30分钟学会ggplot2

肖凯

xccds1977@gmail.com

太极剑法和ggplot2

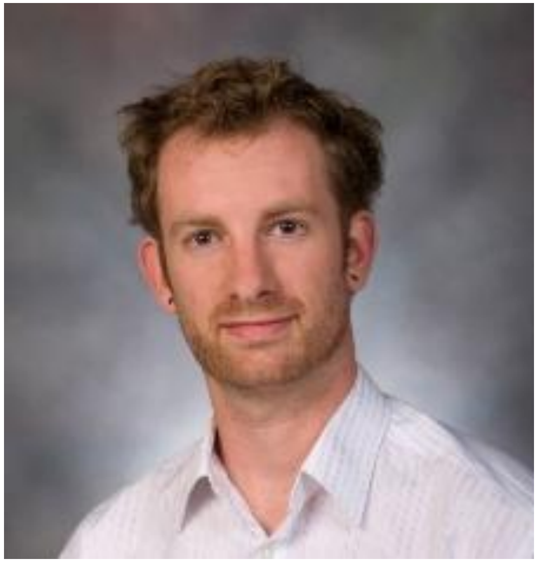
- 招无定式
- 潜力无穷
- 需要忘记
- 容易学习

内容提要：

- 简介
- 基本概念
- 简单示例
- 进阶示例
- 学习资源

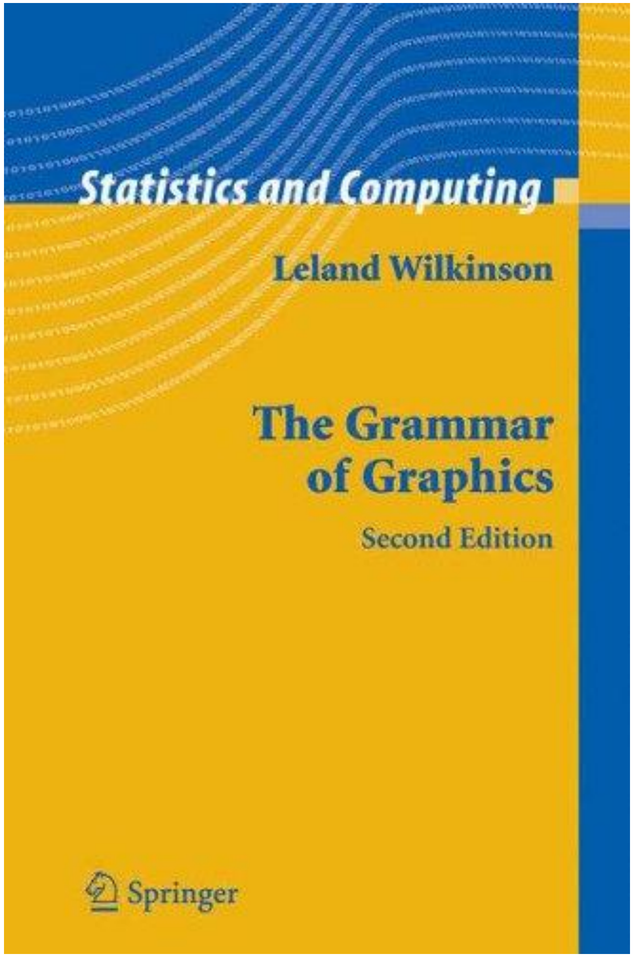
ggplot2简介

- 由Hadley Wickham于2005年创建
- 于2012年四月进行了重大更新，最新版本0.91
- 作者目前的工作是重写代码，简化语法，方便用户开发和使用



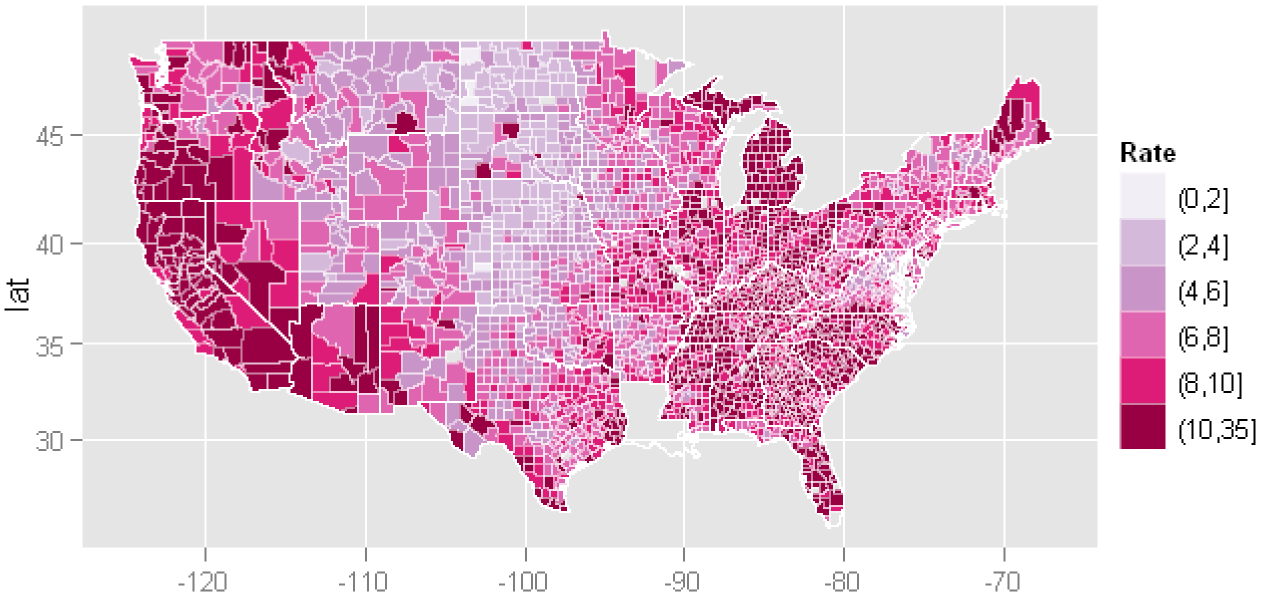
ggplot2简介

- ggplot2 is a plotting system for R
- based on the《The Grammar of Graphics》
- which tries to take the good parts of base and lattice graphics and none of the bad parts
- It takes care of many of the fiddly details that make plotting a hassle
- It easy to produce complex multi-layered graphics



为什么要使用ggplot2

- 用户能在更抽象层面上控制图形，使创造性绘图更容易；
- 采用图层的设计方式，有利于结构化思维；
- 图形美观，同时避免繁琐细节。

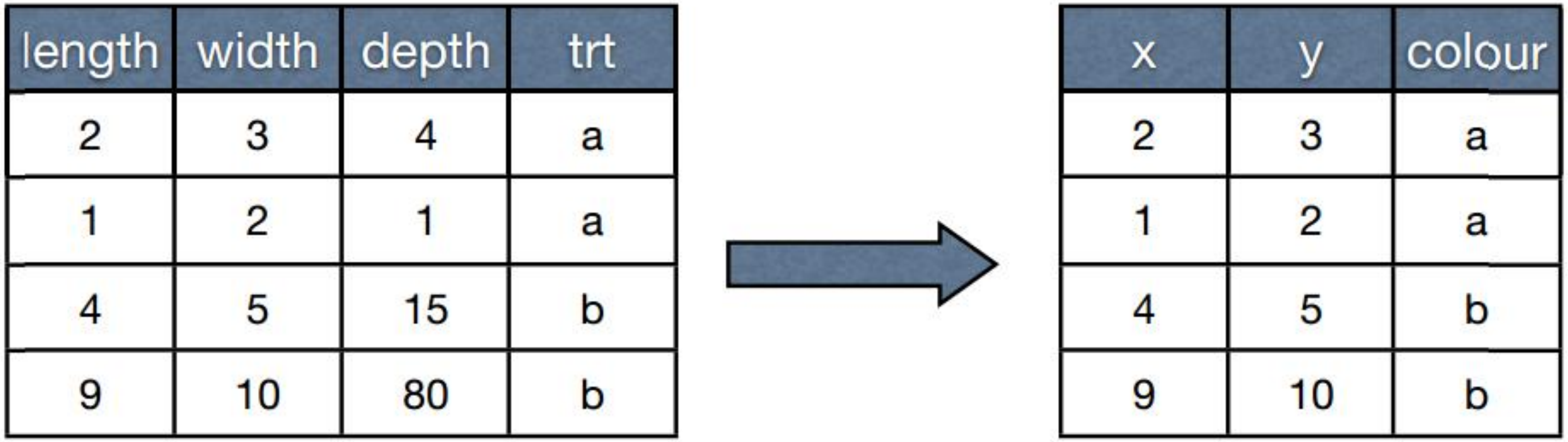


ggplot2的基本概念

- 数据（Data）和映射（Mapping）
- 标度（Scale）
- 几何对象（Geometric）
- 统计变换（Statistics）
- 坐标系统（Coordinate）
- 图层（Layer）
- 分面（Facet）

数据（Data）和映射（Mapping）

将数据中的变量映射到图形属性。映射控制了二者之间的关系。

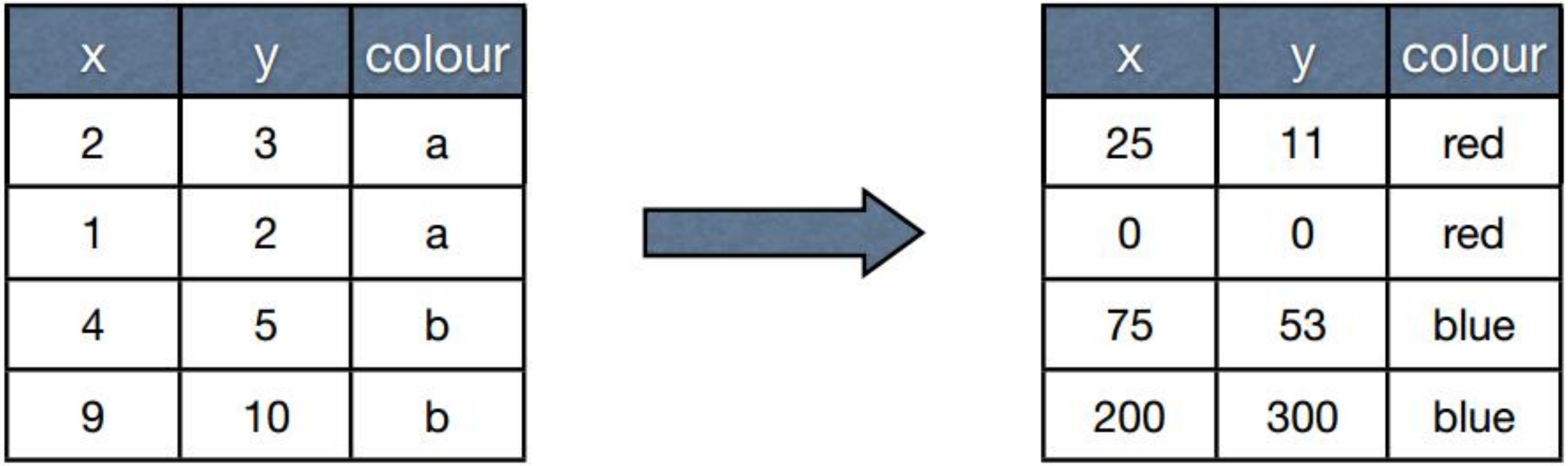


length	width	depth	trt
2	3	4	a
1	2	1	a
4	5	15	b
9	10	80	b

x	y	colour
2	3	a
1	2	a
4	5	b
9	10	b

标度（Scale）

标度负责控制映射后图形属性的显示方式。具体形式上来看是图例和坐标刻度。Scale和Mapping是紧密相关的概念。

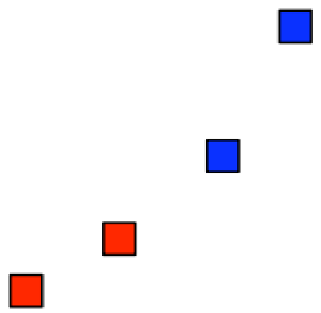


x	y	colour
2	3	a
1	2	a
4	5	b
9	10	b

x	y	colour
25	11	red
0	0	red
75	53	blue
200	300	blue

几何对象（Geometric）

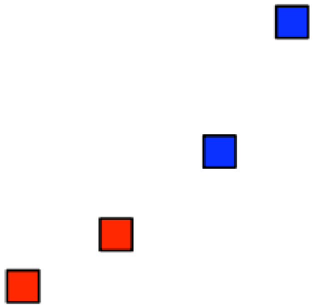
几何对象代表我们在图中实际看到的图形元素，如点、线、多边形等。



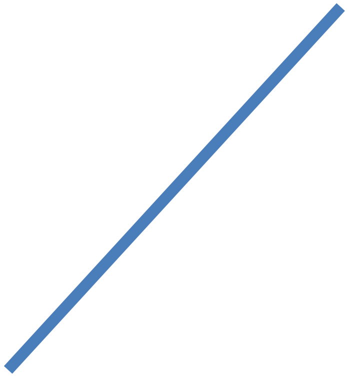
Geoms

统计变换（statistics）

对原始数据进行某种计算，例如对二元散点图加上一条回归线。



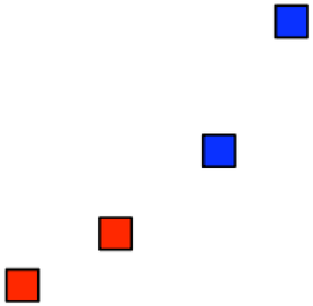
Geoms



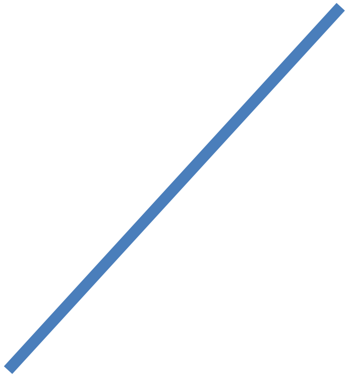
Stat

坐标系统（Coordinate）

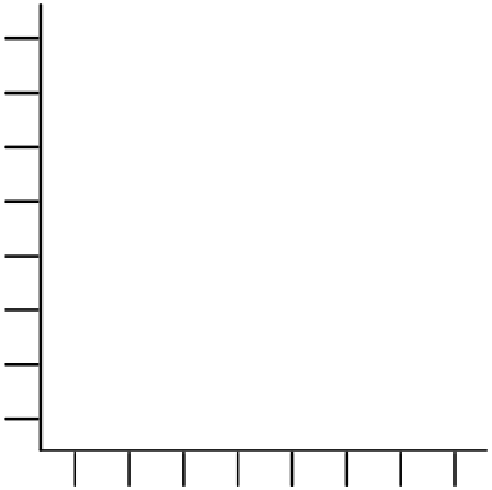
坐标系统控制坐标轴并影响所有图形元素，坐标轴可以进行变换以满足不同的需要。



Geoms



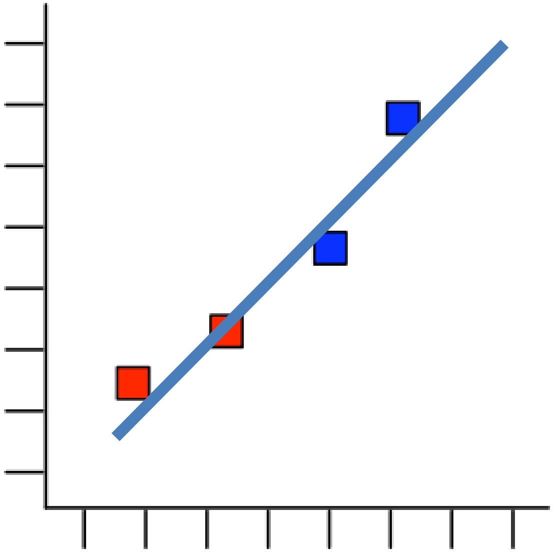
Stat



Coord

图层（Layer）

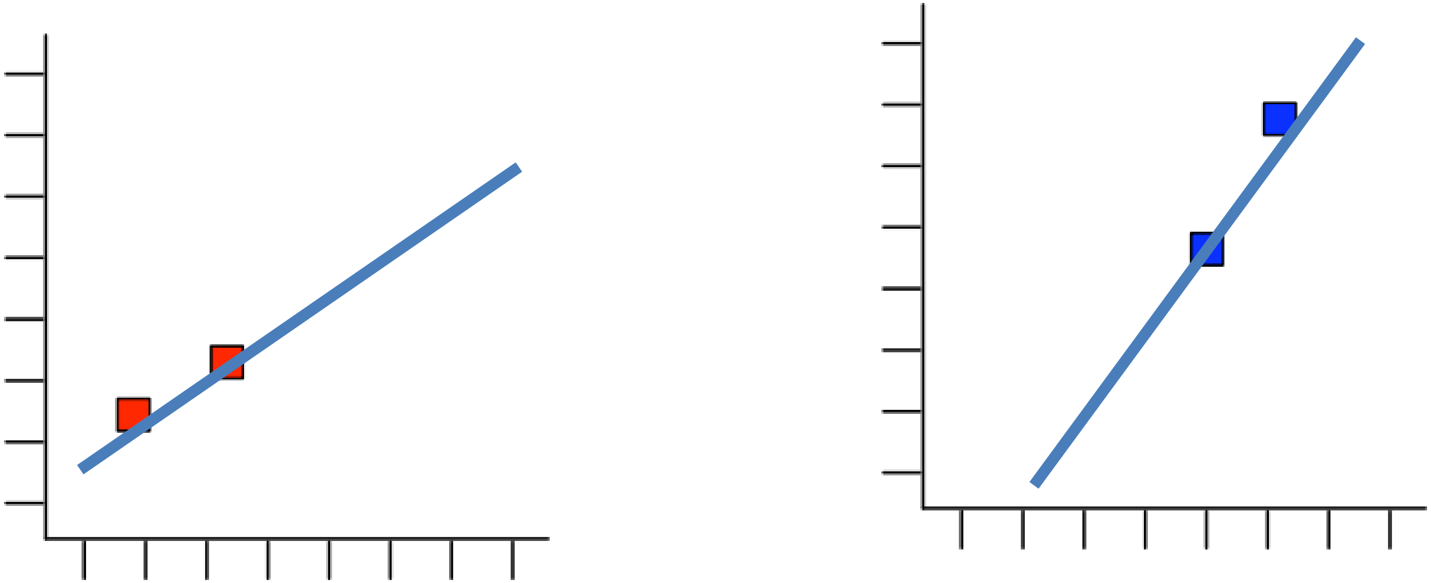
数据、映射、几何对象、统计变换等构成一个图层。图层可以允许用户一步步的构建图形，方便单独对图层进行修改。



分面（Facet）

条件绘图，将数据按某种方式分组，然后分别绘图。

分面就是控制分组绘图的方法和排列形式。



ggplot2的基本概念

- 数据（Data）和映射（Mapping）
- 标度（Scale）
- 几何对象（Geometric）
- 统计变换（Statistics）
- 坐标系统（Coordinate）
- 图层（Layer）
- 分面（Facet）

简单示例

- 散点图
- 直方图
- 条形图

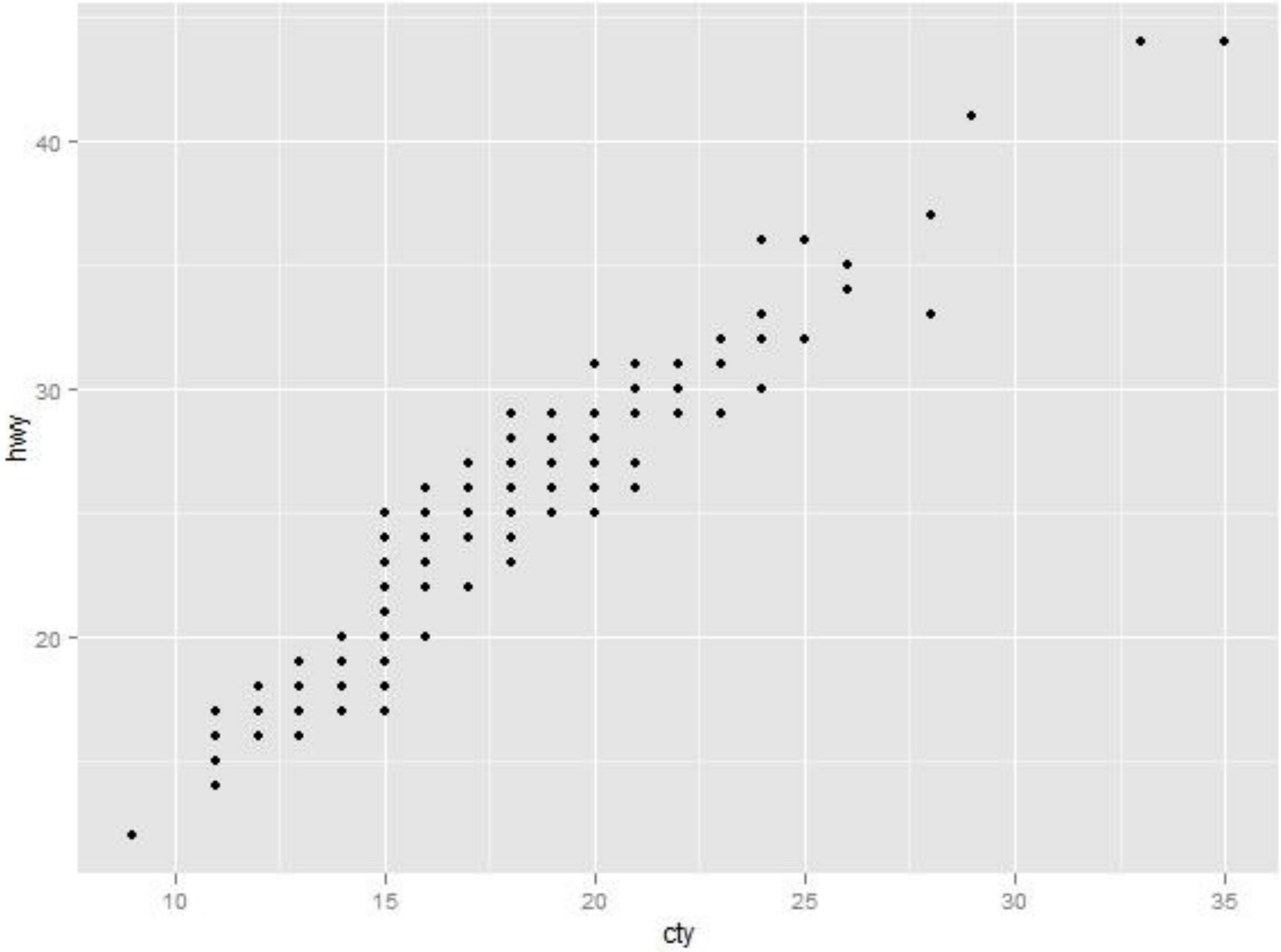
- 饼图
- 箱线图
- 二维直方图

示例数据

```
> str(mpg)
'data.frame':        234 obs. of  14 variables:
 $ manufacturer: Factor w/ 15 levels "audi","chevrolet",..:
 $ model       : Factor w/ 38 levels "4runner 4wd",..:
 $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
 $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999
 $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
 $ trans       : Factor w/ 10 levels "auto(av)","auto(l3)",..:
 $ drv         : Factor w/ 3 levels "4","f","r":
 $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
 $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
 $ fl          : Factor w/ 5 levels "c","d","e","p",..:
 $ class       : Factor w/ 7 levels "2seater","compact",..:
```

```
> library(ggplot2)
> p <- ggplot(data=mpg, mapping=aes(x=cty, y=hwy))
> p + geom_point()
```

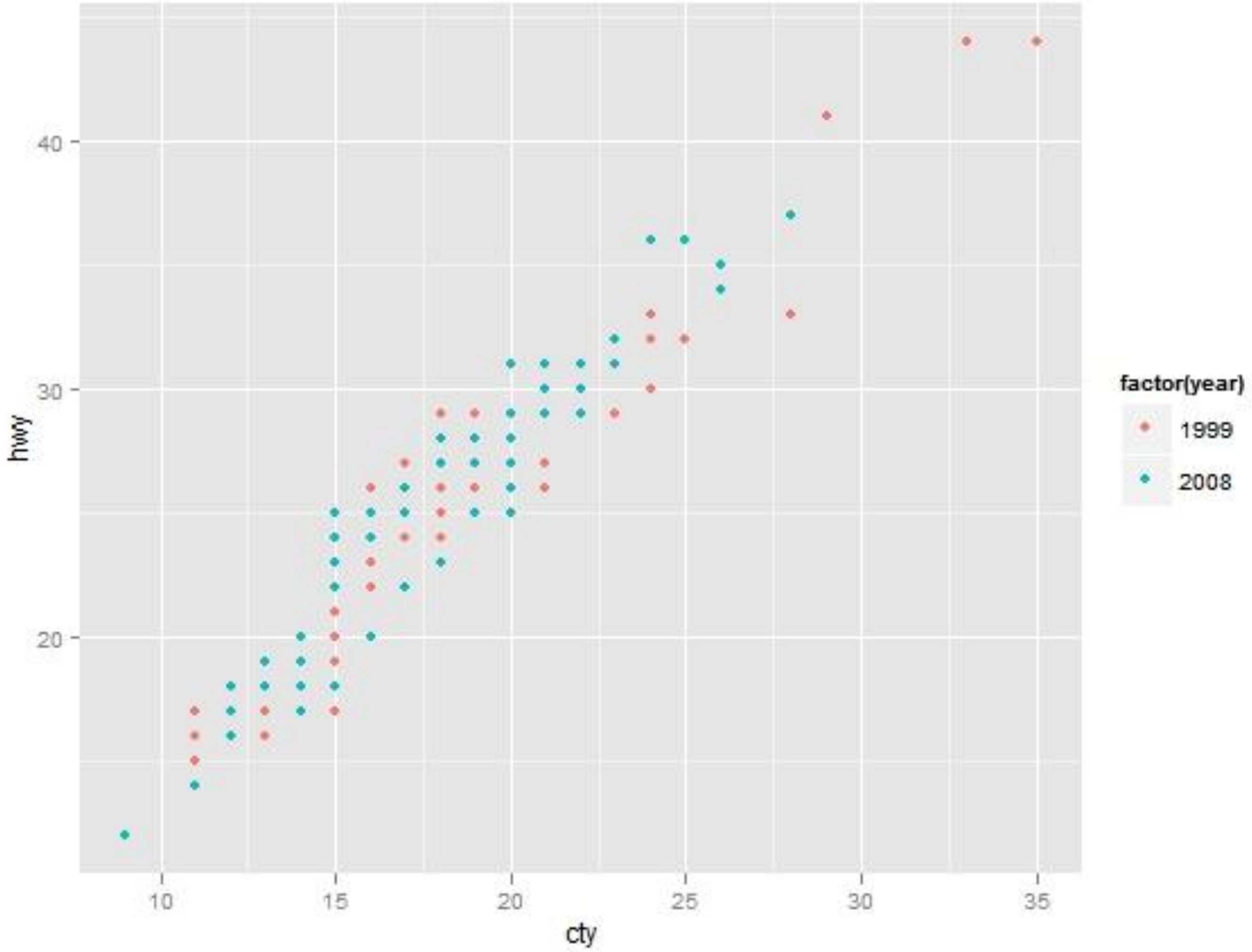
aesthetics



```
> summary(p)
data: manufacturer, model, displ, year, cyl, trans, drv, cty, hwy,
fl, class [234x11]
mapping: x = cty, y = hwy
faceting: facet_null()

> summary(p+geom_point())
data: manufacturer, model, displ, year, cyl, trans, drv, cty, hwy,
  fl, class [234x11]
mapping:  x = cty, y = hwy
faceting: facet_null()
-----------------------------------
geom_point: na.rm = FALSE
stat_identity:
position_identity: (width = NULL, height = NULL)
```

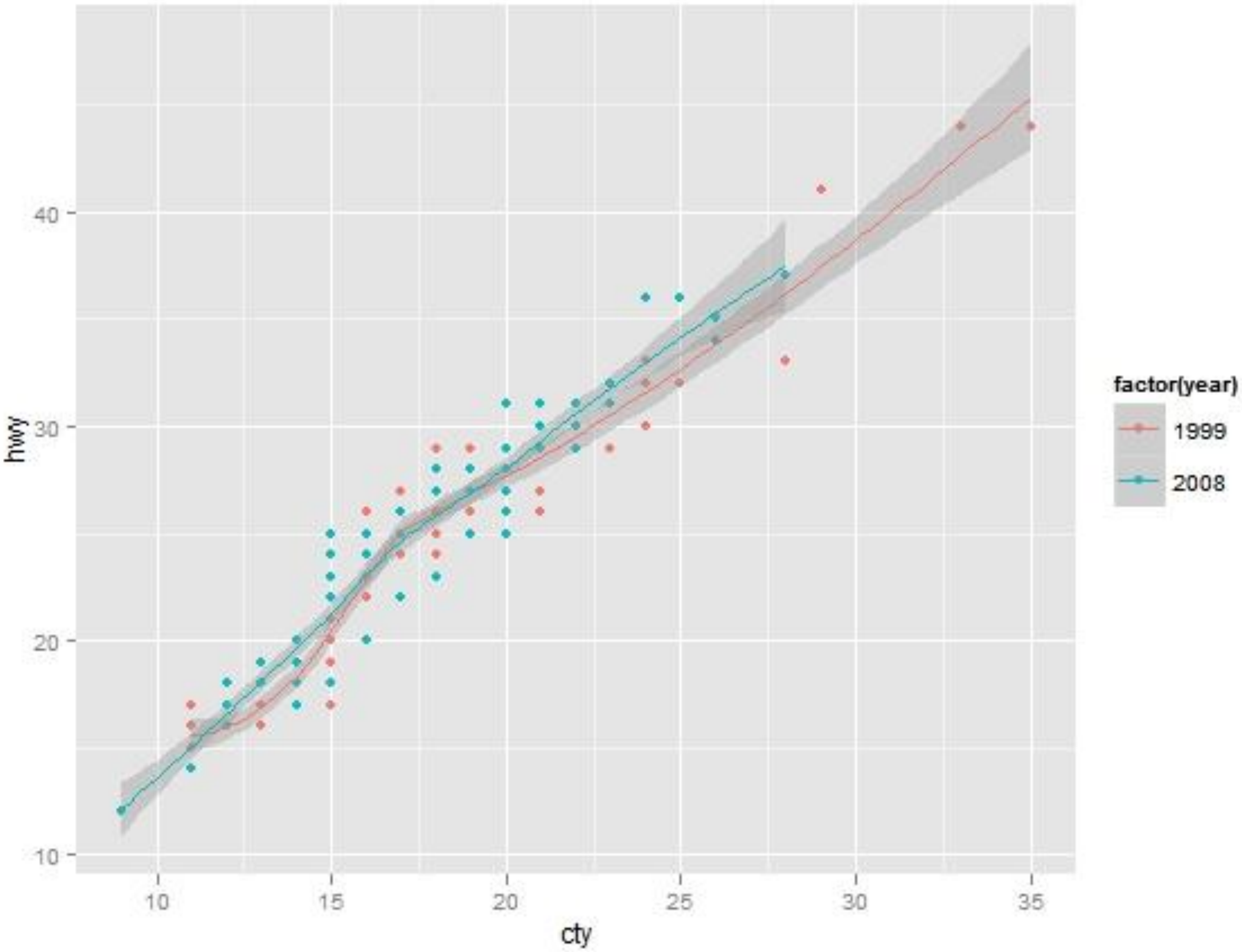
将年份映射到颜色属性

```
> p <- ggplot(mpg,
          aes(x=cty, y=hwy, colour=factor(year)))
> p + geom_point()
```

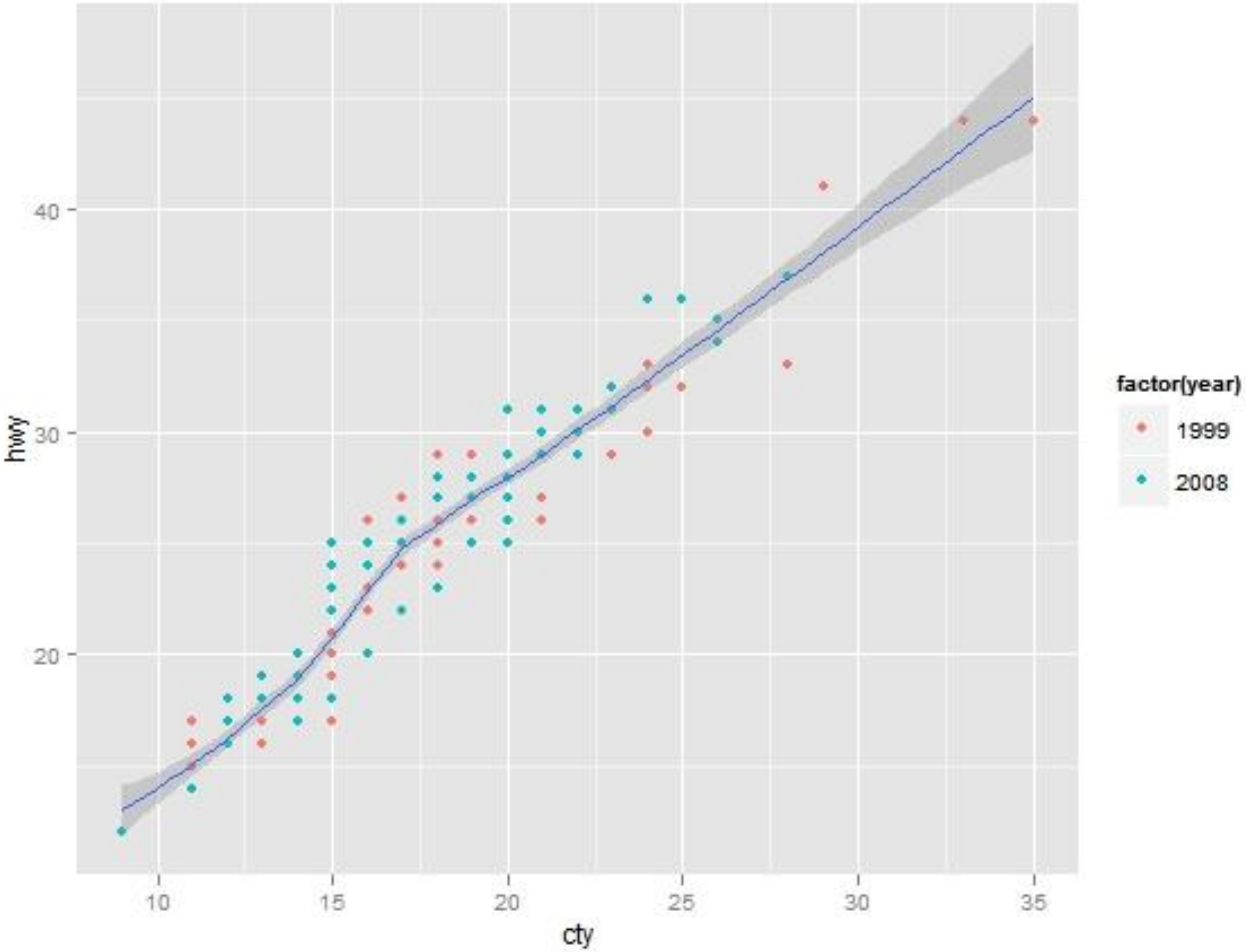


增加平滑曲线

```
> p + geom_point() + stat_smooth()
```



```
> p <- ggplot(mpg, aes(x=cty,y=hwy))
  p + geom_point(aes(colour=factor(year)))+
      stat_smooth()
```



两种等价的绘图方式

```
 > p <- ggplot(mpg, aes(x=cty,y=hwy))
   p + geom_point(aes(colour=factor(year)))+
       stat_smooth()
```

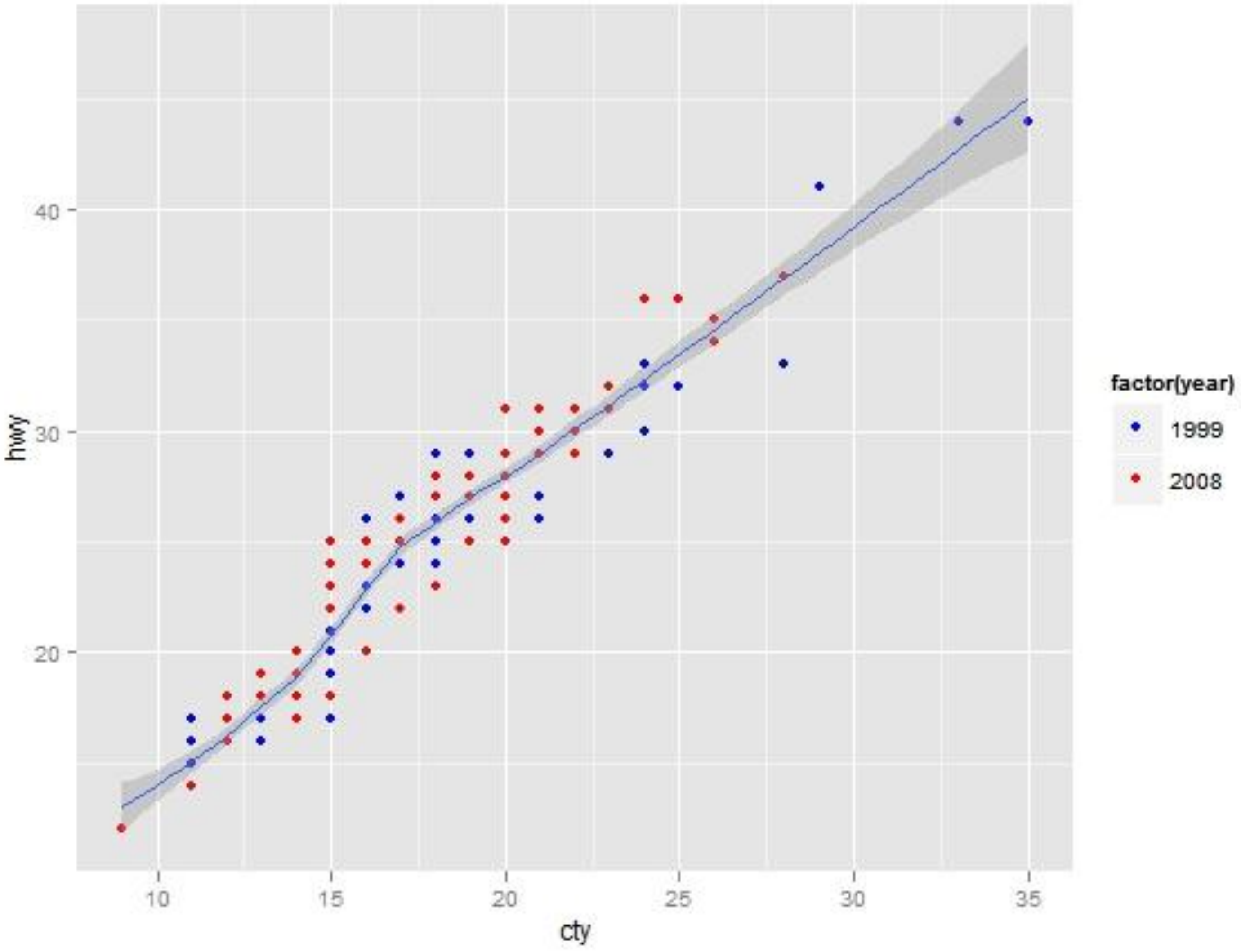
```
>  d <- ggplot() +
        geom_point(data=mpg, aes(x=cty, y=hwy, colour=factor(year)))+
        stat_smooth(data=mpg, aes(x=cty, y=hwy))
>  print(d)
```

此时除了底层画布外，有两个图层，分别定义了geom和 stat。

```
> summary(d)
data: [0x0]
faceting: facet_null()
-----------------------------------
mapping: x = cty, y = hwy, colour = factor(year)
geom_point: na.rm = FALSE
stat_identity:
position_identity: (width = NULL, height = NULL)

mapping: x = cty, y = hwy
geom_smooth:
stat_smooth: method = auto, formula = y ~ x, se = TRUE,
n = 80, fullrange = FALSE, level = 0.95, na.rm = FALSE
position_identity: (width = NULL, height = NULL)
```

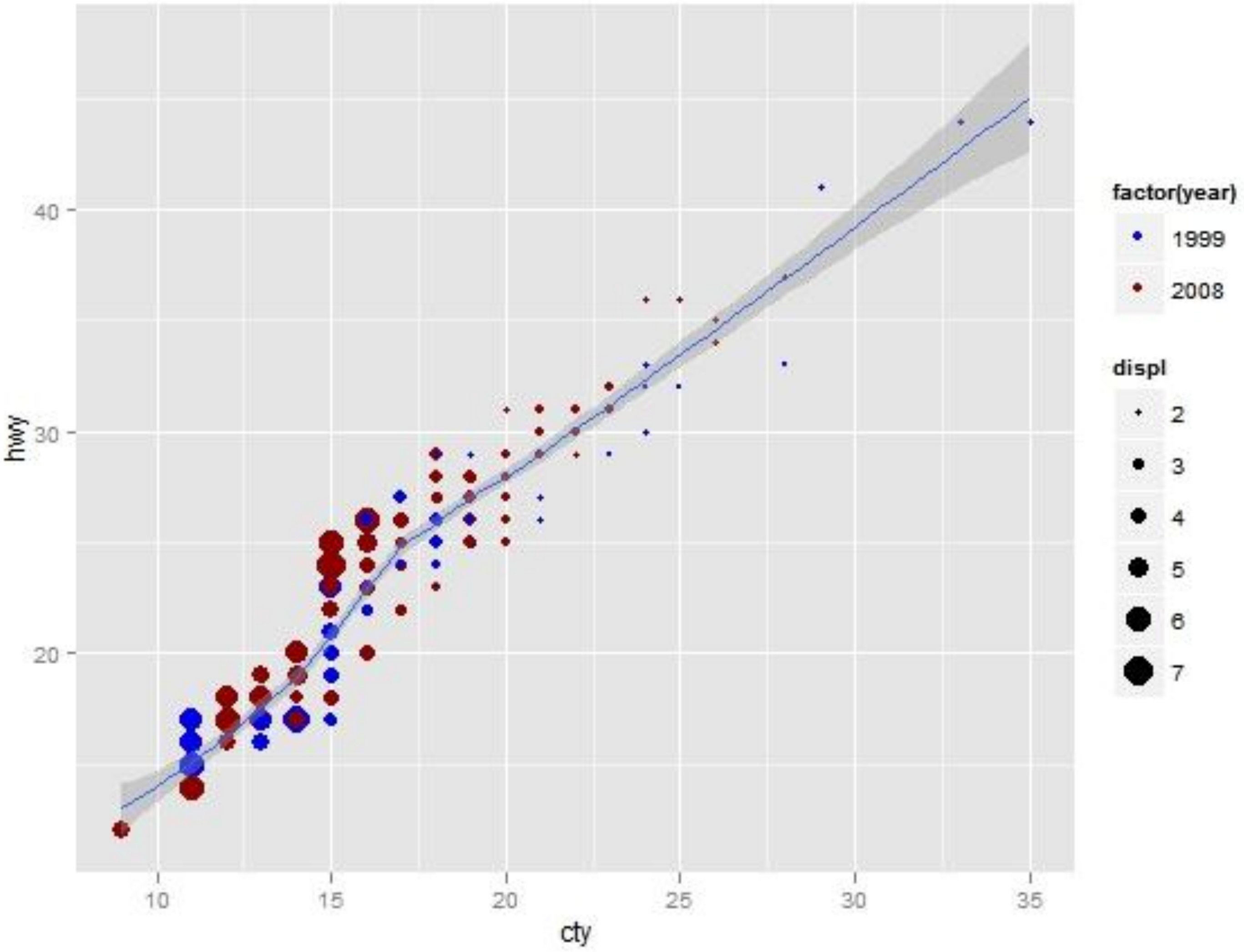
用标度来修改颜色取值

```
> p + geom_point(aes(colour=factor(year)))+
        stat_smooth()+
        scale_color_manual(values =c('blue','red'))
```

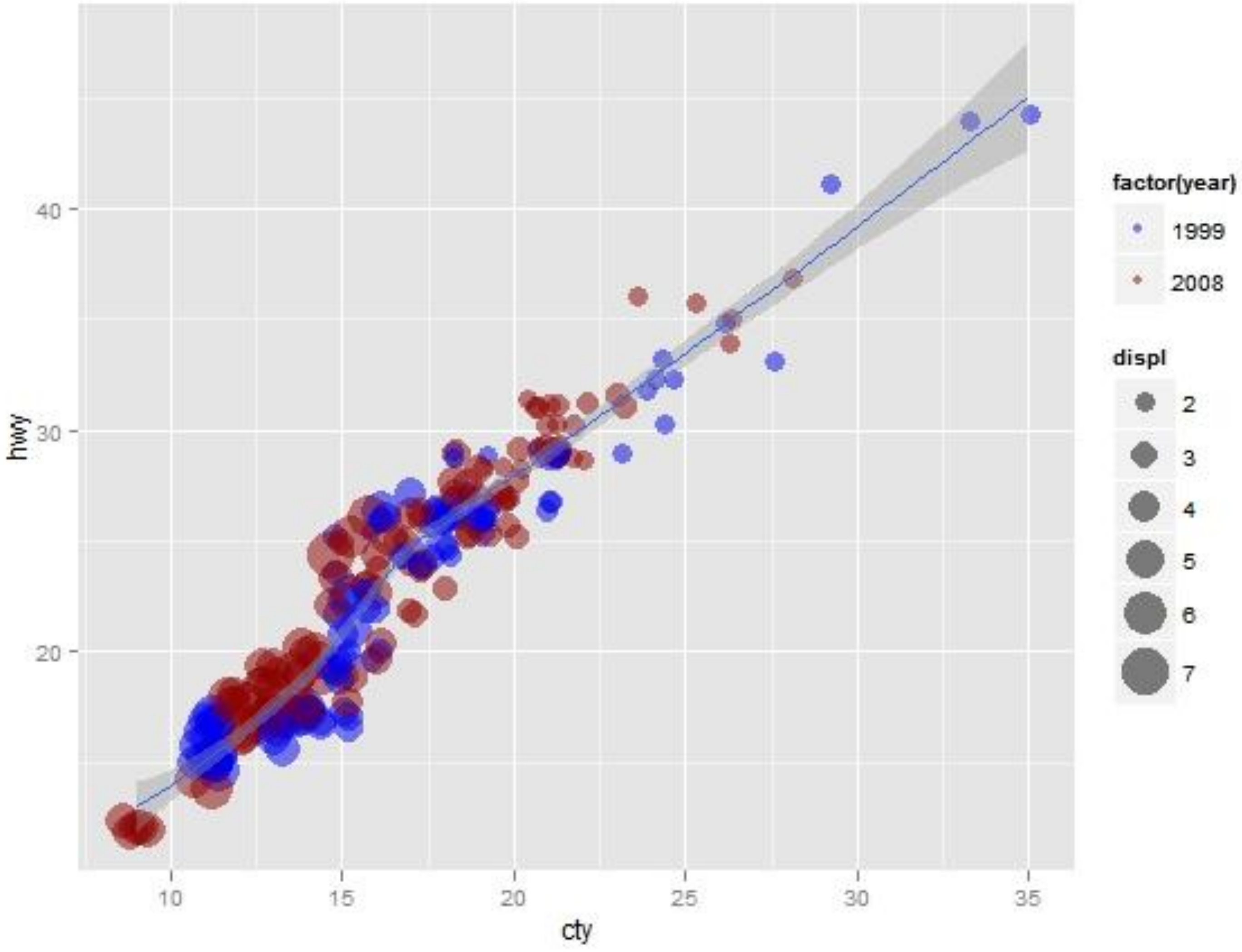


将排量映射到散点大小

```
> p + geom_point(aes(colour=factor(year),size=displ))+
      stat_smooth()+
      scale_color_manual(values =c('blue2','red4'))
```

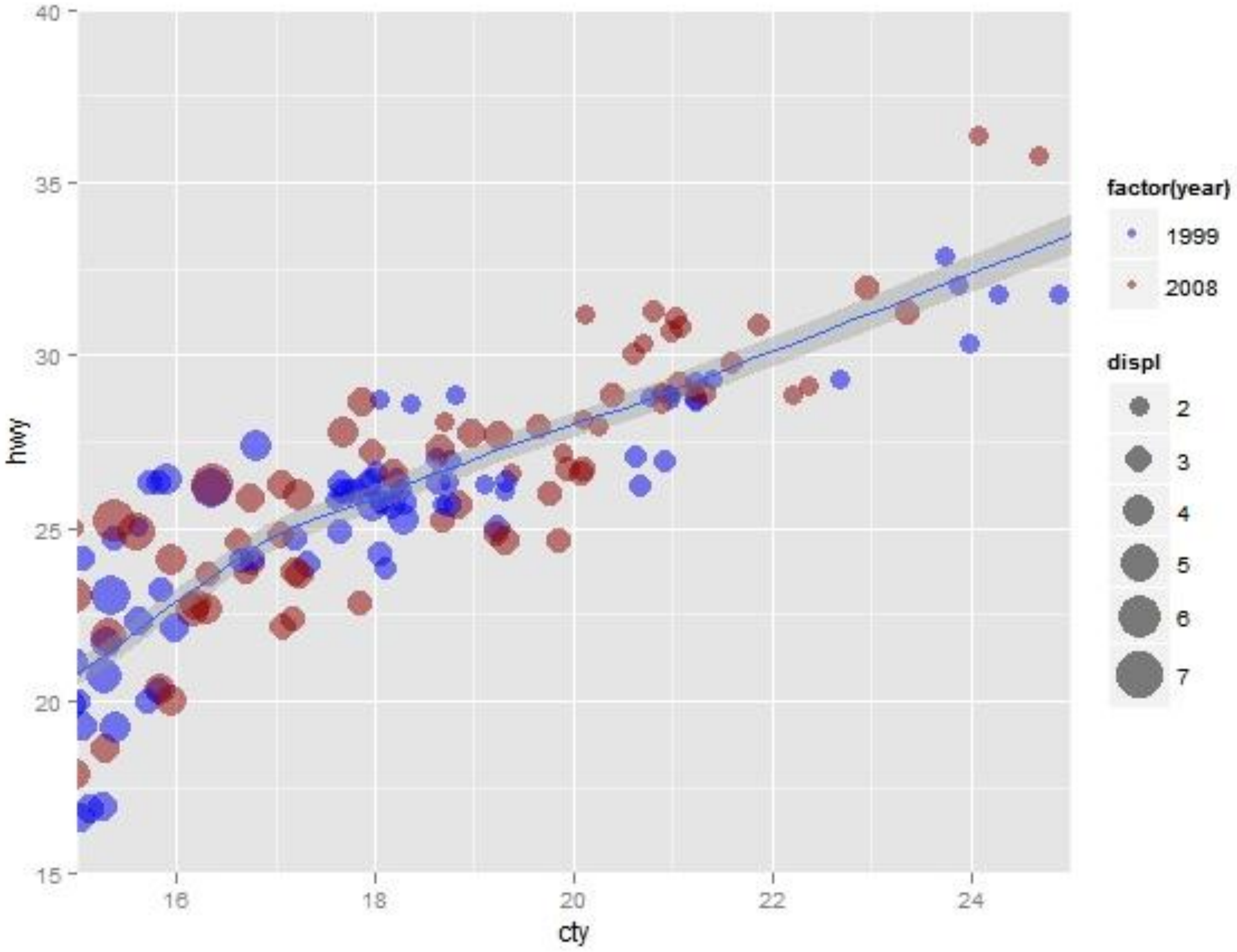


```
> p + geom_point(aes(colour=factor(year),size=displ),
          alpha=0.5,position = "jitter") +  stat_smooth()+
      scale_color_manual(values =c('blue2','red4'))+
      scale_size_continuous(range = c(4, 10))
```



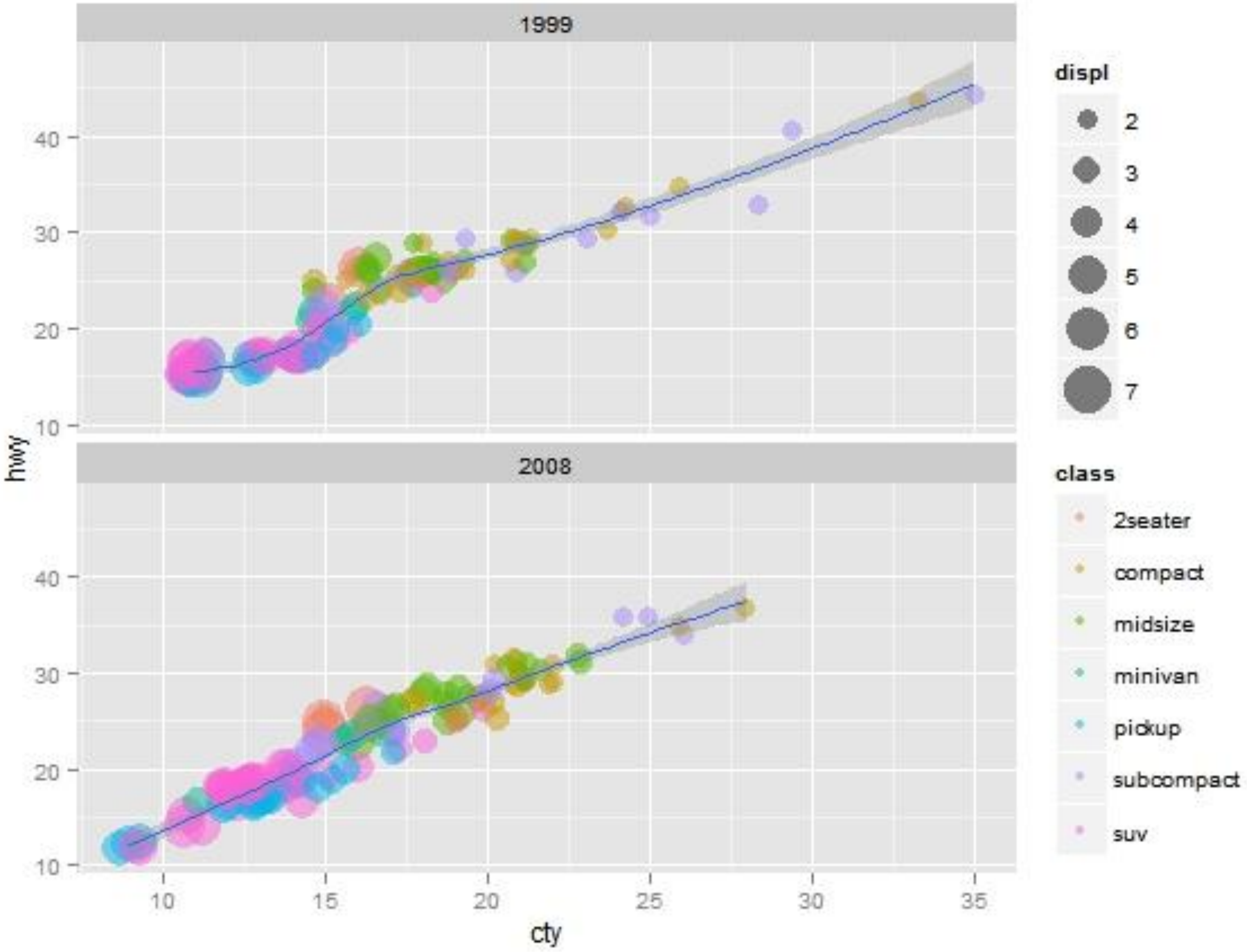
```
>p + geom_point(aes(colour=factor(year),size=displ),
        alpha=0.5,position = "jitter")+    stat_smooth()+
    scale_color_manual(values =c('blue2','red4'))+
    scale_size_continuous(range = c(4, 10))+
    coord_cartesian(xlim = c(15, 25),ylim=c(15,40))
```

用坐标控制图形显示的范围



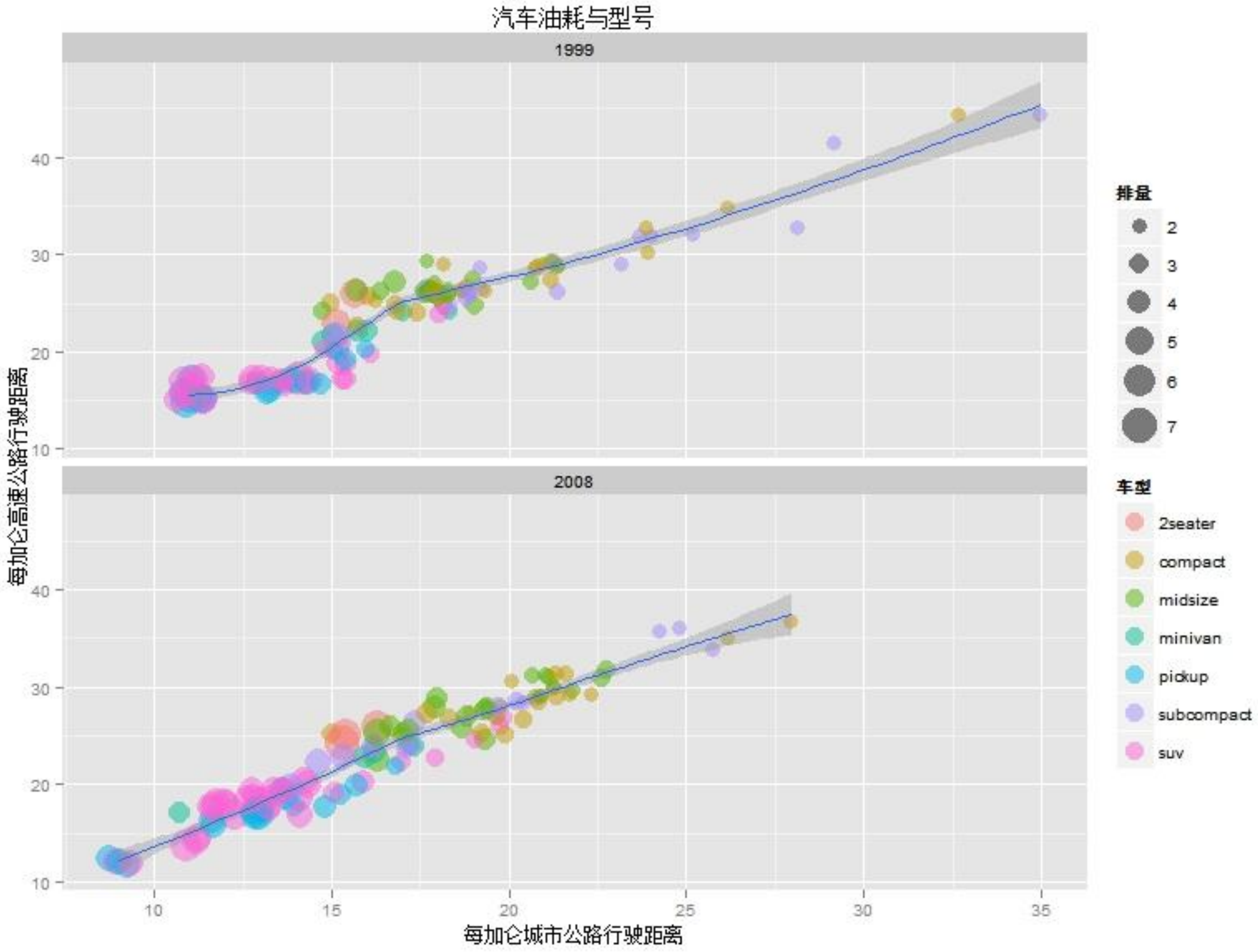
利用facet分别显示不同年份的数据

```
>p + geom_point(aes(colour=class, size=displ),
       alpha=0.5, position = "jitter")+   stat_smooth()+
    scale_size_continuous(range = c(4, 10))+
    facet_wrap(~ year, ncol=1)
```



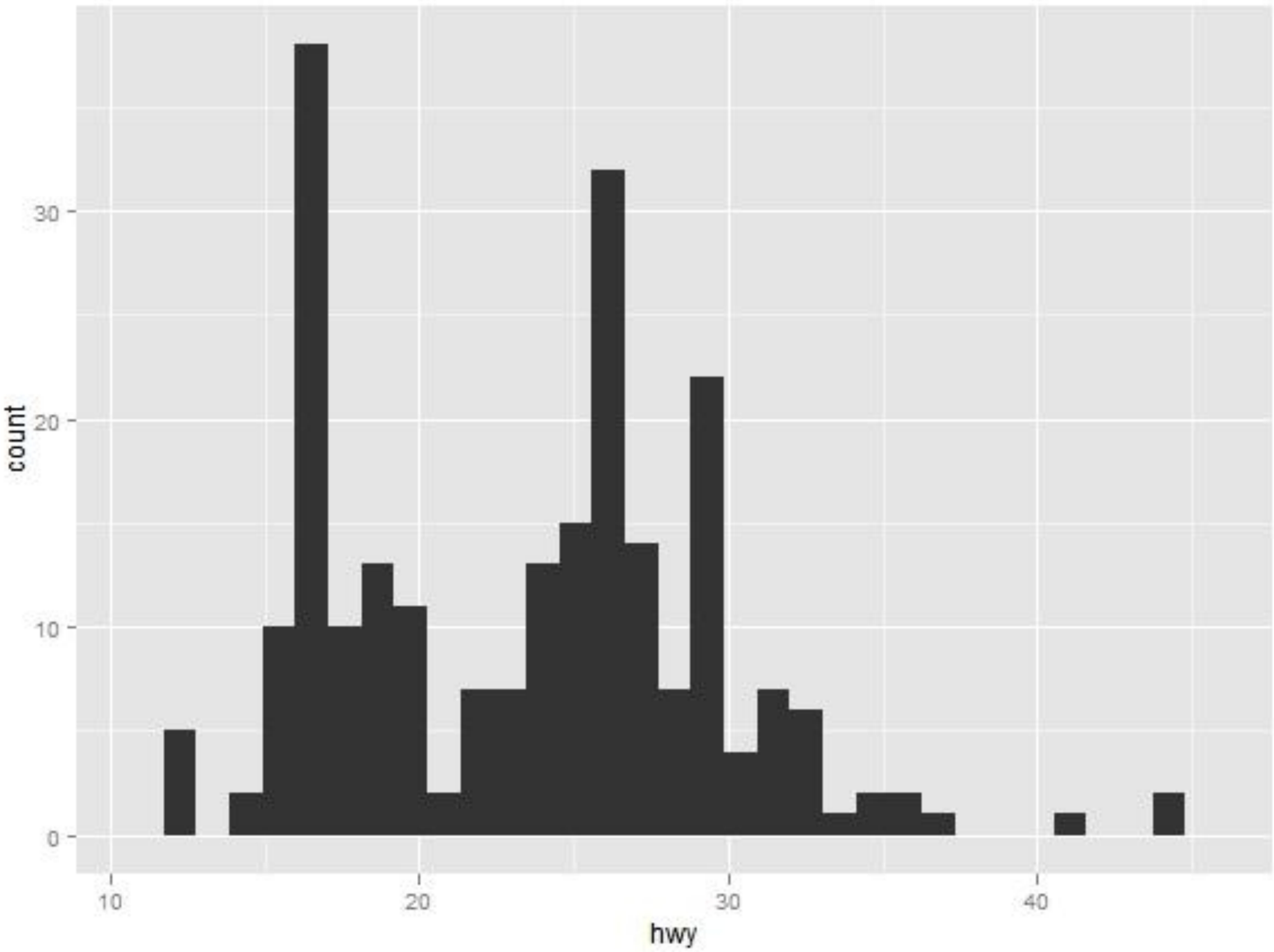
增加图名并精细修改图例

```
> p <- ggplot(mpg, aes(x=cty, y=hwy))
> p + geom_point(aes(colour=class,size=displ),
           alpha=0.5,position = "jitter")+
      stat_smooth()+
      scale_size_continuous(range = c(4, 10))+
      facet_wrap(~ year,ncol=1)+
      opts(title='汽车油耗与型号')+
      labs(y='每加仑高速公路行驶距离',
           x='每加仑城市公路行驶距离')+
      guides(size=guide_legend(title='排量'),
          colour = guide_legend(title='车型',
          override.aes=list(size=5)))
```


汽车油耗与型号
1999
2008
每加仑高速公路行驶距离
每加仑城市公路行驶距离
10
20
30
40
10
15
20
25
30
35
排量
2
3
4
5
6
7
车型
2seater
compact
midsize
minivan
pickup
subcompact
suv

直方图

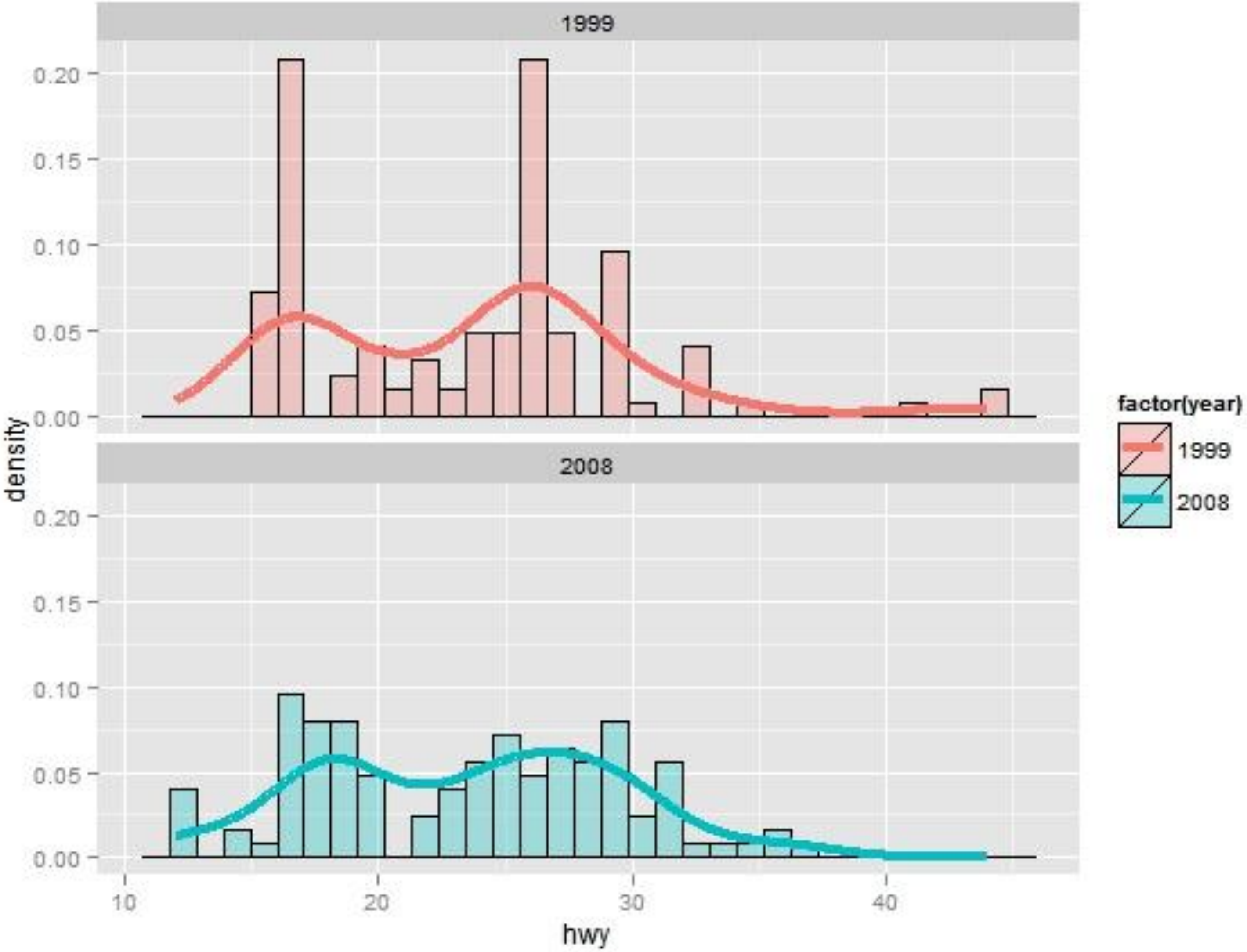
```
> P <- ggplot(mpg,aes(x=hwy))
  p + geom_histogram()
```



直方图的几何对象中内置有默认的统计变换

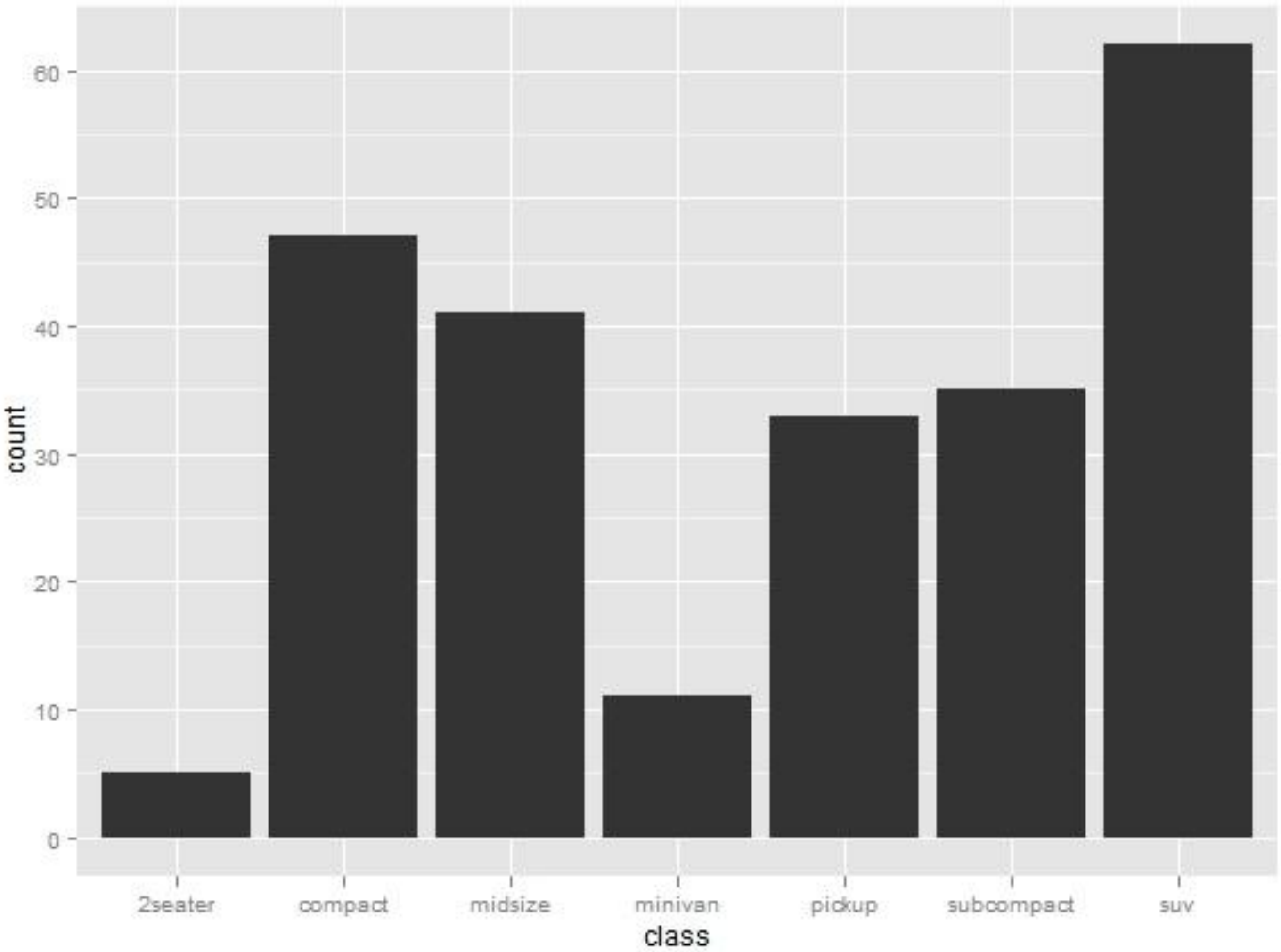
```
>  summary(p + geom_histogram())
 data: manufacturer, model, displ, year, cyl, trans,    drv,
cty, hwy, fl, class [234x11]
mapping: x = hwy
faceting: facet_null()
-----------------------------------
geom_histogram:
stat_bin:
position_stack: (width = NULL, height = NULL)
```

```
> p + geom_histogram(aes(fill=factor(year),y=..density..), alpha=0.3,colour='black')+
  stat_density(geom='line',position='identity',size=1.5, aes(colour=factor(year)))+
  facet_wrap(~year,ncol=1)
```



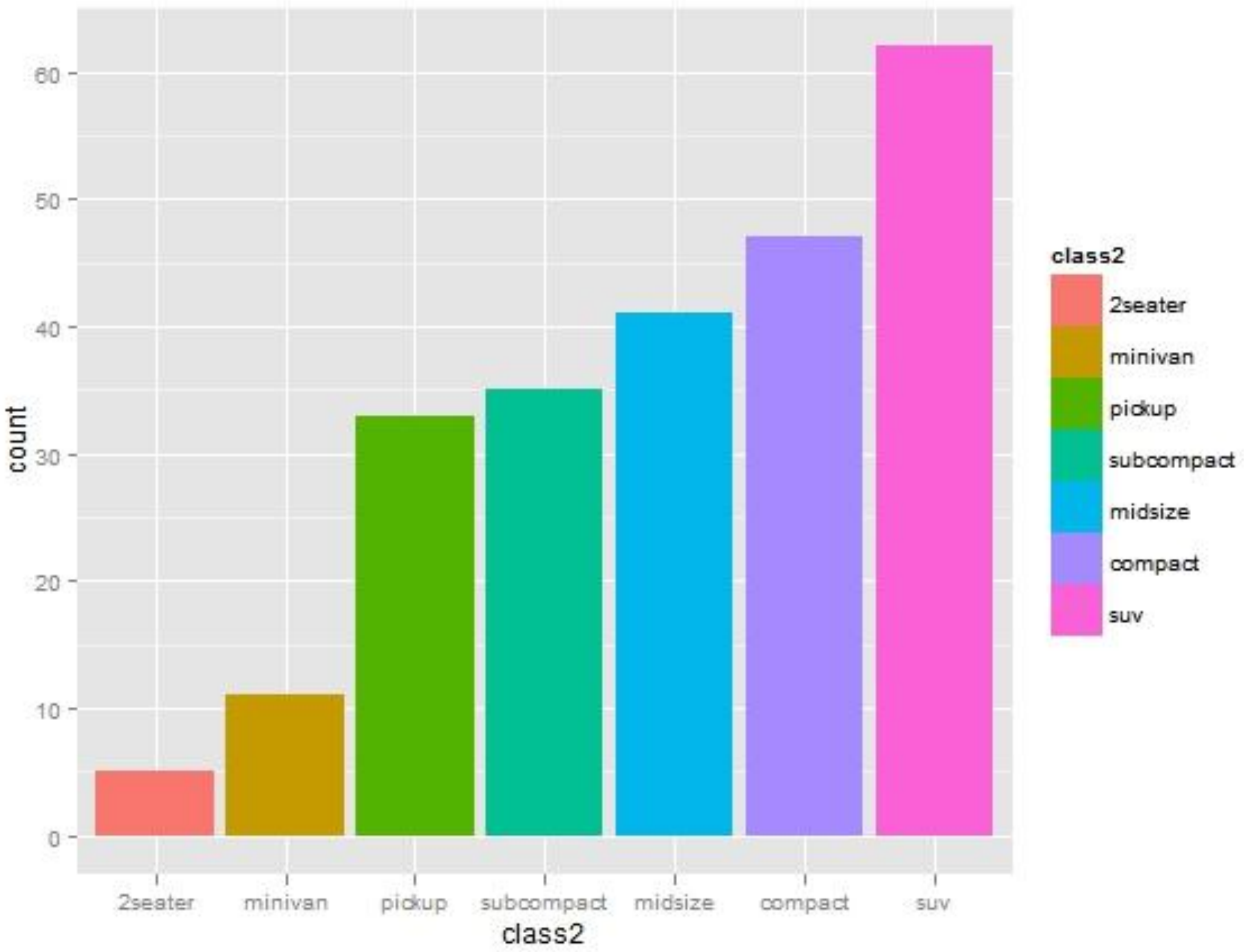
条形图

```
>  p <- ggplot(mpg, aes(x=class))
   p + geom_bar()
```



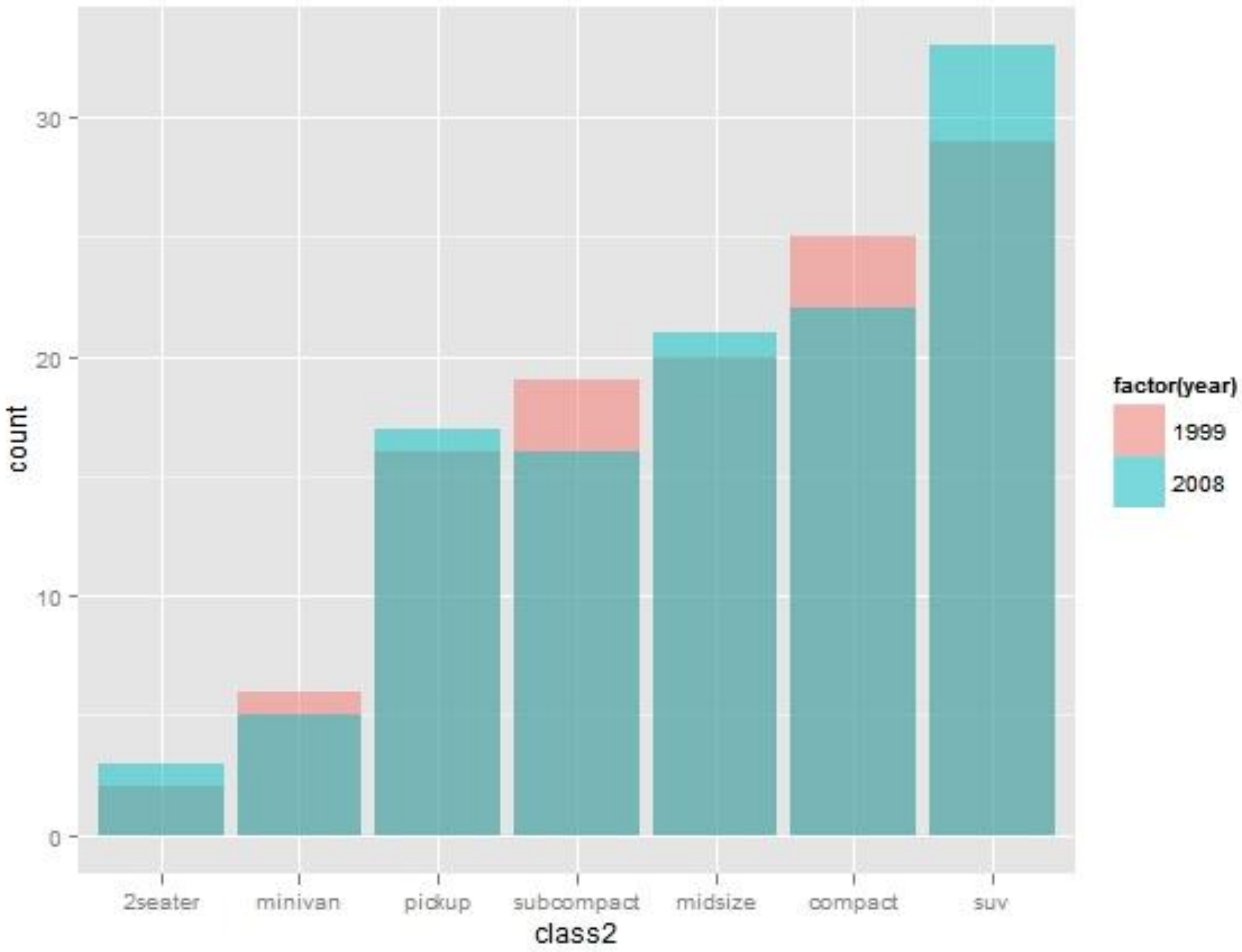
```
>  class2 <- mpg$class;  class2 <- reorder(class2,class2,length)
> mpg$class2  <-  class2
>  P  <-  ggplot(mpg, aes(x=class2))
>  p  +  geom_bar(aes(fill=class2))
```

根据计数排序后
绘制的条形图



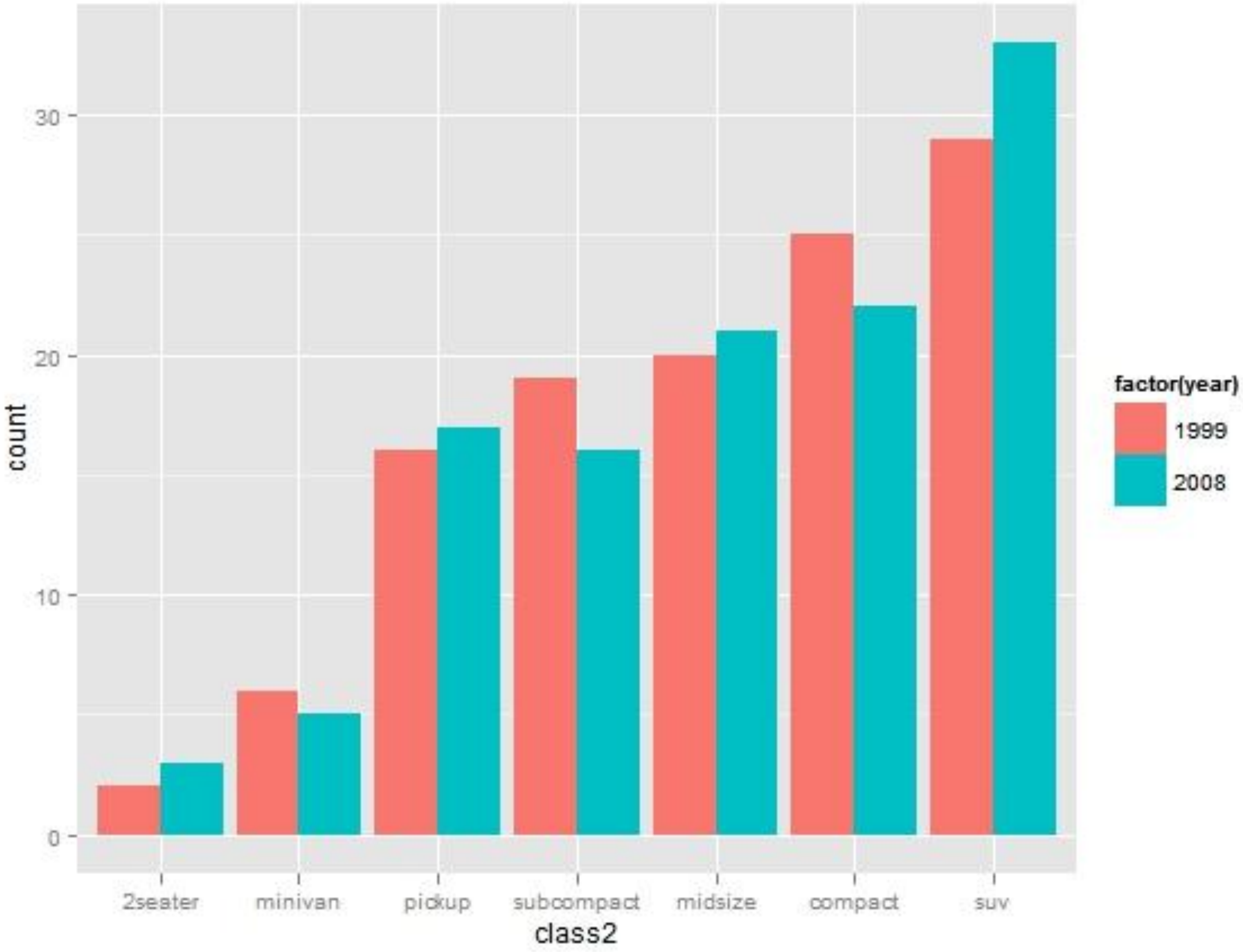
根据年份分别绘制条形图，position控制位置调整方式

```
> p <- ggplot(mpg, aes(class2,fill=factor(year)))
  p  + geom_bar(position='identity',alpha=0.5)
```



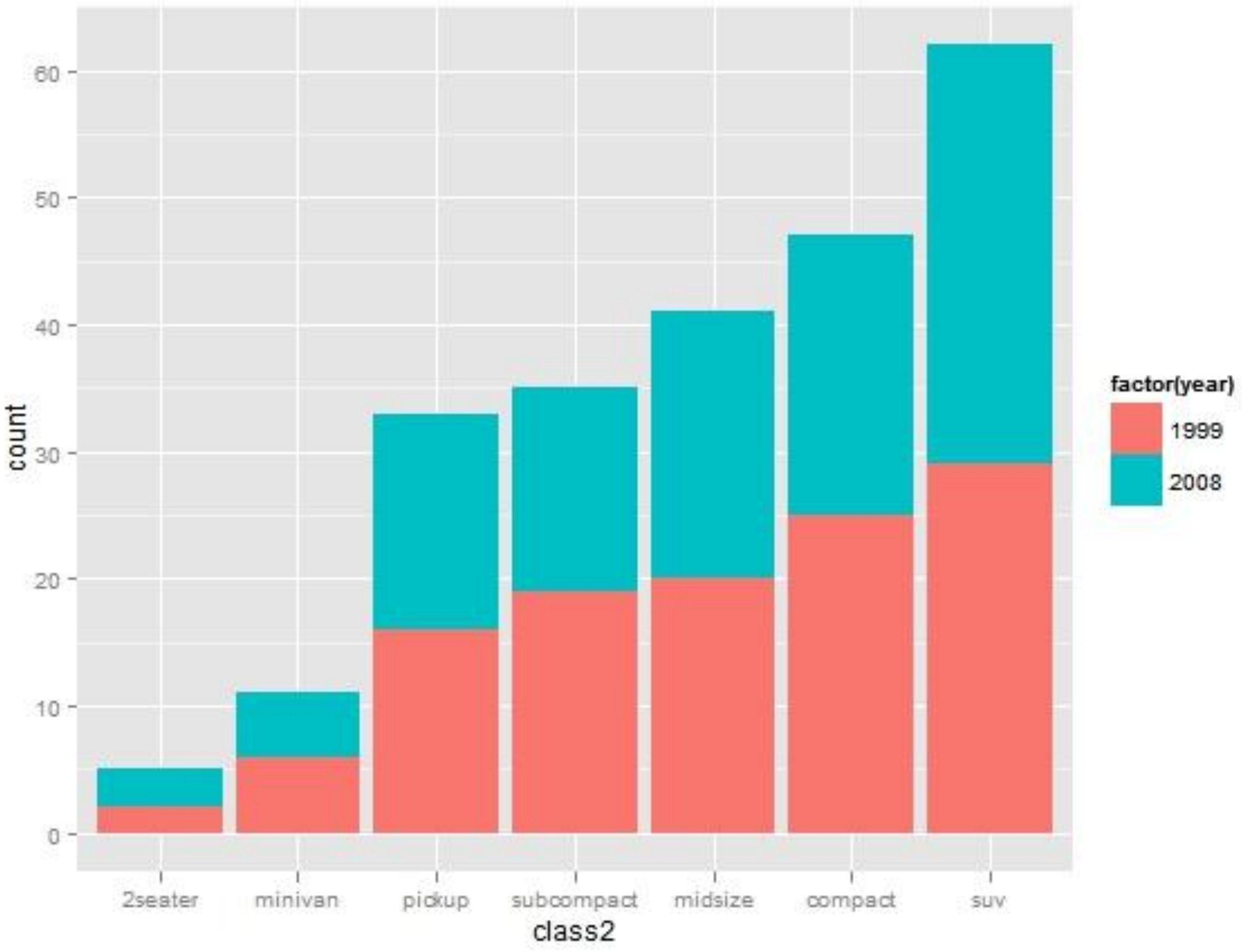
并立方式

```
> P + geom_bar(position='dodge')
```



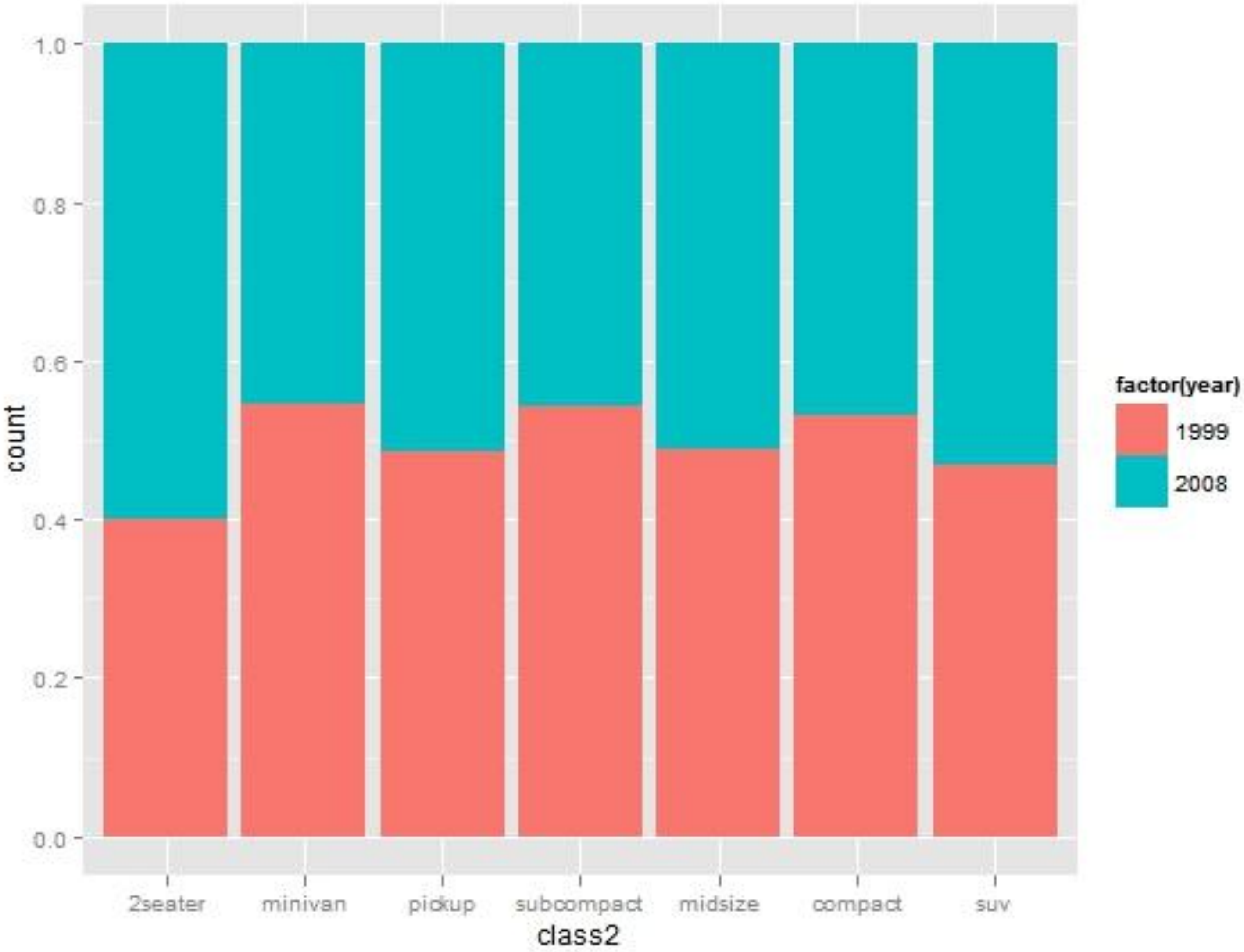
叠加方式

```
> p+geom_bar(position='stack')
```



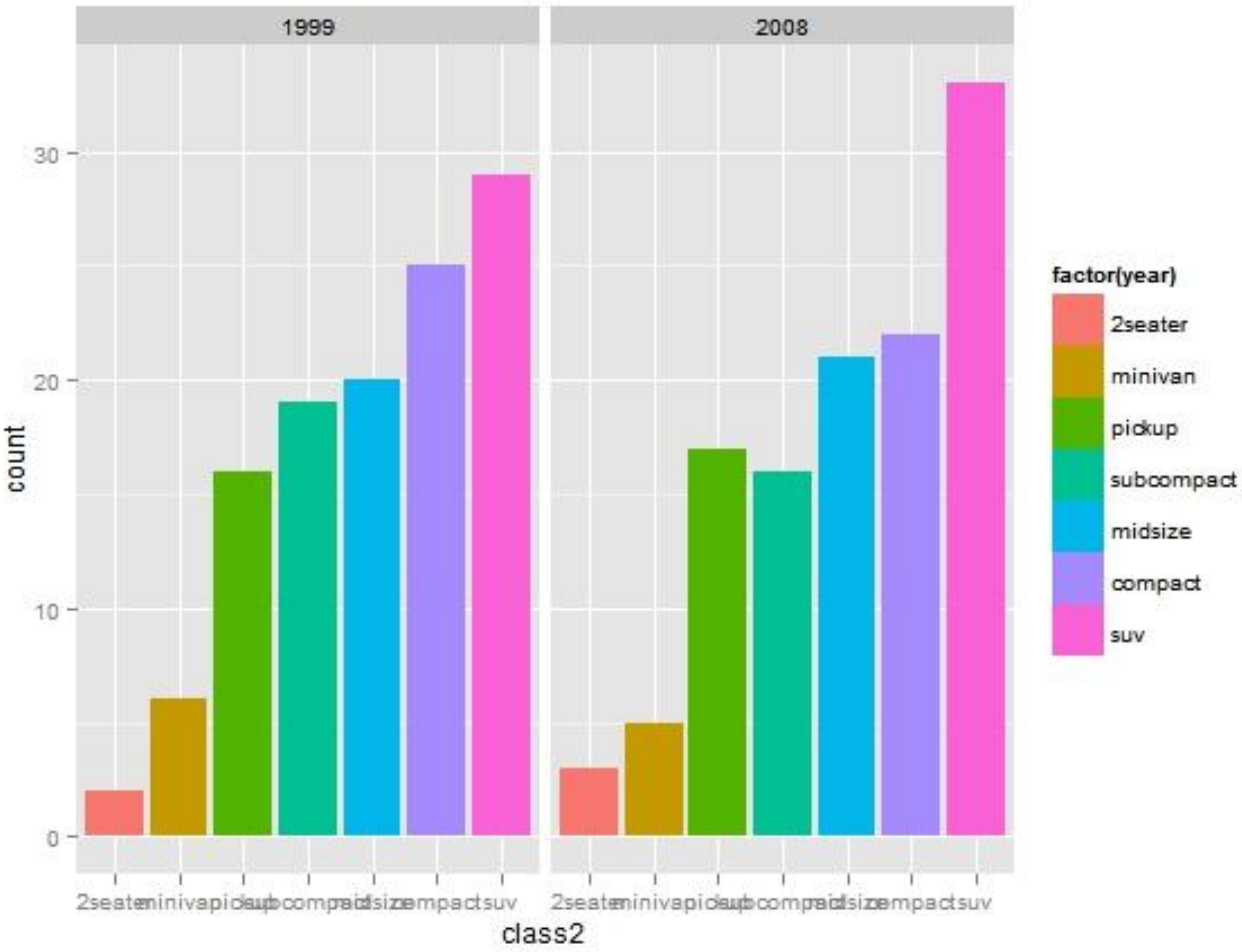
相对比例

```
> p+geom_bar(position='fill')
```



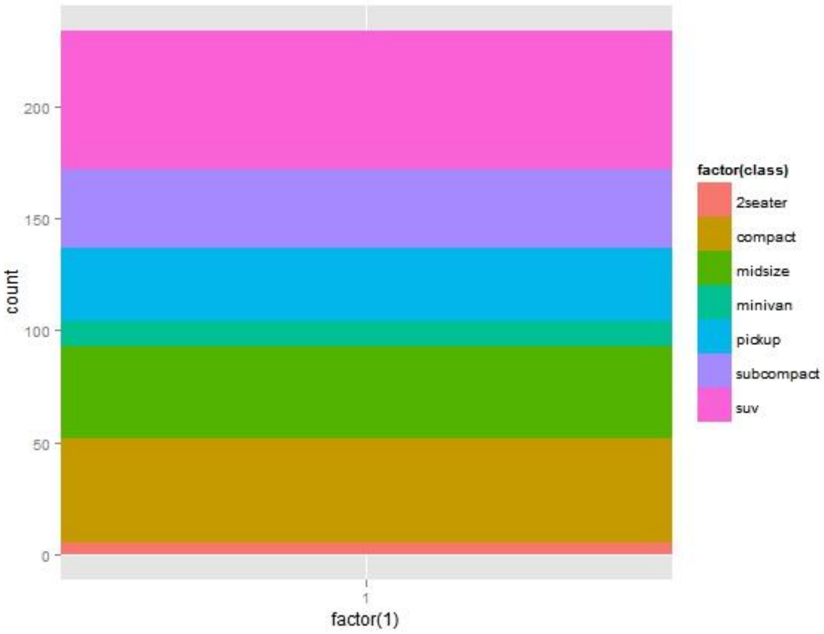
分面显示

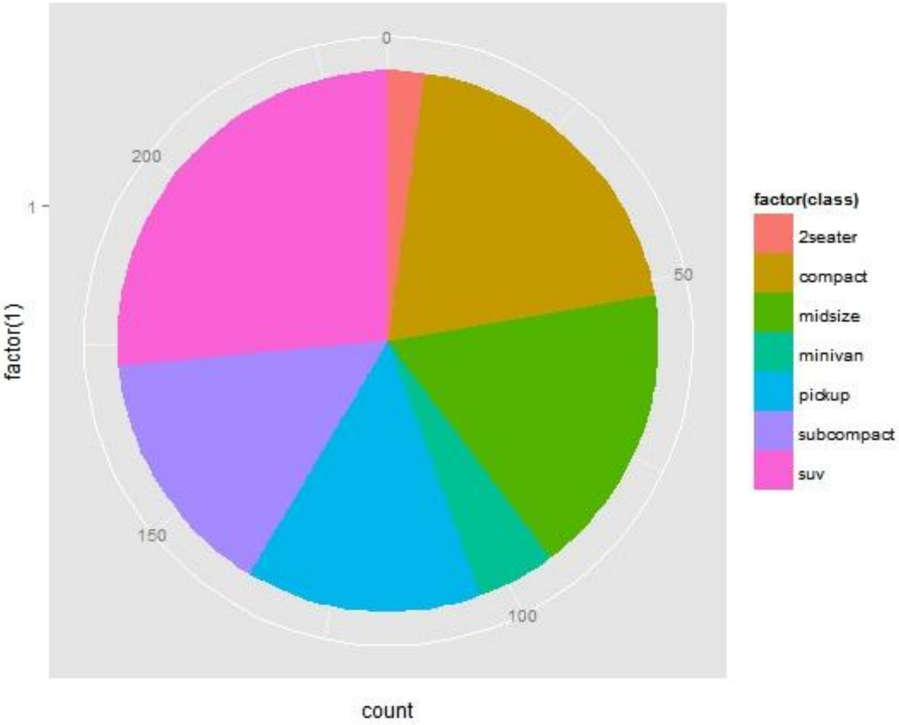
```
> p+ geom_bar(aes(fill=class2))+facet_wrap(~year)
```



饼图

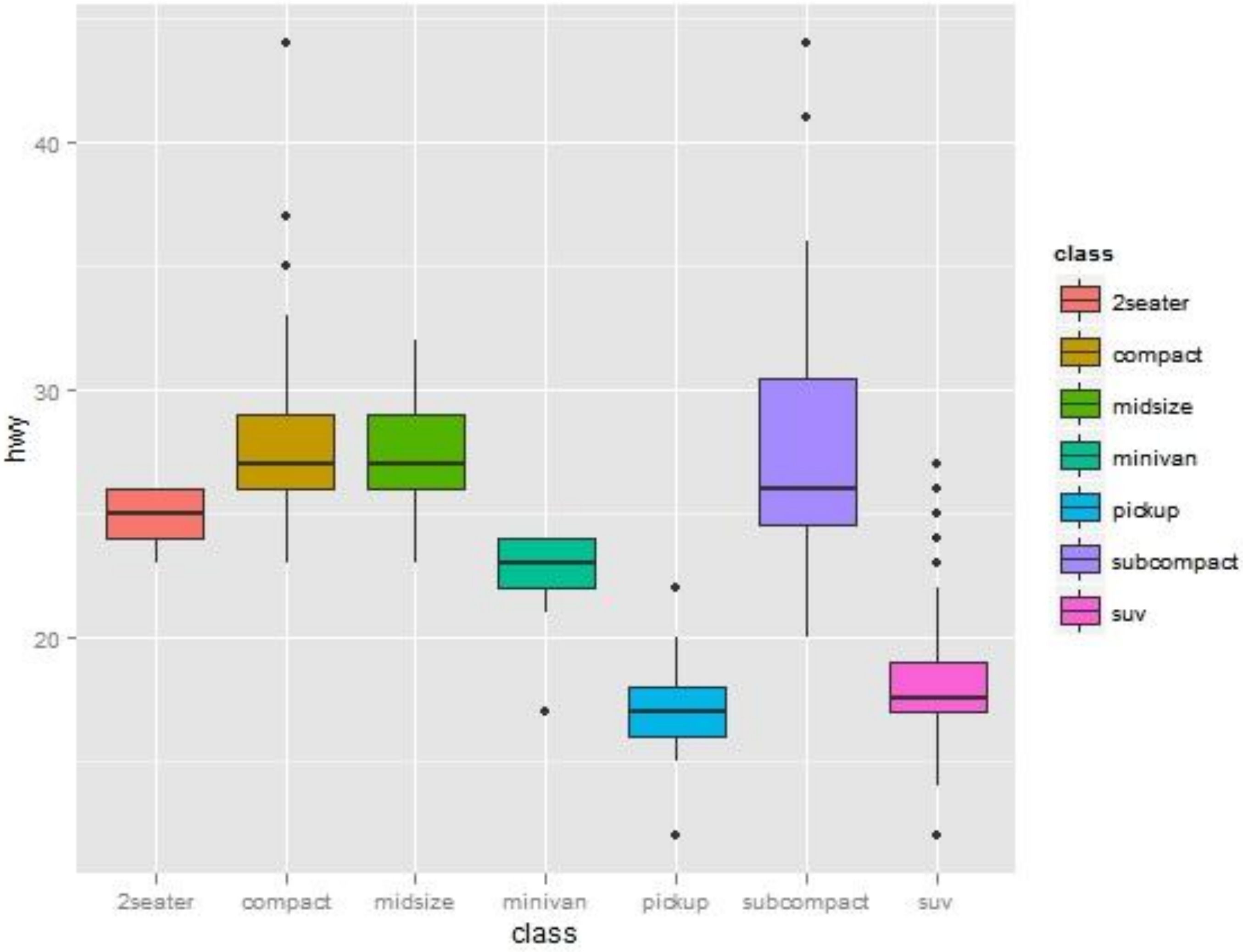
```
> p <- ggplot(mpg, aes(x = factor(1), fill = factor(class))) +
        geom_bar(width = 1)
  p + coord_polar(theta = "y")
```



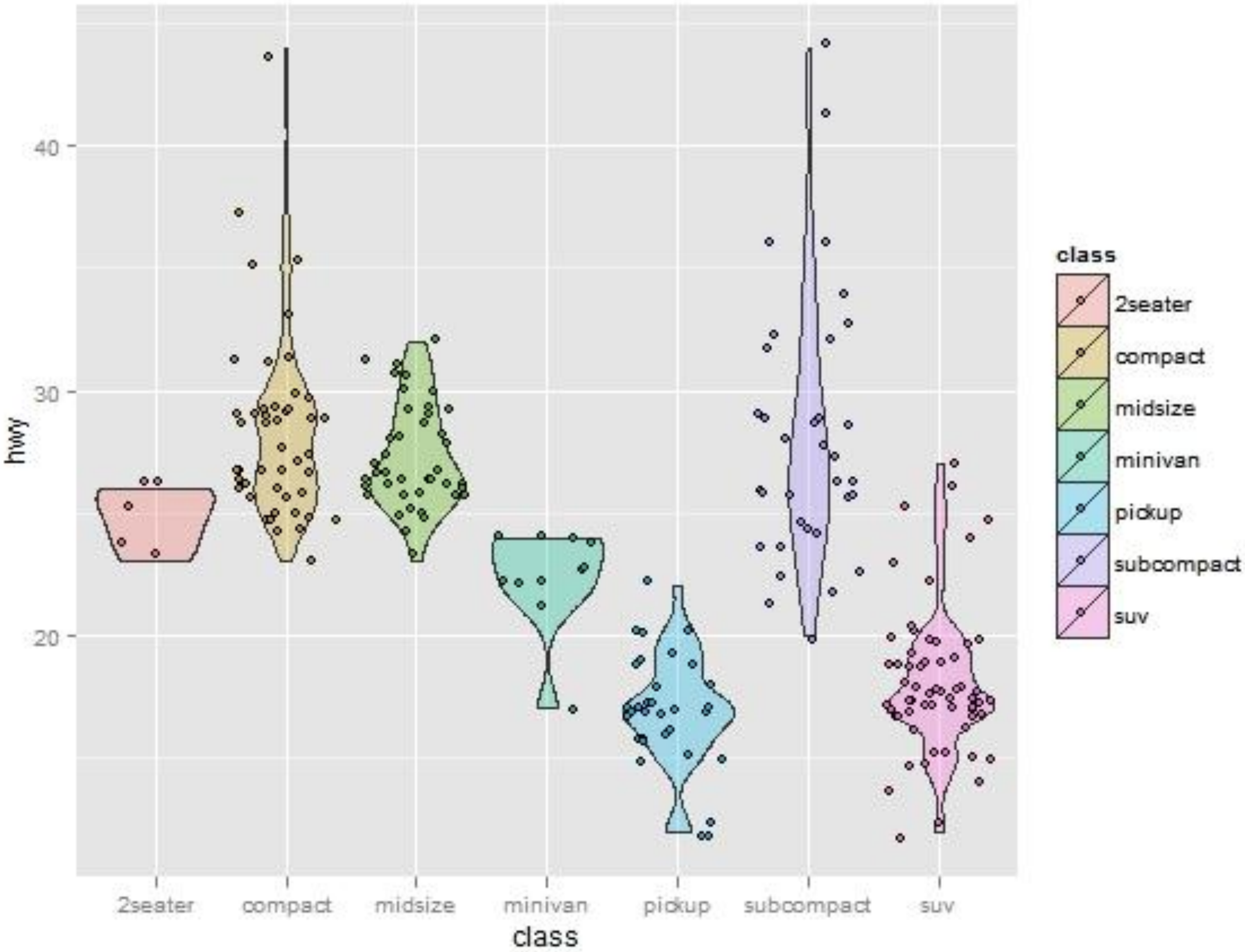


箱线图

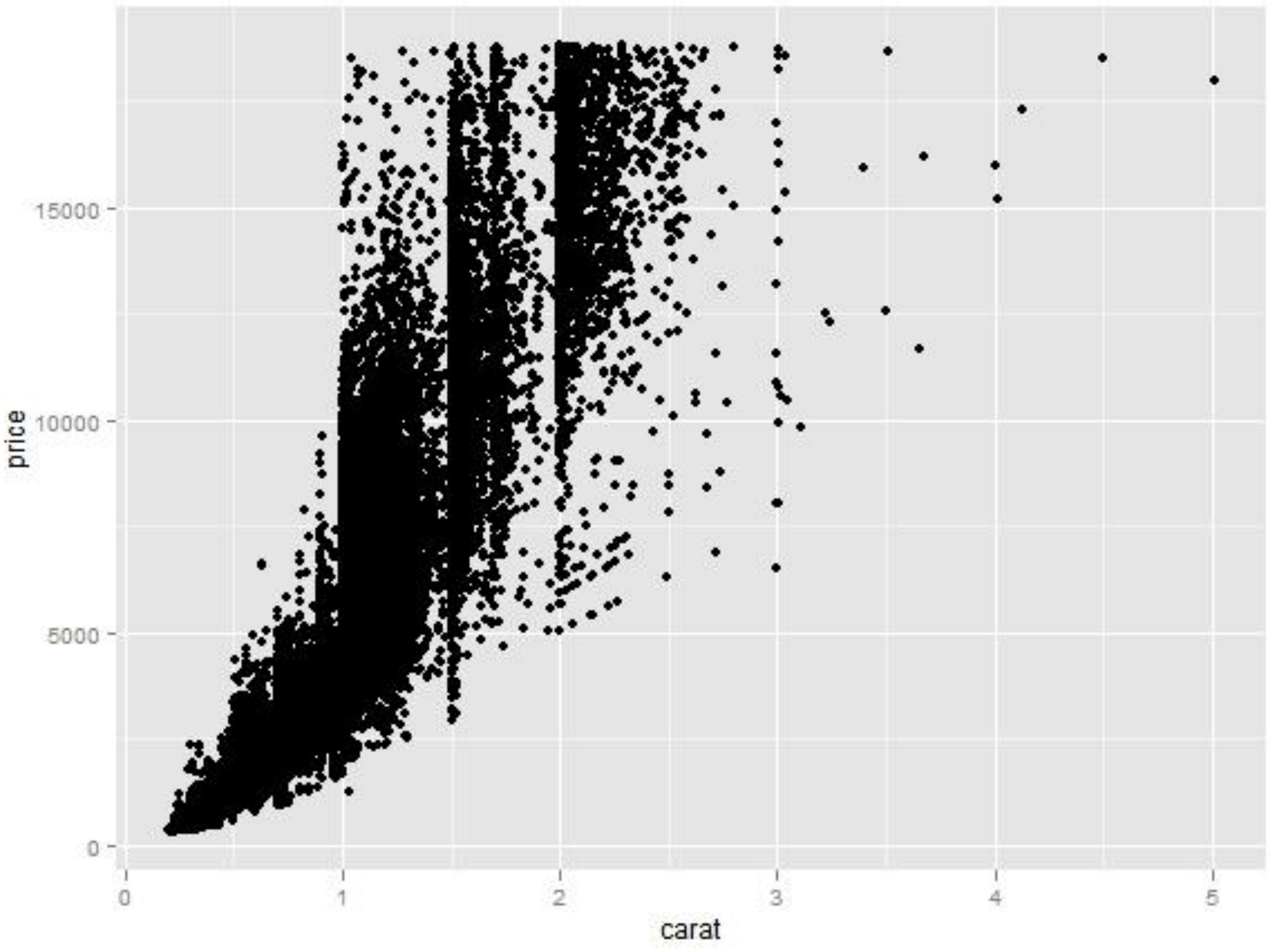
```
>  p <- ggplot(mpg, aes(class,hwy,fill=class))
  p+geom_boxplot()
```



```
>  P + geom_violin(alpha=0.3,width=0.9)+
      geom_jitter(shape=21)
```



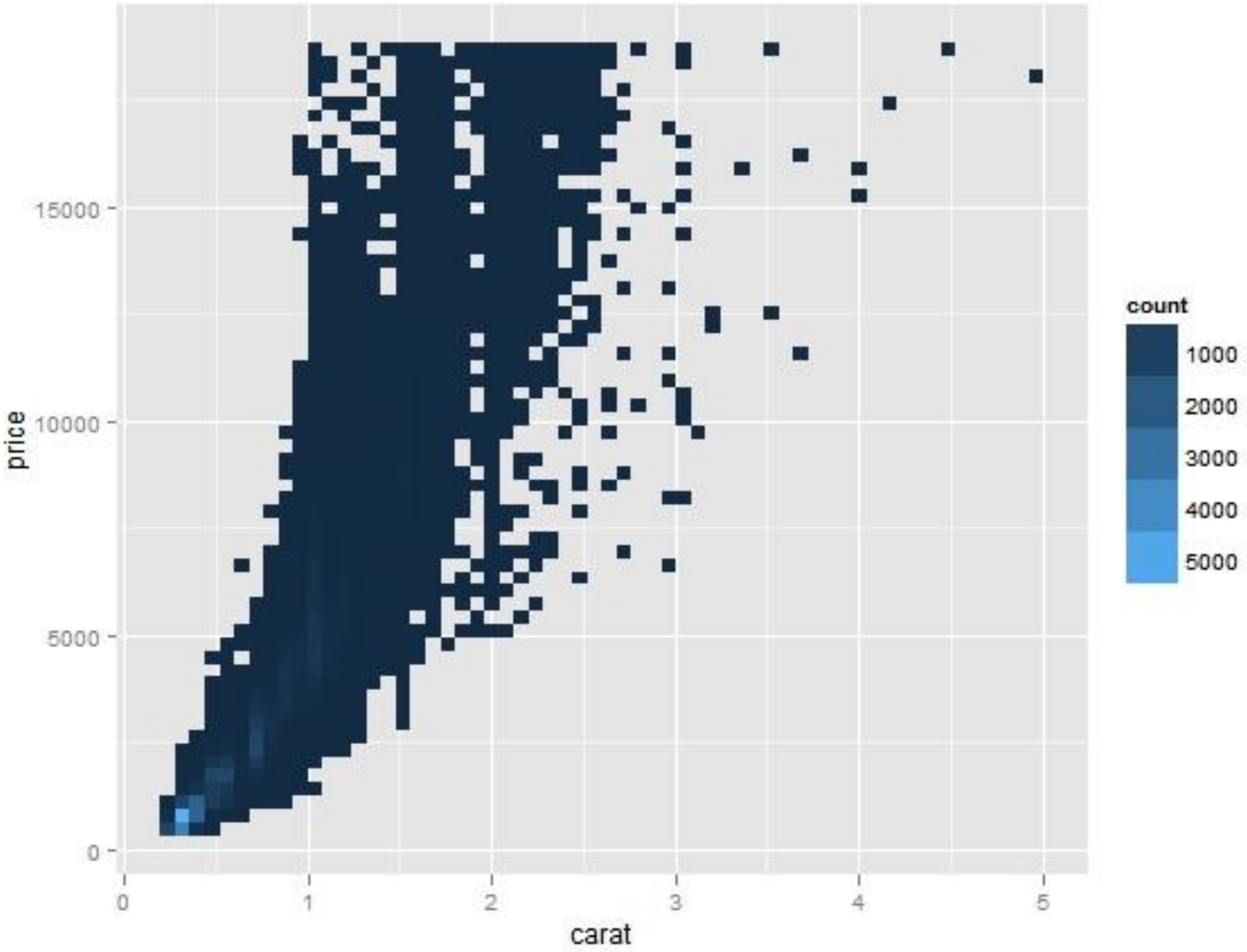
```
>  p <- ggplot(diamonds,aes(carat,price))
   p + geom_point()
```



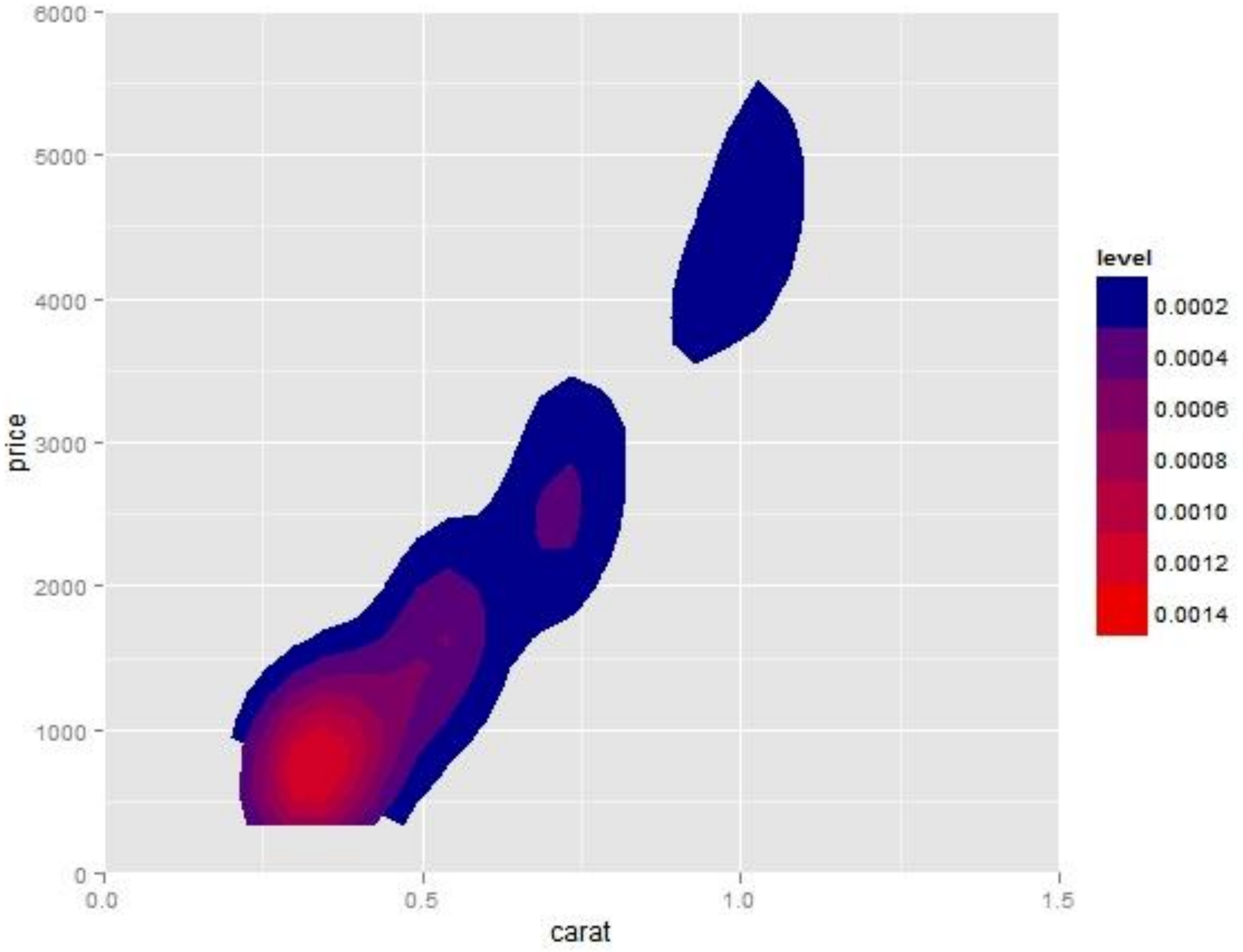
观察密集散点的方法

- 增加扰动（jitter）
- 增加透明度（alpha）
- 二维直方图（stat_bin2d）
- 密度图（stat_density2d）

```
>  p + stat_bin2d(bins = 60)
```



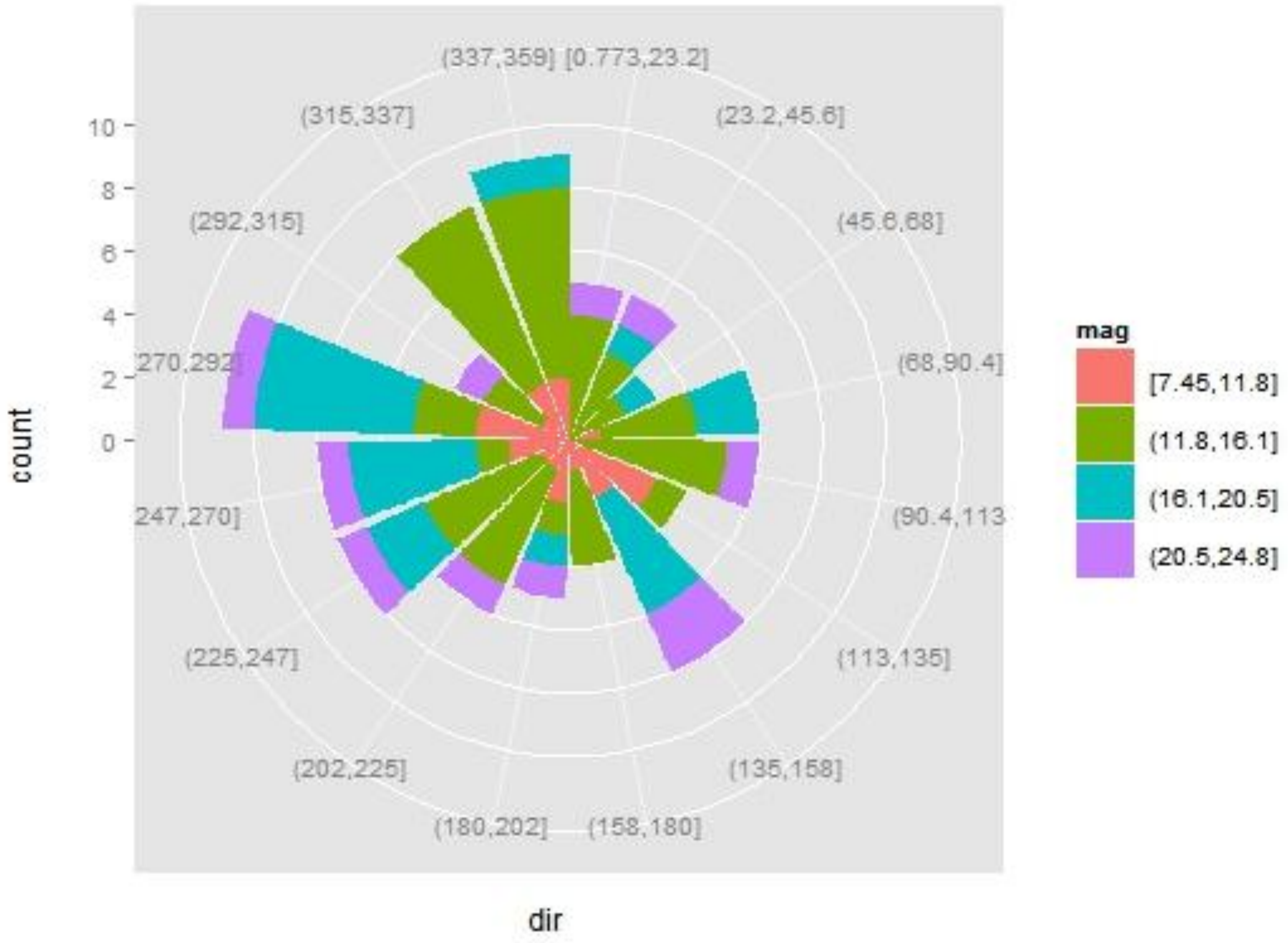
```
> p + stat_density2d(aes(fill = ..level..), geom="polygon") +
   coord_cartesian(xlim = c(0, 1.5),ylim=c(0,6000))+
   scale_fill_continuous(high='red2',low='blue4')
```



进阶示例

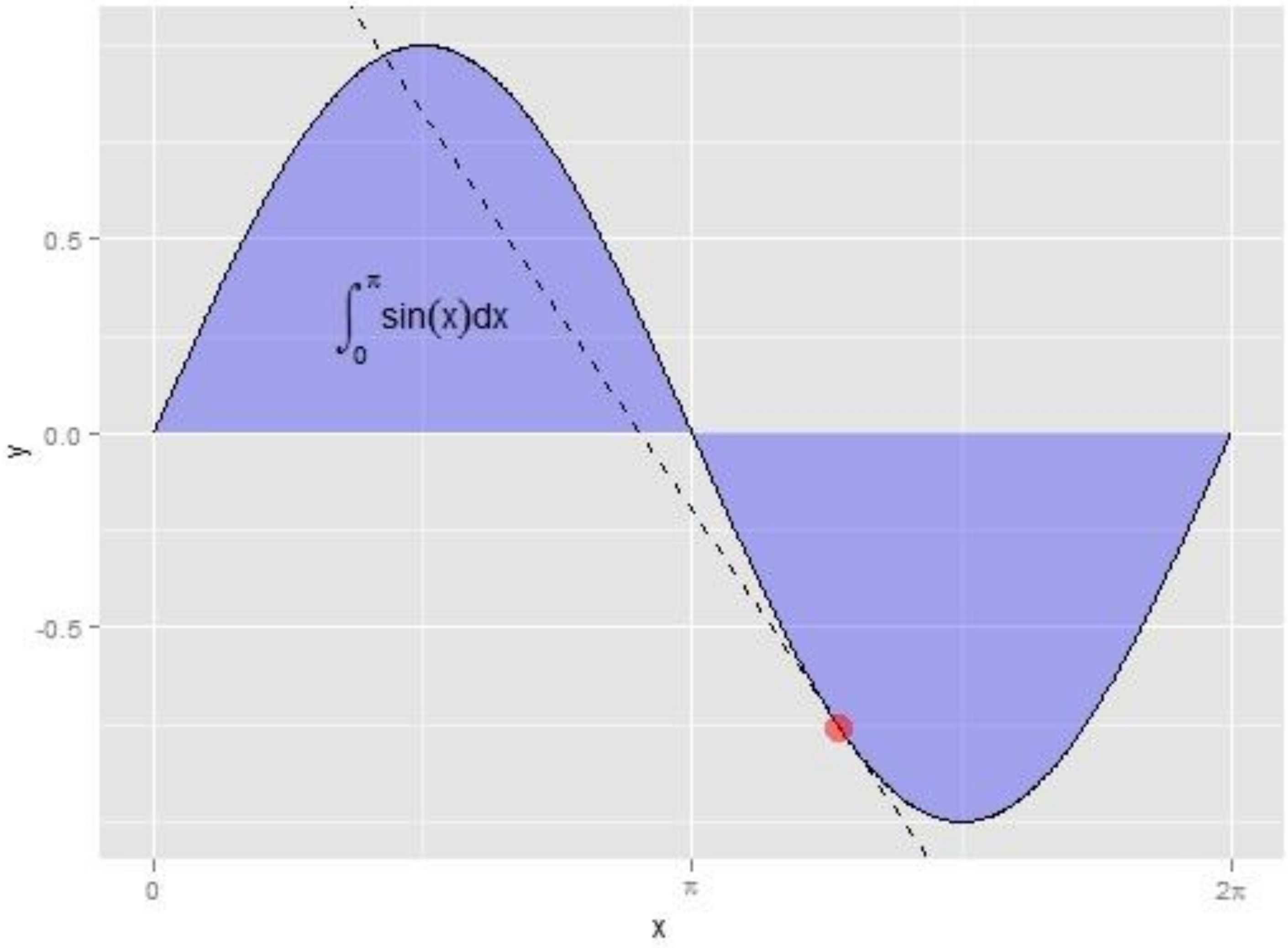
- 风向风速
- 插入数学符号
- 时间序列
- 水资源分布
- OpenStreetMap
- 日历热图

风向风速玫瑰图


(337,359]
[0.773,23.2]
(23.2,45.6]
(45.6,68]
(68,90.4]
(90.4,113
(113,135]
(135,158]
(158,180]
(180,202]
(202,225]
(225,247]
247,270]
270,292]
(292,315]
(315,337]
10
8
6
4
2
0
count
dir
mag
[7.45,11.8]
(11.8,16.1]
(16.1,20.5]
(20.5,24.8]

```
#随机生成100次风向，并汇集到16个区间内
dir <- cut_interval(runif(100,0,360),n=16)

#随机生成100次风速，并划分成4种强度
mag <- cut_interval(rgamma(100,15),4)
sample <- data.frame(dir=dir,mag=mag)

#将风向映射到X轴，频数映射到Y轴，风速大小映射到
填充色，生成条形图后再转为极坐标形式即可
p  <-  ggplot(sample,aes(x=dir,y=..count..,fill=mag))
p + geom_bar()+ coord_polar()
```

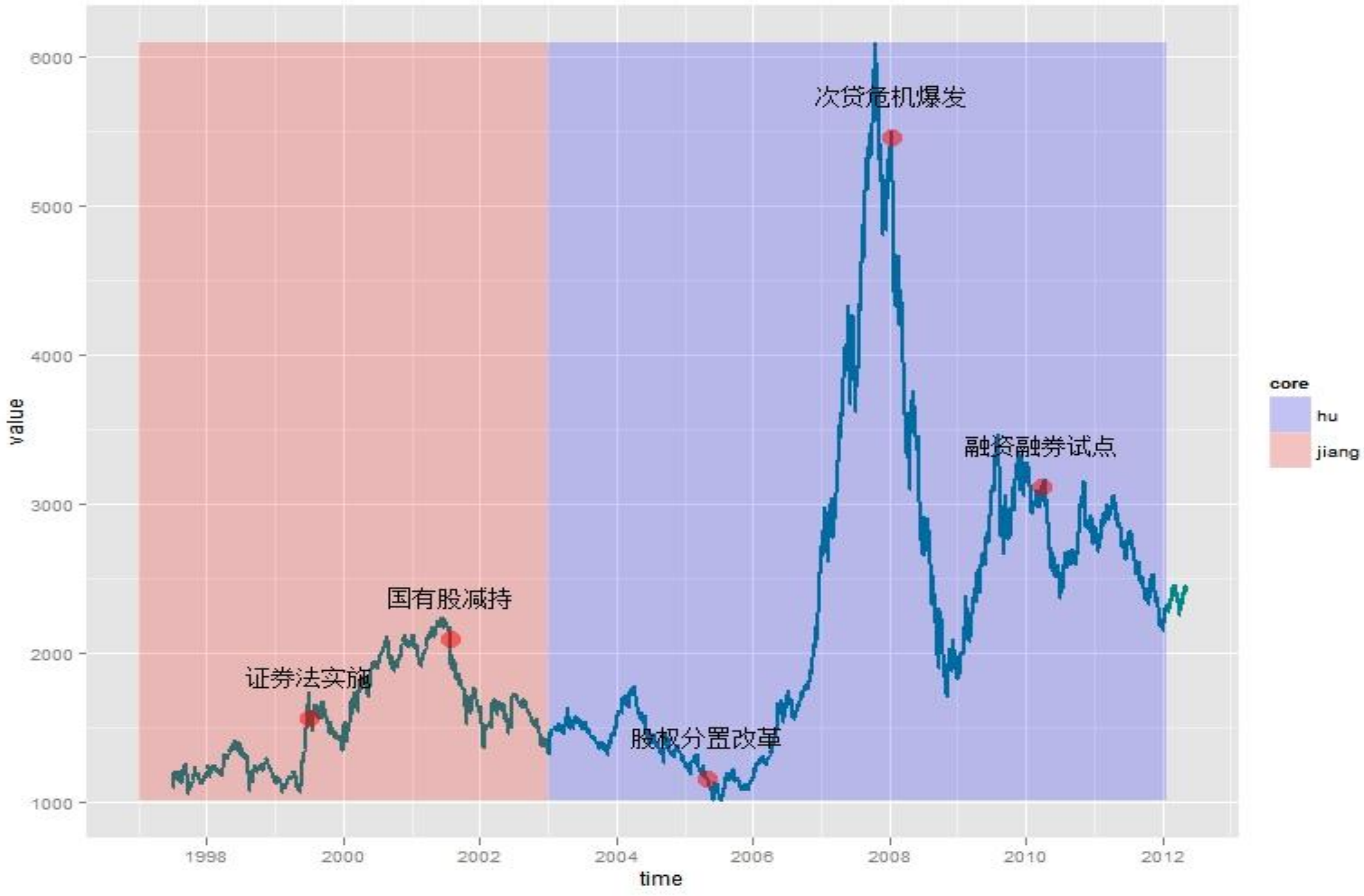
插入数学符号



使用的是老版本ggplot2，新版未测试

```
intercept <- sin(4)-slope*4
x <- seq(from=0,to=2*pi,by=0.01)
y <- sin(x)
p <-  ggplot(data.frame(x,y),aes(x,y))
p +  geom_area(fill=alpha('blue',0.3))+
 geom_abline(intercept=intercept,slope=slope,linetype=2)+
scale_x_continuous(breaks=c(0,pi,2*pi),
labels=c('0',expression(pi),expression(2*pi)))+
geom_text(parse=T,aes(x=pi/2,y=0.3,label='integral(sin(x)*
dx, 0, pi)'))+
 geom_line()+
geom_point(aes(x=4,y=sin(4)),size=5,colour=alpha('red',0.5
))
```

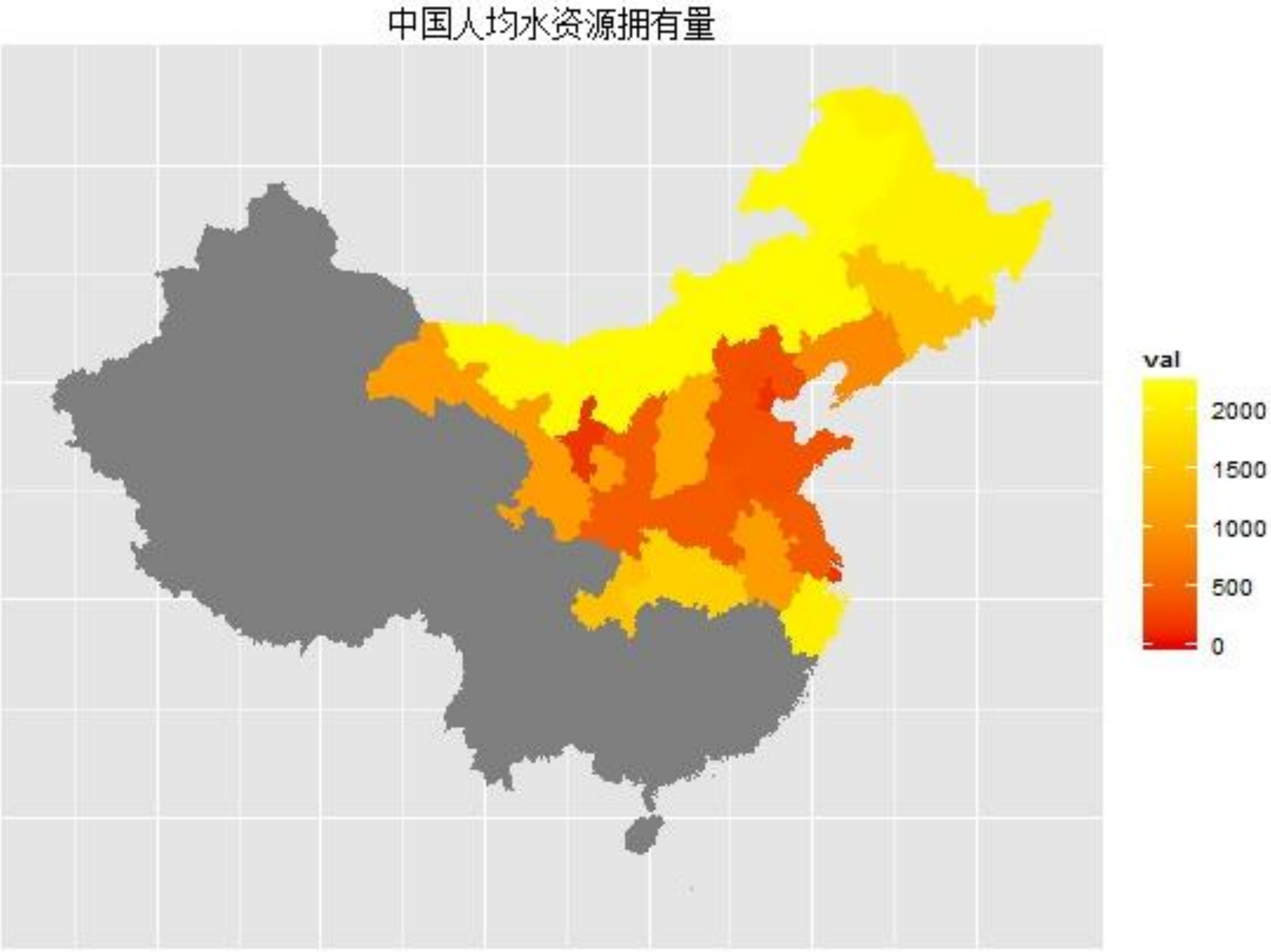
时间序列



6000
5000
4000
3000
2000
1000
value
次贷危机爆发
融资融券试点
国有股减持
证券法实施
股权分置改革
1998
2000
2002
2004
2006
2008
2010
2012
time
core
hu
jiang

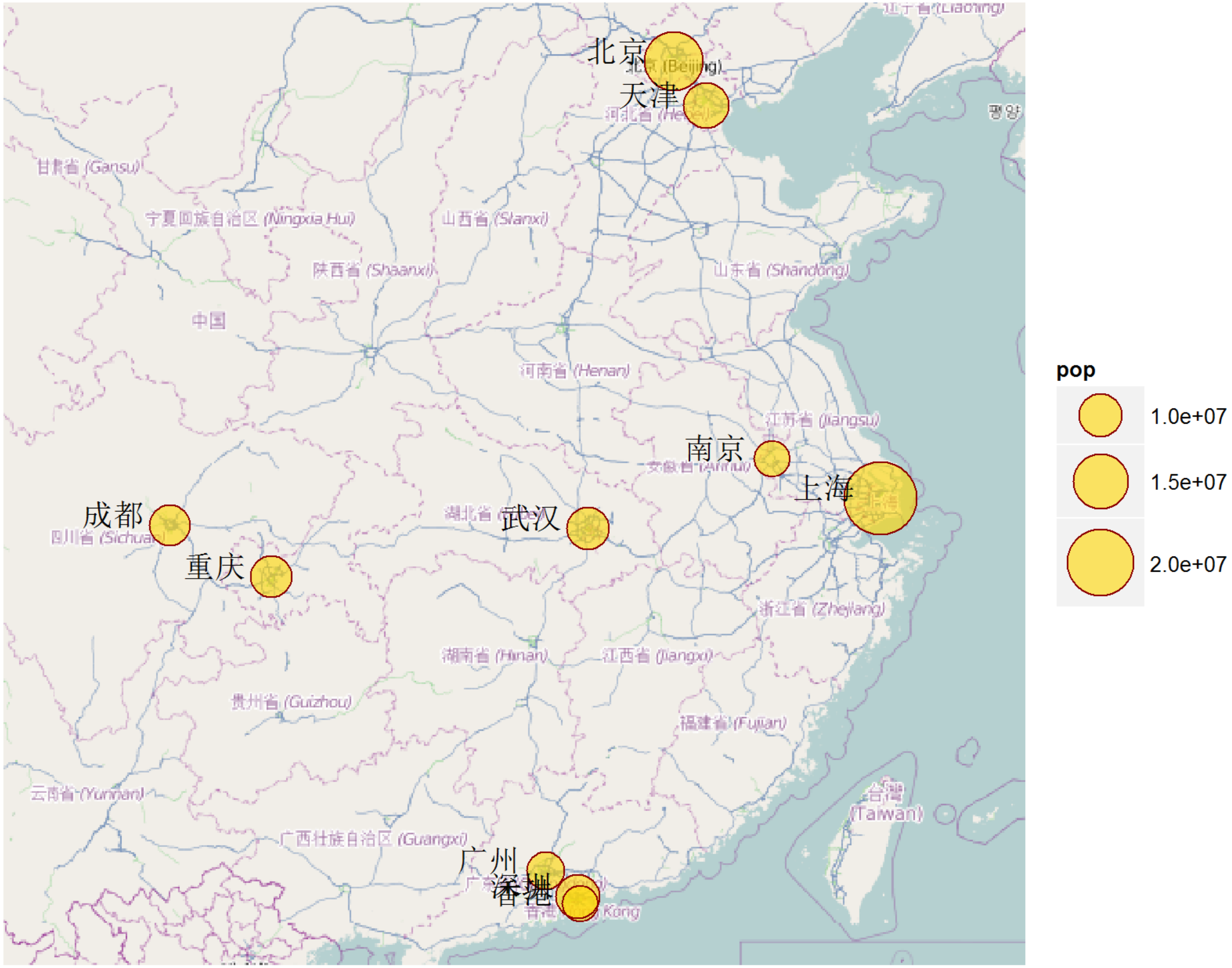
```
library(quantmod)
library(ggplot2)
getSymbols('^SSEC',src='yahoo',from = '1997-01-01')
close <- (Cl(SSEC))
time <- index(close)
value <- as.vector(close)
yrng <- range(value)
xrng <- range(time)
data <- data.frame(start=as.Date(c('1997-01-01','2003-01-01')),end=as.Date(c('2002-12-30','2012-01-20')),core=c('jiang','hu'))
timepoint <- as.Date(c('1999-07-02','2001-07-26','2005-04-29','2008-01-10','2010-03-31'))
events <- c('证券法实施','国有股减持','股权分置改革','次贷危机爆发','融资融券试点')
data2 <- data.frame(timepoint,events,stock=value[time %in% timepoint])

p <- ggplot(data.frame(time,value),aes(time,value))
p + geom_line(size=1,colour='turquoise4')+
geom_rect(alpha=0.2,aes(NULL,NULL,xmin = start, xmax = end, fill = core),ymin = yrng[1],ymax=yrng[2],data = data)+
scale_fill_manual(values = c('blue','red'))+
geom_text(aes(timepoint, stock, label = events),data = data2,vjust = -2,size = 5)+
geom_point(aes(timepoint, stock),data = data2,size = 5,colour = 'red',alpha=0.5)
```

水资源分布

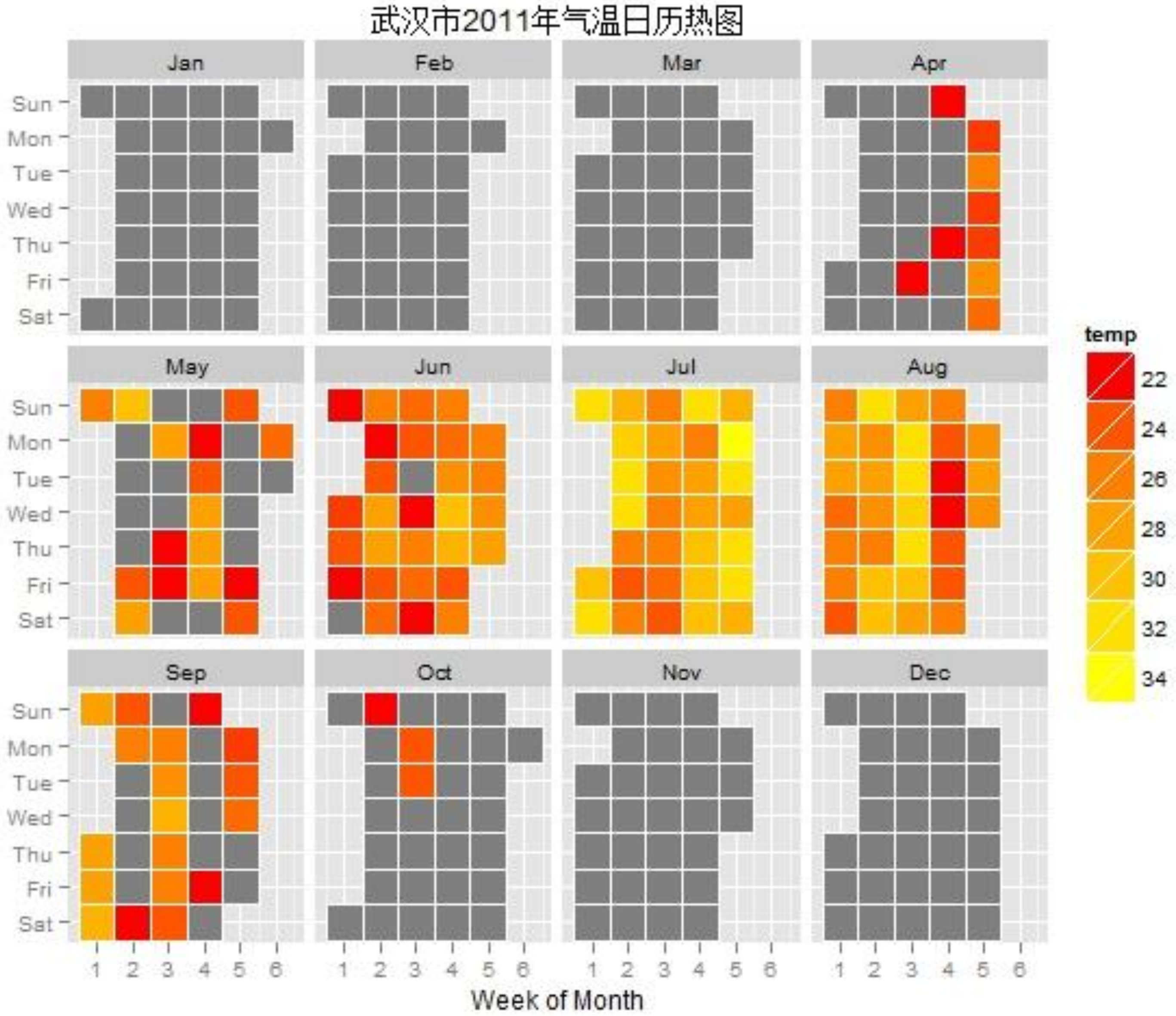


```
library(ggplot2)
library(gpclib)
library(maptools)
load(url("http://gadm.org/data/rda/CHN_adm1.RData"))
water  <- c(1085,325,1473,3524,1079,2935,3989,2790,4147,358,2046,434
   ,1652,2490,451,3362,1467,871,2145,182,1000,12278,448,377,
   182,1221,3135,152,4976,10000,5298,2005)

gpclibPermit()
china.map  <-  fortify(gadm,region='ID_1')
vals  <-   data.frame(id =unique(china.map$id),val=water)

ggplot(vals, aes(map_id = id)) +
    geom_map(aes(fill = val), map =china.map) +
    expand_limits(x = china.map$long, y = china.map$lat) +
   scale_fill_continuous(limits=c(0,2200),low = 'red2',high ='yellow',
        guide = "colorbar") +
  opts(title='中国人均水资源拥有量',
       axis.line=theme_blank(),axis.text.x=theme_blank(),
       axis.text.y=theme_blank(),axis.ticks=theme_blank(),
       axis.title.x=theme_blank(),
       axis.title.y=theme_blank()) +
  xlab("") + ylab("")
```

OpenStreetMap



日历热图


武汉市2011年气温日历热图
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Sun
Mon
Tue
Wed
Thu
Fri
Sat
1
2
3
4
5
6
Week of Month
temp
22
24
26
28
30
32
34

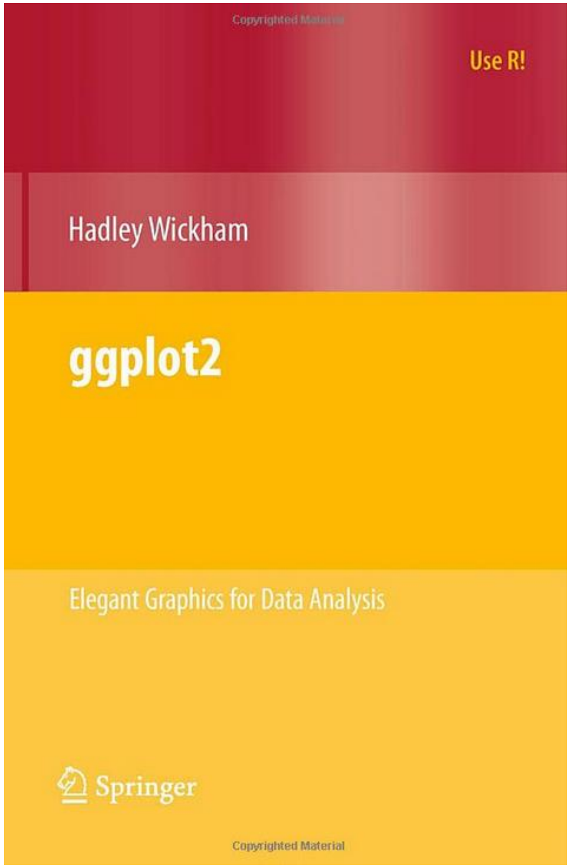
学习资源

- 教材 （中文版即将面世）
- 官方网站

http://had.co.nz/ggplot2/

- 0.9新功能说明

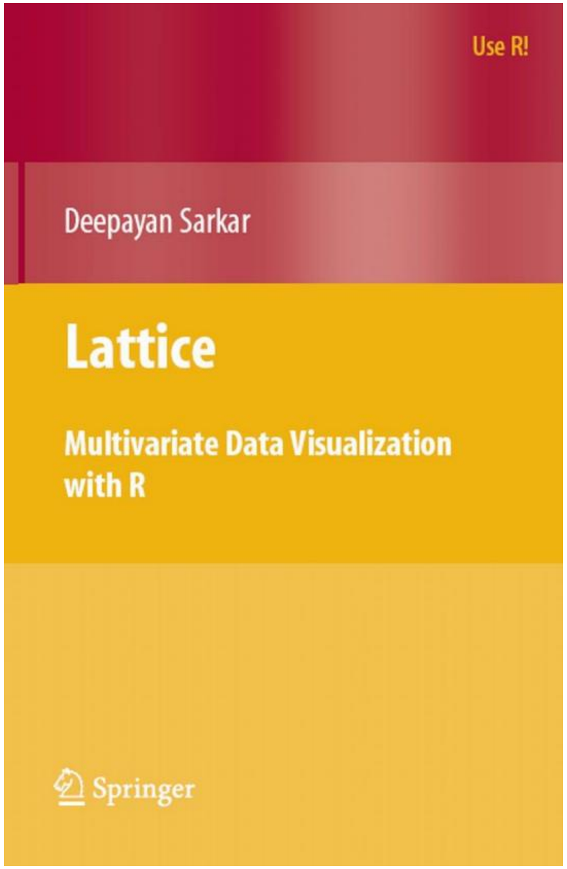
http://cloud.github.com/downloads/hadley/ggplot2/guide-col.pdf



学习资源

http://learnr.wordpress.com

该博客将所有Lattice作的图全部用ggplot2重画了一遍。



Use R!

Deepayan Sarkar

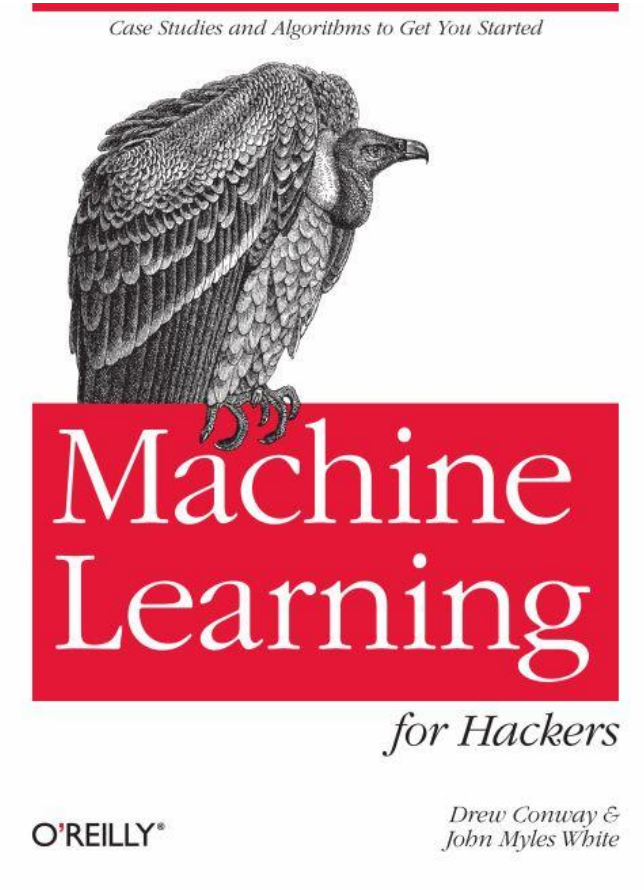
Lattice

Multivariate Data Visualization with R

Springer

学习资源

该书中所用图形均为ggplot2包绘制。



学习资源

- http://wiki.stdout.org/rcookbook/Graphs/
- http://r-blogger.com
- http://Stackoverflow.com
- http://xccds1977.blogspot.com （需科学上网）
- http://r-ke.info/
- http://www.youtube.com/watch?v=vnVJJYi1mbw

谢谢