Data Science and Big Data Analytics

Discovering, Analyzing, Visualizing and Presenting Data

EMC²

数据科学与大数据分析

数据的发现 分析 可视化与表示

[美] EMC Education Services 著
曹逾 刘文苗 李枫林 译
孙宇熙 主审

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

Data Science and
Big Data Analytics

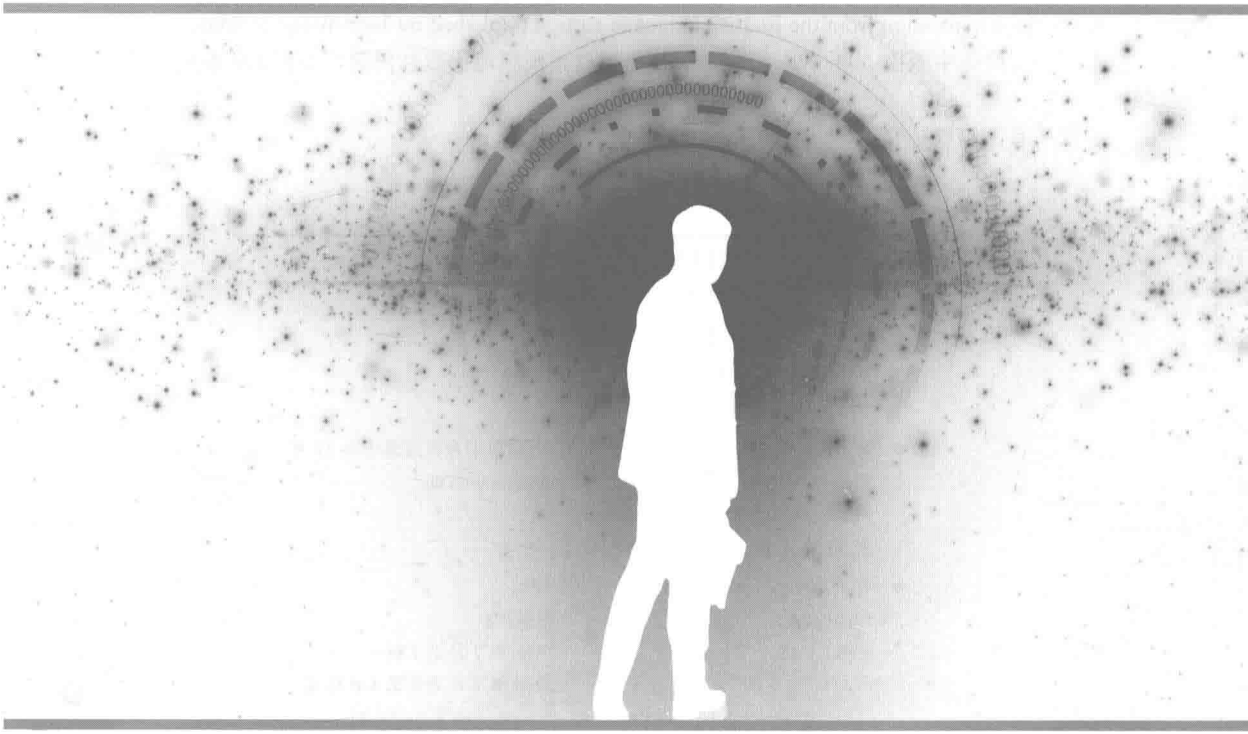
Discovering, Analyzing, Visualizing and Presenting Data

EMC²

数据科学与大数据分析

数据的发现 分析 可视化与表示

[美] EMC Education Services 著
曹逾 刘文苗 李枫林 译
孙宇熙 主审



人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

数据科学与大数据分析 ：数据的发现 分析 可视化与表示 / 美国EMC教育服务团队著 ；曹逾，刘文苗，李枫林译. -- 北京 ：人民邮电出版社，2016.7
ISBN 978-7-115-41637-7

Ⅰ. ①数… Ⅱ. ①美… ②曹… ③刘… ④李… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第019027号

版权声明

◆ 著 [美] EMC Education Services
译 曹 逾 刘文苗 李枫林
主 审 孙宇熙
责任编辑 傅道坤
责任印制 焦志炜

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：800×1 000 1/16
印张：23.5 彩插：8
字数：515 千字 2016 年 7 月第 1 版
印数：1 – 3 000 册 2016 年 7 月北京第 1 次印刷

著作权合同登记号 图字：01-2015-2823 号

定价：69.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316
反盗版热线：(010)81055315

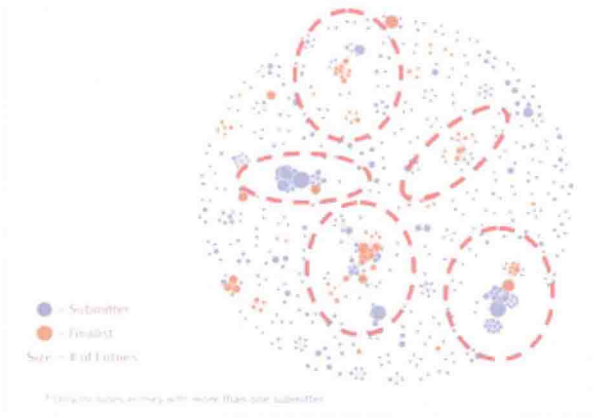


图 2.10　创新想法提交者和入围者的社交图谱 [27] 可视化

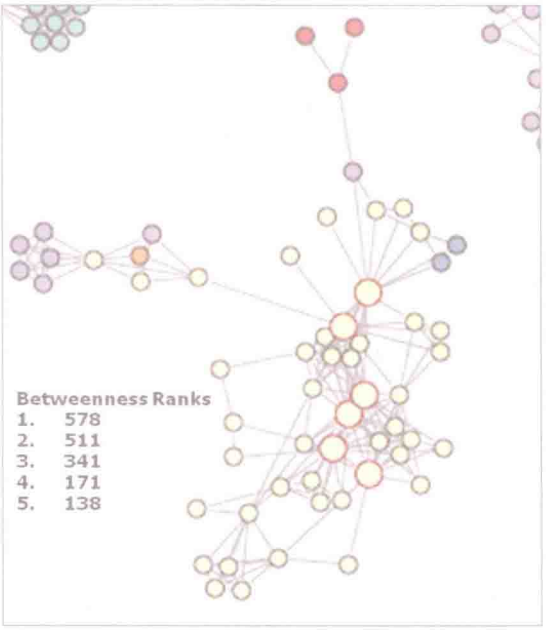


图 2.11　最具影响力创新者的社交图谱可视化

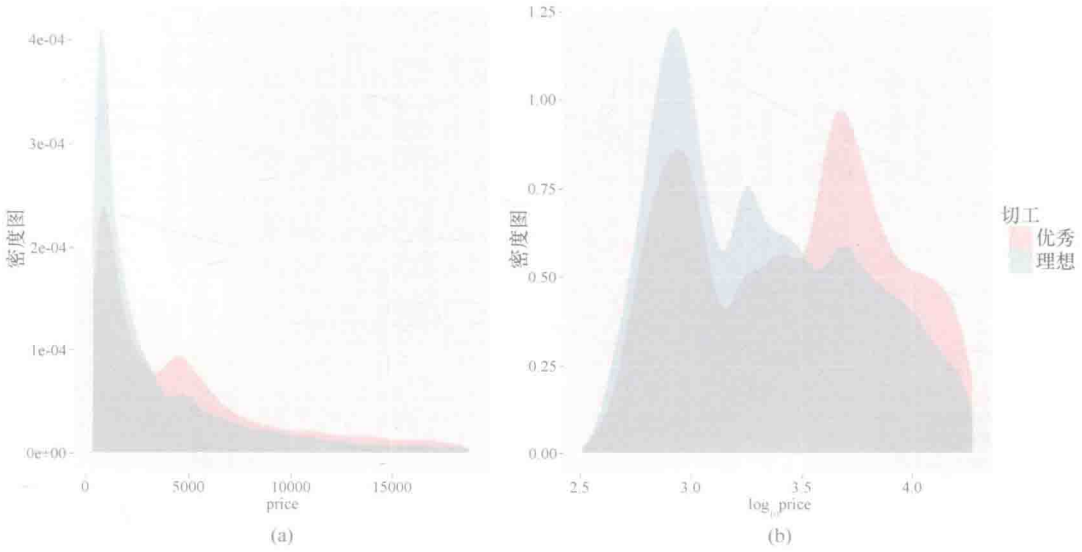


图 3.12　（a）钻石价格密度图和（b）钻石价格对数密度图

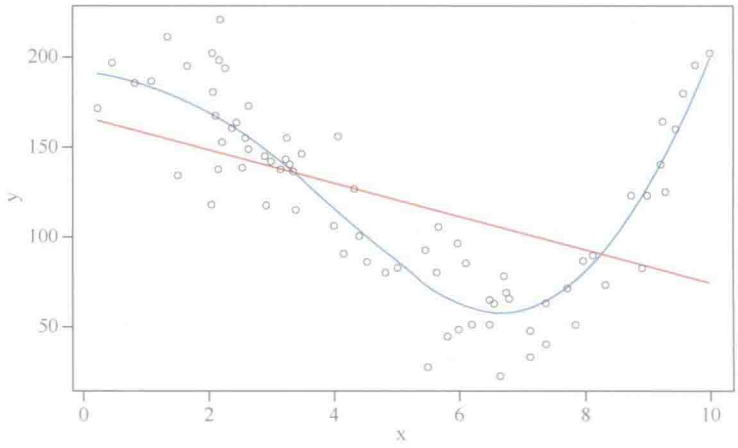


图 3.13　用回归研究二个变量

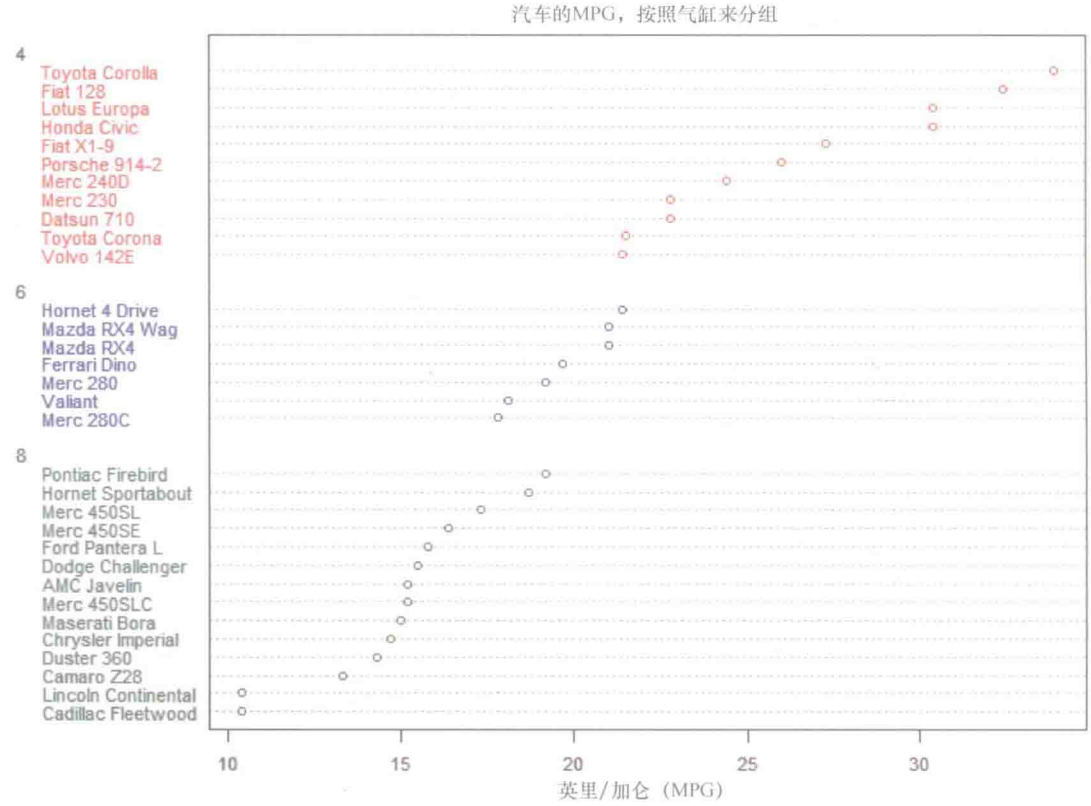


图 3.14　点图用于可视化多个变量

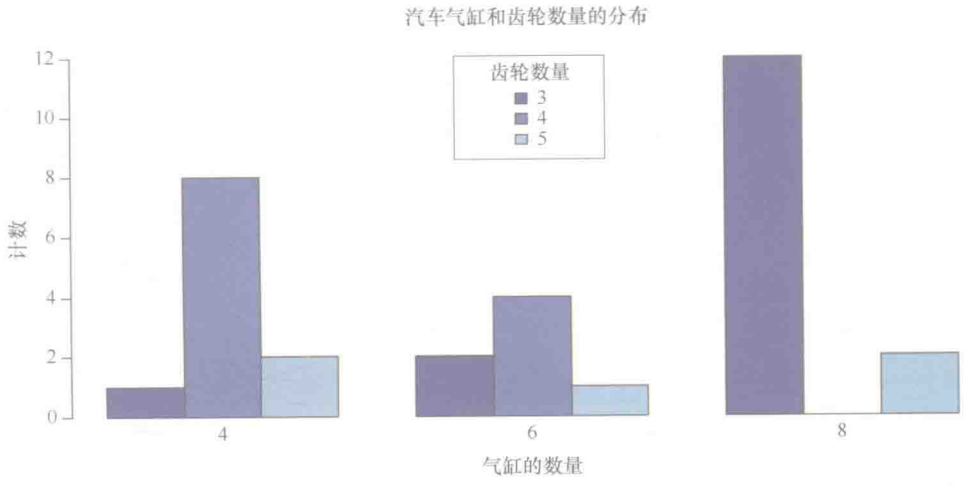


图 3.15　通过条状图来可视化多个变量

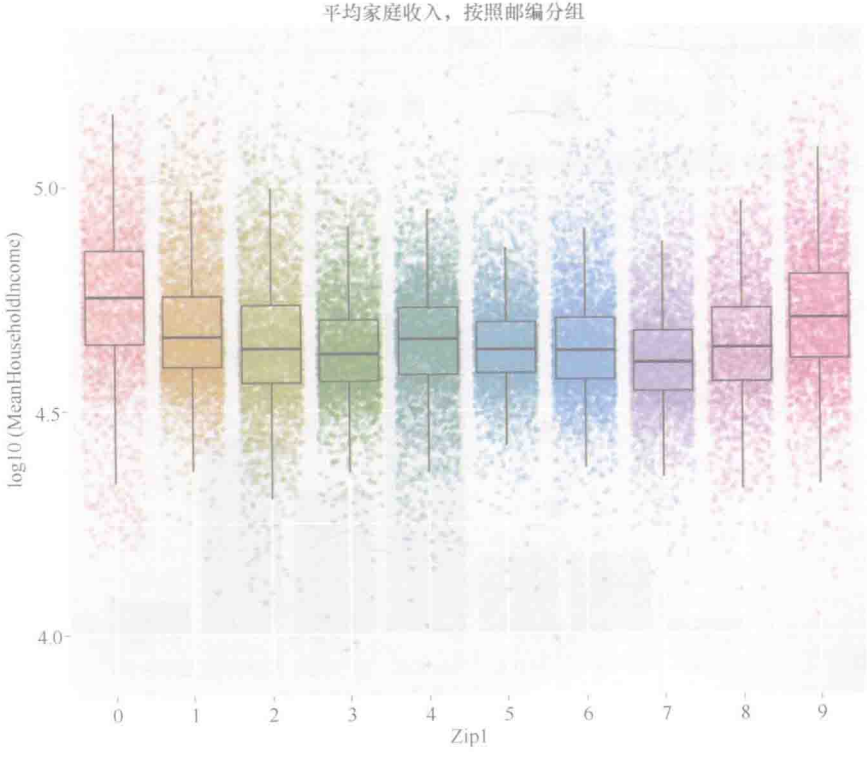


图 3.16　箱线图表示中位数家庭收入和地理区域的关系

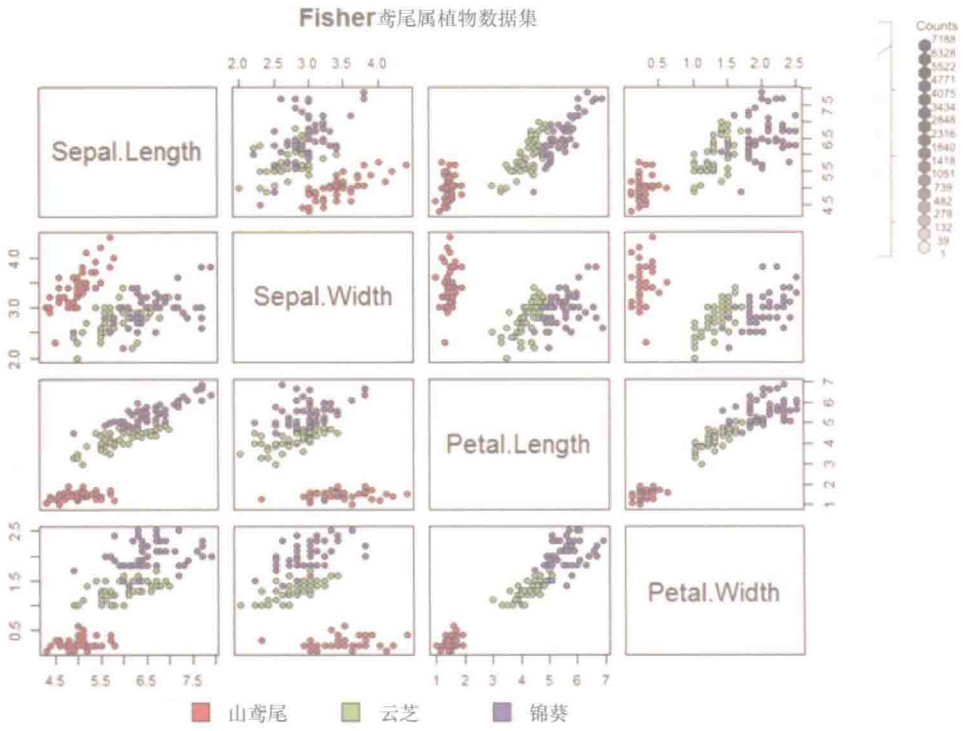


图 3.18 Fisher[13] 鸢尾属植物数据集的散点图矩阵

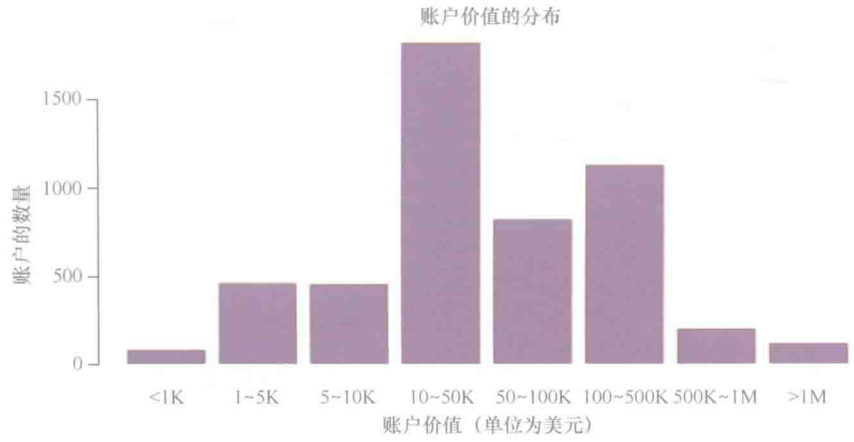


图 3.21 直方图更适合利益相关者查看

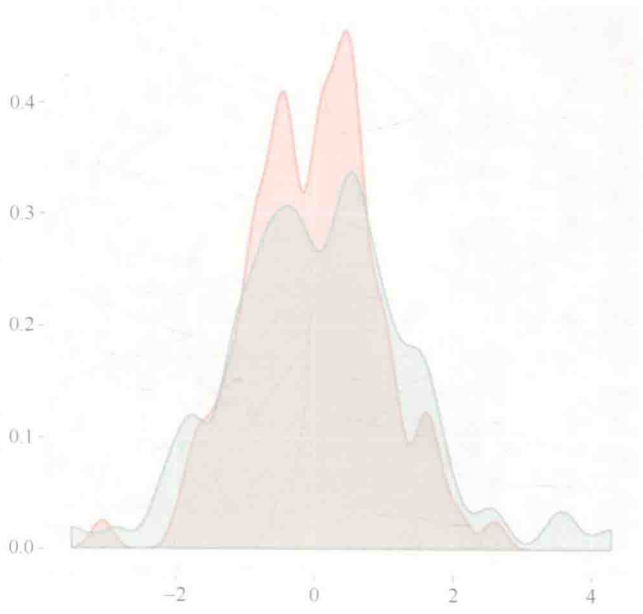


图 3.22　两个样本数据的分布

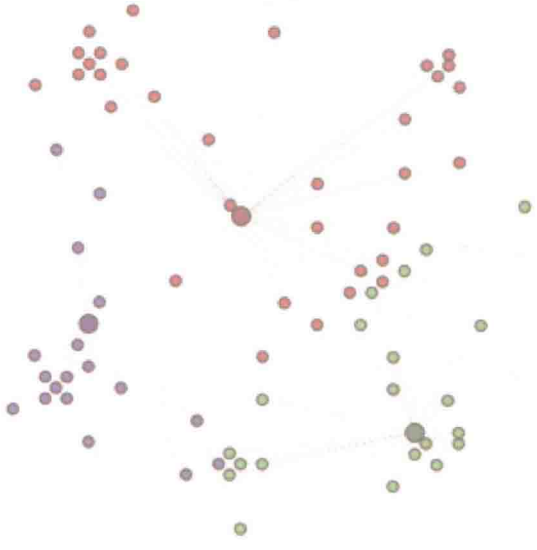


图 4.1　当 k=3 时可能的 k 均值聚类簇

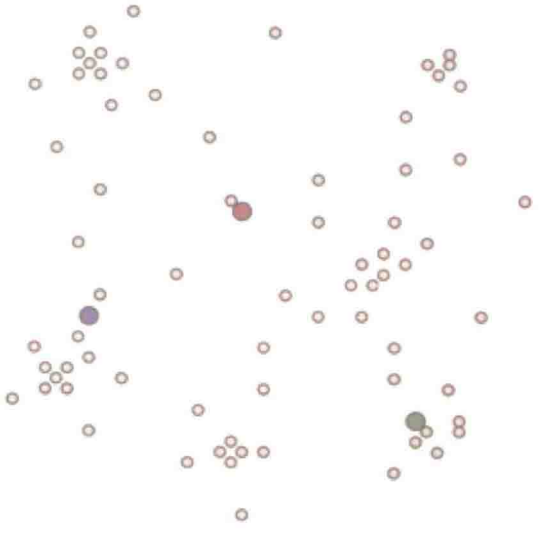


图 4.2　质心的初始起点

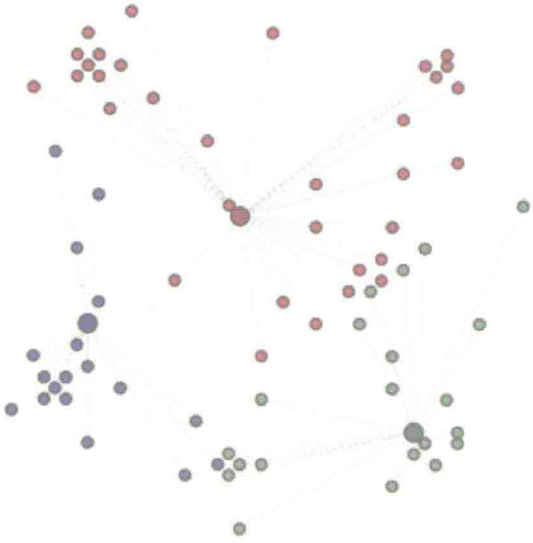


图 4.3　被关联到最近质心的点

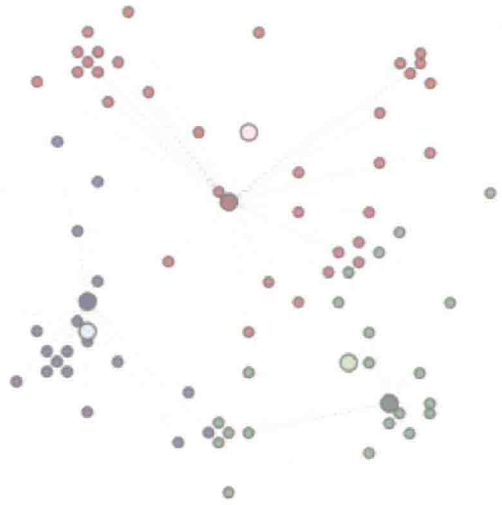


图 4.4　计算每个簇的质心

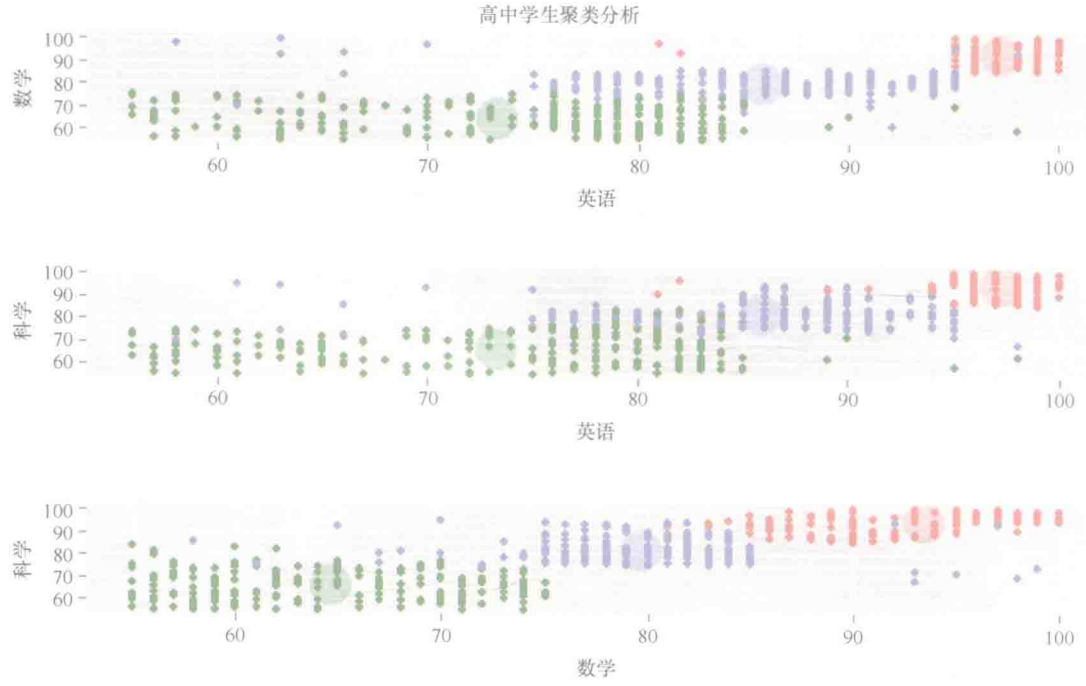


图 4.6　识别出的学生聚类图

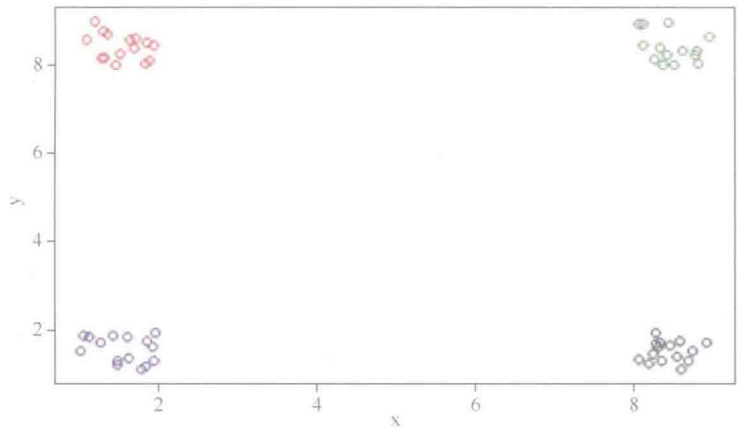


图 4.7　高度分离的簇示例

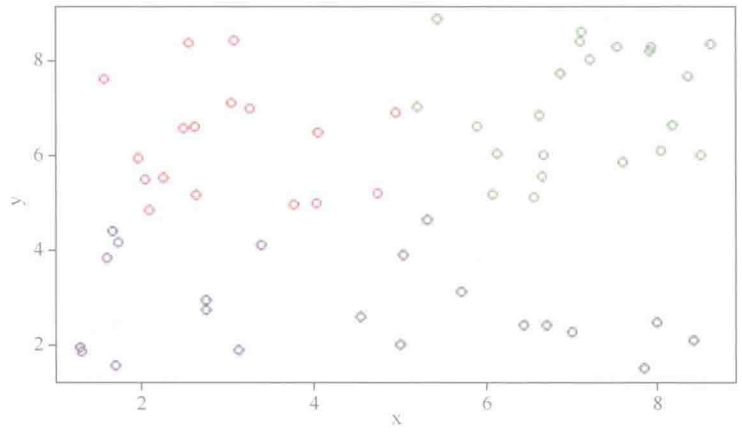


图 4.8　界限不那么明显的簇示例

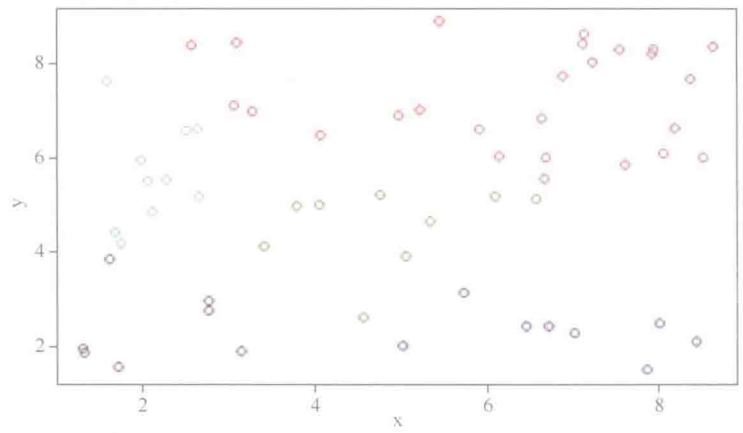


图 4.9　把图 4.8 的点划分成 6 个簇

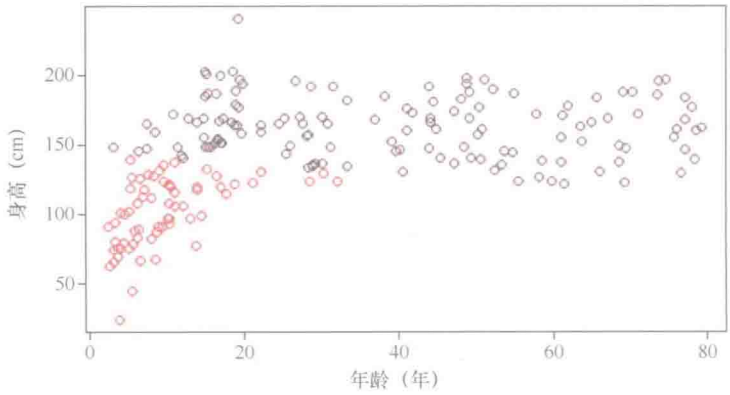


图 4.11　用厘米表示身高的聚类簇

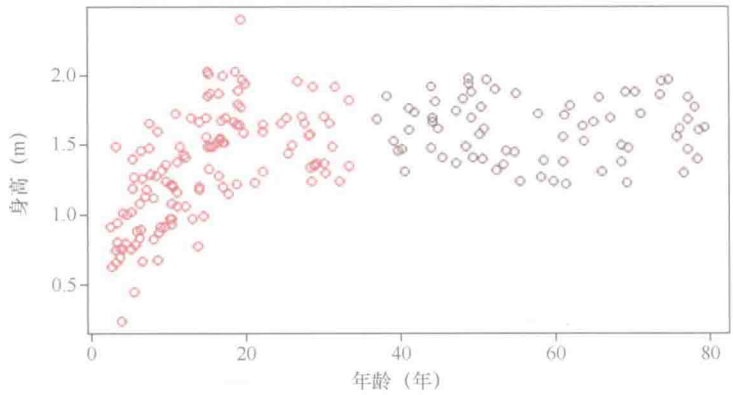


图 4.12　用米表示身高的聚类簇

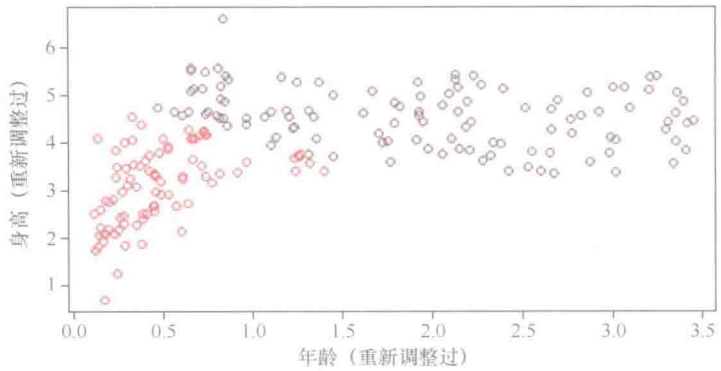


图 4.13　重新调整属性后的聚类簇

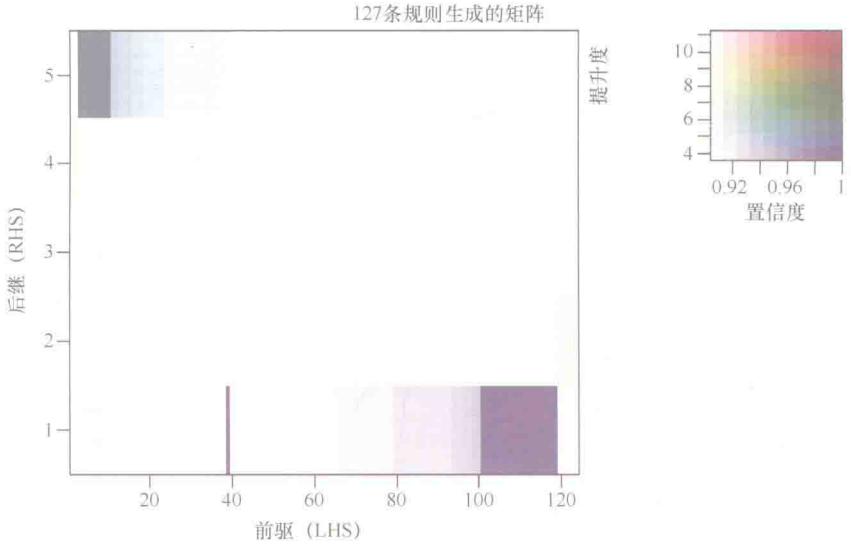


图 5.5　LHS 和 RHS 中的矩阵可视化，使用提升度和置信度进行了填色

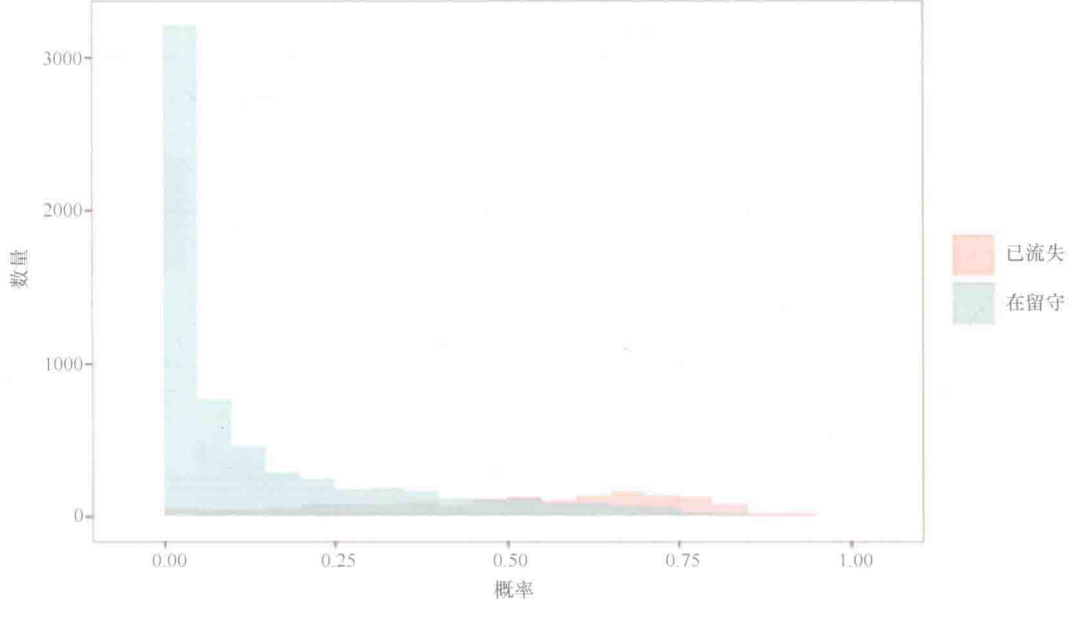


图 6.17　用户数与估计的流失率的对比

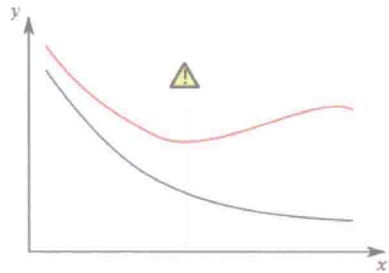


图 7.7　一个过度拟合的模型，它在训练集上运行良好，但是对未见过的数据则表现糟糕

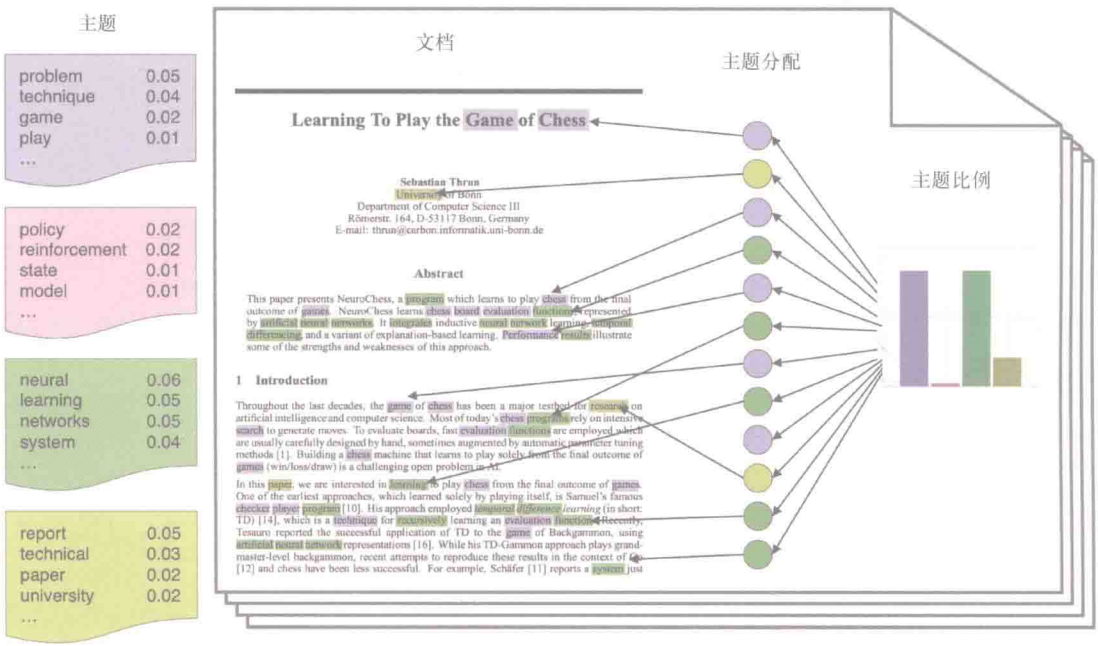


图 9.4　LDA 背后的直觉

图 9.5　10 个主题在 Cora 数据集中 9 篇科学文档上的分布

图 9.15　5 星评论（左图）和 1 星评论（右图）的 5 个主题

图 9.16　与 bPhone 相关的 tweet 的情感分析

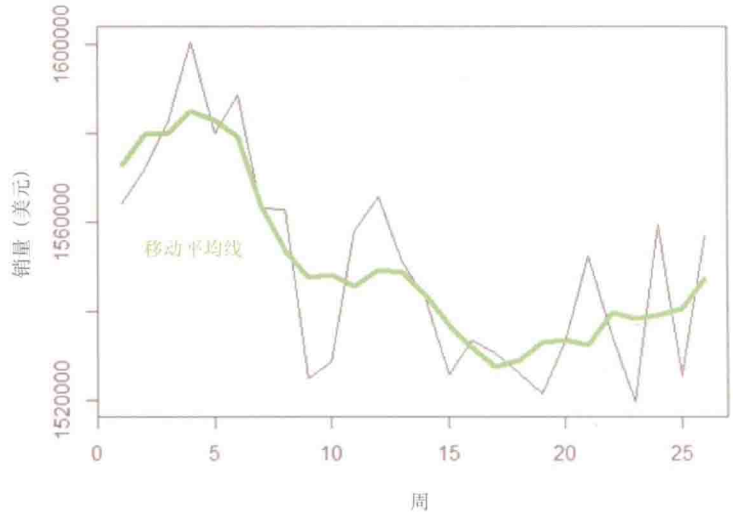


图 11.3　带有移动平均线的周销量

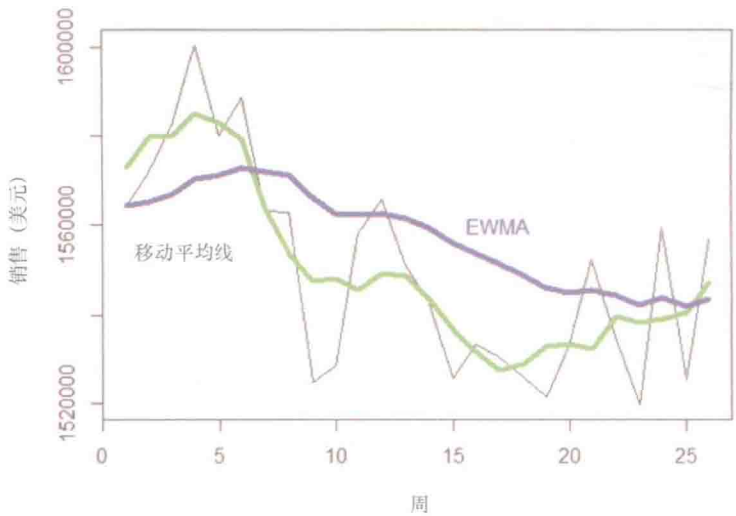


图 11.4　每周销售的移动平均线与 EWMA

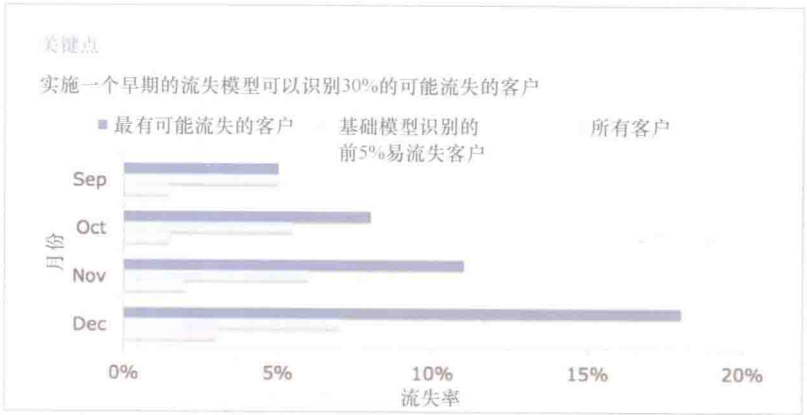


图 12.12 数据科学项目的关键论点示例，以条形图显示

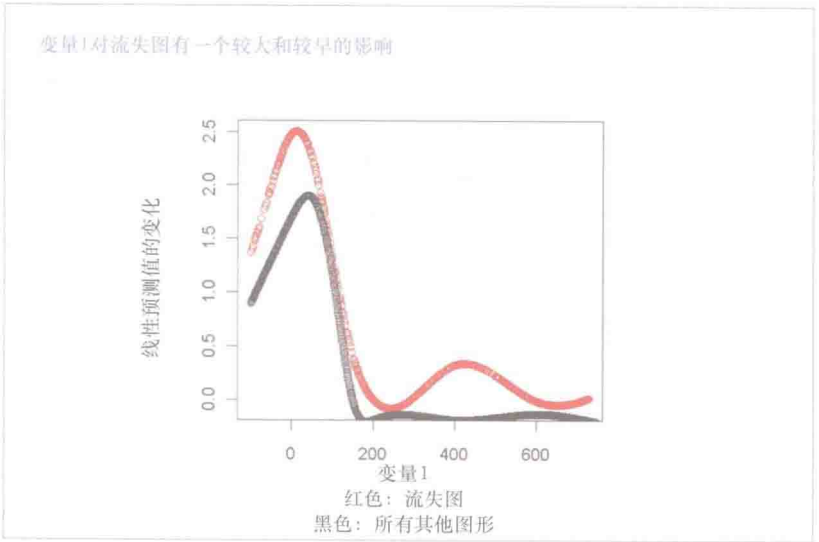


图 12.14 比较两个数据变量的模型细节

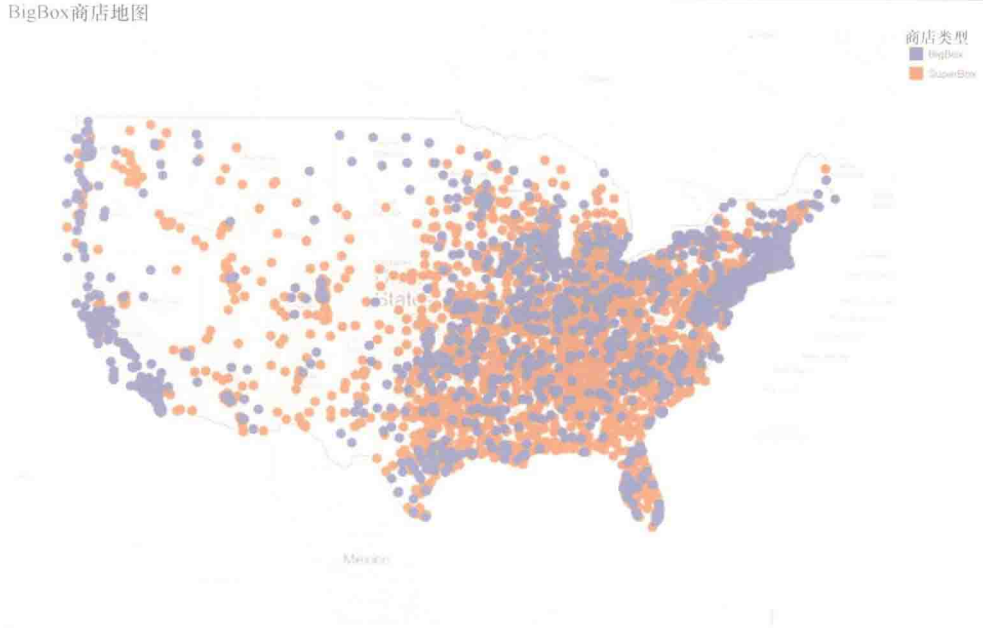


图 12.18 以地图形式显示的 45 年开店数据

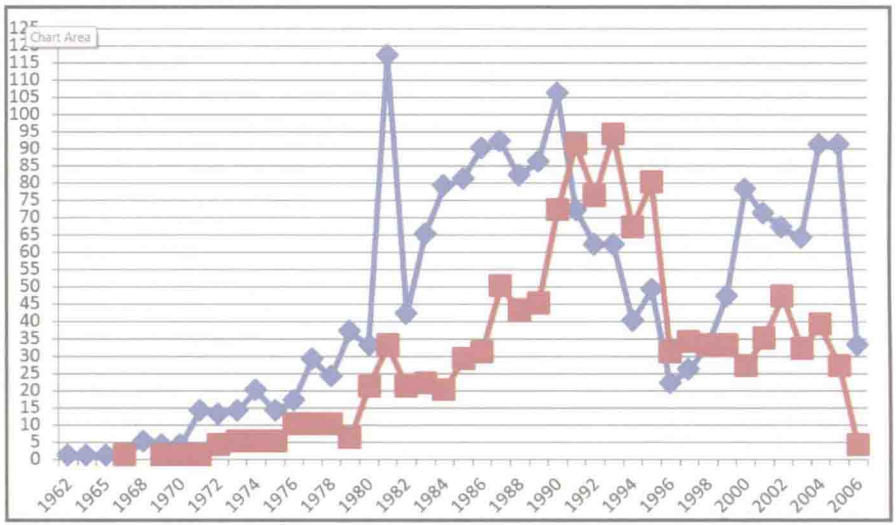


图 12.28　如何清理图形，例 1（清理之前）

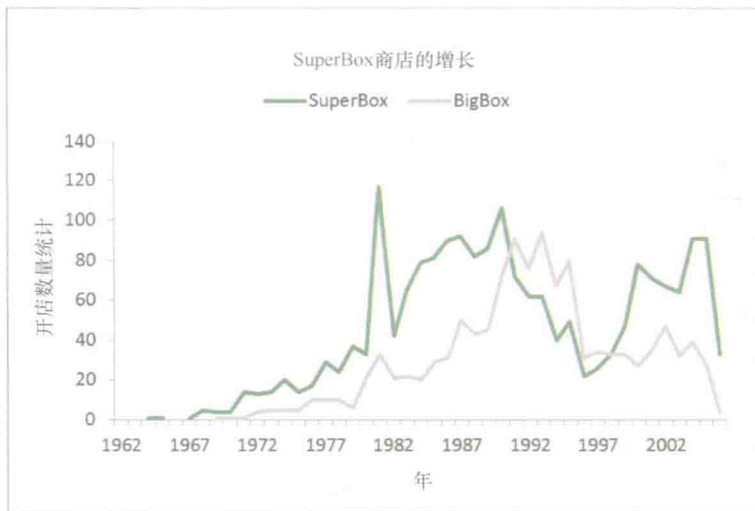


图 12.29　如何清理图形——例 1（清理之后）