统计图形概览

与其在 R 下的实现

中南大学数学院

高涛 joegaotao@gmail.com

李程 cn.cheneylee@gmail.com

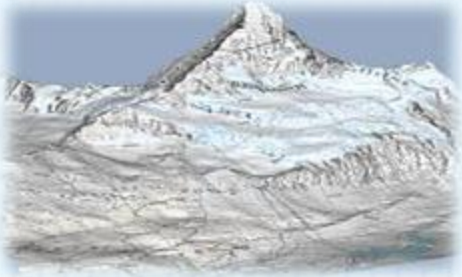
OUTLINE

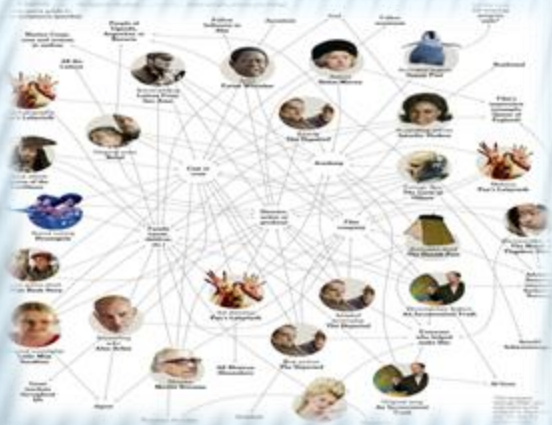
- 可视化概述
- 统计图形概览与其在 R 下实现
- 统计图形的欣赏与批判
- 总结与展望

OUTLINE

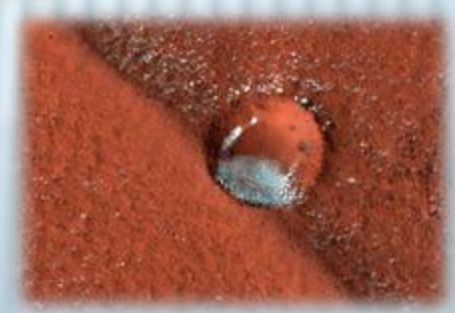
- 可视化概述
- 统计图形概览与其在R下实现
- 统计图形的欣赏与批判
- 总结与展望

1 什么是可视化？






GOD
THE GREAT DEEP
THE ANCIENT HEBREW CONCEPTION OF THE UNIVERSE

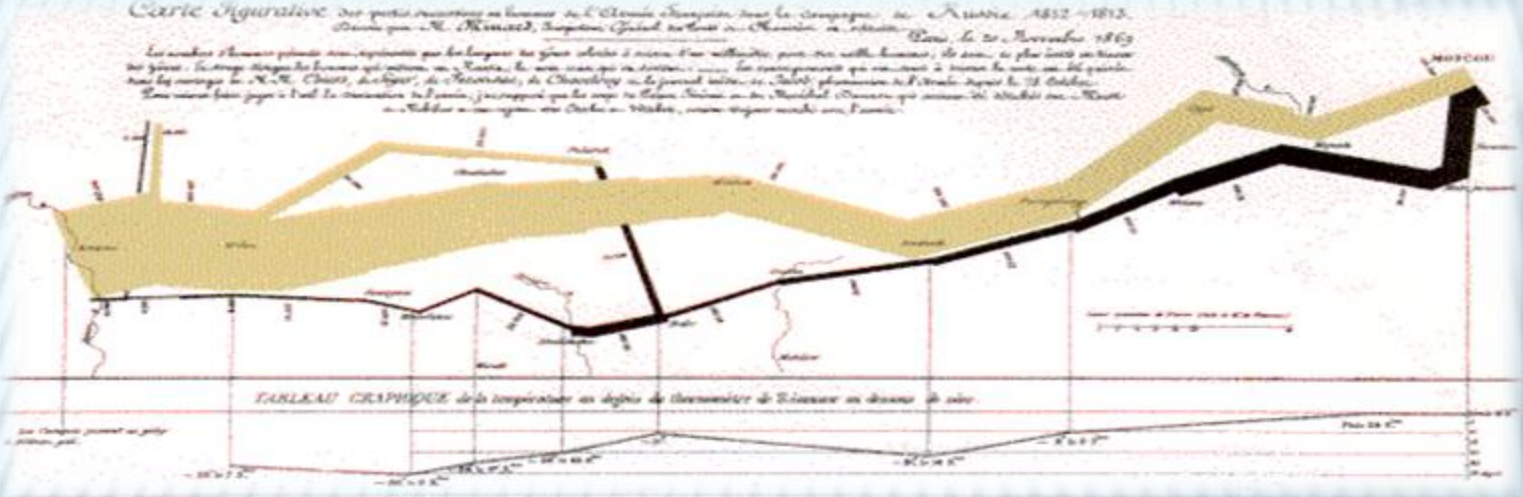



Google

1 什么是可视化?

可视化(Visualization)是利用计算机图形学和图像处理技术，将数据转换成图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术.

2 可视化目标



- 记录：保存数据
- 分析：得出数据内在的关系
- 呈现：将数据隐藏的信息直观的呈现

The purpose of visualization is to **convey information to people through graphical means.**

3 可视化意义

化枯燥数据为直观图形

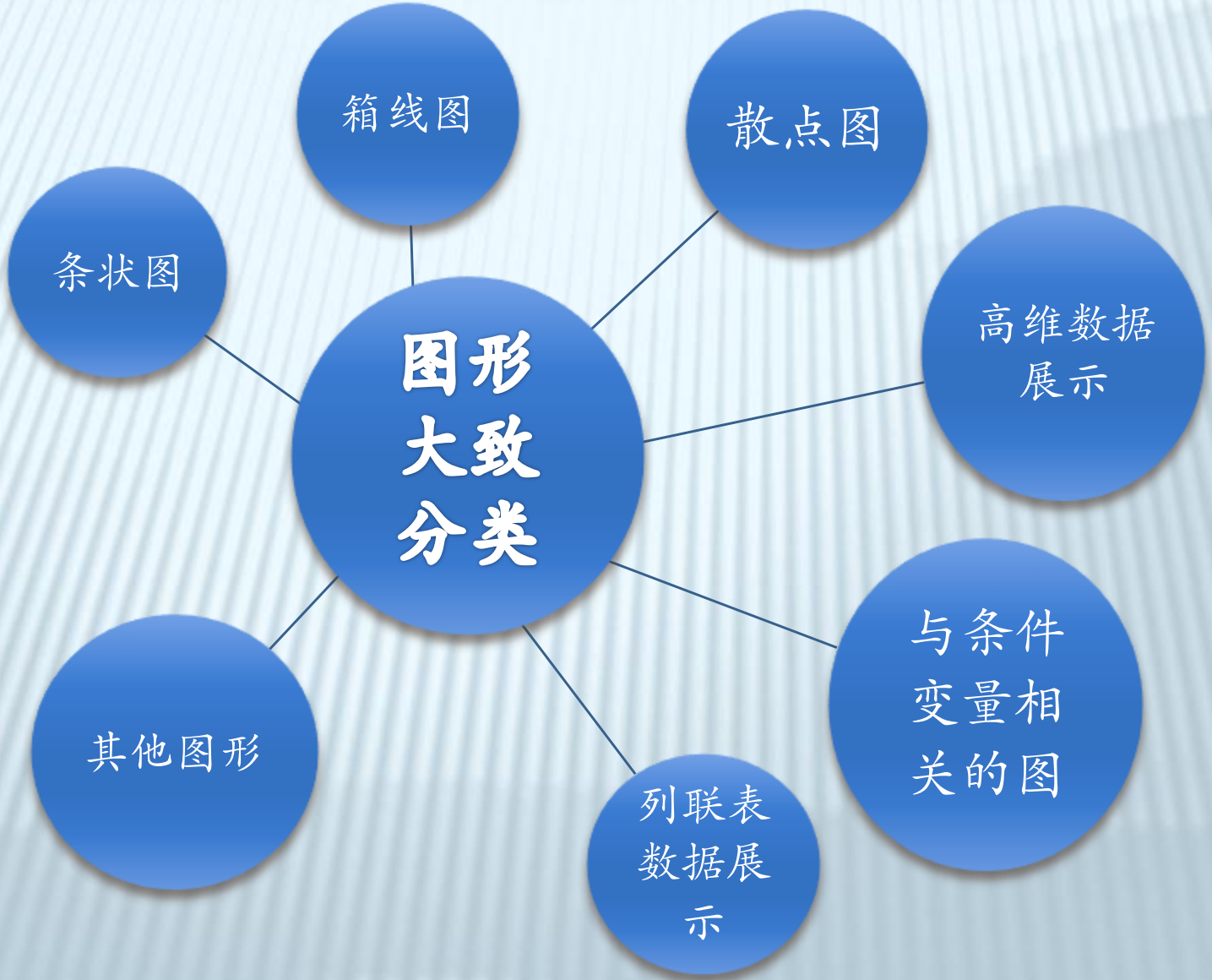
为发现和理解科学规律提供有力工具

直观性　敏锐性　开拓性

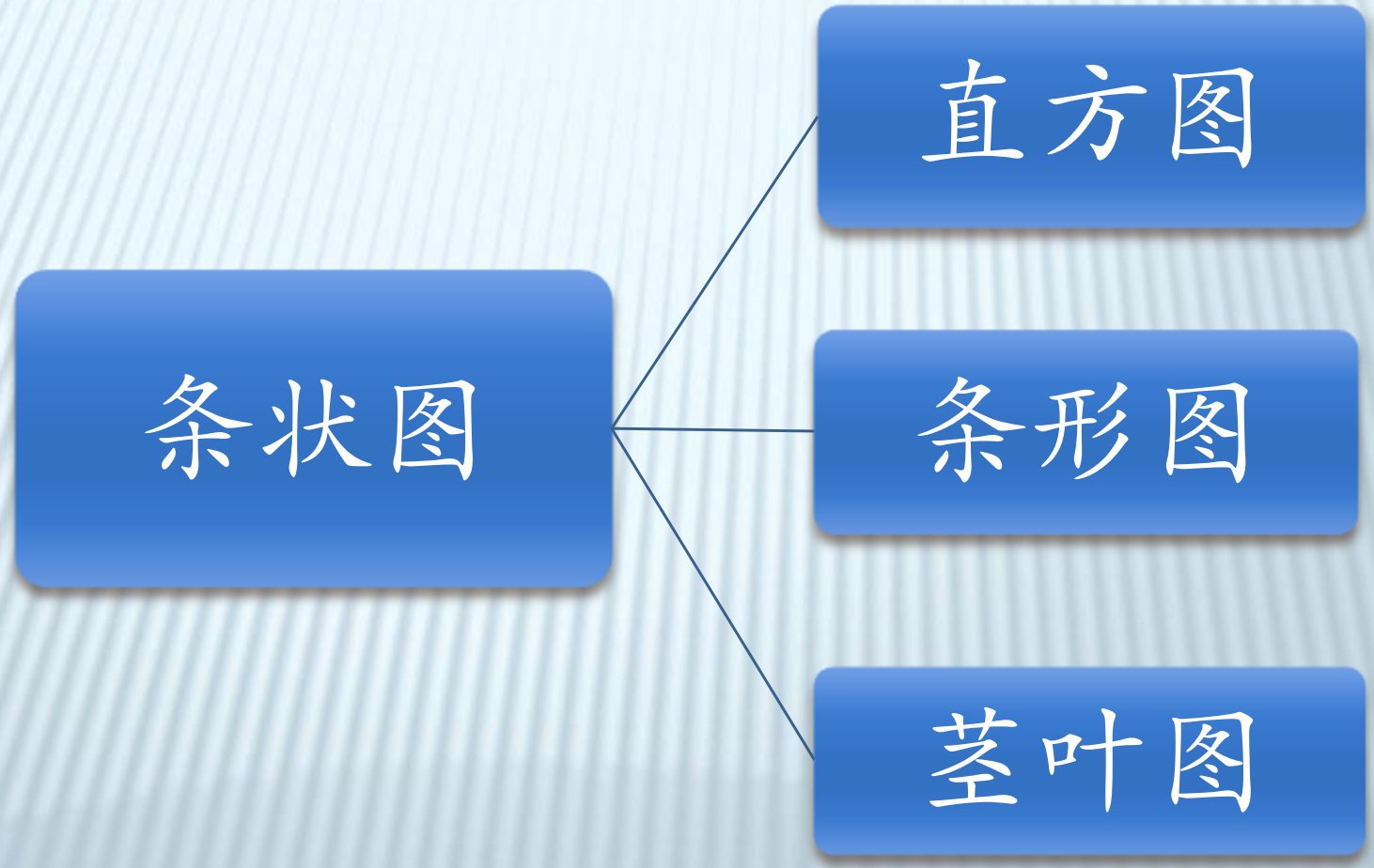
OUTLINE

- 可视化概述
- 统计图形概览与其在 R 下实现
- 统计图形的欣赏与批判
- 总结与展望

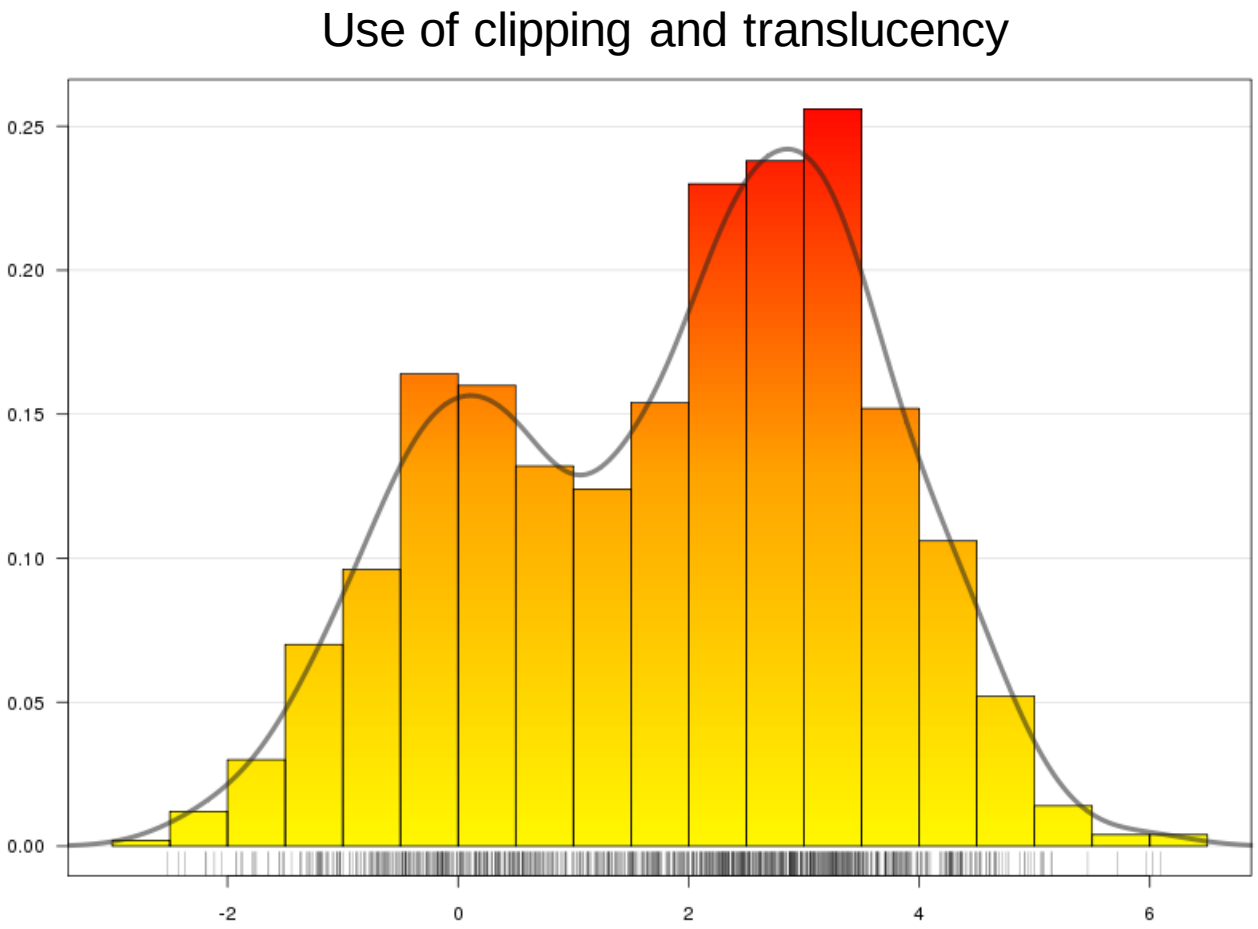
4 统计图形概览与其在R下的实现



4.1 条形图



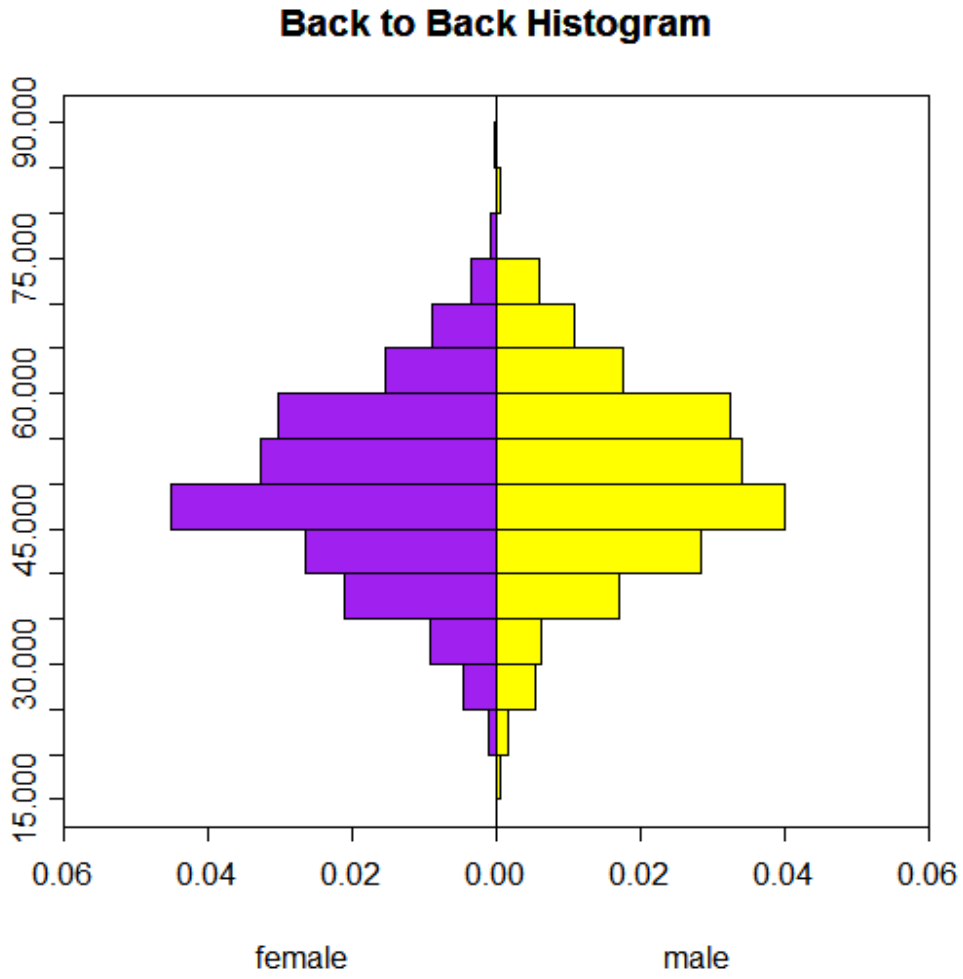
4.1.1 直方图



graphics包

clip()
rug()

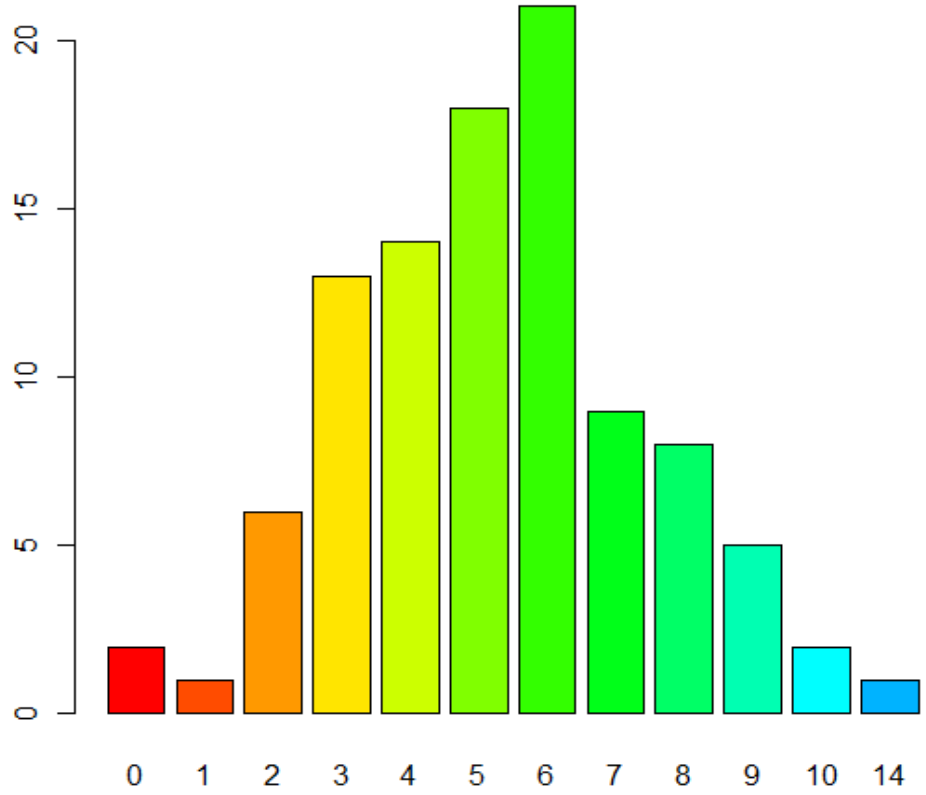
4.1.1 直方图



Himsc包

histbackback(x, y,...)

4.1.2 条形图

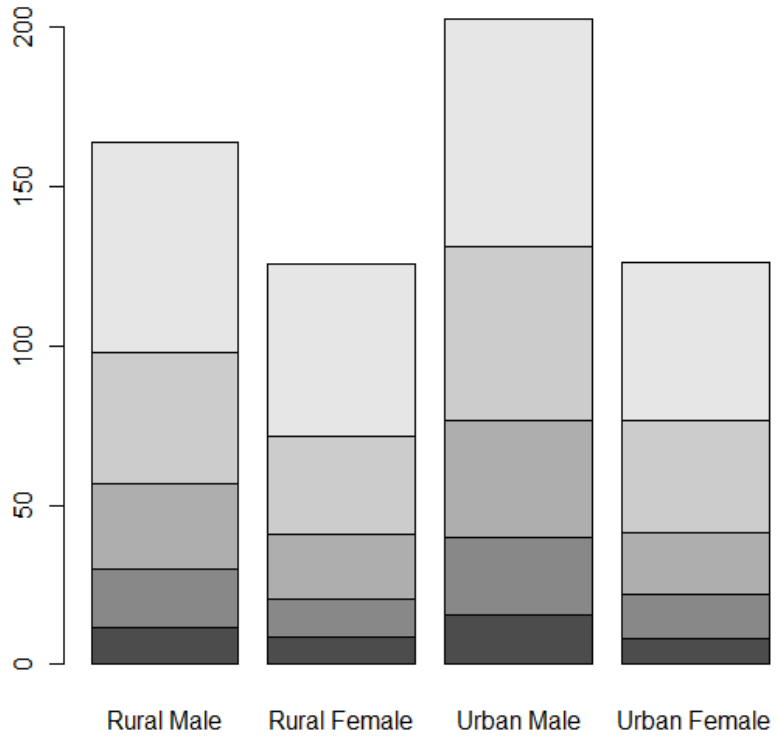


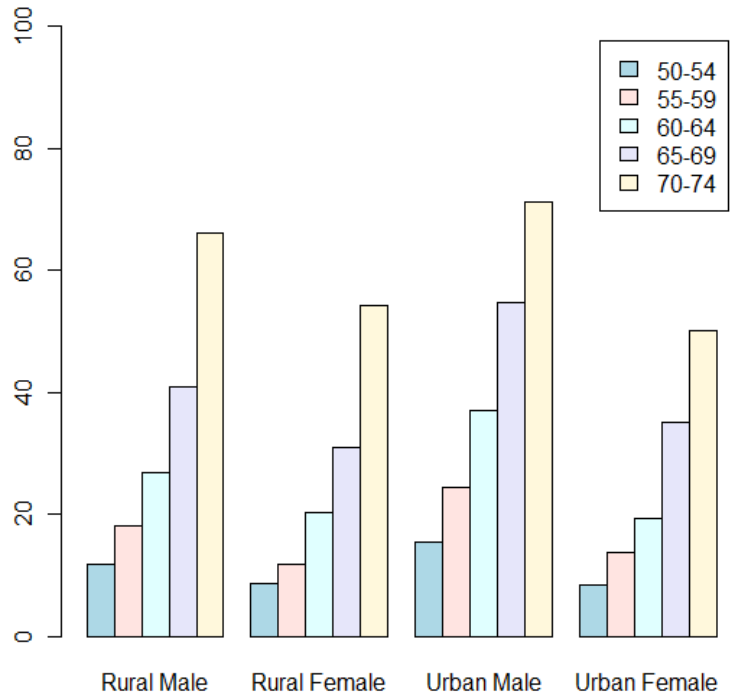
graphics包

barplot(height, ...)

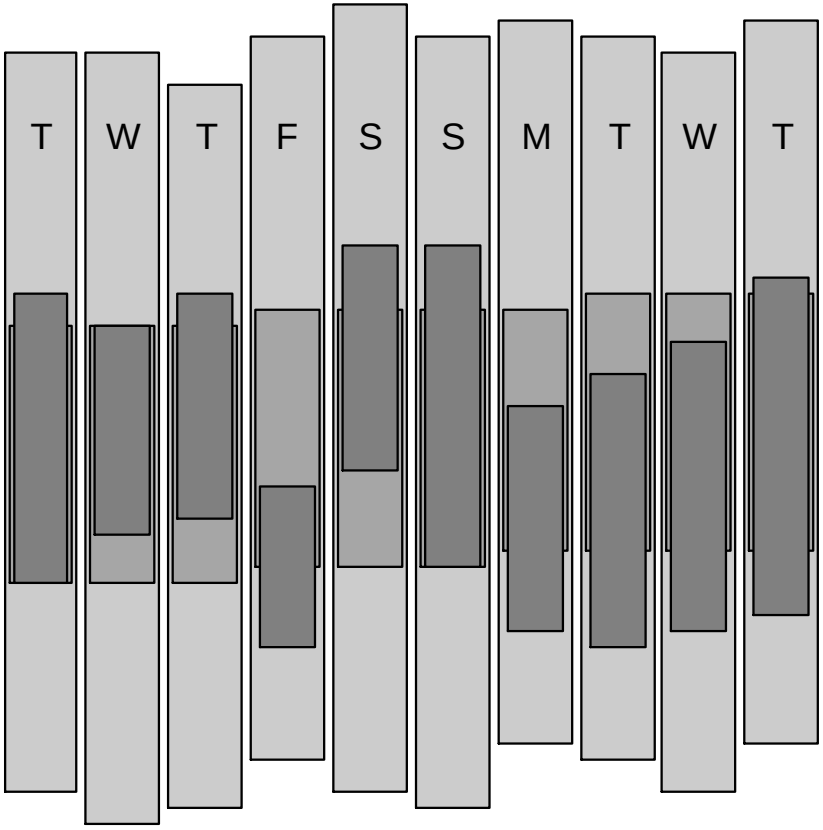
grDevices包

4.1.2 条形图


0
50
100
150
200
Rural Male
Rural Female
Urban Male
Urban Female


0
20
40
60
80
100
50-54
55-59
60-64
65-69
70-74
Rural Male
Rural Female
Urban Male
Urban Female

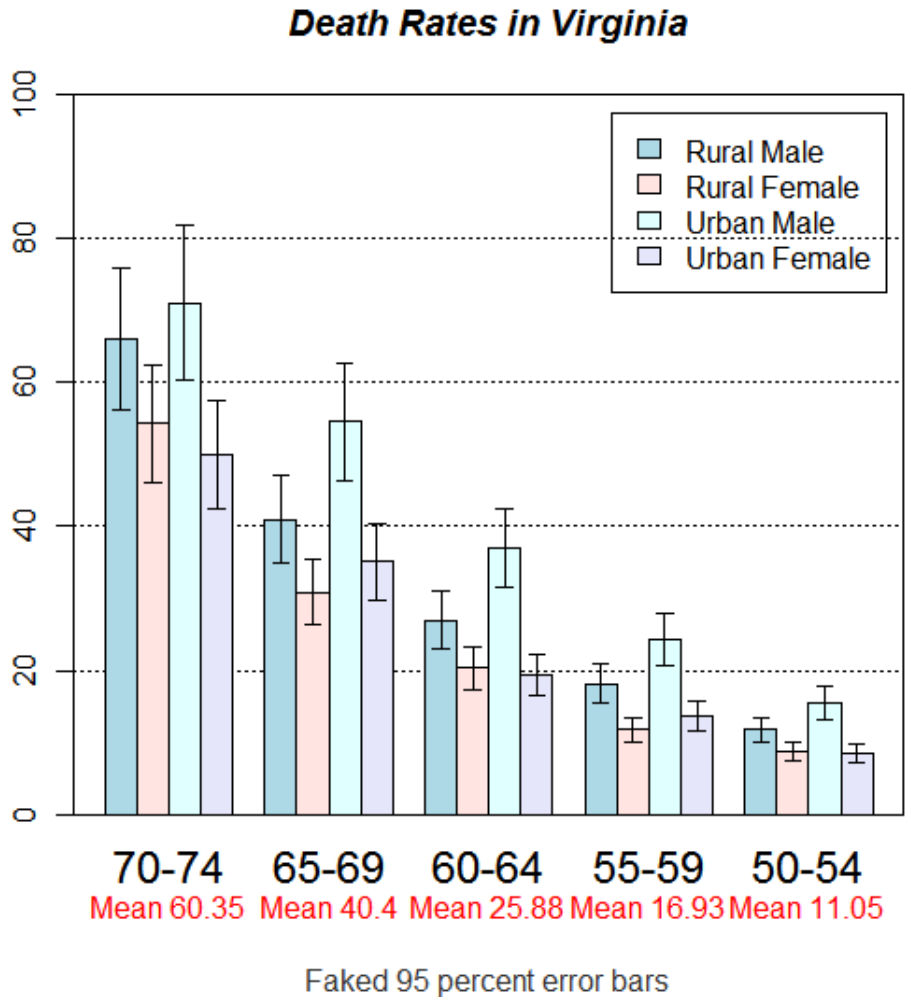
4.1.2 条形图



UsingR包

superbarplot(x,
names=,…)

4.1.2 条形图




```
library(gplots)
hh <- t(VADeaths)[, 5:1]
ci.l <- hh * 0.85ci.u <- hh * 1.15
…
mp <- barplot2(hh,…, plot.ci=TRUE,
        ci.l = ci.l, ci.u = ci.u,
        plot.grid=TRUE)
…
box()
```

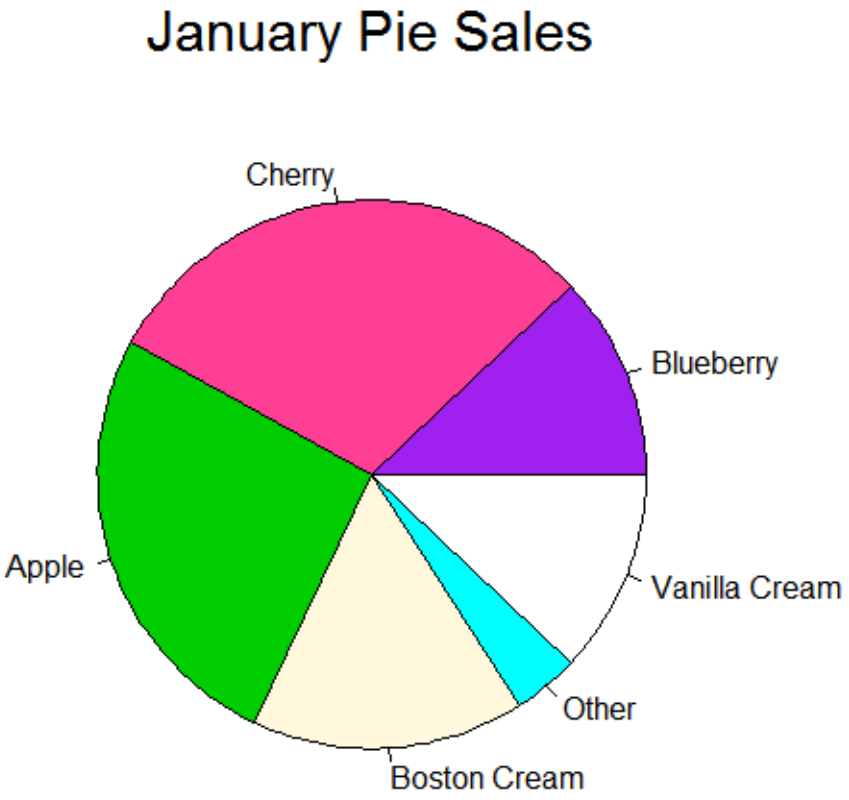
4.1.3 茎叶图

```
> stem(islands)

  The decimal point is 3 digit(s) to the right of the |

   0 | 000000000000000000000000000000111111222338
   2 | 07
   4 | 5
   6 | 8
   8 | 4
  10 | 5
  12 |
  14 |
  16 | 0

> stem(log10(islands))

  The decimal point is at the |

  1 | 1111112222233444
  1 | 5555556666667899999
  2 | 3344
  2 | 59
  3 |
  3 | 5678
  4 | 012
```

graphics包

stem(x, scale = 1, width = 80, atom = 1e-08)

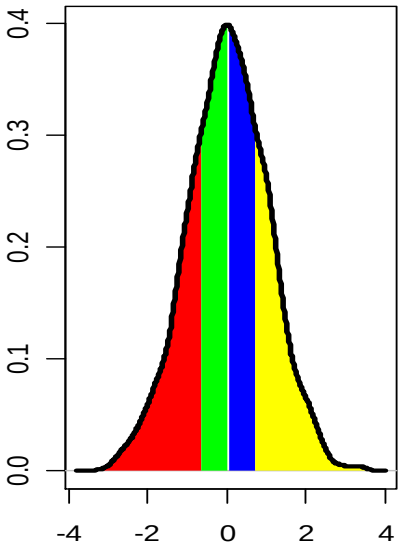
4.2 饼图

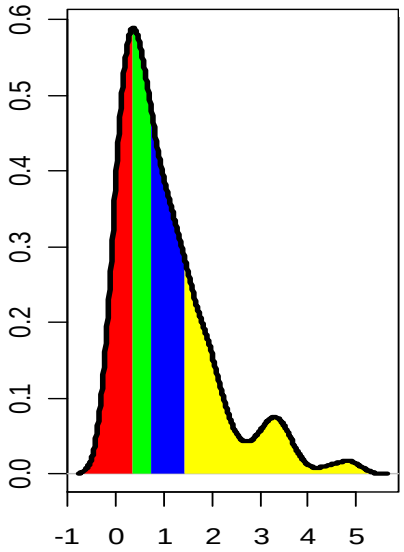


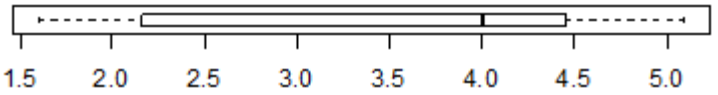
graphics包

pie(x, …)

4.3 箱线图



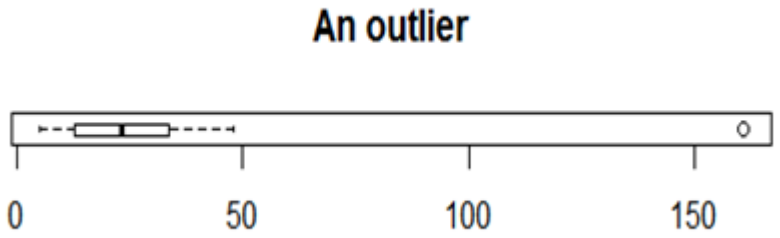




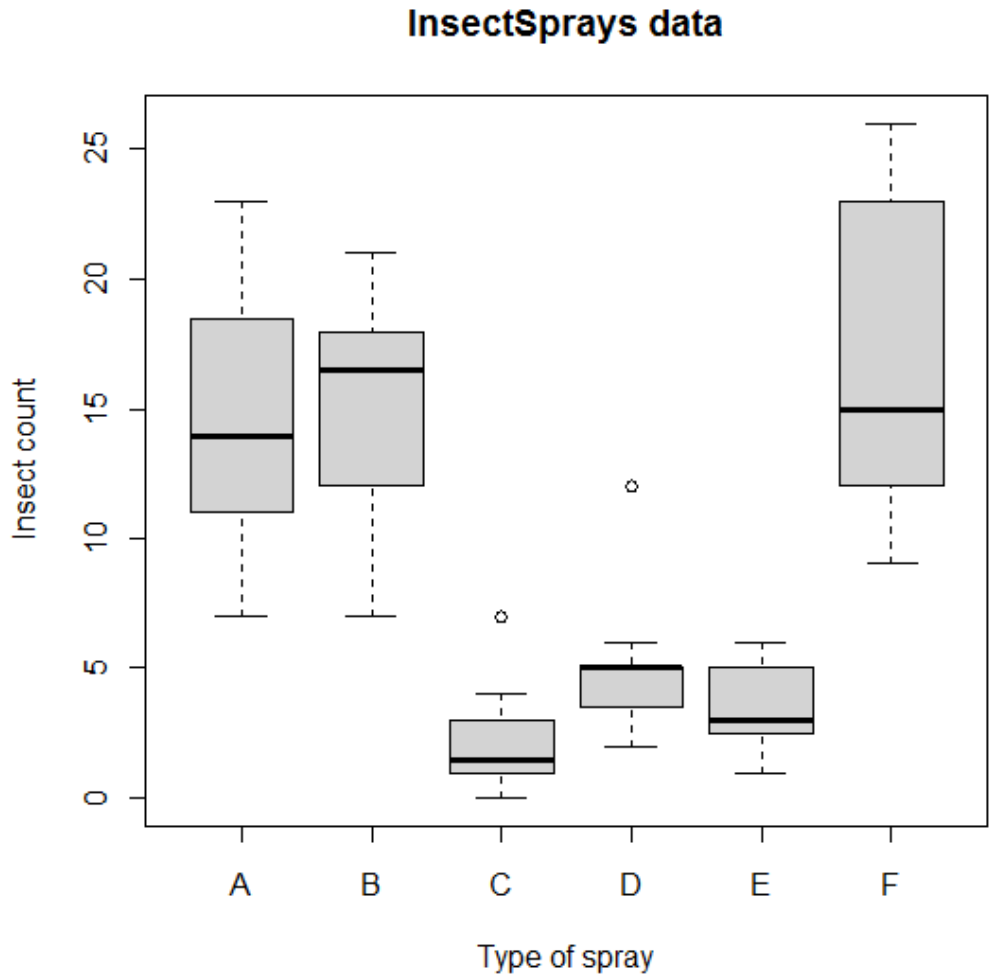
数据的分布

评估数据对称性

直观显示离群点



4.3 箱线图

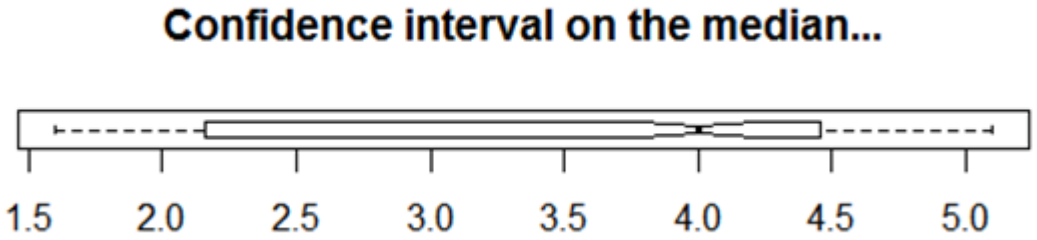


- 常规箱线图

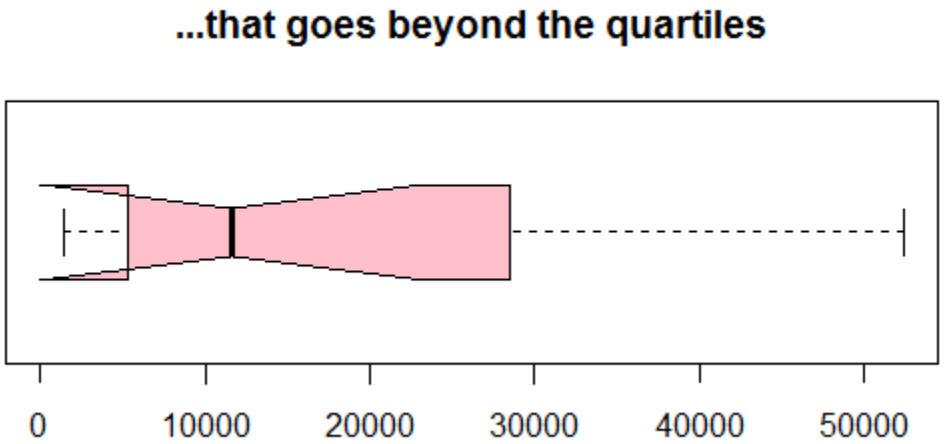
graphics包

boxplot(x, ...)

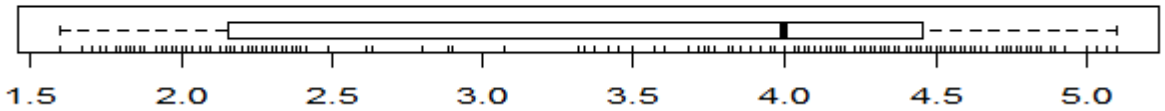
4.3 箱线图



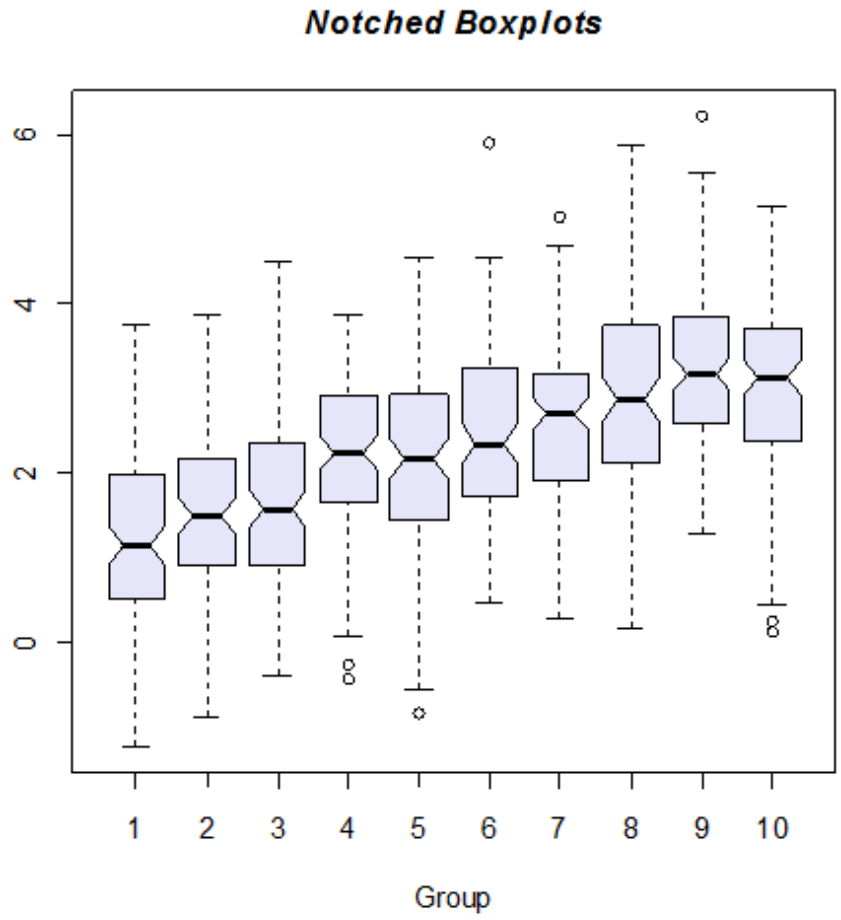
绘制中位数置信区间



添加坐标须



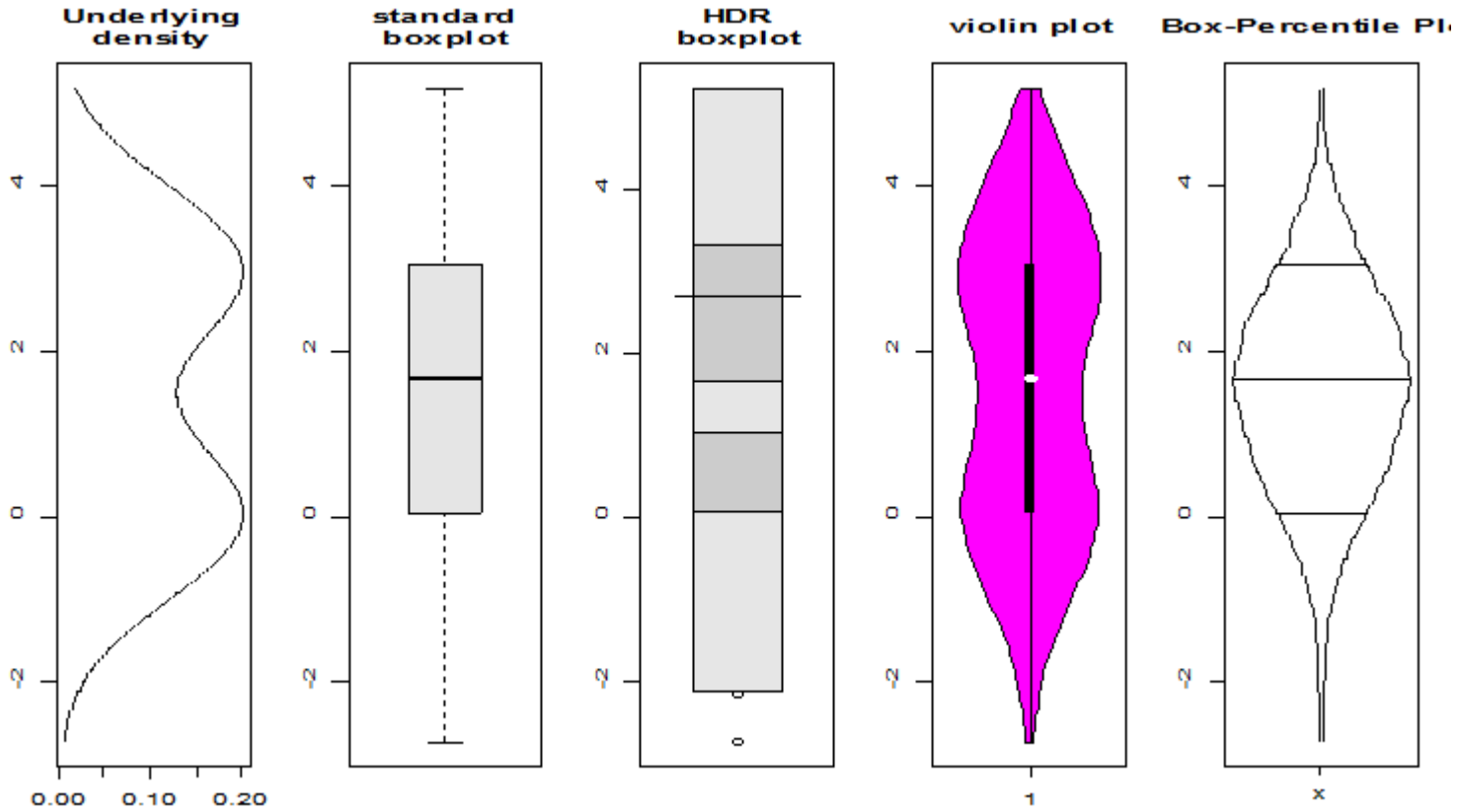
4.3 箱线图



graphics包

boxplot(formula, notch=TURE, ...)

4.3 箱线图



- The boxplot friends

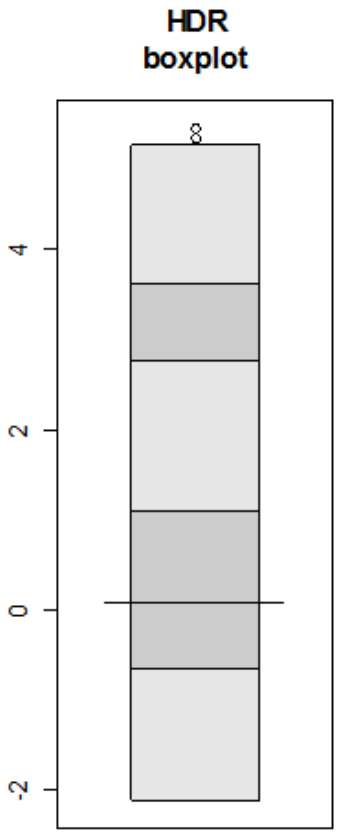
Package "hdrcde","vioplot" &"Hmisc" needed

4.3 箱线图

- **Highest Density Region (HDR) plot**

Hdrcde包

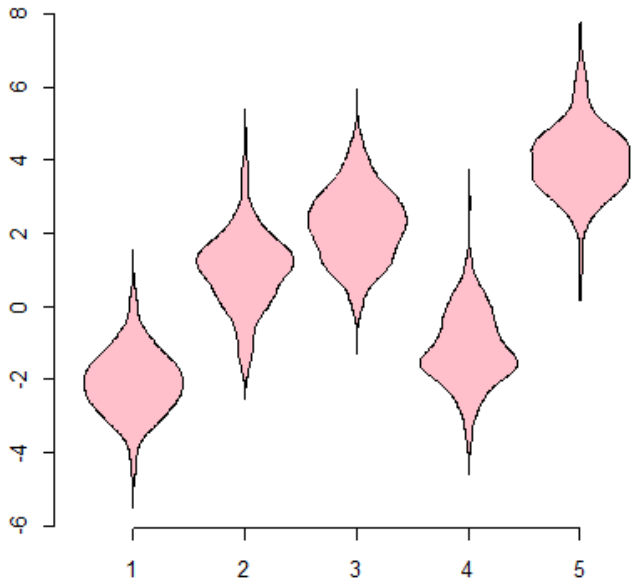
```
hdr.boxplot(x,
 prob = c(99, 50),
h=hdrbw(BoxCox(x,lambd
a),mean(prob)), lambda=1,
 boxlabels = "",
col = gray((9:1)/10),
main="", xlab="", ylab="",
pch=1, ...)
```

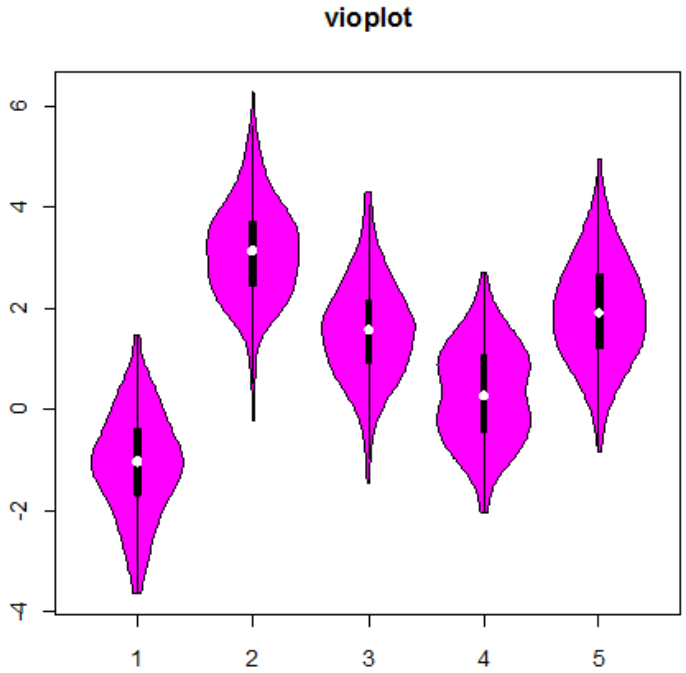


4.3 箱线图

- **The Violin plot**

Vioplot

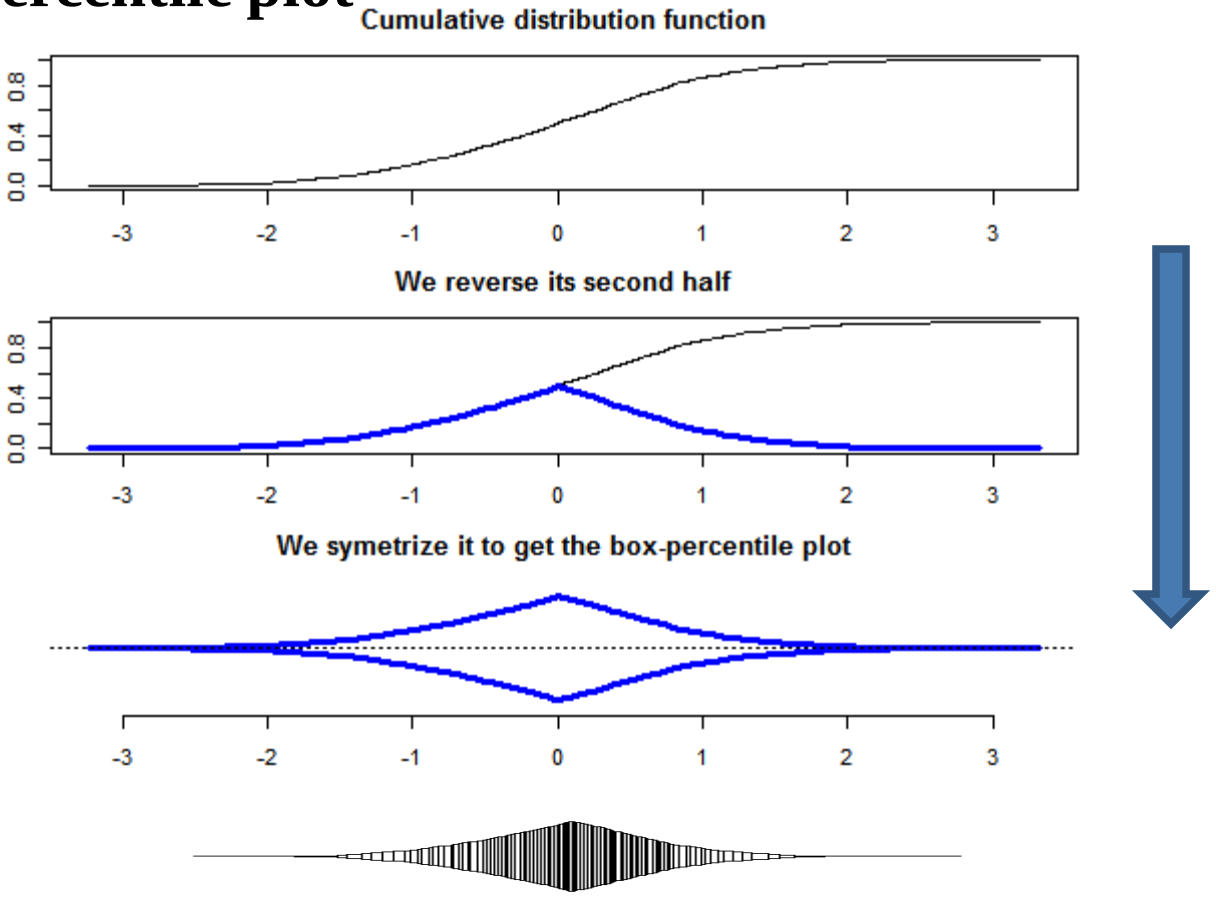




UsingR包

4.3 箱线图

- **The Box-Percentile plot**

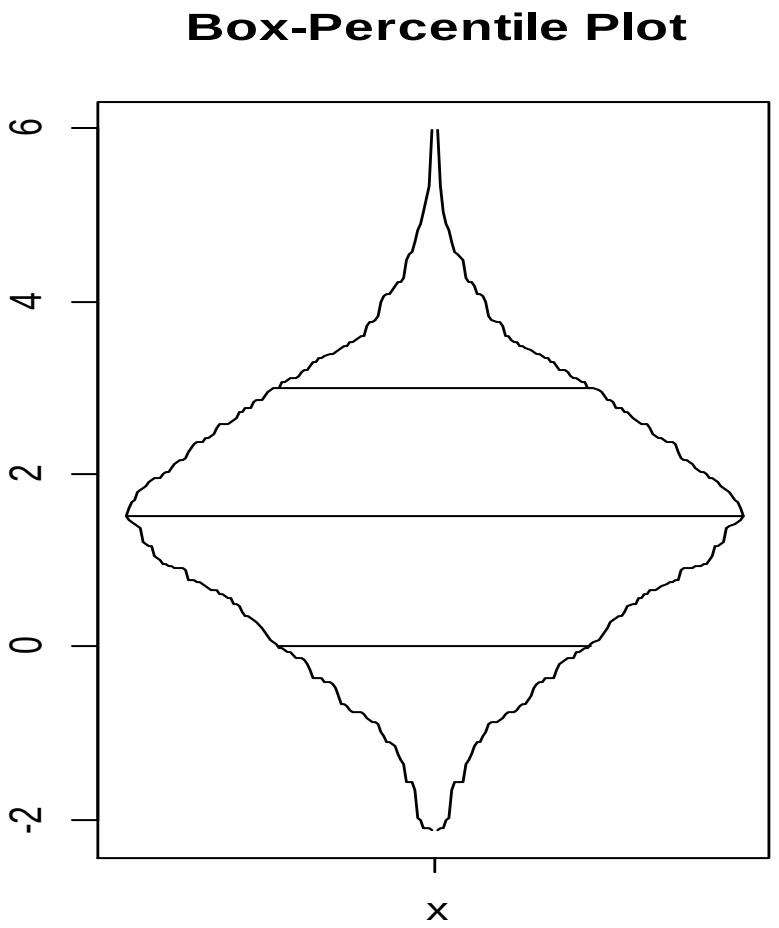


4.3 箱线图

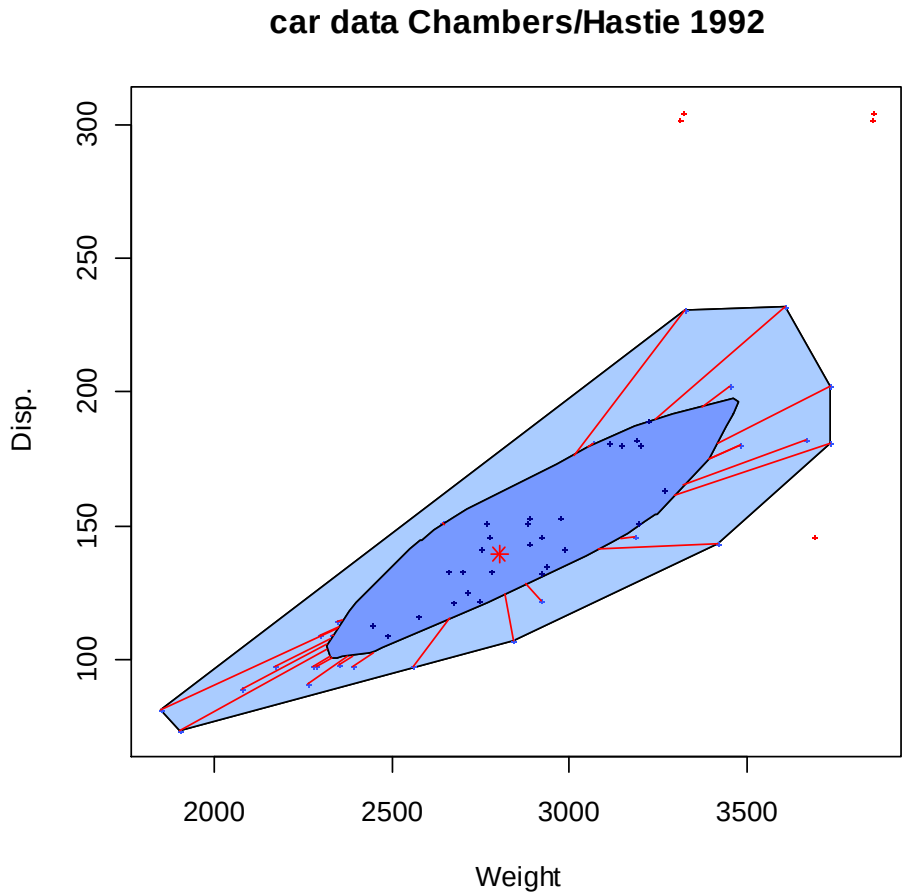
- **The Box-Percentile plot**

Hmisc包

bpplot(x)

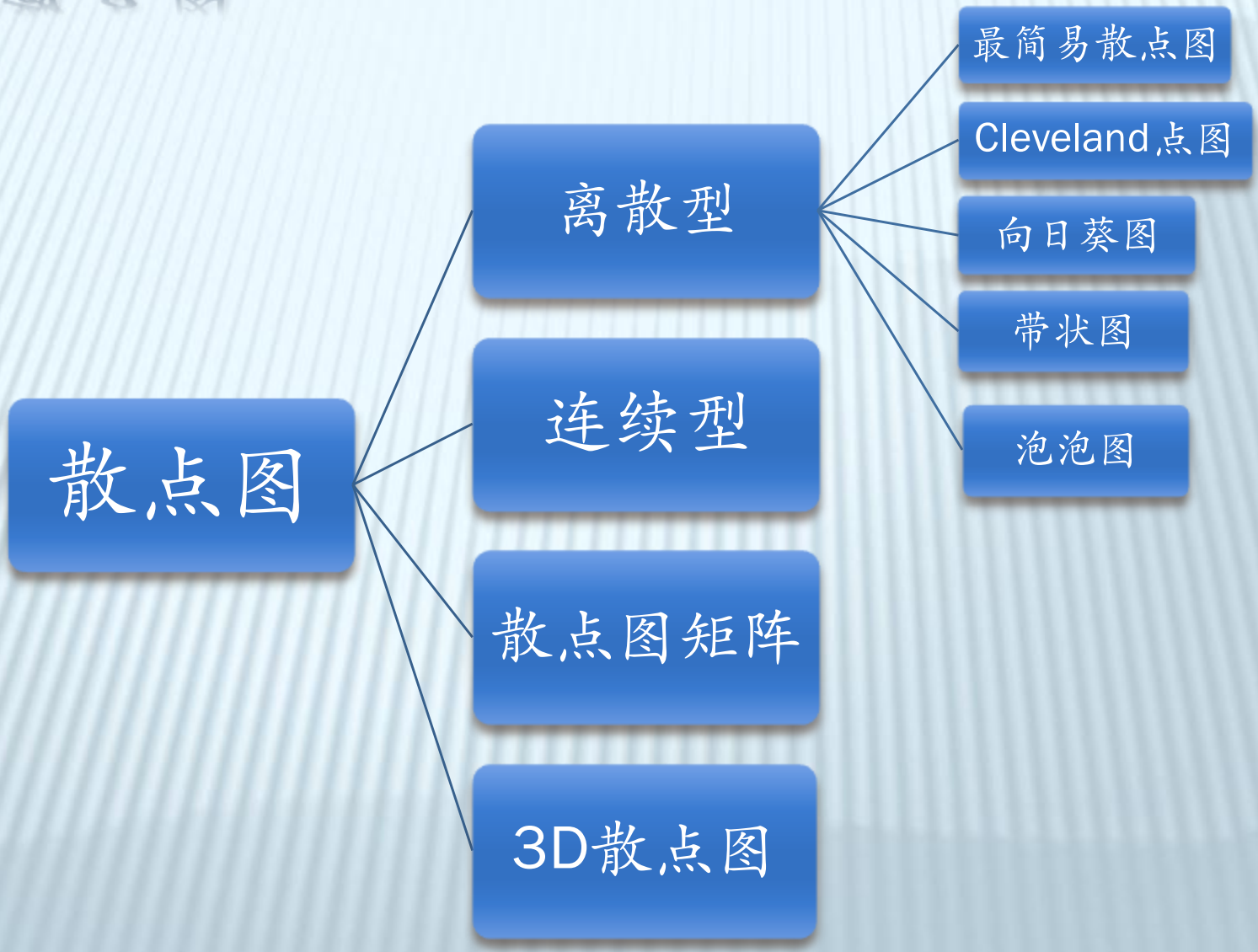


4.3 箱线图

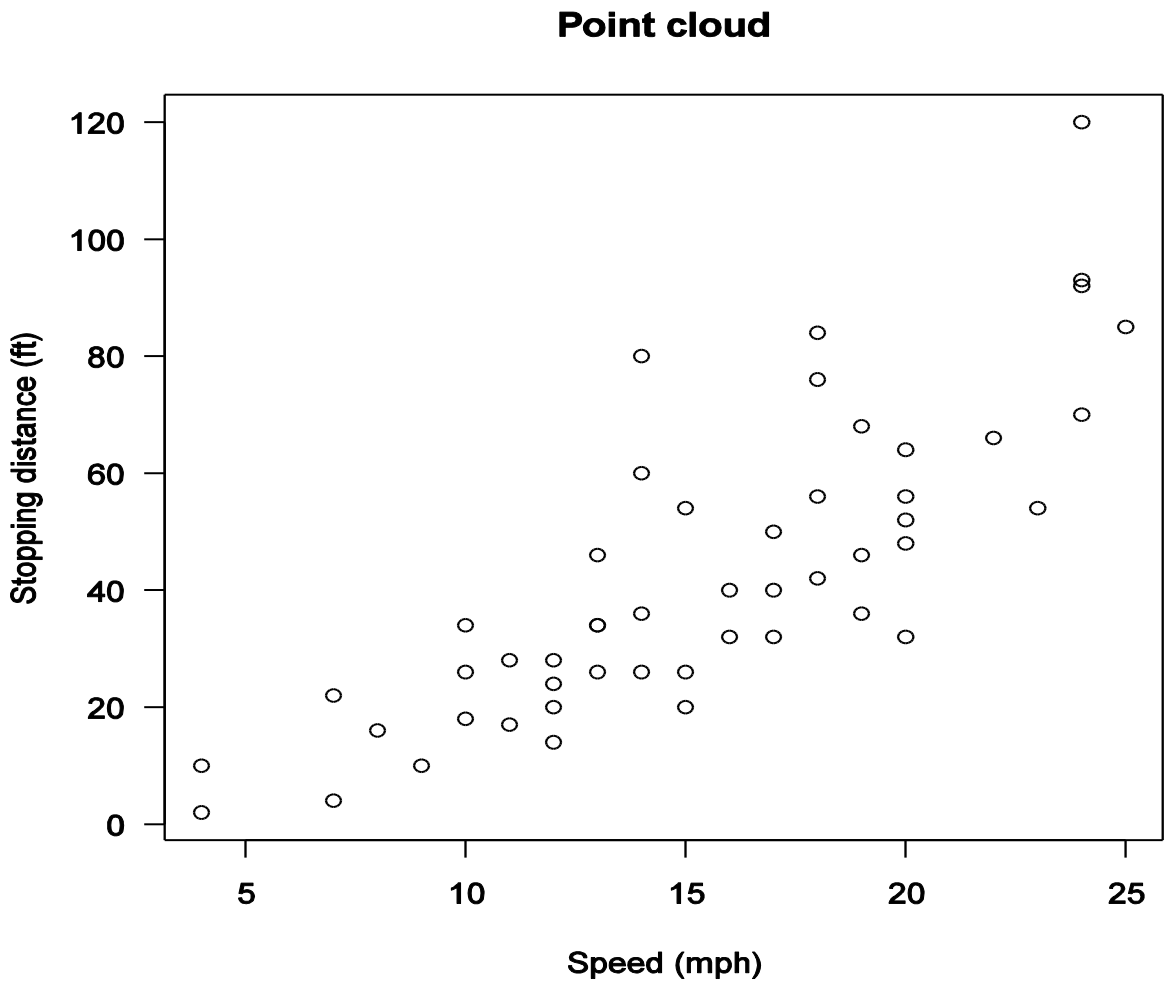


bagplot

4.4 散点图



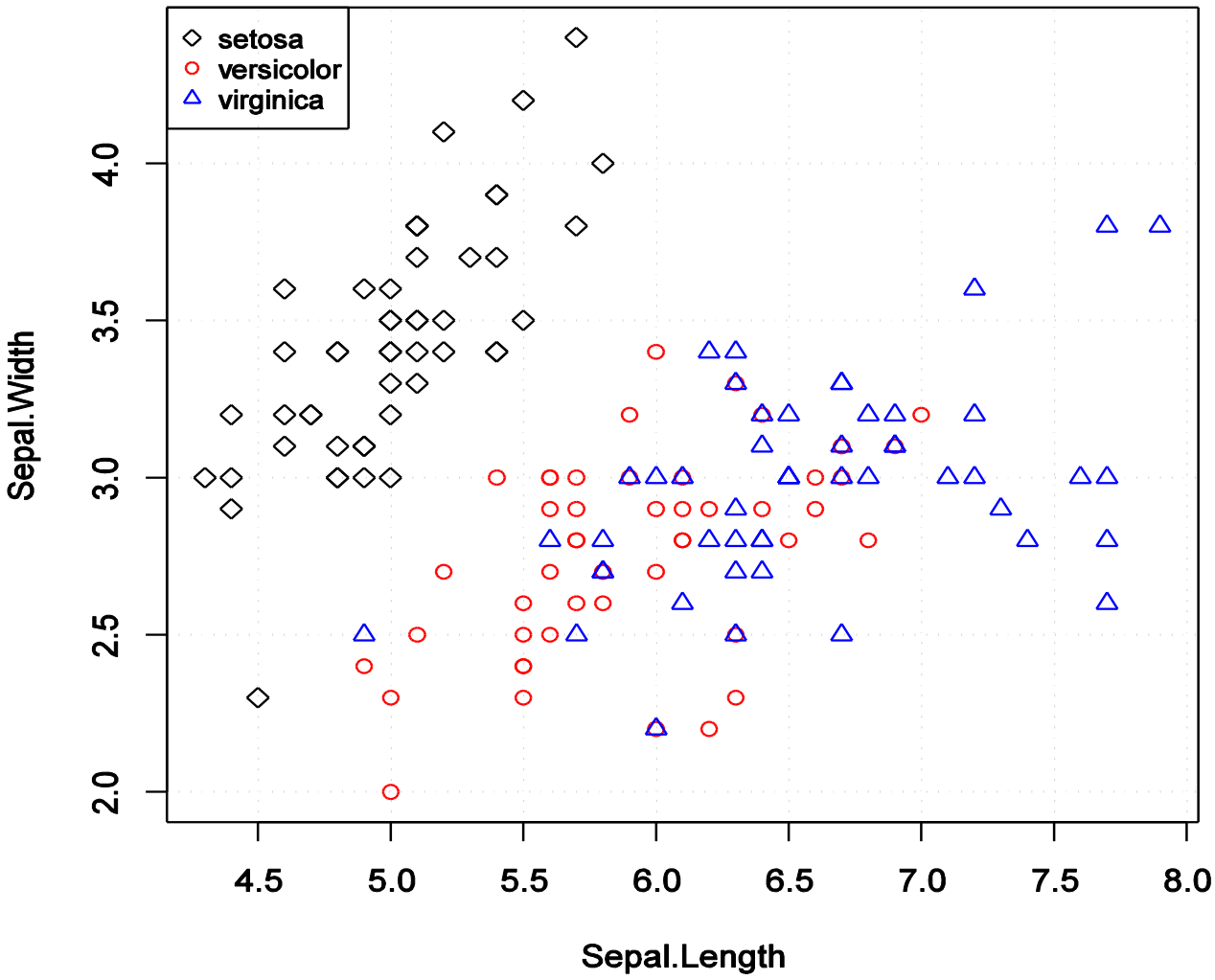
4.4.1 离散型



- 最简易散点图

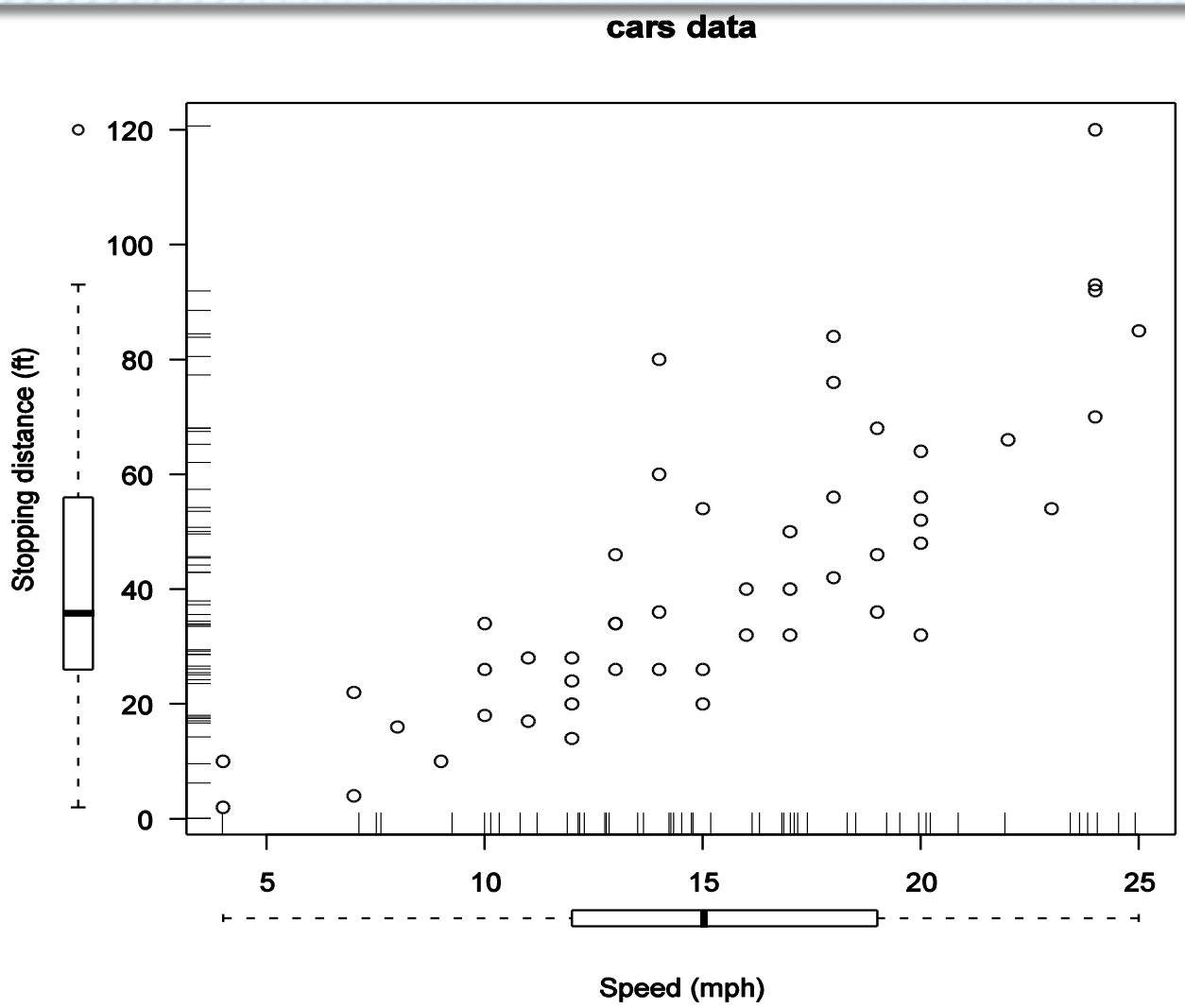
graphics包
plot(x,y, …)

4.4.1 离散型



改变其中的参数，如pch和col等

4.4.1 离散型

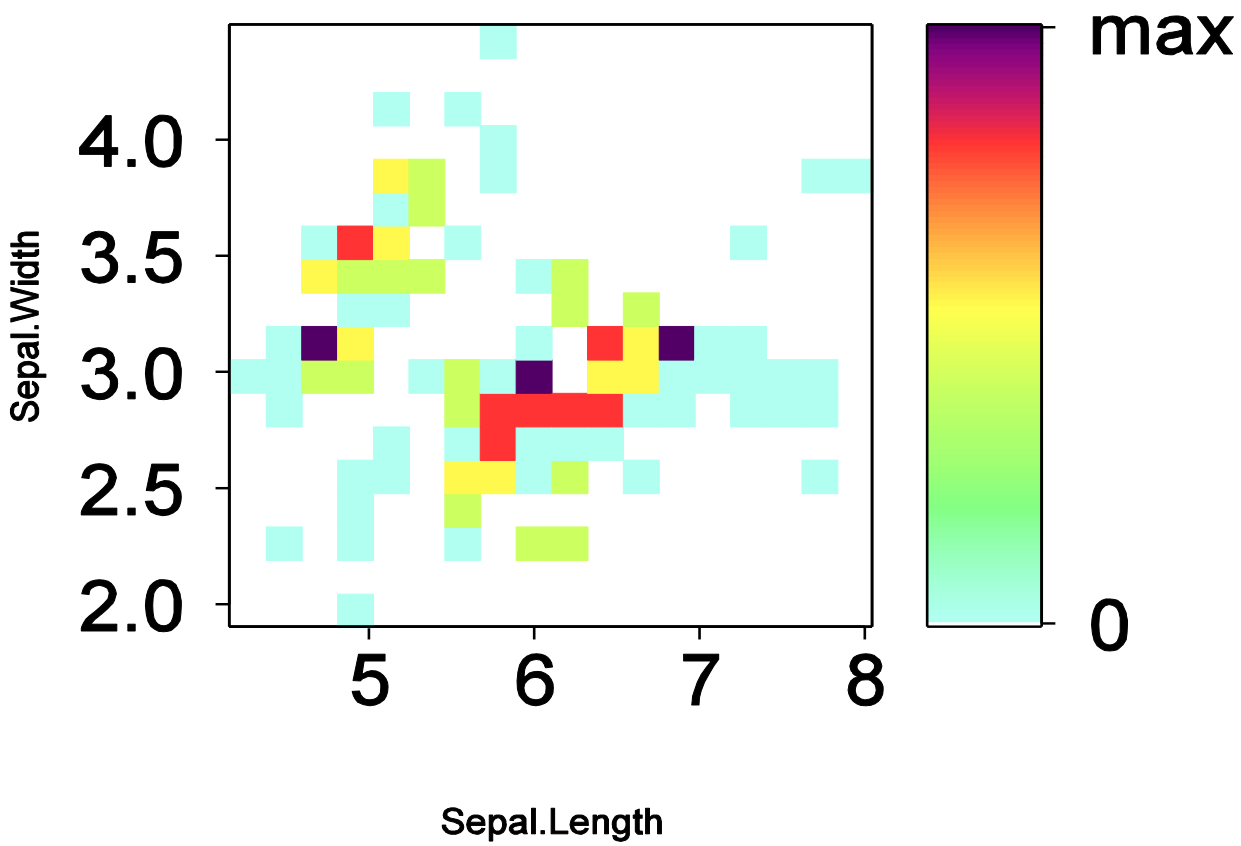


改变其中的参数，如pch和col等

或者添加其他的元素，如rug()等

再或者与箱线图等联用

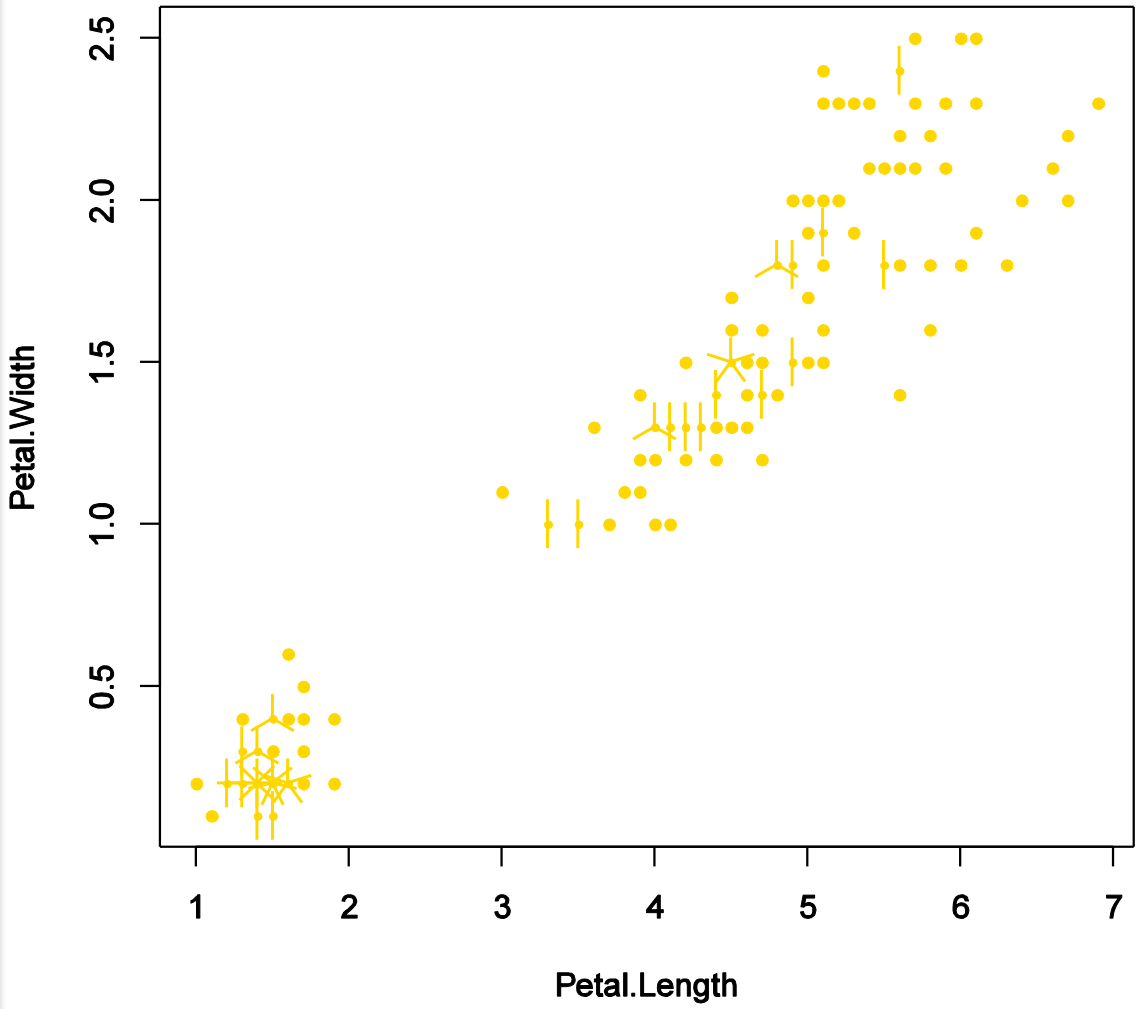
4.4.1 离散型



IDPmisc包
iplot( )

用颜色来展示数据的密度，很适合大数据展示

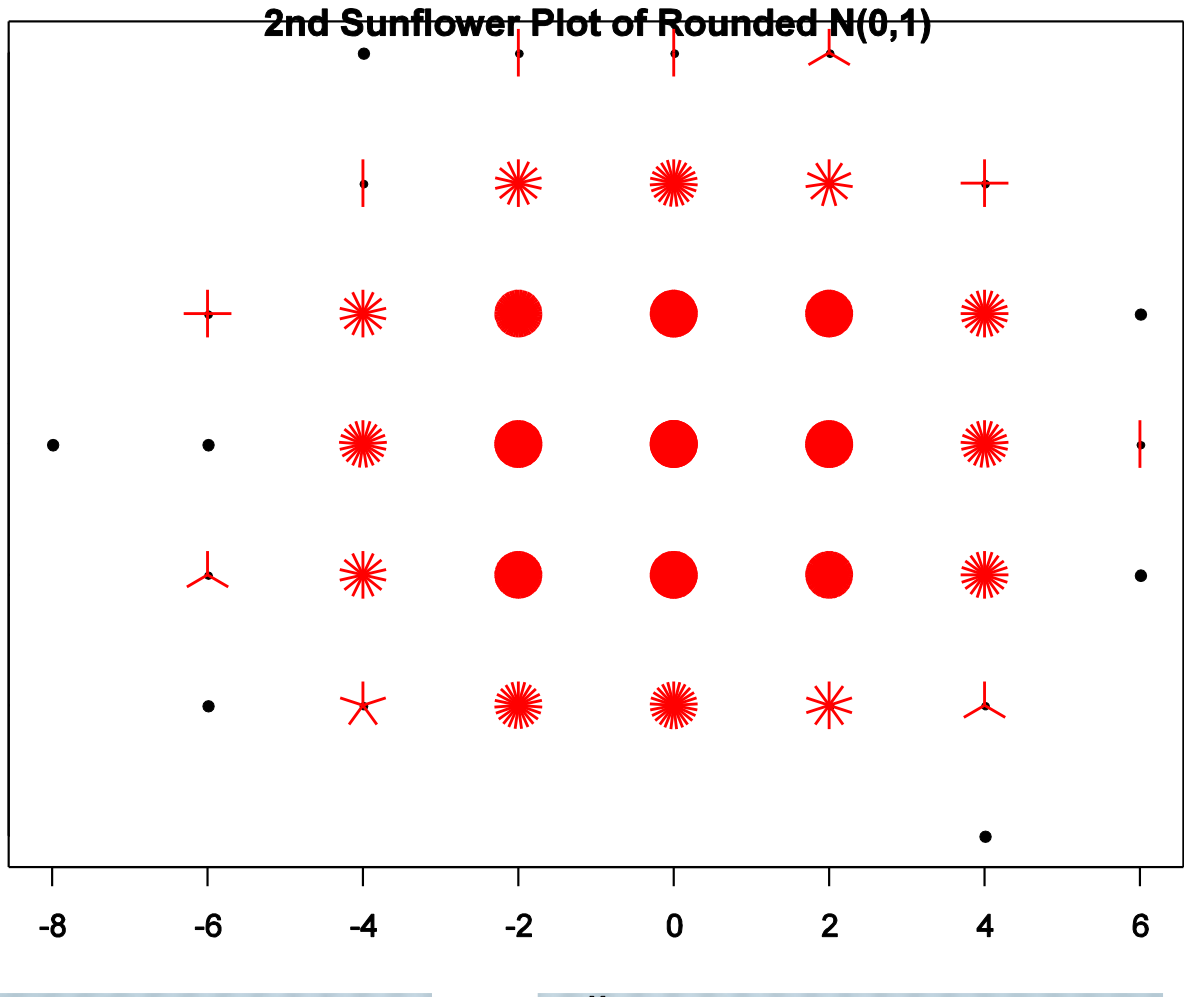
4.4.1 离散型



- 向日葵图

graphics包
sunflower(x,y, …)

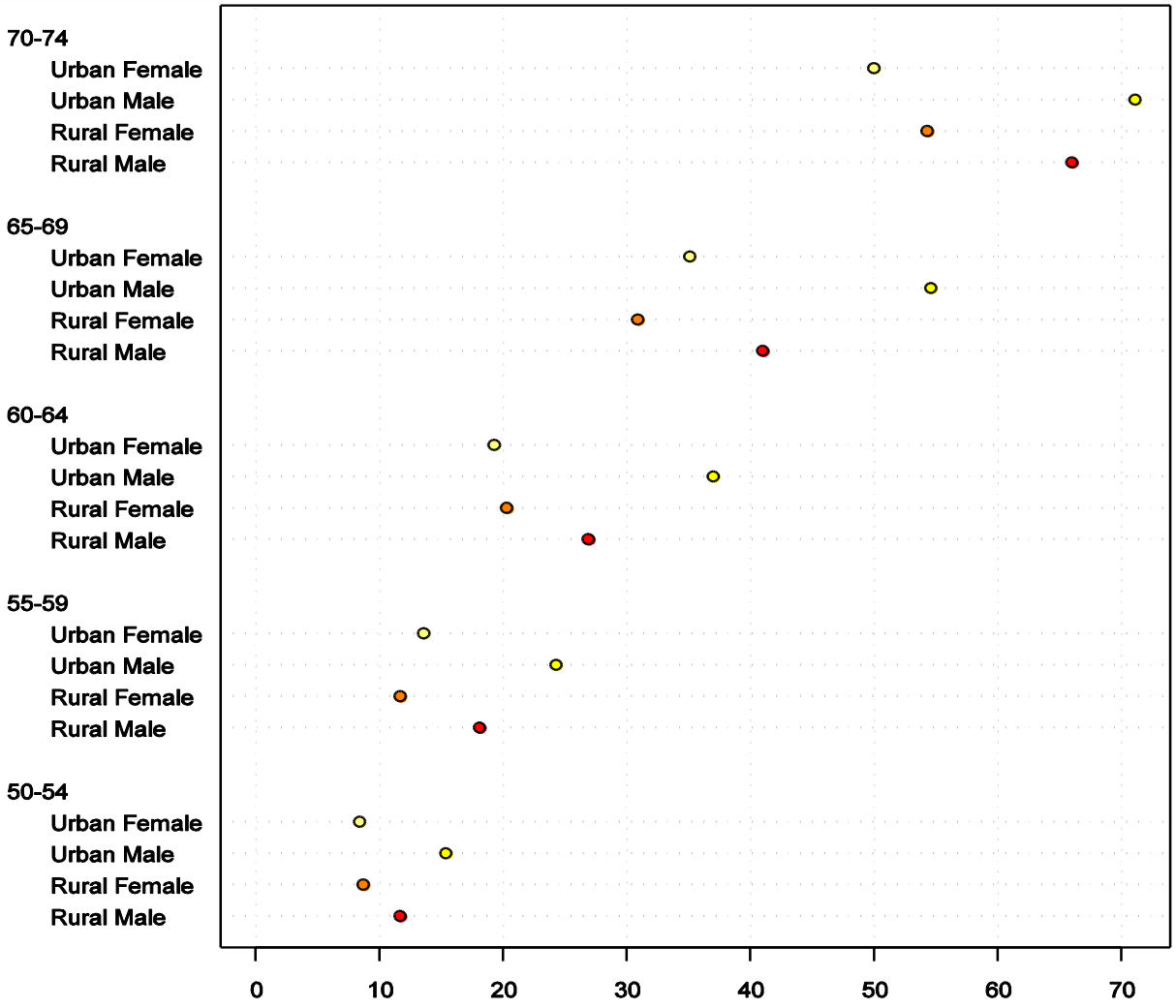
4.4.1 离散型



- 向日葵图

数据密度小时，花瓣很形象展示数据；
数据密度过大时，花瓣就成了实心圆点。

4.4.1 离散型

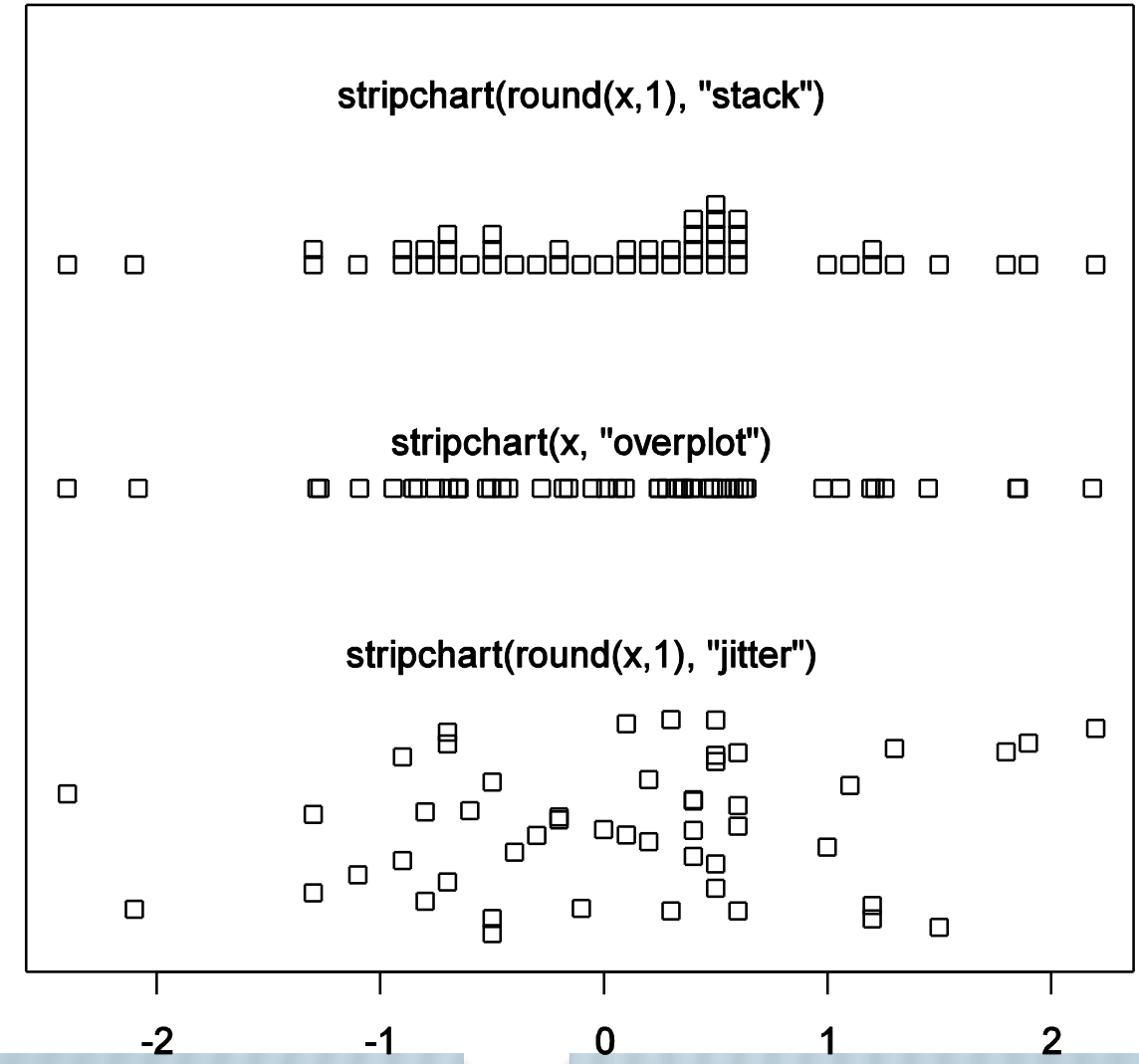


- Cleveland点图

graphics包
dotchart (x,y, …)

清晰、简洁、
可比性强

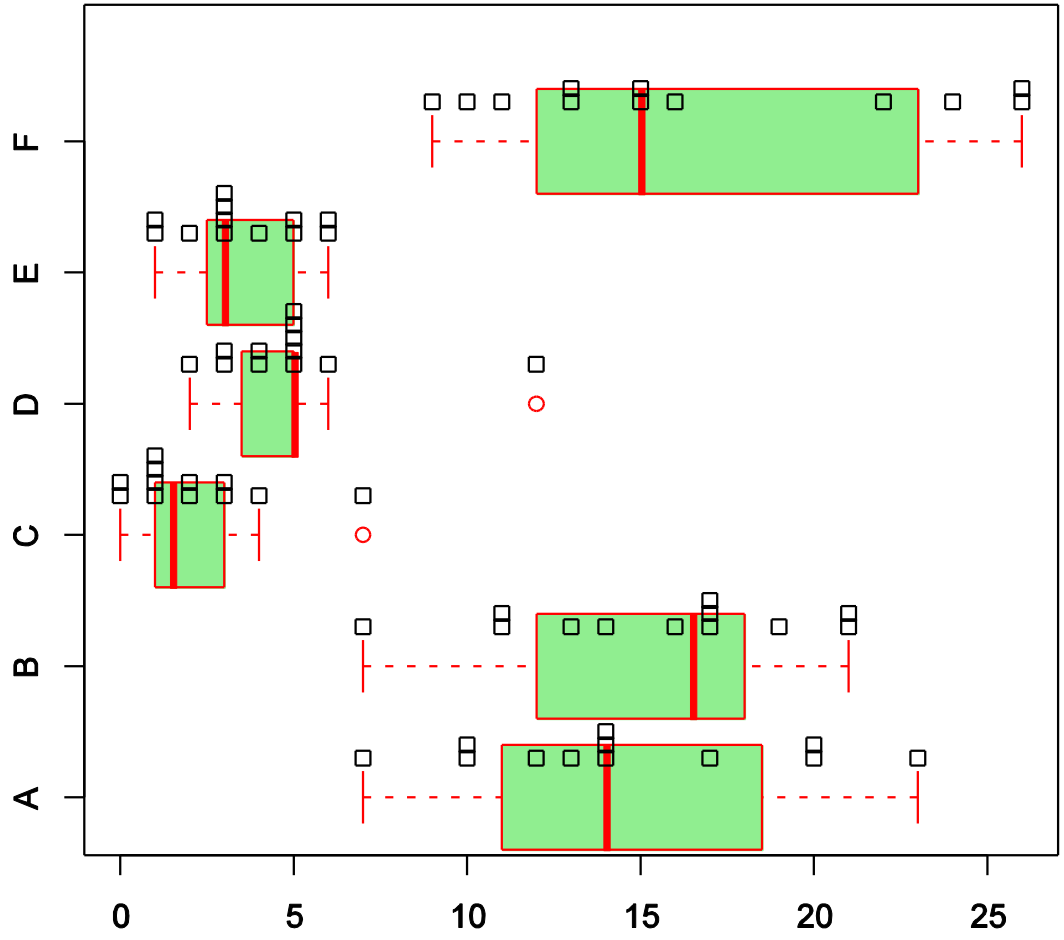
4.4.1 离散型



- 带状图

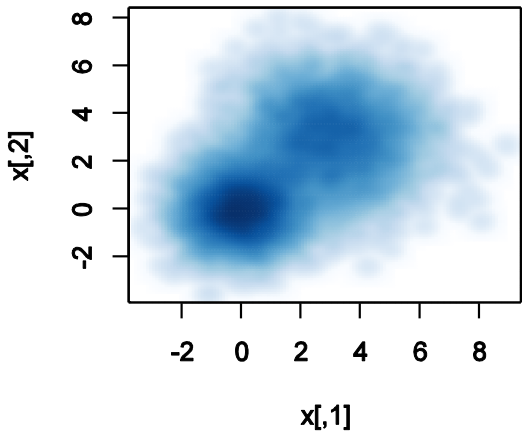
graphics包
stripchart( )
(x,method=… , …)

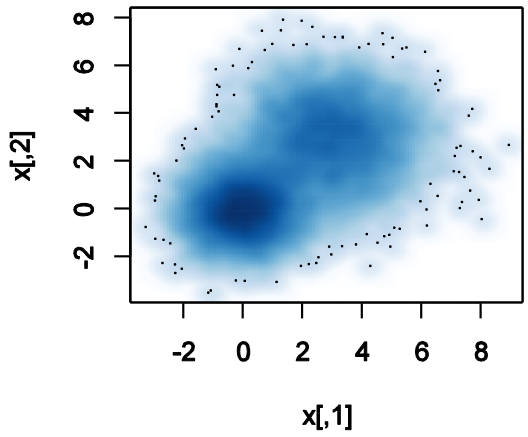
4.4.1 离散型



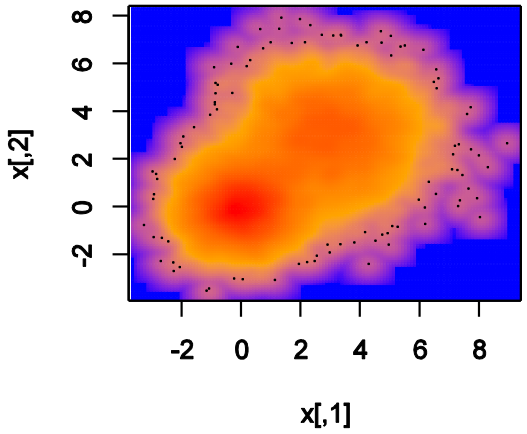
与坐标轴须有共通之处，
与其他图联用更全面展示数据。

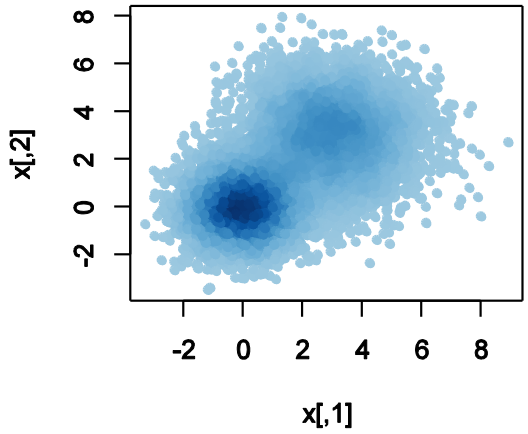
4.4.2 连续型





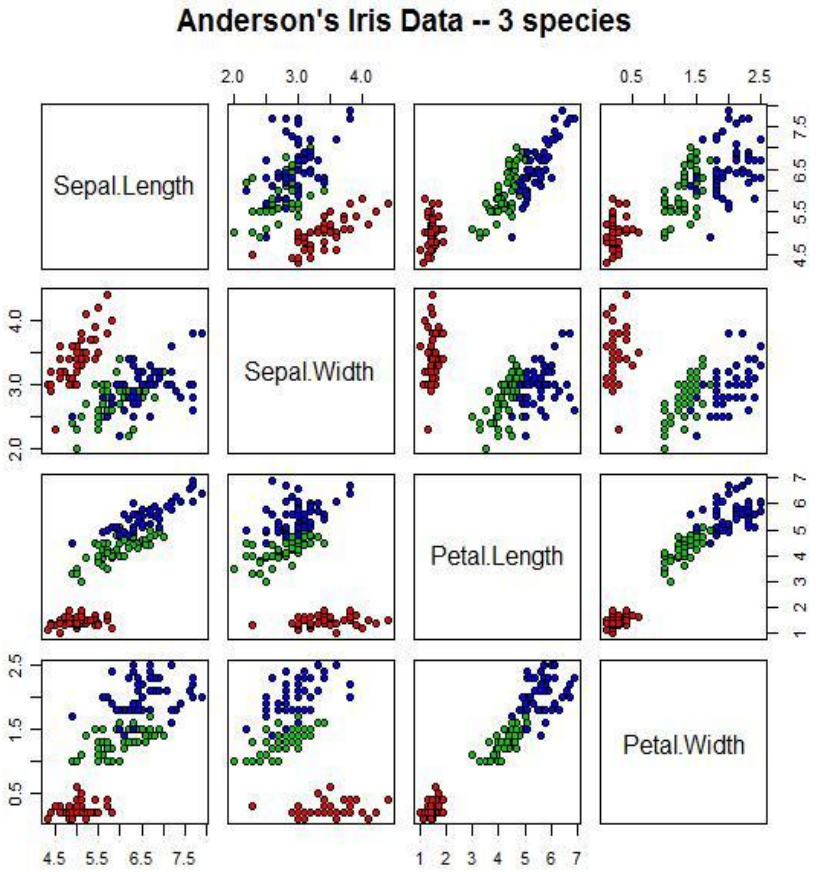
graphics包
smoothScatter( )



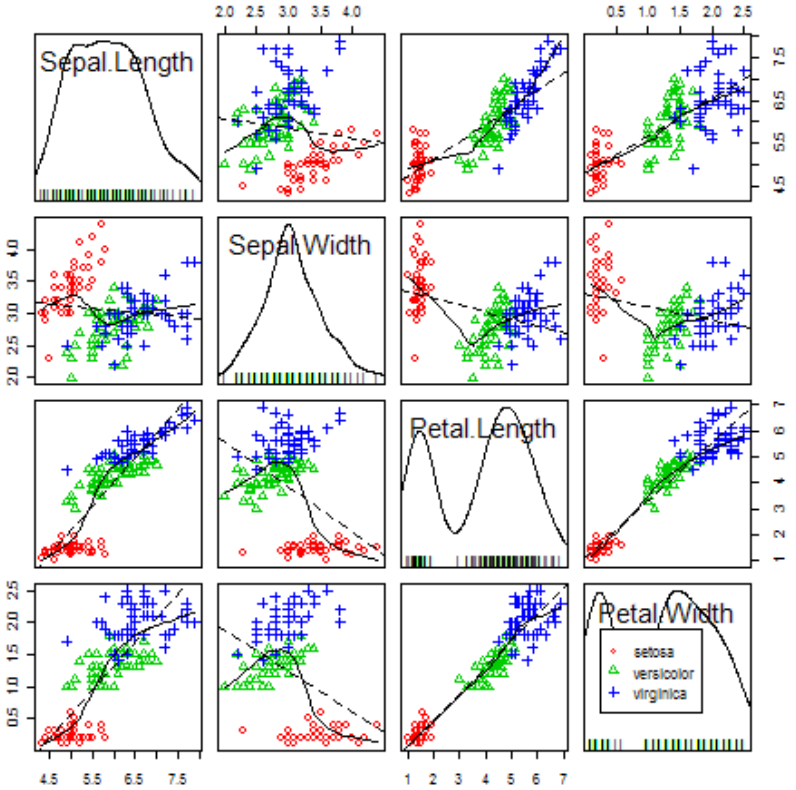


颜色代表
点密度

4.4.3 散点图矩阵

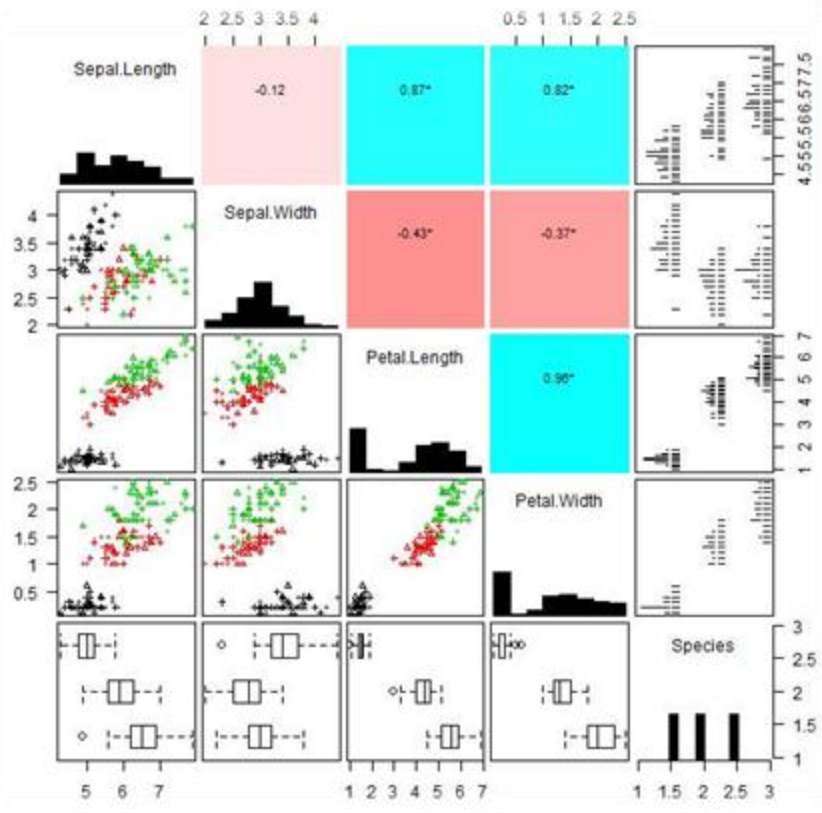


graphics包pairs( )

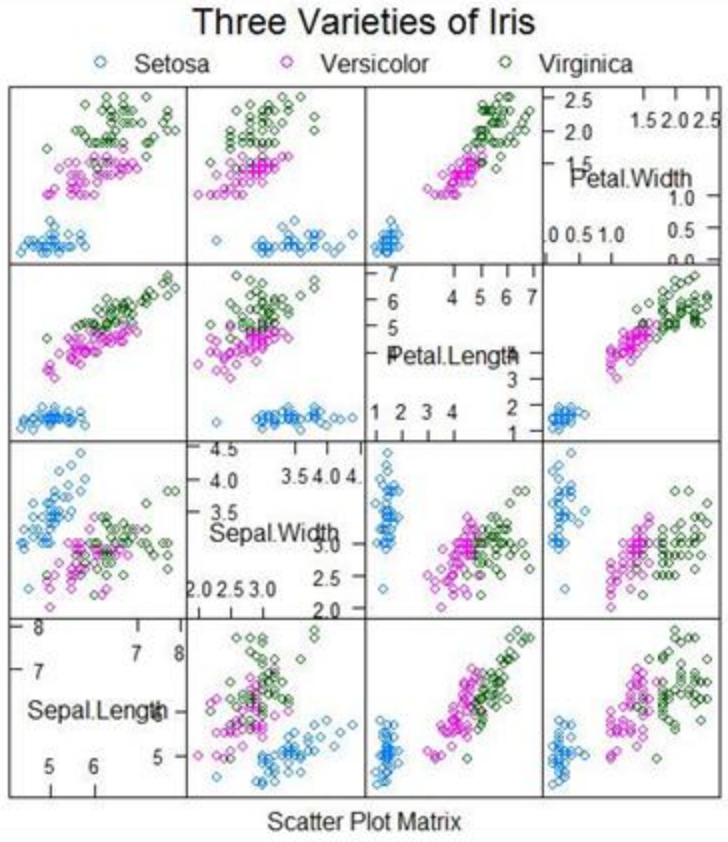


car包spm( )

4.4.3 散点图矩阵

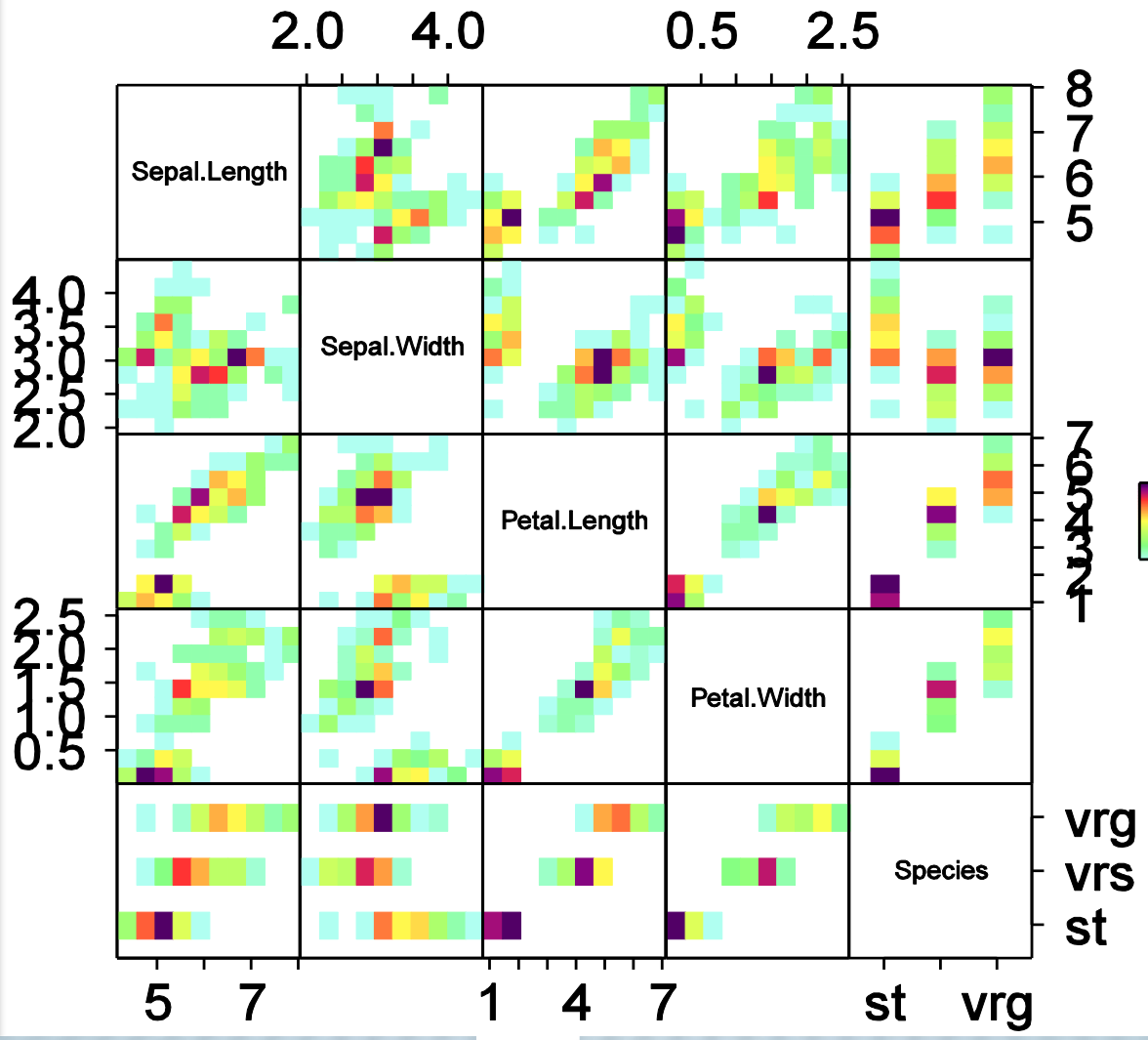


YaleToolkit包gpairs()



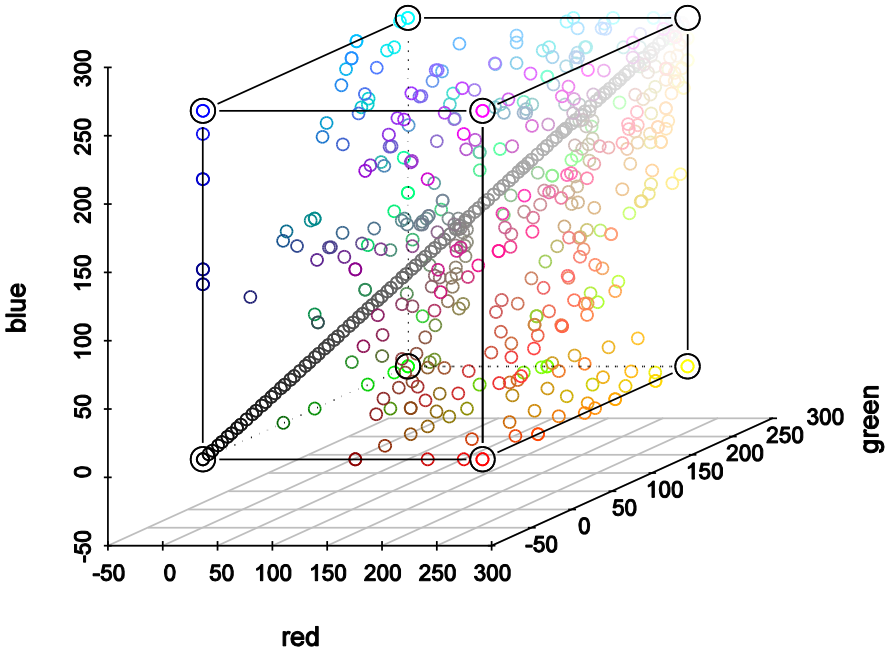
Lattice包splom()

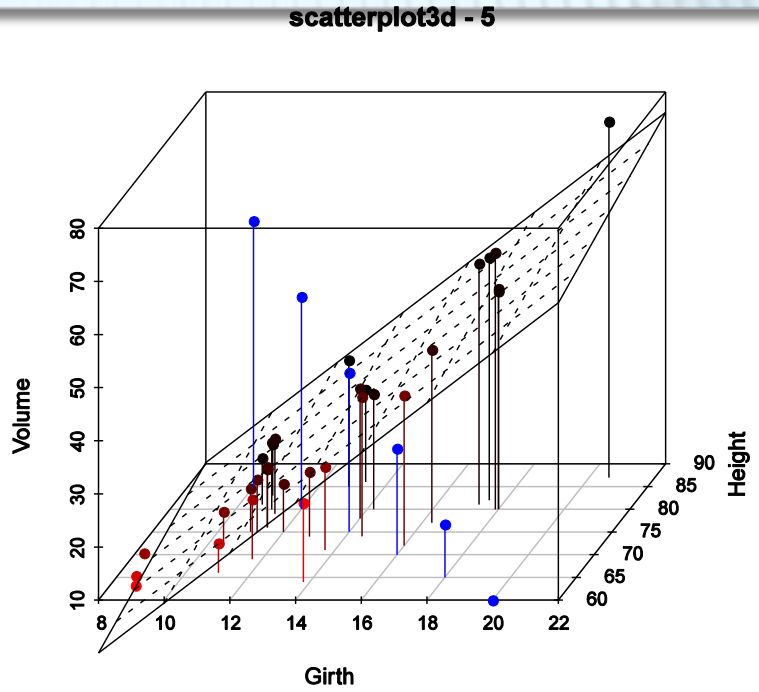
4.4.3 散点图矩阵



IDPmisc包
ipairs()

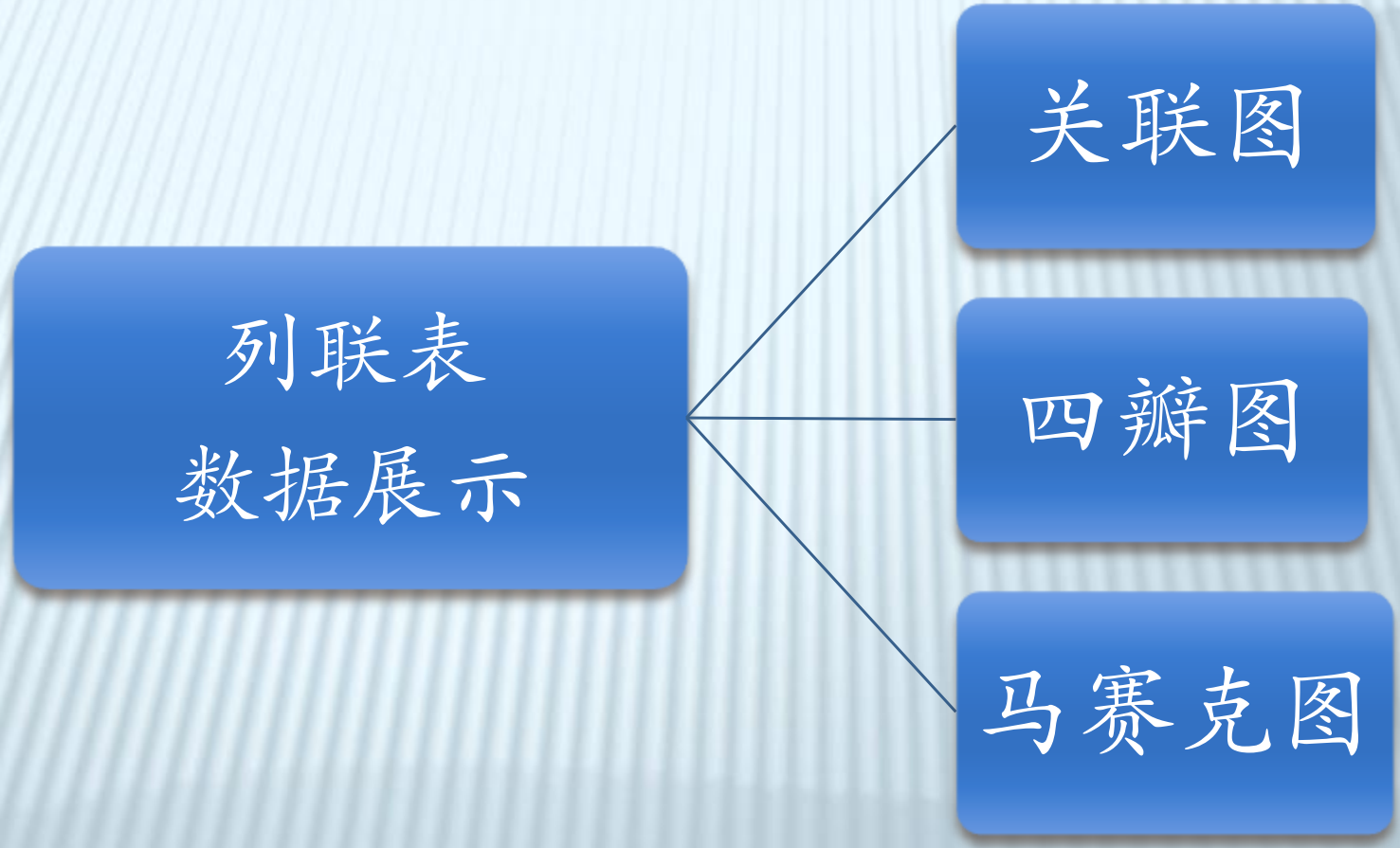
4.4.4 3-D散点图



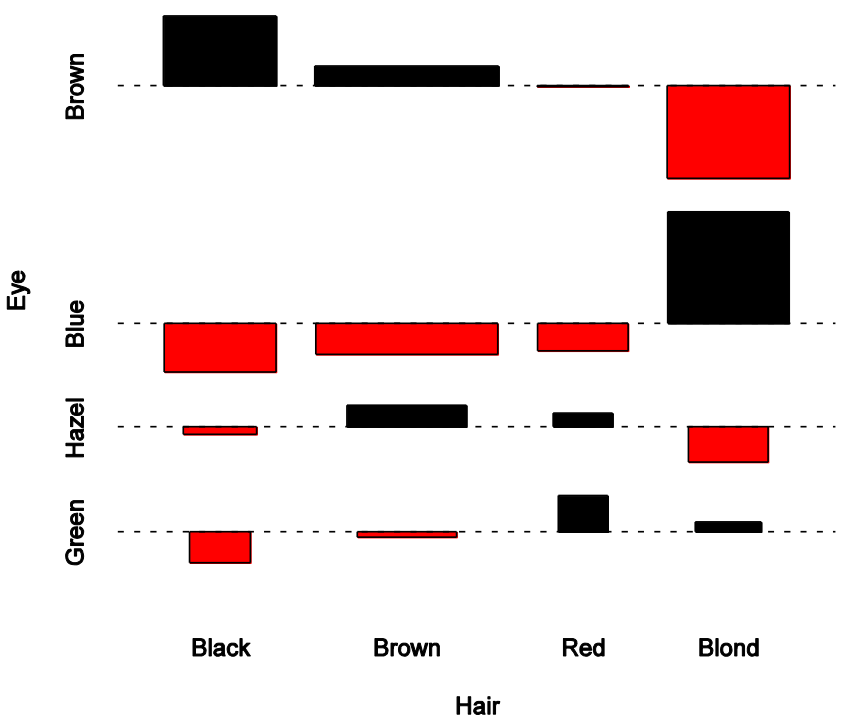


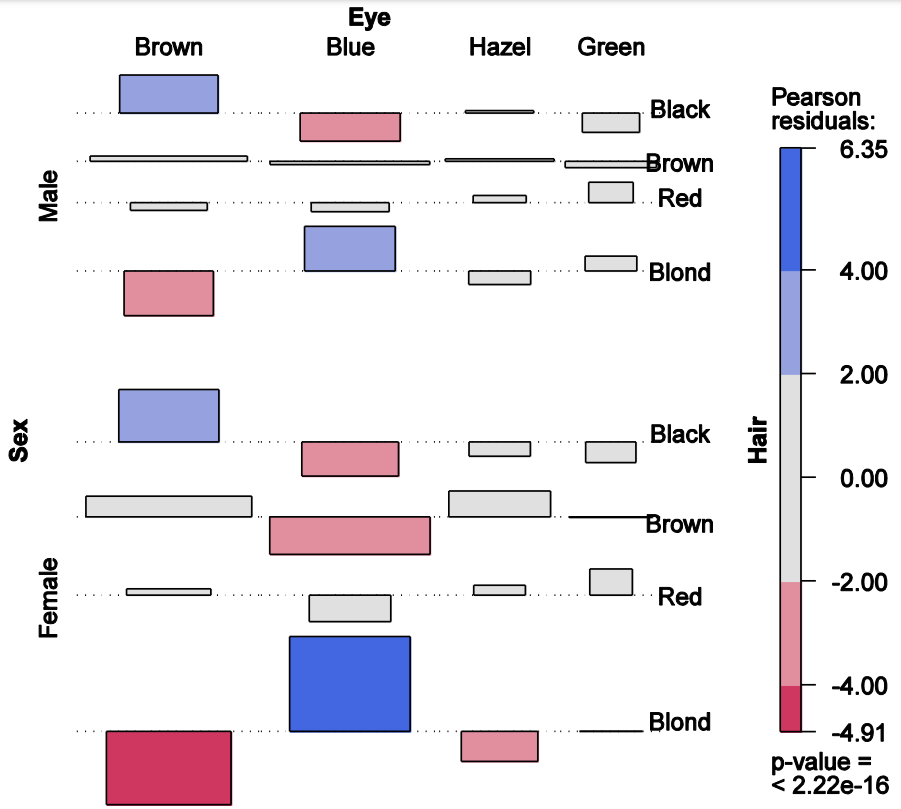
scatterplot3d包scatterplot3d()二维散点图的三维扩展

4.5 列联表数据展示



4.5.1 关联图

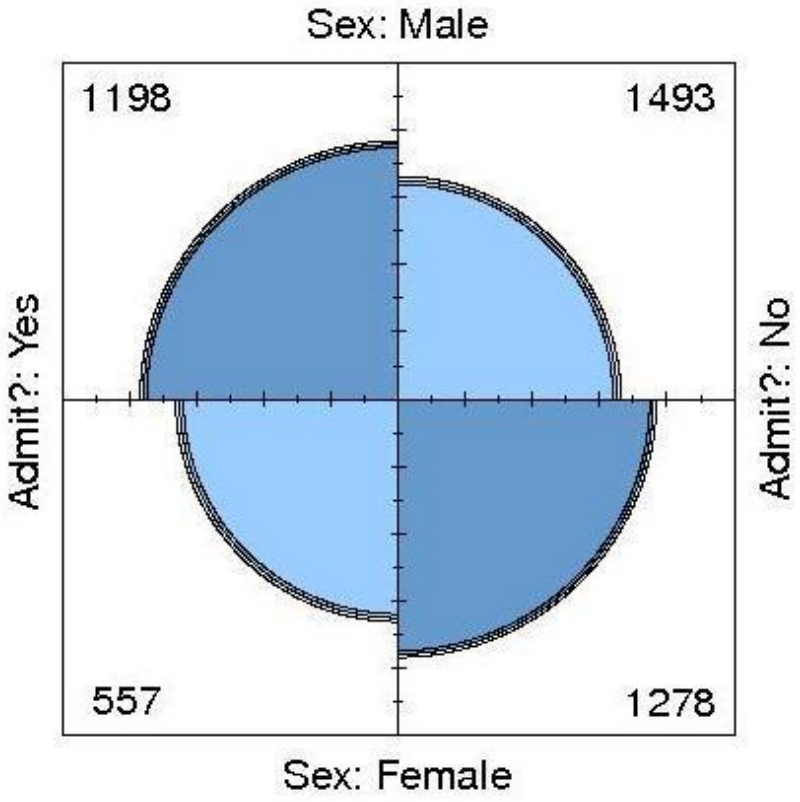




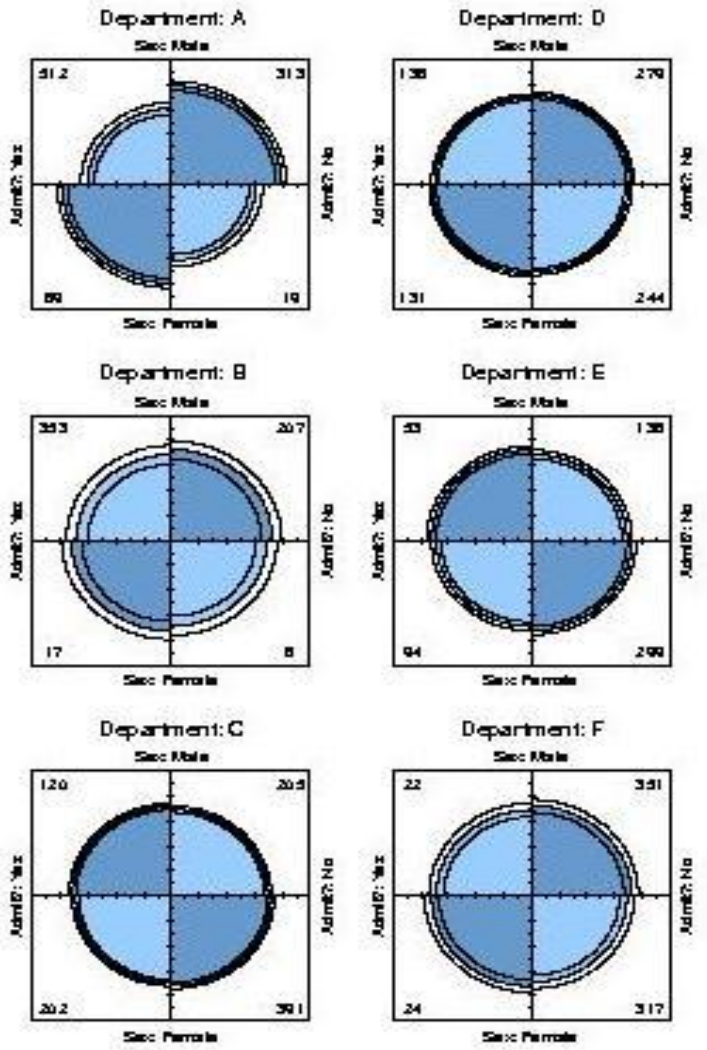
graphics包assocplot()

vcd包assoc()

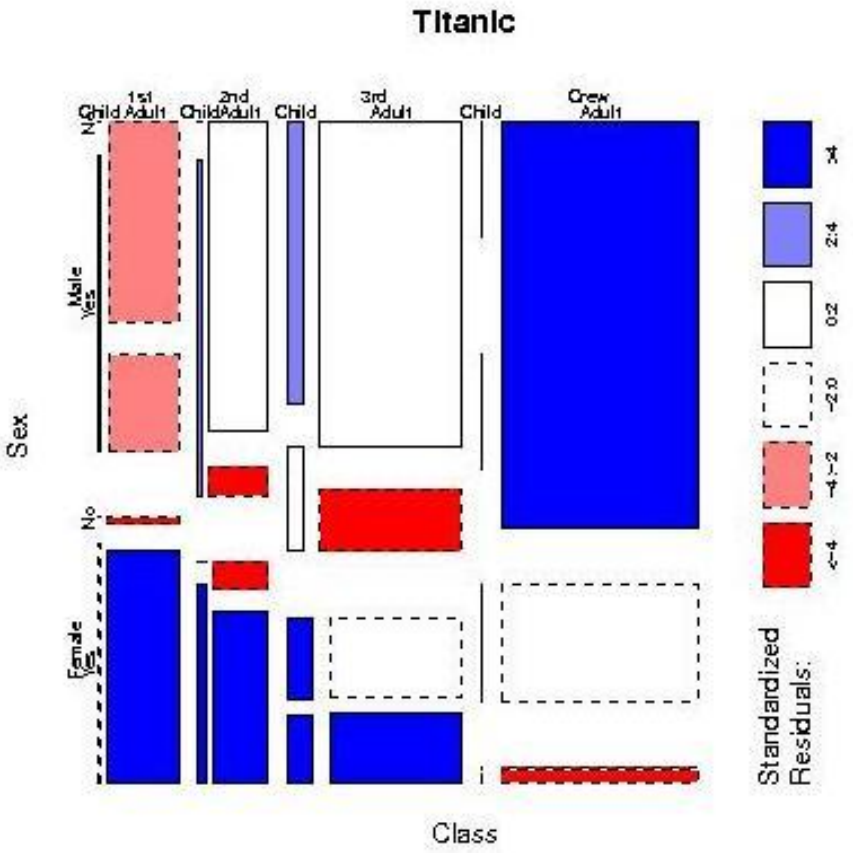
4.5.2 四瓣图



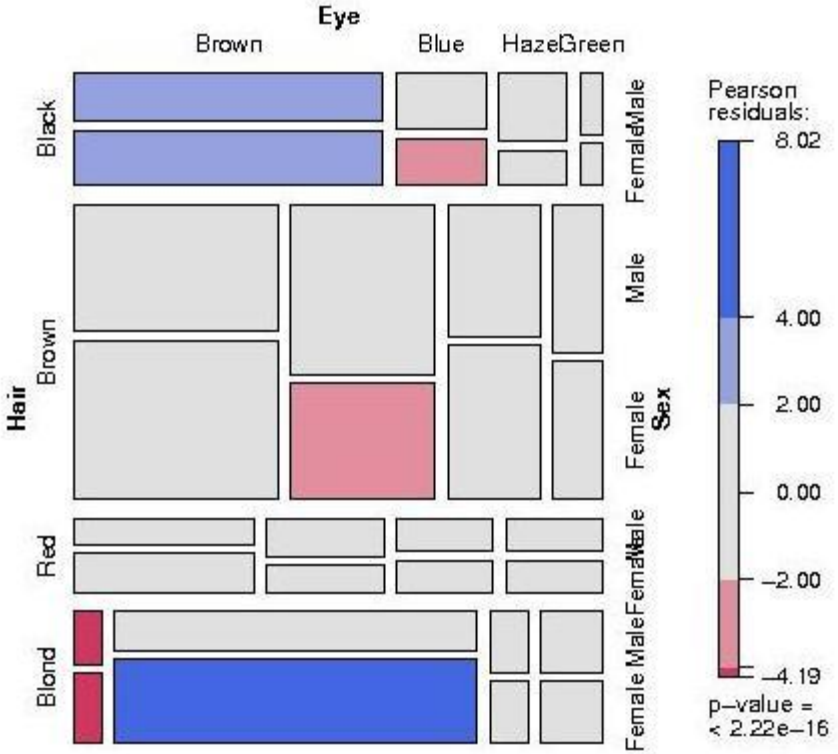
graphics包
fourfoldplot( )



4.5.3 马赛克图

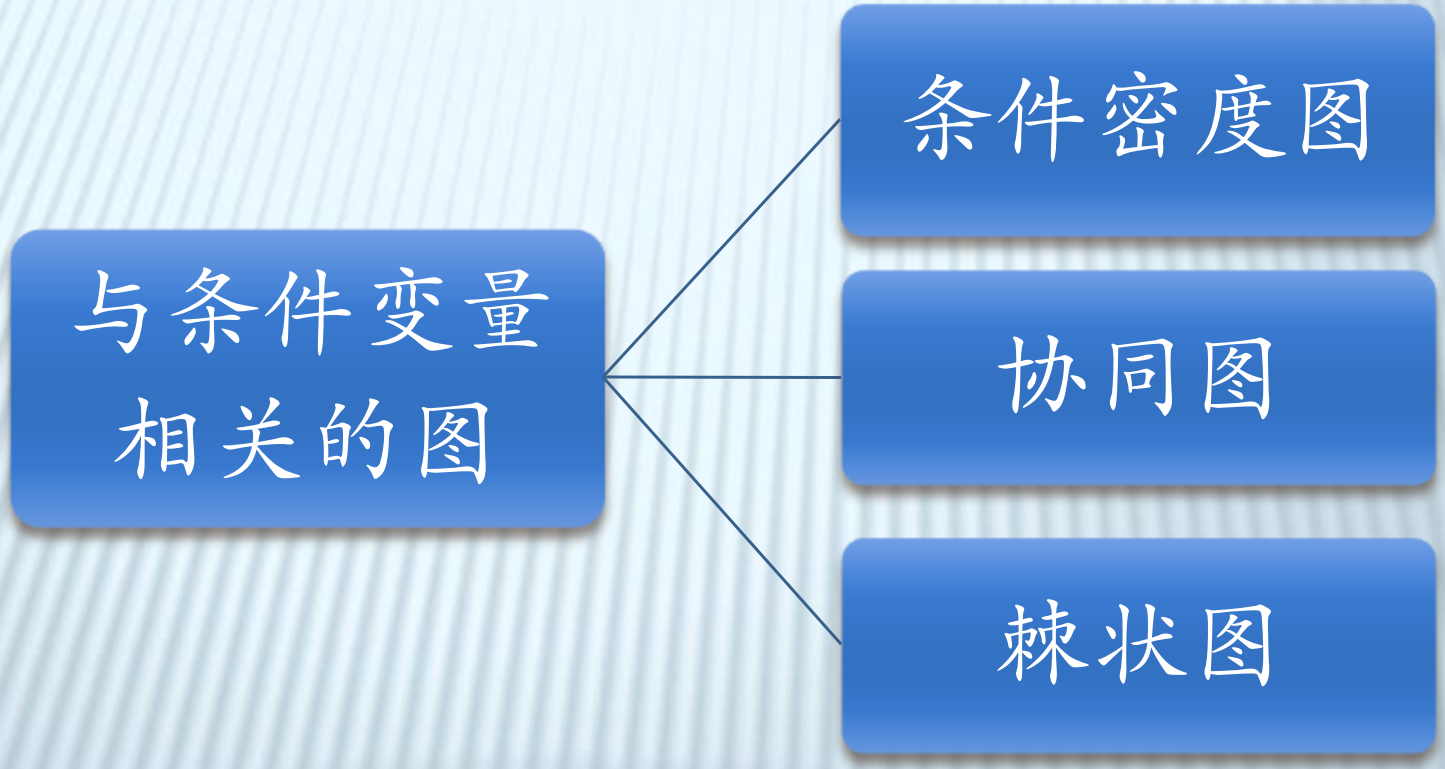


graphics包mosaic()

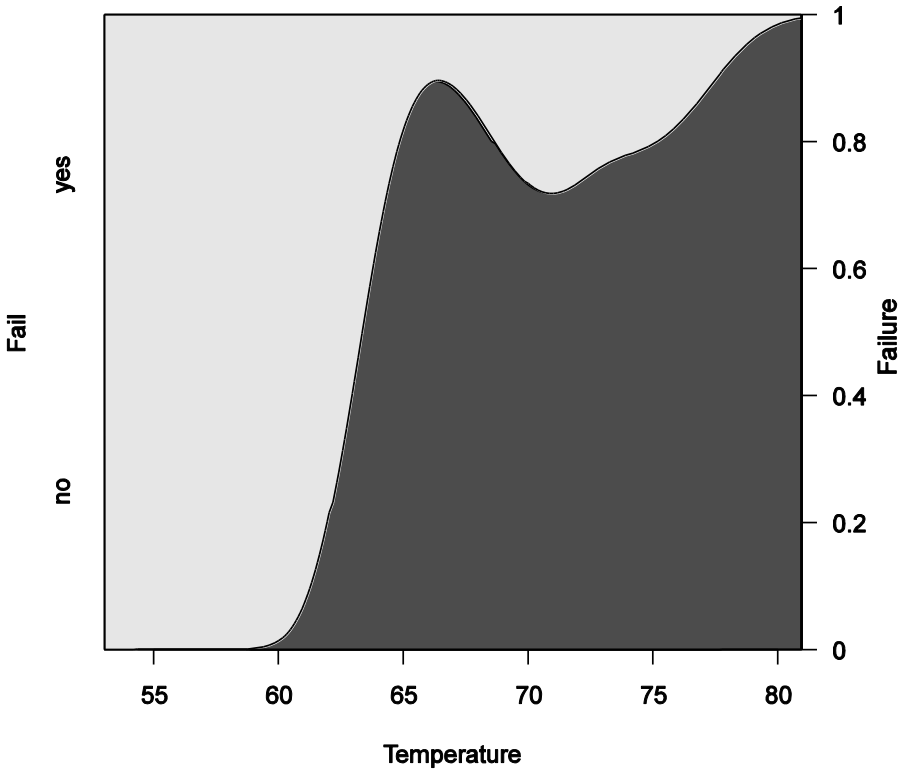


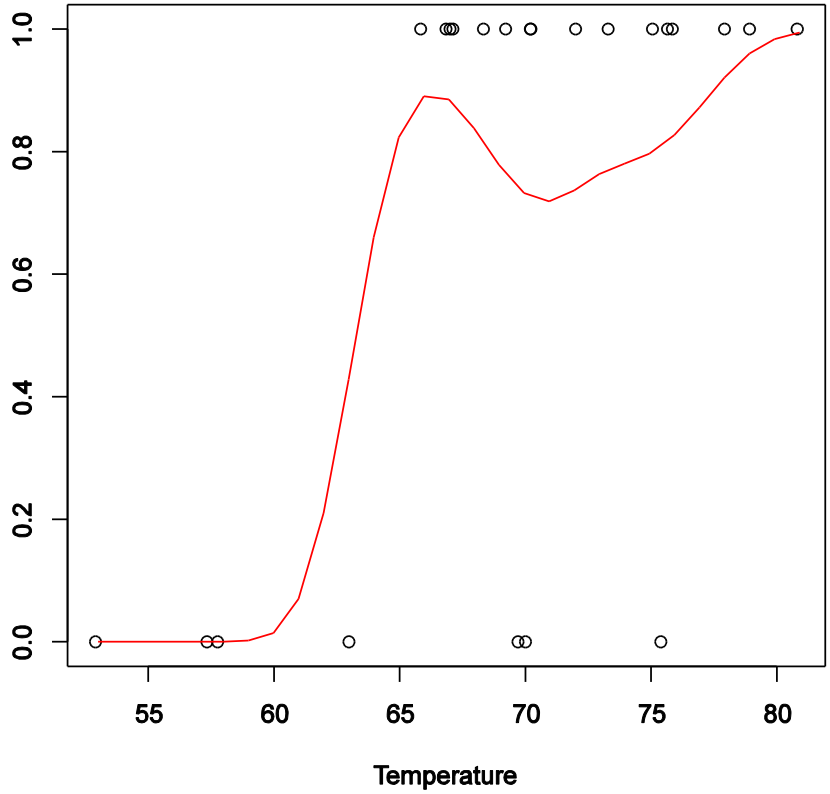
vcd包mosaicplot()

4.6 与条件变量相关的图



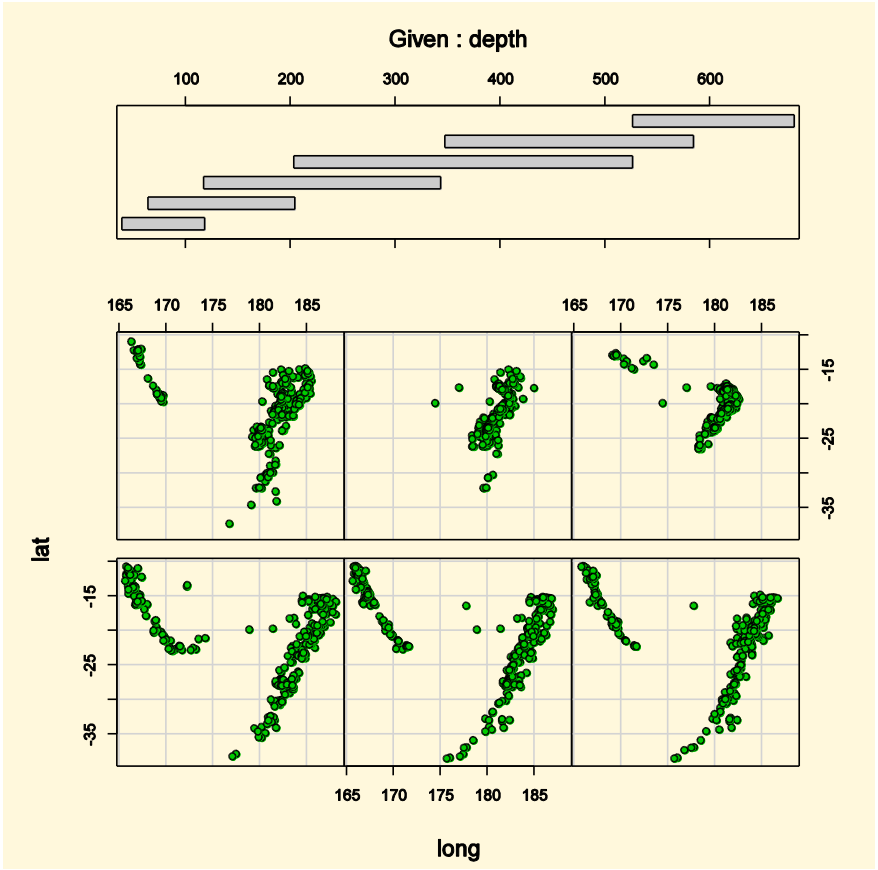
4.6.1 条件密度图

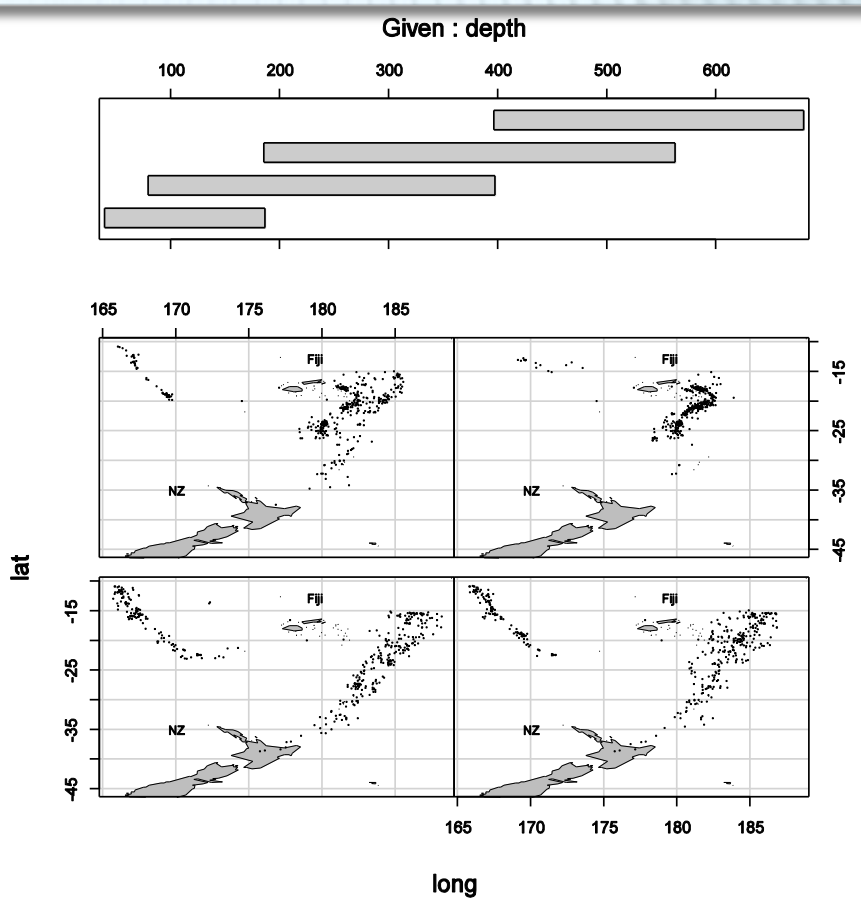




graphics包cdplot()

4.6.2 协同图

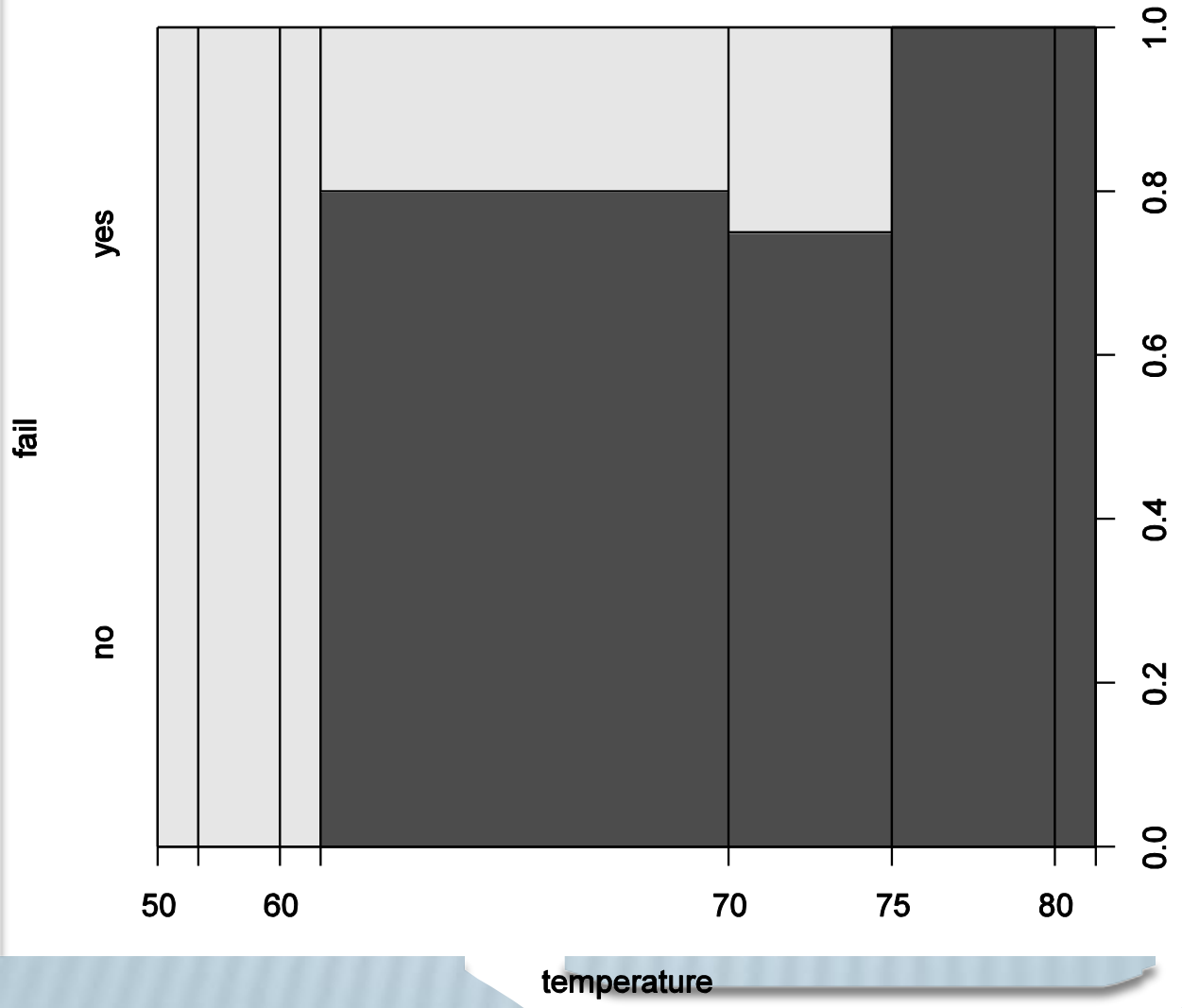




graphics包

coplot( )

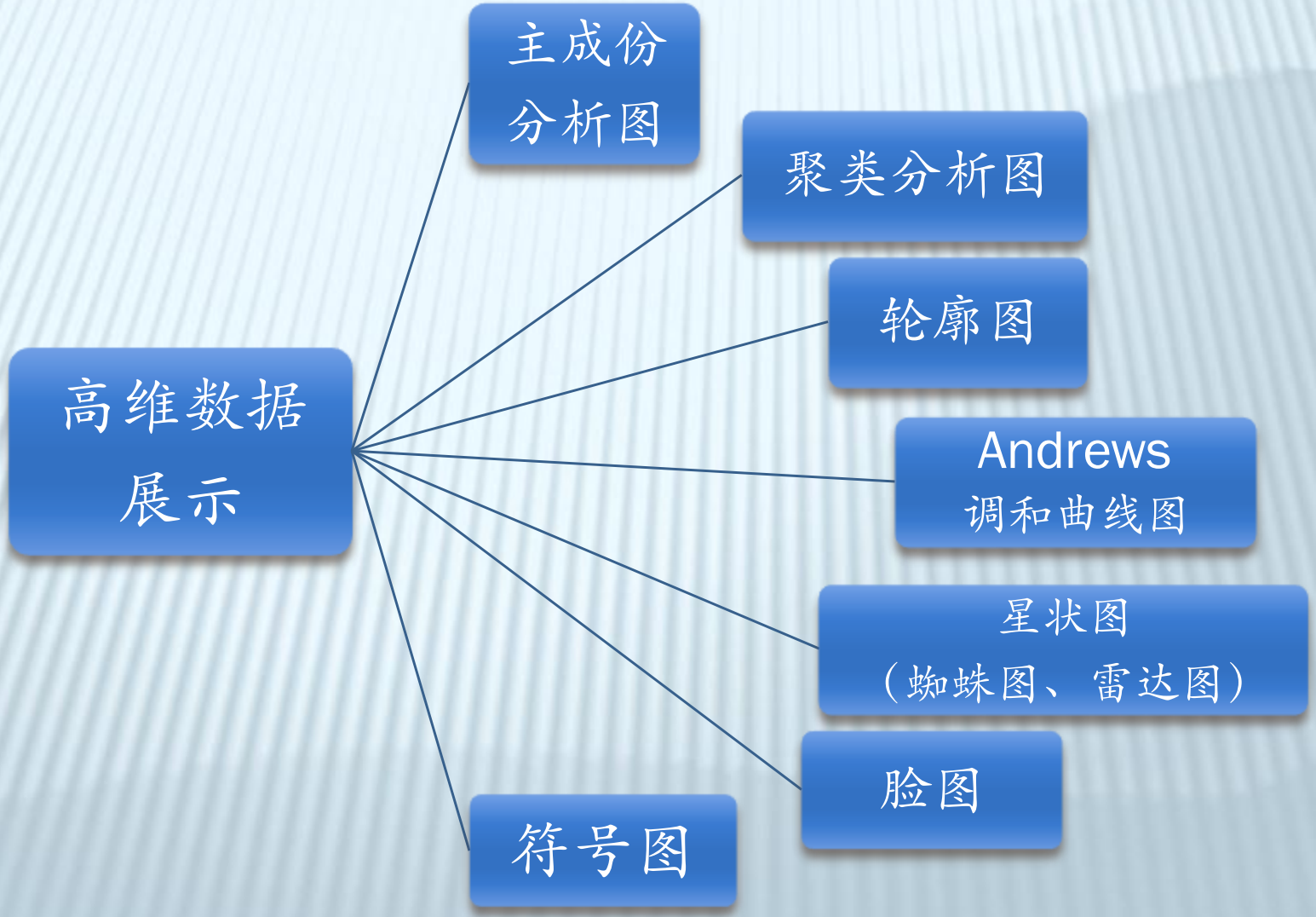
4.6.3 棘状图



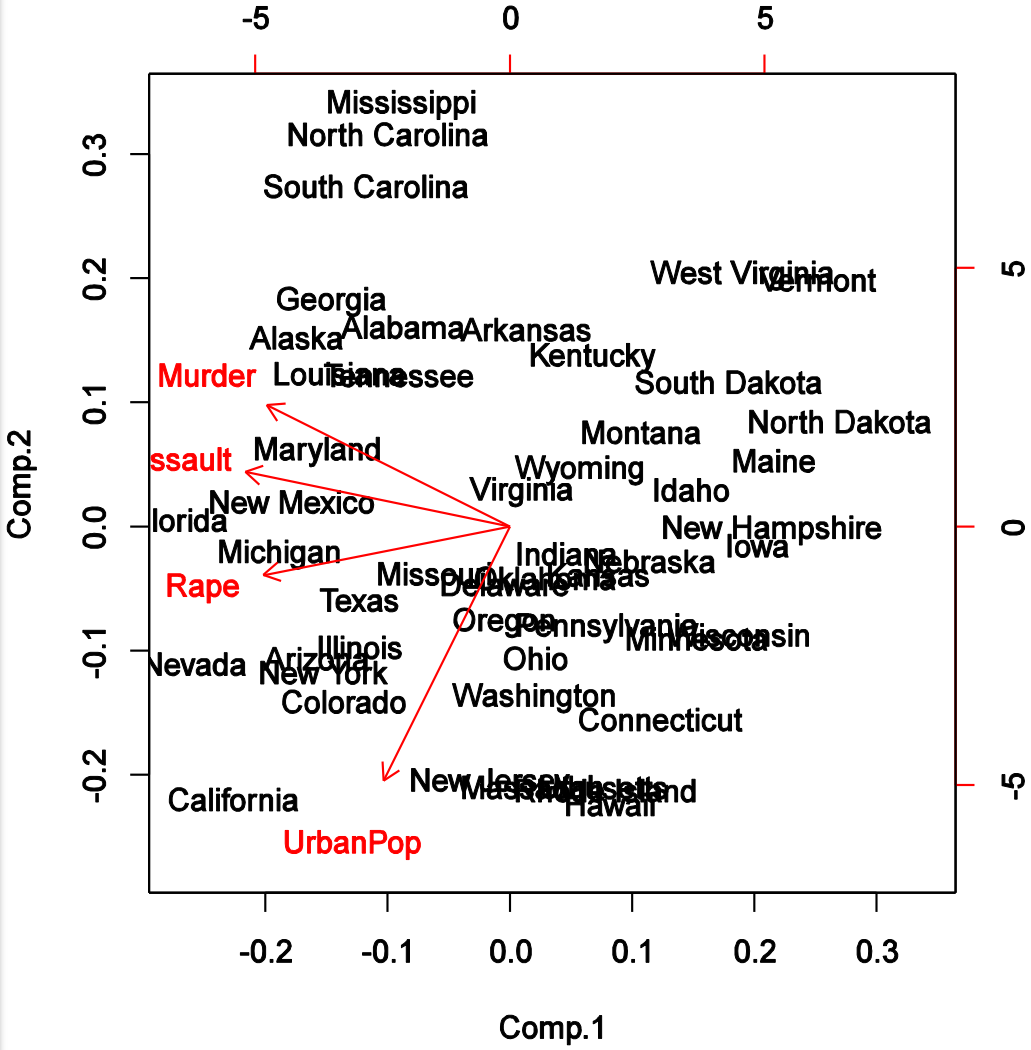
vcd包

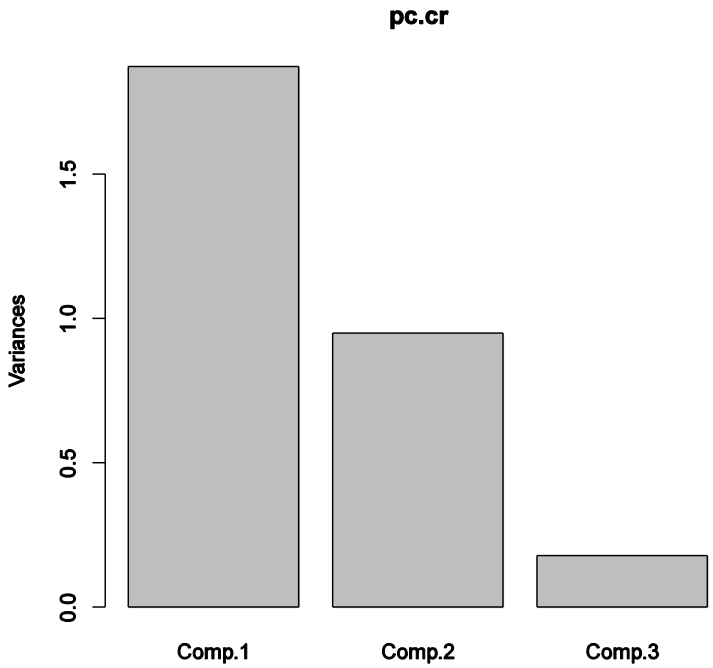
spine()

4.7 高维数据展示



4.7.1 主成份分析图(数据降维)

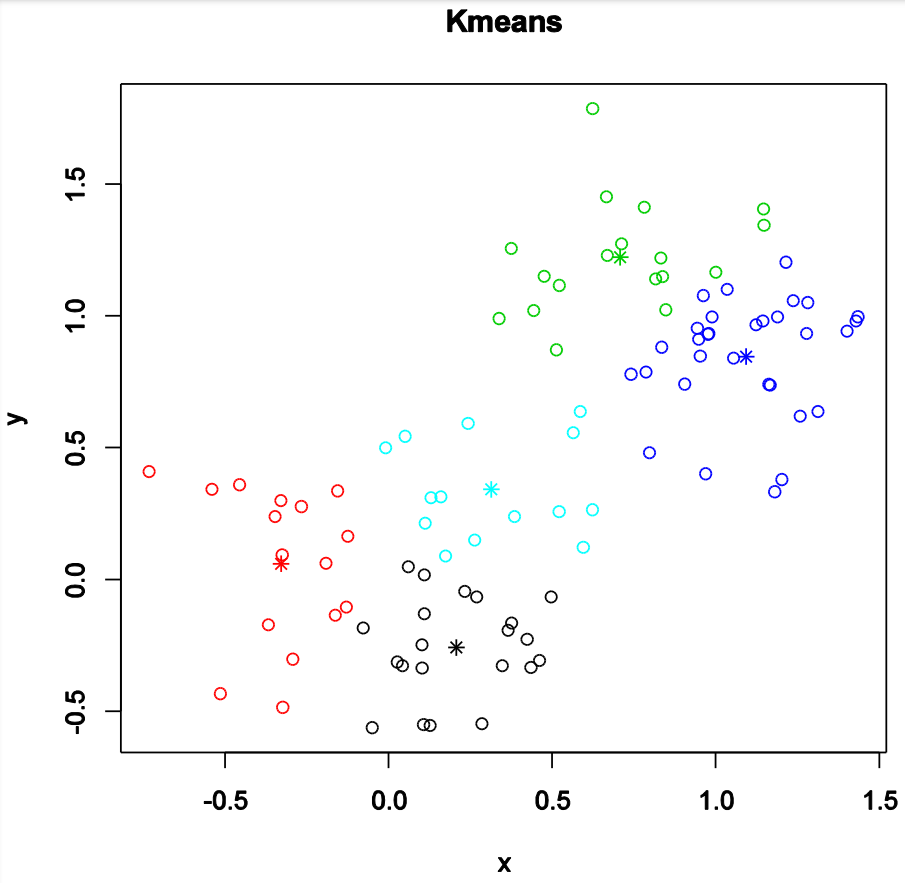




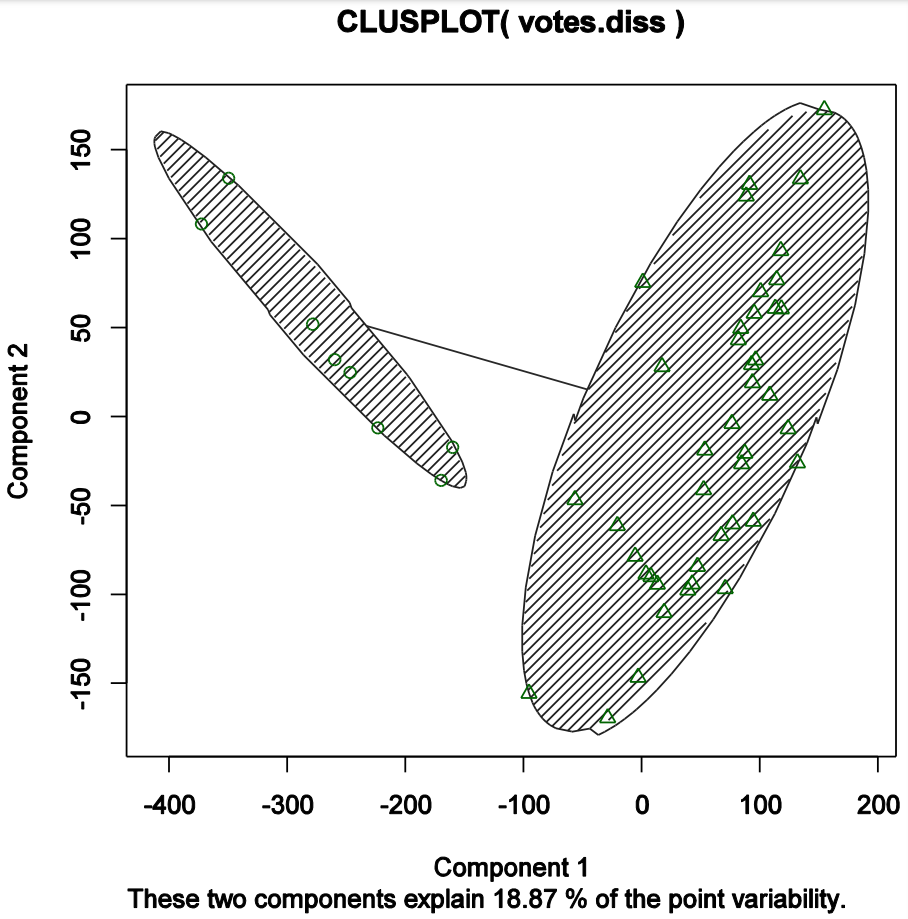
stats包

princomp( )

4.7.2 聚类分析图

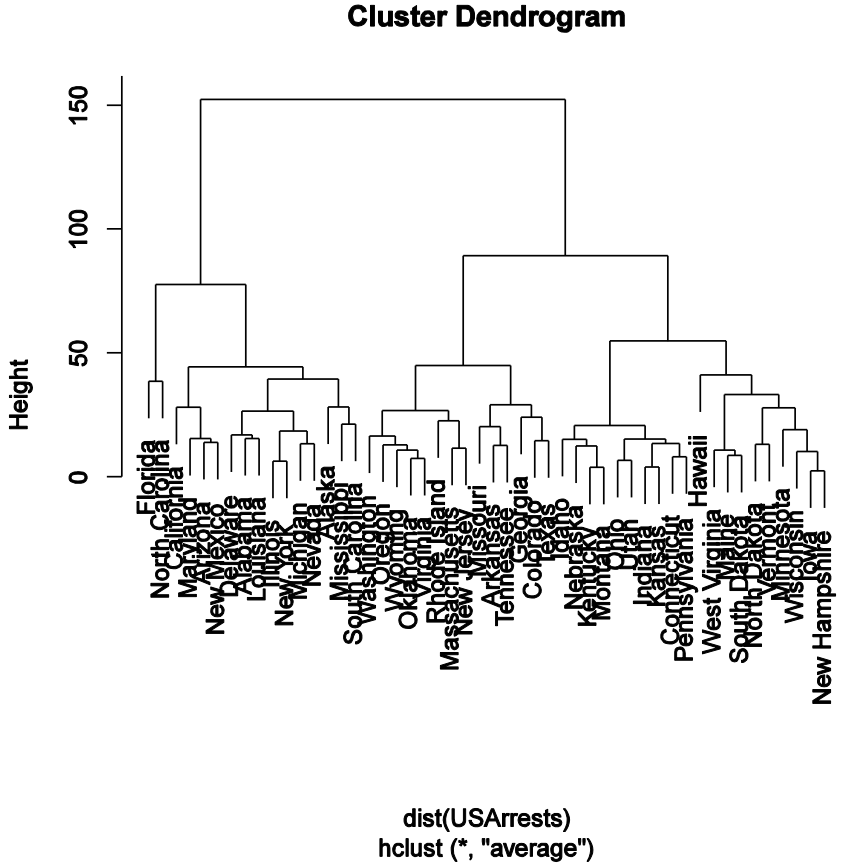


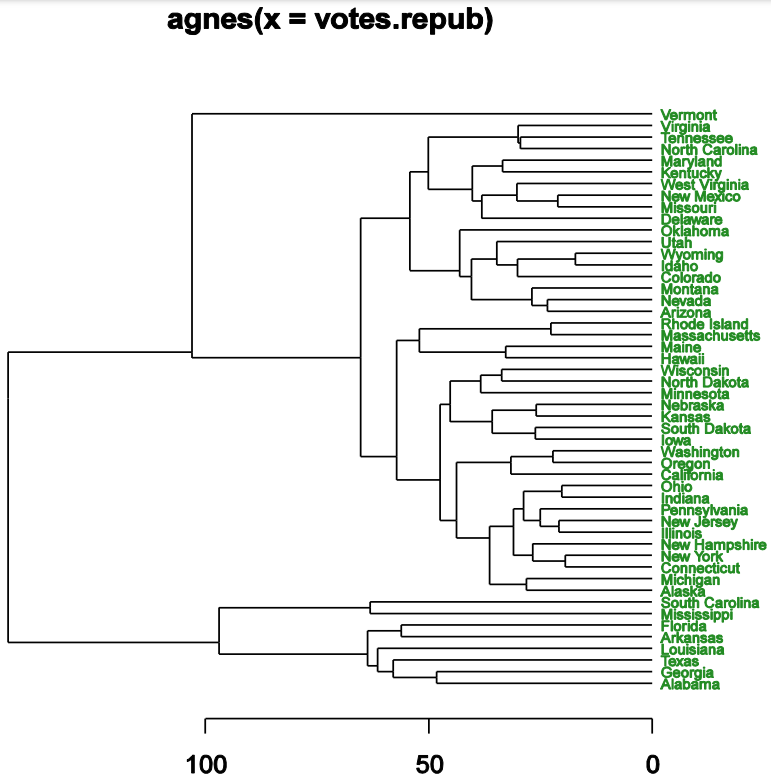
stats包kmeans()



cluster包clusplot()

4.7.2 聚类分析图

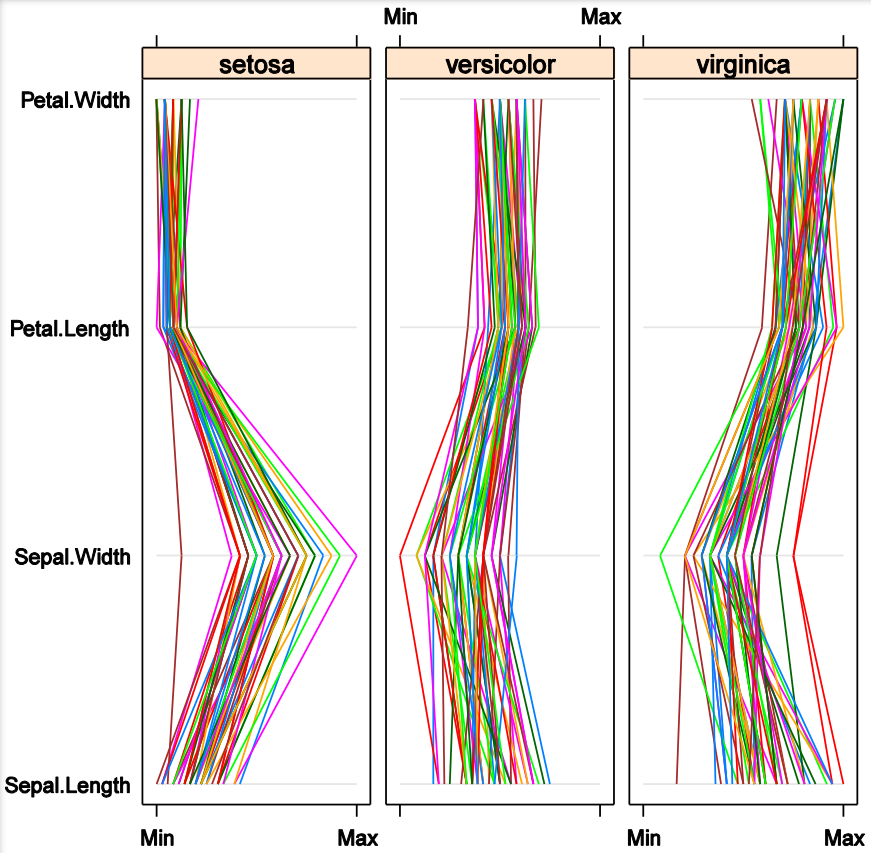


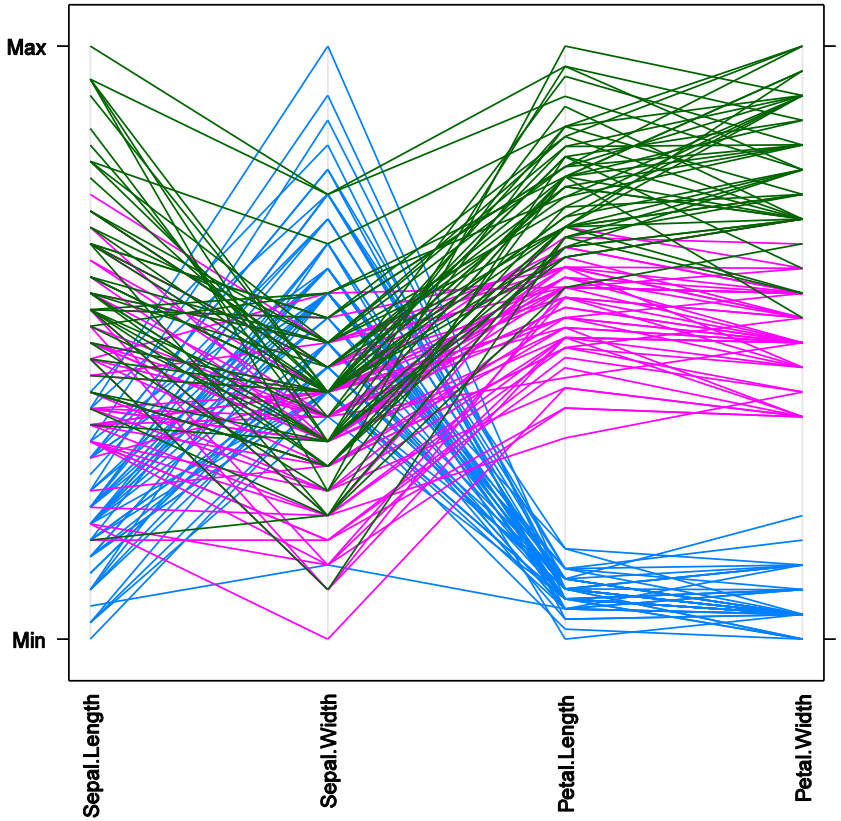


stats包 hclust()

cluster包pltree.twins()

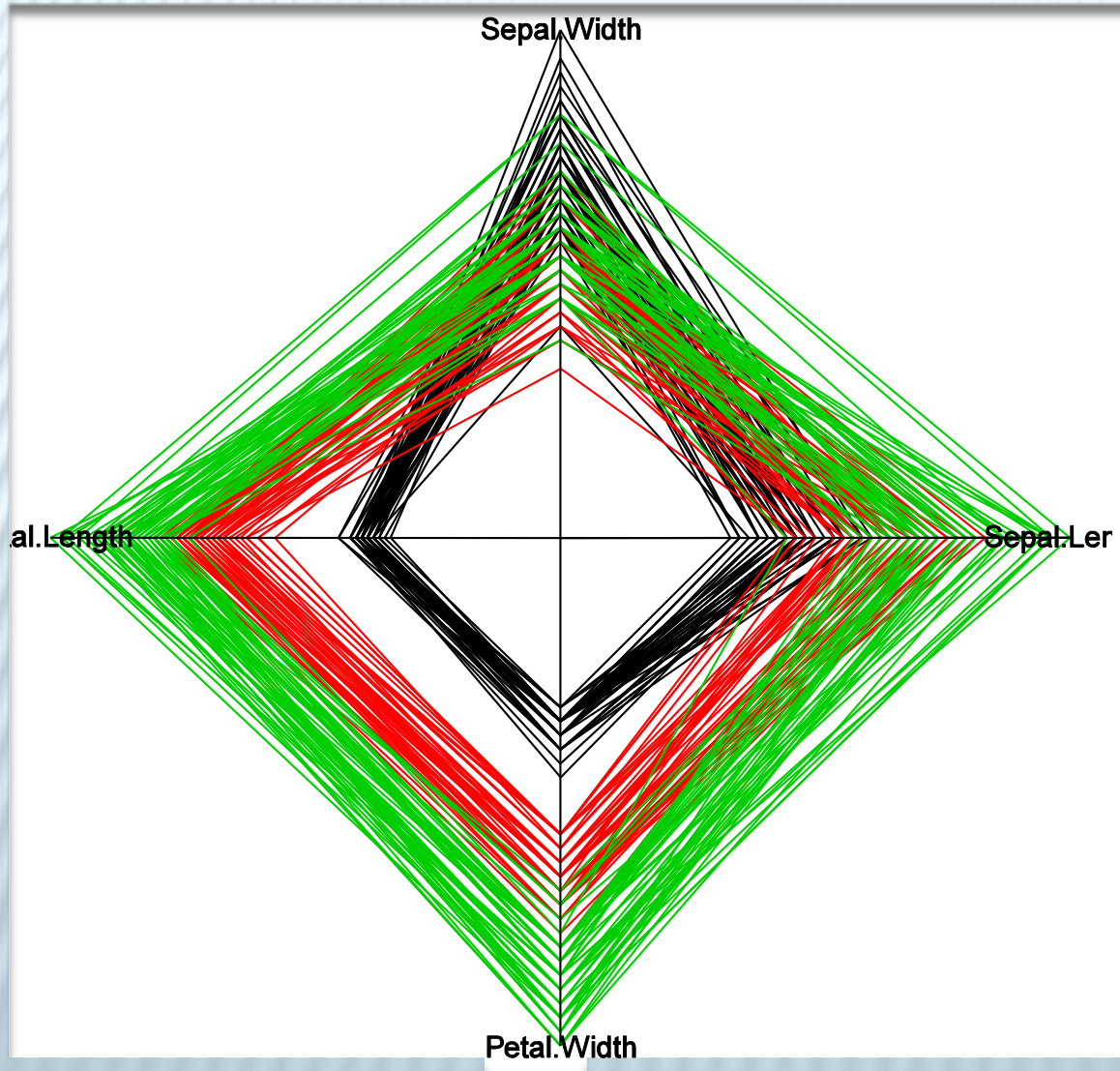
4.7.3 轮廓图(平行坐标图)





lattice包parallel()

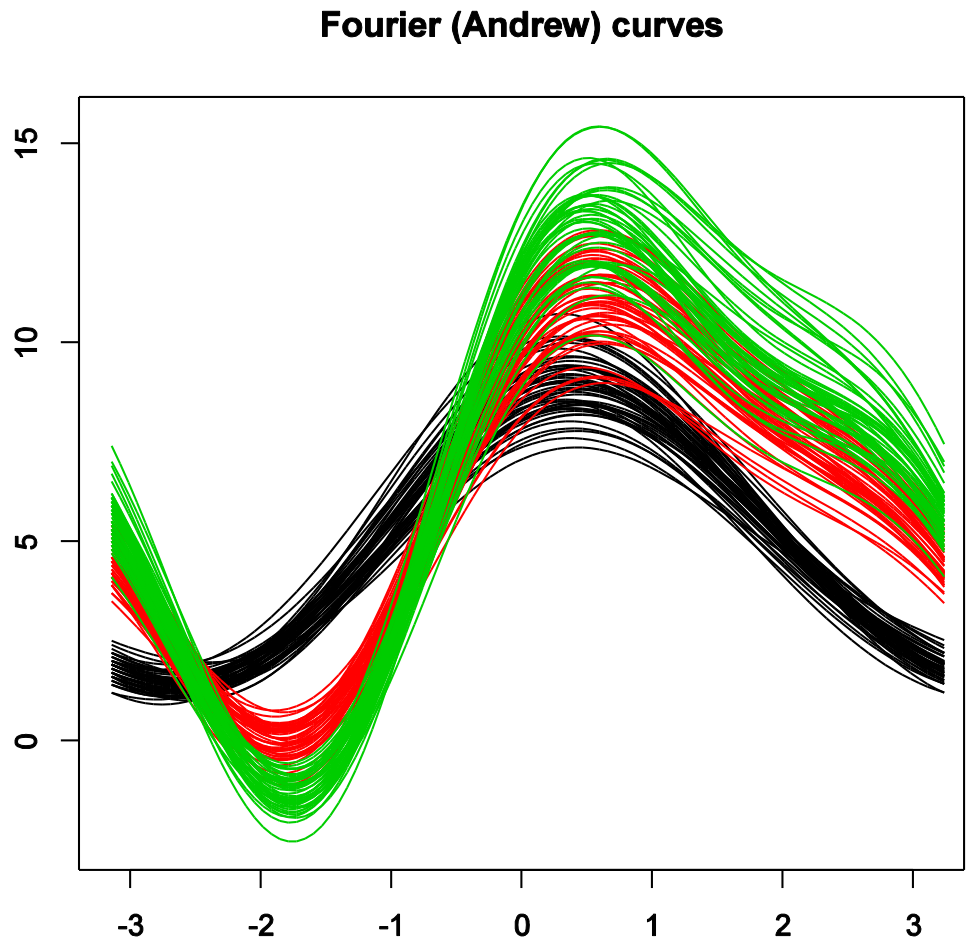
4.7.3 轮廓图



自己写函数画轮廓图，绘图函数多matplot()。此处为极坐标轮廓图。

```
polar_parallel_plot <-
function (d, col =
par("fg"),
type = "l", lty = 1, ...) {
  d <- as.matrix(d)……
……
}
```

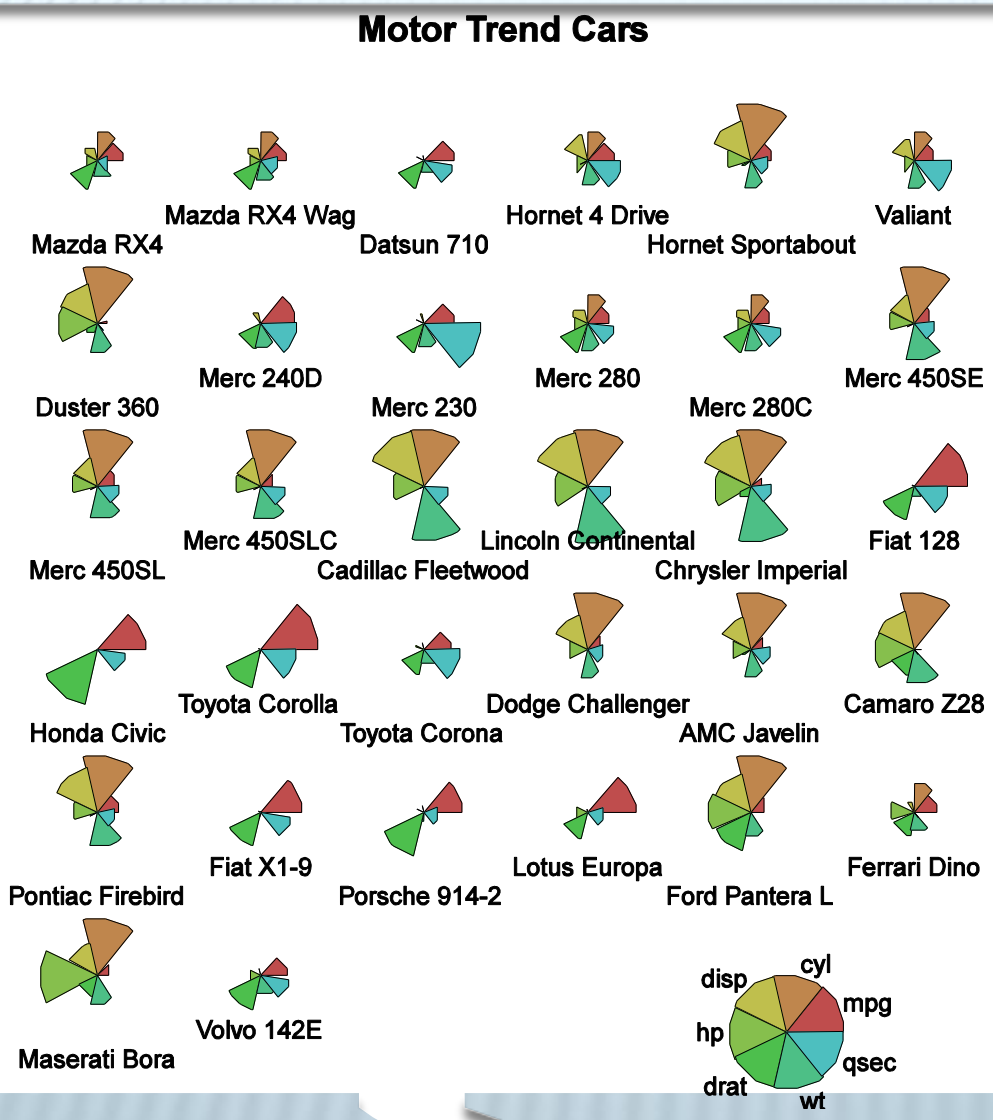
4.7.4 ANDREWS调和曲线图



需要自己写函数

```
x <- seq(-pi, pi, length=100)
y <- apply(as.matrix(iris[,1:4]),
           1,
           function (u) u[1] + u[2] * cos(x) +
                        u[3] * sin(x) + u[4] *
cos(2*x))
matplot(x, y,
        type = "l",
        lty = 1,
        col = as.numeric(iris[,5]),
        xlab = "", ylab = "",
        main = "Fourier (Andrew) curves")
```

4.7.5 星状图(蜘蛛图、雷达图)

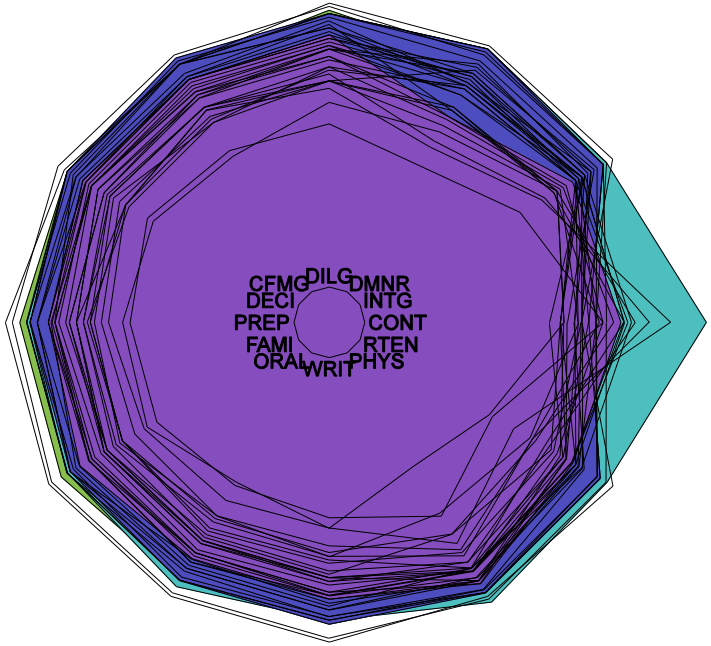


graphics包

stars(x,…)

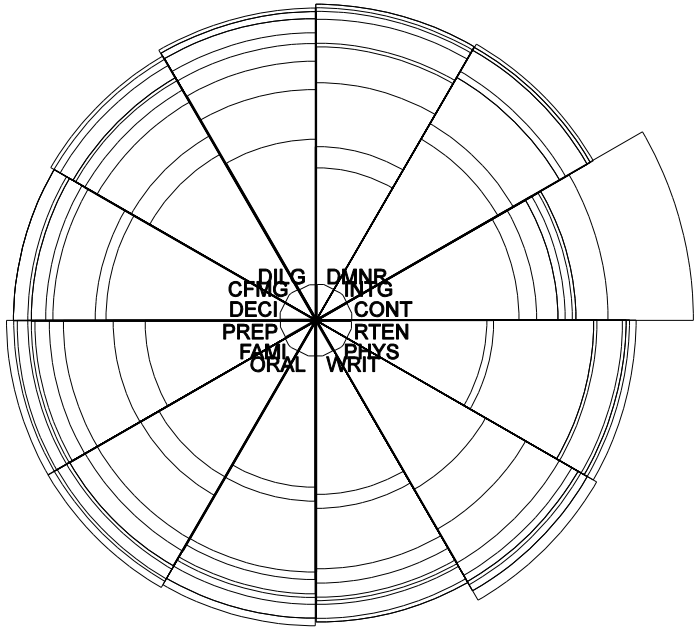
4.7.5 星状图(蜘蛛图、雷达图)

US Judges rated



蜘蛛图

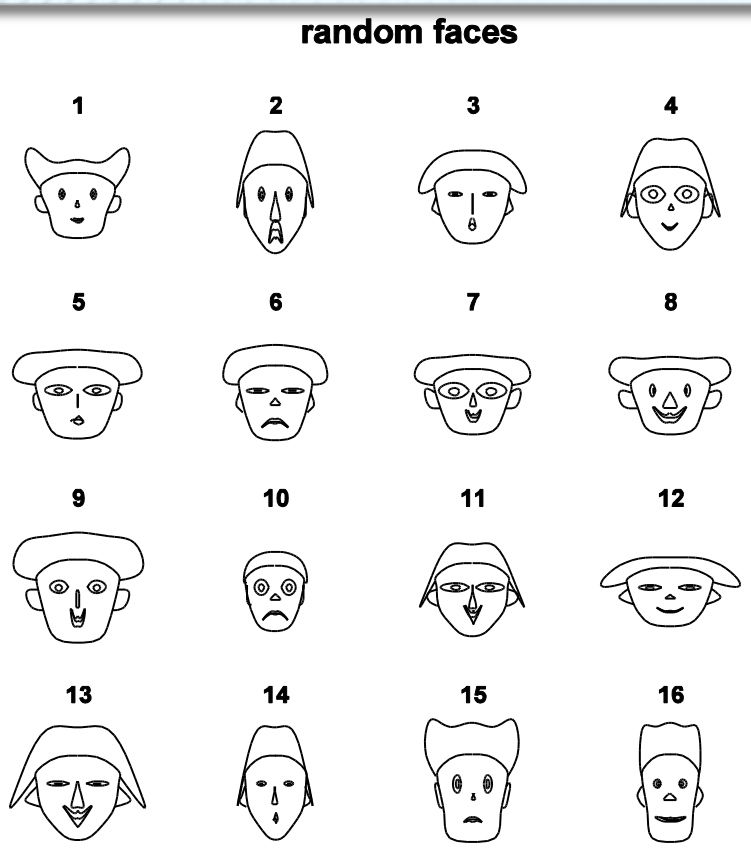
US Judges 1-10

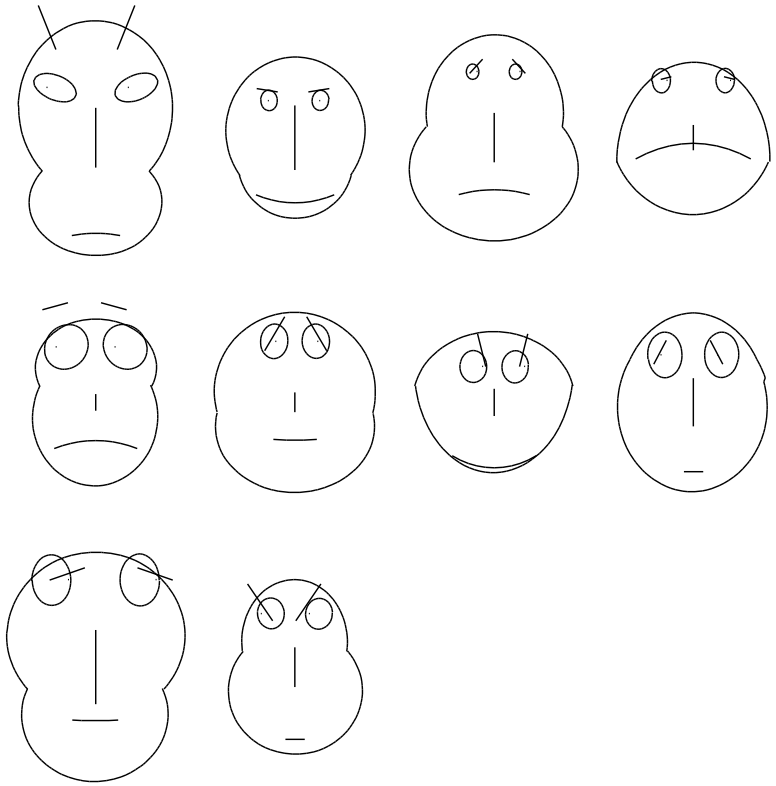


雷达图

改变stars()中参数locations

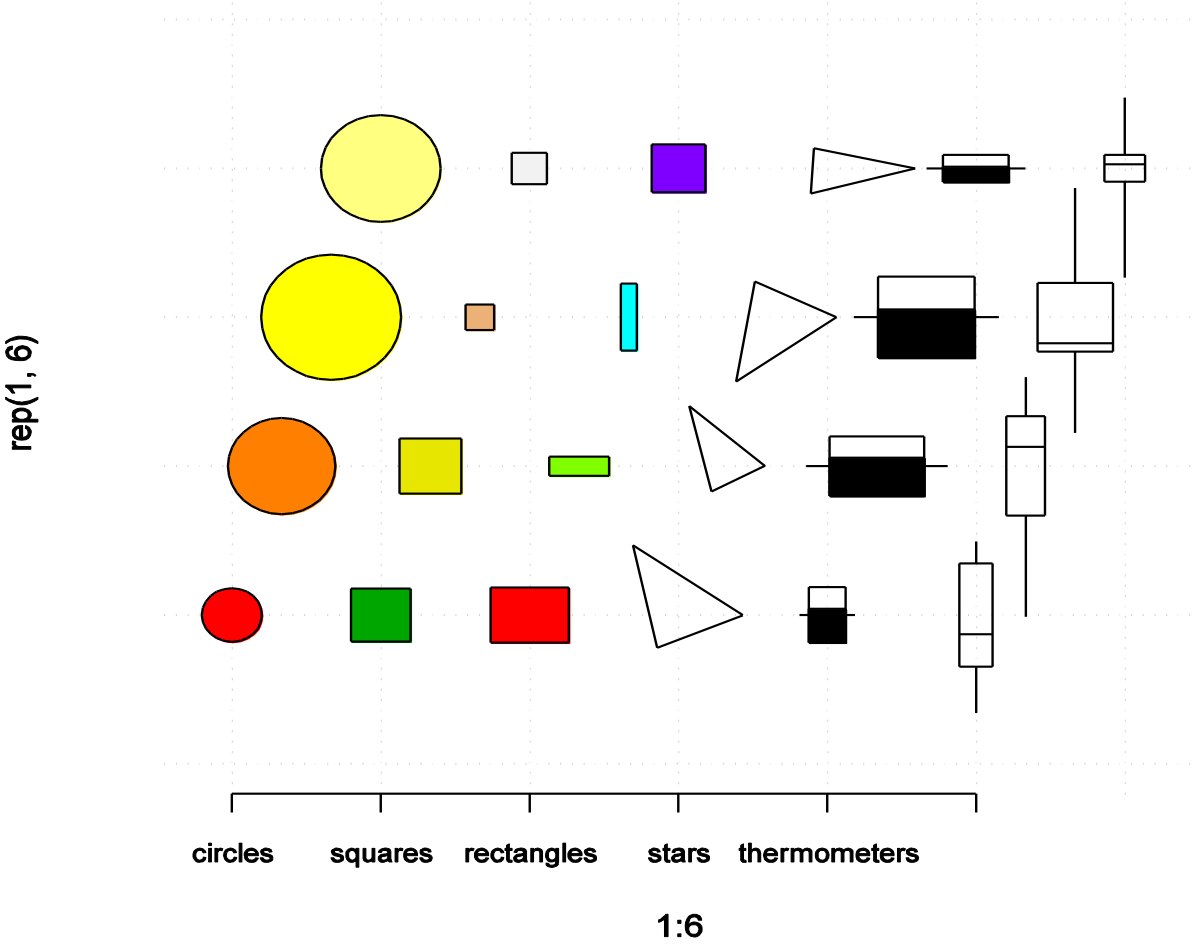
4.7.6 脸图





TeachingDemos包faces()或者faces2()

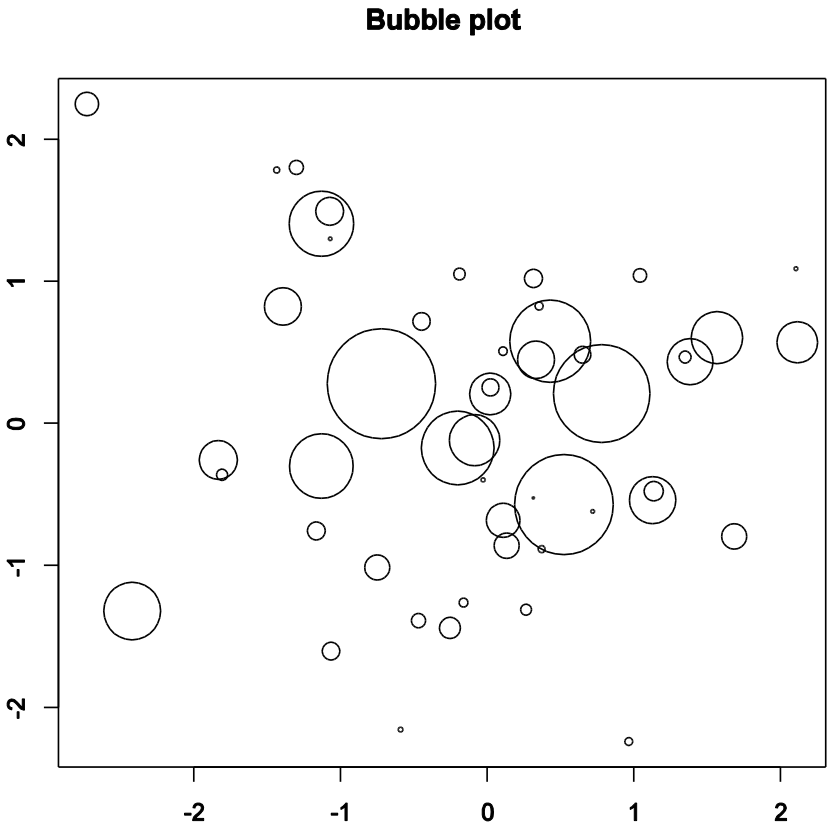
4.7.7 符号图

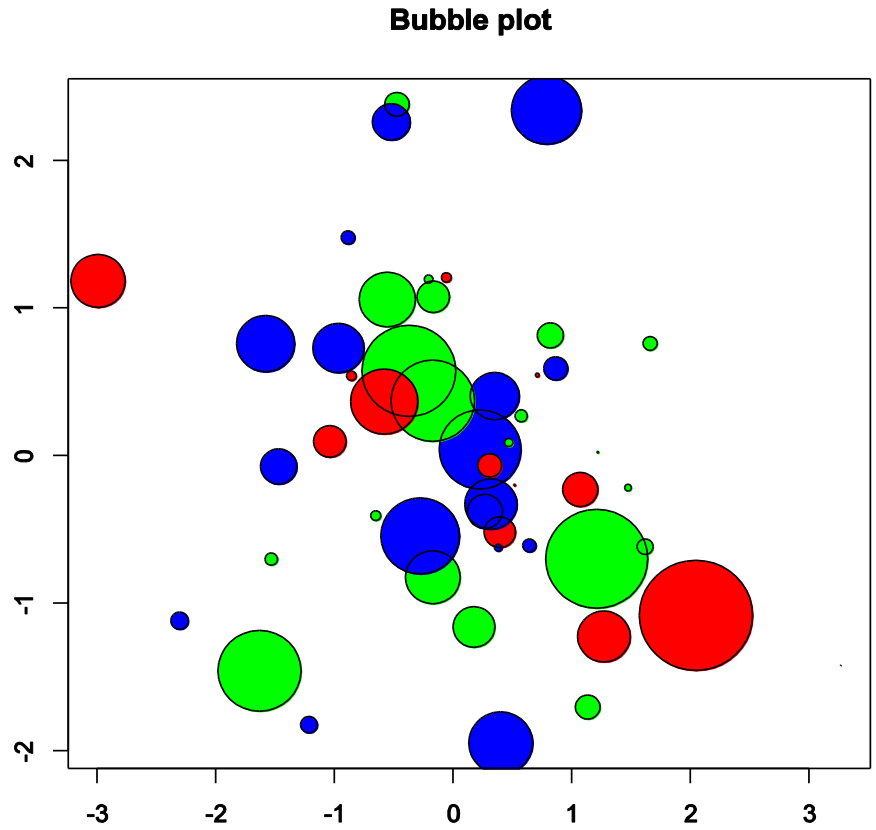


graphics包
symbols()

修改其中形状参数即可得不同符号。

4.4.1 离散型

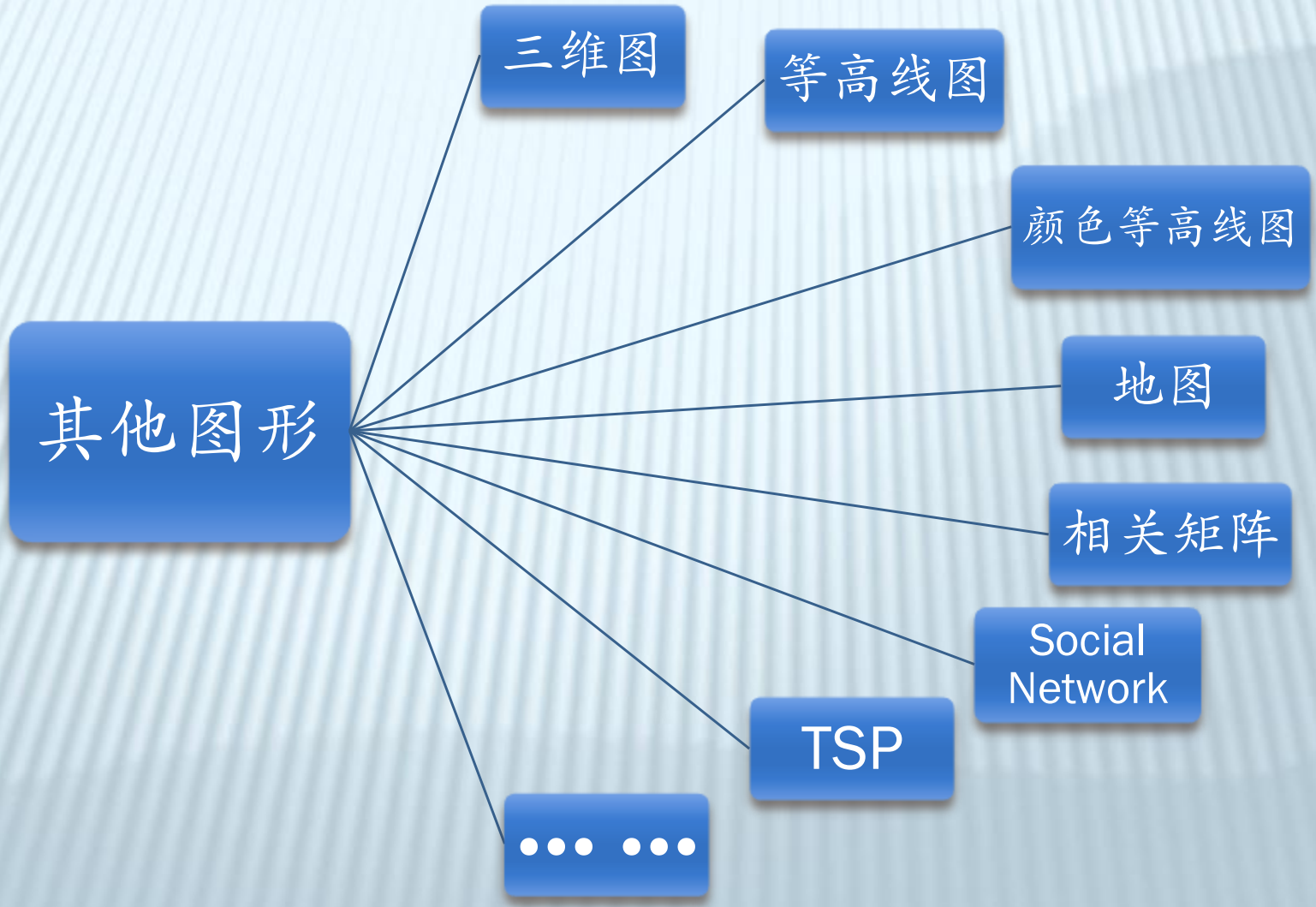




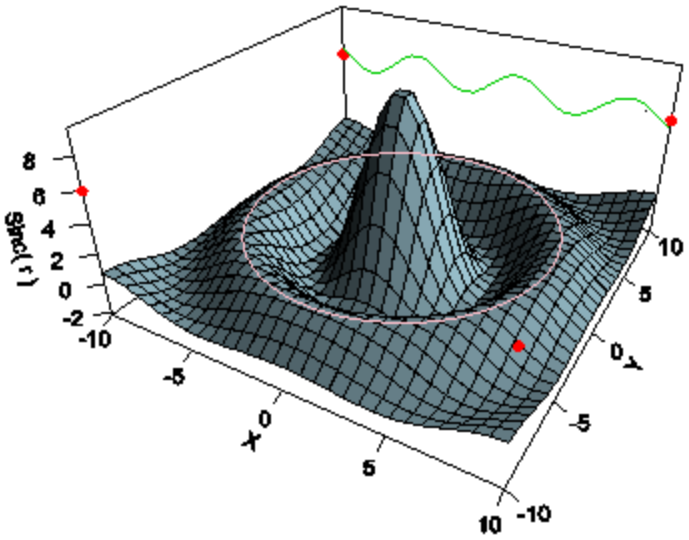
plot()　对三个或者三个以上变量作图

- 泡泡图

4.8 其他图形

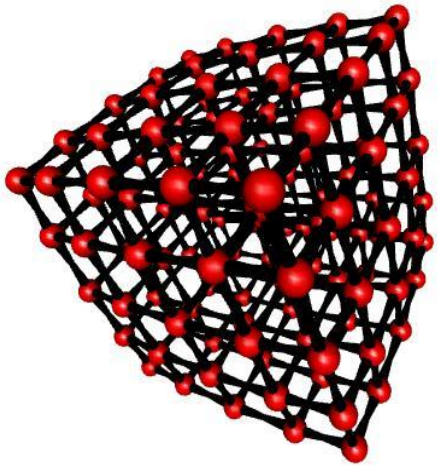


4.8.1 三维图



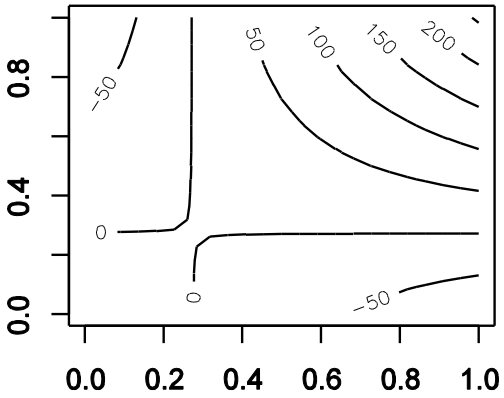
graphics包persp()

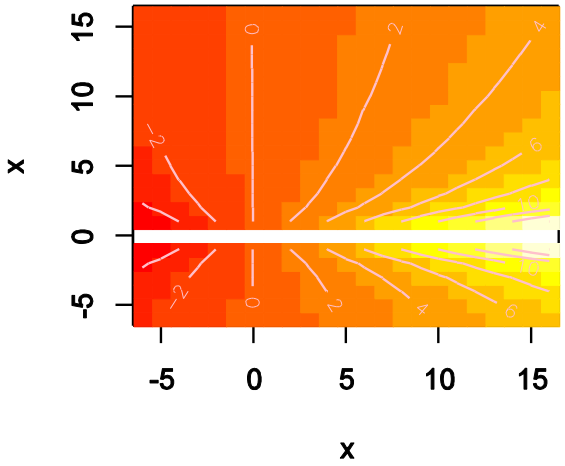
rgl包persp3d()

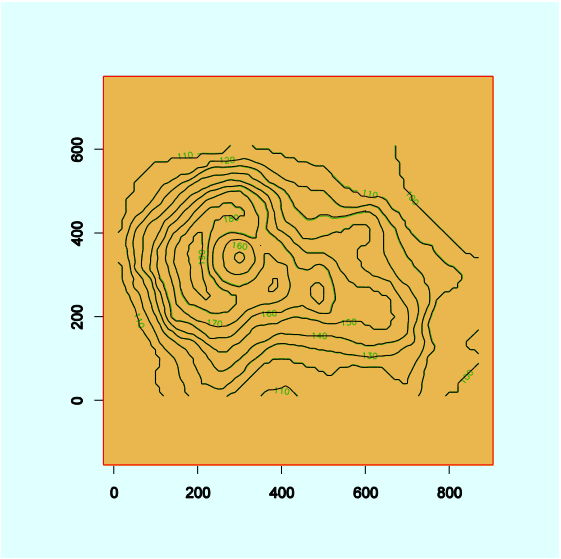


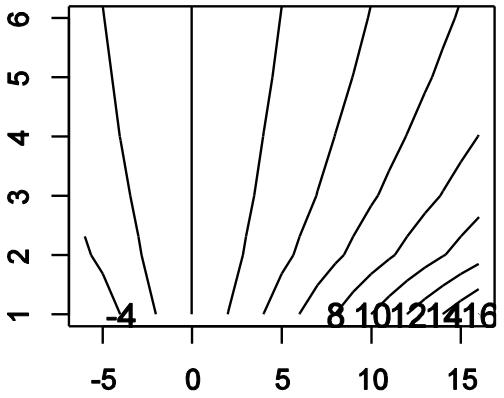
sna包gplot3d

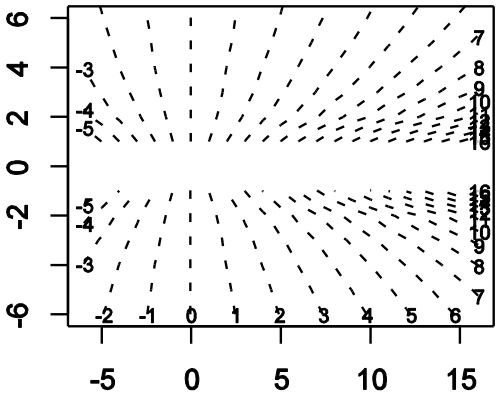
4.8.2 等高线图





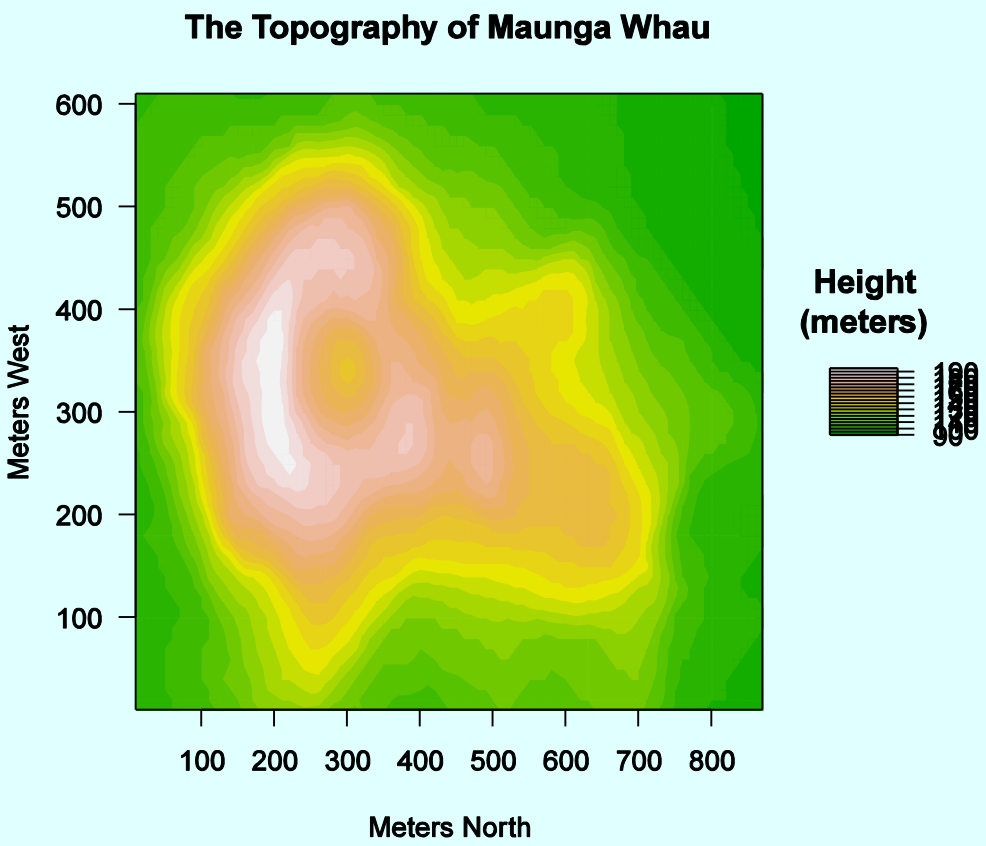


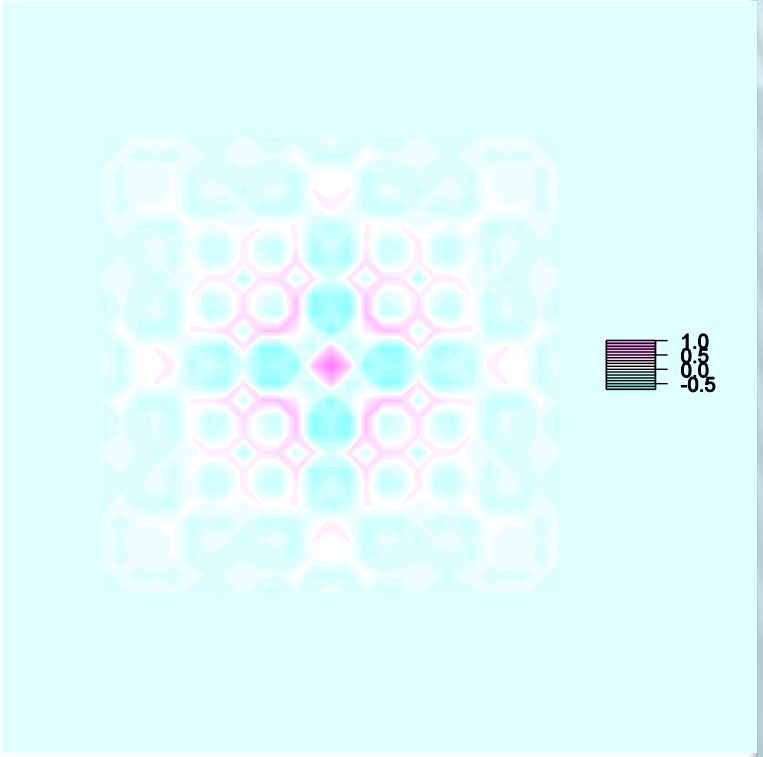




graphics包
contour()

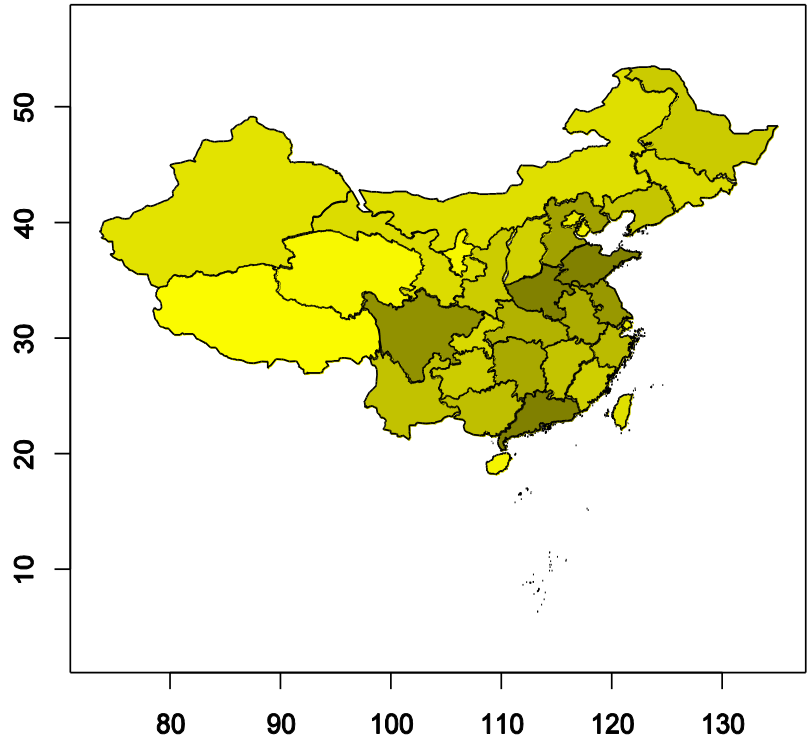
4.8.2 颜色等高图(层次图)

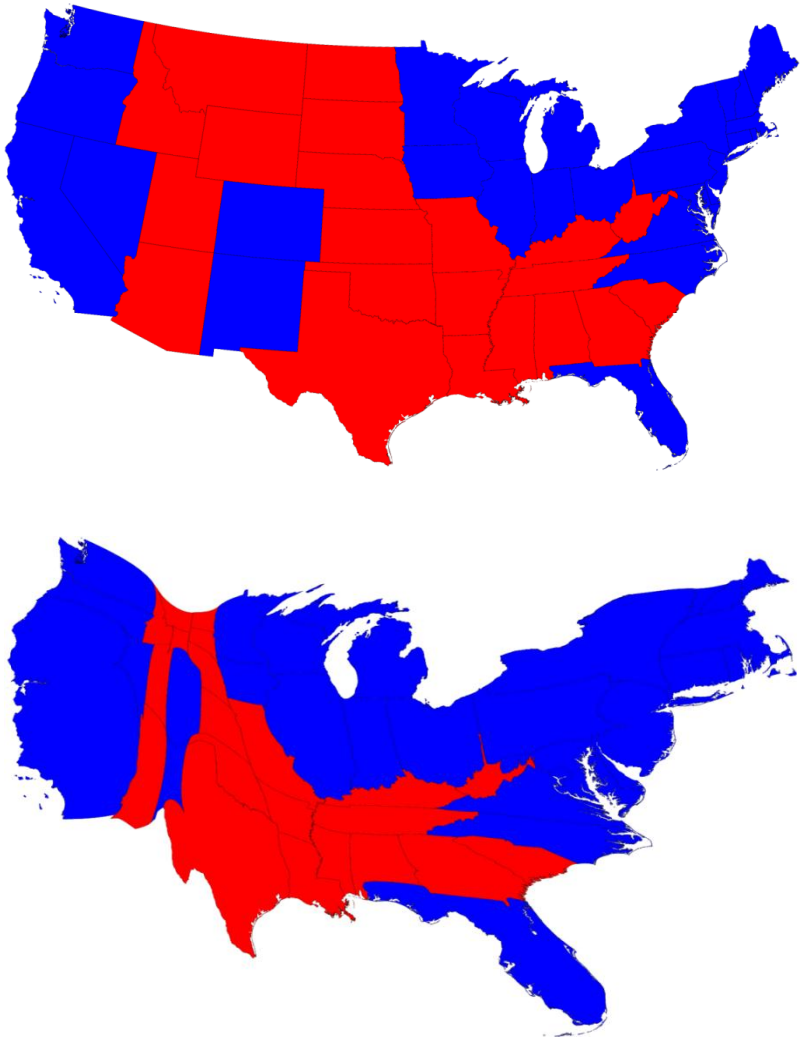




graphics包
filled.contour()

4.8.3 地图

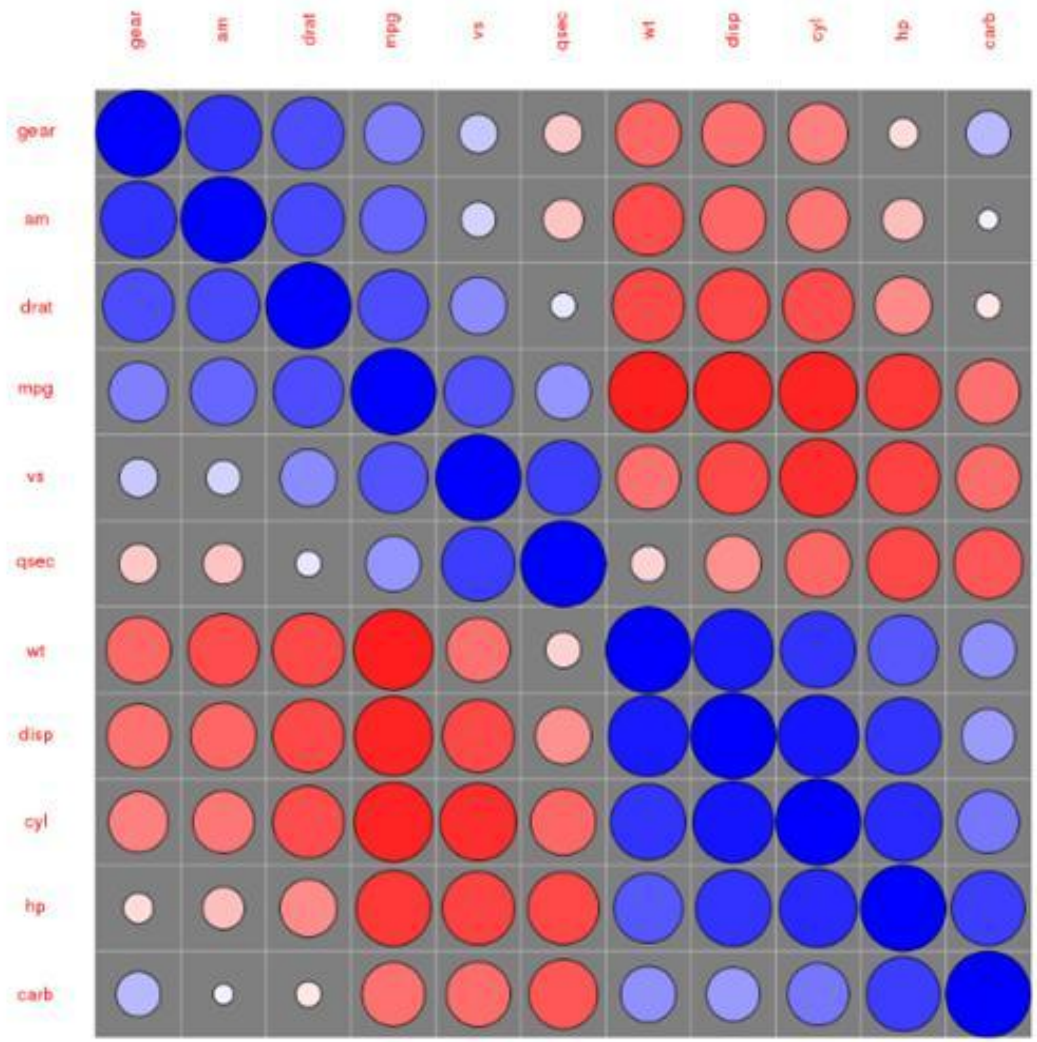




maps包、mapdata包map()

maptools包自己设置函数画地图

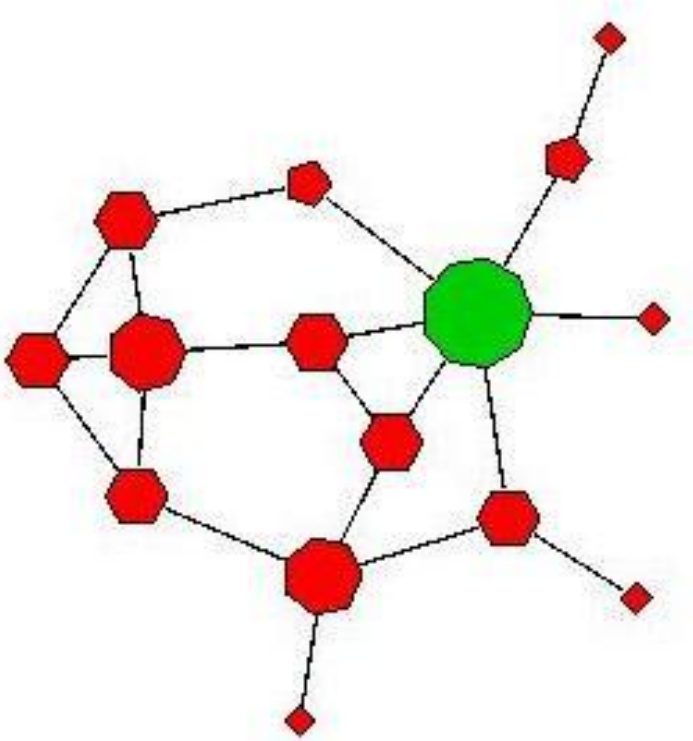
4.7.7 相关矩阵图



ellipse包
plotcorr()

魏太云的
corrplot包

4.7.7 SOCIAL NETWORK图



egrm、network、sna包等等

4.7.7 TSP图

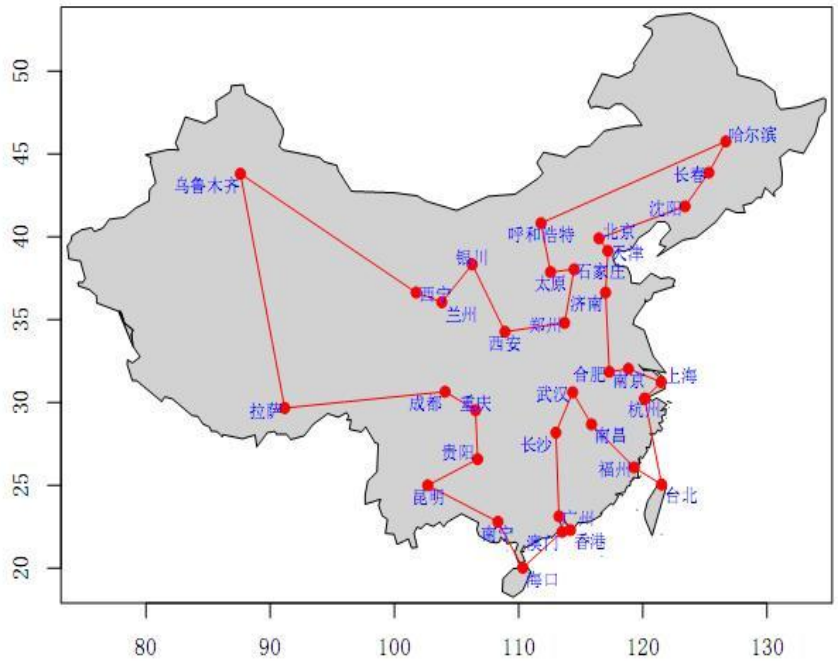
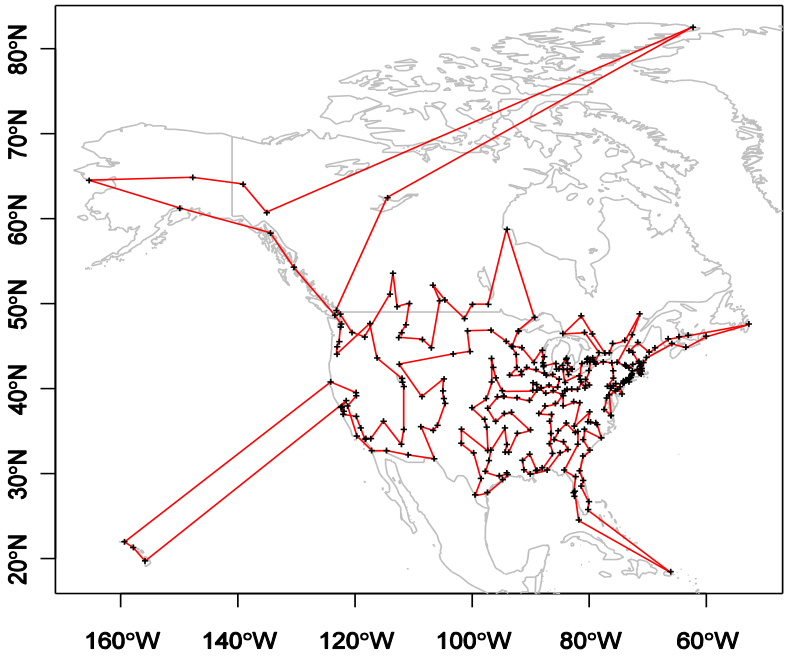
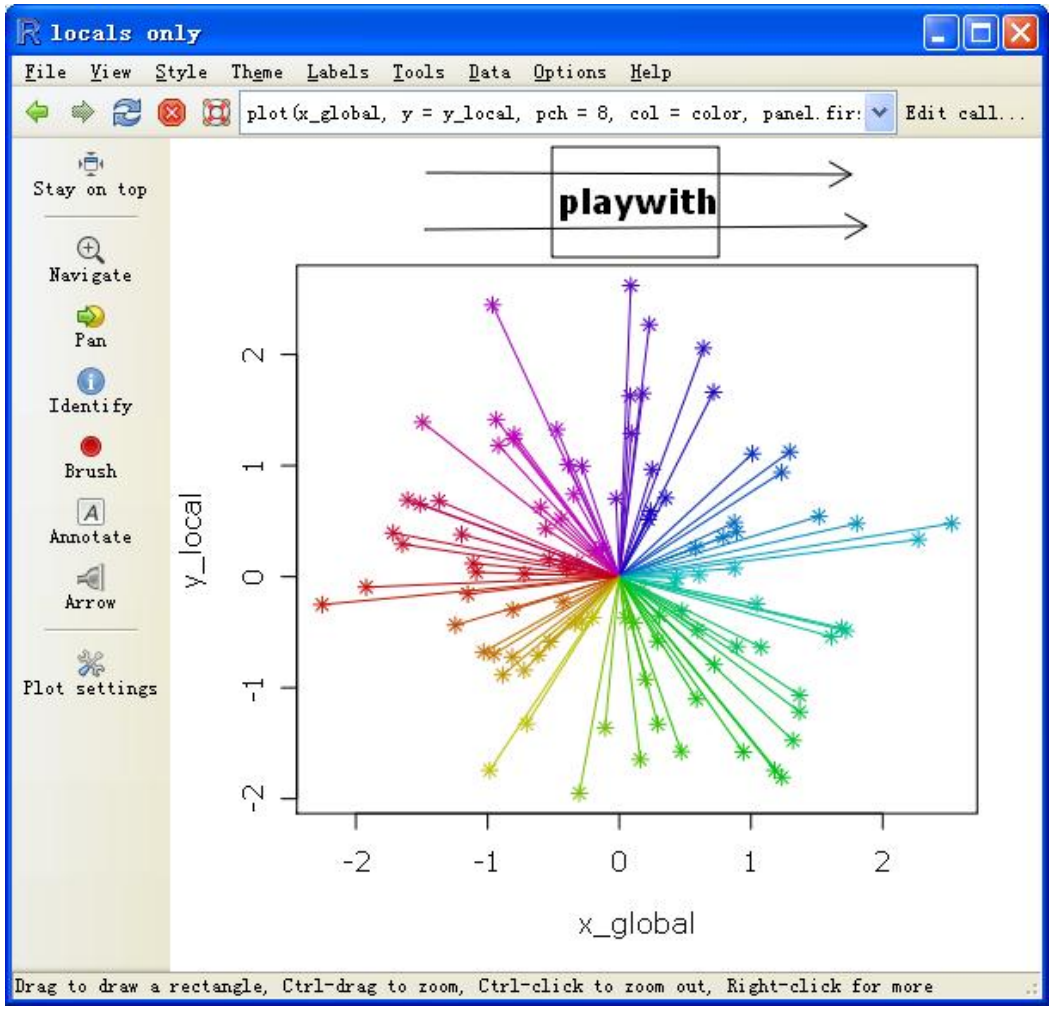


图 5: 走遍中国线路



TSP包 TSP()、solve_TSP()函数等等

4.7.8 playwith包



playwith包

省去了写代码来添加简单图示的麻烦，如箭头、矩形和文本等等

OUTLINE

- 可视化概述
- 统计图形概览与其在 R 下实现
- 统计图形的欣赏与批判
- 总结与展望

5 统计图形的欣赏与批判

What is Graphical excellence?

An excellent graph…

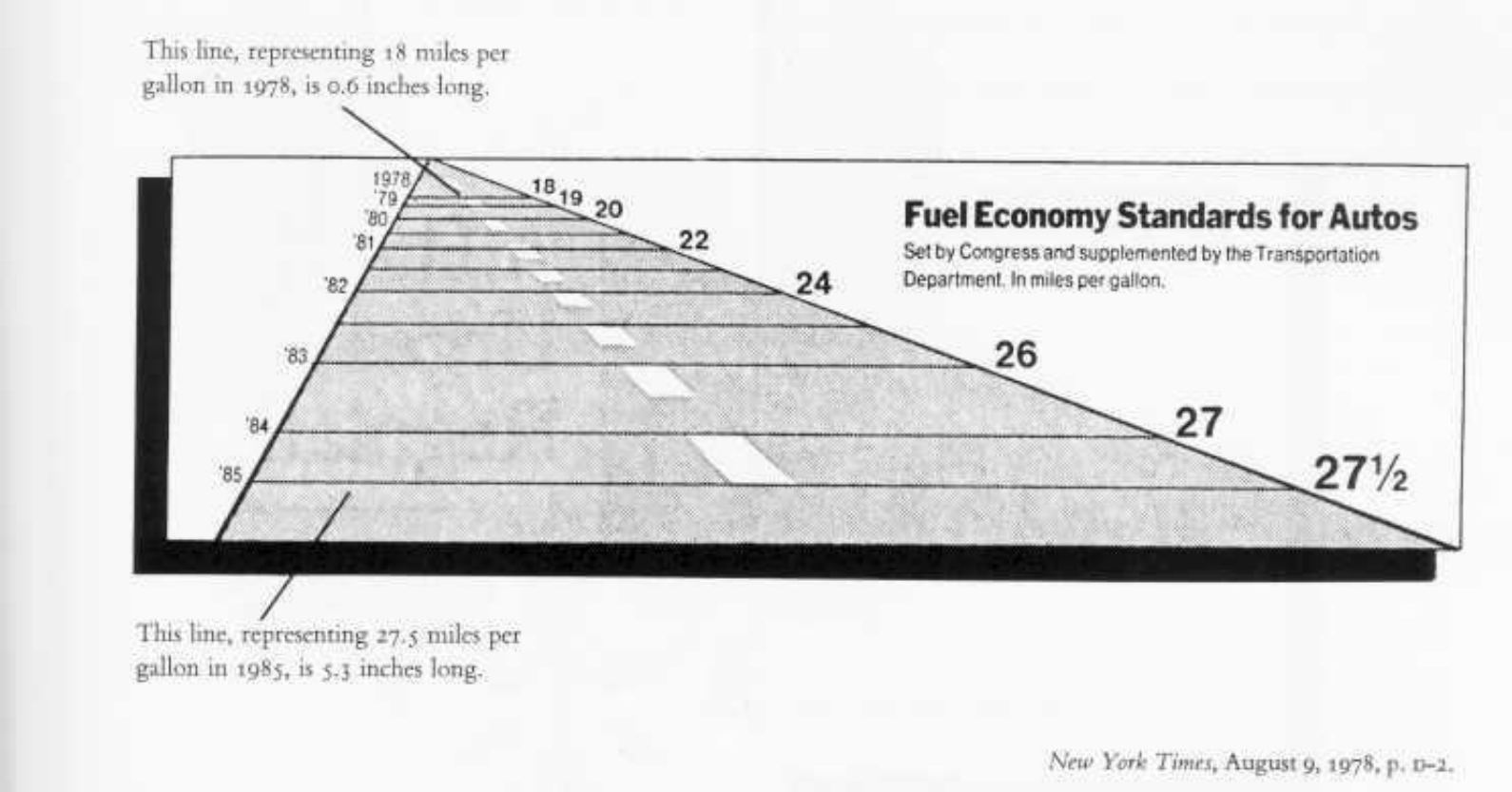
- shows the data
- makes the viewer think about the subject not the graph
- doesn't distort the data
- helps the eye make comparisons
- is visually efficient, showing a lot of info with a little ink

5 统计图形的欣赏与批判

And what is bad graph?

5 统计图形的欣赏与批判（批判篇）

The Lie Factor

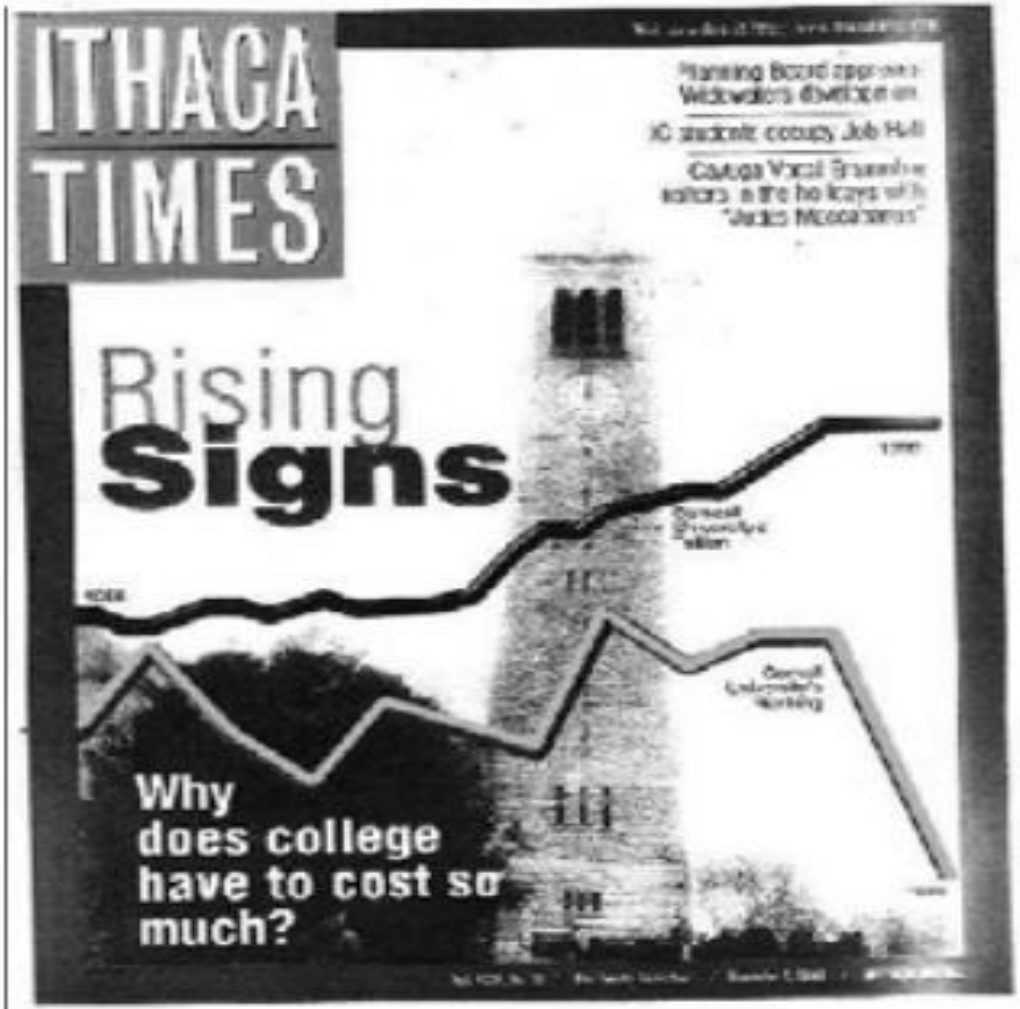


New York Times, August 9, 1978, p. D-2.

5 统计图形的欣赏与批判（批判篇）

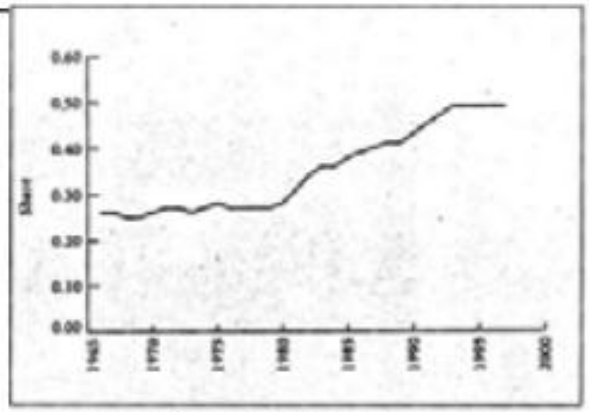
- This graph shows how one American dollar in 1958 had shrunk to a value of 44 cents in 1978 (due to the effects of rising prices or inflation).
- If you think carefully, this means that one American dollar in 1978 could buy just under half as much as it could in 1958.
- Here, the artist decreased the length by half, so that decreases the area by a factor of 4.
- You may argue that this problem goes unnoticed by people when they look at a pictograph like this one, so it is not particularly important. However, the fact is that subconsciously many people interpret the dollar to have lost far more of its value than is the case.
- *It is also worth noting that the pictograph appeared during an American presidential election campaign in a leading newspaper…*

5 统计图形的欣赏与批判（批判篇）

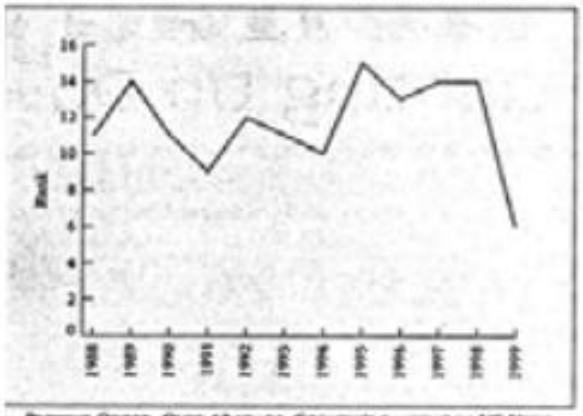


- from the cover of the *Ithaca Times* (Dec. 7, 2000)
- The cover story, "Why does college have to cost so much?" shows a large graph superimposed on a scene from the Cornell campus.
- There are two jagged lines running across the graph, one labeled "Cornell's Tuition" and the other "Cornell's Ranking".
- The tuition graph shows a steady rise, and the ranking graph, after some early meandering, plummets to an all time low.
- The clear impression is that students are paying more for far less.
- What's wrong with these pictures?

5 统计图形的欣赏与批判（批判篇）



By the Numbers: Over 35 years, Cornell's tuition has taken an increasingly larger share of its median student family income.



Picking Order: Over 12 years, Cornell's ranking in *US News & World Report* has risen and fallen erratically.

- More careful reading of the whole article (buried several pages into the paper) reveals a different story:
- The ranking graph covers an 11 year period, the tuition graph 35 years, yet they are shown simultaneously (the same apparent width) on the same horizontal "scale".
- The vertical scale for tuition and ranking could not possibly have common units, but the ranking graph is placed under the tuition graph creating the impression that cost exceeds quality.
- And here is the masterstroke: the sharp "drop" in the ranking graph over the past few years actually represents the fact that Cornell's rank has IMPROVED from 15th TO 6th ... (a lower ranking is actually good!)

OUTLINE

- 可视化概述
- 统计图形概览与其在 R 下实现
- 统计图形的欣赏与批判
- 总结与展望

6 总结与展望

一图胜千言

将最关键的信息用最能激发视觉感知的形式表现出来

参考文献

- 谢益辉 《现代统计图形》2008-08-04
- 薛毅、张立萍《统计建模与R软件》
- http://cos.name/
- http://www.math.yorku.ca/SCS/friendly.html
- http://zoonek2.free.fr/UNIX/48_R/all.html
- http://www.math.yorku.ca/SCS/Gallery/
- http://addictedtor.free.fr/graphiques/
- ……

非常感谢！