# Contents

**数据分析与应用（选修，32 学时）** **1**
课程简介 . . . . . 1
为何选择 R . . . . . 1
R 可以做什么 . . . . . 2
R 语言的知识体系 . . . . . 2
R 语言中文社区 . . . . . 3
开发环境安装 . . . . . 3
R 的使用 . . . . . 3

# 数据分析与应用（选修，32 学时）

- 教师：李俊平，计算机工程学院计算机信息管理专业，ljp@szpt.edu.cn；
- 考核方式：平时 60%，期末数据分析报告 40%；

## 课程简介

- 大数据背景介绍；
- R 语言和 Python 语言在数据分析中的地位；
- 以 R 语言为蓝本讲解数据分析的基本步骤和技巧；

## 为何选择 R

R 语言作为统计学一门语言，一直在小众领域闪耀着光芒。直到大数据的爆发，R 语言变成了一门炙手可热的数据分析的利器。随着越来越多的工程背景的人的加入，R 语言的社区在迅速扩大成长。现在已不仅仅是统计领域，教育，银行，电商，互联网.... 都在使用 R 语言。

与起源于贝尔实验室的 S 语言类似，R 也是一种为统计计算和绘图而生的语言和环境，它是一套开源的数据分析解决方案，由一个庞大且活跃的全球性研究型社区维护。但是，市面上也有许多其他流行的统计和制图软件，如 Microsoft Excel、SAS、IBM SPSS、Stata 以及 Minitab。为何偏偏要选择 R？

R 有着非常多值得推荐的特性。

- 多数商业统计软件价格不菲，投入成千上万美元都是可能的。而 R 是免费的！如果你是一位教师或一名学生，好处显而易见。
- R 是一个全面的统计研究平台，提供了各式各样的数据分析技术。几乎任何类型的数据分析工作皆可在 R 中完成。
- R 囊括了在其他软件中尚不可用的、先进的统计计算例程。事实上，新方法的更新速度是以周来计算的。
- R 拥有顶尖水准的制图功能。如果希望复杂数据可视化，那么 R 拥有最全面且最强大的一系列可用功能。
- R 是一个可进行交互式数据分析和探索的强大平台。
- 从多个数据源获取并将数据转化为可用的形式，可能是一个富有挑战性的议题。R 可以轻松地从各种类型的数据源导入数据，包括文本文件、数据库管理系统、统计软件，乃至专门的数据仓库。它同样可以将数据输出并写入到这些系统中。R 也可以直接从网页、社交媒体网站和各种类型的在线数据服务中获取数据。
- R 是一个无与伦比的平台，在其上可使用一种简单而直接的方式编写新的统计方法。它易于扩展，并为快速编程实现新方法提供了一套十分自然的语言。
- R 的功能可以被整合进其他语言编写的应用程序，包括 C++、Java、Python、PHP、Pentaho、SAS 和 SPSS。这让你在继续使用自己熟悉语言的同时在应用程序中加入 R 的功能。
- R 可运行于多种平台之上，包括 Windows、UNIX 和 Mac OS X。这基本上意味着它可以运行于你所能拥有的任何计算机上。
- 如果你不想学习一门新的语言，有各式各样的 GUI（Graphical User Interface，图形用户界面）工具通过菜单和对话框提供了与 R 语言同等的功能。

## R 可以做什么

R 应用最热门的领域：

- 统计分析：包括统计分布，假设检验，统计建模
- 金融分析：量化策略，投资组合，风险控制，时间序列，波动率
- 数据挖掘：数据挖掘算法，数据建模，机器学习
- 互联网：推荐系统，消费预测，社交网络
- 生物信息学：DNA 分析，物种分析
- 生物制药：生存分析，制药过程管理
- 全球地理科学：天气，气候，遥感数据
- 数据可视化：静态图，可交互的动态图，社交图，地图，热图，与各种 Javascript 库的集成。

## R 语言的知识体系

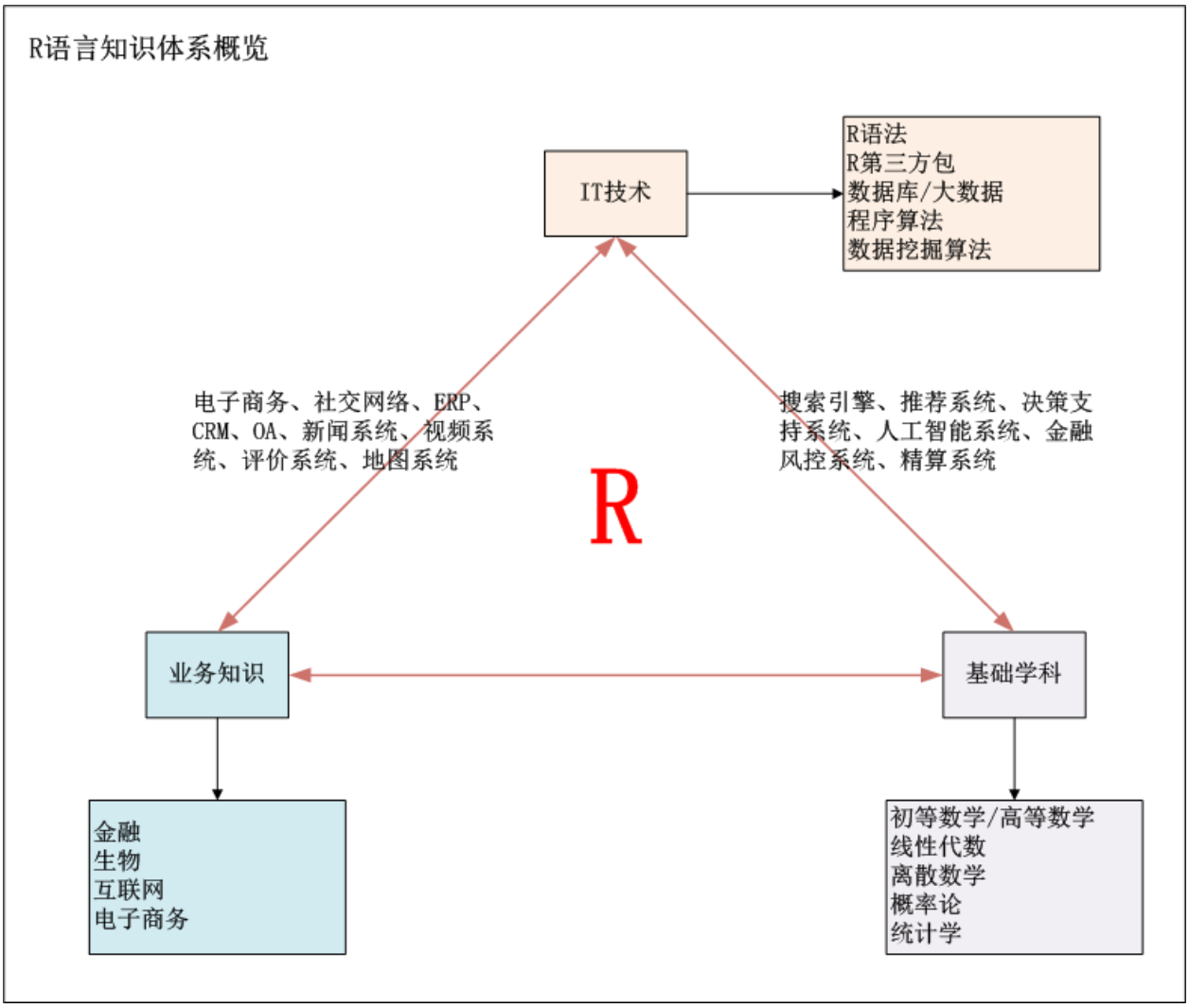



Figure 1:

R 语言知识体系结构分为 3 个部分：IT 技术 + 业务知识 + 基础学科。

- IT 技术：是计算时代必备的技术之一，R 语言就是一种我们应该要掌握技术。

- 业务知识：是市场经验和法则，不管你在什么公司，都会有自己的产品、销售、市场等，你要了解你的公司产品有什么，客户是谁，怎么才能把产品卖给你的客户。
- 基础学科：是我们这十几年在学校学的理论知识，当初学的时候并不知道是为了什么，毕业后如果你还能掌握一些知识并实际运用，那么这将是你最有价值的竞争力。

每个部分知识单独看都有其局限性，但如果能把知识两两结合起来，就构成了我们现在社会的各种技术创新点。

- IT 技术 + 业务知识：创造了阿里巴巴的电子商务帝国，腾讯全生态链的社交网络。
- IT 技术 + 基础学科：创造了 Google 搜索的神话，华尔街金融不败的帝国。

## R 语言中文社区

- 统计之都，中国大陆最权威的 R 语言组织，不仅积累了大量高质量的 R 语言文章，并主办了七届中国 R 语言会议。统计之都团队成员，还参与翻译了《R 语言编程艺术》、《R 语言实战》、《ggplot2: 数据分析与图形艺术》、《R 语言核心技术手册 (第 2 版)》、《R 数据可视化手册》、《R 语言统计入门 (第 2 版)》等多本图书。
- 炼数成金论坛，以数据分析为主题，设有 R 语言板块，提供在线的 R 语言入门培训。
- 人大经济论坛，以经管教育为主题，设有 R 语言板块，以线下培训为主。

## 开发环境安装

- 安装 R；
- 安装 RStudio（方便地进行编程，丰富的提示功能，RNoteBook 的强大）；
- 设置 CRAN 的镜像为国内网站（安装包的时候网速快）；
- 创建一个 RNotebook 文件，按提示安装相应的包；
- 安装常用的包：swirl、ggplot2、xlsx、dplyr、stringr 等。

## R 的使用

- 赋值：<-；例如：x <- 5
- 注释：#；例如：# 这是一行注释
- 一个非常好的在线学习环境：swirl 包

### 新手上路

让我们通过一个简单的虚构示例来直观地感受一下这个界面。假设我们正在研究生理发育问题，并收集了 10 名婴儿在出生后一年内的月龄和体重数据。我们感兴趣的是体重的分布及体重和月龄的关系。

10 名婴儿的月龄和体重

| 年龄（月） | 体重（kg） | 年龄（月） | 体重（kg） |
|---|---|---|---|
| 01 | 4.4 | 09 | 7.3 |
| 03 | 5.3 | 03 | 6.0 |
| 05 | 7.2 | 09 | 10.4 |
| 02 | 5.2 | 12 | 10.2 |
| 11 | 8.5 | 03 | 6.1 |

可以使用函数 c() 以向量的形式输入月龄和体重数据，此函数可将其参数组合成一个向量或列表。然后用 mean() 、sd() 和 cor() 函数分别获得体重的均值和标准差，以及月龄和体重的相关度。最后使用 plot() 函数，从而用图形展示月龄和体重的关系，这样就可以用可视化的方式检查其中可能存在的趋势。函数 q() 将结束会话并允许你退出 R。

```
# 第一个实例：婴儿体重和月龄的关系
age <- c(1,3,5,2,11,9,3,9,12,3)
```

```
weight <- c(4.4,5.3,7.2,5.2,8.5,7.3,6.0,10.4,10.2,6.1)
mean(weight)
```

```
## [1] 7.06
```

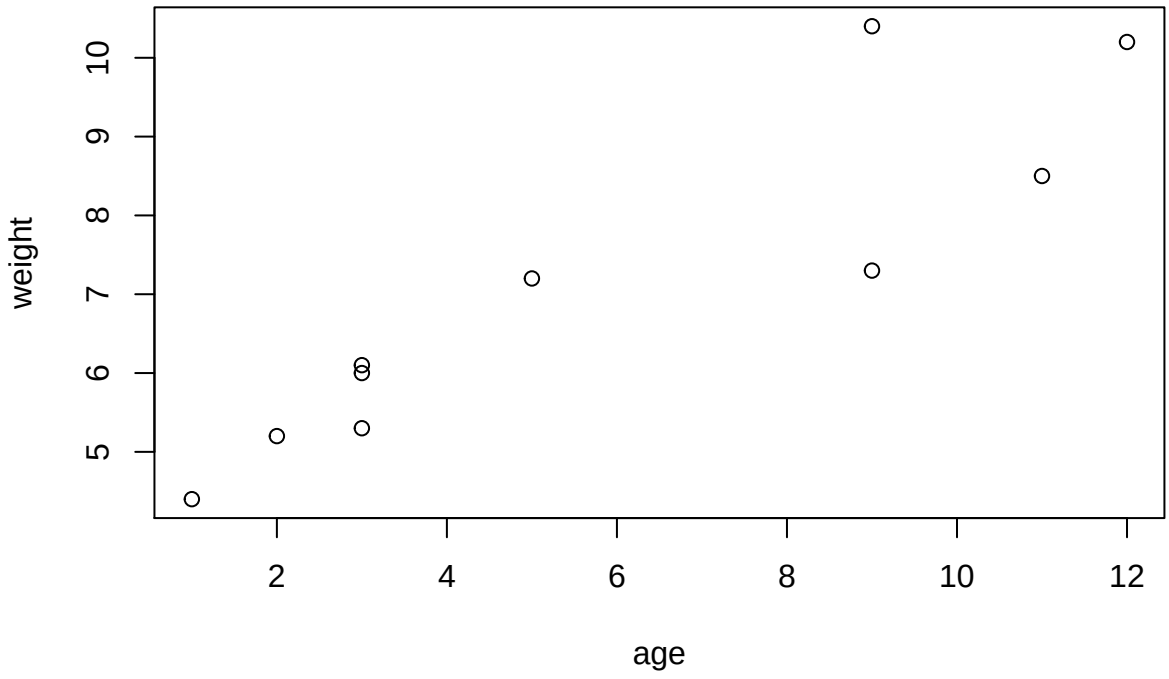
```
sd(weight)
```

```
## [1] 2.077498
```

```
cor(age,weight)
```

```
## [1] 0.9075655
```

```
plot(age,weight)
```




```
# q()
```

从代码清单和结果中可以看到，这 10 名婴儿的平均体重是 7.06kg，标准差为 2.08kg，月龄和体重之间存在较强的线性关系（相关度 =0.91）。这种关系也可以从图散点图中看到。不出意料，随着月龄的增长，婴儿的体重也趋于增加。

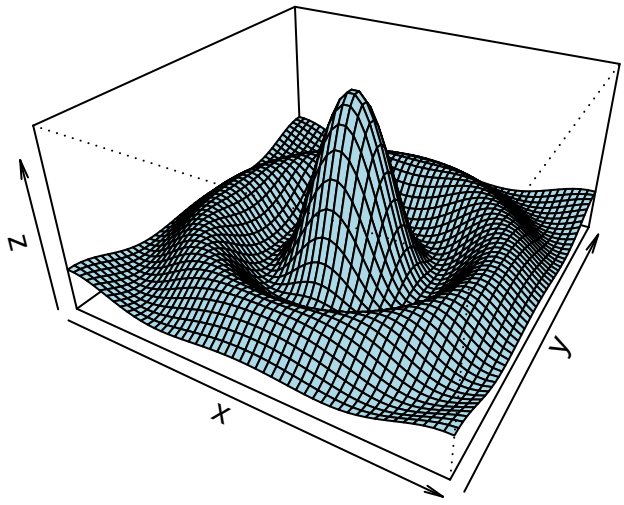
若想大致了解 R 能够作出何种图形，在命令行中运行 demo() 即可。其他的演示还有 demo(Hershey) 、demo(persp) 和 demo(image)。要看到完整的演示列表，不加参数直接运行 demo() 即可。

```
# 这里以三维图形为例
demo(persp)
```

```
##
##
##   demo(persp)
##   ---- ~~~~~
```

```
##
## > ### Demos for  persp()  plots   -- things not in  example(persp)
## > ### ------------------------
## >
## > require(datasets)
##
## > require(grDevices); require(graphics)
##
## > ## (1) The Obligatory Mathematical surface.
## > ##     Rotated sinc function.
## >
## > x <- seq(-10, 10, length.out = 50)
##
## > y <- x
##
## > rotsinc <- function(x,y)
## + {
## +     sinc <- function(x) { y <- sin(x)/x ; y[is.na(y)] <- 1; y }
## +     10 * sinc( sqrt(x^2+y^2) )
## + }
##
## > sinc.exp <- expression(z == Sinc(sqrt(x^2 + y^2)))
##
## > z <- outer(x, y, rotsinc)
##
## > oldpar <- par(bg = ”white”)
##
## > persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = ”lightblue”)
```

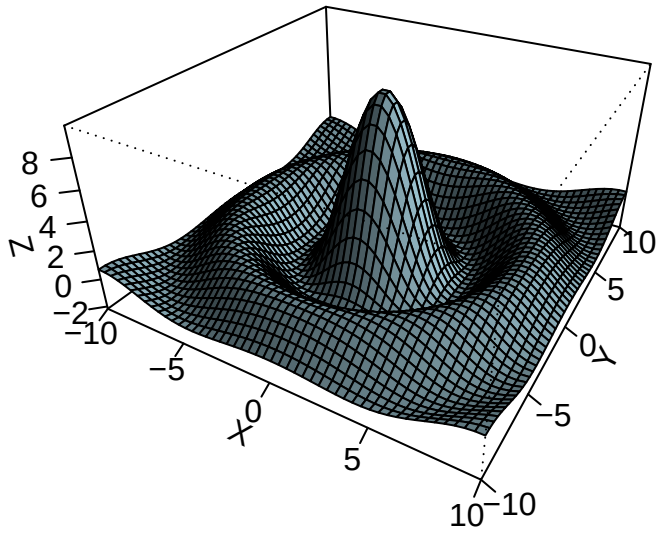
$$z = \mathrm{Sinc}(\sqrt{x^2 + y^2})$$




.

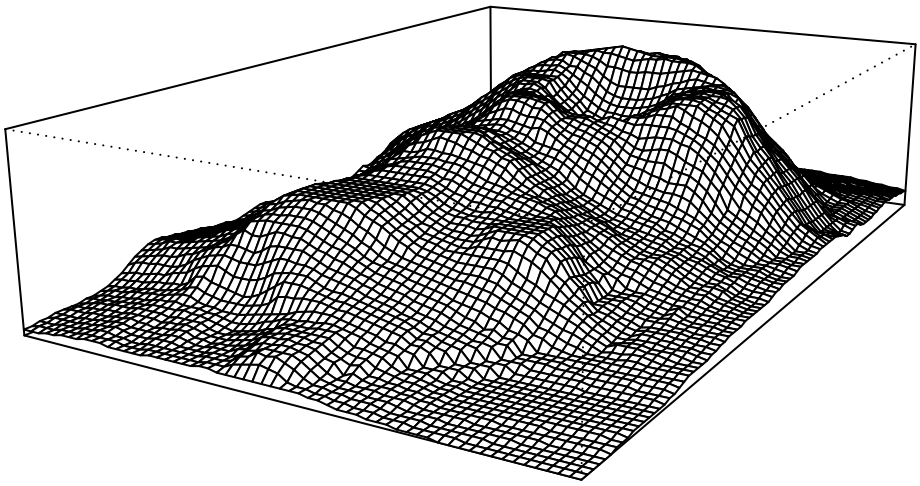
```
##
## > title(sub=".")## work around persp+plotmath bug
##
## > title(main = sinc.exp)
##
## > persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue",
## +       ltheta = 120, shade = 0.75, ticktype = "detailed",
## +       xlab = "X", ylab = "Y", zlab = "Z")
```

$$z = \mathrm{Sinc}(\sqrt{x^2 + y^2})$$

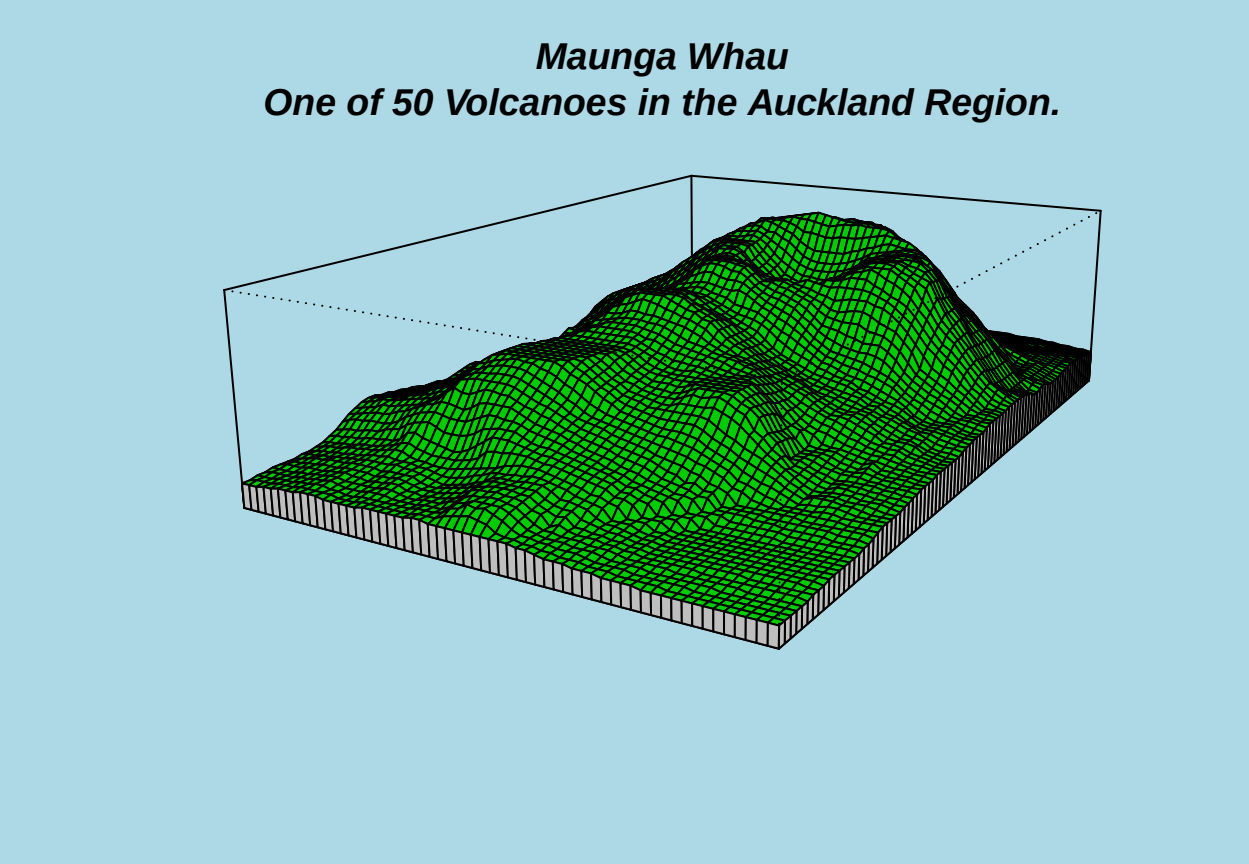



.

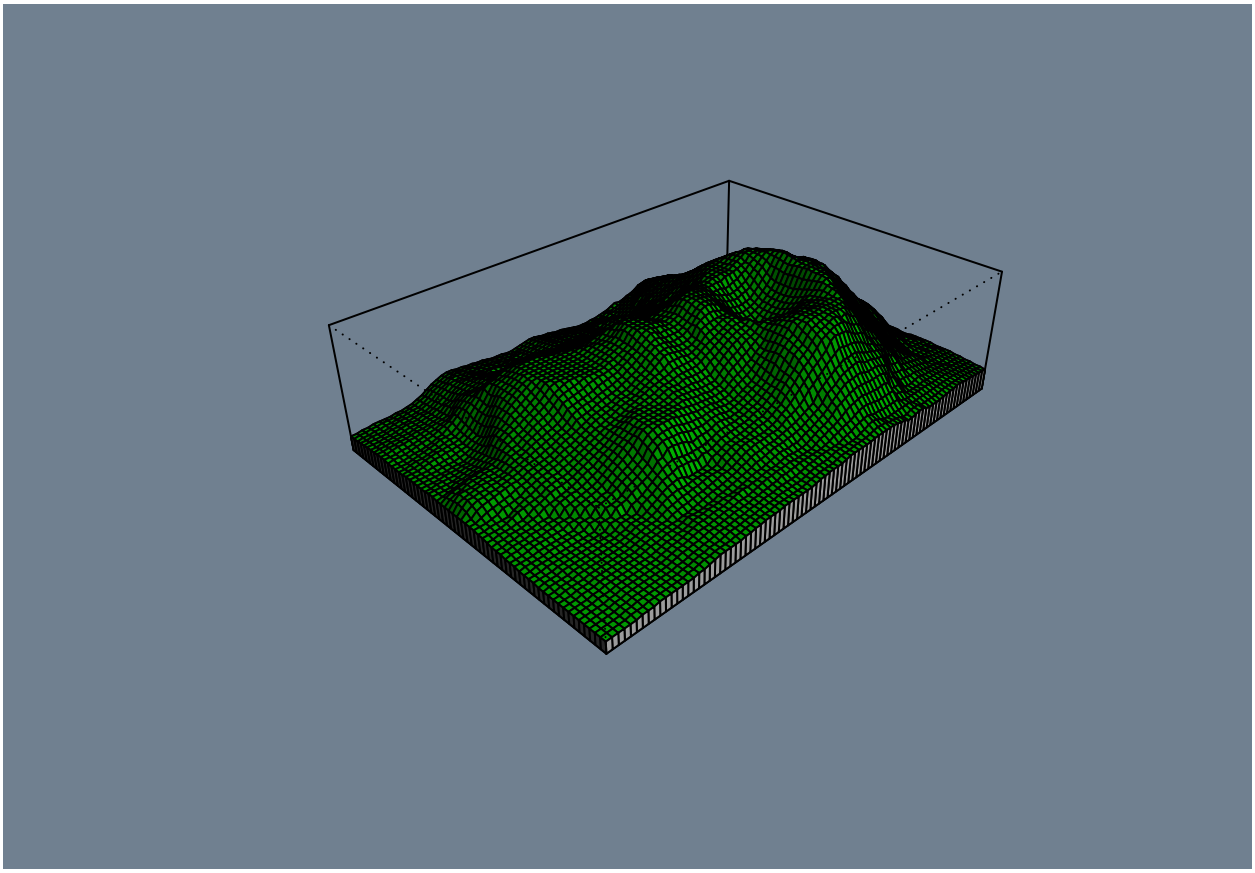
```
##
## > title(sub=".")## work around persp+plotmath bug
##
## > title(main = sinc.exp)
##
## > ## (2) Visualizing a simple DEM model
## >
## > z <- 2 * volcano        # Exaggerate the relief
##
## > x <- 10 * (1:nrow(z))   # 10 meter spacing (S to N)
##
## > y <- 10 * (1:ncol(z))   # 10 meter spacing (E to W)
##
## > persp(x, y, z, theta = 120, phi = 15, scale = FALSE, axes = FALSE)
```



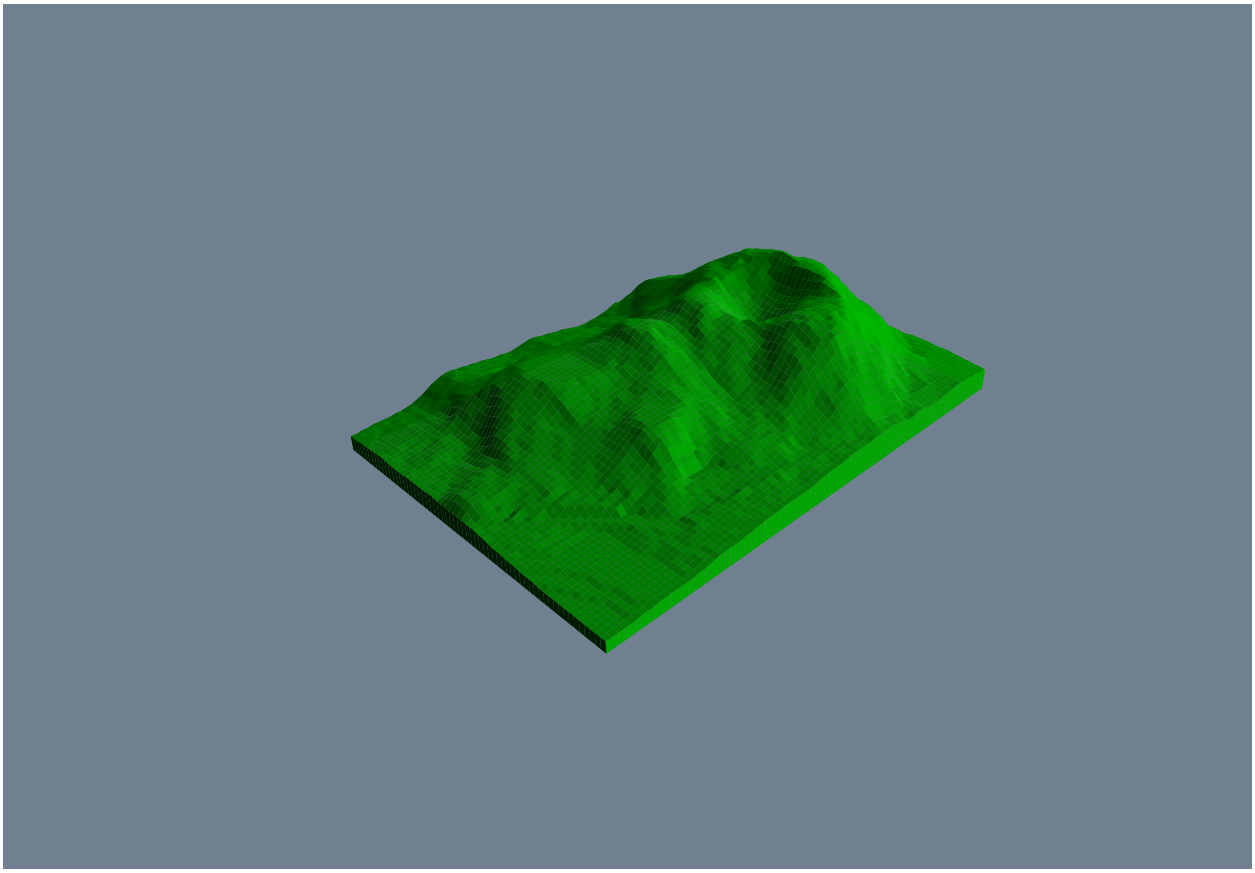
```
##
## > ## (3) Now something more complex
## > ##     We border the surface, to make it more "slice like"
## > ##     and color the top and sides of the surface differently.
## >
## > z0 <- min(z) - 20
##
## > z <- rbind(z0, cbind(z0, z, z0), z0)
##
## > x <- c(min(x) - 1e-10, x, max(x) + 1e-10)
##
## > y <- c(min(y) - 1e-10, y, max(y) + 1e-10)
##
## > fill <- matrix("green3", nrow = nrow(z)-1, ncol = ncol(z)-1)
##
## > fill[ , i2 <- c(1,ncol(fill))] <- "gray"
##
## > fill[i1 <- c(1,nrow(fill)) , ] <- "gray"
##
## > par(bg = "lightblue")
##
## > persp(x, y, z, theta = 120, phi = 15, col = fill, scale = FALSE, axes = FALSE)
```



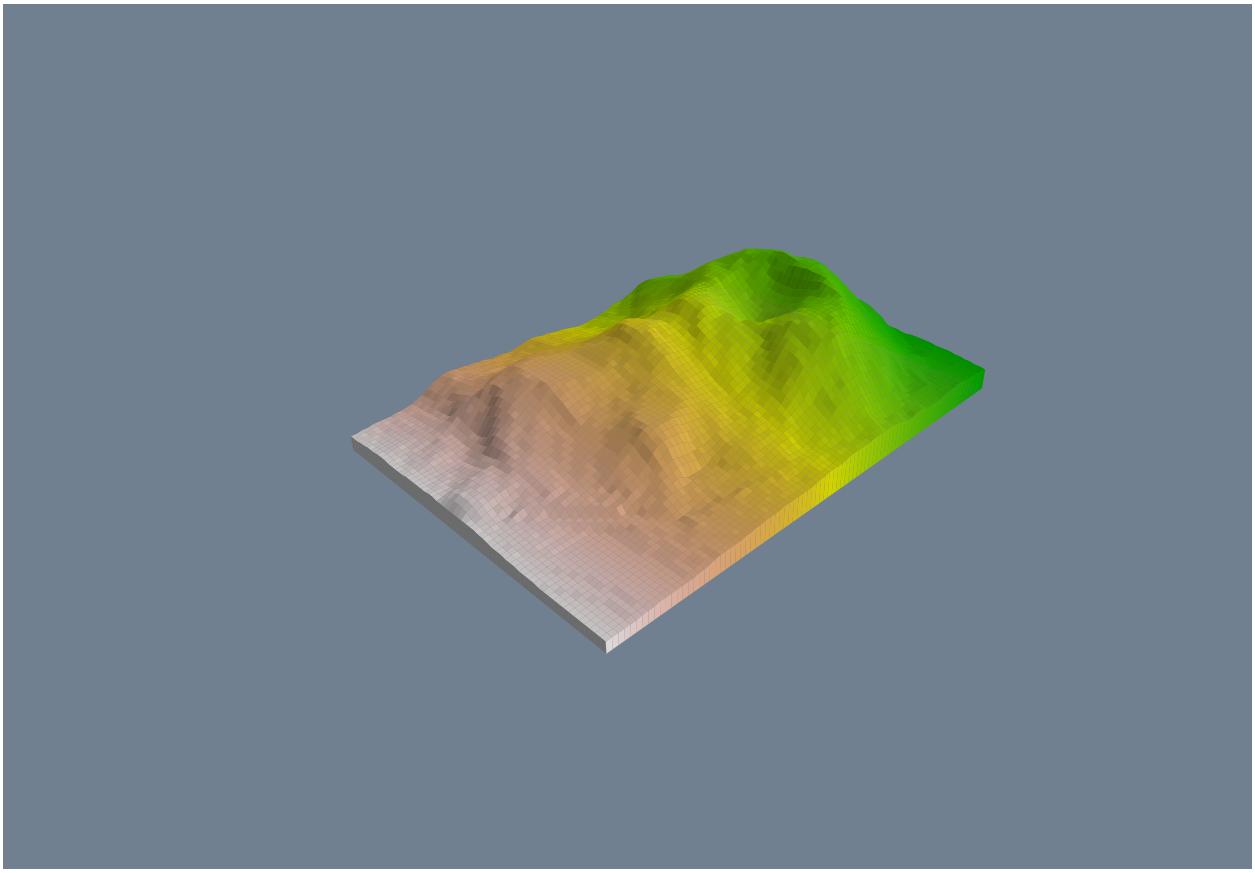

```
##
## > title(main = "Maunga Whau\nOne of 50 Volcanoes in the Auckland Region.",
## +       font.main = 4)
##
## > par(bg = "slategray")
##
## > persp(x, y, z, theta = 135, phi = 30, col = fill, scale = FALSE,
## +       ltheta = -120, lphi = 15, shade = 0.65, axes = FALSE)
```



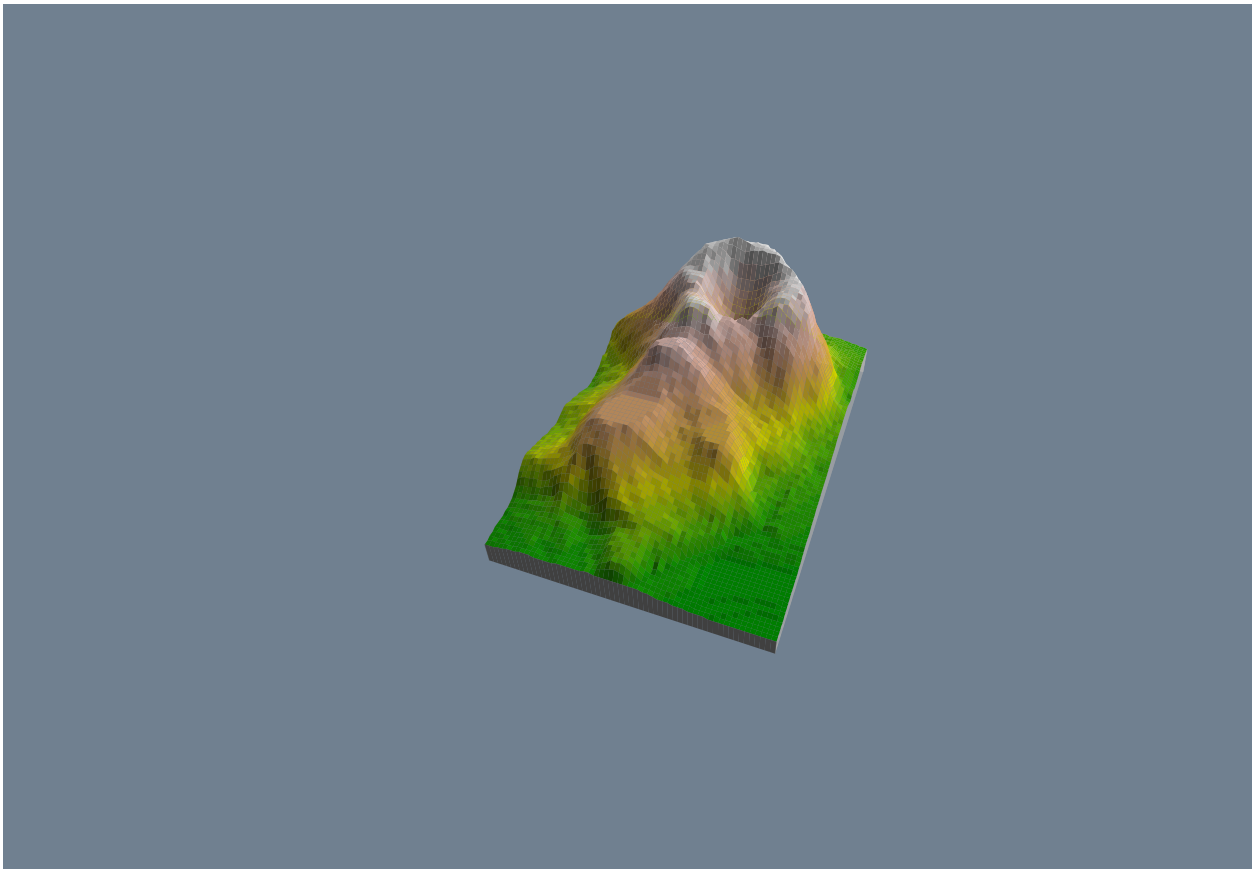
```
##
## > ## Don't draw the grid lines :  border = NA
## > persp(x, y, z, theta = 135, phi = 30, col = "green3", scale = FALSE,
## +       ltheta = -120, shade = 0.75, border = NA, box = FALSE)
```



```
##
## > ## 'color gradient in the soil' :
## > fcol <- fill ; fcol[] <- terrain.colors(nrow(fcol))
##
## > persp(x, y, z, theta = 135, phi = 30, col = fcol, scale = FALSE,
## +       ltheta = -120, shade = 0.3, border = NA, box = FALSE)
```



```
##
## > ## 'image like' colors on top :
## > fcol <- fill
##
## > zi <- volcano[ -1,-1] + volcano[ -1,-61] +
## +            volcano[-87,-1] + volcano[-87,-61]  ## / 4
##
## > fcol[-i1,-i2] <-
## +     terrain.colors(20)[cut(zi,
## +                            stats::quantile(zi, seq(0,1, length.out = 21)),
## +                            include.lowest = TRUE)]
##
## > persp(x, y, 2*z, theta = 110, phi = 40, col = fcol, scale = FALSE,
## +       ltheta = -120, shade = 0.4, border = NA, box = FALSE)
```



```
##
## > ## reset par():
## > par(oldpar)
```

**获取帮助**

R 中的帮助函数:

| 函数 | 功能 |
|---|---|
| help.start() | 打开帮助文档首页 |
| help(“foo”) 或?foo | 查看函数 foo 的帮助（引号可以省略） |
| help.search(“foo”) 或??foo | 以 foo 为关键词搜索本地帮助文档 |
| example(“foo”) | 函数 foo 的使用示例（引号可以省略） |
| RSiteSearch(“foo”) | 以 foo 为关键词搜索在线文档和邮件列表存档 |
| apropos(“foo”, mode=“function”) | 列出名称中含有 foo 的所有可用函数 |
| data() | 列出当前已加载包中所含的所有可用示例数据集 |
| vignette() | 列出当前已安装包中所有可用的 vignette 文档 |
| vignette(“foo”) | 为主题 foo 显示指定的 vignette 文档 |

函数 help.start() 会打开一个浏览器窗口，我们可在其中查看入门和高级的帮助手册、常见问题集，以及参考材料。

```
#R 语言的帮助
#? seq
#help(”seq”) # 同上
#help.search(”mean”) # 可以使用类似于百度搜索的关键字
```

```
example("seq")
```

```
##
## seq> seq(0, 1, length.out = 11)
##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
##
## seq> seq(stats::rnorm(20)) # effectively 'along'
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##
## seq> seq(1, 9, by = 2)     # matches 'end'
## [1] 1 3 5 7 9
##
## seq> seq(1, 9, by = pi)    # stays below 'end'
## [1] 1.000000 4.141593 7.283185
##
## seq> seq(1, 6, by = 3)
## [1] 1 4
##
## seq> seq(1.575, 5.125, by = 0.05)
##  [1] 1.575 1.625 1.675 1.725 1.775 1.825 1.875 1.925 1.975 2.025 2.075
## [12] 2.125 2.175 2.225 2.275 2.325 2.375 2.425 2.475 2.525 2.575 2.625
## [23] 2.675 2.725 2.775 2.825 2.875 2.925 2.975 3.025 3.075 3.125 3.175
## [34] 3.225 3.275 3.325 3.375 3.425 3.475 3.525 3.575 3.625 3.675 3.725
## [45] 3.775 3.825 3.875 3.925 3.975 4.025 4.075 4.125 4.175 4.225 4.275
## [56] 4.325 4.375 4.425 4.475 4.525 4.575 4.625 4.675 4.725 4.775 4.825
## [67] 4.875 4.925 4.975 5.025 5.075 5.125
##
## seq> seq(17) # same as 1:17, or even better seq_len(17)
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
```

**工作空间**

工作空间（workspace）就是当前 R 的工作环境，它存储着所有用户定义的对象（向量、矩阵、函数、数据框、列表）。在一个 R 会话结束时，你可以将当前工作空间保存到一个镜像中，并在下次启动 R 时自动载入它。各种命令可在 R 命令行中交互式地输入。使用上下方向键查看已输入命令的历史记录。这样我们就可以选择一个之前输入过的命令并适当修改，最后按回车重新执行它。当前的工作目录（working directory）是 R 用来读取文件和保存结果的默认目录。我们可以使用函数 getwd() 来查看当前的工作目录，或使用函数 setwd() 设定当前的工作目录。如果需要读入一个不在当前工作目录下的文件，则需在调用语句中写明完整的路径。记得使用引号闭合这些目录名和文件名。

```
getwd() # 获取当前工作目录
```

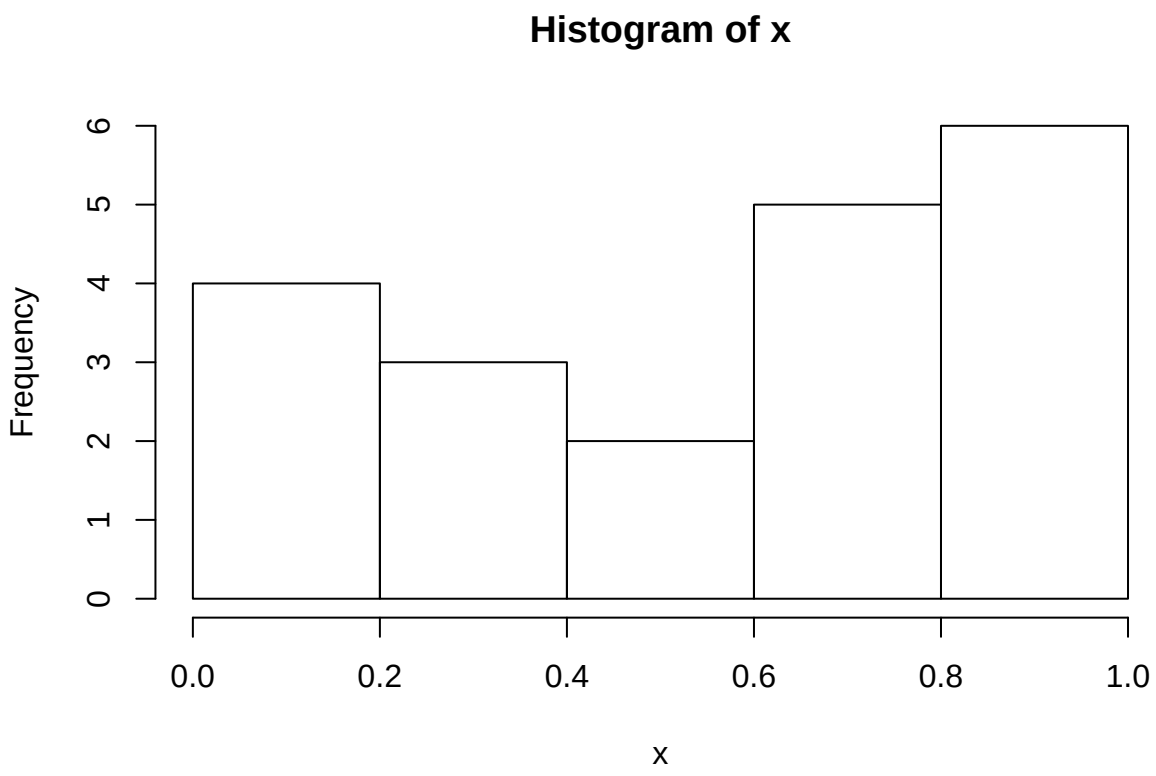
```
## [1] "E:/课程/数据分析与应用/DataAnalysis"
```

```
#setwd("F:/code/R") # 改变当前工作目录

x <- runif(20)
summary(x)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.004504 0.261100 0.649000 0.553400 0.818600 0.991000
```

```
hist(x)
```




```
#savehistory()
save.image()
```

查看当前工作空间中有哪些变量：

```
ls()
```

```
##  [1] "age"      "fcol"     "fill"     "i1"       "i2"       "oldpar"
##  [7] "rotsinc"  "sinc.exp" "weight"   "x"        "y"        "z"
## [13] "z0"       "zi"
```

**输入和输出**

- 输入：source("sample.R")
- 输出
  - 文本输出：sink("filename") # 输出重定向
  - 图形输出：jpeg("filename")、bmp("filename")、pdf("filename")

**包**

- 什么是包
- 包的安装：install.packages("包的名字")
- 包的载入：library(包的名字)
- 包的使用：help(package="包的名字")

看看目前工作空间中加载了哪些包：

```
search()
```

```
## [1] ".GlobalEnv"        "package:stats"     "package:graphics"
## [4] "package:grDevices" "package:utils"     "package:datasets"
## [7] "package:methods"   "Autoloads"         "package:base"
```

**示例实践**

我们将以一个结合了以上各种命令的示例结束本章。以下是任务描述。(1) 打开帮助文档首页，并查阅其中的 “Introduction to R”。(2) 安装 vcd 包（一个用于可视化类别数据的包，你将在第 11 章中使用）。(3) 列出此包中可用的函数和数据集。(4) 载入这个包并阅读数据集 Arthritis 的描述。(5) 显示数据集 Arthritis 的内容（直接输入一个对象的名称将列出它的内容）。(6) 运行数据集 Arthritis 自带的示例。如果不理解输出结果，也不要担心。它基本上显示了接受治疗的关节炎患者较接受安慰剂的患者在病情上有了更多改善。(7) 退出。

```
#help.start()
#install.packages("vcd")
help(package="vcd")
library(vcd)
```

```
## Loading required package: grid
```

```
help(Arthritis)
```

```
## starting httpd help server ...
```

```
##  done
```

```
knitr::kable(Arthritis)
```

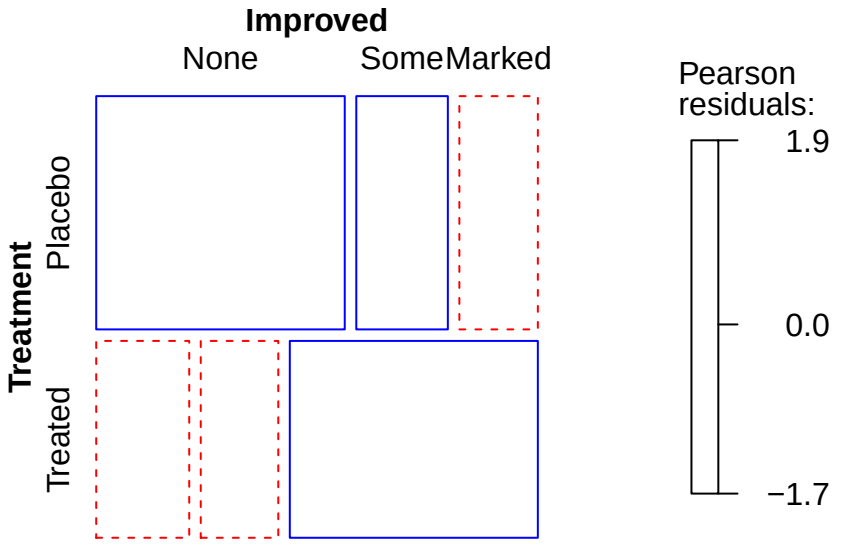
| ID | Treatment | Sex | Age | Improved |
|---|---|---|---|---|
| 57 | Treated | Male | 27 | Some |
| 46 | Treated | Male | 29 | None |
| 77 | Treated | Male | 30 | None |
| 17 | Treated | Male | 32 | Marked |
| 36 | Treated | Male | 46 | Marked |
| 23 | Treated | Male | 58 | Marked |
| 75 | Treated | Male | 59 | None |
| 39 | Treated | Male | 59 | Marked |
| 33 | Treated | Male | 63 | None |
| 55 | Treated | Male | 63 | None |
| 30 | Treated | Male | 64 | None |
| 5 | Treated | Male | 64 | Some |
| 63 | Treated | Male | 69 | None |
| 83 | Treated | Male | 70 | Marked |
| 66 | Treated | Female | 23 | None |
| 40 | Treated | Female | 32 | None |
| 6 | Treated | Female | 37 | Some |
| 7 | Treated | Female | 41 | None |
| 72 | Treated | Female | 41 | Marked |
| 37 | Treated | Female | 48 | None |
| 82 | Treated | Female | 48 | Marked |
| 53 | Treated | Female | 55 | Marked |
| 79 | Treated | Female | 55 | Marked |
| 26 | Treated | Female | 56 | Marked |
| 28 | Treated | Female | 57 | Marked |

| ID | Treatment | Sex | Age | Improved |
|---|---|---|---|---|
| 60 | Treated | Female | 57 | Marked |
| 22 | Treated | Female | 57 | Marked |
| 27 | Treated | Female | 58 | None |
| 2 | Treated | Female | 59 | Marked |
| 59 | Treated | Female | 59 | Marked |
| 62 | Treated | Female | 60 | Marked |
| 84 | Treated | Female | 61 | Marked |
| 64 | Treated | Female | 62 | Some |
| 34 | Treated | Female | 62 | Marked |
| 58 | Treated | Female | 66 | Marked |
| 13 | Treated | Female | 67 | Marked |
| 61 | Treated | Female | 68 | Some |
| 65 | Treated | Female | 68 | Marked |
| 11 | Treated | Female | 69 | None |
| 56 | Treated | Female | 69 | Some |
| 43 | Treated | Female | 70 | Some |
| 9 | Placebo | Male | 37 | None |
| 14 | Placebo | Male | 44 | None |
| 73 | Placebo | Male | 50 | None |
| 74 | Placebo | Male | 51 | None |
| 25 | Placebo | Male | 52 | None |
| 18 | Placebo | Male | 53 | None |
| 21 | Placebo | Male | 59 | None |
| 52 | Placebo | Male | 59 | None |
| 45 | Placebo | Male | 62 | None |
| 41 | Placebo | Male | 62 | None |
| 8 | Placebo | Male | 63 | Marked |
| 80 | Placebo | Female | 23 | None |
| 12 | Placebo | Female | 30 | None |
| 29 | Placebo | Female | 30 | None |
| 50 | Placebo | Female | 31 | Some |
| 38 | Placebo | Female | 32 | None |
| 35 | Placebo | Female | 33 | Marked |
| 51 | Placebo | Female | 37 | None |
| 54 | Placebo | Female | 44 | None |
| 76 | Placebo | Female | 45 | None |
| 16 | Placebo | Female | 46 | None |
| 69 | Placebo | Female | 48 | None |
| 31 | Placebo | Female | 49 | None |
| 20 | Placebo | Female | 51 | None |
| 68 | Placebo | Female | 53 | None |
| 81 | Placebo | Female | 54 | None |
| 4 | Placebo | Female | 54 | None |
| 78 | Placebo | Female | 54 | Marked |
| 70 | Placebo | Female | 55 | Marked |
| 49 | Placebo | Female | 57 | None |
| 10 | Placebo | Female | 57 | Some |
| 47 | Placebo | Female | 58 | Some |
| 44 | Placebo | Female | 59 | Some |
| 24 | Placebo | Female | 59 | Marked |
| 48 | Placebo | Female | 61 | None |
| 19 | Placebo | Female | 63 | Some |

| ID | Treatment | Sex | Age | Improved |
|---|---|---|---|---|
| 3 | Placebo | Female | 64 | None |
| 67 | Placebo | Female | 65 | Marked |
| 32 | Placebo | Female | 66 | None |
| 42 | Placebo | Female | 66 | None |
| 15 | Placebo | Female | 66 | Some |
| 71 | Placebo | Female | 68 | Some |
| 1 | Placebo | Female | 74 | Marked |

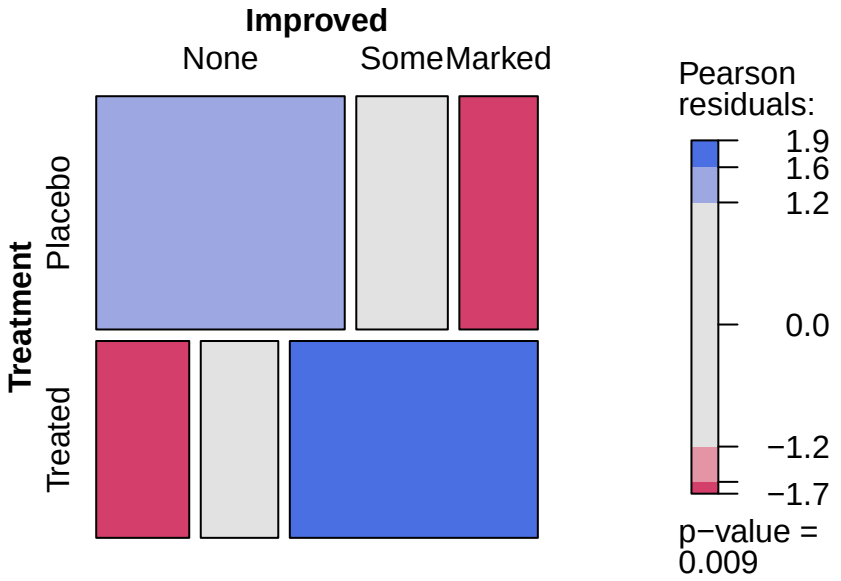
```
example(Arthritis)
```

```
##
## Arthrt> data("Arthritis")
##
## Arthrt> art <- xtabs(~ Treatment + Improved, data = Arthritis, subset = Sex == "Female
##
## Arthrt> art
##           Improved
## Treatment None Some Marked
##   Placebo   19    7      6
##   Treated    6    5     16
##
## Arthrt> mosaic(art, gp = shading_Friendly)
```




```
##
```

```
## Arthrt> mosaic(art, gp = shading_max)
```




```
# q()
```