Занятие 2
Data frame
Построение графиков

26 февраля 2014

Содержание лекции

- 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

	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

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                  # Обращение к столбцу
[1] 2 3 5               по имени, можно
> colnames(df)          использовать tab!
[1] "n" "s" "b"
> rownames(df)          # Важно, что это
[1] "1" "2" "3"         имена строк, а не
> dim(df)               числа!
[1] 3 3
```

Использование data()

> mtcars

	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

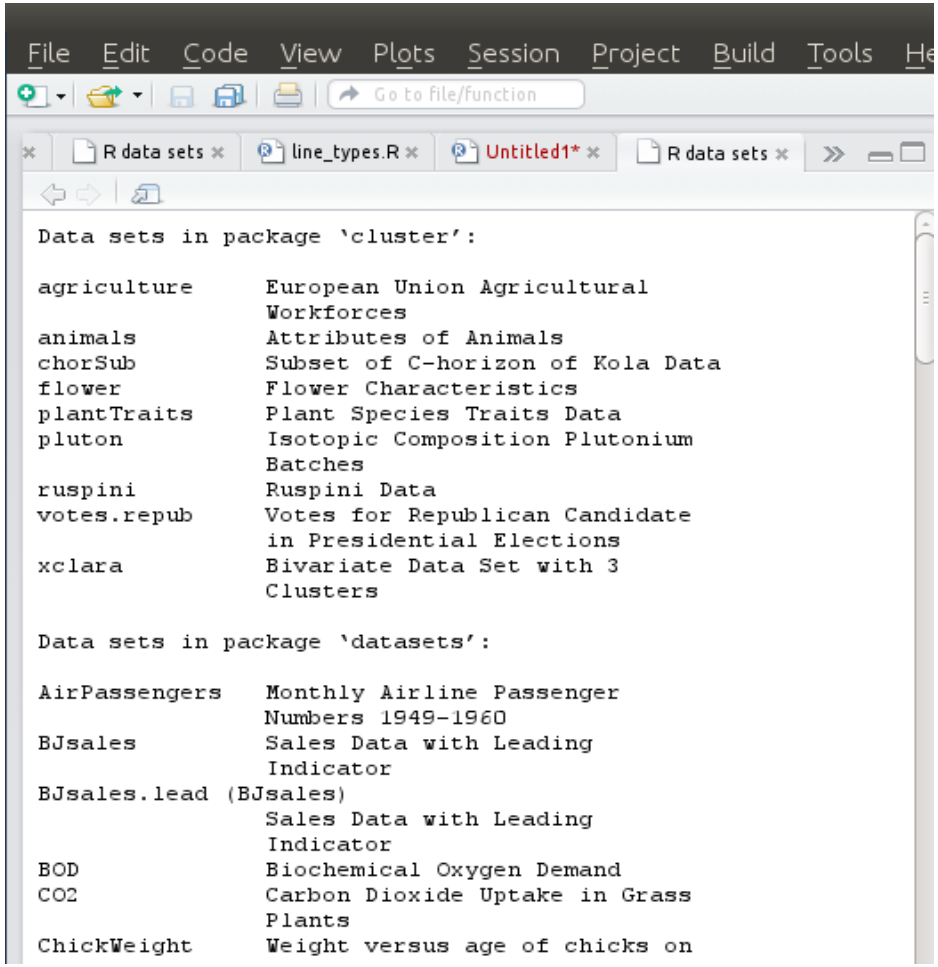
Командой data() можно посмотреть, какие выборки загружены для использования

> data()

!

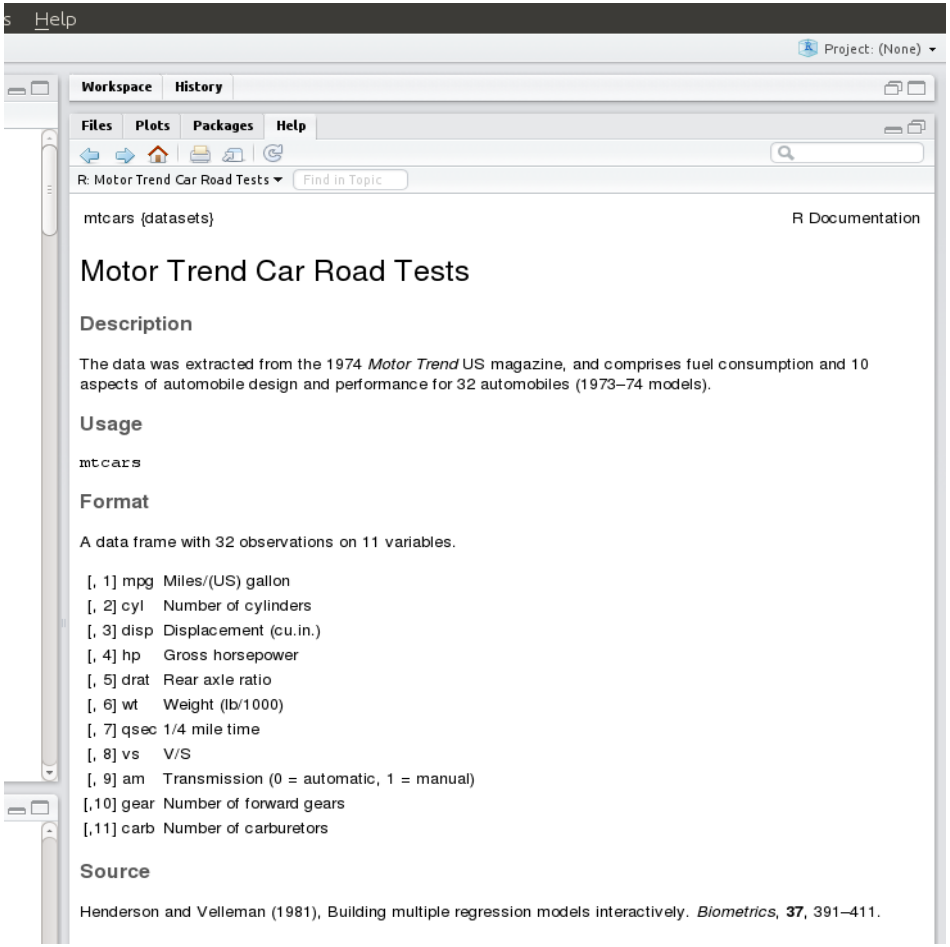
Использование data()

> data()



Можно узнать о доступной выборке более подробно

> ?mtcars



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 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
```

Добавить столбец

```
> dim(mtnew)
[1] 33 11
> num<-1:33
> mtnew<-cbind(mtnew, num)        #добавляем столбец
> mtnew[30:33,]
               mpg cyl disp  hp drat   wt  qsec vs am gear carb num
Ferrari Dino  19.7   6  145 175 3.62 2.77 15.50  0  1    5    6  30
Maserati Bora 15.0   8  301 335 3.54 3.57 14.60  0  1    5    8  31
Volvo 142E    21.4   4  121 109 4.11 2.78 18.60  1  1    4    2  32
Lada          21.0   6  150 120 4.00 2.50 16.46  1  1    4    4  33
```

Добавить строку

```
> mtnew<-mtcars

> dim(mtnew)
[1] 32 11
> mtnew[1,]
          mpg cyl disp  hp drat   wt  qsec vs am gear carb
Mazda RX4  21   6  160 110  3.9 2.62 16.46  0  1    4    4

> newcar<-c(21, 6, 150, 120, 4.0, 2.5, 16.46, 1, 1, 4, 4)#работает только если все данные одного типа!!!!

> newcar<-data.frame(mpg=21, cyl=4, disp=100, hp=80, drat=1, wt=2, qsec=16, vs=1,am=0, gear=4, carb=1) # data.frame из 1 строки

> mtnew<-rbind(mtnew, newcar)      #добавляем строку

> rownames(mtnew)[33]<-"Lada"      #присваиваем ей имя

> mtnew[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  150 120 4.00 2.50 16.46  1  1    4    4
```

Логические условия и order

```
> mtcars1 <- mtcars[mtcars$cyl>4 & mtcars$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

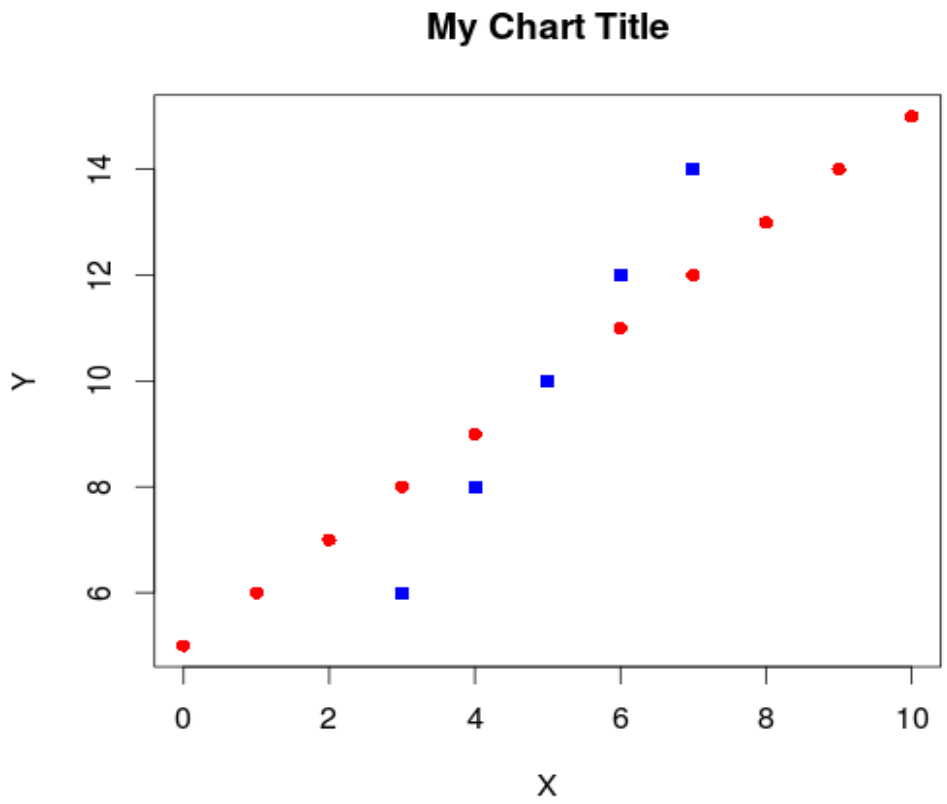
```
> mtcars1[order(mtcars1$drat),]
```

	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

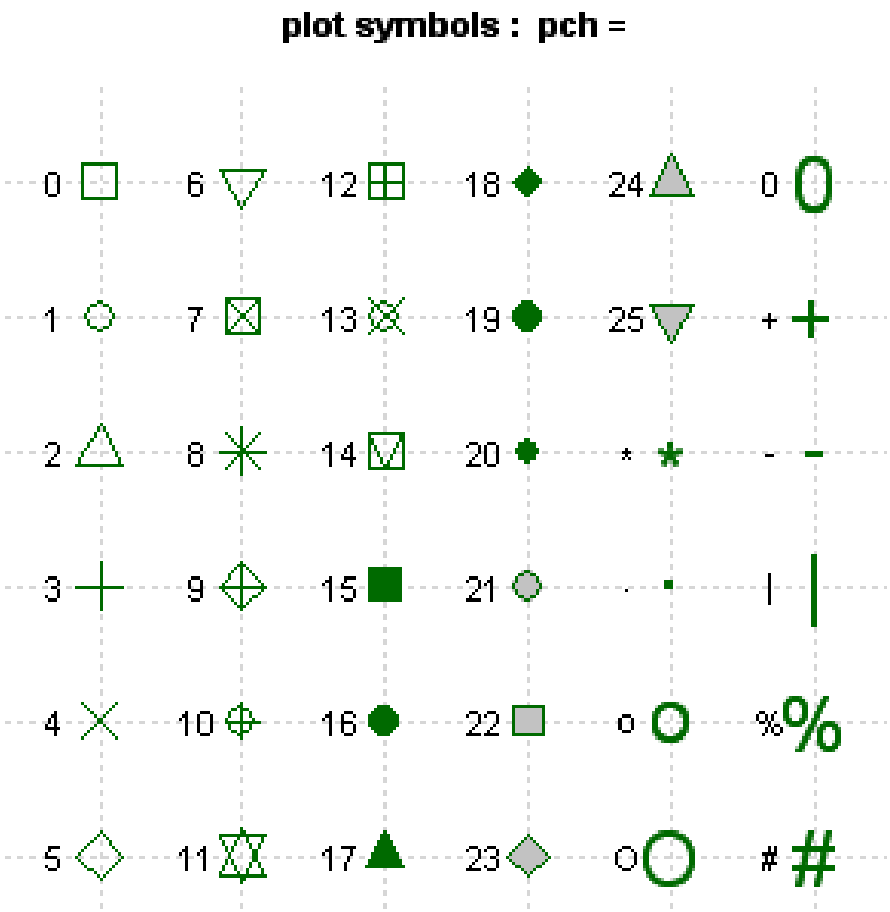
ГРАФИКА

Параметр pch

```
>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")
> y2_data<-x_data*2
> lines(x_data, y2_data, pch=15, col="blue", type="p")
```

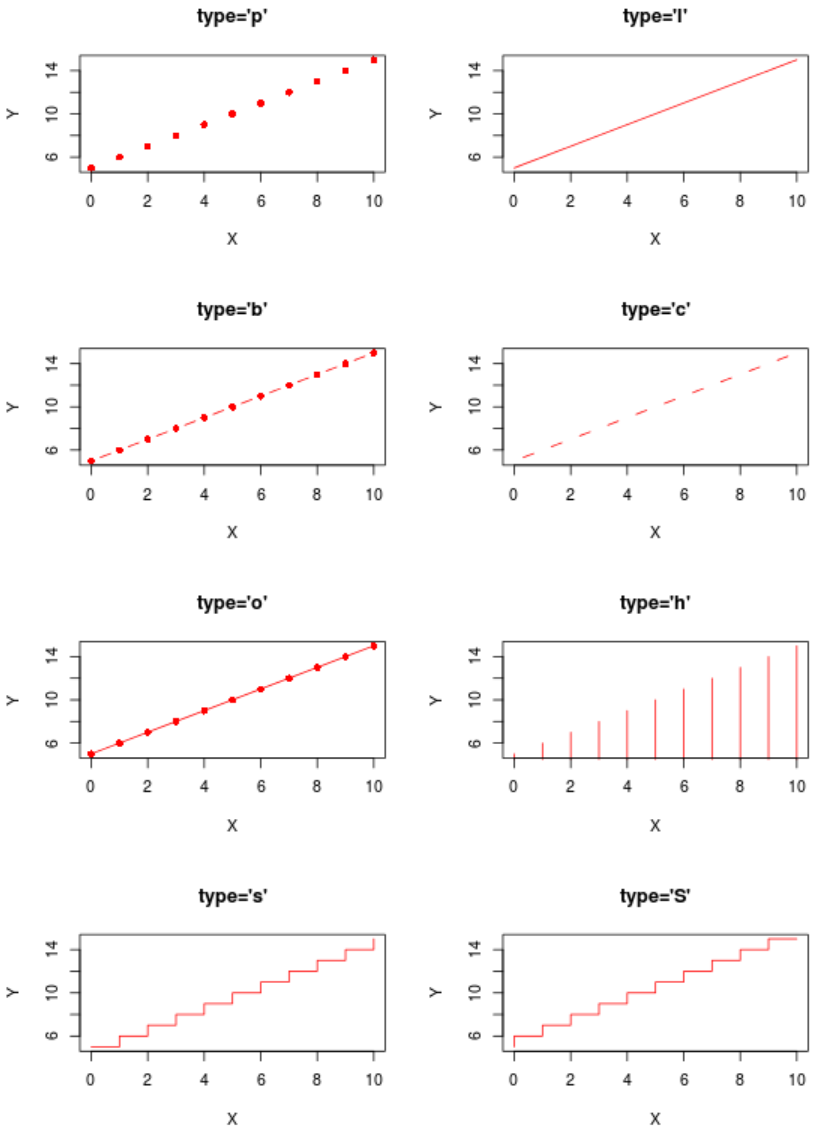


Параметр pch



- В R существует 25 символов для графиков
- Символы 15 – 20 могут быть залиты выбранным цветом
- Символы 21- 25 могут быть залиты выбранным цветом (col) и обведены рамкой (bg)

Параметр type




```
plot(x_data, y_data,   xlab ="X", ylab = "Y", pch=16, col = "red", type='p')
```

"p" точки

"l" сплошная линия

"b" точка-тире

"c" тире

"o" точки на линии

"h" вертикалиные линии (вроде гистограммы)

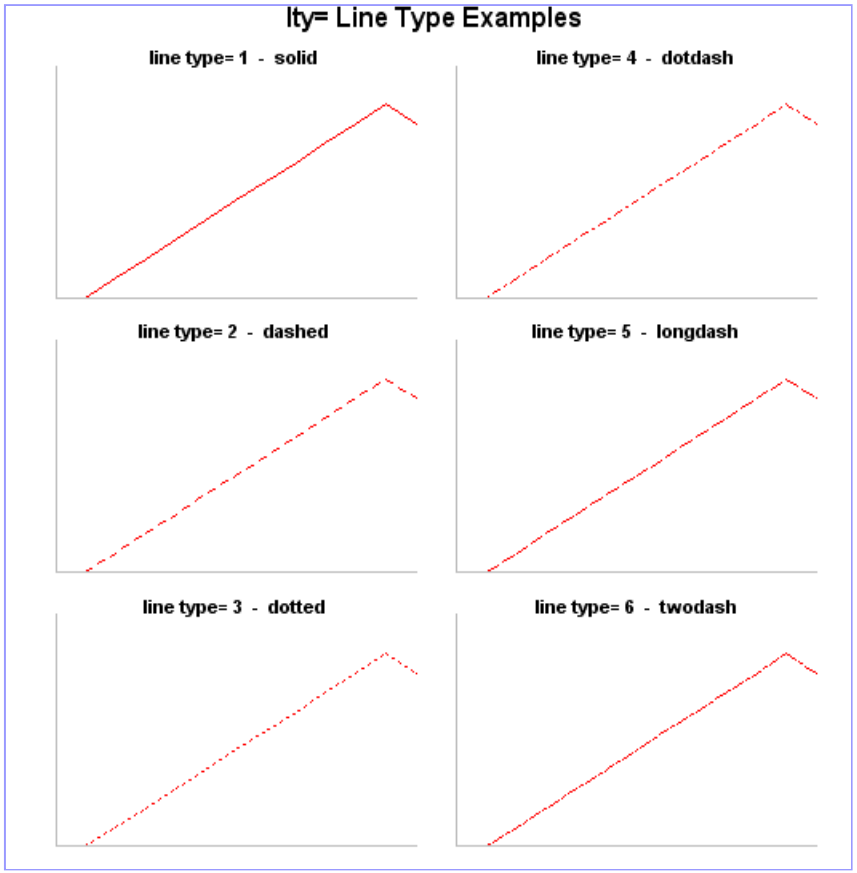
"s" лесенкой

"S" другой лесенкой

"n" прозрачная

Параметр lty

В R существует 7 типов линий




```
plot(x_data, y_data, xlab ="X", ylab = "Y", pch=16, col = "red", type='l', lty=2)
```

0 – «прозрачная линия»

1 – «сплошная»

2 – «пунктирная»

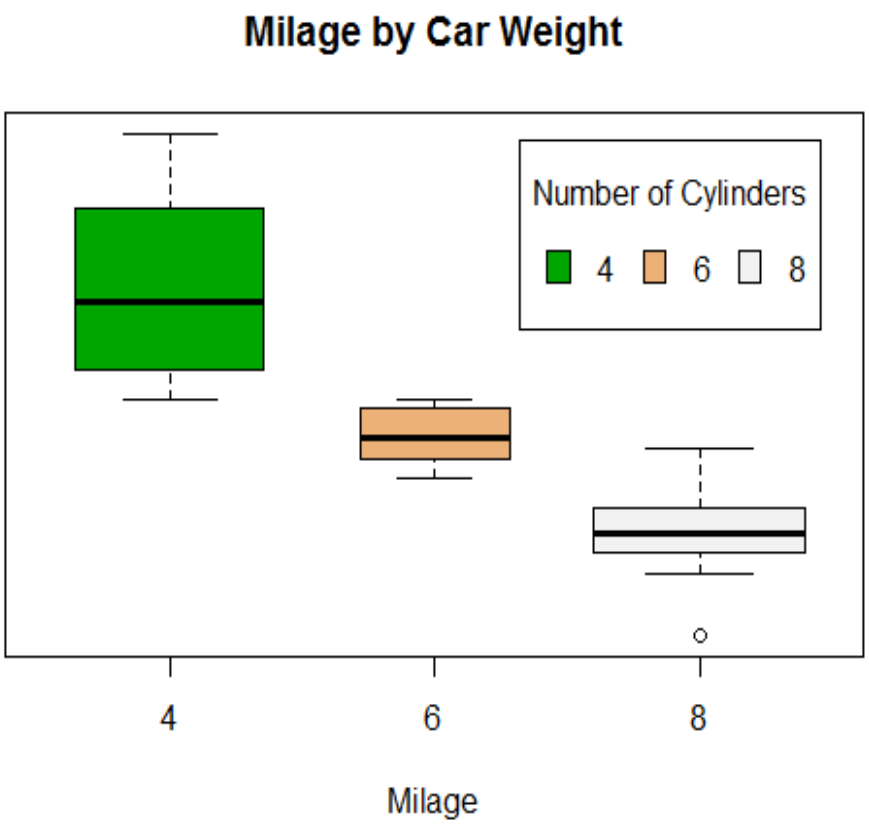
3 – «точками»

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)
```

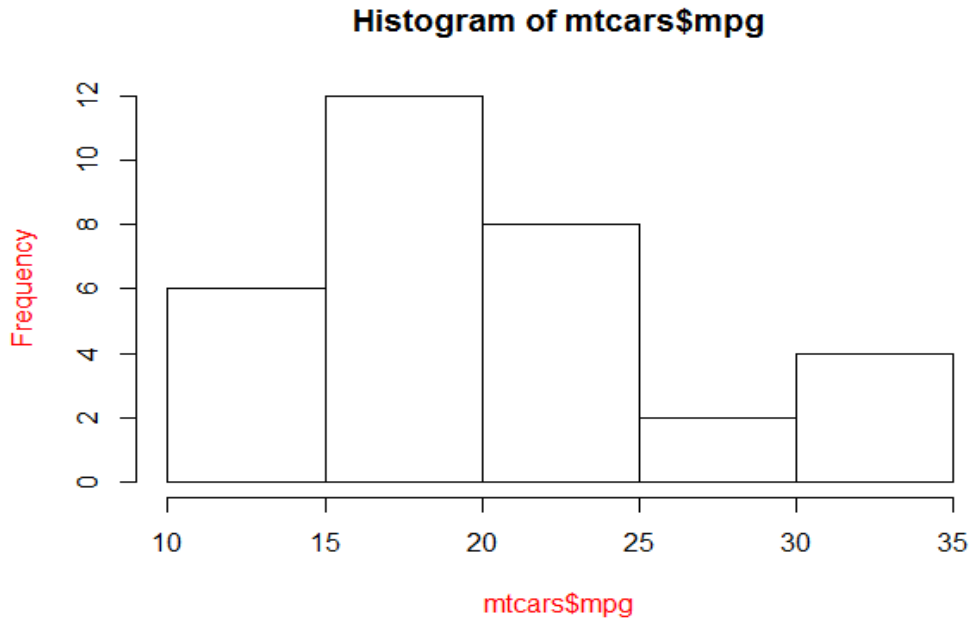
Графический параметр par()

> par() # просмотреть текущие значения графических параметров

! >old_par<-par(no.readonly = TRUE) #прежде чем менять настройки, рекомендуем сохранить старые

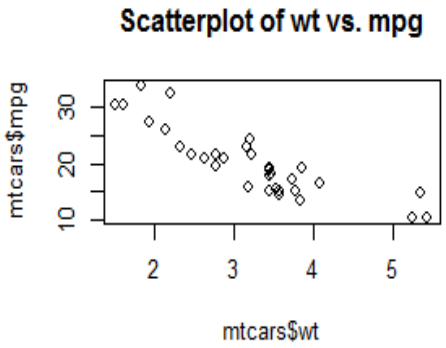
> par(col.lab="red") # сделать красными подписи к осям

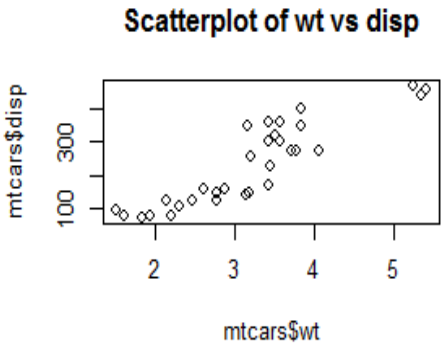
> hist(mtcars$mpg) # график рисуется с новыми настройками

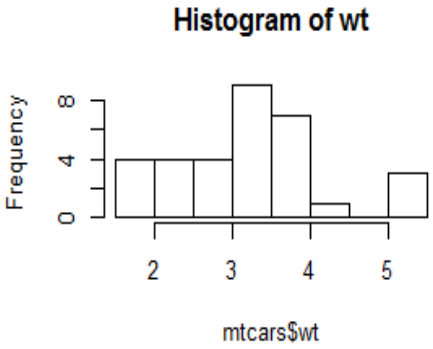


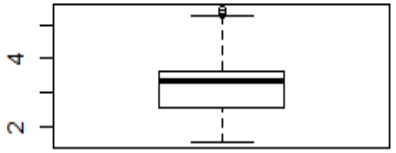
>par(old_par) #восстанавливаем старые настройки

Комбинация графиков



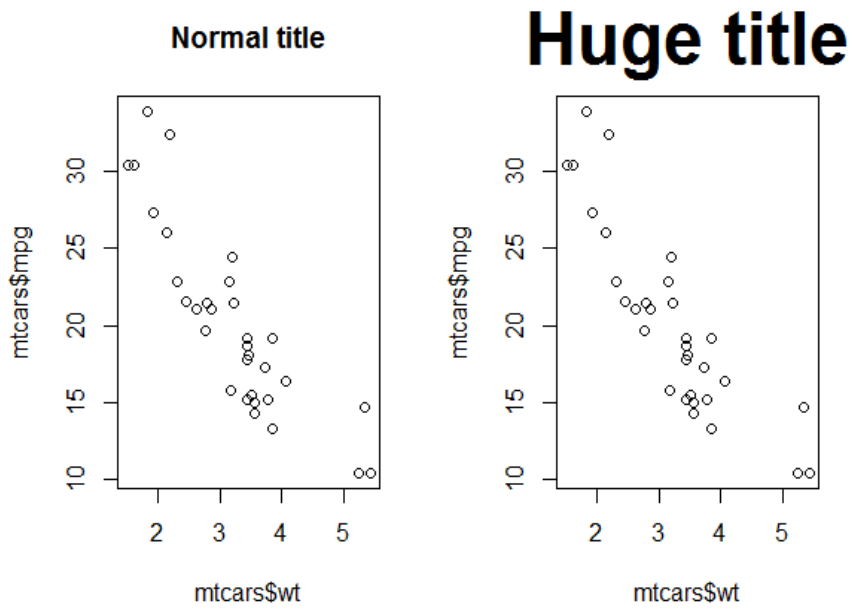







```
> par(mfrow=c(2,2))
> plot(mtcars$wt,mtcars$mpg,
main="Scatterplot of wt vs.
mpg")
> plot(mtcars$wt,mtcars$disp,
main="Scatterplot of wt vs
disp")
> hist(mtcars$wt,
main="Histogram of wt")
> boxplot(mtcars$wt,
main="Boxplot of wt")
```

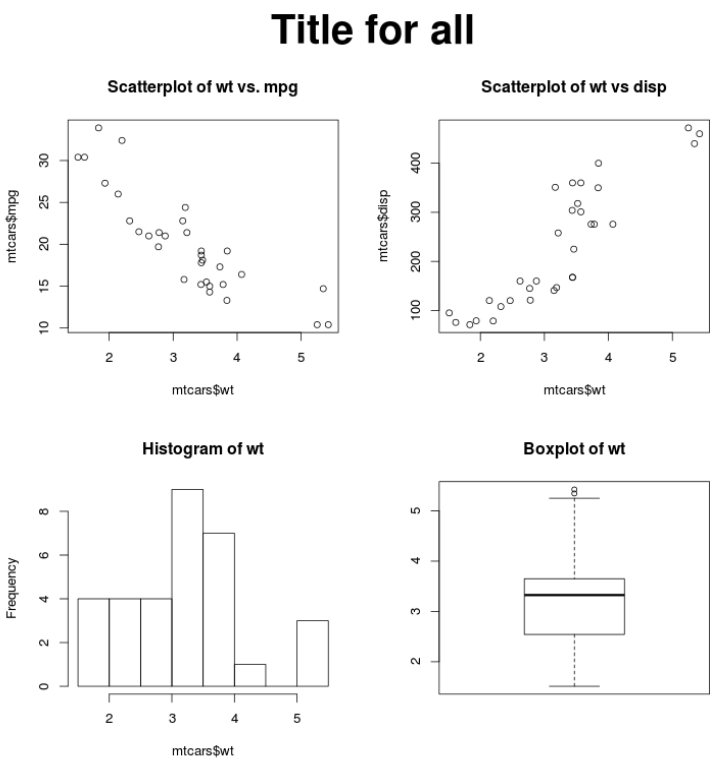
Размер текста и символов



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)
```

Общий заголовок к комбинации графиков




```
# нужен отступ сверху!
>par(mfrow=c(2,2), oma = c( 1, 1, 3, 1 ))
>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")
```

```
# рисуем заголовок вне графиков (во внешней рамке)
>title('Title for all',cex.main = 3, outer = TRUE)
```

Цвета

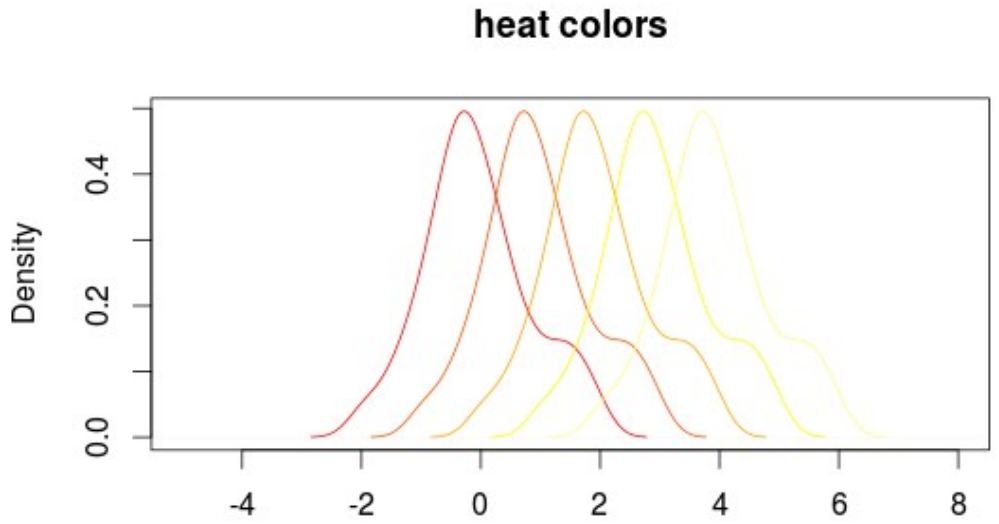
опция	описание
col	Цвет по умолчанию (может быть вектором)
col.axis	Цвет текста по осям
col.lab	Цвет подписей к осям
col.main	Цвет заголовков

Можно использовать функции выбора палитры

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

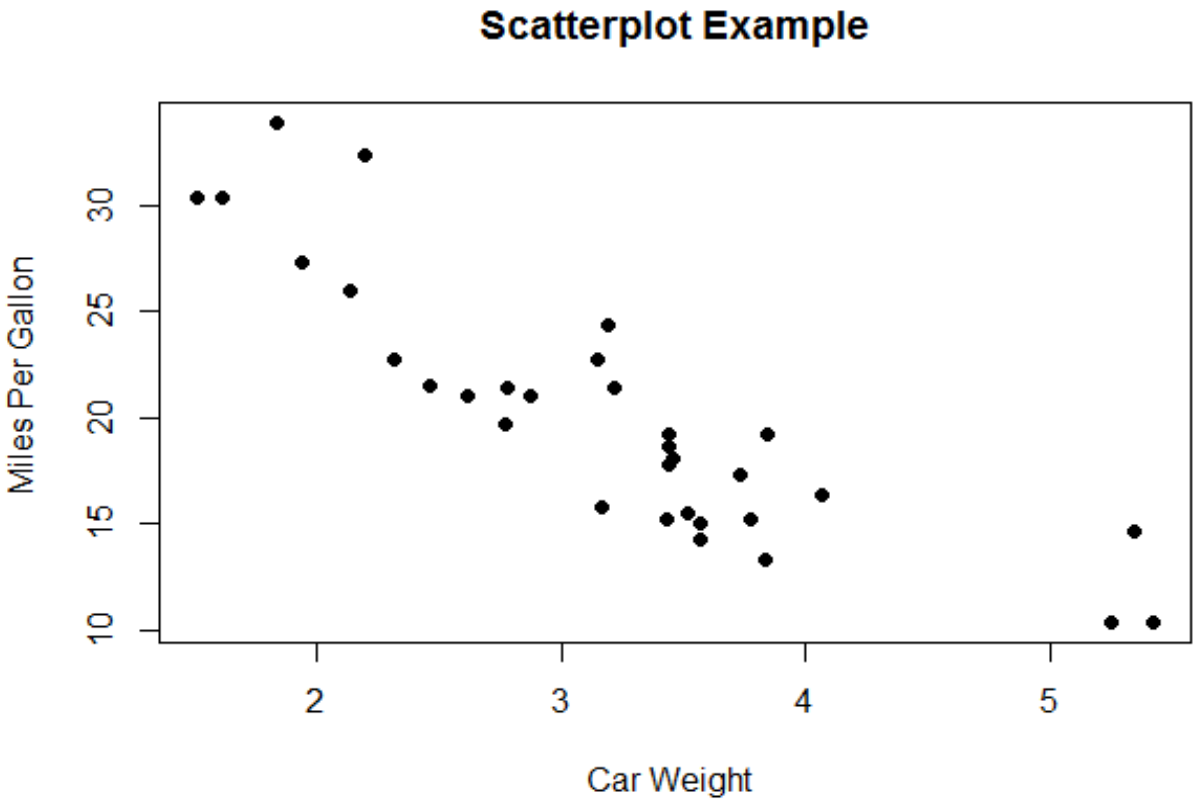
для создания вектора цветов

Цвета



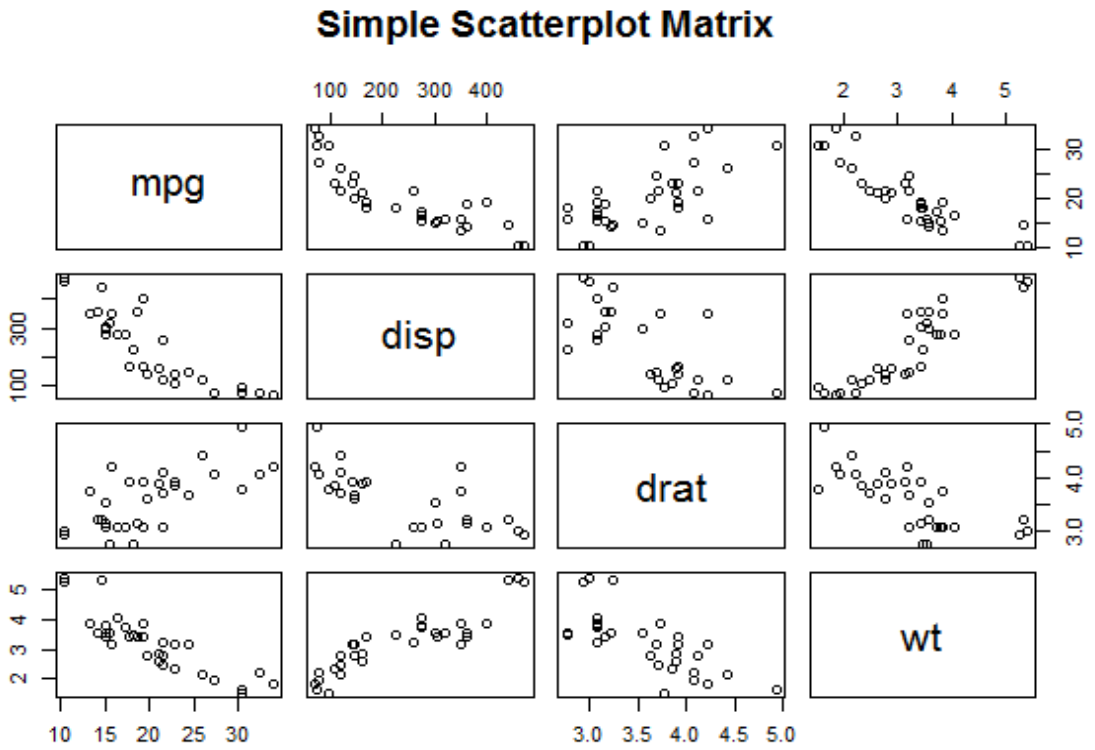

```
> x1 <- rnorm(100) ; x2 <- x1+1 ; x3 <- x2+1
> x4 <- x3+1 ; x5 <- x4+1
> ourCol <- heat.colors(5)
> plot(density(x1), col=ourCol[1], xlim=c(-5,8),
main="heat colors", xlab="")
> lines(density(x2), col=ourCol[2])
> lines(density(x3), col=ourCol[3])
> lines(density(x4), col=ourCol[4])
> lines(density(x5),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

Scatterplots




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

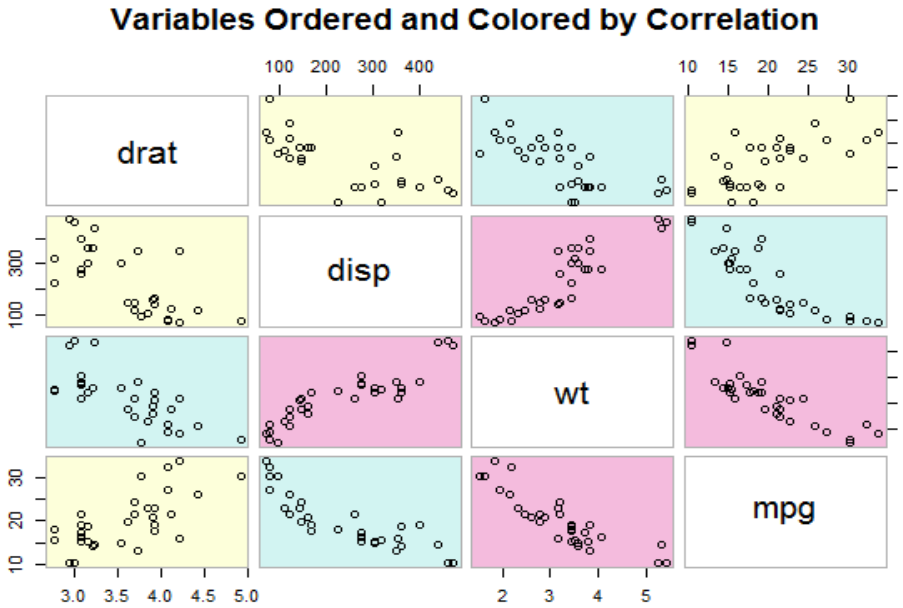
Scatterplot: матрицы

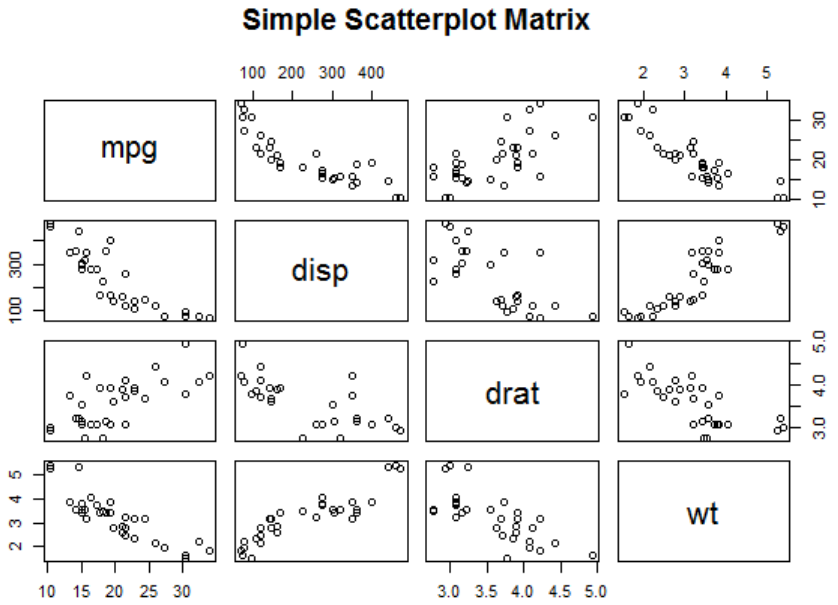


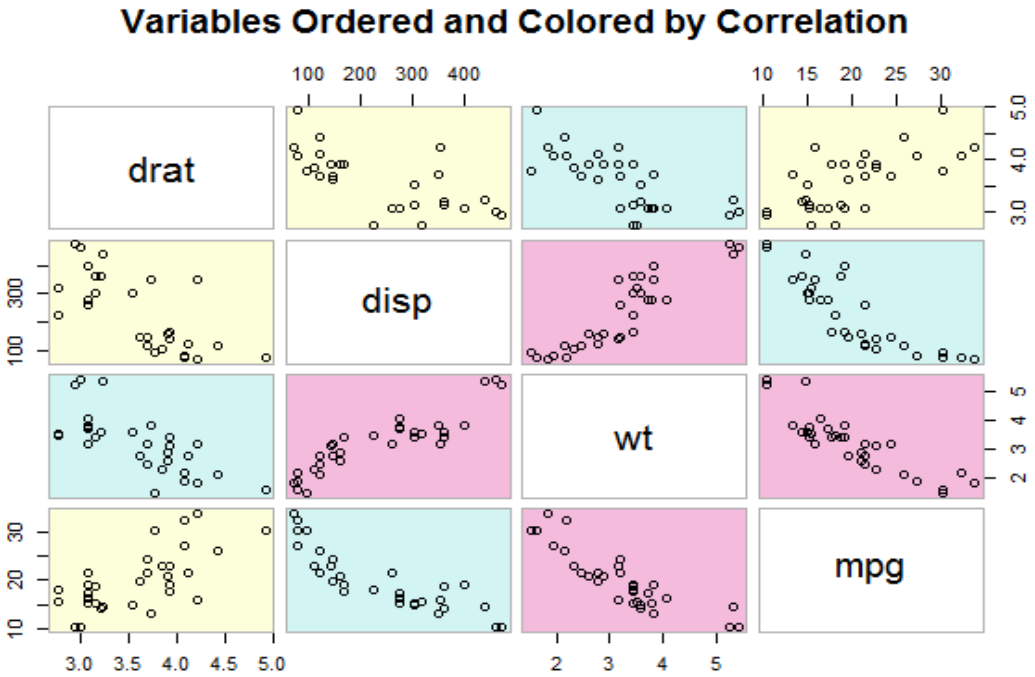

```
> pairs(mtcars[,c(1,3,5,6)], main="Simple Scatterplot Matrix")
#или то же самое:
> pairs(~mpg+disp+drat+wt,
data=mtcars, main="Simple Scatterplot Matrix")
```

Другие scatterplots

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

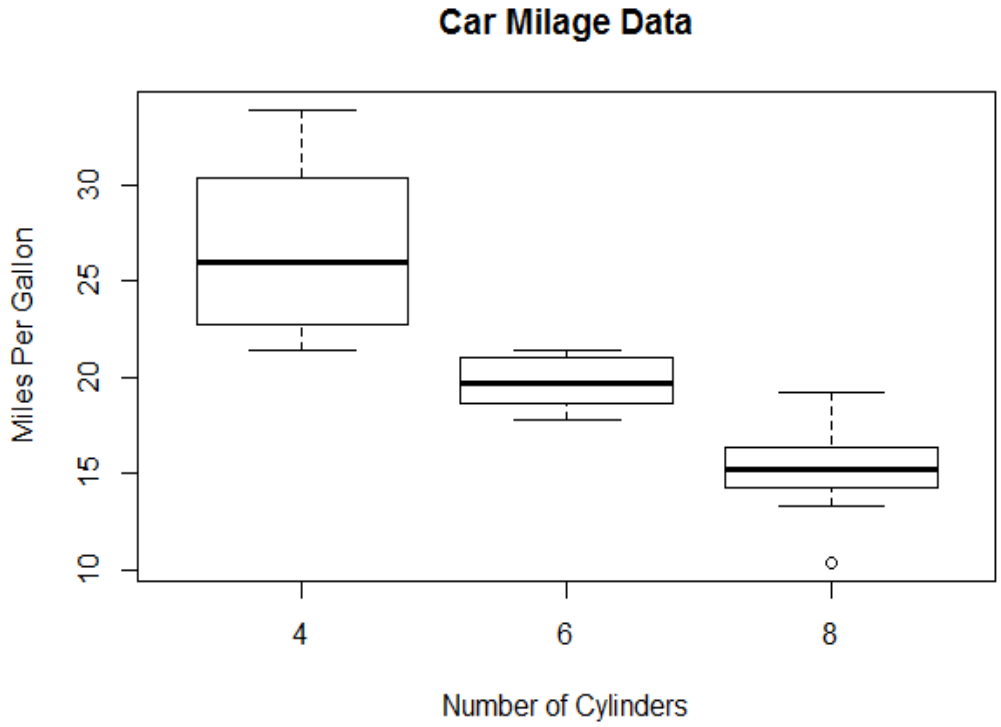







```
> library(gclus)
> dta <- mtcars[,c(1,3,5,6)]
> dta.r <- abs(cor(dta))  #запоминаем матрицу модулей корреляций
> dta.col <- dmat.color(dta.r)  #строим матрицу цветов по матрице корреляций
> dta.o <- order.single(dta.r) #новый порядок объектов, более скоррелированные идут подряд
> cpairs(dta, dta.o, panel.colors=dta.col, gap=.5, main= "Variables Ordered and Colored by Correlation" )
```

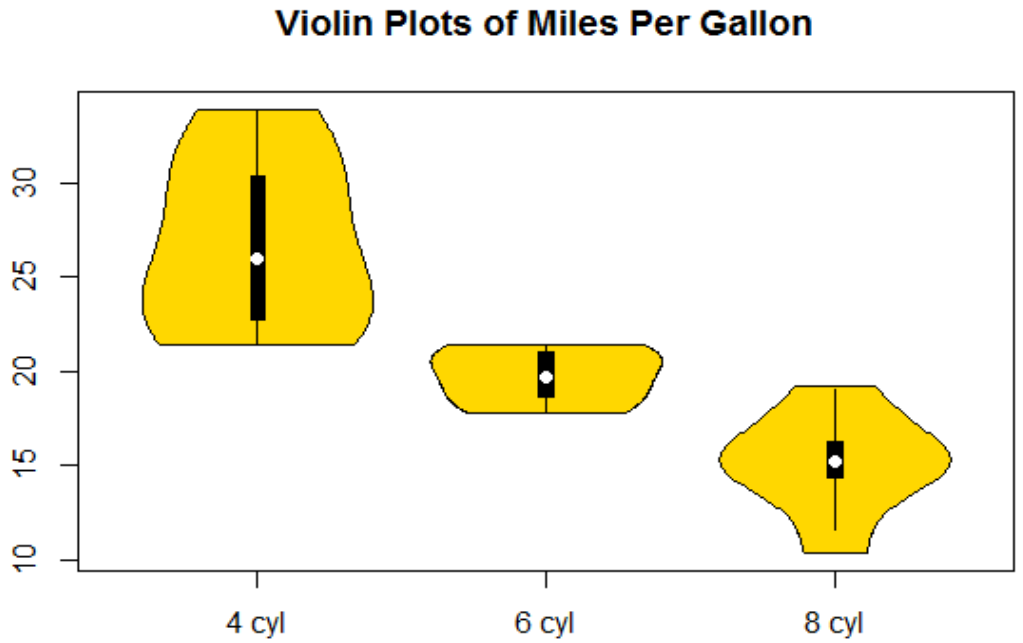
Boxplots




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

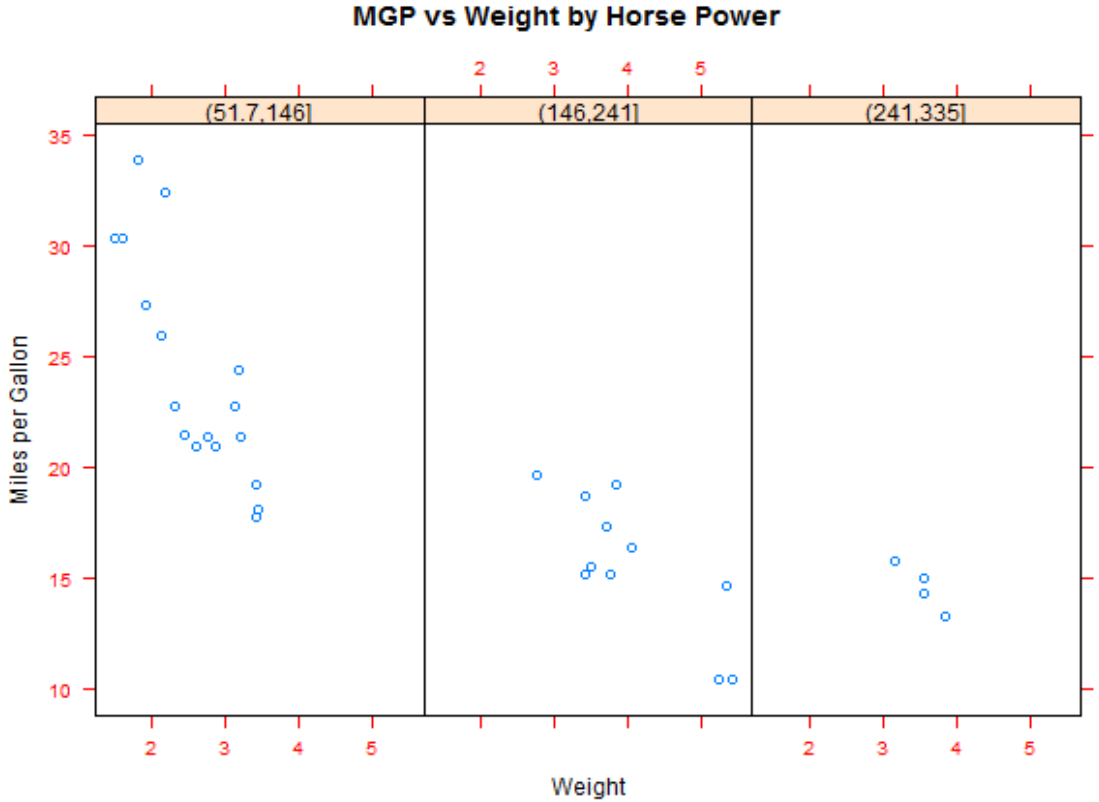
Violin Plot: комбинация boxplot и графика плотности распределения

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



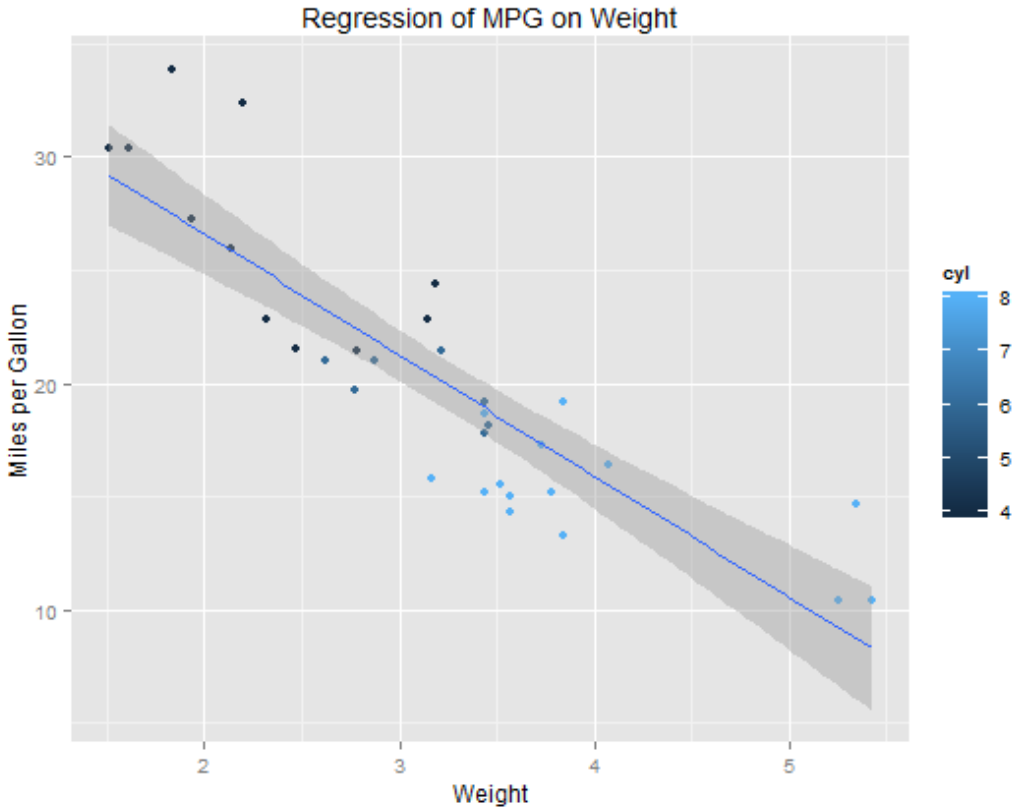

```
> library(vioplot)
> x1 <- mtcars[mtcars$cyl==4,]$mpg
> x2 <- mtcars[mtcars$cyl==6,]$mpg
> x3 <- mtcars[mtcars$cyl==8,]$mpg
> vioplot(x1, x2, x3, names=c("4 cyl", "6 cyl",
"8 cyl"), col="gold")
title("Violin Plots of Miles Per Gallon")
```

Возможности lattice




```
> library(lattice)
> hp <- cut(mtcars$hp,3) # разбиваем график на три в зависимости от количества лошадиных сил машин
> xyplot(mtcars$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")
```

Что еще можно добавить на график

`grid(nx, ny)`	Добавить сетку
`axis(side n,)`	Добавить ось к графику
`box(which=, )`	Добавить рамку вокруг графика, картинки и т.п.
`legend`	Добавить легенду

`arrows(x,y)`
`lines(x, y)`
`points(x,y)`
	Добавить стрелки, линии, точки

`abline(a,b)` `abline(h= or v=)`	Добавить прямую с заданным углом наклона и смещением, либо вертикальную или горизонтальную
`segments(x0,x1,y0, y1)`	Добавить отрезки мезжу точками
`polygon(x,y)`	Многоугольник по точкам
`text(x,y, "note")`	Текст на графике в заданной точке

Больше графиков по ссылкам

- http://www.statmethods.net/advgraphs/
- http://www.sr.bham.ac.uk/~ajrs/R/r-gallery.html

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

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

> which(c(FALSE, TRUE, FALSE, TRUE)) #как работает команда which
[1] 2 4

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

> mtcarsA <- na.omit(mtcarsNew)   #или просто уберем все строки,
                                  содержащие NA

> dim(mtcarsNew)
[1] 33 11
#проверим, изменилось ли число строк
> dim(mtcarsA)
[1] 32 11
```