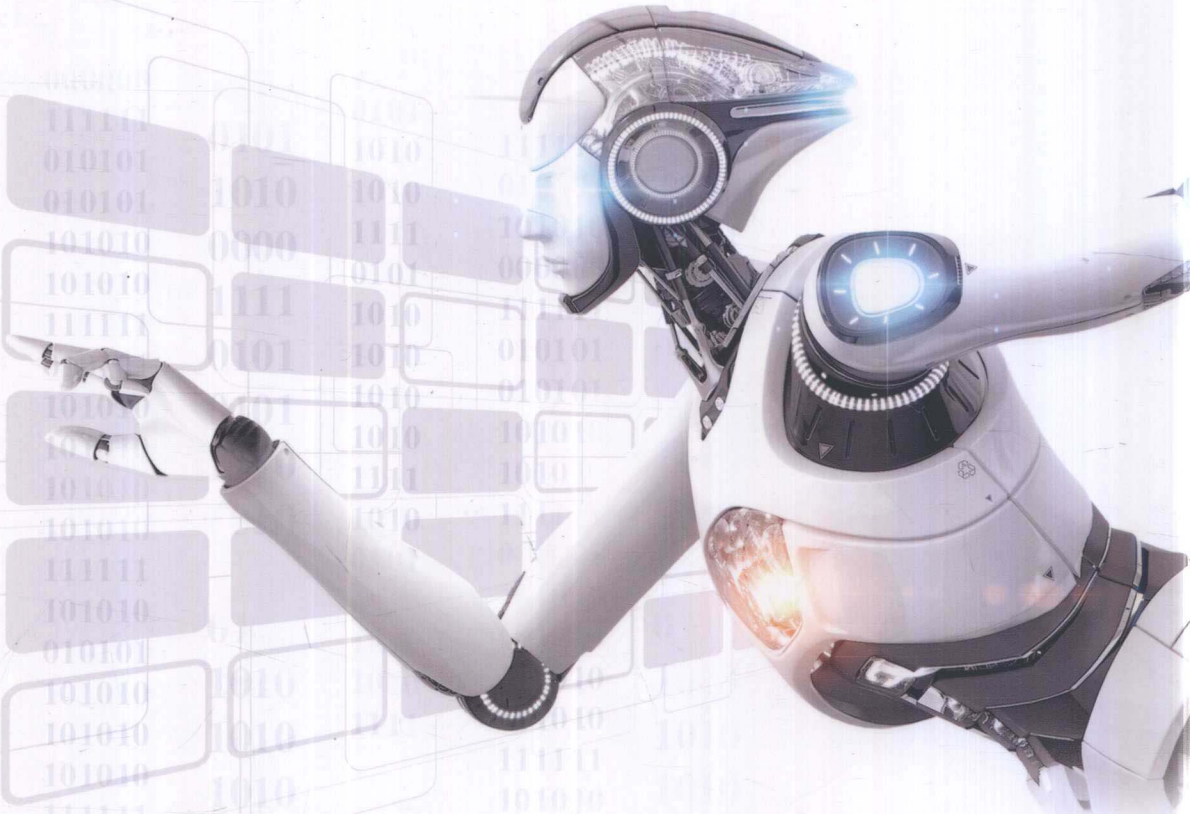

Python 数据分析与应用：从数据获取到可视化

Python应用编程丛书

黑马程序员　编著

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

Python应用编程丛书

Python 数据分析与应用：从数据获取到可视化

黑马程序员　编著

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

内 容 简 介

本书采用理论与案例相结合的形式，以 Anaconda 为主要开发工具，系统、全面地介绍了 Python 数据分析的相关知识。全书共分为 9 章，第 1 章介绍了数据分析的基本概念，以及开发工具的安装和使用；第 2~6 章介绍了 Python 数据分析的常用库及其应用，涵盖了科学计算库 NumPy、数据分析库 Pandas、数据可视化库 Matplotlib、Seaborn 与 Bokeh；第 7、8 章主要介绍了时间序列和文本数据的分析；第 9 章结合之前所学的技术开发了一个综合案例，演示如何在项目中运用所学的知识。除了第 1 章外，其他章节都包含了很多示例和综合案例，通过动手操作和练习，可以帮助读者更好地理解和掌握所学的知识。

本书适合作为高等院校计算机相关专业的大数据技术类课程教材，也可以作为大数据技术爱好者入门用书。

图书在版编目（CIP）数据

Python 数据分析与应用：从数据获取到可视化/黑马程序员编著．—北京：中国铁道出版社，2019.1
（Python 应用编程丛书）
ISBN 978-7-113-25145-1

Ⅰ．①P… Ⅱ．①黑… Ⅲ．①软件工具－程序设计 Ⅳ．①TP311.561

中国版本图书馆 CIP 数据核字(2018)第 301026 号

书　　名：Python 数据分析与应用：从数据获取到可视化
作　　者：黑马程序员　编著

策　　划：秦绪好　翟玉峰　　**读者热线**：（010）63550836
责任编辑：翟玉峰　贾淑媛
封面设计：王　哲
封面制作：刘　颖
责任校对：张玉华
责任印制：郭向伟

出版发行：中国铁道出版社（100054，北京市西城区右安门西街 8 号）
网　　址：http://www.tdpress.com/51eds/
印　　刷：中煤（北京）印务有限公司
版　　次：2019 年 1 月第 1 版　2019 年 1 月第 1 次印刷
开　　本：787 mm×1 092 mm　1/16　**印张**：17　**字数**：409 千
书　　号：1～2 000 册
书　　号：ISBN 978-7-113-25145-1
定　　价：52.00 元

PREFACE

前言

随着大数据时代的到来，数据已经成为与物质资产和人力资本同样重要的基础生产要素，如何从数据里面发现并挖掘有价值的信息成为一个热门的研究课题。基于这种需求，数据分析技术应运而生。数据分析是有目的地收集、整理、加工和分析数据，提炼出有价值信息的一个过程，它可以帮助企业或个人预测未来趋势和行为，规避风险，使得商务和生产活动具有前瞻性。

Python 在数据分析、探索性计算、数据可视化等方面都有非常成熟的库和活跃的社区，从 21 世纪开始，在行业应用和学术研究中使用 Python 进行数据分析的势头越来越猛，对于要往数据分析方向发展的读者而言，学习 Python 数据分析是一个不错的选择。

为什么学习本书

本书站在初学者的角度，循序渐进地介绍了学习数据分析必备的基础知识，以及一些比较优秀的数据分析工具，帮助读者具备数据分析的相关技能，能够独立编写项目，以胜任 Python 数据分析工程师相关岗位的工作。

本书在讲解时，采用需求引入的方式，循序渐进地介绍了数据分析工具的基本使用，同时对一些比较特殊的时间序列和文本数据的分析进行了拓展讲解，提高了读者的开发兴趣和开发能力。

作为开发人员，要想真正掌握一门技术，离不开多动手练习，所以本书在绘声绘色讲解知识的同时，不断地增加案例，有针对某个知识点的示例程序，也有针对某章的案例，最大程度地帮助读者真正掌握 Python 数据分析的核心技术。

如何使用本书

本书基于 Python 3，系统全面地讲解了 Python 数据分析的基础知识，全书共 9 章，具体章节内容如下。

第 1 章主要是带领大家了解数据分析，包括数据分析产生背景、什么是数据分析、数据分析的应用场景、数据分析的流程、开发工具的基本使用及常见数据分析工具等。通过本章的学习，希望大家能够对数据分析有一个初步的认识，并为后续章节的学习准备好开发环境。

第 2 章主要针对科学计算库 NumPy 进行讲解，包括创建数组、数据类型、数组运算、

索引和切片操作、转置和轴对称、通用函数、使用数组处理数据、线性代数模块及随机数模块等，并结合酒鬼漫步的案例，讲解如何使用 NumPy 数组参与简单的运算。希望读者能熟练使用 NumPy 包，为后面章节的学习奠定基础。

第 3 章主要介绍 Pandas 的基础功能，包括数据结构分析、索引操作、算术运算与数据对齐、数据排序、统计计算与描述、层次化索引和读写操作，并结合北京高考分数线的分析案例，讲解如何使用 Pandas 操作数据。通过对本章的学习，希望大家可以用 Pandas 实现简单的操作，为后续深入学习打好扎实的基础。

第 4 章进一步介绍了 Pandas 的数据预处理，包括数据清洗、数据合并、数据重塑和数据转换，并结合预处理部分地区信息的案例，讲解了如何利用 Pandas 预处理数据。数据预处理是数据分析中必不可少的环节，希望大家要多加练习，并能够在实际场景中选择合理的方式对数据进行预处理操作，另外，还可以参考官网提供的文档深入学习。

第 5 章继续介绍了 Pandas 的聚合与分组运算，包括分组聚合的原理、分组操作、数据聚合及其他分组级运算，并结合运动员基本信息的案例，讲解如何在项目中应用分组与聚合运算。大家在学习与理解的同时，要多加练习，可根据具体情况选择合理的技术进行运用即可。

第 6 章主要介绍了几个数据可视化工具，包括 Python 2D 绘图库 Matplotlib、绘制统计数据的库 Seaborn 和交互式可视化的库 Bokeh，并结合某年旅游景点的案例，讲解如何使用 Matplotlib 库绘制图表辅助分析。希望通过本章的学习，读者可以体会到在数据分析中运用可视化工具的好处。

第 7 章围绕着时间序列数据分析进行了介绍，包括创建时间序列、时间序列的索引和切片操作、固定频率的时间序列、时间周期与计算、重采样、滑动窗口及时序模型 ARIMA，并结合预测股票收盘价的案例，讲解了在项目中如何用时序模型对时间序列数据进行预测分析。通过对本章内容的学习，读者应该掌握处理时间序列数据的一些技巧，并灵活加以运用。

第 8 章主要针对文本数据分析进行讲解，包括文本数据分析的工具、文本预处理、文本情感分析、文本相似度和文本分类，并结合商品评价分析的案例，讲解了如何利用 NLTK 与 jieba 预处理和分析文本数据。希望通过对本章知识的学习，读者可以理解文本数据分析的原理，以便后续能基于机器学习更深入地去探索。

第 9 章是一个完整的实战项目，用于统计分析当前北京租房的信息，包括数据收集、预处理数据、数据分析，以及利用图表展现数据。希望通过对本章的学习，读者能够灵活地运用数据分析的技术，具备开发简单项目的能力。

在学习过程中，读者一定要亲自实践本书中的案例代码。如果不能完全理解书中所讲知识，

读者可以登录博学谷平台，通过平台中的教学视频进行深入学习。学习完一个知识点后，要及时在博学谷平台上进行测试，以巩固学习内容。

另外，如果读者在理解知识点的过程中遇到困难，建议不要纠结于某个地方，可以先往后学习。通常来讲，通过逐渐深入的学习，前面不懂和疑惑的知识点也就能够理解了。在学习编程的过程中，一定要多动手实践，如果在实践的过程中遇到问题，建议多思考，理清思路，认真分析问题发生的原因，并在问题解决后总结出经验。

致　谢

本书的编写和整理工作由传智播客教育科技股份有限公司完成，主要参与人员有吕春林、高美云、王晓娟、孙东等。全体人员在近一年的编写过程中付出了很多辛勤的汗水，在此一并表示衷心的感谢。

意见反馈

尽管我们付出了最大的努力，但书中难免会有不妥之处，欢迎各界专家和读者朋友们来信给予宝贵意见，我们将不胜感激。您在阅读本书时，如发现任何问题或有不认同之处，可以通过电子邮件与我们取得联系。

请发送电子邮件至：itcast_book@vip.sina.com。

黑马程序员

2018 年 11 月 12 日于北京

CONTENTS

目 录

第 1 章 数据分析概述 1

1.1 数据分析的背景 1
1.2 什么是数据分析 2
1.3 数据分析的应用场景 2
1.4 数据分析的流程 3
1.5 为什么选择 Python 做数据分析 4
1.6 创建新的 Python 环境——Anaconda 5
1.6.1 Anaconda 发行版本概述 5
1.6.2 在 Windows 系统中安装 Anaconda 5
1.6.3 通过 Anaconda 管理 Python 包 7
1.7 启用 Jupyter Notebook 9
1.7.1 启动 Anaconda 自带的 Jupyter Notebook 9
1.7.2 Jupyter Notebook 界面详解 10
1.7.3 Jupyter Notebook 的基本使用 13
1.8 常见的数据分析工具 16
小结 17
习题 17

第 2 章 科学计算库 NumPy 19

2.1 认识 NumPy 数组对象 19
2.2 创建 NumPy 数组 21
2.3 ndarray 对象的数据类型 22
2.3.1 查看数据类型 22
2.3.2 转换数据类型 23
2.4 数组运算 24
2.4.1 矢量化运算 24
2.4.2 数组广播 25
2.4.3 数组与标量间的运算 25
2.5 ndarray 的索引和切片 26
2.5.1 整数索引和切片的基本使用 26
2.5.2 花式(数组)索引的基本使用 28
2.5.3 布尔型索引的基本使用 29
2.6 数组的转置和轴对称 30
2.7 NumPy 通用函数 32
2.8 利用 NumPy 数组进行数据处理 34
2.8.1 将条件逻辑转为数组运算 34
2.8.2 数组统计运算 34
2.8.3 数组排序 35
2.8.4 检索数组元素 36
2.8.5 唯一化及其他集合逻辑 36
2.9 线性代数模块 37
2.10 随机数模块 38
2.11 案例——酒鬼漫步 39
小结 40
习题 40

第 3 章 数据分析工具 Pandas 42

3.1 Pandas 的数据结构分析 42

3.1.1 Series 42
3.1.2 DataFrame 44
3.2 Pandas 索引操作及高级索引 46
3.2.1 索引对象 46
3.2.2 重置索引 47
3.2.3 索引操作 49
3.3 算术运算与数据对齐 53
3.4 数据排序 54
3.4.1 按索引排序 54
3.4.2 按值排序 55
3.5 统计计算与描述 56
3.5.1 常用的统计计算 57
3.5.2 统计描述 58
3.6 层次化索引 59
3.6.1 认识层次化索引 59
3.6.2 层次化索引的操作 64
3.7 读写数据操作 68
3.7.1 读写文本文件 68
3.7.2 读写 Excel 文件 70
3.7.3 读取 HTML 表格数据 72
3.7.4 读写数据库 73
3.8 案例——北京高考分数线统计分析 77
2.8.1 案例需求 77
2.8.2 数据准备 77
2.8.3 功能实现 78
小结 81
习题 81

第 4 章 数据预处理 83

4.1 数据清洗 83
4.1.1 空值和缺失值的处理 83
4.1.2 重复值的处理 88
4.1.3 异常值的处理 90
4.1.4 更改数据类型 94
4.2 数据合并 96
4.2.1 轴向堆叠数据 96
4.2.2 主键合并数据 99
4.2.3 根据行索引合并数据 103
4.2.4 合并重叠数据 105
4.3 数据重塑 106
4.3.1 重塑层次化索引 106
4.3.2 轴向旋转 109
4.4 数据转换 110
4.4.1 重命名轴索引 110
4.4.2 离散化连续数据 112
4.4.3 哑变量处理类别型数据 113
4.5 案例——预处理部分地区信息 115
4.5.1 案例需求 115
4.5.2 数据准备 115
4.5.3 功能实现 116
小结 123
习题 123

第 5 章 数据聚合与分组运算 125

5.1 分组与聚合的原理 125
5.2 通过 groupby() 方法将数据拆分成组 126
5.3 数据聚合 132
5.3.1 使用内置统计方法聚合数据 132
5.3.2 面向列的聚合方法 132
5.4 分组级运算 136
5.4.1 数据转换 136
5.4.2 数据应用 138
5.5 案例——运动员信息的分组与聚合 141

5.5.1 案例需求 141
5.5.2 数据准备 141
5.5.3 功能实现 142
小结 146
习题 147

第 6 章 数据可视化 149

6.1 数据可视化概述 149
6.1.1 什么是数据可视化 149
6.1.2 常见的图表类型 150
6.1.3 数据可视化的工具 154
6.2 Matplotlib——绘制图表 155
6.2.1 通过 figure() 函数创建画布 155
6.2.2 通过 subplot() 函数创建单个子图 157
6.2.3 通过 subplots() 函数创建多个子图 158
6.2.4 通过 add_subplot() 方法添加和选中子图 160
6.2.5 添加各类标签 161
6.2.6 绘制常见图表 162
6.2.7 本地保存图形 167
6.3 Seaborn——绘制统计图形 168
6.3.1 可视化数据的分布 168
6.3.2 用分类数据绘图 174
6.4 Bokeh——交互式可视化库 178
6.4.1 认识 Bokeh 库 178
6.4.2 通过 Plotting 绘制图形 179
6.5 案例——画图分析某年旅游景点数据 180
6.5.1 案例需求 181
6.5.2 数据准备 181
6.5.3 功能实现 181
小结 185
习题 185

第 7 章 时间序列分析 187

7.1 时间序列的基本操作 187
7.1.1 创建时间序列 187
7.1.2 通过时间戳索引选取子集 189
7.2 固定频率的时间序列 191
7.2.1 创建固定频率的时间序列 191
7.2.2 时间序列的频率、偏移量 193
7.2.3 时间序列的移动 195
7.3 时间周期及计算 196
7.3.1 创建时期对象 196
7.3.2 时期的频率转换 198
7.4 重采样 198
7.4.1 重采样方法（resample） 199
7.4.2 降采样 200
7.4.3 升采样 201
7.5 数据统计——滑动窗口 203
7.6 时序模型——ARIMA 206
7.7 案例——股票收盘价分析 207
7.7.1 案例需求 207
7.7.2 数据准备 207
7.7.3 功能实现 208
小结 213
习题 214

第 8 章 文本数据分析 216

8.1 文本数据分析工具 216
8.1.1 NLTK 与 jieba 概述 216

8.1.2 安装 NLTK 和下载语料库 217
8.1.3 jieba 库的安装 219
8.2 文本预处理 220
8.2.1 预处理的流程 220
8.2.2 分词 221
8.2.3 词性标注 223
8.2.4 词形归一化 224
8.2.5 删除停用词 226
8.3 文本情感分析 227
8.4 文本相似度 229
8.5 文本分类 232
8.6 案例——商品评价分析 235
8.6.1 案例需求 235
8.6.2 数据准备 236
8.6.3 功能实现 236
小结 240
习题 240
第 9 章 数据分析实战——北京租房数据统计分析 242
9.1 数据来源 242
9.2 数据读取 243
9.3 数据预处理 244
9.3.1 重复值和空值处理 244
9.3.2 数据转换类型 246
9.4 图表分析 247
9.4.1 房源数量、位置分布分析 248
9.4.2 户型数量分析 255
9.4.3 平均租金分析 258
9.4.4 面积区间分析 260
小结 262

第 1 章
数据分析概述

学习目标

- 了解数据分析的背景及应用场景。
- 掌握什么是数据分析以及数据分析的流程。
- 会创建 Python 环境，会使用 Anaconda 管理 Python 包。
- 会简单使用 Jupyter Notebook。
- 认识常见的数据分析工具。

近些年，随着网络信息技术与云计算技术的快速发展，网络数据得到了爆发性的增长，人们每天都生活在庞大的数据群体中，这一切标志着人们进入了大数据时代。在大数据环境的作用下，能够从数据里面发现并挖掘有价值的信息变得愈发重要，数据分析技术应运而生。

数据分析可以通过计算机工具和数学知识处理数据，并从中发现规律性的信息，以做出具有针对性的决策。由此可见，数据分析在大数据技术中扮演着不可估量的角色，接下来，我们就正式进入数据分析的学习吧！

1.1 数据分析的背景

半个世纪以来，随着计算机技术全面地融入社会生活，信息爆炸已经积累到一个开始引发变革的程度，它不仅使得世界上充斥着比以往更多的信息，而且增长速度也在逐步加快，驱使着人们进入了一个崭新的大数据时代。互联网（社交、搜索、电商）、移动互联网（微博）、物联网（传感器、智慧地球）、车联网、GPS、医学影像、安全监控、金融（银行、股市、保险）、电信（通信、短信）都在疯狂产生着数据。到目前为止，无论是线下的大超市还是线上的商城，每天都会产生 TB 级以上的数据量。

以前，人们得不到想要的数据，是因为数据库中没有相关的数据，然而，现在人们依旧得不到想要的数据，主要的原因就是数据库里面的数据太多了，而缺乏一些可以快速地从数据库

中获取利于决策的有价值数据的操作方法。世界知名的数据仓库专家阿尔夫·金博尔说过：“我们花了多年的时间将数据放入数据库，如今是该将它们拿出来的时候了。”

数据分析就可以从海量数据中获得潜藏的有价值的信息，帮助企业或个人预测未来的趋势和行为，使得商务和生产活动具有前瞻性。例如，创业者可以通过数据分析来优化产品，营销人员可以通过数据分析改进营销策略，产品经理可以通过数据分析洞察用户习惯，金融从业者可以通过数据分析规避投资风险，程序员可以通过数据分析进一步挖掘出数据价值。总之，数据分析可以使用数据来实现对现实事物进行分析和识别的能力。

在大数据时代中，数据处理技术得到了突飞猛进的发展，我们终于拥有了发现及挖掘隐藏在海量数据背后的信息，并且将这些信息转化为知识及智慧的能力，数据开始了从量变到质变的转化过程。

不管你从事什么行业，掌握了数据分析能力，往往在岗位上更有竞争力。

1.2 什么是数据分析

数据分析的数学基础在 20 世纪早期就已确立，但直到计算机的出现才使得实际操作成为可能，并使数据分析得以推广。

数据分析，是指使用适当的统计分析方法（如聚类分析、相关分析等）对收集来的大量数据进行分析，从中提取有用信息和形成结论，并加以详细研究和概括总结的过程。

数据分析的目的在于，将隐藏在一大批看似杂乱无章的数据信息中的有用数据集提炼出来，以找出所研究对象的内在规律。在统计学领域中，数据分析可以划分为如下三类：

（1）描述性数据分析：从一组数据中可以摘要并且描述这份数据的集中和离散情形。

（2）探索性数据分析：从海量数据中找出规律，并产生分析模型和研究假设。

（3）验证性数据分析：验证科研假设测试所需的条件是否达到，以保证验证性分析的可靠性。

其中，描述性数据分析隶属于初级数据分析，常见的分析方法有对比分析法、平均分析法、交叉分析法，而探索性和验证性数据分析属于高级数据分析，常见的分析方法有相关分析、因子分析、回归分析等。

1.3 数据分析的应用场景

随着大数据的应用越来越广泛，应用的行业也越来越多，我们每天都可以看到一些关于数据分析的新鲜应用，从而帮助人们获取有价值的信息。例如，网购时经常发现电商平台向我们推荐商品，往往这类商品都是我们最近浏览的，之所以电商平台能够如此了解用户的需求，主要是根据用户上网行为轨迹的相关数据进行分析，以达到精准营销的目的。接下来，我们一起来看看数据分析在一些领域中的应用。

1. 营销方面的应用

据杜克大学的一项研究显示，是习惯而非有意识的决策促成了我们每天 45% 的选择，这意

味着只要了解了习惯的形式，就可以更简单地控制它们。通过分析消费者的购物行为，便能够精准地预测下一步的消费，塔吉特公司便是一个最成功的例子。塔吉特公司给每名顾客分配了一个顾客码——利用它密切关注顾客所购买的物品，并且通过会员卡和购买方式获得个人信息。通过对消费者的购买信息进一步研究其购买习惯，发现各类有价值的目标群体，确认顾客人生中的特殊时刻，因为这时他们的购物习惯会变得特别灵活，适时地广告或优惠券将使他们开始全新的购物方式。

2. 医疗方面的应用

数据分析应用的计算能力可以让我们能够在几分钟内就可以解码整个 DNA，并且让我们可以制定出最新的治疗方案，同时可以更好地去预测疾病，就好比人们戴上智能手表就可以产生数据一样。数据分析同样可以帮助病人及早预防和预测疾病的发生，做到早治疗、早康复。大数据技术目前已经在医院应用监视早产婴儿和患病婴儿的情况，通过记录和分析婴儿的心跳，医生针对婴儿的身体可能会出现的不适症状做出预测，这样可以帮助医生更好地救助患儿。

3. 零售方面的应用

在美国零售业曾经有这样一个传奇故事，某家商店将纸尿裤和啤酒并排放在一起销售，结果纸尿裤和啤酒的销量双双增长！为什么看起来风马牛不相及的两种商品搭配在一起，能取到如此惊人的效果呢？后来经过分析发现，这些购买者多数是已婚男士，这些男士在为小孩购买纸尿裤的同时，会给自己购买一些啤酒。发现这个秘密后，沃尔玛超市就大胆地将啤酒摆放在纸尿裤旁边，这样顾客购买起来更方便，销量自然也会大幅上升。之所以讲“啤酒 – 纸尿裤”这个例子，其实是想告诉大家，挖掘数据潜在的价值，是零售业竞争的核心竞争力。

4. 网络安全方面的应用

传统的网络安全主要依靠静态防御及处理病毒的流程发现威胁、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以使用数据分析技术，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. 交通物流方面的应用

物流是指物品从供应地流向接受地的活动，包括运输、搬运、储存、保管、包装、装卸、流通加工和物流信息处理等基本功能，以满足社会的需求。用户可以通过业务系统和 GPS 定位系统获得数据，使用数据构建交流状况预测分析模型，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。

1.4　数据分析的流程

数据分析是基于商业目的，有目的地进行收集、整理、加工和分析数据，提炼出有价值的信息的一个过程。整个过程大致可分为五个阶段，具体如图 1–1 所示。

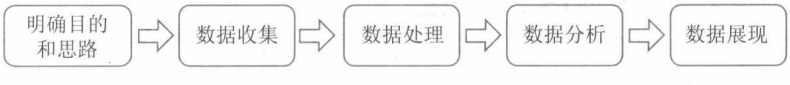

图 1–1　数据分析的过程

关于图 1-1 中流程的相关说明具体如下。

1. 明确目的和思路

在进行数据分析之前，我们必须要搞清楚几个问题，比如：数据对象是谁？要解决什么业务问题？并基于对项目的理解，整理出分析的框架和思路。例如，减少新客户的流失、优化活动效果、提高客户响应率等，不同的项目对数据的要求是不一样的，使用的分析手段也是不一样的。

2. 数据收集

数据收集是按照确定的数据分析思路和框架内容，有目的地收集、整合相关数据的一个过程，它是数据分析的基础。

3. 数据处理

数据处理是指对收集到的数据进行清洗、加工、整理，以便开展数据分析，它是数据分析前必不可少的阶段。这个过程是数据分析整个过程中是最耗时的，也在一定程度上保证了分析数据的质量。

4. 数据分析

数据分析是指通过分析手段、方法和技巧对准备好的数据进行探索、分析，从中发现因果关系、内部联系和业务规划，为商业提供决策参考。

到了这个阶段，要想驾驭数据开展数据分析，就要涉及工具和方法的使用，其一是要熟悉常规数据分析方法及原理，其二是要熟悉专业数据分析工具的使用，比如 Pandas、Maltlab 等，以便进行一些专业的数据统计、数据建模等。

5. 数据展现

俗话说：字不如表，表不如图。通常情况下，数据分析的结果都会通过图表方式进行展现，常用的图表包括饼图、折线图、条形图、散点图等。借助图表这种展现数据的手段，可以更加直观地让数据分析师表述想要呈现的信息、观点和建议。

1.5 为什么选择 Python 做数据分析

近年来，数据分析正在改变我们的工作方式，数据分析的相关工作也越来越受到人们的青睐。很多编程语言都可以做数据分析，比如 Python、R、Matlab 等，Python 凭借着自身无可比拟的优势，被广泛地应用到数据科学领域中，并逐渐衍生为主流语言。选择 Python 做数据分析，主要考虑的是 Python 具有以下优势：

1. 语法简单精练，适合初学者入门

比起其他编程语言，Python 的语法非常简单，代码的可读性很高，非常有利于初学者的学习。例如，在处理数据的时候，如果希望将用户性别数据数值化，也就是变成计算机可以运算的数字形式，这时便可以直接用一行列表推导式完成，十分简洁。

2. 拥有一个巨大且活跃的科学计算社区

Python 在数据分析、探索性计算、数据可视化等方面都有非常成熟的库和活跃的社区，这使得 Python 成为数据处理的重要解决方案。在科学计算方面，Python 拥有 Numpy、Pandas、

Matplotlib、Scikit-learn、IPython 等一系列非常优秀的库和工具，特别是 Pandas 在处理中型数据方面可以说有着无与伦比的优势，并逐渐成为各行业数据处理任务的首选库。

3. 拥有强大的通用编程能力

Python 的强大不仅体现在数据分析方面，而且在网络爬虫、Web 等领域也有着广泛的应用，对于公司来说，只需要使用一种开发语言就可以使完成全部业务成为可能。例如，我们可以使用爬虫框架 Scrapy 收集数据，然后交给 Pandas 库做数据处理，最后使用 Web 框架 Django 给用户做展示，这一系列的任务可以全部用 Python 完成，大大地提高了公司的技术效率。

4. 人工智能时代的通用语言

在人工智能领域中，Python 已经成为了最受欢迎的编程语言，这主要得益于其语法简洁、具有丰富的库和社区，使得大部分深度学习框架都优先支持 Python 语言编程。比如当今最火热的深度学习框架 TensorFlow，它虽然是使用 C++ 语言编写的，但是对 Python 语言支持最好。

5. 方便对接其他语言

Python 作为一门胶水语言，能够以多种方式与其他语言（比如 C 或 Java 语言）的组件“黏连”在一起，可以轻松地操作其他语言编写的库，这就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 语言。

1.6 创建新的 Python 环境——Anaconda

Python 的开发环境中拥有诸如 NumPy、Pandas、Matplotlib 等功能齐全的库，能够为数据分析工作提供极大的便利，不过，库的管理及版本问题不能让数据分析人员专注于数据分析，而是将大量的时间花费在解决包配置与包冲突等问题上。

基于上述需求，可以使用 Anaconda 进行开发，它是一个集成了大量常用扩展包的环境，能够避免包配置或兼容等各种问题。

1.6.1 Anaconda 发行版本概述

Anaconda 是一个可以便捷获取和管理包，同时对环境进行统一管理的发行版本，它包含了 conda、Python 在内的超过 180 个科学包及其依赖项。

Anaconda 发行版本具有以下特点：

（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 OS X，支持 Python 2.6、2.7、3.4、3.5、3.6，可以自由切换。

在此，我们推荐数据分析的初学者安装 Anaconda 进行学习。

1.6.2 在 Windows 系统中安装 Anaconda

以 Windows 系统为例，向读者介绍如何从 Anaconda 官方网站下载合适的安装包，并成功

安装到计算机上。在浏览器的地址栏中输入 https://www.anaconda.com/download/ 进入 Anaconda 的官方网站，如图 1-2 所示。

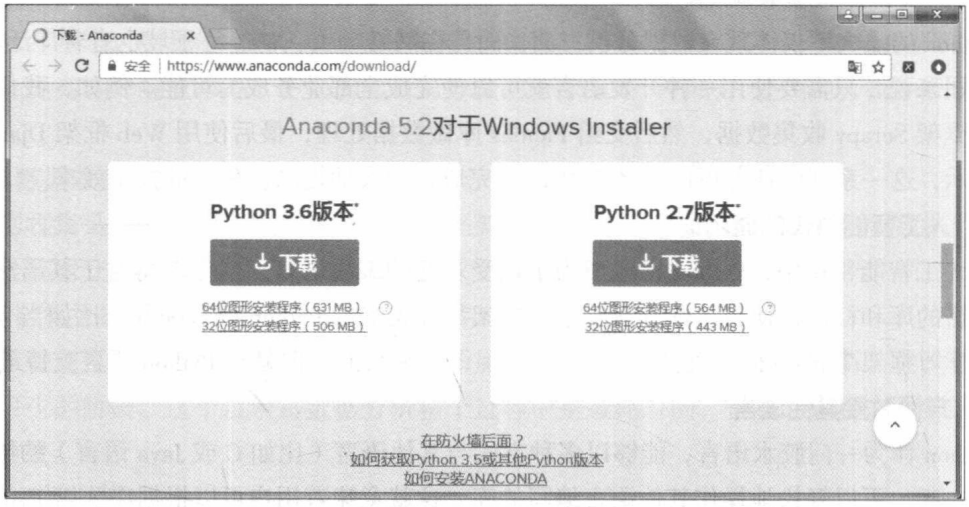

图 1-2　Anaconda 官网首页

图 1-2 的首页中展示了适合 Windows 平台下载的版本，大家选择合适的版本单击下载即可。这里，我们下载 “Python 3.6 版本” 下的 “64 位图形安装程序（631 MB）” 。

下载完以后，就可以进行安装了。Anaconda 的安装是比较简单的，直接按照提示选择下一步即可。为了避免不必要的麻烦，建议采用默认安装路径，在指定完安装路径后，继续单击 “Next” 按钮，窗口会提示是否勾选如下复选框选项，如图 1-3 所示。

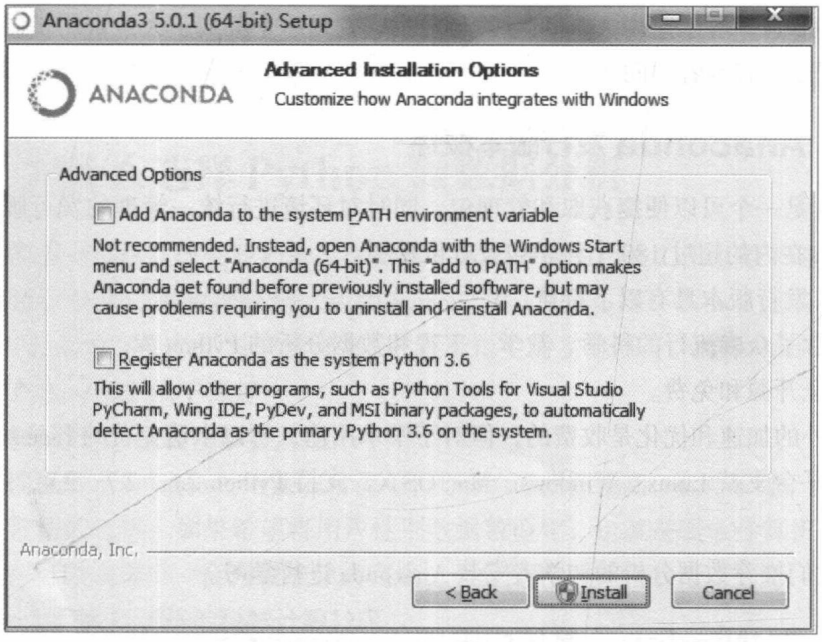

图 1-3　Anaconda 安装选项

在图 1-3 中，第 1 个复选框表示是否允许将 Anaconda 添加到系统路径环境变量中，第 2 个复选框表示 Anaconda 使用的 Python 版本是否为 3.6。勾选两个复选框，单击“Install”按钮，直至提示安装成功。

安装完以后，在系统左下角的“开始”菜单→“所有程序”中找到 Anaconda3 文件夹，可以看到该目录下包含了多个组件，如图 1-4 所示。

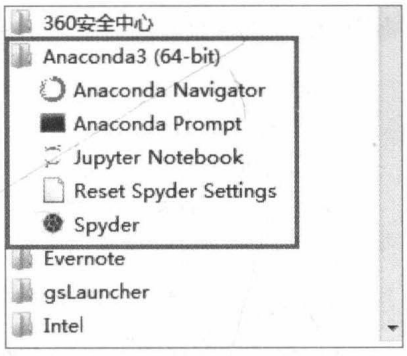

图 1-4　Anaconda3 的目录结构

关于图 1-4 中 Anaconda3 目录下的组件说明如下：

（1）Anaconda Navigator：用于管理工具包和环境的图形用户界面，后续涉及的众多管理命令也可以在 Navigator 中手动实现。

（2）Anaconda Prompt：Anaconda 自带的命令行。

（3）Jupyter Notebook：基于 Web 的交互式计算环境，可以编辑易于人们阅读的文档，用于展示数据分析的过程。

（4）Spyder：一个使用 Python 语言、跨平台的科学运算集成开发环境。

单击图 1-4 中的“Anaconda Navigator”图标，若能够成功启动 Anaconda Navigator，则说明安装成功，否则说明安装失败。Anaconda Navigator 成功打开后的首页界面如图 1-5 所示。

图 1-5　打开 Anaconda Navigator

1.6.3　通过 Anaconda 管理 Python 包

Anaconda 集成了常用的扩展包，能够方便地对这些扩展包进行管理，比如安装和卸载包，这些操作都需要依赖 conda。conda 是一个在 Windows、Mac OS 和 Linux 上运行的开源软件包管理系统和环境管理系统，可以快速地安装、运行和更新软件包及其依赖项。