

# 第 2 章

## 深入认识数据分析



### 本章要点

了解数据形成过程与数据处理

理解字段、记录和数据表

认识 Excel 处理的数据类型

第一步：明确数据分析的目的和思路

第二步：获取需要分析的数据

第三步：对收集的数据进行处理

第四步：分析数据，获得有用信息

第五步：选择合适的的数据呈现方式

第六步：撰写数据分析结果报告

平均数指标

频数与频率指标



### 学习目标

通过对第 1 章的学习，清楚了数据分析这个行业及其相应的职位，那么到底数据分析师如何来分析数据？作为零基础的新手而言，应该从哪里开始入门？在这章及本书的后面章节，都将围绕整个数据分析流程以及最初级的数据分析方法来带领大家快速上手。

| 知识要点        | 学习时间  | 学习难度  |
|-------------|-------|-------|
| 充分理解数据      | 30 分钟 | ★★★   |
| 掌握数据分析的流程   | 50 分钟 | ★★★★★ |
| 认识数据分析的误区   | 20 分钟 | ★★    |
| 了解基本的数据分析指标 | 30 分钟 | ★★★★  |

# 2.1

## 充分理解数据

数据是数据分析工作主要操作的元素，因此，有必要对数据进行一个全面的认识，这样才能从各种繁杂的信息中快速提炼出有用的数据。

### 2.1.1 了解数据形成过程与数据处理

从狭义上来说，数据就是数字，但是从广义上来说，数据是具有一定意义的文字、字母、数字符号的组合、图形、图像、视频及音频等，也是客观事物的属性、数量、位置及其相互关系的抽象表示。那么数据是怎么形成的呢？通常，人们把客观存在的事物以数据的形式保存到计算机中，需要经历从现实世界→信息世界→数据世界这3个过程。

(1) 现实世界。现实世界是客观的、人们头脑之外的世界。它由事物和事物之间的联系组成。将现实世界信息化，就可以进入信息世界。

(2) 信息世界。信息世界是现实世界在人们头脑中的反映，人们把它用文字或符号记载下来。在信息世界中，与数据库技术相关的术语有实体、属性、码、域、实体型、实体集和联系等。信息世界数据化之后，就可以进入数据世界。

(3) 数据世界。数据世界又称为机器世界。信息世界的信息在机器世界中以数据形式存储，在这里，每一个实体用记录表示，相应于实体的属性用数据项（又称字段）来表示，现实世界中的事物及其联系用数据模型来表示。

从3个过程中可以看到，信息是客观事物转变为数据的重要过程，二者之间也是不可分离的：信息依赖数据来表达，数据则生动具体地表达出信息。

对从信息中提取出来的数据进行处理，实际就是对数据进行管理，即对数据进行分类、组织、编码、存储、检索和维护。在计算机系统中，数据管理通常使用数据库管理系统完成，而在信息化的当今社会，数据管理主要是通过数据库技术来完成的。

数据库技术是信息系统的一个核心技术，它是一门综合学科，涉及操作系统、数据结构、算法设计、程序设计和数据管理等多方面知识，它的不断发展使得人们可以科学地组织存储数据、高效地获取和处理数据。

从第1章也可以了解到，作为一个合格的数据分析师，需要掌握的数据处理软件

技能有很多，其实大多数数据分析利用 Excel 进行已经绰绰有余了，因此在本书中主要研究利用 Excel 功能来完成数据处理需要掌握的各种技术。当然，如果你要在数据分析师的道路上越走越远，你所掌握的就应越多，如你懂 Hadoop，那么就可能离大数据更近一点。

### 2.1.2 理解字段、记录和数据表

信息以数据的形式最终得以保存下来，其中离不开字段、记录和数据表这 3 个组成部分，下面从数据分析的角度来理解这 3 个概念。

**字段。**字段是事物或现象的某种特征，如“姓名”“性别”“籍贯”和“居住地址”等都是字段。

**记录。**记录是事物或现象某种特征的具体表现，如姓名可以是张三、李四和王五，性别可以是男性或者女性。

**数据表。**将字段和记录组织在一起形成的关系结构就是数据表。

从概念中可以看到，记录、字段和数据表是一个有机的整体，三者之间相互依赖才得以存在。

在 Excel 表格中，表的每一列称为一个字段，一个记录可以包含若干个互不相同的字段，字段在记录中的先后顺序不影响数据的实际意义。表的每一行数据称为一个记录，一个表可以包含多条互不相同的记录，记录在表中的先后顺序不影响数据的实际意义，如图 2-1 所示。

The figure shows an Excel spreadsheet titled '员工考核成绩表' (Employee Performance Score Table). The spreadsheet has 14 rows and 10 columns. The columns are labeled: 编号 (ID), 部门 (Department), 员工姓名 (Employee Name), 办公应用 (Office Application), 电脑操作 (Computer Operation), 管理能力 (Management Ability), 礼仪素质 (Etiquette Quality), 企业文化 (Corporate Culture), 企业制度 (Corporate System), and 总分 (Total Score). The rows contain data for 14 employees. Annotations with arrows point to specific parts of the table: '字段' (Field) points to the '办公应用' column, '字段' points to the '管理能力' column, '数据表' (Data Table) points to the entire table area, and '记录' (Record) points to the row containing employee 刘星星 (Liu Xingxing).

|    | A     | B   | C    | D    | E    | F    | G    | H    | I    | J   |
|----|-------|-----|------|------|------|------|------|------|------|-----|
|    | 编号    | 部门  | 员工姓名 | 办公应用 | 电脑操作 | 管理能力 | 礼仪素质 | 企业文化 | 企业制度 | 总分  |
| 2  | BH002 | 人事部 | 张嘉利  | 82   | 75   | 73   | 80   | 83   | 80   | 473 |
| 3  | BH003 | 财务部 | 何时韦  | 71   | 91   | 82   | 83   | 70   | 84   | 481 |
| 4  | BH013 | 人事部 | 兰慧芳  | 83   | 75   | 82   | 93   | 77   | 73   | 483 |
| 5  | BH008 | 行政部 | 王豆豆  | 77   | 79   | 94   | 82   | 82   | 72   | 486 |
| 6  | BH001 | 人事部 | 萧天宝  | 90   | 95   | 72   | 84   | 72   | 79   | 492 |
| 7  | BH005 | 财务部 | 钟嘉惠  | 93   | 83   | 71   | 78   | 81   | 86   | 492 |
| 8  | BH009 | 后勤部 | 刘星星  | 79   | 94   | 79   | 78   | 82   | 81   | 493 |
| 9  | BH006 | 后勤部 | 何思伴  | 83   | 80   | 80   | 83   | 88   | 80   | 494 |
| 10 | BH012 | 后勤部 | 杨天雄  | 77   |      | 91   | 77   | 81   | 76   | 494 |
| 11 | BH007 | 行政部 | 高雅婷  | 81   |      | 93   | 83   | 80   | 83   | 501 |
| 12 | BH011 | 财务部 | 邓丽梅  | 76   | 88   | 87   | 91   | 80   | 83   | 505 |
| 13 | BH004 | 后勤部 | 马田东  | 94   | 87   | 85   | 84   | 86   | 71   | 507 |
| 14 | BH010 | 人事部 | 赵大宝  | 86   | 73   | 93   | 93   | 70   | 94   | 509 |

图 2-1

表格是我们经常看到的，但是从数据分析的角度出发，为了方便后期的数据分析操作，在设计数据表时，还需要注意以下几点。

- (1) 在 Excel 中的第一行必须是标题，即全部是字段名称，且不能存在重复字段。
- (2) 从第二行开始，都是记录行。
- (3) 无论是标题行还是记录行，都不能存在合并单元格的情况。
- (4) 数据表的结构要以一维结构存在，不能设置为二维结构。

其实一维表和二维表在表格的外观结构上来看没有什么区别，都是第一行为标题行，第二行还是为记录行，唯一的区别是表格的每一列的字段是否是一个独立的参数。如果每一列都是独立的参数那就是一维表，如果某些列中每一列都是同类参数那就是二维表。下面来看图 2-2 所示表格。

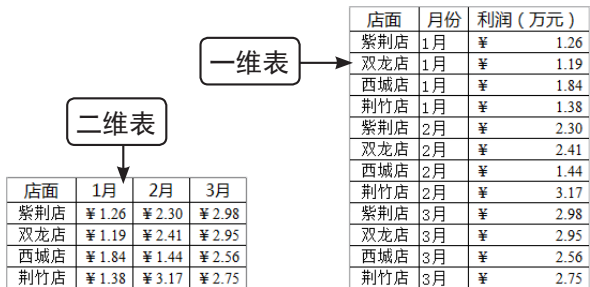


图 2-2

在图 2-2 中，左表就是一个典型的二维表，右表是标准的一维表，因为左侧表格中，1 月、2 月和 3 月都是属于针对月份来进行描述的，因此属于同种参数。一般情况下，通常制作的表格都是按照二维结构来进行存储的，主要是因为这种结构简单清晰，便于查看，但是不利于数据分析，此时就需要将二维结构的表格转化为一维表格，要达到这一目的，在 Excel 中可以利用数据透视表功能来实现，有关数据透视表的具体使用方法将在本书的第 6 章进行详细介绍。

### 2.1.3 认识 Excel 处理的数据类型

Excel 可处理的数据类型有多种，如文本、数值、日期和时间及货币数据等，要看到 Excel 可以处理哪些类型的数据，直接选择任意一个单元格，单击鼠标右键，在弹出的快捷菜单中选择“设置单元格格式”命令，在打开的“设置单元格格式”对话框的“数字”选项卡的“分类”列表框中即可查看到，如图 2-3 所示。

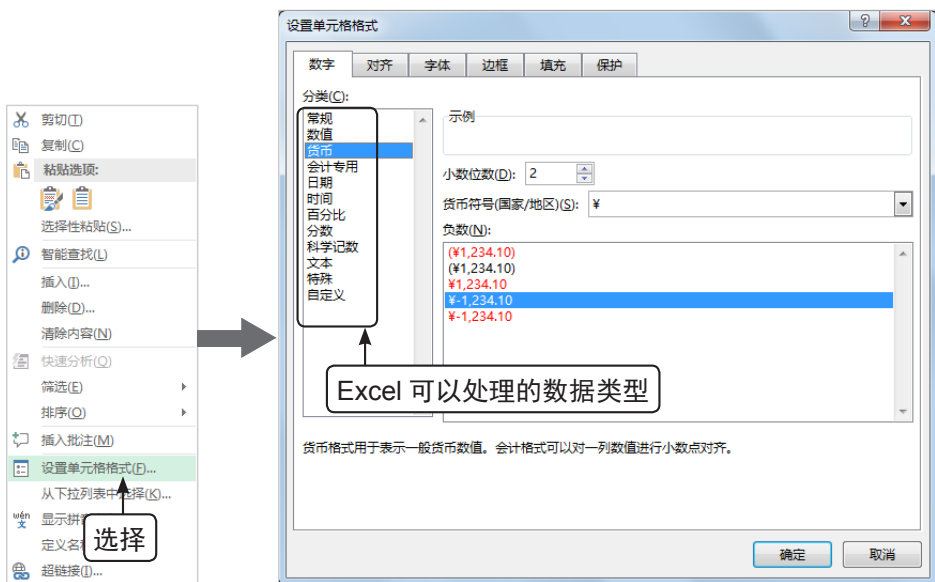


图 2-3

从图 2-3 中可以看到 Excel 能够处理的数据类型有很多，但是最终都可归结为如图 2-4 所示的几种类型，而且在这几种类型中，部分数据之间还存在相互的转化关系，如将数值数据添加上货币符号就转化为货币数据等。

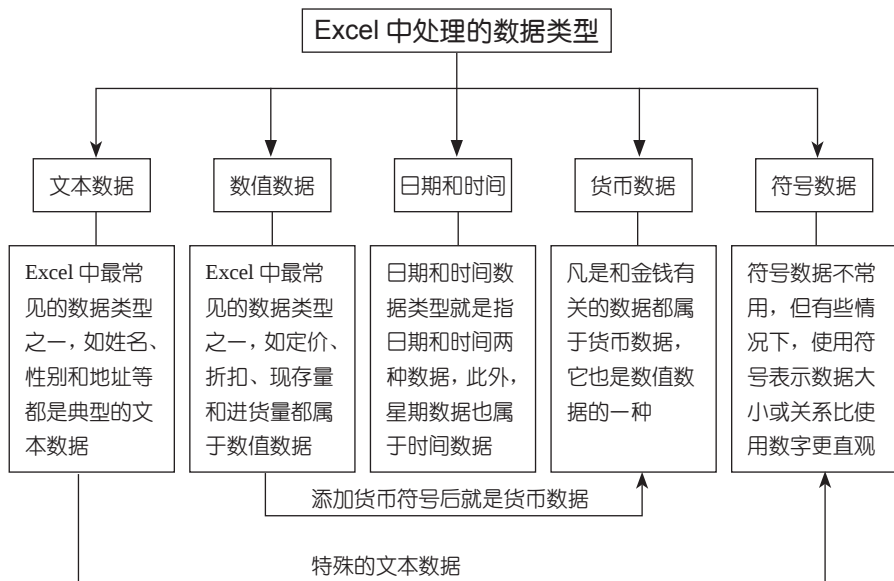


图 2-4

# 2.2

## 掌握数据分析的流程

一个完整的数据分析流程应该包括 6 个步骤，即明确数据分析目的、获取数据、处理数据、分析数据、展现数据和撰写数据分析结果报告。

### 2.2.1 第一步：明确数据分析的目的和思路

对于一只盲目航行的船来说，所有的风都是逆风。

——法国 哈伯特

做任何事情都要求目标明确，当然，数据分析也是一样。专业的数据分析师在进行分析之前通常都会思考：为什么要开展数据分析？通过这次数据分析我需要解决什么问题？一般的数据分析者在进行数据分析之前，更多考虑的是要用什么公式，要用什么图表，图表怎么设计才漂亮等。下面先来看一下图 2-5。

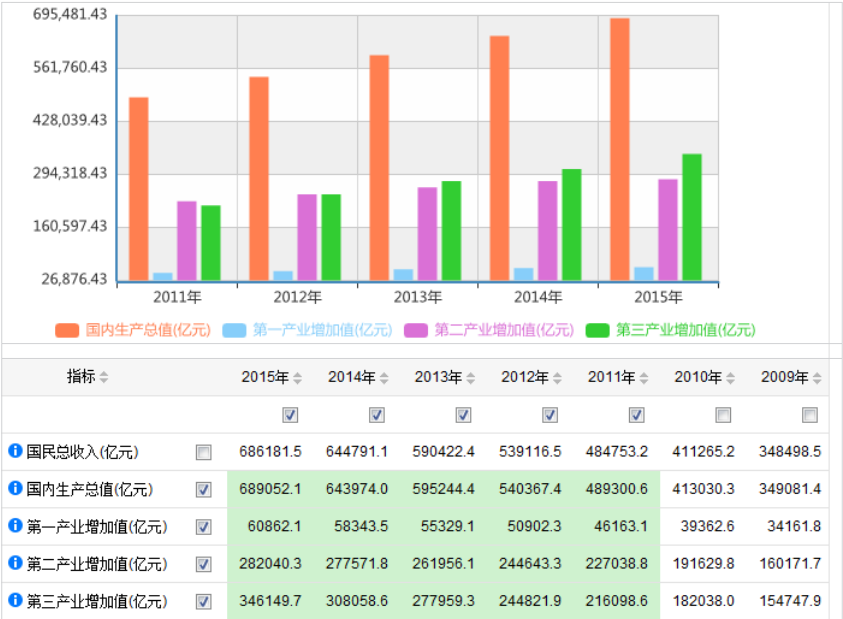


图 2-5

在图 2-5 中显示的是 2011—2015 年国内生产总值和三大产业的统计数据，针对该图表来模拟一下专业数据分析师和一般数据分析者思考问题的方式，如图 2-6 所示。

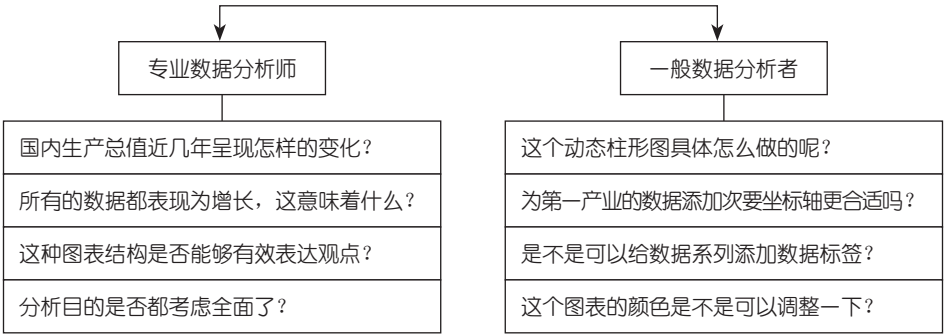


图 2-6

通过这种简单的对比就可以发现，专业的数据分析师是根据分析目的来寻求解决方法或者得到需要的目标结果，其整个过程不会偏离最初预想的方向；而一般数据分析者是更多地考虑外在表现形式，缺乏目标性。即使图表效果设置得再好看，如果与分析目的相违背，那么得到的数据结果是没有任何意义的，还有可能会为决策者提供错误的数据，从而最终造成不可预想的后果。

在确定分析目标后，还应该理清分析思路，搭建整体的分析框架，即将分析目标划分成多个小点，并确定合适的数据分析方法，这样才能更好地指导我们全面地开展数据分析工作，确保数据分析的有效性。

2.2.2 第二步：获取需要分析的数据

数据分析师是基于数据来开展工作的，因此，在进行数据分析之前，首先需要数据，那么数据应该从什么地方来呢？归纳起来可以有 3 个来源，分别是公司内部数据库、市场调查和公开数据源。

1. 从公司内部数据库获取数据

公司从建立以后，发生的各种数据都会保存起来，包括公司的发展规模、业务数据、客户资源和在职员工信息等，日积月累，这些数据会越来越多，也越来越完善，它是数据分析最直接的数据来源。充分利用这些数据信息，可以分析公司过往的发展情况，从

而更好地把握未来。

## 2. 通过市场调查获取数据

市场调查是指运用科学的方法，有目的地系统搜集、记录和整理有关市场营销的信息和资料，分析市场情况，了解市场现状及其发展趋势，为市场预测和营销决策提供客观和正确的资料。

通过这种方法得到数据，其目的性更强，而且数据更准确。因此，公司在制订营销计划之前，都会采用这种方法来获取第一手资料。根据调查目的的不同，市场调查又划分为4种形式，分别是探测性调查、描述性调查、因果关系调查和预测性调查。下面分别进行详细介绍。

**探测性调查。**这种调查一般是针对问题或范围不明确的情况下选择的一种市场调查方法。例如，企业近期销量明显下降，到底是什么原因导致的这个结果无法确定，可以通过探测性调查去找出问题，从而为问题究竟应该如何解决提供进一步的资料及信息。要进行探测性调查，可以从3个方面来获取资料，分别是现存资料、请教有关人士和参考以往类似的实例。

**描述性调查。**描述性调查是市场调查中最常见的形式，如市场潜力研究、市场占有率研究和销售渠道研究等，它是通过找出相关变量来描述调查对象的特征，它可以说明“怎样”或“如何”的问题，但是这种方法得到的数据只能是一种现象描述，并不说明变量中哪个是因，哪个是果，也不解释“为什么”的问题。

**因果关系调查。**因果关系调查是在描述性调查的基础上进行的，它主要是通过各种方法来确定描述性调查中已收集到的变量之间具体存在的关系，以便为解决实际问题提供进一步的数据分析基础。

**预测性调查。**市场需求的估计对每个公司来说都是十分重要的，因为如果公司对未来的市场需求不了解或者估计错误，都会给公司带来很大的风险。预测性调查主要是根据描述性调查与因果关系调查所提供的资料，来建立预测分析的模型，从而最终对市场进行评估。

对于第一手资料的获取，所耗费的金钱和时间都是相当巨大的。但是这些资料对数据分析师而言，是非常重要的数据。公司要获得这类数据，可以委托专业的调查公司来进行，也可以设立市场研究部门，专门负责此项工作。需要特别提醒，市场调查数据也会存在误差，因此其数据仅供参考。

### 3. 从公开数据源获取数据

公开数据源的来源有两个方面：一是公开发行的刊物，如“中国统计年鉴”“世界经济年鉴”和“世界发展报告”等；二是网络，如国家相关部门统计信息、各种信息查询平台以及专业的第三方机构等。

尤其随着互联网的发展，从网络上获取公开的数据源越来越方便，下面列举一些常见的信息查询平台及专业的第三方数据统计机构供读者参考。

(1) 中国国家统计局网站 (<http://www.stats.gov.cn/tjsj/>)。它主要包括国家经济宏观数据、社会发展和民生相关重要数据及信息，非常全面且定期发布统计出版刊物，实用性强，其统计数据页面效果如图 2-7 所示。



图 2-7

(2) 中国经济数据库 (<https://www.ceicdata.com/zh-hans/products/china-economic-database>)。这里介绍的中国经济数据库是由司尔亚司数据信息有限公司 (CEIC) 提供的，该公司成立于 1992 年，公司提供的有关世界发达经济体和发展中经济体的信息是最广泛且最精确的。不论是用于研究还是用于投资，该公司的“中国数据库”都可以为用户提供大量有关中国市场各种部门和行业经济活动的信息，其页面效果如图 2-8 所示。



图 2-8

(3) 国家数据平台 (<http://data.stats.gov.cn/index.htm>)。它的数据来自国家统计局，但排版更清晰简洁，包括国计民生各个方面的月度数据、季度数据、年度数据、各地区数据、部门数据以及国际数据，其页面效果如图 2-9 所示。



图 2-9

(4) 艾瑞网 (<http://www.iresearch.cn/>)。它是艾瑞咨询精心打造的国内首家新经济门户网站，其定期发布互联网相关数据及报告，在国内的互联网咨询服务方面报告相对出色。该网站的页面效果如图 2-10 所示。



图 2-10

### 2.2.3 第三步：对收集的数据进行处理

数据处理的基本目的是从大量的、可能是杂乱无章的、难以理解的数据中抽取并推导出对于某些特定的人们来说是有价值和有意义的信息，并将其形成适合数据分析的样式。

在进行数据分析之前，首先要做的就是数据处理。如果数据处理工作不到位，则会影响数据的分析方向。下面先看一个案例。

#### 案例陈述

2017 年 3 月 14 日，国家统计局发布了关于 1—2 月份的工业生产数据，具体资料如下。

2017 年 1—2 月份，全国规模以上工业增加值同比实际增长 6.3%，增速较上年 12 月份加快 0.3 个百分点，延续上年以来稳中向好的运行态势。工业生产增速有所加快的主要原因有以下几个方面。

一是工业出口加快。1—2 月份，规模以上工业出口交货值同比增长 8.8%( 上年同期为下降 4.8%)，较上年 12 月份加快 4.9 个百分点。其中，汽车、电子行业出口交货值同比分别增长 18.1% 和 11.1%，较上年 12 月分别加快 11.1 和 4.5 个百分点。

二是行业面、产品面总体向好。1—2 月份，41 个工业大类行业中，增速较上年 12 月份加快或降幅收窄的行业有 31 个，占比达到 75.6%；主要工业产品产量中，增速较上年 12 月份加快或降幅收窄的产品占比达到 55.9%。其中，光电子器件、锂离子电池、工业机器人、SUV 和集成电路等新兴产品产量保持了较高的增速，工程机械和载货汽车等产品产量受益于基础设施建设继续保持快速增长。

三是装备制造业和高技术产业支撑作用明显。1—2 月份，装备制造业和高技术产业增加值分别同比增长 11.9% 和 12.6%，增速分别高于规模以上工业 5.6 和 6.3 个百分点，较上年 12 月份分别加快 1.8 和 0.4 个百分点。

总体来看，今年以来，工业产品供需状况继续改善，工业产品价格总体稳中有升，企业预期向好，呈现新兴产业继续较快增长和部分传统产业有所复苏的良好态势。

在以上这段数据源中，给我们传递的信息量非常大，既有对总体工业生产数据变化的统计，也有对影响工业生产增速加快的原因分析，并且每个方面的分析还有具体的数据。如果不根据分析目的来处理这段资料中的数据，那么有可能导致数据分析的侧重点和方向错误。

当然，将第一手资料处理为适合数据分析样式的数据是最基本的数据处理方法，在数据处理过程中，有时还会对数据进行二次处理，如清理重复数据、检查数据的完整性以及对数据进行计算等，以此来确保数据的正确、完整和有效。

#### 2.2.4 第四步：分析数据以获得有用信息

数据分析是在数据处理的基础上，利用合适的数据分析方法和工具，将处理的数据进行深入分析，从而得到有价值的信息，并形成最终结论的过程。

数据分析的方法和工具很多，数据分析师在进行数据分析之前，首先要根据分析目的选择最合适的数据分析方法，并确定用什么工具来完成，做足准备后，在正式进入到数据分析过程时，才能够从容地进行数据分析和研究。

一般情况下的数据分析工作，直接利用 Excel 工具就可以完成，如最值数据的分析、筛选符合条件的数据和透视分析数据等，而复杂的数据分析就需要采用专业的数据分析

软件来完成。

需要特别说明的是，这里所说的数据分析并不是数据挖掘，虽然二者看起来都是通过已有的数据得到结果，但是二者之间也存在很大的区别，详见表 2-1 所列。

表 2-1 数据分析与数据挖掘的区别

| 区别项     | 数据分析                                    | 数据挖掘                                       |
|---------|-----------------------------------------|--------------------------------------------|
| 分析的数据量  | 数据量不会太大                                 | 数据量极大                                      |
| 分析是否有约束 | 有约束，在进行数据分析之前，首先要提出某个假设，然后按照假设来建立方程或者模型 | 没有约束，因此在进行数据挖掘时，不需要假设，直接将手上的资料往有用的方向进行收集即可 |
| 分析的对象   | 一般都是数字化的数据                              | 除了数字化的数据，还可以分析声音和图形等                       |
| 分析的目标   | 目标清晰                                    | 目标不清晰，要依靠挖掘算法来找出隐藏在大量数据中的规则、模式和规律等         |
| 分析的结果   | 分析结果明确                                  | 结果不是特别容易解释，因为挖掘出来的数据是对信息的价值评估及其着眼于未来的预测    |

通过表 2-1 可以发现，数据分析比数据挖掘更简单，作为初学者而言，需要先掌握数据分析，再进一步来学习数据挖掘。

2.2.5 第五步：选择合适的数据呈现方式

数据分析完成后，将其以可视化的方式呈现出来是数据分析师必须具备的能力，那么如何确保数据的呈现方式最合适呢？可以从以下几个方面来考虑。

(1) 工具如何选择。PowerPoint( 可缩写为 PPT)、Excel 和 Word 都是不错的展现工具，任意一个工具用好都很强大，但是具体怎么选择还需要看演示场景。一般情况下，在大型会议上展示数据则 PPT 最合适，如果汇报说明则 Word 最实用，数据较多时 Excel 更方便。本书主要介绍如何用 Excel 中提供的各种图形化表达功能来让数据直观和清晰地展示出来。

(2) 受众对象是谁。根据目标对象选择数据的呈现方式，一般而言，领导层喜欢读图、看趋势和要结论，因此数据呈现方式应重点突出结果，而执行层喜欢看数字、读文字和看过程，因此数据呈现方式要注重过程尽量阐述得详细和细致。

(3) 形式有哪些。俗话说“文不如表，表不如图”，图形化的表达是最能让数据直观展示的方式，其表现效果不仅清晰，而且生动，是最容易理解的一种表现形式，如图 2-11 所示。

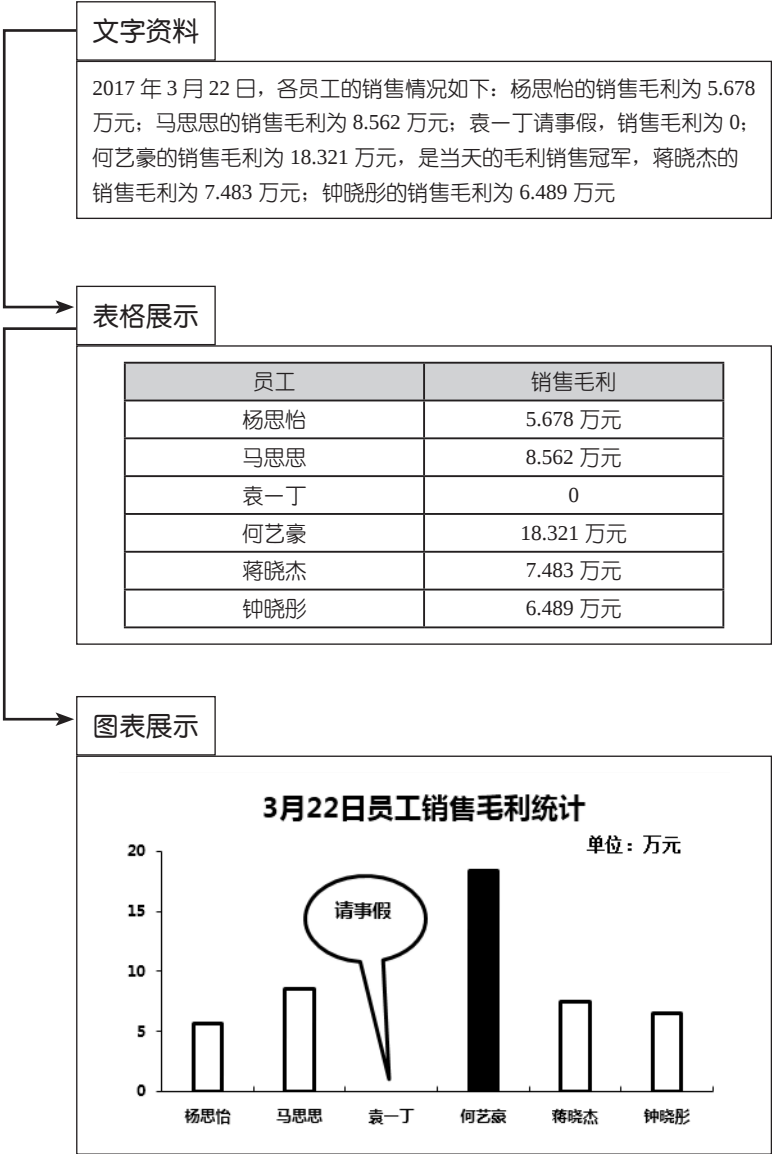


图 2-11

最重要的一点是，数据展现永远辅助于数据内容，有价值的数据报告才是关键。

### 2.2.6 第六步：撰写数据分析结果报告

我们平常理解的数据分析工作完成是得到数据分析结果，但是正规情况下，一轮完整的数据分析工作，是以完成数据分析结果报告的撰写为标志。报告写完，即宣告此次数据分析工作完成。

数据分析报告是整个数据分析过程的总结。通过撰写数据分析报告，可以将数据分析工作的起因、采用的数据分析方法和数据分析工具、数据分析的过程、得到的数据分析结果以及根据数据分析结果给出的建议完整地呈现出来，让决策者全方位地了解公司的运营情况，对公司的运营质量进行更准确的评估，从而依据严谨的数据分析结果来制订更科学的运营计划。

以上就是数据分析的全流程，一次科学完整的数据分析工作必须包括这6个步骤，在本节中只是对每个过程进行了一个粗略的介绍，在本书的后面将以 Excel 为数据分析工具，按照这个流程详细介绍每个过程涉及的数据分析技术。

## 2.3

### 认识数据分析的误区

数据分析过程中也会存在很多的分析误区，要确保数据的正确性和客观性，数据分析师就要避免进入这些分析误区，下面具体来认识一下数据分析有哪些误区。

#### 1. 以点代面进行数据分析导致结果错误

以点代面是指用某一类型的数据代替了全部类型的数据。在进行数据分析时，要切忌用这种以点代面的手段，因为这样得到的数据结果具有片面性，最终得到的分析结果有可能就是错误的。下面通过一个小案例来加深理解。

#### 案例陈述

在第二次世界大战时期，英国空军提出希望增加飞机的装甲厚度的想法，但是考虑到对飞机全部装甲加厚会降低飞机飞行的灵活性，所以最终决定只增加受攻击最多部位的装甲。后来相关工作人员对中弹飞机进行了统计分析，发现大部分飞机的机翼弹孔比较多，所以得到受攻击最多的部位是机翼，最终决定只增加机翼的装

甲厚度。而事实上，如果机头中弹的飞机几乎就被摧毁了，并没有飞回来。

上面这个案例就是典型的根据部分数据代替全部数据来分析，正确的统计分析应该是对全部飞机进行分析，损毁的飞机虽然不存在了，但它仍然是全面进行数据分析的重要数据源。缺少了这部分数据，得到的数据结果就会是错误的。

因此，数据分析师在统计数据和分析数据时，一定要时刻想着是否全面考虑了各种问题和情况，只有多考虑多重视这个问题，才能提高分析结果的正确性。

## 2. 鲜明数据误导我们将偶然因素放大

有些数据分析师在分析数据时，如果分析的某个因素比较鲜明，则会放大这个因素带来的影响，这种分析方法是不正确的。比如你身边有 10 个朋友在炒股，赚钱的有 8 个，是否就说明炒股赚钱的概率就是 80% 呢？其实不然，事实上散户炒股的整体概率是 8 赔 1 平 1 赚，即 80% 的散户炒股是赔钱，只有 10% 的散户炒股在赚钱。

因此，在分析数据时，不能被太过鲜明的偶然事件或者数据所蒙蔽；否则你会忽略背后一直存在的整体概率。当遇到这种情况后，分析师一定要保持冷静，多参考一些历史数据或平均情况，切忌放大偶然因素而轻易做出判断。

## 3. 把简单问题复杂化

在分析数据时，明明有现成的、简单的且很适用的方法，但是很多数据分析师却不采用，而是把时间花费在数据算法上，如利用回归分析和因子分析等高级分析方法来进行数据分析，他们总觉得有分析模型的才是专业的，结果才可信，从而将简单的问题复杂化。其实不然，高级的数据分析方法不一定是最好的，能够以简单的方法有效解决问题才是最好的。

## 4. 业务知识与数据分析联系不上

做数据分析工作的人一般是统计学、计算机或者数学专业出身的，对于算法技术非常熟悉，但是他们大多都缺乏从事营销和管理方面的工作经验，在进行数据分析工作时，不能很好地将业务知识与统计分析结合起来，从而导致分析结果都是一些零散的数据，虽然数据分析报告做得很专业，但是实用性不强。

因此，作为数据分析师，要多结合业务方面进行分析，才能得出更加切合实际的数据分析结果，为决策层制订运营计划提供有力的支撑。这也更加说明了合格的数据分析师必须要懂业务和管理。

以上简单介绍了几种常见的数据分析误区。数据分析是复杂庞大的工程，每一个细节都会关系到公司业务的发展，随着业务变化，数据的变化也就增大，遇到的问题也会增多，所以作为数据分析师保持独立思考、不武断，多次验证才是重要的。

## 2.4

## 了解基本的数据分析指标

数据分析指标是数据分析过程中的重要元素。透过数据分析指标可以帮助数据分析师打开思路，将抽象的分析具体化，从多个角度对数据进行深度解读。

表 2-2 和表 2-3 分别为互联网行业与银行业在进行数据分析时涉及数据分析指标。

表 2-2 互联网行业进行数据分析时涉及的指标

| 流量类指标   | 订单产生效率指标 | 总体销售业绩指标 | 客户价值类指标  | 市场营销活动指标 | 广告投放指标   |
|---------|----------|----------|----------|----------|----------|
| 独立访客数   | 总订单数量    | 成交金额     | 累计购买客户数量 | 新增访问人数   | 新增访问人数   |
| 页面访问数   | 访问到下单转化率 | 销售金额     | 新客户数量    | 新增注册人数   | 新增注册人数   |
| 人均页面访问数 | —        | 客户单价     | 新客户获取成本  | 总访问次数    | 总访问次数    |
| —       | —        | 销售毛利     | 消费频率     | 订单数量     | 订单数量     |
| —       | —        | 毛利率      | 最近一次购买时间 | 下单转化率    | UV 订单转化率 |
| —       | —        | —        | 重复购买率    | ROI      | 广告投资回报率  |

表 2-3 银行业进行数据分析时涉及的指标

| 业务发展效率指标    | 业务发展质量指标  | 业务发展效益指标           | 资产负债比指标  | 经营状况类指标 | 经营成果类指标            |
|-------------|-----------|--------------------|----------|---------|--------------------|
| 个人网上银行客户渗透率 | 个人网银客户动户率 | 柜面交易替代情况           | 备付金比例    | 流动比率    | 利润率                |
| 企业网上银行客户渗透率 | 企业网银客户动户率 | 电子渠道交易 / 收入占比分析    | 资产流动性比例  | 速动比率    | 资本金利润率             |
| 手机银行客户渗透率   | 手机银行客户动户率 | 电子渠道交易 / 收入趋势分析    | 存贷比例     | 资本风险比率  | 成本率                |
| 支付宝客户渗透率    | 自助设备开机率   | 各支行电子渠道交易 / 收入排名分析 | 对流动负债依存率 | 固定资产比率  | 各支行电子渠道交易 / 收入排名分析 |
| 银联在线客户渗透率   | 自助设备空钞率   | 电子渠道营业概况地理位置分析     | 中长期贷款比例  | —       | 综合费用率              |

续表

| 业务发展效率指标     | 业务发展质量指标    | 业务发展效益指标 | 资产负债比指标 | 经营状况类指标 | 经营成果类指标 |
|--------------|-------------|----------|---------|---------|---------|
| 电话 POS 客户渗透率 | 自助设备长短款及吞卡数 | —        | 拆借资金比例  | —       | —       |
| —            | —           | —        | 资本利润率   | —       | —       |

从表 2-2 和表 2-3 中可以看到，行业划分得越细致，其使用的数据分析指标越具有行业特性，而在本节中，主要是从统计学的角度介绍一些常见的数据分析指标。

### 2.4.1 平均数指标

一般情况下，提到平均数都是指的算术平均数，又称为均值，它是统计学中最基本和最常用的一种平均指标。算术平均数又分为几何算术平均数和加权算术平均数。下面分别进行介绍。

#### 1. 几何算术平均数

几何算术平均数是指全部数据累加除以数据的个数。例如，某销售小组有 5 名销售员，“五一”节的销售额分别为 5027 元、7654 元、8452 元、6754 元和 15040 元，求该日平均销售额。

平均销售额  $= (5027 + 7654 + 8452 + 6754 + 15040) / 5 = 8585.4$  (元)

#### 2. 加权算术平均数

加权算术平均数主要用于处理经分组整理的数据。设原始数据被分成  $K$  组，各组的组中的值为  $X_1, X_2, \dots, X_k$ ，各组的频数分别为  $f_1, f_2, \dots, f_k$ ，加权算术平均数的计算公式为

$$M = \frac{X_1 \times f_1 + X_2 \times f_2 + \dots + X_k \times f_k}{f_1 + f_2 + \dots + f_k}$$

例如，某人射击 10 次，其中一次射中 10 环，两次射中 9 环，3 次射中 8 环，4 次射中 7 环，求这个人平均射中的环数。

平均射中的环数  $= (10 \times 1 + 9 \times 2 + 8 \times 3 + 7 \times 4) / 10 = 8$  (环)

### 2.4.2 频数与频率指标

频数也称为次数，频率又称为相对次数。简单地理解，在一组数据中，每个数据

出现在某个类别中的次数称为频数，而每个数据出现的次数与总次数的比值称为该数据的频率，即表示该类别在总数中出现的频繁程度。下面通过一个案例来进行了解。

案例陈述

随机抽取某城市 30 天的空气质量状况，统计结果如表 2-4 所示，现在要分析在这 30 天中空气质量状况为优的频率是多少。

表 2-4 空气质量状况统计

|          |    |    |    |     |     |     |
|----------|----|----|----|-----|-----|-----|
| 污染指数 $W$ | 40 | 70 | 90 | 110 | 120 | 140 |
| 天数 $t$   | 3  | 5  | 10 | 7   | 4   | 1   |

其中， $W \leq 50$  时，空气质量为优； $50 < W \leq 100$  时空气质量为良； $100 < W \leq 150$  时，空气质量为轻度污染。根据统计表的数据分析出这 30 天空气质量状况如表 2-5 所示。

表 2-5 空气质量状况分析

| 空气状况 | 天数 | 频数 | 频率          |
|------|----|----|-------------|
| 优    | 3  | 3  | $3/30=0.1$  |
| 良    | 15 | 15 | $15/30=0.5$ |
| 轻度污染 | 12 | 12 | $12/30=0.4$ |

通过分析可以知道，在这 30 天中，空气质量状况为优的频率为 0.1。

2.4.3 绝对数与相对数指标

绝对数是统计中常用的总量指标，它是反映客观现象总体在一定时间和地点条件下的总规模和总水平的综合指标，是现象“量”的具体表现。

相对数在统计中又称为比例指标，它是指两个有联系的指标对比计算得到的数值，用以反映客观现象之间的数量联系程度。相对数根据相互对比指标的性质和所能发挥的作用不同，又可分为结构相对数、比例相对数、比较相对数、强度相对数、计划完成程度相对数和动态相对数，分别介绍如下。

(1) 结构相对数：将同一总体内的部分数值与全部数值对比求得比例，用以说明事物的性质、结构或质量。

(2) 比例相对数：将同一总体内不同部分的数值对比，表明总体内各部分之间的比例关系。

(3) 比较相对数：将同一时期两个性质相同的指标数值对比，说明同类现象在不同空间条件下的数量对比关系。

(4) 强度相对数：将两个性质不同但有一定联系的总量指标对比，用以说明现象的强度、密度和普遍程度。

(5) 计划完成程度相对数：这是某一时期实际完成数与计划完成数对比，用以说明计划完成程度。

(6) 动态相对数：将同一现象在不同时期的指标数值对比，用以说明发展方向和变化的速度。

下面通过一个案例来详细理解绝对数和相对数概念。

## 案例陈述

---

甲和乙来自同一个村子，两人大学毕业后甲选择了在大城市里面打拼，而乙选择了回农村教书。一晃五年过去了，乙始终坚持在自己的教书岗位上，领着每个月800元的工资，即使这样每月还是会拿出200元的工资来资助家庭贫困的孩子或者购买教学用具。

在去学校的路上有一条河，河上没有桥，每逢下雨涨水，乙还会挨个背着孩子们过河上学，五年以来从不间断。甲在大城市也是努力拼搏，五年过去了，也小有成就，赚取了100万元的资产，他捐出了10万元在河上修建了一座桥。后来村民在桥头修了一座石碑以纪念甲的善举。

我们从绝对数的角度来分析这件事，甲付出的是10万元，乙付出的是他五年工资的 $\frac{1}{4}$ ，一共付出 $200 \times 12 \times 5 = 12000$ （元），从数据上来看，显然甲的付出比乙的大，因此村民在桥头修石碑纪念甲是合理的。

现在我们从相对数的角度来分析这件事，甲虽然付出10万元，但仅仅是他全部财富的 $\frac{1}{10}$ ，虽然乙只付出了12000元，但是这是他全部财富的 $\frac{1}{4}$ ，除了金钱上的支持以外，在甲修桥之前，乙在孩子们身上付出的精力更多，由此来判定乙

的贡献远远大于甲，因此村民在桥头修石碑更应该纪念乙。

在现实的数据分析中，人们往往更多的是以绝对数的角度来分析事情，而忽视那些默默付出的人，这种做法是不全面的。

相对数和绝对数指标像一对龙凤胎儿女，男孩女孩都同样重要，因此数据分析要将绝对数和相对数结合起来综合考虑，这样分析结果才更准确。就如上面的案例，甲和乙都应该值得被纪念，因为他们都付出了，我们无法精确地衡量到底谁付出得更多，因为分析的角度不同，得到的结果就不同。

2.4.4 其他常见数据分析指标

除了前面介绍的一些基本指标以外，还有一些相对比较容易理解的数据分析指标，具体见表 2-6。

表 2-6 其他常见数据分析指标

| 指标  | 描述                                                     | 举例                                                                     |
|-----|--------------------------------------------------------|------------------------------------------------------------------------|
| 峰值  | 增长曲线的最高点 ( 顶点 )                                        | 据有关部门预测，中国总人口 2033 年将达峰值 15 亿                                          |
| 拐点  | 在数学上指改变曲线向上或向下方向的点。在统计学中指趋势开始改变的地方，出现拐点后的走势将保持基本稳定     | 经过了 2007 年的股市和楼市疯涨后，在 2008 年出现拐点，楼市伴随着股市大跳水                            |
| 增量  | 也称增长量，指数值的变化方式和程度                                      | 3 增大到 5，则 3 的增量为 +2 3 减少到 1，则 3 的增量为 -2                                |
| 增速  | 也称增长速度，是反映社会经济现象增长程度的相对指标，它是报告期增长量与基期发展水平之比            | 以 2016 年的国民生产总值数字与 2015 年的国民生产总值数字对比以计算这段时间的增速时，2016 年即为报告期，2015 年即为基期 |
| 百分比 | 百分比表示一个数是另一个数的百分之几的数，也叫百分率，通常用于表达其比例关系                 | 0.05 和 0.2 分别是数，可分别转化为百分数，即 5% 和 20%；另外，20% 是 5% 的 4 倍，用百分比表示则是 400%   |
| 百分点 | 用以表达不同百分数之间的“算术差距” ( 即差 ) 的单位                          | 20% 比 5% 多出 15%，用百分点描述为 20% 比 5% 多 15 个百分点                             |
| 倍数  | 属于相对数，是指一个数除以另一个数所得的商。倍数一般是表示数量的增长或上升幅度，不适用于表示数量的减少或下降 | $A \div B = C$ ，A 就是 C 的倍数                                             |

续表

| 指标 | 描述                                                                                                         | 举例                                                                                                                     |
|----|------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| 番数 | 属于相对数，是指原来数量的 2 的 $N$ 次方                                                                                   | 如原始数据为 $A$ ，翻一番就是原来数的 $2^1=2$ 倍，即 $A \times 2$ ；翻二番就是原来数乘以 $2^2=4$ ，即 $A \times 4$ ；翻三番就是原来数乘以 $2^3=8$ ，即 $A \times 8$ |
| 比例 | 属于相对数，是指总体中各部分的数值占全部数值的比例，通常反映总体的构成和结构                                                                     | 技术部共有 20 人，本科学历的员工 12 人，本科以上学历的人 8 人，则本科学历的员工比例为 $12:20=3:5$                                                           |
| 比率 | 属于相对数，是指不同类别数值的对比，它反映的不是部分与整体之间的关系，而是一个整体中各部分之间的关系。这一指标经常会用在社会经济领域                                         | 技术部共有 20 人，本科学历的员工 12 人，本科以上学历的人 8 人，则本科学历的员工与本科以上学历的员工的比率为 $12:8=3:2$                                                 |
| 同比 | 是指与历史同时期进行比较得到的数据，即指与上年的同期水平对比。该指标主要反映的是事物发展的相对情况，其计算公式为：( 某个指标的值 - 同期这个指标的值 ) / 同期这个指标的值 $\times 100\%$   | 2017 年 3 月公司订单量为 50 笔，2016 年 3 月公司订单量为 38 笔，则订单量同比增长 $(50-38)/38 \times 100\% = 31.6\%$                                |
| 环比 | 是指与前一个统计期进行比较得到的数值，即指与同一年连环的两期对比。该指标主要反映的是事物逐期发展的情况，其计算公式为：( 某个指标的值 - 上期这个指标的值 ) / 上期这个指标的值 $\times 100\%$ | 2017 年 3 月公司订单量为 50 笔，2017 年 2 月公司订单量为 45 笔，则订单量环比增长 $(50-45)/45 \times 100\% = 11.1\%$                                |