Data Analysis Tools and Practice(Using R)

2018.05.17

图形复习

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10

先用电脑完成
40分钟 然后誊抄纸上

- I、某银行三个分行随机抽取7个VIP客户的账户余额数据如下表:

  ✳ 检验样本在A因素的三个水平下，均值是否存在差异，输出结果并作出箱线图验证你的结论。

分行	账户余额（万元）						
A_1	103	101	98	110	105	100	106
A_2	113	107	108	116	114	110	115
A_3	82	92	84	86	84	90	88

- 2、城市道路交通管理部门为了研究不同路段、不同时段的拥堵情况，分别在三个路段和高峰期、非高峰期进行试验，每一个水平下测量5次，共获得30个行车时间的数据，如下表所示：

  ✳根据表中信息构造数据集traffic，并作出两个因子下各水平的箱线图

  ✳绘制两个因素的交互效应图

  ✳假设两个因素有相互作用进行双因素方差分析并输出结果

课堂测试10

先用电脑完成
*40*分钟　然后誊抄纸上

	路段 I	路段 II	路段III
高峰期	25 24 27 25 25	19 20 23 22 21	29 28 31 28 30
非高峰期	20 17 22 21 17	18 17 13 16 12	22 18 24 21 22

- 3、为研究三种肥料对苹果树产量的影响有无差异，每种肥料各施用于8棵树，测得每棵树的初始产量和增量，利用协方差分析三种肥料的效果是否相同

	肥料	观察值
初始产量	*A*	15 13 11 12 12 16 14 17
	B	17 16 18 18 21 22 19 18
	C	22 24 20 23 25 27 30 32
产量增量	*A*	85 83 65 76 80 91 84 90
	B	97 90 100 95 103 106 99 94
	C	89 91 83 95 100 102 105 110

上次课程内容回顾

- 方差分析定义和概念
- aov()函数
- 单因素方差分析
- 多重比较
- 单因素协方差分析
- 双因素方差分析
- 重复测量方差分析

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

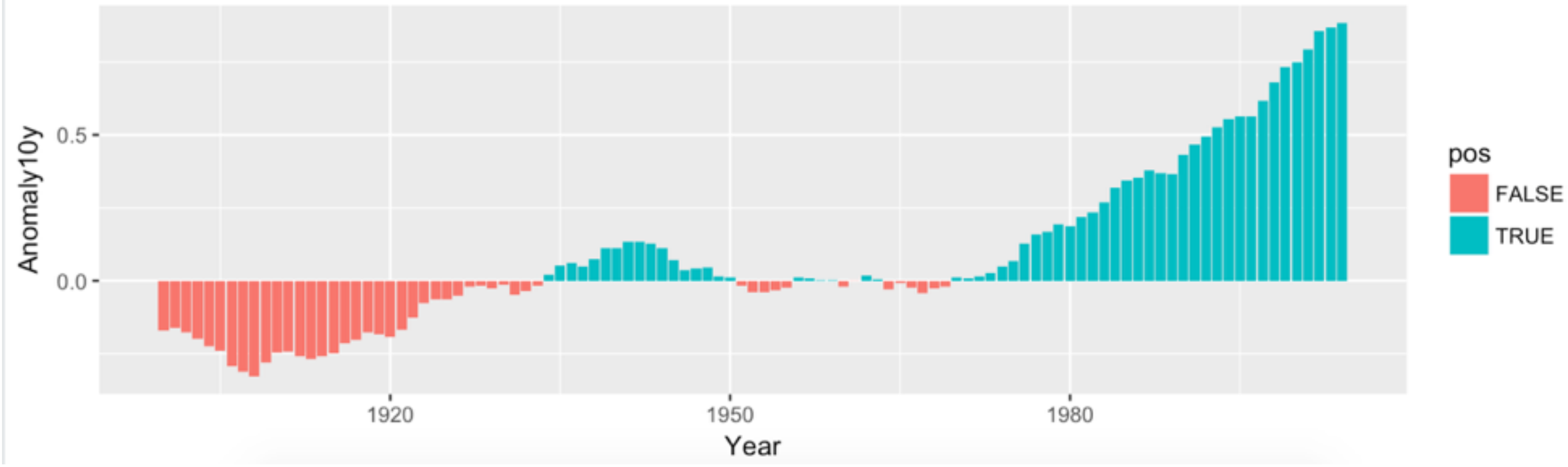

```
library(gcookbook) # For the data set
library(ggplot2)

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

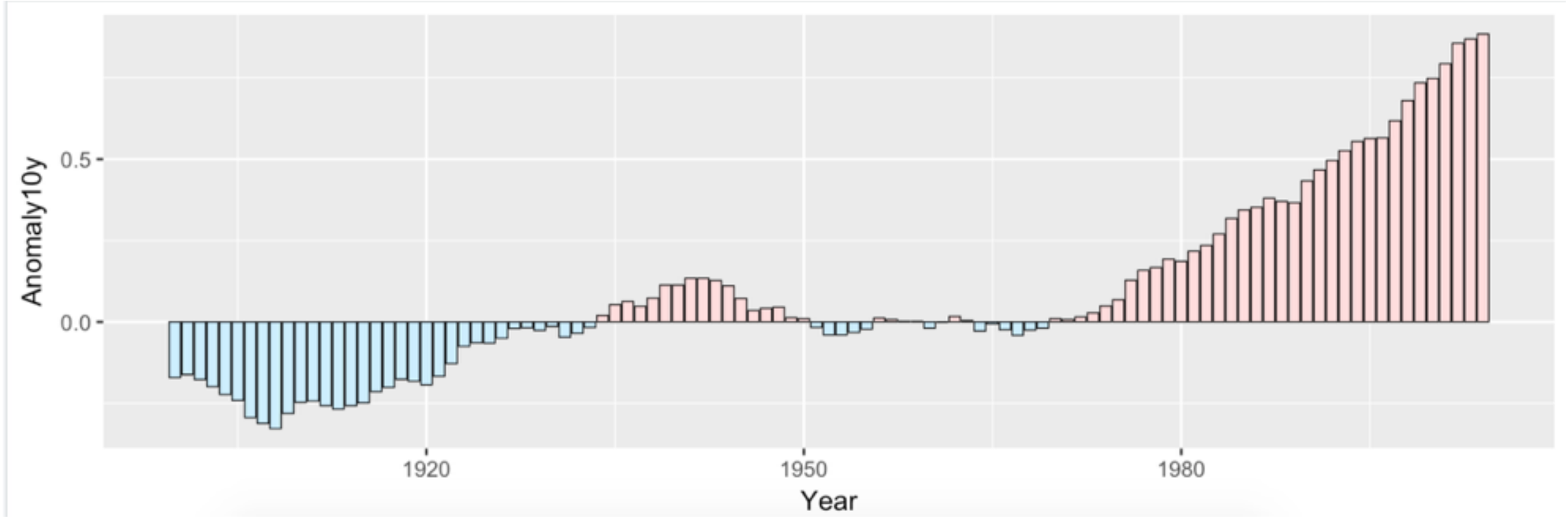
对正负条形图分别着色


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



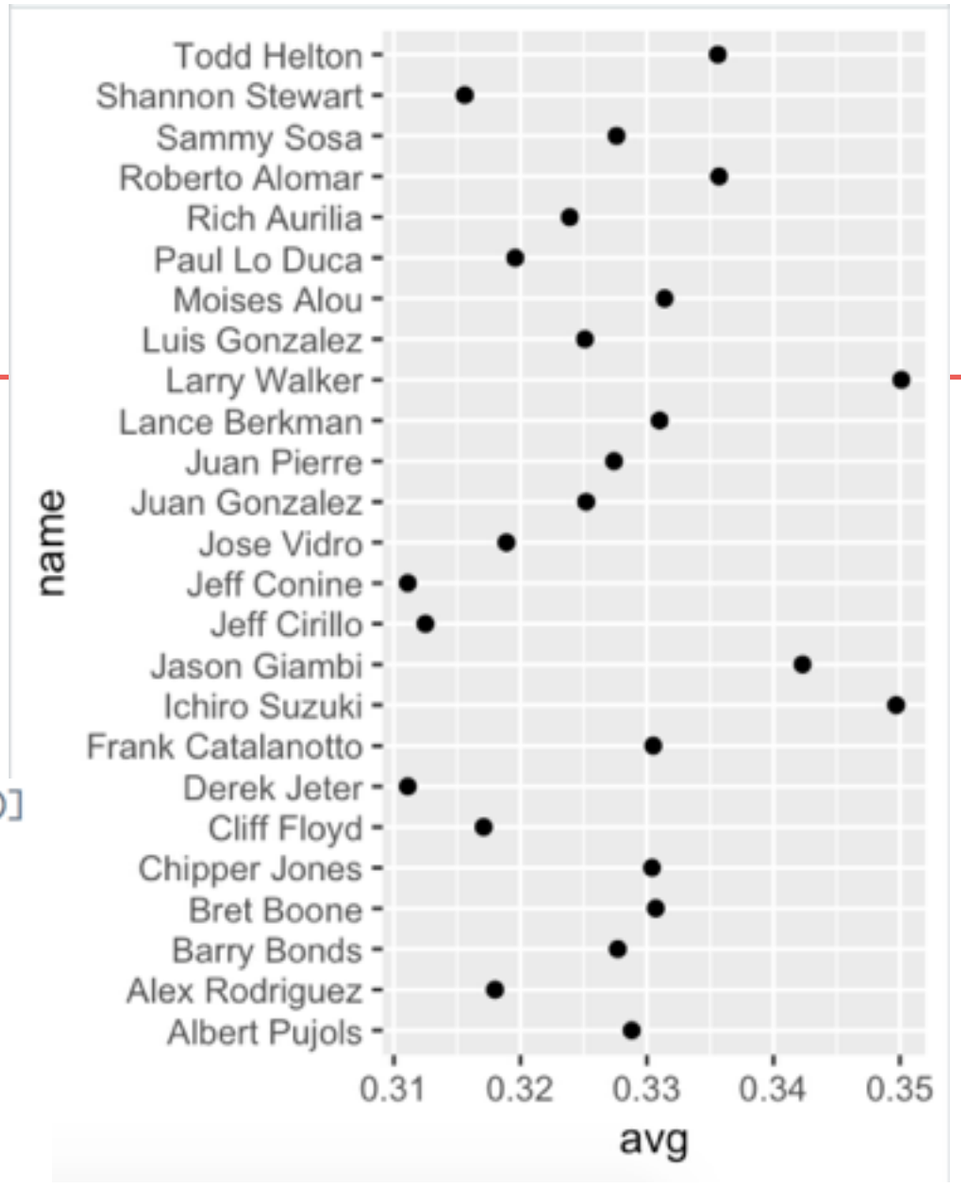
对正负条形图分别着色


```
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)
```

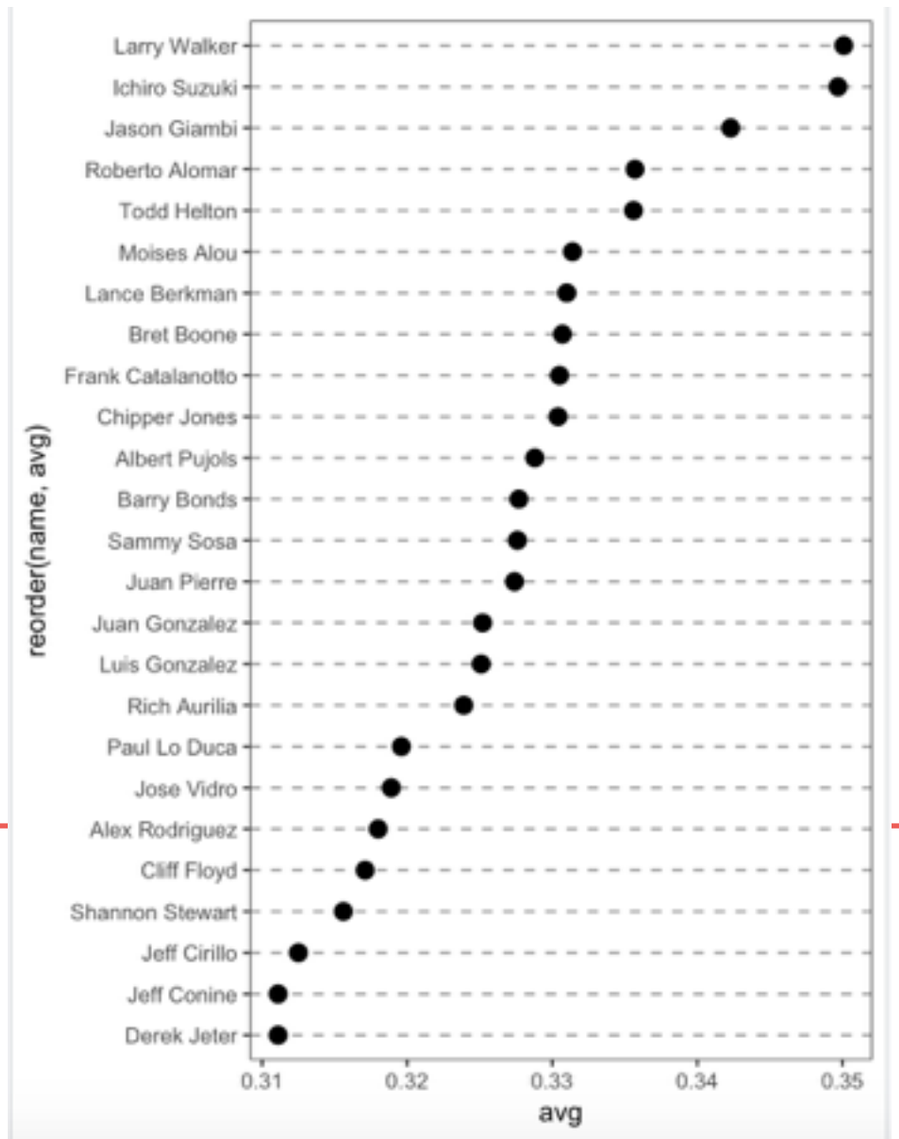


克利夫兰(Cleveland)点图


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

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



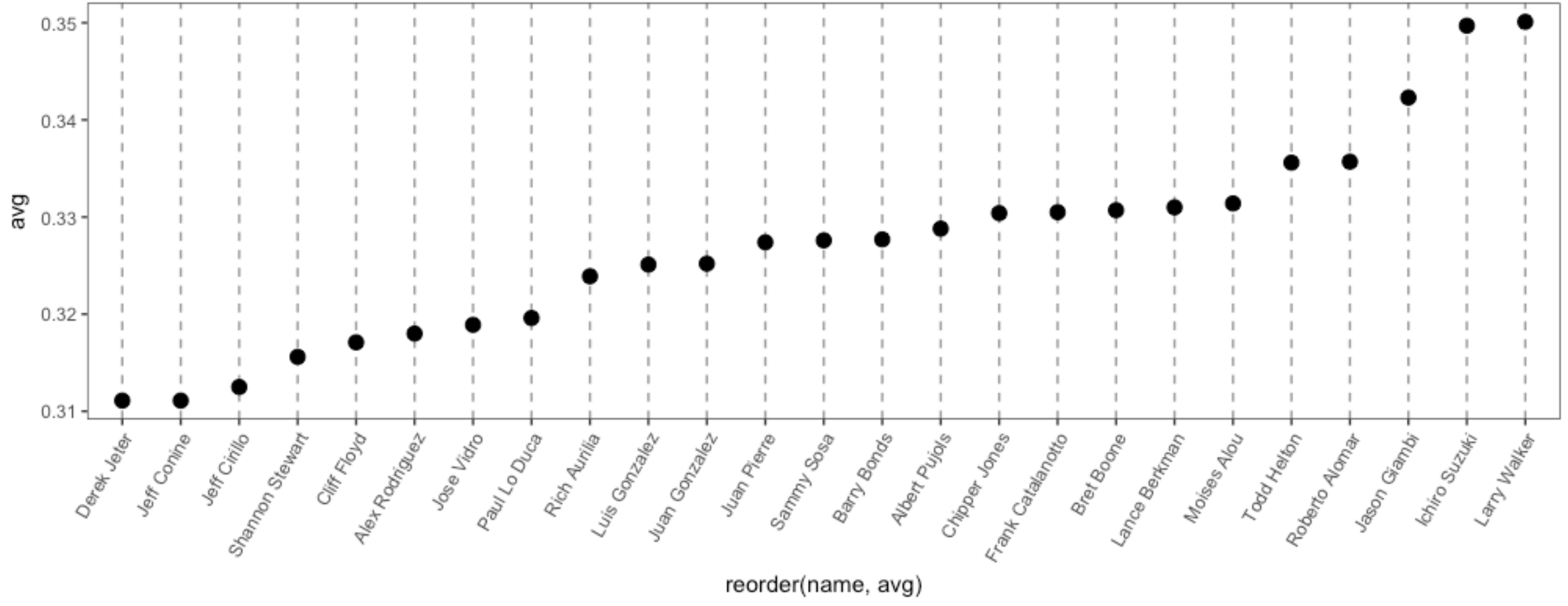

```
tophit[, c("name", "lg", "avg")]

            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"))
```

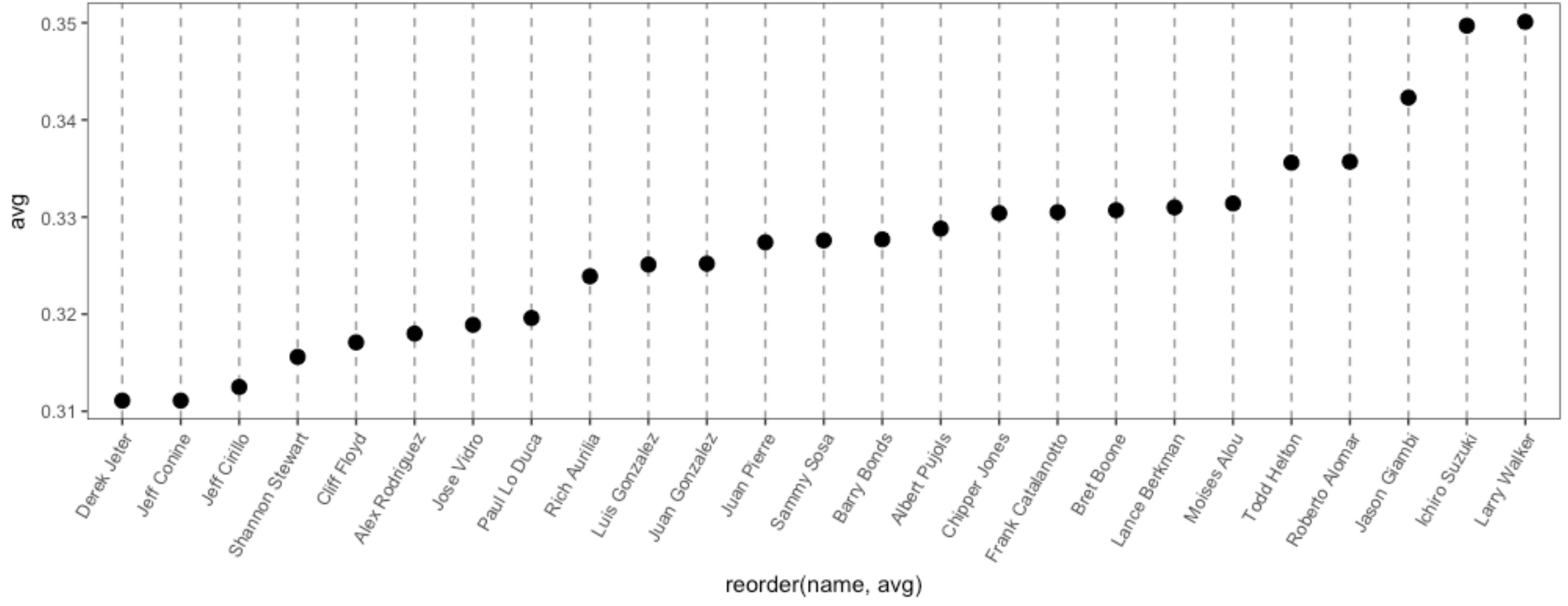
克利夫兰(Cleveland)点图

```
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"))
```



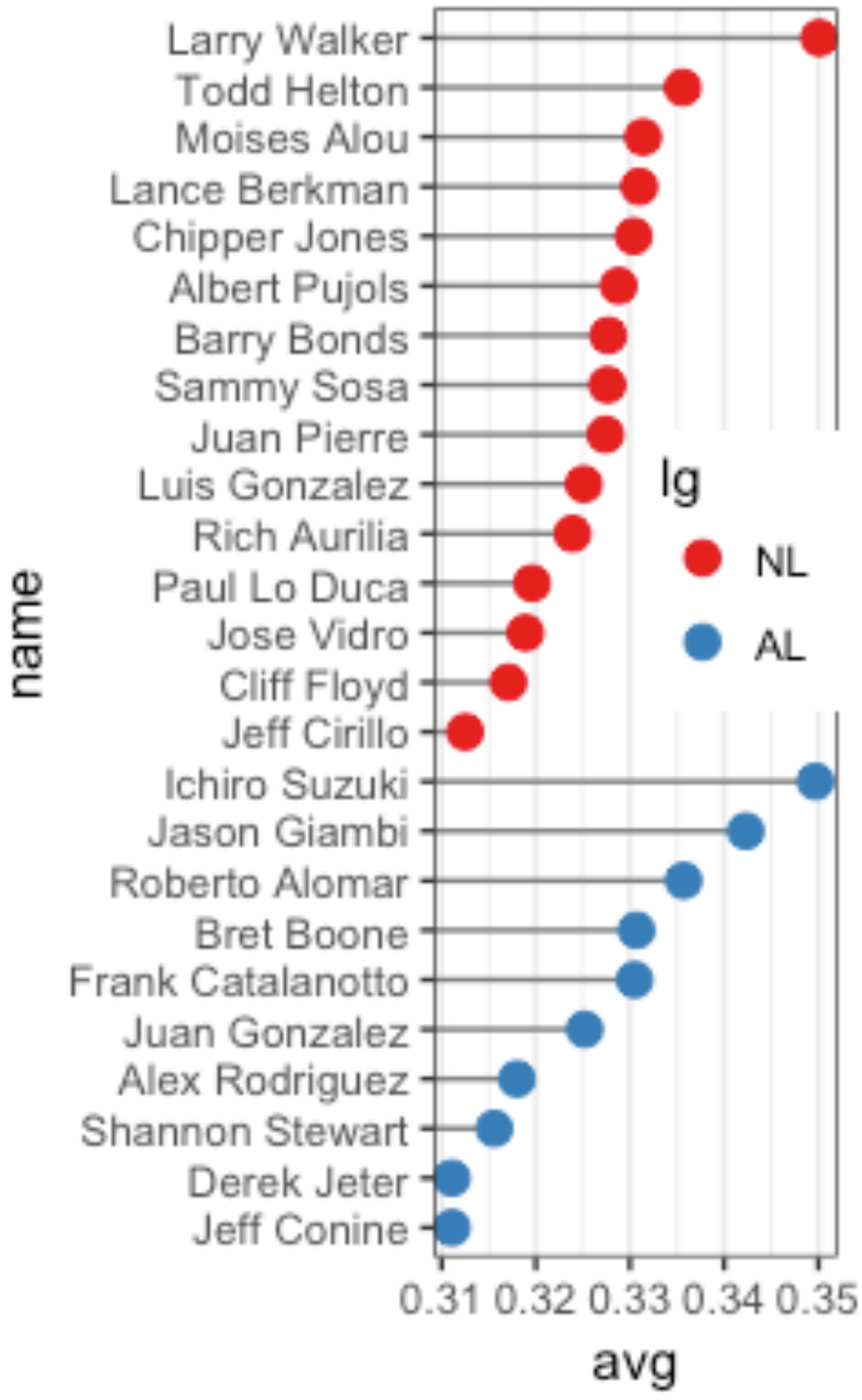
克利夫兰(Cleveland)点图

```
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"))
```



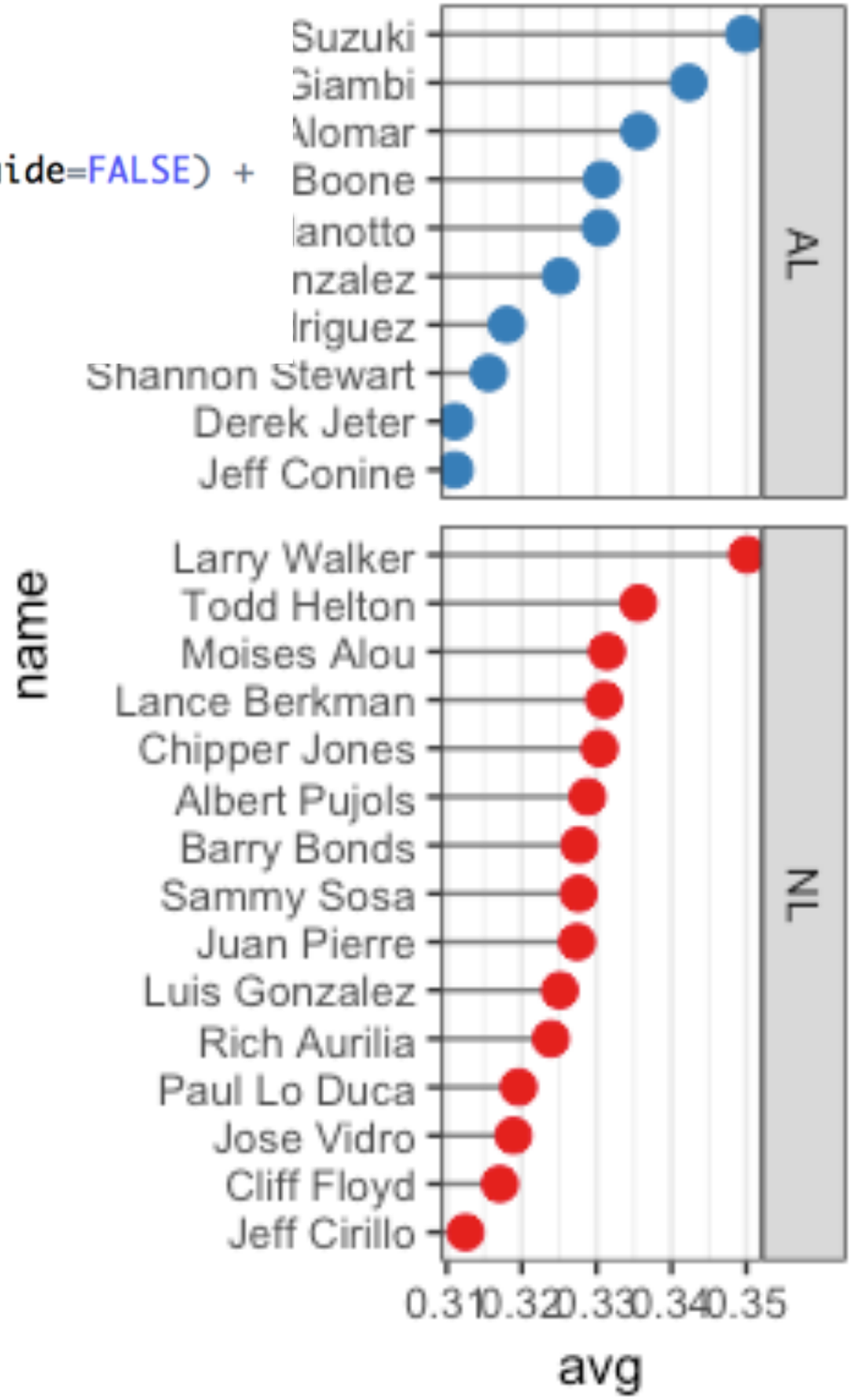
克利夫兰(Cleveland)点图

```
# 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

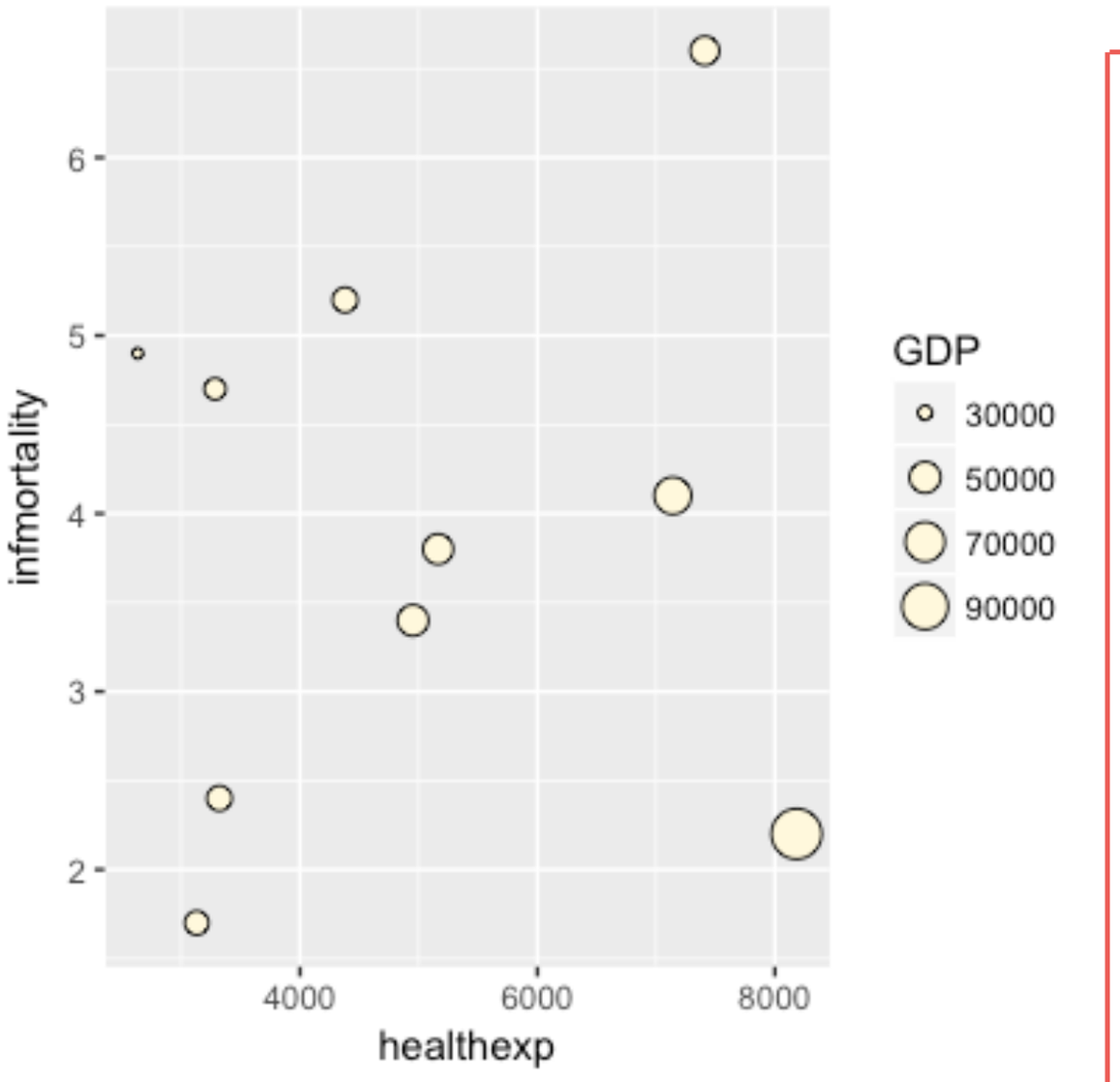

```
library(gcookbook) # For the data set

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

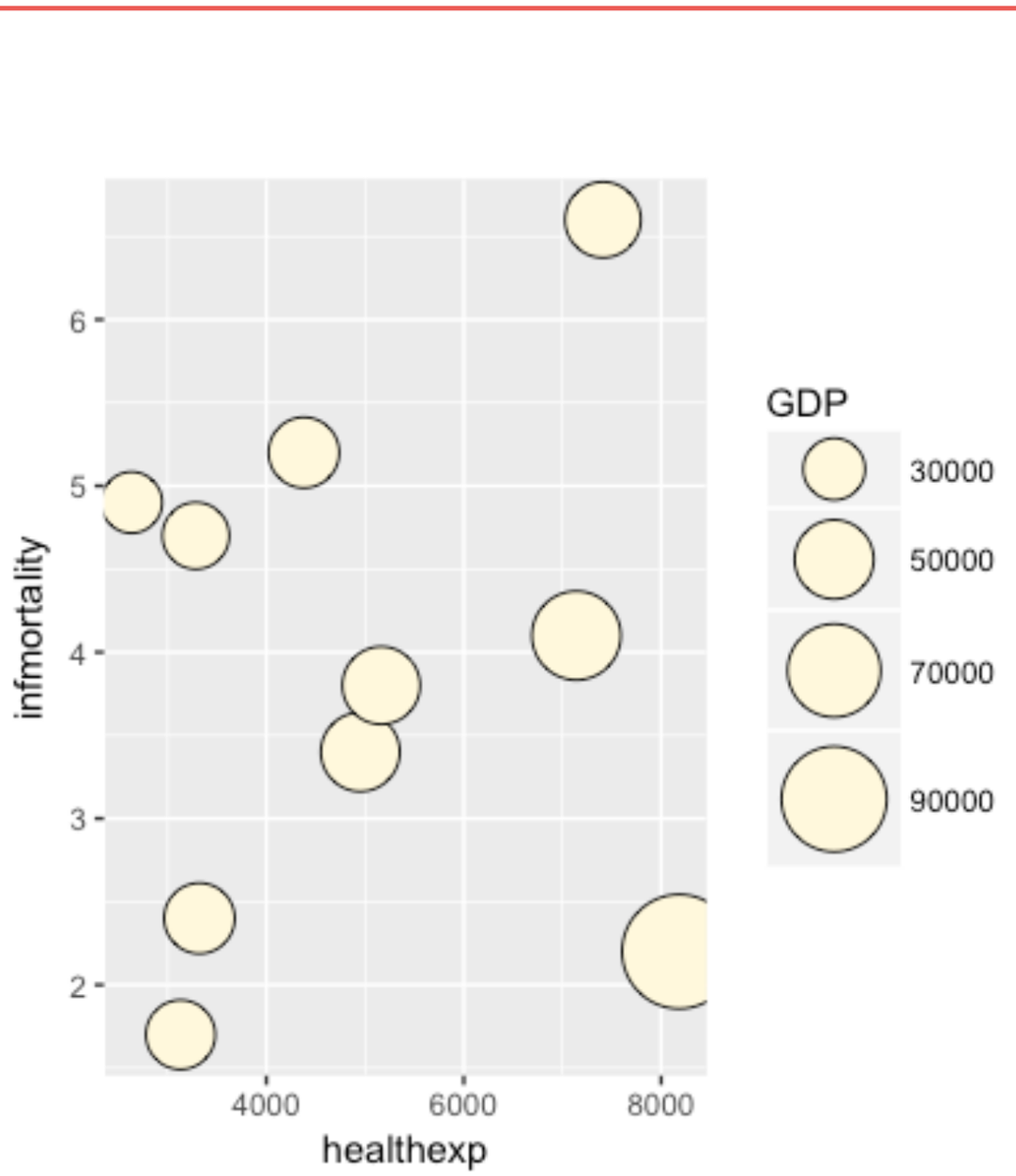
```
cdat
```

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")
```

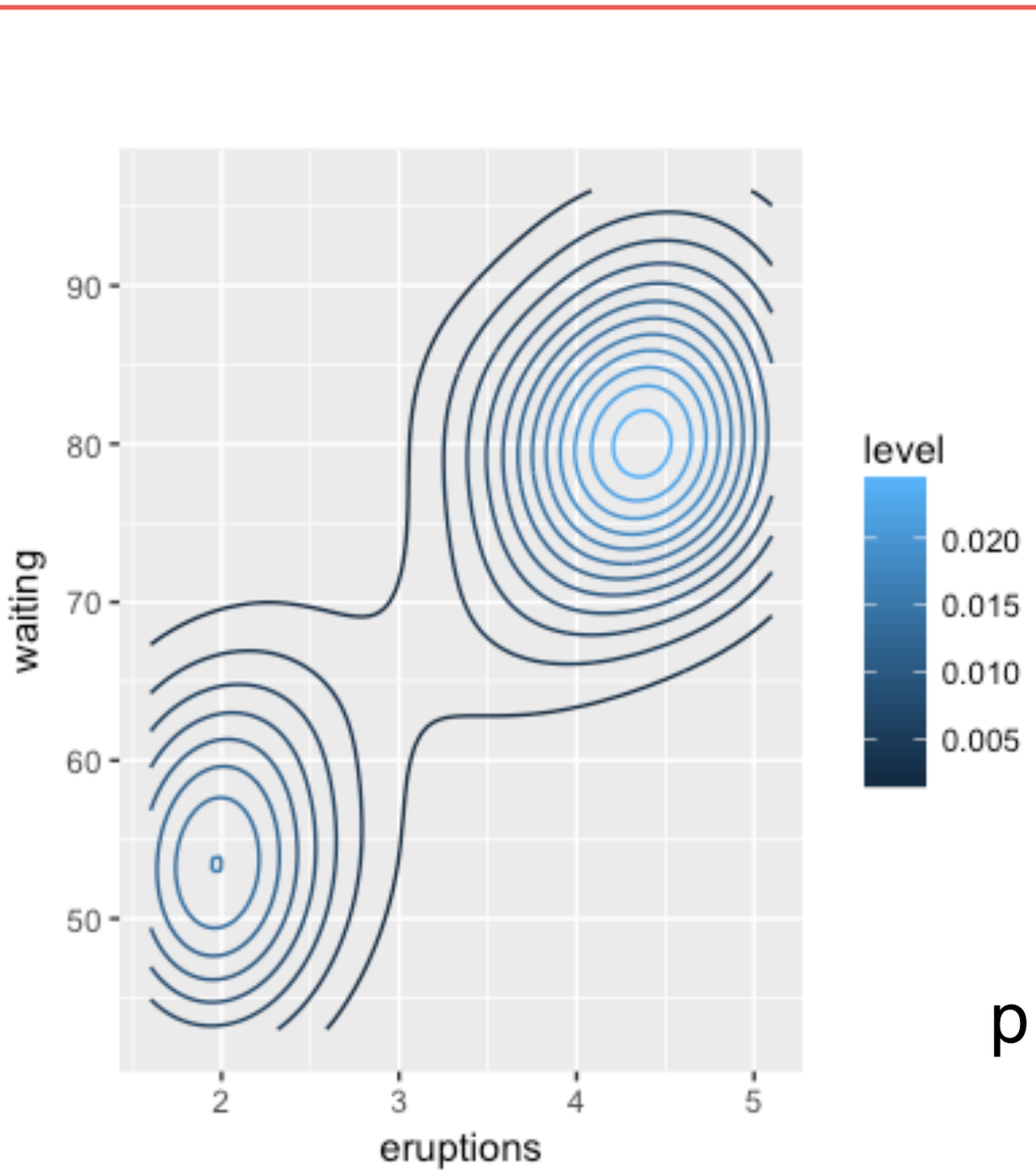


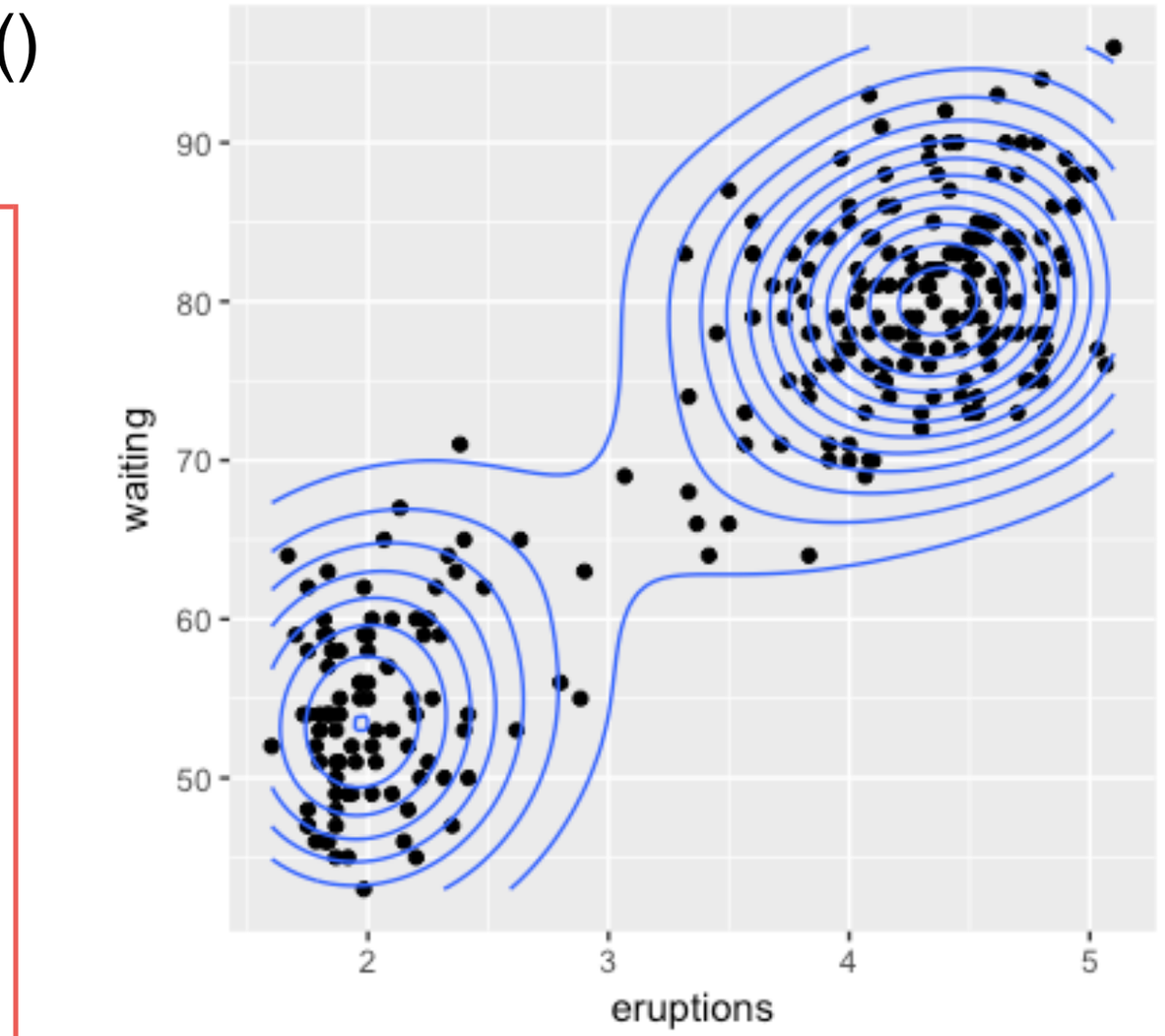



```
p + scale_size_area(max_size=15)
```

绘制二维数据的密度图


```
p <- ggplot(faithful, aes(x=eruptions, y=waiting))
p + geom_point() + stat_density2d()
```





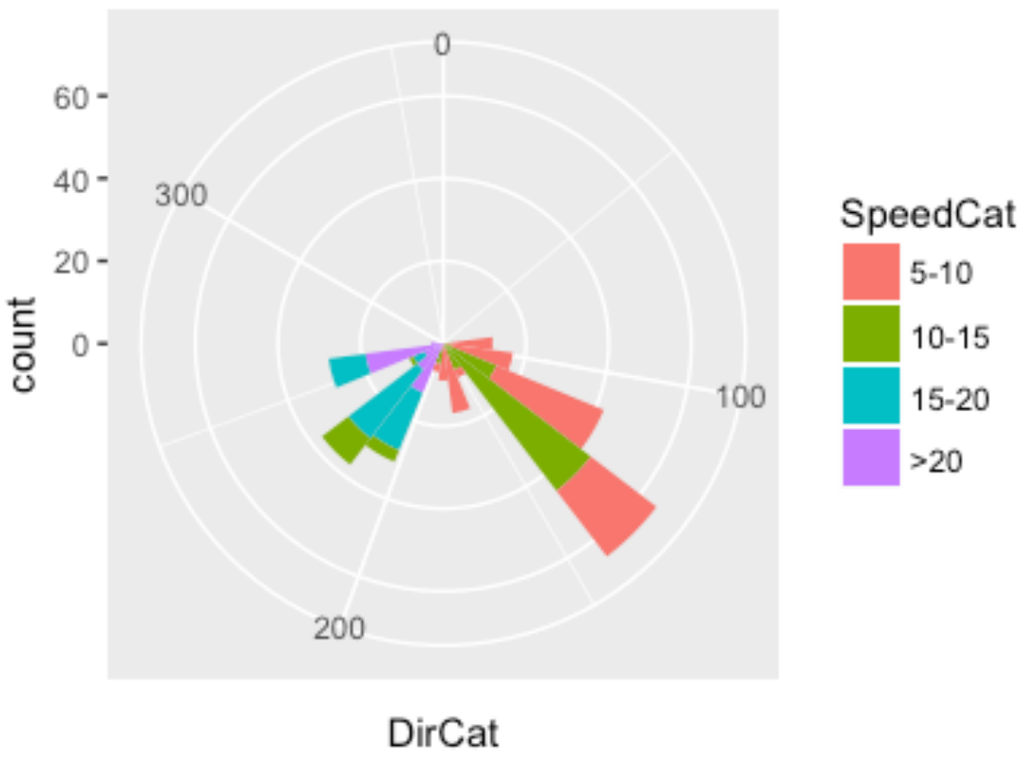

```
p + stat_density2d(aes(colour=..level..))
```

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

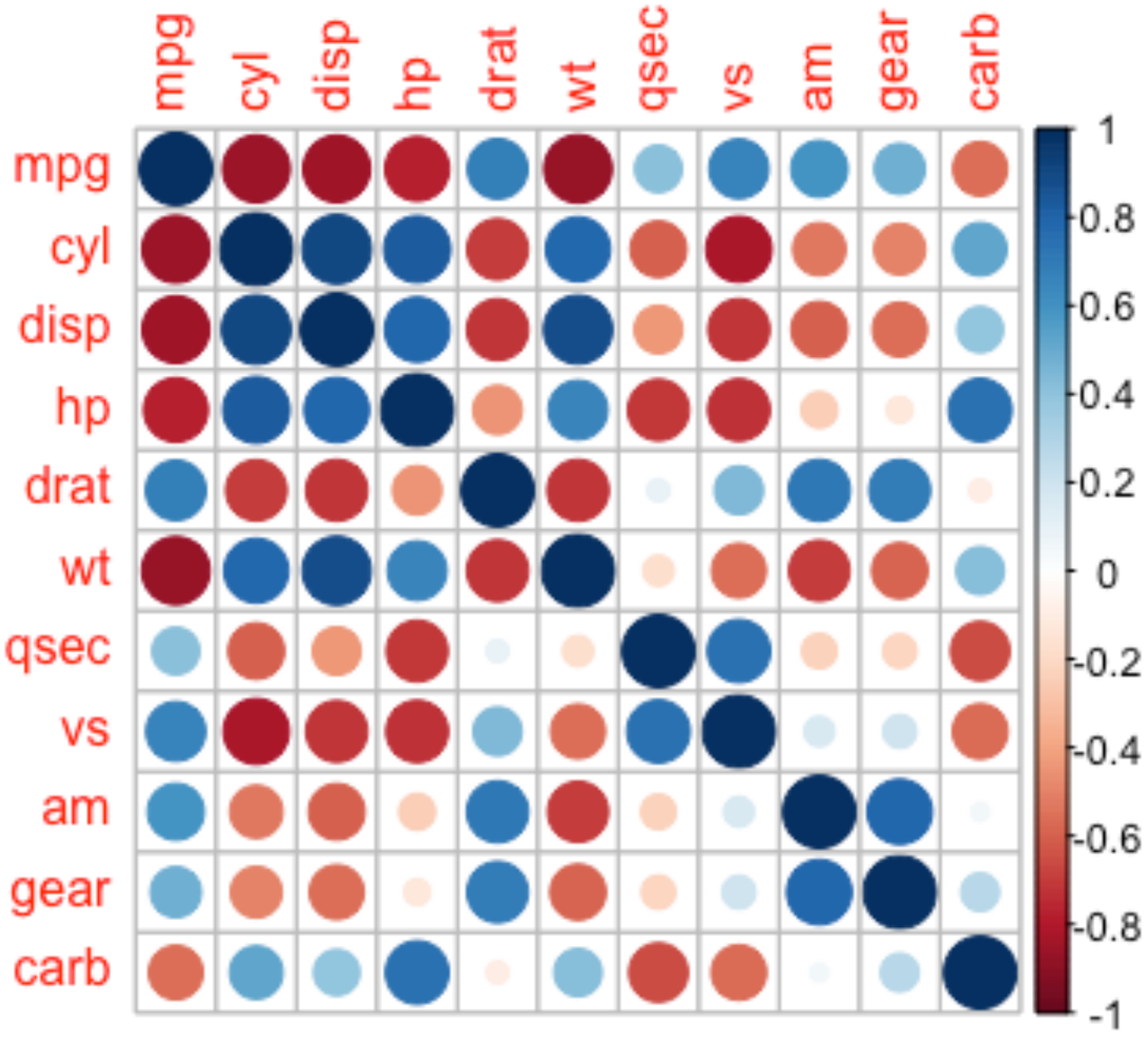

```
mcor <- cor(mtcars)
# Print mcor and round to 2 digits
round(mcor, digits=2)
```

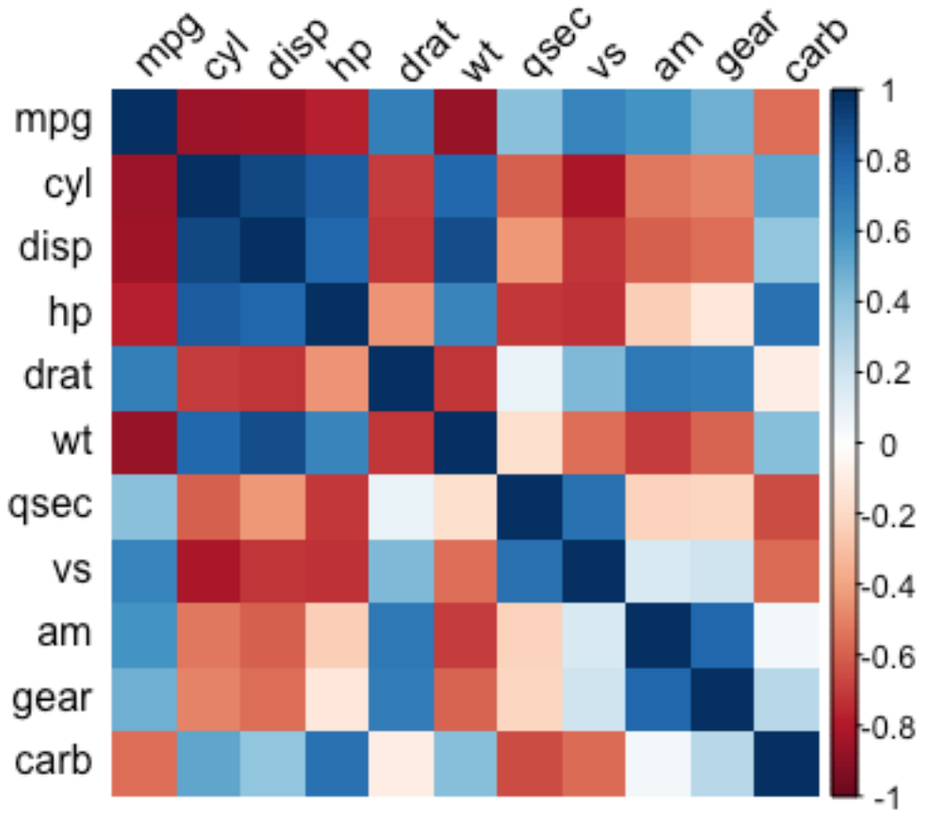
```
       mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
hp   -0.78  0.83  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
```

绘制相关矩阵图

corrplot(mcor)

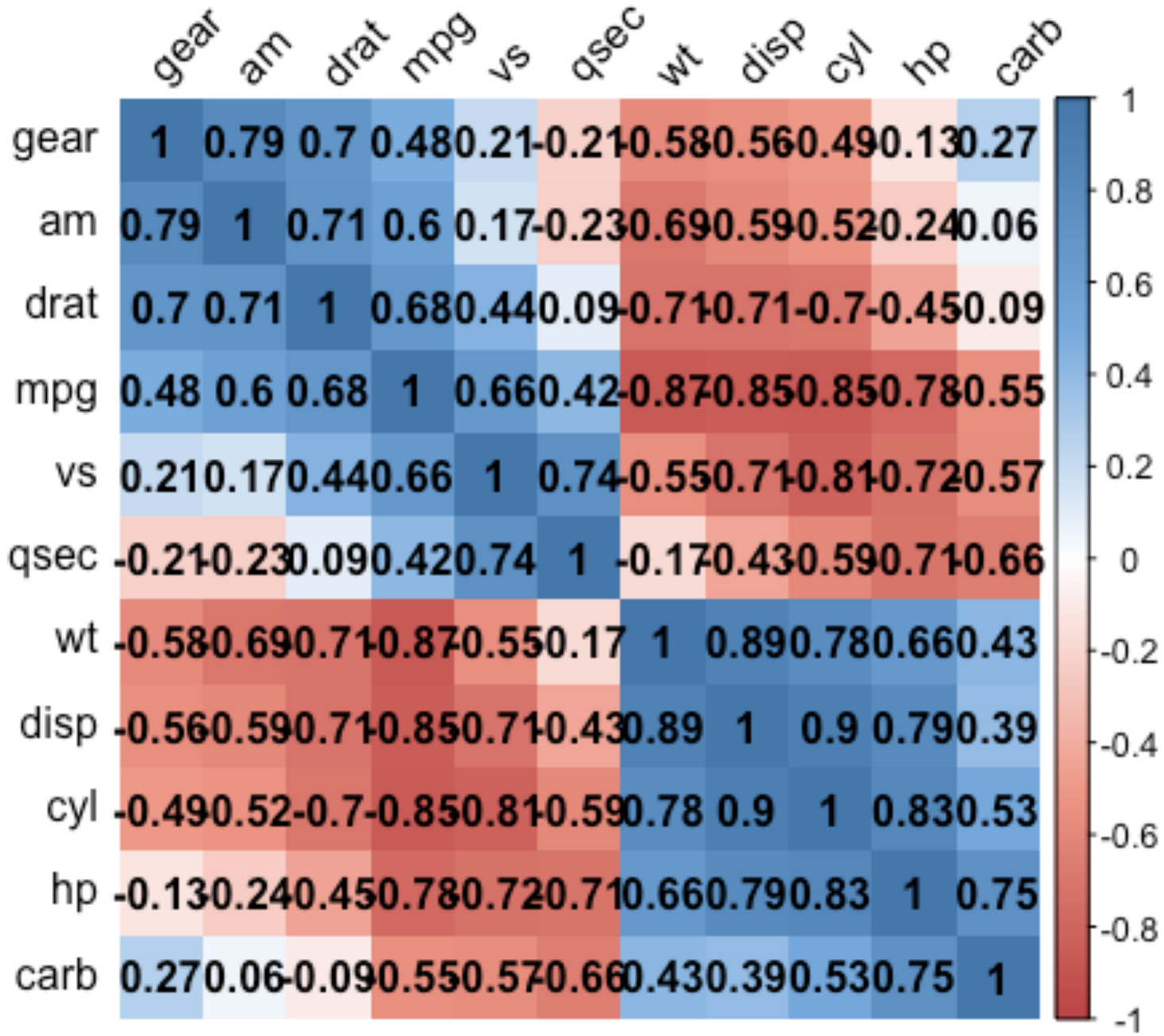





```
corrplot(mcor,
method="shade",
shade.col=NA,
tl.col="black",
tl.srt=45)
```

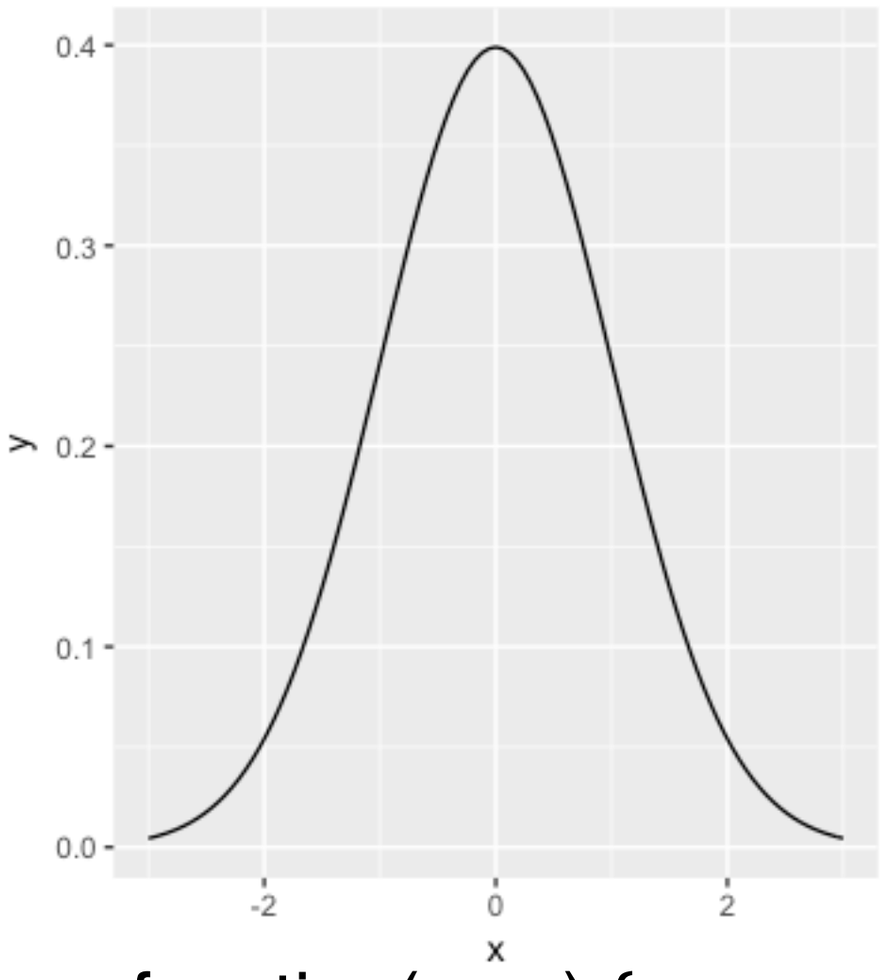
绘制相关矩阵图

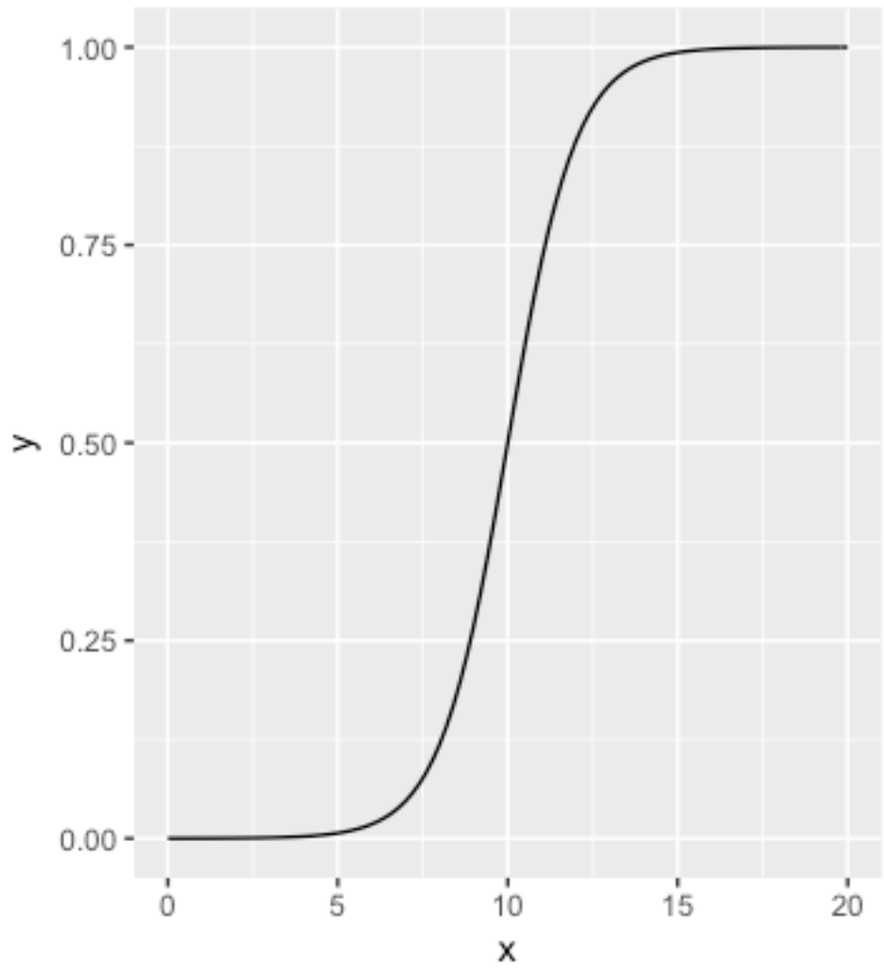
```
col <-
colorRampPalette(c("#BB
4444", "#EE9988",
"#FFFFFF", "#77AADD",
"#4477AA"))

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



绘制函数曲线


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

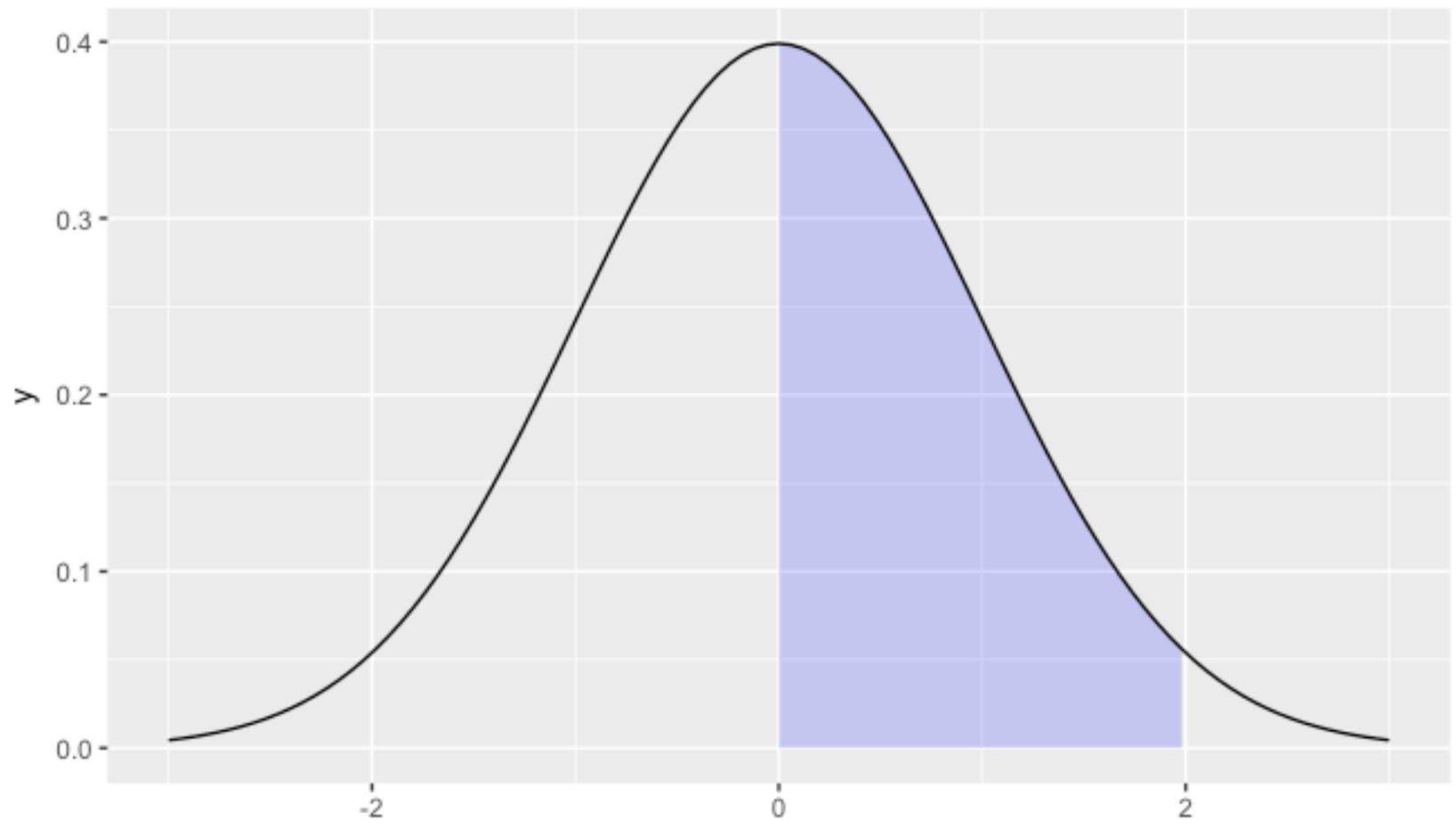





```
myfun <- function(xvar) {
    1/(1 + exp(-xvar + 10))
}
ggplot(data.frame(x=c(0, 20)), aes(x=x)) + stat_function(fun=myfun)
```

绘制函数曲线

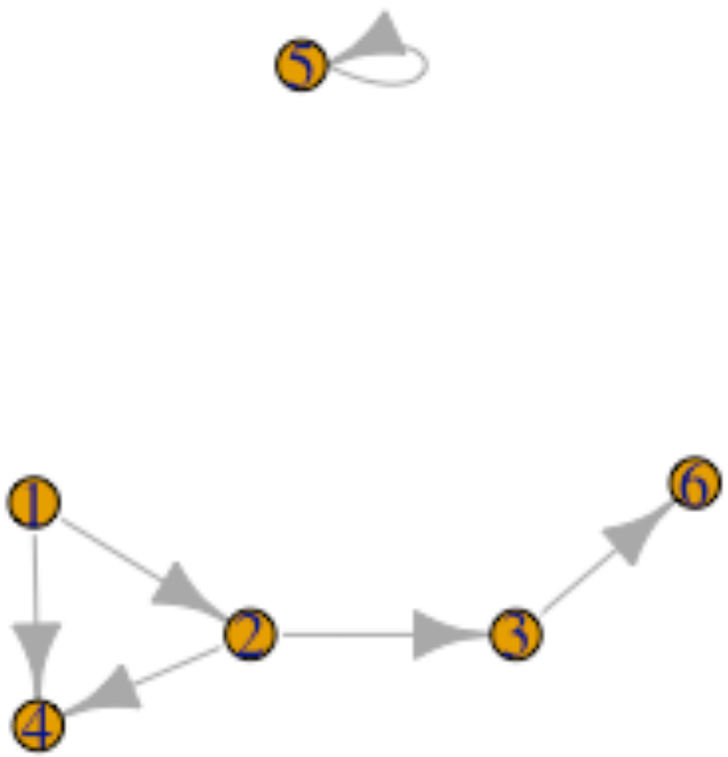

```
# Return dnorm(x) for 0 < x < 2, and NA for all other x
dnorm_limit <- function(x) {
    y <- dnorm(x)
    y[x < 0  |  x > 2] <- NA
     return(y)
}
```



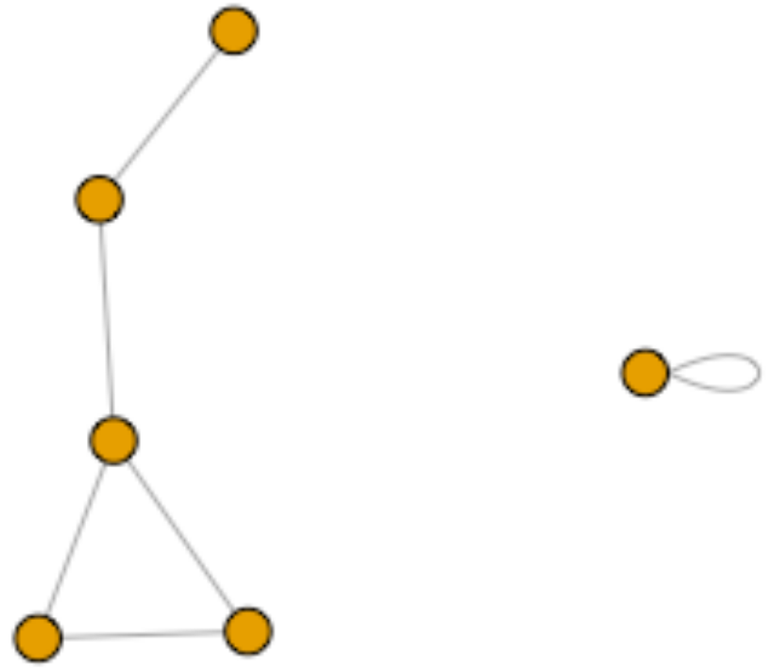

```
# ggplot() with dummy
data
p <-
ggplot(data.frame(x=c(-
3, 3)), aes(x=x))

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

stat_function(fun=dnor
m)
```

绘制网络图


```
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)
```

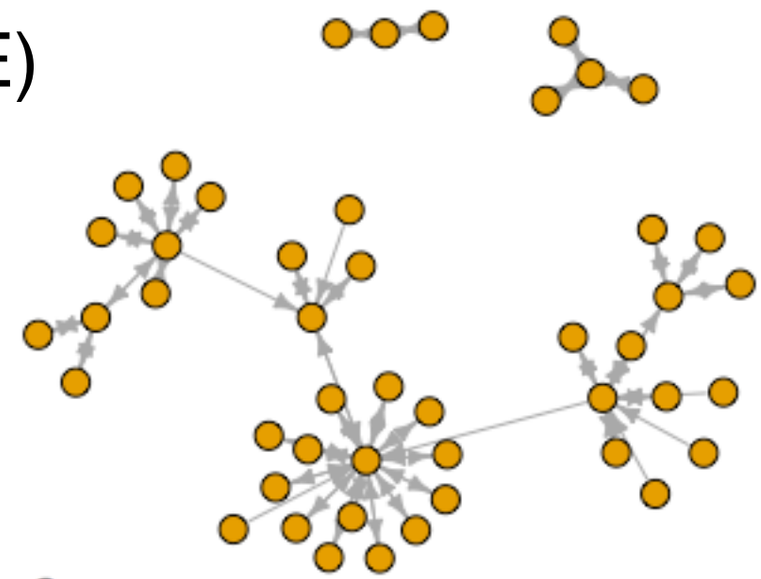



```
# For an undirected
graph
gu <- graph(c(1,2, 2,3,
2,4, 1,4, 5,5, 3,6),
directed=FALSE)
# No labels
plot(gu, vertex.label=NA)
```

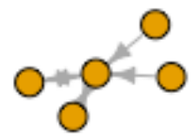


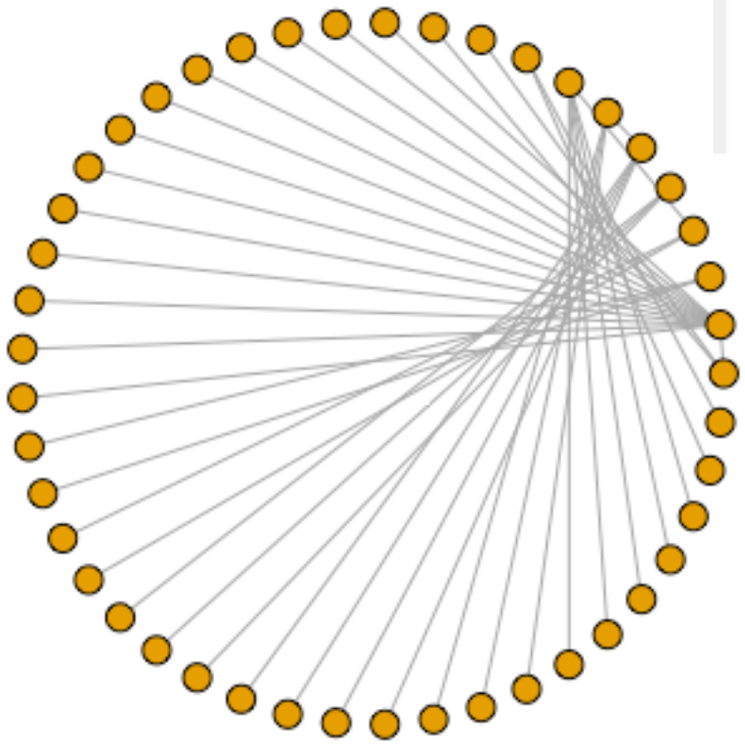
绘制网络图

```
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 - 0046

阅读所有章节，
运行所有代码

注解	小提琴图
坐标系	热图
图例	三维图
分面	谱系图
颜色	向量图
输出	马赛克图
……	……

确定分组和包

谢谢！

孙惠平

sunhp@ss.pku.edu.cn