Data Analysis Tools and Practice(Using R)

2020.03.26

ggplot2画图

北京大学 软件与微电子学院
School of Software and Microelectronics, Peking University

Huiping Sun(孙惠平)
sunhp@ss.pku.edu.cn

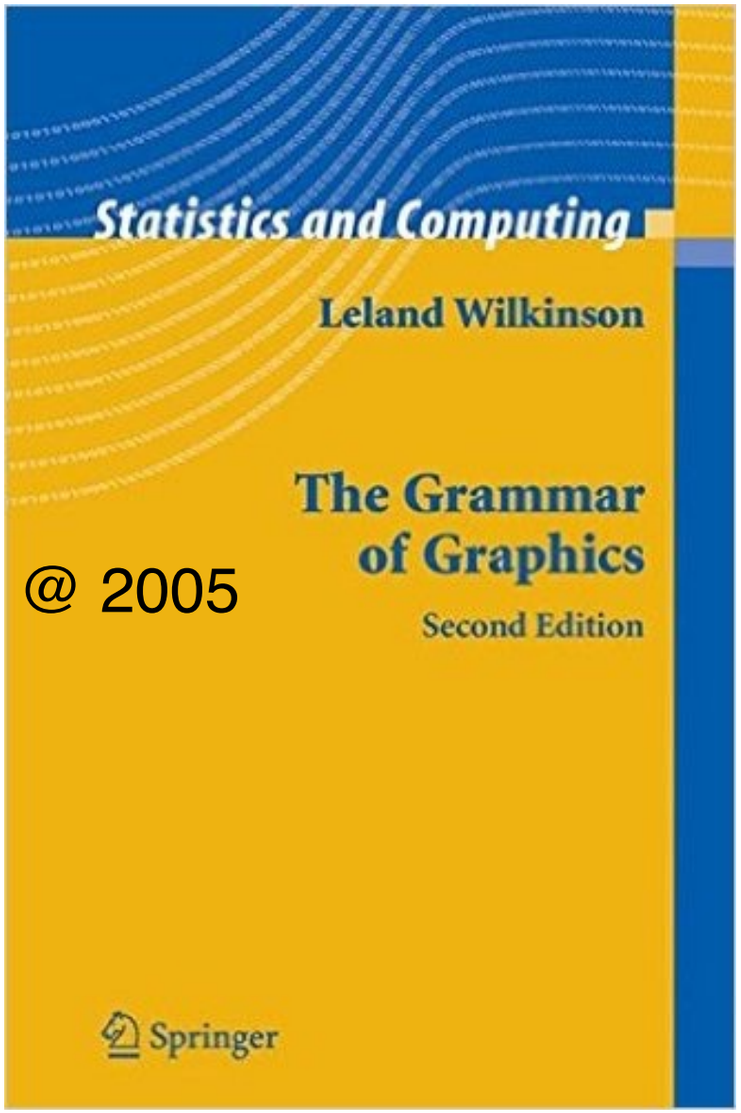
上次课程内容回顾

- 图例：
  - ✴ 坐标; 边界标注; 标注(mar);horiz=TRUE;
- 线图：
  - ✴ grid(); abline(); line(); lm(); arrows();
- 条形图：
  - ✴ 堆积(beside); horiz=TRUE;
  - ✴ 显示数字; 宽度、颜色和边界; 显示标注; 增加误差线
- 散点图：
  - ✴ point();type="n"; corplot(); 增加抖动;
- 其余：
  - ✴ point(); par(); corplot(); axis(); mtext(); jitter();

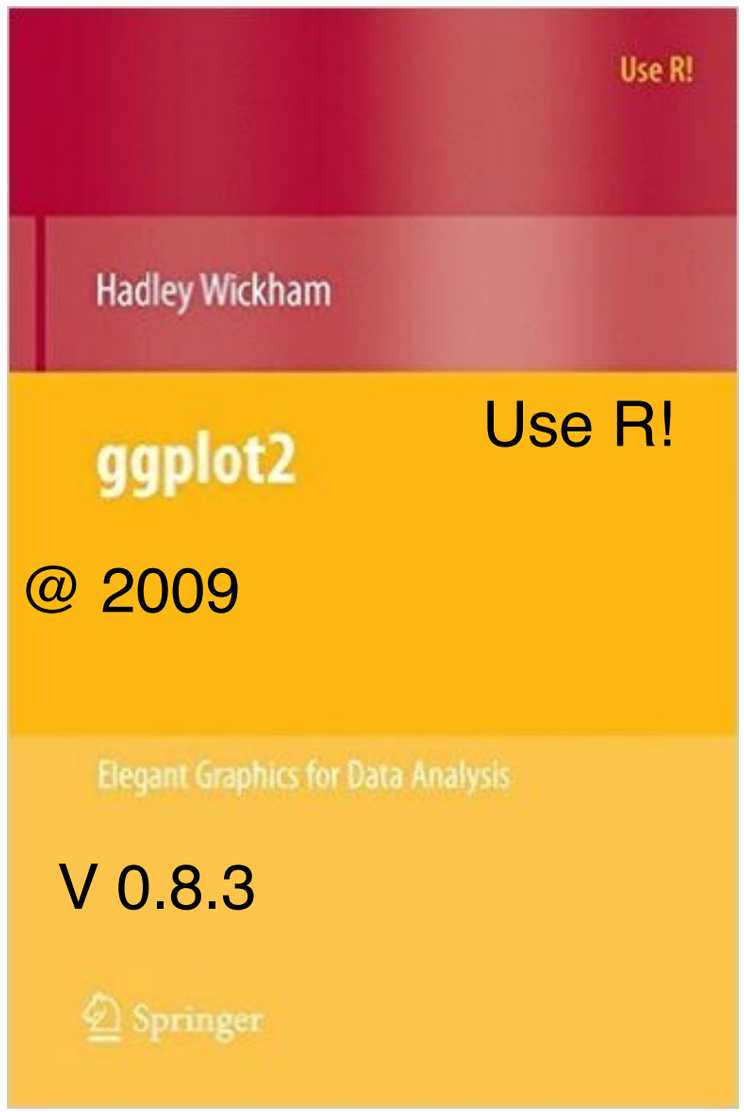
ggplot2简介

ggplot2

https://cran.r-project.org/web/packages/ggplot2/index.html V 2.2.1

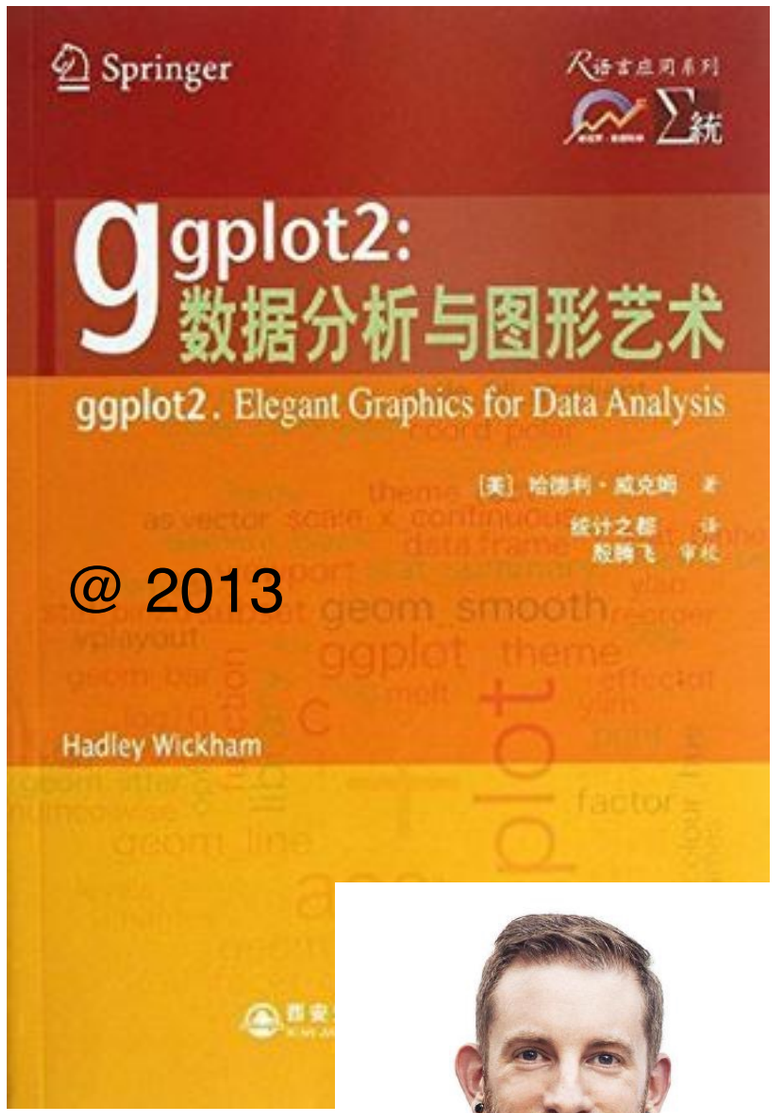


@ 2005



@ 2009

V 0.8.3



@ 2013

- graphics、grid、lattice
- ggplot2

http://hadley.nz/

为什么使用ggplot2

- 函数繁杂，语法复杂
- “笔纸”工作方式，不能增减
- 自动化低
- 主次不分

忘记一切

- 有理论基础，支持一套图形语法
- 采用图层的设计方式，可增减
- 媲美商业数据化软件的作图效果
- 使用简单，定制容易（主题）

从头开始

```
install.packages("ggplot2")
```

图形语法

- 数据（data）

  映射（mapping）⟷ 图形属性（aesthetic attributes）

- 几何对象（geometric object）
- 统计变换（statistical transformation s）
- 标度（scale）
- 坐标系（coordinate system）
- 分面（facet）

qplot

钻石数据集

carat	cut	color	clarity	depth	table	price	x	y	z
0.2	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.2	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.2	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.2	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.2	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
0.2	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48

carat：克拉重量

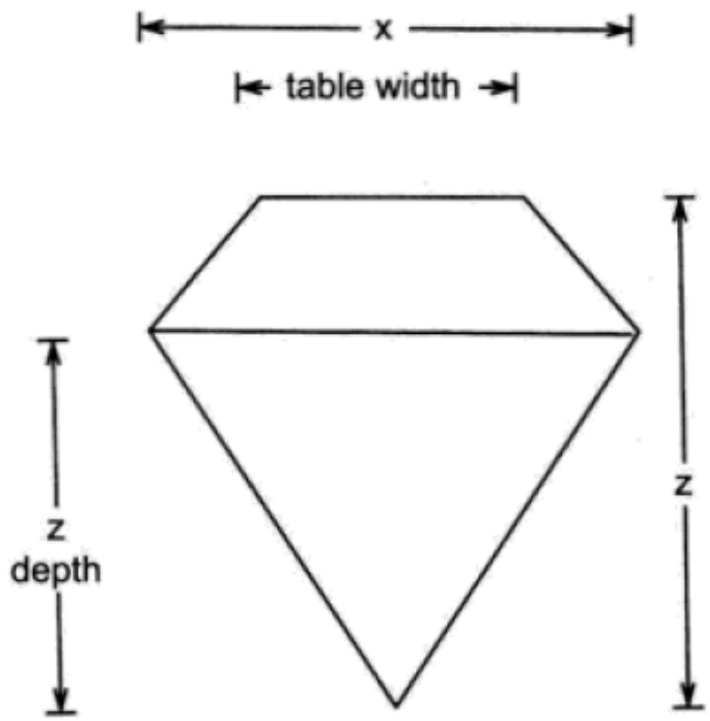
cut：切工

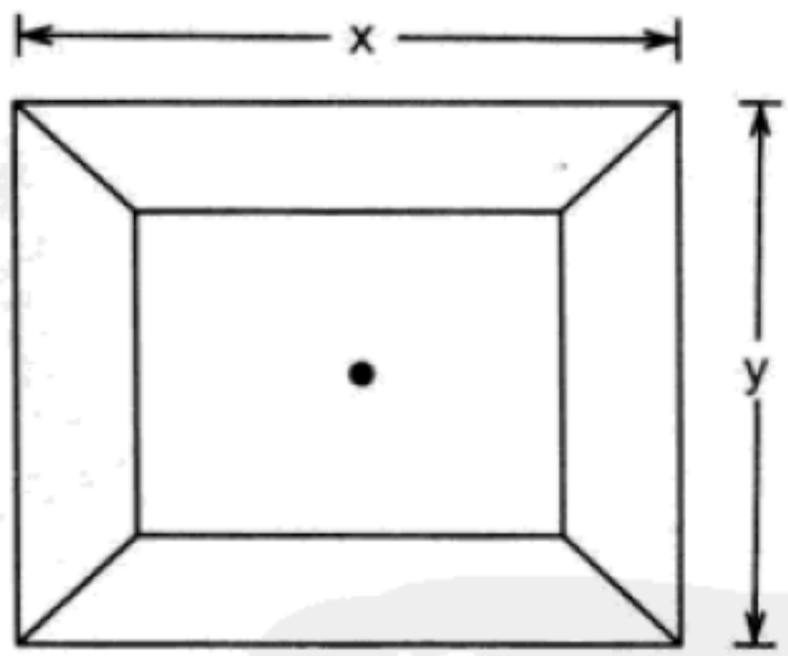
color：颜色

clarity：净度

depty：深度

table：钻面宽度

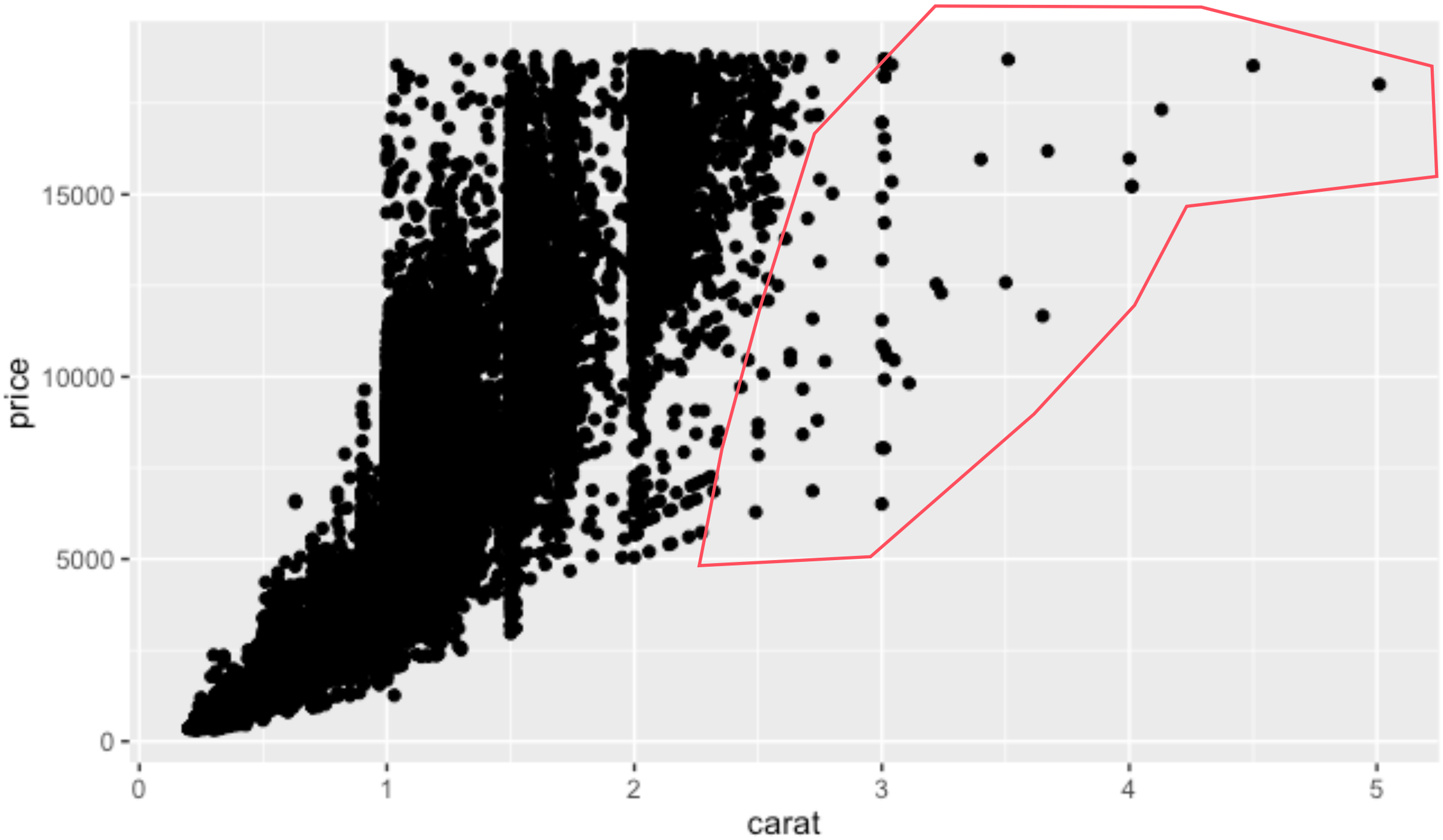




depth = z depth / z * 100
table = table width / x * 100

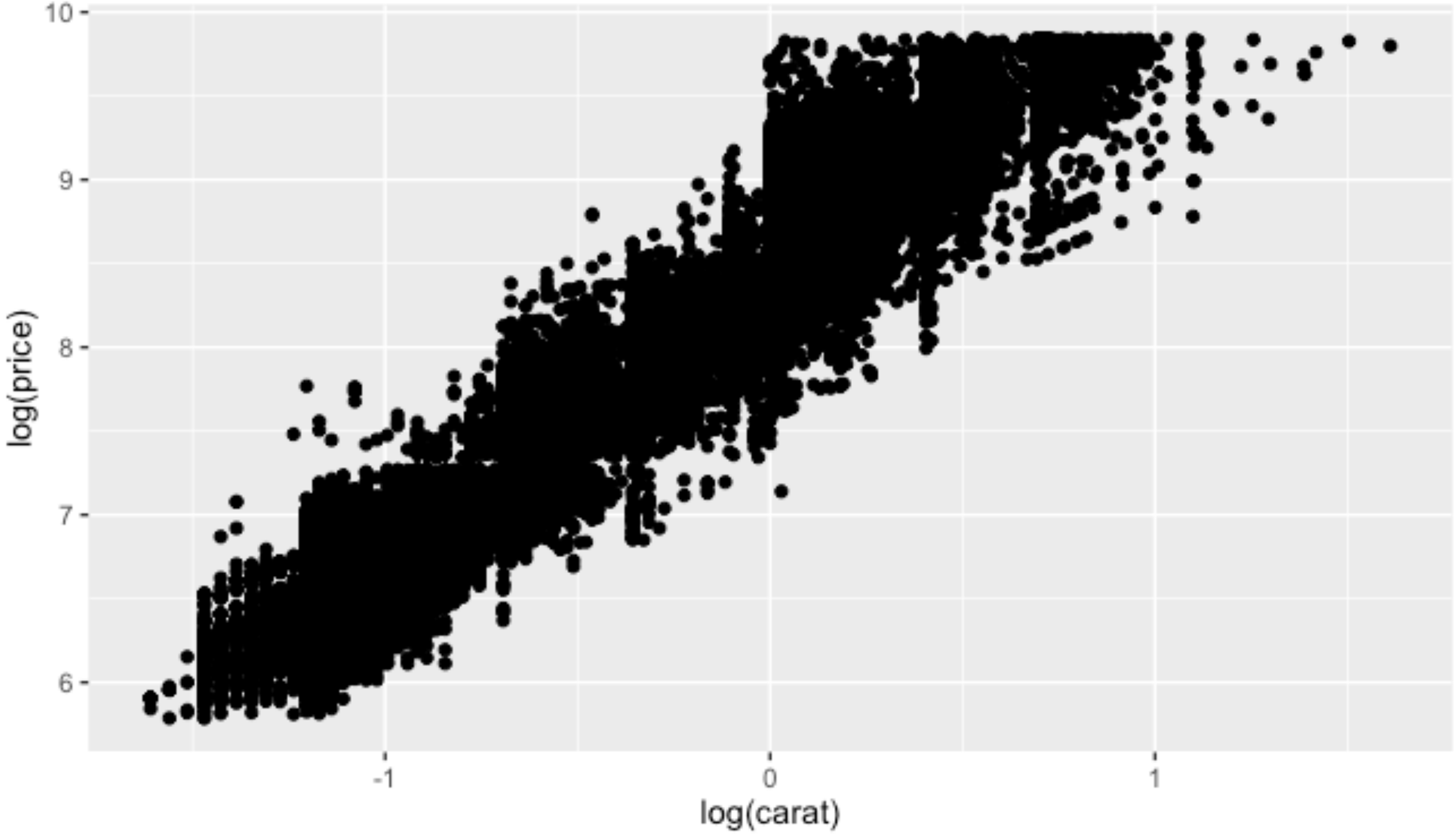
散点图

qplot(carat, price, data = diamonds)



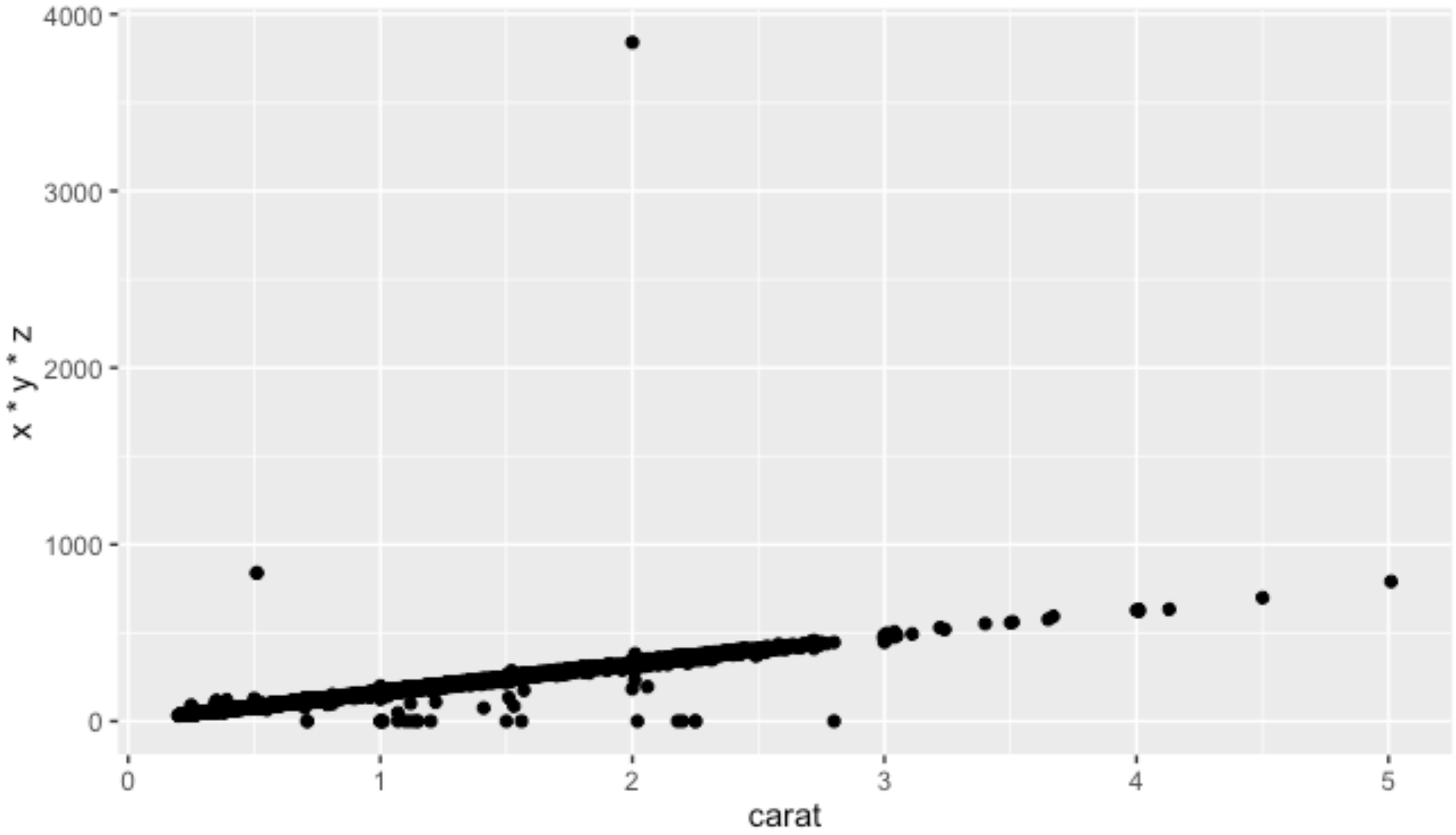
对数变换

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



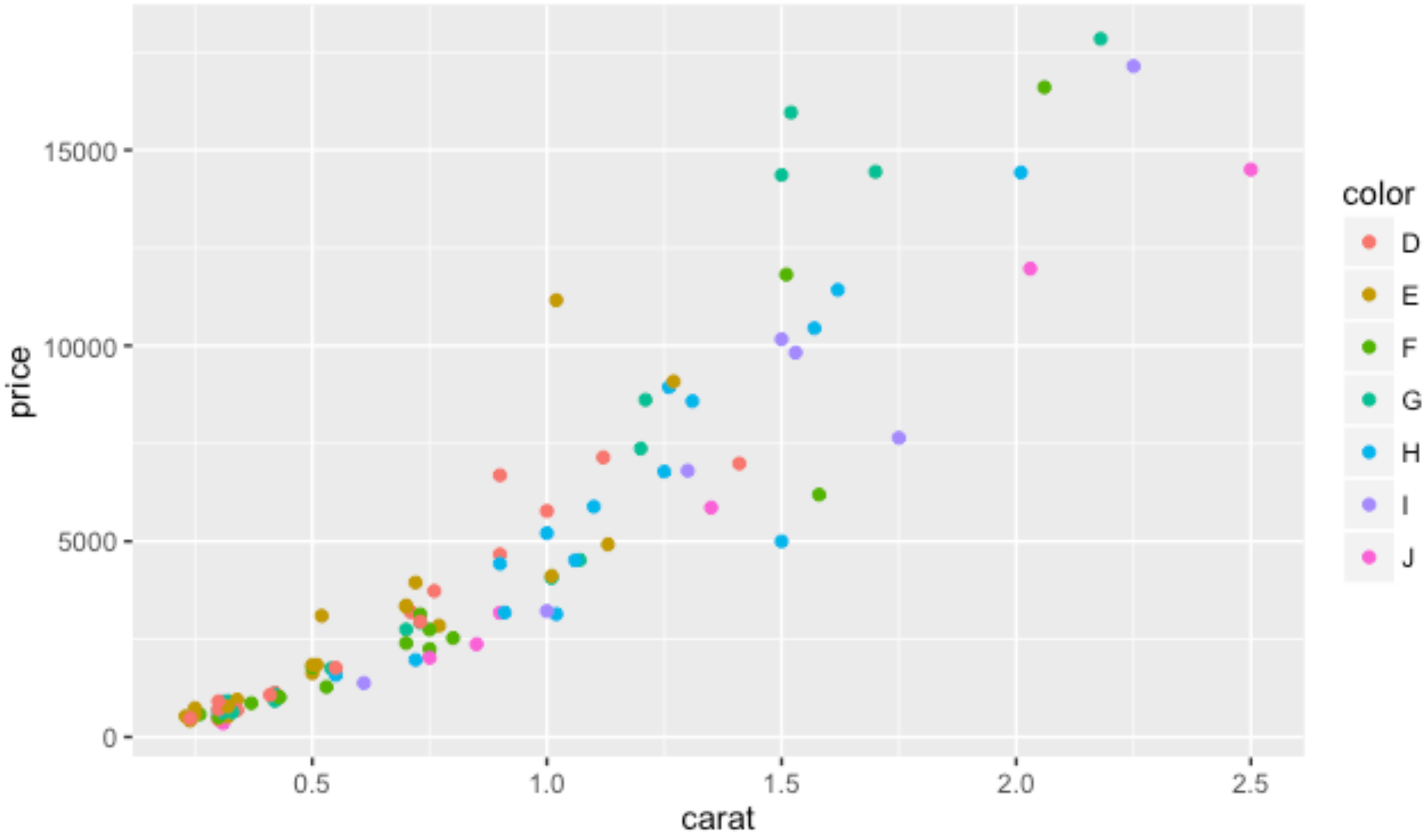
体积 vs. 体重

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



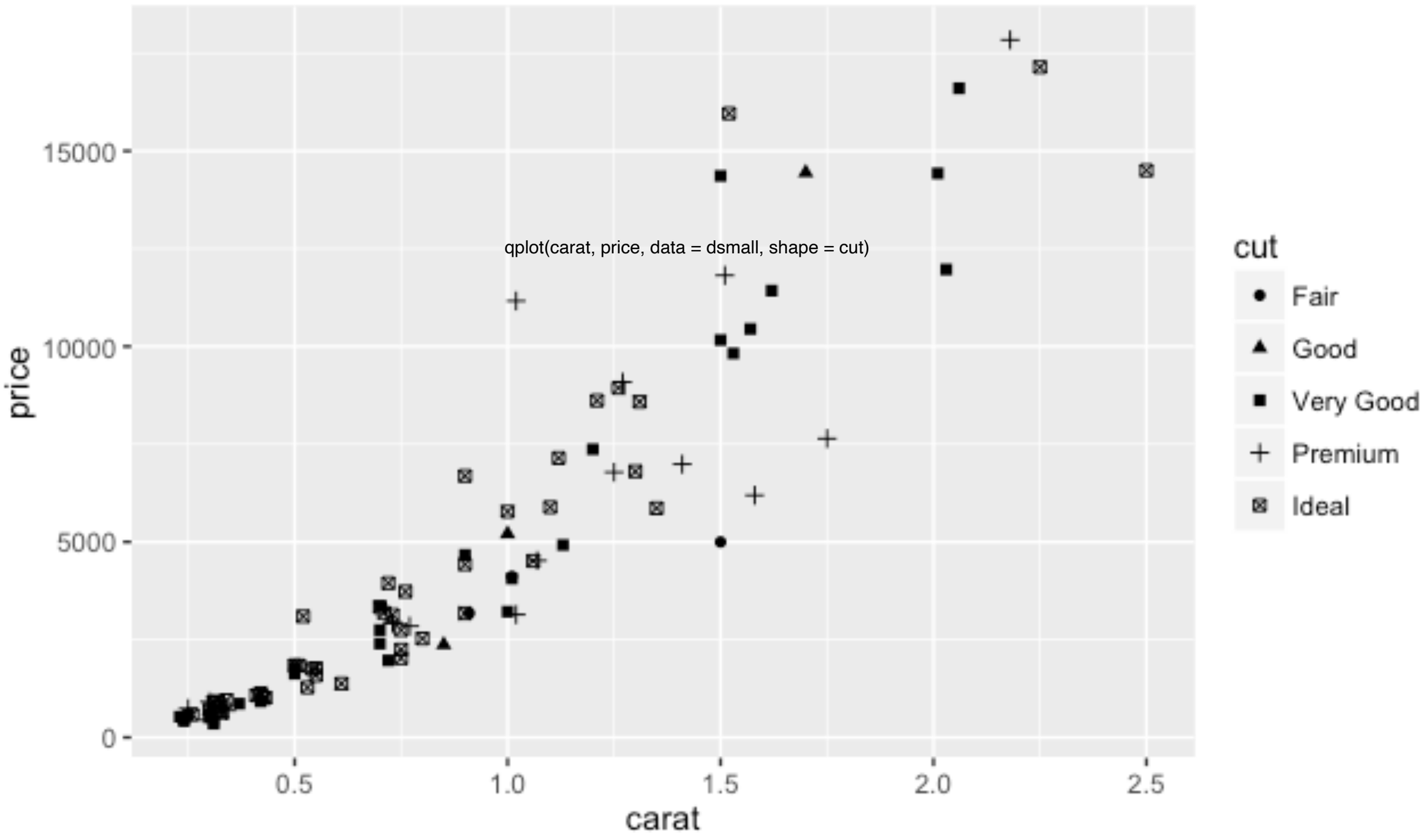
颜色属性

```
set.seed(1410)     dsmall <- diamonds[sample(nrow(diamonds), 100), ]

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



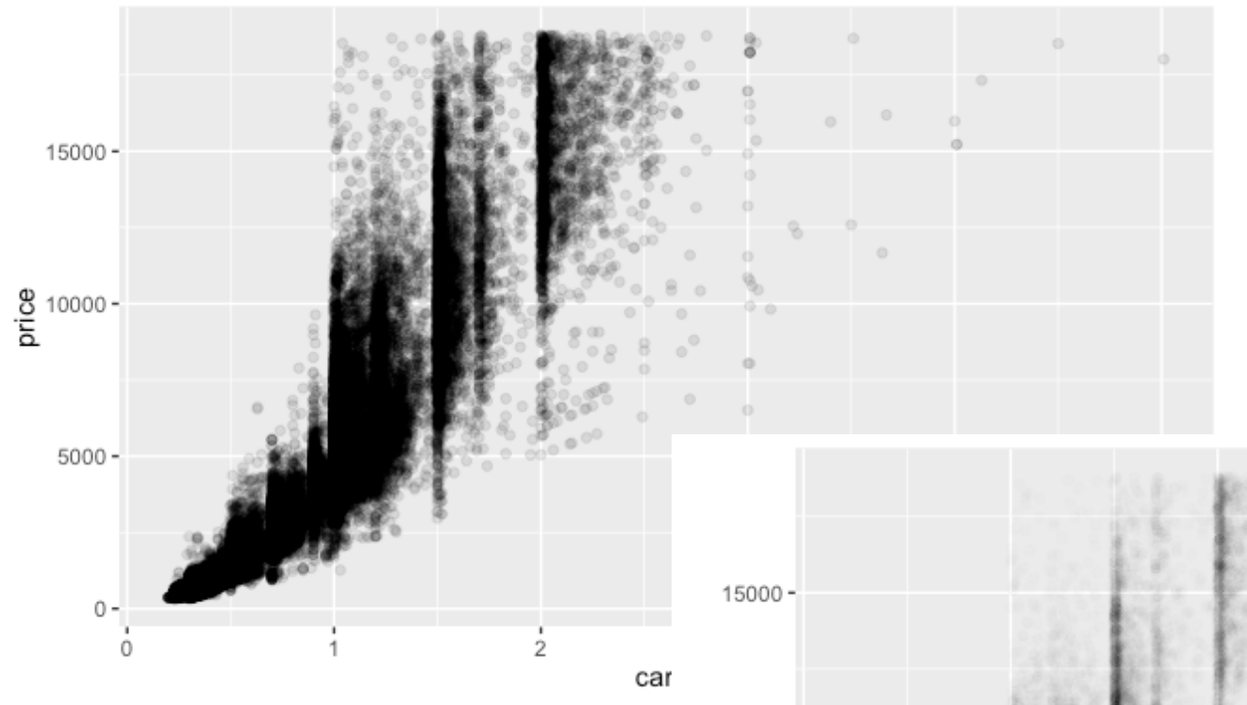
切工属性

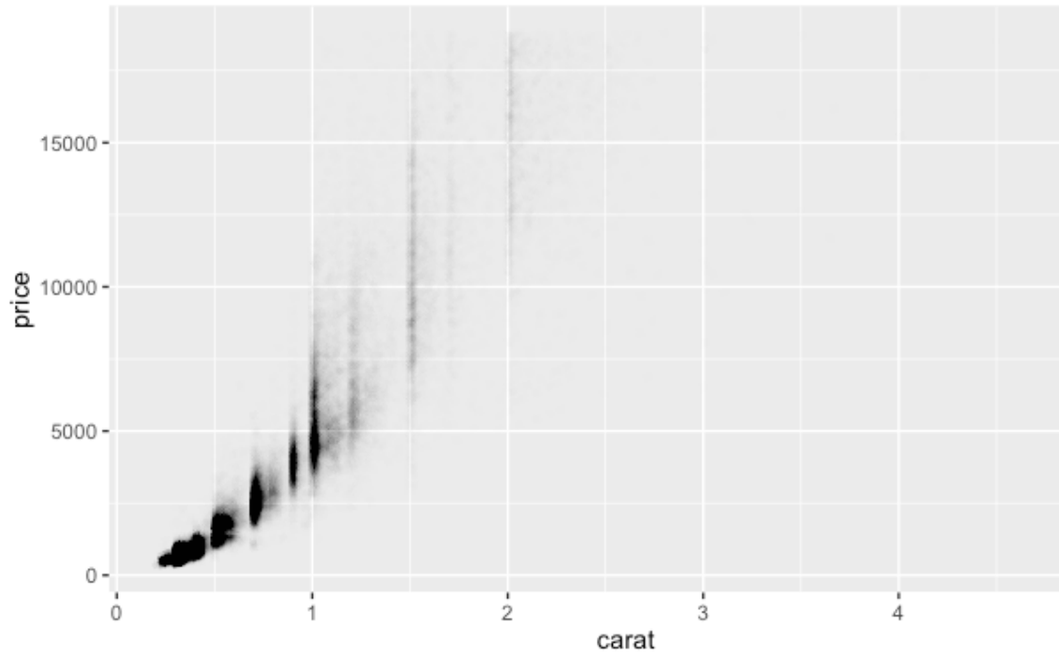
qplot(carat, price, data = dsmall, shape = cut)

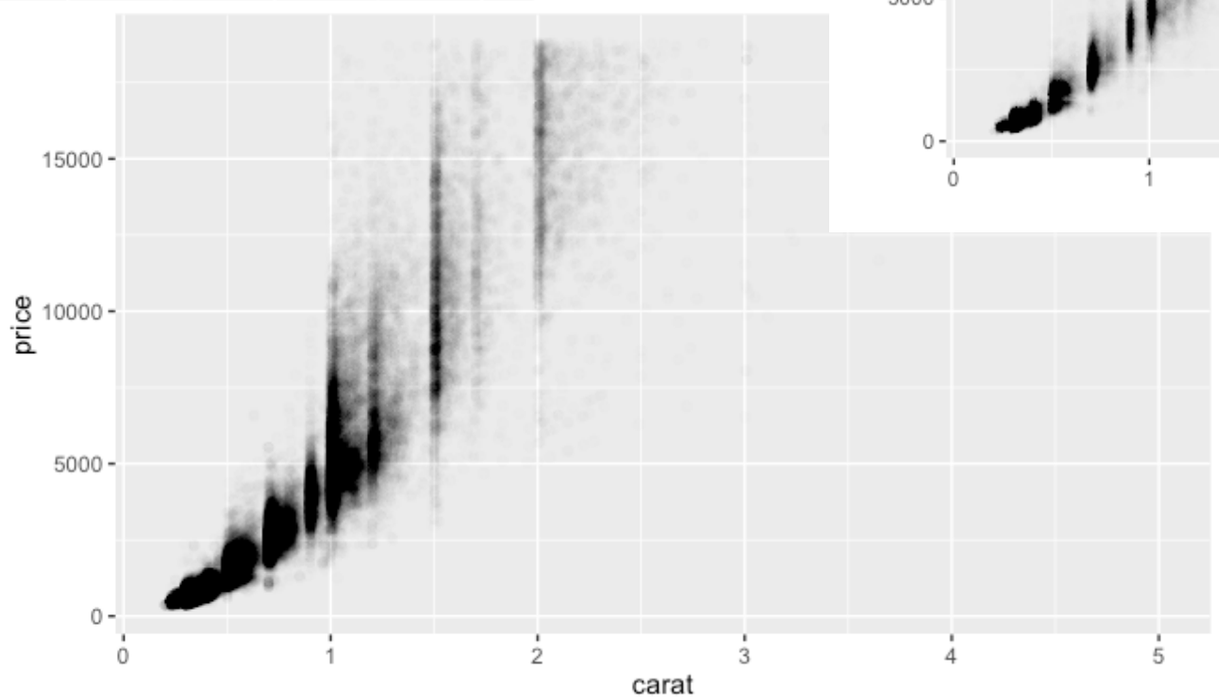


Alpha取值

```
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))
```







集合对象 geom

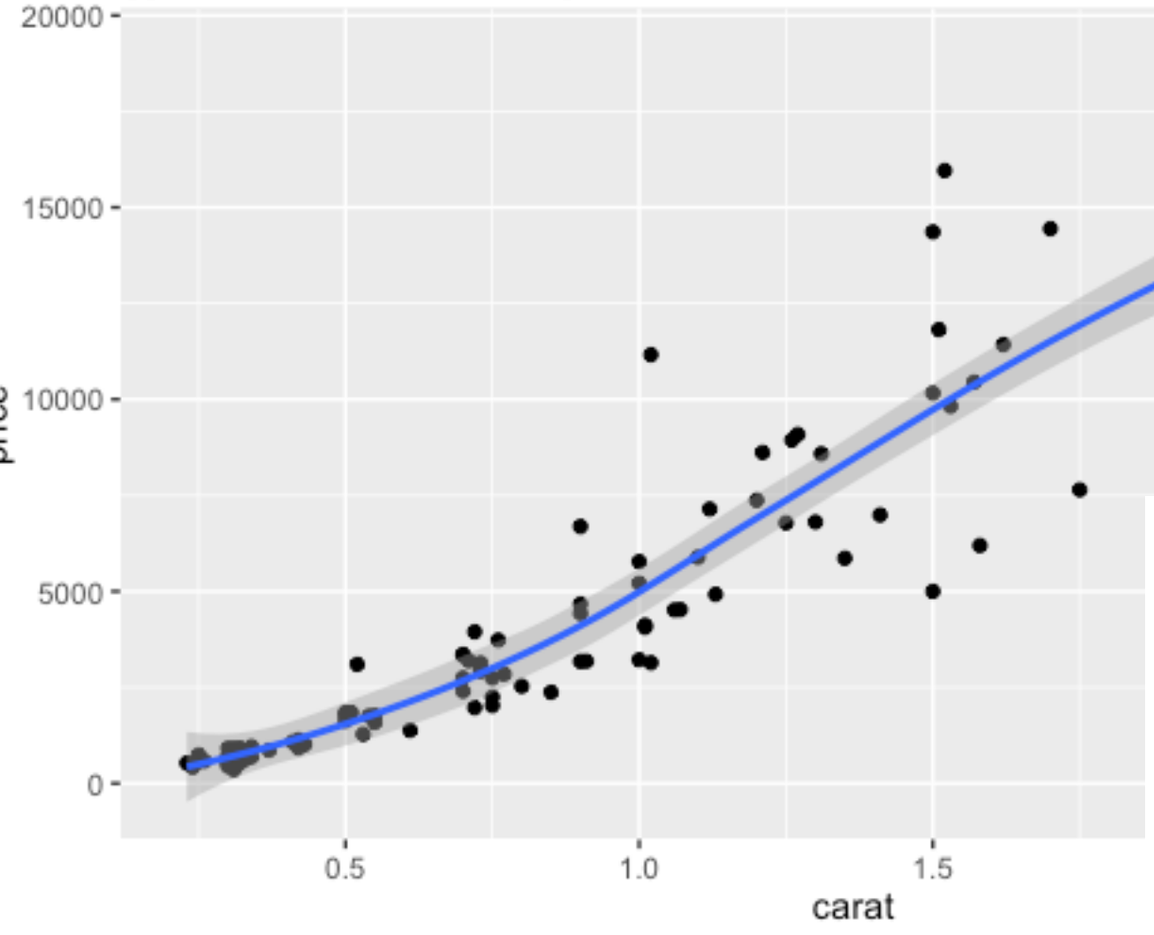
- point：散点图
- smooth：平滑曲线和标准误
- boxplot：箱线图

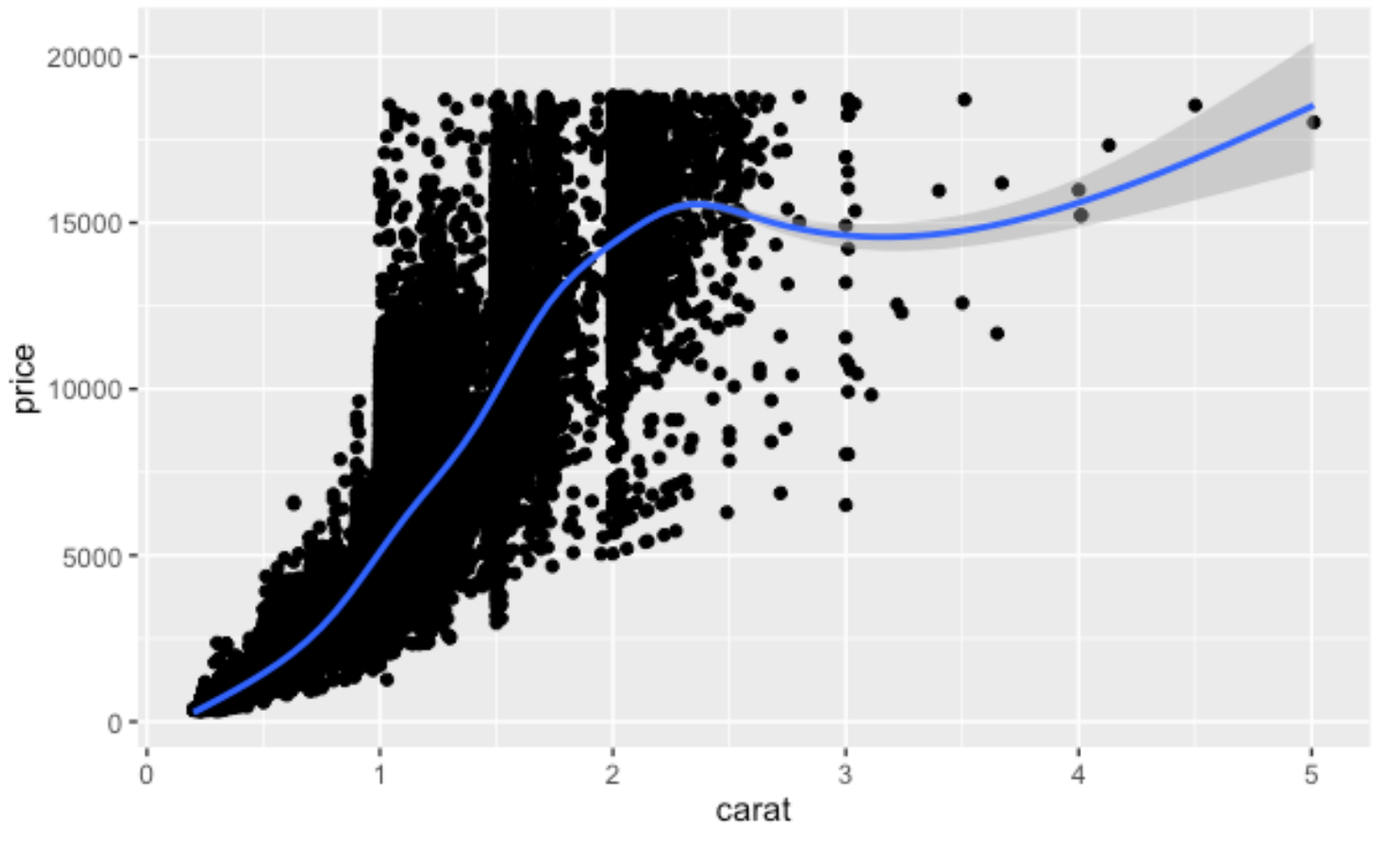
geom = “point”

- path、line：连线（曲线图、路径图）
- histogram：直方图
- freqpoly：频率多边形
- density：密度曲线
- bar：柱状图（条形图）

平滑曲线

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

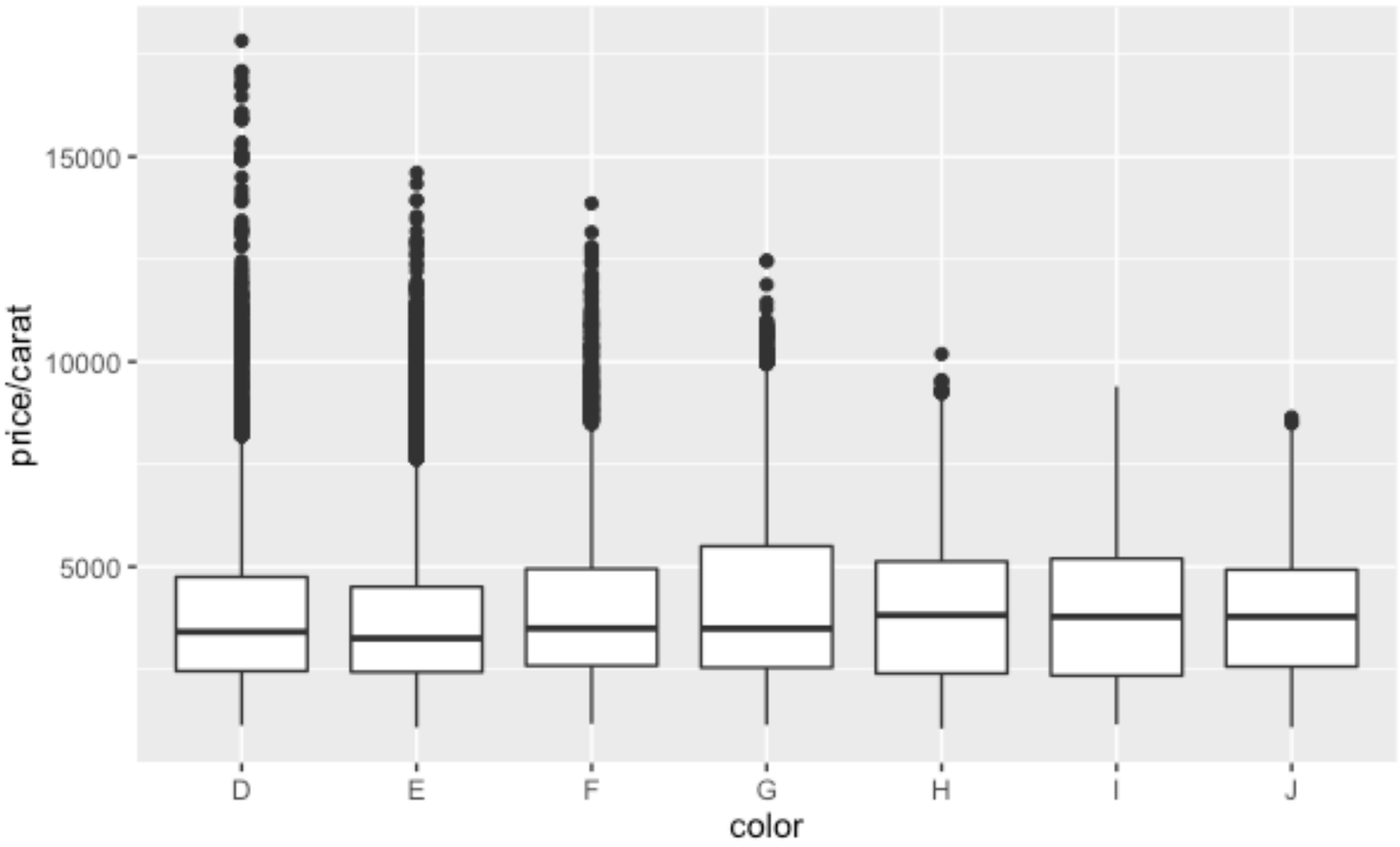





```
qplot(carat, price, data = diamonds, geom = c("point", "smooth"))
```

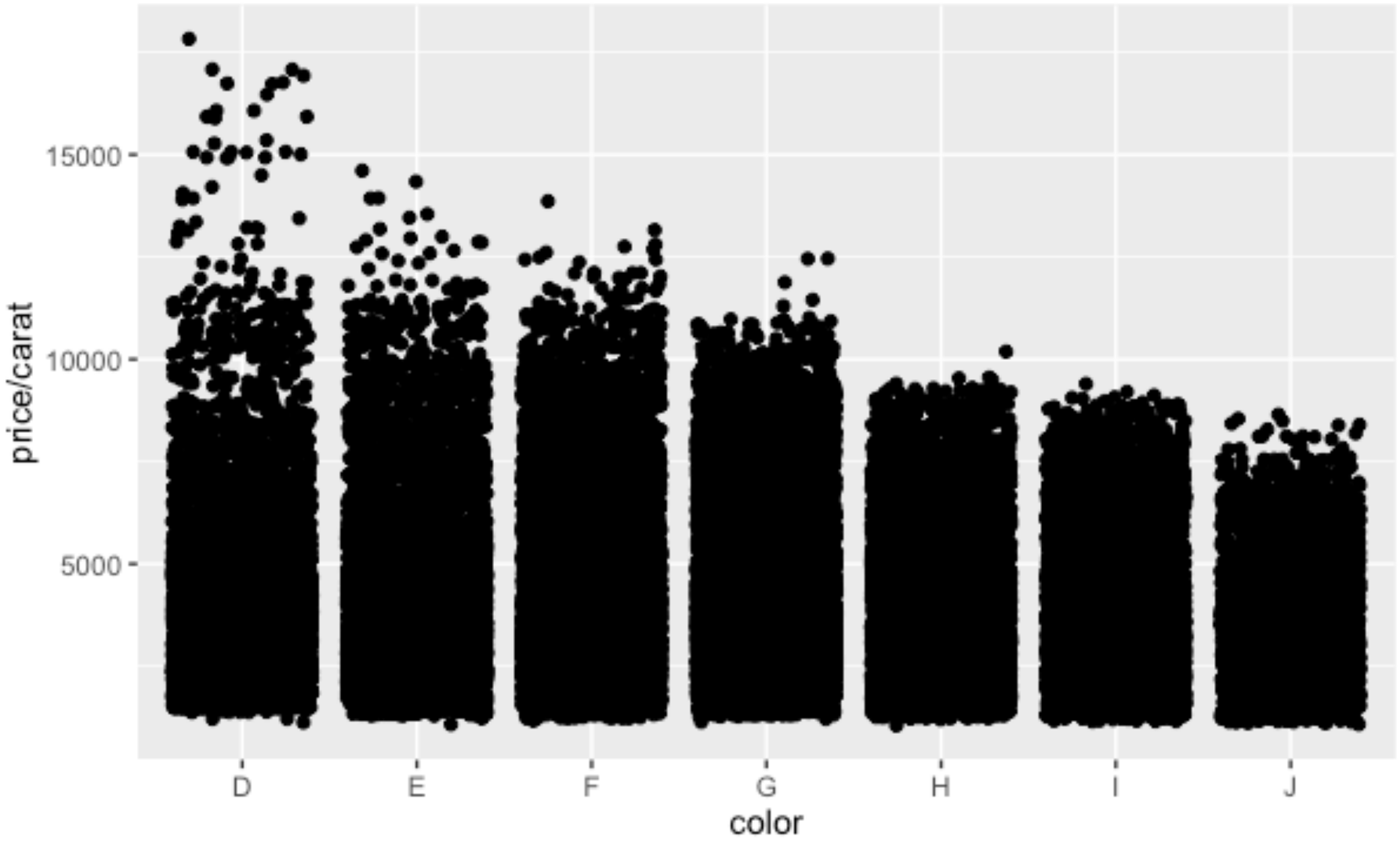
箱线图

qplot(color, price / carat, data = diamonds, geom = "boxplot")



扰动点图

qplot(color, price / carat, data = diamonds, geom = "jitter")

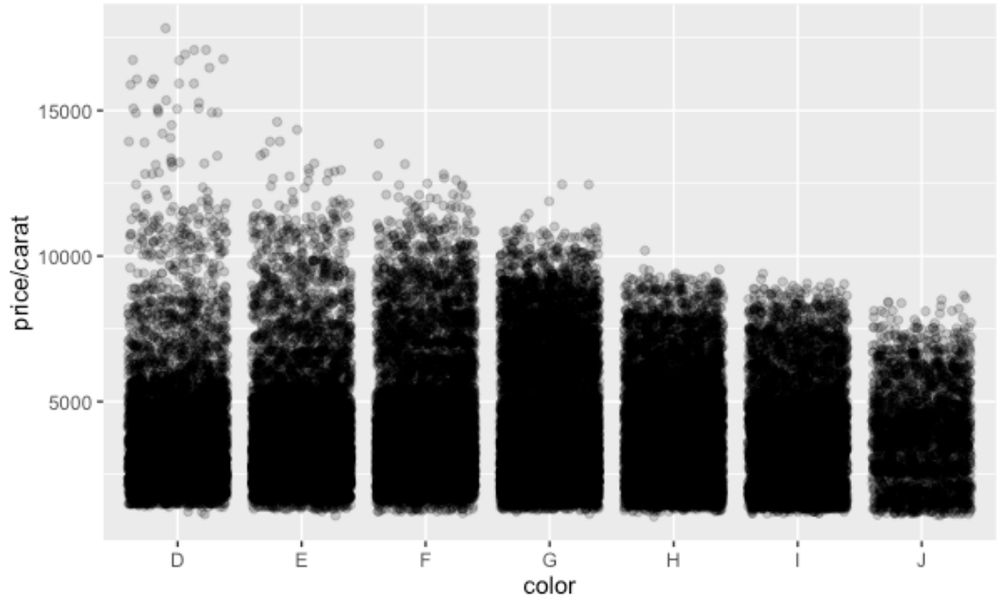


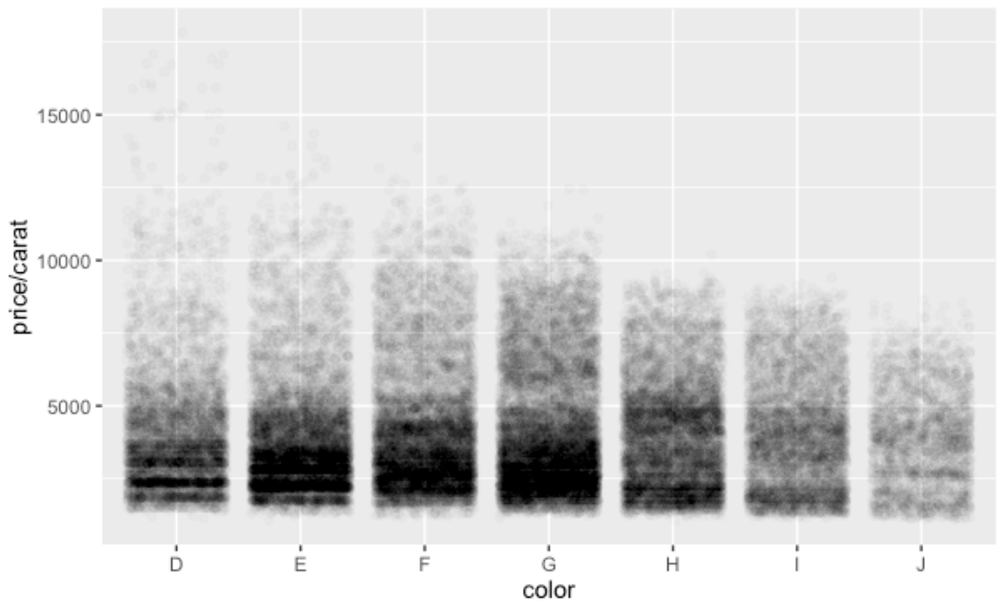
扰动点图的透明度

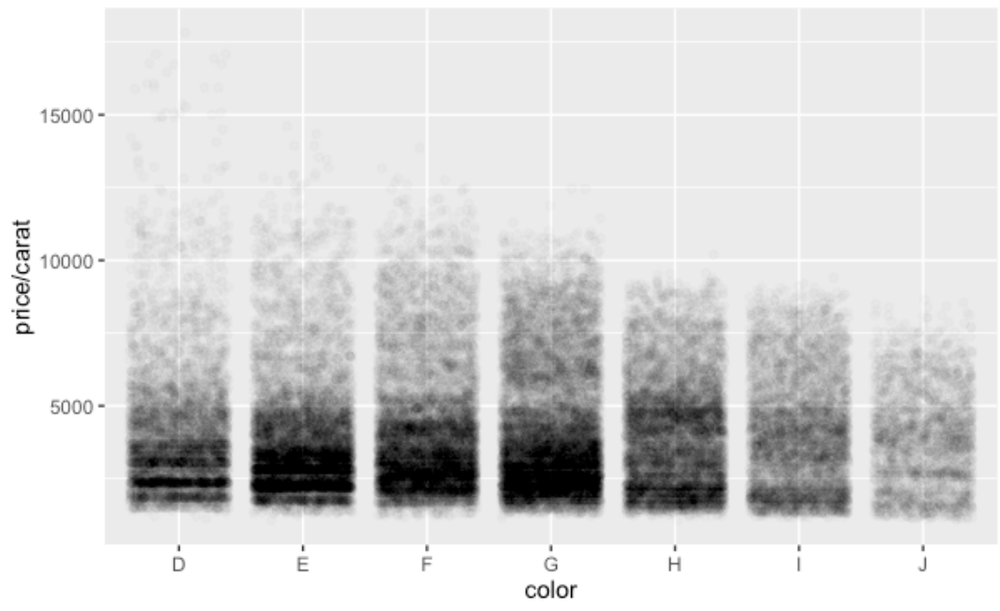
qplot(color, price / carat, data = diamonds,geom = "jitter",alpha = I(1 / 5))

qplot(color, price / carat, data = diamonds, geom = "jitter",alpha = I(1 / 50))

qplot(color, price / carat, data = diamonds, geom = "jitter",alpha = I(1 / 200))

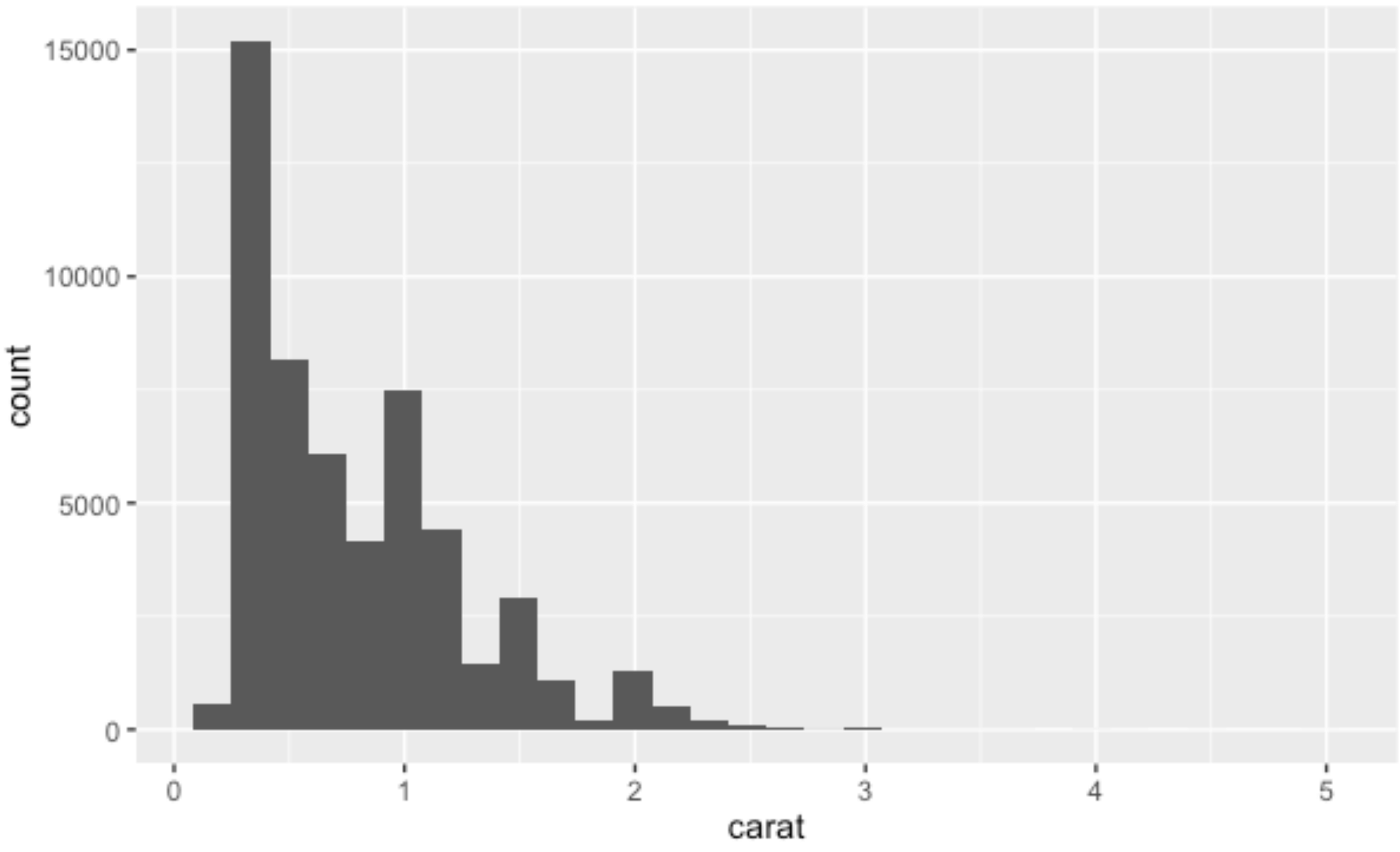






直方图

```
qplot(carat, data = diamonds, geom = "histogram")
```

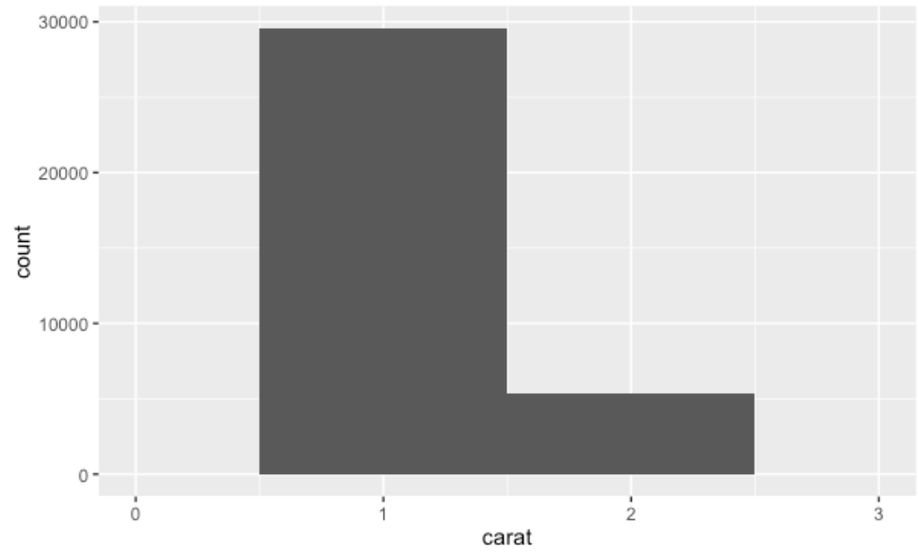


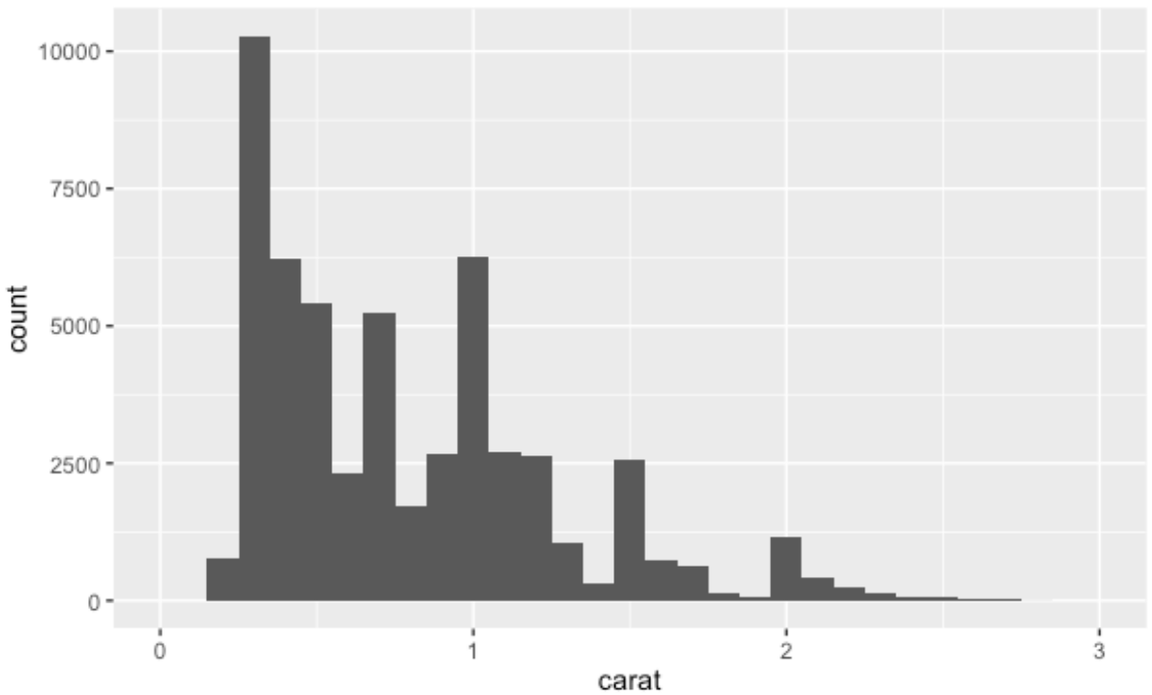
直方图的区间

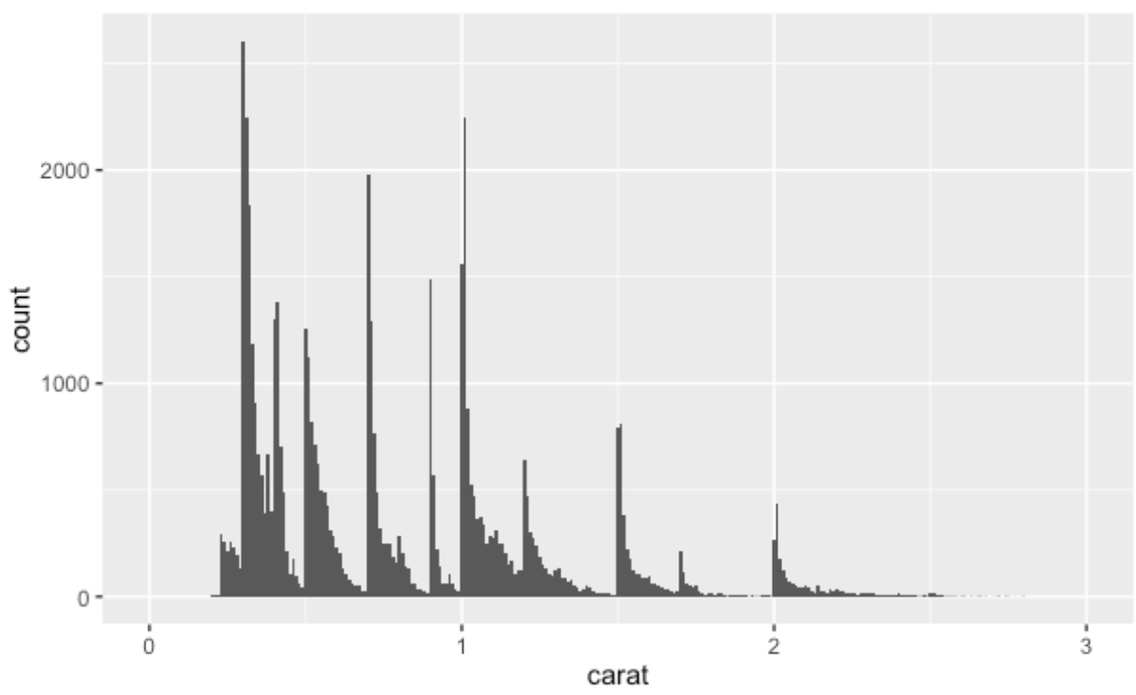
qplot(carat, data = diamonds, geom = "histogram", binwidth = 1, xlim = c(0,3))

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))

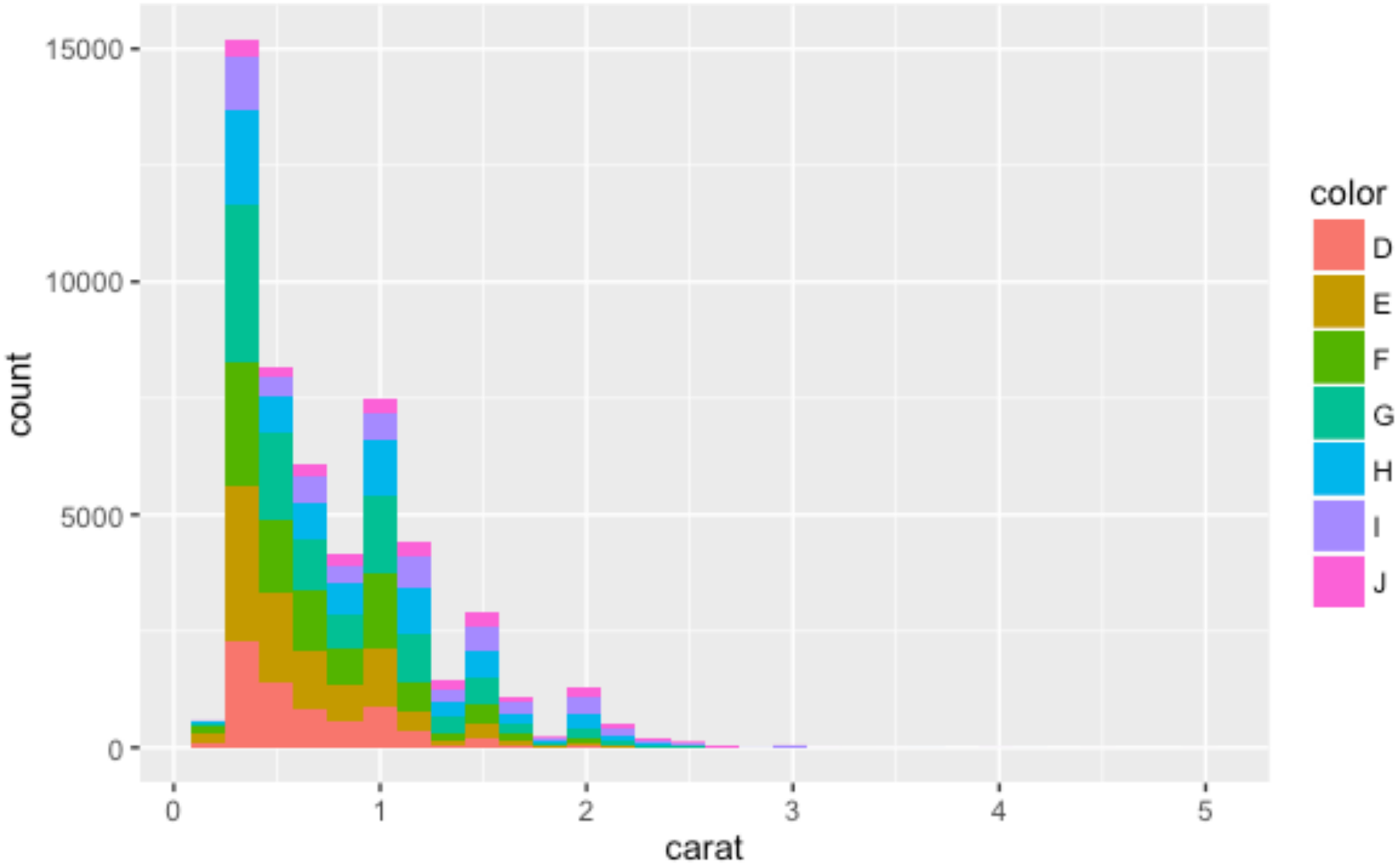






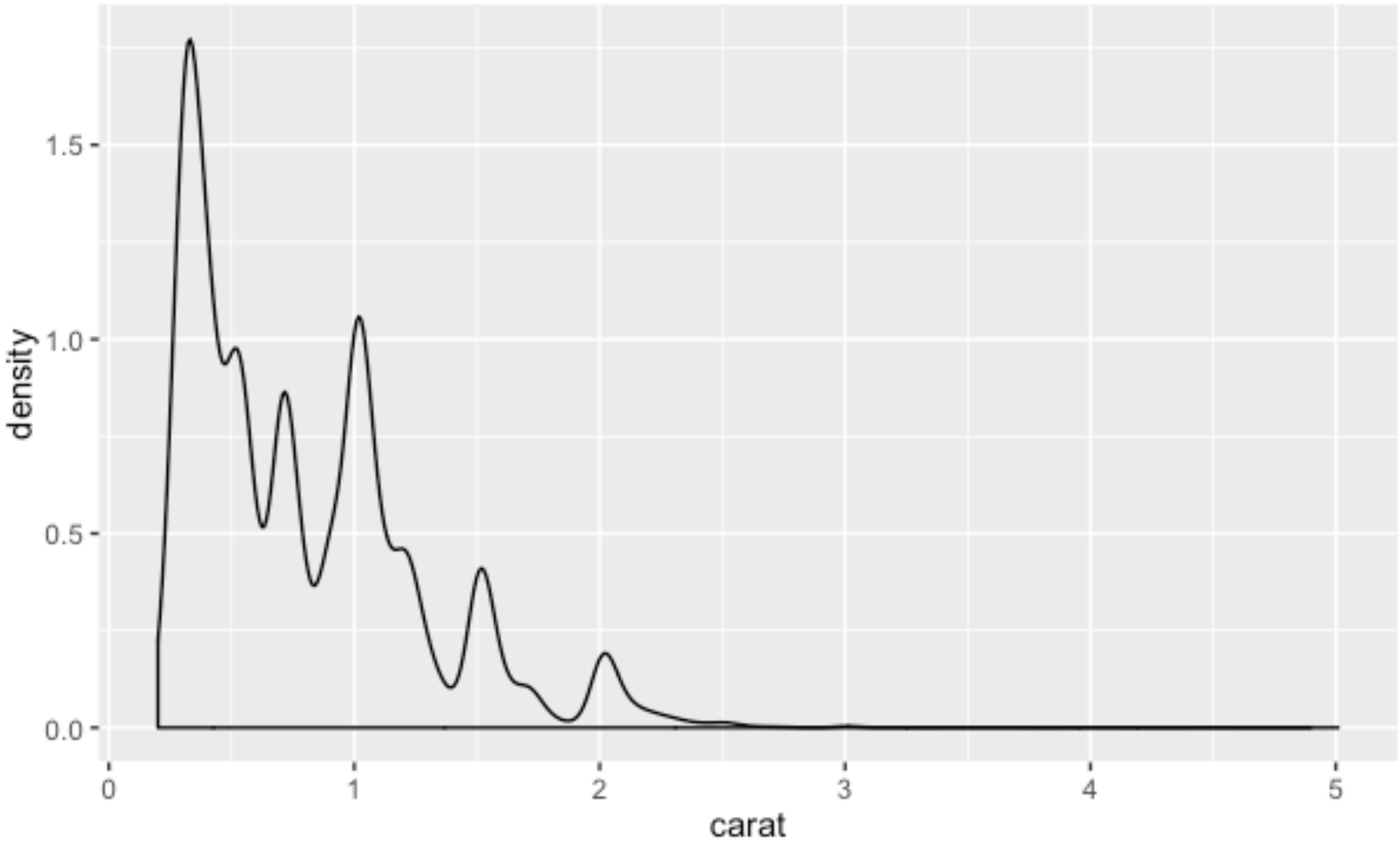
直方图的颜色

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



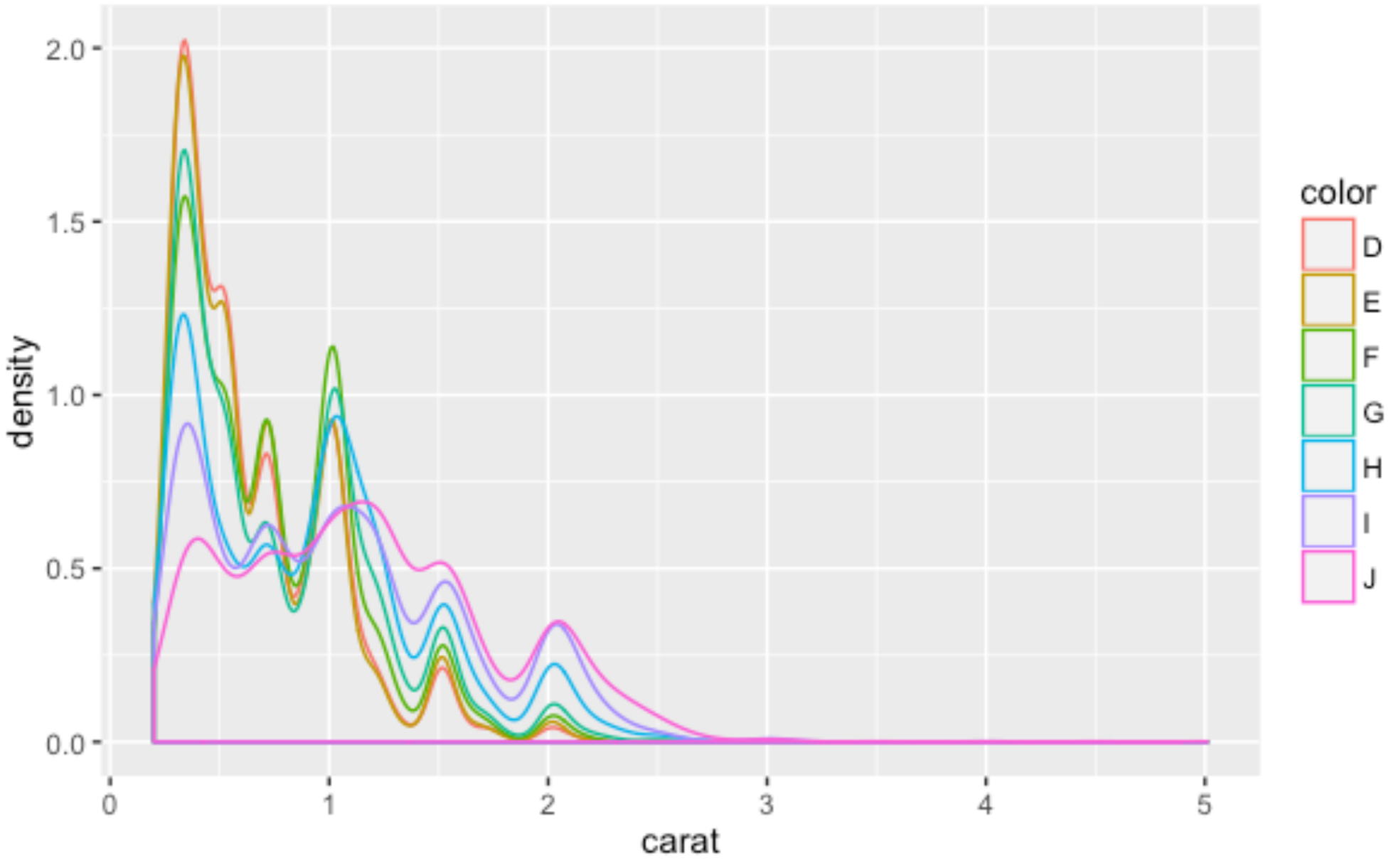
密度曲线图

```
qplot(carat, data = diamonds, geom = "density")
```



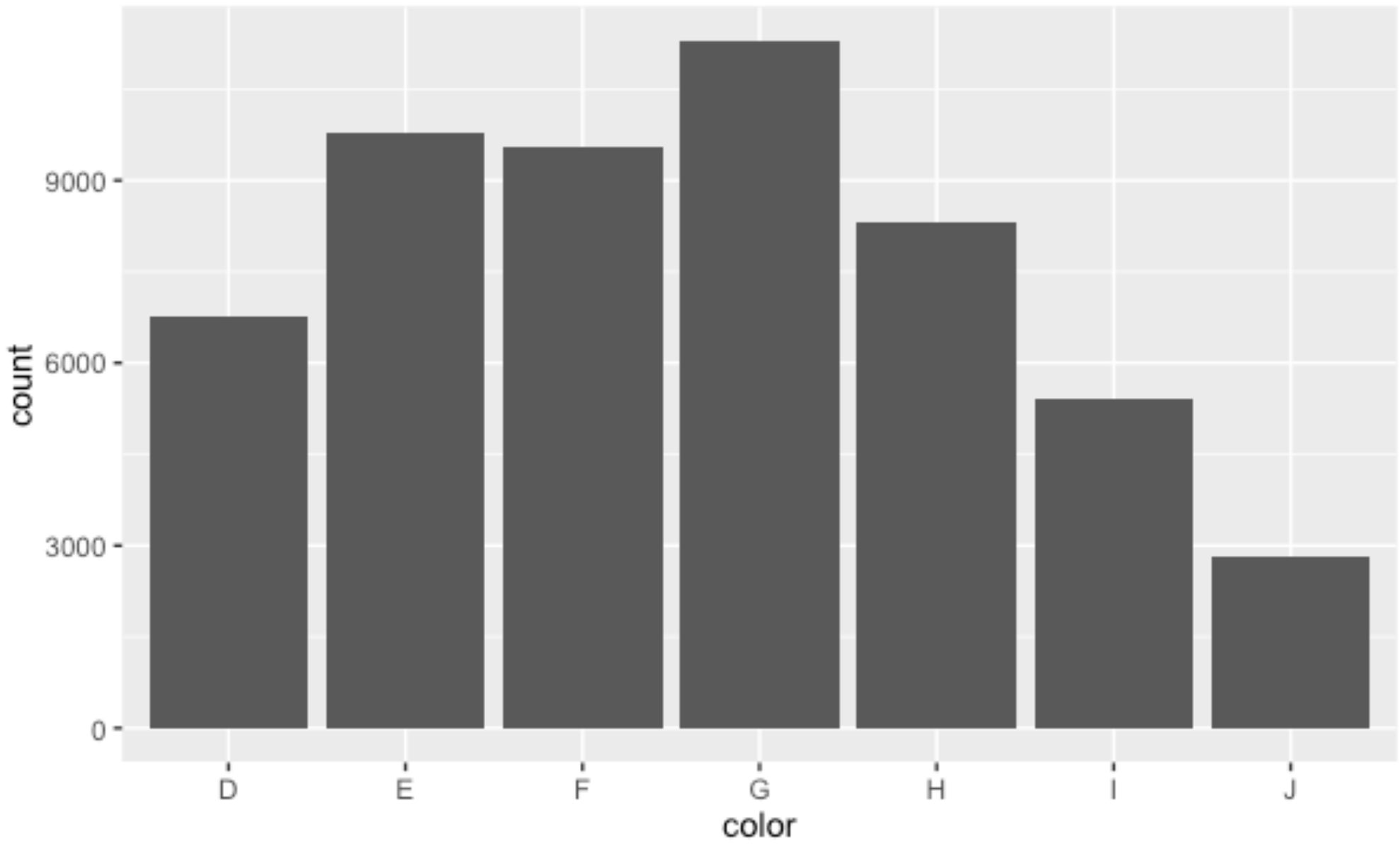
密度曲线图的颜色

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



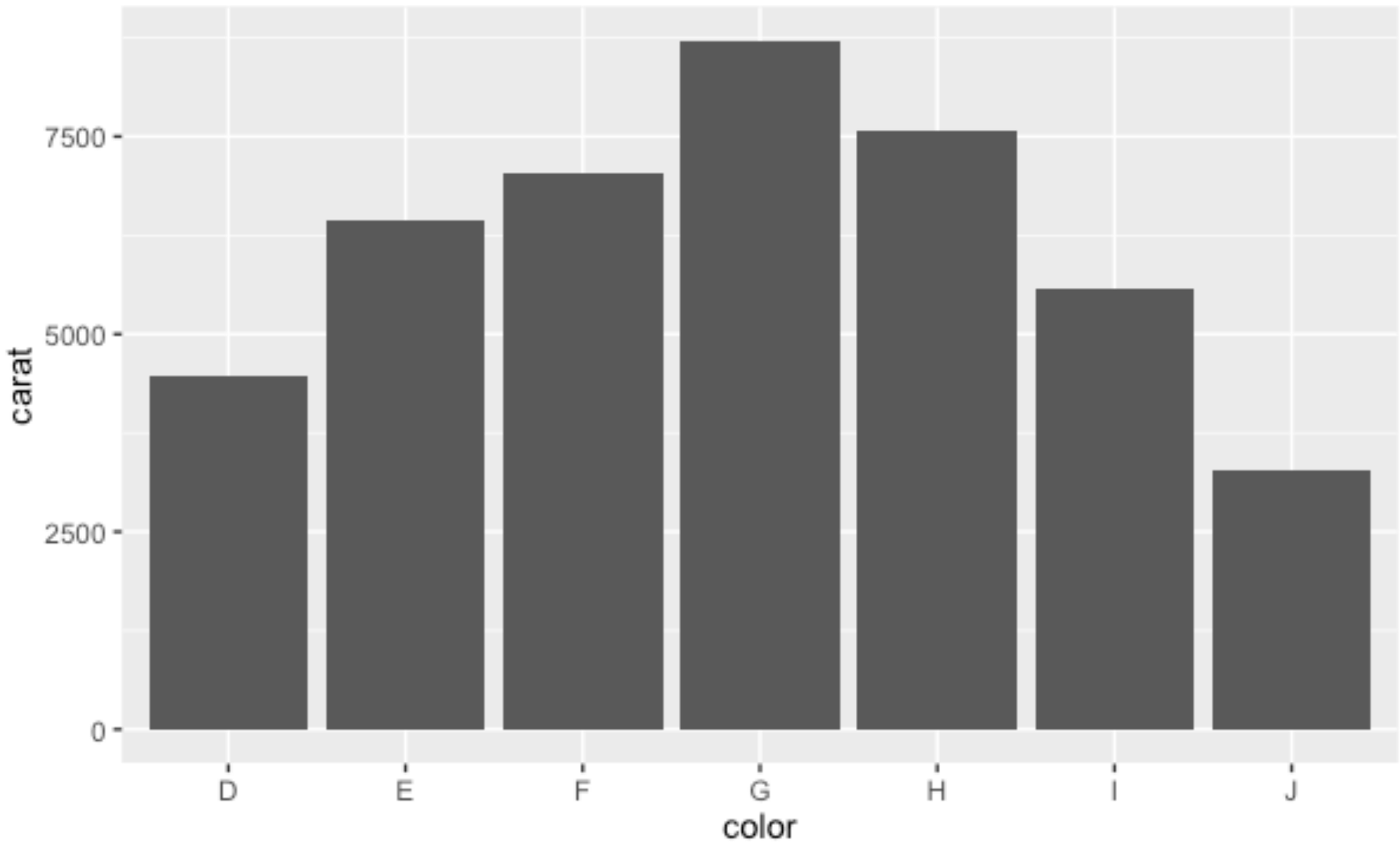
条形图

qplot(color, data = diamonds, geom = "bar")

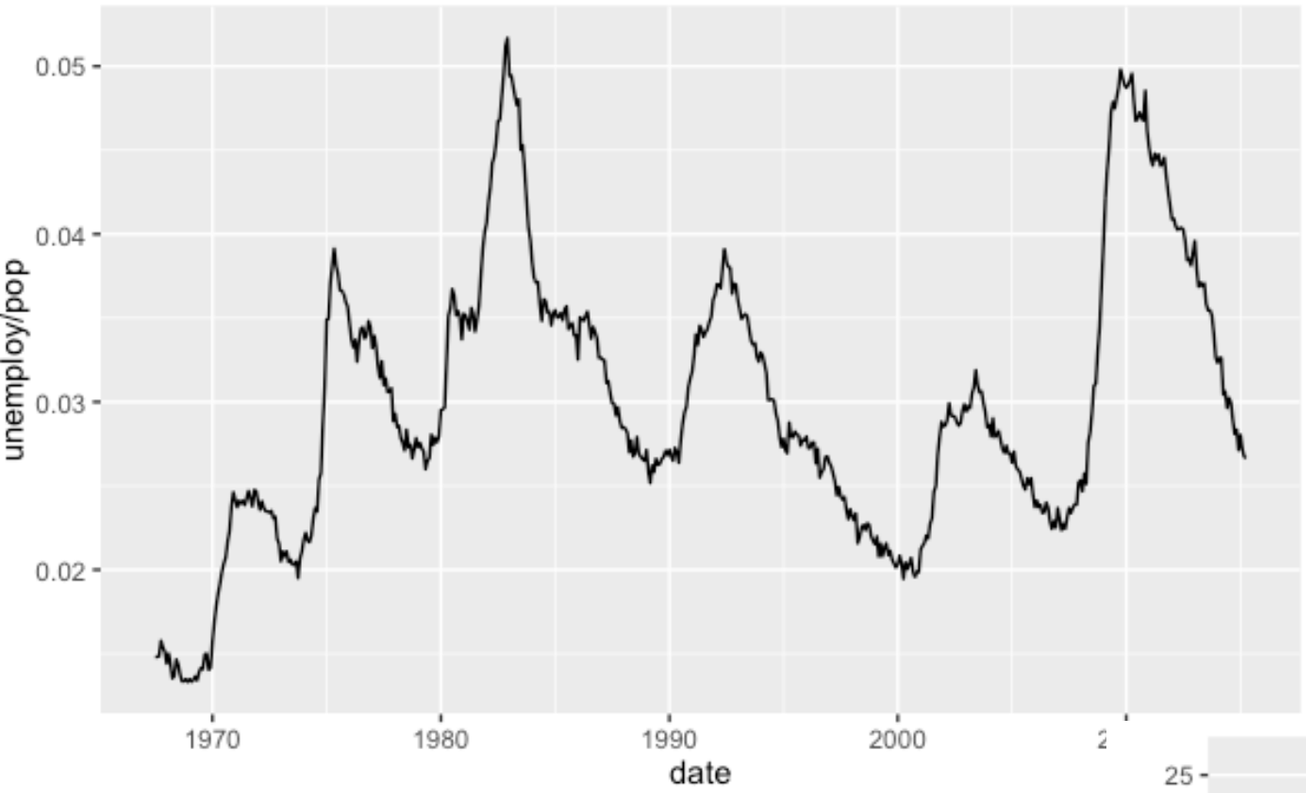


求和

```
qplot(color, data = diamonds, geom = "bar", weight = carat) +
scale_y_continuous("carat")
```

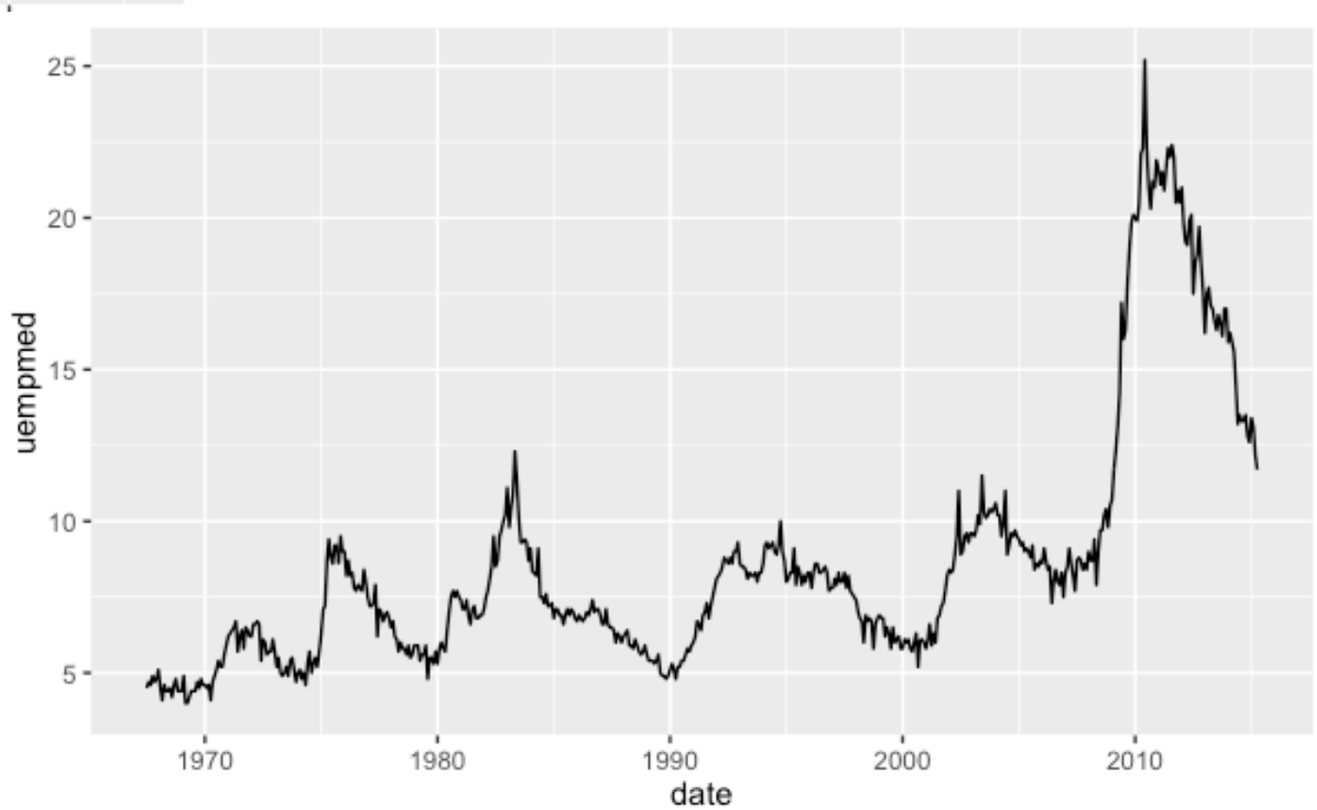


曲线图/折线图

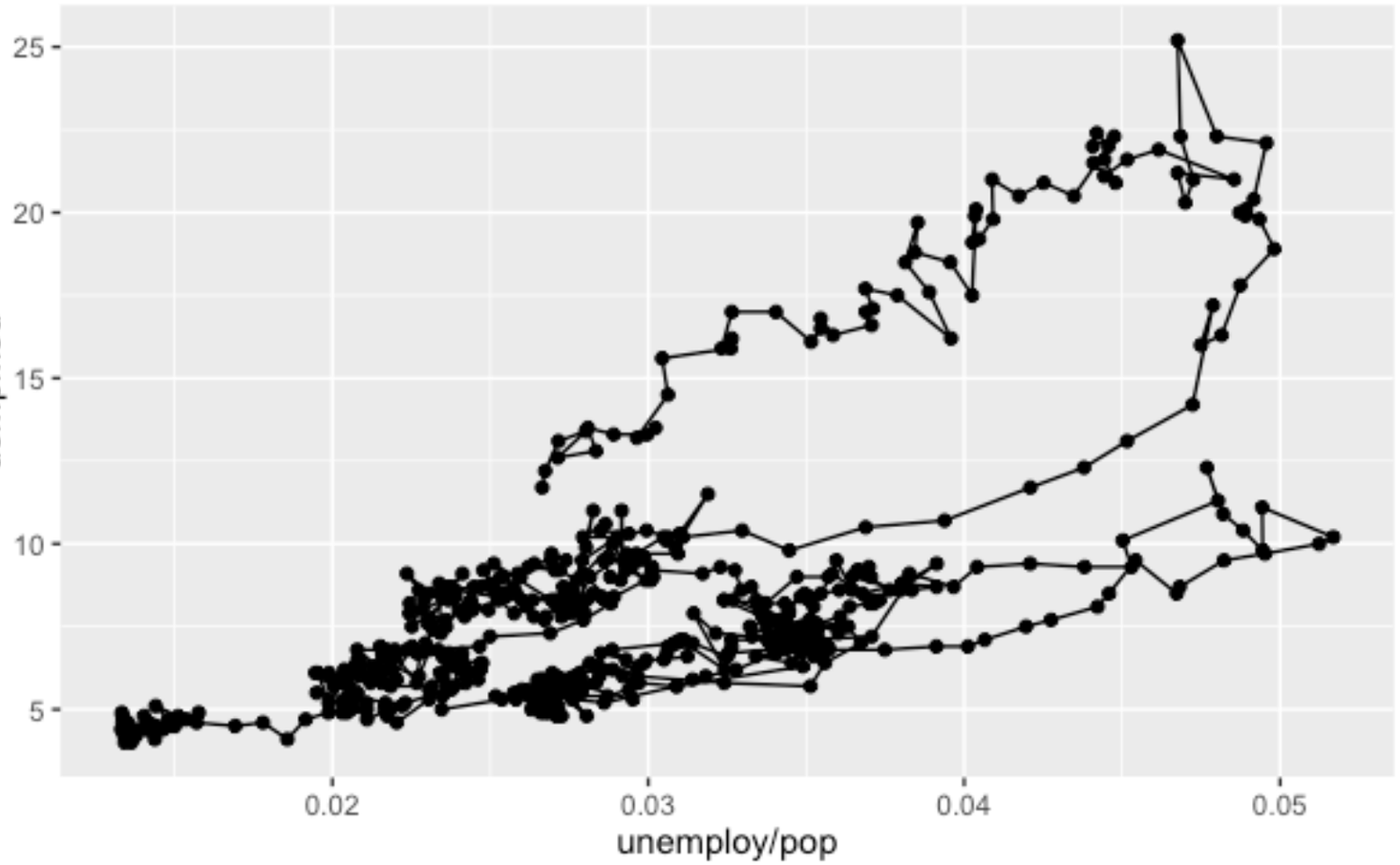


qplot(date, unemploy / pop, data = economics, geom = "line")

qplot(date, unempmed, data = economics, geom = "line")

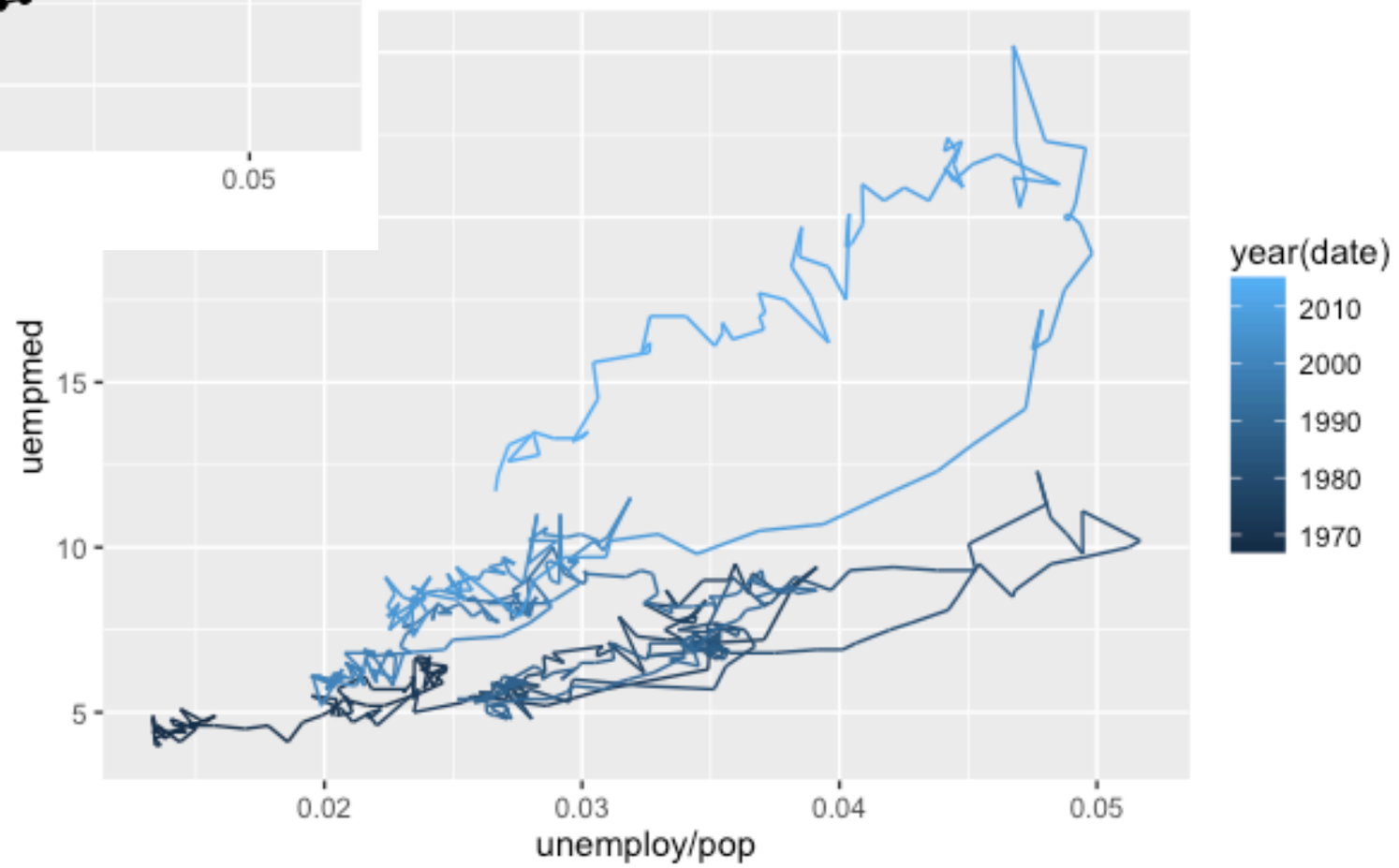


路径图



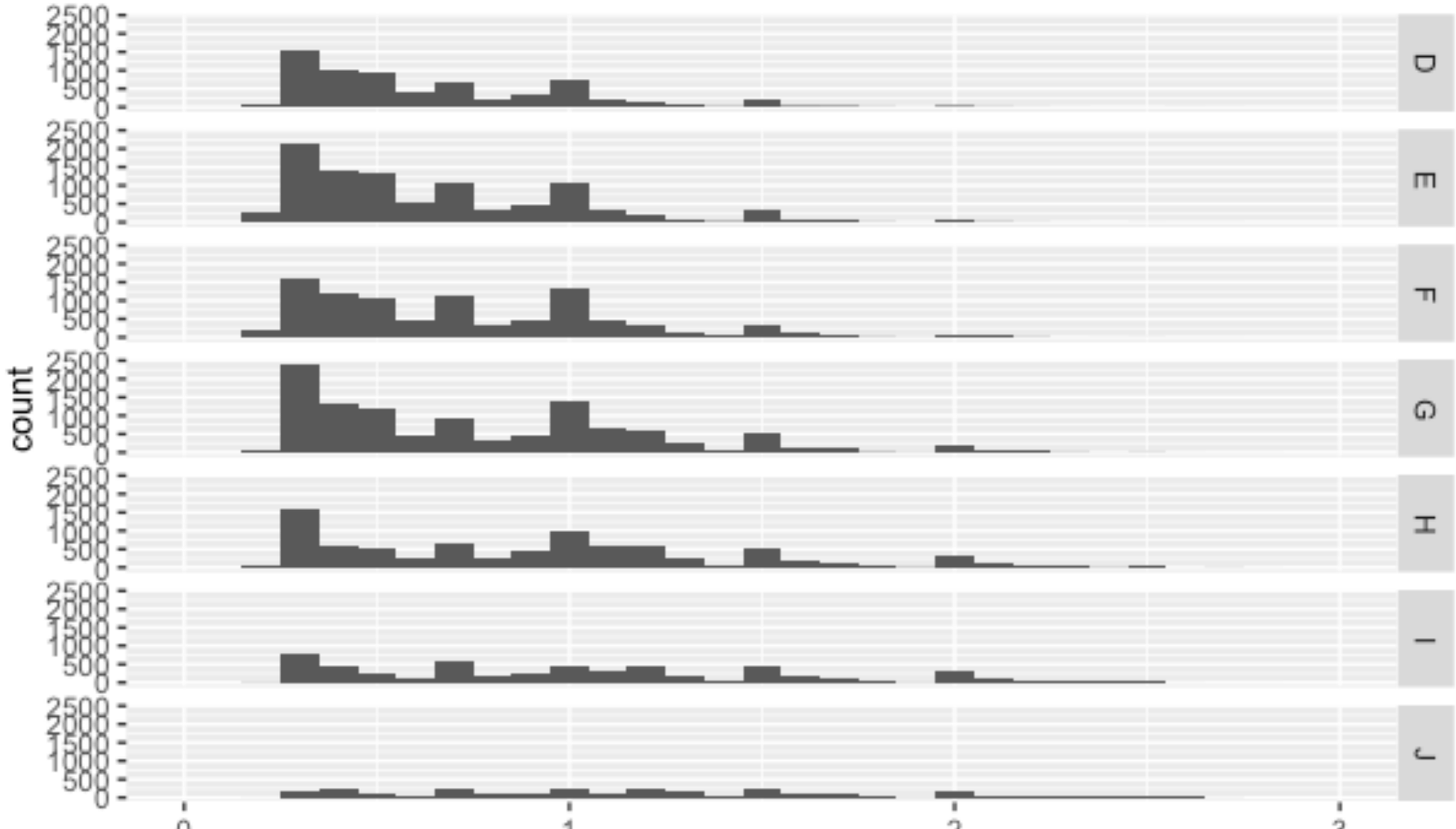

```
year <- function(x) as.POSIXlt(x)
$year + 1900

qplot(unemploy / pop, uempmed,
data = economics,
  geom = c("point", "path"))
```

```
qplot(unemploy / pop,
uempmed, data = economics,
geom = "path",
colour = year(date))
```



分面

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



其余参数

- xlim
- ylim
- log
- main
- xlab
- ylab

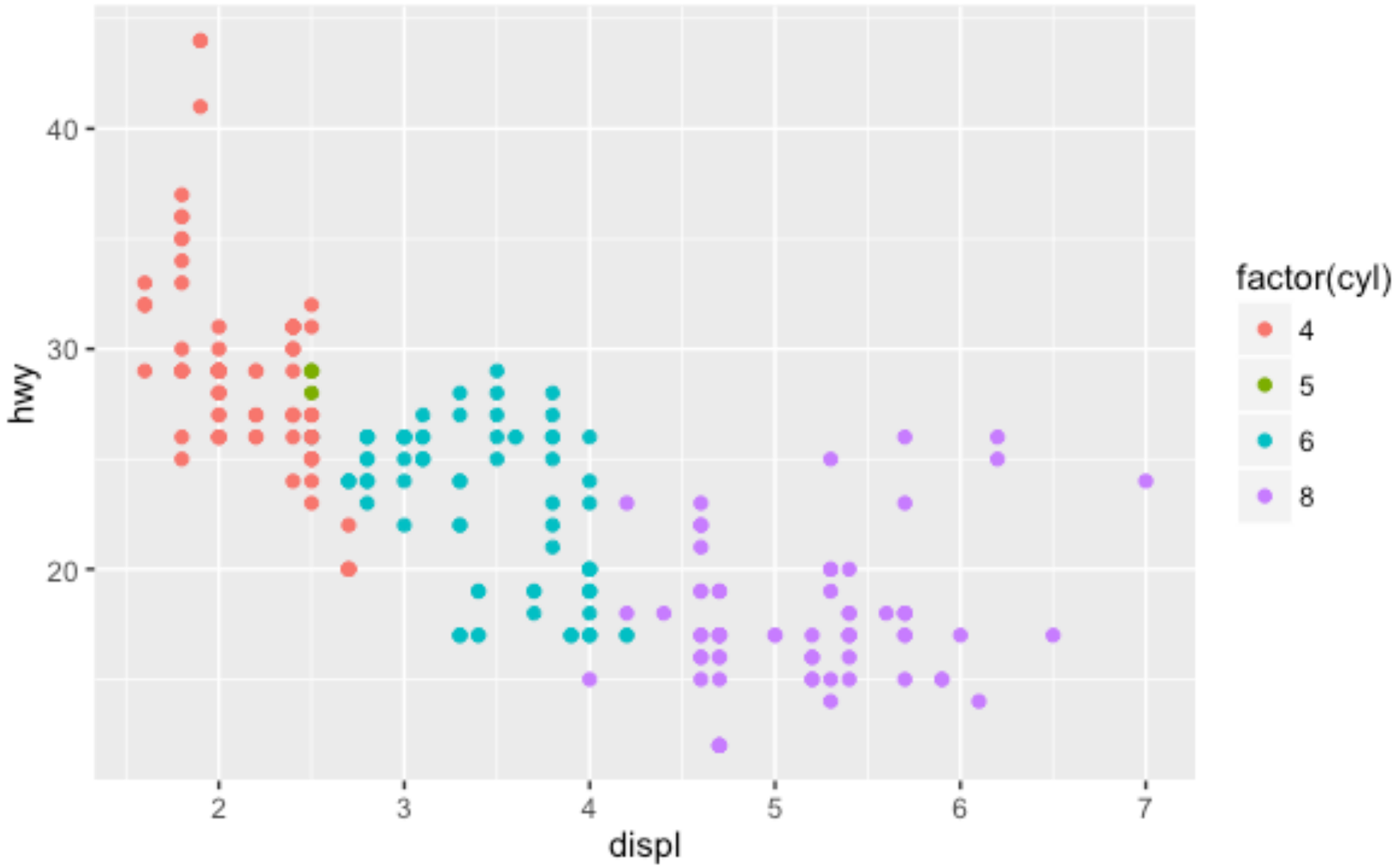
语法突破

MPG数据集

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
audi	a4	1.80	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.80	1999	4	manual(m5)	f	21	29	p	compact
audi	a4	2.00	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2.00	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.80	1999	6	auto(l5)	f	16	26	p	compact
audi	a4	2.80	1999	6	manual(m5)	f	18	26	p	compact
audi	a4	3.10	2008	6	auto(av)	f	18	27	p	compact
audi	a4 quattro	1.80	1999	4	manual(m5)	4	18	26	p	compact
audi	a4 quattro	1.80	1999	4	auto(l5)	4	16	25	p	compact
audi	a4 quattro	2.00	2008	4	manual(m6)	4	20	28	p	compact

散点图

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



数据到属性的映射

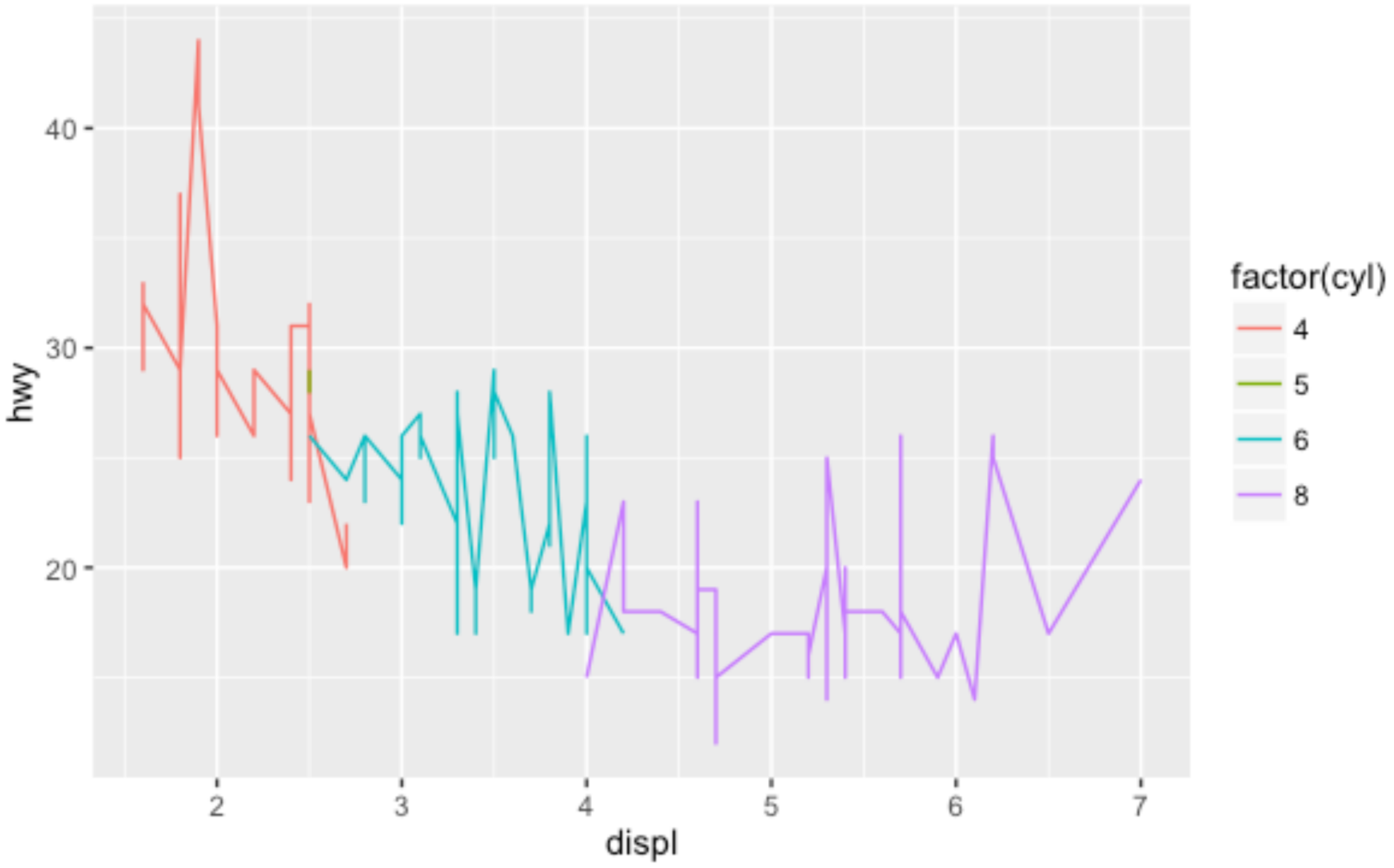
Disp映射到x坐标，hwy映射到y坐标，cyl映射到颜色

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

其余映射

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



标度变换

- 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

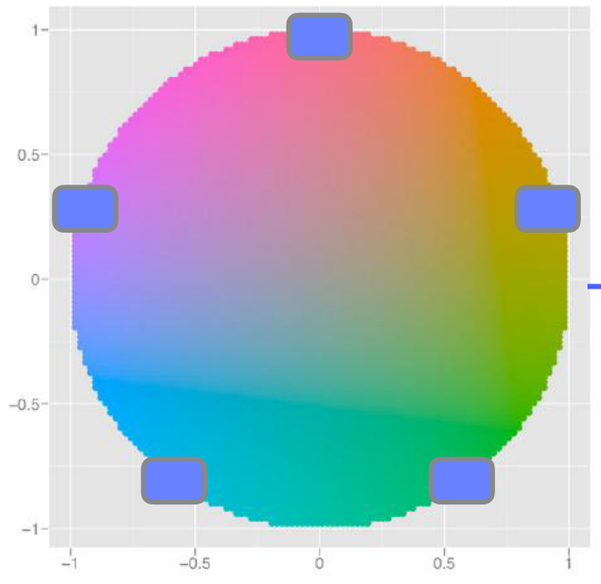
✳将水平坐标x映射到[0,1]区间。这里不使用具体像素值的原因是grid包替我们完成最终的转换

✳将垂直坐标y映射到[0,1]区间

✳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

✳颜色的scaling

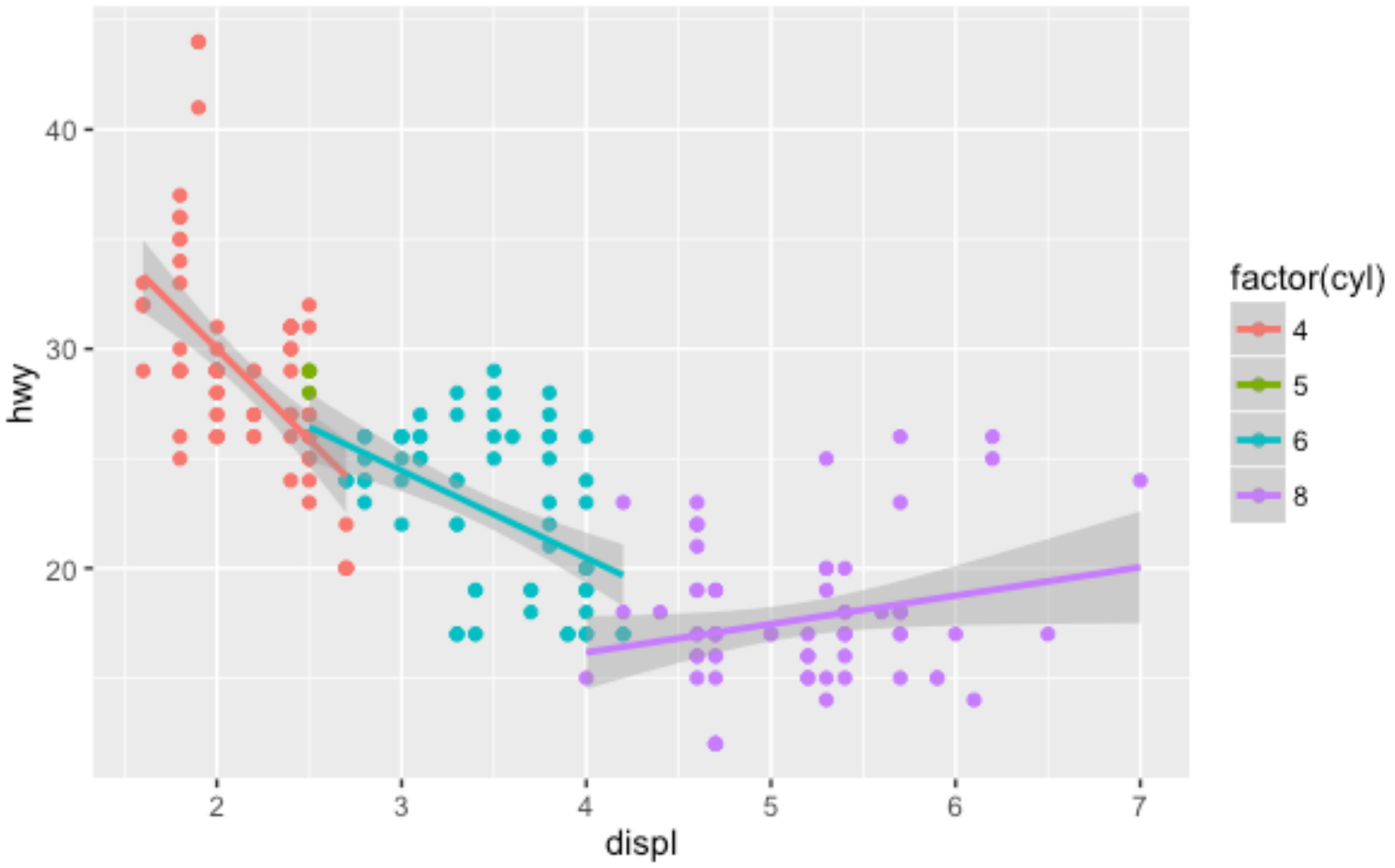
x	y	colour
1.8	29	4
1.8	29	4
2.0	31	4
2.0	30	4
2.8	26	6
2.8	26	6
3.1	27	6
1.8	26	4
1.8	25	4
2.0	28	4



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

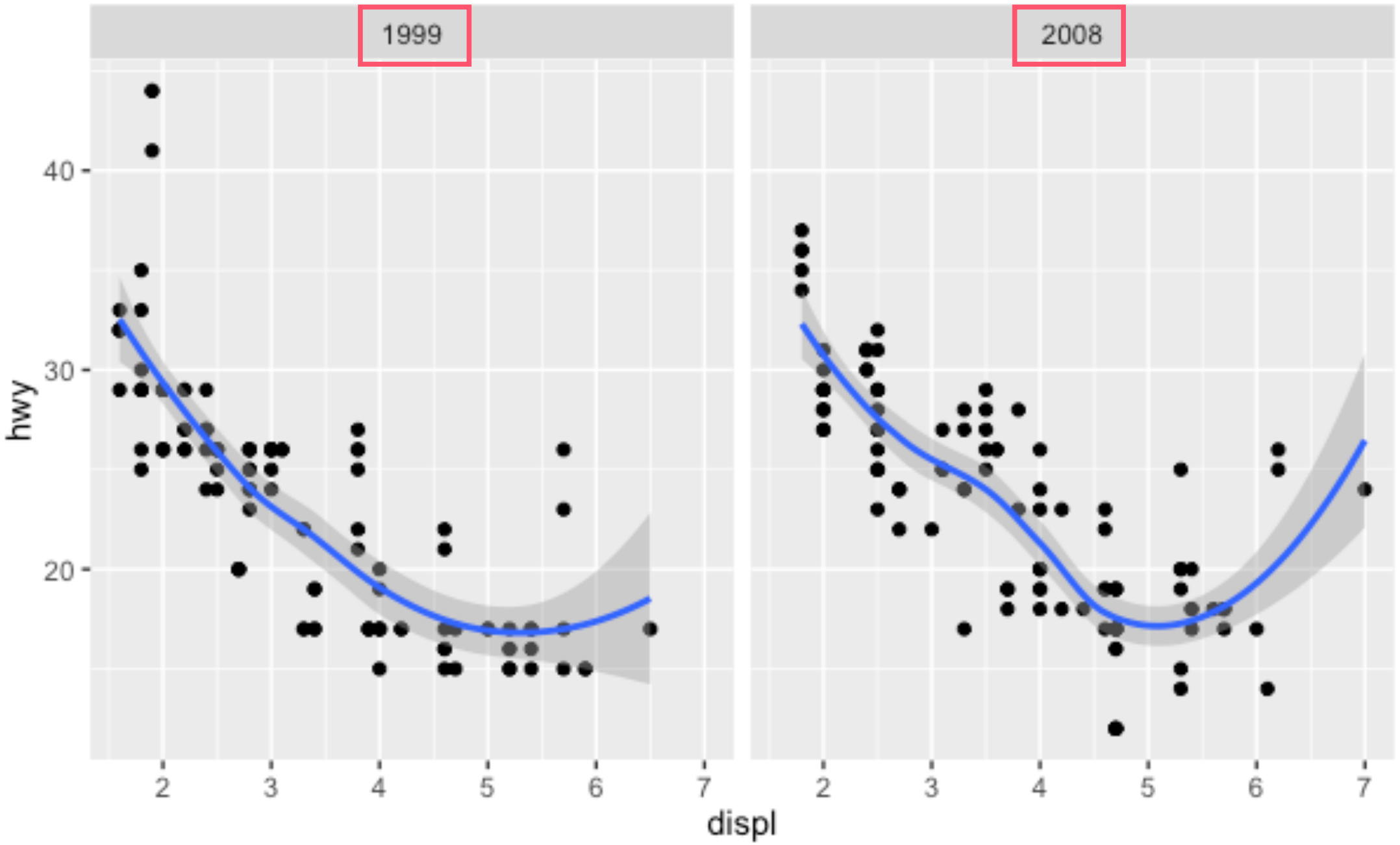
复杂图形

```
qplot(displ, hwy, data=mpg, colour=factor(cyl)) +
geom_smooth(data= subset(mpg, cyl != 5), method="lm")
```



含有分面的复杂图形

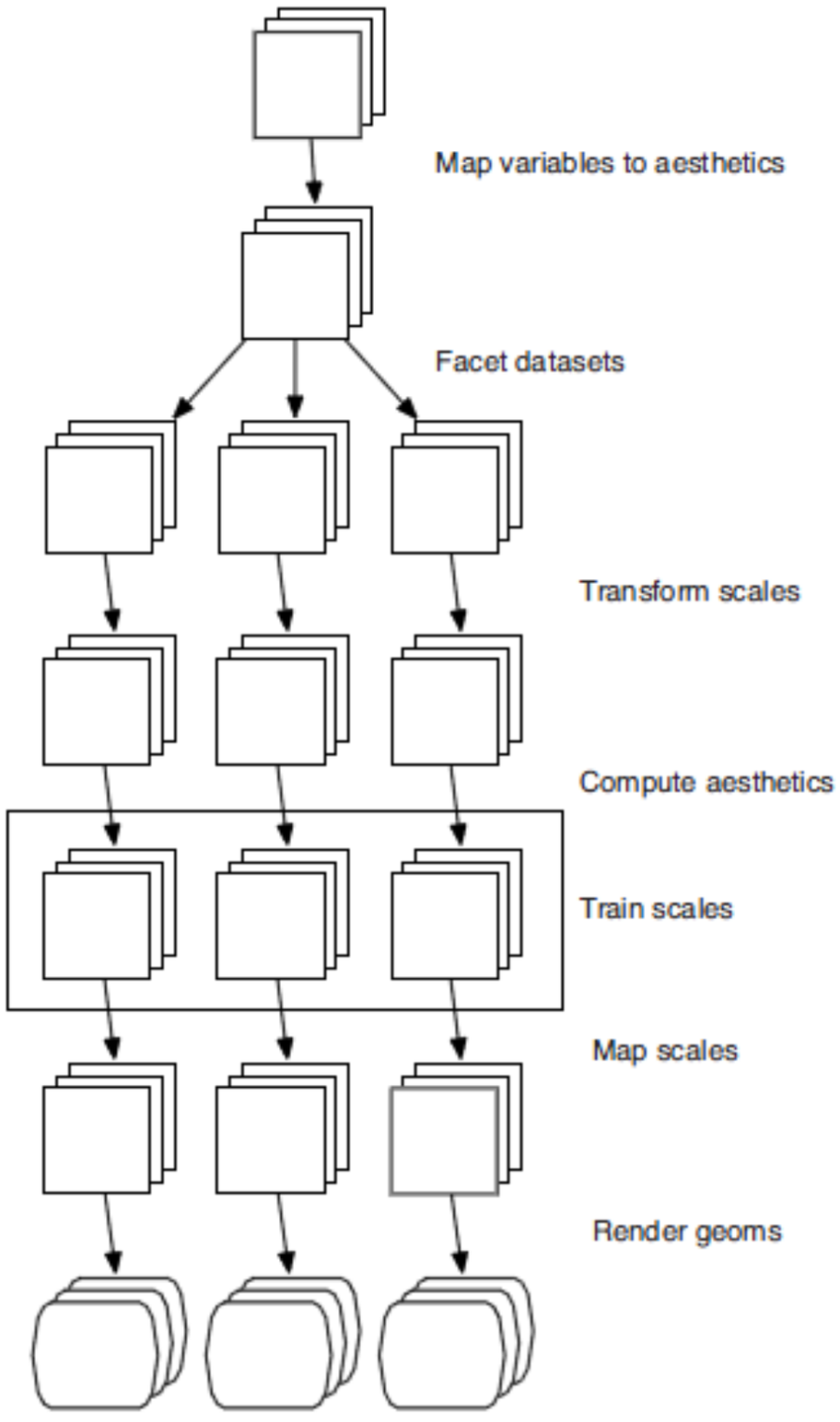
qplot(displ, hwy, data=mpg, facets = . ~ year) + geom_smooth()



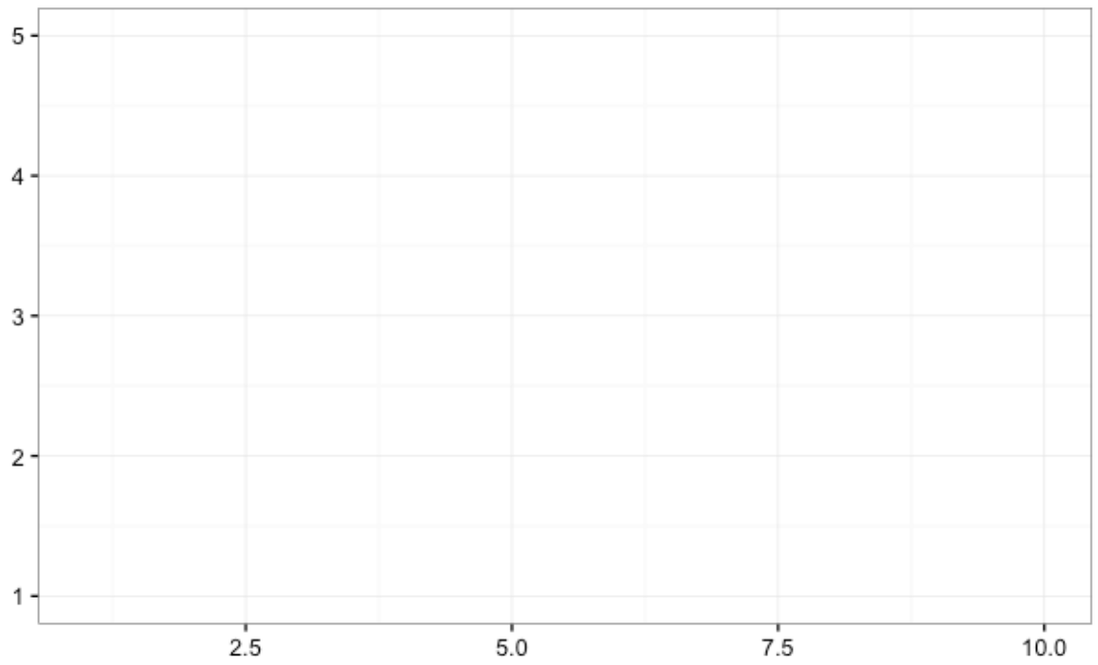
绘图过程

- 将变量映射到图形属性
- 对数据进行分面处理
- 标度转换
- 计算图形属性
- 标度训练
- 标度影射
- 渲染几何对象

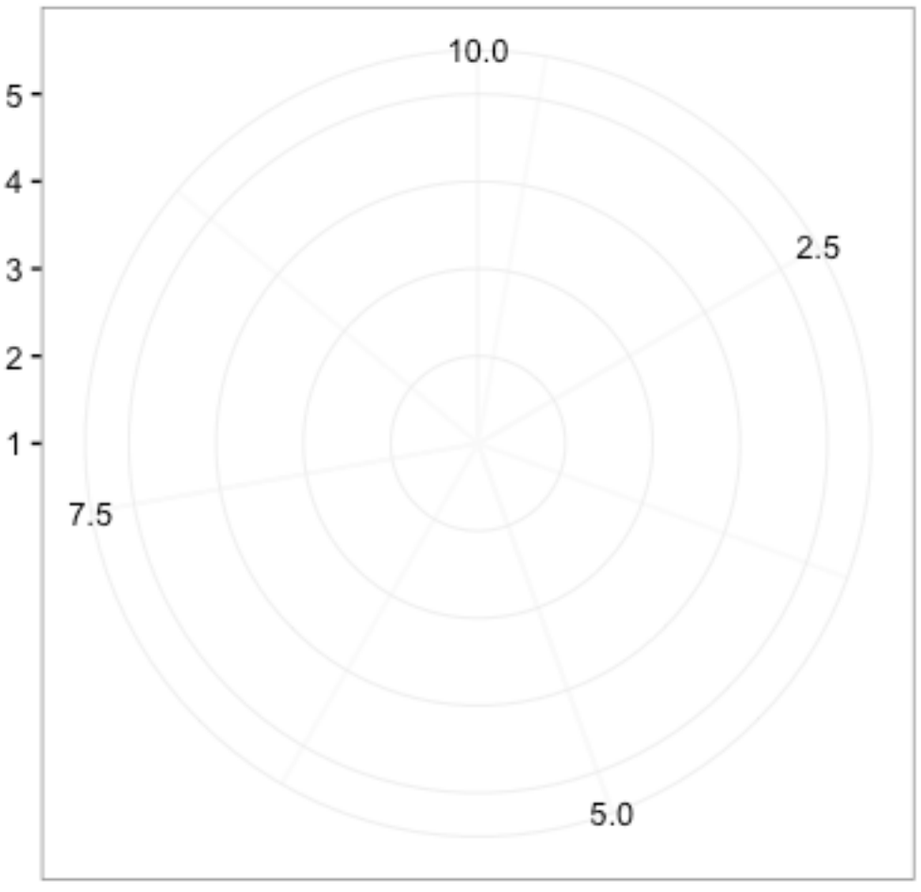
图层



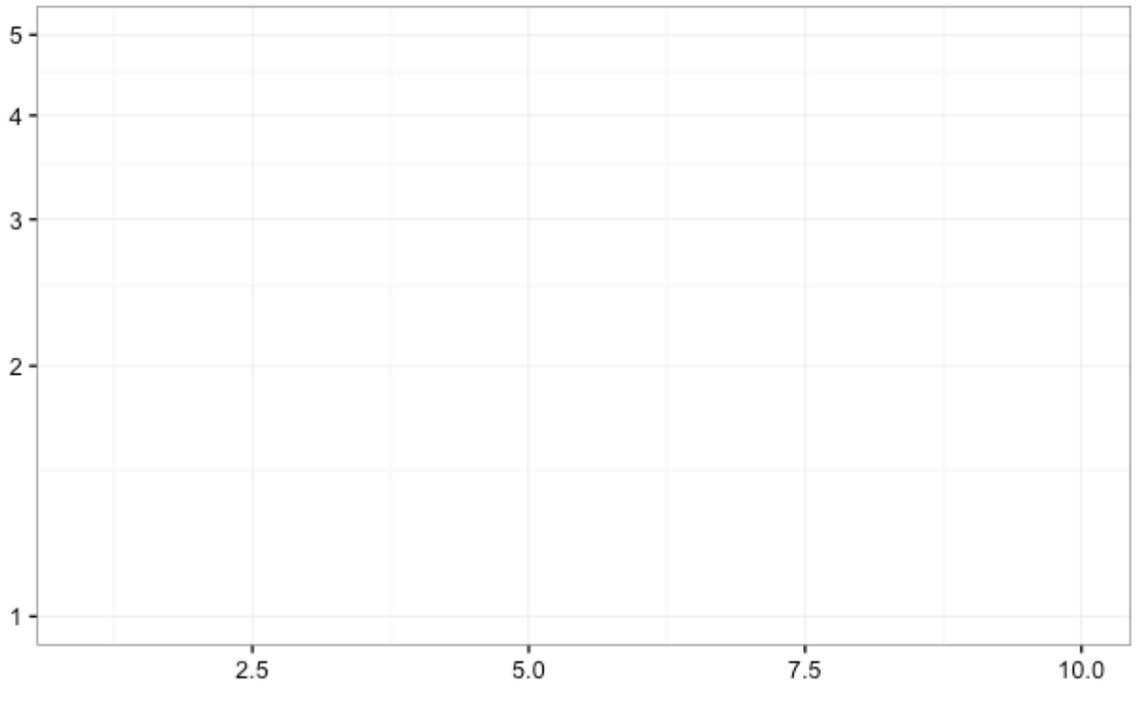
坐标系



笛卡尔



极坐标



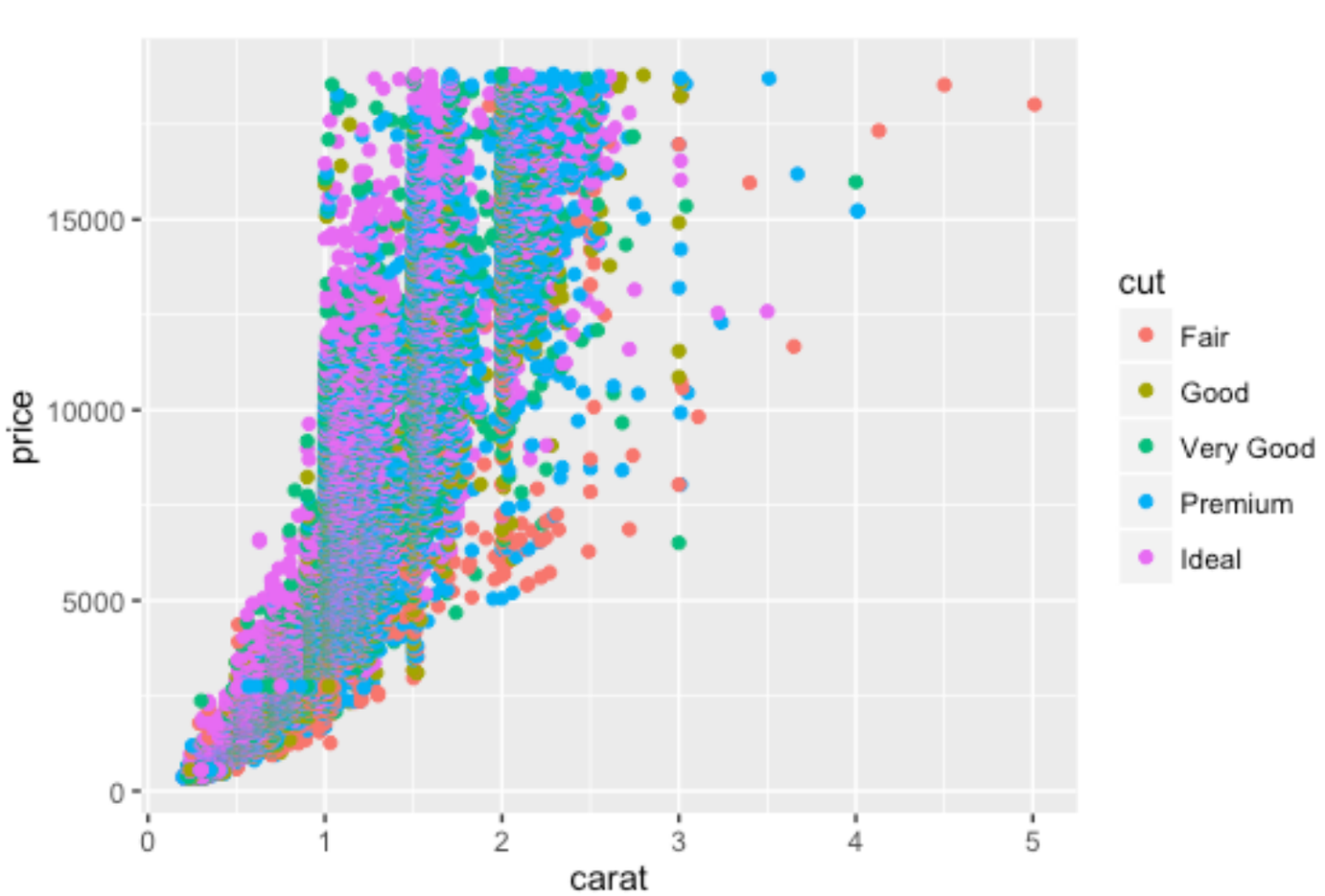
半对数

用图层构建图形

ggplot()

```
ggplot(data = NULL,
       mapping = aes(),
       …,
       environment = parent.frame())
```

layer()
自己查帮助




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

```
p <- p + layer(geom = “point",
               stat = "identity",
               position = "identity"
               )
p
```

ggplot_xxx()

```
geom(mapping = NULL,
        data = NULL,
        stat = “identity”
        position = “identity”
        …,
        na.rm = FALSE,
        show.legend = NA,
        inherit.aes = TRUE
        )
```

见教材ggplot2的58页

- geom_point()
- geom_line()
- geom_path()
- geom_bar()
- geom_histogram()
- geom_smooth()
- geom_density()
- geom_jitter()
- geom_text()
- geom_hline()
- geom_vline()
- geom_blank()
- geom_area()
- geom_abline()
- …

geom_xxx()和stat_xxx()

```
stat(mapping = NULL,
     data = NULL,
     geom/stat = “”
     position = “identity”
     …,
     na.rm = FALSE,
     show.legend = NA,
     inherit.aes = TRUE
)
```

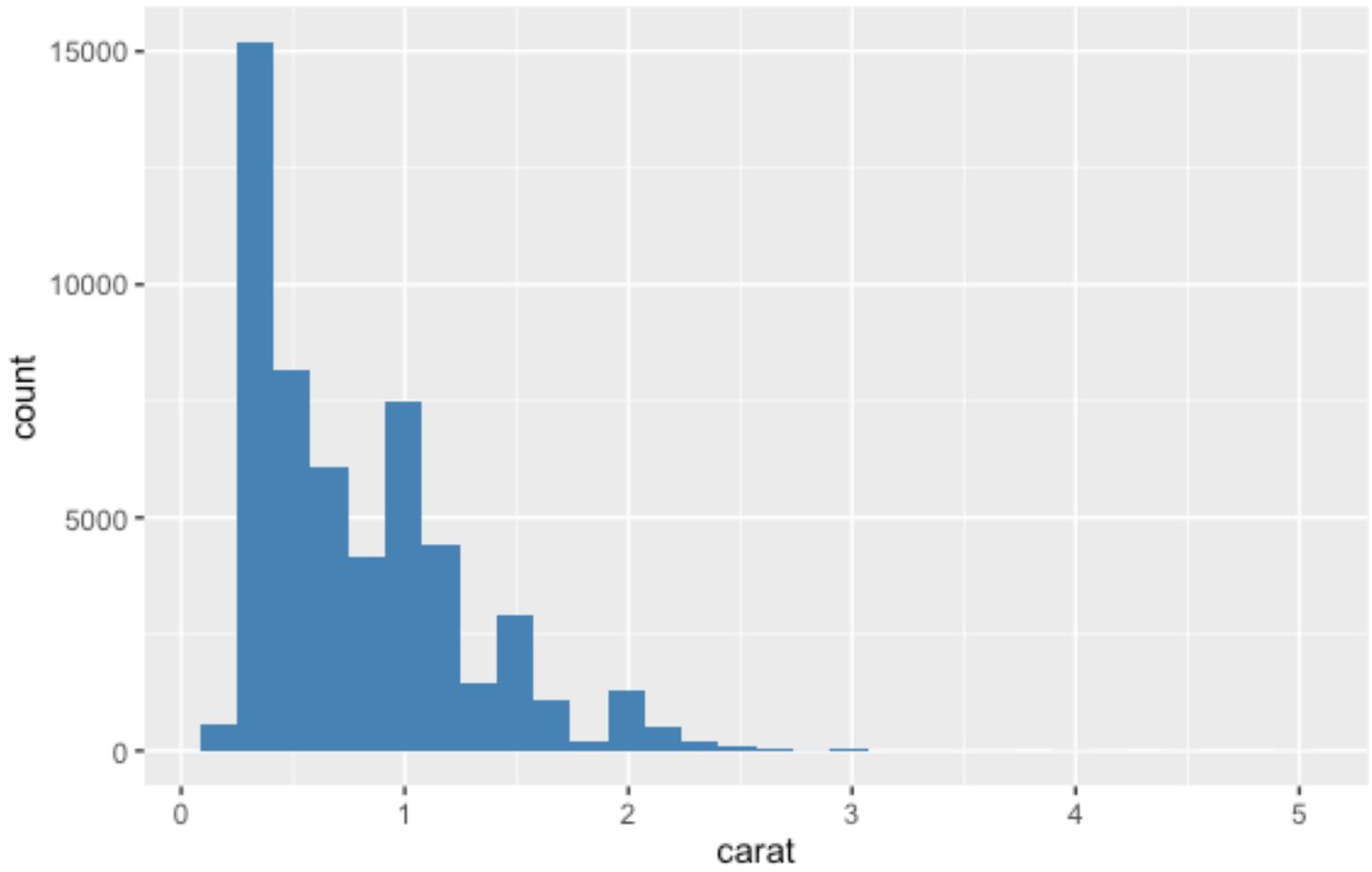
见教材ggplot2的60页

stat_identity()
stat_smooth()
stat_function()
stat_boxplot()
stat_density()
stat_quantile()
stat_sum()
stat_summary()
stat_unique()
stat_bin()
stat_bindot()
…

layer()和geom_xxx()

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

```
p <- ggplot(diamonds,
                    aes(x = carat))
p <- p + geom_histogram(bins = 30,
                                    fill = "steelblue")
p
```



summary()

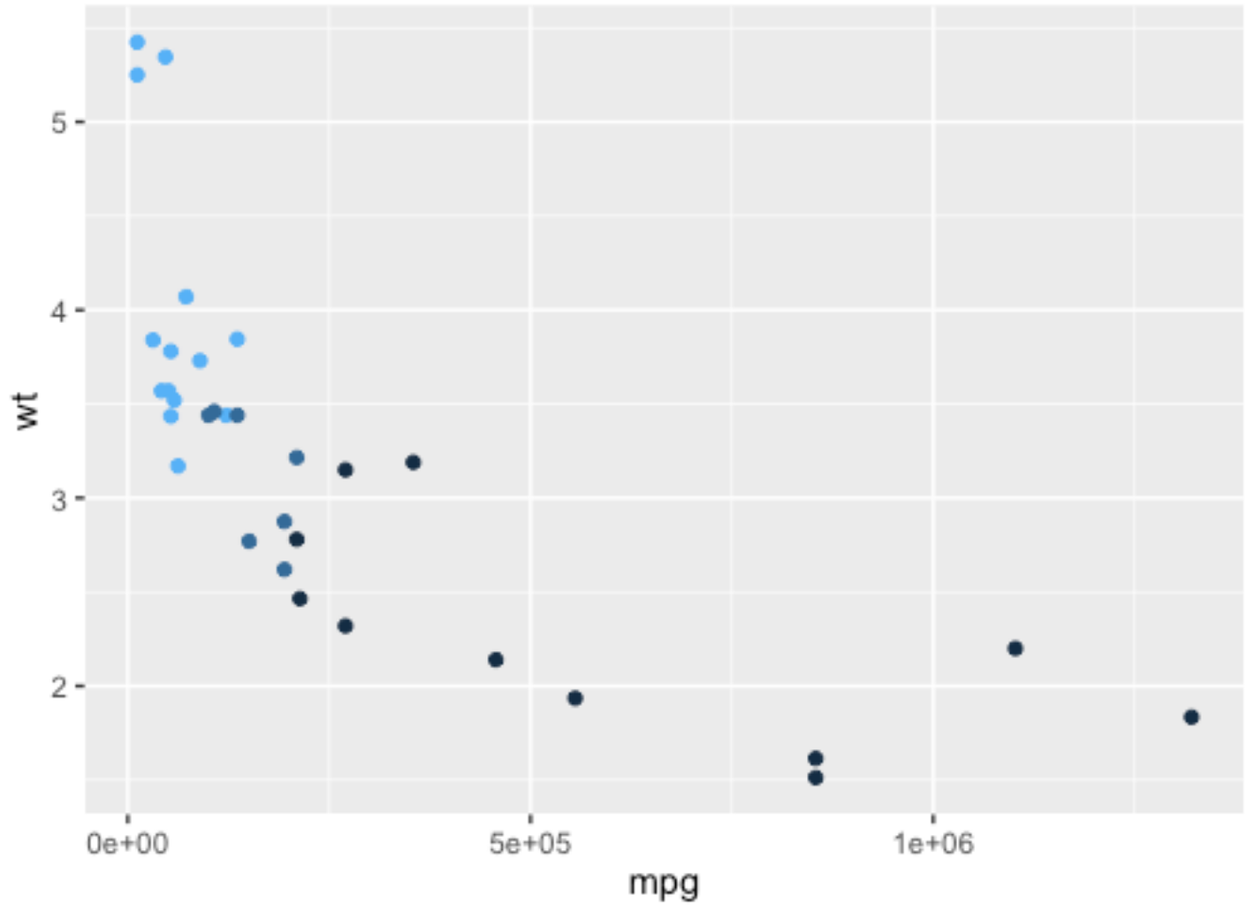
```
> p <- ggplot(msleep, aes(sleep_rem / sleep_total, awake))
> summary(p)
data: name, genus, vore, order, conservation, sleep_total, sleep_rem,
  sleep_cycle, awake, brainwt, bodywt [83x11]
mapping:  x = sleep_rem/sleep_total, y = awake
faceting: facet_null()
```

```
> 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: na.rm = FALSE
position_identity
```

ggplot2 I

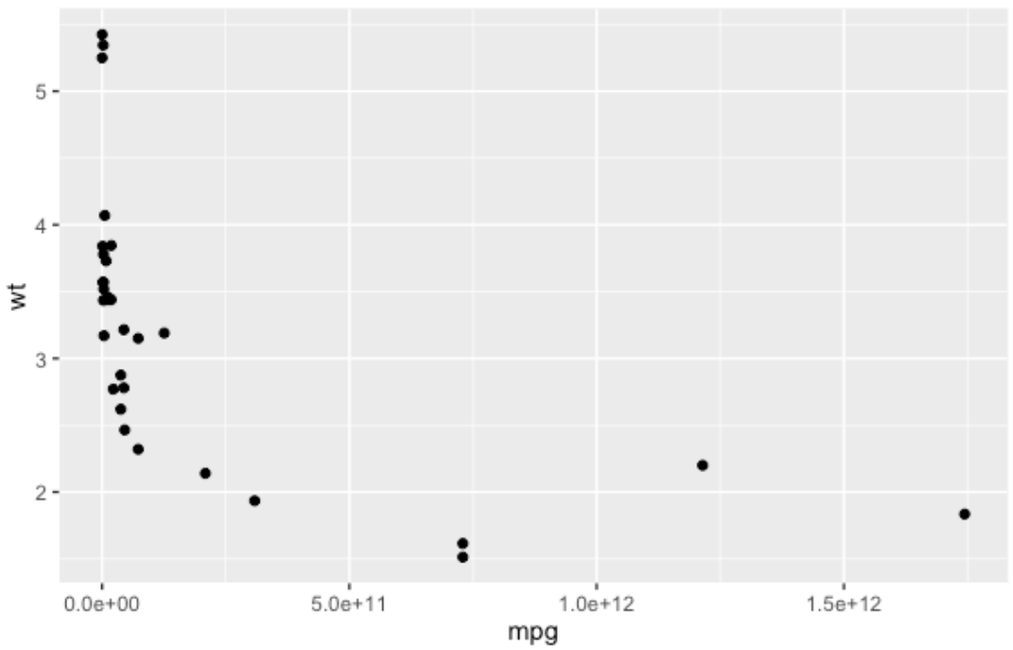
%+%

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




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

aes()

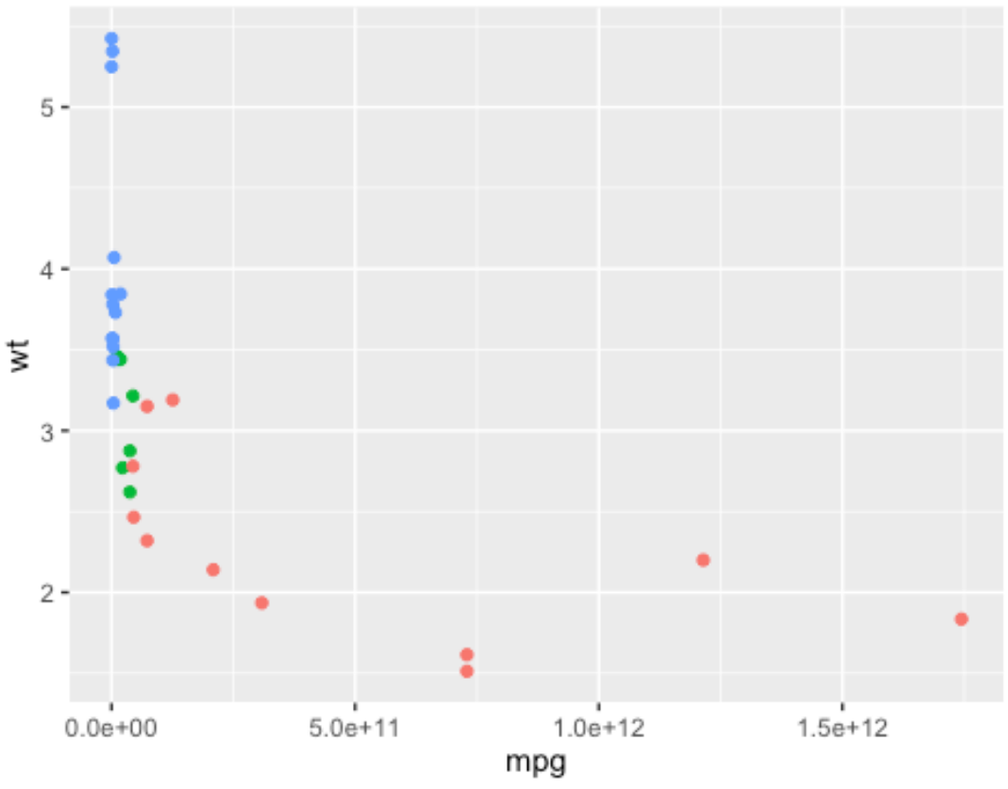


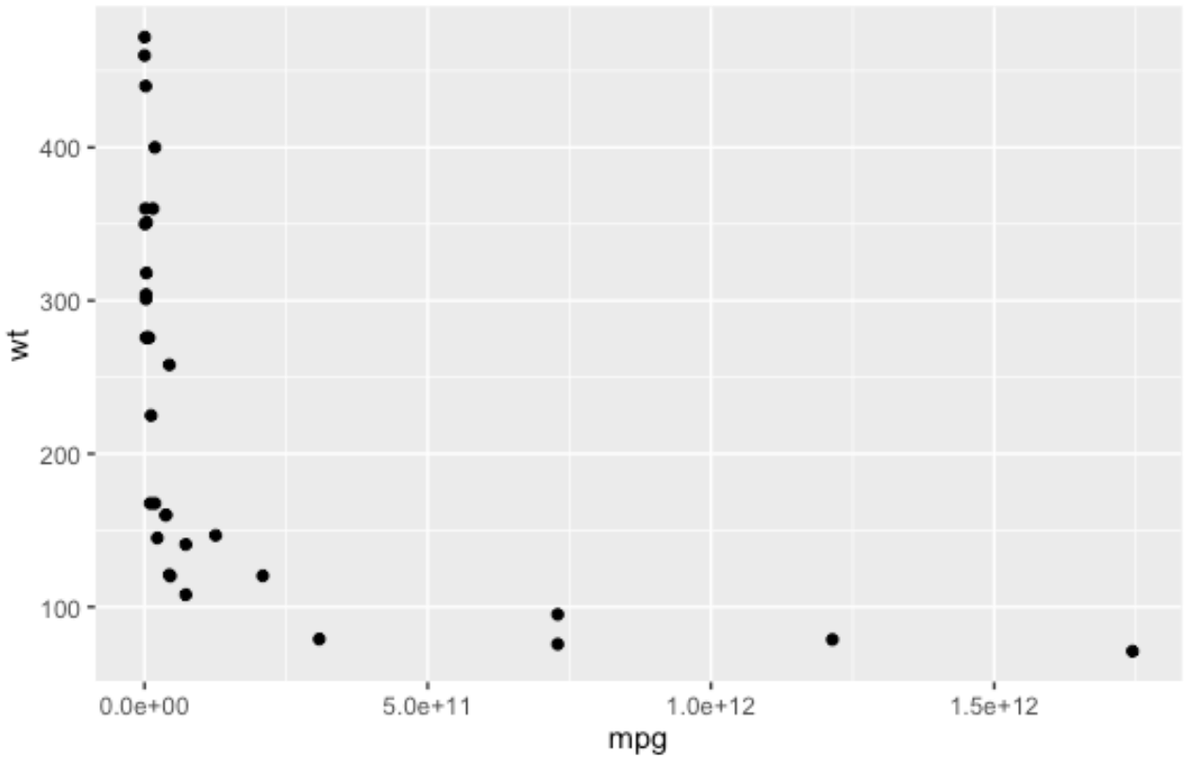

```
p <- ggplot(mtcars, aes(x = mpg, y = wt))
```

```
p + geom_point()
```

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

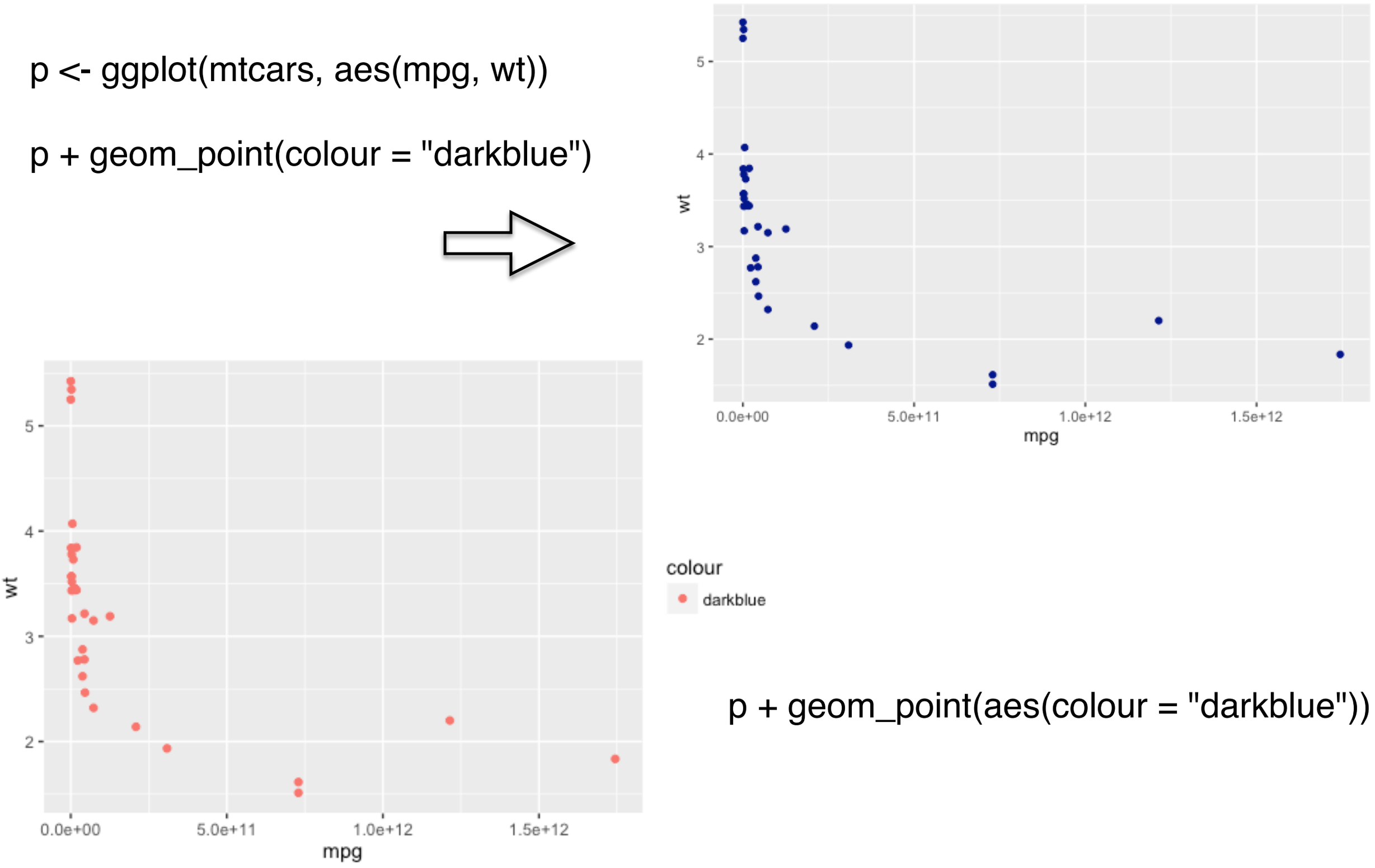
```
p + geom_point(aes(y = disp))
```





图形属性vs. 图层属性

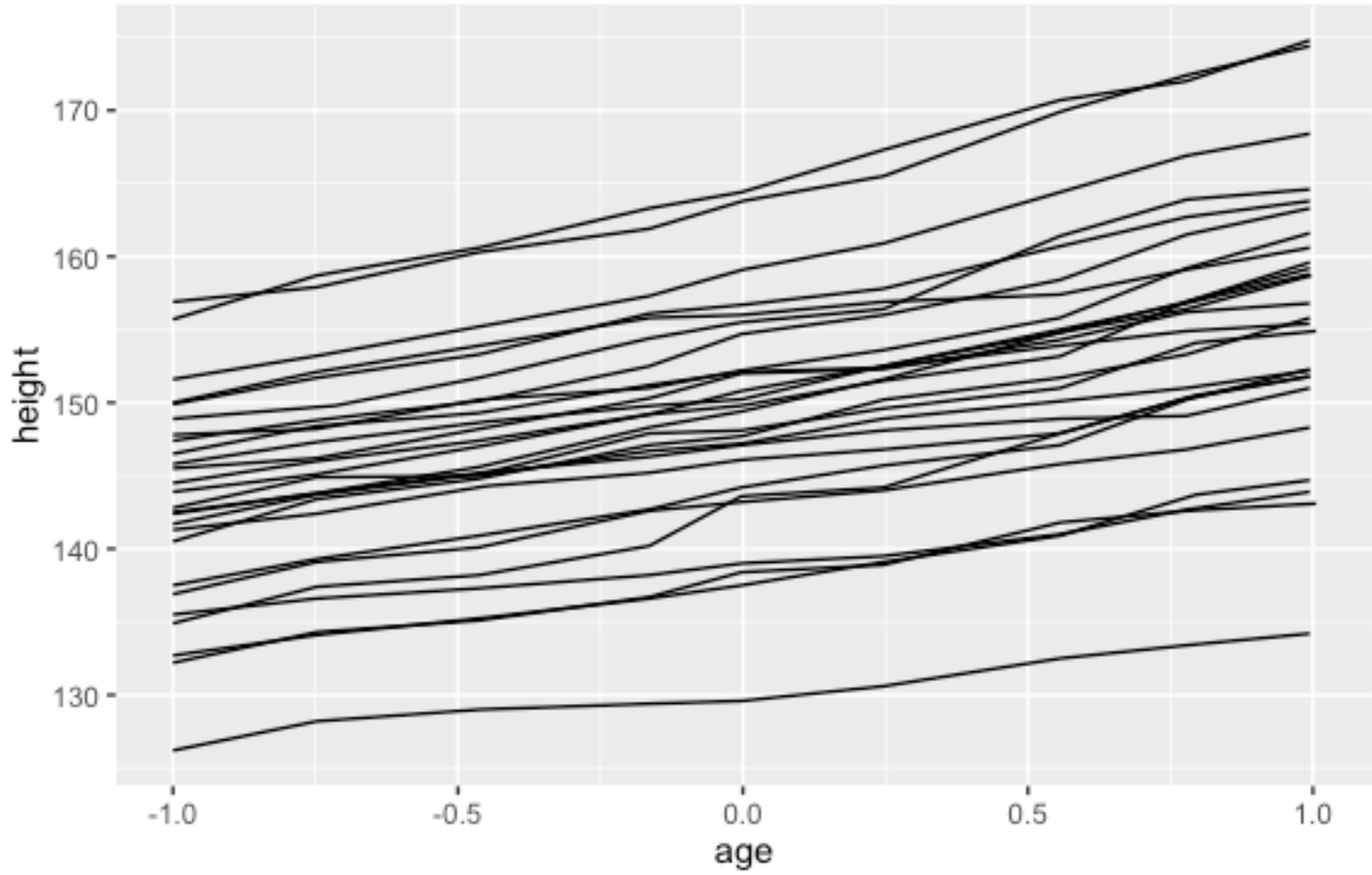
```
p <- ggplot(mtcars, aes(mpg, wt))

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

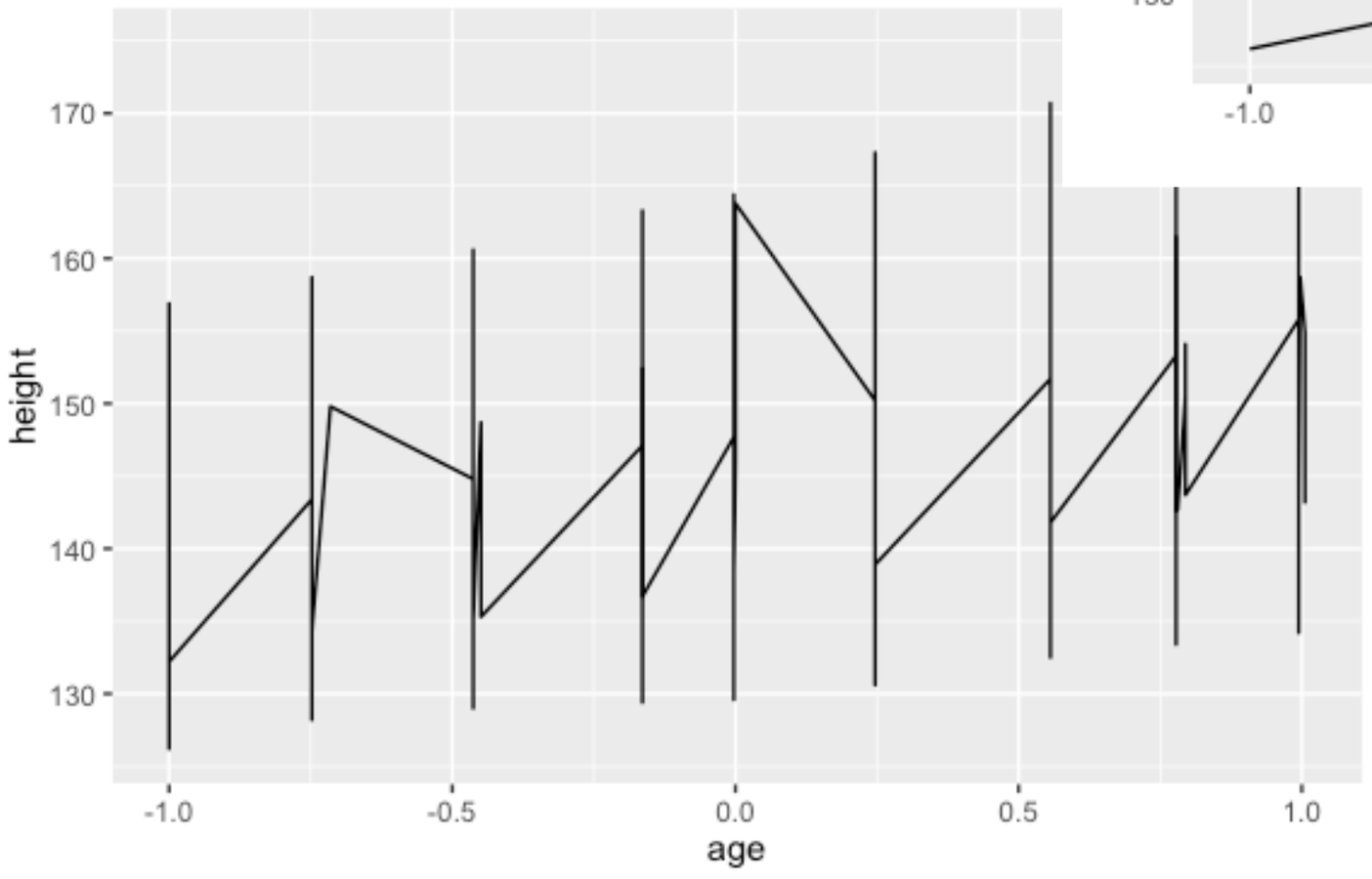



```
p + geom_point(aes(colour = "darkblue"))
```

分组

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

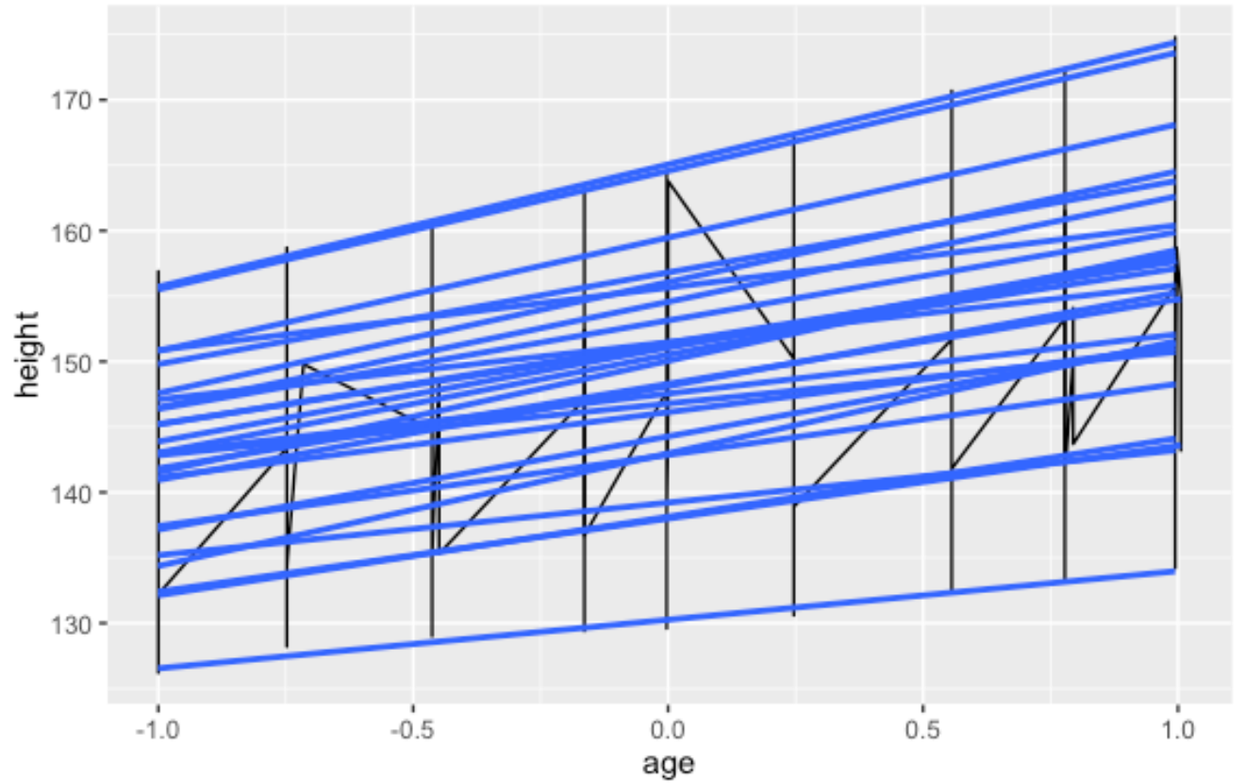


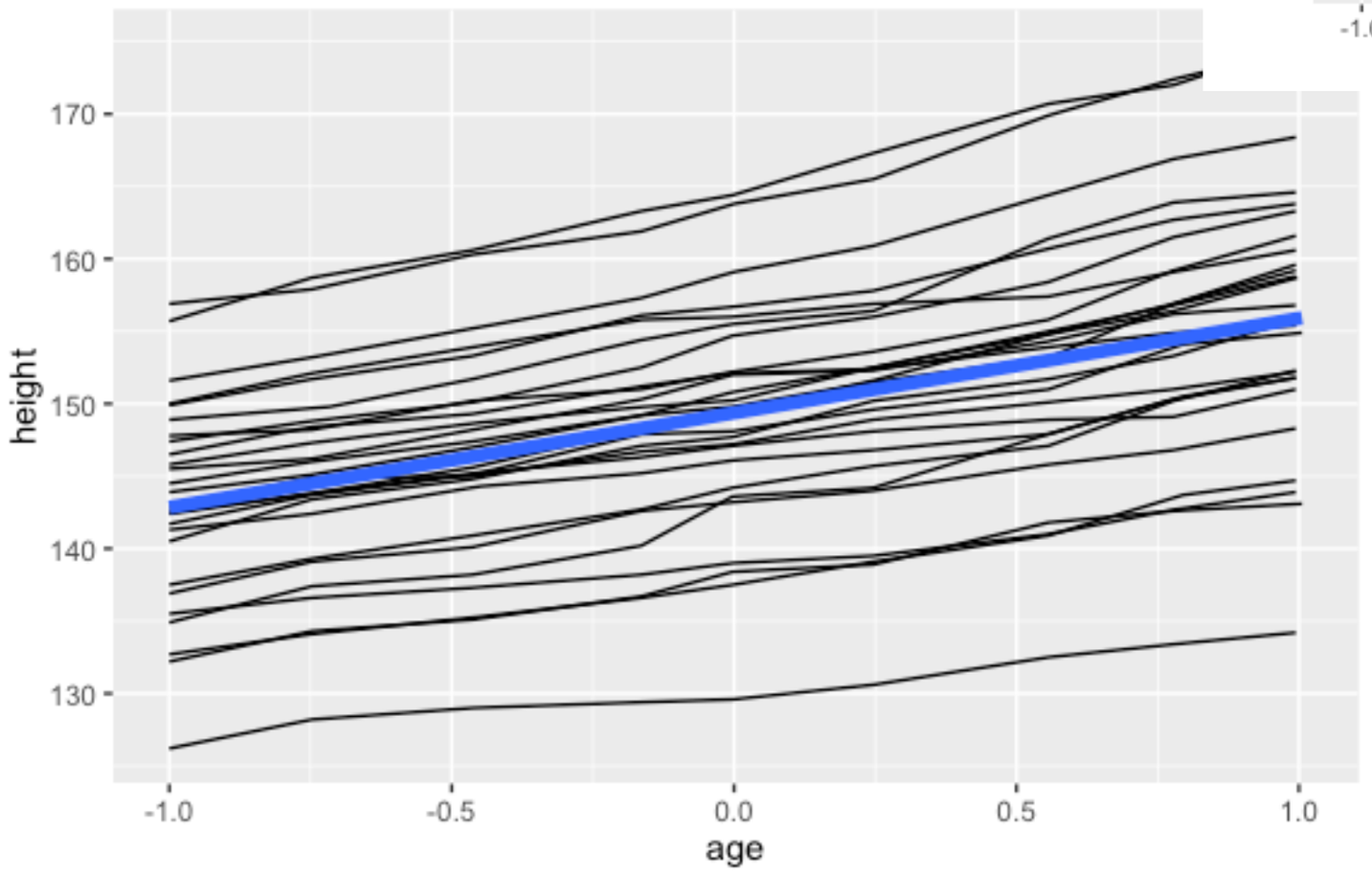



```
p <- ggplot(Oxboys,
            aes(age,
                height,
                group = 1)
            )
+ geom_line()
```

分组

```
p <- ggplot(Oxboys,
            aes(age,
                height,
                group = Subject)
            )

p + geom_smooth(aes(group = Subject),
                method="lm",
                se = F)
```






```
p <- ggplot(Oxboys,
            aes(age,
                height,
                group = Subject)
            )

p + geom_smooth(aes(group = 1),
                method="lm",
                se = F)
```

提问时间！

孙惠平

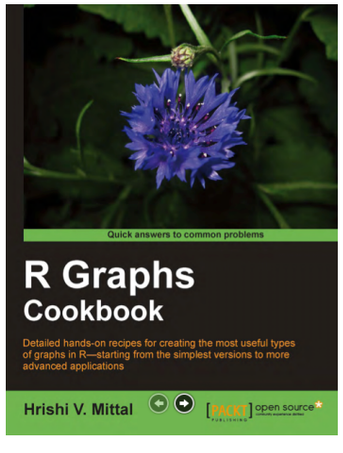
sunhp@ss.pku.edu.cn

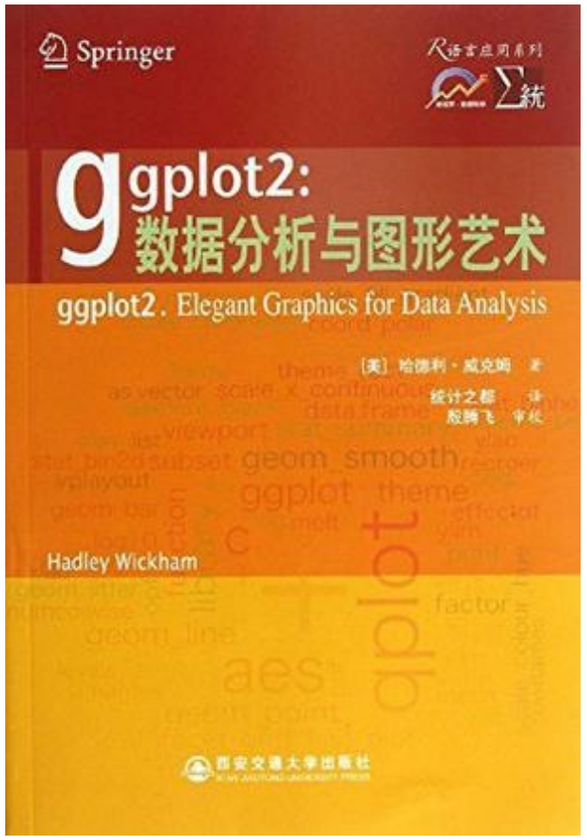
R语言卡片

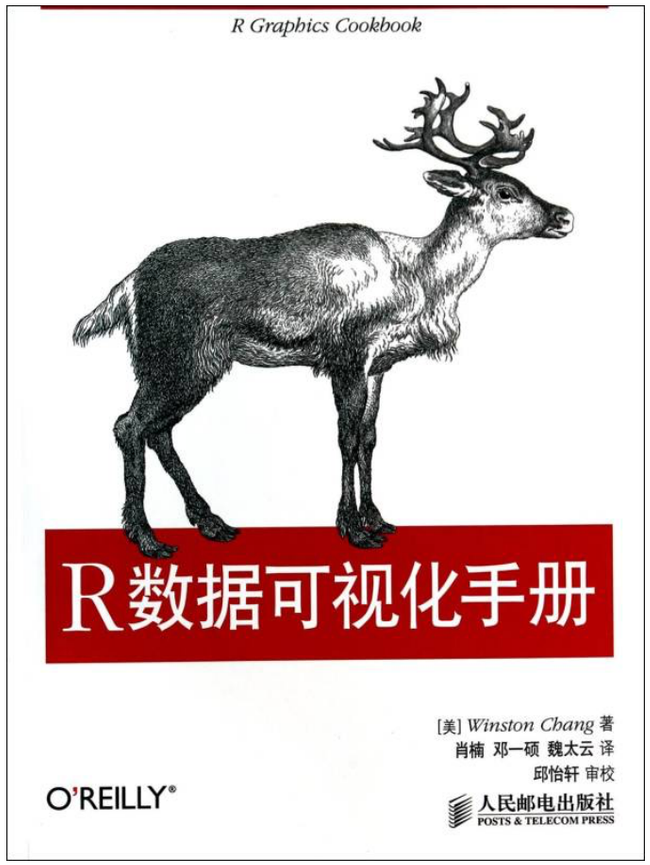
练习

练习 - 0024

- ggplot2的1-4章，熟悉所有例子。
- R数据可视化手册的2-5章，熟悉所有例子。
- 教材RIA（第二版）的第19章，熟悉所有例子。







练习 - 0025

- dapengde_DummyR_PM25.csv是2003 年 8 月在北京城区的三个高度（8 米，100 米，325 米）测得的 PM2.5 的质量浓度日变化的统计数据，共 4 列 25 行。
  - ➡ 请画出一条折线表示h8和time的关系，要求是"time"和"pm2.5" 分别是x轴的名称和y轴的名称，lty=1（表示line的type为1，表示直线）y轴的范围是0到200.
  - ➡ 在上图增加一条折线(使用lines()函数)表示h100和time的关系，要求颜色为红色，线型为虚线(lty=2)
  - ➡ 在上图中增加图例来表示上边画的两条折线，其中图例位置为（x=15，y=180）位置处，内容为8m和100m,两条折线分别为黑色直线和红色虚线。
  - ➡ 画出x轴，刻度指定为和时间相对应的24个小时。
  - ➡ 与h8和h100两条折线相对应，画出其对应的y轴均值的水平线。

- 基本绘图、qplot、ggplot

练习 - 0026

- 使用数据集airquality回答下列问题
  1）使用str()函数来观察airquality这个数据的变量有那些:
  2）用函数计算第三个变量（风速）的平均值，最小值，最大值和标准差:
  3）使用pdf("mygraph.pdf") 将上面的图形保存到你的作业文件夹（本地硬盘）
  4）用plot()函数创建风速与风度的散点图：添加回归曲线和标题“Weather in NYC”：

- 使用R自带的数据集cars画出散点图，颜色设置为彩虹色，形状为编码为1:10的图形。主标题为“speed and diantance”，主标题颜色为蓝色，主标题缩放比例为1.5，字体为2，副标题为“scatter plot”，副标题颜色为灰色，主标题缩放比例为1.2

- 画出数据框cars的speed列的频率直方图，主标题为“speed hist”，主标题颜色为蓝色，主标题缩放比例为1.5，字体为2，副标题为“histogram exercise”，副标题颜色为灰色，主标题缩放比例为1.2，y轴范围为0到0.1
  添加密度曲线，要求颜色为红色，线段类型为虚线，宽度为2

- 使用R数据集VADeaths,查看这个数据集，画出各个年龄段死亡率的箱型图，要求并排排列，颜色为前4个彩虹色，添加图例，图例名称为VADeaths的列名，y轴范围为0到100，主标题为“VADeaths barplot”，主标题颜色为蓝色，主标题字体为2，副标题为“barplot exercise”，副标题颜色为灰色，主标题缩放比例为1.5

- 基本绘图、qplot、ggplot

练习 – 0027

INTERACTIVE COURSE

Data Visualization with ggplot2 (Part 1)

Start Course For Free | Bookmark

5 hours | 14 Videos | 62 Exercises | 141,097 Participants | 5,250 XP

提交方式和上节课一样！

https://www.datacamp.com/courses

INTERACTIVE COURSE

Data Visualization with ggplot2 (Part 2)

Start Course For Free | Bookmark

5 hours | 11 Videos | 55 Exercises | 51,466 Participants | 4,750 XP

谢谢！

孙惠平

sunhp@ss.pku.edu.cn