数据可视化

高级绘图：ggplot2

本节内容

- ggplot2简介
- 从qplot开始入门

简介

ggplot2

- ggplot2是R的相对比较年轻的一个包。
- Wilkinson在“Grammar of Graphics”（2005）一书中给出一套图形语法。核心思想：

  A statistical graphic is a mapping from data to aesthetic attributes (colour, shape, size) of geometric objects (points, lines, bars).
- 基于该图形语法，Hadley Wickham在爱荷华州立大学读博士期间完成了ggplot2（是他博士论文主题之一）。

ggplot2的特点

- ggplot2的核心理念是将绘图与数据分离，数据相关的绘图与数据无关的绘图分离
- 按图层作图
- 保有命令式作图的调整函数，使其更具灵活性
- 将常见的统计变换融入到了绘图中

ggplot2的安装和使用

- 安装：

  install.packages("ggplot2")
- 使用：

  library(ggplot2)

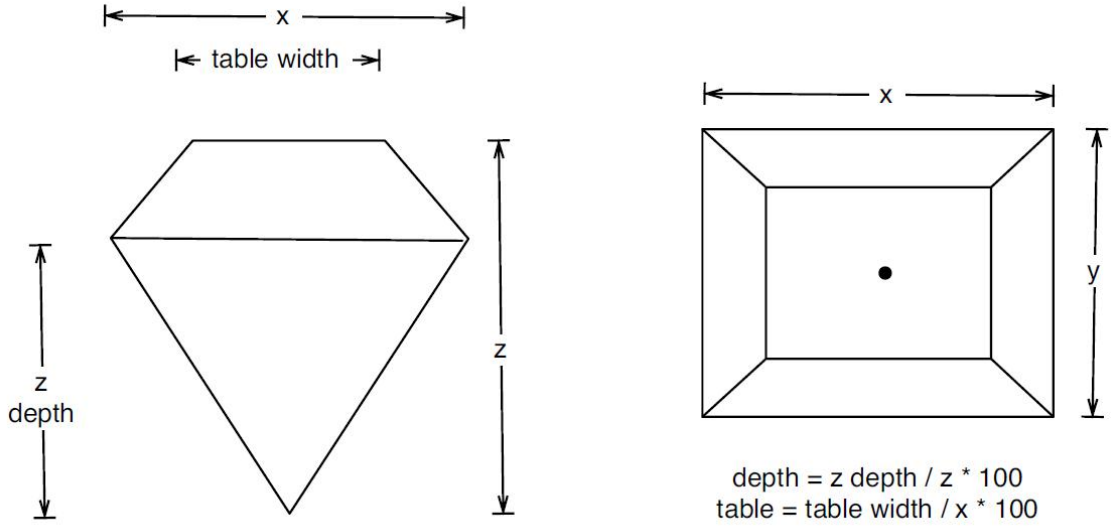
从qplot开始入门

qplot()

- R语言作者Hadley知道我们接受一种新事物不会太容易，所以设计了个qplot函数。
- qplot 即quick plot，能快速对数据进行可视化分析。
- 它的用法和R base包的plot函数很相似，主要作用是让用户在不知不觉中洗脑。

数据集（1/2）

- diamonds数据集：54000颗钻石的价格和品质（四个“C”：克拉重量carat、切工cut、颜色color、净度clarity；五个物理指标：深度depth、钻面宽度table、x、y、z。）



数据集（2/2）

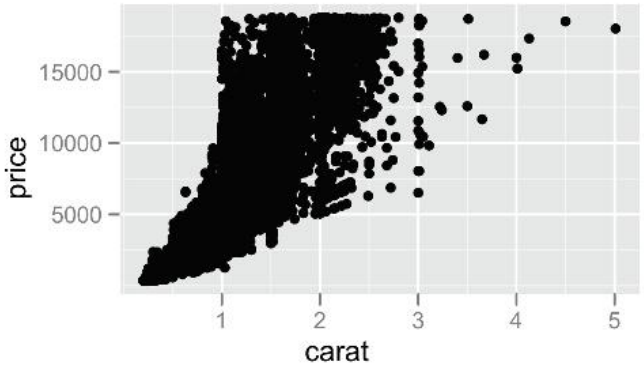
- dsmall：diamonds的一个容量为100的随机样本。
- 代码：

```
set.seed(1410) #让样本可重复
dsmall <- diamonds[sample(nrow(diamonds), 100), ]
```

qplot()基本用法（1/2）

- 与plot()类似，qplot()前两个参数是x和y，分别代表图中所画对象的x坐标和y坐标。data参数可选，如果指定，那么qplot首先会在该数据框内查找变量名。
- 例：钻石价格和重量的关系

qplot(carat, price, data = diamonds)



- 可以看出变量之间有很强的相关关系和一些明显的异常值。
- 相关关系似乎是**指数**型的。

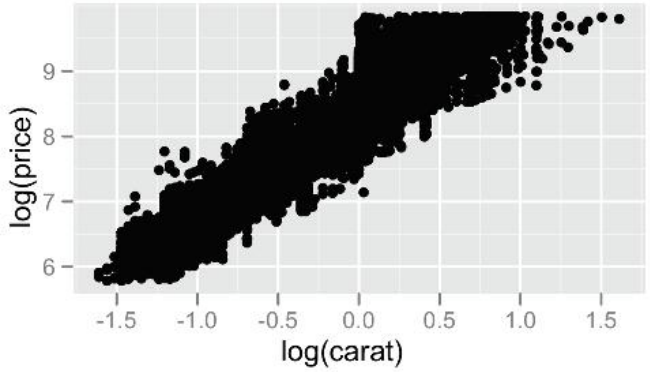
qplot()基本用法（2/2）

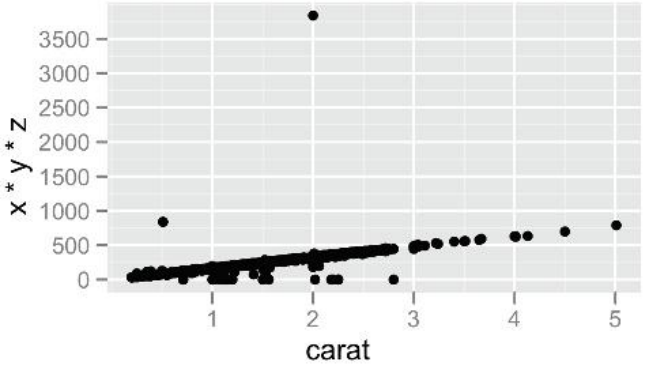
- 取log之后就接近线性了（左下图）

  qplot(**log**(carat), log(price), data = diamonds)

- 钻石的体积和重量应该是线性关系（钻石的密度），但仍存在异常点（右下图）

  qplot(carat, x * y * z, data = diamonds)





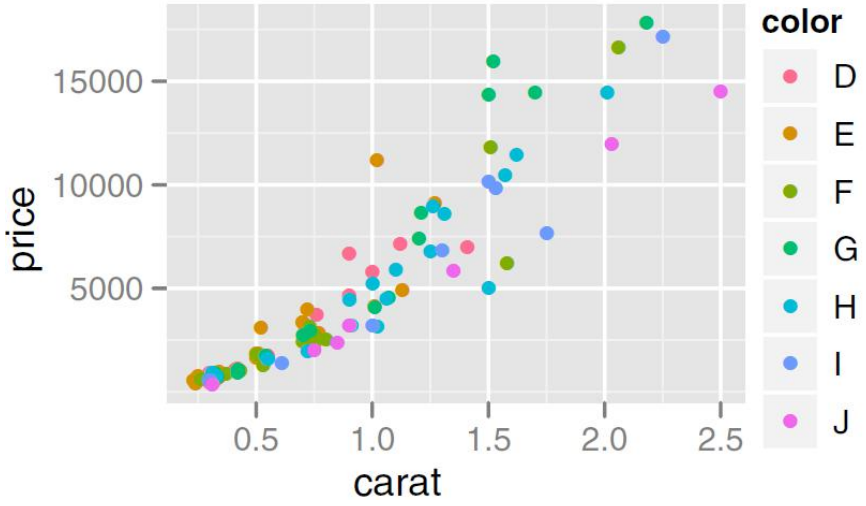
颜色、大小、形状等属性

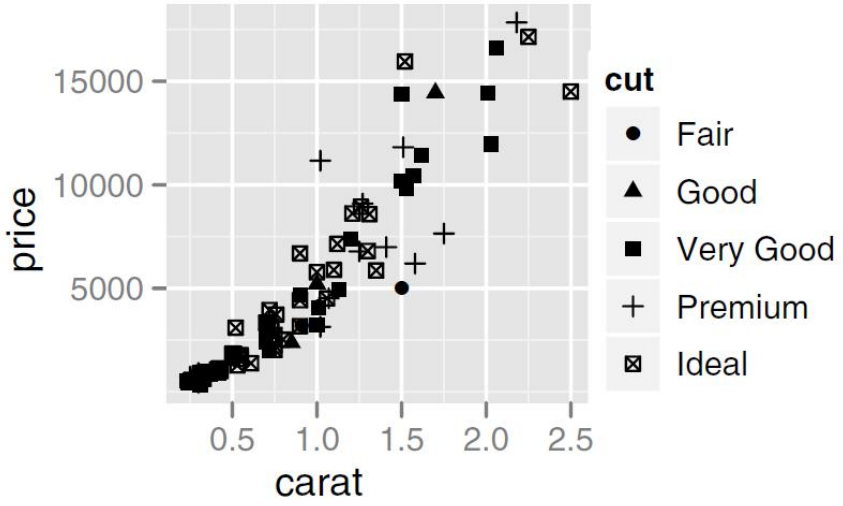
- qplot与plot的第一个不同在于给图中的点设定颜色（大小、或形状）是采用了不同方式。
  - 在plot中，用户需要将数据的分类变量（例如苹果、香蕉、桃子）转换为plot可以理解的形式（例如red, yellow, green）。
  - 而qplot可以自动完成。

颜色、大小、形状等属性：例1

qplot(carat, price, data = dsmall, **colour =** color)

qplot(carat, price, data = dsmall, **shape =** cut)



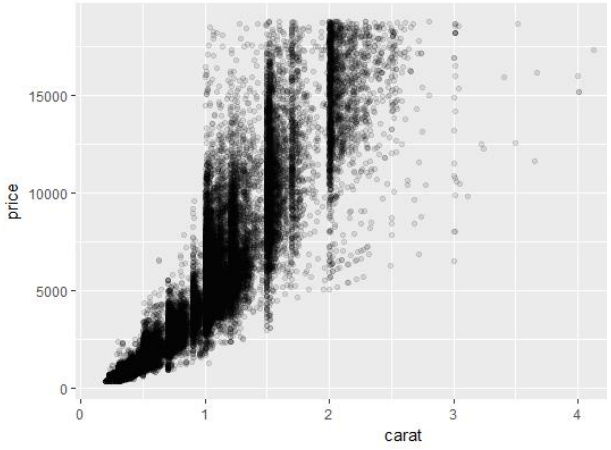


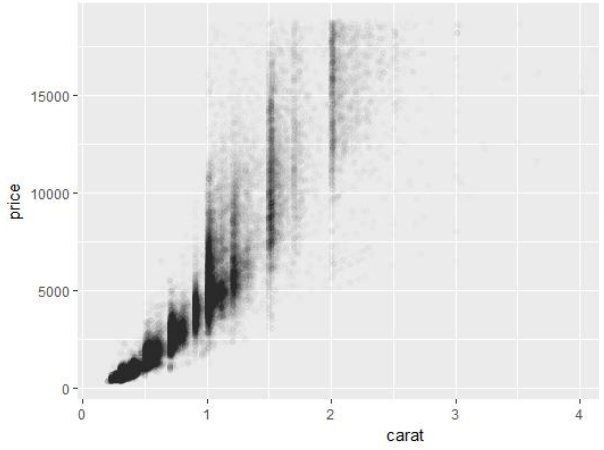
颜色、大小、形状等属性：例2

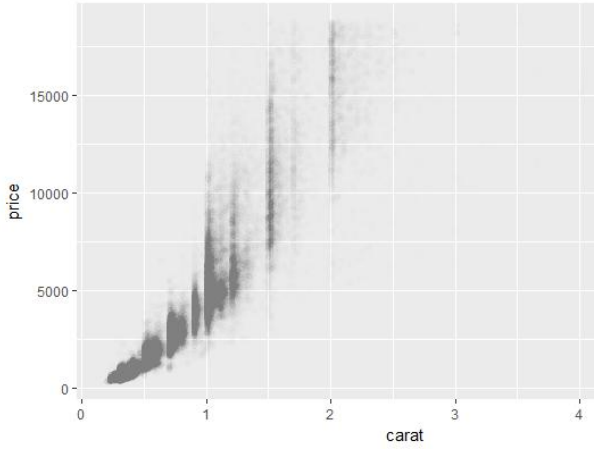
不写I()被当做字符处理

- alpha用来创建半透明的颜色，其取值从0（完全透明）到1（完全不透明）。例：

```
qplot(carat, price, data = diamonds, alpha = I(1/10))
qplot(carat, price, data = diamonds, alpha = I(1/100))
qplot(carat, price, data = diamonds, alpha = I(1/200))
```







几何对象（简写为**geom**）

- 除了散点图，通过改变几何对象，qplot几乎可以画出任何一种类型的图形。
- 几何对象描述了应该用何种对象来对数据进行展示，其中有些几何对象关联了相应的统计变换。

geom参数（1/2）

- **geom =** "point" 散点图（默认选项）
- geom = "smooth" 拟合平滑曲线
- geom = "boxplot" 箱线图
- geom = "path" 或 geom = "line" 数据点间绘制连线

geom参数（2/2）

- 对于一维的分布，geom的选择由变量类型指定
  - 连续变量
    - geom = "histogram" 直方图（一维数据默认选项）
    - geom = "freqpoly" 频率多边形
    - geom = "density" 密度曲线
  - 离散变量
    - geom = "bar" 条形图

平滑曲线（smoother）（1/2）

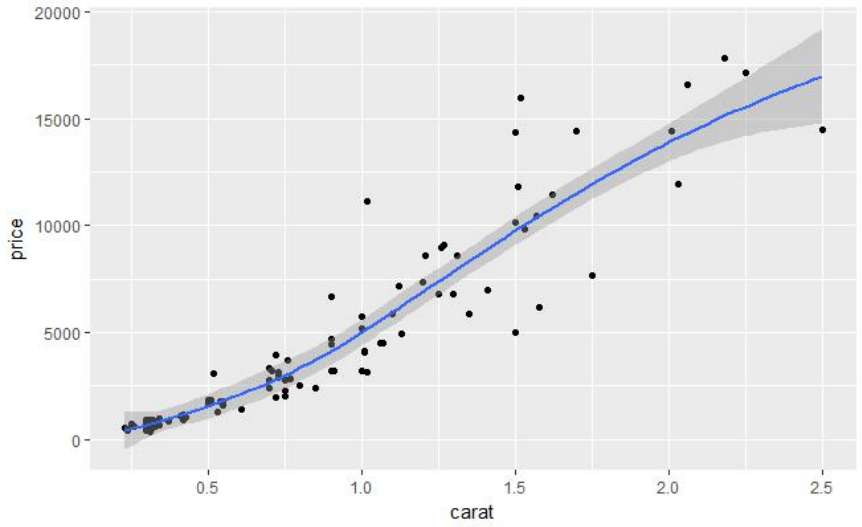
- 如果在散点图中有非常多的数据点，那么数据展示的趋势可能并不明显。在这种情况下应该在图中添加一条平滑曲线。
- 用法：利用c()将多个几何对象组成一个向量传递给geom。

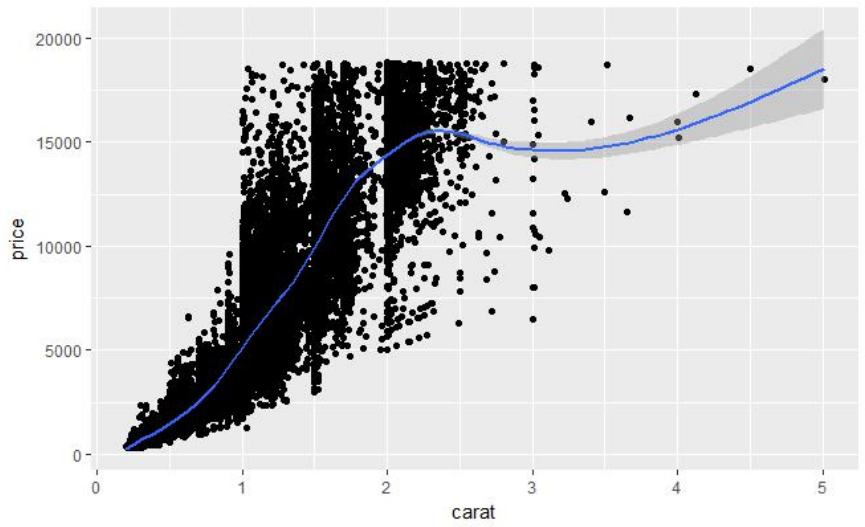
qplot(carat, price, data = dsmall, **geom = c("point", "smooth")**)

平滑曲线（2/2）

qplot(carat, price, data = dsmall, geom = c("point", "smooth")) #采样数据集

qplot(carat, price, data = diamonds, geom = c("point", "smooth")) #完整数据集





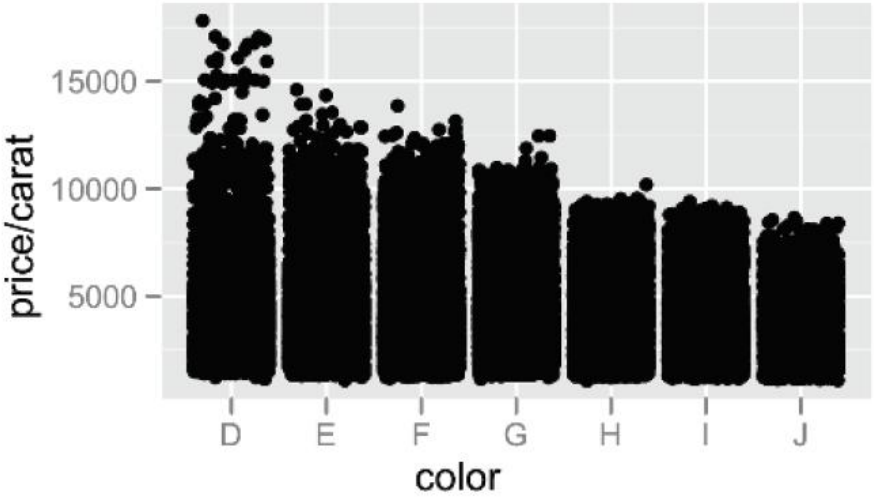
箱线图（boxplots）和扰动点图（ jittered points）（1/2）

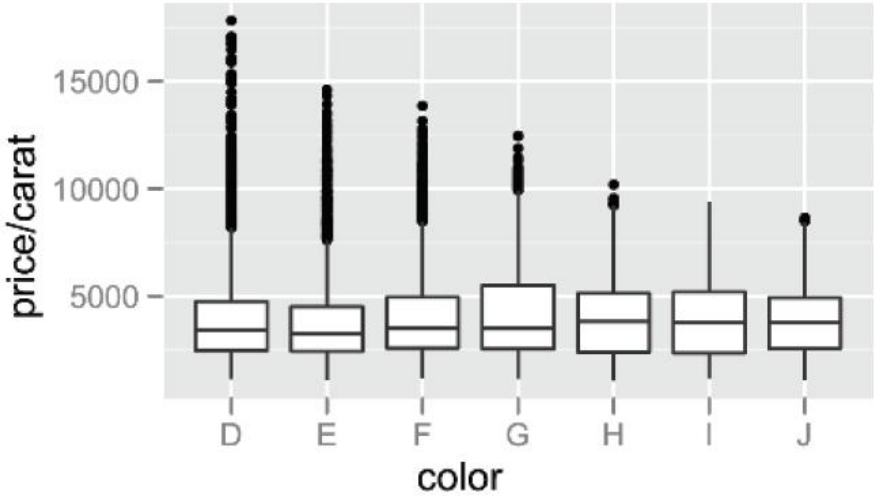
- 如果一个数据集中包含了一个分类变量和多个连续变量，可以使用箱线图或扰动点图来知道连续变量如何随着分类变量的变化而变化。

箱线图和扰动点图（2/2）

qplot(color, price / carat, data = diamonds, **geom = "jitter"**) #扰动点图

qplot(color, price / carat, data = diamonds, **geom = "boxplot"**) #箱线图





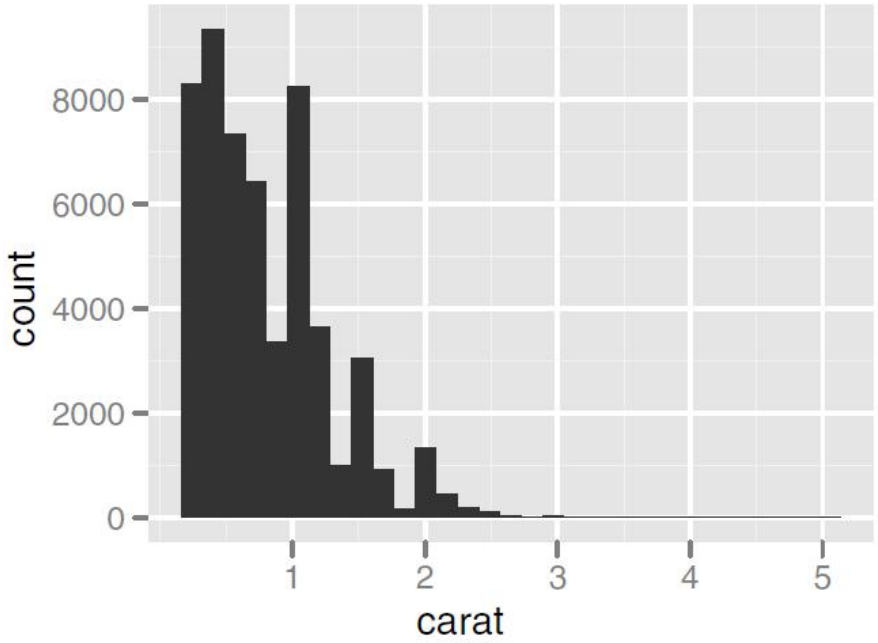
- 箱线图的信息更充分：显示分布的中位数和四分位数都没有太大变化。

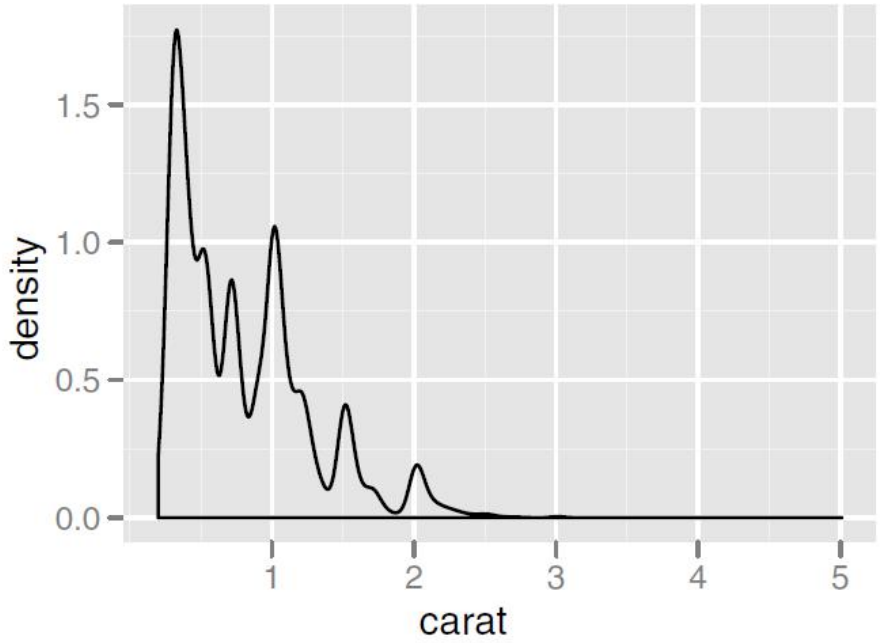
直方图（histogram）和密度曲线图（density plots）（1/5）

- 直方图和密度曲线图可以展示**单个变量的分布**（但它们不太容易在不同组之间进行比较）。

直方图和密度曲线图（2/5）
例：展示钻石重量的分布

```
#直方图
qplot(carat, data = diamonds, geom = "histogram")
#密度曲线图
qplot(carat, data = diamonds, geom = "density")
```





直方图和密度曲线图（3/5）

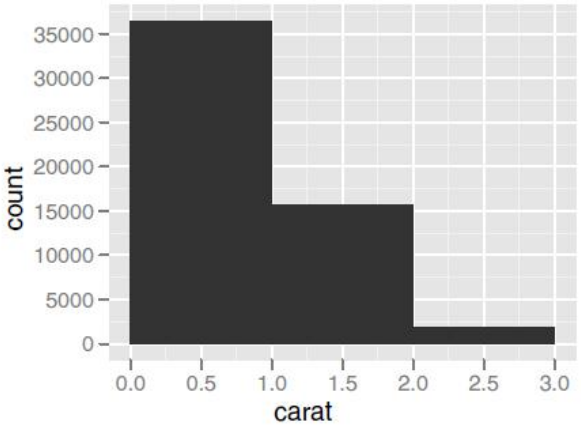
- 对于**密度曲线图**，**adjust**参数控制曲线**平滑程度**（取值越大越平滑）。
- 对于**直方图**，通过**binwidth**参数设定**组距**来调节平滑度。
- 绘制图形时，对平滑程度进行试验非常重要。

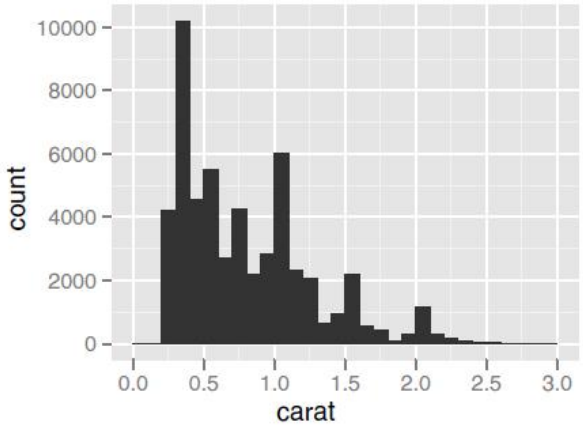
直方图和密度曲线图（4/5）
不同组距的直方图

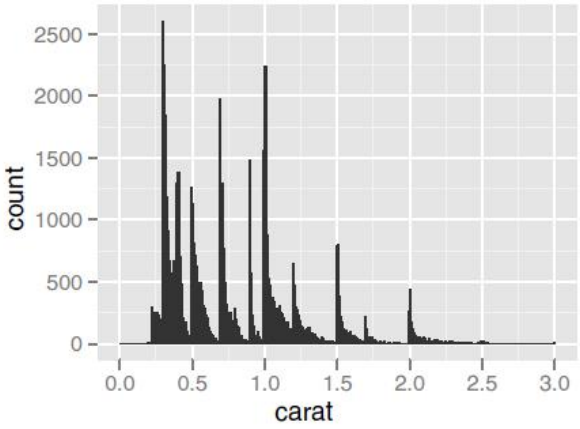
```
qplot(carat, data = diamonds, geom = "histogram",
binwidth = 1, xlim = c(0,3)) #横坐标只显示0到3

qplot(carat, data = diamonds, geom = "histogram",
binwidth = 0.1, xlim = c(0,3))

qplot(carat, data = diamonds, geom = "histogram",
binwidth = 0.01, xlim = c(0,3))
```





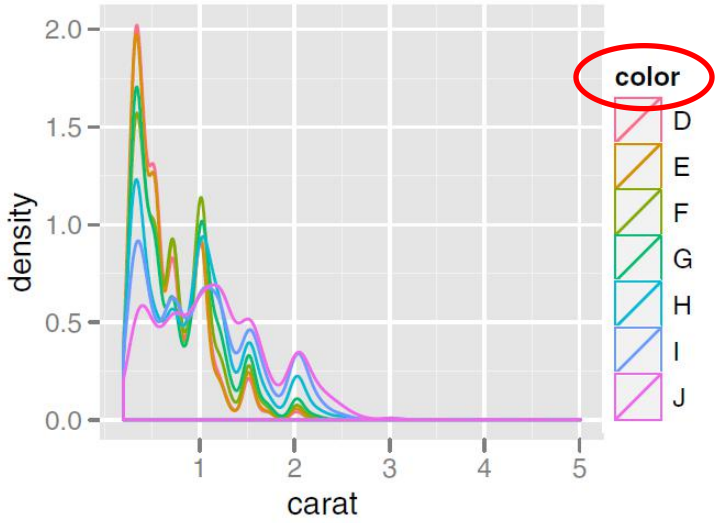


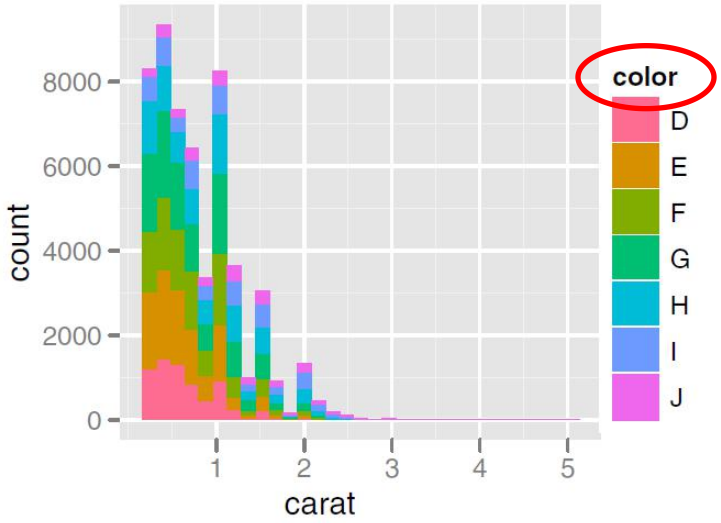
直方图和密度曲线图（5/5）

- 要在不同分组之间对分布进行对比，只需再加上一个**图形映射**（aesthetic mapping）

qplot(carat, data = diamonds, geom = "density", **colour =** color)

qplot(carat, data = diamonds, geom = "histogram", **fill =** color)



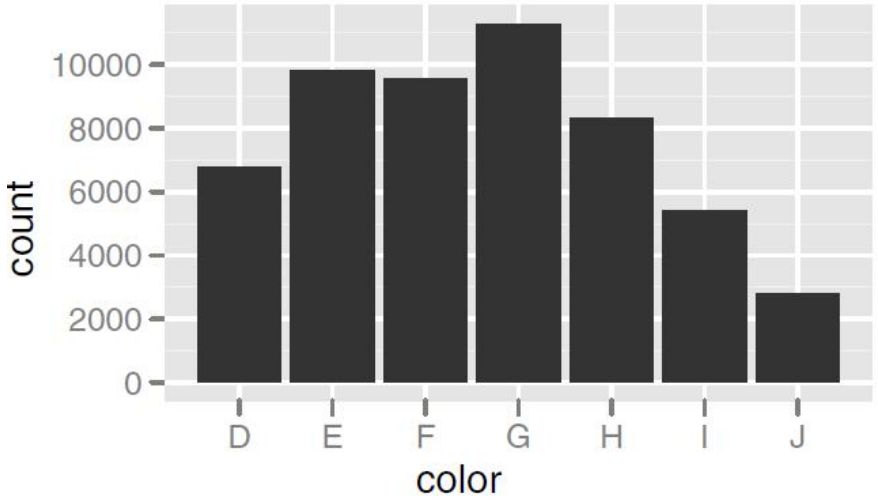


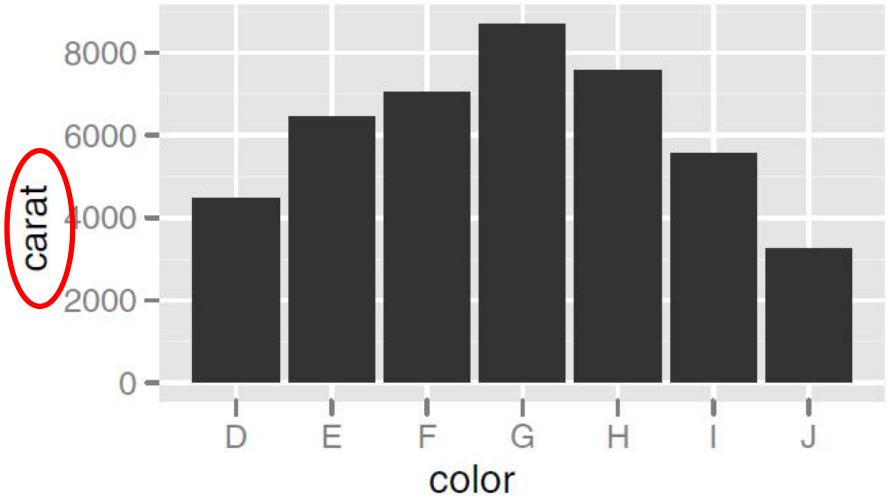
条形图（bar charts）（1/2）

- 条形图几何对象会计算每一个分类下的观测数量，因此不需要像在基础绘图中安逸预先对数据进行处理。
- 如果数据已经进行了汇总，或者想用其他方式对数据进行分组处理，那么可以使用weight几何对象。

条形图（bar charts）（2/2）

```
qplot(color, data = diamonds, geom = "bar")
#下例是按重量加权的条形图；纵坐标改为crart
qplot(color, data = diamonds, geom = "bar",
weight = carat) + scale_y_continuous("carat")
```





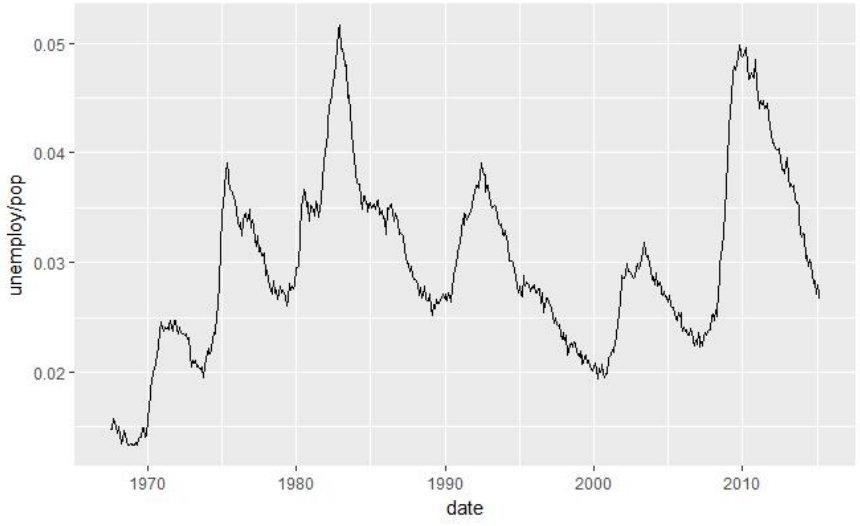
时间序列的线条图（line plots）和路径图（path plots）（1/3）

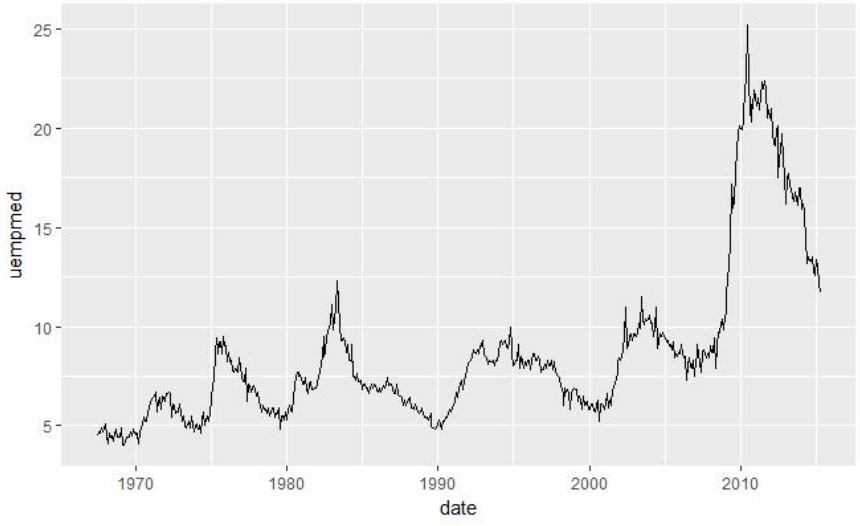
- 线条图和路径图常用于可视化时间序列数据。
- **线条图**的x轴一般是时间，它展示了单个变量随时间变化的情况。
- **路径图**展示了两个变量随时间联动的情况，时间反映在点的顺序上。

时间序列的线条图和路径图（2/3）
例：线条图

- economics包含美国过去几十年的经济数据

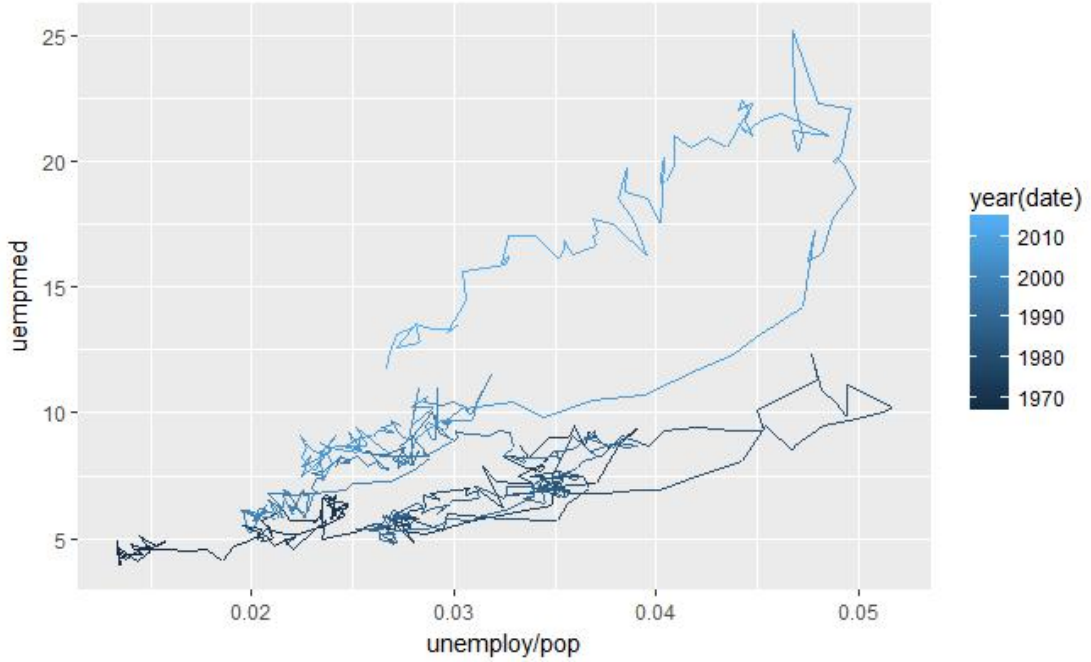
```
qplot(date, unemploy / pop, data = economics, geom = "line") #失业比例
qplot(date, uempmed, data = economics, geom = "line") #失业时间（星期数）中位数
```





时间序列的线条图和路径图（3/3）
例：路径图

```
year <- function(x) as.POSIXlt(x)$year + 1900
qplot(unemploy / pop, uempmed, data = economics, geom = "path", colour = year(date))
```



- 可以看出失业率和失业时间长度是相关的。
- 10年左右失业时间长度与失业率比较高；近几年有所降低。

分面（faceting）（1/2）

- 除了利用颜色和形状来比较不同分组，还可以用**分面**：将数据分割成若干子集，然后创建一个图形的矩阵。
- qplot()默认的分面方法是拆分成若干个窗格，通过形如facets = row-var ~ col_var的表达式进行指定。如果只想指定一行或一列，可以使用“.”作为占位符，例如row-var ~ .会创建一个单列多行的图形矩阵。

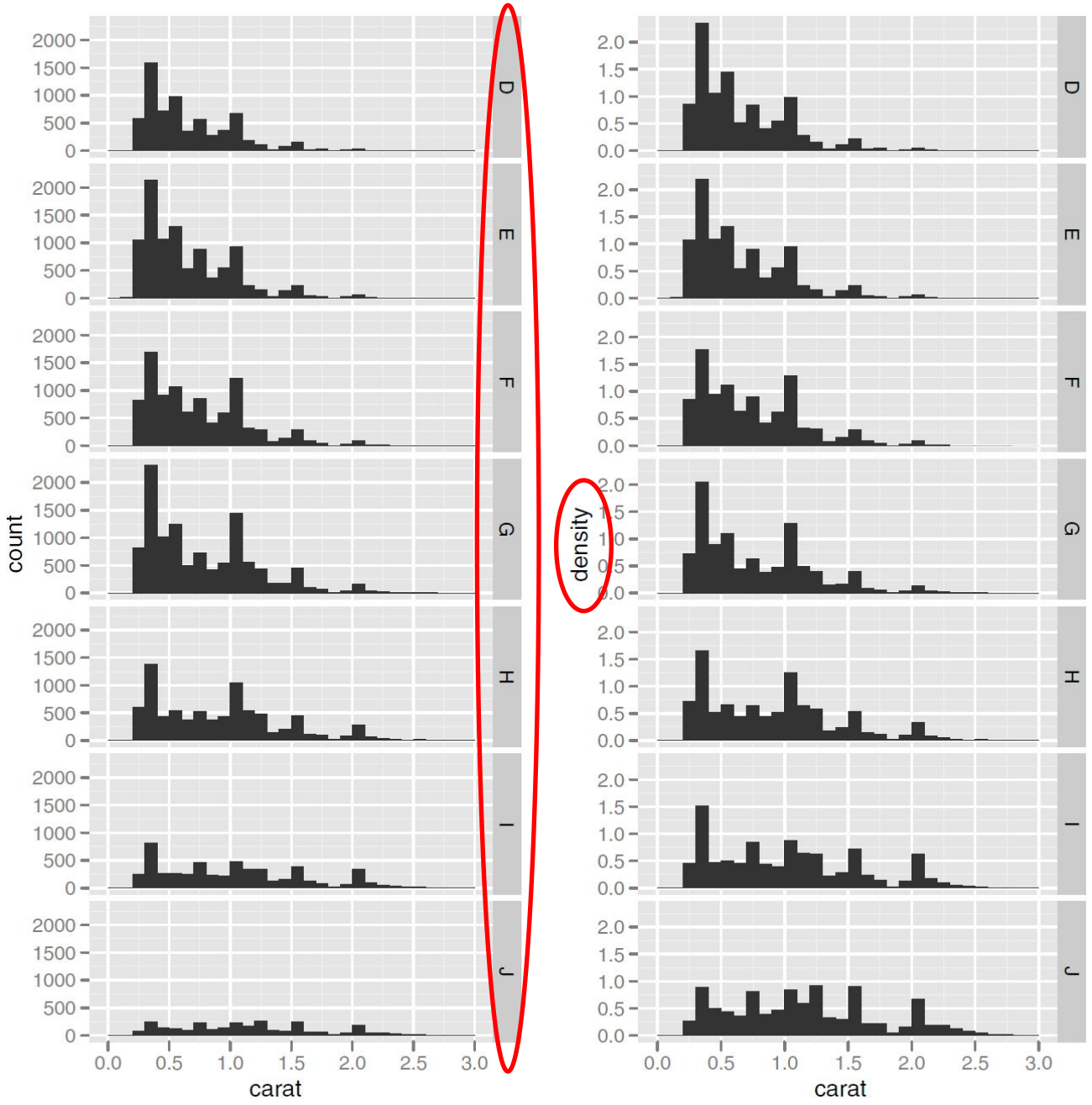
分面（2/2）

#针对属性color，创建单列多行分面

```
qplot(carat, data = diamonds, facets = color ~ .,
geom = "histogram", binwidth = 0.1, xlim = c(0, 3))
```

#..density..表示将密度（density: proportions of the whole）而不是频数映射到y轴。使用密度可以使得比较不同组的分布时不会受该组样本量大小的影响。

```
qplot(carat, ..density.., data = diamonds, facets =
color ~ ., geom = "histogram", binwidth = 0.1, xlim
= c(0, 3))
```



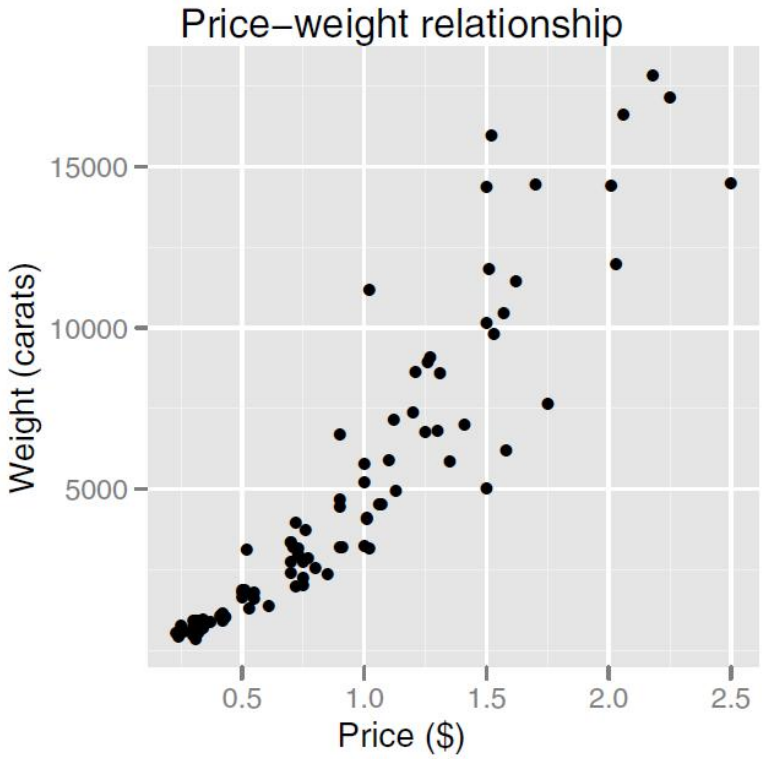
- 针对不同的color分面。
- 左图展示的是频数，右图是密度。
- 高质量的钻石（颜色D）在小尺寸的分布是偏斜的，随着品质下降，重量分布变得越来越平坦。

其他选项（1/4）

- xlim, ylim：设置x轴和y轴的显示区间，例如，xlim=c(0, 20) 和 ylim=c(-0.9, -0.5)
- log：log="x" 表示对x轴取对数；log="xy" 表示对x轴和y轴都取对数。
- main：图形主标题。可以是字符串也可以是数学表达式。
- xlab, ylab：设置x和y轴的标签文字，可以是字符串或数学表达式。

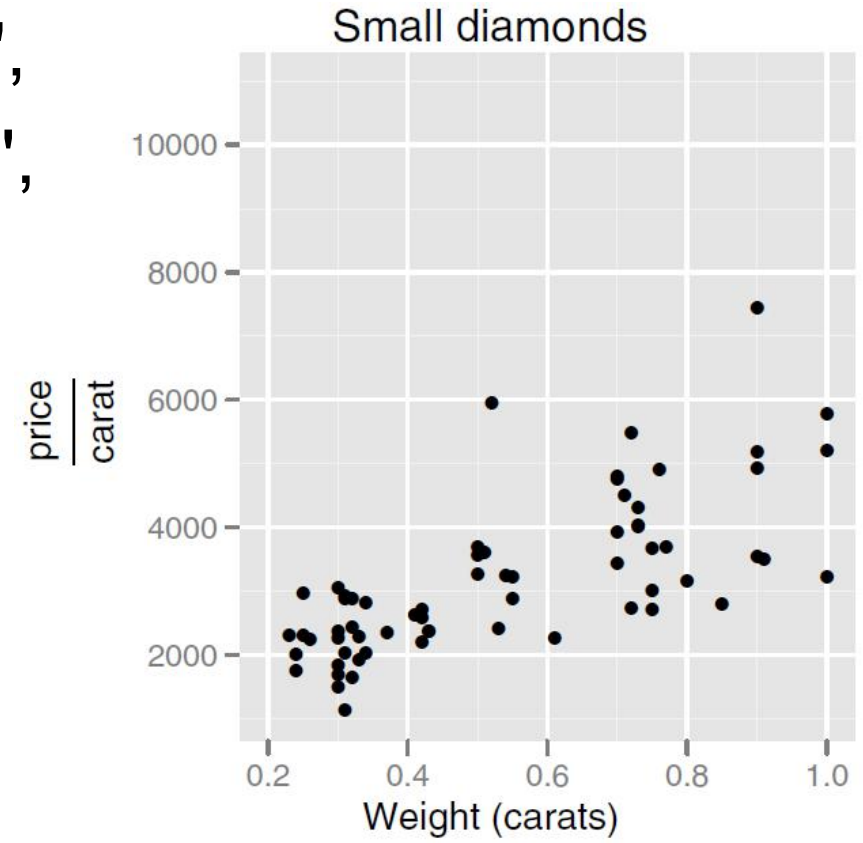
其他选项（2/4）

```
qplot(
  carat, price, data = dsmall,
  xlab = "Weight (carats)", ylab = "Price ($)",
  main = "Price-weight relationship"
)
```



其他选项（3/4）

```
qplot(
   carat, price/carat, data = dsmall,
   ylab = expression(frac(price,carat)),
   xlab = "Weight (carats)",
   main="Small diamonds",
   xlim = c(.2,1)
)
```



其他选项（4/4）

qplot(carat, price, data = dsmall, log = "xy")