Data Analysis Tools and Practice(Using R)

2021.04.13

R数据可视化手册

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

课堂测试时间

课堂测试07

40分钟 先用电脑完成 然后誊抄纸上

- 使用ggplot2里的画图函数完成以下的练习：

  - ✳ 1、将数据集Big_Mart_Dataset.csv,加载到R空间，将数据框命名为mart,查看mart的维度和基本结构。

  - ✳ 2、画Item_MRP和Item_Visibility的关系图，要求: (1)指定颜色属性为Item_Type; (2)设置x轴的标度(scale)，x轴名字为Item Visibility", x轴刻度为0-0.35以0.05为间隔的数值序列；设置y轴的标度(scale)，y轴名字为Item MRP , y轴刻度为0-270以30为间隔的数值序列; (3)设置图形主题为theme_bw，图形标题为Scatterplot。

  - ✳ 3、在2基础上，根据因子类型的列Item_Type进行分面。

  - ✳ 4、画列变量Item_MRP的直方图，要求：(1)每个小圆柱体的宽度为2, (2)设置x轴的标度(scale)，x轴名字为Item MRP, x轴刻度为0-270以30为间隔的数值序列；设置y轴的标度(scale)，y轴名字为Count,y轴刻度为0-200以20为间隔的数值序列;(3)设置标题为”Histogram”

课堂测试07

先用电脑完成

40分钟　然后誊抄纸上

- 使用ggplot2里的画图函数完成以下的练习：

  - ✳ 5、画出列变量Outlet_Establishment_Year的条形图，要求(1): 填充色为”red”; (2): 主题为theme_bw和theme_gray;(3): 设置x轴的标度(scale)，x轴名字为Establishment_Year, x轴刻度为1985-2010为间隔的数值序列；设置y轴的标度(scale)，y轴名字为Count , y轴刻度为0-1500以150为间隔的数值序列; (4): 设置标题为Bar Chart，翻转坐标轴

  - ✳ 6、画出Outlet_Location_Type堆叠的条形图 (1): 使用 Outlet_Type设置填充色; (2): 设置图形的标题为Stacked Bar Chart，x轴的名称为Outlet Location Type", y轴的名称为Count of Outlets

  - ✳ 7、画Outlet_Identifier以Item_Outlet_Sales为分类变量的箱型图;(1): 填充色为红色; (2): y轴名称为“Item Outlet Sales”, 坐标为0-15000以150为间隔的数值序列; (3): 设置标题为”Box Plot", x 轴坐标为"Outlet Identifier

  - ✳ 8、画列变量Item_Outlet_Sales面积图表 要求: (1)统计变换为 "bin", bin的宽度为30, 填充色为”steelblue”;(2)x轴的标度为0-11000以1000间隔的数值序列;(3)图形标题为”Area Chart", x 轴命名为 "Item Outlet Sales", y轴命名为 “Count”。

上次课程内容回顾

- ggplot(), 图层
  - ✳ data; mapping; geom; stat; position; aes(); layer();
- geom_xxx：
  - ✳ point; path; bar; histogram; smooth; density; jitter; tile; area; polygon;
  - ✳ line; vline; hline; abline; rect; text; arrow;
- stat_xxx：
  - ✳ identity; smooth; function; boxplot; density; quantile; sum; unique;
  - ✳ stat_bin; stat_bin2d; stat_binhex; stat_density2d; stat_summary;
- 其余：
  - ✳ fill; bins; colour; group; labs; binwidth; shape; alpha; maps;

参考图书

R Graphics Cookbook

R数据可视化手册

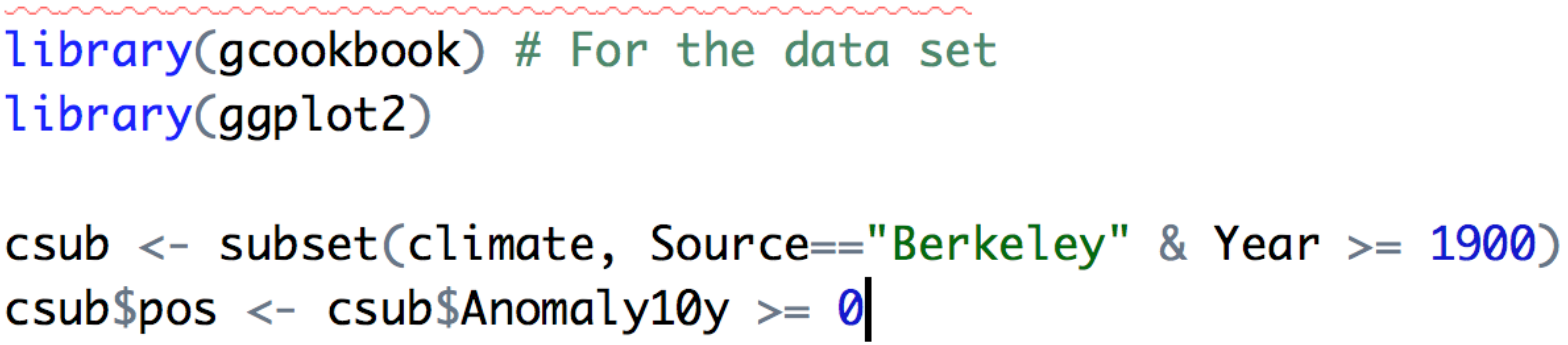
[美] *Winston Chang* 著
肖楠 邓一硕 魏太云 译
邱怡轩 审校

O'REILLY®

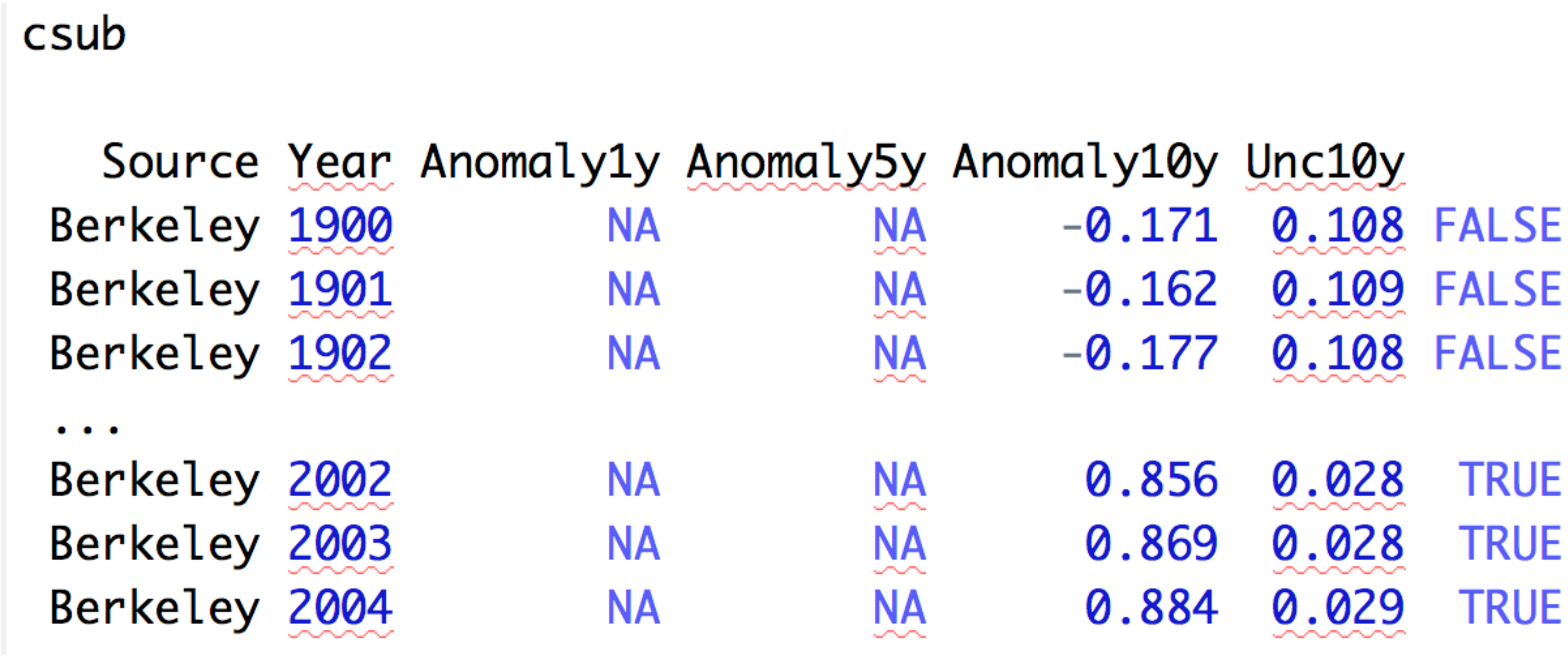
人民邮电出版社
POSTS & TELECOM PRESS

对正负条形图分别着色

3.5 P24



```r
library(gcookbook) # For the data set
library(ggplot2)

csub <- subset(climate, Source=="Berkeley" & Year >= 1900)
csub$pos <- csub$Anomaly10y >= 0
```



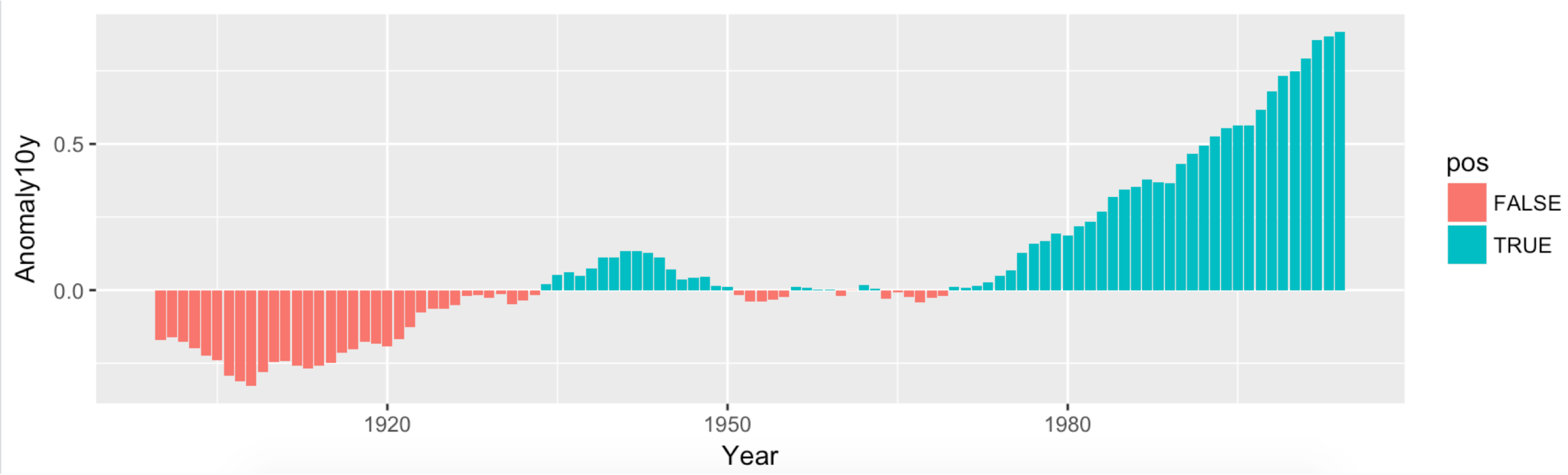
```
csub

   Source Year Anomaly1y Anomaly5y Anomaly10y Unc10y
 Berkeley 1900        NA        NA     -0.171  0.108 FALSE
 Berkeley 1901        NA        NA     -0.162  0.109 FALSE
 Berkeley 1902        NA        NA     -0.177  0.108 FALSE
 ...
 Berkeley 2002        NA        NA      0.856  0.028  TRUE
 Berkeley 2003        NA        NA      0.869  0.028  TRUE
 Berkeley 2004        NA        NA      0.884  0.029  TRUE
```

RGCook

对正负条形图分别着色

3.5 P24

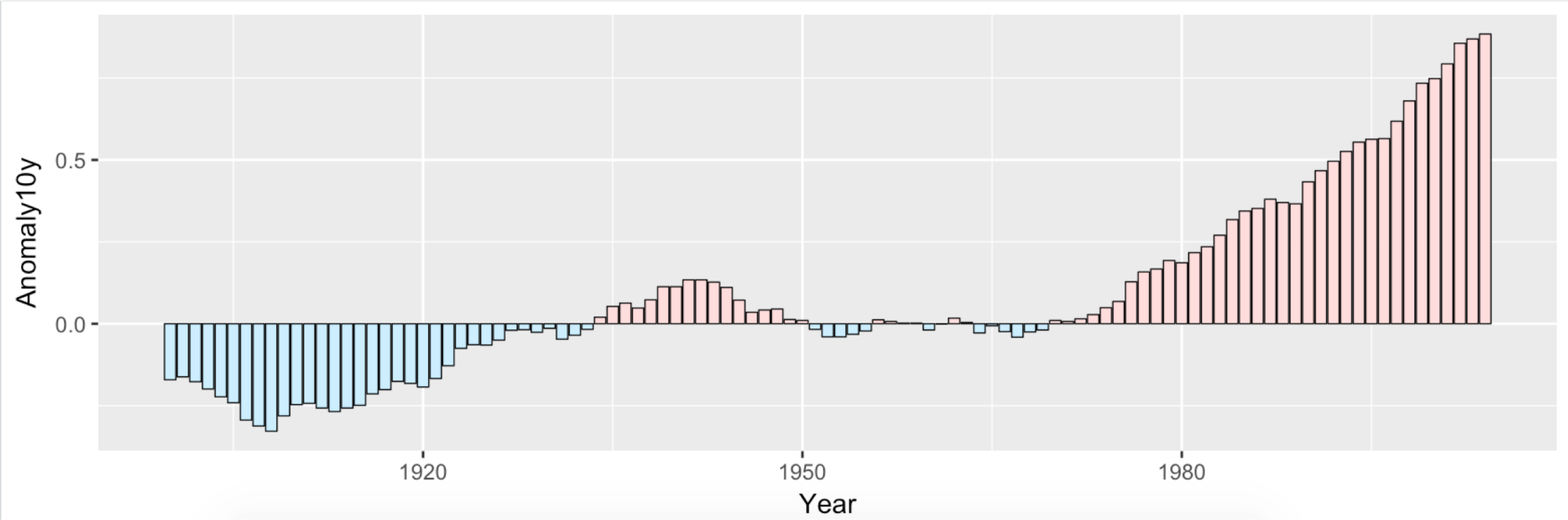
```
ggplot(csub, aes(x=Year, y=Anomaly10y, fill=pos)) +
  geom_bar(stat="identity", position="identity")
```



对正负条形图分别着色

3.5 P24

```
ggplot(csub, aes(x=Year, y=Anomaly10y, fill=pos)) +
    geom_bar(stat="identity", position="identity", colour="black", size=0.25) +
    scale_fill_manual(values=c("#CCEEFF", "#FFDDDD"), guide=FALSE)
```

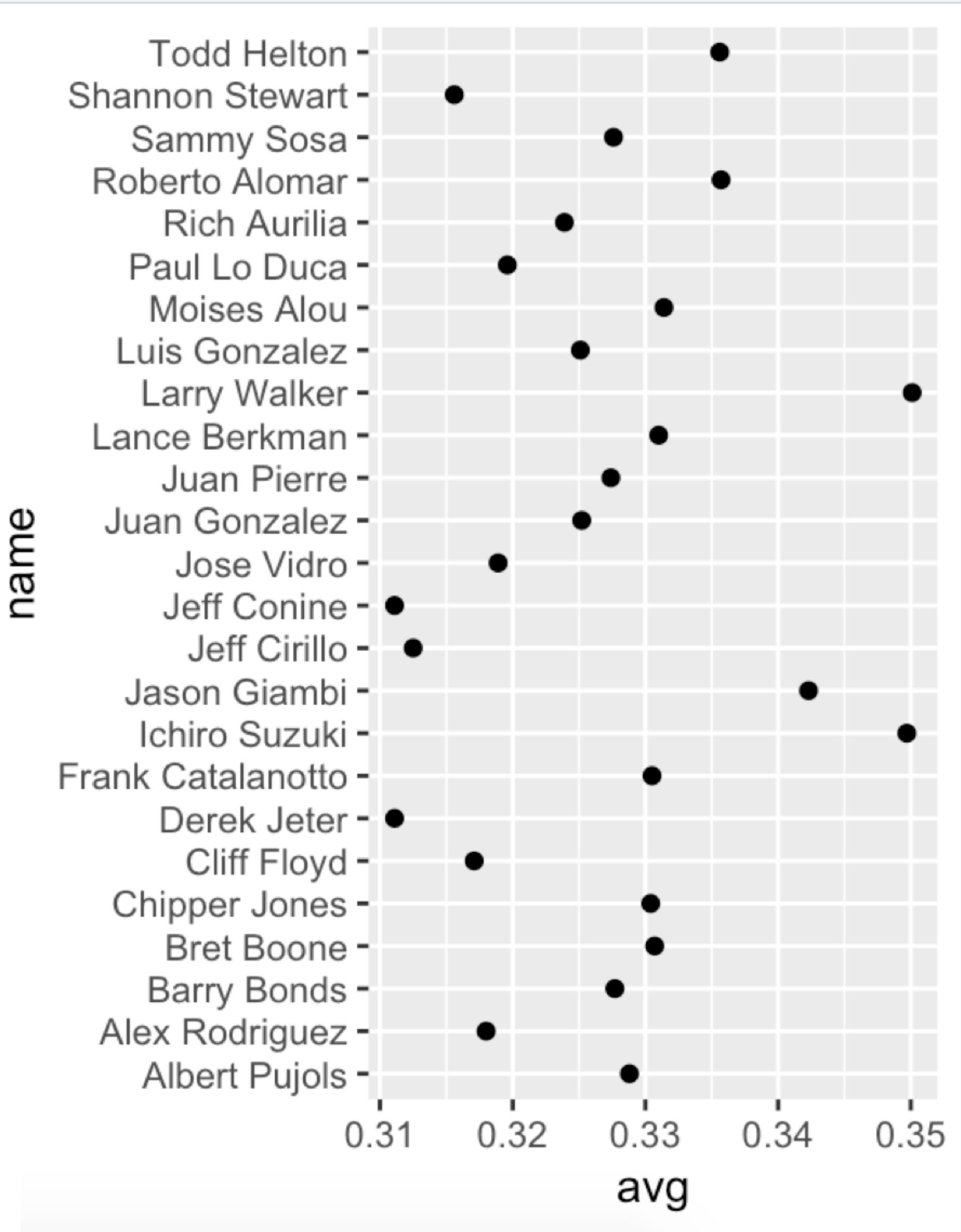


RGCook

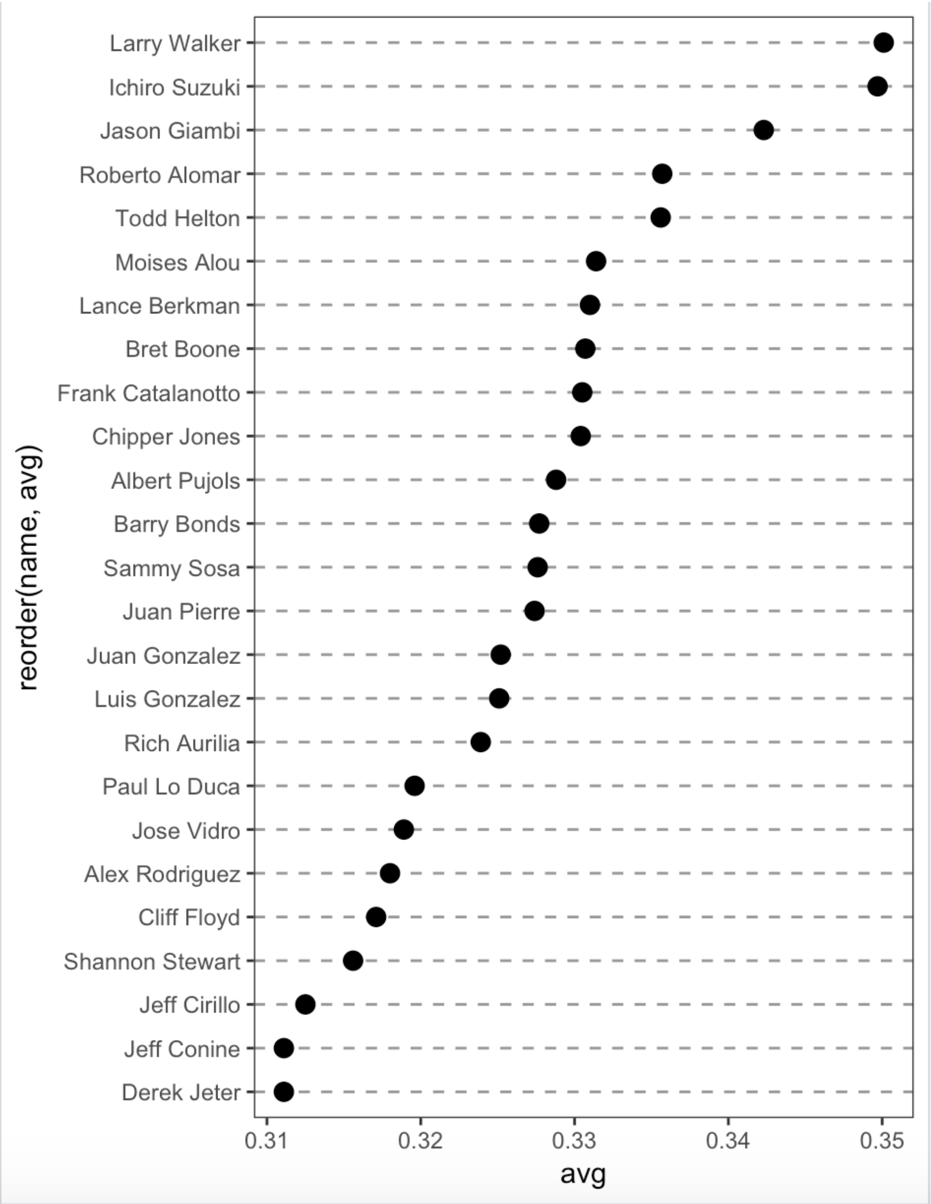
克利夫兰(Cleveland)点图

3.10 P37

```
library(gcookbook) # For the data set
tophit <- tophitters2001[1:25, ] # Take the top 25

ggplot(tophit, aes(x=avg, y=name)) + geom_point()
```



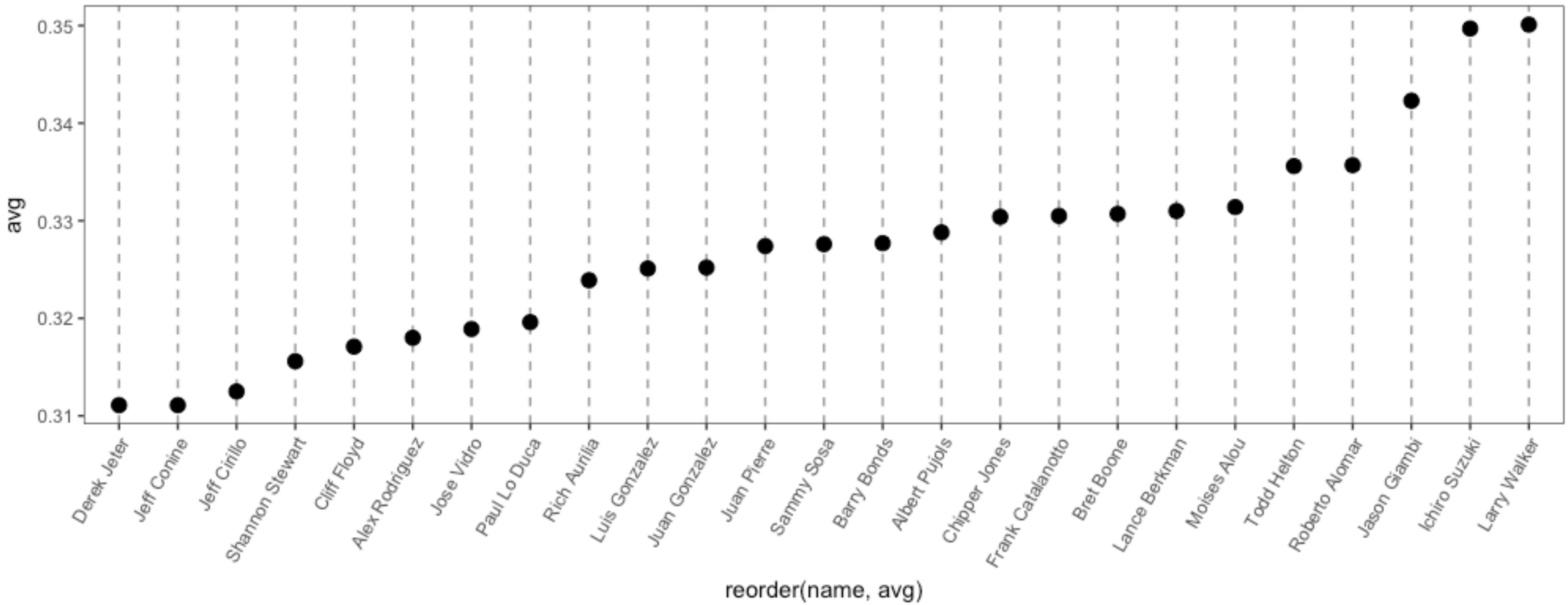
```
tophit[, c("name", "lg", "avg")]

            name lg    avg
   Larry Walker NL 0.3501
  Ichiro Suzuki AL 0.3497
   Jason Giambi AL 0.3423
...
    Jeff Conine AL 0.3111
    Derek Jeter AL 0.3111
```



```
ggplot(tophit, aes(x=avg, y=reorder(name, avg))) +
   geom_point(size=3) +                          # Use a larger dot
   theme_bw() +
   theme(panel.grid.major.x = element_blank(),
         panel.grid.minor.x = element_blank(),
         panel.grid.major.y = element_line(colour="grey60", linetype="dashed"))
```

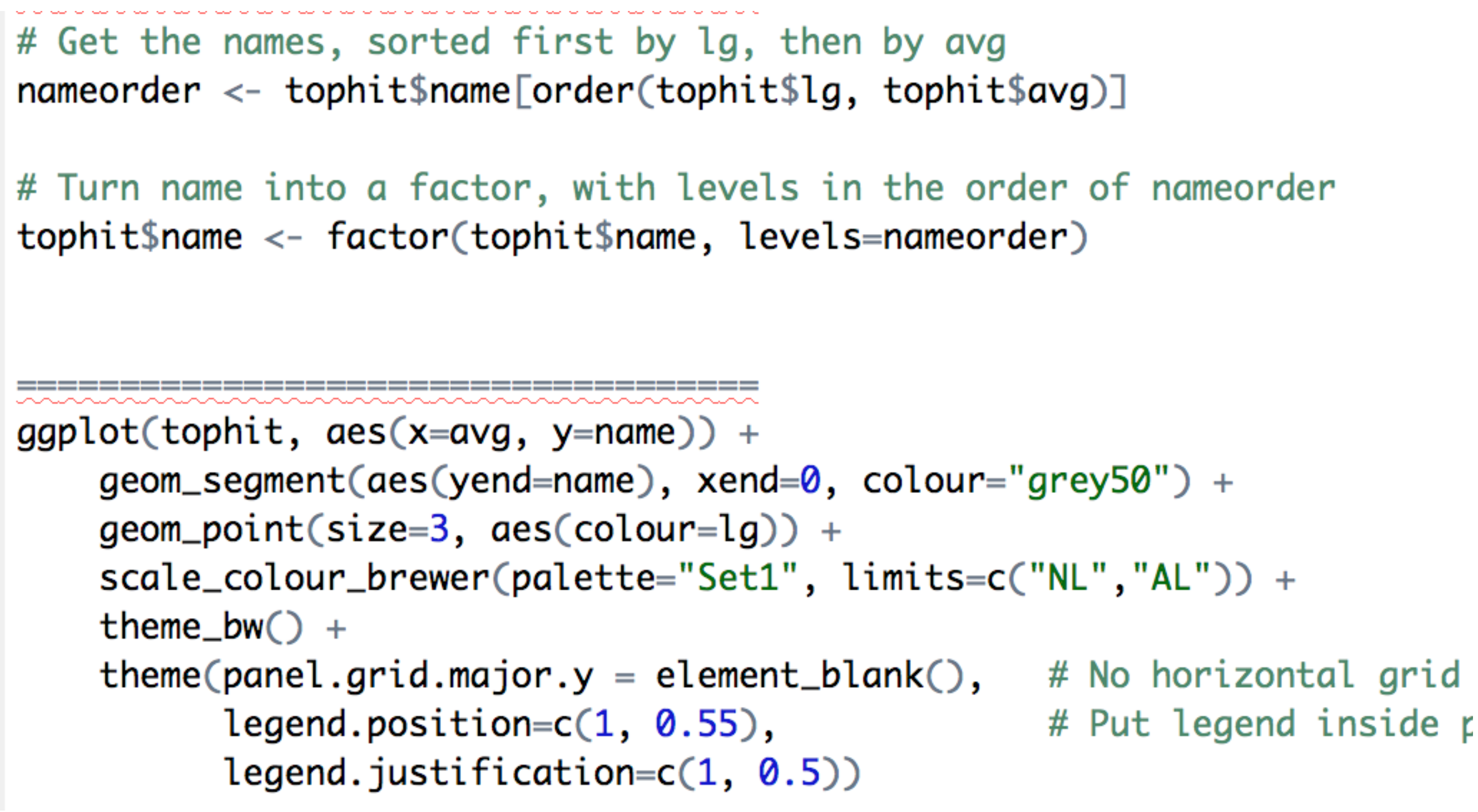
克利夫兰(Cleveland)点图

```
ggplot(tophit, aes(x=reorder(name, avg), y=avg)) +
    geom_point(size=3) +                            # Use a larger dot
    theme_bw() +
    theme(axis.text.x = element_text(angle=60, hjust=1),
          panel.grid.major.y = element_blank(),
          panel.grid.minor.y = element_blank(),
          panel.grid.major.x = element_line(colour="grey60", linetype="dashed"))
```

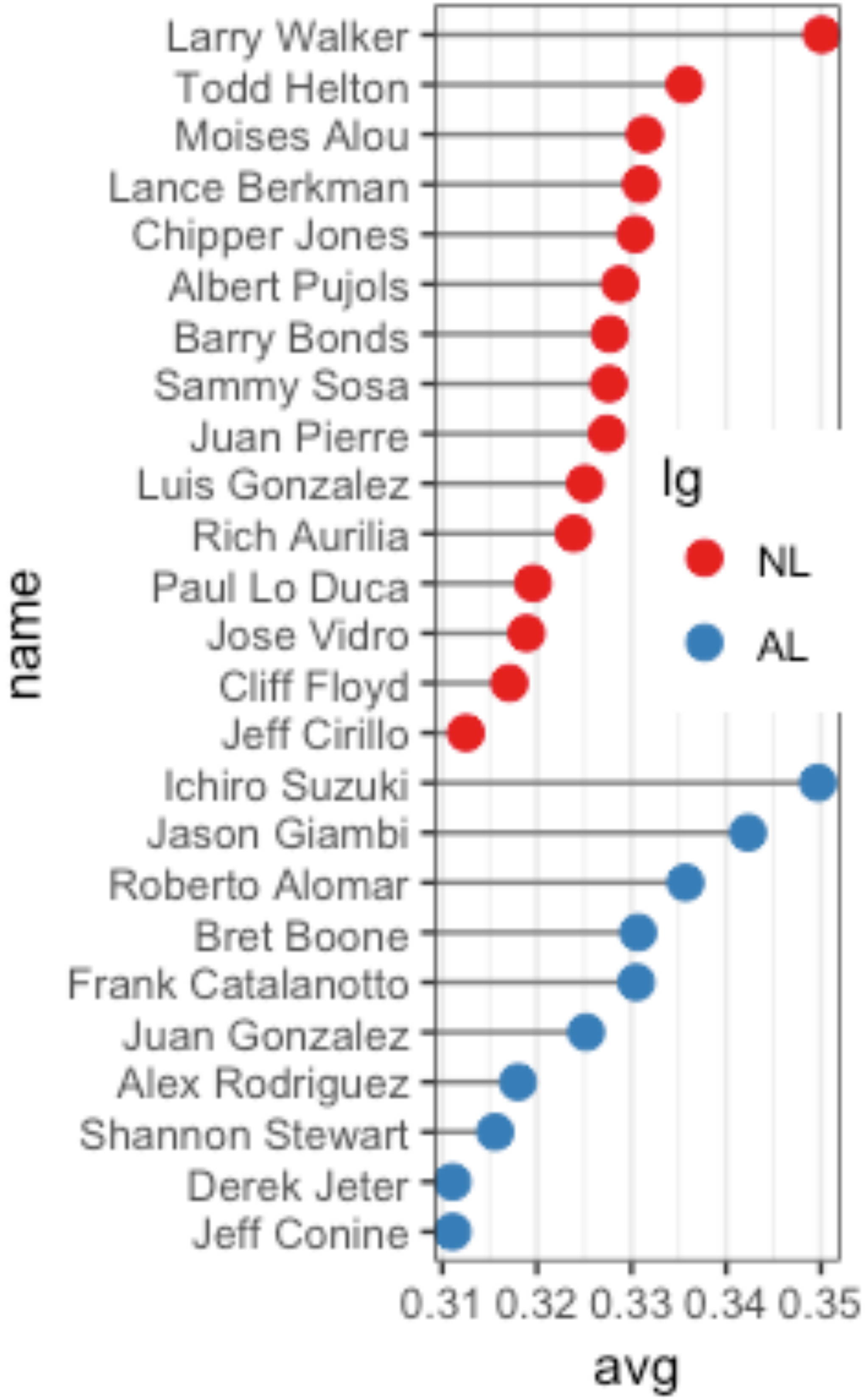


RGCook

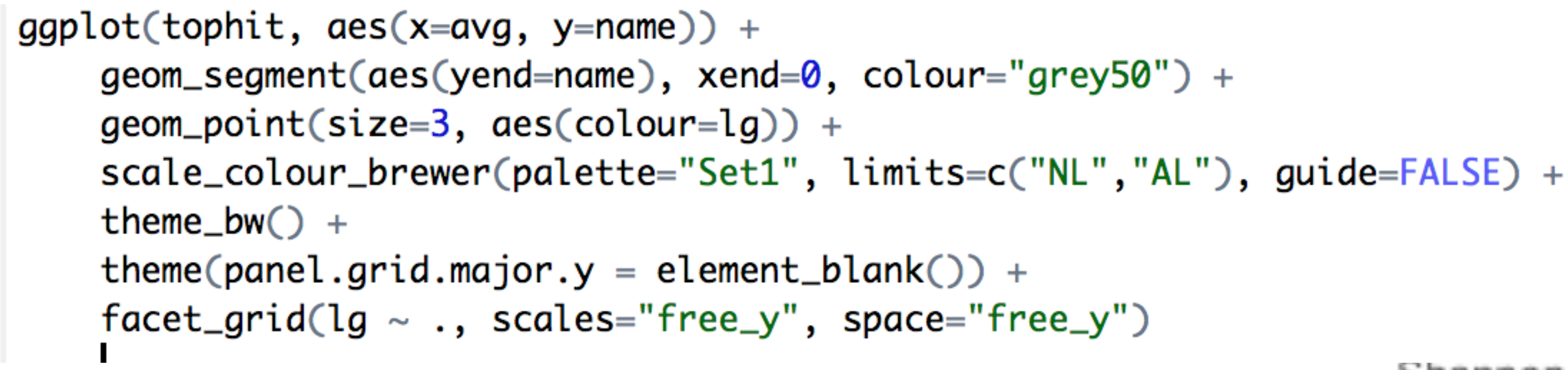
克利夫兰(Cleveland)点图



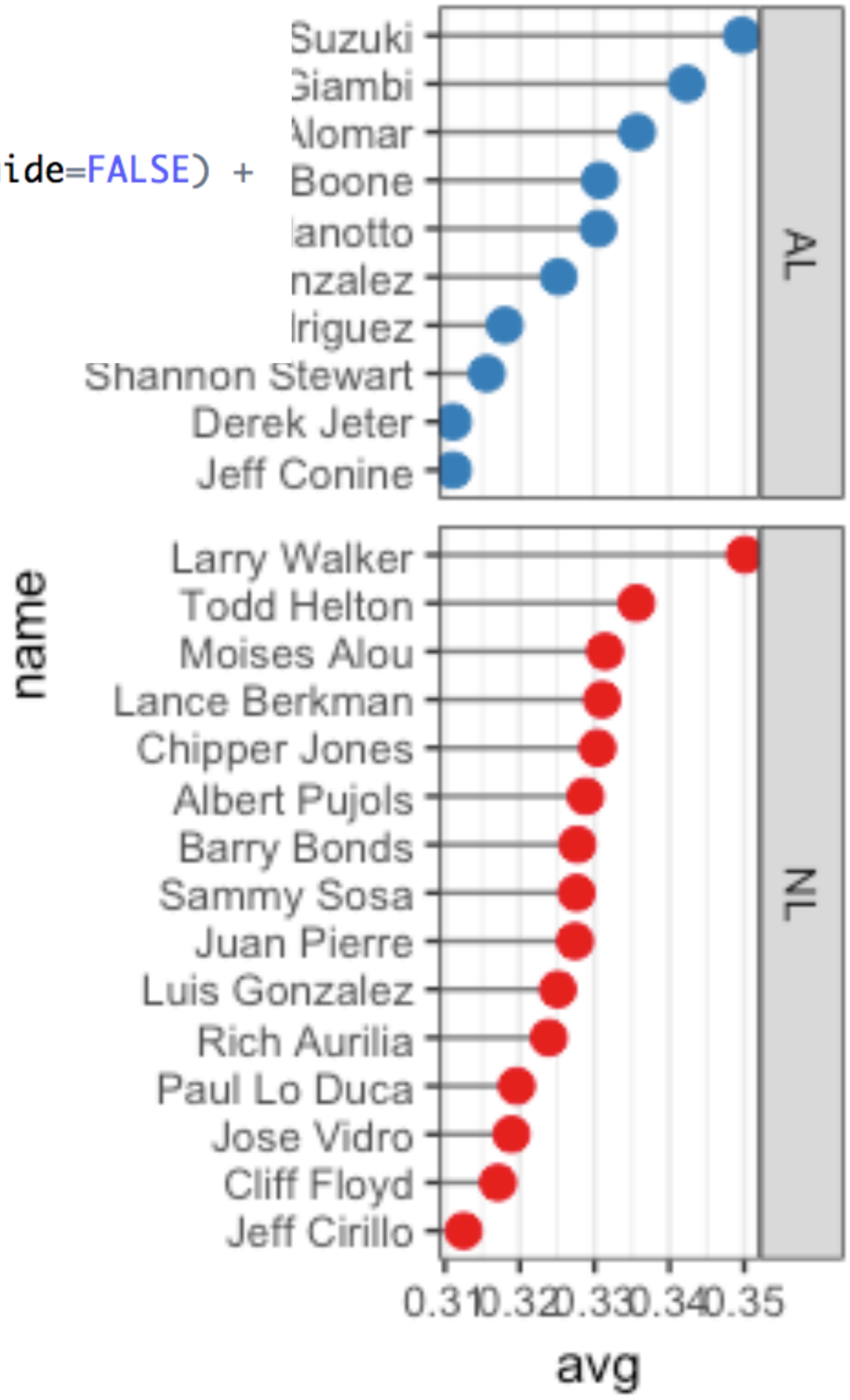
```
# Get the names, sorted first by lg, then by avg
nameorder <- tophit$name[order(tophit$lg, tophit$avg)]

# Turn name into a factor, with levels in the order of nameorder
tophit$name <- factor(tophit$name, levels=nameorder)



====================================
ggplot(tophit, aes(x=avg, y=name)) +
    geom_segment(aes(yend=name), xend=0, colour="grey50") +
    geom_point(size=3, aes(colour=lg)) +
    scale_colour_brewer(palette="Set1", limits=c("NL","AL")) +
    theme_bw() +
    theme(panel.grid.major.y = element_blank(),   # No horizontal grid
          legend.position=c(1, 0.55),             # Put legend inside p
          legend.justification=c(1, 0.5))
```



克利夫兰(Cleveland)点图

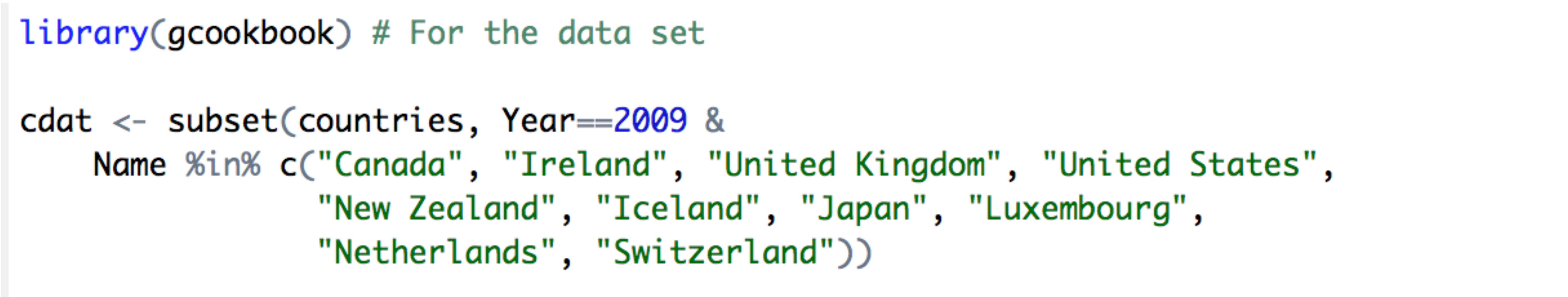


```
ggplot(tophit, aes(x=avg, y=name)) +
    geom_segment(aes(yend=name), xend=0, colour="grey50") +
    geom_point(size=3, aes(colour=lg)) +
    scale_colour_brewer(palette="Set1", limits=c("NL","AL"), guide=FALSE) +
    theme_bw() +
    theme(panel.grid.major.y = element_blank()) +
    facet_grid(lg ~ ., scales="free_y", space="free_y")
```

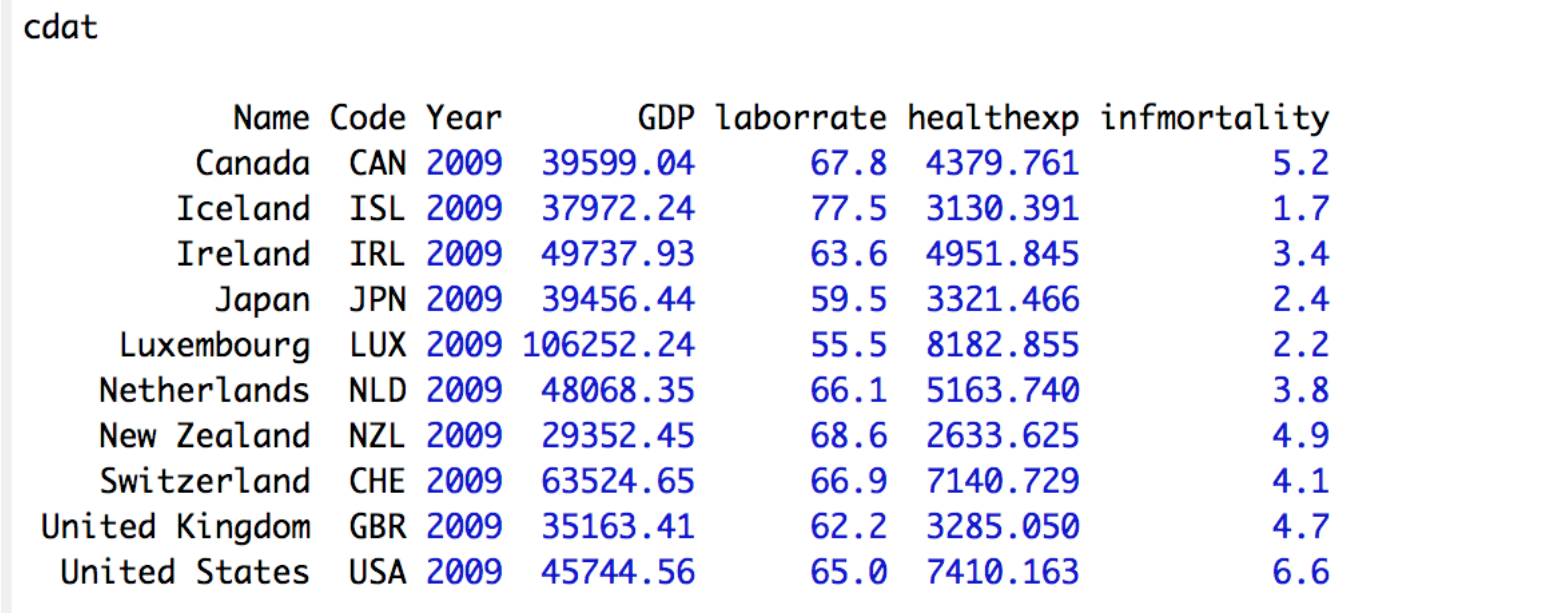


绘制气泡图

5.12 P92



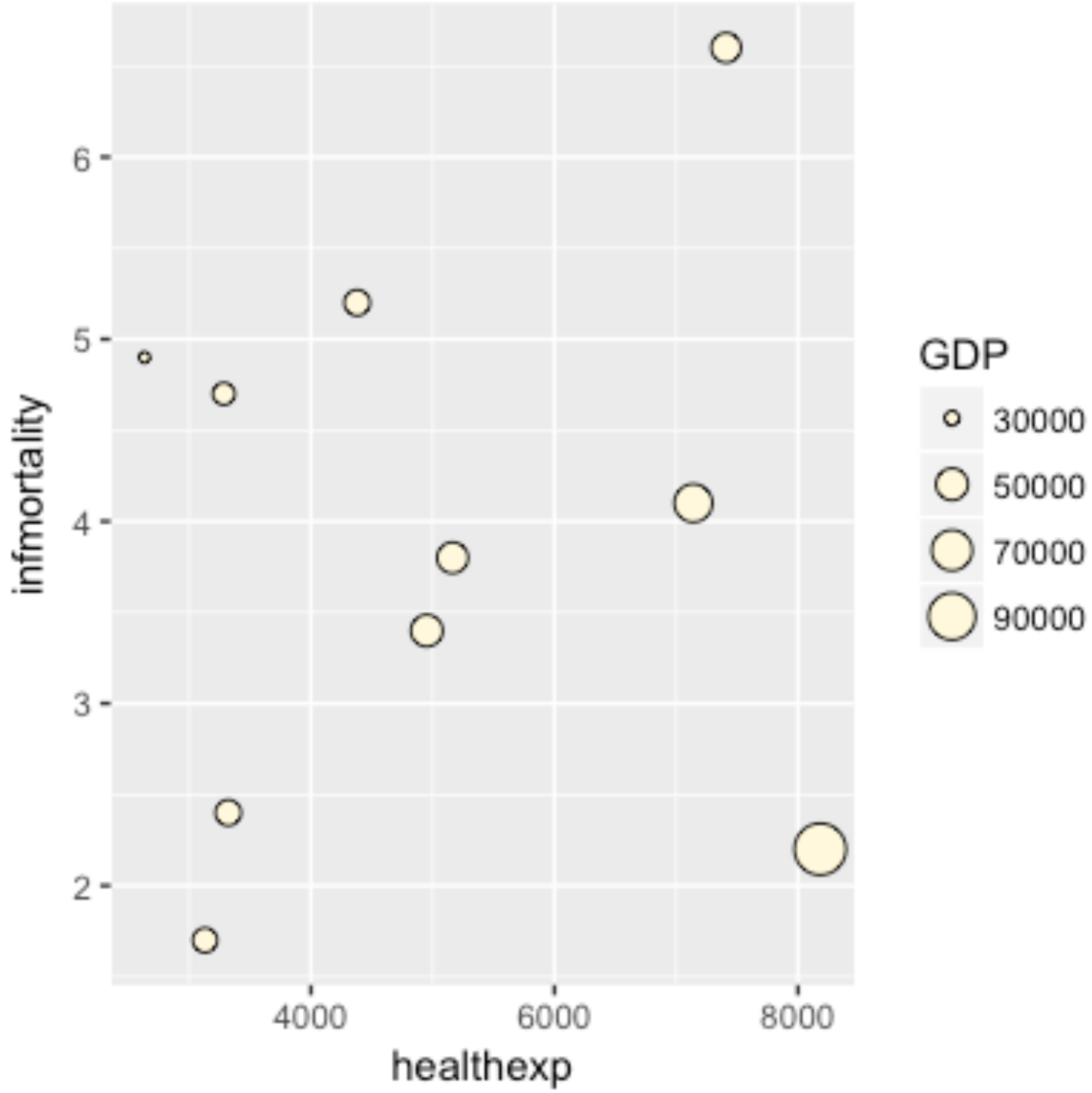
```
library(gcookbook) # For the data set

cdat <- subset(countries, Year==2009 &
    Name %in% c("Canada", "Ireland", "United Kingdom", "United States",
                "New Zealand", "Iceland", "Japan", "Luxembourg",
                "Netherlands", "Switzerland"))
```

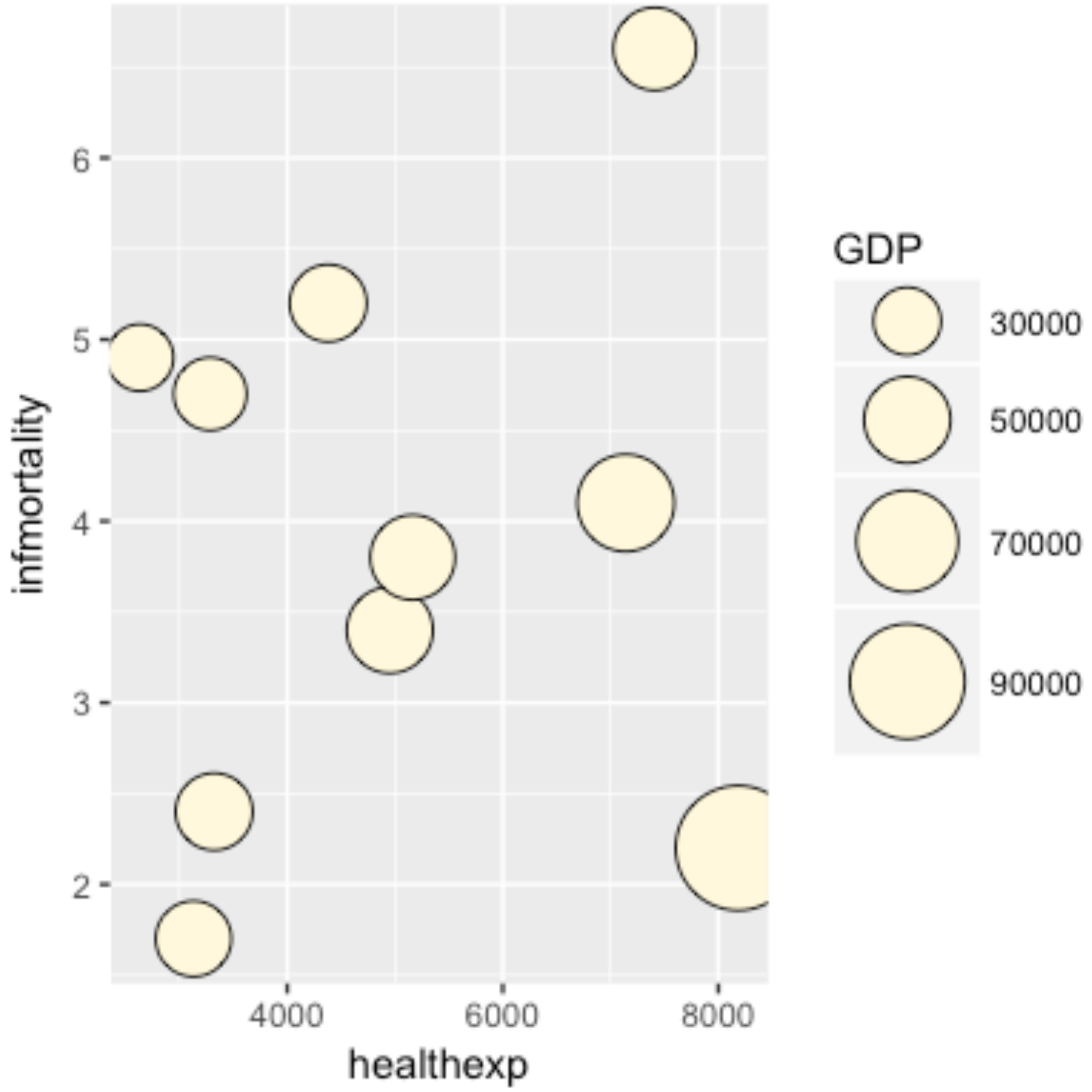


```
cdat

             Name Code Year       GDP laborrate healthexp infmortality
           Canada  CAN 2009  39599.04      67.8  4379.761          5.2
          Iceland  ISL 2009  37972.24      77.5  3130.391          1.7
          Ireland  IRL 2009  49737.93      63.6  4951.845          3.4
            Japan  JPN 2009  39456.44      59.5  3321.466          2.4
       Luxembourg  LUX 2009 106252.24      55.5  8182.855          2.2
      Netherlands  NLD 2009  48068.35      66.1  5163.740          3.8
      New Zealand  NZL 2009  29352.45      68.6  2633.625          4.9
      Switzerland  CHE 2009  63524.65      66.9  7140.729          4.1
   United Kingdom  GBR 2009  35163.41      62.2  3285.050          4.7
    United States  USA 2009  45744.56      65.0  7410.163          6.6
```

绘制气泡图

```
p <- ggplot(cdat, aes(x=healthexp, y=infmortality, size=GDP)) +
    geom_point(shape=21, colour="black", fill="cornsilk")
```





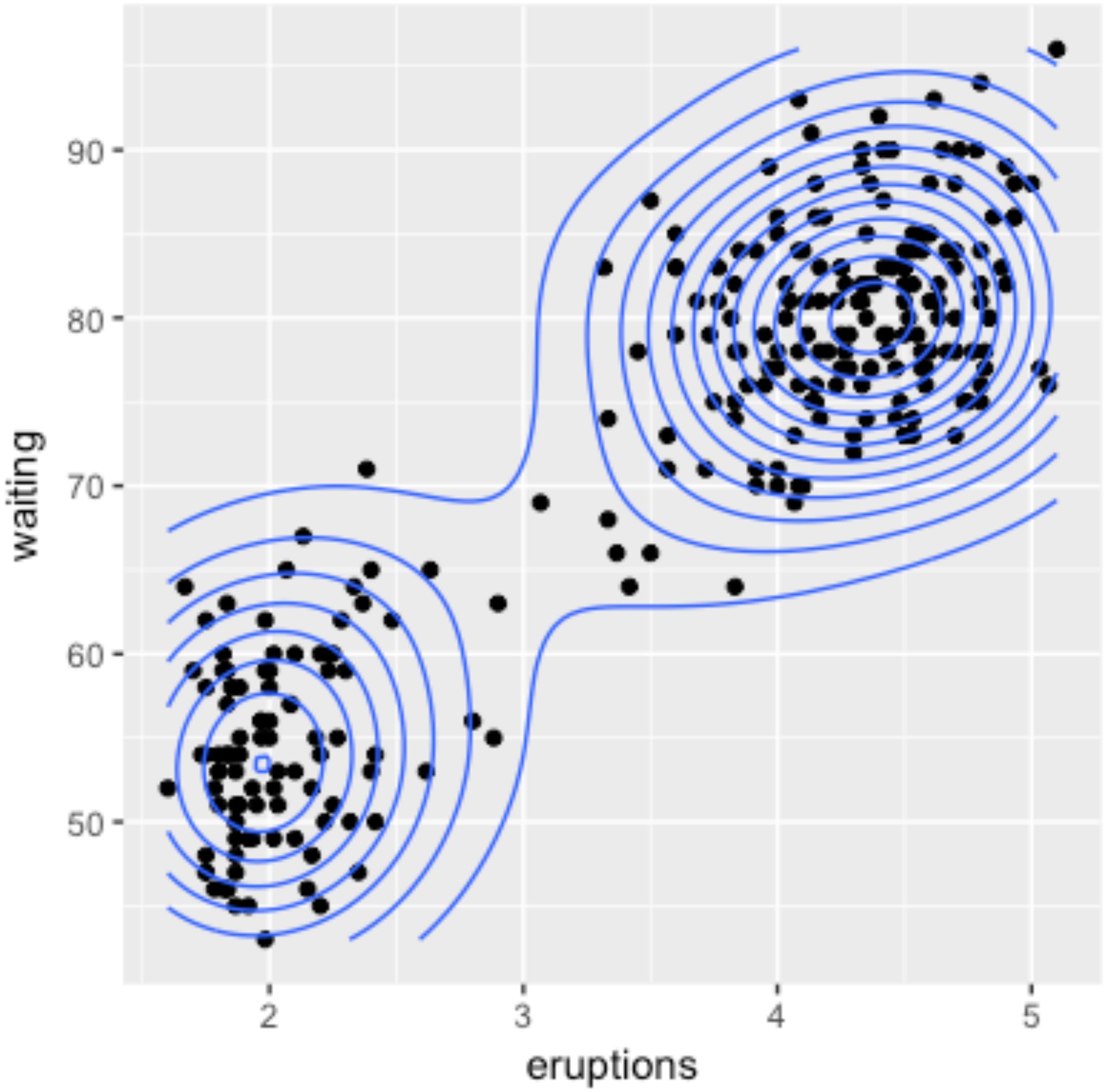
```
p + scale_size_area(max_size=15)
```

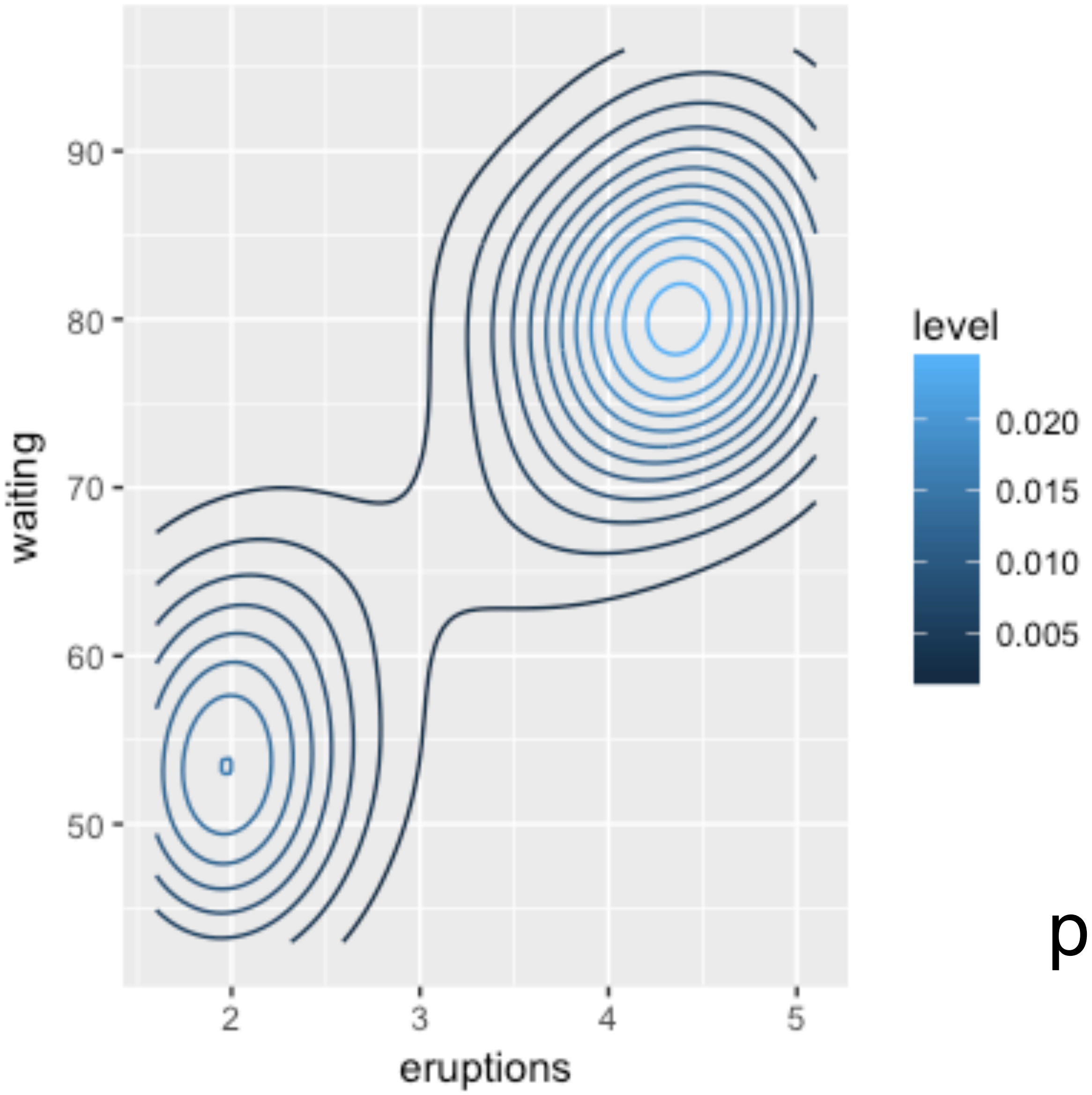
绘制二维数据的密度图

6.12 P120

```
p <- ggplot(faithful, aes(x=eruptions, y=waiting))
```

```
p + geom_point() + stat_density2d()
```





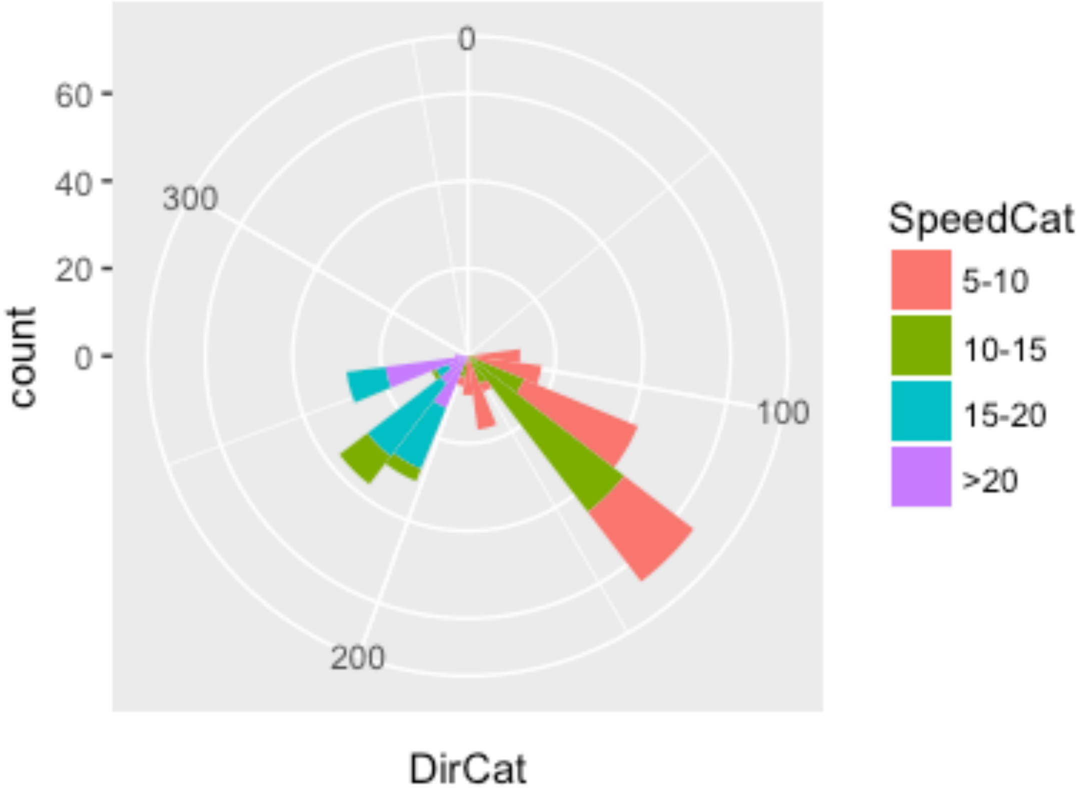
```
p + stat_density2d(aes(colour=..level..))
```

绘制环形图形

8.16 P165

library(gcookbook) # For the data set wind

```
TimeUTC  Temp WindAvg WindMax WindDir SpeedCat DirCat
      0  3.54    9.52   10.39      89    10-15     90
      5  3.52    9.10    9.90      92     5-10     90
     10  3.53    8.73    9.51      92     5-10     90
 ...
   2335  6.74   18.98   23.81     250      >20    255
   2340  6.62   17.68   22.05     252      >20    255
         6.22   18.54   23.91     259      >20    255
```

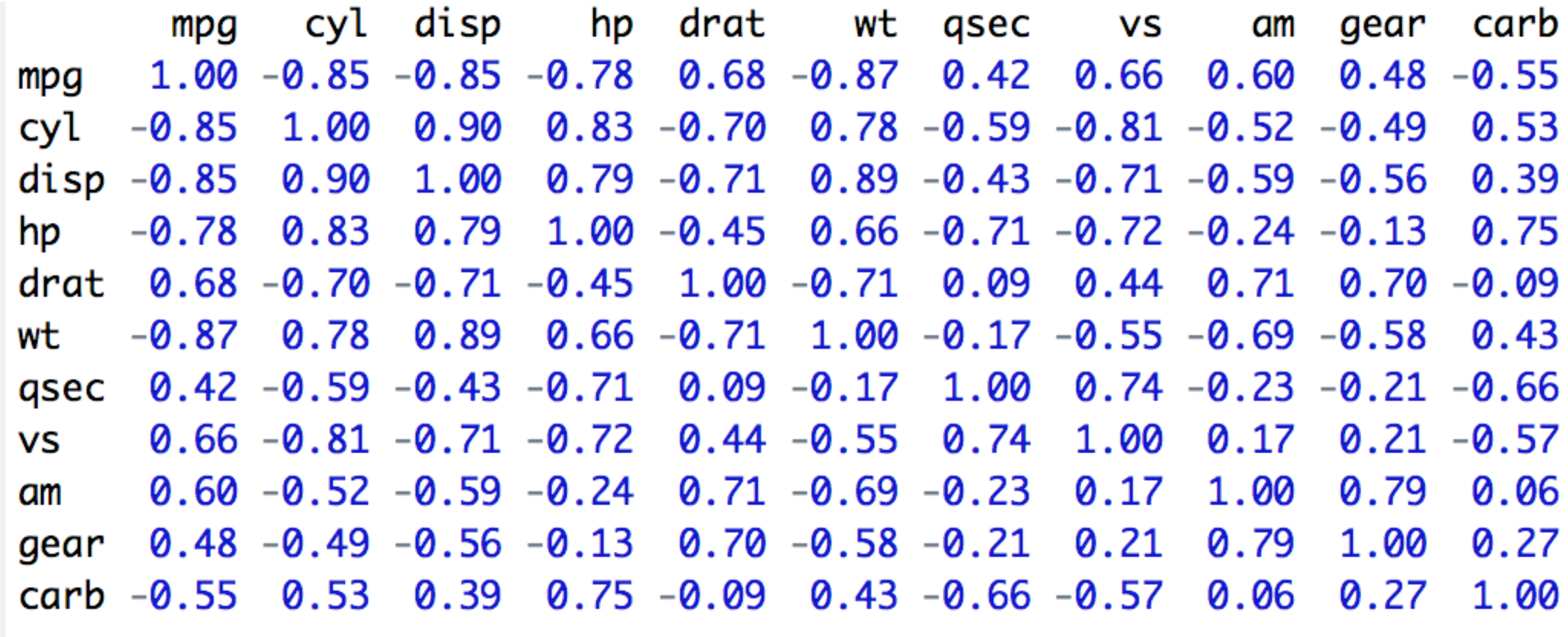


```
ggplot(wind, aes(x=DirCat, fill=SpeedCat)) +
    geom_histogram(binwidth=15, origin=-7.5) +
    coord_polar() +
    scale_x_continuous(limits=c(0,360))
```

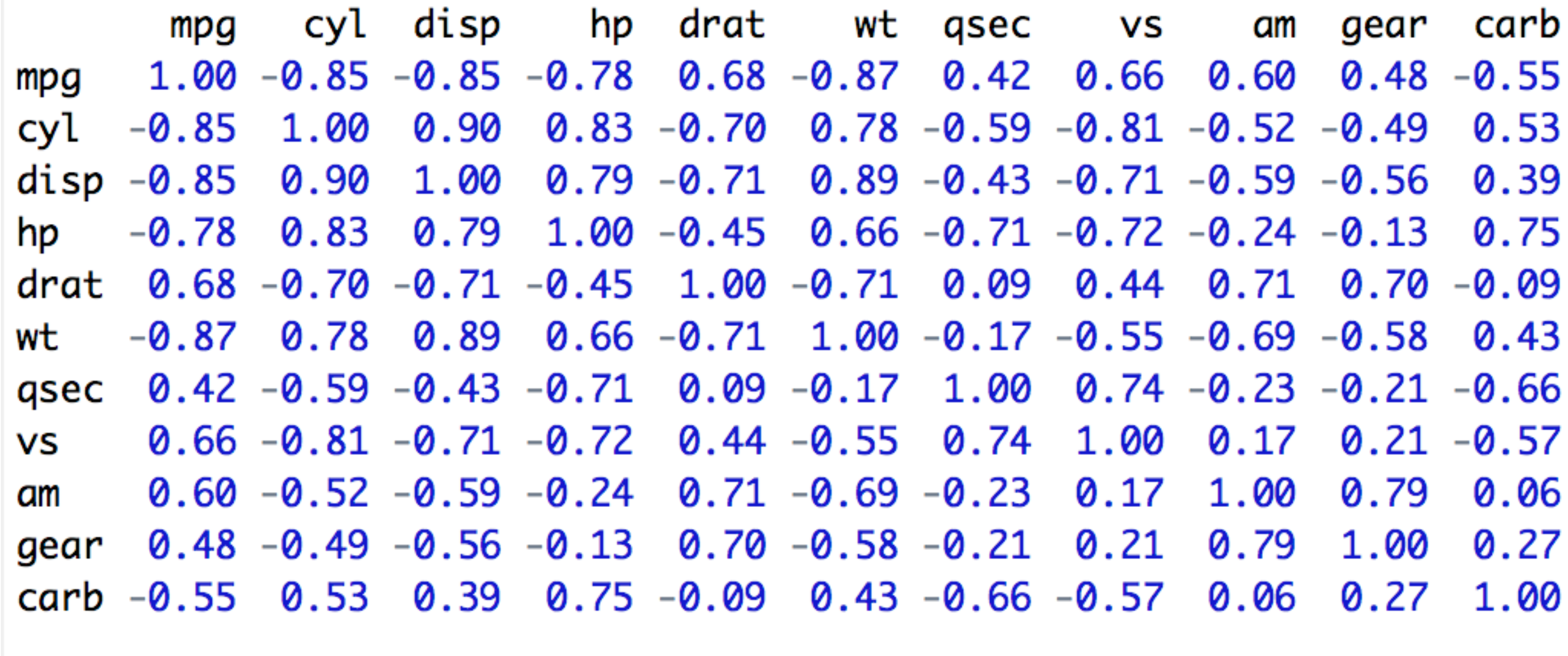
绘制相关矩阵图

13.1 P221

```
mcor <- cor(mtcars)
# Print mcor and round to 2 digits
round(mcor, digits=2)
```



```
       mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

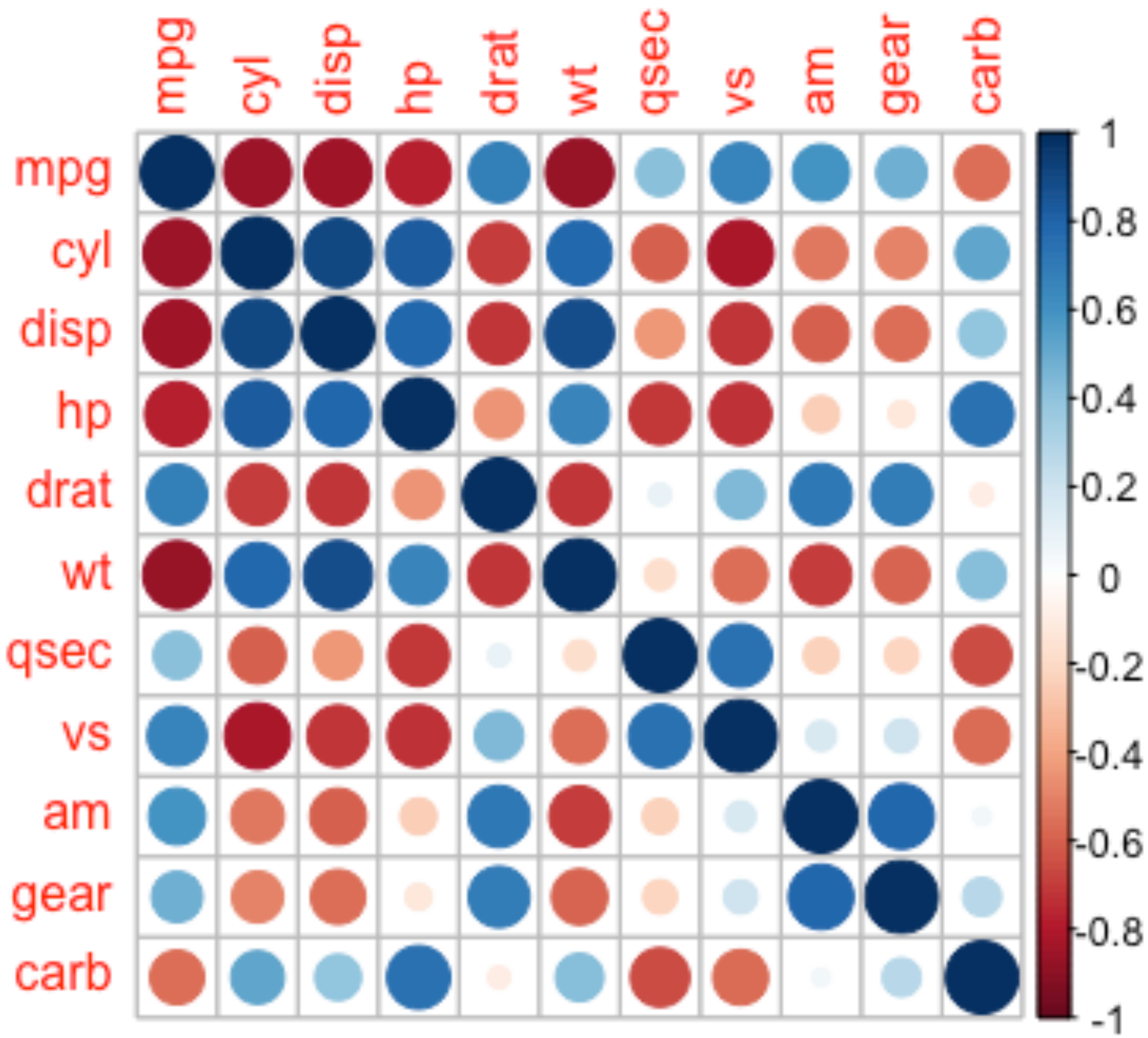


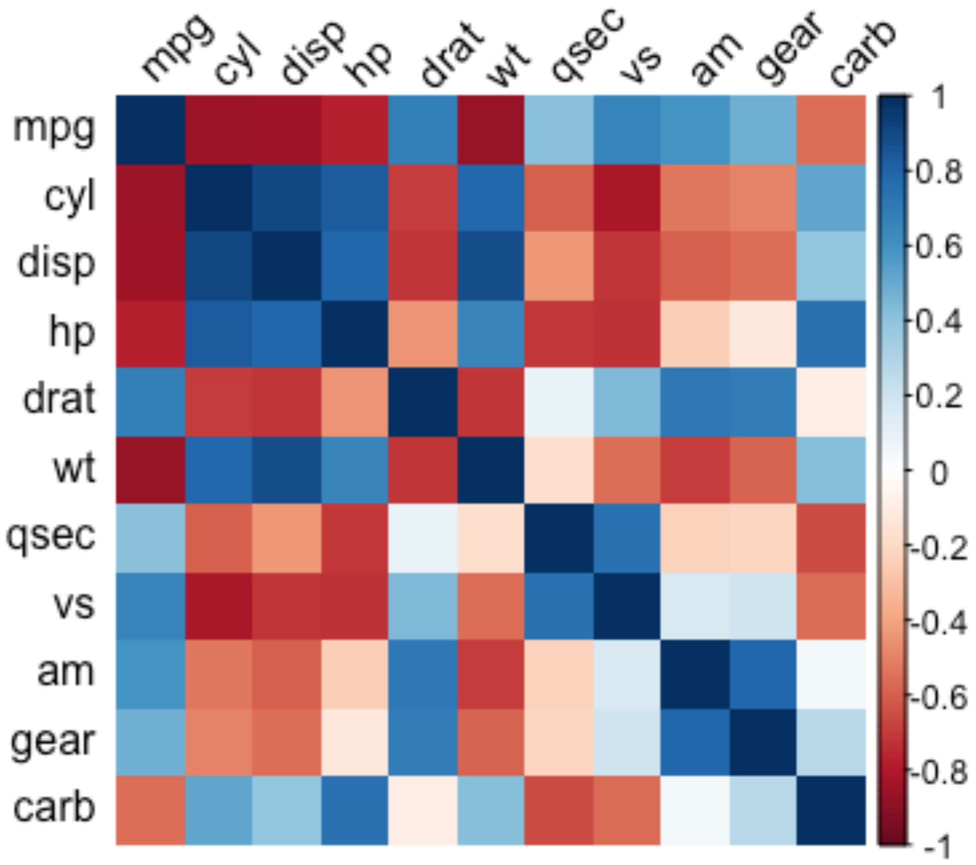
```
       mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

RGCook

绘制相关矩阵图

corrplot(mcor)

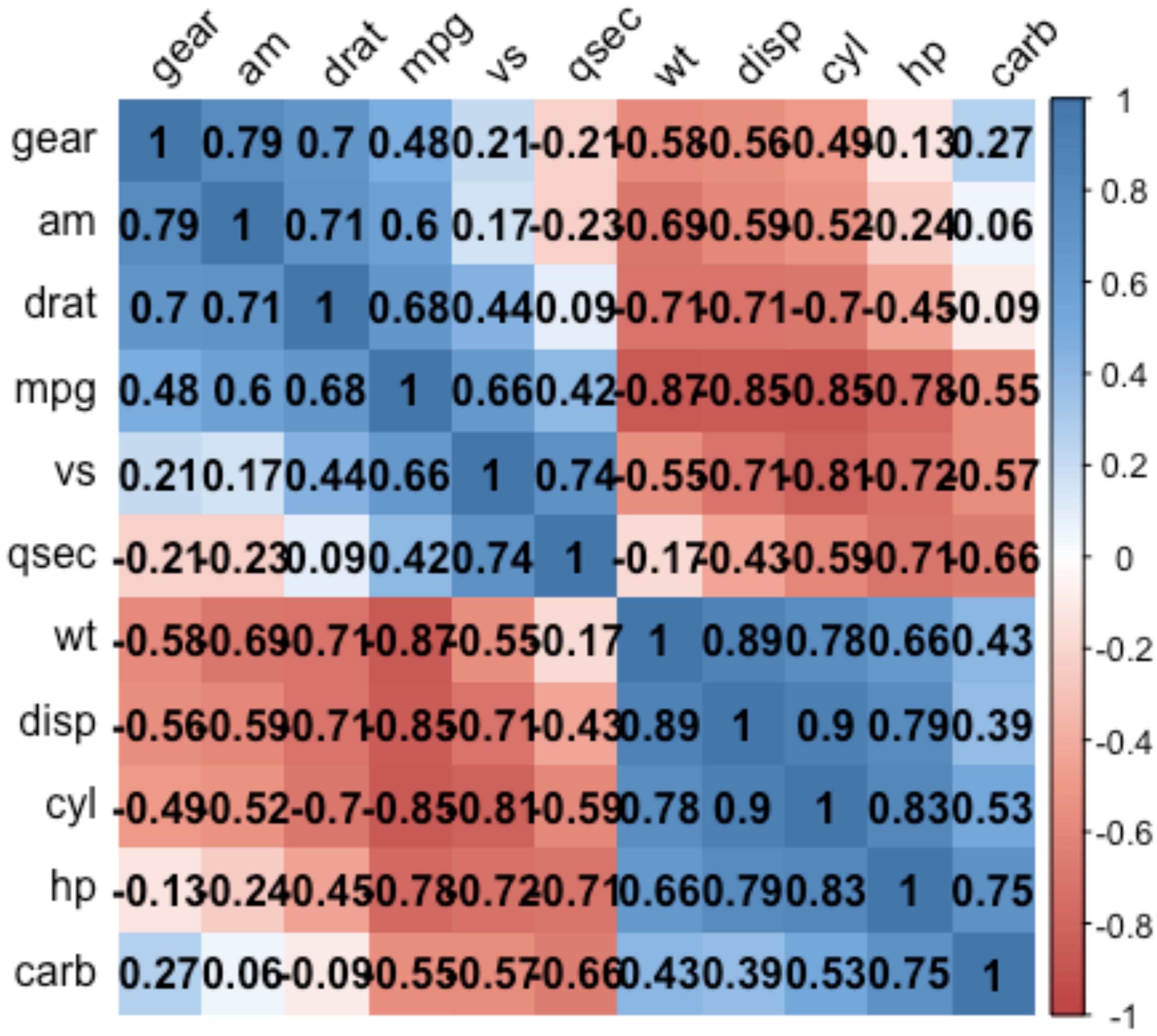




```
corrplot(mcor,
method="shade",
shade.col=NA,
tl.col="black",
tl.srt=45)
```

绘制相关矩阵图

```
col <-
colorRampPalette(c("#BB
4444", "#EE9988",
"#FFFFFF", "#77AADD",
"#4477AA"))

corrplot(mcor,
method="shade",
shade.col=NA,
tl.col="black",
tl.srt=45,
 col=col(200),
addCoef.col="black",
addcolorlabel="no",
order="AOE")
```

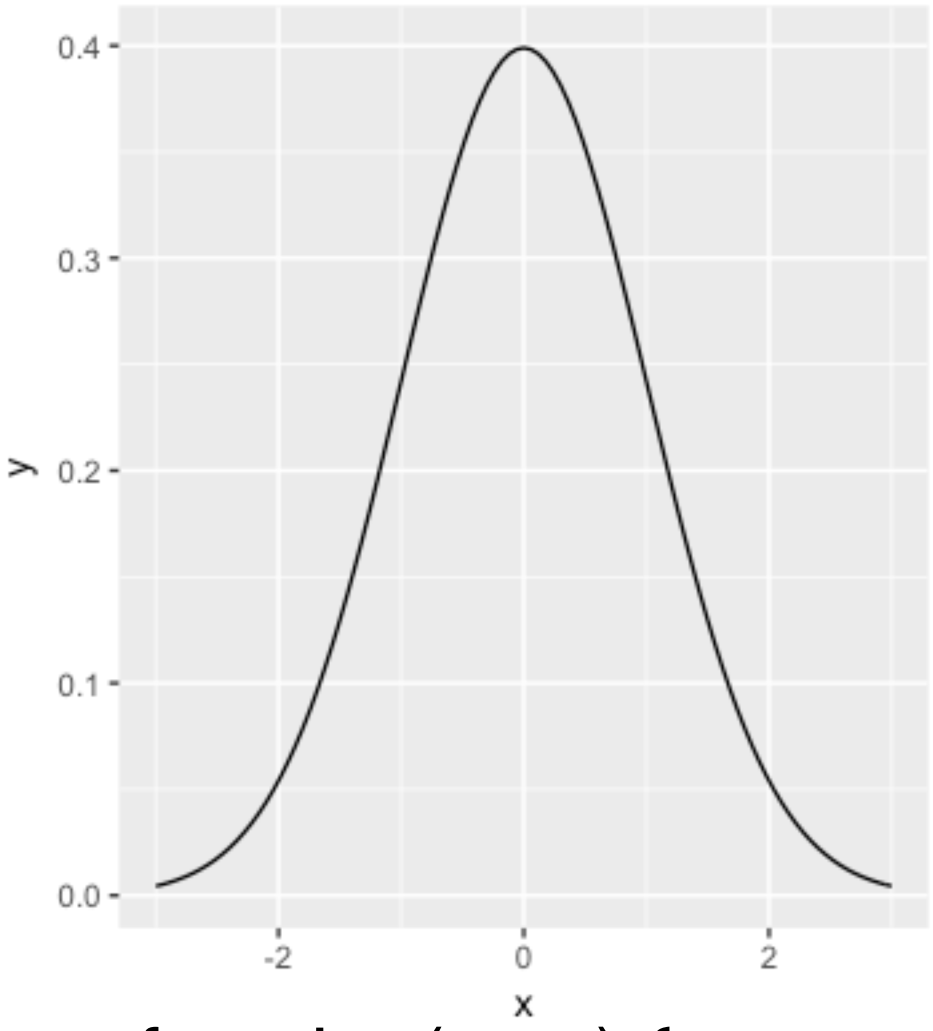


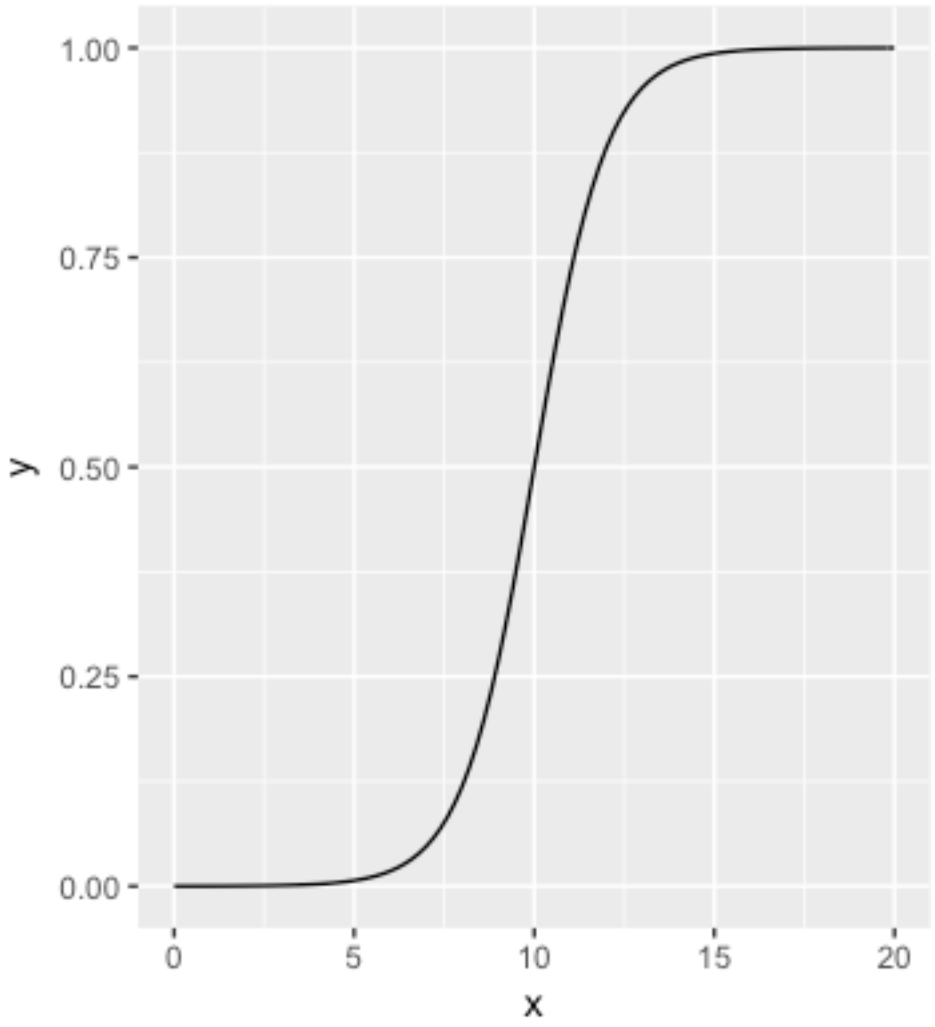
	gear	am	drat	mpg	vs	qsec	wt	disp	cyl	hp	carb
gear	1	0.79	0.7	0.48	0.21	-0.21	-0.58	-0.56	-0.49	-0.13	0.27
am	0.79	1	0.71	0.6	0.17	-0.23	-0.69	-0.59	-0.52	-0.24	0.06
drat	0.7	0.71	1	0.68	0.44	0.09	-0.71	-0.71	-0.7	-0.45	-0.09
mpg	0.48	0.6	0.68	1	0.66	0.42	-0.87	-0.85	-0.85	-0.78	-0.55
vs	0.21	0.17	0.44	0.66	1	0.74	-0.55	-0.71	-0.81	-0.72	-0.57
qsec	-0.21	-0.23	0.09	0.42	0.74	1	-0.17	-0.43	-0.59	-0.71	-0.66
wt	-0.58	-0.69	-0.71	-0.87	-0.55	-0.17	1	0.89	0.78	0.66	0.43
disp	-0.56	-0.59	-0.71	-0.85	-0.71	-0.43	0.89	1	0.9	0.79	0.39
cyl	-0.49	-0.52	-0.7	-0.85	-0.81	-0.59	0.78	0.9	1	0.83	0.53
hp	-0.13	-0.24	-0.45	-0.78	-0.72	-0.71	0.66	0.79	0.83	1	0.75
carb	0.27	0.06	-0.09	-0.55	-0.57	-0.66	0.43	0.39	0.53	0.75	1

绘制函数曲线

13.2 P224

```
p <- ggplot(data.frame(x=c(-3,3)), aes(x=x))
p + stat_function(fun = dnorm)
```





```
myfun <- function(xvar) {
   1/(1 + exp(-xvar + 10))
}
ggplot(data.frame(x=c(0, 20)), aes(x=x)) + stat_function(fun=myfun)
```

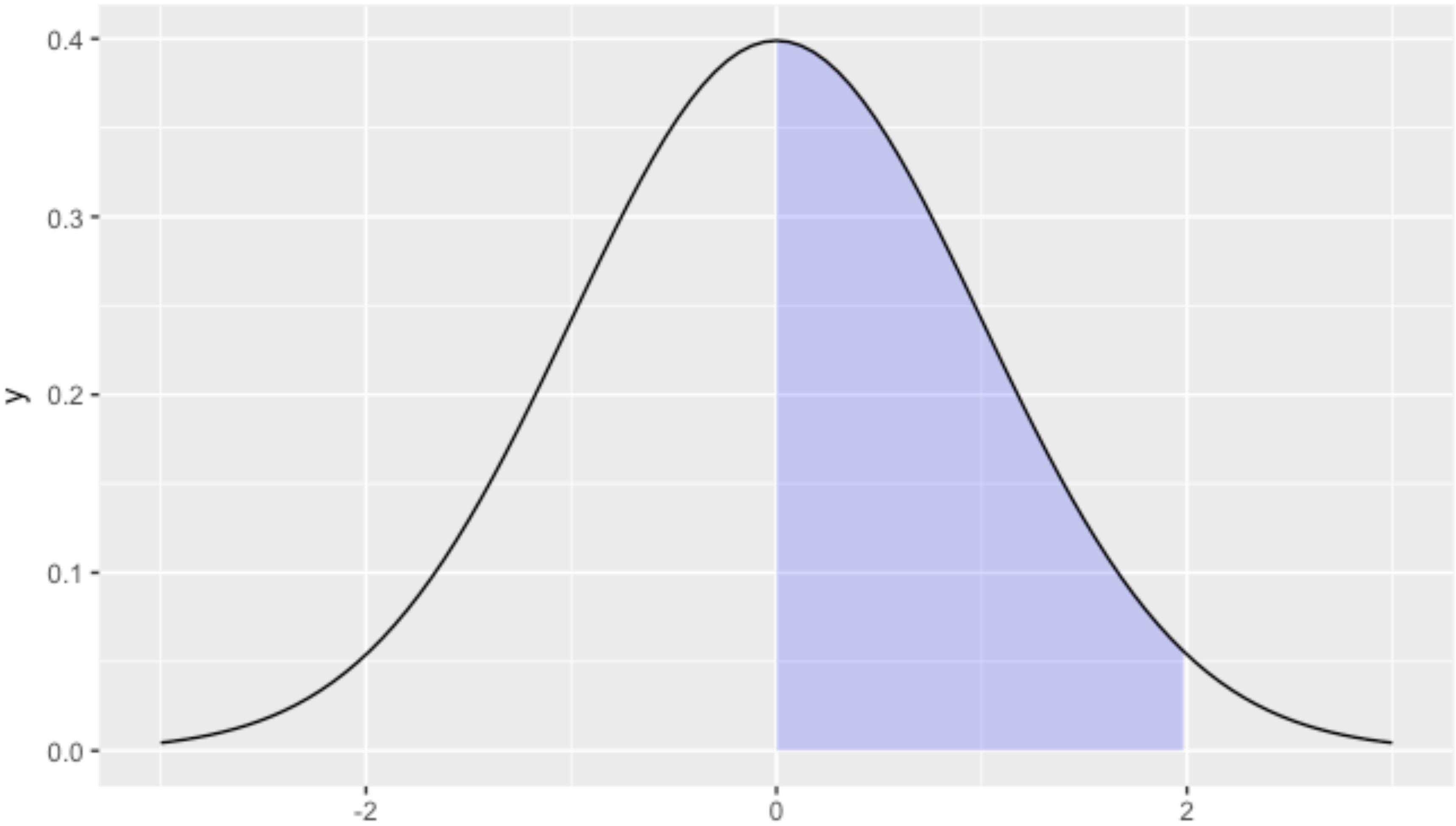
RGCook

绘制函数曲线

13.2 P224

```
# Return dnorm(x) for 0 < x < 2, and NA for all other x
dnorm_limit <- function(x) {
   y <- dnorm(x)
   y[x < 0  |  x > 2] <- NA
    return(y)
}
```

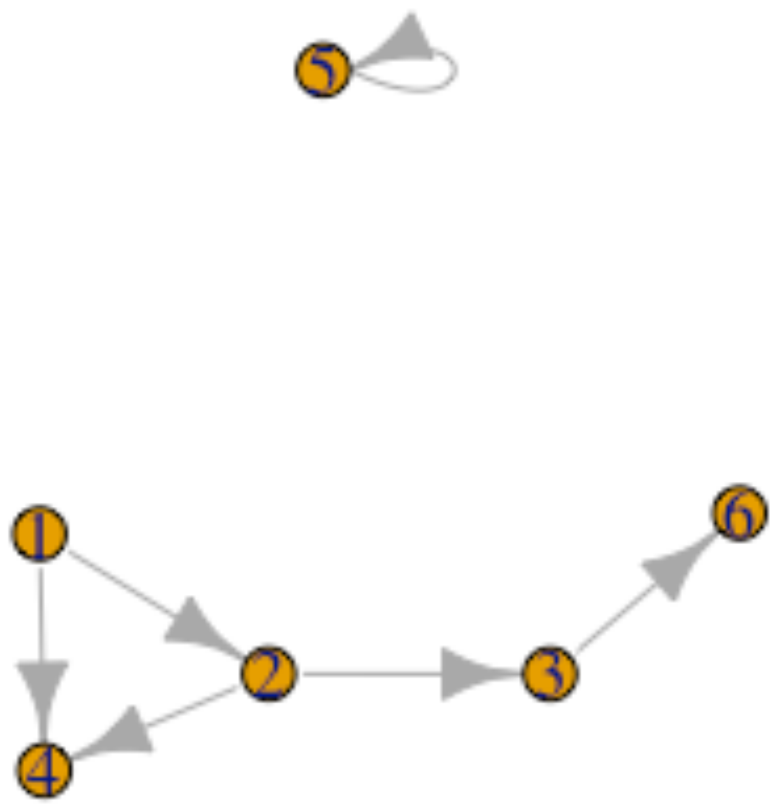
```
# ggplot() with dummy data
p <- ggplot(data.frame(x=c(-3, 3)), aes(x=x))

p + stat_function(fun=dnorm_limit, geom="area", fill="blue", alpha=0.2) +

stat_function(fun=dnorm)
```



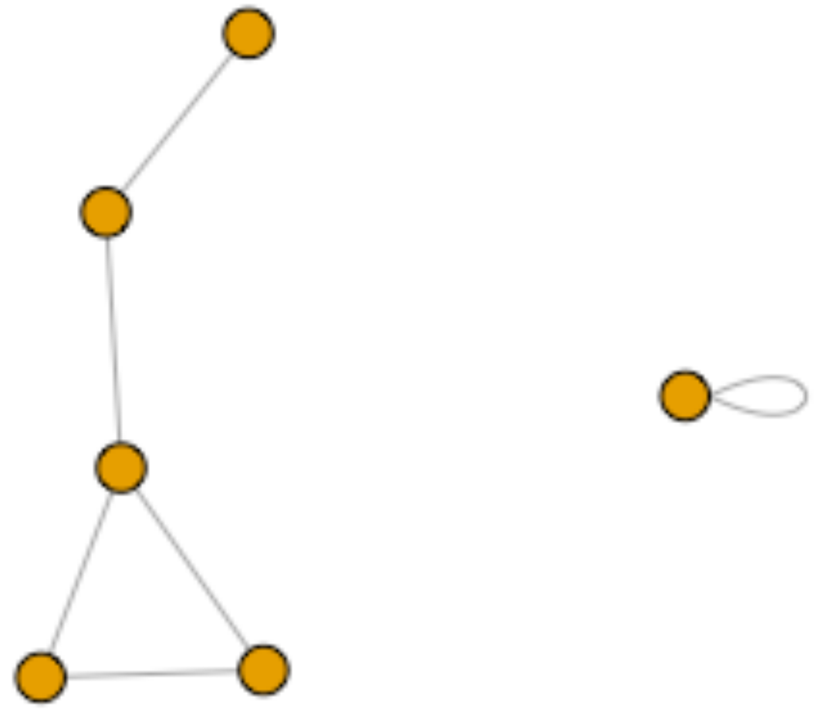
绘制网络图

13.4 P227

```
library(igraph)
# Specify edges for a directed
graph
gd <- graph(c(1,2, 2,3, 2,4, 1,4,
5,5, 3,6))
plot(gd)
```

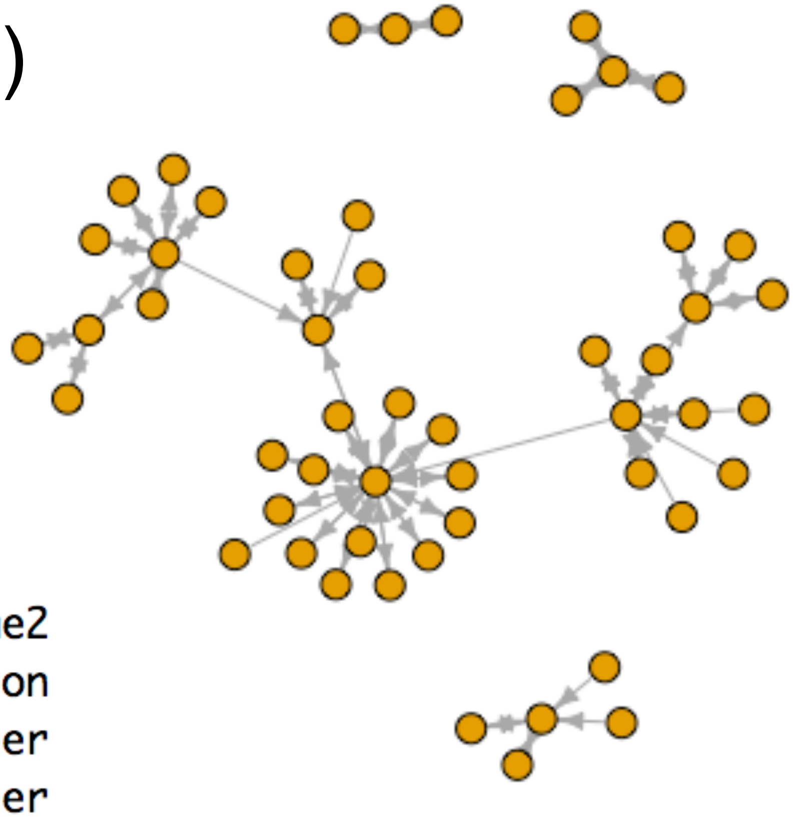


```
# For an undirected
graph
gu <- graph(c(1,2, 2,3,
2,4, 1,4, 5,5, 3,6),
directed=FALSE)
# No labels
plot(gu, vertex.label=NA)
```

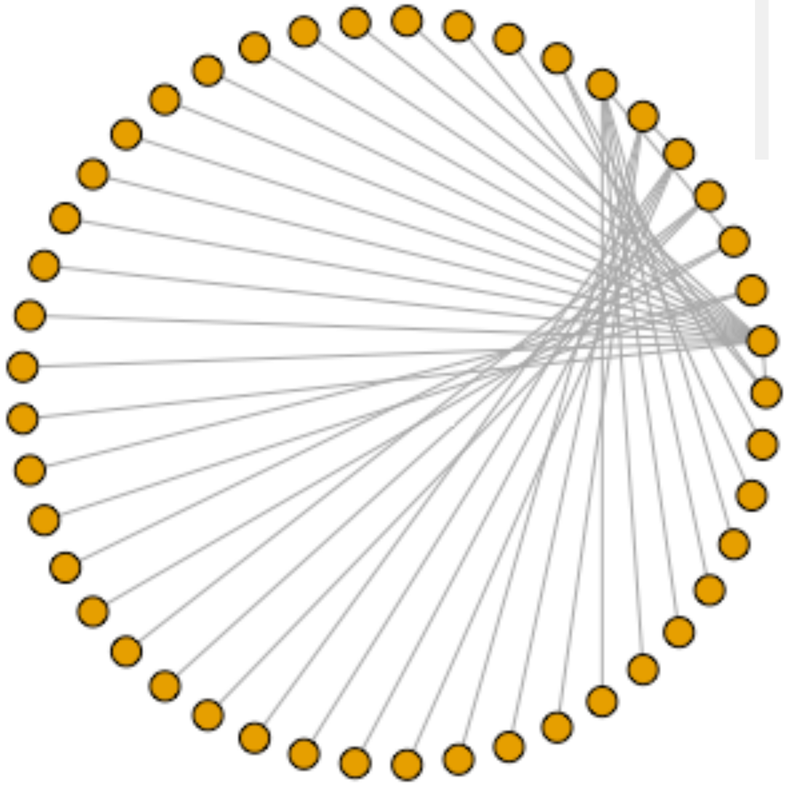


绘制网络图

```
library(gcookbook)
g <- graph.data.frame(madmen2, directed=TRUE)
par(mar=c(0,0,0,0))

plot(g, layout=layout.fruchterman.reingold,
vertex.size=8, edge.arrow.size=0.5,
    vertex.label=NA)
```



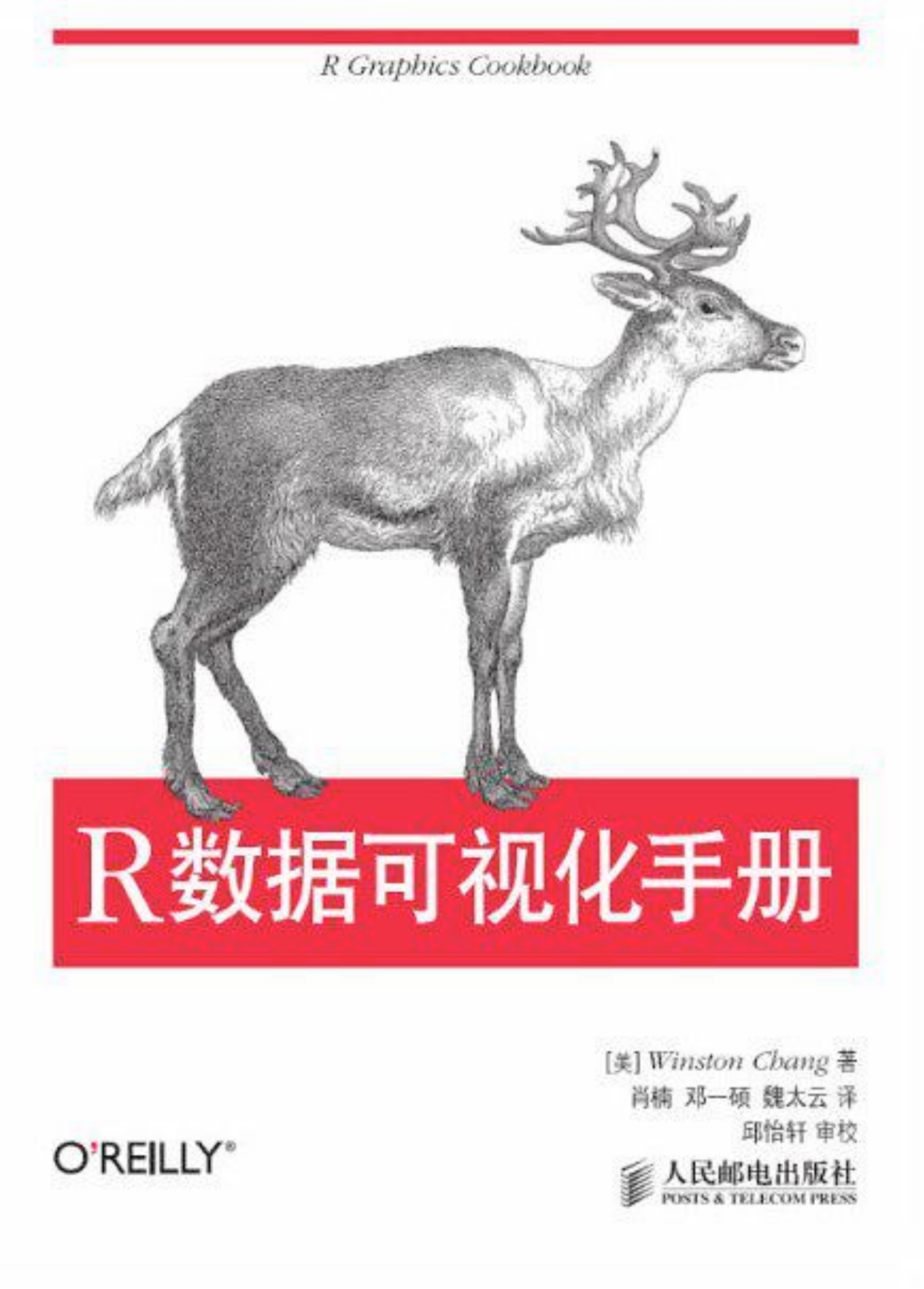
```
       Name1               Name2
 Abe Drexler         Peggy Olson
     Allison          Don Draper
 Arthur Case        Betty Draper
...
```



```
g <- graph.data.frame(madmen,
directed=FALSE)
par(mar=c(0,0,0,0))
# Remove unnecessary margins
plot(g, layout=layout.circle, vertex.size=8,
vertex.label=NA)
```

练习

练习 - 0029

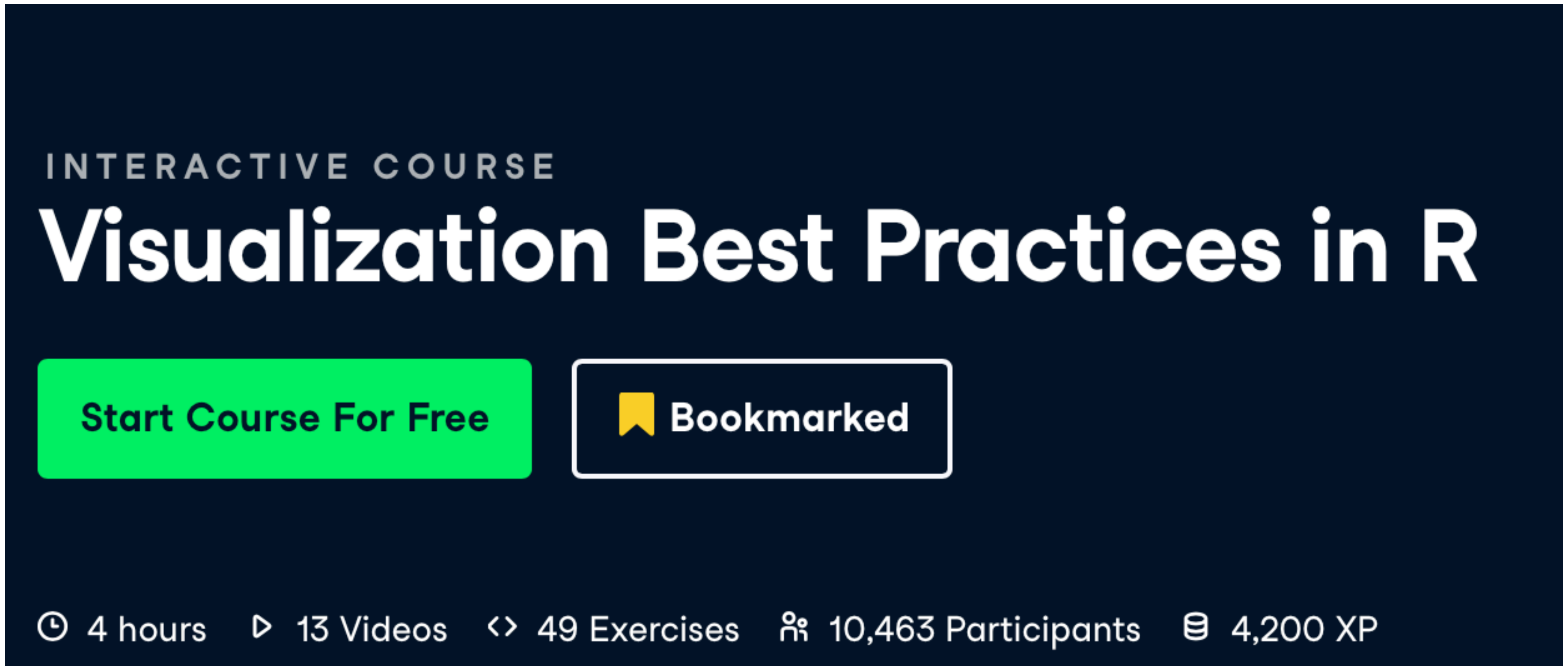


阅读所有章节，运行所有代码

注解
坐标系
图例
分面
颜色
输出
… …

小提琴图
热图
三维图
谱系图
向量图
马赛克图
… …

练习 - 0030



INTERACTIVE COURSE

Visualization Best Practices in R

4 hours · 13 Videos · 49 Exercises · 10,463 Participants · 4,200 XP

提交方式和上节课一样!

https://www.datacamp.com/courses

谢谢！

孙惠平

sunhp@ss.pku.edu.cn