第 1 章
数据可视化简介

本章学习目标：

- 掌握大数据的概念。
- 了解大数据的发展现状和趋势。
- 了解数据可视化的基本概念。
- 掌握数据可视化技术的基础思想。

本章首先介绍什么是大数据及大数据的基本特征；然后根据大数据产生的缘由对它进行分析；接着介绍大数据的发展现状和趋势；最后详细介绍了数据可视化的基本概念及可视化技术的应用。

1.1 大数据概述

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。

在维克托·迈尔·舍恩伯格及肯尼斯·库克耶编写的《大数据时代》一书中，对于大数据，并不是使用随机分析法（抽样调查）这种捷径对抽样数据进行分析处理，而是对所有数据进行分析处理。

大数据的 5V 特点（IBM 提出）如下：Volume（大量）、Velocity（高速）、Variety（多样）、Value（低价值密度）、Veracity（真实性）。

对于大数据，研究机构 Gartner 给出了如下定义：“大数据”是需要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。麦肯锡全球研究所给出的定义如下：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，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

1. 大数据产生的原因

大多数的技术突破来源于实际的产品需要。大数据最初诞生于谷歌的搜索引擎中，随着 Web 2.0 时代的发展，互联网上的数据量呈现爆炸式的增长，为了满足信息搜索的需要，对大规模数据的存储提出了非常高的要求。

当数据量、数据的复杂程度、数据处理的任务要求等超出了传统数据存储与计算能力时，称之为“大数据（现象）”。可见，计算机科学与技术是从存储和计算能力视角来理解大数据

的——大数据不仅仅是数据存量的问题，还与数据增量、复杂度和处理要求（如实时分析）有关。

大量信息带来的问题如下：

- 信息过量，难以消化。
- 信息真假难以辨识。
- 信息安全难以保证。
- 信息形式不一致，难以统一处理。
- 缺乏挖掘数据背后隐藏的知识的手段，导致“数据爆炸但知识贫乏”现象。

2. 基本单位

在计算机存储中，最小的基本单位是 bit，最大的是单位是 DB；按从小到大的顺序给出所有单位如下：bit、B、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB。它们按照进率 1024（2 的 10 次方）来计算，如图 1-1 所示。

1 B=8 bit
1 KB = 1,024 B=8192 bit
1 MB = 1,024 KB = 1,048,576 B
1 GB = 1,024 MB = 1,048,576 KB
1 TB = 1,024 GB = 1,048,576 MB
1 PB = 1,024 TB = 1,048,576 GB
1 EB = 1,024 PB = 1,048,576 TB
1 ZB = 1,024 EB = 1,048,576 PB
1 YB = 1,024 ZB = 1,048,576 EB
1 BB = 1,024 YB = 1,048,546 ZB
1 NB = 1,024 BB = 1,048,546 YB
1 DB = 1,024 NB = 1,048,546 BB

图 1-1　计算机存储单位

3. 大数据技术

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的计算机进行处理，必须采用分布式架构。分布式架构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。大数据需要特殊的技术，以有效地处理大量的容忍经过时间内的数据。适用于大数据的技术，包括大规模并行处理（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。

1.2　大数据与大数据技术发展历程

1. 大数据发展历程

在全球范围内，以电子方式存储的数据（简称为电子数据）总量空前巨大。2011 年，电子数据总量达到 1.8ZB（ZettaByte，泽字节，代表的是十万亿亿字节），相比 2010 年同期增加了 1ZB，统计结果表明，每经过两年就可以增加一倍。

回顾大数据的发展历程，大数据总体上可以划分为以下 4 个阶段：萌芽期、成长期、爆发期和稳步发展期。

（1）萌芽期（1980—2008 年）：大数据术语被提出，相关技术概念得到一定程度的传播，但没有得到实质性发展。同一时期，随着数据挖掘理论和数据库技术的逐步成熟，一批商业

智能工具和知识管理技术开始被应用。1980 年，未来学家托夫勒在其所著的《第三次浪潮》一书中首次提出“大数据”一词，将大数据称赞为“第三次浪潮的华彩乐章”。2008 年 9 月，《自然》杂志推出了“大数据”封面专栏。

（2）成长期（2009—2012 年）：大数据市场迅速成长，互联网数据呈爆发式增长，大数据技术逐渐被大众熟悉和使用。2010 年 2 月，肯尼斯·库克尔在《经济学人》上发表了长达 14 页的大数据专题报告《数据，无所不在的数据》。2012 年，牛津大学教授维克托·迈尔·舍恩伯格的著作《大数据时代》开始在我国风靡，推动了大数据在我国的发展。

（3）爆发期（2013—2015 年）：大数据迎来了发展的高潮，世界各个国家纷纷布局大数据战略。2013 年，以百度、阿里、腾讯为代表的国内互联网公司各显身手，纷纷推出创新性的大数据应用。2015 年 9 月，国务院发布《促进大数据发展行动纲要》，全面推进国大数据发展和应用，进一步提升创业创新活力和社会治理水平。

（4）稳步发展期（2016 年至今）：大数据应用渗透到各行各业，大数据价值不断凸显，数据驱动决策和社会智能化程度大幅提高，大数据产业迎来快速发展和大规模应用实施。2019 年 5 月，《2018 年全球大数据发展分析报告》显示，中国大数据产业发展和技术创新能力有了显著提升。这一时期学术界在大数据技术与应用方面的研究创新也不断取得突破，截至 2020 年，全球以“big data”为关键词的论文发表量达到 64,739 篇，全球共申请大数据领域的相关专利 136,694 项。

随着我国大数据战略谋篇布局的不断展开，国家高度重视并不断完善大数据政策支撑，大数据产业迅速发展，大致经历了 4 个阶段，如图 1-2 所示，正逐步从数据大国向数据强国迈进。

图 1-2　大数据发展历程

2. 大数据技术发展历程

大数据技术是指从数据采集、清洗、集成、存储、展示到分析，进而从各种各样的巨量数据中快速获得有价值信息的全部技术。目前所说的大数据有双重含义，它不仅指数据本身的特点，也包括采集数据的工具、平台和数据分析系统等技术。

在大数据时代，传统的软件已经无法处理和挖掘大量数据中的信息。谷歌在 2004 年前后相继发布了分布式文件系统（GFS）、大数据分布式计算框架——MapReduce、大数据 NoSQL 数据库——Big Table。受到谷歌的启发，2004 年 7 月，Doug Cutting 和 Mike Cafarella 在 Nutch 中实现了类似 GFS 的功能，也就是 HDFS 的前身。2005 年 2 月，Mike Cafarella 在 Nutch 中实现了 MapReduce 的最初版本。图 1-3 所示为大数据技术发展的全过程。

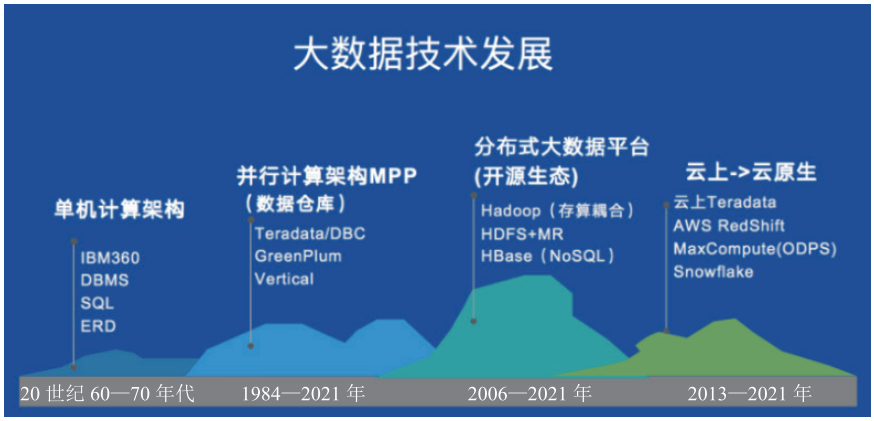


图 1-3　大数据技术发展

在大数据的生命周期中，数据采集处于第一个环节。根据 MapReduce 产生数据的应用系统分类，大数据的采集主要有 4 种来源：管理信息系统、Web 信息系统、物理信息系统、科学实验系统。对于不同的数据集，可能存在不同的结构和模式，如文件、XML 树、关系表等，表现为数据的异构性。对多个异构的数据集，需要做进一步集成处理或整合处理，将来自不同数据集的数据收集、整理、清洗、转换后，生成一个新的数据集，为后续查询和分析处理提供统一的数据视图。针对管理信息系统中异构数据库集成技术、Web 信息系统中的实体识别技术和 Deep Web（又称不可见网、隐藏网，是指互联网上那些不能被标准搜索引擎索引的非表面网络内容）集成技术、传感器网络数据融合技术等，业界人员已经做了很多研发工作，并取得了较大的进展，也推出了多种数据清洗和质量控制工具。这些工具包括美国 SAS 公司的 Data Flux、美国 IBM 公司的 Data Stage、美国 Informatica 公司的 Informatica Power Center 等。

3. 大数据处理流程

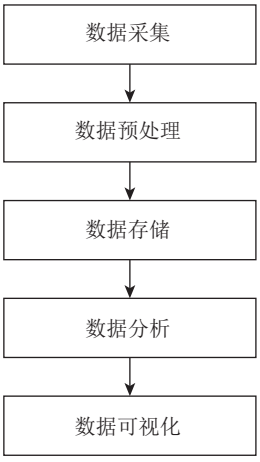


图 1-4　大数据处理流程

一般来说，大数据处理流程包括数据采集、数据预处理、数据存储、数据分析、数据可视化，如图 1-4 所示。

（1）数据采集。数据采集又称数据获取，通过 RFID 射频数据、传感器数据、社交网络数据、移动互联网数据等方式获得各种类型的结构化、半结构化及非结构化的海量数据。

（2）数据预处理。要分辨清楚哪些数据采用批处理就可以了、哪些数据是有实时处理价值的。实时处理对技术要求高，毕竟集群资源是有限的，需要合理利用计算资源。

（3）数据存储。数据存储是一个使用存储库持久地存储和管理数据的集合，其中不仅包括数据仓库，还包括简单的存储类型，如简单的文件、电子邮件等。

（4）数据分析。将多份数据查询出来，互相关联合并，生成一张新的表单，然后可以在新表单的基础上进行查询或者再与其他数据关联合并。

（5）数据可视化。数据可视化即数据的图形表示，旨在以更易于掌握和理解的有效方式传达大量海量数据。从某种意义上说，数据可视化是原始数据和图形元素之间的映射，它决定了这些元素的属性如何变化。可视化通常是通过使用图表、折线、点、条形图和地图来进行的。

1.3　数据可视化简介

本节主要对数据可视化进行介绍。

1.3.1　数据可视化的概念与分类

人类的创造性不仅取决于人类的逻辑思维，还取决于人类的形象思维。将数据映射为视觉符号，充分利用人们的杰出视觉来获取大数据中蕴含的信息。只有将大数据变成形象可视化之后才能激发人的形象思维与想象力。由于数据可视化的范围不断扩大，导致数据可视化成为一个不断发展与动态变化的概念。

数据可视化包括 3 种类型：科学可视化、信息可视化和数据可视化。

- 科学可视化是解释大量数据的最有效手段，因而首先应用在科学与工程计算领域中，并发展为科学可视化的研究领域。其主要过程是建模和渲染。
- 信息可视化是跨学科领域的大规模非数值型信息资源的视觉展现，帮助人们理解和分析数据。信息可视化致力于创建以直观方式传达抽象信息的手段和方法。
- 数据可视化是关于数据的视觉表现形式的研究。将大型数据集中的数据以图形图像形式表示，主要借助图形化手段，清晰有效地传达与沟通信息。

1.3.2　数据信息的展示方式

数据显示是将系统内部或外部存储器中的数据以可见或可读的形式进行输出，包括数据值直接显示、数据表显示、各种统计图形显示等形式。在地理信息系统中，反映空间信息的数据还能以图形或图像的形式显示。数据显示除与数据本身有关外，还与显示设备有关。对于高分辨彩色显示器、彩色绘图机等，不仅显示精度高，还可利用不同颜色表示不同数值。对于单色显示器、打印机等，需设计不同显示符号来表示不同数值，以增强显示效果。

1. 列表

将实验数据按一定规律用列表方式表达出来是记录和处理实验视觉最常用的方法。表格的设计要求对应关系清楚、简单明了、有利于发现相关量之间的物理关系，实验数据统计表格示例如图 1-5 所示，根据实际需求还可以列出除原始数据外的计算栏目和统计栏目等。

成分名称	含量	成分名称	含量	成分名称	含量
可食部	100	水分（g）	0.1	能量（kCal）	899
能量（KJ）	3761	蛋白质（g）	0	脂肪（g）	99.9
碳水化合物（g）	0	膳食纤维（g）	0	胆固醇（mg）	0
灰分（g）	0.1	维生素A（mg）	0	胡萝卜素（mg）	0
视黄醇（mg）	0	硫胺素（μg）	0	核黄素（mg）	0
尼克酸（mg）	0	维生素C（mg）	0	维生素（T）（mg）	42.06
a-E	17.45	（β-γ）-E	19.31	δ-E	5.3
钙（mg）	12	磷（mg）	15	钾（mg）	1
钠（mg）	3.5	镁（mg）	2	铁（mg）	2.9
锌（mg）	0.48	硒（mg）	0	铜（mg）	0.15
锰（mg）	0.33	碘（mg）	0		

图 1-5　实验数据

2. 作图

作图法可以直观地表达物理量之间的变化关系。从图线上还可以简单快速地计算出实验需要的某些结果，如直线的斜率和截距值等，读出没有进行观测的对应点（内插法），或在一定条件下从图线的延伸部分读取到测量范围以外的对应点（外推法），如图 1-6 所示。

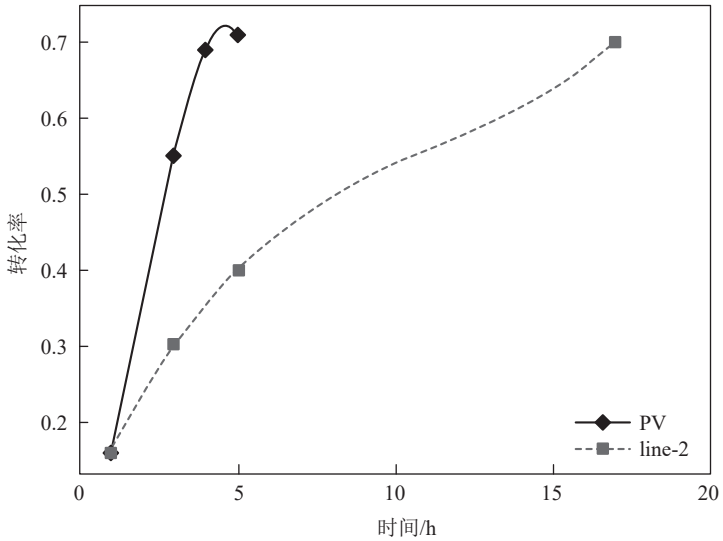


图 1-6　作图法

3. 图表

（1）直方图

直方图是将一个变量的不同等级的相对频数用矩形块标绘的图表。直方图又称柱形图、质量分布图，是一种统计报告图。直方图中一系列高度不等的纵向条纹或线段表示数据分布的情况，如图 1-7 所示。

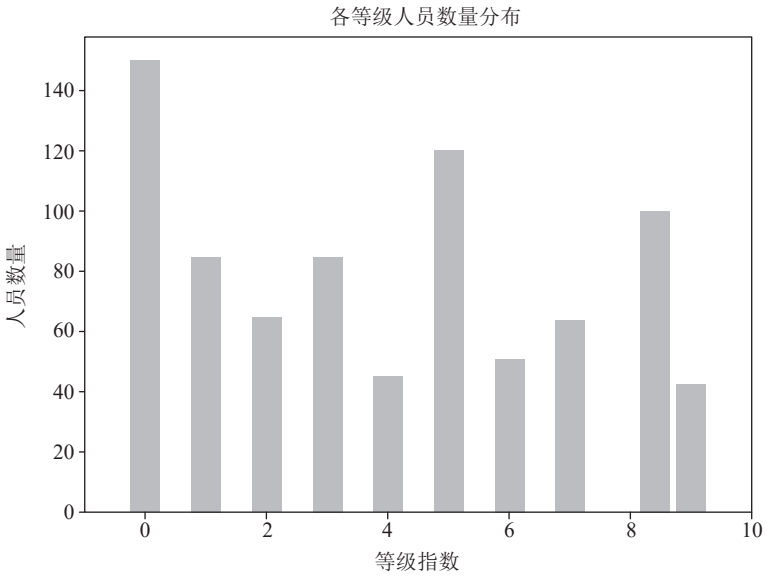


图 1-7　直方图

（2）散点图

散点图表示因变量随自变量变化的大致趋势，据此可以选择合适的函数对数据点进行拟合，如图 1-8 所示。

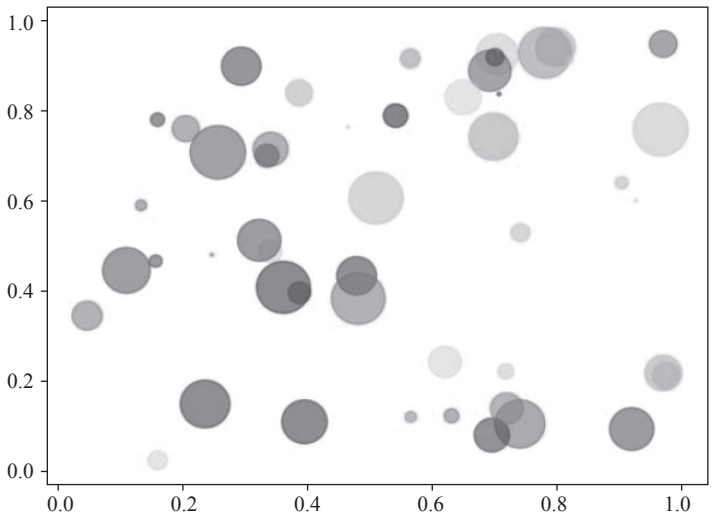


图 1-8　散点图

1.4　数据可视化技术

本节主要介绍数据可视化技术。

1.4.1　数据可视化技术的概念

数据可视化技术是一个处于不断演变之中的概念，其边界在不断扩大。数据可视化技术主要是指技术上较为高级的技术方法，这些技术方法允许利用图形、图像处理、计算机视觉及用户界面，通过表达、建模和对立体、表面、属性及动画的显示，对数据加以可视化解释。与立体建模之类的特殊技术方法相比，数据可视化所涵盖的技术方法要广泛得多。

在获得计算机图形学发展后，先后经历了科学可视化、信息可视化和数据可视化 3 个阶段。最初由科研人员提出科学建模和数据的可视化需求，进入 20 世纪 90 年代后，出现大量单机数据可视化需求，Excel 是这个时期的代表，互联网时代各种产品兴起，大数据爆发式增长，促使数据可视化技术飞速发展。

借助图形化的手段，数据可视化可以清晰、快捷、有效地传达与沟通信息。从用户的角度来看，数据可视化可以让用户快速抓住要点信息，让关键的数据点经由用户的眼睛快速直达用户的心灵深处。数据可视化通常具备如下 3 个特点：准确性、创新性、简洁性。

1.4.2　常用的可视化技术方法

1. 面积可视化

面积可视化对同一类图形（如柱状、圆环等）的长度、高度和面积加以区别，以清晰地表达不同指标对应的值之间的对比情况，如图 1-9 所示。

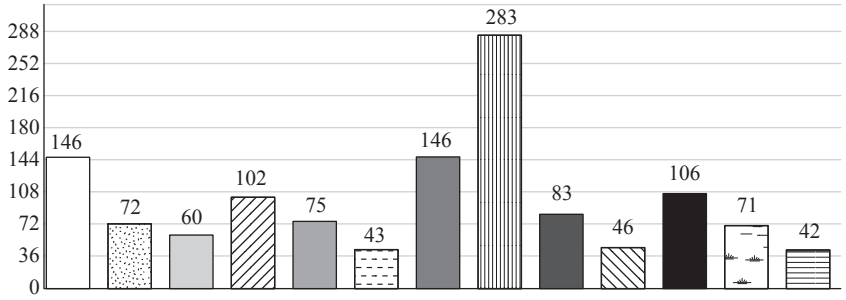


图 1-9　柱形图

2. 颜色可视化

颜色可视化通过颜色的深浅来表达指标值的强弱和大小，是数据可视化设计的常用方法，用户一眼便可看出哪部分指标的数据值更突出。

图 1-10 所示为上海二手房的房屋分布热力图，通过对上海地图单位的划分，用不同的颜色来代表不同的房屋分布密度，全上海的二手房状况便尽收眼底了。

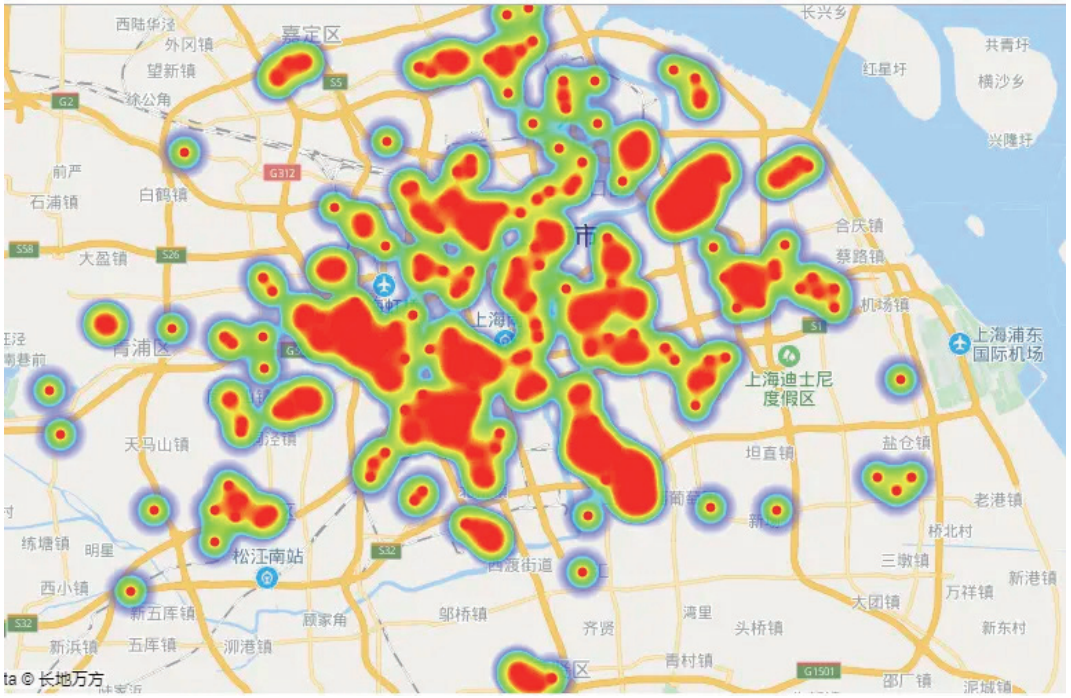


图 1-10　颜色可视化图

3. 图形可视化

在设计指标及数据时，结合有对应实际含义的图形，会使数据图表展现得更加生动，更便于用户理解图表要表达的主题。

图 1-11 所示为各类不同型号的手机在某个月的销售数量，可以让用户第一眼就能看到这些图的大小，直观而清晰。

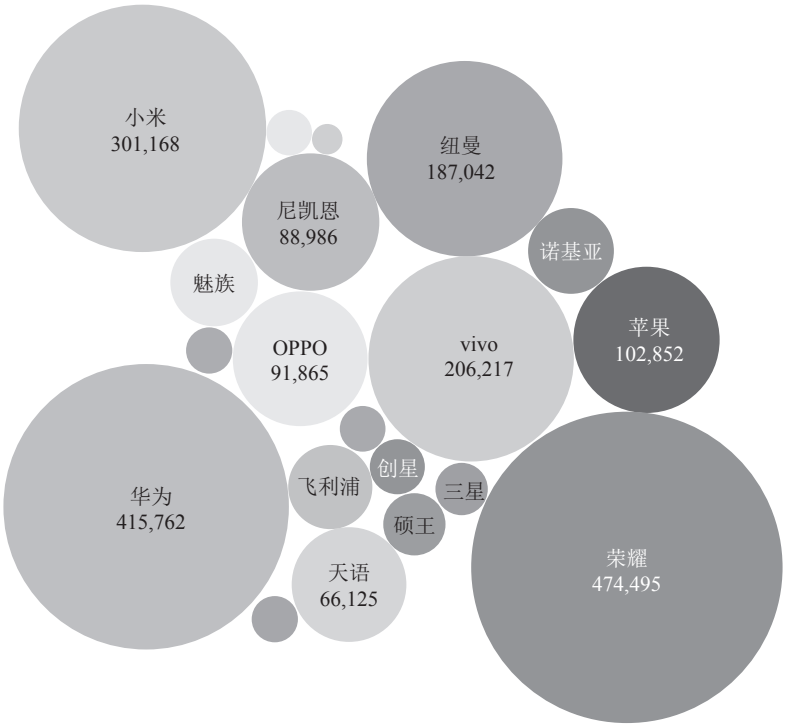


图 1-11　图形可视化

4. 地域可视化

当指标数据要表达的主题与地域有关联时，一般会选择以地图为大背景。这样用户可以直观地了解整体的数据情况，同时也可以根据地理位置快速定位到某一地区来查看详细数据。

5. 概念可视化

通过将抽象的指标数据转换成熟悉的容易感知的数据，使用户更容易理解图形要表达的意义。

图 1-12 所示为一张广告图，用了概念转换的方法，让用户清晰地感受到用纸量之多。如果只是描述擦手纸的量及堆积可达高度，观看者可能还没有什么概念，但当看到用纸的堆积高度比世界最高建筑还高，同时需砍伐 500 多棵树时，想必观看者的节省纸张甚至禁用纸张的情怀便油然而生了。由此可见用概念转换的方法是多么重要和有效。

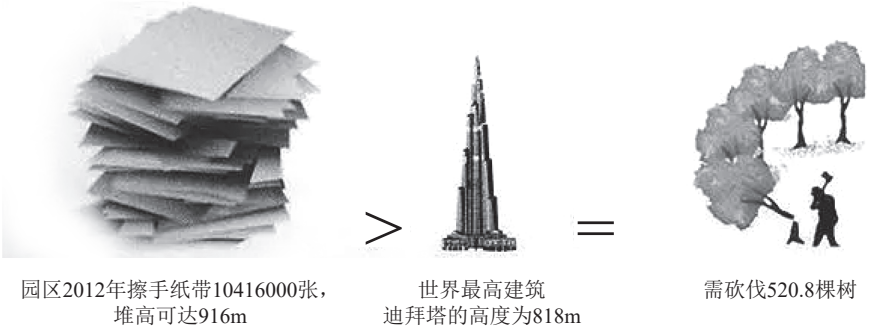


图 1-12　概念图

1.5 数据挖掘可视化

数据挖掘是指在大量的数据中挖掘出信息，通过认真分析来揭示数据之间有意义的联系、趋势和模式。数据挖掘技术是指为了完成数据挖掘任务所需要的全部技术。

与数据挖掘相近的词还有数据融合、数据分析和决策支持等。这个定义包括如下几层含义：数据源必须是真实的、大量的、含噪声的：发现的是用户感兴趣的知识；发现的知识要可接受、可理解、可运用：并不要求发现放之四海皆准的知识，仅支持发现特定的问题。

数据挖掘的目标是建立一个决策模型，根据过去的行动数据来预测未来的行为。例如，分析一家公司的不同用户对公司产品的购买情况，进而分析出哪类客户会对公司的产品有兴趣。在讲究实时、竞争激烈的网络时代，若能事先破解消费者的行为模式，是公司获利的关键因素之一。数据挖掘是一门交叉学科，它涉及数据库、人工智能、统计学、可视化等不同的学科和领域。

数据挖掘是从大量数据中寻找规律的技术，主要包括数据准备、规律寻找和规律表示 3 个步骤，因此，认为数据挖掘必须包括以下因素：

- 数据挖掘的本源：大量、完整的数据。
- 数据挖掘的结果：知识、规则。
- 结果的隐含性：需要一个挖掘过程。

下面介绍 4 种数据挖掘可视化方法。

1.5.1 文本挖掘

文本挖掘是将不同的文档进行比较之后，对文档的重要性和相关性进行排列，以整理出文档的模式和趋势。文本挖掘的处理过程包括对大量文档集的内容进行预处理、特征提取、结构分析等，其中不仅需要处理结构化和非结构化文档数据，还需要处理复杂的语义关系。文本挖掘流程如图 1-13 所示。

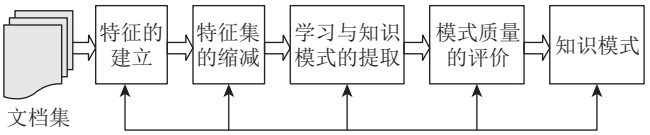


图 1-13 文本挖掘流程

文本是指从文档中抽取关键信息，用简洁的形式对文档内容进行摘要或解释，使得用户不需要浏览全文就可以了解文档或文档集合的总体内容。文本总结在有些场合非常有用，例如，搜索引擎在向用户返回查询结果时，通常需要给出文档的摘要。目前，绝大部分搜索引擎采用的方法是简单地截取文档的前几行。在对文档进行特征提取前，需要先进行文本信息的预处理：对英文而言，需进行词干提取（Stemming）处理；中文的情况则不同，因为中文的词与词之间没有固有的间隔符，所以，需要进行分词处理。在中文信息处理领域，对中文自动分词研究已经比较多了，提出了一些分词方法，如最大匹配法、逐词遍历匹配法、最小匹配法等。

1.5.2　Web 挖掘

Web 挖掘是指从与 WWW 有关联的资源及行为中获取有用的模式和隐含的信息。Web 中含有大量信息和超链接信息、Web 页面的访问和使用信息，这些信息是进行 Web 挖掘的重要资源。

Web 挖掘流程图如图 1-14 所示。

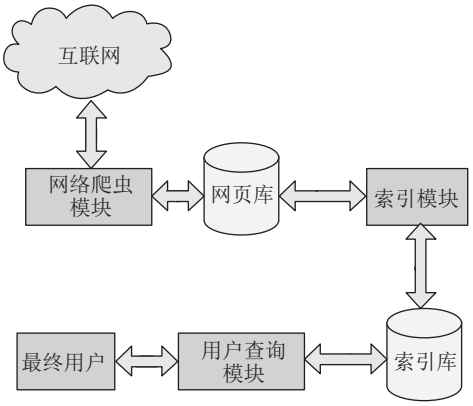


图 1-14　Web 挖掘流程图

Web 挖掘可以分为 3 类：Web 内容挖掘、Web 结构挖掘和 Web 使用记录挖掘。其中，Web 结构挖掘的对象是 Web 本身的超链接，包括页面内部的结构及页面之间的结构。对于给定的 Web 文档集合，运用引用分析方法找到同一网站内部及不同网站之间的链接关系，通过算法发现它们之间的链接情况的有用信息。挖掘 Web 结构信息对于导航用户浏览行为、改进站点设计、评价页面的重要性等都非常重要。Web 结构挖掘通常需要 Web 的全局数据，因此，在个性化搜索引擎或主题搜索引擎研究领域得到了广泛的应用。

1.5.3　多媒体数据挖掘

多媒体数据类型包括图形、视频、音频时空数据和超文本等，隐藏了大量有价值的知识，而多媒体数据的挖掘是综合分析大量多媒体数据的试听特征与语义，利用多媒体的时间、空间、视觉特性、视听对象及运动特性，挖掘具有一定价值的、能够理解的知识模式，找出实践的趋势及关联性。

因此，多媒体数据集中必定存在关于信息主体的特征、属性及它们之间的关系，或者存在某些人们从直观上无法得到的模式。多媒体数据挖掘是一种智能的数据分析，旨在从特定的多媒体数据集中发现必要的结果来用于决策、对策及融合分析。例如，在 MMMiner（MultiMediaMiner，多媒体挖掘）中查找包含人脸的所有图像，用户界面上就会逐步显示不同国家、不同肤色、不同表情的人脸，而不是显示一些猴脸或马脸。由此可见，多媒体挖掘就是从大量的多媒体数据集中，通过综合分析视听特性和语义，发现隐含的、有效的、有价值的、可理解的模式，得出事件的趋向和关联，为用户提供问题求解层次的决策支持能力。

1.5.4　时空数据挖掘

时空数据挖掘是指从海量、高维、高噪声和非线性的时空数据中提取隐含的、人们事先不知的、潜在的有用信息及知识的过程。时间维度和空间维度的存在增加了时空数据挖掘的复杂性。

随着传感器网络、手持移动设备等的普遍应用，遥感卫星和地理信息系统等的显著进步，人们获取了大量的地理科学数据。这些数据内嵌于连续空间，并且随时间动态变化，具有很大程度的特殊性和复杂性。实际上，在很多应用领域，如交通运输、气象研究、地震救援、犯罪分析、公共卫生与医疗等，在问题求解过程中需要同时考虑时间和空间两方面因素。随着信息技术的发展，人们已经不满足于单纯的时空数据的存储和展现，而是需要更先进的手段帮助理解时空数据的变化。如何从这些复杂、海量、高维、高噪声和非线性的时空数据中挖掘出隐含

的时空模式，对这些模式进行分析从而提取出有价值的信息并用于商业活动，是对时空数据挖掘及分析技术的一项极大的挑战。

根据时空对象的不同，时空预测有不同的分类，大致可以分为 3 类：面向时空数据的位置和轨迹预测、密度和事件预测、结合空间的时间序列预测。

1. 面向时空数据的位置和轨迹预测

面向时空数据的位置预测主要是基于时空对象的特征构建预测模型来预测时空对象所在的具体空间位置。对于实时物流、实时交通管理、基于位置的服务和 GPS 导航等涉及时空数据的应用而言，预测单个或者一组对象未来的位置或目的地是至关重要的，它能使系统在延误的情况下采取必要的补救措施，避免拥堵，提高效率。

2. 密度和事件预测

某个区域的对象密度定义为在给定时间点该区域内对象数与该区域大小之比。这是一些对象随时间变化而呈现出的一个全局特征。面向时空数据的密度预测主要应用于实时交通管理，会给及时改善交通拥堵带来很大助益。例如，交通管理系统通过密度预测，可以识别出道路中的密集区域，从而帮助用户避免陷入交通阻塞，并采取有效措施，及时缓解交通拥堵。此外，面向时空数据的事件预测可以根据历史数据（时间序列），结合地理区域密度估计（发现重要特征和时空地点）来预测给定空间位置和时间范围内的概率密度，譬如，基于过去犯罪事件发生的地点、时间和城市经济等特征预测给定区域和时间段内犯罪发生的概率，进而检测犯罪率发展趋势，有效降低城市犯罪率。

3. 结合空间的时间序列预测

结合空间的时间序列预测是从时间的角度来考虑时空数据。与传统的时间序列不同的是，与空间有关的时间序列彼此不是独立的，而是和空间相关的。例如，可以首先构造时间序列模型以获取每个独立空间区域的时间特性；然后构造神经网络模型拟合隐含的空间相关性；最后基于统计回归，结合时间和空间进行预测，获得综合预测。

1.6 本章习题

一、填空题

1. 大数据指在一定时间内无法使用常规软件工具进行_____、_____和_____的数据集合。

2. 麦肯锡全球研究所给出的大数据定义是一种规模大到在_____、_____、_____、_____方面超出了传统数据库软件工具能力范围的数据集合。

3. 从技术上看，大数据与_____的关系就像一枚硬币的正反面一样密不可分。

4. 大数据最小的基本单位是_____。

5. 数据可视化信息展示方式包括_____、_____（两种）。

6. 可视化被分为 3 种类型，如_____、_____、_____。

7. 大数据可以分为_____、_____、_____、_____等领域。

8. FMEA 是一种可靠性设计的重要方法，是以_____和_____的组合。

9. 常用可视化技术方法有_____、_____、_____、_____、_____。

10. 数据挖掘可视化有_____、_____。

二、问答题

1. 大数据结构可以从哪 3 个层面展开？
2. 大数据基本单位有哪些？（写出 4 种）
3. 设计方式 FMEA 是什么？
4. 大数据信息展示方式的列表是什么意思？
5. 信息展示方式的图表有哪些？举例 1 ～ 2 种。
6. 简介数据可视化的 3 种类型。
7. 常用可视化技术方法有哪些？
8. 数据挖掘可视化中的文本挖掘是什么？

第 2 章
Excel 图表可视化

本章学习目标：

- 掌握 Excel 的基本操作。
- 掌握数据采集、数据存储、数据处理和数据分析。
- 学会使用 Excel 实现数据可视化的案例。

本章首先介绍 Excel 的功能和基本操作；然后介绍如何使用 Excel 实现可视化。

2.1 Excel 简介

Microsoft Office Excel 是电子数据表程序（进行数字和预算运算的软件程序），是最早的 Office 组件。Excel 内置了多种函数，可以对大量数据进行分类、排序、绘制图表等。

1. Excel 的功能

Excel 的功能如图 2-1 所示。

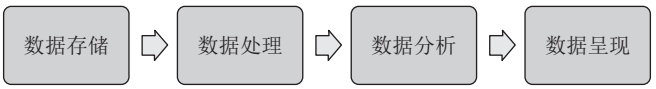


图 2-1　Excel 的功能

2. Excel 中的重要概念

在 Excel 中有以下几个重要概念：

- 文件类型：Excel 的几种常用文件类型包括 XLS/XLSX 工作簿文件、XLW 工作区文件。
- 工作簿：Excel 环境中用来存储并处理工作数据的文件，即 Excel 文档。
- 工作表：显示在工作簿窗口中的表格。
- 单元格：工作表的基本单位，由行标和列标唯一确定。

3. Excel 界面

Excel 界面展示如图 2-2 所示。

4. Excel 单元格数字格式

在 Excel 中，系统数字格式效果如图 2-3 所示。除了系统定义的数字格式，用户还可以根据规定自定义数字格式，例如，yyyy/mm/dd 表示 2023/03/01，而 yyyy/mm/dd 表示 2023/Jan/Thu，具体的自定义格式可根据实际情况制定。

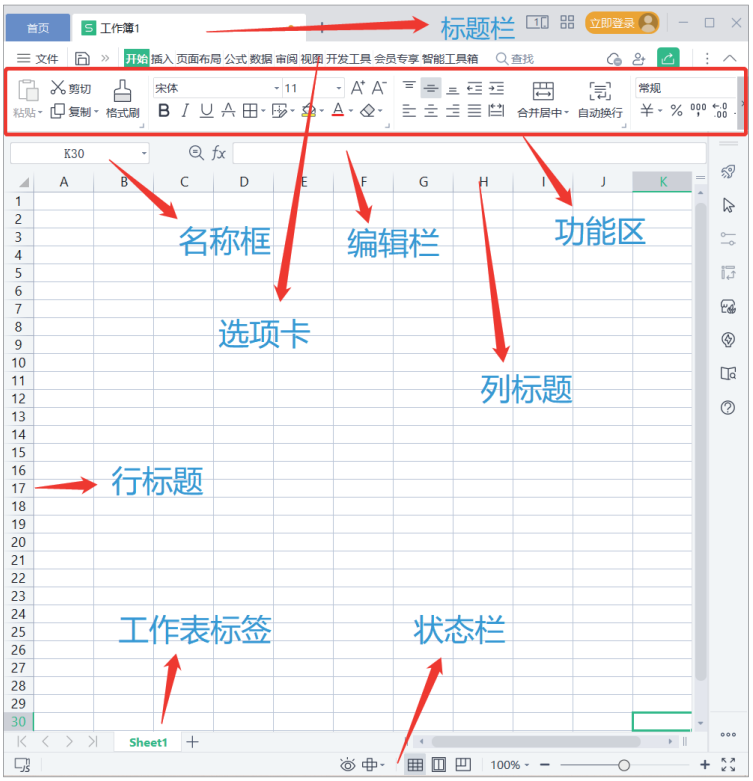


图 2-2　Excel 界面

类型	原格式	转变后的格式
数值	−25636	−25636.00
货币	10000	￥10,000.00
会计专用	1555	￥1,555.00
日期	49814	2030−11−27
百分比	0.11	11.00%
分数	0.6	3/5
科学计数	120014	1.20E+0.5
文本	2422	2422
特殊	25368	二万五千三百六十八

图 2-3　Excel 中的系统数字格式效果

自定义数字格式如图 2-4 所示。

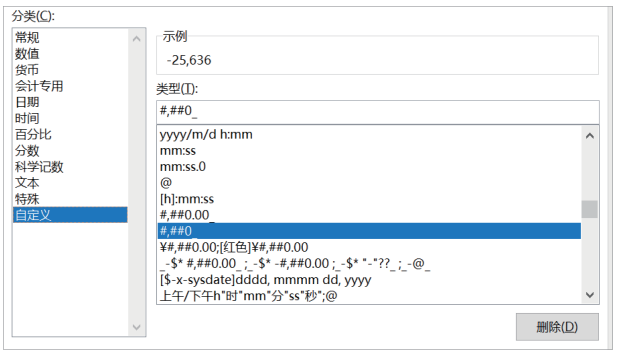


图 2-4　Excel 自定义数字格式设置

5. Excel 的基本操作

（1）查找与替换

- 按值查找或按格式查找：使用 Ctrl+F 组合键调出“查找和替换”对话框，根据值进行操作。
- 模糊查找：使用 Ctrl+F 组合键调出“查找和替换”对话框，然后使用通配符（*、?、～）进行操作。通配符的含义如下：
  - *：全通配。
  - ?：单字符统配。
  - ～：将 * 和 ? 转换为普通字符，而不作为通配符使用。

（2）定位

- 定位：通过名称框定位单元格及区域位置。
- 定义名称：在 Excel 中用户可选中工作表中常用的区域并定义名称，之后可在名称框中输入名称进行定位。
- 使用定位条件：按 Ctrl+G 组合键，再按 Alt+S 组合键，可根据条件进行定位操作。

（3）筛选

- 自动筛选：在“开始”选项卡中选择“排序和筛选”→“筛选”命令，进行数值筛选 / 文本筛选等。
- 高级筛选：在“数据”选项卡的“排序和筛选”选项组中单击“高级”按钮，将弹出“高级筛选”对话框，可以使用高级筛选完成一些自动筛选不能完成的逻辑运算，如“或”关系。

（4）分类汇总

分类汇总即把资料进行数据化后，先按照某一标准进行分类，然后在分完类的基础上对各类别的相关数据分别进行求和、求平均数、求个数、求最大值、求最小值等方法的汇总。

6. Excel 图表元素说明

下面先来认识图表中的元素。图表中的元素可用“七块积木 + 一个核心”概括，七块积木分别为①图表标题、②坐标轴标题、③图例、④数据标签、⑤模拟运算表、⑥坐标轴（线条 + 刻度 + 标签）、⑦网格线；一个核心为主次坐标系，如图 2-5 所示。

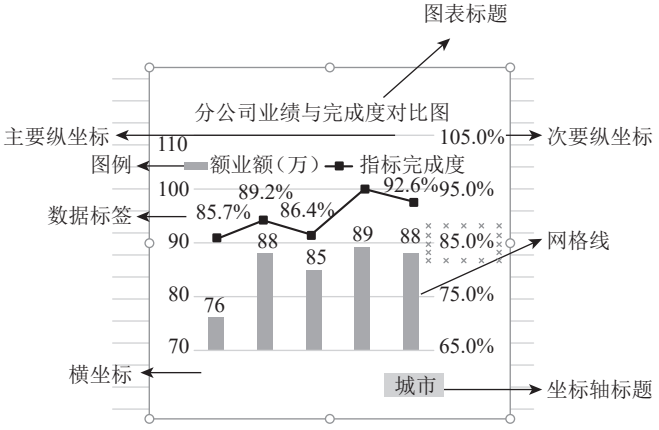


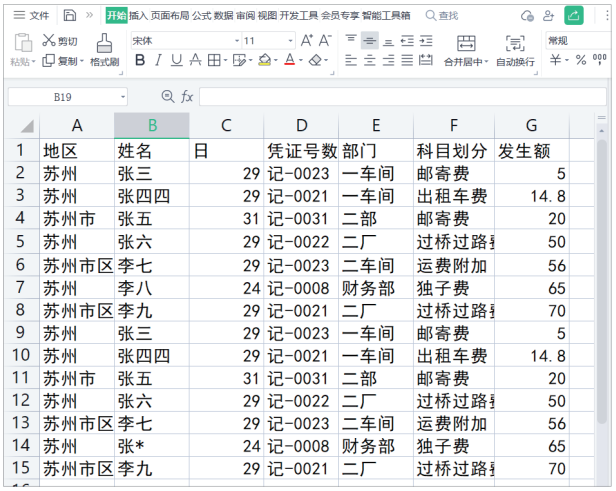
图 2-5　图表元素说明

2.2　Excel 实现可视化案例

使用 Excel 实现可视化的操作步骤如下：

Step01 向工作表中输入数据或自行添加类似数据。数据内容如图 2-6 所示。

Step02 右击任一数据，在弹出的快捷菜单中选择“设置单元格格式”命令，比较不同数据的单元格数字格式有何不同，如图 2-7 所示。



	A	B	C	D	E	F	G
1	地区	姓名	日	凭证号数	部门	科目划分	发生额
2	苏州	张三	29	记-0023	一车间	邮寄费	5
3	苏州	张四四	29	记-0021	一车间	出租车费	14.8
4	苏州市	张五	31	记-0031	二部	邮寄费	20
5	苏州	张六	29	记-0022	二厂	过桥过路	50
6	苏州市区	李七	29	记-0023	二车间	运费附加	56
7	苏州	李八	24	记-0008	财务部	独子费	65
8	苏州市区	李九	29	记-0021	二厂	过桥过路	70
9	苏州	张三	29	记-0023	一车间	邮寄费	5
10	苏州	张四四	29	记-0021	一车间	出租车费	14.8
11	苏州市	张五	31	记-0031	二部	邮寄费	20
12	苏州	张六	29	记-0022	二厂	过桥过路	50
13	苏州市区	李七	29	记-0023	二车间	运费附加	56
14	苏州	张*	24	记-0008	财务部	独子费	65
15	苏州市区	李九	29	记-0021	二厂	过桥过路	70

图 2-6　输入数据

类型	原格式	转变后的格式
数值	-25636	-25636.00
货币	10000	¥10,000.00
会计专用	1555	¥　1,555.00
日期	44354	2021-6-7
时间	0.69805	16:45:12
百分比	0.12	12.00%
科学计数	124565	1.25E+05
文本	2456	2456
特殊	56452	056452

图 2-7　不同的数字格式

Step03 在“开始”选项卡的“编辑”选项组中单击“查找和选择”按钮，在弹出的下拉菜单中选择“替换”命令，弹出“查找和替换”对话框。

在“查找内容”文本框中输入“苏州”，在“替换为”文本框中输入“苏州市”，勾选“单元格匹配”复选框，如图 2-8 所示。单击“全部替换”按钮，将“苏州”全部替换为“苏州市”。同理，将“苏州市区”也全部替换为“苏州市”，结果如图 2-9 所示。

图 2-8　替换操作

	A	B	C	D	E	F	G
	地区	姓名	日	凭证号数	部门	科目划分	发生额
	苏州市	张三	29	记-0023	一车间	邮寄费	5
	苏州市	张四四	29	记-0021	一车间	出租车费	14.8
	苏州市	张五	31	记-0031	二部	邮寄费	20
	苏州市	张六	29	记-0022	二厂	过桥过路	50
	苏州市	李七	29	记-0023	二车间	运费附加	56
	苏州市	李八	24	记-0008	财务部	独子费	65
	苏州市	李九	29	记-0021	二厂	过桥过路	70
	苏州市	张三	29	记-0023	一车间	邮寄费	5
	苏州市	张四四	29	记-0021	一车间	出租车费	14.8
	苏州市	张五	31	记-0031	二部	邮寄费	20
	苏州市	张六	29	记-0022	二厂	过桥过路	50
	苏州市	李七	29	记-0023	二车间	运费附加	56
	苏州市	张*	24	记-0008	财务部	独子费	65
	苏州市	李九	29	记-0021	二厂	过桥过路	70

图 2-9　替换的结果

Step04 任选几组数据，右击，在弹出的快捷菜单中选择“插入批注”命令，如图 2-10 所示。

在“开始”选项卡的“编辑”选项组中单击“查找和选择”按钮，在弹出的下拉菜单中选择“定位条件”（Alt+S）命令，在弹出的“定位条件”对话框中选择“批注”单选按钮，然后单击“确定”按钮。结果使用了定位工具，定位到了前面插入批注的单元格。

添加批注后，单元格的格式如图 2-11 所示。

地区	姓名	日	凭证号数	部门	科目划分	发生额
苏州	张三	29	记-0023	一车间	邮寄费	5.00
苏州	张四四	29	记-0021	一车间	出租车费	14.80
苏州市	张五	31	记-0031	二部	邮寄费	20.00
苏州	张六	29	记-0022	二厂	过桥过路费	50.00
苏州市区	李七	29	记-0023	二车间	运费附加	56.00
苏州	李八	24	记-0008	财务部	独子费	65.00
苏州市区	李九	29	记-0021	二厂	过桥过路费	70.00
苏州	张三	29	记-0023	一车间	邮寄费	5.00
苏州	张四四	29	记-0021	一车间	出租车费	14.80
苏州市	张五	31	记-0031	二部	邮寄费	20.00
苏州	张六	29	记-0022	二厂	过桥过路费	50.00
苏州市区	李七	29	记-0023	二车间	运费附加	56.00
苏州	张*	24	记-0008	财务部	独子费	65.00
苏州市区	李九	29	记-0021	二厂	过桥过路费	70.00

听阴天说什么：
故障

图 2-10　插入批注

	A	B	C	D	E	F	G
1	地区	姓名	日	凭证号数	部门	科目划分	发生额
2	苏州	张三	29	记-0023	一车间	邮寄费	5.00
3	苏州	张四四	29	记-0021	一车间	出租车费	14.80
4	苏州市	张五	31	记-0031	二部	邮寄费	20.00
5	苏州	张六	29	记-0022	二厂	过桥过路费	50.00
6	苏州市区	李七	29	记-0023	二车间	运费附加	56.00
7	苏州	李八	24	记-0008	财务部	独子费	65.00
8	苏州市区	李九	29	记-0021	二厂	过桥过路费	70.00
9	苏州	张三	29	记-0023	一车间	邮寄费	5.00
10	苏州	张四四	29	记-0021	一车间	出租车费	14.80
11	苏州市	张五	31	记-0031	二部	邮寄费	20.00
12	苏州	张六	29	记-0022	二厂	过桥过路费	50.00
13	苏州市区	李七	29	记-0023	二车间	运费附加	56.00
14	苏州	张*	24	记-0008	财务部	独子费	65.00
15	苏州市区	李九	29	记-0021	二厂	过桥过路费	70.00

图 2-11　添加注释

Step05 筛选数据。

选中第一行数据，单击“数据”选项卡中的“筛选”图标按钮，表头出现下拉框。单击“部门”后的下三角按钮，在下拉列表中只勾选“财务部”复选框，将部门为“财务部”的数据筛选出来，如图 2-12 所示。

	A	B	C	D	E	F	G
1	地区	姓名	日	凭证号	部门	科目划	发生额
4	苏州市	李八	24	记-0008	财务部	独子费	65
5	苏州市	张*	24	记-0008	财务部	独子费	65

图 2-12　筛选功能

Step06 数据分析（分类汇总、函数与公式）。

继续使用前面已经排好序的数据。单击“数据”选项卡中的“分类汇总”图标按钮，在“分类字段”下拉列表框中选择“部门”，在“汇总方式”下拉列表框中选择“求和”，在“选定汇总项”列表框中勾选“发生额”复选框，如图 2-13 所示。单击“确定”按钮，这样工作表中就会显示汇总后的数据。还可以通过切换左上角的序号来查看汇总明细，如图 2-14 所示。

Step07 认识函数与公式。

函数的使用方法：以等号开头，函数名在中间，以括号结尾，括号中间写参数。

函数公式能够在编辑栏中显示，如图 2-15 所示。

简单的函数及其功能如下：

- SUM：求和。
- AVERAGE：求平均。
- MAX：求最大值。
- MIN：求最小值。

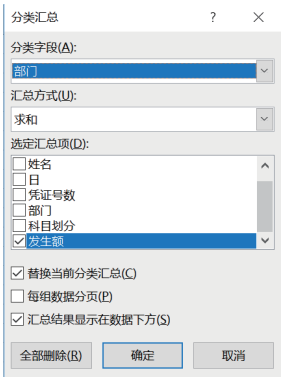
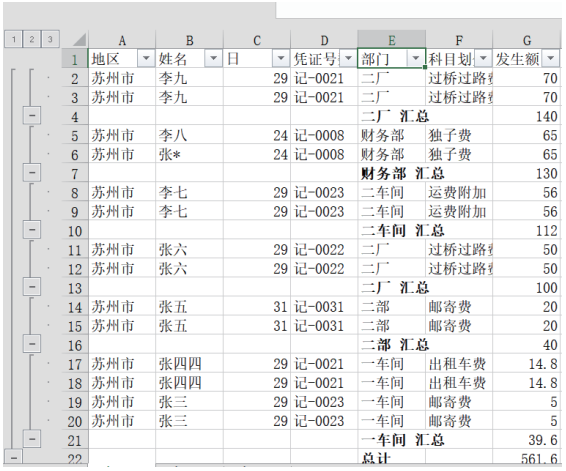


图 2-13　分类汇总



地区	姓名	日	凭证号	部门	科目划	发生额
苏州市	李九	29	记-0021	二厂	过桥过路	70
苏州市	李九	29	记-0021	二厂	过桥过路	70
				二厂 汇总		140
苏州市	李八	24	记-0008	财务部	独子费	65
苏州市	张*	24	记-0008	财务部	独子费	65
				财务部 汇总		130
苏州市	李七	29	记-0023	二车间	运费附加	56
苏州市	李七	29	记-0023	二车间	运费附加	56
				二车间 汇总		112
苏州市	张六	29	记-0022	二厂	过桥过路	50
苏州市	张六	29	记-0022	二厂	过桥过路	50
				二厂 汇总		100
苏州市	张五	31	记-0031	二部	邮寄费	20
苏州市	张五	31	记-0031	二部	邮寄费	20
				二部 汇总		40
苏州市	张四四	29	记-0021	一车间	出租车费	14.8
苏州市	张四四	29	记-0021	一车间	出租车费	14.8
苏州市	张三	29	记-0023	一车间	邮寄费	5
苏州市	张三	29	记-0023	一车间	邮寄费	5
				一车间 汇总		39.6
				总计		561.6

图 2-14　汇总明细

Step 08 数据可视化。

向工作表中输入数据，选中数据，切换到“插入”选项卡，在“图表”选项组中选择“条形图”→“二维条形图”→“堆积条形图”，Excel 根据选中的数据自动生成一张堆积条形图，如图 2-16 所示。

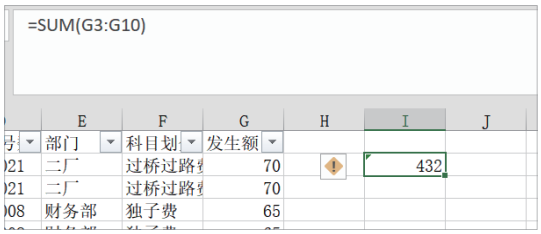


图 2-15　Excel 公式

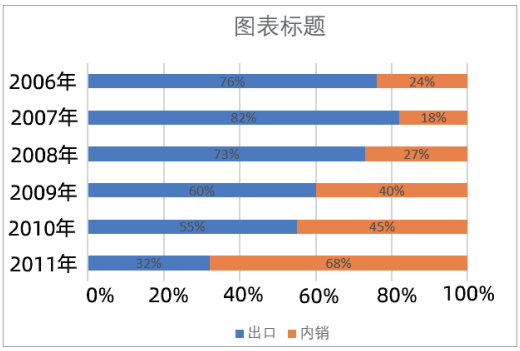


图 2-16　堆积条形图

Step 09 设置次要坐标系。

在图 2-16 中选中内销数据，右击，在弹出的快捷菜单中选择“设置数据系列格式”命令，在弹出的“设置数据系列格式”对话框中选择“次坐标轴”单选按钮，设置次要坐标，如图 2-17 所示。

Step 10 将垂直坐标轴的标签设置为高。选中坐标轴，右击，在弹出的快捷菜单中选择“设置坐标轴格式”命令，在弹出的“设置坐标轴格式”对话框中将“标签位置”设置为“高”，如图 2-18 所示。

Step 11 设置坐标轴数字格式。在“设置坐标轴格式”对话框中对坐标轴数字格式进行设置，具体设置如图 2-19 所示。

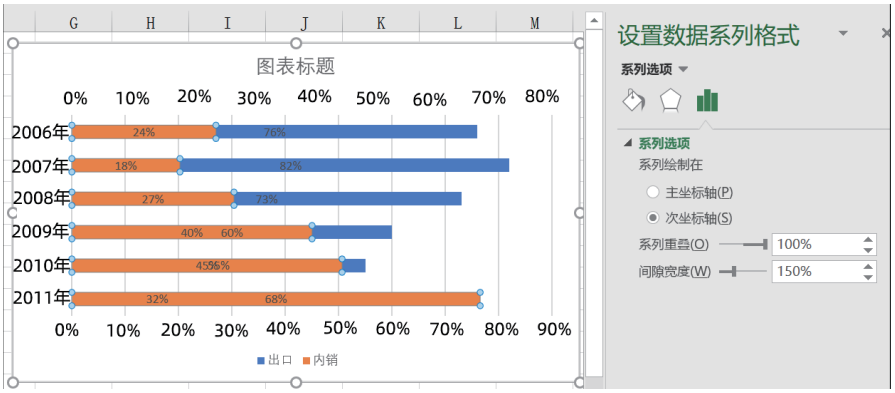


图 2-17　设置次要坐标系

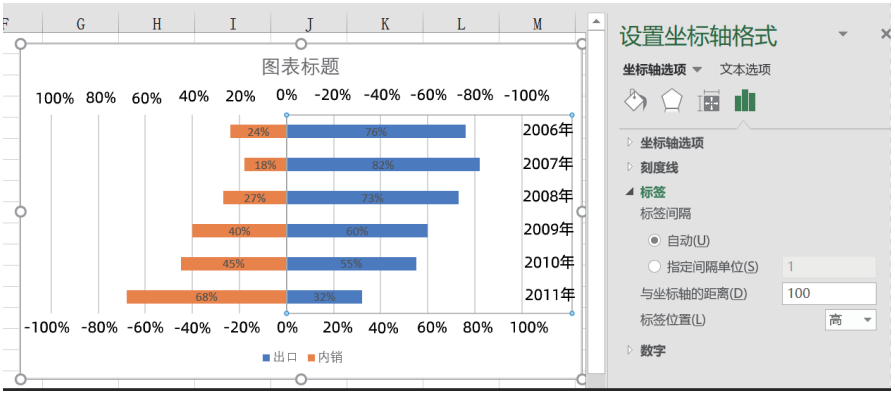


图 2-18　设置标签

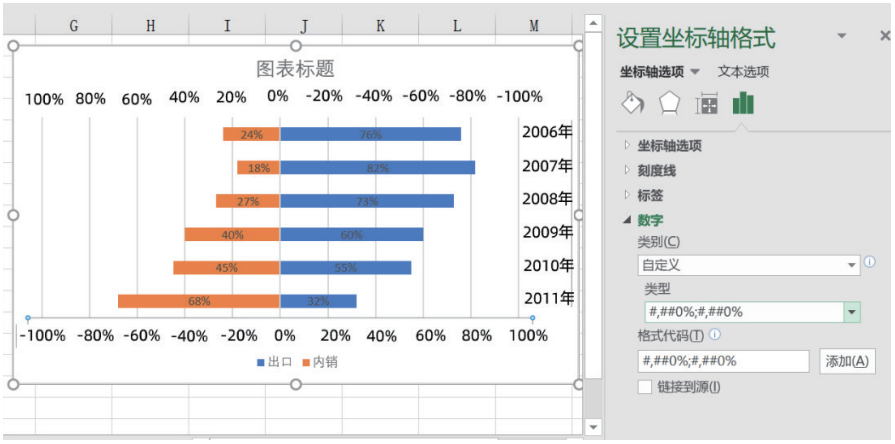


图 2-19　设置坐标轴数字格式

Step 12 设置主要坐标轴刻度线为外部，删除次要坐标轴与网格线，具体设置如图 2-20 所示。

Step 13 设置图表区背景为图片。选中图表区，右击，在弹出的快捷菜单中选择“设置图表区域格式”命令，在弹出的对话框中选择“图片或纹理填充”单选按钮，选择素材图片，如图 2-21 所示。

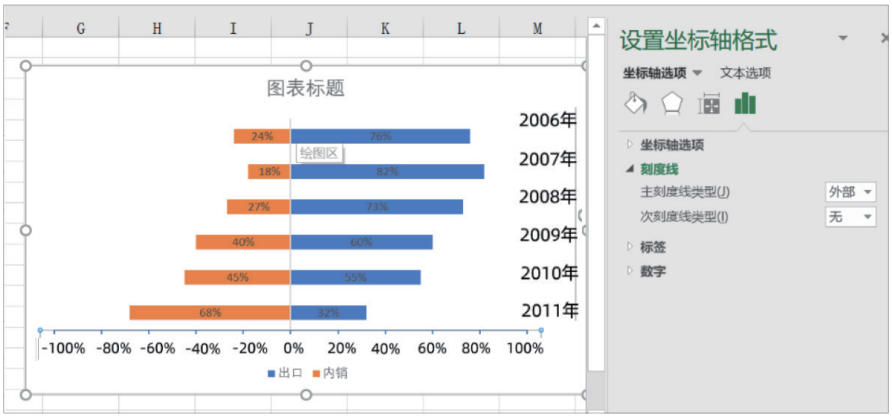


图 2-20　设置坐标轴刻度格式

图 2-21　设置背景图片

背景图无固定要求，读者可选择自己喜欢的素材图片来进行设置。

Step 14 调整条形颜色与间距，具体设置如图 2-22 所示。

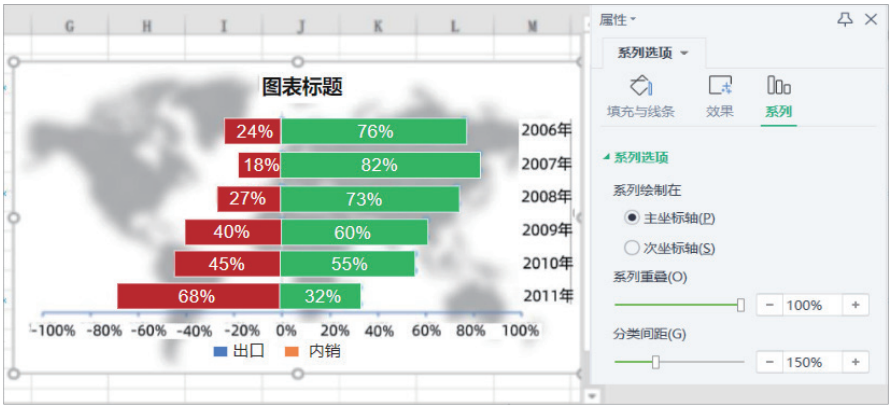


图 2-22　设置条形颜色与间距

Step 15 美化图表（整理图例、标题等），效果如图 2-23 所示。

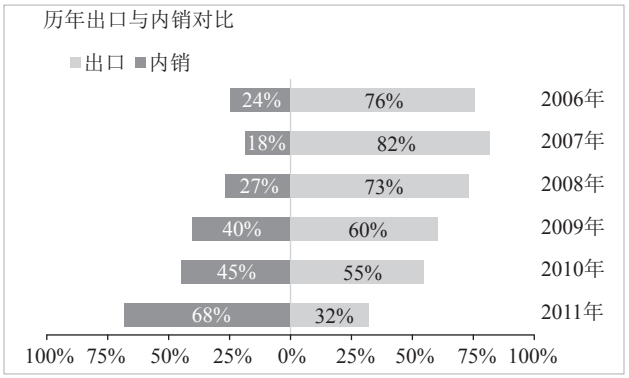


图 2-23　美化结果

2.3　本章习题

一、单选题

1. Excel 中下面哪个数据格式能表示时间 2023/04/01？（　　）

A. yy/m/dd　　B. yyyy/mm/dd　　C. yyy/m/dd　　D. yyy/mmm/ddd

2. 将两组数据值域差距较大的数据绘制在一张图上，应该（　　）。

A. 调整坐标轴值域　　B. 删除坐标轴　　C. 添加次坐标系　　D. 调整刻度线

3. Office 组件中属于表格处理的是（　　）。

A. Excel　　B. Word　　C. PowerPoint　　D. Outlook

4. Excel 不能做什么？（　　）

A. 数据存储　　B. 数据分析　　C. 数据处理　　D. 数据挖掘

5. Excel 的工作簿扩展名为（　　）。（多选）

A. .xls　　B. .xlsx　　C. .xsl　　D. .xslx

二、判断题

1. Excel 中使用函数都应该以“=”开头。

2. Microsoft Office Word 是文字处理软件。它被认为是 Office 的主要程序。它在文字处理软件市场上拥有统治份额。

3. Microsoft Office Outlook 是微软公司推出的一款网页设计、制作、发布、管理的软件。

4. 分类汇总即把资料进行数据化后，先按照某一标准进行分类，然后在分完类的基础上对各类别进行汇总。

三、填空题

1. Excel 能完成的功能为_____、_____、_____和_____四部分。

2. 查找与替换的快捷键为_____；定位工具的快捷键为_____。

四、简答题

1. 简单介绍一下 Microsoft Office PowerPoint。

2. Excel 图表中的元素由哪些部分组成？

第 3 章 Tableau Desktop

本章学习目标：

- 了解 Tableau 的基本原理和技术。
- 了解 Tableau 的数据分析。
- 掌握 Tableau 的可视化分析。
- 了解 Tableau 的预测分析。
- 掌握 Tableau 仪表板的分享与发布。

本章首先介绍大数据处理软件 Tableau 的基本特征；然后介绍此软件如何安装与使用，包括 Tableau 数据处理、计算；接着重点介绍 Tableau 的可视化操作，如图表可视化、地图可视化；最后介绍 Tableau 的预测分析，以及如何分享所创建的可视化图像。

3.1 Tableau 简介

本节主要介绍 Tableau 的基本知识，使读者能够入门 Tableau。

3.1.1 Tableau 概述

Tableau 是用于数据可视分析的商业智能工具。用户可以创建和分发交互式、可共享的仪表板，以图形和图表的形式描绘数据的趋势、变化和密度。Tableau 可以连接到文件、关系数据源和大数据源来获取和处理数据。该软件允许数据混合和实时协作，这使得它非常独特。在企业、学术机构及许多政府机构，都会使用 Tableau 进行视觉数据分析。

Tableau 算是数据可视化比较容易入门的软件，只需简单地拖曳，就可以将各种类型的数据以多种图表的形式反映出来，然后将它们嵌入文档或者网页中。即便不是专门从事数据可视化方面的工作，也有必要学习 Tableau，可以通过它将数据组织好以后放入 Word 或者 PPT 中。

数据可视化技术是 Tableau 的核心，主要包括以下两个方面：

- 独创的 VizQL 数据库。Tableau 的初创合伙人是来自斯坦福大学的数据科学家，他们为了实现卓越的可视化数据获取与后期处理，并没有像普通数据分析类软件那样简单地调用和整合现行主流的关系型数据库，而是革命性地进行了大尺度的创新，独创了 VizQL 数据库。
- 用户体验良好且易用的表现形式。Tableau 提供了一个非常新颖且易用的使用界面，使得处理规模巨大的、多维的数据时，可以即时从不同角度看到数据所呈现出的规律。

Tableau 通过数据可视化技术，使得数据挖掘变得平民化，而其自动生成和展现出的图表，也丝毫不逊色于互联网美术编辑的水平。正是这个特点奠定了其广泛的用户基础（用户总数年均增长 126%），带来了高续订率（90% 的用户选择续订其服务）。

Tableau 主要有 3 个平台，分析数据使用的是 Tableau Desktop，直接从官网上下载 Tableau Desktop 的免费个人版即可。Tableau Desktop 的安装是一个非常直接的过程，需要接受许可协议并提供安装的目标文件夹。

3.1.2 Tableau 的工作界面

1. Tableau 工作区

Tableau 工作区包含菜单、工具栏、“数据”窗格、卡和功能区，以及一个或多个工作表。表可以是工作表、仪表板或故事。有关仪表板或故事工作区的详细信息，请参见创建仪表板或故事工作区。注：Tableau 中的卡是可视化特性，用来快速分析大量数据，帮助用户可视化和理解数据，从而帮助用户作出更好的决策。Tableau 中的故事工作区是一个可视化界面，用于帮助用户记录和展示他们的故事，创建和分享可视化演示文稿。

工作区区域如图 3-1 所示。

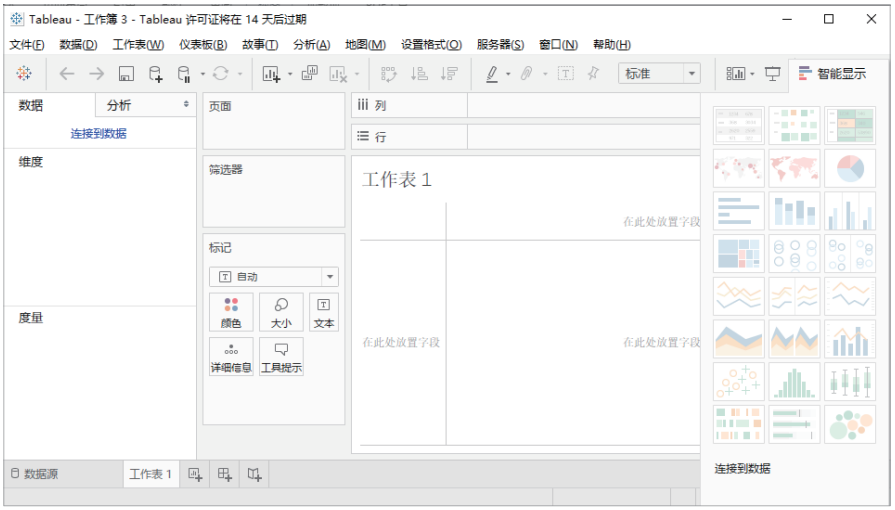


图 3-1　工作区

2. 工作簿区域

Tableau 的工作簿如图 3-2 所示，图中各字母标签的含义如下：

- A：工作簿名称。工作簿包含工作表，后者可以是工作表、仪表板或故事工作区。相关详细信息可参见工作簿和工作表。
- B：卡和功能区。将字段拖到工作区中的卡和功能区，可以将数据添加到视图中。
- C：工具栏。工具栏中提供快捷访问命令及分析和导航工具。
- D：单击此图标将转到“开始”页面，可以在其中连接数据。
- E：侧栏。在工作表中，侧栏区域包含“数据”窗格和“分析”窗格。
- F：单击此选项卡可以转到“数据源”页面并查看数据。
- G：状态栏。显示有关当前视图的信息。

- H：工作表标签。标签表示工作簿中的每张工作表，是工作表、仪表板或故事工作区。

图 3-2　工作簿

3. 显示和隐藏侧栏

在编辑工作表时，侧栏包含“数据”窗格和“分析”窗格。根据要在视图中进行的操作，这时可能会看到不同的窗格，包括“数据”“分析”“故事”“仪表板”“布局”和“格式”。关于侧栏最重要的一点是可以在工作区中展开和折叠此区域。

要在 Tableau Desktop 中隐藏侧栏，可单击侧栏中的折叠按钮，如图 3-3 所示。

图 3-3　单击折叠按钮隐藏侧栏

要在 Tableau Desktop 中显示侧栏，可单击工作区左下方状态栏中的展开按钮，如图 3-4 所示。

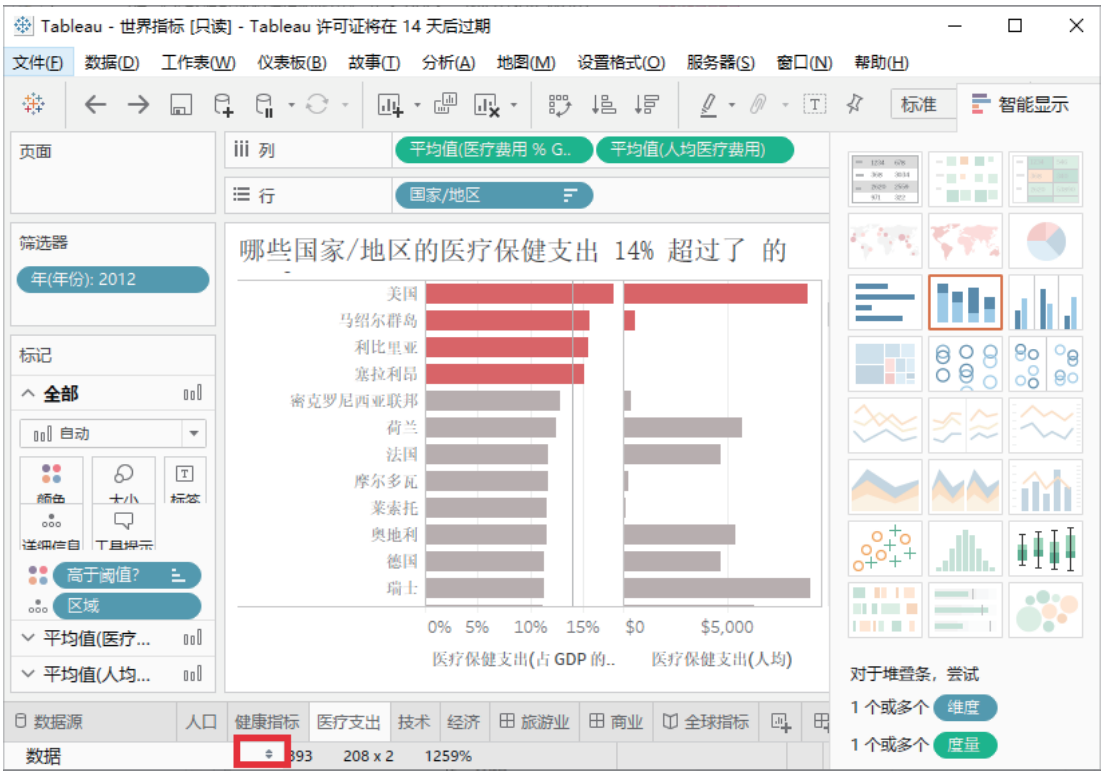


图 3-4　单击展开按钮显示侧栏

4. 状态栏信息

状态栏位于 Tableau 工作区的底部。它显示菜单项说明及当前视图的有关信息。例如，图 3-5 所示的状态栏显示该视图拥有 788 个标记，在 197 行和 4 列中显示，它还显示视图中所有标记的 SUM（度量值）为 13,448.53。可以通过单击“窗口”→“显示状态栏”来隐藏状态栏。有时，Tableau 会在状态栏的右下角显示警告图标，以指示错误或警告。

图 3-5　状态栏

5. 导入数据

打开 Tableau 软件，在界面左侧的“连接”中选择“更多…”选项，将弹出“打开”对话框，在对话框中可以选择需要导入的数据，如图 3-6 所示。

6. 导出 Tableau 数据

导出 Tableau 数据可以复制从视图中选择的数据，以便在其他应用程序中使用，也可以将工作表中的所有数据导出到 Access 数据库中。

将数据复制到剪贴板，在视图中选择要复制的数据，右击，在弹出的快捷菜单中选择“拷贝”→“数据”命令，打开其他应用程序（如 Word 或者 Excel），然后选择“编辑”→“粘贴”命令。操作页面如图 3-7 所示。

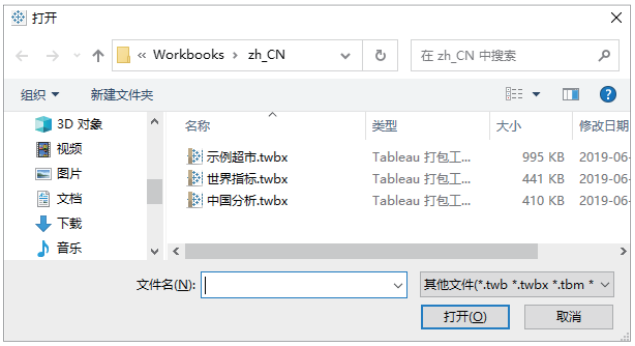


图 3-6　导入数据

注意：还可以按交叉表形式复制数据，以便在 Microsoft Excel 中使用，在快捷菜单中选择“复制”→“交叉表”命令即可。

7. 导出 Tableau 数据图像

要导出 Tableau 数据图像，可以复制视图以便在其他应用程序中使用，也可以将视图导出为图像文件。

将视图复制到剪贴板：在视图中选中要复制的图像，依次选择“工作表”→“复制”→“图像”命令，弹出“复制图像”对话框，选择要包括的内容和图例布局（如果视图包含图例），然后单击“复制”按钮，如图 3-8 所示，打开其他应用程序（如 Word 或者 Excel），然后选择“编辑”→“粘贴”命令。

注意：在某些应用程序中，可以选择“选择性粘贴”命令并指定图像的格式。

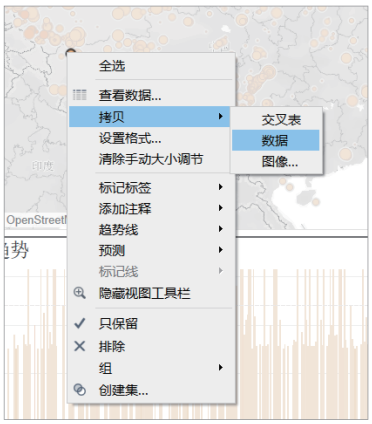


图 3-7　拷贝数据

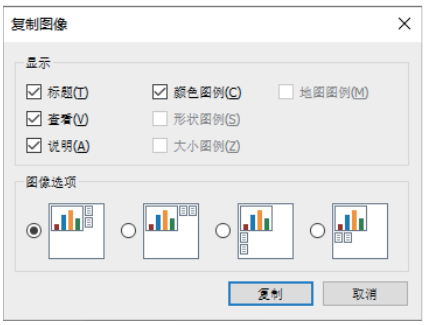


图 3-8　“复制图像”对话框

3.2　Tableau 功能介绍

本节主要对 Tableau 的功能进行介绍。

3.2.1　可视化分析概述

可视化分析主要应用于海量数据关联分析，可辅助人工操作进行数据关联分析，并作出完

整的分析图表。由于可视化分析所涉及的信息比较分散、数据结构有可能不统一，而且通常以人工分析为主，再加上分析过程的非结构性和不确定性，所以，不易形成固定的分析流程或模式，很难将数据调入应用系统中进行分析挖掘。

大脑对视觉信息的处理优于对文本的处理。在数据处理的过程中，使用可视化的方法（如图表、图形等）将数据的趋势、差别、相关性等关系展示出来，将有助于人们对数据的理解与分析，特别是一些高维度、高复杂度的数据。

3.2.2 Tableau 的可视化分析

可视化界面化繁为简，用户不仅可以通过命令行输入指令，还可以通过图形界面自由拖放完成数据的查询；而通过关系型或非关系型数据库来管理、查询和展示数据可以更加高效地完成数据管理的工作。

当能够充分发挥可视化和数据库的能力时，就会从数据工作者变成数据的思考者，可以轻松完成数据的查询，随心所欲地使用数据。目前，面临的问题有以下 3 点：

（1）低效的信息展示能力。

（2）有限的数据探索能力。

（3）痛苦的用户界面。

Tableau 软件的开发思想是分析和可视化不能分家，而且必须变成一种可视化分析的过程。可视化分析具体来说包括以下几方面。

1. 数据挖掘 / 探索

可视化分析就是要支持分析推理。可视化分析的目的就是回答数据与事实相关的问题。为了支持分析，仅仅存取和报告数据是不够的，分析需要在整个过程中有强大的运算能力的支持。通常的分析包括如下几项：①筛选，突出重要的数据；②分类排序，优先次序；③分组，聚合来归纳数据；④快速完成运算，让数字变得有意义。

2. 数据可视化

可视化分析意味着分析过程支持可视化思维，也就是要采用最优的方式对信息进行可视化，从而展示数据的变化。

简单来说，可视化分析就是用视觉的方式去发现数据、发掘数据。可视化分析主要有以下两个特点：

（1）数据的变化即刻就可以看到，例如，单击一系列数据的均值，马上就能呈现出来。

（2）观看数据变化的方式即刻就能转换，例如，用直方图显示的一组数据，马上就可以用折线图来展示。

3. Tableau 条形图

条形图（Bar Chart）是用条形的高度或长短来表示数据多少的图形。条形图可以横置，也可以纵置，纵置时称之为柱形图。

描绘条形图的要素有 3 个，分别是组数、组宽度、组限。

- 组数：把数据分成几组，指导性的经验是将数据分成 5 ～ 10 组。
- 组宽度：通常来说，每组的宽度是一致的。
- 组限：分为组下限和组上限，并且一个数据只能在一个组限内。

4. Tableau 饼图

饼图的英文学名为 Sector Graph，常用于统计学模块。2D 饼图为圆形，如图 3-9 所示。饼图可以分为简单的饼图和下钻饼图。

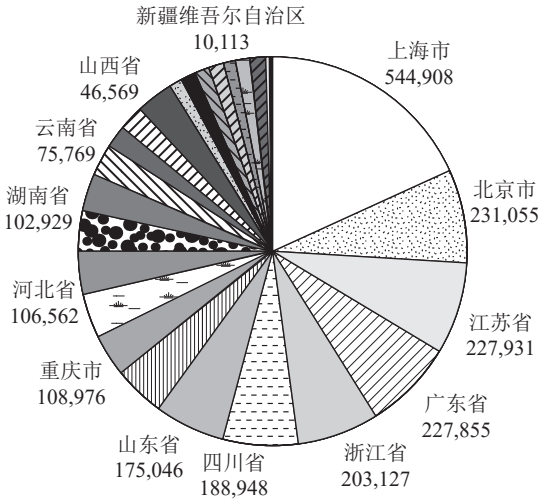


图 3-9　饼图

5. Tableau 折线图

工作表中的列或行中的数据可以绘制到折线图中。折线图可以显示随时间（根据常用比例设置）而变化的连续数据，因此，非常适用于显示在相等时间间隔下数据的变化趋势。在折线图中，类别数据沿水平轴均匀分布，所有值数据沿垂直轴均匀分布，如图 3-10 所示。

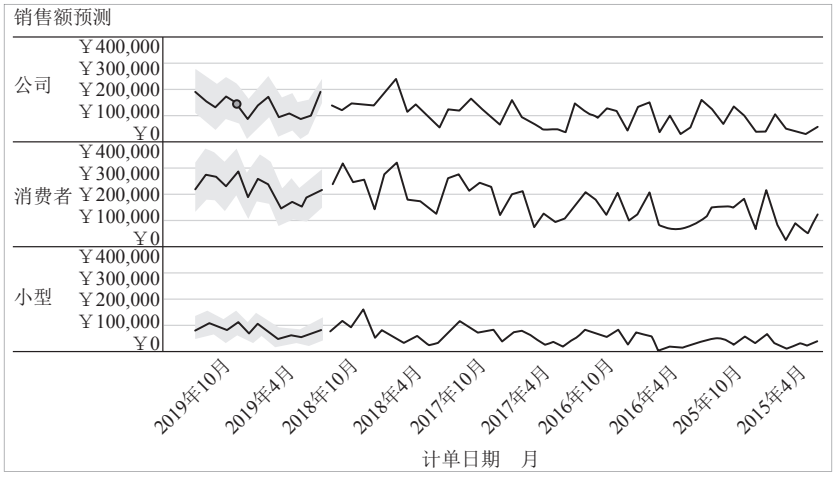


图 3-10　折线图

3.2.3　Tableau 的数据处理与计算

1. 运算符

运算符是一个符号，用于通知编译器执行特定的数学或逻辑运算。Tableau 有多个运算符用于创建计算字段和公式。

可用的运算符的类型和运算的顺序（优先级）如下：①常规运算符；②关系运算符；③逻辑运算符。

图 3-11 所示为运算符对数字、字符串及日期数据类型的示例。

2. Tableau 函数

任何数据分析都涉及大量的计算。在 Tableau 中，计算编辑器用于将计算应用于正在分析的字段。Tableau 具有许多内置函数，它们有助于创建复杂计算的表达式。

① 数字函数：用于数值计算的函数，它们只接收数字作为输入。

② 字符串函数：用于字符串操作，如图 3-12 所示。

运算符	描述	例子
+	添加两个数字 或连接两个字符串	7+3 利润+销售额 \'abc\'+\'def\' '=\'abcdef\'
-	减去两个数字 或从日期中减去天数	False

图 3-11　运算符

LEN(string)	返回字符串的长度	LEN("Tableau")=7
LTRIM(string)	返回删除了任何前导空格的字符串	LTRIM("Tableau")=7
UPPER(string)	返回字符串，所以字母都为大写	UPPER("Tableau")="TABLEAU"

图 3-12　字符串函数示例

③ 日期函数：Tableau 有各种日期函数来执行涉及日期的计算。所有日期函数都使用 date_part，它是一个字符串，表示日期的一部分，如年、季、月或日。

3. Tableau 数值计算

Tableau 中的数值计算使用公式编辑器中提供的大量内置函数来完成。

3.2.4　Tableau 的地图分析

地理数据有许多形状和格式。打开 Tableau Desktop 时，在首页左侧的“连接”窗格中显示可用的连接器，可以通过这些连接器来连接数据，也可以通过连接到空间文件来处理地理数据，还可以连接存储在电子表格、文本文件中服务器上的位置数据。

空间文件（如 Shapefile 或 geoJSON 文件）包含实际几何图形（如点、线或多边形），而文本文件或电子表格包含经纬度坐标格式的点位置，或者包含在引入 Tableau 时连接到的地理编码的指定位置。

1. 地图概念

数据存放在地图上的原因有很多，也许是数据源中有一些位置数据，也许是有人认为地图真的可以让数据很受欢迎。这两个都是创建地图可视化项的足够好的理由，但一定要记住，地图像任何其他类型的可视化项一样，有特定用途——它们可以回答空间问题。

2. 空间问题

在 Tableau 中制作地图是因为有空间问题，所以，需要使用地图来了解数据中的变化趋势或模式。

空间问题示例如下：

①哪个省 / 自治区 / 直辖市的农贸市场最多？

②中国的肥胖率高发区在哪里？

③所在城市的每条地铁线路中哪个地铁站是最繁忙的？

④人们按照当地的自行车共享计划在哪里借出和归还自行车？

3. Tableau 中的地图类型

使用 Tableau 可以创建以下通用地图类型：

①比例符号地图。

②面量图（填充地图）。

③点分布图。

④热图（密度图）。

⑤流线图（路径图）。

⑥蜘蛛图（起点至终点图）。

3.2.5　Tableau 的预测分析

1. Tableau 预测的工作原理

Tableau 中预测使用的是一种称为“指数平滑”的技术，该预测算法可以在数据中寻找能够延续到未来的模式，也可以在缺少日期的情况下，使用度量和整型维度来创建预测。Tableau 可以帮助用户分析数据，使用指数平滑技术可以有效获取数据并制定未来趋势。

所有预测算法都是实际数据生成过程的简单模型。为获得高质量预测，DGP 中的简单模式必须与模型所描述的模式很好地匹配。质量指标用于衡量模型与 DGP 的匹配程度，如果质量指标较低，则由置信区间测量的精度将比较低，因此，需要重新调整模型以获得更高的质量指标。

2. 指数平滑方法

Tableau 自动选择最多 8 个模型中最佳的一个，以生成最高质量的预测。Tableau 会优化每个模型的平滑参数。优化方法是全局的，因此，选择的本地最佳的平滑参数也可能是全局范围内最佳的。不过，初始值参数将根据最佳做法进行选择，但不会进一步优化，所以，初始值参数可能不是最佳的。Tableau 提供的 8 个模型属于 OTexts 网站上的“指数平滑方法的分类”中所介绍的那些模型的一部分。

指数平滑模型通过对一个固定的时间序列的过去值的加权平均值，以迭代方式来预测该序列的未来值。最简模型从上一个实际值和上一个级别值来计算下一个级别值或平滑值。该方法之所以是指数方法，是因为每个级别的值都受前一个实际值的影响，影响程度呈指数下降，即值越新权重越大。

3. Tableau 的闭合式方程

当可视化项中的数据不足时，Tableau 会自动尝试以更精细的时间粒度进行预测，然后将预测聚合回可视化项的粒度。Tableau 提供可以由闭合式方程进行模拟或计算的预测区间。所有具有累乘组件或具有集合预测的模型都具有模拟区间，而所有其他模型则使用闭合式方程。

4. Tableau 预测度量

当要预测的度量在进行预测的时间段内呈现出趋势或季节性时，带有趋势或季节性的组件的指数平滑模型十分有效。趋势就是数据随时间增加或减小的趋势。季节性是指值的重复和可预测的变化，如每年中各季节的温度波动。

通常，时间序列中的数据点越多，所产生的预测就越准确。如果要对季节性建模，则具有足够的数据就尤为重要，因为模型越复杂，就需要越多的数据才能达到合理的精度级别。如果使用两个或更多不同 DGP 生成的数据进行预测，得到的预测质量将较低，因为一个模型只能匹配一个。

5. 模型类型

在“预测选项”对话框中，可以选择 Tableau 用于预测的模型类型。对于大多数视图而言，“自动”设置通常是最佳设置。如果选择“自定义”，则可以单独指定趋势和季节特征，

选择“无”“累加”或“累乘”。

累加模型是对各模型组件的贡献进行求和，而累乘模型是将一些组件的贡献相乘。当趋势或季节性受数据级别的影响时，采用累乘模式可以显著提高数据的预测质量。

6. 粒度与修剪

在创建预测时，需要选择一个日期维度，指定度量日期值所采用的时间单位。Tableau 日期支持一系列时间单位，如年、季度、月和天。为日期值选择的单位称为日期的粒度。度量中的数据通常并不与粒度单位完全一致。不完整季度的值会被预测模型视为完整季度，而不完整季度的值通常比完整季度的值要小。如果允许预测模型考虑此数据，则产生的预测将不准确。解决方法是修剪该数据，从而忽略可能会误导预测的末端周期。使用“预测选项”对话框中的“忽略最后”选项来移除部分周期。默认设置是修剪一个周期。

7. Tableau 季节性

Tableau 针对一个季节周期进行测试，具有对于估计预测的时间系列而言最典型的时间长度。如果按月聚合，Tableau 将寻找 12 个月的周期；如果按季度聚合，Tableau 将寻找 4 个季度的周期；如果按天聚合，Tableau 将寻找每周季节性。因此，如果按月时间系列中有一个 6 个月的周期，Tableau 可能会寻找一个 12 个月模式，其中包含两个类似的子模式；然而，如果按月时间系列中有一个 7 个月的周期，Tableau 可能找不到任何周期。幸运的是，以 7 个月为周期的情况并不常见。

3.2.6 Tableau 的仪表板

仪表板是若干视图的集合，可以同时比较各种数据。举例来说，如果有一组需要每天都审阅的数据，那么可以创建一个一次性显示所有视图的仪表板，而不是导航到单独的工作表。

像工作表一样，可以通过工作簿底部的标签访问仪表板。工作表和仪表板中的数据是相连的，当修改工作表时，包含该工作表的任何仪表板也会更改，反之亦然。工作表和仪表板都会随着数据源中的最新可用数据一起更新。

默认情况下，Tableau 仪表板设置为使用固定大小，如果保留此设置，则务必按照将要查看的大小来构建可视化项。也可以将“大小”设置为“自动”，这会使 Tableau 根据屏幕大小自动适应可视化项的总体尺寸。这意味着，如果设计 1300×700 像素的仪表板，则 Tableau 将调整它的大小以适应小型显示器 , 有时这会使视图或滚动条挤成一团。“范围”调整功能可以帮助避免这一点。

应该将仪表板中包括的视图数限制为两个或三个。如果加的视图太多，则详细信息中可能会丢失视觉清晰度和重点。如果发现故事工作区的范围需要超过两个或三个视图，则可以创建更多的仪表板。发布仪表板后，如果视图太多，则可能会影响仪表板性能。有关性能的更多详细信息，请参阅“加快可视化项的速度”部分的内容。

3.2.7 Tableau 的分享与发布

当发布到 Tableau Public 时，这些视图可公开访问，这意味着将与能够访问 Internet 的任何人共享本地视图及本地基础数据。共享机密信息时，可考虑使用 Tableau Server 或 Tableau Online。

Tableau Desktop 能以全新的方式查看数据，在短短几分钟内完成关键业务问题的探索与回答，然后与他人分享见解。Tableau 提供了共享和协作选项，可以轻松实现这个目标，无论数

据在本地、云端还是分散在混合部署中，只需选择能够在现有数据环境中实现最佳集成效果的 Tableau 选项。

1. 发布数据源的组成部分

将数据源发布到 Tableau Online 或 Tableau Server 是维护单一数据源必不可少的一个步骤，也有助于在同事之间共享数据，包括与那些不使用 Tableau Desktop 但在 Web 编辑环境中编辑工作簿的同事共享数据。发布的数据源的更新将及时推送到所有关联的工作簿，这样有助于保持数据的一致性。

2. 选择使用用户筛选器生成工作簿

在 Tableau 服务器上，用户可以通过缩略图视图来浏览 Tableau 内容，这些缩略图图像是基于工作簿和工作表而生成的。如果工作簿包含用户筛选器，则可以指定选择哪个用户的筛选器来创建缩略图。如果希望缩略图图像显示销售额预测的所有区域，则只有被允许查看所有区域的用户生成这些缩略图。

在以下情形中，将会显示一个通用图像来替代缩略图视图：

- 被选中的用户没有查看数据的权限。
- 数据来自数据源筛选器、用户计算、模拟所得或来自其他用户。

3.3　Tableau 可视化案例

先来看一下本案例的可视化背景。3W 咖啡是一家总部位于北京市海淀区的公司，主要经营咖啡、饮料、美食、茶饮等业务，业务分布中国、美国、英国、荷兰等国家。在北京市开设有鲁谷店、门头沟店、西单店、大兴店等多个分店，消费人员多集中于商务精英、学生、退休人员。现在手中有一份 3W 咖啡的销售数据，需要对数据进行可视化处理，数据源如图 3-13 和图 3-14 所示。

日期	客户id	性别	国籍	职业	产品分类	产品名称	等级	价格	购买数量	费额()	门店名称	店长
2019/1/1	A010	女	中国	无业游民	饮料	产品13	三级	6.4	3	19.2	3W咖啡府井店	Nancy
2019/1/1	A020	女	美国	学生	咖啡	产品15	三级	48	4	192	3W咖啡西单店	Jerry
2019/1/1	A009	女	英国	学生	茶饮	产品22	三级	3.6	8	28.8	3W咖啡西单店	Jerry
2019/1/1	A012	女	美国	上班族	茶饮	产品17	三级	13.6	1	13.6	3W咖啡鲁谷店	Jerry
2019/1/1	A004	男	中国	退休	冰饮	产品10	一级	65.7	6	394.2	3W咖啡门头沟店	Simon
2019/1/1	A008	男	中国	学生	美食	产品15	二级	79.2	3	237.6	3W咖啡长城店	Jim
2019/1/1	A021	男	荷兰	无业游民	美食	产品11	一级	16.1	1	16.1	3W咖啡王府井店	Nancy
2019/1/1	A013	男	美国	无业游民	饮料	产品15	一级	30	4	120	3W咖啡大兴店	Bob
2019/1/1	A016	男	英国	无业游民	茶饮	产品21	一级	3.5	3	10.5	3W咖啡王府井店	Nancy

图 3-13　各门店目标销售额的数据源 1

门店	目标额
3W咖啡鲁谷店	670266
3W咖啡门头沟店	732623
3W咖啡公主坟店	592647
3W咖啡西单店	575295
3W咖啡望京店	737179
3W咖啡王府井店	420331
3W咖啡大兴店	571544
3W咖啡长城店	768084

图 3-14　各门店目标销售额的数据源 2

1. 折线图

现在通过展示一个月中每天销售情况的折线图来表现本月销售额的浮动趋势，横坐标代表

日期，纵坐标代表销售量，如图 3-15 所示。

Step01 设置筛选器。工作表界面的“标记”功能的上方也有一个“筛选器”，单击加入筛选器后，选择“月”格式、日期字段。

Step02 设置行列名。在筛选器右上方可以看到行、列选项，对应的就是坐标 x 轴和 y 轴，在行中设置“天（日期）”，在列中设置“总和（消费额（元））”。

Step03 选择颜色。选择“标记”下方的颜色选项，可以选择你想要在图中显示什么样的颜色，例如，此面积图中选择的就是蓝色。

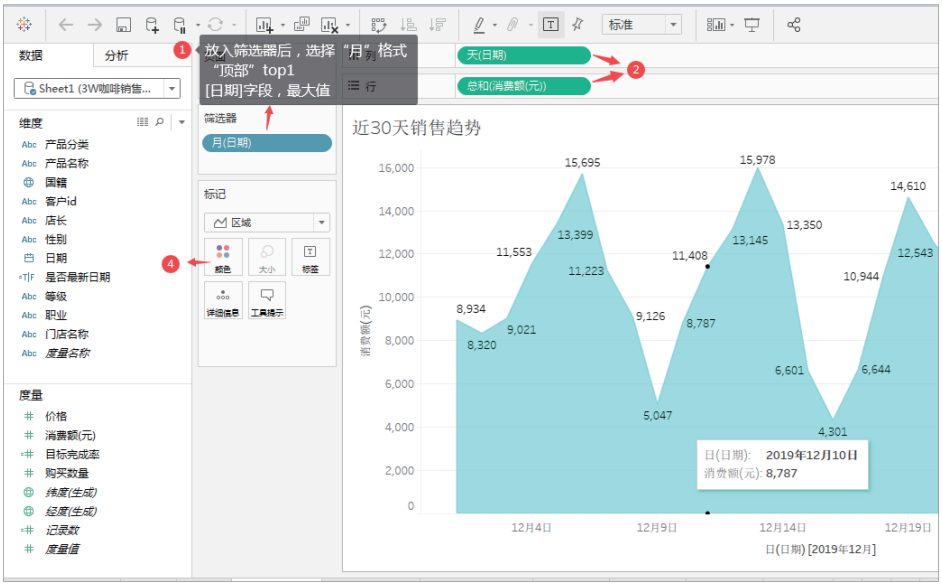


图 3-15　绘制折线图

2. 标靶图

绘制标靶图，分别展示鲁谷店、门头沟店、西单店、大兴店等各个分店在一年之内要达成指定目标的完成比例，如图 3-16 所示。

Step01 “连接数据”命令。选择“数据”→“连接数据”命令。

Step02 设置行列名称。在筛选器右上方可以看到行、列选项，对应的就是坐标 x 轴和 y 轴，在行中设置“门店名称”，在列中设置“总和（消费额（元））”。

Step03 选择标靶图。在工作表右上方可以看到“智能显示”，单击选择其中的标靶图。

Step04 检查字段。检查行、列选项中的字段名称有没有写错。

Step05 添加“完成率”字段。在工作表界面，单击左下方度量中的“完成率”字段。

Step06 选择颜色。单击颜色下方的“完成率”，对其进行颜色选择，让其标签突出完成情况。

Step07 检查设置情况。检查以上步骤是否全都设置成功。

3. 饼图

绘制饼图，展示到该咖啡店消费的男士与女士的占比情况，如图 3-17 所示。

Step01 选择维度。在工作表界面，单击左侧维度下的“性别”维度。

Step02 选择饼图。在工作表右上方可以看到“智能显示”，单击选择其中的饼图。

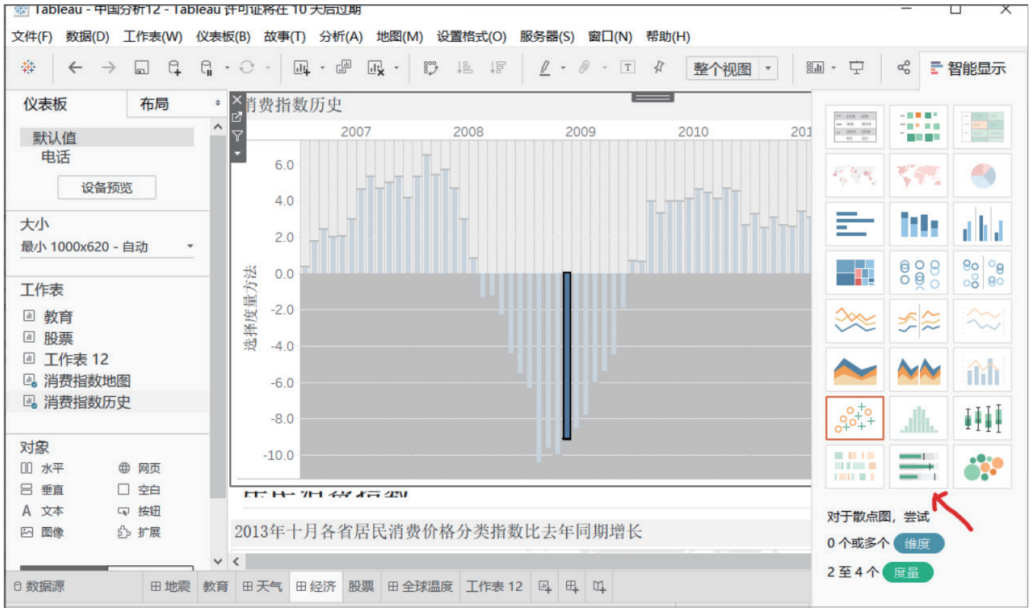


图 3-16　绘制标靶图

Step 03 选择标签。在“标记”功能下方选择“标签”选项。

Step 04 将标签放入图形中。拖动标签放到图中，就可以得到图 3-17 所示的效果。

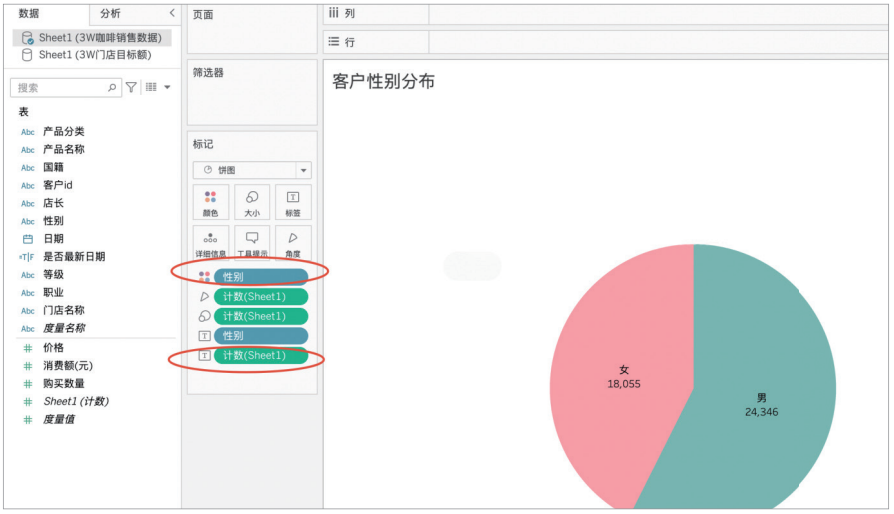


图 3-17　饼图

4. 环形图

该咖啡店的消费人群有商务精英、学生、退休人员等，各个人群的职业分布可用环形图展示，如图 3-18 所示。

Step 01 选择维度。在工作表界面上单击左侧维度下的“职业”维度。

Step 02 选择环形图。在工作表右上方可以看到“智能显示”，单击选择其中的环形图。

Step 03 选择标签。在“标记”功能下方选择“标签”选项。

Step 04 将标签放入图形中。拖动标签放到图中，就可以得到图 3-18 所示的效果。

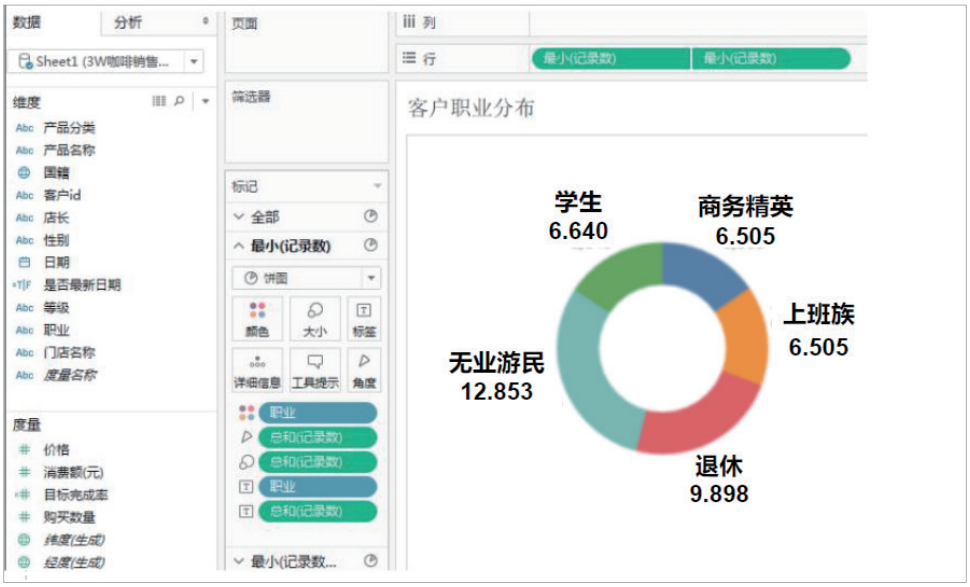


图 3-18　环形图

3.4　本章习题

一、单选题

1. Tableau 是用于（　　）数据的商业智能工具。

A. 可视分析　　B. 用户分析　　C. 商业分析　　D. 网络分析

2. 在企业、学术机构及许多政府机构，都会使用 Tableau 进行视觉数据分析。作为领先的数据（　　）工具。

A. 图像化　　B. 界面化　　C. 可视化　　D. 图表化

3. Tableau 工作区包含菜单、工具栏、“数据”窗格、卡和功能区，以及一个或多个（　　）。

A. 工作表　　B. 视图表　　C. 数据表　　D. 画图表

4. （　　）包含工作表，后者可以是工作表、仪表板或故事。

A. 内存　　B. 工作簿　　C. 外存　　D. 工作主页

5. Tableau 工作区包含菜单、工具栏、“（　　）”窗格、卡和功能区，以及一个或多个工作表。

A. 数据　　B. 内部　　C. 外部　　D. 内置

6. 通过决定尺寸和度量来选择要分析的数据。尺寸是描述性数据，而度量是（　　）数据。

A. 数字　　B. 文字　　C. 图表　　D. 单元格

7. 一个打开的 Tableau，得到开始页面显示各种（　　）。

A. 信息　　B. 图表　　C. 数据源　　D. 链接源

8. 编辑工作表时，侧栏包含“数据”窗格和“分析”窗格。根据要在视图中进行的操作，可能会看到不同的窗格（“数据”“分析”“故事”“仪表板”“布局”“（　　）”）。关于侧栏最重要的一点是可以在工作区中展开和折叠此区域。

A. 外存　　B. 内存　　C. 形式　　D. 格式

二、问答题

1. 什么是 Tableau?

2. 简单介绍一下 Tableau 的使用方法。

3. Tableau 有什么功能？

4. 什么是 Tableau 仪表板？

5. Tableau 启用突出显示有哪些方法？请简要说明。

6. 怎么制作 Tableau 制作仪表板？例如现已在 Tableau 中制作完成了条形图及饼图，如何使用仪表板展示出来？