Занятие 2
Data frame
Построение графиков

20 февраля 2013

Overview

- Что такое data frame
- Создание своего data frame и использование готовых
- Subsetting и функция order
- Графики
- Работа с NA

Что такое data frame

- Структура данных: таблица из нескольких векторов (по столбцам), в разных столбцах могут быть данные разных типов

- Как создать свой data frame?

```
> n <- c(2, 3, 5)
> s <- c("aa", "bb", "cc")
> b <- c(TRUE, FALSE, TRUE)
> df <- data.frame(n, s, b)
```

Или короче:

```
> df <- data.frame(n=c(2, 3, 5),  s=c("aa", "bb", "cc"),
b= c(TRUE, FALSE, TRUE))
```

Основные команды

```
> df <- data.frame(n=c(2, 3, 5),  s=c("aa", "bb", "cc"), b= c(TRUE, FALSE, TRUE))
> df
  n  s     b
1 2 aa  TRUE
2 3 bb FALSE
3 5 cc  TRUE
> df$n
[1] 2 3 5
> colnames(df)
[1] "n" "s" "b"
> rownames(df)
[1] "1" "2" "3"
> dim(df)
[1] 3 3
```

Обращение к столбцу по имени, можно использовать tab!

Важно, что это имена строк, а не числа!

Использование data()

```
> mtcars[1:12,]
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3

Командой data() можно посмотреть, какие data sets загружены для использования

Выбор строк, столбцов, ячеек

```
> mtcars[12,2]
```

строка 12, столбец 2

```
[1] 8
```

```
> mtcars[8,]
```

```
           mpg cyl disp hp drat wt qsec vs am gear carb
Merc 240D 24.4 4 146.7 62 3.69 3.19 20 1 0 4 2
```

```
>  mtcars[1:3,]
```

строки 1 - 3, все столбцы

```
               mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

Выбор строк, столбцов, ячеек

```
> mtcars[,2]
```
все строки, столбец 2

```
[1] 6 6 4 6 8 6 8 4 4 6 6 8 8 8 8 8 4 4 4 4 8 8 8 8 4 4 4 4 8 6 8 4
```

```
> mtcars[c(1,13),]
```
строки 1 и 13, все столбцы

```
                mpg cyl  disp  hp drat   wt  qsec vs am gear carb
Mazda RX4      21.0   6 160.0 110 3.90 2.62 16.46  0  1    4    4
Merc 450SL     17.3   8 275.8 180 3.07 3.73 17.60  0  0    3    3
```

```
> mtcars[c(1,3,7,13),1]
```
строки 1, 3, 7 и 13, столбец 1

```
[1] 21.0 22.8 14.3 17.3
```

Логические условия и order

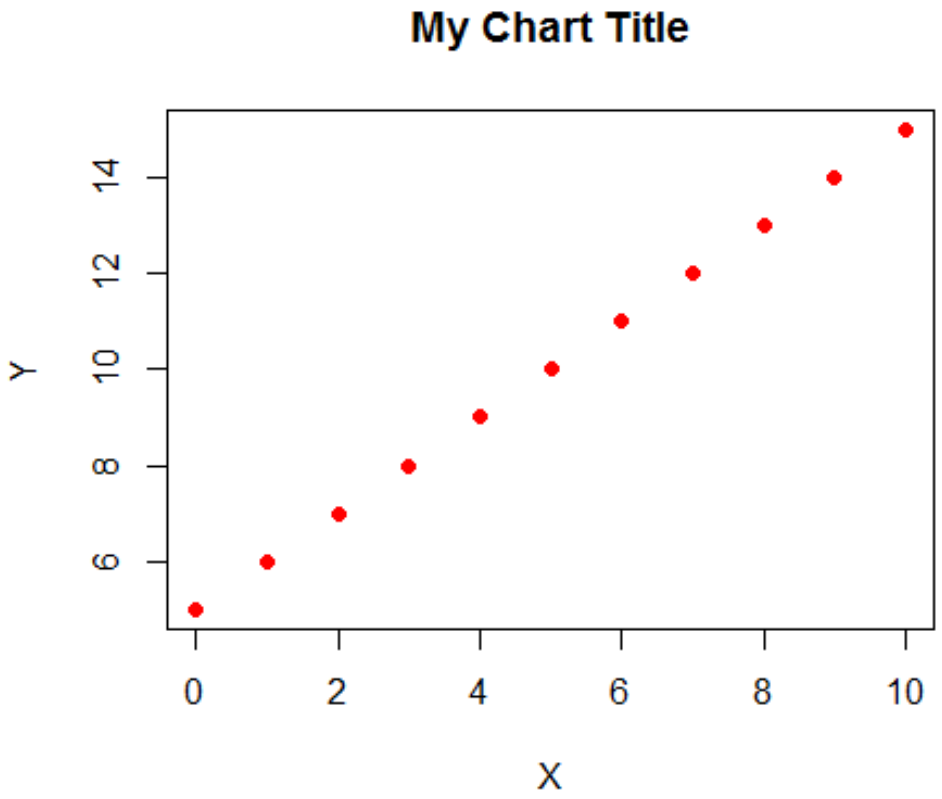
```
> mtcars1 <- mtcars[mtcars$cyl>5 & mtcars$cyl<8,]
> mtcars1
                mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4      21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Hornet 4 Drive 21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Valiant        18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Merc 280       19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C      17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Ferrari Dino   19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
```

```
> mtcars1[order(mtcars1$drat),]
                mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Valiant        18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Hornet 4 Drive 21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Ferrari Dino   19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
Mazda RX4      21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Merc 280       19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C      17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
```

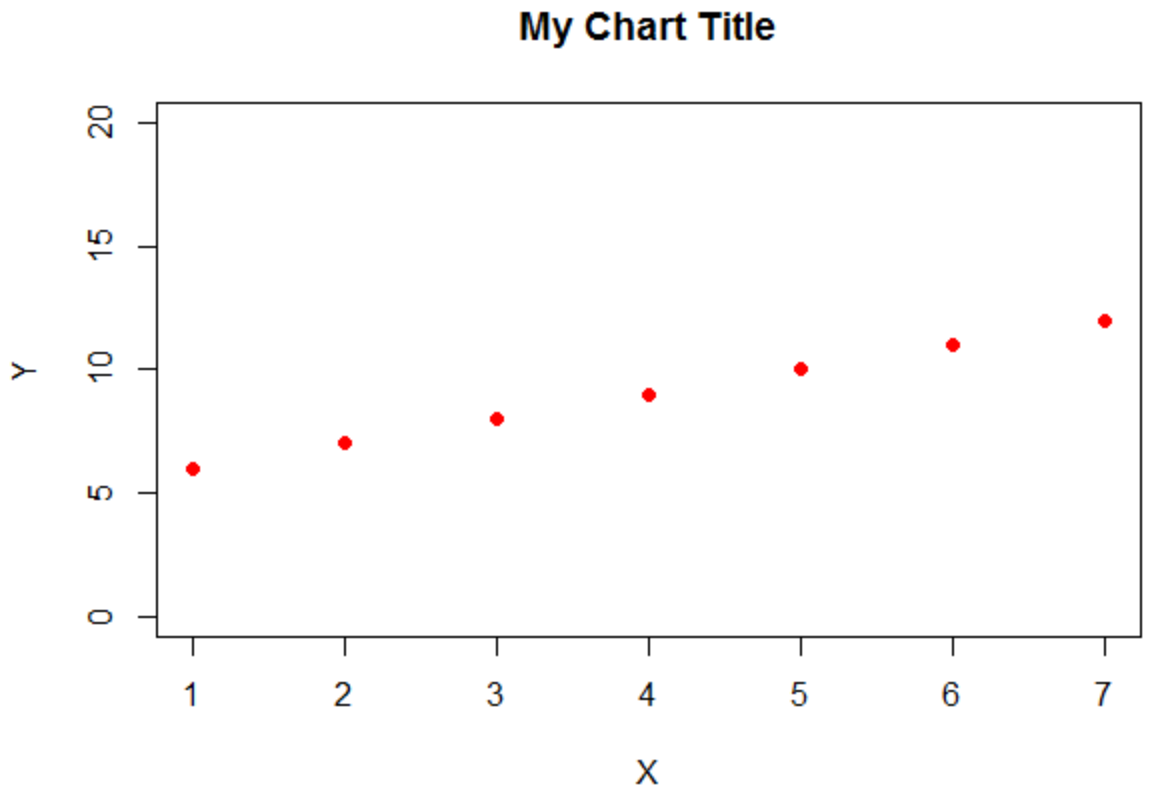
ГРАФИКА

Самый простой график

```
>x_data <- c(0:10)
>y_data <- x_data +5
>plot(x_data, y_data,  main = "My Chart Title", xlab ="X", ylab = "Y", pch=16, col = "red")
```



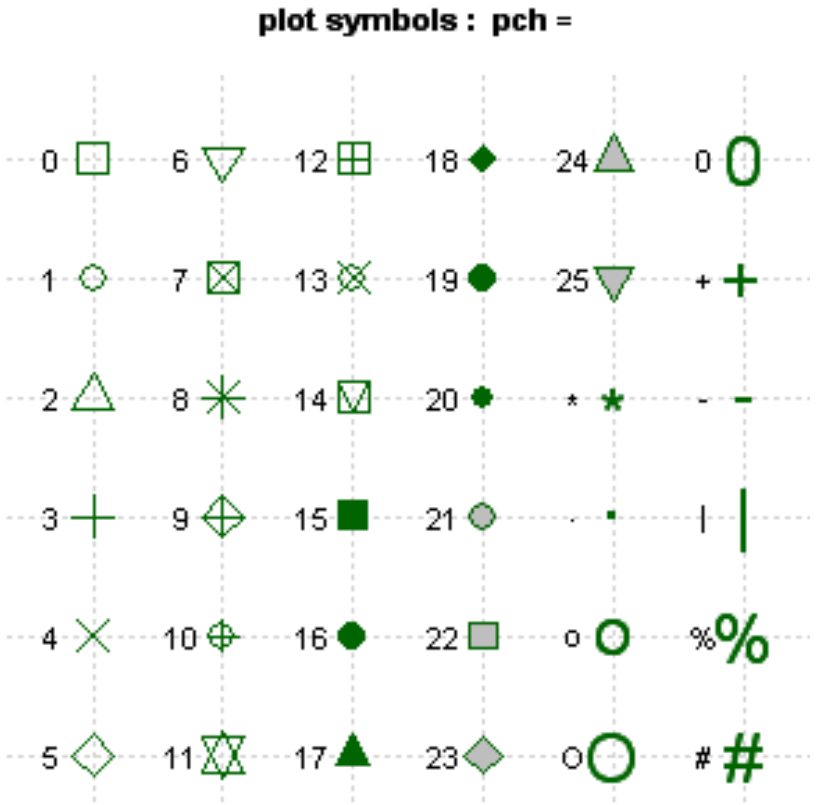
Параметры xlim, ylim



>plot(x_data, y_data, main = "My Chart Title", xlab ="X", ylab = "Y", pch=16, col = "red", xlim=c(1,7), ylim=c(0, 20))

Параметр pch

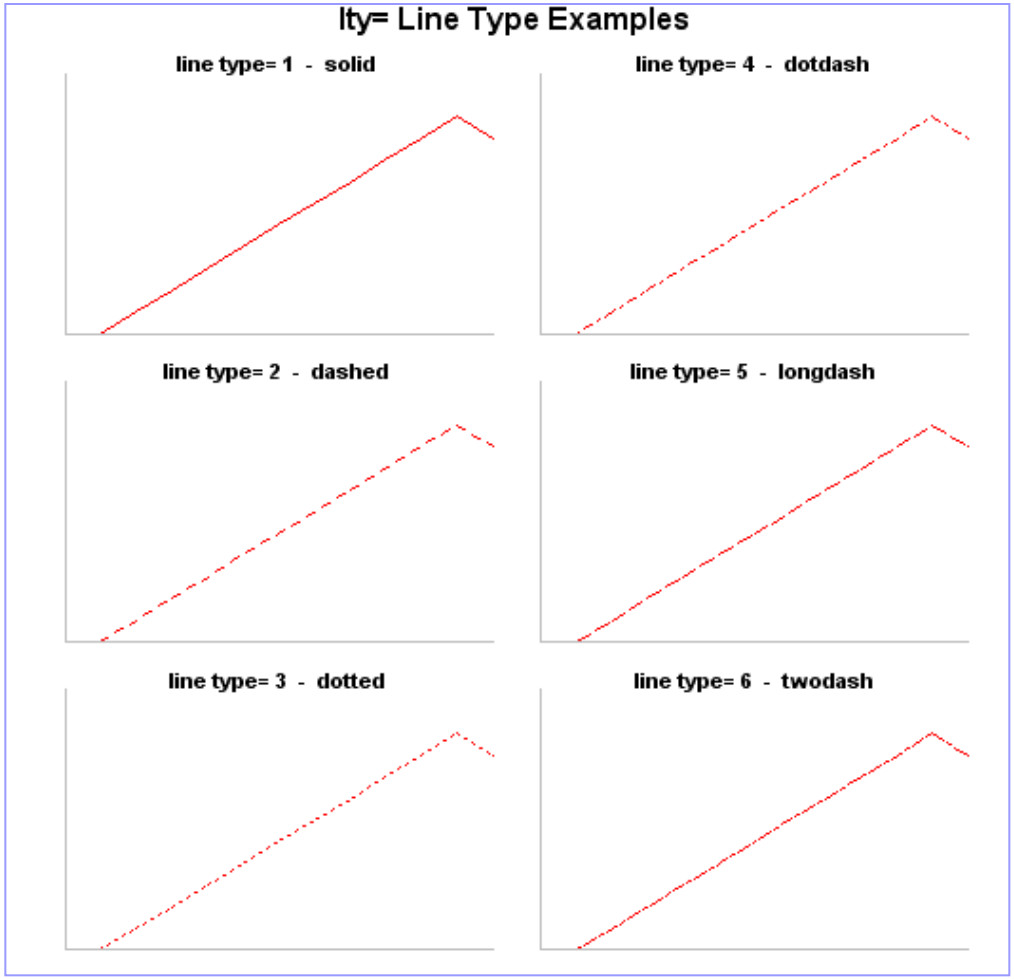
- В R существует 25 символов для графиков
- Символы 19 – 20 могут быть залиты выбранным цветом
- Символы 21: 25 могут быть залиты выбранным цветом (col) и обведены рамкой (bg)



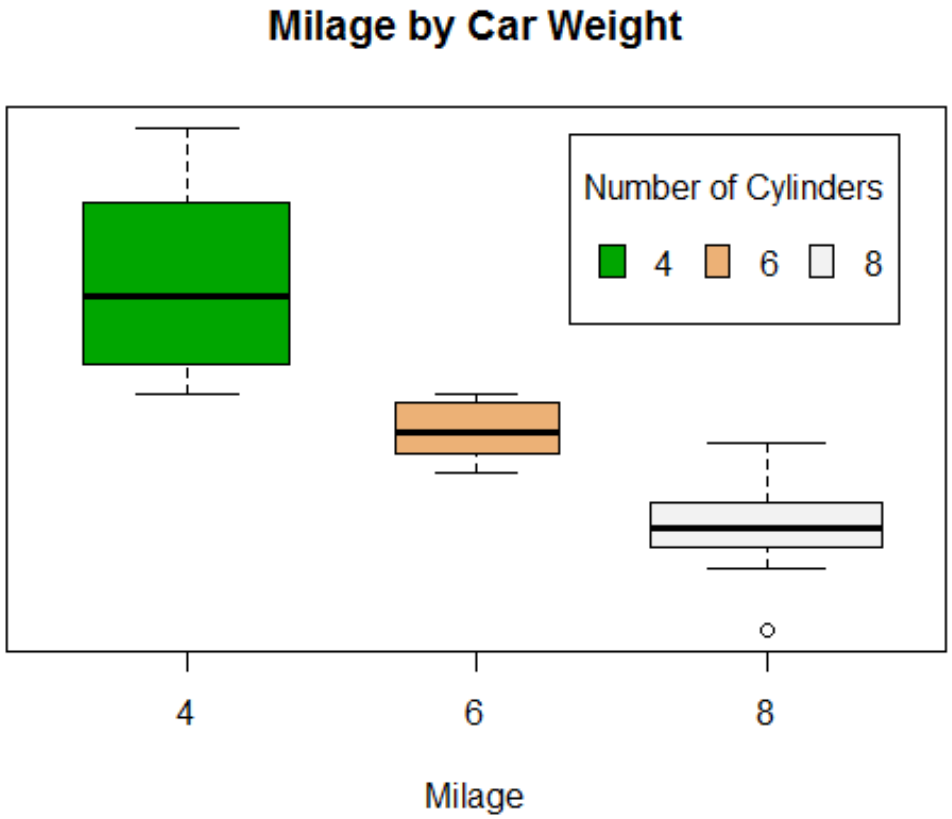
Параметр lty

- В R существует 7 типов линий

- 0 – «прозрачная линия»
  1 – «сплошная»
  2 – «пунктирная»
  3 – «точками»
  4 – «точка-тире»
  5 – «длинное тире»
  6 – «двойное тире»



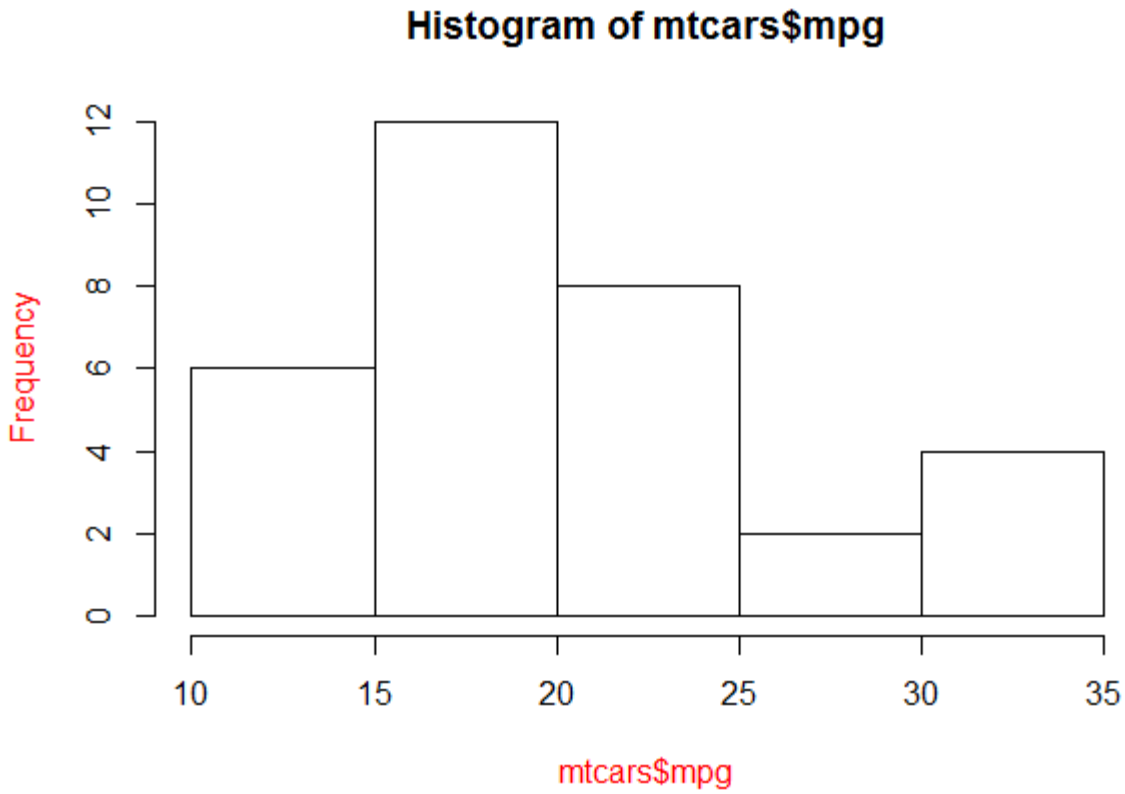
Параметр legend



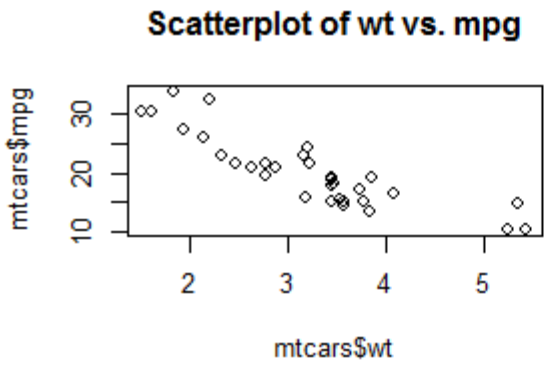

```
> boxplot(mtcars$mpg~mtcars$cyl, main="Milage by Car Weight", yaxt="n",
xlab="Milage", col=terrain.colors(3), varwidth=T)
> legend("topright", inset=.05, title="Number of Cylinders", c("4","6","8"),
fill=terrain.colors(3), horiz=TRUE)
```

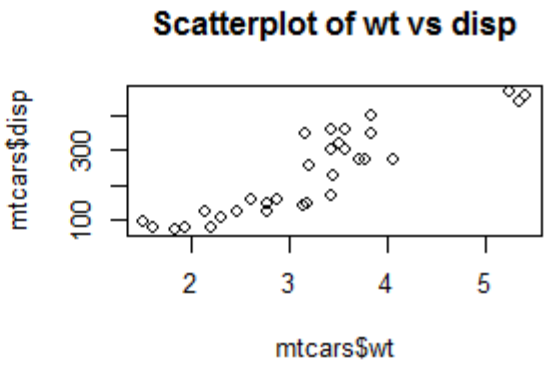
Графический параметр par()

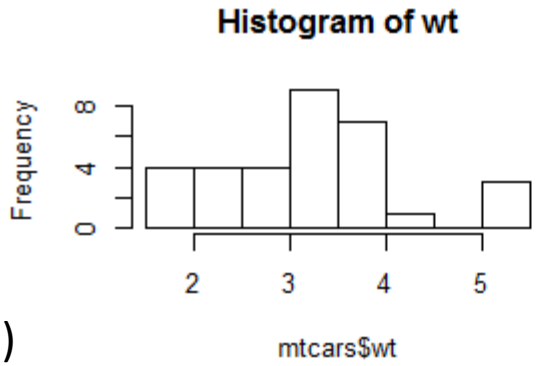
```
> par()                          # просмотреть текущие настройки
> par(col.lab="red")             # сделать красными подписи к осям
> hist(mtcars$mpg)               # нарисовать график с новыми настройками
```

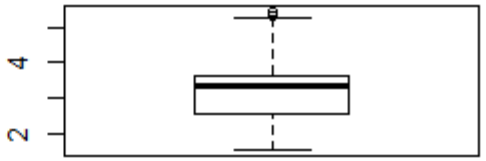


Комбинация графиков





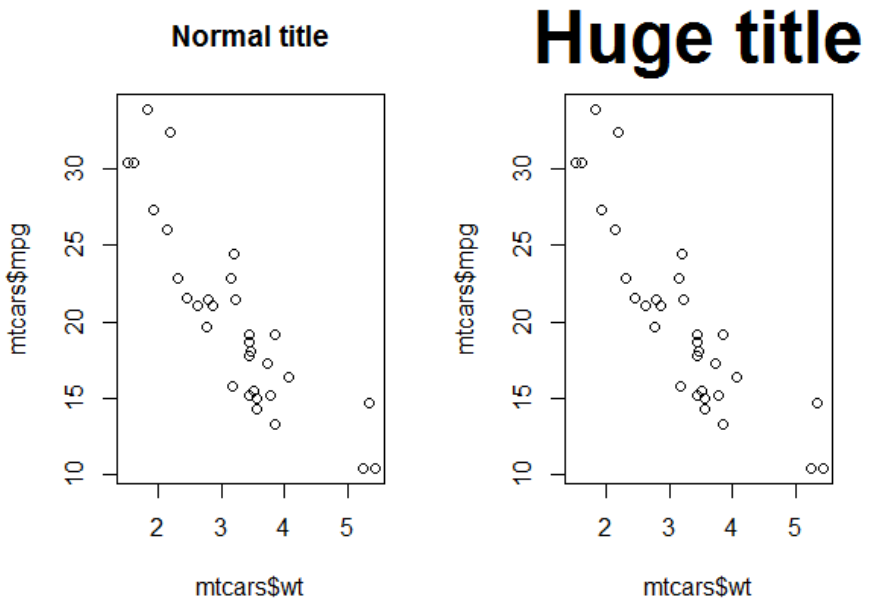





```
> par(mfrow=c(2,2))
> plot(mtcars$wt,mtcars$mpg,
main="Scatterplot of wt vs. mpg")
> plot(mtcars$wt,mtcars$disp,
main="Scatterplot of wt vs disp")
> hist(mtcars$wt, main="Histogram of wt")
> boxplot(mtcars$wt, main="Boxplot of wt")
```

Размер текста и символов

опция	описание
cex	Размер текста и символов относительно размера по умолчанию
cex.axis	Увеличение текста по осям
cex.lab	Увеличение подписей к осям
cex.main	Увеличение заголовков




```
> par(mfrow=c(1,2))
> plot(mtcars$mpg ~ mtcars$wt,
main="Normal title")
> plot(mtcars$mpg ~ mtcars$wt,
main="Huge title", cex.main=3)
```

Цвета 1/2

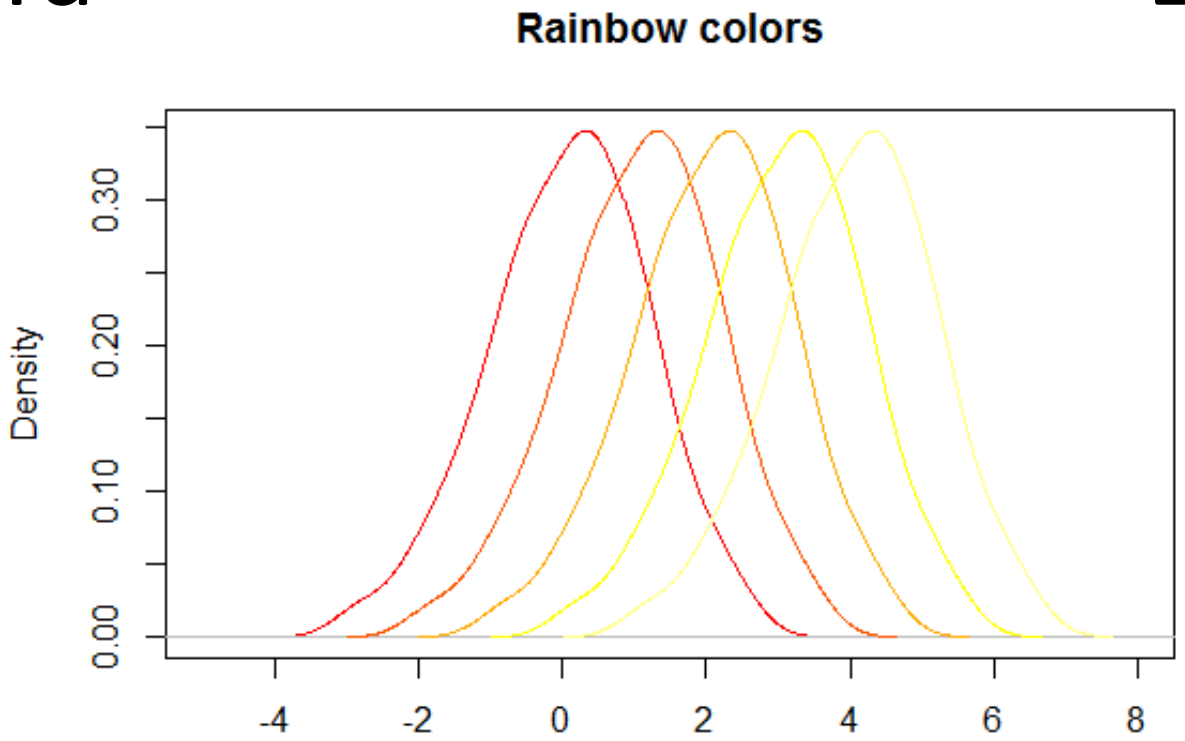
опция	описание
col	Цвет по умолчанию (может быть вектором)
col.axis	Цвет текста по осям
col.lab	Цвет подписей к осям
col.main	Цвет заголовков

Можно использовать функции

rainbow(*n*), **heat.colors(*n*)**, **terrain.colors(*n*)**, **topo.colors(*n*)** и **cm.colors(*n*)**

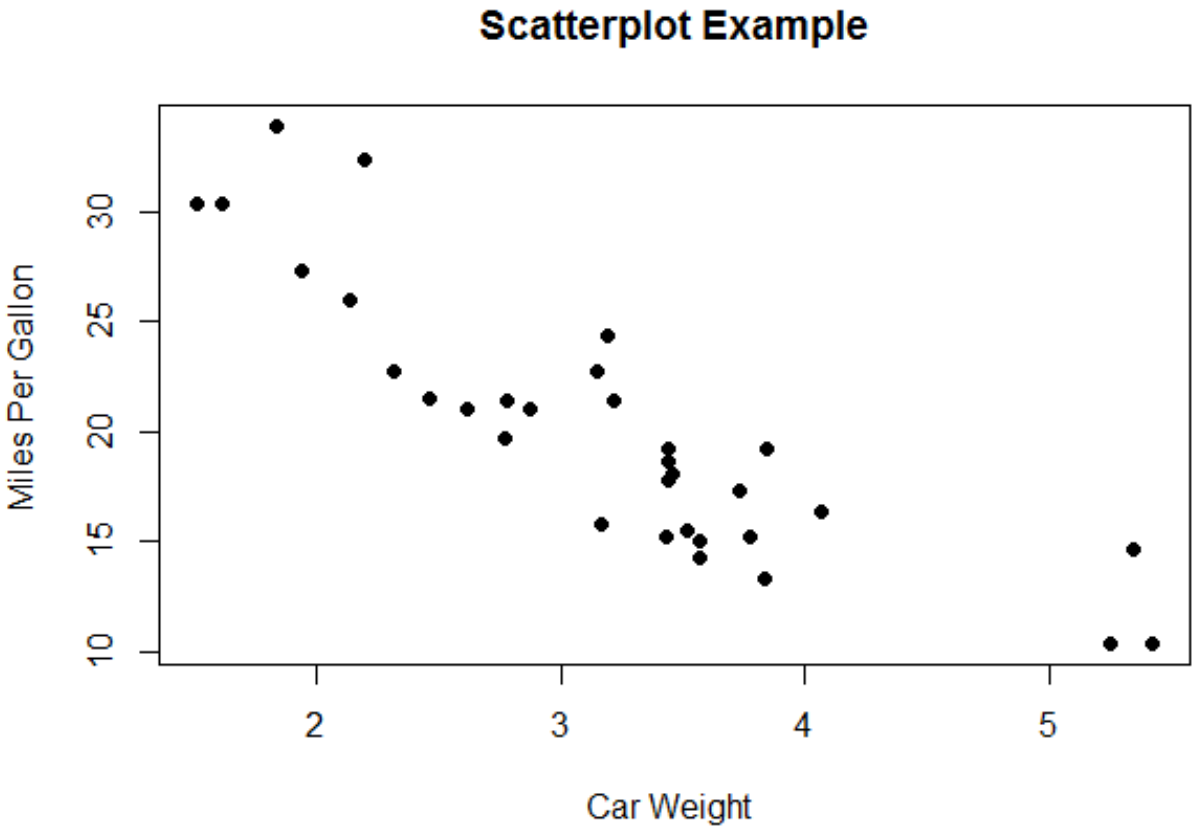
для создания вектора цветов

Цвета 2/2



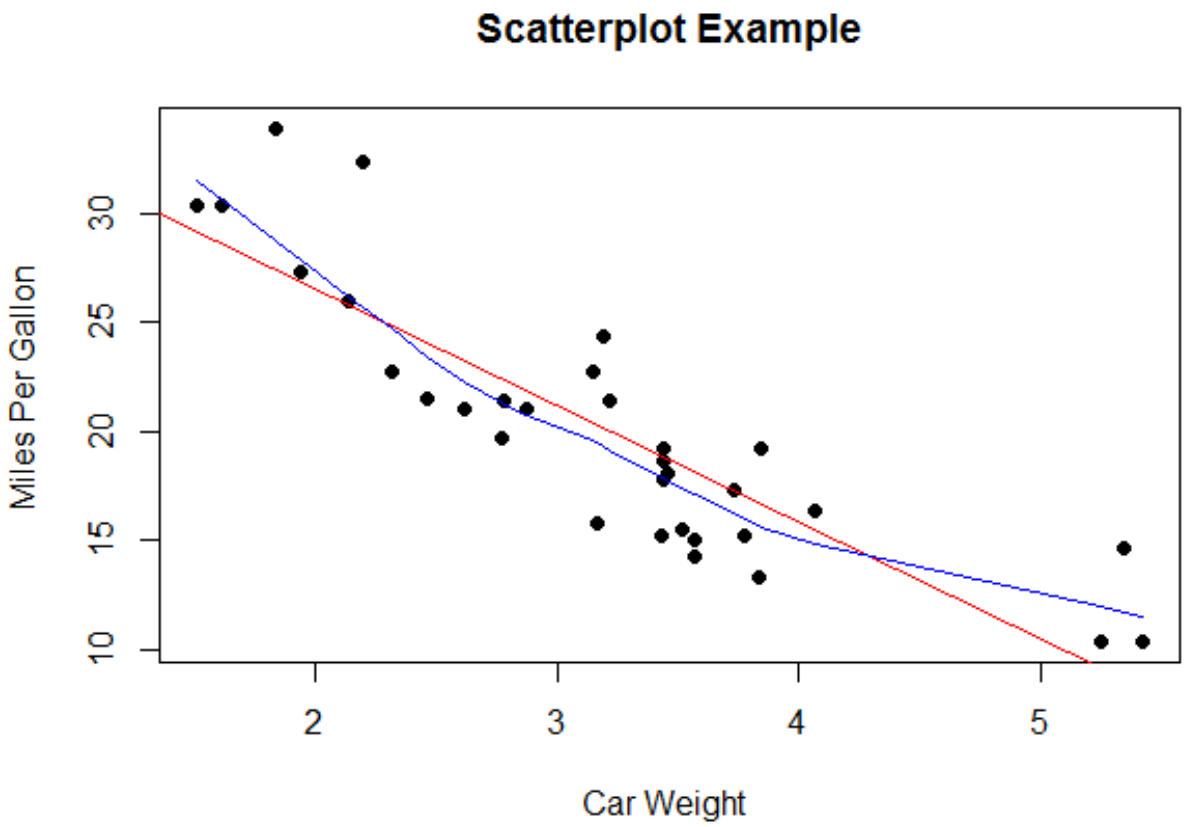

```
> x1 <- rnorm(100) ; x2 <- x1+1 ; x3 <- x2+1 ; x4 <- x3+1 ; x5 <- x4+1
> ourCol <- heat.colors(5)
> plot(density(x1), col=ourCol[1], xlim=c(-5,8), main="Rainbow colors", xlab="")
> lines(density(x2), col=ourCol[2])
> lines(density(x3), col=ourCol[3])
> lines(density(x4), col=ourCol[4])
> lines(density(x5), col=ourCol[5])
```

Scatterplots



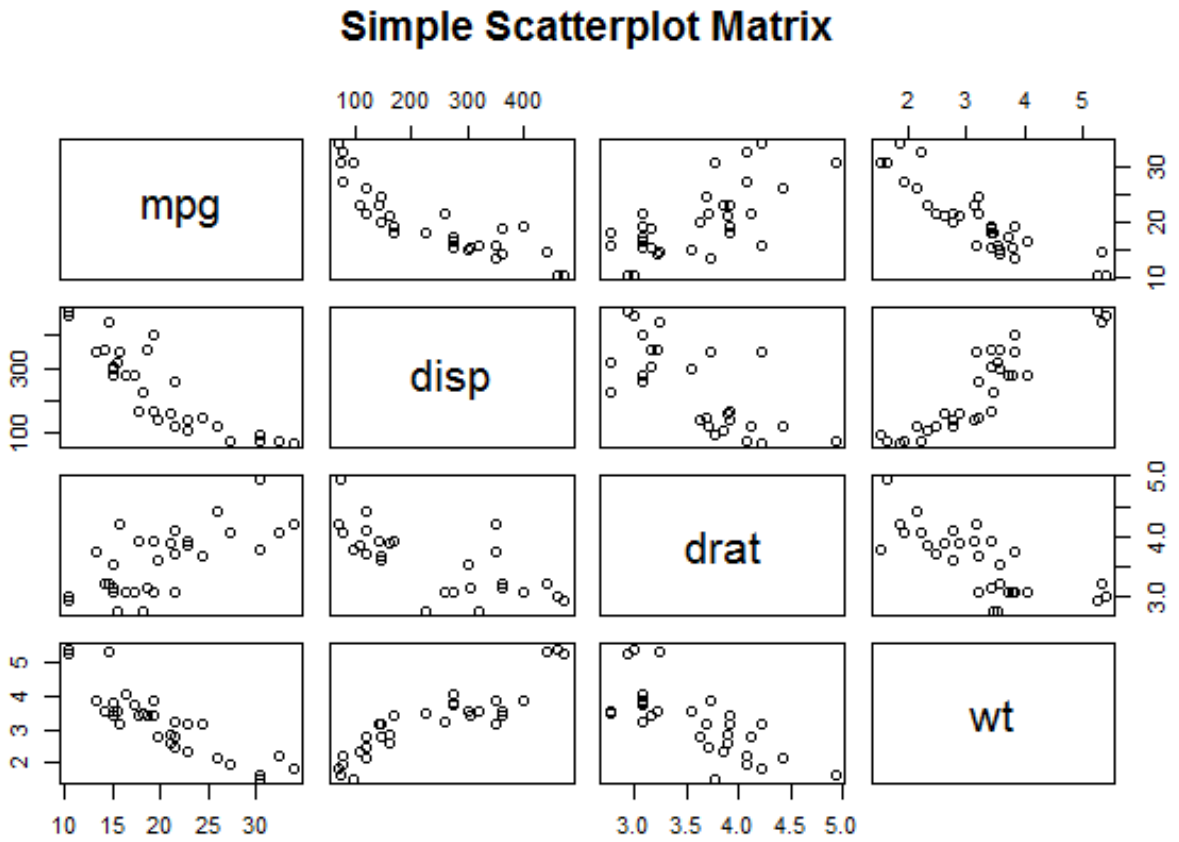

```
> plot(mtcars$wt, mtcars$mpg, main="Scatterplot
Example",    xlab="Car Weight ", ylab="Miles Per Gallon ", pch=19)
```

Scatterplots: добавим линии




```
> abline(lm(mtcars$mpg~mtcars$wt), col="red")
> lines(lowess(mtcars$wt, mtcars$mpg), col="blue")
```

Scatterplot: матрицы



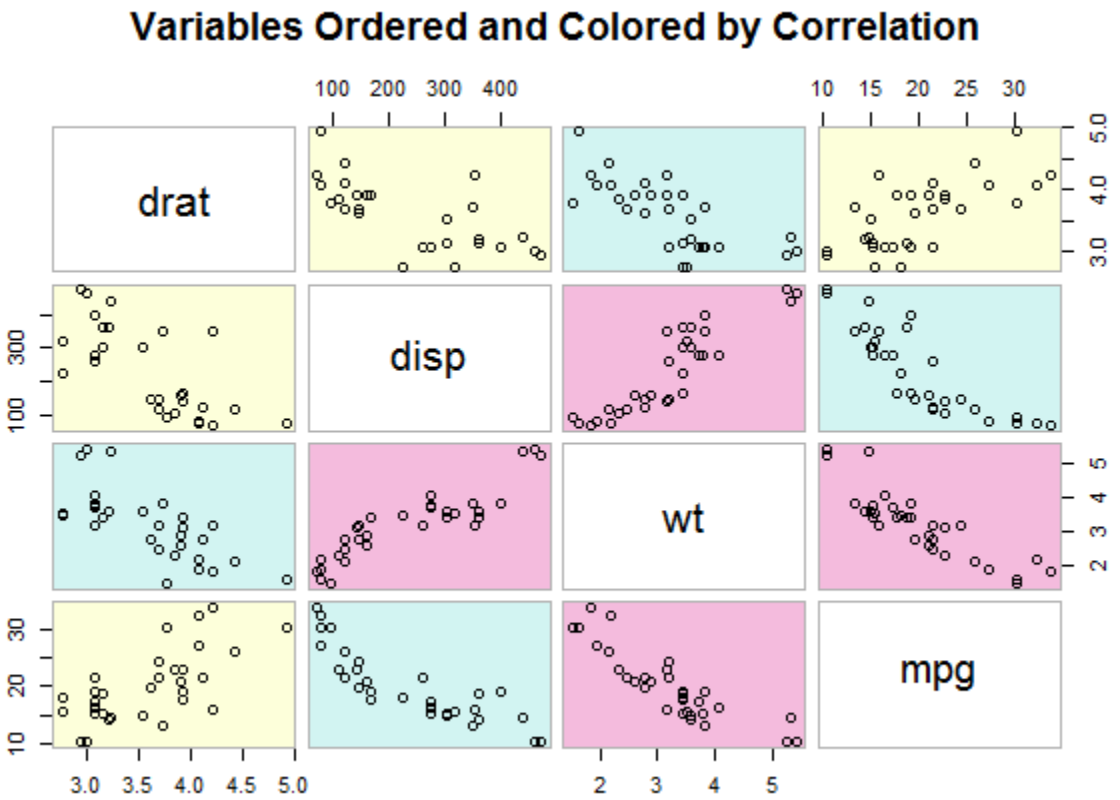

```
> pairs(mtcars[,c(1,3,5,6)])
```

#или то же самое:

```
> pairs(~mpg+disp+drat+wt, data=mtcars, main="Simple Scatterplot Matrix")
```

Другие scatterplots

gclus package позволяет группировать переменные таким образом, чтобы переменные с большими корреляциями были ближе к диагонали. Цвета соответствуют коэффициенту корреляции.




```
> library(gclus)
> dta <- mtcars[,c(1,3,5,6)]
> dta.r <- abs(cor(dta))
> dta.col <- dmat.color(dta.r)
> dta.o <- order.single(dta.r)
> cpairs(dta, dta.o, panel.colors=dta.col, gap=.5,
main="Variables Ordered and Colored by Correlation" )
```

> dta.r

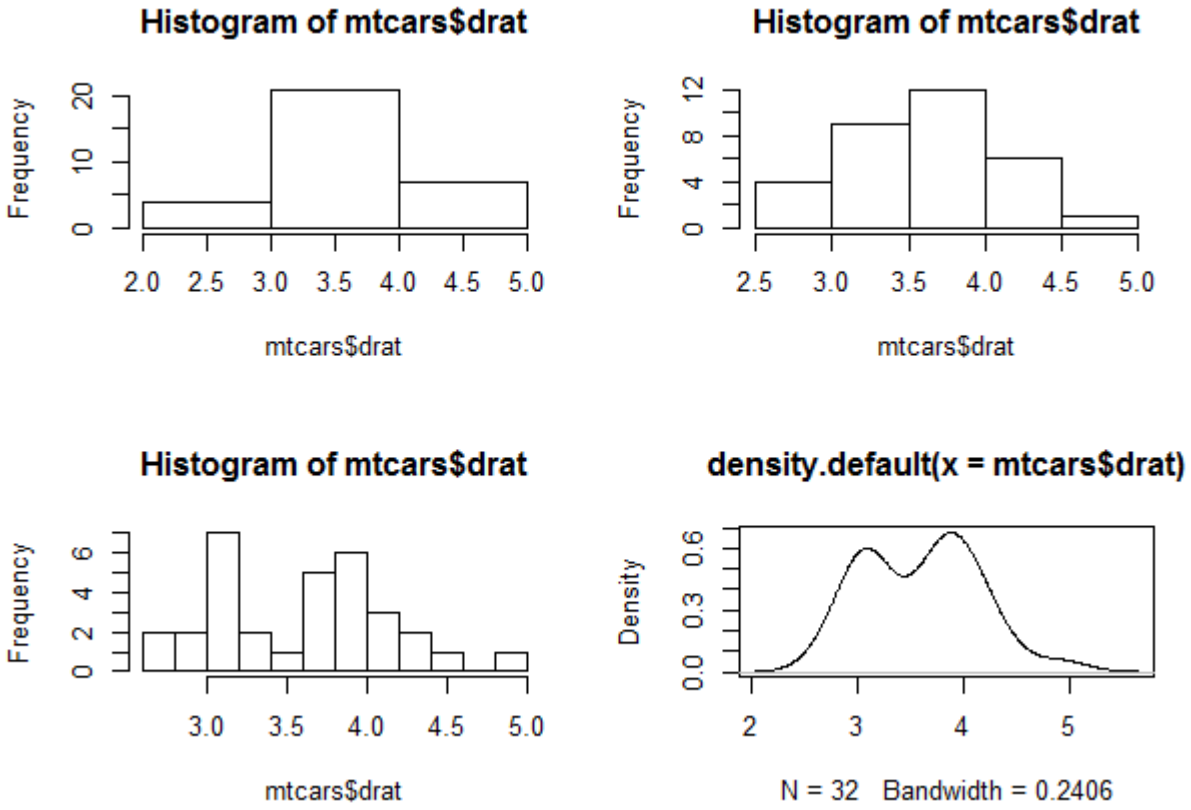
```
            mpg       disp       drat         wt
mpg  1.0000000 0.8475514 0.6811719 0.8676594
disp 0.8475514 1.0000000 0.7102139 0.8879799
drat 0.6811719 0.7102139 1.0000000 0.7124406
wt   0.8676594 0.8879799 0.7124406 1.0000000
```

dmat.color: метод, который берет на вход матрицу с корреляциями, возвращает матрицу цветов

> dta.col

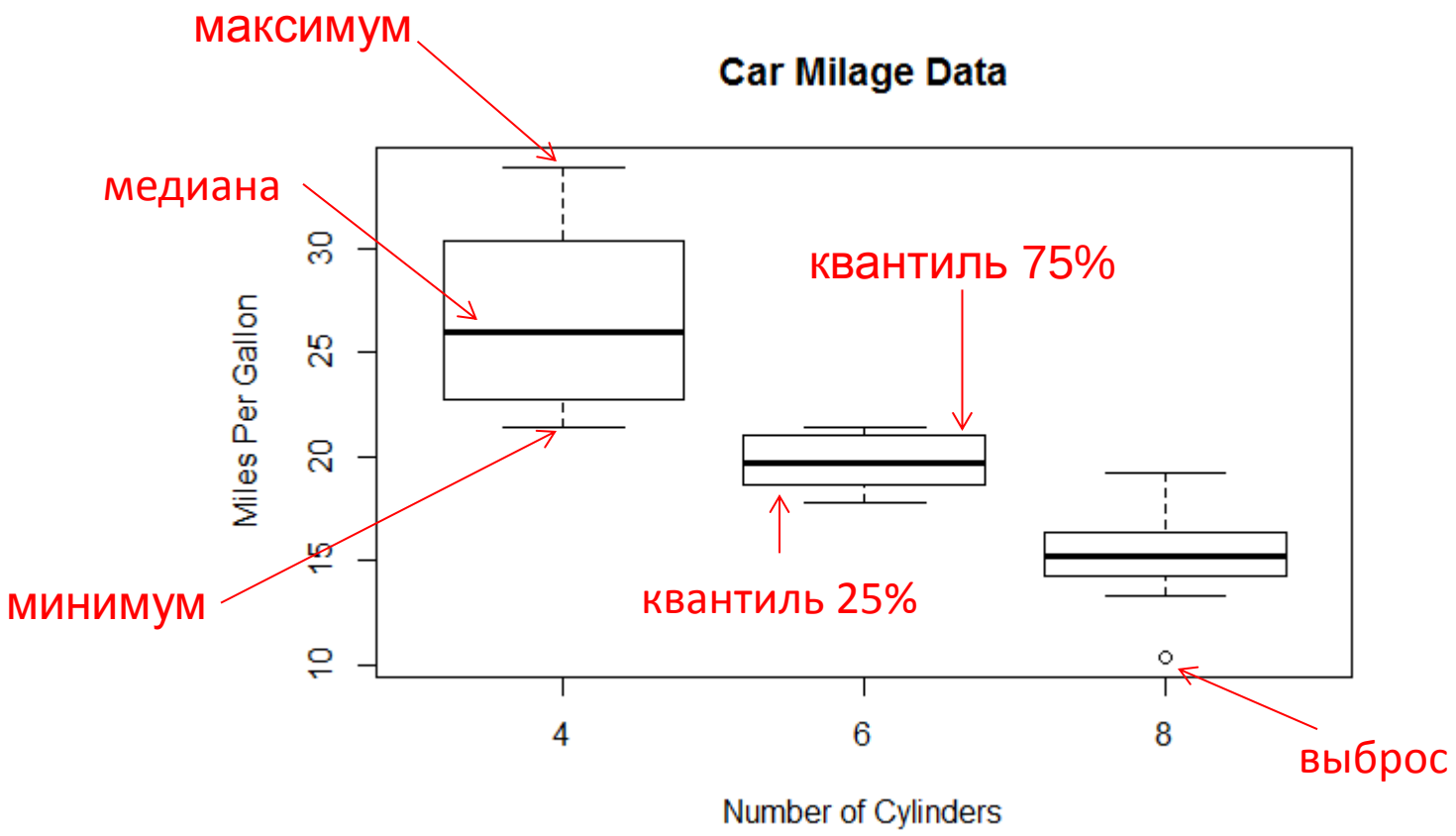
```
     mpg         disp        drat        wt
mpg  NA          "#D2F4F2"   "#FDFFDA"   "#F4BBDD"
disp "#D2F4F2"   NA          "#FDFFDA"   "#F4BBDD"
drat "#FDFFDA"   "#FDFFDA"   NA          "#D2F4F2"
wt   "#F4BBDD"   "#F4BBDD"   "#D2F4F2"   NA
```

Гистограммы




```
> par(mfrow=c(2,2))
> hist(mtcars$drat, breaks=3)
> hist(mtcars$drat, breaks=5)
> hist(mtcars$drat, breaks=12)
> plot(density(mtcars$drat))
```

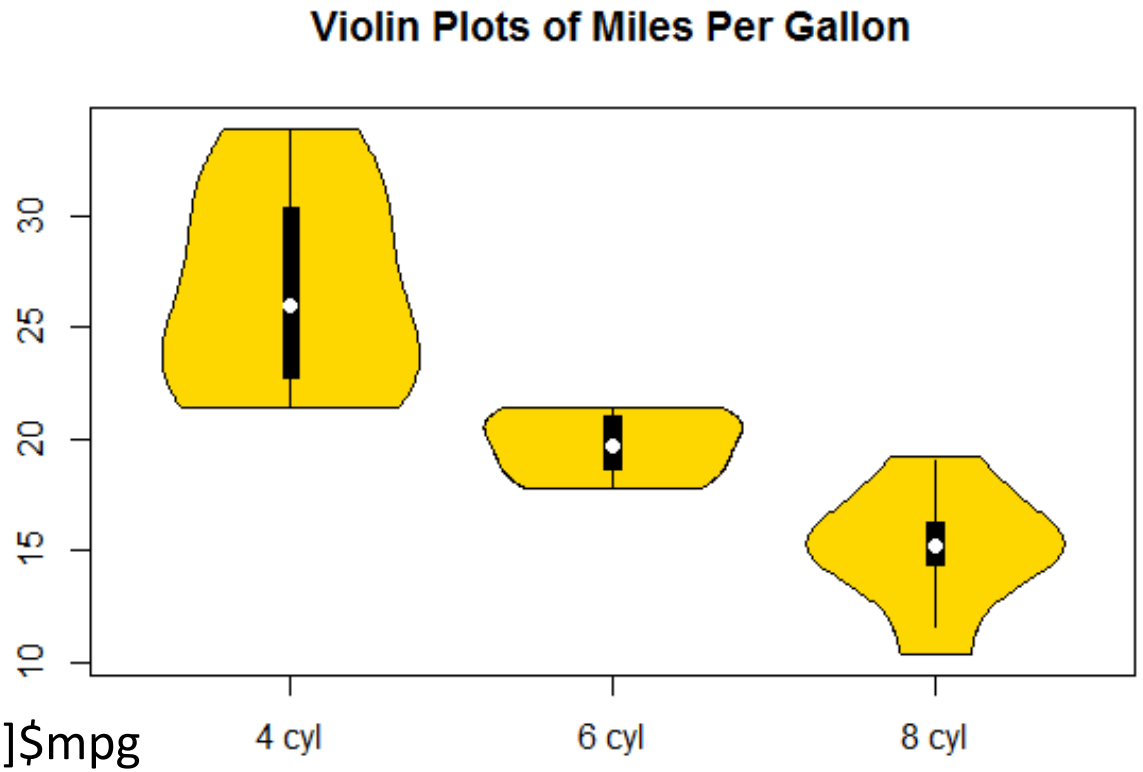
Boxplots




```
> boxplot(mpg~cyl,data=mtcars, main="Car Milage Data",    xlab="Number of Cylinders",
ylab="Miles Per Gallon")
```

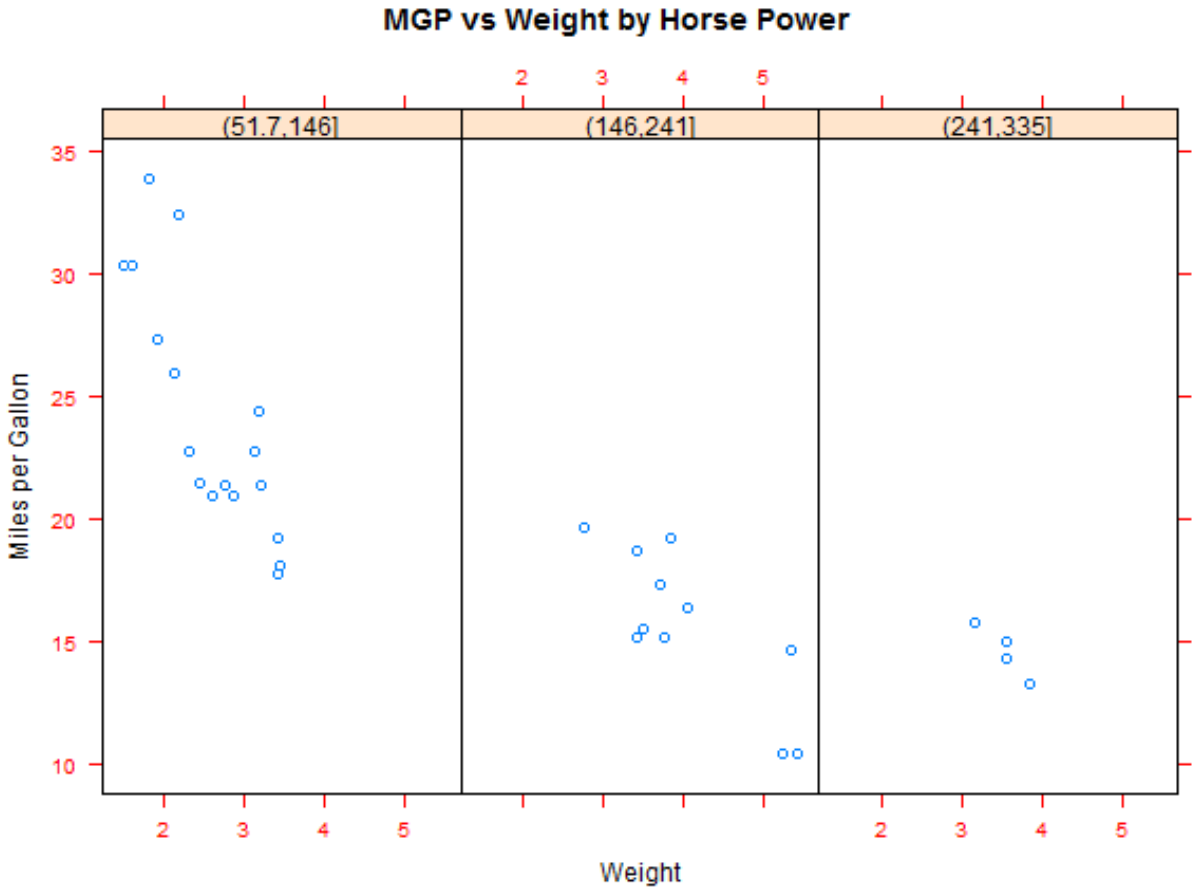
Violin Plot: комбинация boxplot и графика плотности распределения

«The violin plot is like the lovechild between a density plot and a box-and-whisker plot.»



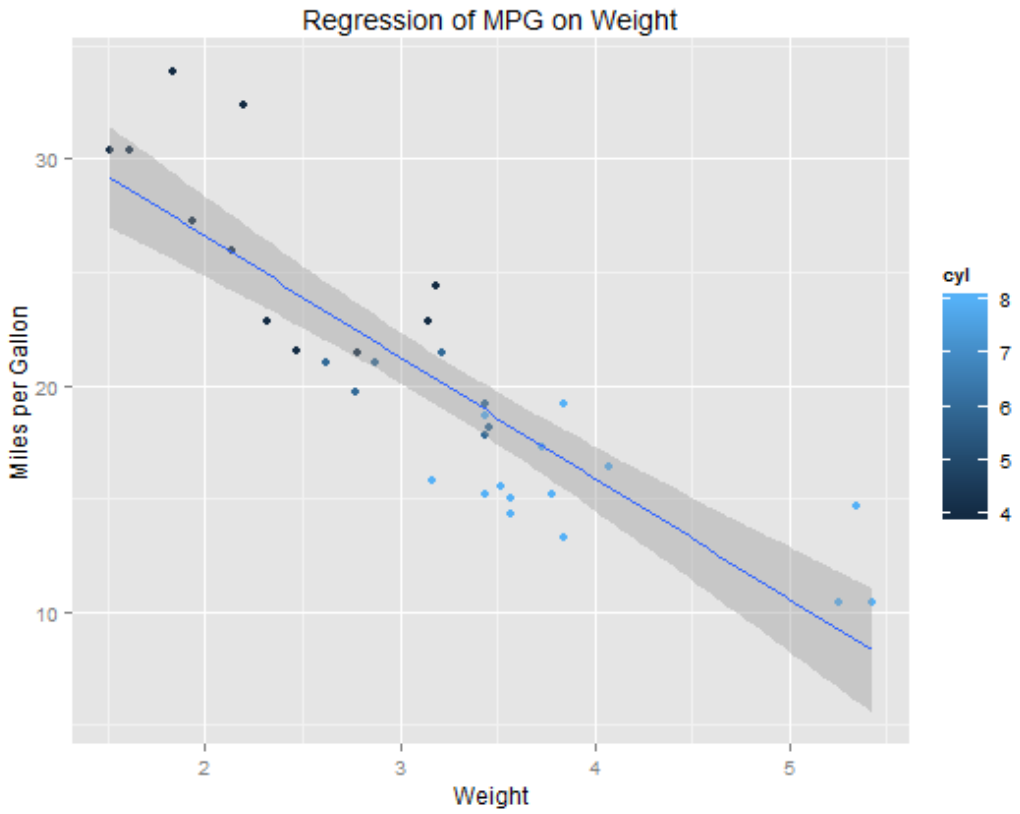

```
> library(vioplot)
> x1 <- mtcars[mtcars$cyl==4,]$mpg
> x2 <- mtcars[mtcars$cyl==6,]$mpg
> x3 <- mtcars[mtcars$cyl==8,]$mpg
> vioplot(x1, x2, x3, names=c("4 cyl", "6 cyl", "8 cyl"), col="gold")
title("Violin Plots of Miles Per Gallon")
```

Возможности lattice




```
> library(lattice)
> hp <- cut(mtcars$hp,3) # divide horse power into three bands
> xyplot(mtcars$mpg~mtcars$wt|hp, scales=list(cex=.8, col="red"), xlab="Weight",
ylab="Miles per Gallon", main="MGP vs Weight by Horse Power")
```

Возможности ggplot2




```
> qplot(wt, mpg, data=mtcars, geom=c("point", "smooth"), method="lm", formula=y~x,
color=cyl, main="Regression of MPG on Weight", xlab="Weight", ylab="Miles per
Gallon")
```

Больше графиков по ссылкам

- http://www.statmethods.net/advgraphs/
- http://gallery.r-enthusiasts.com/thumbs.php

Работа с missing data 1/2

```
> newRow <- mtcars[1,]

> rownames(newRow) <- "Lada"

> newRow[4] <- NA

> mtcarsNew <- rbind(mtcars, newRow)

> mtcarsNew[30:33,]
               mpg cyl disp  hp drat   wt  qsec vs am gear carb
Ferrari Dino  19.7   6  145 175 3.62 2.77 15.50  0  1    5    6
Maserati Bora 15.0   8  301 335 3.54 3.57 14.60  0  1    5    8
Volvo 142E    21.4   4  121 109 4.11 2.78 18.60  1  1    4    2
Lada          21.0   6  160  NA 3.90 2.62 16.46  0  1    4    4

> mean(mtcarsNew$hp)
[1] NA

> any(is.na(mtcarsNew$hp))
[1] TRUE
```

Работа с missing data 2/2

```
> mean(mtcarsNew$hp, na.rm=TRUE)
[1] 146.6875

> which(is.na(mtcarsNew$hp))
[1] 33

> which(c(FALSE, TRUE, FALSE, TRUE))
[1] 2 4

> mtcarsA <- na.omit(mtcarsNew)

> dim(mtcarsNew)
[1] 33 11

> dim(mtcarsA)
[1] 32 11
```

! #как работает команда which

#или просто уберем все строки, содержащие NA

#проверим, изменилось ли число строк

Что еще можно добавить на график

grid(nx, ny**)**	Add grid lines to current plot. NA stop grid in corresponding direction
axis(side n,**)**	Add axis at side n to current plot
box(which=, **)**	Add box around current plot, figure or outer margin area depending on which specified
legend	Add legend to current plot
arrows(x,y**)** **lines(**x, y**)** **points(**x,y**)**	Add arrow line, line or points to current plot. type = can be used to specify style ("p","b", "l", etc)
abline(a,b**)** **abline(**h= or v=**)**	Add line to current plot. a is intercept, b is slope. h/v for horizontal/ vertical line
segments(x0,x1,y0, y1**)**	Add line segment(s) between pairs of points
polygon(x,y**)**	Add polygon defined by vectors x and y
text(x,y, "note"**)**	Add text to current plot at x & y