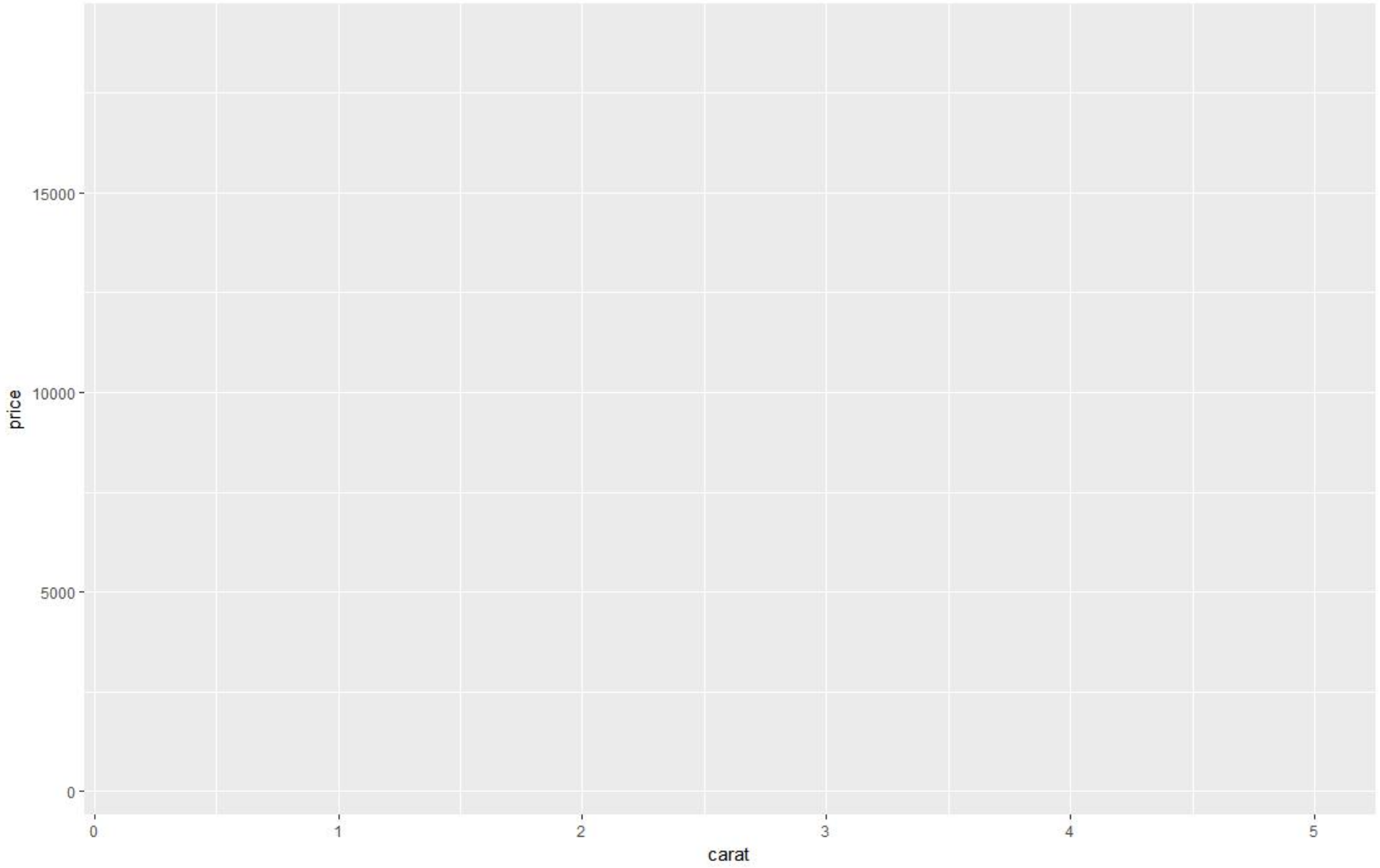
用图层构建对象

简介

- qplot()的局限性在于它只能使用一个数据集和一组图形属性映射，解决这个问题的办法就是使用图层。每个图层可以有自己的数据集和图形映射，附加的数据元素可通过图层添加到图形中。

创建绘图对象

- qplot()做了很多幕后工作，在整个过程中使用了很多默认的绘图参数。如果想手动创建图形对象，就需要使用ggplot()函数。
- ggplot()有两个主要的参数：**数据**和**图形属性映射**。
  - 参数“数据”指定绘图所用的默认数据集，**必须是data frame**（data.table也可以）。
  - 图形属性映射的设定方法与qplot()相似，只需**将图形映射属性和变量名放到函数aes()的括号里**即可。
- p <- ggplot(diamonds, **aes(carat, price, colour = cut)**)
- p #这个**图形对象在加上图层之前无法显示**，因此现在什么也看不见（见下页）。


15000
10000
5000
0
price
0
1
2
3
4
5
carat

图层

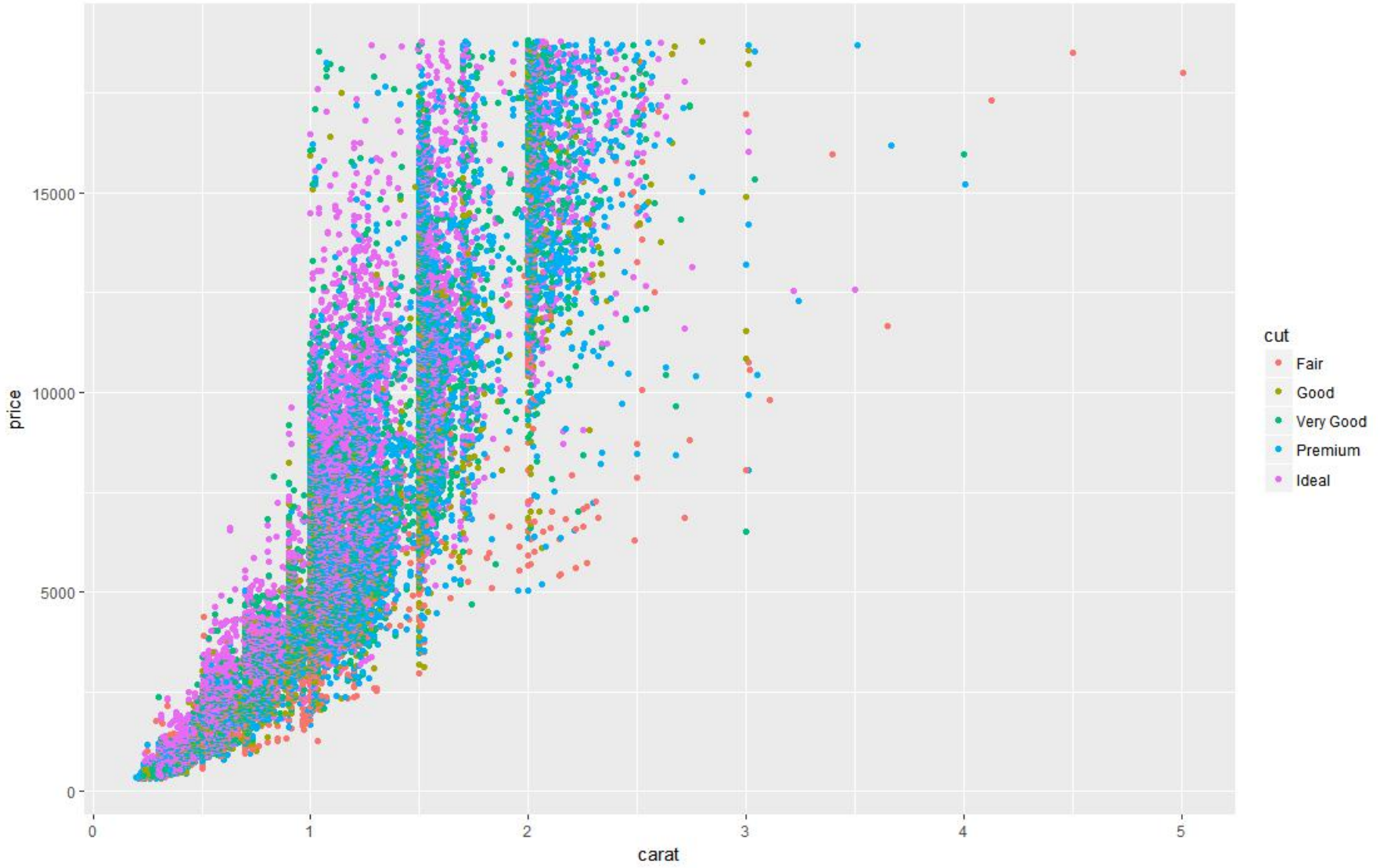
图层

- 最简单的图层只设定一个几何对象。

```
p <- ggplot(diamonds, aes(carat, price, colour = cut))
p <- p + layer(geom = "point",stat="identity",position="identity")
p #图在下页
```

新版**ggplot2**的**layer()**与教材已不一致。
不写这两个参数无法运行。

- 注意我们是用“+”来添加图层的。


price
15000
10000
5000
0
0
1
2
3
4
5
carat
cut
Fair
Good
Very Good
Premium
Ideal

图层参数

- 图层设定的参数很多：

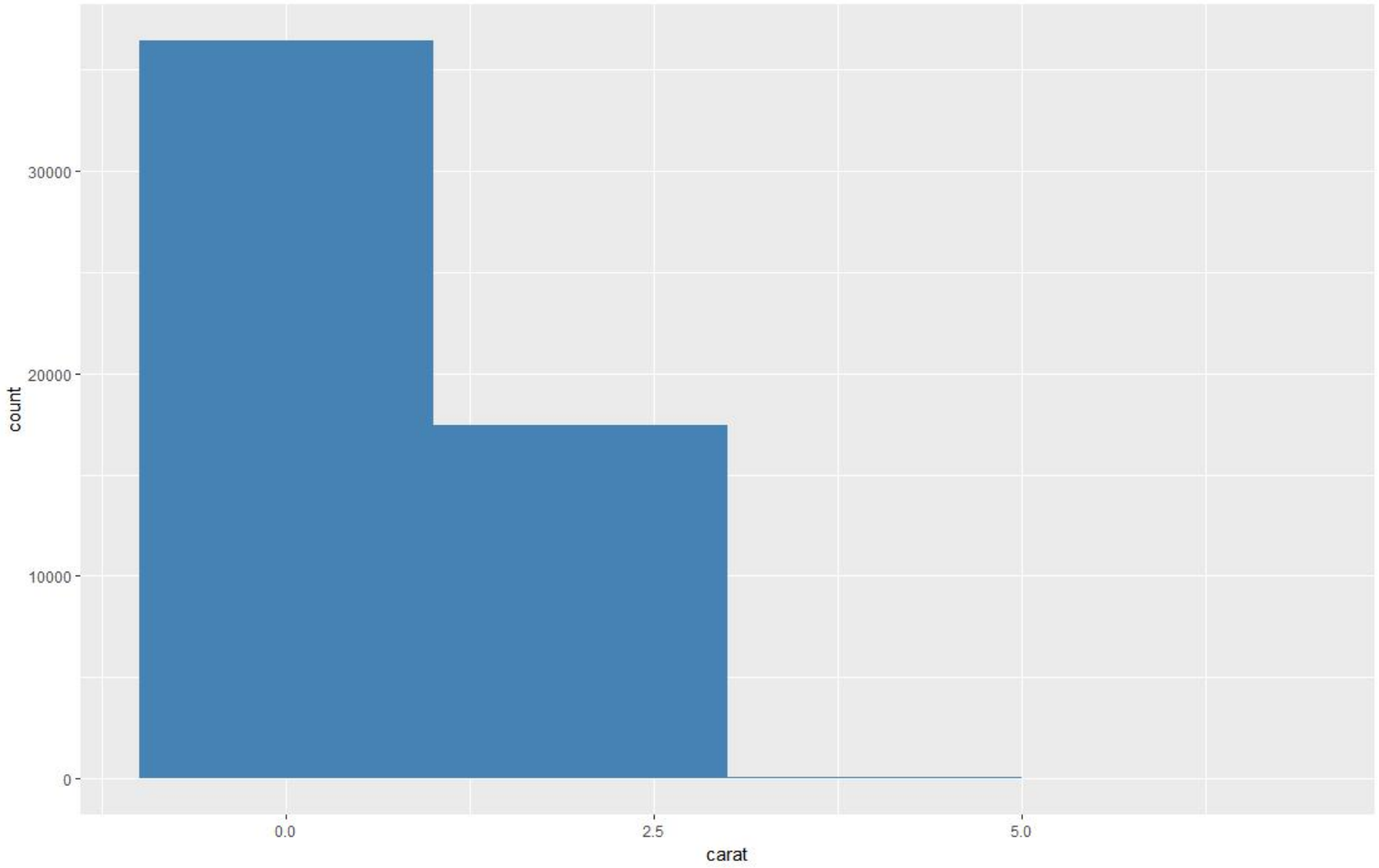
  layer(geom, params, stat, data, mapping, position)

- 例：

```
p <- ggplot(diamonds, aes(x = carat))
p <- p + layer(
    geom = "bar",
    params = list(fill = "steelblue", binwidth = 2),
    stat = "bin",
    position="identity"
)
p #图在下页
```

针对条形图图层的参数设定

fill = "steelblue", binwidth = 2



快捷函数（shortcut）

- 图层的参数设定细致但过于繁琐。可以用快捷函数（shortcut)来简化之前的代码。
- 每一个几何对象都对应一个默认的统计变换和位置参数，而每一个统计变换都对应着一个默认的几何对象参数，所以对于一个图层我们只需设定**stat**或**geom**参数即可。
- 上页图可用以下代码生成：

```
p+geom_histogram(binwidth = 2, fill =
"steelblue")
```

快捷函数的形式和参数

- 所有的快捷函数都有相同的形式：以geom或stat开头。
  - **geom**_XXX(mapping, data, ..., geom, position)
  - **stat**_XXX(mapping, data, ..., stat, position)
- 参数：
  - mapping：图形属性映射，通过aes()设定。
  - data：可以修改默认数据集
  - geom或stat参数，如上例中的binwidth = 2, fill = "steelblue"
  - position：选择一种调整对象重合的方式。
- 注意：参数data和mapping在ggplot()和图层函数中的位置是相反的。因为在图形对象中一般先设定数据集，而在图层函数中是审定图形属性而不是数据集。**要写清参数名**，而不要以来参数相对位置。

ggplot()和qplot()的等价性

- 例一：

```
qplot(sleep_rem / sleep_total, awake, data = msleep)
#等价于
ggplot(msleep, aes(sleep_rem / sleep_total, awake)) +
geom_point()
```

- 例二：

```
#也可以给qplot加图层:
qplot(sleep_rem / sleep_total, awake, data = msleep) +
  geom_smooth()
#等价于
ggplot(msleep, aes(sleep_rem / sleep_total, awake)) +
  geom_point() + geom_smooth()
```

图形对象

- 图形对象可以存储到一个变量里。
- summary() 可以查看图形的结构。
  - 首先给出图形的默认设置，然后给出每个图层的信息。
  - 例：

```
p <- ggplot(msleep, aes(sleep_rem / sleep_total, awake))
summary(p)
p <- p + geom_point()
summary(p)
```

图层对象

- 图层是普通的R对象，也可以存储到变量里。
- 例：图层bestfit可用于不同的数据：
  - **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**

图形属性映射（aesthetic mapping）

图形属性映射（aesthetic mapping）

- aes()用来将数据变量映射为图形属性（aesthetics）。**例：**

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

- 这里将x坐标映射为weight，y坐标映射为height，颜色映射为age。前两个参数中x=和y=可省略，会自动匹配。
- 注意：不要使用指定数据集以外的变量，因为这样无法将绘图所用数据都封装到一个对象里。
- 也可以使用变量的函数值作为参数。例：

  aes(weight, height, colour = sqrt(age))

图和图层（1/2）

- 默认的图形属性映射可以在图形对象初始化时设定，或者之后用“+”修改。
- 例：
  - p <- ggplot(mtcars, aes(x = mpg, y = wt))
  - p + geom_point()
  - p + geom_point(aes(colour = factor(cyl))) #用factor(cyl)修改颜色
  - p + geom_point(aes(y = disp)) #用disp修改y坐标值

图和图层（2/2）

- 图层的图形属性的添加、修改和删除。

Operation	Layer aesthetics	Result
Add	`aes(colour = cyl)`	`aes(mpg, wt, colour = cyl)`
Override	`aes(y = disp)`	`aes(mpg, disp)`
Remove	`aes(y = NULL)`	`aes(mpg)`

分组（group）

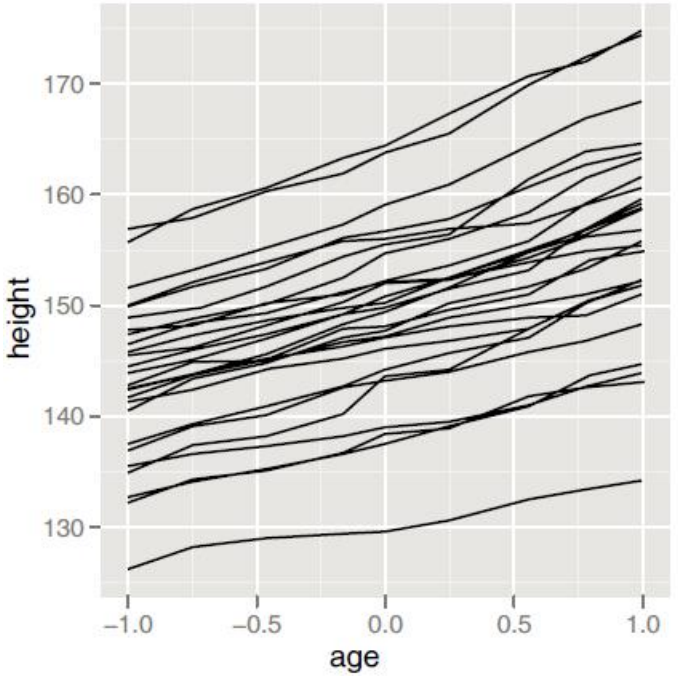
- 所有离散型变量的交互作用被设计为分组的默认值。但如果没能正确分组或图中没有离散型变量，就需要自定义分组结构。

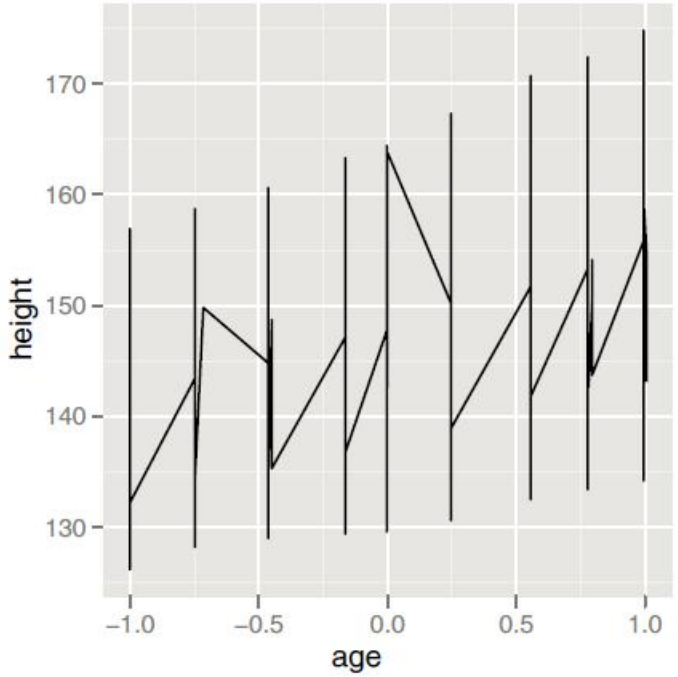
数据集

- 数据集：nlme包中的Oxboys：26名男孩（subject）在9个不同的时期（Occasion）所测定的身高（height）和中心化后的年龄（age）。
- library(nlme)
- head(Oxboys)

多个分组与单个图形属性

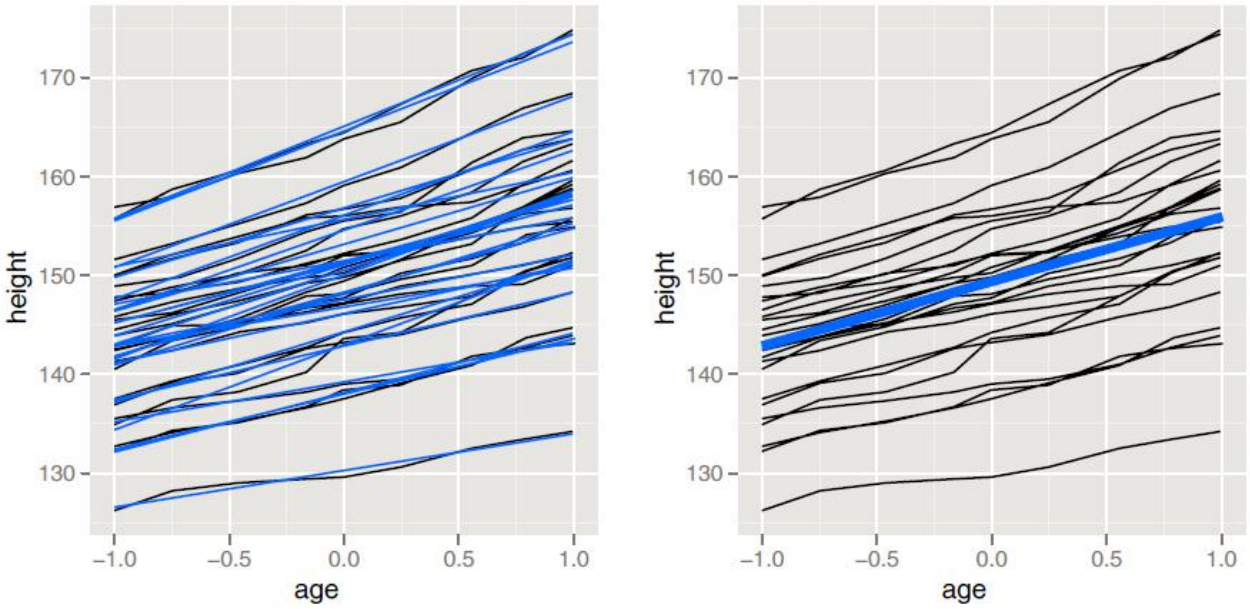
- qplot(age, height, data=Oxboys, group = Subject, geom="line") #左图：指定Subject（每个男孩）为分组标量，每条线对应一个男孩。
- qplot(age, height, data=Oxboys, geom="line") #右图，无分组标量，等同于group = 1。此时图无意义。





不同图层上的不同分组

- 有时我们想整合不同水平下的数据来对统计汇总信息进行图形绘制。
- 例：在上页例子基础上，根据所有男孩的年龄和身高在图中添加一条光滑线条。
- p <- ggplot(Oxboys, aes(age, height, group = Subject)) + geom_line()
- p + geom_smooth(aes(group = Subject), method="lm", se = F) #左图，无意间给每个男孩添加了一条光滑线条，这不是我们想要的。
- p + geom_smooth(aes(**group = 1**), method="lm", size = 2, se = F) #右图



匹配图形属性和图形对象（1/4）

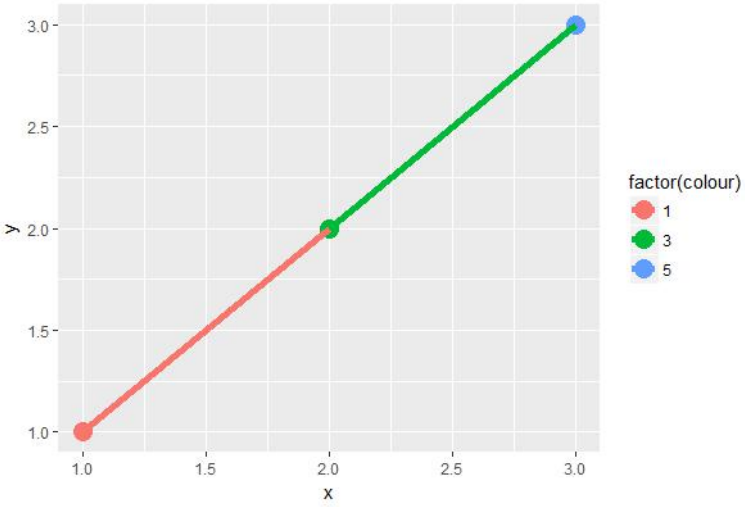
- 群组几何对象的另一重要议题是，如何将个体的图形属性映射给整体的图形属性。
- 对于个体几何对象这不是问题，因为每一条观测都被一个单一的图形元素所表示。但高密度的数据会使得区别单个点变得困难或不可能。

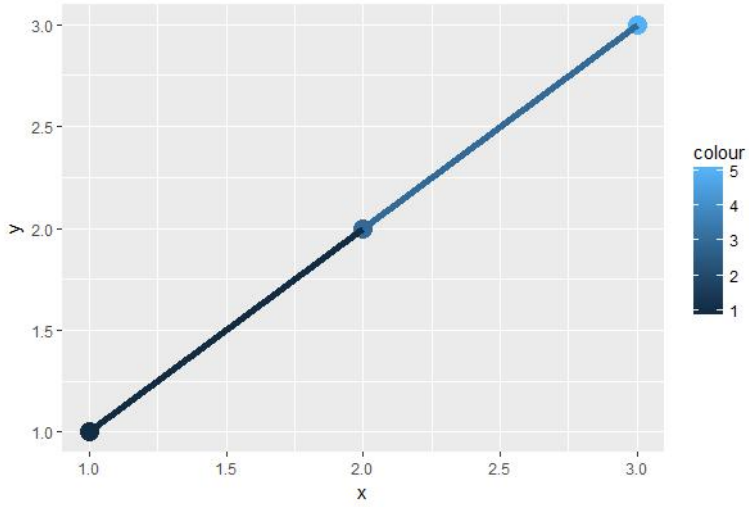
匹配图形属性和图形对象（2/4）

- **线条和路径**遵循**差一原则**：观测点比线段数码多一，第一条线段将使用第一条观测的图形属性，第二条线段将使用第二条观测的图形属性，以此类推。这意味着最后一条观测的图形属性不会被用到。

```
> df
  x y colour
1 1 1      1
2 2 2      3
3 3 3      5
```

  - df <- data.frame(x = 1:3, y = 1:3, colour = c(1,3,5))
  - qplot(x, y, data=df, colour=**factor**(colour), size = I(5)) + geom_line(aes(group = 1), size = 2) #左图：颜色是**离散**的
  - qplot(x, y, data=df, colour=colour, size = I(5)) + geom_line(size = 2) #右图：尽管颜色是**连续**的，但默认条件下，R不会对相邻颜色进行插补。插补见下页。





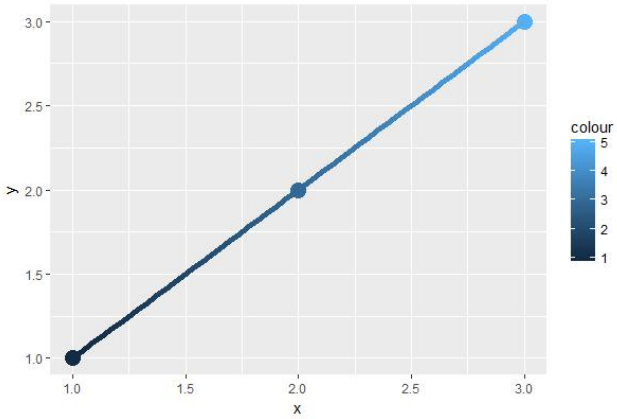
匹配图形属性和图形对象（3/4）

- 对于连续型变量，如果希望线段平稳地从一种图形属性变换到另一种图形属性，可以使用线性插值法。

```
xgrid <- with(df, seq(min(x), max(x), length = 50))
interp <- data.frame(
  x = xgrid,
  y = approx(df$x, df$y, xout = xgrid)$y,
  colour = approx(df$x, df$colour, xout = xgrid)$y
)
qplot(x, y, data = df, colour = colour, size = I(5)) + geom_line(data = interp, size = 2)
```

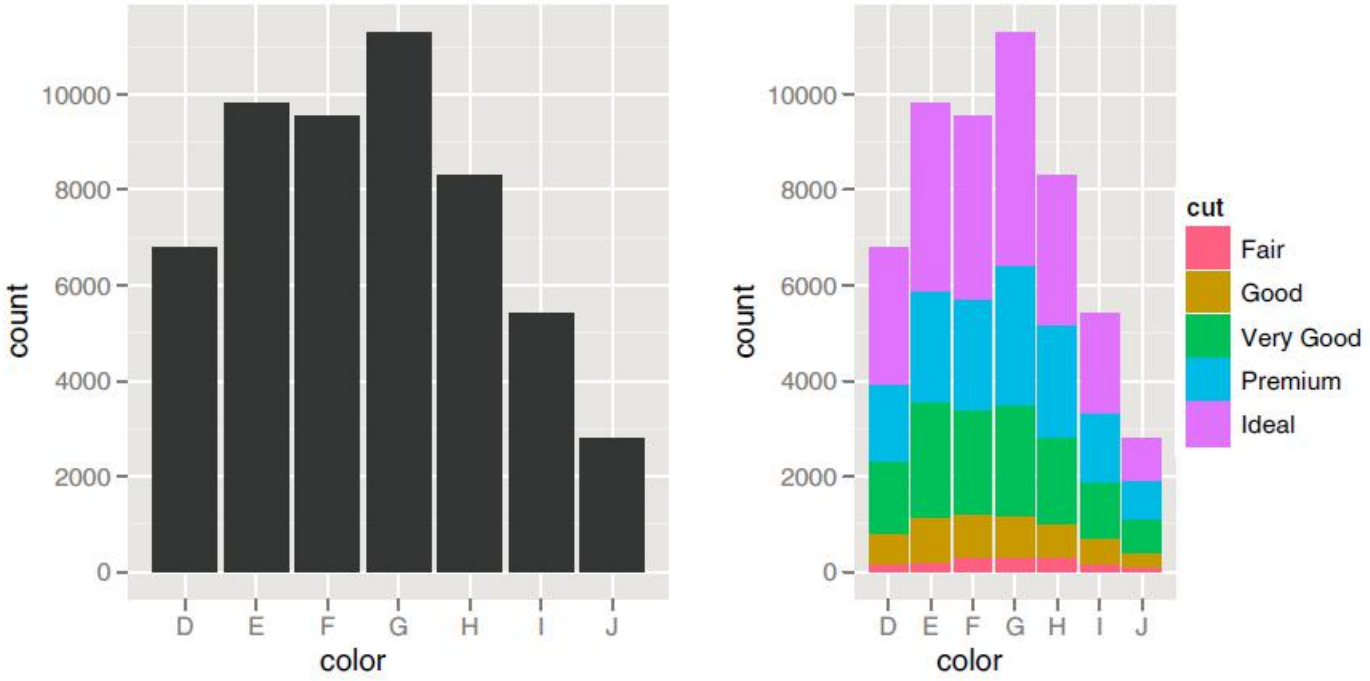
```
>    df
  x y colour
1 1 1      1
2 2 2      3
3 3 3      5
```

```
> interp
          x        y   colour
1  1.000000 1.000000 1.000000
2  1.040816 1.040816 1.081633
3  1.081633 1.081633 1.163265
4  1.122449 1.122449 1.244898
5  1.163265 1.163265 1.326531
6  1.204082 1.204082 1.408163
7  1.244898 1.244898 1.489796
8  1.285714 1.285714 1.571429
9  1.326531 1.326531 1.653061
10 1.367347 1.367347 1.734694
11 1.408163 1.408163 1.816327
```



匹配图形属性和图形对象（4/4）

- 当映射对象是离散型的变量时，它将默认地把群组几何对象分解成更小的块。
  - 例：一个条形图（左）按组分解后得到的叠加条形图（右），两者轮廓相同。



几何对象

几何对象

- 几何图形对象（**geom**）负责图层的渲染和图的类型。
- 例如使用point geom会生成散点图，使用line geom会生成折线图。
- 每个几何对象都有一组它能识别的图形属性和一组绘图所需的值。
- 有些几何对象主要是它们参数不同。
  - 例如同样是“矩形”，tile geom设定的是中心位置、长、宽；而rect geom设定的是上、下、左、右。
- 每一个几何对象都有一个默认的统计变换，并且每一个统计变换都有一个默认的几何对象。

ggplot2中的几何对象

Name	Description
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

翻译：

abline	线，由斜率和截距决定
area	面积图 (area plot)
bar	条形图，以 x 轴为底的矩形
bin2d	2 维热图
blank	空白，什么也不画
boxplot	箱线图
contour	等高线图
crossbar	带有水平中心线的盒子图
density	光滑密度曲线图
density2d	二维密度等高线图
dotplot	“点直方图”，用点来表示观测值的个数
errorbar	误差棒
errorbarh	水平的误差棒
freqpoly	频率多边形图
hex	用六边形表示的 2 维热图
histogram	直方图
hline	水平线
jitter	给点添加扰动，减轻图形重叠问题
line	按照 x 坐标的大小顺序依次连接各个观测值
linerange	一条代表一个区间的竖直线
map	基准地图里的多边形
path	按数据的原始顺序连接各个观测值
point	点，用来绘制散点图
pointrange	用一条中间带点的竖直线代表一个区间
polygon	多边形，相当于一个有填充的路径
quantile	添加分位数回归线
raster	高效的矩形瓦片图
rect	2 维的矩形图
ribbon	色带图，连续的 x 值所对应的 y 的范围
rug	边际地毯图
segment	添加线段或箭头
smooth	添加光滑的条件均值线
step	以阶梯形式连接各个观测值
text	文本注释
tile	瓦片图
violin	小提琴图
vline	竖直线

默认的统计变换和图形属性，黑体图形属性是必须声明的参数

Name	Default stat	Aesthetics
abline	abline	colour, linetype, size
area	identity	colour, fill, linetype, size, **x**, **y**
bar	bin	colour, fill, linetype, size, weight, **x**
bin2d	bin2d	colour, fill, linetype, size, weight, **xmax**, **xmin**, **ymax**, **ymin**
blank	identity	
boxplot	boxplot	colour, fill, **lower**, **middle**, size, **upper**, weight, **x**, **ymax**, **ymin**
contour	contour	colour, linetype, size, weight, **x**, **y**
crossbar	identity	colour, fill, linetype, size, **x**, **y**, **ymax**, **ymin**
density	density	colour, fill, linetype, size, weight, **x**, **y**
density2d	density2d	colour, linetype, size, weight, **x**, **y**
errorbar	identity	colour, linetype, size, width, **x**, **ymax**, **ymin**
freqpoly	bin	colour, linetype, size
hex	binhex	colour, fill, size, **x**, **y**
histogram	bin	colour, fill, linetype, size, weight, **x**
hline	hline	colour, linetype, size
jitter	identity	colour, fill, shape, size, **x**, **y**
line	identity	colour, linetype, size, **x**, **y**
linerange	identity	colour, linetype, size, **x**, **ymax**, **ymin**
path	identity	colour, linetype, size, **x**, **y**
point	identity	colour, fill, shape, size, **x**, **y**
pointrange	identity	colour, fill, linetype, shape, size, **x**, **y**, **ymax**, **ymin**
polygon	identity	colour, fill, linetype, size, **x**, **y**
quantile	quantile	colour, linetype, size, weight, **x**, **y**
rect	identity	colour, fill, linetype, size, **xmax**, **xmin**, **ymax**, **ymin**
ribbon	identity	colour, fill, linetype, size, **x**, **ymax**, **ymin**
rug	identity	colour, linetype, size
segment	identity	colour, linetype, size, **x**, **xend**, **y**, **yend**
smooth	smooth	alpha, colour, fill, linetype, size, weight, **x**, **y**
step	identity	colour, linetype, size, **x**, **y**
text	identity	angle, colour, hjust, **label**, size, vjust, **x**, **y**
tile	identity	colour, fill, linetype, size, **x**, **y**
vline	vline	colour, linetype, size

统计变换

统计变换

- 统计变换（**stat**）通常以某种方式对数据信息进行汇总。
- 为了阐明在图形的意义，一个统计变换必须是一个位置尺度不变量，即

$$f(x + a) = f(x) + a \text{ 且 } f(b \cdot x) = b \cdot f(x)$$

- 这样才能保证当改变图形的标度时，数据变换保持不变。

ggplot2中的统计变换

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 scatterplots
summary	Summarise y values at every unique x
unique	Remove duplicates

生成变量（1/2）

- 统计变换可以将图形属性映射成新变量。
- 例如，用于绘制直方图的stat_bin统计变换会生成以下新变量：
  - count，每个组里观测值的数目
  - density，每个组里观测值的密度（占整体比例）
  - x，组的中心位置
- 这些生成变量可以被直接调用。

生成变量（2/2）

- 例：直方图默认的条形高度是观测值的频数。下面代码可以使用传统的密度作为高度。
- ggplot(diamonds, aes(carat)) + geom_histogram(aes(y = **..density..**), binwidth = 0.1)

生成变量的名字必须要用..围起来，这样是为了防止混淆原数据集中的变量和生成变量。

位置调整（1/2）

- 位置调整是对该层中的元素进行微调。位置调整一般多见于处理离散型数据，连续型数据一般很少出现完全重叠的问题（并且出现问题用位置调整也解决不了）。
- 五种位置调整参数

Adjustment	Description
dodge	Adjust position by dodging overlaps to the side
fill	Stack overlapping objects and standardise have equal height
identity	Don't adjust position
jitter	Jitter points to avoid overplotting
stack	Stack overlapping objects on top of one another

位置调整（2/2）

应用条形图的三种位置调整：

```
dplot <- ggplot(diamonds, aes(clarity, fill = cut))
dplot + geom_bar(position = "stack") #左
dplot + geom_bar(position = "fill") #中
dplot + geom_bar(position = "dodge") #右
```