- **STATE_CODE** a numeric vector
- **STATE** a factor with levels North Carolina
- **COUNTY_CODE** a numeric vector
- **COUNTY** a factor with levels Durham
- **SITE_LATITUDE** a numeric vector
- **SITE_LONGITUDE** a numeric vector

**Source**

### Poker

**Examples**

```r
data(pm25.2011.durham)
## maybe str(pm25.2011.durham); plot(pm25.2011.durham) ...
```

**Description**

Poker winnings (and losses) for 50 days by a professional poker player.

**Usage**

```r
data(poker)
```

**Format**

A data frame with 49 observations on the following variable.

- **winnings**  
  Poker winnings and losses, in US dollars.

**Source**

Anonymity has been requested by the player.

**References**


**Examples**

```r
data(poker)
histPlot(poker$winnings)
```

---

### Possum

**Description**

Data representing possums in Australia and New Guinea. This is a copy of the data set by the same name in the DAAG package, however, the data set included here includes fewer variables.

**Usage**

```r
data(possum)
```
**Format**

A data frame with 104 observations on the following 8 variables.

- **site**: The site number where the possum was trapped.
- **pop**: Population, either **vic** (Victoria) or **other** (New South Wales or Queensland).
- **sex**: Gender, either **m** (male) or **f** (female).
- **age**: Age.
- **headl**: Head length, in mm.
- **skullw**: Skull width, in mm.
- **totall**: Total length, in cm.
- **taill**: Tail length, in cm.

**Source**


**References**


**Examples**

```r
data(possum)
par(mfrow=1:2)
plot(possum$headl, possum$skullw)
densityPlot(possum$totall, possum$sex, key=c('f','m'),
xlab='total length (cm)')
legend('topright', col=c('black', 'red'), lty=1:2, legend=c('f', 'm'))
```

---

**president**  
*United States Presidential History*

**Description**

Summary of the changes in the president and vice president for the United States of America.

**Usage**

```r
data(president)
```
**Format**

A data frame with 67 observations on the following 5 variables.

- potus  President of the United States
- party  Political party of the president
- start  Start year
- end    End year
- vpotus Vice President of the United States

**Source**

Presidents of the United States (table) – infoplease.com (visited: Nov 2nd, 2010)

http://www.infoplease.com/ce6/history/A0840075.html

**Examples**

```r
data(president)
```

---

**Description**

Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. Researchers at University of Texas, Austin collected data on teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors. The scatterplot below shows the relationship between these variables, and also provided is a regression output for predicting teaching evaluation score from beauty score.

**Usage**

```r
data("prof.evaltns.beauty.public")
```

**Format**

A data frame with 463 observations on the following 64 variables.

- tenured  a numeric vector
- profnumber  a numeric vector
- minority  a numeric vector
age a numeric vector
beautyf2upper a numeric vector
beautyflowerdiv a numeric vector
beautyfupperdiv a numeric vector
beautym2upper a numeric vector
beautymlowerdiv a numeric vector
beautymupperdiv a numeric vector
btystdave a numeric vector
btystdf2u a numeric vector
btystdf1 a numeric vector
btystdfu a numeric vector
btystdmd2u a numeric vector
btystdm1 a numeric vector
btystdmu a numeric vector
class1 a numeric vector
class2 a numeric vector
class3 a numeric vector
class4 a numeric vector
class5 a numeric vector
class6 a numeric vector
class7 a numeric vector
class8 a numeric vector
class9 a numeric vector
class10 a numeric vector
class11 a numeric vector
class12 a numeric vector
class13 a numeric vector
class14 a numeric vector
class15 a numeric vector
class16 a numeric vector
class17 a numeric vector
class18 a numeric vector
class19 a numeric vector
class20 a numeric vector
class21 a numeric vector
class22 a numeric vector
class23 a numeric vector
class24 a numeric vector
class25 a numeric vector
class26 a numeric vector
class27 a numeric vector
class28 a numeric vector
class29 a numeric vector
class30 a numeric vector
courseevaluation a numeric vector
didevaluation a numeric vector
female a numeric vector
formal a numeric vector
fullddept a numeric vector
lower a numeric vector
multipleclass a numeric vector
nonenglish a numeric vector
onecredit a numeric vector
percentevaluating a numeric vector
profevaluation a numeric vector
students a numeric vector
tenuretrack a numeric vector
blkandwhite a numeric vector
btystdvariance a numeric vector
btystdavepos a numeric vector
btystdaveneg a numeric vector

Source


Examples

data(prof.evaltns.beauty.public)
## maybe str(prof.evaltns.beauty.public) ; plot(prof.evaltns.beauty.public) ...
Election results for the 2008 U.S. Presidential race

Description

Election results for the 2008 U.S. Presidential race

Usage

data(prRace08)

Format

A data frame with 51 observations on the following 7 variables.

state State name abbreviation
stateFull Full state name
nObama Number of votes for Barack Obama
pObama Proportion of votes for Barack Obama
nMcCain Number of votes for John McCain
pMcCain Proportion of votes for John McCain
e1Votes Number of electoral votes for a state

Details

In Nebraska, 4 electoral votes went to McCain and 1 to Obama. Otherwise the electoral votes were a winner-take-all.

Source

Presidential Election of 2008, Electoral and Popular Vote Summary, collected on April 21, 2011 from


Examples

```r
### Obtain 2010 US House Election Data ===#
data(houseRace10)
hr <- table(houseRace10[, c("abbr", "party1")])
nr <- apply(hr, 1, sum)

### Obtain 2008 President Election Data ===#
data(prRace08)
pr <- prRace08[prRace08$state != "DC", c("state", "pObama")]
hr <- hr[as.character(pr$state),]
(fit <- glm(hr ~ pr$pObama, family=binomial))
```
res.demo.1

Simulated data for regression

Description

Simulated data for regression.

Usage

data("res.demo.1")

Format

A data frame with 100 observations on the following 3 variables.

x a numeric vector

yLin a numeric vector

yFanBack a numeric vector

Examples

data(res.demo.1)

## maybe str(res.demo.1); plot(res.demo.1) ...
Description

Simulated data for regression.

Usage

data("res.demo.2")

Format

A data frame with 300 observations on the following 3 variables.

- x  a numeric vector
- yFan a numeric vector
- yLog a numeric vector

Examples

data(res.demo.2)
## maybe str(res.demo.2); plot(res.demo.2) ...

Description

Cherry Blossom 10 mile run data, 2009

14 variables for all 14,974 10 mile participants in the 2009 Cherry Blossom Run (run10_09) and 9 variables for all 16,924 participants in 2012.

Usage

data(run10)

data(run10_09)
**Format**

The `run10_09` data frame summarizes 14,974 observations on the following 14 variables. The `run10` (2012 data) summarizes 16,924 observations on 9 variables, which are featured with an asterisk.

- **place** * Finishing position. Separate positions are provided for each gender.
- **time** * The total run time. For run10, this is equivalent to netTime.
- **netTime** The run time from the start line to the finish line.
- **pace** * The listed pace for each runner.
- **age** * Age.
- **gender** * Gender.
- **first** First name.
- **last** Last name.
- **city** Hometown city.
- **location** * Hometown city. (run10 data only.)
- **state** * Hometown state. (For run10, this may also list a country.)
- **country** Hometown country.
- **div** Running division (age group).
- **divPlace** * Division place, also broken up by gender.
- **divTot** * Total number of people in the division (again, also split by gender).

**Source**

~~ cherryblossom.org ~~

**References**

~~ OpenIntro Statistics (openintro.org) ~~

**Examples**

data(run10)

### men's times

```r
histPlot(run10$time[run10$gender == 'M'])
```

### times by gender

```r
densityPlot(run10$time, run10$gender, key=c('M','F'))
legend('topright', lty=2:1, col=c('red','black'),
       legend=c('M','F'))
```

### Examine Sample

```r
data(run10Samp)
```
**satGPA**

**Description**

SAT and GPA data for 1000 students at an unnamed college.

**Usage**

```r
data(satGPA)
```

**Format**

A data frame with 1000 observations on the following 6 variables.

- **sex**  Gender of the student.
- **SATV** Verbal SAT percentile.
- **SATM** Math SAT percentile.
- **SATSum** Total of verbal and math SAT percentiles.
- **HSGPA** High school grade point average.
- **FYGPA** First year (college) grade point average.

---

**sat.improve**

*Simulated data for SAT score improvement*

**Description**

Fake data for score improvements from students who took a course from an SAT score improvement company.

**Usage**

```r
data("sat.improve")
```

**Format**

A data frame with 30 observations on the following variable.

- **sat.improve**  a numeric vector

**Examples**

```r
data(sat.improve)
## maybe str(sat.improve) ; plot(sat.improve) ... 
```
Public Opinion with SCOTUS ruling on American Healthcare Act

Description
On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46

Usage
data("scotus_healthcare")

Format
A data frame with 1012 observations on the following variable.

  response Response values reported are agree and other.

Source
Examples

data(scotus_healthcare)
table(scotus_healthcare)

------
senateRace10

Description

Election results for the 2010 U.S. Senate races

Usage

data(senateRace10)

Format

A data frame with 38 observations on the following 23 variables.

id  Unique identifier for the race, which does not overlap with other 2010 races (see *govRace10* and *houseRace10*)
state State name
abbr State name abbreviation
name1 Name of the winning candidate
perc1 Percentage of vote for winning candidate (if more than one candidate)
party1 Party of winning candidate
votes1 Number of votes for winning candidate
name2 Name of candidate with second most votes
perc2 Percentage of vote for candidate who came in second
party2 Party of candidate with second most votes
votes2 Number of votes for candidate who came in second
name3 Name of candidate with third most votes
perc3 Percentage of vote for candidate who came in third
party3 Party of candidate with third most votes
votes3 Number of votes for candidate who came in third
name4 Name of candidate with fourth most votes
perc4 Percentage of vote for candidate who came in fourth
party4 Party of candidate with fourth most votes
votes4 Number of votes for candidate who came in fourth
name5 Name of candidate with fifth most votes
perc5 Percentage of vote for candidate who came in fifth
party5 Party of candidate with fifth most votes
votes5 Number of votes for candidate who came in fifth
**sinusitis**

### Source

Data was collected from MSNBC.com on November 9th, 2010.

### Examples

```r
data(senateRace10)
table(senateRace10$party1)
histPlot(senateRace10$perc1, xlab="Winning candidate vote percentage")
```

---

### sinusitis

**Sinusitis and antibiotic experiment**

### Description

Researchers studying the effect of antibiotic treatment for acute sinusitis to one of two groups: treatment or control.

### Usage

```r
data("sinusitis")
```

### Format

A data frame with 166 observations on the following 2 variables.

- **group**: a factor with levels `control` `treatment`
- **self.reported_improvement**: a factor with levels `no` `yes`

### Source


### Examples

```r
data(sinusitis)
## maybe str(sinusitis) ; plot(sinusitis) ...
### sleep.deprivation

*Survey on sleep deprivation and transportation workers*

**Description**

The National Sleep Foundation conducted a survey on the sleep habits of randomly sampled transportation workers and a control sample of non-transportation workers.

**Usage**

```r
data("sleep.deprivation")
```

**Format**

A data frame with 1087 observations on the following 2 variables.

- **sleep**: a factor with levels `<6` `6-8`
- **profession**: a factor with levels `bus` `taxi` `limo` `drivers` `control pilots` `train operators` `truck drivers`

**Source**


**Examples**

```r
data(sleep.deprivation)
## maybe str(sleep.deprivation); plot(sleep.deprivation) ...  
```

### smoking

*UK Smoking Data*

**Description**

Survey data on smoking habits from the UK. The data set can be used for analyzing the demographic characteristics of smokers and types of tobacco consumed.

**Usage**

```r
data(smoking)
```
**Format**

A data frame with 1691 observations on the following 12 variables.

- **gender** Gender with levels Female and Male.
- **age** Age.
- **maritalStatus** Marital status with levels Divorced, Married, Separated, Single and Widowed.
- **highestQualification** Highest education level with levels A Levels, Degree, GCSE/CSE, GCSE/O Level, Higher/Sub Degree, No Qualification, ONC/BTEC and Other/Sub Degree
- **nationality** Nationality with levels British, English, Irish, Scottish, Welsh, Other, Refused and Unknown.
- **ethnicity** Ethnicity with levels Asian, Black, Chinese, Mixed, White and Refused Unknown.
- **grossIncome** Gross income with levels Under 2,600, 2,600 to 5,200, 5,200 to 10,400, 10,400 to 15,600, 15,600 to 20,800, 20,800 to 28,600, 28,600 to 36,400, Above 36,400, Refused and Unknown.
- **region** Region with levels London, Midlands & East Anglia, Scotland, South East, South West, The North and Wales
- **smoke** Smoking status with levels No and Yes
- **amtWeekends** Number of cigarettes smoked per day on weekends.
- **amtWeekdays** Number of cigarettes smoked per day on weekdays.
- **type** Type of cigarettes smoked with levels Packets, Hand-Rolled, Both/Mainly Packets and Both/Mainly Hand-Rolled

**Source**


**Examples**

```r
data(smoking)  # Load the data frame
str(smoking)    # Display the structure of the data frame
histPlot(smoking$amtWeekends)  # Create a histogram of weekends
histPlot(smoking$amtWeekdays)  # Create a histogram of weekdays
table(smoking$smoke, smoking$gender)  # Create a table
mosaicplot(~ smoke + maritalStatus, data = smoking)  # Create a mosaic plot
barplot(sort(table(smoking$maritalStatus), decreasing = TRUE))  # Sort and plot marital status
```
Description

Daily stock returns from the S&P500 for 1990-2011 can be used to assess whether stock activity each day is independent of the stock's behavior on previous days. We label each day as Up or Down (D) depending on whether the market was up or down that day. For example, consider the following changes in price, their new labels of up and down, and then the number of days that must be observed before each Up day.

Usage

data("sp500.seq")

Format

A data frame with 2948 observations on the following variable.

race  a factor with levels 1 2 3 4 5 6 7+

Source

http://www.google.com/finance

Examples

data(sp500.seq)
## maybe str(sp500.seq) ; plot(sp500.seq) ...

Description

1,302 UCLA students were asked to fill out a survey where they were asked about their height, fastest speed they have ever driven, and gender.

Usage

data("speed.gender.height")
stanford

Format

A data frame with 1325 observations on the following 4 variables.

- `x`: a numeric vector
- `speed`: a numeric vector
- `gender`: a factor with levels `female` `male`
- `height`: a numeric vector

Examples

data(speed, gender, height)
## maybe str(speed, gender, height) ; plot(speed, gender, height) ...

Stanford heart transplant experiment

Description

The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not.

Usage

data("stanford")

Format

A data frame with 103 observations on the following 2 variables.

- `outcome`: a factor with levels `alive` `dead`
- `group`: a factor with levels `control` `treatment`

Source


Examples

data(stanford)
## maybe str(stanford) ; plot(stanford) ...
Description

Nutrition facts for several Starbucks food items

Usage

data("starbucks")

Format

A data frame with 77 observations on the following 7 variables.

item Food item.
calories Calories.
fat a numeric vector
carb a numeric vector
fiber a numeric vector
protein a numeric vector
type a factor with levels bakery bistro box hot breakfast parfait petite salad sandwich

Source

http://www.starbucks.com/menu/nutrition (March 10, 2011)

Examples

data(starbucks)
## maybe str(starbucks) ; plot(starbucks) ...

Description

Scores range from 57 to 94.

Usage

data("stats.scores")
stem.cell

Format

A data frame with 20 observations on the following variable.

scores a numeric vector

Examples

data(stats.scores)
## maybe str(stats.scores) ; plot(stats.scores) ...

---

stem.cell

Embryonic stem cells to treat heart attack (in sheep)

Description

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? Each sheep in the study was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured in the study. A positive value corresponds to increased pumping capacity, which generally suggests a stronger recovery.

Usage

data("stem.cell")

Format

A data frame with 18 observations on the following 3 variables.

trmt a factor with levels ctrl esc
before a numeric vector
after a numeric vector

Source

http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(05)67380-1/abstract

Examples

data(stem.cell)
## maybe str(stem.cell) ; plot(stem.cell) ...
Description

Stem cells were used as the treatment for sheep that had suffered a heart attack.

Usage

```r
data("stem.cells")
```

Format

A data frame with 18 observations on the following 3 variables.

- `trmt` Factor indicating the treatment group.
- `before` Heart pumping function, pre-experiment.
- `after` Heart pumping function, post-experiment.

Details

The animals used were sheep, and the embryonic stem cells were from mice. The control had no treatments. The ESC group had stem cells transplanted to their hearts.

Source


Examples

```r
data(stem.cells)
change <- stem.cells$after - stem.cells$before
summary(lm(change ~ trmt, stem.cells))
```
### Description

An experiment that studies effectiveness of stents in treating patients at risk of stroke with some unexpected results. These data represent the results 30 days after stroke.

### Usage

```r
data("stent30")
```

### Format

A data frame with 451 observations on the following 2 variables.

- **group**: a factor with levels control treatment
- **outcome**: a factor with levels no event stroke

### Source


### Examples

```r
data(stent30)
## maybe str(stent30) ; plot(stent30) ...
```
Source


Examples

data(stent365)
## maybe str(stent365); plot(stent365) ...

student.housing Community college housing (simulated data, 2015)

Description

These are simulated data and intended to represent housing prices of students at a college.

Usage

data("student.housing")

Format

A data frame with 175 observations on the following variable.

price Monthly housing price, simulated.

Examples

data(student.housing)
set.seed(5)
generate.student.housing <- data.frame(
  price = round(rnorm(175, 515, 65) + exp(rnorm(175, 4.2, 1))))
hist(student.housing$price, 20)
t.test(student.housing$price)
mean(student.housing$price)
 sd(student.housing$price)
 identical(student.housing, generate.student.housing)
student.sleep

Description
A simulated data set for how much 110 college students each slept in a single night.

Usage
data("student.sleep")

Format
A data frame with 110 observations on the following variable.

hours  Number of hours slept by this student (simulated).

Source
Simulated data.

Examples
data(student.sleep)

set.seed(2)
x <- exp(c(rnorm(100, log(7.5), 0.15),
          rnorm(10, log(10), 0.196)))
x <- round(x - mean(x) + 7.42, 2)
identical(x, student.sleep$hours)

sulphinpyrazone

Description
Experiment data for studying the efficacy of treating patients who have had a heart attack with Sulphinpyrazone.

Usage
data("sulphinpyrazone")
Format

A data frame with 1475 observations on the following 2 variables.

- group: a factor with levels control treatment
- outcome: a factor with levels died lived

Source


Examples

data(sulphinpyrazone)
## maybe str(sulphinpyrazone) ; plot(sulphinpyrazone) ...

---

table

<table>
<thead>
<tr>
<th>supreme.court</th>
<th>Supreme Court approval rating</th>
</tr>
</thead>
</table>

Description

Summary of a random survey of 976 people.

Usage

data("supreme.court")

Format

A data frame with 976 observations on the following variable.

- answer: a factor with levels approve not

Source


Examples

data(supreme.court)
## maybe str(supreme.court) ; plot(supreme.court) ...
Description

A random sample was taken of nearly 10% of UCLA courses. The most expensive textbook for each course was identified, and its new price at the UCLA Bookstore and on Amazon.com were recorded.

Usage
data(textbooks)

Format

A data frame with 73 observations on the following 7 variables.

depabbr  Course department (abbreviated).
course  Course number.
isbn  Book ISBN.
uclanew  New price at the UCLA Bookstore.
amaznew  New price on Amazon.com.
more  Whether additional books were required for the course (Y means "yes, additional books were required").
diff  The UCLA Bookstore price minus the Amazon.com price for each book.

Details

The sample represents only courses where textbooks were listed online through UCLA Bookstore’s website. The most expensive textbook was selected based on the UCLA Bookstore price, which may insert bias into the data; for this reason, it may be beneficial to analyze only the data where more is "N".

Source

This data was collected by David M Diez on April 24th.

References

See Section 5.1 of the Open Intro Statistics textbook: http://www.openintro.org/
Examples

data(textbooks)
### an improper analysis ###
boxPlot(textbooks$uclaNew, xlim=c(0.5,2.5))
boxPlot(textbooks$amazonNew, add=2)
axis(1, at=1:2, labels=c('UCLA Bookstore', 'Amazon'))
t.test(textbooks$uclaNew, textbooks$amazonNew)

### a reasonable analysis ###
# the differences are moderately skewed
# the sample size is sufficiently large to justify t test
histPlot(textbooks$diff)
t.test(textbooks$diff)

tgSpending

Thanksgiving spending, simulated based on Gallup poll.

Description

This entry gives simulated spending data for Americans during Thanksgiving in 2009 based on findings of a Gallup poll.

Usage

data(tgSpending)

Format

A data frame with 436 observations on the following 1 variable.

spending  Amount of spending, in US dollars.

Examples

data(tgSpending)
histPlot(tgSpending$spending)

tips

Tip data

Description

A simulated data set of tips over a few weeks on a couple days per week. Each tip is associated with a single group, which may include several bills and tables (i.e. groups paid in one lump sum in simulations).
Usage

data(tips)

Format

A data frame with 95 observations on the following 5 variables.

- **week**: Week number.
- **day**: Day, either Friday or Tuesday.
- **nPeop**: Number of people associated with the group.
- **bill**: Total bill for the group.
- **tip**: Total tip from the group.

Details

This data set was built using simulations of tables, then bills, then tips based on the bills. Large groups were assumed to only pay the gratuity, which is evident in the data. Tips were set to be plausible round values; they were often (but not always) rounded to dollars, quarters, etc.

Source

Simulated data set.

References

http://www.openintro.org/

Examples

data(tips)
par(mfrow=c(2,2))
boxPlot(tips$tip, tips$day)
densityPlot(tips$tip, tips$week, key=1:3)
legend('topright', lty=1:3, col=c('black', 'red', 'blue'), legend=1:3)
dotPlot(tips$tip)
densityPlot(tips$tip, tips$day)
legend('topright', col=c('black','red'), lty=1:2,
legend=c('Tuesday', 'Friday'))
### toohey

**Simulated polling data set**

**Description**

Simulated data for a fake political candidate.

**Usage**

```r
data("toohey")
```

**Format**

A data frame with 500 observations on the following variable.

- `voteFor` a factor with levels `no` `yes`

**Examples**

```r
data(toohey)
## maybe str(toohey) ; plot(toohey) ...
```

### tourism

**Turkey tourism**

**Description**

Summary of tourism in Turkey.

**Usage**

```r
data("tourism")
```

**Format**

A data frame with 47 observations on the following 3 variables.

- `year` a numeric vector
- `visitor_count_tho` a numeric vector
- `tourist_spending` a numeric vector

**Source**

Association of Turkish Travel Agencies, Foreign Visitors Figure & Tourist Spendings By Years. [http://www.tursab.org.tr/en/statistics/foreign-visitors-figure-tourist-spendings-by-years_1083.html](http://www.tursab.org.tr/en/statistics/foreign-visitors-figure-tourist-spendings-by-years_1083.html)
toy.anova

Simulated data set for ANOVA

Description
Simulated data set for getting a better understanding of intuition that ANOVA is based off of.

Usage
data("toy.anova")

Format
A data frame with 70 observations on the following 3 variables.
groups a factor with levels I II III
x1 a numeric vector
x2 a numeric vector

Examples
data(toy.anova)
## maybe str(toy.anova) ; plot(toy.anova) ...

transplant

Transplant consultant success rate (fake data)

Description
Summarizing whether there was or was not a complication for 62 patients who used a particular medical consultant.

Usage
data("transplant")

Format
A data frame with 62 observations on the following variable.
outcome a factor with levels complications okay
Examples

data(transplant)
    ## maybe str(transplant); plot(transplant) ...

treeDiag

Description

Construct beautiful tree diagrams

Usage

treeDiag(main, p1, p2, out1 = c("Yes", "No"), out2 = c("Yes", "No"),
    textwd = 0.15, solwd = 0.2, SBS = c(TRUE, TRUE), showSol = TRUE,
    solSub = NULL, digits = 4, textadj = 0.015, cex.main = 1.3,
    col.main = "#999999", showWork = FALSE)

Arguments

main Character vector with two variable names, descriptions, or questions
p1 Vector of probabilities for the primary branches
p2 List for the secondary branches, where each list item should be a numerical
    vector of probabilities corresponding to the primary branches of p1
out1 Character vector of the outcomes corresponding to the primary branches
out2 Character vector of the outcomes corresponding to the secondary branches
textwd The width provided for text with a default of 0.15
solwd The width provided for the solution with a default of 0.2
SBS A boolean vector indicating whether to place text and probability side-by-side
    for the primary and secondary branches
showSol Boolean indicating whether to show the solution in the tree diagram
solSub An optional list of vectors corresponding to p2 to list alternative text or solutions
digits The number of digits to show in the solution
textadj Vertical adjustment of text
cex.main Size of main in the plot
col.main Color of main in the plot
showWork Whether work should be shown for the solutions

Value

No value is given.
trees

Author(s)

David M Diez, Christopher D Barr

References

OpenIntro Statistics, Chapter 2.

See Also

histPlot

Examples

# Examples
# generic with random probabilities

treeDiag(c('Flight on time?','Luggage on time?'),
         c(.8,.2), list(c(.97,.03), c(.15,.85)))

treeDiag(c('Breakfast?','Go to class'), c(.4,.6),
         list(c(.4,.36,.34),c(.6,.3,.1)), c('Yes','No'),
         c('Statistics','English','Sociology'), showWork=TRUE)

treeDiag(c('Breakfast?','Go to class'), c(.4,.11,.49),
         list(c(.4,.36,.24),c(.6,.3,.1),c(.1,.4,.5)),
         c('one','two','three'), c('Statistics','English','Sociology'))

treeDiag(c('Dow Jones rise?','NASDAQ rise?'),
         c(0.53, 0.47), list(c(0.75, 0.25), c(0.72, 0.28)),
         solSub=list(c("(a)"), "(b)", c("(c)", "(d)")), solwd=0.08)

trees

Summary of 32 trees

Description

Variables include girth, height, and volume.

Usage

data("trees")

Format

A data frame with 31 observations on the following 3 variables.

Girth Diameter, in inches
Height Height, in feet.
Volume Volume, in cubic feet.
**Source**


**Examples**

```r
data(trees)
## maybe str(trees); plot(trees) ...
```

<table>
<thead>
<tr>
<th>Year</th>
<th>Unemployment rate, in percent</th>
<th>usdata</th>
</tr>
</thead>
</table>

**Description**

A compilation of two data sets that provides an estimate of unemployment from 1890 to 2010.

**Usage**

data(unempl)

**Format**

A data frame with 121 observations on the following 3 variables.

- **year** Year
- **unemp** Unemployment rate, in percent
- **usData** 1 if from the Bureau of Labor Statistics, 0 otherwise

**Source**

The data are from Wikipedia at the following URL accessed on November 1st, 2010:

Below is a direct quotation from Wikipedia describing the sources of the data:

Examples

data(unempl)

### Time Series Plot of Data

```r
COL <- c("#DDEEBB", "#EEEDBB", "#BBDEEE", "#FFD5DD", "#FFC5CC")
plot(unempl$year, unempl$unemp, type="n")
rect(0, -50, 3000, 100, col="#E2E2E2")
rect(1914.5, -1000, 1918.9, 1000, col=COL[1], border="#E2E2E2")
rect(1929, -1000, 1939, 1000, col=COL[2], border="#E2E2E2")
rect(1939.7, -1000, 1945.6, 1000, col=COL[3], border="#E2E2E2")
rect(1955.8, -1000, 1965.3, 1000, col=COL[4], border="#E2E2E2")
rect(1965.3, -1000, 1975.4, 1000, col=COL[5], border="#E2E2E2")
abline(h=seq(0,50,5), col="#F8F8F8", lwd=2)
abline(v=seq(1900, 2000, 20), col="#FFFFFF", lwd=1.3)
lines(unempl$year, unempl$unemp)
points(unempl$year, unempl$unemp, pch=20)
legend("topright", fill=COL,
c("World War I", "Great Depression", "World War II",
"Vietnam War Start", "Vietnam War Escalated"),
bg="FFFFFF", border="FFFFFF")
```

---

**unemploy.pres**

**President’s party performance and unemployment rate**

---

**Description**

Covers midterm elections.

**Usage**

data("unemploy.pres")

**Format**

A data frame with 29 observations on the following 5 variables.

<table>
<thead>
<tr>
<th>year</th>
<th>a numeric vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>potus</td>
<td>The president in office.</td>
</tr>
<tr>
<td>party</td>
<td>President’s party.</td>
</tr>
<tr>
<td>unemp</td>
<td>Unemployment rate.</td>
</tr>
<tr>
<td>change</td>
<td>Change in House seats for the president’s party.</td>
</tr>
</tbody>
</table>

**Source**

Wikipedia.

**Examples**

data(unemploy.pres)

## maybe str(unemploy.pres); plot(unemploy.pres) ...
urban.owner  Summary of many state-level variables

Description
Census data for the 50 states plus DC and Puerto Rico.

Usage
data("urban.owner")

Format
A data frame with 52 observations on the following 28 variables.

state  State
total_housing_units_2000  Total housing units available in 2000.
total_housing_units_2010  Total housing units available in 2010.
pct_vacant  a numeric vector
occupied  Occupied.
pct_owner_occupied  a numeric vector
pop_st  a numeric vector
area_st  a numeric vector
pop_uurban  a numeric vector
poppct_uurban  a numeric vector
area_uurban  a numeric vector
areapct_uurban  a numeric vector
popden_uurban  a numeric vector
pop_u  a numeric vector
poppct_u  a numeric vector
area_u  a numeric vector
areapct_u  a numeric vector
popden_u  a numeric vector
pop_uc  a numeric vector
poppct_uc  a numeric vector
area_uc  a numeric vector
areapct_uc  a numeric vector
popden_uc  a numeric vector
pop_rural  a numeric vector
poppct_rural  a numeric vector
area_rural  a numeric vector
areapct_rural  a numeric vector
popden_rural  a numeric vector
Source

US Census.

Examples

data(urban.owner)
## maybe str(urban.owner) ; plot(urban.owner) ...

urban.rural.pop    State summary info

Description

Census info for the 50 US states plus DC.

Usage

data("urban.rural.pop")

Format

A data frame with 51 observations on the following 5 variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>US state.</td>
</tr>
<tr>
<td>UrbanIn</td>
<td>a numeric vector</td>
</tr>
<tr>
<td>UrbanOut</td>
<td>a numeric vector</td>
</tr>
<tr>
<td>RuralFarm</td>
<td>a numeric vector</td>
</tr>
<tr>
<td>RuralNonfarm</td>
<td>a numeric vector</td>
</tr>
</tbody>
</table>

Source

US census.

Examples

data(urban.rural.pop)
## maybe str(urban.rural.pop) ; plot(urban.rural.pop) ...
WriteTable

Write a Data Frame or Matrix to a Text File

Description

The arguments are identical to those of write.table, except the defaults differ: the separator is a tab, does not include quotation marks, and does not use row names by default.

Usage

WriteTable(x, file = "", append = FALSE, quote = FALSE, sep = "\t",
          eol = "\n", na = "NA", dec = ".", row.names = FALSE,
          col.names = TRUE, qmethod = c("escape", "double"),
          fileEncoding = "")

Arguments

x The object to be written, preferably a matrix or data frame. If not, it is attempted to coerce x to a data frame.

file T either a character string naming a file or a connection open for writing. "" indicates output to the console.

append Logical. Only relevant if file is a character string. If TRUE, the output is appended to the file. If FALSE, any existing file of the name is destroyed.

quote A logical value (TRUE or FALSE) or a numeric vector. If TRUE, any character or factor columns will be surrounded by double quotes. If a numeric vector, its elements are taken as the indices of columns to quote. In both cases, row and column names are quoted if they are written. If FALSE, nothing is quoted.

sep Separator for values.

eol The character(s) to print at the end of each line (row).

na The string to use for missing values in the data.

dec The string to use for decimal points in numeric or complex columns: must be a single character.

row.names Either a logical value indicating whether the row names of x are to be written along with x, or a character vector of row names to be written.

col.names Either a logical value indicating whether the column names of x are to be written along with x, or a character vector of column names to be written. See the section on CSV files for the meaning of col.names = NA.

qmethod A character string specifying how to deal with embedded double quote characters when quoting strings.

fileEncoding Character string: if non-empty declares the encoding to be used on a file (not a connection) so the character data can be re-encoded as they are written.

See Also

write.table
Examples

## Not run:
## To write a CSV file for input to Excel one might use
x <- data.frame(a = I("a \\ " quote"), b = pi)
write.table(x, file = "foo.csv", sep = ",", col.names = NA,
            qmethod = "double")
## and to read this file back into R one needs
read.table("foo.csv", header = TRUE, sep = ",", row.names = 1)
## NB: you do need to specify a separator if qmethod = "double".

## End (Not run)

---

Data Frame

Exxon Mobile stock data

Description


Usage

data("xom")

Format

A data frame with 98 observations on the following 7 variables.

Date  Date.
Open  a numeric vector
High  a numeric vector
Low   a numeric vector
Close  a numeric vector
Volume a numeric vector
Adj.Close a numeric vector

Source

http://finance.yahoo.com

Examples

data(xom)
## maybe str(xom); plot(xom) ...
Description

An experiment conducted by the MythBusters, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously influenced into yawning if another person near them yawns. 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a group where there wasn’t a person yawning near them (control).

Usage

data("yawn")

Format

A data frame with 50 observations on the following 2 variables.

result a factor with levels not yawn yawn
group a factor with levels ctrl trmt

Source

MythBusters, Season 3, Episode 28.

Examples

data(yawn)
## maybe str(yawn) ; plot(yawn) ...

Description

Select variables from YRBSS.

Usage

data("yrbss")
Format

A data frame with 13583 observations on the following 13 variables.

- age  Age, in years.
- gender  Gender.
- grade  School grade.
- hispanic  Hispanic or not.
- race  Race / ethnicity.
- height  Height, in meters (3.28 feet per meter).
- weight  Weight, in kilograms (2.2 pounds per kilogram).
- helmet_QRm  How often did you wear a helmet when biking in the last 12 months?
- text_while_driving_SPd  How many days did you text while driving in the last 30 days?
- physically_active_Wd  How many days were you physically active for 60+ minutes in the last 7 days?
- hours_tv_per_school_day  How many hours of TV do you typically watch on a school night?
- strength_training_Wd  How many days did you do strength training (e.g. lift weights) in the last 7 days?
- school_night_hours_sleep  How many hours of sleep do you typically get on a school night?

Source

CDC’s Youth Risk Behavior Surveillance System (YRBSS)

Examples

data(yrbss)
table(yrbss$physically_active_7d)

data("yrbss.samp")

Description

A sample of the yrbss data set.

Usage

data("yrbss.samp")
Format

A data frame with 100 observations on the following 13 variables.

age  Age, in years.
gender  Gender.
grade  School grade.
hispanic  Hispanic or not.
race  Race / ethnicity.
height  Height, in meters (3.28 feet per meter).
weight  Weight, in kilograms (2.2 pounds per kilogram).
helmet_12m  How often did you wear a helmet when biking in the last 12 months?
text_while_driving_30d  How many days did you text while driving in the last 30 days?
physically_active_7d  How many days were you physically active for 60+ minutes in the last 7 days?
hours_tv_per_school_day  How many hours of TV do you typically watch on a school night?
strength_training_7d  How many days did you do strength training (e.g. lift weights) in the last 7 days?
school_night_hours_sleep  How many hours of sleep do you typically get on a school night?

Source

CDC’s Youth Risk Behavior Surveillance System (YRBSS)

Examples

data(yrbss.samp)
table(yrbss.samp$physically_active_7d)
Index

*Topic 2008
prRace08, 120

*Topic Abbreviation
abbr2state, 7

*Topic Algebra
ArrowLines, 11
CCP, 31
dlsegments, 49
lsegments, 92

*Topic Bayes Theorem
treeDiag, 144

*Topic Cartesian Coordinate Plane
CCP, 31

*Topic Conditional probability
treeDiag, 144

*Topic Data tube
makeTube, 93

*Topic Graphics
myPDF, 105

*Topic Kernel smoothing
makeTube, 93

*Topic LaTeX
contTable, 37

*Topic Least squares
makeTube, 93

*Topic Line Segment
ArrowLines, 11
dlsegments, 49
lsegments, 92

*Topic PDF
myPDF, 105

*Topic Plane
CCP, 31

*Topic Plotting
myPDF, 105

*Topic Regression
makeTube, 93

*Topic Save
myPDF, 105

*Topic State
abbr2state, 7

*Topic Tree diagram
treeDiag, 144

*Topic United States
prRace08, 120

*Topic axis
buildAxis, 24

*Topic categorical data
heartTr, 76

*Topic contingency tables
heartTr, 76

*Topic control axis
buildAxis, 24

*Topic customize axis
buildAxis, 24

*Topic datasets, ball bearings, inference on means
ballBearing, 17

*Topic datasets, college credits, inference on means
credits, 44

*Topic datasets, correlation, regression
gradesTV, 73

*Topic datasets, histogram, distribution
infMortRate, 86
tgSpending, 140

*Topic datasets, iPod, inference on means
ipod, 87

*Topic datasets, regression
gifted, 67

*Topic datasets, smoking
smoking, 128

*Topic datasets
acs12, 8
ageAtMar, 10
<table>
<thead>
<tr>
<th>ami_occurrences</th>
<th>gpa</th>
<th>gpa_study_hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>71</td>
<td>72</td>
</tr>
<tr>
<td>association.1.3</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>association.4.6</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>association.7.12</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>assortive.mating</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>avandia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>babies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>babies.crawl</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>bac</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bdims</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>births</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>books</td>
<td></td>
<td></td>
</tr>
<tr>
<td>burger</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>cancer.in.dogs</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>cards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cars</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>cats</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>cchHousing</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>census</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>cherry</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>chick.wts</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>china</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>classData</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>COL</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>corr.match</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>county</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>county.w.sm.ban</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>countyComplete</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>cpr</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>diamonds</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>dream</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>drug.use</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>ebola.survey</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>elmhurst</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>email</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>email50, 58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>epa2012, 60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>exams</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>fcid</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>fheights</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>friday</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>full.body.scan</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>gear.company</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>gender.discrimination</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>global.warming.pew</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>goog</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>gov.poll</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>govRace10</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>health_coverage</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>healthcare.law.survey</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>healthcare.survey</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>heartTr</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>helium</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>house</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>houseRace08</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>housing</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>hsb2</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>husbands.wives</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>immigration</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>iris</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>jury</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>leg_mari</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>male.heights</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>mammals</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>mammogram</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>marathon</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>marioKart</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>migraine</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>MLB</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>mlbBat0.10</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>nba.heights</td>
<td>106</td>
<td></td>
</tr>
<tr>
<td>ncbirths</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>nuclear.survey</td>
<td>109</td>
<td></td>
</tr>
<tr>
<td>offshore.drilling</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>orings</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>oscars</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td>outliers</td>
<td>112</td>
<td></td>
</tr>
<tr>
<td>pm25.2011.durham</td>
<td>114</td>
<td></td>
</tr>
<tr>
<td>poker</td>
<td>115</td>
<td></td>
</tr>
<tr>
<td>possom</td>
<td>115</td>
<td></td>
</tr>
<tr>
<td>president</td>
<td>116</td>
<td></td>
</tr>
<tr>
<td>prof.evaltns.beauty.public</td>
<td>117</td>
<td></td>
</tr>
<tr>
<td>prRace08</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>res.demo.1</td>
<td>121</td>
<td></td>
</tr>
<tr>
<td>res.demo.2</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>run10</td>
<td>122</td>
<td></td>
</tr>
<tr>
<td>sat.improve</td>
<td>124</td>
<td></td>
</tr>
<tr>
<td>satGPA</td>
<td>124</td>
<td></td>
</tr>
<tr>
<td>scotus_healthcare</td>
<td>125</td>
<td></td>
</tr>
<tr>
<td>senateRace08</td>
<td>126</td>
<td></td>
</tr>
<tr>
<td>sinusitis</td>
<td>127</td>
<td></td>
</tr>
<tr>
<td>sleep.deprivation</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>sp500.seq</td>
<td>130</td>
<td></td>
</tr>
</tbody>
</table>
INDEX

157

speed, gender, height, 130
stanford, 131
starbucks, 132
stats, scores, 132
stem, cell, 133
stem, cells, 134
stent, 135
stent, 365, 135
student, housing, 136
student, sleep, 137
sulphinpyrazone, 137
supreme, court, 138
textbooks, 139
tips, 140
toohey, 142
tourism, 142
toy, anova, 143
transplant, 143
trees, 145
unempl, 146
unemploy, pres, 147
urban, owner, 148
urban, rural, pop, 149
xom, 151
yawn, 152
yrbss, 152
yrbss, samp, 153

*Topic dot plot
dotPlotStack, 52

*Topic election
prRace08, 120

*Topic for loop
loop, 92

*Topic heart transplant
heartTr, 76

*Topic index
loop, 92

*Topic linear model
lmPlot, 90

*Topic looping
loop, 92

*Topic message
loop, 92

*Topic myPDF
myPDF, 105

*Topic package
openintro-package, 5

*Topic president

prRace08, 120

*Topic randomization tests
heartTr, 76

*Topic residuals
lmPlot, 90

*Topic stacked dot plot
dotPlotStack, 52

*Topic table
contTable, 37

abbr2state, 7
acs12, 8
ageAtMar, 10
ami, occurrences, 10
ArrowLines, 11, 32, 50, 93
arrows, 11, 31
association, 1.3, 12
association, 4.6, 13
association, 7.12, 13
assortive.mating, 14
avandia, 14

babies, 15
babies, crawl, 16
bac, 17
ballBearing, 17
bdims, 18
births, 20
books, 21
boxPlot, 5, 21, 25, 46, 47, 51, 55, 79
Braces, 23
buildAxis, 5, 24, 109
burger, 26

cancer, in, dogs, 27
cards, 28
cars, 5, 28, 38
cat, 37
cats, 30
cchousing, 30
CCP, 12, 31, 50, 93
census, 32
cherry, 33
chick, wts, 34
china, 34
ChiSquareTail, 35
classData, 36
COL, 36
contTable, 37
corr.match, 38
county, 8, 39, 57, 59
county.w.sm.ban, 40
countyComplete, 8, 39, 42
cpr, 44
createEdaOptions (edaPlot), 55
credits, 44
CT2DF, 45
densityPlot, 5, 22, 25, 46, 51, 55, 79
diamonds, 48
d1segments, 12, 24, 32, 49, 93
dotPlot, 5, 22, 25, 47, 50, 52, 55, 62, 79
dotPlotStack, 52
dream, 53
drug.use, 53
ebola.survey, 54
edaPlot, 5, 55, 106
elmhurst, 55
email, 38, 39, 56, 58, 59
email150, 39, 57, 58
e-mail_test (email), 56
email_test (email), 56
epa2012, 60
exams, 61
fa1deColor, 62
fcid, 63
fheights, 64
fitNormal (edaPlot), 55
friday, 64
full.body.scan, 65
gear.company, 66
gender.discrimination, 66
Gestation, 76
gifted, 67
global.warming.pew, 68
goog, 69
gov.poll, 69
govRace10, 70, 81, 126
gpa, 71
gpa_study_hours, 72
gradesTV, 73
gsearch, 74
guessMethod (edaPlot), 55
health_coverage, 76
healthcare.law.survey, 74
healthcare.survey, 75
heartTr, 76
helium, 77
histPlot, 5, 22, 25, 47, 51, 52, 55, 78, 145
house, 80
houseRace10, 70, 81, 126
housing, 83
hsb2, 83
husbands.wives, 84
immigration, 85
infMortRate, 86
ipod, 87
iris, 87
jury, 88
leg_mari, 88
lines, 11, 24
linResPlot, 89
lmPlot, 90, 94
loop, 92
l1segments, 12, 32, 50, 92
makePlotIcon (edaPlot), 55
makeTube, 90, 91, 93, 113
male.heights, 95
mammals, 96
mammogram, 97
marathon, 98
marioKart, 5, 38, 98
migraine, 100
MLB, 101
mlbBat10, 102
MosaicPlot, 45, 104
myPDF, 5, 92, 105
myPNG, 5
myPNG (myPDF), 105
nba.heights, 106
ncbirths, 107
normTail, 5, 35, 108
nuclear.survey, 109
offshore.drilling, 110
openintro (openintro-package), 5
openintro-package, 5
orings, 110
oscars, 111
outliers, 112
plotNothing (edaPlot), 55
PlotWLine, 113
pm25.2011.durham, 114
points, 52
poker, 115
possum, 5, 38, 115
president, 116
prof.evaiTNs.beauty.public, 117
prRace08, 120

res.demo.1, 121
res.demo.2, 122
run10, 5, 122
run10_09 (run10), 122
run10Samp (run10), 122

sat.improve, 124
satGPA, 5, 124
scotus_healthcare, 125
senateRace10, 70, 81, 126
setTxtProgressBar, 92
sinusitis, 127
sleep.deprivation, 128
smoking, 128
sp500.seq, 130
speed.gender.height, 130
stanford, 131
starbucks, 132
state2abbr (abbr2state), 7
stats.scores, 132
stem.cell, 133
stem.cells, 134
stent30, 135
stent365, 135
student.housing, 136
student.sleep, 137
sulphinpyrazone, 137
supreme.court, 138

text, 31, 32
textbooks, 5, 139
tg$Spending, 140
tips, 140
toohey, 142
tourism, 142
toy.anova, 143
transplant, 143
treeDiag, 144
trees, 145
txtProgressBar, 92

unempl, 146
unemploy.pres, 147
urban.owner, 148
urban.rural.pop, 149

write.table, 150
WriteTable, 150
xom, 151

yawn, 152
yrbss, 152, 153
yrbss.samp, 153