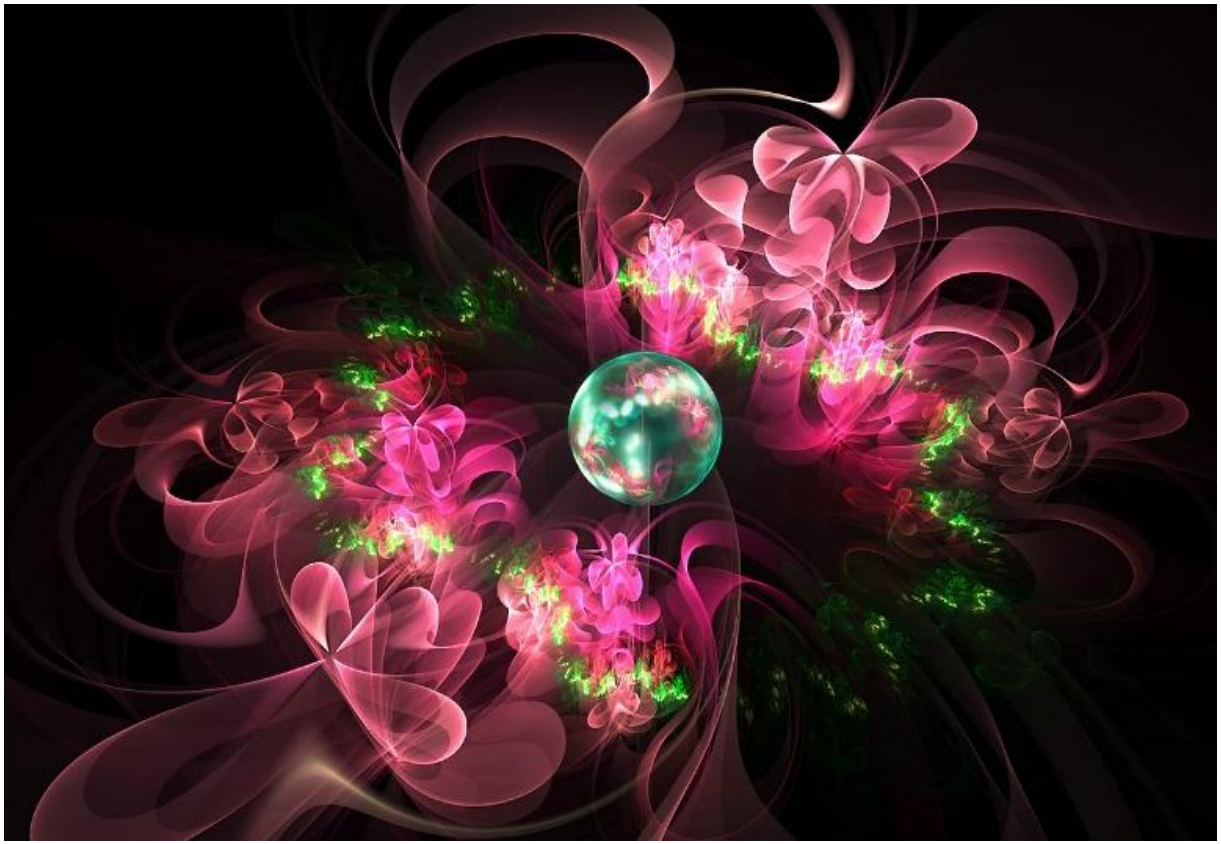


R数据可视化—gplot2包 第2周

2013.2.8

法律声明

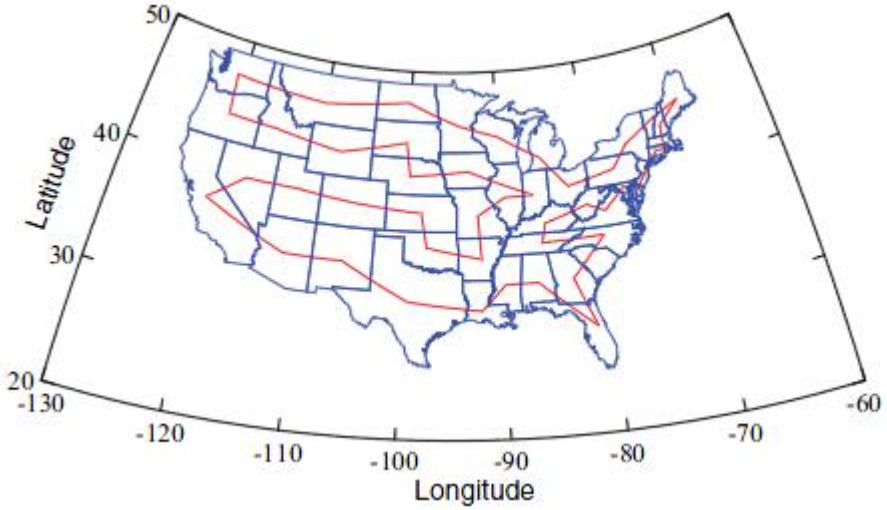
【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

主要内容

- 理解GG的主要概念
- 使用图层的想法进行绘图



Fuel economy数据集

manufacturer	model	disp	year	cyl	cty	hwy	class
audi	a4	1.8	1999	4	18	29	compact
audi	a4	1.8	1999	4	21	29	compact
audi	a4	2.0	2008	4	20	31	compact
audi	a4	2.0	2008	4	21	30	compact
audi	a4	2.8	1999	6	16	26	compact
audi	a4	2.8	1999	6	18	26	compact
audi	a4	3.1	2008	6	18	27	compact
audi	a4 quattro	1.8	1999	4	18	26	compact
audi	a4 quattro	1.8	1999	4	16	25	compact
audi	a4 quattro	2.0	2008	4	20	28	compact

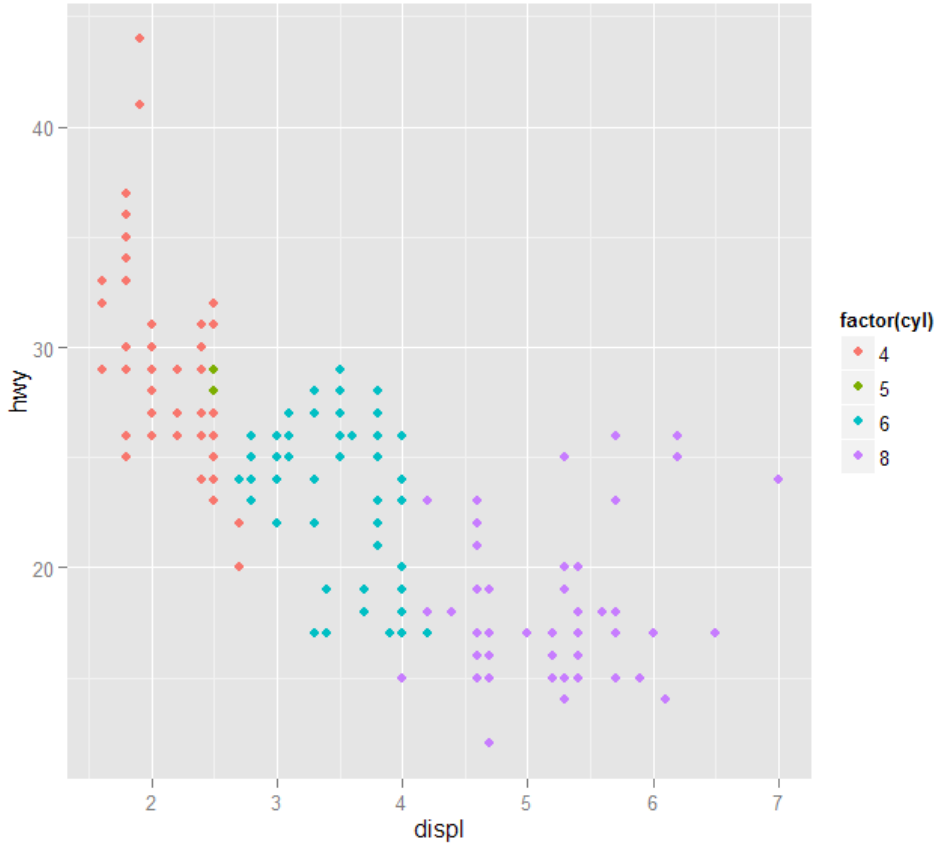
Table 3.1: The first 10 cars in the `mpg` dataset, included in the `ggplot2` package. `cty` and `hwy` record miles per gallon (mpg) for city and highway driving, respectively, and `displ` is the engine displacement in litres.

Fuel economy数据集

```
> mpg
    manufacturer                model displ year cyl      trans drv cty hwy fl      class
1           audi                   a4   1.8 1999   4   auto(l5)   f  18  29  p    compact
2           audi                   a4   1.8 1999   4 manual(m5)   f  21  29  p    compact
3           audi                   a4   2.0 2008   4 manual(m6)   f  20  31  p    compact
4           audi                   a4   2.0 2008   4   auto(av)   f  21  30  p    compact
5           audi                   a4   2.8 1999   6   auto(l5)   f  16  26  p    compact
6           audi                   a4   2.8 1999   6 manual(m5)   f  18  26  p    compact
7           audi                   a4   3.1 2008   6   auto(av)   f  18  27  p    compact
8           audi              a4 quattro   1.8 1999   4 manual(m5)   4  18  26  p    compact
9           audi              a4 quattro   1.8 1999   4   auto(l5)   4  16  25  p    compact
10          audi              a4 quattro   2.0 2008   4 manual(m6)   4  20  28  p    compact
11          audi              a4 quattro   2.0 2008   4   auto(s6)   4  19  27  p    compact
12          audi              a4 quattro   2.8 1999   6   auto(l5)   4  15  25  p    compact
13          audi              a4 quattro   2.8 1999   6 manual(m5)   4  17  25  p    compact
14          audi              a4 quattro   3.1 2008   6   auto(s6)   4  17  25  p    compact
15          audi              a4 quattro   3.1 2008   6 manual(m6)   4  15  25  p    compact
16          audi              a6 quattro   2.8 1999   6   auto(l5)   4  15  24  p    midsize
17          audi              a6 quattro   3.1 2008   6   auto(s6)   4  17  25  p    midsize
18          audi              a6 quattro   4.2 2008   8   auto(s6)   4  16  23  p    midsize
19     chevrolet     c1500 suburban 2wd   5.3 2008   8   auto(l4)   r  14  20  r        suv
20     chevrolet     c1500 suburban 2wd   5.3 2008   8   auto(l4)   r  11  15  e        suv
21     chevrolet     c1500 suburban 2wd   5.3 2008   8   auto(l4)   r  14  20  r        suv
22     chevrolet     c1500 suburban 2wd   5.7 1999   8   auto(l4)   r  13  17  r        suv
23     chevrolet     c1500 suburban 2wd   6.0 2008   8   auto(l4)   r  12  17  r        suv
24     chevrolet               corvette   5.7 1999   8 manual(m6)   r  16  26  p    2seater
25     chevrolet               corvette   5.7 1999   8   auto(l4)   r  15  23  p    2seater
```

散点图

qplot(displ, hwy, data = mpg, colour = factor(cyl))



2013.2.8

上图中的装饰属性

- 点
- 点的位置
- 点的大小
- 点的颜色

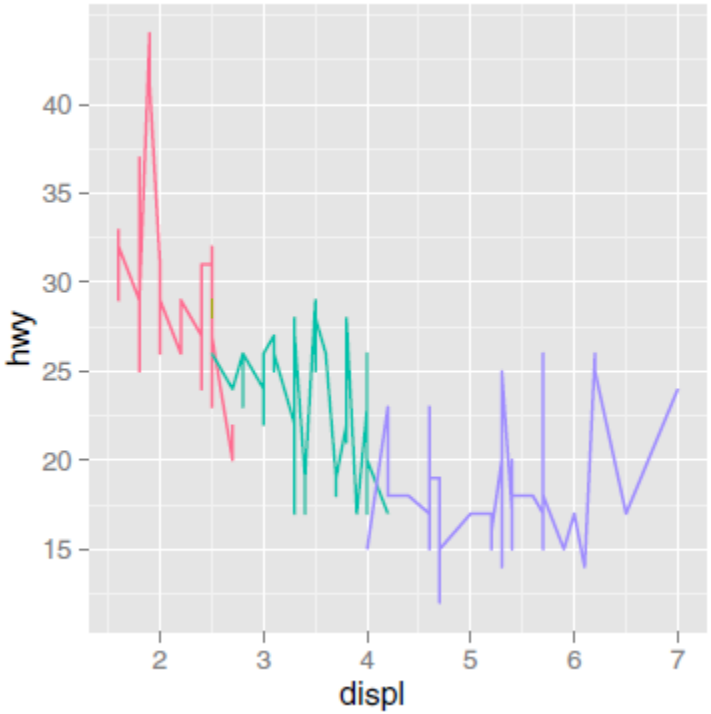
数据到装饰属性的映射（Mapping）

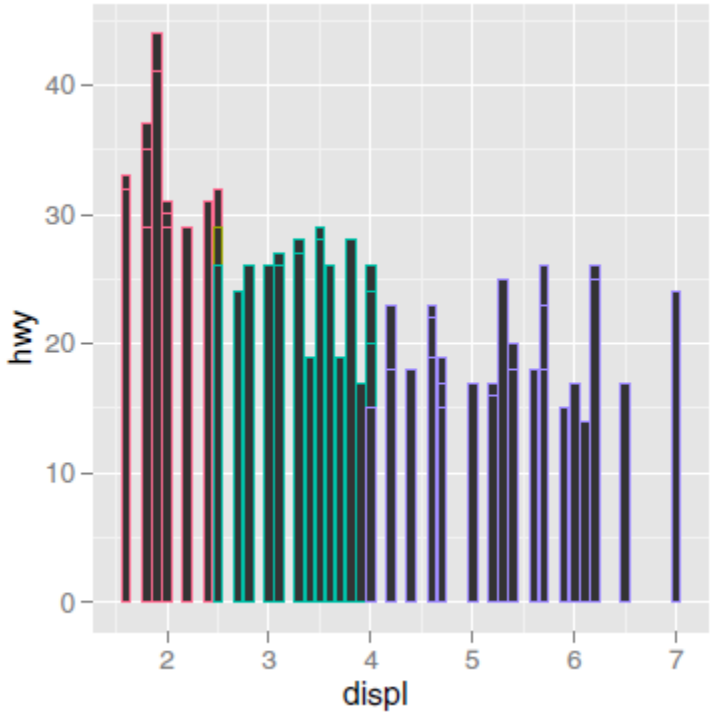
- Disp映射到x坐标，hwy映射到y坐标，cyl映射到颜色

manufacturer	model	disp	year	cyl	cty	hwy	class
audi	a4	1.8	1999	4	18	29	compact
audi	a4	1.8	1999	4	21	29	compact
audi	a4	2.0	2008	4	20	31	compact
audi	a4	2.0	2008	4	21	30	compact
audi	a4	2.8	1999	6	16	26	compact
audi	a4	2.8	1999	6	18	26	compact
audi	a4	3.1	2008	6	18	27	compact
audi	a4 quattro	1.8	1999	4	18	26	compact
audi	a4 quattro	1.8	1999	4	16	25	compact
audi	a4 quattro	2.0	2008	4	20	28	compact

x	y	colour
1.8	29	4
1.8	29	4
2.0	31	4
2.0	30	4
2.8	26	6
2.8	26	6
3.1	27	6
1.8	26	4
1.8	25	4
2.0	28	4

其它的映射方法



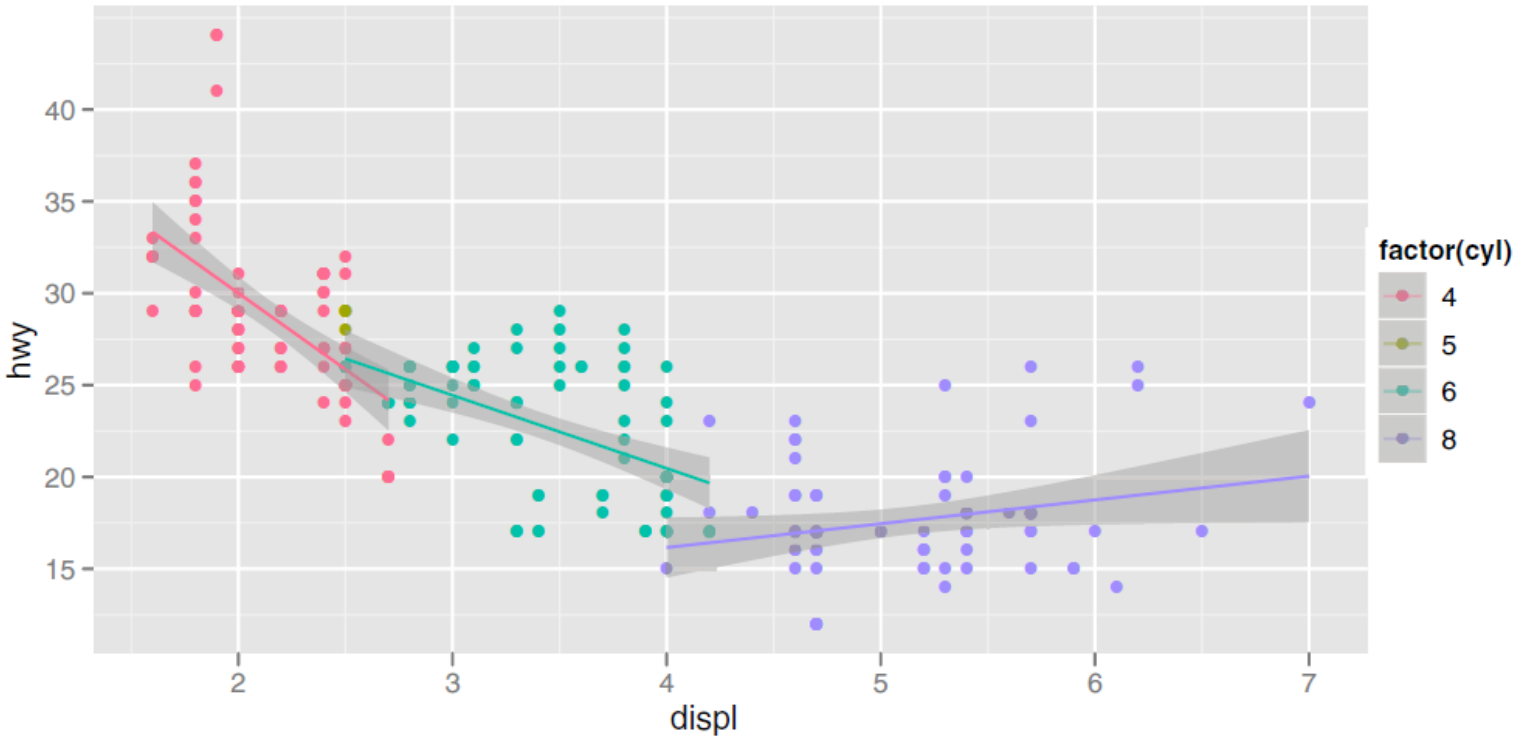


几何对象（geom）决定统计图的类型

Named plot	Geom	Other features
scatterplot	point	
bubblechart	point	size mapped to a variable
barchart	bar	
box-and-whisker plot	boxplot	
line chart	line	

难以命名的统计图

- 一般由基本统计图组合而成



Scaling

- 把数据从其计量单位（例如油耗的升数，里程等）转化为计算机能识别的显示要素（例如像素，颜色等）的过程，称为Scaling
- 在右图中有几项scaling

1) 将水平坐标x映射到[0,1]区间。这里不使用具体像素值的原因是grid包替我们完成最终的转换
2) 将垂直坐标y映射到[0,1]区间
3) 由坐标系统(coord)根据x,y的组合最终定位，常见的坐标系统包括直角坐标系，极坐标系，球面映射等
4) 颜色的scaling

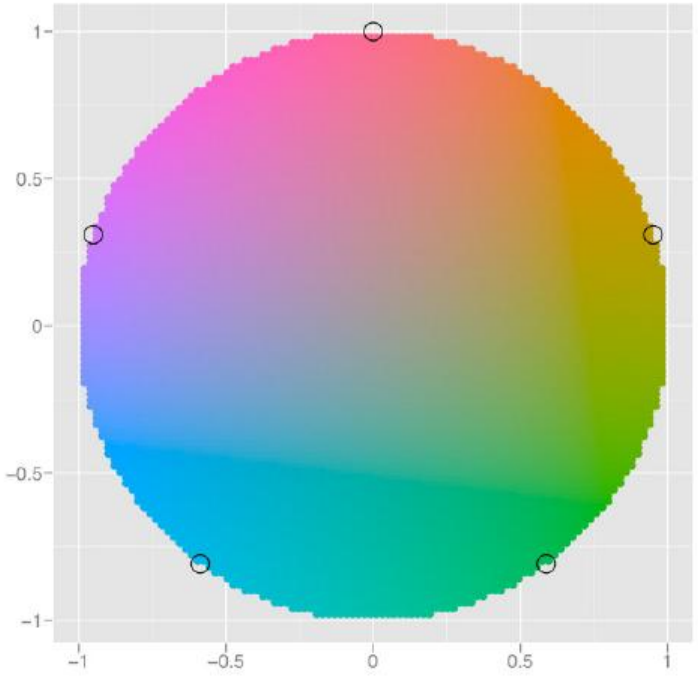
x	y	colour
1.8	29	4
1.8	29	4
2.0	31	4
2.0	30	4
2.8	26	6
2.8	26	6
3.1	27	6
1.8	26	4
1.8	25	4
2.0	28	4

Scaling

x	y	colour	size	shape
0.037	0.531	#FF6C91	1	19
0.037	0.531	#FF6C91	1	19
0.074	0.594	#FF6C91	1	19
0.074	0.562	#FF6C91	1	19
0.222	0.438	#00C1A9	1	19
0.222	0.438	#00C1A9	1	19
0.278	0.469	#00C1A9	1	19
0.037	0.438	#FF6C91	1	19
0.037	0.406	#FF6C91	1	19
0.074	0.500	#FF6C91	1	19

离散值的颜色scale

- 5个离散值时的缺省算法示意图



Scales画出的内容

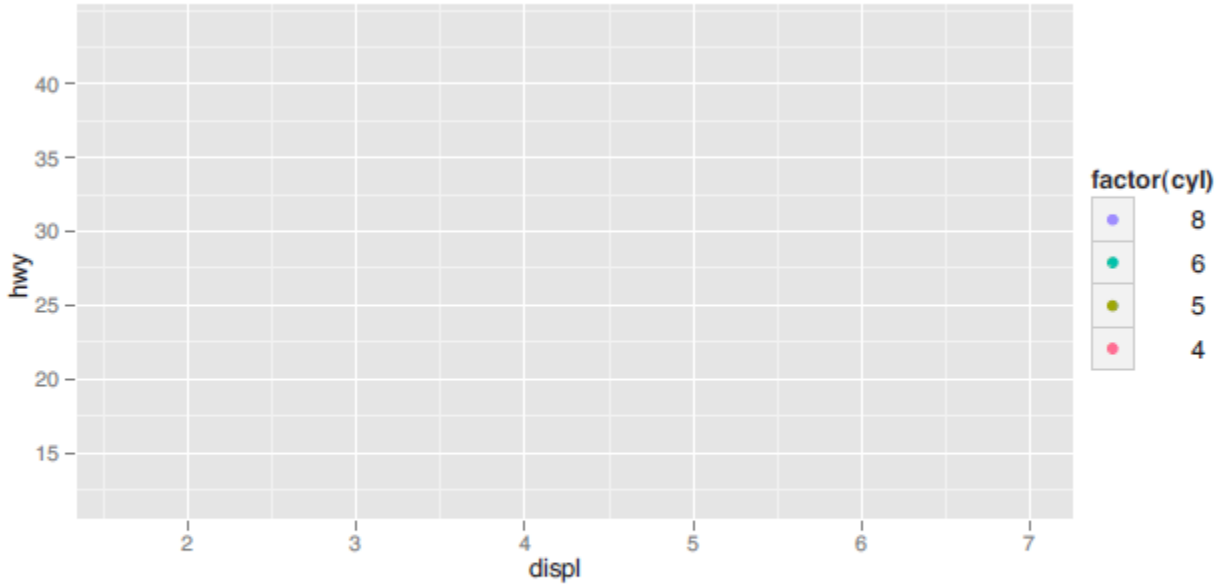
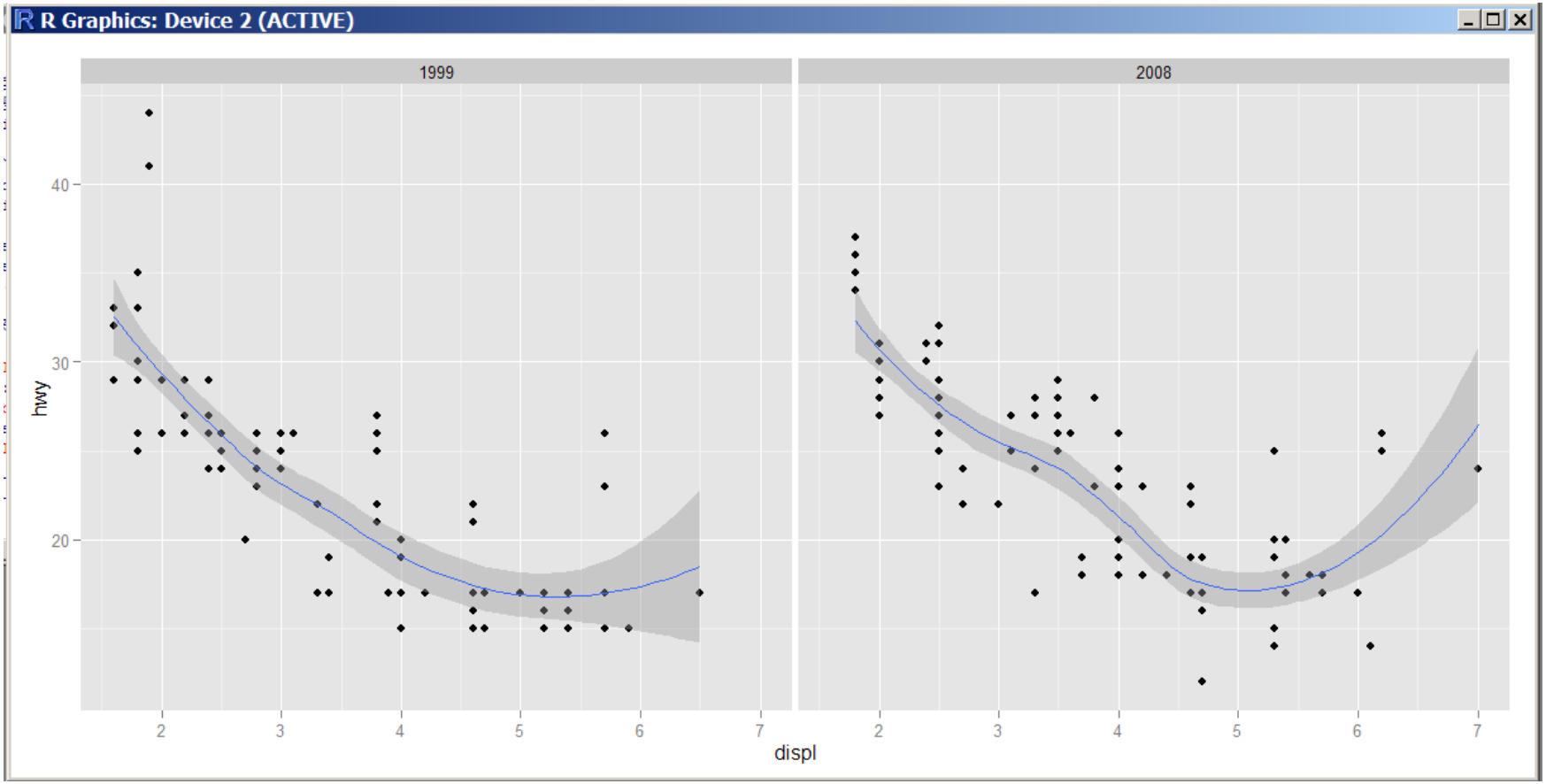


Fig. 3.5: Contributions from the scales, the axes and legend and grid lines, and the plot background. Contributions from the data, the point geom, have been removed.

更复杂的例子

qplot(displ, hwy, data=mpg, facets = . ~ year) + geom_smooth()



2013.2.8

解释

- Layers
- Facets

总结：基于图层概念的绘图过程

Layer的组成

1 数据到装饰属性的映射

2 统计变换

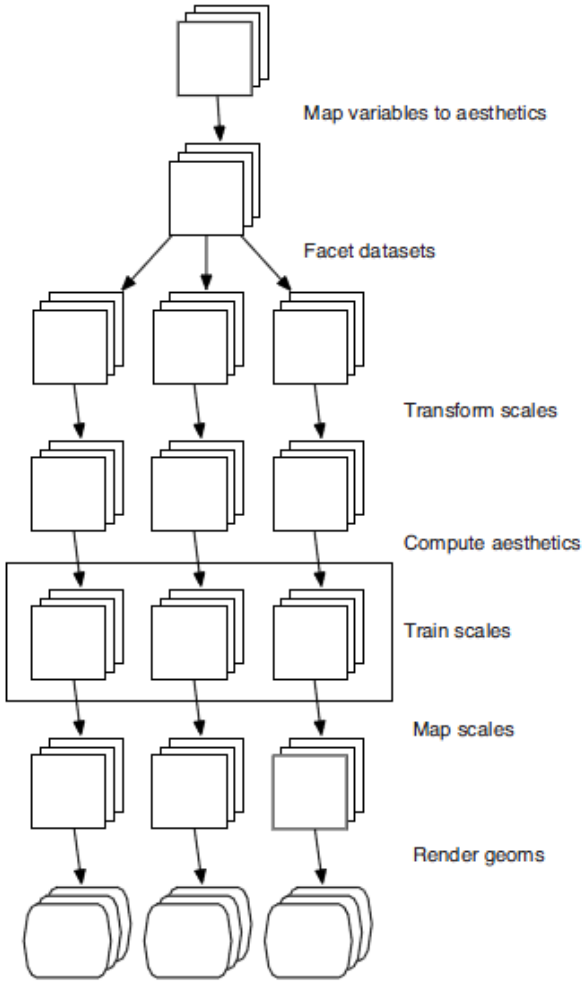
3 几何对象

4 位置变换

Scale

Coord

Faceting



2013.2.8

ggplot函数

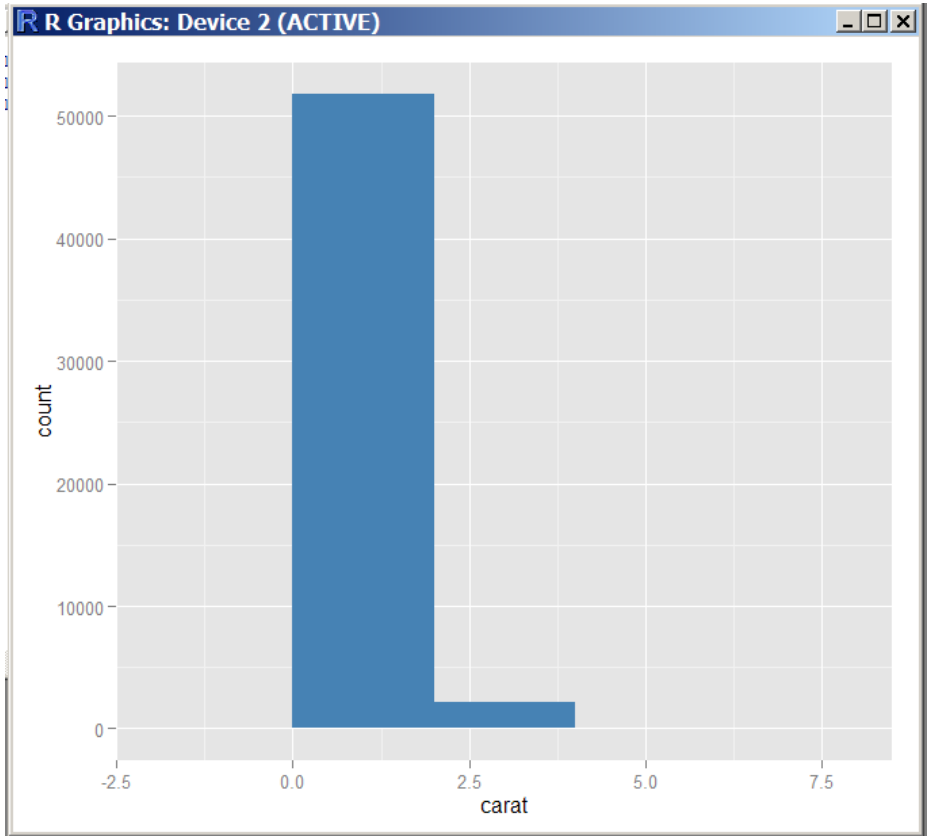
p <- ggplot(diamonds, aes(carat, price, colour = cut))

```
> qplot(displ, hwy, data = mpg, colour = factor(cyl))
>
>
> p <- ggplot(diamonds, aes(carat, price, colour = cut))
> summary(p)
data: carat, cut, color, clarity, depth, table, price, x, y, z
  [53940x10]
mapping:  x = carat, y = price, colour = cut
faceting: facet_null()
>
```

图层（layer）

- layer函数: layer(geom, geom_params, stat, stat_params, data, mapping,position)

```
p <- ggplot(diamonds, aes(x = carat))
p <- p + layer(
geom = "bar",
geom_params = list(fill = "steelblue"),
stat = "bin",
stat_params = list(binwidth = 2)
)
p
```



2013.2.8

捷径函数

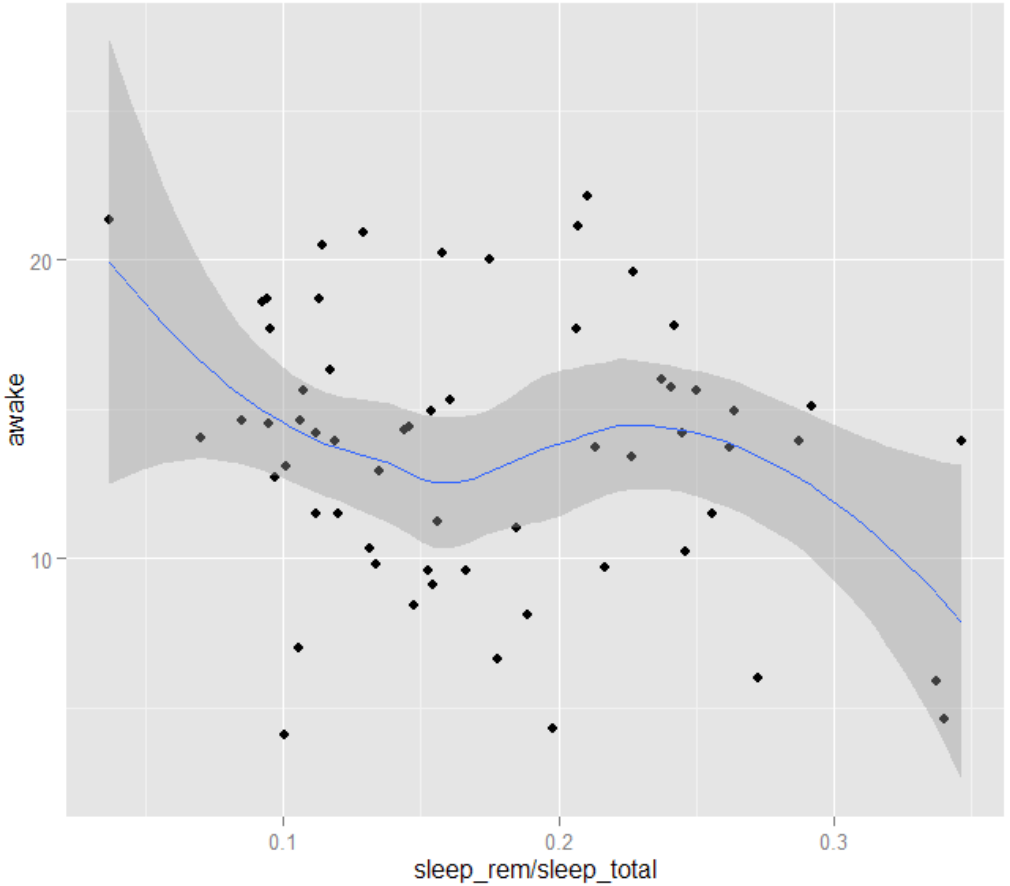
- 样例： geom_histogram(binwidth = 2, fill = "steelblue")
- 函数名一般以“geom_”或“stat_”开头
- 参数： geom_XXX(mapping, data, ..., geom, position)或stat_XXX(mapping, data, ..., stat, position)

参数

- mapping：可选，指出到装饰属性的映射，通常使用aes( )函数
- data：可选，指出数据集，将覆盖ggplot函数中所指定的缺省数据集
- geom或stat：可选，可以用于覆盖指定geom的缺省stat，或覆盖指定stat的缺省geom
- position：可选，用于指出调整重叠对象的方法

例子

```
ggplot(msleep, aes(sleep_rem / sleep_total,
    awake)) +geom_point()
# which is equivalent to
qplot(sleep_rem / sleep_total, awake, data =
    msleep)
# You can add layers to qplot too:
qplot(sleep_rem / sleep_total, awake, data =
    msleep) +geom_smooth()
# This is equivalent to
qplot(sleep_rem / sleep_total, awake, data =
    msleep,geom = c("point", "smooth"))
# or
ggplot(msleep, aes(sleep_rem / sleep_total,
    awake)) +geom_point() +
    geom_smooth()
```



把作图对象存储到变量

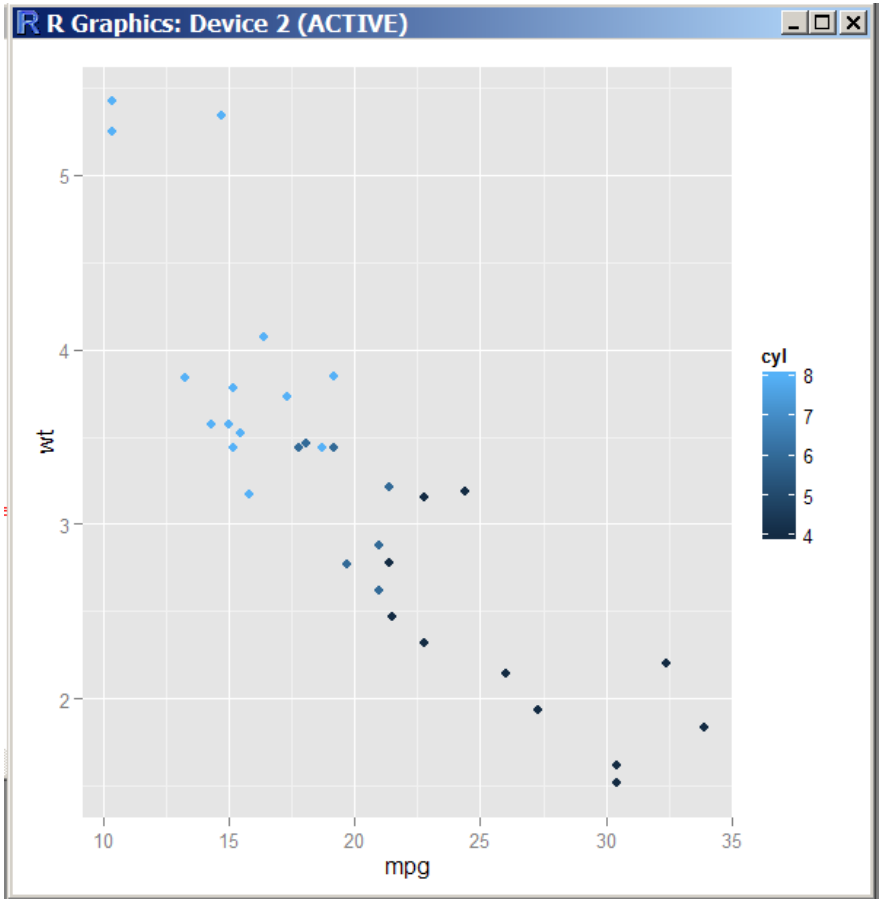
```
> library(ggplot2)
Use suppressPackageStartupMessages to eliminate package startup messages.
> p <- ggplot(msleep, aes(sleep_rem / sleep_total, awake))
> summary(p)
data: name, genus, vore, order, conservation, sleep_total, sleep_rem, sleep_cycle, awake,
  brainwt, bodywt [83x11]
mapping:  x = sleep_rem/sleep_total, y = awake
faceting: facet_null()
> p <- p + geom_point()
> summary(p)
data: name, genus, vore, order, conservation, sleep_total, sleep_rem, sleep_cycle, awake,
  brainwt, bodywt [83x11]
mapping:  x = sleep_rem/sleep_total, y = awake
faceting: facet_null()
-----------------------------------
geom_point: na.rm = FALSE
stat_identity:
position_identity: (width = NULL, height = NULL)

>
```

把作图对象存储到变量

```
bestfit <- geom_smooth(method = "lm", se = F,
  colour = alpha("steelblue", 0.5), size = 2)
qplot(sleep_rem, sleep_total, data = msleep) + bestfit
qplot(awake, brainwt, data = msleep, log = "y") + bestfit
qplot(bodywt, brainwt, data = msleep, log = "xy") + bestfit
```

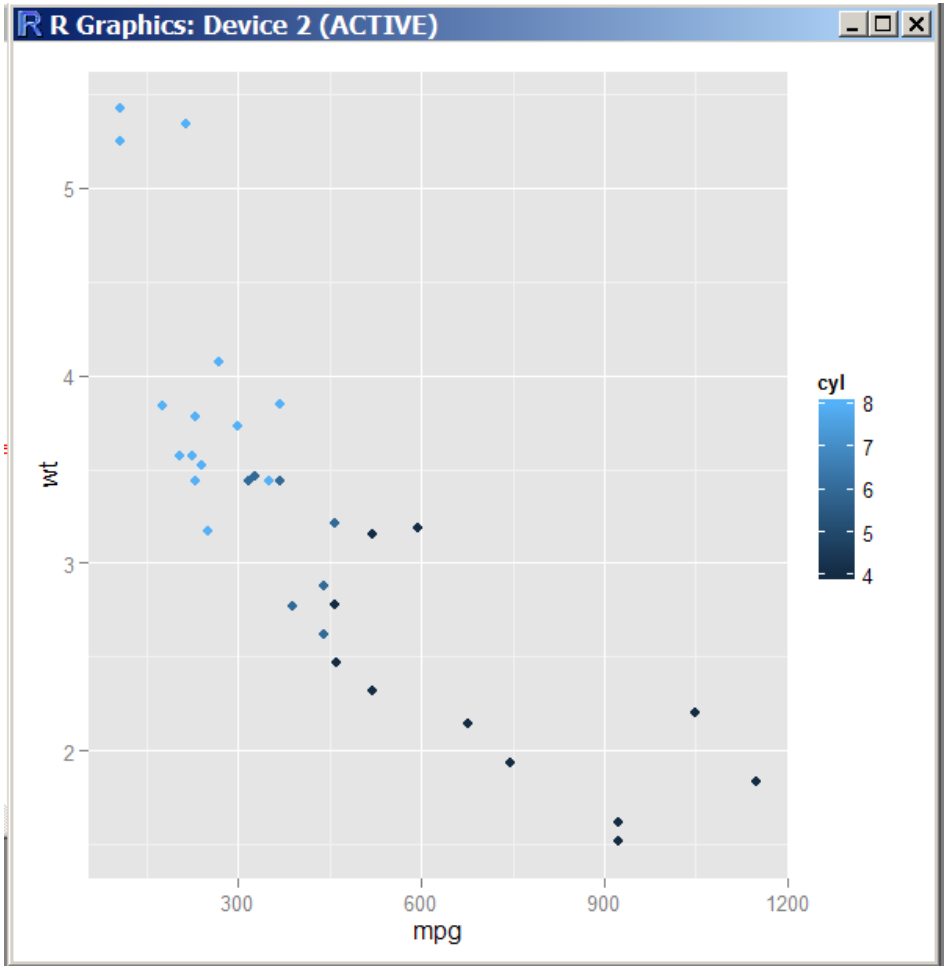
Data

```
p <- ggplot(mtcars, aes(mpg, wt,
    colour = cyl)) + geom_point()
p
```



Data

```
mtcars <- transform(mtcars, mpg =
    mpg ^ 2)
p %+% mtcars
```



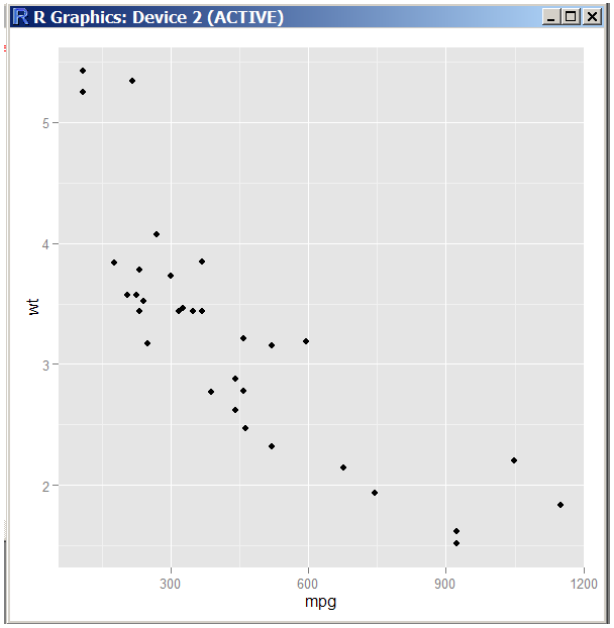
2013.2.8

Aesthetic mappings

- aes(x = weight, y = height, colour = age)
- aes(weight, height, colour = sqrt(age))

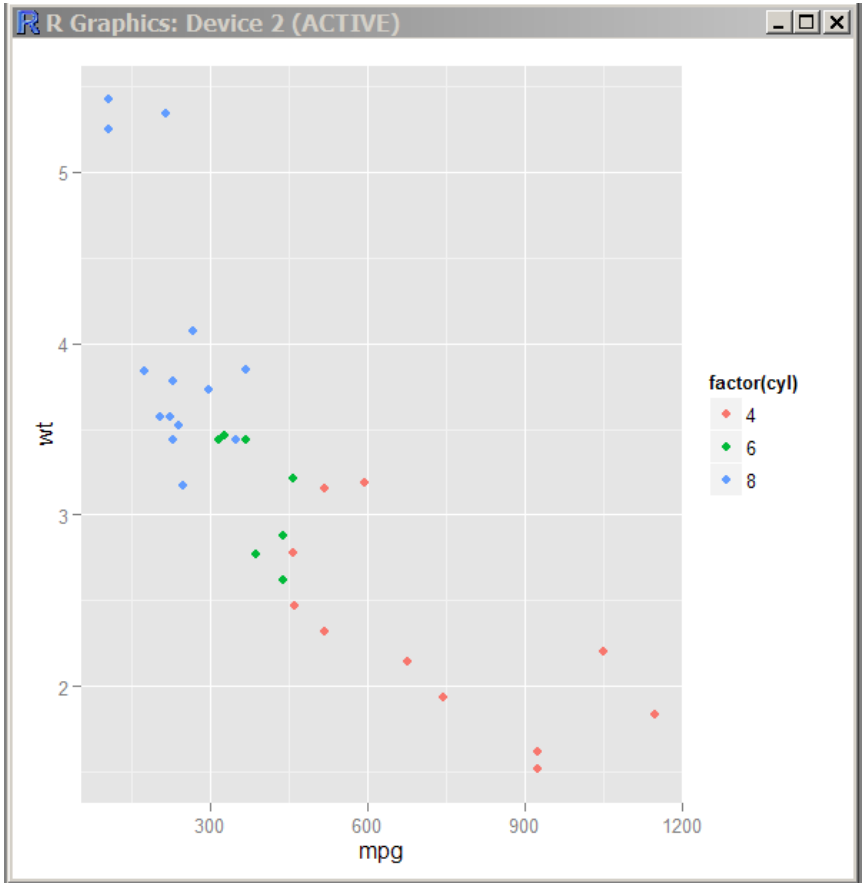
Plots and layers

```
>
> p <- ggplot(mtcars)
> summary(p)
data: mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb [32x11]
faceting: facet_null()
> p <- p + aes(wt, hp)
> summary(p)
data: mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb [32x11]
mapping:  x = wt, y = hp
faceting: facet_null()
> p <- ggplot(mtcars, aes(x = mpg, y = wt))
> p + geom_point()
>
```



续

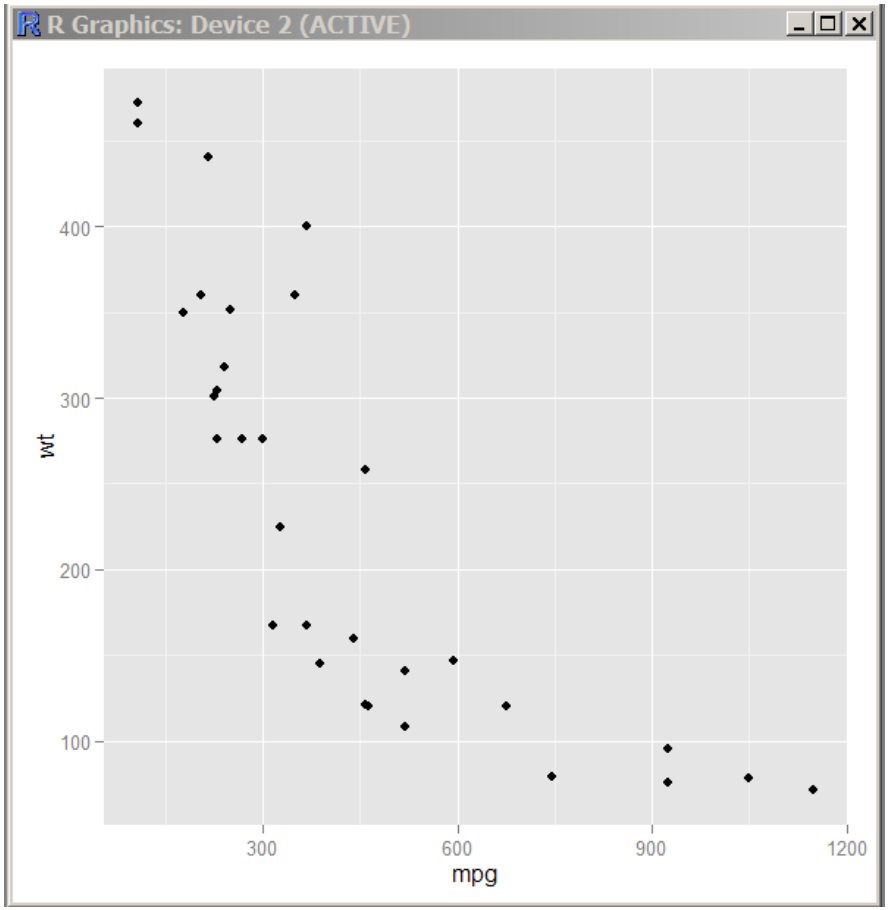
p + geom_point(aes(colour = factor(cyl)))



2013.2.8

续

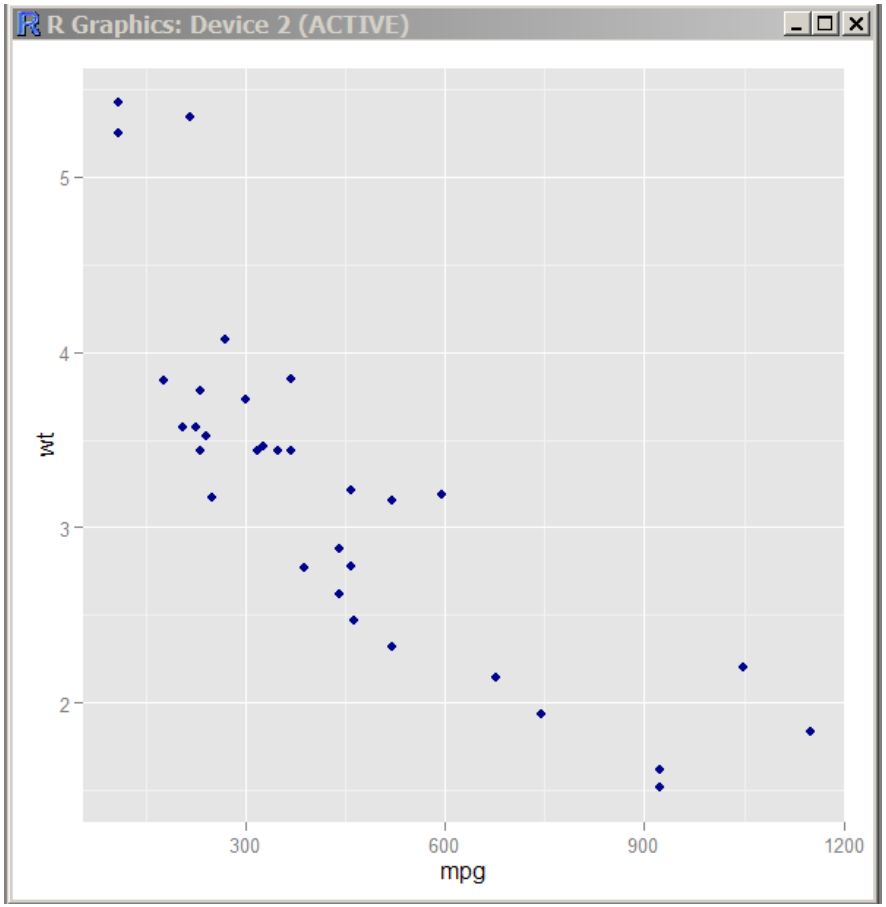
p + geom_point(aes(y = disp))



Sets与Maps

Sets:

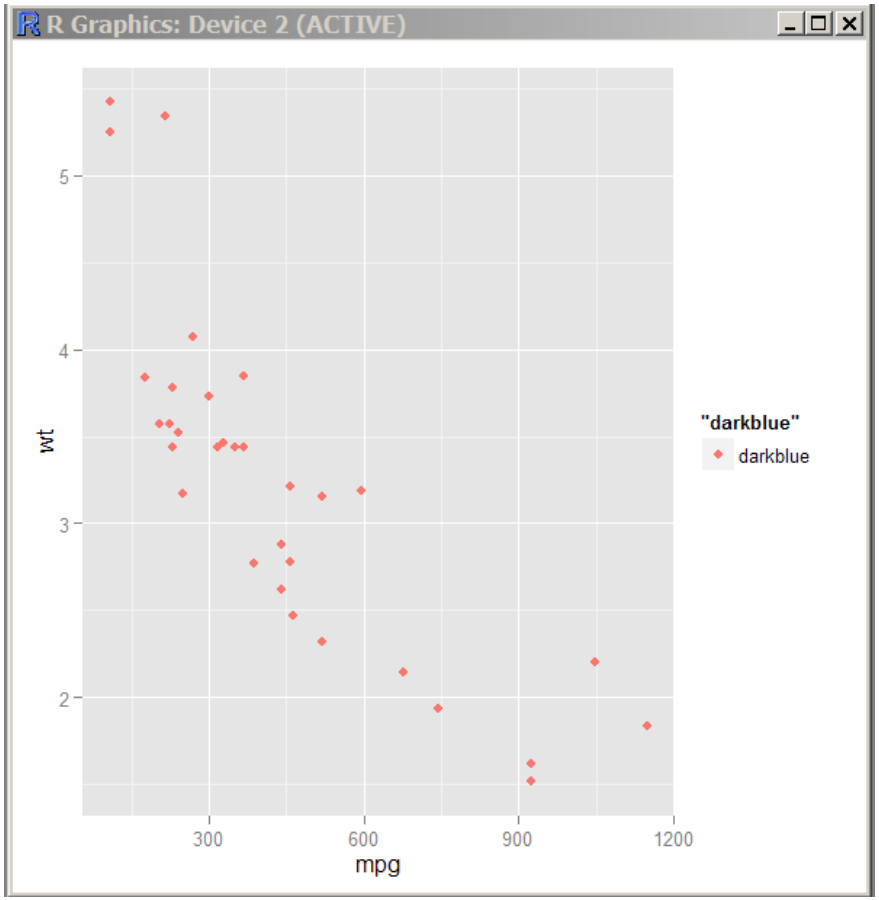
```
p <- ggplot(mtcars, aes(mpg, wt))
p + geom_point(colour = "darkblue")
```



Sets与Maps

Maps:

p + geom_point(aes(colour = "darkblue"))



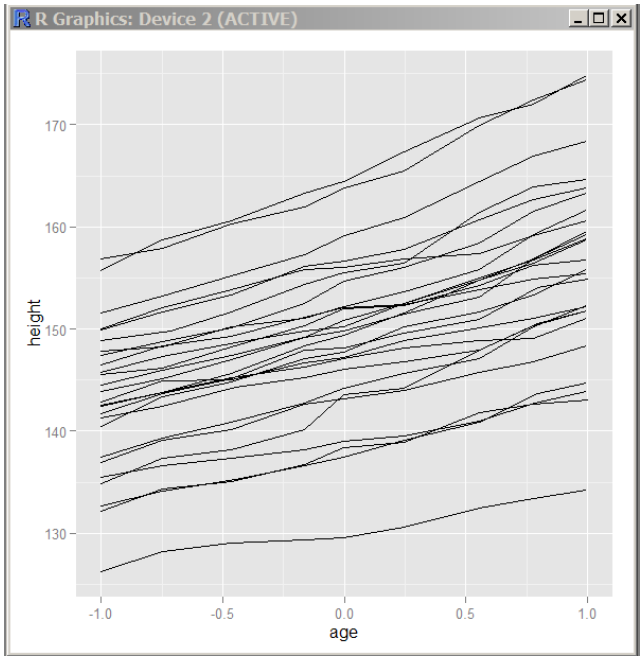
Groups

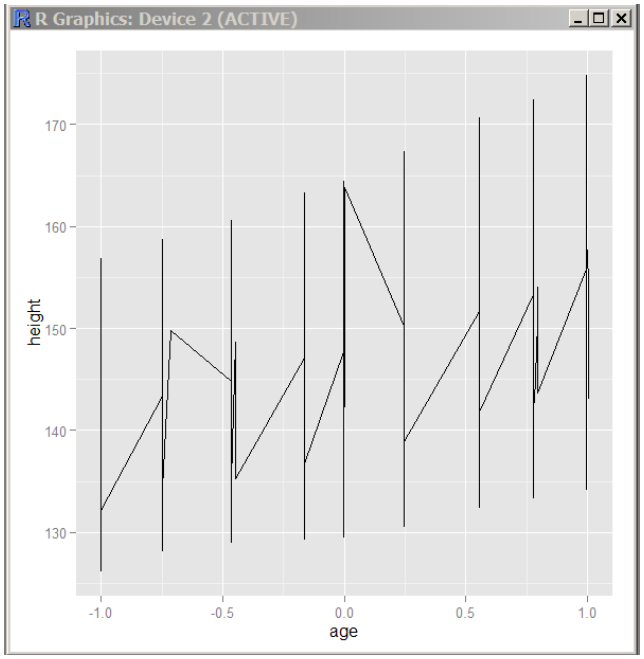
- Oxboys数据集

```
> Oxboys
Grouped Data: height ~ age | Subject
    Subject      age height Occasion
1         1 -1.0000 140.50        1
2         1 -0.7479 143.40        2
3         1 -0.4630 144.80        3
4         1 -0.1643 147.10        4
5         1 -0.0027 147.70        5
6         1  0.2466 150.20        6
7         1  0.5562 151.70        7
8         1  0.7781 153.30        8
9         1  0.9945 155.80        9
10        2 -1.0000 136.90        1
11        2 -0.7479 139.10        2
12        2 -0.4630 140.10        3
13        2 -0.1643 142.60        4
14        2 -0.0027 143.20        5
15        2  0.2466 144.00        6
16        2  0.5562 145.80        7
17        2  0.7781 146.80        8
```

多分组单一装饰属性

```
p <- ggplot(Oxboys, aes(age, height, group = Subject)) + geom_line()
P
p <- ggplot(Oxboys, aes(age, height, group = 1)) + geom_line()
p
```

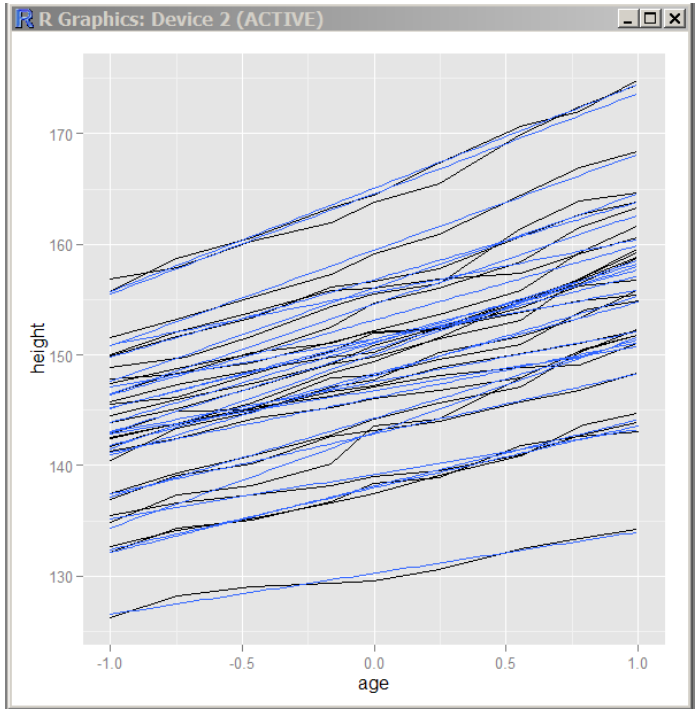


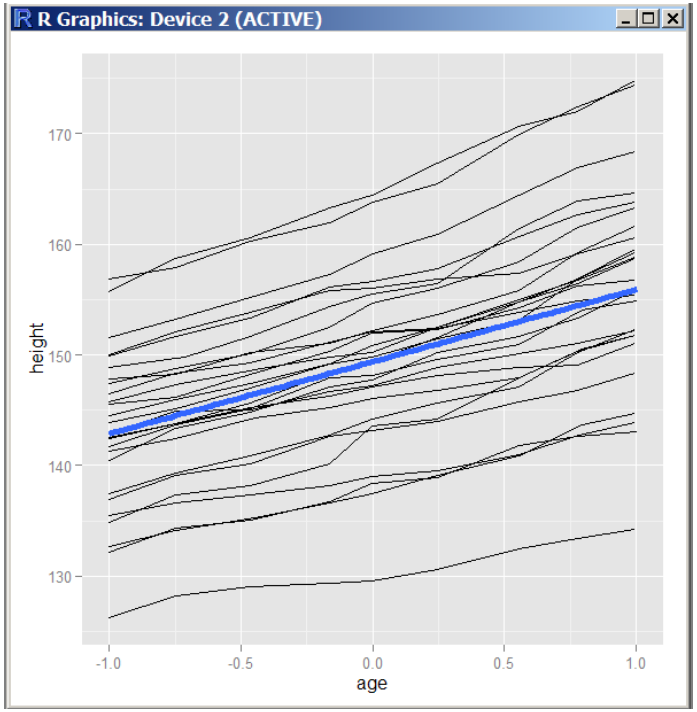


2013.2.8

续

```
p <- ggplot(Oxboys, aes(age, height, group = Subject)) + geom_line()
p + geom_smooth(aes(group = Subject), method="lm", se = F)
p + geom_smooth(aes(group = 1), method="lm", size = 2, se = F)
```





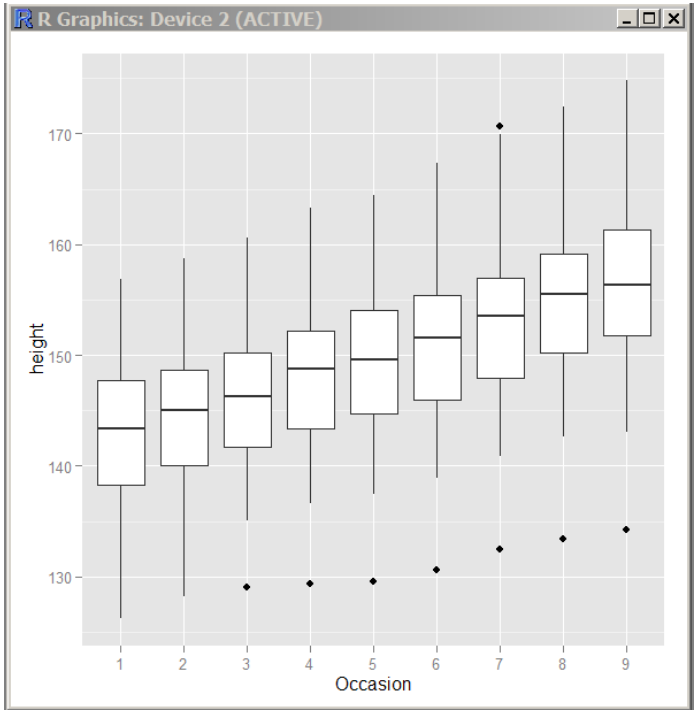
2013.2.8

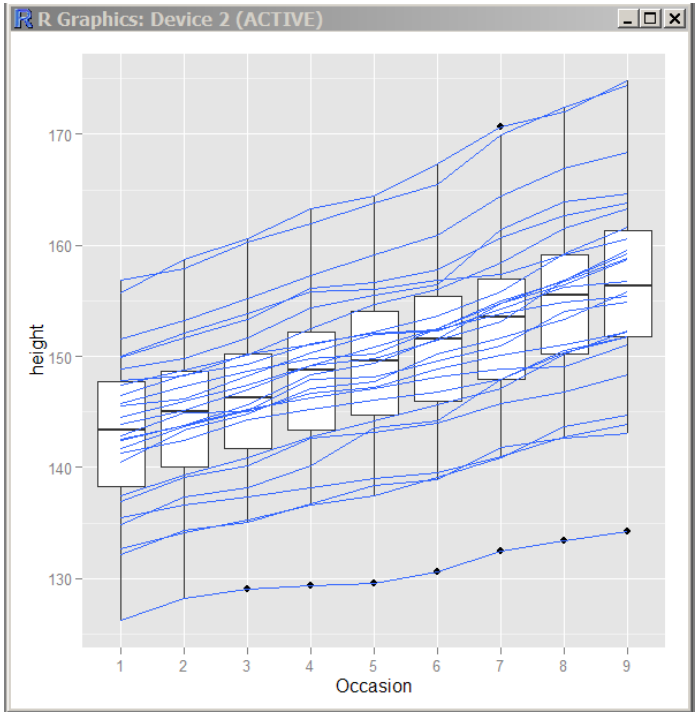
覆盖缺省分组

boysbox <- ggplot(Oxboys, aes(Occasion, height)) + geom_boxplot()

Boysbox

boysbox + geom_line(aes(group = Subject), colour = "#3366FF")





2013.2.8

Geoms

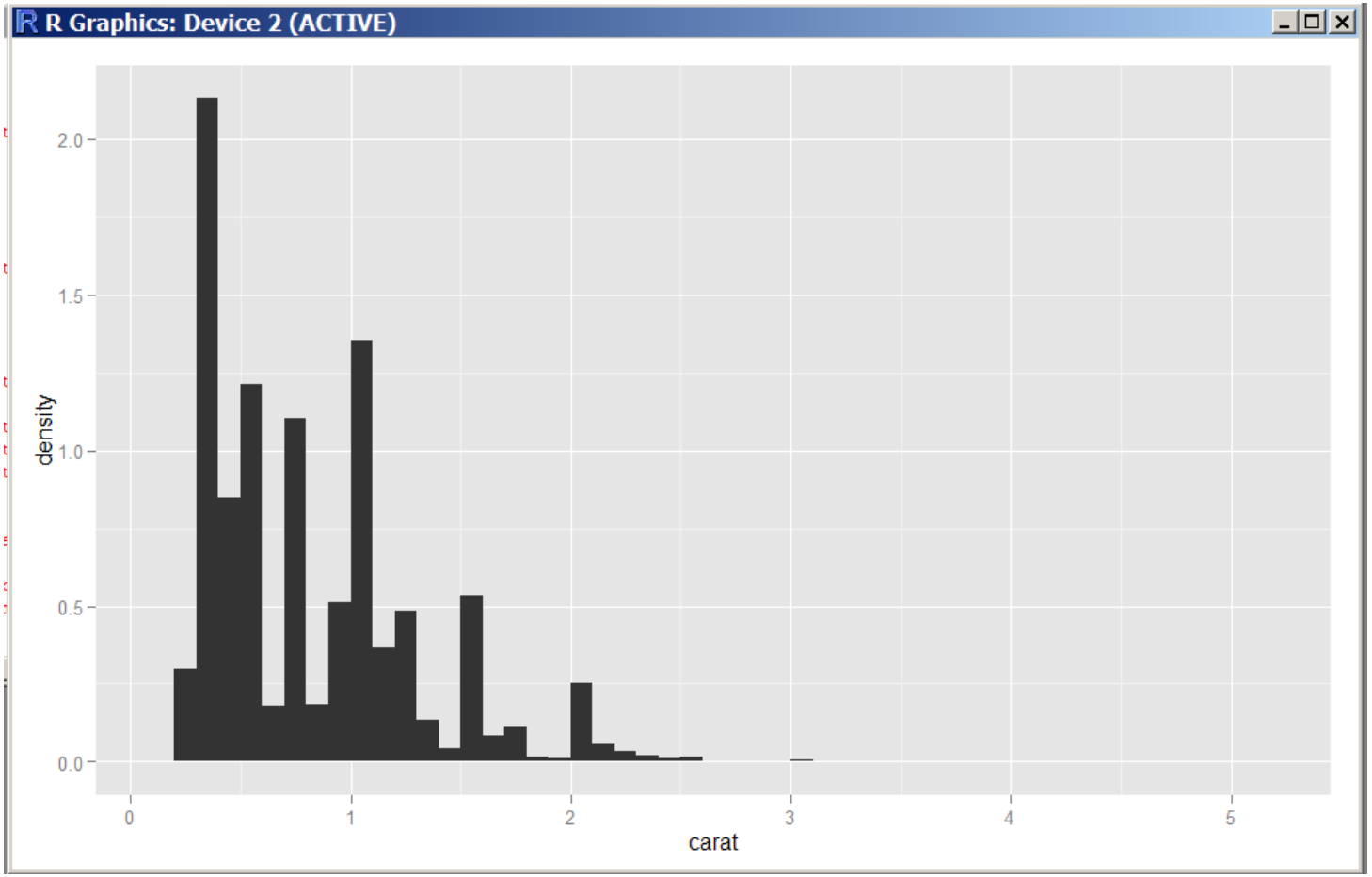
abline	Line, specified by slope and intercept
area	Area plots
bar	Bars, rectangles with bases on y-axis
blank	Blank, draws nothing
boxplot	Box-and-whisker plot
contour	Display contours of a 3d surface in 2d
crossbar	Hollow bar with middle indicated by horizontal line
density	Display a smooth density estimate
density_2d	Contours from a 2d density estimate
errorbar	Error bars
histogram	Histogram
hline	Line, horizontal
interval	Base for all interval (range) geoms
jitter	Points, jittered to reduce overplotting
line	Connect observations, in order of x value
linerange	An interval represented by a vertical line
path	Connect observations, in original order
point	Points, as for a scatterplot
pointrange	An interval represented by a vertical line, with a point in the middle
polygon	Polygon, a filled path
quantile	Add quantile lines from a quantile regression
ribbon	Ribbons, y range with continuous x values
rug	Marginal rug plots
segment	Single line segments
smooth	Add a smoothed condition mean
step	Connect observations by stairs
text	Textual annotations
tile	Tile plot as densely as possible, assuming that every tile is the same size
vline	Line, vertical

Stats

Name	Description
bin	Bin data
boxplot	Calculate components of box-and-whisker plot
contour	Contours of 3d data
density	Density estimation, 1d
density_2d	Density estimation, 2d
function	Superimpose a function
identity	Don't transform data
qq	Calculation for quantile-quantile plot
quantile	Continuous quantiles
smooth	Add a smoother
spoke	Convert angle and radius to xend and yend
step	Create stair steps
sum	Sum unique values. Useful for overplotting on scatter-plots
summary	Summarise y values at every unique x
unique	Remove duplicates

例子

ggplot(diamonds, aes(carat)) + geom_histogram(aes(y = ..density..), binwidth = 0.1)



2013.2.8

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间