大数据应用与技术丛书

现代数据科学

(R 语言·第 2 版)

本杰明·S. 鲍默(Benjamin S. Baumer)
[美] 丹尼尔·T. 卡普兰(Daniel T. Kaplan) 著
尼古拉斯·J. 霍顿(Nicholas J. Horton)

张小明 郭华 张骞允 译

清华大学出版社

北 京

北京市版权局著作权合同登记号 图字：01-2020-6255
Modern Data Science with R, Second Edition/ by Benjamin S. Baumer, Daniel T. Kaplan and Nicholas J. Horton / ISBN: 9780367745448
Copyright@ 2021 by CRC Press.
Authorized translation from English language edition published by CRC Press, a member of the Taylor & Francis Group. All rights reserved.本书原版由Taylor & Francis出版集团旗下CRC Press出版公司出版，并经其授权翻译出版。版权所有，侵权必究。
Tsinghua University Press is authorized to publish and distribute exclusively the Chinese (Simplified Characters) language edition. This edition is authorized for sale in the People's Republic of China only, excluding Hong Kong, Macao SAR and Taiwan. No part of the publication may be reproduced or distributed by any means, or stored in a database or retrieval system, without the prior written permission of the publisher. 本书中文简体翻译版授权由清华大学出版社独家出版。此版本仅限在中华人民共和国境内（不包括中国香港、澳门特别行政区和台湾地区）销售。未经出版者书面许可，不得以任何方式复制或发行本书的任何部分。
Copies of this book sold without a Taylor & Francis sticker on the cover are unauthorized and illegal.
本书封面贴有Taylor & Francis公司防伪标签，无标签者不得销售。
版权所有，侵权必究。举报：010-62782989，beiqinquan@tup.tsinghua.edu.cn

图书在版编目(CIP)数据

现代数据科学 : R 语言 : 第 2 版 / (美) 本杰明 • S.鲍默 (Benjamin S. Baumer)，(美) 丹尼尔 • T.卡普兰(Daniel T. Kaplan)，(美) 尼古拉斯 • J. 霍顿(Nicholas J. Horton) 著；张小明，郭华，张骞允译. —北京：清华大学出版社，2022.1
(大数据应用与技术丛书)
书名原文: Modern Data Science with R, Second Edition
ISBN 978-7-302-59879-4

I. ①现… II. ①本… ②丹… ③尼… ④张… ⑤郭… ⑥张… III. ①统计分析－统计程序 IV. ①C819

中国版本图书馆 CIP 数据核字(2022)第 011522 号

责任编辑：王 军
封面设计：孔祥峰
版式设计：思创景点
责任校对：成凤进
责任印制：刘海龙

出版发行：清华大学出版社
网 址：http://www.tup.com.cn，http://www.wqbook.com
地 址：北京清华大学学研大厦 A 座 邮 编：100084
社 总 机：010-83470000 邮 购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京同文印刷有限责任公司
经 销：全国新华书店
开 本：170mm×240mm 印 张：34.5 字 数：815 千字
版 次：2022 年 3 月第 1 版 印 次：2022 年 3 月第 1 次印刷
定 价：128.00 元

产品编号：086617-01

关于作者

Benjamin S. Baumer 是 Smith 学院统计与数据科学专业的副教授。自 2004 年成为纽约大都会队第一位全职统计分析师以来，他一直是一名应用数据科学家。Benjamin 是 *The Sabermetric Revolution* 和 *Analyzing Baseball Data with R* 的合著者，获得过美国棒球研究学会颁发的 2019 年 Waller 教育奖和 2016 年突出贡献者奖。

Daniel T. Kaplan 是 Macalester 学院 DeWitt Wallace 数学和计算机科学系的名誉教授，是多本统计建模和统计计算教科书的作者。Daniel 获得过 2006 年 Macalester 学院卓越教学奖和 2017 年 CAUSE 终身成就奖。

Nicholas J. Horton 是 Amherst 学院的技术和社会(统计和数据科学)系的教授，是美国统计学会(ASA)和美国科学促进协会(AAAS)的高级会员，美国国家科学院应用与理论统计委员会的联合主席，多项国家教学奖的获得者，一系列统计计算书籍的作者。Nicholas 积极参与数据科学课程，帮助学生“用数据思考”。

前　言

背景和动机

不断增长的数据量和复杂性给分析人员提出了新挑战，他们需要能够对复杂的数据集进行转换来回答重要的统计问题。一份关于本科生数据科学的调查报告[National Academies of Science, Engineering, and Medicine, 2018]指出，数据科学正在彻底改变科学和工作场所。他们将数据科学家定义为“主要从事复杂和海量数据资源分析的知识工作者”。

Michael I. Jordan 将数据科学描述为计算思想和推理(统计)思想的结合。如果没有能够整理(wrangle)或维护(marshal)我们周围日益丰富和复杂的数据的技能，分析师将无法使用这些数据做出更好的决策。

对具备这些技能的毕业生的需求非常迫切。根据公司评级网站 Glassdoor 的报告，2016—2019 年间，每年“数据科学家”都是美国薪酬最高的工作[Columbus, 2019]。

新的数据技术使从比以往更多的来源提取数据成为可能。流线型的数据处理库使数据科学家能够表达如何将这些数据重组为适合分析的形式。数据库系统能使存储和检索越来越多的数据成为可能。最先进的工作流工具有助于形成良好的文档和可再现的分析。现代统计和机器学习方法允许分析者拟合和评估模型，并进行有监督或无监督的学习以提取现实中内在的信息。现代数据科学需要将统计、计算、数据以及通信技能进行紧密结合。

目标受众

本书是为那些想要学习适当技能来处理复杂数据科学项目和“用数据思考”(由谷歌的 Diane Lambert 创造)的读者而写的。使用数据解决问题的愿望是本书介绍的方法的核心。

我们承认，不可能在一本书中涵盖所有这些主题的细节层次：许多章节可以有效地构成一门课程或一系列课程的基础。相反，我们的目标是为现实世界数据的分析打下基础，并确保分析师看到统计和数据分析的作用。阅读本书后，读者将极大地扩展他们处理这些数据的技能，并对他们不断地学习新技术的能力树立新的信心。

本书最初是为了支持一个学期(13 周)的数据科学高年级课程。我们还认为，本书对于相关学科的更高年级的学生也非常有用，而且对于那些想要提高数据科学技能的分析师也很有用。

同时，没有编程或统计经验的普通读者可以阅读本书的第 I 部分。

本书的主要特点

关注案例研究和扩展示例

我们以一系列复杂的、现实世界的扩展案例研究和示例为特色，这些案例研究和示例来自广泛的应用领域，包括政治、交通、体育、环境科学、公共卫生、社交媒体和娱乐。这些丰富的数据集需要使用复杂的数据抽取技术、现代数据可视化方法和精确的计算方法。

背景是这些问题的核心，我们撰写这本书的目的是促进统计思维、数据相关技能和沟通能力的平行发展。每章侧重的应用扩展示例各不相同，而章节练习题可让读者巩固和完善该章中所学的知识。

结构

这本书有三个主要部分和补充附录。第 I 部分介绍数据科学，包括数据可视化的介绍、数据管理(或“整理”)的基础和职业道德伦理。第 II 部分扩展了基础统计的关键建模概念，包括回归建模、分类、预测、统计基础和仿真。第III部分介绍更高级的主题，包括交互式数据可视化、SQL、关系数据库、地理空间数据、文本挖掘和网络科学。

最后，我们在附录中介绍本书的 R 包、R 语言、RStudio、算法思想的一些关键方面、可再现性分析、回归方法，以及如何安装一个本地 SQL 数据库。

相关材料

除了许多例子和扩展的案例研究外，本书第 2～20 章的每章末尾都包含了练习题以及在线提供的补充练习。许多练习都是开放式的，旨在让学生探索在解决数据科学问题时的创造力(可从出版商处获得教师参考答案手册)。

本书的网站是 https://mdsr-book.github.io/mdsr2e，网站内容包括本书目录、各章全文、参考文献、主题和 R 索引。教师网站是 https://mdsr-book.github.io/，网站内容包含代码示例、补充练习、附加练习和勘误表。

第 2 版的变化

数据科学发展迅速，而且自从我们写了第 1 版以来，已发生了很多变化。我们已经更新了所有章节的内容，以详细解释其中的许多变化，并采用了最先进的 R 程序包。

首先，对使用地理空间数据的内容进行了扩展，并分为两章(第 17 章和第 18 章)。第 17 章的重点是处理地理空间数据，第 18 章的重点是地理空间计算。现在，这两章都使用 sf 包和

ggplot2 中的新函数 geom_sf()。这些变化使学生能够更深入地了解地理空间数据分析。

其次，关于规整数据的一章(第 6 章)经历了重大修订。增加了新的关于 list-column 的一节，而关于迭代的部分已经扩展为完整的一章(第 7 章)。第 7 章全部使用了 purrr 包提供的函数式编程风格。这些变化有助于学生养成一种弹性思维习惯：如果复制和粘贴代码超过两次，那么可能有一种更有效的方法。

另外，关于监督学习的内容被分为两章(第 10 章和第 11 章)，并更新为使用 tidymodels 系列包。第 10 章概括介绍模型评估，第 11 章介绍几种模型。tidymodels 生态系统以与 tidyverse 一致的方式为拟合、解释和评估各种机器学习模型提供了相应的语法。这些更改显著减少了本章代码的认知开销。

其他几章的内容做了较小但具有实质性的一些修订。本书中的所有代码都经过了修改，以更加符合 tidyverse 语法和风格。我们还修订了第 1 版的练习题和解答，并添加了新的练习题。每章的代码现在都可从配书网站上找到。这本书已被移植到 bookdown，因此可以在以下网站上找到完整版本：https://mdsr-book.github.io/mdsr2e。

技术的关键角色

虽然许多工具可有效地用于开展数据科学研究，分析技术也在迅速变化，但 R 和 Python 已经成为两个非常强大且可扩展的环境。虽然对于一个数据科学家来说，能够使用多种技术进行分析是很重要的，但我们在本书选择使用 R 和 RStudio 来避免认知过载。我们描述一个强大的和有机联系的工具集，可以在一个学期的范围内学完，这为数据整理和探索提供了基础。

我们充分利用 RStudio 环境。这个强大且易于使用的前端为 R 添加了许多特性，包括对包的支持、代码补全、集成帮助、调试器和其他编码工具。根据我们的经验，RStudio 的使用能极大地提高 R 用户的生产率，并且通过紧密集成可再现的分析工具，有助于避免容易出错的“剪切粘贴”工作流。我们的学生和同事发现 RStudio 有一个非常舒适的界面。使用该界面不需要具备 R 或 RStudio 方面的知识或经验，附录对此提供了简单介绍。

如前所述，我们全面整合了 tidyverse 的许多实质性改进。tidyverse 是一套基础的程序包，提供了与 R 更一致的接口[Wickham，2019h]。tidyverse 软件包中嵌入的许多设计决策解决了传统上使 R 难以用于数据分析的问题。这些决策允许新手用户更快地取得进展并养成良好习惯。

我们使用了一个可再现的分析系统(knitr)来生成本书中的示例代码和输出。该书的网站上提供了从这些文件中提取的代码。我们详细讨论了这些系统的原理和用法。特别是，我们认为与 RStudio 紧密集成的 knitr 和 markdown 包应该成为每个 R 用户的工具箱的一部分。无法想象，在没有它们的情况下如何从事一个项目。

现代数据科学是一项团队工作。为能充分参与，分析师必须能够提出问题，找出解决问题的数据，将其融入计算环境，建立模型，进行探索，然后交流结果。这是一个迭代过程，需要

统计和计算技能的混合。

如何使用本书

本书的材料迄今为止已经支撑了 Amherst 学院、Smith 学院和 Macalester 学院，以及世界上其他学校的多门课程。据我们所知，这些课程包括数据科学的一门中级课程(Smith 学院，2013 年和 2014 年；Amherst 学院，自 2017 年以来的学年)、数据科学的一门入门课程(Smith 学院，2016 年)和高级数据分析的一门顶级课程(Amherst 学院，多个学年)。

Smith 学院数据科学入门课程没有任何先决条件，它主要包括以下材料子集。

- 数据可视化：三周，包括第 1～3 章。
- 数据整理：五周，包括第 4～7 章。
- 职业伦理道德：一周，包括第 8 章。
- 数据库查询：两周，包括第 15 章。
- 地理空间数据：两周，包括第 17 章和第 18 章的部分内容。

Amherst 学院的中级课程采用了[Baumer，2015b]的方法，该方法要求先修一些统计学和计算机科学课程，还有一个期末综合项目。本课程一般包括以下章节：

- 数据可视化：两周，包括第 1～3 章以及第 14 章。
- 数据整理：四周，包括第 4～7 章。
- 职业伦理道德：一周，包括第 8 章。
- 无监督学习：一周，包括第 12 章。
- 数据库查询：一周，包括第 15 章。
- 地理空间数据：一周，包括第 17 章和第 18 章的部分内容。
- 文本挖掘：一周，包括第 19 章。
- 网络科学：一周，包括第 20 章。

Amherst 学院的“顶级”课程更深入地回顾了其中大部分内容：

- 数据可视化：三周，包括第 1～3 章以及第 14 章。
- 数据整理：两周，包括第 4～7 章。
- 职业伦理道德：一周，包括第 8 章。
- 仿真：一周，包括第 13 章。
- 统计学习：两周，包括第 10～12 章。
- 数据库：一周，包括第 15 章和附录 F。
- 文本挖掘：一周，包括第 19 章。
- 空间数据：一周，包括第 17 章。
- 大数据：一周，包括第 21 章。

我们期望本书可作为其他各类课程的主要资料，无论是否有额外的补充材料。

第Ⅰ部分的内容，特别是第 3 章介绍的 ggplot2 可视化概念和第 4 章介绍的 dplyr 数据整理操作是基本内容，是第Ⅱ部分和第Ⅲ部分内容的前提。第Ⅲ部分内容中的每一个主题都彼此独立，与第Ⅱ部分内容中的材料无关。因此，尽管大多数教师希望在所有课程中涵盖第Ⅰ部分的大部分或全部内容，但第Ⅱ部分和第Ⅲ部分的内容几乎可以完全自由地添加。

第Ⅱ部分内容中的材料旨在让初学者了解统计学(即基本的推理和线性回归)知识，让学生走进更丰富多彩的统计建模和统计推理世界。

彩图

本书是黑白印刷，无法显示彩图效果。读者可扫描封底二维码，下载彩色图片。

参考资料

正文所涉及参考资料的形式是[*]，即方括号中加编号。读者可扫描封底二维码，下载“参考资料”文档，从中找到编号对应的资料。

致谢

感谢 Informa CRC/Chapman 和 Hall 的 John Kimmel 对我们的支持和指导。感谢 Jim Albert、Nancy Boynton、Jon Caris、Mine Çetinkaya-Rundel、Jonathan Che、Patrick Frenett、Scott Gilman、Maria-Cristiana Gîrjău、Johanna Hardin、Alana Horton、John Horton、Kinari Horton、Azka Javaid、Andrew Kim、Eunice Kim、Caroline Kusiak、Ken Kleinman、Priscilla (Wencong) Li、Amelia McNamara、Melody Owen、Randall Pruim、Tanya Riseman、Gabriel Sosa、Katie St. Clair、Amy Wagaman、Susan (Xiaofei) Wang、Hadley Wickham、J. J. Allaire，感谢 RStudio 的开发者们、匿名评论员、Smith 和 Amherst 学院的多个班级，感谢其他许多对 R 环境、RStudio 环境、手稿提出有用建议的人士。Rose Porta 对于校对和简化从 Sweave 到 R Markdown 的过渡过程发挥了重要作用。Jessica Yu 将第 1 版的大部分练习转换为基于 etude 的新格式，并添加了标签。

尤其需要提出的是，我们非常感谢 Cory、Maya 以及 Julia 的耐心和支持。

目　录

第 I 部分　数据科学简介

第 1 章　序言：为什么有数据科学？ ……2
1.1　数据科学是什么 …… 3
1.2　案例学习：棒球资料统计分析的演变 …… 5
1.3　数据集 …… 6
1.4　扩展资源 …… 7

第 2 章　数据可视化 ……8
2.1　2012 年联邦大选周期 …… 8
2.1.1　这两组数据有区别吗？ …… 10
2.1.2　图形变化 …… 11
2.1.3　检查变量之间的关系 …… 12
2.1.4　网络 …… 14
2.2　组成数据图形 ……15
2.2.1　数据图形分类 …… 15
2.2.2　颜色 …… 17
2.2.3　剖析数据图形 …… 19
2.3　数据图形的重要性：挑战者号 ……21
2.4　创建有效的演示 ……25
2.5　更广阔的数据可视化世界 ……25
2.6　扩展资源 ……27
2.7　练习题 ……28
2.8　附加练习 ……29

第 3 章　图形语法……30
3.1　数据图形语法 ……30
3.1.1　画面 …… 31
3.1.2　刻度 …… 33
3.1.3　指南 …… 35
3.1.4　方面 …… 35
3.1.5　层次 …… 36
3.2　R 中的规范数据图形 ……38
3.2.1　单变量显示 …… 38
3.2.2　多元显示 …… 40
3.2.3　地图 …… 46
3.2.4　网络 …… 46
3.3　扩展示例：历史婴儿名字 ……47
3.3.1　至今仍活着的人口的百分比 …… 48
3.3.2　最普通的女性名字 …… 53
3.4　扩展资源 ……56
3.5　练习题 ……56
3.6　附加练习 ……58

第 4 章　在一张表中整理数据…… 59
4.1　数据整理语法 ……59
4.1.1　select()和 filter()…… 59
4.1.2　mutate()和 rename() …… 62
4.1.3　arrange() …… 64
4.1.4　用 group_by()进行 summarize() …… 66
4.2　扩展示例：Ben 在大都会队(Mets)的时间……67
4.3　扩展资源 ……75
4.4　练习题 ……76
4.5　附加练习 ……79

第 5 章 多张表的数据整理 ····· 80
5.1 inner_join() ····· 80
5.2 left_join() ····· 82
5.3 扩展示例：Manny Ramirez ····· 83
5.4 扩展资源 ····· 90
5.5 练习题 ····· 90
5.6 附加练习 ····· 92

第 6 章 数据规整 ····· 93
6.1 规整数据 ····· 93
6.1.1 动机 ····· 93
6.1.2 规整的数据是什么 ····· 95
6.2 重塑数据 ····· 101
6.2.1 用于从宽到窄以及从窄到宽转换的数据动词 ····· 103
6.2.2 pivot_wider()函数 ····· 103
6.2.3 pivot_longer()函数 ····· 103
6.2.4 list-column ····· 104
6.2.5 示例：中性姓名 ····· 108
6.3 命名约定 ····· 110
6.4 数据获取 ····· 111
6.4.1 数据表友好的格式 ····· 111
6.4.2 API ····· 115
6.4.3 清洗数据 ····· 115
6.4.4 示例：日本核反应堆 ····· 121
6.5 扩展资源 ····· 124
6.6 练习题 ····· 124
6.7 附加练习 ····· 127

第 7 章 迭代 ····· 128
7.1 向量化操作 ····· 128
7.2 利用 dplyr 实现 across() ····· 131
7.3 map()函数族 ····· 132
7.4 在一维向量上迭代 ····· 133
7.4.1 迭代已知函数 ····· 133
7.4.2 迭代任意函数 ····· 134
7.5 在子组上迭代 ····· 135
7.5.1 示例：期望获胜率 ····· 136
7.5.2 示例：年度领导者 ····· 138
7.6 仿真 ····· 140
7.7 扩展示例：与 BMI 相关的因子 ····· 142
7.8 扩展资源 ····· 145
7.9 练习题 ····· 145
7.10 附加练习 ····· 145

第 8 章 数据科学伦理学 ····· 146
8.1 引言 ····· 146
8.2 真实的谎言 ····· 146
8.2.1 坚持你的立场 ····· 147
8.2.1 全球温度 ····· 147
8.2.3 COVID-19 报道 ····· 148
8.3 数据科学在社会中的作用 ····· 149
8.4 职业道德的一些设置 ····· 150
8.4.1 CEO ····· 150
8.4.2 就业歧视 ····· 150
8.4.3 “Gaydar” ····· 151
8.4.4 种族预测 ····· 151
8.4.5 数据爬取 ····· 152
8.4.6 可重复的电子表格分析 ····· 152
8.4.7 药物危害 ····· 152
8.4.8 法律谈判 ····· 153
8.5 道德行为的一些指导性原则 ····· 153
8.5.1 CEO ····· 154
8.5.2 就业歧视 ····· 154
8.5.3 “Gaydar” ····· 154
8.5.4 种族预测 ····· 154
8.5.5 数据爬取 ····· 155
8.5.6 可重复的电子表格分析 ····· 155
8.5.7 药物危害 ····· 155
8.5.8 法律谈判 ····· 156
8.6 算法偏差 ····· 157

8.7 数据与泄露 …… 157
8.7.1 重新识别和泄露避免 ……157
8.7.2 安全的数据存储 ……158
8.7.3 数据爬取和使用条款 ……158
8.8 再现性 …… 159
8.9 集体的道德准则 …… 160
8.10 职业道德行为准则 …… 161
8.11 扩展资源 …… 161
8.12 练习题 …… 162
8.13 附加练习 …… 163

第Ⅱ部分　统计与建模

第 9 章　统计基础 …… 166
9.1 样本和总体 …… 166
9.2 样本统计 …… 169
9.3 自举 …… 173
9.4 异常值 …… 176
9.5 统计模型：方差解释 …… 178
9.6 混淆和解释其他因素 …… 181
9.7 p 值的风险 …… 184
9.8 扩展资源 …… 186
9.9 练习题 …… 187
9.10 附加练习 …… 187

第 10 章　预测建模 …… 188
10.1 预测建模 …… 189
10.2 简单的分类模型 …… 190
10.3 评价方法 …… 197
10.3.1 权衡偏差与方差 ……197
10.3.2 交叉验证 ……197
10.3.3 混淆矩阵和 ROC 曲线 ……198
10.3.4 定量响应的预测误差测量 ……200
10.3.5 示例：收益模型评估 ….201
10.4 扩展示例：谁患有糖尿病？ …… 204
10.5 扩展资源 …… 207
10.6 练习题 …… 207
10.7 附加练习 …… 208

第 11 章　监督学习 …… 209
11.1 非回归分类器 …… 209
11.1.1 决策树 …… 210
11.1.2 随机森林 …… 217
11.1.3 最近邻 …… 218
11.1.4 朴素贝叶斯 …… 221
11.1.5 人工神经网络 …… 222
11.1.6 集成方法 …… 224
11.2 参数调整 …… 224
11.3 示例：收入模型 redux 的评价 …… 226
11.4 扩展示例：这次谁得了糖尿病？ …… 230
11.5 正则化 …… 235
11.6 扩展资源 …… 237
11.7 练习题 …… 238
11.8 附加练习 …… 239

第 12 章　无监督学习 …… 240
12.1 聚类 …… 240
12.1.1 层次聚类 …… 241
12.1.2 *k*-means …… 245
12.2 降维 …… 246
12.2.1 直觉法 …… 248
12.2.2 奇异值分解 …… 249
12.3 扩展资源 …… 254
12.4 练习题 …… 254
12.5 附加练习 …… 255

第 13 章　仿真 …… 256
13.1 逆向推理 …… 256

13.2 扩展示例：癌症分组 …… 257
13.3 随机化函数 …… 259
13.4 仿真可变性 …… 261
13.4.1 部分计划的会合 ……261
13.4.2 工作报告……262
13.4.3 餐厅健康卫生等级 ……264
13.5 随机网络 …… 267
13.6 仿真关键原则 …… 267
13.6.1 设计……267
13.6.2 模块化……267
13.6.3 再现性和随机数种子 ….267
13.7 扩展资源 …… 270
13.8 练习题 …… 270
13.9 附加练习 …… 271

第Ⅲ部分 数据科学专题

第 14 章 动态定制数据图形 ……274
14.1 使用 D3.js 和 htmlwidgets 丰富 Web 内容 …… 274
14.1.1 Leaflet ……275
14.1.2 Plot.ly……275
14.1.3 DataTables ……276
14.1.4 dygraphs……277
14.1.5 streamgraphs ……277
14.2 动画 …… 278
14.3 flexdashboard …… 279
14.4 基于 Shiny 的交互式 Web 应用程序 …… 281
14.4.1 示例：披头士名字的交互展示 ……281
14.4.2 反应式编程的更多知识 ……282
14.5 ggplot2 图形的定制 …… 285
14.6 扩展示例：吃热狗 …… 289
14.7 扩展资源 …… 294
14.8 练习题 …… 294
14.9 附加练习 …… 295

第 15 章 使用 SQL 查询数据库…… 296
15.1 从 dplyr 到 SQL …… 296
15.2 平面文件数据库 …… 300
15.3 SQL …… 302
15.4 SQL 数据操作语言 …… 303
15.4.1 SELECT...FROM …… 305
15.4.2 WHERE …… 307
15.4.3 GROUP BY…… 310
15.4.4 ORDER BY…… 312
15.4.5 HAVING …… 313
15.4.6 LIMIT …… 315
15.4.7 JOIN …… 316
15.4.8 UNION …… 321
15.4.9 子查询…… 321
15.5 扩展示例：FiveThirtyEight 航班 …… 323
15.6 SQL 与 R …… 331
15.7 扩展资源 …… 331
15.8 练习题 …… 331
15.9 附加练习 …… 333

第 16 章 数据库管理…… 334
16.1 构建高效的 SQL 数据库 ….. 334
16.1.1 创建新的数据库…… 334
16.1.2 创建表…… 335
16.1.3 键…… 336
16.1.4 索引…… 337
16.1.5 查询计划…… 338
16.1.6 分区…… 340
16.2 更改 SQL 数据 …… 340
16.2.1 更改数据…… 341
16.2.2 增加数据…… 341
16.2.3 从文件导入数据…… 342

16.3 扩展示例：创建数据库 …… 342
16.3.1 抽取 …… 342
16.3.2 转换 …… 342
16.3.3 载入 MySQL 数据库 …… 343
16.4 可扩展性 …… 346
16.5 扩展资源 …… 346
16.6 练习题 …… 346
16.7 附加练习 …… 347

第 17 章 使用地理空间数据 …… 348
17.1 动机：地理空间数据有什么了不起的？ …… 348
17.2 空间数据结构 …… 351
17.3 制作地图 …… 353
17.3.1 静态地图 …… 353
17.3.2 投影 …… 354
17.3.3 基于 leaflet 的动态地图 …… 360
17.4 扩展示例：国会选区 …… 362
17.4.1 选举结果 …… 362
17.4.2 国会选区 …… 365
17.4.3 整合所有数据 …… 367
17.4.4 使用 ggplot2 …… 368
17.4.5 使用 leaflet …… 369
17.5 有效的地图：如何避免撒谎 …… 370
17.6 投影多边形 …… 371
17.7 有效利用其他技术 …… 373
17.8 扩展资源 …… 374
17.9 练习题 …… 374
17.10 附加练习 …… 374

第 18 章 地理空间计算 …… 375
18.1 地理空间操作 …… 375
18.1.1 地理编码、路线和距离 …… 375
18.1.2 几何运算 …… 378
18.2 地理空间聚合 …… 384
18.3 地理空间联接 …… 386
18.4 拓展示例：MacLeish 中的路径海拔 …… 387
18.5 扩展资源 …… 391
18.6 练习题 …… 391
18.7 附加练习 …… 392

第 19 章 文本数据 …… 393
19.1 使用 Macbeth 的正则表达式 …… 393
19.1.1 解析苏格兰戏剧文本 …… 393
19.1.2 Macbeth 中的生与死 …… 397
19.2 扩展示例：分析 arXiv.org 的文本数据 …… 399
19.2.1 语料库 …… 402
19.2.2 词云 …… 404
19.2.3 情感分析 …… 405
19.2.4 bigrams 和 N-grams …… 407
19.2.5 文档词项矩阵 …… 408
19.3 获取文本 …… 412
19.4 扩展资源 …… 415
19.5 练习题 …… 416
19.6 附加练习 …… 418

第 20 章 网络科学 …… 419
20.1 网络科学引言 …… 419
20.1.1 定义 …… 419
20.1.2 网络科学简史 …… 420
20.2 扩展示例：Kristen Stewart 的六度空间理论 …… 424
20.2.1 收集好莱坞数据 …… 424
20.2.2 构建好莱坞网络 …… 427
20.2.3 建立 Kristen Stewart 预言 …… 430
20.3 PageRank …… 433

20.4 扩展示例：1996 年男子大学篮球赛 …… 435
20.5 扩展资源 …… 442
20.6 练习题 …… 442
20.7 附加练习 …… 443

第 21 章 结束语：走向“大数据” ……444
21.1 大数据的一些概念 …… 444
21.2 更大数据的工具 …… 446
21.2.1 大数据的数据和内存结构 ……446
21.2.2 编译……447
21.2.3 并行和分布式计算 ……447
21.2.4 SQL 的替代方案 ……454
21.3 R 的替代方案…… 455
21.4 结束语 …… 456
21.5 扩展资源 …… 456

第Ⅳ部分 附录

附录 A 本书使用的包…… 458
附录 B R 和 RStudio 简介 …… 465
附录 C 算法思维…… 483
附录 D 可再现性分析和工作流程…… 495
附录 E 回归建模…… 504
附录 F 安装数据库服务器…… 524

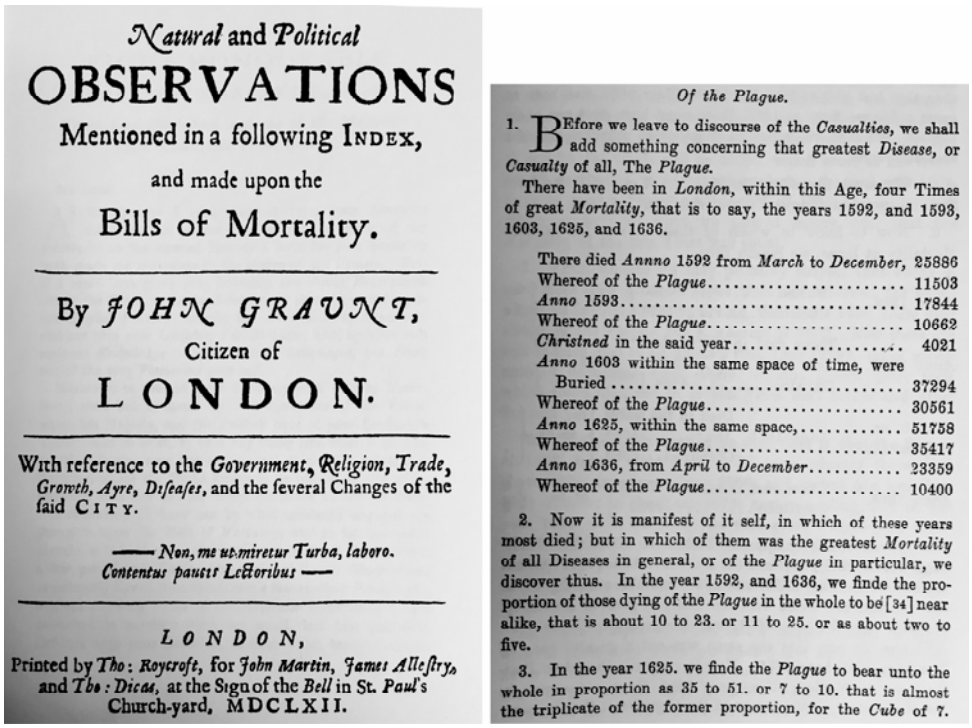
第Ⅰ部分

数据科学简介

第 1 章

序言：为什么有数据科学?

通常来说，信息是我们想要的，但数据是我们拥有的。将数据转换为信息的技术可以追溯到数百年前。一个重要的里程碑是 1592 年，约翰·格伦特(John Graunt)在伦敦发表了每周“死亡清单”，见图 1.1。这些“清单”是列表形式，它是由各个单独的事件数据汇总而成的，以一种人类更容易理解的形式向读者进行展示，然而这些列表是手工构造的。



Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,

and made upon the

Bills of Mortality.

By JOHN GRAUNT,

Citizen of

LONDON.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Ayre*, *Diseases*, and the several Changes of the said CITY.

—— *Non, me ut miretur Turba, laboro.*
Contentus paucis Lectoribus ——

LONDON,

Printed by *Tho*: *Roycroft*, for *John Martin*, *James Allestry*, and *Tho*: *Dicas*, at the Sign of the *Bell* in St. *Paul*'s Church-yard, MDCLXII.

Of the Plague.

1. BEfore we leave to discourse of the *Casualties*, we shall add something concerning that greatest *Disease*, or *Casualty* of all, The *Plague*.

There have been in *London*, within this Age, four Times of great *Mortality*, that is to say, the years 1592, and 1593, 1603, 1625, and 1636.

There died *Annno* 1592 from *March* to *December*,	25886
Whereof of the *Plague*.....	11503
Anno 1593.....	17844
Whereof of the *Plague*.....	10662
Christned in the said year.....	4021
Anno 1603 within the same space of time, were Buried	37294
Whereof of the *Plague*.....	30561
Anno 1625, within the same space,.....	51758
Whereof of the *Plague*.....	35417
Anno 1636, from *April* to *December*.....	23359
Whereof of the *Plague*.....	10400

2. Now it is manifest of it self, in which of these years most died; but in which of them was the greatest *Mortality* of all Diseases in general, or of the *Plague* in particular, we discover thus. In the year 1592, and 1636, we finde the proportion of those dying of the *Plague* in the whole to be [34] near alike, that is about 10 to 23. or 11 to 25. or as about two to five.

3. In the year 1625. we finde the *Plague* to bear unto the whole in proportion as 35 to 51. or 7 to 10. that is almost the triplicate of the former proportion, for the *Cube* of 7.

(a) 标题页　　(b) 瘟疫节录

图 1.1　摘自格伦特的死亡清单

在过去几个世纪中，随着数据量的增加，人们引入了机器来加快制表速度。其中，迈出的

重要一步是赫尔曼·霍勒里斯(Herman Hollerith)为 1890 年美国人口普查开发的打孔卡和电子制表系统。这一成功如此之大，以至于霍勒里斯创立了 IBM 公司，该公司在当今电子计算机的发展中产生了至关重要的作用。

另一方面，在 19 世纪末，统计方法开始迅速发展。这些方法在解释数据方面非常重要，但它们本质上与机器处理数据无关。当时，经过几代的发展，学生已经学会了手工处理少量数据并进行统计操作。

如今，我们经常会面临规模非常庞大的数据集，它们只能由机器处理。在这个“大数据时代”，数据被特定设备和计算机网络汇集，生成这些数据的环境多种多样，如基因组数据、卫星对地球的观测数据、网络用户的录入数据、销售交易数据等。这些大数据为使用数据挖掘、机器学习、数据可视化等技术来发现和表征数据模式带来了新机遇。这些技术都需要计算机来处理，这些处理需要执行的操作包括：数据清理、组合来自多个数据源的数据，以及将数据整理为特定形式以适合作为可视化和建模等数据整合处理的输入。

在写作本书时，我们希望能帮助人们理解数据整理(为可视化和其他现代统计解释技术准备数据的过程)，并掌握使用这些数据通过建模和可视化方法回答统计问题的技巧。这样做不可避免地涉及进行统计推理以及熟练运用各种计算方法和算法的能力。

美国国家科学院“本科生数据科学”共识报告指出：数据科学这一重要的新领域涵盖了一系列被称为“数据敏锐性”的能力。数据敏锐性的关键组成部分包括数学、计算学和统计学基础、数据管理和存储、数据描述和可视化、数据建模和评估、工作流程和可再现性、沟通和团队合作、特定领域的处理，以及道德问题的解决。他们建议所有学生都应能从数据科学的“意识和能力”中受益。

为执行复杂的计算，是否需要对计算机程序进行深入研究？我们认为事实并非如此。

首先，在过去的半个世纪中，已经开发出了一系列便于使用的数据操作，这些操作可被当成复杂的数据整理过程的基础。诀窍不是精通编程，而是学会根据这些操作进行思考。本书的大部分内容旨在帮助你领会这种思想。

其次，可以利用软件的最新发展来大大减少使用这些数据操作所需的编程量。我们已经可以使用这类软件，尤其是 R 语言以及 dplyr 和 ggplot2 程序包，并且只需要重点关注一小部分能以简洁明了的方式完成数据整理任务的功能。编程语法严格一致，只需要通过少量实践，你就能调整本书中的代码来解决自己的问题。经验丰富的 R 程序员将能注意到本书中 R 语句的独特风格，包括始终如一地专注于少量函数集以及对 pipe(管道)运算符的广泛使用。本书的第Ⅰ部分着重于数据整理和数据可视化，该部分是数据科学的关键组成部分。

1.1　数据科学是什么

数据科学具有广阔的应用前景，它被认为是从数据中提取有意义信息的科学。这个简单定义包含了几个关键思想。首先，数据科学是一门科学，是一门源于数学的严谨学科，它结合了统计学和计算机科学中的关键元素。加州大学伯克利分校的迈克尔·乔丹(Michael Jordan)将数

据科学描述为统计学和计算机科学中经典知识的无缝融合：

计算机科学不仅是编程，它为表达可计算的结构创造了有效的抽象表示方法，并且在这些抽象的表示形式上提出了一系列算法。同样，统计学不仅是估计和检验的集合，它是抽样、模型、分布和决策等常用概念相互作用的科学。[数据科学]基于这样的思想，即这些思维方式相互支持[Pierson，2016]。

其次，在有原始数据领域专业知识的背景下，数据科学的应用效果非常好。这些领域可能涉及从天文学到动物学的各种学科，在这些领域中，商业和医疗保健是特别重要的两个。

第三，数据和信息之间的区别是数据科学存在的依据，因此，数据科学家总是努力地将现有的丰富数据转换为看似稀疏的可操作信息。

许多统计学家都会说：“我们在这方面已经有了一个学科：这就是统计学！”数据科学家和统计学家的目标是相同的：都希望从数据中提取有意义的信息。许多统计技术最初都在数据稀缺、收集困难或昂贵的环境中产生，统计学家专注于发明在数据量稀少的情况下能够最大限度地提高推理能力的方法。这些技术通常都很巧妙，涉及复杂的数学理论知识，并且在一个世纪以来已被经验科学证明具有不可估量的价值。尽管几位最具影响力的早期统计学家将计算理论视为统计学的组成部分，但确实，统计学理论的发展主要是为我们尚无法计算的问题找到数学近似解[Cobb，2007]。

如今，我们从数据中提取有价值的信息的方式在两方面存在差异，这两方面都主要归因于计算理论的进步：

(1) 我们能够比以前计算更多的东西。

(2) 拥有比以前更多的数据。

第一个变化意味着 20 世纪统计学教学中普遍讲授的某些技术(例如 t 检验、方差分析)已被概念上更简单的计算技术所取代，直到微机革命(例如自举方法、枚举测试)，这种现象才有所改变。第二个变化意味着我们现在收集的许多数据都是可观察的，它们不是来自于设计好的实验，并且它们并不是真正地随机抽样的。这使为这些数据开发可行的概率模型具有非常大的挑战性，而反过来又使形式化的统计推断成为更具挑战性的问题(也许不太相关)。某些情况下(例如临床试验和 A / B 测试)，模型参数的有效估计仍然是一个目标，推论统计仍然是解决问题的主要工具。但在一系列学术、政府和工业环境中，最终目标可能是一个可用于预测的模型、数据的交互式可视化或允许用户对数据进行切片和切分以进行简单比较的 Web 应用程序。在本书的第 II 部分，我们将更深入地探讨与统计推断和建模有关的问题。

现代数据日益复杂和异构，这意味着每个数据分析项目都需要定制。简而言之，现代数据分析人员需要能够读写计算机指令，即从中构建数据分析项目的所谓“代码”。本书的第 I 部分介绍数据可视化和数据整理的基本能力，这是现代数据科学家必须掌握的两种基本技术。这些章节着重于传统的数据二维表示：数据表中的行和列，数据图形中的水平线和垂直线。在第III部分，我们将探索各种非传统数据类型(例如空间、文本、网络、大数据)和交互式数据图形。

通过学习本书，你将获得被称为大数据的“先驱者”的计算技能[Horton et al., 2015]。在第

21 章，我们介绍了一些用于处理真正的大数据的工具。人在走路之前必须先学会爬行，我们认为对于大多数人而言，此处学习的技能与你可能会遇到的各种问题更紧密相关。

1.2　案例学习：棒球资料统计分析的演变

棒球分析的发展(通常称为棒球资料统计分析)在许多方面概括了其他领域分析的发展。尽管领域知识在数据科学中非常重要，但本节不需要许多的棒球知识[1]。

在棒球运动中使用统计知识已有很长的历史，一部分原因是比赛数据本身是自然离散的，另一部分原因是亨利·查德威克(Henry Chadwick)在 20 世纪初期就开始发布比赛得分[Schwarz, 2005]。由于这些原因，人们开始积累了丰富的棒球运动数据。

但是，尽管堆积了越来越多的棒球数据，但对数据的分析并不普遍。也就是说，现存的数据提供了一种保存记录的途径，其结果是比赛历史上的一些数字元素占据了运动员的整个职业生涯(例如，Babe Ruth 的 714 个本垒打)。但是目前尚不清楚人们从数据中学到了多少有关棒球比赛的知识。知道 Babe Ruth 比 Mel Ott 击出更多的本垒打可以告诉我们一些有关两名运动员的信息，但不能提供有关比赛本身的任何见解。

1947 年的杰基·罗宾逊(Jackie Robinson)的新秀赛季中，布鲁克林道奇队(Brooklyn Dodgers)总经理布兰奇·里奇(Branch Rickey)做出了另一项重大创新：他聘请艾伦·罗斯(Allan Roth)作为棒球界的第一位统计分析师。罗斯通过对棒球数据的分析得出了布鲁克林道奇队可赢得更多比赛的见解。尤其是，罗斯说服里奇，击球手以任意方式(例如击打、步行)到达第一垒的频率比击球手通过命中(过去，可能仍然是最常引用的击球统计数据)到达第一垒的频率能更好地表明击球手的价值。支持这一观点的逻辑既来源于罗斯对棒球比赛的理解(我们称之为领域知识)，又基于他对棒球数据的统计分析。

在接下来的 50 年中，各种各样的人为棒球分析做出了许多重要贡献，最著名的是“棒球资料统计分析之父”比尔·詹姆斯(Bill James) [James, 1986]；这些人中的大多数没有接受过统计学方面的正规培训。他们依赖的有力工具就是电子表格，重要的是他们能够利用自己的创造力、领域知识以及对特定问题的敏锐观察力来进行有趣的发现。

2003 年出版的 *Moneyball*[Lewis, 2003]展示了比利·比恩(Billy Beane)和保罗·迪波德斯塔(Paul DePodesta)如何使用统计分析来分析奥克兰运动家队(Oakland A's)，该分析引发了棒球 front office 管理方式的革命[Baumer and Zimbalist, 2014]。在接下来的十年中，数据的规模迅速扩大，以至电子表格不再适合存储所有可用的数据，更不用说数据分析了。如今，许多专业运动团队都拥有由统计学或计算机科学博士领导的研发小组[Baumer，2015a]。考虑到美国职业棒球大联盟每年的收入估计超过 80 亿美元，这一点不足为奇。

1 棒球的主要规则如下：由九名球员组成的两支球队交替在包含四垒(一垒、二垒、三垒或主场)的区域冲刺以得分。防守队投球，而进攻队中的一名成员则站在本垒旁击球。进攻球员在按顺序冲过其他垒后越过本垒板，即可得分。

下一代棒球分析家要有所贡献的话，将需要编码能力。推动艾伦·罗斯和比尔·詹姆斯工作的创造力和领域知识仍然是成功的必要特征，但仅仅靠这两点已经不足以保证成功。在这方面，棒球没有什么特别的，现在类似的大量数据还可以在许多其他领域获得，包括天文学、卫生服务研究、基因组学和气候学等。对于所有应用领域的数据科学家来说，创造力、领域知识和技术能力绝对至关重要。

1.3 数据集

本书使用了许多数据集。较小的数据集可通过 R 语言的 mdsr(见附录 A)或 mosaic 程序包获得。本书使用的其他一些数据可以直接从 Internet 提取，这些数据的 URL 包含在正文中。这里有一些更大的、更复杂的数据集，我们可以重复使用，并在这里给出一些说明。

航空公司延误数据集(Airline Delays)

自 1987 年 10 月以来，美国运输统计局已收集了超过 1.69 亿次国内航班的数据。我们开发了 airlines 程序包，以允许 R 语言用户以最小的代价下载和处理这些数据。有关如何安装数据库的说明，请参见附录 F。这些数据最初用于 2009 ASA 数据博览会[Wickham, 2011]。nycflights13 程序包包含了这个数据的子集(仅包含 2013 年离开纽约三个最著名机场的航班)。

棒球数据集(Baseball)

Lahman 数据集由自述数据库日志记录员肖恩·拉曼(Sean Lahman)维护。数据集由一组志愿者编辑而成，包含了可追溯至 1871 年的完整的季节赛记录；数据通常每年更新一次。它既可作为预打包的 SQL 文件，又可作为 R 包下载[Friendly et al., 2020]。

婴儿名字数据集(Baby Names)

R 的 babynames 程序包提供了来自美国社会保障局[Wickham，2019c]的有关单个婴儿姓名流行程度的数据，这些数据可用于跟踪某些姓名随时间而变化的流行情况。

联邦选举委员会数据集(Federal Election Commission)

Fec 程序包[Baumer and Gjekmarkaj, 2017]可访问联邦选举委员会维护的近期联邦选举的竞选支出数据。这些数据包括个人对委员会的贡献、委员会代表的支出，还包括作为总统、参议院和众议院的候选人的支出，以及有关这些委员会委员和候选人的信息。fec12 和 fec16 包以简化的形式提供单个选举周期的信息[Tapal et al., 2020b]。

麦克里什数据集(MacLeish)

Ada 和 Archibald MacLeish 野外研究站是史密斯学院(Smith College)拥有并运营的 260 英亩土地。教师、学生和当地社区的成员都使用它进行环境研究、户外活动和娱乐。R 中的 Macleish 程序包允许你使用 etl 框架从 Macleish 野外研究站下载和处理天气数据(时间序列形式)[Baumer

et al., 2020]。它还包含表示空间点邻近区域信息的 shapefile。

电影数据集(Movies)

互联网电影数据集(IMDb)存储了关于各种电影的大量信息[IMDB.com，2013]。将 IMDb 数据导入 SQL 的最简单方法是使用开源的 Python 程序包 IMDbPY [Alberani, 2014]。

餐厅违规信息数据集(Restaurant Violations)

mdsr 程序包包含了纽约市卫生局生成的餐厅健康检查数据。

1.4　扩展资源

每章都有一个扩展资源列表，这些资源可作为给定主题的权威参考，可提供更多详细信息或建议要探索的其他材料。数据科学和数据分析的定义比比皆是。更多例子请参见 Donoho (2017)、De Veaux et al. (2017)、Cobb (2015)、Horton and Hardin (2015)、Hardin et al. (2015)、Finzer (2013)、Provost and Fawcett (2013)以及 Cleveland (2001)。有关数据敏锐度组成部分的更多信息，请参见 National Academies of Science, Engineering, and Medicine (2018)。

第 2 章

数据可视化

数据图形提供了一种最易于访问、引人注目的和最具表现力的模式来研究和描绘数据中的模式。本章将让读者理解一个设计精美的数据图形的重要性，并介绍一种方便读者理解其组成部分的分类法。如果你是第一次看到这种材料，你将永远不会再用以前的方式看数据图形了，你的眼睛将很快成为一个更加敏锐的镜头。

2.1 2012 年联邦大选周期

美国每四年一次的总统选举引起了民众的广泛兴趣。最杰出的候选人在正式选举前两年就宣布自己的候选人资格，并开始筹集数亿美元来组织一场全国性竞选活动。从许多方面来说，成功进行总统大选的经验本身就证明了担任总统所需的领导能力和组织能力。

各方面的政治声音都对金钱在政治活动中的重大影响持批评态度。虽然公民个体对个人候选人的捐款受到各种限制，但美国最高法院在《公民联合会诉联邦选举委员会案》中的判决允许公司(非营利性组织或其他组织)进行无限的政治支出。这样，委员会体系(最著名的是 PAC)可以接受无限制的捐款，并代表(或反对)特定候选人或多个候选人使用这些捐款。因此，解开复杂的竞选支出网络是一个能引起人们极大兴趣的课题。

进行拆解是数据科学中的一项重要工作。联邦选举委员会(FEC)维护着一个网站，不仅记录了个人对候选人和委员会的所有捐款(200 美元或更多)，而且包括委员会代表(和反对)候选人进行开支的日志。当然，FEC 还保留有关哪些候选人赢得选举，以及获得多少选票的数据。这些数据源是独立的，需要一定的技巧才能将它们组合在一起。我们将在第 4 章和第 5 章介绍这些技能，但是在本章，我们将主要关注从这些数据中可收集到的信息的图形化展示。在此阶段，我们的重点是如何更好地展示某些数据以传递更加清晰(正确)的信息，从而帮助用户做出明智决定。

最基本的问题之一就是：每个候选人筹集了多少钱？但是，混乱的竞选财务网络使这个简单问题甚至也难以回答，而且更重要的是，它提供的信息也许比我们想象的更糟。一个更好的问题是：哪个候选人身上花费的钱最多？在图 2.1 中，我们展示了一张条形图，该图显示了在 2012 年联邦选举周期中的大选阶段，委员会花在特定候选人身上的金额(百万美元)，人物对象包括总统、参议院和众议院的候选人。图 2.1 仅包括竞选花费至少为 400 万美元的候选人。

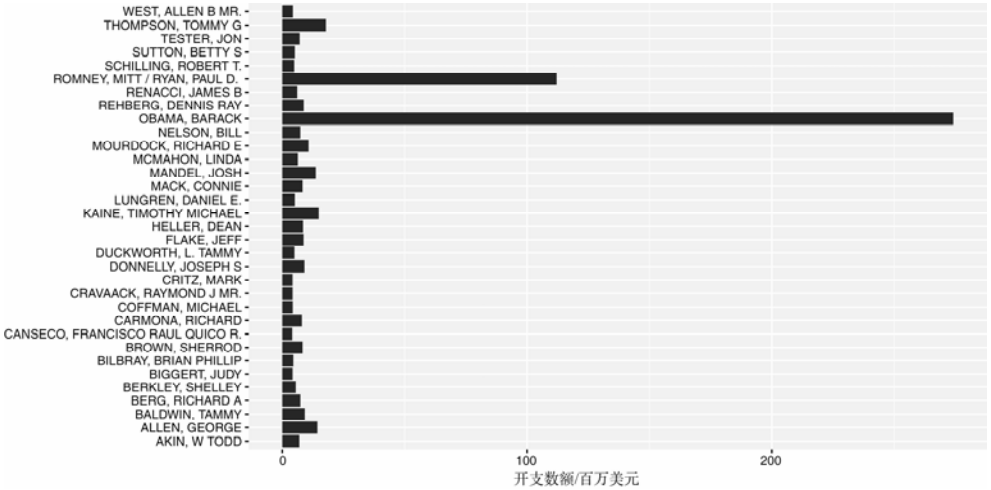


图 2.1　2012 年联邦选举周期大选阶段单个候选人身上的开支。图中只统计了开支超过 400 万美元的候选人

从图 2.1 可以清楚地看出，巴拉克·奥巴马(Barack Obama)总统的竞选活动花费的资金远远超过其他所有候选人，尤其是比他的共和党挑战者罗姆尼(Mitt Romney)花费的资金增加了一倍以上。但是，委员会不仅限于花钱支持候选人，还可以花钱攻击特定候选人(即花在攻击性广告上)。在图 2.2 中，我们根据资金是花费在攻击某个候选人身上还是花费在支持某个候选人身上，将图 2.1 所示的包含相同支出的数据分开。

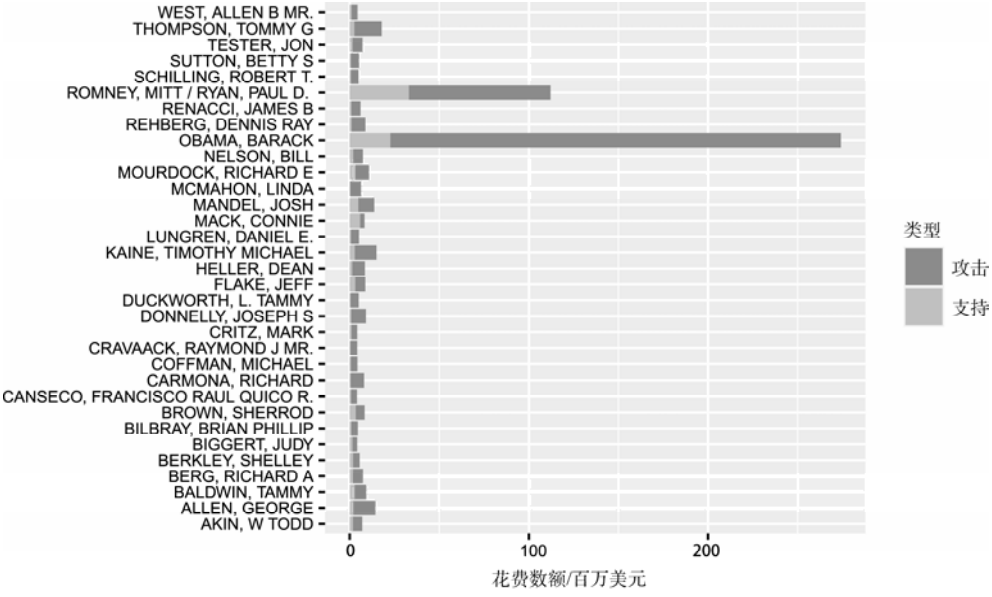


图 2.2　2012 年联邦选举周期大选阶段花在单个候选人身上的钱，按支出类型进行细分，只统计了开支超过 400 万美元的候选人

在这些选举中，大部分资金都花在了攻击其他候选人上，特别是奥巴马总统竞选的 2.74 亿美元中有 2.51 亿美元是用来攻击其他竞选候选人的。同样，米特 • 罗姆尼(Mitt Romney)竞选活动的大部分资金都不是用来支持自己的，但是罗姆尼在竞选活动中用于攻击其他候选人的支出百分比(70%)远低于奥巴马的这一比例(92%)。

图 2.1 和图 2.2 的区别在于，在后者中，我们使用了不同灰度将第三个变量(支出类型)引入图中，这使我们可以进行清晰的比较，从而极大地改变了我们可能从前一张图中得出的结论。特别是，在图 2.1 中，看起来奥巴马总统的竞选支出使罗姆尼的竞选支出相形见绌，然而事实恰恰相反。

2.1.1 这两组数据有区别吗？

由于奥巴马花费在攻击其他竞选人上的资金比罗姆尼的这笔花费要多得多，因此你可能从图 2.2 得出结论，共和党在这次选举周期中筹款更成功。在图 2.3 中，我们可以看出确实是这种情况，因为用于支持共和党(REP)候选人的资金要比支持民主党(DEM)候选人的资金多，用于攻击民主党候选人的资金也比攻击共和党人的多。从图 2.3 可看到，几乎所有的资金都花费在了民主党或共和党身上。

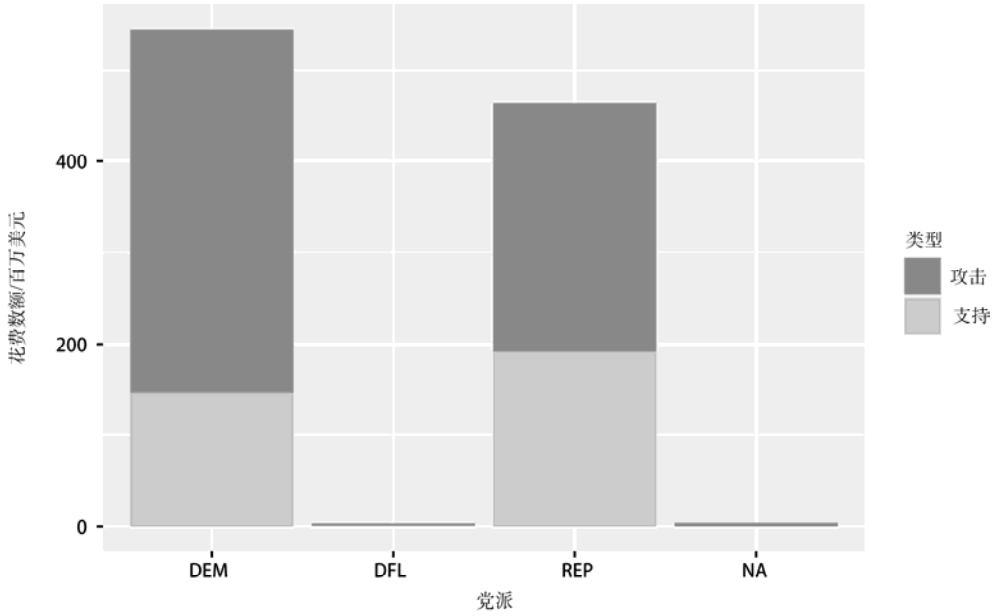


图 2.3 2012 年联邦选举期间花费在单个候选人身上的资金(按党派划分)

但是，花费在候选人身上的资金是否真的因党派而有所不同？这个问题有点棘手。如上所述，在这次选举周期中，总统选举主导了政治捐赠。罗姆尼在试图就任现任总统时面临严重的劣势。这种情况下，所要竞选的职位是一个容易迷惑人的变量。通过按竞选的职位进一步细分图 2.3 中的捐款，可以在图 2.4 看到，虽然花费了更多的资金来支持政府的三个参议院的所有

共和党候选人，但只有在总统选举中，才花费了更多的资金来攻击民主党候选人。实际上，用于攻击共和党众议院和参议院候选人的资金稍多。

请注意，图 2.3 和图 2.4 显示了相同的数据。在图 2.4 中，我们有一个附加变量，它为分析竞选财务的奥秘提供了重要线索。即使两张图都是“正确的”，但通过在图 2.4 中引入一个变量可使它能够比图 2.3 揭示更多信息。在本章中，我们逐步开发一个框架用于创建规范的数据图形。

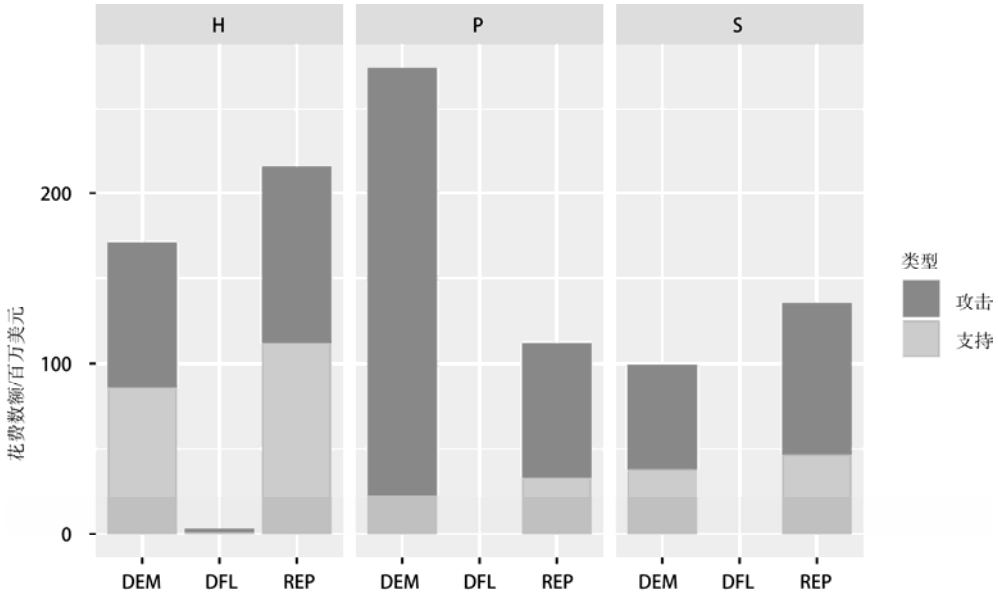


图 2.4　在 2012 年联邦选举周期的大选阶段，各政党单个候选人身上花费的资金，按竞选职位(众议院、总统或参议院)进行细分

2.1.2　图形变化

总统大选期间出现了一个问题，有人指出罗姆尼的竞选得到了一些富人的支持，而奥巴马的支持则来自整个经济领域的人群。如果这是真的，那么我们希望能看到两位候选人之间捐赠金额的分布有所不同。特别是，我们希望能在图 2.5 所示的直方图中看到这一点。该直方图总结了 2012 年超过 100 万笔捐款，他们捐给了支持每个候选人(支持奥巴马的是美国奥巴马和 2012 奥巴马胜利基金；支持罗姆尼的是 2012 罗姆尼胜利和罗姆尼担任总统基金)的两个主要委员会。我们确实在图 2.5 中看到了有关此结论的一些证据，奥巴马的确收到了一些较小的捐款，但该证据远不足以定论。首先，两个候选人都收到了很多小笔捐款，且只收到了几笔大额捐款。水平轴上的比例尺很难观察到实际发生的事情。第二，直方图很难并排放置进行比较。最后，我们将总统选举的两个阶段(即初选与大选)的所有捐赠集中在一起。

在图 2.6 中，我们通过以下方法纠正这些问题：①使用密度曲线代替直方图，以便可以

直接比较捐赠金额的分布；②在水平刻度线上绘制捐赠金额的对数以使用户关注重要的数据；③按选举阶段划分捐赠数据。图 2.6 可以使我们得出一些更加细微的结论。右面板支持以下主张：在初选阶段奥巴马的捐赠来自广大民众。在选举的这一阶段，奥巴马的捐款似乎大部分是小额捐款。但在大选阶段，对这些竞选活动的捐款的分布实际上没有差异。

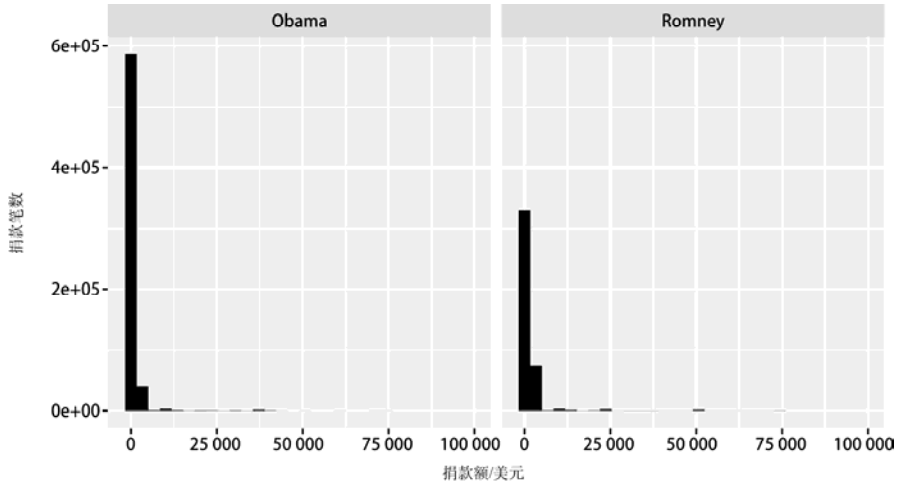


图 2.5 个人向 PAC 的捐款，PAC 用来支持 2012 年大选的两位主要总统候选人

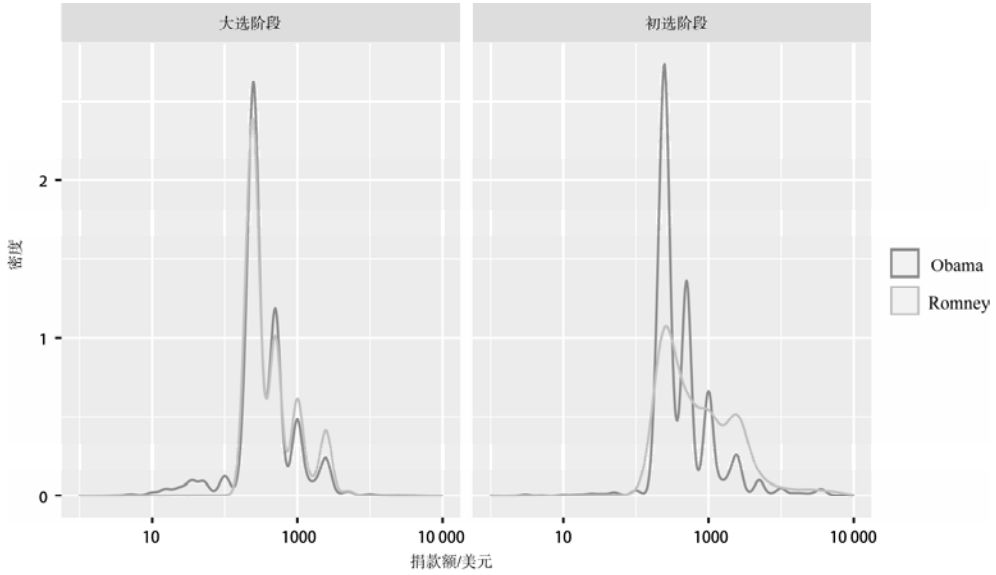


图 2.6 按选举阶段划分的个人向 PAC 的捐款，PAC 用来支持 2012 年大选的两位主要总统候选人

2.1.3 检查变量之间的关系

很自然，公民联合会决策提出的最大问题是关于资金在选举中的影响。如果竞选支出是无

限的，这是否意味着竞选支出最多的候选人将获得最多的选票？解决这个问题的一种方法是将每次选举中每个候选人所花费的资金与候选人获得的票数进行比较。统计人员想知道这两个数量之间的相关性，即当一个数量很高时，另一个数量也可能很高吗？

由于美国众议院的所有 435 名成员都是每两年选举一次，而各区的选民人数大致相同，因此众议院选举提供了一个很好的数据集进行这种比较。在图 2.7 中，我们给出了一个简单的散点图，将民主党候选人花费的美元数与该候选人在每次众议院选举中获得的票数相关联。

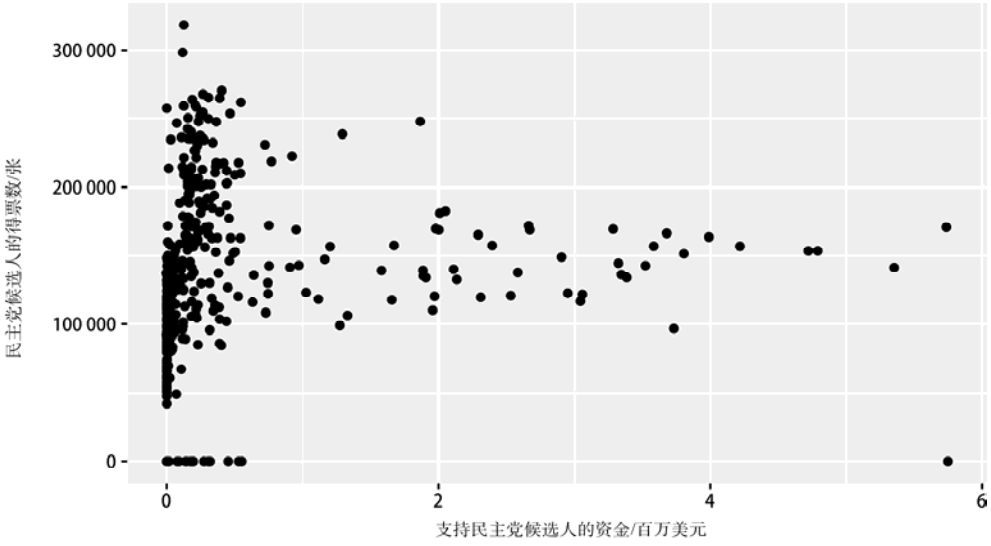


图 2.7　在 2012 年众议院选举中，民主党的花费与获得的选票数量之间关系的散点图

图 2.7 中描述的两个数据量之间的关系非常弱。从竞选支出中受益更多的候选人似乎没有获得更多选票。但是，图 2.7 中的比较具有误导性。在两个轴上，重要的不是数量，而是百分比。尽管国会每个选区的人口都相似，但并不相同，并且投票人数会因多种因素而有所不同。通过比较投票百分比，我们可以控制每个地区的投票人数规模。同样，关注花费的总金额而不是花费的百分比是没有意义的。在图 2.8 中，我们进行了相同的比较，但两个坐标轴均按百分比进行了缩放。

图 2.8 捕获了许多细微差别，这些细微差别在图 2.7 中看不到。首先，支持候选人花费的资金的百分比与他们获得的选票的百分比之间似乎存在正相关。但是，这种关系对实际发生选举的散列点的中心最为重要。在该区域之外，一位候选人赢得了 55%以上的选票。这种情况下，通常只花很少的钱。这些情况被认为是“安全的”众议院选举，你可以在图中看到这些点，因为它们中的大多数都接近 x = 0 或 x = 1 两条线，并且这些点非常小，例如，右下角是俄亥俄州第 8 区，当时的众议院议长约翰·博纳(John Boehner)竞选连任成功，花钱最多(超过 1100 万美元)的选举也发生在俄亥俄州。在第 16 区，共和党现任议员吉姆·雷纳奇(Jim Renacci)差点击败了本来是第十三区议员的民主党挑战者贝蒂·萨顿(Betty Sutton)，这场竞争是通过十年一次

的重新分区而实现的(请参见第 17 章)，在这次选举中，有 51.2%的资金用于支持萨顿，但她只赢得了 48.0%的选票。

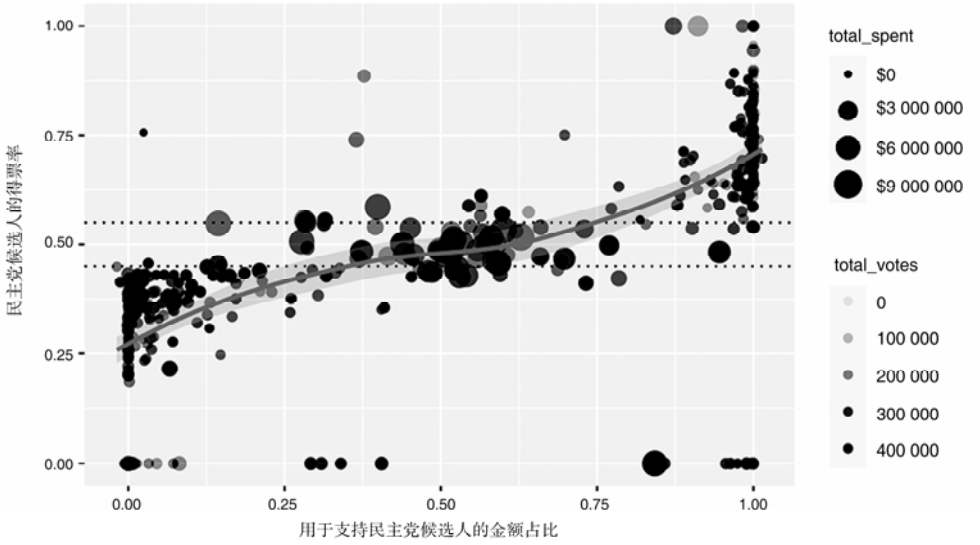


图 2.8 在 2012 年众议院选举中用于支持选举的美元比例与民主党获得的选票比例之间关系的散点图。每个点代表一个区。每个点的大小与该区选举的总支出成正比，每个点的透明度与该选区的总票数成正比

在图的中心，圆点较大，表示在这些有竞争的选举上花费了更多的资金。当然这是有道理的，因为为政治生涯而竞争的候选人更有可能积极筹款。然而，在有竞争的选举中，更多的财政支持与赢得更多选票两方面的相关性的证据显得很薄弱。

2.1.4 网络

并非所有变量之间的关系都可以通过散点图合理地表达。变量关联的另一种方式是网络形式(我们将在第 20 章详细讨论)。这种情况下，竞选资金具有网络结构，在这个网络中个人向委员会捐赠资金，然后委员会代表候选人花钱。尽管国家竞选资金网络过于复杂，无法在此处详细显示，但在图 2.9 中，我们显示了马萨诸塞州候选人的资金网络。

在图 2.9 中，我们看到从委员会支出中受益最大的两个竞选活动是共和党人米特·罗姆尼(Mitt Romney)和斯科特·布朗(Scott Brown)的竞选。这并不奇怪，因为罗姆尼竞选总统，并获得了共和党全国委员会的大量捐款，而布朗则在参议院中拥有自己的席位，以反对强大的挑战者伊丽莎白·沃伦(Elizabeth Warren)。两人都在选举中失败了，蓝色圆点表示马萨诸塞州的国会代表团，他们都是民主党人(请注意，本书为黑白印刷，后同)。

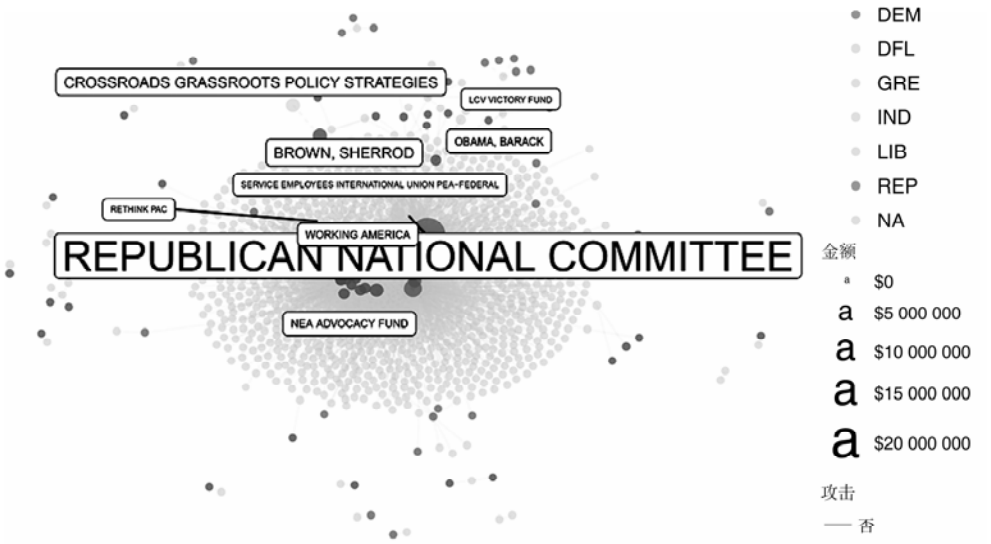


图 2.9　2012 年联邦选举中，马萨诸塞州候选人竞选资金网络。每条边代表 PAC 对候选人的贡献

2.2　组成数据图形

前《纽约时报》实习生和 FlowingData.com 的创建者邱南森(Nathan Yau)作了一个类比，即创建数据图形就像做饭一样：任何人都可以学习键入图形命令并在计算机上生成图形。同样，任何人都可以用微波炉加热食物。将高质量的可视化与简单的可视化区分开和将优秀厨师与新手区分开需要相同的要素：精通工具、关于其组成成分的知识，具有洞察力和创造力[Yau，2003]。在本节中，我们介绍一个源于科学研究的用于理解数据图形的框架。我们希望，通过吸收这些思想，可使你的数据图形更完美。

2.2.1　数据图形分类

文献 Yau(2013)中提出的分类体系提供了一种系统方式来思考数据图形如何传达特定信息，以及如何对其进行改进。图形的补充语法[Wilkinson et al., 2005]由哈德利・威克姆(Hadley Wickham)在 ggplot2 图形包[Wickham, 2016]中实现，该包使用的术语略有不同。为叙述清晰，我们将对 ggplot2 的讨论推迟到第 3 章。进一步用烹饪进行类比，就是你必须先学会品尝，然后才能学会烹饪。

在此框架中，可根据四个基本元素来理解数据图形：视觉线索、坐标系、刻度和背景。在接下来的内容中，我们阐述了这一愿景，并附加了一些其他内容(多面板图形、图层和动画)。本节可以使细心的读者具有系统地分解数据图形的能力，从而可以对其中的内容进行更严格的分析。

1. 视觉线索

视觉线索是可以吸引观众的图形元素。它们是数据图形的基本构建块，选择使用哪种视觉线索来表示哪些数量是数据图形编辑者要考虑的核心问题。表 2.1 标识了九种不同的视觉线索，表中列出了线索是用于编码数字值还是分类值。

表 2.1 视觉线索及其含义

视觉线索	变量类型	问题
位置	数字	在哪里与其他事物相关？
长度	数字	多大(在一个维度中)？
角度	数字	有多宽？与别的东西平行吗？
方向	数字	在什么斜率上？在一个时间序列中，是上升还是下降？
形状	分类	属于哪个组？
面积	数字	有多大(在二维空间中)？
体积	数字	有多大(在三维空间中)？
阴影	任一种	到什么程度？有多严重？
颜色	任一种	到什么程度？有多严重？

对图形感知的研究(可追溯到 20 世纪 80 年代中期)表明，人类对程度差异的感知能力按特定顺序下降[Cleveland and McGill，1984]。也就是说，人类非常擅长准确地感知位置的差异(例如，一根杆比另一根杆高多少)，但不擅长感知角度的差异。这就是许多人喜欢条形图而不是饼图的原因之一。我们对颜色差异的感知能力较差，这是许多数据科学家从热图(heat map)获取的信息较少的主要原因。

2. 坐标系

数据点如何组织？虽然我们可以使用任意个数的坐标系，但最常见的是三种：

直角坐标系 这是熟悉的(x, y)矩形坐标系，它具有两个相互垂直的坐标轴。

极坐标系 直角坐标系的径向模拟，该坐标系中的点由其半径ρ和角度θ标识。

地理坐标系 这是一个越来越重要的坐标系，在该坐标系中，我们可在地球的曲面上标记位置，但是我们将在平坦的二维平面中表示这些位置，第17章将讨论这种空间分析。

选择一个合理的坐标系对于准确地表示一个人的数据至关重要。例如，在直角坐标平面上显示诸如航空公司航线的空间数据可能导致真实情况的严重失真(请参见第 17.3.2 节)。

3. 刻度

刻度将值转换为视觉线索。刻度的选择通常至关重要，它的核心问题就是数据图形中的距离如何转化为数量上有意义的差值。每个坐标轴都可以有自己的刻度，通常我们有三种不同的选择。

数字 数字值通常以线性、对数或百分比刻度进行设置。注意，对数刻度不具有这样的性质，即坐标轴上任何位置上1cm的距离表示的数值相等。

分类　分类变量可能没有顺序(例如民主党、共和党或独立党派)，也可能是顺序变量(例如从不、以前或现在吸烟的人)。

时间　具有某些特殊性质的数字值。首先，由于存在日历的原因，它可以用一系列不同的单位进行划分(例如年、月、日等)。第二，它可以是周期性的，例如作为一个“环绕几圈”的刻度。时间的使用也是非常广泛的，我们必须仔细考虑它。

刻度容易引起误解，因为它有可能完全扭曲任何图形中数据点的相对位置。

4. 背景

数据图形的目的是帮助用户进行有意义的比较，但是一个不好的数据图形却可能适得其反：它可能将用户的注意力集中到毫无意义的人为组件上，或者忽略重要的具有外部知识的相关组件。背景可以以标题、字幕、轴标签、参考点或线的形式添加到数据图形中；这些标题或字幕解释要显示的内容，轴标签可以清楚地说明如何描述单位和比例，参考点或线能提供相关的外部信息。虽然我们应该避免在数据图形中添加过多注释，但必须提供适当的背景信息。

5. 多面板图形、图层和动画

创建数据图形的主要挑战之一就是如何将多变量信息浓缩成二维图像。虽然三维图像偶尔有用，但它们往往比其他图形都更令人困惑。相反，以下是将更多变量合并到二维数据图形中的三种常见方法。

多面板图形(small multiples)：面板图形也称为 facet，单个数据图形可由具有相同的基本图的几个多面板图形组成，每个小子图像中都有一个(离散)变量的变化。

图层：有时在现有数据图形上绘制一个新图层是合理的方法。这个新层可提供上下文或比较，但用户可以有效地解析多少层是有限制的。

动画：如果时间是附加变量，那么一幅动画有时可以有效地表达该变量蕴含的变化信息。当然，这在打印的页面上不起作用，用户不可能一次看到所有数据。

2.2.2　颜色

颜色是最不生动但最容易被误解和误用的视觉线索之一。在选择颜色时，有一些关键思想对数据科学家来说是非常重要的。

首先，如上所述，一种颜色及其单色相近色调是最难察觉的两个视觉线索。因此，尽管颜色和阴影对于包含级别较少的分类变量很有用，但并不是特别合适的表示数字变量的方式，特别是这些变量值的微小差异对区分事实很重要的情况。这意味着，虽然颜色可以在视觉上吸引人类，但通常不如我们认为的那样丰富。对于两个数值变量，很难想到颜色和阴影比位置更有用的例子。颜色最有效的地方就是，一旦用尽了两个位置信息，可以用颜色表示散点图上的第三个或第四个数字。

其次，大约有 8%的人(大多数是男性)患有某种形式的色盲。通常他们无法准确地看到颜色，尤其是不能区分红色和绿色。使问题更加复杂的是，这些人中的许多人都不知道自己是色盲。

因此，对于专业图形，需要仔细考虑使用哪种颜色。在 2015 年的一场比赛中，美国国家橄榄球联盟(NFL)就出现过这一问题：布法罗比尔(Buffalo Bills)穿着全红球衣，纽约喷气机队(New York Jets)穿着全绿球衣，这使色盲球迷无法区分一个球队与另一个球队！

专业提示 1：为防止遇到色盲的问题，请避免在数据图形中对比红色和绿色。你的绘图不需要看起来有圣诞节氛围！

值得庆幸的是，有了 ColorBrewer 网站(和 R 包)的创建者辛西娅·布鲁尔(Cynthia Brewer)的研究，我们摆脱了需要创建此类智能调色板的困境。布鲁尔为单个变量中的三种不同类型的数字数据创建了各种颜色的色盲安全调色板。

序列数据 数据的顺序只有一个方向。正整数是序列型的，因为它们只能向上增加，即它们不能越过 0(因此，如果将 0 编码为白色，则任何较深的灰色阴影都表示更大的数字)。

发散型数据 数据的顺序有两个方向。在选举预测中，我们通常会根据各州对总统的投票方式看到各州的颜色。由于红色与共和党人相关联，蓝色与民主党人相关联，因此以红色或蓝色为涂色的州在比例轴的两端。但是其中以任何一种方式出现的“摇摆状态”可能会显示为紫色、白色，或者是介于红色和蓝色之间的其他中性色(请参见图 2.10)。

RdBu (divergent)

图 2.10 发散的红蓝色调色板(可扫描封底二维码下载彩图，后同)

定性数据 数据没有顺序，我们只需要颜色即可区分不同的分类。

RColorBrewer 程序包提供了直接在 R 中使用这些调色板的功能。图 2.11 说明了内置在 RColorBrewer 中的不同调色板，包括序列、定性和发散等类型的调色板。

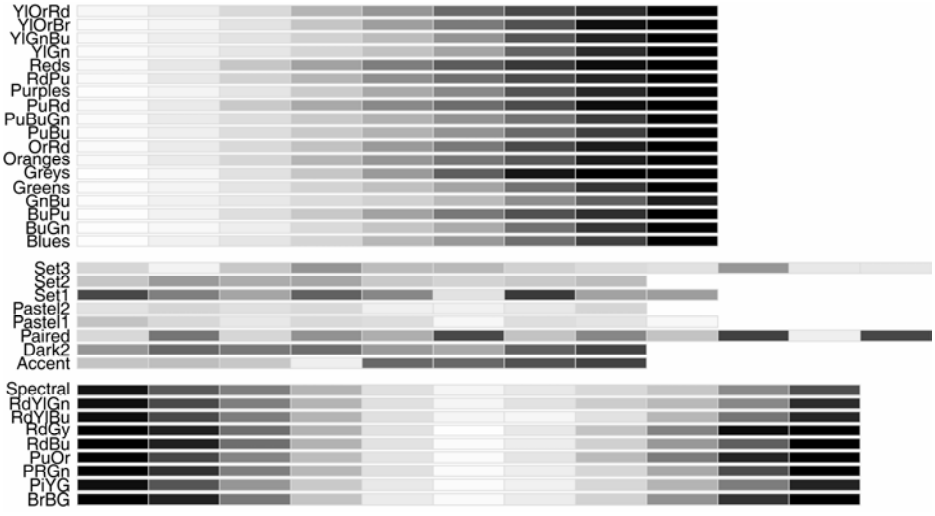


图 2.11 内置在 RColorBrewer 中的不同调色板

专业提示 2：花一些额外的时间使用精心设计的调色板。那些以色彩为主要工作的人可能会比自己的色彩选择更好。

viridis 程序包还提供了其他出色的具有不同感知的调色板。这些模板模仿 Python 的 matplotlib 绘图库中使用的模板。在 ggplot2 中还可以通过 scale_color_viridis()函数访问 viridis 的调色板。

2.2.3　剖析数据图形

稍加练习，你就可以学会根据上述分类法来剖析数据图形。例如，你的基本散点图可以使用具有线性刻度的笛卡儿平面中的位置来显示两个变量之间的关系。接下来，我们在一系列简单的数据图形中识别视觉线索、坐标系和刻度。

(1) 图 2.12 中的条形图显示了 1994—1995 年 SAT 考试的数学部分的平均分数(分数范围为 200～800)，图中的这些州至少有 2/3 的学生参加了 SAT 考试。

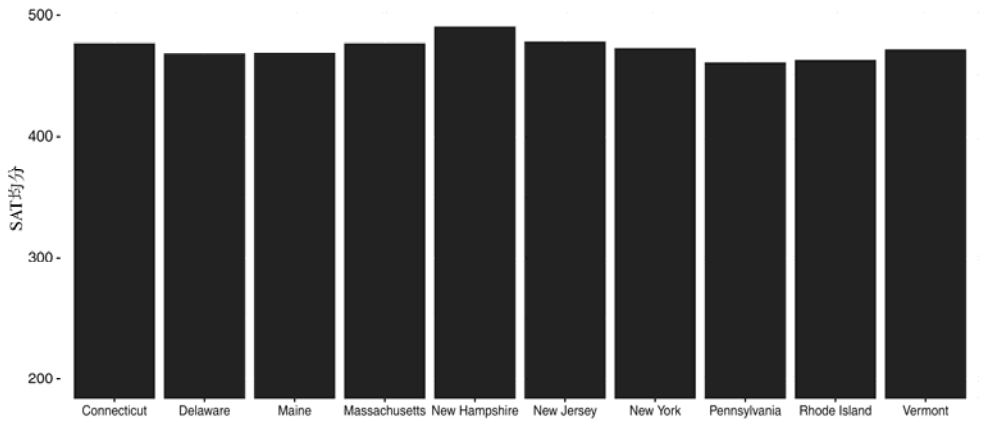


图 2.12　至少 2/3 的学生参加考试的州中平均 SAT 分数条形图

该图基于线性刻度，利用位置信息的视觉线索在纵轴上表示 SAT 数学分数。分类变量 state 分布在水平轴上。尽管 state 是按字母顺序排序的，但将变量 state 视为序列数据显然是不合适的，因为在 SAT 数学分数的所在环境中，排序没有意义。尽管如前所述，水平坐标没有意义，但这里还是采用笛卡儿坐标系。分数所在环境由轴标签和标题提供。还要注意，由于在每个部分的 SAT 上可能的最低分是 200，所以纵轴已被约束为从 200 开始。

(2) 接下来，我们考虑一个时间序列数据，该序列数据显示了 100 米男女自由泳比赛中世界纪录保持时间的情况。图 2.13 显示的时间是创造新纪录的年份的一个函数。

在某种程度上，这只是一个散点图，它在笛卡儿平面内使用纵轴和横轴上的位置分别表示游泳比赛时间和编年序列时间。纵轴上的数字刻度是线性的，以秒为单位，且横轴上的数字刻度也是线性的，以年为单位。但是，这里还有更多事情要做。颜色被用作区分分类变量 sex 的视觉线索。此外，由于这些点是通过线条连接的，因此使用变量 direction 表示比赛世界纪录发

生时间的情况(自由泳世界纪录只会变得更快，因此方向总是朝下)。甚至有人可能会争辩说可以使用角度变量 angle 来比较世界纪录在时间和/或性别方面的下降情况。实际上，这种情况下，形状变量 shape 也被用来区分 sex。

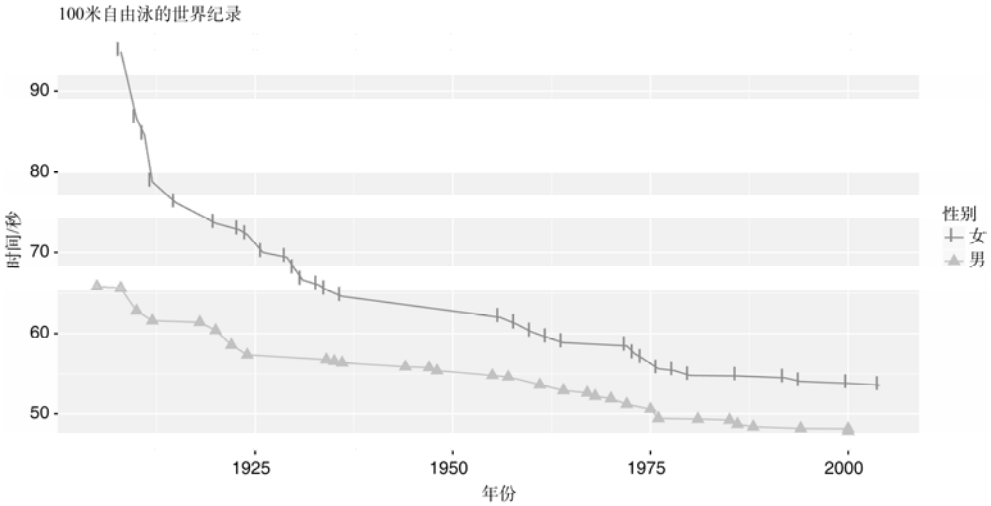


图 2.13 100 米自由泳的世界纪录时间散点图

(3) 接下来，我们在图 2.14 中显示了两个饼图，展示了在 HELP(健康评估与初级保健的联系)临床试验中受试者滥用的不同麻醉毒品[Samet et al., 2003]。它表明每个受试者都与一种主要麻醉毒品(酒精、可卡因或海洛因)有关。在右侧，我们看到有住所的参与者(没有在庇护所或街道过夜)的麻醉毒品分布相当均匀，而在左侧，我们看到了在外留宿一晚或多晚的无家可归者的麻醉毒品分布(更可能是以酒精为主要滥用麻醉毒品)。

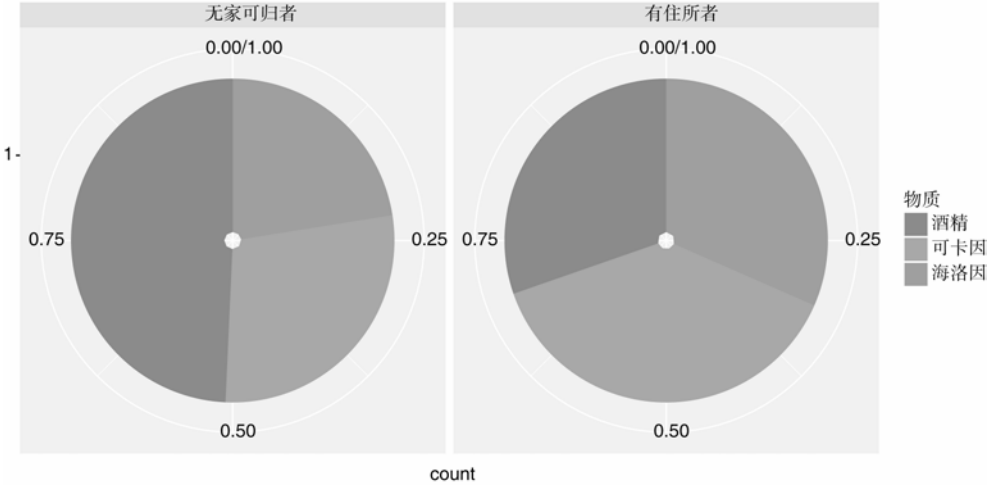


图 2.14 饼图显示了 HELP 研究中的参与者的滥用麻醉毒品的细目分类，即根据无家可归者的情况进行分类，和图 3.13 进行比较

该图形使用径向坐标系和 color 视觉线索来区分分类变量 substance 的三个级别。视觉线索 angle 用于量化使用每种麻醉毒品的患者比例的差异。你能否从图中准确地确定这些百分比？

专业提示 3：除非在多面板图形中，否则请勿使用饼图。

```
# A tibble: 3 x 3
  substance Homeless          Housed
  <fct>     <chr>             <chr>
1 alcohol   n = 103 (49.3%)   n = 74 (30.3%)
2 cocaine   n = 59 (28.2%)    n = 93 (38.1%)
3 heroin    n = 47 (22.5%)    n = 77 (31.6%)
```

在这个示例中，显示比例的简单表格比此类图形能更有效地传达真正的差异。请注意，该示例仅显示了六个数据点，因此任何图形都可能是没有必要的。

(4) 最后，在图 2.15 中，我们显示了一个 choropleth 图，该图显示了 2010 年马萨诸塞州人口普查区域的人口数据。

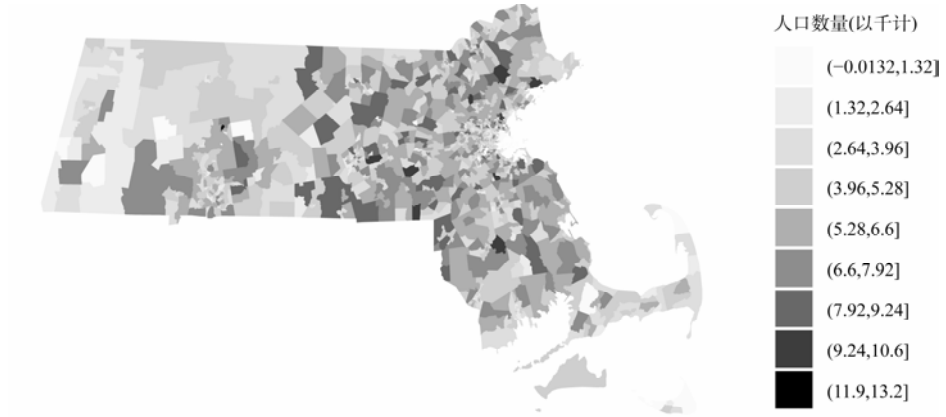


图 2.15　基于 2018 年美国社区调查的马萨诸塞州人口普查区的人口分布 choropleth 图

显然，这里使用的是地理坐标系，在纵轴和横轴上分别表示经度和纬度(该图不是投影图：有关投影系统的更多信息，请参见第 17 章)。利用阴影表示人口数量，但这里的规模更复杂。十个蓝色阴影被映射到人口普查区域人口数的十分位，并且由于这些区域中的人口分布是右偏的，因此每个阴影并不对应于相同宽度的人口数量范围，而是对应于在该人口数量范围内的相同数量的区域。这里的标题、字幕和图例提供了有用的背景信息。

2.3　数据图形的重要性：挑战者号

1986 年 1 月 27 日，曾为 NASA 航天飞机提供固体火箭发动机(SRM)的 Morton Thiokol 公司的工程师们建议 NASA 推迟挑战者号航天飞机的发射，因为他们担心第二天发射时的寒冷天

气会危害将火箭固定在一起的橡胶 O 型环的稳定性。工程师们提供了 13 张图表，这些图表在两小时的电话会议中进行了审查，会议参与者包括工程师、经理和 NASA 官员。由于缺乏有说服力的证据，工程师们的建议被否决了，发射工作如期进行。结果，O 型环完全按照工程师们担心的方式失败，发射 73 秒后挑战者爆炸，机上 7 名宇航员全部死亡[Tufte, 1997]。

除了悲惨的人员生命损失外，这一事件还给美国宇航局和美国太空计划造成了毁灭性打击。随之而来的麻烦包括：NASA 停滞了两年半，并成立了 Rogers 委员会以专门研究该灾难。显而易见的是，Morton Thiokol 公司的工程师非常正确地确定了温度与 O 型环损坏之间的关键因果关系。他们使用统计数据分析和科学的物理知识得出结论：橡胶 O 型环在低温下会变脆(这个因果关系由著名的物理学家和 Rogers 委员会成员 Richard Feynman 在听证会上使用一杯水和一些冰块进行了著名的演示[Tufte, 1997])。因此，工程师们可以使用他们的领域知识来识别关键的薄弱环节；这种情况下，领域知识就是火箭科学及其数据分析。

发射的失败及其可怕的后果就是一种说服力，他们只是没有以令人信服的方式将证据提供给最终决定继续发射的 NASA 官员。30 多年后，这一悲剧仍然警醒着世人。在这次发射之前，讨论是否启动发射的证据是以手写数据表(或“图表”)的形式展现的，这些数据都没有图形化的形式。在对该事件的深入评论中，Edward Tufte 创建了一个功能强大的散点图，类似于图 2.16 和图 2.17，该图可以从工程师们当时所拥有的数据中直接获得，但这张图比原始的数据要有效得多[Tufte, 1997]。

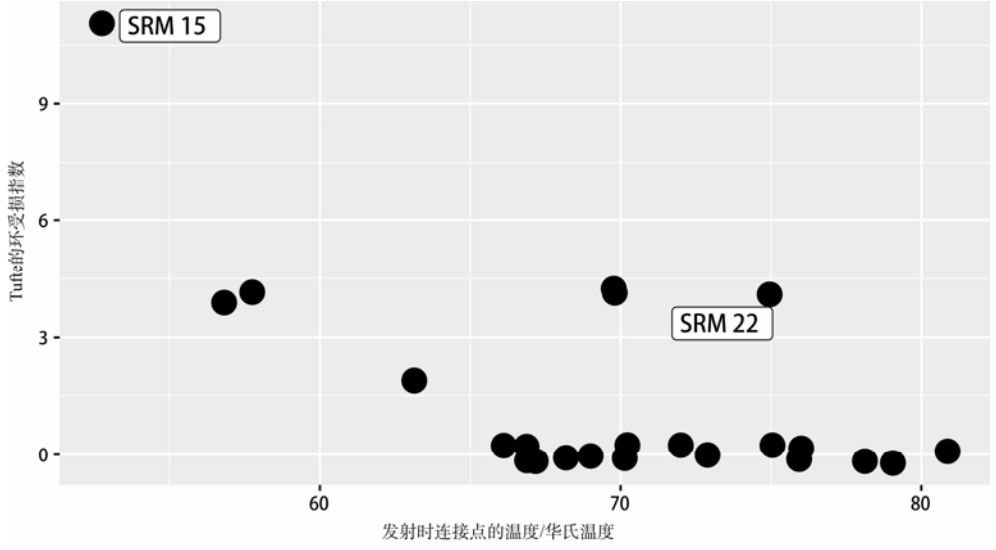


图 2.16 更平滑的散点图，显示了固体火箭发动机上温度和 O 型环损坏之间的关系。这些点是半透明的，因此较暗的点表示具有相同值的多个观测值

图 2.16 显示了环境温度与固体火箭发动机上的 O 型环损坏之间的明确关系。为了演示对 1986 年 1 月 27 日的预测温度进行的戏剧性外推，Tufte 在散点图中扩展了水平轴(图 2.17)以包括预测温度，巨大的间隙使这个外推问题显而易见。图 2.18 和图 2.19 显示了两个 Morton Thiokol

数据图形的重绘版[Tufte, 1997]。

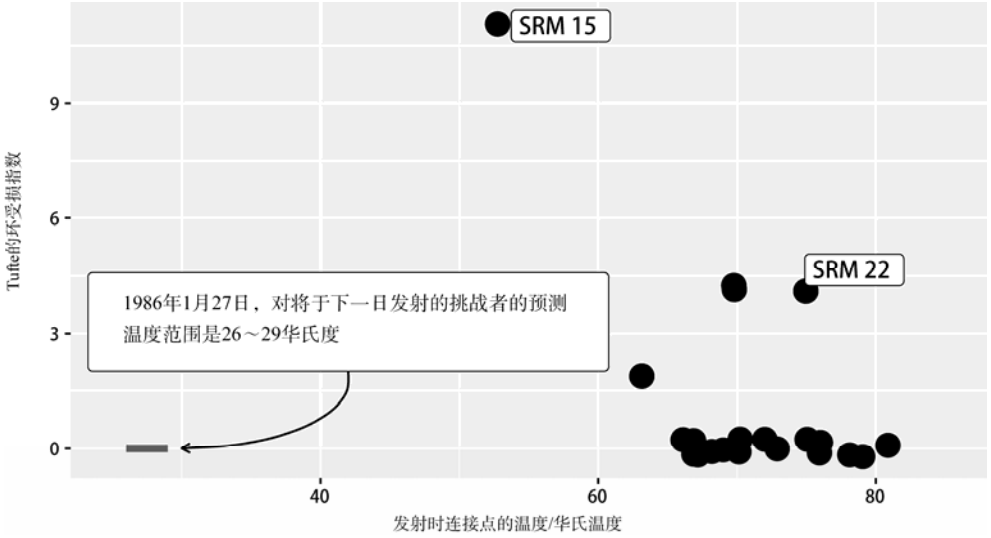


图 2.17　重新绘制的 Tufte 散点图，展示了温度与固体火箭发动机 O 型环损坏之间的关系

HISTORY OF O-RING TEMPERATURES (DEGREES - F)

MOTOR	MBT	AMB	O-RING	WIND
OM-1	68	36	47	10 MPH
OM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

图 2.18　Morton Thiokol 的工程师们在挑战者号发射前一晚的电话会议上向 NASA 提交的 13 张原始图表之一，这是一张数据密集型图表

Tufte 对工程师们的失败进行了全面批判[Tufte, 1997]，包括了许多对数据科学家具有指导意义的意见。

缺乏作者资格　任何图表上都没有标注姓名，这导致缺乏责任感。没有人愿意对任意一张图表中的数据承担责任。驳斥一群无名人士的论点要比驳斥一个或一群有名望的人士更容易。

单变量分析　工程师们提供了几张数据表，但实际上这些表基本上都是单变量的。也就是说，它们只提供了有关单个变量的数据，但没有说明两个变量之间的关系。请注意，尽管图 2.18 确实显示了两个不同变量的数据，但很难以表格形式看到两者之间的联系。由于此处的关键联系为温度高低和 O 型环是否会损坏，因此缺乏双变量分析可能是工程师演讲中最致命的漏洞。

轶事证据　由于样本量如此之小，因此轶事证据反驳起来尤其具有挑战性。这种情况下，基于两个观察结果进行了虚拟比较。虽然工程师们认为 SRM-15 在最冷的发射日期受到的损害

最大(请参见图 2.17)，但美国宇航局官员反驳说，SRM-22 在一个较温暖的发射日期受到的损害是第二大的。当在图 2.17 中考虑所有数据时，这些轶事证据就会支离破碎。很明显，SRM-22 是一个偏离一般模式的离群值，但工程师们从未在上下文中展示所有数据。

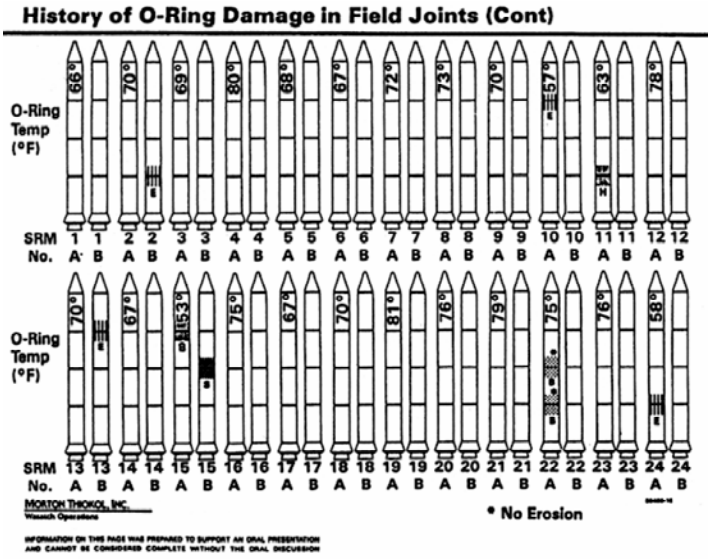


图 2.19 挑战者爆炸后在国会听证会上提供的证据

省略的数据 出于某些原因，工程师们选择不提供其他 22 次飞行的数据，这些数据占发射总量的 92%。这可能是由于时间限制，然而积累的证据的急剧减少在促成上述轶事证据方面发挥了重要作用。

混乱 毫无疑问，他们是在争分夺秒地工作，很可能是协同工作，工程师们并不总是清楚两种不同类型的损坏：腐蚀和窜气。如果不能明确界定这些术语，可能会影响美国宇航局官员的理解。

外推 最具说服力的是，未能包括完整数据的简单散点图，掩盖了“惊人的推断”[Tufte, 1997]，该推断是证明发射合理性所需的。底线是，预测的发射温度(26～29 华氏度)比以前任何时候都要低得多，因此，任何关于 O 型环损伤与温度的关系的模型都是没有测试过的。

专业提示 4：当观察结果较多时，数据图形往往比表格更具启示性，时刻注意考虑其他表达方式以改善沟通。

Tufte 指出了工程师们的首要问题是未能将数据与什么联系起来。某些数据可以被理解为与某事物相关的概念，这也许是统计推理的基础和决定性特征。我们将沿着这条线索贯穿全书。

专业提示 5：始终确保图形显示利用适当的轴标签、附加文本说明和标题进行清晰的描述。

在本章中，我们将这个悲剧事件作为对数据可视化进行深入研究的动机。这为应用数据科学家说明了一个重要的真理：只是正确是不够的，你必须要有足够的说服力。请注意，图 2.19

包含的数据与图 2.17 中的数据相同，但格式表现信息更明显。对大多数人来说，图形化的解释是非常有说服力的。因此，要成为一名成功的数据分析师，必须掌握数据可视化的基础知识。

2.4　创建有效的演示

对数据科学家来说，进行有效的演示是一项重要技能。无论这些演示是在学术会议上、教室里、会议室里，还是在舞台上，与听众交流的能力都有不可估量的价值。虽然有些人在聚光灯下会表现得很自然、很舒畅，但每个人都可以提高演示的质量。

下面列出一些一般性的建议[Ludwig, 2012]。

规划时间　通常你只有很少的时间来演讲，并且只有额外的几分钟来回答问题。如果你的发言时间太短或太长，会让你显得毫无准备。排练几次，以便更好地把握时机。还要注意，你在实际演讲时可能比在排练时说话更快。加快语速并不是一个非常差的策略，最好提前删减材料。你可能很难在 *x* 分钟内看完 *x* 张幻灯片。

专业提示 6：通过加快语速来加快进度并不是一个好策略，你最好提前删减材料，或者转到关键的幻灯片或结论。

每张幻灯片上不要写太多内容　你不能指望别人去阅读你的幻灯片，因为如果观众在阅读你的幻灯片，那么他们就不会听你的。你希望幻灯片能为所说的内容提供视觉上的提示，而不是替代你所说的话。把注意力集中在图形显示和思想要点列表上。

把问题放在上下文中　记住，在大多数情况下听众对你的主题知之甚少或一无所知。失去听众最简单的方法就是直接深入到需要具备一定领域知识的技术细节中。在演讲开始前花几分钟时间向听众介绍主题的背景知识，并对你正在学习的内容介绍一些动机。

大声说清楚　记住，在大多数情况下你对话题比房间里的任何人都了解得多，所以要自信地说话和表达！

讲一个故事，但不一定是整个故事　你期望能在 *x* 分钟内把你所知道的关于话题的一切都告诉听众是不现实的。你应该努力以清晰的方式传达主要思想，但不要停留在细节上。如果听众能够理解你研究的问题是什么，你是如何解决的，以及你的发现有什么样的含义，你的演讲就会成功。

2.5　更广阔的数据可视化世界

到目前为止，我们对数据可视化的讨论仅限于静态的二维数据图形。然而，还有其他许多方法可以可视化数据。虽然第 3 章侧重于静态数据图形，但第 14 章介绍了几种用于交互式数据可视化的高级工具。更广泛地说，视觉分析领域关注的是交互式视觉界面构建所隐含的科学，目的是增强人们的数据推理能力。最后，我们还形成了数据艺术。

总之，我们拥有数据艺术(data art)。你可以用数据做很多事情。一方面，你可能会专注于

预测特定响应变量的结果。这种情况下，你的目标是非常明确的，你的成功是可以量化的。在另一方面，是被称为数据艺术的项目，在这些项目中，对数据所能做的事情的意义是难以理解的，但能以一种新方式洞察数据，这种体验本身就是有意义的。

考虑一下 Memo Akten 和 Quayola 的表单(Forms)，它们的灵感来自英联邦运动会运动员的身体移动。通过视频分析，这些移动被转换成三维数字对象，如图 2.20 所示。请注意左上角的图像是如何在一个游泳运动员跳水后浮出水面时荡起的。如果把这些看成电影，那么表单就是数据艺术中引人注目的一幕。

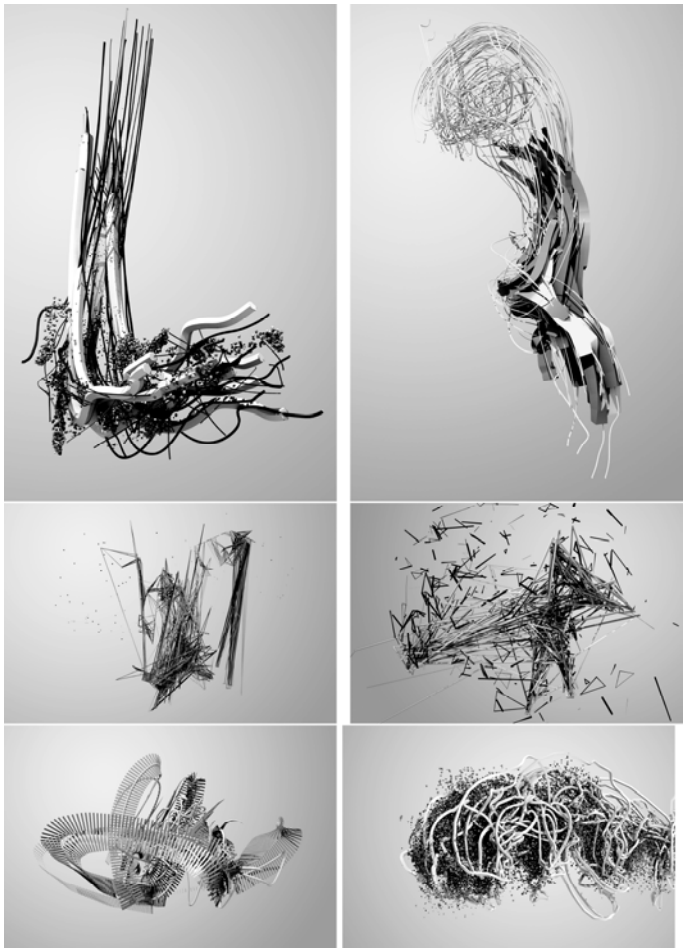


图 2.20 来自表单的静态图像，由 Memo Akten 和 Quayola 提供。每张图片都代表一个运动员在英联邦运动会上所做的移动，但都被重新想象成一组移动的三维数字对象

专业提示 7：查看 Vimeo 上备忘录的表单(流程)。

成功的数据艺术项目需要艺术天赋和技术能力。“我们面前是推销员的房子”是一个现场版的、不断更新的对在线市场 eBay 的探索。这套装置是由统计学家 Mark Hansen 和数码艺术

家 Jer Thorpe 创作的，当你进入 eBay 园区时，它被投射到一个大屏幕上。

专业提示 8：在我们面前的是推销员的房子——Vimeo 上 blprnt 的三个场景。

屏幕首先显示 Arthur Miller 的经典戏剧《推销员之死》，然后“阅读”第 1 章的文本内容。在此过程中，会有多个名词(例如冰箱、椅子、床、奖杯等)是从文本中提取出来的。对于连续场景中的每一个场景，屏幕显示然后转移到一个地理位置，eBay 在该位置正出售文本描述中与这个名词相关的东西，其中详细列举了价格和拍卖信息(请注意，这些描述并不总是完美的。在视频中，搜索“冰箱”会发现前芝加哥熊队防守端成员 William "Refrigerator" Perry 的 T 恤)。

接下来，选择一个出售此类物品的城市，收集附近出售的美国文学经典书籍。一个商品被选中，通过“阅读”那本书的第一页可以循环回到开始处。这个过程无限期地继续下去。在描述展览时，Hansen 谈到了“一个数据集读取另一个数据集”，正是数据和文学的相互作用使得这些数据艺术项目非常强大。

最后，我们考虑 Mark Hansen 的另一个合作，这次是 Ben Rubin 和 Michele Gorman。在《莎士比亚机器》中，37 个数字液晶显示刀片排列成一个圆圈，每个刀片对应莎士比亚的一个戏剧。每个刀刃上都显示从这些剧本的文本中挑选出来的字符模式。首先显示连字号对，然后找出布尔对，再找出修饰名词的冠词和形容词(如“the holy father”)。以这种方式，莎士比亚的艺术杰作被粉碎成公式化模块。在第 19 章中，我们将学习如何使用正则表达式查找《莎士比亚机器》的数据。

专业提示 9：观看 Vimeo 上 Ben Rubin、Mark Hansen、Michele Gorman 演出的《莎士比亚机器》。

2.6　扩展资源

与数据可视化相关的问题贯穿全文，它们将是第 3 章、第 14 章和第 17 章的重点。

没有阅读 Tufte 的《定量信息的视觉展示》[Tufte, 2001]，数据图形教育是不完整的，该著作也包含了 John Snow 的霍乱地图的描述(请参见第 17 章)。关于挑战者号事件的完整描述，请参见[Tufte, 1997]。Tufte 还出版了另外两本具有里程碑意义的著作[Tufte, 1990, 2006]，并对 PowerPoint 的缺点进行了理性评论[Tufte, 2003]。Cleveland 和 McGill 的工作为 Yau 的分类学提供了基础[Yau, 2013]。Yau 的文章[Yau, 2011]列举了许多发人深省的数据可视化例子，尤其是数据艺术。图形语法首先由 Wilkinson 描述[Wilkinson et al. 2005]， Wickham 基于这个语法实现了 ggplot2[Wickham 2016]。

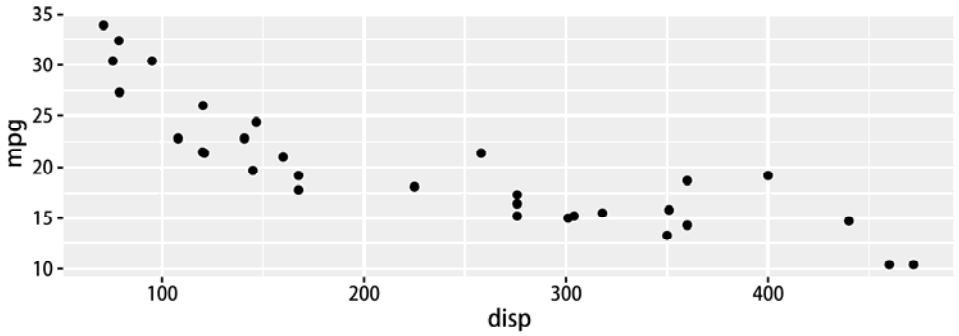
许多重要的数据图形是由 Tukey[Tukey, 1990]开发的。Andrew Gelman[Gelman et al. 2002]也在统计期刊上写过关于数据图形的具有说服力的文章。Gelman 讨论了一组典型的数据图形，Tufte 也对这些数据图形建议修改。Nolan 和 Perrett 讨论了数据可视化任务和评分的标准[Nolan and Perrett, 2016]。Steven J. Murdoch 创建了一些 R 函数来绘制 Tufte 在[Tufte, 2001]中描述的修

改后的图表，这些也出现在 ggthemes 包[Arnold, 2019a]中。

Cynthia Brewer 的调色板可在 http://colorbrewer2.org 下载，并可通过 R 的 RColorBrewer 包调用；她的工作在[Brewer (1994), Brewer (1999)]中有更详细的描述。viridis[Garnier，2018a]和 viridisLite[Garnier，2018b]程序包为 R 提供了类似 matplotlib 的调色板。Ram 和 Wickham 创造了异想天开的调色板[Ram and Wickham, 2018]，令人想起 Wes Anderson 与众不同的电影。Technically Speaking 是国家自然科学基金会资助的一个项目，提供了一些演讲建议，并为学生提供了一些教学视频[Ludwig, 2012]。

2.7 练习题

问题 1(易)：考虑以下数据图形。



如果汽车有自动变速器，am 变量的值为 0；如果汽车有手动变速器，am 变量的值为 1。如何根据变速箱类型区分图中的车辆？

问题 2(中)：从 https://www.edwardtufte.com/tufte 选择一个科学笔记本条目(如 Making better inferences from statistical graphics)。写一篇关于这个条目所说明的图形原理的简短个人思考。

问题 3(中)：找出最近两年在报纸或互联网上发表的两张图表。

a. 确定一个你觉得有吸引力的图形显示。该显示的哪些方面表现得很好，这些好的方面与本章中确立的原则有何关系？包括显示内容的屏幕截图以及你的解决方案。

b. 确定一个你觉得不那么引人注目的图形显示。该显示的哪些方面表现不好？有没有改进显示效果的方法？给出解决方案以及显示内容的屏幕截图。

问题 4(中)：在同行评议的期刊上找到过去两年发表的两篇科学论文(Nature 和 Science 是不错的选择)。

a. 确定一个你觉得有吸引力的图形显示。该显示的哪些方面表现得很好，以及这些方面与本章中确立的原则有何关系？包括你的解决方案以及显示内容的屏幕截图。

b. 确定一个你觉得不那么引人注目的图形显示。该显示的哪些方面不好？有没有改进显示的方法？包括你的解决方案及显示内容的屏幕截图。

问题 5(中)：考虑一下《纽约时报》在 http://www.nytimes.com/2012/04/15/sunday-review/

coming-soon-taxmageddon.html 上的一篇文章，分析与 Taxmageddon 相关的两个图形。第一个是“谁的税率涨或跌”，第二个是“谁从减税中获益最多”。

a. 仔细查看这两个图形。讨论一下它们传达了什么。图形说明了什么故事？
b. 根据本章描述的分类法评估两个图形。比例合适吗？是否一致？清楚地贴上了标签？变量维度是否超过数据维度？
c. 这些图形是否有误导性？

问题 6(中)：考虑关于节育方法的数据图形(http://tinyurl.com/nytimes-unplanned)。

a. 在每个图形的 y 轴上显示的数量是多少？
b. 列出数据图形中显示的变量，以及每个变量的单位和一些典型值。
c. 列出数据图形中使用的视觉线索，并解释每个视觉线索如何链接到每个变量。
d. 仔细检查图形。用文字描述一下你认为数据图形传达了什么信息。不要只是汇总数据，在问题的背景下解释数据的含义(注意，信息对人类来说是有意义的，它与数据不是一回事)。

2.8　附加练习

可从 https://mdsr-book.github.io/mdsr2e/ch-vizI.html#datavizI-online-exercises 获得。