WILEY

大数据应用与技术丛书

Data Analysis Using SQL and Excel，Second Edition

数据分析技术(第2版)

——使用SQL和Excel工具

[美] Gordon S. Linoff 著
陶佰明 译

清华大学出版社

大数据应用与技术丛书

数据分析技术

(第2版)

——使用SQL和Excel工具

[美] Gordon S. Linoff 著

陶佰明 译

清华大学出版社

北 京

Gordon S. Linoff
Data Analysis Using SQL and Excel,Second Edition
EISBN：978-1-119-02143-8

北京市版权局著作权合同登记号 图字：01-2016-1649

图书在版编目(CIP)数据

数据分析技术：使用 SQL 和 Excel 工具：原书第 2 版 / (美) 戈登 S.林那夫(Gordon S. Linoff) 著；陶佰明 译. —北京：清华大学出版社，2017
(大数据应用与技术丛书)
书名原文：Data Analysis Using SQL and Excel, Second Edition
ISBN 978-7-302-46139-5

Ⅰ.①数… Ⅱ.①戈… ②陶… Ⅲ.①关系数据库系统－应用－统计数据－统计分析 ②表处理软件－应用－统计数据－统计分析 Ⅳ.①O212.1

中国版本图书馆 CIP 数据核字(2017)第 010900 号

责任编辑：王　军　李维杰
封面设计：牛艳敏
版式设计：牛静敏
责任校对：曹　阳
责任印制：宋　林

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：010-62770175　　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 刷 者：清华大学印刷厂
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：39.5　　字　　数：961 千字
版　　次：2017 年 3 月第 1 版　　印　　次：2017 年 3 月第 1 次印刷
印　　数：1～3000
定　　价：98.00 元

产品编号：064477-01

译 者 序

数据分析，目的在于从海量数据中集中信息，总结、探索出信息中的模式。这个模式可以用于学术研究，可以用于决定市场的运作方向，甚至是影响一个行业的走向。因此，数据分析在实际业务中的角色举足轻重。在数据分析中，有两项技能是分析师必备的，分别是计算机相关技术和数据分析方法论。对于计算机从业人员，或许编写复杂的 SQL 语句轻而易举，然而，他们对数据分析方法——例如朴素贝叶斯模型——可能一无所知；而对于统计学家，即便是最简单的查询语句，可能都是让人头疼的事情。

而本书中，计算机技术与统计学的结合，正是本书最大的特色之一。纵观全书，所有的实例都是以统计学的方法进行分析，然后使用 SQL 予以实现。本书同时包含了对分析结果的展示。使用 Excel，以表格和图表的形式展示 SQL 的结果。直观并形象地体现出数据中的模式。本书介绍的关于 Excel 的使用方法和很多实用技巧，值得我们学习和使用。本书结构合理，构思巧妙。作者以实际业务中遇到的问题为起点，逐步引入对知识点的阐述。内容浅显易懂，难度逐渐提高，拥有不同技术等级的读者都可以从本书中受益。本书的作者有丰富的技术背景和行业经验，对于每一个知识点，都精心设计出一个与理论相关的实际业务问题。通过对分析问题、解决问题、展示结果的内容介绍，使读者在不知不觉中了解业务逻辑、理解理论知识、掌握解决问题和展示结果的相关技术。本书偏向对实际业务案例的介绍，毫不枯燥。

本书的翻译工作历经 8 个月，主要内容由我本人单独完成。这其中要特别感谢清华大学出版社的编辑们对本书的校验及审阅。同时，如下人员也参与了本书的翻译：孙宇佳、汪刚、李超华、金桂凤、姜静、王田田、孙玉亮、侯珊珊、高俊英、李显琴、邵丹、孙亚芳、佟秀风、田春红、陶玉霞、孙艳良、王洪波、张连廷、孙永福、孙守学，在此一并表示感谢。此外，整个翻译过程中，妻子默默地陪伴也是支撑我完成本书的动力，我由衷感谢她。

由于译者水平有限，翻译工作中可能会有不准确的内容，如果读者在阅读过程中发现失误和遗漏之处，望多多包涵，并欢迎批评指正。敬请广大读者提供反馈意见，读者可以将意见发到 bmtao0807@gmail.com，我会仔细查阅读者发来的每一封邮件，以求进一步提高今后译著的质量。

序　言

Gordon Linoff 与我共同编著了 3 本半书(如果 *Data Mining Techniques*，*Second Edition* 也算是新书的话，那么就是 4 本书，而且完成这本书要做的工作并不比其他书籍少)。在此之前，我们从未彼此单独著书，因此我也必须承认，当我看到本书封面上 Gordon 的名字旁边没有我的名字时，我的心中泛起阵阵悔意。然而，当记忆中关于写作生活的回忆潮水般涌来时，这份悔意很快就消失了。在写书时，假期里的生活是与键盘相伴，而不是在湖上泛舟，生活中的机会一个个被错过，人际关系变得越来越紧张。更重要的是，这本书只有 Gordon Linoff 可以撰写。他独有的天赋和经验体现在每个章节中。

我与 Gordon 初次见面是在 Thinking Machines Corporation 公司(一家曾经的巨型机制造商)，我们在 20 世纪 80 年代末和 90 年代初一起就职于这家公司。作为 Gordon 担当过的职位之一，Gordon 曾经负责管理并行关系型数据库的实现，使它支持在非常庞大的数据库上做复杂的分析查询。从根本上讲，这个数据库的设计与当时的其他关系型数据库系统不同，因为这些数据库并不支持事务处理。前者是对扫描和联接大型数据表的需求，而后者是对快速查找或更新记录的需求。放弃后者的需求，选择支持事务处理，使数据库在用于分析时，变得更加干净、高效。关于 Gordon 这部分的背景介绍，旨在指明 Gordon 对用于数据分析的 SQL 语言有深入了解。

正如数据库的设计，用于回答重要问题的设计，其结构不同于用于处理很多单个事务的数据库的设计，而用于回答重要问题的书籍，也需要不同的 SQL 使用方法。很多关于 SQL 的书籍是为了数据库管理员而撰写。其他书籍是为让读者能够做一些简单的报表。还有一些数据，尝试从每一个细节介绍 SQL 的特殊用法。而本书面向数据分析师、数据挖掘者以及任何想从大型数据库中抽取最大量信息的读者。本书的目的不是解决所有数据库使用者的需求，这使本书的内容更专注于愿意阅读本书的读者。简而言之，这是一本关于如何使用数据库的书籍。

相对于 Gordon 的数据库技术背景，更重要的是他作为数据挖掘顾问以来积累的多年经验。这些经验能使他深度理解不同种类的实际业务，并根据业务提出问题，再从已有数据中回答问题。多年来对数据库的探索，使 Gordon 能够本能地嗅到处理问题的方法，这其中可能要跨越很多不同的业务领域：

- 如何利用地理数据。邮政编码字段中包含的内容可能超乎你的认知。通过邮政编码可以获取经度和纬度值，使用经度和纬度，可以计算距离。邮政编码字段可以和人口统计局的数据做联接以获取重要特性，例如人口密度、中等收入、公共救助区的

人口比例等。

- 如何利用日期。订单日期、派送日期、报名日期、生日。公司的数据充满了日期。一旦理解了如何将日期转换为任期，如何按星期分析购买数据，如何追踪完成时间的趋势，这些日期就变得非常有用。当你理解了如何使用这些日期分析时间到事件的问题时，例如距离下一次购买的时间，或客户关系的剩余时间，你就会发现日期更加有用。
- 如何直接在 SQL 中建立数据挖掘模型。本书所展示的 SQL 使用方式，可能是你从未想象过的，包括为超市购物车的分析生成关联规则、建立回归模型、实现朴素贝叶斯模型和分数集。
- 如何为数据挖掘工具准备数据。虽然很多情况下可以使用 SQL 和 Excel 的组合解决问题，然而最终你可能还是会使用一些特殊的数据挖掘工具。这些工具需要特殊格式的数据，即客户签名。本书展示了如何创建这些数据挖掘工具所需要的数据。

本书包含海量的示例，并且它们都是真实数据。这是值得一提的。虚假的数据集将导致虚假的结果集。对于学生来说，这是很容易让他们郁闷的。在真实生活中，你越是理解业务环境，就越能够使用数据挖掘返回正确的结果。项目专家为你开了一个好头。你了解了哪些变量是具有预测性的，并以此判断出新的变量。虚假数据不会给你带来好的思路，因为应该存在于数据库中的模式消失了，相反，不应该存在的模式被虚拟出来。真实数据很难处理，因为真实数据可能会揭露出更多问题。因此，很多书籍和课程使用人工构建出来的数据集。此外，本书中的数据集都可在网站上免费下载，网址为 www.wiley.com/go/dataanalysisusingsqlandexcel2e。

在本书的编著过程中，我审阅书中的内容。在此期间，我对 SQL 和 Excel 的理解更深了，从中受益匪浅。在示例中，对于相当复杂的查询的思考练习，极大地提高了我对 SQL 工作方式的理解程度。通过阅读本书，我不再畏惧嵌套查询、多路联接、庞大的 CASE 语句，以及这门语言中其他令我畏惧的地方。在过去几十年的合作中，每当有 SQL 和 Excel 的问题出现时，我经常向 Gordon 寻求帮助。而现在，我可从本书中找到答案。你也可以。

—Michael J. A. Berry

作者简介

长时间以来，Gordon S.Linoff 一直在致力于处理数据库、大数据和数据挖掘。在高效使用数据的实践中，他拥有数十年的经验，被认为是数据挖掘领域中的专家。

当 Gordon 还是 MIT 的一名学生时，他最初在康柏手提电脑(Compaq Portable)上开始使用电子表格(康柏手提电脑世界上第一批便携电脑)。几年后，他在 Thinking Machines Corporation 公司管理一个开发小组，主要任务是为决策并支持建立一个大型的并行关系型数据库。

Thinking Machines 倒闭后，在 1998 年，他与他的好友兼前同事 Michael J.A.Berry（2012 年离世）创建了 Data Miners。从那以后，他在不同的公司里工作，参与了大量不同种类的项目。通过统计和业务分析软件的领导者 SAS Institute，他传授了超过一百节关于数据挖掘和生存分析的课程。他也是 Stack Overflow 的热心贡献者，特别关注数据库相关的问题，而且他是 2014 年最高分数的持有者。

与 Michael Berry 一起，Gordon 编写了几本关于数据挖掘的很有影响力的书籍，其中包括 *Data Mining Techniques for Marketing, Sales, and Customer Support*，本书是数据挖掘领域第一部实现第三版的书籍。

Gordon 和与他共处 25 年的妻子 Giuseppe Scalia 居住在纽约。

致　谢

尽管本书的封面上只有我的名字，但是很多人曾经帮助过我完成此书，以及帮助我更宏观地理解数据、分析数据和展示数据。

1990 年，我与 Michael Berry 初次相见。后来我们共同建立了 Data Miners，他在所有战线上都有功劳。他审阅章节、测试示例中的 SQL 脚本、帮助分析数据。他的洞察力和调试能力使示例更为精准。同时也要赞扬他的妻子 Stephanie Jack 的耐心以及无私分享 Michael 的时间。

最初著书的想法来自 Nick Drake，那时他还在 Datran Media 工作。作为一个统计学家，Nick 一直在寻找一本能够帮助他使用数据库来做数据分析的书籍。与此同时，我在 Wiley 的编辑 Bob Elliott 也喜欢这个主意。

纵观所有章节，对于数据处理的理解基于数据流。这个概念是 Ab Initio Corporation 的 Craig Stanfill 在很久以前向我提出的，那时我们还一起在 Thinking Machine Corporation 工作。

一直以来，我从不同的人身上学到了很多。来自 SAS Institute 的 Anne Milley 首次建议我学习生存分析。此后，我很多关于这个主题的知识都承自于 Will Potts，目前他效力于 CapitalOne。Brij Masand 帮助我延伸这个想法并在用于预测的应用程序上得到了实现。针对将生存分析应用于计算客户价值的领域，Chi Kong Ho 和他的团队在《纽约时报》上给予了非常有价值的反馈。

来自《纽约时报》的 Stuart Ward 和 Zaiying Huang 花费了大量时间阐述和讨论统计学概念。同样来自《纽约时报》的 Harrison Sohmer 也教会我很多 Excel 的使用技巧，其中一些技巧已经纳入本书。

来自微软的 Jamie MacLennan 和 SQL Server 团队在产品问题上给我提供了很多帮助。

在过去几年中，我一直是 Stack Overflow 的主要贡献者。一直以来，我学到了大量的关于 SQL 以及如何解释其概念的知识。还有少数未曾见面的人，同样在不同的方面给了我很大帮助。Richard Stallman 创造了编译器 emacs 和 Free Software Foundation；emacs 为日历表提供了基础。书中的几个章节使用了来自 Applications Professional,Inc 的 Rob Bovey 创建的 X-Y 图表标签器。密苏里人口数据统计中心的人们建立了我们所用的人口统计数据。Juice Analytics 启发了第 5 章中关于工作表菜单栏的示例(感谢 Alex Wimbush 帮我指出他们的方向)。Frontline System 的 Edwin Straver 回答了关于 Solver 的一些问题。

多年来，很多同事、朋友和学生都为我提供了灵感、问题和答案。有太多的人士，以至于我无法一一列出，但是我要特别感谢 Eran Abikhzer、Christian Albright、Michael Benigno、

Emily Cohen、Carol D'Andrea、Sonia Dubin、Lounette Dyer、Victor Fu、Josh Goff、Richard Greenburg、Gregory Lampshire、Mikhail Levdanski、Savvas Mavridis、Fiona McNeill、Karen Kennedy McConlogue、Steven Mullaney、Courage Noko、Laura Palmer、Alan Parker、Ashit Patel、Ronnie Rowton、Vishal Santoshi、Adam Schwebber、Kent Taylor、John Trustman、John Wallace、David Wang 和 Zhilang Zhao。同时还要感谢来自 SAS Institute Training 团队的人们，多年以来，他们帮助我组织、审阅、发起了关于数据挖掘的课程，使得我有机会与数据挖掘领域不同有趣的人相遇。

同样要感谢我的朋友和亲人们：我的母亲、父亲、姐姐 Debbie、哥哥 Joe，我的姻亲 Raimonda Scalia、Ugo Scalia 和 Terry Sparacio，以及朋友 Jon Mosley、Paul Houlihan、Leonid Poretsky、Anthony DiCarlo 和 Maciej Zworski，在撰写本书时，我曾拜访过他们，而且他们给予我空间和时间以完成这项任务。另外，我的猫 Luna 曾经蜷缩在我身边很长时间，它或许也会怀念我的写作时光。

最后，没有包含对贤妻 Giuseppe Scalia 的感谢是不完整的。在完成所有 7 本书的过程中，我的妻子确保我的头脑是清醒的。

谢谢每一个人！

前　言

本书的第 1 版使用我们熟悉的工具 SQL 和 Excel，从实用的角度解释数据分析。这本书的指导原则是从问题出发，同时从业务角度和技术角度提供解决方案，以指导读者。这个方法被证明是非常成功的。

从第 1 版到现在已经过去了 10 年，这期间已经发生了很多变化，工具本身也发生了很多变化。例如，当年的 Excel 还没有功能区，而且在当时的数据库中，窗口函数也非常罕见。一些工具，如 Python 和 R，以及 NoSQL 数据库变得越来越常见，它们改变了分析师赖以生存的工具世界。然而，随着技术延伸到大大小小的各项业务中，关系型数据库在今天仍然被广泛使用，而且 SQL 也变得更加至关重要。对于很多商务人士，Excel 工具仍然是做报表和展示的理想之选。大数据不再是未知的领域，它是我们每天都会面临的问题、挑战和机遇。

根据底层软件的变化，在第 2 版中对本书的内容做了调整和更新，同时包含了更多的示例和技术，以及增加了关于数据库性能的一整章新内容。同时，我一直在努力保持本书第 1 版的优势。本书仍然围绕着数据、分析和展示的原则——少见地将三个功能放在一起处理。示例围绕着所提出的问题，同时讨论了这些问题的业务相关性和技术实现。示例使用的是真实的代码。数据、代码以及 Excel 示例都可以在配套网站上找到。

撰写这本书的最初动机来源于我的一个同事——Nick Drake，他是受过培训的统计学家。曾经，他一直在寻找一本书，关于介绍如何使用 SQL 编写可用于数据分析的复杂查询。当时，基于 SQL 的书籍，要么介绍 SQL 的基础查询结构，要么介绍数据库的工作原理。严格地讲，没有从分析数据的角度介绍 SQL 的书籍，也没有基于回答数据问题的书籍。在统计学的众多书籍中，没有一本书能够面对这样一个事实提出解决方案：统计学所用的数据，多数都存储于关系型数据库中，而本书则填补了这一空白。

笔者与 Michael Berry 一起撰写的其他关于数据挖掘的书籍，侧重于高级算法和案例学习。相比之下，本书侧重于“操作方式”。首先描述了存储在数据库中的数据，然后继续完成准备数据和生成结果集的过程。书中穿插的内容，是我在这个领域多年经验的结晶，解释了结果集被应用的可能方式，以及为什么有些事情有效果，而有些事情无效。书中示例非常具有实践性，它们所使用的数据都在本书的配套网站上(www.wiley.com/go/dataanalysisusingsqlandexcel2e)。

关于数据仓库和分析数据库的一个老生常谈的话题是它们实际上没有做任何事。是的，它们存储数据，能够将不同来源的数据汇集在一起，并整理数据使数据变得清晰。是的，

它们定义业务维度，存储关于客户的事务，还可能总结重要的数据(是的，所有这些都非常重要！)然而，数据库中的数据存储在旋转的硬盘上，而且数据在计算机内存中的数据结构非常复杂。对于如此多的数据，信息却很少。

我们如何探索这些数据(特别是描述客户的数据)？很多关于统计学建模和数据挖掘的华丽算法都有一条简单的规则："无用输入，无用输出"。即使是最复杂的技术，也只有当数据是好数据时，结果才是好的。数据是理解客户、产品以及市场的中心。

本书中的章节覆盖了数据的不同方面，同时包含了 SQL 和 Excel 支持的重要的数据分析技术。这些数据分析技术的范围涵盖了很多内容，从最初的探索性数据分析到生存分析，从超市购物车分析到朴素贝叶斯模型，从简单的动画到线性回归。当然，本书不可能涵盖所有的数据分析技术。本书所介绍的方法历经时间的考验，被认为是有用的且适用于很多不同的领域。

最后，只有数据和分析还不够，还必须将结果展示给正确的观众。为完整地探索数据值，需要将数据转化为故事和情景、图表、数据指标和透视图。

本书内容和技术综述

本书侧重于三个关键的技术领域，这些技术用于将数据转化为可操作的信息：

- 关系型数据库存储数据。获取数据的最基本的语言是 SQL(注意，变种的 SQL 也用于 NoSQL 数据库)。
- Excel 工作表是展示数据的最常见工具。或许，Excel 最强大的功能是绘图，它能够将包含数字的列转换为图片。
- 统计学是数据分析的基础。

这三种技术一并介绍，是因为它们是彼此相关的。SQL 回答"我们如何访问数据？"统计学回答："数据是如何相关的？"而使用 Excel 可以方便地向人们展示和证明我们所发现的结论。

关于数据处理的描述围绕着 SQL 语言。在实际业务中，Oracle、PostgresSQL、MySQL、IBM DB2，以及微软的 SQL Server 等都是常见的数据库，它们存储海量的业务数据事务信息。好消息是所有的关系型数据库都支持 SQL 作为查询语言。然而，正如英国和美国被称为是"拥有共同语言的两个国家"一样，每种数据库支持一些与众不同的 SQL 方言。附录列出了如何使用不同的 SQL 方言实现一些常见的功能。

相似地，也有其他华丽的展示工具和专业的制图包。然而，对于一台用于工作的电脑，安装 Excel 或类似的电子表格工具是再常见不过的事情了。

统计学和数据挖掘技术通常并不需要高级工具。其中一些非常重要的技术，可以使用 SQL 和 Excel 轻易地实现，包括生存分析、相似模型、朴素贝叶斯模型和关联规则。事实上，本书中介绍的方法通常比这些工具中的方法更强大，因为书中的方法更接近数据，因此它们更精准，而且容易定制。对这些技术的介绍涵盖了基础思想和深度扩展，这是在其

他工具中所没有的内容。

本书章节描述了不同的技术，在熟悉工具和数据的前提下，为数据建模和数据探索提供扎实的知识介绍。本书同时强调，当简单工具遇到瓶颈时，高级工具是非常有用的。

内容结构

本书的 14 章可以分为 4 部分。前 3 章介绍 SQL、Excel 和统计学的核心概念。中间 7 章讨论特别适合使用 SQL 和 Excel 的数据探索和数据分析技术。在后续的 3 章中，从统计学和数据挖掘的角度，介绍了关于建模的更正式的思想。最后，新增的第 14 章讨论编写 SQL 查询时的性能问题。

每一章都通过不同的视角，介绍使用 SQL 和 Excel 做数据分析的方方面面，包括：

- 使用数据分析的基础示例
- 分析师需要回答的问题
- 详解数据分析技术的工作原理
- 实现技术的 SQL 语法
- 以表格或图表展示结果，以及如何在 Excel 中创建它们

SQL 是一门精准的语言，以至于有时难以读懂。数据流程图通常有助于理解 SQL 的工作原理。这些数据流程图是 SQL 引擎实际处理数据的合理预测，当然，实际上的数据处理细节由数据库引擎决定。

结果以表格或图表的形式展现，分布在本书的所有章节中。此外，本书强调了 Excel 的一些重要特征，介绍了 Excel 图表的一些有趣用法。每一章都有技术专栏，通常讲述某项技术的重要方面或与正文内容相关的一些有趣历史背景。

章节引导

第 1 章“数据挖掘者眼中的 SQL”从数据分析的角度介绍 SQL，这是 SQL 语言的查询部分，使用 SELECT 查询从数据库中获取数据。

第 1 章介绍了描述数据结构的实体-关系图——表、列，以及它们彼此间的关系。该章同时介绍了用于描述查询处理过程的数据流程图；通过数据流程图，能够可视化地理解数据的处理过程。本章介绍了全书中使用到的一些重要功能——例如联接、聚合和窗口函数。

此外，第 1 章还描述了全书示例所使用的数据集(该数据集也可以从网站自行下载)。数据包括存储零售数据的表，存储手机客户数据的表，以及其他描述邮政编码和日历的引用表。

第 2 章“表中有什么？开始数据探索”介绍使用 Excel 做数据探索和结果展现。在 Excel 的众多功能中，或许最有用的功能就是绘图了。正如一句古老的中国谚语所说，“百闻不如一见”。Excel 的绘图依据是数据。这样的图表不仅美观有用，同时在 Word 文档、PPT 展

示、电子邮件、网站中也非常实用。

图表并非终点，它们只是探索数据分析的一个方面。此外，本章还介绍了在表格中汇总列，以及使用 Excel 生成 SQL 查询的有趣想法。

第 3 章“不同之处是如何不同”介绍了一些描述性统计学的核心概念，例如平均值、P 值和卡方检测。本章的目的是展示如何将这些技术应用于数据表中的数据上。至于这些统计学内容和统计学测试方法的选择，是由它们的实用性决定的。同时，本章侧重介绍这些知识的使用方法，而不是它们的理论内容。多数的统计学测试方法都可以使用 Excel(甚至 SQL)来实现。

SQL 技术

一些技术非常适合使用 SQL 和 Excel。

第 4 章“发生的地点在何处？”介绍了地理数据以及如何将地理信息纳入数据分析中。地理信息首先是位置，以经度和纬度描述。位置也可以用不同等级的地理信息描述，例如人口普查区、邮政编码区域，以及其他我们熟悉的国家和省份，这些数据都可从人口统计局(或是其他相似的政府机构)获取。这一章也讨论了如何使用不同地理等级比较结果集。最后，不包含地图的地理信息是不完整的。使用基础的 Excel 功能，可以创建非常初级的地图。

第 5 章“关于时间”讨论了客户行为的另一个关键特征：什么时候发生。该章描述了如何访问数据库中的日期和时间，以及如何使用这些信息来帮助理解客户。该章包含的示例，可以用于准确地比较不同年份的数据，并从历史上计算每天的活跃客户数量。该章最后介绍 Excel 中的一个简单的动画——也是本书中唯一一处使用 Visual Basic 的地方。

第 6 章和第 7 章介绍了用于理解客户随时间变化的最重要的数据分析技术。在传统的统计学中，生存分析根深蒂固，而且它也很适合处理与客户相关的问题。

第 6 章“客户的持续时间有多久？使用生存分析理解客户和他们的价值”介绍了风险率和生存率的基本思想，解释了如何使用 SQL 和 Excel 简单地计算它们。或许令人感到惊讶的是，在使用生存分析时，并不需要复杂的统计学工具。第 6 章后续介绍了生存分析应用在实际业务中的重要性，例如平均客户生命周期。然后讲解如何将这些片段拼接在一起，形成对客户值计算的预测。

第 7 章“影响生存率的因素：客户任期”扩展讨论三个不同的领域。第一，它解决了在以客户为中心的数据库中的重要问题：左截断(left-truncation)。第二，它介绍了生存分析领域中的一个非常有趣的思想：竞争风险。这个思想考虑了一个事实，即客户是因不同原因而离开的。第三，将生存分析应用在分析前和分析后。即当客户在其生命周期内发生一些事情时，我们如何量化所发生的事情，例如量化客户加入忠诚计划之后的影响，或量化一次失败的主要计费方法。

第 8 章至第 10 章使用 SQL 和 Excel 介绍如何理解客户正在购买的内容。

第 8 章“多次购买以及其他重复事件”介绍了关于购买事件的所有事——什么时候发生，在哪里发生，发生频率——除了购买的东西。该章介绍了 RFM，一种理解客户购买行为的传统技术。同时介绍了随时间推移，在识别客户时的种种问题。即使是在我们查看详细的购买信息之前，我们也能发现很多关于购买的信息。

在第 9 章“购物车里有什么？购物车分析”中，产品成了焦点。该章介绍了随时间推移，针对购买行为的探索性分析。该章包括了如何识别驱动客户行为的产品，同时介绍了 Excel 中一些有趣的可视化方法。

第 10 章“关联规则”转移到对关联规则的正式讨论。关联规则是指被同时购买或按序购买的产品组合。在 SQL 中建立关联规则是相当复杂的。本章讨论的方法扩展了传统的关联规则分析，介绍更有效的替换指标，并展示如何生成不同事物的组合。例如，单击会导致一次购买行为(使用网站的一个实例)。在本章中解释的关联规则技术，比数据挖掘工具中的技术更强大，因为这里的技术是可以扩展的，并使用支持度、置信度和提升度之外的指标。

建模技术

接下来的 3 章讨论统计学和数据挖掘的建模技术和方法。

第 11 章“SQL 数据挖掘模型”介绍了数据挖掘的建模思想，以及建模相关的名词。同时讨论了一些重要的模型类型，这些模型适用于处理业务问题和 SQL 环境。相似性模型找到与给定示例相似的事物。查找模型使用查找表返回模型评分。

该章同时介绍了一种更复杂的建模技术，即朴素贝叶斯模型。这门技术可以总结不同业务维度的信息来估算未知的数值。

第 12 章“最佳拟合线：线性回归模型”介绍了一种更传统的统计学技术：线性回归。该章介绍了不同种类的线性回归，包括多项式回归、加权回归、多维回归和指数回归。这些内容以 Excel 图表的形式介绍，同时包含 R^2 值，用于衡量模型与数据的拟合度。

对回归的介绍同时用到了 Excel 和 SQL。虽然 Excel 中有几种内置的功能可以处理回归问题，但 Solver 比这些内置功能更强大。本章从线性回归的角度介绍了 Solver(Solver 是可与 Excel 绑定的免费加载项)。

第 13 章“为进一步分析数据创建客户签名”介绍了客户签名。客户签名是一个数据结构，它总结了客户在某个特定的时间点的数据。客户签名在建模时非常强大。

在介绍该章时认识到虽然 SQL 和 Excel 都非常强大，但有时还需要一些更复杂的工具。很多情况下，客户签名是总结客户信息的正确方法，而且 SQL 是完成这类总结的强大工具。

性能

编写 SQL 查询的一个原因是性能——通过至少完成一些分析工作，可以将已有的硬件

资源分配给关系型数据库。编写一本关于通用 SQL 而非指定数据库的书籍，其缺点就是缺少关于特定数据库的一些技巧和提示。

令人欣慰的是，很多关于编写 SQL 的最佳实践能够普遍提升查询在不同数据库中的执行速度。第 14 章“性能问题：高效使用 SQL”致力于这个话题。其中特别讨论了索引和如何利用索引，同时还介绍了编写查询的不同方法？——以及为什么有些方法的性能更好。

本书读者对象

本书面向不同技术等级的各类读者。

技术方面不足的管理者，特别是那些负责理解客户或业务单元的管理者。通常情况下，这样的人精通 Excel，然而，他们所需要的数据存储于关系型数据库中。为了帮助他们，本书中的示例提供了有用的结果集。这些示例十分详尽，不仅展示了业务问题，同时展示了技术方法和结果。

另一部分读者，他们的工作是理解数据和客户，通常他们的职位描述中包含“分析师”字样。这些人通常使用 Excel 和其他工具，有时直接访问数据仓库或一些以客户为中心的数据库。本书能帮助他们提高 SQL 查询技巧，展示好的图表示例，以及介绍生存分析和关联规则，以便他们理解客户和业务。

一部分重要的读者是数据科学家，他们精通诸如 R 或 Python 这样的工具，但是他们发现需要学习其他的工具。在业务世界中，以编程为中心的工具可能并不足以解决问题，分析师可能会发现他们不得不直接处理关系型数据库中的数据，并以 Excel 形式展现给用户。

技术等级更高的是统计学家，他们通常使用有特殊功能的工具，例如 SAS、SPSS、R 和 S-plus。然而，数据存储于数据库中。本书可以在 SQL 技术方面为他们提供帮助，并提供数据分析示例以帮助他们解决业务问题。

此外，数据库管理员、数据库设计者和架构师应该会发现本书是非常有趣的。在不同章节中展示的查询，说明了人们对数据的使用方式和方法。这些查询应该可以促进数据库管理员和设计者创建更适合使用的高效数据库。

建议所有的读者，即使是技术专家，阅读或至少浏览前 3 章内容。这些章节全部从分析海量数据的视角，介绍 SQL、Excel 和统计学知识。这个视角与平常所读书籍的视角不同。在这些章节中，有相当一部分的内容和想法贯穿全书，例如样本数据、数据流、SQL 语法和格式转换、出色的图标绘制。

需要的工具

本书是独立的——读者应该可以直接通过书中的内容阅读并学习。

本书中的所有 SQL 语句都经过测试(在微软 SQL Server 数据库上，少量查询在其他数据库(PostgresSQL)上测试)。可以从网上下载数据集和结果，网址为 www.wiley.com/go/data-

analysisusingsqlandexcel2e。对于想要尝试的读者，我们建议下载数据并执行书中的示例代码。

本书中，多数示例是与数据库供应商无关的，因此，它们(或稍作修改后)应该可以在所有的关系型数据库中执行。这里不建议使用 Microsoft Access 或 MySQL，因为它们缺少窗口函数——窗口函数是分析性查询的关键功能。

如果没有数据库，可以下载一些程序包；数据库供应商通常会提供一些免费的单机版本。例如，SQL Server Express 是微软提供的免费 SQL Server 版本，Oracle 也提供免费版本的 Oracle 数据库，可以从 www.postgres.org 下载 PostgresSQL 数据库，其他数据库也有它们的免费版本。

网站内容介绍

配套网站(www.wiley.com/go/dataanalysisusingsqlandexcel2e)上包含本书使用的数据集。这些数据集包含如下信息：

- 引用表。共有 3 个引用表，其中两张表包含人口统计信息(来自于人口统计局 2000 年的统计数据)，另一张表包含关于日期的日历信息。
- Subscribers 数据集，用于描述移动电话公司的客户子集。
- Purchases 数据集，用于描述客户购买模式的数据集。

下载这些数据的同时，还可以下载将数据导入 SQL Server 和其他数据库的使用说明。

此外，配套网站的其他页面包含更多的信息。例如，将数据导入常见数据库中的脚本，包含 SQL 查询的工作表，以及本书中使用 Excel 生成的所有表格和图表。

总　结

本书起源于一个同事的问题，他询问是否有一本关于使用 SQL 做数据分析的参考书。然而，所需要的并不是简单的关于 SQL 的参考书，即使它侧重介绍使用 SQL 做数据查询的实际使用。

对于数据分析，不能凭空学习 SQL。一个 SQL 查询，不管它编写的多么精妙，通常不是一个业务问题的完整解决方案。业务问题，需要被转换为可以使用查询回答的问题。然后需要将结果展示出来，通常以表格或 Excel 图表的形式。

笔者想要扩展这个观点。在现实世界中，也不能凭空学习统计学知识。曾经，收集数据不仅花费时间且难以操作。现在，数据量非常足够。例如，本书的配套网站，只需要轻点几下，就能上传几 GB 的数据。数据分析的问题不再局限于几个统计学方法，同时包括管理和抽取数据。

本书将三个核心概念融入到解决问题这一条线中。在笔者的数据挖掘生涯中，笔者发现 SQL、Excel 和统计学是分析数据的关键性工具，比某些特殊的技术更加重要。希望本书可以帮助读者改进他们的技术，并为他们理解客户和理解业务提供新思路。

目　录