模块一：搭建数据可视化环境

学习情境 1.1 选择数据可视化工具

学习情境描述

年底了，项目组成员基本完成了本年度的项目开发，准备休息休息。乐于好学的小王想趁此时间学习学习新技术。这天，项目经理告诉小王，明年部分项目需要在前端实现数据的可视化展示，让小王在休息的同时，预研一下数据可视化的相关技术和工具。小王听到后非常开心，因为在数字经济环境下，数据价值愈发凸显，数据资源化、数据资产化、数据要素化的趋势愈加显著。然而，伴随着数据的快速增长和集聚，数据要发挥自身的价值，需要一系列的加工处理，而数据可视化就是很重要的一项技术。小王以前在读一篇文章时，看到这样一句话，“我们沉浸在数据的海洋中，却渴望着知识的淡水”令他难忘，在他内心也播下了“挖掘数据价值”的种子。一切真是机缘巧合，终于让小王有了真正接触数据领域工作的机会。于是，小王想先了解目前有哪些数据可视化工具，然后再选择适合自身项目开发的工具和环境。

学习目标

➢ 知识目标：

(1)了解数据可视化的概念。

(2)了解数据可视化常用的工具。

(3)了解数据可视化的基本理论。

➢ 能力目标：

(1)能够利用网络工具检索数据可视化信息。

(2)能够通过官网获取权威和最新信息。

(3)具备对不同的数据可视化工具进行鉴别和评价的能力。

➢ 素质目标：

(1)培养自学的能力。

(2)培养信息获取的能力。

(3)培养英文阅读的能力。

学习任务

1.完成通过百度、知乎、豆瓣等互联网平台检索数据可视化及其工具。

2.完成通过思维导图等工具对不同的数据可视化工具进行比较分析，获取每种工具的特点、优势和应用场景。

资 讯

引导问题：

1.生活中，我们通过拍照来记录美好时刻，请问手机中存放的照片属于数据吗？

2.什么是数据可视化技术？

3.数据可视化工具有哪些？

计 划

1.制订工作方案。（见表 1-1）

表 1-1　　工作方案

步骤	工作内容
1	
2	
3	
4	
5	
6	
7	
8	

2.写出数据可视化的输出结果。

3.列出工具清单。（见表 1-2）

表 1-2　　工具清单

序号	工具名称	工具版本	备注

4.列出技术清单。（见表 1-3）

表 1-3　　技术清单

序号	技术名称	技术版本	备注

决　策

1.根据引导、构思、计划等，各自阐述自己的设计方案。

2.对其他人的设计方案提出自己不同的看法。

3.教师结合大家完成的情况进行点评，并选出最佳方案。

知识准备

1. 什么是数据

(1)数据的产生推动计算的发展

人类最早数据记录的产生，可以追溯到三万年前的旧石器时代。那时的人类祖先就开始在岩石、洞穴上，绘制描述自然生活的壁画。法国肖维岩洞壁画(图 1-1)是人类已知最早的史前艺术，创作年代约为公元前 30000 年至公元前 28000 年，壁画的内容大多为动物和捕猎的人类，贴近生活，反映了该时期的部分生活风貌。我国的贺兰山岩画(图 1-2)同样记录了远古人类 3000～10000 年前放牧、狩猎、祭祀、征战等生产生活场景，成为研究远古人类文化史、原始艺术史的文化宝库。

图 1-1　肖维岩洞壁画

图 1-2　贺兰山岩画

到了新石器时代，一些早期的社会形态逐渐形成，人们对于记录的需求日益增强，出现了各种各样的记事方法，其中使用较多的为结绳记事。在我国，《易经·系辞》是有关结绳记事的最早文献记载，其中提到“上古结绳而治，后世圣人易之以书契，百官以治，万民以查”。直到近代，一些民族依旧沿用了结绳记事。例如，我国哈尼族、瑶族、独龙族、高山族等少数民族直到20世纪50年代，依然保留这种记事方法。除了结绳记事，还有刻木记事、编贝记事、积石记事等。这些记事方法使数据信息在人类大脑以外的地方得到保存。

据考古发现，人类真正意义上的文字，是公元前3200年左右乌鲁克古城中刻有象形符号的泥板文书。最古老的文字外形并不像楔形，只是一些平板图形。随着人类文明的发展和交流范围的扩张，原始图形无法满足应用需求，于是苏美尔人逐渐简化符号，增加其意义，使得象形符号逐渐过渡为以音节表意的抽象楔形文字。事实上，汉字也起源于图画，之后从图画逐渐抽象为图案符号，再由图案符号逐渐抽象为具有意义的文字单元，这一过程持续了几千年。目前，学术公认的最早汉字，是殷商时期刻在龟甲和兽骨上的甲骨文以及铸造在青铜器上的金文，存在的时期约为公元前17世纪至公元前11世纪。

随着文字产生的另一个事物，就是数字。当文字产生后，随之产生了各式各样的数学符号。发展到后期，产生的较为成熟且一直沿用至今的，便是由印度人发明、阿拉伯人改造并传播到西方的阿拉伯数系。印度数字在公元前3世纪就已经出现，经过阿拉伯人的使用流通之后，随着阿拉伯鼎盛时期的远征传入欧洲。1202年，数学家Fibonacci发布著作《计算之书》，标志着印度数字在欧洲获得认可。后来，人们就将其称为阿拉伯数字系统，也是今日最为常见的全球通用的一种数字形式。数字的出现使得人类对事物的描述开始变得数量化、精准化，为一系列高级计算方式的诞生提供了可能，这也是为什么有人会说“数学每往前前进一小步，人类文明就往前前进一大步”。

数字的出现，也推动了计算的发展。回顾历史，算盘是人类历史上最早用来计算的专业工具。在中国，算盘大约可以追溯到汉朝时期的一种更为简单的工具——算筹。算盘由算筹在实际应用的长期过程中改进而来，并于宋元时期广泛流行。使用算盘的计算称为珠算，珠算有对应于四则运算的相应法则。在这个时期，人类就已经具有了通过工具计算数据量较大且较为复杂难解的问题的能力。后来，欧洲逐渐产生了一些机械计算器。1642年，法国数学家布莱士·帕斯卡(Blaise Pascal)创造出了第一台机械计算机，通过齿轮运作实现加法运算。1673年，德国数学家、微积分发明人之一戈特弗里德·威廉·莱布尼兹(Gottfried Wilhelm Leibniz)创造出了步进计算器(图1-3)，成为第一台可以进行加减乘除四则运算的机械计算机。

在接下来的两个世纪里，众多发明家在不同时期对步进计算器进行了连续的迭代和更新。他们不断改进这一机械计算工具，使其更加精确和高效。到了19世纪30年代，英国数学家、发明家查尔斯·巴贝奇(Charles Babbage)提出通用数字计算机的设计思想。1812年，20岁的巴贝奇从法国人杰卡德发明的提花编织机中获得启发，经过十年的钻研，于1822年制造出了第一台“差分机”，并在1823年开始建造第二台“差分机”。1832年，巴贝奇设计了一种名为“分析机”的计算设备，这是一台通用计算机，提出了近乎完整的计算机设计方案，查尔斯·巴贝奇也因此被誉为计算机之父。随后，英国数学家阿达·洛芙莱斯(Ada Lovelace)为分析机编写了一份假想程序，并预言“未来会诞生一门全新的、强大的、专为分析

所用的语言”，她也因此成为世界上第一位程序员。

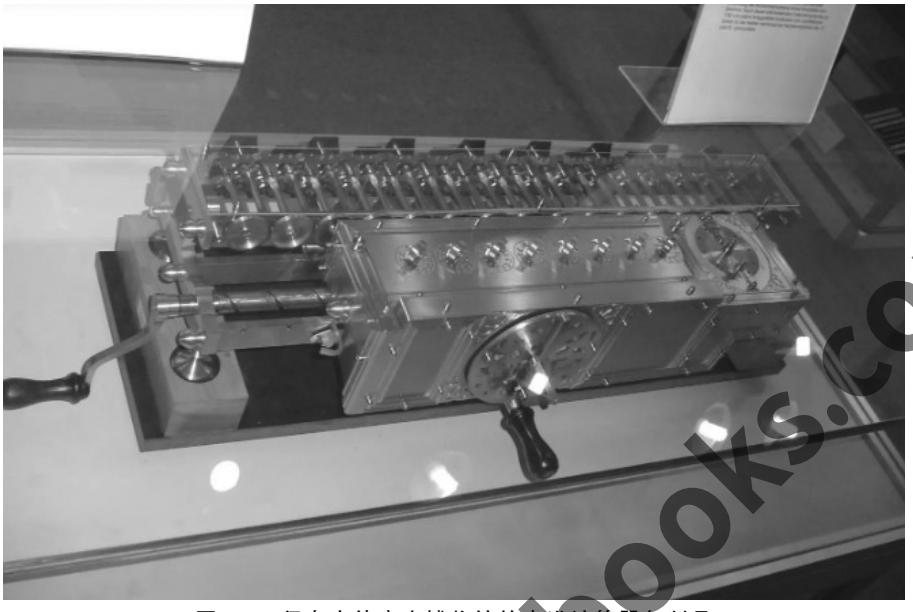

图 1-3　保存在德意志博物馆的步进计算器复制品

1848 年，英国数学家乔治·布尔(George Boole)创立了二进制代数学，为现代二进制计算机铺平了道路。到了 1854 年，布尔提出了符号逻辑的思想，进一步推动了计算机领域的发展。布尔的创新奠定了计算机科学的基础，他的理论为逻辑运算和信息处理提供了坚实的数学基础，为现代计算机的逻辑设计和运算方式奠定了基石。布尔的发现不仅推动了数学和计算机科学的发展，还对数理逻辑、电子工程和信息技术产生了深远的影响。1913 年，麻省理工学院的教授范内瓦·布什(Vannevar Bush)制造出了第一台模拟式计算机微分分析仪。1939 年，美国人约翰·阿塔那索夫(John Vincent Atanasoff)和他的学生克利夫·贝瑞(Clifford E. Berry)研制了人类第一台电子计算机。此后，计算机一次次改进和优化，为人类迎接信息时代提供了高效的工具。

20 世纪 50 年代，通信领域的学者们开始意识到，不同计算机用户之间也有通信的需求，于是他们开始对分散网络、排队论、分组交换等展开研究。1960 年，美国国防部高级研究计划局，出于冷战考虑创建了 ARPA 网络。此后，网络技术日益进步，ARPA 网络逐渐成为互联网发展的基础。1973 年，ARPA 网络被扩展为互联网，接入了来自英国和挪威的计算机。在互联网几十年的发展过程中，ARPA 的罗伯特·卡恩(Robert Elliot Kahn)和斯坦福的温顿·瑟夫(Vint Cerf)提出了 TCP/IP，蒂姆·伯纳斯-李(Tim Berners-Lee)在瑞士欧洲核子研究组织构建了万维网项目。如今，互联网已经达到了高度普及的程度，世界各地的人们都在互联网上分享、下载、上传各种类型的数据，庞大的互联网数据正成为一种全新的数据的表现形式，相应的并行计算、分布式计算、集群计算和云计算技术等的出现，为当下数据科学的研究提供了坚实的基础和重要的支撑。

(2)数据及其数据的 DIKW 模型

当下，翻开书本、打开电脑或手机，甚至不必自己去搜寻，就已经有各种各样的数据源源不断地向我们涌来。大至政府发布的各种人口数据、税务数据，小至物价数据、天气数据，以

及自身身体的健康数据等，今天的我们，已经生活在一个离不开数据的世界。数据的产生及数据的应用，将我们带进了数据科学时代。

今天，我们身处数据的漩涡，却又不能准确说出数据到底是什么，我们对数据既熟悉又陌生。当然，不同的学科对数据的定义也是不同的。统计学中的数据，是指为了找出问题背后的规律而需要的、与问题相关的变量的观测值，是对客观现象进行计量的结果；计算机科学中的数据，是指所有能输入计算机并被计算机程序处理的，具有一定意义的数字、字母、符号等的统称。但可以看出，从某个学科来讲，数据的概念被狭义化了。从广义的视角来看，数据是对客观事物的一种符号化描述，具体呈现的形式有文本、图形、图像、声音、视频等。比如，书本上的内容是一种数据，载体是纸张，符号是文本或图像等；商店里面各种商品的价格是一种数据，载体是商品价格单，符号是文本、数字和图像；刻在岩石上的壁画是一种数据，载体是石头，符号是图像；监控拍摄的录像是一种数据，载体是磁盘等数字化设备，符号是视频。因此，凡是记录客观世界里各种事物的符号表示都是数据。数据是对真实世界的简化描述，只能无限逼近，但永远无法完全反映真实世界。

数据中蕴含着有价值的信息，通过对有价值信息的提取，进而可以帮助人们做出正确的决策，以便解决现实遇到的问题和困难。在此过程中，衍生出了数据的 DIKW 模型，D 代表数据(Data)，I 代表信息(Information)、K 代表知识(Knowledge)、W 代表智慧(Wisdom)，如图 1-4 所示。

DIKW 模型呈现金字塔形状。在该模型中，数据(Data)位于最底层，是对客观事物的数量、属性、位置、状态及其相互关系等的具体表示，属于“原始材料”；信息(Information)在数据层之上，是数据的语义化解释，具有一定的时效性，是已经过加工处理，并对决策有指导作用的数据流；知识(Knowledge)在信息层之上，是经过人类长期选择与积累的、具有价值的信息，相比信息更具抽象性、逻辑性和价值性；智慧(Wisdom)位于模型的最顶层，是人类所具备的、基于已有知识和相关信息对问题进行分析和解决的能力，这种能力运用的结果是将有价值的信息挖掘出来，并使之成为已有知识结构的一部分，进而促进智慧的产生。

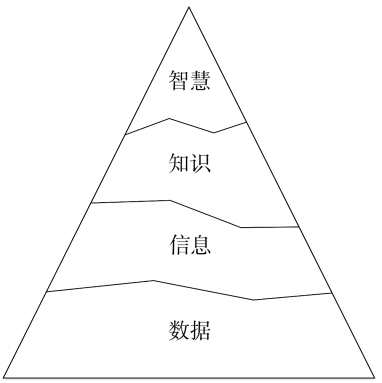

图 1-4　数据的 DIKW 模型

2. 数据可视化技术

(1)理解数据可视化

在数据可视化领域，爱德华·塔夫特(Edward Tufte)被誉为“数据界的列奥纳多·达·芬奇(Leonardo da Vinci)”。他的一大贡献就是聚焦于将每一个数据做成图示物。塔夫特的信息图形不仅能传达信息，甚至被很多人看作是艺术品。塔夫特指出，可视化不仅能作为商业工具发挥作用，还能以一种视觉上引人入胜的方式传达数据信息。

信息对接收者的价值受到很多因素的影响。除了数据本身的完整性、准确性和时效性之外，数据的展现方式也将影响信息传递的准确性和接收者的认知，如人们常说的“一图胜千言”，即图形能够以更直观的形式呈现出复杂的数据所要表达的信息，方便接收者对数据

进行进一步的分析与应用。特别是面对以高速(Velocity)、大量(Volume)和多样化(Variety)3V为典型特征的大数据,迫切需要利用数据可视化,帮助决策者理解复杂数据背后隐藏的信息。

数据可视化就是使用图形化手段表达数据的变化、联系或者趋势的方法,将数据转换为图形图像显示出来,其目的是清晰有效地传达与沟通信息,让用户更好地理解和使用数据。按过程来讲,数据可视化主要是记录信息、分析推理并进行信息传播与协同的过程。简而言之,数据可视化意味着以可视化图表的形式来显示数据信息,实现发现、分析、预测、监控和决策的目的。

数据可视化与信息图形、信息可视化、科学可视化以及统计图形密切相关。当前,在研究、教学和开发领域,数据可视化乃是一个极为活跃而又关键的方面。“数据可视化”这个术语实现了成熟的科学可视化领域与较年轻的信息可视化领域的统一。

(2)数据可视化的作用

数据可视化是借助图形化手段,清晰有效地传达与沟通信息,帮助企业从数据中提取信息,从信息中获取价值。具体而言,数据可视化的作用包括以下几点:

①快速理解信息

根据美国宾夕法尼亚大学医学院的研究人员估计,通常情况下,人类视网膜“视觉输入(信息)的速度可以和以太网的传输速度相媲美”。人类视网膜中大约包含1 000 000个神经元细胞,算上所有的细胞,人类视网膜能以大约10兆字节每秒的速度传达信息。丹麦的著名科学作家陶·诺瑞钱德证明了人们通过视觉接收的信息比其他任何一种感官都多。如果人们通过视觉接收信息的速度和计算机网络相当,那么通过触觉接收信息的速度就只有它的10%。人们的嗅觉和听觉接收信息的速度更慢,大约是触觉接收速度的10%。同样,人们通过味蕾接收信息的速度也很慢。换言之,人们通过视觉接收信息的速度比其他感官接收信息的速度快了10~100倍。因此,可视化能传达庞大的信息量也就容易理解了。如果包含大量数据的信息被压缩成了充满知识的图片,那么接收这些信息的速度会更快。

数据可视化正是利用人类天生技能来增强数据处理和组织效率。通过使用图形能够以清晰、一致的方式查看大量数据,快速理解这些数据隐含的信息。比如,查尔斯·米纳德(Charles Joseph Minard)绘制的拿破仑行军图(Napoleon March Map),通过一张图呈现出了丰富的信息,如图1-5所示。

这幅图通过两个维度(2维图形)展现了6种数据类型,分别是拿破仑军队的数量、行进的路程、维度、经纬度、行进方向和特定日期或事件的位置。图的横轴是一个时间轴(刻度并不均匀),可以对应每个日期,下面的部分标明了返程时每一天的温度。图的主干部分用今天的话来说是带状图,用来表示每个时刻、每个位置的军队人数,其中淡黄色的带状区域表示向莫斯科行进的军队,黑色的带状区域表示返回巴黎的军队,带状的宽度表示了当时军人的数量。带状区域向横轴的投影对应了时刻,同时根据当时所在位置的经纬度,可以计算其指向莫斯科的方向(角度),所以带状区域的方向就是当时的真实行军方向。这样,对于带状区域上的任一点,可以对应一个地点和时间,于是可以对应一个关键的事件(比如经过某个地点)。如果行进的过程中有分兵,那么也能通过分支的区域画出来。

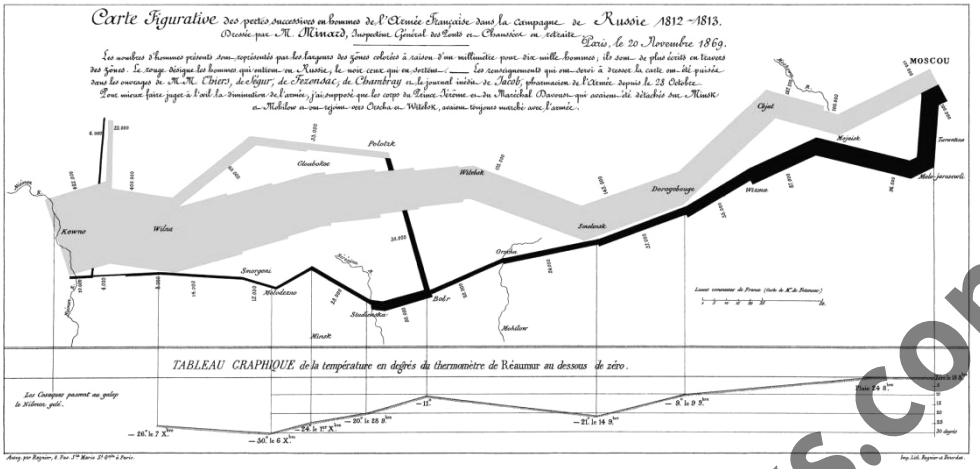

图 1-5　拿破仑行军图

这幅图的历史背景如下：1812 年的拿破仑在欧洲势如破竹，整个欧洲基本都在他的掌控之下，却唯独无法突破英国的防线。为了拿下英国，他计划让其他欧洲国家停止与英国进行贸易。俄国沙皇亚历山大眼看法国在拿破仑的带领下实力剧增，不免担心，便拒绝暂停对英贸易，此举彻底激怒了拿破仑，也直接引发了拿破仑率 60 万士兵于当年 6 月开始进攻俄国。两国军队实力相差悬殊，当时俄国的军队人数仅有 20 余万，于是他们便一路退回俄国境内，并烧毁沿路的一切物资，让法国军队越走越远但又得不到补给。到了 10 月，军粮严重不足且伤亡惨重的法国军队被一路引到了莫斯科，拿破仑终于意识到他必须开始返回了，因为俄国寒冷的冬天已经来临。但在返回途中，大量饥寒交迫的法国士兵死亡，最终仅剩不到 6 万人撤回到华沙。

战争虽然结束了，但它反复被人们研究和讨论。乘兴而来败兴而归，可以精准地概括拿破仑在这场战争中的状态。近 50 年后，查尔斯·米纳德在 80 岁高龄创新地运用由他发明的展现人员流动的图表再现了俄法战争。这张图恰恰直观地描绘出了当年的“乘兴”和“败兴”。该图有效的统计信息展现外加历史人文背景，加之信息量丰富，达到“一图胜千言”，被信息设计的先驱者、视觉化大师爱德华·塔夫特(Edward Tufte)誉为是迄今为止最好的统计图表。

同样，安斯库姆(F. J. Anscombe)四重奏的例子很好地说明数据可视化对快速理解信息的作用。

1973 年，统计学家 F. J. Anscombe 构造出了四个数据集，它们的平均数、方差、相关系数和线性方程式完全一致，这些数据集具有几乎相同的简单描述性统计数据。单从这些统计数字来看，四组数据所反映的实际情况非常相近，然而事实上，这四组数据有着天壤之别。通过绘制这四组数据的散点图(如图 1-6 所示)，可以看出点的分布完全不同。

从图中可以看出，第一张是线性关系图，数据是大多数人看到上述统计数字的第一反应，是最“正常”的一组数据；第二张是曲线关系图，数据所反映的事实上是一个精确的二次函数关系，只是在错误地应用线性模型后，各项统计数字与第一组数据恰好都相同；第三张是异常值图，数据描述的是一个精确的线性关系，只是这里面有一个异常值，它导致上述各个统计数字尤其是相关度值的偏差；第四张为极端异常值图，展现了一个更极端的例子，其

异常值导致了均值、方差、相关度、线性回归线等所有统计数字全部的偏差。

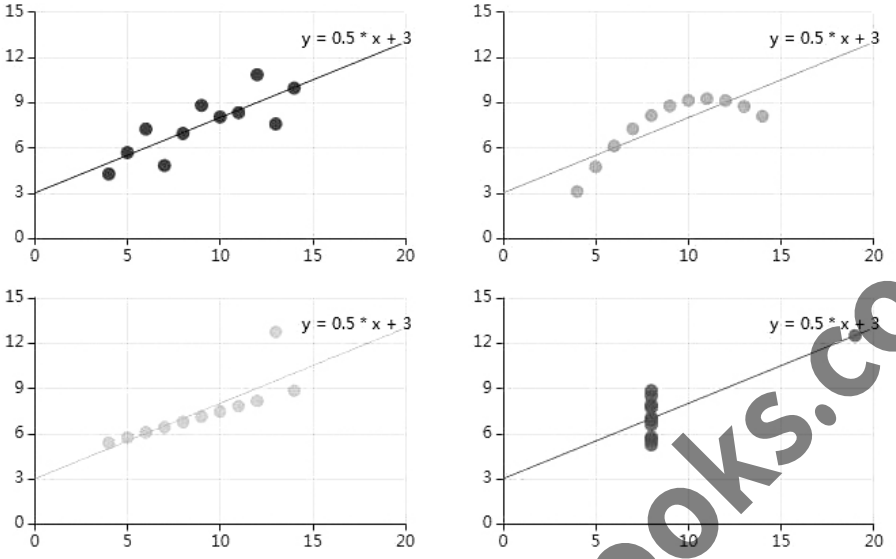

图 1-6　Anscombe 四组数据散点图

②挖掘有价值信息

数据之间以及数据描述的对象之间，往往存在隐含的、难以直接观测的信息，数据可视化以图形方式可以揭示出这些隐含的、难以观测的信息，从而发现问题的根源，帮助人们做出科学的决策。比如，1854 年英国布洛德(Broad)大街大规模暴发霍乱，当时了解微生物理论的人很少，人们不清楚霍乱的传播途径，而“瘴气传播理论”是当时的主导理论。内科医生约翰·斯诺(John Snow)对这种理论表示了怀疑，于 1855 年发表了关于霍乱传播理论的论文，图 1-7 所示即其主要依据。图的正中心沿东西方向的街道即为布洛德大街，黑色条形表示死亡的地点，死亡人数越多条形越长。这幅图形揭示了一个重要现象，死亡发生地都在街道中部一处水井(水泵)周围，市内其他水源周围极少发现死者。经过进一步调查，他发现这些死者都饮用过这里的井水，结合其他证据得出饮用水传播的结论，于是移掉了该水泵的把手，霍乱最终得到控制。

这幅图被信息设计的先驱者、视觉化大师爱德华·塔夫特(Edward Tufte)选为经典案例，用来说明数据可视化的重要性。在这个例子里，斯诺医生亲自上门统计死亡病例，并记录具体的住址，每发生一起死亡病例，他就在该地址上画一条黑线，多条黑线堆叠起来就形成了条形，一个地址条形越长说明死亡人数越多。这样所有的数据都能够画在地图上，并且一目了然地发现死亡病例的分布规律，然后推断可能的原因，最终结合他作为医生的专业背景提出假设，然后通过移除水泵把手的方式来尝试解决问题，果然控制住了疫情的传播。这是一个经典的科学分析的案例，也是数据可视化的成功典范。因为它揭示了问题的根本原因，并启发了解决方案。此外，在点图和热图还没有完全开创的时代，这种早期的尝试是令人难以置信的创新。这个解决方案之所以被发现，是因为分析师突破了可视化技术的界限，创造了一些有用的新东西。

在人们通常的印象中，数据可视化是为了漂亮和炫酷，所以时常会走入一种追求新技术的极端。但从这个近 200 年前的例子可以发现，可视化的本质是数据分析，用一种直观的方

式来发现数据中隐藏的规律才是数据可视化的真正意义。

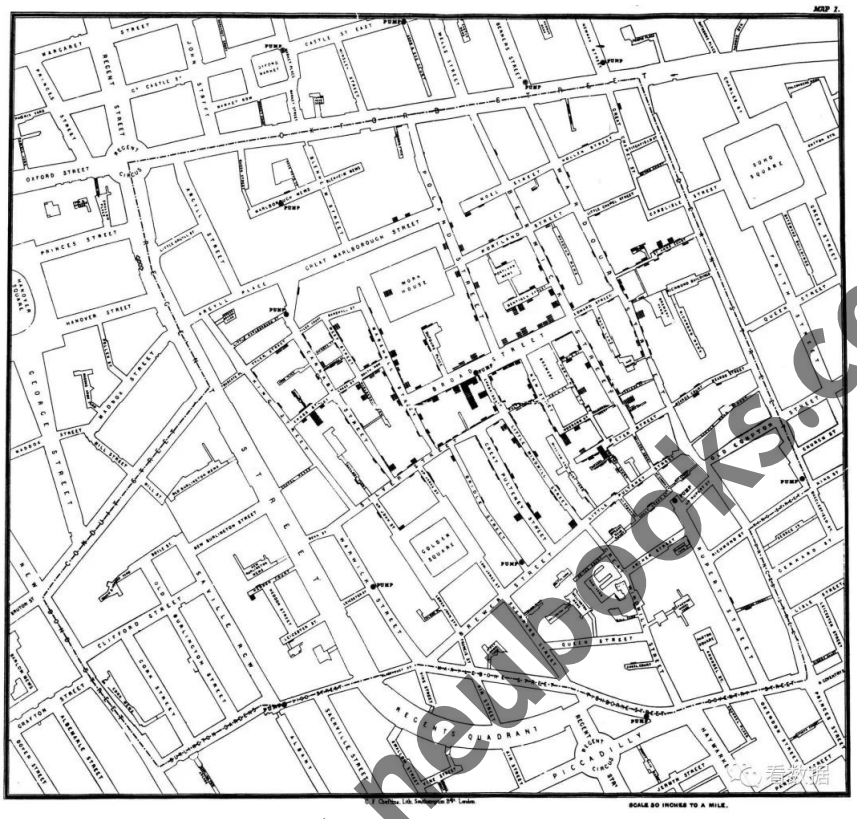

图 1-7　霍乱地图

③证实假设或猜想

除了帮助快速理解信息、识别关系和模式之外，数据可视化还能够帮助论证假设或者猜想。常常通过 A/B 测试的模式来验证假设，对于分析问题过程中的推论进行验证假设，从而发现根本原因。南丁格尔发明的玫瑰图（Nightingale's rose diagram）就是利用数据可视化证实假设的经典案例。

长得像饼图又不是饼图，这种有着极坐标的统计图有着一个美丽的名字——南丁格尔玫瑰图，又名鸡冠花图（Coxcomb Chart）或极坐标区域图（Polar area diagram），该图的发明者为南丁格尔（Florence Nightingale）女士，她是一位护士，也是一名统计学家，更是英国皇家统计学会的第一位女性会员。南丁格尔玫瑰图将柱图转化为更美观的饼图形式，是极坐标化的柱图，其夸大了数据之间差异的视觉效果，适合展示数据原本差异小的数据。

19 世纪 50 年代，英国、法国、土耳其和俄国进行了克里米亚战争。南丁格尔主动申请担任战地护士。当时的医院卫生条件极差，受伤战士死亡率高达 42%，直到 1855 年卫生委员会来到医院改善整体的卫生环境后，受伤战士的死亡率才戏剧性地降至 2.5%。当时的南丁格尔注意到这件事，认为政府应该改善战地医院的条件来拯救更多年轻的生命。

出于对资料统计的结果会不受人重视的忧虑，她发展出一种色彩缤纷的图表形式，让数据能够更加让人印象深刻，如图 1-8 所示。这张图表用以表达军医院季节性的死亡率，从整体上来看，这张图是用来说明、比较战地医院伤患因各种原因死亡的人数，每块扇形代表着

各个月份中的死亡人数，面积越大代表死亡人数越多。

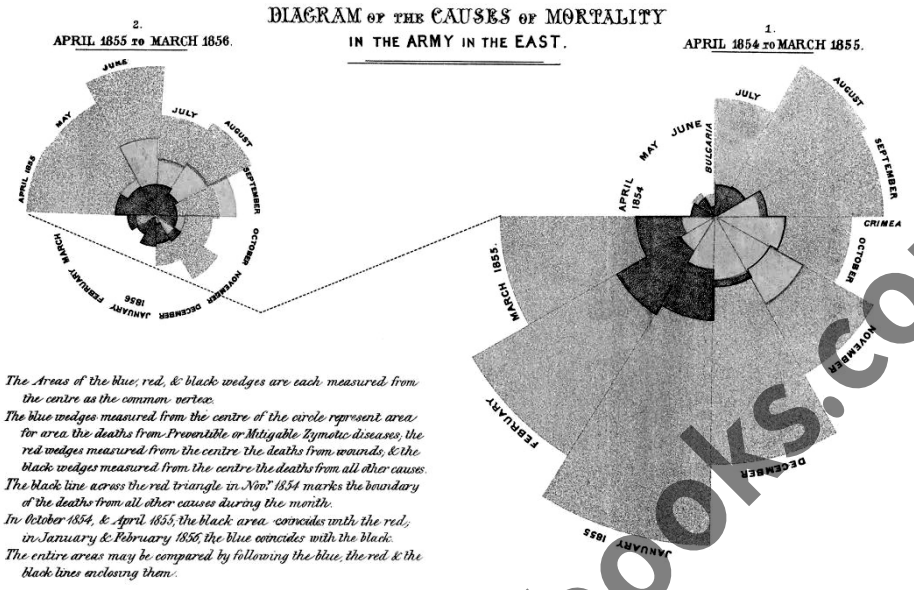

图 1-8　南丁格尔玫瑰图

对应此图，各色块圆饼区均由圆心往外的面积来表现数字；浅灰色区域代表死于原本可避免的感染的士兵数；深灰色区域代表因受伤过重而死亡的士兵数；黑色区域代表死于其他原因的士兵数；1854 年 10 月、1855 年 4 月的深灰色区域和黑色区域恰好面积相等；1856 年 1 月与 2 月的浅灰色区域和黑色区域恰好面积相等；1854 年 11 月深灰色区域中的黑线指出该月的黑色区域大小。

由图可知，左、右两个玫瑰图被时间点“1955 年 3 月”所隔开。左、右两个玫瑰图都包含了 12 个月的数据。其中，右侧较大的玫瑰图展现的是 1854 年 4 月至 1855 年 3 月的数据；而左侧的玫瑰图展现的则是 1855 年 4 月至 1856 年 3 月的数据。通过对两幅图面积大小的对比，可以轻易地得出结论：

第一，蓝色的区域的面积明显大于其他颜色，这意味着大多数的伤亡并非直接来自战争，而是来自糟糕医疗环境下的感染。

第二，卫生委员到达后(1955 年 3 月)，死亡人数明显的下降。

这幅图让政府相关官员了解到改善医院的医疗状况可以显著地降低英军的死亡率。南丁格尔的方法打动了当时的高层，包括军方人士和维多利亚女王本人，于是医事改良的提案才得到支持，甚至挽救了千万人的生命。这种新型的图表也由此得名，因为外形很像一朵绽放的玫瑰，它也被称为“南丁格尔玫瑰图”。

3. 数据可视化工具

数据可视化工具分为可视化软件工具和可视化编程语言。可视化软件工具是在第三方软件的基础上，通过所提供的工具及模板对导入的数据进行可视化。此类工具对用户的技术要求相对较低，但是往往需要花费一定成本购买软件的使用权。可视化编程语言是通过编程的方式实现可视化，该方式所使用的编程语言大都是免费的，但是需要用户具备一定的程序语言基础，技术门槛相对较高。

(1)可视化软件工具

可视化软件工具主要有 Excel、Tableau、Sugar BI 等。

①Excel

Excel 是 Microsoft 公司开发的一款电子表格软件，直观的界面、出色的计算功能和图表工具，再加上成功的市场营销，使 Excel 成为最流行的个人计算机数据处理软件。作为一个入门级工具，Excel 拥有强大的函数库，是快速分析数据的理想工具。Excel 的图形化功能虽然可以满足大部分基础应用场合，但是算不上功能强大，并且在制作可视化图表时，图表中的颜色、线条和样式可选择的范围有限，这也意味着用 Excel 很难制作出符合专业出版物和网站需要的数据图。在 Excel 2010 以上的版本中可以加载 Power Pivot 等一系列程序。这些程序为 Excel 添加了更多的数据模型和新功能，如动态图表、数据透视图等，从而使开发者制作更好的可视化图表。

Excel 提供的可视化图表部分类型如图 1-9 所示。用户可以根据数据的类型和数据可视化的目的，选择合适的图表类型。

图 1-9　Excel 提供的可视化图表类型

柱形图和条形图是从属性视角可视化数量数据的常用图表，用长条的长短表示数量变量取值的大小，而不同的长条表示不同的属性，包括簇状柱形图、堆积柱形图、百分比堆积柱形图、三维簇状柱形图、三维堆积柱形图等；折线图可以清晰地展示数据随相应因素的走势，包括基础折线图、堆积折线图、百分比堆积折线图、带数据标记的折线图等；饼图是用来表示各要素分别占整体的百分比的图表，包括基础饼图、三维饼图、复合饼图、复合条饼图、圆环图等；散点图反映的是两个数量数据之间的相关关系或整体的分布情况，包括基础散点图、带平滑线的散点图、带平滑线和数据标记的散点图、带直线的散点图、带直线和数据标记的散点图等；面积图是指在折线图内部填充颜色的图表，包括基础面积图、堆积面积图和百分

比堆积面积图等。

②Tableau

Tableau 与大多数商务智能工具一样，通过可视化方式进行数据分析。不同于传统的 BI 软件，Tableau 是一款“轻”BI 工具。可以在 Tableau 的拖放界面中可视化任何数据，探索不同的视图，甚至可以轻松地将多个数据库组合在一起。Tableau 不需要任何复杂的脚本，旨在轻松创建和分发交互式数据仪表板，通过简单而有效的视觉效果来提供对动态、变化趋势和数据密度分布的深入描述。

Tableau 提供了多种具有鲜明特征的可视化功能，实现数据发现和深入洞察的智能方式。丰富的可视化类型库包括“文字云”和“气泡图”，可为 Tableau 提供独特的高级别理解。树图和树形图为视觉效果提供上下文信息，后者通常用于描述分类数据，重点关注最相关的信息。Tableau 是面向企业级的可视化工具，分为 Desktop 版和 Server 版。Desktop 版又分为个人版和专业版，个人版只能连接到本地数据源，专业版还可以连接到服务器上的数据库；Server 版主要是用来处理仪表盘，上传仪表盘数据，进行共享，各个用户通过访问同一个 Server 就可以查看到其他同事处理的数据信息。除此之外，Tableau 还可以与 Amazon Web Services（AWS）、MySQL、Hadoop、Teradata 以及 SAP 等平台或系统协作，使之成为一个能够创建详细图形和展示直观数据的多功能工具，从而辅助企业中的各级管理人员做出决策。

③SPSS

SPSS 是世界上应用最广泛的专业统计软件之一，其全称为 Statistical Product and Service Solutions，意为“统计产品与服务解决方案”。

SPSS 包括了各种成熟的统计方法和模型，为用户提供了全方位的统计方法，如方差分析、回归分析、多元统计分析方法、生存分析方法等，方法体系覆盖全面。在数据准备方法方面，SPSS 提供了各种数据准备与数据整理技术。在结果报告方面，SPSS 提供了自由灵活的表格功能，使得制表变得更加简单、直接。同时，SPSS 可绘制各种常用的统计图形，如条图、线图、饼图、直方图、散点图等多种图形，以对数据进行全面直观的展示。

SPSS 之所以有广大的用户群，不仅因为它是一种权威的统计学工具，也因为它是一款非常简单易学的软件。人机界面友好、操作简单（图 1-10），使得统计分析人员对它“情有独钟”，事实上，不断地增强其易用性几乎是近十几年来 SPSS 的核心改进方向。另外，SPSS 也向高级用户提供了编程功能，使分析工作变得更加节省时间和精力。

④Sugar BI

Sugar BI 是基于百度云推出的敏捷 BI（商业智能）和数据可视化在线平台，目标是解决报表和大屏的数据 BI 分析与可视化问题。Sugar BI 提供界面优美、体验良好的交互设计，通过拖拽图表组件可快速搭建数据可视化页面，并对数据进行快速的分析。

Sugar BI 支持连接多种数据源，包括 Excel/CSV、MySQL、SQL Server、PostgreSQL、Oracle、GreenPlum、Kylin、Hive、Spark SQL、Impala、Presto、Vertica 等，还可以通过 API、静态 JSON 方式绑定可视化图表的数据，简单灵活。大屏与报表的图表数据源可以复用，便于用户为同一套数据搭建不同的展示形式。

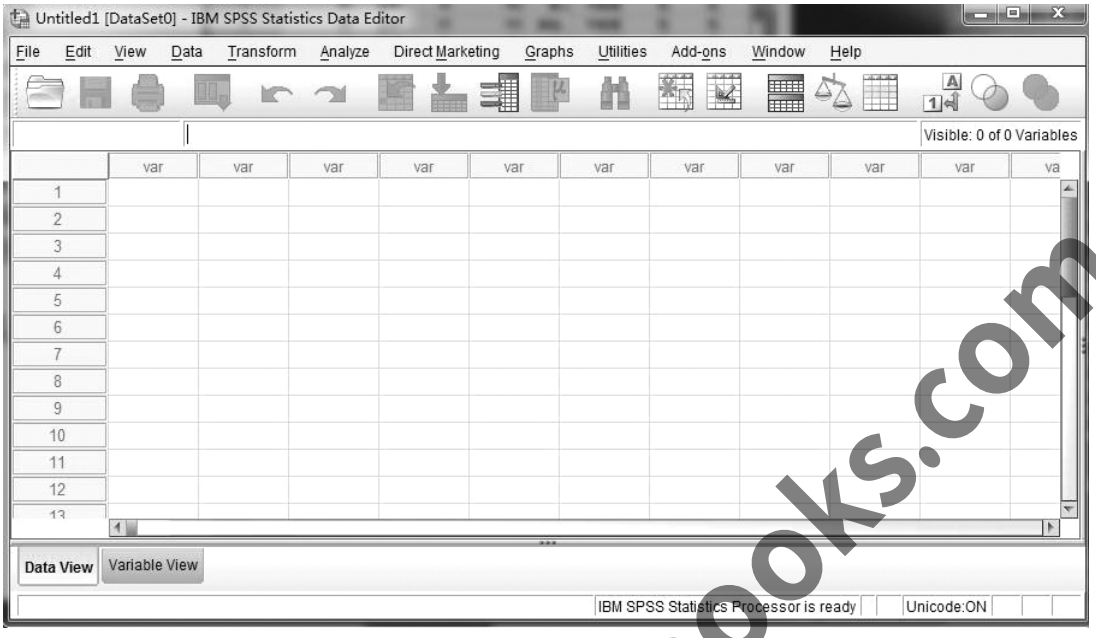

图 1-10　SPSS 工作界面

(2)可视化编程语言

①ECharts

ECharts 是一个使用 JavaScript 实现的开源可视化库，能兼容市面上的大部分浏览器，如 IE、FireFox、Chrome、Safari 等浏览器，能流畅地运行在 PC 和移动端的设备上。ECharts 的底层依赖图形库 ZRender，提供直观、生动、可交互、可个性化定制的数据可视化图表。ECharts 最初由百度团队开源，并于 2018 年初捐赠给 Apache 软件基金会(ASF)，成为 ASF 孵化级项目。2021 年 1 月 26 日，ECharts 成为 Apache 顶级项目。

ECharts 提供个性化了很多常规的图形，如折线图、柱状图、散点图、饼图、K 线图等，其中盒形图用于统计，地图、热力图、线图用于地理数据的可视化，关系图、树形图、旭日图用于关系数据的可视化，平行图用于多维数据的可视化，还有漏斗图用于 BI、仪表盘的可视化，并且支持图与图之间的混搭。

ECharts 支持 HTML5 中的 Canvas 技术、SVG(4.0+)、VML 的形式渲染图表。VML 也可以兼容低版本 IE，SVG 使移动端不再为内存担忧，Canvas 可以轻松应对大数据量和特效的展现。不同的渲染方式为用户提供了更多选择，使得 ECharts 在各种场景下都有很好的表现。

ECharts 能够展现十万级的数据，也提供了对流加载(4.0+)的支持，可以使用 WebSocket 或者对数据分块后加载，加载多少渲染多少，能大大节省时间，无须等待所有数据加载完才进行绘制。此外，ECharts 针对客户端交互做了细致的优化，例如，在 PC 端可以用鼠标在图中缩放(用鼠标滚轮)、平移等。在移动端小屏上则可以用手指在屏幕的坐标系中缩放、平移。

②Python

Python 由 Guido van Rossum 于 1989 年底开始设计与开发，于 1991 年发行第一个公开

发行版本。Python是一门免费、开源的跨平台高级动态编程语言，支持命令式编程、函数式编程，完全支持面向对象程序设计，拥有大量功能强大的内置对象、标准库和扩展库。经过多年的发展，Python已经渗透到计算机科学与技术、统计分析、移动终端开发、科学计算可视化、逆向工程与软件分析、图形编程与图像处理、人工智能、游戏设计与策划、网站开发、数据采集、大数据处理、密码学、系统运维、音乐编程、计算机辅助教育、医药辅助设计、天文信息处理、化学、生物等几乎所有专业和领域。著名搜索引擎Google的核心代码使用Python实现，迪士尼公司的动画制作与生成采用Python实现，大部分UNIX和Linux都内建了Python环境支持，豆瓣网使用Python作为主体开发语言进行网站架构和有关应用的设计与开发，网易大量网络游戏的服务器端代码超过70%采用Python进行设计与开发，易度的PaaA企业应用云端开发平台和百度云计算平台BAE也都大量采用了Python语言，美国宇航局使用Python实现了CAD/CAE/PDM库及模型管理系统，雅虎公司使用Python建立全球范围的站点群，微软公司的集成开发环境Visual Studio 2015开始默认支持Python语言，开源ERP系统Odoo完全采用Python语言开发，引力波数据是用Python进行处理和分析的，类似应用不胜枚举。

基于Python的数据可视化是通过其扩展库来实现的。一般来讲，Python可视化是基于NumPy库、Pandas库、Matplotlib库、Seaborn库、Pyecharts库等第三方库来实现的。

- NumPy库

NumPy库是Python做数据处理的底层库，是高性能科学计算和数据分析的基础，掌握NumPy的数据处理功能是利用Python做数据运算和机器学习的基础。NumPy最核心的部分是N维数组对象，即Ndarray对象，它具有矢量算术能力和复杂的广播能力，可以执行一些科学计算。Ndarray对象同时拥有对高维数组的处理能力，这是数值计算不可或缺的重要特性。NumPy库对Ndarray的基本操作包括数组的创建、索引和切片、运算、转置和轴对称等。NumPy数组中可以保存任何类型的数据，如整数、字符串、浮点数等。

- Pandas库

Pandas库是Python下著名的数据分析库，主要功能是进行大量的数据处理。Series和DataFrame是Pandas库的两类主要的数据结构。其中，Series是一维的，DataFrame是二维的。基于Pandas库，可以完成数据读取、数据整理和数据可视化等主要步骤。在多数情况下，数据可视化的数据来源于外部数据，如CSV文件、Excel文件、Json文件和数据库文件等。

- Matplotlib库

Matplotlib库是Python中最流行的数据可视化库，功能十分强大，绘图风格类似MATLAB。Matplotlib通过pyplot模块提供了一套绘图API，将众多绘图对象构成的复杂结构隐藏在这套API内部，用户只需要调用pyplot模块提供的方法，以渐进的方式快速绘图，并设置图表的各种细节，如创建画布、在画布中创建一个绘图区、在绘图区上画几条线、给图像添加文字说明等，而且可以输出为PNG或PDF等多种文件格式。

- Seaborn库

Seaborn库是在Matplotlib库基础上的高级API封装，它提供了一个高级界面来绘制有吸引力的统计图形，可以使数据可视化更加方便、美观。

● Pyecharts 库

Pyecharts 是基于 ECharts 的类库，用于生成 ECharts 图表的库，是将 Python 与 ECharts 相结合的数据可视化工具。使用 Pyecharts 库可以制作多种不同的图表，基于 Web 浏览器进行显示。

③R

新西兰人罗斯·伊哈卡(Ross Ihaka)和罗伯特·杰特曼(Robert Gentleman)创造了 R，该语言因两位作者名字的首位字母而得名。R 语言专注于统计模型和数据分析与可视化，编程容易、资源丰富、入门门槛低。它的一个设计理念是“人类的时间永远比机器的时间更宝贵”，在数据科学领域，R 语言虽然经常因为运算性能稍弱被诟病，但它一直是数据科学领域编程最容易的语言，可以用最少的代码来解决复杂的分析问题。

R 语言是贝尔实验室开发的 S 语言的一种实现，著名的 C 语言、UNIX 系统也是贝尔实验室开发的。R 是属于 GNU 系统的一个自由、免费、源代码开发的软件，在多个商业领域有着广泛的应用。

R 语言具有完善的数据类型，如向量、矩阵、因子、数据集、一般对象等，支持缺失值，代码像伪代码一样简洁、可读。强调交互式数据分析，支持复杂算法描述，图形功能强。实现了经典的、现代的统计方法，如参数和非参数假设检验、线性回归、广义线性回归、非线性回归、可加模型、树回归、混合模型、方差分析、判别、聚类、时间序列分析等。

R 程序安装后，启动界面如图 1-11 所示。

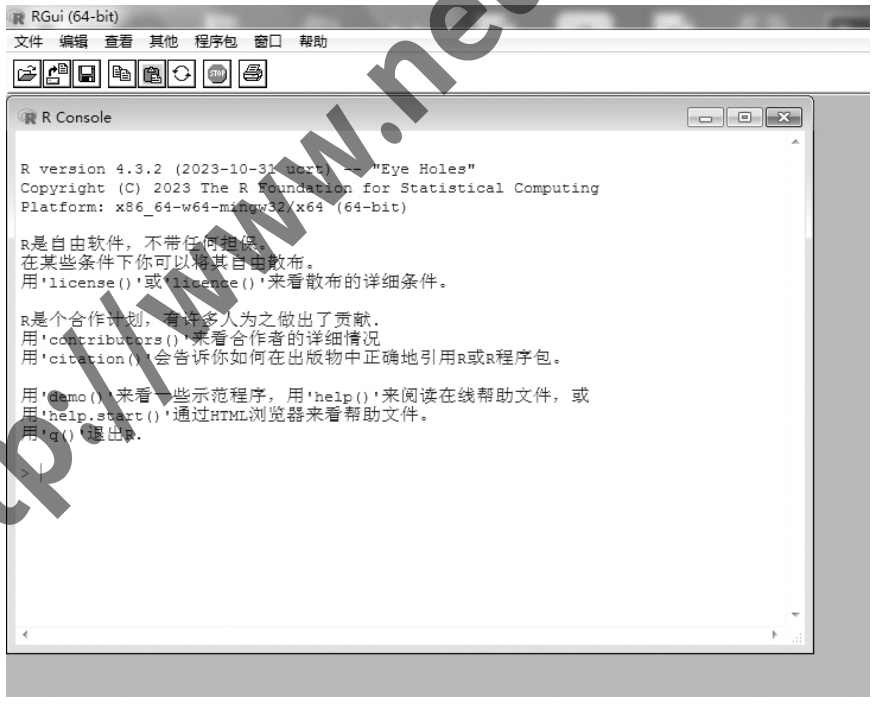

图 1-11　R 工作界面

目前，R 语言已超越仅仅是流行的强有力开源编程语言的意义，成为统计计算和图表呈现的软件环境，并且还处在不断发展的过程中。R 语言的核心开发团队完善了其核心产品，

这将推动其进入一个令人激动的全新方向。无数的统计分析和数据挖掘人员利用R语言开发统计软件并实现数据分析。

Google首席经济学家哈尔·范里安(Hal Ronald Varian),有一句经典名言形容R语言"The Great beauty of R is that you can modify it to do all sorts of things. And you have a lot of prepackaged stuff that's already available, so you're standing on the shoulders of giants."大意是,R语言之美在于,你可以通过修改很多高手已经写好的程序包,解决各种各样的问题。因此,当你使用R语言时,你已经站在巨人肩膀上了。

成立国家数据局,助推数字经济发展

2023年10月25日,国家数据局正式揭牌。国家数据局的设立,折射出数据在今天的经济体系中已经变得非常重要。以数据为核心的数字经济,是如今世界经济发展的核心领域。我国提出要大力发展数字经济,牢牢抓住数字技术发展主动权,把握新一轮科技革命和产业变革发展先机。2019年,《中共中央关于坚持和完善中国特色社会主义制度、推进国家治理体系和治理能力现代化若干重大问题的决定》首次将"数据"列为生产要素。2020年,《中共中央、国务院关于构建更加完善的要素市场化配置体制机制的意见》首次将数据作为一种新型生产要素写入中央文件,与土地、劳动力、资本、技术等传统要素并列。2022年1月,国务院印发了《"十四五"数字经济发展规划》,从顶层设计上明确了数字经济及其重点领域发展的总体思路、发展目标、重点任务和重大举措。2022年12月,中共中央、国务院印发《关于构建数据基础制度更好发挥数据要素作用的意见》(以下简称"数据二十条"),通过"一条主线、四项制度、四项措施",充分实现数据要素价值、促进数字经济发展。

数字经济是继农业经济、工业经济之后的新经济形态,从某种程度上来说,数字经济的外在表现之一是平台经济,其二是算力经济。数据驱动的平台化模式引领各行各业衍生出更多的新业态和新模式,成为推动价值创造和价值聚集的重要载体。我国是一个数据大国,具有数字经济的发展优势。但不可否认的是,数据要素市场的发展仍然面临着较大难题,数据要素基础制度尚未明确。除此之外,数据也可能涉及国家主权和国家安全利益,因此各国立法都对数据跨境行为设定了数据安全审查流程。由此可见,数据是数字经济时代的复杂客体,关涉多个方面的利益,因此在理论层面数据面临着确权难题,进而制约了我国数据要素市场的发展。

面对数字经济时代的现实挑战,党中央、国务院高屋建瓴、审时度势,《党和国家机构改革方案》明确提出组建国家数据局,负责协调推进数据基础制度建设、数据资源整合共享和开发利用等职能。这是一次具有重要意义的机构改革,优化了数据管理体制,将国家数据局作为数据发展的宏观统筹核心,确保了数字经济发展中的分工科学、职责明确、目标明确,有利于形成有效的目标约束机制。

任务实施

按照本学习情境所涉及的知识点，本任务将展示“选择数据分析工具”的具体实施过程。在学习情境描述中要求“需要在前端实现数据的可视化展示”，而且前端呈现的数据是动态的，所以可视化软件工具显然是不合适的。而在前端开发的三大技术中，HTML 主要用于搭建网页结构，CSS 用于美化页面，JavaScript 用于实现交互。因此基于编程语言的数据可视化工具需要无缝对接 JavaScript。目前，基于 JavaScript 的数据可视化工具也比较丰富，具有代表性的如图 1-12 所示。

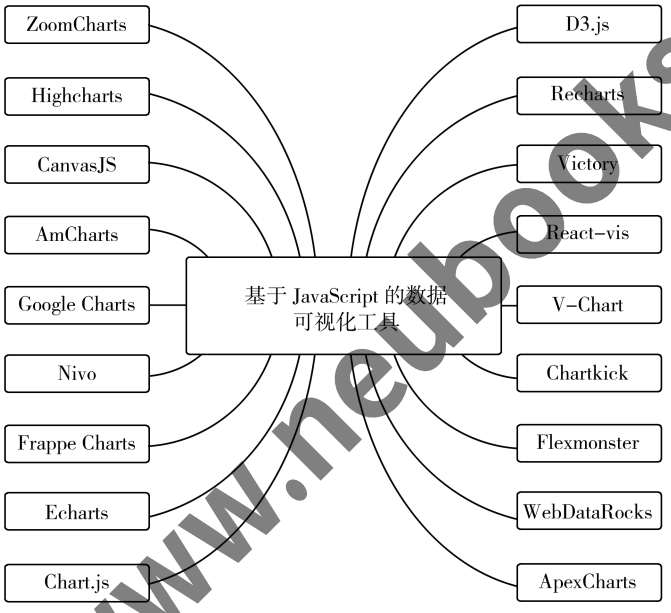

图 1-12　基于 JavaScript 的数据可视化工具

现对各工具的主要特性总结如下：

D3.js：D3 不仅用于在 Web 浏览器中生成动态、交互式数据可视化，还用于动画、数据分析、地理和数据实用程序。使用可缩放矢量图形（SVG）、HTML5 和级联样式表（CSS）标准。D3 将数据绑定到 DOM 及其元素，能够通过更改数据来操作可视化效果。

Recharts：Recharts 是一个使用 React 且流行的数据可视化库，在后台使用 D3 并公开声明性组件，Recharts 非常轻巧，可以渲染 SVG 元素以创建漂亮的交互式图表，可以制作各种不同类型的图表，同时仍然允许根据需要进行高度的自定义。

Victory：Victory 是一个用于构建交互式数据可视化的 ReactJS 库。它依靠 D3 进行强大的底层数学运算，依靠 Radium 进行内联样式管理，依靠 React 进行其他大多数操作。Victory 专门为 React 和 React Native 设计的模块化图表组件，以实现轻松的跨平台图表绘制。

React-vis：React-vis 是 Uber 公司开源的数据可视化库，基于 React 和 D3，React-vis 库

的安装和使用非常简单，拥有 Uber 支持的优秀文档，可以快速创建流行的图，如折线图、面积图、条形图、饼图、树状图等。React-vis 很轻，动画简单且流畅，还允许基于现有元素编写自定义图表。

V-Chart：V-Chart 基于 Vue2. x 和 echarts，用于创建具有简单数据配置的通用图表。

Chartkick：Chartkick 是一个图表绘制工具，UI 美观、使用简单。

Flexmonster：Flexmonster 用于可视化业务数据的强大 JavaScript 工具，可以在 Web 应用程序上可视化复杂的业务信息。支持 Microsoft Analysis Services OLAP 多维数据集、Mondrian、icCube、Salesforce、SAP、SQL（MS SQL、MySQL 和许多其他）静态或数据库 CSV 文件。

WebDataRocks：WebDataRocks 是一个用 JavaScript 编写的可嵌入的 Web 数据透视表，可以在 Web 应用程序中使用，并根据数据构建交互式报表。

ApexCharts：ApexCharts 是一个带有 Vue. js 和 React 包装器的 SVG 图表库，支持缩放、平移、滚动数据，支持在图表上放置信息性注释等。

Chart. js：Chart. js 是一个免费的 JavaScript 库，用于制作基于 HTML 的图表，是最简单的 JavaScript 可视化库之一，并且附带许多内置图表类型，允许使用 HTML5 Canvas 元素构建响应式图表。

Echarts：Echarts 是一个由百度开源的数据可视化库，凭借着良好的交互性、精巧的图表设计，得到了众多开发者的认可。

Frappe Charts：Frappe Charts 是一个开源 SVG 图表组件，受类似 GitHub 的视觉效果启发的软件包，支持折线图、条形图和其他类型的图表。

Nivo：Nivo 是一个基于 D3 和 React 构建的框架，提供了多种不同类型的组件来呈现数据。Nivo 提供了许多自定义选项和三个渲染选项：Canvas、SVG 和基于 API 的 HTML。

Google Charts：Google Charts 是一个纯粹的基于 JavaScript 的图表库，旨在通过添加交互式图表功能来增强 Web 应用程序。Google Charts 提供了各种各样的图表，例如：折线图、样条图、面积图、条形图、饼图等。

AmCharts：AmCharts 是一组基于 JavaScript 的数据可视化库，包括常规图表类型，如 Serial、Pie 等，以及股票图表和地图等高级版本。Amcharts 可以从简单的 CSV 或 XML 文件提取数据，也可以从动态数据读取生成，比如 PHP、. NET、Ruby on Rails 和 Perl，以及其他许多编程语言。

CanvasJS：CanvasJS 是一个易于使用的 HTML5 和 JavaScript 图表库，可以在 Canvas 元素上构建。CanvasJS 是一个核心图表创建器库，使用户能够创建丰富的 UI 仪表板和图表，这些仪表板和图表可以在所有设备上工作，而不会影响 Web 应用程序的功能或维护。用户可以使用 CanvasJS 创建响应式、动态、可渲染、轻量级和丰富的 UI 仪表板。

Highcharts：Highcharts 是一个用纯 JavaScript 编写的一个图表库，能够很简单便捷地在 Web 网站或是 Web 应用程序添加有交互性的图表。

ZoomCharts：ZoomCharts 是一个 JavaScript/HTML 图表库，只需最少量的代码就可向应用程序添加视觉丰富的交互式图表，可以将 ZoomCharts 与任何服务器端编程语言（包括 .NET、PHP、Java、Ruby 等）和任何客户端框架（包括 AngularJS、jQuery 等）一起使用。

综上，从应用生态的成熟性、支持图表的丰富性、用户交互的友好性，以及图表展现的美观性、开发使用的开源性和免费性等多个角度考虑，最终选择国内百度公司的 ECharts 作为数据可视化工具。

各自完成学习情境的任务并展示结果，介绍任务的完成情况。作品展示前应准备阐述材料，并完成自查和互查，自查单如表 1-4 所示。

表 1-4　　自查单

学习情境 1.1	选择数据可视化工具		
检查项目	检查标准	分值	得分
信息检索	能够使用网络工具准确检索数据可视化相关信息	20	
归纳整理	能够对不同的数据可视化工具进行分类	20	
比较分析	能够综合分析不同工具的优缺点	20	
工作结果	能够根据实际需求，选择最优的数据可视化工具	40	
合计		100	

学生展示过程中，以个人为单位，对以上学习情境的结果进行互查。互查单如表 1-5 所示。

表 1-5　　互查单

学习情境 1.1		选择数据可视化工具							
评价项目	分值	等级				评价对象			
		优（100%）	良（80%）	中（60%）	差（40%）	1	2	3	4
计划合理	10								
方案准确	10								
工作质量	20								
工作效率	20								
工作完整	10								
工作规范	10								
成果展示	20								
合计	100								

评 价

教师对学生的工作过程和工作结果进行综合评价，如表 1-6 所示。

表 1-6　　综合评价表

学习情境 1.1		选择数据可视化工具		
评价项目		评价标准	分值	得分
考勤(20%)		无迟到、早退、旷课现象	20	
工作过程(40%)	方案制作	能有效制订整个工作方案	10	
	工具运用	能有效使用相关工具	10	
	工作态度	态度端正、工作认真、积极向上	10	
	职业素养	一丝不苟、精益求精	10	
工作结果(40%)	工作完整	能按时完成任务	10	
	工作质量	能按计划完成任务	10	
	材料准备	能准备成果展示所需要的材料	10	
	成果展示	能准确表达、汇报工作成果	10	

拓展思考

1. 数据可视化的一般过程是什么？
2. 数据可视化与数据分析之间的关系是什么？
3. 数据可视化的图形要素有哪些？

学习情境 1.2　了解数据可视化基本图形

学习情境描述

小王结合项目实际最终选择好了可视化工具之后，接下来他要思考的是如何将数据有效地通过图表的方式呈现出来。通过查询资料后发现，现实中存在的图形类型实在是太多了，有柱状图、散点图、雷达图、马赛克图等，这让他眼花缭乱，疑惑重重，到底是根据数据决定图形呢，还是通过图形选择数据呢？并且不同图形有其自身的特色和优势，在项目开发中，到底应怎样确定并使用合适的可视化图形呢？

学习目标

➢ 知识目标：

(1)了解数据可视化中会涉及哪些图形。

(2)了解每种可视化图形的基本特征。

(3)了解基本的数据类型。

➢ 能力目标:

(1)能够根据实际选择相应的可视化图形。

(2)能够把握数据类型与可视化图形之间的关系。

(3)具备对不同的可视化图形依据数据类型进行归纳总结的能力。

➢ 素质目标:

(1)培养学生欣赏图形美的能力。

(2)培养多学科交叉的复合型能力。

(3)培养学生综合分析的能力。

学习任务

1.完成对数据可视化基本图形的理解。

2.完成通过思维导图或表格等工具建立数据类型和数据可视化基本图形之间的关系。

资 讯

引导问题:

1.小明想把自己 8 次的考试成绩以图形的方式展现出来,到底是选择使用折线图还是雷达图?

2.数据可视化的基本图形有哪些?

3.数据类型与数据可视化图形之间存在什么关系?

计 划

1.制订工作方案。(见表 1-7)

表 1-7 工作方案

步骤	工作内容
1	
2	
3	
4	
5	
6	
7	
8	

2.分析数据与数据可视化图形之间的关系。

3.列出工具清单。(见表 1-8)

表 1-8　　工具清单

序号	工具名称	工具版本	备注

4.列出技术清单。(见表 1-9)

表 1-9　　技术清单

序号	技术名称	技术版本	备注

决　策

1.根据引导、构思、计划等,各自阐述自己的设计方案。

2.对其他人的设计方案提出自己不同的看法。

3.教师结合大家完成的情况进行点评,并选出最佳方案。

知识准备

1. 直方图

直方图(histogram)与核密度图(Kernel density plot)是观察数据分布特征的常用图形，它们可以直观地展示数据分布的形状是否对称、偏斜的方向和程度等。

将数据分组后，在 x 轴上用矩形的宽度表示每个组的组距，在 y 轴上用矩形的高度表示每个组的频数或密度，多个矩形并列在一起就是直方图。

直方图的主要作用有：

(1)能够显示各组频数或数量分布的情况；

(2)易于显示各组之间频数或数量的差别。通过统计直方图还可以观察和估计哪些数据比较集中，异常或者孤立的数据分布在何处。

以黄石国家公园喷泉数据 geyser(Venables and Ripley, 2002)为例，制作了喷泉喷发间隔时间的分布情况，如图 1-13 所示。

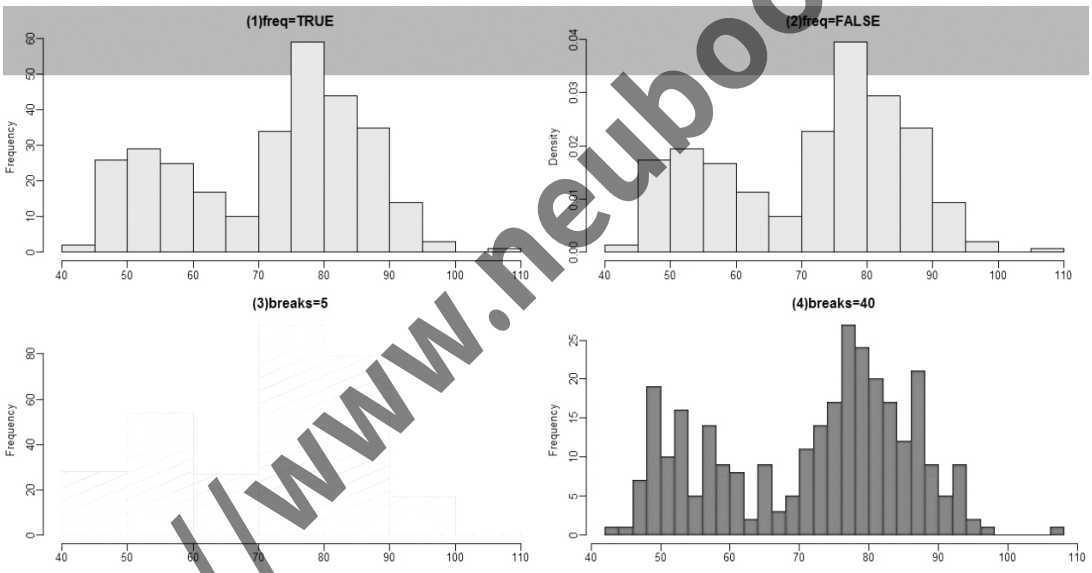

图 1-13　喷泉喷发间隔时间直方图

图 1-13 中，(1)和(2)中的直方图看起来形状完全一样，区别仅仅是前者为频数图，后者为密度图。二者在统计量上仅相差一个常数倍，但密度直方图的一个便利之处在于它可以方便地添加密度曲线，用以辅助展示数据的统计分布；(3)和(4)的区别在于区间划分段数，我们可以很清楚地看出区间划分的多少对直方图的直接影响。这里需要特别指出的是，直方图的理论并非想象中或看起来的那么简单，窗宽也并非可以任意选择，不同的窗宽或区间划分方法会导致不同的估计误差。

2. 折线图

折线图(line chart)是描述时间序列最基本的图形，它主要用于观察和分析时间序列随时间变化的形态和模式。折线图的 x 轴是时间，y 轴是变量的观测值。

图 1-14 展示了 2018 年某地的 AQI、PM2.5、PM10 和臭氧浓度等指标的折线图。

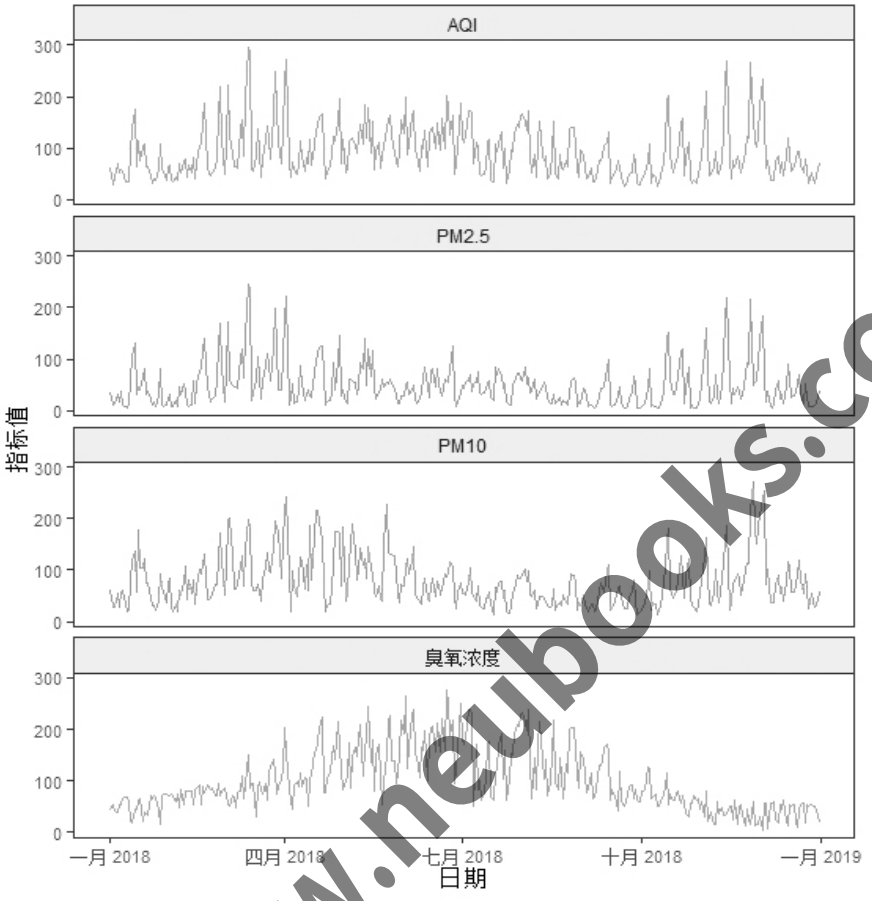

图 1-14　空气污染指标折线图

从图可以看出，AQI、PM2.5、PM10 三个指标的全年变化特征十分相似，均有一个共同特点，就是有两个峰值，即在 3 月到 4 月有一个波峰，11 月左右也有一个波峰，而 7 到 10 月处于波谷。这说明气温较低时，AQI、PM2.5、PM10 的值都较高，空气质量也较差；而气温较高时，空气质量较好。臭氧浓度的变化刚好相反，在 7 月左右有一个波峰，表明气温较高时，臭氧浓度相对较高，空气质量相对较好；而气温较低时，臭氧浓度也较低，空气质量相对较差。除臭氧浓度外，其他三个指标都没有趋势性特征或固定的模式，基本上为随机波动。

3. 散点图

散点图（scatter graph、point graph、X-Y plot、scatter chart 或 scattergram）是分析变量间关系的常用工具，如果涉及两个变量时，那么可以绘制普通散点图；如果涉及两个以上变量时，那么可以绘制散点图矩阵或相关系数矩阵。散点图可以提供三类关键信息：

（1）变量之间是否存在数量关联趋势；

（2）如果存在关联趋势，那么其是线性还是非线性的；

（3）观察是否有存在离群值，从而分析这些离群值对建模分析的影响。

通过观察散点图上数据点的分布情况，我们可以推断出变量间的相关性。如果变量之间不存在相互关系，那么在散点图上就会表现为随机分布的离散的点；如果存在某种相关

性，那么大部分的数据点就会相对密集并以某种趋势呈现。数据的相关关系主要分为正相关（两个变量值同时增加）、负相关（一个变量值增加另一个变量值下降）、无相关、线性相关、指数相关等。那些离集群较远的点我们称为离群点或者异常点（outlier）。

如图 1-15 所示为鸢尾花数据 iris 所做的散点图矩阵散点图矩阵。

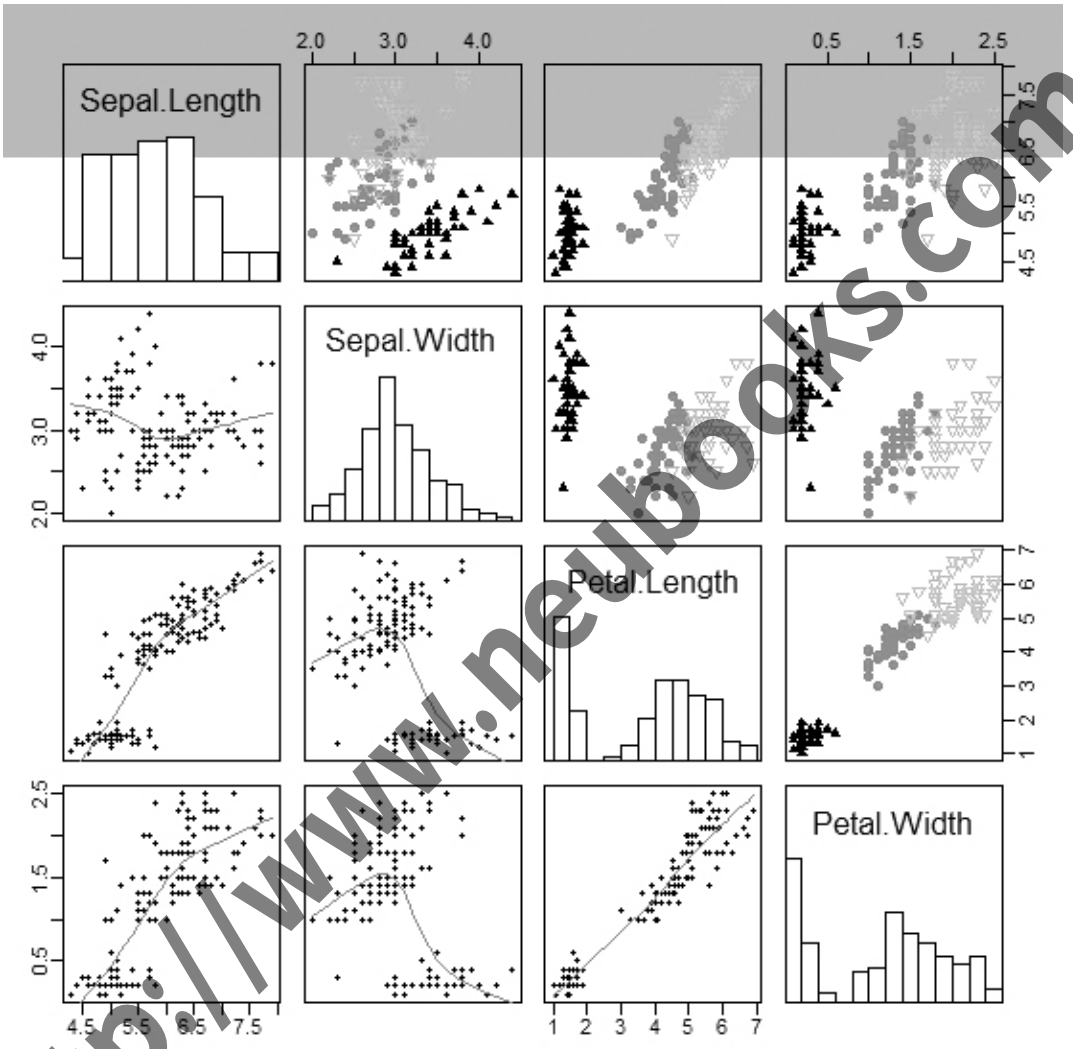

图 1-15　鸢尾花数据散点图矩阵

4. 柱形图

柱形图是以高度或长度的差异来显示统计指标数值的一种图形。柱形图简明、醒目，是一种常用的统计图形。

柱形图一般用于显示一段时间内的数值变化或显示各项之间的比较情况。另外，柱形的高度反映了数值的大小。柱形越“矮”，数值越小；柱形越“高”，数值越大。需要注意的是，柱形的宽度与相邻柱形的间距决定了整个柱形图视觉上的美观程度。若柱形的宽度小于间距，则会使人们的注意力集中在空白处而忽略了数据，所以合理地选择宽度很重要。

如图 1-16 所示为不同品牌的汽车的每加仑燃料所行英里数。

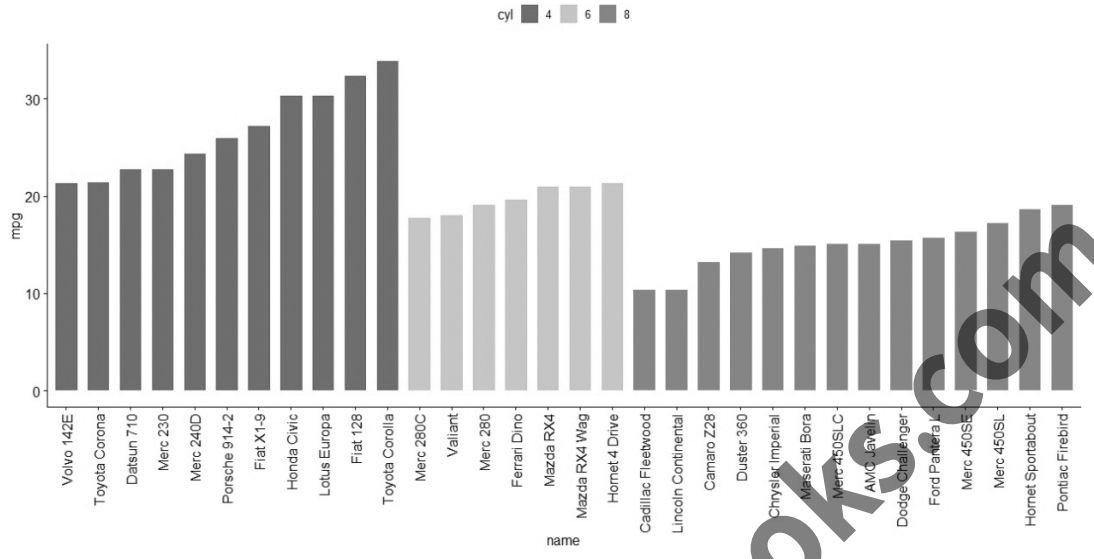

图 1-16　不同品牌汽车的 mpg 柱形图

5. 箱线图

箱线图是展示数据分布的另外一种图形，它不仅可以反映一组数据分布的特征，如分布是否对称、是否存在离群点等，还可以比较多组数据的分布特征，这也是箱线图的主要用途。如图 1-17 所示为 6 项空气污染指标的箱线图。

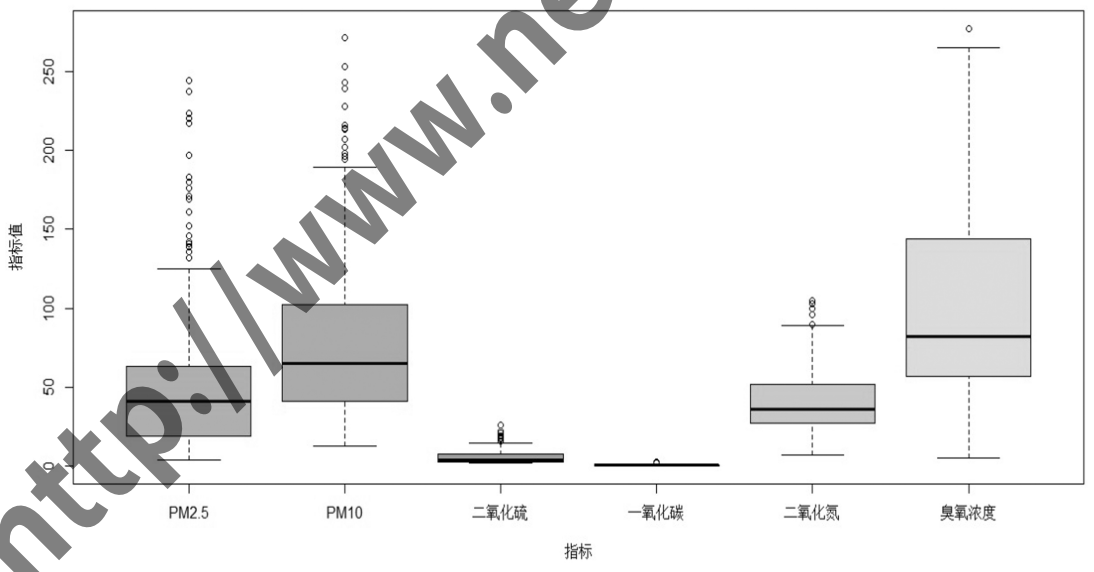

图 1-17　空气污染指标箱线图

6. 饼图

饼图(pie chart) 被广泛地应用于各个领域，用于表示不同分类的占比情况，通过弧度大小来对比各种分类。饼图通过将一个圆饼按照分类的占比划分成多个切片，整个圆饼代表数据的总量，每个切片(圆弧)表示该分类占总体的比例，所有切片(圆弧)的加和等于 100%。如图 1-18 所示为运动偏好的统计。