高等院校计算机应用系列教材

Excel 在数据分析中的应用

（第 2 版）

陈　斌　主　编

吕洪柱　张银霞　张桂香　迟立颖　张光妲　副主编

清華大學出版社

北　京

内 容 简 介

在当今数字化时代，人们在日常生活和工作中要面对大量的数据，如何有效地利用这些数据改善生活、改进工作成为人们关心的问题，而数据挖掘技术的发展为人们打开了一扇窗，提供了各种实用工具用于进行数据分析与处理。但这些工具往往专业性强，对使用者要求高，很难普及，而 Excel 是大众普遍使用的电子表格制作软件，它不仅能够保存数据，在数据处理中也有着良好的表现，特别是它提供的函数、图表、数据透视表、数据分析工具、规划求解工具等，能够有效解决各类复杂的数据分析与处理问题，为日常生活和工作中的数据处理提供了重要支持。

本书以数据分析与处理的过程为主线，以统计学为基础，通过 Excel 的各种功能实现对各种数据的处理，主要内容包括数据分析与处理概述、数据收集与预处理、Excel 函数在数据分析中的应用、数据管理、数据图表化展现、抽样与参数估计、方差分析、时间序列分析、相关分析、常用统计分布图形分析、回归图形分析、Excel 在数据挖掘中的应用。

本书内容丰富，结构清晰，采用了从原理到实践的讲解方式，并给出了大量的案例。同时，本书附赠 PPT 教学课件、案例源文件和结果文件，以便于教学。

本书既可作为高等院校数据分析相关课程的教材，也可作为企事业单位数据分析人员的参考用书。

图书在版编目(CIP)数据

Excel 在数据分析中的应用 / 陈斌主编. -- 2 版.
北京 : 清华大学出版社, 2025. 2. -- (高等院校计算机应用系列教材). -- ISBN 978-7-302-68046-8

Ⅰ. TP391.13

中国国家版本馆 CIP 数据核字第 2025YP6666 号

责任编辑：刘金喜
封面设计：高娟妮
版式设计：思创景点
责任校对：成风进
责任印制：曹婉颖

出版发行：清华大学出版社
网　　址：https://www.tup.com.cn，https://www.wqxuetang.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：三河市天利华印刷装订有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张：**18.75　　**字　　数：**503 千字
版　　次：2021 年 2 月第 1 版　2025 年 3 月第 2 版　　**印　　次：**2025 年 3 月第 1 次印刷
定　　价：69.80 元

产品编号：106703-01

前　言

随着大数据技术的发展，人们对身边大量数据的价值认识不断加深，数据的价值不仅仅是专业人士关注的问题，普通人对其也愈发重视。人们对数据进行分析与处理的意识增强，使得一些数据分析处理软件受到青睐。然而，专业的分析软件的专业性强且分析结果不易理解，因此对使用者的要求较高。相比之下，Excel电子表格制作软件不仅操作简单、易学易用，而且数据整齐、美观，还能像数据库操作一样对表格中的数据进行各种复杂的计算与处理分析。Excel提供的大量函数可以对数据进行拆分、合并、分类、汇总等操作，实现对数据的预处理；通过图表、数据透视表、数据透视图等功能，使数据图表化更简单；同时，内置的数据分析工具、规划求解工具等为采用统计学方法解决问题提供了强有力的支撑，可以让人们更加轻松地进行数据处理和统计分析。

本书以数据分析与处理的过程为主线，在简单介绍统计学知识的基础上结合具体实例讲解了应用Excel实现数据分析与处理的过程。本书共分为12章，按照由浅入深、循序渐进的思路进行编排。

第1章——数据分析与处理概述，内容包括数据分析与处理的概念、处理过程及大数据的应用；

第2章——数据收集与预处理，介绍了数据的分类、收集方法、清理、集成及转换；

第3章——Excel函数在数据分析中的应用，介绍了数据分析中的常用函数及实际的应用案例；

第4章——数据管理，介绍了数据排序、筛选、分类汇总及合并计算等数据管理操作；

第5章——数据图表化展现，介绍了数据图表、数据透视表及数据透视图等图表化工具；

第6章——抽样与参数估计，介绍了抽样与参数估计的基本概念、利用Excel实现抽样推断的方法，以及不同条件下的区间估计的实现方法；

第7章——方差分析，介绍了单因素方差分析和双因素方差分析；

第8章——时间序列分析，内容包括统计对比分析、移动平均分析、指数平滑分析、趋势外推分析及季节调整分析；

第9章——相关分析，内容包括简单相关分析、多元相关分析及等级相关分析；

第10章——常用统计分布图形分析，介绍了概率函数、正态分布、泊松分布、指数分布、卡方分布等图形分析；

第11章——回归图形分析，介绍了一元线性回归、多元线性回归及非线性回归等图形分析；

第12章——Excel在数据挖掘中的应用，内容包括聚类分析和判别分析。

本书内容丰富、实例典型，采用了由浅入深、理论与实际操作相结合的讲解方式，在内容编写上注重实用性和可操作性，通过大量实例让读者直观、快速地掌握Excel的功能，实现对数据的分析与处理。此外，本书还配有大量的练习题供读者对所学知识加以巩固。本书针对性和实用性较强，适用于数据分析与处理的初学者，既可作为高等院校数据分析相关课程的教材，也可作为企事业单位人员学习数据分析的参考用书。

本书由陈斌担任主编。本书的编写分工如下：迟立颖编写第1章、第4章，张桂香编写第2章、

第 8 章，张光妲编写第 3 章，张银霞编写第 5 章、第 7 章，李敬有编写第 6 章，陈斌编写第 9 章，吕洪柱编写第 10 章、第 11 章的第 1 节和第 2 节，以及第 12 章，王桂英编写第 11 章的第 3 节，吕洪柱和张银霞审阅了全书，并提出了宝贵意见。

在经过几轮教学实践后，编者对书中存在的疏漏和不妥之处进行了修改，特别是对一些陈旧的实例进行了更新。在本书的编写过程中，获得了同行及专家的支持，并参考了部分网络资料，在此一并表示感谢。

本书编者力图在书中对 Excel 在数据分析中的应用方法进行完美呈现，但限于作者水平，书中难免有不足之处，欢迎广大读者批评指正。

本书 PPT 课件、案例源文件和习题答案可通过扫描下方二维码获取。

教学资源下载

服务邮箱：476371891@qq.com

编　者

2024 年 11 月

目　录

第 1 章

数据分析与处理概述

随着人工智能、数据挖掘、大数据等概念不断为人们所熟悉，不同行业领域的人们开始考虑一个共同的问题：如何在自身已有的繁杂数据中快速找到有价值的信息。数据分析与处理技术可以有效地实现这个目标。

本章将介绍数据、数据分析与处理的概念，并结合实例展示数据分析与处理的基本过程，以及大数据的相关概念、技术及应用。

1.1 数据分析与处理简介

在介绍数据分析与处理相关知识之前，我们先来看一个案例。

全球知名家用电器和电子产品零售商百思买，其销售产品种类近 4 万种，产品价格也随各地区消费水平和市场条件而不同。由于产品种类繁多且促销活动频繁，产品价格一年中可变动 4 次，结果导致每年的调价次数高达 16 万次，这让公司高管觉得非常难以管理。鉴此，公司成立了一个定价团队，希望通过分析消费者的详细购买记录，提高定价的准确度和响应速度。

定价团队的分析围绕以下三个关键维度展开。

(1) 数量。定价团队收集了上千万消费者的购买记录，从客户的购买习惯、常用产品等多个维度进行分析，了解客户对每种产品定价的最高接受能力，从而给出产品的合理定价。

(2) 多样性。除了分析购买记录数据，定价团队还应考虑社交媒体发布的产品促销数据，如消费者点赞、留言、促销优惠券等多种类型数据的影响。

(3) 速度。为了实现价值最大化，团队对所获得的数据进行了实时处理，他们能够根据一个消费者既往的产品购买记录，为当时身处卖场该产品销售区的此消费者推送优惠信息、赠送优惠券，为客户带来惊喜。

通过上述活动，定价团队提高了定价的准确度和响应速度，为百思买增加了数千万美元的利润。

上面是一个非常典型的大数据分析案例，通过数据分析有效地帮助零售商提高了商品销售额并带来了非常可观的利润。那么到底什么是数据分析呢？

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，对它们加以汇总、理解并消化，以求最大限度地利用数据所包含的信息，发挥数据的作用。数据分析是为了提取有用信息并得出结论而对数据加以详细研究和概括总结的过程。

数据分析的数学基础在20世纪早期就已确立，但直到计算机出现时才得以实际应用，并使得数据分析得以推广。因此，数据分析是数学与计算机科学相结合的产物。

数据分析的目的是把隐藏在一大批杂乱无章的数据中的有用信息提炼出来，从而找出所研究对象的内在规律。在实际应用中，数据分析可帮助人们做出判断，以便采取适当行动。数据分析是有组织、有目的地收集数据、分析数据，使之成为有用信息的过程。这一过程是质量管理体系的支持过程。例如，在某产品的整个生命周期，从市场调研、售后服务到最终处置的各个过程都需要适当进行数据分析，以提升有效性。再如，设计人员在设计一个新产品之前，要通过广泛的设计调查，分析所得数据以判定设计方向。因此，数据分析在工业设计中具有极其重要的地位。如今，数据分析技术不断发展，促使企业在管理方面做到科学务实、脚踏实地，进而做出正确、合理的决策。

1.2 数据分析与处理的概念

在对数据分析与处理有一个大致的了解之后，本节对数据分析与处理的概念加以介绍。

1.2.1 什么是数据

数据(data)，在拉丁文中是“已知”的意思，代表对某件事物的描述，一般指描述事物的符号记录，如图形、声音、文字、数值等，是构成信息和知识的原始材料。例如，数据60代表的信息可以是某个人的体重为60公斤，也可以是一个同学某门功课的成绩为60分，还可以是课堂学生人数为60人。

数据的表现形式还不能完全表达其内容，需要经过解释。数据的解释是指对数据含义的说明，数据的含义称为数据的语义，数据与其语义是密不可分的。

在计算机科学中，数据是指能够输入电子计算机并被计算机程序处理的具有一定意义的数字、字母、符号和模拟量等的通称。计算机存储和处理的对象十分广泛，表示这些对象的数据也随之变得越来越复杂。

1.2.2 什么是数据的分析与处理

数据的分析与处理，包括数据分析和数据处理两个部分。数据分析的基本目的是从大量的、杂乱无章的、难以理解的数据中抽取并推导出对某些特定的人们来说有价值、有意义的数据。数据处理是指对数据进行采集、存储、检索、加工、转换和传输。数据分析与处理是系统工程和自动控制的基本环节，它贯穿于社会生产和社会生活的各个领域。数据分析与处理技术的发展及其应用，极大地影响着人类社会发展的进程。

计算机数据处理主要包括以下8个方面。

(1) 数据采集，采集所需的信息。

(2) 数据转换，把信息转换成机器能够接收的形式。

(3) 数据分组，指定编码，按有关信息进行有效的分组。

(4) 数据组织，整理数据或用某些方法安排数据，以便进行处理。

(5) 数据计算，进行各种算术和逻辑运算，以便得到进一步的信息。

(6) 数据存储，将原始数据或计算的结果保存起来，供以后使用。

(7) 数据检索，按用户的要求找出有用的信息。

(8) 数据排序，把数据按一定要求排出次序。

数据的复杂性使得人们难以用传统的方法对其进行描述与度量，需要将高维图像等多媒体数据降维后进行度量与处理，利用上下文关联进行语义分析，从大量动态或模棱两可的数据中提取信息，导出可理解的内容，过程中可能会运用统计和分析、机器学习、数据挖掘等技术。因此，我们要注重分析数据的相关关系，而不是因果关系。

数据分析与处理的结果应是可视化的、直观的，以便于洞察。目前，尽管计算机智能化有了很大进步，但只能针对小规模、有结构或类结构的数据进行分析，无法实现深层次的数据挖掘，现有的数据挖掘算法在不同行业中难以通用。

1.3 数据分析与处理的过程

进行数据分析的主要目的是让数据说话，作为行动的向导，以提供决策的依据。那么，如何进行有效的数据分析呢？数据分析与处理的过程又是怎样的呢？

1.3.1 数据分析与处理的实现过程

在进行数据分析时，运用统计方法应遵循如下原则：坚持用数据说话的基本观点；有目的地收集数据；掌握数据的来源；认真整理数据。

1. 数据分析流程

数据分析流程主要包括五大步骤，如图 1.1 所示。

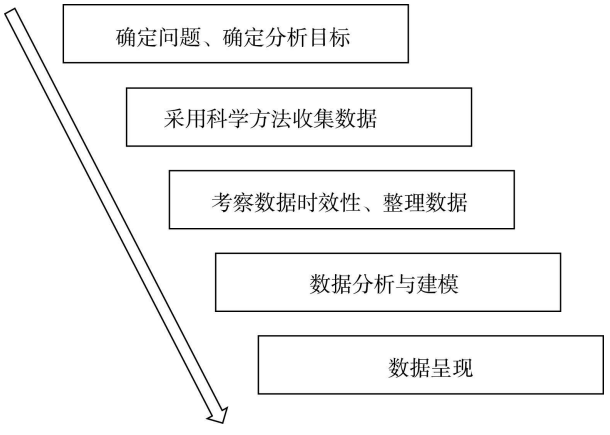


图 1.1　数据分析流程

第一步，确定问题、确定分析目标。比较典型的场景是针对企业的数据进行分析。例如，企业通常会有销售数据、用户数据、运营数据、产品生产数据等，我们需要从这些数据中获得哪些有用的信息，以对策略的制定进行指导呢？又如，我们需要做一份市场调研或行业分析，那么我们需要获得关于这个行业的哪些信息呢？

第二步，采用科学方法收集数据。数据可以通过网络爬虫爬取、本地数据导出、物联网设备采

集或人工录入等方式采集。在确定了分析目标后，要根据客户的需求，构建数据源并采集数据。

第三步，考察数据时效性、整理数据。现实世界中的数据大多是不完整、不一致的脏数据，无法直接进行数据分析，或者分析结果不尽如人意，把这些影响分析的脏数据处理好，才能获得更加精确的分析结果。数据预处理有多种方法：数据清理、数据集成、数据转换、数据归约等。

第四步，数据分析与建模。数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，对数据加以详细研究和概括总结，提取有用信息，形成结论的过程。这一过程也是质量管理体系的支持过程。在实际生活中，数据分析可帮助人们做出判断，以便采取适当的行动。数据模型是对信息系统中客观事物及其联系的数据描述，它描述了复杂数据之间的整体逻辑结构关系。数据模型不但提供了整个组织赖以收集数据的基础，它还与组织中的其他模型一起，精确恰当地记录业务需求，并支持信息系统不断发展和完善，以满足不断变化的业务需求。

第五步，数据呈现。出具分析报告，提出解决意见或建议。分析结果最直接的形式是统计量的描述和统计量的展示。数据分析报告不仅是对分析结果的直接呈现，还是对相关情况的全面认识。

2. 数据的可视化

数据可视化起源于 20 世纪 60 年代出现的计算机图形学，人们使用计算机创建图形图表，将提取出来的数据进行可视化，并将数据的各种属性和变量呈现出来。随着计算机硬件的发展，人们创建了更复杂、规模更大的数字模型，发展了数据采集设备和数据保存设备，同样也需要更高级的计算机图形学技术及方法来创建这些规模庞大的数据集。随着数据可视化平台的拓展、应用领域的增加、表现形式的不断变化，增加了实时动态效果展示、用户交互使用等功能，数据可视化像所有新兴概念一样，其边界不断扩大。

数据可视化是关于数据视觉表现形式的科学技术研究。这种数据的视觉表现形式被定义为一种以某种概要形式抽提出来的信息，包括相应信息单位的各种属性和变量。图 1.2(本图来源于互联网)为某商城停车场的停车收费信息可视化，用不同的图表展示了停车信息与收费信息的统计结果。

图 1.2　某商城停车场的停车收费信息可视化

数据可视化旨在借助图形化手段，清晰有效地传达与沟通信息。但是，这并不意味着数据可视化就一定因为要实现可视化功能而令人感到枯燥乏味，或者是为了看上去绚丽多彩而显得极其复杂。为了有效地传达思想观念，美学形式与功能需要齐头并进，通过直观地传达关键的方面与特征，从而实现对于相对稀疏而又复杂的数据集的深入洞察。然而，设计人员往往并不能很好地把握设计与功能之间的平衡，从而得到的是华而不实的数据可视化形式，无法达到其主要目的，也就无法传达与沟通信息。

人们熟悉的饼图、直方图、散点图、柱状图等是原始的统计图表，它们是数据可视化的基础和常见应用。作为一种统计学工具，它们可用于创建一条快速认识数据集的捷径，并成为一种令人信服的沟通手段。

数据可视化技术包含以下几个基本概念。

(1) 数据空间，指由 n 维属性和 m 个元素组成的数据集所构成的多维信息空间。

(2) 数据开发，指利用一定的算法和工具对数据进行定量的推演和计算。

(3) 数据分析，指对多维数据进行切片、切块、旋转等动作，剖析数据，从而能多角度多侧面地观察数据。

(4) 数据可视化，指将大型数据集中的数据以图形图像的形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程。

针对数据可视化，人们已经提出了许多方法，这些方法根据其可视化的原理可以划分为基于几何的技术、面向像素的技术、基于图标的技术、基于层次的技术、基于图像的技术和分布式技术等。

3. 某化妆品公司销售和宣传数据分析

下面通过一个简单的例子来说明数据分析与处理的过程。

1) 背景材料

某化妆品公司专注于青年女性消费者这一核心客户群体，目前正在尝试通过增加社会网络广告投入来扩大市场影响力，然而，这个新做法是否成功尚待观察。该公司的产品在青年女性消费者中的销售潜力巨大，请从分析师的角度进行分析，得出提高产品销量的方法。

2) 数据整理

表 1.1 是收集整理的该公司某年 1 月份到 6 月份的产品销售数据汇总表。

表 1.1　产品销售数据汇总表

某化妆品公司产品销售数据汇总						
月份	目标销售额	总销售额	广告费	社会网络费	总销量	单价
1 月	$5,290,000	$5,280,000.00	$1,056,600	$0	2,640,000	$2.00
2 月	$5,600,000	$5,499,000.00	$950,500	$105,600	2,749,500	$2.00
3 月	$5,729,000	$5,469,000.00	$739,200	$316,900	2,734,500	$2.00
4 月	$5,968,000	$5,480,380.90	$528,000	$528,000	2,884,411	$1.90
5 月	$6,217,000	$5,532,999.50	$316,800	$739,200	2,912,105	$1.90
6 月	$6,480,000	$5,554,000.20	$316,800	$739,200	2,923,158	$1.90

3) 数据图表分析

从表 1.1 中很难看出问题所在，因此需要对此表进行图表分析，如图 1.3 所示。

4) 主观分析结果呈现

根据数据图表分析可以得出以下结论。

(1) 6 月份的销量相对 1 月份的销量略有上升，但是成绩平平。
(2) 总销售额从 3 月份开始与目标销售额拉开差距。
(3) 4 月份之后降价，总销售额依然无明显提高。
(4) 总销售额未达到目标销售额可能与广告费用调整有关。
(5) 降价无益于总销售额达标。

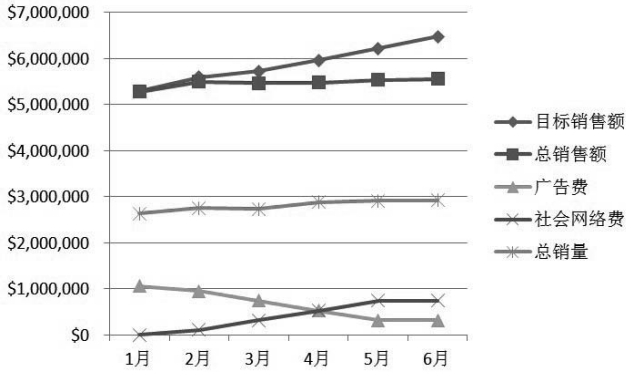


图 1.3　数据图表分析

根据以上分析给出建议：将宣传费用比例调整至 1 月份的水平，观察后续的效果。

1.3.2　数据分析与处理的案例

【案例 1.1】利用 Excel 录入各种类型的数据，并对数据进行有效性设置。

1. 案例的数据描述

现有包含学号、姓名、出生日期、班级、三门科目成绩、总分等信息的学生信息表(见表 1.2)。试对表中的数据进行有效性检查：学号为 10 位长度的文本，出生日期为 2000/1/1 至 2010/1/1 之间，各科成绩值范围在 0～100。

表 1.2　学生信息表

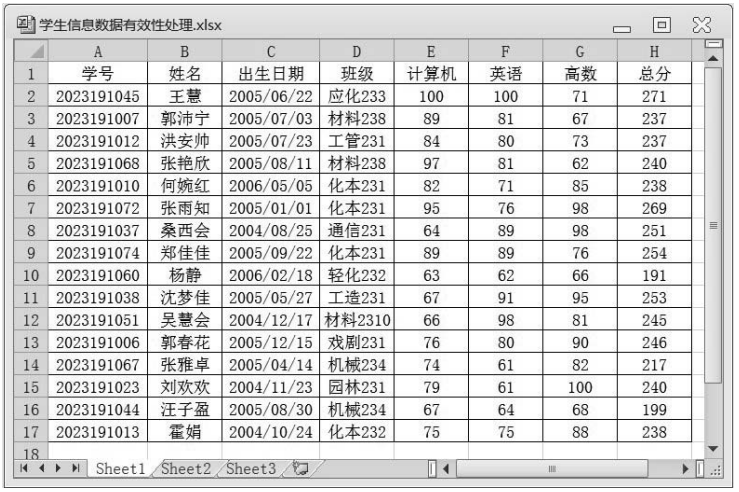
学号	姓名	出生日期	班级	计算机	英语	高数	总分
2023191045	王慧	2005/6/22	应化 233	100	100	71	271
2023191007	郭沛宁	2005/7/3	材料 238	89	81	67	237
2023191012	洪安帅	2005/7/23	工管 231	84	80	73	237
2023191068	张艳欣	2005/8/11	材料 238	97	81	62	240
2023191010	何婉红	2006/5/5	化本 231	82	71	85	238
2023191072	张雨知	2005/1/1	化本 231	95	76	98	269
2023191037	桑西会	2004/8/25	通信 231	64	89	98	251
2023191074	郑佳佳	2005/9/22	化本 231	89	89	76	254
2023191060	杨静	2006/2/18	轻化 232	63	62	66	191
2023191038	沈梦佳	2005/5/27	工造 231	67	91	95	253
2023191051	吴慧会	2004/12/17	材料 2310	66	98	81	245
2023191006	郭春花	2005/12/15	戏剧 231	76	80	90	246

(续表)

学号	姓名	出生日期	班级	计算机	英语	高数	总分
2023191067	张雅卓	2005/4/14	机械 234	74	61	82	217
2023191023	刘欢欢	2004/11/23	园林 231	79	61	100	240
2023191044	汪子盈	2005/8/30	机械 234	67	64	68	199
2023191013	霍娟	2004/10/24	化本 232	75	75	88	238

2. 案例的操作步骤

(1) 新建一个 Excel 工作簿，将其命名为“学生信息数据有效性处理”，并在工作表 Sheet1 中输入相应的文字和数据，如图 1.4 所示。



	A	B	C	D	E	F	G	H
1	学号	姓名	出生日期	班级	计算机	英语	高数	总分
2	2023191045	王慧	2005/06/22	应化233	100	100	71	271
3	2023191007	郭沛宁	2005/07/03	材料238	89	81	67	237
4	2023191012	洪安帅	2005/07/23	工管231	84	80	73	237
5	2023191068	张艳欣	2005/08/11	材料238	97	81	62	240
6	2023191010	何婉红	2006/05/05	化本231	82	71	85	238
7	2023191072	张雨知	2005/01/01	化本231	95	76	98	269
8	2023191037	桑西会	2004/08/25	通信231	64	89	98	251
9	2023191074	郑佳佳	2005/09/22	化本231	89	89	76	254
10	2023191060	杨静	2006/02/18	轻化232	63	62	66	191
11	2023191038	沈梦佳	2005/05/27	工造231	67	91	95	253
12	2023191051	吴慧会	2004/12/17	材料2310	66	98	81	245
13	2023191006	郭春花	2005/12/15	戏剧231	76	80	90	246
14	2023191067	张雅卓	2005/04/14	机械234	74	61	82	217
15	2023191023	刘欢欢	2004/11/23	园林231	79	61	100	240
16	2023191044	汪子盈	2005/08/30	机械234	67	64	68	199
17	2023191013	霍娟	2004/10/24	化本232	75	75	88	238
18								

图 1.4　输入原始数据

(2) 选定单元格区域“A2:A17”，选择“数据”选项卡，单击“数据工具”组中的“数据有效性”选项下的“数据有效性”按钮。在打开的“数据有效性”对话框的“设置”选项卡中进行有效性的设置：“允许”为“文本长度”；“数据”为“等于”；“长度”为“10”，如图 1.5 所示。设置完成单击“确定”按钮。

图 1.5　学号有效性设置

(3) 选定单元格区域“C2:C17”，选择“数据”选项卡，单击“数据工具”组中的“数据有效性”选项下的“数据有效性”按钮。在打开的“数据有效性”对话框的“设置”选项卡中进行有效性的设置：“允许”为“日期”；“数据”为“介于”；“开始日期”为“2000/1/1”；“结束日期”为“2010/1/1”，如图 1.6 所示。设置完成单击“确定”按钮。

(4) 选定单元格区域“E2:G17”，选择“数据”选项卡，单击“数据工具”组中的“数据有效性”选项下的“数据有效性”按钮。在打开的“数据有效性”对话框的“设置”选项卡中进行有效性的设置：“允许”为“整数”；“数据”为“介于”；“最小值”为“0”；“最大值”为“100”，如图 1.7 所示。设置完成单击“确定”按钮。

图 1.6　出生日期有效性设置

图 1.7　各科成绩有效性设置

3. 案例的结果分析

本案例依次设置了文本的长度、日期的起止及整数的有效范围。数据的有效性设置可以对数据自动进行检测，提高了数据输入或检查时的正确率。

【案例 1.2】对数据进行分列操作。

1. 案例的数据描述

现有三组数据，如表 1.3～表 1.5 所示。试对第一组数据进行固定宽度分列操作；对第二组数据进行分隔符号分列操作；对于第三组数据，可借助分列操作完成日期型数据的转换，如将日期格式“2023.1.1”转换为 2023/1/1 或 2023-1-1。

表 1.3　固定宽度分列原数据

人名	年假	薪水
甲	12	50,000
乙	20	50,000
丙	10	40,000

表 1.4　分隔符号分列原数据

北京市海淀区中关村东路 1 号院
哈尔滨市呼兰区公证处
齐齐哈尔市龙沙区光复街 5 号
北京市海淀区中关村东路 2 号院
哈尔滨市呼兰区公证处
齐齐哈尔市龙沙区光复街 6 号
北京市海淀区中关村东路 3 号院
哈尔滨市呼兰区公证处
齐齐哈尔市龙沙区光复街 7 号
北京市海淀区中关村东路 4 号院

表 1.5　分列完成数据转换

日期格式
2023.1.10
2023.1.11
2023.1.12
2023.1.13
2023.1.14
2023.1.15
2023.1.16
2023.1.17
2023.1.18
2023.1.19
2023.1.20
2023.1.21
2023.1.22
2023.1.23
2023.1.24
2023.1.25
2023.1.26

2. 案例的操作步骤

1) 对数据进行固定宽度分列操作

(1) 新建一个 Excel 工作簿，将其命名为“固定宽度分列”，并在表格中输入相应的文字和数据，如图 1.8 所示。

图 1.8　输入原始数据

(2) 选定单元格区域“A1:A4”，选择“数据”选项卡，单击“数据工具”组中的“分列”按钮，打开“文本分列向导-第 1 步，共 3 步”对话框，选择最合适的文件类型“固定宽度”，单击“下一步”按钮，如图 1.9 所示。

图 1.9　文本分列向导第 1 步

(3) 在“文本分列向导-第 2 步，共 3 步”对话框的“数据预览”中，首先在要建立分列处单击，然后按住分列线(有箭头的垂直线)将其拖至指定位置，即设置字段宽度(列间距)，最后单击“下一步”按钮，如图 1.10 所示。若要清除分列线，则可双击分列线。

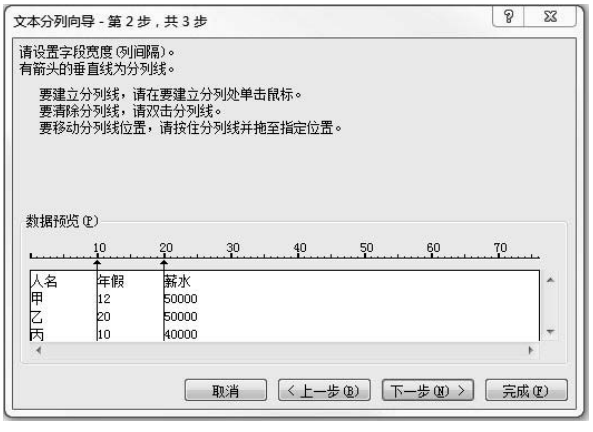


图 1.10　文本分列向导第 2 步

(4) 在“文本分列向导-第 3 步，共 3 步”对话框中：选择“人名”列，选择“列数据格式”为“文本”；选择“年假”列，选择“列数据格式”为“常规”；选择“薪水”列，选择“列数据格式”为“常规”；单击“高级”按钮，可进一步设置数据的数字格式；设置“目标区域”为A1；单击“完成”按钮，如图 1.11 所示。

分列后的结果如图 1.12 所示。

图 1.11　文本分列向导第 3 步

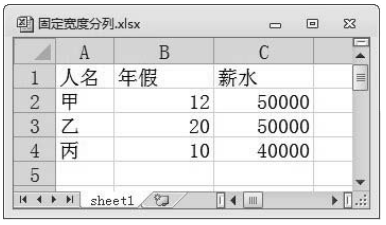


图 1.12　分列结果

2) 对数据进行分隔符号分列操作

(1) 新建一个 Excel 工作簿，将其命名为“分隔符号分列”，并在表格中输入相应的文字和数据，如图 1.13 所示。

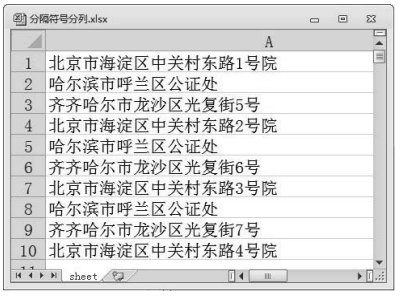


图 1.13　输入原始数据

(2) 选定单元格区域“A1:A10”，选择“数据”选项卡，单击“数据工具”组中的“分列”按钮，打开“文本分列向导-第 1 步，共 3 步”对话框，选择最合适的文件类型“分隔符号”，单击“下一步”按钮，如图 1.14 所示。

图 1.14　文本分列向导第 1 步

(3) 在“文本分列向导-第 2 步，共 3 步”对话框中，选择“分隔符号”为“其他”，填写“市”字，以“市”字进行分列，单击“下一步”按钮，如图 1.15 所示。

图 1.15　文本分列向导第 2 步

(4) 在“文本分列向导-第 3 步，共 3 步”对话框中，选择第 1 列数据，选择“列数据格式”为“文本”，设置“目标区域”为A1，单击“完成”按钮，如图 1.16 所示。

图 1.16　文本分列向导第 3 步

分列后的结果如图 1.17 所示。

如果想进一步将 B 列中的“区”分列出来，则选定单元格区域“B1:B10”，重复以上文本分列向导的 3 个步骤，在第 2 步中填写“区”字进行分列。分列后可在最上面加入一行，写入各列字段名称。最终分列结果如图 1.18 所示。

图 1.17　按“市”分列的结果

图 1.18　按“市”“区”分列的结果

3) 借助分列操作完成日期型数据的转换

(1) 新建一个 Excel 工作簿，将其命名为“分列转换数据”，并在表格中输入相应的文字和数据，如图 1.19 所示。

(2) 选定单元格区域“A2:A18”，选择“数据”选项卡，单击“数据工具”组中的“分列”按钮，打开“文本分列向导-第 1 步，共 3 步”对话框。由于要进行数据类型转换，因此直接单击两次“下一步”按钮。在“文本分列向导-第 3 步，共 3 步”对话框中，选择“列数据格式”为“日期—YMD”类型，设置“目标区域”为A2，单击“完成”按钮，结束类型转换，转换结果如图 1.20 所示。如果想转换为“2023-1-1”的格式，则需要先在系统中设置“区域和语言”中的日期和时间的默认格式为“yyyy-m-d”，再使用分列方式进行类型转换。

分列转换数据.xlsx

	A
1	日期格式
2	2023. 1. 10
3	2023. 1. 11
4	2023. 1. 12
5	2023. 1. 13
6	2023. 1. 14
7	2023. 1. 15
8	2023. 1. 16
9	2023. 1. 17
10	2023. 1. 18
11	2023. 1. 19
12	2023. 1. 20
13	2023. 1. 21
14	2023. 1. 22
15	2023. 1. 23
16	2023. 1. 24
17	2023. 1. 25
18	2023. 1. 26

sheet1

图 1.19　输入原始数据

分列转换数据.xlsx

	A
1	日期格式
2	2023/1/10
3	2023/1/11
4	2023/1/12
5	2023/1/13
6	2023/1/14
7	2023/1/15
8	2023/1/16
9	2023/1/17
10	2023/1/18
11	2023/1/19
12	2023/1/20
13	2023/1/21
14	2023/1/22
15	2023/1/23
16	2023/1/24
17	2023/1/25
18	2023/1/26

sheet1

图 1.20　日期类型数据分列转换结果

3. 案例的结果分析

本案例实现了按分隔符号分列、按固定宽度分列和数据转换。分隔符号分列适用于数据源带有某些特定符号(如逗号、冒号、空格等)的情况，汉字也可以作为分隔符来使用。以固定宽度分列主要适用于数据源比较整齐划一、排列比较有规律的数据分列情况。

【案例 1.3】根据公司提供的销售数据进行数据分析，为公司提出改进销售的方案。

1. 案例的数据描述

根据某化妆品公司提供的近 6 个月的产品销售情况表(见表 1.6)、宣传投入表(见表 1.7)和目标销售额表(见表 1.8)进行数据分析，从而为公司提出改进销售的方案。

表 1.6　产品销售情况表

某化妆品公司产品销售情况				
月份	销售地区	单价	销售数量	销售额
1 月	销售一区	$2.0	904,000	$1,808,000.0
1 月	销售二区	$2.0	744,000	$1,488,000.0
1 月	销售三区	$2.0	992,000	$1,984,000.0

(续表)

某化妆品公司产品销售情况				
月份	销售地区	单价	销售数量	销售额
2 月	销售一区	$2.0	945,500	$1,891,000.0
2 月	销售二区	$2.0	760,000	$1,520,000.0
2 月	销售三区	$2.0	1,044,000	$2,088,000.0
3 月	销售一区	$2.0	953,000	$1,906,000.0
3 月	销售二区	$2.0	752,500	$1,505,000.0
3 月	销售三区	$2.0	1,029,000	$2,058,000.0
4 月	销售一区	$1.9	986,300	$1,873,970.0
4 月	销售二区	$1.9	768,000	$1,459,200.0
4 月	销售三区	$1.9	1,130,111	$2,147,210.9
5 月	销售一区	$1.9	965,100	$1,833,690.0
5 月	销售二区	$1.9	745,005	$1,415,509.5
5 月	销售三区	$1.9	1,202,000	$2,283,800.0
6 月	销售一区	$1.9	961,080	$1,826,052.0
6 月	销售二区	$1.9	750,078	$1,425,148.2
6 月	销售三区	$1.9	1,212,000	$2,302,800.0

表 1.7　宣传投入表

某化妆品公司产品宣传投入		
月份	广告费	社会网络费
1 月	$1,056,600	$0
2 月	$950,500	$105,600
3 月	$739,200	$316,900
4 月	$528,000	$528,000
5 月	$316,800	$739,200
6 月	$316,800	$739,200

表 1.8　目标销售额表

某化妆品公司产品目标销售额	
月份	目标销售额
1 月	$5,290,000
2 月	$5,600,000
3 月	$5,729,000
4 月	$5,968,000
5 月	$6,217,000
6 月	$6,480,000

2. 案例的操作步骤

(1) 新建一个 Excel 工作簿，将其命名为“某化妆品销售分析”，并依次建立“产品销售情况”工作表、“宣传投入”工作表和“目标销售额”工作表，并在表格中输入相应的文字和数据，如

图 1.21～图 1.23 所示。

某化妆品销售分析.xls [兼容模式]

	A	B	C	D	E
1	某化妆品公司产品销售情况				
2	月份	销售地区	单价	销售数量	销售额
3	1月	销售一区	$2.0	904000	$1,808,000.0
4	1月	销售二区	$2.0	744000	$1,488,000.0
5	1月	销售三区	$2.0	992000	$1,984,000.0
6	2月	销售一区	$2.0	945500	$1,891,000.0
7	2月	销售二区	$2.0	760000	$1,520,000.0
8	2月	销售三区	$2.0	1044000	$2,088,000.0
9	3月	销售一区	$2.0	953000	$1,906,000.0
10	3月	销售二区	$2.0	752500	$1,505,000.0
11	3月	销售三区	$2.0	1029000	$2,058,000.0
12	4月	销售一区	$1.9	986300	$1,873,970.0
13	4月	销售二区	$1.9	768000	$1,459,200.0
14	4月	销售三区	$1.9	1130111	$2,147,210.9
15	5月	销售一区	$1.9	965100	$1,833,690.0
16	5月	销售二区	$1.9	745005	$1,415,509.5
17	5月	销售三区	$1.9	1202000	$2,283,800.0
18	6月	销售一区	$1.9	961080	$1,826,052.0
19	6月	销售二区	$1.9	750078	$1,425,148.2
20	6月	销售三区	$1.9	1212000	$2,302,800.0

产品销售情况 / 宣传投入 / 目标销

图 1.21 “产品销售情况”工作表

某化妆品销售分析.xls [兼容模式]

	A	B	C
1	某化妆品公司产品宣传投入		
2	月份	广告费	社会网络费
3	1月	$1,056,600	$0
4	2月	$950,500	$105,600
5	3月	$739,200	$316,900
6	4月	$528,000	$528,000
7	5月	$316,800	$739,200
8	6月	$316,800	$739,200
9			

宣传投入 / 目标销售

图 1.22 “宣传投入”工作表

(2) 按月份汇总销售额。将各月份的目标销售额、总销售额、广告费、社会网络费、总销量、单价等合并到一个表格中。插入一张新的工作表(按 Shift+F11 键)，将其命名为“销售数据汇总”，选定“目标销售额”工作表中的单元格区域“A2:B8”并复制到“销售数据汇总”工作表的“A2:B8”区域。

某化妆品销售分析.xls [兼容模式]

	A	B
1	某化妆品公司产品目标销售额	
2	月份	目标销售额
3	1月	$5,290,000
4	2月	$5,600,000
5	3月	$5,729,000
6	4月	$5,968,000
7	5月	$6,217,000
8	6月	$6,480,000

目标销售额

图 1.23 “目标销售额”工作表

(3) 在“产品销售情况”工作表中增加一列“总销量”，并在“F3”“F6”“F9”“F12”“F15”和“F18”单元格内利用求和函数 sum()分别计算 1 月份至 6 月份 3 个销售区的产品总销量，如图 1.24 所示。将“总销量”列中的数据复制到“销售数据汇总”工作表的“F2:F8”区域。

(4) 在“产品销售情况”工作表中增加一列“总销售额”，并在“G3”“G6”“G9”“G12”“G15”和“G18”单元格内，利用求和函数 sum()分别计算 1 月份至 6 月份 3 个销售区的产品总销售额，如图 1.24 所示。将“总销售额”列中的数据复制到“销售数据汇总”工作表的“C2:C8”区域。

某化妆品销售分析.xls [兼容模式]

	A	B	C	D	E	F	G
1	某化妆品公司产品销售情况						
2	月份	销售地区	单价	销售数量	销售额	总销量	总销售额
3	1月	销售一区	$2.0	904000	$1,808,000.0		
4	1月	销售二区	$2.0	744000	$1,488,000.0	2640000	$5,280,000.0
5	1月	销售三区	$2.0	992000	$1,984,000.0		
6	2月	销售一区	$2.0	945500	$1,891,000.0		
7	2月	销售二区	$2.0	760000	$1,520,000.0	2749500	$5,499,000.0
8	2月	销售三区	$2.0	1044000	$2,088,000.0		
9	3月	销售一区	$2.0	953000	$1,906,000.0		
10	3月	销售二区	$2.0	752500	$1,505,000.0	2734500	$5,469,000.0
11	3月	销售三区	$2.0	1029000	$2,058,000.0		
12	4月	销售一区	$1.9	986300	$1,873,970.0		
13	4月	销售二区	$1.9	768000	$1,459,200.0	2884411	$5,480,380.9
14	4月	销售三区	$1.9	1130111	$2,147,210.9		
15	5月	销售一区	$1.9	965100	$1,833,690.0		
16	5月	销售二区	$1.9	745005	$1,415,509.5	2912105	$5,532,999.5
17	5月	销售三区	$1.9	1202000	$2,283,800.0		
18	6月	销售一区	$1.9	961080	$1,826,052.0		
19	6月	销售二区	$1.9	750078	$1,425,148.2	2923158	$5,554,000.2
20	6月	销售三区	$1.9	1212000	$2,302,800.0		

产品销售情况 / 宣传投入 / 目标销售额 / 销售数据汇总

图 1.24 3 个销售区的产品总销量和产品总销售额

(5) 将“宣传投入”工作表中“B2:C8”区域中的数据复制到“销售数据汇总”工作表的“D2:E8”区域。

(6) 在“销售数据汇总”工作表的“G2:G8”区域增加一列“单价”，将“产品销售情况”工作表中的单价复制到“G3”至“G8”单元格中。最终的“销售数据汇总”工作表如图 1.25 所示。

某化妆品销售分析.xls [兼容模式]

	A	B	C	D	E	F	G
1	某化妆品公司产品销售数据汇总						
2	月份	目标销售额	总销售额	广告费	社会网络费	总销量	单价
3	1月	$5,290,000	$5,280,000.0	$1,056,600	$0	2640000	$2.00
4	2月	$5,600,000	$5,499,000.0	$950,500	$105,600	2749500	$2.00
5	3月	$5,729,000	$5,469,000.0	$739,200	$316,900	2734500	$2.00
6	4月	$5,968,000	$5,480,380.9	$528,000	$528,000	2884411	$1.90
7	5月	$6,217,000	$5,532,999.5	$316,800	$739,200	2912105	$1.90
8	6月	$6,480,000	$5,554,000.2	$316,800	$739,200	2923158	$1.90

宣传投入 | 目标销售额 | 销售数据汇总

图 1.25　“销售数据汇总”工作表

(7) 绘制目标销售额、总销售额、广告费、社会网络费、总销量折线图。选择“销售数据汇总”工作表“A2:F8”单元格区域，在“插入”选项卡中，执行“图表”→“折线图”→“带数据标记的折线图”命令，生成的图表如图 1.3 所示。

3. 案例的结果分析

根据折线图分析其中存在的关系如下。

(1) 2 月份的总销售额相对 1 月份的总销售额略有上升，但是成绩平平。

(2) 总销售额从 3 月份开始与目标相去甚远。

(3) 广告费用和社交网络费总量持平，但是广告费逐渐减少，社交网络费逐渐增加。

(4) 4 月份后价格下降，但是总销售额没有明显变动。

根据以上分析得出结论：可能是广告费用和社交网络费的调整影响了销量，因此建议将宣传费用比例调整至 1 月份的水平，观察后续的效果。

1.3.3　数据分析师

数据分析师是数据师的一种，指的是不同行业中，专门从事行业数据收集、整理、分析，并依据数据进行行业研究、评估和预测的专业人员。

1. 数据分析师的作用

这是一个用数据说话的时代，也是一个依靠数据竞争的时代。世界 500 强企业中，有 90%以上的企业都建立了数据分析部门。IBM、微软、Google 等知名公司都积极投资数据业务，建立数据部门，培养数据分析团队。各国政府和越来越多的企业意识到数据和信息已经成为企业的智力资产和资源，数据的分析和处理能力正逐渐成为关键的技术手段。

互联网本身具有数字化和互动性的特征，这种特征给数据收集、整理、研究带来了革命性的突破。以往“原子世界”中数据分析师要花较高的成本(资金、资源和时间)获取支撑研究、分析的数据，数据的丰富性、全面性、连续性和及时性都比互联网时代差很多。

与传统的数据分析师相比，互联网时代的数据分析师面临的情况不是数据匮乏，而是数据过剩。因此，互联网时代的数据分析师必须学会借助技术手段进行高效的数据处理。更为重要的是，互联

网时代的数据分析师要不断在数据研究的方法论方面进行创新和突破。

2. 企业数据分析师

一个企业的数据分析师不仅是一个数据分析者，还应是帮助决策者做出正确决定的人。那么，一个好的企业数据分析师应该帮助企业解决哪些问题？答案如下。

(1) 轻视数据。让决策者接纳、理解数据分析的结果并不是一件简单的事情。数据分析师(无论是金融工程师还是数据专家)要避免决策者对数据产生轻视的现象。在现实中，决策者必须寻求的是“多重变量的一致性直觉判断与数据分析”。换句话说，决策者应该以一种互补的方式，综合利用数据分析和直觉观察判断来形成一个整体观点，而不是过于依赖数据或仅仅通过观察。

(2) 决策偏差。无论是数据分析师还是决策者，都需要掌握直觉观察和数据分析的平衡，数据分析师必须意识到，任何观点，哪怕是来自所有人的认同和铁一般的事实，都可能带有潜在的偏差。另一个需要注意的问题是情感偏见。例如，当一个决策者暴露在公众视野中，长期被“粉丝”、客户和社交媒体包围，面对舆论监督和媒体审查等，这些外部的噪声必然会对决策产生某种影响，这时就产生了情感偏见，这正是决策者需要规避的。

(3) 精准沟通。另一个重要的主题是沟通方式的重要性。决策者正在试图寻求更明确的方法来接受数据分析师的想法。他们希望数据分析师能用简单易懂的语言和他们对话，从而可以轻松把握数据的意义。

(4) 数据转化。很多人指出，决策者们并“不熟悉”数据分析之类的科学方法。因此，必须改变对话策略，实现数据分析师和决策者的双向对话，即进行有效的数据转化，这将促进两种不同思维之间的相互理解。

3. 企业数据分析师应具备的技能

数据分析师应该拥有如下技能。

(1) 具有较高的业务水平和地位，能够和决策者进行平等沟通。

(2) 掌握数据分析知识或有愿意学习这类知识的动力，并能够与数据工程师进行有效沟通。

(3) 能从容地向高管、同事和下属传递和表达想法。

(4) 具有很强的学习能力。

(5) 能够理解并转化同事们的问题和解决办法。

(6) 对质量标准的高要求和对细节的关注。

(7) 具有独立的组织能力，能够凝聚高管和工程师。

数据分析师可以通过举例进行类比，让决策者产生共鸣，或者向决策者表达问题，而不是结论。特别是对那些持怀疑态度的决策者，要注意不要一开始就表现得过分自信。数据分析师可以通过数据分析提出问题，让决策者自己去想答案，而不是直接告诉他们。通过数据分析师对数据的成功转化，可以弥补决策者和数据工程师之间的沟通障碍，进而解决数据和现实之间的差异。

1.4 数据分析模型

在进行数据分析时，肯定要用到数据分析模型。在进行数据分析之前，需要先搭建数据分析模型，根据模型中的内容，对不同的数据指标进行细化分析，最终得到想要的分析结果或结论。

数据模型就是对现实世界进行抽象化的数据展示，它在满足抽象的同时，越简单越好。统计数据视角下的数据模型通常指的是统计分析、大数据挖掘、深度学习、人工智能技术等实体模型，这些模型是从科学研究视角来界定的。数据模型分析主要包括以下几方面。

1. 异常数据检测

现实世界的数据一般是不完整的、有噪声的、不一致的和冗余的。预处理就是对数据进行预先处理，以提高数据质量。数据集中的异常数据通常被称为异常点、异常值或孤立点等。这些数据的特征或规则与大多数数据不一致，表现出“异常”的特征。检测这些数据的方法称为异常检测。

噪声数据是指一个测量变量中的随机错误或偏离期望的孤立点值。产生噪声的原因很多，如数据输入时的人为错误或计算机错误、网络传输中的错误、数据收集设备的故障等。例如，在输入工资值时，输入了-6368.00，这明显是一个错误的数据。

不完整数据是指实际应用系统中，由于系统设计得不合理或使用过程中的某些因素，某些属性值缺失或值不确定的数据。

不一致数据的产生通常是由于原始数据来源于多个不同的应用系统或数据库，这些数据的信息来源多样，采集和加工的方法有别，数据描述的格式也各不相同，缺乏统一的分类标准和信息的编码方案，因此难以实现信息的集成共享，很难直接用于数据挖掘。例如，“年龄：20”也可以表示为“生日：2004 年 10 月 1 日”；不同的惯用语“齐齐哈尔大学”或“齐大”都代表“齐齐哈尔大学”；不同的计量单位“50KG”“50 公斤”“110 磅”都代表体重。对于这种不一致的数据需要进行预处理，制定统一的标准和编码方案。

同一事物在数据库中存在两条或多条完全相同的记录，或者相同的信息冗余地存在于多个数据源中，称为重复数据。原始数据中通常记录着事物的较为全面的属性，而在数据分析中，这些属性并不是都有用，只需要一部分属性即可得到有用的信息，而且无用属性的增加还会导致无效归纳，把分析结果引向错误的结论。

2. 数据降维

在对大量的数据进行数据挖掘时，往往会面临“维度灾害”。数据集的维度在无限地增加，但计算机的处理能力有限，而且数据集的多个维度之间可能存在共同的线性关系，这会造成学习模型的可扩展性不足，乃至许多时候优化算法的结果无效。因此，人们必须减少维度总数并减少各维度间的共线性危害。

数据降维也称为数据归约或数据约减，进行数据降维有助于减少数据计算和建模中涉及的维数。有两种数据降维思想：一种是基于特征选择的降维，另一种是基于维度变换的降维。

3. 回归分析

在现实生活中，常常需要定量分析并确定两个或两个以上变量间的依存关系，这种分析方法称为回归分析(regression analysis)，该分析方法在统计学中被广泛应用。回归分析按照涉及的自变量的多少，可分为一元回归分析和多元回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。如果在回归分析中只包括一个自变量和一个因变量，且两者的关系可用一条直线近似表示，这种回归分析就称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。

在现实生活中，非线性关系是大量存在的，因此，非线性回归函数比线性回归函数能够更准确

地描述客观现象之间的回归关系。例如，在农作物产量与施肥量之间，随着施肥量的增加，粮食亩产量呈增加趋势，当施肥量达到一定的饱和点后，粮食亩产量不仅不会增加，反而会下降。又如，在商品价格保持不变的情况下，随着广告费支出的增加，商品销售量会呈线性增加趋势，但是当市场对该商品的需求趋于饱和时，再增加广告费支出，对商品销售量就不会产生显著影响，商品销售量会相对趋于稳定。因此，如果要分析施肥量对粮食亩产量的影响，或者分析广告支出费用对商品销售量的影响，就应考虑采用非线性回归模型。

非线性回归分析必须解决两个主要问题：一是如何确定非线性回归函数的具体形式；二是如何估计函数中的参数。非线性回归关系模型包括多项式模型、对数模型、幂函数模型和指数模型，对它们进行回归分析主要有两种方法：一是通过绘制散点图添加趋势线拟合出相应的回归方程；二是先将非线性关系线性化，然后利用回归分析工具进行线性回归分析。

在回归分析中，不仅需要确定变量间的相互依存关系，即确定回归函数，还需要检验估计的参数、评价方程拟合效果等。因此，回归分析是一个系统的分析过程。

4. 聚类

人们常说“物以类聚，人以群分”，这是聚类分析的基本思想。聚类分析法是大数据挖掘算法的基础，它是将很多具备“类似”特点的统计数据划分为一致类型的分析方式。大量数据集中必定有相似的数据点，基于这一假设可以区分数据，并且可以找到每个数据集(分类)的特征。

下面介绍几种常见的聚类算法。

(1) K-Means 均值聚类是非常知名的聚类算法。K-Means 算法的优势在于它的速度非常快，由于人们所要计算的是点和簇中心之间的距离，这已经是非常少的计算量了，因此它具有线性的复杂度 O(n)。

(2) Mean-Shift 是一种基于滑动窗口的聚类算法。也可以说，它是一种基于质心的算法，即它通过计算滑动窗口中的均值来更新中心点的候选窗口，以此达到找到每个簇中心点的目的。在接下来的处理阶段中，对这些候选窗口进行滤波，以消除近似或重复的窗口，找到最终的中心点及其对应的簇。

(3) 基于密度的噪声应用空间聚类(DBSCAN)。与其他聚类算法相比，DBSCAN 具有很多优点。首先，它根本不需要确定簇的数量。不同于 Mean-shift 算法，当数据点不同时，DBSCAN 会将它们单纯地引入簇中，将异常值识别为噪声。其次，它能够很好地找到任意大小和任意形状的簇。DBSCAN 的主要缺点是，当数据簇密度不均匀时，它的效果不如其他算法好。这是因为当密度变化时，用于识别邻近点的距离阈值 ε 和 minpoints 的设置将随着簇而变化。在处理高维数据时也会暴露这种缺点，这是因为难以估计距离阈值 ε。

(4) 使用高斯混合模型(GMM)的期望最大化(EM)聚类。相较于 K-means 算法，高斯混合模型能处理更多的情况。对于 GMM，我们假设数据点是高斯分布的，这是一个限制较少的假设，而不是用均值来表示它们是圆形的。这样，我们利用两个关键参数来描述簇的形状：均值和标准差。以二维空间为例，由于 GMM 在 x 和 y 方向上分别具有标准偏差，这意味着这些簇可以是任意方向的椭圆形。因此，每个高斯分布都对应一个簇的特定形状和位置。

5. 统计指数

统计指数是一种反映社会经济现象数量变动的相对数，体现在综合反映所研究的社会经济现象复杂总体数量的时间变动和空间对比状况。复杂总体是指不同度量单位或性质各异的若干事物组成的、数量不能直接加总或不可以直接加总的总体。因此，在实际应用中，统计指数不仅解决了复杂

总体数据太大的问题，更重要的是，对于不能直接加总或不能直接对比的复杂总体，编制统计指数能够反映和研究它们的变动方向和变动程度，以总结所有成员的综合变化。

指数的编制有特定的方法，因为指数要反映的是多个部分的变动问题，这里主要涉及横向如何加总和纵向如何对比的问题。在统计学中，按照指数的编制方式划分，指数主要分为“先综合、后对比”的综合统计指数和“先对比、后综合”的平均统计指数。

6. 抽样与参数估计

在实际工作中，常常需要对某一总体样本的特性进行分析，从成本和可行性的角度考虑，一般并不对总体的所有样本进行逐一检测，而是通过一定的方法抽取其中的一部分，通过抽出的部分来推断总体样本的特征。例如，要检验某种工业产品的质量，我们只需从中抽取一小部分产品进行检验，并用计算出来的合格率来估计全部产品的合格率，或者根据合格率的变化来判断生产线是否出现了异常。这便是统计工作中常用的抽样推断方法。

统计抽样推断是统计学研究的重要内容，是按照随机性原则，从研究对象中抽取一部分进行观察，并根据得到的观察数据，对研究对象的数量特征做出具有一定可靠程度的估计和推断，以达到认识总体的一种统计方法。它包括两大核心内容：参数估计和假设检验。两者都是根据样本资料，运用科学的统计理论和方法，对总体特征做出判断。其中，参数估计是对所要研究的总体参数进行合乎数理逻辑的推断，而假设检验是对先前提出的某个陈述进行检验，以判断真伪。

7. 方差分析

在统计学中，当需要对两个以上的总体均值进行检验时(即需要检验两个以上的总体是否具有相同的均值时)，需要进行方差分析。方差分析又称为变异数分析或 F 检验，是研究一个或多个可分组自变量与一个连续因变量之间的统计关系，并测定自变量对因变量的影响和作用的一种统计分析方法。简单而言，方差分析就是利用实验数据，分析各个因素对某事物、某指标的影响是否显著的一种统计分析方法。方差分析的目的是通过数据分析找出对该事物有显著影响的因素、各因素之间的交互作用，以及显著影响因素的最佳水平等。

在方差分析中通常要有以下两个假定：①各个观察值是独立的，即各组观察数据是从相互独立的总体中抽取的；②每个总体都应服从正态分布且方差相等，即各组观测数据是从具有相同方差的正态分布总体中抽取的简单随机样本。

按照总体均值仅受一个因素影响还是受两个因素影响来划分，方差分析可分为单因素方差分析和双因素方差分析。

8. 相关分析

自然界和人类社会中的许多事物或现象，彼此之间都是相互联系、相互依赖和相互制约的。例如，国内生产总值与财政收入之间、家庭消费支出与收入之间、人的身高与体重之间、农作物产量与施肥量之间、商品销售量与广告投入费用之间等，无不存在着一定的联系。现象之间的这种联系最终都要通过相互之间的数量对应关系反映出来，因此现象之间的联系必然表现为变量之间的依存关系。变量之间的依存关系有两种不同的类型：一种是函数关系，另一种是相关关系。前者是指变量之间存在的严格确定的依存关系，后者是指变量之间存在的不确定的依存关系。

具体来说，相关分析用于研究现象之间是否存在某种依存关系，并对具体有依存关系的现象进行相关方向和相关程度的探讨，是研究随机变量之间的相关关系的一种统计方法。从不同的角度来

划分，相关关系可以分为多种类型。根据研究变量的多少，可分为简单相关和多元相关，简单相关是指两个变量之间的相关关系，多元相关是指 3 个或 3 个以上变量之间的相关关系；根据变量关系的形态，可分为线性相关和非线性相关，可以通过散点图呈直线或曲线加以判断；根据变量值变动方向的趋势，可分为正相关和负相关，如果变量同增同减，则为正相关，反之则为负相关；根据相关程度的不同，又可分为完全相关、不完全相关和非相关。

9. 时间序列

客观事物永远处于不断的发展变化之中，我们要认识事物的本质及其变化的规律，展望其发展前景，不仅应该在客观事物的相互联系和相互制约中进行研究，还要在它们的发展变化中进行研究。这一任务正是通过编制时间序列进行时间序列分析来完成的。

时间序列分析是一种用于研究数据随时间变化规律的方法，常被应用于回归预测。时间序列分析的基本原则基于事物的连续性原理。连续性原理是指客观事物的发展过程遵循一种内在的、规律性的连续变化模式。在特定条件下，只要影响这些规律性的外部条件不发生根本性的变化，事物的基本发展趋势往往会持续至未来。因此，时间序列分析通过捕捉和利用数据中的时间依赖性，来预测未来的变化趋势。时间序列分析的形式有对比分析、移动平均分析、指数平滑分析、趋势外推分析和季节调整分析。

1.5 大数据的分析处理

大数据(big data)是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要使用新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》中指出，大数据处理指不使用随机分析法(抽样调查)这样的捷径，而是对所有数据进行分析处理。

1.5.1 大数据时代——你的一天

先来看一看未来大数据时代你的一天是如何度过的。

(1) 7:00，你被手机闹钟叫醒。昨晚你带着一款小型可穿戴设备睡觉，这个设备连接着你手机里的一款大数据 App，打开它就可以看到你昨晚睡觉时的翻身次数、心跳和血压状况。根据监测结果，它建议你今天出门之前多喝点橙汁等果汁饮品来补充维生素。

(2) 7:15，在你刷牙洗脸时，早餐机自动热好早餐。

(3) 7:30，在你吃早餐时，手机开始自动播报订阅的隔夜新闻，提醒日程安排。音响的屏幕显示当日的天气预报：有雨，不适合洗车，空气污染程度低，适合开窗透气。

(4) 7:55，先打开手机，控制车辆开启空调，调节好温度，然后设定目标路线，车内大数据系统会自动进行今天的交通预测，并根据大数据计算最佳的出行路线。

(5) 8:00，下楼，汽车已经根据指示自动驾驶，提前到达小区门口等待。出发上班，进入自动驾驶模式，车辆开始播放音乐，座椅自动躺平，开始简单的肩颈按摩。

(6) 8:25，到达公司。视频监视系统自动识别人物特征，车辆直接进入公司，下车后，车辆进入

自动泊车模式，自行到车库寻找车位。

(7) 8:30，大数据会将昨天遗留的工作内容和今天的工作安排发到你的手机。办公桌上已经没有计算机，直接使用手机将资料投影到办公桌前的一块玻璃上，并在投影中可以使用虚拟触摸操作，办公数据直接通过互联网存储到网络共享空间。

(8) 办公时，大数据系统自动发出信息，某商城打折，价格极其优惠，建议购买，于是手机下单，30 分钟后，无人机携带你购买的商品送达。

(9) 12:00，大数据会自动根据你之前的用餐记录，推荐你到一个餐馆用餐，并已经推荐好菜单。同时告诉你餐馆附近有多少车位，算出你可能会花费的拥堵时间、到了是否还有车位等可能性。

(10) 18:00，回到家，你的可穿戴设备告诉你，今天你在室内和室外的时间分别是多少，你一天内走了多少步，消耗了多少卡路里。

从 7 点被闹钟叫醒起床到 18 点下班回家，整整一天的活动都是在大数据的各种 App 的控制和记录中完成的。大数据时代方便了人们的生活，同时将人们的点点滴滴都记录下来，成为海量数据中的一部分。大数据离不开云处理，其为大数据提供了弹性可拓展的基础设备，是产生大数据的平台之一。自 2013 年开始，大数据技术已开始和云计算技术紧密结合，预计未来两者的关系将更为密切。

除此之外，物联网、移动互联网等新兴计算形态，也将一齐助力大数据革命，让大数据营销发挥出更大的影响力。随着大数据的快速发展，就像计算机和互联网一样，大数据很有可能是新一轮的技术革命。随之兴起的数据挖掘、机器学习和人工智能等相关技术，可能会改变数据世界中的很多算法和基础理论，实现科学技术上的突破。

1.5.2　大数据概述

大数据是一个体量特别大、数据类别特别多的数据集，无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理。

麦肯锡全球研究院给出的大数据的定义是，一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有数据规模大、数据流转速度快、数据类型多样和价值密度低四大特征。

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。

适用于大数据的技术，包括大规模并行处理(massively parallel processing，MPP)数据库、数据挖掘、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

1. 数据的存储设备

在计算机被发明之前，人类的数据大多以纸张和书籍的形式存储，全世界有很多的图书馆存储了海量的人类文明资料。随着电子设备的出现，人类的数据存储方式也发生了改变，从个人的硬盘到网络服务器，再到云，由纸张上的文字符号转变为以 0 和 1 表示的数字符号。这些改变，使得数据的意义发生了天翻地覆的变化，数据实时在线，24 小时可得，复制、传播、整合更加方便，保存和分享的成本越来越低。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台计算机进行处理，必须采用分布式架构(它的特色在于对海量数据进行分布式数据挖掘)，但它

必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术。

随着云时代的来临，大数据也得到了越来越多的关注。分析师团队认为，大数据通常用来形容一个公司创造的大量非结构化数据和半结构化数据，将这些数据下载到关系数据库用于分析时会花费过多时间和金钱。大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要借助类似于 MapReduce 的框架来向数十、数百甚至数千台计算机分配工作。

2. 数据的存储容量

存储容量是指存储器可以容纳的二进制信息量，用存储器中存储地址寄存器(MAR)的编址数与存储字位数的乘积表示。所有信息都是以“位”(bit)为单位传递的，一位就代表一个 0 或 1。每 8 位组成一字节(Byte)。一字节是什么概念呢？一个英文字母就占用一字节，也就是二进制的 8 位，一个汉字占用两字节。一般，位简写为小写字母 b，字节简写为大写字母 B。

存储容量的计算如下。

1Byte(字节)=8bit(位)

1KB(千字节)=1024B=2^{10}B

1MB(兆字节)=1024KB=2^{20}B

随着信息量的增大，有更大的单位表示存储容量，比兆字节更大的还有：吉字节(gigabyte，GB)、太字节(terabyte，TB)、拍字节(petabyte，PB)、艾字节(exabyte，EB)、泽字节(zettabyte，ZB)和尧字节(yottabyte，YB)、千亿亿亿字节(brontobyte，BB)等。

1GB(吉字节)=1024MB=2^{30}B

1TB(太字节)=1024GB=2^{40}B

1PB(拍字节)=1024TB=2^{50}B

1EB(艾字节)=1024PB=2^{60}B

1ZB(泽字节)=1024EB=2^{70}B

1YB(尧字节)=1024ZB=2^{80}B

1BB(千亿亿亿字节)=1024YB=2^{90}B

3. 大数据有多大

据统计，2019 年移动互联网接入流量 12 200 000 万 GB，互联网宽带接入用户 44 928 万户。阿里云曾帮助用户抵御全球互联网史上最大的 DDoS 攻击，峰值流量达到每秒 453.8Gb/s。阿里巴巴、百度、腾讯这样的互联网巨头，数据量已经接近 EB 级。一个 8Mb/s 的摄像头，一小时能产生 3.6GB 的数据，一个城市每月产生这样的数据达上亿 GB。在医院，一个病人的 CT 影像数据量达几十 GB，全国每年需保存这样的数据达上百亿 GB。

国际数据公司(IDC)的研究结果表明，2008 年全球产生的数据量为 0.49ZB，2009 年的数据量为 0.8ZB，2010 年增长为 1.2ZB，2011 年的数据量更是高达 1.8ZB，相当于全球每人产生 200GB 以上的数据，2012 年，数据量已经从 TB(1024GB=1TB)级别跃升到 PB(1024TB=1PB)、EB(1024PB=1EB)，甚至 ZB(1024EB=1ZB)级别，2015 年的数据达到 7.9ZB，2020 年数据量已达到 35ZB，如图 1.26 所示。

根据 IDC 监测，人类产生的数据量(数据圈)正在呈指数级增长，大约每两年翻一番，全球每年产生的数据将从 2020 年的 35ZB 增长到 2025 年的 175ZB，相当于每天产生 491EB 的数据。预计在 2025 年中国数据圈增至 48.6ZB，占全球 27.8%，成为最大数据圈。如果把 175ZB 全部存在 DVD 光盘中，那么 DVD 光盘叠加起来的高度将是地球和月球距离的 23 倍(月地距离约 38.4 万千米)，或者

绕地球 222 圈(一圈约为 4 万千米)。假设网速为 25Mb/s，一个人要下载完 175ZB 的数据，需要 18 亿年。

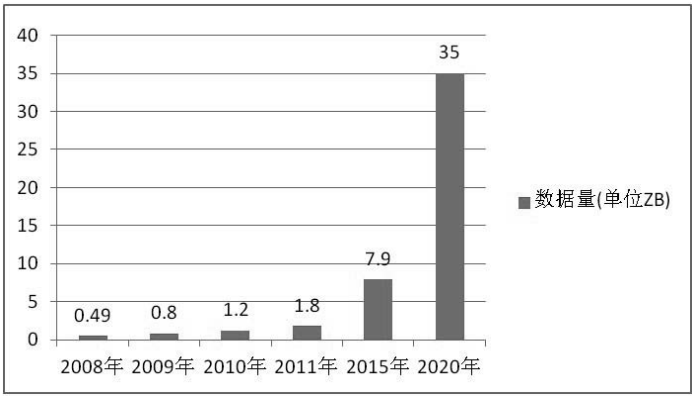


图 1.26　全球数据量统计图

4. 大数据的特征

大数据其实就是海量资料，这些海量资料来源于世界各地随时产生的数据。在大数据时代，任何微小的数据都可能产生不可思议的价值。大数据的 4 个特征分别为 volume(规模性)、variety(多样性)、velocity(高速性)、value(价值性)，一般称为 4V。

(1) volume(规模性)。大数据的特征首先就体现为“大”，在早前的 Map 3 时代，一个小小的 MB 级别的 Map 3 就可以满足人们的需求，然而随着时间的推移，存储单位从过去的 GB 扩展到了 TB，乃至现在的 PB、EB 级别。随着信息技术的高速发展，数据开始爆发性增长。社交网络(微博、推特、脸书)、移动网络、各种智能工具、服务工具等，都成为数据的来源。2019 年，脸书日活跃用户数为 15.9 亿，每天产生的日志数据以 PB 计。在当今时代，迫切需要智能的算法、强大的数据处理平台和新的数据处理技术来统计、分析、预测和实时处理如此大规模的数据。

(2) variety(多样性)。广泛的数据来源决定了大数据形式的多样性。任何形式的数据都可以产生作用，目前应用最广泛的就是推荐系统，如淘宝、网易云音乐、今日头条等，这些平台都会通过对用户的日志数据进行分析，从而进一步推荐用户喜欢的东西。日志数据是结构化明显的数据，还有一些数据结构化不明显，如图片、音频、视频等，这些数据因果关系弱，需要人工对其进行标注。

(3) velocity(高速性)。大数据的产生非常迅速，主要通过互联网传输。生活中每个人都离不开互联网，也就是说每个人每天都在向大数据提供大量的资料。这些数据是需要及时处理的，因为花费大量资本去存储作用较小的历史数据是非常不划算的，对于一个平台而言，保存的数据一般为过去几天或一个月之内产生的，更早的数据就要及时清理，不然存储代价太大。基于这种情况，大数据对处理速度有非常严格的要求，服务器中大量的资源都用于处理和计算数据，很多平台都需要做到实时分析。数据无时无刻不在产生，谁的速度更快，谁就有优势。

(4) value(价值性)。这也是大数据的核心特征。现实世界所产生的数据中，有价值的数据所占比例很小。相比于传统的小数据，大数据最大的价值在于通过从大量不相关的各种类型的数据中挖掘出对未来趋势与模式预测分析有价值的数据，通过机器学习方法、人工智能方法或数据挖掘方法深度分析，发现新规律和新知识，并运用于农业、金融、医疗等各个领域，从而最终达到改善社会治理、提高生产效率、推进科学研究的目的。

5. 大数据的应用

现代社会是一个高速发展的社会，科技发达，信息流通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个高科技时代的产物。

有人把数据比喻为蕴藏能量的煤矿，而露天煤矿与深山煤矿的挖掘成本肯定不一样。与此类似，大数据并不在于“大”，而在于“有用”。价值含量、挖掘成本比数量更为重要。对于很多行业而言，如何利用这些大规模数据是赢得竞争的关键。

提到大数据，也许大部分人会联想到庞大的服务器集群，或者联想到销售商提供的一些个性化的推荐和建议。如今大数据应用的深度和广度远不止这些，大数据已经在人类社会实践中发挥着巨大的优势，其利用价值也超出我们的想象。下面就来介绍大数据的五大应用领域。

1) 政治领域

政治网络营销对于政治选举以及其他政治活动具有重要作用。在网络技术不断发展、信息应用不断创新的过程中，政治网络营销以其双向、交互、共享、快速、广泛、经济、便捷等特点，成为政党和政治团队之间竞争的重要手段。

一个经典的案例就是大数据帮助美国前总统奥巴马成功实现连任。奥巴马的数据团队对数以千万计的选民邮件进行了大数据挖掘，精确预测出了更可能拥护奥巴马的选民类型，并进行了有针对性的宣传，从而帮助奥巴马成为美国历史上在竞选经费处于劣势下实现连任的总统。

2) 金融领域

在大数据发展如火如荼的迅猛盛况下，互联网金融应运而生且茁壮成长。大数据分析与应用在金融领域的应用与发展，给越来越多的公司带来了更多的收益和对未来规划越来越可靠的数据支撑。支付宝、京东金融等都在依托大数据分析与应用推出越来越符合大众需求的金融产品。

金融大数据应用已经成为行业热点趋势，在交易欺诈识别、精准营销、消费信贷、信贷风险评估、供应链金融、股市行情预测、股份预测、骗保识别、风险定价等涉及银行、证券、保险、支付清算和互联网金融等多领域的具体业务中，金融大数据得到了广泛应用。

例如，大数据在证券交易方面的应用。证券交易实时性要求高、数据规模大，目前沪深两市每天 4 个小时的交易时间会产生 3 亿条以上逐笔成交数据。通过对历史和实时数据的挖掘创新，可以创造和改进量化交易模型，并将之应用于基于计算机模型的实时证券交易过程中。

3) 电子商务领域

通过大数据进行市场营销不仅能够有效节约企业或电子商务平台的营销成本，还能够通过大数据实现营销的精准化，达成精准营销。通过大数据对消费者的消费偏好进行分析，可在消费者输入关键词之后，提供与消费者消费偏好匹配程度较高的产品，节约了消费者寻找商品的时间成本，使交易双方实现快速的对接，实现电子商务平台或企业营销的高效化。在数据化时代，针对消费者进行针对性的营销能够实现精准营销，提升产品的下单率及电子商务的营销效率。

对于电子商务平台来讲，往往都会针对用户提供一些推荐和导购服务。通过大数据的分析和挖掘，能够实现导购服务的个性化。针对消费者的年龄、性别、职业、购买历史、购买商品种类、查询历史等信息，对消费者的消费意向、消费习惯、消费特点进行系统性的分析，从而针对消费者个人制定个性化的推荐和导购服务。

大数据的运用能够抵消电子商务虚拟性带来的影响，提升竞争力，挖掘更多的潜在消费者。针对消费者的消费偏好进行适宜的广告推广，可以提升产品的广告转化率，便于提供个性化的导购服务。

对于一些大型的电子商务平台来讲，产品种类繁多，想要提升消费者的下单率就要分析消费

者的消费偏好，主动进行商品推送。这种通过大数据进行分析的方式不仅能提升消费者对产品的浏览量，还能针对消费者的消费需求提供商品的推送，提升消费者的用户体验，进而提升消费者的忠诚度。

大数据的分析不仅能够帮助电子商务平台提升下单率和销售额，还能将大数据的分析作为产品和服务向中小型的电子商务商家进行销售。这样不仅能够提升平台的收益，还能帮助商家了解消费者的消费偏好、消费者对于该类产品的喜好等信息，帮助商家及时针对大部分消费者的消费偏好及市场的动态，针对产品的性能等进行研发和调整。

经典案例之“聚划算”商品预测系统。“聚划算”中筛选商品是关键的一环，之前这个项目的运营人员平均每人每天要审核 200 个商品。为提升选品效率，增加爆款量，该系统根据“聚划算”和淘宝的属性、卖家的属性、品牌属性，自动筛选高销量的商品，从海量的卖家中寻找合适的卖家，并进行合理的定价和库存管理，这使得平均销量提高了 64%。

4) 教育领域

美国独立研究机构布鲁金斯学会发布的报告中指出，大数据使得查探关于学生表现和学习途径的信息成为可能，不用依赖阶段测验表现，导师就可以分析学生懂什么及每个学生最有效的技术是什么，通过聚焦大数据的分析，教师可以用更微妙的方式研究学生的学习状况。

可以说，大数据贯穿基础教育、高等教育，甚至终生教育。教育大数据分布在包括教育管理、教学资源、教学行为、教学评估等在内的综合教育系统的始末。大数据的思维和理念可以为优化教育政策、创新教育教学模式、变革教育测量与评价方法等理论研究提供客观依据和新的研究视角，能够更好地推动教育领域的变革。

经典案例之改善教育。在美国，高中生退学率高达 30%，33%的大学生需要重修，46%的大学生无法正常毕业。美国联邦政府教育部2012年参与了一项耗资2亿美元的公共教育中的大数据计划，旨在通过大数据分析来改善教育。大数据分析已经被应用到美国的公共教育中，成为教学改革的重要力量。比尔·盖茨曾说过，利用数据分析教育大数据能够提高学生的学习成绩，拯救美国的公立学校系统。教育技术未来发展的关键在于数据。

5) 生活娱乐

在物联网时代，大数据将成为文化娱乐产业的核心资产。在大数据的具体应用上，娱乐产业的生态全球化商业模式至关重要。全球化首先是内容的全球化，电影公司应当充分利用生态圈和互联网大数据红利。任何一个国家的公司可以在全世界寻求合作伙伴，将更多元化的平台、更蓬勃的生态、更优质的内容带到全世界。对于文化娱乐产业来说，内容是核心，只要把握好全球用户的需求，生产出符合用户需求的内容，就能吸引全球的用户，市场巨大。2013 年的《纸牌屋》使 big data 这个时髦的概念被影视产业所熟知，如果说刚开始还只是一个噱头大于实质的探索阶段，那么如今的大数据在影视圈可以说大有作为，它可以为娱乐项目的投资决策、演员组合、剧本修改、营销策略提供实际帮助。

经典案例之《红海行动》。经统计，2018 年中国大陆共有 334 部电影登陆院线，其中进口影片为 94 部，国产影片为 240 部。国产片占比为 72%。电影频道融媒体中心公开发布的《2018 中国电影年度调查报告》显示，2018 年国产电影市场份额较 2017 年整体增长 9%左右。国产电影市场占有率的增长，代表了观众对国产电影信任度的上升，说明观众越来越愿意走进影院观看国产电影。全年电影票房前十名中，前四名均为国产影片。截至 2018 年 12 月 31 日，在票房排名前十的影片中，影片《红海行动》以 36.51 亿元领跑票房排行榜，其次为《唐人街探案 2》33.98 亿元、《我不是药神》

31 亿元、《西虹市首富》25.4 亿元。观众的消费行为越来越趋于成熟和理性，因此一部电影想要获得较好的经济效益，不仅要选符合市场口味的题材，还要坚持质量为上，并在宣传上下功夫。《红海行动》之所以成为 2018 年度最具代表性的票房、口碑、热度综合实力较高的电影，是因为该电影彰显了中国军人刚毅果敢的英勇形象，观众看得热血沸腾，电影口碑一路飙升；战斗场景还原、角色设置、演员选取和气氛渲染等多方面的制作使得该影片在质量上达到了观众对于电影的高标准要求；上映期间借助微博、路演、品牌合作等多方面进行营销也为其票房做出了一定的贡献。

6. 大数据的安全隐忧

美国的情报部门通过一个代号为“棱镜”的项目，从多家知名互联网公司获取电子邮件、在线聊天内容、照片、文档、视频等网络私人数据，跟踪用户的一举一动。

2013 年，美国中央情报局前职员斯诺登向媒体的爆料，引起一片哗然。根据他提供的资料，被卷入“棱镜门”事件的公司包括多家 IT 行业巨头。在“棱镜门”事件开始发酵之后，这些公司先是赶紧出面否认与美国政府的监视项目进行过合作，并相继发表声明，呼吁政府采取更透明的态度，以证明他们的“清白”。

大数据为人们带来便捷的同时不可避免地对人们的隐私构成了威胁。

人们所有的网络行为对于服务提供商来说都是透明的。人们既想借助互联网平台与别人交流，又不想被窥探，这是完全不可能的。网络隐私安全未来将是一个巨大的问题。

在数据的应用方面，相关法律法规的制定变得越来越重要。作为用户，需要明确界定自己在数据的使用方面具有哪些权利和义务；作为企业和政府，需要明确可以在多大程度上或以怎样的方式使用用户的数据。

1.5.3 数据挖掘

1. 数据挖掘技术的产生

20 世纪 90 年代，随着数据库系统的广泛应用和网络技术的高速发展，数据库技术也进入了一个全新的阶段，即从过去仅管理一些简单数据发展到管理由计算机产生的图形、图像、音频、视频、电子档案、Web 页面等多种类型的复杂数据，并且数据量越来越大。数据库在给人们提供丰富信息的同时，也体现出明显的海量信息特征。信息爆炸时代，海量信息给人们带来了许多负面影响，比较明显的就是难以提炼有效信息，过多无用的信息必然会产生信息距离(信息状态转移距离，是对一个事物信息状态转移所遇到障碍的测度，简称 DIST 或 DIT)并导致有用知识的丢失。这也就是约翰·内斯伯特所说的“信息丰富而知识贫乏”窘境。因此，人们迫切希望能对海量数据进行深入分析，发现并提取隐藏在其中的信息，以更好地利用这些数据。但仅以数据库系统的录入、查询、统计等功能，无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势，更缺乏挖掘数据背后隐藏知识的手段。正是在这样的条件下，数据挖掘技术应运而生。

2. 数据挖掘的对象

数据挖掘是指从大量的数据中通过算法搜索隐藏于其中的信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统(依靠过去的经验法则)和模式识别等诸多方法来实现上述目标。

数据的类型可以是结构化的、半结构化的，甚至是非结构化的。发现知识的方法可以是数学的、

非数学的，也可以是归纳的。最终被发现的知识可以用于信息管理、查询优化、决策支持及数据自身的维护等。

数据挖掘的对象可以是任何类型的数据源，可以是关系数据库，此类型包含结构化数据；也可以是数据仓库、文本、多媒体数据、空间数据、时序数据、Web 数据，此类型包含半结构化数据甚至非结构化数据。

3. 数据挖掘的步骤

在实施数据挖掘之前，明确采取什么样的步骤、每一步都做什么、达到什么样的目标都是很有必要的。有了好的计划才能保证数据挖掘有条不紊地实施并取得成功。

数据挖掘的步骤主要包括定义问题、建立数据挖掘库、分析数据、准备数据、建立模型、评价模型和实施。

(1) 定义问题。在进行数据挖掘之前，首要任务就是了解数据和业务问题。必须对目标有一个清晰明确的定义，即决定到底想干什么。例如，为了提高电子信箱的利用率，想做的可能是“提高用户使用率”，也可能是“提高一次用户使用的价值”。为解决这两个问题而建立的模型几乎是完全不同的，必须做出决定。

(2) 建立数据挖掘库。建立数据挖掘库包括以下几个步骤：数据收集、数据描述、选择、数据质量评估和数据清理、合并与整合、构建元数据、加载数据挖掘库、维护数据挖掘库。

(3) 分析数据。分析的目的是找到对预测输出影响最大的数据字段，决定是否需要定义导出字段。如果数据集包含成百上千的字段，那么浏览分析这些数据将是一件非常耗时和累人的事情，这时，用户需要选择一个具有友好界面且功能强大的工具软件来完成这件事。

(4) 准备数据。这是建立模型之前的最后一步，可以把此步骤分为四个部分：选择变量、选择记录、创建新变量、转换变量。

(5) 建立模型。建立模型是一个反复的过程。需要仔细考察不同的模型以判断哪个模型对面对的问题最有用。先用一部分数据建立模型，然后再用剩下的数据测试和验证这个模型。有时还有第三个数据集，称为验证集，因为测试集可能受模型的特性影响，这时需要一个独立的数据集来验证模型的准确性。训练和测试数据挖掘模型需要将数据至少分成两个部分，一个用于模型训练，另一个用于模型测试。

(6) 评价模型。模型建立好之后，必须评价得到的结果、解释模型的价值。从测试集中得到的准确率只对用于建立模型的数据有意义。在实际应用中，需要进一步了解错误的类型和由此带来的相关费用的多少。经验证明，有效的模型并不一定是正确的模型。造成这一点的直接原因就是模型建立中隐含的各种假定，因此，直接在现实世界中测试模型很重要。我们可先在小范围内应用，取得测试数据，觉得满意之后再大范围推广。

(7) 实施。模型建立并经验证之后，有两种主要的使用方法：第一种是提供给分析人员做参考；另一种是把此模型应用到不同的数据集上。

4. 数据挖掘分析方法

数据挖掘分为有指导的数据挖掘和无指导的数据挖掘。有指导的数据挖掘是利用可用的数据建立一个数据模型，这个模型是对一个特定属性的描述。无指导的数据挖掘是在所有的属性中寻找某种关系。具体而言，分类、估值和预测属于有指导的数据挖掘；关联规则和聚类属于无指导的数据挖掘。

(1) 分类。先从数据中选出已经分类的训练集，在该训练集上运用数据挖掘技术建立一个分类模型，再用该模型对没有分类的数据进行分类。

(2) 估值。与分类类似，但估值最终的输出结果是连续型数值，估值的量并非预先确定。估值可以作为分类的准备工作。

(3) 预测。它是通过分类或估值来进行的，通过分类或估值的训练得出一个模型，如果对于检验样本组而言该模型具有较高的准确率，则可将该模型用于对新样本的未知变量进行预测。

(4) 相关性分组或关联规则。其目的是发现哪些事情总是一起发生。

(5) 聚类。它是自动寻找并建立分组规则的方法，它通过判断样本之间的相似性，把相似样本划分在一个簇中。

5. 数据挖掘的经典算法

目前，数据挖掘的算法主要包括遗传算法、决策树法、神经网络法、粗糙集法、模糊集法、关联规则法等。

(1) 遗传算法。遗传算法模拟了自然界中生物的进化过程，包括繁殖、交配和基因突变等现象，是一种基于进化理论的机器学习方法，它通过遗传结合、遗传交叉变异及自然选择等操作来生成和优化解决问题的规则。它的基本观点是“适者生存”，具有隐含并行性、易于和其他模型结合等性质。遗传算法的主要优点是可以处理许多数据类型，同时可以并行处理各种数据；缺点是需要的参数太多，编码困难，一般计算量比较大。遗传算法常用于优化神经元网络，能够解决其他技术难以解决的问题。

(2) 决策树法。此算法是根据对目标变量产生效用的不同而建构分类的规则，通过一系列的规则对数据进行分类的过程，其表现形式类似于树形结构的流程图。典型的算法是 John Ross Quinlan 于 1986 年提出的 ID3 算法，之后在 ID3 算法的基础上又提出了极其流行的 C4.5 算法。决策树法的优点是决策制定的过程是可见的，不需要长时间构造过程，描述简单，易于理解，分类速度快；缺点是很难基于多个变量组合发现规则。决策树法擅长处理非数值型数据，而且特别适用于进行大规模的数据处理。决策树提供了一种展示类似在什么条件下会得到什么值这类规则的方法。例如，在贷款申请中，要对申请的风险大小做出判断。

(3) 神经网络法。此算法模拟生物神经系统的结构和功能，是一种通过训练来学习的非线性预测模型，它将每一个连接看作一个处理单元，试图模拟人脑神经元的功能，来完成分类、聚类、特征挖掘等多种数据挖掘任务。神经网络的学习方法主要表现在权值的修改上，其优点是具有抗干扰、非线性学习、联想记忆功能，对于复杂情况能得到精确的预测结果；缺点首先是不适合处理高维变量，不能观察中间的学习过程，具有“黑箱”性，输出结果也难以解释，其次是需较长的学习时间。神经网络法主要应用于数据挖掘的聚类技术中。

(4) 粗糙集法。粗糙集法也称为粗糙集理论，是由波兰数学家 Z Pawlak 在 20 世纪 80 年代初提出的，是一种新的处理含糊、不精确、不完备问题的数学工具，可以处理数据约简、数据相关性发现、数据意义的评估等问题。粗糙集法的优点是算法简单，在处理过程中不需要任何关于数据的预备知识，可以自动找出问题的内在规律；缺点是难以直接处理连续的属性，需先进行属性的离散化。因此，连续属性的离散化问题是制约粗糙集理论实用化的难点。粗糙集理论主要应用于近似推理、数字逻辑分析和化简、建立预测模型等问题。

(5) 模糊集法。模糊集法是指利用模糊集合理论对问题进行模糊评判、模糊决策、模糊模式识别和模糊聚类分析。模糊集合理论用隶属度来描述模糊事物的属性。系统的复杂性越高，模糊性就

越强。

(6) 关联规则法。关联规则反映了事物之间的相互依赖性或关联性。著名的算法是 Rakesh Agrawal 等人提出的 Apriori 算法，该算法的思想是，首先找出频繁性至少和预定意义的最小支持度一样的所有频集，然后由频集产生强关联规则。最小支持度和最小可信度是为了发现有意义的关联规则而给定的两个阈值。在这个意义上，数据挖掘的目的就是从源数据库中挖掘出满足最小支持度和最小可信度的关联规则。

本章小结

本章对数据分析与处理的相关概念进行了阐述，使读者对数据分析与处理有一个基本的认知，理解数据分析模型的有关内容，了解大数据相关的前沿技术及应用。本章主要内容包括数据分析与处理的概念、数据分析与处理的过程、数据分析模型，以及数据分析在大数据处理中的应用。本章的重点是数据分析与处理的过程、数据分析模型、大数据特点。本章的难点是数据分析与处理的过程。

习题 1

一、填空题

1. 大数据的(　　)特性，是指大数据中蕴含着巨大的价值，但是价值密度较低，呈现碎片化、离散化。

2. 根据数据的结构，可以将数据分为(　　)、半结构化数据和非结构化数据。

3. (　　)是为了提取有用信息和形成结论而对数据加以详细研究和总结的过程。

4. 数据按表现形式可分为：数字数据和(　　)数据。

5. 数据挖掘的对象可以是(　　)类型的数据源。

6. 数据可视化是关于数据(　　)表现形式的科学技术研究。

7. (　　)是指用适当的统计方法对各种数据资料进行全面分析，以求最大化地开发数据资料的功能，发挥数据的作用。

二、选择题

1. 预处理的方法不包括(　　)。

A. 数据清理　　B. 数据集成　　C. 数据转换　　D. 数据分析

2. (　　)是对数据进行采集、存储、检索、加工、变换和传输。

A. 数据分析　　B. 数据处理　　C. 数据压缩　　D. 数据应用

3. 一个企业数据分析师应具备的技能是(　　)。

A. 具有很好的学习能力

B. 能够理解并转化同事们的问题和解决办法

C. 对质量标准的高度要求和对细节的关注

D. 以上都是

4. 下列选项中，属于大数据应用领域的是(　　)。

A. 政治领域　　B. 金融领域　　C. 电子商务领域　　D. 教育领域

5. 当前世界上的数据量属于(　　)数量级。

A. 泽字节　　B. 艾字节　　C. 太字节　　D. 拍字节

6. 下列选项中，不属于大数据特点的是(　　)。

A. 规模性　　B. 多样性　　C. 灵活性　　D. 高速性

7. 数据可视化方法不包括(　　)。

A. 面积和尺寸可视化　　B. 颜色可视化

C. 数据压缩　　D. 地域空间可视化

8. 下列选项中，不属于非结构化数据的是(　　)。

A. 视频　　B. 音频　　C. 图像　　D. Excel 数据表

三、综合题

1. 如何在分列后的城市名称后面加上“市”字？

2. 进行数据分析的目的是什么？

3. 数据挖掘的步骤是什么？

4. 大数据的主要特征是什么？

5. 合并操作：在如图 1.27 所示的对应的单元格 C2、C3、H2、H3 中显示文本合并结果和日期合并结果。

图 1.27　合并操作表