DATA ANALYST ' S BIBLE

数据分析师

养成宝典

程显毅　曲平　李牧◎编著

只要 学会数据思维 >>>>>> 数据分析任你摆布

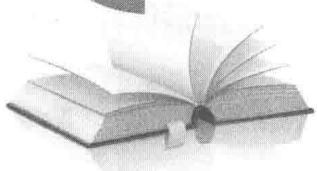
只要 有想法肯动脑 >>>>>> 有无基础皆可学会

只要 懂得指标设计 >>>>>> 项目落地信手捏来

机械工业出版社
CHINA MACHINE PRESS

DATA ANALYST'S BIBLE

数据分析师养成宝典



程显毅　曲平　李牧◎编著

机械工业出版社
CHINA MACHINE PRESS

在数据为主导的今天，对于一种已经成型的模型，“怎么用”通常不是问题，用个软件或者编几行程序就能得到结果了，问题一般都出在模型“什么时候用”和“用完了，然后呢”。本书就集中讨论后面两件事情。

本书共27章，分为业务理解篇（第1~4章）、指标设计篇（第5~7章）、数据建模篇（第8~16章）、价值展现篇（第17~19章）和实战进阶篇（第20~27章）。业务理解篇的目的是让读者建立正确的思维观，理解数据，熟悉业务；指标设计篇学习把数据转换为专家数据的一些技巧；数据建模篇以R语言为计算平台实施数据分析全过程；价值展现篇主要讨论如何撰写有价值的数据分析报告；实战进阶篇通过对8个经典案例的分析，使读者能够把学到的思维方法、实施工具应用到解决实际问题中，把数据变成价值。

本书可供数据科学相关技术人员阅读，也可作为高等院校数据科学相关专业的教材或培训教材，以及数据分析爱好者的参考读物。

图书在版编目（CIP）数据

数据分析师养成宝典/程显毅，曲平，李牧编著．—北京：机械工业出版社，2018.3
ISBN 978-7-111-59510-6

Ⅰ．①数… Ⅱ．①程… ②曲… ③李… Ⅲ．①数据处理
Ⅳ．①TP274

中国版本图书馆CIP数据核字（2018）第059214号

机械工业出版社（北京市百万庄大街22号 邮政编码 100037）
策划编辑：汤 枫 责任编辑：汤 枫
责任校对：张艳霞 责任印制：孙 炜

北京中兴印刷有限公司印刷

2018年4月第1版·第1次印刷
184mm×260mm·20印张·485千字
0001-3000册
标准书号：ISBN 978-7-111-59510-6
定价：69.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务
服务咨询热线：(010)88379833
读者购书热线：(010)88379649

网络服务
机 工 官 网：www.cmpbook.com
机 工 官 博：weibo.com/cmp1952
教育服务网：www.cmpedu.com
金 书 网：www.golden-book.com

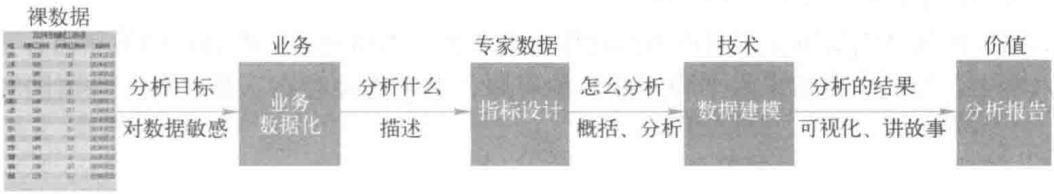
封面无防伪标均为盗版

如何使用本书

随着大数据时代的到来，企业管理者对数据价值的重视程度越来越高，他们渴望从企业内部数据、外部数据中获得更多的信息财富，并以此为依据，帮助自己做出正确的战略决策。如今在数据分析师的岗位上，大多数员工都是非统计专业出身，远远达不到专业数据分析要求，如何能够快速找到突破口，帮助对数据分析有兴趣的人员全面掌握数据分析技巧，基于此，本书旨在帮助读者解决如下困惑：

学习前的困惑	学习后将收获什么
零基础入门数据分析领域	只要有数据思维，数据分析任你摆布
不会编程	只要有想法，R 语言帮你搞定
对行业业务流程不了解	项目实际操作从业务思路到落地技能全掌握
不会写数据分析报告	掌握了前三项技能，写数据分析报告是小意思

全书分为 5 篇：业务理解篇、指标设计篇、数据建模篇、价值展现篇和实战进阶篇，从数据到价值的演化如下图所示。



业务数据化是把数据变为价值的先决条件，目的是建立正确的思维观，理解数据，熟悉业务；对数据描述、概括和分析是把数据变为价值的核心，是一个数据分析项目的实施过程；数据分析报告是数据价值的最终形态，好的数据分析报告是企业决策的重要依据，专业的数据分析报告能体现你的职场价值。

如果把整个数据分析过程比作开饭店，业务数据化相当于菜谱，虽然是纸上谈兵，但也是不可缺少的一环；描述、概括和分析相当于烧菜的厨艺，这是开饭店的先决条件，菜烧得好可以品尝，不能保证盈利；撰写数据分析报告相当于开饭店的理由，关键是说清楚如何让饭店盈利？盈利多少？菜谱、厨艺、开饭店理由三者相辅相成，缺少任何一环，盈利的目标都难以达成。

本书的特点如下：

1. 落地实用

全书分为 5 篇，共 27 章，每一章的内容都从实际业务出发，书中所提供的思维方法、分析框架、数据指标设计与操作步骤都可以直接运用到工作当中。

2. 重道轻术

“术”是数据分析方法和工具，“道”强调了如何对数据敏感，如何把数据分析工作融

入商业思考，弥补许多人只懂理论脱离实践的不足。

3. 零距离接触行业前沿

本书以 R 语言为计算平台，无论你是什么专业，无论你是否有编程基础，无论你是否学过统计，要想成为一名数据分析师，本书能帮到你。

4. 体系完整

近年来，数据分析师可谓是大数据时代最热门的职业，相关的资料五花八门，让读者无所适从。从学科体系来看，无非包括三个层次：理论、工具和技巧。但由于数据分析的特殊性，依赖于思维和业务，所以，市场上成体系的书籍并不多见，大多是讲理论和工具，本书试图在数据分析完整的体系上做些探索。

在本书的编写过程，得到了许多人的支持，再次表示感谢：

感谢南通大学-南通智能信息技术联合研究中心给予的资金资助。

感谢硅湖职业技术学院在培训、实验方面所给予的支持。

感谢南通大学教材建设资金资助。

感谢我的学生沈佳杰、谢璐、胡海涛、姚泽峰、周春瑜、孙丽丽、杨琴和赵丽敏在资料整理方面所做的贡献。

其次，感谢我的妻子和儿女们，正是你们的鼓励和支持，我才会走到今天，你们的鼓励和陪伴永远是我前进的动力。

最后，特别要感谢我的母亲和已故的父亲，感谢你们的养育之恩。仅以此书献给健在的母亲，希望母亲健康，健康，更健康。

数据分析领域发展迅猛，对许多问题作者并未做深入研究，一些有价值的新内容也来不及收入本书。加上作者知识水平和实践经验有限，书中难免存在不足之处，敬请读者批评指正。

程显毅

此为试读，需要完整PDF请访问：www.ertongbook.com

目　录

数据建模篇

价值展现篇

实战进阶篇

第 0 章　说在前面的话

俗话说“内行看门道，外行看热闹”。我们每天都在接触各式各样的数据，这些数据在一般人眼中就是数字而已，但在数据分析师看来，它们蕴含着取之不尽、用之不竭的宝藏。数据来源形式多样，数据质量参差不齐，数据分析师的工作就是对这些数据进行清晰整理，从中分析出有价值的结论与规律。

0.1　大数据分析案例

（1）大数据反腐倡廉

大数据则是反腐倡廉的“术”，而且是最直接最有效的“术”，比指望官员主动申报自己所有财产要靠谱得多。

首先，需要建立一张全国人口信息表（注意，是“一张”包含 13 亿多条记录的大数据）；然后，建立一张全国官员信息表，根据全国人口信息表，再建立起一张全国官员社会关系表。

要注意，建立官员社会关系表，就要用到本书讲的数据分析，从全国人口信息表中，挖掘出官员的各种社会关系。

有了全国官员信息表和官员社会关系表，这只是第一步，对他们的行为进行监控，才是关键，也就是大数据技术中的“用户行为分析”。

比如，可以监控官员及其社会关系的存取款、信用卡消费、股票基金、信托投资、出入境记录等，以银行为例，从银行系统中实时或近实时地获取官员及其社会关系的存取款记录、信用卡消费记录，并建立分析系统，从中发现官员贪腐的蛛丝马迹。

当这些监控分析系统运作建立起来以后，最高人民检察院、中纪委的同志们，就可以安心地在监控室里，看着大屏幕，静静等待系统发出的告警。

必须要指出的是，上述技术都是成熟的、可行的。

（2）大数据与房价

我国住建部建立的全国联网的个人房产信息，其实这就是一张大数据表，住建部完全可以建立两张表：全国居民个人房产信息表（以居民为索引）、全国房产信息登记表（以房产为索引），相互校验，相信一定可以发现不少问题。

重要的是，在（1）中提到的社会关系的分析手段，在这里仍然必不可少，至少要分析出以直系亲属为单位共同拥有的房产。

（3）大数据与智慧农业

为了解决全国各地各类农产品滞销的问题，可以建立一个全国性的农产品种植销售一体化的大数据平台，农民通过手机终端，就可以从这个大数据平台中看到全国每种农产品的种植面积，也需要上报自己的种植面积。

同时，如（1）和（2）中所述，最关键的是，这个大数据平台需要根据统计出的每种农产品的历年销售情况和区域，给出当年的销售预测，这样，就可以较好地向农民预警，避免农民一窝蜂地跟风种植“热销”农产品。

此外，经销商也可以从这个平台上看到农产品的种植情况和区域。

凡此种种，大数据分析就是用来消除信息孤岛，消除信息不对称带来的种种弊端。

0.2 数据分析

数据分析指的是将数据转化为价值的一个完整过程。作为一个完整过程，数据分析应该有很多环节。用看病来类比数据分析，是一个不错的例子，如图 0.1 所示。

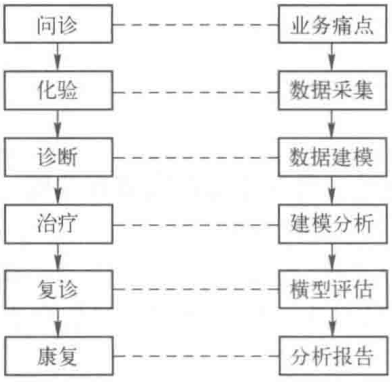


图 0.1 数据分析过程与看病过程类比

为了理解数据分析，首先要弄清楚数据分析与其他相关概念的区别。数据分析还没有公认的定义，百度的解释是：数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。

在使用中，数据分析可帮助人们做出判断，以便采取适当行动。

下面用宾州大学知名的 Dennis Lin 教授提到过的一个例子，显示数据分析与我们到底有多么紧密相关。这是一封大数据情书，信中写道：

亲爱的齐：

我们的感情，一年来正沿着健康的道路蓬勃发展。这主要表现在：

1）我们共通信 121 封，平均 3.01 天一封。其中你给我的信 51 封，占 42.1%；我给你的信 70 封，占 57.9%。每封信平均 1502 字，最长的达 5215 字，最短的也有 624 字。

2）约会共 98 次，平均 3.7 天一次。其中你主动约我 38 次，占 38.7%；我主动约你 60 次，占 61.3%。每次约会平均 3.8 小时，最长达 6.4 小时，最短的也有 1.6 小时。

3）我到你家看望你父母 38 次，平均每 9.4 天一次，你到我家看望我父母 36 次，平均 10 天一次。

以上充分证明一年来的交往我们形成了恋爱的共识，我们爱情的主流是互相了解、互相关心、互相帮助的，是平等的、互利的。

这封情书就是一个现实生活中利用数据进行分析的范例，如果情书通篇只谈我有多么爱

此为试读，需要完整PDF请访问：www.ertongbook.com

你，往往是一封空洞的书信。但是如果在情书中加入量化的数据，能够大大增加情书的说服力。

为了理解数据分析，接下来比较一下与数据分析相关的概念。

0.2.1 数据分析不同于信息化系统

信息化是以现代通信、网络、数据库技术为基础，将所研究对象的各要素汇总至数据库，供特定人群生活、工作、学习、辅助决策等和人类息息相关的各种行为相结合的一种技术，使用该技术后，可以极大地提高各种行为的效率，为推动人类社会进步提供极大的技术支持。

数据分析与信息化系统对比见表0.1。

表0.1 数据分析与信息化系统对比

指　标	数据分析	信息化系统
目的	把数据转化为价值	规范管理信息流
核心	数据思维	业务
驱动	数据	用户
人类文明的阶段	第四阶段（数据科学）	第三阶段（计算科学）
对业务的理解	数据业务化	业务数据化

0.2.2 数据分析不同于统计分析

统计分析是指运用统计方法及与分析对象有关的知识，从定量与定性的结合上进行的研究活动。它是继统计设计、统计调查、统计整理之后的一项十分重要的工作，是在前几个阶段工作的基础上通过分析达到对研究对象更深刻的认识。它又是在一定的选题下，集分析方案的设计、资料的搜集和整理而展开的研究活动。系统、完善的资料是统计分析的必要条件。

运用统计方法、定量与定性的结合是统计分析的重要特征。随着统计方法的普及，不仅统计工作者可以搞统计分析，各行各业的工作者都可以运用统计方法进行统计分析。只将统计工作者参与的分析活动称为统计分析的说法严格说来是不正确的。提供高质量、准确而又及时的统计数据和高层次、有一定深度、广度的统计分析报告是统计分析的产品。从一定意义上讲，提供高水平的统计分析报告是统计数据经过深加工的最终产品，这里的深加工指数据挖掘方法。

统计分析还是就数据分析数据，还不能讲数据的故事。数据分析与统计分析对比见表0.2。

表0.2 数据分析与统计分析对比

指　标	数据分析	统计分析
方法	统计+机器学习	纯统计
报告	讲故事	报表式

（续）

指　标	数据分析	统计分析
结果	价值	信息
执行与反馈	干完活后需要用数据监测是否达到既定目标？如果达到了，关键因素是什么？如果没达到，问题出在哪里？	活干完即结束，没有反馈

0.2.3　数据分析不同于数据挖掘

在许多时候，数据分析和数据挖掘常常一起出现，许多人容易把这两个概念搞混淆。

所谓数据挖掘（Data Mining，DM）是指从大量不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、有用的信息和知识的过程。其表现形式为概念、规则、模式等形式。挖掘的结果是数据分析报告的素材，挖掘得越深，数据故事讲得就越精彩。数据挖掘技术是做数据分析达人的基本功。

数据分析与数据挖掘对比见表0.3

表0.3　数据分析与数据挖掘对比

指　标	数据分析	数据挖掘
重心	偏向业务	偏向于算法
字面理解	对已有对象的全面描述、刻画、梳理后得出结论	对对象的剖析、分解、透视，发现不为人知的价值
比喻	分析沙子结构，用图	用铲子，挖沙子，看看沙子里埋的东西
目的性	极强，指导决策	找关系、做分类、搞聚类
数据来源	各种渠道	数据库
时效性	像一把枪，指哪打哪	搞武器研究，前期投入高，时间跨度长

在企业运转过程中，数据分析和数据挖掘的需求持续不断，两者相辅相成，不可或缺，同等重要。

0.2.4　数据分析不同于数据管理

随着计算机技术的发展，数据管理经历了人工管理、文件系统和数据库系统三个发展阶段。在数据库系统中所建立的数据结构，更充分地描述了数据间的内在联系，便于数据修改、更新与扩充，同时保证了数据的独立性、可靠性、安全性与完整性，减少了数据冗余，故提高了数据共享程度及数据管理效率。

数据管理只依赖于数据本身，与业务场景、思维习惯无关。数据管理是一种技能，而数据分析是一种艺术。

数据管理数据源一般要求数据是结构化的，数据分析的数据源可以是结构化、半结构化和非结构化的。

数据分析不同于数据管理，数据分析输入的是数据，输出是用于决策的数据分析报表，而数据管理输入的是数据，输出的还是数据。

0.2.5　数据分析不同于商业智能

数据分析只是一种工具（一种系统化分析问题的方式），可以很简单，也可以很复杂。

商业智能则是一种产品/服务，这个产品/服务可能包含报表、分析、管理等利用计算机和编程技术自动化一些商业过程的行为。

举例子：水果店老板利用商业智能做出来的报表或仪表盘观测自己商店的人流量、购买量、购买时间，及时调整自己的库存和销售节奏。

过去人们做生意，依靠的是直觉和经验。现在在计算机的帮助下，可以利用数据分析减少试错，减少错误决策带来的成本，明白生意好的因由。而商业智能将这一切尽可能地自动化和简化。

商业智能常常被理解为企业内部现有数据转化为指导商业决策的平台或系统。类似于ERP、CRM 等系统一样的企业级信息化应用。常见的系统有 Business Object、Cognos 和 Hyperon 等。

从企业分工的角度来讲，通常商业智能部（BI）会涵盖大数据产品、数据分析和数据仓库 3 个部分。所以，数据分析仅仅是 BI 中的一个部分。

数据分析应用于各个部门，通常更多是零散的应用和局部的应用；BI 通常是企业级的应用，更宏观。

数据分析通常针对某个问题，运用一定的方法进行分析、归纳、演绎并得出结论；商业智能更多侧重于流程化、规范化和智能化的应用。

数据分析的工具包括 R、SAS 等挖掘工具，也包括 Webtrekk、GA 等统计分析工具，更包含 Excel 等初级工具，只要能实现分析都可以使用；BI 通常包括 SAP、Oracle、甲骨文等大型公司提供的工具，一般小工具都不能应用。

0.2.6 数据分析的内容

数据分析的内容可根据业务需求有所侧重，图 0.2 给出了分析内容的 9 个方面。

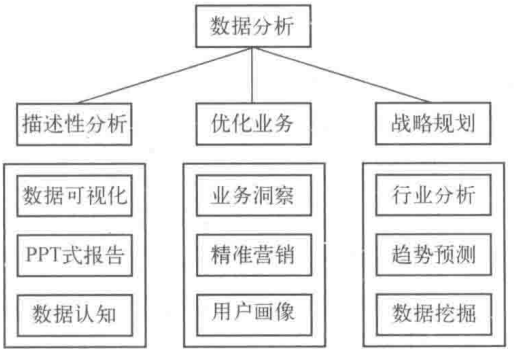


图 0.2 数据分析目标的深度示意图

（1）数据可视化

数据分析不使用图表是难以想象的，数据可视化的作用、技术、工具可参考第 18 章。

下面以客户咨询情况分析为例，说明可视化的必要性：

“在 1205692 件客户咨询中，咨询话音基本业务 423058 次，占咨询总量的 35.09%；咨询新业务 367978 次，占咨询总量的 30.52%；咨询终端 2635 次，占咨询总量的 0.22%；咨询服务及营业网点 99109 次，占咨询总量的 8.22%；咨询网络 26896 次，占咨询总量的

2.23%；咨询卡类业务 7792 次，占咨询总量的 0.65%；咨询计费原则 4636 次，占咨询总量的 0.38%；咨询营销活动 211312 次，占咨询总量的 17.53%；咨询其他业务 62276 次，占咨询总量的 5.16%。”

上面的文字描述可以用图 0.3 表示。

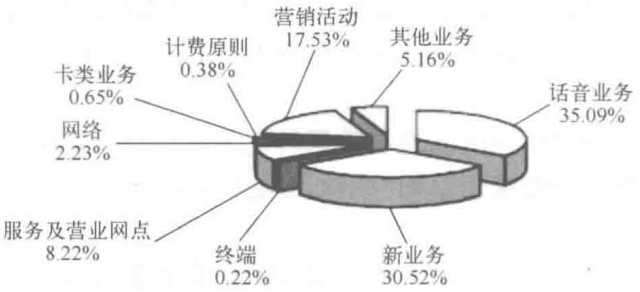


图 0.3　客户咨询情况分析可视化示例

（2）PPT 式报告

在程序员的世界里，讲究“No more talk，Show me the code”，在数据分析师世界里，讲究“Show me the report”。PPT 式报告是了解情况的最简形式，好的数据分析报告是企业决策的重要依据，专业的数据分析报告能体现分析师的职场价值。具体细节参见 0.4.4 节和第 17 章。

（3）数据认知

当拿到一个数据集时，你通常会怎么做？你脑子里好不容易蹦出的那个答案正确吗？这个问题或许能让不少人尴尬。分析与探索是对数据的认知，将遵循如下顺序：数据源质量→数据类型→数据集质量→平均水平→数据分布→量变关系→多维交叉。细节参考 7.2 节和 7.3 节。

（4）业务洞察

业务洞察可以为组织提供快速的评估和路线图，帮助组织识别机遇和规划转型路径以实现其分析举措和目标。业务分析可以通过分析，帮助组织开启实现价值和竞争优势的新途径。

（5）精准营销

精准营销大致意思就是充分利用各种新式媒体，将营销信息推送到比较准确的受众群体中，从而既节省营销成本，又能起到最大化的营销效果。这里的新式媒体，一般意义上指的是除报纸、杂志、广播、电视之外的媒体。

（6）用户画像

用户画像是对现实世界中“用户”的“数学建模”。

一方面是描述用户，没有说人，是说明它跟业务密切相关，它是从业务中抽象出来的，因此来源于现实，高于现实。

另一方面，用户画像是一种模型，是通过分析挖掘用户尽可能多的数据信息得到的，它是从数据中来，但对数据做过了抽象，比数据要高，后面所有用户画像的内容都是基于这个展开的。比如月光族，这个是挖掘分析出来的，不是说原来的数据中包含月光族这个标签。

（7）行业分析

行业是由许多同类企业构成的群体。如果只进行企业分析，虽然可以知道某个企业的经营和财务状况，但不能知道其他同类企业的状况，无法通过比较知道企业在同行业中的位置。而这在充满着高度竞争的现代经济中是非常重要的。另外，行业所处生命周期的位置制约着或决定着企业的生存和发展。

（8）趋势预测

趋势是指市场运动的方向，有三个方向：上升方向、下降方向和水平方向。

趋势的类型（规模）分为：

主要趋势（一年以上）；

次要趋势（三个星期到数月）；

短暂趋势（两三个星期）。

（9）数据挖掘

数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。细节可参考第8~16章。

0.3 数据分析师

0.3.1 什么是数据分析师

数据分析师是一个随着大数据兴起而崛起的新兴的工作岗位，是专门从事行业数据搜集、整理、分析，并依据数据制作业务报告、提供决策、管理数据资产、评估和预测的专业人员。

很多人并不知道数据分析师在做什么？从下面数据分析师和其家人的一段对话就可对这一岗位有所了解。

家人："数据分析？分析什么东西？"

我："哪里有数据，哪里就有我们，什么都可以分析。"

家人："是软件工程师吗？会编程吗？"

我："……不是，不太会。"

家人："那是管理层吗？"

我："还……还不到那个级别。"

家人："那是商务人员？做市场或销售？"

我："……也不是，不过我们辅助他们作决策。"

家人："决策不都是老板说了算吗？你们到底做什么？"

"小陈，你能给我发一个去年一年的汽车品牌页面的访问量吗？最好是以国家、行业、公司规模作为纬度的，浏览量和UV（Unique Visitor，指访问某个站点或点击某条新闻的不同IP地址的人数）都要。"在数据分析师眼中，这样的场景早已司空见惯。

数据分析，被很多部门漏看了"分析"二字，"分析"的本质是对数据敏感。

对数据足够敏感的公司的优势在于，运营过程中产生大量数据，这些数据可以通过一些

手段转化为决策的动力。

产品、营销、销售等部门，都会有不同的需求。例如，产品经理最关心的，是 AB 测试的数据，用以决定产品的效果；营销团队，在乎营销渠道反馈与结果的数据，以便设计下一个营销战略；销售则关心用户的购买率、保留以及追加销售时机等。数据可以直接为其提供服务。

数据分析师到底在做什么呢？

把数据整理地干干净净、整整齐齐，这仅仅是第一步，很多时候，商务部门人员无法直接理解表格数据。那么数据分析师需要把数据通过浅显易懂的图表形式展现出来，如饼状图、曲线图、柱状图等，并给出结论和建议。

相比产品、技术、财务、人力等各个职能明确的部门而言，数据分析师的工作不局限于某一个领域，它更像一个内部咨询机构，它的工作贯穿于公司的业务之中，需要解决每一个部门，乃至高管们提出的分析需求与战略问题。

很多即将步入职场的年轻人也许都想试一试，成为一名看上去高大上的数据分析师，然而心中却不免有一些疑问：

数据分析师对于学历要求是不是很高？

是不是只有统计学、数学专业的人才能做数据分析师？

我是一个文科生，能做数据分析师吗？

我是一个没有工作经验的应届毕业生，能做数据分析师吗？

……

本书会给你答案：

零基础玩转数据分析，皆有可能。

0.3.2 基本要求

数据分析师的基本要求如图 0.4 所示。

从图 0.4 可知，正确的思维习惯、对数据敏感程度，是成为数据分析师的先决条件，其次才是你的“硬件”条件。

1）懂业务。从事数据分析工作的前提就是需要懂业务，即熟悉行业知识、公司业务及流程，最好有自己独到的见解，若脱离行业认知和公司业务背景，分析的结果只会是脱了线的风筝，没有太大的使用价值。业务知识是架起理论和实际应用的桥梁。

假如你在互联网公司工作，却连 PV（Page View，即页面浏览量或点击量，通常是衡量一个网络新闻频道或网站甚至一条网络新闻的主要指标）、UV 为何物都不做功课，未免太粗心了吧。

2）懂管理。数据分析师所面临的工作通常都是以项目形式展开的，数据分析师对自己所参与的项目需要承担对进度、成本和质量的控制。如果不熟悉管理理论，就很难搭建数据分析的框架，对后续的数据分析结论也很难提出有指导意义的分析建议。

3）懂分析。即掌握数据分析基本原理与一些有效的数据分析方法，并能灵活运用到实践工作中，以便有效的开展数据分析。基本的分析方法有对比分析法、分组分析法、交叉分析法、结构分析法、漏斗图分析法、综合评价分析法、因素分析法和矩阵关联分析法等。高级的分析方法有相关分析法、回归分析法、聚类分析法、判别分析法、主成分分析法、因子

此为试读，需要完整PDF请访问：www.ertongbook.com