Data Analysis Tools and Practice(Using R)

2020.03.11

R基本图形I

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

R包介绍

- 选择一个R扩展包，做10-20分钟的课堂介绍，包括包的作用，示例，2道习题；
- 组团自愿，人数不要太多或太少；
- 包的选择可以检索官方网站，也可以搜索。

计划4月份开始介绍

Available Packages

Currently, the CRAN package repository features 10338 available packages.

Table of available packages, sorted by date of publication

Table of available packages, sorted by name

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages fi

CRAN Task Views allow you to browse packages by topic and provide tools to automatically install all packages

Package Check Results

All packages are tested regularly on machines running Debian GNU/Linux, Fedora, OS X, Solaris and Windows

The results are summarized in the check summary (some timings are also available). Additional details for Wind

Writing Your Own Packages

The manual Writing R Extensions (also contained in the R base sources) explains how to write new packages an

Repository Policies

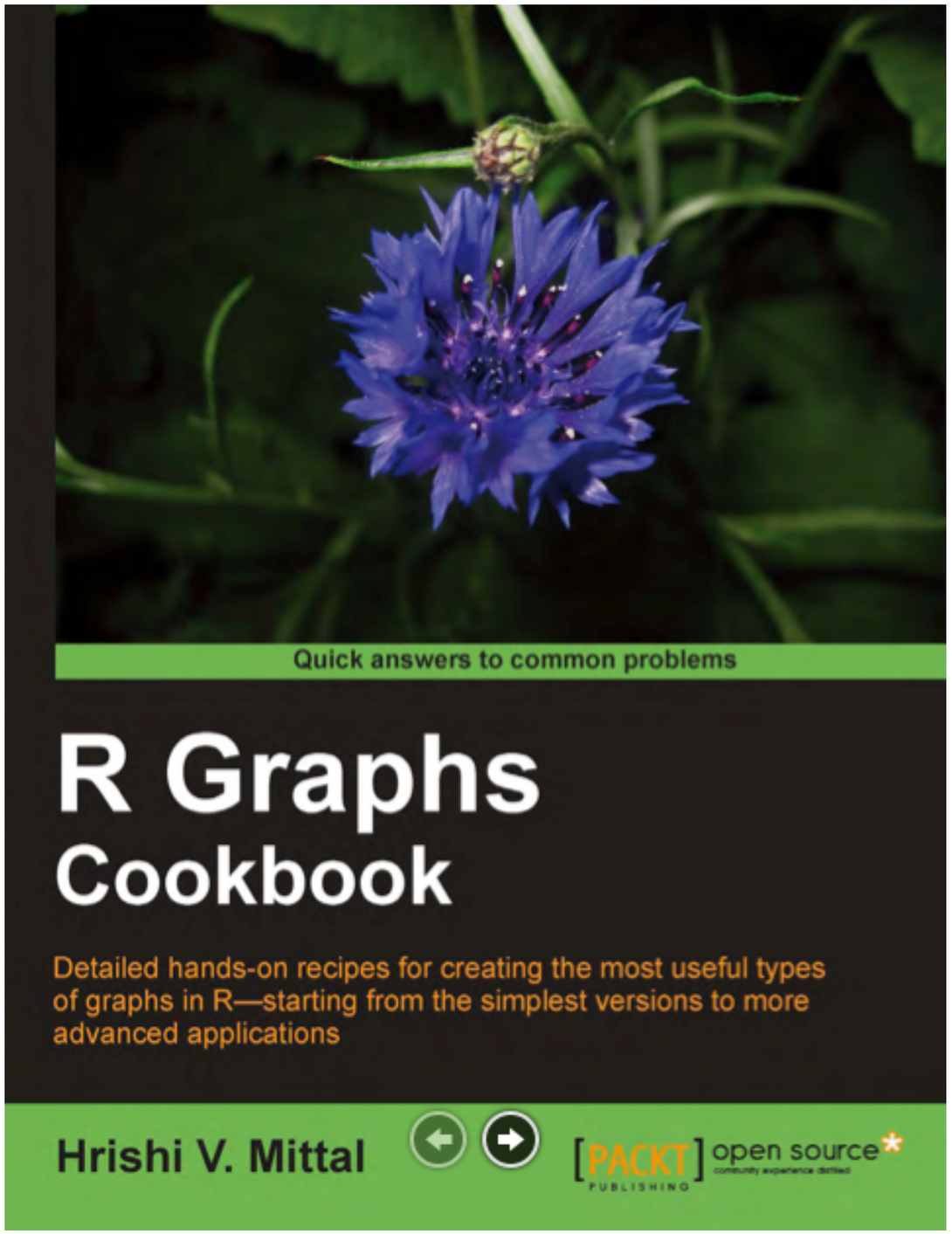
The manual CRAN Repository Policy [PDF] describes the policies in place for the CRAN package repository.

https://cran.r-project.org/web/views/

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
ExtremeValue	Extreme Value Analysis
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
MetaAnalysis	Meta-Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
NumericalMathematics	Numerical Mathematics
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods
ReproducibleResearch	Reproducible Research
Robust	Robust Statistical Methods
SocialSciences	Statistics for the Social Sciences
Spatial	Analysis of Spatial Data
SpatioTemporal	Handling and Analyzing Spatio-Temporal Data
Survival	Survival Analysis
TimeSeries	Time Series Analysis
WebTechnologies	Web Technologies and Services
gR	gRaphical Models in R

https://cran.r-project.org/web/packages/

参考教材


Quick answers to common problems
R Graphs
Cookbook
Detailed hands-on recipes for creating the most useful types of graphs in R—starting from the simplest versions to more advanced applications
Hrishi V. Mittal
[PACKT] open source
PUBLISHING

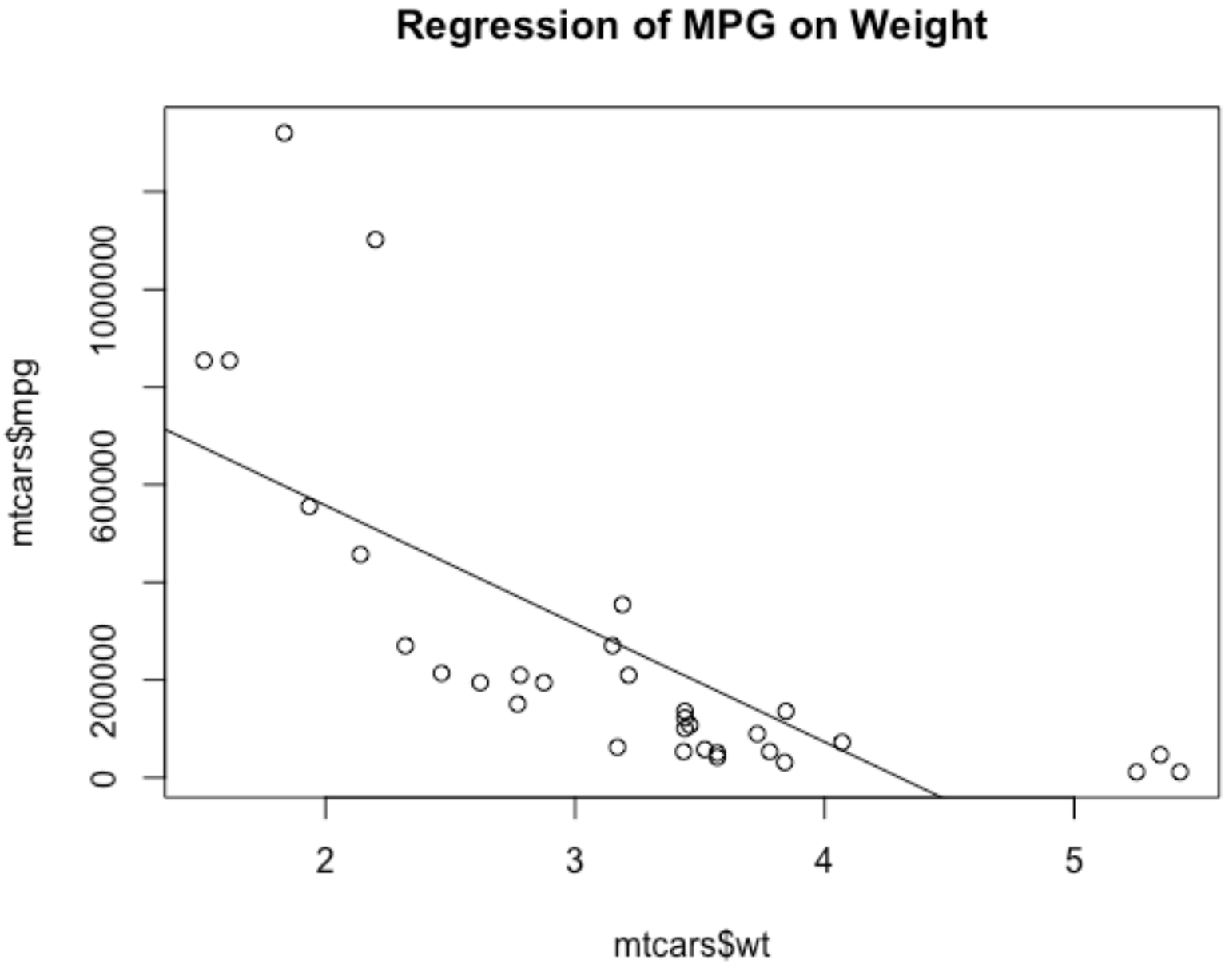

R Graphics Cookbook
R数据可视化手册
[美] Winston Chang 著
肖楠 邓一硕 魏太云 译
邱怡轩 审校
O'REILLY
人民邮电出版社
POSTS & TELECOM PRESS

基本图形

图形参数

一个例子

```
> plot(mtcars$wt, mtcars$mpg)
> abline(lm(mtcars$mpg ~ mtcars$wt))
> title("Regression of MPG on Weight")
```



见RiA课本44页

输入输出

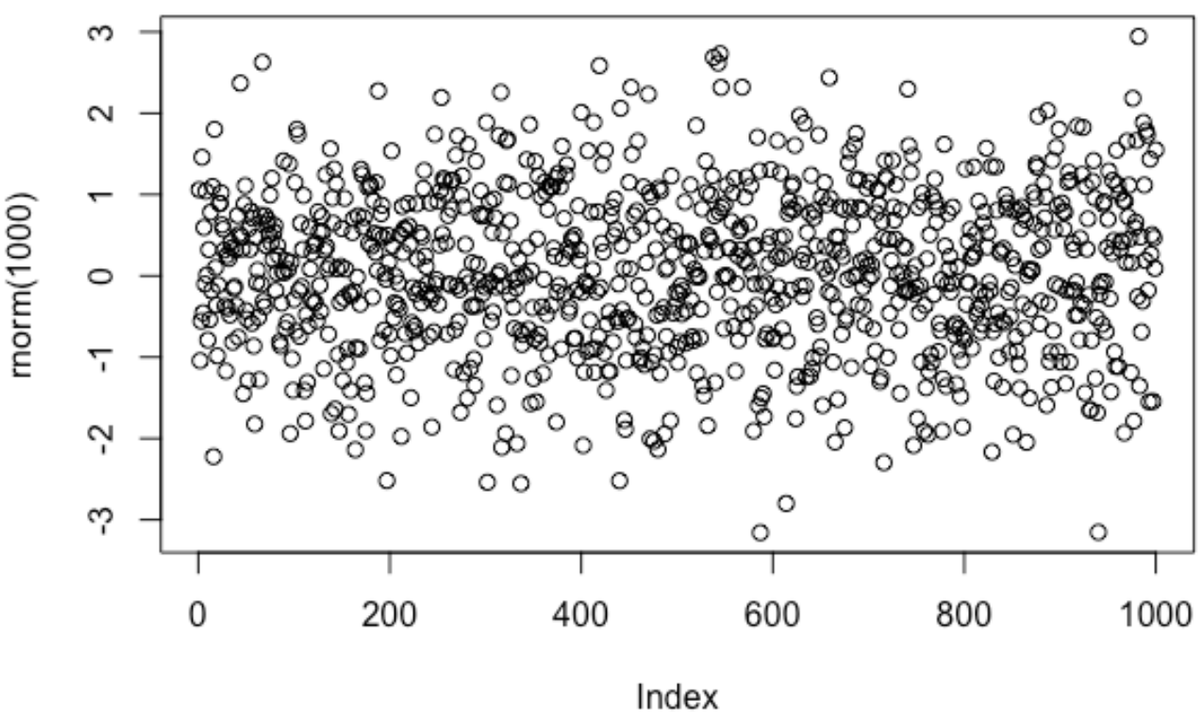
```
png("scatterplot.png")
plot(rnorm(1000))
dev.off()
```

```
png("scatterplot.png", height=600, width=600)
plot(rnorm(1000))
dev.off()
```

```
png("scatterplot.png", height=4, width=4.units="in")
plot(rnorm(1000))
dev.off()
```

```
png("scatterplot.png",res=600)
plot(rnorm(1000))
dev.off()
```

```
pdf("scatterplot.pdf")
plot(rnorm(1000))
dev.off()
```



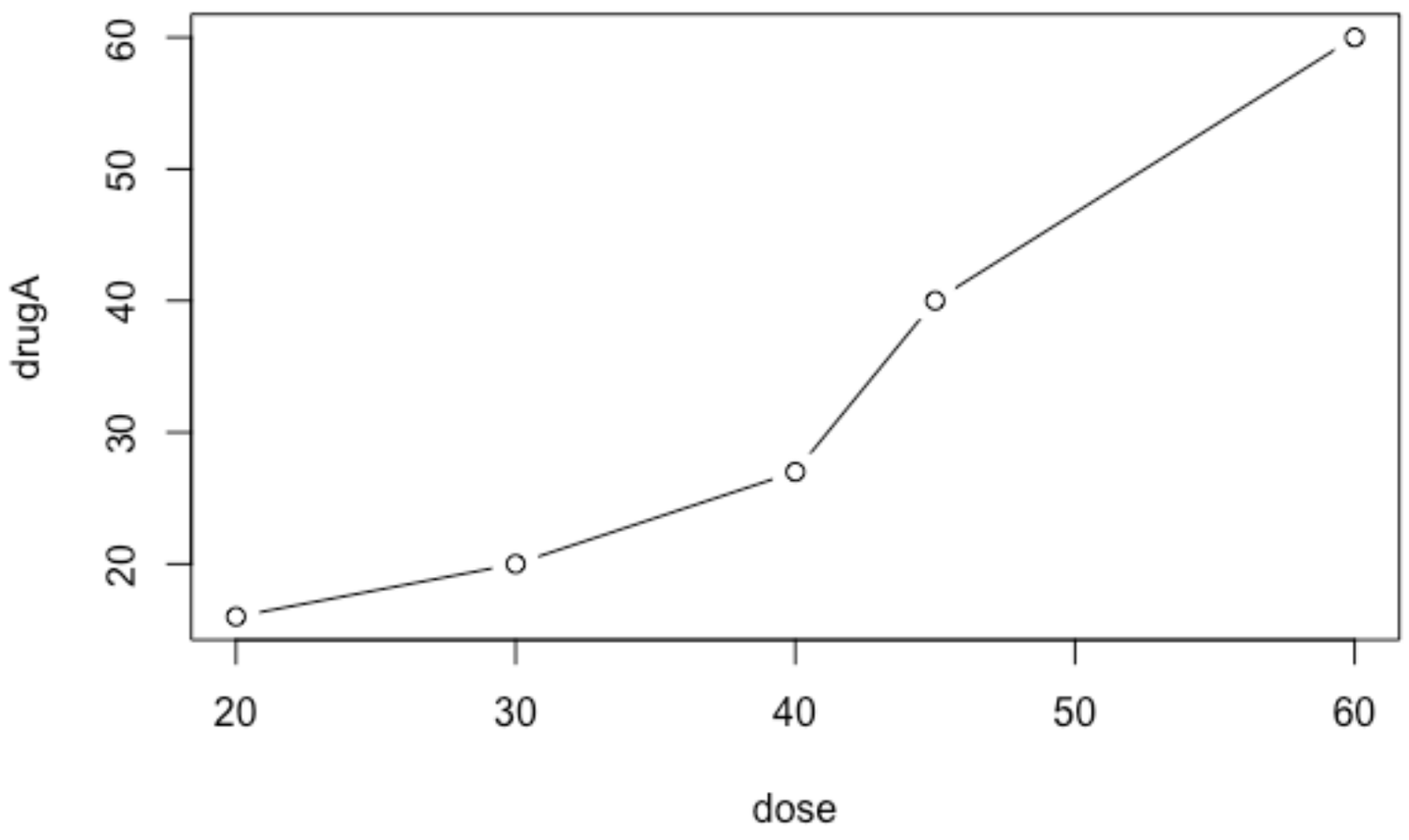
另一个例子

表3-1　病人对两种药物五个剂量水平上的响应情况

剂　　量	对药物A的响应	对药物B的响应
20	16	15
30	20	18
40	27	25
45	40	31
60	60	40

```
> dose <- c(20, 30, 40, 45, 60)
> drugA <- c(16, 20, 27, 40, 60)
> drugB <- c(15, 18, 25, 31, 40)
> plot(dose, drugA, type = "b")
```

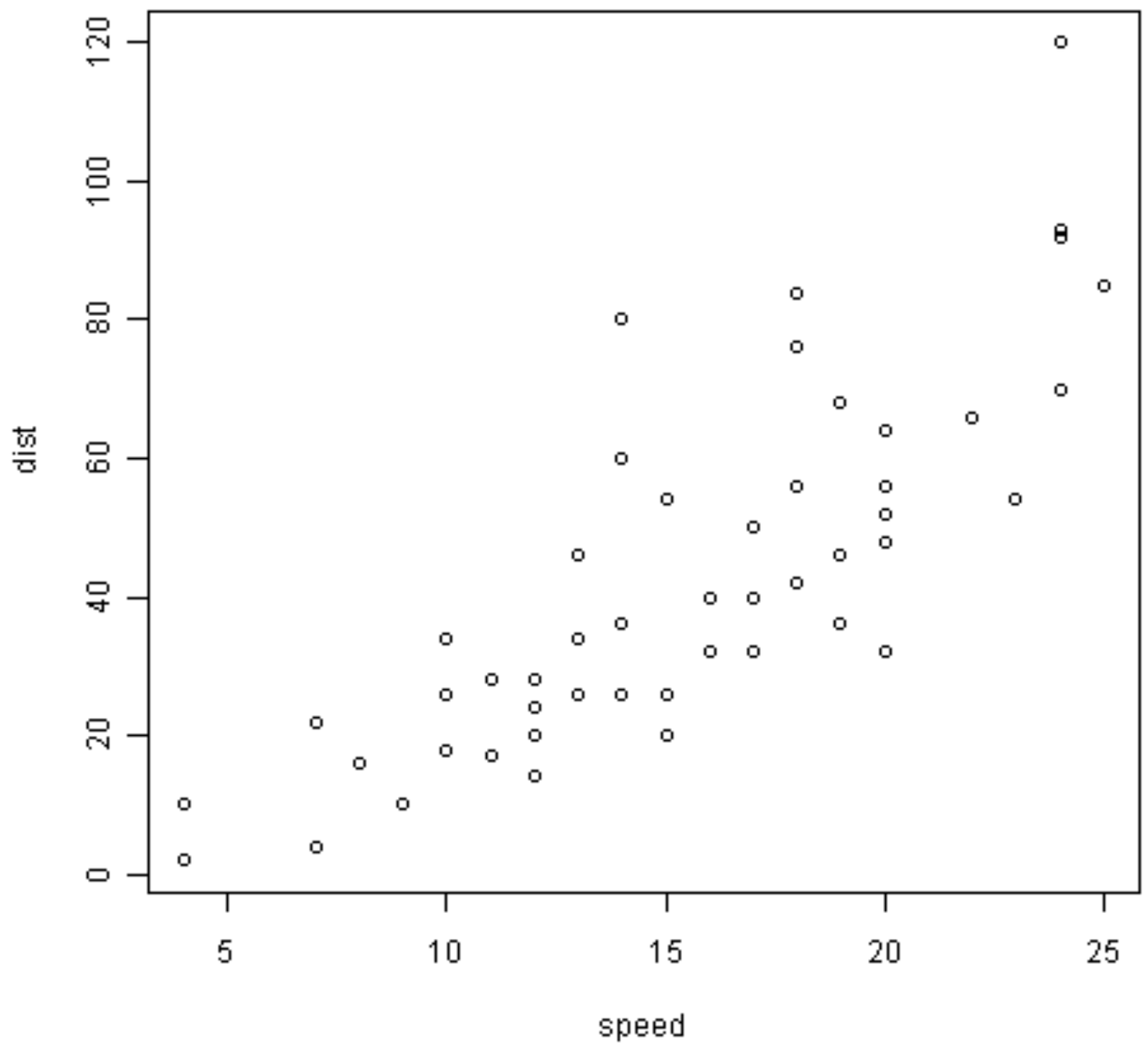
见RiA课本46页



散点图

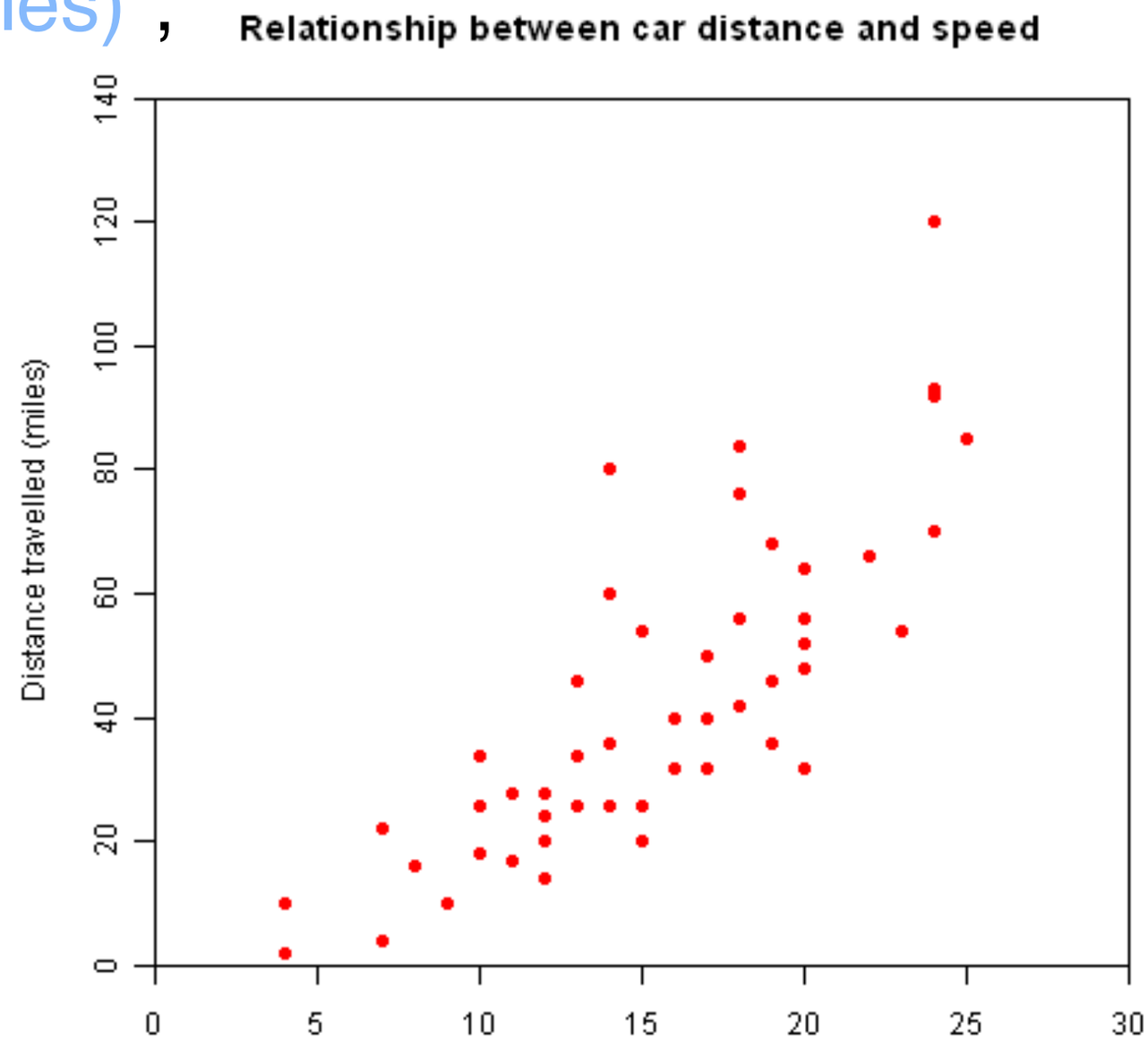
```
> cars
   speed dist
1      4    2
2      4   10
3      7    4
4      7   22
5      8   16
6      9   10
7     10   18
8     10   26
9     10   34
10    11   17
11    11   28
12    12   14
13    12   20
14    12   24
15    12   28
16    13   26
17    13   34
18    13   34
19    13   46
20    14   26
21    14   36
22    14   60
23    14   80
24    15   20
25    15   26
26    15   54
27    16   32
28    16   40
29    17   32
30    17   40
31    17   50
32    18   42
33    18   56
34    18   76
35    18   84
36    19   36
37    19   46
38    19   68
39    20   32
40    20   48
41    20   52
42    20   56
43    20   64
44    22   66
45    23   54
46    24   70
47    24   92
48    24   93
49    24  120
50    25   85
```

plot(cars$dist~cars$speed)



散点图

```
plot(cars$dist~cars$speed,
main="Relationship between car distance & speed",
xlab="Speed (miles per hour)",
ylab="Distance travelled (miles)",
xlim=c(0,30),
ylim=c(0,140),
xaxs="i",
yaxs="i",
col="red",
pch=19)
```

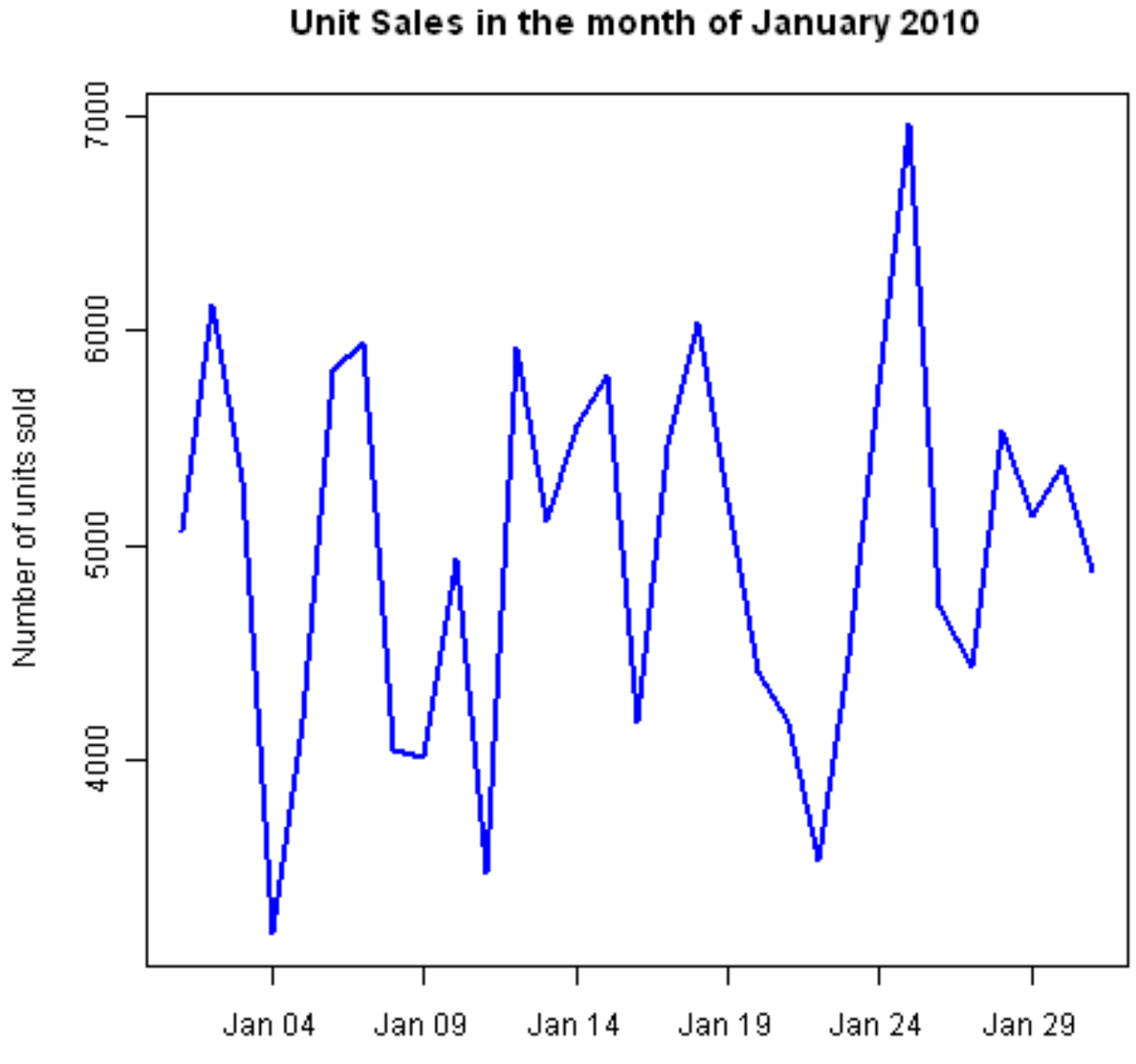


折线图

```
sales  <- read.csv("dailysales.csv",header=TRUE)

plot(sales$units~as.Date(sales$date,"%d/%m/%y"),
type="l",main="Unit Sales in the month of January 2010",
xlab="Date",ylab="Number of units sold",col="blue")
```

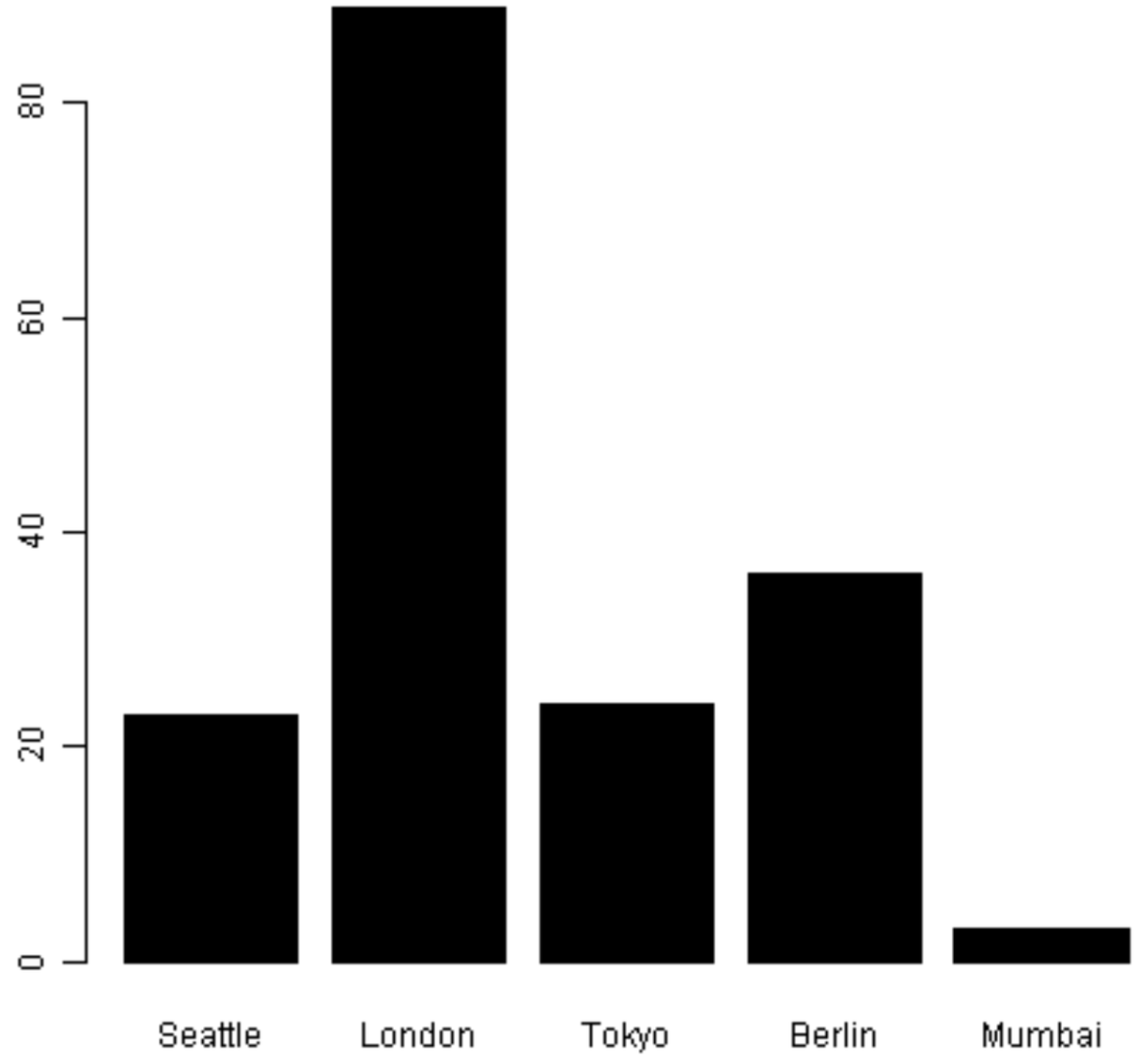
	date	units
1	01/01/2010	5063.782
2	02/01/2010	6115.308
3	03/01/2010	5305.093
4	04/01/2010	3184.974
5	05/01/2010	4181.691
6	06/01/2010	5815.504
7	07/01/2010	5947.141
8	08/01/2010	4048.948
9	09/01/2010	4003.134
10	10/01/2010	4937.259
11	11/01/2010	3470.477
12	12/01/2010	5915.390
13	13/01/2010	5111.493
14	14/01/2010	5563.198
15	15/01/2010	5790.271



条形图

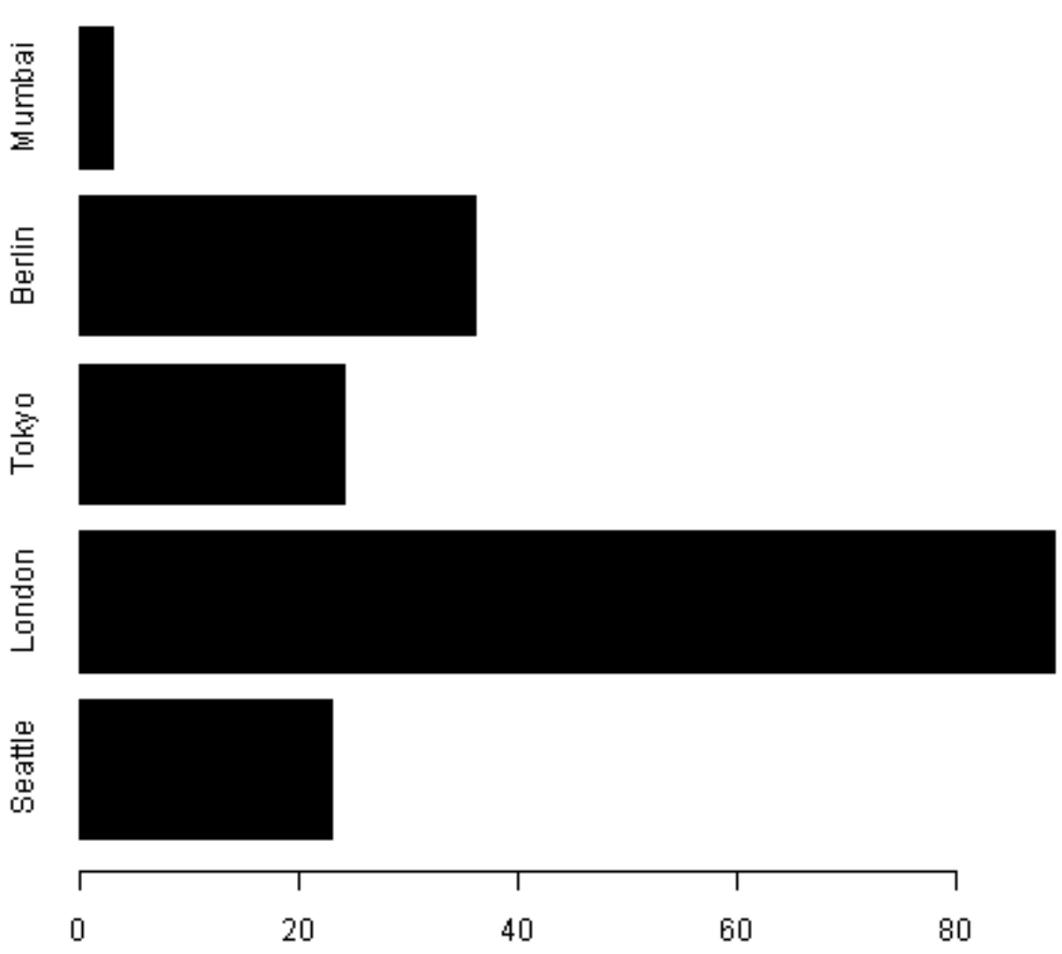
```
sales<-read.csv("citysales.csv",header=TRUE)
barplot(sales$ProductA,names.arg= sales$City,col="black")
```

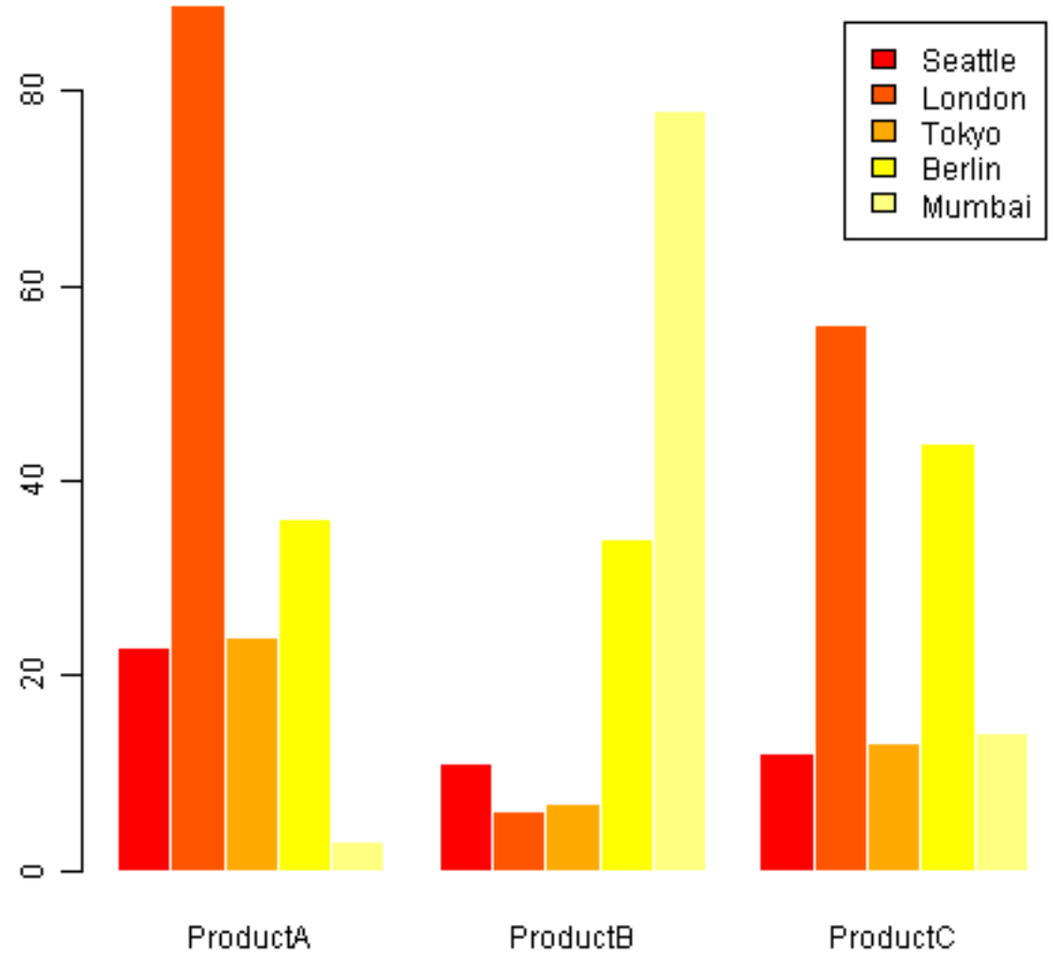
	City	ProductA	ProductB	ProductC
	lumn 0: rownames			
1	Seattle	23	11	12
2	London	89	6	56
3	Tokyo	24	7	13
4	Berlin	36	34	44
5	Mumbai	3	78	14



条形图

```
barplot(sales$ProductA,names.arg= sales$City,
horiz=TRUE,col="black")
```

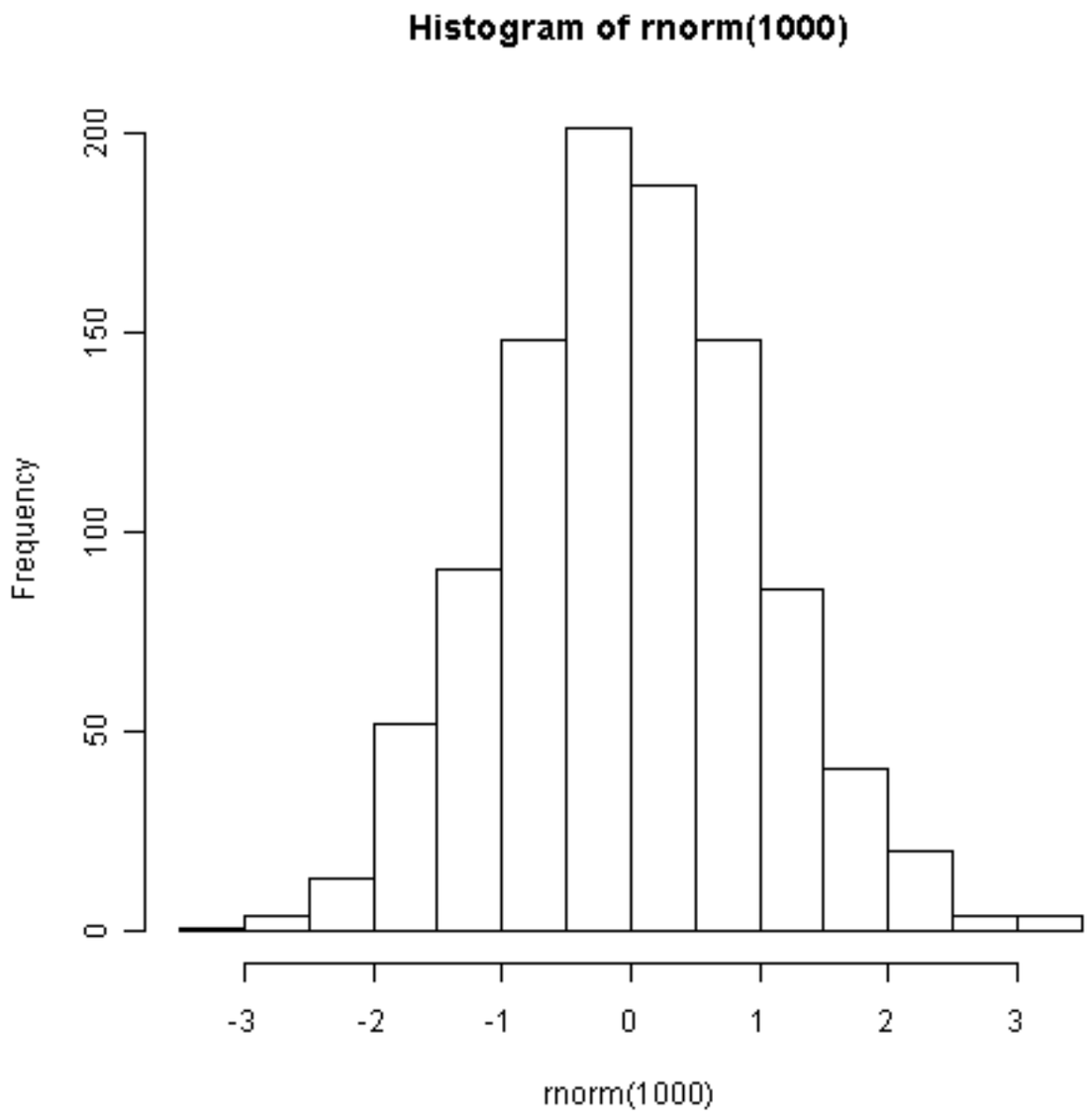




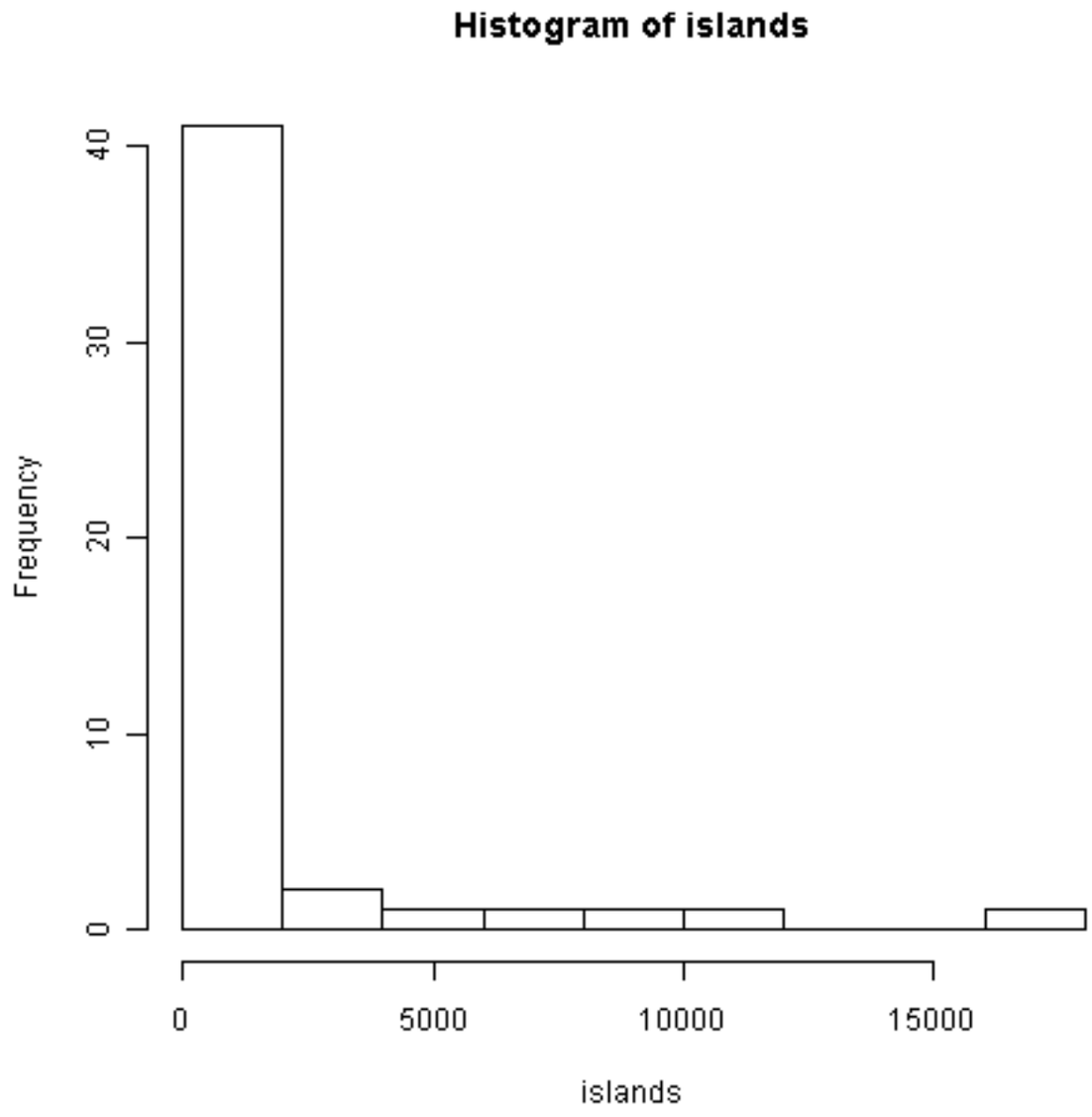

```
barplot(as.matrix(sales[,2:4]), beside=
TRUE,legend=sales$City,col=heat.colors(5),border="white")
```

直方图

hist(rnorm(1000))

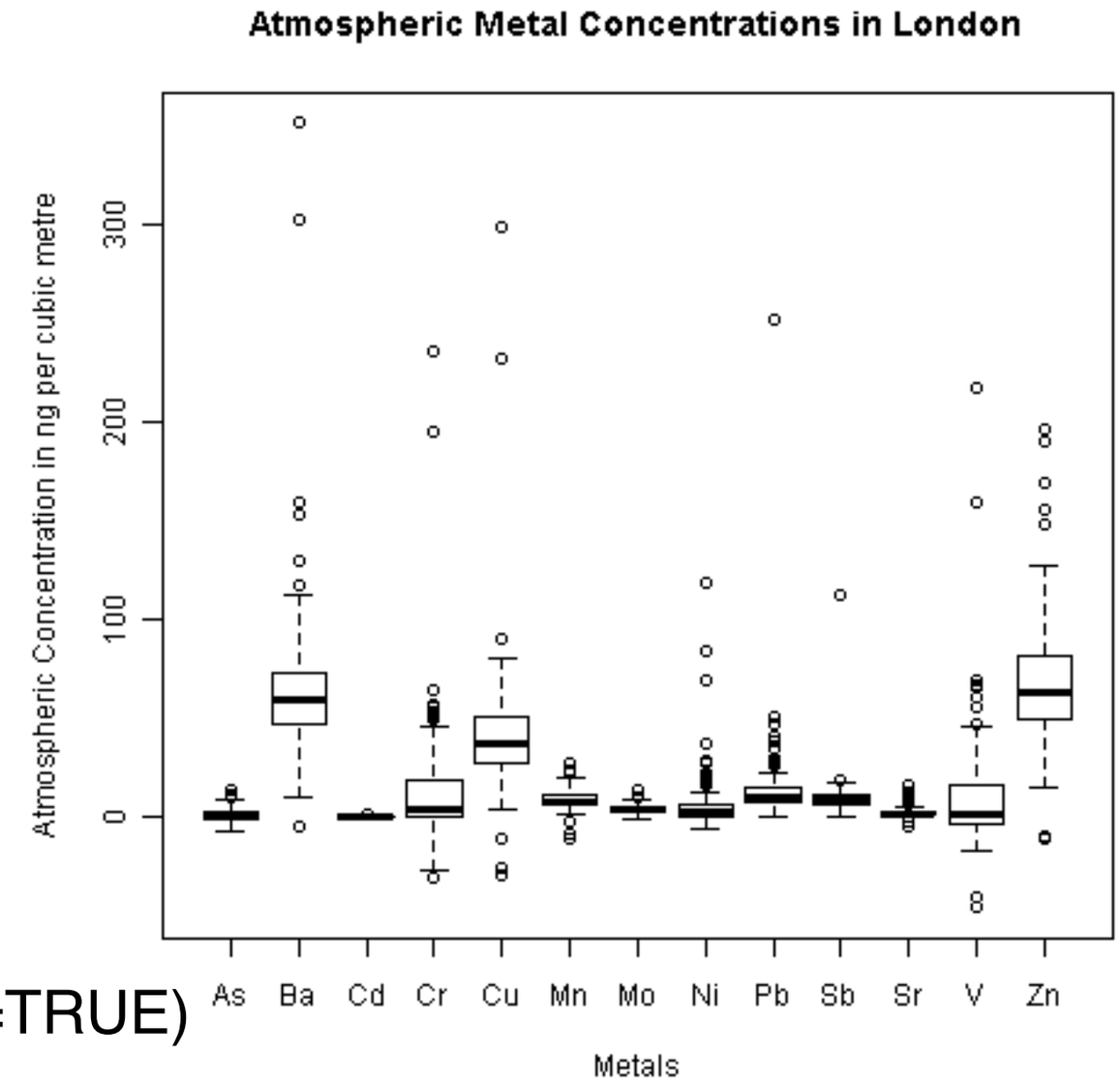


hist(islands)



箱线图

```
metals<-read.csv("metals.csv",header=TRUE)
boxplot(metals,xlab="Metals",ylab="Atmospheric Concentration in ng per cubic metre", main="Atmospheric Metal Concentrations in London")
```




```
copper<-read.csv("copper_site.csv",header=TRUE)
boxplot(copper$Cu~copper$Source, xlab="Measurement Site",ylab="Atmospheric Concentration of Copper in ng per cubic metre",main="Atmospheric Copper Concentrations in London")
```

图形 函数

plot()	画图
barplot()	条形图
pie()	饼图
hist()	直方图
boxplot()	箱线图

基本图形

图形参数

颜色

- 名字：red、blue、black
- 数字：2、4、1
- 十六进制：#FF0000, #0000FF, #000000
- rgb：rgb(1,0,0), (0,0,1),(0,0,0)

- colors()
- colours()

- rainbow()
- heat.colors()
- terrain.colors()
- top.colors()
- cm.colors()
- gray(0:n/n)

见RiA课本49页

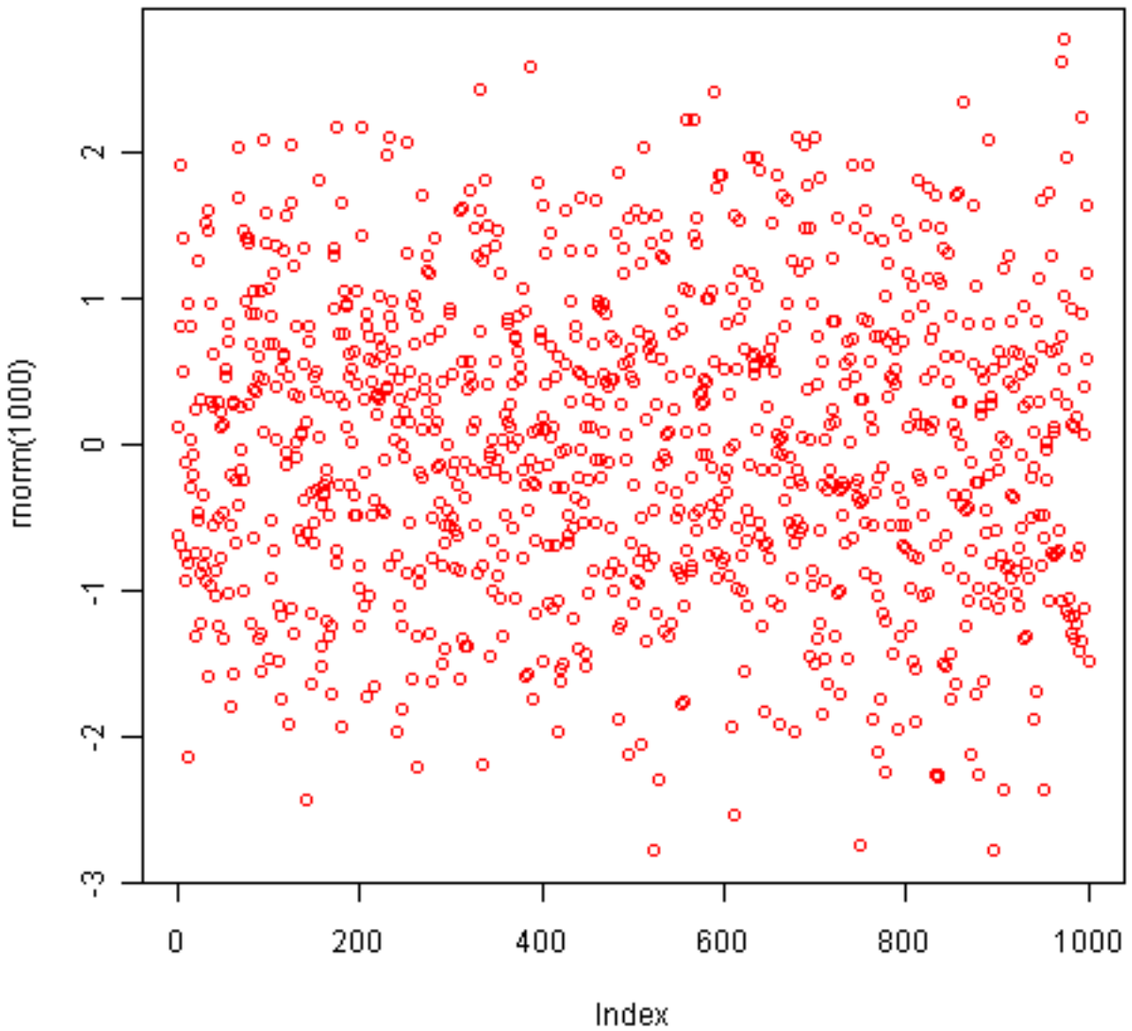
- col的参数可以是一个颜色向量
- palette()　c("red","blue","green","orange")
  palette(c("red","blue","green","orange"))

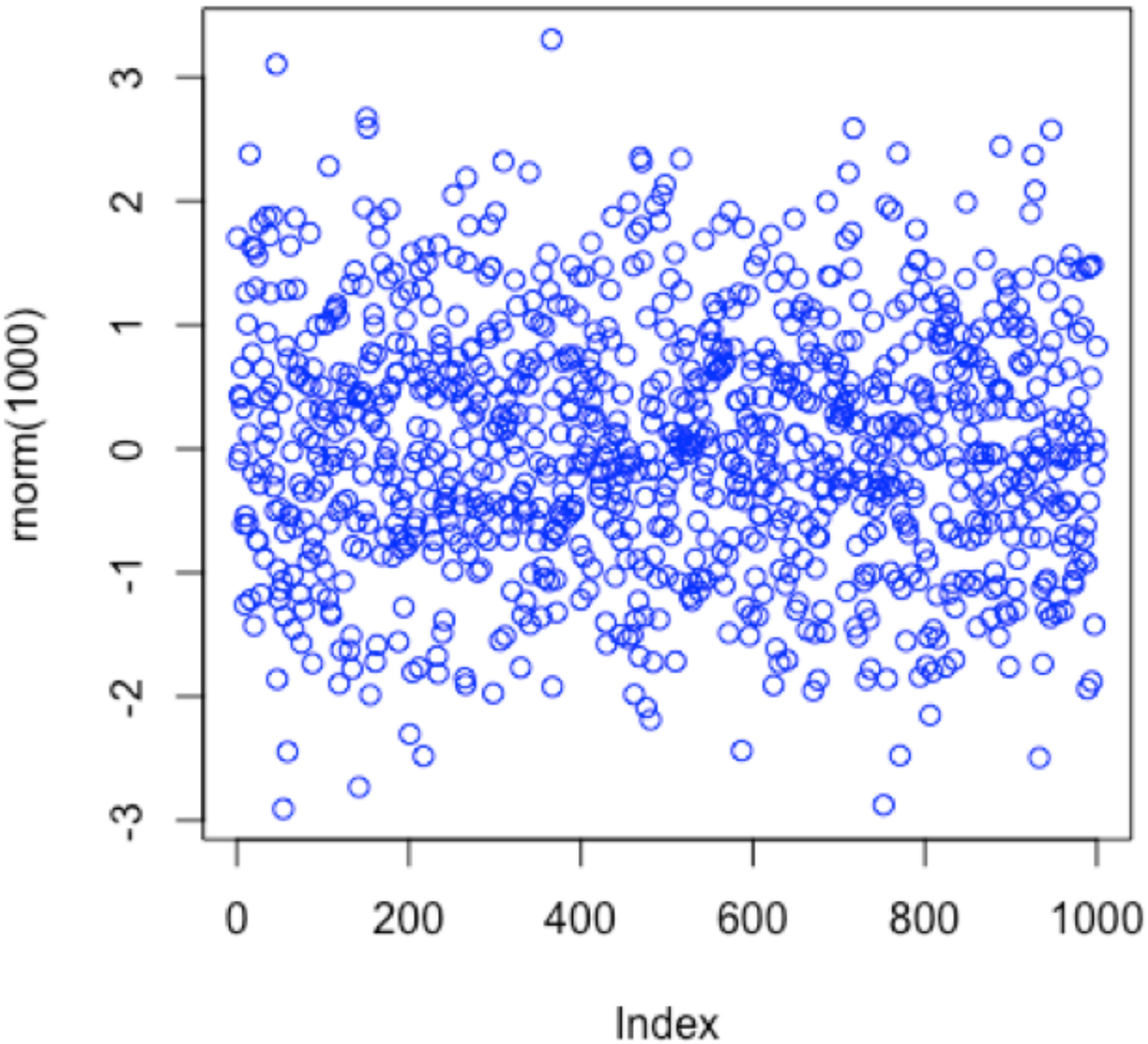
见PACKT课本52页

调整颜色

```
plot(rnorm(1000),col="red")
plot(rnorm(1000),col="blue")
```

自己练习
颜色的各种表示方法

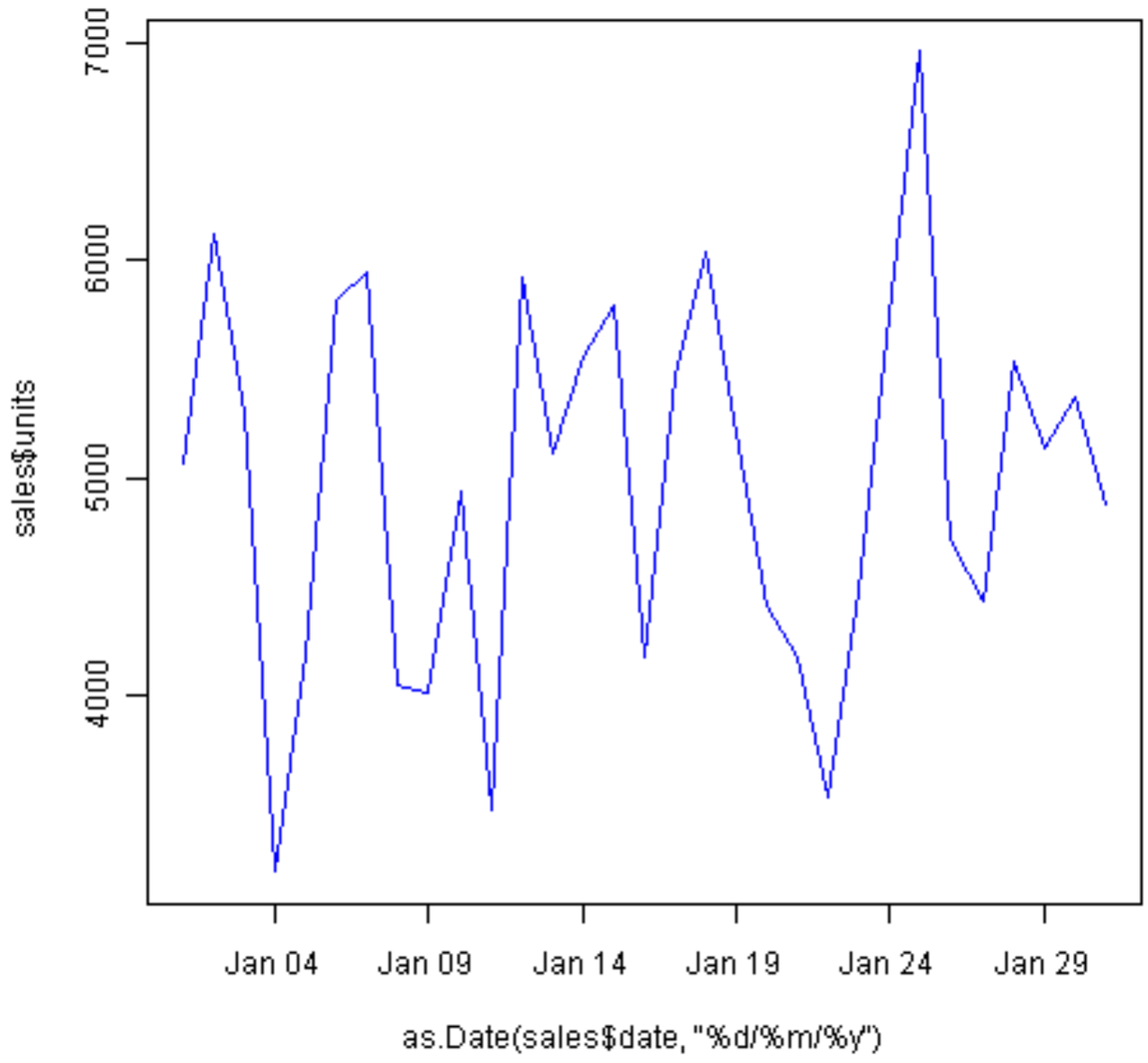




调整颜色

```
Sales <- read.csv("dailysales.csv",header=TRUE)
 plot(Sales$units~as.Date(Sales$date,"%d/%m/%y"),
 type="l", col="blue")
```

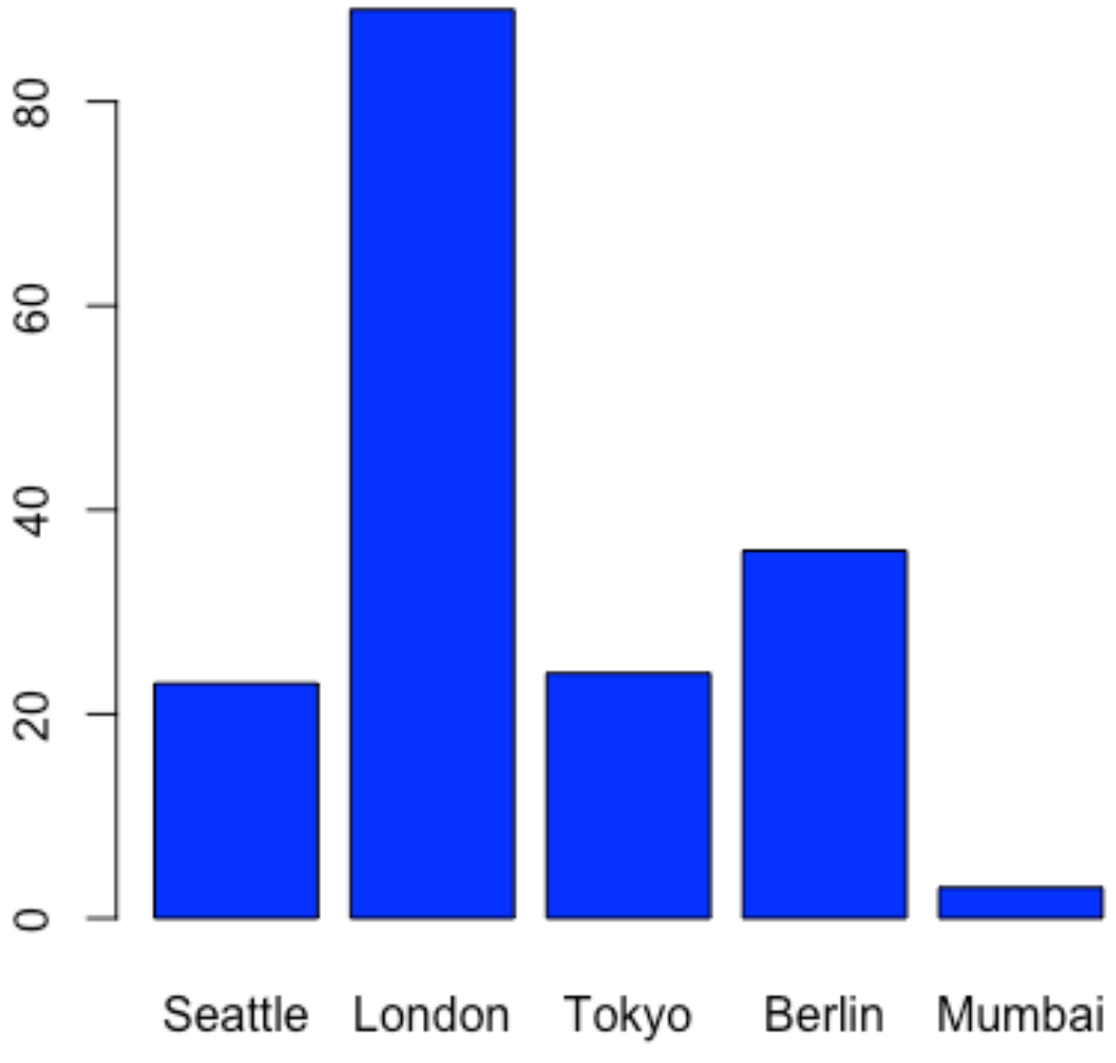
	date	units
1	01/01/2010	5063.782
2	02/01/2010	6115.308
3	03/01/2010	5305.093
4	04/01/2010	3184.974
5	05/01/2010	4181.691
6	06/01/2010	5815.504
7	07/01/2010	5947.141
8	08/01/2010	4048.948
9	09/01/2010	4003.134
10	10/01/2010	4937.259
11	11/01/2010	3470.477
12	12/01/2010	5915.390
13	13/01/2010	5111.493
14	14/01/2010	5563.198
15	15/01/2010	5790.271



调整颜色

```
CitySales <- read.csv("citysales.csv",header=TRUE)
barplot(CitySales$ProductA,names.arg= CitySales$City,
col="blue")
```

	City	ProductA	ProductB	ProductC
1	Seattle	23	11	12
2	London	89	6	56
3	Tokyo	24	7	13
4	Berlin	36	34	44
5	Mumbai	3	78	14



调色板

```
CitySales <- read.csv("citysales.csv",header=TRUE)

barplot(CitySales$ProductA,names.arg= CitySales$City,
col="blue")
```

```
barplot(as.matrix(CitySales[,2:4]), beside=T,
col=c("red","blue","green","orange","pink"),
border="white")

barplot(as.matrix(CitySales[,2:4]), beside=T,
col=c("red","blue","green","orange"),
border="white")
```

```
heat.colors(5)

barplot(as.matrix(CitySales[,2:4]), beside=T,
col=heat.colors(length(CitySales$City)),
border="white")
```

见
PACKT
课本
46-47页
52-53页

自己练习
调色板的各种
表示方法
palette()

颜色的其余参数

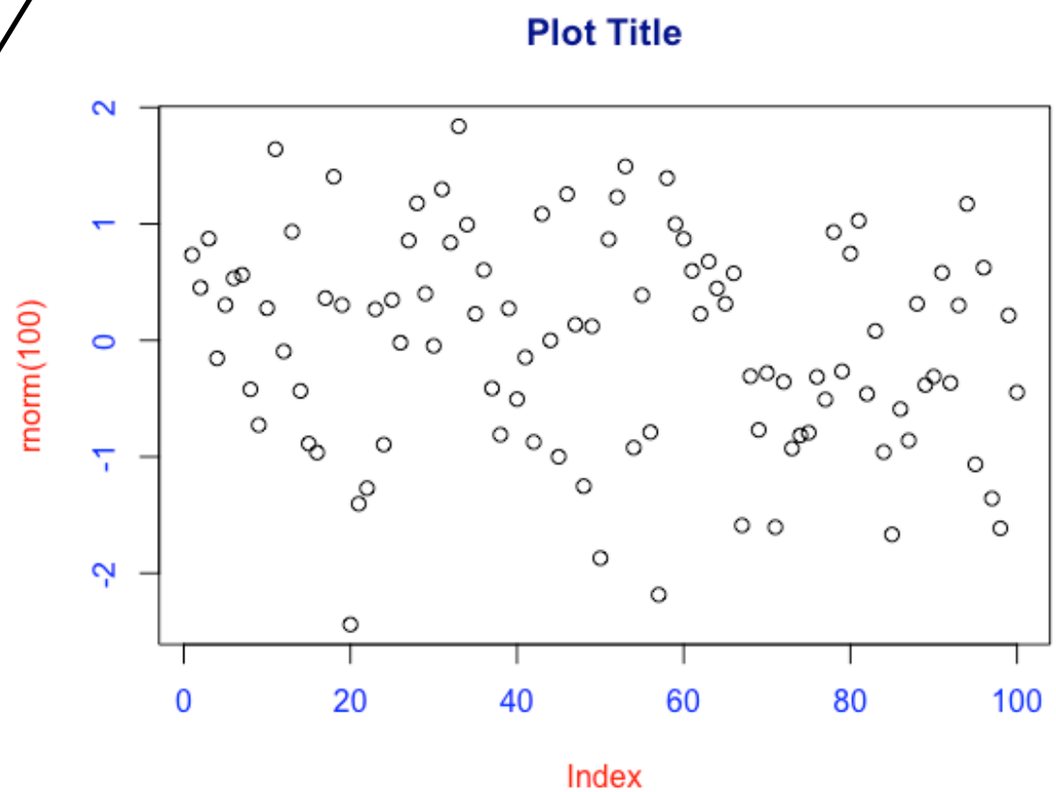
- col.axis： 坐标轴刻度文字的颜色
- col.lab ： 坐标轴标签（名称）的颜色
- col.main： 标题颜色
- col.sub ： 副标题颜色

见
PACKT
课本
48-51页

自己练习
这些颜色参数的表示方法

- fg ： 图形的前景色
- bg ： 图形的背景色

```
plot(rnorm(100),
main="Plot Title",
col.axis="blue",
col.lab="red",
col.main="darkblue")
```



字体

- font ： 字体样式
  - 1：常规
  - 2：粗体
  - 3：斜体
  - 4：粗斜体
  - 5：符号字体
- font.axis： 坐标轴刻度字体样式
- font.lab ： 坐标轴标签（名字）字体样式
- font.main： 标题字体样式
- font.sub ： 子标题字体样式
- family ： 绘制文字的字体族
  - serif
  - sans
  - mono
  - … …
  - windowsFonts()
  - quartzFonts()
  - pdfFonts()

见RiA课本51页

见PACKT课本46-47页

符号与线条

- pch： 绘制适应的符号
- cex： 符号的大小
- lty ： 线条类型
- lwd： 线条宽度

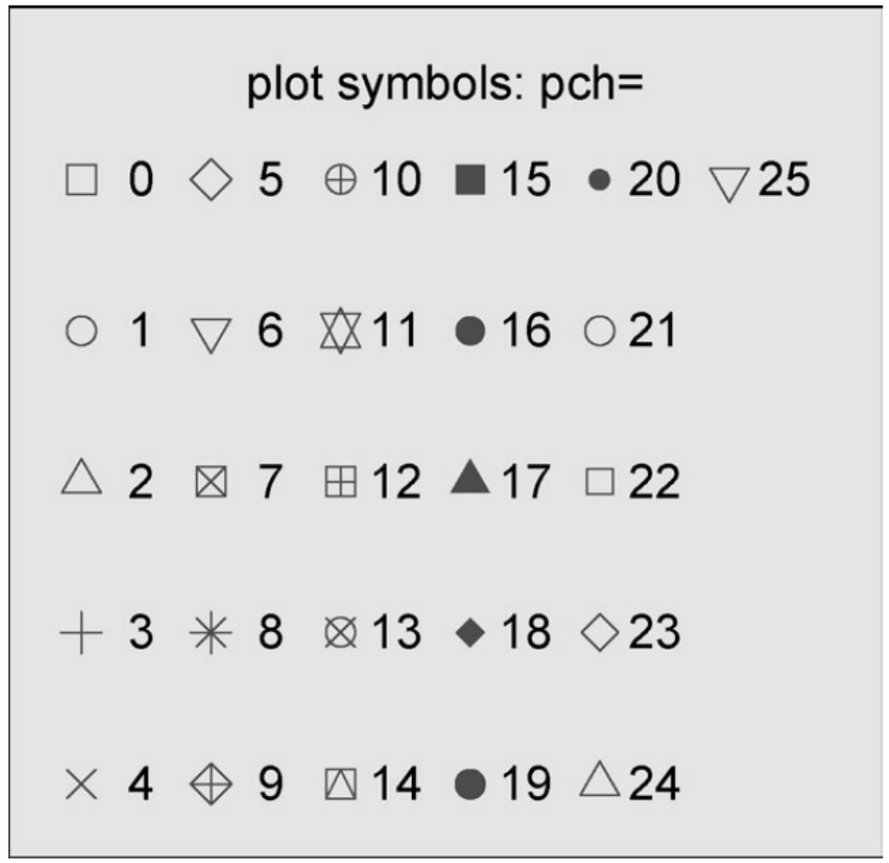


图3-4　参数pch可指定的绘图符号

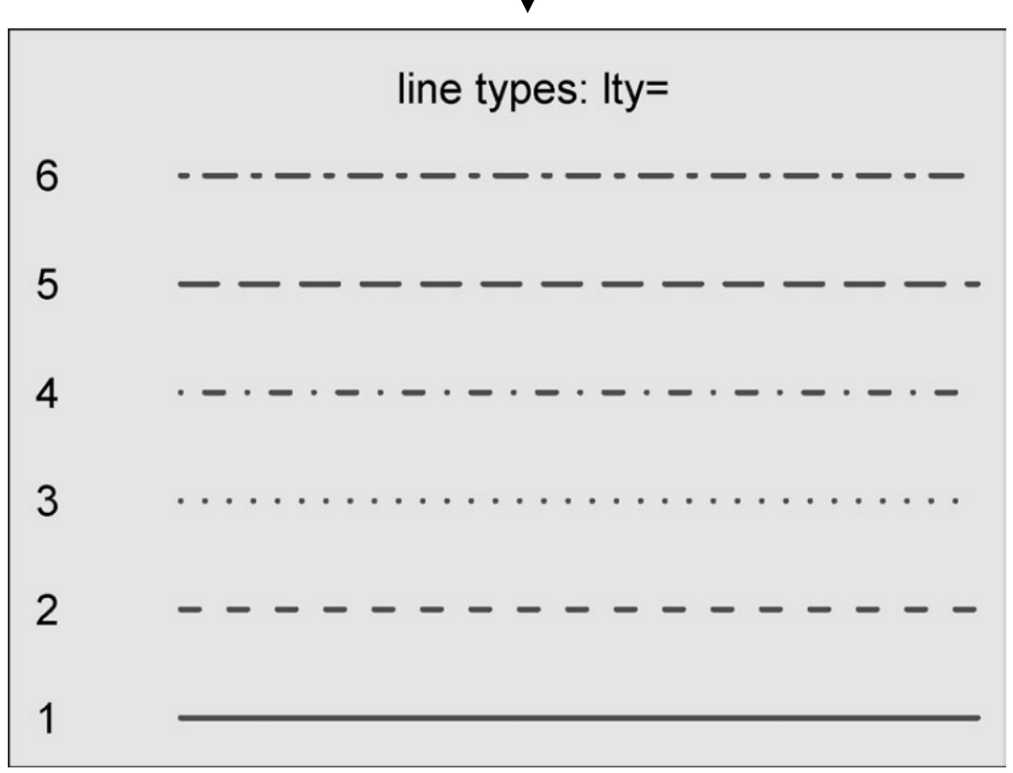


图3-5　参数lty可指定的线条类型

见
RiA
课本
48-49页

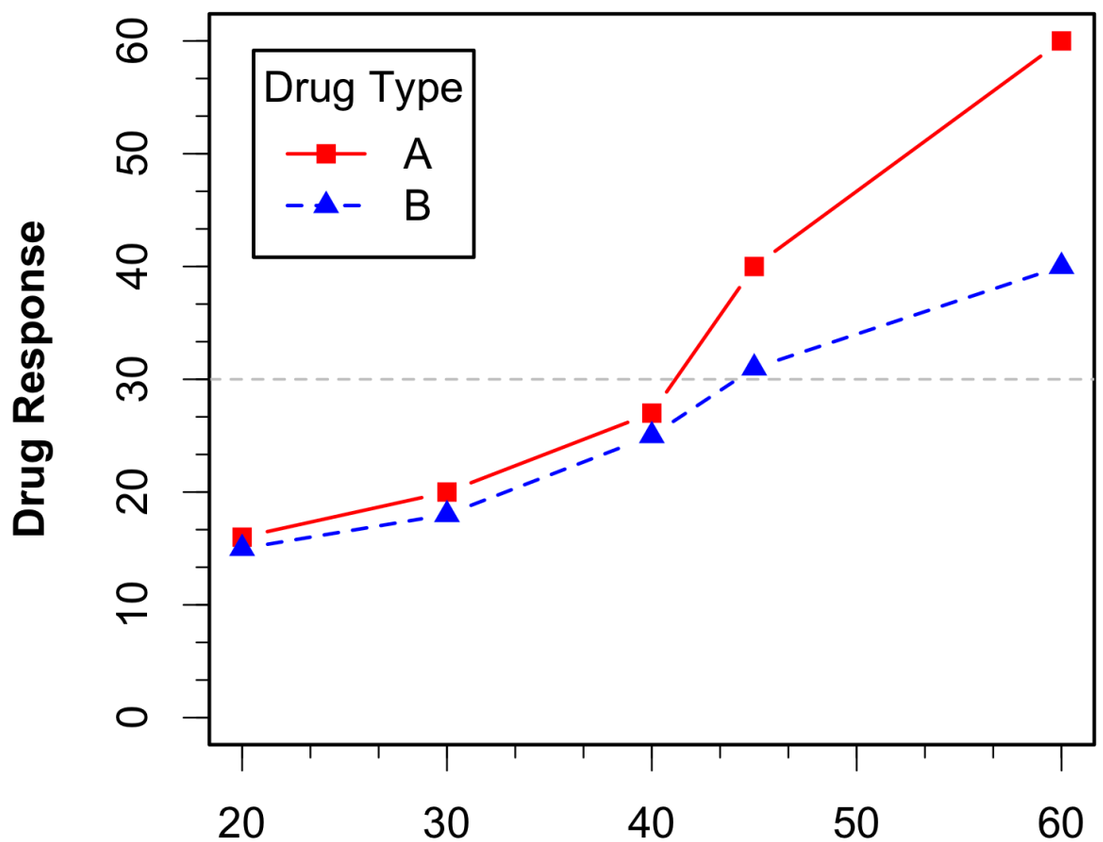
见
PACKT
课本
56-59页

图例

- legend(*location, title, legend, …*)
  - ➡ location：位置
  - ➡ title　　：图例标题
  - ➡ legend　：图例标签向量

见 RiA 课本 57页

- bottom
- bottomleft
- left
- topleft
- top
- topright
- right
- bottomright
- center



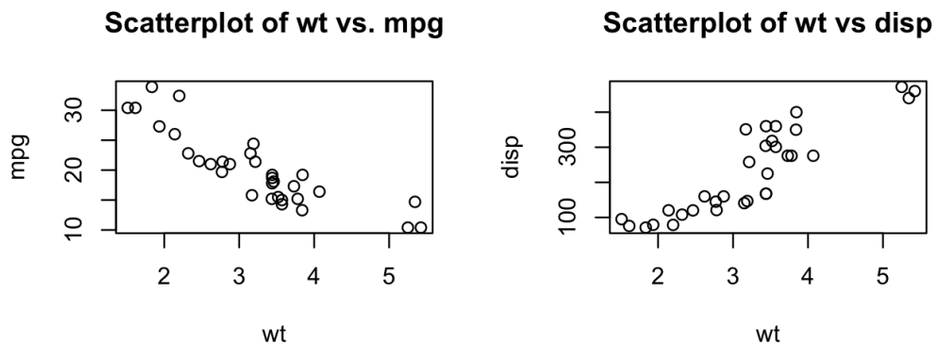

```
legend(
"topleft",
inset = 0.05,
title = "Drug Type",
c("A", "B"),
lty = c(1, 2),
pch = c(15, 17),
col = c("red","blue"))
```

图形组合

- par(mfrow=c(nrows,ncols)), 按行填充
- par(mfcol=c(nrows,ncols)), 按列填充

见 RiA 课本 61-63页

- layout(mat)
- layout(matrix(c(1,1,2,3), 2, 2, byrow=TRUE))



其余函数和参数

- titile()：添加标题
- abline()：添加参考线
- text()：将文本添加到图形
- mtext()：同上
- line()：在图形上划线
- log=“x”,y,xy：log坐标

提问时间！

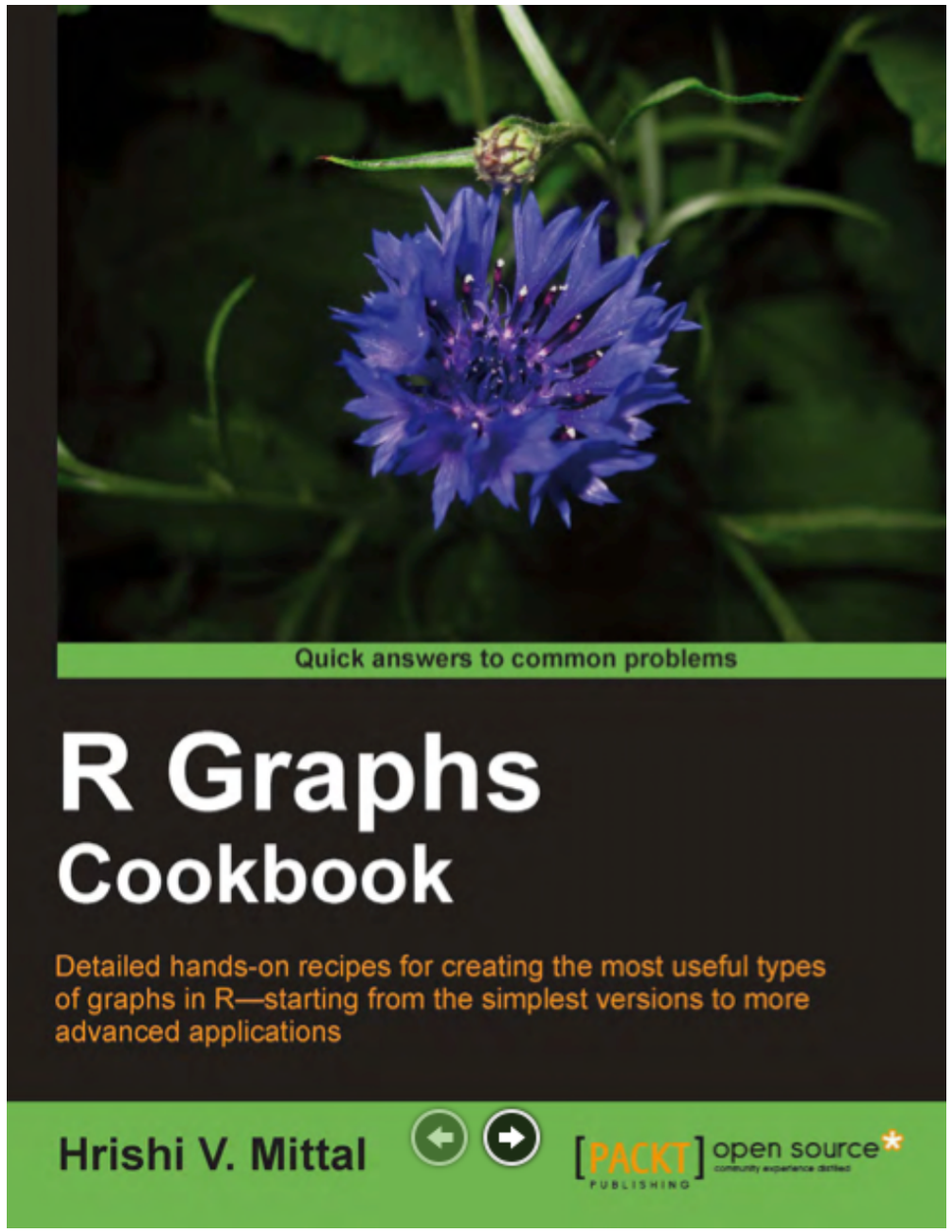
孙惠平

sunhp@ss.pku.edu.cn

练习

练习 - 0014

第三、六章



第一、二章

练习 - 0015

- 模拟产生100个学号（1300022001到1300022100）
- 模拟产生三个科目的成绩，要求第一科最大值99，最小值70；第二科平均值81，sd=7，最大值100；第三科平均值83，sd＝18，最大值100
- 把学号和三科成绩组成一个数据框，显示数据框内容
- 求每个学生的总分、平均分
- 针对三科成绩、总分、平均分，分别做饼图、直方图、条形图，箱线图
- 分别用par和layout把多个图放在一个图中显示：同一个数据的不同类的图形，不同数据的同一类，不同数据的不同图形

练习－0016

- 某校测的19名学生的四项指标：性别、年龄、身高（cm）、体重（磅），具体见0016_student.CSV，要求：
  - ✳ 绘出体重对于身高的散点图
  - ✳ 绘出不同性别情况下，体重与身高的散点图
  - ✳ 绘出不同年龄段的体重与身高的散点图
  - ✳ 绘出不同性别和不同年龄段的体重与身高的散点图

- 0016_height01.txt，画直方图
- 0016_height02.txt，画箱式图
- 0016_marriage.txt，画散点图
- 0016_language.txt，画条形图（母语和日常使用）
- 0016_language.txt，画饼图（世界主要语种使用人数比例）

练习 – 0017

- 从0017_grade.csv中读取两班成绩
- 计算每个班级的均值和标准方差
- 计算每个人的标准化成绩，添加到数据中，写到0017grade.txt中
- 分别画出来两班成绩和标准成绩的箱线图
- 在一张图中画出两班成绩和标准成绩的箱线图

练习 - 0018

plot(rnorm(1000),**col=“red"**)

- 使用上面的语句，练习颜色的各种表示方法
- 使用Par和layout函数，分别现实不同颜色的多个图形组合，2*2，3*3，1*1*2*3等

课件第12页，citysales.csv

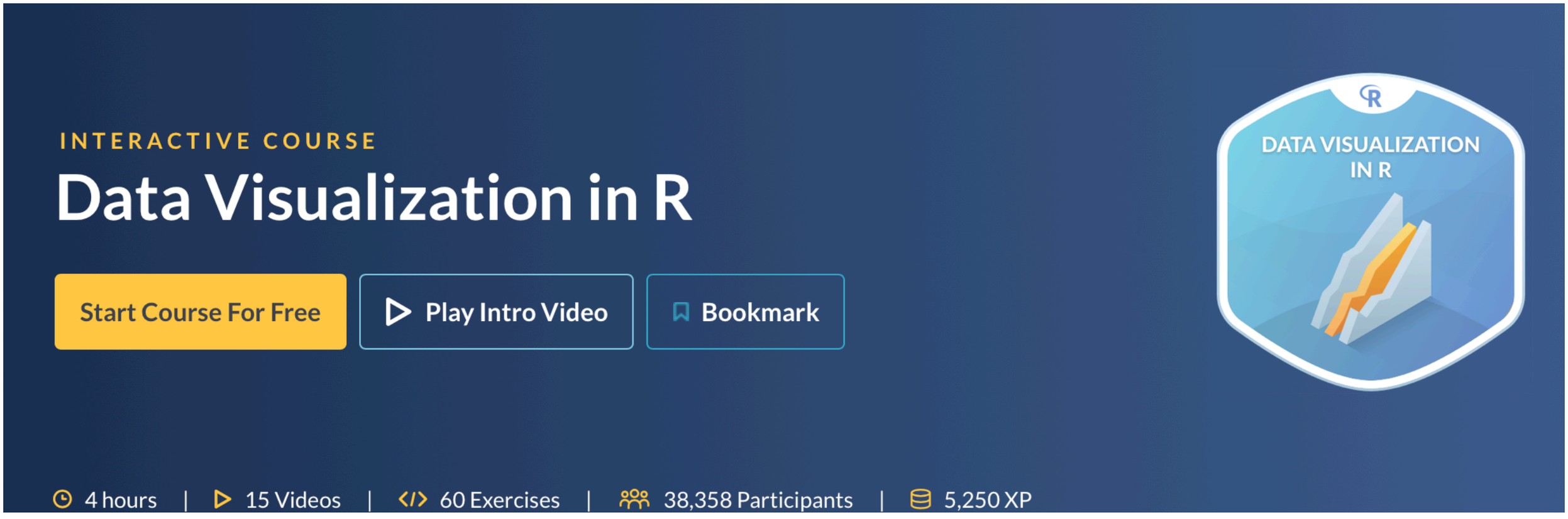
- 输入现有代码，看显示结果
- 用rainbow、top.colors、cm.colors、gray、terrian.colors替换heat.colors，看执行效果
- 练习课件第23页的颜色参数
- 添加图例

练习 - 0019

cityrain.csv

- 用不同颜色画出不同城市的线图
- 用不同符号画出不同城市的线图
- 用不同颜色画出不同城市的散点图
- 用不同符号画出不同城市的散点图
- 分别加上图例
- 用par和layout把前面四个图放在一张图中，分别为2*2，1*4，1+2+1

练习 – 0020



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