数据分析与决策技术丛书

HZ BOOKS
华章IT

[PACKT]
PUBLISHING

R Data Mining Blueprints

R语言数据挖掘

实用项目解析

[印度] 普拉迪帕塔·米什拉（Pradeepta Mishra） 著

黄芸 译

结合现实中广泛应用的数据案例讲解R语言数据挖掘技术

配有大量代码和图片

机械工业出版社
China Machine Press

数据分析与决策
技术丛书

R Data Mining Blueprints

R语言数据挖掘
实用项目解析

[印度] 普拉迪帕塔·米什拉（Pradeepta Mishra） 著
黄芸 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

R 语言数据挖掘：实用项目解析 /（印）普拉迪帕塔 · 米什拉（Pradeepta Mishra）著；黄芸译 . —北京：机械工业出版社，2017.3
（数据分析与决策技术丛书）
书名原文：R Data Mining Blueprints

ISBN 978-7-111-56520-8

I. R… II. ①普… ②黄… III. ①程序语言 – 程序设计 ②数据采集 IV. ① TP312 ② TP274

中国版本图书馆 CIP 数据核字（2017）第 070089 号

R 语言数据挖掘：实用项目解析

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：吴晋瑜 责任校对：殷 虹
印 刷：北京市荣盛彩色印刷有限公司 版 次：2017 年 5 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：12.5
书 号：ISBN 978-7-111-56520-8 定 价：49.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

The Translator's Words 译 者 序

在这个信息爆炸的时代中，无论是个人还是企业，都是数据的产生者，同时也是数据价值的受益者。对于已经积累了大量数据的企业来说，通过数据挖掘来提升投资回报率（Return on Investment，ROI）或商业价值已成为刻不容缓的目标。

R 语言凭借其健康的开源工具生态及简单易上手的语言特性，广泛应用于统计领域，并获得了数据分析爱好者们的青睐。R 语言的主要用户群或许未曾想到，也正如数据挖掘人士未曾想到的是，用作统计分析工具的 R 语言也可以成为数据挖掘的利器。R 语言的语言特性使其不仅适合数据分析人员使用，也适合所有试图从数据中获取个人在意的信息或者企业关注的业务价值的各行业人员使用。

本书是一本介绍使用 R 语言进行数据挖掘的指南书。既然是指南书，也就不要求读者有多么深厚的统计基础以及丰富的编程经验。本书将对所涉及的理论知识进行简单的介绍，清晰地列出相关公式与使用技术时的注意要点，还配有大量代码和图片，以帮助读者通过实践加深对概念的理解。为了给读者营造出一种清晰的数据挖掘项目流程感，本书按照“数据处理——数据探索——建立应用模型”这样的顺序组织编写，以求做到简洁而不失细节。此外，本书对数据处理中的棘手问题（譬如时间格式、缺失值的处理）均做出了详细指导，且由于数据探索在项目中的重要性，亦从统计角度到可视化角度给出了讲解。针对应用模型的建立，本书选取了现实中常见的模型进行介绍，由简单的回归模型开始，到应用广泛的购物篮分析、推荐系统构建，再到较复杂的神经网络模型。

本书的一大特色是结合了现实中广泛应用的数据案例，如零售业、制造业、信用评分、医疗业等的数据案例。通过本书的学习，读者不仅能够掌握一定的技术实战能力，也能从中得到一些有关业务应用的启发，最终学以致用。

黄 芸

前　言 Preface

随着数据规模和种类的增长，应用数据挖掘技术从大数据中提取有效信息变得至关重要。这是因为企业认为有必要从大规模数据的实施中获得相应的投资回报。实施数据挖掘的根本性原因是要从大型数据库中发现隐藏的商机，以便利益相关者能针对未来业务做出决策。数据挖掘不仅能够帮助企业降低成本以及提高收益，还能帮助他们发现新的发展途径。

本书将介绍使用R语言（一种开源工具）进行数据挖掘的基本原理。R是一门免费的程序语言，同时也是一个提供统计计算、图形数据可视化和预测建模的软件环境，并且可以与其他工具和平台相集成。本书将结合R语言在示例数据集中的应用来阐释数据挖掘原理。

本书将阐述数据挖掘的一些主题，如数学表述、在软件环境中的实现，以及如何据此来解决商业问题。本书的设计理念是，读者可以从数据管理技术、探索性数据分析、数据可视化等内容着手学习，循序渐进，直至建立高级预测模型（如推荐系统、神经网络模型）。本书也从数据科学、分析学、统计建模以及可视化等角度对数据挖掘这一概念进行了综述。

本书内容

第1章　带领读者初识R编程基础，借助真实的案例帮助读者了解如何读写数据，了解编程符号和语法指令。这一章还给出了供读者动手实践的R脚本，以更好地理解书中的原理、术语以及执行特定任务的深层原因。之所以这样设计，是为了让没有太多编程基础的读者也能使用R来执行各种数据挖掘任务。这一章将简述数据挖掘的意义以及

它与其他领域（诸如数据科学、分析学和统计建模）的关系，除此之外，还将展开使用R进行数据管理的讨论。

第2章　帮助读者理解探索性数据分析。探索数据包括数据集中变量的数值描述和可视化，这将使得数据集变得直观，并使我们能对其快速定论。对数据集有一个初步的理解很重要，比如选择怎样的变量进行分析、不同变量之间的关联，等等。创建交叉二维表有助于理解分类变量之间的关系，对数据集实施经典统计检验来验证对数据的种种假设。

第3章　涵盖从基础的数据可视化到调用R语言中的库实现高级的数据可视化。观察数字和统计能从多个侧面“告诉”我们关于变量的“故事”，而当图形化地了解变量和因子之间的关系时，它将展示另一个“故事”。可见，数据可视化将揭示数值分析和统计无法展现的信息。

第4章　帮助读者学习利用回归方法的预测分析基础，包括线性和非线性回归方法在R中的实现。读者不仅可以掌握所有回归方法的理论基础，也将通过R实践获得实际动手操作的经验。

第5章　介绍了一种产品推荐方法——购物篮分析（MBA）。这种方法主要是将交易级的商品信息关联，从中找出购买了相似商品的客户分类，据此推荐产品。MBA还可以应用于向上销售和交叉销售中。

第6章　介绍了什么是分类、聚类是如何应用到分类问题的、聚类用的是什么方法等内容，并对不同的分类方法进行了对比。在这一章，读者将了解使用聚类方法的分类基础知识。

第7章　涵盖以下内容及相应的R语言实现：推荐系统是什么，实现推荐的工作原理、类型和方法，使用R语言实现商品推荐。

第8章　使用R语言和一个实际数据集实现主成分分析（PCA）、奇异值分解（SVD）和迭代特征提取等降维技术。随着数据的量与类的增长，数据的维度也在随之增长。降维技术在不同领域都有很多应用，例如图像处理、语音识别、推荐系统、文本处理等。

第9章　讲解了多种类型的神经网络、方法，以及通过不同的函数来控制人工神经网络训练的神经网络变体。这些神经网络执行标准的数据挖掘任务，例如：采用基于回归的方法预测连续型变量，利用基于分类的方法预测输出水平，利用历史数据来预测数值变量的未来值，以及压缩特征从而识别重要特征以执行预测或分类。

准备工作

为了学习本书附带的例子和代码，读者需要从 https://cran.r-project.org/ 下载 R 软件（也可以从 https://www.rstudio.com/ 下载 R Studio），然后安装。没有特定的硬件要求，只需要一台至少 2GB RAM 的计算机，适用于任何操作系统，包括 MAC、Linux 和 Windows。

读者对象

本书适用于刚开始从事数据挖掘、数据科学或者预测建模的读者，也适用于有中等统计与编程水平的读者。基本的统计知识对于理解数据挖掘是必需的。阅读前几章并不需要编程知识。本书将讲解如何使用 R 语言进行数据管理和基本的统计分析。本书亦适用于学生、专业人员及有志成为数据分析师的读者。

排版约定

在本书中，为了区分不同内容，字体风格也会随之变化。以下是字体风格示意：

书中的代码、文件名、文件扩展名、路径名、URL 地址、用户输入、推特标签看起来会是这样：“在处理 ArtPiece 数据集时，我们将通过一些与业务相关的变量来预测一个艺术作品是否值得购买。”

所有命令行的输入或输出在书中显示如下：

```
>fit<- neuralnet(formula = CurrentAuctionAveragePrice ~ Critic.Ratings +
Acq.Cost + CollectorsAverageprice + Min.Guarantee.Cost, data = train,
hidden = 15, err.fct = "sse", linear.output = F)
> fit
Call: neuralnet(formula = CurrentAuctionAveragePrice ~ Critic.Ratings +
Acq.Cost + CollectorsAverageprice + Min.Guarantee.Cost, data = train,
hidden = 15, err.fct = "sse", linear.output = F)
1 repetition was calculated.
Error Reached Threshold Steps
1 54179625353167 0.004727494957 23
```

作者的话

如果读者对于本书所涉及的内容有疑问，可以在 Twitter 上搜索 @mishra1_PK，我非常乐意为大家提供帮助。

非常感谢我的妻子 Prajna 和女儿 Aarya，也要感谢我的朋友和工作中的同事在我完成本书的过程中给予我的支持与鼓励。

关于审稿人

Alexey Grigorev 是一名熟练的数据科学家和软件工程师，有超过 5 年的专业经验。他现在是 Searchmetrics Inc 的一名数据科学家。在日常工作中，他热衷于使用 R 和 Python 进行数据清洗、数据分析和建模工作。他也是 Packt 出版的其他数据分析书籍的审稿人，比如《测试驱动的机器学习》与《掌握 R 数据分析》。

目　录 Contents

译者序

前言

第 1 章　使用 R 内置数据进行数据处理 ····· 1

1.1　什么是数据挖掘 ····· 2

1.2　R 语言引论 ····· 4

1.2.1　快速入门 ····· 4

1.2.2　数据类型、向量、数组与矩阵 ····· 4

1.2.3　列表管理、因子与序列 ····· 7

1.2.4　数据的导入与导出 ····· 8

1.3　数据类型转换 ····· 10

1.4　排序与合并数据框 ····· 11

1.5　索引或切分数据框 ····· 15

1.6　日期与时间格式化 ····· 16

1.7　创建新函数 ····· 17

1.7.1　用户自定义函数 ····· 17

1.7.2　内置函数 ····· 18

1.8　循环原理——for 循环 ····· 18

1.9　循环原理——repeat 循环 ····· 19

1.10　循环原理——while 循环 ····· 19

1.11　apply 原理 ····· 19

1.12　字符串操作 ····· 21

1.13　缺失值（NA）的处理 ····· 22

小结 ····· 23

第 2 章　汽车数据的探索性分析 ····· 24

2.1　一元分析 ····· 24

2.2　二元分析 ····· 30

2.3　多元分析 ····· 31

2.4　解读分布和变换 ····· 32

2.4.1　正态分布 ····· 32

2.4.2　二项分布 ····· 34

2.4.3　泊松分布 ····· 34

2.5　解读分布 ····· 34

2.6　变量分段 ····· 37

2.7　列联表、二元统计及数据正态性检验 ····· 37

2.8　假设检验 ····· 41

2.8.1　总体均值检验 ····· 42

2.8.2　双样本方差检验 ····· 46

2.9 无参数方法 …… 48
2.9.1 Wilcoxon 符号秩检验 …… 49
2.9.2 Mann-Whitney-Wilcoxon 检验 …… 49
2.9.3 Kruskal-Wallis 检验 …… 49
小结 …… 50

第 3 章 可视化 diamond 数据集 …… 51

3.1 使用 ggplot2 可视化数据 …… 54
3.1.1 条状图 …… 64
3.1.2 盒状图 …… 65
3.1.3 气泡图 …… 65
3.1.4 甜甜圈图 …… 66
3.1.5 地理制图 …… 67
3.1.6 直方图 …… 68
3.1.7 折线图 …… 68
3.1.8 饼图 …… 69
3.1.9 散点图 …… 70
3.1.10 堆叠柱形图 …… 75
3.1.11 茎叶图 …… 75
3.1.12 词云 …… 76
3.1.13 锯齿图 …… 76
3.2 使用 plotly …… 78
3.2.1 气泡图 …… 78
3.2.2 用 plotly 画条状图 …… 79
3.2.3 用 plotly 画散点图 …… 79
3.2.4 用 plotly 画盒状图 …… 80
3.2.5 用 plotly 画极坐标图 …… 82
3.2.6 用 plotly 画极坐标散点图 …… 82
3.2.7 极坐标分区图 …… 83
3.3 创建地理制图 …… 84
小结 …… 84

第 4 章 用汽车数据做回归 …… 85

4.1 回归引论 …… 85
4.1.1 建立回归问题 …… 86
4.1.2 案例学习 …… 87
4.2 线性回归 …… 87
4.3 通过逐步回归法进行变量选取 …… 98
4.4 Logistic 回归 …… 99
4.5 三次回归 …… 105
4.6 惩罚回归 …… 106
小结 …… 109

第 5 章 基于产品数据的购物篮分析 …… 110

5.1 购物篮分析引论 …… 110
5.1.1 什么是购物篮分析 …… 111
5.1.2 哪里会用到购物篮分析 …… 112
5.1.3 数据要求 …… 112
5.1.4 前提假设 / 要求 …… 114
5.1.5 建模方法 …… 114
5.1.6 局限性 …… 114
5.2 实际项目 …… 115
5.2.1 先验算法 …… 118
5.2.2 eclat 算法 …… 121
5.2.3 可视化关联规则 …… 123
5.2.4 实施关联规则 …… 124
小结 …… 126

第 6 章　聚类电商数据 …… 127

6.1　理解客户分类 …… 128

6.1.1　为何理解客户分类很重要 … 128

6.1.2　如何对客户进行分类 …… 128

6.2　各种适用的聚类方法 …… 129

6.2.1　*K* 均值聚类 …… 130

6.2.2　层次聚类 …… 135

6.2.3　基于模型的聚类 …… 139

6.2.4　其他聚类算法 …… 140

6.2.5　聚类方法的比较 …… 143

参考文献 …… 143

小结 …… 143

第 7 章　构建零售推荐引擎 …… 144

7.1　什么是推荐 …… 144

7.1.1　商品推荐类型 …… 145

7.1.2　实现推荐问题的方法 …… 145

7.2　前提假设 …… 147

7.3　什么时候采用什么方法 …… 148

7.4　协同过滤的局限 …… 149

7.5　实际项目 …… 149

小结 …… 157

第 8 章　降维 …… 158

8.1　为什么降维 …… 158

8.2　降维实际项目 …… 161

8.3　有参数法降维 …… 172

参考文献 …… 173

小结 …… 173

第 9 章　神经网络在医疗数据中的应用 …… 174

9.1　神经网络引论 …… 174

9.2　理解神经网络背后的数学原理 …… 176

9.3　用 R 语言实现神经网络 …… 177

9.4　应用神经网络进行预测 …… 180

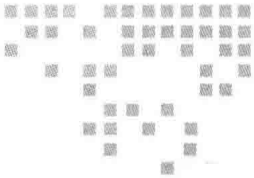
9.5　应用神经网络进行分类 …… 183

9.6　应用神经网络进行预测 …… 185

9.7　神经网络的优缺点 …… 187

参考文献 …… 187

小结 …… 187



第 1 章 Chapter 1

使用 R 内置数据进行数据处理

本书主要介绍在 R 语言平台上实现数据挖掘的方法和步骤。因为 R 是一种开源工具，所以对各层次的学习者而言，学习使用 R 语言进行数据挖掘都会很有意思。本书的设计宗旨是，读者可以从数据管理技术着手，从探索性数据分析、数据可视化和建模开始，直至建立高级预测模型，如推荐系统、神经网络模型等。本章将概述数据挖掘的原理及其与数据科学、分析学和统计建模的交叉。在本章，读者将初识 R 编程语言基础，并通过一个真实的案例，了解怎样读取和写入数据，熟悉编程符号和理解句法。本章还包含了 R 语言脚本，可供读者动手实践，以加深对原理和术语的理解，领会数据挖掘任务的来龙去脉。本章之所以这样设计，是为了让那些编程基础薄弱的读者也可以通过执行 R 语言命令来完成一些数据挖掘任务。

本章将简述数据挖掘的意义以及它与其他领域（如数据科学、分析学和统计建模）的关系，还会就使用 R 进行数据管理的话题展开讨论。通过学习本章的内容，读者应掌握以下知识点：

- 了解 R 语言中所使用的各种数据类型，包括向量和向量运算。
- 数据框的索引及因子序列。
- 数据框的排序与合并以及数据类型的转换。
- 字符串操作以及数据对象格式化。

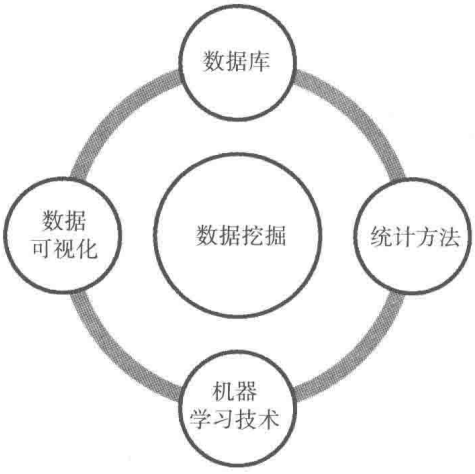
- 缺失值（NA）的处理方法。
- 流控制、循环构建以及 apply 函数的应用。

1.1 什么是数据挖掘

数据挖掘可以定义为这样的过程：从现有数据库中“解读”出有意义的信息，然后加以分析，并将结果提供给业务人员。从不同数据源分析数据，进而归纳出有意义的信息和洞见——这属于统计知识的探索，不仅有助于业务人员，也有助于多个群体，如统计分析员、咨询师和数据科学家。通常，数据库中的知识探索过程是不可预知的，对探索结果也可以从多个角度进行解读。

硬件设备、平板、智能手机、计算机、传感器等电子设备的大规模增长促使数据以超前的速度产生与收集。随着现代计算机处理能力的提升，可以对增长的数据进行预处理和模型化，以解决与商业决策过程相关的各种问题。数据挖掘也可以定义为利用统计方法、机器学习技术、可视化和模式匹配技术从离散的数据库和信息资源库中进行知识密集型搜索。

零售商店内所有物品的条形码、制造业所有货物的射频识别标签、推特简讯、Facebook 上的贴子、遍布城市用于监控天气变化的传感器、录像分析、基于观看信息统计的视频推荐……这些结构化和非结构化数据的增长创造了一个催生各种各样的工具、技术和方法的生态系统。前文提到应用于各种数据的数据挖掘技术，不仅提供了有用的数据结构信息，也就企业未来可采取的决策提出了建议。



数据挖掘包括以下几个步骤：

1）从数据库和数据仓库中抽取需要的数据。

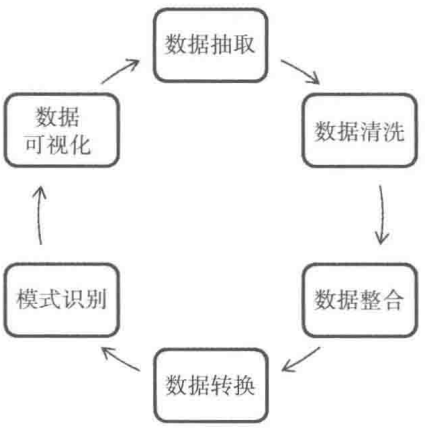
2）检查数据，删除冗余特征和无关信息。

3）有时需要与其他未关联数据库中的数据相合并。所以，需要找到各个数据库的共同属性。

4）应用数据转换技术。有时，一些属性和特征需要包含在一个模型中。

5）对输入的特征值进行模式识别。这里可能会用到任何模式识别技术。

6）知识表达。其中包括把从数据库中提炼出来的知识通过可视化方式展示给利益相关者。



在讨论了数据挖掘的流程和核心组成之后，我们也需注意到实施数据挖掘时可能遇到的挑战，比如运算效率、数据库的非结构化以及怎样将其与结构化数据结合、高维数据的可视化问题，等等。这些问题可以通过创新的方法来解决。本书在项目实践中会涉及一些解决方法。

它是怎么与数据科学、分析和统计建模关联的

数据科学是个很宽泛的话题，其中也包含了一些数据挖掘的概念。根据之前对数据挖掘的定义，即它是从数据中发现隐藏模式，找出有意思的关联并能提供有用的决策支持的过程，可知数据挖掘是数据科学项目的子集，涉及模式识别、特征提取、聚类以及监督分类等技术。分析学和统计建模包含了很多预测模型——基于分类的模型，通过应用这些方法解决实际业务问题。数据科学、分析学和统计建模、数据挖掘这些术语之间明显是有重叠的，所以不应该把它们看作完全独立的术语。根据项目要求和特定的业务问题，它们重叠的部分可

能有所不同。但总的来说，所有概念都是相关联的。数据挖掘过程也包括基于统计和机器学习方法来提取数据，提取自动化规则，也需要利用好的可视化方法来展示数据。

1.2 R 语言引论

本节将开始使用基础的 R 编程知识来做数据管理和数据处理，其中也会讲到一些编程技巧。R 可以从 https://www.r-project.org/ 下载。用户可以基于自己的操作系统下载和安装 R 二进制文件。R 编程语言作为 S 语言的扩展，是一个统计计算平台。它提供高级预测建模、机器学习算法实施和更好的图表可视化。R 还提供了适用于其他平台的插件，比如 R.Net、rJava、SparkR 和 RHadoop，这提高了它在大数据场景下的可用性。用户可以将 R 脚本移植到其他编程环境中。关于 R 的详细信息，读者请参考：

```
https://www.r-project.org/.
```

1.2.1 快速入门

启动 R 时的信息如下图所示。所有输入 R 控制台的都是对象，在一个激活的 R 会话中创建的对象都有各自不同的属性，而一个对象附有的一个共同属性称作它的类。在 R 中执行面向对象编程有两种比较普遍的方法，即 S3 类和 S4 类。S3 和 S4 的主要区别在于前者更加灵活，后者是更结构化的面向对象编程语言。S3 和 S4 方法都将符号、字符和数字当作 R 会话中的一个对象，并提供了可使对象用于进一步计算的功能。

```
R version 3.2.1 (2015-06-18) -- "World-Famous Astronaut"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

>
```

1.2.2 数据类型、向量、数组与矩阵

数据集可分为两大类型：原子向量和复合向量。在 R 语言中，原子向量可以分为 5

种类型，即数值或数字型、字符或字符串型、因子型、逻辑型以及复数型；复合向量分为 4 种类型，即数据框、列表、数组以及矩阵。R 中最基本的数据对象是向量，即使将单数位数字赋给一个字母，也会被视为一个单元素向量。所有数据对象都包含模式和长度属性，其中模式定义了在这个对象里存放的数据类型，长度则定义了对象中包含的元素个数。R 语言中的 c() 函数用于将多种元素连接成一个向量。

让我们来看 R 中不同数据类型的一些示例：

```
> x1<-c(2.5,1.4,6.3,4.6,9.0)
> class(x1)
[1] "numeric"
> mode(x1)
[1] "numeric"
> length(x1)
[1] 5
```

在上述代码中，向量 *x*1 是一个数值型向量，元素个数是 5。class() 和 mode() 返回相同的结果，因此都是在确定向量的类型：

```
> x2<-c(TRUE,FALSE,TRUE,FALSE,FALSE)
> class(x2)
[1] "logical"
> mode(x2)
[1] "logical"
> length(x2)
[1] 5
```

在上述代码中，向量 *x*2 是由 5 个元素组成的一个逻辑型向量。逻辑型向量的元素或值可以写成 T/F 或者 TRUE/FALSE。

```
    > x3<-
c("DataMining","Statistics","Analytics","Projects","MachineLearning")
    > class(x3)
    [1] "character"
    > length(x3)
    [1] 5
```

在上述代码中，向量 *x*3 代表了一个长度为 25 的字符型向量。该向量中的所有元素都可以用双引号（" "）或单引号（' '）调用。

```
a <- c('Male','Male','Female','Female','Male','Male',
+ 'Male','Female','Male','Female')
> mode(a)
[1] "character"
> factor(a)
[1] Male Male Female Female Male Male Male Female
[9] Male Female
Levels: Male Female
```

因子是数据的另一种格式，因子型向量中列出了多种分类（也称“水平”）。在上述代码中向量 *a* 是一个字符型向量，它的两个水平 / 分类以一定频率重复。as.factor() 命令用于将字符型向量转换成因子数据类型。使用该命令后，我们可以看到它有 5 个水平：Analytics、DataMining、MachineLearning、Projects 和 Statistics。table() 命令可用于显示因子变量频数表的计算结果：

```
> x<-data.frame(x1,x2,x3)
> class(x)
[1] "data.frame"
> print(x)
  x1    x2              x3
1 12  TRUE       Analytics
2 13 FALSE      DataMining
3 24  TRUE MachineLearning
4 54 FALSE        Projects
5 29  TRUE      Statistics
```

数据框是 R 中另一种常见的数据格式，它可以包含所有不同的数据类型。数据框是一个列表，其中包含了多个等长的向量和不同类型的数据。如果只是从电子表格导入数据集，那么该数据类型将默认为数据框。之后，每个变量的数据类型均可更改。因此，数据框可定义为由包含不同类型的变量列组成的一个矩阵。在前面的代码中，数据框 x 包含了三种数据类型：数值型、逻辑型和字符型。大多数真实数据集会包含不同的数据类型，比如，零售商店里存储在数据库中的客户信息就包括客户 ID、购买日期、购买数量、是否参与了会员计划等。

关于向量的一个要点：向量中的所有元素必须是同类型的。如果不是，R 会进行强行转换。例如，在一个数值型向量中，如果有一个元素是字符型，该向量的类型会从数值型转换成字符型。代码如下所示：

```
> x1<-c(2.5,1.4,6.3,4.6,9.0)
> class(x1)
[1] "numeric"
> x1<-c(2.5,1.4,6.3,4.6,9.0,"cat")
> class(x1)
[1] "character"
```

R 是区分大小写的，比如，“cat”与“Cat”，它们是不同的。所以，用户在给向量分配对象名字时必须格外注意。

有时，要记住所有对象名字不总是那么容易，示例如下：

```
> ls()
 [1] "a"                "centers"          "df"               "distances"
```