R数据可视化—ggplot2包 第1周

2013.2.8

法律声明

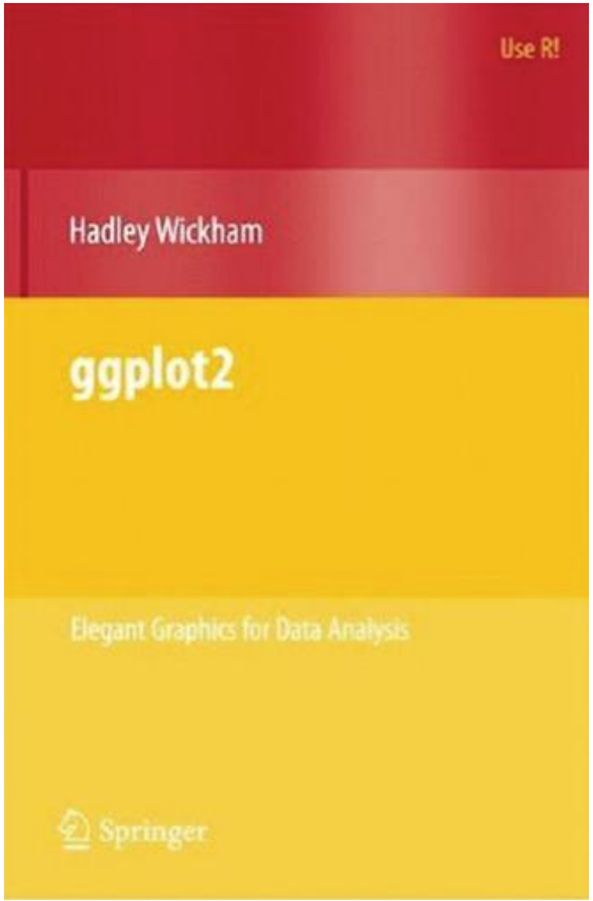
【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

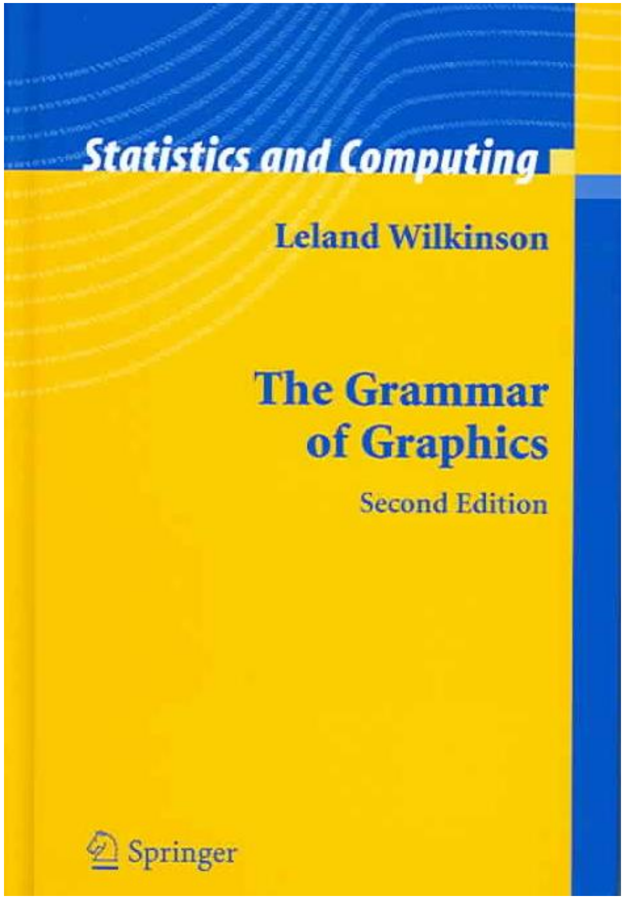
课程详情访问炼数成金培训网站

http://edu.dataguru.cn

2013.2.8

参考书


Use R!
Hadley Wickham
ggplot2
Elegant Graphics for Data Analysis
Springer


Statistics and Computing
Leland Wilkinson
The Grammar of Graphics
Second Edition
Springer

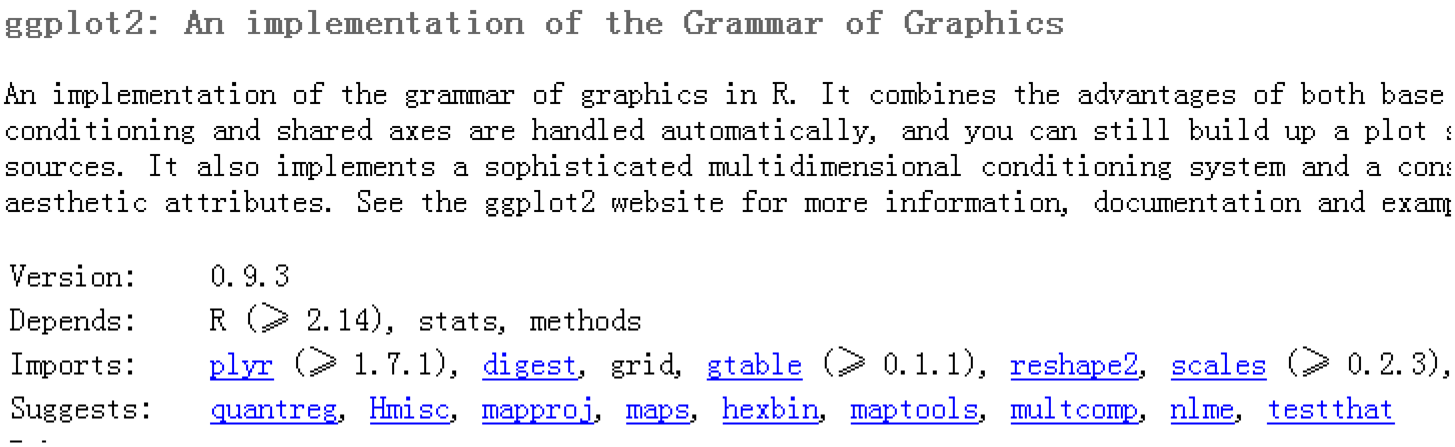
2013.2.8

ggplot2包简介

- 由Hadley Wickham于2005年创建
- 具有理论基础的图形包，基于《the Grammar of Graphics》(Wilkinson, 2005)，这也是它名称的由来
- 能媲美商业数据可视化软件的作图效果，使用“层”的概念，容易上手，可以非常简单地画出复杂的统计图表
- 于2012年初进行了重大更新，最新版本0.9.3

网上资源

- 官网：http://had.co.nz/ggplot2
- CRAN下载：http://cran.r-project.org/web/packages/ggplot2/
- 本书网页：http://had.co.nz/ggplot2/book
- 讨论组：http://groups.google.com/group/ggplot2



ggplot2: An implementation of the Grammar of Graphics

An implementation of the grammar of graphics in R. It combines the advantages of both base conditioning and shared axes are handled automatically, and you can still build up a plot sources. It also implements a sophisticated multidimensional conditioning system and a con aesthetic attributes. See the ggplot2 website for more information, documentation and exam

```
Version:     0.9.3
Depends:     R (≥ 2.14), stats, methods
Imports:     plyr (≥ 1.7.1), digest, grid, gtable (≥ 0.1.1), reshape2, scales (≥ 0.2.3),
Suggests:    quantreg, Hmisc, mapproj, maps, hexbin, maptools, multcomp, nlme, testthat
```

the grammar of graphics

- Wilkinson 在2005年所写的关于统计图形的总结性抽象
- 主要观点： **a statistical graphic is a mapping from data to aesthetic attributes (colour, shape, size) of geometric objects (points, lines, bars)**
- 基本概念：数据（Data）和映射（Mapping），标度（Scale），几何对象（Geometric），统计变换（Statistics），坐标（Coordinate），图层（Layer），面（Facet）

```
DATA: longitude, latitude = map(source("World"))
TRANS: bd = max(birth-death, 0)
COORD: project.mercator()
ELEMENT: point(position(lon*lat), size(bd), color(color.red))
ELEMENT: polygon(position(longitude*latitude))
```

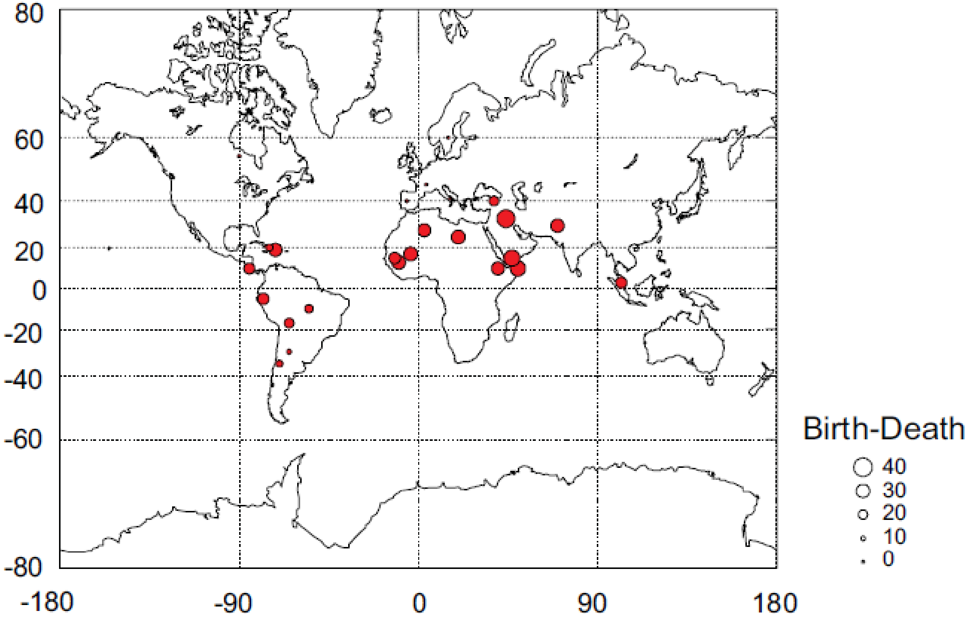


Figure 1.5 Excess birth (vs. death) rates in selected countries

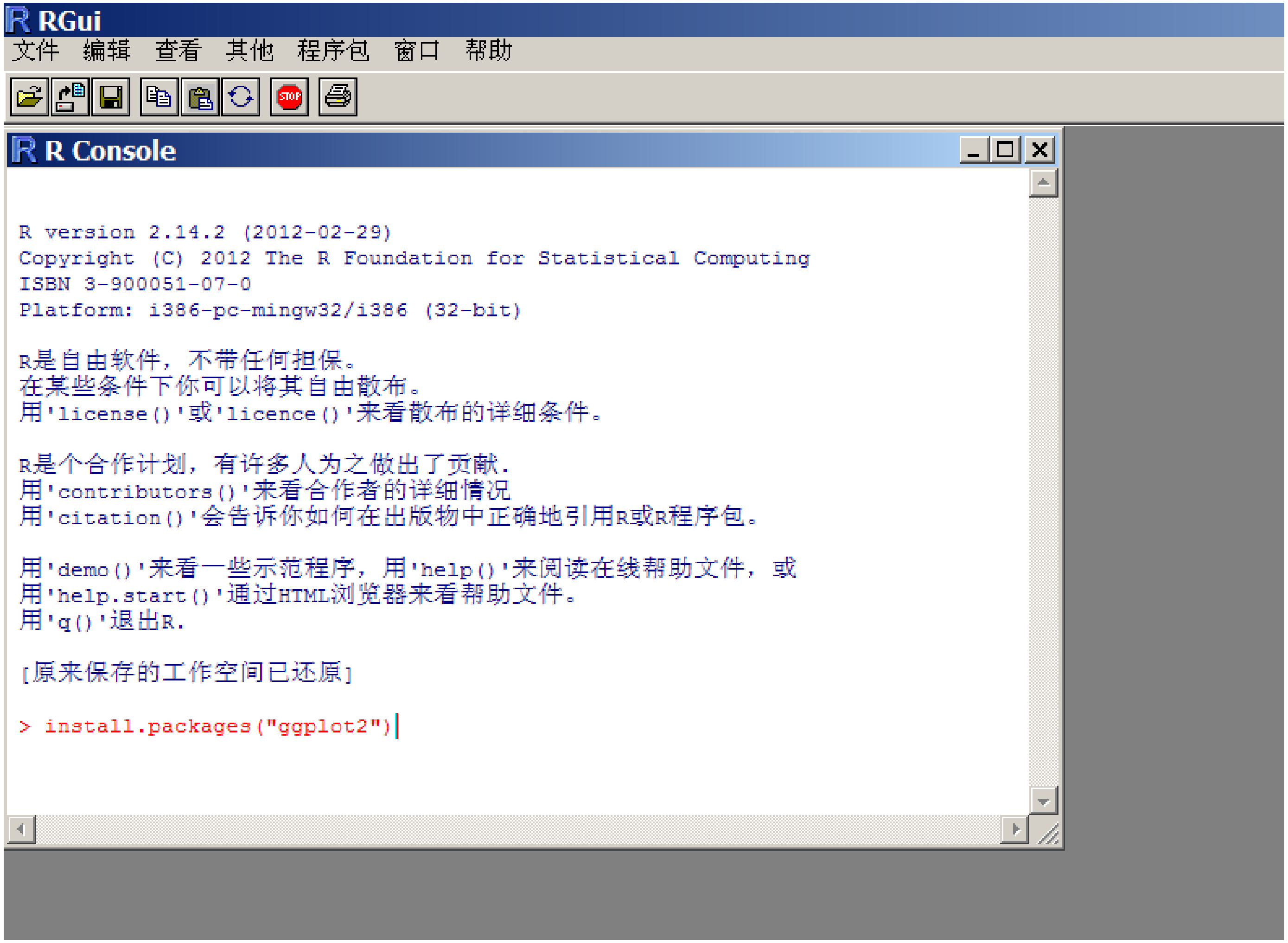
ggplot2 vs R基础作图方式

- 函数繁杂，语法（参数）繁多，用户需要记忆的东西太多，非常依赖用户的经验
- 传统的“笔-纸”工作模式，只能追加，不能删除或修改现有内容
- 自动化程度低，一些常见的图形也要用户手工描画（例如“图例”）
- 主次不分，主要图形和次要部件全部堆放在一段代码里面，通过代码难以识别图形的主要部分和层次结构
- 忘记一切，从头开始

R其它作图包

- Grid包。 Paul Murrell在2000年基于他之前在攻读博士学位时的工作所创建。允许多坐标系，可修改的图形对象等特点，使其更容易实现复杂的图形
- Lattice包。 Deepayan Sarkar在2008年创建。基于grid包实现了the trellis graphics system of Cleveland。
- R的图形包概述： http://cran.r-project.org/web/views/Graphics.html

安装ggplot2包



```
R version 2.14.2 (2012-02-29)
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)

R是自由软件，不带任何担保。
在某些条件下你可以将其自由散布。
用'license()'或'licence()'来看散布的详细条件。

R是个合作计划，有许多人为之做出了贡献.
用'contributors()'来看合作者的详细情况
用'citation()'会告诉你如何在出版物中正确地引用R或R程序包。

用'demo()'来看一些示范程序，用'help()'来阅读在线帮助文件，或
用'help.start()'通过HTML浏览器来看帮助文件。
用'q()'退出R.

[原来保存的工作空间已还原]

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

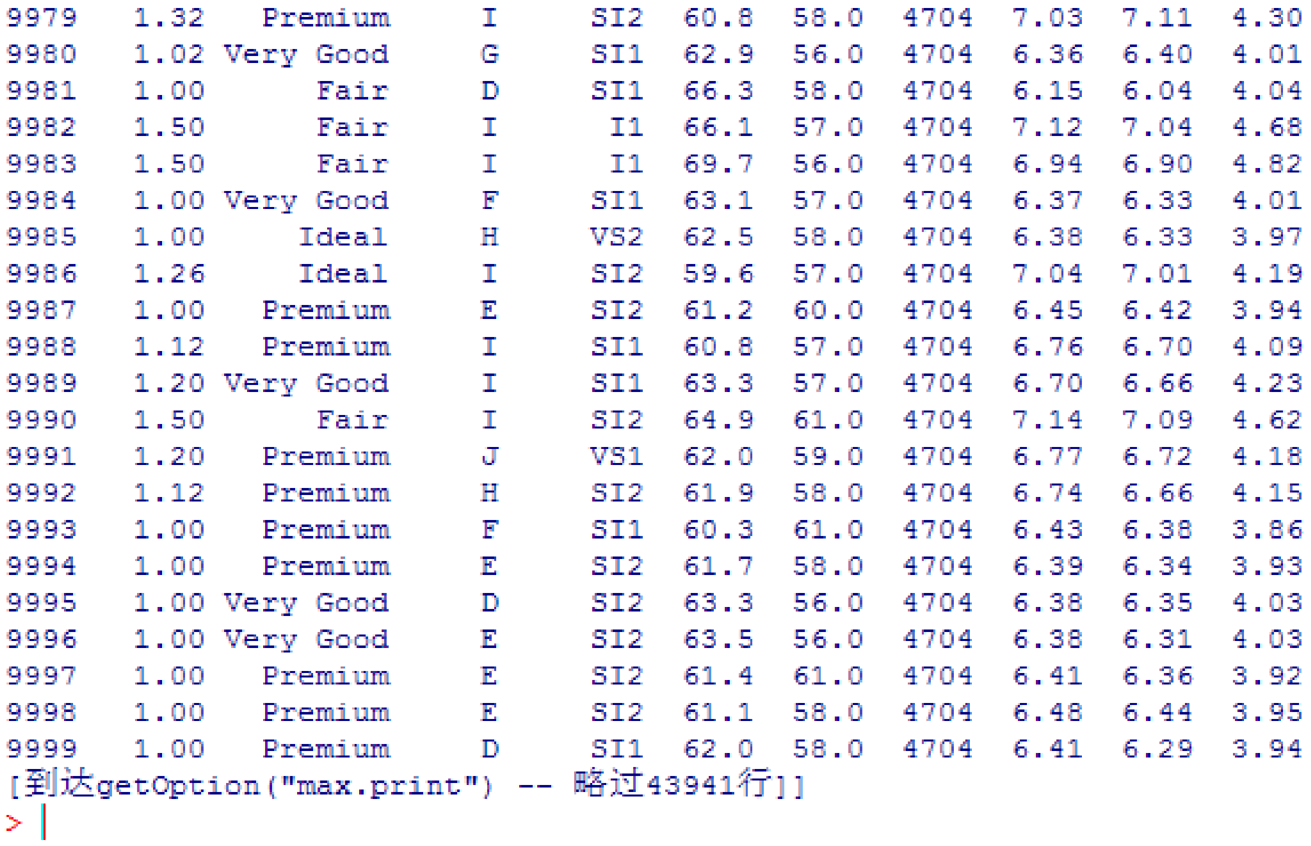
2013.2.8

从qplot函数开始

- qplot的基本用法
- 把变量映射到装饰属性（颜色，形状，大小等）
- 使用不同的几何对象实现几种常见图形
- 使用facet实现分组作图
- 一些基本选项
- qplot vs plot

钻石数据集

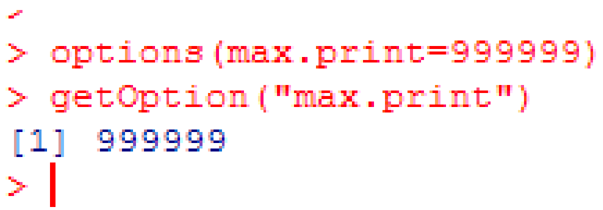
- Ggplot2包自带，包含有54000多条钻石信息



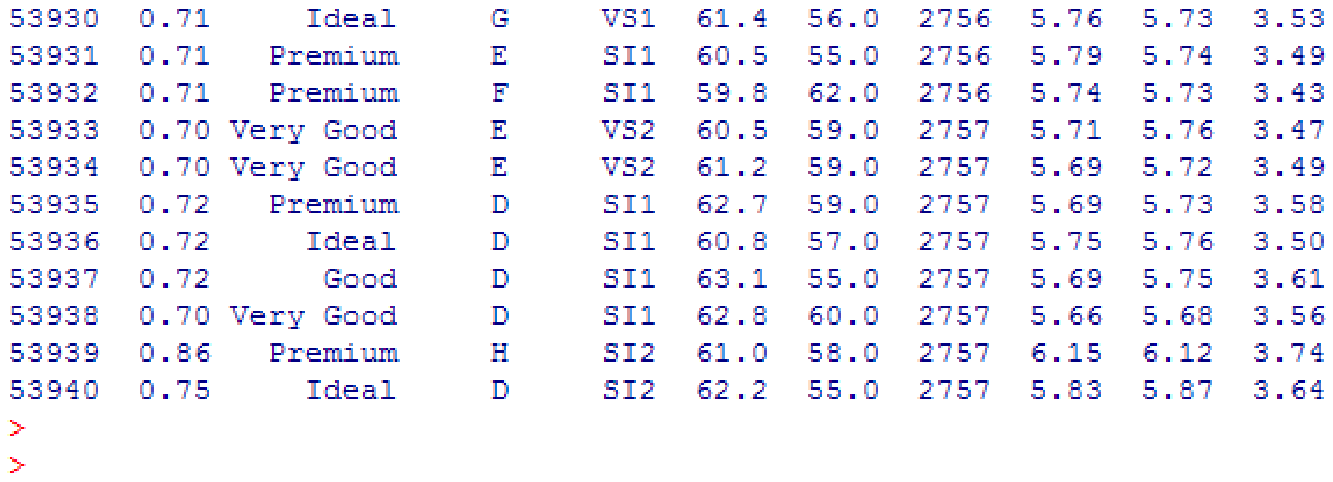
```
9979  1.32   Premium     I     SI2  60.8  58.0  4704  7.03  7.11  4.30
9980  1.02 Very Good     G     SI1  62.9  56.0  4704  6.36  6.40  4.01
9981  1.00      Fair     D     SI1  66.3  58.0  4704  6.15  6.04  4.04
9982  1.50      Fair     I      I1  66.1  57.0  4704  7.12  7.04  4.68
9983  1.50      Fair     I      I1  69.7  56.0  4704  6.94  6.90  4.82
9984  1.00 Very Good     F     SI1  63.1  57.0  4704  6.37  6.33  4.01
9985  1.00     Ideal     H     VS2  62.5  58.0  4704  6.38  6.33  3.97
9986  1.26     Ideal     I     SI2  59.6  57.0  4704  7.04  7.01  4.19
9987  1.00   Premium     E     SI2  61.2  60.0  4704  6.45  6.42  3.94
9988  1.12   Premium     I     SI1  60.8  57.0  4704  6.76  6.70  4.09
9989  1.20 Very Good     I     SI1  63.3  57.0  4704  6.70  6.66  4.23
9990  1.50      Fair     I     SI2  64.9  61.0  4704  7.14  7.09  4.62
9991  1.20   Premium     J     VS1  62.0  59.0  4704  6.77  6.72  4.18
9992  1.12   Premium     H     SI2  61.9  58.0  4704  6.74  6.66  4.15
9993  1.00   Premium     F     SI1  60.3  61.0  4704  6.43  6.38  3.86
9994  1.00   Premium     E     SI2  61.7  58.0  4704  6.39  6.34  3.93
9995  1.00 Very Good     D     SI2  63.3  56.0  4704  6.38  6.35  4.03
9996  1.00 Very Good     E     SI2  63.5  56.0  4704  6.38  6.31  4.03
9997  1.00   Premium     E     SI2  61.4  61.0  4704  6.41  6.36  3.92
9998  1.00   Premium     E     SI2  61.1  58.0  4704  6.48  6.44  3.95
9999  1.00   Premium     D     SI1  62.0  58.0  4704  6.41  6.29  3.94
[到达getOption("max.print") -- 略过43941行]]
> 
```

2013.2.8

处理max.print问题



```
> options(max.print=999999)
> getOption("max.print")
[1] 999999
> 
```



```
53930  0.71      Ideal     G     VS1  61.4  56.0  2756  5.76  5.73  3.53
53931  0.71    Premium     E     SI1  60.5  55.0  2756  5.79  5.74  3.49
53932  0.71    Premium     F     SI1  59.8  62.0  2756  5.74  5.73  3.43
53933  0.70  Very Good     E     VS2  60.5  59.0  2757  5.71  5.76  3.47
53934  0.70  Very Good     E     VS2  61.2  59.0  2757  5.69  5.72  3.49
53935  0.72    Premium     D     SI1  62.7  59.0  2757  5.69  5.73  3.58
53936  0.72      Ideal     D     SI1  60.8  57.0  2757  5.75  5.76  3.50
53937  0.72       Good     D     SI1  63.1  55.0  2757  5.69  5.75  3.61
53938  0.70  Very Good     D     SI1  62.8  60.0  2757  5.66  5.68  3.56
53939  0.86    Premium     H     SI2  61.0  58.0  2757  6.15  6.12  3.74
53940  0.75      Ideal     D     SI2  62.2  55.0  2757  5.83  5.87  3.64
>
>
```

数据含义

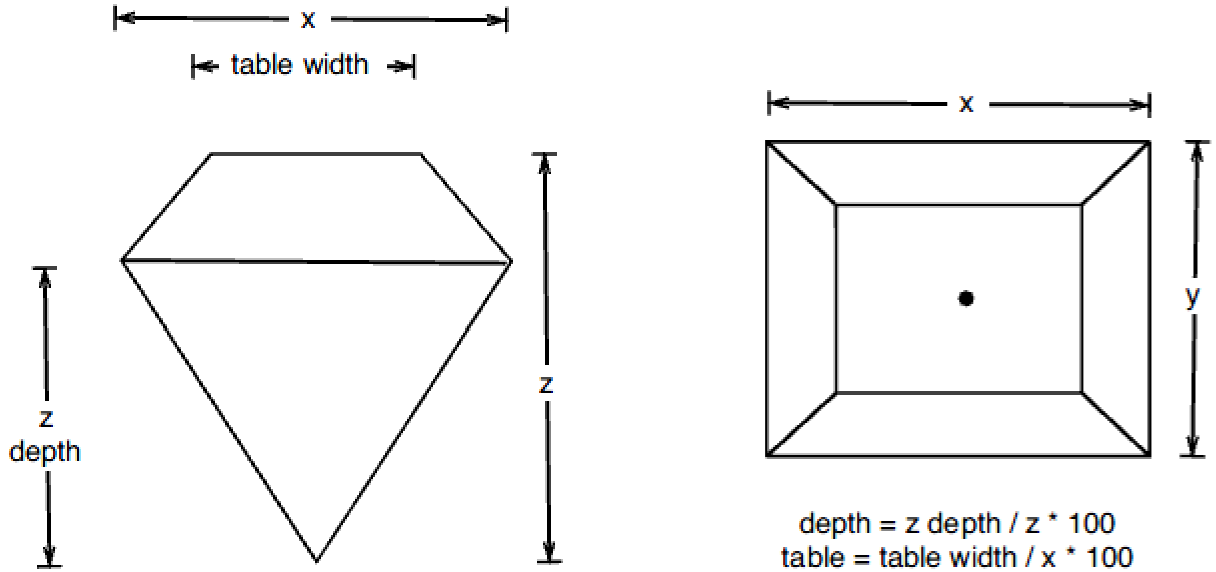
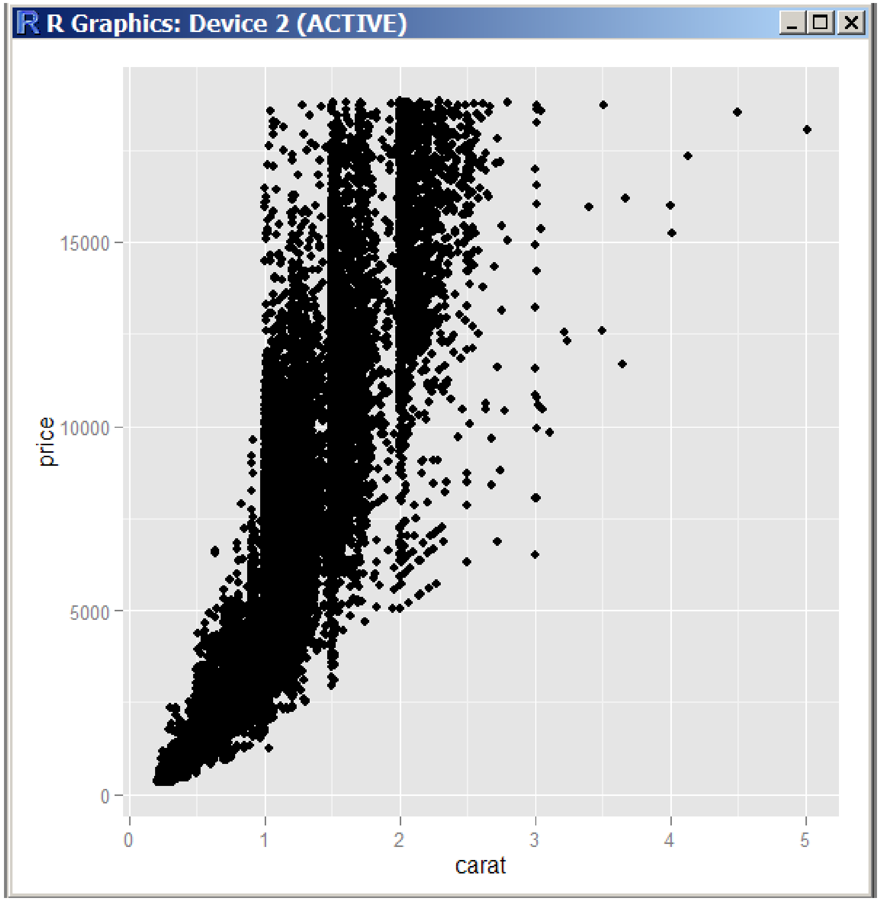


Fig. 2.1: How the variables x, y, z, table and depth are measured.

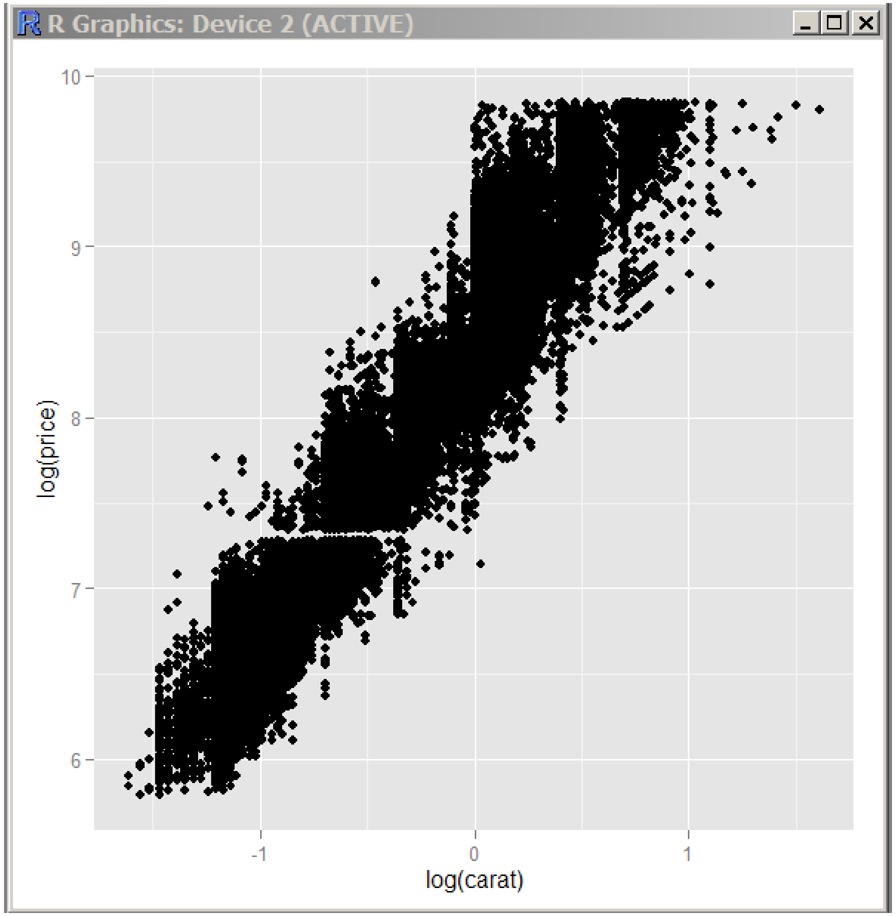
第一个散点图

qplot(carat, price, data = diamonds)



对数变换

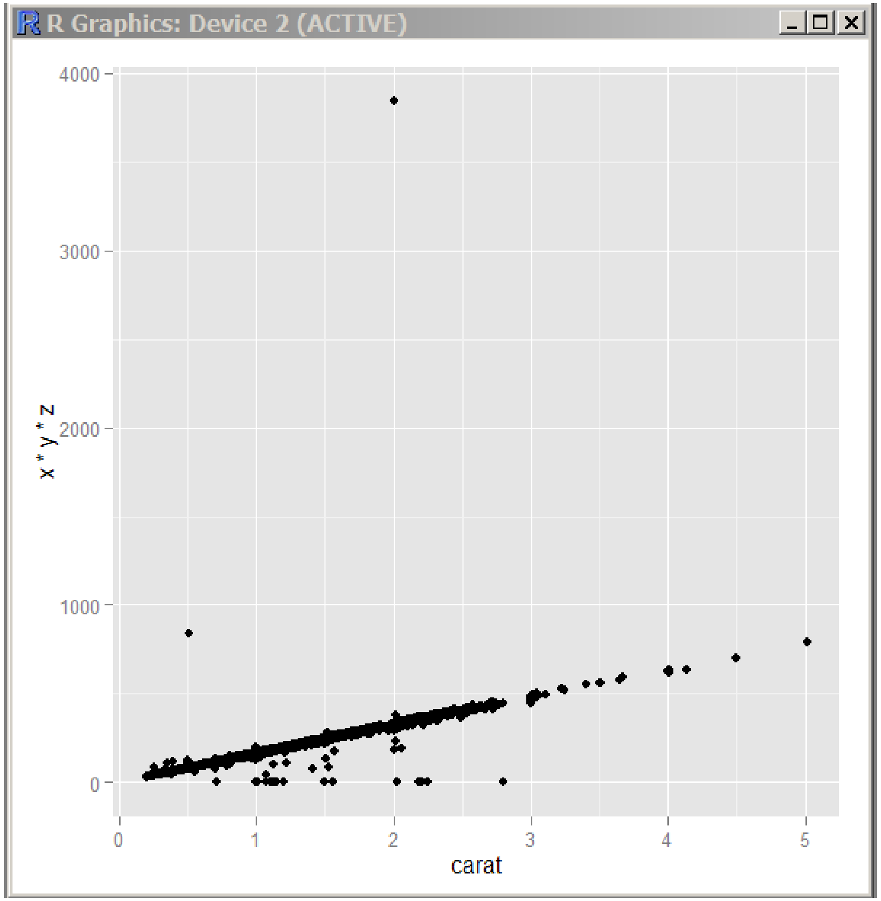
qplot(log(carat), log(price), data = diamonds)



2013.2.8

体积 vs 重量

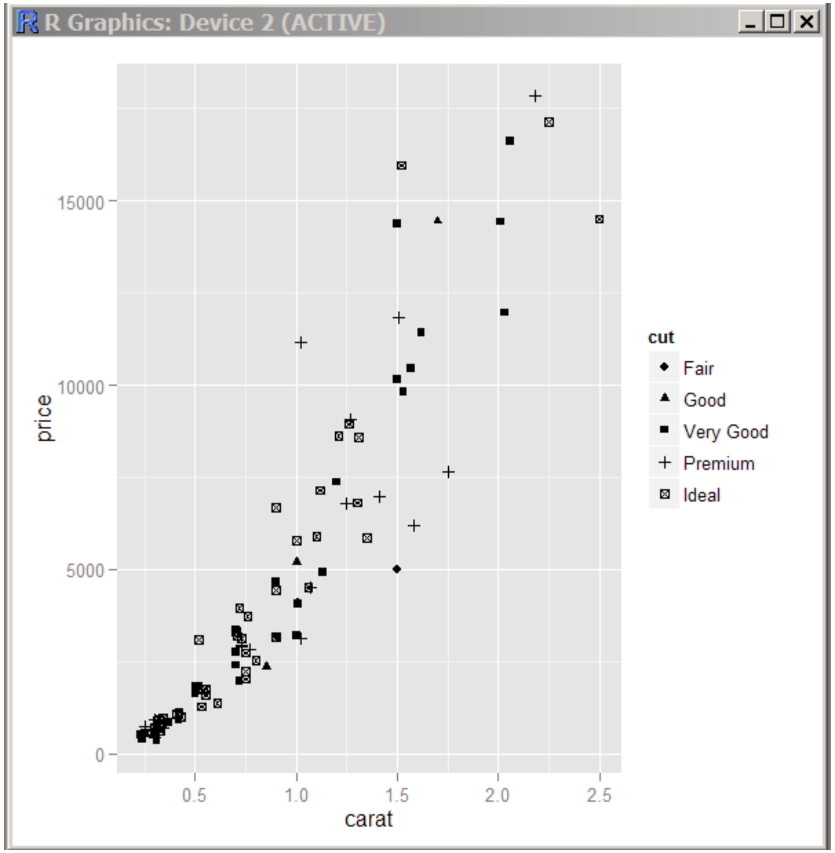
qplot(carat, x * y * z, data = diamonds)

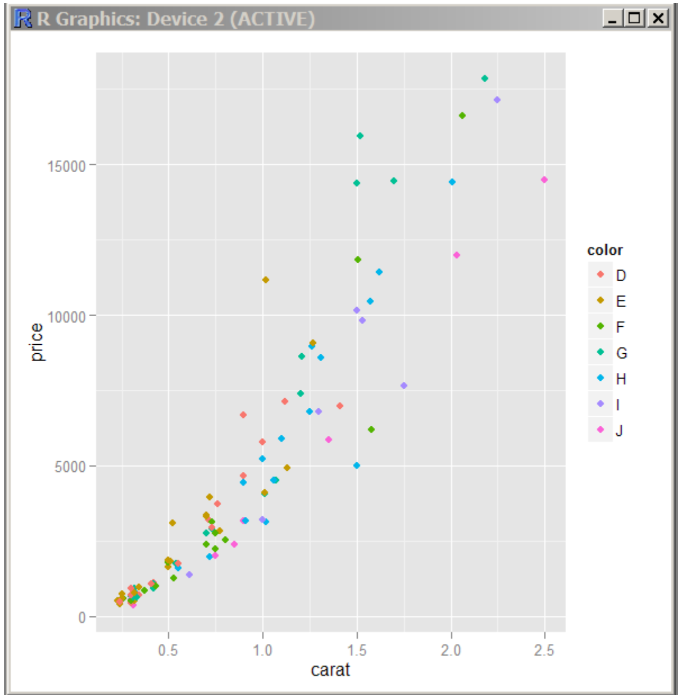


2013.2.8

装饰属性

```
set.seed(1410) # Make the sample reproducible
dsmall <- diamonds[sample(nrow(diamonds), 100), ]
qplot(carat, price, data = dsmall, colour = color)
qplot(carat, price, data = dsmall, shape = cut)
```



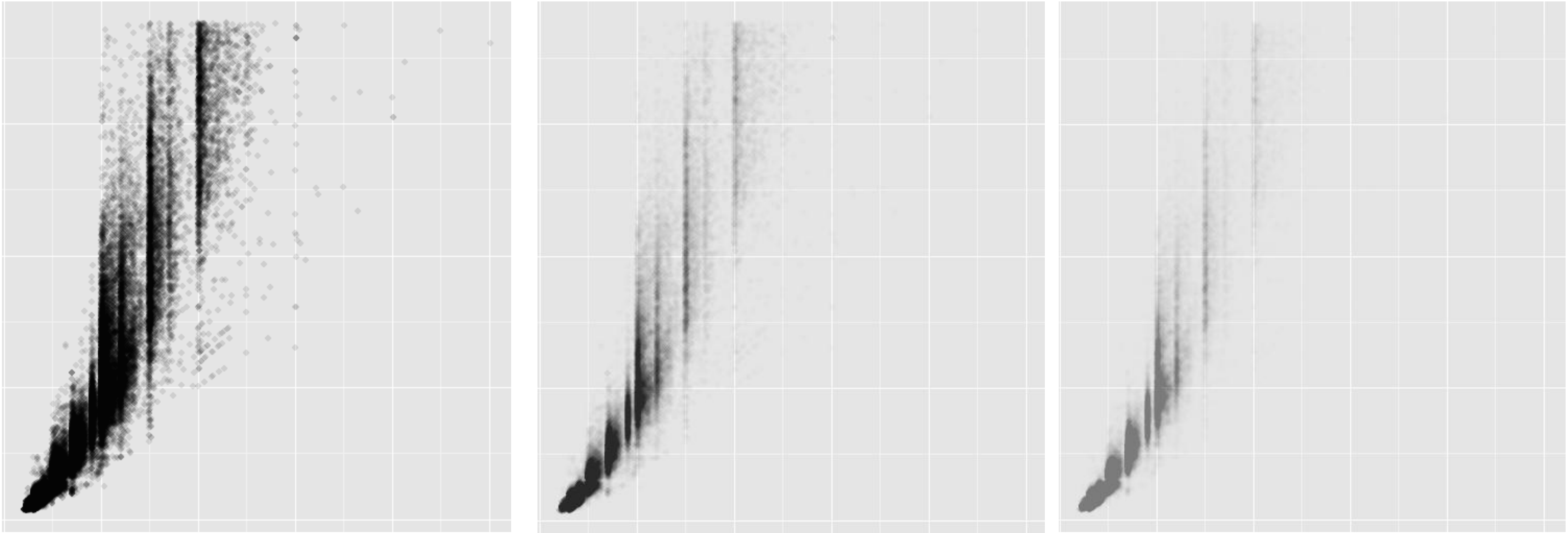


Alpha值

```
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))
```

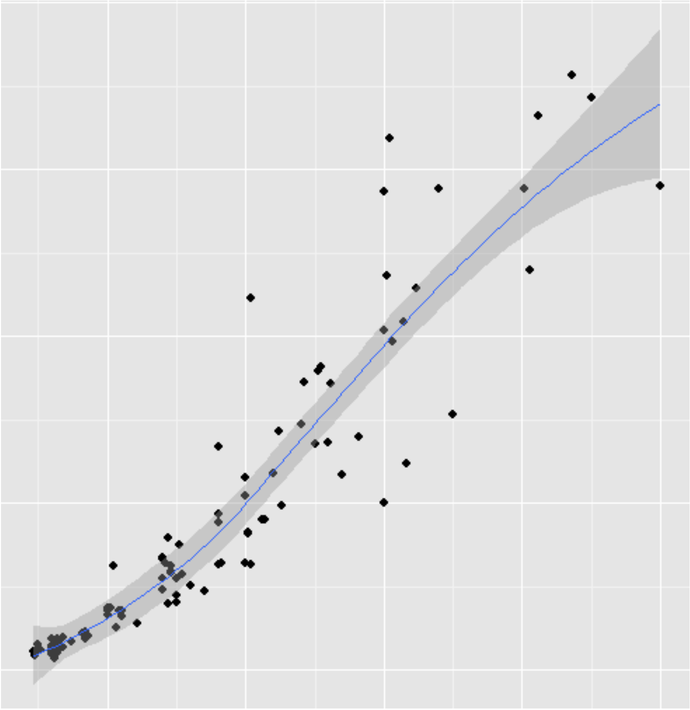


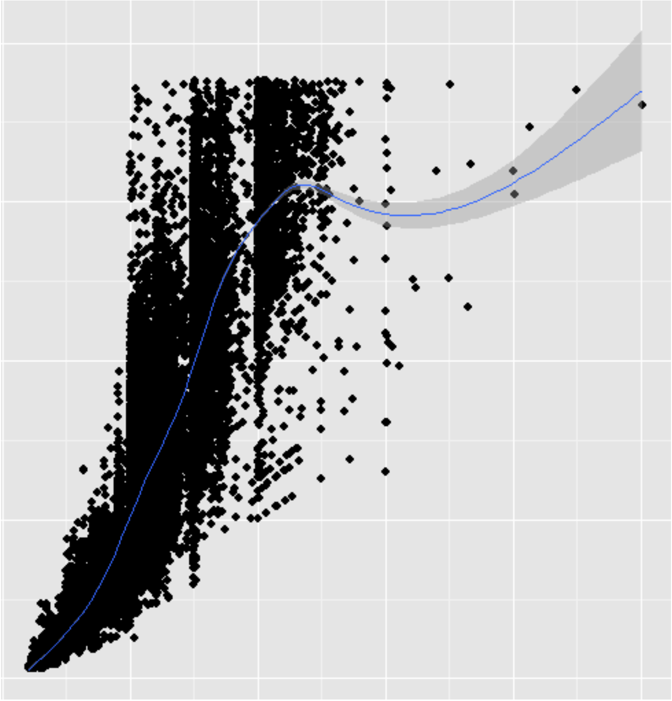
几何对象

- geom = “point” ，画散点图，当提供x,y时为缺省选项
- geom = “smooth” ，画平滑曲线及标准误
- geom = “boxplot” ，画箱线图
- geom = “path” 或geom = “line” ，画连线
- geom = “histogram” ，画直方图，当只提供x时为缺省选项
- geom = “freqpoly” ，画频率多边形
- geom = “density” ，画密度曲线
- geom = “bar” ，画柱形图

平滑曲线

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

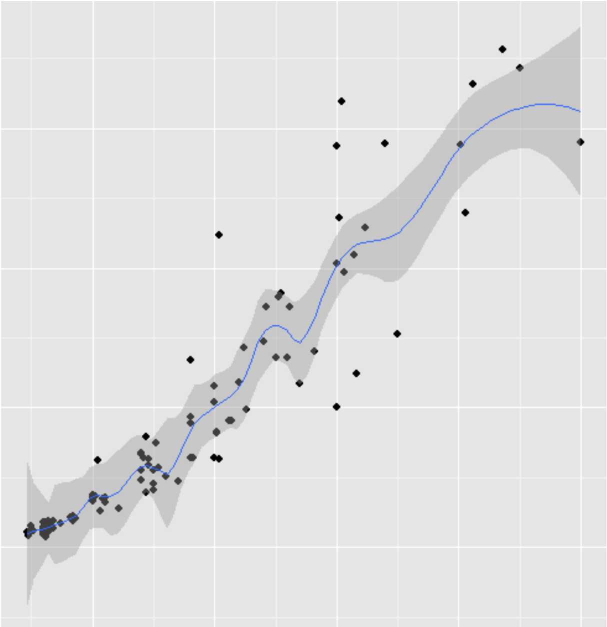


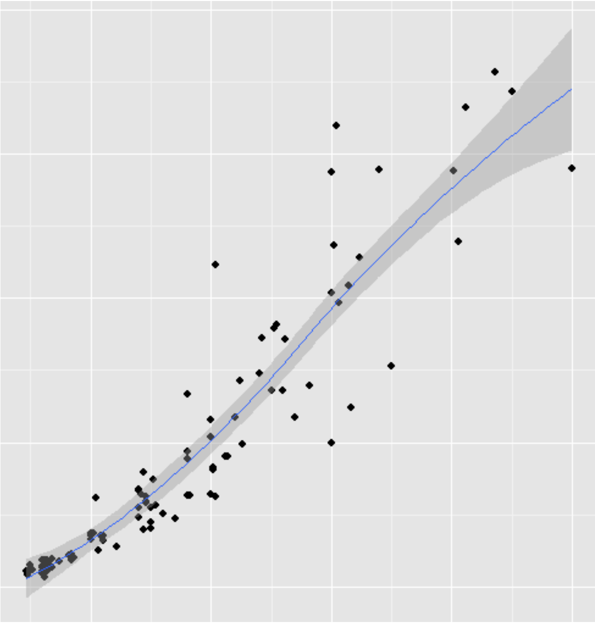


多项式拟合

- method = “loess” ，对于较小的n为缺省拟合方式(n<1000)
- 弯曲程度取决于span

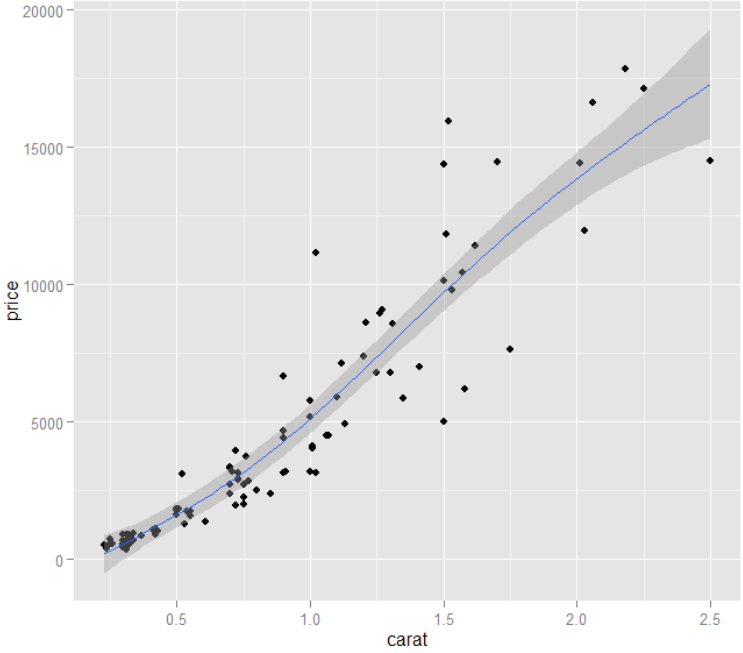
```
qplot(carat, price, data = dsmall, geom = c("point", "smooth"),span = 0.2)
qplot(carat, price, data = dsmall, geom = c("point", "smooth"),span = 1)
```





GAM

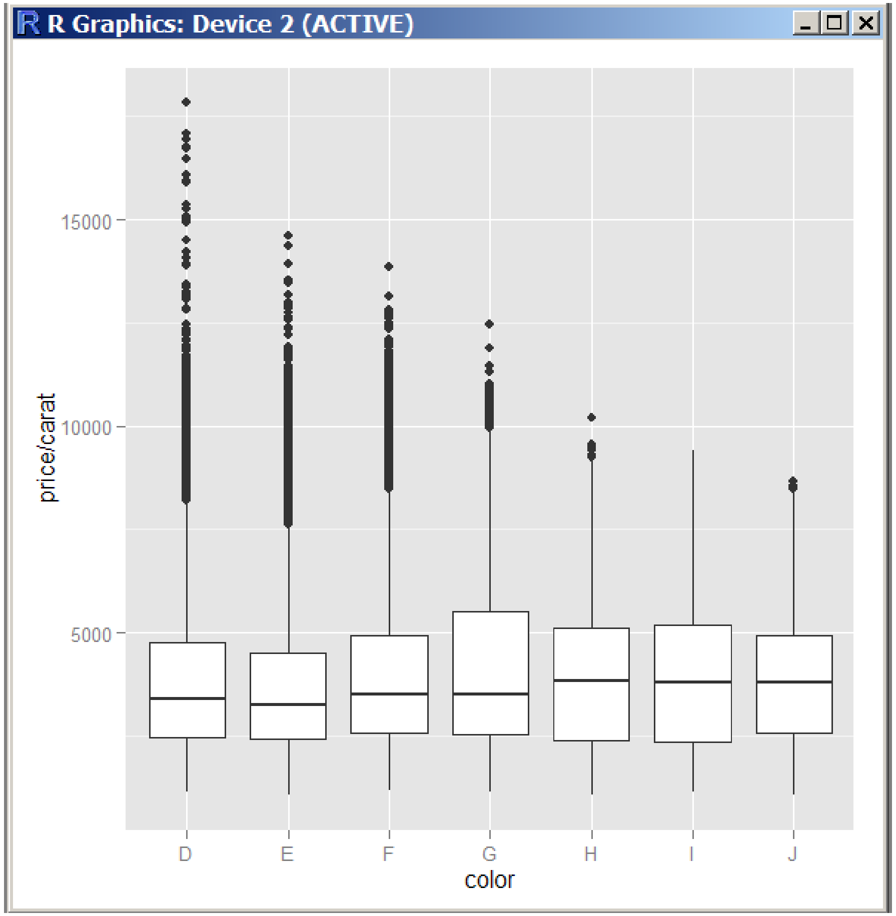
```
qplot(carat, price, data = dsmall, geom = c("point", "smooth"),method = "gam",
    formula = y ~ s(x))
qplot(carat, price, data = dsmall, geom = c("point", "smooth"),method = "gam",
    formula = y ~ s(x, bs = "cs"))
```



2013.2.8

箱线图

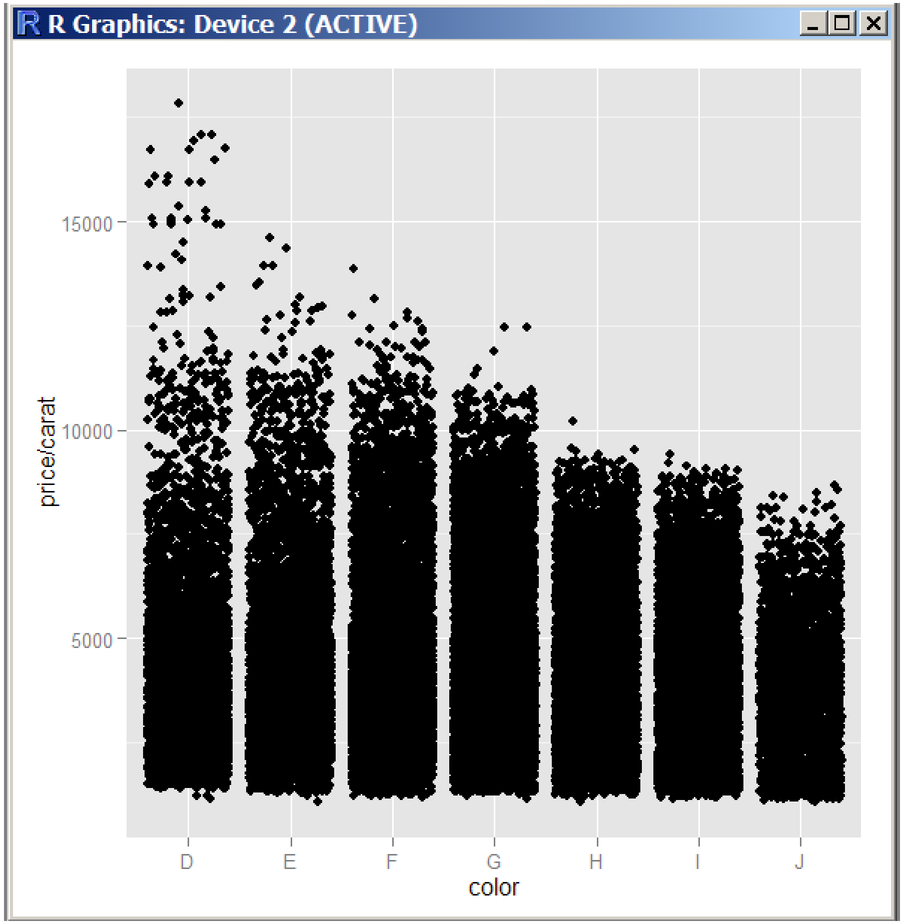
qplot(color, price / carat, data = diamonds, geom = "boxplot")



2013.2.8

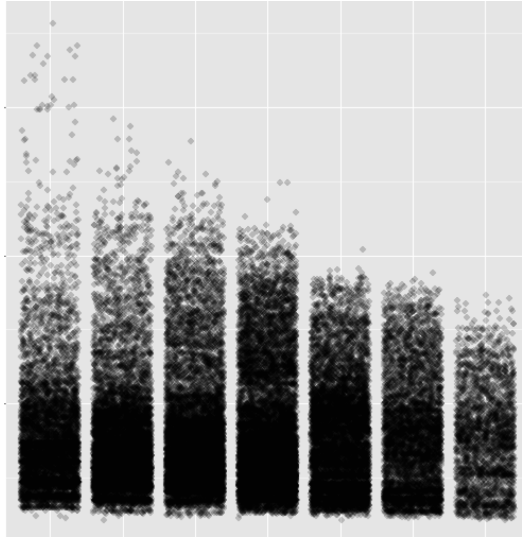
jitter

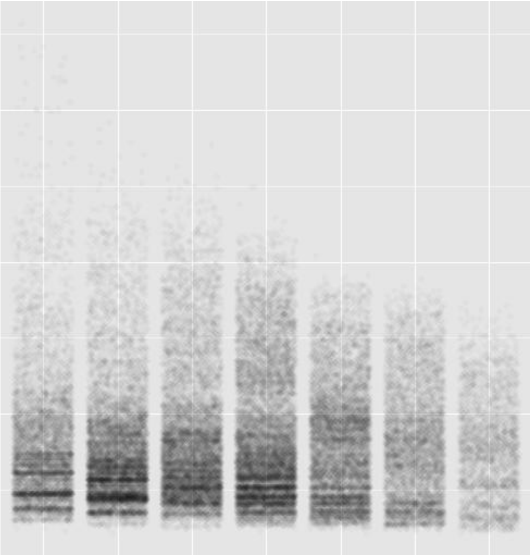
qplot(color, price / carat, data = diamonds, geom = "jitter")

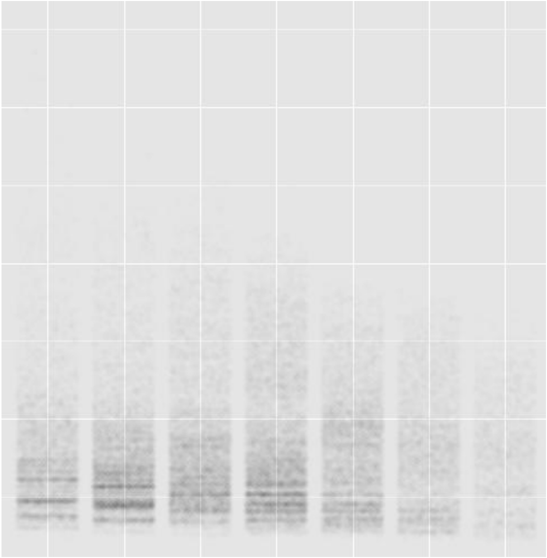


jitter

```
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))
```

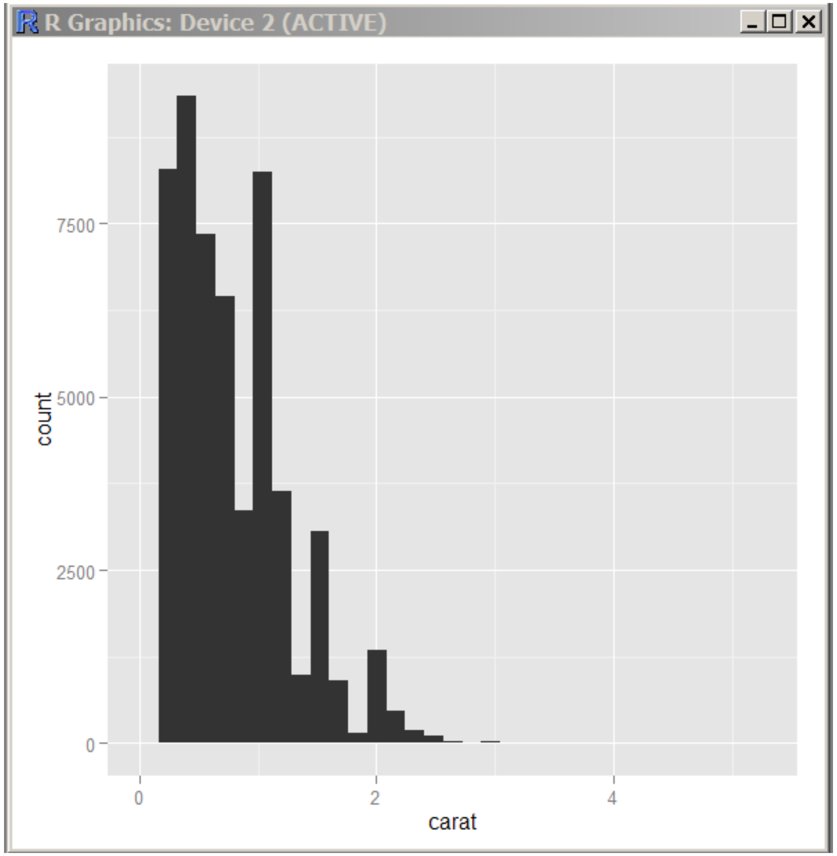






直方图

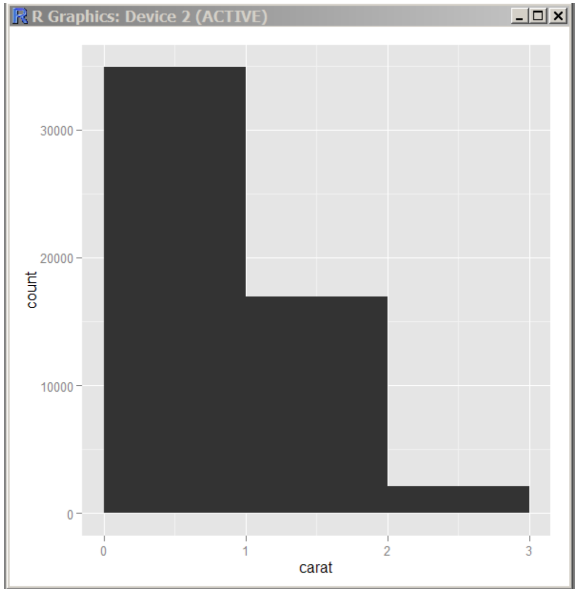
qplot(carat, data = diamonds, geom = "histogram")

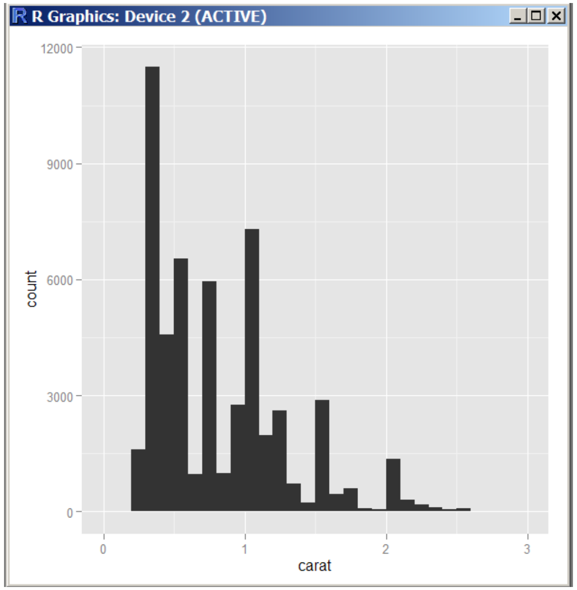


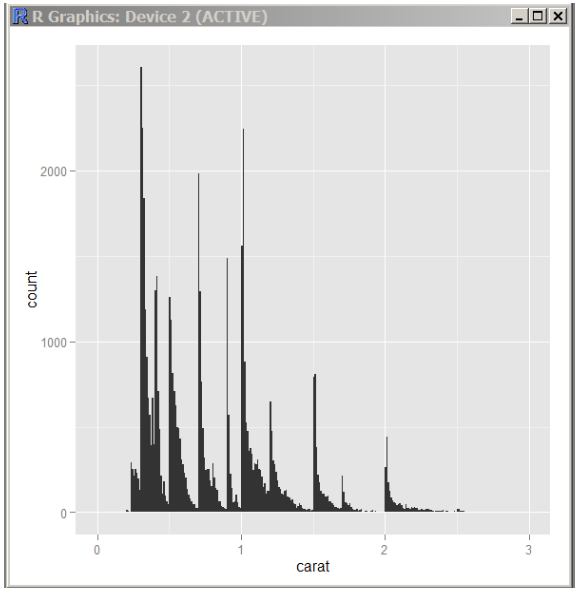
2013.2.8

设置直方图的区间

```
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))
```

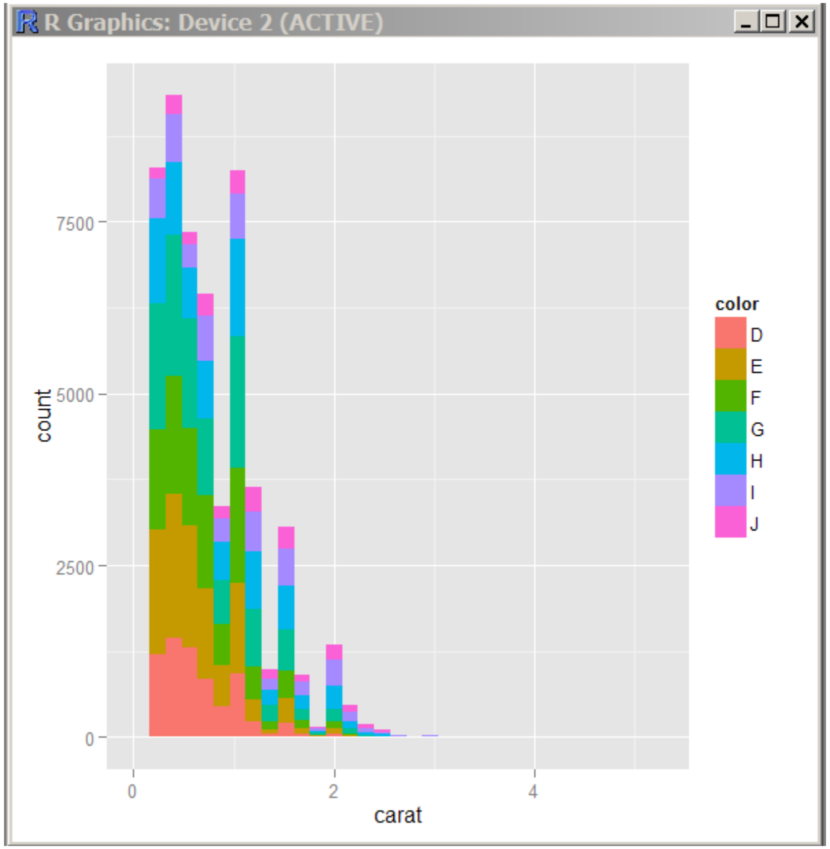






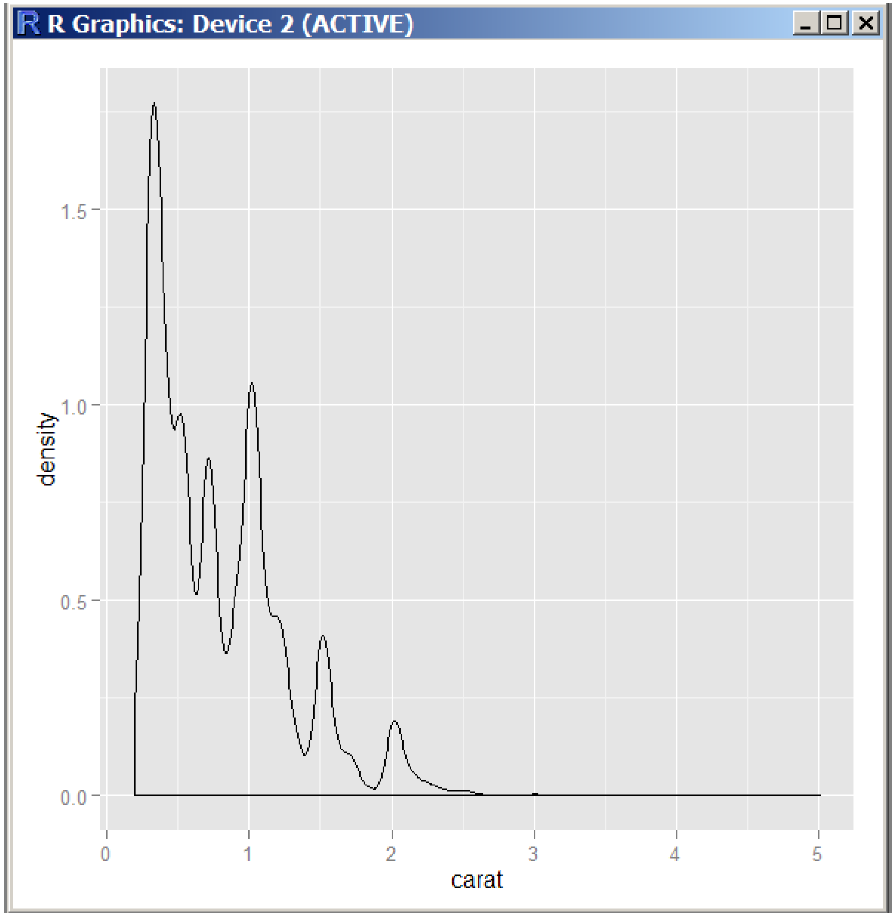
设置色彩

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



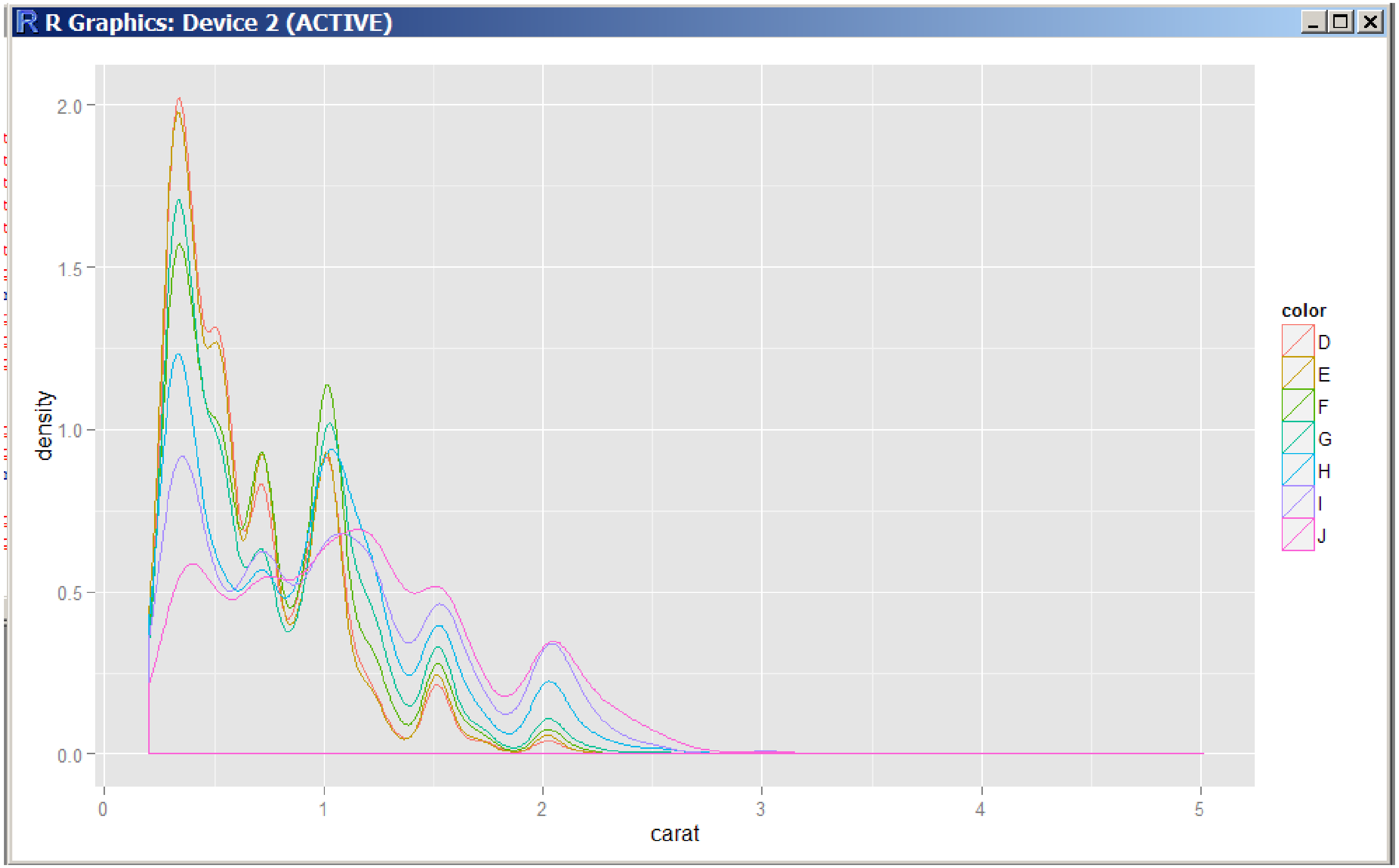
密度曲线图

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



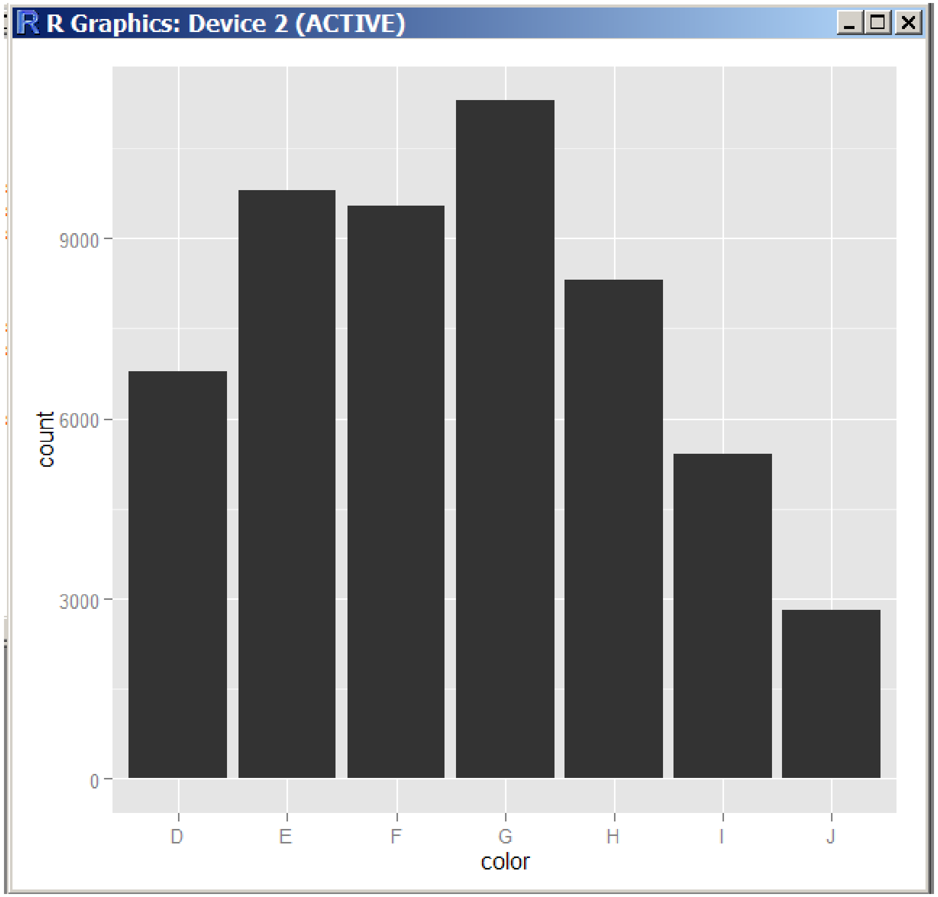
设置彩色

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



柱状图

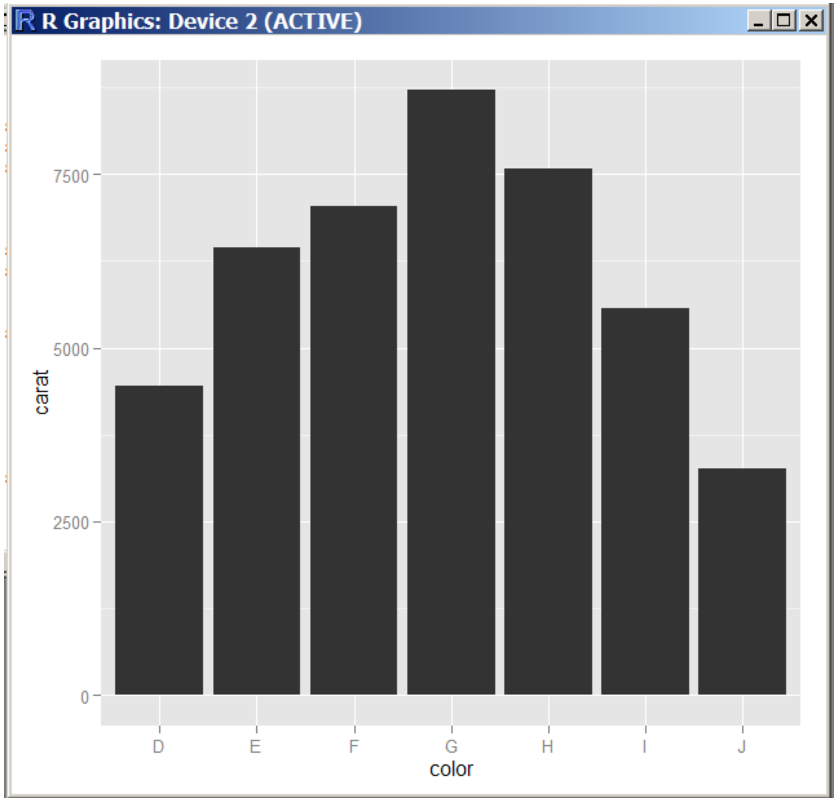
qplot(color, data = diamonds, geom = "bar")



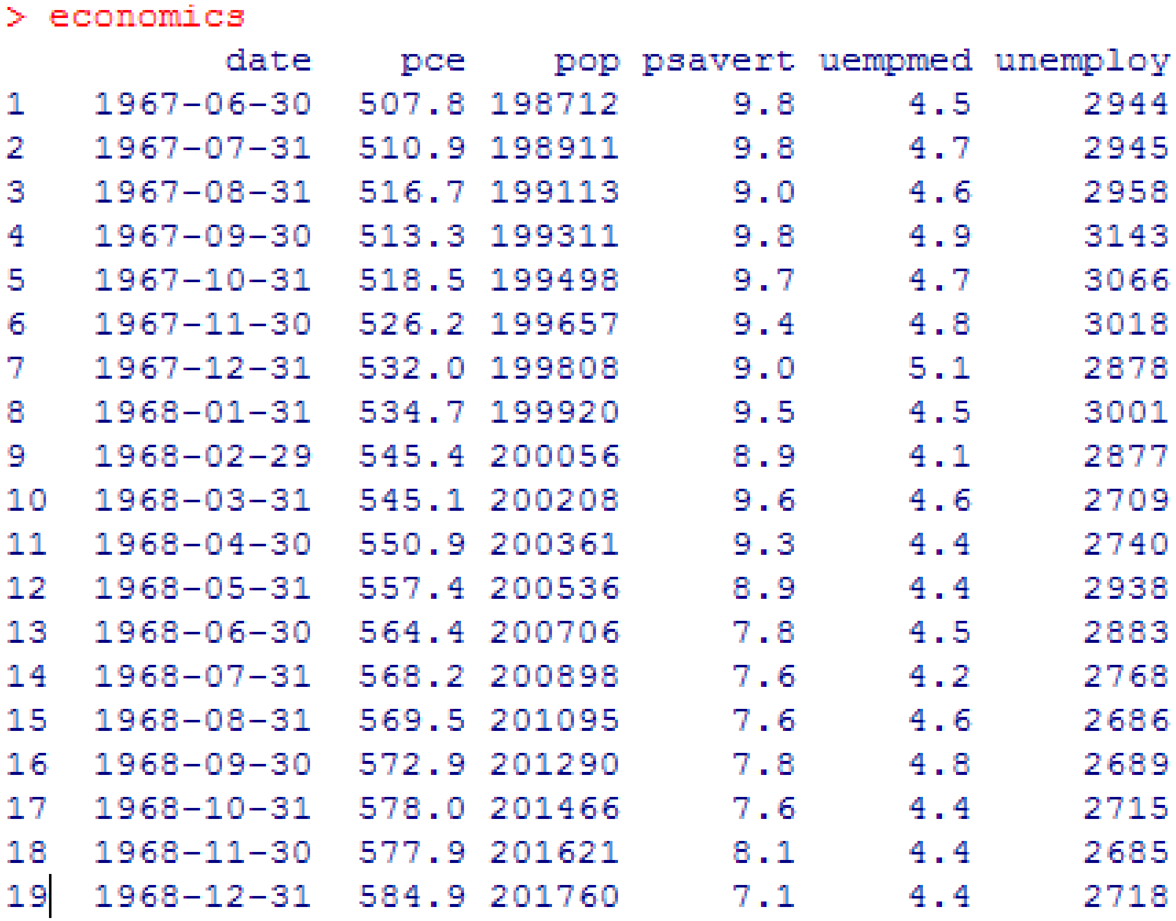
2013.2.8

求和

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



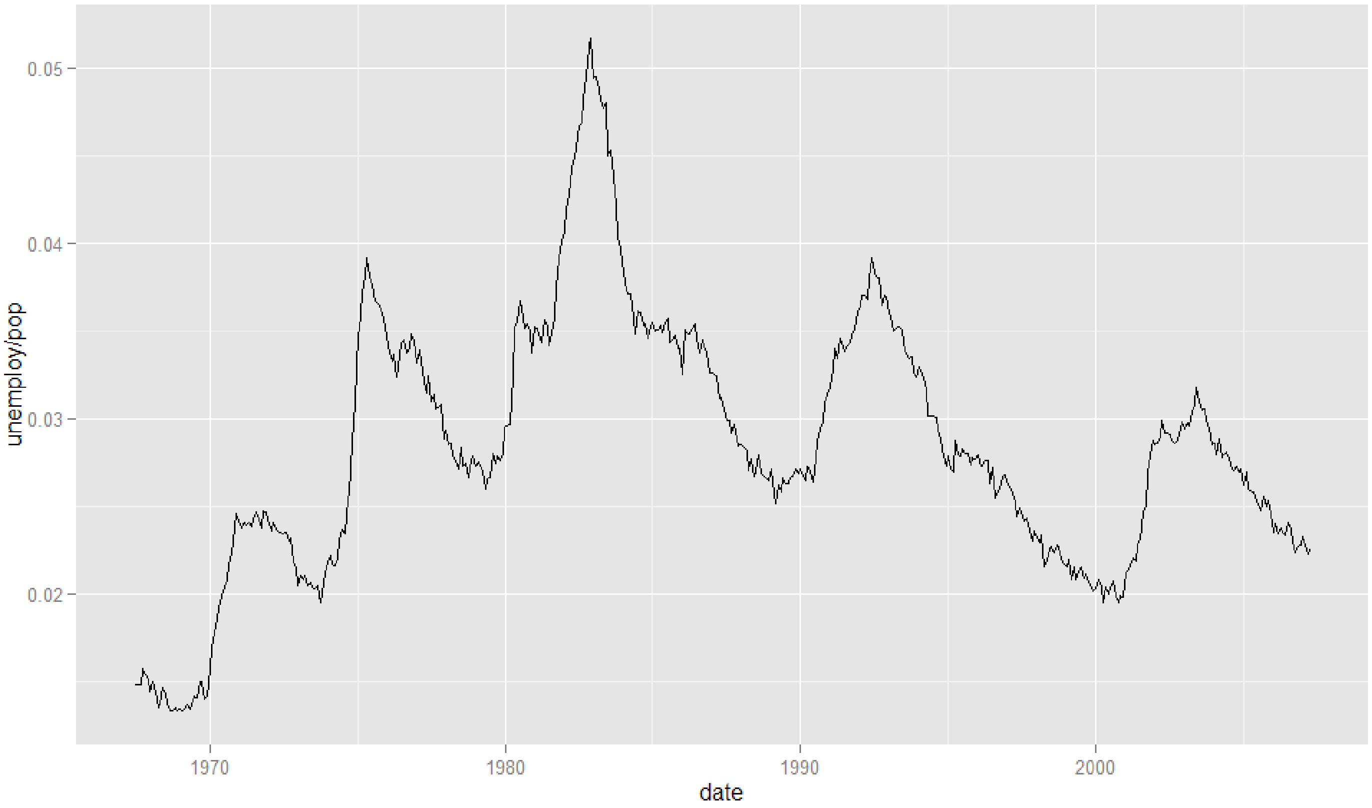
economics数据集



```
> economics
         date   pce    pop psavert uempmed unemploy
1  1967-06-30 507.8 198712     9.8     4.5     2944
2  1967-07-31 510.9 198911     9.8     4.7     2945
3  1967-08-31 516.7 199113     9.0     4.6     2958
4  1967-09-30 513.3 199311     9.8     4.9     3143
5  1967-10-31 518.5 199498     9.7     4.7     3066
6  1967-11-30 526.2 199657     9.4     4.8     3018
7  1967-12-31 532.0 199808     9.0     5.1     2878
8  1968-01-31 534.7 199920     9.5     4.5     3001
9  1968-02-29 545.4 200056     8.9     4.1     2877
10 1968-03-31 545.1 200208     9.6     4.6     2709
11 1968-04-30 550.9 200361     9.3     4.4     2740
12 1968-05-31 557.4 200536     8.9     4.4     2938
13 1968-06-30 564.4 200706     7.8     4.5     2883
14 1968-07-31 568.2 200898     7.6     4.2     2768
15 1968-08-31 569.5 201095     7.6     4.6     2686
16 1968-09-30 572.9 201290     7.8     4.8     2689
17 1968-10-31 578.0 201466     7.6     4.4     2715
18 1968-11-30 577.9 201621     8.1     4.4     2685
19 1968-12-31 584.9 201760     7.1     4.4     2718
```

用连线图表现时间序列

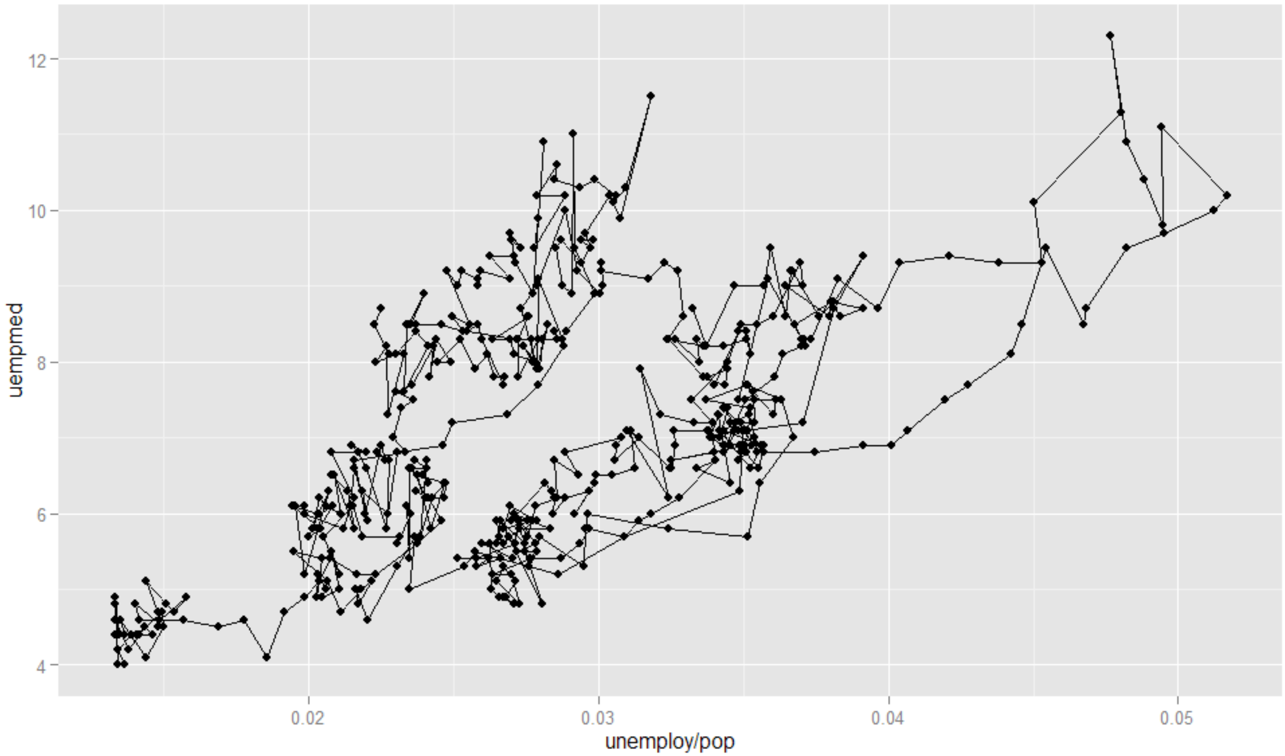
```
qplot(date, unemploy / pop, data = economics, geom = "line")
```



路径表达方式

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

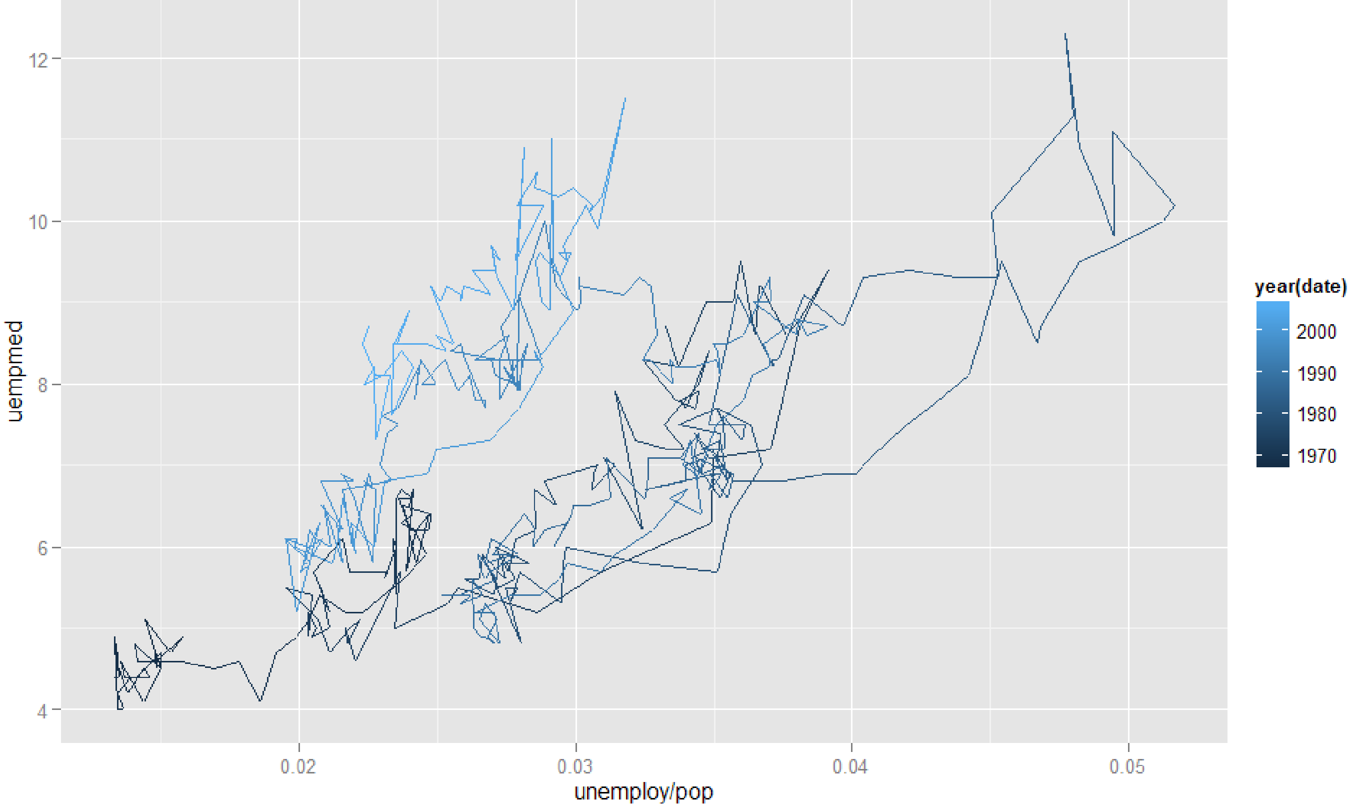
qplot(unemploy / pop, uempmed, data = economics,geom = c("point", "path"))



2013.2.8

彩色路径

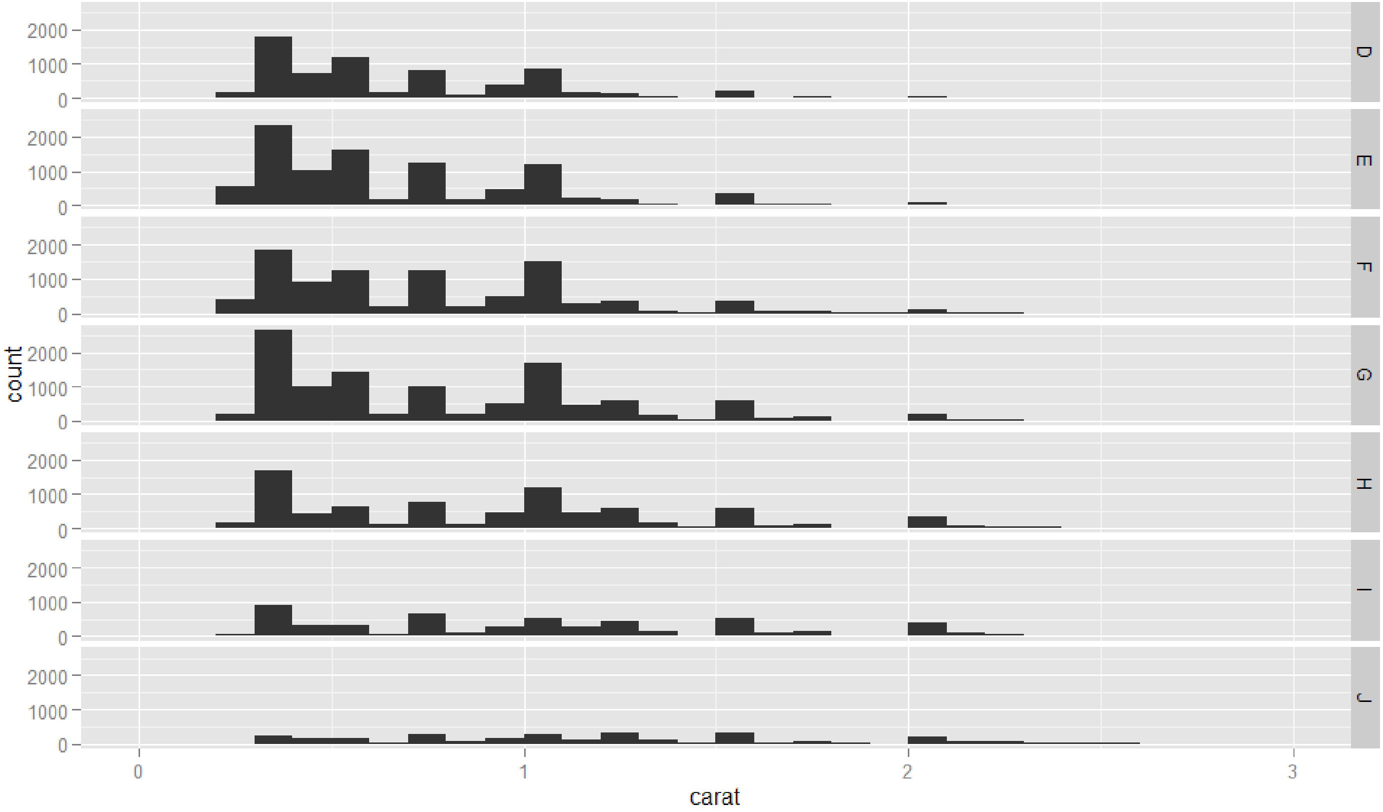
qplot(unemploy / pop, uempmed, data = economics,geom = "path", colour = year(date)) + scale_area()



2013.2.8

Facet

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



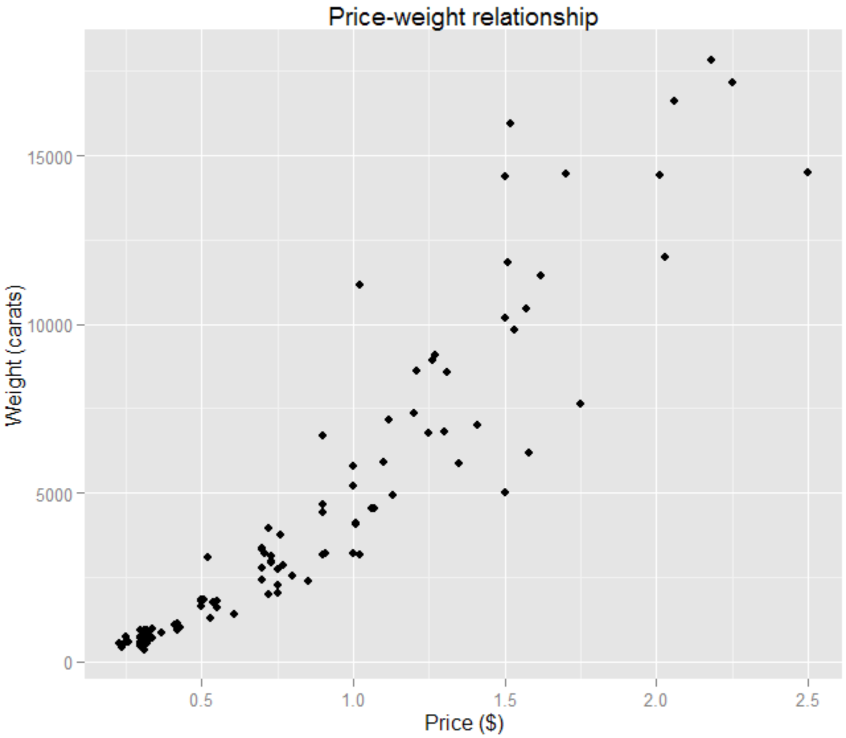
2013.2.8

其它选项

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

xlab, ylab

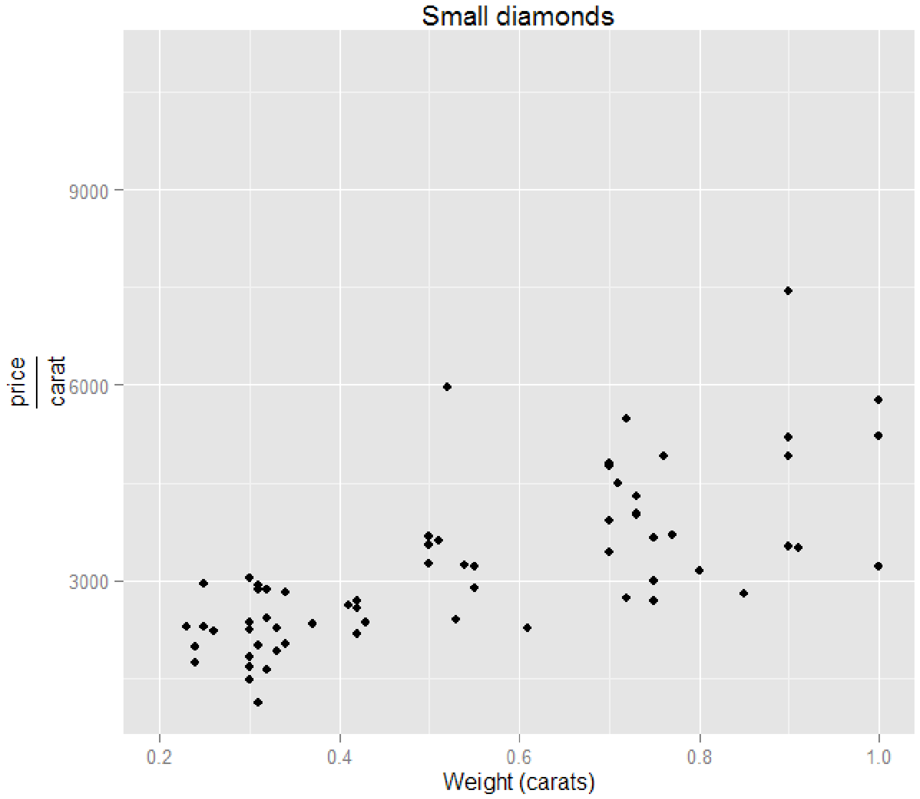
qplot(carat, price, data = dsmall,xlab = "Price ($)", ylab = "Weight (carats)",main = "Price-weight relationship")



2013.2.8

xlim, ylim

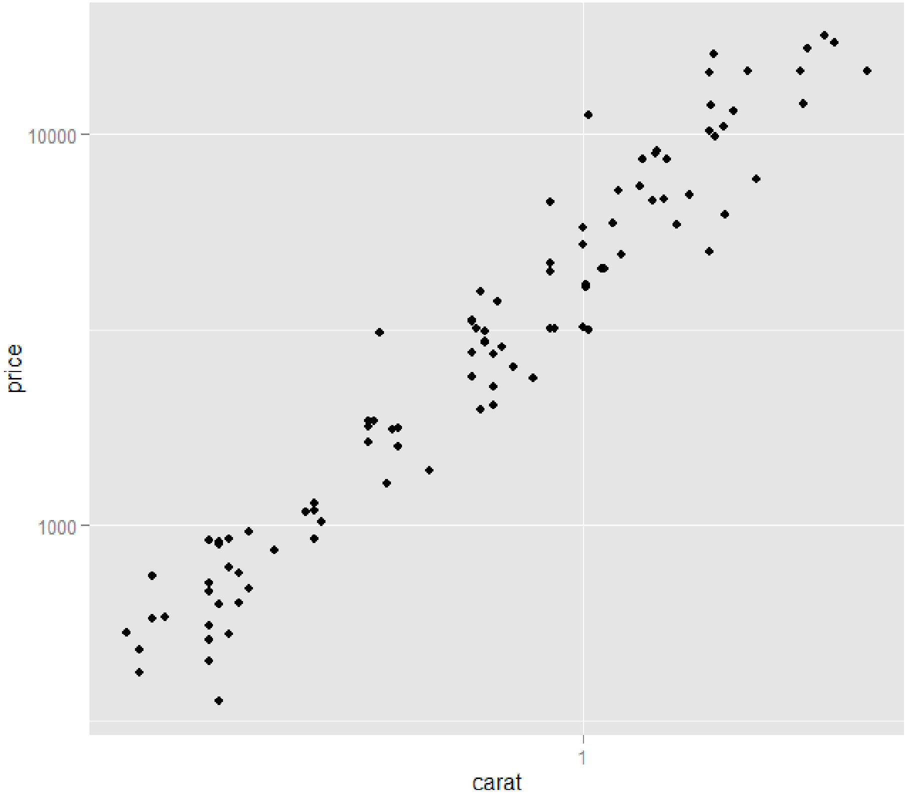
```
qplot(carat, price/carat, data = dsmall,ylab = expression(frac(price,carat)),xlab =
    "Weight (carats)",main="Small diamonds",xlim = c(.2,1))
```



2013.2.8

log

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



qplot vs plot

- 不是generic，因此不能像plot那样可以处理很多不同的R对象（ggplot才可以）
- I( )的使用
- 推荐使用ggplot2的装饰属性名
- 使用图层追加图形元素

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