第 1 章 Python 数据分析概述

当今社会，网络和信息技术开始渗透进人类日常生活的方方面面，产生的数据量也呈现指数型增长的态势。现有数据的量级已经远远超越了目前人力所能处理的范畴。如何管理和使用这些数据，逐渐成为数据科学领域中一个全新的研究课题。Python 语言在最近十年发展迅猛，大量的数据科学领域的从业者使用 Python 完成数据科学相关的工作，其中最突出的就是数据分析师。

学习目标

（1）掌握数据分析的概念与流程。
（2）了解数据分析的应用场景。
（3）了解 Python 在数据分析领域的优势。
（4）了解 Python 数据分析常用类库。
（5）掌握 Windows/Linux 系统中 Anaconda 的安装。
（6）掌握 Jupyter Notebook 的常用功能。

任务 1.1 认识数据分析

任务描述

数据分析作为大数据技术的重要组成部分，近年来随着大数据技术逐渐发展和成熟。数据分析技能，被认为是数据科学领域中数据从业人员需要具备的技能之一。与此同时，数据分析师也成了时下最热门的职业之一。数据分析技能的掌握是一个循序渐进的过程。明确数据分析概念、分析流程和分析方法等相关知识是迈出数据分析的第一步。

任务分析

（1）掌握广义的数据分析和狭义的数据分析的概念。
（2）掌握典型的数据分析流程。
（3）了解七大类常见的数据分析应用场景。

1.1.1 掌握数据分析的概念

数据分析是指用适当的分析方法对收集来的大量数据进行分析，提取有用信息和形成结论，对数据加以详细研究和概括总结的过程。随着计算机技术的全面发展，企业生产、收集、存储和处理数据的能力大大提高，数据量与日俱增。而在现实生活中，需要把这些繁多、复杂的数据通过统计分析进行提炼，以此研究出数据的发展规律，进而帮助企业管理层做出决策。

广义的数据分析包括狭义数据分析和数据挖掘。狭义的数据分析是指根据分析目的，采用对比分析、分组分析、交叉分析和回归分析等分析方法，对收集的数据进行处理与分析，提取有价值的信息，发挥数据的作用，得到一个特征统计量结果的过程。数据挖掘则是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，通过应用聚类模型、分类模型、回归和关联规则等技术，挖掘潜在价值的过程。

图 1-1 所示为广义数据分析的概念。广义数据分析是指依据一定的目标，通过统计分析、聚类、分类等方法发现大量数据中的目标隐含信息的过程。

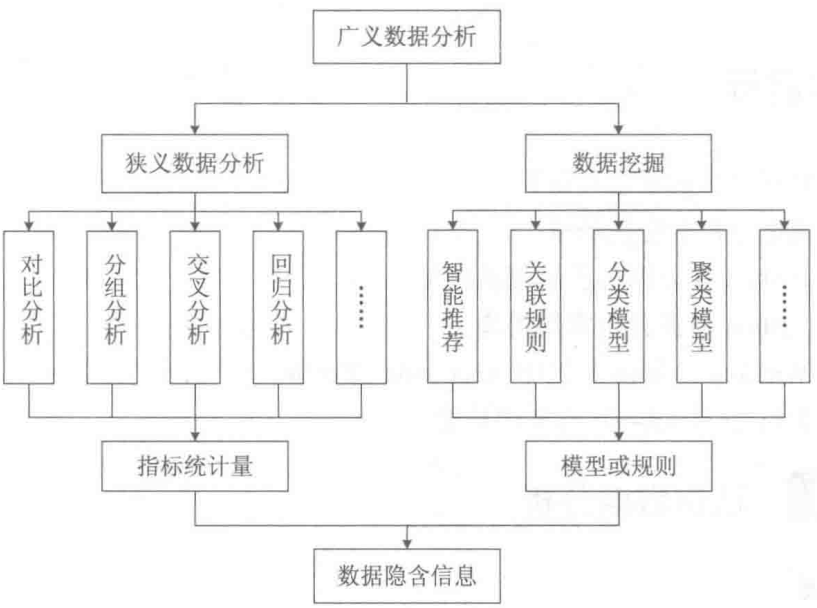


图 1-1 广义数据分析的概念

1.1.2 掌握数据分析的流程

数据分析已经逐渐演化为一种解决问题的过程，甚至是一种方法论。虽然每个公司都会根据自身需求和目标创建最适合的数据分析流程，但数据分析的核心步骤是一致的。图 1-2 所示是一个典型的数据分析的流程。

1. 需求分析

需求分析一词来源于产品设计，主要是指从用户提出的需求出发，挖掘用户内心的真实意图，并转化为产品需求的过程。产品设计的第一步就是需求分析，也是最关键的一步，因为需求分析决定了产品方向。错误的需求分析可能导致在产品实现过程中走入错误方向，甚至对企业造成损失。

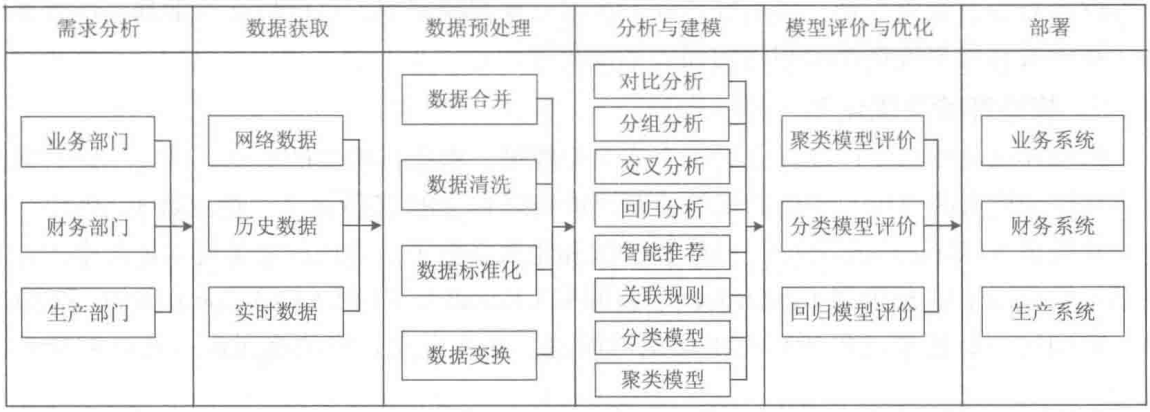


图 1-2 数据分析流程

数据分析中的需求分析是数据分析环节的第一步，也是非常重要的一步，决定了后续的分析方向和方法。数据分析中的需求分析的主要内容是，根据业务、生产和财务等部门的需要，结合现有的数据情况，提出数据分析需求的整体分析方向、分析内容，最终和需求方达成一致意见。

2. 数据获取

数据获取是数据分析工作的基础，是指根据需求分析的结果提取、收集数据。数据获取主要有两种方式：网络数据与本地数据。网络数据是指存储在互联网中的各类视频、图片、语音和文字等信息；本地数据则是指存储在本地数据库中的生产、营销和财务等系统的数据。本地数据按照数据时间又可以划分为两部分：历史数据与实时数据。历史数据是指系统在运行过程中遗存下来的数据，其数据量随系统运行时间的增加而增长；实时数据是指最近一个单位时间周期（月、周、日、小时等）内产生的数据。

在数据分析过程中，具体使用哪种数据获取方式，依据需求分析的结果而定。

3. 数据预处理

数据预处理是指对数据进行数据合并、数据清洗、数据标准化和数据变换，并直接用于分析建模的这一过程的总称。其中，数据合并可以将多张互相关联的表格合并为一张；数据清洗可以去掉重复、缺失、异常、不一致的数据；数据标准化可以去除特征间的量纲差异；数据变换则可以通过离散化、哑变量处理等技术满足后期分析与建模的数据要求。在数据分析的过程中，数据预处理的各个过程互相交叉，并没有明确的先后顺序。

4. 分析与建模

分析与建模是指通过对比分析、分组分析、交叉分析、回归分析等分析方法，以及聚类模型、分类模型、关联规则、智能推荐等模型与算法，发现数据中的有价值信息，并得出结论的过程。

分析与建模的方法按照目标不同可以分为几大类。如果分析目标是描述客户行为模式的，可采用描述型数据分析方法，同时还可以考虑关联规则、序列规则和聚类模型等。如果分析目标是量化未来一段时间内某个事件发生概率的，则可以使用两大预测分析模型，即分类预测模型和回归预测模型。在常见的分类预测模型中，目标特征通常都是二元数据，

例如欺诈与否、流失与否、信用好坏等。在回归预测模型中，目标特征通常都是连续型数据，常见的有股票价格预测和违约损失率预测等。

5. 模型评价与优化

模型评价是指对于已经建立的一个或多个模型，根据其模型的类别，使用不同的指标评价其性能优劣的过程。常用的聚类模型评价指标有 ARI 评价法（兰德系数）、AMI 评价法（互信息）、V-measure 评分、FMI 评价法和轮廓系数等。常用的分类模型评价指标有准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 值（F1 Value）、ROC 和 AUC 等。常用的回归模型评价指标有平均绝对误差、均方误差、中值绝对误差和可解释方差值等。

模型优化则是指模型性能在经过模型评价后已经达到了要求，但在实际生产环境应用过程中，发现模型的性能并不理想，继而对模型进行重构与优化的过程。在多数情况下，模型优化和分析与建模的过程基本一致。

6. 部署

部署是指将数据分析结果与结论应用至实际生产系统的过程。根据需求的不同，部署阶段可以是一份包含了现状具体整改措施的数据分析报告，也可以是将模型部署在整个生产系统的解决方案。在多数项目中，数据分析师提供的是一份数据分析报告或者一套解决方案，实际执行与部署的是需求方。

1.1.3 了解数据分析应用场景

企业使用数据分析解决不同的问题，实际应用的数据分析场景主要分为以下 7 类。

1. 客户分析（Customer Analytics）

客户分析主要是根据客户的基本数据信息进行商业行为分析，首先界定目标客户，根据客户的需求、目标客户的性质、所处行业的特征以及客户的经济状况等基本信息，使用统计分析方法和预测验证法分析目标客户，提高销售效率。其次了解客户的采购过程，根据客户采购类型、采购性质进行分类分析，制定不同的营销策略。最后还可以根据已有的客户特征进行客户特征分析、客户忠诚度分析、客户注意力分析、客户营销分析和客户收益分析。通过有效的客户分析能够掌握客户的具体行为特征，将客户细分，使得运营策略达到最优，提升企业整体效益等。

2. 营销分析（Sales and Marketing Analytics）

营销分析囊括了产品分析、价格分析、渠道分析、广告与促销分析这 4 类分析。产品分析主要是竞争产品分析，通过对竞争产品的分析制定自身产品策略。价格分析又可以分为成本分析和售价分析。成本分析的目的是降低不必要的成本；售价分析的目的是制定符合市场的价格。渠道分析是指对产品的销售渠道进行分析，确定最优的渠道配比。广告与促销分析则能够结合客户分析，实现销量的提升、利润的增加。

3. 社交媒体分析（Social Media Analytics）

社交媒体分析是以不同的社交媒体渠道生成的内容为基础，实现不同社交媒体的用户分析、访问分析和互动分析等。用户分析主要根据用户注册信息、登录平台的时间点和平

时发表的内容等用户数据，分析用户个人画像和行为特征；访问分析则是通过用户平时访问的内容分析用户的兴趣爱好，进而分析潜在的商业价值；互动分析根据互相关注对象的行为预测该对象未来的某些行为特征。同时，社交媒体分析还能为情感和舆情监督提供丰富的资料。

4. 网络安全（Cyber Security）

大规模网络安全事件的发生，例如 2017 年 5 月席卷全球的 WannaCry 病毒，让企业意识到网络攻击发生时预先快速识别的重要性。传统的网络安全主要依靠静态防御，处理病毒的主要流程是发现威胁、分析威胁和处理威胁。这种情况下，往往在威胁发生以后才能做出反应。新型的病毒防御系统可使用数据分析技术，建立潜在攻击识别分析模型，监测大量网络活动数据和相应的访问行为，识别可能进行入侵的可疑模式，做到未雨绸缪。

5. 设备管理（Plant and Facility Management）

设备管理同样是企业关注的重点。设备维修一般采用标准修理法、定期修理法和检查后修理法等方法。其中，标准修理法可能会造成设备过剩修理，修理费用高；检查后修理法解决了修理费用成本问题，但是修理前的准备工作繁多，设备的停歇时间过长。目前企业能够通过物联网技术收集和分析设备上的数据流，包括连续用电、零部件温度、环境湿度和污染物颗粒等多种潜在特征，建立设备管理模型，从而预测设备故障，合理安排预防性的维护，以确保设备正常作业，降低因设备故障带来的安全风险。

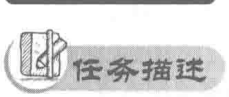
6. 交通物流分析（Transport and Logistics Analytics）

物流是物品从供应地向接收地的实体流动，是将运输、储存、装卸搬运、包装、流通加工、配送和信息处理等功能有机结合起来而实现用户要求的过程。用户可以通过业务系统和 GPS 定位系统获得数据，使用数据构建交通状况预测分析模型，有效预测实时路况、物流状况、车流量、客流量和货物吞吐量，进而提前补货，制定库存管理策略。

7. 欺诈行为检测（Fraud Detection）

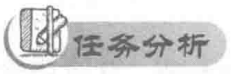
身份信息泄露及盗用事件逐年增长，随之而来的是欺诈行为和交易的增多。公安机关、各大金融机构、电信部门可利用用户基本信息、用户交易信息和用户通话短信信息等数据，识别可能发生的潜在欺诈交易，做到提前预防、未雨绸缪。以大型金融机构为例，通过分类模型分析方法对非法集资和洗钱的逻辑路径进行分析，找到其行为特征。聚类模型分析方法可以分析相似价格的运动模式。例如对股票进行聚类，可能发现关联交易及内幕交易的可疑信息。关联规则分析方法可以监控多个用户的关联交易行为，为发现跨账号协同的金融诈骗行为提供依据。

任务 1.2 熟悉 Python 数据分析的工具



Python 已经有将近 30 年的历史。在过去的将近 30 年中，Python 在运维工程师群体中受到广泛欢迎，然而却极少有企业将 Python 作为生产环境的首选语言。在最近几年，这一

情况有所改变。随着云计算、大数据以及人工智能技术的快速发展，Python 及其开发生态环境正在受到越来越多的关注。2011 年 1 月，在 TIOBE 编程语言排行榜中，它被评为 2010 年度语言。在 2017 年 5 月的编程语言排行榜中，Python 首次超越 C#，跃居第四。Python 已经成为整个计算机世界最重要的语言之一，更是数据分析的首选语言。



（1）了解数据分析常用的 Python、R 和 MATLAB 工具。

（2）了解使用 Python 工具进行数据分析的优势。

（3）了解 7 个 Python 数据分析常用类库。

1.2.1 了解数据分析常用工具

目前主流的数据分析语言有 Python、R、MATLAB 这 3 种。其中，Python 具有丰富和强大的库。它常被称为胶水语言，能够把用其他语言制作的各种模块（尤其是 C/C++）很轻松地连接在一起，是一门更易学、更严谨的程序设计语言。R 语言则是用于统计分析、绘图的语言和操作环境。它属于 GNU 系统的一个自由、免费、源代码开放的软件。MATLAB 的作用是进行矩阵运算、绘制函数与数据、实现算法、创建用户界面和连接其他编程语言的程序等，主要应用于工程计算、控制设计、信号处理与通信、图像处理、信号检测、金融建模设计与分析等领域。

3 种语言均可以进行数据分析。表 1-1 从语言学习难易程度、使用场景、第三方支持、流行领域和软件成本 5 方面比较了 Python、R、MATLAB 这 3 种数据分析工具。

表 1-1 Python、R、MATLAB 这 3 种工具对比

	Python	R	MATLAB
语言学习难易程度	接口统一，学习曲线平缓	接口众多，学习曲线陡峭	自由度大，学习曲线较为平缓
使用场景	数据分析、机器学习、矩阵运算、科学数据可视化、数字图像处理、Web 应用、网络爬虫、系统运维等	统计分析、机器学习、科学数据可视化等	矩阵运算、数值分析、科学数据可视化、机器学习、符号计算、数字图像处理、数字信号处理、仿真模拟等
第三方支持	拥有大量的第三方库，能够简便地调用 C、C++、Fortran、Java 等其他程序语言	拥有大量的包，能够调用 C、C++、Fortran、Java 等其他程序语言	拥有大量专业的工具箱，在新版本中加入了对 C、C++、Java 的支持
流行领域	工业界＞学术界	工业界≈学术界	工业界≤学术界
软件成本	开源免费	开源免费	商业收费

1.2.2 了解Python数据分析的优势

结合1.2.1小节的不同数据分析工具的对比可以发现，Python是一门应用十分广泛的计算机语言，在数据科学领域具有无可比拟的优势。Python正在逐渐成为数据科学领域的主流语言。Python数据分析主要包含以下5个方面优势。

（1）语法简单精练。对于初学者来说，比起其他编程语言，Python更容易上手。

（2）有很多功能强大的库。结合在编程方面的强大实力，可以只使用Python这一种语言去构建以数据为中心的应用程序。

（3）功能强大。从特性观点来看，Python是一个混合体。丰富的工具集使它介于传统的脚本语言和系统语言之间。Python不仅具备所有脚本语言简单和易用的特点，还提供了编译语言所具有的高级软件工程工具。

（4）不仅适用于研究和原型构建，同时也适用于构建生产系统。研究人员和工程技术人员使用同一种编程工具，会给企业带来非常显著的组织效益，并降低企业的运营成本。

（5）Python是一门胶水语言。Python程序能够以多种方式轻易地与其他语言的组件“粘接”在一起。例如，Python的C语言API可以帮助Python程序灵活地调用C程序。这意味着用户可以根据需要给Python程序添加功能，或者在其他环境系统中使用Python。

1.2.3 了解Python数据分析常用类库

1. IPython

IPython是Python科学计算标准工具集的组成部分，它将其他所有相关的工具联系在一起，为交互式和探索式计算提供了一个强健而高效的环境。同时，它是一个增强的Python Shell，目的是提高编写、测试、调试Python代码的速度。IPython主要用于交互式数据并行处理，是分布式计算的基础架构。

另外，IPython还提供了一个类似于Mathematica的HTML笔记本、一个基于Qt框架的GUI控制台，具有绘图、多行编辑以及语法高亮显示等功能。

2. NumPy

NumPy是Numerical Python的简称，是一个Python科学计算的基础包。NumPy主要提供了以下内容。

（1）快速高效的多维数组对象ndarray。

（2）对数组执行元素级计算以及直接对数组执行数学运算的函数。

（3）读/写硬盘上基于数组的数据集的工具。

（4）线性代数运算、傅里叶变换及随机数生成的功能。

（5）将C、C++、Fortran代码集成到Python的工具。

除了为Python提供快速的数组处理能力外，NumPy在数据分析方面还有另外一个主要作用，即作为算法之间传递数据的容器。对于数值型数据，使用NumPy数组存储和处理数据要比使用内置的Python数据结构高效得多。此外，由低级语言（比如C和Fortran）编写的库可以直接操作NumPy数组中数据，无须进行任何数据复制工作。

3. SciPy

SciPy基于Python的开源代码，是一组专门解决科学计算中各种标准问题域的模块的

集合，特别是与 NumPy、Matplotlib、IPython 和 pandas 这些核心包一起使用时。SciPy 主要包含了 8 个模块，不同的模块有不同的应用，如用于插值、积分、优化、处理图像和特殊函数等。模块的内容如表 1-2 所示。

表 1-2　SciPy 的模块及其简介

模块名称	简　介
scipy.integrate	数值积分和微分方程求解器
scipy.linalg	扩展了由 numpy.linalg 提供的线性代数求解和矩阵分解功能
scipy.optimize	函数优化器（最小化器）以及根查找算法
scipy.signal	信号处理工具
scipy.sparse	稀疏矩阵和稀疏线性系统求解器
scipy.special	SPECFUN［这是一个实现了许多常用数学函数（如伽马函数）的 Fortran 库］的包装器
scipy.stats	检验连续和离散概率分布（如密度函数、采样器、连续分布函数等）的函数与方法、各种统计检验的函数与方法，以及各类描述性统计的函数与方法
scipy.weave	利用内联 C++代码加速数组计算的工具

4. pandas

pandas 是 Python 的数据分析核心库，最初被作为金融数据分析工具而开发出来。pandas 为时间序列分析提供了很好的支持。它提供了一系列能够快速、便捷地处理结构化数据的数据结构和函数。Python 之所以成为强大而高效的数据分析环境与它息息相关。

pandas 兼具 NumPy 高性能的数组计算功能以及电子表格和关系型数据库（如 SQL）灵活的数据处理功能。它提供了复杂精细的索引功能，以便便捷地完成重塑、切片和切块、聚合及选取数据子集等操作。pandas 将是本书中使用的主要工具。

5. Matplotlib

Matplotlib 是最流行的用于绘制数据图表的 Python 库，是 Python 的 2D 绘图库。它非常适合创建出版物中用的图表。Matplotlib 最初由 John D.Hunter（JDH）创建，目前由一个庞大的开发团队维护。Matplotlib 的操作比较容易，用户只需用几行代码即可生成直方图、功率谱图、条形图、错误图和散点图等图形。Matplotlib 提供了 pylab 的模块，其中包括了 NumPy 和 pyplot 中许多常用的函数，方便用户快速进行计算和绘图。Matplotlib 与 IPython 结合得很好，提供了一种非常好用的交互式数据绘图环境。绘制的图表也是交互式的，读者可以利用绘图窗口中工具栏中的相应工具放大图表中的某个区域，或对整个图表进行平移浏览。

6. scikit-learn

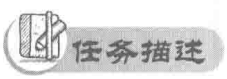
scikit-learn 是一个简单有效的数据挖掘和数据分析工具，可以供用户在各种环境下重复使用。而且 scikit-learn 建立在 NumPy、SciPy 和 Matplotlib 基础之上，对一些常用的算法方法进行了封装。目前，scikit-learn 的基本模块主要有数据预处理、模型选择、分类、聚类、数据降维和回归 6 个。在数据量不大的情况下，scikit-learn 可以解决大部分问题。对

算法不精通的用户在执行建模任务时，并不需要自行编写所有的算法，只需要简单地调用 scikit-learn 库里的模块就可以。

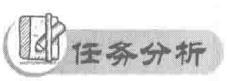
7. Spyder

Spyder（前身是 Pydee）是一个强大的交互式 Python 语言开发环境，提供高级的代码编辑、交互测试和调试等特性，支持 Windows、Linux 和 OS X 系统。Spyder 包含数值计算环境，得益于 IPython、NumPy、SciPy 和 Matplotlib 的支持。Spyder 可用于将调试控制台直接集成到图形用户界面的布局中。Spyder 的最大优点就是模仿 MATLAB 的“工作空间”，可以很方便地观察和修改数组的值。Spyder 的界面由许多窗格构成，用户可以根据自己的喜好调整它们的位置和大小。当多个窗格出现在一个区域时，将使用标签页的形式显示。界面包含了“Editor”“Object inspector”“Variable explorer”“File explorer”“Python Console”“History log”和“IPython Console”等区域，方便用户灵活运用 Python。

任务 1.3 安装 Python 的 Anaconda 发行版



Python 拥有 NumPy、SciPy、pandas、Matplotlib 和 scikit-learn 等功能齐全、接口统一的库，能为数据分析工作提供极大的便利。库的管理以及版本问题，使得数据分析人员并不能够专注于数据分析，而是将大量的时间花费在与环境配置相关的问题上。基于上述原因，Anaconda 发行版应运而生。



（1）了解 Python 的 Anaconda 发行版。

（2）在 Windows 和 Linux 系统中安装 Anaconda。

1.3.1 了解 Python 的 Anaconda 发行版

Anaconda 发行版 Python 预装了 150 个以上的常用 Packages，囊括了数据分析常用的 NumPy、SciPy、Matplotlib、pandas、scikit-learn 和 IPython 库，使得数据分析人员能够更加顺畅、专注地使用 Python 解决数据分析相关问题。

Python 的 Anaconda 发行版主要有以下几个特点。

（1）包含了众多流行的科学、数学、工程和数据分析的 Python 库。

（2）完全开源和免费。

（3）额外的加速和优化是收费的，但对于学术用途，可以申请免费的 License。

（4）全平台支持 Linux、Windows、Mac；支持 Python 2.6、2.7、3.4、3.5 和 3.6，可自由切换。

因此，推荐数据分析初学者（尤其是 Windows 系统用户）安装此 Python 发行版。读者只需要到 Anaconda 官方网站（http://continuum.io/downloads）下载适合自身的安装包即可。

1.3.2 在 Windows 系统中安装 Anaconda

进入 Anaconda 官方网站，下载 Windows 系统中的 Anaconda 安装包，选择 Python 3.0

以上版本。安装 Anaconda 的具体步骤如下。

（1）单击图 1-3 所示的“Next”按钮进入下一步。

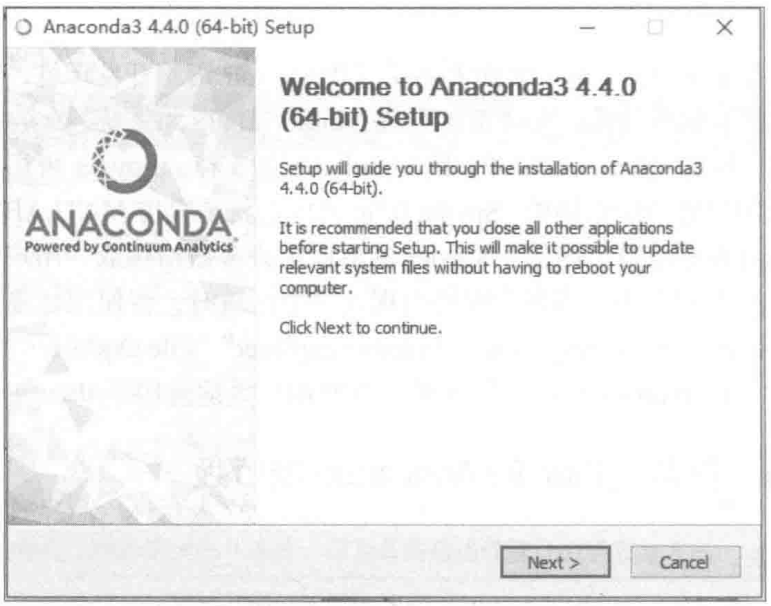


图 1-3　Windows 系统安装 Anaconda 步骤 1

（2）单击图 1-4 所示的“I Agree”按钮，同意上述协议并进入下一步。

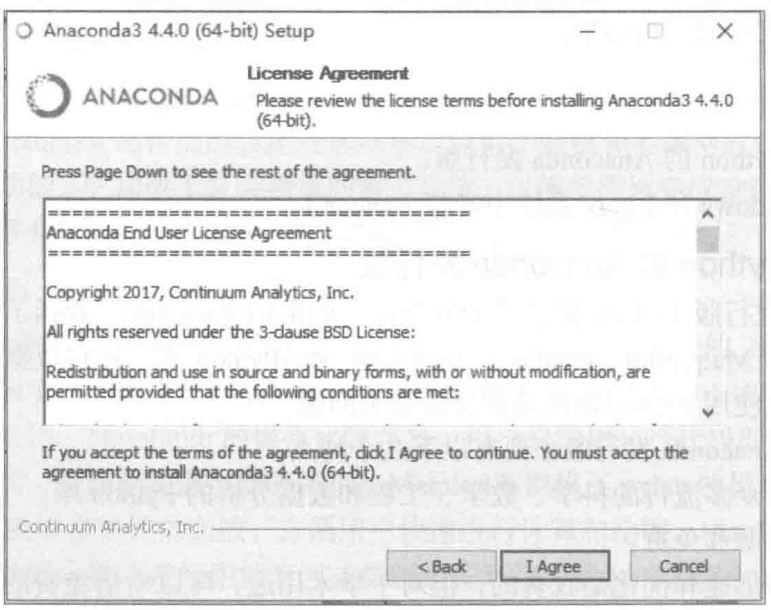


图 1-4　Windows 系统安装 Anaconda 步骤 2

（3）选择图 1-5 所示的“All Users(requires admin privileges)”单选按钮，进入下一步。

（4）单击“Browse”按钮，选择在指定的路径安装 Anaconda，如图 1-6 所示，选择完成后单击“Next”按钮，进入下一步。

（5）图 1-7 中的两个复选框分别代表了允许将 Anaconda 添加到系统路径环境变量中、Anaconda 使用的 Python 版本为 3.6。勾选后，单击“Install”按钮，等待安装结束。

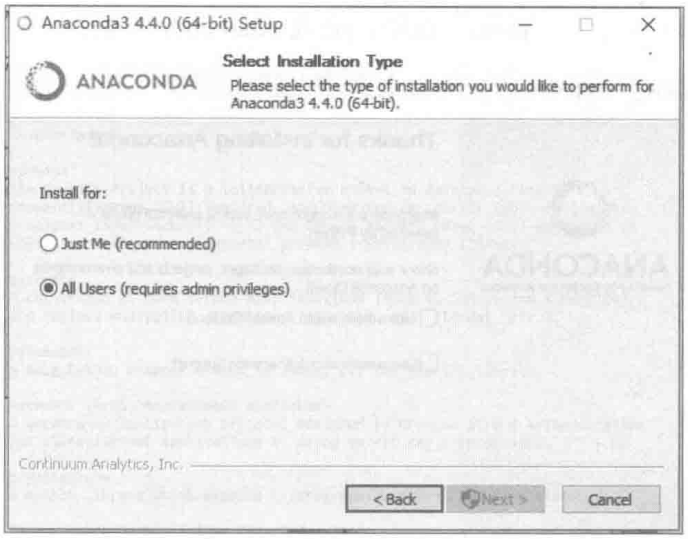


图 1-5 Windows 系统安装 Anaconda 步骤 3

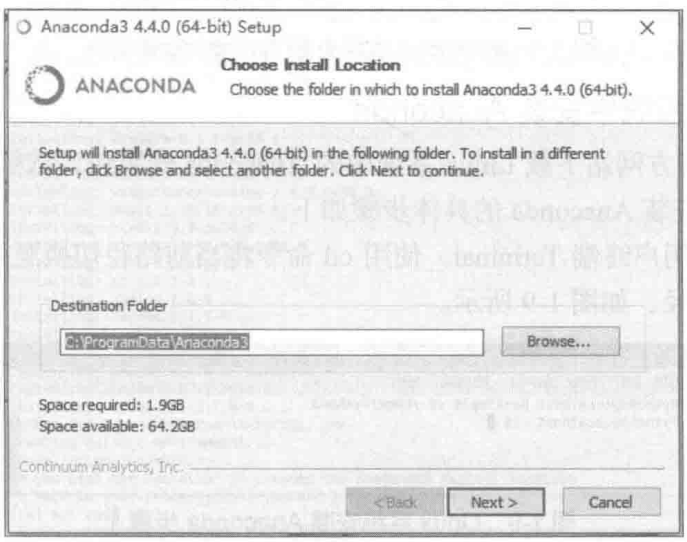


图 1-6 Windows 系统安装 Anaconda 步骤 4

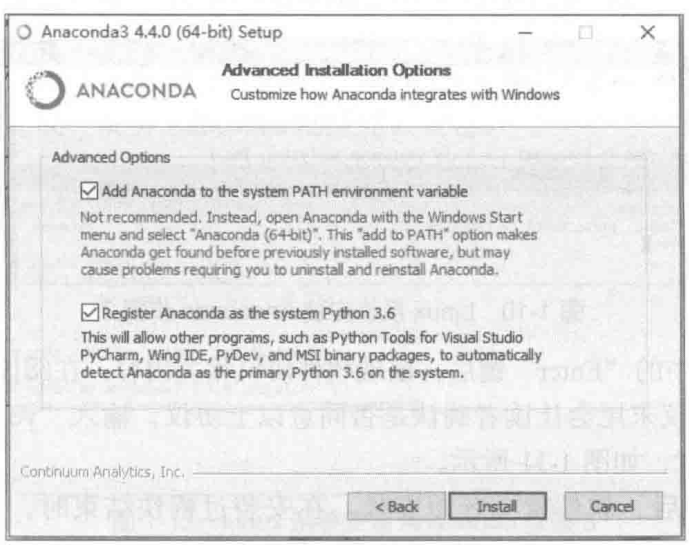


图 1-7 Windows 系统安装 Anaconda 步骤 5

（6）单击图 1-8 所示的“Finish”按钮，完成 Anaconda 安装。

图 1-8　Windows 系统安装 Anaconda 步骤 6

1.3.3　在 Linux 系统中安装 Anaconda

从 Anaconda 官方网站下载 Linux 系统中的 Anaconda 安装包，选择 Python 3.0 以上版本。Linux 系统中安装 Anaconda 的具体步骤如下。

（1）打开一个用户终端 Terminal。使用 cd 命令将当前路径切换至系统下 Anaconda 安装包所在的文件路径，如图 1-9 所示。

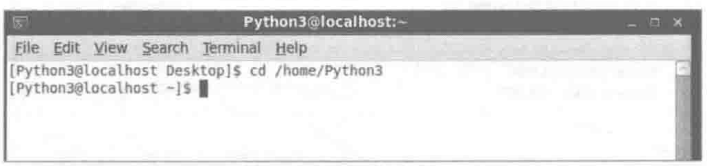


图 1-9　Linux 系统安装 Anaconda 步骤 1

（2）输入代码“bash Anaconda3-4.4.0-Linux-x86_64.sh”，进行安装，如图 1-10 所示。

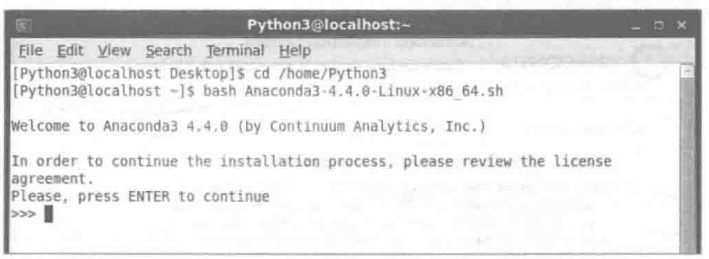


图 1-10　Linux 系统安装 Anaconda 步骤 2

（3）按下键盘中的“Enter”键后，出现软件协议相关内容，在阅读时连续按“Enter”键读取全文，在协议末尾会让读者确认是否同意以上协议，输入“yes”并按下键盘中的“Enter”键确认同意，如图 1-11 所示。

（4）同意协议后，软件就会开始安装。在安装过程快结束时，将提示读者是否将 Anaconda 的安装路径加入到系统当前用户的环境变量中，输入“yes”并按下键盘中的

“Enter”键确认同意，如图 1-12 所示。

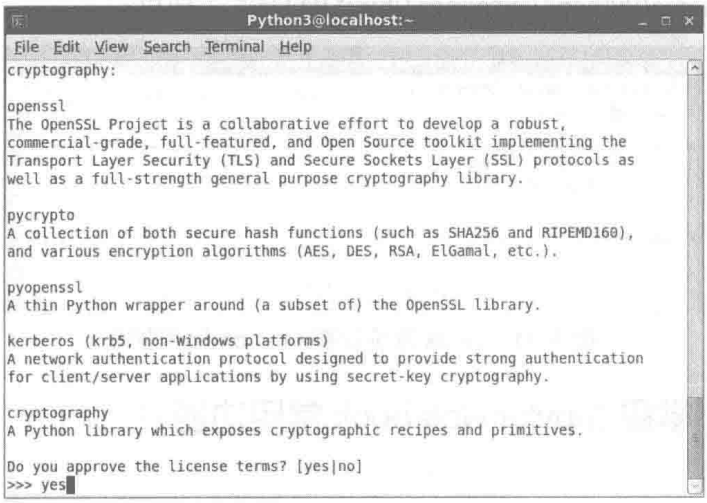


图 1-11　Linux 系统安装 Anaconda 步骤 3

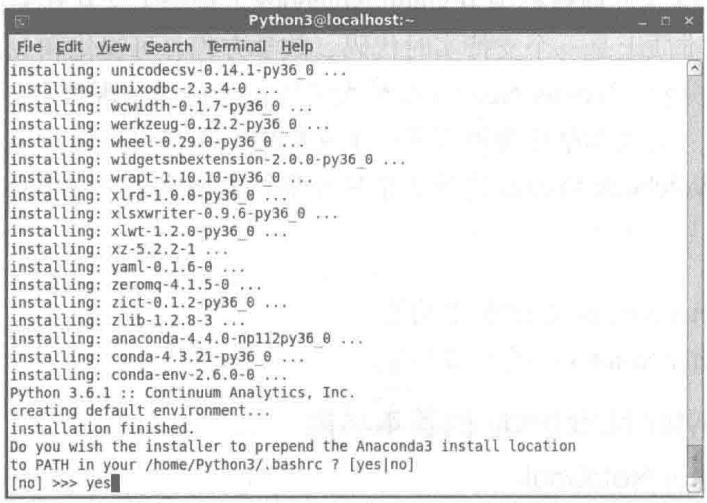


图 1-12　Linux 系统安装 Anaconda 步骤 4

（5）等待安装完成，完成后使用 Linux 系统的文本编辑器 VIM 或者 gedit 查看当前用户的环境变量。输入命令“vi /home/Python3/.bashrc”来查看文档，出现图 1-13 所示界面，表示环境变量配置完成，说明 Anaconda 已经完成安装。

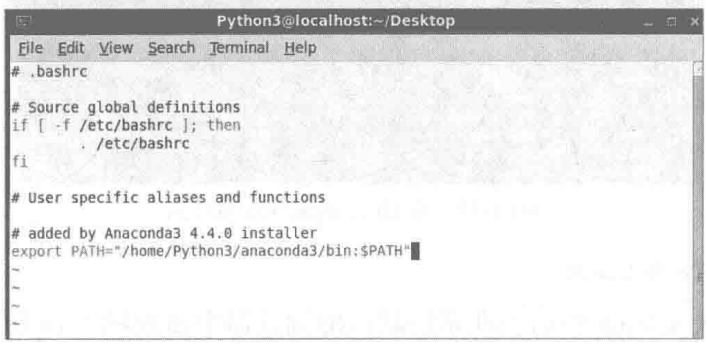


图 1-13　Linux 系统安装 Anaconda 步骤 5

（6）如果未配置完成，在图 1-14 所示的界面末尾添加 Anaconda 安装目录的环境变量“export PATH= " /home/Python3/anaconda3/bin:$PATH " ” 即可。

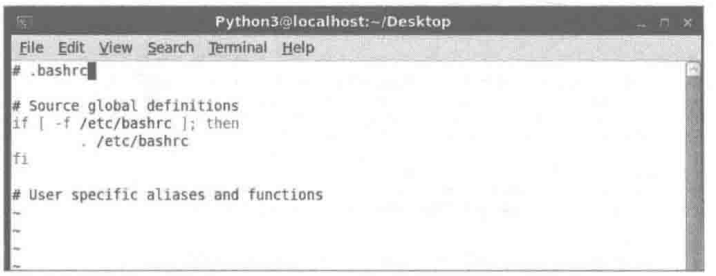


图 1-14　Linux 系统安装 Anaconda 步骤 6

任务 1.4　掌握 Jupyter Notebook 常用功能

任务描述

Jupyter Notebook（此前被称为 IPython Notebook）是一个交互式笔记本，支持运行 40 多种编程语言。它本质上是一个支持实时代码、数学方程、可视化和 Markdown 的 Web 应用程序。对于数据分析，Jupyter Notebook 最大的优点是可以重现整个分析过程，并将说明文字、代码、图表、公式和结论都整合在一个文档中。用户可以通过电子邮件、Dropbox、GitHub 和 Jupyter Notebook Viewer 将分析结果分享给其他人。

任务分析

（1）掌握 Jupyter Notebook 的基本功能。

（2）掌握 Jupyter Notebook 的高级功能。

1.4.1　掌握 Jupyter Notebook 的基本功能

1. 启动 Jupyter Notebook

在安装完成 Python、配置好环境变量并安装了 Jupyter Notebook 后，在 Windows 系统下的命令行或者在 Linux 系统下的终端输入命令“jupyter notebook”，即可启动 Jupyter Notebook，如图 1-15 所示。

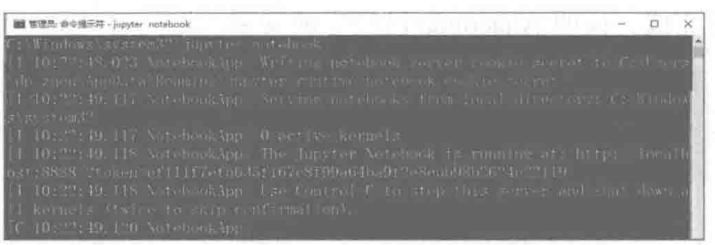


图 1-15　启动 Jupyter Notebook

2. 新建一个 Notebook

打开 Jupyter Notebook 以后会在系统默认的浏览器中出现图 1-16 所示的界面。单击右上方的“New”下拉按钮，出现下拉列表，如图 1-17 所示。

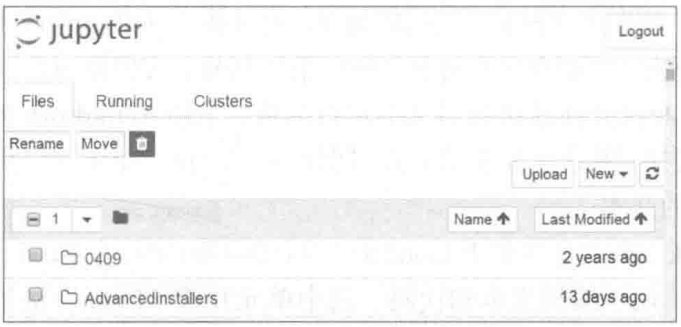


图 1-16　Jupyter Notebook 主页

图 1-17　New 下拉列表

在下拉列表中选择需要创建的 Notebook 类型。其中，“Text File”为纯文本型，“Folder”为文件夹，“Python 3”表示 Python 运行脚本，灰色字体表示不可用项目。选择“Python 3”选项，进入 Python 脚本编辑界面，如图 1-18 所示。

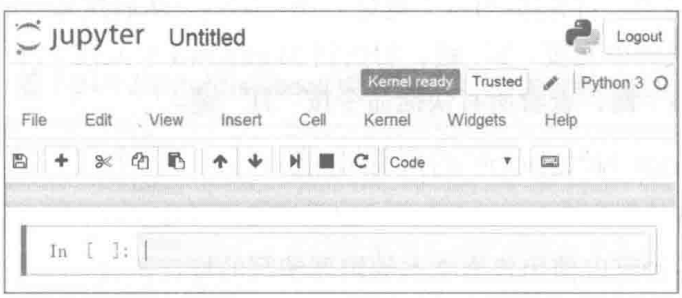


图 1-18　Jupyter Notebook Python 3 脚本编辑界面

3. Jupyter Notebook 的界面及其构成

Notebook 文档由一系列单元（Cell）构成，主要有两种形式的单元，如图 1-19 所示。

图 1-19　Jupyter Notebook 的两种单元

（1）代码单元。这里是读者编写代码的地方，通过按“Shift+Enter”组合键运行代码，其结果显示在本单元下方。代码单元左边有“In [ ]:”编号，方便使用者查看代码的执行次序。

（2）Markdown 单元。在这里可对文本进行编辑，采用 Markdown 的语法规范，可以设置文本格式，插入链接、图片甚至数学公式。同样，按“Shift+Enter”组合键可运行 Markdown 单元，显示格式化的文本。

Jupyter Notebook 编辑界面类似于 Linux 的 VIM 编辑器界面，在 Notebook 中也有两种模式。

（1）编辑模式。用于编辑文本和代码。选中单元并按“Enter”键进入编辑模式，此时单元左侧显示绿色竖线，如图 1-20 所示。

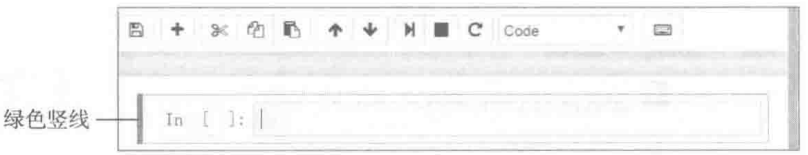


图 1-20　编辑模式

（2）命令模式。用于执行键盘输入的快捷命令。通过按“Esc”键进入命令模式，此时单元左侧显示蓝色竖线，如图 1-21 所示。

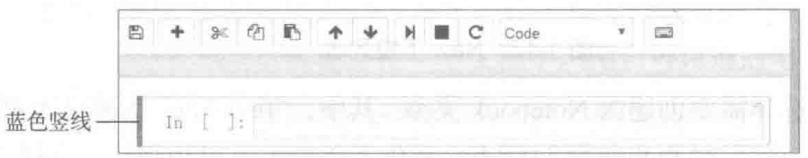


图 1-21　命令模式

如果要使用快捷键，首先按“Esc”键进入命令模式，然后按相应的键实现对文档的操作。例如，切换到代码单元按“Y”键，切换到 Markdown 单元按“M”键，在本单元的下方增加一单元按“B”键，查看所有快捷命令按“H”键。

1.4.2　掌握 Jupyter Notebook 的高级功能

1. Markdown

Markdown 是一种可以使用普通文本编辑器编写的标记语言。通过简单的标记语法，它可以使普通文本内容具有一定的格式。Jupyter Notebook 的 Markdown 单元比基础的 Markdown 的功能更加强大，下面将从标题、列表、字体、表格和数学公式编辑 5 个方面进行介绍。

（1）标题

标题是标明文章和作品等内容的简短语句。读者写报告或者写论文时，标题是不可或缺的，尤其是论文的章节等，需要使用不同级别的标题。Markdown 作为一款排版的工具，一般使用类 Atx 形式，在首行前加一个“#”字符代表一级标题，加两个“#”字符代表二级标题，以此类推。图 1-22 和图 1-23 所示分别为 Markdown 的代码和展示效果。

（2）列表

列表是一种由数据项构成的有限序列，即按照一定的线性顺序排列而成的数据项的集合。列表一般分为两种：一种是无序列表，使用一些图标标记，没有序号，没有排列顺序；另一种是有序列表，使用数字标记，有排列顺序。Markdown 对于无序列表，可使用星号、

加号或者减号作为列表标记；Markdown 对于有序列表，则使用数字“.”“ ”（一个空格）表示。图 1-24 和图 1-25 所示分别为列表的代码和运行结果。

```
# 一级标题
## 二级标题
### 三级标题
#### 四级标题
##### 五级标题
###### 六级标题
```

图 1-22 Jupyter Notebook 中的 Markdown 的标题代码

一级标题

二级标题

三级标题

四级标题

五级标题

六级标题

图 1-23 Jupyter Notebook 中的 Markdown 的标题展示

```
* Python
+ Python2
- Python3

1. Python
2. Python2
3. Python3
```

图 1-24 Jupyter Notebook 中的 Markdown 的列表代码

- Python
- Python2
- Python3

1. Python
2. Python2
3. Python3

图 1-25 Jupyter Notebook 中的 Markdown 的列表展示

（3）字体

文档中为了凸显部分内容，一般对文字使用加粗或斜体格式，使得该部分内容变得更加醒目。对于 Markdown 排版工具而言，通常使用星号“*”和下划线“_”作为标记字词的符号。前后有两个星号或下划线表示加粗，前后有 3 个星号或下划线表示斜体。图 1-26 和图 1-27 所示分别为加粗/斜体的代码和运行结果。

```
Python数据分析
**Python数据分析**
***Python数据分析***
__Python数据分析__
___python数据分析___
```

图 1-26 Jupyter Notebook 中的 Markdown 的加粗/斜体代码

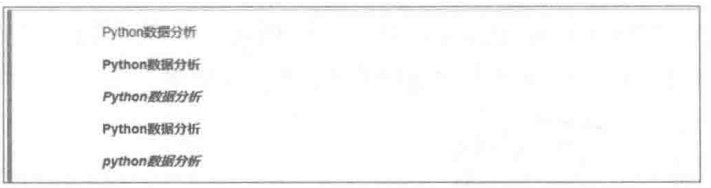


图 1-27　Jupyter Notebook 中的 Markdown 的加粗/斜体展示

（4）表格

使用 Markdown 同样也可以绘制表格。代码的第一行表示表头。第二行分隔表头和主体部分。从第三行开始，每一行代表一个表格行。列与列之间用符号“|”隔开，表格每一行的两边也要有符号“|”。图 1-28 和图 1-29 所示分别为表格的代码和运行结果。

图 1-28　Jupyter Notebook 中的 Markdown 的表格代码

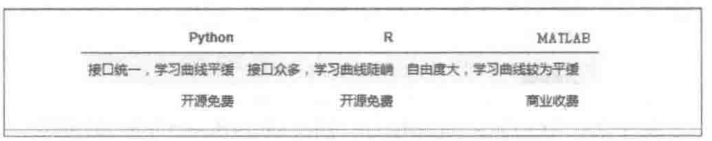


图 1-29　Jupyter Notebook 中的 Markdown 的表格展示

（5）数学公式编辑

LaTeX 是写科研论文的必备工具，不但能实现严格的文档排版，而且能编辑复杂的数学公式。在 Jupyter Notebook 的 Markdown 单元中也可以使用 LaTeX 来插入数学公式。在文本行中插入数学公式，应使用两个“$”符号，例如质能方程“$E = mc^2$”。如果要插入一个数学区块，则使用两个“$$”符号，比如用“$$ z = \frac{x}{y} $$”表示式（1-1）。

$$z = \frac{x}{y} \tag{1-1}$$

在输入上述公式的 LaTeX 表达式后，运行结果如图 1-30 所示。

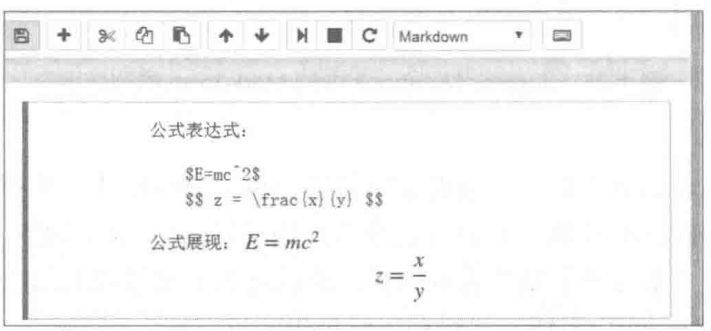


图 1-30　Jupyter Notebook 中的 Markdown 的 LaTeX 语法示例

2. 导出功能

Notebook 还有一个强大的特性，就是导出功能。可以将 Notebook 导出为多种格式，如 HTML、Markdown、reST、PDF（通过 LaTeX）等格式。其中，导出 PDF 功能，可以让读者不用写 LaTeX 即可创建漂亮的 PDF 文档。读者还可以将 Notebook 作为网页发布在自己

的网站上。甚至，可以导出为 reST 格式，作为软件库的文档。导出功能可通过选择“File”→“Download as”级联菜单中的命令实现，如图 1-31 所示。

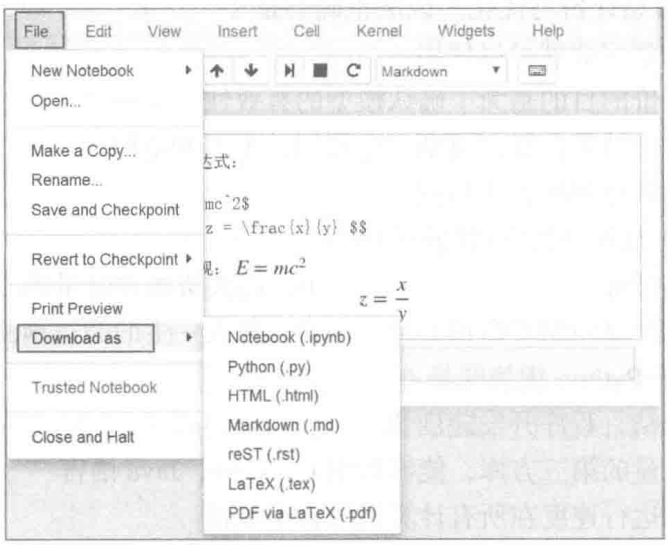


图 1-31 导出功能级联菜单

小结

本章根据目前的数据分析发展状况，将数据分析具象化，首先介绍了数据分析的概念、流程、目的以及应用场景，阐述了使用 Python 进行数据分析的优势，列举说明了 Python 数据分析重要类库的功能。紧接着阐述了 Anaconda 的特点，实现了在 Windows 和 Linux 两个系统中安装 Anaconda 数据分析环境。最后展现了 Python 数据分析工具 Jupyter Notebook 的优异特性及使用方法。

课后习题

1. 选择题

（1）下列关于数据和数据分析的说法正确的是（　　）。

A. 数据就是数据库中的表格

B. 文字、声音、图像这些都是数据

C. 数据分析不可能预测未来几天的天气变化

D. 数据分析的数据只能是结构化的

（2）下列关于数据分析流程的说法错误的是（　　）。

A. 需求分析是数据分析最重要的一部分

B. 数据预处理是能够建模的前提

C. 分析与建模时只能够使用数值型数据

D. 模型评价能够评价模型的优劣

（3）下列关于分析与建模流程的说法错误的是（　　）。

A. 传统的统计对比分析不属于分析与建模流程

B. 分析与建模的模型选择要根据需求确定

C．分析与建模时可以选择多个模型，同时分析

D．分析与建模工作是数据分析的核心

（4）下列关于模型评价与优化的说法正确的是（　　）。

A．模型构建完成就可以使用

B．模型评价的目的是为了确认模型的有效性

C．模型评价结果良好，模型一定可用，不需要重构

D．所有的模型评价方法相同

（5）下列不属于数据分析应用场景的是（　　）。

A．产品销量分析　　B．码头货物吞吐量预测

C．计算机硬盘使用寿命预测　　D．某人一生的命运预测

（6）下列不属于 Python 优势的是（　　）。

A．语法简洁，程序开发速度快

B．拥有大量的第三方库，能够调用 C、C++、Java 语言

C．程序的运行速度在所有计算机语言中最快

D．开源免费

（7）Jupyter Notebook 不具备的功能是（　　）。

A．Jupyter Notebook 可以直接生成一份交互式文档

B．Jupyter Notebook 可以安装 Python 库

C．Jupyter Notebook 可以导出 HTML 文件

D．Jupyter Notebook 可以将文件分享给他人

（8）【多选】下列关于 Jupyter Notebook 的描述错误的是（　　）。

A．Jupyter Notebook 有两种模式

B．Jupyter Notebook 有两种单元形式

C．Jupyter Notebook Markdown 无法使用 LaTeX 语法

D．Jupyter Notebook 仅仅支持 Python 语言

（9）【多选】下列关于 Python 数据分析库的描述错误的是（　　）。

A．NumPy 的在线安装不需要其他任何辅助工具

B．SciPy 的主要功能是可视化图表

C．pandas 能够实现数据的整理工作

D．scikit-learn 包含所有算法

（10）【多选】下列属于 Anaconda 主要特点的是（　　）。

A．包含了众多流行的科学、数学、工程、数据分析的 Python 包

B．完全开源和免费

C．支持 Python 2.6、2.7、3.4、3.5、3.6，可自由切换

D．额外的加速和优化是免费的

2．操作题

（1）在自用计算机上完成 Python Anaconda 发行版安装。

（2）使用 Jupyter Notebook 创建一个 Hello World 程序，并导出为 HTML 文件。